IntechOpen

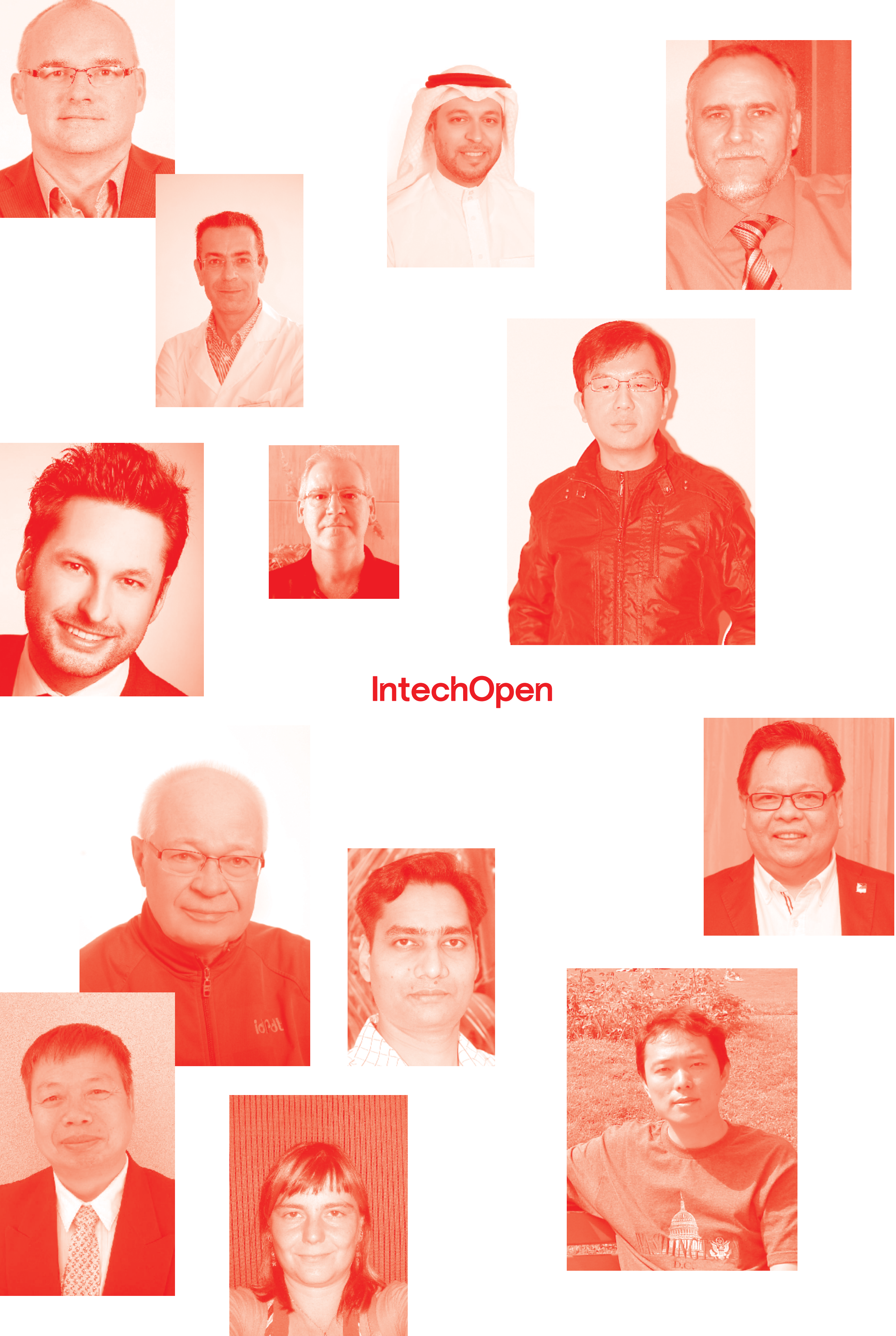
Data Integrity and Quality

Edited by Santhosh Kumar Balan

Data Integrity and Quality

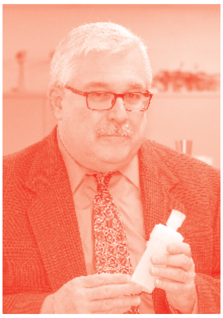
Edited by Santhosh Kumar Balan

Published in London, United Kingdom



IntechOpen

Supporting open minds since 2005



Data Integrity and Quality
http://dx.doi.org/10.5772/intechopen.90975
Edited by Santhosh Kumar Balan

Contributors
Heru Susanto, Vardan Mkrttchian, Yulia Vertakova, Arsen Simonyan, Manuel Doblare, Patricia Carrera-Lasfuentes, Vega Rodrigalvarez-Chamarro, Rafael Del-Hoyo-Alonso, Rocío Aznar-Gimeno, Angel Lanas, Sreekantha Desai Karanam, Raja Vittal Rao Kulkarni, Rajani Sudhir Kamath, Bantwal Hebbal Sinakatte Pai Karthik Pai, Adithya B., Santhi G., Aurora González-Vidal, Tomás Alcañiz, Alfonso Ramallo, Antonio Skarmeta, Santhosh Kumar Balan

Notice
Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2021 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales,
registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Data Integrity and Quality
Edited by Santhosh Kumar Balan
p. cm.
Print ISBN 978-1-83968-798-3
Online ISBN 978-1-83968-799-0
eBook (PDF) ISBN 978-1-83968-800-3

Meet the editor

Dr. B. Santhosh Kumar is currently a professor at Guru Nanak Institute of Technology, Hyderabad. His research interests include data science, machine learning, blockchain technology, and data mining. He has more than 90 publications to his credit, including 49 journal articles and 41 conference papers. He has delivered fifteen guest lectures, webinars, and keynote speeches on various occasions. He received twelve awards from various professional bodies. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and appointed as an ACM Distinguished Speaker. He is also a reviewer for reputed journals including *IEEE Transactions*, *IEEE Access*, *ACM Transactions*, and others.

Contents

Preface **XIII**

Section 1
Introduction **1**

Chapter 1 **3**
Introductory Chapter: Data Integrity and Quality
by Santhosh Kumar Balan

Section 2
Data Integrity **9**

Chapter 2 **11**
Data Integrity Management for Laboratory of the Control of Lifecycle of Domestic Russian Tour Products
by Vardan Mkrttchian, Yulia Vertakova and Arsen Symonyan

Chapter 3 **21**
Analysis and Curation of the Database of a Colo-Rectal Cancer Screening Program
by Rocio Aznar-Gimeno, Patricia Carrera-Lasfuentes, Vega Rodrigalvarez-Chamarro, Rafael del-Hoyo-Alonso, Angel Lanas and Manuel Doblare

Section 3
Data Interaction **41**

Chapter 4 **43**
Big Data Integration Solutions in Organizations: A Domain-Specific Analysis
by Sreekantha Desai Karanam, Rajani Sudhir Kamath, Raja Vittal Rao Kulkarni and Bantwal Hebbal Sinakatte Karthik Pai

Section 4
Data Quality **75**

Chapter 5 **77**
Quality of Information within Internet of Things Data
by Tomás Alcañiz, Aurora González-Vidal, Alfonso P. Ramallo and Antonio F. Skarmeta

Section 5
Applications **93**

Chapter 6 **95**
DNA Computing Using Cryptographic and Steganographic Strategies
by Adithya B. and Santhi G.

Chapter 7 **115**
Revealing Cyber Threat of Smart Mobile Devices within Digital Ecosystem: User Information Security Awareness
by Heru Susanto

Preface

Data is one of the essential resources for an organization to perform well. We are living in an era that is highly data-driven. From decision-making processes to enhancing customer experiences, data is involved in almost all business activities. Out of all the Petabytes and Exabytes of data, it is the responsibility of organizations to get the most benefit out of all the volumes of data residing in humungous databases. This is where data integrity and quality are of utmost importance. Ensuring the integrity and quality of data enriches the insights of the business operations performed. However, confidentiality and safety have been a source of public concern. Quick modifications in technologies such as the Internet and electronic trade along with the implementation of more cultured schemes for gathering, assessing, and making use of private data have made confidentiality a key dispute among the public and government. The domain of data integrity is thus attracting much attention.

To easily differentiate the two terms, one can ask themselves the following questions. Data integrity answers the questions such as: When was the data created? What is its lifetime? Are the entries consistent? Data quality answers questions such as: Is the data relevant? Is the data complete? Is it unique? This quest will help one identify the key differences between the two techniques to be employed to churn the useful information from the large volumes of data.

Section 1 introduces and explains the concepts of data integrity and data quality. Section 2 includes two chapters: "Data Integrity Management for Laboratory of the Control of Lifecycle of Domestic Russian Tour Products" and "Analysis and Curation of the Database of a Colo-Rectal Cancer Screening Program." Section 3 includes one chapter on "Big Data Integration Solutions in Organizations: A Domain-Specific Analysis." Section 4 includes one chapter on "Quality of Information within Internet of Things Data." The final section includes two chapters: "DNA Computing Using Cryptographic and Steganographic Strategies" and "Revealing Cyber Threat of Smart Mobile Devices within Digital Ecosystem: User Information Security Awareness."

In writing this book, I have been fortunate to be assisted by technical experts in many of the subdisciplines that make up the field of data integrity and quality. First and foremost, praises and thanks to God, the almighty, for his showers of blessings throughout my efforts to complete this work. I would like to express my sincere gratitude to the management at Guru Nanak Institute of Technology for providing me with all sorts of support while I completed this book. I am indebted to the principal, vice-principal, and head of the Department of Computer Science & Engineering for their guidance and encouragement. I am also grateful to my colleagues at the GMR Institute of Technology.

I am extremely grateful to my parents for their love, prayers, and sacrifices and for educating and preparing me for my future. I am very much thankful to my wife J. Nandhini, my son B. S. Haarish Athithiya, and my daughter B.S Josita Varshini for their love, understanding, prayers, and continual support as I worked on this book. I also give special thanks to my friends who inspire me, and finally, I wish to acknowledge the staff at IntechOpen, especially Author Service Manager Maja Bozicevic, without their continuous support and valuable suggestions we would not be able to finish this book project.

Dr. Santhosh Kumar Balan
Professor,
Department of Computer Science and Engineering,
Guru Nanak Institute of Technology,
Hyderabad, Telangana, India

Section 1

Introduction

Chapter 1

Introductory Chapter: Data Integrity and Quality

Santhosh Kumar Balan

1. Introduction

The significance of data could be employed to escalate the income incision expenses or both. The data integrity related software is one of the numbers of an assessment tool for evaluating the information. It permits the users to assess the confidentiality of the information is rising regularly. Due to these diverse analyses are performed on confidentiality safeguarded data integrity the design of fresh scheme permits the mining data while attempting to safeguard the confidentiality of the users. Many of these schemes are intended for user's confidentiality but still, others are intended on the confidentiality of the organization. It is broadly termed as data exploration in repositories which is the significant mining of hidden, conventionally indefinite and probably needful data from the information in the repositories. The data exploration is required to make logic and usage of information. Even though the data exploration in repositories are regularly regarded as replacements, and it is crucially a segment of the data exploration process. Normally the data integrity, for instance, the information or the data exploration is the process of assessing information from diverse viewpoints and abstracting them into needful data from diverse angles, classification and abstraction of recognized association.

Precisely it is the process of locating synchronization or prototypes among dozens of domains in immense relational repositories. Though, it is relatively fresh domain the technology still remains the same. The organizations are making use of potent computers for straining through immense of supermarket scanner information and assess market trends over years. Therefore, the constant improvements in processing the computational power, disk storages and arithmetic software are significantly escalating the precision of the assessment while governing the expenses. The data integrity of fresh and potential prototypes in immense information sets is a domain in the ignition. The major feature is to enhance the safety for instance identification of interferences. The subsequent feature is the possible safety risks imposed on the opponent has its abilities. The confidentiality related risks have gained the care of multimedia, legislators, government firms, trade and confidential promoters. The data integration edges resume to develop, there are diverse prevailing issues examined. The assembly might choose to contemplate in relation to the planning and inaccuracy. These problems comprise but are not restricted to the quality of the information, interoperability and stealing of assignment and confidentiality. Along with the added features, the technology-based abilities are crucial where other features might also govern the victory of the results [1].

2. Data quality

The Data Quality which we are maintaining in the data integrity comprises the usage of cultured information assessment tools to explore conventionally unfamiliar, lawful prototypes and associations in immense information sets. These tools could comprise arithmetic prototypes, numerical routines and machine learning schemes. Therefore, it comprises gathering, arranging and preserving information which comprises assessment and forecast. It could be accomplished on information denoted in measurable, text-based, visual, image or hypermedia patterns. The applications could make use of selective metrics to assess the information. They comprise relationship orders or route assessment, categorization, grouping and estimation. Diverse firms gathers and perfectly voluminous extents of information. The schemes could be used quickly on the conventional software and hardware platforms for improving the values of the prevailing resources and could be combined with the fresh products and systems due to their availability online. The repositories and information repositories are becoming more and more attractive and make use of the immense volume of information which requires being assessed efficiently. The data exploration in repositories could be entailed as the exploration of attractive, hidden and conventionally unfamiliar data from the immense repositories.

The data integrity repositories might be reasonable instead than a physical subgroup of the information depository offered that the information depository Database Management Systems which could aid the supplementary supply requirements of information mining. If it is possible, it is better to leave a distinct it's repository. In general usage, the terms data integrity and data quality are used interchangeably. However, they often have few significant differences between each other. Data integrity validates that the data and ensures that it remains unaltered throughout its life cycle. Numerous operations such as storing, retrieving, updating, etc., are performed very often on data. The techniques ensure that, irrespective of all the operations performed, the data is maintained, just as how it was inputted. The data encryption, backup, access controls, validation are few practices that maintain data integrity. On the other hand, data is labeled as quality data if it is relevant and complete and is suitable to the intended purpose. As per the standards, data quality is defined in three different perspectives such as from a consumer's perspective, from a business perspective, and from a standard-based perspective.

The Data quality is a multi-layered problem which symbolizes the immense disputes in data integrity. The quality of information states the precision and fullness of the information. The quality of the information could also be bothered based on the framework and reliability of the information which is being assessed. The existence of redundant reports, the missing information policies, the properness of revisions and human faults could crucially influence the efficiency of more intricate it's schemes which are delicate to the elusive variations which might prevail over the information. In order to enhance the quality of information, it is roughly mandatory to refine the information which comprises the eradication of redundant reports, standardizing the values employed to symbolize data in the repository [2].

3. Confidentiality safeguarded data integrity

In relation to the quality of the information, the problem is the synchronization of various repositories and its software. The synchronization denotes the capability of a computational system or information to work with other systems or the information employing usual principles or operations. Synchronization is a crucial segment of the immense determination to enhance the linked association and data distribution using e-government and native confidentiality edges. For Data Integrity, the synchronization of repositories and software is crucial to allow the exploration and assessment of diverse repositories consequently and aids in assuring the comforts of its actions of diverse firms. It attempts to partake the benefits of the prevailing inherited repositories or that are opening the initially shared attempts with other firms or extents of government might practice synchronization issues. Likewise, as the firms progress onward with the generation of fresh repositories and data distribution attempts, they will require resolving the synchronization problems during their phases of implementation to better assure the efficiency of their schemes [3–6].

The Data Integrity has influenced crucial attention, especially over the past years with its immense varieties of applications. In terms of safety concerns, it is considered advantageous in challenging diverse sorts of risks to the computational system. Therefore, the similar methodologies could be employed to generate probable risks related to safety. Moreover, the collection of data and assessment attempted by the government firms and trade elevates the anxieties related to confidentiality which inspires the confidentiality safeguarding in data integrity. The feature of confidentiality safeguarding is that it shall be capable to make use of various schemes without monitoring the values of private information. But still, the disputes are being explored. An additional feature is that the use of its schemes, the opponent can gain access private data which cannot be attained using request tools risking the confidentiality of peoples. Diverse preliminary analyses are available in confidentiality safeguarded Data Integrity. Conversely, there are several problems which require more analyses in the conception of data integrity from both confidentiality and safety initiatives [7, 8].

4. Applications

The analytical illustration offers trade buying system greatest of the products from the preceding year one could forecast the level of products which requires goods for the impending periods. The authentication could verify on the ailments such as viral with the exception that it is probable to locate the acknowledgment and withdrawal identification in terms of scams. It is employed for diverse objectives in both the private and public firms. The organization like banking, insurance, medicals and purchasing normally make use of data integrity to minimize the expenses, improve analysis and escalates trades. Consider the insurance and banking organization employing data integrity applications for identifying scams and aid in threat evaluation. The usage of user-related information gathered over the present periods the firms could design prototypes which forecasts the threats prevailing to the users in terms of credits or regarding the privileges during accident might be false and shall be inspected more carefully.

The medical society roughly makes use of data integrity to aid the analysis of the efficiency of the scheme or medicines. The medical firms make use of data integrity of the chemical substances and genetic components to aid the governance of studies on fresh management for ailments. The vendors could employ the data gathered using attraction programs to evaluate the efficiency of choosing the items and position related choices, voucher offers and the frequency of items bought regularly. The firms like telephone service suppliers and music clubs could make use to generate a segment assessment to examine which users are probable to continue as users and which ones are probably to migrate to the opponent [9, 10].

Author details

Santhosh Kumar Balan
Department of Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India

*Address all correspondence to: bsanthosh.csegnit@gniindia.org

References

[1] Tina Hui, Ensuring Data Quality and Integrity in Financial Management Reporting for Medical Imaging Operations, Journal of Medical Imaging and Radiation Sciences, Volume 50, Issue 3, Supplement, 2019, Page S13, ISSN 1939-8654, doi:10.1016/j.jmir.2019.06.034.

[2] T. Hongxun et al., "Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory, " 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 248-252, doi:10.1109/ICCCBDA.2018.8386521.

[3] I. Taleb, M. A. Serhani and R. Dssouli, "Big Data Quality: A Survey," 2018 IEEE International Congress on Big Data (BigData Congress), 2018, pp. 166-173, doi: 10.1109/BigDataCongress.2018.00029.

[4] MI Svanks, Integrity analysis: Methods for automating data quality assurance, Information and Software Technology, Volume 30, Issue 10,1988, Pages 595-605, ISSN 0950-5849, https://doi.org/10.1016/0950-5849(88)90116-4.

[5] Alan R. Simon, Steven L. Shaffer, Chapter 9 - Data Quality and Integrity Issues, Editor(s): In the Morgan Kaufmann Series in Data Management Systems, Data Warehousing And Business Intelligence For e-Commerce, Morgan Kaufmann, 2002, Pages 193-208, ISBN 9781558607132, https://doi.org/10.1016/B978-155860713-2/50012-7.

[6] Nikita R. Nikam, Priyanka R. Patil, et. al., "Data Integrity: An Overview", International Journal of Recent Scientific Research, Vol. 11, Issue, 06 (A), pp. 38762-38767, June, 2020

[7] World Health Organization, working document QAS/19.819, Guideline on Data Integrity, October 2019, 14-16.

[8] Boritz, J. IS practitioners views on Core concepts of information integrity, Int. J. Account. Inf. Syst, Elsevier.

[9] https://www.veracode.com/blog/2012/05/what-is-dataintegrity.

[10] https://digitalguardian.com/blog/what-data-integrity-dataprotection-101

Section 2

Data Integrity

Chapter 2

Data Integrity Management for Laboratory of the Control of Lifecycle of Domestic Russian Tour Products

Vardan Mkrttchian, Yulia Vertakova and Arsen Symonyan

Abstract

A primary responsibility of domestic tour products is to provide safe and efficacious products of appropriate quality to consumers by assuring decisions are based on accurate, reliable, truthful, and complete data. This task in this chapter is solved on the basis of the authors' works on Avatar-Based Intellectual Managing for Innovation Technologies Transfer in the tourism industry of the Republic of Armenia, where citizens of the Russian Federation travel with their internal passports, that is, they use it as an internal tour product. The chapter describes the entire algorithm of the system, shows the results of the study, which guarantee Data Integrity Management for Laboratory of the Control of Lifecycle of domestic Russian tour products.

Keywords: Data integrity, management system, laboratory of control, lifecycle, domestic tour products, Industrial Internet of Things

1. Introduction

The COVID-19 lockdown has led to the closure of tour industry throughout the world an estimated 2.6 Billion people across the world are affected due to the same; nearly 186 countries around the world have stopped working due to this global pandemic. The rest argue that information technology will help tour industry and will eventually be a part of the regular living. However the calendar was disrupted and there was a need to stay on the touristic domestic service not only for people but also touristic industry. Now the best way to do so was to make use of online platforms, online to promote tour products during this pandemic. A primary responsibility of domestic tour products is to provide safe and efficacious products of appropriate quality to consumers by assuring decisions are based on accurate, reliable, truthful, and complete data. This task in this chapter is solved on the basis of the authors' works on Avatar-Based Intellectual Managing for Innovation Technologies Transfer in the tourism industry of the Republic of Armenia [1], where citizens of the Russian Federation travel with their internal passports, that is, they use it as an internal tour product.

2. The data integrity management for laboratory of the control of lifecycle of domestic Russian tour products

A world-class laboratory has been created at Sochi State University, which involves the use of the original Mkrttchian's technology - multi-agent models of intelligent digital twin-avatars [2, 3]. The subject of the laboratory is fully consistent with the section of the priorities of the scientific and technological development of the Russian Federation, specified in paragraph 1.1 of this document: "the transition to advanced digital, intelligent production technologies, robotic systems, new materials and design methods, intelligence ". The scientific profile of the laboratory is aimed at "the transition to advanced digital, intelligent production technologies" in such an important area for Russia as tourism. In the course of scientific and scientific-practical activities, "the creation of systems for processing large amounts of data, machine learning and artificial intelligence" will be carried out, as provided for by the Priorities of Scientific and Technological Development of the Russian Federation. This study is fits into a moment of operational uncertainty and theoretical redevelopment of the nature of tourism in a society marked by geopolitical turmoil and declining international security, as well as rapid changes at the global level, including the pandemic (COVID-19), which is currently posing new challenges for the sector. Today, it is more relevant and appropriate than ever to reflect on them, with the new, digital energy of blockchain technology, using a fundamental approach to digitalizing the decentralized lifecycle management of the domestic Russian tour product with problem-oriented digital twin avatars, supply chain, volumetric hybrid and federated-consistent blockchain. The goal of the project is theoretical study and practical implementation, in the form of basic models and software modules, artificial intelligence algorithms in managing the life cycle of an internal Russian tour product. Why at the State Sochi University, using the scientific potential of the head and responsible executors of the project, the Laboratory for digitalization and management of tour products, using multi-agent models of intelligent digital twins-avatars, is being created, the purpose of these studies is to solve a scientific problem in terms of creating an integrated scientific and methodological approach to modeling and design of monitoring systems, diagnostics and management of distributed cyber-physical objects and processes in the network segments of the Industrial Internet of Things based on the convergence of engineering technologies, data mining and in-depth analysis of processes, predictive modeling and machine learning. The objectives of the research are related to the development of new models, methods and a set of tools for digital transformation of monitoring, diagnostics and management of distributed cyber-physical objects during the transition to the digital economy within the framework of the fourth industrial revolution (Industry 4.0). The results of design research are needed to synthesize the architecture of a new generation of intelligent cyber-physical systems, which represents a multi-agent computing ecosystem. It is designed to provide decision support processes based on monitoring events and processes at distributed cyber-physical objects of the Russian tourism industry. In such systems, there are many cyber-physical objects that receive a huge amount of sensory data that cannot be processed by humans in real time. Currently, there are no ready-made integrated solutions for modeling and designing distributed monitoring and control systems for cyber-physical objects. Despite advances in engineering and knowledge management, the use of this approach for the synthesis of cyber-physical monitoring and control systems is still poorly developed. Such systems work with a variety of distributed cyber-physical objects, which are, in most cases, measuring devices with sensors that collect and accumulate sensor data for transmission to a processing center via a telecommunications network.

The results of data analysis are used for predictive modeling of the dynamics of the development of processes at cyber-physical objects and for making management decisions. Cyber-physical monitoring and control systems are needed to automate the decision-making process based on data mining. The relevance of the project is associated with the need to develop and develop new universal mechanisms for modeling and designing cyber-physical systems using new control technologies and in-depth analysis of processes at controlled objects of the Russian tour product. For in-depth analysis of processes, it is necessary to develop automated technologies for collecting, storing and intelligent analysis of data obtained from controlled cyber-physical objects of the Russian tour product. The scientific novelty of design research consists in the creation of a new scientific and methodological approach to the modeling and design of cyber-physical systems for monitoring and controlling distributed objects and processes in the network segments of the Industrial Internet of Things, as well as the methodology for distributed monitoring, diagnostics and recovery of these systems during their operation. Scientific and practical significance lies in the creation of new technologies and software and tools for the synthesis of cyber-physical systems for monitoring and controlling distributed objects and processes on the Internet of Things. For in-depth analysis of processes, it is necessary to develop automated technologies for collecting, storing and intelligent analysis of data obtained from controlled cyber-physical objects of the Russian tour product. The scientific novelty of design research consists in the creation of a new scientific and methodological approach to the modeling and design of cyber-physical systems for monitoring and controlling distributed objects and processes in the network segments of the Industrial Internet of Things, as well as the methodology for distributed monitoring, diagnostics and recovery of these systems during their operation. Scientific and practical significance lies in the creation of new technologies and software and tools for the synthesis of cyber-physical systems for monitoring and controlling distributed objects and processes on the Internet of Things. A new generation cyber-physical monitoring and control system is implemented in the form of a hyper-converged component-based architecture of a reconfigurable ecosystem, which performs the functions of multi-agent processing of large amounts of sensor data in a computing grid of sensor node controllers based on a fog (edge) computing model. The main scientific problem solved in the research process is associated with the synthesis of a new approach to modeling and designing cyber-physical systems for monitoring, diagnostics and control of distributed objects and processes in the network segments of the Industrial Internet of Things. Optimization of management is one of the central tasks facing the Russian economy. Currently, there is a gradual transition in control systems from simple automation to technologies of “smart” or “smart”, and the concept of “digital twin” is central to the development of the corresponding systems. The existing experimental systems have a number of obvious bottlenecks - cyber vulnerability, fragmentation, binding to a specific tour product, etc. The use of intelligent avatar technology for the development of twins can eliminate bottlenecks, which is detailed in three fundamental monographs of the project manager. As a result of the implementation of the proposed scientific research, new scientific, scientific, technical and technological digital solutions will be created that will provide an innovative and digital transformation of product tour management, as well as the development of a typical multi-agent system of intelligent avatars for effective management. The expected results correspond to the world level of scientific research in this area, as they relate to the development of new approaches to monitoring and control of complex geographically distributed cyber-physical systems, which are the basis for the implementation of intelligent cyber-physical systems of a new generation. The results of the project are components for the creation and implementation of new technologies and systems within the framework of the fourth industrial revolution, the transition to a digital economy, digital transformation of management processes and decision support. The public and social significance of the project is determined by the fact that the results of the project are intended for the implementation and development of new intelligent cyber-physical systems for managing the tour product of the Russian Federation. The main scientific problem solved in the research process is related to the synthesis of a new approach to modeling and designing cyber-physical systems for monitoring, diagnostics and management of distributed objects and processes of creating and implementing an internal tour product in the network segments of the industrial Internet of Things. The expected results correspond to the world level of scientific research in this area, as they relate to the development of new approaches to monitoring and managing complex geographically distributed cyber-physical systems, which are the basis for the implementation of intelligent cyber-physical systems of a new generation. During the implementation of the project, a new multi-agent approach will be developed. Modeling and design of modern cyber-physical systems in the industrial Internet of Things. The results of the project are components for the creation and implementation of new technologies and systems within the framework of the fourth industrial revolution, the transition to a digital economy, digital transformation of management processes and decision support. The results can be used to synthesize new intelligent monitoring systems, which prove the practical significance of design research, as well as the versatility of the developed models, methods and technologies, which in turn will allow the use of tools for creating various geographically distributed cyber-physical systems for creating and implementing an internal tour product. In the process of preparing this article, a new technology for "Data Integrity and Quality" was developed. The main area of research is the synthesis and development of proactive monitoring technologies for managing the risks of events in geographically distributed systems of internal tourism products, the following specific results:

1. The concept is a methodology of proactive monitoring of events in geographically distributed systems of the environment based on the collection, consolidation and predictive analysis of big data on normal and emergency situations in order to identify possible causes and factors of influence for predictive modeling and risk assessment of the occurrence and development of negative events, which includes: - methodology for collecting and processing big data from distributed sensors, measuring devices.

2. The method of consolidation of large heterogeneous data on critical events, accidents and emergency situations with different time and geospatial labels, including elimination of duplicates, validation, information noise cleaning, formalization in the form of vector and graph models, classification and clustering in the space of signs and influence factors, collection of statistics, retrospective analysis to establish correlations with similar incidents in the past (event patterns).

3. Model and method of ensuring information security for the protection of large sensor data in distributed information storage on sensor and mobile data collection nodes and in communication channels of the telecommunications environment in the process of collection, transmission and storage based on distributed ledger technologies (blockchain).

4. Methods of presenting information about events in the form of time series of event characteristics and time series of dynamics of possible factors for assessing the risks of occurrence and development of negative events, determining correlations with a number of factors and patterns of their influence.

5. The method of comparative analysis (benchmarking) of time series of event characteristics with time series of influencing factors to determine possible correlations between them, selection and assessment of the sensitivity and the degree of influence of factors on the occurrence of accidents and emergency situations.

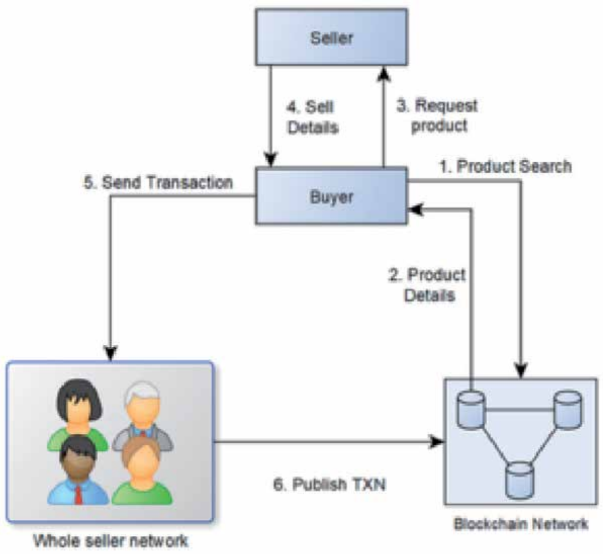


Figure 1.
Diagram on problem-oriented digital twins-avatars, supply chain, 3D-hybrid, federated & coordinated blockchain and domestic tour product whole seller monitoring for data integrity management.

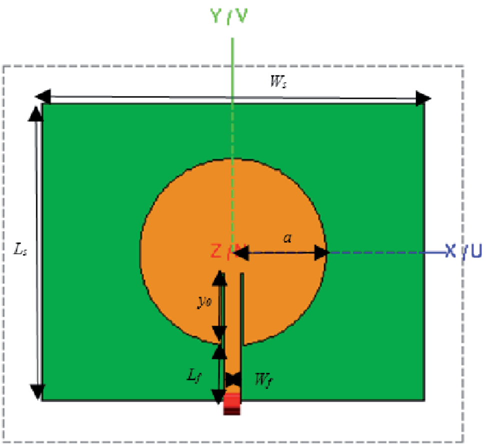


Figure 2.
One geometry of the designed basic antenna for monitoring data integrity management.

6. Predictive model for predictive assessment of the risks of occurrence and development of similar critical events, accidents and emergency situations.

7. A method of visualizing the results of proactive monitoring with geospatial and temporal reference on mobile communications.

The main result of design research is the creation and development of an approach to proactive volume monitoring of the domestic tour products using

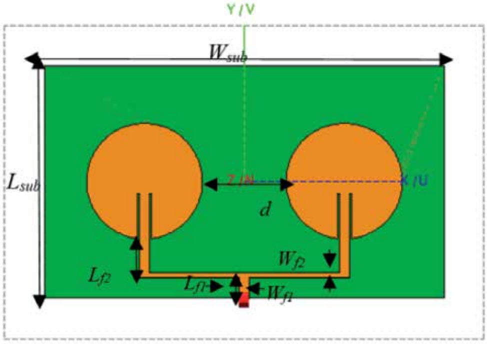


Figure 3.
Two geometry of the designed basic antenna for monitoring data integrity management.

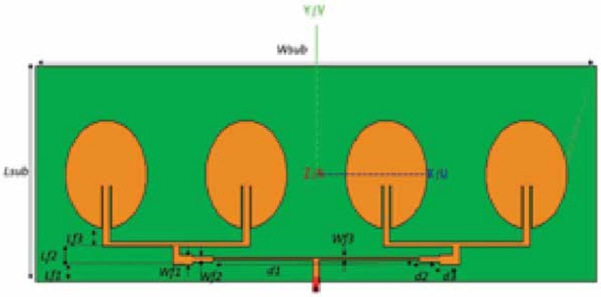


Figure 4.
Four geometry of the designed basic antenna for monitoring data integrity management.

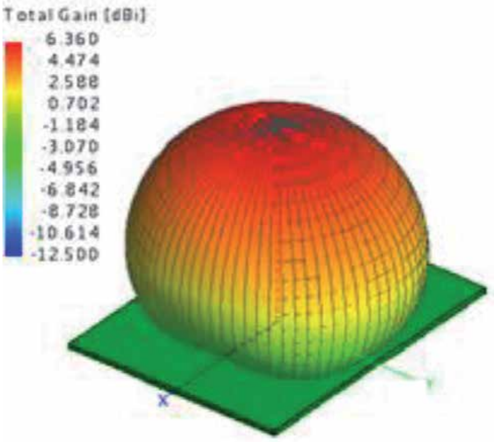


Figure 5.
3D normal blockchain for monitoring data integrity management.

Blockchain Network- with problem-oriented digital twins-avatars, supply chain, 3D-Hybrid, federated & coordinated blockchain [1–3] (**Figure 1**) and the developed antenna array [2, 4–16] (**Figures 2**–**4**) and Blockchain was development from 3D normal, 3D hybrid, and 3D hybrid, federated & coordinated (**Figures 5**–**7**) [2].

This chapter presents the design reader array antenna operates, for correspond to IoT applications for monitoring domestic tour product. The antenna was performed using federative blockchain. The configuration presents a good performance in terms of gain and bandwidth compared with the designed single and array antenna. The increase in the number of radiating elements improves the antenna performances especially the gain. This structure will be a good solution for IoT

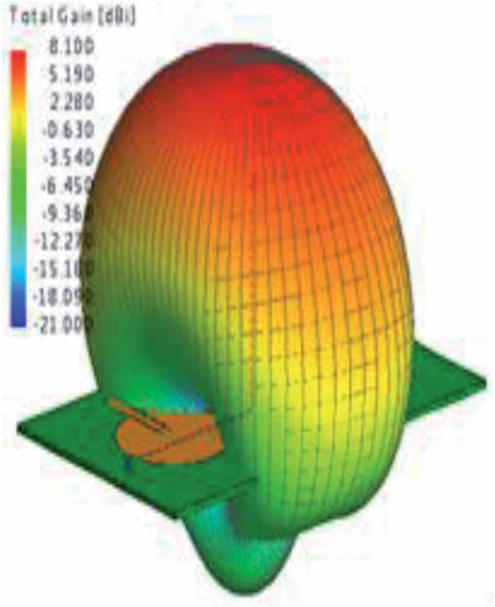


Figure 6.
3D hybrids blockchain for monitoring data integrity management.

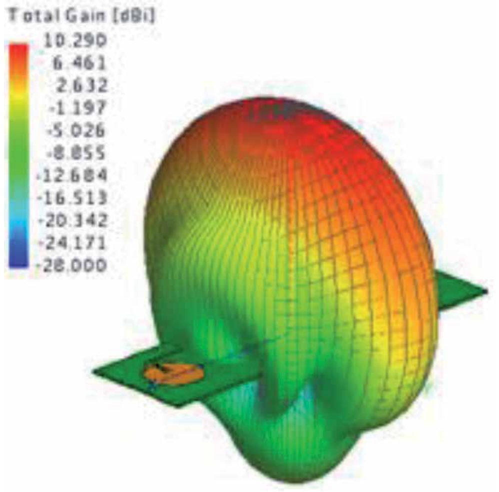


Figure 7.
3D hybrids, federated and coordinated blockchain for monitoring data integrity management.

applications. As perspective of this work, fabrication and measurement should be done to confirm the simulated results.

3. Conclusion

The chapter describes the entire algorithm of the system, shows the results of the study, which guarantee Data Integrity Management for Laboratory of the Control of Lifecycle of domestic Russian tour products. As a result of the research, key technologies of the of "Industry 4.0" era were identified, their characteristics and role in use were given. Conclusions are made that the introduction of these technologies will favorably affect productivity, revenue growth, employment and investment. In the conclusion the detailed description of various areas of using the Internet of things in activity of the domestic tourist organizations is resulted. The study allows us to conclude that the digitalization of the touristic sector will entail the release of better products. In addition, Industry 4.0 will lead to the creation of more flexible systems, the participants of which will exchange information via the Internet, which in turn will significantly increase labor efficiency and reduce costs in domestic tour production processes and data Integrity management.

Author details

Vardan Mkrttchian[1]*, Yulia Vertakova[2] and Arsen Symonyan[3]

1 HHH University, Sydney, Australia

2 Southwest State University, Kursk, Russia

3 Sochi State University, Sochi, Russia

*Address all correspondence to: hhhuniversity@gmail.com

References

[1] Mkrttchian V, Chernyshenko S, Ivanov M, Avatar-Based Intellectual Managing for Innovation Technologies Transfer in Nationals Entrepreneurships of Armenia. In: Mehdi Khosrow-Pour. Encyclopedia of Organizational Knowledge, Administration, and Technology (5 Volumes): IGI Global, USA; 2020. p. 1468-1479. DOI: 10.4018/978-1-7998-3473-1.ch101

[2] Chernyshenko V., Vertakova Y., Mkrttchian V. Development and Implementation of Adaptive Trade Policy in the Era of Digital Globalization Based on Virtual Exchange of Intellectual Knowledge// Avatar-Based Models, Tools, and Innovation in the Digital Economy. Chapter 8, – Hershey, USA: IGI Global, P. 131-140 (2020).

