Title Page

Title: A model for Identity Resolution in Big Data Digital Forensics using deep learning algorithms for the South African networking environment.

Table of Contents

List of Tables

Table of Figures

[Figure .1.1 Big data sources xiii](#_Toc111496454)

[Figure 1.2 Dissertation Chapters Layout 23](#_Toc111496902)

Declaration

Dedication

Acknowledgments

Abstract

Part I

INTRODUCTION

# Chapter 1: INTRODUCTION

## 1.1 INTRODUCTION

The digital environment or cyberspace has taken over our world today.  In today's technological era, we see digital use in every area of human life, in business, marketing, education, and medical, almost everything that exists uses cyberspace or online one way or the other. The digital environment is a global networking system that can be used on most devices nowadays and has become an essential part of our lives. Almost all organizations increasingly rely on digital networks for critical services {Brake, 2015}. The internet has made many positive contributions to our lives, allowing more productivity and comfort, and its use continues to grow. Kepios analysis reveals that internet users have more than doubled over the past 10 years, climbing from 2.18 billion at the start of 2012 to 4.95 billion at the start of 2022. This results in a compound annual growth rate (CAGR) of 8.6 percent for the past decade. Invariably increased use of the internet has brought voluminous big data that covers a wide spectrum, ranging from text, audio, images, click-through logs, Web videos, and EEG signals, to surveillance videos to mention but a few. This is made easily possible with the proliferation of sensor-rich mobile devices. As such multimedia data has been generated, published, and spread explosively, and has become an indispensable part of today’s big data. However, cyber digital crimes still exist and continue to rise emancipating from the rise of internet users. Studies note that such large-scale multimedia data has generated challenges and opportunities for intelligent multimedia analysis regarding management, retrieval, recognition, identification, categorization, visualization, and generation {Wei Zhang, 2019}.

Perpetrating criminal activities and other anti-social behaviors such as cyberbullying, sexting and other online-related crimes remain unresolved due to improper identification of the cyber attackers and victims for forensic investigation purposes. The legal system has challenges in keeping pace with today’s dynamic global networked technological environment and emerging crimes, identifying criminals in a digital environment as it is characterized by anonymity, complex hardware, and software, and exponential data growth (Mary C. (Kay) Michel, 2018). Responding to the crisis of cybercrimes is even more complicated by ambiguity in cyberspace, an example is having different multiple users with identical or very similar images or names. This is primarily with regards to users taking responsibility and their intent of cyber behavior like data errors, cyber fraud, and intentional or unintentional deceptions. According to {Brake, 2015} ambiguity about who is responsible for a cyberattack exacerbates the risk that most countries amid a geopolitical crisis will misattribute an attack, unduly retaliate or expand a crisis, or be unable to attribute an attack at all to the right users’ profiles responsible, thereby preventing or delaying a response of digital investigations, law enforcement and weakening their deterrence, cyber integrity, and credibility. Therefore, an effective means of automatically identifying cyber criminals, attackers, and victims for investigations and prosecutions will aid mitigate cybercrimes and combating violent behavior. This is paramount as one of the major reasons crime perpetrators are never brought to justice is insufficient identification of attackers and victims in the evidence noting that user footprints are traceable in cases of cyber-criminal offenses.

Some researchers call it entity resolution. According to {Wang, 2015} identity resolution aims to identify different descriptions that refer to the same real-world identity or entity appearing within or across a data source. Here, we see no existence of a unique identifier in the data source. The importance of identity resolution in today’s modern digital technological society cannot be stressed enough. Massimo Leone (2021) emphasizes that identification is a primary necessity in our societies today to unmask impostures, disguising, and forgery identities in digital networks. Identity resolution often plays an indispensable role in both civil proceedings and criminal case investigations. For example, in the case of a digital investigation, the investigator’s collation of any potential digital evidence data about an attacker or victim during the digital forensic investigation process must ultimately be accurately linked to the same right real-world individuals. This can be challenging as the identity representation of real-world persons can be generated from big data heterogeneous sources through networks by analyzing diverse sources where the references to real-world persons are mentioned like the web, IoT, cloud, Databases, and social media. Identity resolution is the process of resolving multiple instances of data representing the same entity into a single record for that entity which is one of the most crucial processes in [master data management](https://www.sciencedirect.com/topics/computer-science/master-data-management) (David Plotkin, 2014).

Identity resolution is known to uncover identity records that are co-referent to the same real-world individual that is due to duplicate, false, and unresolved identity records that relate to attackers and/or victims due to data ambiguity, errors, or intentional that are available in today’s heterogeneous big data, hindering digital forensic evidence for investigations. Identity resolution is not only caused by duplicate individual entry errors or data ambiguities but also intentional identity fraud and deception, which tends to be hidden and concealed. It may occur in big data digital forensics due to massive amounts of data from diverse heterogenous sources and across multiple data sets; containing ambiguous, duplicate, or incomplete identity data which significantly limits the ability to find true identities of persons in investigations related to digital forensics. This situation may further cause misunderstanding, confusion, and errors in evidence analysis during a digital forensic investigation process when attempts are made to harmonize data emanating from different sources (Piasecki, 2008). Being able to resolve identities and duplicate identity records of attackers and victims in heterogeneous big data sources is very crucial for improving the quality of the data obtained. Identity resolution results in two kinds of data matches: deterministic which is derived from known first-party data and probabilistic which is a less resolute prediction of a user’s identity having a probable possibility as it is not given and unknown.

For cybersecurity to be achieved, digital forensic experts must be able to determine whether an individual is who they purport to be which is not a trivial task as criminals may assume multiple identities using either fraudulent or legitimate means, committing identity crimes like identity fraud, identity theft, and identity deception, to enhance the chances of success in their missions. Digital forensic experts acknowledge that automation and artificial intelligence can be a solution to deal with the increasing complexity and volume of today's big data digital evidence (E. Casey, 2012) to eradicate cyber deception. If we devise a mechanism to detect the duplicate identity of attacker and victims; automatically enable integration and linkage of digital evidence to the attacker and target victims in the South African networked environment; it becomes much easier, faster, and useful in incarcerating any attacker, enabling successful prosecution as the data or information presented stands a much greater chance of being accepted in court for events and incidents related to a digital crime. This would eradicate the challenge of acquiring and examining appropriate evidence relevant to the case under investigation interpretation, descriptions, and representation of the same or related digital forensic data or information. Digital forensics is the use of scientific methods for identifying, preserving, extracting, and documenting digital evidence derived from digital sources to enable successful prosecution, answering the when, what, who, where, how, and why events and incidents of a digital crime. Nirkhi, Dharaskar and Thakre (2012) outlined the different techniques traditionally used in a digital forensics’ investigation like imaging used to image or copy the media to be examined; hashing used to quickly identify a file and provide an authenticity that an image or file was not modified, and carving used to scan disk blocks that don’t belong to examined files to find deleted data recovering deleted files and allowing fragmented files to be recovered with more accuracy. However, one of the limitations is media and files could be modified and the integrity of the evidence is compromised.

Several current applications have also been used in digital forensics. These include identifying correlations in forensic data (association), discovering, and sorting forensic data into groups based on similarity (classification), locating groups of latent facts (clustering), and discovering patterns in data that may lead to useful predictions (forecasting). While these techniques are ideal, they have become limited and very challenging with today’s emergence of big data as these current investigation tools are designed for limited or reduced digital information in several devices. Deep learning is a sub-field of artificial intelligence that has penetrated so many fields like the medical, cybersecurity, customer relationship CRM, agriculture, and public management amongst others; greatly enlightening identity resolution for big data digital forensics with its ability to learn big data with heterogenous large sizes using techniques like clustering, classification, feature attributes extraction, etc.

The concept of big data requires investigation tools that manage large sizes of data with heterogeneity from numerous sources and with the requirement of high-speed data processing and data variety. Big Data is usually sourced from various large mediums like social media, network or process logs, emails, health care, transport, e-commerce, mobile phones, satellite communications, GPS systems, and so on (Suneeta Satpathy and Chandrakant Mallick, 2019). This makes these big data both structured, semi-structured, unstructured, and/or even mixed. It is even more complicated as big data holds duplicate and false identity records which are connected with attackers and/or victims due to data ambiguity, data errors, or intentional deceptions. This greatly hampers the digital forensic process to acquire and examine appropriate evidence relevant to a particular case under investigation and impedes the efficiency of criminal investigations creating a need for identity resolution of big data digital forensics. For this reason, there is a vital need to develop models of identity resolution in the big data digital forensic domain that can assist in interpreting, describing, or even presenting the most common big data representations of digital forensic data in any court of law or during civil proceedings. Big data digital forensic uses very large-scale datasets to establish evidence, identification, collection, organization, and fact presentation. Rawat et.al (2021) describe big data analytics as the art of processing, storing, and gathering large amounts of data for future examination. The use of big data and deep learning allows for the processing and analysis of large datasets of different formats like structured and unstructured data example, text, audio, and video in a short time used for digital evidence in computer forensic processes to determine attack patterns and profiles of attackers to establish counterattack strategies and uncover hidden insights Andrade et.al (2020). For example, Google has access to big data through its Chrome browser search engine for autocomplete services to its users and uses this data to train algorithms that improve search tasks allowing a better understanding of what users are trying to search for, recommending search suggestions, correcting misspellings, and parsing sentences.

Distinguishing the uniqueness of an individual from several identity attributes is a real and recurring problem that appears across different fields.

Now that the scene of this dissertation has been set and the aim established, the reader should note that the remainder of this chapter is structured as follows:

* Section 1.2 focuses on identifying the problems to be addressed by this research.
* Section 1.3 covers the goals, the objectives and describes the tactics that are employed in addressing the identified problems.
* Section 1.4 explains the research methodology that is used for this dissertation.
* Section 1.5 discusses the motivation for the study
* Finally, the overall structural layout of this research is presented in Section 1.6, with a short conclusion to this chapter in Section 1.7.

