WHO RUNS THE WORLD: DATA

Editors
Sevinç GÜLSEÇEN, Sushil SHARMA, Emre AKADAL

İSTANBUL UNIVERSITY PRESS

WHO RUNS THE WORLD: DATA

EDITORS

Prof. Dr. Sevinç GÜLSEÇEN
Prof. Dr. Sushil SHARMA
Dr. Emre AKADAL

İSTANBUL UNIVERSITY PRESS

Published by
Istanbul University Press
Istanbul University Central Campus
IUPress Office, 34452 Beyazıt/Fatih
Istanbul - Turkey

www.iupress.istanbul.edu.tr

WHO RUNS THE WORLD: DATA
Editors: Sevinç Gülseçen, Sushil Sharma, Emre Akadal

e-ISBN: 978-605-07-0743-4

DOI: 10.26650/B/ET06.2020.011

Istanbul University Publication No: 5273
Informatics Department Publication No: 5

Published Online in December, 2020

It is recommended that a reference to the DOI is included when citing this work.

This work is published online under the terms of Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/

This work is copyrighted. Except for the Creative Commons version published online, the legal exceptions and the terms of the applicable license agreements shall be taken into account.

CONTENTS

CHAPTER 1
MATERIAL INFORMATION CARRIERS: HISTORICAL DEVELOPMENT
Róbert JÁGER 1

CHAPTER 2
ASTRONOMICAL DATA
Hulusi GÜLSEÇEN, Hasan H. ESENOĞLU 13

CHAPTER 3
DATA STORAGE IN THE DECENTRALIZED WORLD:
BLOCKCHAIN AND DERIVATIVES
Enis KARAARSLAN, Enis KONACAKLI 37

CHAPTER 4
DATA IN THE CONTEXT OF INDUSTRY 4.0
Fatma Önay KOÇOĞLU, Denizhan DEMİRKOL 71

CHAPTER 5
BIG DATA GOVERNANCE
Malgorzata PANKOWSKA 93

CHAPTER 6
A CORE PROBLEM WITH HUMAN DATA PROCESSING: EPISTEMIC
CIRCULARITY IN ACTION
Mehmet Selim DERİNDERE 107

CHAPTER 7
DATA PRE-PROCESSING IN TEXT MINING
Tuğçe AKSOY, Serra ÇELİK, Sevinç GÜLSEÇEN 123

CHAPTER 8
BIG DATA IN EDUCATION: A CASE STUDY ON PREDICTING E-LEARNING
READINESS OF LEARNERS WITH DATA MINING TECHNIQUES
Zeki ÖZEN, Elif KARTAL, İlkim Ecem EMRE 145

CHAPTER 19
THE VALUE OF DATA FOR IMPROVING EFFECTIVENESS OF
CAMPUS COURSES: THE CASE OF HYBRID MOOCS
Oğuz AK, Selim YAZICI, Sevinç GÜLSEÇEN 165

CONTENTS

CHAPTER 10
INTELLIGENT TUTORING OF LEARNERS IN E-LEARNING SYSTEMS AND MASSIVE OPEN ONLINE COURSES (MOOC)
Yacine LAFIFI, Asma BOUDRIA, Atef LAFIFI, Moadh CHERAITIA 177

CHAPTER 11
SMART HOUSE: DATA GATHERING AND ANALYSIS
Natalija LEPKOVA 193

CHAPTER 12
PRIVACY FOR ENTERPRISES IN THE DATA AGE
Bilgin METİN, Enes YILMAZ, Erdi ŞEKERCİLER 209

CHAPTER 13
DATA COLLECTION APPROACHES FOR ARTIFICIAL INTELLIGENCE APPLICATIONS IN HEALTHCARE
Murat GEZER, Çiğdem SELÇUKCAN EROL 227

CHAPTER 14
THE TECHNOLOGICAL TRANSFORMATION PROCESS FROM ELECTRONIC INTELLIGENCE TO CYBER INTELLIGENCE
Ahmet Naci ÜNAL 239

CHAPTER 15
AUTOMATIC MEASUREMENT OF THE MORPHOLOGICAL CHARACTERISTICS OF HONEYBEES WITH A COMPUTATIONAL PROGRAM
Zlatin ZLATEV, Veselina NEDEVA, Ivanka ZHELYAZKOVA 261

ABOUT THE CONTRIBUTORS

Dr. Sevinç Gülseçen, is a Professor and Head of the Department of Informatics at Istanbul University and also Director of Computer Applications and Research Center, Istanbul, Turkey. She has her bachelor degree in math and astronomy from Faculty of Natural Sciences and her doctoral degree in artificial neural networks from Faculty of Business, Istanbul University.

Dr. Gulsecen has her primary teaching and research interests in system analysis and design, constructivist learning, e-learning, computer-mediated communications, e-government, community and social informatics, and knowledge management. She has published her research in more than 50 papers in several national and international journals, conference proceedings and edited books such as International Journal of E-adoption, IEEE Technology and Society Magazine, Educational Technology and Society, Euroasia Journal of Mathematics, Science and Technology Education, International Review of Research in Open and Distance Learning, Education and Information Technologies, Technologies for Enhancing Pedagogy, Engagement and Empowerment in Education: Creating Learning-Friendly Environment.

Dr. Gülseçen has been member of Istanbul University International Academic Relations Board, Istanbul University Institute of Science Academic Board and Turkish Informatics Society. Dr. Gulsecen has served on the editorial board of several refereed journals and has reviewed manuscripts and conference proceedings. She has served as a moderator and session chair at national and international conferences. She was a Turkey delegate to USA, England, Italy, Germany, Poland, Slovakia, Lithuania, Bulgaria and China for her professional meetings and work. She has also worked with business professionals for her consulting and research.

Dr. Gülseçen is a co-founder and co-chair of the international conference named "FutureLearning: Innovations in Learning for the Future".

Dr. Sushil Sharma, is currently an associate dean and a professor of computer information systems (CIS) in the Miller College of Business at Ball State University. Dr. Sharma has co-authored/edited/co-edited twelve (12) books and published over ninety (90) refereed research papers in the most reputed national and international information systems' journals. In addition, he has published forty-five (45) refereed chapters in various books and has presented and published over 160 papers in various national and international conferences. Dr. Sharma's research has appeared in several highly ranked journals in the MIS field, including Decision Support Systems, Communications of the Association for Information Systems, European Journal of Information Systems, Information Systems Frontiers, Journal of Information Privacy & Security (JIPS), Electronic Commerce Research Journal, and Information Management and Computer Security.

His primary research interests are in computer information systems security, e-Learning, e-Government, computer-mediated communications, human computer interaction (HCI) and community and social informatics. He has taught several graduate and undergraduate courses on a variety of subjects including database management, ERP systems (SAP), electronic commerce, computer and network security, management information systems, systems analysis and design, distributed data processing systems, computer networking, knowledge management and other information systems related subjects. Since 2007, Dr. Sharma has served as the editor-in-chief/co-editor-in-chief of the International Journal of E-Adoption. He has been a guest editor, associate editor, reviewer and member of the editorial boards for several national and international journals in the area of MIS.

ABOUT THE CONTRIBUTORS

Dr. Emre Akadal, is a researcher at the Department of Informatics at Istanbul University. He has a bachelor's degree in Physics from the Faculty of Science and a master and doctoral degree in Informatics from the Institute of Graduate Studies in Science at Istanbul University.
His primary research interests are in evolutionary algorithms, database design, internet-based programming, and human-computer interaction. Dr. Akadal has research papers over ten in related national and international journals. He has served as the co-editor of the International Journal of Acta INFOLOGICA, since 2017. Also, he has been the manager of Istanbul University Human-Computer Interaction Laboratory. Dr. Akadal has been taken part in the team of several projects about information management.

Doc. JUDr. PhDr. Róbert Jáger, PhD. PhD. studied law, philosophy and cognitive sciences. Pedagogically, he focuses on teaching the legal history and legal philosophy. Scientifically, he focuses mainly on the development of the oldest law in the territory of Central Europe. Author uses linguistic methods to reconstruct the law at a time when there was no written record of legal norms. He currently points to the need for critical thinking in interpreting information and text.

Assoc. Prof. Hulusi Gülseçen, graduated from Astronomy and Space Science Department of Faculty of Science of Ege University in 1980. He received his Master of Science degree in 1983 and his PhD degree in 1989 from Institute of Graduate Sciences of Istanbul University with the theses on "The Commensurability Problem in Asteroids" and "The Effects of Corriolis Force in Sunspots" respectively. Dr. Gulseçen is currently working at Astronomy and Space Sciences Department of Faculty of Science of Istanbul University. His primary researh and teaching interest is in the topics of double stars, data analysis and astroinformatics. His research has appeared in journals such as Publications of the *Astronomical Society of Australia, New Astronomy, Education and Information Technologies*. He is lecturing both in Astronomy and Space Sciences Department and Informatics Department of Istanbul University respectively. In 2011, Dr. Gülseçen was visiting schollar at Astronomy Department of Ball State University in Muncie, Indiana (USA).

Assoc. Prof. Hasan Hüseyin Esenoğlu, (PhD) was born in 1963 in Eskisehir (Turkey). After his graduation from Istanbul University, Faculty of Science, Department of Astronomy and Space Sciences in 1982, he started his MSc in Physics Education by attending the English preparatory class in Marmara University. He completed his master's degree in Astrophysics in 1990 at the same department of Istanbul University with the thesis titled "Observational Properties of Two Cataclysmic Binary Stars". After completing his PhD at the same university, he started to work as guest astronomer at the Asiago Astrophysical Observatory of Padova University (Italy) under the scholarship of the Italian Ministry of Foreign Affairs. By preparing a thesis titled "Spectral and Photometric Analysis of Novae: A New Classification Method", he graduated from Stellar Astrophysics PhD program in 1996. Dr. Esenoğlu also visited King Saud University Science Faculty at Astronomy Unit of Physics Department (Riyat, Saudi Arabia) as a lecturer for three years, and the National Observatory of The Scientific and Technological Research Council of Turkey (TUBITAK) (TUG, Antalya, Turkey) as the Chief Researcher for five years. Dr. Esenoğlu primary research interest is on classical novae, cataclysmic variable stars and meteorites. He actively continue follow-up observations with ground based telescopes for alert sources of GAIA

and SRG (in the near future) satellites. His research has appeared in well known Astronomy journals (SCI) and he also published papers on popular science.

Enis Karaarslan, is currently Assistant Professor at the Mugla Sitki Kocman University (MSKU) Computer Engineering Department. He established NetSecLab for network/security education and research. He has been studying the potentials of the blockchain in BcRG (Blockchain Research Group) since 2017. He is one of the founders of the Artificial Intelligence (AI) discipline in the MSKU. His research areas are computer networks, cybersecurity, blockchain, data science, disaster management, and digital twin. He has over 50 papers to his name. He likes biking, hiking, traveling and reading.

Enis Konacaklı, has been managing CIS and IT teams, systems, services, from acquisition through delivery on through their lifecycle sustainment for 18 years. He is currently the Director of CIS of Eskişehir Air Base. He has been studying the potential benefits of the blockchain technology for National security and Military implementations as a member of BcRG (Blockchain Research Group) since 2017. His research areas are computer networks, cybersecurity, blockchain.

Fatma Önay Koçoğlu, (PhD) graduated from Istanbul University (IU) Faculty of Science, Department of Mathematics and Sakarya University Faculty of Engineering, Department of Industrial Engineering. With her dissertations on Data Mining, she received her master's degree in 2012 and doctorate in 2017 from IU Institute of Graduate Studies in Sciences, Department of Informatics. In 2011, she was a graduate student at Université de Technologie de Compiégne (France), Computer Engineering Department for 6 months. She is continuing her second doctorate in IU-Cerrahpaşa Institute of Graduate Studies, Industrial Engineering Program. She has been working as a Research Assistant at IU Informatics Department since 2010. Her research areas are Artificial Intelligence, Data Mining, Machine Learning, Decision Support Systems, Facility Location and Allocation Problems, Mathematical Modeling and Optimization. She has many scientific studies published in national and international fields. She worked as a researcher in projects supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) and the other Higher Education Institutions. She also worked as an instructor in the scientific training programs organized by Higher Education Institutions and Non-Governmental Organizations. She speaks English fluently and basic French.

Denizhan Demirkol, received his bachelor's degree in Information Systems and Technologies from Yeditepe University. He completed his master's degree in Yeditepe University, Department of Management Information Systems with honors degree. Demirkol continues his doctorate education at Istanbul University, Institute of Science and Engineering, Department of Informatics. He has been working as a research assistant at Aydın Adnan Menderes University Department of Management Information Systems since 2016. He has been working as a research assistant at Istanbul University Informatics Department due to his doctorate education. He gave lectures in the field of "Human Computer Interaction" at "SMK University of Applied Social Science" within the scope of Erasmus teaching mobility. During his PhD education, he started his studies in Data Science and continues to study Human –Computer Interaction, Artificial Intelligence, Machine Learning and Data Mining.

ABOUT THE CONTRIBUTORS

Malgorzata Pankowska, is Professor of Social Science and Chair of the Department of Informatics at the University of Economics in Katowice, Poland. She received the qualification in econometrics and statistics from the University of Economics in Katowice in 1981, the Ph.D. degree in 1988 and the Doctor Habilitatus degree in 2009, both from the University of Economics in Katowice. She was visiting professor at ISLA Braganca in Portugal, Trier University in Germany, ICHEC in Brussels, Belgium, VGTU in Vilnius, Lithuania, Istanbul University in Turkey, Ionian University in Corfu, Greece, Universidad de Ibague, Ibague-Tolima, Colombia, and Lapland University in Kemi-Tornio, Finland. She is in the Board of Information System Audit and Control Association (ISACA) Katowice Chapter. List of publications and more information on website: http://web.ae.katowice.pl/pank/

Mehmet S. Derindere, works with executives and teams to develop effective problem solving practices using knowledge. He has worked in several private organizations as trainer, business analyst, project manager, auditor and manager. Currently interested in cognitive and organizational traps that ensnare individuals as well as teams, prevent effectiveness and reduce quality of life.

Tugce Aksoy, graduated with a B.A. in Translation and Interpreting from Istanbul University in June 2018. She is currently studying in M.Sci program in Informatics at Istanbul University.
She completed her scientific preparation program and attended various courses such as statistics and programming during that period.
She works as a Vendor Manager in a translation company and continues to do academic work. Tugce has also worked as a translator for several years.
Tugce's research topics are NLP (Natural Language Processing), text mining, and statistics.

Serra Çelik, took her Phd in Quantitative Methods Department at School of Business in Istanbul University. She currently works in Department of Informatics in Istanbul University. Serra's research topics are multivariate data analysis, statistical learning and data mining. She gives lectures master level courses as Data Analysis, Quantitative Methods in Business. Serra recently works on Digital Transformation and Natural Language Processing. She has a web site about her academic life (serra.ist).

Zeki Özen, is a research assistant at Istanbul University, Informatics Department. He graduated from Sakarya University Computer Programming Department in 2002 and Istanbul University, Faculty of Science, Department of Physics in 2008. He received his MSc in 2012. In 2013, Dr. Ozen was at Ball State University, Miller College of Business, Information Systems and Operations Management (IN, USA) as a visiting scholar for 2 months. He received his Ph.D. degree in 2016 from Istanbul University, Informatics Department with his thesis titled *"Developing a security system for authentication based on keystroke dynamics using artificial neural networks"*. Research areas of Dr. Ozen are artificial intelligence, machine learning, data mining, biometric authentication systems, relational database systems, and source code comparison/similarity.

ABOUT THE CONTRIBUTORS

Elif Kartal, started her undergraduate education at Istanbul University (IU), Faculty of Science, Department of Mathematics in 2004. In 2008, she started her master's degree in IU Informatics Department and was accepted to the same department as a Research Assistant. She completed her master's degree in the IU Informatics Department with her thesis titled *"Software project cost estimation with artificial neural networks"* in 2011, and her doctorate with her thesis titled *"Machine learning techniques based on classification and a study on cardiac risk assessment"* in 2015. In 2013, Dr. Kartal was at Ball State University, Miller College of Business, Information Systems and Operations Management (IN, USA) as a visiting scholar for 2 months. From August 2019 to March 2020, Dr. Kartal was a visiting scholar at the University of Miami, College of Arts and Sciences, Department of Computer Science (FL, USA), and worked on computational neuroscience, deep learning, and computer vision with Assoc. Prof. Odelia Schwartz (Ph.D.). This postdoctoral research supported by TUBITAK-BIDEB 2219-International Postdoctoral Research Scholarship Program. Machine learning, data mining, and artificial intelligence are the main research areas of Dr. Kartal.

İlkim Ecem Emre, is a research assistant at Marmara University, Management Information Systems Department. She graduated from the same department in 2015. She has started the graduate program at Istanbul University Informatics Department in 2015 and graduated in 2017, where she wrote a thesis titled *"Analysis of effects of acute rheumatic fever in childhood on heart disease with data mining".* She is currently a Ph.D. student at İstanbul University Informatics Department. Her research areas are data mining, machine learning, and medical informatics.

Oğuz Ak, Ph.D., Instructor at Bogazici University, During his education period he get various courses like educational technology, pedagogy, information technology and informatics. He has a B.S. in Computer Education and Educational Technologies & M.A. in Management Information Technologies at Bogazici University. And Ph.D. in Informatics at Istanbul University. He worked as a distance education coordinator in a private university for a year. He has been working as an academician at Bogazici University more than 10 years. He is currently making researches about distance education, educational technologies and digital games.

Selim Yazıcı, Ph.D., Professor at Istanbul University, Selim YAZICI is Professor of Management and Organization at Department of Business Administration at Istanbul University, Faculty of Political Sciences. He received his BSc in Mechanical Engineering from Yildiz Technical University and his MSc and PhD in Business Administration from Istanbul University. He is also Adjunct Professor at Marmara University, Banking and Insurance Institute, Department of Insurance (2009-2016) and Ozyegin University, Graduate School of Business, Financial Engineering and Risk Management Program (2018-…). He is lecturing on E-Learning, Entrepreneurship, Financial Technologies, Digital Insurance, Project Management, *Business Continuity Management*, Management, Organizational Behavior and International Business both at undergraduate and graduate levels. His research interests are *E-Learning*, Digital Transformation, Disruptive Technologies, Entrepreneurship, Startups, Financial Technologies (FinTech), Insurance Technologies (InsurTech), Business Continuity Management, Team Building, Leadership and Experiential Learning. He is the author of four books published in Turkish: Business Continuity

Management (2013), E-Learning (2004), E-Insurance (2002), and *Learning Organizations* (2001). He is the co-founder of *FinTech Istanbul*, a platform for supporting financial technology (FinTech) startups and building a strong ecosystem in Turkey. He has also designed a "FinTech 101 Training Program" for entrepreneurs, corporates, and investors which is the first program of its kind in Turkey. Selim YAZICI is a member of Turkish Quality Association (KalDer), Academy of International Business (AIB) and The Chamber of Mechanical Engineers (TMMOB).

Yacine Lafifi, is currently working as a Full Professor at the Computer Science Department of Guelma University, Algeria. Since November 2018, he is the Vice-Rector of Development, Foresight and Orientation at Guelma University. Also, he is a senior researcher at LabSTIC laboratory (Guelma University, Algeria) where he is the leader of the "Web Technology and Intelligent Systems" team. He works in e-learning research field since 1997. He received his PhD in computer science from the University of Annaba (Algeria) in 2007. He has several published papers in conferences and journals. Furthermore, he is an editorial board member of many international journals (International Journal of Web-Based Learning and Teaching Technologies, Journal of Theoretical and Applied, SAGE Open journal, Electronic Journal of e-Learning, Acta INFOLOGICA, Frontiers of Contemporary Education, and International Journal of Information and Communication Technology Education). Currently, he works on e-tutoring environments, e-Learning, CSCL, recommender systems, Artificial intelligence in education, Intelligent agents, e-technology, MOOC and human tutoring systems.

Asma Boudria, is currently a PhD student in computer science at LabSTIC Laboratory (Guelma University, Algeria). She has worked for three years for implementing massive open online courses at LabSTIC laboratory. She is preparing a PhD thesis about the learning and evaluation in MOOC. Her current research concerns include e-learning systems, MOOC and human tutoring systems, personalization in MOOC and social learning in MOOC.

Atef Lafifi, obtained his master's degree in Computer Science in June 2017 from the University of Guelma (Algeria), specialty: Computer Systems. He has a good knowledge of website development in general and e-learning systems in particular. He masters several programming languages such as JAVA and PhP. He worked on tutoring in MOOCs.

Moadh Cheraitia, obtained his master's degree in Computer Science in June 2017 from the University of Guelma (Algeria), specialty: Computer systems. He masters several programming languages such as JAVA and PhP. He worked on tutoring in MOOCs.

Natalija Lepkova, doctor of technological sciences, Associate Professor at the Department of Construction Management and Real Estate, Civil Engineering Faculty, Vilnius Gediminas Technical University, Lithuania. She has 17 years of experience in teaching at university, specializing in facilities management, quality management systems, BIM in facility management. She published 2 books, 6 chapters in books and a number of journal papers in the mentioned fields. Participated in international projects, such as INTEREG Project LONGLIFE, also COST. She is a member of the editorial board of the following journals: International Women Online

ABOUT THE CONTRIBUTORS

Journal of Distance Education; Turkish Online Journal of Distance Education; Baltic Journal of Real Estate Economics and Construction Management, BIM Arabia. She is a member of: GYODER, IFMA Czech Republic Chapter, IFMA, Lithuanian construction engineers union, National Association of Distance Education (NADE).

Bilgin Metin, firstly received the B.Sc. degree in Electronics and Communication Engineering from Istanbul Technical University, Istanbul, Turkey. Later, he worked in the leading companies in the private sector, and he handled the design and installation of computer networks and cybersecurity projects. He received M.Sc. and Ph.D. degrees in Electrical and Electronics Engineering from Bogazici University, Istanbul. During M.Sc. and Ph.D. studies, he worked as a consultant in the private sector for designing, implementing, and supporting data communications and cybersecurity projects such as the International Terminal of Antalya Airport in Turkey.
In 2007, He started to work as an assistant professor for Management Information Systems Dept. at Bogazici University. He received the title of Associate Professor in 2014. His research interests include cybersecurity, IT Governance, data privacy, and telecommunication circuit design. He published more than 100 papers in international journals and conferences. Assoc. Prof. Bilgin Metin has been on the organization committee of many national and international conferences.
He has been also the head of the Bogazici University MIS Cybersecurity Center Project since 2017. He is carrying out national and international projects, also giving training and consultancy to various national and international institutions.

Enes Yılmaz, was graduated from Bogazici University, Management Information Systems Department in 2017. Currently, they continue their career in the information technology area.

Erdi Şekerciler, was graduated from Bogazici University, Management Information Systems Department in 2017. Currently, they continue their career in the information technology area.

Murat Gezer, has received PhD degree from Istanbul University, Natural Sciences Institute Electrics and Electronics Engineering. He is a working in Istanbul University Informatics Department. His teaching and research interests focus on Image Processing, Signal Processing, Computer Vision, datamining and data science programming. Furthermore detailed CV of Dr. Gezer, which has many national and international publications, book chapters and projects, can be accessed from https://www.researchgate.net/profile/Murat_Gezer

Dr. Çiğdem Selçukcan Erol, received her PhD degree from İstanbul University, Institute of Science in 2010. Dr. Erol is working in Istanbul University Informatics Department since 2003. She received the title of associate professor in Management Information Systems in 2018. Her research interests in bioinformatics, data mining, machine learning and artificial intelligence in health. Furthermore detailed CV of Dr. Erol, which has many national and international publications, books and projects.

ABOUT THE CONTRIBUTORS

Ahmet Naci Ünal, graduated from Gazi University. Then, he got his master's and PhD degrees at the Defense Technologies Department of the Institute of Science and Technology at Istanbul University. Between the years 1991-2012, he worked as an officer lecturer in the Turkish Air Force. He retired in 2012 at his own request. Ahmet Naci ÜNAL started to work as an assistant professor at Bahçeşehir University in 2013. He is also the Director of the Cyber Security Implementation and Research Center at Bahçeşehir University and the Coordinator of the Cyber Security Graduate Program in the Graduate School of Natural and Applied Sciences at Bahçeşehir University. He is specialized in electronic-based defense technologies, decision support systems and cyber security concepts. He has various books, articles and symposium papers published in these fields.

Zlatin Zlatev, is an Associate Professor at the Department of Technics and technologies, Trakia University, Bulgaria. He holds a Bachelor's degree in Electrical Engineering at the Trakia University and a master's and a doctorate in Process automation at the Ruse University, Bulgaria. Research interests: Computer-Aided Design, Informatics, Image processing, and analysis. Assoc. Prof.Zlatev is author of more than 100 publications and more than 95 citations. List of publications and more information on the website: https://www.researchgate.net/profile/Zlatin_Zlatev

Veselina Nedeva, PhD, is an Associate Professor in Informatics and Computer Science at Trakia University – Stara Zagora, Faculty of Technics and Technologies – Yambol, Bulgaria. She graduated Master's degree in Informatics from the University of Economics – Varna, and Ph.D. in Application of Computing in Economics at the same University. A dedicated research fellow with thirty years of experience in computer science, program languages, databases, information systems, computational data, e-learning, and distance learning. Extensively published in science journals and scientific conferences and author of more than 105 scientific publications and 130 citations.

Ivanka Zhelyazkova, is Professor of Apiculture and Head of sub-depatrment Beekeeping and Sericulture of the department of Animal husbandry-Nonruminants and other animals at the Faculty of Agriculture in Trakia University, Stara Zagora, Bulgaria. She received the PhD degree in 1999 and the DSc degree in 2013, both from the Trakia University in Stara Zagora. Main scientific interests: Biology of the honeybee and technologies for raising bee colonies; Oportunities for using bees and bee products as bioindicators for pollution of the environment. Prof. Zhelyazkova is author of more than 100 publication and more than 230 citations.

PREFACE
Who Runs the World: DATA

In ancient times, possessing a land and gold was the most important asset in the world. In the 21st century, "Data" is becoming the most important asset. As data activities continue to increase in speed, scale and variety, data Science and Data Analytics is becoming the new phenomenon of the 21st century. In fact, Data is considered to be the gold of the 21st Century and "data" is changing the face of our world. The organizations are becoming bigger and bigger and need a large amount of data to process for creating a business insight or intelligence. As more and more companies worldwide start offering products and services online, companies need to not only process information quickly but also to get insight about the needs, expectations, transactional behavior and responses of the customers. This mandates the companies to deal with "Data" effectively. Nowadays, "Data" is in various forms such as; texts, documents, online books, music, videos on number of different platforms such as; social media, transactional websites and many other internet based online forums.

The "Data" explosion is creating newer exciting opportunities for companies and individuals at the same time, it is also creating concerns and challenges. This edited book presents the research work of several researchers who are working in the data science related areas. The book comprises 15 chapters. A brief abstract of each chapter is provided below:

Material Information Carriers: Historical Development
Róbert Jáger, Matej Bel University, Slovakia

In this study, we will briefly try to describe what material information carriers were like in each period of the development of human society, what the advantages or disadvantages of these carriers were, and how society changed with the change of the material information carriers themselves. In conclusion, we will highlight an interesting fact of current development: the digitization of material information carriers, the separation of the information itself from its material carrier, and the risk that it will face in the future. In the first part of this study we describe Material information carriers in prehistoric times and in antiquity. Specifically, we focus on the development and use of material information carriers in Ancient Egypt, Mesopotamia and compare them with the development in European communities, which reached a similar level of civilization in the later period. In the second part, we pay attention

to the issue of using material information carriers in the Middle Ages. We show how the material carriers changed in the given period. In the third part of this study, we pay attention to the issue of using material information carriers in modern times. In particular, we point out the risks of the current state of information storage in information systems. The study primarily uses methods of description, analysis, synthesis, comparison, abstraction, and generalization.

Astronomical Data

Hulusi Gülseçen, İstanbul University, Turkey
Hasan H. Esenoğlu, İstanbul University, Turkey

Space telescopes have increased the quality of data collection for today's astronomy. In parallel to this, obtaining high quality data with high technology and good resolution focal plane detectors in accordance with the developments in material science in the ground-based observations has been achieved. With the new generation of ground based and space observations, global campaigns also brought continuity in data acquisition and increased performance. Finally, the fact that theoretical outputs can be made to allow in today's technology, for example, the detection of gravitational waves in the universe and these add new ones to the existing data. In addition, there has been a significant increase in data archiving, reduction and processing together with the number and variety of data collection tools. Astronomers have been able to overcome the facilitation in these processes in their own way: manpower waste has been reduced with autonomous telescopes, the data has been transformed into informatics (astroinformatics) with pipelines, the workload has been reduced to large masses by establishing a virtual observatory, and finally smart applications have been opened with the provided big data and new open areas have been reached with a future such as data mining. In this way, there has been progress in solving many astronomical events in the universe. This chapter is orginized in two subsections. In first, we are discussing how to solve problems in astronomy by using big data. In the second, we mention about big data sources in astronomy. The importance of data in astronomy, sources of data, big data in regards to the discovery of universe and analyzing data are the topics discussed in these subsections.

Data Storage in the Decentralized World: Blockchain and Derivatives

Enis Karaarslan, Mugla Sitki Kocman University, Turkey.
Enis Konacaklı, Eskisehir Technical University, Turkey.

We have entered an era where the importance of decentralized solutions has become more obvious. Blockchain technology and its derivatives are distributed ledger technologies that keep the registry of data between peers of a network. This ledger is secured within a successive over looping cryptographic chain. The accomplishment of the Bitcoin cryptocurrency proved that blockchain technology and its derivatives could be used to eliminate intermediaries and provide security for cyberspace. However, there are some challenges in the implementation of blockchain technology. This chapter first explains the concept of blockchain technology and the data that we can store therein. The main advantage of blockchain is the security services that it provides. This section continues by describing these services.. The challenges of blockchain; blockchain anomalies, energy consumption, speed, scalability, interoperability, privacy and cryptology in the age of quantum computing are described. Selected solutions for these challenges are given. Remarkable derivatives of blockchain, which use different solutions (directed acyclic graph, distributed hash table, gossip consensus protocol) to solve some of these challenges are described. Then the data storage in blockchain and evolving data solutions are explained. The comparison of decentralized solutions with the lcentralized database systems is given. A multi-platform interoperable scalable architecture (MPISA) is proposed. In the conclusion we include the evolution assumptions of data storage in a decentralized world.

Data in the Context of Industry 4.0

Fatma Önay KOÇOĞLU, İstanbul University, Turkey
Denizhan DEMİRKOL, Aydın Adnan Menderes University, Turkey

Today, every sector, not least industry, has been affected by the development of technology. With the breakthrough development of technology, Industry 4.0 has emerged with the concept of big data. Data is the most important element in the process of creating information. This study aims to deal with the subject of Industry 4.0 which has attracted great interest in the global field in the context of big data. Studies concerning Industry 4.0 and related data are examined in our study through a systematic literature review. Web of Science database and "industry 4.0 and data" keywords were used for our article search. A preliminary evaluation was performed for 20 articles meeting the objective of this study which were selected for

detailed examination. When the studies on Industry 4.0 and data are analyzed, we can determine that studies with big data, digitalization, internet of things, digital twin, cyber-physical systems, smart factories and cloud computing are prominent. Moreover, when the countries where the articles were published were analyzed, it was found that China was the most cited and studied country in this field. It is believed that the results of this examination will enlighten people working in this field and direct future studies.

Big Data Governance

Malgorzata Pankowska, University of Economics in Katowice, Poland

Information processing in a traditional way focuses on relatively stable structured data, repeatable processes as well as on operations in Business Intelligence systems. However, nowadays more and more popular, big data, defined as huge volumes of data available in varying degrees of complexity, generated at different velocities, and varying degrees of ambiguity, cannot be processed using traditional methods and technologies. Some people argue that suitable IT (Information Technology) infrastructure for big data processing is not yet widely developed nor implemented to discuss the big data architecture implementation benefits, risks, and opportunities. Nevertheless, this paper is to present the big data governance issues. Particularly, within the proposed theme, the author discusses the big data system architecture and development strategy. The last part of the paper includes a proposal of a big data architecture model as well as a design of balanced scorecard objectives and measures specification to support the big data governance at public services business organizations. As usual, there are two main research methods, i.e., literature review and the analysis of case studies. The first provides an overview of the existing knowledge and the second permits for contextualization of the proposed models. Beyond that, the paper includes definitions of the key concepts and enables to extend the knowledge base in the research area.

A Core Problem with Human Data Processing: Epistemic Circularity in Action

Mehmet Selim Derindere, İstanbul University, Turkey

Managers are expected to solve critical problems of our society in an efficient manner and in ways so that the problems remain solved. In order to accomplish this, the managers are provided with vast amounts of resources including mountains of data and a wide variety of problem-solving methods available. On the other hand, the effectiveness of social and organizational problem solving is far from satisfactory and this lack of effectiveness is

ubiquitous. One reason of this ineffectiveness we claim has to do with how the human mind works. The inherent capabilities and limitations of human mind coupled with social-cognitive skills lead to sub-par problem solving. An especially counterproductive problem solving approach used by managers is setting and attempting to solve problems using erroneous cognitive skills that not only fails to include relevant data but also uses the existing data in a counterproductive manner. The very data processing skills of managers make problem solving a dead end for the actors involved at great cost to them and to the society.

This chapter looks at a core human data processing problem that renders the available data and techniques ineffective. Epistemic Circularity disregards all the disconfirming or threatening data and fails to include it in the problem solution. Epistemic Circularity thus renders the revelant data useless in developing effective solutions. Easy knowledge, a product of epistemic circularity, leads to ineffective problem solving which in many cases result in exacerbated problems and counterproductive consequences.

Data Pre-processing in Text Mining

Tuğçe Aksoy, İstanbul University, Turkey
Serra Çelik, İstanbul University, Turkey
Sevinç Gülseçen, İstanbul University, Turkey

The fact that any kind of user has the ability to generate data with great ease at any time causes an increase in the importance of data mining. Considering the reality that the vast majority of the available data is composed of unstructured data and that the data in the text type is outnumbering, it proves the increasing interest in text mining and the abundance of studies in this field. However, in order to be able to examine an unstructured data type like text, which is quite different from machine language, it is necessary to make this data more structured and make the machine work. At this point, the data pre-processing step, which covers a large part of the entire text mining process, is of great importance. In this chapter, it is aimed to explain the text pre-processing phase on a basic level by supporting this using visuals. In doing so, it is primarily planned to focus on text mining and to explain in detail the characteristics of the data processed. In this context, it is aimed to explain the data pre-processing steps followed in order to overcome these difficulties by examining the difficulties created by the data in question. As a result, this chapter is a descriptive review of the data pre-processing phase in text mining, which covers some of the studies previously conducted on this subject.

Big Data in Education: A Case Study on Predicting E-learning Readiness of Learners with Data Mining Techniques

Zeki Özen, İstanbul University, Turkey
Elif Kartal, İstanbul University, Turkey
İlkim Ecem Emre, Marmara University, Turkey

Since the term "personalized learning" became popular, smart features have begun to be integrated into the e-learning environment. Data mining and machine learning algorithms are used to analyze big data stored in an e-learning system to make predictions to improve course quality or learners' performance. From the learners' perspective, it might now be considered possible for everybody to benefit from e-learning by considering their personal interests or their own specific development plan as long as the course contents are available in the system. In addition, in an e-learning environment there is no limitation on the time and place where a course can be attended and a program completed. However, it is just not that simple. Today not the only, but by far the most important, requirement is still the readiness of the learners to study in an e-learning system. The aim of this chapter is to predict the e-learning readiness of learners using data mining techniques and to provide feedback for institute managers and admin personnel of e-learning systems which are intended to be used in an institution. According to the results of this study, the highest accuracy value (0.831) is obtained with C4.5 Decision Tree Algorithm. While students, who agree and strongly agree with the statement "My studying/research area is appropriate for e-learning" are classified as ready to attend an e-learning course, students who disagree with the same statement are classified as not ready to attend an e-learning course. Students who strongly disagree with the statements "My studying/research area is appropriate for e-learning" and "E-learning is better than face to face learning", are also classified as not ready to attend an e-learning course.

The Value of Data for Improving Effectiveness of Campus Courses: The Case of Hybrid MOOCs

Oğuz Ak, Boğaziçi University, Turkey
Selim Yazıcı, İstanbul University, Turkey
Sevinç Gülseçen, İstanbul University, Turkey

In recent years with the advances in technology, learners started to learn various concepts in informal learning environments apart from the official traditional learning programs.

We describe such learning environments as part of the Personal Learning Environment (PLE) approach. One great resource for these environments is using Massive Open Online Course (MOOC). Learners can learn any subject by enrolling in MOOCs easily and develop themselves by reaching their personal learning goals. But in such an informal learning environment, it would be hard to manage the learning process. Learners need some ability to manage this process that is called "self-regulation". There seem to be some problems in both fully face to face learning (like difficulties in following courses), and fully online MOOCs (like lack of interaction). So, a midway approach is a hybrid MOOC that is a combination of both methods. Literature and author experiences indicated that this method would make learning more effective. However, there is a need for improving the method with proper data management. We provide a list of data collection methods in hybrid MOOCs and explain how this data helped us to improve the learning process. In the PLE approach, students need data to shape their learning process, similarly instructors need to obtain data with various strategies and reshape the course structure by using this data. We think that in education, data usage is somehow limited, but it is required for making it more efficient.

Intelligent Tutoring of Learners in E-learning systems and Massive Open Online Courses (MOOC)

Yacine Lafifi, University 8 May 1945, Guelma
Asma Boudria, University 8 May 1945, Guelma
Atef Lafifi, University 8 May 1945, Guelma
Moadh Cheraitia, University 8 May 1945, Guelma

In the last few years, many terms related to learning environments have emerged. Each one of these terms is distinguished by a set of criteria such as the target audience, the duration of learning, the type and nature of the educational content, the manner of dissemination of knowledge, etc. Unfortunately, lack of support for the learners seems to be a serious frequently faced problem in these environments, which requires special attention. Among proposed solutions, tutoring seems to be a convenient candidate for this problem. Tutoring involves offering assistance to learners that are in need for help. Regardless the nature of this assistance (pedagogical, social, etc.), it can be delivered in many forms: advice, guidance or even recommendation. And while tutoring has been applied for decades in traditional e-learning environments, its application in new systems such as Massive Open Online Courses (MOOC) is still under study. In fact, a considerable number of studies driven on MOOCs had reported the problem of learners' dropout. Several reasons can be listed as

causes of such a problem. Among these reasons, we can find learners' isolation as well as learners' loss of motivation. This same problem has been reported by researchers working in the field of Computer-based Environments for Human Learning.
In this article, we propose a new vision on how to apply an intelligent tutoring process in human learning systems in general and in MOOCs in particular. This new vision is based on the behaviors and skills of learners. This activity can take many forms and can be carried out by different types of actors (teachers, learners, etc.).

Smart House: Data Gathering and Analysis
Natalija Lepkova, Vilnius Gediminas Technical University, Lithuania

In modern society, the concept of "smart house" is increasingly being heard. At present, it is generally acceptable that a smart house has efficient building management, local management and business management systems. A smart house increases the business value of the environment created by the adaptability and flexibility provided by the location and the communication systems. There are many opinions on how we should understand the concept of a smart house. Some people believe this is a modern home audience, others think it's a fully designed home cable system. There are some who guess that this reflects modern telecommunication systems, etc. Everything that has been mentioned really reflects only part of the "smart home" possibilities. Smart House introduces a modern, robust automated system that allows to integrate all of the main operating subsystems such as: energy supply, supply of gas and water, lighting system, heating systems, microclimate systems, other remote controls. The smart houses are often pointed as one of the main constituents of smarter living environments. The chapter provides the smart house definition, criteria defining smart building, smart house technology explanation, examples of smart houses in different countries, smart house data model, building progress and analysis of smart building automation and control systems (applying SWOT analysis method).

Privacy for Enterprises in Data Age
Bilgin Metin, Bogazici University, Turkey
Enes Yılmaz, Bogazici University, Turkey
Erdi Şekerciler, Bogazici University, Turkey

The world we live in is now becoming increasingly virtual. We all interact with this new age which we can describe as the digital age. We shop online, we communicate with people via

social media, we are informed at any time through the devices that are in our hands about goings-on, whether we like it or not, we have become a part of this globalized and digitalized world. Data can be described as the structure of the digitalized world. In each interaction between us and the tools which we use, we create data or we cause data transferring or we can be a small part of a large data collection because of our presence in a platform on the internet. Certainly, this close relationship can reveal our private life in some situations. Most of the time, we are exposed to situations where our private information is collected, used, and processed without our permission. Sometimes we cannot even notice the violation of one of the most fundamental rights and freedoms we can define as privacy. This literature survey study is based on the fundamentals of information security, and it seeks answers to these questions: Why does our personal information need protection? What kind of information should be protected? What is the situation regarding the data privacy in Turkish and world law? What kind of laws have been passed upon the privacy of tax from past to today? What are the perspectives, opinions on protection of personal data in Turkey and Europe? What is the importance of data privacy for the business sectors? We also believe that this study will raise awareness on this matter.

Data Collection Approaches for Artificial Intelligence Applications in Healthcare

Murat Gezer, İstanbul University, Turkey
Çiğdem Selçukcan Erol, İstanbul University, Turkey

As in all other fields, research in the field of artificial intelligence is rapidly continuing in the field of health. As a result of this research, the importance of data comes to the fore. In this study, which includes data collection approaches in the field of health, we aim to emphasize the importance of data in this field and to contribute to the more conscious handling of the data to be used in artificial intelligence applications at every stage. For this purpose, the definition of data and how to distinguish information and knowledge are mentioned. The characteristics of data and data collection methods are also mentioned, and an attempt is made to emphasize the importance of health data collection in artificial intelligence research. As a result of this study, we believe that all personnel working in data-related departments and the health field, where the moment is vital, must receive training on collecting, storing, sharing data, and data security in particular. In our study we emphasize that especially the people who produce and consume data must have the awareness and morality for every step of data collection and handling, and that this issue should be prioritized in the field of health.

Technological Transformation Process from Electronic Intelligence to Cyber Intelligence

Ahmet Naci Ünal, Bahcesehir University, Turkey

Throughout the whole of human history, concepts such as defense, security, safety and intelligence have been very important for human beings on a personal level, and for human communities in general. The sociological transformation that came as a result of these processes played a key role in the development of science and technology. Thanks to the developments in electronic science during the 20th century, systems using electromagnetic energy have come to the fore.

This development process which started with systems such as telephones, radios and radars has eventually been used in many different areas such as air defense systems, guided missiles, early warning receivers, communication systems, and computers. For this reason, the control and active use of what can be called the electromagnetic spectrum in short, has been an important factor in all kinds of activities. By the 21st century, almost all of the systems used in this process began to operate in a cyberspace environment and became software controlled. The concept of the target intelligence needed in this transformation process has changed dimensions and shifted from electronic intelligence to cyber intelligence.

This study will focus on the transformation of the electronic intelligence process, which is an indispensable element of the 20th century, into the concept of cyber intelligence in the 21st century.

Automatic Measurement of The Morphological Characteristics of Honeybees by A Computational Program

Zlatin Zlatev, Trakia University, Bulgaria
Veselina Nedeva, Trakia University, Bulgaria
Ivanka Zhelyazkova, Trakia University, Bulgaria

The use of Big data related to the breeding of honey bees, when administered and processed effectively, will encourage the development of knowledge-based beekeeping, create new markets and business opportunities and further encourage the development of this industry. There have been attempts to fully automate the process of measuring the morphological characteristics of bees (at this stage there are conversions for Measuring wings), but this process for other parts are still completed manually. A survey was made of the possibilities

to automate the process of measuring the morphological characteristics in honeybees and the proposed algorithm and program to implement it. Color characteristics of parts of the bee body - tergite and proboscis, through which they can be separated from the background of the image, are analyzed and measured. Distances are determined between the values of the colour components of the object and background. From statistical analysis, it is found that S and V colour components of the HSV colour model are appropriate for the separation of an object from the background . Algorithms and a program in Matlab environment for separating tergite and proboscis from the background of the image and definition of their main sizes are developed. From the analysis of the results, it is found that the major influence on the accuracy of the measurement is of the bee in the image.

We are deeply indebted to several colleagues and students who have helped to complete this book. We owe a special debt of gratitude to several reviewers who provide valuable suggestions to authors for improving their manuscripts. We also acknowledge helpful advice from Istanbul University Press. I am sure readers will like the contributions of various researchers presented in this book and in last but not the least, we would like to thank the sections' authors, the management and staff of Istanbul University for their support and contributions in the emergence of this work.

CHAPTER 1

MATERIAL INFORMATION CARRIERS: HISTORICAL DEVELOPMENT

Róbert JÁGER*

*JUDr. PhDr. PhD. PhD. Matej Bel University, Faculty of Law, Department of History of State and Law, Banská Bystrica, Slovakia

e-mail: robert.jager@umb.sk

DOI: 10.26650/B/ET06.2020.011.01

Abstract

In this study, we will briefly try to describe what material information carriers were like in each period of the development of human society, what the advantages or disadvantages of these carriers were, and how society changed with the change of the material information carriers themselves. In conclusion, we will highlight an interesting fact of current development: the digitization of material information carriers, the separation of the information itself from its material carrier, and the risk that it will face in the future.

Keywords: Data, Material information, Material information carriers, History

Introduction

Material information carriers are an integral part of the development of human society. The historical development of human society is based primarily on the knowledge that we have been left with by our predecessors, and secondly, on the medium through which this knowledge has been recorded. From the earliest times of our existence, information that was to be long-term or permanently preserved was inevitably linked to a particular material carrier (stone, clay tablet, papyrus, wax plates, birch bark, animal skin, parchment, and later paper which became the primary carrier of information in the modern era, and in the current period and the period to come, an era of digital information media is emerging). But what do all these material information carriers have in common? Using the current vocabulary of modern language, we can conclude that in all cases, the material information carriers mentioned had a nature of an "external memory" (similar to an external storage disk on a computer). They were supposed to be used to preserve information, whether for later use of the person who wrote the information, or for another person who later gets access to the material information carrier.

1. Material information carriers in prehistoric times and in antiquity

Human history is usually divided into the prehistoric period and the historic period (and the latter period is divided into antiquity, the middle ages and the modern era). The prehistoric period is dated "from the earliest time of man's existence" to the period around the 31st century BC (but this time-frame refers primarily to the history of Egypt and the lands of Mesopotamia). It is characteristic of the prehistoric period that we get to know it primarily on the basis of sources of knowledge other than written sources. [1] It is characteristic of the historic period that we recognize it primarily on the basis of written sources. [2] To put it

1 Prehistoric period is examined primarily by the methods of archaeology.

2 The historical period is known primarily by examining the written sources of knowledge: the examined are information recorded as written. However, it must not be excluded that some aspects of historical times can also be explored on the basis of archaeological methods: for example, the Middle Ages can be explored by historical methods (that is, written recorded information from a given period is examined), but archaeological methods are also used, towns are explored, dwellings, burial grounds, human remains, and written recorded information are complemented by information obtained archaeologically.
The main difference between prehistoric and historic times is that while historic times can also be explored by archaeological methods of exploration, the prehistoric period cannot be explored by historical methods, i.e., by examining written information, as these did not exist in the prehistoric age (as there was no script). The dividing moment between the historical and the prehistoric is thus the emergence and use of the script. However, since the origin and use of scriptures in different cultures fall within a different period, we must specifically describe the historical and prehistoric periods in the history of each nation. For example, while the Egyptian landmark is about the 31st century BC, in Slovak history this is the landmark of the 9th century AD. (The difference between the prehistoric and historical times between the Near East civilizations and our Slavic ancestors was almost four millennia.)

simply, we talk about the historic period from the time when we can get to know the history of the culture in question on the basis of written records of that culture. If a given company did not use written information records in a particular period, in history such a period is considered prehistoric.

People also differ from the animal kingdom by the fact (among other things) that they can exchange information effectively (they can communicate effectively verbally). It is this ability that is considered to be one of the factors that have enabled people to survive and spread, almost into all parts of the earth.[3] However, for the purpose of getting to know our oldest history it is also essential that people have tried to capture information by transmitting information to persons who were not present at the time of publication. The first link between information and the material carrier thus arises.

Although written scripts were not used in prehistoric times, graphic representations of certain scenes have been found (and these were used fairly frequently). In Europe alone we have more than 350 locations with discovered Paleolithic art. Wall paintings (not only in caves) exist from a period of about 14,000 years ago, some of them even from about 40,000 years ago (Coward, 2010).[4] Although it may seem at first sight that they are exclusively works of art, there are now also opinions that these images also contain some "communiqués", that is, they are not just art, but also contain information that was meant to be kept for a longer period. If it really is so, although this information is not quite clear to us, it is remarkable that such messages have been preserved for so long.[5]

In addition to the above-mentioned examples, whose information content is not yet accessible to us, we are slightly better off with other works of the period just before the prehistoric era ends . For example, in ancient Egypt we have a period of the so-called Zero Dynasty (a label for a series of rulers of Egypt in the period prior to the rule of Minya - the official "first" ruler of Egypt). We have, for example, a preserved "scraped" image in a stone wall showing a male figure with something reminiscent of a royal crown, and one foot stands on the body of the enemy, and the other hand maces the enemy. At the head of the male figure

3 In practice, it may have looked like this: in a period of great drought, an old man who cannot walk will tell the younger members of his family where they can find water in the dry season. This vital information will help family members survive. Animals - not using speech, cannot submit such information. Although an animal parent may show his offspring a source of water, but in the case of his immobility, the animal is unable to provide information about the water source to its offspring. Just providing information through speech makes us unique (among other things).

4 Coward, (2010)

5 *„Most of the art objects of prehistoric art were linked to the spiritual ideas of prehistoric people and their symbolic world. However, this does not mean that there were no exclusively ornamental objects."* Šída, P. (2007)

there is something that resembles a scorpion. Although this image is just a sketch, and although a certain amount of imagination is required to understand these "lines" (or "scratches") creating the image described above, Egyptologists say it is perhaps one of the first depictions of the Scorpion king, one of the rulers of the so-called Zero Dynasty.[6] Although this image does not have the artistic value of the images of the rulers of the Old Empire, the message is similar: here, I am the ruler, and I will defeat the enemies that may intend to seize my land (almost all later portrayals of Egyptian pharaohs have such a meaning).

The graphical representation described above (e.g. the ruler defeating the enemy) was an intermediate step in creating a script of sorts. The oldest form of script only involved depicting symbals (it is assumed that they were made to record the economy: a sort of proto-accounting about things in warehouses or stock), an intermediate stage was a graphic representation of real facts (a sovereign versus an enemy) and later a picture script was created in Egypt, also in the cultures of Mesopotamia, or the oldest Chinese script).[7]

It is interesting to note that we can observe a similar development for our Slavic ancestors in the period preceding the use of Latin script in Great Moravia. Even in the period texts of Great Moravia it is said that before the Slavs adopted Christianity, they used various "features and cuts".[8]

So what were the oldest carriers of written information? The oldest carrier of information was stone walls. However, for practical reasons, it was necessary to gradually replace the stone. The stone wall could not be moved or sent to another recipient. Stone slabs were used,

6 We do not have the existence of this ruler in the written documents of the Ancient Egyptian period, but the present Egyptology considers his existence to be - at least - very likely.

7 On the development of the Egyptian script from the original signs serving mainly for economic and administrative purposes, see e.g.: Manley (2004)

8 This suggests that Slavs had been using "signs" to record important data for a long time. However, there were two fundamentally different types of signs, on the one hand, features and notches (which are a primitive level of writing development), and on the other hand, the graphemic features of the Greek and Latin alphabets. It should be emphasized that there is a vast history gap between these developmental stages of the graphical record, which can only be filled by inserting into the expected development line the transformation of "features" into ideograms, then into syllable script and finally a long stage of evolutionary change of the syllable to phonetic script ... (Slavs probably "skipped" this long development: by being in contact with advanced ethnicities using Latin and Greek scriptures, they simply took over these scripts./a note by R. J/). „*From the history of the evolution of the scriptures of great civilizations, it is well known that individual nations used the scriptures of advanced cultures to create their own alphabetic systems according to their pattern. This is how the Phoenician alphabet originated from the Egyptian model, and then from the Phoenician the Hebrew or Greek alphabet and from the Greek the Latin, but it was always centuries-old processes closely linked to the development of specific languages.* "Kralčák (2014). Some authors tried to fill in the above-mentioned gap between signs and complex letters by the so-called runic alphabet that Slavs allegedly used. Any attempts to present evidence of Slavic runes later turned out to be fake and forgeries. *Vales's book*, whose runes are noticeably similar to later Cyrillic, also has such a character, and the Cyrillic reader can read these texts without much difficulty. For more details on this subject, see, for example: Jáger (2017)

which were lighter and could be sent to the recipient, but were still impractical because the information had to be engraved on the stone slab, which was both time-consuming and involved a lot of physical effort . Although stone slabs could be transported, they were still relatively difficult to carry in larger quantities, and were still relatively fragile. It was also difficult to archive more records of information recorded only on stone slabs. Therefore, the stone was gradually replaced by a material that did not have these shortcomings. What replaced the stone depended mainly on what material was available in the country. In the countries of Mesopotamia it was clay, in Egypt it was papyrus.[9]

In the lands of Mesopotamia, where there was plenty of quality clay, clay tablets played the role of information carrier. These tablets were formed by the wet processing of the clay, shaping the clay into the desired (rectangular) shape and, when the clay was still soft, the required information was engraved with a wooden digger. After drying the clay (in the sun), this carrier was relatively easy to carry, was durable, light and well archived. In this form, parts of the oldest Epic of Gilgamesh have been preserved ,[10] as well as official records or bills, or insurance contracts.

The character of this material carrier was also determined by the development of the script. The original script used in the lands of Mesopotamia looked similar to Egyptian hieroglyphs ("pictorial" script, signs originally resembling what they expressed), but in clay tablets the round shapes were difficult to engrave by the chisel, so the shapes of the characters in the sign system gradually became more abstract and "more angular" in shape. One could describe these as "wedge shaped", that is, by the imprint of a chisel into the clay (the sun sign no longer had a round shape but a square shape). Thus, the very nature of the material carrier also determined the development of the script. From an historical point of view, clay tablets could be considered an effective material carrier of information, used in the countries of Mesopotamia for several millennia, and they were able to preserve information about the bearers of the original culture in sufficient quality and quantity to the present. However, the disadvantage of such tablets was that after the entry was completed and dried, no changes or corrections could be made to the record.

Also in Egypt, a material information carrier was used which came from aproduct that was readily available in the Nile Valley: papyrus. Its production was time-saving and

9 Parchment, smoothed animal skin was also used at the same time.

10 The main theme of the Gilgamesh Epic is the search for eternal life and the meaning of life, but Gilgamesh was just inspired by the death of his friend Enkidua. Only after his death did he realize that he, too, was mortal, and found the meaning of life itself in the search for eternal life and immortality. See more: Zamarovský (2002)

inexpensive. It had properties similar to today's paper, was light and durable. Papyrus proved to be easily archived and transmitted to the recipient of the information. In addition it was easy to write on, and round shapes of more complicated characters were easy to write compared to the clay tables. Therefore, in Egyptian script we do not see the process of individual characters "becoming angular", as can be seen in the wedge shaped script. The hieroglyphic sign system evolved so that even more abstract terms could be recorded by it. If we used hieroglyphs in Slovak, we could give the following example. The "mouse" sign and the "ladybird" sign would be read together as a mouse (in Slovak *myš*) + ladybird (in Slovak *lienka*), i.e. *myšlienka* - idea.[11]

The papyrus itself, as a carrier of information for three millennia of Egyptian history, did not see much change. Its appearance was similar in all developmental periods of Egyptian history. The main advantage of papyrus as an information carrier is its light weight, and durability when properly stored. It is also significant that it can be archived in large quantities. In its flourishing period in the 1st century BC, the Library of Alexandria had more than two million books in the form of papyrus scrolls.[12] Like the clay tablets, it can be stated that the Egyptian papyrus was a very effective material carrier of information. One of its advantages was that it could be reused. The old text could be scraped off the papyrus carefully, and it was possible to write on it again.

From Egypt, papyrus also spread to Greece and to the Roman Empire (and thus practically to many European countries). Unlike Egypt, however, much written information on papyrus has not been preserved in European countries. This is mainly due to the way of burying. Most of the papyrus was stored in tombs in Egypt, where it "survived" the millennia. However there was a different way of burying in Europe, where the body was burned and the burial equipment was not used so that papyrus scrolls could be found untouched. Although Egyptian papyrus enjoyed great popularity in European countries, its price was extremely high as it had to be imported only from Egypt (papyrus does not grow in Europe). Papyrus was not like other material information carriers, which were common products found in every home (such

11 However, the hieroglyphic sign system was extremely difficult to memorize, so two simpler sign systems (demotic and hieratic script) are emerging, which were significantly easier to learn and for practical administrative purposes. Hieroglyphs were used throughout Egyptian history, but in the earlier times they were used only as a script to record religious texts, for the practical purposes of everyday life, the above-mentioned scriptures were used.

12 The Egyptians demanded that every ship that entered the port of Alexandria brought 'books' by way of a city entry tax, which they had to hand over to the library upon arrival. These were rewritten onto papyrus scrolls, copies were archived, and the original books were returned to the owner. The library was burned down during the siege of Alexandria by Caesar. The library was restored and continued operating until the 4th century, when a crowd of Christians plundered it and burned it down, for its "ungodly" character.

as today's paper), but it was a significant commodity that only wealthy scholars or government officials could afford.

Also, the high price of papyrus caused the use of parchment (fur-free animal skin, finely worked to a thin thickness) to be developed as a material information carrier in European countries during ancient times and in the Middle Ages. Parchment was already used in ancient times in countries where there was plenty of fur, and Egyptian papyrus was not imported there or was too expensive. However, the parchment had a disadvantage compared to papyrus: when rolled up and deposited for a longer time in archives it dried up and it was not possible to unfold it without cracking or breaking. Therefore, the parchment was not rolled, but was stored as paper is today "put one on top of another". Once, someone sewed several parchments on one side, and the first book was made. As this happened in Byblos (in Lebanon today), the first book was called the "Bible" due to the place in which this happened. Thus, the term bible originally referred to any book, and later on, the word of God began to be called the Bible.[13]

In addition to the material information carriers mentioned above, others have also been used. In Greece, for example, expensive papyrus was not used to teach writing or for temporary administrative records, but waxy petals, stored in a wooden case. This wooden case with wax petals was called "diplomas". The wax was simply engraved with letters, and when it was not necessary to archive the recorded information, the wax was warmed up, and it was smoothed again, allowing the material carrier to be used again. Such records were also used in the state administration: the messenger, who was given the mandate to negotiate with foreign countries, carried with him "credentials" in the form of wax plates. Therefore, the area of public relations dealing with international relations has also become known as diplomacy. (Králik, 2016)

In describing material information carriers in antiquity, we will point out another interesting aspect. Information in the ancient world was not commonly available to "everyone" as it is now. Restricted access to information in antiquity was due not only to the high cost of the material information carrier and the illiteracy of the vast majority of the population, but also to the deliberate non-disclosure of much information. For example, in Egypt, knowledge (education) was practically accessible only to monks who lived in the temple. If someone wanted to gain access to their knowledge (such as mathematics, astronomy, or geography), they had to join their monastic community, and after several years of staying in that community, the knowledge was made available to them. These were closed communities

13 Also papyrus sheets were similarly connected, so the name bible does not only refer to "parchment" books. Cf. Králik (2016)

with strict rules, similar to sects. It was forbidden to make the monastic community's knowledge available to "uninitiated" persons.

Even Pythagoras, when he wanted to study mathematics, had to enter a monastic community in Egypt that made their knowledge of mathematics available to him, and after his return to Greece he presented this knowledge as "his" discoveries. Also, his school was atypical in Greek terms. His school was a closed association with strict rules whereby his students did not see him for several years (he would speak in the dark and only behind a curtain), and both teachers and students were very superior to other people who did not belong to their community, and thus, other people had no access to their information and knowledge (De Crescenzo 2006, 2004). This character of his school was inspired by the character of the monastic communities in Egypt. Thus, access to information in antiquity was not as automatic as it is in modern times and now.

2. Material information carriers in the Middle Ages

The production of medieval books was concentrated mainly around monasteries, centres of medieval education. It is logical, since, in the period under review, almost nobody was literate apart from the clergy (not even monarchs or civil servants in the state administration). Books in the Middle Ages were rewritten in monastic scriptoria, with a great emphasis on the aesthetic side of the text. The text of the book would be rewritten for several years, and after the text had been rewritten, the text was decorated with hand paintings. When the text was finished, the cover of the book was decorated with gold platelets inlaid with gems. Books in the Middle Ages had an extremely high price: the average book in the Middle Ages had the value of a middle-class car today. Only a very wealthy man could possess just one book in the Middle Ages, and the possession of a library with several books was a sign of immense luxury. That is why the books in the medieval libraries were carefully guarded and books were often chained to prevent them from being stolen.(Turošík 2016, 2016a)

Another interesting aspect of medieval books, such as the contemporary primary information carrier, is the number of titles available in the Middle Ages. Given that in early Christianity resistance to "pagan" literature arose, non-Christian books ceased to be transcribed in the Middle Ages over time, and were gradually forgotten. In the early and high Middle Ages, approximately 50 book titles were available in the average library. Compared to the many books in the ancient Library of Alexandria, this ratio is woefully low. While the number of book titles in medieval Christian Europe was minimal, a different situation entirely existed in the Muslim world. There was no resistance to Greek and Roman literature in the

early Middle Ages, and so many works of ancient philosophers (such as Aristotle's works) were preserved mainly in libraries of Muslim countries. It was not until the 13th century that these works were translated into Latin and made available for knowledge in Europe. Thus, if Muslims had not been tolerant of ancient literature, today's philosophy would be very impoverished.

It is interesting to note here that although the Muslim and Christian worlds fought frequent wars against each other during the High Middle Ages, it was precisely the area of science that was the point of contact between the two cultures. For example, the Spanish Cordoba was a place where scholars from both the Christian and Muslim world came to exchange their information. However, in Europe, it is overlooked that, in particular, information from scientists from the Muslim world was more valuable in exchanging this information, as Arab science was much more advanced at that time, as compared to science in Europe.

When describing books as the primary information carrier in the Middle Ages, it should be mentioned that since the vast majority of the medieval population was illiterate (about 99 percent of the population), it was necessary to disseminate information by other means. For example, rulings of monarchs that were written on parchment were published so that they were read in a public place (for example, in the market square during a market). The recording of the ruling of the monarch itself and its "publication" in writing would be ineffective for the illiteracy of the population. Another frequent means of communicating information (especially of a religious nature) was also wall paintings in churches expressing, for example, scenes from the lives of saints, or scenes from the crucifixion of Christ. For ordinary people, these wall paintings were a full-fledged source of information and at the same time they were the only source of "written" character information. Of course, in addition to these sources of publication of "written" information, clerics who spoke the message of God served the faithful with the information contained in Christian written documents.

3. Material carriers of information in modern times

The Renaissance was a great turning point for the literacy of the population. In this period, secular schools also began to be founded. If someone wanted to learn and have access to information recorded on material carriers in writing, he no longer had to enter the monastery, but only to attend a secular school. The number of literate people grew slowly, and with it a number of secular books. Another important milestone was the discovery of the printing press. The price of books began to decline rapidly (but books were still expensive for the poorest compared to the present). Even so, the average townspeople or merchants could buy

at least one book (most often the Bible). Moreover, the process of increasing the literacy of the population and lowering the price of books had started. The climax of this process can only be seen in the 19th century. Only in this period did the majority of the population become literate and the price of books acceptable to ordinary people. This was a period in which information was available to a wide range of the population.

In the 20th century, the phenomenon of mass media can be observed. Even in the interwar period, newspapers were becoming everyday information resources for everyday people. (By comparison, in the Middle Ages, an ordinary person had access to as much international information as could fit into one newspaper.)

The radio also becomes a mass medium of the interwar period. Unlike newspapers that were only able to make new information available the next day, the radio could convey information immediately. Although the radio was indeed a revolutionary means of broadcasting information at that time, its use in broad sections of the population was hampered by the fact that not all households of the period had installed and were able to use electricity. The access to electricity of most households became commonplace only after the Second World War.

After World War II, the use of the television grew rapidly. Unlike the radio, the television did not only transmit voice but also video information. Information from different cultures or companies was easier to convey to members of other cultures since this information was also associated with image information. To a large extent, the radio in the past had to use its own imagination of the recipients to know what they were hearing on the radio. For the first time in their lives, television viewers could see the presidents or rulers of their countries, whom they had previously only read about in newspapers, or whose voices they had only known from the radio. Politicians, rulers and people of public life became very close to the ordinary population.

In 1952, the US company IBM introduced the first magnetic tape storage system, putting electronic computing to use. In the 1990s, the Internet arrived : the US scientist's original network of information transfer marks the digital era we live in today. How did material information carriers change in the digital era? First and foremost, it is now easy and cost-effective to copy information from one carrier to another. By reducing the cost of a material information carrier, the price of the information itself also decreases considerably, and information is available to a wide range of recipients.

However, it is now possible to disseminate information without having a material carrier. The Internet environment creates an almost unlimited number of platforms for accessing or

exchanging information. The advantage is also the speed at which the information reaches its addressee. The distance between the originator and the recipient of the information no longer plays any role. Despite the obvious positives that the digital era certainly brings, problems arise, such as digital security of information. Since most of the information is in digital form, it is important to ensure its protection as well. The legislation of individual countries must also respond (and react) to this requirement. Cyber security is becoming a very topical issue in today's world (Rosenoer, 1997).

One of the serious problems highlighted in particular by archivists is that in the future, the exclusive storage of information on digital material carriers may pose the risk that future generations will not be able to access information contained on older material carriers (floppy disks, if the floppy disk drives will no longer work on computers, or their computers will not use them at all), and if the digital media that is to archive the information is damaged, the information collected for archiving may be lost forever.

"Unlike the previous decades, there is currently almost no physical record of most of the digital material we own. We are taking pictures on digital cameras, but few people want to take pictures, the lifetime of CDs will not last more than a few decades. We may know less about the beginning of the 21st century than about the early 20th century", says Rick West, managing data at Google. "The beginning of the 20th century is still largely archived in paper and film formats, which are mostly accessible. Much of what we do now are things we put into digital content. And that will disappear after some time. It is not something we have translated from analog to digital container", as Google's data manager concluded."[14]

Although these problems may seem petty, at least they need to be considered and addressed when setting up digital information systems to serve the long-term archiving of cultural wealth and information we currently undeniably have. It would certainly be a tremendous loss if we lost the wealth of accessible amount of information that we have gathered to date.

References

Coward, F. (2010) Paleolitické jaskynné umenie [Paleolithic cave art]. In Predhistória. Bratislava: Ikar.

De crescenzo, L. (2006) Příběhy středověké filozofie [Stories of medieval philosophy]. Praha: Leda.

De crescenzo, L. (2004) Příběhy řecké filozofie [Stories of Greek philosophy]. Praha: Leda.

Gábriš, T. (2014) Cyber Law. Bratislava: Komenius University.

14 (Autor´s name is not given). Vytvárame digitálny temný vek. In Pravda, online document https://spravy.pravda.sk/ekonomika/clanok/453654-vytvarame-digitalny-temny-vek-varuju-vedci/.

Jáger, R. (2017) Metod - zakladateľ (právnického) vzdelávania na Veľkej Morave? (niekoľko úvah o vzdelávaní na Veľkej Morave) [Method - the founder of (legal) education in Great Moravia? (some thoughts on education in Great Moravia)]. In Školy, osobnosti, polemiky : pocta Ladislavu Vojáčkovi k 65. Narodeninám. Brno : The European society for history of law, pp. 206-219.

Kralčák, Ľ. (2014) Pôvod hlaholiky a Konštantínov kód [The origin of Glagolitic and the Code of Constantine]. Martin: Matica slovenská.

Králik, Ľ. (2016) Stručný etymologický slovník slovenčiny [Brief etymological dictionary of Slovak]. Bratislava: Veda.

Manley, B. (ed) (2004) Sedmedesiat veľkých záhad starého Egypta [Seventy Great Mysteries of Ancient Egypt]. Bratislava: Slovart.

Rejzek, J. (2015) Český etymologický slovník [Czech etymological dictionary]. Praha: Leda.

Rosenoer, J. (1997) Cyber law: the law of the internet. London: Springer.

Schneiderová, A. (2011) Kultúrne - špecifické slová a súvisiace prekladové postupy [Cultural - specific words and related translation procedures]. In Notitiae Novae Facultatis Iuridicae Universitatis Matthiae Beli Neosolii. Vol. 16. Banská Bystrica : Univerzita Mateja Bela, Právnická fakulta. pp. 278-288.

Schneiderová, A. (2000) Lingvistické vlastnosti anglického právneho textu – náčrt [Linguistic features of English legal text - sketch]. In Notitiae Novae Facultatis Iuridicae Universitatis Matthiae Beli Neosolii. Banská Bystrica : Univerzita Mateja Bela. Vol. 4.

Schneiderová, A. (2012) Skopos theory in the translation process. In Journal of modern science. Józefów : Wydawnictvo Wyższej Szkoły Gospodarki Euroregionalnej im. Alcide De Gasperi. Vol. 2, No. 13, pp. 71-76.

Turošík, M. (2016) Inštitút spravodlivej vojny v rímskom práve [Institute of Just War in Roman Law]. In Historia et theoria iuris. Vol. 8, No.1, pp.103-109.

Turošík, M. (2016a) Kúpna zmluva [Contract of sale]. In Obchodný zákonník : veľký komentár. Bratislava: Eurokódex, 2016.

Šída, P. a kol. (2007) Lovci mamutov [Mammoth hunters]. Martin: Otovo nakladatelství.

Zamarovský, V.(2001) Gilgameš [Gilgamesh]. Bratislava: Perfekt.

Vytvárame digitálny temný vek [We are creating a digital dark age]. In Pravda, online document https://spravy.pravda.sk/ekonomika/clanok/453654-vytvarame-digitalny-temny-vek-varuju-vedci/.

Grant Support: APVV-16-0362: *Privatization of criminal law - substantive, procedural, criminal and organizational-technical aspects*.

CHAPTER 2

ASTRONOMICAL DATA

Hulusi GÜLSEÇEN*, Hasan H. ESENOĞLU**

*Asos. Prof., İstanbul University, Science Faculty, Astronomy and Spaces Sciences Department, İstanbul, Turkey
E-mail: hgulsecen@istanbul.edu.tr

**Asos. Prof., İstanbul University, Science Faculty, Astronomy and Space Sciences Department, İstanbul, Turkey
E-mail: esenoglu@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.02

Abstract

Space telescopes have increased the quality of data collection for today's astronomy. In parallel to this, obtaining high quality data with high technology and good resolution focal plane detectors in accordance with the developments in material science in the ground-based observations has been achieved. With the new generation of ground based and space observations, global campaigns also brought continuity in data acquisition and increased performance. Finally, the fact that theoretical outputs can be made to allow in today's technology, for example, the detection of gravitational waves in the universe and these add new ones to the existing data. In addition, there has been a significant increase in data archiving, reduction and processing together with the number and variety of data collection tools. Astronomers have been able to overcome the facilitation in these processes in their own way: manpower waste has been reduced with autonomous telescopes, the data has been transformed into informatics (astroinformatics) with pipelines, the workload has been reduced to large masses by establishing a virtual observatory, and finally smart applications have been opened with the provided big data and new open areas have been reached with a future such as data mining. In this way, there has been progress in solving many astronomical events in the universe. This chapter is orginized in two subsections. In first, we are discussing how to solve problems in astronomy by using big data. In the second, we mention about big data sources in astronomy. The importance of data in astronomy, sources of data, big data in regards to the discovery of universe and analyzing data are the topics discussed in these subsections.

Keywords: Astroinformatics, Astrostatistic, Astronomy, Big data, Machine learning, Processing, Reduction

1. Introduction

Astronomy is the study of physics, chemistry, and evolution of celestial objects and phenomena that originate outside the Earth's atmosphere, including supernova explosions, gamma ray bursts, and cosmic microwave background radiation (Zhang and Zhao, 2015). Since astronomy is a science that studies celestial bodies, the objects in space can only be investigated by examining the light coming or reflected from them. Thus, the only source astronomers have is light.

There are many difficulties when investigating a celestial body. The Earth, the Sun and the Solar system are constantly in motion. Also, more distant celestial bodies such as stars and galaxies are constantly in motion. For this reason, the location of a celestial body at the time of observation, the position of the earth and the time of observation are very important.

Studies with celestial bodies must be reduced to a heliocentric coordinate system. Observation time (which is one of the main parameters of astronomical data sets) should also be reduced to HJD (Heliocentric Julian Day).

The time and the coordinates of both the celestial body and the detector (telescope, satellite, CCD, etc.) are indispensable parameters of a data set regardless of the wavelength in which the field of astronomy is studied.

We can roughly divide astronomy into three areas of study. These are astrometric, photometric and spectroscopic studies. Roughly, we can classify the celestial objects to be observed as the sun, the objects of solar system, stars, Milky Way, galaxies, and galaxy groups. These celestial bodies are observed with different devices at different wavelengths of the electromagnetic spectrum. The classes of astronomy in terms of wavelength can be made as follows: gamma-rays astronomy, x-rays astronomy, ultraviolet astronomy, optical astronomy, infrared astronomy and radio astronomy.

Astrophysics is the branch of astronomy that studies the physics of the universe, in particular, the nature of celestial objects rather than their positions or motions in space. Astrophysics typically uses many disciplines from physics, including mechanics, electromagnetism, statical mechanics, thermodynamics, quantum mechanics, relativity, nuclear, particle physics, and atomotic and molecular physics to solve astronomical issues (Zhang and Zhao, 2015).

The occurrence times and life span of the events taking place in space also vary greatly. For example, gamma-ray bursts last for a few seconds while solar eruptions and binary star

eclipses last for a few minutes and several hours to years, respectively. The lives of stars last from ten million years to several billion years.

Gamma-ray bursts (GRBs) in Astronomy are flashes of gamma-rays associated with extremely energetic explosions that have been observed in distant galaxies. They are the brightest electromagnetic events known to occur in the universe after the big bang. Bursts can last from milliseconds to several minutes. The initial burst is usually followed by a longer-lived afterglow emitted at longer wavelengths (x-ray, ultraviolet, optical, infrared, microwave and radio). Targets of Opportunity (ToO) are astronomical objects undergoing unexpected/unpredictable transient phenomena and proposed for observation. The observations are normally urgent because of the transient nature of the event and may require even an immediate intervention at the telescope. ToO include objects that can be identified before the onset of such phenomena (e.g. dwarf novae, x-ray binaries) as well as objects which cannot be identified in advance (e.g. novae, supernovae, gamma-ray bursts). Modules have been developed for fast telescopes that respond to GRB alerts robotically in collaboration with the coordination of data networks. An example of deployed T60 at the TUBITAK National Observatory (Antalya, Turkey) was carried out by embedded software of the robotic telescope (Dindar et al., 2015). The telescope responds to GRB triggers transmitted from the Goddard Space Flight Center alert system thanks to this autonomy. It uses the Gamma-Ray Explosion Coordinates Network - GCN (formerly known as the BATSE Coordinates Distribution Network, BACODINE) while doing this. There are also some pipelines designed for Gaia alerts (http://gsaweb.ast.cam.ac.uk/alerts/alertsindex) similar to GRB alerts. One of these is “AlertPipe” which is responsible for real-time detection and classification of anomalies and transient astrophysical phenomena. The pipeline works within the Gaia data processing stream.

Recent advances in satellite and CCD technology have allowed for a more detailed examination. Dark energy, dark matter and exoplanet research have been accelerated thanks to these developments in technology.

Advances in computer technology, the enormous expansion of new storage capacities, the diversity and organization of astronomical data have led to the addition of two new fields of study to astronomy. In particular, data mining, machine learning and artificial intelligence applications have started to be used in astronomy studies.

Finding solutions to the problems in astronomy with big data and subjects of big data in astronomy are discussed under the relevant subheadings below.

2. Solving Problems in Astronomy with Big Data

Statistics plays an essential role in data-rich astronomy. Scientific insights cannot be extracted from massive datasets without statistical analysis. The statistical challenges are not simple; image analysis, time series analysis, nonlinear regression, survival analysis, and multivariate classification are all critically important (Feigelson and Babu, 2012).

Data in a method called DFS (Distributed File System) is placed wherever there is a free computer on Earth. For example, a part of the picture you upload to Facebook can be held on a computer in China and the other part can be held on a computer in Canada. Hadoop combines these two pieces of information in milliseconds when you click to view.

Astronomy was developed in two main areas namely "Astrostatistics" and "Astroinformatics". Astrostatistics can be summarized as the application of the science of statistics to the sciences of astronomy and astrophysics. Astroinformatics can be defined as computer programs and analysis methods developed to process big data from telescopes.

For example, CALTECH's space telescope GALEX (The Galaxy Evolution Explorer) 30 TB, Australia's SkyMapper (Southern Sky Survey) 500 Terabyte, and NASA JPL's Hawaii telescope PanSTARRS 40 PB is generating data while the data produced by Palomar Observatory is 3 TB. The amount of data generated reaches almost zettabytes when we combine all the telescopes in the world.