[3] Mkrttchian V., Chernyshenko S. *Digital Intelligent Design of Avatar-Based Control with Application to Human Capital Management*//International Journal of Human Capital and Information Technology Professionals. - Volume 12, Issue 1, P. 19-32 (2021)

[4] Alami, A.El, Ghazaoui, Y., Das, S., Bennani, S. D., & Ghzaoui, M. E. (2019). Design and Simulation of RFID Array Antenna 2x1 for Detection System of Objects or Living Things in Motion. Procedia Computer Science, *151*, 1010-1015. https://doi.org/10.1016/j.procs.2019.04.142

[5] Alami, Ali El, Bennani, S. D., Bekkali, M. E., & Benbassou, A. (2005). DESIGN, ANALYSIS AND OPTIMIZATION OF A NEW STRUCTURE OF MICROSTRIP PATCH ANTENNA FOR RFID APPLICATIONS. Vol., *63*, 7.

[6] Balanis, C. A. (2005). *Antenna theory: Analysis and design* (3rd ed). John Wiley. Chen, Z. N., & Qing, X. (2010). Antennas for RFID applications. *2010* International Workshop on Antenna Technology (IWAT), 1-4. https://doi.org/10.1109/IWAT.2010.5464865

[7] Dong, Y., Choi, J., & Itoh, T. (2017). Folded Strip/Slot Antenna with Extended Bandwidth for WLAN Application. IEEE Antennas and Wireless Propagation Letters, *16*, 673-676. https://doi.org/10.1109/LAWP.2016.2598276

[8] Fahmy, A., Altaf, H., Al Nabulsi, A., Al-Ali, A., & Aburukba, R. (2019). Role of RFID Technology in Smart City Applications. *2019 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 1-6. https://doi.org/10.1109/ICCSPA.2019.8713622

[9] Giay, Y., & Alam, B. R. (2018). Design and Analysis 2.4 GHz Microstrip Patch Antenna Array for IoT Applications using Feeding Method. *2018 International Symposium on Electronics and Smart Devices (ISESD)*, 1-3. https://doi.org/10.1109/ISESD.2018.8605455

[10] Ikram, T., Najiba, E. A. E. I., Jorio, M., & Slimani, A. (2017). A high gain 1*2 array RFID reader MPA for indoor localization applications. *2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1-6. https://doi.org/10.1109/WITS.2017.7934639

[11] Katoch, S., Jotwani, H., Pani, S., & Rajawat, A. (2015). A compact dual band antenna for IOT applications. *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 1594-1597. https://doi.org/10.1109/ICGCIoT.2015.7380721

[12] Khardioui, M., Bamou, A., El Ouadghiri, M. D., & Aghoutane, B. (2020). Implementation and Evaluation of an Intrusion Detection

System for IoT : Against Routing Attacks. In M. Ezziyyani (Éd.), *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)* (Vol. 92, p. 155-166). Springer International Publishing. https://doi.org/10.1007/978-3-030-33103-0_16

[13] Nate, K. A., Hester, J., Isakov, M., Bahr, R., & Tentzeris, M. M. (2015). A fully printed multilayer aperture-coupled patch antenna using hybrid 3D/inkjet additive manufacturing technique. *2015* European Microwave Conference (EuMC), 610-613. https://doi.org/10.1109/EuMC.2015.7345837

[14] Nate, K., & Tentzeris, M. M. (2015). A novel 3-D printed loop antenna using flexible NinjaFlex material for wearable and IoT applications. *2015 IEEE 24th Electrical Performance of Electronic Packaging and Systems (EPEPS)*, 171-174. https://doi.org/10.1109/EPEPS.2015.7347155

[15] Ouazzani, O., Bennani, S. D., & Jorio, M. (2017). Design and simulation of 2*1 and 4*1 array antenna for detection system of objects or living things in motion. *2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1-5. https://doi.org/10.1109/WITS.2017.7934640

[16] Varum, T., Duarte, M., Matos, J. N., & Pinho, P. (2018). Microstrip Antenna for IoT/WLAN applications in Smart Homes at 17GHz. *12th European Conference on Antennas and Propagation (EuCAP 2018)*, 116 (4 pp.)-116 (4 pp.). https://doi.org/10.1049/cp.2018.0475

Chapter 3

Analysis and Curation of the Database of a Colo-Rectal Cancer Screening Program

Rocio Aznar-Gimeno, Patricia Carrera-Lasfuentes, Vega Rodrigalvarez-Chamarro, Rafael del-Hoyo-Alonso, Angel Lanas and Manuel Doblare

Abstract

Data collection in health programs databases is prone to errors that might hinder its use to identify risk indicators and to support optimal decision making in health services. This is the case, in colo-rectal cancer (CRC) screening programs, when trying to optimize the cut-off point to select the patients who will undergo a colonoscopy, especially when having insufficient offer of colonoscopies or temporary excessive demand. It is necessary therefore to establish "good practice" guidelines for data collection, management and analysis. With the aim of improving the redesign of a regional CRC screening program platform, we performed an exhaustive analysis of the data collected, proposing a set of recommendations for its correct maintenance. We also carried out the curation of the available data in order to finally have a clean source of information that would allow proper future analyses. We present here the result of such study, showing the importance of the design of the database and of the user interface to avoid redundancies keeping consistency and checking known correlations, with the final aim of providing quality data that permit to take correct decisions.

Keywords: colo-rectal cancer screening program, health data, data analysis and curation, data coherence, data integrity

1. Introduction

Big Data and Artificial Intelligence are revolutionizing medicine, although they require large amounts of data [1]. Healthcare is an information-intensive activity that produces large quantities of structured (laboratory data) and non-structured (images, texts, etc.) data, from laboratories, wards, operating theaters, primary care organizations. Also, the amount of these data will surely highly increase in the near future due to the interconnection of medical devices via the Internet of Things [2].

Electronic Health Record Databases (EHRD) quality and interoperability [3] is one hot topic in Health Data Science. However, the data captured in EHRD is not available just from well-designed and maintained databases controlled by an administrator and curated by an IT Department, but, contrarily, it is composed of non-unified, redundant, and often replicated information that come from numerous independent e-health service providers (hospital, primary services, regional governments). Therefore, to assure the quality of the data, new processes are required from origin to final service generation through an appropriate data governance.

We can expect that new technologies such as blockchain will reduce this problem, introducing data interoperability, and security [4] and by the adoption of international standards for EHRD (Reference Model, ISO/DIS 13606–2, OpenEHR) [5–8]. However, we are just in the time when leveraging the power of health data to improve clinical or administrative decisions still requires an important effort to ensure the requested data quality. In this regard, many research studies discuss different approaches to improve quality and curation [9] and to deploy new advanced services [2].

There have been several works addressing this problem of data analysis and curation of health information systems [10–13]. This chapter focuses on data quality analysis of electronic medical records and, in particular, of the database of the colorectal cancer screening programme of the Spanish region of Aragón.

The colo-rectal cancer (CRC) screening program of Aragón started in 2014 and, as many other similar programs in the world, is based on the result of a fecal immunohisto-chemical test (FIT) and is focused on medium risk population (ages between 50 and 69 years, without family history of CRC and without the presence of colon diseases, colectomy, or irreversible terminal diseases such as Alzheimer's). The result of this test determines whether it is necessary to perform a colonoscopy (positive cases) or the patient will be screened again after a predefined period. The objective of the program is to diagnose the colo-rectal cancer in its early stage and/or to remove precancer polyps before they may evolve to potential malignant tumors.

One of the difficulties/limitations that this type of programs usually encounters is the insufficient offer of colonoscopies, or the excessive demand derived from positive FIT cases in the invited population. It is necessary therefore to analyze the historical information to define a set of data-based risk indicators than can support the decision-making process in public health services, trying to set the least harmful criteria for selecting the patients who will finally undergo the colonoscopy. It is then clear the importance of the quality of the information stored concerning the clinical data of the patients participating in the program as well as the information of the tests carried out, the results of the colonoscopies and the associated pathological data.

Data collection, if performed by humans, is prone to filling errors. These potential errors can be reduced by a proper design of the database and of the user interface, avoiding redundancies, keeping consistency and checking known correlations. Therefore, it is necessary to establish "good practice" guidelines for data collection and management. With the aim of improving the redesign of the current platform, we carried out an exhaustive analysis of the data collected in the Aragón's regional colo-rectal screening program from 2014 to 2018. This analysis revealed considerable data noise, so we proposed a list of recommendations to improve their quality. We also carried out a curation process of the available data in order to have a clean source of information that would allow proper future analyses.

The recommendations arose from the identification of a series of assumptions and restrictions that the platform should contain to comply with the integrity, coherence and consistency of the data and, therefore, to mitigate the noise. They covered from the default value of each variable, its range, its mandatory character and the redundancy control, to other types of suggestions that include possible constraints on their values, relationships between variables, creation of new variables that may facilitate the analysis and possible warnings or alerts that could help the user to perform a correct data filling.

The chapter is organized as follows. In the first section, the database analyzed is described. Section 2 introduces some basic principles with respect to data integrity, consistency and coherence that any data manager must adhere to in order to ensure the data quality and presents some examples of the data analysis undertaken and a number of recommendations with the aim of complying with these principles. Decisions taken retrospectively are then introduced into the data healing process in order to obtain a clean source of information from which to draw knowledge for further analyses. Finally, the last section includes the conclusions of the entire study.

2. Description of the database

All the information of the CRC screening program in the region of Aragón (Spain) is stored in a centralized database that is fed from other external databases and from the personal information of the patient that is filled by hospital staff through a user-interface (UI) tool. This UI is a web application on which the patient information is displayed and managed. The information that contains comes from the different public hospitals in Aragón: San Jorge Hospital, Barbastro Hospital, Miguel Servet University Hospital, Lozano Blesa University Clinical Hospital, Ernest Lluch Hospital, Obispo Polanco Hospital of Teruel and Alcaniz Hospital. Therefore, the staff who use the platform have different roles, belong to different hospitals and have different degree of training. This means that the application must be as intuitive as possible, as well as to comply with sufficient checks to handle the data relationships correctly, reducing possible errors as much as possible during the data collection. This translates the problem to the good design of the database.

Furthermore, it is quite common in public institutions to have several contracts with different private companies in a short period of time. This usually implies waste of time in understanding, adapting and changing the structure to the way of working of each of company. Therefore, it makes even more sense to establish a good database design and architecture that allows its correct growth and maintenance.

In particular, this section explains the characteristics of the database of the colo-rectal cancer screening program of Aragón between 2014, when the program started, and 2018. In the following sections, the inconsistences found in its design and, therefore, in the data quality are exposed and a series of recommendations (and “good practices”) are proposed to comply with data integrity, coherence and consistency as much as possible.

The existing database model is here explained in inverse order to the actual development of the database. First the final result is explained (relational model) and then the underlying model (entity-relationship model) is discussed [14].

2.1 Relational model

The database tables analyzed contain the following data information which is extracted from different sources of information:

- Patient: Basic demographic information on target patients. This information is extracted from the User Database (BDU) of the corresponding health area.

- Exclusion: Information on temporary or permanent exclusions to the program, which are similar to exclusion criteria of other CRC screening programs in Spain. Exclusions may be due to family history of

CRC, presence of colonic disease, colectomy, irreversible disease (e.g., Alzheimer), previous negative FIT result, or previous negative colonoscopy outcome. This information is automatically dumped into the table from different health system databases, such as OMI-AP (clinical information of patients attended in primary care), CMBDH (clinical information of patients attended in hospital), HP-His (clinical information of ambulatory patients) and BDU (User Database).

- Correspondence: Information from the letters sent to patients along their stay in the program. The process of sending letters is carried out manually by administrative staff through the platform, according to the hospital's criteria, using the target population (60–69 years), excluding those in the exclusion table. Administrative staff is in charge of setting dates, choosing the number of patients to send the letter and gathering positive results. This process is time costly and prone to errors, requiring additional validation.

- Test: Information about the tests performed on patients throughout the program. In particular, the tests carried out are the following:

    ○ Fecal immunochemical test (FIT): The result of this test comes from several laboratories, whose information is automatically uploaded to the table. This implies the need for a homogenization process for the information provided by the different labs, which might also provoke misunderstandings and associated errors.

    ○ Colonoscopy: The anatomo-pathological results of this test comes from several pathology laboratories, whose information is translated to the tables by health staff, which may also imply additional errors. Regarding the findings, the tables distinguish between the information about polyps and cancer lesions detected in the colonoscopy.

The whole information regarding the test procedure, preparation, exploration and findings is analyzed and entered into the platform by health professionals with different roles.

In summary, the information in the database comes from different external databases whose information is automatically dumped as well as from data filling by hospital staff with different roles. In these situations where several agents are involved and different information is crossed, we must ensure a good database design, proper data integration and an appropriate data checking and validation.

2.2 Entity-relationship model

The entity-relationship model facilitates the representation of the relationships between the entities. The main objectives of having an entity-relationship model are [14, 15]:

- To allow a high degree of independence between the application/platform and the internal representation of data.

- To provide a solid basis for addressing data consistency and redundancy.

Figure 1 shows the entity-relationship diagram of our database that represents the relationships between the entities. For simplicity, these relationships are shown

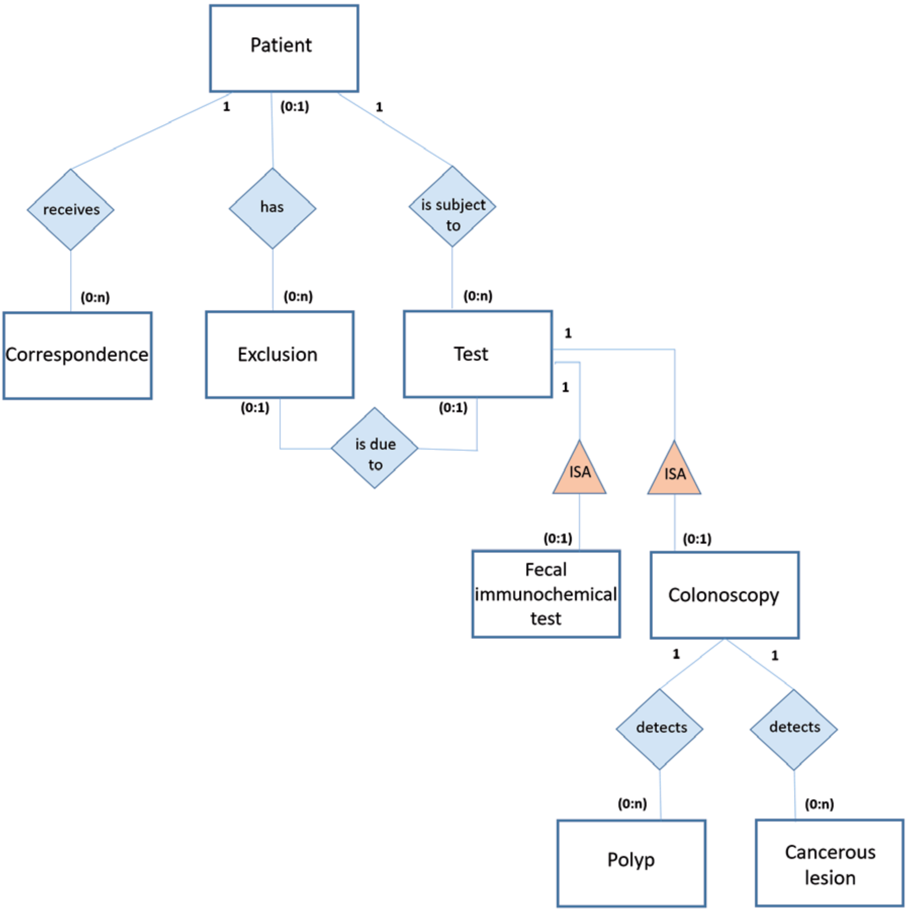


Figure 1.
Entity–relationship model using Chen notation.

in the diagram by means of the Chen notation [16]. The main relationships among data that may be found in the database of the colo-rectal cancer screening program are:

- One patient belonging to the target population may not have any invitation since their invitation to the program has not been processed yet, or may have received more than one invitation to the program over the years (1–0:n).

- A patient may (or may not) be excluded from the program (0:1–0:n) for several reasons. As mentioned above, this exclusion may (or may not) be due to a negative FIT result or colonoscopy findings (0:1–0:1).

- The target patient decides whether (or not) to undergo an FIT and also a colonoscopy if the FIT result is positive (1–0:n).

- In each colonoscopy, findings can be detected (or not) such as cancer lesions and/or polyps (1–0:n).

The fields for each entity are presented below. In total, the database contains about 140 fields.

Patient entity. The patient entity contains 12 attributes with basic patient demographic information in addition to his/her identifier. These fields are related to the date of birth, sex, the round in which the program is at the time in which the patient was enrolled, as well as the place of residence, the health district and the hospital to which the patient belongs.

Exclusion entity. The exclusion entity contains 6 attributes related to the period of exclusion from the program (date of exclusion and, in the case of temporary exclusion, the date of inclusion), the reason for exclusion, a number that determines the priority of exclusion, a binary field that determines whether the exclusion was entered manually and a free text field for comments. If a patient had more than one exclusion, the one with the highest priority prevails. As shown in **Figure 1**, in addition to these fields, the table contains the unique exclusion identifier, the patient identifier, and the test identifier if the exclusion was due to such test.

Correspondence entity. The correspondence entity contains 8 attributes which are as follows: the time when the correspondence was sent (date and time), the type of correspondence sent to the patient (invitation to the program, FIT result, date of scheduled colonoscopy), the round in which the patient was enrolled at the time when the correspondence was received, a binary field that determines if the patient agreed to participate in the program, a binary field that determines if the test recipient was received successfully, a binary field that determines if the patient was included in the program on demand and a free text field for notes. In addition, the table contains the unique identifier of the correspondence and the patient identifier.

Test entity. The entity related to the tests of the screening program contains a large number of attributes (>80). For simplicity, we present here only a high-level description with additional detail for the most relevant aspects. Specifically, the entity contains attributes related to the patient's condition prior to the test and after ending the patient's cycle (round, if the cycle ended, the reason for such ending and the patient's situation after finishing the round).

Regarding the FIT inheritance table, the fields are associated with the date of the interview at primary care, the date of the test and the result of the test. In particular, a field for continuous values of blood concentration in feces (ng/ml), a binary field that determines whether the test was positive and fields that determine whether the sample and the test were correct.

Concerning the colonoscopy inheritance table, it contains a big number of fields related to the following information: basic colonoscopy information (date and time of the scheduled colonoscopy, whether it was performed or not, actual date and time of the colonoscopy and the reason for being performed), colonoscopy preparation (drugs, tolerance, modality, colonic preparation and Boston scale [17]), the process during the colonoscopy (tolerance, which zone was reached, the duration, adequacy of the colonoscopy, etc.), the treatment used during the colonoscopy (type of sedation, type of endoscopic treatment, etc.), the findings found during the colonoscopy (main result: normal colonoscopy, non-neoplastic pathology, polyps, polyposis, cancer, cancer associated with polyposis; risk degree: no risk, low risk, medium risk, high risk and cancer; number of polyps, adenomas, cancer lesions), with the possible complications after the operation (type of complication, whether hospitalization was required and if the patient passed away within the following 30 days…) and possible repetitions of the colonoscopy if required.

Polyp entity. The polyp entity contains, in addition to the identifier of each polyp and the colonoscopy test identifier, 12 attributes concerning the order, size, histology, dysplasia, shape and location of the detected polyp as well as the method for the polypectomy performed, the treatment, the removal performed, etc.

Cancer lesion entity. The cancer lesions entity contains, in addition to the identifier of each lesion and the colonoscopy test, 12 attributes related to the order, size, histology, location of the lesion detected, as well as the stage of the cancer lesion, presence of occluding structure, the type of primary resection and the type of chemotherapy or radiotherapy if applied.

The entity relation model is therefore clear and well defined. However, we should not forget that, in these situations where several agents are involved and different information is crossed, we must ensure proper data integration from a correct database design. This is analyzed in the next section.

3. Analysis and recommendations

An incorrect design of the database and/or the platform often ends up with deficiencies, noise and mistakes in the data, which might prevent a rigorous analysis. In order to have information of sufficient quality to guide appropriate clinical decisions, it is necessary to follow basic principles for data collection and management.

The underlying overall objective of the chapter, and of this section in particular, is to establish from a general perspective, a set of basic principles regarding the integrity, consistency and coherence of data that must be met by any data management system. In particular, for our case study, we thoroughly analyzed whether each of these principles was met and, if not, we proposed a series of recommendations to mitigate the noise of the data and improve the quality of data management. In this analysis it was fundamental to work in a multidisciplinary team with biomedical, statistical and database experts.

The derived recommendations correspond to prospective improvement actions related to data filling, platform characteristics and database design. However, if the information is intended to be used retrospectively, it is also necessary to carry out an additional data curation action. In the following section we explain this curation process in our case study.

In summary, the underlying objective is to highlight the importance of an effective data governance [18, 19], a concept that refers to the ability of an organization to guarantee high quality data throughout its lifecycle, ensuring principles such as availability, easy use, consistency, integrity and security of data. The data manager must ensure such data governance principles and processes.

This concept is crucial as organizations rely more and more on data analysis to optimize their processes and to take relevant decisions [20]. In our particular case, quality data are essential to extract statistical information such as the screening program indicators, or to carry out studies with the objective of improving the overall healthcare system. Some examples are the establishment of the cut-off point to undergo a colonoscopy, a risk analysis to identify risk factors and decisions taken to minimize the undetected lesions, but always based on data evidence.

3.1 Principles

Some of the basic principles, related to data integrity, coherence and consistency, that we analyzed are the following:

- Information utility: All fields defined in the database (or variable to be introduced in the platform) must be filled for some entity (or record).

- Maintenance of consistency: The database or data manager must ensure the stability of the information to any change in the procedure/process and/or to any data dump from an external database.

- Redundancy control: Each register should be uniquely identified. A good database design avoids having more than one field identifying the same event.

- Clarity of the data dictionary: The information of each field of the database (or variable to be introduced in the platform) must be clearly established without any doubt for any user. The information of the field is related to:

    - ○ Name of the field

    - ○ Description of the field (unambiguity of the information)

    - ○ Mandatory

    - ○ Data type

    - ○ Default value

    - ○ Range of values

    - ○ Primary key or foreign key

    - ○ Table to which it belongs

- Management of relationships: The relationships between the fields of the database (or variables in the platform) must be established clearly.

- Control fields in the tables: Fields that identify the creation date, last change date, deletion date, deletion bit, creation user, last change user/process and deleted user/process allow to control the process changes in the data management.

Without loss of generality, we present in this section the analysis of these principles for the fields analyzed and introduce general and specific recommendations to comply with these principles and guarantee good data quality.

Information utility. First, the completeness of the fields in the database tables (about 140 fields) was analyzed. We detected two fields that were not filled, and eight fields defined in the database that were not filled for any entity. The latter can be variables that were defined at the beginning but were never used. To comply with the principle of useful information and to maintain a clean database, these variables should be removed.

Consistency of the information. First and regarding procedure changes, we detected some variables that were no longer used from a certain time; in particular, examples are the binary field that determines whether the patient agrees to participate in the program and the field that determines whether the patient was included in the program on request. **Figure 2** displays the time graph that represents the completeness of these variables over time, where the value 1 indicates completion. As can be seen, from mid-2016, the variables were not completed even once. In this case, and in order to maintain the consistency of the information, from that date of change, the variable should disappear from the platform and the values in the database should be filled to null by default.

Another example concerns the type of correspondence. Currently (from 2017) 3 types of correspondence are delivered: invitations to the program, notification of negative FIT result and notification of positive FIT result along with the scheduled

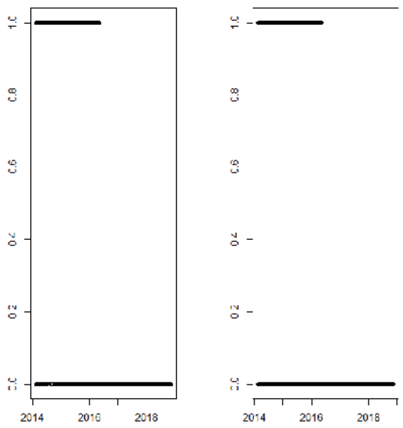


Figure 2.
Time chart of completeness.

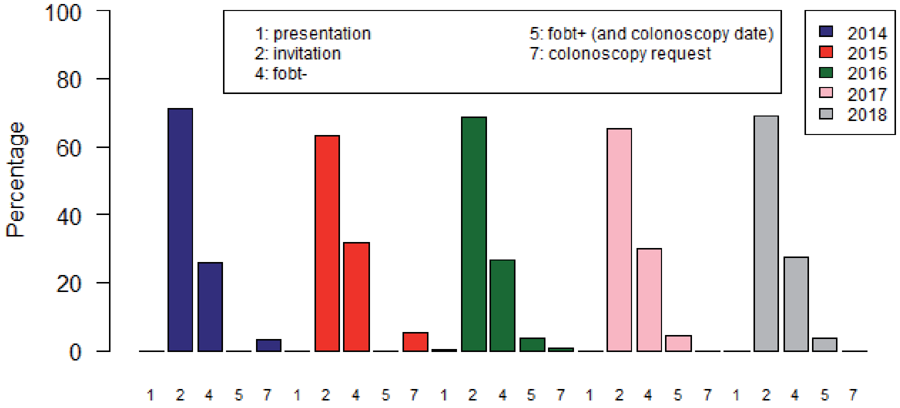


Figure 3.
Distribution of type of correspondence over the years.

date for the colonoscopy. However, previously, the procedure for the FIT positive result and the colonoscopy request was different, as shown in **Figure 3**.

Therefore, currently the value of the field corresponding to the positive FIT notification together with the scheduled date for the colonoscopy (value = 5) refer to a different type of correspondence than in previous years. These changes in the definition of the fields are not recommended since they do not ensure the stability of the information. If they are eventually made, they should be documented and to keep in mind that the historical information should be translated into its current equivalent.

As discussed in the previous section, some of the information in the screening program were extracted from external databases. Therefore, it is also important to analyze its source and the quality of such external sources. Consequently, the quality of the information in the external databases was analyzed and some shortcomings were identified.

For example, in the exclusion process, those exclusions due to findings of cancer lesions in the colonoscopy were considered as temporary exclusions (for 10 years), when it should be a permanent exclusion since the patient as part of the "high risk" group is transferred to the digestive service specialists. Another deficiency found was related to the date of exclusion and, consequently, of inclusion in the program. In particular, in the interview in primary care (OMI database), when a patient fulfills some reason for exclusion, the date of exclusion, that is stored is the one of the interview, and not the actual date when the reason for exclusion was detected. This is important since an erroneous date of inclusion leads to the patient being (falsely) part of the target population at a certain time or the opposite, i.e., not being part of the target population when he/she should be.

These shortcomings involve importing incorrect information into the database and, at best, manual human correction. Ideally, these deficiencies should be corrected from these external databases but since this control can be more difficult and limited, the recommendation regarding our database in these cases would be to correctly identify the possible cases and relationships that must be met (requirements of the screening program database). Also, it is important to make a procedure where only those records that do not induce conflict are updated in the database while the other cases should be reported to allow the user their modification and import/store them correctly. Finally, the automatic generation of reports is also desirable.

In addition to the above deficiencies, in particular in the information from the laboratory database (fecal immunohistochemical test), some records were detected whose information in some fields was crushed or deleted. An incremental import of the information from the external databases would guarantee and ensure the correct storage of the manual changes and would prevent their deletion (since the original unmodified information would not be reloaded).

Redundancy control. Another basic principle for a good database (or data manager) design is the control of redundancy. In particular, the database analyzed does not fully comply with this principle as it contains several fields that identify the same event and, therefore, with redundant/repeated information.

Some examples are the fields related to the result of the FIT: on the one hand, there is a binary field to determine whether the test is positive (> = 117 ng/ml) or negative and, on the other hand, the field representing the quantitative value of the test (ng/ml). Another example detected was the variables related to colonic preparation: colonic preparation in the left colon, colonic preparation in the right colon, colonic preparation in the transverse colon and the Boston scale (from 0 to 9). The latter is, by definition, the sum of the values of the three previous ones.

These examples are fields with a deterministic relationship, where some are the result of the information of others. Therefore, if redundancy control is not fulfilled and the redundant fields are maintained, at least it should be guaranteed that these relations are fulfilled in a deterministic way both in the database and in the platform, self-calculating the fields and/or restricting their values according to the information of the rest of the related fields. However, the analysis carried out revealed that these relationships were not considered in the platform or in the database. This can be observed in **Tables 1** and **2**. **Table 1** shows the qualitative variable of the FIT and the transformation to a categorical variable from the quantitative variable given the current cut-off point (117 ng/ml). **Table 2** shows the variable

Cuantitative/Cualitative	**Negative FIT**	**Positive FIT**	**Total**
Concentration < 117 ng/ml	59.95	0.02	59.97
Concentration ≥ 117 ng/ml	0.02	10.99	11.01
Empty value	27.42	1.59	29.01
Total	87.39	12.6	100

Table 1.
Contingence table (%): Fetal occult blood concentration.

Left+right+transverse/ Boston scale	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Empty**
0	6557	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	0	0	0	0	0	0	0
2	1	0	10	0	0	0	0	0	0	0	0
3	16	0	0	38	0	0	0	0	0	0	0
4	40	0	0	0	65	0	0	0	0	0	0
5	55	0	0	0	0	158	0	0	0	0	0
6	280	0	0	0	0	0	884	0	0	0	0
7	234	0	1	0	0	0	0	450	0	0	2
8	394	0	1	0	0	0	0	0	673	0	1
9	589	0	0	0	0	0	0	0	0	1100	2
Empty	1462	3	10	39	41	97	411	259	297	314	80211

Table 2.
Contingence table (absolute frequency): sum of the colonic preparation of the left, right and transverse part of the colon vs. Boston scale.

calculated as the sum of the values of the variables of the colonic preparation of the left, right and transversal part of the colon versus the qualitative variable of the Boston scale.

If the database and the platform guarantee these deterministic relations of redundant fields, the above tables should be diagonal matrices. However, the analysis showed this weakness in data consistency and showed that neither the platform nor the database considers these relationships, permitting the user an unrestricted completion of those fields, which could lead to data inconsistencies.

Therefore, the analysis carried out showed not only a lack of redundancy control, but also a lack of consistency in the data management. As a recommendation, redundant information should either be removed or, if not, these deterministic restrictions should be established both in the platform and in the database in such a way as to ensure that the relevant information entered is consistent and not confusing.

Data dictionary and relations. As mentioned above, in order to be clear about the meaning of each of the tables and their fields, it is advisable to prepare a priori a data dictionary where each of the fields of each table and the relationships to be established between them are clearly defined. The minimum information to establish, whenever possible, is the following: name and description of the field, mandatory (or not) field, type of data, default value and range of values. However, the analysis performed showed the non-existence of an explicit data dictionary.

An example of ambiguity in the definition would be the variable "round" which appears in both the correspondence table and the test table as well as in the patient's table. Its name is ambiguous since its meaning leads to confusion, having two possible alternatives: it indicates either the patient's round in the program or the current hospital's call round.

It would be natural to think that the "round" variable in the correspondence table refers to the program round and the "round" variable in the test and patient tables refers to the patient round. However, after an exhaustive analysis of these variables, it was concluded that no clear definition of the variable could be extracted from either table. Specifically, if it were "round by patient" the following basic hypothesis should be fulfilled: if a patient has round 2, he/she must also have had round 1. However, patients with round 2 who had not had round 1 were detected in the data. If it were "round by program" the following basic hypothesis should be fulfilled: for the same patient the variable "round" should be increasing over time, that is, if a patient has round 2 at a certain moment, at later dates he or she should have round greater than or equal to 2. However, this was not fulfilled either. Therefore, as commented, we concluded that there is no clear definition of the variable and its completion may be ambiguous. Furthermore, this variable is of great importance since it allows the temporal follow-up of the patient in the program. Thus, the recommendation is crucial in this particular case: to establish a consistent definition of the round variable that is implemented both in the database and in the platform.

Another field information to be established is its mandatory character, if any. The information collected in these fields is the minimum information required to have quality information. In our case, the mandatory variables would be those containing the minimum information of the screening program. However, the analysis showed that there were no mandatory fields established (neither in the platform nor in the database). As an example, each FIT should have the minimum information of its date and its quantitative result (ng/ml), however this does not always happen, as shown in **Figure 4**.

Some fields, like those above, are of a permanent mandatory character, while others may have this nature depending on the values of other fields. For example, the field indicating the findings found in the colonoscopy should be mandatory if the colonoscopy was performed. A good database design should establish this obligation permanently or with restrictions in all necessary cases. As far as the platform is concerned, this obligation should also be established in such a way that all completed information cannot be saved if the minimum necessary information is not filled in.

The data type is another kind of fundamental information to be established for each field. The analysis carried out revealed a lack of consistency in this regard: there are several free text fields in the platform with no restrictions. In particular, the field that determines the time of the next colonoscopy is a free text field that provokes that each user is free to interpret the type of data, filling it in with three different types of data: strings, integers or dates. Examples are the following:

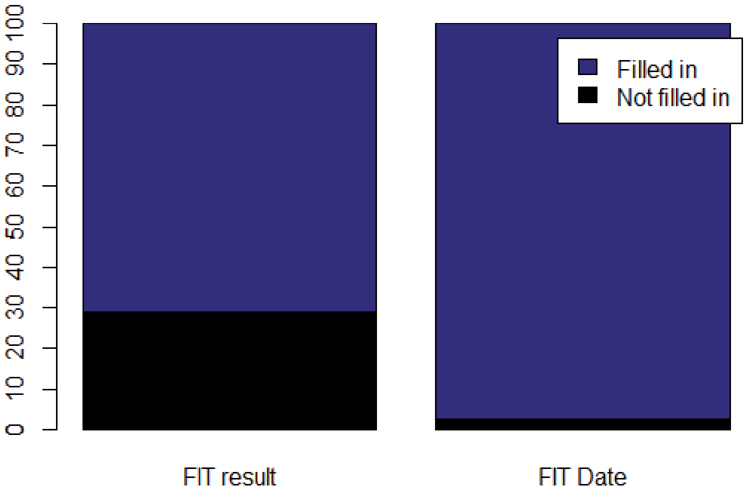


Figure 4.
FIT filling distribution.

"In 5 years", "Not required", "-5 years", "3", "09/05/2016", etc. This means that, in order to use this information, it is necessary to standardize it in the same format, which entails certain difficulties and limitations. For example, the user who filled the value "3" may have referred to months or years and, if this is not established in the type of data or in the definition, this information cannot be used in an analysis. In addition to this, the normalization process of a text-type field takes a great effort [21]. In our particular case, for example, one user entered "In 5 years", another filled the field ("Not required") when the message shows that it should not have been filled (he misuses it as a note) and another uses the mathematical minus sign ("- 5 years") which could mean that the next colonoscopy should be performed in less than 5 years or it could be a simple filling error.

It is therefore necessary to consciously establish the type of data to avoid problems of ambiguity that are difficult to deal with. In addition, the number of free text fields should be limited and, a training effort should be made for the staff who handle and fill the data in order to standardize and unify their interpretation.

The analysis also found that there were fields without a fixed default value or an inadequate default value either in the database or in the platform. For example, it was observed that in some numerical fields both the value 0 and the null value were used indistinctly as default values, which leads to ambiguity in the interpretation of the information for the value 0 which may indicate either the value itself or its default value.

An appropriate default value should be established for each field to avoid ambiguity in the subsequent interpretation of the information and to ensure adequate data quality.

In addition to a clear definition of the field, its mandatory nature, its data type and its default value, restricting the field to a certain range of values is also important in the definition of the data dictionary as it limits the information to possible values and mitigates noise, i.e. possible filling errors and ambiguity problems. If this restriction of the range of possible values is not contemplated, a series of warnings or alerts should be at least implemented to notify the user of an outlier value and the need to revise it.

At this point, both the platform and the database showed weaknesses as there is also a lack of constraints in this regard and no alerts were implemented. An example can be seen in **Table 3** that shows the distribution of values (minimum, quartiles (Q1, Q2, and Q3), mean and maximum value) in the field "weight (kg)" of the patient. On the one hand, weights of 0 kgs in the screening program are not possible, while at least 50% of the filled values took this value. This is an error that can potentially come from not setting the default value or from an incorrect default value, as mentioned above. This would imply that the value 0 was taken incorrectly as default value, distorting the statistics. On the other hand, very high weights (e.g. 81,700 kg) are also inconsistent and may come from human error in the filling process. This example shows that if the fields included a range of possible concrete values or outlier alerts, these errors would be mitigated and, consequently, better quality data is got.

One of the key principles for ensuring consistency in data is to look at the relationships between fields through appropriate constraints. Some of these

Min	Q1	Q2	Mean	Q3	Max	Num. NA's
0	0	0	34	70	81,700	80,415

Table 3.
Patient weight distribution.

relationships may be deterministic, such as those between fields that identify the same event and, as discussed above, in these cases, the constraint must be clear, and preferentially the associated values should be self-calculated. Other relations correspond to restrictions in the values of a field depending on the value or values of other fields and others to restrictions related to the mandatory filling of a field depending on the filling of others.