To this effect, the next section presents the problem statement of this research. The identified problem is then broken-down into smaller elements that are addressed in the course of this dissertation.

## 1.2 PROBLEM STATEMENT AND RESEARCH QUESTIONS

This research study recognizes the importance of identity resolution for big data digital forensics to foster digital forensic investigation processes in the South African networked environment. It is always assumed that identity attributes like names, videos, audio, and pictures are unique which is not true as diverse people can have the same data attributes causing ambiguity. This is a well-known problem in digital data. One of the challenges encountered by law enforcement investigative agents is to verify cyber-crime events with the evidence of identities involved. As a result, numerous criminal cases remain unexplored, or the crimes that are detected remain unsolved due to a lack of sufficient evidence placing the real attackers or victim's identity at the scene of the crime. The question that still exists is discovering automatic ways of identifying the real user of a user attributes of an identity class or authentic unique identities in digital data. This is now of utmost importance as Mary Catherine Michel **(**2020**)** claims that with today’s technological systems,digital fraud and assuming another person’s identity attributes is very easy for example, another user's photos can be simply stolen and copied to create a new profile. Given the growing scale of data and the increased number of user accounts, it is simply impossible to manually link suspicious attackers and victim individuals and track their latest status. Therefore, it is highly desirable to develop novel methodologies that can automatically link multiple accounts of the same individuals. In order words, creating a model that can help automatically resolve user identities in today’s networked heterogeneous diversified big data environment will greatly enhance digital forensic investigations.

Research shows that the traditional software packages for digital forensics analysis and examination like file retriever, registry viewer, log parser, and so on, run in a standalone computer environment. Whereas with big data, lots of systems are interconnected and interrelated with each other. This makes the traditional software packages and methods no longer relevant to meet the current needs of volume, variety, velocity, ambiguity, complexity, and other characteristics of big data. Hence, a new direction is urgently needed to adapt to the distributed big data system as big data forensics investigation is usually run on distributed heterogenous systems. Jie Song and Jin Li (2020) noted that one of the major challenges encountered in their study is a lack of a powerful and accurate artificial intelligence system that is essential for quickly and efficiently finding criminal clues or evidence in massive data. This reduces the workload of investigators. They also reiterated that ensuring standardization in the process of big data digital forensics will positively guarantee the reliability of forensic investigation results and improve work efficiency. This could be done in line with the data, processing procedures, tools, software, technologies, system structure, workflow, and components, in the way of system deployment and technical solutions. Therefore, the main problem tackled in this research study is the lack of a theoretical model specifically designed for big data identity resolution in the digital forensic domain.

The vitality of this problem is that duplicate and false identity records are very common in big data due to so many reasons such as unintentional data errors and deceptions that cause security problems to digital systems. Eoghan Casey(2019)explained some consequences that data errors and omissions in digital forensics can cause. Examples like imprisoning innocent people, leaving dangerous criminals free to perpetrate additional crimes, and the continual victimization of organizations and people targeted by offenses. Some of these identity crimes are falsifying passports and baptismal certificates to facilitate their financial operations and execution of attacks, either in the real world or in cyberspace. Srivastava and Roychoudhury (2020) observed that one major challenge in developing a reliable and scalable matching scheme for online identities is the non-availability of the required information or having contradictory information for the same user across networks. To mitigate these risks, it is necessary to harmonize identity resolution in big data digital forensics in the South African networking environment and to strengthen knowledge management throughout forensic ecosystems. Duplicate identity records in big data of different data formats need to be resolved, this will help to eradicate deceptive criminal identities, and eliminate challenges in acquiring and examining appropriate evidence relevant to cases under investigation. It will also enable the examination of appropriate identities for accurate intellectual interception of digital data as inadequacies due to an individual having multiple identities can easily mislead intelligence and law enforcement investigators.

Cybersecurity studies show that digital evidence is very volatile and fragile, hence the improper handling of this evidence can alter it. As such adequate protocols are followed to ensure data isn’t modified and has good data quality storage. With this, as a database administrator, one must be confident that there are no duplicate records of identity data from server logs and networking hardware. For example, if an investigator is required to collate data about an attacker or a victim from customer data profiles with diverse sources like receiving the name and address of a user identity from government records or third-party data providers without any data linkage between both data sources. All “Caleb Odi”, “CalebO.co”, and Caleb O are assumed to point to the same user identity whereas “Caleb Odusan” is an entirely different identity. This can be so challenging as it is also the case for other identity attributes like address, first-name, last-name, emails, phone number, videos, and images where there are some variations mostly in heterogenous big datasets.

The problem area identified in this research study can be broken down further into the following research questions:

* When does identity ambiguity occur in big data digital forensics and what are the current efforts undertaken to resolve them?
* What are the big data digital forensics influencing factors of identity resolution?
* Of what significance is the whole process of identity resolution in big data digital forensics to computer professionals, law enforcement agencies, and practitioners in digital forensics?
* What types of practical methods currently exist that can help to resolve identities in big data digital forensics?
* How are deep learning algorithms applied to big data identity resolution digital forensics?

Note that the study presented in this research thesis and as reflected in the research questions above is motivated by the need to prepare for the prevention and combat of cybercrime in the era of big data. The study aims to propose a model that can assist in identity resolution for big data digital forensics in a South African networked environment where an investigator collates data for victims or attackers in big data. Also, to investigate ways that will enable the forensic investigator to automatically detect deceptive criminal identities for effective prosecution in digital crime case investigation for digital forensics. The questions give insight into the degree of and use of identity resolution systems in big data digital forensics. The accomplishments, challenges, and opportunities surrounding big data identity resolution digital forensics are identified, so that a redundant solution by this dissertation is avoided. Instead, an improved solution is developed upon aspects that are beneficial to big data identity resolution digital forensics.

The issues identified in the problem statement are solved by a few objectives which are discussed in the following section.

## 1.3 GOAL AND OBJECTIVES

Knowing that there has been quite very little research done on identity resolution for big data digital forensics from research on existing works of literature, the primary objective of this study is thus, to develop a model for resolving identities in big data digital forensics in South African network environment using deep learning algorithms. Also, the other contributions of this research study attempt to demonstrate the proposed model using a prototype that addresses duplicate identity records of attackers and victims in heterogeneous big data sources, identify false identity records, resolves data errors, representations, and identities that are available in different data formats like text, image, audio, video. To achieve this study's objectives, we first review identity theories and existing identity resolution techniques to understand the state-of-the-art, followed by an introduction to our proposed collective identity resolution technique.

The objectives of this study are shown more specifically in the bulleted list below.

* Presents a literature review on existing identity resolutions techniques, to conduct a research survey that provides information on the current state of big data identity resolutions for digital forensics in the South Africa network environment and on automatic and dynamic data annotation for identity resolution for big data digital forensics
* Establishes the significance of automatic and dynamic data annotation for identity resolution to computer professionals, law enforcement agencies, and digital forensic practitioners.
* Designs a big data identity resolutions digital forensics model
* Develops a prototype of big data identity resolutions digital forensics model as a proof of concept. The following section outlines the methods that will be applied to achieve the objectives listed above.

## 1.4 RESEARCH METHODOLOGY

The methods included in this research are the literature review, design science methodology, modeling, and prototyping. This study approaches answering research questions through empirical and theoretical methods. Also, the study aims to explore and use some big data analytical tools like Spark, Tensor flow, and Hadoop, experimenting with which tools are best appropriate. These tools are used along aside deep learning algorithms through an inductive approach. Quantitative data collection is used to collate data from heterogenous sources. A more detailed description of these methods follows.

* Literature review: A state-of-the-art literature review is conducted on identity resolution, big data digital forensics, cyber security, identity resolution in big data digital forensics, and the use of deep learning algorithms in the South African networking context. The literature review provides the necessary background and helps identify the similarities, challenges, and opportunities from the relevant literature. Subsequently, the knowledge acquired from the literature review is used to design the conceptual model.
* Data collection techniques: Data is collected and sourced from the South Africa networking data repository and some other sources. This is done numerically and empirically as data is large, unstructured, heterogeneous, having diverse dimensionality and complexity in nature which causes these data ambiguity and misleading. The study collects data from online and offline sources and integrates it into a single system with considerations of the speed, accuracy, and security of the data. This is done through simulation with API. The majority of these big data digital forensic investigation involves the sourcing of digital information stored on digital devices like computers, mobile, game console, cell phone, hard disk, router, social media, webpages, IoT, and other various digital media storage mediums for a sole purpose of civil, criminal, and digital forensic investigation. The processes usually entail identification, acquisition, preservation, examination, and presentation of the findings to the court of law and other stakeholders by forensic expert investigators for decisions making.
* Big data analytical tools: The data is analyzed using big data tools and technologies like Tensor flow. Apache Hadoop, Apache HBase, and Apache ZooKeeper for data storage and management. Excel for data cleaning. Hadoop MapReduce, Pig, Hive, and Apache Spark, for data analysis. Sqoop, Flume, Mahout, Impala, Oozie, NoSQL databases, and stream mining. Deep learning algorithms are adapted to Map Reduce with considerations on the techniques that are best suited and appropriate for identity resolution like k-means. Also, data is organized with deep learning algorithms as they are suitable to address unstructured, volume, and variety of big data (Nagwa M Elaraby et.al, 2016).
* Design science research methodology: This methodology is suitable and appropriate because it allows cycling back to earlier activities and earlier processes of the research step processes outlined below. The study aims to propose a research six-step process:
* Identify a problem and motivate to show research importance.
* Define research objectives of a solution showing improved research model and artifact accomplishment.
* Design and development of the improved model.
* Demonstrate a model to solve practical problems with suitable context for metrics and analysis.
* Evaluate, observe and communicate its effectivity and efficiency, iterating back to the model design for better results.
* Modeling: A visual representation of the idea that describes the semantics of the system is shown from the knowledge obtained from the data analysis and the statistical results from the survey. The model presents an abstraction of the proposed system, whose function resolves user identity in heterogeneous big data for digital forensics investigations.
* Prototyping: A software prototype for the designed conceptual model is developed using the relevant programming languages, application programming interfaces, and integrated big data development environments. Prototyping allows for a software artifact to be developed and tested as the specifications of the system are gradually obtained. The feasibility and shortcomings of the artifact are also identified and used to simulate real-world situations.