The International Virtual Observatory Association (http://www.ivoa.net) established for this purpose is designed to combine information from telescopes all around the world with Hadoop to establish an environment accessible to every astronomer. A virtual observatory has been set up and all data from telescopes so far has been shared, and when astronomers want to analyze a region, they can access information from telescopes on a single screen and make virtual observations. In short, the virtual observatory makes it easier for scientists to make science.

All observational data in the world and in space were collected and opened to share with the virtual observatory. How will this data be processed? The database mining programs which were developed for them come into play here. CALTECH claims the existence of the ninth planet because it analyzes the data it receives from them and computes mathematics. The main astronomy data analysis programs are:

1. StatCodes (http://astrostatistics.psu.edu/statcodes/)
2. VOStat (http://astrostatistics.psu.edu:8080/vostat/)

3. Weka (http://www.cs.waikato.ac.nz/ml/weka/index.html)
4. AstroML (http://github.com/astroML/astroML)
5. DAME (Data Mining & Exploration) (http://dame.dsf.unina.it/)
6. Auton Lab (http://www.autonlab.org/autonweb/2.html)

These programs and these data alone are not something that astronomers can handle. They need computer engineers, Big Data experts and statisticians to do these tasks. 7 different organizations operate to bring them together. The complete list (ASAIP) is available at: http://asaip.psu.edu.

Data was collected from the telescopes, a virtual observatory was opened, this information was accessed and analyzed with software. The results of these studies was evaluated at an annual conference. The name of this conference is Astronomical Data Analysis Software and Systems. One-year studies have been reviewed and the methods in the analyses compared over the last 6 years at the conference held every year. The website of this conference (ADASS) is available at: http://www.adass.org.

As a result, it can be said that Big Data is a technology that will hold the future in both normal life and astronomy. Programmers are already calling Big Data "Oil of the Future" (Zhang and Zhao, 2015).

3. Astroinformatics

Astroinformatics is an interdisciplinary field of astronomy, astrophysics and informatics that uses information and communications technologies to solve the big data problems in astronomy (Zhang and Zhao, 2015).

In this subheading, let us mention the characteristics of the data scientist who handles the data so we can understand more comprehensively how data becomes information. Then we will relate it with astronomy. An inquisitive mind set enables the data scientist to solve complex problems. Data scientists would normally work in multi-disciplinary teams. This means that you would normally develop an area of expertise and then work in a team to solve problems. There is no one qualification path that will enable you work as a data scientist. For example, some data scientists have a statistics or mathematics academic background, whereas others have combined statistics with computer science and computer programming. Data scientists would often have training in mathematics and statistics, modelling, and computer science and then learn specific technology skills and programming languages to be able to

complete data analysis tasks. There are different approaches to becoming a data scientist and it can be quite confusing once you start reading and talking to people about what matters. The following list of core skills areas is intended as a map to the various skills sets:

1. Fundamentals (including mathematics and data modelling)
2. Statistics (including probability theory, exploratory data analysis, hypothesis testing and regression)
3. Programming (Computer programming languages such as Python, statistical programmes such as R and commercial packages such as SPSS and Hadoop)
4. Machine Learning (knowing which techniques to apply using Python and R)
5. Text Mining/Natural Language Processing (text analysis, packages such as WEKA)
6. Data Visualization (using statistical packages to visualise and present data)
7. Big Data (including Hadoop: it is a free database program and the most widely used framework for distributed file system processing)
8. Data Getting (including data formats, data discovery, and data integration)
9. Data Munging (knowing how to clean data to be able to analyse it)
10. Toolbox (programs and packages that you should be familiar with)

There are specific techniques that have to be learned within each of these areas (https://www.unisa.ac.za/static/corporate_web/Content/About/Service%20departments/DCCD/Documents/career_data_science_math_stats_unisa.pdf).

The Virtual Observatory (VO) became real. The Virtual Observatory is the vision that astronomical datasets and other sources should work within as an uninterrupted whole. Many projects and data centres worldwide are working towards this goal. For example, the International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible. Examples of virtual observatory science are:

1. Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, x-ray, etc (to learn about the large-scale structure in the universe and the structure of our galaxy).
2. Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources (e.g., extremely distant or unusual quasars, new types, etc.).

3. Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations.

VO has also been successful because: all data are collected in a digital form, the computer- and data-enthusiast community, some appropriate standard formats, large data collections from funded-agency supported archives, an established culture of data sharing, a community initiative driven by the needs of an exponential data growth, federal agency support/funding, and data have no commercial value or privacy issues.

The positive aspects of the Virtual Observatory are: progress on interoperability, progress on standards etc., a global data grid of astronomy, empowering a broad community, some useful web services, community training, and outreach better than most other fields.

There are also negative aspects to VO: failing on data exploration and mining tools, should this be the level science reaches?, How much can be done effective science with little VO?, a slow community participation, and finally its own end (Djorgovski, 2017). What's better? Answer: Astroinformatics, as it can be bridged from the virtual observatory to astroinformatics as follows:

1. A bridge field connecting astronomy with computer science and engineering.
2. A mechanism for a broader community inclusion both as contributors and as consumers.
3. A mechanism for an interdisciplinary data science methodological sharing with other fields (Djorgovski, 2017).

The fields of Astrostatistics and Astroinformatics are vital for dealing with the actual big data issues in astronomy. The new disciplines of Astrostatistics and Astroinformatics have emerged in order to cope with the various challenges and opportunities offered by the exponential growth of astronomical data volumes, rates, and complexity.

The size of data repositories with the rapid growth of data volume from a variety of sky surveys has increased from gigabytes into terabytes and petabytes. Astroinformatics has appeared at an opportune time to deal with the challenges and opportunities generated by the massive data volume, rates, and complexity from new generation telescopes. This field of study uses data mining tools to analyze large astronomical repositories and surveys. Its many advantages are not only an efficient management of data resources but also the development of new valid tools intended for astronomical problems.

Different scientific areas have similar requirements concerning the ability to handle massive and distributed data sets and to perform complex knowledge discovery tasks on them. Data mining specialists have developed a lot of software and tools for solving various data mining tasks in different fields. Currently, there exist many successful application examples in the fields of business, medicine, science, and engineering. Researchers from astronomy, statistics, informatics, computers, and data mining are collaborating to focus on developing data mining software and tools for use in astronomy. Certainly some data mining tools from other fields may be directly used to overcome astronomical problems.

The arrival of the big data era in astronomy has led to a collaboration boom between astronomers, statisticians, computer scientists, data scientists, and information scientists. Collaboration is the only solution for scientists faced with difficulties and challenges caused by big data. Because of this, various organizations (see Table 2), for example, the International Astro-Statistical Association (IAA, affiliated to the International Statistical Institute), the American Astronomical Society Astroinformatics and Astrostatistics Working Group (AAS/WGAA) were established. Other groups are the Union Working Group in Astro-Statistics and Astroinformatics (IAU/WGAA), the Planned Large Synoptic Survey Telescope (LSST/ISSC) Consortium of Information and Statistics Sciences, the American Society of Statistics in Astroistatistics (ASA/IGA) and the IAA Study Cosmoistatistics Group (Zhang and Zhao, 2015).

4. Big Data in Astronomy

At present, the continuing construction and development of ground-based and space-born sky surveys ranging from gamma rays and x-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era. Astronomical data, already amounting to petabytes, continue to increase with the advent of new instruments. Astronomy, like many other scientific disciplines, is facing a data tsunami that necessitates changes to the means and methodologies used for scientific research. This new era of astronomy is making dramatic improvements in our comprehensive investigations of the Universe. Much progress is being made in the study of such astronomical issues as the nature of dark energy and dark matter, the formation and evolution of galaxies, and the structure of our own Milky Way. Astronomy research is changing from being hypothesis-driven to being data-driven to being data-intensive (Zhang and Zhao, 2015).

What is big-data in astronomy and astrophysics? Some of the big data providers in astronomy (ground and space based telescopes) are given with their abbreviations in Figure 1. They are also presented in tables and other forms.

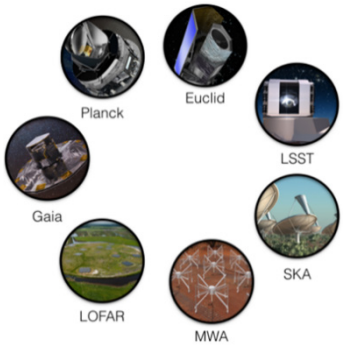


Figure 1: Wide and deep data and observations (McEwen, 2016)

The sky is not the limit for big data! We often hear terms such as "big data" and "data deluge". And it doesn't get much bigger than astronomy and satellite data! (https://adacs.org.au/wp-content/uploads/2018/01/10AstronomyThings.pdf)

Big data cannot be categorized into existing technological dimensions like data mining, algorithms and machine learning, or artificial intelligence. Big data are interconnected with those technologies and takes a new form during this process. As artificial intelligence becomes smarter, more autonomous and opaque, big data are transformed in novel ways. Without big data and the abundance of data available, none of the current improvements in technology would be possible. Big data are entangled in a complex way with data mining, algorithms and machine learning, and artificial intelligence. Big data enable those technologies to be better. On the other hand, big data are enabled by these technologies. Big data contribute to a cycle of technology and can be depicted as in Figure 2.

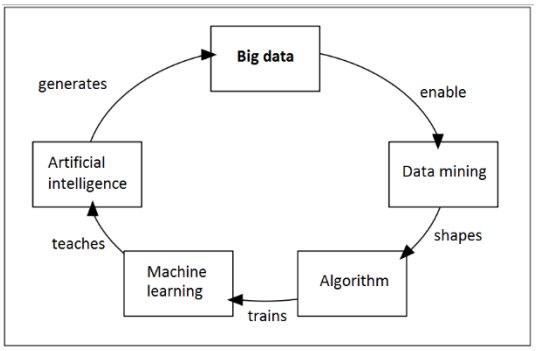


Figure 2: Big Data Technology Cycle (Scholz, 2017)

Researchers are parsing big data produced by the Hubble Space Telescope, the Large Hadron Collider and numerous other sources to learn more about the nature and origins of the

universe. These processes all involve large amounts of information that were once too vast and messy for even computers to analyze. Now that data can be mined for patterns and insights, some of which could spawn major advances in everything from theoretical physics to basic government services. In other words, big data are a chance to take all the things we don't know we know and finally know them (Marks, 2011).

5. Importance of Data in Astronomy

Rapid advances in technology and terabytes of data obtained in a day in astronomy have led to the discovery of new celestial bodies. Thus, in astronomy, data mining applications and algorithms, new decision trees and neurol networks were needed in astronomy for the rapid clustering and classification of these celestial bodies. Table 1 shows data mining tasks and their applications in astronomy (Zhang and Zhao, 2015).

Table 1. Applied approaches as well as their applications for he main data mining tasks in astronomy

Data Mining Tasks	Applied Approaches	Applications in Astronomy
Classification	Artificial Neural Networks (ANN)	Known knowns: - Spectral classification (stars, galaxies, quasars, supernovae) - Photometric classification (star and galaxies, stars and quazars, supernovae) - Morphological classification of galaxies - Solar activity
	Support Vector Machines (SVM)	
	Learning Vector Quantization (LVQ)	
	Decision Trees	
	Random Forest	
	K-Nearest Neighbors	
	Naive Bayesian Networks	
	Radial Basis Function Network	
	Gaussian Process	
	Decision Table	
	ADTree	
Regression	Artificial Neural Networks (ANN)	Known unkowns: - Photometric redshifts (galaxiex, quazars) - Stellar physical parameter measurement ([Fe/H], T_{eff}, logg)
	Support Vector Regression (SVR)	
	Decision Trees	
	Random Forest	
	K-Nearest Neighbor Regression	
	Kemel Regression	
	Principal Component Regression (PCR)	
	Gaussian Process	
	Least Squared	
	Regression Random	
	Forest	
	Partial Least Squares	

Clustering	Prencipal Component Analysis (PCA)	Unkown unknowns: - Classification - Special/rare object detection
	DBScan	
	K-Means	
	OPTICS	
	Cobweb	
	Self Organizing Map (SOM)	
	Expectation	
	Maximization	
	Hierarchical Clustering	
	AoutuClass	
	Gaussian Mixture Modeling (GMM)	
Outlier Dedection or Anomaly Detection	Prencipal Component Analysis (PCA)	Unkown unknowns: - Special/rare/ object detection
	K-Means	
	Ecpectation	
	Maximization	
	Hierarchical Clustering	
	One-Class SVM	
Time-Series Analysis	Artificial Neural Networks (ANN)	Known unkowns: - Novel detection - Trend prediction
	Support Vector Machines (SVM)	
	Random Forest	

Good organization was needed because ground-based and space-based observations led to collection of very large data. Sky surveys have been made available to astronomers by these organizations. Some of the astrostatistics and astroinformatics organizations are given in Table 2 (Zhang and Zhao, 2015).

Table 2. Astrostatistics and astroinformatics organizations

Organization	**Under community or project**	**Foundation Time**	**Chair**
International Astrostatistics Association (IAA)	The International Statistical Institude (ISI)	August 2012	Joseph Hilbe
IAU Working Group in Astrostatistics and Astroinformatics	The International Astronomical Union (IAU)	August 2012	Eric Feigelson
AAS Working Group in Astroinformatics and Astrostatistics	The American Astronomical Society (AAS)	June 2012	Zeljko Ivezic
ASA Interest Group in Astrostatistics	The American Statistical Association (ASA)	March 2014	Jessi Cisnewski
LSST Informatics and Statistics Science Collaboration	The Large Synoptic Survey Telescope (LSST)	Under construction	Kirk Borne
IAA Working Group on Cosmostatistics (renamed Cosmostatistics initiative, short for COIN)	The International Astrostatistics Association (IAA)	April 2014	Rafael de Souze

These organizations are also available on the ASAIP, Astrostatistics and Astroinformatics Portal of http://asaip.psu.edu. The ASAIP web site will provide links and resources to organizations devoted to the advancement of statistical and computational methodology for

astronomical research. It is intended to promulgate the organizations' work and assist in cross-fertilization between various organizations and interested individuals (Zhang and Zhao, 2015).

The broadest organizations are those associated with international societies:

1. International Astrostatistics Association (IAA): https://asaip.psu.edu/organizations/iaa
2. Special Interest Group in Astrostatistics (https://www.isi-web.org/index.php/news-from-isi/128-isi-astrostatistics-committee-and-network) within The International Statistical Institude (ISI): https://www.isi-web.org
3. Commission on Astroinformatics and Astrostatistics (https://www.iau.org/science/scientific_bodies/commissions/B3/info; under construction) within The International Astronomical Union (IAU): https://www.iau.org
4. The Astrominer Task Force (https://asaip.psu.edu/organizations/ieee-astrominer-task-force) of the Institute of Electrical and Electronics Engineers Computational Intelligence Society (IEEE-CIS): https://cis.ieee.org/data-mining-tc.html

The important U.S. national organizations also use these portions of the ASAIP web site:

1. The Working Group in Astroinformatics and Astrostatistics (https://asaip.psu.edu/organizations/aas-working-group-in-astroinformatics-and-astrostatistics) within the American Americal Society (AAS): https://aas.org
2. The Interest Group in Astrostatictics (https://asaip.psu.edu/organizations/asa-interest-group-in-astrostatistics) within the International Astrostatistics Association (IAA): https://www.amstat.org

One Project-level organization is using this web site:

1. The Informatics and Statistics Science Collaboration (https://asaip.psu.edu/organizations/lsst-informatics-and-statistics-Science-collaboration) of Large Synoptic Survey Telescope (LSST): https://www.lsst.org
2. International Astrostatistics Association (IAA): https://asaip.psu.edu/organizations/iaa
3. ISI Astrostatistics Special Interest Group: https://asaip.psu.edu/organizations/isi-astrostatistics-special-interest-group-sigastro

4. IAU Commission on Astroinformatics and Astrostatistics: https://asaip.psu.edu/organizations/iau-commission-on-astroinformatics-and-astrostatistics
5. IEE CIS Task Force on Mining ComplexAstronomical Data: https://asaip.psu.edu/organizations/ieee-astrominer-task-force
6. AAS Working Group in Astroinformatics and Astrostatistics: https://asaip.psu.edu/organizations/aas-working-group-in-astroinformatics-and-astrostatistics
7. ASA Interest Group in Astrostatistics: https://asaip.psu.edu/organizations/asa-interest-group-in-astrostatistics
8. LSST Informatics and Statistics Science Collaboration: https://asaip.psu.edu/organizations/lsst-information-and-statistical-science-collaboration
9. The Virtual Observatory (VO): http://www.ivoa.net; http://www.euro-vo.org; https://heasarc.gsfc.nasa.gov/vo/summary
10. The All Sky Virtual Observatory (ASVO): http://www.asvo.org.au/

The ASAIP provides searchable abstracts to recent papers in the field, several discussion forums, various resourches for research, brief articles by experts, lists of meetings, and access to various web resources such as on-line courses, books, jobs and blogs.

6. Data Sources in Astronomy

There are many sky surveys for ground-based and space-based observations at different wavelengths of the electromagnetic spectrum. The most important ones are given in Table 3 according to the data volumes (added to the table given by Zhang and Zhao, 2015). Similarly, foreseen data measurements of similar or different surveys are given in Figure 3. Kremer et al. (2017) provide detailed information about the data size of some large space surveys obtained overnight (see Figure 4). The data growth in only one area, for example the radio region, is given in Figure 5.

Table 3. Data volumes of different sky survey projects

Sky Survey Projects	**Data Volume**
Very Large Telescope (VLT, per night in 1998)	10 GB
The Hubble Space Telescope (HST; per week)	20 GB
Visible and Infrared Telescope for Astronomy (VISTA, per night in 2009)	315 GB
GAIA satellite (since 2014.07.25)	69585 GB
The Polamar Digital Sky Survey (DPOSS)	3 TB
The Two Micron All-Sky Survey (2MASS)	10 TB
Green Bank Telescope (GBT)	20 TB
The Galaxy Evolution Explorer	30 TB
The Sloan Digital Sky Survey (SDSS; 200 GB per night in 2000)	40 TB
Thirty Meter Telescope (TMT; per night in 2022)	90 TB
Sky Mapper Southern Sky Survey	500 TB
The Panoramic Survey Telescope and Rapid Responce System (PanSTARRS)	~40 PB expected
Cherenkov Telescope Array (CTA; by 2030)	~100 PB expected
The Square Kilometre Array (SKA Observatory; 500 and 1000 GB per second for low and mid, respectively)	~300 PB expected
The Large Synoptic Survey Telescope (LSST; ~2 TB per hour, 15-30 TB per night)	~4.6 EB expected

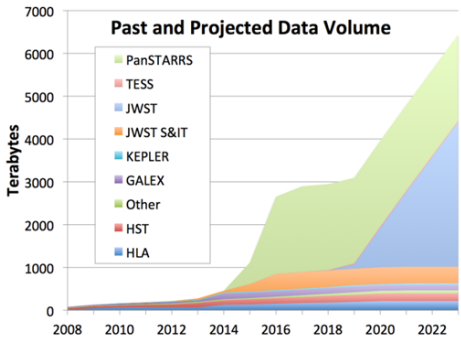


Figure 3: Past and projected growth in the data volume hosted by STScI's Mikulski Archive for Space Telescopes (MAST). In 2016, MAST's holdings exceeded 2.5 Petabytes (https://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf)

.

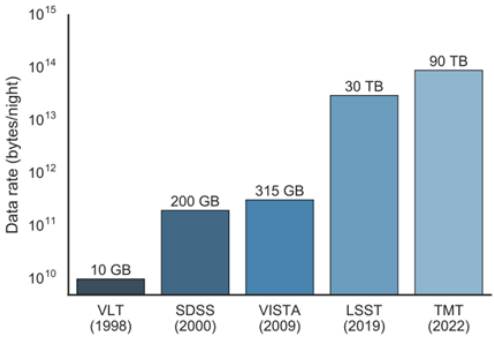


Figure 4: Increasing data volumes of existing and upcoming telescopes: Very Large Telescope (VLT), Sloan Digital Sky Survey (SDSS), Visible and Infrared Telescope for Astronomy (VISTA), Large Synoptic Survey Telescope (LSST) and Thirty Meter Telescope (TMT) (Kremer et al., 2017)

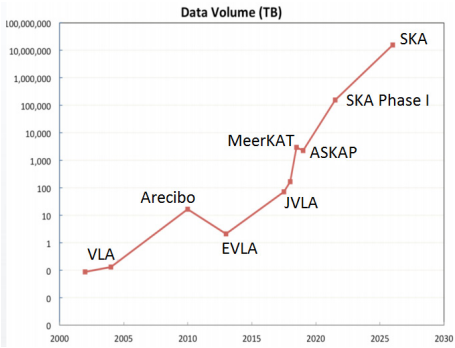


Figure 5: Exponential growth of radio data volumes. The horizontal axis is the year and the vertical is Terabayt (Raynard, 2017)

For developing data-rich astronomy it can be said that telescope + instrument is just a front end for data systems of live actions. Another example of a Big Data science driven by the advances in computing/information technology is presented in Figure 6.

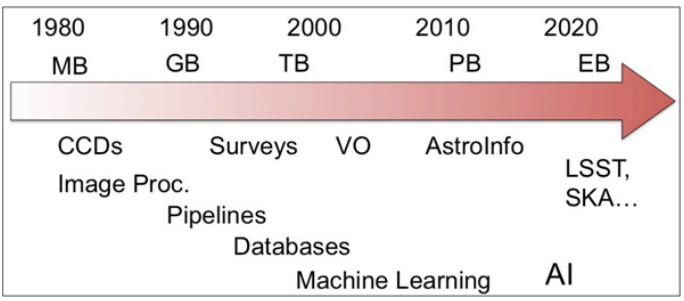


Figure 6: Evolving data-rich astronomy (Djorgovski, 2017)

Given the amount of data in the future archives, we expect that server-side analyses will be commonplace for the users, thus an advanced scripting capability must be supported (Big Data @ STScI). Accordingly, taking into account the number of tiles, years, and filters, the disk space needs for each survey can then be estimated (see Table 4; https://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf). Not surprisingly, for the surveys with large plate scales (PTF, ASASSN, ATLAS), the disk space needs for the stacks and difference images are small and on the order of a few Terabytes (TB). For the surveys with small plate scales (PS1 and LSST), the stacks and difference images are on the order of 1 Petabyte (PB), which goes beyond simple local storage systems, but is feasible on department or campus-wide computing centers. The input data volume in Table 4 seems to be significantly larger. For PTF and ASASSN, it is still feasible to store the input images locally. However, for ATLAS, PS1, and LSST, the input image set is on the order of a PB (ATLAS, PS1) and 17 PBs (LSST) due to the many epochs and/or small plate scales. In these cases, most likely the input images need to be accessed as needed via the Internet. In particular for LSST, this requires excellent connectivity to ensure the data transfer, data reduction and data analysis can be completed on timescales of just a few months (Morgan, 2018).

Table 4. Disk space and computing time requirements for the different surveys (sorted from top to bottom according to input data in the first column)

Survey Name	Input Image Data Volume (TB)	Stacked Image Data Volume (TB)	Stacked Image Processing Time (CPU Days)	Difference Image Data Volume (TB)	Difference Image Processing Time (CPU Days)
LSST	17,107	855	21,094	770	18,984
PanSTARRS	876	219	1080	164	4,050
ATLAS	475	0.8	2,344	0.5	52
PTF	51	3.4	253	2.6	253
ASASSN	6.8	0.04	33	0.03	3

In addition, there are many astronomical archives on the internet about the published articles in astronomy and the properties of celestial bodies. The information about the telescopes mentioned here can be learned from their related sites and was not also written in order not to convert this chapter into the basic astronomy. Some of these are:

1. The Sloan Digital Sky Survey: https://www.sdss.org
2. The Very Large Telescope array (VLT): https://www.eso.org/public/teles-instr/paranal-observatory/vlt
3. Thirty Meter Telescope (TMT): https://www.tmt.org
4. Square Kilometre Array (SKA): https://www.skatelescope.org

5. James Webb Space Telescope (JWST): https://www.jwst.nasa.gov
6. EUCLID Space Telescope: http://www.euclid-ec.org
7. PLANCK satellite: http://www.esa.int/planck
8. The International Event Horizon Telescope (EHT): https://eventhorizontelescope.org
9. Transiting Exoplanet Survey Satallite (TESS): https://www.nasa.gov/tess-transiting-exoplanet-survey-satellite
10. The Murchison Widefield Array (MWA): http://www.mwatelescope.org
11. The Galactic and Extra-Galactıc All-Sky MWA Survey (GLEAM): http://www.mwatelescope.org/gleam
12. The Galactıc and Extra-Galactic All-Sky MWA Extended Survey (GLEAM-X): http://www.mwatelescope.org/gleam-x
13. The Low-Frequency Array (LOFAR): http://www.lofar.org
14. The Solar & Heliospheric Observatory (SOHO): https://sohowww.nascom.nasa.gov
15. Sunspotter: https://www.sunspotter.org
16. Visible and Infrared Survey Telescope for Astronomy (VISTA): http://www.vista.ac.uk
17. Aladin Sky Atlas: https://aladin.u-strasbg.fr
18. SIMBAD Astronomical Database – CDS (Strasbourg): http://simbad.u-strasbg.fr/simbad
19. VizieR: http://vizier.cfa.harvard.edu
20. SAO/NASA ADS, Astrophysics Data System (ADS): http://cdsads.u-strasbg.fr
21. Minor Planet Center (MPC): https://minorplanetcenter.net/iau/mpc.html
22. SETI Institute: https://www.seti.org
23. The Galaxy Zoo website: https://www.galaxyzoo.org ; https://www.zooniverse.org/projects/zookeeper/galaxy-zoo
24. Much of the Kepler data for exoplanet discovery is publicly available through Mikulski Archive for Space Telescopes: http://archive.stsci.edu/kepler
25. Kepler Spacecraft: https://www.nasa.gov/mission_pages/kepler/main/index.html
26. Debrecen Sunspot Data archive: http://fenyi.solarobs.unideb.hu/ESA/HMIDD.html

The main purpose of the Gaia (http://sci.esa.int/gaia) in Table 3 is to examine our galaxy and star contents and to provide high precision astrometric and photometric parameters. The satellite will also conduct many other investigations, in particular observing millions of small galaxies of the local universe, a significant number of supernovae, and interactive binary systems as a large part of transient events.

The Gaia satellite (https://www.cosmos.esa.int/web/gaia/mission-numbers) continues to provide precise location data. Continuity in the acquisition of observational data within the framework of Gaia campaigns with ground-based telescopes has provided and so increased the performance of the data. A recent example of this was the binary microlensing event of the Gaia16eye object (Wyrzykowski et al., 2020). Gaia16aye is the first of its kind, the first activity discovered by the Gaia space mission and discovered in the direction of the Northern Galactic Disc. The light curve exhibited five different maximum brightnesses of up to 11 magnitudes, and the event was elaborated with approximately 25000 data points collected by the telescope network organized by the Gaia team for 500 days. This study demonstrated the potential of the microlensing method to question the mass function of dark objects, including black holes, in other directions from the Galactic overhang. This also emphasizes the importance of long-term coordinated observations with a network of heterogeneous telescopes.

The observation of all asteroids by a single observatory is not possible because their number, being more than a million, is too big to handle. For this reason, it is necessary that all astronomers and sky enthusiasts on Earth should work together. All observations of asteroids, comets and natural satellite observations of the Solar System by observers or amateur astronomers are collected in the Minor Planet Center (MPC; https://minorplanetcenter.net/iau/mpc.html). The ephemeris information and the assigned trajectory parameters are published on the web address which is open to all internet users worldwide. At the same time the accuracy of the parameters obtained can be controlled by anyone interested in this field (Kaynar, 2019).

7. Big Data for Explorations of the Universe

Machine Learning techniques have begun to find applications in astronomy, but mainly for “clerical” tasks, such as error checking, and bulk classification. This leaves vast scope for harnessing Machine Learning for more interesting tasks that enable new scientific discoveries. Historically, major discoveries have often relied on serendipity; an expert examines new data with an eagle eye and an open mind. However effective this approach has been in the past, it does not scale. New astronomical datasets are too massive and complex for any individual or group of experts to look at every aspect, object, or measurement. Yet modern algorithms for

sequencing, classification, or anomaly detection can provide us with methods to uncover new phenomena (Mazeh & Poznanski, 2018).

Within the next few years, image analysis and machine learning systems that can process terabytes of data in near real-time with high accuracy will be essential (Gómez-Vargas, 2018). There are great opportunities for making novel discoveries, even in databases that have been available for decades. The volunteers of Galaxy Zoo have demonstrated this multiple times by discovering structures in SDSS images that have later been confirmed to be new types of objects. These volunteers are not trained scientists, yet they make new scientific discoveries. Even today, only a fraction of the images of SDSS have been inspected by 12 humans. Without doubt, the data still hold many surprises, and upcoming surveys, such as LSST, are bound to image previously unknown objects. It will not be possible to manually inspect all images produced by these surveys, making advanced image analysis and machine learning algorithms of vital importance. One may use such systems to answer questions like how many types of galaxies there are, what distinguishes the different classes, whether the current classification scheme is good enough, and whether there are important sub-classes or undiscovered classes. These questions require data science knowledge rather than astrophysical knowledge, yet the discoveries will still help astrophysics tremendously. In this new data-rich era, astronomy and computer science can benefit greatly from each other. There are new problems to be tackled, novel discoveries to be made, and above all, new knowledge to be gained in both fields (Kremer et al., 2017).

Big Data is transforming how astronomers make discoveries. The next game-changer is likely lurking in the data we already have but it will take scientists years to uncover it. Earlier in 2018, astronomers stumbled upon a fascinating finding: Thousands of black holes likely exist near the center of our galaxy. The x-ray images that enabled this discovery weren't from some state-of-the-art new telescope. Nor were they even recently taken - some of the data was collected nearly 20 years ago. The researchers discovered the black holes by digging through old, long-archived data. Discoveries like this will only become more common, as the era of "big data" changes how science is done. Astronomers are gathering an exponentially greater amount of data every day so much that it will take years to uncover all the hidden signals buried in the archives. Sixty years ago, the typical astronomer worked largely alone or in a small team. They likely had access to a respectably large ground-based optical telescope at their home institution. Their observations were largely confined to optical wavelengths more or less what the eye can see. That meant they missed signals from a host of astrophysical sources, which can emit non-visible radiation from very low-frequency radio

all the way up to high energy gamma-rays. For the most part, if you wanted to do astronomy, you had to be an academic or eccentric rich person with access to a good telescope. Old data were stored in the form of photographic plates or published catalogs. But accessing archives from other observatories could be difficult and it was virtually impossible for amateur astronomers. Today, there are observatories that cover the entire electromagnetic spectrum. No longer operated by single institutions, these state-of-the-art observatories are usually launched by space agencies and are often joint efforts involving many countries. With the coming of the digital age, almost all data are publicly available shortly after it is obtained. This makes astronomy very democratic - anyone who wants to can reanalyze almost any data set that makes the news (you too can look at the Chandra data that led to the discovery of thousands of black holes!). These observatories generate a staggering amount of data. For example, the Hubble Space Telescope (HST), operating since 1990, has made over 1.3 million observations and transmits around 20 GB of raw data every week, which is impressive for a telescope first designed in the 1970s. The Atacama Large Millimeter Array (ALMA) in Chile now anticipates adding 2 TB of data to its archives every day. The archives of astronomical data are already impressively large. But things are about to explode. Each generation of observatories is usually at least 10 times more sensitive than the previous, either because of improved technology or because the mission is simply larger. Depending on how long a new mission runs, it can detect hundreds of times more astronomical sources than previous missions at that wavelength. For example, compare the early EGRET gamma ray observatory, which flew in the 1990s, to NASA's flagship mission Fermi, which turns 10 in 2018. EGRET detected only about 190 gamma ray sources in the sky. Fermi has seen over 5,000. The Large Synoptic Survey Telescope (LSST), an optical telescope currently under construction in Chile, will image the entire sky every few nights (Estévez, 2016). It will be so sensitive that it will generate 10 million alerts per night on new or transient sources, leading to a catalog of over 15 petabytes after 10 years. The Square Kilometre Array (SKA), when completed in 2020, will be the most sensitive telescope in the world, capable of detecting airport radar stations of alien civilizations up to 50 light-years away (Scaife, 2016 and 2019). In just one year of activity, it will generate more data than the entire internet. These ambitious projects will test scientists' ability to handle data. Images will need to be automatically processed meaning that the data will need to be reduced down to a manageable size or transformed into a finished product. The new observatories are pushing the envelope of computational power, requiring facilities capable of processing hundreds of terabytes per day. The resulting archives all publicly searchable will contain 1 million times more information that what can be stored on your typical 1 TB backup disk. The data deluge will

make astronomy become a more collaborative and open science than ever before. Thanks to internet archives, robust learning communities and new outreach initiatives, citizens can now participate in science. For example, with the computer program Einstein@Home, anyone can use their computer's idle time to help search for gravitational waves from colliding black holes. It's an exciting time for scientists, too. Astronomers often study physical phenomena on timescales so wildly beyond the typical human lifetime that watching them in real-time just isn't going to happen. Events like a typical galaxy merger (which is exactly what it sounds like), can take hundreds of millions of years. All we can capture is a snapshot, like a single still frame from a video of a car accident. However, there are some phenomena that occur on shorter timescales, taking just a few decades, years or even seconds. That's how scientists discovered those thousands of black holes in the new study. It's also how they recently realized that the x-ray emission from the center of a nearby dwarf galaxy has been fading since first detected in the 1990s. These new discoveries suggest that more will be found in archival data spanning decades. In Meyer's (2018) work, she used Hubble archives to make movies of "jets" of high-speed plasma ejected in beams from black holes. She used over 400 raw images spanning 13 years to make a movie of the jet in nearby galaxy M87. That movie showed, for the first time, the twisting motions of the plasma, suggesting that the jet has a helical structure. This kind of work was only possible because other observers, for other purposes, just happened to capture images of the source she was interested in, back when she was in kindergarten. As astronomical images become larger, higher resolution and ever more sensitive, this kind of research will become the norm (Meyer, 2018).

8. Ways of Processing Big Data in Astronomy

Now and the next decade promises to be an exciting time for astronomers. Large volumes of astronomical data are continuously being collected from highly productive space missions. These data have to be efficiently stored and analyzed in such a way that astronomers maximize their scientific return from these missions. Recognizing the need to better handle astronomical datasets, we designed ASTROIDE, a distributed data server for astronomical data. We analyze the peculiarities of the data and the queries in cosmological applications and design a new framework where astronomers can explore and manage vast amounts of data. ASTROIDE introduces effective methods for efficient astronomical query execution on Spark through data partitioning with HEALPix and customized optimizer. ASTROIDE offers a simple, expressive and unified interface through ADQL, a standard language for querying databases in astronomy. Experiments have shown that ASTROIDE is effective in processing astronomical data, scalable and outperforms the state-of-the-art (Brahem et al., 2018).

Mehta et al. (2017) presented the first comprehensive study of large-scale image analytics on big data systems. They surveyed the different paradigms of large-scale data processing platforms using two real-world use cases from domain sciences. While they could execute the use cases on these systems, their analysis shows that leveraging the benefits of all systems requires deep technical expertise. Overall, they argue that current systems provide good support for image analytics, but they also open new opportunities for further improvement and future research.

Zhang et al. (2016) investigated the idea of leveraging the modern big data platform for many-task scientific applications. Specifically, they built Kira (https://github.com/BIDS/Kira), a flexible, scalable, and performant astronomy image processing toolkit using Apache Spark running on Amazon EC2 Cloud. They also presented the real world Kira Source Extractor application, and use this application to study the programming flexibility, dataflow richness, scheduling capacity and performance of the surrounding ecosystem. They also demonstrated that Apache Spark can integrate with a pre-existing astronomy image processing library. This allows users to reuse existing source code to build new analysis pipelines. They believe that Apache Spark's flexible programming interface, rich dataflow support, task scheduling capacity, locality optimization, and built-in support for fault tolerance make Apache Spark a strong candidate to support many-task scientific applications. Apache Spark is one (popular) example of a Big Data platform. They learned that leveraging such a platform would enable scientists to benefit from the rapid pace of innovation and large range of systems and technologies that are being driven by widespread interest in Big Data analytics. Their experience with Kira demonstrates that data intensive computing platforms like Apache Spark are a performant alternative for many-task scientific applications.

Examples of other special methods for analyzing big data include: Bayesian analysis, MCMC sampling, hierarchical probabilistic (Bayesian) models, variable selection, experimental design, machine learning, optimisation, wavelets, sparsity, compressed sensing, and finally Astrostatics and Astroinformatics.

We conclude the end of this chapter of the book with this last sentence: collaborations between astronomers, statisticians and information scientists have begun, but need to be expanded. The International Statistical Institute and similar astronomical organisations should be promoting to continue these collaborations (Feigelson & Babu, 2012). Bigger data is not always better data and may big data be with you (Scholz, 2017).

9. Conclusion

Thanks to the big data in astronomy and Machine Learning algorithms, there have been great advances in Astroinformatics and Astrostatistics studies. Very valuable information has been obtained about micro and macro structures of the universe. Wide horizons have been opened to astronomers about dark matter, dark energy, supernovae, novae and galaxies. It is tried to develop algorithms that can make predictions about solar eruptions by observing the atmosphere of our sun continuously. In addition, the research continues with large data following the trajectory movements of asteroids threatening our world. Through surveys such as GAIA, LSST and TMT, there will be many new developments and discoveries in Physics, Astrophysics, Astronomy, and Cosmology.

In this chapter we tried to guide those who want to study Astronomy, Astroinformatics and Astrostatistics. We can say that the field of astronomy is at the top of the big data in the world. We hope the references section of this chapter will guide enthusiasts. Everything about big data in astronomy is almost impossible to write here. For this reason, the article titled "the Astro2020 Science White Paper. The Next Decate of Astroinformatics and Astrostatistic" prepared by Aneta Siemiginowska (2019) together with 34 authors is a good reference for those who wish to work in these fields.

Acknowledgments. We thank Demir IT Company (Eskisehir, Turkey) for providing computer support. HHE thanks TUBITAK National Observatory (TUG) for the experience from several projects on follow-up observations of the Gaia satellite since started scientific operations in mid-2014.

References

Brahem, M., Yeh, L., Zeitouni, K. (2018). *ASTROIDE: A Unified Astronomical Big Data Processing Engine over Spark*. A Preprint, October 25.

Dindar, M., Helhel, S., Esenoglu, H., Parmaksizoglu, M. (2015). A new software on TUG-T60 autonomous telescope for astronomical transient events. *Experimental Astronomy*, 39(1), 21–28

Djorgovski, S., G. (2017). *Astronomy in the Era of Big Data- From Virtual Observatory to Astroinformatics and beyond*. TIARA Summer School on Astrostatistics and Big Data Taipei, Taiwan, September.

Gómez-Vargas, G. (2018). *First Ideas to Connect Astronomical Data, Deep Learning and Image Analysis, Accelerating the search of dark matter with machine learning*. Lorentz Center, Leiden, January.

Estévez, P. (2016). *Big Data Era Challenges and Opportunities in Astronomy: How SOM/LVQ and Related Learning Methods Can Contribute?* WSOM 2016 Houston, TX, January 8.

Feigelson, E. D. & Babu, G. J. (2012). Big data in astronomy. *Significance*, The Royal Statistical Society, August.

Kaynar, S. (2019). *Determination of the Trajectory of Selected Several Near Earth Asteroids and Investigation Their Physical Properties*. Akdeniz University Graduate School of Natural and Applied Sciences Department of Physics Master Thesis (October).

Kremer, J., Kristoffer, S. S., Gieseke, F., Steenstrup, K. P., Igel, C. (2017). Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, *IEEE Intelligent Systems*, 32, 16–22, March–April (https://arxiv.org/abs/1704.04650).

Marks, J. (2011). *5 Things You Need to Know About Big Data*. NetApp. Veteran Data Solutions-VetDS

Mazeh, T. & Poznanski, D. (2018). *Big Data and Exo-Planets*, Proposal for a research group in Astronomy.

McEwen, J. (2016). *Big-Data in Astronomy and Astrophysics Extracting Meaning from Big-Data*.

(https://indico.hephy.oeaw.ac.at/event/86/session/3/contribution/1/material/slides/0.pdf)

Mehta, P., Dorkenwald, S., Zhao, D., Kaftan, T., Cheung, A., Balazinska, M., Rokem, A. (2017). *Comparative Evaluation of Big-Data Systems on Scientific Image Analytics Workloads*. Andrew Connolly, Jacob Vanderplas, Yusra AlSayyad, Proceedings of the VLDB Endowment, Vol. 10, No. 11.

Meyer, E. (2018). Big Data is Transforming How Astronomers Make Discoveries, *The Conversation*, May 15 (https://theconversation.com/the-next-big-discovery-in-astronomy-scientists-probably-found-it-years-ago-but-they-dont-know-it-yet-95280?xid=PS_smithsonian)

Morgan, H. (2018). *Large Synoptic Survey Telescope (LSST) Scaling Issues and Network Needs*, Pacific Northwest Gigapop Meeting October 23.

Raynard, L. (2017). *Radio Astronomy & SDGs A Justification or Solution?* South African Radio Astronomy Observatury (SARAO), September 4.

Scaife, A. (2019). *Big Telescope, Big Data: Towards Exa-Scale With the SKA, Numerical algorithms for high-performance computational science*. Royal Society 8–9 April.

Scaife, A. (2016). *Big Telescope, Big Data: Indirect Imaging in the SKA Era*, IAU Astroinformatics, Sorrento.

Scholz, T. M. (2017). *Big data in organizations and the role of human resource management: A complex systems theorybased conceptualization*. Econstor, Personalmanagement und Organisation, No. 5, Peter Lang International Academic Publishers (http://hdl.handle.net/10419/182489)

Siemiginowska, A., Eadie, G., Czekala, I. with 33 authors. (2019). *Astro2020 Science White Paper: The Next Decade of Astroinformatics and Astrostatistics*. 15 March. (https://arxiv.org/abs/1903.06796)

Wyrzykowski, L. et al. (2020). Full Orbital Solution for the Binary System in the Northern Galactic Disk Microlensing Event Gaia16aye, *Astronomy & Astrophysics*, in pressed (633, A98.)

Zhang, Y., Zhao, Y. (2015). Astronomy in the Big Data Era. *Data Science Journal*, 14(11), 1–9 (https://datascience.codata.org/articles/10.5334/dsj-2015-011).

Zhang, Z., Barbary, K., Nothaft, F. A., Sparks, E. R., Zahn, O., Franklin, M. J., Patterson, D. A., Perlmutter, S. (2016). Kira: Processing Astronomy Imagery Using Big Data Technology, *IEEE Transactions on Big Data*, DOI: 10.1109/TBDATA.2016.2599926.

CHAPTER 3

DATA STORAGE IN THE DECENTRALIZED WORLD: BLOCKCHAIN AND DERIVATIVES

Enis KARAARSLAN*, Enis KONACAKLI**

*Assistant Professor, Mugla Sitki Kocman University, Department of Computer Engineering, Mugla, Turkey.
E-mail: enis.karaarslan@mu.edu.tr
**Eskisehir Technical University, Department of Computer Engineering, Eskisehir, Turkey.
E-mail: enisk@eskisehir.edu.tr

DOI: 10.26650/B/ET06.2020.011.03

Abstract

We have entered an era where the importance of decentralized solutions has become more obvious. Blockchain technology and its derivatives are distributed ledger technologies that keep the registry of data between peers of a network. This ledger is secured within a successive over looping cryptographic chain. The accomplishment of the Bitcoin cryptocurrency proved that blockchain technology and its derivatives could be used to eliminate intermediaries and provide security for cyberspace. However, there are some challenges in the implementation of blockchain technology. This chapter first explains the concept of blockchain technology and the data that we can store therein. The main advantage of blockchain is the security services that it provides. This section continues by describing these services.. The challenges of blockchain; blockchain anomalies, energy consumption, speed, scalability, interoperability, privacy and cryptology in the age of quantum computing are described. Selected solutions for these challenges are given. Remarkable derivatives of blockchain, which use different solutions (directed acyclic graph, distributed hash table, gossip consensus protocol) to solve some of these challenges are described. Then the data storage in blockchain and evolving data solutions are explained. The comparison of decentralized solutions with the lcentralized database systems is given. A multi-platform interoperable scalable architecture (MPISA) is proposed. In the conclusion we include the evolution assumptions of data storage in a decentralized world.

Keywords: Data, Data storage, Distributed ledger technology, Security, Cryptology, Blockchain, Scalability, Blockchain derivatives, Directed acyclic graph, Gossip consensus protocol, Sidechain

1. Introduction

We are now entering an era where people seek solutions for eliminating intermediaries. The processes can be made faster, while they became less bureaucratic. These solutions can be possible with decentralized solutions; blockchain technology and its derivatives. We mean the "blockchain frameworks" which implement this technology, when we use the term "blockchain technology".

Decentralized solutions are important, as they establish trust without using any intermediary. They do not depend on a central node and are more fault-tolerant and resistant to attacks than traditional solutions. These solutions work as peer-to-peer (P2P), which allows direct communication between peers via the Internet (Karaarslan & Adiguzel, 2018). BitTorrent is one of the most successful implementations of the P2P file-sharing protocol (Alves et al., 2018).

Decentralized solutions can be used to eliminate intermediaries like banks, notary, etc. Bitcoin (BTC) cryptocurrency is a working example of how it can be done. As described in (Brennan et.al, 2018), "cryptocurrencies are only the beginning". Autonomous codes are devised to make the processes autonomous and work without intermediaries. Decentralized applications (Dapp) allow us to have answers within a distributed and secured network (Karaarslan & Adiguzel, 2018).

This chapter aims to describe blockchain technology and to show the differences in its purpose and design. In section 2 we start with a brief explanation of blockchain technology. Blockchain technology fundamentals and security services are described. Data storage in blockchain is addressed here. The challenges of blockchain technology and some remarkable solutions are described in Section 3. Blockchain anomalies, energy consumption, scalability, speed; interoperability, privacy, and cryptology challenges in the age of quantum computing are addressed here. The decentralized derivatives (Tangle, Hashgraph, Holochain) and their technological differences are described in Section 4. Data storage in decentralized systems is covered in Section 5. Evolving data solutions for the decentralized systems and hybrid solutions are given here. Decentralized solutions are compared with centralized databases. A multi-platform interoperable scalable architecture (MPISA) is proposed in Section 6. Finally, results and conclusions are given.

2. Blockchain Technology

Blockchain is atechnology, which is used for the keeping of a list of records in a (semi-) decentralized manner. These records contain information about any transaction or any

program code (smart contract) which allows a system to work autonomously (Alves et al., 2018; Ali et al., 2017). The records are aggregated in data structures and called blocks. These blocks are linked to each other using cryptographic techniques and thus form a chain structure. The registry, which keeps this blockchain, is called the ledger. Blockchain keeps the ledger distributed and is also called distributed ledger technology (DLT). The ledger is kept in several devices, which are called nodes. These nodes are connected using P2P protocols. These nodes can act as servers or clients at the same time and form a decentralized system. Nodes with different hardware can have different functions, which are summarized in Table 1 (Barnas, 2016). These nodes use consensus protocols to make a common decision on operations, such as the choice regarding who will write the new block. The new block is written by the selected node and then distributed to all nodes.

Table 1. Blockchain Node types

Node Type	***Function***	***Examples***
Full Node	Keep full copy of the blockchain, Generate blocks, Validate blocks, Validate transactions, Generate new transaction and broadcast.	Servers or personal computers with sufficient hardware resources
Partial/Half Node	Keep only partial copy of the blockchain, Validate blocks, Validate transactions, Validate old records as peer support, Generate new transaction and broadcast.	Laptops or alike
Simple Node	Validate new transactions, Generate new transaction and broadcast.	IoT or limited capacity mobile devices

Blockchain is not a suitable solution for all computational issues and neither for all data storage problems. The need of a blockchain solution is discussed in detail in Wüst and Gervais' paper (Wüst & Gervais, 2018) and also summarized in Fig. 1. A blockchain solution is suitable under the following conditions:

- If the domain has a dataset which is to be shared with more than one party,
- Where there is low trust between parties and there is no trusted third-party to ensure trust,
- In cases of a need for auditing.

In a scenario of a supply chain, a company may want to track all the processes in the supply chain and even make it transparent to its users. As it is shown in Fig. 2, it becomes

complex even in a scenario of two companies (a producer and a consumer). All the transport means, authorities, banks, and others need to share the data or generate transactions during this process. Blockchain is a good solution in a scenario like that, where there are many parties that have to trust each other (Mohan, 2019).

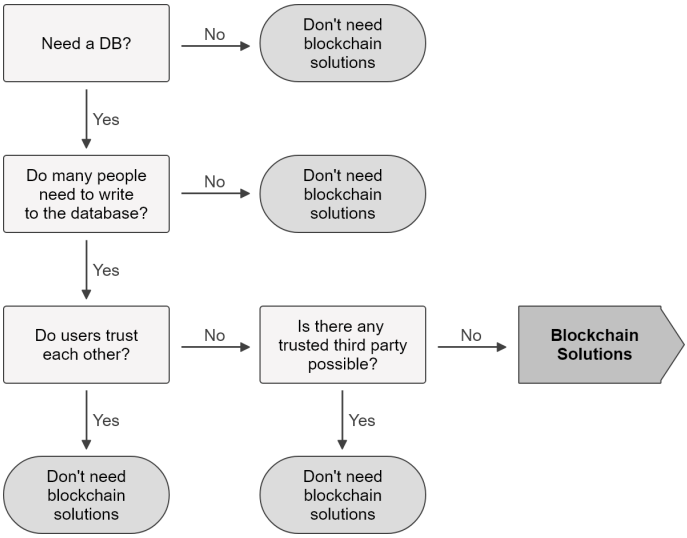


Figure 1: Do you need a blockchain?

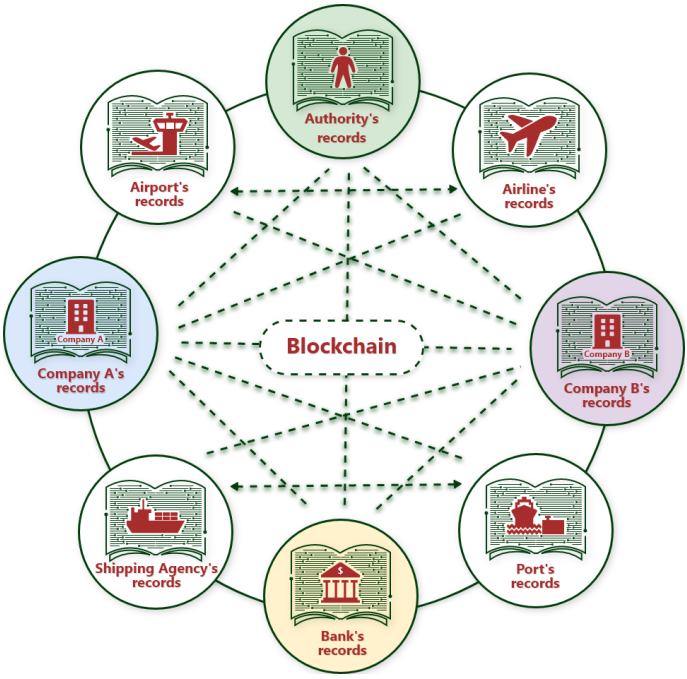


Figure 2: Multi-party data access scenario

The blockchain solution will help in keeping the records of the transactions. The records should be reachable at all times, unmodifiable and inerasable (White et al., 2017). The system will work in an autonomous way, which will ensure trust in the system. Full trust, complete privacy, and decentralization should be aimed at when creating such decentralized systems (Karaarslan & Akbaş, 2016).

Different blockchain implementations, which depend on the anonymity and trustworthiness of the validator (node), are possible, as shown in Fig. 3 (Gür et al., 2019). These are:

- Allowing nodes to join the network with or without permission (permissionless),
- Allowing public or private access to the ledger,

Different consensus protocols such as proof of work (PoW), proof of stake (PoS), proof of authority (PoA), practical byzantine fault tolerance (PBFT) and such like are preferred in accordance with the anonymity and trustworthiness of the node.

There are also hybrid blockchain implementations, which allow different types to work together to achieve a function. Some implementations can have public and private ledgers together. Implementations like federated (consortium) blockchain allow multiple organizations to share information privately between parties (Bauer, 2015).

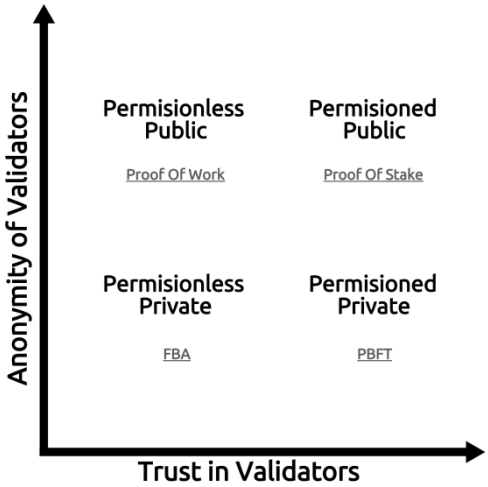


Figure 3: Blockchain implementation types per anonymity/trust of validators

Cryptocurrency implementations mostly use permissionless validators and public blockchain. The users are anonymous or pseudonymous. The term 'permissionless' states that any node can enter or leave the system without permission. The trust in the validator is

low as the nodes are anonymous (Karaarslan & Akbaş, 2016). Cryptocurrency implementations depend on using their own currencies to run transactions on their systems. Bitcoin can be called the first blockchain implementation of this type, which has been active since 2008. Bitcoin is the proof-of-concept that this type of system can work and have value. Satoshi has proposed a model in his paper (Nakamoto, 2008), where the system generates a new crypto coin per block and gives an award to the owner of the node that will write the block. Ethereum (ETC) introduced a framework where new blockchain applications can be developed. Smart contracts are used which are in the form of an autonomous software code on the blockchain.

The steps in the process of making a value (cryptocurrency) transfer in such a blockchain network are given in Fig. 4. In this scenario, Fatih wants to make a value (cryptocurrency) transfer to Eylul. Most cryptocurrency systems use "mining pools", which orchestrate such a process. The nodes in the P2P network validate the transactions (account balance check, double spending check) and collect the validated transaction data. According to the protocol used, nodes collect information of variable number of different transactions in a specified time. PoW consensus protocol is used to select the node which will write the new block. PoW depends on a calculation to solve a puzzle like a mathematical problem. The node, which solves the problem, will first be selected. The selected node will form and write the block and advertise it in the network (Karaarslan & Akbaş, 2016). The fairness of the node selection and the security of such a process results in high energy and time-consuming operations (Gür et al., 2019). PoW and alike consensus protocols, have a bad reputation on high energy usage, which is said to affect climate change. The blocks are transparent and that means the transaction information is visible through web interfaces, which are called explorers such as the block explorer. These web interfaces also show detailed information about that cryptocurrency system.

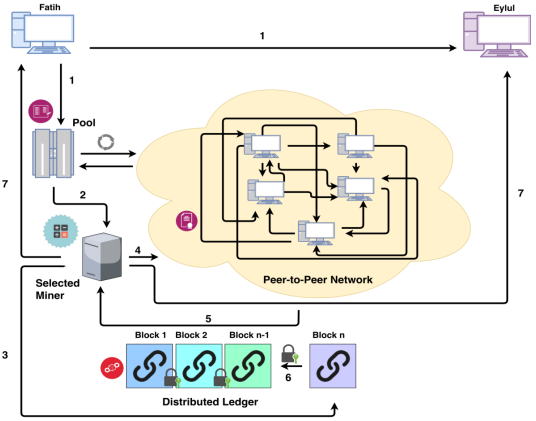


Figure 4: Transaction steps of a value transfer in cryptocurrency implementations

Different consensus protocols can be developed and deployed, which consume less energy and are faster than PoW. POS and alike consensus protocols are being tested with cryptocurrency implementations (Zheng et al., 2018). This will be covered in section 3.2.

The needs of enterprise implementations are different from cryptocurrency implementations. The identity of the users is known. Permissioned validators and private/public blockchains are mostly used. Different parties of the blockchain system supply the validator nodes. The validators are trusted and not anonymous, which means they are under the control of the management. PoW consensus protocol is not necessary. PBFT, PoA and similar consensus protocols are preferred in this type of implementation (Zheng et al., 2018). Hyperledger Fabric, R3 Corda (Valenta & Sandner, 2017) and Quorum can be given as examples.

Hyperledger Fabric is widely used in production (Hyperledger, 2018) and in academia (Androulaki et.al, 2018; Nasir et al, 2018). Such an implementation scenario in Hyperledger Fabric is given in Fig. 5, which consists of a customer and his/her IoT device, two companies, and one authority. The customer can be subscribed to different companies and there is also one authority that these companies have to share their data with. The IoT device of the customer sends a summary of collected data to the blockchain network. Each company creates a group (channel) among themselves. Different consensus protocols and different types of nodes can be used in each group. These nodes, rest server, and CA server can all be installed as Docker containers. Each node has limited authority. The owner of each transaction is identified in its own certificate authority. These groups also share data with the authority, which is labeled as Auth. C Peer in this case (Gür et al., 2019).

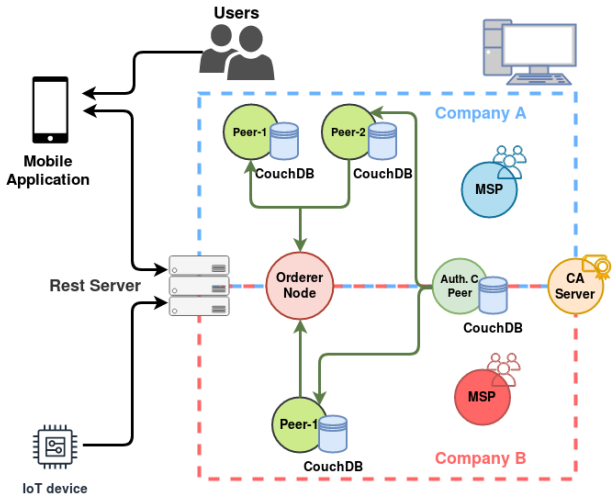


Figure 5: Hyperledger Fabric enterprise blockchain solution

Many cloud services have started to provide environments of blockchain as a service (BaaS), which serve cloud services to build Dapps. By way of example, IBM, Alibaba, Huawei, and many others provide Hyperledger Fabric based BaaS (Mohan, 2019).

A comparison of the security services provided by blockchains, central databases and distributed databases are given in Table 2 (Bozic et al., 2016). Data integrity, availability and fault tolerance services can best be provided with blockchain.

The integrity of the data is established by the design of the DLT. Each block is connected to the previous one using its hash value. Hash functions (SHA-256, Keccak-256 … etc) are one-way functions that form the fingerprint of the input data. This data structure makes the binding so strong that, when an attacker wants to change block n, the blocks starting from the nth block, until the last block, have to be modified and rewritten according to the change. It should also be noted that the selected node of each block is also recorded in the ledger and such an attempted attack will easily be detected (Karaarslan & Akbaş, 2016).

The availability and security of the system depends on the number of nodes and their distribution in the network. More nodes will make the system stronger against the attacks. Then taking control of the majority of the nodes by the attacker will be harder and the compromised nodes will not be able to misguide the block creation process. If the nodes are more widely distributed in different networks, the network will also be stronger against DDoS attacks. Fault tolerance is the ability of the blockchain to correct any misuse and errors. This is implemented by using consensus protocols.

Privacy is not a design concern in most implementations such as cryptocurrencies. Privacy and security in Bitcoin are investigated in (Conti et al., 2018). Different implementations can have different levels of privacy. Transparent records do not mean the privacy level is low. Personal data is not revealed, transactions are only traceable with the public addresses. Transparency property is used to prevent any possible fraud and misuse. It can be used to enable safer environments (Ölmez & Karaarslan, 2019).

Table 2. Comparison of the security services

	Blockchain	***Central Database***	***Distributed Database***
Integrity	High	Average	Average
Availability	High	Low	Average
Fault tolerance	High	Low	High
Privacy	Variable*	High	Average

** Privacy is not by design. Mainly depends on the implementation*

There are nascent standardization efforts, which focus on narrow aspects of blockchain (Mohan, 2019). IEEE Blockchain Initiative has just started several blockchain standardization efforts focusing on areas like agriculture, medicine and IoT (IEEE, 2019). ISO/TC 307 technical committee is working on blockchain and distributed ledger technologies (ISO, 2019). W3C community group is working on the Web Ledger Protocol, which will describe the format and protocol of decentralized ledgers on the web (W3C, 2019).

3. Meeting the Challenges

Despite the opportunities of blockchain technology, the challenges of blockchain are still notable for discussion. The challenges can be summarized as follows:

- Blockchain anomalies,
- Energy consumption,
- Scalability and speed,
- Interoperability,
- Privacy,
- Cryptology challenges in the age of quantum computing.

Some note-worthy solutions proposed and studied are given in the subsections.

3.1. Blockchain Anomalies

Some anomalies may result in the addition of conflicting blocks and the formation of new branches of the chain in PoW based blockchains. The conditions, which may lead to these anomalies, are covered in Natoli and Gramoli's study (Natoli & Gramoli, 2016). This can cause usability, integrity and performance problems (Mohan, 2019). Blockchain implementations should give deterministic guarantees on these conditions. Implementations can be adapted and smart contracts can be written to overcome these types of anomaly (Natoli & Gramoli, 2016).

3.2. Energy Consumption

Mining operations of the conventional PoW based blockchain systems require expensive hardware and a very high degree of energy consumption (Flipo & Berne, 2017; Trautman & Molesky, 2019). Energy efficient solutions, which will replace or minimize the usage of the conventional PoW based blockchain systems, are being experimented. Different node selection algorithms are proposed which are based on random choice or on the cryptocurrency amount of the miners (Rosic, 2017).

POS consensus protocol has started to be preferred in cryptocurrency implementations instead of the PoW approach. The nodes have to deposit a predefined amount of cryptocurrency and show their commitment to the system and become a trusted validator. The system does not require a calculation-based competition, rather it randomly chooses from the validators. The possibility of being selected is directly proportional to the amount of cryptocurrency. The system will consume much less electricity and be much faster with POS (Sayeed & Marco-Gisbert, 2018; Opray, 2017).

Current business blockchain frameworks such as Hyperledger and R3 Corda are token-free platforms and are far more energy efficient as they eliminate this extravagant process. Other blockchain derivatives, such as Hashgraph, Holochain, and Tangle, are also energy efficient and resource friendly DLT systems.

3.3. Scalability and speed

Scalability is the ability to handle large volumes of transactions at high speeds. This basically depends on the following factors:

- Consensus: The nodes have to agree on the validity of the transaction. Adding information to a block with POW consensus protocol is a very slow process in the conventional cryptocurrency architectures. Creating a block can take around 10 to 60 minutes in Bitcoin (Bitinfocharts, 2019); it takes about 15 seconds in Ethereum (Etherscan, 2019). All new blocks are broadcasted and verified by all nodes in a typical blockchain network.
- Storage: Storage capacity is the biggest concern when implementing blockchain. The exponential growth of the block size creates a performance problem. Keeping the whole data in every node can be unfeasible and impractical in many solutions.

This brings out the scalability problem since the broadcast traffic and the size of the ledger data stored in the nodes increases exponentially because of the nature of the blockchain architecture. Moreover, lightweight devices like Internet of Things (loT) do not have sufficient resources for this. Many solutions have started to use all nodes for validation of the transactions, and only use some (full nodes) for storing all the data. Maintaining only the summary or link of the data in the nodes, keeping the data in the DSN architecture is also being implemented. Vitalik Buterin, co-founder of Ethereum, once claimed that a blockchain solution can have a maximum of two characteristics out of the three core characteristics (decentralization, security and scalability). This is also called the scalability/blockchain

trilemma, which is shown in Fig. 6. An attempt to solve the scalability problem will result in sacrificing on decentralization or security (Gomez, M., 2017).

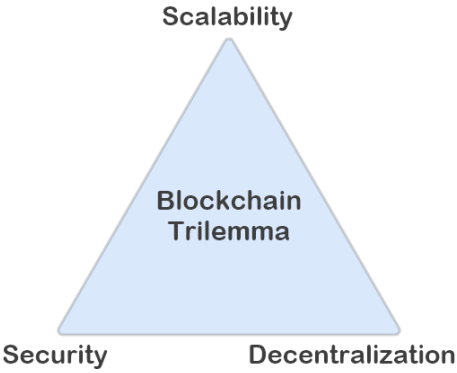


Figure 6: Scalability/Blockchain Trilemma

Scalability solutions can be covered in four layers; hardware, network, blockchain and application (EUBlockchain, 2019a). These solutions are summarized in Table 3. However, the throughput values are estimations to show the effect of each solution.

Using better machines or having faster communication has a limited effect on scalability. Hardware upgrades can be in limited amounts, and this type can perform best only in consortium chains and dPoS consensus protocol (EUBlockchain, 2019a). Higher bandwidths may be available from the telecom providers, but that does not mean that faster communication is possible.

There are various scalability solutions in blockchain layer such as adjusting block size, adjusting block interval, sharding, using different consensus mechanisms and decentralized derivatives. Solutions like directed acyclic graph (DAG), distributed hash table (DHT) can also be used. These solutions have different requirements and are difficult to compare with each other on performance (EUBlockchain, 2019a). These will be covered in detail in the next sections.

Table 3. Scalability measures in blockchain layers

Layer	Solution	Throughput	Limitations
Hardware	Using better machines	Up to 5-10x	Not for large networks Best only in consortium chains, dPoS
Network	Faster communication links	Up to 5x	Not affordable in all areas
Blockchain	Adjust block size Adjust block interval Sharding Different consensus mechanisms Decentralized derivatives	Up to 10-20x	Difficult to compare on performance
Application	Off-chain Sidechain	Up to 10,000 to 100,000x	Depends on interoperable tools

Significant parts of the data and computation can be transferred to conventional systems to make the processes faster. Structures like off-chain and side chain can be used to increase the throughput. Direct channels can be established between parties (EUBlockchain, 2019a). Solutions will be described in detail in the next sections.

3.4. Interoperability

Interoperability of blockchain infrastructures has emerged as a newborn challenge for the blockchain community in recent years. Although blockchain technology has been designed and established for removing the intermediaries and trusted third parties, users of different blockchain systems cannot easily transfer digital assets between each other without using an intermediary. For example, if a user wants to transact some data or a digital asset, secured and processed in Hyperledger Fabric network, to a R3 Corda network client, this user first has to register to the Hyperledger Fabric network, then decrypt the secured data, and then register on R3 Corda to use this network's functionality and put the aforementioned data into R3 Corda network. This creates a great amount of wasted time and processes. It becomes a necessity to ensure the interoperability of different blockchain architectures even between different companies or industries.

We will testify that different blockchain architectures will be able to communicate and share digital assets in the near future. Mechanisms like QuickX should be used to enable cross transactions. Sidechains have been proposed as a promising mechanism that allows transactions from one blockchain to another. It is not only a DLT technology but also a potential architecture for enabling the interoperability of the blockchain technologies (Ray, 2018).

3.5. Privacy

Privacy is another challenging issue that emerges from the nature of the blockchain methodology. In a permissionless blockchain architecture, all parties have the right to download the ledger, which implies that they have the right to explore the entire history of the recorded transactions. Implementing "the right of privacy" is a challenge in these architectures. Special care must be taken, when working with PII (Personally Identifiable Information). It is a good practice not to store PII on the blockchain and let the user handle his/her own data.

Zero Knowledge Proof (ZKP) can be integrated into blockchain systems to ensure privacy. The user can be given the total control of his/her data. ZKP can be used to validate any process (like identity check) without revealing any information about it (Goldreich, 2019; Korkmaz et al., 2019).

3.6. Cryptology challenges in the age of Quantum Computing

Quantum computing and the parallel processing power it promises threaten the security of the current public-key-based algorithms and blockchain systems. Quantum computing is an earthshaking technology that can be used to break ciphers and expose secrets that are secured by the current cryptographic algorithms (Piscini et al., 2018). Symmetric algorithms appear to be secure against quantum computers (and Grover's algorithms) by simply increasing the associated key sizes. Commonly used public-key cryptographic algorithms (based on integer factorization and discrete log problem) such as RSA, DSA, Diffie-Hellman Key Exchange, ECC, ECDSA will be vulnerable to Shors algorithm and will no longer be secure (Cromwell, 2015).

Researchers are studying post-quantum blockchain (PQB) and secure cryptocurrency schemes based on PQB systems, which can resist quantum computing attacks. This area is still under progress (Gao et al., 2019).

4. Decentralized Derivatives

There are Blockchain derivatives that intend to solve the problematic issues of this technology and offer individual solutions for specific aforementioned challenges (Schueffel, 2017). These derivatives are basically distributed ledger technologies that have different consensus protocols and architectures other than conventional blockchains. Directed acyclic graph (DAG) and distributed hash table (DHT) aim to perform the benefits of blockchain with better performance. Sidechain implementations offer to solve scalability solutions. Gossip protocol aims to reach a faster consensus than the counterparts do. These solutions are described, then the platforms that use these solutions are compared.

4.1. Directed Acyclic Graph

Changing the manner of the transaction validation process using distributed acyclic graphs is a new and effective approach, which creates new solutions for the scalability and speed problems of traditional blockchains. IOTA Tangle and Byteball are well-known examples, which put this methodology into practice (Wang et al., 2018).

The graphs are a representation of the connected peers through which information can be passed from one peer to another along different edges in a multidimensional space. They are great tools for traversing between various connections of individual units of data. A peer initially communicates with the closest peer according to pre-defined rules. They may be directed or undirected. Fig. 7 shows various graph types.

DAG is a non-looping graph that joins edges to turn in a pre-defined direction. Each square stands for a separate transaction in Fig. 7. The transactions are validated by the recently validated transactions in the way through the DAG branches.

DAGs stand out as promising DLT structure, enabling promising applications that can compete with classical blockchains (Jiab, Bouric, Guntad, & Roubaude). The use of DAG structures in distributed networks aims to solve the speed, cost, and scalability challenges of classical blockchain architecture.

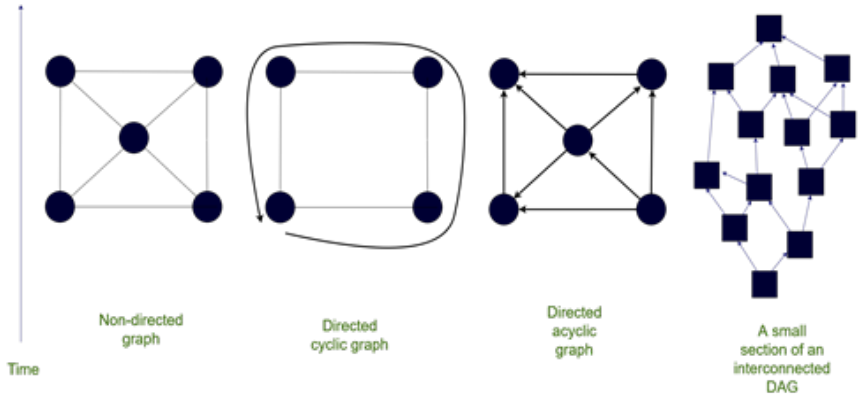


Figure 7: DAG and other graph types

4.2. Distributed Hash Table

DHT is a set of distributed storage systems that provides lookup and storage schemes for the peers, which store and retrieve data, identified by key values in the network. Distributed hash table establishes a distributed routing table in a very large and distributed network. There is no central authority, and peers can join and leave the network at any time in a distributed network. They are connected together through an overlay network. The nodes store and share the data by coordinating with each other (Dufel, 2017).

Fig. 8 shows the dictionary-like structure of the DHT usage. DHT allows the nodes to find any given key in the key-space. It maps the whole network by key values. Key value is the ID of the node that is calculated by hashing the node's IP and port combinations. This key identifies every separate node, and the node's position in the DHT indicates separate independent node which keeps related data. If a node leaves the network, the algorithm automatically shifts the abandoned key value to another peer, which is not addressed with any key. Nodes can make a search for the related node and find its data using this easy-to-implement structure. (Dufel, 2017).

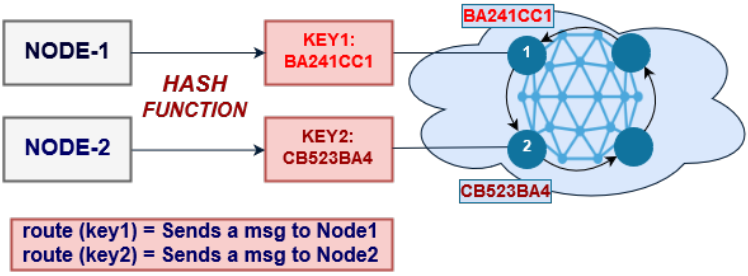


Figure 8: Mapping network using key values

4.3. Side Chain

Sidechain (child-chain) is a solution which allows making a partial copy and a separate branch of the (main/parent) blockchain, which is bound to counterpart(s). The original blockchain is called the mainchain and all additional blockchains are called the sidechains. Sidechains are used to allow cryptocurrencies and other digital assets to be processed in a separate private blockchain and then be securely transferred back to the original blockchain (Halpin & Piekarska, 2017).

Sidechain uses two-way pegging mechanisms to allow two separate chains bound to each other and transfer assets in between. In a crypto currency transfer scenario which is shown in Figure 9, a user on the parent chain initially sends its cryptocoins to an output address (Musungate et.al, 2019). The first step is a lock box, which locks the sent cryptocoins so the user cannot spend them. After acceptance of the transaction, an equivalent amount of crypto coins is delivered to the side chain. The user can spend the coins after that step. The reverse process is performed when moving back from a sidechain to the mainchain.

Every sidechain is responsible for its own security. Since each sidechain is independent, if it is hacked, the damage will be enclosed within that chain and will not affect the main chain.

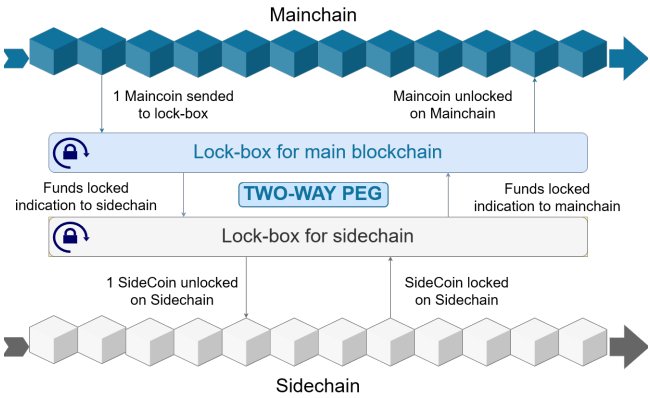


Figure 9: Transaction process of sidechain and two-way pegging

4.4. Gossip Consensus Protocol

Gossip is a communication protocol that is an agent for nodes to interact each other at high speed (Baird et al., 2018). When a node gossips, it randomly selects a peer and shares the received new information with it. The selected peer does the same thing, and this process continues until the information is passed on to all the connected nodes. It works just like social gossiping and information passes through the whole network in this way. The transactions are validated by the previous successful transactions rather than the mining process of currency based blockchain architectures. This protocol can run successfully on DAG and DHT networks to achieve high speed transactions (Zhenyu, Gaogang, Zhongcheng, Yunfei, & Xiaodong, 2018).

4.5. Comparison of the DLT Technologies

The DLT systems covered in this section are typically designed to deal with a registry of data that is distributed across a network. They are more transparent and robust compared with the conventional centralized database systems. The basic idea of blockchain derivatives is to form alternative decentralized systems that can solve the structural challenges and overcome the architectural limitations of traditional blockchains. Table 4 shows the comparison of the DLT derivatives.

Even though there are many similarities among the derivatives of DLT systems, there are also some architectural differences (El Ioini & Pahl, 2018). The foremost blockchain designs were created to be permissionless, but the DLT derivatives are predominantly permissioned. Everyone can join a permissionless network, however only accepted parties may access the network in the permissioned DLT alternatives. This difference also influences the size of the

network. Blockchain networks, which are used for well-known cryptocurrencies such as Bitcoin, aim to expand to provide a more secure environment. In a permissioned DLT network, the number of parties involved tend to be smaller, as this number does not have such an effect on the security of the systemin this case.

Projects such as Hashgraph, Holochain and Tangle are promising platforms, which are creating new types of distributed ledger technologies. These projects share the common aspects of distributed, consensus, flexible, and peer-to-peer platforms. Although being fast DLT architectures, they use their own consensus protocols and the data structures. Hashgraph and Tangle solve the scalability problem with DAG.

Hashgraph and Holochain use the gossip consensus protocol. Hashgraph is a patented permissioned DLT network, which can handle over 250,000 TPS. The validation of the network requires at least ⅔ of the nodes to receive gossip and confirm transactions. Holochain uses distributed hash table (Anwar, 2018). It aims to create a new distributed Internet structure, trying to establish a new secure generation of cloud computing framework. Trust is established using the computing power of the peers. It is estimated to have an immense scalability rate depending on the holochain networks expansion. Each user peer keeps its own data and transactions information (Harris-Braun, Luck.,& Brock, 2018).

Bitcoin processes 3–7 transactions per second (TPS) and Ethereum can handle 10–20 TPS, however Hashgraph promises to process hundreds of thousands of TPS (Kerner, 2018). Tangle's consensus mechanism hashcash has a high theoretical limit on the TPS throughputs (IOTA, 2019). Tangle can reach up to 800 TPS rates (Kerner, 2018). Despite their potential advantages, their capabilities have not been tested as traditional blockchain systems.