In relation to this, the analysis carried out found a lack of constraints in the fields which may lead to data inconsistencies, sometimes difficult to correct. For example, if this principle were fulfilled, the following should occur: if the FIT concentration is greater than or equal to 117 ng/ml (cut-off point), the FIT result variable should not be "positive"; however, this restriction was not always considered. A characteristic example related to the restriction of values depending on the value of another field is the one of the monitoring dates that should follow a chronological order (e.g. date of invitation<date of sample reception<FIT result date<colonoscopy date...), however, these inequalities were not always met. Other example is the following: if the field that determines whether the colonoscopy was performed is equal to "No", then the variables related to the colonoscopy should not be filled.

Establishing these constraints, both in the database and in the platform by activating or not the fields in the platform, is fundamental to avoid possible inconsistencies in the data which, on some occasions, can be remedied by curing the data and, on other cases, it is unfeasible to know what the real information in the data is. These restrictions can also be accompanied by alerts or warnings in the platform to help the user and avoid mistakes.

These constraints must be implemented not only in the data filling but also in the deletion, that is, they must guarantee that when the user the value of a variable, the data related to such value must be deleted. For example, if the variable indicating whether a colonoscopy was performed changes its value from "Yes" to "No", then all variables related to the colonoscopy should be set to their default or null value, thus deleting their last filled-in values.

In summary, the analysis showed that a conscious establishment of the values for each field, the data dictionary and a good training of the staff who handle the data is crucial. The more limited and defined the information to be entered is, the better the data will be processed, resulting in fewer errors and less problems of ambiguity, many of which are difficult to deal with subsequently. In addition, the implementation of alerts in the platform could also help to mitigate those filling errors. It is also crucial to thoroughly analyze all possible relationships between all fields in the database and to establish these constraints in the database or in the data manager.

This section has highlighted the inconsistencies, incoherence and errors (some difficult to fix) that can occur in a database if it does not comply with the basic principles of good data management, especially when different agents are involved (external databases, staff with different roles, etc.). As a first conclusion, a good data governance is required to guarantee data quality permitting the extraction of reliable knowledge.

4. Data curation process

The recommendations suggested are referred to improvement measures to comply with the basic principles for a correct design of the database, with the aim of improving the quality of data in the future. However, on many occasions, such data are needed to be used retrospectively. In such cases, a previous curation process is required to eliminate as many errors as possible. In our particular case, the information in the CRC screening database was used to obtain annual indicators of the screening program [22] and to analyze different scenarios for decision making based on the FIT cut-off point, the colonoscopies offered, the target population and the risk factors in order to minimize undetected lesions.

The data curation process carried out was done in the most conservative way possible, and it was mainly based on the relationships that can be established between the fields, recalculating the inconsistent values according to the values of the most reliable/secure reference fields and, if this was not possible, either setting the values that produce inconsistencies to null or finally by removing the whole record from the data set to be analyzed. To carry out this process it was necessary to have a multidisciplinary team composed of statisticians and clinical staff so that statistical knowledge and decision making was supported by knowledge on the environment.

This section explains the main steps and difficulties of the curation process carried out chronologically in order to obtain a clean dataset. The variables presented are the most representative and important ones to carry out the studies required: related to FIT, colonoscopy and follow-up.

4.1 Fecal Immunohistochemical test (FIT)

As commented in the previous section, the relationship between the quantitative variable of the FIT concentration and the qualitative variable (negative, positive FIT…) is not fulfilled in a deterministic way as it should be. Below we present the most representative cases of incoherence and how they were cured:

- Records with concentration = 117 ng/ml but with "negative" FIT result and records with negative concentration were detected. Cooperation with health staff was key here. With regard to the first case, after several meetings, it was concluded that they were values from the laboratory where the report specified the value at "-117", referring to "less than 117". For the second, it was deduced that a hospital considered the cut-off point of 117 ng/ml as negative. In order to standardize the information from all hospitals it was decided to re-establish the value of the quantitative concentration at 116 ng/ml in these records.

- Records were detected with concentration > 117 ng/ml but with negative FIT result. It was found that these records were two tests for the same patient in which the second test overwrote the qualitative value of FIT but maintained the old value of the quantitative one. In these cases, it was decided to follow the more conservative decision and take the information of the first one (with its quantitative value) and recalculate the qualitative value.

- Records with concentration < 117 ng/ml were detected but with positive FIT results: This was one of the errors that were not possible to re-establish and, therefore, after several meetings, it was decided to establish the values of the concentration at null and to save the identification of those patients in order to establish the correct value of the concentration in the future in case they are identified.

- Once the concentration values were corrected, the qualitative variable of the FIT was recalculated in the cases where the concentration was different from zero.

4.2 Colonoscopy

The variable that indicates the performance of the colonoscopy (colonoscopy variable) is also strictly related to the result of the FIT, and, obviously, with the variables associated with the colonoscopy. Below we present the inconsistencies found and the steps followed for their curation:

- Records were detected in which the FIT result was negative, and the colonoscopy variable took the value "Yes". After studying it, it was concluded that this error was likely due to an overwriting or crushing of the data. Since this error does not allow the recovery of realistic values from the record, it was decided to remove these records from the data set.

- Records were detected that had a value in the variable that indicates the result of the colonoscopy (normal colonoscopy, polyps, cancerous lesion...) and with a completed colonoscopy date, and yet the colonoscopy variable did not take the value "Yes". These cases are a clear example of error because the colonoscopy variable was not defined as mandatory in the database. In these cases the value of the colonoscopy variable was reset to "Yes".

- All records with no colonoscopy value were reset to 0. This is due to the lack of definition of the default value of this field in the database.

- Records with no information on colonoscopy-related variables were detected, yet the colonoscopy variable was equal to "Yes". This error was possible thanks to the lack of constraints in the database. In these cases the colonoscopy variable was reset to "No".

- The variable that determines the reason for the exploration in the cases in which the colonoscopy variable was equal to "No" was reset to null.

4.3 Follow-up

When the records have all their process information filled, they should have the "end of cycle" variable filled to "Yes". Therefore, in the analysis we must take the records that have finished their cycle ("end of cycle" = "Yes"). However, this variable was not defined as mandatory in the database and, therefore, presents inconsistencies that were solved according to the following rules:

- All records with the colonoscopy variable equal to "No" should be closed cycles and, therefore, the variable "end of cycle" should take the value "Yes".

- All records with the variable that determines the reason for the end of the cycle completed refer to the patient's closed cycles and, therefore, the variable "end of cycle" should take the value "Yes".

- All colonoscopies prior to the last year (2018) with "end of cycle" equal to "No" would really be considered as closed cycle ("end of cycle" = "Yes") since the information should take less than one year to be filled in.

The variables related to dates are important since they allow the information recorded to be followed up by years. Therefore, their completion should be ensured as far as possible, in particular the date of the sample result and the date of the colonoscopy which are the most crucial ones. The dates considered for completion are the following: date of invitation to the program, date of interview at primary care, date of reception of the sample, date of result of FIT, date scheduled for the colonoscopy, date of colonoscopy. Chronological completion was done as follows:

- In the records where the FIT result date was not completed, it was set as follows in order of preference: result date = colonoscopy date –1-month, result date = sample receipt date, result date = OMI date, result date = colonoscopy schedule date-1 month, result date = program invitation date+1 month.

- In the records where the colonoscopy date was not completed, it was established considering the same date as the date programmed or considering the result date of the FIT-1 month.

In this section the most important aspects of the curation process carried out in order to obtain a clean data set with reliable information on which to carry out future analyses have been introduced. Despite they are specific, it has been shown, the wide range of potential bugs that may appear due to a wrong design of the database.

5. Conclusions

Good data management and consequently good data quality must comply with some basic principles. This is especially important when those data are used to support decision makers in public health services. In this chapter, we analyze the database of a regional CRC screening program to identify the weaknesses in the process of data collection, providing some guidelines for future maintenance. We also identified incorrections in the database design that may lead to data errors. General and specific recommendations were suggested to meet the requirements of data integrity, consistency and coherence.

However, most of these recommendations are forward-looking suggestions, i.e. they will improve the quality of future data from the moment they are considered. Simultaneously, and in order to be able to exploit the information retrospectively, it was necessary to make a data curation of the historical information. To do this, a clean-up process was followed in the most conservative way possible, re-establishing values, cleaning-up some data and discarding repetitive or non-essential data, trying to eliminate as many errors as possible and guarantee good quality data both prospectively and retrospectively. This process is time costly and tedious, but it is an essential first step in data governance to extract reliable knowledge and taking correct decisions.

In summary, this analysis showed the importance of data quality and curation to get a robust, consistent and reliable database, as well as the need for a good design of the data acquisition process and, finally, a proper and coherent maintenance system, especially in health systems where the decisions derived from the analysis of databases may be critical.

Acknowledgements

We acknowledge the help provided by the Department of Health of the Aragón Government for allowing us the access to anonymized patient data and to the EU Program for Employment and Social Innovation (EaSI) for partial financial support.

Author details

Rocio Aznar-Gimeno[1], Patricia Carrera-Lasfuentes[2],
Vega Rodrigalvarez-Chamarro[1], Rafael del-Hoyo-Alonso[1], Angel Lanas[2]
and Manuel Doblare[3]*

1 Technological Institute of Aragon (Itainnova), Zaragoza, Spain

2 Aragon Institute of Health Research (IISAragon) and CIBERehd, Zaragoza, Spain

3 Aragon Institute of Health Research (IISAragon) and CIBERbbn, Zaragoza, Spain

*Address all correspondence to: mdoblare@unizar.es

References

[1] Abadi, D., et al. The Beckman report on database research. Communications ACM, 2016, vol. 59(2), p. 92-99

[2] da Costa, C.A., Pasluosta, C.F., Eskofier, B., Bandeirada, D., Rodrigoda, S. and Righi, R. Internet of Health Things: Toward intelligent vital signs monitoring in hospital wards. Artificial intelligence in medicine, 2018 vol. 89, p. 61-69

[3] Bhalla, S., Sachdeva, S. and Batra, S. Semantic interoperability in electronic health record databases: Standards, architecture and e-health systems. In 5th International Conference on Big Data Analytics, Hyderabad, India, 2017. Lecture Notes in Computer Science book series (LNCS, volume 10721)

[4] Biswas, S., Sharif, K., Li, F., Latif, Z., Kanhere, S.S. and Mohanty, S.P. Interoperability and Synchronization Management of Blockchain-Based Decentralized e-Health Systems, in IEEE Transactions on Engineering Management, 2020, vol. 67(4), p. 1363-1376, doi: 10.1109/TEM.2020.2989779.

[5] Dipak, K., Beale, T. and Sam Heard. The openEHR foundation. Studies in health technology and informatics, 2005, vol. 115, p. 153-173. PMID: 16160223.

[6] Pathak, J., Bailey, K.R., Beebe, C.E., Bethard, S., Carrell, D.S., Chen, P.J., … and Chute, C.G. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association, 2013, vol. 20(e2), ep. 341-e348 doi: 10.1136/amiajnl-2013-001939.

[7] Sachdeva, S. and Bhalla, S. Semantic interoperability in standardized electronic health record databases. J. Data Inf. Qual. (JDIQ), 2012 vol. 3(1), p. 1 https://doi.org/10.1145/2166788.2166789

[8] Hoffman, S. and Podgurski. A. Big bad data: law, public health, and biomedical databases. The Journal of Law, Medicine & Ethics, 2013 vol. 41, p. 56-60 https://doi.org/10.1111/jlme.12040

[9] Batra, S. and Sachdeva, S. Pre-Processing Highly Sparse and Frequently Evolving Standardized Electronic Health Records for Mining. Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning. IGI Global, 2020. P. 8-21 doi: 10.4018/978-1-7998-2742-9.ch002

[10] Satti, F. A., Ali, T., Hussain, J., Khan, W. A., Khattak, A. M., and Lee, S. Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability. Computing, 2020, vol. 102(11), 2p. 409-2444. https://doi.org/10.1007/s00607-020-00837-2

[11] Pezoulas, V. C., Kourou, K. D., Kalatzis, F., Exarchos, T. P., Venetsanopoulou, A., Zampeli, E., … and Fotiadis, D. I. Medical data quality assessment: On the development of an automated framework for medical data curation. Computers in biology and medicine, 2019, vol. 107, p. 270-283. doi: 10.1016/j.compbiomed.2019.03.001

[12] Feder, S.L. Data quality in electronic health records research: quality domains and assessment methods. Western journal of nursing research, 2018, vol. 40(5), p. 53-766. doi: 10.1177/0193945916689084

[13] Weiskopf, N. G., and Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.

Journal of the American Medical Informatics Association, 2013, vol. 20(1), p. 44-151. doi:10.1136/amiajnl-2011-000681

[14] Elmasri, R. and Navathe, S.B. (eds) The relational data model and relational database constraints. In Fundamentals of Database Systems, Pearson Addison-Wesley, 2013. ISBN-0133970779

[15] Codd E.F. A Relational Model of Data for Large Shared Data Banks. In: Software Pioneers (Broy M., Denert E. (eds)). Springer Verlag, 2002 https://doi.org/10.1007/978-3-642-59412-0_16

[16] Chen, P.P-S. The entity-relationship model—toward a unified view of data. ACM Transactions on Database Systems, 1976, vol. 1(1), p. 9-36. Doi:10.1145/320434.320440

[17] Calderwood, A.H. and Jacobson, B.C. Comprehensive Validation of the Boston Bowel Preparation Scale. Gastrointestinal Endoscopy, 2010 vol. 72(4) p. 686-692. Doi: 10.1016/j.gie.2010.06.068.

[18] Dama International. Dama-DMBOOK: Data Management Body of Knowledge. Technics Publications, LLC, 2017 ISBN-1634622340

[19] Khatri, V. and Brown, C.V. Designing data governance. Communications of the ACM, 2010, vol. 53, no 1, p. 148-152. Doi: 10.1145/1629175.1629210

[20] Wieten, E., Schreuders, E.H., Nieuwenburg, S.AV., Hansen, B.E., Lansdorp-Vogelaar, I., Kuipers, E.H., Bruno, M.J. and Spaander, M.C.W. Effects of increasing screening age and fecal hemoglobin cutoff concentrations in a colo-rectal cancer screening program. Clinical Gastroenterology and Hepatology, 2016, vol. 14, no 12, p. 1771-1777. Doi:10.1016/j.cgh.2016.08.016

[21] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M. and Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics, 2017, vol. 73, p. 14-29. Doi: 10.1016/j.jbi.2017.07.012

[22] Llop, E.S., Cano del Pozo, M., García Montero, J.I., Carrera-Lasfuentes, P. and Lanas A. Colo-rectal cancer screening program in Aragon (Spain): preliminary results Gaceta sanitaria, 2018, vol. 32, no 6, p. 559-562. doi: 10.1016/j.gaceta.2017.05.014

Section 3

Data Interaction

Chapter 4

Big Data Integration Solutions in Organizations: A Domain-Specific Analysis

Sreekantha Desai Karanam,
Rajani Sudhir Kamath, Raja Vittal Rao Kulkarni
and Bantwal Hebbal Sinakatte Karthik Pai

Abstract

Big Data Integration (BDI) process integrates the big data arising from many diverse data sources, data formats presents a unified, valuable, customized, holistic view of data. BDI process is essential to build confidence, facilitate high-quality insights and trends for intelligent decision making in organizations. Integration of big data is a very complex process with many challenges. The data sources for BDI are traditional data warehouses, social networks, Internet of Things (IoT) and online transactions. BDI solutions are deployed on Master Data Management (MDM) systems to support collecting, aggregating and delivering reliable information across the organization. This chapter has conducted an exhaustive review of BDI literature and classified BDI applications based on their domain. The methods, applications, advantages and disadvantage of the research in each paper are tabulated. Taxonomy of concepts, table of acronyms and the organization of the chapter are presented. The number of papers reviewed industry-wise is depicted as a pie chart. A comparative analysis of curated survey papers with specific parameters to discover the research gaps were also tabulated. The research issues, implementation challenges and future trends are highlighted. A case study of BDI solutions implemented in various organizations was also discussed. This chapter concludes with a holistic view of BDI concepts and solutions implemented in organizations.

Keywords: master data management (MDM), internet of things (IoT), business intelligence (BI), software as a service (SAAS), machine learning (ML), artificial intelligence (AI)

1. Introduction

Accenture company has conducted a survey on the implementation of BDI solutions in organizations. The survey outcome revealed that 92% of managers are happy with the results obtained from BDI solutions and 89% of managers agree that big data integration and analytics is very vital for their business planning to leverage competition. The Internet Trends Report from KPCB's by Mary Meeker discovered the decreasing trends in the cost of hardware technology in the past twenty years, the cost of computing has been reduced by 33%, 38% storage cost reduction and 27% bandwidth costs reduction year after year. The major challenges faced in BDI processing are data selection, gathering, storing, communication, searching, visualization, ensuring privacy, and security of data.

The efficiency in handling big data drives effective decision making. The advancement in computing infrastructures, algorithms and innovative technologies have boosted the big data management and analytics domain and reduced the investment costs to deliver the best value for businesses.

1.1 Motivation and significance of BDI study

The business experts have agreed that big data would mean big value. The digital transformation of business operations is enhancing customer experience and reducing costs. Consumers would like to access personalized data and carry out business on the go. Online processing of bigdata using analytical platforms in the organizations can make the information accurate, standardized, and actionable. Acquiring insights from big data leverage the companies to make more informed business decisions with improved efficiency, and to design more BDI applications. The revolution in computing and digitalization has also increased the potential of cyber-attacks. The cyber threats by hackers are ever increasing and becoming more and more complex day by day. ML and DL techniques have been significantly applied to design intelligent and secure BDI solutions for automating business processes. ML projects are receiving the maximum funding since 2019, compared to all other AI projects combined. Walmart corporation has implemented BDI solutions for acquiring business intelligence and taking real-time business decisions. Many leading fast-food based companies such as McDonald's, KFC, Pizzahut are using BDI solutions for designing their marketing strategies to discover the hanging business trends. The Casinos are also utilizing the BDI solutions to enhance their revenues in the recent years and to attract and inspire customers for regular visits. The hotel industry uses BDI applications to predict customer behavior, food habits and demands. Tourists today are also using digital solutions to collect information on all issues related to tourism. BDI has been applied in the healthcare industry for rendering quality healthcare services, decreasing the wastage of money and time. The governments are using BDI for developing smart city public services. BDI has empowered e-commerce industries such as Amazon, Flipkart, etc. by providing data insights and analytical reports. The integration of AI, BDI and visualization tools helped meteorologists to predict weather conditions precisely. BDI solutions have been applied successfully in modern agriculture. BDI solutions have also empowered digital marketing for the success of every business. The above facts and applications have motivated the researchers to study the BDI in detail.

1.2 International market potential

According to global forecasts BDI solutions market size is estimated to reach US$ 12.24 billion by 2022 at a Compound Annual Growth Rate (CAGR) of 13.7%. The market survey by Dresner Advisory Assc. in the year 2020 has revealed that 80% of organizations are considering BDI solutions as critical for decision-making activities and 60% of them prefer to deploy BDI solutions on cloud platforms. International Data Corporation (IDC) has predicted that the global data-sphere would be about 175 zettabytes by 2025. IDC has estimated that several billion IoT devices and embedded systems would generate, gather, communicate a wealth of IoT data and carryout analytics every day throughout the world. IDC has also predicted that by 2025 about six billion customers or 75% of the global population would be communicated by using online and real-time data every day. The share of real-time data would be about 30% in global data as estimated by IDC.

1.3 Overview of BDI technologies

1.3.1 BDI process types

BDI is the process of consolidating data from multiple applications and creating a unified view of data assets. BDI is the main component of various mission-critical data management projects, such as building an enterprise data warehouse, migrating data from one or multiple databases to another, and synchronizing data among applications. BDI directs at furnishing an integrated and consistent view of data coming from external and internal data sources.

1.3.1.1 Data consolidation

Big data consolidation is the process of consolidating or integrating data from various data sources to make a centralized data store or repository. This is an amalgamated data store used for diverse purposes, such as data analysis and reporting. It can also execute for downstream applications as a data source.

1.3.1.2 Data federation

A Data Federation is a data integration technique. Data federation is used to integrate the data and simplify the approach for consuming by the users and front-end applications. In data federation, distributed data with various data models are combined into a unified data model that features a virtual database.

1.3.1.3 Data propagation

It is another technique for data integration. Data would be propagated from an enterprise data warehouse to different data marts after the needed transformations.

1.3.2 BDI technologies

1.3.2.1 Extract, transform, load (ETL)

ETL is the best-known data integration technology. ETL is a process of data integration that includes extraction of data from a source system and it's loading after transformation to a target destination.

1.3.2.2 Enterprise information integration (EII)

This data integration technology is used to deliver curated data-sets on an on-demand basis. EII is a technology that admits developers and business users alike to treat a range of data sources as if they were one database and represent the incoming data in novel ways.

1.3.2.3 Enterprise data replication (EDR)

EDR is a real-time data consolidation method that includes moving data from one storage system to another. In its simplest form, having the same schema, EDR involves shifting a data-set from one database to another database.

1.3.3 Bigdata integration platforms

1.3.3.1 Adeptia connect

Enterprise BDI tools provided by Adeptia may be utilized by other than technical business users. Adeptia Connect has an easy user interface to coordinate with all data interfaces and external connections. It also involves a no-code approach and self-service partner onboarding that allows partners and users to view, set up and coordinate data connections. The platform brags a suite of Cloud Services Integration and pre-built connections along with protocol support and B2B standards.

1.3.3.2 Alooma platform

Alooma provides a data pipeline service that combines with prevalent data sources. The Alooma platform contains security from end-to-end level, which ascertains that every event is securely moved to a data warehouse (HIPAA, SOC2, and EU-US Privacy Shield certified). The solution reacts to the changes in data in real-time to ascertain that no such events have vanished. Users can select to carry out changes automatically or get notified and do on-demand changes. This tool also automatically reduces the data volume to make control customizable.

1.3.3.3 Boomi AtomSphere

It is a Dell Technologies company's Boomi's flagship product. AtomSphere supports an integration process between cloud platforms, software-as-a-service applications, and on-prem systems. The visual interface is used by AtomSphere is used to configure application integrations. Wherever it is needed, the solution's runtime tool, Boomi Atom, allows integrations to be deployed. Based on use case and functionality, the AtomSphere platform is also available in various editions.

1.3.3.4 Integrator.io

Celigo company provides an Integration Platform called Integrator.io as a service product. The solution enables organizations to synchronize the data, connect applications, and automate processes. Celigo bundles an integration wizard that involves visual field mapping interface, an API assistant, and drop-down menus. This tool also provides integration templates which are reusable pre-configured on the integrator.io marketplace, permitting users to have their library of reusable, standalone flows.

1.3.3.5 Cleo Integration Cloud

The Cleo Integration Cloud accords organizations to connect to SaaS applications and enterprises with a range of connectors and APIs. This tool automatically accepts, transforms, orchestrates, connects and integrates all B2B data types from any source and to any target, and can be implemented via several different methods. Cleo Integration Cloud can also be engrafted for Information Services organizations or SaaS and can be used as an administered service to divest complex integrations to the vendor's experts.

1.3.3.6 Denodo Platform

The Denodo Platform provides data virtualization for integrating multi-structured sources of data from database management systems, a wide variety of other big data, cloud, documents, and enterprise sources. Connectivity support involves legacy data, flat files, relational databases, packed applications, CML, and emerging data types including Hadoop. The only data virtualization solution, Denodo, is to be represented as a virtual image on AWS Marketplace of Amazon.

1.3.3.7 Diyotta data integration suite

A unified data integration platform, Diyotta, that combines with data warehousing environments and modern data lake. The native processing capabilities and drag-and-drop user interface to build this product. Diyotta enables faster data movement, shorter development times, and reusability all over the enterprise to simplify future development. Diyotta touts the industry's first data integration software to leverage modern data processing platforms like Snowflake, Google BigQuery, Hadoop, and Amazon Redshift.

1.3.3.8 IBM products - InfoSphere information server

IBM provides several distinct data integration tools in both cloud and on-prem deployments, and for every enterprise use case virtually. Its on-prem data integration suite has tools for modern data integration synchronization, data virtualization) and traditional (replication and batch processing) requirements.

IBM also provides a range of connectors and pre-built functions. The mega-vendors cloud integration product is considered as one of the most excellent in the marketplace.

1.3.3.9 Informatica products - an intelligent data platform

Informatica's data integration tools portfolio covers both cloud deployments and on-prem for several enterprise use cases. The vendor integrates governance functionality and advanced hybrid integration with self-service business access for different analytical functions. Augmented integration is possible via a metadata-driven AI engine, and Informatica's CLAIRE Engine, that enforces machine learning. Informatica touts interoperability in strong in nature.

1.3.3.10 Microsoft Products - SQL Server Integration Services (SSIS).

The company's SQL Server Integration Services (SSIS), traditional integration tools, is integrated inside the SQL Server DBMS platform. Microsoft also promotes two cloud SaaS products: Microsoft Flow and Azure Logic Apps. Flow is adhoc integrator-centric and integrated into the overarching Azure Logic Apps solution.

1.3.3.11 Oracle products - data integration cloud service

Oracle provides a full spectrum of data integration tools for modern ones as well as conventional use cases, in both cloud and on-prem deployments.

The company's product portfolio includes services and technologies that permit organizations for data enrichment and full lifecycle data movement. Oracle data integration allows permanent and uninterrupted access to data across heterogeneous systems via transformation, bidirectional replication, bulk data movement, data services, metadata management, and data quality for product and customer domains.

1.3.3.12 SAP products - data services

SAP provisions clouds and on-prem integration functionality by two primary channels. Traditional capabilities are provided through a data management platform, SAP Data Services, that gives capabilities for data cleansing, integration, and quality. SAP Cloud Platform provides Integration Platform as a Service features are existing in it. Integration of processes and data between cloud apps, third-party applications, and on-prem solutions are arranged through SAP's Cloud Platform.

1.4 Organization of the chapter

This chapter has been framed into seven sections. Section 1 explains the introduction, sub section 1.1 discusses the motivation and significance of the study. Sub section. 1.2 shows international market potential and 1.3 presents an overview of big data technologies and taxonomy. Sub section 1.4 Organization of the chapter, 1.5 summarizes the authors' research contribution. 1.6 Illustrates the list of acronyms used. The review of recent literature is described in Section 2. Section 2 is further divided into four subsections. 2.1 subsection describes the papers reviewed from one technology domain. The highlights and findings from each paper are tabulated. Sub section 2.2 deals with a comparative analysis of survey papers with specific parameters. Section 3 shows the architecture of BDI. 4th section deals with the research issues, challenges. Section 5. Presents the case studies of BDI solutions from various organizations. Section 6. Outlines the findings and conclusion. Section 7 is the references of the papers reviewed.

1.5 Research contribution

1. This study has revealed that various technologies, systems, techniques, algorithms are applied for implementing business intelligence systems across the world. These papers have been further classified technology-wise and presented as a pie chart in **Figure 1**.

2. A table of acronyms is presented in **Table 1**

3. **Figure 2** presents taxonomy of concepts applied in BDI techniques in various applications

4. The overview of the organization of this chapter, section-wise is shown **Figure 3**.

5. In each concept of taxonomy, the existing literature has been mapped to several issues as shown in **Figure 4**.

6. Research issues, challenges and future directions of BDI technologies are discussed

7. A case study of five BDI based solutions implemented healthcare, retail, finance and tourism domains are discussed

8. The set of a curated survey papers are compared with specific factors such as architecture, open issues and challenges, applications, taxonomy and security to understand the scope of coverage each paper and to understand the research gaps **Tables 2** and **3**

1.5.1 Table of acronyms

This section shows a list of all the acronyms used in this chapter for easy reference is presented **Table 1**.

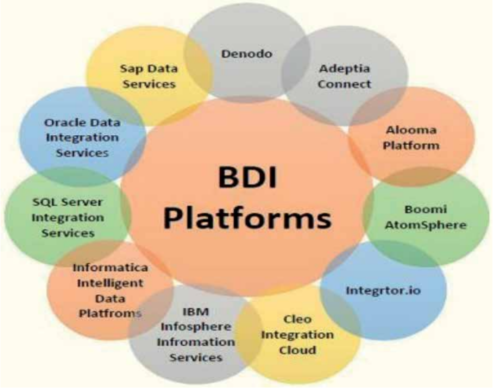

BDI Platforms
Denodo
Sap Data Services
Adeptia Connect
Oracle Data Integration Services
Alooma Platform
SQL Server Integration Services
Boomi AtomSphere
Informatica Intelligent Data Platfroms
Integrtor.io
IBM Infosphere Infromation Services
Cleo Integration Cloud

Figure 1.
Bigdata integration platforms.

Acronym	Description	Acronym	Description
AI	Artificial Intelligence	BDI	Big Data Integration
AWS	Amazon Web Services	AIG	American International Group
B2B	Business to Business	CAGR	Compound Annual Growth Rate
CCTV	Closed Circuit TV	ETL	Extract, Transform, Load
EII	Enterprise Information Integration	EDR	Enterprise Data Replication
HIPAA	Health Insurance Portability and Accountability Act	ICT	Information and Communication Technology
IoT	Internet of Things	MDM	Master Data Management
IoMT	Internet of Medical Things		
SOC 2	Service organization control is an auditing procedure	API	Application Programming Interface

Table 1.
List of acronyms used in this chapter.

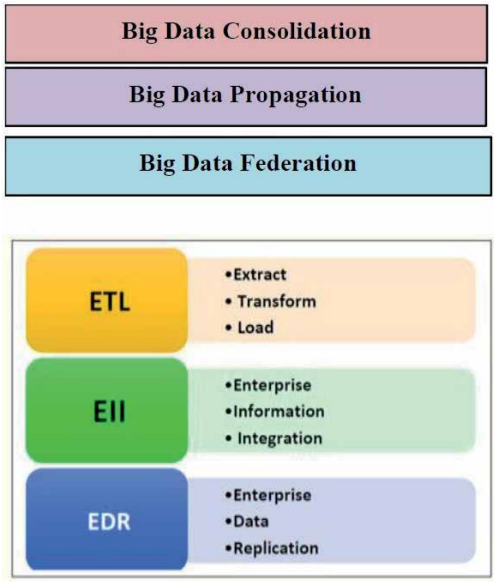


Figure 2.
Taxonomy of big data concepts.

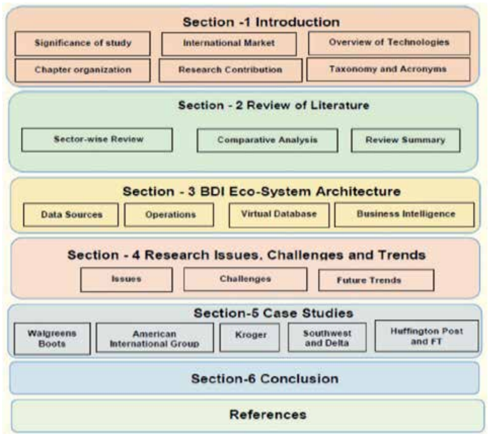


Figure 3.
Organization of the chapter.

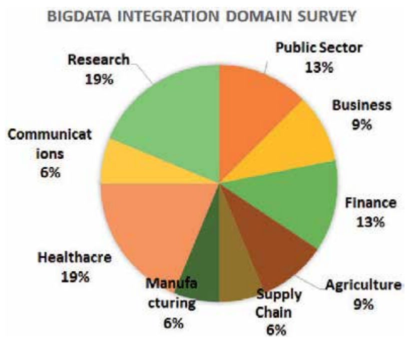


Figure 4.
Sector-wise reviewed papers.

Ref. No.	Methods Used	Results	Applications
1	Interviewing of experts and content analysis approach was used as a qualitative technique for data gathering.	The principal problems identified during the study are the hardship of administration, the ineffectiveness of human resource, politics, standards and absence of executives	Data fusion solutions for public sector
2	Comparisons, experiments and questionnaires	The decorum status of high recognition of positive viewpoint is around 65%. Negative perspective has an absence of etiquette knowledge accounted for 56%	Talent refinement of college students
3	Semantic data model, Resource description system, SPARQL, semantic query language	Semantic extract transform load system produces semantic information that would possibly be distributed on the web as Web of data.	Innovative Big data applications in fuel economy, household transportation and vehicles
4	Exploration of BDI based on stage, features and semantic meaning	Big data problems are resolved by appropriate BDI methods.	Open cross-domain Big data
5	A study on analysis of industrial needs and potential applications of BDI in the public sector	A set of open research questions such as scalability of data required in real-time applications	Labor agency, Online gambling operations, Public Safety, and Predictive policing

Table 2.
Public sector review summary.

Ref. No.	Methods Used	Results	Applications
6	Review of literature on BDI, Business Intelligence and Cloud Computing	It is possible to integrate technologies to the need of SMEs	Small and Medium-sized organizations
7	Developing a deployable data integration tool that handles technical issues.	Shortage of machine learning examples, the requirement of clarifying business owners' outcomes and the expense of involving domain experts.	Scalable data integration challenges in the enterprise
8	The amalgamation of diverse sources in the Hadoop environment with HDFS. Spark computation model with a Hive database as a distributed data warehouse	A prototype of a data integration system	E-Commerce Domain
9	A study on data collection, analytics implementation, and benefits of BDI	BDI implementation in business organizations improves business performance	Performance improvement in business organization
10	Randomly selected articles on big data are reviewed to analyze the role of big data in business	As per the study, 63% of business reported that the implementation of big data is useful to business	Decision making in business

Table 3.
Business sector review summary.

2. Review of recent literature

Authors have selected papers from highly reputed research journals from IEEE, Elsevier, Science Direct and Springer publications. About fifty-five papers covering big data integration concepts and applications are reviewed. This section presents the findings and highlights from each reviewed paper which are organized domain-wise.

2.1 Review of the public sector

Hasliza et al. analyzed the fundamental problems and difficulties encountered by the BDI solutions in the public sector [1]. The discovery of the right dimensions and factors are important to find the solutions to these problems. Zhang has reported the BDI solutions for professional procedure amalgamation in modern decorum [2]. The comparisons, experiments and questionnaires concerning the BDI concepts are discussed. Bansal has proposed the use of semantic technologies for the distribution of information in the contest of semantic ETL [3]. This information is open and the data was gathered from various sources. Zheng et al. have presented significant standards, classification of strategies and models of BDI process [4]. The real BDI issues are discussed using these models.

Authors have classified BDI techniques based on different combinations of strategies such as stage, feature and semantics. Munne has explained the technological trends for current social and economical status. This paper highlighted BDI technologies and applications in the public sector [5]. **Table 2** presents a summary of the highlights of the papers reviewed the public sector.

2.2 Review from business sector literature

The study by Camargo et al. revealed the possibility of incorporating and implementing BDI technologies to the needs of small and medium scale enterprises [6]. Some companies are offering open source BDI tools for organizations for intelligent business decision making. Stonebraker et al. discussed the difficulties in the scalability of BDI solutions today and in near future [7]. This analysis was carried out using the past five years' data from large enterprises. The integration of data from heterogeneous sources in a distributed environment was explored by Sazontev et al. [8]. The authors have explained the process of BDI framework development and its methodology. Alsghaier et al. discussed BDI process in Hadoop platform for business organizations. Authors focused on the implementation and benefits of big data analytics in business organizations [9].

Alam et al. have reviewed the role of BDI in the business sector [10]. **Table 3** presents a summary of highlights from the business sector review.

2.3 Review of the finance sector

Fikri et al. presented a BDI approach combined with distributed datasets of financial ontology and a real-time data stream [11]. This model was associated with classic ETL. This model was suitable for handling BDI in real-time. Bucea-Manea-Tonis illustrated the use of predictive logic in deductive frameworks to integrate different sets of data types [12]. Chen et al. have proposed a framework for managing data with heterogeneity problems [13]. This unified data model was adaptable to different data sources by setting up panoramic data. Hussain and Prieto discussed the analysis of industrial needs, constraints and potential applications of BDI to insurance and finance sectors [14]. Authors have mapped the requirements to research queries. The paper by Avi and Kamaruddin reported the role of BDI in insurance, finance and banking sectors [15]. Authors have highlighted the benefits of cutting-edge technologies associated with BDI in the financial sector. **Table 4** presents a summary of highlights from the finance sector review.

Ref. No.	Methods Used	Results	Applications
11	Combining distributed datasets of financial ontology and real-time stream	The data integration pipeline in real-time. The use of Apache Spark enhances short time frames for quality and availability reporting	Data integration in real-time, Financial reporting
12	Predicate logic in deductive systems	Integrates different kinds of data types	E-Commerce applications
13	Integrating heterogeneous data from multiple sources, Big data ETL in a distributed environment	Better performance in processing data integration from multiple sources	Power dispatching and control system
14	A review of industrial needs, constraints and application	Highlights the challenges in providing an effective technological solution	Manipulation recognition, threat management in finance and insurance sectors
15	A comprehensive review of the banking sector in terms of digital banking, analytics, mobile banking. Use cases of the latest technologies in the banking sector	Various business problems are solved by using the latest technological trends and big data analytics in the banking industry	Big data analytics for the banking industry

Table 4.
Finance sector review summary.

2.4 Review of agriculture sector

BDI and data analytics concepts are emphasized by Nabrzyski et al. [16]. This proposed solution incorporates the execution of complex queries on various datasets. These data sets contain the layers of raster and geospatial data. Kim and Tam (2020) have proposed a data integration estimator [17]. This is a classification technique with non-parametric and overlapping units which recognizes and corrects misclassification errors. Saggi and Jain (2018) have reported a data analytics solution for organizations [18]. This solution performs an exhaustive realistic analysis. The components of BDI application platforms are discussed. Authors have thrown light on past, current research issues and future directions.

Ribarics (2016) explained the importance of big data in agricultural sector [19]. The author has highlighted the need for using technological innovations in farming. Sarker et al. (2020) discussed the impact of BDI in digital farming [20].

The study results showed that big data analytics helps the farmers in crop management and yield forecasting. This study also revealed that BDI in farming is not fully established. **Table 5** presents a summary of highlights from the agriculture sector review.