## 1.5 MOTIVATION FOR THIS STUDY

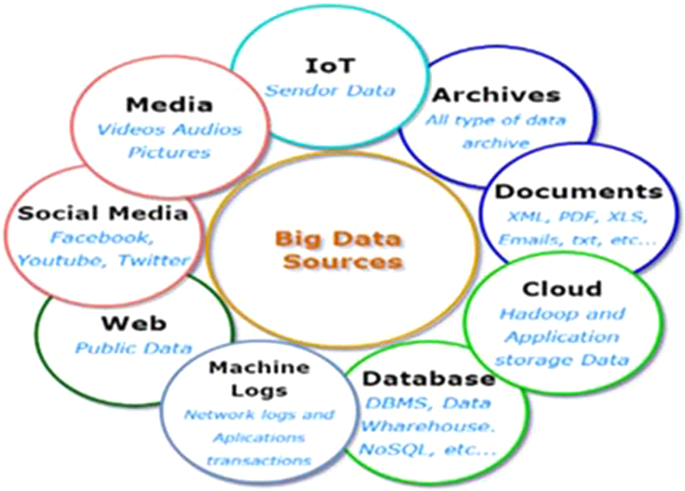
The research undertaken for this study is primarily motivated by the realization that, as the evolution in digital technology goes on, some challenges also erupt. This is due to increased flooding of heterogenous, and voluminous digital data generated daily; computers and other digital systems like social media, smartphones, wearables, or any other digital device people use to interact with the world around them. The continuous changes in technology as devices become more connected through all kinds of technology and computer networks (Afcea, 2014) and the consequences of not resolving identity ambiguity, data errors, and user identities which are available in different data formats are all very challenging to manage. It becomes even more serious as computer crime techniques are also becoming more sophisticated and well-coordinated, especially in a networking environment. It is important in our world today to have automatic ways of properly identifying a victim or criminal amidst so many identities in big data for forensic.

Duplicate and false identity records are very common in electronic systems, identity management systems, and other systems. This is due to so many reasons like data ambiguity, unintentional errors, and even intentional deceptions. {Wang, 2015} observed that the main reasons for these duplicate and false identity records are the lack of sufficient verification or validation during the data entry processes. Also, attackers employ techniques that conceal their identity such as cryptography, and remailers which are intermediary mail servers that act as gates between the sender and recipient of the email to hide their identity to commit a crime. As a result, it becomes very challenging to resolve the identities of an attacker and targeted victims for digital attacks in network environments, especially in big data digital environments. This weakens digital forensics investigations as digital crimes remain unsolved or the potential proof is inconclusive due to a lack of concrete or sufficient unique identity evidence that automatically puts the offender(s) or victims at the scene of the alleged crime. This hinders the prosecution of the crime perpetrators. Therefore, finding an effective solution to address this issue is extremely crucial as it will greatly reduce cybercrime and insecurity. (Pengfei Hu, 2017) cited that in the face of increased use of the internet of things, identification and resolution technology are a prerequisite for realizing and mapping identity consistency from the physical to the cyber space in the deployment and operation of digital applications and services.

By using a mobile device application as a tool to curb crime, communities are empowered to identify and fight crime in their neighborhoods. The motivation for this research is therefore to employ the techniques of crowdsourcing and the social media method of ’capture and upload’ to empower and motivate communities to engage in the fight against crime and thus build safer South African communities.

Also, the increasing heterogeneity and voluminosity of data in our world today – which increases per second of the day due to its source from various large mediums like social media, network or process logs, cloud, web, emails, IoT, databases, health care, transport, e-commerce, mobile phones, satellite communications, GPS systems and so on (Suneeta Satpathy and Chandrakant Mallick, 2019) has rendered identity resolution inevitable. According to (Manasa Priya Koduri, 2021) with phasing out third-party cookie support for user identification, organizations are faced with the use of first-party data hence the importance of an automatic way to resolve attackers and victim identities in South African networks.

Figure 1.1, for example, from the digital world shows the growing heterogenous diverse big data sources in the world by variety. These statistics make the identity resolution in big data digital forensics such a rich field for all manner of criminal activities.



Copied from: journalofbigdata.springeropen.com (2021)

Figure 2.1 Big data sources

The gap in this study originates from the fact that the emergence of big data today in digital forensics has its characteristics that differentiates it from other identity resolution studies, and this warrants the need to be studied separately. Howard Munkhondya et.al. (2020) observed that the increased dependency on information technology (IT) with its security apprehensions is one of the rousing factors for digital forensics as digital forensics has the potential to either strengthen the security functionality of a network through iterative investigation processes. Digital forensics is known to support legal proceedings and follow established guidelines that ensure the availability, authenticity of data, and court admissibility of digital elements and evidence. A critical component of digital forensic cases involves extracting information and data from diverse devices and the methods for collection and examination of these digital devices are constantly changing. Studies show that advanced threat actors now use memory implants as malware that resides and lives only in the memory of digital systems. This causes attacks leveraging memory malware and converting operations to avoid detection. Studies also show that the dimension of potential digital evidence supports has grown exponentially, be it email, text messages, photos, graphic images, documents, files, images, video clips, audio clips, databases, internet browsing history, and so on, with increased speed and volatility mostly in the connection of internet of things in the networked environment. Cloud services amongst others are now sources of potential evidence in a vast range of investigations and the network traffic is greatly expanded in cyber security. This makes it almost impossible to effectively sift through vast amounts of data quickly without data ambiguity. The necessity of identity resolution in big data digital forensics is now paramount. Intelligent analysis, deep learning, and the use of identity resolution automatic algorithms can profit from sophisticated analysis of such digital forensic investigations, particularly in a networking environment. Some of these challenges of big data evidence highlight the necessity of revising tenets on identity resolution and procedures firmly established in digital forensics. Therefore, most tools and libraries are not suitable and/or well validated for today's big data digital forensic work, so there still exists a wide space for the development of today's innovative tools leveraging deep learning methods in the South African networking environment.

As a way to restrain the growth of digital crimes emancipating from the rise of internet users, identity resolution is forming a significant part of big data digital forensics. Identity resolution tries to recognize an entity by connecting the growing volume of user identifiers and relationships to one individual identity as he or she interacts across different channels, big data sources, and devices. Identity resolution attempts to get answers regarding digital crime incidents by encouraging a unified view that differentiates a specific user from another individual identity of digital users. Identity resolution brings together identifiers across online and offline touchpoints. This is the only way to truly understand individual internet users, their footprints, and their attributes.

Though there are some works done on big data in digital forensics (Suneeta Satpathy and Chandrakant Mallick (2019) they are limited as they are not in any way related to identity resolution. Other research work on identity resolution in big data (Huqing Wang and Zhixin Sun, (2021), (Yuzheng Ren et.al., 2020), Nihel Kooli (2018), (Adeyemi R. Ikuesan and H.S. Venter, 2019; Jonghyun Kim et.al., 2021), Tanner Fry (2020) and (James Brusseau, 2019). But these studies are not related to digital forensics. Andrade et.al (2020) conducted a literature review on big data for cyber security, the authors observed that there are few contributions regarding threat hunting and cyber-deception and the use of big data for this purpose must be enhanced. This is the purpose of this study as it is aimed at proposing a new theoretical model for identity resolution of big data forensic investigation using deep learning in a South African networked environment. This is so important as he further reiterated that just implementing a big data architecture isn’t enough to solve the problem of large amounts of data. Identity resolution in heterogeneous big data to identify duplicate records, data errors in identity representations, false identity records, and resolving identities helps in managing the data quality measures for digital forensics. Finding an effective solution to this problem is extremely critical in fighting crime and terrorism to enforce national security. Criminals and terrorists often assume fake identities using either fraudulent or legitimate means to hide their true identities. Identity resolution for big data digital forensics has become a core issue in solving digital forensic security problems. The emphasis of our study lies in proposing a model for an effective security evaluation system for identity resolution that can be applied to all levels of big data digital forensics in the South African networked environment.

However, Hans Henseler, Jop Hofste and Maurice van Keulen (2013) researched identity resolutions in digital forensics. The study is limited as it involved the use of small data. {Nickson M. Karie, 2019} on the other hand diverged deep learning cognitive computing techniques into cyber or digital forensics and proposed a generic deep learning cyber forensics framework (DLCF) to realize effectiveness during forensic investigations. Developing a model of identity resolution for big data digital forensics has not been met at this current stage and is yet to be explored. With internet accessibility widening, the emergence of big data that is heterogenous, voluminous, having diversity and large sizes of data with heterogeneity from different sources, and with the requirement of high speed of data processing reiterates the importance of this study as big data poses new challenges to security analysts who must process this data to determine attack patterns or anomalies that allow identifying threats or security attacks and more crime taking on a digital aspect.