Sidechain has a very high potential to enhance scalability and TPS values depend on the platform used. It can also be used to provide interoperability between different blockchains. There are several platforms which are testing sidechains. A promising project is Plasma (Saini, 2018). Plasma is the child-chain solution of Ethereum. High TPS values are aimed at by allowing each Dapp to use its own chain (Poon, & Buterin, 2017).

Table 4. Comparison of the DLT Technologies

	Blockchain	**Plasma**	***Tangle***	***Hashgraph***	***Holochain***
Structure	P2P	P2P	Directed acyclic graph	Directed acyclic graph	Distributed hash table
Platform	Bitcoin	Sidechain	IOTA	Hedra Swirlds	Holo
Transaction per second (Tps)	4 to 7	More than billions	500 to 800	More than 200.000	More than millions
Consensus	PoW	POS	PoW: hashcash	Virtual voting	DNS Validation Rules
Decentralized	Yes but using mining pools make it semi decentralized	Depends on the implementation	Semi-Centralized	Semi-Centralized	Decentralized
Licence	Open Source	Open Source	Open Source	Patented	Open Source
Maturity	Proven and been used since 2008	Experimental	Experimental	Experimental (Public use since 2018)	Experimental (Alpha1 rel. in 2018)

5. Data Storage in Decentralized Systems

Blockchain is not a place to store all kinds of different data. As mentioned above, it is a registry where the records (logs) of the transactions are kept. A transaction can be a record of any process and may also contain codes, which allows the autonomous working of a system. A transaction can also give a link to the cloud storage where the actual data exits. Data may be stored in different forms. There are evolving data solutions to solve scalability and interoperability problems. Selected solution proposals are covered first. Then the hybrid solutions, which are formed by using different solutions together, are covered. This section will continue with the comparison of decentralized solutions with centralized databases.

Table 5. Comparison of Blockchain with Evolving Data Solutions

Category	Solution	Throughput	Cost Power& Resource	Capacity Block width/ size	Advantage	Disadvantage
Basic Blockchain (Bitcoin)	PoW	Low	High	Basic	Proved to work in trustless environment, Protection against DDoS attacks	Scalability, Computationally Expensive, Needs high computational power, High energy and processing costs, %51 Attack
Alternate Consensus Protocols	POS PoA	High	High	Low	APX 0 runtime cost, High transaction speeds	The node who has the steak controls the network
	Raft-based consensus	High	Low	Low	Handle multiple problems, Easy to implement	Lacking enough live tests
	Gossip	High	Low	Low	APX 0 transaction fee and waiting time	Lacking enough live tests
On-chain	Big block	High	Low	High	High capacity transmissions	Centralization of mining pools, High Orphan block rate
	Segwit	High	Low	-	Various Possible Bitcoin solutions	Fungibility occurrence
	Sharding	High	-	Low	Low capacity burden Parallel processing	%1 Attack
Off-chain	Lightning network	High	Low	Low	APX 0 transaction fee and waiting time	P2P Payment channels
	Raiden network	High	Low	Low	General purpose channel	P2P Payment channels
Child-chain	Plasma	High	-	Low	Parent-child blockchain tree	High costs of verification
Inter-chain	Side Chain	High	Low	Low	Blockchain interoperability and cross transactions	Application boundaries
DSN	IPFS Gaia Storj	High	Low	High	More secure Flexible Reduced rate of data failures and outages	-
Decentralized Derivatives New Solutions	DAG (Nano, IOTA & Byteball)	High	Low	Low	Better scalability No miners Quantum resistant cryptography	Lacking enough live tests
	DHT	High	Low	Low	APX 0 runtime cost, Very high transaction speeds	Lacking enough live tests

5.1. Evolving Data Solutions

Data storage solutions are evolved to solve the scalability, interoperability, and privacy of blockchain in several forms and are given in Table 5. The solutions are shown as follows (Kim et.al, 2018):

- Using alternative consensus protocols
- On-chain: Storing all the data on the main-chain. Solutions such as sharding, making blocks bigger are possible.
- Off-chain: Off-chain is storing the data outside the blockchain, processing it and writing the summary on the blockchain. There are challenges to reach the manipulation resistance, verifiability and privacy (Eberhardt & Tai, 2017), (Lightning network, Raiden network).
- Child & parent chains: The records of the child-chain are processed and written to the parent-chain in this type of implementation (Plasma).
- Inter-chain: This is used to provide communication and join the functionality of different blockchains. Structures such as Atomic swaps and side-chain are used.
- DSN: Using blockchain and cloud storage together.
- Other structures (DAG, DHT … etc)

Enterprise solutions have different needs and expectations than the cryptocurrency systems. Different consensus protocols (PoA, POS, raft-based consensus, Istanbul BFT, etc.) are being implemented to make the consensus phase faster and reach higher transactions per second (TPS) rates by eliminating the mining processes, while ensuring confidentiality. By way of example, Quorum, which is based on Ethereum, does not use POW/POS consensus protocols, but instead supports multiple consensus protocols to support enterprise needs. It supports alternative consensus protocols like raft-based consensus, Istanbul BFT (IBFT) (Baliga et.al. 2018).

Sharding and making blocks bigger are both possible on-chain solutions. The Big block is the basic process to enhance the block size. Making blocks bigger enlarges the transmission limit, but big blocks need extremely high processing powers, which will also increase the transmission cost (Clifford, 2017). Since the propagation speed becomes limited, this process increases the probability of orphan blocks appearing. Big blocks are not efficient at the current stage because of these disadvantages. Sharding is the process of dividing a database

into smaller segments. It is also called horizontal partitioning. Sharding is a controversial issue in blockchain and there are different views. Vitalik Buterin and Beniamin Mincu believe that sharding can be one of the solutions. Vitalik Buterin once expressed the concept of the 'sharding' model as the creation of hundreds of different universes, each of which being different account spaces. According to him, the transaction will affect the things only in the universe it belongs to. He claims that thousands of transactions per second can be achieved without any special server, nor with consortium chains (Gomez, M., 2017). Beniamin Mincu, the CEO of the Elrond Network, claims that sharding is needed to reach the throughput capacity that is needed to rival networks like VISA and states that some challenges are single-shard takeovers, cross-shard communication and data validity (Cointelegraph, 2019).

Lightning network and raiden network can be given as examples of off-chain solutions. A consensus process will not be used in lightning network when two parts trust each other. Transactions will be quicker and will not be recorded on the chain (Karaarslan & Adigüzel, 2018; Poon & Dryja, 2016).

Plasma can be given as an example of the child-chain solution in Ethereum. Each Dapp will use its own chain in the Plasma solution (Poon, & Buterin, 2017).

Atomic swaps and sidechain are inter-chain solutions that are established to enable cross transactions and blockchain interoperability. Atomic Swap is the peer-to-peer currency exchange between different blockchain networks, without the need for a mediator. Sidechain was covered in Section 4. It has a very high potential of enhancing scalability and can also be used to provide interoperability between two separate blockchains (Musungate et.al, 2019).

Using blockchain and cloud storage together forms decentralized storage, which is also called a decentralized cloud storage network (DSN). Data can be stored and shared without having to trust third parties (Wilkinson et.al, 2014). This solution is used to overcome storage limits and also provide personal data storage and privacy. A DSN network can have advanced privacy, security and data control as it has the following characteristics (Karaarslan & Adiguzel, 2018):

- More secure as it uses client-side encryption,
- Flexible as there are speed and low-cost advantages with proper implementation,
- Integrity and availability of the data is ensured with proof of retrievability,
- Reduced rate of data failures and outages.

Examples of DSN can be given as follows (Karaarslan & Adiguzel, 2018):

- Gaia: It is used by Blockstack. When the user uses the decentralized application and any data is needed to be written, it serves to save this data on the existing cloud infrastructure. Data is written in encrypted or signed form (Ali et.al, 2017).
- Storj: Storj works as a P2P cloud storage network. An open source software project called Metadisk provides a set of tools to make Storj easily integrated with legacy systems (Wilkinson et.al, 2014).

Other token-free DLT derivatives, such as Hashgraph, Holochain, and Tangle achieve better scalability and TPS rates by using different structures (DAG, DHT) while eliminating mining operations.

Different multi platform solutions are possible. Outstanding ones are shown as follows:

- Using hybrid blockchain solutions which involve public and private blockchain solutions working together,
- Using inter-chain structures like sidechain to make different decentralized solutions working together.
- Using blockchain with decentralized cloud storage network (DSN)
- Alternative cloud storage platforms which use blockchain as an awarding system.

5.2. Comparison of Decentralized Solutions with Centralized Databases

The differences between decentralized solutions and databases is in their design and purpose. This topic is widely investigated in (Tabora V., 2018). Blockchains are distributed systems, which hold replicated databases on several different nodes. Special consensus protocols are used to ensure these replicas are trusted (Murthy C., 2016).

Database systems are becoming more complex with the ever-increasing usage of different data types, big data, and cloud infrastructure. There are many characteristics to classify them. Firstly, it is important to talk about the data management models like relational and non-relational. Relational databases are the most commonly used database types in the world. However, non-relational databases are also becoming popular with the rising storage needs of unstructured data and the increasing usage of the machine learning processes which use them. By way of example, No-SQL databases are also becoming widespread and are mostly used for rapid development or used to store large amounts of data that have little or no structure. Blockchain is a non-relational database but there are also exceptions. A recent

blockchain system called postchain (Botsford A., 2019) seems to be the first blockchain system which uses the relational model.

The general characteristics of decentralized solutions with relational databases are given in Table 6. There are also variations but it is outside the scope of this section. The comparison of these solutions is shown in Table 6 and is summarized in the following paragraphs.

Database is deployed in client/server model, however blockchain system is decentralized. There is mostly one party involved in relational database. Consistency is hard and expensive to achieve in relational database when there is more than one party. Blockchain solutions are best suited for multi-party solutions and satisfy consistency as all nodes have the full copy of the dataset. A blockchain system will directly identify and correct possible inaccurate records. Companies, authorities, banks, transportation companies and such like can be a part of this multi-party network (Schlapkohl, 2019).

Security services such as availability, integrity, and fault tolerance are highly supported with blockchain systems. Database systems may be deployed to serve these services, but we can say that it will not be as effective as blockchain systems. Users trust databases that they will work right, but no one can be sure since administrators have full control of the system. Even competitor companies need to share data between each other. They do not need to trust each other, but need to trust the shared data. Trusted third parties can also be used to ensure trust but their trustworthiness is also questionable (Karaarslan & Adiguzel, 2018). Blockchain systems work by ensuring trust without using any intermediaries. Trust is established using autonomous code and consensus protocols.

The attackers try to delete all possible evidence on the compromised system after any attack. Digital forensics become difficult when logs are deleted. Any change attempt on the blockchain ledger is also kept on the immutable ledger, so the details of the incident (who, when and what) will be detected. This will also have a deterrent effect on attackers.

Cryptocurrencies use public chains that have transaction records transparent to everyone. They allow everyone to see and query all transaction records on the system. Enterprise solutions use private or hybrid chains and queries that give reading access only in that domain. Databases do not give such a service.

Data management is relational in databases, blockchain is non-relational. The user accounts are created on the database system and administered. Security is mostly implemented by giving roles on the database system such as the database tables they can reach and their

permissions. However, blockchain works autonomously; consensus protocols and smart codes (autonomous codes) define how the system works. There are no users on the system. Decentralized identity management systems (IDMS) (EUBlockchain, 2019b) or such like may be used, however these do not define any user roles on the system. Permissioned blockchains are also possible where there is an access control layer. This layer is used to permit specified actions to the defined users. This property is different from the relational database permission process.

Blockchain is distributed by default. Database systems are installed as standalone by default, but may also be deployed asdistributed. However, the amount of nodes that the blockchain solutions can reach is mostly not possible in distributed database solutions. Only allowed nodes can be added to the distributed architecture of relational databases. Nodes can be permissioned or permissionless depending on which decentralized technology is used. Redundancy is only possible to the level where relational database is distributed. All full nodes have the latest copy and data redundancy is satisfied in the blockchain implementations.

Sharding is available in relational databases when the data is distributed in several servers. Sharding is a controversial issue in blockchain.

Parallelization is limited in relational databases. Cloud adaptability is high with decentralized databases. Big data handling capability is limited in relational databases; however, decentralized solutions are more suitable for big data operations, especially when used along with the cloud infrastructure. Relational databases generally handle small data better. However, decentralized solutions handle big data better. Scalability for variable data sizes is rigid in relational databases, but elastic in decentralized solutions (Demir et.al, 2018).

Databases support high volume transactions at a fast processing rate. Blockchain implementations have to validate transactions and this comes at the cost of speed. The solutions which use PoW consensus protocols support low volume transactions at a slow processing rate. Higher volume transactions and faster processing rates are possible when different consensus protocols like PoS, PoA are used. These faster consensus protocols mostly need trusted nodes. We can say that blockchain should not be used when transaction speed is a concern. However, there are studies on low-latency solutions. Data analytics is supported with databases, however blockchain can be described as poorly supported in this concept.

Blockchain systems are said to have problems in the areas of data size, synchronization, energy consumption, interoperability and scalability. There are many studies and many new

solution proposals on these areas. Some of these proposals are given in the previous sections of this chapter. Databases are widely used in various projects. However blockchain solutions have a value when there is a need for establishing trust between parties without any intermediaries involved and a need for data verification.

Table 6. Comparison of Blockchain (and derivatives) with Relational Database

	Relational Database	**Blockchain (and Derivatives)**
Centralization	Centralized	Decentralized
Party Involved	Mostly one	More than one party
Consistency (multiple party)	Hard and expensive to achieve	Consistent (full copy)
Security Services (Availability, Integrity, Fault Tolerance)	Poorly supported (by default)	Highly supported
Trust	Trusted 3rd party	Trust without intermediary Trust on smart code, consensus
Forensics	Difficult (if logs are deleted)	Easier (unalterable records)
Transparency of transaction data	No	Yes (public chains) Partial (private or federated chains)
Data management system	Relational model	Non-relational
Management Method	Administrated	Autonomous
User Control Method	Permissioned	Permissionless, permissioned
Distributed Deployment	Possible	Distributed by default
Node Add Method	Permissioned	Permissionless, permissioned
Redundancy	Possible (when distributed)	All full nodes have the latest copy
Sharding	Suitable (when distributed)	Controversial
Parallelization	Limited	Suitable
Cloud Adaptability	Limited	High
Big data handling	Limited	Suitable (with cloud)
Scalability for variable data sizes	Rigid	Elastic
Read/write Speeds	Faster for small data	Faster for big data
Transaction Volume	High volume	Low to Average*
Transaction Speed	Fast	Slow to Average*
Data Analytics	Supported	Limited
Problems	Single Point of Failure Administration Issues Security Issues	Energy Consumption Interoperability Scalability
Best When	High volume of data Fast processing need Quick query need	Data verification needed Establishing Trust without intermediaries

** Changes according to the level of decentralization*

6. Proposed Model: MPISA

Many decentralized computation and storage solutions use different technologies and are used in different domains. Just like connecting different communication networks to form the Internet, different solutions can be inter-connected and their services can be associated. Different decentralized solutions generally use different platforms that are suitable for that domain. By way of example, a scenario may require co-working of a PoW-based cryptocurrency and a PoA-based supply chain solution. Hence, there is a need for a unifying platform that will solve the interoperability problem.

We propose a model called MPISA, whose name is a portmanteau of "Multi-Platform Interoperable Scalable Architecture". We aim to show how multiple platforms can be used together and help developers in solving scalability and interoperability issues. The MPISA model is shown in a two blockchain platform scenario in Figure 10. In this scenario, the two blockchain platforms have their own P2P network and a mainchain as the main blockchain. Each platform uses its own sidechain structure for the scalability issues.

Common data such as digital identities or general preferences can be kept in the shared data storage. Such a system will help in preventing unnecessary re-entrance of such data in different parties and also in preventing inconsistencies. Any change of this data will require only one update and will be available to all parties instantaneously. These will decrease the maintenance costs of this data across systems (Houlding, 2019). The data can be kept in a cloud or distributed storage. It can be reached through a decentralized identity management system. Such a platform can be designed to keep user credentials safely. The Dapps will be able to check the user identity through this system. Zero-knowledge proof can also be integrated into this system to ensure the privacy of the parties. Users can keep their credentials on this shared platform without revealing their private data.

The Dapps have relevant APIs to grant access to their associated blockchain platform. Smart contracts are used in the data storage operations. The blockchain only keeps the records of the transactions made, but the associated data is not kept in the ledger. A cloud or distributed storage is used for storing and retrieving data. Data can be reached using the data locations in the ledger records.

The most challenging component of the model is the interoperability platform. The Dapps will be able to reach different blockchains and their associated data using this platform. Sidechains or atomic swaps can be used to enable interoperability between the chains.

However, sidechain solution proposals are mostly proof of concept and experimental (Johnson et.al, 2019). This area and the scalability issues are still open for development.

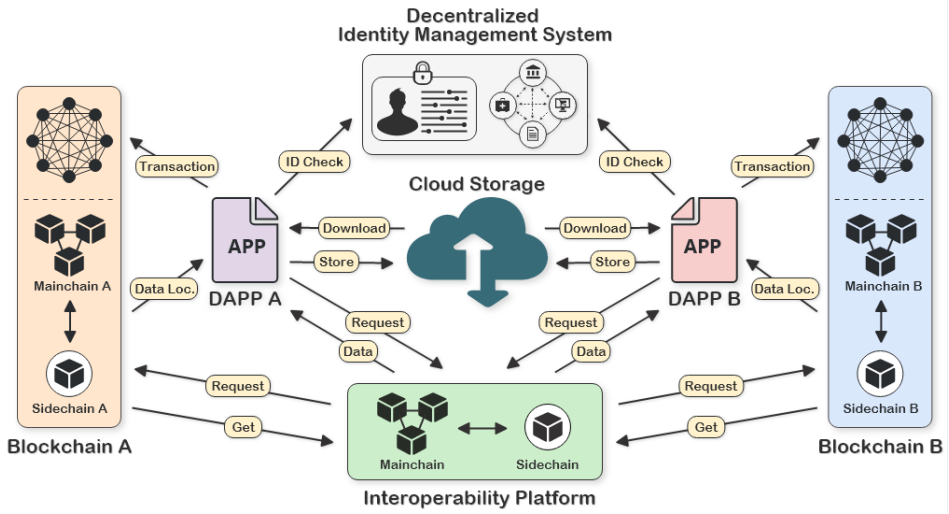


Figure 10: Multi-platform interoperable scalable system (MPISA) scenario

7. Conclusion

This study aims to describe the decentralized ledger technology and its usage as a data storage to the data scientists and to give a contribution to academia by making the concept easier to understand. Scalability measures in blockchain layers are given in Table 4. Blockchain technology is compared with evolving data solutions in Table 5 and is compared with the relational database in Table 6. We believe that if the data scientists could understand this technology better, they would be able to be a part of the work to solve the challenging issues that come with it.

Blockchain is not a place to store all kinds of different data; it is a registry where the records of transactions are stored. Blockchain is currently the most effective secure way of keeping these records as a ledger, which are distributed in a network. It will help in sharing the data between different parties and enable collaboration. These DLT based solutions ensure the trust without intermediaries. Smart contracts allow autonomous working of the system.

Decentralized systems can be designed to provide common data such as digital identities or general preferences. Such a system will reduce the time needed for the data synchronization across parties and decrease the maintenance costs of this data. We recommend keeping the data in a cloud or distributed storage. Data will be reachable using the data locations in the ledger records.

Nowadays more decentralized application prototypes have become the new focus, and we are now talking about projects which have started to move towards production alongside legacy systems(Brennan et al, 2018). Finance and supply chain are some of the widely used fields of blockchain technology. There are many fields such as health data exchange, know your customer (KYC), smart governance and fraud detection that fit perfectly with the benefits of blockchain technology. We will see blockchain based open global trade digitization platforms in the future, which will ensure secure and instant access to end-to-end supply chain information (Mohan, 2019).

Despite the obvious opportunities of blockchain technology, the challenges of blockchain are still in need of discussion. The most notable architectural challenges are scalability and privacy problems. Other challenges include energy consumption, interoperability, cryptology challenges in the age of quantum computing. These problems should be solved to achieve better implementations in the field. In particular, we should work on scalability problems. Possible solutions should not have an effect on security and decentralization. However, we also believe that some enterprise solutions may also have some centralized parts.

There are evolving data solutions for the decentralized storage challenges. Several solutions such as making blocks bigger, sharding, using more than one chain, and using a decentralized cloud storage network have been proposed. Using solutions like directed acyclic graph (DAG), distributed hash table (DHT) also seem promising.

DAG, DHT, sidechain, gossip protocol and such like can be used to solve the scalability problems of blockchain. Platforms such as Tangle, Hashgraph and Holochain which use these solutions are compared. We think that these derivatives are important for the evolution of decentralized systems. The decentralized solutions promise better TPS rates than the traditional blockchain systems and are likely be preferred in the near future if no security flaws are noticed in their implementations. However, their capabilities have not been tested much and their sustainability has not been tested as long as the known blockchain technologies.

Measures for the privacy of data should be taken. It is a good practice not to store PII on the blockchain and let the user handle his/her own data. Zero Knowledge Proof (ZKP) can be integrated to blockchain systems to ensure privacy.

We proposed a multi-platform interoperable scalable architecture (MPISA) model. We plan to study scalability and interoperability technologies, which can be used to make such a system possible.

In the near future, interoperability will be one of the most important necessities of the business blockchain platforms for benefiting inter-sectoral business solutions with the wide usage of DLT. Sidechain is the potential structure for enhancing the scalability of existing blockchain implementations and a chance for ensuring the interoperability of blockchain technologies. However, it adds more complexity and should be well designed and implemented.

Blockchain immutability may also be argued. Controlled rewriting of blockchain records with chameleon-hashing may be applicable in some cases (Derler et.al, 2019). Different approaches may be appropriate for different implementation areas. Some domains such as the Internet of Things have domain specific characteristics such as frequent data transfers with small content. Domain specific solutions should be developed. IOTA, based on Tangle, is a candidate for a solution; however, it still has many issues which need to be solved.

Blockchain and artificial intelligence (AI) can be used together to complement each other. Revolutionary improvements are possible (Dinh & Thai, 2018; Salah et.al., 2019).

Supercomputers and quantum technology will be further key elements that will shape future implementations of blockchain. Post-quantum blockchain and secure cryptocurrency schemes, which can resist quantum computing attacks, should be studied.

IEEE, ISO and W3C are working on new standards. We need more standardization efforts on blockchain and decentralized systems. We would like to emphasize that blockchain and its derivatives are still evolving. New advanced approaches and better benchmark systems (Gutierrez C., 2019) are being developed. The promises of decentralized implementations are so evident that the challenges should be studied, and more attention should be given to this field.

Acknowledgements

We would like to thank MSKU Blockchain Research Group (http://wiki.netseclab.mu.edu.tr/index.php?title=MSKU_BcRG) members (especially Cemal Dak, Şafak Öksüzer, Ahmet Önder Gür) for their contribution to the graphics used in this chapter.

References

Ali, M., Shea, R., Nelson, J., & Freedman, M. J. (2017). Blockstack: A new decentralized internet. Whitepaper, May.

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., & Muralidharan, S. (2018). Hyperledger fabric: a distributed operating system for permissioned blockchains. In Proceedings of the Thirteenth EuroSys Conference (p. 30). ACM.

Alves, G., Cavalcante, E., & Batista, T. (2018). On the Use of New Blockchain-based Technologies for Securely Distributing Data. 81 - 88.

Botsford A. (2019). What is relational blockchain and why should you use it?. Retrieved from https://blog.chromia.com/what-is-relational-blockchain-and-why-should-you-use-it/

Anwar H. (2018). Blockchain vs Hashgraph vs DAG vs Holochain. Retrieved from http://www.101blockchains.com.

Baird, L., Harmon, M., & Madsen, P. (2018). Hedera: A governing council & public hashgraph network. The trust layer of the internet, whitepaper, 1.

Barnas, N. B. (2016). Blockchains in national defense: Trustworthy systems in a trustless world. Blue Horizons Fellowship, Air University, Maxwell Air Force Base, Alabama.

Bitinfocharts (2019). Bitcoin Block Time historical chart. Retrieved from https://bitinfocharts.com/comparison/bitcoin-confirmationtime.html

Bozic, N., Pujolle, G.,& Secci, S. (2016). A Tutorial on Blockchain and Applications to Secure Network Control-Planes. IEEE 3rd Smart Cloud Networks & Systems (SCNS), pp. 1-8.

Harris-Braun, E., Luck, N., & Brock, A. (2018). Holochain-scalable agentcentric distributed computing. Alpha, 1, 1-14.

Baliga, A., Subhod, I., Kamat, P., & Chatterjee, S. (2018). Performance evaluation of the quorum blockchain platform. arXiv preprint arXiv:1809.03421.

Brennan, C. , Zelnick, B. , Yates, M. & Lunn, W. (2018). Cryptocurrencies are only the beginning, Credit Suisse Blockchain Revolution Series

Clifford, J. (2017). Understanding the block size debate. Retrieved from https://medium.com/scalar-capital/understanding-the-block-size-debate-351bdbaaa38

Cointelegraph (2019). Sharding explained. Retrieved from https://cointelegraph.com/explained/sharding-explained

Conti, M., Kumar, E.S., Lal,C., & Ruj, S. (2018). A Survey On Security and Privacy Issues of Bitcoin. IEEE Communications Surveys & Tutorials, 20(4), pp.3416-3452.

Cromwell, B. (2015). What Is Post-Quantum Cryptography And What Does It Mean For Us?. Retrieved from https://blog.learningtree.com.

Demir E., Senocak T., Gezer N., Çabuk U. C. (2018). A Preliminary Study on Suitable Database Types for E-Voting Systems, (ICENS) 2018, vol.4, pp.288

Derler, D., Samelin, K., Slamanig, D., & Striecks, C. (2019). Fine-Grained and Controlled Rewriting in Blockchains: Chameleon-Hashing Gone Attribute-Based. IACR Cryptology ePrint Archive, 2019, 406.

Dinh, T. N., & Thai, M. T. (2018). Ai and blockchain: A disruptive integration. Computer, 51(9), 48-53.

Dufel M. (2017, 27 Dec). Distributed Hash Tables And Why They Are Better Than Blockchain For Exchanging Health Records. Retrieved from https://medium.com/@michael.dufel_10220/distributed-hash-tables-and-why-they-are-better-than-blockchain-for-exchanging-health-records-d469534cc2a5

Eberhardt, J., Tai, S. (2017) On or off the blockchain? Insights on off-chaining computation and data. European Conference on Service-Oriented and Cloud Computing. Springer, Cham.

El Ioini, N., & Pahl, C. (2018). A review of distributed ledger technologies. Springer, Cham, In OTM Confederated International Conferences, On the Move to Meaningful Internet Systems, (pp. 277-288).

Etherscan (2019). Ethereum Block Time History , Retrieved from http://etherscan.io/chart/blocktime

EUBlockchain (2019). Scalability, Interoperability and Sustainability of Blockchains. [Report]. Retrieved from https://www.eublockchainforum.eu/sites/default/files/reports/report_scalaibility_06_03_2019.pdf

EUBlockchain (2019) Blockchain and digital identity. [Report]. Retrieved from https://www.eublockchainforum.eu/sites/default/files/report_identity_v0.9.4.pdf

Flipo, F.,& Berne, M. (2019) The Bitcoin and Blockchain: Energy Hogs. Retrieved from https://theconversation.com

Gao, Y. L., Chen, X. B., Chen, Y. L., Sun, Y., Niu, X. X., & Yang, Y. X. (2018). A secure cryptocurrency scheme based on post-quantum blockchain. IEEE Access, 6, 27205-27213.

Goldreich, O. (Ed.). (2019). Providing Sound Foundations for Cryptography: On the work of Shafi Goldwasser and Silvio Micali. Morgan & Claypool.

Gomez, M. (2017). Ethereum Co-Founder Vitalik Buterin Weighs in on Blockchain Improvement & Scaling Issues. Cryptovest. Retrieved from https://cryptovest.com/news/ethereum-co-founder-vitalik-buterin-weighs-in-on-blockchain-improvement--scaling-issues/

Gutierrez C. (2019). Hyperledger Caliper to Provide Benchmarking for Blockchain Systems. Retrieved from https://www.altoros.com/blog/hyperledger-caliper-to-provide-benchmarking-for-blockchain-systems/

Gür Ö., Öksüzer Ş.,& Karaarslan E. (2019). Blockchain Based Metering and Billing System, ICSG 2019. [Accepted to be indexed] IEEE Explore.

Halpin H., & Piekarska M.(2017, July 3). Introduction to Security and Privacy on the Blockchain, 2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). http://dx.doi.org/ 10.1109/EuroSPW.2017.43

Harris-Braun, E., Luck, N., & Brock (2018). A. Holochain: Scalable Agent-Centric Distributed Computing.

Houlding D. (2019). A Data Centric View of Blockchain. Retrieved from https://www.linkedin.com/pulse/data-centric-view-blockchain-david-houlding-cissp-cipp/

Hyperledger (2018). Five Hyperledger Blockchain Projects Now in Production. Retrieved from https://www.hyperledger.org/blog/2018/11/30/six-hyperledger-blockchain-projects-now-in-production

IEEE (2019). IEEE Blockchain Standards. Retrieved from https://blockchain.ieee.org/standards

IOTA (2019). The Tangle. Retrieved from https://docs.iota.org/docs/getting-started/0.1/network/the-tangle

ISO (2019). ISO/TC 307 technical committee on blockchain and distributed ledger technologies. Retrieved from https://www.iso.org/committee/6266604.html

Jiab, Q., Bouric, E., Guntad, R., & Roubaude, D. (2018, Nov). Network causality structures among Bitcoin and other financial assets: A directed acyclic graph approach. The Quarterly Review of Economics and Finance (70), 203-213.

Johnson, S., Robinson, P., & Brainard, J. (2019). Sidechains and interoperability. arXiv preprint arXiv:1903.04077.

Karaarslan, E., & Adiguzel, E. (2018). Blockchain Based DNS and PKI Solutions. IEEE Communications Standards Magazine 2.3 (2018): 52-57.

Karaarslan E., & Akbaş, M.F. (2016). Blok Zinciri Tabanlı Siber Güvenlik Sistemleri [Blockchain Based Cyber Security Systems]. Uluslararası Bilgi Güvenliği Mühendisliği Dergisi, 3(2), 16 - 21, http://dx.doi.org/10.18640/ubgmd.373297.

Kerner, L. (2018, Mar 25). Is The Future Of Blockchains DAGs ? - 5 Takeaways From The Hashgraph Event In NYC on March 13th [Web log post]. Retrieved from https://medium.com/crypto-oracle/is-the-future-of-blockchains-dags-5-lessons-from-the-hashgraph-event-in-nyc-on-march-13th-ff0f7e0fa510

Kim, S., Kwon, Y., & Cho, S. (2018). A survey of scalability solutions on blockchain. In 2018 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1204-1207). IEEE.

Korkmaz, U., Altunlu, H. İ., Özkan, A., & Karaarslan, E. Sustainable Member Motivation System Proposal for NGOs: NGO-TR. UBMYK 2019, 2019

Mohan C. (2019). State of Permissionless and Permissioned Blockchains: Myths and Reality, BlueTalks @ Rio BNDES

Murthy, C. (2016). Blockchain DB-unked, Presentation Slides, Retrieved from https://ripple.com/files/db-unked.pdf

Musungate, B. N., Candan, B., Çabuk, U. C., & Dalkılıç, G. (2019). Sidechains: Highlights and Challenges. ASYU 2019

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Nasir, Q., Qasse, I.A., Talib M.A., & Nassif A.B. (2018). Performance analysis of hyperledger fabric platforms. Security and Communication Networks, Volume 2018, Article ID 3976093.

Natoli, C., & Gramoli, V. (2016). The blockchain anomaly. In 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA). 310-317. IEEE.

Opray, M. (2017). Could a blockchain-based electricity network change the energy market. The Guardian. July, 13.

Ölmez A.C., Karaarslan, E. (2019). Blockchain Based Adoption and Fostering System Proposal for Animal Shelters: BAdopt. UBMYK 2019

Piscini, E., Dalton, D., & Kehoe, L. (2017). Deliotte, Blockchain & Cyber Security.

Poon J., & Buterin V., (2017, August 11). Plasma: Scalable Autonomous Smart Contracts, [Working draft]. Retrieved from https://plasma.io/

Poon, J., & Dryja, T. (2016). The bitcoin lightning network: Scalable off-chain instant payments.

Ray, S. (2018, Jan 22). What are Sidechains? [Web log post]. Retrieved from https://hackernoon.com/what-are-sidechains-1c45ea2daf3

Rosic, A. (2017). Proof of work vs proof of stake: Basic mining guide. Blockgeeks blog.

Saini V. (2018, April 26). Retrieved from https://hackernoon.com/13-sidechain-projects-every-blockchain-developer-should-know-about-804b65364107

Salah, K., Rehman, M. H. U., Nizamuddin, N., & Al-Fuqaha, A. (2019). Blockchain for AI: review and open research challenges. IEEE Access, 7, 10127-10149.

Sayeed, S., & Marco-Gisbert, H. (2018). On the Effectiveness of Blockchain against Cryptocurrency Attacks. Proceedings of the UBICOMM.

Schlapkohl K., (2019). What's the difference between a blockchain and a database?. Retrieved from https://www.ibm.com/blogs/blockchain/2019/01/whats-the-difference-between-a-blockchain-and-a-database/

Schueffel, P. (2017). Alternative Distributed Ledger Technologies Blockchain vs. Tangle vs. Hashgraph - A High-Level Overview and Comparison (December 15, 2017). http://dx.doi.org/10.2139/ssrn.3144241.

Tabora V. (2018). Databases and Blockchains, The Difference Is In Their Purpose and Design. Retrieved from https://hackernoon.com/databases-and-blockchains-the-difference-is-in-their-purpose-and-design-56ba6335778b

Trautman, L. J., & Molesky, M. J. (2019). A Primer for Blockchain. University of Missouri-Kansas City Law Review, Forthcoming.

Wang, W., Hoang, D.T., Xiong, Z., Niyoto, N., Wang, P., Hu, P., &Wen, Y. (May 7, 2018) A Survey on Consensus Mechanisms and Mining Management in Blockchain Networks. arXiv preprint arXiv:1805.02707v1.

W3C (2019). The Web Ledger Protocol 1.0, Draft Community Group Report 18 June 2019. Retrieved from https://w3c.github.io/web-ledger/

White,M.,Killmeyer,J.,&Chew,B. (2017). Understanding basics of blockchain in government. Retrieved from https://dupress.deloitte.com.

Wilkinson, S., Boshevski, T., Brandoff, J., & Buterin, V. (2014). Storj a peer-to-peer cloud storage network.

Wüst, K.,& Gervais, A. Do you need a Blockchain. Retrieved from https://eprint.iacr.org

Valenta, M., & Sandner, P. (2017). Comparison of Ethereum Hyperledger Fabric and Corda, Frankfurt:Frankfurt School Blockchain Center, Jun. 2017.

Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services, 14(4), 352-375.

Zhenyu, L., Gaogang, X., Zhongcheng, L., Yunfei, Z.,& Xiaodong, D. (2008). DHT-Aid, Gossip-Based Heterogeneous Peer-to-Peer Membership Management. 2008 5th IEEE Consumer Communications and Networking Conference. http://dx.doi.org/ 10.1109/ccnc08.2007.70.

CHAPTER 4

DATA IN THE CONTEXT OF INDUSTRY 4.0

Fatma Önay KOÇOĞLU*, Denizhan DEMİRKOL**

*Dr., İstanbul University, Informatics Department, İstanbul, TURKEY
E-mail: fonayk@istanbul.edu.tr

**Research Assistant, Aydın Adnan Menderes University, Söke Business Faculty, Management Information Systems Department, Aydın, TURKEY
E-mail: denizhan.demirkol@adu.edu.tr

DOI: 10.26650/B/ET06.2020.011.04

Abstract

Today, every sector, not least industry, has been affected by the development of technology. With the breakthrough development of technology, Industry 4.0 has emerged with the concept of big data. Data is the most important element in the process of creating information. This study aims to deal with the subject of Industry 4.0 which has attracted great interest in the global field in the context of big data. Studies concerning Industry 4.0 and related data are examined in our study through a systematic literature review. Web of Science database and "industry 4.0 and data" keywords were used for our article search. A preliminary evaluation was performed for 20 articles meeting the objective of this study which were selected for detailed examination. When the studies on Industry 4.0 and data are analyzed, we can determine that studies with big data, digitalization, internet of things, digital twin, cyber-physical systems, smart factories and cloud computing are prominent. Moreover, when the countries where the articles were published were analyzed, it was found that China was the most cited and studied country in this field. It is believed that the results of this examination will enlighten people working in this field and direct future studies.

Keywords: Big data, Industry 4.0, Literature review

1. Introduction

Nowadays, the importance of knowledge is an undeniable fact. Decision-makers need the right knowledge in order to make decisions with minimum errors, to develop effective strategies and business models. Knowledge should be revealed by making maximum use of availableresources. For this purpose, the most important source is data, which is the building block of knowledge. Data is obtained from various sources, recorded and analyzed in the best way. Much effort is being put into developing various methods for transforming data into meaningful and useful information. On the one hand, the information and communication technologies developing day by day significantly increase both the computing power and memory capacity of computers, and on the other hand, different devices, platforms, and applications that enable data collection are being developed. All these improvements allow for the storage of much larger amounts of different types of data. It is possible to say that data is collected in many areas from health to education, from security to economy, from tourism to transportation. Data has been continuously recorded and stored in virtual environments through many systems such as patient tracking, appointment and pharmacy systems, student information systems of primary, secondary and higher education institutions, information and document management systems used in public/private institutions or organizations, social media platforms, geographic information systems, seismic record collectors, satellite and air vehicles. This has led to an increase in the amount of data stored day by day, and data has begun to pile up in stacks. In addition to the increase in quantity, non-structural data, which is irregular, non-conforming, and contains text, sound, and image content, have also been stored as well as structural data, which is generally recorded in tables and conforms to a predetermined pattern. As a result of all these developments, the concept of big data has been mentioned. In this direction, it can be said that data-based concepts and methods are shifted to the big data axis.

On the other hand, one of the areas affected by the development of information and communication technologies is industry. Industry has passed through three main periods and reached the fourth period mentioned today. In addition to the developments in the third period, the integration of high technologies such as robotics and digital technology into production systems, and the use of the Internet more effectively in these systems brought about a new era in the field of industry. There has been a need for a systematic approach in order to integrate the systems (especially smart systems) equipped with the new technologies mentioned above in production areas in order to make production continuous in the factories and to minimize breakdown (Bagheri, Yang, Kao, & Lee, 2015). In this direction, the fourth period, which is mentioned together with the concepts of cyber-physical systems, the internet

of objects, smart factories, etc., continues to evolve today. This period is called the 4th Industrial Revolution or Industry 4.0. Within the scope of Industry 4.0, the use of ubiquitous information and communication technology (ICT) infrastructure is considered to contribute greatly to sustainable production. (Stock & Seliger, 2016). Industry 4.0, also known as Intelligent Manufacturing, Industrial Internet, or Integrated Industry (Hofmann & Rüsch, 2017), is defined as a network approach in which components and machines become intelligent and are part of a standardized network that is built to strictly defined Internet standards (Kolberg & Zühlke, 2015). One of the factors that accelerated the transition to Industry 4.0 is the surprising increase in data volume and computing power and capabilities (Baur & Wee, 2015).

This study aims to deal with the subject of Industry 4.0 which has attracted great interest in the global field in the context of big data. In this respect, firstly, the basic concepts of Industry 4.0, data and big data have been discussed and a literature review has been conducted. In the literature research, Industry 4.0 studies carried out within the framework of both general applications and data-big data have been examined, and the importance of data and big data concepts for Industry 4.0 and its future have been discussed in the discussion and conclusion section.

2. Industry 4.0

With the opportunities provided by developing technology for the benefit of human beings, it is seen that there are shifts to digital environments in many areas from finance to education, and from health to safety in terms of habits and business processes. Shopping on digital platforms, the use of internet banking, the recording of physical documents on computers, the use of e-mail instead of mail, the selection of students' courses , the announcement of exam results, and the sharing of lecture notes through management information systems can all be shown as examples of this situation. This is referred to as digitization or digital transformation. Digital transformation is a process that results in differentiation in the fields of enterprises such as products, organizational structures, and automation processes with the integration of digital technologies into business models (Matt, Hess, Benlian, & Wiesbock, 2016). Digital transformation is the deep and rapid transformation of business activities, processes, competencies, and models by using the changes and opportunities brought about by digital technologies in order to fully strengthen the impacts of these technologies on society in a strategic and priority manner (Demirkan, Spohrer, & Welser, 2016). The potential benefits of digitization are enormous, especially in terms of sales or productivity increases and innovations in value creation (Matt, Hess, & Benlian, 2015).

One of the areas where digital transformation shows its effect is industry. Industry has reached the present day by passing through different periods together with the systems developed according to the conditions of the day and included in the production process. These periods are called the first, second, third and fourth industrial revolutions. The systems that affect the transformation can be counted as steam-powered mechanical systems, electrical energy-powered systems, computer-based automation systems, and finally, intelligent systems that can communicate with each other via the internet and decide on their own.

Globalization has lifted international borders and it has led to a shift in the competition to a different dimension in all areas. Increased product and process complexity, variable market characteristics, shortened product, market, technology, and innovation cycles are some of the challenges in the competitive environment (Rennung, Luminosu, & Draghici, 2016). In this respect, various strategies that are innovative and increase productivity are developed and different investment proposals are evaluated in order to survive under tough competitive conditions. The purpose of assessing savings through various investment options or making strategic decisions is to generate more revenue in the future. In this case, one of the important points is to provide the highest return with minimum risk. On the other hand, The development of information and communication technologies and the increasing importance of information have attracted the attention of countries in order to make new initiatives especially in the field of industry. Also, the advantages and disadvantages of shaping strategies and investments around changing social habits, knowledge and new technologies have started to be discussed. Especially with the integration of internet and sensor technology to existing production systems, a new production model has been proposed. This model, which is referred to as Industry 4.0, was set out in 2011 by a proposal file submitted to the German government by a working group under the direction of Robert Bosch GmbH and Henning Kagermann. With the final report announced at the Hannover Fair in Germany in 2013, it has become a topic of interest on the world agenda. Factors that accelerate the transition to Industry 4.0 include increased amount of stored data, increased computing power of computers, improved analytical and business intelligence solutions, improved interfaces in human-machine interaction, and the ease of transforming digital guidelines into the physical world (Baur & Wee, 2015).

Although it seems to be the development of computerized manufacturing systems that led to the third industrial revolution with new technologies, Industry 4.0 is based on a network model that includes not only the automation of value chain elements but also the integration of these tools with continuous communication and real-time features (Hopali & Vayvay,

2018). In this direction, Industry 4.0 has a concept shaped around the concepts of cyber-physical systems, internet of things and services, wireless communication, industrial internet, intelligent production and cloud-based production (Almada-Lobo, 2015; Vaidya, Ambad, & Bhosle, 2018). Industry 4.0 is a model associated with data exchange based on connectivity between new technologies on the one hand, and with automation on the other hand (Fantoni, Chiarello, Fareri, Pira, & Guadagni, 2018). To put it more clearly, it is a production model in which a real factory has one-to-one representation in a virtual environment, where machines, robots and people can communicate with each other via the internet and intelligent machines (robots) can manage themselves by analyzing the data, which is collected from different sources with the help of sensors, in decision systems. In this way, flexible and fast solutions can be produced and resources are used more efficiently. The Industry 4.0 model can be summarized as follows (Koçoğlu, 2018):

- Forming intelligent factories with a modular structure consisting of sensors that can detect the environment and intelligent robots carrying out production activities,
- Creating a cyber-physical system in which a virtual object of every object in physical structure is created and communication between objects and people is provided via the internet,
- Recording the data flowing into the system,
- Obtaining high quality and efficient production with less error by processing this big data.

With the new production system proposed within the scope of Industry 4.0, a higher production automation level is targeted by optimizing production management and with the production safety and training of employees (Wu & Duan, 2018). Industry 4.0 aims to make factories smart enough in terms of adaptability, resource efficiency and improved integration of supply and demand processes (Varghese & Tandur, 2014). The benefits of Industry 4.0 with the subjects on the use of idle data, production time and personalization are strengthening this model (Schmidt et al., 2015).

In order to make Industry 4.0 more meaningful, the functions of new technologies used for this production model and their roles in the system must be known. In this context, the most important components of Industry 4.0 are explained below.

Cyber-Physical Systems (CPS): This is an essential part of the Industry 4.0 model. Cyber-Physical Systems are the integration of physical processes with the virtual processes and physical processes are monitored and controlled through embedded computers, sensors,

various software and networks (Lee, 2008). These two systems communicate over the cyber-physical system and thus work synchronously with each other. The ability of these systems to interact with the physical world and expand their capabilities through computing, communication, and control is crucial to future technological advances (Baheti & Gill, 2011).

Internet of Things (IoT): Communication in the future will not only be between people. Similarly, access to information will not only be requested by people. On behalf of people, machines will try to communicate with other machines and collect data (Tan & Wang, 2010). All this communication will take place via the internet. Internet of things whose architecture is technically based on data communication tools, primarily RFID, aims to facilitate the exchange of information between all objects defined on the network (M. Wu, Lu, Ling, Sun, & Du, 2010). In other words, all objects (human or machine) defined in the cyber-physical system will use the internet of things for communication.

Smart Factory: The change in the close connection and communication between products, machinery, transport systems and people by means of the other technologies mentioned above also indicates a change in the existing production logic (Hofmann & Rüsch, 2017). The factories of the future will be much more than a system where production resources are interconnected and where they automatically exchange information. According to this, a sufficiently intelligent system will emerge that can predict the problems that may arise and determine the required maintenance times, control the production process and manage the machines (Qin, Liu, & Grosvenor, 2016). With smart factories, the aim is to realize a flexible and adaptable intelligent production process that is able to adapt quickly to change, is based on automation, manages the machine by reducing human intervention and uses resources efficiently.

Cloud Computing: Instead of meeting their hardware, system and software needs within the framework of the resources in the enterprise because of disadvantages in terms of cost, flexibility, complexity of infrastructure, and storage of data, companies provide these needs from outside. Cloud computing is a flexible and inexpensive technology that provides services including infrastructure, software, hardware, platforms, and other information technology infrastructure resources when needed. Users can use the services provided to them according to application requirements and based on access to computer and storage systems (Zhou, Liu, & Zhou, 2015).

Human-Computer Interaction (HCI): With the Industry 4.0 model, the integration of new technologies into production systems is becoming more complex due to a growing

communication network. In addition to this, it is clear that there will be changes in the duties and responsibilities of people within this system. Also, it should be mentioned that it will be a large system where people will interact with computers and machines. The field of study on interaction and communication between computer and human is called human-computer interaction. Human-computer interaction studies, which started with the aim of creating highly usable interface designs, have been expanded with the inclusion of tasks, actions, explanations, reasons, and discussions in the study (Fischer, 2001). Human-computer interaction is concerned with how people benefit from existing systems and devices, as well as the design of new interactive systems and devices that will enhance human performance and experience (Carroll, 2006). As a part of human-computer interaction, virtual and augmented reality concepts gain importance in the Industry 4.0 model. Virtual Reality (VR) is a technology that allows the user to simulate and interactively explore the behavior of a CPS-based production system (Gorecky, Schmitt, Loskyll, & Zühlke, 2014), while Augmented Reality (AR) is a technology that integrates real and virtual object (world) images using various computer graphics. These two technologies maximize human interaction with the computer system via a display interface. For example, computer-based training for factory employees can be carried out in a one-to-one simulation of a real factory created using VR or AR technologies, thus the effectiveness of the training can be increased.

Big data is also among the aforementioned components of Industry 4.0. In this study, the concept of big data is discussed in more detail below.

3. Big Data

Before the concept of big data, the concepts of data-information-knowledge should be considered. Data is a collection of unprocessed characters that describe any state, object, or event and hold their values. Information is defined as data when its form and benefit are increased as a result of various processes such as calculation, merging, categorizing and summarizing (Ackoff, 1989). The data alone does not make any sense. The meaning and benefit of information are limited. Therefore, data needs to be transformed into knowledge in order to make it meaningful, in other words, useful for predicting, developing strategies or making decisions. The content obtained as a result of the integration of information through a certain filter, analysis and synthesis, individual experience and expert opinions for the purpose of creating benefit and value is called knowledge. A significant portion of the information that is considered important is used by recording and is considered to have been transformed into knowledge only when it becomes of value for individuals or institutions (Dinçmen, 2010). As can be seen, information can be obtained as a result of the conversion

of data, whereas knowledge can only be built on the information obtained from the data. The link between data, information, and knowledge is represented by the knowledge hierarchy pyramid (Figure 1). When the pyramid is examined, value and meaning increase in the direction from data to knowledge , as indicated by the definitions, and decrease in the opposite direction.

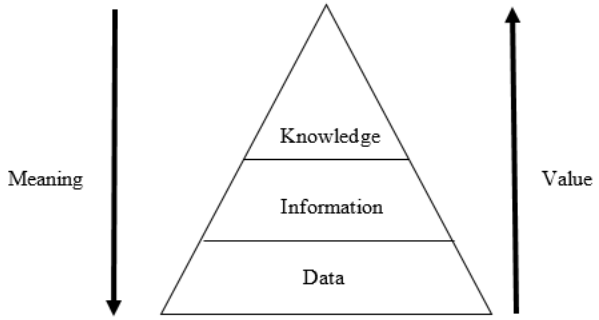


Figure 1: Knowledge hierarchy pyramid (Chaffey and Wood, 2005)

The rapid development of information and communication technologies has increased the computing power of computers and has enabled the creation of larger storage areas for the recording of data. Accordingly, all kinds of data have been stored in order to reach knowledge. Every transaction we perform in the market vault or online banking, all kinds of results obtained in health screening (blood values, MR / BT images, physical examination results, etc.), camera images obtained in different areas such as shopping malls, schools, hospitals, squares, call centers' voice recordings, such as the values produced by the machines in production lines in factories, all these are examples of this recorded data. In addition to the data currently being recorded, companies that discover the importance of the data will start using real-time data from various descriptive devices to understand their workflows in detail, develop innovative products and services, and keep pace with changes (Davenport, Barth, & Bean, 2012). By taking this situation into consideration, it is clear that the size of the recorded data increases day by day. Big data has emerged in connection with the challenges that organizations face in trying to cope with rapidly growing data sources (Villars, Eastwood, & Olofson, 2011). Big data is the term used to describe data sets that cannot be managed by existing methodologies due to both the size and complexity of the structure (Fan & Bifet, 2013). Big data is a data set obtained from different sources, which is continuously increasing in volume, structured in terms of structure, semi-structured and unstructured data and difficult to manage. As can be seen from the definition of big data, there are three main components.

These are volume, velocity, and variety. In addition to these components, veracity, validity, variability, volatility, visualization and value are among the other components (Demchenko, de Laat, & Membrey, 2014; Owais & Hussein, 2016). The quality of this data set is far from perfect due to its large volume, speed, and heterogeneity (Saha & Srivastava, 2014), and its analysis is not possible with standard methods. However, the big data obtained by combining the data collected from various incompatible systems can be analyzed at an advanced level with the developed big data technologies (Witkowski, 2017).

Obtaining knowledge using big data has become a key way for businesses to know themselves and their competitors better and perform better (Manyika et al., 2011). As in all fields, in production also, it is important to process the data correctly and convert it into useful knowledge. In the industrial field, big data emerges with the combination of data generated by values such as vibration-pressure, which can be measured by various sensors installed as a result of intelligent and cyber-physical systems in the manufacturing process, data obtained through communication protocols between all assets in the system and historical data (Lee, Kao, & Yang, 2014). The big data obtained in this direction can be used to increase the quality of pruduction, reduce resource wastage and maintenance costs, shorten the material supply and production time, faciliate more efficient planning, and increase efficiency.

Until this section, the Industry 4.0 model, its components and big data concepts have been discussed and basic explanations have been given. In the next section , the literature research and its results are shared.

4. Methodology

Within the scope of the study, research was carried out on the examination of articles through systematic review. The article selection was made on the Web of Science (WOS) database.

The articles were determined according to the index status of the journal, the number of citations received by the articles and the keywords which had previously been determined.

Articles published between 2013-2019 were searched using the WOS database. Since the concept of industry 4.0 became official as of 2013, 2013 was chosen as the start of publication date.

- “Industry 4.0 and data” keywords were used for article search. Keywords were searched in topics of the articles.

- The articles selected for examination were published in journals indexed by Social Sciences Citation Index (SSCI), Sciences Citation Index Expanded (SCI-Expanded), Arts & Humanities Citation Index (A&HCI), Book Citation Index- Science (BKCI – S), Book Citation Index- Social Sciences & Humanities (BKCI-SSH) and Emerging Sources Citation Index (ESCI).
- "Highly Cited in the Field" and "Hot Papers in Field" options were selected and all articles were listed in the WOS Database.
- The number of articles was limited to 30. Therefore, only the first 30 articles were selected for evaluation.
- A preliminary evaluation was performed for 30 articles. For the purpose of the study, only 20 articles met the objective of this study. As a result of the preliminary evaluation these 20 articles were selected for detailed examination.

The selected articles were examined in terms of general research subject, application sector, applied methods and acquired general results. All articles were summarized under these terms.

5. Findings

Twenty articles published between 2013 and 2019 were examined in this study. The summary of the article evaluation is shared below and a summary table is given in Appendix.

Ardito, Petruzzelli, Panniello, & Garavelli (2019) conducted research on the digital technologies used to show the connection between supply chain management and marketing processes and to manage related interfaces. Real samples and patent analysis methods were used for this aim. United States Patent and Trademark Office (USPTO) database was used to find the categories defined within the scope of the research. Patent count analysis was performed at technology level. Annual patent count was used as a measurement of innovation efforts in time. A tendency for a sharp increase was observed for patents especially since 2013. This article becomes prominent as one of the studies supporting the integration of supply chain management and marketing regarded as a key success factor to survive in a competition environment and to acquire an outstanding financial performance. From a manufacturer's perspective, Gu, Guo, Hall, & Gu (2019) suggested an integrated architecture to acquire an efficient and productive Extended Producer Responsibility (EPR) through a case study approach. Based on the selected case study, Haier's architecture of a smart refrigerator facility of integrated information systems and facilitated life cycle management

was observed. The objective of Xu & Duan, (2019) was to better remark the importance of cyber-physical systems and big data for Industry 4.0. They stated that Industry 4.0 was especially supported by the German government and since it is a new concept formed by an advanced production vision, there are few studies in the literature on the big data and cyber-physical systems which are among the constituents of this concept. Cyber-physical systems are critical constituents representing the efficient and productive management of data acquired from the basic infrastructure of Industry 4.0 and the big data with the data acquired from cyber-physical systems. In the study, they emphasized the necessity of big data technologies which provide for the processing of the big data and which consistently improve the scalability, safety and efficiency of cyber-physical systems.

Ivanov, Dolgui, & Sokolov (2019) investigated the impact of digitalization and Industry 4.0 on the ripple effect and supply chain disruption risk control analytics of the supply chain. This study demonstrated that digital technologies increased the responsiveness and capacity elasticity of demand at the proactive stage. It was also stated that this condition may have a positive effect on the decreases in risk, thus decreasing inventory in the ripple effect control. It was detected that depending on additive production, shorter delivery times and digitalization would increase the effect on inventory control. Wan et al. (2019) presented a multi-dimensional data indexing scheme which was designed to solve the range query and is effective in terms of energy and time. This scheme collects multi-characteristic sensor data in an energy and time efficient manner. Hierarchical intranet storage was used to provide quick query responses. Through performance evaluation, the efficiency of the suggested scheme compared to other data structures was also proven. In their study, Tao & Qi (2019) focused on the fact that new information technologies particularly promote smart manufacturing, and they aimed to present a general framework on the subject. They stated that new technologies such as internet of things (IoT), cloud computing, big data, mobile Internet and cyber-physical systems (CPS) would especially play a significant role in promoting smart manufacturing. Service-oriented smart manufacturing (SoSM) framework using the afore-mentioned technologies was also suggested in the research. It was stated that the suggested framework had a critical importance particularly in the facilitation of smart manufacturing.

Reyna, Martín, Chen, Soler, & Díaz (2018) conducted a study to detect the difficulties in IoT's integration with blockchain and to analyze potential future advantages. Key points through which Blockchain technology may facilitate the improvement of IoT applications were determined and also an evaluation was performed to prove the applicability of using blockchain knots in IoT devices. At the end of the study, it was predicted that Blockchain

would be revolutionary in IoT. Li (2018) compared Germany's "Industry 4.0" and China's "Made-in-China 2025" and estimated China's locus in "Made-in-China 2025". Data and information from the World Bank Data and China's National Bureau of Statistics were used to analyze the potential of advancing the plan from "Made-in-China" to "Designed-in-China". The acquired results show that China is not the workforce market with the lowest cost in the current situation. It was also found that China is not the strongest player in the high technology arena and that well-established industrialized countries such as USA, Germany and Japan actively use digital technology to build new industrial environments, to manufacture new products and to improve their well-established trademarks. But it was stated that China has an ascending course in terms of manufacturing potential, research development and human capital investment subjects. In their research, de Sousa Jabbour, Jabbour, Foropon, & Godinho Filho (2018) suggested a framework providing for the integration of two industrial waves (industry 4.0 and sustainability) and promising the restructuring of present manufacturing and consumption habits. It was emphasized that Industry 4.0 and industrial sustainability had improved separately until today and that the integration of these two factors is at the beginning stage. In the study, 11 critical success factors, which should be carefully considered by institutions especially during the concurrent application of Industry 4.0 and environmentally-sustainable manufacturing, were pointed out. These factors were listed as management and leadership, being ready for organizational change, decisiveness of senior management, strategic cooperations, training and capacity increase, keeping high motivation of the employees, teamwork, building a company culture, communication, project management and understanding regional differences. Finally, it was stated in the study that Industry 4.0 may increase environmentally sustainable manufacturing especially by improving green products, green production phases and green supply chain management as never seen before.

Moeuf, Pellerin, Lamouri, Tamayo-Giraldo, & Barbaray (2018) carried out a literature research on the currently applied researches performed within the concept of Industry 4.0 in terms of SMEs. The results show that SMEs did not use all available resources to apply Industry 4.0 and generally limited themselves to Cloud Computing and Internet of Things constituents of Industry 4.0. From a different standpoint, it was detected that SMEs adopted Industry 4.0 concepts only to follow industrial phases but there were no actual applications for manufacture planning. Finally, it was determined that Industry 4.0 projects in SMEs remained as cost-oriented initiatives and did not transform into an actual work model. In the research conducted by Qi & Tao (2018) to determine the role of the mutual use of digital twin

and big data in the improvement of smart manufacturing, it was concluded that both technologies played a significant role in promoting smart manufacturing. It was stated that digital twin concept would lead to the preparation of cyber-physical integration basis by providing the manufacturers with the chance to manage real-time matchings between a physical object and its digital representation. Some interesting aspects of this study were the facts that through the correct analysis of big data, digital twin-driven smart manufacturing can be performed in a more sensible and predictive way and would be more rational in many aspects and advantageous for absolute manufacturing management. As a result, digital twin and big data were detected as two factors complementing each other to assist smart manufacturing. In their research, Chen et al., (2018) aimed to determine the hierarchical architecture of a smart factory and then to analyze its different layers. The main constituents of the basic technologies at physical source layer, network layer and data application layer in the smart factory were analyzed. It was emphasized that investigating the key technologies would not mean the integration of information technology only as the main structure, but it should also cover traditional disciplines such as control theory, mechanic technology, material and energy. It was stated that with the progression of big data technology, the product quality and manufacturing efficiency of databased virtual manufacturing mode would increase, and the energy consumption would decrease. It was also stated that big data-dependent smart manufacturing would provide the acceleration of the industrial revolution. Müller, Kiel, & Voigt (2018) examined the factors improving the application of Industry 4.0 and the role of sector opportunities and difficulties in terms of sustainability. A research model covering the opportunities on Industry 4.0 and the difficulties primarily faced in its application was suggested in the study. To test the model, PLS-SEM (Partial Least Squares Structural Equation Modeling) was applied on a sample acquired from 746 German manufacturing companies from five industrial sectors. The results show that as much as strategic and operational ones, environmental and social opportunities also have positive impulses on the application of Industry 4.0. On the other hand, it was observed that the difficulties with competitive power and future applicability, as well as organizational and manufacturing suitability, prevented the progress.

In their research which was conducted to suggest and apply a big data solution for active preventive maintenance in manufacturing environments, Wan et al. (2017) built the system architecture required primarily for active preventive maintenance. Analyzing the collection method according to data characteristics in big data production, cloud data processing including cloud layer architecture, real time active maintenance mechanism and offline

prediction and analysis method was performed. Prototype platform was analyzed, and the suggested active preventive maintenance method was compared to the traditional method. As a result, it was shown that the suggested model had the potential of accelerating the Industry 4.0 application. The aim of the study performed by Sikorski, Haughton, & Kraft (2017) was to investigate the applications of blockchain technology related to the 4th Industrial Revolution (Industry 4.0) and to present a sample of chemistry industry using block chain to facilitate machine-machine interactions. In the presented scenario, two electricity producers trading with each other through a block chain and an electricity consumer are present. The producers present energy trading (in kWh) in foreign exchange (USD) in a dataflow. The consumer reads and analyzes the offers and tries to meet the energy demand with minimum cost. When an offer is accepted, it is applied as an atomic exchange. This study contributes to the clear application of the described scenario and its technical details. In current literature, Liao, Deschamps, Loures, & Ramos (2017) stated that the efforts to systematically examine the condition of the industrial revolution wave are still inadequate. For this aim, examining the academic progresses in Industry 4.0, they covered this deficit in their research. A systematic literature search was performed to analyze academic articles on Industry 4.0. Results acquired through the general data analysis and specific data analysis of the articles included in the research were shown and discussed. These results not only summarized the available research activities, but also covered the present deficiencies and potential researches for the future, offering a research suggestion. It was stated that the findings of this inspection can be used as the basis of future studies on Industry 4.0 and related subjects. Starting from the target to develop data-based manufacturing information system architecture, Theorin et al. (2017) developed The Line Information System Architecture (LISA) designed to provide flexible factory integration and data use. The main targets of LISA were stated as focusing on the integration of device and services at all levels and supporting constant recoveries in the information visualization and control in addition to the integration of new smart services.

In their research, Wan et al. (2016) presented a new concept suggestion for industrial environments by presenting software-defined Internet of Things to make the industrial wireless networks more flexible. IoT architecture was analyzed in detail and the information interaction between different devices was explained. As a result, the interface of a software-based IoT architecture was designed to manage physical systems. In their research, Wang, Wan, Zhang, Li, & Zhang (2016) presented a smart factory framework combining industrial network, cloud and supervisory control terminals with smart workshop objects. Then they classified smart objects based on different agent types and defined a coordinator in the cloud.

Based on this model, a smart mechanism was suggested to provide the cooperation of agents and crashing was prevented through improving the decision making of the agents in the model and the actions of the coordinator.

Zhan et al. (2015) stated an increase in the use of Evolutionary Computation (EC) algorithms. It was stated that it would be advantageous to investigate the general role of EC in the development of cloud computing and its planning for the big data via the Internet. It was also asserted that deep learning in EC algorithms for predictive data analytics for cloud computing scheduling would be one noteworthy theme in the Industry 4.0 stage. At the beginning of Industry 4.0, planning for big data and cyber-physical cloud computing was explored. It was found that research in this field was just in its earliest stages and newer issues would arise with the fast improvement of cloud computing, big data and Internet of Things.

5. Discussion and Conclusion

Having useful information is very important for enterprises to understand their competitive environment and their position in this environment, to evaluate their internal operations correctly and to make the necessary strategic decisions with minimum errors. Realizing this situation, institutions and organizations have continuously tried to record all kinds of data they can obtain. Therefore, besides resources such as labor, materials and equipment, information has been among the important sources. Such an increase in the importance of information has resulted in continuous data collection from all possible sources. Developing information technologies and increasing expectations led to the recording of data with a different structure and an increase in volume. This resulted in a large volume of data sets that are difficult to cope with and the concept of big data was put forward. Moreover , the industry has now entered a new transformation process which is called Industry 4.0. The basic logic of Industry 4.0 is based on the representation and communication of people, machines, robots and all other physical components in the production system through cyber-physical systems and the internet of objects. Within this structure, data will be generated both by sensors and during the communication process with the internet of objects. Therefore, within the scope of Industry 4.0, data and even big data become one of the important components. Smart factories targeted at Industry 4.0 and efficient self-decision systems can only be obtained by processing this data and producing the accurate knowledge.

Since the official launch of the concept in 2013, the number of studies carried out within the scope of Industry 4.0 has been increasing. In this study, it was seen that the articles

examined are mainly made up of cyber-physical systems, smart factories, digitalization, internet of things, cloud computing, and digital twin. When the evaluation is made on the basis of big data, it was seen that the subject is considered as one of the important components of Industry 4.0 but that it is not in the focus of the studies. This situation is considered to be of importance because the studies related to Industry 4.0 are still in the conceptual, basic application and model installation stage. In the second part of the study, four steps of Industry 4.0 were mentioned. The third and fourth of these steps are the collection of data, following the integration of robot and sensor technologies into the production systems of cyber-physical systems and the completion of the infrastructure of Industry 4.0. Therefore, after the establishment of the Industry 4.0 infrastructure and production started to be performed within the scope of this model, the data produced will increase and the methods and technologies required for processing this big data will increase This is an indication that big data studies will gain more importance and speed.

In some of the studies examined, it was seen that predictive maintenance particularly comes to the forefront. Predictive maintenance is used to monitor the equipment, to estimate possible failures before the failure occurs, to take the necessary measures and to reduce the negative effects of time, cost and efficiency. In predictive maintenance, big data analysis methods and technologies are utilized due to the developments in the industrial field, the machines that have turned into a complex structure and the increasing data size as a result. Some of important results of Industry 4.0 and big data are the reduction of maintenance costs with the information obtained from the analyzes and the self-determination of the maintenance times of the machines without the need for any control.

.

The concept of Industry 4.0 is a production model introduced in Germany and the first studies were carried out in European countries, especially in Germany. Other industrialized countries that do not want to fall behind in the competition have been included in this industrial movement. America, Japan, Canada, and China are among these countries. However, some of these countries have evaluated the proposed Industrial 4.0 model from their own perspectives and have proposed new models. Society 5.0 is one of these models, which Japan specifically suggests. Our research shows that most of the articles that are obtained as a result of the article review conducted in this study are based in China. It is clear that China, which has taken important steps in the field of industry and economy, also gives importance to Industry 4.0. When evaluated in general, it is evident that there is a lot of interest in this new production model, called Industry 4.0 or otherwise, and that the industry

will shift rapidly towards this new model on a global basis . Accordingly, scientific research articles show that different countries from all regions of the world are interested in this subject and try to improve themselves in this field. For countries that do not want to be left behind in this movement of change, it can be said that it will be a tight competition.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1), 3-9.

Almada-Lobo, F. (2015). The ındustry 4.0 revolution and the future of manufacturing execution systems (MES). *Journal of Innovation Management*, 3(4), 16–21. https://doi.org/10.24840/2183-0606_003.004_0003

Ardito, L., Petruzzelli, A. M., Panniello, U., & Garavelli, A. C. (2019). Towards industry 4.0: Mapping digital technologies for supply chain management-marketing integration. *Business Process Management Journal*, 25(2), 323–346. https://doi.org/10.1108/BPMJ-04-2017-0088

Bagheri, B., Yang, S., Kao, H.-A., & Lee, J. (2015). Cyber-physical systems architecture for self-aware machines in ındustry 4.0 environment. *IFAC-PapersOnLine*, 48(3), 1622–1627. https://doi.org/10.1016/j.ifacol.2015.06.318

Baheti, R., & Gill, H. (2011). Cyber-physical systems. In T. Samad & A. Annaswamy (Eds.), *The Impact of Control Technology*. https://doi.org/10.1109/icmech.2019.8722929

Baur, C., & Wee, D. (2015). Industry 4.0 is more than just a flashy catchphrase. A confluence of trends and technologies promises to reshape the way things are made. *McKinsey Quarterly*, (Jun), 1–5.

Carroll, J. M. (2006). Human–computer interaction. In *Encyclopedia of Cognitive Science*. https://doi.org/10.1002/0470018860.s00545

Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M., & Yin, B. (2018). Smart factory of ındustry 4.0: key technologies, application case, and challenges. *IEEE Access*, 6, 6505–6519. https://doi.org/10.1109/ACCESS.2017.2783682

Chaffey D. & Wood, S. (2005). *Business information management: Improving performance using ınformation systems*. FT Prentice Hall:Harlow.

Davenport, T. H., Barth, P. F. P., & Bean, R. V. (2012). How 'big data' is different. *MIT Sloan Management Review*, 54(1), 22–24.

De Sousa Jabbour, A. B. L., Jabbour, C. J. C., Foropon, C., & Godinho Filho, M. (2018). When titans meet – Can industry 4.0 revolutionise the environmentally-sustainable manufacturing wave? The role of critical success factors. *Technological Forecasting and Social Change*, 132, 18–25. https://doi.org/10.1016/j.techfore.2018.01.017

Demchenko, Y., de Laat, C., & Membrey, P. (2014). *Defining architecture components of the big data ecosystem*. 2014 International Conference on Collaboration Technologies and Systems (CTS), 104–112. https://doi.org/10.1109/CTS.2014.6867550

Demirkan, H., Spohrer, J. C., & Welser, J. J. (2016). Digital ınnovation and strategic transformation. *IT Professional*, 18(6), 14–18. https://doi.org/10.1109/MITP.2016.115

Dinçmen, M. (2010). *Bilgi yönetimine giriş.* [Introduction to Knowledge Management] In M. Dinçmen (Ed.), Bilgi yönetimi ve uygulamaları, Papatya Yayıncılık:İstanbul, Turkey, ISBN:978-605-4220-15-1.

Fan, W., & Bifet, A. (2013). Mining big data: Current status, and forecast to the future. *SIGKDD Explor. Newsl*., 14(2), 1–5. https://doi.org/10.1145/2481244.2481246

Fantoni, G., Chiarello, F., Fareri, S., Pira, S., & Guadagni, A. (2018). *Defining industry 4.0 professional archetypes: A data-driven approach*. In Terence Hogarth (Ed.), Economy, Employment and Skills: European, Regional And Global Perspectives In An Age Of Uncertainty (p. 298). Italy: Fondazione Giacomo Brodolini.

Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction, 11,* 65–86.

Gorecky, D., Schmitt, M., Loskyll, M., & Zühlke, D. (2014). *Human-machine-interaction in the industry 4.0 era*. 2014 12th IEEE International Conference on Industrial Informatics (INDIN) (pp.289–294). https://doi.org/10.1109/INDIN.2014.6945523

Gu, F., Guo, J., Hall, P., & Gu, X. (2019). An integrated architecture for implementing extended producer responsibility in the context of industry 4.0. *International Journal of Production Research*, 57(5), 1458–1477. https://doi.org/10.1080/00207543.2018.1489161

Hofmann, E., & Rüsch, M. (2017). Industry 4.0 and the current status as well as future prospects on logistics. *Computers in Industry*, 89, 23–34. https://doi.org/10.1016/j.compind.2017.04.002

Hopali, E., & Vayvay, Ö. (2018). Industry 4.0 as the last industrial revolution and its opportunities for developing countries. Analyzing the Impacts of Industry 4.0 in Modern Business Environments, 65–80. https://doi.org/10.4018/978-1-5225-3468-6.ch004

Ivanov, D., Dolgui, A., & Sokolov, B. (2019). The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, 57(3), 829–846. https://doi.org/10.1080/00207543.2018.1488086

Koçoğlu, F. Ö. (2018). "Endüstri 4.0" konusu üzerine r programlama dili ile bibliyometrik analiz [Bibliometric Analysis on the topic of "Industry 4.0" with r pogramming language] in *Modern Dönemde Edebiyat, Eğitim, İktisat ve Mühendislik* (pp. 859–889). Ankara, TR:Berikan Yayınevi.