2.5 Review of literature on BDI in smart cities

Kaur and Kushwaha (2018) were motivated by different applications of BDI and IoT integration in smart cities [21]. The earlier researchers reviewed the critical data analysis issues. Huang et al. (2014) have proposed HiperFuse solution for addressing BDI challenges and automating the BDI process [22]. Nuaimi et al. (2015) reviewed the prospects, issues and advantages of BDI in smart cities. This study discussed the BDI challenges faced in smart cities [23]. Gomes et al. (2016)

Ref. No.	Methods Used	Results	Applications
16	Data acquisition and semantic integration, statistical data analysis, data visualization, data query language, and geospatial data techniques	Data integration and big data analytics solution are discussed	Agriculture decision support system. Helps the policymakers to implement restoration strategies
17	Identifying overlapping units, matching variables, and classification methods	Estimation of the missing data stratum, independent probability of sample infinite population	Agricultural census data analysis of Australia for the year 2015–2016
18	Characteristics of BDA, architecture, technologies, the relationship between value creation and BDA, applications	Big data analytics framework for value creation	Smart city, cybersecurity, agriculture and healthcare domains
19	Summary of Oracle's strategic white paper on Big data applications	Big data analytics as technological innovation in farming	Farming and food production
20	The comprehensive review reveals the impact of big data infarming	Farming is not fully equipped with big data technologies.	Big data analytics helps the farmer in crop management and forecasting

Table 5.
Agriculture sector review summary.

demonstrated a smart city project model using BDI solutions in Brazil [24]. This project proposed a model that can be hosted in big data servers.

Alshawish et al. (2016) discussed the role and potential of BDI solutions in smart cities [25]. The authors have explained the complete process of BDI applications in smart cities.

This study has incorporated some real-world examples of smart city components. **Table 6** presents a summary of highlights from the smart cities review.

2.6 Manufacturing

Ahmed et al. (2016) have proposed a Generating Attributes with Rolled Paths (GARP) algorithm that creates a mining table attributes from multiple data sources [26]. The experiments were carried out on the U.S. consumer electric retailer dataset and revealed that classification accuracy was improved by using GAPR. Bennani et al. (2014) have reported a guided BDI solution with Service Level Agreement (SLA) for querying data from multiple clouds [27]. The methodologies and algorithms designed are applied to energy utilization. Product planning, product design, manufacturing and maintenance process are reviewed in terms of concepts and applications. Qi and Tao (2018) provided a 360-degree review of big data in smart manufacturing [28]. Product planning, product design, manufacturing and maintenance processes are reviewed in terms of concepts and applications. Hufnagel et al. (2015) demonstrated a distributed integration model applicable to the manufacturing industry [29]. This research has created the user-oriented integration platform using a modular approach. O'Donovan et al. (2015) reported a detailed review of BDI implementation in the manufacturing sector. This study has provided a detailed review of big data research in manufacturing [30]. **Table 7** presents a summary of highlights from the manufacturing sector review.

Ref. No.	Methods Used	Results	Applications
21	Various technologies for the handling of big data and IoT integration	A new data architecture that supports IoT and other data resources	Critical data analysis solution for IoT and Big Data
22	Data mixing planner, domain-specific data models, robust type inference, and declarative interface	Automates the data integration process and leverages key capabilities	Website visitors income analysis, retail business analytics
23	Literature survey on prospects, issues and advantages of big data technologies in smart cities	Big data applications for smart use of data and operations in smart cities	Effective management of smart city resources
24	Design of smart city project model using big data in Brazil	This software can be used in big data servers	Software for smart city project in Brazil
25	Collecting data from networks, processing data with various stages and visualization data	Big data-driven smart city improves smart city applications	Smart Energy, Smart public safety and Smart traffic systems.

Table 6.
Smart cities sector review summary.

Ref. No.	Methods Used	Results	Applications
26	Automatic generation of discriminant features, aggregation of information from multiple resources	Classification accuracy improvement and discriminant feature generation. Mitigates the impact of class imbalance	Consumer electronic retailer in Circuit City U.S.
27	The economic model of the cloud referred for lookup, aggregation and correlation in SLA data integration, handling SLA interoperability and collaboration	A distributed data as a service for SLA guided data aggregation framework	Energy consumption applications, data integration of political campaign and electronics
28	Compare and contrast of digital twin and big data. Product planning, product design, manufacturing and maintenance process are reviewed in terms of concepts and applications	Digital twin and big data have great significance in smart manufacturing	Smart manufacturing in workshop or factory
29	Featuring missing connection between successful business integration concept and proven graphical description	User-oriented integration platform using a modular approach	Workflows and product life cycles in the manufacturing industry
30	Captured the status of big data research in manufacturing, and compared the secondary research studies	Usage of big data technologies in manufacturing for maintenance and diagnosis	Various manufacturing domain

Table 7.
Manufacturing sector review summary.

2.7 Review of healthcare sector

Hardiman has explored BDI methodologies for Omics data and network algorithm development [31]. The objective was to channel the gap between phenotype and genotype which were not applied earlier. These researchers used spectrometry permitted geneticists, deep sequencing technologies, biostatisticians and biologists. Bhandari et al. have explained HGBEnviroScreen in their paper [32].

This is an EJ mapping tool providing the key services online to local decision-makers and communities. This study has resulted in multiple risk factors leading to the largest vulnerability census tracts. These risk factors lead to natural disaster, social vulnerability and flooding. Shayne et al. have carried out a comprehensive study of integration solutions for big medical data [33]. This study has covered the applications, tools and technologies of BDI in the healthcare domain. Eftekhari et al. have proposed software as a service architecture [34]. This provides backend infrastructure for database access operations on data from different data sources. This methodology was approved with a proof-of-concept prototype developed on the OpenStack cloud architecture. Vidal et al. have presented a knowledge-driven framework [35]. This framework extracts knowledge from short text and unstructured data.

This framework used controlled vocabularies and ontologies to clarify the extracted entities and relations. Husain et al. have reported SOCR data dashboard design, implementation, and testing. SOCR does exploratory questioning of multi-source and heterogeneous and datasets [36]. **Table 8** presents a summary of highlights from the healthcare sector review.

Ref. No.	Methods Used	Results	Applications
31	Network algorithms and Gene ontology path are followed	Chanel the gap between phenotype and genotype on a scale using high throughput techniques	Biomedical, clinical and Omics data integration
32	Five domains data collected at HGB region for the year 1990 and designed EJ mapping tool for community online services	Online services for decision-makers and community by EJ mapping tool	Usage of result in a community action plan by community partners
33	Usage of various tools, techniques and applications of data integration in the healthcare domain. Analysis of integration techniques abilities to handle speed, variety and uncertainty.	Strength and weaknesses of various solutions, and its findings	Healthcare big data integration
34	Designing Big data store by collecting data from multiple sources. Web interface and RESTful APIs for the integration of RDBMSs with non-relational databases. The queries on such remote databases by proof of concept.	SaaS framework for integrating multiple data sources performing operations such as data access, querying and visualization	Ad-hoc querying of health care datasets
35	Data integration of multiple data resources, Knowledge-driven framework for data description that uses knowledge graph	Ontologies and unified schema as a knowledge graph for describing integrated data	Discovery of interactions among drugs in treatments with much faster running time prescribed to lung cancer patients
36	Human-machine interface for integration of data from heterogeneous resources in a secure and scalable way	Human-machine interactions customization	Service-oriented infrastructure for healthcare data.

Table 8.
Healthcare sector review summary.

2.8 Review of communication sector

Cheng et al. proposed a remote sensing data management system [37]. This system is distributed multisource and followed the MongoDB model. The remote sensing, data integration and access are examined by designing a set of experiments.

Wang et al. have described the major aspects of BDI such as characteristics, advantages, platform architecture, and application areas in telecommunication [38]. This research can be extended by improving multiple levels of protection technologies in the big data platform. Yayah et al. explained a few use cases of machine learning implementation in big data platforms [39]. Scalability and extensibility are the parameters used for the evaluation of BDI technologies. Nwanga et al. studied the impact of big data analytics in mobile phone industry [40]. This study has revealed that BDI solutions and big data analytics has an impact on the growth of the telecommunication industry by adding huge data insights. **Table 9** presents a summary of highlights from the communication sector review.

2.9 Review of supply chain sector

Antonio et al. presented a literature survey of simulation techniques in supply chain risks [41]. The authors have highlighted the significance of BDI in supply chain systems. This analysis has concluded that the problem at hand is simplified without complexity in modeling. This study has complied with industry 4.0 standards. Ostrowski et al. have explored the potential of semantic web technologies by demonstrating a case study in the supply chain [42]. Authors have identified the system for supporting data from multiple sources. This study was carried out by semi-automated mapping using shared domain ontology. Awwad et al. provided a review on applications, advantages and issues of BDI technologies for the supply chain management [43]. The supply chain risk management was carried out using data analytics by making a proactive decision. Lia and Liu have illustrated a data-driven framework for supply chain management [44]. The various circumstances of the supply chain

Ref. No.	Methods Used	Results	Applications
37	Multi-Source BDI framework, Spatial Segmentation Indexing Model, integration based on distributed storage	Scalable storage data integration architecture, latest technical support and development	Professional remote sensing big data
38	Introduced and reviewed major aspects of big data such as characteristics, advantages, platform architecture, and application areas in telecommunication	Internal data applications enhance the efficiency of big data applications. External cooperation provides better services.	Development in telecommunication organization
39	Integrating machine learning tools in the Hadoop platform	Adoption and improvement of big data in the telecommunication industry	Telco, Retail, Financial Services and Energy sector
40	Comprehensive study and analysis of the impact of big data in the mobile industry	Big data analytics has impact on the growth of the telecommunication industry	Growth of the telecommunication industry with help of big data

Table 9.
Communication sector review summary.

Ref. No.	Methods Used	Results	Applications
41	Literature survey of simulation techniques for the analysis and synthesizing risks in supply chains	Analyzed the impact of risks in supply chains	Supply chain management
42	Semantics with annotation in Ontology federation process, Integration of data from multiple resources using shared domain ontology	Integration of data from multiple resources using semi-automated mapping	Supply chain risk detection
43	Detailed review on applications, advantages and issues of big data technologies for supply chain management	Infrastructure and human skillset need to be improved, new and effective techniques need to be developed	Supply chain in manufacturing and logistics
44	Design and development of a data-driven framework for supply chain management	Multiple working modes of big data in supply chain management	Power split device in hybrid vehicles
45	A detailed survey of various supply chain operations reference model with opportunities and challenges.	Studies revealed that the supply chain process is having higher importance	The supply chain and manufacturing products

Table 10.
Supply chain sector review summary.

are accommodated by enabling multiple working modes. Benabdellah et al. discussed the impact of big data on supply chain management [45].

The survey of various supply chain operation reference model presented their applications and challenges. **Table 10** presents a summary of highlights from the supply chain sector review.

2.10 Review of research domain

A review by Li presented BDI technology applications for the analysis of Chinese and Russian dance components with modern features [46]. Arputhamary and Arockiam presented the prominence of BDI by identifying the open problems and the same is extended to proceed with future research in the big data environment [47]. Kadadi et al. have surveyed BDI methods and their interoperability [48]. Authors have also explored its usage in big data setup and the corresponding challenges. Ostrowski and Kim have presented a BDI strategy based on ontology [49]. BDI strategy was implemented in Apache Spark prototyping environment that generates ontology versions using rule-based translations. Sottovia et al. have described the Research Alps project pipeline. This project was funded by the EU Commission [50]. They have created an open dataset providing Alpine area research centre details. Portugal et al. have presented a high-level spatial–temporal architectural framework for massive data integration, analysis and provenance management [51]. This methodology was applied for BDI analysis. **Table 11** presents a summary of the highlights of the research review.

2.11 Review of recent advancements in BDI

Large scale implementation of BDI solutions is a very complex and difficult process than automating data transformation processes. To reduce the complexity, the organizations should implement the procedures for data discovery, semantic or

Ref. No.	Methods Used	Results	Applications
46	Big Data Technology	Combination of Chinese and Russian cultural and modern features	Dance elements
47	Importance of BDI issues and challenges are identified	The existing techniques and approaches are inefficient to handle the problems.	Possible research directions
48	Addressing challenges of BDI such as Data accommodation, Data irregularity, Query optimization, Extensibility, ETL processing	Big data integration architecture	BDI within the organization and inter organizations
49	Ontology-based data integration, creation of new ontology versions by using rule-based translation	Multiple data sources 'Semi-automated mapping	Large scale Big data applications
50	M-STEP and entity matching method and functional framework to deal with hierarchical data instances	Open dataset providing Alpine area research centres details	Research Alps project funded by EU commission
51	Domain experts focusing on appropriate analysis steps, high-level models linked with code produces middleware	Model-driven techniques resulting in data integration and analysis	Provenance information

Table 11.
Research domain review summary.

business comprehension of data, metadata management, structured and unstructured data management, and transformation. Integrating unstructured and semi-structured data enables organizations to manage modern data sources containing text, images, and video. A survey was conducted by AtScale Inc. in collaboration with Cloudera and ODPi.org reveals that most of the organizations are selecting multi-cloud strategies for BDI implementation. Data virtualization and data governance are their top priorities [52]. This survey has collected data from 150 data practitioners where the respondents are from multiple industries around the world. The online magazine "Smarter with Gartner" has reported that top ten technology trends in data analytics require essential investments [53].

This article revealed that the combination of machine learning algorithms and data technologies could help the medical and public health experts to discover new possible treatments. The article entitled "2020 CRN Big Data 100" published in "Data Integration Solutions Review" enlists the emerging big data tool vendors [54]. This list provides the details of data integration software, tools, platforms and vendors. A data-driven technique for a hybrid BDI using multilayer perceptron was discussed in this research [55]. A customized multilayer perceptron model was constructed using time-based parameters. The fields applied in optimization analysis are also used in the error matrix through additional neural network model. Research results revealed that this solution captures the variations in state variables. BDI project implementation for COVID-19 analytics was discussed [56]. This project was funded by the European Union research fund. This platform combines information from multiple sources such as world news, social media, published science and health data from healthcare institutions. The project design was co-created with industry, academia, health professionals, and policymakers to align with innovative technologies. This project successfully provides useful and actionable information to public health authorities.

Authors	**Year**	**Study Objectives**	**Advantages**	**Disadvantages**	**1**	**2**	**3**	**4**	**5**	**6**
Fikri et al	2019	To get solutions for financial data real-time integration issues and interpretation of data	This real-time data integration solution resolves earlier issues of classic ETL tools	This solution cannot be integrated with a hot production setting	Y	Y	Y	Y	N	Y
Cheng et al	2020	To design BDI distributed architecture for remote sensing data, where these data from multiple sources	Improvement in performance with a distributed architecture for remote sensing data	Time and resource complexity to handle various pre-processing steps in data integration	Y	Y	Y	N	N	Y
Authors	Year	Study Objectives	Advantages	Disadvantages	1	2	3	4	5	6
Bhandari et al	2020	To develop EJ screening, an adaptable and community-based tool for the region Houston Galveston Brazoria	Risk factor identification and understanding among the communities	reducing environmental disparities and improving their health and well-being	Y	Y	N	N	N	Y
Vieira et al	2020	To conduct a literature survey of simulation methods used for handling risks in the supply chain with an emphasis on data integration	Simplification of the problem in the absence of complex modeling	It is required to focus on supply chain real cases	N	Y	Y	Y	N	Y
Stonebraker et al	2018	To explore issues of BDI related to scalability in enterprises at Tamr region	Automation by machine learning and rule-based approach for augmenting	Involves high cost for domain experts, shortage of training data	N	Y	Y	Y	N	Y
Ahmed et al	2016	To aggregate data from local and external resources, to generate mining table from these, automatic generation of potential discriminant features	Classification accuracy improvement and thus mitigates the impact of class imbalance	Time complexity is linear and it is required to reduce computation time with an efficient method	Y	Y	N	Y	N	Y

Authors	Year	Study Objectives	Advantages	Disadvantages	1	2	3	4	5	6
Bansal	2014	BDI by designing a Semantic Extract-Transform-Load architecture	Publishing semantic data on the internet and thus contribute to the web of data	It is required to understand the heterogeneity of data i.e. ontology engineering	Y	Y	Y	N	N	Y
Dhayne et al	2019	To study healthcare data integration methods, tools, and applications	Wide range of healthcare data integration concepts, techniques and tools are covered	Data integration in the healthcare sector could not be done efficiently using traditional way	Y	Y	Y	Y	Y	Y
Sazontev et al	2019	To develop a prototype of a big data integration system	Useful for e-commerce data integration domain	Lacks in methods for schema alignment	Y	Y	N	N	N	Y
Chen et al	2015	To accomplish data integration of back-end datasets in a complete manner	Data movement is faster than that of Spark thus achieved optimization	Integration of more Spark modules is not supported	Y	Y	N	N	N	Y
Zheng et al	2015	To summarize categories and its subcategories of data integration techniques	Extensive details of big data integration solutions for communities	Since BDI methods behave differently in different applications it's difficult to select the best data fusion technique	Y	Y	Y	Y	N	Y
Huang et al.	2014	To automate the data integration process and leverage key capabilities	A more agile process for compelled analysis by generating a subset of data	HiperFuse modules are implemented separately yet to be integrated	Y	Y	N	N	N	Y
Portugal et al	2016	To perform spatial and temporal data analysis for assisting domain experts	High-level representations by domain specific languages, data analysis and integration by model-driven techniques	provenance technologies need to be used in related spatial–temporal approaches	Y	Y	Y	N	N	Y

Authors	Year	Study Objectives	Advantages	Disadvantages	1	2	3	4	5	6
Saggi et al	2018	To bridge the gap by big data processing and analytics	A comprehensive review of big data projects in terms of analytics, management, and machine learning	It is required to carry out empirical research based on qualitative and quantitative methods	Y	Y	N	Y	N	Y
Kim et al	2020	Survey sample data approach to handle big data integration	Recognition of overlapping units and correction of misclassification errors	Statistical inference variance estimation with non-parametric propensity score tuning is not covered	Y	Y	N	Y	N	Y

Table 12.
Comparative Analysis of curated survey papers with specific parameters.

2.12 Comparative analysis of survey papers with specific parameters

Authors have selected fifteen BDI survey papers for comparative study. The specific feature parameters such as 1. Architecture, 2. Applications, 3. Open Issues and Challenges 4. Taxonomy, 5. Security, 6. Future Directions are used for comparing these survey papers. The Comparative analysis results are shown in **Table 12**.

3. The architecture of the BDI ecosystem

The outline architecture of the BDI ecosystem is shown in **Figure 5**. This architecture has four major components. These components are Data Sources, Data Operations, Virtual Databases and Business Intelligence. This architecture also shows the operations performed by each of these components. The business Big data would be collected from various distributed sources in different formats and sizes in Data Sources component. The Data Operations component shows the different operations which are performed on this heterogeneous Big data.

The Big data gathered from various types of physically distributed databases are integrated to form a unified logical virtual database. Business intelligence information is extracted from this virtual data source by performing the operations stated in Business Intelligence Component. This intelligent information would be used for real-time intelligent business decision-making process across the organization.

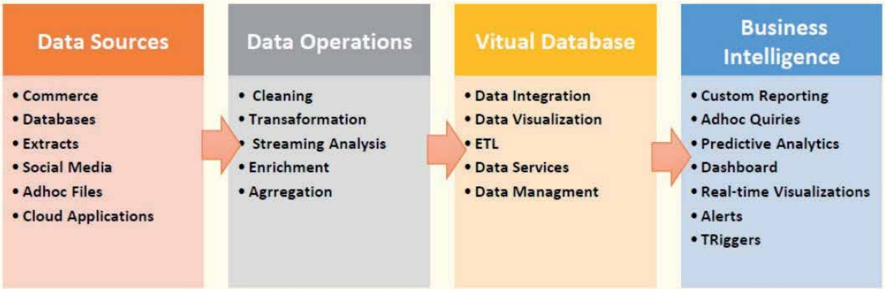


Figure 5.
The architecture of the BDI ecosystem.

4. BDI research issues, challenges and future directions

The research issues, challenges and future directions related to BDI implementation are discussed in the following sections.

4.1 BDI research issues

1. Scalability - Scalable architectures for parallel big data processing

2. Real-time big data analytics - Stream big data processing of text, image, and video

3. Deployment of the IoMT, IoT and CCTVs systems in smart environments would capture big data continuously. Processing multimedia big data in real-time with low latency and high accuracy

4. The balancing of big data processing load at the edges and distributed to the hybrid cloud securely
5. Implementing real-time, complex big data analytics in the cloud by reducing the cost of operations
6. Ensuring authorization, authentication, security and privacy at the edges and cloud.
7. Efficient storage and transfer of big data in real-time
8. Efficient modeling of uncertainty with unlabeled big data
9. Management of graphical big databases
10. Social media analytics using efficient graphical processing.
11. Quantum computing for big data analytics
12. Building context-sensitive large scale systems

4.2 BDI challenges

1. Extracting actionable information from BDI solutions
2. Synchronization of data across heterogeneous data sources
3. Lack of comprehension and management of uncertainty
4. Effective anonymization of sensitive fields in the largescale data systems
5. Support for scalable privacy preservation during BDI processing
6. Generating process models that learn with a smaller number of data samples
7. Building context-sensitive large scale systems:
8. BDI Talent shortage

4.3 BDI trends

1. Everyone is adopting Software as a Service (SAAS)
2. Self-service has evolved to self-sufficiency
3. Shared data, visualizations and storytelling are consumed by all
4. Now constant updating of business-ready data is very vitaa
5. Support for advanced analytics with different perspectives
6. It is critical to gather and create alternative big data
7. Every business is undergoing re-engineering process

8. The measures for competition, surveillance and security are constantly redefined

9. Collaboration has to coalesce earlier in the chain

10. The great digital switch may force a generational shift in analytics.

4.4 BDI advantages

1. Improved e-commerce sales and operations efficiency

2. Creating efficient marketing strategies

3. Increased security enforcement

4. Improving fraud prevention;

5. Enhancing user experience

6. Increased profits

5. BDI organizational case studies

Authors have presented a set of real-life case studies of BDI solutions implemented successfully across the business domains in organizations. **Figure 6** shows the domains and tolls used in that domain for illustration.

5.1 Walgreens Boots Alliance Company

It is a global leader in retail and wholesale pharmacy business operating in the U.S. and Europe and has more than 170 successful business years of serving humanity. Walgreens Boots Alliance, Inc., declared its IT collaboration with Microsoft and Adobe to introduce a world's best digital platform for enhanced customer experience and data insights to offer truly customized healthcare, adhere to their healthcare plans and shopping services as stated by their global chief marketing officer Vineet Mehra. The BDI systems can manage 7.5 billion medical transactions 100 million citizens providing a singular, unified view of the customer information about demographics, registration, diagnoses, procedures, and data from managed-care plans.

This BDI digital platform helps the customers to access key services of pharmacy, beauty and other categories on daily basis. Data security and privacy are important principles in the design of Microsoft's trusted cloud platform. Walgreens has introduced personally customized prescription understanding for patients at Walgreens, Dynamics 365 Customer Insights would serve as WBA's Customer Data Platform (CDP) provided by Microsoft. CDP provides a unified, 360-degree perceptions of the customer and reveals the details to leverage personal experience. Adobe's Customer Experience Management (CXM) solutions leverage Walgreens to offer supreme customer experience, with end-to-end platform for analysis, managing content, customization, campaign composition and many more. Walgreens Company also extends collaboration with Tata Consultancy Services to build highly scalable, maintainable and world-class unified IT operating platform to enable digital transformation, innovation and automation of services offerings. Walgreens Boots Alliance also collaborates with Hortonworks to offer excellent customer satisfaction.

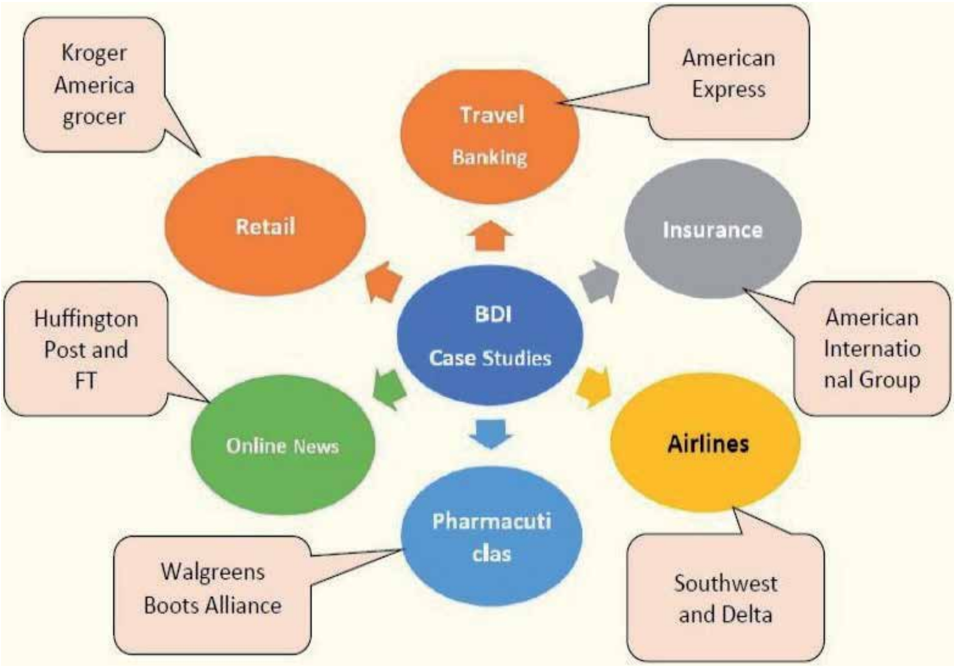


Figure 6.
Case study domains.

5.2 The American International Group (AIG)

AIG Data Services Pvt. Ltd. is a 100% owned subsidiary of American International Group Inc. It is a Fortune 500 company with revenues of the US $70 billion. AIG drives the best decision-making through BDI solution sutilizing business and customer big data across 130+ countries and 64,000 employees which is ever-growing. AIG has implemented sophisticated prediction models with 115 variables to analyze the past business transactions to forecast the potential trends. AIG identified 24% accounts in the Australian market that are about close in next four-month time. AIG has applied BDI tools and visualization systems to discover the frauds by detecting the false claims and adjuster handwritten notes to detect probable frauds. These tools offer insights into insurance claims and enhance machine learning algorithms. AIG creates data profiles and assesses vital data elements against pre-defined data quality standards on important business data for important applications. Today big data is distributed across the globe and facts are available across multiple sources. The team responsible for data sourcing uses ETL tools to provide a unified virtual version of these facts collected from various data sources. AIG has implemented Netezza and data virtualization technologies on Cisco Information Server. AIG also utilizes Hadoop, R, Python, SAS and other open-sourced/licensed tools to implement BDI solutions to beat the competitors. AIG uses tools such as QlikView, Tableau, Cognos and Micro strategy for data visualization.

5.3 Kroger - America's grocer company

Kroger has nearly 2,800 stores in 35 states under twenty-four banners with and annual sales exceeding 121.1 billion. Kroger today ranks as one of the world's largest retailers. Kroger with its joint venture with Dunnhumby is leveraging BDI solutions. Dunnhumby is a technology solutions provider company for retail industry.

These solutions are implemented using the latest techniques, algorithms, procedures and applications. The Kroger company gathers and processes the data from about 770 million consumers. Kroger has implemented BDI solutions for extracting more actionable information for profitability, customer loyalty. Kroger claims that 95% of sales are from the loyalty card. Kroger achieved about $12 billion in revenues by BDI implementing and analytics solutions since 2005.

5.4 Southwest and Delta Airline company

This company has encashed on customer loyalty and relationships by providing boundless service through social channels and other data exchange mechanisms. Southwest utilizes speech analytics to help and enhance the exchanges between service professionals and customers. Southwest applied BDI solutions to understand customer online behavior and activities, increasing offers for customers and driving growth in customer satisfaction year after year. Delta has applied BDI solutions to support most painful travel condition that results in lost baggage. This company tracks the data about baggage and became the first airline company to permit customers to trace their baggage from smartphones. This company checks about 130 million baggage every. Delta is branding its self as a customer friendly services by permitting customers to download their apps over 11 million times and provides best customers with baggage secure services company.

5.5 Huffington Post and FT is an online news service company

This company has become number one online news site in the United States. According to this report, the company's leadership believes in running the business based on big data. This involves enhancing the user experience in real-time through recommendations, moderation, social trends, and personalization. This company optimizes its portal in many ways, and its analytics platform powers the entire analytical process. Huffington Post utilizes data to comprehend and serve the customers well, make targeted advertising, and design innovative products based on information gathered. Their CEO informed that BDI solutions have transformed its business by intelligent and real-time decision making. This company utilized many data points to enhance relevance in their communications, analyze customer content preferences, and personalize the content all to keep traffic and visitors always. The BDI also benefits the company to comprehend the time of day consumption based on both mobile channels and PC.

6. Results discussion

This chapter reviewed the literature on BDI tools and applications in diverse industries and presented the highlights from each domain. All most all organizations are gathering a huge quantity of big data in real-time, Online and offline modes. Managing, real-time processing this big data to extract useful business information for intelligent decision making is the real challenge. The big data processing systems are empowered by big data integration and analytics platforms. BDI systems are facing the challenges in integrating and synchronization of heterogeneous big data from multiple distributed sources. The lack of comprehension and management of uncertainty in big data is another challenge faced in the big data processing. BDI processing should ensure context-sensitivity and extracting the semantics in the distributed data processing. Research on designing effective machine and deep learning algorithms is going on in the BDI domain. BDI Processing uses Hadoop environment with HDFS, Spark computation model with a Hive database as a distributed data warehouse. The use of Apache Spark enhances short time frames for quality and availability reporting. BDI processing involves data acquisition, semantic integration, statistical data analysis, data visualization, data query language, geospatial data techniques. Big data analytics framework enables us to create business value.

The economic model of the cloud promotes BDI processing by providing online services for decision-makers and the business community. Usage of various tools, techniques and applications of BDI leverages the ability to handle speed, variety and uncertainty. Knowledge-driven framework for BDI describes a knowledge graph. Human-machine interface for BDI integrates data from heterogeneous resources in a secure and scalable way. Ontologies and unified schema as a knowledge graph for describing integrated data. Multi-Source BDI is a framework for integrating data in the distributed storage environment. Authors have discussed BDI research issues and challenge data accommodation, data irregularity, query optimization, extensibility, ETL processing. Remote sensing data. Real-time big data analytics processes stream of big text, image, and videos generated from IoMT, IoT and CCTVs systems. Implementing real-time, complex big data analytics in the cloud using BDI process reduces the cost of operations. This paper discussed five case studies of BDI applications implemeneted in the in world-class organizations.

7. Conclusion

This chapter discussed the importance of BDI process implemented in diverse organizations for providing valuable insights into business data. These insights into the data enable the manager to take intelligent and well-informed rational decisions. An extensive study of literature on BDI applications deployed in diverse domains across the world was carried out and highlights are discussed. The intelligent and autonomous BDI systems are designed using AI, Blockchain, Big data, 5G, Fog and cloud technologies. The comparative analysis of specific parameters was carriedout on curated to survey papers to identify the research gaps and future opportunities in the BDI domain. The five case studies from fortune 500 companies have discussed the insights about how BDI is empowering business decision making leveraging quality, trust, security, flexibility, efficiency and also reduce the cost of operations. The authors attempted to provide a holistic view of BDI concepts and applications. Authors concluded that BDI plays a vital role in the diverse organizations at present and in near future also.

Author details

Sreekantha Desai Karanam[1]*, Rajani Sudhir Kamath[2], Raja Vittal Rao Kulkarni[2] and Bantwal Hebbal Sinakatte Karthik Pai[1]

1 NMAM Institute of Technology, Nitte, Karnataka, India

2 CSIBER, Kolhapur, Maharastra, India

*Address all correspondence to: sreekantha@nitte.edu.in

IntechOpen

References

[1] Hasliza, N., Hassana, M., Ahmada, K. & Salehuddina, H. (2020). Diagnosing the Issues and Challenges in Data Integration Implementation in Public Sector, International Journal Advanced Science Engineering Information Technology, 10(2).

[2] Zhang, Y. (2020). The Integration of Professional Ethics of Modern Etiquette Students under the Background of Big Data, Journal of Physics: Conference Series 1574.

[3] Bansal, S. K. (2014). Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration, IEEE International Congress on Big Data, 978-1-4799-5057-7/14 © 2014 IEEE, DOI 10.1109/BigData.Congress.2014.82

[4] Zheng, Y. (2015). Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions On Big Data, 1(1).

[5] Munné R. (2016). Big Data in the Public Sector. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3_11.

[6] Camargo-Perez, J. A., Puentes-Velasquez, A. M., & Sanchez-Perilla, A. L. (2019). Integration of big data in small and medium organizations: Business intelligence and cloud computing, J. Phys.: Conf. Ser. 1388 012029.

[7] Stonebraker, M., & Ilyas, I. F. (2018). Data Integration: The Current Status and the Way Forward, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

[8] Sazontev, V., & Stupnikov, S. (2019). An Extensible Approach for Materialized Big Data Integration in Distributed Computation Environments, Ivannikov Memorial Workshop (IVMEM), 978-1-7281-4623-2/19/ ©2019 IEEE DOI 10.1109/IVMEM.2019.00011

[9] Alsghaier, H., Akour, M., Shehabat, I., & Aldiabat, S. (2017). The Importance of Big Data Analytics in Business: A Case Study. American Journal of Software Engineering and Applications, 6(4), 111-115.

[10] Alam, J. R., Sajid, A., Talib, R., & Niaz, M. (2014). A Review on the Role of Big Data in Business. International Journal of Computer Science and Mobile Computing, 3(4), 446-453.

[11] Fikri, N., Rida, M., Abghour, N., Moussaid, K., & Omri, A. I. (2019). An adaptive and real-time based architecture for financial data integration. Journal of Big Data, 6(97).

[12] Bucea-Manea-Tonis, R. (2018). Deductive systems for Big data integration, Journal of Economic Development, Environment and People, 7(1).

[13] Chen, W., Wang, R., Wu, R., Tang, L., & Fan, J. (2016). Multi-source and Heterogeneous Data Integration Model for Big Data Analytics in Power DCS [Paper Presentation]. International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

[14] Hussain K., Prieto E. (2016). Big Data in the Finance and Insurance Sectors. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3_12

[15] Avi V., Kamaruddin S. (2017). Big Data Analytics Enabled Smart Financial Services: Opportunities and

Challenges. In: Reddy P., Sureka A., Chakravarthy S., Bhalla S. (eds) Big Data Analytics. BDA 2017. Lecture Notes in Computer Science, vol 10721. Springer, Cham. https://doi.org/10.1007/978-3-319-72413-3_2

[16] Nabrzyski, J., Liu, C., Vardaman, C., Gesing, S., & Budhatoki, M. (2014). Agriculture Data for All - Integrated Tools for Agriculture Data Integration, Analytics and Sharing. IEEE International Congress on Big Data. 978-1-4799-5057-7/14 © 2014 IEEE DOI 10.1109/BigData.Congress.2014.117

[17] Kim, J. K., & Tam, S. (2020). Data integration by combining big data and survey sample data for finite population inference. arXiv:2003.12156v3.

[18] Saggi, M. K., & Jain, S. (2018). A survey towards the integration of big data analytics to big insights for value-creation. Information Processing & Management, 54.

[19] Ribarics, P. (2016). Big Data and its impact on agriculture. Eco cycles, 2(1), 33-34.

[20] Sarker, M. N., Islam, M. S., Murmu, H., & Rozario, E. (2020). Role of Big Data on Digital Farming. International Journal of Scientific & Technology Research, 9(04).

[21] Kaur, H., & Kushwaha, A. S. (2018). A Review on Integration of Big Data and IoT. 4th International Conference on Computing Sciences. 978-1-5386-8025-4/18/$31.00 ©2018 IEEE DOI 10.1109/ICCS.2018.00040.

[22] Huang, E., Quiroz, A., & Ceriani, L. (2014). Automating Data Integration with HiperFuse [Paper Presentation] 2014 IEEE International Conference on Big Data

[23] Nuaimi, E. A., Neyadi, H. A., Mohamed, N., & Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6(25).

[24] Gomes, E., Dantas, M. A., Macedo, D. D., Rolt, C. D., Brocardo, M. L., & Foschini, L. (2016). Towards an Infrastructure to Support Big Data for a Smart City Project [Paper Presentation]. 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, 2016, pp. 107-112, DOI: 10.1109/WETICE.2016.31

[25] Alshawish, r. A., Alfagih, S. M., & Musbah, M. S. (2016). Big data applications in smart cities. 2016 International Conference on Engineering & MIS (ICE), Agadir, 2016, pp. 1-7, DOI: 10.1109/ICEMIS.2016.7745338

[26] Ahmed, F., Samorani, M., Bellinger, C., & Zaiane, O. R. (2016). Advantage of Integration in BigData: Feature Generation in Multi-Relational Databases for Imbalanced Learning, 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005-7/16/$31.00 ©2016 IEEE

[27] Bennani, N., Ghedira-Guegan, C., Musicante, M. A., & Vargas-Solar, G. (2014). SLA-Guided Data Integration on Cloud Environments [Paper Presentation]. 2014 IEEE International Conference on Cloud Computing, Alaska, United States. 934-935.

[28] 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.

[29] Hufnagel, J., & Vogel-Heuser, B. (2015). Data integration in manufacturing industry: Model-based integration of data distributed from ERP to PLC [Paper Presentation]. 2015 IEEE 13th International Conference on Industrial Informatics (INDIN),

Cambridge, 2015, pp. 275-281, DOI: 10.1109/INDIN.2015.7281747.

[30] O'Donovan, P., Leahy, K., Bruton, K., & T. J. O'Sullivan. (2015). Journal of Big Data, 2(20). DOI 10.1186/s40537-015-0028-x

[31] Hardiman, G. (2020). An Introduction to Systems Analytics and Integration of Big Omics Data, Genes, 11(245).