In the end, this research is building on the DLCF framework, and the research contribution aims to extend the evidence analysis part of automatically resolving user profiles in multimedia files such as text, images, audio, and video to realize effectiveness during rightly identifying attackers or victims through their attributes for forensic investigations using deep learning approaches. The contribution of this paper is thus, developing a model for determining or resolving identities in big data digital forensics to help digital forensic investigators to manage the investigation process concerning automatically determining user identity profiles.

With the aggregation of multiple data sets brought together from different heterogeneous sources, the data tends to have duplicates, inconsistencies, and inaccuracies when consolidated together into a system. With the diversity and voluminosity of big data today, the identity resolution approach of simply matching with rule-based and statistical methods is insufficient and ineffective during an aggregation phase. There is an automated need to link and consolidate identity profile information with a high level of confidence in the big data environment to easily detect cyber attackers and victims. The motivation for this research is, therefore, to employ the techniques of deep learning to automatically identify user profiles of victims or criminals amongst so many identities in big data for forensics be it text, video, audio, or images to fight against crime and thus build safer South African networks.

## 1.6 DISSERTATION LAYOUT

This dissertation consists of five parts that are broken down into eleven chapters. Part One comprises the introduction chapter (Chapter 1), while Part Two contains the three background chapters of the study (Chapters 2, 3, and 4). Chapter 2 discusses existing literature on identity resolution. Chapter 3 focuses on big data then chapter 3 delves into digital forensics. Part three constitutes two modeling chapters (Chapters 5 and 6). Part Four comprises three chapters that explain the model prototype (Chapters 7, 8, and 9) which concentrates on the design, the implementation, and the identity resolution problem in big data digital forensics. Finally, Part 5 entails two chapters for the evaluation containing Chapters 9 and 10 which focus on the critical evaluation and conclude the thesis respectively.

The current chapter (Chapter 1) sets the scene of the study by introducing the research problem, the motivation, the objectives of the study, and the research questions are also outlined. However, the rest of the thesis is organized as shown in Figure 1.1, followed by an outlook for each of the chapters.

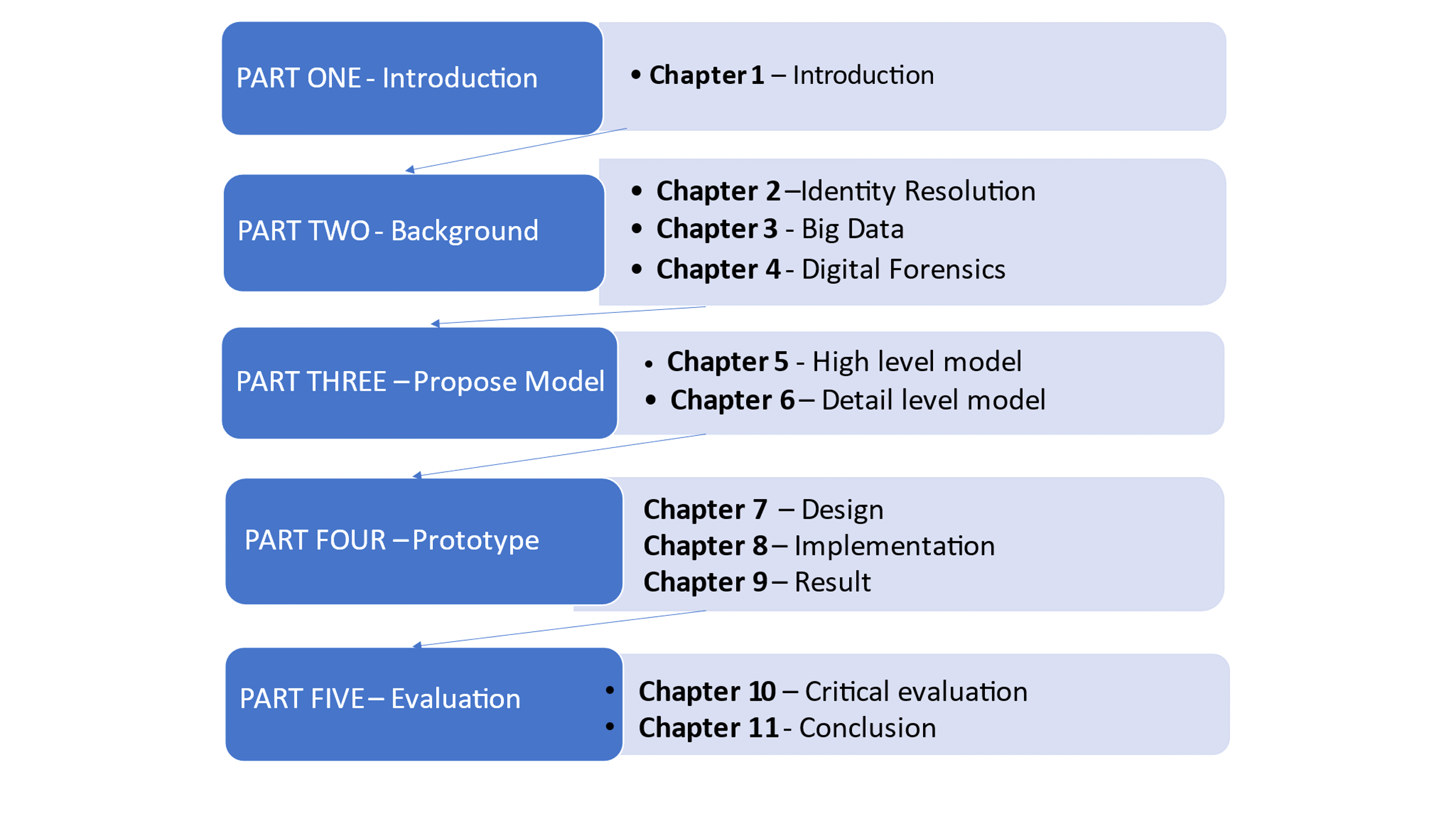


Figure 1.2 Dissertation Chapters Layout

Chapters 2, 3, and 4 make up Part Two of this thesis and cover the background and literature reviews for identity resolution, digital forensics, and big data concepts respectively. The chapters provide the necessary background information for the areas of interest discussed in the dissertation. In Chapter 2, for example, the reader is provided with a comprehensive background of identity resolution, how it works, and the techniques of identity resolution. Nonetheless, special reference is also made here to the focus of the study, namely the identity resolution in big data digital forensics within the South African network environment.

In line with the problem statement, trying to eliminate cybercrime that is caused owing to users falsifying identity attributes, digital forensics practitioners are faced with trying to resolve user identities in data. The study investigates how identity resolution assists digital forensics. Digital forensics and various challenges are discussed in Chapter 3. Digital forensics entails the process of preservation, identification, extraction, and documentation of computer evidence available in digital devices which can be used by the court of law for civil and criminal investigations.

Aligned with the research problem as well, due to the challenges of increased emergence of heterogeneous large data that we experienced in our everyday living, which brings about managing large sizes of data with heterogeneity from diverse sources with the requirement of high speed of data processing; we are faced with the problems of data ambiguity, errors or intentional deception duplicate and false identity records of attacker and victims that are available in these heterogeneous big data sources in a networked environment. Big data and some analytical techniques used in big data digital forensics are discussed in Chapter 4. Big data digital forensics is a multi-level recursive parsing technology as the digital forensic processes have different steps with the security architecture that is complex and the security field covering a wide range of areas. Also, the processing capacity of South African network environment apparatuses connected to big data is very different and there might be regulatory or cultural components of digital forensic processes that can influence results as well. These must be considered in investigating identity resolution in big data digital forensics.

Chapters 5 to 6 make up Part Three of this thesis and covers the proposed model of the study. Chapter 5 presents the reader with a conceptualized high-level concept of the model, almost like a black box of the model. While Chapter 6 expands more on the model with detailing and explaining. The model is also meant to resolve identities in a big data digital forensic data source that occurs in a South African networking environment. Where such a model can also be used to develop new techniques for resolving identities in big data digital forensics.

To assess the feasibility and implementation of the proposed model, a prototype is developed and discussed in Part 4. Note that Part 4 covers three chapters which are chapters 7, 8, and 9. In Chapter 7, for example, the reader is provided with the design of the model for identity resolution in big data digital forensics in the South African networking environment. The model will be designed based on the information gathered from the literature reviews (Chapters 2, 3, and 4). The model will present an abstraction of the proposed system, the function of which is to facilitate a new theoretical model for big data digital forensic investigation using deep learning in the South Africa networked environment. The chapter will also illustrate the architectural design of identity resolution in big data digital forensics in the South African networking environment based on the conceptual model.

Chapter 8, on the other hand, discusses the implementation of the model in an attempt to demonstrate an understanding of the proposed model for identity resolution in big data digital forensics in the South African network environment. The experiments conducted to test the implementation and accuracy of the model prototype are also explained in Chapter 8. The prototype attempts to resolve identity as a result of data ambiguity, errors, or intentional deception in these heterogeneous big data sources in a networked environment. However, the experimental results are based on the individual methods used to develop the prototype in this study and Chapter 9 organizes and explains the results obtained from the study.

The last chapters (chapters 10 and 11) make up Part Five which involves the evaluation and conclusion of this thesis with an explanation of the extent to which the research problem has been addressed as well as all the accomplishments. Chapter 10 does a critical evaluation and Chapter 11 concludes the study and points out possible areas for future research based on the current study.

## 1.7 CHAPTER CONCLUSION

Chapter 1 is the introduction chapter and states the problem statement of the research. In this chapter, the researcher introduces and explains the primary area of focus in this study, as well as identifies the problem statement. The motivation for the study and the objectives are also highlighted to show the necessity for this study. The whole of Chapter 1 is generally meant to show the scope of the study covered in this dissertation.