Kolberg, D., & Zühlke, D. (2015). Lean automation enabled by industry 4.0 technologies. *IFAC-PapersOnLine*, 48(3), 1870–1875. https://doi.org/10.1016/j.ifacol.2015.06.359

Lee, E. A. (2008). *Cyber physical systems: Design challenge*s. 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC) (pp.363–369). https://doi.org/10.1109/ISORC.2008.25

Lee, J., Kao, H.-A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP*, *16*, 3–8. https://doi.org/10.1016/j.procir.2014.02.001

Li, L. (2018). China's manufacturing locus in 2025: With a comparison of "Made-in-China 2025" and "Industry 4.0." *Technological Forecasting and Social Change*, *135*, 66–74. https://doi.org/10.1016/j.techfore.2017.05.028

Liao, Y., Deschamps, F., Loures, E. de F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0—A systematic literature review and research agenda proposal. *International Journal of Production Research*, *55*(12), 3609–3629. https://doi.org/10.1080/00207543.2017.1308576

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity (pp. 1–143). Retrieved from McKinsey Global Institute website: https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

Matt, C., Hess, T., & Benlian, A. (2015). Digital transformation strategies. *Business & Information Systems Engineering*, *57*(5), 339–343. https://doi.org/10.1007/s12599-015-0401-5

Matt, C., Hess, T., Benlian, A., & Wiesbock, F. (2016). Options for formulating a digital transformation strategy. *MIS Quarterly Executive*, *15*(2). Retrieved from https://aisel.aisnet.org/misqe/vol15/iss2/6

Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., & Barbaray, R. (2018). The industrial management of SMEs in the era of industry 4.0. *International Journal of Production Research*, *56*(3), 1118–1136. https://doi.org/10.1080/00207543.2017.1372647

Müller, J. M., Kiel, D., & Voigt, K.-I. (2018). What drives the implementation of industry 4.0? The role of opportunities and challenges in the context of sustainability. *Sustainability*, *10*(1), 247. https://doi.org/10.3390/su10010247

Owais, S. S., & Hussein, N. S. (2016). Extract five categories CPIVW from the 9V's characteristics of the big data. *International Journal of Advanced Computer Science and Applications*, *7*(3). https://doi.org/10.14569/IJACSA.2016.070337

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. IEEE Access, 6, 3585–3593. https://doi.org/10.1109/ACCESS.2018.2793265

Qin, J., Liu, Y., & Grosvenor, R. (2016). A Categorical Framework of Manufacturing for Industry 4.0 and Beyond. Procedia *CIRP*, *52*, 173–178. https://doi.org/10.1016/j.procir.2016.08.005

Rennung, F., Luminosu, C. T., & Draghici, A. (2016). Service provision in the framework of industry 4.0. *Procedia - Social and Behavioral Sciences*, *221*, 372–377. https://doi.org/10.1016/j.sbspro.2016.05.127

Reyna, A., Martín, C., Chen, J., Soler, E., & Díaz, M. (2018). On blockchain and its integration with IoT. Challenges and opportunities. *Future Generation Computer Systems*, *88*, 173–190. https://doi.org/10.1016/j.future.2018.05.046

Saha, B., & Srivastava, D. (2014). *Data quality: The other face of big data*. 2014 IEEE 30th International Conference on Data Engineering (pp.1294–1297). https://doi.org/10.1109/ICDE.2014.6816764

Schmidt, R., Möhring, M., Härting, R.-C., Reichstein, C., Neumaier, P., & Jozinović, P. (2015). Industry 4.0 - potentials for creating smart products: Empirical research results. In W. Abramowicz (Ed.), *Business Information Systems* (pp. 16–27). https://doi.org/10.1007/978-3-319-19027-3_2

Sikorski, J. J., Haughton, J., & Kraft, M. (2017). Blockchain technology in the chemical industry: Machine-to-machine electricity market. *Applied Energy*, *195*, 234–246. https://doi.org/10.1016/j.apenergy.2017.03.039

Stock, T., & Seliger, G. (2016). Opportunities of sustainable manufacturing in industry 4.0. *Procedia CIRP*, *40,* 536–541. https://doi.org/10.1016/j.procir.2016.01.129

Tan, L., & Wang, N. (2010). Future internet: The internet of things. *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, *5*, V5-376-V5-380. https://doi.org/10.1109/ICACTE.2010.5579543

Tao, F., & Qi, Q. (2019). New IT driven service-oriented smart manufacturing: Framework and characteristics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(1), 81–91. https://doi.org/10.1109/TSMC.2017.2723764

Theorin, A., Bengtsson, K., Provost, J., Lieder, M., Johnsson, C., Lundholm, T., & Lennartson, B. (2017). An event-driven manufacturing information system architecture for Industry 4.0. *International Journal of Production Research*, *55*(5), 1297–1311. https://doi.org/10.1080/00207543.2016.1201604

Vaidya, S., Ambad, P., & Bhosle, S. (2018). Industry 4.0 – A glimpse. *Procedia Manufacturing*, *20*, 233–238. https://doi.org/10.1016/j.promfg.2018.02.034

Varghese, A., & Tandur, D. (2014). *Wireless requirements and challenges in industry 4.0*. 2014 International Conference on Contemporary Computing and Informatics (IC3I) (pp.634–638). https://doi.org/10.1109/IC3I.2014.7019732

Villars, R. L., Eastwood, M., & Olofson, C. W. (2011). Big data: What it is and why you should care. Retrieved from http://www.tracemyflows.com/uploads/big_data/idc_amd_big_data_whitepaper.pdf

Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H., & Vasilakos, A. V. (2017). A manufacturing big data solution for active preventive maintenance. *IEEE Transactions on Industrial Informatics*, *13*(4), 2039–2047. https://doi.org/10.1109/TII.2017.2670505

Wan, J., Tang, S., Shu, Z., Li, D., Wang, S., Imran, M., & Vasilakos, A. (2016). Software-defined industrial internet of things in the context of industry 4.0. *IEEE Sensors Journal*, 1–1. https://doi.org/10.1109/JSEN.2016.2565621

Wan, S., Zhao, Y., Wang, T., Gu, Z., Abbasi, Q. H., & Choo, K.-K. R. (2019). Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things. *Future Generation Computer Systems*, *91*, 382–391. https://doi.org/10.1016/j.future.2018.08.007

Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: A self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, *101*, 158–168. https://doi.org/10.1016/j.comnet.2015.12.017

Witkowski, K. (2017). Internet of things, big data, industry 4.0 – Innovative solutions in logistics and supply chains management. *Procedia Engineering*, *182*, 763–769. https://doi.org/10.1016/j.proeng.2017.03.197

Wu, M., Lu, T.-J., Ling, F.-Y., Sun, J., & Du, H.-Y. (2010). Research on the architecture of internet of things. *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, *5*, 484–487. https://doi.org/10.1109/ICACTE.2010.5579493

Wu, Y., & Duan, Y. (2018). "Made in China": Building chinese smart manufacturing image. *Journal of Service Science and Management*, *11*(06), 590–608. https://doi.org/10.4236/jssm.2018.116040

Xu, L. D., & Duan, L. (2019). Big data for cyber physical systems in industry 4.0: A survey. *Enterprise Information Systems*, *13*(2), 148–169. https://doi.org/10.1080/17517575.2018.1442934

Zhan, Z.-H., Liu, X.-F., Gong, Y.-J., Zhang, J., Chung, H. S.-H., & Li, Y. (2015). Cloud computing resource scheduling and a survey of its evolutionary approaches. *ACM Computing Surveys*, *47*(4), 1–33. https://doi.org/10.1145/2788397

Zhou, K., Taigang Liu, & Lifeng Zhou. (2015). *Industry 4.0: Towards future industrial opportunities and challenges*. 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (pp.2147–2152). https://doi.org/10.1109/FSKD.2015.7382284

Appendix:

Author- Year	Country	Keywords	Field of Application
Ardito, L., Petruzzelli, A. M., Panniello, U., & Garavelli, A. C. (2019).	Italy	innovation, marketing, internet of things, patent analysis, cloud computing, supply chain management, big data analytics, industry 4.0, cyber security	supply chain management marketing integration
Gu, F., Guo, J., Hall, P., & Gu, X. (2019).	China	Industry 4.0, extended producer responsibility, information sharing, integration, life cycle management, smart factory, sustainability	an integrated architecture presentation to implement extended manufacturer responsibility in the context of Industry 4.0
Xu, L. D., & Duan, L. (2019).	United States of America	Industry 4.0, IoT, cloud computing, cyber-physical systems, big data,data science, industrial information integration, engineering	importance of cyber-physical systems for Industry 4.0
Ivanov, D., Dolgui, A., & Sokolov, B. (2019).	Germany, France, Russia	supply chain dynamics, supply chain risk management, supply chain resilience, supply chain design, supply chain engineering, Industry 4.0, additive manufacturing, blockchain, big data analytics, ripple effect	the impact of digitization and Industry 4.0 on the supply chain management
Wan, S., Zhao, Y., Wang, T., Gu, Z., Abbasi, Q. H., & Choo, K.-K. R. (2019).	China, Germany, United Kingdom, United States of America	range query processing, multi-dimensional data indexing, voronoi diagram, IoT energy efficiency	multi-dimensional data indexing approaches in applications of Industry 4.0 and internet of things (IoT)
Tao, F., & Qi, Q. (2019).	China	big data, cloud computing, cyber-physical integration, internet of things (IoT), manufacturing service, smart manufacturing.	new IT driven service-oriented smart manufacturing: framework and characteristics
Reyna, A., Martín, C., Chen, J., Soler, E., & Díaz, M. (2018).	Spain	internet of things, blockchain, smart contract, trust	blockchain and IoT integration applications
Li, L. (2018).	United States of America	Made-in-China 2025, Industry 4.0,emerging economy, cyber-physical systems (CPS), internet of things (IoT) manufacturing capability, human capital, R&D, smart factory, collaborative robots	comparison of "Made-in-China 2025" and "Industry 4.0"
de Sousa Jabbour, A. B. L., Jabbour, C. J. C., Foropon, C., & Godinho Filho, M. (2018).	France, Brazil	industrial wave overlap, Industry 4.0, sustainable operations, environmentally-sustainable manufacturing,critical success factors	critical success factors of Industry 4.0's potential to change the environmentally sustainable production wave

Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., & Barbaray, R. (2018).	France, Canada	production control, Industry 4.0; smart manufacturing, operational improvement, SME, SMB	industrial management practices of SMEs in Industry 4.0 period
Qi, Q., & Tao, F. (2018).	China	big data, digital twin, smart manufacturing, comprehensive comparison, convergence.	comparison of smart manufacturing and big data for digital twin and Industry 4.0
Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M., & Yin, B. (2018).	China, United Kingdom, Germany	smart factory, big data, cloud computing, cyber-physical systems, industrial Internet of Things.	the role of Industry 4.0 in smart factory applications
Müller, J. M., Kiel, D., & Voigt, K.-I. (2018).	Germany	Industry 4.0; industrial Internet of Things, sustainability; implementation structural equation modeling, German industry sectors	Industry 4.0 in the context of sustainability
Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H., & Vasilakos, A. V. (2017).	China, United States of America, Sweden	big data, cyber-physical systems, Industry 4.0, intelligent manufacturing, preventive maintenance.	manufacturing applications of big data solution for active preventive maintenance
Sikorski, J. J., Haughton, J., & Kraft, M. (2017).	United Kingdom, Singapore	blockchain technology,chemical industry, electricity market,machine-to-machine communications	blockchain technology in chemical industry and its applications in machine-machine interaction
Liao, Y., Deschamps, F., Loures, E. de F. R., & Ramos, L. F. P. (2017).	Brazil	the fourth Industrial revolution; Industry 4.0; systematic literature review; qualitative research; quantitative research; research agenda	the past, present and future of Industry 4.0 - a systematic literature review
Theorin, A., Bengtsson, K., Provost, J., Lieder, M., Johnsson, C., Lundholm, T., & Lennartson, B. (2017).	Sweden	automation; agile manufacturing; manufacturing information systems; service-oriented manufacturing systems, event-driven architecture	developing event-based manufacturing information system architecture for Industry 4.0
Wan, J., Tang, S., Shu, Z., Li, D., Wang, S., Imran, M., & Vasilakos, A. (2016).	China Saudi Arabia Sweden	Industry 4.0, industrial wireless networks, industrial Internet of Things, software-defined networks, cyber physical systems.	software-defined industrial internet of things in the context of Industry 4.0
Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016).	China	Industry 4.0,smart factory,cyber-physical system,multi-agent system, deadlock prevention	evaluation of smart factories in the context of Industry 4.0
Zhan, Z.-H., Liu, X.-F., Gong, Y.-J., Zhang, J., Chung, H. S.-H., & Li, Y. (2015).	China United Kingdom	cloud computing, resource scheduling, evolutionary computation, genetic algorithm, ant colony optimization, particle swarm optimization	evolutionary approaches in cloud computing resource planning

CHAPTER 5

BIG DATA GOVERNANCE

Malgorzata PANKOWSKA*

*Professor, University of Economics in Katowice, Faculty of Informatics & Communication, Department of Informatics, Katowice, Poland

E-mail: pank@ue.katowice.pl

DOI: 10.26650/B/ET06.2020.011.05

Abstract

Information processing in a traditional way focuses on relatively stable structured data, repeatable processes as well as on operations in Business Intelligence systems. However, nowadays more and more popular, big data, defined as huge volumes of data available in varying degrees of complexity, generated at different velocities, and varying degrees of ambiguity, cannot be processed using traditional methods and technologies. Some people argue that suitable IT (Information Technology) infrastructure for big data processing is not yet widely developed nor implemented to discuss the big data architecture implementation benefits, risks, and opportunities. Nevertheless, this paper is to present the big data governance issues. Particularly, within the proposed theme, the author discusses the big data system architecture and development strategy. The last part of the paper includes a proposal of a big data architecture model as well as a design of balanced scorecard objectives and measures specification to support the big data governance at public services business organizations. As usual, there are two main research methods, i.e., literature review and the analysis of case studies. The first provides an overview of the existing knowledge and the second permits for contextualization of the proposed models. Beyond that, the paper includes definitions of the key concepts and enables to extend the knowledge base in the research area.

Keywords: Big data, System architecture, ArchiMate, Data governance, Balanced scorecard

1. Introduction to Big Data Issues

The term of big data was coined to describe large data repositories because of the explosive increase of the amount of global data. Its tremendous growth has come from people's daily life, especially in relation to the Internet, social media and mobile devices. In general, big data concerns the datasets, which could not be perceived, acquired, managed, and processed by traditional software tools within an acceptable time and scope. Big data includes data from blogs, tweets, social networking sites, news feeds, discussion boards, video sites, web logs in various, semi-structured formats, as well as machine-generated data, from RFIDs and other sensors, such as optical, acoustic, thermal, seismic, chemical and medical devices. Other examples include shoppers' smartphones, road sensors, GPS devices, TV set boxes, and video cameras. Gathering big data requires modifying the traditional view of the data warehouses. Business organizations are now involved in mixing structured, unstructured and streaming data that often has low latency requirements and still supports queries. Therefore, they need new IT for processing. Business organizations recognize that there is a wealth of data in open social media. Nonetheless, social media data is unstructured and requires different technologies to process and extract useful information.

The objective of data governing for a business organization is to process data as quickly as it is technologically possible while keeping the quality as high as it is practically possible. Traditional approach to data processing concerns processing of transactions and the Internet application data as well as mainframe, OLTP (online transaction processing) system or ERP (enterprise resource planning) system data. For years, data has been stored in transaction and database systems, and in data warehouses. The new approach to data processing is to be creative, holistic, and intuitive, because data is from different sources. According to Krishnan (2013) the basic premises of big data architecture modelling and implementation cover business model transformation by globalization and connectivity, personalization of services, communication media convergence and new sources of data because of the advances in mobile technology, large-scale data processing networks, commoditization of hardware, virtualization, cloud and fog computing. Considering that big data is neither structured, nor does it have a finite state and volume, the complexity of big data relies on the following (Dong & Srivastava, 2015, Gupta & Singla, 2017):

- Data volume, which reflects the amount of unique data converting low-density data into high-density data that has value. The amount of data varies from organization to organization ranging from terabytes to petabytes;

- Data velocity, which is the rate for receiving data. The highest velocity data streams directly into computer memory instead of being written to disk. Many mobile Internet applications enable real-time evaluation and action.
- Data variability that is dissimilar from variety. Almost the same message is transferred in different communication channels, eventually to the same person but in a slightly different form. The meaning of the message constantly changes if the information context changes. Furthermorethe information recipients change their locations or other specific characteristics.
- Data variety, which means that data comes in all types of formats from structured, numeric data in traditional databases to unstructured text documents, emails, video, audio and financial transactions.
- Data value, which is expected to be calculated. Value is derived by a range of quantitative and investigative techniques. The cost of data storage and computation has rapidly decreased because of IT development, but finding the data value requires creation and implementation of analytical processes and involves business analysts, users, and executives. They are learning to ask the right questions, recognizing patterns, and predicting phenomena or behaviours.
- Data veracity that concerns the exactness and precision of data as well as the dependability of information. Because of many types of enormous data, quality and exactness are less controllable, however, the big data investigation innovation allows to work with these different kinds of data as reliable.
- Data visualization, which is critical for data usage. Using charts and graphs to visualize large amount of data is nowadays much more effective than some years ago.

Chen et al. (2014) emphasize obstacles in the development of big data applications:

- Data representation. Many datasets have certain levels of heterogeneity in type, structure, semantics, granularity, organization, provenance, and accessibility, therefore data stewards and managers need techniques and tools to integrate the data.
- High degree of redundancy in datasets, therefore redundancy reduction and data compression tools are expected to be developed.
- Data analytics models, techniques, and tools are constantly required to be developed and implemented because of data volume and velocity.

Big data analytics can be defined as a combination of traditional analytics and data mining techniques along with huge volumes of data to create a platform to analyze, model, and predict the behaviour of customers, markets, products, services, and the competition to enable the achievement of competitive advantage on the market. Traditional data analysis means to use appropriate statistical methods to analyze massive data to concentrate, extract, and refine the useful data from the chaotic. Many traditional data analysis methods are still valid in the new big data environment, e.g., cluster analysis, factor analysis, correlation analysis, regression analysis, but also new techniques of data mining and decision support are implemented. According to Schmarzo (2013), the biggest difference between the business intelligence (BI) analyst and data scientist is the environment, in which they work. BI professional is working in a highly structured data warehouse environment. This environment is typically product or market driven, with highly centralized management of IT services and service level agreements (SLA) implementation in order to ensure timely generation of managerial dashboards and reports. On the other hand, data scientists create separate data sandboxes to load whatever data they can get on both internal and external data sources, and later on they are involved in data cleansing, profiling, transformation, creation of new metrics and models, and testing. Morabito et al. (2015) formulated the process of big data analytics, which included six steps and required specific policies and procedures due to the characteristics of big data:

- Identification of key priorities and recognition of the business context, and setting up the analytics goals.
- Selection of the appropriate data for the analysis.
- Enhancement of data reliability by defining missing data, or removal of irrelevant data and outliers, as well as compilation of data coming from different sources.
- Data mining to verify hypotheses and to extract the meaningful signals.
- Evaluation of data processing results and pattern interpretations.
- Visualisation and reporting on the achieved results.

Data science takes advantage of big data because of its exceptional scale and possibilities to process heterogeneous data, i.e., texts, images, graphs, or sounds. Data analytics allow for deeper insights into the data and improving the quality of products and services offered by business organizations through its multidisciplinary functionality. There is a need to emphasize the difference between big data and data science. Big data is a term

used to concern the exponential growth and availability of data, which can be structured or unstructured. Data science is a research field on knowledge drawn from large volumes of heterogeneous data (i.e., video, audio, text and image). Data science is connected with data analysis, statistics, machine learning and data mining as well as knowledge discovery in databases. Although business analytics cover the use of data-driven insight to generate value, the big data architecture governance is necessary to enable business analytics as well as data science research.

2. Big Data Architecture

Big data analytics is accepted as a very attractive research domain with significant impact on industrial and scientific domains. Belcastro et al. (2017) identified the key research sub-fields, which cover programming models for big data analytics, data storage scalability, data availability by cloud service providers, data interoperability and openness, data quality and usability improvement, integration of big data analytics frameworks, development of tools for massive social network analysis, local mining and distributed model coordination, and in-memory analysis. That research challenges are required to be supported by appropriate system architecture. According to Azarmi (2016) the system architecture modelling should take into account some common issues to create the right sizing. The important ones comprise of defining the appropriate size of daily data input, the structure of the ingested data, the average number of events ingested per second, the retention period, the required availability of data, the expected indexing throughput, the number of visualization users, or the centralization of logs management. System architecture developers, bearing these requirements, already proposed some referential models of system architecture. Therefore, the NIST model can be considered as fundamental for big data architecture. As Heisterberg and Verma (2014) argue that people determine business architecture, process-application architecture, and tools-technology architecture, the NIST model defines three cloud service models appropriate for big data:

- Infrastructure as a Service (IaaS). This includes the storage, servers, and network as the base. The distributed file systems are part of this layer, therefore the big data is also stored in cloud repositories.
- Platform as a Service (PaaS). The NoSQL data storages and distributed caches that can be logically queried using query languages form the platform layer of big data. The layer includes NoSQL and relational databases.

- Software as a Service (SaaS). Specific industries, like health, retail, e-commerce, energy, or banking can build packaged applications that serve a specific business need and leverage the data for cross-cutting data functions.

Heisterberg and Verma (2014) argue that big data architecture is expected to address all type of data coming from various data sources, such as enterprise applications. There is data generated from ERP (enterprise resource planning) systems, CRM (customer relationship management) systems, SCM (supply chain management) systems, e-commerce transactions, HR (human resources) and payroll transactions. Beyond that, there are records from call centers, web logs, smart meters and manufacturing sensors data, equipment logs, and trading systems data generated by machine and computer systems. Companies oriented towards social networking collect social media data, which covers customer feedback streams, microblogging sites like Twitter, and social media platforms like Facebook data.

Big data system architecture is to process business vision and strategy into effective enterprise by creating, communicating and improving the key principles and models that describe the enterprise's future state and enable its continuous transformation. Unhelkar (2018) emphasized some essential advantages of big data architecture development, such as creating a positive impact on the agility of business, expanding new horizons for data analytics and technologies, collaborations by shareable architecture and by involvement in cloud computing. Beyond that, sustainable computing and environmental considerations in business operations are also opportunities made possible through big data architecture. Figure 1 covers big data system architecture model for public services business organization. The proposed model of architecture for realizing big data solutions includes heterogeneous infrastructures, databases, data repositories, and visualization and analytics tools. Many open source frameworks, databases, Hadoop distributions, and analytics tools are available on the market, however, introducing big data requires firstly to answer the question of why it is necessary. The answer is included in the Motivation layer in the model architecture in Figure 1. In general, the proposed model consists of four layers in ArchiMate language and in OMG free software tool. Everything starts from the top layer, which is the Motivation layer, covering the identification of business stakeholders, goals, drivers, principles, and assessments. The model of architecture is to be always allocated in certain context. In this case study, the public service business organization context is proposed. The second layer in Figure 2 is the Business layer covering business processes of public service organization as well as big data management processes. Data management consists of two primary groups of activities, i.e., the management of organization-wide conceptual data models and the

management of organization-wide data standards. Data manager is deputed to be responsible for organization-wide coordination, focusing on the goals and plans for data quality management among responsible organization units. The third Application layer in Figure 1 compromises all applications for data processing. The traditional approach to data management is based on centralized assembling all the company data. However, big data is retained in distributed systems instead of a centralized one. As it is presented in big data architecture model, data is stored in databases, data marts, data warehouses as well as in data lakes. The last term refers to the container for raw data. Data lake is a system that stores data from a single source. However, today data lake is considered a general, enterprise-wide data repository for data from multiple sources (Quix and Hai, 2019).

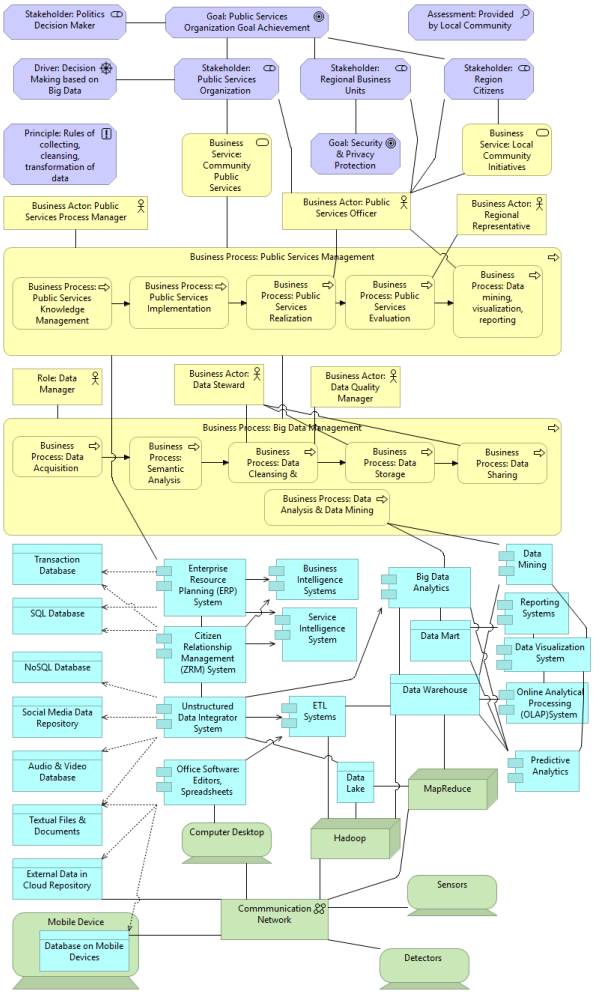


Figure 1: Big data architecture model for public service business organization

3. Big Data Balanced Scorecard

Big data governance leaders provide a framework for setting data-usage policies and controlling if information remains accurate, consistent and accessible. The strong point in their activity is to design the big data architecture model and implement it. However, this model is usually supplemented by additional elaborations, i.e., business strategy and IT strategy. Both of them are supported by a balanced scorecard (BSC), a tool proposed by Kaplan and Norton (2004).

In this paper, big data governance is understood as a developed and enforced set of processes to ensure that important data assets are formally and consistently managed across different heterogeneous platforms to ensure the required level of data quality. Melgarejo Galvan and Navarro (2017) identified problems typical for data governance. These problems particularly refer to the speed of processing, confidentiality of data, ignoring public sources, which contain unstructured, but valuable data . The other problem concerns unstandardised and inconsistent data, which should be cleaned up before the process of analysis. In general, big data governance is oriented towards the increase of processed data value. The process requires that users know what questions they want to answer and for what decisions. This knowledge permits to uncover new opportunities that impact business activities, reduce costs, and mitigate risks across all operational and financial aspects of an organization's value chain.

According to Plotkin (2014), data governance is the authority and the exercise of decision making. Simultaneously, it is considered as a system of decision rights and accountabilities for processes, executed according to agreed models. DAMA International Guide to the Data Management Body of Knowledge (Cupoli et al., 2014) defines the data governance main domains as follows: data architecture, data quality, data modeling and design, data storage and operations, data security, data warehousing and business intelligence, data integration and interoperability, and meta-data management. Also, Smallwood (2014) emphasized big data governance as a key concern in the environment of big data because of data risk management, privacy and security, business operations long-term digital preservation, and business intelligence including sub disciplines like document management, knowledge management, business continuity and disaster recovery. Big data governance can be considered as an organizational challenge, i.e., roles have to be identified, stakeholders have to be assigned to roles and responsibilities, and business processes need to be established to organize various issues around data governance. Some of these issues are already solved at the stage of big data architecture modeling and implementation. Big data governance should

be supported by appropriate techniques and tools. Beyond the above mentioned subtopics, big data governance requires defining data ownership, stewardship and protection. Anilkumar et al. (2017) emphasize meta-data and master data management, data dictionaries and standards maintenance, audit validation, and data life cycle management. Van Helvoirt and Weigand (2015) argue that master data management (MDM) adds a new value to the data, because MDM focuses on establishing integration and interoperability of heterogeneous databases and applications in a business oriented manner. Following Van Helvoirt and Weigand (2015), big data governance is presumed to be more than just achieving compliance, because it is necessary to adopt practices and principles that increase data quality and trust. A valuable data quality standard series is ISO 8000, which focuses on data characteristics and exchange in terms of vocabulary, syntax, semantics, encoding, provenance, accuracy and completeness. Standard ISO/TS 8000-1:2011 contains an introduction to ISO 8000. It covers a statement of the scope of ISO 8000, principles of data quality, the high-level data architecture of ISO 8000, a description of the ISO 8000 structure, and a summary of the content of the other parts of the general data quality series of parts of ISO 8000. Standard ISO 8000-2:2018 Data quality –Part 2: Vocabulary enables to create, collect, store, maintain, transfer, process and present data to support business processes effectively. Standard ISO 8000:150 includes a framework for data quality management. This model comprises processes, named data operations, data quality monitoring, and data quality improvement, which are in general oriented towards constant improvement of quality management. These processes are connected with particular roles, i.e., data manager, data administrator and data technician, who are responsible for process activities. According to Unhelkar (2018) quality practices in big data domain cover data profiling, cleansing and standardizing the data, applying syntax, semantics, and aesthetic checks to data, using standard architectural reference models and data patterns, controlling the business processes quality, continuous testing, and using agile techniques in developing high-quality analytics. Data cleansing consists of finding errors in data, removing unnecessary duplications, inconsistencies or incomplete values. The data is corrected by replacing values generated or deleted in the worst case (Melgarego Galvan and Navarro, 2017).

Although business analytics concerns the use of data-driven research to generate value, the big data architecture and big data governance are necessary to enable the business analytics as well as the data science research. In business organization the value is created through leveraging people, processes, data and technology. These components should be included in big data governance models. Encompassing all of these elements is the

organization culture, as a system of shared values. People are the professionals and their skills are involved in applying business analytics. Processes are a series of activities linked to achieve an outcome, like information required by decision maker. All these assets, i.e., people, processes, data and technology are applied to achieve value. As Chi (2015) notices, while collecting data is not difficult, value creation out of the data is still often questionable. Although the data intensive scientific discoveries are more and more published, and huge big data repositories are developed, business organization still have a question, if they really need the data for decision making. Therefore, this paper covers a proposal of a balanced scorecard (BSC) as a tool to support the decision making of strategic investment in big data architecture implementation. In the proposed balanced scorecard (Figure 2) four perspectives are included:

- Financial Perspective covering financial measures like return on investment (ROI), revenues, costs, and business model to increase adherence to audit and corporate responsibility, reduce time to market, and reduce complaints.
- Social Perspective concerning leadership support, professional involvement, strategic alliances and partnerships for enhancing business analytics capabilities, training and continuous learning.
- Process Perspective including descriptive, prescriptive, and predictive analytics, used to understand and solve specific problems.
- Technology Perspective conveying information in reports and dashboards, and including cross-department applications, substantial IT infrastructure, data marts, data lakes, data warehouses, Hadoop and MapReduce technology, visualization tools, cloud storage, mobile applications, and advanced machine learning tools.

As well as corporate governance, the big data governance is the organizational capacity to control the formulation and implementation of big data strategy and in this way ensure the fusion of business and big data. Van Grembergen and DeHaes (2008) argue that IT management is to be included in the IT governance process. They emphasize the role of the governance process as value creation. They assume that focus area for big data governance is driven by stakeholder values like strategic business – information technology alignment (BITA), resource management, risk management and performance management.

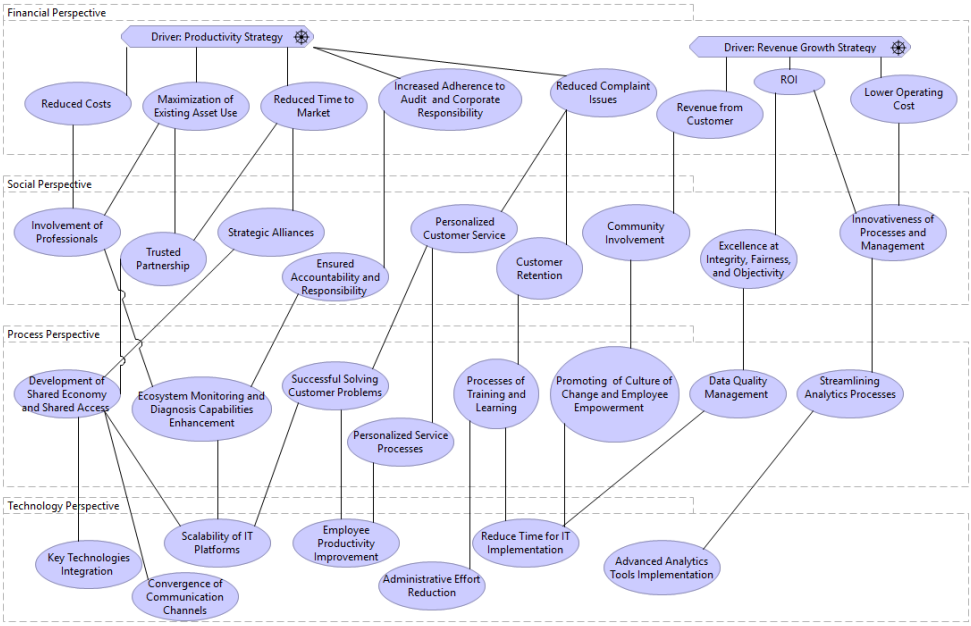


Figure 2: Big Data Balanced Scorecard Perspective Objectives

Balanced scorecard is defined to balance short term cost savings with long term performance, optimally match big data resources to business requirements, manage big data security and risk to business (Kaplan and Norton, 2004). Taking into account that BSC focuses on the value of assets and activities provided, as well as on eliminating non-value adding activities and processes, in Figure 2 the proposed values for big data governance are presented. The proposals are to emphasize what seems to be extremely important from each of the particular perspectives. The proposed valuable objectives can be linked among perspectives and in this way the compatibility of perspectives can be presented. Therefore, the balanced scorecard offers a framework for the development of big data implementation strategies for creating values. Proposed balanced scorecard is offered as a referential model, which is expected to encourage public services business organization to consider its decision on the big data architecture development, investment and implementation. If the proposed objectives can be achieved then the big data architecture should be successfully implemented.

4. Conclusion

Stimmel (2015) argues that although analyzing big data ensures advantages for many business organizations, providing effective governance is not an easy task. In the paper, big data governance was considered as a discipline, not a particular activity. This discipline is

value oriented and the balanced scorecard could be applied as a referential model to support that discipline decision making.

As proposed by Kaplan and Norton (2004) the BSC can be applied when a value creation is not direct. Intangible assets such as knowledge, big data, business analytics have an indirect impact on financial and social outcomes such as increased productivity, revenues, lowered costs and time to markets, and higher profits and customer satisfaction. Although values are contextual, referential models can be considered as helpful in that values' contextualization for each particular case.

References

Anilkumar, R., Deshmukh, R.R., Emmanuel, M. (2017) Big Data Predictive Analysis for Detection of Prostate Cancer on Cloud-Based Platform: Microsoft Azure, Privacy and Security Policies. In Tamane, S., Kumar Solanki, V., Dey, N. (eds.) *Privacy and Security Policies in Big Data*. IGI Global, Hershey, 259-278

Azarmi, B. (2016) *Scalable Big Data Architecture, A practitioner's guide to choosing relevant big data architecture*. Springer NY.

Belcastro, L., Marozzo, F., Talia, D., Trunfio, P. (2017) Big Data Analysis on Clouds. In Zomaya A.Y., Sakr S. (eds.) *Handbook of Big Data Technologies*. Springer Cham, 101-142.

Chen, M., Mao, S., Zhang, Y., Leung, V.M. (2014) *Big Data, Related technologies, challenges and future prospects*. Springer, Cham Heidelberg.

Chi, C-H.(2015) Behaviour Informatics: Capturing Value Creation in the Era of Big Data. In Intan, R., Chi, C-H., Palit, H.N., Santoso, L.W. (eds.) *Intelligence in the Era of Big Data*. Springer Verlag Berlin, XIV-XVI.

Cupoli, P., Earley, S., Henderson, D. (2014) *DAMA -DMBOK2 Framework, The Data Management Association*. https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf. Accessed July 12, 2019.

Dong, X.L., Srivastava, D. (2015) *Big Data Integration.* Morgan & Claypool Publishers, Waterloo.

Gupta, M., Singla, N. (2017) Evolution of Cloud in Big Data With Hadoop on Docker Platform. In Tamane, S., Kumar Solanki, V., Dey, N. (eds.) *Privacy and Security Policies in Big Data*. IGI Global, Hershey, 41-64.

Heisterberg, R., Verma, A.(2014) *Creating Business Agility, How Convergence of Cloud, Social, Mobile, Video, and Big Data Enables Competitive Advantage*. Wiley, Hoboken.

ISO/TS 8000-1:2011, Data quality –Part 1:Overview, https://www.iso.org/standard/50798.html. Accessed July 12, 2019.

ISO/TS 8000:150:2011, Data quality –Part 150: Master data: Quality management framework, https://www.iso.org/standard/54579.html. Accessed July 12, 2019.

ISO 8000-2:2018 Data quality –Part 2: Vocabulary, https://www.iso.org/standard/76563.html. Accessed July 12, 2019.

Kaplan, R.S., Norton, D.P. (2004) *Strategy Maps, converting intangible assets into tangible outcomes*. Harvard Business School Press, Boston.

Krishnan, K.(2013) *Data Warehousing in the age of Big Data*. Morgan Kaufmann, Elsevier, Amsterdam.

Melgarejo Galvan, A.R., Rocio Clavo Navarro, K. (2017) Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies. In Lossio-Ventura, J.A., Alatrista-Salas, H. (eds.) *Information Management and Big Data*. Springer Heidelberg, 120-133.

Morabito, V. (2015) *Big Data and Analytics, Strategic and Organizational Impacts.* Springer Cham.

Plotkin, D. (2014) *Data Stewardship, An Actionable Guide to Effective Data Management and Data Governance*. Elsevier, Amsterdam.

Quix, Ch., Hai, R. (2019) *Data Lake*. In Sakr, S., Zomaya, A.Y (eds.) *Encyclopedia of Big Data Technologies*. Springer Nature, Cham, 552-559.

Schmarzo, B. (2013) *Big Data, Understanding How Data Powers Big Business.* Wiley, Indianapolis.

Smallwood, R.F. (2014) *Information Governance*. John Wiley and Sons, Hoboken.

Stimmel, C.L. (2015) *Big Data Analytics Strategies for the Smart Grid*. CRC Press, Taylor & Francis Group, London.

Unhelkar, B. (2018) *Big Data Strategies for Agile Business, Framework, Practices and Transformation Roadmap*. CRC Press, Boca Raton, London.

Van Grembergen, W., DeHaes, S. (2008) *Implementing Information Technology Governance, Models, Practices, and Cases*. IGI Publishing, Hershey, New York.

van Helvoirt, S., Weigand, H. (2015) Operationalizing Data Governance via Multi-level Metadata Management. In Janssen, M., Mäntymäki, M., Hidders, J., Klievink, B., Lamersdorf, W., van Loenen, B., Zuiderwijk, A. (eds.) *Open and Big Data Management and Innovation*. Springer, Cham, 160-172.

CHAPTER 6

A CORE PROBLEM WITH HUMAN DATA PROCESSING: EPISTEMIC CIRCULARITY IN ACTION

Mehmet Selim DERİNDERE*

*İstanbul University, Informatics Department, İstanbul, Turkiye
E-mail: mehmetderindere@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.06

Abstract

Managers are expected to solve critical problems of our society in an efficient manner and in ways so that the problems remain solved. In order to accomplish this, the managers are provided with vast amounts of resources including mountains of data and a wide variety of problem-solving methods available. On the other hand, the effectiveness of social and organizational problem solving is far from satisfactory and this lack of effectiveness is ubiquitous. One reason for this ineffectiveness we claim has to do with how the human mind works. The inherent capabilities and limitations of human mind coupled with social-cognitive skills lead to sub-par problem-solving. An especially counterproductive problem-solving approach used by managers is setting and attempting to solve problems using erroneous cognitive skills that not only fails to include relevant data but also uses the existing data in a counterproductive manner. The very data processing skills of managers make problem-solving a dead end for the actors involved at great cost to them and to the society. This chapter looks at a core human data processing problem that renders the available data and techniques ineffective. Epistemic Circularity disregards all the disconfirming or threatening data and fails to include it in the problem solution. Epistemic Circularity thus renders the relevant data useless in developing effective solutions. Easy knowledge, a product of epistemic circularity, leads to ineffective problem solving which in many cases result in exacerbated problems and counterproductive consequences.

Keywords: Epistemic Circularity, Managerial problem solving, Difficult problems, Data dismissal, Knowledge loss

Introduction

All human beings require information about their environment in order to survive and flourish. Norbert Wiener wrote "the world may be viewed as a myriad of To Whom It May Concern messages" (in Watzlawick, 1977). These messages construe the data that can be harvested for constructing theories that explain how the world works and how we should act in order to survive and flourish in this world. One of the main goals of our cognitive mechanisms is to maintain an accurate representation of our environment, at least the relevant aspects of it (Sperber & Mercier, 2017). The data chosen for constructing, testing and utilizing our worldviews in turn depends on the theories we have in our minds. As Einstein stated, "it is the theory which decides what we can observe" (in Heisenberg, 1972). Thus, there is a continuous circular process in action. In this process, data is used to feed, test, affirm or disconfirm our theories and these theories are used to construct our behavioral worlds as well as decide which data is chosen and how it is utilized.

In order to act effectively and solve problems in persistent and productive ways this circular reasoning process itself must be productive. A productive process uses hard, valid data and makes explicit inferences. Premises in a productive process are explicit and the conclusions are publicly testable (Argyris, 1991). Using productive reasoning processes is especially critical when working with social world problems characterized by their uniqueness, uncertainty and instability (Friedman, 2001) where the available data is mostly ambiguous and scattered and information is mostly vague, inconsistent and incongruent (Argyris & Schön, 1978).

Paradoxically, in practice, the main role of reasoning is not to motivate us in reaching grounded conclusions but to explain and justify existing or after the fact conclusions and beliefs (Sperber & Mercier, 2017,p.121). This paradox is caused by the limited information processing capacity (Simon, 1994) as well as the cognitive and social skills of human beings (Argyris, 1993).

Information processing capacities of the human mind makes it impossible for a single mind to gather, process and utilize all the required data (Knight, 2014). Based on empirical studies Simon (1971) asserts that the human mind constructs a reality with what limited information it gathers. On the other hand information systems used to manage gather and present information enough to override human data processing capabilities (Ackoff, 1967; Eppler & Mengis, 2004). This data flood swamps the decision-maker and reduces the likelihood that the managers will make informed decisions (Argyris, 1987).

One way managers reduce the likelihood of making good use of data in making informed decisions is when they use reasoning processes that dismiss data that can potentially challenge or disconfirm their erroneous assumptions and perspectives regarding the situation at hand. Kahneman illustrates that even if actors have access to information that contradicts with their beliefs, they can elect to disregard such information (Kahneman, 2011). Epistemic Circularity is a good example of how managers reason in ways that inhibits effective problem solving.

After providing a definition and example for Epistemic Circularity, I will illustrate how Epistemic Circularity leads to erroneous Easy Knowledge used by managers to solve the problems in ways that fail to solve the problems effectively and hide the source of ineffectiveness.

1. Epistemological Circularity

Epistemic Circularity is a kind of cognitive bias (fallacy) in which the reliability of a source of belief or conclusion relies on premises that are themselves based on that source (Lammenranta, 2006). When the reliability of a belief-source such as perception, intuitive reason, introspection, memory or reasoning is established based on the very beliefs produced by the same belief-source rather than the belief is said to be epistemologically circular (Bondy & Delaplante, 2011).

One example of epistemically circular arguments is a track-record argument as in follows (ibid). Suppose S1 is a belief source such as perception, reasoning or memory and an actor wants to find reasons for the reliability of his or her belief-source S1 using the following inductive argument.

S1 formed a perceptual belief p1 at t1 (a given time) and p1 is true

S1 formed a perceptual belief p2 at t2 and p2 is true

S1 formed a perceptual belief p3 at t3 and p3 is true

S1 formed a perceptual belief pn at tn and pn is true

Therefore, my belief source S1 is a reliable source of belief

Since the premises (p1..pn) formed by the use of the very faculty whose reliability the actor is trying to establish, the argument is circular. In an epistemically circular argument, the reliability of a source of belief is defended by relying on premises that are themselves based on the very same source (ibid.).

Coined by William Alston (1986), Epistemic Circularity points out that the basic sources of beliefs such as perception, introspection, intuitive reasoning, memory and reasoning are

reliable except by using epistemically circular arguments. Contemporary accounts of knowledge and justification that enables us to gain knowledge and justified beliefs are based on such arguments. Using epistemic circularity allows us to know the premises of an argument which is epistemically circular even without knowing the conclusion and using the argument leads to the knowledge of the conclusion (Lammenranta, 2006).

Bergman provides the following Juror case as an example epistemic circularity:

> *Juror #1: You know that witness named Hank? I have doubts about his trustworthiness.*
>
> *Juror #2: Well perhaps this will help you. Yesterday I overheard Hank claiming to be a trustworthy witness.*
>
> *Juror #1: So Hank claimed to be trustworthy did he? Well, that settles it then. I'm now convinced that Hank is trustworthy (Bergmann, 2006, p.180).*

In this case, the Juror #1 takes Hank's claim that he is reliable as evidence of Hank's reliability. However, without an independent reason and publicly verifiable evidence that leads us to believe Hank is reliable there is no reason to trust any of Hank's claims let alone his claim that he is reliable (Bondy & Delaplante, 2011).

Knowledge acquired through epistemic circularity is called "easy knowledge." It is the knowledge that is produced using processes whose reliability is justified with the process itself. In the example above, Juror #1's belief that Hank is a reliable witness is easy knowledge.

Leaving philosophical discussion regarding problems with easy knowledge such as its dialectical ineffectiveness to others (Jenkins, 2011; Schmitt, 2004; Van Cleve, 2003; Vogel, 1987), here will illustrate some practical problems in everyday organizational life.

2. A Managerial Problem-Solving Case

This actual case is about a manager who had tried to correct the problematic attitude of his subordinate engineer numerous times. The manager frames the problem as the subordinate being lazy and irresponsible and had several attempts to rectify the situation. Frustrated with his inability to "set him straight" this case illustrates his one last attempt to solve the problem.

The case is written using the left-hand case method developed by Chris Argyris and his colleagues (Argyris, Putnam, & Smith, 1985; Argyris & Schön, 1978). The case captures how the actor frames a problematic situation, what is his strategy to go about solving the problem and the actual dialogue as well as the thoughts and feelings during the solution attempt.

When writing the case, the manager provides an explanation of the problem (Question1) and the strategy he or she used to solve the problem (Q2). Then as in a script, he writes the actual dialogue (Q3) on the right-hand side column. The thoughts and feelings that are not communicated during the dialogue are written on the left-hand column. Left hand column provides data about the reasoning and self-censorship processes during the actual dialogue.

Table 1: Left-hand case. An attempt to solve a problem

Q1: In one paragraph, provide some context for the problem. What is the problem? Who is the other actor? (1) One of my subordinate engineers (S) has been very lax in his attitude. (2) He is an irresponsible guy who always neglects his duties. (3) His irresponsibility is affecting all the team in a negative way. I am sick of his attitude.	
Q2: What is your strategy to resolve a problem? What do you want to achieve while dealing with this problem? (4) I talked to him dozens of times and tried to get him to act responsibly. I clearly explain his duties and how important it is for all of us that he shoulders his share of responsibility. (5)Whenever I talk to him, I try to clarify his responsibilities and get him to act. (6) He always says he will get his act together, but he never delivers.	
Q3: What happened when you spoke with the others?	
My uncommunicated thoughts and feelings	**The conversation**
(7) You are just avoiding my question.	**Manager (M):** (8) Why didn't you fix the (clients) problem equipment? **Engineer (E):** (9) His equipment is fine.
(10) You are still dodging my question.	**M:** (11) But, they still complain. **E:** (12) They are always like that; complain just for the sake of it.
(13) You did not even see the equipment.	**M:** (14) But I confirm that they have a problem, didn't you see the equipment? **E:** (15) According to their claim
(16) I am going to blow my top	**M:** (interrupting him**):** Hold on! Hold on! Hold on! (17) Their claim?!
(20) Oh my God!	(18) Have you ever seen the equipment? **E:** (19) I was about to go to their site. **M:** (21) You are joking right?
(23) I knew you would say that.	**E:** (22) My other work has just finished **E:** (24) Which one? **O:** (25) (the other clients)
(27) You never do it right.	**M:** (26) You didn't finish that until now? **E:** (28) It was a greater problem that we anticipated. **M:** (29) Yes, It is always a bigger problem with you. **E:** (30) You know; you can help me. **M:** (31) Sorry, cannot do that. **E:** (32) Why not? **M:** (33) It is not my job, it is your job. **E:** (35) Okay, there's no need to shout at me. I am going as we planned. **M:** (36) I am coming with you.
(34) I would help you if I could ever believe that you are sincere and serious.	**E:** Look at that! (37) The equipment is configured incorrectly. **M:** (38) See, (the client) could not use critical equipment for such a ridiculous reason. **E:** (39) Next time, I will be faster.
(40) I wish, I only wish	

3. Analysis

3.1. Ladder of Inference

When analyzing the case data, the ladder of inference is used. The ladder of inference is a conceptual tool that schematically represents how human beings select data from their environment and make inferences (Argyris et al., 1985, p.57). The process of selecting data and imposing meaning to the selected data is mostly an automatic process and works beyond the consciousness (Senge, 1990).

The first rung of the ladder consists of the relatively directly observable data which can be recorded with a video or tape recorder. In our research, these are the recorded conversations as well as the written documents. This is "hard data" because regardless of what they believe, parties can agree on what the data is. The second rung represents the inferences made about the meanings embedded in the words. This inference process occurs often in milliseconds regardless of whether they agree with the meanings. Then people impose their own meanings on the actions that they believe the other person intends (Argyris, 1993, p.57). People make attributions which are causal explanations about others' intentions or goals. Or they make evaluations of the effectiveness of the behavior.

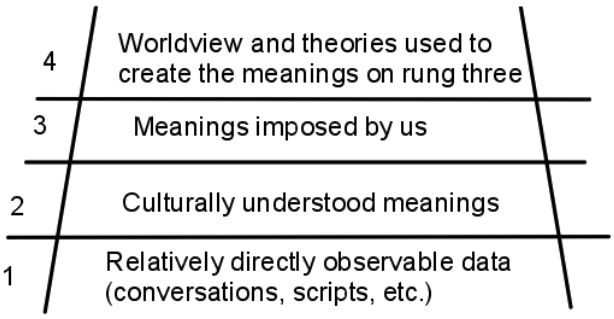


Figure 2. Ladder of Inference. Adopted from Argyris (1993)

In the next step (rung 3) the actor imposes his own meaning on the actions he believes the other person intended. He attributes intentions to the other person or evaluates his actions or intentions against his own -mostly tacit- criteria.

When going up on the ladder of inference the inferences, meanings, judgements and conclusions get more abstract and brief. A great amount of data that can be interpreted in a number of ways is summarized in service of brief conclusions. This requires testing the conclusions and inferences in each step with the other party. Otherwise the further steps will be wrong or untested conclusions. Using such an approach is enacting the basic scientific rule

"beginning with data independent of the observer and testing all inferences" (Burns & Okey, 1985; Lederman, Schwartz, Abd-El-Khalick, & Bell, 2002) in interpersonal, social contexts.

The case in this chapter provides many examples of the usage of the ladder of inference during the interaction. For example, the engineer's answers to the superior's question (sentence 9: "his computer is fine") is the hard data of the case and is on rung 1. The superior imposes his own meaning to this answer and thinks "instead of addressing my concern he is talking about the equipment." Culturally, replying with some other thing instead of answering the question can be understood as avoiding a subject (rung 2). When the engineer answers the uttered question (sentence 8) instead of answering the real and unuttered question (which is his irresponsibility) the manager imposes the meaning "trying to dodge my question" (sentence 7) (rung 3). Then the manager causally explains all the engineer's actions with his attitude and character and evaluates him as lax and irresponsible.

3.2. Analyzing the Data and Reasoning Process

In this problem-solving case, the main source of data and information of the researcher is the manager. Since both the researcher and the manager are subject to distortions of data and a number of biases (Robinson & Lai, 2006) it is critical that data and information is checked independent of the reasoning of the provider. The hard data in this case is the utterances and quoted thoughts and feelings of the clients.

In the analysis of the case, the focus consisted of three initial questions. Does the client's reasoning process contain all the relevant data provided by the client himself? Does the data explain what the client claims is happening? If the answer to first and second questions are negative, then what are the consequences?

One important set of information the researcher surfaced during the interview with the manager is related to the history of the problem. The following excerpts are the attempts by the researcher to reveal more related data.

Table 2: An excerpt from the interaction between the researcher and the manager	
Researcher (R): (in our talk) you stated that this problem has been going on for some time and you said that you attempted to get him straight several times. Do I remember correctly?	Checks if the data he recalls is true.
Client (M): Yes. I talked to him maybe dozens of times.	
R: Is it fair to say that this (written) case is an example of those "dozens of times" of your talk?	Confirms the data
M: Certainly. There are some little differences, but it is all the same. He never gets his act together.	Checks whether his inference is valid.
R: Do you remember the first time you talked to him? Was it in any way different from this (written case)?	Confirms the data.
M: No, no. It was related to his irresponsibility again. I never saw him taking charge of his job.	
R: What I am trying to find out is how did you initially reach the conclusion that this guy has a "lax attitude" and is "irresponsible?"	Explains his intention. Tries to get at the data that constitutes the basis for the manager's inferences
M: Believe me with this guy it is always like this. It was, again a similar incident like this. The client complained that there was a problem and it was related to him. It is this guy that is the problem.	Does not provide the data. Repeats his initial belief.

The client claims that the problem is the engineer himself. That is, he explains the problems through attributions about the personality or attitude of the engineer. The following questions seeks data that would disconfirm this claim by through data indicating that other engineers are also prone to similar problems.

Table 3: An excerpt from the interaction between the researcher and the manager	
R: Do other engineers, if you remember, have similar problems with the clients similar to this incident? And I am asking this question because I am trying to understand what is it that he does or says that leads you to believe that the problem is "in" him and not related to some other circumstances.	*Asks for data* *Shares his intent.*
M: Of course. Ours is a sensitive operation and since we are building custom solutions to changing requirements, we have many similar cases. It is part of the job.	*Provides data.*
R: If similar cases happen with other engineers, what makes this engineer in your view different? If the problems (as shown in this case) are common occurrences with other engineers, what leads you to conclude that this particular engineer is "irresponsible?"	*Asks for data*
M: Believe me he is irresponsible. I know this guy and I talked to him dozens of times. He is plain lazy and just won't change.	*Provides self-referential conclusions.*

4. Epistemic Circularity in the Manager's Reasoning and Actions

As revealed during the interview, whatever event happened led the manager to conclude that this particular subordinate is lazy and irresponsible also happens with other subordinates, but he does not label them as lazy and irresponsible. The directly observable data that is available from both cases and the recordings can be interpreted in ways that show the engineer is responsible or at least doing what he can, given the situation, in order to fulfill his duties or

having the same problems as other engineers. So far, the manager has not provided any conclusive data that can lead others to conclude that this specific engineer is irresponsible or lazy in an independently verifiable way. However, this lack of supporting data and occurrences of similar cases with other engineers that indicates the problems are not exclusive to this particular engineer does not lead the manager to refute or review his beliefs about the engineer. It seems that the only supporting evidence of his beliefs are his beliefs. In this case it can be concluded that the manager's reasoning is epistemically circular.

It can be hypothesized that the reasoning process is somehow triggered by some action of the subordinate, but the manager seems unaware of what particular set of action triggers it. In addition, he is unaware of the reasoning process that leads him to his current conclusions and actions. Here is a manager who has a strong belief explaining the subordinate's behavior while being detached from his own reasoning process upon which the belief is based on. Bavelas's research shows that once a tentative explanation has taken hold on one's mind, information to the contrary does not lead to corrections but to more elaborate explanation of the conclusions (in Watzlawick, 1977). This is an expected consequence of epistemic circularity.

Based on the interaction between the manager and the engineer, a partial map of the manager's epistemically circular reasoning can be drawn.

Table 4: A partial map of the ongoing problem

How the Manager thinks	How the Manager acts.	Consequences
Manager believes that the engineer is irresponsible and lazy. Manager believes that the engineer doesn't understand that he is the problem. Manager believes that he should set the engineer straight. The basis for the manager's beliefs is that he believes that the engineer is irresponsible. The manager's previous attempts (which are same as this one) prove that he is right.	He acts based on his beliefs. The manager doesn't explain how he reasons about the situation. Hides his beliefs. Acts as if he is not doing so. He asks questions and uses the engineer's answer to confirm his (unrevealed) conclusion that the engineer is irresponsible.	The engineer has no access to the manager's reasoning and knowledge to take informed action. The errors continue to exist. The same situation repeats itself. Manager's beliefs are reinforced. Manager's frustration accumulates.

Based on the data regarding the interaction between the manager and the engineer the following hypothesis are set. These are similar to hypothesis used in existing research (Schön & Argyris, 1992):

1. As the manager tries to control the actions of the subordinate and control and censor the feelings of the subordinate then subordinate will possibly feel controlled or he is in a win/lose situation.
2. As the manager tries to hide his feelings and thoughts unilaterally in order to save face but fails to hide them then the subordinate will probably feel that the manager is hiding something and acting defensively.
3. As the manager controls the situation and the task then the subordinate will feel little freedom of choice and internal commitment.
4. The manager is going to claim that the subordinate does not understand the situation or that the subordinate understands what he (the manager) is trying to say but acts as if he did not understand.
5. If the self-feeding, self-reinforcing vicious circle keeps going, then the accumulated stress and frustration will possibly lead to the termination of the engineer's job.

Differing from the managers reasoning that leads him to believe that the engineer "is" the problem, I believe that in addition to the reasoning processes of the manager (and engineer), this pattern of interaction is causally responsible for the counterproductive problem-solving process. For further discussion one can refer to (Argyris, 1993; Dent, 2003; Friedman, 2001; Watzlawick, Beavin, & Jackson, 1967; Watzlawick, Weakland, & Fisch, 2011).

5. Consequences of Epistemically Circular Reasoning and Action

An important part of the analysis is finding out the consequences of the reasoning of the participants and the interactions for the problem. The consequences may be intended or unintended by the client. If there is a mismatch between the intended consequences and the actual consequences, this signifies an outcome gap (Rudolph, Taylor, & Foldy, 2001)

5.1. Consequences for the Manager

There are some interesting consequences for the reasoning of the manager. Firstly, the stated goal of the meeting in the case is to get the subordinate to understand the "situation" and take responsibility. However, neither this meeting nor the previous ones achieved this

goal. This is a fundamental ineffectiveness in the core organizational processes which is managerial problem solving (Mintzberg, 2005)

Secondly, the manager feels frustrated, but he tries to suppress his negative evaluations and feelings about the engineer. This makes it difficult for him to genuinely test these evaluations and the sources of feelings because in order to genuinely test these he has to reveal and discuss them and make them testable. A genuine test is one that can actually disconfirm these ideas (Popper, 2002).

Argyris observes that one of the most frequent strategies used when trying to tell the other of his poor performance is to ease-in (Argyris, 2010, p.44). Easing-in is a strategy used when attempting to avoid the threat associated with direct confrontations that have the possibility of escalating. The actor using easing-in strategy tries to get the other to understand what he is actually trying to get at. The actor using the strategy is tacitly saying, "I want you to understand something but if I say it directly, I am afraid that the problem will escalate. So, I will ask you some questions and if you give the 'right' answers, you will understand what I am hiding." Argyris (ibid.) claims that this strategy allows the principal actor to show concern to the other. However, in doing so he covers up his actual views, acts as if he is not doing so and creates an interaction based on the rule "threatening issues as well as negative feelings shall not be addressed openly!" The subordinate also has to cover up his actual feelings and acts as if not doing so. This ensures that the erroneous reasoning processes and conclusions and assumptions are not checked.

Such a strategy contains several risks. Firstly, the strategy relies on the engineer to be willing to give the "right answers." When the subordinate fails to provide the right answers, the manager gets frustrated and angry for both the initial problem and for his lack of understanding.

Another critical risk is by using this strategy: the manager hides what he is trying to achieve and acts as if he is not hiding anything. Even though the engineer understands that there is something hidden going on, he will not be willing to discuss something that his boss made undiscussable.

5.2. Consequences for the Engineer

Based on the data, it is possible to predict that the subordinate will fail to take responsibility because he does not have access to the "actual" problem since the manager hides it and makes it undiscussable. Now the engineer is in a "double bind" (P. Watzlawick et al., 1967).

If he takes the correct action, it is because the manager pushes him not because he takes responsibility. If he keeps doing what he has been doing then he causes problems to himself and to his manager. In such a situation there is no way the engineer can prove that he can act responsibly. The only way he can get escape this double bind is for either the engineer or the manager to make what the manager made undiscussable discussable.

5.3. Consequences for the organization

In order to be effective, the problem-solving efforts should be based on valid data and sound processes (Hasenfeld & Furman, 1994). The organizational problem examined in the case above is based on ambiguous data and fuzzy logic. Furthermore, the manager's solution attempts until this date not only failed to solve the problem but also aggravated the problem by reinforcing his beliefs and increasing his frustration. As the problem situation repeats itself the manager has the more "evidence" confirming his beliefs thus his beliefs about the engineer are reinforced. But as we have seen these beliefs are based on erroneous reasoning and fuzzy data. This results in a "self-sealing" process. Self-sealing processes are "conjectures that cannot be refuted" (Paul Watzlawick, 1977). But Popper has shown that refutability is one indispensable condition for scientific explanation (Popper, 2002) and sound reasoning. In such a self-sealing process which uses epistemically circular reasoning, the assumptions and beliefs of the manager regarding the problem are reinforced no matter what.

As illustrated in the case, the engineer has no way of refuting the assumptions and beliefs of the manager because they are either hidden from the engineer. The manager either does not reveal his attributions about the engineer by keeping these attributions on the left-hand column or he communicates them in ways that make them undiscussable and the undiscussability undiscussable (Argyris, 1988). No matter what the engineer does he will fail to change the manager's view.

For example, the manager thinks that he would help the engineer if he believed that he was sincere and serious (34). But by keeping this in his left-hand column private he covers it up and act as if he is not doing so. The engineer do not have access to this thought. Even if he had access he had to learn how he could prove that he is sincere and serious. Thus they are trapped in circular process of their own making (Argyris, 2010).

The manager in this case is trying to solve a problem, which in the given context, is very real and has dire consequences for the organization. In his interactions with this engineer and others with whom he is having similar problems, the manager is affected by the accumulated stress. Whenever a similar problem comes up the issue deepens. Usually such problems are never solved until one of the parties has had enough.

6. Conclusion

Managers trying to solve difficult problems containing ambiguous data which actually or potentially refutes their conclusions, usually neglect or disregard such data and use knowledge based on epistemic circularity. This results in getting trapped in misdiagnosed problems that has great cost it terms of time, energy and focus (Senge, 1990). These traps decrease the quality of life in the organization as the mental effort of organization members is spent on these problems (Schwarz, 2013). The behavioral world around difficult problems in an organization with such traps turns into a swamp (Razer & Friedman, 2016). The relevant data is dismissed and lost in this swamp.

Epistemically circular processes swallow significant amount of data and lead to organizational vicious cycles. In order to be able to access, contain, process and utilize all available data, managers need to obtain the required socio-cognitive skills that will enable them to critically look and reflect on their reasoning and problem-solving processes. If utilized, the data wasted in those vicious cycles can play a critical role in developing a rich understanding and overcoming the difficult and complex organizational and social problems. Managers who learn the skill of overcoming circularity and develop productive data processes can turn difficult and disconcerting data into productive resources.

Research shows that while it is difficult to change the way one reasons or sees the world, it is not impossible (Arieli, Friedman, & Agbaria, 2009; Le Fevre & Robinson, 2015; Wolfberg & Dixon, 2000; Wright, 1962). In order to better utilize the available data and information the actors must first become aware of how their reasoning works and be able to "see" and contain the data they have been dismissing. Becoming aware of this previously dismissed data, that potentially or actually disconfirm their "easy knowledge", opens the path for producing more valid information regarding the reality and the situation thus enabling effective and responsible action.

There are a number of approaches that prove useful. Argyris intervened in organizational systems to get the actors to see the symmetrical inconsistencies and injustices in their reasoning and actions (Argyris, 1982, 2004, 2007). Razer and Friedman surfaces the emotional component that leads to exclusive relations and blindness towards disconfirming information by attempting to provide alternative interpretations of the actual actions (Razer & Friedman, 2016). Such alternative interpretations may enable the actor an emotional relief and open the possibility of adopting a more productive stand. Rudolph et al. teaches the actors to see that reality is not an objective truth but a constructed phenomenon, that they are causally responsible for the

consequences in this constructed behavioral world and if they can reflect and change their 'frames' and actions they can alter the consequences they produce (Rudolph et al., 2001). Robinson helped educators improve their practice by becoming aware of their tacit constraint sets and develop more coherent and effective practices (Robinson and Lai,2006).

Watzlawick once claimed that ''The Situation is Hopeless but not Serious'' (Watzlawick, 1993). When observing efforts that produce actionable knowledge which enables productive change in social systems based on valid data, I came to believe that the situation is serious but not hopeless.

References

Ackoff, R. L. (1967). Management misinformation systems. *Management Science*, *14*(4), 147–156.

Argyris, C. (1982). *Reasoning, Learning and Action: Individual and Organization*. San Francisco: Jossey-Bass.

Argyris, C. (1987). Bridging economics and psychology: The case of the economic theory of the firm. *American Psychologist*, *42*(5), 456–463.

Argyris, C. (1988). Crafting a Theory of Practice: The Case of Organizational Paradoxes. In *Paradox and transformation: Toward a theory of change in organization and management* (pp. 137–162). Cambridge , MA: Ballinger.

Argyris, C. (1991). Teaching Smart People How to Learn. *Harvard Business Review*, *69*(3), 4–15.

Argyris, C. (1993). *Knowledge for Action A Guide to Overcoming Barriers to Organizational Change*. San Francisco: Jossey-Bass.

Argyris, C. (2004). *Reasons and Rationalizations*. New York: Oxford University Press.

Argyris, C. (2007). Double Loop Learning in Organizations: A Theory of Action View. In K. G. Smith & M. A. Hitt (Eds.), *Great Minds in Management: The Process of Theory Development* (1st Editio, p. 624). Boston: Oxford University Press.

Argyris, C. (2010). Organizational Traps: Leadership, Culture, Organizational Design. In *Organizational Traps: Leadership, Culture, Organizational Design*. Oxford: Oxford University Press.

Argyris, C., Putnam, R., & Smith, D. M. (1985). *Action Science* (1st Editio). New York: Jossey-Bass, Inc.

Argyris, C., & Schön, D. A. (1978). *Organizational Learning : A Theory of Action Perspective*. New York: Addison-Wesley Publishing.

Arieli, D., Friedman, V. J., & Agbaria, K. (2009). The paradox of participation in action research. *Action Research*, *7*(3), 263–290.

Bergmann, M. (2006). *Justification Without Awareness: A Defense of Epistemic Externalism*. Oxford: Clarendon Press.

Bondy, P., & Delaplante, K. (2011). Against epistemic circularity. In *OSSA Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation* (pp. 1–8). Windsor.

Burns, J. C., & Okey, J. R. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching*, *22*(2), 167–177.

Dent, E. B. (2003). The Interactional Model: an Alternative To the Direct Cause and Effect Construct for Mutually Causal Organizational Phenomena. *Foundations of Science*, *8*, 295–314. https://doi.org/10.1023/A

Eppler, M. J., & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society*, *20*(5), 325–344.

Friedman, V. J. (2001). Action science: Creating communities of inquiry in communities of practice. *Handbook of Action Research, Thousand Oaks: Sage*, 131–143.

Hasenfeld, Y., & Furman, W. (1994). Intervention research as an interorganizational exchange. In *Intervention research: Design and development for human service* (pp. 297–311).

Heisenberg, W. (1972). *Physics and beyond : encounters and conversations*. New York: Harper & Row.

Jenkins, C. S. I. (2011). Reflective Knowledge and Epistemic Circularity. *Philosophical Papers*, *40*(3), 305–325.

Knight, F. H. (2014). *Risk, Uncertainty and Profit* (2014 Repri). Cambridge: Pantianos Classics.

Le Fevre, D. M., & Robinson, V. (2015). The Interpersonal Challenges of Instructional Leadership. *Educational Administration Quarterly*, *51*(1), 58–95.

Lederman, N. G., Schwartz, R. S., Abd-El-Khalick, F., & Bell, R. L. (2002). Preservice teachers' understanding and teaching of the nature of science: An intervention study. *The Canadian Journal of Science, Mathematics, and Technology Education*, *1*(2), 135–160.

Mintzberg, H. (2005). *Managers Not MBAs: A Hard Look at the Soft Practice of Managing and Management Development*. San Francisco: Berrett-Koehler Publishers.

Popper, K. R. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge* (2nd Editio).

Razer, M., & Friedman, V. J. (2016). *From Exclusion to Excellence: Building Restorative Relationships to Create Inclusive Schools*. Rotterdam: Sense Publishers.

Robinson, V., & Lai, M. K. (2006). *Practitioner Research for Educators*. California: Corwin Press.

Rudolph, J. W., Taylor, S. S., & Foldy, E. G. (2001). Collaborative Off-Line Reflection: a Way to Develop Skill in Action Science and Action Inquiry. In P. Reason & H. Bradbury (Eds.), *Handbook of Action Research: Participative Inquiry and Practice* (pp. 405–412). London: Sage Publications.

Schmitt, F. F. S. (2004). What Is Wrong with Epistemic Circularity? In *Philosophical Issues* (Vol. 14, pp. 379–402).

Schwarz, R. (2013). Smart Leaders, Smarter Teams: how you and your team get unstuck to get results. San Francisco: Jossey-Bass.

Schön, D. A., & Argyris, C. (1992). *Theory in Practice: Increasing Professional Effectiveness* (Revised Ed). San Francisco, CA: Jossey-Bass.

Senge, P. (1990). *The fifth discipline. The art and practice of the learning organization.* New York: Doubleday.

Simon, H. A. (1971). Designing organizations for an information-rich world. In Greenberger (Ed.), *Computers, Communications and the Public Interest*. Baltimore: John Hopkins University Press.

Simon, H. A. (1994). Bottleneck of Attention: Connecting Thought with Motivation. In W. D. Spaulding (Ed.), *Integrative Views of Motivation, Cognition and Emotion* (pp. 1–22). Lincoln, NE: University of Nebraska Press.

Sperber, D., & Mercier, H. (2017). *The Enigma of Reason: A New Theory of Human Understanding*. Penguin UK.

Van Cleve, J. (2003). Is Knowledge Easy--or Impossible? Externalism as the Only Alternative to Skepticism. In S. Luper (Ed.), *The Skeptics: Contemporary Essays*. Hampshire: Ashgate.

Vogel, J. (1987). Tracking, Closure, and Inductive Knowledge. In S. Luper-Foy (Ed.), *The Possibility of Knowledge: Nozick and His Critics* (pp. 197–215). Lanham: Rowman & Littlefield.

Watzlawick, P., Beavin, J., & Jackson, D. D. (1967). *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes*. New York: Norton.

Watzlawick, Paul. (1977). *How real is real? Confusion, disinformation, communication*. New York: Vintage Books.

Watzlawick, Paul. (1993). *The Situation Is Hopeless But Not Serious*. New York: W.W Norton & Company Inc.

Watzlawick, Paul, Weakland, J., & Fisch, R. (2011). *Change: Principles of Problem Formulation and Problem Resolution*. New York: W.W Norton & Company Inc.

Wolfberg, A., & Dixon, N. M. (2000). Speaking Truth to Power: Nurturing a Reflective Culture at the U.S. Defense Intelligence Agency. *Reflections*, *7*(4), 1–13.

Wright, J. C. (1962). Consistency and complexity of response sequences as a function of schedules of noncontingent reward. *Journal of Experimental Psychology*, *63*(6), 601–609.

CHAPTER 7

DATA PRE-PROCESSING IN TEXT MINING

Tuğçe AKSOY*, Serra ÇELİK, Sevinç GÜLSEÇEN*****

*Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: aksoy0tugce@gmail.com

**Dr., Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: serra.celik@istanbul.edu.tr

***Prof. Dr., Istanbul University, Informatics Department, Istanbul, Turkey
E-mail: gulsecen@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.07

Abstract

The fact that any kind of user has the ability to generate data with great ease at any time causes an increase in the importance of data mining. Considering the reality that the vast majority of the available data is composed of unstructured data and that the data in the text type is outnumbering, it proves the increasing interest in text mining and the abundance of studies in this field. However, in order to be able to examine an unstructured data type like text, which is quite different from machine language, it is necessary to make this data more structured and make the machine work. At this point, the data pre-processing step, which covers a large part of the entire text mining process, is of great importance. In this chapter, it is aimed to explain the text pre-processing phase on a basic level by supporting this using visuals. In doing so, it is primarily planned to focus on text mining and to explain in detail the characteristics of the data processed. In this context, it is aimed to explain the data pre-processing steps followed in order to overcome these difficulties by examining the difficulties created by the data in question. As a result, this chapter is a descriptive review of the data pre-processing phase in text mining, which covers some of the studies previously conducted on this subject.

Keywords: Text mining, Pre-processing, Linguistics

1. Introduction

This chapter is a compilation study on the basic level of data pre-processing steps in text mining. After a broad description of the concept of data mining, it is aimed to elaborate on text mining, to examine the pre-processing phase by explaining the characteristic features of the text data type and the difficulties it brings.

The concept, which emerged as Knowledge Discovery from a Database (KDD) and then evolved into data mining, involves the processing of existing raw data into usable knowledge. With the rapid development of technology, the production of huge amounts of data in seconds and the emergence of the concept of big data creates the need to accelerate the developments in the field of data mining. In the data mining process, which includes steps such as data selection, pre-processing, transformation, data mining and interpretation, the pre-processing step is the most important and the most time-consuming step. The fact that the amount of data produced every second increases greatly demonstrates the importance of the pre-processing step. The International Data Corporation (IDC) estimates that about 90% of all data is unstructured (Gantz & Reinsel, 2011). This again reveals the concept of pre-processing. Since the data pre-processing step is so important, it is aimed to introduce the subject at the beginning level and shed some general light on the subject. This study is a compilation study that combines the studies carried out in this field from past to present by making a literature review and explaining the pre-processing steps in detail with schemes.

The structure of the chapter is as follows: In the second section, the concept of data mining will be explained in detail and the basic concepts of data mining, which are classification, clustering and association will be explained and their connection with text mining will be elaborated. The importance of text mining, characteristics of data processed and why it is difficult to process will be explained. While doing this, questions such as what is text, a word and a paragraph are answered. In the third section, the text pre-processing step will be defined and each method used in this step will be examined individually and in detail. While doing this, basic examples will be presented with a better understanding of these methods with visuals. In the fourth section, the conclusion paragraph will be presented and a final evaluation will be made and the importance of the pre-processing process will be re-emphasized.

2. The Definitions of Data Mining and Text Mining

Structured data is data in a tabular form which has some relational rows (records) and columns (variables). The type of each variable is pre-defined - numerical, categorical, logical, date, image, and so on. It is the easiest data to work on. It is defined as a type of data in which

both structured and unstructured data coexist. Unstructured data is information that either does not have a predefined data model or is not organized in a pre-defined manner. It includes data types such as text, sound, image and video (Eberandu, 2016). It is the most difficult data to work with and needs to be transformed into structured form before it is processed. By its very nature, the concept of data mining can be defined in many different ways. The reason for this is that the process may vary according to the type of data and the needs of the person in many different fields and sectors. However, according to Han, Kamber, & Pei (2011), data mining means the process of extracting interesting patterns from large amounts of data and discovering information. This is because the main idea in data mining would not be efficient if huge amounts of raw data available worldwide are not processed. What is meant here is that the data becomes knowledge through the necessary steps and that this knowledge is exchanged by people and made available for a specific purpose. The data may be numerical, image, audio, video and text.

There are basically five steps in data mining (Fayyad, Piatetsky-Shapiro & Smith, 1996). These are the selection of data, pre-processing of data, transformation of data, data mining and interpretation. In the selection of the data, the data to be processed is determined and the information to be obtained from this data is predicted. In the data pre-processing phase, it is aimed to make the selected data suitable for performing the necessary mathematical operations and creating a model. In this step, deficiencies in the existing raw data are tried to be eliminated, the noise is removed and the data is made much more structured. At the end of this process, the attributes that are contained in the data are determined and by making use of these attributes, models are tried to be created by methods such as classification, clustering, association and regression performed at next step, which is data mining. Finally, there is the stage of interpreting the performance rates by calculating how well the models represent the data.

A more detailed review of the data mining methods before the data pre-processing step will help to understand how important the data pre-processing phase is for the next steps.

2.0.1. Classification

In general, the classification method is the process of categorizing the elements in the data. A dataset has certain variables e.g. inputs (predictors) and an output (class). Outputs are categorical or numerical results according to classicifaction algorithms such as Classification and Regression Trees (CART), and Random Forest. As a preliminary preparation of the classification method, the list of categories is determined so that the classification system and the elements of the data are split as sample data for each category (Jo, 2018).

In order to apply the classification algorithms, the data must first be divided into training and test sets (Mitchell, 1997). The training set is used as sample data to create the classification capacity by using machine learning algorithms, and in the test set, the data elements are classified and the differences between the real and classified labels are observed (Jo, 2018).

Duda, Hart & Stark (2000) divide the classification into soft and hard classification in their studies. According to this distinction, the conditional probability of the class is calculated in the soft classification method and the class estimate is made according to the greatest probability. On the other hand, in the hard classification method, the classification boundaries are determined directly without estimating the probability of the class. In addition, Duda et al. (2000) have classified the classification method horizontally and hierarchically. In the horizontal classification method (Figure 2), the categories are predetermined as a single list, whereas in the hierarchical classification method (Figure 1), clustered categories exist in a number of categories and the categories have the characteristics of a tree model.

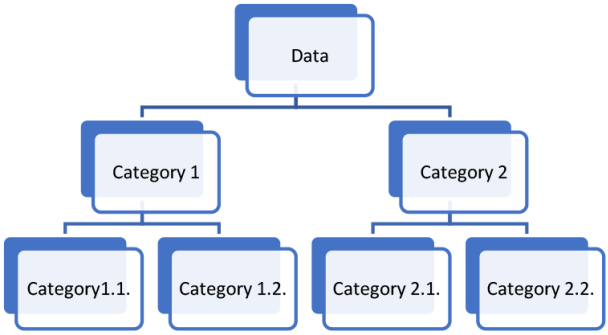


Figure 1: Hierarchical Classification Method

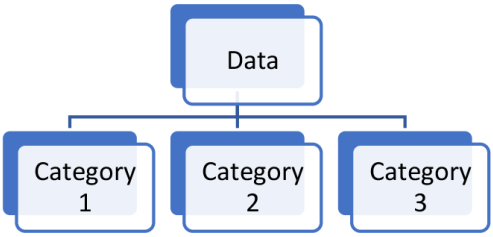


Figure 2: Horizontal Classification Method

2.0.2. Clustering

Clustering is the process of splitting a group of various data elements into subgroups, that is, clusters with similar properties. At this stage, first, unlabeled data elements are provided

and the similarity measures between them are calculated. The elements are subcategorized according to the similarities between them. The most important purpose of clustering is to maximize the similarity of elements in each cluster and minimize the similarity between clusters (Jo, 2006).

Duda et al. (2000) divide clustering into hard and soft clustering. In hard clustering, all elements can be collected in one cluster, whereas in soft clustering, each element can be clustered in more than one cluster (Figure 3).

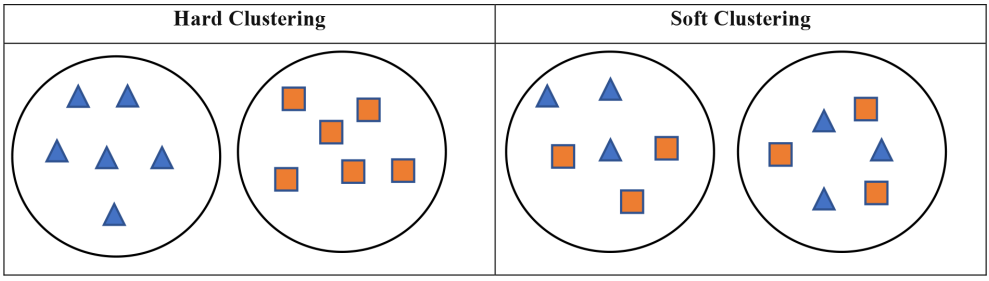


Figure 3: Differences Between Hard and Soft Clustering

Clustering is also divided into horizontal clustering and hierarchical clustering (Duda et al., 2000). In horizontal clustering, clusters are created as a single list while in the hierarchical clustering method, clusters are created in the tree model, and clusters can also have subsets.

The clustering method can automate the predetermination of categories, which is a prerequisite for the classification method. Horizontal or hierarchical clusters constitute the category list to be used in the classification method (Jo, 2006).

2.0.3. Association

Association is considered to be the extraction of the association between data elements in the if-then form (Jo, 2018). In other words, the association rule is a method aimed at revealing the related variables and determining the magnitude of the connection between them. It often involves identifying repetitive patterns and making predictions through them. It also provides great benefits in predictions in the fields such as purchasing, marketing and campaigning (Jo, 2018).

Confidence and support measures are frequently used when creating association rules. Support determines the rate at which a relationship is repeated in the entire data set, while confidence reveals the possibility of coexistence with two variables. If two variables are independent of each other, there is no association between them (Jo, 2018).

2.0.4. Regression

Regression is the process of estimating the output values of each data element. In other words, it is the process of calculating continuous data by examining the input data. Unlike the classification method which provides discrete values as outputs, regression gives continuous values as outputs (Jo, 2018). Regression gives numerical results based on a linear model such as shown in Figure 5. Regression works as a classification method if the dependent (output) attribute is categorical (as in the Logistic Regression model).