[32] Bhandari, S., Lewis, P., Craft, E., Marvel, s. W., Reif, D. M., & Chiu, W. A. (2020). HGBEnviroScreen: Enabling Community Action through Data Integration in the Houston–Galveston–Brazoria Region, Int J Environ Res Public Health, 17(4): 1130.

[33] Dhayne, H., Haque, R., Kilany, R., & Taher, Y. (2019). In Search of Big Medical Data Integration Solutions - A Comprehensive Survey. IEEE Access, 7.

[34] Eftekhari, A., Zulkernine, F., & Martin, P. (2016). BINARY: A Framework for Big Data Integration for Ad-hoc Querying, 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005-7/16/©2016 IEEE

[35] Vidal, M., & Sakor, A. (2019). Semantic Data Integration Techniques for Transforming Big Biomedical Data into Actionable Knowledge, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS).

[36] Husain, S., Kalinin, A., Truong, A., & Dinov, D. (2015). SOCR Data Dashboard: An integrated Big Data archive mashing Medicare, labour, census and econometric information. Journal of Big Data, 2(13).

[37] Cheng, Y., Zhou, K., Wang, J., & Yan, J. (2020). Big Earth Observation Data Integration in Remote Sensing Based on a Distributed Spatial Framework. Remote Sens. 12, 972.

[38] Wang, Z., Wei, G., Zhan, Y., & Sun, Y. (2017). Big data in telecommunication operators: data, platform and practices. Journal of Communications and Information Networks, 2(3). DOI: 10.1007/s41650-017-0010-1

[39] Yayah, F. C., Ghauth, K. I., & Ting, C. (2017). Adopting Big Data Analytics Strategy in the Telecommunication Industry. Journal of Computer Science & Computational Mathematics. 7(3). DOI: 10.20967/jcscm.2017.03.002

[40] Nwanga, M. E., Onwuka, E. N., Aibinu, A. M., & Ubadike, O. C. (2015). Impact of Big Data Analytics to the Nigerian Mobile Phone Industry. Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE), March 3-5, 2015.

[41] Antonio, A. C., Luis, M. S., Santos, M. Y., Guilherme, A. B., & Jose, A. O. (2020). Supply chain data integration: A literature review. Journal of Industrial Information Integration 19 100161.

[42] Ostrowski, D., & Kim, M. (2017). Semantic-Based Framework for Big Data Integration [Paper Presentation]. 2017 IEEE 11th International Conference on Semantic Computing

[43] Awwad, M., Kulkarni, P., Bapna, R., & Marathe, A. (2018). Big Data Analytics in Supply Chain: A Literature Review. Proceedings of the International Conference on Industrial Engineering and Operations Management, Washington DC, USA, September 27-29.

[44] Lia, Q., Liu, A. (2019). Big Data-Driven Supply Chain Management, Procedia CIRP 81 ScienceDirect 52nd CIRP Conference on Manufacturing Systems, 1089-1094.

[45] Benabdellah, A. C., Benghabrit, A., Bouhaddou, I., & Zemmouri, E. M. (2016). Big Data for Supply Chain

Management: Opportunities and Challenges. International Journal of Scientific & Engineering Research, 7(11).

[46] Li, J. (2020). Research on the Integration of Chinese and Russian Original Ecological Dance Elements and Modern Elements Based on Computer Big Data Analysis. Journal of Physics: Conference Series 1578.

[47] Arputhamary, B. & Arockiam, L. (2015). Data Integration in Big Data Environment. Bonfring International Journal of Data Mining, 5(1), 1-5.

[48] Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of Data Integration and Interoperability in Big Data. 2014 IEEE International Conference on Big Data, 978-1-4799-5666-1/14/$31.00 ©2014 IEEE

[49] Ostrowski, D., Rychtyckyj, N., MacNeille, P., Kim, M. (2016). Integration of Big Data Using Semantic Web Technologies. 2016 IEEE Tenth International Conference on Semantic Computing, 978-1-5090-0662-5/16 © 2016 IEEE DOI 10.1109/ICSC.2016.101

[50] Sottovia, P., Paganelli, M., Guerra, F., & Vincini, M. (2019). Big Data Integration of Heterogeneous Data Sources: the Research Alps CaseStudy. 2019 IEEE International Congress on Big Data (BigData Congress), 978-1-7281-2772-9/19 ©2019 IEEE DOI 10.1109/BigDataCongress.2019.00027

[51] Portugal, I., David, P. A., & Cowan, D. (2016). Towards a Provenance-Aware Spatial-Temporal Architectural Framework for Massive Data Integration and Analysis, 2016 IEEE International Conference on Big Data (Big Data).

[52] AtScale Inc, Big Data & Analytics Maturity 2020 Survey Results, https://www.atscale.com/wp-content/uploads/2020/02/2020-Big-Data-Analytics-Survey-Results.pdf

[53] Laurence Goasduff, Gartner Top 10 Trends in Data and Analytics for 2020, https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020/ posted on 19th October 2020 and retrieved on 24th December 2020

[54] Timothy King, Data Integration Solutions News, 2020 CRN Big Data 100: 14 Data Integration Tools Companies to Consider, https://solutionsreview.com/data-integration/2020-crn-big-data-100-data-integration-companies-to-consider/ posted on 30th April 2020 and retrieved on 24th December 2020

[55] Lilan Huang, Hongze Leng, Xiaoyong Li, Kaijun Ren, Junqiang Song, Dongzi Wang, A Data-Driven Method for Hybrid Data Assimilation with Multilayer Perceptron, Big Data Research 23 (2021) 10017, https://doi.org/10.1016/j.bdr.2020.100179

[56] Joao Pita Costa, Marko Grobelnik, Flavio Fuart, and Luka Stopar, Meaningful Big Data Integration for a Global COVID-19 Strategy, IEEE Computational Intelligence Magazine, November 2020

Section 4

Data Quality

Chapter 5

Quality of Information within Internet of Things Data

Tomás Alcañiz, Aurora González-Vidal, Alfonso P. Ramallo and Antonio F. Skarmeta

Abstract

Due to the increasing number of IoT devices, the amount of data gathered nowadays is rather large and continuously growing. The availability of new sensors presented in IoT devices and open data platforms provides new possibilities for innovative applications and use-cases. However, the dependence on data for the provision of services creates the necessity of assuring the quality of data to ensure the viability of the services. In order to support the evaluation of the valuable information, this chapter shows the development of a series of metrics that have been defined as indicators of the quality of data in a quantifiable, fast, reliable, and human-understandable way. The metrics are based on sound statistical indicators. Statistical analysis, machine learning algorithms, and contextual information are some of the methods to create quality indicators. The developed framework is also suitable for deciding between different datasets that hold similar information, since until now with no way of rapidly discovering which one is best in terms of quality had been developed. These metrics have been applied to real scenarios which have been smart parking and environmental sensing for smart buildings, and in both cases, the methods have been representative for the quality of the data.

Keywords: IoT, QoI, outliers, interpolation, data quality, data integrity

1. Introduction

The emergence of Internet of Things (IoT) deployments has allowed millions of connected, communicating, and exchanging objects to be embedded seamlessly around the world, generating large amounts of data through sensor monitoring on a timely basis.

The data flow between the physical and the digital world through artificial intelligence can expand the computer's awareness of the surrounding environment, thereby obtaining the ability to act on behalf of humans through ubiquitous services.

In this IoT-based environment, the basis for making wise decisions and providing services is the data collected by sensors and actuators. If the data quality is poor, these automated decisions may be incorrect, ranging from sensor failure to deliberately providing false information with malicious intent. Data quality (DQ) is therefore needed to attract users to participate and accept IoT paradigms and services.

Data Quality refers to how well data meet the requirements of data consumers [1]. In a similar manner, Quality of Information (QoI) relates to the ability to judge whether information is adequate for a particular purpose [2, 3].

From such a well-known and accepted definition, we understand that it refers to a perception or an evaluation of the suitability of the data to fulfill its purpose in a given context, subject to the requirements of the consumer. On the literature, the quality of the data is determined by factors such as availability, usability, reliability accuracy, completeness, relevance, and novelty [4].

According to [5], ensuring data quality is crucial when deploying and leveraging devices, given that:

- Decision-making is only possible if the data available are correct and appropriate.

- Serious problems are practically unapproachable without an adequate data source.

The way to tackle this problem is through the use of so-called data quality metrics which are calculated in order to validate the Quality of the Information (QoI).

The aim of this chapter is to define some metrics for DQ and calculate them in IoT scenarios in order to test their viability.

2. Data integrity

Data integrity refers to the accuracy and reliability of data. The data must be complete, without variations or compromises from the original, which is considered reliable and accurate. Therefore, this term is closely related to the quality of the data and this in turn to the quality metrics [6].

There are several types of data integrity [7]:

- Physical integrity is the protection of the integrity and accuracy of data as they are stored and extracted. That is, it is related to the physical layer of the systems. In the context of IoT, a physical integrity problem comes from the physical degradation of the sensors, whether due to a breakdown or sabotage.

- Logical integrity preserves the data without any change, since it is used differently in a relational database. Logical integrity protects data from human errors and also from hackers, but in a very different way than physical integrity.

- The integrity of the entity is based on the creation of primary keys, or unique values, that identify data to ensure that it is not listed more than once and that there is no field in a table considered null. It is a feature of relational systems that store data in tables that can be linked and used in very different ways. In an IoT scenario, an entity integrity problem can arise in case of a sensor failure which produces redundant measurements or by a human failure in which two different sensors are assigned the same identifier, which produces redundancies in databases.

- Referential integrity is a series of processes that ensure that data is stored and used consistently. The rules built into the database structure about how foreign keys are used to ensure that only appropriate data changes, additions, or deletions occur.

- Domain integrity is the set of processes that guarantee the veracity of each data in a domain. In this context, a domain is a set of acceptable values that a column can contain. You can incorporate restrictions and other measures that limit the format, type, and amount of data entered. Due to an error in the IoT devices, one of them could be entering data that does not correspond to the correct type in a column of a database, such as saving a number when a date should be saving or a date in a format that is not adequate.

- User-defined integrity comprises the rules and constraints created by the user to suit their particular needs. Sometimes entity, referential, and domain integrity are not enough to safeguard data. Often times, specific corporate rules need to be considered and incorporated into measures regarding data integrity. In an IoT scenario, a sensor may be giving acceptable values, that is, that they respect the rest of the integrity criteria, however, it may not be meeting a necessary criterion for the correct functioning of the system, such as a sensor that collects percentage values and that you are receiving a value greater than 100.

In this section, Data Integrity has been defined, however, it is necessary to note what is the difference between this term and the term Data Quality. Data quality is related to the reliability of the information, which is necessary for planning and decision making for a specific operation. Whereas, the integrity of the data guarantees the reliability of the data in physical and logical terms.

3. Data quality metrics

In this section, we describe the metrics that have been defined to calculate and annotate the QoI for IoT data. Those were previously described on [8].

3.1 QoI basic metrics

The first set of metrics is based on a descriptive analysis. This approach was also used on the IoTCrawler framework [9]. It proposes to integrate quality measures and analysis modules to rate data sources to identify the best fitting data sources to get the needed information. The first step before implementing some quality analysis modules is to identify quality measures, which can be used to rate data sources and the delivered/produced data for their Quality of Information. To measure the QoI, we propose to use the so-called QoI Vector, which is defined in Eq. (1) and gathers the information belonging to all the metrics proposed in this framework

$$\vec{Q} = \left\langle q_{cmp}, q_{tim}, q_{pla}, q_{art}, q_{con} \right\rangle \tag{1}$$

The elements of the vector are defined as follows:

- Completeness (q_{cmp}): it represents the percentage of missing or the unusable data.

$$q_{cmp} = 1 - \frac{M_{miss}}{M_{exp}} \tag{2}$$

where M_{miss} is the sum of missing values and M_{exp} is the sum of expected values of an incoming dataset.

- Timeliness (q_{tim}): refers to the expected time of accessibility and availability of information. In other words, it represents how long is the time difference between the data capture and the reality event happening. It is crucial in critical IoT applications such as traffic safety. Its definition is:

$$q_{tim} = 1 - \frac{T_{age}}{W} \tag{3}$$

where T_{age} is the difference between the expected time and the time taken by the sensor $(T_{real} - T_{exp})$, and W is the proper time of the system, which is chosen arbitrarily.

- Plausibility (q_{pla}): shows if received data is coherent according to the probabilistic knowledge of the variables that are being measured. Sensor annotations or meta-data are used to determine an expected value range of an incoming measurement.

$$q_{pla} = \prod P_{Annotations}(\nu) \tag{4}$$

The range of Plausibility value is defined between 0 and 1.

- Artificiality (q_{art}): this metric determines the inverse degree of the used sensor fusion techniques and defines if this is a direct measurement of a singular sensor, an aggregated sensor value of multiple sources or an artificially interpolated value.

- Concordance (q_{conc}): describes the agreement between information of the data source and the information of other independent data sources, which report correlating effects. The Concordance analysis takes any given sensor x_0 and computes the individual concordances, $C(x_0, x_i)$, with a finite set of n sensors $(i = 1, \ldots, n)$.

$$q_{con}(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) c(x_0, x_i) \tag{5}$$

with λ as a weight-function

$$\lambda_i(x_0) = \frac{1}{d(x_0, x_i)} \tag{6}$$

And $d(x_a, x_b)$ propagation and infrastructure-based distance function between sensor location x_a and x_b or sensors a and b.

All the metrics exposed in this section take values between 0 and 1, with the value 1 being the ideal case in which the quality of the data is maximum and 0 the opposite case.

These metrics represent the simplest ones that can be calculated in this kind of IoT scenarios. However, it is possible to go further and compute some metrics that give us a deeper knowledge of the IoT system.

3.2 Oultier-based metrics using heuristics

Since these metrics provide us basic information, it is possible to go further and obtain a series of metrics that can be useful. These new metrics come from the hand of Machine Learning (ML), in this case the search for outliers.

In machine learning, an outlier is an observation that diverges from an overall pattern. The number of outliers in an indicator of data quality.

In the literature, there are usually considered 4 types of basic outliers for time series: additive outliers, level shifts, temporary changes and innovational outliers, see [10, 11] for a complete description.

A metric similar to the case of q_{cmp} can be defined, only taking into account the values that are considered outliers instead of the missing ones. The percentage of outliers in the studied sensor is named q_{out} (see Eq. (7)). In order to obtain which of these values are considered outliers it is useful an Autoregressive Integrated Moving Average (ARIMA) based framework [12]. It can also determine if the oulier is innovational, additive, level shift, temporary changes or seasonal level shifts.

$$q_{out} = 1 - \frac{M_{out}}{M_{total}} \tag{7}$$

where M_{out} is the sum of outlier values on the features of the sensor and M_{total} is the sum of total features.

As important as determining whether an instance is an outlier or not is knowing how much it deviates from what would be the expected value corresponding to the normal behavior of the time series. For that purpose it is necessary to impute the data of the time series that are considered outliers as if they were missing values, in order to know what this expected value would be. Then the difference between the value and the imputation is another metric that has been computed by dividing the difference of each sensors value by the mean, median or mode of the values and then calculate their mean, median or mode ($q_{mean}, q_{median}, q_{mode}$).

$$q_{mean} = mean(|\hat{x}_i - x_i|) \tag{8}$$

$$q_{median} = median(|\hat{x}_i - x_i|) \tag{9}$$

$$q_{mode} = mode(|\hat{x}_i - x_i|) \tag{10}$$

where i corresponds to those indices of the features that present an anomalous behavior, while $\hat{x}_i$ and x_i represent the imputed value that follows the expected behavior and the value of the outlier respectively. This metric takes values between 0 and 1, with 1 being the ideal case.

Unsupervised methods are also adequate for oultier detection, so we propose q_{prob}. This metric corresponds to the probability of belonging to a certain cluster that has been computed using Gaussian Mixture Models (GMM), which consists of representing in the most faithful way possible the data points by adding some Gaussian distributions. It informs quantitatively of the anomalous values. The number of clusters or Gaussians distributions is an hyperparameter and it could be chosen in different ways. In the experiments we used silhouette coefficient.

$$q_{prob} = \sum_{i=1}^{k} G_i(x_j) \tag{11}$$

where k is the number of clusters or distributions used, G_i corresponds to distribution i and x_j is the vector taken by the sensor. Because this metric is probabilistic, it takes values between 0 and 1, in such a way that the closer the value is to 1, the more quality the instance has.

Another way to determine if the data series exhibits anomalous behavior is by using so-called AutoEncoders. An AutoEncoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The objective of these autoencoders is to learn a representation of the data to be studied, with the aim of eliminating noise, however it is possible to use this tool to detect anomalous values. AE are a specific type of feedforward neural networks where the input is the same as the output. They compress the input into a lower-dimensional code and then reconstruct the output from this representation.

The metric based on AE [13] informs us about how the correlations between the different variables of the system behave. Given that, the metric q_{rec} is based on the difference between the input and the output value of the AE, in such a way that the greater the reconstruction error, the less concordance there will be between the variables [14].

$$q_{rec} = \sum_{i=1}^{N} |x'_i - x_i| \tag{12}$$

where x_i correspond to the features of the data taken by the sensors, N is the number of total features. On the other hand x'_i is the value of the vector of variables reconstructed by the AE. Sometimes $|x'_i - x_i|$ is known as a reconstruction error and is represented as E_{rec}. Since this metric is based on a difference between two values, it can take any real value greater than 0, in such a way 0 is the value with the highest quality.

3.3 Geospatial-based metrics

Considering sensors' location is also highly relevant for knowledge extraction. In this sense, we also provide two metrics that use interpolation methods for assessing how well a sensor is coordinated and correlated with its peers according to their distance. The used models are Inverse Distance Weighting (IDW) [15] and Bayesian Maximum Entropy (BME) [16]. IDW is a deterministic estimation method in which, assuming that the near sensors are more similar, a weighted average of available values at known points is used to calculate unknown data points. BME is a knowledge-based probabilistic modeling framework for spatial and temporal information. It allows various knowledge bases to be used as prior information, and the determination rules for hard (high precision) and soft (low precision) data are logically incorporated into the modeling. Like previously, we calculate the difference between the interpolated and the real measure, and the average value will become the metric, named:

- q_{inv_mean} and q_{inv_med} for IDW.
- q_{BME_mean} and q_{BME_med} for BME.

4. Examples of implementation

In this section, 3 different IoT scenarios are introduced, in which the previous metrics are computed and highlight the possible drawbacks.

4.1 Parking data

This data was collected from 5 private parking sensors located in the city of Murcia[1], Spain.

First, the variables that are useful for our goal had to be chosen: the timestamp and the parking occupation measurements and aggregated the data in 10 minutes intervals.

This aggregation can generate redundancies on the timestamps, so the result has been averaged. Storing information about this aggregation process will be useful for the Artificiality metric.

NA (not available) instances have been kept since due to their importance in obtaining some quality metrics (Completeness). Given that the data is not measured periodically, a lot of missing values are generated at this point. For illustrative purposes, a new variable called real_time was computed, which adds a random delay to the timestamps, simulating that the data needs some time to be stored. These are some highlights:

- Completeness: it consists on counting instance by instance the percentage of non-absent values there are.

- Timeliness: the random time lag that is included in the data (T_{age}) is used, so when it is divided by the arbitrary aggregation time W (600 seconds, in this case) it shows the time that data takes to be available, as follows

$$q_{tim} = 1 - \frac{T_{age}}{W} \tag{13}$$

- Plausibility: if the data of each parking lot belongs to the interval $[0, C_i]$, this measure will be said to be plausible and will receive a value of 1. The values of C_i are: 330, 312, 305, 162 and 220 respectively.

- Artificiality: due to aggregation over time, the number of instances used for computing the mean and therefore the aggregated value were considered. Thus, if a data was obtained by means of two data-points taken in the same time frame, its metric of artificiality will be $\frac{1}{2}$.

- Concordance: the geostatistical metrics have been used for covering this concept.

- Outliers: given the amount of missing data, the ARIMA framework could not be used for detecting outliers in this dataset.

A subset of the quality metrics and data values are shown in **Table 1**. Where Park101, …, Park105 are the parkings' ids, as we can see there are many instances

[1] Their locations are stored in the following web address http://mapamurcia.inf.um.es/

that cannot be correct, that information is condensed in the quality metrics. Whereas **Figure 1** shows the histograms of all basics metrics that could be computed for the parking dataset. In **Figure 2** the histogram of outlier-based metrics is shown.

Parking data geospatial-based metric's histograms are shown in **Figure 3**. As it was said above, the calculation of these metrics replace the calculation of the concordance metric, because they provide information about the correlation of the different sensors, in this case, the lower the value of the metric, the better.

4.2 Luminosity data

In this section, the monitored luminosity from 4 sensors located in the Pleiades building of the University of Murcia was studied.

timestamp	Park$_{101}$	Park$_{102}$	Park$_{103}$	Park$_{104}$	Park$_{105}$
11:50:00	NA	163.33	NA	NA	117.5
12:00:00	NA	10000	NA	NA	116.5
12:10:00	NA	163.00	NA	10000	116.5
12:20:00	NA	165.00	NA	NA	118.0
12:30:00	NA	166.00	NA	NA	120.0
12:40:00	−1	166.50	NA	NA	119.0

Table 1.
Parking observations (number of cars) subset.

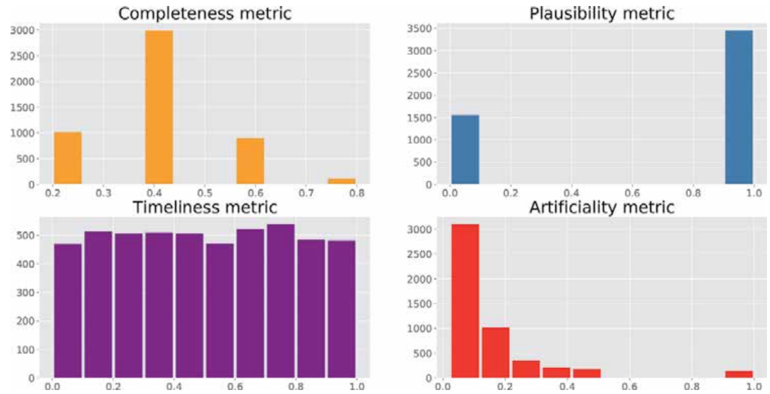


Figure 1.
Parking basic metric's histograms.

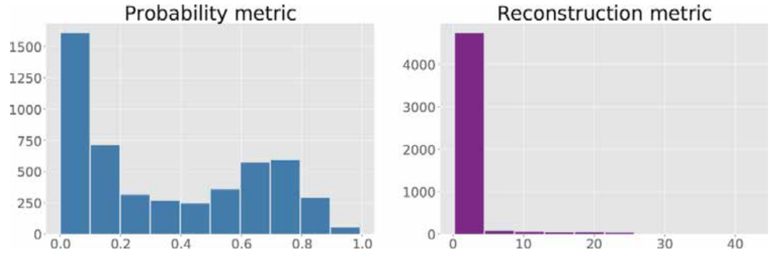


Figure 2.
Parking outlier-based metric's histograms.

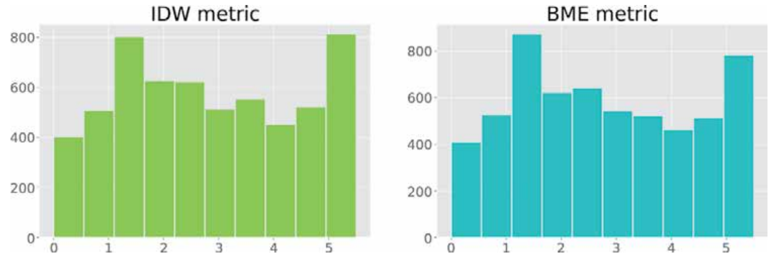


Figure 3.
Parking data geospatial-based metric's histogram.

Time	S_1	S_2	S_3	S_4
18:00:00	20	55	10	80
18:10:00	25	70	20	40
18:20:00	NA	70	10	NA
18:40:00	20	95	10	65
18:50:00	30	30	20	60
19:10:00	20	75	10	280

Table 2.
Luminosity (lumens) data subset.

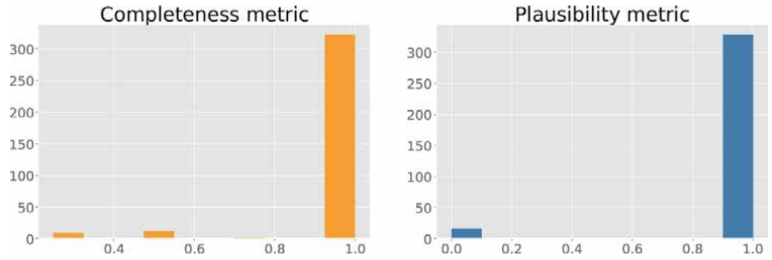


Figure 4.
Luminosity basic metric's histograms.

First, the data is aggregated using the timestamp as in the previous section, choosing a 10 minutes aggregation time. **Table 2** shows the aggregated values and also some of the computed metrics.

Figure 4 shows the histograms of all metrics that could be computed for the luminosity dataset together with basic statistics. The timeliness metric could not be calculated, since there are no signs of any lag in the data's storage. Also, the artificiality value always takes the value of 1 because the timestamps of the data are far apart. The rest of metrics are included in **Figure 5**.

By last, in **Figure 6** the geospatial luminosity's metrics can be seen. As in the case of parking, these metrics replace the concordance metric.

4.3 Pollution data

Given that the only way to calculate concordance on previous datasets has been through spatial interpolation due to poor dataset quality, a dataset of high quality has been used to compare the values that this metric takes in this situation and when they are added to it some imperfections.

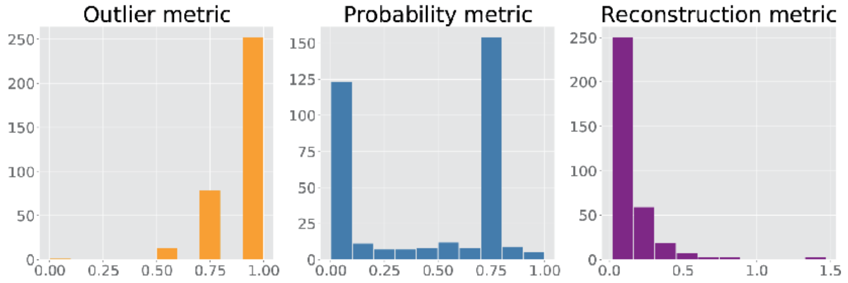


Figure 5.
Luminosity outlier-based metric's histograms.

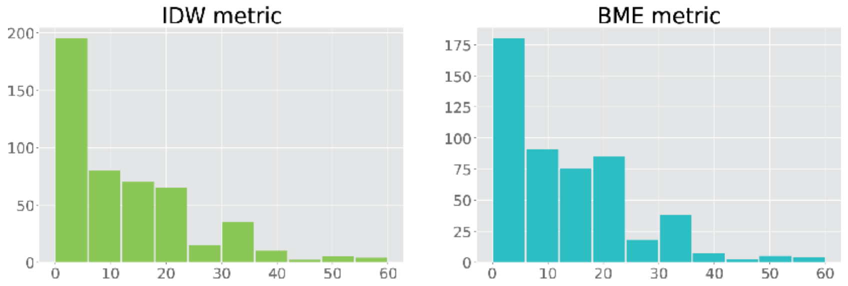


Figure 6.
Luminosity data geospatial-based metric's histogram.

Ozone	Particulate matter	Carbon monoxide	Sulfur dioxide	Nitrogen dioxide
0.18	−0.23	−1.03	−1.73	−1.66
0.29	−0.17	−1.05	−1.67	−1.68
0.31	−0.21	−1.03	−1.77	−1.70
0.22	−0.30	−0.99	−1.73	−1.66
0.27	−0.23	−1.03	−1.83	−1.77

Table 3.
Pollution data subset.

As can be seen in **Table 3**, this dataset has five variables that inform on the pollution of the atmosphere every five minutes, the data values are scaled.

Now the data are given, one way to calculate the concordance metric is to calculate the correlation between a value and the previous one, in such a way that if when the data is taken properly this value will be very close to 1, while if the data suffers any problem, this value will move away from 1. This is shown in **Figure 7**, in which we have the original dataset on the left side and the same dataset on the right side to which anomalous values have been added randomized, as it can be seen, the agreement values change significantly.

It should be noted that if the rest of the metrics are calculated in the case of the unaltered dataset, they will take perfect values, that is, they will always indicate a high quality of the dataset.

For this dataset, the rest of the metrics have been calculated, however, the results have not been added, since the dataset presents a high quality and therefore the results are not of great interest since the histograms of the metrics take the ideal behavior.

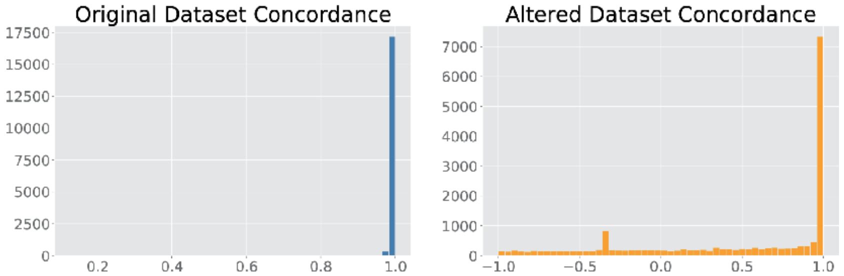


Figure 7.
Pollution dataset concordance comparison.

V1	V2	V3	V4	V5
−0.606470	−0.143913	−0.348654	2.468968	0.199896
−0.543575	0.073679	−0.223220	−0.594037	0.223077
−0.543575	−0.143913	0.090365	−0.594037	0.223077
−0.543575	−0.216444	0.090365	−0.733265	0.199896
−0.606470	−0.796689	0.090365	−0.872493	0.176716

Table 4.
Data without context subset.

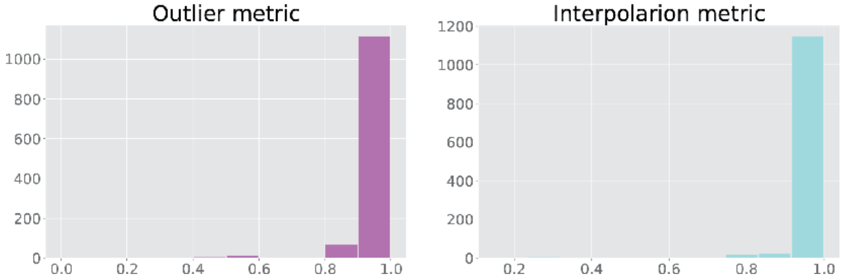


Figure 8.
Data without context outlier-based metric's histograms (I).

4.4 Data without context

For demonstration purposes, we propose to compute the quality metrics in a dataset whose context, origin and meaning are unknown. It is a dataset in which we have no knowledge about what the columns represent, how the data was collected and the timestamp of the observations. In such scenario, the only basic metric that can be computed is completeness. However, outlier-based metrics are very useful, since they consider the variables as plain time series without taking into account their physical meaning. **Table 4** shows a subset of the dataset, that presents 5 unknown variables, with 1200 instances.

Figure 8 shows the histograms of the outlier-based metrics, while **Table 5** shows the value taken by the metrics for a small data subset.

Similarly, the probabilistic and reconstruction metrics can be calculated here, since they do not assume any kind of knowledge of the data. In **Figure 9** the histogram of both metric is shown.

q_{outlier}	q_{inter}
1.00	1.00
0.90	0.80
0.85	0.80
1.00	1.00
1.00	1.00

Table 5.
Data without context subset.

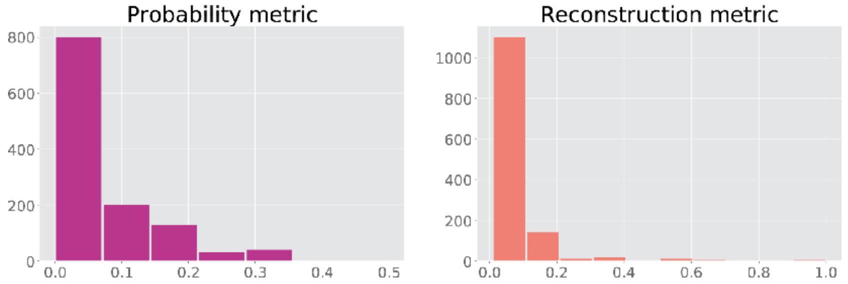


Figure 9.
Data without context outlier-based metric's histograms (II).

5. Conclusions

The proliferation of datasets thanks to the new paradigm of the Internet of Things, is populating repositories and open data platforms with data that could be of great use for the scientific community and the technologists to catalyze the growth of scientific knowledge and to make proliferate the creation of new technological solutions. Although all data has value, a point has been reached in which it is necessary to rapidly recognize the quality of a dataset, or a data stream, ideally on an only manner.

In this chapter, several concepts have been combined in order to measure the quality of data from IoT-based real-time streams (tested on real-world) sensor systems.

Three sets of quality assurance methods, descriptive, analytic and geometrical have been developed that can be used as levels of a given evaluation, or independently depending on the nature of the datasets to be evaluated.

It has been shown that the metrics can be an standard on the calculation of data quality and the majority can be applied independently on the problem context. At the same time, basic concepts that must be present in any system in which the quality of the data is to be guaranteed have been reviewed. Furthermore, it has been shown how it is possible to obtain quality metrics when knowledge about the data is limited.

The applications of this technology are linked to the proliferation of open data portals. There exist many initiatives and organizations that are working towards publishing data as open. The main funding body for engineering and physical sciences research in the UK, the Engineering and Physical Sciences Research Council (EPSRC) is supporting the management and provision of access to research data. They claim that *publicly funded research data should generally be made as widely and* *freely available as possible in a timely and responsible manner*[2]. Other initiatives are the EU Open Data Portal[3] at European level or the national-level ones such as Open Data Aarhus[4]. In that sense, the selection of data sources becomes more complicated given the great amount of data that researchers and practitioners have access to. Our system provides an easy, understandable and quick way to make an informed decision for choosing between several data sources based on data quality.

As future work, we are considering several technologies in order to make our metrics available to researchers and businesses. We consider that they have the potential to become a standard for measuring data quality.

Acknowledgements

This work has been sponsored by MINECO through the PERSEIDES project (ref. TIN2017-86885-R), by ERDF funds of project UMU-CAMPUS LIVING LAB EQC2019-006176-P by the European Comission through the H2020 IoTCrawler (contract 779852), and DEMETER (grant agreement 857202) EU Projects. It was also co-financed by the European Social Fund (ESF) and the Youth European Initiative (YEI) under the Spanish Seneca Foundation (CARM).

[2] https://epsrc.ukri.org/about/standards/researchdata/

[3] https://data.europa.eu/euodp/en/home

[4] www.opendata.dk/city-of-aarhus

Author details

Tomás Alcañiz, Aurora González-Vidal*, Alfonso P. Ramallo
and Antonio F. Skarmeta
Department of Information and Communication Engineering, Faculty of Computer Science, University of Murcia, Murcia, Spain

*Address all correspondence to: aurora.gonzalez2@um.es

IntechOpen

References

[1] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

[2] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young, "Building principles for a quality of information specification for sensor information," in *2009 12th International Conference on Information Fusion.* IEEE, 2009, pp. 1370–1377.

[3] C. H. Liu, J. Fan, J. W. Branch, and K. K. Leung, "Toward qoi and energy-efficiency in internet-of-things sensory environments," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 4, pp. 473–487, 2014.

[4] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data science journal,* vol. 14, 2015.

[5] C. Liu, P. Nitschke, S. Williams, and D. Zowghi, "Data quality and the internet of things," Computing, vol. 102, 02 2020.

[6] C. Lagoze, "Big data, data integrity, and the fracturing of the control zone," Big Data & Society, vol. 1, 11 2014.

[7] M. Celma, J. C. Casamayor, and L. Mota, *Bases de datos relacionales*. Alhambra, 2003, ch. 1, pp. 1–16.

[8] A. González-Vidal, T. Alcañiz, T. Iggena, E. Bin Illyas, and A. F. Skarmeta, "Domain agnostic quality of information metrics in iot-based smart environments," in *Intelligent Environments 2020: Workshop Proceedings of the 16th International Conference on Intelligent Environments*, vol. 28. IOS Press, 2020, p. 343.

[9] D. Kuemper, T. Iggena, R. Toenjes, and E. Pulvermueller, "Valid. iot: a framework for sensor data quality analysis and interpolation," in *Proceedings of the 9th ACM Multimedia Systems Conference.* ACM, 2018, pp. 294–303.

[10] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," Journal of the American Statistical Association, vol. 88, no. 421, pp. 284–297, 1993.

[11] Javier López-de-Lacalle. Detection of Outliers in Time Series. 2019. R package version 0.6-8. https://CRAN.R-project.org/package=tsoutliers

[12] A. González-Vidal, J. Cuenca-Jara, and A. F. Skarmeta, "Iot for water management: Towards intelligent anomaly detection," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT).* IEEE, 2019, pp. 858–863.

[13] M. Zolotukhin, T. Hmlinen, T. Kokkonen, and J. Siltanen, "Increasing web service availability by detecting application-layer ddos attacks in encrypted traffic," in *23rd International Conference on Telecommunications (ICT)*, 2016.

[14] N. García, T. Alcañiz, A. González-Vidal, J. B. Bernabé, D. Rivera, and A. Skarmeta, "Distributed real-time slowdos attacks detection over encrypted traffic using artificial intelligence," *Journal of Network and Computer Applications*, vol. 173, p. 102871, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804520303362

[15] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An experimental comparison of ordinary and universal kriging and inverse distance weighting," Mathematical Geology, vol. 31, no. 4, pp. 375–390, 1999.