The next chapter (Chapter 2) presents the background of identity resolution. Chapter 2 is also meant to introduce the reader to some of the basic facts about identity resolution digital forensics, as well as how it works and the techniques of identity resolution. Special reference is made to the focus of the study: the identity resolution in big data digital forensics within the South African network environment.

Part 2

BACKGROUND

# Chapter 2: BIG DATA

## 2.1 INTRODUCTION

Since the invention of computer devices, the proliferation of cloud computing (Botta et al., 2016), the evolving development of the internet technology, web data, social media data, sensors data, advances, and the spread of smart devices like IoT; the era of big data with data explosion is here with us. These data are mainly multimedia in nature and used as input or output of applications such as images, video, audio, and text. According to {Diana P. Tobón V, et.al, 2022} a great amount of multimedia data is generated daily which complicates the data analysis and storage as they are mostly characterized as heterogeneous – coming from different sources, unstructured, and large. This consequently has created many challenges {Ibrar Yaqooba, 2016}. These growing amounts of data originate from various sources that are not organized, unstructured and unordered, including data from machines, devices, or sensors, and massive public and private data sources as well (McAfee et al., 2012). This big data is so important for organizations as they unveil new patterns, and trends, allowing them to gain additional insights, add value to businesses, improvement data-driven decisions making, providing solutions and answers to pressing business issues. However, in the case of cybercrime incidents and investigations, these continuously increasing types and amounts of data needs to be processed {Jie Song and Jin Li, 2020}.

Since this dissertation deals with the identification of user profiles in big data digital forensics to be used by law enforcement agents (LEAs) and digital forensic investigators in their battle against crime, it is necessary to discuss what big data entails and how it has emerged over time. In discussing big data, therefore, there is a need to firstly introduce the concept(s) of big data.

## 2.2 DEFINING BIG DATA AND ITS CHARACTERISTICS.

According to Philip Chen & Zhang, 2014) big data is a collection of large amounts of complex data that cannot be managed efficiently by state-of-the-art data processing technologies. Hence surpassing the current ability of humans to manually design appropriate data storage and analytic systems to manage this large amounts of data effectively. IEEE computer society (2022) remarkablys confirms that today digital world is multimodal, which is the combining different modalities of data such as text, audio, images, videos, animations, drawings, depth, 3D, biometrics, interactive content and so on. They call it multimodal big data analytics, a need today.

{Ahmed Oussous, et al., 2018} define big data as a huge growing data set that includes heterogeneous formats containing both structured, unstructured and semi-structured data. As such having a complex nature that requires powerful technologies and advanced algorithms is a necessity.

{Stanier and Isitor, 2016} on the other hand define big data from three approaches – the comparative approach, the attribute-based approach, and the environmental approach definition. The comparative approach although a very limited approach discusses big data in terms of the differences with relational systems, relative processing capacities, and the opportunities offered by new technology. It compares big data emergence to traditional systems and their inefficiencies. The attribute-based approach on the other defines big data in the context of its characteristics volume, velocity, veracity, value, and so on. While the environmental approach defines big data in line with its architectural design, models, and framework, mainly for efficient storage, manipulation, and analysis.

In this dissertation, however, the focus is on digital and multimedia data sciences, which is referred to as multimedia big datasets. Therefore, in the context of this research, big data is a voluminous-scaled acquisition of data through various types of systems, such that is scientifically derived, mathematically proven, and algorithmic methods are used in the acquisition, preservation, validation, identification, analysis, interpretation, and documentation of huge datasets with varying data types from diverse sources for the purpose of facilitating or furthering future decisions.

According to {Jianqing Fan, et.al., 2014} the emergence of big data has brought so many new opportunities to our modern society and challenges to data scientists as well. The essence of big data originated from the need of large companies, such as Yahoo, Google, and Facebook, Microsoft, IBM, to analyze inevitable large amounts of data (Garlasu et al., 2013). The type of information being created today is no more the usual traditional database-driven data which we refer to as structured data rather the data includes so many types such as documents or text, images, audio, video, and social media contents known as unstructured data or big data {Jasmine Zakir, et.al., 2015}. For data to be termed as big data, it must have some characteristics provided that define or associate any data as big data which some call the v’s. Data researchers, scientists, and experts define big data regarding different views of the “v’s” characteristics. The characteristics are aimed at comprehending the conceptual foundation, vision, and trend of big data. Due to the characteristics of big data, the technology used to process this big data like Hadoop ecosystem technology needs a high processing speed to process data quickly and huge size storage to ensure timely and effectivity in analysis {Jie Song and Jin Li, 2020}.

From looking at these definitions it shows that there are important characteristics that define big data. These characteristics have evolved through the years from 3v’s to 5v’s to 7v’s characteristics as big data continually emerge. Literature firstly defined and characterized big data as having 3 main v’s – volume, velocity, and variety {Demchenko et, al. 2014} But even more characteristics have been added by some studies to better define big data. The 3v’s were found too limiting and insufficient to define Big Data.

Big Data cannot be understood only in terms of the data alone but must be seen in the light of other contexts like its environment that is iterative and potentially open-ended {Stanier and Isitor, 2016}. Big data was then extended in the context of other perspectives like the scientific view to include value and veracity making it 5v’s characterized {Demchenko, et.al, 2013}. Big data was also stretched from the commercial data view incorporating veracity, variability, and visualization. Today we now see a 7v’s approach to big data, which extends the study to the 3Vs with veracity, validity, volatility, and value {Khan, et.al, 2014 as detailed in the next section. Other studies added even more characteristics to better define big data which include the vision or purpose of big data, verifying that the processed data follow certain specifications, validating that the main purpose is fulfilled, and complexity which explains the difficulty in organizing and analyzing big data and its data relationships and immutability {Gandomi and Haider, 2015}. The next section focuses on big data characteristics.

### Characteristics of Big Data

The majority of these big data v’s characteristics {Ibrar Yaqooba, 2016} are briefly explained below:

1. Volume: This attribute of big data relates to data generated by humans, corporations, or sensors from different sources, the size, and the scale of the data. The magnitude of data is very large certainly in petabytes and zettabytes. So many studies perceive volume characteristics as the most visible big data characteristic.
2. Velocity: This attribute refers to the speed of incoming and outgoing data. The data generated is very fast and continuous coming from different data sources. {Bedi, et.al, 2014} define velocity as the rapidity with which data is received, stored, processed, and analyzed. This is sometimes referred to as real-time, near real-time, and streaming of data.
3. Variety: This attribute refers to how the data generated is derived from diverse sources, in different variations or forms such as some are text, images, video, audio, multimedia, logs, or mail to mention a few. Big Data always arise from multiple sources, from outside and inside organizations, including both the traditional relational data and all other archetypes of unstructured data sources. As such data entails both structured, semi-structured, and unstructured data types. Managing the different varieties of big data comes with its challenges, especially with unstructured data.
4. Value: This attribute refers to the worth of the hidden insights or understandings derived inside big datasets. Data must be converted for value, uncover trends, and extracted to improve organizations. The added value of big data is always retrieved at the end-product of every analysis of the data.
5. Veracity: This refers to the reliability, certainty, and trustworthiness of the data about how data is collected, methods with which data is processed, its infrastructure used, and how data is sourced. Data must be as accurate and authentic as possible. While some studies define veracity in a scientific way to mean how uniform and consistent ought to be, veracity simply means the truthfulness of the data {Bedi, et.al, 2014}.
6. Variability or Volatility: This refers to how constantly changing the data flow, streams or data sources of big data are. Big datasets are randomly gathered, they lack consistency or a fixed pattern in the structure and sources of the data. This invariably causes inconsistency, disorderliness and unpredictable rapid changes that makes the data complex and hard to manage. For example, incoming data streams producing tremendous datasets that are varying and without structure.
7. Validity: This attribute means the correctness and accuracy of the data with regard to the storage for analytics. The quality of the data must be placed under consideration when storing the data for use as for effective use, the data must be logically and factually sound, correct, and truthful so to avoid wrong analysis for decisions.

Although there are multiple definitions of big data, the current dissertation draws from these definitions (among others) and understands big data to be derived from multiple sources, having very large datasets (in terabytes) in real-time, combined in both structured, semi-structured and unstructured data types used to facilitate analytics and data insights. As such frameworks, systems, technologies, applications, and models are designed with considerations of big data characteristics, attributes and to accommodate large storage capacity, parallel processing, and real-time analysis of different heterogeneous data sources {Ahmed Oussous, et.al, 2018}

### Big Data Types and Structures

No matter the way data is sourced and presented, they are still very useful for retrieving data insights and analytics. Big data is mainly used in platforms like the cloud where resources, infrastructures, and services reside, which makes big data technologies accessible and affordable to any enterprise. These cloud platforms are managed and maintained by big data cloud storage providers like AWS, Google, Oracle, and Microsoft. Big data sources extend beyond the traditional corporate database of structured data type to include emails, mobile device outputs, and environment sensor-generated data where data is no longer restricted to structured database records but rather includes unstructured data that possess different data dimensions, having no standard formats or uniform definition. . Data types for big data are generally in different varieties ranging from structured, semi-structured and unstructured data types. Brief definitions and examples of these big data types are:

1. Structured data types: This data type is carefully arranged and organized in a predictable order. An example is databases as data is arranged in tables, predefine sorted out in rows and columns. This type of data has no challenges when analyzed as they are easy to analyze since it is ordered and properly arranged.
2. Semi-structured data types: This data typically comes in a combination of both ordered and unordered data types. Examples of these unstructured data types are JSON, XML, and CSV as they allow you to file any type of data types
3. Unstructured data types: This data type is unorganized, never in any order, and has no pre-defined structuring. As such effective mechanisms and techniques must be used to clean and extract meaningful insights from unstructured data types. Unstructured data are always complex having a combination of different data types like images, text, audio, and sensor data. One most common potential source of unstructured data is data generated from social media files like Facebook posts and tweets as they are multimedia data types. Examples are files like JPG, PDF, MP3, DOC, and PPT. Noting that unstructured datatypes are very large, unclear, ambiguous, disorganized, and noisy, it has to undergo the pre-processed steps of filtering, cleaning, and separating data to be able to discover useful information and recognize data patterns. {Rai, 2017} observed that unstructured data types contain a lot of useful information hidden however, the main challenge of unstructured data types is in the process of separating, cleaning, and analyzing data types, some important portions of the data are likely to be discarded and lost.