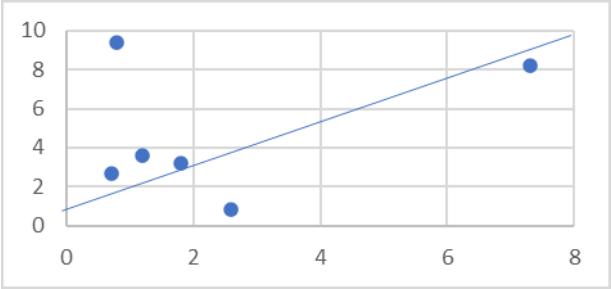


Figure 4: Regression Analysis

2.1. Text Mining

In general, text mining is the process of extracting quality information from the text by making use of numerical methods and techniques. In different studies, it is also defined as the process of extracting implicit information from text data (Feldman & Sager, 2007). It is a more specific sub-branch of data mining. In text mining, it is aimed to obtain information that is not yet known from very large unstructured text data. According to Kalra and Aggarval (2018), information retrieval is actually text mining. Another view takes a more technical approach to text mining and defines text mining as a set of statistical and computer science techniques developed specifically for analyzing text data (Zanini & Dhawan, 2015). Examined in detail, it is understood that text mining is not actually a new concept and it is an extension of data mining. Therefore, many algorithms used in data mining can also be used in text data, and therefore in text mining. The only difference is that while data mining deals with structured, quantitative data, text mining deals with unstructured or semi-structured data. In fact, the goal is to extract meaningful numerical indexes from the text that the computer can understand. While doing this, statistical methods are used extensively. With text mining, the information contained in the text can be categorized and clustered to obtain results such as word frequency distribution, and distributions. Then association and predictive analyses can be applied to words. For this reason, calculating classification, clustering and

association rules of the text is the basic function of text mining (Jo, 2018). This is the reason for the detailed descriptions and definitions made in the previous section. If the concepts of classification, clustering and association are understood, the importance of data pre-processing phase in text mining will be more easily realized.

The process in text mining is no different from data mining. Only the types of data being processed differ.

Figure 5: The Text Mining Process (Gaikwad, Chaugule & Patil, 2014)

Looking at the history of text mining, the National Center for Text Mining (NaCTeM) is the first public institution established worldwide (Zanini & Dhawan, 2015). It was established by the UK's Joint Information Systems Commission (JISC). The first activities of text mining were observed in the 1980s, and at first text mining was only dealing with data in databases and data warehouses. Nowadays, with the developing technology, there has been an intense interest in this field - unstructured data has reached a very high rate of 90%, and there are many types of text data such as text messages, e-mails, social media activities, blog contents and internet searches that are of interest to text mining.

Text mining is used in classifying, organizing and summarizing documents, estimating and developing contextual suggestion systems. Recently, text mining has been increasingly used in epidemiology, economics and education, and applied research to gain insights into the market and consumer, particularly in relation to businesses (Zanini & Dhawan, 2015). The information obtained from this research is very useful in decision-making processes.

The types of text mining can be divided into seven groups according to their function:

1) Document Classification
2) Document Clustering
3) Information Retrieval
4) Web Mining
5) Information Extraction
6) Natural Language Processing
7) Term Extraction

2.1.1. Difficulties

The actual source of text data is, in fact, language, and it is not possible, at least for the time being, to establish a universal model in the text mining process, as each language has its own characteristics. It is very difficult to create a generally valid model that can be adapted to the different uses of a single language, let alone between languages. The fact that each language has different syntax, and the ability of a single language to produce a wide range of syntaxes within and especially in daily use, makes this process even more challenging. In addition to syntax, there is great ambiguity in language. A word can have different meanings depending on the context of the sentence in which it is used or different words can be used for the same meaning. Adopting this ambiguity and diversity of language into a mathematical model needs a complex structure and intensive language knowledge. It is even more challenging to process data because of the fact that the majority of the data that text mining deals with includes text messages, social media content, and e-mails which are also generated without much attention to grammatical rules such as spelling and punctuation. In order to carry out studies such as classification and clustering on text data, it is necessary to determine the attributes of the data elements, and even if the data is a single paragraph text, almost every word will come as a separate attribute, and processing these attributes causes a great deal of time and space consumption. The ambiguity of the language can be examined in four categories (Sheeba & Vivekanandan, 2012):

i) Homophony

Those are words that have the same spelling but have different meanings depending on the context of the sentence.

Example:

I left my phone.

My phone is on the left side of the table.

ii) Synonymy

Those words have different spelling but have the same meaning.

Example:

The small child was sleeping.

This kid is so smart.

iii) Polysemy

A word has different but interrelated meanings depending on the context of the sentence.

Example:

The woman's face was beautiful.

You have to face with the consequences.

He sat facing the door.

iv) Hyponymy

It means semantic inclusion between lexical units.

Example:

Dog - Animal

In this example, the dog has a hyponymy relationship with the animal species.

All these ambiguities make it difficult to conduct semantic analyses in the text mining process, to correctly label the words according to the word type and many other processes.

2.1.2. What is text?

Text is an unstructured data type consisting of arrays called words (Salton, 1998). According to Jo (2018), the text is a collection of sentences or paragraphs written in natural language. In the first definition, the concept is discussed in a technical context, while in the second definition, there is a more linguistic perspective. If both definitions are considered, the text consists of elements of very different dimensions. Words are the basic units of the text and the words come together in accordance with grammatical rules to form sentences and sentences form logical paragraphs (Jo, 2018). The emphasis here is important. Although it is stated in the previous sections of the chapter that the language can vary in syntactic terms, this diversity takes place within the framework of certain rules. Therefore, the words that make up sentences have to comply with certain rules. It is somewhat easier to produce models within the framework of these rules. However, there are no grammatical rule limitations when it comes to paragraphs. The important thing is that the combined sentences follow each other with a certain logic and sub-context (Jones and Manu, 1999). This requires significant semantic analysis and semantics is one of the most challenging concepts in the field of natural language processing.

The text includes variables such as the size, author and title of the text as well as the paragraphs that form it , and is considered short text if the text consists of a single paragraph,

medium-length text if it consists of a single group of paragraphs, and long text if it consists of multiple groups of paragraphs (Jo, 2018).

2.1.3. What is a sentence?

A sentence is a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses (https://www.lexico.com/en/definition/sentence). A sentence can only consist of a single complement, that is, it can only have the subject and predicate, or it can consist of more than one complement. Sentences begin with uppercase letters in most languages and end with punctuation, such as dots, question marks, exclamation points, and triple dots. All of these elements are objects of the pre-processing step and will be explained in more detail in the next section. In addition, the words in the sentence are separated by spaces in many languages, and in fact, this feature of languages is similar to the tokenization process (Manning, Raghavan, & Schutze, 2009).

2.1.4. What is a word?

The word is considered as the basic unit of a text. Since the text data processed in text mining will form a very long string as a whole, in the pre-processing phase, the text is segmented into a word list and the other processes are performed through this list. The reason for choosing the word as the smallest unit is that the smallest unit that can be evaluated as meaningful and an attribute is also the word and, as mentioned earlier, the main purpose of text mining is to extract meaningful information from unstructured data.

In text mining, there is a term called stop-words, which means unnecessary words. Stop-words are the words that are necessary in grammatical terms and that are very common in the text but do not make any sense alone. Due to the fact that they are frequently included in the text due to grammatical rules, they cause wrong measurement results in making necessary statistical measurements and therefore stop-words are extracted from the text in the pre-processing phase. This will be explained in more detail in the next section.

3. A Detailed Description of Pre-processing Steps

3.1. What is data pre-processing?

The process of data pre-processing in text mining consists of the steps of converting the original text data into a raw data structure that distinguishes important textual features between text categories (Srividhya & Anitha, 2010). There are many methods of data pre-processing in text mining and they all try to make documents or document collections

structured in some way. As many different methods emerged during the effort to make the text structured, these methods evolved over time. Data pre-processing is an important part of the natural language processing system as well as text mining because at this stage, characters, words and sentences are determined and the characteristics of these elements are specified and transferred to the next stage, then they are used in information retrieval or machine translation systems. The raw data, which is formed after the data that is planned to be processed is collected, goes through the visualization stage in order to understand its structure. Correlation matrices can be used to perform this process, especially in text mining, and thus the similarity between the attributes that make up the text can be examined and more reliable results can be obtained in the following steps. After the data is better understood through visualization, the actual data pre-processing step is initiated and data cleaning is performed. In the context of text mining, clearing of data involves steps such as removing stop-words, punctuations and special characters (Kalra & Aggarwal, 2018) and will be explained in more detail in this section. Performing the data cleaning process earlier will reduce the size of the data to be dealt with in the next steps, thus achieving more optimal results in terms of time and space. This is because not only are stop-words eliminated in the data pre-processing phase, but also words in multiple forms are reduced to a single form (Kadhim, 2018). Obviously, due to the detailed processing of the pre-processing step, the phase covers about 50% to 80% of the entire text mining process. Then, in order to carry out other text mining steps following the data pre-processing step, the attributes created as a tabular form are labeled and the task of each item in the text is determined.

For the reasons already mentioned, an efficient pre-processing step should effectively represent the text in terms of both space (storage) and time (information retrieval) requirements**,** as well as good retrieval performance (precision and recall) (Giagole, Patil & Chaudhari, 2013). It is understood that the purpose of the data pre-processing phase is to present the text as an attribute vector by separating each text into individual words and to establish a relationship between the obtained attributes and the text.

Finally, to summarize why the pre-processing phase is important in text mining: First, it reduces the file size of the text data because stop-words correspond to approximately 20% to 30% of the total number of words in the text, and stemming reduces the index size by almost 40% to 50% (Gurusamy & Kannan, 2014). Secondly, it is important to make information retrieval systems work more effectively. Because stop-words are useless in searches and text mining, they can cause confusion in information retrieval systems, and stemming is used to match similar words in the text file (Gurusamy & Kannan, 2014).

Feldman & Sanger (2006) divide the pre-processing phase into two types of methods: task-oriented pre-processing methods and other methods. In the task-oriented pre-processing method, there is usually a problem that needs to be solved, such as extracting titles or authors from a PDF file, and structured text descriptions are made by creating tasks and sub-tasks through this problem (Feldman & Sanger, 2006). Other pre-processing methods consist of formal methods that are created to analyze complex structures and that can also be used on texts produced in natural language, and they include classification schemes, probabilistic models and rule-based system approaches (Feldman & Sanger, 2006). Although the two methods differ from each other, the aim of both is to store the most meaningful information as an attribute and exclude unnecessary elements, and both methods deal with unstructured or semi-structured data.

Feldman & Sanger (2006) grouped the data pre-processing step according to their functions in a different way and divided the process into three subprocesses as shown in Figure 6:

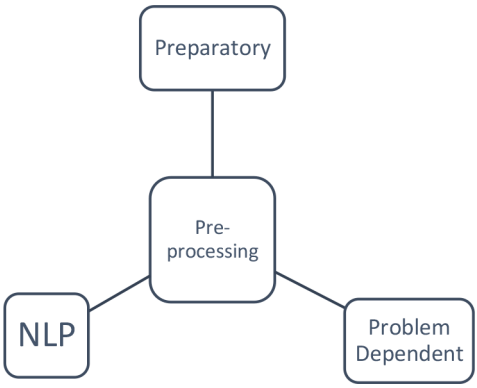


Figure 6: Feldman and Sanger (2006) Data Pre-processing Stages

At the preparatory stage, the raw data is first transformed into the appropriate structured form for the subsequent linguistic processing, and in doing so, divides the text into paragraphs, columns or tables. Steps such as determining words, POS tagging, syntactic parsing and morphological analysis are involved in the natural language processing. The output of this process is generally meaningless to the end user and is used as input to the next stage, the problem dependent stage. In the problem dependent stage**,** a final semantic representation of the data is created. In text mining, this process is usually completed with classification and information retrieval methods.

If we take a closer look at the linguistic steps at the stage of natural language processing, it is observed that linguistic methods are given different priority orders and even some of the studies include the methods that other studies do not mention. For example, while Brants (2003) argues about the natural language processing taking place during the information retrieval process, he lists methods such as stemming, POS tagging, compound recognition, de-compounding, chunking, and word sense disambiguation. However, Lourdusamy and Abraham (2018) list the pre-processing steps as tokenization, stop-word filtering, POS tagging, stemming, document indexing, grammatical parsing, text summarization, TF-IDF and chunking in their study. Kalra & Aggarwal (2018), on the other hand, believe that creating the vector space model of the attributes obtained in the list is another pre-processing step. In their work, Srividhya & Anitha (2010) include document indexing, and listing the attributes of the texts and measuring the term weighting of them in the pre-processing phase. TF-IDF is not described as a separate pre-processing step for them, but as a method that is frequently used in calculating term weights.

Although data pre-processing steps differ in the studies, it should be remembered that it is necessary to choose these steps according to the content of the text to be processed and the priority order of these steps should be determined according to the text. In line with this context, the following table summarizes the data pre-processing steps as a general review of all studies:

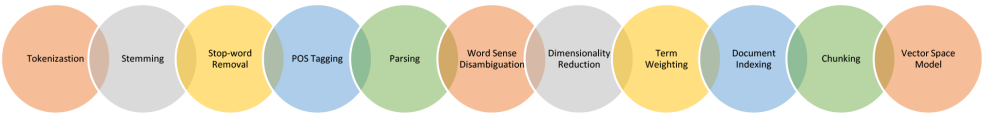


Figure 7: Text Pre-processing Steps

3.2. Text pre-processing steps

3.2.1. Tokenization

The first step in the text pre-processing phase is tokenization. Since we work on words - which are considered to be the smallest meaningful units in the text - they must be tokenized first. In general, the text is separated into words by spaces or punctuation marks and is organized in a list at this step. According to Jo (2018), the tokenization step is a prerequisite for stop-word removal and stemming. As the sub-stages of the tokenization step, capital letters can also be converted to small letters and special characters, symbols and numbers can be removed. However, converting the first letter of the words to lower case may affect the

chance of getting healthy results in the text mining process where there is a purpose such as selecting proper names. Such steps should be carried out taking into account similar situations. In one study, the removal of punctuation marks was considered as one of the basic steps in tokenization (Kalra & Aggarwal, 2018). Feldman & Sanger (2006) argue that this step is not limited only to identifying words, but in general, parsing the main text into paragraphs, sentences, phrases, words, and even morphemes that are accepted as the smallest meaningful units in linguistics, but the most commonly preferred parsing is on a sentence and word level. However, the main problem in parsing sentences is the determination of the beginning and end of sentences accurately. This is because the full stop, often used at the end of the sentence, can be used not only for this function, but also for abbreviations of titles in languages such as English. The full stop at the end of ordinal numbers in Turkish also causes great problems and inaccurate tokenization. Srividhya & Anitha (2010) divided the tokenization step into three phases: The first is to convert the text into a word count, i.e. to create a bag of words. Then, the necessary cleaning and filtering operations should be performed, that is, spaces, special characters and symbols should be removed and finally the text is converted into an attribute list, that is, separated into words, terms or properties. It is understood from these stages that in this study, the cleaning and filtering stage is accepted as a basic step in tokenization.

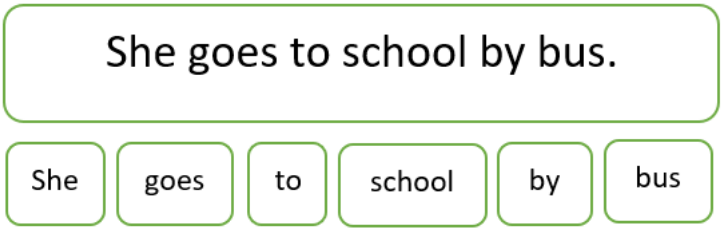


Figure 8: Tokenization

3.2.2. Stemming

In the stemming phase, the attributes in the list created in the previous step are separated from their prefixes and suffixes and converted to a stem, which is the nominative version of words. Although there are some differences between the meanings of the derivative and nominative words, they have relative meanings, and therefore, in order to avoid having to deal with a lot of data and improve performance, only one of the words with the same root goes on to the next steps and the derivative ones are excluded. However, in order to carry out this step, it is necessary to have an intensive linguistic knowledge and to develop language-

specific methods. Because each language has different grammar rules and each prefix or suffix is added to the words in accordance with these rules, the removal process should be performed according to those. Otherwise, the root words that do not actually exist can be obtained as outputs and may negatively affect the results of the study. This step is usually applied to nouns, verbs and adjectives (Kowalski & Maybury, 2000).

Lourdusamy & Abraham (2018) divide the problems that may occur in the stemming step: over stemming and under stemming. In the case of over stemming, two separate words having two different roots are stemmed to the same root while in the case of under stemming, two different words having the same root are stemmed to two different roots.

Example:

Membership (Input) -> Member - ship = Member (Output) [Name]

Discovered (Input) -> Discover - ed = Discover (Output) [Verb]

Historical (Input) -> Historic - al = Historic (Output) [Adjective]

Different methods are also mentioned in the stemming step and are divided into linguistic/dictionary-based stemming and Porter-style stemming (Brants, 2003). Accuracy of the linguistic/dictionary-based stemming is much higher in methods, but cost is also proportionally increased and its usage area is narrower, because language-specific methods need to be developed. In the Porter-style method, lower accuracy rates are obtained, but in the same direction, the cost is lower and healthy results can be achieved in the field of information retrieval with these methods.

3.2.3. Stop-word Removal

Stop-words are very high frequency words that do not contain any content information (Zhai & Massung, 2016). They consist of grammatical words of the language. They are necessary for the formation of sentences that meet the rules of grammar and do not have any meaning by themselves. Since the main purpose of text mining is to extract meaningful information from the text being processed, stop-words should be removed. These may be words in the language such as conjunctions, prepositions and pronouns. For example; but, however, because, with, like, it, this and these. Srividhya & Anitha (2010) argue that the step of removing stop-words should be done immediately after the tokenization step because the more the elimination of unnecessary data is carried out in the first stage, the higher the performance of the model, in other words, it provides more optimum results in terms of time and space. A list of pre-formed stop-words specific to a language is loaded into the system

and in case of matching with the existing words in the text, these words are removed (Kowalski & Maybury, 2000).

3.2.4. POS Tagging

POS tagging is the process of labeling the words in the text that are present as input according to their tasks in the sentence. POS tagging is a very important step for identifying neighboring words by labeling language-specific elements of sentences such as nouns, verbs, adjectives, prepositions, conjugations, adverbs, and analyzing syntactic structure and observing the relationship between words (Lourdusamy & Abraham, 2018). However, there are also studies that consider this step as an additional process (Jo, 2018). But there are different opinions objecting those studies, too. POS tagging plays a very important role in many natural language processing areas such as speech recognition, machine translation, information retrieval and information extraction (Singh, 2018). POS tagging is generally examined in two categories: rule-based approaches and statistical approaches. Rule-based approaches require an advanced linguistic expertise and a comprehensive collection that requires labor and cost. In addition, since it is necessary to create separate corpora specific to each language, it is not possible to have a universal characteristic. Although the problem of universality has not been solved, a transformation-based approach has been proposed as an alternative to this approach and is intended to automatically learn from the corpora. On the other hand, statistical methods benefit from Decision Trees and the Hidden Markov Model and are not specific to a particular language, they are universal. The data obtained after POS tagging can be used in a different function as stemmed and labeled words can represent separate dimensions in the vector space model (Brants, 2003). Thus, the model can give much more detailed results about the data. These labels provide information about the semantic properties of the text (Feldman & Sanger, 2006).

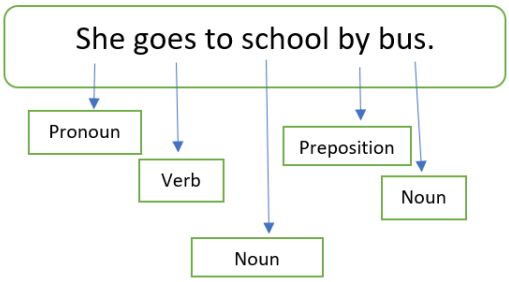


Figure 9: POS Tagging (Lourdusamy and Abraham, 2018)

3.2.5. Parsing and Chunking

Parsing and chunking steps are often given together in studies. Chunking separates the words in the sentence into basic phrases. Examples of these are noun phrases, adjective phrases and verb phrases. The phrases obtained by chunking form one dimension of the vector space model (Brants, 2003). Unlike parsing, chunking is used to create a hierarchical structure between elements of a sentence. Parsers are the comprehensive linguistic analysis of the text. These parsers are called syntactical parsing and divided into two categories depending on grammatical formalism: constituency and the dependency parser (Singh, 2018). The constituency parser separates the phrases in a sentence according to a hierarchical order and visualizes the relationship between the phrases. The root of the tree starts with "S", which represents the Sentence. Each sentence has a noun phrase and verb phrase, and they are referred to as "NP" and "VP", respectively (Zhai & Massung, 2016). In other branches, the labels of the words that were implemented in the previous step according to the content of the sentence are included in this tree. Shallow parsing, which is another method of parsing, is preferred when speed is important and forms a general model by separating only the noun and verb phrases without analyzing all the phrases in the sentence. This decomposition method is generally used to examine semantic relationships of phrases in sentences and to create functional classification labels. Dependency parser examines a sentence by reviewing the dependency of the words in pairs. Each dependency represents a linguistic function. These parsers see language as a set of relationships between words and create a graph for each sentence.

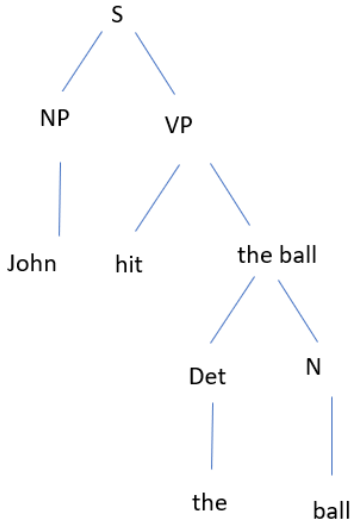


Figure 10: Parsing and Chunking (Zhai & Massung, 2016)

3.2.6. Word Sense Disambiguation

Word sense disambiguation is the process of distinguishing the correct meaning of the word within the context. When used in information retrieval, words are replaced by meaning in the vector space model (Brants, 2003). It is a step that needs to be done in sentences containing words that have different meaning according to the context. The problem of monophony mentioned in the previous section is the greatest example of this.

3.2.7. Dimensionality Reduction

Just as the removal of stop-words serves to remove words with very high frequency from the data, dimensionally reduction is intended to remove words with very low frequency. Document frequency is an important concept in this process and represents the number of documents in which a term exists. The most preferred method for dimensionally reduction is document frequency threshold. Words that are present in less than "m" documents are not considered as an attribute and are thus removed. "m" is a pre-determined threshold. This method is based on the idea that words that do not appear in a text at a certain frequency do not have an informative feature.

3.2.8. Term Weighting

Term weighting means calculating the weight of each word and assigning it an importance value. Each word in the text has different levels of importance (Salton & Buckley, 1988). Determining the term weights of the words before the stop-words removal and dimensionally reduction steps make those steps perform more effectively. There are three main factors that influence the importance of terms in a text: Term Frequency, IDF (Inverse Document Frequency) and Document Length Normalization (Karbasi & Bughanem, 2006). Therefore, Term Frequency and TF-IDF are the most commonly used term weighting methods. Term weights of the words are calculated according to their frequency of occurrence in the text. IDF is calculated according to the frequency of words found in all documents in the document database. The TF-IDF model is very popular in text classification and almost all models used for this process are variations of TF-IDF (Chrisholm & Kolda, 1998).

If all existing documents are called "D",

"w" is a word,

"d" is a document,

"w_d" is weight,

and the formula is:

$$w_d = f_{w,d} * \log(|D|/f_{W,D})$$

$f_{w,d}$ in the formula represents the frequency of "w" word in the "d" document. |D| is the size of the dataset and $f_{W,D}$ represents the frequency of "W" words in the "D" documents. The result of the TF / IDF measurement is a vector in which various terms exist together with their weights.

3.2.9. Document Indexing

Document indexing is defined as the process of converting a text into a list of words (Kowalski & Maybury, 2000). As mentioned in the first section, there are different opinions about the steps to be followed in the data pre-processing phase and this difference is mostly observed in the document indexing step. While document indexing is a term that expresses the data pre-processing phase in some studies, in others, it is considered as a framework covering the steps of dimensional reduction, term weighting and forming a vector space model. Therefore, in some studies, it is argued that document indexing has the same function as tokenization and in other studies, it is defined as a process in which keywords representing the documents in the best manner are selected and these keywords are appointed as the weight of these documents in the vector space model (Srivighya & Anitha, 2010). In this respect, document indexing is a step closely related to the weighting and dimensionally reduction.

3.2.10. Vector Space Model

The list of words obtained after tokenization is not yet suitable for numerical processing. Therefore, these words need to be converted to a numerical value. And each word should be converted to a term vector. The term vector facilitates further processing by giving numerical values to each word in the text. There are three ways to convert words into term vectors: Term Frequency, Term Occurrence and TF-IDF. The TF-IDF model is the most commonly preferred among them and it gives more weight to important terms and less weight to less important words. Vector values are observed between 0 and 1. In this context, 0 indicates that the word has no significance in the context of the text, while 1 indicates that the term is relevant to the text.

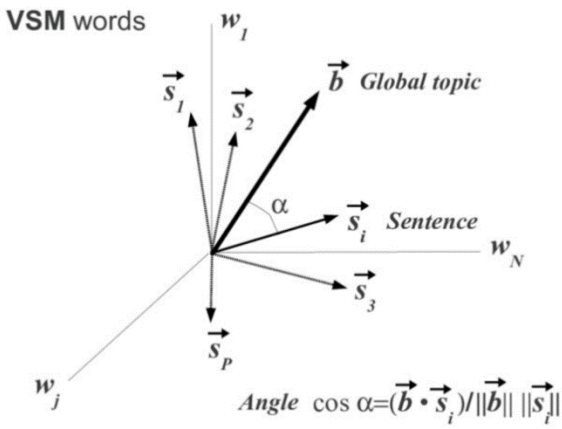


Figure 11: Vector Space Model (Moreno, 2012)

4. Conclusion and Evaluation

The aim of this chapter is to explain the data processing steps performed in text mining, and to convey the background information necessary to understand the subject and to support them with visuals. Firstly, explanations about the data (which is the main object of all these operations) are made and then the concept of data mining is examined. In this stage, classification, clustering, association and regression steps are mentioned among those models that are frequently used in data mining, what information is needed in these processes and what preliminary steps are needed to obtain healthy results are explained. Then, the concept of text mining, which is a sub-category of data mining, is tried to be defined and the fields that text mining is used are explained and the difficulties posed by the text data are mentioned. The reasons for the difficulty of expressing natural language with numerical algorithms are tried to be explained. Then, the text data involved in the text mining process was examined in more depth and the answers to the questions such as "what is text, a sentence and a word?" were sought. After providing all the necessary preliminary information, the data pre-processing steps were explained in detail and supported with visuals by emphasizing how important the data pre-processing stage plays in the whole text mining process.

It is understood from all the studies examined that the data pre-processing stage has a very important role in the text mining process. The data pre-processing step, which has a portion of about 50% to 80% of the entire text mining process, is of great importance in structuring the unstructured text data and creating the necessary models by means of algorithms. However, in this step, the methods chosen according to the type and purpose of

the text and the order of priority of these methods may vary. Although these differences exist, each method is of great importance in the process in which they are involved, and each step is a continuation or prerequisite of another. Therefore, before starting the whole text mining process, the available data should be examined very well, the content of the data should be known, the objectives should be determined by establishing solid foundations and necessary pre-processing steps should be followed in this direction. When all these conditions are fulfilled, it is foreseen that healthier text mining results will be obtained.

References

Brants, T. (2003, Ocak). *Natural Language Processing in Information Retrieval.* Conference: Computational Linguistics in the Netherlands.

Chrisholm, E. & Kolda, T.F. (1998). New Term Weighting Formulas for The Vector Space Method in Information Retrieval, Technical Report, Oak Ridge National Laboratory.

Duda, R.O., Hart, P.E. & Stark, D.G. (2000). Pattern Classification. Access Address: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.320.4607&rep=rep1&type=pdf

Eberandu, A.C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Emerging Trends & Technology in Computer Science, 38(1), 46-50. DOI: 10.14445/22312803/IJCTT-V38P109

Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, NY: Cambridge University Press.

Gantz, J. & Reinsel, D. (2011). Extracting Value from Chaos, IDC Iview.

Gaikwad, S.V., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. International Journal of Computer Applications, 85(17), 42-45.

Giagole, P.C., Patil, L.H. & Chaudhari, P.M. (2013). Pre-processing Techniques in Text Categorization. International Journal of Computer Applications. Access Address: https://pdfs.semanticscholar.org/ff34/7657082e70347a916548a9fe567ab791162a.pdf

Gurusamy, V. & Kannan, S. (2014). Pre-processing Techniques for Text Mining. Date: 18 February 2018, https://www.researchgate.net/publication/273127322_Pre-processing_Techniques_for_Text_Mining

Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques. USA: Elsevier Inc.

Jo, T. (2018). Text Mining: Concepts, Implementation and Big Data Challenge. Poland: Polish Academy of Science.

Jo, T. (2006). The Implementation of Dynamic Document Organization Using the Integration of Text Clustering and Text Categorization. University of Ottawa. http://dx.doi.org/10.20381/ruor-19708

Jones, K.S., & Manu, I. (Ed.). (1999). Automatic Summarizing: Factors and Directions in Advanced Automatic Summarization (pp.1-12). Cambridge, MA: MIT Press.

Kadhim, A.I. (2018). An Evaluation of Pre-processing Techniques for Text Classification. International Journal of Computer Science and Information Security, 16(6).

Kalra, V. & Aggarwal, R. (2018). Importance of Text Data Pre-processing & Implementation in RapidMiner. Proceedings of The First International Conference on Information Technology and Knowledge Management, (pp. 71-75). DOI: 10.15439/2018KMK6

Karbasi, S. & Boughanem, M. (2006). Document Length Normalization Using Effective Level of Term Frequency in Large Collections. Advances in Information Retrieval, Lecture Notes in Computer Science, 3936/2006, 72-83.

Kowalski, G.J. & Maybury, M.T. (2000). Information Storage and Retrieval Systems: Theory and Implementation. Boston: Kluwer Academic.

Lourdusamy, R. & Abraham, S. (2018). A Survey on Text Pre-processing Techniques and Tools. International Journal of Computer Sciences and Engineering, 6(3).

Manning, C.D., Raghavan, P., Schutze, H. (2009). Introduction to Information Retrieval. Cambridge, NY: Cambridge University Press.

Mitchell, T. (1997). Machine Learning. McGraw, NY: Hill Companies.

Moreno, J. (2012). Artex is AnotheR TEXt summarizer. CoRR, abs/1210.3312

Salton, G. (1998). Automatic Text Pre-processing: Transformation, Analysis and Retrieval of Information by Computer. Tokyo: Addison Weseley Publishing Company.

Salton, G. & Buckley, C. (1988). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5), 513-523.

Sheeba, J. & Vivekanandan, K. (2012). Improved Unsupervised Framework for Solving Synonym, Homonym, Hyponym & Polysemy Problems from Extracted Keywords and Identify Topics in Meeting Transcripts. International Journal of Computer Science, Engineering and Applications (IJCSEA), 2(5), 85-92.

Singh, S. (2018). Natural Language Processing for Information Retrieval. arXiv:1807.02383 [cs.CL]

Srividhya, V. & Anitha, R. (2010). Evaluating Pre-processing Techniques in Text Categorization. International Journal of Computer Science and Applications, 2010.

Zanini, N. & Dhawan, V. (2015). Text Mining: An Introduction to Theory and Some Applications. Research Matters: A Cambridge Assessment Publication, 19, 38-44.

Zhai, C., Massung, Z. & Özsu, M.T. (Ed.). (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool Publishers.

CHAPTER 8

BIG DATA IN EDUCATION: A CASE STUDY ON PREDICTING E-LEARNING READINESS OF LEARNERS WITH DATA MINING TECHNIQUES

Zeki ÖZEN*, Elif KARTAL, İlkim Ecem EMRE*****

*Ph.D., İstanbul University, Informatics Department, İstanbul, Turkey
E-mail: zekiozen@istanbul.edu.tr

**Ph.D., İstanbul University, Informatics Department, İstanbul, Turkey
E-mail: elifk@istanbul.edu.tr

***Res. Assist., Marmara University, Faculty of Business Administration,
Business Informatics Department, İstanbul, Turkey
E-mail: ecem.emre@marmara.edu.tr

DOI: 10.26650/B/ET06.2020.011.08

Abstract

Since the term "personalized learning" became popular, smart features have begun to be integrated into the e-learning environment. Data mining and machine learning algorithms are used to analyze big data stored in an e-learning system to make predictions to improve course quality or learners' performance. From the learners' perspective, it might now be considered possible for everybody to benefit from e-learning by considering their personal interests or their own specific development plan as long as the course contents are available in the system. In addition, in an e-learning environment, there is no limitation on the time and place where a course can be attended and a program completed. However, it is just not that simple. Today not the only, but by far the most important, the requirement is still the readiness of the learners to study in an e-learning system. The aim of this chapter is to predict the e-learning readiness of learners using data mining techniques. This chapter aims to provide feedback for institute managers and admin staff of e-learning systems which are intended to be used in an institution. Moreover, this section of the book contains one of the applications of big data analysis in education. Therefore, the main topic of this study is examined in terms of both classification and clustering techniques in order to provide a wider perspective to readers while using the sample application.

According to the results of this study, the highest accuracy value (0.831) is obtained with C4.5 Decision Tree Algorithm. While students, who agree and strongly agree with the statement "My studying/research area is appropriate for e-learning" are classified as ready to attend an e-learning course, students who disagree with the same statement are classified as not ready to attend an e-learning course. Students who strongly disagree with the statements "My studying/research area is appropriate for e-learning" and "E-learning is better than face to face learning", are also classified as not ready to attend an e-learning course. Furthermore, the statement "My studying/research area is appropriate for e-learning" is at the top of the obtained decision tree which indicates that it is an effective and directly related attribute which expresses student opinions about attending an e-learning course.

Keywords: Big data, Classification, Clustering, Data mining, Education

1. Introduction

E-learning is defined as "*instruction delivered on a digital device that is intended to support learning*" (Clark & Mayer, 2016). The e-learning concept can quite simply be argued on the basis of its two main pillars: technology and learners. From a technology perspective, e-learning is a serious investment. Technology directly affects techniques, tools, and applications that are used in e-learning. For training or educational purposes, e-learning management systems are used. There are many different open source and commercial alternatives for learning management systems (LMS), content management systems (CMS), virtual classrooms (VC), etc., that help to build and organize an e-learning environment. Course contents can be in various forms such as videos, text documents, podcasts, presentations, and such like, as long as they are compatible with the system. Learners can interact with each other and instructors via video and messaging features of the system. In addition, they can use their social media accounts or e-mails on the system for communication. Today, there are many other areas such as virtual reality and robotics that help instructors to improve the course contents and also the e-learning environment. The e-learning environment is also expected to work properly on desktop PCs, laptops, tablets, and mobile phones. Furthermore, any limitation of e-learning environment design mostly depends on the demands of the users. Improvements can be carried out using eye-tracking tools in terms of human-computer interaction for the evaluation phase of the system.

Since the term "personalized learning" became popular, smart features have been integrated into the e-learning environment as well. Data mining and machine learning algorithms are used to analyze big data stored in the system to make predictions to improve course quality or learners' performance. Big data means not only data with high volume, but it also indicates the production speed (velocity) of the data and the variety of the data

resources (Zikopoulos et al., 2011). Big data in education includes courses, course scores of the students, content information (time to complete, repetition time, pause points, last access time, etc.), system information (the most frequent usage time, the most frequently used browser, tools to access the system, etc.), face recognition data / keystroke dynamics data of the users, social media shares, etc. (Özen, Kartal, & Emre, 2017).

From the learners' perspective, it might be thought that today everybody can benefit from e-learning by considering personal interests or development plans as long as the course content is available in the system. In addition, in an e-learning environment there is no limitation on the time and place where a course can be attended and a program completed. However, it is just not that simple. Today not the only, but by far the most important, requirement is still the readiness of the learners to study in an e-learning system. The aim of this chapter is to predict the e-learning readiness of learners using data mining techniques. This chapter aims to provide feedback for institute managers and admin staff of e-learning systems which are intended to be used in an institution. . Moreover, this section of the book contains one of the applications of big data analysis in education. Therefore, the main question examined by this study is addressed using both classification and clustering techniques in order to provide a wider perspective to readers while demonstrating the sample application.

The following section gives a quick overview of e-readiness and e-learning readiness. The third section will focus on the study method (understanding the dataset, data pre-processing stage, and data mining techniques and models). The results obtained from the models are given in the findings section and an interpretation of the results and intended future research are discussed in the last section.

2. E-readiness and E-learning Readiness

E-readiness is "*a measure of the quality of a country's Information and Communications Technology (ICT) infrastructure and the ability of its consumers, businesses and governments to use ICT to their benefit*" (Economist Intelligent Unit, 2009). In other words, it is not enough merely to have perfect software and ICT infrastructure in a country, but also the citizens of that country should have the necessary skills to use ICT. One of the most noticeable research studies in the literature is "E-readiness Rankings". These reports were published by The Economist, Economist Intelligence Unit with the collaboration of IBM Institute for Business Value. In these reports, the selected countries were compared and ranked in terms of the e-readiness concept. The following definition from the Economist Intelligent Unit (2005) also helps to define the concept of e-readiness:

> *"A country's e-readiness is essentially a measure of its e-business environment, a collection of factors that indicate how amenable a market is to Internet-based opportunities". ... "E-readiness is not simply a matter of the number of computer servers, websites and mobile phones in the country (although these naturally form a core component of the rankings), but also such things as its citizens' ability to utilize technology skillfully, the transparency of its business and legal systems, and the extent to which governments encourage the use of digital technologies."*

This description refers to the concept that e-readiness is not only about technological infrastructures or opportunities, but also the ability/skills of the users and the other fields that support or affect these technologies. These reports have been published as "E-readiness Rankings" since 2000. However, in 2010 the institutions decided to change the name of the research and called it "Digital Economy Ranking". The explanation about this change is reported in the 2010's rankings as quoted below (Economist Intelligent Unit, 2010):

> *"Since 2000, the Economist Intelligence Unit has assessed the world's largest economies on their ability to absorb information and communications technology (ICT) and use it for economic and social benefit. Previously titled the "e-readiness rankings", in 2010 the study is being renamed as the "digital economy rankings", to reflect the increasing influence of ICT in economic (and social) progress". … "Given the prevalence of Internet-connected consumers, businesses and governments, and the indispensable role that digital communications and services now play in most of the world's economies, we believe that the countries in our study have achieved, to one degree or another, a state of e-readiness. The study's new title, the "digital economy rankings", captures the challenge of maximizing the use of information and communications technology (ICT) that countries face in the years ahead".*

It seems that the concept of e-readiness was considered in the light of its technical and technological aspects from 2000 to 2010. In 2010, it was decided to consider the effects of technological developments on the economies, and thus the name of the report changed to "Digital Economy Rankings". Table 1 shows the rankings of Turkey between 2002-2010. In these reports, all countries are given a score of e-readiness and a ranking is made accordingly. Between 2002 and2008, the e-readiness score of Turkey steadily increased , however in 2008 it started to decrease. Furthermore, it can be seen on the table that through the years there has not been a major change in Turkey's ranking.

Table 1. E-readiness rankings of Turkey between 2002-2010

	E-Readiness Rankings								***Digital Economy Ranking***
	2002	***2003***	***2004***	***2005***	***2006***	***2007***	***2008***	***2009***	***2010***
E-readiness score (of 10)	4.37	4.63	4.51	4.58	4.77	5.61	5.64	5.34	5.24
E-readiness rank	40	39	45	43	45	42	43	43	43
Total number of countries	60	60	64	65	68	69	70	70	70

Source: *(Economist Intelligent Unit, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010)*

In a report on the most recent ranking study of The International Institute for Management Development (IMD) World Competitiveness Center (2019), The Digital Competitiveness Ranking Results of Turkey among 63 countries can be seen in terms of overall, knowledge, technology, and future readiness between 2015 and 2019 (Table 2). The report explains knowledge as the "*know-how necessary to discover, understand and build new technologies*", technology as the "overall context that enables the development of digital technologies", and future readiness as the "*level of country preparedness to exploit digital transformation*".

Table 2. IMD World Digital Competitiveness Ranking 2019 Results

	2015	***2016***	***2017***	***2018***	***2019***
overall	52	50	52	52	52
knowledge	59	58	60	59	60
technology	48	48	49	45	48
future readiness	42	42	40	42	41

Source: (IMD World Competitiveness Center, 2019)

With the help of e-readiness reports and different studies which focus on the e-readiness of specific groups or countries (Purcell & Toland, 2004; Rizk, 2004; Al-Solbi & Mayhew, 2005; Ifinedo & Davidrajuh, 2005; Princely Ifinedo, 2005; Beig, Montazer, & Ghavamifar, 2007; Zaied, Khairalla, & Al-Rashed, 2007), the e-learning readiness of learners can be defined as the grasp of essential know-how in the field of ICT and the ability to use e-learning systems sufficiently. Furthermore, e-learning readiness includes the psychological acceptance and demand of the learners, as well. According to studies in Turkey and around the world,

e-learning readiness is a concept that has captured the attention of industries, schools, and universities. Studies in the literature can be classified according to the focus on their content. Some of them investigate the e-learning readiness from the perspective of the instructor and others from that of the learners.

There are many studies that have been conducted at the higher education level. Eslaminejad, Masood, & Ngah (2010) examined e-learning readiness from the perspective of the instructors. Hung et al. (2010) searched for the underlying dimensions of students' readiness for online learning. Mafenya (2013) conducted an observation of the pedagogical readiness of first-year students at the University of South Africa. A model was developed by Ha, Jm, & An (2014) in order to assess the e-learning readiness of lecturers in higher education institutions and the study was carried out with lecturers at the University of Nairobi. Parkes et al. (2015) investigated the readiness of students by surveying staff and students with previous learning experiences with e-learning. Paturusi et al. (2015) investigated lecturers' and students' readiness for e-learning at the University of Sam Ratulangi in Indonesia. Rasouli et al. (2016) investigated the readiness of art students in public Iranian Universities (Alzahra, Tarbiat Modares, and Tehran). Rohayani, Kurniabudi, & Sharipuddin (2015) reviewed the literature in order to investigate the factors that affect e-learning readiness in higher education by other researchers.

The study of Bonanno (2011) examines the usage of ICT technologies in education and this includes e-learning. The readiness in this study is investigated from a larger perspective. Akaslan & Law (2011a, 2011b) investigated the e-learning readiness of academic staff and students in the departments associated with the subject of electricity in different universities in Turkey. Soydal et al. (2011) made an assessment of e-learning readiness of academic staff at Hacettepe University. Ünal et al. (2013) conducted a study to assess the e-readiness of students at Hacettepe University. Doğan (2013) tried to observe the readiness in e-learning of lecturers at Osmangazi University. Korkmaz et al. (2015) investigated readiness and satisfaction in e-learning and their impact on academic achievement at Amasya University. Yurdugül (2016) concluded that the concept of e-learning is now more involved in education and thus the issue of eliminating deficiencies in the e-learning readiness level of university students has come to the fore. In the aforementioned study, the authors conducted a case study on university students and developed a scale for measuring e-learning readiness. As stated in that study there are other scales in the literature. Yilmaz (2017) conducted a study in a state university in order to observe the e-learning readiness of the students specifically in the flipped classroom model. Table 3 gives a brief summary of e-learning readiness studies in Turkey in terms of education levels and instructor/learner perspectives.

Table 3. Studies about e-learning readiness in education in Turkey

Title of the Study	*Reference*	*Education Level*	*Perspective*
Developing an instrument to assess teachers' readiness for technology-enhanced learning	(Bonanno, 2011)	-	Readiness of instructors
Measuring teachers' readiness for e-learning in higher education institutions associated with the subject of electricity in Turkey	(Akaslan & Law, 2011b)	University	Readiness of instructors
Measuring Student E-Learning Readiness: A Case about the Subject of Electricity in Higher Education Institutions in Turkey	(Akaslan & Law, 2011a)	University	Learner's readiness
Are Turkish universities ready for e-learning: A case of Hacettepe University Faculty of Letters	(Soydal et al., 2011)	University	Readiness of instructors
The Scale of Online Learning Readiness: A Study of Validity and Reliability	(Yurdugül & Sırakaya, 2013)	University	Learner's readiness
Evaluating E-Learning Readiness of Faculty of Letters of Hacettepe	(Moftakhari, 2013)	University	Readiness of instructors and learners
Students Readiness for E-Learning: An Assessment on Hacettepe University Department of Information Management	Ünal et al., (2013)	University	Readiness of learners
The examination of E-readiness levels of academicians	(Doğan, 2013)	University	Readiness of instructors
Two Concepts that Have to be Considered in the Transition of Vocational Colleges to E-Learning Model: Student's Computer Self Efficacy and E-Learning Readiness	(Pınar, Selçuk, & Dağ, 2014)	University, vocational college	Readiness of learners
E-Learning Readiness among Academic Staff in the Department of Information Science at the University of South Africa	(Ncube, Dube, & Ngulube, 2014)	University	Readiness of instructors
Assessing E-learning Readiness of Learners in Turkey	(Sharma, Gülseçen, Özen, & Kartal, 2014)	University	Readiness of learners
Assessing E-learning Readiness of Instructors in Turkey	(Sharma, Gülseçen, Özen, & Kartal, 2015)	University	Readiness of instructors
Students E-learning Readiness and Satisfaction Levels and Effects on the Academic Achievement	(Korkmaz et al., 2015)	University, vocational college	Readiness of learners
An investigation of Pre-service Teachers' Readiness for E-learning at Undergraduate Level Teacher Training Programs: The Case of Hacettepe University	(Yurdugül, 2016)	University	Readiness of learners
Iraqi Nursing Faculty Attitudes toward E-Learning a National Survey	(AL-Fayyadh & Mohammad, 2016)	University	Readiness of instructors
E-Learning readiness amongst nursing students at the Durban University of Technology	(Coopasami, Knight, & Pete, 2017)	University	Readiness of learners
Exploring the role of e-learning readiness on student satisfaction and motivation in the flipped classroom	(Yilmaz, 2017)	University	Readiness of learners

E-learning readiness is not only observed for educational purposes, but also for organizations/institutions (Aydin & Tasci, 2005; Lopes, 2007; Mercado, 2008; Schreurs, Ehlers, & Sammour, 2008; Schreurs, Moreau, & Ehlers, 2008; Darab & Montazer, 2011; Schreurs & Al-Huneidi, 2012; Azimi, 2013; Okinda, 2014; Kuruliszwili, 2015; Cathy, 2016; Doculan, 2016) or countries (Abas, Kaur, & Harun, 2004; Minges, 2005).

2. Method

This section covers the understanding of data, data preparation, and modeling steps of CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000). The evaluation step of CRISP-DM, which is related to evaluating the performance of the data mining models obtained, is also given in the Findings section.

2.1. Data Understanding and Data Preparation

In this study, data mining techniques are performed on "readiness" dataset, which is a subset of data that was used earlier for assessing the e-learning readiness of the learners in Turkey by Sharma, Gülseçen, Özen, & Kartal (2014). There are 667 observations (between the ages of 18 and 69, 330 male and 337 female) and 20 attributes in the "readiness" dataset. The statement "I am ready to attend an e-learning course" is taken as the target attribute of the survey (e-learning readiness status of a student); other statements are taken as predictive attributes. Age is numeric, gender, and e-learning readiness are binary. The rest of the attributes are coded according to the 5-Likert Scale (1=Strongly Disagree, 2=Disagree, 3=Have No Idea, 4=Agree, 5=Strongly Agree). Predictive attributes (except age) are treated as numeric in clustering analysis and as categorical in classification analysis. In addition, survey items are divided into three groups, namely "ICT skills" (S[1-5]), "e-learning experience" (E[1-6]), and "personal e-learning assessment" (A[1-8]). Figure 1 shows the class distribution of the target attribute which is designed based on the status of students regarding e-learning readiness. Table 4, Table 5, and Table 6 indicate the frequency (f) and the percentage (%) distribution of 5-Likert Scale attributes in terms of ICT skills, e-learning experience, and personal e-learning assessment. In this study, the authors focused particularly on predicting participants who are not ready for e-learning.

Dataset is normalized using min-max normalization technique at the data pre-processing stage. Data mining analyses are performed with C4.5 Decision Tree Algorithm for classification purpose and k-Means Algorithm for clustering purposes.

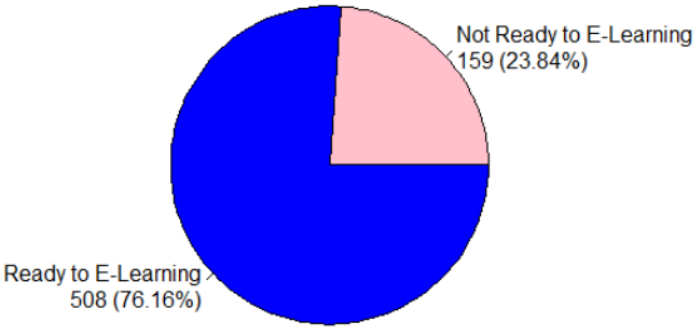


Figure 1: Class distribution of the target attribute

Table 4. Frequency (f) and percentage (%) distribution of 5-likert scale skill attributes

Attributes		***Strongly Disagree***		***Disagree***		***Have No Idea***		***Agree***		***Strongly Agree***	
		f	***%***	***f***	***%***	***f***	***%***	***f***	***%***	***f***	***%***
S1	I am good at using computer/internet	19	2.8	22	3.3	5	0.7	224	33.6	397	59.5
S2	I use my smartphone to communicate with my instructors outside the classroom	142	21.3	139	20.8	80	12	149	22.3	157	23.5
S3	I use social media for my courses	52	7.8	81	12.1	41	6.1	254	38.1	239	35.8
S4	I have a good e-learning background	93	13.9	133	19.9	178	26.7	176	26.4	87	13
S5	I have required IT infrastructure for e-learning	77	11.5	59	8.8	109	16.3	180	27	242	36.3

Table 5. Frequency (f) and percentage (%) distribution of 5-likert scale experience attributes

Attributes		*Strongly Disagree*		*Disagree*		*Have No Idea*		*Agree*		*Strongly Agree*	
		f	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*
E1	I have instructors which live in different cities/ countries	231	34.6	165	24.7	101	15.1	96	14.4	74	11.1
E2	I ask questions to instructors by e-mail	118	17.7	78	11.7	64	9.6	198	29.7	209	31.3
E3	I have joined a video conference before	245	36.7	151	22.6	65	9.7	93	13.9	113	16.9
E4	I want to join different courses at different universities	30	4.5	30	4.5	42	6.3	159	23.8	406	60.9
E5	I have used smartboard before	216	32.4	116	17.4	54	8.1	105	15.7	176	26.4
E6	I have attended an online course before	217	32.5	109	16.3	47	7	124	18.6	170	25.5

Table 6. Frequency (f) and percentage (%) distribution of 5-likert scale assessment attributes

Attributes		*Strongly Disagree*		*Disagree*		*Have No Idea*		*Agree*		*Strongly Agree*	
		f	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*
A1	I want lecture notes be shared electronically	35	5.2	27	4	20	3	147	22	438	65.7
A2	I prefer online exams because they are time saving and secure	107	16	95	14.2	92	13.8	143	21.4	230	34.5
A3	E-learning course content is different to face to face course content	59	8.8	71	10.6	99	14.8	232	34.8	206	30.9
A4	E-learning is better than face to face learning	174	26.1	186	27.9	176	26.4	69	10.3	62	9.3
A5	My studying/research area is appropriate for e-learning	84	12.6	101	15.1	124	18.6	221	33.1	137	20.5
A6	I prefer e-learning instead of face to face learning	164	24.6	185	27.7	121	18.1	102	15.3	95	14.2
A7	My instructors have enough IT skills for e-learning	86	12.9	107	16	258	38.7	131	19.6	85	12.7
A8	My university has the required IT infrastructure for e-learning	115	17.2	81	12.1	231	34.6	128	19.2	112	16.8

2.2. Modeling

In this study, our first stage of research involved the use of C4.5 Decision Tree Algorithm, developed by Quinlan (1993). It is one of the supervised learning algorithms that is used for classification problems. The aim of the C4.5 Decision Tree Algorithm is to find the best split node (attribute) using information gain and gain ratio and to create a set of rules in a tree form (Han & Kamber, 2006). These rules are used to predict the class label of an unlabeled observation. In addition, since it is the variable that provides the best partition, the algorithm also reveals the order of importance of the predictive attributes.

Stratified 5-fold cross-validation is used as a model performance evaluation technique for C4.5 Algorithm. In other words, the dataset is randomly divided into five pieces and each piece is used as testing dataset and the rest of the pieces are used as training dataset recursively. Each time accuracy value, error rate, sensitivity, specificity, positive predictive value, negative predictive value, and F-Measure are used as model performance evaluation metrics. At the end of the analysis, the mean of these metrics is used to evaluate the total model performance. In addition, the ratio of the class values of the target attribute is preserved in training and test datasets by means of stratified sampling.

The target attribute used in classification was based on the self-assessment of the respondents about their e-learning readiness. However, in similar analyzes to be performed with data collected through an e-learning system, a target attribute may not always be available in the dataset. Therefore, k-Means Algorithm, which is one of the clustering techniques, was used. This algorithm was also helpful to see how coherent the self-assessment of the participants was.

k-Means Algorithm is performed in order to cluster instances according to their similarities. This algorithm is one of the unsupervised learning algorithms that can be used when the target attribute does not exist in the dataset. The main aim of the algorithm is to cluster instances by minimizing the distance within clusters and maximizing distances between clusters. Since the dataset is intended to split into two groups, namely ready or not ready for e-learning, the cluster number k is determined as 2 for the k- Means Algorithm. After the k-Means Algorithm is performed, the class labels (ready or not ready for e-learning) are matched with the clusters according to the correlation between predictive attributes and the target attribute. Therefore, it can be said that all observations are labeled using k-Means Algorithm. Finally, new class labels are compared with the real class labels of the target attribute (Balaban & Kartal, 2015).

All the analyses are performed using R programming language (cran.r-project, 2019) on RStudio (RStudio, 2019). The following R packages are used: caret (Kuhn, 2018), clusterSim (Walesiak & Dudek, 2016), RWeka (Hornik, Buchta, & Zeileis, 2009; Witten & Frank, 2005), stats (R Core Team, 2018), and xlsx (Dragulescu & Arendt, 2018).

3. Findings

The performance of C4.5 Decision Tree Algorithm is given in Table 7. Performance evaluation metrics are calculated for each fold, then average performance results are given in the last column.

Table 7. Performance of C4.5 Decision Tree Algorithm

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*
Accuracy	0.836	0.806	0.865	0.820	0.827	0.831
Error Rate	0.164	0.194	0.135	0.180	0.173	0.169
Sensitivity	0.469	0.469	0.548	0.469	0.469	0.485
Specificity	0.951	0.912	0.961	0.931	0.941	0.939
Positive Predictive Value	0.750	0.625	0.810	0.682	0.714	0.716
Negative Predictive Value	0.851	0.845	0.875	0.847	0.848	0.853
F-Measure	0.577	0.536	0.654	0.556	0.566	0.578

Figure 2 shows the C4.5 decision tree obtained from the best cross-validation fold (Fold 3). Decision rules can be extracted from this decision tree such as:

- **Rule 1:** IF a student "strongly disagrees" with the statement "My studying/research area is appropriate for e-learning" AND "E-learning is better than face to face learning", the E-learning Readiness Status of the Student is NO.
- **Rule 2:** IF a student has "no idea" about the statement "My studying/research area is appropriate for e-learning" and "E-learning is better than face to face learning" AND strongly agrees with the statement "I want to join different courses at different universities", the E-learning Readiness Status of the Student is YES.
- **Rule 3:** IF a student "agrees" with the statement "My studying/research area is appropriate for e-learning", the E-learning Readiness Status of the Student is YES.
- **Rule 4:** IF a student "strongly agrees" with the statement "My studying/research area is appropriate for e-learning", the E-learning Readiness Status of the Student is YES.

Two models were performed using C4.5 Decision Tree Algorithm. C4.5 is first performed using real class labels as target attribute and secondly using cluster labels obtained using k-Means Algorithm as real target attribute. As seen in Table 8, specificity and negative predictive values are better than sensitivity and positive predictive values, because of the imbalanced dataset used in the analyses.

Table 8. Performance comparison of C4.5 decision tree models

Performance Evaluation Metric	*Classification with real class labels as target attribute*	*Classification with cluster labels as target attribute*
Accuracy	0.831	0.700
Error Rate	0.169	0.300
Sensitivity	0.485	0.868
Specificity	0.939	0.648
Positive Predictive Value	0.716	0.435
Negative Predictive Value	0.853	0.940
F-Measure	0.578	0.580

```
J48 pruned tree
------------------

My_studying/research_area_is_appropriate_for_e-learning = Strongly_Disagree
|   E-learning_is_better_than_face_to_face_learning = Strongly_Disagree: No (49.0/7.0)
|   E-learning_is_better_than_face_to_face_learning = Disagree: Yes (11.0/4.0)
|   E-learning_is_better_than_face_to_face_learning = Have_No_Idea
|   |   I_want_to_join_different_courses_at_different_universities = Strongly_Disagree: No (2.0)
|   |   I_want_to_join_different_courses_at_different_universities = Disagree: Yes (0.0)
|   |   I_want_to_join_different_courses_at_different_universities = Have_No_Idea: Yes (0.0)
|   |   I_want_to_join_different_courses_at_different_universities = Agree: Yes (1.0)
|   |   I_want_to_join_different_courses_at_different_universities = Strongly_Agree: Yes (4.0/1.0)
|   E-learning_is_better_than_face_to_face_learning = Agree: Yes (3.0/1.0)
|   E-learning_is_better_than_face_to_face_learning = Strongly_Agree: No (1.0)
My_studying/research_area_is_appropriate_for_e-learning = Disagree
|   I_want_to_join_different_courses_at_different_universities = Strongly_Disagree: No (5.0/1.0)
|   I_want_to_join_different_courses_at_different_universities = Disagree: No (5.0)
|   I_want_to_join_different_courses_at_different_universities = Have_No_Idea
|   |   Gender = Male: Yes (2.0)
|   |   Gender = Female: No (6.0/1.0)
|   I_want_to_join_different_courses_at_different_universities = Agree
|   |   I_have_a_good_e-learning_background = Strongly_Disagree: No (4.0)
|   |   I_have_a_good_e-learning_background = Disagree: Yes (11.0/2.0)
|   |   I_have_a_good_e-learning_background = Have_No_Idea: No (2.0/1.0)
|   |   I_have_a_good_e-learning_background = Agree
|   |   |   I_have_used_smart_board_before = Strongly_Disagree: No (0.0)
|   |   |   I_have_used_smart_board_before = Disagree: Yes (2.0)
|   |   |   I_have_used_smart_board_before = Have_No_Idea: No (0.0)
|   |   |   I_have_used_smart_board_before = Agree: No (3.0)
|   |   |   I_have_used_smart_board_before = Strongly_Agree: No (0.0)
|   |   I_have_a_good_e-learning_background = Strongly_Agree: Yes (1.0)
|   I_want_to_join_different_courses_at_different_universities = Strongly_Agree: Yes (41.0/9.0)
My_studying/research_area_is_appropriate_for_e-learning = Have_No_Idea
|   I_want_to_join_different_courses_at_different_universities = Strongly_Disagree: No (3.0/1.0)
|   I_want_to_join_different_courses_at_different_universities = Disagree: No (5.0/1.0)
|   I_want_to_join_different_courses_at_different_universities = Have_No_Idea
|   |   I_have_instructors_which_live_in_different_cities/countries = Strongly_Disagree
|   |   |   My_university_has_required_IT_infrastructure_for_e-learning = Strongly_Disagree: No (2.0)
|   |   |   My_university_has_required_IT_infrastructure_for_e-learning = Disagree: Yes (1.0)
|   |   |   My_university_has_required_IT_infrastructure_for_e-learning = Have_No_Idea: Yes (4.0)
|   |   |   My_university_has_required_IT_infrastructure_for_e-learning = Agree: Yes (0.0)
|   |   |   My_university_has_required_IT_infrastructure_for_e-learning = Strongly_Agree: Yes (0.0)
|   |   I_have_instructors_which_live_in_different_cities/countries = Disagree: Yes (4.0/1.0)
|   |   I_have_instructors_which_live_in_different_cities/countries = Have_No_Idea: No (2.0)
|   |   I_have_instructors_which_live_in_different_cities/countries = Agree: No (1.0)
|   |   I_have_instructors_which_live_in_different_cities/countries = Strongly_Agree: Yes (0.0)
|   I_want_to_join_different_courses_at_different_universities = Agree
|   |   Age <= 21: No (7.0/2.0)
|   |   Age > 21: Yes (18.0/2.0)
|   I_want_to_join_different_courses_at_different_universities = Strongly_Agree: Yes (60.0/8.0)
My_studying/research_area_is_appropriate_for_e-learning = Agree: Yes (169.0/14.0)
My_studying/research_area_is_appropriate_for_e-learning = Strongly_Agree: Yes (105.0/3.0)

Number of Leaves  :     39

Size of the tree :      50
```

Figure 2: C4.5 decision tree obtained from the best cross-validation fold

4. Discussion and Conclusion

This study aimed to predict the e-learning readiness of learners using data mining techniques. The intention was to provide feedback for institute managers and admin staff of e-learning systems planned to be used in an institution.

First, C4.5 Decision Tree Algorithm was used to predict the e-learning readiness status of the learners in the classification analysis. An approximate accuracy value of 83% was obtained. Classification results showed us that this study method was very effective and directly related to student opinions about e-learning readiness. In the Findings Section, Rule 3 and Rule 4 (if a student "agrees" with the statement "My studying/research area is appropriate for e-learning", the e-learning readiness status of the student was labeled as YES, if a student "strongly agrees" with the statement "My studying/research area is appropriate for e-learning", the e-learning readiness status of the student was labeled as YES.) can be seen a simple proof of this.

Moreover, from the beginning of studies on e-learning, a lot of work has been done regarding the comparison of face to face learning with e-learning. In this study, when this kind of comparison is considered with studying area of a student, in other words, if a student "strongly disagrees" with the statement "My studying/research area is appropriate for e-learning" and "E-learning is better than face to face learning", the e-learning readiness status of a student is predicted as NO.

Furthermore, the preference of a student about to join different courses in different universities is a significant factor. The possibility of studying together with faculty members from different universities, even from different countries, makes e-learning more attractive. If a student has "no idea" about the statement "My studying/research area is appropriate for e-learning" and "E-learning is better than face to face learning" and strongly agrees with the statement "I want to join different courses at different universities", the e-learning readiness status of a student is predicted as YES.

The sensitivity value in classification (0.485) is lower than the specificity (0.939) in the other model. The frequency difference between class labels of the target attribute is seen as the reason for a low sensitivity value in classification. Class labels of the target attribute should be taken on balance for further studies in order to obtain better performance results.

Secondly, k-Means Algorithm was used to cluster dataset without the target attribute based on the self-assessment of the respondents about e-learning readiness. Samples were

grouped by considering the similarities (in other words dissimilarities or distances) between them. Clusters obtained from the k-Means Algorithm were replaced with the real class labels (ready or not ready for e-learning). After that, performance evaluation metrics were calculated using cluster labels as target attributes. The accuracy value 70% was obtained. At this point, it can be seen that there was no such big difference between the results of the two models.

The authors believe that the results of the study will be beneficial for the feasibility of further study of an e-learning project and that this present study is a good example of big data analyses in the education field. Also, the dataset is only limited to the higher education students in this study. Dataset can be extended with students in primary and secondary schools or employees in the public and private sectors. In addition, some resampling methods such as oversampling, undersampling, and other such different methods can be used to eliminate imbalanced data problems just before the data mining analyses. Furthermore, other data mining techniques such as Naive Bayes Classifier, Binary Logistic Regression, k-Nearest Neighbor Algorithm, Support Vector Machines, etc. can be used to improve the performance of the prediction model.

5. Acknowledgments

This study was supported by the Scientific Research Projects Coordination Unit of Istanbul University. Project number 26089.

References

Abas, Z. W., Kaur, K., & Harun, H. (2004). E-learning readiness in Malaysia. *A National Report Submitted to the Ministry of Energy, Water and Communications*.

Akaslan, D., & Law, E. L.-C. (2011a). Measuring Student E-Learning Readiness: A Case about the Subject of Electricity in Higher Education Institutions in Turkey. In H. Leung, E. Popescu, Y. Cao, R. W. H. Lau, & W. Nejdl (Eds.), *Advances in Web-Based Learning—ICWL 2011* (pp. 209–218). https://doi.org/10.1007/978-3-642-25813-8

Akaslan, D., & Law, E. L.-C. (2011b). Measuring teachers' readiness for e-learning in higher education institutions associated with the subject of electricity in Turkey. *2011 IEEE Global Engineering Education Conference (EDUCON)*, 481–490. https://doi.org/10.1109/EDUCON.2011.5773180

AL-Fayyadh, S. A., & Mohammad, Q. Q. (2016). Iraqi Nursing Faculty Attitudes toward E-Learning a National Survey. *IOSR Journal of Nursing and Health Sciences*, *5*(3), 57–63.

Al-Solbi, A., & Mayhew, P. J. (2005). Measuring e-readiness assessment in saudi organisations preliminary results from a survey study. *From E-Government to m-Government*, 467–475.

Aydin, C. H., & Tasci, D. (2005). Measuring readiness for e-learning: Reflections from an emerging country. *Journal of Educational Technology & Society*, *8*(4). Retrieved from http://www.jstor.org/stable/jeductechsoci.8.4.244

Azimi, H. M. (2013). Readiness for implementation of e-learning in colleges of education. *Journal of Novel Applied Sciences*, *2*(12), 769–775.

Balaban, M. E., & Kartal, E. (2015). *K-Ortalamalar Algoritmasıyla Ülkelerin Bilişim Alanında Kümelenmesi [Clustering of Countries in Informatics with k-Means Algorithm]*. 112–117. ATO Congresium, Ankara, Turkey: Türkiye Bilişim Derneği.

Beig, L., Montazer, A. P. G. A., & Ghavamifar, A. (2007). Adoption a Proper Tool For E-Readiness Assessment in Developing Countries (Case Studies: İran, Turkey and Malaysia). The *Journal of Knowledge Economy & Knowledge Management (JKEM)*, *2*(1). Retrieved from http://dergipark.ulakbim.gov.tr/beyder/article/view/5000098823

Bonanno, P. (2011). Developing an instrument to assess teachers' readiness for technology-enhanced learning. *2011 14th International Conference on Interactive Collaborative Learning*, 438–443. https://doi.org/10.1109/ICL.2011.6059622

Cathy, J.-S. (2016). *Building a Tool for Determining E-learning Readiness of Organizations: A Design and Development Study* (Doctoral Thesis, Virginia Polytechnic Institute and State University). Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/70912/James-Springer_CD_D_2016.pdf?sequence=2

Clark, R. C., & Mayer, R. E. (2016). *E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*. John Wiley & Sons.

Coopasami, M., Knight, S., & Pete, M. (2017). E-Learning readiness amongst nursing students at the Durban University of Technology. *Health SA Gesondheid*, *22*, 300–306. https://doi.org/10.1016/j.hsag.2017.04.003

cran.r-project. (2019). The Comprehensive R Archive Network. Retrieved October 3, 2019, from https://cran.r-project.org/

D. Doculan, J. A. (2016). E-Learning Readiness Assessment Tool for Philippine Higher Education Institutions. *International Journal on Integrating Technology in Education*, *5*(2), 33–43. https://doi.org/10.5121/ijite.2016.5203

Darab, B., & Montazer, Gh. A. (2011). An eclectic model for assessing e-learning readiness in the Iranian universities. *Computers & Education*, *56*(3), 900–910. https://doi.org/10.1016/j.compedu.2010.11.002

Doğan, Ş. (2013). *The examination of E-readiness levels of academicians* (Master's Thesis, Eskişehir Osmangazi Üniversitesi). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Dragulescu, A. A., & Arendt, C. (2018). *xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files*. Retrieved from https://CRAN.R-project.org/package=xlsx

Economist Intelligent Unit. (2002). *The 2002 e-readiness rankings*. Retrieved from https://www.westerncape.gov.za/text/2004/2/ereadiness_2002.pdf

Economist Intelligent Unit. (2003). *The 2003 e-readiness rankings*. Retrieved from http://graphics.eiu.com/files/ad_pdfs/eready_2003.pdf

Economist Intelligent Unit. (2004). *The 2004 e-readiness rankings*. Retrieved from http://graphics.eiu.com/files/ad_pdfs/err2004.pdf

Economist Intelligent Unit. (2005). *The 2005 e-readiness rankings*. Retrieved from http://graphics.eiu.com/files/ad_pdfs/2005ereadiness_ranking_wp.pdf

Economist Intelligent Unit. (2006). *The 2006 e-readiness rankings*. Retrieved from http://graphics.eiu.com/files/ad_pdfs/2006ereadiness_ranking_wp.pdf

Economist Intelligent Unit. (2007). *The 2007 e-readiness rankings Raising the bar*. Retrieved from http://graphics.eiu.com/files/ad_pdfs/2007ereadiness_ranking_wp.pdf

Economist Intelligent Unit. (2008). *E-readiness rankings 2008 Maintaining momentum*. Retrieved from http://graphics.eiu.com/upload/ibm_ereadiness_2008.pdf

Economist Intelligent Unit. (2009). *E-readiness rankings 2009 The usage imperative*. Retrieved from http://graphics.eiu.com/pdf/e-readiness%20rankings.pdf

Economist Intelligent Unit. (2010). *Digital Economy Rankings 2010 Beyond E-readiness*. Retrieved from https://www-935.ibm.com/services/us/gbs/bus/pdf/eiu_digital-economy-rankings-2010_final_web.pdf

Eslaminejad, T., Masood, M., & Ngah, N. A. (2010). Assessment of instructors' readiness for implementing e-learning in continuing medical education in Iran. *Medical Teacher*, *32*(10), e407-412. https://doi.org/10.3109/0142159X.2010.496006

Ha, O., Jm, N., & An, W. (2014). E-Learning Readiness Assessment Model In Kenyas' Higher Education Institutions: A Case Study Of University Of Nairobi. *International Journal of Scientific Knowledge*, *5*(6), 29–42.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA, USA: Morgan Kaufmann.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics*, *24*(2), 225–232. https://doi.org/10.1007/s00180-008-0119-7

Hung, M.-L., Chou, C., Chen, C.-H., & Own, Z.-Y. (2010). Learner readiness for online learning: Scale development and student perceptions. *Computers & Education*, *55*(3), 1080–1090. https://doi.org/10.1016/j.compedu.2010.05.004

Ifinedo, P., & Davidrajuh, R. (2005). Digital divide in Europe: Assessing and comparing the e-readiness of a developed and an emerging economy in the Nordic region. *Electronic Government, an International Journal*, *2*(2), 111–133. https://doi.org/10.1504/EG.2005.007090

Korkmaz, Ö., Çakır, R., & Tan, S. S. (2015). Students E-learning Readiness and Satisfaction Levels and Effects on the Academic Achievement. *Journal of Kirsehir Education Faculty*, *16*(3). Retrieved from http://kefad2.ahievran.edu.tr/archieve/pdfler/Cilt16Sayi3/JKEF_16_3_2015_219-241.pdf

Kuhn, M. (2018). *caret: Classification and Regression Training*. Retrieved from https://CRAN.R-project.org/package=caret

Kuruliszwili, S. (2015). E-learning Readiness of Organization and Employees. *International Journal of Electronics and Telecommunications*, *61*(3), 245–250. https://doi.org/10.1515/eletel-2015-0032

Lopes, C. T. (2007). Evaluating E-Learning Readiness In A Health Sciences Higher Education. *El2007 Proceedings of the IADIS International Conference on E-Learning*. Presented at the IADIS International Conference on e-Learning, Portugal.

Mafenya, P. N. (2013). An Investigation of First-Year Students' Pedagogical Readiness to E-Learning and Assessment in Open and Distance Learning: An University of South Africa Context. *Mediterranean Journal of Social Sciences*, *4*(13), 353.

Moftakhari, M. M. (2013). *Evaluating E-Learning Readiness of Faculty of Letters of Hacettepe* (Master's Thesis). Hacettepe University, Ankara.

Mercado, C. (2008). Readiness Assessment Tool for an E-Learning Environment Implementation. *Fifth International Conference on E-Learning for Knowledge Based Society*, 183–187.

Minges, M. (2005). *Evaluation of e-readiness indices for Latin America and the Caribbean*. Retrieved from http://repositorio.cepal.org//handle/11362/31929

Ncube, S., Dube, L., & Ngulube, P. (2014). E-Learning Readiness among Academic Staff in the Department of Information Science at the University of South Africa. *Mediterranean Journal of Social Sciences*, *5*(16), 357–366. https://doi.org/10.5901/mjss.2014.v5n16p357

Okinda, R. A. (2014). Assessing E-Learning Readiness at the Kenya Technical Teachers College. *Journal of Learning for Development-JL4D*, *1*(3).

Özen, Z., Kartal, E., & Emre, İ. E. (2017). Eğitimde Büyük Veri [Big Data in Education]. In H. F. Odabaşı, B. Akkoyunlu, & A. İşman (Eds.), *Eğitim Teknolojileri Okumaları 2017* (1st ed., pp. 183–204). Retrieved from http://www.tojet.net/e-book/eto_2017.pdf

Parkes, M., Stein, S., & Reading, C. (2015). Student preparedness for university e-learning environments. *The Internet and Higher Education*, *25*, 1–10. https://doi.org/10.1016/j.iheduc.2014.10.002

Paturusi, S., Chisaki, Y., & Usagawa, T. (2015). Assessing Lecturers and Student's Readiness for E-Learning: A preliminary study at National University in North Sulawesi Indonesia. *GSTF Journal on Education (JEd)*, *2*(2). Retrieved from http://dl6.globalstf.org/index.php/jed/article/view/1160

Pınar, İ., Selçuk, A. G., & Dağ, B. (2014). Two Concepts that Have to be Considered in the Transition of Vocational Colleges to E-Learning Model: Student's Computer Self Efficacy and E-Learning Readiness. *Electronic Journal of Occupational Improvement and Research*, *2*(3), 50–60.

Princely Ifinedo, U. of J. (2005, March 26). Measuring Africa's e-readiness in the global networked economy: A nine-country data analysis. Retrieved June 4, 2017, from International Journal of Education and Development using ICT, Vol. 1, No. 1, 2005 website: http://ijedict.dec.uwi.edu/viewarticle.php?id=12.&layout=html

Purcell, F., & Toland, J. (2004). Electronic Commerce for the South Pacific: A Review of E-Readiness. *Electronic Commerce Research*, *4*(3), 241–262. https://doi.org/10.1023/B:ELEC.0000027982.96505.c6

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Rasouli, A., Rahbania, Z., & Attaran, M. (2016). Students' Readiness for E-Learning Application in Higher Education. *Malaysian Online Journal of Educational Technology*, *4*(3), 51–64.

Rizk, N. (2004). E-readiness assessment of small and medium enterprises in Egypt: A micro study. *Topics in Middle Eastern and North African Economies*, *6*. Retrieved from http://ecommons.luc.edu/cgi/viewcontent.cgi?article=1055&context=meea

Rohayani, A. H. H., Kurniabudi, & Sharipuddin. (2015). A Literature Review: Readiness Factors to Measuring e-Learning Readiness in Higher Education. *Procedia Computer Science*, *59*(Supplement C), 230–234. https://doi.org/10.1016/j.procs.2015.07.564

RStudio. (2019). RStudio—RStudio. Retrieved October 3, 2019, from https://rstudio.com/

Schreurs, J., & Al-Huneidi, A. M. (2012). E-Learning Readiness in Organizations. *International Journal of Advanced Corporate Learning (IJAC)*, *5*(1), 4–7.

Schreurs, J., Ehlers, U.-D., & Sammour, G. (2008). E-learning Readiness Analysis (ERA): An e-health case study of e-learning readiness. *International Journal of Knowledge and Learning*, *4*(5), 496–508.

Schreurs, J., Moreau, R., & Ehlers, U. (2008). *Measuring e-learning readiness*. Retrieved from https://doclib.uhasselt.be/dspace/handle/1942/8740

Sharma, S. K., Gülseçen, S., Özen, Z., & Kartal, E. (2014, May 5). *Assessing E-learning Readiness of Learners in Turkey* (S. Gülseçen, Z. Ayvaz Reis, & Ç. Selçukcan Erol, Eds.). İstanbul Üniversitesi, İstanbul, Türkiye: İstanbul Üniversitesi Yayınları.