[16] A. González-Vidal, P. Rathore, A. S. Rao, J. Mendoza-Bernal, M. Palaniswami, and A. F. Skarmeta-Gómez, "Missing data imputation with bayesian maximum entropy for internet of things applications," *IEEE Internet of Things Journal,* 2020.

Section 5

Applications

Chapter 6

DNA Computing Using Cryptographic and Steganographic Strategies

Adithya B. and Santhi G.

Abstract

Information protection and secrecy are major concerns, especially regarding the internet's rapid growth and widespread usage. Unauthorized database access is becoming more common and is being combated using a variety of encrypted communication methods, such as encryption and data hiding. DNA cryptography and steganography are used as carriers by utilizing the bio-molecular computing properties that have become more common in recent years. This study examines recently published DNA steganography algorithms, which use DNA to encrypt confidential data transmitted through an insecure communication channel. Several DNA-based steganography strategies will be addressed, with a focus on the algorithm's advantages and drawbacks. Probability cracking, blindness, double layer of security, and other considerations are used to compare steganography algorithms. This research would help and create more effective and accurate DNA steganography strategies in the future.

Keywords: DNA, Cryptography, Steganography, Bio-Molecular, DNA Computing

1. Introduction

The concept of security refers to the prevention of unauthorized access to information. In today's computer science, encryption's primary goal is to prevent confidential data from being altered, lost, hacked, or compromised by a third party [1]. Encryption and concealment of information are among the most widely used methods in networking and information security. Encryption and concealment of information (both similar concepts) are commonly used to keep communications secure [2, 3] fact that both methods have the same purpose. Still, their development and use are vastly different. Cryptography alters the sense of coded writing, while steganography is a covert way of writing that conceals the encrypted message's nature. Thus, in data transmission through an insecure public medium, the science of steganography is more reliable, necessary and often preferred over encryption [4, 5].

Various steganography systems, as well as their criteria, are discussed in this article based on the literature. Different systems use different strategies for embedding data, each with a set of benchmarks to evaluate performance and determine its advantages and disadvantages. Vulnerability to adversary attack is one of the three common criteria. To avoid arousing suspicion, the embedded data must be kept undetectable both visually and statistically. A fully reliable system with comparable carrier and stego file statistics should be considered during the message embedding process [5, 6]. The carrier's power, known as the amount of data concealed within it, is the second common prerequisite. The development of a steganography technique could allow more sensitive data to be hidden within the carrier while maintaining the properties of the stego file [1, 5]. A successful steganography strategy should keep enough information in its embedding capability [6]. Imperceptibility is the third common prerequisite, which is characterized as having a high embedding potential and the ability to resist intruders. The stego carrier should ideally be devoid of visual artifacts and the greater the stego carrier's fidelity should be better [2].

The masking theory is typically modeled by a pair of algorithms: embedding and extraction, as seen in **Figure 1**. The embedding algorithm produces a stego file containing the private data by merging two folders, secret and vector data, with an optional key. On the other hand, the extraction algorithm is used to recover the secret data from the stego file [7]. Steganography is a method of concealing data that does not require the use of a key. Its protection depends on the privacy of the algorithm. As a result, it is known as a less reliable approach [8, 9]. Another way to hide information is to hide confidential data, which uses one key for all operations (embedding and extraction). One of the most important benefits of this type is its rapid stage in all procedures [10, 11]. Unlike previous patterns, public steganography uses two keys for embedding and extraction: embedding and the other for extracting. The biggest value of this type is the durability of the system. The identification of the other key could be a concern if one of the keys is identified by a third party [10, 12]. On the other hand, this model is 100–1000 times slower than private steganography [13].

Several applications represent a container for confidential data. In steganography schemes, these programs are used as cover objects or carriers. Per carrier has its own set of characteristics that aid in the data concealment process. The carrier's field availability determines the amount of confidential information needed to hide

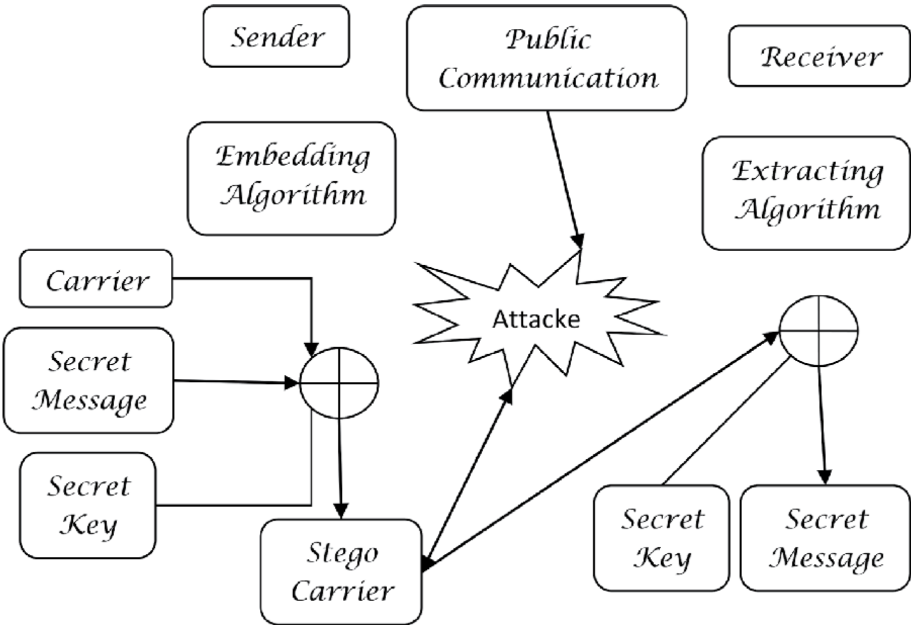


Figure 1.
Block diagram of steganography system.

data within each carrier. Text, audio, video, and photographs are examples of multimedia used to hide records. Text can be obscured by changing the text's layout, inserting an nth character from the text, or changing any of the rules, such as spacing. Text can also be hidden using a code made up of letters, lines, and page numbers. However, this process is insecure [2]. The biggest benefit of this carrier is that it does not take a lot of memory and is quick to switch.

In contrast to other carriers, it has a very limited number of redundant data [10, 14]. The use of inaudible frequencies and a small shift in the binary sequence of an audio file can be used to hide data in audio files [2, 15]. Data masking in video files is more efficient and effective due to the wide available space. Allowing data to be hidden within multiple video frames [16]. Uncompressed and compressed video are the two main formats of video in which data can be hidden. Digital images have been common carriers for masking confidential information due to their high redundancy, high capacity in images, low effect on exposure, and ease of manipulation [15, 17]. DNA is a relatively recent vector that has been used in the field of steganography. In this article, we look at the data hidden in DNA.

2. Deoxyribonucleic acid (DNA)

The most important molecular structure in biology is deoxyribonucleic acid (DNA), which encodes the information required to generate and direct all chemical elements in the human body. As a result, DNA has been suggested as a possible candidate for computational purposes [18].

2.1 DNA structure

DNA is described as a living creature's genetic blueprint. Each body cell has its DNA collection and a polymer made up of monomers called deoxyribose nucleotides, consisting of three components, as seen in **Figure 2** [19].

The human body is made up of trillions of cells, each with its purpose. As seen in **Figure 3**, each cell has a nucleus that comprises several chromosomes. The majority of DNA is present in a nucleus, which is known as nucleus DNA, and the remainder is found in mitochondria, which is known as mitochondria DNA (mtDNA). Each cell's activity is regulated by DNA. DNA chromosome is made up of a DNA molecule of genes. A gene is the entire genetic makeup of an organism, containing information from all chromosomes [20].

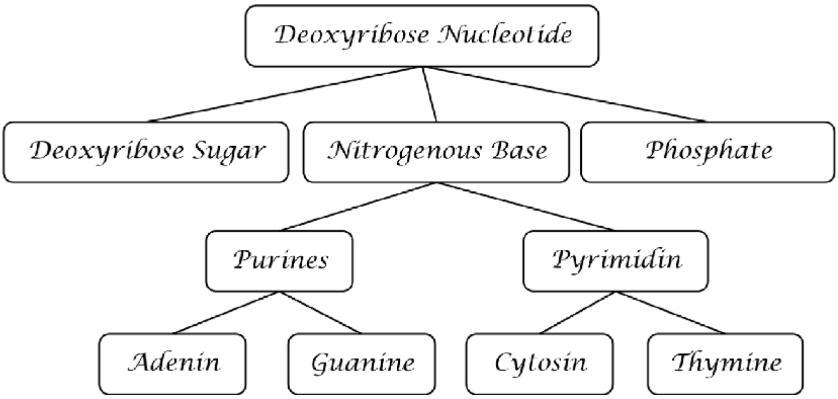


Figure 2.
Structure of deoxyribonucleic acid.

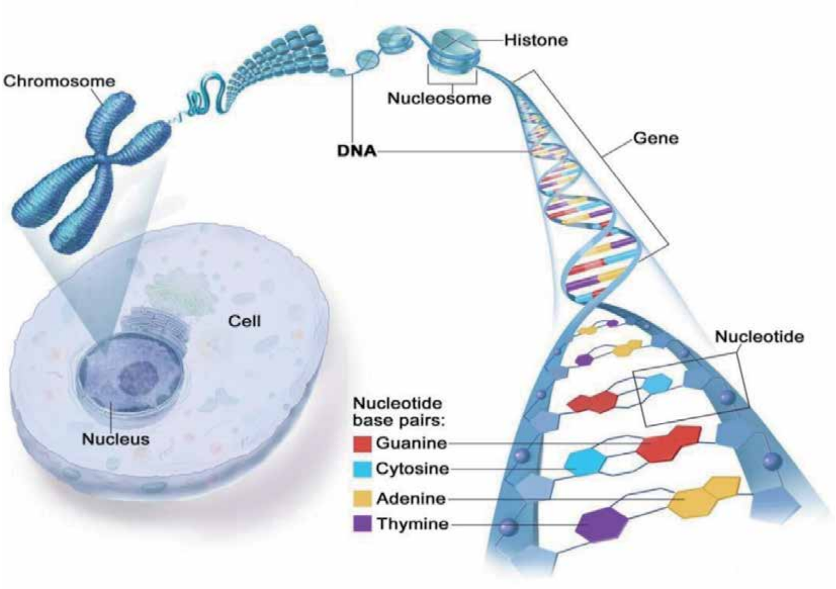


Figure 3.
Gene development cycle.

In 1953, Watson and Crick discovered DNA structure, a form of genetic material. DNA is a long molecule present in all living things' body cells. DNA is a kind of bacterial plasma that contains all lifestyles. It is made up of two simple bands that are twisted around each other in a double helix (see **Figure 4**). Each DNA chain is made up of nucleotides, which are small subunits. The four chemical bases in the chain DNA are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C), which bind to sugar and phosphates in the backbone to complete the nucleotide. Purines (A and G) and Pyrimidine (T and C) are the two DNA bases in biology. Continuously (A) is bound to (T) by two hydrogen bonds, and (C) is bound to (G) by three hydrogen bonds [19, 21]. Transcription is the method for producing RNA, which is an intermediate copy of DNA instructions. Adenine (A), Cytosine (C), Uracil (U), and Guanine (G) are the four bases that makeup RNA. All 64 codons are represented in **Figure 5**. The STOP codons do not necessarily symbolize any amino acids but rather indicate the protein chain's end. The twenty amino acids are determined by the remaining 61 codons. Some amino acids are coded by several codons [11]. As a result of this codon duplication, it is possible to change the genetic sequence while keeping it functional [11, 23, 24].

2.2 DNA computing

Currently, biology methods are used in a variety of fields. DNA is a relatively new biological technology that is used in a variety of applications [25]. This is because DNA computing can solve a variety of NP-complete problems, in which the computation time increases dramatically.

There has been a considerable amount of research in this field, with significant progress made on DNA and the immune system [19]. Leonard Adelman conducted the first experiment in DNA computing (bio-molecular computing) in 1994, in which molecular biology instruments were used to solve a portion of the standard path of the Hamiltonian puzzle. Computing with molecules directly was discovered at the time, and it was regarded as a new discipline in terms of science defense [26]. The satisfaction problem (SAT), an NP-complete problem, was solved using DNA

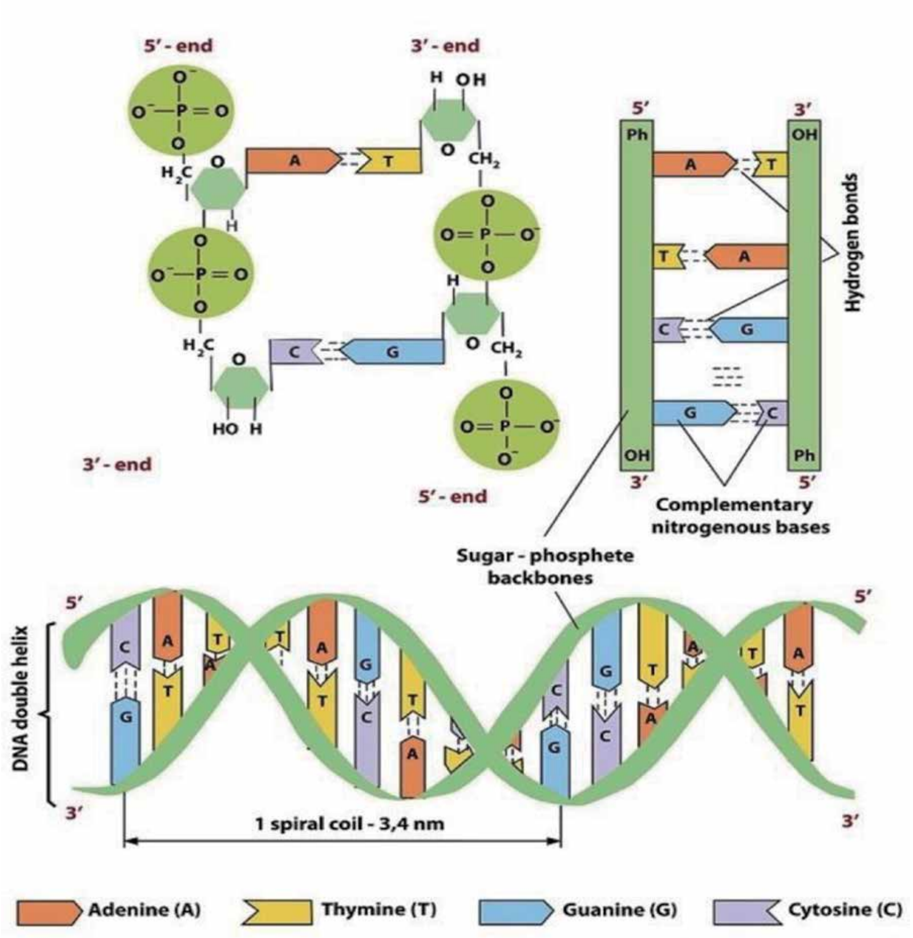
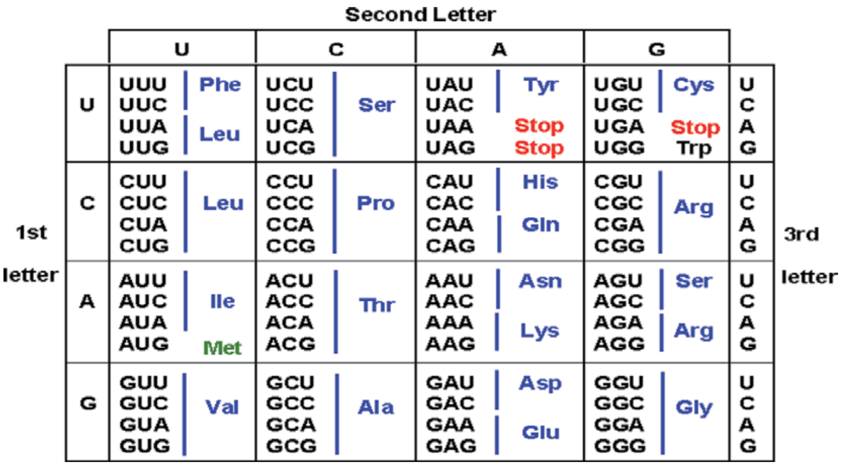


Figure 4.
Helical structure of DNA [20].



	Second Letter								
1st letter	U		C		A		G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 5.
Codon and amino acid table [22].

DNA base	Binary code
A	00
C	01
G	10
T	11

Table 1.
Binary code of DNA.

computing in a 1995 study by Lipton. The offered approach took advantage of DNA's parallelism and its computational and storage capacities [19]. In 1997, Ogihara and Ray discovered that DNA could be used to simulate AND and OR gates [27]. Clelland [28] proposed the first successful experiment of a DNA steganography technique for concealing sensitive data using DNA microdots.

2.3 Binary code of DNA

A, C, G, and T are the four chemical bases that make up each DNA chain. A is biologically related to T, while C is related to G. T The synthesis of DNA rules can be modified in binary arithmetic by changing input judgments, such as assuming that T is related to C or T is related to G [29]. Researchers would use a binary encoding rule to translate a hidden message into DNA rules before mixing it with sequenced DNA to store data in DNA particles. For each rule (A), researchers may use the corresponding binary form: binary formulas can be "00," "01," "10," or "11." This can be expressed as in **Table 1**. The encoding of DNA and its random properties make it an ideal candidate for both coding and coding. As a result, converting DNA into the binary form will result in 4! = 24 different encoding methods [30, 31]. On DNA bases, logical operations such as addition, subtraction, XOR, AND, OR, and NOT are possible.

3. Comparative study

The aim of the comparison presented in this study is to ensure that researchers are aware of the shortcomings in current steganography systems, thus inspiring future advances in this field. **Table 2** compares the strengths and disadvantages of existing algorithms in terms of security problems such as chance of intrusion, double security layer, blindness, and more.

The derived comparison in **Table 2** aims to clarify the proposed DNA's strengths and weaknesses using data hiding algorithms. Encrypting sensitive data into encryption data before embedding, rather than including the initial data format, improves confidentiality [13, 18, 23, 34, 36, 38, 41, 44, 45, 47–49, 51, 52, 55, 57–60]. Playfair technology, adopted in [58], is the most promising encryption technology combined with DNA-based data masking technology. A thorough comparison of several encryption methods, including vigenere and Playfair, AES, and RSA ciphers, has been done in their work. Any of them was paired with a replacement tool for hiding data in DNA. The findings revealed that the Playfair cipher is not only quick and easy to use, but it also has a high level of protection and ability.

The blindness trait, which eliminates the need to give the original DNA connection to the recipient, is the primary function supported by DNA-based data masking techniques. The main goal of the blindness feature is to improve protection and avoid any intruder way of detecting it, as shown in [11, 18, 25, 35, 41, 43, 48, 49, 51, 57, 58, 62].

S.No	Reference	Strengths	Weaknesses
1	[24]	*Insertion Technique* • High embedding capacity. • Simple to bring into practice. • Modification rate is low.	• Length of Stego DNA is longer than length DNA of comparison. • The payload does not equal zero. • In the extraction process, multiple data is needed. • The amino acid function is not preserved. • The algorithm is not blind. • Increase the level of redundancy • Steganography method for purely obscuring results.
		Complementary Technique • Simple to bring into practice. • To break the hidden data, attackers must have a ton of information.	• The payload does not equal zero. • Modification rate is high. • The algorithm is not blind. • Steganography method for purely obscuring results. • After the embedding process, the length of DNA is modified.
		Substitution Technique • High embedding capacity. • Simple to bring into practice. • The payload is set to zero. • In contrast to previous approaches, this one is more efficient, dynamic, and performs better.	• The amino acid function is not preserved. • The algorithm is not blind. • Steganography method for purely obscuring results. • Modification rate is high.
2	Ref [21]	• The payload is set to zero. • High embedding capacity. • Simple to bring into practice. • Maintain the biological DNA's features. • To increase the degree of secrecy and complexity, the consequence of hiding data in the cloud is being implemented.	• The DNA reference determines the level of security. • Increase the size of the message. • The algorithm is not blind. • Steganography method for purely obscuring results.
3	Ref [25]	• Build a steganography method that is reversible. • Preserve the DNA's versatility. • The algorithm is blind. • A secret key is employed.	Does not encrypt confidential information when storing it.
4	Ref [18]	• To provide security, a map was created between DNA codons and amino acids. • Before hiding, use the playfair cipher to encrypt the hidden letter. • Improve the playfair cipher by changing it to 5*5 to prevent its pitfalls, such as the diagraphs and hidden text form remaining after encryption. • Adding a second layer of protection. • Algorithm for the blind. • Capacity and time efficiency are also improved. • Provide a high risk of cracking. • It is necessary to use a hidden key.	• Increase the length of the stego DNA. • The biological DNA's versatility is not preserved. • It must send many data to the recipient in order to retrieve the hidden message from Stego DNA. • The payload is not empty.
5	Ref [32]	• The usefulness of the initial replacement process has been improved.	• The biological DNA's versatility is not preserved.

S.No	Reference	Strengths	Weaknesses
		• The communication performance of a data hiding device on the internet can be enhanced. • In terms of power and protection, providing better results. • TLSM has been enhanced to allow secret data to be hidden in any series of letters or symbols. • The Base-t TLSM and the Extended TLSM (ETLSM) are two methods proposed to increase the efficiency of the TLSM. • Capacity has been expanded.	• It needs to submit multiple data, including DNA reference, Stego DNA, secret message site collection, table code, to extract the secret message from Stego DNA. • Modification rate is high. • The algorithm is not blind. • Steganography method for purely obscuring results.
6	Ref [13]	• Proposed a protocol for masking encrypted data to limit the use of public keys while maintaining the highest level of reliability. • The payload is set to zero. • A wide embedding capacity. • Using the cutting-edge technology of DNA data hiding, the secret key is hidden inside the DNA reference for added confidentiality.	• The biological DNA's versatility is not preserved. • The algorithm is not blind.
7	Ref [33]	• If the length of stego DNA is not extended, the payload is zero. • Algorithm is simple. • The ability to cover has been enhanced. Reduce the pace of modification. • In hiding, substitution form is used.	• The biological DNA's versatility is not preserved. • If the DNA comparison includes a number of repeated nucleotides, the modification rate would be high. • Both the sender and the receiver should be aware of the un-blind algorithm, as well as injective mapping and complementary rules. • Algorithm for simply obfuscating results.
8	Ref [34]	• Flexible algorithm that is easy to execute. • Encrypt a hidden message using a revamped Playfair algorithm that incorporates DNA and amino acids. • After the hiding process, the length of DNA does not extend. • It is necessary to use a hidden key. • In hiding, substitution form is used.	• The biological DNA's versatility is not preserved. • Algorithm that is not blind.
9	Ref [35]	• The algorithm employs three keys. • In terms of modification volume, the first and third techniques of Ref [24] have been improved. • The stego DNA is not expanded. • The algorithm is blind.	• There is no encryption method used. • Only nucleotides with marks equal to zeros after conversion to binary are used to hide hidden records.
10	Ref [36]	• It's easy to bring into effect. • Low rate of modification. • The length of stego DNA is not increased. • To encrypt hidden data before hiding it, one of the most efficient encryption techniques (RSA) is used. • A public key is employed.	• The biological DNA's versatility is not preserved. • Save the location of each DNA base that contains the hidden data and submit it to the receiver for extraction. • The hidden data's size has been expanded. • Algorithm that is not blind.

S.No	Reference	Strengths	Weaknesses
			• Cracking with a low probability
11	Ref [36]	• High embedding capacity. • Simple to bring into practice.	• The biological DNA's versatility is not preserved. • Algorithm that is not blind. • Algorithm for simply obfuscating results. • Cracking with a low probability
12	Ref [11]	• It's easy to bring into effect. • Ensure the biological DNA's functionality is maintained. • Low rate of modification. • The algorithm is blind. • The secret key is hidden in the DNA guide, which adds to the protection. • After hiding sensitive details, the DNA reference is not extended.	• Due to the use of LSB in the hiding operation, the potential is low. • There was no encryption on the confidential data until it was hidden. • Cracking with a low probability.
13	Ref [23]	• Exhibit DNA amino acids to encrypt hidden records. • Before hiding the secret key inside the DNA reference, encrypt it using the RSA algorithm. • The public key is used, and the capability is high. • Cracking with a high probability.	• Algorithm that is not blind. • A high degree of modification. • The payload is not empty. • The versatility of amino acids is not maintained.
14	Ref [37]	• Preservation of protein translation in the protein coding DNA (PcDNA). • Data encoding is consistent and near optimal. • Keep track of the codon statistics. • Embedding data came close to being perfect. • Embedding data in DNA in a reliable and effective manner. • A secret key is employed.	• Estimation that is difficult. • Unconstrained ncDNA hiding can be estimated by intruders.
15	Ref [38]	• Before using the Playfair algorithm to hide hidden data, encrypt it. • High-level surveillance. • Since hiding in an audio at the last stage would not draw attackers. • Hide the secret data and translate it into an audio file so that it is impossible to show that all data is inside the audio. • Provide two layers of concealment. • A secret key is employed.	• The hidden data must be extracted using several data sources. • The algorithm is not blind.
16	Ref [38]	• Key area is wide enough to resist negative intruders using brute force. • Before hiding secret data in host text, encrypt it. • The algorithm is blind. • The embedding power ratio is 100 percent. • Provide two layers of concealment. • Chebyshev maps are used to establish DNA references. • In hiding, the substitution method is used.	• Calculation is difficult.

S.No	Reference	Strengths	Weaknesses
17	Ref [39]	• Ref [40] algorithm's hidden key was modified to use the secret key. As well as keeping all of Ref [40] high points.	• Pure steganography algorithm. • Complex calculation.
18	Ref [41]	• The initial replacement technique's capability and protection have been increased. • The algorithm is blind. • Method of replacement has been improved.	• Pure steganography algorithm. • The biological DNA's versatility is not preserved. • If multiplied by 6, if the result is not equal to zero, additional zeros are added. • The length of Stego DNA is extended.
19	Ref [42]	• High embedding capacity. • Simple to bring into practice. • Secret data is sent in the (ABCD) format.	• Pure steganography algorithm. • Cracking with a low probability • Algorithm is not blind. • The receiver should obtain a random DNA sequence and a complementary pair law. • There is no encryption on the data until it is embedded. • Cracking with a low probability • Steganography method for purely obscuring results.
20	Ref [43]	• Only the correct value of Stego DNA is sent to the recipient. • High level protection. • Hackers have a tough time spotting the seeds of the random numbers generated. • Hackers have a hard time deciding how many packets to split, in addition to the number of DNA message bits and binary in each packet. • The secret message bits and DNA comparison bits are randomly combined. • The algorithm is blind. • A secret key is employed. • Cracking with a high probability	• Redundancy has been increased. • The message size has been increased. • The DNA functionality is not preserved. • Increase the size of stego DNA.
21	Ref [44]	• A secret key is employed. • Until hiding a secret document, encrypt it with RC4. • Exceptional ability. • Providing a safe environment. • Provide two layers of concealment. • Build DNA from a picture.	• During the extraction process, the algorithm needs several keys.
22	Ref [45]	• A secret key is employed. • Classified data protection has increased dramatically. • Extra grids of different sizes may be used to store additional data. • BASE64 encoding is used to encrypt confidential info. • Provide two layers of concealment. • Secret text is used to build DNA.	• Complex calculation.
23	Ref [46]	• A secret key is employed. • High levels of protection. • High capacity. • Since the key of prime duration is between 20 and 40, the possible prime range is 420–440.	• The extraction header and data extractions are two aspects of the algorithm.

S.No	Reference	Strengths	Weaknesses
		• Increased payload capability thus reducing image distortion. • Until being hidden, sensitive data is encrypted using RC4 encryption. • Provide two layers of concealment. • Develop DNA from the cover image.	
24	Ref [47]	• A secret key is employed. • Ensure that there are two levels of protection. • AES-128 is used to encrypt secret files. • AES has provided a strong degree of protection. • Before and after encryption, separate operations such as XOR and HASH-512 were performed on sensitive data. • Microdot has DNA embedded it to improve security.	• Several types of data are needed during the extraction process. • The DNA functionality is not maintained.
25	Ref [48]	• Modification rate is low. • After embedding confidential details, the DNA reference does not extend. • It makes use of two DNA references. • The initial DNA reference's usefulness was preserved. • Algorithm for blind people. • The non-labeled nucleotides do not shift. • High ability. • Until embedding plain text, encrypt everything. • Cracking with a high probability.	• Steganography method for purely obscuring results. • The receiver should be sent substitution rules. • Only uppercase letters, lowercase letters, 0, …., 9, period, and dots) are allowed in plain text. • It cannot have any other punctuation marks in it.
26	Ref [49]	• In the suggested algorithm, three DNA references are used. • Before hiding the plain text, encrypt it. • A secret key is employed. • Cracking with a high probability. • The algorithm is blind.	• Modification rate is high. • The biological DNA's versatility is not preserved.
27	Ref [50]	• Any programming language can be used to execute it. • To translate a hidden message to DNA format, build a random codon table. • Because of the insertion technique, there is a lot of duplication.	• There is no encryption. • May not keep records of an organism's life knowledge. • After embedding, lengthen the DNA reference. • The algorithm is not blind. • Algorithm for purely hiding records.
28	Ref [51]	• The algorithm is blind. • A secret key is employed. • Encrypt the hidden message using Playfair's algorithm. • After hiding the hidden data, there was no extension to the DNA reference. • In concealment, the replacement form is used. • Modification rate is poor. • The initial DNA reference's usefulness was preserved.	• Cracking with a low probability • The alteration rate would be high if the DNA comparison has several repetitive bases.
29	Ref [52]	• A secret key is employed. • High embedding capacity.	• The biological DNA's versatility is not preserved.

S.No	Reference	Strengths	Weaknesses
		• Using a modified Playfair algorithm, encrypt a secret letter. • After the hiding process, the length of DNA does not extend. • Easy, fast to implement, and performs better than Ref [32]. • Ref [32] hiding mechanism has been improved. • In hiding, the substitution form is used.	• The algorithm is not blind. • Cracking with a low probability
30	Ref [53]	• Technique that is almost imperceptible. • Before hiding a hidden message, encrypt it. • Provide two layers of concealment.	• The algorithm is not blind. • Algorithm for purely hiding records. • Only one part of the cover image is used to hide the DNA message.
31	Ref [54]	• A secret key is employed. • Without distorting the picture, two secret images may be hidden within it. • Provide two layers of concealment.	• The algorithm is not blind. • On secret records, no encryption technique was used.
32	Ref [55]	• Protection has been improved. • By reducing picture noise bits, the double carrier has been improved. • Enable for a fair amount of space. • Using a two-dimensional 2D logistic map with many parameters. • RC4 is a cryptographic algorithm that is used to encrypt sensitive information. • Provide two layers of concealment. • Image is used to create DNA. • A secret key is employed. • In hiding, the substitution form is used.	• Multiple data are required in the embedding and extraction processes.
33	Ref [56]	• Technique that is almost imperceptible. • This is an effective method. • By hiding in a random video frame, you can have protection. • Provide two layers of concealment.	• The algorithm is not blind. • Algorithm for purely hiding records. • The extraction method necessitates the use of numerous data sources.
34	Ref [57]	• The algorithm is blind. • Method that is both safe and efficient. • Until embedding, encrypt hidden data using the RSA algorithm. • Provide two layers of concealment. • A public key is employed.	• The biological DNA's versatility is not preserved.
35	Ref [58]	• Keeping track of an organism's life records. • The length of stego DNA is not increased. • The hidden data is encrypted using XOR and PRBG. • Reed-Solomon (RS) programming is used to measure and correct errors.	• It's not easy to put into practice. • Modification rate is high.
36	Ref [59]	• The hidden data and the key may be of any form and dimension.	• Algorithm is not blind.

S.No	Reference	Strengths	Weaknesses
		• Until hiding, using various encryption methods and analyzing them to choose the best one. • The normal key is used to select English characters to create more stable playfair cipher network. • There is no redundancy in the operation. • Strong results in a limited period of time. • In hiding, the substitution form is used.	• The amino acid functionality is not maintained. • A high degree of modification. • Cracking with a low probability
37	Ref [60]	• Using the vigenere or playfair cipher, encrypt hidden info. • The sum of data that is hidden is doubled. • High levels of security. • Until submitting to the recipient, the DNA connection will be hidden in a microdot on a piece of paper. • If the paper is unsafe, recreate a new key and sequence DNA, and the hiding process will start again. • Maintain the DNA sequence's functionality while avoiding mutations.	• Different data sets are sent to the receiver for retrieval. • Non-coding area has a high degree of alteration.
38	Ref [61]	• High-level security. • Random key generator for two levels of randomness. • It is necessary to use a hidden key. • The risk of cracking is incredibly high.	• Algorithm is not blind. • The functionality of DNA is not maintained. • The payload is not empty.

Table 2.
A comparison of the strengths and weaknesses of DNA steganography techniques.

This is accomplished by minimizing the requisite data that is transmitted to the recipient as much as possible. One of the strengths is to biologically preserve the DNA relationship's original features during the inclusion step while maintaining a fair data load. The reference DNA is used to mask hidden data while preserving protein processing functions. As shown in [11, 21, 25, 37, 48, 51, 52, 58, 60], some DNA characteristics such as silent mutation and codon repetition can mask details and alter the genetic sequence without changing the protein chain.

After most data-masking algorithms, the carrier can experience some distortion. Data masking techniques take care of embedding and embedded data; that is why it is communicated invisibly. As a result, it is important to minimize conveyor distortion. When data is entered into a string of stego DNA, the sequenced DNA's length and the degree of change are used to determine stego DNA precision. The low rate of change and lack of expansion rate results in high-quality DNA, which attracts less interest from potential attackers. [11, 33, 35, 36, 48, 51] reaches a low modulation frequency. Moreover, the expansion rate characteristic of DNA stego is not achieved at [11, 13, 21, 33–36, 48, 51, 58], which means that the payload is equal to zero.

It is recommended to use a two-stage steganography technique to hide sensitive data with more detail than previous data masking methods. Using two separate vectors in the same manner, increases confidentiality and makes it difficult for criminals to ingest or recover hidden data. Several methods [38, 44, 46, 54–57, 62] used the ref. DNA with another multimedia player to cover the hidden data. Some built DNA from cover images or confidential information, as shown in [44–46, 55, 62], while others used a random sample or selected from an online database, as shown in [38, 54–57].

The main factor is one of the most important aspects of data masking strategies. Data masking schemas are centered on the key used and can be classified into three categories. As shown in [21, 24, 32, 33, 40–42, 48, 50, 53, 56], pure data masking is less reliable because it does not use any key. As a result, using a key increases the device's usability by complicating the data-masking mechanism attack. Even if the perpetrators figure out what data-masking scheme is being used, they are unable to retrieve it. The carrier's sensitive information is not protected by the key. The secret is only in the hands of the sender and receiver. As a result, it is advisable to use a strong key when encrypting files, which ensures a more stable method. The second form is the hidden key [11, 13, 18, 25, 34, 35, 37–39, 43–47, 49, 51, 52, 54, 55, 58–61], which was accomplished in [11, 13, 18, 34, 35, 37–39, 43–47, 49, 51, 52, 54, 55, 58, 59]. The third form is classified as a public key, as shown by [23, 36, 57]. The public key is more secure than the private key in general, but it is still slower.

The probability of splitting the code and accessing confidential, sensitive data is known as the algorithm-cracking potential. Studying the probability of a striatum fracture aims to identify the variables that ensure that the risk of rupture is reduced. The likelihood of a leak is determined by the inclusion of certain unknown variables in the algorithm used to mask sensitive data, not by the amount of attempts made before the attacker gained access to the secret data. High probability penetration leads to high protection of the data-masking strategy described in [18, 23, 43, 48, 49, 58, 61]. The replacement strategy is believed to be a more powerful means of concealing data in DNA. The DNA sequence length can be preserved using this process as long as the payload is kept at zero. It also has more power as seen in [32–34, 41, 51, 52, 55, 59, 62], because it substitutes certain DNA nucleotides with cached data blocks or other nucleotides based on confidential data.

Capacity is a vital aspect of any data masking strategy, and it is one of the main criteria for data masking techniques. A steganography strategy must have broad data anonymization potential. This capacity can be measured in absolute terms, such as the hidden message's volume (for example, the data embedding rate, the bit per pixel, the bit per non-zero discrete cosine, the conversion factor, or the ratio of the secret message to a medium). The strength of DNA is calculated in bits per nucleotide (bpn). Thus, one of the main concerns for researchers in this area is improving the potential of secret results, which has previously been accomplished in [13, 18, 21, 23, 32, 33, 40–42, 44, 46, 48, 52, 55, 58–60].

As a result, it can be inferred that the primary goal of DNA-based double-layer masking algorithms is to encode sensitive data before hiding it in a high-power, blind, bio-stored, low moderation rate, load-free algorithm, not a pure method, with a high probability crack. In [48, 51, 58] suggested a low moderation rate, preservation of stretch length DNA for contrast, blindness, preservation of DNA versatility, double layer of security, high strength, and not a pure algorithm.

4. Conclusions

An increase in storage demand has generated a massive demand for creating new and evolving strategies for storing large amounts of data. DNA has recently been recognized as an efficient data carrier with the additional benefit of dependable data storage. DNA's bio-molecular computing capabilities are being used in cryptography and steganography. This research compares some recent DNA-based steganography algorithms and points out their security flaws. Each algorithm's advantages and disadvantages are also listed. Some crucial issues are discussed in terms of chance breaking, double layer security, single and double hiding layers, blindness, biologically retained DNA, alteration rate, an extension of DNA comparison, not a pure algorithm, substituting operation, and capacity. This study's comparison aims to provide researchers with the information they need to perform future tasks on more effective and accurate stable DNA steganography techniques.

Conflict of interest

"The authors declare no conflict of interest."

Author details

Adithya B.[1]* and Santhi G.[2]

1 Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India

2 Department of Information Technology, Pondicherry Engineering College, Puducherry, India

*Address all correspondence to: adithya27.07@pec.edu

References

[1] Singh G. A study of encryption algorithms (RSA, DES, 3DES and AES) for information security. International Journal of Computer Applications, 2013; 67(19).