On the other hand {Samira Pouyanfar, et.al., 2018} observed that multimedia data are always embedded within structured and unstructured data types and must be captured or represented likewise. As such the authors categorized multimedia data for analytics into five groups of multimedia data, i.e. visual data, audio data, text data, sensor data, and other data; according to the data type content.

1. Visual data: This type of data is daily generated in large amounts due to increased use of the internet, the presence of IoT, mobile technologies, and increased cloud computing. Visual data includes data with image sequences like images and video data information. Some systems perform image classification with visual data like Microsoft ImageNet. Examples of visual data application extraction from big data sources are retrieved from video surveillance systems, autonomous vehicles, video and image retrieval systems, and healthcare. {Diana P. Tobón V, et.al., 2022} assert that convolutional neural networks (CNNs) are the most popular models used for image feature learning for analytics.
2. Audio data: This type of data is in audio, speech recognition, music, sounds, and aural data information. Examples of audio data applications from big data sources are social media, industrial machines, and medical devices example wearable technologies for measuring and analyzing human heartbeats rate and breathing sounds.
3. Text data: Multimedia data also include textual data such as metadata, web pages, social network feeds, and survey data. This type of data is analyzed with text analytics techniques and frameworks such as SparkText, information extraction (IE), sentiment analysis {Lei Zhang and Bing Liu, 2017} summarization, and Question Answering (QA) to extract relevant information from large-scale unstructured text sources.
4. Sensor data: According to {Samira Pouyanfar, et.al., 2018} the main power of the IOTs today are sensors and actuators which are embedded in smartphones, wearables, tablets, and other connected objects. They are used extensively everywhere. Examples of such sensor data applications sources in big data include monitoring environmental conditions, meteorological patterns, airplane engines, industrial equipment, and patient vital statistics.
5. Other data types: Multimedia data are not only limited to visual, audio, and text data types but also include user-generated content, relationship connections, clicks, likes, comments, news blogs, geographic data, genomics datasets, and so on which are mostly found in social media networks.

## 2.3 BIG DATA APPLICATIONS

Today big data is applied in various areas for analytics and retrieving insights. {Ahmed Oussous et.al, 2018} observed that so many organizations from different sectors depend progressively on the knowledge that is obtained from heterogeneous sourced huge volumes of data daily generated at an unparalleled rate, for business purposes. Studies show that the world is now tending towards adopting a smart city ideology where big data applications are used to support smart components to improve human lives {Eiman Al Nuaimi, et.al., 2015}. The authors believe that the best way to achieve the smart city concept is to rely on cloud computing platforms for collecting, storing, and analyzing big data in real time. The continuous generation of this huge, diversified data is due to so many factors like the explosive use of the internet of things (IoT), the proliferation of cloud computing, as well as the spread of smart devices, used today. However, for effective use of these data, powerful systems and distributed applications are needed and used to manage and support such multiple connection systems. Although today we now have different big data systems and devices that handle any type of big data as batch-processing, stream processing, or SQL processing. Management mainly in its storage capacity, scalability, flexibility, the performance needed, and enabling parallel processing of very large volumes of data context. To enable an efficient ability to extract knowledge from big data, numerous models, programs, software, hardware, technologies, algorithms, and approaches used to process big data have been designed and proposed. Amongst other sources of big data, most big data types are multimedia data that grows very fast as every individual connected to the internet generates multimedia data from various sources {Yousafzai, et al., 2016}, such as text, images, audio, video, and graphic objects. This dissertation is mainly focused on multimedia data type sources of big data applications. Big data applications are:

* **Social media data**: This platform aims to enable communication amongst people. They capture and post varieties of data like snapshots, tweets, videos, audio files, and images. The use of social media causes an increasing surge in data generation such as Twitter, Facebook, Youtube, LinkedIn, Instagram, and Whatsapp. Social media big data are practically applied and used for marketing purposes and analytics {Wendy and David, 2017; Jie Yang et.al., 2022}, for hospitality and tourism {Yung-Chun Chang et.al., 2019; {A.K. Siti-Nabiha, et.al., 2021}, for disaster management {Tan Yigitcanlar, et.al., 2022; {Jedsada Phengsuwan, et.al., 2021}, disruptive technology and so much more. The majority of analyzed social media data are mainly done using sentiment and content analysis as methods.
* **Black box data of aircraft and automobiles**: This captures varieties of information that’s happening inside the aircraft from voices, the recordings of the microphone and earphones to the performance of the aircraft. {Devi and Prabha, 2019} proposed a solution by analyzing previous accidents for safety and security concerns. The system has features that assist in reducing the number of accidents by providing an automatic accident alert system informing the nearest hospital, and the nearest traffic authority, by providing advanced web tracking, the coordinates of the accident location, and exact physical address for immediate medical attention.
* **Environmental sensor data**: These are time series data and are mostly crucial to managing in real-time. Hence, they require dynamic analysis based on real-time data. Example sources of sensor data are weather data from satellite images, scientific data, radar data such as vehicle movements, and data collected from smart grids operation and GPS. { Paolis, et.al., 2018} collected sensor biomedical data about the breath of patients to produce alerts that warn doctors in cases of problematic situations. All data is captured from these sensors and sent to data servers for storage. This is achieved through multiple connections among smart meters, sensors, control centers, and other infrastructures.
* **Agricultural purpose data**: Big data is applied to the food supply chains and farming production resolves. For example {Akshit Singha, et.al., 2018}used Twitter data to identify supply chain management issues in the food industries. Another study by {Wittingham et al., 2020} drew data from 522 public Twitter accounts to ascertain how individual differences of users like their personality traits and values influence the perception of whether natively improved food items are considered safe or not.
* **Internet of Things (IoT):** They are connected appliances, wearable health monitors, smart cities, and devices or systems. IoT is one of the main drivers of big data in our world today due to its high variability of objects and appliances involved. Examples are tracking vehicle’s locations with sensors, log files, wireless adapters, and geographical positioning systems (GPS) and storing this various information. {Sayed Khushal Shah, et.al., 2018} used IoT sensor audio records to propose a home automation system for safety in domestic violence situations. The audio data is immediately sent to a machine learning server for classification.
* **Transports and logistics data**: According to {Rajaraman, 2016} radiofrequency identification (RFID) and GPS are used to public track buses to explore the data for optimizing bus routes, and improving roads and transport services. {Muhammad Awais Javed, et.al., 2019} emphasized the importance, the benefits of data analytics for transport systems, extracting useful and relevant information for traffic management and efficient mobility.
* **Crime data**: Big data is used and applied for crime investigations {Mingchen Feng, et.al., 2019}. According to {Andrii Shalaginov, et.al., 2019} some novel technologies and big data sources that facilitate timely crime investigations as the use of smart devices, drones, the internet of things (IoT), opensource intelligence, social network analysis such as Twitter and Facebook. This is because they can be used as intelligent linkage graphs for understanding user social behavior, and tracking unusual and criminal patterns {S. A. Catanese, et.al., 2011}.
* **Google search engine data**: Optimised searches are continuously done daily causing massive amounts of web-based data collection. These data are analyzed to ensure improved visibility, easy search, and findability on the web. {Ioannis C. Drivas, et.al., 2020} proposed a novel computational development methodology to increase organic search engine visits based on the impact of multiple search engine optimization strategy factors. The authors used and explored 171 cultural heritage websites, their retrieved data analytics about their performance, and user experience for the data analytics. Big data analytics technologies that are used in big data applications for discovering data patterns and insights are elaborated on in the next section.

## 2.4 BIG DATA ANALYTICS TECHNOLOGIES

Data is always in its raw form when collected and is not useful to any organization. As a result, data processing is needed for collecting the raw data and translating it into usable information for better business strategies and decision-making. The raw data must be collected, filtered, sorted, cleaned, processed, analyzed, and even stored, before they are presented in a readable understandable format. The Oracle research (2020) explains that organizations are more able to access so much data today than ever before but this huge accessed data are of no value unless they know how to put big data to work. Big data analytics technologies are used for this purpose, to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences to mention but a few. According to {Andrii Shalaginov, et.al., 2019} big data is now with certainty perceived as a value for key insights on data for efficient business strategic decision making.

Big data analytics technologies and techniques have emerged due to traditional data techniques and platforms becoming less efficient {Ahmed Oussous, et al., 2018}. For example, they show slow responsiveness and lack of scalability to size data as a result of having limited storage capacity; lack of flexibility and rigid management tools, poor performance, inaccuracy in analyzing, and becoming too expensive to maintain. This brought about a need for big data context and technologies. Big data analytic tools and technologies enable effective work with both structured and unstructured data coming from multiple heterogenous sources. They provide analytic algorithms and models that are easy to integrate with other technologies in an organization. Data mining tools are used for processing the data and feature analysis. Examples of data processes are data cleaning, data transformation, data reduction, and compression. Examples, where feature analysis is used, are for analyzing videos, audio, texts, motion, and spatiotemporal data. These new adaptive tools and technologies became a necessity as research shows that big data requires significant resources, new methods, and powerful technologies. With considerations for the multimodality and complex nature of big data representations and complex for efficient usage and real-time accuracy in the processing, analyzing, securing, and storage of data. Also, ensuring it competently provides granular access to today’s massive evolving data sets. Big Data Analytics Pipelines are also needed to be enabled to build and deploy big data analytics in the real world, from the data configuration, collection, verification, feature extraction, analysis tools, having the right infrastructure, and monitoring tools and components. Some of the popular tools used for data sharing use sharing systems like the cloud, online file sharing systems, and wireless data sharing.