Sharma, S. K., Gülseçen, S., Özen, Z., & Kartal, E. (2015). Assessing E-learning Readiness of Instructors in Turkey. *İstanbul Journal of Innovation in Education*, *1*(3), 13–28.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Soydal, İ., Alır, G., & Ünal, Y. (2011). Are Turkish universities ready for e-learning: A case of Hacettepe University Faculty of Letters. *Information Services & Use*, *31*(3–4), 281–291. https://doi.org/10.3233/ISU-2012-0659

The International Institute for Management Development (IMD) World Competitiveness Center. (2019). *IMD World Digital Competitiveness Ranking 2019 Results*. Retrieved from https://www.imd.org/globalassets/wcc/docs/release-2019/digital/imd-world-digital-competitiveness-rankings-2019.pdf

Ünal, Y., Alır, G., & Soydal, İ. (2013). Students Readiness for E-Learning: An Assessment on Hacettepe University Department of Information Management. *International Symposium on Information Management in a Changing World*, 137–147. Retrieved from https://link.springer.com/chapter/10.1007/978-3-662-44412-2_13

Walesiak, M., & Dudek, A. (2016). *clusterSim: Searching for Optimal Clustering Procedure for a Data Set*. Retrieved from https://CRAN.R-project.org/package=clusterSim

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Yilmaz, R. (2017). Exploring the role of e-learning readiness on student satisfaction and motivation in flipped classroom. *Computers in Human Behavior*, *70*, 251–260. https://doi.org/10.1016/j.chb.2016.12.085

Yurdugül, H. (2016). An investigation of Pre-service Teachers' Readiness for E-learning at Undergraduate Level Teacher Training Programs: The Case of Hacettepe University. *Hacettepe University Journal of Education*, 1–1. https://doi.org/10.16986/HUJE.2016022763

Yurdugül, H., & Sırakaya, D. A. (2013). The Scale of Online Learning Readiness: A Study of Validity and Reliability. *Education and Science*, *38*(169).

Zaied, A. N. H., Khairalla, F. A., & Al-Rashed, W. (2007). Assessing e-readiness in the Arab countries: Perceptions towards ICT environment in public organisations in the State of Kuwait. *The Electronic Journal of E-Government*, *5*(1), 77–86.

Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. New York, USA: McGraw-Hill Osborne Media.

CHAPTER 9

THE VALUE OF DATA FOR IMPROVING EFFECTIVENESS OF CAMPUS COURSES: THE CASE OF HYBRID MOOCS

Oğuz AK*, Selim YAZICI, Sevinç GÜLSEÇEN*****

*Ph.D., Boğaziçi University, Faculty of Education, Computer Education and Educational Technology, İstanbul, Turkey
E-mail: oguz.ak@boun.edu.tr

**Prof, İstanbul University, Faculty of Political Sciences, Department of Business Administration, İstanbul, Turkey
E-mail: selim@istanbul.edu.tr

***Prof, İstanbul University, Informatics, Informatics, İstanbul, Turkey
E-mail: gulsecen@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.09

Abstract

In recent years with the advances in technology, learners started to learn various concepts in informal learning environments apart from the official traditional learning programs. We describe such learning environments as part of the Personal Learning Environment (PLE) approach. One great resource for these environments is using Massive Open Online Course (MOOC). Learners can learn any subject by enrolling in MOOCs easily and develop themselves by reaching their personal learning goals. But in such an informal learning environment, it would be hard to manage the learning process. Learners need some ability to manage this process that is called "self-regulation". There are some problems in both fully face to face learning (like difficulties in following courses), and fully online MOOCs (like lack of interaction). So, a midway approach is a hybrid MOOC that is a combination of both methods. Literature and author experiences indicated that this method would make learning more effective. However, there is a need for improving the method with proper data management. We provide a list of data collection methods in hybrid MOOCs and explain how this data helped us to improve the learning process. In the PLE approach, students need data to shape their learning process, similarly instructors need to obtain data with various strategies and reshape the course structure by using this data. We think that in education, data usage is somehow limited, but it is required for making it more efficient.

Keywords: Hybrid massive open online courses, Massive open online course, Educational data analysis, Data for evaluation, Data for quality of education

1. Introduction

In recent years, online learning degrees have been created in addition to traditional classroom learning programs. Many people have been learning with the help of computers and the internet for decades, and they have gotten degrees. The Personal Learning Environment (PLE) approach -apart from the institutional degrees- is another way of learning. In this approach, learners set their learning goals and reach the learning content that they need and learn by themselves (Initiative, 2009). The last generation seems to use this method as well (Hollands & Tirthali, 2014).

Of course, there are some challenges in using this method. One of the main challenges is that students need to have self-regulation skills. Self-regulation is -in brief- a skill for managing individual learning processes (Zimmerman, 2000). Learners need to determine what they are required to learn and how to manage that process. Today, learners that have this skill can reach almost all knowledge within structured mediums like massively open online courses (MOOCs). MOOCs are online courses that can usually be accessed for free, without any prerequisites and are offered by credible universities around the world. Currently, there are really huge amount of MOOC resources that can easily be reached and millions of students are taking them (Liyanagunawardena, Adams, & Williams, 2013; Shah, 2015). Moreover, it is also known that the current generation can learn differently (Prensky, 2001). In addition, they can manage their learning process easily by using various tools like the Learning Management Systems (LMS).

Although some students can successfully manage their personal learning environments, some cannot. Similarly, although some students perform well in traditional learning courses, some do not. To help poorly achieving students in both environments a hybrid approach should be used. In particular, some online materials can be implemented into learning courses and both type of students could benefit from it. In this chapter we will demonstrate the use of hybrid Massive Open Online Courses (h-MOOCs) (Perez-Sanagustin, Hilliger, Alario-Hoyos, Kloos, & Rayyan, 2017).

The aim of this chapter is to clarify a personal learning environment method and how it could support traditional learning in the h-MOOC form. One important requirement for managing this process and maximizing its effectiveness is the use of good data management. It is because, data usage will enlighten learners and the course providers in such informal learning environments. Different types of data could be collected in h-MOOC applications with various methods like observations, interviews with students, logs from the tools etc. So,

we will discuss how to realize the inclusion of MOOCs into classroom learning and improve the process with the use of data.

2. An Alternative Learning Approach: Personal Learning Environments

2.1. What is Personal Learning Environment (PLE) Approach?

The personal learning environment approach consists of tools, services, and communities that create a learning environment with the aim of reaching students' personal learning goals (Initiative, 2009). So, to learn anything, participants do not have to follow a degree path, but can learn by themselves through this approach. Actually, this method existed before the computer and internet era, but the current tools and services make it a more feasible way of learning. And it seems people are using this method.

Since the invention of computers and the internet, learning management systems (LMS) have been using. With LMS, instructors can manage learning periods and create valuable data to better follow students' development. Personal Learning Environment (PLE) applications may be similar to LMS, but actually they are different. LMS is primarily designed for course management but PLE's main aim is to manage the learning period (Ebner & Taraghi, 2010; Initiative, 2009). It is important to differentiate the PLE approach and tools as well. PLE tools are software designed for using as part of the PLE approach.

To better understand this method, we need to discuss the place where learning occurs. Traditionally schools were addressed as the primary learning place. However, a school is not the only area, there is significant learning happening outside schools (Humanante-Ramos, García-Peñalvo, & Conde-González, 2015). Students can learn from other sources like books, media, friends, and family. To support this fact, Banks et al. (2007) stated that students only get about 19% of all learning gains in the classroom within their first 12 school years. A significant amount of learning occurs outside. In the PLE approach, our aim is to better manage the outside learning. Students need to have the ability to manage their own learning periods by using quality learning resources. In outside learning, descriptive data is needed for delivering these good information resources to students. In this way, learners could better decide their learning requirements and find the related resources. For example, when anyone buying a product from the internet he should follow some steps: First, he needs to decide what is required and then analyzes the products to see if the quality and specifications fit the requirements. In online sales, most stores try to provide data that best describes the product, they try to get feedback from users to describe how well the product functions. Manufacturers may use this data to produce better products next time. The PLE approach is similar to this

marketing example. Learning is a requirement and a person needs to have the ability to manage their learning needs and find resources that best fit their requirements. So, using good data in the PLE approach is critical.

Currently there are different physical and digital learning resources, but MOOCs seem to be the one that is well structured (but still need more development) and easy to access. So, in a PLE approach learners could use various learning resources including MOOCs. In this chapter we will try to better understand MOOC resources and in particular their integration with face to face courses using data.

2.2. Self-Regulation in Personal Learning Environments Approach

Self-regulation is an ability that includes thoughts, emotions, and actions to help a learner reach their personal goals (Zimmerman, 2000). Recently the term was mostly related with online environments (e.g. Kuo et al., 2014). It is because, students need to have more autonomy and responsibility skills in an online learning environment (Artino, 2008). Normally if students learn from instructors face to face, it is easier to communicate and ask questions, but in an online environment they have limited interaction with the content provider. It seems, it is one of the most difficult part of this approach, but good self-regulation skills could solve this problem.

Moore (1993) defined the "transactional distance" concept as the psychological and communicative distance between learner and instructor, and it is clearly high in online learning environments. However, in such environments like MOOCs this distance can be closed by a leaners' autonomy skill (Shearer, Gregg, Joo, & Graham, 2014). So self-regulation is a critical skill. To get this skill, students need to learn how to study by themselves from an early age. It seems once students have this skill they can manage learning by themselves for their whole life. We think 'learning how to dress' as an example: First, a child gets instructions of how to dress from an early age. Parents give instructions on how to put on a piece of clothing step by step, and never allow the child to try their own way. With this learning approach, every time the child faces a new type of dress, they would have difficulty and require parent support. In this example a parent's help is similar to an instructors' presentation in the classroom. Students need to learn their own way. If they don't find methods to learn by themselves, they would require an instructor's help in every learning process and this will limit learning.

3. Massive Open Online Courses (MOOCs)

3.1. A Good Resource for Learning

If we consider the profiles of current personal learners, we are faced with two major types of learners. The first type is the adult learners that are trying to improve their job skills (Castano-Munoz, Kreijns, Kalz, & Punie, 2017) and the second type is the students that are trying to complete their school learning. Both of them require information which they could get from various tools including the internet. Of course, there are other learner profiles as well, for example parents who need to learn "how to care for their child" or a person who would like to learn "how to cook" etc.

There are lots of learning opportunities today. Historically, learning has evolved with the advances in technology. Especially TV, radio, computers, and the internet have affected learning processes (Liyanagunawardena et al., 2013). After MIT's open Courseware project in 2001 a new era of "open learning" was started. It was because previously reaching good courses or course materials was limited as they were part of some learning degrees. In 2008 the first massive open online course was created (Liyanagunawardena et al., 2013) and since then there has been a massive open online course movement around the world.

Today, more than five thousand MOOCs are created, and more than 60 million learners are already registered for them. So, they are being actively used. Currently, there are two types of MOOCs; xMOOCs and cMOOCs. xMOOCs usually consist of video lectures and evaluated assignments (Siemens, 2013) and today most of the MOOCs in the USA are in this form (Daniel,2012). On the other hand, cMOOCs use a connectivist approach, which includes social interaction and online tools to create knowledge networks (Siemens, 2013). Any learner can reach a MOOC from any part of the world with internet access usually for free. It is a good opportunity because the MOOC content is provided by top universities in the world.

Hollands and Tirthali (2014) asked the reasons for providing MOOCs to MOOC administrators. They listed a set of reasons that includes; increasing the reach to learning, improving brand, economic reasons, innovation, and research about education. Interestingly, they did not talk much about the use of data or improving the quality of learning by using it. This would be the one missing point most educationalists have about this method of learning. Learning is not a process of content providing and presentation, and administrators need to acquire better use of the data inside the courses.

3.2. MOOC in Personal Learning Environment Approach

We stated that the PLE approach needed tools, services, and communities. As tools, online learning environments provide some software that includes LMS functionality. In addition, there are lots of services and communities that support learners like google docs, google drive, forums, discussions networks etc. Any person can use them to achieve their personal learning goals, but it may be hard for some learners to manage lots of different components. Especially for more traditional learners, a MOOC could be a better alternative because they are designed as a whole.

Any learner can take a MOOC to reach their own personal learning goal. Some platforms even offer MOOC degrees that are not like traditional learning programs. They are a kind of learning path recommended for some main learning targets. Learners can follow these paths and improve their ability in a specific field. The difference between these paths and traditional ones is that they are not mandatory, learners do not have to follow a strict schedule, and they have a certain prerequisite to access the learning content including entrance exams. As a result, MOOC is a good learning resource for a personal learning environment approach.

4. Hybrid Approach with the Use of Data

4.1. Hybrid Massive Open Online Courses to Improve Classroom Learning

Blended learning is a method for using both face to face and online learning together. Means et al. (2009) stated that blended learning usually results in better outcomes than solely online or solely face to face learning. So, although personal learning is usually outside official learning programs, they could be implemented into them, because in reality learners already try to use online resources to support their school learning.

There are different types of blended learning approaches. Kloos et al. (2015) stated the ways as: local digital prelude, flipping classroom canned digital teaching with remote tutoring, face to face and connected teaching, live teaching with remote tutoring, and face to face teaching with remote tutoring. Similarly, Perez-Sanagustin et al. (2017) showed four types of MOOC integration in the classroom as: MOOC as a service (MOOC can be recommended and not related to curriculum), MOOC as an added value (a MOOC that is related with curriculum is recommended), MOOC as a replacement (face to face course is replaced by a MOOC), and MOOC as an operator (face to face course is operated by a MOOC).

There are some h-MOOC examples in the literature. Konstan et al. (2015) applied a "MOOC as an added value" type of implementation. The application resulted in good learning gains and in general better student perception. Similarly, Bruff et al.(2013)'s "MOOC as an operator" application study showed that students similarly liked this application. In other studies, researchers reported usually more positive effects than the negative ones (Robinson, 2016; Swinnerton, Morris, Hotchkiss, & Pickering, 2017). They all showed that students in the current generation would benefit from this application.

Moreover, in one study, students used some common PLE devices and more than 90% of them stated that they would continue to use the platform after graduation (Tsui & Sabetzadeh, 2014). It seemed that PLE tools supported lifelong learning.

4.2. Data Collection in Hybrid MOOC Applications

In MOOCs, there were too many learners, so it was hard to give personal feedback and evaluation needed to automatize (Daradoumis, Bassi, Xhafa, & Caballé, 2013). Current automatic evaluation systems have limited interaction and seem not good enough. Similarly, peer review -which is a valuable evaluation type--is limited again because the learners were not professionals (Daradoumis et al., 2013). However, we think that automatic evaluation systems could evaluate learners, the problem is not related to its nature, but it is related to the limitations of current technologies.

Liyanagunawardena et al. (2013) analyzed 45 studies and found data collection techniques such as email interviews, focus groups, Moodle (an LMS) logs, discussion forums, blogs, and observations. By using a combination of techniques, MOOCs can be better evaluated. But in the case of hybrid MOOCs, data collection is somehow different. Actually, it is hard to reach MOOC usage data, but with alternative data collection methods students' data could be collected. We think it is better to evaluate what was learned than what was reached. For example, to understand student gains from a MOOC chapter, a very basic assignment about the related MOOC content could be given to students. To do so, short answers or writing some comments for discussion forum entries would be good.

While in MOOCs some quantitative data like questionnaires are collected, in hybrid MOOCs there is an opportunity to use qualitative measurements as well, because there is more interaction with student and instructor in face to face sessions. For example, the instructor can ask for feedback about the MOOC content and make observations about students understanding.

5. A Case Study: Data Analysis from a Hybrid Massive Open Online Course

In our case, an introductory database management course was given to 3 different groups of higher education students with a hybrid MOOC approach for three consecutive years. In each group there were about 30 students (there were 92 students in three years, 41% female and 59% male) who made good scores in the national university entrance exam in Turkey. Participants were in their 3rd year at university (average age is about 23). Participants followed half of the course from a MOOC while they followed the other half face to face. The MOOC course was an introductory database course that was given by Stanford University. Each week, participants studied two hours from MOOCs and the other two hours in class, face to face. While students studied theoretical concepts from the MOOC, in face to face sessions they learned a general summary of the theory which was mainly applied to small example projects.

In this process, it was recommended that learners study some MOOC content every week during their own time and place. During this period some data collection techniques were used and in general students' satisfaction and learning improvement was observed.

In this hybrid MOOC application, some qualitative and quantitative data was collected with different methods. As a quantitative method, students completed some database diagram assignments during the term, and entered an exam. Moreover, they submitted small discussion post entries on homework each week about the given MOOC content. As qualitative data, very quick verbal surveys at the beginning of face to face meetings were done. The instructor asked some basic questions about the weeks' content like "how can you merge 2 tables with SQL's 'select' statement with alternative ways? Did you learn from the MOOC?" and he observed their level of understanding during the face to face learning. One advantage of hybrid MOOC application over a completely online MOOCs was to get weekly qualitative data from students. If they had some difficulty they could freely ask their instructor in meetings. So, it is a good idea to follow students' level of learning in the face to face meetings.

Moreover, in the hybrid MOOC application instructors can evaluate assignments and have a better understanding about what they learned. This is much more difficult in fully online MOOCs. In fully online MOOCs, the course is usually limited to content presentation and basic assignments. Some students may not learn the concepts although they completed a MOOC. It is because, usually in MOOCs they do not have to mirror their learning with projects, but in h-MOOC there is such an opportunity. In fully face to face courses, students need to take notes during class or instructor should share course content to allow them to

study on their own time. In this method even though students connect with the content, their effectiveness is limited compared to a recorded course. In MOOCs there is usually a recorded and carefully planned course content. Student could work by themselves with this and can repeat the content if they did not understand a part.

After three consecutive years of h-MOOC application, the course instructor -depending on his observations of the courses- perceived that students were learning subjects much more easily in the course hours as compared to fully face to face. For example, if the course instructor taught a new concept in a face to face section, because many students were learning the content for the first time, they had difficulties in understanding. But in h-MOOC sections, they performed clearly better because they already viewed new course content before the session that made them familiar with the concepts. Moreover, it is known that students learn at different speeds. Previously, although many students understood the course, some of them had problems and they asked questions during the classes. These questions were slowing down the learning process of the whole class in face to face sessions. But in h-MOOC, the instructor reminded them that they would repeat a part of the face to face course from MOOC if they did not understand. If they still had difficulties, there was an 'office hours' option as well.

According to instructor observations in the courses and small discussions with the students, the instructor perceived another good side of the h-MOOC application as having flexibility for both instructors and students. As students usually have complicated schedules, if they do not have to enter a course, it is observed that some prefer to do the work themselves rather than follow the course. This was especially valid if they could find some equivalent online courses like physics or computer science. In this manner, when they are given the opportunity to take some part of the courses on their free time, students were observed to be happy because of this flexibility. This was valid for the instructor. Because he assigned some online courses, he had additional time for managing the course quality and giving more feedback during office hours.

In fully online MOOCs, one problem seemed to be the lack of interaction (Daradoumis et al., 2013). It was also observed in our three years of study. But this was not a big problem in h-MOOCs because if students did not understand some part of the online form they could ask questions in the class. Actually, two types of students were observed: The first type of students prefer to search the content over the net and find answers to their questions on MOOCs. They seem to have high self-regulation skills. The second type of students prefer to ask the instructor questions instead of searching over the net. This type of students may not be happy

about using the MOOCs but only a few complaints were observed over the three years. We think that using h-MOOC can serve both types of students. In fully online MOOCs high dropout rates were observed (Liyanagunawardena et al., 2013). One reason would be that they were looking for extra interaction. On the other hand, although there is high interaction in fully face to face courses, some students hardly focus on the course, because they want to learn at their own pace. So, because h-MOOC type includes both methods together they would better help both groups of students.

6. Conclusion

In this chapter we mentioned that MOOC is a good type for a personal learning environment approach and also a good resource for data collection to improve learning outcomes. In addition, a PLE approach can be used in face to face learning with the help of MOOCs. So, a combination of MOOC and face to face sections that is called as hybrid MOOCs (PerezSanagustin et al., 2017) could be an effective way of learning.

In related literature it is understood that using a hybrid MOOC would result in mainly positive outcomes in terms of achievement and student feelings. Data collection is a critical part of the learning period because it will shape the overall process. Especially in the PLE approach, good data would be very helpful for students to allow them to shape their own learning process.

In h-MOOCs there is a need for special data collection approaches. It includes short interviews with students, observing their learning level, and evaluating homework. This data can help instructors to manage the course content. Actually, an instructor does not have to use only one MOOC resource (Bruff et al., 2013), instead he could take different chapters from different MOOCs and even alternative web resources. So, in time, the course content is getting better with the help of data. In this process, an instructor always needs to look at the effectiveness of the specific material that he used. In addition, an instructor should be encouraged to share this data in various platforms. In learning, there is a need for sharing learning experiences about such cases to make the learning process more effective.

Instructors can use LMS or similar software to manage their courses. It is good to assign the online courses, assignments, and discussion sections in a proper way. Many LMS, like Moodle or Edmodo, could be used for this aim. Instructors could share the required course content with URL to related sources. By using such LMS, students will have a live portfolio for their future life. Whenever they need, they could open required course from this portfolio.

Although it seems frustrating at the beginning, instructors could easily adopt this kind of application and over time it would become routine for them. Overall, because it gives some flexibility to both students and instructors, and it helps students to learn the content easier, this method makes learning more efficient. In addition, with the h-MOOC method, a course could be driven with more data and it has an advantage to better track the process. Of course, there could be problems as well. Some students, especially the ones that are more comfortable with the traditional education system and do not have high self-regulation skills may not like these methods. But, in the current generation, the number of this type of students is probably very limited. Finally, although the method works well in a Database Management Course, it may not be a proper way for learning in a different type of course. There is a need to research and evaluate the effectiveness of it in different courses like the social sciences, history, and the arts.

References

Banks, J. A., Au, K. H., Ball, A. F., Bell, P., Gordon, E. W., Gutiérrez, K., . . . Mahiri, J. (2007). Learning in and out of school in diverse environments: Life-long, life-wide, life-deep. Seattle: The LIFE Center and the Center for Multicultural Education, University of Washington.

Bruff, D. O., Fisher, D. H., McEwen, K. E., & Smith, B. E. (2013). Wrapping a MOOC: Student perceptions of an experiment in blended learning. Journal of Online Learning and Teaching, 9(2), 187.

Castano-Munoz, J., Kreijns, K., Kalz, M., & Punie, Y. (2017). Does digital competence and occupational setting influence MOOC participation? Evidence from a cross-course survey. Journal of Computing in Higher Education, 29(1), 28-46. doi:10.1007/s12528-016-9123-z

Daradoumis, T., Bassi, R., Xhafa, F., & Caballé, S. (2013). A review on massive e-learning (MOOC) design, delivery and assessment. Paper presented at the P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on.

Ebner, M., & Taraghi, B. (2010). Personal learning environment for higher education–a first prototype. Paper presented at the World conference on educational multimedia, hypermedia and telecommunications.

Hollands, F. M., & Tirthali, D. (2014). MOOCs: Expectations and Reality. Full Report. Online Submission.

Humanante-Ramos, P. R., García-Peñalvo, F. J., & Conde-González, M. Á. (2015). Personal learning environments and online classrooms: An experience with university students. IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, 10(1), 26-32.

Initiative, E. L. (2009). things you should know about… Personal Learning Environments. Educause..

Kloos, C. D., Muñoz-Merino, P. J., Alario-Hoyos, C., Ayres, I. E., & Fernández-Panadero, C. (2015). Mixing and blending MOOC Technologies with face-to-face pedagogies. Paper presented at the Global Engineering Education Conference (EDUCON), 2015 IEEE.

Konstan, J. A., Walker, J. D., Brooks, D. C., Brown, K., & Ekstrand, M. D. (2015). Teaching Recommender Systems at Large Scale: Evaluation and Lessons Learned from a Hybrid MOOC. Acm Transactions on Computer-Human Interaction, 22(2), 23. doi:10.1145/2728171

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. The International Review of Research in Open and Distributed Learning, 14(3), 202-227.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. US Department of Education.

Moore, M. G. (1993). Theory of transactional distance. Theoretical principles of distance education, 1, 22-38.

Perez-Sanagustin, M., Hilliger, I., Alario-Hoyos, C., Kloos, C., & Rayyan, S. (2017). H-MOOC framework: reusing MOOCs for hybrid education. *Journal of Computing in Higher Education*, 29(1), 47-64. doi:10.1007/s12528-017-9133-5

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon,* 9(5), 1-6.

Robinson, R. (2016). Delivering a medical school elective with massive open online course (MOOC) technology. *Peerj*, 4, 9. doi:10.7717/peerj.2343

Shah, D. (2015). MOOCs in 2015: Breaking Down the Numbers. Retrieved from https://www.edsurge.com/news/2015-12-28-moocs-in- 2015-breaking-down-the-numbers

Shearer, R., Gregg, A., Joo, K., & Graham, K. (2014). Transactional Distance in MOOCs: A critical analysis of dialogue, structure, and learner autonomy.

Siemens, G. (2013). Massive open online courses: Innovation in education. *Open educational resources: Innovation, research and practice*, 5, 5-15.

Swinnerton, B. J., Morris, N. P., Hotchkiss, S., & Pickering, J. D. (2017). The Integration of an Anatomy Massive Open Online Course (MOOC) into a Medical Anatomy Curriculum. *Anatomical Sciences Education*, 10(1), 53-67. doi:10.1002/ase.1625

Tsui, E., & Sabetzadeh, F. (2014). Lessons Learnt from and Sustainability of Adopting a Personal Learning Environment & Network (Ple&N). *International Association for Development of the Information Society*.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective.

CHAPTER 10

INTELLIGENT TUTORING OF LEARNERS IN E-LEARNING SYSTEMS AND MASSIVE OPEN ONLINE COURSES (MOOC)

Yacine LAFIFI*, Asma BOUDRIA, Atef LAFIFI***, Moadh CHERAITIA******

*LabSTIC Laboratory, University 8 may 1945 Guelma, 24000, Algeria
e-mail: lafifi.yacine@univ-guelma.dz
**LabSTIC Laboratory, University 8 may 1945 Guelma, 24000, Algeria
e-mail: boudria.asma@univ-guelma.dz
***Computer science department, University 8 may 1945 Guelma
BP 401 Guelma, 24000, Algeria
e-mail: saiflafifi@gmail.com
****Computer science department, University 8 may 1945 Guelma
BP 401 Guelma, 24000, Algeria
e-mail: moadhcheraitiaa@gmail.com

DOI: 10.26650/B/ET06.2020.011.10

Abstract

In the last few years, many terms related to learning environments have emerged. Each one of these terms is distinguished by a set of criteria such as the target audience, the duration of learning, the type and nature of the educational content, the manner of dissemination of knowledge, etc. Among problems encountered in these environments, the lack of support for learners seems to be a serious problem that requires special attention. Tutoring can be seen as a good solution for this problem; it consists of giving assistance to learners who are in need for help. Regardless the nature of this assistance (pedagogical, social, etc.), it can be delivered in many forms: advice, guidance or even recommendation. If tutoring has been applied for a long time in traditional e-learning environments, its application in new systems such as Massive Open Online Courses (MOOC) is still under study. In fact, a considerable number of studies driven on MOOCs had reported the problem of learners' dropout. Several reasons can be listed as causes of such a problem. Among these reasons, we can find learners' isolation as well as learners' loss of motivation. This same problem has been reported by researchers working in the field of Computer-based Environments for Human Learning. In this article, we propose a new vision on how to apply an intelligent tutoring process in human learning systems in general and in MOOCs in particular. This new vision is based on the behaviors and skills of learners. This activity can take many forms and can be carried out by different types of actors (teachers, learners, etc.).

Keywords: E-learning, MOOCs, Intelligent tutoring, Social tutoring, Traces

1. Introduction and Motivation

Computer-based systems and networks have been in widespread use since the emerge of the internet and the technological boom of the early 21st century, and as dedicated to human learning they keep growing faster each year. These systems use a set of informational and educational resources that serve thousands of learners' needs. As uniquely designed for colossal number of learners, Massive Open Online Courses (MOOCs) allow the acquisition of knowledge easily with the advantage of being transmitted in different forms, audible, textual, or visual (Jordan, 2014; Khalil & Ebner, 2016; Zhang et al., 2016). These courses are freely accessible without any enrollment conditions in most cases due to participants' turnout in such systems. Moreover, learners are quite offered a set of tools that help carrying out the required activities and duties that can be traced or even tracked simultaneously. Such specialized use of MOOCs makes it highly sustainable for learners' to have accessible data that is to be traced whether implicitly or explicitly.

Many results can be found just by typing the word MOOC on any search engine, the thing that stands for the provability of several academic and professional bodies' work on this axis. In some developed countries, the development of MOOCs was one of the goals of their governments and their policies for research and development. Therefore, several educational institutions and companies come to offer their own MOOCs for they have some characteristics that encouraged thousands or even hundreds of thousands of learners to enroll online.

What really distinguishes these MOOCs from other learning formats is that they provide free access though its limited nature, demand no high degrees or even qualified-knowledgeable geeky learners, and immediately attract the attention of participants who face some difficulties on the audible and textual level.

The gathered data on the number of participants' access and enrollment in such systems sumps up its efficiency, however, the majority of participants would shy away and keep the programmed training unfinished although the splendid display that these MOOCs demonstrate. Many researches have shown that the percentage of learners who accomplished their courses is around 10%. Their withdrawal generated a sort of fear (Boyatt et al., 2014).

As an attempt to look for the reasons behind this quitting, results have shown that the main cause is their isolation and the lack of an adequate follow-up. The letter became very crucial and demands an immediate resolution.

In the wake of these findings, this framework is ultimately conducted to present an approach that offers an intelligent tutoring for learners using an e-learning system through different ways that work on eliminating the participants' isolation, and further guide them according to their needs and social profiles. It aims also to empower the teachers and tutors' roles as supporters, and interaction makers. The latter can intervene to answer all different questions asked from the part of the enrollers whether they are pedagogical, social, administrative, technical, etc. (Bendjebar et al., 2016). Moreover, the process involves real identification of the participants' categories and their exact level/situation (very serious situation, serious situation, rather serious situation, etc.), so that the smart intervention take place, and assign them with the appropriate type of tutoring. We come to raise these questions, under what category learners are to be classified during the learning process? How can we associate this system to meet the supported needs? What is the possible means that serve best the proposed intervention? These questions are best answered under full examination, which is the main goal of this approach.

To validate our approach, we have implemented a system called 'TutMOOC' (http://www.mooc24.net/MC/) dedicated to learning "algorithmic" subject. TutMOOC has been tested for one month. The obtained results were very promising and encouraging.

This paper is divided into three parts; section 2 is a full presentation of previous researches on tutoring and social tutoring in MOOCs. Section 3 involves a very detailed demonstration of the proposed approach. The last section is highly dedicated to the conclusion and further perspectives.

2. Related Works

Traditional learning environments involve backing up tasks, and one of the most functioning is tutoring. It takes several forms: supervision, pedagogical and psychological support, reinforcement of knowledge, answers to questions, motivation, etc. The main objectives of this activity are essentially to increase learners' motivation, break their isolation and increase opportunities to improve learners' levels of knowledge.

This tutoring task is performed by actors called tutors. They follow and accompany a number of learners. In some cases, learners themselves may be tutors; this is known as peer-tutoring. In all cases, tutors have several skills and perform a set of functions.

Several research studies have investigated the detection and the description of the tutor's roles and functions in human learning environments that can support hundreds of learners.

The reader can consult (Lafifi et al., 2010) to get an idea about the extraction and the desciption of the roles of online tutors in e-learning environments and how to assign roles to tutors.

In the case of MOOCs that contain thousands of learners and more, the tutoring task becomes very delicate and requires a special attention. From our findings, only few works that focused on figuring out what is exactly with applying tutoring in MOOCs, and why it is not working; these systems turned to be creating of more problems when testing it and the main reason is the huge number of learners enrolling in such online environments.

We found a first attempt that was the subject of a startup, Livementor (mentor or tutor live). Livementor (https://www.livementor.com/about-LiveMentor) is an online tutoring service that allows the user to benefit from an instant support, to have an excellent tutor according to the needs of the applicants and to ensure this task of tutoring at any time. The advantage of this system is that it can be applied to several types of learners (schoolchildren, students, young entrepreneurs, etc.). Nevertheless, the services offered by this system are not free.

Another system is Openclassrooms (https://openclassrooms.com/), which offers learners a video conferencing mentor. According to the developers of this platform, the tutor is "an expert" who adapts learning to the level of learners and "a tutor" to motivate them and lead them to achieve their goals.

In (Collect et al., 2017), the authors proposed peer tutoring to be applied in MOOCs. This tutoring technique has been applied in the POEM platform (Personalised Open Education for the Masses platform of the UNESCO UniTwin Digital Campus Systems). The authors indicated the impossibility of managing teacher-learner interactions in MOOCs given the very high number of learners. According to the authors, POEM offers each learner a tutor who is only another learner with a higher level in the same course. In this sense, the learner can ask questions to his tutor. Of course, if the tutor can not answer a question, he can send it to his own tutor, and so on until a question is sent to a teacher in case of difficulty.

3. Intelligent tutoring in e-Learning and MOOC environments

The main objective of this work is to offer learners a variety of mechanisms and tools of social tutoring to support and motivate them in order to reduce the risk of quitting the training. "Social tutoring" can be defined as the adoption of social indicators in the activity of human tutoring. In other words, social tutoring is attached to the application of tutoring through the

interactions between the learners themselves and by adopting some indicators, terms and tools used in social networks. The ultimate goal is to create a friendly and a social environment in which promoted learning is in turn essential for successful tutoring, that is to say, social tutoring means taking into account the social aspect when applying the tutoring task. Moreover, this smart social tutoring states that each learner's situation is associated with an appropriate tool to intervene before the learner's situation worsens. Indeed, for example if the learner is in a situation very close to giving up, the system offers a specific tutoring (semi-collaborative or TWISA (a term used by Algerian farmers during the periods of work of the land .... ..)).

In order to benefit from these tools and mechanisms, we must take into account information about learners. These information concern all the actions taken by the learners during the course follow-up once the learning, evaluation, tutoring and interaction with other human actors in the system have taken place. Furthermore, to make use of these gathered data, we propose to group them according to the nature of the activity and present them in a model or profile. For taking into account all these functionalities, the proposed approach is divided into three phases: the information collection phase (modeling), the zone detection phase and the intelligent tutoring phase (c.f. figure 1).

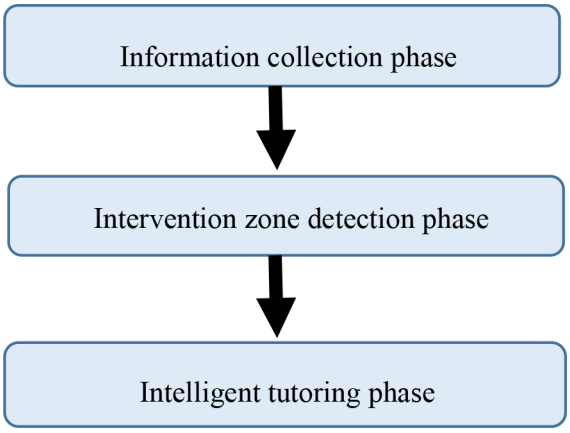


Figure 1: General description of the proposed approach

For validating our contributions, the proposed approach was supported by a system dedicated to learning "the algorithmic" subject, called "TutMOOC". The general architecture of this system is presented in Figure 2.

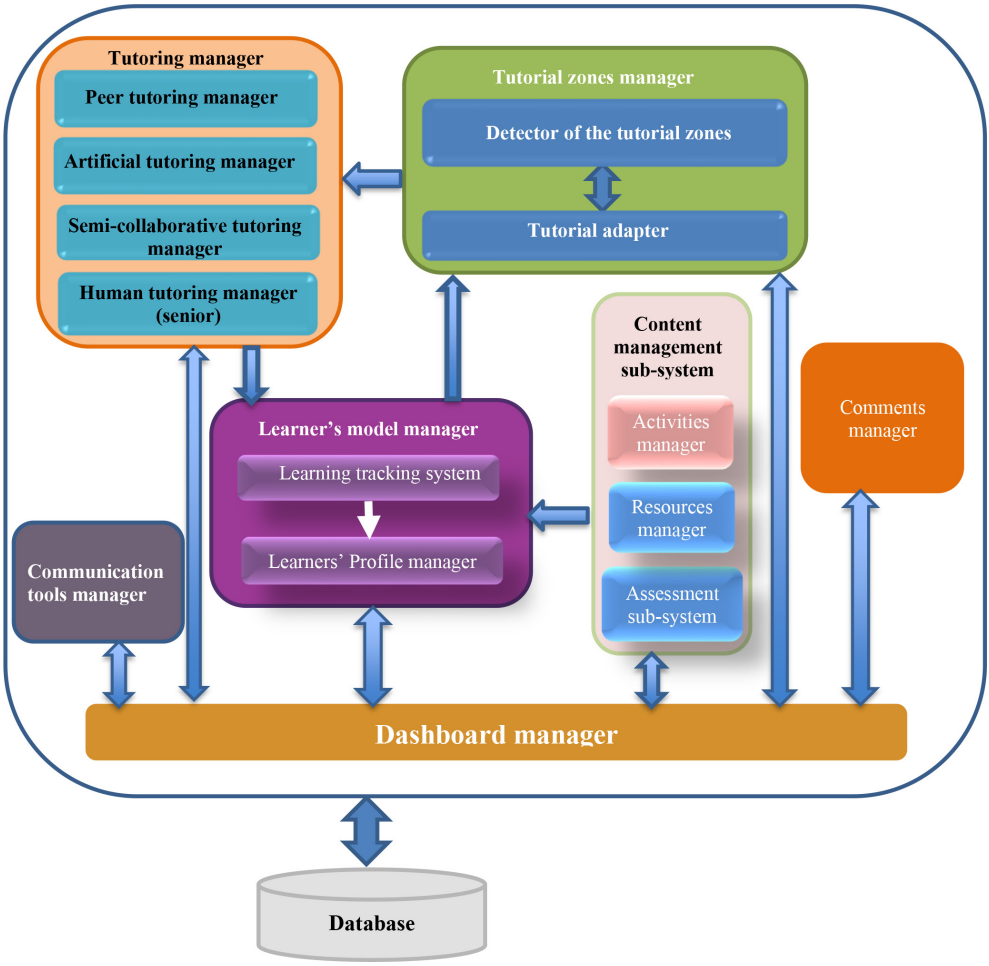


Figure 2: General architecture of the system

3.1. The collection of information (modeling)

In this phase, the system collects all the actions (traces) of the learners. These traces may be the result of the learner/learner and learner/system interactions. From these traces the system will build the model of the learner. The latter will be modeled by the quadruplet (CP, AP, BP, TP) where CP designates the cognitive profile, AP denotes the attendance profile, BP is the behavioral profile and finally TP designates the tutorial profile.

3.1.1. Learner Model

As already mentioned, our learning model consists of the cognitive profile, the attendance profile, the behavioral profile and finally the tutorial profile. Evidently, these names of profiles do exist and are cited by several researchers, even in our previous research works, except that in this work is the way to represent these profiles and the formulas used to calculate them are different. Indeed, we have proposed a model of the learner that contains: the cognitive profile that is modeled by tokens, the behavioral profile that is modeled by stars and the tutorial profile that is represented by sticks. We have associated with each interval a color to indicate the situation where the learner is.

Table 1 shows the proposed color classification.

Table 1. Classification of colors associated with the intervention zones

Profile	Number of	Representation	Red Zone	Orange Zone	Yellow Zone	Green Zone	Blue Zone
Cognitive	Tokens	●	0...25	25...45	45...60	60..85	85..100
Behavioral	Stars	☼	0..25	25..45	45..60	60..85	85..100
Tutorial	Sticks	|||||||	0..20	20..40	40..60	60..80	80..100

3.1.1.1. Cognitive profile

A. Level of knowledge:

This profile refers to the level of knowledge of the learner. It is calculated according to the learner's answers on a questionnaire designed for this purpose. Depending on the obtained responses, this profile can be: Very good, good, average, weak or very weak. We used the same formulas adopted in our previous research. Therefore, according to the final obtained grade (i.e. score); we obtain the associated cognitive level. Therefore, we propose the following levels:

- if (score <25%) then the cognitive_level := "very low".
- if (25 % <= score <45%) then the cognitive_level := "low".
- if (45% <= score <60%) then cognitive_level := "average".
- if (60% <= score <85%) then the cognitive_level := "good".
- if (85% <= score <= 100%) then the cognitive_level := "very good".

After the enrollment of a learner, he must pass a test in the form of multiple choices questions to know his initial level; once his level is upgraded the learner makes other tests. In our proposal, we used tokens to represent this level.

B. Attendance:

It indicates the learner's attendance towards the evaluation activities proposed by the teaching staff in the system or the MOOC. These activities may include homework, quizzes, questionnaires or any other type of assessment.

In any e-learning system or a MOOC, the teachers' staff proposes activities to be carried out by the learners to test their knowledge. In our proposal, we associate a coefficient of importance to each activity. This coefficient is designated by the staff when designing the evaluation activity. This coefficient can take an integer value between 1 and 3 (3: Very Important, 2: Important and 1: Not important). For example, "the exam" is a very important activity whereas "the assignment can be either small or important according to the staff.

Table 2. Coefficients of importance associated with some pedagogical activities

Type of activity	Degree of importance	Coefficient
Examination	Very important	3
Quiz	Important	2
Duty	Not important	1

This work shows the importance of the individual effort of each learner. Willingness and serious work can guide the learner to develop his level.

We propose a simple mathematical equation to calculate the attendance as follows (it is noted by A).

A = (number of completed activities) / (total number of scheduled activities) (1)

According to the value of A, we deduce if a learner is serious (when A is close to 1) or not (when the value of A is close to 0).

To update the cognitive profile of the learner, we use the following formula:

$$CP(Aj) = \frac{\sum_{i=1}^{n} Coefi * Mark\ (Activityi)}{\sum_{i=1}^{n} Coefi * Max - Mark(Activityi)}$$

With :

- ***Mark (activityi)*** is the grade associated with the pedagogical activity **i**. This grade is between 0 and 20 as in the notation adopted by institutions of higher education in Algeria.

- ***Coefi***: is the coefficient of the activity i.

- ***n***: total number of the activities.

3.1.1.2. Behavioral profile

It represents the behavior of the learner in the system. Notably, this profile indicates the degree of the interactions made by the learner. In other words, the learner's contributions using different communication tools offered determine his behavior. This profile is calculated using a set of indicators such as the number of questions posted on the forum, the number of messages sent, etc. To model this profile, we used "the stars" that indicate the degree of behavior of the learner. Each action performed by the learner is associated with a certain number of stars. We propose to use the following indicators:

Table 3. Indicators associated with the behavioral profile of the learner

Action / indicator	Number of proposed stars	Maximum number of stars
Number of system access (access)	1	20
Message sent via email (message-email)	1	20
Number of words "I like"(I-like)	1	20
Number of topics posted on the forum(subjects)	1	20
Public comment posted on the course(comments)	1	20

To enhance this social tutoring to work more in favor of the learners, we asked for their opinions and eventually the comments ranged from being positive to negative. According to the number of positive opinions (by quota of 10 in our case), the number of stars increases to reach a maximum (20 stars for our case). Similarly, depending on the number of negative reviews, the number of stars decreases to a minimum.

Finally, we can calculate the number of stars for a learner **j** using the following formula. This number is written as **Bj** as follows:

$$Bj=nb(accesj) + nb(commentsj) + nb(message_emailj) + nb(I\text{-}likej) + nb(subjectj)$$

Depending on the number of obtained stars, the associated zone and the corresponding color are obtained. We used the following rules to obtain the behavioral profile of a learner j (BPj) and the intervention zone:

- if (Bj <25) then BP_j: = “Isolated” and zone: = “red”.
- if (25 <= Bj <45) then BP_j: = “not very dynamic” and zone: = “orange”.
- if (45 <= Bj <60) then BP_j: = “moderately dynamic” and zone: = “yellow”.
- if (60 <= Bj <85) then BP_j: = “dynamic” and zone: = “green”.
- if (85 <= Bj <= 100) then BP_j: = “strongly dynamic” and zone: = “blue”.

3.1.1.3. Tutorial profile

It defines the learner’s ability to perform assistance tasks for other learners. In other words, it shows whether or not the learner can act as a tutor and, if so, what are his or her roles. All this information is then stored in his tutorial profile. This last one includes two types of information: the role of the tutor and the number of sticks acquired.

Of course, the learner-tutor with more skills and knowledge is expected to help and answer other learner’s (his or her peers) requests for help, known as peer tutoring. During their time in the system, learners can support other learners by helping them to find appropriate learning resources, perform suitable tasks correctly, as well as motivate and encourage them. In our approach, we propose to adopt a stick-based approach.

Every request for help (tutoring) is rewarded with a number of sticks. In other words, each learner is assigned a number of sticks (nb = 100 in our case). This number will be updated according to the requests for assistance addressed to other learners and the replies given by the learner-tutor to the requests for assistance issued by the other learners.

If the answer sent by a tutor-learner is appreciated by the learner requesting the tutoring, the number of corresponding sticks is subtracted from the learner’s account and added to the participant’s account (number_sticks_response = 4 in our case). In the opposite scenario, i.e. if the tutor learner’s answer was not appreciated by the learner, the number of sticks allocated is divided on both (i.e. 2 sticks in our case of application).

Within our system, any learner can be considered as a tutor as long as he/she answers at least one request for assistance and this answer is satisfied. In other words, a learner can be a tutor if he or she possesses a number of sticks that is greater or equal to 104 (100 sticks at the beginning plus 4 sticks acquired after a satisfying answer). We have also proposed a ranking system for tutor-learners so that they can be promoted to senior tutors. This ranking is based on the number of sticks belonging to each of them. We have followed this ranking with regarding the number of sticks acquired.

- If number_sticks >= 104 then Tutor_Type: = “Beginner”.
- If number_sticks >= 200 then Tutor_Type: = “Experienced”.
- If number_sticks >= 300 then Tutor_Type: = “Senior”.

Figure 3 bellow shows a screenshot from by TutMOOC system that indicates the tutorial profile of a learner, where his zone is blue.

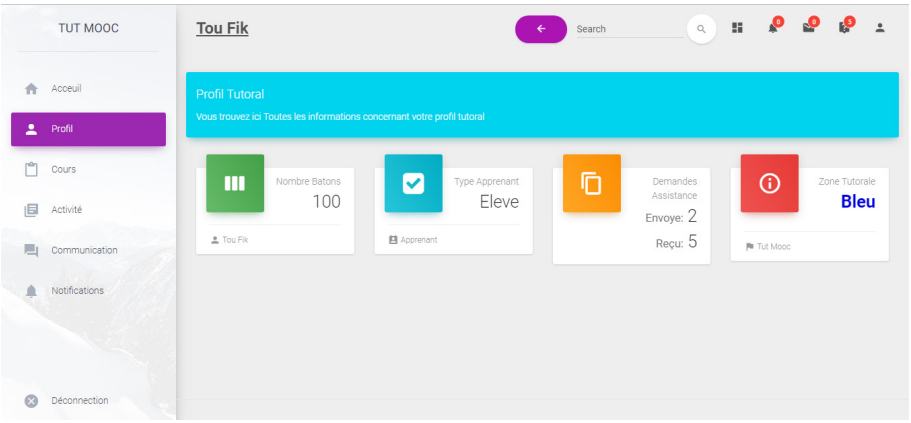


Figure 3. Tutorial profile of a learner in TutMOOC system

3.2. Detection of the critical intervention zone

This phase makes it possible to detect if the learner needs a tutorial intervention through the detection of a tutorial intervention zone. These areas are detected from the traces left by the learners by carrying out the different pedagogical activities (learning, assessment, tutoring ...) as well as the interactions done between them. These areas indicate the degree of need for the assistance of the concerned learner. Thus, this stage designates an entry state which represents the traces of the learner and an output (or the exit) one that demonstrates its exact situation or the zone of tutorial intervention.

Nevertheless, in a difficult situation, these profiles (cognitive, behavioral or tutorial) do not have the same importance. We assume that learners who have difficulties in their learning have priority over learners who have a weak tutorial profile (because of the lack of sticks for example). For this reason, we have identified two types of zones: **a primary zone and a secondary zone**. The first zone contains both Cognitive and Behavioral profiles (total activities performed on the platform), while the second one is called the secondary zone and contains the tutorial profile.

Each zone has a coefficient, which can be modified by the staff (administrator, teachers and tutors). In our case, we assigned the value “3” as a coefficient for the primary zone since this part is more important for the learner because it represents his dynamism and his learning activities, and the value “1” for the secondary zone. We used a set of mathematical formulas to calculate the zone of each learner (they will be the subject of another research paper).

For example, from the value of the cognitive profile (CP), we detect the zone and its color as follows:

- If (Cognitive_ Profile <25%) then zone: = “red”.
- If (25 % <= Cognitive_ Profile <45%) then zone: = “orange”.
- If (45% <= Cognitive_ Profile <60%) then zone: = “yellow”.
- If (60% <= Cognitive_ Profile <85%) then zone: = “green”.
- If (85% <= Cognitive_ Profile <= 100%) then zone: = “blue”.

3.3. Intelligent tutorial intervention

After determining the learner’s situation (the critical intervention area), we apply a series of tutoring rules to identify the means or tools to be adopted for social tutoring.

3.3.1. Proposed tutoring type

There are several types of tutoring, which vary according to the level of the tutor.

A. Human tutoring: it can be split up in two categories:

- **Tutoring by a teacher**: for this category, tutors are teachers. They have more knowledge with more experience than learners who are seeking tutoring. The concept of this type of tutoring is simple: we assign a teacher to each learner or group of learners who will act as their tutor. He only deals with pedagogical problems, so he cannot reach all the objectives of the tutoring.
- **Tutoring by a specialist**: Since the previous category is not enough to achieve all tutoring goals, experts suggested the creation of a special job called “specialized tutor for learner follow-up “. As a result, the tutor has a specific job (role). In this case, the tutor must undergo a training in order to be able to carry out this activity.

B. Peer Tutoring: In some cases, learners can help their peers (i.e. the tutor is another learner). This tutoring is called peer tutoring. Thus, the tutoring requester (who requests tutoring) and the tutor are both learners.

C. Semi-Collaborative Tutoring: The system assigns a main primary tutor for the learner in question. This tutor will then create a group of tutors who will work together to help the learner requesting tutoring or help. The appointed tutor (lead tutor) may contact other tutors to assist him/her with specific requests. When selecting the main tutor, the system computes a degree of freedom for each tutor and takes the tutor with the highest degree of freedom.

D. Tutoring by an Animated Conversational Agent (Automatic tutoring):

This type of tutoring is intended for learners who are not in a critical situation. It consists of selecting a set of questions predefined by the administrative staff of the system/MOOC. The answer to the question is also predefined by the same staff. It is the same principle as the FAQ (Frequently Asked Questions). In our system, we proposed reading the associated response by a lively animated conversational agent. We have offered another option to the learners in our system, which is to add the most popular questions to the database of questions that are predefined by the staff. In other words, learners can ask questions or ask for help. If the answers sent to these learners are appreciated by other learners using the statement "I like", this question can be added to the questions database so that it can be used by the animated conversational agent or the artificial tutor.

3.3.2. Tutorial rules:

In this section, we give a set of tutoring rules that can be applied depending on the critical area of intervention detected for each learner. Rules can be of the form: **If zone then tutoring_Type**. In what follows, we provide the appropriate type of tutoring detection algorithm based on the intervention zone associated with each learner seeking help (Algorithm1).

Algorithm1: Algorithm of the social adapter;

Entry: learner's-zone

Output: Type of tutoring (Type_Tut)

begin

If learner's-zone = "red" then Type_Tut: = "Semi-collaborative tutoring"

Otherwise if learner's-zone = "orange" then Type_Tut: = "Human tutoring"

Otherwise if learner's-zone = "yellow" then Type_Tut: = "Peer Tutoring"

Otherwise if learner's-zone = "green" then Type_Tut: = "Tutoring by animated conversational agents"

End.

4. Conclusion and future works

The field of learning kept evolving to include different types of supporting the learners. This evolution was highly demonstrated by MOOCs. The latter is one of the most significant systems that has seen such an interest and been the main focus of many researchers in recent years. These MOOCs seek to look for tools and techniques to improve the skills of learners and facilitate their different teaching tasks. Although it has attracted many learners' interest who considered their enrollment in these spaces highly beneficial, the system has recorded high rates of the participants' withdrawals for some unknown reasons at the beginning.

Such quitting was very disappointing for MOOC advocates. This situation has led many researchers to come up with immediate solutions or simply to try to adopt approaches already implemented in human learning environments. Indeed, some researchers have proposed techniques such as the personalization of MOOCs (Lefevre et al., 2016; Clerc et al., 2015), the adaptation of educational pathways (Sun et al., 2015; Bakki et al., 2015), motivation of learners in MOOCs (Mangenot, 2014; Hew and Cheung, 2014; Bakki et al., 2015) or in other e-learning systems, etc.

Therefore, the purpose of this research was to provide learner's support functions in such environments (i.e. tutoring), that is made throughout the learning process and based on the needs of learners and their skills and behaviors. These are represented by a learner model that encompasses a number of profiles. It is from his profiles that the degree of a learner's need for assistance is detected. This degree makes it possible to identify a tutorial intervention zone that will be used to designate the type of tutoring offered to the concerned learner. In our proposal, we identified four zones of tutorial intervention. Each zone is associated with a different color and a type of tutoring. For example, if the detected situation of the learner is very serious (close to quitting), tutoring by a senior tutor (experienced) is recommended. Therefore, depending on the situation of the learner (i.e. his tutorial zone), a type of tutoring

is proposed. Special emphasis has been placed on a special type of tutoring, which is the peer-tutoring. This activity can be done by the learners themselves under certain conditions.

Another type of offered tutoring is the artificial tutoring. It is dedicated to situations that are not serious. It holds a set of pre-set questions with their answers. In such cases, the nature of the questions differs, and most of the times according to the learners' opinions. They can be added with their best answers to the database of the questions used by the artificial tutor.

All these ideas have been implemented in a system dedicated to learning the concepts of the "algorithmic" subject. This system offers several services facilitating the tasks of the various human actors namely: teachers, tutors and learners. In addition, it has several features related to the management of different learner's profiles, the safeguarding of tutorial zones, the management of various educational resources, the follow-up of tutorial assistance requests, etc. Special attention has been given to the peer-tutoring process. Indeed, a particular management is associated with the requests for assistance issued by the learners to the learners-tutors as well as the answers and the announced appreciations.

Finally, through the carried out experimentation on the implemented system, several remarks and suggestions were made. We further seek to improve this system, and we suggest adding a learner annotations tracking tool, customizing and proposing an alternative algorithmic tool to predict at least the learners' withdrawal from their tracks just before completing their training.

References

Bakki, A., Oubahssi, L., Cherkaoui, C., George, S., & Mammass, D. (2015). MOOC: Assister les enseignants dans l'intégration des ressorts de motivation dans les scénarios pédagogiques. 7éme Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2015), Agadir, Morocco. pp.450-452.

Bendjebar, S., Lafifi, Y., Zedadra, A. (2016). Automatic Detection of Tutoring Styles Based on Tutors' Behavior. International Journal of Distance Education Technologies, IJDET, 14(2): 79-97

Boyatt, R., Joy, M., Rocks, C., Sinclair, J. (2014). What (use) is a MOOC? L. Uden et al. (eds.), The 2nd International Workshop on Learning Technology for Education in Cloud. Springer Proceedings in Complexity. Springer, Dordrecht

Clerc, F., Lefevre, M., Guin, N., & Marty, J. C. (2015). Mise en place de la personnalisation dans le cadre des MOOCs. In 7ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2015) (pp. 144-155), Agadir, Morocco.

Collet, C., Seereekissoon, R., Abotsi, I., Michaud-Maret, M., Scius-Bertrand, A., Tillich, E. and Parrend, P. (2017). POEM-COPA Collaborative Open Peer Assessment. In: Bourgine P., Collet P., Parrend P. (eds.) First Complex Systems Digital Campus World E-Conference 2015. Springer Proceedings in Complexity. Springer, Cham.

Jordan, K. (2014). Initial Trends in Enrolment and Completion of Massive Open Online Courses. International Review of Research in Open and Distance Learning, 15(1), 133–160.

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, *12*, pp. 45-58., 2014

Khalil, M., & Ebner, M. (2016). What can Massive Open Online Course (MOOC) Stakeholders Learn from Learning Analytics? Learning, Design, and Technology. An International Compendium of Theory, Research, Practice, and Policy. Springer. Accepted, in print.

Lafifi, Y., Azzouz, K., Faci, H., Herkas, W. (2010). Dynamic Management of Tutor's Roles in an Online Learning System. International Journal of Learning Technology (IJLT), InderScience Publication, Vol 5, N 2, pp. 103-129.

Lefevre, M., Guin, N., Marty, J. C., & Clerc, F. (2016). Supporting Teaching Teams in Personalizing MOOCs Course Paths. In *European Conference on Technology Enhanced Learning* (pp. 605-609). Springer International Publishing. 2016.

Mangenot, F. (2014). MOOC: hypothèses sur l'engouement pour un objet mal identifié. *Distances et médiations des savoirs. Distance and Mediation of Knowledge*, *2*(7). 2014.

Zhang, Q., Peck, K.L., Hristova, A. (2016). Exploring the communication preferences of MOOC learners and the value of preference-based groups: Is grouping enough?. Education Tech Research Dev, 64(4): 809. doi:10.1007/s11423-016-9439-

CHAPTER 11

SMART HOUSE: DATA GATHERING AND ANALYSIS

Natalija LEPKOVA*

*Associate Professor, Vilnius Gediminas Technical University, Faculty of Civil Engineering, Department of Construction Management and Real Estate, Vilnius, Lithuania

e-mail: natalija.lepkova@vgtu.lt

DOI: 10.26650/B/ET06.2020.011.11

Abstract

In modern society, the concept of "smart house" is increasingly being heard. At present, it is generally acceptable that a smart house has efficient building management, local management and business management systems. A smart house increases the business value of the environment created by the adaptability and flexibility provided by the location and the communication systems. There are many opinions on how we should understand the concept of a smart house. Some people believe this is a modern home audience, others think it's a fully designed home cable system. There are some who guess that this reflects modern telecommunication systems, etc. Everything that has been mentioned really reflects only part of the "smart home" possibilities. Smart House introduces a modern, robust automated system that allows to integrate all of the main operating subsystems such as: energy supply, supply of gas and water, lighting system, heating systems, microclimate systems, other remote controls. The smart houses are often pointed as one of the main constituents of smarter living environments.

The chapter provides the smart house definition, criteria defining smart building, smart house technology explanation, examples of smart houses in different countries, smart house data model, building progress and analysis of smart building automation and control systems (applying SWOT analysis method).

Keywords: Smart house, Smart house technology, Smart house data model

Introduction

1. Smart House Definition

In modern society, the concept of "smart house" is increasingly being heard (from "Smart Home" to"Intelligent Building").

At present, it is generally acceptable that a smart house has efficient building management, local management and business management systems. A smart house increases the business value of the environment created by the adaptability and flexibility provided by the location and the communication systems.

There are many opinions on how we should understand the concept of a smart house. Some people believe this is a modern home audience, others think it's a fully designed home cable system. There are some who guess that this reflects modern telecommunication systems, etc. Everything that has been mentioned really reflects only part of the "smart home" possibilities. Smart House introduces a modern, robust automated system that allows to integrate all of the main operating subsystems:

- Energy supply;
- Supply of gas and water;
- Lighting system;
- Heating systems;
- Microclimate systems;
- Other remote controls.

The term "intelligent building" was first used by UTBS Corporation (United Technology Building Systems Corporation) in 1981 in the USA. About 2 years later, their efforts became a reality and the City Place Building in Hartford (Connecticut, USA) was named as the world's first intelligent building (Azari et al., 2016). Since then, various definitions have been proposed for intelligent buildings. The initial definitions only focused on the technological aspects without taking the user's requirements into account (Powell, 1990).

The word "smart" has recently become an umbrella term for innovative technology that possesses some degree of artificial intelligence. The key attributes of smart technology are the ability to acquire information from the surrounding environment and react accordingly. The long-term objective of smart technology is to improve the well-being of people and as

such it has become the backbone for such an innovative concept as the “smart home” (Marikyan et al., 2019).

A “smart home” is a residence equipped with smart technologies aimed at providing tailored services for users. Smart technologies make it possible to monitor, control and support residents, which can enhance the quality of life and promote independent living (Marikyan et al., 2019).

Smart houses are often pointed as one of the main constituents of smarter living environments (GhaffarianHoseini et al., 2013).

An “intelligent building” combines innovations and technology with skillful management to maximise return on investment (Clements-Croome, 2004).

With the operating costs of a non-domestic building being significant when compared to the capital cost and a “shifting culture towards value rather than initial cost” (Clements-Croome, 2011) it is suggested that a more suitable representation of this driver would be its ability to maintain value over a long period of time under changing use and external conditions; its longevity. Therefore the three distinct drivers for building progression are: (1) longevity; (2) energy and efficiency; and (3) comfort and satisfaction (Buckman et al., 2014).

Although the possibilities are endless when it comes to smart technology and home automation, every smart home is composed of one or more of the 5 elements:

- Energy
- Security
- Atmosphere
- Convenience
- Entertainment

These elements represent the areas of our homes and our lives that can benefit from smarter technology. The perfect smart home incorporates and seamlessly integrates all 5 elements (5 elements of smart homes, 2019).

Criteria defining smart building presented in Table 1.

Table 1. Criteria defining smart building (based on Azari *et al.,* 2016)

Safety and security		**Comfort**			**Environment and energy**	
Safety	**Security**	**Indoor quality**	**User Comfort**	**Response ability**	**Energy factors**	**Environmental Factors**
1) Compliance with regulations; 2) Emergency escape capability; 3) Time for total egress; 4) Fire detection and fighting; 5) Earthquake monitoring; 6) Wind load monitoring; 7) Structural monitoring	1) Area under supervision and monitoring (Number of input and output covered by cameras) 2) Automatic and remote control or monitoring	1) Indoor air quality; 2) Occupant wellbeing and health; 3) Domestic hot water supply; 4) Gas supply; 5) Amount of fresh air; 6) Daylight utilization; 7) Odor level	1) Easy access for installation, maintenance; 2) Ease of control; 3) Acoustical comfort; 4) Thermal comfort; 5) Visual comfort	1) Response to change in temperature; 2) Response to change in sunlight	1) Total energy consumption; 2) Energy saving & conservation; 3) Energy recovery; 4) Energy regeneration; 5) Pollution related to fuel consumption	1) Method of cooling, heating and ventilation; 2) Natural ventilation possibility; 3) Natural light usage possibility; 4) Ease of cleaning

Table 2 presents the factors measuring the building intelligence.

Table 2. Factors measuring the building intelligence (Clements-Croome, 2004).

Fluid intelligence factors (owners/landlords)	**Crystallised intelligence factors (occupiers/tenants)**
General function measure (durability; environmental resource; reliability; response to environmental change)	Specific function measure (how well does a building fulfil the client's brief for specific occupiers)
Adaptability (ease of change of use)	Flexibility (ability to respond to short-term change demands of occupants)
Capital utilization (initial cost of construction)	Fixed asset run rate (fixed occupancy costs)
Environmental impact (energy, water, waste, pollution)	Efficiency (outputs/inputs ratios; service charges)
Social impact (parking; access; safety)	Effectiveness (benefits and increased value from occupier satisfaction)

2. Smart home technology

SMART home technology uses devices connected to the Internet of things (IoT) to automate and monitor in-home systems. It stands for Self-Monitoring Analysis and Reporting Technology. The technology was originally developed by IBM and was referred to as Predictive failure analysis. The first contemporary SMART home technology products

became available to consumers between 1998 and the early 2000s. SMART home technology allows users to control and monitor their connected home devices from SMART home apps, smartphones, or other networked devices. Users can remotely control connected home systems whether they are home or away. This allows for more efficient energy and electric use as well as ensuring your home is secure. SMART home technology contributes to health and well-being enhancement by accommodating people with special needs, especially older people. SMART home technology is now being used to create SMART cities. A Smart city functions similar to a SMART home, where systems are monitored to more efficiently run the cities and save money (Smart home technology, 2019).

SMART home technology devices can range in the following: Wireless speaker systems; Thermostats; Home security & monitoring systems; Domestic robots; Smoke/CO detectors; Lighting; Home energy use monitors; Door locks; Refrigerators; Laundry machines; Water detectors.

As of 2015, the most common piece of SMART home technology in the United States were wireless speaker systems with 17 percent of people having one or more. SMART thermostats were the second most prevalent piece of SMART home technology with 11 percent of people using the device. A 2012 consumer report that pulled data from the National Association of Home Builders looked for what SMART home devices homeowners wanted most and found that top five were wireless security systems (50%), programmable thermostats (47%), security cameras (40%), lighting control systems and wireless home audio systems (39%), and home theater and multi-zone HVAC systems (37%) (Smart home technology, 2019).

3. Smart house data model

Data collection and abundance depend on the number and complexity of the systems installed in the building. Figure 1 presents the building progress and the difference can be seen between primitive buildings and possibility to control them and smart buildings and their systems.

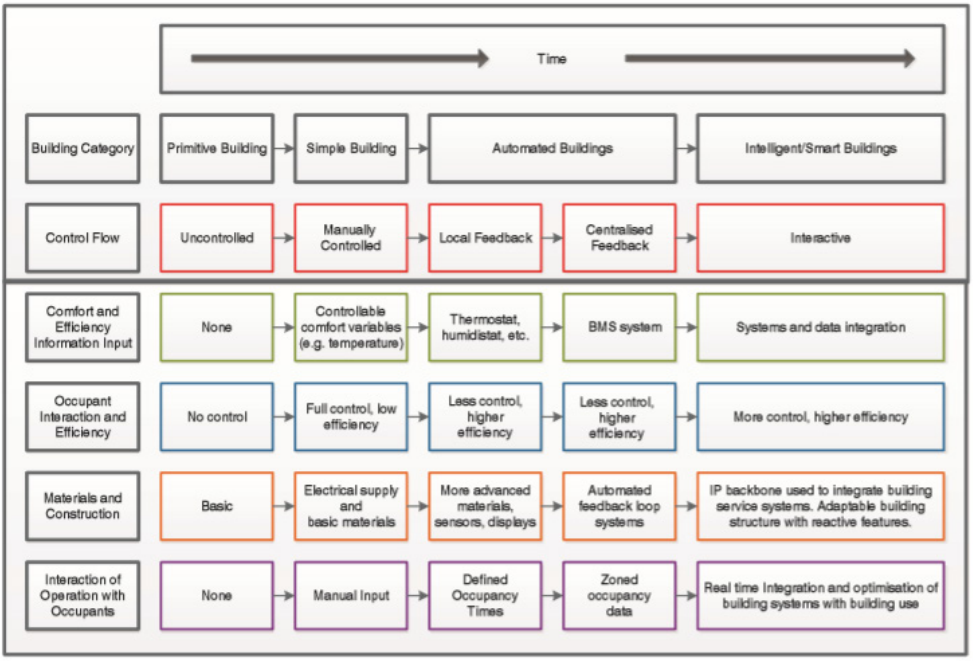


Figure 1: Building progress (Buckman *et al.*, 2014)

Smart houses have a lot of data. The main problem is to gather that data and to respond quickly to any requirements of the buildings users or failures of the systems.

The data model (presented in Table 3) contains information, which can be divided to main 3 parts: information about the building (structures, materials, systems and their locations), information about user (location, activity in the building, age and etc.) and information about object (located in building: furniture, appliance and etc.).

Table 3. Data model of building (based on Lertlakkhanakul *et. al.*, 2008)

Building data	**User data**	**Object data**
Building (ID, domain, site)	User: name, location, age, sex	Furniture: location, orientation, property
Building plan (story): wall (ID, type, surface); column (ID, location, type, width, height); beam (ID, type, width, height); slab (ID, are, type); stair (type, start point, ID, angle); space (ID, type, location, floor, ceiling, surface)	Command list: activity, command (object, function, parameter)	Appliance: location, orientation, property
Opening in wall: door (location, orientation, property, behaviors), window (location, orientation, property, behaviors), hole door (location, orientation, property, behaviors)	Context: space, time, user, object (ID, space, event (ID, event), users, functions (ID, function), event.	Light: location, orientation, property

The information provided in Table 3 can be received from the building BIM model (if during the construction such a model was created). Also the information can be received from motion detectors, sensors and other devices. For receiving and managing mentioned information in Table 3, it is necessary to use special software. Software defined smart building is shown in Table 4.

Table 4. Software defined smart building (created by author based on Younus *et al.,* 2019).

Object of identification	**Software and devices for identification**
Energy consumption and supply	Energy management and supply software
Water consumption	Water management software
HVAS systems	HVAS maintenance services
Fire	Fire alarm
Light consumption	Lighting sensor
Indoor air quality	Indoor air quality sensors
Security	Security camera
Enterprise systems	Enterprise systems integration software
Intrusion	Intrusion detection sensors
Sink	Sink node detectors

4. Smart house examples

4.1. Smart house examples in Lithuania

In Lithuania the smart house concept is rather new. The first multi-apartment building in Lithuania, which recognizes residents by fingerprints, emerged in Vilnius Antakalnis district (the title of project "Rūtų 21", completed in 2018, see Figure 2). The house is equipped with heating, lighting and security automation systems and is heated by geothermal energy. The project "Rūtų 21" is not the first smart residential building in Lithuania. Smart home automation solutions are offered by the "Penki Kontinentai" group in the "Loft Town" project in Vilnius (capital of Lithuania). There are also several other real estate developers. However, the house on Antakalnis Rūtų Street is the first apartment building in Lithuania with a biometric entrance control system, which is installed on both the main and all apartment doors. So far, biometric access control systems based on face, iris, or fingerprint scanning technology have generally been used in offices, banks and other commercial premises. In the housing market, biometric technologies are just coming (Degutis, 2018).

Figure 2: Rūtų 21 project façade (Baltic Sothebys realty, 2019)

Fingerprints. The "Rūtų 21" project uses an access control system based on the scanning technology of the papillary lines (lines visible to the fingerprints, see Figure 3). However, it works differently from fingerprint readers on phones. Darius Krasinskas, Head of Sales at Šviesos Studija, said it used a solution from the Austrian company Ekey Biometric Systems GmbH, which translates fingerprints into a standard binary system for computers. In other words, units and zeros. The prints themselves, according to him, are not saved anywhere, so there should be no question of the EU General Data Protection Regulation (which appeared in May of 2018). In this case, only those binary system codes that are still encrypted are stored. When a user adds the finger, the ekey system scans the human papillary lines, turns them into code, aligns with the codes already in the system, and unlocks the door if a match is found. Outdoor and apartment doors are unlocked with the owner's finger. And all access rights are controlled by phone (see Figure 4). Every owner of the apartment can install a gadget on the phone and independently manage access rights. For example, it is possible to enter the nanny's or cleaner's fingerprint data into the system and to allow them to enter the house only at specific hours. The *ekey system* is much safer and better than the ones used in phones and some sports clubs (Degutis, 2018).

Figure 3: Fingerprints, the "Rūtų 21" project (Degutis, 2018).

Remote control. The house is equipped with an electric car charging station and lighting, security and heating automation solutions are installed. The building is not connected to the city's heat networks - it is drilled with ten 100 meter deep wells and a geothermal heating recuperation system with air cooling equipment. The basement is equipped with a Danish Danfoss smart heating system that adapts itself to the ambient temperature and can be remotely adjusted. The apartment-mounted thermostats can also be operated remotely. Additional preferred automation solutions, such as automatic blinds or curtain control, the ability to turn on or off household appliances or power outlets at a distance, or to receive a message about an open window, must be ordered by customers. The building is also equipped with a semi-intelligent security system consisting of camera and motion sensors. The cameras turn on only when the latter are triggered. This greatly facilitates work for security company staff, especially at night. The apartment has a total of 7 luxury apartments ranging from 120 square meters. m. to over 200 sq. m. Apartment prices range from 750,000 to 1.3 million Eur. (range from 4.5 to 5 thousand euros per sq.m). The customer of the project "Rūta 21" is UAB "Promo Vision". The construction manager and technical supervisor (main contractor) are "Keista" UAB, designed by "Vilniaus architektūros studija". Smart home solutions were implemented by the "Šviesos studija" company, operating under the brand Think Light (Degutis, 2018).

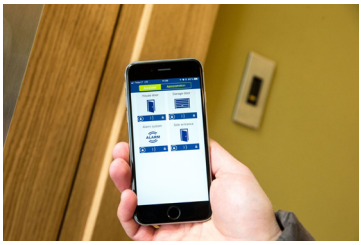


Figure 4: Mobile remote control application, the "Rūtų 21" project (Degutis, 2018).

Smart house systems and their elements can be found in different buildings in Lithuania such as: arenas, library, hotels, parking lots. For example:

- Hotels: Holiday Inn Vilnius, Crowne Plaza Vilnius and etc.;
- Sports and Entertainment Center Siemens Arena, Forum Palace in Vilnius;
- Akropolis shopping centers;
- office building: Quadrum in Vilnius and etc.;
- M. Mažvydas Library (BMS) in Vilnius and others.

4.2. Smart house examples in different countries

Examples of active smart homes are The Aware Home, Gator Tech Smart Home, Toyota Dream House Papi and NICT's Ubiquitous Home are shown in Figure 5 (Lertlakkhanakul *et al.,* 2008).

For active smart homes, The Aware Home is one of the first generation laboratory houses for the elderly developed at Georgia Institute of Technology. The research home was simultaneously inhabited by elderly people as well as tested and monitored by researchers. The research goal was to apply ubiquitous computing for everyday activities. Another similar project is Gator Tech Smart House developed by Mobile and Pervasive Computing Laboratory at University of Florida. With extensible technology based on OSGi framework, the goal of this context-aware home was to create an 'off-the shelf' smart house which the average user can buy, install, and monitor without the aid of engineers. Compared with The Aware Home, Gator Tech Smart House is more appliance-oriented. Various smart functions for smart home appliances, home security system and home assistant service have been developed. In Japan, the same movement in context-aware home has been well recognized at Toyota Dream House Papi (Lertlakkhanakul et al., 2008).

Figure 5: Physical Smart Host Test-beds: (A) The Aware Home, (B) Gator Tech Smart House, (C) NICT Ubiquitous Home, (D) Toyota Dream Home Papi (GhaffarianHoseini *et al.,* 2013; Lertlakkhanakul *et al.,* 2008).

Another example of a smart house is R128 house in Stuttgart (Germany). House R 128 (Sobek House) is a modernist single-family house in Stuttgart, Germany, built by architect Werner Sobek in 1999/2000 (Trulove, 2003). This four-level residence is sited on a steep parcel of land (see Figure 6). Access via a bridge leads to the top floor. The house features a modular and recyclable design, is completely glazed and has no interior dividing walls. It is computerized and meets its own energy requirements completely (House R128).

The building is shaped like a cube, has four levels and is wrapped by a glass shield. All components can be segregated for recycling. There are no walls or closed rooms (apart from the bathrooms) and only a few pieces of furniture. The transparency is supposed to create the impression that one lives outdoors exposed to nature. Owing to its passive solar architecture with triple-glazed walls, the house needs no energy for heating. Seasonal temperature shifts are balanced by a seasonal thermal energy store. Electricity is generated by solar cells. Every item in the house is computer-controlled (House R128). Using advanced smart technology, the house is devoid of switches, door handles, and other such fittings normally associated with comfortable residential living. Various functions in the house are controlled using non-touch sensors, voice control, or touch screens. Operations such as controlling lights, opening and closing windows, watering the garden and setting room temperatures use especially developed house control software. Activation is by means of touch screens provided on each floor and in the guest annex. Bathrooms and toilets feature reflecting photocell units permitting a swipe of the hand to control opening and closing of doors and flushing of toilets. Showers and washbasins also have non-touch controls. The refrigerator has a concealed microwave sensor that responds to close-up hand movements to open and close the door (Trulove, 2003).

Figure 6: Smart house R128 in Stuttgart (Germany) (Trulove, 2003)

5. Smart house adaptation to the needs of disabled people

If a person is at home and for some time, does not take coffee or no longer switches on the light in the toilet, as they usually do, the smart house sends this information-warning to social workers who can call and ask the resident of the house whether everything is fine, without causing any inconvenience to anyone. And only after a person does not answer the call several times does the social worker go to the disabled person's home. Motion sensors will turn on the lights when the inhabitant is indoors and switch off when the person leaves. Climate control installed in the smart home will ensure that the air is at the right temperature and will not dry out, and that the rooms are well ventilated - the automated houses even know how many people are in the room and what kind of climate they need to make them feel comfortable.

6. Analysis of smart building automation and control systems

Based on literature analysis and collaboration with smart house companies the SWOT analysis of smart building automation and control systems is presented in Table 5.

Table 5. SWOT Analysis of Smart Building Automation and Control Systems	
Strengths of Smart Building Automation and Control Systems	**Weaknesses of Smart Building Automation and Control Systems**
1) efficient use of energy and other resources in buildings; 2) user-friendly work environment is created for building users; 3) Reducing the operating and maintenance costs of the building; 4) Effective functioning of interconnected building engineering and other systems; 5) Safety of building users is ensured (fire detector, leakage of water or gas)	1) inadequate customer requirements, ignorance; 2) inadequate functioning of the system due to lack of qualified specialists; 3) system failures due to improperly selected equipment and defective materials; 4) difficult to reconcile equipment from different manufacturers; 5) lack of communication between stakeholders (inconsistencies of opinions); 6) human factor errors
Possibilities of Smart Building Automation and Control Systems	**Threats of Smart Building Automation and Control Systems**
1) more accurate, efficient, cost-effective system management due to technological progress; 2) competition between customers (building users) leads to the installation of intelligent building automation and management systems in newly built buildings; 3) unlimited functionality of intelligent building automation and control system; 4) one user interface on different devices	1) reduced system security due to intrusion into system threats; 2) the lack of a common definition (clarity) of the intelligent building automation and management system; 3) the untapped potential of intelligent building automation and control systems; 4) technology aging; 5) the equipment is expensive, there are not many manufacturers, so there is limited choice

It should be noted that the functionality of smart building automation and control systems depends on customer requirements (expectations) and investment. If the customer is reluctant to invest in the capabilities provided by the Intelligent Building Automation and Management System, the customer's expectations will not be met. Therefore, it can be said that the functionality (capabilities) of intelligent building automation and control systems depends on the customer's investment. All surveyed companies say that customer expectations do not exceed the capabilities of intelligent building automation and management systems, and emphasizes that the system's functionality (capability) is limitless.