[2] Subhedar MS, Mankar VH. Current status and key issues in image steganography: A survey. Computer science review, 2014; 13:95–113.

[3] Hamed G, et al. Comparative study for various DNA based steganography techniques with the essential conclusions about the future research. 11th International Conference on Computer Engineering & Systems (ICCES); IEEE; 2016.

[4] Amin MM, et al. Information hiding using steganography. 4th National Conference on Telecommunication Technology; IEEE; 2003.

[5] Al-Mohammad A. Steganography-based secret and reliable communications: Improving steganographic capacity and imperceptibility [thesis]. Brunel University, School of Information Systems, Computing and Mathematics; 2010.

[6] Santoso KN, et al. Information Hiding in Noncoding DNA for DNA Steganography. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences; 2015; 98(7):1529–1536.

[7] Kumari P, Kapoor R. Image Steganography for Data Embedding & Extraction using LSB Technique. International Journal of Computer Applications & Information Technology; 2016; 9(2):192.

[8] Ashok J, et al. Steganography: an overview. International Journal of Engineering Science and Technology; 2010; 2(10):5985–5992.

[9] Nickfarjam AM, Azimifar Z. Image steganography based on pixel ranking and Particle Swarm Optimization. 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP); IEEE; 2012.

[10] Sheelu AB. An Overview of Steganography. IOSR Journal of Computer Engineering (IOSR-JCE); 2013; 11(1):15–19.

[11] Khalifa A. LSBase: A key encapsulation scheme to improve hybrid crypto-systems using DNA steganography. 8th International Conference on Computer Engineering & Systems (ICCES); IEEE; 2013.

[12] Jain S, Bhatnagar V. Analogy of various DNA based security algorithms using cryptography and steganography. International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT); IEEE; 2014.

[13] Torkaman MRN, Kazazi NS, Rouddini A. Innovative approach to improve hybrid cryptography by using DNA steganography. International Journal of New Computer Architectures and their Applications (IJNCAA); 2012; 2(1):224–235.

[14] Bansod S, Bhure G. Data encryption by image steganography. Int. J. Inform. Comput. Technol; Int. Res. Publ. House; 2014; 4:453–458.

[15] Singh KU. Video steganography: text hiding in video by LSB substitution. International Journal of Engineering Research and Applications; 2014;4(5): 105–108.

[16] Chandel B, Jain S. Video Steganography: A Survey. IOSR Journal of Computer Engineering (IOSR-JCE); 2016; 18(1):11–17.

[17] Yang Y. Information analysis for steganography and steganalysis in 3D polygonal meshes [thesis]. Durham University; 2013.

[18] Atito A, Khalifa A, Rida S. DNA-based data encryption and hiding using playfair and insertion techniques. Journal of Communications and Computer Engineering; 2012; 2(3):44.

[19] Al-Wattar AHS, Mahmod R, Zukarnain ZA, Udzir N. Review Of Dna And Pseudo Dna Cryptography. International Journal of Computer Science and Engineering (IJCSE); 2015; 4(4):65–76.

[20] Tornea O. Contributions to DNA cryptography: applications to text and image secure transmission [Thesis]. Université Nice Sophia Antipolis; 2013.

[21] Abbasy MR, et al. DNA base data hiding algorithm. International Journal of New Computer Architectures and their Applications (IJNCAA); 2012; 2(1): 183–192.

[22] Adithya B, Santhi G. Bio-inspired Deoxyribonucleic Acid based data obnubilating using Enhanced Computational Algorithms. In: Proceedings of the International Conference on Computer Networks, Big Data and IoT; Springer; 2020. p. 597–609

[23] Skariya M, Varghese M. Enhanced double layer security using RSA over DNA based data encryption system. International Journal of Computer Science& Engineering Technology (IJCSET); 2013; 4(6):746–750.

[24] Shiu H, et al. Data hiding methods based upon DNA sequences. Information Sciences; 2010; 180(11): 2196–2208.

[25] Mousa H, et al. Data hiding based on contrast mapping using DNA medium. Int. Arab J. Inf. Technol.; 2011; 8(2): 147–154.

[26] Adleman LM. Molecular computation of solutions to combinatorial problems. Nature; 1994; 369:40.

[27] Ogiwara M. Simulating Boolean Circuits on DNA Computers. In Proceedings of the 1st International Conference on Computational Molecular Biology; ACM Press; 1997.

[28] Clelland CT, Risca V, Bancroft C. Hiding messages in DNA microdots. Nature; 1999; 399(6736):533–534.

[29] Sureshraj D, Bhaskaran VM. Automatic DNA sequence generation for secured cost-effective multi-cloud storage. 2012.

[30] Singh A, Singh R. Information hiding techniques based on DNA inconsistency: An overview. 2nd International Conference on Computing for Sustainable Global Development (INDIACom); IEEE; 2015.

[31] Bhateja A, Mittal K. DNA Steganography: Literature Survey on its Viability as a Novel Cryptosystem. Journal of Computer Science and Engineering; 2015; 2(1):8–14.

[32] Taur J.-S, et al. Data hiding in DNA sequences based on table lookup substitution. International Journal of Innovative Computing, Information and Control; 2012; 8(10):6585–6598.

[33] Guo C, Chang C-C, Wang Z-H. A new data hiding scheme based on DNA sequence. Int. J. Innov. Comput. Inf. Control; 2012; 8(1):139–149.

[34] Khalifa A, Atito A. High-capacity DNA-based steganography. 8th International Conference on Informatics and Systems (INFOS); IEEE; 2012.

[35] Huang YH, Chang CC, Wu CY. A DNA-based data hiding technique with

low modification rates. Multimedia tools and applications; 2014; 70(3):1439–1451.

[36] Mitras BA, Abo A. Proposed steganography approach using DNA properties. International Journal of Information Technology and Business Management; 2013; 14(1):96–102.

[37] Haughton D, Balado F. Biocode: Two biologically compatible algorithms for embedding data in non-coding and coding regions of dna. BMC bioinformatics; 2013; 14(1):121.

[38] Shyamasree C, Anees S. Highly secure DNA-based audio steganography. International Conference on Recent Trends in Information Technology (ICRTIT); IEEE; 2013.

[39] Haughton D, Balado F. Security study of keyed DNA data embedding. Global Conference on Signal and Information Processing (GlobalSIP); IEEE; 2013.

[40] Bhattacharyya D, Bandyopadhyay SK. Hiding secret data in dna sequence. International Journal of Scientific &Engineering Research; 2013; 4(2).

[41] Agrawal R, Srivastava M, Sharma A. Data hiding using dictionary based substitution method in DNA sequences. 9th International Conference on Industrial and Information Systems (ICIIS); IEEE; 2014.

[42] Menaka K. Message encryption using DNA sequences. World Congress on Computing and Communication Technologies (WCCCT); IEEE; 2014.

[43] Manna S, et al. Modified technique of insertion methods for data hiding using DNA sequences. International Conference on Automation, Control, Energy and Systems (ACES); IEEE; 2014.

[44] Das P, Kar N. A DNA based image steganography using 2d chaotic map. International Conference on Electronics and Communication Systems (ICECS); IEEE; 2014.

[45] Majumdar A, Sharma M, Kar N. An Improved Approach to Steganography using DNA Characteristics. IEEE; 2014. p. 138–145

[46] Das P, Kar N. A highly secure DNA based image steganography. International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE); IEEE; 2014.

[47] Chaudhary H, Bhatnagar V. Hybrid approach for secure communication of data using chemical DNA. 5th International Conference on Confluence the Next Generation Information Technology Summit (Confluence); IEEE; 2014.

[48] Ibrahim FE, Abdalkader H, Moussa M. Enhancing the Security of Data Hiding Using Double DNA Sequences. Industry Academia Collaboration Conference (IAC).

[49] El-Latif EIA, Moussa MI. Chaotic Information-hiding Algorithm based on DNA. International Journal of Computer Applications (0975–8887); 2015; 122 (10).

[50] Yamuna M, Elakkiya A. Codons in Data Safe Transfer. International Journal of Engineering Issues; 2015; (2): 85–90.

[51] Hamed G, et al. Hybrid technique for steganography-based on DNA with n-bits binary coding rule. 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR); IEEE; 2015.

[52] Marwan S, Shawish A, Nagaty K. An Enhanced DNA-based Steganography

Technique with a Higher Hiding Capacity. Bioinformatics; 2015.

[53] Manisha B, Mohit P. Double Layered Dna Based Cryptography. IJRET: International Journal of Research in Engineering and Technology; 2015; 4 (4):2321–7308.

[54] Chakraborty S, Bandyopadhyay KS. Data Hiding by Image Steganography Appling DNA Sequence Arithmetic. International Journal of Advanced Information Science and Technology (IJAIST); 2015; 44(44).

[55] Das P, et al. An Improved DNA based dual cover steganography. Procedia Computer Science; 2015; 46: p. 604–611

[56] Indora S. Cascaded DNA cryptography and steganography. International Conference on Green Computing and Internet of Things (ICGCIoT); IEEE; 2015.

[57] Tank RM, Vasava HD, Agrawal V. DNA-Based Audio Steganography. Oriental journal of Computer Science and Technology; 2015; 8:43–48.

[58] Santoso K, et al. Sector-based DNA information hiding method. Security and Communication Networks; 2016; 9 (17):4210–4226.

[59] Marwan S, Shawish A, Nagaty K. DNA-based cryptographic methods for data hiding in DNA media. Biosystems; 2016; 150:110–118.

[60] Marwan S, Shawish A, Nagaty K. Utilizing DNA Strands for Secured Data-Hiding with High Capacity. International Journal of Interactive Mobile Technologies; 2017; 11(2).

[61] Malathi P, et al. Highly Improved DNA Based Steganography. Procedia Computer Science; 2017; 115: p. 651–659

[62] Liu H, Lin D, Kadir A. A novel data hiding method based on deoxyribonucleic acid coding. Computers & Electrical Engineering; 2013; 39(4):1164–1173.

Chapter 7

Revealing Cyber Threat of Smart Mobile Devices within Digital Ecosystem: User Information Security Awareness

Heru Susanto

Abstract

In recent years, the number of mobile device users has increased at a significant rate due to the rapid technological advancement in mobile technology. While mobile devices are providing more useful features to its users, it has also made it possible for cyber threats to migrate from desktops to mobile devices. Thus, it is important for mobile device users to be aware that their mobile device could be exposed to cyber threats and that users could protect their devices by employing cyber security measures. This study discusses how users in responded to the smart mobile devices (SMD) breaches. A number of behavioural model theories are used to understand the user behaviour towards security features of smart mobile devices. To assess the impact of smart mobile devices (SMD) security and privacy, surveys had been conducted with users, stressing on product preferences, user behaviour of SMD, as well as perceptions on the security aspect of SMD. The results was very interesting, where the findings revealed that there were a lack of positive relationships between SMD users and their level of SMD security awareness. A new framework approach to securing SMD is proposed to ensure that users have strong protection over their data within SMD.

Keywords: smart mobile device, awareness, behaviour, cyber threat, cyber security, cyber crime

1. Introduction

Technology has been known to continually evolve since centuries ago, creating new innovations and infrastructures that changes how economies functions and overall improves standards of living within societies, and there have been no signs of it slowing down. One of the radical innovations within this century is the introduction of mobile phones. When mobile device was first introduced in 1973, mobile devices was a bulky communication device that was considered as a luxury good that aren't affordable to everyone and has limited features.

However, as the years goes by, emerging technology and innovation has successfully created an upgraded version of mobile phones which is known as smart mobile devices. Unlike mobile phones, smart mobile devices have more features to users, where it acts more than just a medium of communication, but as a device capable of storing data, capturing pictures of memorable moments and much more.

In other words, smart mobile devices had been delivering great number of benefits to everyone that it has deeply integrated itself within society's livelihood. Within this era of digitalization, more users are actively using their smart mobile devices to conduct activities such as paying bills online, managing their finances or do some online shopping as a direct result of the incremental upgrades that had been made which includes increased storage, power and speed. In addition, smart mobile devices are also a part of a large computing environment that encompasses a child's legacy, a user's persona and the keys to a user's home or digital life which includes any items in the house that is connected to the user's mobile device. But for every good thing that has been created, there's always a downside to it. New technology such as smart mobile devices also provides opportunities that could be reaped and exploited by malicious entities with ill intentions, hence exposing users and organisations to potential cyber threats such as hacking or data breaches [1–3]. According to a research made by comScore in 2016 (**Figure 1**), they have revealed that across the different age brackets of users in the UK and Canada, the average usage time of mobile devices is considerably higher than the average usage time of desktops.

A statistic developed by Broadbandsearch discovers that the total percentage of users accessing global websites through their mobile devices has increased exponentially within the last five years, ranging between the year 2013 to 2018. In 2013, the amount of combined web traffic from mobile devices was at 16.2% and the value has surged year by year, where by the combined web traffic was at a high value of 52.2% in 2018. Another research made by Statista (**Figure 2**) reveals that the number of mobile users will continue to increase within the next 5 years, where it estimated that the number of smart mobile device users will grow from 6.8 billion in 2019 to 7.33 billion in 2023.

Overall, as users spend more time with their mobile devices, these mobile devices would have accumulated and stored tremendous amounts of private data and information that it eventually becomes an attractive target for malicious groups or entities such as hackers. McAfee in their "Mobile Threat Report 2019" discovered that these cyber criminals will keep creating tactics to bypass mobile security in order to execute their cyber threat activities for the sake of one common goal, which is to maximise their income and profits.

There are various cyber threats that are affecting smart mobile devices and one of the cyber threats commonly found in mobile devices is malware. Malware had been infecting mobile devices for more than a decade and the increasing number of mobile devices infected by it is certainly very concerning. **Figure 3** shows that malware cases has increased by approximately 310% since 2016, thus reinforcing the undeniable fact that these malware authors have continued to adapt and create

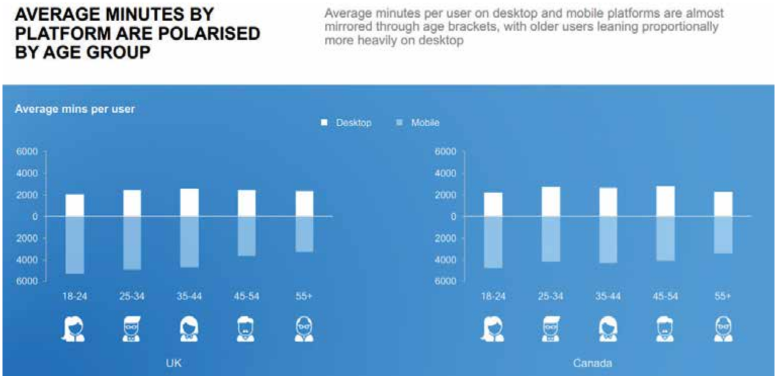


Figure 1.
Average usage time of Mobile devices.

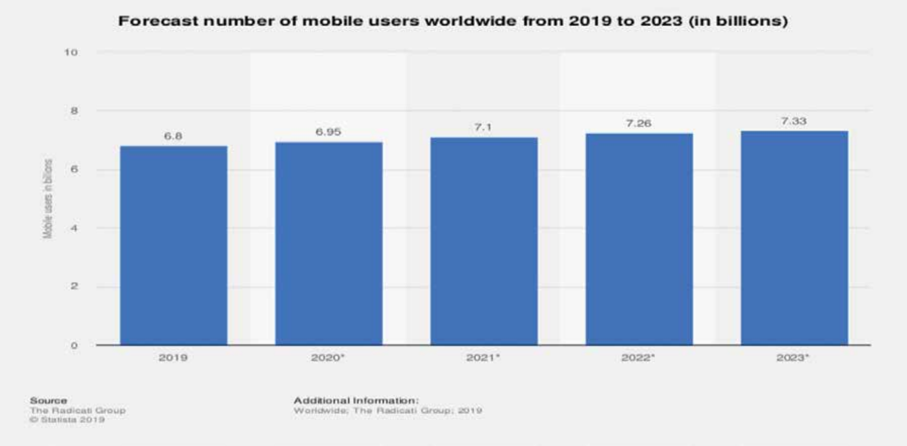


Figure 2.
Number of Mobile users.

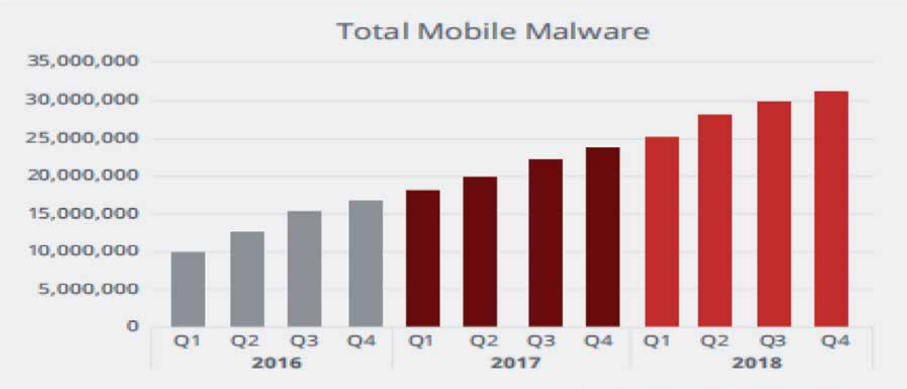


Figure 3.
Mobile malware.

new tactics against any challenges it encounters [2]. Malware, in the face of cyber security and countermeasures, instead has become more serious, dangerous and harder to detect, and it shows no signs of decline. Hence, all of this shows how important it is for users all over the world to protect and keep their smart mobile devices secure and safe from becoming a victim or prey to cyber criminals through implementation of cyber security in mobile devices.

Significance of the study.

Various researchers have highlighted the importance of protecting sensitive data from cyber threats by implementing cyber security measures within smart mobile devices. Awareness of the risk of mobile threats invading and stealing personal information from their mobile devices and the methods of mitigating this risk through mobile security methods.

Aim and objectives of the study.

The aim of this study is to measure the level of smart mobile device security and privacy awareness. The specific research objectives are shown as follows:

1. To reveal the different types of smart mobile devices usage.

2. To explore the attitudes of users towards smart mobile device security

3. To discover the category, costs and impact of cyber threat incident on smart mobile devices

4. To propose a new framework approach to securing SMD

5. To ensure users have strong protection over their data and the type of cyber security required to combat cyber threats on SMD comparing with security standard.

Structure of the study.

This first section has outlined the background, significance of the study, the aims and objectives of the study as well as the limitations of the study. The rest of the paper will be structured in the following way. The second section will present the literature review related to cyber threats and cyber security. The third section will discuss on the methods utilised in collection of data. The fourth section will be the discussion on the survey findings and how it relates to SMD security. Finally, the last section will be on the recommendation of a new framework as well as conclusion.

2. Literature review

The topic of cyber threat and cyber security on mobile devices had been greatly debated by various researchers around the world. Thus, this section will be reviewing numerous literatures on cyber threats and cyber security in the context of mobile devices as follows:

- Definition of a mobile device

- Cyber threats in mobile devices

- Cyber security in mobile devices

- Related works

Definition of a mobile device.

Mobile device can only be defined when the two aspects such as the software and hardware aspect are explained together [3–5]. A mobile device will usually have a small form that has a non-removable data storage and are equipped with an operating system that is particularly different from the operating system of laptops or desktops. A mobile device should be equipped with at least one wireless network interface such as cellular network or Wi-Fi for the purpose of connectivity and communication. Also, a mobile device should be able to obtain and install applications through various ways such as app stores, websites or other third party sources. There are also other common features of a mobile device but are optional such as the ability to connect to multiple wireless or area network interface and the ability to connect to real-time location services through the use of Global Positioning System (GPS). Other features of a mobile device also includes the microphone which allows voice to be recorded, a built-in camera that allows mobile device to record or capture pictures as well as a removable data storage.

Cyber threats in mobile devices.

Mobile threats can be classified into four different categories of mobile threats which are (1) application-based threat, (2) physical threat, (3) network-based threat and (4) system-based threat. The first category is application-based threats. Outdated or unpatched third-party applications in mobile devices poses risks as hackers may exploit the vulnerabilities within those applications. Users that are still using an old mobile device that has a lack of software updates, an untimely patch update or a cease in support for older operating systems are at risk of being compromised by hackers through the holes within the software or the OS itself [4, 6, 7]. Additionally, there are various application-based threats which consist of (a) Malware, (b) Spyware, (c) Privacy threats and (d) Vulnerable applications.

a. Malware is referred as malicious software that is operated by hackers to obtain access to a mobile device and perform illegal criminal activities. It requires the hacker to install malware into the mobile device through many devious ways that are very difficult to track or trace. Malware could be used to alter or execute actions without the owner's permission, such as sending prompt or subitaneous text messages to contacts, charging phone bills or acquiring successful control over the mobile device.

b. Spyware is known as a program used by hackers which utilises private or confidential data without consent for illicit motives, whereby it usually targets sensitive information such as owner's list of contacts, phone call records, text messages, real-time location, gallery images, browser history and email addresses.

c. Privacy threat could occur when hackers alters or erase the mobile device's data using a software applications that are not somewhat program codes. The sensitive data within the mobile device are visible to the attacker and can easily be exploited for different purposes that are ill-natured.

d. Vulnerable application refers to applications that have holes that could be exploited by hackers to sneak into the mobile device and gaining full control over the mobile device. Once control over the mobile device is established, hackers can easily acquire personal or sensitive information, execute unfavourable activities against user's will such as rendering certain services useless and force download unknown applications without authorization.

The second category is physical threat. The authors refer physical threat as a security incident which involved a mobile device being stolen or lost in the process. Mobile devices have a greater chance of being lost or stolen than laptops due to the features of mobile devices such as its small size, lightweight and easy to carry, thus making it the perfect target of attackers. An attacker that had successfully obtained physical access to a mobile device will proceed with other malicious activities that are conducted from the attacker's computer. The mobile device will display an image of a malicious system that is trying to install a harmful software or attempting for data extraction. Further added that mobile devices could be misused by attackers in several ways such as creating a fake identity by using the personal information contained inside the mobile device or by selling sensitive or confidential data to the black-market for profit motives [7, 8].

The third category is network-based threats. Most of the mobile devices used by consumers are usually connected to wireless network interface such as Wi-Fi or Bluetooth, the use of these network interfaces carries certain risks. It makes mobile devices vulnerable towards malicious activities such as wireless eavesdropping that is performed using off-the -shelf software such as Aircrack-ng Suite or Wifite. Attackers could exploit the network to plant malwares on mobile devices unnoticeably whenever a mobile device is connected to a wireless or cellular network. Once the malware has been installed, it will give attackers free access to the mobile device, allowing them to modify or extract any confidential or sensitive information within the mobile device. Also, attackers can use the method of Wi-Fi sniffing to conduct criminal activities by reading, monitoring or altering any unencrypted data that is travelling in the same network [9–11]. Mobile device manufacturers introduce unintentional flaws or vulnerabilities into their own devices such as the incident with Samsung's Android SwiftKey keyboard which was discovered to be susceptible to eavesdropping attempts. Another similar incident occurred with Apple devices, specifically the iPhone's Operating System (iOS) where the "No iOS Zone" flaw causes any iOS devices within range to automatically connect to a malicious fabricated network and constantly crashes those devices. In addition Web-based threats are known as threats that involve user's interaction with online services through the access of the Internet, which could be divided into smaller categories which are (a) Phishing scams, (b) Drive-By downloads and (c) Browser exploit [5, 8, 12, 13].

a. Phishing scams happens when users are being delivered through their email, text messages or social media links that appears to originate from a legitimate company or organisations when in reality it is a scam. The main purpose is to trick users, individuals or organisation, into disclosing sensitive or confidential information such as debit/credit card number or passwords.

b. Drive-by-downloads occurs when a hacker obtains illegal access to a mobile device as a result of a user opening up a web page or clicking on a link found on a website. It will then trigger an automatic download of malicious applications that wasn't consented by the user.

c. Browser exploit is described as a devastating code that allows hackers to exploit the unsecured data within the mobile operating system. It could also be described as malicious software that aims to alter a mobile browser's settings without any consent that is usually triggered when a user had visited unsafe websites.

Another research classifies cyber threats into two different aspects which are the technical aspect and the management aspect of mobile security. The technical aspect of mobile device cyber threat are quite similar to what was described by previous researchers, where it consisted of device security threats, network security threats, services security threats and content security threats. However, there was one factor that wasn't touched on in the two previous research but was present within this research work, and it was concerning on the management aspect of cyber threat in mobile devices [1, 14, 15].

The management aspect of mobile device cyber threat studies the threats that are associated with the security policy of mobile devices, which can be broken down into three categories namely (a) application distribution environment security threat, (b) law institutional security threat and (c) domestic and foreign enterprise environment security threat.

Cyber security in mobile devices.

However, the study revealed measurement of users that may possibly undertake in order to protect their personal data stored inside their mobile device [16–18]. One of it is by using password or PIN lock features to ensure that only the user can access the device and prevent outsiders from accessing it. Also, users should only connect their mobile device to wireless networks that are protected by a password and avoid connecting the device to public networks as public networks raises the chances of the user being compromised. Users that have their devices with Bluetooth enabled should set it to non-discoverable to other users so that attackers will not be able to sneak in and steal the user's sensitive data and the user's contact number should never be revealed easily to other people as it might be used to execute ill-intent activities.

Basic steps that users can exercise to protect their mobile devices from cyber threats such as (1) Regular or prompt update of operating system, (2) Device rooting or jailbreaking prevention, (3) Mobile applications management and (4) Mobile antivirus.

1. Regular or prompt update of operating system - when mobile devices run on outdated operating system (OS) such as Android or iOS, these devices are much more vulnerable to cyber attacks, such as the entry of malicious applications into the mobile device. This situation could have been prevented if the latest operating system had been updated on time and without delay.

2. Device rooting or jailbreaking prevention - When users decided to root or jailbreak the operating system on their mobile device for certain personal reasons, users should remind themselves the gravity and consequences of it because at the moment they do so, the responsibility of the privacy and security of their mobile device have transferred from the developers to the users themselves. Users should also be informed that cyber threats such as spyware are more likely target devices that are rooted or jailbroken.

3. Mobile applications management - Users should install mobile applications from trusted and secure source such as Apple store or Google store and avoid installing from untrusted sites from the internet. By downloading applications from trusted sources, users do not need to worry about security as the applications are scanned for any vulnerabilities before installed and the installed applications will automatically be updated to fix any vulnerabilities in the future.

4. Mobile antivirus - Installing a mobile antivirus may seem ineffective for Apple devices as Apple ensures that it will not be allowing any applications from gaining any permission it needs to execute any damage. It may seem redundant to install a mobile antivirus in Android devices as Android restricts any app installation from sources other than Google store but for users that tend to install applications from outside sources, an antivirus will protect the device to some extent from unknown threats originating from the installed applications.

Moreover, a set of security solution was proposed as function as it can be implemented by organisations and enterprises to manage mobile device security [19–22]. The first solution is by creating a general policy that includes the restrictions on the use of mobile devices within the organisation such as restrictions on user access and application access tools and hardware such as cameras, removable storages such as USB flash drive and hard disk drive (HDD) as well as to local OS services, for instance inbuilt email, web browser, contact and calendars. The policy also includes guidelines on the management of wireless network management such as Wi-Fi or Bluetooth and additionally limits personnel's access to organisation's services based on the mobile device's brand, model, software client version and OS status (ensures device is not rooted or jailbroken). Any suspicious actions will be monitored, detected and reported back to the management and once it has been found that the actions has violated the general policy, further actions and reprimandation will automatically take place accordingly.

The second solution concerns on the data storage and communication within the organisation. The management should strongly encrypt organisation's confidential data that are contained within the built-in storage as well as the removable media storage and any device that will be reissued to other personnel must first be wiped to clean the data previously stored in it. Additionally, if any of the organisation's device is assumed to be lost or stolen by unknown instigators that by any chance cannot be trusted, the management should initiate remote wipe on the device to prevent confidential information from being harvested by malicious attackers [16, 23–25]. Another way to prevent the mobile device from being accessed illegally is by implementing a configuration that has wipe feature within its devices that will automatically factory-resets all the data within after it detected several failed authentication attempts. The organisation should also aim at having a secure data communication between organisation and mobile devices by encrypting it using Virtual Private Network (VPN) or other encryption tools that suits their needs.

The third solution is based on the device and user authentication. A user authentication step should be implemented before any personnel could access the organisation's data and resources, which could be in the form of password or other various authentication such as token-based or domain authentication. The organisation should also include certain parameters for password characters, password length and the maximum number of retries allowed before the device is locked out or wiped. In cases where a user has requested a password reset or was locked out of the mobile device, the administrator should be able to restore the user's access to it remotely. Any device that is suspected to be accessed in an unsecured location should be remotely locked under the supervision of the administrator and any device that is in an inactive state for a certain period of time should be locked automatically by the device itself.

The final solution involves restrictions on various aspects of mobile applications. The management should restrict the list of app stores that can be accessed by personnel to download mobile applications or instead, the management could issue applications from a chosen application store. In addition, the installation of certain applications should also be restricted through the process of whitelisting and blacklisting. There should also be a restriction on what device location are permissible for the application to access such as storage access or camera access. The digital signatures found in applications should be verified to ensure that the applications installed are from a safe and trusted source and that the code wasn't altered in any way.

3. Methodology

The evaluation method which has been utilised by countless researchers in obtaining research data known as the questionnaire method was implemented in this study. A random sampling method has also been chosen and implemented as a method of collecting the research data in this study. The nature of questionnaires asked will be focused on the topic of cyber threats, cyber security and its relationship with SMD. Through the employment of the random sampling method, a set of questionnaires have been distributed within the duration of approximately three months to the targeted group of respondents. Other platforms as well as social media had also been utilised to distribute the online survey such as WhatsApp and Instagram. The target respondents of this study are focusing particularly on the youths which include the generation-Z strictly. This particular group of respondents have been chosen as they represent the majority of mobile device users that are technologically literate. A variety of respondents with different gender, background and educational level had taken part in the study. A total of 109 respondents have participated in the online questionnaire where almost all of the respondents are within the age range of 20–29 years old which matches the targeted group of respondent previously mentioned before.

4. Findings and analysis

The data analysis will be made according to each of the section that has been created within the survey questionnaire as follows:

- Demographics
- General section
- Password security
- Application security
- Email and Account security
- Personal security
- Knowledge and Attitude towards mobile security

Demographics.

In this section, the questions asked the respondents about their gender, their age group, their current status as well as their present educational level.

Referring to **Figure 4**, out of 114 respondents that participated in the survey, 67.5% of them are female respondents and consequently 32.5% of them are male respondents, thus highlighting that a majority of the respondents are female. When looking at the age range of the respondents who have answered the survey questions, a large number of them are within the age range of 20–29 years old which contributes to a high 83.3% of total respondents. The rest of the respondents originated from two other age groups where 15.8% of the respondents are aged below 18 years old while the remaining percentage are within the age group of 30–39 years old.

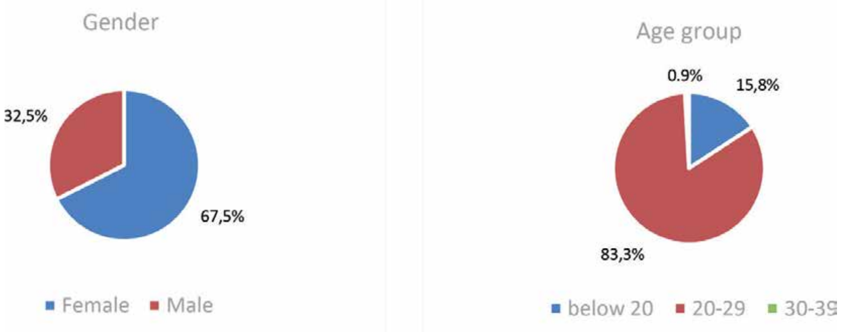


Figure 4.
Respondents demographic.

It has been observed in **Figure 5**, that amongst the respondents, 49.1% of them are students from Universiti Teknologi Brunei, 13.2% of them are students from Universiti Brunei Darussalam, 5.3% of them are from Politeknik Brunei and 11.4% of the respondents came from various other public or a private higher institutions such as Institute of Brunei Technical Education (IBTE), Laksamana College (LCB), Cosmopolitan College of Commerce & Technology and Micronet International College. Additionally, the survey also received responses from non-university students where it comprises of 7.9% from high school students, 7.9% from the working population as well as 5.3% from the unemployed population.

General Section.

In this section, the questions that were asked were focused on finding out the type of mobile device the youths are generally using, their general purpose of using a mobile device as well as the frequency of internet connectivity amongst the youths.

Figure 6 shows the questions that were asked within the general section of the survey questionnaire. When respondents were asked about the type of smart mobile device they are currently using, there are 86 respondents that uses Android devices which contributes to 75.4% of the chart, 24 respondents that uses Apple devices which contributes to 21.1% of the chart and a small number of respondents which is 4 respondents uses both Android and Apple devices thus contributing to 3.5% of the chart. An assumption was made in this survey whereby every respondents that answers the survey has at least one mobile device, which is the reason why the question directly asks its respondents the type of mobile device used. It can be seen

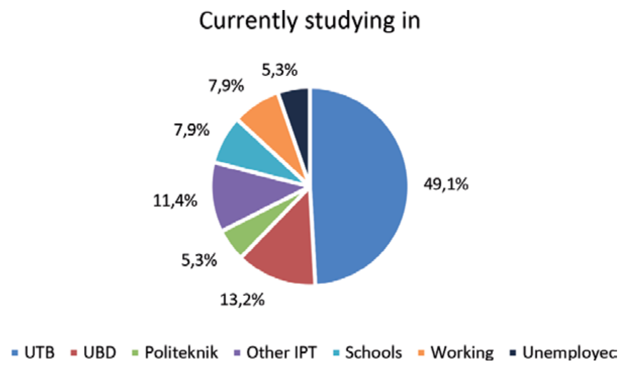


Figure 5.
Education background.

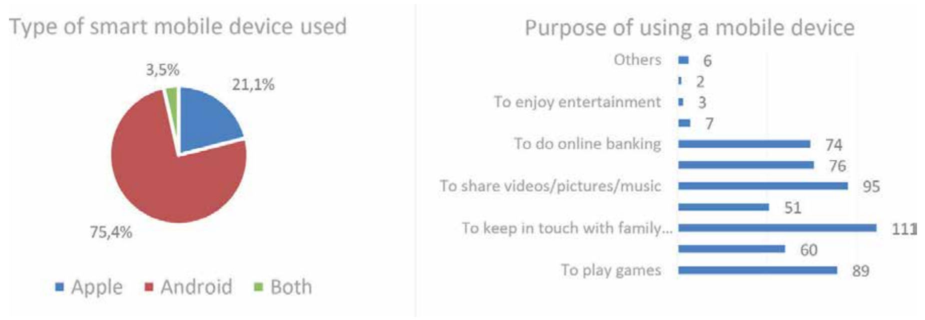


Figure 6.
Smart Mobile device usability.

that majority of the respondents favours the android devices compared to the apple devices which is proven by the results shown on the charts.

Also, when the respondents were asked about their general purpose of using a mobile phone, the respondents have chosen various purposes as the survey questionnaire allowed them to choose more than one purpose. The chart shows that 111 out of 114 respondents which agreed that one of the main purpose of using a mobile device is to keep in touch with family and friends. The chart also showed that 95 respondents uses a mobile device to share videos, picture or music, 89 respondents uses their mobile device to play games, 76 respondents uses their mobile device to make online transactions, 74 respondents uses their mobile device to perform online banking, 60 respondents utilises their mobile device to make professional and business contacts and 51 respondents uses their mobile device to create new friends.

There were also some minority purposes chosen by the respondents where 7 respondents believes their purpose of using a mobile device is to go on social media platforms, 3 respondents uses their mobile device to watch entertainment, 2 respondents uses their mobile devices to take pictures and lastly, one respondent each believes that their purpose of using a mobile device is to either surf the internet, read news, listen to radio, download video, create digital notebooks or doing some phone modification. This means that most of the respondents feel that it is safe to use their mobile device to do important activities such as maintaining communication as well as making online/bank transactions. It also acts as an indicator that the respondents felt it is secure enough to send and share videos, music or pictures amongst themselves and their friends through their mobile devices.

Figure 7 shows that when the respondents are asked about how frequent they are connected to the internet, 100% of the respondents agreed that they are constantly connected to the internet and when asked about how they are connected to the internet in which they are given three choices, 95.4% of the respondents are connected through the Wi-Fi medium, 93.6% of the respondents are connected through their mobile data (cellular connection) and 25.7% of the respondents are connected through the use of hotspot. Another assumption was made while doing the survey which believes that almost all respondents are constantly connected to the internet and this assumption was proven through the survey results whereby 100% of the respondents stated that they are frequently connected to the internet. It is inevitable for the users of mobile device to be constantly connected to the internet because within this technological era, the only way to maintain communication and receive information is through the use of internet.

Password Security Section.

In this section, the questions asked were focused on discovering respondent's behaviour and habit in regard to the security of their mobile device and the network they are using to surf the internet.

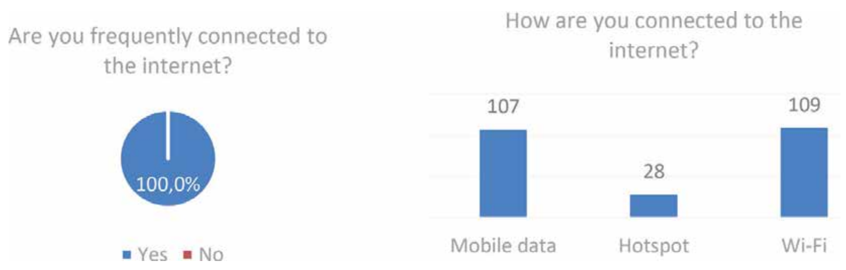


Figure 7.
Internet connectivity types.