Steps in processing big data are:

* Data collection is always the first process where raw data is gathered from defined heterogenous and accurate sources to ensure that the results are valid, relevant, and usable.
* Data ingestion and acquisition is the process of getting the data into Hadoop for processing. Tools used are Sqoop, Flume, or Apache Storm.
* Prepare the data: Raw data is always messy and unorganized so there is cleaning, sorting, and filtering of the raw data to remove every unnecessary and inaccurate data, errors, duplication, miscalculations, or missing data. To transform data into an appropriate fit for further analysis and processing.
* Input data: For the raw data to be usable, it must be converted into machine readable form and fed into the machine processing unit by an input device such as keyboards, scanners, mouse, digitizers, microphones, joystick and so on.
* Process data depends on the source of the data like if the data is in batch or bits, in real-time within seconds, if data is online and continuous, if data is multiprocess in parallel such as weather forecasts or if data is time-shared example, where multiple users share one computer simultaneously as data is placed in time slots to the several users. Also, the use of the data and the processing methods like machine learning, deep learning, and artificial intelligence algorithms, to generate the desired output are significant.
* Data mining is a process of discovering insights within a database. The main tools for data mining are Teradata and Rapid miner.
* Output data: As soon as data processing is completed, data is transmitted and displayed in a readable form like graphs, tables, vector files, audio, video, documents, images, videos, and so on to the user. The outputs are stored for further processing when needed like in the case of deep neural networks.
* Storing the data for further use is very important as this allows for quick access and retrieval of information whenever it is needed.
* Data reporting is used to show reports of data insights discovered from processing the data for analytics. One of the primary tools used is the PowerBI.

Big data analytics platforms like Apache Hadoop and Apache Spark platforms have been added to an ecosystem that completely enables people to perform different types of big data analytics tasks ranging from handling the data processing, analysis, resource management, coordination, and data input. The ecosystem is not limited as different capabilities are made easy like any data type input process is taken like batch or stream, any language can be used whether java, python or SQL. Frameworks of big data have emerged for handling big data analytics. Big deep learning leverages Apache Spark and has a data framework for SQL, Spark, streaming data, machine learning pipelines, and distributed training architectures. The distributed training has a parameter server architecture that handles the gradient and weight parameters for each network node and neuron. It also has libraries like Caffe, Torch, and Tensorflow. Apache Spark examples are BigDL, TensorFlow-on-Spark, Deep Learning Pipelines for Spark, SparkNet, and so on.

In general, many technologies, tools, frameworks, and software for analyzing big data types are open source. Some are used for storing the data before analysis.

### Big Data Preprocessing Technologies

According to {Farhad Mehdipour, et.al., 2016} big data processing is a set of tools, techniques, or programming models to access large-scale data to extract useful information and insights for supporting and providing decisions. Big data processing today now exceeds only data acquisition and data cleaning but extends to other five steps which are data extraction, data transformation, data loading, data visualization or analytics, and applied machine learning.

* Data extraction entails collecting and extracting information from diverse resources through many unstructured and structured data streams such as enterprise applications, web pages, sensors, marketing tools, transactional records, and so on.
* Data transformation entails modifying the data into required formats that are understandable by users so it helps in building different insights and visualizations. Examples of transformation techniques that are used are aggregation, normalization, feature selection, binning and clustering, and concept hierarchy generation.
* Data loading involves using the big data extract transform and load (ETL) process of automating load converted or modified data in batch, streamed, or real-time into a centralized database system. Before loading the data, index the database is indexed and constraints are removed to make the process more efficient.
* Data visualization enables data analysis, allowing a better understanding of data with the data patterns and correlations or relationships between the attributes.  Also, it allows visualize huge datasets and create dashboards that summarize the data.
* Applying machine learning: This final stage permits the creation of models that can learn to evolve in the response to adding new input(s) probably in the future. The learning algorithms allow for a more quickly analyzing of large amounts of data.

Some big data processing technologies that are used are:

1. **Apache Spark**: This is a low latency, distributed data processing framework running as a library on a java node. It has many Spark tasks running on a cluster such as YARN clusters, Mesos clusters, Kubernetes clusters, or standalone modes, each working on a separate data partition. One important feature of spark is its ability to capture a wide range of processing workloads that previously needed separate engines since it performs in-memory computations which enables fast data processing when compared to traditional disk processing, such as including SQL, streaming, machine learning, and graph processing.
2. **MapReduce**: This is used for data processing in different languages. It can shrink large amounts of data into small sets so its easier to analyze
3. **Apache Hadoop**: This is a scalable open-source computation framework that has two main components which are the MapReduce execution engine and distributed file system called Hadoop Distributed File System (HDFS). According to {White 2012} Hadoop allows partitioning of the computation processes across many host servers even when they are not necessarily high-performance computers. Hadoop is mainly used to access the systems input format API using HBase, and HDFS. Hadoop has many advantages such as its high flexibility, scalability, low cost, providing job schedules for balancing data, resource and task loads, and reliability for managing and efficiently processing large volumes of both structured and unstructured datasets {Ticiana L. Coelho da Silva, et.al., 2018}.
4. **YARN:** stands for Yet Another Resource Negotiator developed from Hadoop and is an architecture that decouples the programming model from the resource management infrastructure as it delegates scheduling functions to per-application components.
5. **Apache Flink**: This is open-source, mostly used in streaming execution and batch processing workloads for distributed and high-performing applications such as real-time analytics, and continuous data pipelines. It leverages in-memory storage for improving the performance of the runtime execution.
6. **Hyracks/ASTERIX**: This is a partitioned-parallel software platform mainly designed to run data-intensive computations on large shared-nothing clusters. Hyracks provides fast data ingestion and a scalable information management system that supports the storage, querying, and analysis of large collections of semi-structured nested data objects. Hyracks delivers a more flexible user model that allows quick fault tolerance and recovery performance.

### Big data storage and management tools and technologies:

Studies show that one main challenge of big data is the ability to store and process data within a specific time. Some of the popular tools used for data storage and querying/analysis and management of big data are Apache Hadoop, Microsoft HDInsight, Talend, NoSQL databases, Hive, and Apache Zookeeper.

1. **Apache Hadoop**: Hadoop is a big data framework extensively used to store data, and handle the volume and complexity of big data. It comprises several open source tools that are used for processing, managing, and analyzing huge data types. Hadoop has environments that can be configured to process real-time data types as they could be used for distributed storage and processing of big data as clusters. This enables tasks to be completed much faster. Hadoop is known to split big data and distribute across many nodes in a cluster {Rai, 2017}, scaling from a single server to thousands of systems hence creating scalable storage of data as a Hadoop distributed file system.
2. **NoSQL databases**: This means Not Only SQL data and is suitable for storing massive amounts of unstructured data types. Examples of popular NoSQL databases here are MongoDB, Apache Cassandra, neo4j, and Apache HBase.

* MongoDB – compatible and ideal for fast real-time data. MongoDB is popularly used, very flexible, and can integrate with any third-party tool. Examples are e-commerce applications like eBay and Amazon.
* Cassandra – This is mainly used in Apache Spark. It enables efficient management of large data, good fault tolerance, low latency, and supports data replication for scalability,

.

1. **Hbase**: This is a distributed non-relational database. It is an open-source project that is built on top of HDFS. According to {Ahmed Oussous, et.al., 2018} Hbase supports high table-update rates, providing flexibly structured data hosting for very large tables in a big table-like format and scaling out horizontally in distributed clusters. The author further emphasizes that HBase is even more flexible than relational databases as it allows many attributes to be grouped into column families so that the elements of a column family are all stored together.
2. **Azure data lake**: This is provided by Microsoft for data storage in the cloud.

### Big data cleaning tools

Data needs to be cleaned to remove errors, inconsistencies, or missen data, defined, reshaped, and well structured for user understanding. Examples of data cleansing tools used here are Microsoft Excel, Open Refine, and Trifacta Wrangler.

1. **Microsoft Excel**: Data is arranged in rows and columns and mathematical functions are used.
2. **Open Refine**: This is used for cleaning messy data, exploring large data sets, and transforming it from one format to another. The cleaned data is then uploaded into a database.
3. **Trifacta Wrangler**: This is a connected desktop application used on large data, supports cloud and other deployment channels for discovering, structuring, cleaning, enriching, and publishing data of any shape and size.

### Big data visualization tools

These are useful ways of showing representations of the complex discovered data insights so it is easy for users to understand. Examples of popular tools used for visualization are Tabluea, PowerBI, Plotly, and IBM Watson Analytics.

1. **Tableau**: This is a secure and end-to-end analytics data visualization platform used mainly in business intelligence industries for business analytics. It can extract data from any database platform ranging from PDF, Excel, Amazon web services, Oracle, and so on. Tableau can be connected to Hadoop.
2. **PowerBI**: This is a business analytics platform provided by Microsoft. It provides interactive visualization and business intelligence capabilities through easy and simple interfaces for users to do analytics. PowerB can connect to Excel spreadsheets, connect to both cloud-based data, and on-premises data warehouses, and uses SQL.
3. **Amazon quick site**: This tool is for visualization mainly in cloud data.
4. Clicksins and Cysins for visualisation

### Big data analytics tools and technologies

This process entails asking questions to understand the data by answering questions regarding the data. Tools used are Hive, Pig, Hadoop Map Reduce, and Apache Spark.