JSC "Jung Vilnius" company representative states that the customer (building user) does not make use of the possibilities offered by the facade blinds management and other intelligent building automation and control systems.

Another important moment is the customer's competence in choosing the intelligent building automation and management system. For example, if the customer's requirements (expectations) are rational, then there is a discussion between the customer and other stakeholders to best meet the customer's expectations and maximize the benefits. On the contrary, if the customer's requirements (expectations) are inadequate due to ignorance,

therefore the realization of the customer's expectations is complicated and extremely expensive.

At the stage of selecting the intelligent building automation and management system, the problem of the irrationality of the customer's requirements is encountered, and in the implementation phase with the problem of the compatibility of the equipment of different manufacturers and the errors of the human factor.

Scientific publications analyze the benefits of various technologies in the field of automation and management of buildings and the possibilities of their implementation.

In addition, the problems of today's intelligent building automation and management systems - the intrusion into the system from the outside - are particularly relevant.

To sum up, the use of innovation is indispensable because of the untapped potential of intelligent building automation and management systems. The threat of intrusion into intelligent building automation and management systems must be eliminated by investing in cyber security.

Conclusions

The main benefits of using smart house technology are:

1. Security. Using current technology, electricity is supplied to all homes in order to be placed on the market. If user cut down set or need to be encrypted. Meanwhile, the "smart house" must contain information about the product that is on the market and the "smart home". Additionally, maybe sensors detect leakage of gas and water.
2. Comfort. Smart Home Technology supplies power at different voltage levels and ensures a comfortable environment (temperature, humidity and etc.). This technology can be automatically controlled.
3. Economy. Smart Home Technology controls the power supply for each device. Traditionally, each device is provided with usage information. The "smart home" controller can be installed to use tools to use them for use. Such regulation will guarantee lower operating costs.

References

Azari, K. T., Asadian, E., Ardebili, A. V. (2016). Evaluation of Multi-criteria Selection Factors of Intelligent Buildings. *2nd International Congress of Technology, Management and Social Sciences-16 (ICTMS-16), An International Journal of Engineering Sciences*, Special Issue ICTMS-16, 31–37.

Baltic Sothebys Realty. (2019). Rūtų 21 project. Internet access: https://lt.balticsothebysrealty.com/nekilnojamasis-turtas/parduodamas-butas-vilniaus-apskritis-vilnius-rutu-g-166674/

Buckman, A. H., Stephen, M. M., Beck, B. M. (2014). What is a Smart Building? *Smart and Sustainable Built Environment 3(2),* 92–109. https://doi.org/10.1108/SASBE-01-2014-0003

Clements-Croome, D. (Editor). (2004). Intelligent buildings: design, management and operation. London: ThomasTelford, 408 p.

Clements-Croome, D. (2011). Sustainable intelligent buildings for people: a review. *Intelligent Buildings International*, 3(2), pp. 67-86.

Degutis, G. (2018). Offers an apartment for a million in a smart home (in Lithuianian). Published 2018-03-10. Internet access: https://www.vz.lt/informacines-technologijos-telekomunikacijos/2018/03/10/siulo-buta-uz-milijona-ismaniame-name#22

GhaffarianHoseini, A., Dahlan, N.D., Berardi, U., GhaffarianHoseini, A., Makaremi, N. (2013). The essence of future smarthouses: From embedding ICT to adapting to sustainability principles. *Renewable and Sustainable Energy Reviews, 24*, 593–607. http://dx.doi.org/10.1016/j.rser.2013.02.032

House R128. Wikipedia. Internet access: https://en.wikipedia.org/wiki/House_R_128

Marikyan, D., Papagiannidis, S., Alamanos, E. (2019). A systematic review of the smart home literature: A user perspective.

Lertlakkhanakul, J., Choi, J.W., Kim, M.Y. (2008). Building data model and simulation platform for spatial interaction management in smart home. *Automation in Construction, 17*, 948–957. doi:10.1016/j.autcon.2008.03.004

Powell, J.A. Intelligent design teams design intelligent building. (1990). *Habitat International, 14,* 83–94. https://doi.org/10.1016/0197-3975(90)90038-3

5 elements pf smart homes. (2019). Internet access: http://www.modernsmarthome.com/the-5-elements-of-smart-homes/

Smart home technology (2019). Wikipedia. Internet access: https://en.wikipedia.org/wiki/SMART_home_technology, viewed 2019 March 14.

Trulove, J. G. 2003. The Smart House. Harper Collins Publishers, 192 p.

Younus, M.U., Islam, S., Ali, I., Khan, S., Khan, M.K. (2019). A survey on software defined networking enabled smart buildings: Architecture, challenges and use cases. *Journal of Network and Computer Applications*, *137,* 62–77.

CHAPTER 12

PRIVACY FOR ENTERPRISES IN THE DATA AGE

Bilgin METİN* Enes YILMAZ*, Erdi ŞEKERCİLER*

*Bogazici University, Department of Management Information Systems, Istanbul, Turkey
e-mail:bilgin.metin@boun.edu.tr

DOI: 10.26650/B/ET06.2020.011.12

Abstract

The world we live in is now becoming increasingly virtual. We all interact with this new age which we can describe as the digital age. We shop online, we communicate with people via social media, we are informed at any time through the devices that are in our hands about goings-on, whether we like it or not, we have become a part of this globalized and digitalized world. Data can be described as the structure of the digitalized world. In each interaction between us and the tools which we use, we create data or we cause data transferring or we can be a small part of a large data collection because of our presence in a platform on the internet. Certainly, this close relationship can reveal our private life in some situations. Most of the time, we are exposed to situations where our private information is collected, used, and processed without our permission. Sometimes we cannot even notice the violation of one of the most fundamental rights and freedoms we can define as privacy. This literature survey study is based on the fundamentals of information security, and it seeks answers to these questions: Why does our personal information need protection? What kind of information should be protected? What is the situation regarding the data privacy in Turkish and world law? What kind of laws have been passed upon the privacy of tax from past to today? What are the perspectives, opinions on protection of personal data in Turkey and Europe? What is the importance of data privacy for the business sectors? We also believe that this study will raise awareness on this matter.

Keywords: Privacy, Cyber security, Digital transformation

Introduction

1. What is Privacy? and Why We Need It

In today's digital world, the meaning of treasure has changed. 100 years ago, the notion of treasure was gold, petrol or maybe being a landowner. The transformation of the world in the last decades has affected our values, our standards, and concepts. Today, there are more valuable things than gold or petrol. Today's treasure is data (deMontjoye, Wang, Pentland, Anh & Datta, 2012).

Almost all the enterprises, institutions in various sectors such as the finance sector, health sector, energy sector use modern computerized techniques for digitalization based on collecting data. We are used to interact with these information systems every day. For example, while shopping, and in the hospitals, our personal information is collected. While we download a smart phones application, some personal access permissions should be allowed. While popular applications of digital life such as mobile banking, e-commerce, mobile signature, etc. make life easier for us, they let public or private organizations to observe our data. Companies, banks, hospitals, and government institutions log our behaviors, movements and store our private data, but they are expected to be careful and respectful of the privacy.

Privacy is one of the fundamental rights included in most of the contracts, agreements and countries' constitutions (Dülger, 2018). It aims to protect information which defines who we are, what we do, what we think and what we believe (Bignami, 2007), (Dumortier & Goemans, 2000). Privacy/Confidentiality keeps information that is conducted from data processing or correspondence between involved parties during an operation wanted to be hidden from unrelated third parties. Privacy creates a strict line between our personal life and the world, so it provides personal freedom/liberty and self-respect by blocking excessive interference of others. McFarland (2012), and Phelps et al. (2000) emphasize that privacy is a term that contains four dimensions: (1) intrusion (invading a person's loneliness), (2) disclosure (publicly revealing private facts), (3) false light (false public portrayal) and (4) appropriation (using personal identity without permission). If our privacy is protected, we have control on who has access to our personal life, such as our secrets, locations, telephone number and credentials. We can reduce publicly known private information and protect this information from unauthorized/illegal use of power as a consequence of privacy (Wagner DeCew, 1986).

Technology has improved dramatically in the last two or three decades, with the Internet becoming an important part of our lives. We witness novel technological infrastructures and devices, like IoT improved with the help of modern information technology systems such as ERP and cloud systems every passing day. With these technological improvements, the world has become more and more connected and digitalized, enabling people and organizations to reach their data from anywhere easily and collect others' information with new systems and applications (Chiper Cloud, 2015).

New technological devices and applications provide an opportunity to collect other people's personal information without their knowledge. For example, most of the people use smart devices like mobile phones, and some applications and features provided by them such as messages, shopping online, Google search, and calendar tasks. It is impossible to keep pace with the digitalizing world without using one of these functions. In normal circumstances, people set passwords to their devices in order to protect personal information confidentiality (Hoven et al., 2014). However, many powerful organizations such as Google and Facebook collect vast amount of information about us like location, profile, financial data, habits, e-mail, health situation, psychometric and political interests. Even if all this information such as political interest or health situation is not collected directly, this data can be obtained by combining different data collected via cookies. Therefore, it can be said that our control is low over which information is collected about us (de Montjoye et al., 2012).

With new laws, technologies and leaked information, the government has legal power to watch not only terrorists, but also all citizens. This monitoring activity encompasses call records/history, internet surfing, e-mails, social networking accounts, etc. (Bignami, 2007; Privacy International, 2016). Also, some powerful organizations bend the law in order to collect our private information. This situation causes a debate about an individual right to privacy versus the financial interest of corporations and the security concerns of government. Even worse, although people are the main related party, they barely know this issue because it is conducted in a confidential manner. Companies push the limits of the law regarding processing personal data in order to gain extra benefits, such as increasing their profit and the number of active customers. Companies need to offer their customers more customized campaigns and opportunities than other companies to be able to achieve this goal. Taking into consideration the fact that companies need even a piece of information related to a particular person, and sole information or anonymized information are useless to provide personalized offers, they need to monitor, brand, categorize and profile this data (Privacy International, 2016).

Technological developments have a lot of advantages for people and organizations, taking into consideration the fact that they make our lives easier. However, they also carry high risk potential when taken into account the increasing number of cyber-attacks to corporations like banks and telecommunication companies which have huge amounts of valuable personal data including individual's name, address, social security number, date of birth, alien registration number, taxpayer identification number, government passport number, driver's license information, mother's maiden name, or biometric information. This data can be used for identity theft, which is one of the popular cybercrimes in the digitalizing world (Chiper Cloud, 2015), (Allison et al., 2005).

Businesses using consumers' personal information may lead to competitive advantages over other companies. For example, financial data, search background in the company website, and purchase habits, are used to provide a better and targeted service (Dülger, 2019a). However, they also use this information out of their business scope. Note only businesses, but with the help of new technologies, now everyone can collect personal data. Since everyone gathers data, everyone can become a potential target, and this situation negatively affects the data privacy. Therefore, data security is a subject that needs to be argued, and we should make a distinction between data security and data privacy. These two are used as synonyms, however, these are just related concepts. Data security is a policy to ensure data privacy. Data security is about the confidentiality, availability and integrity of data, namely, it keeps data accurate and reliable, and it contains processes and implementations to ensure the data is not used by unknown individuals (CSX, 2015). Data security is also engaged in making plans, that is, gathering required information, keeping information and deleting information, so it helps to obey the legal restrictions. As for that, data privacy is the usage of data in an appropriate way. There are some legal obligations about data privacy. For example, usage of private information must be on an agreement with a company and the owner of the information. The information also cannot be sold and disclosed without permission. Therefore, industries, enterprises or individuals working with data must have a real data security policy to provide privacy of the data. When we provide a lot of personal data to these companies and other third parties, unwarranted disclosure of the data has the potential risk to be victim of cybercrime (O'Brien, 2019).

Some important companies have been hit with a series of security breaches over the past year. For example, the cyberattack on the Marriott hotel chain that collected personal details of roughly 500 million guests was exposed (New York Times, 2018). Furthermore, more than 540 million records of Facebook users were publicly exposed on Amazon's cloud computing

service. It exposed 146 gigabytes of Facebook user data, including account names, IDs and details about comments and reactions to posts (CBSnews, 2019). The main cause of this catastrophic result is unauthorized disclosure of personal information. This event increases public awareness about the importance of protecting privacy information, and people start to put pressure on the government and organizations to protect their personal information (Pascual, Marchini & Miller, 2016).

As a result of this, when we consider all these issues, privacy of personal data is not a need; it can be thought as a necessity or even an essential right regarding people, government and organizations. Therefore, a government introduces a wide range of legislation and regulation to create a balance between protecting private information and providing better service. Considering businesses reputation and trustworthiness, they also give adequate impotence to the privacy issue, so every company must develop and apply strong privacy policies and procedures (International Telecommunication Union, 2006).

2. Privacy of Personal Data in Law

Personal information has had an important place in the EU for a long time. The beginning of the data protection law can be seen in the European Convention on Human Rights which is an international agreement signed in 1950 related to the protection of private information. According to this agreement, people show respect to others' personal life, and no one can interfere with personal life except in the case of legal and democratic issues. With the advent of new technological developments, an agreement which protects individuals from processing their information automatically, was prepared in 1981. Turkey also signed this agreement to build good a relationship with EU countries (Ersoy, 2007), (Keser et al., 2014).

In the 1970s, governments used database systems to store citizens' personal information. Western European States, notably Germany, had some unpleasant experiences related tostoring and processing of personal data limitlessly (Küzeci, 2010). The first legal act considering the data protection of private data started to be implemented in the 1970s. This law mainly focuses on government establishment and some private sector organizations such as telecommunication and banks which collect confidential information of people. However, police forces/ law enforcement were outside of the scope of the law because of national security issues, so there is no restriction for the legal power to collect and monitor personal information of the citizens (Bignami, 2007). During these years, similar laws were applied in the US. The law named Fair Information Practice Principles (FIIPs) played an important role in creating privacy lines in the US in different areas, such as Health, Education and Wealth,

and also in other countries (Solove & Hartzog, 2014). In addition to these countries, Turkey also came up with the Central Population Administration System project (CPAS), which was created in 1973, and became effective in 2002. Protecting and using personal data in accordance with the purpose of collecting data can be one of the most important applications performed in this area (Ersoy, 2007).

After the advance of the internet and the development of new technologies, people have shared a lot of information in the online environment; therefore, in any criminal situation, police can use this information such as phone calls, cameras, and traffic data to locate or identify us. Considering the fact that we no longer live without using new technologies, countries have created new regulations, directives or acts which, to some extent, control the private sector, government and law enforcement.

In Europe, a second law related to data protection was the Data Protection Directive, that is the first legal act which was prepared in 1990, and became effective in 1995, to protect personal information (Bignami, 2007). The directive covers private organizations and Government Corporations which serve citizens by using their private information in order to prevent the illegal use of personal information (Bignami, 2007). The directive's main purpose is to create a common standard between EU countries considering data processes (Keser et al, 2014). Turkey has also performed similar studies about data protection during these years. When the related Commission did not complete the draft related to processing personal information automatically, the Ministry of Justice sent the Protection Personal Data Law which was prepared again, to relevant Ministries, and new resolution was sent to the prime minister's office in 2004. This resolution included a common application applied all around the world related to the Data Protection Law (Ersoy, 2007). With the 12th of September, 2010 plebiscite, a provision was added to article 20 of the constitution. According to this law, everybody has a right to have protection of personal data. Usage of this data depends on the permission of the owner of the data (Turan, 2016).

Data privacy issues have increased after the spread of the Internet and the development of new technologies like smart-phones, IoT devices, etc. New technological improvements make the data gathering, sharing, and analyzing processes easier so organizations have a substantial amount of information (Ersoy, 2007; Keser, et al., 2014). This situation makes organizations the targets of cyber criminals/hackers, and naturally, the number of cybercrimes related to data protection privacy increases (Ersoy, 2007). According to Hewlett Packard (HP) research, the number of cyber-attacks occurred in 2011 increased by 56%, and 86% of web applications are not safe, considering inscription and interface issues. Also, Symantec

research shows that cyber criminals produce more malicious code, viruses, etc. to get information from the online environment (Henkoğlu & Yılmaz, 2013).

Considering new coming threats, the Data Protection Directive which was signed with the EU, was not reliable anymore because every single country in the EU can adjust these directives according to their domestic law. Therefore, there is a need for a more common and widely accepted regulation. In 2012, the European Concil issued a new regulation applied in all the EU countries in the same way. In Europe, the protection in question is provided with two systems. The first one is the European Council. At the level of the Council, protection is provided by the European Convention on Human Rights article 8 and Section 108 of the Agreement: Right to respect for private and family life, home and correspondence. According to this Agreement, personal data should be:

- Fair and collected legally
- Collected for legal purposes and used for this purpose
- Relevant and sufficient, should not be exceeded
- Accurate and current
- Kept as long as necessary according to the collecting purpose (Dedeoğlu, 2004)

The second one is the General Data Protection Regulation (GDPR). The GDPR was adopted on 14 April 2016 from the Data Protection Directive 95/46/EC. The GDPR has been recognized as law, and became enforceable beginning on 25 May 2018.

In addition to this, in 2013, the EU stated that Turkey did not have a framework and appliance related to the privacy of personal data, and this situation had a negative impact on the relationship between Turkey and the EU (Keser et al, 2014). As a result of this situation, in 2016, the Turkish government accepted the data protection law in March 3, and published the regulation on April 07 (KVKK, 2016). With the law, the Protection of Personal data, scope, purpose and descriptions were clearly specified. Moreover, there are some principles of processes of data like deleting, destroying, anonymizing and transferring.

3. Privacy in Enterprises

The application of the personal data protection law differs in sectorial basis in some countries. Generally, many developed/developing countries such as the EU countries and Turkey use the same regulation or procedure for all sectors, so it provides a whole perspective (Strahilevitz, 2013), (Solove & Hartzog, 2014). Even though there are no sectorial rules and laws in this type of countries related to privacy concerns, governments establish different types of regulatory and supervisory authorities in order to determine the working principles of these sectors and market rules in general meaning, and enforce these conditions on specific sectors. For example, in Turkey, there is the Banking Regulation and Supervision Agency (BDDK), Capital Markets Board (SPK), Public Procurement Authority (KİK), Competition Authority (RK), Public Oversight, Accounting and Auditing Standards Board (KGMDSK), Energy Market Regulatory and Supervisory Authority (EPDK), Radio and Television Supreme Council(RTÜK), Information and Communication Technologies Authority (BTK), the Tobacco and Alcohol Market Regulatory Authority (TAPDK) and the Biosecurity Council (BK). These regulatory authorities are not for handling privacy concerns; they are just ensuring that these sectors do not act against the law within business processes (Çırakoğlu, 2016).

However, some countries like the US apply a different law and data protection act according to sectors. There are multiple laws and regulations related to how organizations protect the personal data of their customers. Although there are no laws and regulations which cover privacy of personal data, all organizations should follow their sectorial rules while they process personal data (Strahilevitz, 2013; Solove& Hartzog, 2014). This sectorial basis approach can be thought of as an inadequate application because of the lack of ability to see the whole picture, so in this approach, the country can have difficulties in taking immediate or real time actions in case of a new situation (Strahilevitz, 2013). Despite that, some others think that it is a good application of the data protection law because sector specific cases are handled easily by laws, and this type of regulation proposes customized rules and application methods for each unique sector (Nissenbaum, 2014; Schwartz, 2004). Some of the laws in force are the Health Insurance Portability and Accountability Act (HIPAA), the Gramm-Leach Bliley Act, the Fair Credit Reporting Act, the Privacy Act of 1974, the Telephone Records and Privacy Protection Act, the Electronic Communication Privacy Act, the Family Education Rights and Privacy Act, and the Payment Card Industry Data Security Standard.

For compliance with the related privacy laws, organizations should examine their risky personal data processing actions and focus on how to collect, maintain and process personal

data for mitigating or terminating these risks with a Privacy Impact Assessment (ISO/IEC 29134, 2017), (IPC, 2015), (ICO, 2015), (SEC, 2007), (HIQA, 2017).

3.1 Sectoral Appliance

3.1.1. Health

In the health sector, the usage of Information Technology (IT) is gradually increasing. Yu emphasized that Artificial intelligence (AI) is modifying medical practice using digitized data acquisition, machine learning and computing infrastructure. AI applications are expanding into areas that were previously thought to be only the province of human experts. (Yu et al., 2018)

The usage of IT provides organizations with using data more efficiently and in a meaningful way. In the US, using health data brings almost a $300 billion financial acquisition to this sector. (Kayyali et al, 2013). It does not just have a monetary advantage, but it also helps the health sector diagnose illness before the progression of the disease by using big data related to the history in the health records. Also, hospitals are connected to the same network, and doctors or physicians get a lot of information related to the same disease in order to apply the best treatment and early diagnosis, so they can have different perspectives and, if they miss something, they can capture the information by means of this system (Kayyali et al , 2013). A similar system, called epSOS (Smart Open Services for European Patients), is used in 12 EU countries (Linden, 2009).

In addition to this, the use of big data brings along a lot of developments in the medical field. For example, doctors suggest many hypotheses for the cure of different diseases. In order to find the right treatment method, they need more data and subject groups (Wang, 2018). Considering the fact that using patients' health data, which are private, has numerous advantages, it is not a surprise that the amount of information collected increases dramatically. For example, in 2012, the amount of health records in electronic systems to make data sharing easy increased from 30% to 75% (Keser et al., 2014). The increase of data sharing causes a dilemma between privacy of personal data and applying the best treatment. Therefore, governments establish some laws and regulations considering privacy concerns.

In some countries, there are specific laws related to the health sector, such as HIPAA for the US government. HIPAA basically creates limitations in sharing or using medical records collected via health providers such as hospitals and doctors. Some medical records cannot be transferred until the patient expresses consent. This law also allows health providers to keep

medical records in an electronic environment and ensures the protection of this type of data by conducting some security protocols. There is also some regional/ state legislation in order to protect health records (Solove & Hartzog, 2014).

3.1.2. Education

In education institutions, with the help of technology, the usage of data including personal data is quite high. The advance of online courses and technological applications make monitoring students and taking necessary actions related to students possible. Also, this type of education method will be useful for people with special needs. In the US, the Department of Education tries to prepare education plans by analyzing data which are collected via the online education program. In addition to this, online environment schools will able to examine teacher's performance, and this transparency will create a competitive environment so the problematic issue resulting from teachers' education methodology will be determined. As a result of this, the quality of education increases in parallel with the usage of personal data. In Turkey, there is a project called FATIH (Increase Opportunities and Technology Improvement Act). The main objective is developing an education system by analyzing students' data collected via tablet computers (Keser et al., 2014).

This type of data usage comes with the question of whether educators should collect and use data that identify students' or teachers' profiles. If data is transfered into an anonymized form, it will not be a problem, but in some situations, it can be used for different versions. Therefore, there are some laws related to this issue in almost all countries. In the US, this legislation is tied to a specific regulation called FERPA. This legislation mandates schools/ colleagues to inform students and their families in order to protect the privacy of education records. Thanks to this law, students review their records with respect to the accuracy of these records. Also, this law prevents students' records to be preserved, not only informal channels but also formal channels (Solove & Hartzog, 2014), (Walch, 2011).

3.1.3. Finance

In the finance sector, IT technology is used intentionally and collects a lot of public and personal information (Pendley, 2018). The use of IT technology and collecting personal data have a substantial amount of advantages in this sector, such as determining some fraudulent activities while analyzing the user's past habits. For example, if a user does not use his/her credit card abroad, and then someday, the user uses the credit card abroad for the first time, the bank informs the user about the expenditure and requires mobile or digital confirmation. Also, banks analyze users' data in order to offer specific promotions. With the help of this

data, banks predict some critical incidents, including problems occurring through economic crises, and they determine and take the necessary action before the incident takes place. In addition to this, they can calculate users' credibility, and determine which users are in high risk groups so they can save more money and make good investments. With the advent of mobile devices and applications, the amount of data collection has increased; therefore, analyzing users' patterns also increases and becomes more accurate. In parallel with the increase in data collection, some extra security measures are taken by authorities (Keser et al., 2014).

Cloud systems also have an important role in data collection and analyzing issue. However, in some countries like Turkey, although cloud systems are used in different areas, in the finance sector, there are some restrictions. BDDK mandates banks to hold their primary (main sources of the data are kept) and secondary (backup of the information system) systems domestically in order to keep their citizens information private. When we consider the importance and characteristic features of personal data used in the banking sector, this sector has a higher risk than others, considering cyber-attacks and the use personal data. Therefore, authorities legislate for this sector (Keser et al., 2014).

In the US, there are specific laws and regulations related to the finance sector different from the GDPR conducted in the EU. One of them is the GLBA, which covers legislations for the banking sector in order to protect customers' personal information, including financial data. According to this act, banks must inform their customers about privacy policies and procedures conducted within the bank. Also, this law provides customers with opt-out channels to some extent. Thanks to the opt-out channels, customers prevent banks from sharing their information with third parties (Code, 1999), (Solove & Hartzog, 2014).

The other law is FCRA, which covers some rules about both customers' financial information and credit information to prevent inadvertent disclosure. This law mandates banks to ask consent from their customers when sharing the financial information with third parties. In some situations, such as the processes of employment, this information is shared with some institutions (employers). Even if the user gives consent for sharing their information with third parties, employers also accept they will use this information within the scope of the legislation (Stokes, 1999).

3.1.4. Telecommunication

The intense usage of technological devices collecting personal data requires regulations and procedures for the purpose of protecting the privacy of people. In the telecommunications

sector, there is a lot of information relevant to customers such as personal information (ID, IP and address), traffic information (subscriber usage), bill information, location information, etc. Telecommunication companies use this information to provide better service to their customers, and access these people in case of an emergency using their location information. The data collected by these companies can include unnecessary information about the customer. In order to prevent this, authorities enforce some regulations. In Turkey, there is no specific law for this sector; however, there is the BTK, which protects customer's rights and information security in a general meaning (Onur, 2013).

In the US, there are the TRPPA and ECPA laws. TRPPA basically prevents telecommunications and similar companies from selling and sharing telephone records, call logs, etc., and applying monetary sentences to dissuade these institutions. There are some exceptions in the case of criminal cases for police forces (TRPPA, 2007). The other act is ECPA, which protects customers' information collected during transactions, storage such as e-mail, etc. in the electronic environment. The law mandates companies to obtain customers' consent in order to track their behavior and usage statistics (Solove & Hartzog, 2014).

3.1.5. Public/State

In public sectors, a substantial amount of data, including both public, like weather conditions and private, such as address information are collected by corporations (Aggarwal, 2019). In some countries, this information is shared with the private sector as open data in order to obtain financial earnings. However, this data sharing and processing must have some limitations and restrictions (Keser et al., 2014).

In the US, for federal agencies and public institutions, there is a specific law called the Privacy Act of 1974. With this law, the government provides protection to citizens with unauthorized disclosure of personal information compiled about a specific person by legal forces without permission of the people. Also, people are authorized to see their records in order to control whether their records are correct or not (Privacy Act, 1974).

3.2. Disclosure of Personal Information

Rapidly growing technology makes an enormous data warehouse available on the Internet. This warehouse also includes our personal information. For example, information on a Facebook account (Dülger, 2019b) almost equals the information that can be obtained by an intelligence agency with a long term study, through Twitter, it is possible to see where a person is, what he/she is doing, even what he/she is thinking (Küzeci, 2010). Some social

networking sites have more numbers than the population of most of countries. These sites provide an opportunity to access a person's information and photos, and to be seen by other people (Kılınç, 2012).

Undoubtedly, people accept to share their personal information, to record their speech and behaviors in order to make their life easier, increase their life quality and use their personal rights. This personal information can be stored, analyzed and spread. Besides, there is no way to know how this information is used. That is why the unauthorized disclosure of the information about private life arises (Kılınç, 2012), (İzgi, 2014). Various services offered on the Internet cause the disclosure of personal information publicly. Now, organizations have to take a stand against a wide range of risks and threats, like ways of e-fraud, information theft, hackers, information leakage, internal attacks, etc. (Karaarslan et al., 2010).

Rising risks and threats lead to security problems because personal information can only be declared, processed, stored and transferred within the permission of the owner of the information (Kaya, 2011). Although lots of organizations try to find a solution in terms of their perspectives, personal information can be released or stolen in many ways. Personal information can even be obtained by querying some information in the government's systems (Ketizmen & Ülküderner, 2007). Problems about the disclosure of personal information are controversial in the EU also. For example, in 2013, the disclosure of personal data of German citizens was obtained by the UK and the US intelligence agencies as one of the important propaganda materials of Merkel's opponents (Ceran, 2014).

Since information and communication technologies are commonly used in daily life, unauthorized disclosures of the information can be seen in many fields that are engaged in data, like banking, e-commercial, healthcare, education, etc. (Henkoğlu & Yılmaz, 2013). Some important private information can be gathered from various information sources. In addition to this, by using social engineering techniques, valuable information can be obtained (Ketizmen & Ülküderner, 2007). Although most of the organizations take some precautions against the leakage of information, there are deficiencies in the precautions. The most important deficiencies are firstly, technical factors and secondly, human factors (Henkoğlu & Yılmaz, 2013).

3.2.1. Classical (offline) identity theft: Identity thieves employ all kinds of methods to obtain personal data. One of the methods is parrying the security process, making use of human errors and social engineering. Influence, forcing, developing deceptive relationships can be counted as ways of social engineering. Other classical methods are dumpster diving, pretexting, shoulder surfing, skimming and business theft.

3.2.2. Online identity theft: These methods are unlimited because each passing day, a new one arises. The most important and common methods for online identity theft are:

- Malwares: Software that are installed into computers, mobile phones, and smart devices.
- Some deceptive e-mails and websites:
  - *Phishing:* Some e-mails that are disguised as coming from an enterprise, banks, government, or mirror web sites, namely a fake copy of a real website. These e-mails and websites are used for stealing users' personal information, like card information, passwords, etc.
  - *Spam:* Some e-mails come involuntarily and include harmful contents.

3.2.3. Hacking: It is another way of getting personal identity. By using hacking methods, personal information can be stolen through the system's gaps (OECD, 2008), (New York Times, 2018;) (CBSNews, 2019).

Abuse of personal data is another title for the disclosure of personal information. It is distributing and selling the commercial and occupational secrets to obtain an advantage. The data of banks, governments and hospitals have a commercially incredible value. Moreover, the release of this type of data can be a step for committing significant crimes (Karimi & Korkmaz, 2013).

4. Conclusion

Today, it is almost impossible for our information to be hidden behind closed doors. In every aspect of our lives, we encounter situations where information is recorded and used. Laws are the only regulatory factors that will prevent our personal data from being used without our permission.

With this study, we have revealed the examination of the law of data protection which has been enacted recently in our country, in terms of what its scope is, which situations are included, how it will be implemented, what it will bring, and so on. In terms of information security, this is a great development for our country to reach the standards of Europe in the data protection issue. In this context, the law on the protection of personal data, which removes many uncertainties and protects our privacy in the digital world, should be regarded as a milestone. Unfortunately, we should say that neither our people nor our organizations have the necessary and sufficient knowledge about this law and its benefits. Particularly, all

small and large organizations should be aware of this law and its sanctions. In this way, we can preclude the abuse of personal data. Although legal regulations for the protection of personal data are not strong enough to provide complete protection against rapidly evolving technology, it is the most important assurance we have to rely on.

Acknowledgement:

The authors thank Meltem Mutlutürk for her help writing the manuscript.

References

de Montjoye, Y. A., Wang, S. S., Pentland, A., Anh, D. T. T., & Datta, A. (2012). On the Trusted Use of Large-Scale Personal Data. *IEEE Data Eng. Bull.*, *35*(4), 5-8.

Bignami, F. (2007). Privacy and law enforcement in the European union: the data retention directive. Chi. J. Int'l L., 8, 233.

Dumortier, J., & Goemans, C. (2000). Data privacy and standardization. In CEN/ISSS Open Seminar on Data Protection, disponible sur https://www. law. kuleuven. be/icri/publications/90CEN-Paper. pdf.

McFarland, M. "Definitions of Privacy." Internet: www.scu.edu/ethics/focus-areas/internet-ethics/resources/what-is-privacy/, Jun. 01, 2012 [Oct. 06, 2016].

Phelps, J., Nowak, G., & Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. Journal of Public Policy & Marketing, 19(1), pp. 27-41

Wagner DeCew, J. (1986). The scope of privacy in law and ethics. Law and Philosophy, 5(2), 145-173.

Chiper Cloud, (2015). "Global guide to data protection".Internet: http://pages.ciphercloud.com/global-guide-to-data-protection-laws-landing-page.html, Nov. 20, 2015 [Oct. 02, 2016].

Hoven, J. V. D., Blaauw M., Pieters W.& Warnier M. "Privacy and Information Technology." Internet: http://plato.stanford.edu/entries/it-privacy/ , Nov. 20, 2014 [Nov.. 02, 2019].

Privacy International, (2016). "The Global Surveillance Industry". Internet: https://privacyinternational.org/explainer/1632/global-surveillance-industry, [Nov. 02, 2019].

Allison, S. F., Schuck, A. M., & Lersch, K. M. (2005). Exploring the crime of identity theft: Prevalence, clearance rates, and victim/offender characteristics. Journal of Criminal Justice, 33(1), 19-29.

O'Brien, S (2019). "The Difference Between Data Privacy and Data Security". Internet: https://blog.cygilant.com/blog/the-difference-between-data-privacy-and-data-security, Oct. 22, 2019 [Nov. 2 , 2019]

CBSnews (2019) https://www.cbsnews.com/news/millions-facebook-user-records-exposed-amazon-cloud-server/ [July 31, 2019]

International Telecommunication Union (2006) "Research on legislation in data privacy, security and the prevention of cybercrime" Place des Nations CH-1211 Geneva, Switzerland (p. 69)

Keser, L., Kaya, M. B., & Kınıkoğlu, B. (2014). Türkiye'de Kişisel Verilerin Korunmasının Hukuki ve Ekonomik Analizi. [Legal and Economic Analysis of the Personal Data Protection in Turkey] https://www.tepav.org.tr/upload/files/1421853130-9.Turkiyede_Kisisel_Verilerin_Korunmasinin_Ekonomik_ve_Hukuki_Analizi.pdf [Nov. 2, 2019]

Solove, D. J., & Hartzog, W. (2014). The FTC and the new common law of privacy. Columbia Law Review, pp. 583-676.

Ersoy, E. (2007). Gizlilik, Bireysel Haklar, Kişisel Verilerin Korunması [Privacy, Individual Rights, Protection of Personal Data]. Akademik Bilişim Konferansı 2007.

Turan M. (2016). Kişisel Verilerin Korunması [Protection of Personal Data] Türkiye Kalkınma Bankası Yayını, vol. 80, pp. 2-3 April-June, 2016

Henkoğlu, T., & Yılmaz, B. (2013). Avrupa Birliği (AB) Bilgi Güvenliği Politikaları [European Union (EU) Information Security Policies]. Türk Kütüphaneciliği, 27(3), 451-471.

Dedeoğlu, G. (2004). Gözetleme, Mahremiyet ve İnsan Onuru [Surveillance, Privacy and Human Dignity]. TBD Bilişim, 89, 36.

KVKK (2016). Türkiye'de Kişisel Verilerin Korunmasının Hukuki ve Ekonomik Analizi [Legal and Economic Analysis of the Protection of Personal Data in Turkey] Internet:http://www.resmigazete.gov.tr/, Apr. 07, 2016 [Oct. 08, 2016].

Çırakoğlu, M. (2016). Düzenleyici Ve Denetleyici Kurulların Denetlenme Şekillerinin İdari Vesayet Bakımından Değerlendirilmesi [Evaluation of the Ways of Inspection of Regulatory and Supervisory Boards in terms of Administrative Guardianship.]. Yıldırım Beyazıt Hukuk Dergisi, (2).

Strahilevitz, L. (2013). Toward a positive theory of privacy law. Harvard Law Review, 113(1).

Nissenbaum, H. (2014). Respect for Context as a Benchmark for Privacy Online: What it Is and Isn't. Cahier de prospective, 19.

Schwartz, P. M. (2004). Property, privacy, and personal data. Harvard Law Review, 2056-2128.

Walch, D. (2011). Family Educational Rights and Privacy Act. Harmony, 503, 594-6000.

Code, U. S. (1999). Gramm-Leach-Bliley Act. Gramm-Leach-Bliley Act/AHIMA, American Health Information Management Association.

Stokes, R. (1999). Fair Credit Reporting Act., internet: https://www.consumer.ftc.gov/articles/pdf-0111-fair-credit-reporting-act.pdf [Nov. 2, 2019]

Onur, A (2013). Impact of Telecominications Regulation on Data Protection. İstanbul Bilgi Üniversitesi Sosyal Bilimler Enstitüsü Bilişim ve Teknolojileri Hukuku.

Privacy Act, (1974)." Privacy Act of 1974". Internet: https://foia.state.gov/Learn/PrivacyAct.aspx, Sep. 9, 2000 [Oct. 16 , 2016]

Küzeci, E. (2010). Kişisel Verilerin Korunması [Protection of Personal Data]. Turhan Kitabevi.

Kılınç, D. (2012). Anayasal Bir Hak Olarak Kişisel Verilerin Korunması [Protection of Personal Data as a Constitutional Right], Anakara Üniversitesi Hukuk Fakültesi Dergisi, 61 (3) 2012:1089-1169

İzgi, M. C. (2014). Mahremiyet Kavramı Bağlamında Kişisel Sağlık Verileri [Personal Health Data in the Context of the Privacy Concept]. Türkiye Biyoetik Dergisi, 1(1).

Karaarslan, E., Koç, S., & Akın, G. (2010). Vatandaşlık Numarası Bazlı E-Devlet Sistemlerinde Kişisel Veri Mahremiyeti Durum Saptaması [Personal Data Privacy Status Determination in Citizenship Number Based E-Government Systems] İzmir Bilişim Hukuk Kurultayı, 1-8.

Kaya, C. (2011). Avrupa Birliği Veri Koruma Direktifi Ekseninde Hassas (Kişisel) Veriler ve İşlenmesi [Sensitive (Personal) Data and Processing on the Axis of the European Union Data Protection Directive]. İstanbul Üniversitesi Hukuk Fakültesi Mecmuası,vol. 69(1-2), 317-334.

Ketizmen, M., & Ülküderner, M. (2007). E-devlet uygulamalarında kişisel verilerin korun(ma)ması [protection (failure) of personal data in e-government applications]. XII."Türkiye'de İnternet" Konferansı.

Ceran A. (2014). Kişisel Verilerin Korunması: Avrupa ve Türkiye [Personal Data Protection: Europe and Turkey]. Iktisadi Kalkınma Vakfı Değerlendirme Notu, vol.104.

Karimi, O. & Korkmaz, A. (2013). Kişisel Verilerin Korunmasi [Personal Data Protection]. 18. Türkiye'de İnternet Konferansı inet-tr'13, İstanbul Üniversitesi, 9-11 Aralık 2013, İstanbul, Türkiye.

OECD (2008). OECD Policy Guidance on Online Identity Theft, Internet: http://www.oecd.org/sti/consumer/40879136.pdf , [Nov. 2, 2019]

ISO/IEC 29134 (2017), ISO/IEC 29134:2017**,** Guidelines for privacy impact assessment, internet: https://www.iso.org/obp/ui/#iso:std:iso-iec:29134:ed-1:v1:en

ICO (2015). Conducting privacy impact assessments code of practice. Internet: https://ico.org.uk/media/about-the-ico/consultations/2052/draft-conducting-privacy-impact-assessments-code-of-practice.pdf [Oct. 02 , 2019]

SEC (2007). Privacy Impact Assessment (PIA) Guide. Privacy Office of Information Technology. Internet: https://www.sec.gov/about/privacy/piaguide.pdf [Nov. 2, 2019]

IPC (2015). " Information and Privacy Commissioner of Ontario: ", Planning-for-Success Privacy Impact Assessment Guide, Internet: https://www.ipc.on.ca/wp-content/uploads/2015/05/Planning-for-Success-PIA-Guide.pdf [02 Nov. 2019]

HIQA (2017). Guidance on Privacy Impact Assessment in health and social care, Health Information and Quality Authority, Internet: https://www.hiqa.ie/sites/default/files/2017-10/Guidance-on-Privacy-Impact-Assessment-in-health-and-social-care.pdf, Oct. 2017 [Nov. 02 , 2019]

GDPR (2016). Regulatıon (Eu) 2016/679 Of The European Parliament And Of The Council Act. Official Journal Of European Union, (65).

CSX (2015). " Cyber Security Nexus Cyber Security Fundamentals". Internet: https://www.isaca.org/cyber, Jan. 1, 2015 [Oct. 28 , 2016], pp 77-82.

New York Times (2018), https://www.nytimes.com/2018/12/11/us/politics/trump-china-trade.html, *New York Times*, Dec. 11, 2018, [july 31 2019]

Pascual A., Marchini K. & Miller S. "2016 Identity Fraud: Fraud Hits an Inflection Point." Internet:www.javelinstrategy.com/coverage-area/2016-identity-fraud-fraud-hits-inflection-point , Feb. 02, 2016 [Oct. 02, 2016].

Lindén, F. (2009). epsos, smart open services for European patients from strategies to services health as the enabler for cross-border healthcare. *Infrastructures for Health Care*, *23*.

Dülger, M.V. (2019a). Kişisel Verilerin Korunması Hukuku [Personal Data Protection Law]. İstanbul: Hukuk Akademisi Yayıncılık

Dülger, M. V. (2019b). First Major Breach of the GDPR: France Fined Google€ 50.000. 000. *Available at SSRN 3331321*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331321

Dülger, M. V. (2018) İnsan Hakları ve Temel Hak ve Özgürlükler Bağlamında Kişisel Verilerin Korunması [Protection of Personal Data in the Context of Human Rights and Fundamental Rights and Freedoms]. İstanbul Medipol Üniversitesi Hukuk Fakültesi Dergisi 5 (1), Bahar 2018

Kayyali, B., Knott, D., & van Kuiken S. (2013), The 'big data' revolution in healthcare: Accelerating value and innovation, McKinsey Global Institute Report, https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care [Nov. 02, 2019].

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3-13.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719-731.

Pendley, J. A. (2018). Finance and Accounting Professionals and Cybersecurity Awareness. *Journal of Corporate Accounting & Finance*, *29*(1), 53-58.

Aggarwal, A. K. (2019). Opportunities and challenges of big data in public sector. In *Web Services: Concepts, Methodologies, Tools, and Applications* (pp. 1749-1761). IGI Global.

CHAPTER 13

DATA COLLECTION APPROACHES FOR ARTIFICIAL INTELLIGENCE APPLICATIONS IN HEALTHCARE

Murat GEZER*, Çiğdem SELÇUKCAN EROL**

*Dr, İstanbul University, Informatics Department,İstanbul, Turkey
e-mail: murat.gezer@istanbul.edu.tr

**Assoc. Prof. Dr., İstanbul University, Informatics Department, İstanbul, Turkey
e-mail: cigdems@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.13

Abstract

As in all other fields, research in the field of artificial intelligence is rapidly continuing in the field of health. As a result of this research, the importance of data comes to the fore. In this study, which includes data collection approaches in the field of health, we aim to emphasize the importance of data in this field and to contribute to the more conscious handling of the data to be used in artificial intelligence applications at every stage. For this purpose, the definition of data and how to distinguish information and knowledge are mentioned. The characteristics of data and data collection methods are also mentioned, and an attempt is made to emphasize the importance of health data collection in artificial intelligence research.

As a result of this study, we believe that all personnel working in data-related departments and the health field, where the moment is vital, must receive training on collecting, storing, sharing data, and data security in particular. In our study we emphasize that especially the people who produce and consume data must have the awareness and morality for every step of data collection and handling, and that this issue should be prioritized in the field of health.

Keywords: Artificial intelligence, Data collection approaches, Data, Healthcare, Machine learning

Introduction

1. Data, Information, and Knowledge

In today's information age, almost every sector has realized the value of data. However, there is much competition in converting data into value (information/knowledge) faster than the competitors. Speed tests are performed at every stage of the process of transforming data into knowledge. No matter how fast you are, you can only go as far as your data. Even if you use the best, most up-to-date algorithms, computers, and technology and work with the best experts, you can only go as far as your data. Therefore, we decided to focus on data on this chapter.

"Data are symbols that represent the properties of objects and events" (Ackoff, 1999). These symbols are the smallest building blocks of knowledge. They are raw information such as numbers, letters, sounds, images, videos that we use frequently and are familiar with. Farmers do not collect raw fruit; they know that this has no or very low value. They wait for it to maturate. After the fruit matures, they collect and sell it as soon as possible. At this point, with many internal and external factors, time is a very critical factor. On the way from data to knowledge, time is also critical. Collecting early or belatedly may not be useful. In order to let data mature, it must be processed. The process of processing data occurs according to an algorithm consisting of a series of steps. We implement algorithms through computers and work with experts; in other words, there are people at every stage of this process. Although it may seem quite simple, it is a complex process, just like the maturing of fruit. So how do we distinguish data from knowledge? Unfortunately, it is not understood simply by biting, as in the case of fruit, and does not turn into knowledge directly. There is another step that we call the information step. This process includes three stages: Data - Information - Knowledge. In sum, data shows that something abstract or concrete exists. But we don't know what that is. When we get the answer to the question "what is this," in other words, when we give the meaning of data, we transform data into information. For example, when we ask "what is 3", it can be answered that 3 is a number. In this case, while 3 represents the data, the number is not information. The fact that 3 is a number characterizes the data itself, whereas we expect it to give meaning to the events or objects it represents. When the answers to the question "what is 3" are like the following examples, data transforms into information;

1. The number of patients waiting in the line
2. A period expressing how many days left for surgery

3. The amount expressing how much medicine I should take
4. The amount of new x-ray devices purchased for the hospital

3 represents different elements such as human, time, medicine, and device, respectively, but this symbol can be the answer to all or even more questions. Information gives us the meaning of the data; in other words, 3. It gives an identity. Three as a number is not just a number anymore. For example, 3 is a person in the first answer! So, is this information valuable? Yes, of course, it is. We often consume this information quickly. But there is one more thing that is more valuable and less available than information: Knowledge.

Knowledge includes the answer to the question "how" and allows us to understand the relationship between multiple information. It contains a process and a result. Let's suppose we have information about a regional power outage in three days. This information, when combined with the information that the generator in a hospital in the same area is out of order, turns into a decision that would delay the operation to be performed in three days at that hospital. And this decision creates a correlation between the knowledge about the process and the information about how a surgical operation can be performed. And as a result, it is decided that surgery should be postponed. Thus, the information about how a surgical decision was taken includes the relationship of information related to the operation process, such as the suitability of the patient to the operation, the availability of an empty operating room, and the availability of the physician and assistant health personnel at that time.

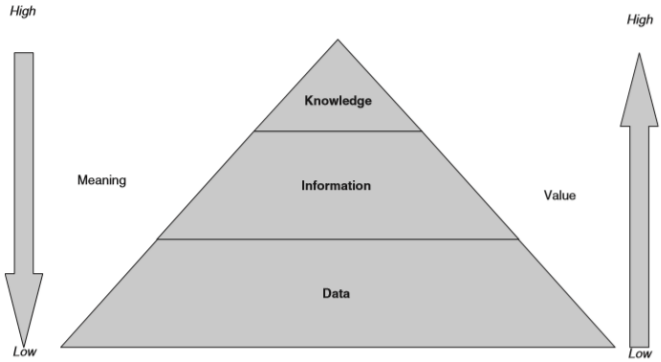


Figure 1: Data, information, and knowledge- Meaning and Value (Chaffey and Wood, 2005 in Rowley, 2007).

As shown in Figure 1, there is a value relationship between the data-information-information pyramid. Knowledge is more valuable than information, and information is more

valuable than data. The more knowledge you access, the more valuable it gets. On the other hand, while the knowledge has the most meaning, meaning decreases when stepping down to data. According to the information you obtain in this chapter, examine the relationship between "3" and "postponing the surgery" again.

1.1. Data sources

Returning to our main subject, we process data to transform it into more valuable information and knowledge. But before processing it, we need to collect and store the data. So why should I take the trouble? Because I have a purpose or a problem. For example, I aim to improve the quality of service in the hospital, or I have encountered a problem in my area of expertise, and I want to solve this problem. First of all, I need to collect data for my purpose or problem. Where can I collect this data when it comes to health and what are my data sources?

1- Patient and patient's relative: A great variety of data can be obtained from the patient and patient's relative such as radiography - image data, blood gas analysis - numeric data, voice recording during anamnesis - audio data, x-ray report - text data, using diabetes pump - sensor data, genetic testing data. We may even include social media data. Although it does not belong to a single patient or a relative, it is possible to obtain data for a drug or disease from thousands of patients and/or their relatives through social media.

2- Electronic health records: Records kept in hospital information systems.

3- Internet: Databases, articles, and similar resources that were previously developed and open to the internet for your purpose or problem.

We need to store the data after collecting it according to our purposes. However, there are a few things to be aware of: Is this data related to my purpose? Is it accurate, reliable, and up-to-date? Is there any legal obstacle for me to share this data? Does the owner of the data give his/her consent?

1.2. Data Types

We want to express the process of storing data using the metaphor of tidying up a room. Your room is very messy, and you start to tidy it up. If you take everything in the room and throw it in your wardrobe, does that equate to tidying up your room? Yes, your room will certainly be tidied up. But what happens to the things you threw into the wardrobe, and how long does it take you to find something when you need it? Or will you even be able to find it?

The process of collecting and storing data is similar to that. If you classify your items one by one and place them where they belong, it will be easier to access them when you need them. Even if you put them into boxes, it will be a waste of time searching which item is in which box as time goes by. Therefore, it is also useful to label the boxes. We made a metaphor about your item, which stands for your data and the box you place it. When we substitute the data for the item again, you will have data about your data, and we call it **metadata**.

We often use databases to store data. We often use organized, ordered, i.e., structured data in databases (NoSQL databases have been used in recent years to store unstructured data). We group our data into structured, semi-structured and unstructured data. For example, an x-ray report of a patient or an e-mail from the head physician is called unstructured data.

(A) UNSTRUCTURED DATA

Dear colleagues,
Your 37-year-old patient with ID number 1 has a body mass index of 25.6, and the mean blood sugar value for three months is 6.1%. Your 25-year-old patient with ID number 2 has a body mass index of 27.8 and a hemoglobin a1c value of 6.7%. It is submitted for consideration.
Kind regards

(B) STRUCTURED DATA

ID	Age	BMI	Hb1ac (%)
1	37	25.6	6.1
2	25	27.8	6.7

Figure 2: Datatypes; Structured and unstructured data

The unstructured data (A) in Figure 2 is converted to structured data (B) and stored in the databases. Apart from these, there is a semi-structured data type that is usually used on websites. This type of data, as its name signifies, has a format between structured and unstructured data. It is stated as labels (Figure 3).

```
<Hospital>
        <Patient ID=1>
                <Age> 37 </Age>
                <BMI> 25.6 </BMI>
                < Hb1ac> 6.1 </ Hb1ac>
        </Patient>
        <Patient ID=2>
                <Age> 25 </Age>
                <BMI> 27.8 </BMI>
                < Hb1ac> 6.7 </ Hb1ac>
        </Patient>
</Hospital>
```

Figure 3: Semi-structured data

Also, the data types are divided into two according to the purpose of collection: primary data type and secondary data type. The data that the researcher collects for the first time for a particular purpose is the primary data, while the secondary data is the data formed by converting it and making it ready for use again. Figure 4 presents a comparison of the two data types.

Primary data	**Secondary Data**
Data collected for research	Data collected in the past
The source of the data is certain	The source of the data is uncertain
Helps us find the solution to the problem	Supports finding the solution to the problem
Data is collected on demand; therefore, it can be structured according to the needs.	
The cost of data collection can be high	It has more relevant costs

Figure 4: Primary data vs. secondary data

2. Health Data Collection Methods

Although the issue of data collection in each sector has country-specific regulatory rules and management challenges, the complexity of data in the health sector and the strict regulations make collecting data difficult (WHO,2003). Since it is related to the privacy of the patient, health data is considered as personal data in the "sensitive data" group (Dülger, 2015). In addition, some of the health data can be categorized as trade secret data. In this respect, the method of collecting and storing health data is important. Health data and a wide range of health indicators for a community are used to assess the costs of measurable clinical health services. Also, scientific studies provide clinical comparisons and can be used to identify the measures needed. Collected data should include all findings related to the patient's condition. These include diagnosis, treatment, and other events that have occurred.

Despite all the difficulties, how can we make the data collected from the data sources usable? Then how can we store this data securely? Nowadays, it is important that the data collected for artificial intelligence research should be usable. Usability is considered an issue that needs to be solved in data collection approaches. Since health data is increasingly used in the field of artificial intelligence, the accuracy and reliability of the collected data is important for the production of useful information from that data.

The choice of data collection methods depends on the purpose of the study (evaluation), the questions to be answered, and the sources from which the data will be collected. In the field of health, data is generated and collected by doctors, nurses, health technicians, health officers, patients, and technological devices. Data collection methods are divided into two

categories, namely *the primary data collection method* and *secondary data collection method,* depending on the type of data (primary and secondary) (Hox 2005).

Primary data collection methods are used for data which is designed, collected, and analyzed to answer a specific research question in the research and these are divided into two groups: *qualitative data collection* and *quantitative data collection*.

In health care, data is collected by devices (Laboratory instruments, Scanning devices, Electroencephalography, etc.), as a written document (surveys, forms, anamnesis, etc.) or as computer inputs (barcode scans, data typing, audio input, text input, etc.). Today, data is collected mostly through digital channels and with the help of numerous applications available on the market. In their study, Sarkies et al. (2015) stated that data collection in the field of health care is done by *observational data*, *retrospective data extraction*, and *retrospective review* methods.

Observational data can be defined as the collection of information that corresponds to the effects of a patient's laboratory values, behavior and life changes, demographic characteristics such as malignancy, and being ill. *Retrospective data extraction* is described as extracting knowledge from previous studies that correspond to a certain impact in a health information system. Retrospective analysis can be defined as scanning all written documents and saving them into an electronic environment after the patient is discharged (Sarkies 2015).

The advantages of the primary data collection method are that the data is collected for the intended purpose, the quality of the collected data is already under the inspection of the research team, and additional data collection is possible when needed. The issue of disadvantages in the primary data collection process is one of the main problems that the researcher has to solve. These disadvantages can be explained as the time required for data collection, ethics committee permission requirements, and costs. The advantages of the secondary data collection are shorter data collection time and lower cost. However, additional data needed for the research can hardly be obtained, and data quality cannot be monitored during data collection. These can be considered as the disadvantages of secondary data collection.

3. Data Quality and Usage

The development of artificial intelligence and the performance of machine learning algorithms depend on the usage of large data sets. Understanding whether the collected data can be used in health research, and artificial intelligence research in particular, concerns data

quality and information governance. The criteria for data quality are validity and reliability (Otieno-Odawa, 2014).

Validity acknowledges the true accuracy of a piece of data. Validity is defined as the concordance of the data to be collected with different instruments, i.e., the data does not affect the results. In other words, there should be no uncertainty. Any deviation in the accuracy of the data will cause the results to be inaccurate. Reliability is not just a feature of the measurement tool. It is a feature of the measurement tool and the results of the tool. If the collected data is obtained in the same way in repeated cases, the data collection method or medium can be considered reliable. Also, reliable data must be understandable. Figure 5 shows the meaning and relationship between the reliability and validity of the data generated as a result of a certain number of shots on a target board (Troachim, 2006). The collected data in Figure 5a is assembled with the same reliability each time, but incorrect measurement of the data is performed systematically. In Figure 5b, the collected data were randomly spread on the target. It was rarely appropriately collected, but in one case, the correct answer was received. Figure 5c shows that the collected data cannot be adequately obtained under any circumstances, and the measurement tools are defective, and finally, Figure 5d shows that the data collection is accurate and that the measurement tools work correctly.

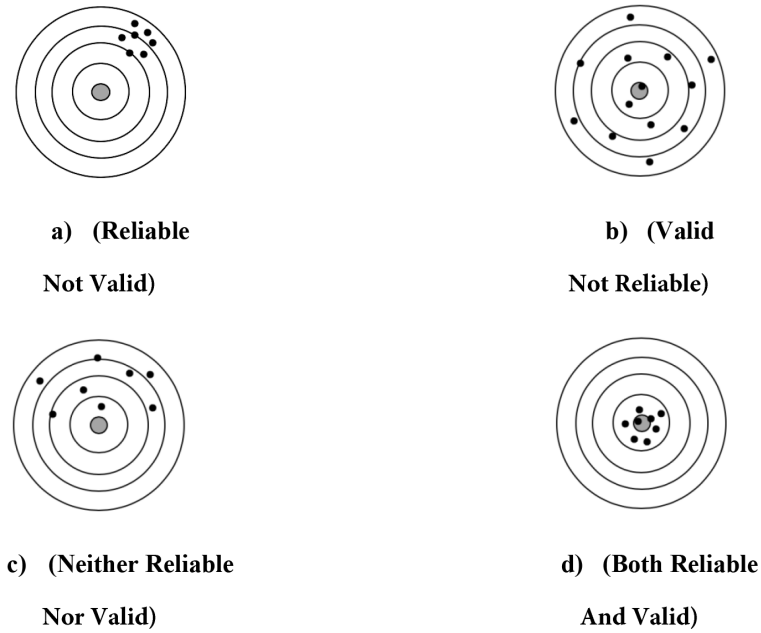


Figure 5: Relationship between Validity and Reliability (Trochim, 2006).

In addition to the above two criteria, the accuracy, completeness, legibility, relevance, timeliness, and accessibility features of the data are examined in the field of health during the measurement of the quality of data for data processing (WHO 2003, Kirch 2008).

Accuracy: This is defined as the state of data when it meets the gold standard criteria and is measured with reliability and validity (Kirch 2008).

Completeness: This indicates whether the data that should be included in the dataset is missing or not. It is especially important for the accurate decision making of artificial intelligence systems. That is why it should contain all records related to activities about the collected health data/health records, and it should be documented completely and properly (WHO 2003, Kirch 2008).

Legibility: As the data collected during patient registration is **legible**, it becomes possible for data users to comment on the subject. The handwriting to be used in the records should be legible, no undefined coding should be used, and abbreviations should be in accordance with standards (WHO 2003, Kirch 2008).

Relevance: This can be defined as the relevance of an attribute collected in the data content to the subject to be studied. This needs to be taken into account, especially when collecting data for artificial intelligence algorithms, as unnecessary attributes will affect performance (WHO 2003, Kirch 2008).

Timeliness: This is defined as the retention of data in such a way as to include time-dependent changes (WHO 2003, Kirch 2008). Clinical information should be documented as soon as actions are performed. Each activity should be recorded during treatment. Postponement of data entry may cause the skipping of information and errors.

Accessibility: This indicates that the data should be accessible by authorized bodies and persons when needed (WHO 2003, Kirch 2008). If data is not available when it is required, information loses its value.

Data security and information management are also essential factors since data contains sensitive personal information. Therefore, it is necessary to create a very precise balance in data quality and information management (AoRMC, 2019). These balance factors can be listed as clinical considerations, ethical concerns and practical issues arising from the processing of the data (AoRMC 2019). For example, in clinical examinations, the usage of artificial intelligence algorithms in a doctor-patient relationship will involve potential third parties. Confidentiality and security of the data obtained during the treatment will become

questionable. In this case, ethical concerns may arise about who is the owner of the data. Who can be the owner of the data - the patient (i.e., source), the collector (i.e., doctor), the processor (i.e., system), system manager or system owner? Therefore, regulations such as GDPR (General Data Protection Regulations) allow the deletion of personal data. However, in this case, it should be examined whether it is practically possible to remove this data from the artificial intelligence algorithm.

4. Conclusion

In artificial intelligence studies, machines learn from any data such as audio, image, text, signal, etc.. The data we collect and store from different sources for our purpose goes through a very intensive process that we call data preprocessing to use it in artificial intelligence applications. Collecting the data deliberatively in accordance with specific standards provides faster and more efficient execution of preprocessing; hence, artificial intelligence applications.

In this study, we discussed the data collection approaches that are critical in data preprocessing. In addition to these approaches, we believe that following certain standards while collecting and storing data is very important, especially in the field of health, in which there is so much data, and every second is crucial. We believe that all personnel working in data-related departments must receive training on collecting, storing, sharing data and data security in particular. In artificial intelligence research, while we mainly focus on machines, data, algorithms and applications, we ignore the human factor which produces and consumes it. We want to draw attention to the fact that health data, in particular, must be used by conscious and ethically-aware producers/consumers.

References

AoRMC, (2019) Academy of Royal Medicine Colleges Report: Artificial Intelligence in Colleges.

Ackoff, R. L. (1999) Ackoff's Best. New York: John Wiley & Sons, pp 170 – 172.

Ackoff, R.L. (1989) From data to wisdom, Journal of Applied Systems Analysis 16 3–9.

Chaffey, D., Wood, S. (2005). Business Information Management: Improving Performance using Information Systems (FT Prentice Hall, Harlow).

Doğan G.,(2019). Veri toplama araçları [Data Collection Tools]. https://acikveri.ulakbim.gov.tr/acik-veri-acik-bilim/bolum-2-arastirma-verisi-hazirlama-sureci/2-5-veri-toplama-araclari/. (erişim tarihi: 01.8.2019)

Dülger M. V. (2015), Sağlık Hukukunda Kişisel Verilerin Korunması ve Hasta Mahremiyeti [Protection of Personal Data In Health Law And Patient Privacy], İstanbul Medipol Üniversitesi Hukuk Fakültesi Dergisi 1)20, 43-80.

Jox j.j., Boeije H. R., (2005) Data Collection Primary vs. Secondary, Encyclopedia of Social Measurement, Volum 1, Elsevier.

Kirch W,(2008) Encyclopedia of Public Health, Springer.

Lim, C., Kim, K. H., Kim, M. J., Heo, J. Y., Kim, K. J., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. International Journal of Information Management, 39, 121-135.

Rowley, J., (2007). The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180.

Trochim, W. M., (2006). The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006).

Sarkies, M. N., Bowles, K. A., Skinner, E. H., Mitchell, D., Haas, R., Ho, M., … Haines, T. P. (2015). Data collection methods in health services research: hospital length of stay and discharge destination. Applied clinical informatics, 6(1), 96–109. DOI:10.4338/ACI-2014-10-RA-0097

Otieno-Odawa, C. F., & Kaseje, D. O. (2014). Validity and reliability of data collected by community health workers in rural and peri-urban contexts in Kenya. BMC health services research, 14 Suppl 1(Suppl 1), S5. DOI:10.1186/1472-6963-14-S1-S5

WHO Library Cataloguing in Publication Data (2003) Improving data quality: a guide for developing countries. World Health Organization, Geneva

CHAPTER 14

THE TECHNOLOGICAL TRANSFORMATION PROCESS FROM ELECTRONIC INTELLIGENCE TO CYBER INTELLIGENCE

Ahmet Naci ÜNAL*

*Assist. Prof., Bahcesehir University, Faculty of Engineering and Natural Sciences, Software Engineering, Istanbul, Turkey

e-mail: ahmetnaci.unal@vs.bau.edu.tr

DOI: 10.26650/B/ET06.2020.011.14

Abstract

Throughout the whole of human history, concepts such as defense, security, safety and intelligence have been very important for human beings on a personal level, and for human communities in general. The sociological transformation that came as a result of these processes played a key role in the development of science and technology.

Thanks to the developments in electronic science during the 20th century, systems using electromagnetic energy have come to the fore.

This development process which started with systems such as telephones, radios and radars has eventually been used in many different areas such as air defense systems, guided missiles, early warning receivers, communication systems, and computers.

For this reason, the control and active use of what can be called the electromagnetic spectrum in short, has been an important factor in all kinds of activities. By the 21st century, almost all of the systems used in this process began to operate in a cyberspace environment and became software controlled. The concept of the target intelligence needed in this transformation process has changed dimensions and shifted from electronic intelligence to cyber intelligence.

This study will focus on the transformation of the electronic intelligence process, which is an indispensable element of the 20th century, into the concept of cyber intelligence in the 21st century.

Keywords: Signal intelligence, Electronic intelligence, Cyber intelligence, Decision support systems, Cyber ecurity

Introduction

1. Intelligence

The notion of intelligence is almost as old as humanity and its first mention in a written source occurs in Sun Tzu's work entitled "The Art of War". While intelligence has different definitions in various cultures, it is basically a science that covers the processes of gathering information (on any topic), as well as processing, analyzing and reporting on this information. It was mainly considered a social science branch since the last quarter of the 20th century. However, from the end of the 20th century, the development of information technology, the change of information production and the transformation of the defense systems in particular have turned this area into an interdisciplinary one.

Schleher (1999, p.11) has divided the concept of intelligence into eight basic sub-disciplines within itself, as shown in Figure 1.

In Figure 1, we can say that intelligence is generated through the combination of these eight sub-disciplines in order to observe, detect, record and transmit different forms of data. The common denominator in this process is the use of different technologies for a common purpose.

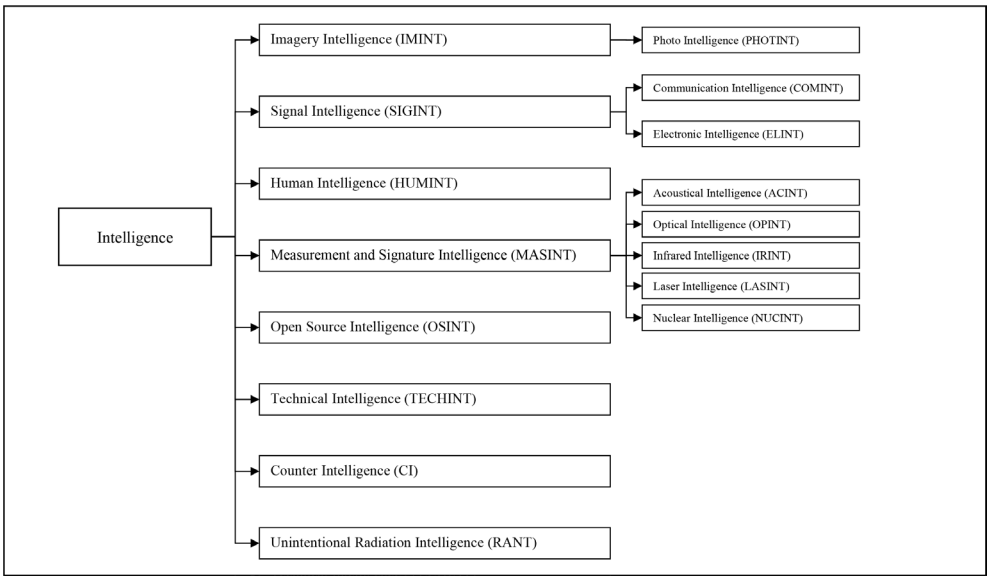


Figure 1: Sub-disciplines of intelligence (Schleher, 1999, p.11)

2. Electronic Intelligence

It is possible to associate the first use of electromagnetic energy in communication with the telephone invented by Graham Bell at the end of the 19th century. Then, radio communication tools developed in early 20th century began to expand this development into different dimensions.

The introduction of telephones and radios, and the effective use of these inventions in terms of security and defense increased the desire to know the source and content of the communication. The scientific studies carried out as a result of this desire were named Electronic Warfare (EW) especially during World War II and in the post-war period.

As stated by Schleher (1994), EW was defined as "the use of the electromagnetic spectrum in order to deceive the enemy, to locate their units and facilities, to disrupt their communication facilities and to destroy the command, control and target detection systems of the enemy" and as shown in Figure 2, can be divided into four sub-sections called Intelligence, Electronic Support Measures (ECM), Electronic Counter Measures (ECM) and Electronic Counter Countermeasures (ECCM).

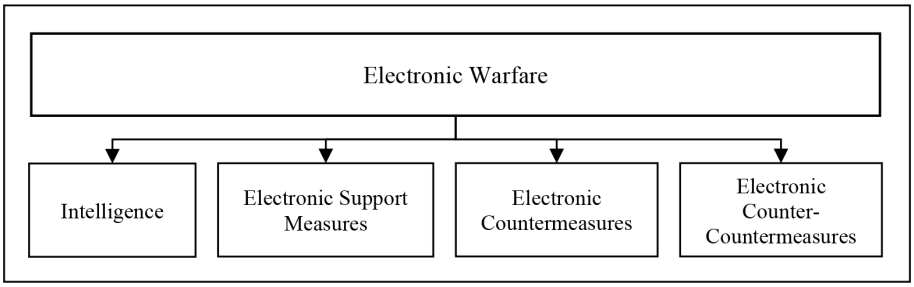


Figure 2: Previous structuring of electronic warfare (Schleher, 1994, p.7)

Here, in addition to intelligence:

- Electronic support measures include the activities of eliciting electromagnetic transmissions of the enemy in order to organize EW activities for offensive and defensive campaigns against them,
- Electronic countermeasures include activities to prevent or reduce the enemy's effective use of the electromagnetic environment,
- Electronic counter countermeasures include activities that ensure the efficient use of electromagnetic energy by friendly forces despite the enemy's electronic countermeasures.

Described by Hoisington (1994) "the use of electromagnetic systems and directed energy for military purposes to monitor the electromagnetic spectrum, to collect information, to control, to attack the enemy and, where necessary, to prevent the enemy's use of the electromagnetic spectrum" has in modern times gained the structure shown in Figure 3.

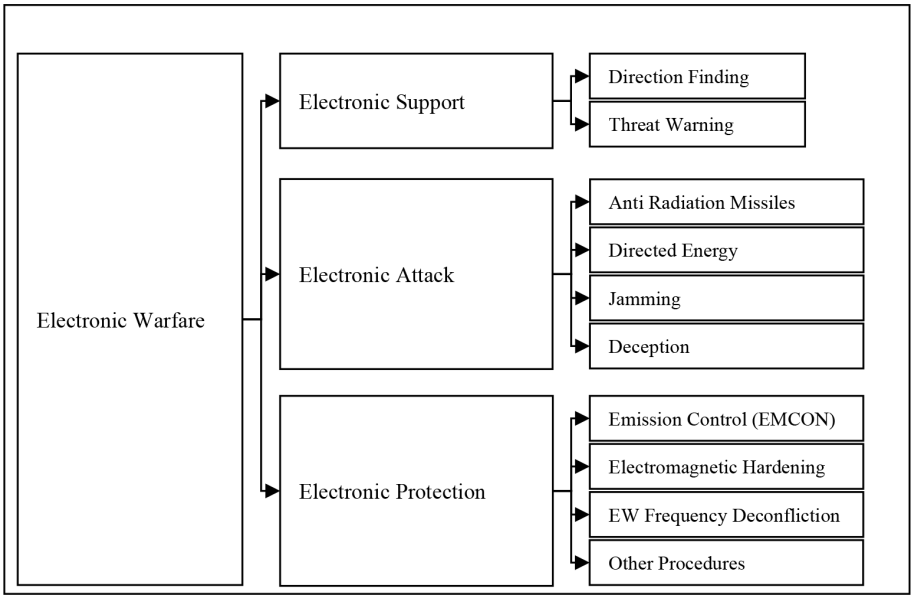


Figure 3: Sections of EW. (Hoisington, 1994, p. 1a-17)

When Figure 3 is examined, it can be seen that EW is studied under three main sections: Electronic Support (ES), Electronic Attack (EA) and Electronic Protection (EP). These sections are summarized below (Scleher, 1999: 2):

- Electronic Support (ES) includes the activity of searching for, detecting and identifying the sources of deliberate/undeliberate electromagnetic energy for emergency threat detection,
- Electronic Attack (EA) includes the reduction or elimination of combat capabilities of hostile electromagnetic systems, or offensive campaigns against them with directed energy weapons,
- Electronic Protection (EP) covers the protection of personnel, facilities and equipment from the effects of the friendly or hostile EW activities and includes reducing or eliminating their effectiveness.

When the structures in Figures 2 and 3 are examined, both the hostile systems and the friendly systems to be protected appear to be electromagnetic based defense systems.

Therefore, the most important factor in any EW activity is the identification of threats from electromagnetic sources, analysis of these threats and the prevention of these signs. The most important factor in this context is efficient intelligence efforts.

Within the scope of EW, intelligence activities are carried out in two stages. The first one is carrying out an effective SIGINT activity as shown in Figure 1, followed by the analysis of the signal intelligence obtained through this activity, and the monitoring of hostile signal parameters identified through the ES systems shown in Figure 3.

2.1. SIGINT Activities

Signal Intelligence (SIGINT) is a generic term that usually covers Communication Intelligence (COMINT) and Electronic Intelligence (ELINT).

"COMINT is defined as intelligence derived from potentially hostile communications by other than the intended recipients. ELINT is defined as intelligence information that is the product of activities in the collection and processing, for subsequent intelligence purposes, of potentially hostile, non-communications electromagnetic radiations which emanate from other than nuclear detonations and radioactive sources" (Schleher, 1994:8).

2.2. Electronic Support (ES) Activities

The purpose of ES activity is to detect, analyze and identify the transmissions from hostile electromagnetic systems and to inform the operator or electronic assault systems of the target location, type and attack stage. While performing all these activities, it is attempted collect detailed data about the hostile electromagnetic system.

One of the most sought-after features in ES systems is the capability of platforms where these systems are installed to detect threat systems before they are detected by threat prevention systems. The ability to perform this detection depends on comparing each detected signal with a predetermined sequence of parameters. These basic parameters (as depicted in the Electronic Warfare and Radar Systems Engineering Handbook, p. 5-8.1) are as follows:

- Radio Frequency (RF)
- Amplitude (Power)
- Direction of Arrival (DOA) - also called Angle of Arrival (AOA)
- Time of Arrival (TOA)
- Pulse Repetition Interval (PRI)

- PRI Type
- Pulse Width (PW)
- Scan Type and Rate
- Lobe Duration (Beam Width)

The basic working flow diagram of the warning receivers used to separate these parameters are shown in Figure 4.

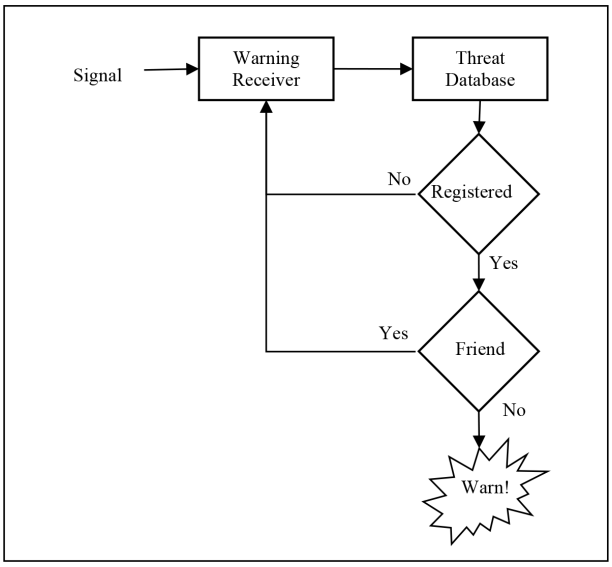


Figure 4: Block diagram of a warning receiver

Here, the signal parameters of threats expected to be present in the operational environment where the task will be performed are recorded in the threat library in the warning receiver system prior to the mission, and the system is programmed to detect signals with these characteristics. The parameters of the detected signals are measured and compared to the actual threat parameters previously defined in the system's threat library to determine whether the detected signals indicate real threats and the location of these threats. If these signals are not friendly, a warning is given.

When we consider the entire area where electromagnetic transmission takes place, it is not possible to detect all frequencies with a single system with the current technological means. The main reason for this is that the power in the platform is constant and that this power is also used by other electronic systems. The limited availability of power supplies causes the electronic support system to not be able to cover the entire frequency band, and

only serve against threat systems that have been identified or found on the field of operations. That this activity takes place within the shortest amount of time is just as important as the effectiveness of the activity. For this reason, fast and effective data analysis is of great importance.

3. Decision Support Systems (DSS)

Decision-Making (D-M) is not a coincidental action that results in choosing one of the multiple alternative modes of action. The basis for decision-making is the ability to process information in a variety of ways. D-M in this context is defined as:

"The act of identifying "N" alternative modes of action and selecting one of them as a decision" (Holsapple, 2008, p.26).

"The process of selecting one of the many action options that can be fulfilled in order to achieve a defined goal" (Harrison, 1995, p.4).

Today's world is globalized, virtualized and has become more complicated in every way. In a complex world, it has become difficult to predict and decide on the future, while access to information needed for these activities has become easier. However, one of the consequences of accessing so much information is the difficulty of evaluating this information with accurate data analysis.

At this point, it is necessary to produce estimations and decision options by providing the information needed by the decision-maker from the information that is not fully known, reliable and accurate in terms of intelligence science. This is possibly the most difficult and time-consuming task of intelligence activities. The fact that these activities have to be carried out at a strategic level and in a limited time determined in order to reach the right information makes the conditions even more difficult. The need for properly classified, reliable and well-proven data banks and mathematical models to meet the needs of decision-makers grows even further in extreme situations, especially in the case of different events taking place at the same time.