In **Figure 8**, when the respondents were asked about how frequent their network password are changed, 43.0% of the chart which accounts to 49 respondents rarely changes their network password, 26.3% of the chart which accounts for 30 respondents changes their password sometimes, 23.7% of the chart which accounts for 27 respondents never changed their password, 4.4% of the chart which accounts for 5 respondents frequently changes their password and finally, only 2.6% of the chart which accounts for 3 respondents always changes their password.

It is important to change the network password regularly as it is one of the simple measures to avoid people from silently stealing the user's network data unconsciously. Few of the dangers of not changing the network password regularly is that the users might be exposed to network attacks such as sniffing or eavesdropping and there might also be illicit entities or hackers that had previously obtained the password to enter the network, hacked into the user's network and using the network to perform unlawful actions.

In **Figure 9**, when the respondents were asked about what type of security measure they have implemented to their mobile device, the responses received were divided into few categories. About 40.4% of the respondents uses fingerprint protection only, 24.6% of the respondents uses password protection only, 11.4% of the respondents employs solely pattern protection, 2.6% of the respondents uses face protection measure and 5.3% does not employ any kind of protection measure. There were also respondents that utilises multiple protection measures for their mobile device where 12.3% of the respondents uses a combination of two protection measures and 3.5% of the respondents uses a combination of three protection measures.

It can be seen that many users tend choose fingerprint lock compared to other security measures such as password or pattern. It is considered as the best option because fingerprint is a unique identifier of each individual person and since it is

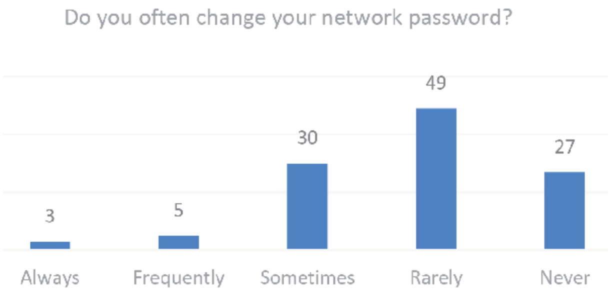


Figure 8.
Changing the passport.

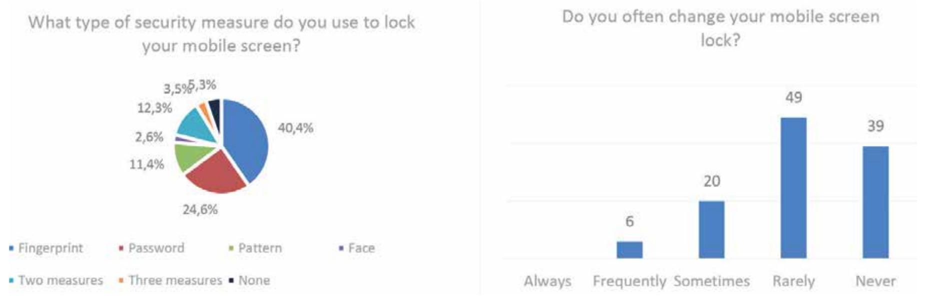


Figure 9.
Security measuring.

hard to obtain someone else's fingerprint, it makes it difficult for attackers to gain access to the mobile device. Password and pattern are also good measures of security lock for mobile devices, but it has a disadvantage. As all smart mobile devices have touch screen input, any act of accessing the mobile device by using the pattern or password lock will inevitably leave smudges or residues of the screen which could be used by attacker to retrace the pattern and access the device. But then again it will take the attacker some time to figure out the pattern correctly, hence it is still better for a mobile device to be protected by a security measure rather than having none at all in order to reduce the risk of being breached.

The final question asked within this sub-section was aimed at discovering how frequent does the respondents change their mobile screen lock or protection measure and it was seen that 43.0% of the respondents which contributes to 49 respondents rarely changes their mobile screen lock, 34.2% of the respondents which is equivalent to 39 respondents never changed their mobile screen lock, 17.5% of the respondents which contributes to 20 respondents alter their mobile screen lock sometimes and 5.3% of the respondents which is equivalent to 6 respondents frequently changes their mobile screen lock.

It can also be see that many respondents have never change their mobile screen lock or rarely do so. The act of changing the mobile screen lock regularly is particularly important to users that implement password and pattern lock measures. It will reduce the user's risk of being breached physically or virtually and if the worst case comes whereby an attacker that aims to breach the device had figured out some of the correct password or pattern, the process of regularly changing the lock screen will ensure that these attackers will fail in their attempt.

Application security section.

In this section, the respondents were asked on questions that were inter-related in nature that aims at revealing respondent's behaviour as well as awareness towards the security of applications.

In **Figure 10**, when the respondents were asked about whether the respondents have installed any mobile applications from unknown sources, 71.9% of the respondents agreed that they have installed applications from unknown sources while the remaining 28.1% of the respondents have never installed applications to their mobile device from unknown sources. When users downloads applications from unknown sources, it means that users are downloading third party applications from third party app stores rather than the official stores such as Google or Apple store. Third party applications are known for being risky because these applications has been created by other creators or programmers and not made by the manufacturer of the mobile device or the operating system of the mobile device. Basically these applications cannot be guaranteed safe and secure for use or free from malware by the mobile device's manufacturer as they came from unknown sources.

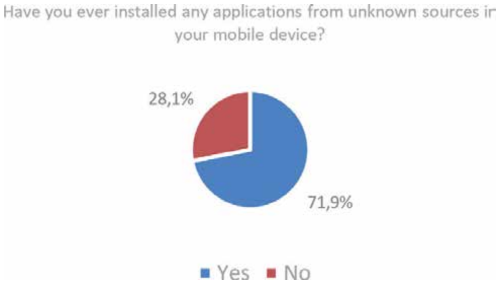


Figure 10.
Apps installation from unknown sources.

Figure 11 shows two charts which reflect the in-depth questions that were aimed towards the security aspect of applications, where one of the questions asked whether they have read the End User-Licence Agreement (EULA) and privacy policy before installing any applications. The chart shows that 56.1% of the respondents have never read it, 36.8% of the respondents have read the policy sometimes and the remaining 7.0% of the respondents always reads the policy prior to installing any applications to their mobile device. When users are being prompted to install any application, there will usually be a window which request the users to deny or agree to the agreement stated in the application's privacy and policy which includes the EULA policy. It is very important for users to read these policies because some these policies will state the purpose and the time period for using the user's personal data and information as well as how users are supposed to use their manufacturer's applications without breaking any of the policies.

The respondents were also asked on whether they have read the application's phone access permissions before installing any application and it was revealed that 46.5% of the respondents always reads it before installing any applications, 36.0% of the respondents reads it sometimes while 17.5% of the respondents never reads it. Prior to installing any application to a mobile device, the application will request the user's permission to access certain folders or areas within the mobile device such as the camera, storage, location and more. Before accepting such permissions, users must first read the "phone access permission" carefully so that can evaluate for themselves whether it is safe to do so instead of just accepting any permissions because there might be instances where some applications requested certain permission that it does not necessarily needs. The act of just accepting any permissions that prompt up could lead in the user handing over their information willingly and unknowingly to shady application developers or fraudulent data miner, which could further result in the exposure and breach of the user's personal information.

Email and Account Security.

In this section, the questions asked were aimed at measuring and assessing respondent's tendency as well as awareness towards the security aspect of email and accounts.

In **Figure 12**, when the respondents were asked a question on whether they would initially check the authenticity of the sender before opening the attachment received. The second chart shows that 67 respondents which would always check the authenticity of the sender before opening the attachment, 41 respondents would sometimes check for the authenticity while the remaining 6 respondents had never checked the authenticity of the sender. This is a good indicator that shows users are taking precautionary measures to protect themselves from harmful attacks that are being orchestrated through the medium of emails and even text messages. These

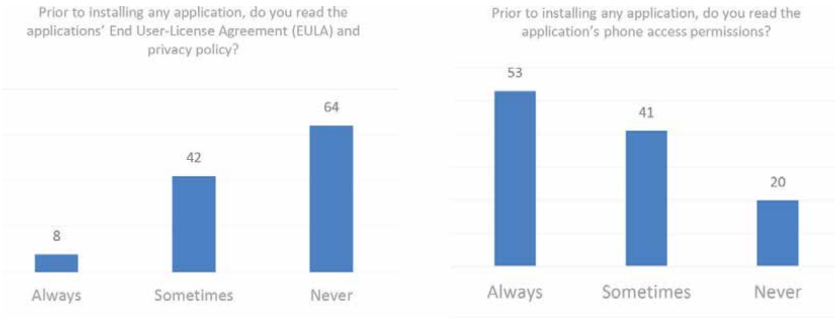


Figure 11.
Security aspects of apps.

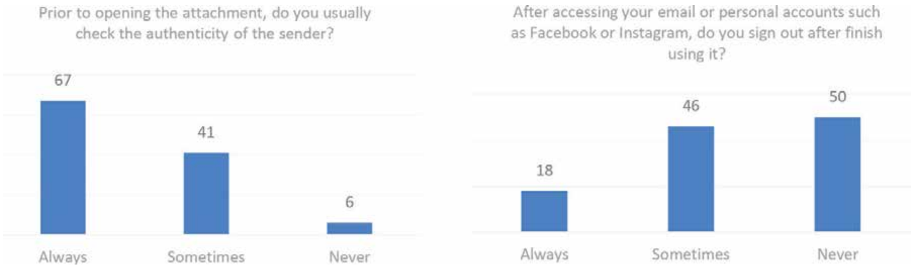


Figure 12.
Sender authenticity.

attackers could be sending emails or attachments that may seem to be from a legit sender such as official corporations initially but it is a scam that aims to trick the users to open, download or click the sent attachment or link. When users successfully downloaded it, the virus will start to spread to other emails or contact list thus as a result endangering other users as well. Hence, this behaviour of checking the sender serves a good measure to protect user's mobile device and the sensitive information contained in it from being harmed.

Additionally, when respondents were asked whether the respondents sign out from their email or personal accounts after using them, the graph showed that only 18 out of 114 respondents always logs out from their accounts after use wile 46 respondents does log out from their accounts sometimes and 50 respondents had never signed out from their accounts after using them. Most users that accessed their personal accounts through their mobile device tend to tend to stay logged in rather than logging out after using them. This might pose risk and dangers to user's personal data and information because if someone unknown such as an intruder gained access to a user's mobile device such as in the case of mobile device theft, these intruders could easily access the user's personal accounts as they are already logged in and it makes it possible for the intruder to steal the user's identity this way. Hence, the behaviour of logging out from personal accounts after using them is a best practice to ensure that the user's sensitive and personal data are constantly protected from any possibilities of malicious actions.

Personal security.

In this section, the questions that were asked to respondents were focused on the security aspect of respondent's personal data in their mobile device and their action in securing it. The result shows that 66.7% of the respondents stores their confidential or sensitive data within their mobile device while 33.3% of the respondents does not store any within their mobile device. Many users considers their mobile as an item that is very near and personal to their owners and it is also regarded very crucial as it usually kept the user's sensitive or confidential information which is proven from the results stated above. Due to the mobility of mobile devices, users tend to keep their confidential or sensitive information within their mobile devices since it allows users to access is much faster compared to other means. It is necessarily not wrong for users to keep their precious data within their mobile devices, but it is recommended for users to set up the most optimum level of security measures to their mobile devices in case it is under a threat of being compromised or breached by malicious entities. Then, when the respondents were asked next on whether they have installed any security software such as anti-virus within their mobile device, only 39.4% of the respondents installs a security software within their mobile device while the remaining 60.6% of the respondents do not equip or install any security software to protect their mobile device.

From the results, it is observed that many users believes that installing security software or applications such as anti-virus is not crucial or important for their mobile device. Thus, when these respondents were asked on their main reason for not installing any security software, various answers have been received where one of responses mention that their mobile device already had a built-in antivirus, hence the absence of need to install another anti-virus. There are few similar responses that were made by different respondents where some of it mentions that some respondents did not know the existence of an anti-virus for a mobile device and some had problems choosing an anti-virus that is trustworthy to be installed. Some of the respondents also stated that it is not necessary to install an anti-virus because they believed that they have not installed any malicious programmes or applications to their mobile device. A few of the respondents are hesitant to install a security software due to the need to pay for it while some are blatantly has no desire to install one at all as they deem anti-virus as unnecessary.

Various kind of behaviours in these responses reflects that many users are having a lack of awareness and knowledge on the importance and benefits of having an anti-virus within their mobile device which is worrying in general. For instance, users are hesitant to install an anti-virus software because they assumed that the built-in anti-virus is sufficient and it makes users think their mobile device is safe enough. But it still does not justify the reason for not installing any security software because there is no operating system that is completely secure and invulnerable from cyber risk and danger. Malicious programs, applications, viruses or malwares could still infiltrate mobile devices through every possible way which is why it is recommended for users to set layers of protection for their mobile device instead of being negligent and over-reliant on the built-in protection within their mobile device.

Then, a question was also asked to the respondents that have installed anti-virus on their mobile device on whether they regularly update their mobile security software. The responses that were received was surprising because out of 44 respondents, 18 respondents relied heavily on the auto-update feature of their mobile device to update their anti-virus, 16 respondents regularly updates their anti-virus and the remaining 10 respondents rarely updates their anti-virus. This indicates that there are over reliance amongst users towards the automatic update function in their mobile device's operating system or the anti-virus itself. Users that rarely updates their anti-virus or only relying on automatic update are usually the type of users that assume that it is "good enough" to have an anti-virus that prevents harmful activities being done to their mobile device.

Knowledge and attitude towards mobile security.

Within this last section of the survey questionnaire, the questions asked are aimed at measuring the respondent's level of knowledge towards mobile security as well as their attitude on the subject of security within mobile devices.

In **Figure 13**, when the respondents were asked on whether they have experience any privacy or security breach on their respective mobile devices, 97 respondents which contributes to 85.1% of the chart had never experienced anything similar before while 17 respondents which is equivalent to 14.9% of the chart have experienced a privacy or security breach on their mobile device beforehand. It can be assumed that these 17 respondents or users might have become victim to privacy or security incidents due user's lack of security measures implementation to their data and mobile device. Thus, due to a prior experience in a privacy or security breach, these users might have increased their knowledge on this matter and started tightening their security measures in their effort to prevent the incident from happening again. On the other hand, it also shows that most of the respondents have never experienced such incidents which might be due to sufficient implementation

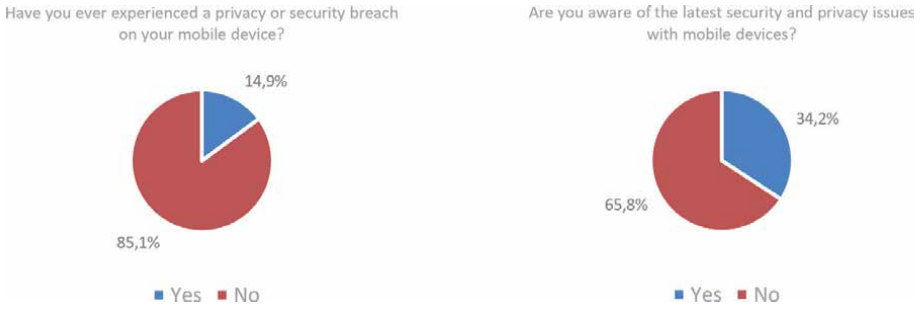


Figure 13.
Smart Mobile devices security breaches.

of some security measures to protect their mobile device and their personal data. But it is gravely reminded for users not become negligent and starts lowering down their security efforts because attacker will always look for those small opportunities to perform malicious activities.

Then, when the respondents were asked on whether they are aware of the latest security as well as privacy issues that is occurring to mobile devices, 65.8% of the respondents are apparently not aware of any mobile security issues while 34.2% of the respondents are aware of the security issues that are trending nowadays and happening to mobile devices. This means that most of the mobile device users are not educating themselves with the latest security incidents and happenings associated with mobile devices. It is a fact that many cyber crime related cases such as privacy and security issues are not being publicised and the society rarely hears anything about these issues but it does not mean that users should not take any initiative in educating themselves especially within this era of digitalization. It is important to be updated with such matters because the information gained by reading, researching and knowing how the security issues happen could turn out to be useful to users. Users which had beforehand known of such incidents could equip themselves with the useful information and when users are encountering a similar incident, users knows the know-how to handle such matters and not be tricked by the scheme created by attackers.

Additionally, the respondents were asked on whether they would be willing to use an application that have previously suffered a privacy or security issue. The chart in **Figure 14**, shows that 46.5% of the respondents might be willing to use such applications, 49.1% of the respondents were not willing to use such applications and the remaining 4.4% of the respondents are willing to use applications that have previously suffered a privacy or security issue. When the respondents were further asked as to why they are not willing to use them, various respondents provided similar answers. Some respondents believed that an application that has been breached or hacked is not secure enough for users to use, some have lost their trust to use a breached app and many respondents believed such application is not secure at all and prioritise over their desire to protect their data and privacy.

A similar in-depth question was also asked to respondents as to why they might give it another chance to use applications that previously had security or privacy issues, there are a number of responses received from various respondents that were similar in nature. Many respondents stated their desire to use the app once the issue had been fixed, resolved or patched while some respondents believed they will do so depending on several factors which are the availability of the app, the necessity of the app, the rating of the app and the severity of the security issue that occurred before. Several respondents also highlighted that they will only use the app if they know or were informed of the true cause of the issue, for instance a security breach that occurred might not have been caused by the developer itself but by the users of

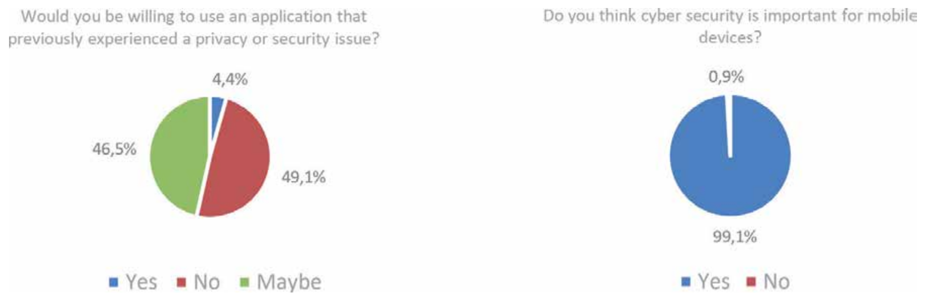


Figure 14.
Apps issues towards security.

the application itself. If it was in such instances, then these respondents are willing to use the application once more.

Lastly, when the respondents were asked on whether they think cyber security is important for mobile devices, 99.1% of the respondents agrees on the importance of cyber security for mobile devices while 0.9% of the respondents disagrees and believes that cyber security is unimportant for mobile devices. Almost all respondents believed that cyber security is important to mobile devices because the respondents believed that cyber security will protect their personal or sensitive data, their privacy as well as confidentiality that were available within their mobile device from being manipulated, misused or taken advantage of. The respondents also believed that the presence of cyber security is the most effective way to fight against cyber threat issues and reduce the number of cyber crime cases throughout the world.

5. Discussion

From the analysis of the survey results in the previous section, it can be seen that there are still mobile device users that aren't taking cyber security measures seriously and continued to stay negligent towards the dangers of cyber threats occurring to mobile devices. Number of reasons behind user's behaviour of not implementing any security measures in general such as:

- User's habit of making irrational thoughts or decisions such as clicking "I accept" without reading what they are actually agreeing to and not contemplating about the consequences of their behaviour

- User's extreme preference for convenience rather than taking the more difficult method namely security

- User's extreme priority on fulfilling user's desire rather than going for security such as downloading applications that they deem very important when there are alternatives that could be a much secure

- The financial costs of opting for security such as purchasing security software such as anti-virus is much greater than the security gains felt by the user.

- Users felt that the level of effort required to fully exercise security measures is too high such as remembering different passwords for different accounts and keeping anti-virus updated regularly.

- Users do not perceive any benefit and believes their behaviour will not affect security at all or users instead justify the cyber risk they perceived such as believing connecting themselves to an unsecure website for a short period of time is a safe action or by thinking there is no possibility of them being attacked or breached

- Users are lacking the knowledge and skills to handle any security issue such as ways and method of handling any fraudulent activities they encountered

- Users do not understand that any behaviour they have conducted will have an impact on the security risks as well as their level of vulnerability to these risks

- Users are simply forgotten to take on security measures due to various distractions encountered while surfing online

In order to increase the level of awareness amongst mobile device users, it is necessary to increase their awareness on how each of their actions and behaviour could affect the security of their mobile devices. In order to do so, a new framework which has been proposed which has been created by analysing and assessing a number of factors that influences user's cyber security behaviour.

Environmental influencers: Design factors.

It has been discovered that creating a good design or interface for a security software or application has an impact on user's willingness to use such applications. Interface design is deemed as a crucial property compared to user participation in regards to security systems in computer. The rationale behind this discovery is that a good design can effectively transmit the correct information in an orderly manner to users, thus allowing users to make an accurate and precise decision in regards to the system's state, structure as well as the security aspect of it. When this is applied to the context of mobile devices, having a good design such as good visualisations and smooth interface allows the security applications to effectively communicate its users with the necessary information useful for users to assess their current security status and risk as well as making informed decisions. Additionally, it can also promote constant interactivity with users which will then eventually result in enticing user's involvement and willingness to use such security applications.

Economic factors.

When users are determining themselves on how to behave, they will usually conduct a cost benefit analysis on the situation. One of the factors that could affect the analysis and consequently their behaviour is the presence of incentives, whether they are in the form of positive incentives such as rewards or any sort of benefits bestowed to the users or the incentives could also be negative in nature such as the cost or punishment for certain conduct or behaviour. When the economic factors of performing insecure actions are seen as acceptable to users, they would perform those risky actions such as visiting insecure websites and dismissing any security risk and credentials that could endanger both the user's mobile device as well as their sensitive information. The relationship between rewards and user's probability of behaving securely and it has been revealed that punishment, rewards as well as control assurance has an impact on user's conformity.

Personal influencers: Knowledge, skills and understanding.

It is very important for users to have the knowledge, understanding as well as the skills in order to defend themselves against fraudulent or unlwaful attacks. It is necessary for users to be equipped with the essential knowledge in order to perform and promote security measures as well as actions. Lack of user's knowledge in regards best security within the security aspect could result in a security failure. But one of the challenges faced by users is that it is quite difficult for users to conform to best practices of protecting their mobile device because users would not know which specific type of attack or risk they will encounter, especially during these times where by the nature of cyber attacks are always changing as attackers could find many different method to perform fraudulent actions. Due to such uncertainties, users tend to rely on their individual heuristics ability or skills that enables users to make quick and efficient judgements as well as decisions within a short period of time. However, it is also noted that even though the use of heuristics is beneficial, it may lead users to create biases.

On the other hand, the availability and delivery of constant and beneficial information is required in order for users to exhibit security behaviours but it is also noted that such information might not be enough to inspire or motivate users to change their behaviours into a more security oriented in nature. Users still exhibit poor security behaviours even after attending cyber awareness training and campaigns and further added that it is not recommended to refer to user's knowledge level as a determiner of good cyber security behaviours.

Perceptions, attitudes and beliefs.

Attitude can be defined as a person's inclination to judge or assess something in a particular way. Attitude has been known to influence an individual's behaviour, whereby individuals and users each own a number of beliefs and attitudes unique to themselves that may affect their behaviour in various aspects which also include their security behaviours. But it is noted that uncomfortable tensions may occur as a result of the misalignment between attitude and behaviours and the only solution to solve it is by undergoing change to one's attitude or behaviour. Within the discussion on the matter of behaviours, it will always involve various different factors that interact together in complex ways and researchers have come up with various models to illustrate such relationships such as:

1. Rational choice based model - One of the main assumption within the rational choice model or also known as rational action model, is that it assumes that users has perfect information. What is meant by perfect information here is that users are assumed to acquire all information regarding every possible choices or alternatives and then users will act on it by behaving in a manner that will provide them with the best outcome out of all possible choices. But an individual's act of processing the obtained information does not necessarily lead to the generation of a rational behaviour and consequently, individuals does not necessarily perform a rational choice in order to achieve the optimum result. Hence, according to this model, it can be said that every mobile device users are already equipped with the cognitive ability and motivation required to make rational decisions when faced with security incidents such as the act of applying facts in assessing cyber incidents. However, it is also noted that there is a challenge in doing so primarily due to the uncertainties of outcome or end results when dealing with anything related to cyber security.

2. Theory and model of planned behaviour - The main motive behind the use of the planned behaviour model amongst researchers is to analyse and describe the behaviours exhibited by individuals that are equipped with the ability to exercise self-control. One of the assumption that has been created within this model is that it assumes that any behaviour is planned and any individuals that plans to act or behave in a certain way will actually commit to it and behave in the way they have initially planned. A fundamental element within the planned behaviour model is behavioural intent, where the intention to behave in a certain way is subject to the attitude towards the expected outcome

of the desired behaviour as well as the evaluation of cost and benefit produce by the behaviour. Thus, according to this model, it can be said that when users believes that by behaving securely and employing security measures to their mobile devices will produce positive outcome to themselves, the users will effectively perform the security behaviour that they had planned in their minds.

3. Protection motivation model - The protection motivation model was created with an intention to aid individuals in resolving and coping with their fear appeals and this model believes that the behaviour of individuals are influenced by two appraisals namely the threat appraisal and the coping appraisal. The threat appraisal refers to user's perception on the gravity of an incident and user's perception on the likelihood of an incident or vulnerability while the coping appraisal refers to user's efficacy of the suggested precautionary behaviour as well as user's perception of their own efficacy (self-efficacy).

4. Learning model - One of the assumptions made within the learning model is that it assumes that behaviour is a process that individuals need to learn and that the learning process is influenced by two different elements which are incentives in the form of punishment or rewards and the social environment surrounding the individual which includes role models.

5. Change models - Change models are known to be built by the assumption that changes in behaviour is a step-by-step process that involves many stages and it does not ever occur in a single step or occasion. Researchers have constructed various change models and some of the most frequently models that have been implemented includes Lewin's 3-stage model of change management and Kotter's 8-step theory of change management.

Social influencers: Social norms at home, workplace and lifestyle.

It is in human nature that every person are bound to be influenced by the people that surrounds them regularly which includes family members, friends, top managers, work colleagues or other various entities that could be labelled as a role model to the particular individual. In other words, the behaviour, norms or beliefs of another person could heavily influence user's behaviour towards SMD security. In the context of organisational workplace, one of the main predictors of employee's behaviour towards the implementation of security policies is how employees perceived the expectation set by the managers on complying with the security measures or policies. The main reason of employees ignoring the instruction of the organisation to employ security practices and measures such as encrypting their email messages is primarily due to employee's not seeing the practices being exercised by fellow peers and managers. Thus, within the context of mobile device security, it is highlighted that user's chances of exhibiting security behaviours are increased exponentially when the entities or role models surrounding the user is exhibiting similar security practices or behaviours. When users feel that they are doing an activity that is similar to their role models or the neighbouring people, it could result in a sort of connection or "aligned" interest that could significantly promote the users of SMD to conduct a set of security behaviour [4, 26].

Generation-Z perception towards SMD Information Security.

It is a widely known fact that generation Z are regarded as the generation of youths that does not remember any strand of moments or memories without the usage of smart mobile devices, and they are considered as the top targets of attackers due to their constant usage of mobile devices. This is where the youth's awareness on cyber threats as well as the best practice of security behaviour on SMD comes into the bigger picture. The generation Z was observed to express concerns on the security of their mobile devices where within the research, it was discovered that about 40% of the Generation Z youths expressed their desire to be able to know the person they are communicating with when making online shopping or retail through authentication so that they are able to trust the person they are interacting with. It further added that generation Z are also concerned on various aspects of security such as the likelihood of their mobile device being hacked and the risk associated with cyber crimes which includes fraud and identity theft.

Even though it seemed that the Generation Z are actively concerned about security issues, there were also evidence from other research which states that the Gen-Z are overconfident in their ability to tackle security issues. In other hand, the Gen-Z assumed to themselves that they are very cyber secure but in reality, it was the vice versa. It was proven so from the survey conducted by the researcher where in one of the questions, the researcher discovered that the 96% of the Gen-Z youths believes that they are able to keep their personal or sensitive online data save but in another question, it was revealed that the 32% of the Gen-Z respondents have not put in little to no effort in creating their passwords. Within another question, it was also revealed that 78% of Gen-Z youths agreed that they had created and used the same passwords for various personal accounts. It is also revealed in recent studies that most of the Gen-Z youths are more open or encouraged to the idea of balancing between their desirability for a greater personalised experience and their concerns on security/privacy issues. However, generation Z are 25% more prone opt for a digital world compared to generation X and boomers, where in that digital world applications and websites have the ability to forecast and deliver what the user requires at any period of time. Here, the Gen-Z youths which accounts to 45% of total Gen-Z respondents were willing to give out their data in order to experience a more personalised environment at the cost of their privacy. The Gen-Z youths founds a website in which the website fails to predict the items they wanted, about 50% of the Gen-Z respondents will halt their activity on that website and stop visiting it.

In conclusion, it can be seen that despite the concerns for cyber security, some of the youths are behaving overconfidently towards their ability to protect themselves from cyber incidents while some behavioural patterns seen in youths presently are their willingness to trade their privacy just for some personalised environment or experience. One way to solve these behavioural problems is by creating or promoting security awareness on their behaviour. There could be different ways to solve them, thus this indicates that there is more work and research to be done on the topic of SMD information security and its relationship with the main users of mobile device, the Generation Z.

Effect of security awareness to customer trust.

Various researches had been established and conducted that was looking into the relationship between awareness of security and customer trust. There are possibility of risk associated with their transaction such as trust as well as privacy risk when a vendor or an organisation requested users to provide information that are considered irrelevant for the transaction such as asking questions on the user's age or gender. Trust amongst public towards organisations or corporations had been deteriorating for the past few years and one of the main rationale behind the drop of public trust is due to the advancement of cyber crime and cyber criminals, especially within this technologically advanced era.

Their study further revealed that there is a strong relationship between cyber security and user's assurance or trust towards an organisation. In the study, 53% of the respondents revealed that their perception of an organisation as a brand that can be trusted were based on the strict and meticulous security measures they had to undergo during the sign in process, and also, 49% of the respondents revealed that their experience of never encountering any security issues acted as an indicator for choosing which organisation to trust. Furthermore, 47% of the users within the study revealed to have stayed with their chosen organisations in which they assume to be more secure that other organisations.

6. Conclusion

In summary, with the increase in SMD usage contributed by the Generation Z youths, it is expected that more cyber attacks or incidents such as malware will be aimed towards the mobile device users. The study is aimed at measuring the level of smart mobile device security and privacy awareness. Firstly, a survey questionnaire was conducted in order to measure the level of awareness of SMD security amongst Generation Z youths.

The major findings of the survey showed that on average 43% of the users rarely changed their network and mobile screen password. It is also found out that over 56.1% users have never read the EULA policy before installing any applications and about 43.8% users always stay logged in after using their personal accounts. Additionally, more than 50% of users have stored sensitive data within their mobile device but 61.4% users have not installed any security software or applications to their mobile device, thus making their sensitive data vulnerable to cyber attacks.

Secondly, a new framework have been proposed in order to increase the awareness level of mobile device users which is based on the security behaviours exhibited by SMD users. Within the framework, it is highlighted that in order to increase the level of SMD security and privacy awareness, users need increase their level of awareness on security behaviours by understanding the importance and rationale behind various cyber security behaviours.

Author details

Heru Susanto[1,2,3]

1 School of Business, Universiti Teknologi Brunei, Brunei

2 Research Centre for Informatics, the Indonesia Institute of Sciences, Indonesia

3 Information Management, Tunghai University, Taiwan

*Address all correspondence to: heru.susanto@utb.edu.bn; heru.susanto@lipi.go.id

References

[1] Choo, Kim-Kwang Raymond. (2011). The Cyber Threat Landscape: Challenges and Future Research Directions. Computers & Security. 30. 719-731. 10.1016/j.cose.2011.08.004.

[2] McAfee Mobile Threat Report. (2019). Retrieved from https://www.mcafee.com/enterprise/en-us/assets/reports/rp-mobile-threat-report-2019.pdf

[3] Park, D. H. Kim, M. S. Kim and N. Park, (2013). "A Study on Trend and Detection Technology for Cyber Threats in Mobile Environment," 2013 International Conference on IT Convergence and Security (ICITCS), Macao, 2013, pp. 1-4.

[4] Sheila, M. & Abdollah, Mohd & Sahib, Shahrin. (2015). Dimension of mobile security model: Mobile user security threats and awareness. International Journal of Mobile Learning and Organisation. 9. 10.1504/IJMLO.2015.069718

[5] Susanto, H., Almunawar, M. N., Leu, F. Y., & Chen, C. K. (2016). Android vs iOS or Others? SMD-OS Security Issues: Generation Y Perception. International Journal of Technology Diffusion (IJTD), *7*(2), 1-18.

[6] Coventry, L., Briggs, P., Blythe, J., & Tran, M. (2014). Using behavioural insights to improve the public's use of cyber security best practices. Government Office for Science.

[7] S. Vashisht, S. Gupta, D. Singh and A. Mudgal, "Emerging threats in mobile communication system," 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Noida, 2016, pp. 41-44.

[8] Thiruvaazhi, U. & Arthi, R.. (2019). Threats to mobile security and privacy. International Journal of Recent Technology and Engineering. 7. 407-412.

[9] Androulidakis, I., & Kandus, G. (2011). Mobile Phone Security Awareness and Practices of Students in Budapest.

[10] Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). "Information Security Policy Complaince: An Empirical Study of Rationality-Based Beleifs and Information Security Awareness," MIS Quartely (34:3), pp. 523-548

[11] Puhakainen, P. (2006). A Design Theory for Information Security Awareness. Unpublished doctoral dissertation, University of Oulu, Oulu, Finland.

[12] Susanto, H., & Almunawar, M. N. (2018). *Information Security Management Systems: A Novel Framework and Software as a Tool for Compliance with Information Security Standard*. CRC Press.

[13] Susanto, H., & Almunawar, M. N. (2015). Managing Compliance with an Information Security Management Standard. In *Encyclopedia of Information Science and Technology, Third Edition* (pp. 1452-1463). IGI Global.

[14] Susanto, H., Yie, L. F., Setiana, D., Asih, Y., Yoganingrum, A., Riyanto, S., & Saputra, F. A. (2020). Digital Ecosystem Security Issues for Organizations and Governments: Digital Ethics and Privacy. In *Web 2.0 and Cloud Technologies for Implementing Connected Government* (pp. 204-228). IGI Global.

[15] Thomson. M.E, and Von Solms, R. (1998) Information security awareness: educating your users effectively, Information Management & Computer Security, Vol. 6 (4), pp.167-173

[16] Leu, F. Y., Ko, C. Y., Lin, Y. C., Susanto, H., & Yu, H. C. (2017). Fall Detection and Motion Classification by Using Decision Tree on Mobile Phone. In *Smart Sensors Networks* (pp. 205-237).

[17] Souppaya, Murugiah, Scarfone, Karen. (2013). NIST Special Publication 800-124 Revision 1, Guidelines for Managing the Security of Mobile Devices in the Enterprise. 10.6028/NIST.SP.800-124r1.

[18] Susanto, H., & Almunawar, M. N. (2016). Security and Privacy Issues in Cloud-Based E-Government. In *Cloud Computing Technologies for Connected Government* (pp. 292-321). IGI Global.

[19] Leu, F. Y., Susanto, H., Tsai, K. L., & Ko, C. Y. (2020). A channel assignment scheme for MIMO on concentric-hexagon-based multi-channel wireless networks. International Journal of Ad Hoc and Ubiquitous Computing, *35*(4), 205-221

[20] Leu, F. Y., Chiang, P. J., Susanto, H., Hung, R. T., & Huang, H. L. (2020). Mobile Physiological Sensor Cloud System for Long-term Care. Internet of Things, 100209.

[21] O'Dea, S. (2020, February 28). Forecast number of mobile users worldwide 2019-2023. Retrieved from https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/

[22] Susanto, H., Yie, L. F., Rosiyadi, D., Basuki, A. I., & Setiana, D. Data Security for Connected Governments and Organisations: Managing Automation and Artificial Intelligence. In *Web 2.0 and Cloud Technologies for Implementing Connected Government* (pp. 229-251). IGI Global.

[23] Susanto, H., Leu, F. Y., Caesarendra, W., Ibrahim, F., Haghi, P. K., Khusni, U., & Glowacz, A. (2020). Managing Cloud Intelligent Systems over Digital Ecosystems: Revealing Emerging App Technology in the Time of the COVID19 Pandemic. Applied System Innovation, *3*(3), 37.

[24] Susanto, H. (2018). Smart mobile device emerging Technologies: an enabler to Health Monitoring system. In *High-Performance Materials and Engineered Chemistry* (pp. 241-264). Apple Academic Press.

[25] Yie, L. F., Susanto, H., & Setiana, D. (2020). Collaborating Decision Support and Business Intelligence to Enable Government Digital Connectivity. In *Web 2.0 and Cloud Technologies for Implementing Connected Government* (pp. 95-112). IGI Global.

[26] Susanto, H., Ibrahim, F., Nazmudeen, S. H., Mohiddin, F., & Setiana, D. (2020). Human-Centered Design to Enhance the Usability, Human Factors, and User Experience Within Digital Destructive Ecosystems. In *Global Challenges and Strategic Disruptors in Asian Businesses and Economies* (pp. 76-94). IGI Global

Edited by Santhosh Kumar Balan

Data integrity is the quality, reliability, trustworthiness, and completeness of a data set, providing accuracy, consistency, and context. Data quality refers to the state of qualitative or quantitative pieces of information. Over five sections, this book discusses data integrity and data quality as well as their applications in various fields.

Published in London, UK

IntechOpen

ISBN 978-1-83968-800-3
9 781839 688003