In this sense, although the data banks are an important help, the use of data in the required subject, the collection and analysis of new data, and in short, the process of gathering intelligence takes time and the need for DSS increases.

The purpose of DSS is to produce and use better information as well as to help decision making (Power, 2009, p.4).

Each DSS is unique, in other words, different from each other. However, each DSS has three basic structures: data, model and knowledge base (Forgionne, 2003, p.14):

- The database includes evaluation, decision options and decision criteria of uncontrollable events directly related to the decision problem.
- The model base is where uniform (graphical, conceptual or mathematical) decision problem models and the results of these models are stored.
- The knowledge base contains the problem information, the formulas that can convert the obtained data into problem parameters, the guide for selecting the decision options, and the problem relations or suggestions on the possible outcomes.

In today's world, considering that DSS activities are conducted via the internet or sensor networks, it is known that in reality, all the said activities are supported by cyberspace.

4. Cyberspace

The internet was developed by the Defense Advanced Research Projects Agency (DARPA) and began to be used for commercial purposes in the early 1990s. After approximately 30 years, the internet is actively used by approximately 4.2 billion people (according to June 2018 data). This number is equal to 55.1% of the world's population (Internetworldstats, 2018). The Internet environment has continued to develop in this thirty-year period, and the results of activities in the virtual environment have undergone a physical transformation. In today's world, a great proportion of internet access is managed by various sensors connected to information systems. The Internet environment, which was formerly called "virtual environment" or "cyber world" is now called "cyberspace" due to its growing domain. Cyberspace is defined by the International Telecommunication Union (ITU) as "the environment that enables communication via computer networks." In the United States Department of Defense Dictionary of Military and Associated Terms (2010:58), it is defined as "a global environment of networked information technology infrastructures, including the Internet, networks, computer systems, embedded processors and controllers." In other words, cyberspace is not only internet based, but is an infinite environment where systems can communicate with each other through different networks.

In the official publication of the US Armed Forces, the Cyberspace Operations Concept Ability Plan 2016-2028, cyberspace is referred to as the fifth dimension in addition to the four dimensions of air, land, sea and space. It is further noted that each of these five dimensions is independent of each other, but that the cyberspace nodes (ports) are connected to each

dimension. This situation expands the environment in which the data is produced, evaluated, stored and shared.

Terms such as data and information are indispensable building blocks of today's world. Vercellis (2009: 6) describes data as the structural code of a phenomenon, while Stair & Reynolds (2009: 4) describe the sum of the regulated attributes (data) as information, and Holsapple (2008: 22) describes information as the main component for decision-making. In this context, the formation process of knowledge is shown in Figure 5 (Sanders, 2016).

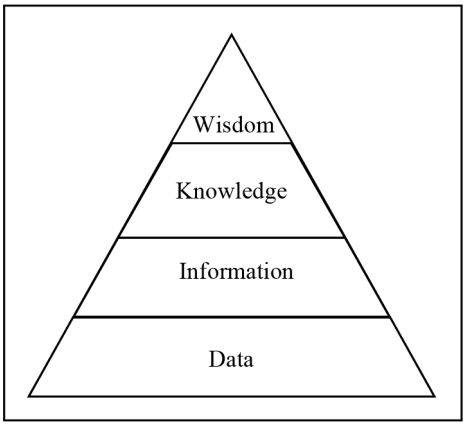


Figure 5: The formation process of knowledge (Sanders, 2016: 2)

When Figure 5 is examined, it can be seen that the data is processed for a specified purpose or purposes and transformed into information, and knowledge is generated by developing and using this information. During this transformation process, access to information, processing of information and storage of information are in the front line, and the information available is used or transmitted to other areas for use. Nowadays, information - with the help of developing technology - enables us to identify the data/information we are interested in very large data/information sets accurately and quickly, perform the analysis/synthesis stage of the data determined with great accuracy, and propagate the new information very quickly.

As the formation process of new information becomes shorter, the technological development process and the cost of reaching this technology also changes in a positive way. In fact, all these developments are related to the production, storage, processing and transmission of information. Therefore, during these activities, the information should not be distorted, lost or have its content changed. In other words, it needs to be protected. This level of protection can be realized on the basis of individuals, societies and countries. This protection highlights the concept of cyber security.

5. Cyber Intelligence

Cyber security endeavors to prevent and protect the enterprise and user assets against security risks in cyberspace environments. General cyber security objectives can be examined under the three basic structures of accessibility, integrity and confidentiality (ITU-T Rec., 2008).

These objectives are described in the Security 101 study by Carnegia Mellon University as follows:

- Accessibility means protection of information and information systems from unauthorized disruption. In short, it is the provision of timely and reliable access to information and information systems.
- Integrity means the protection of information against unauthorized modifications or destruction. Its aim is to ensure that information and information systems are accurate, complete and intact.
- Privacy means protecting information from unauthorized access or disclosure. It allows those who are authorized to access information to do so while preventing unauthorized people from doing the same.

However, all the concepts in with the "cyber" prefix affect all layers of society and the stated objectives create perception of threat for all layers of society. In particular, smartphones, tablet computers that provide easy access to cyberspace and all sensors connected to cyberspace make these threats appear ordinary. This may lead to people becoming the target of cyber intelligence in addition to technological equipment.

The purpose of electronic intelligence summarized in the second part is identifying the parameters of all electromagnetic systems which may pose a threat. For this, firstly it is necessary to define the elements that may pose a threat to us. After this definition, the threat analysis needs to be carried out and measures should be developed according to the determined threat parameters. This applies to almost all types of work related to safety.

The fact that almost all of our activities are carried out in a cyberspace environment nowadays increases the need to know the threats in this environment.

5.1 Cyber Threats

When it comes to the concept of cyber threat, people often think of software-based threats. Although this approach is not wrong, it is quite lacking. When we examine the technologies

we use around us, we see that software is involved in the control or usage of almost every kind of hardware, some devices work with software within the specific hardware called embedded software, Internet of Things (IoT) sensors form their own network, and the usage areas of unmanned devices is expanding constantly. Therefore, hardware factors are also included in cyber threat analysis in addition to software.

Of course, we should not forget about the human factor. Human beings are mostly involved in these systems as end-users, and no matter how secure the system design, users can still compromise this secure system. Since there is no single behavior that will keep people safe in cyberspace, cyber security requires multiple, interrelated behaviors that can be influenced by many factors. For example, the instinct that motivates people to use a strong password can be very different from the one that makes them click on a phishing link (Coventry et al., 2014). On the other hand, while cyber security threats that exploit human behavior are constantly evolving, "human" emerges as the most important factor in most information security violations. A significant number of security vulnerabilities (sharing passwords, opening of unknown e-mails and attachments, etc.) are caused by in-house employees. Such behavior can result with the business becoming vulnerable to hackers as well as business assets vulnerable to threats (Abawajy, 2014). Another element that is directly related to the human factor is the biometric characteristics of humans.

The concept of biometry refers to determining the physiological or behavioral characteristics of a person to identify or verify their identity (Sharma A., Raghuwanshi A., Sharma VK, 2015). Biometric systems measure the physiological and behavioral characteristics of people. In this context, biometric characters used in biometric systems are classified by Unar et al. (2014) and listed in Figure 6:

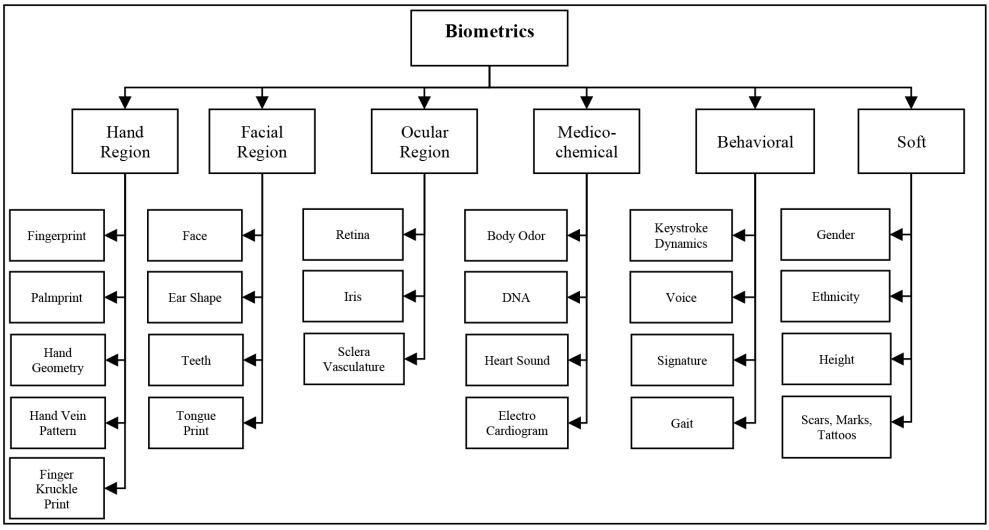


Figure 6: Classification of biometric modalities (Unar, J.A. et al., 2014)

Figure 6 shows the certainties in the physical and behavioral structure of human beings. Similarly, Sharma et al. (2015) examined the more accessible biometric features in their study under the sub-headings of physical and behavioral as shown in Figure 7. While the physiological biometric criteria are classified as face, fingerprint, hand, iris and DNA with the effect of developing technology, behavioral characteristics such as key pressing force, signature and voice under behavior.

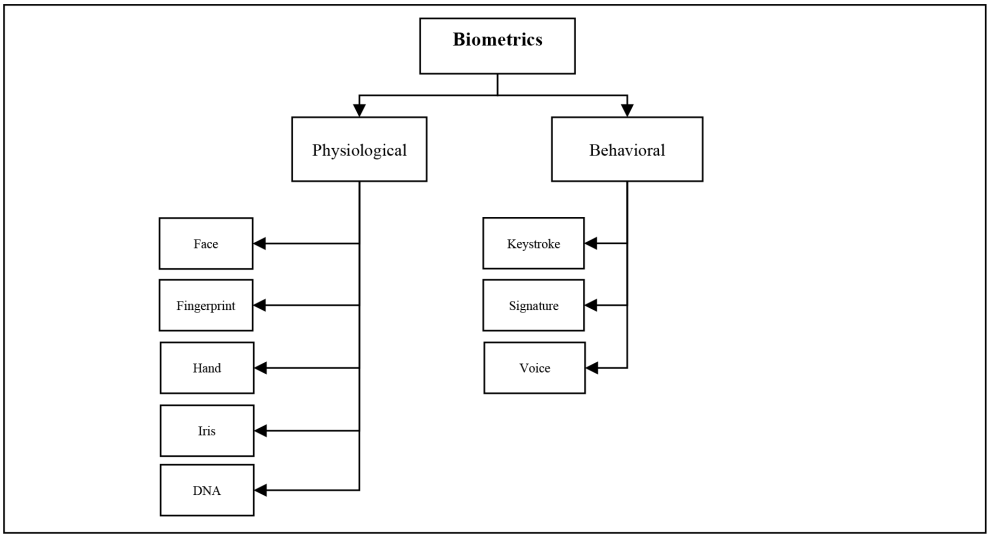


Figure 7: Classification of biometric modalities (Sharma A.K., et al., 2015:4616)

All these biometric features are primarily used in the field of security. However, these qualities also have certain vulnerabilities in their own right. The analysis by Liu and Silverman (2001: 31) who studied these vulnerabilities compared them in terms of various characteristics as shown in Table 1.

Table 1: Comparision of biometrics (Liu and Silverman,2001:31)

Characteristic	Fingerprints	Hand geometry	Retina	Iris	Face	Signature	Voice
Ease of Use	High	High	Low	Medium	Medium	High	High
Error incidence	Dryness, dirt, age	Hand injury, age	Glasses	Poor lighting	Lighting, age, glasses, hair	Changing signatures	Noise, colds, weather
Accuracy	High	High	Very high	Very high	High	High	High
Cost	*	*	*	*	*	*	*
User acceptance	Medium	Medium	Medium	Medium	Medium	Very high	Very high
Required security level	High	Medium	High	Very high	Medium	Medium	Medium
Long-term stability	High	Medium	High	High	Medium	Medium	Medium

* The large number of factors involved makes a simple cost comparison impractical.

In particular, the use of biometric features in cyberspace-related environments emerge as a different aspect of human-induced cyber security errors. For example, suppose that you unlock your mobile phones or information systems with your biometric features such as fingerprint, iris or biometric photo. First of all, these features are stored in the relevant area of the information system and compared to the data stored in each subsequent login. This comparison has become even more sensitive through artificial intelligence applications. Now imagine that sensitive information is obtained by malicious people through cyber methods. Biometric information unique to you is entirely in the hands of others. What can you do? Nothing. In this context, we need to determine the security strategies and access methods of our cyber systems in terms of cyber intelligence.

It is possible to examine such threats under three sub-headings: malware installed or pre-installed in our hardware, malware that can spread via the network, and cyber threats from users:

5.1.1 Malware

Malware is software that can infect information systems through areas such as the network line and external hardware, copy itself and spread to other information systems in communication, and damage the data instantly or in pre-determined intervals automatically

or in a commanded manner. These are called viruses, sworms, trojans, spyware, keyloggers etc. based on their types.

5.1.2 Cyber Threat Methods

Cyber threat methods are methods developed to transmit the malicious software referred to in the previous section into information systems. Social engineering, phishing, DDOS attacks, phishing attacks and IoT sensor attacks are the most commonly used methods.

5.1.2.1 Social Engineering

In today's digital world, people spend most of their time on social media sites. In these networks, they leave their individual traces with the things they share, comments they make to posts by others, or the areas they like. As the number of these comments increase, the feeling, thought or behavior patterns that are almost enough for a personality analysis can be identified. Referred to as Social Network Analysis (SNA) in Hewett's (2011:2732) study, this phenomenon aims to identify the secret characteristics of society as well as the behavioral patterns of individuals and the different roles of individuals or communities involved in the network.

In their study, Sabbagh and Kowalski (2012) mention the "Cultural Risk Theory" which dates back to 1966. According to this theory, it is argued that the way people perceive risks depends on their social norms and cultural roots. According to this finding, people are divided into four groups: individualist, egalitarian, hierarchical and fatalistic. Of these four groups, the individualists perceive their environment well. Therefore, they are not afraid to take risks. Egalitarians perceive their environment temporarily. These persons avoid risks. Hierarchical group members have a tolerant perception of their environment and are willing to take risks within certain limits. Fatalistic people generally have a capricious perception of what is around them and are unresponsive to risks.

Individuals share content on social networks voluntarily and of their own volition, and in doing so, expose their characteristics without being aware of this. Therefore, malicious people who follow these networks can easily perform cyber threat analysis with the help of what these people have shared. In addition, by using this technique, the files in the information system entered can be encrypted by malicious people and just like in real life, ransoms can be demanded in exchange for providing the passwords required to access them.

5.1.2.2 Phishing

Phishing software primarily aims to obtain the passwords people use for banking, finance and e-commerce etc. In this method, generally web pages similar (almost exactly the same)

to those of the organizations that are active in the specified sectors are sent to the e-mail addresses of the users, people are led to believe these are real, convinced that their accounts are at risk, and thus personal information including passwords are obtained in the cyberspace environment.

5.1.2.3 DDOS Attacks

The growth of existing cyberspace usage market with increasing speed and consequently, the intensification of network traffic poses great threats to the use of the system's backbone.

"Denial of Service (DoS) attacks occur when machine or network resources are made unavailable to their intended users by disrupting service, usually by flooding the network with requests, often from botnets. The DoS is particularly damaging to organizations that rely on a web presence. Distributed Denial of Service (DDoS) attacks strike a target from many sources and are harder to stop. Pulse wave DDoS attacks come in short bursts on multiple targets and can last for days. The DoS attacks can be used to mask other attacks. In 2017, the Mirai Internet of Things (IoT) botnet was responsible for the largest DoS attack in history. The attack lent credence to warnings about IoT vulnerabilities and led to massive increases in security of IoT devices" (Kettani & Wainwright, 2019).

5.1.2.4 Ransomware

"A ransomware is a kind of malware which demands a payment in exchange for a stolen functionality" (Gazet: 2010; 77). It also provides service for decrypting files or unlocking a terminal with the exchange of online currency such as bitcoin or moneypack (Ng et al:2018). First, a link or file is sent that will interest the information system user. Afterwards, malware infects the user's information system by clicking on this link or uploading the file. Files in the information system are encrypted. Finally, the ransom message is sent.

6. Comparison of Cyber Threat Intelligence and Electronic Intelligence

EW is effective on electromagnetic systems operating in a physical environment as their activity area. It is important to determine the location, types and other parameters of all previously identified electromagnetic sources emitted by the systems located in the operating environment, in other words, to establish the EW layout (Electronic Order of Battle-EOB). In today's operational environments, this is achieved through a contemporary concept called Network Centric Warfare and three dimensions of the operational field. Figure 8 shows this complex environment.

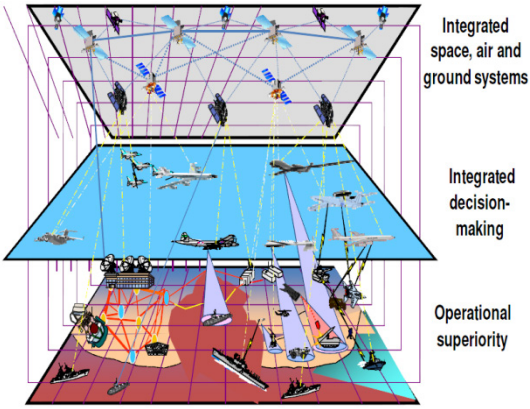


Figure 8: Spectrum utilization environment in the network-centric warfare (Park, at all,2018)

In order to create the dimensions shown in Figure 8, it is important to detect and identify the parameters of hostile electromagnetic systems in the operational environment and develop preventive techniques. For this purpose, it is important to perform signal analysis following an in-depth SIGINT surveillance. When this environment is examined, it can be seen that it is based on a physical plane and electromagnetic transmissions.

As described in the Cyberspace Operations Concept Capability Plan 2016-2028 by the US Military, cyberspace is named as one of the five dimensions that also include air, land, sea and space. It is stated that each of these five dimensions are independent of each other, and only the cyberspace nodes (ports) are connected to each dimension.

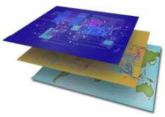
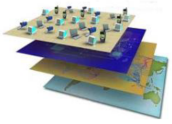
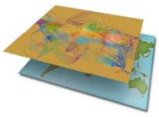
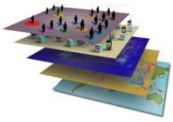
Physical Layer	**Logical Layer**	**Social Layer**
Geographic Components	Logical Network Components	Persona Components
		
Physical Network Components		Cyber Persona Components
		

Figure 9: The three layers of cyberspace
(Cyberspace Operations Concept Capability Plan 2016-2028 by the US Military)

As shown in Figure 9 on the same plan, cyberspace is examined in three main sections named physical, logical and social layers. These three layers are also divided into five different subcomponents:

- The physical layer consists of geographical and network components. Geographical components are the environments in which information systems operate depending on the existing networks. Whereas, physical network components are wired/wireless/optical infrastructures or any technical components that provide access to these infrastructures.
- The logic layer indicates the nodes to which the existing networks are connected. These include computers, smartphones, and all kinds of information systems.
- The social layer consists of both real and virtual individuals. In this context, the numbers of content shared per minute in cyberspace is very high, as shown in Figure 10.

Figure 10: What happens in an internet minute in 2018[1]

These sharing amounts are only numbers from renowned social media platforms. When critical infrastructure systems, IoT devices, and embedded system sensor topologies are also taken into account, you can see how wide the coverage is. In this context, when we examine the structure and volume of information, the term "big data," which can be defined as data produced in different formats, quickly and in large volumes, comes into play. In 2001, the parameters classified in this data collection were composed of three parts: Volume, Velocity and Variety. Whereas in 2017, the parameters consisted of 10 segments in order to analyze the data density in detail: Volume, Velocity, Variety, Value, Veracity, Volatility, Validity, Vulnerability, Variability and Visualization. This increase leads to difficulties in areas such as data storage and analysis,

1 https://www.digitalinformationworld.com/2018/05/infographic-internet-minute-2018.html# [14.02.2019].

information discovery and computational complexity, scalability and visualization of data, as well as information security. This situation is also a big challenge in the process of determining the threat parameters in terms of threat intelligence. In this context, cyber-intelligence is unlikely to be performed in real-time, or even in human control, when compared to SIGINT. For this reason, Information Systems (IS) which are used to reach the desired productivity and efficiency value in today's business world have been transformed into Management Information Systems (MIS) shown in Figure 11 which use different science disciplines efficiently.

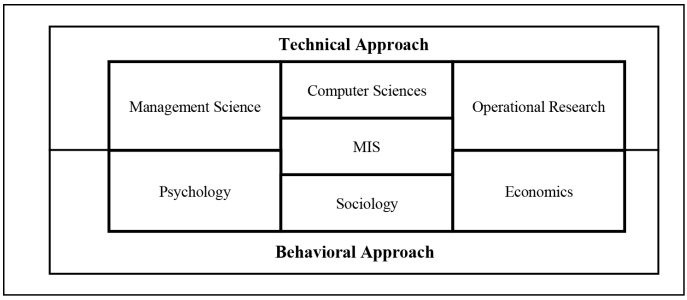


Figure 11: Basic system architecture of a management information system (Laudon, Laudon, 2014:59)

When the structure shown in Figure 11 is examined, it can be seen that MISs are using management science, computer science and operations research as well as psychology, sociology and economics under a behavioral approach. However, software systems that manage MIS need to be Artificial Intelligence (AI) controlled. Although this structure is similar to the Warning Receivers used in post-SIGINT analysis, it does not seem possible to achieve the same speed. This is as shown in Table 2 by Barkay and Dror-Reini (2015).

Table 2: Data Analysis Flow (Barkay and Dror-Reini (2015))

Process	EW SIGINT	Cyber Identity Resolution
Interception (Prevention)	Electromagnetic Signal Reception (Radar, Communication) Measurement of electronic parameters	Receipt of virtual asset activity Identifying features
Geo-Location	Correlation signals in sensors Location estimate	Cyber activity or IP association Location estimate
Association	Real-time signal tracking	Association of virtual assets – Entity resolution
Classification	Classification by signal type	Grouping by characteristics and behavior
Quality measurement	Information and uncertainty quality measurement	
Multiple Hypotheses	Grading of hypotheses and online management Removal of false alarms	
Reporting	Integration with the Intelligence Center Situation Awareness and Encouraging Early Warnings	

When the EW SIGINT and Cyber Identity Resolution processes in Table 2 are examined, it is seen that there are similarities in quality measurement, multiple hypothesis and reporting sections, although there are differences in the prevention, geo-location, association and classification steps.

7. Discussion

With the rapid development of information technology, knowledge production has increased in almost every sector, and even piles of information that can be expressed as information pollution has started. Finding and accessing the right information among these piles is of great importance. The most important problem encountered after this access is the processing of this information. At this point, ISs come into play. Thanks to ISs, the information determined to be valid is processed and this contributes to the correct, effective and efficient operation of initiatives.

However, the most important issue which should be kept in mind in this whole process is securing the information that we have or that is being researched. The cyber security phenomenon that emerged as a result of the security transformation with the use of cyber space suddenly went beyond the physical conditions and influenced community and state structures through smart networks such as social networks, websites, critical infrastructure facilities, e-government, e-bank, and e-health. This area covers all layers of society (age, education level, manager, employee, housewife, etc.). All of these layers are of interest to cyber intelligence. The prevention of any material/spiritual damage during this process depends on the establishment of cyber security awareness as well as taking the basic measures listed below on institutional and individual levels:

- The legislation on cyber security should be kept up to date on a national and international level by closely monitoring the developments in information technology, and include preventive measures to safeguard personal rights and freedoms,
- Establishing specific cyber security strategies according to scope of operation of the institution,
- Implementation policies should be developed in line with the identified strategies and kept up to date through cyber intelligence threat analysis studies,
- Develop short, medium- and long-term cyber security implementation plans within the scope of policies and strategies produced,
- Establishment of “Cyber Security Centers” structures in addition to IT Centers,

- Cyber Security Center employees should have flexible and scientific foresight to keep up with the perceived threats and receive current training to maintain their qualification,
- Since today's cyber-attacks are mostly realized through artificial intelligence software, these activities cannot be detected in time by the targeted information systems. In this context, projects aiming to increase the "threat detection speed" should be developed by giving priority to the use of artificial intelligence methods in target information system protection,
- Practical training sessions should be designed and implemented in all layers of society to increase cyber security awareness in a way that will be interactive between all these layers.

References

Abawajy, J. (2014). User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, Vol. 33, No. 3, 236–247.

http://dx.doi.org/10.1080/0144929X.2012.708787

Barkay, N., Rein, E.D. (2015). *Achieving Cyber Identity Resolution via Electronic Warfare Techniques*. RSA Conference 2015, (s. 30). Singapore, 22-24 July 2015.

Coventry, L., Briggs, P., Blythe, J., Tran, M. (2014). *Using behavioural insights to improve the public's use of cyber security best practices*. London. http://nrl.northumbria.ac.uk/id/eprint/23903/1/14-835-cyber-security-behavioural-insights.pdf

Department of Defense Dictionary of Military and Associated Terms (2010). https://fas.org/irp/doddir/dod/jp1_02.pdf

Electronic Warfare and Radar Systems Engineering Handbook. (2013). http://iaoc.org.il/wp-content/uploads/2017/07/Electronic-Warfare-and-Radar-systems-Engineering-Handbook.pdf

Forgionne, G. (2003). An Architecture for the Integration of Decision Making Support Functionalities (Ed.). *Decision Making Support Systems Achievements and Challenges for the New Decade* içinde (1-19. ss.). IGI Global Publisher of Timely Knowledge. DOI: 10.4018/978-1-59140-045-5

Gazet, A. (2010). Comparative analysis of various ransomware virii. *J Comput Virol 6*, 77–90. https://doi.org/10.1007/s11416-008-0092-2

Harrison, E. (1995). *The Managerial Decision Making Process*. Boston : Houghton Mifflin Company.

Hoisington, D. (1994). *Electronic Warfare Volume I*. Sunnyvale, CA: Lynx Publishing.

Holsapple, C. W. (2008). Decisions and Knowledge (Ed.). *Handbooks on Decision Support Systems 1:Basic Themes* içinde (21-53 ss.). Berlin: Springer. DOI: 10.1007/978-3-540-48713-5

https://www.digitalinformationworld.com/2018/05/infographic-internet-minute-2018.html#

https://www.itu.int/en/ITU-D/Cybersecurity/Documents/Introduction%20to%20the%20Concept%20of%20IT%20Security.pdf

https://www.internetworldstats.com/stats.htm

ITU-T Rec. X.1205 (04/2008) “Overview of cybersecurity”, https://www.itu.int/rec/T-REC-X.1205-200804-I

Kettani, H. and Wainwright, P. (2019). On the Top Threats to Cyber Systems. *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT),* içinde (175-179. ss.) Kahului, HI, USA: doi: 10.1109/INFOCT.2019.8711324.

Laudon, K. C. and Laudon, J. P. (2014). *Management Information Systems Managing the Digital Firm 13th Ed.* Boston: Pearson.

Ng.,C.K., Pan, L., Xiang, Y. (2018). *Honeypot Frameworks and their Applications: A New Framework*. Singapore:Springer. https://doi.org/10.1007/978-981-10-7739-5

Power, D.J. (2009). *Decision Support Basics*. New York: Business Expert Press.

Rattikorn, H. (2011). Toward Identification of Key Breakers in Social Cyber-Physical Networks. *IEEE International Conference on Systems, Man, and Cybernetics Systems*, içinde (2731-2736. ss.) Anchorage, AK, USA: doi: 10.1109/ICSMC.2011.6084086.

Sabbagh, B. A. (2012). ST(CS)2-Featuring Socio-Technical Cyber Security Warning Systems. *2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, içinde (312-316. ss.). Kuala Lumpur: doi: 10.1109/CyberSec.2012.6246110.

Sanders, J. (2016). Defining Terms: Data, Information and Knowledge . *2016 SAI Computing Conference (SAI 2016)*, içinde (223-228. Ss.). London: doi: 10.1109/SAI.2016.7555986.

Schleher, D. C. (1994). *Introduction to Electronic Warfare*. New York: Artech House.

Schleher, D. C. (1999). *Electronic Warfare in the Information Age*. Boston, London: Artech House.

Security 101. https://www.cmu.edu/iso/aware/presentation/security101-v2.pdf

Sharma, A.K., Raghuwanshi, A., Sharma, V.K. (2015). Biometric System- A Review. *International Journal of Computer Science and Information Technologies Vo.6 (5)*, 4616-4619.

Stair, R. and Reynolds, G. (2009). *Fundementals of Information Systems.* UK: Cengage Learning Inc.

The United States Army’s Cyberspace Operations Concept Capability Plan 2016-2028 http://www.fas.org/irp/doddir/army/pam525-7-8.pdf

Unar, J. A., Seng, W.C., Abbasia, A. (2014). A review Of Biometric Technology Along With Trends And Prospects. *Pattern Recognition Volume 47, Issue 8, August 2014.* 2673-2688. https://doi.org/10.1016/j.patcog.2014.01.016

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. West Sussex: John Wiley & Sons Ltd.

CHAPTER 15

AUTOMATIC MEASUREMENT OF THE MORPHOLOGICAL CHARACTERISTICS OF HONEYBEES WITH A COMPUTATIONAL PROGRAM

Zlatin ZLATEV*, Veselina NEDEVA*, Ivanka ZHELYAZKOVA**

*Trakia university, Faculty of "Technics and technologies" Yambol, 38 Graf Ignatiev str., 8602, Yambol, Bulgaria
e-mail: zlatin.zlatev@trakia-uni.bg; vnedeva@yahoo.com

**Trakia University, Agrarian faculty, Stara Zagora, Bulgaria
e-mail: izhel@uni-sz.bg

DOI: 10.26650/B/ET06.2020.011.15

Abstract

The use of Big data related to the breeding of honey bees, when administered and processed effectively, will encourage the development of knowledge-based beekeeping, create new markets and business opportunities and further encourage the development of this industry. There have been attempts to fully automate the process of measuring the morphological characteristics of bees (at this stage there are conversions for Measuring wings), but this process for other parts are still completed manually. A survey was made of the possibilities to automate the process of measuring the morphological characteristics in honeybees and the proposed algorithm and program to implement it. Color characteristics of parts of the bee body - tergite and proboscis, through which they can be separated from the background of the image, are analyzed and measured.Distances are determined between the values of the colour components of the object and background. From statistical analysis, it is found that S and V colour components of the HSV colour model are appropriate for the separation of an object from the background . Algorithms and a program in Matlab environment for separating tergite and proboscis from the background of the image and definition of their main sizes are developed. From the analysis of the results, it is found that the major influence on the accuracy of the measurement is of the bee in the image..

Keywords: Honey bee, Morphometric characteristics, Colour components, Measurement algorithm

Introduction

1. Methods for the automatic measurement of morphological characteristics of honeybees

Big data plays a major role in the development of methods and techniques for the proper breeding of honey bees. Bees are of economic importance, not only for their honey and bee products but also from the point of view of the farm, through the pollination of different crops.

Therefore, the promotion of efficient breeding of honey bees, reducing the cost of rearing, reducing the problems associated with them requires the collection, processing, and analysis of large amounts of data.

The use of Big data related to the breeding of honey bees, when administered and processed effectively, will encourage the development of knowledge-based beekeeping, create new markets and business opportunities and further encourage the development of this industry. It is imperative that they make full use of its unique advantages in promoting the development of science and technology through the methods of obtaining, processing and analyzing big data.

Honeybees are spread across Africa, Europe and parts of Asia. They have a wide variety of species and subspecies, which can be identified by morphological characteristics. These characteristics vary depending on the environmental conditions to which they are adapted and live (Snodgrass, 2010; Abou-Shaara et al., 2012; Abou-Shaara et al., 2013).

In recent years, morphometric analysis as a tool for characterizing honey bees has become essential in the search for solutions to the problem of mortality in bees and the collapse of their colonies.

In traditional beekeeping, more and more modern methods of measurement and management are being implemented through various automated and intelligent systems in order to optimize the processes of bee breeding and solve modern problems such as honey bee mortality.

At the current level of development of science and technology, different sensors are used in bee breeding to measure temperature and humidity - video sensors and strain gauges. Different methods are used to communicate these measurements and control the instruments. The usefulness of the application of these systems in beekeeping is related to the food consumed, the survival of the colony in winter, and the production of nectar. Also, measuring and control systems are suitable for precise management of the microclimate in the hives, which greatly improves the conditions for the existence of bees.

The listed problems, methods and technical tools for solving them are related to the collection, processing, analysis, and transmission of large amounts of data via communication lines. These machine learning and artificial intelligence tasks are connected with large amounts of data to be used effectively.

The separation between subtypes is important for the cultivation and preservation of biodiversity (Abou-Shaara, 2013; Abou-Shaara et al., 2013).

The morphological characteristics are associated with the productive characteristics of the bee colony (El-Aw et al., 2012). The morphological characteristics of the bees can be measured by a variety of methods. Basically, these indicators are used to determine the individual characteristics and subtypes, as well as for determining the hybridization with other subtypes.

The measurements are used to determine the influence of the queen bee used for the purity of the subspecies. Studies show that mainly morphological characteristics are affected by environmental factors. For example, in mountain areas, the proboscis is longer than those of bees reared in lowland regions.

There are studies which search for a correlation between the amount of produced honey and separation of subspecies by several morphological characteristics (Meixner et al., 2007; Miladenovic et al., 2011; Abou-Shaara, 2013). Hind wing and cubital index are the most measured part of the bee – manually and by automatic techniques. The final result of the measurement of morphological characteristics largely depends on the knowledge and experience of experts - based on subjective factors.

A general trend in recent years has been the demand for methods to increase the efficiency of this process, resulting in improved accuracy of the estimates, reducing the time in which they are performed, and especially to minimize the subjectivity in the measurement process. Many new hardware and software tools designed to measure the morphological parameters in bees still do not solve most of the problems associated with measuring accuracy in routine processing steps and operations that are performed visually by humans. In the modern era computer-based measurement methods are preferred to accelerate the process.

Suitable devices for obtaining images are video camera, camera, and scanner (Lazarov, 2016). Regardless of popular literature attempts to fully automate the process of measuring the morphological characteristics of bees, at this stage, it is still done by manual way (Mattu et al., 1984; Roth et al., 1999; Tofilski, 2004; Mostajeran et al., 2006; Abou-Shaara et al., 2013; Zlatev et al., 2017).

The work is organized in the following sequence: An overview of the known solutions for automated measurement of morphological characteristics of honeybees and the results are analysed; The materials and methods that were used in the study are presented; the algorithms for automated measurement of key dimensions at tergite and proboscis are developed and tested; the obtained results are discussed and summarized in conclusion.

In most of contemporary studies mainly front wing is measured. (Miladenovic et al., 2011; Bouga, 2011; Santana et al., 2014; Silvaa et al., 2015). Publications related to the measurement of other parts of bees state that a binocular microscope with an eyepiece micrometer is mainly used. This method is not very high precision, is time-consuming and the measuring depends on the experience (Strauss et al., 1994; Schroder et al., 2002).

El-Aw et al. (2012) offer the measurement of morphological characteristics of bees to be performed with a scanner. They propose Photoshop software to be used for image processing.

A comparative analysis is made between measurements with the proposed simplified method and those with binocular microscope eyepiece measurement with a micrometer.

From the measured three colonies of bees, large differences in measurements were obtained in the first colony between the length of the rostrum and rear wing.

(Lazarov, 2016) states that the results are obtained in increments, then multiplied by a factor depending on the increase to be recalculated in millimeters. Some chitin portions, the length of the front-side length of the proboscis, can not be measured entirely but are divided into 2 parts. All these activities increase the potential for errors. Work on the standard method requires more time to reach the final results. When working with the AutoCAD program , a scanner with high resolution and a computer are required. The dimensions of chitin parts are automatically received in millimeters. The measurements are performed quickly. The objects of measurement are scanned and can be stored for a long time. The results of the control determination sections of graph paper and use of Gauge Block (Certificate of Calibration No.1409914, Mitutoyo Corporation Miyazaki Plant, Japan) gave him a reason to accept that measurements with program AutoCAD are accurate and the program can be successfully applied to determine the morphological characteristics of the body of the worker bees.

(Abou-Shaara et al., 2012) offer a four-step methodology for computing measurements of the morphological characteristics of bees. The first step is collecting samples – taking 15 workers from the colony and exploring 6 colonies. The second step is sample preparation –

bees are frozen or placed in alcohol and prepared on glass slides. The third step involves the measurement – the prepared glass slides are scanned and a computer program is implemented for measurement. The fourth step is data analysis – calculation of the mean values and standard deviations or the use of more complex statistical procedures.

According to the literature, there have been attempts to fully automate the process of measuring the morphological characteristics of bees (at this stage there are conversions for Measuring wings), but this process for other parts is still carried out by hand (Tofilski, 2007). According to the author of (Tofilski, 2008) the development of this measuring system is subject to the interrelated issues - the construction of a model of the measured elements and the construction of an algorithm for operation of the system.

In (Schroder et al., 2002) the authors propose a system for automatic measurement of geometric parameters of the wings of bees. The system consists of a laptop computer and a stereo microscope with an integrated CCD camera connected to the computer via a video adapter installed into the PCMCIA slot. The software offers recognition and measurement of the elements of the front wing and discriminant analysis. The effectiveness of the system proposed by the authors was checked with 469 specimens of 13 species of bees and the reported accuracy in distinguishing species by discriminant analysis was 99.15%. The authors state that training the classifier needs a large amount of information to build a database with information about the species and subspecies of bees.

One of the famous pieces of software for automatic measurement of the wings of insects (including bees) is DrawWing (Tofilski, 2004; Tofilski, 2016). The authors of the software say it has a much better device for discriminant analysis compared to known developments in the field (Strauss et al., 1994; Roth et al., 1999).

Another development related to the automatic measurement of morphological characteristics of insects, including bees is the software MorphoJ (Klingenberg, 2011), which combines the method “Procrustes superimposition” with a number of other methods for analysis of form. The program provides an integrated user interface. The program provides an integrated user interface. The advantage of this software over other renowned software is that it offers multivariate analysis, principal component analysis, discriminant analysis, and multivariate regression. The product is Java-based, which makes it compatible with a variety of operating systems such as Windows, Mac OS, Unix, Linux.

The morphological characteristics of honey bees can be measured for various reasons. Basically, these metrics are used to determine subspecies and individual characteristics, as

well as to determine hybridization with other subspecies. The measurements are used to determine the influence of the used queen bee and the purity of the subspecies.

Table 1 lists the more commonly used methods for automated measurement of the basic morphological characteristics of honey bees. Here are the main controlled dimensions. In the column-defined characteristics, it is described what the measurements of the respective part and the literary source are used for, where this study is presented, which is indicative of the importance of this type of measurement.

Table 1. Measured morphological characteristics of honey bees

Dimensions measured	Automated measurement method	Aim of the study	Reference
Tongue, Proboscis	Microscope with video camera	Subdivision of subspecies. Characteristic of the geographical area. Quantity of honey produced.	(Waddington, 1989; Marghitas et al., 2008)
24 morphological characteristics	For the first time, it uses a combination of a scanner and a vector image processing software	A comparative analysis of automated measurement technique and classical laboratory method	(Lazarov, 2016; Lazarov, 2017)
Fore wing	Stereo microscope calibrated with micrometer	Quantity of honey produced. Subdivision of subspecies. Colony Productivity	(Mosterjeran et al., 2002)
Hind wing, Cubital index	Stereo microscope calibrated with micrometer	Quantity of honey produced	(Mosterjeran, 2002; Mosterjeran, 2006)
Tergite 3 and 4	Comparative analysis of used manual and automated methods and their application in different countries	Subdivision of subspecies	(Burga, 2011)
Metatarsus	Stereo microscope calibrated with micrometer	Quantity of honey produced. Colony Productivity	(Mosterjeran, 2006)
Sternite	Stereo microscope calibrated with micrometer	Different size depending on the measurement season	(Mattu et al., 1984)
Proboscis	The indirect features used are suitable for predicting the functional length of the proboscis	Predicting the functional length of the proboscis	(Waddington et al., 1987)
Length of tongue, wing dimensions, cubic vein, number of hooks, hind leg	Measurement of multiple morphological characteristics of honeybees through a scanner and a raster image processing software	A comparative analysis of automated measurement technique and classical laboratory method	(El-Aw et al., 2012)
Proboscis	The authors measured intertegular distance (as a measure of body size) and proboscis length (glossa and prementum, both individually and combined). Using linear models and model selection, we determined which parameters provided the best estimate of proboscis length.	Allometric relationships makes them a potentially useful tool for estimating ecologically important traits that are otherwise difficult to measure	(Cariveau et al., 2016)

The analysis of known research related to the measurement of morphological characteristics of bees shows that for this purpose the following methods are used:

Classical – using a stereomicroscope and a magnifying glass;

Computer – using software products for general-purpose and specialized.

Development and research into the measurement of morphological characteristics of bees includes improvements to existing or creation of new methods for manual, automated and automatic measurement. The main characteristics, which are the fous of these studies, relate to the measurement of parameters of the wings.

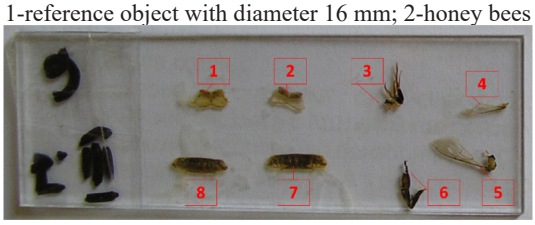
After a review of publications on this topic, it is found that there are few publications in which automatic way measure ment of other body parts of bees such as tergite, foot, proboscis. They are important for determining the subspecies, the productivity of the bee colony, the influence of the geographical area.

2. Material and methods used in this study

Figure 1 shows part of the bees used in this work and one of the prepared samples. The sample consists of the front right wing, rear right wing, right hind leg, third and fourth tergites, sternit and proboscis.

Samples were prepared in the laboratory of the section "Beekeeping" at the Trakia University – Stara Zagora, Bulgaria.

a) part of the bees used
1-reference object with diameter 16 mm; 2-honey bees



b) preparation of the samples
1-Sternite; 2- Sternite; 3- Proboscis; 4- Fore wing; 5- Hind wing; 6- Metatarsus; 7-Tergite; 8-Tergite

Figure 1: Samples used in the study – general view

The verification of the influence of the angle of rotation of the object on the accuracy of measurement is made by the unit of relative measurement error reversed ε,%.

Table 1 is a description of the analysed functions of the distance between the colour components. Those used are the distance of Mahalanobis (Mahalanobis), Euclidean distance (Euclidean), a distance of Manhattan (Cityblock), Chebyshev distance function and Fisher distance (Fisher discriminant ratio) (Tofilski, 2004).

The resulting distances are processed with the method of correspondences analysis (CA) (Kazlacheva, 2011; Kazlacheva et al., 2014) of the software Statistica. Informative colour features are determined on the basis of a certain available methodology (Georgieva et al., 2015; Mladenov et al., 2015; Dimitrova, 2016; Zlatev et al., 2017).

Table 2. Distance functions used in the study

Designation	**Formula**	**Description**
Mahalanobis	${d_{st}}^2 = (x_s - y_t)C^{-1}(x_s - y_t)'$	C – covariance matrix
Euclidean	${d_{st}}^2 = (x_s - y_t)C^{-1}(x_s - y_t)'$	x and y are the compared vectors
Cityblock	$d_{st} = \sum_{j=1}^{n} \lvert x_{sj} - y_{tj} \rvert$	
Chebychev	$d_{st} = max_j\{\lvert x_{sj} - y_{tj} \rvert\}$	max – maximum
Fisher discriminant ratio	$d_{st} = \frac{(\bar{x} - \bar{y})^2}{{SD_x}^2 + {SD_y}^2}$	SD – standard deviation

The OCTAVE (GNU Octave)Program Platform was selected because it offers a high-level programming language, interactive algorithm development environment, visualization, data analysis, and calculations. Octave is compatible with Matlab (The Mathworks Inc.) and is used in many areas such as signal processing, imaging, spectral characteristics, and research systems for automatic control. There are a number of toolboxes containing embedded features, including imaging libraries. Vector and matrix operations (which are key to engineering calculations and image processing), are supported. This programming environment offers a quick algorithm development that focuses the user on the problem solver rather than the details of the program code.

The analysis of the developed algorithm aims to determine the extent to which its output variables are affected by moderate changes to the input data.

Algorithm testing can provide a general assessment of its accuracy, as well as detailed information to overcome errors at different input data values.

There are a number of methods for analyzing algorithms (Tofilski, 2011; Klingenberg, 2011; Abou-Shaara et al., 2013; Zlatev et al., 2017). One of these is by changing the input parameters by ± 10%.

Correspondence analysis. The analysis is performed with a table with frequencies, C, of size mxn where m is the number of rows and n is the number of columns. The vectors w_m and w_n give the marginal probabilities of being the row and column classes respectively, while S gives the joint probability distribution of rows and columns. Therefore M gives deviations from independence. These deviations, squared and appropriately scaled, are summed up to yield the chi-squared statistic in C. The data processing steps with the Correspondence Analysis method are presented in Table 3.

Table 3. Stages of data processing with the Correspondence Analysis method

Stage	Formula	Description
A	$w_m = \frac{1}{n_C} C1$	From Table C weights are calculated by rows w_m
B	$w_n = \frac{1}{n_C} 1^T C$	From Table C weights are calculated by columns w_n
C	$n_c = \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}$	n_c is number of observations, 1 is a vector column of ones with the dimensionality of the data
D	$S = \frac{1}{n_C} C$	Table S is calculated as C is divided by the sum of the elements in it
E	$M = S - w_m w_n$	Table M of S and weights are calculated
F	$W_m = diag\left\{\frac{1}{w_m}\right\}$ $W_n = diag\left\{\frac{1}{w_n}\right\}$ $M = U\Sigma V^*$	Table M is decomposed with a generalized decomposition of singular values. The diagonal elements of W_n are $1/w_n$ and those that are not diagonal are 0, where $U^* W_m U = V^* W_n V = I$
G	$F_m = W_m U \Sigma$ $F_n = W_n V \Sigma$	Factor coefficients for the rows and columns of the matrix C are determined

The influence of the angle of rotation of the object on the measurement accuracy is checked by means of the relative measurement error module ε,%, which is determined by the following relationship (Klingenberg, 2011; Georgiev et al., 2014; Zlatev et al., 2017):

$$\varepsilon = \left| \frac{L_{meas} - L_{ref}}{L_{ref}} \right| . 100, \% \qquad (1)$$

where L_{meas} is a dimension measured by the proposed algorithm; L_{ref} - measurement by reference method.

3. Results obtained and discussion

An algorithm for measuring morphological characteristics of the bees has been developed. The proposed algorithm and its implementation are based on express, contactless measurement of elements from the body of bees using image processing techniques. It should be emphatically stressed that its establishment is not intended to replace or substitute authorized and approved in practice methods for measurement of these dimensions.

Table 4 shows the resulting distances which separate the object from the background using the colour features of six colour models – RGB, HSV, Lab, LCH, XYZ, CMYK.

Table 4. Distances between colour components for object and background

D	Mahalanobis		Euclidean		CityBlock		Minkowski		Chebichev		FDR	
A / CC	T-B	P-B	T-B	P-B	T-B	P-B	T-B	P-B	T-B	P-B	T-B	P-B
R	1,758	1,766	34,006	43,634	42,757	53,726	34,006	43,634	31,004	40,430	2,091	2,600
G	1,752	1,771	34,387	37,785	42,914	47,428	34,387	37,785	31,480	34,394	4,310	6,552
B	1,778	1,746	38,940	38,440	48,387	47,148	38,940	38,440	35,831	35,675	4,538	7,892
H	1,666	1,706	0,388	0,324	0,458	0,393	0,388	0,324	0,368	0,302	0,462	0,953
S	1,795	1,786	0,190	0,268	0,237	0,321	0,190	0,268	0,176	0,255	0,790	1,462
V	1,767	1,768	0,166	0,198	0,209	0,250	0,166	0,198	0,151	0,180	2,577	3,184
L	1,762	1,765	37,517	42,535	47,385	53,327	37,517	42,535	33,979	38,637	3,961	5,642
a	1,698	1,617	5,713	8,254	7,152	9,781	5,713	8,254	5,150	7,757	1,536	0,853
b	1,772	1,755	9,759	10,419	12,406	13,135	9,759	10,419	8,781	9,457	1,093	1,165
L	1,762	1,765	14,712	16,680	18,582	20,913	14,712	16,680	13,325	15,152	3,961	5,642
C	1,768	1,770	7,593	11,214	9,580	13,482	7,593	11,214	6,894	10,611	0,027	0,019
H	1,684	1,732	136,390	115,048	165,118	142,763	136,390	115,048	127,287	105,520	0,757	1,726
X	1,696	1,707	2,370	2,364	2,829	2,807	2,370	2,364	2,236	2,230	2,484	3,288
Y	1,688	1,710	2,552	2,511	2,999	2,922	2,552	2,511	2,430	2,398	2,627	3,513
Z	1,722	1,663	2,554	2,490	2,890	2,714	2,554	2,490	2,479	2,446	2,643	3,362
C	1,782	1,769	33,415	46,333	41,012	54,554	33,415	46,333	31,070	44,363	0,409	1,019
M	1,755	1,704	24,936	27,832	31,349	34,693	24,936	27,832	22,661	25,457	9,278	13,810
Y	1,775	1,723	46,758	37,010	59,227	47,171	46,758	37,010	42,295	33,282	3,545	10,411
K	1,767	1,762	45,293	51,179	56,675	64,237	46,758	37,010	41,408	46,424	2,789	3,669

D-distance; A-area; O-object area; CC-colour component; T-B-tergite-background; P-B-proboscis-background

Figure 2 presents the results of correspondence analysis in removing the tergite from the background. It is seen that with the largest distances are the colour components of the HSV colour model.

Figure 3 presents the results of correspondences analysis for separating the proboscis from the background. In this case, once again, the colour components from the HSV colour model are suitable for this purpose.

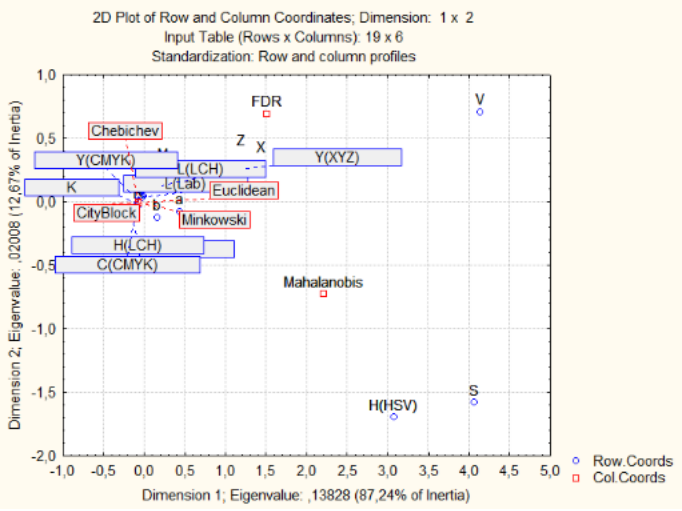


Figure 2: Selection of colour features for separation of tergite from background

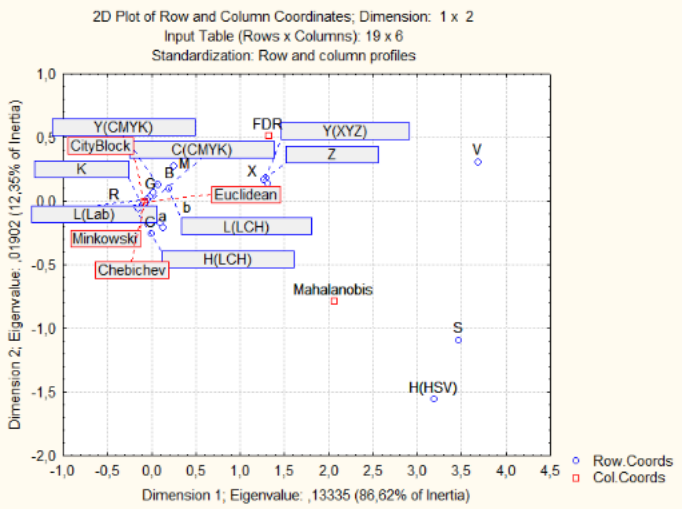


Figure 3: Selection of colour features for separation of proboscis from background

An algorithm for automatic measurement of the main dimensions of tergite. The algorithm for measuring the main dimensions of the tergite is described in Table 5. The original RGB image is transformed into a HSV colour model. Experimentally it has been found that the separation of the tergite from the background in the image is a suitable V (HSV) colour component. The image is transformed into black and white. This conversion is defined as the threshold of binarization, which depends on which pixels will be converted into white and which in black. To remove image noise as points, small objects, it is filtered with a filter of type "Disk". The short and long axes of the tergite are determined by calculation procedures. The distance between the excrescences is defined as defined peaks in the resulting contour around the object. Functions for displaying the results and writing to a file for archiving and subsequent processing of data of the tergite have been introduced.

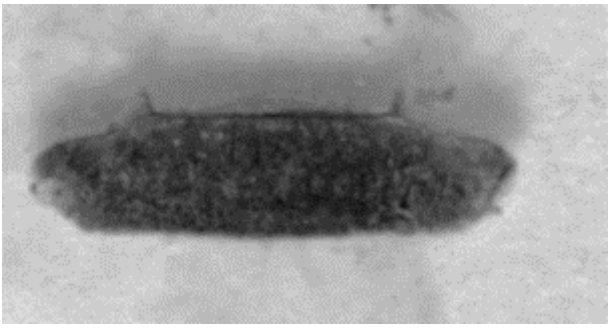
Table 5. An algorithm for measurement of dimensions of tergite

Stage	Description	Pseudocode
A	Loading of the original image	i=imread('Image.jpg')
B	Leveling the object in the image	i=imrotate(i,angle)
C	Conversion in HSV colour model and extraction of V component	i1=rgb2hsv(i); i2=i(:,:,3)
D	Conversion of the image in black and white	i3=im2bw(i2,0.19) The threshold for segmentation is determined experimentally
E	Filtering of the image	h=fspecial('disk',6); i3=imfilter(i3,h)
F	Removing of noises	i3=bwareaopen(i3,1500,4)
G	Obtaining of the object dimensions	stats = regionprops(i3,'all')
H	Finding of the contour of the object	B = bwboundaries(i3);
I	Finding of the long axis of the contour	for k1 = 1:length(B1); b(k1,:) = B1(k1,:); end; s1=b(:,2); t1=-b(:,1); k1=round(length(t1)/4); k2=round(length(t1)/1.33); x1n=s1(k1); y1n=-t1(k1); x2n=s1(k2); y2n=-t1(k2); ds=sqrt((x2n-x1n)^2+(y2n-y1n)^2)
J	Finding of short axis of the contour	for k1 = 1:length(B1); b(k1,:) = B1(k1,:); end; s1=b(:,2); t1=-b(:,1); k1=round(length(t1)/4); k2=round(length(t1)/1.33); x1n=s1(k1); y1n=-t1(k1); x2n=s1(k2); y2n=-t1(k2); ds=sqrt((x2n-x1n)^2+(y2n-y1n)^2)
K	Determining the distance between excrescences	for k1 = 1:length(B1); b(k1,:) = B1(k1,:); end; s1=b(:,2); t1=-b(:,1) k1=round(length(t1)/4); k2=round(length(t1)/1.33); x1n=s1(k1); y1n=-t1(k1); x2n=s1(k2); y2n=-t1(k2); ds=sqrt((x2n-x1n)^2+(y2n-y1n)^2)
L	Visualization of the results	Displaying lines for the main dimensions and displaying measured values, Functions Figure, Line and Text
M	Summarizing the results in a table and saving in a file	The table is stored in the workspace and by function Save is saved in a file with measurements, converted to millimeters, mm.

Figure 4 shows an example of the work of the algorithm to measure the main dimensions of the tergite. The measurements are presented in pixels.

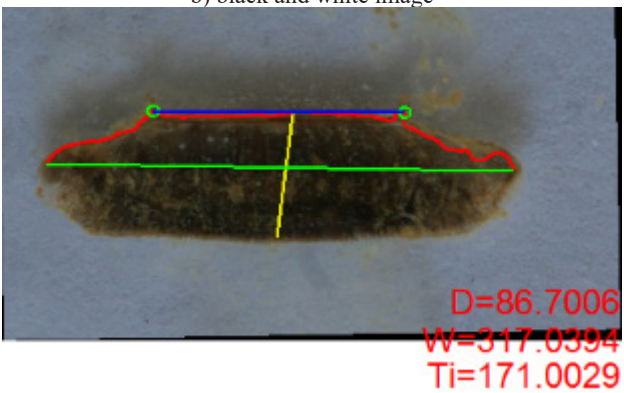
The image from V (HSV) color component is converted to binary one.

After applying the treatments set out in the presented algorithm, the results are displayed.



a) V (HSV) color component

b) black and white image



c) visualization of the results
D-short axis (Longitudinal diameter); W-long axis; Ti-distance between two peaks of the tergite

Figure 4: Stages of the work of the algorithm for automatic measurement of tergitee

The rotation of the object of a certain angle does not affect the two main sizes - long and short axis of the tergite. It affects mainly the measured distance between the two appendages.

Figure 5 shows the results of the algorithm check, with the angle of the object in the image being changed ± 10% relative to the horizontal axis, whereby the measured values are obtained with a small error relative to the reference measurement.

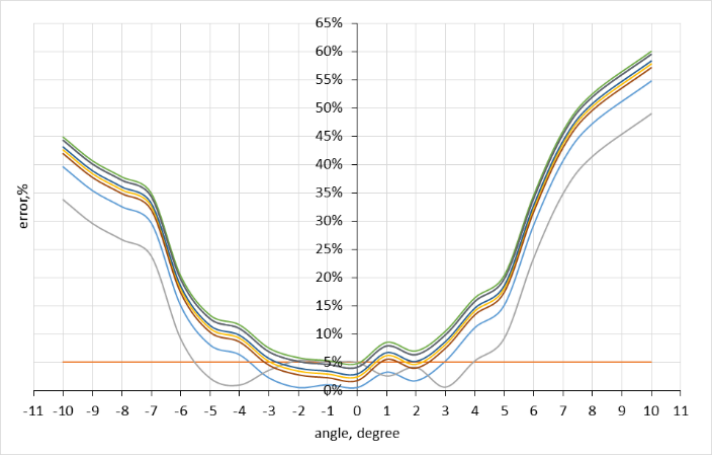
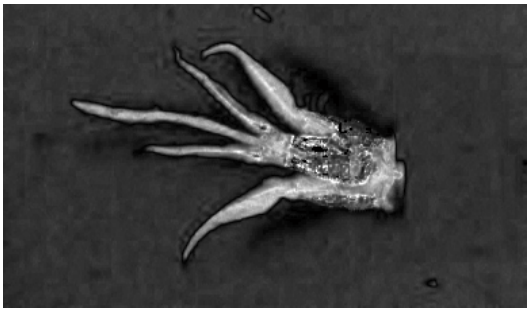


Figure 5: Relative error in altering the angle of tertiary with ±10°

The results of this analysis show that - with an error of up to 5% - the algorithm operates at an angle of rotation of the object relative to the horizontal axis from -2° to +1°.

An algorithm for automatic measurement of the main dimensions in proboscis. The algorithm for measuring the proboscis is built in the same manner as that for determining the main dimensions of tergite. The difference is that instead of the V (HSV), the S (HSV) colour component is used and uses an indexed image, rather than black and white.

Figure 6 shows the steps of operation of the algorithm for measuring the proboscis. The image from S (HSV) color component is converted to binary one. and this two-dimensional image is binarized. The results are displayed, the length of the proboscis is set along the longest axis of the object. It is observed that as in tergite, the change in the angle of the location of the object in the image also affect measurement accuracy. The measurements on the figure are presented in pixels. In the analysis of algorithms it is found that the measurement of the main dimensions in tergite and proboscis of honey bees is accurate to within 5% deviation ±1.5° the longest axis from the object to the horizontal axis of the image.



a) S (HSV) colour component

b) indexed image

c) visualization of the results, W-proboscis length

Figure 6: Stages of algorithm work

Figure 7 shows a graph of the module from a relative error in measuring the length of a shoe according to the angle of the object's position relative to the horizontal axis. An error of up to 5% is obtained by changing the angle of the object relative to the horizontal axis up to ±1.5°.

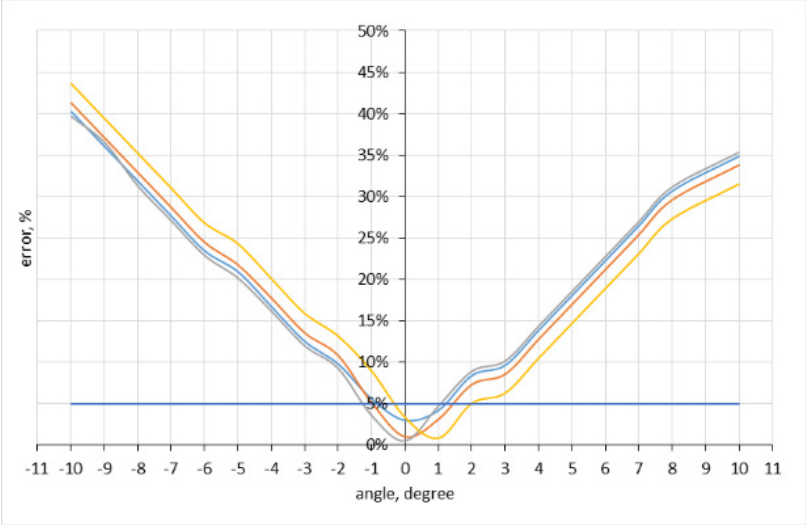


Figure7: Relative error when changing the angle of a proboscis with ±10°

The advantages of automated methods for measuring the basic morphological characteristics of honeybees compared to the standard methods for determining these sizes can be summarized as follow:

✓ Standard measurement methods for morphological characteristics of honey bees include the use of a stereo magnifying glass and a microscope. These methods do not achieve high accuracy and the correctness of measurement will depend on the experience and qualification of the lab worker;

✓ Developed methods for automated measurement of morphological characteristics of honey bees where color digital images of individual parts of a bee, such as a rear right foot, a front wing, a tergite obtained with a video camera or scanner, are handled using accessible software products such as PhotoShop, CorelDraw, ImagePro;

✓ The authors of the known developments offer software products and automatic measurement algorithms that have functions of binarization, extraction of properties and classification of parts of insects, including bees, which can easily be adapted to other measurement software programs;

The advantages of automated systems for measuring the basic morphological characteristics of honey bees are undoubtedly many, as they have, in the main, huge potential, comparatively low cost of equipment, lack of complexity of management, speed, and high

productivity. These provide greater reliability and security when performing measurements which make it easier for farmers and workers to save physical labor and time.

Creating new and optimizing existing mathematical models and statistical methods are also the main areas of work in this field. The implementation of these algorithms is important for retrieving, transforming and using information to determine the basic dimensions of a honey bee to use in behavioral analysis of these insects.

Developments and studies in the measurement of morphological characteristics of bees include the improvement of existing or the creation of new methods for manual, automated and automatic measurement. The main features of these studies are related to the measurement of wing parameters.

(Waddington, 1989) investigated the length of the tergite of bees using an indirect indicator - the morphometric characteristics of the bee left wing and head width. The authors report that the indirect features used are appropriate for predicting the functional length of the scapula. In this way, the size of the difficult to measure puppy can be determined by the relatively easy to measure parts of the bee.

The method proposed by (El-Aw et al., 2012) for measuring multiple morphological characteristics of honey bees (length of the tongue, wing dimensions, cubic vein, number of hooks, hind legs) with scanner and Photoshop software shows that measurements of those parts of bees can be measured with sufficient precision compared to the classic method using a stereo magnifying glass.

The method presented in the present work complements these studies by offering automated measurement of the less-regarded parts of the bees – proboscis, and tergite. The method proposed herein may be further developed to use indirectly to determine the size of a tergite and a proboscis in other more easily measurable parts of the bee.

One such trait, proboscis length in bees, is assumed to be important in structuring bee communities and plant pollinator networks (Cariveau et al., 2016). However, it is difficult to measure and thus rarely included in ecological analyses. We measured intertegular distance (as a measure of body size) and proboscis length (glossa and prementum, both individually and combined).

Using linear models and model selection, we determined which parameters provided the best estimate of proboscis length. We then used coefficients to estimate the relationship between intertegular distance and proboscis length, while also considering family. Using

allometric equations with an estimation for a scaling coefficient between intertegular distance and proboscis length and coefficients for each family, we explain 91% of the variance in species-level means for bee proboscis length among bee species.

The predictive nature of allometric relationships makes them a potentially useful tool for estimating ecologically important traits that are otherwise difficult to measure. Here we take this approach to develop a predictive allometric equation for proboscis length in bees.

There are software products that use techniques for obtaining, processing and image analysis.

Practically, there are proven methods for measuring the size of morphological characteristics in bees, using the software products Corel Draw, AutoCAD, PhotoShop.

The use of software in two main areas – the development of their own programs using programming languages or measurement with those which have a graphical user interface.

The creation of software for processing and analysis of images using programming languages such as C or Delphi, requires knowledge of compiling the programs and experience of measuring, but on the other hand, this way of working is flexible and the algorithm can be modified according to the needs of the particular user.

Using software with a user interface allows for quick and secure measurement and does not require programming knowledge to work with it. These software products have the disadvantage that they lack flexibility and they can be configured according to the requirements of the measurement.

The selection of software and how to work with it depend on the capabilities of those who use it. Essential for this choice are two factors – flexibility and ease of use. In the development of systems for automatic measurement of morphological characteristics of insects and beekeepers, interconnected problems arise – the construction of a model of the measured elements and the construction of an algorithm for the functioning of the system.

The software system proposed here for automated measurement of two major parts of bees partially solves these problems. More studies can be made on the application of the measurement system and other less-measured parts of bees with automated systems. When applying this adapted measurement approach, it is also possible to determine the dimensions of the measurable parts of the bee.

For the most part, the methods used to analyze the behavior of bees are subjective or require considerable processing time. The accuracy of diagnosis is not high and depends on the expert's qualifications. That is why the creation of highly efficient automated technologies for determining the morphological characteristics of bees is a priority goal of current research in this field.

There is little research on the impact of the environment in which honey bees are grown. The question of whether obtaining, processing and analyzing data on the behavior of bees, through their morphological characteristics, can be implemented expressly, efficiently, with a small number of computational operations, remains open and unclear.

An approach for determining the morphological characteristics of honey bees based on color digital imaging data has been adapted based on extracted features and recognition and measurement of the morphological characteristics of bees.

From the study conducted to determine the size of parts of bees, using an image acquisition, processing and analysis technique, it was found that this can be realized with a total error of less than 10%, which is an indication of sufficient accuracy in the analysis of the morphological parts of honey bees.

Analytical dependencies are derived through distance functions. They have been shown to be effective in solving the bee size determination task within the study.

The results were obtained to improve and complement those reported in the available literature. They can be used to refine the approaches and methods used so far to determine the morphological characteristics of bees, as input to determine the causes of their mortality and the collapse of their colonies.

The proposed methods and software tools could be used in the development of mobile applications and methods for remote measurement, in the express determination of the morphological characteristics of bees.

4. Conclusion

Based on a detailed analysis it was found that morphometric measurements are an important criterion in the selection programs of worker bees. The common way to measure chitin body parts in honeybees is through a stereo microscope with an eyepiece micrometer. At the current level of science and technology, semi-automatic measurement of body parts of bees are made. There have been attempts to fully automate the process of measuring the morphological characteristics of bees (at this stage, there are conversions for Measuring wings), but this process for other parts is still done manually.

Bee body parts (tergite and proboscis), are measured and analyzed by color characteristics. these parts can be separated from the background of the image. Separation functions are defined by colour components. These components can be used to separate an object from a background. From statistical analysisit is found that the S and V colour components from the HSV colour model are appropriate for the separation of an object from the background .

The present work is adapted to an automated measurement approach applied to the basic morphological characteristics of honey bees by analyzing color images, which is studied with two main parts of the bees - tergite and proboscis.

Algorithms were developed and programmed in a Matlab environment for separating the tergite and proboscis from the background of the image and definition of their main dimensions through selected colour components of the HSV colour model. These algorithms complement the existing ones. Other parts of the body were measured besides the bee's wing. From the analysis of the results, it is found that the major influence on the accuracy of the measurement is the angle of the bee body part in the image.

The main dimensions in tergite and proboscis of honey bees is accurate to within 5% deviation $\pm 1{,}5^{o}$ the longest axis from the object to the horizontal axis of the image.

The results presented in the paper can be used as a basis for building databases, in part the morphological characteristics of bees. The measurement accuracy and efficiency of the proposed algorithms and procedures can be the basis for obtaining high quality data, suitable for machine learning and solving the problem of data standardization, regardless of the region of cultivation, geographical, features, traditions of breeding and feeding of bees.

Bees are important to the world economy because they pollinate basic crops from agricultural production.

A recent problem is their decline in recent years. In connection with this problem, various technical and technological methods and means for analyzing the behavior of bees have been developed.

These analysis tools generate large amounts of data that need to be processed in order to retrieve information to analyze the causes of the problem of falling bee numbers and finding ways to solve the problem.

A good way to solve the problem of large volumes of data is through the methods of modern Big Data science. Using Big Data Analysis methods, data from multiple sources can

be quickly analyzed and used to retrieve information related to identifying causes of bee mortality and colony decay.

Getting effective solutions to bee-breeding problems involves creating standardized data. This data can be used as a benchmark when compiling effective decision-making algorithms related to effective bee breeding. When compiling databases of benchmarks, it is necessary to take into account the geographical differences in bee rearing sites, genetic differences, specific practices for rearing in different regions. A significant difference is also the language differences - the languages spoken in different countries and the sharing of information through them.

When creating databases of standardized data useful for beekeeping, problems with the organization of data, the form of such data, the quality and informativeness of the data, the rights to use them are more common. It is also necessary to create databases of supporting information for beekeepers.

References

Abou-Shaara, H. (2013). Wing Venation Characteristics of Honey Bees. *J. Apicult*., 28, 79-86.

Abou-Shaara, H., Al-Ghamdi, A., Mohamed, A. (2013). Body morphological characteristics of honey bees. *Agricultura*, 10(1-2), 45-49.

Abou-Shaara, H., Draz, K., Al-Aw, M., Eid, K. (2012). Stability of honey bee morphological characters within open populations. *Bee science*, 12(1), 31-37.

Bouga, M. (2011). A review of methods for discrimination of honey bee populations as applied to European beekeeping – Review article. *Journal of Apicultural Research*, 50(1), 51-84. DOI: 10.3896/IBRA.1.50.1.06

Cariveau, D., et al. (2016). The allometry of bee proboscis length and its uses in ecology. *PloS one*, 11(3), e0151-482. DOI:10.1371/journal.pone.0151482

Dimitrova, A. (2016). Analysis of SEM images of magnetically threated ceramic materials. *Applied scientific journal Innovation and entrepreneurship*, 4(1), 35-43.

El-Aw, M., Draz, Kh., Eid, Kh., Abo-Shara, H. (2012). Measuring the Morphological Characters of Honey Bee (*Apis Mellifera* L.) Using A Simple Semi-Automatic Technique. *Journal of American Science*, 8(3), 558-564.

Georgiev, G., Georgieva, N. (2014). Investigation possibilities for the use of free software for data processing used for accurate measurement details through photogrammetry. *Applied research on technics, technologies and education*, 2(3), 202-210.

Georgieva, K., Kirilova, E., Georgieva, Ts., Daskalov, P. (2015). Selection of informative colour features complexes from digital images of healthy and diseased vine leaves. *Applied research on technics, technologies and education*, 3(4), 289-295.

Kazlacheva, Z. (2011). Use of the correspondence analysis in fashion design. *Textile and apparel*, 7, 191-196. (in Bulgarian)

Kazlacheva, Z., Ilieva, J., Zhekova, M., Dineva, P. (2014). Fashion design on the base of connection between colours and lines. *Applied research on technics, technologies and education*, 2(1), 54-64.

Klingenberg, C. (2011). MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*, 11, 353-357. DOI: 10.1111/j.1755-0998.2010.02924.x

Lazarov, S. (2016). Application of AutoCAD Program To Measure Chitin Body Parts of Worker Bees (*Apis mellifera L.*). *Journal of Ecology and Environment Sciences*, 15(4), 13-19. (in Bulgarian)

Lazarov, S. (2017). Hygiene behavior of worker bees (*Apis mellifera* L.) and its relationship to basic morphological and biochemical features. *PhD thesis*, Trakia University, Stara Zagora, Bulgaria (in Bulgarian)

Marghitas, A., Paniti-Teleky, O., Dezmirean, D., Margaoan, R., Bojan, C., Coroian, C., Laslo, L., Moise, A. (2008). Morphometric differences between honey bees (*Apis mellifera carpatica*) Populations from Transylvanian area. *Zootehnie Si Biotehnologii*, 41(2), 309-315.

Mattu, V., Verma, L. (1984). Morphometric studies on the indian honeybee, apis cerana indica f. effect of seasonal variations. *Apidologie, Springer Verlag*, 15(1), 63-74.

Meixner, D., Miroslaw, W., Jerzy, W., Fuchs, S., Nikolaus, K. (2007). Apis mellifera mellifera range in Eastern Europe – morphometric variation and determination of its limits. *Apidologie*, 38, 1-7. DOI: 10.1051/apido:2006068

Miladenovic, M., Rados, R., Stanisavljevic, L., Rasic, S. (2011). Morphometric traits of the yellow honeybee (*Apis mellifera carnica*) from Vojvodina (Northern Serbia). *Arch. Biol. Sci*., 63, 251-257. DOI: 10.2298/ABS1101251M

Mladenov, M., Penchev, S., Deyanov, M. (2015). Complex assessment of food products quality using analysis of visual images, spectrophotometric and hyperspectral characteristics. *International Journal of Engineering and Innovative Technology (IJEIT)*, 4(12), 23-32.

Mostajeran, M., Edriss, M., Basiri, M. (2002). Heritabilities and correlations for colony traits and morphological characters in honey bee (*Apis mellifera meda*). *Isfahan university of technology, 17 th world congress on genetic applied to livectocks production*, Agust 19-23 Montpellier, France, session 7

Mostajeran, M., Edriss, M., Basiri, M. (2006). Analysis of colony and morphological characters in honey bees (*Apis mellifera meda*). *Pak. J. Biol.Sci*, 14(9), 2685-2688. DOI: 10.3923/pjbs.2006.2685.2688

Roth, V., Steinhage, V., Schröder, S., Cremers, A. (1999). Pattern recognition combining de-noising and linear discriminant analysis within a real world application. *Proceedings of 8th International Conference on Computer Analysis of Images and Patterns*, Ljubljana, 251-266.

Santana, F., Costa, A., Truzzi, F., Silva, F., Santos, S., Francoy, T., Saraiva, A. (2014). A reference process for automating bee species identification based on wing images and digital image processing. *Ecol. Inform*, 24, 248-260. DOI: 10.1016/j.ecoinf.2013.12.001

Schroder, S., Wittmann, D., Drescher, W., Roth, V., Steinhage, V., Cremers, A. (2002). The new key to bees: automated identification by image analysis of wings. *Kevan P & Imperatriz Fonseca VL (eds) - Pollinating Bees - The Conservation Link Between Agriculture and Nature - Ministry of Environment, Brasília*, 209-216.

Silvaa, F., Sellac, M., Francoyb, T., Costaa, A. (2015). Evaluating classification and feature selection techniques for honeybee subspecies identification using wing images. *Computers and Electronics in Agriculture*, 114, 68-77. DOI: 10.1016/j.compag.2015.03.0120168-1699

Snodgrass, R. (1910). The anatomy of the honeybee. *US department of agriculture*, Washington: Government printing office, Issued May 28, 1910.

Strauss, R., Houck, M. (1994). Identification of Africanized honeybees via non-linear multilayer perceptrons. *Proceedings of the IEEE International Conference on Neural Networks*, 5, 3261-3264

Tofilski, A. (2004). DrawWing – a program for numerical description of insect wings. *Journal of Insect Science*, 2004, 1-5. DOI: 10.1673/031.004.1701

Tofilski, A. (2007). Automatic measurements of honeybee wings. in: MacLeod N. (Ed.), *Automated object identification in systematics: theory, approaches, and applications*. CRC Press, Boca Raton, Florida, 289-298.

Tofilski, A. (2008). Using geometric morphometrics and standard morphometry to discriminate three honeybee subspecies. *Apidologie*, 39, 558-563. DOI: 10.1051/apido:2008037

Tofilski, A. (2016 August 8). DrawWing – a software for analysis of insect wing images and extraction of some information about the wings. Retrieved from http://www.drawwing.org/

Waddington, K. (1989). Implications of variation in worker body size for the honeybee recruitment system. *J. Behav*, 2, 91-103.

Waddington, K., Herbst, L. (1987). Body Size and the Functional Length of the Proboscis of Honey Bees. *Entomologist The Florida*, 70(1), 124-128. DOI: 10.2307/3495099

Zlatev, Z., Nedeva, V. (2017). An algorithm for determination of the morphological characteristics of honey bees. *Journal of central European agriculture*, 18(2), 305-308. DOI: /10.5513/JCEA01/18.2.1902