The Hadoop ecosystem is a very popular framework for big data analytical tools because it is open-source and free. It is used by almost every big data analytic company. Hadoop ecosystem has every tool which is needed to create a complete big data analytics system. Tools in Hadoop include:

* Hadoop MapReduce: This tool is a programming model that is used to access big data that is stored in the Hadoop file system (HDFS). It facilitates data processing by splitting large petabytes of data into smaller portions. This is executed inside the server where the data already resides making the process fast and quick.
* Hive and Drill: This is an open-source tool that allows the analysis of large data sets on Hadoop. It is fast and mostly used for summarising, querying, and analyzing data.
* Mahout and Spark MLib for Machine learning and data procesiing,
* Pig for scripting languages,
* Apache Spark is open-sourced tool, used for data analysis. It handles both batch data types and real-time data. It performs with in-memory data flow engine for in-memory data processing.
* Kafke and Strom for streaming,
* Apache Spark MLib, Solr, and Lucene for searching and indexing,
* Oozie for scheduling data tasks,
* HBase used for the NoSQL database instead of MongoDB
* Zookeeper for management and coordination of every single tool in the Hadoop ecosystem,
* Hadoop Yarn for resource management and used to manage all the other tools in the Hadoop ecosystem,
* Hadoop HDFS for storage of all the data coming into Hadoop both structured, semi-structured and unstructured data types. HDFS is the first place where data is placed before data is moved to any other place in the Hadoop ecosystem,
* Sqoop for structured data
* Flume for both semi-structured and unstructured data

Some other big data analytical tools are:

Tensorflow for building and training models especially neural networks.

Amazon red shift – used for data analytics.

## 2.5 BIG DATA CHALLENGES

One of the main qualities of big data is the continual exponential increase in the size of data and distributed sources. However, the current technological capacity to handle and explore these bid data sets is still gradually evolving. Technical or systematic ways of efficiently collecting, integrating, and storing these big data with fewer hardware and software requirements, with the consideration that data is always streamed from different sources are of great concern. This is what the Hadoop distributed file system does as it creates multiple copies of data so when one system fails, there is another to get data on time. It hence enables proper management of big data sets for good analytics, extracting value and reliable insights.

With big data, which is generated from so many sources, data that arrive are unpredictable and uncertain looking complex or complicated, and need to be cleaned, sources verified to ensure reliability and remove noises, errors, or incomplete data before it can be used. As a result of big data containing very large data sets, it could be thought-provoking to ascertain how to decide which data is reliable and useful. Also due to the large size of data involved, the process or technique of data compression is used to reduce the size of storage space. The challenge of ensuring a lossless compression during the process of data reduction where important data necessary for analytics is not lost in the process is very important. Thus, advanced algorithms and efficient methods of data mining are needed to get accurate results, monitor the changes in various fields, and predict future observations.

Another challenge is being able to identify profile descriptions that refer to the same real-world user identity profiles amidst voluminous identical attributes and user profiles from the massive data. This can be very challenging. This is even more elaborated as there are no limitations as to creating profiles on the web, for example, where people create as many possible identity profiles for just one same particular user most times is never linked allowing ambiguity and identity uncertainty in a big data environment. This is a necessity today due to cyber-crimes that increasingly occurs almost daily.

Amongst all these challenges mentioned, assembling meaningful information and competitive advantages from massive amounts of data is very challenging looking at the characteristics of the data in its velocity, speed and streams the data flow distributes in real-time. Also, the variety of data comprising of different formats like text, image, audio, and video coming from diverse sources. According to {Jasmine Zakir, et.al., 2015} in big data, most data is stored in a nonstructured manner, using different languages and formats, which, in most cases, are all compatible. Trying to efficiently extract meaningful insights from such data sources quickly and easily is very challenging. It is important to note that analyzing big data is quite new so managing this big data for analytics and insights with its tools and technologies do require a skill set that is new to most IT departments and organizations, as there is the need to know how to integrate all the relevant internal and external sources of data. As such it is very difficult for users that want to leverage deep learning tools and capabilities to apply the new technology in production.

Another major challenge with big data is according to {Andrii Shalaginov, et.al 2019} that there is a lack of standards and insufficient security awareness in managing the new big data technology landscape the world is in right now. Considering this, one needs to enhance forensics investigation methodologies, employ innovative tools, technologies, and models, and combine threat intelligence with the focus on effectively using and managing big data insights.

## 2.6 BIG DATA IN THE CONTEXT OF CRIME INVESTIGATIONS

As a result of the continuous expansion of data, organized crimes continue to rise simultaneously. Crime here relates to malicious activities and misconducts targeting information and communication technologies such as distributed denial of service attacks, that are enabled by computer systems, and networks and committed using ICT {Andrii Shalaginov, et.al., 2019}. Most times criminal attacks are aimed at compromising confidentiality, integrity, and availability of the computer network systems. As such law enforcement and security agencies are faced with the necessity of improvising advanced technological means to fight the battle against crime. ….. observed that one extremely important part of policing and crime prevention is being able to understand the when, the who, where, the what, and how of a crime. The authors noted that law enforcement agencies are collaborating with research and technology to assist with improving their capacity to manage crime. This has brought about studies on digital forensics (DF) and implementing digital forensic readiness (DFR) to enhance successful cybercrime investigations {Andrii Shalaginov, et.al., 2017}. The authors emphasized the importance of advancing big data analytics for digital forensics towards curbing crimes as experts being able to efficiently correlate data from distinct crimes and crime scenes can be so challenging for experts. They referenced {A. Guarino, 2013} discusses digital forensics as a big data challenge and thus requires complete reorganizing of principles, technologies, models, frameworks, and development of new tools and skills to enhance efficiency in correlating different crime data from different sources to curbing crime. DFR is quite new, it allows businesses to pre-empt and plan how to adequately manage cybercriminal incidents should in case it occurs.

Also, the current improvements to information and communications technology (ICT), smart devices, and internet accessibility for example most daily activities are automated unlike when compared to older computer technologies. Even with the positive improvements that this automated process brings to humanity, they have furthered and improved cyber criminals' ability to attack the security of infrastructures, data, and computer network systems {Andrii Shalaginov, et.al., 2019}. This has invariably led to multiple attack trajectories and vectors where we now see cybercriminals affecting even much more victims all around the world than they could before. Today big data has become very relevant for crime investigations where security professionals can properly retrieve knowledge and extract hidden network structures about criminals. Hidden criminal structures like identifying criminal members and detecting criminal subgroups greatly reduce crime, helping law enforcement and intelligence agencies develop effective strategies that efficiently prevent crimes from happening {M.I. Pramanik, et.al., 2017}. The authors discussed five different data mining technologies that have been used to combat crime which are link analysis, intelligent agents, text mining, neural networks, and Machine learning.

Link analysis shows a representation of the structure of the data as a set of interconnected, linked objects or entities. This invariably reveals hidden relational patterns from entities and their relationships which allows investigators, and analysts to discover useful links among entities or relationships.

NOTE: Tensorflow good for neural networks; uses python – computation involves tensor which is a standard way of representing data in deep learning

* Tensor uses. Steps or life cycle of ML or DL are: collecting the dataset (Majority of the work) – then build my model (few lines of code) – then train the network (one line of code) - then evaluate model (one line of code) – then predict the model (one line of code). Tensorflow is easy to learn.
* Tensorflow is built in Python and C++. Its best for larger projects and more complex workflow
* Other ML frameworks like Tensorflow are Pytorch, Microsoft CNTK and NEXT
* Pytorch – is built in Python, has hardware accelerated components and a highly interactive development model that allows for designing. Pytorch is best choice for fast development projects that needs to be running on a short time.
* Microsoft CNTK is Microsoft cognitive toolkit, it uses graph structure to describe the data flow. It focuses more on creating deep learning neural network as it handles DL neural network jobs fast using API for Python, C++ and C sharp and Java. But its not easy like tensorflow.

## 2.7 CHAPTER CONCLUSION

Having presented a brief overview of big data, its characteristics, types, applications, analytical tools and technologies, and challenges, big data in the context of crime investigations is also discussed as in this dissertation. Although there are tools and technologies to manging big data, there still exist challenges in the use of big data especially involving crimes such as data recovery, cybercrime, fraud, cyber-bullying and intellectual property theft. These crimes need to be investigated in order to prosecute attackers or perpetrators and protect victims. This is only achievable with digital forensics as a method for investigating criminal events involving data in digital format to be used in the court of law to confirm admissible evidence in criminal investigations or in other civil proceedings. In the next chapter, digital forensics is identified and discussed.

# Chapter 3: DIGITAL FORENSICS

## 3.1 INTRODUCTION

**Part of Digital Forensics Chapter**

the types and the amount of data that need to be processed in cybercrime investigations are increasing and finding clues or evidence of crime from the massive data can be very challenging. The emergence of big data technology has brought new directions for digital forensics investigations. It is of practical importance to integrate big data technology and digital forensics to crack down on cybercrime {Jie Song and Jin Li, 2020}.

The first step in the digital forensic process is the ability to rightly identify parties – whether suspects, victims, witnesses or attackers. This first step of resolving the user’s identification is very important as all other steps builds on this first step. As such its key that investigations are done on time so as to allow criminal activities to be prosecuted. This is the crux of this dissertation.

propose a conceptual model to associate big data forensics investigations and provide new insights to investigate criminal incidents. - BIG DATA FORENSICS: CHALLENGES AND APPROACHES Dr.A.CHANDRABOSE T.INDHU RANI

## 3.2 DEFINING DIGITAL FORENSICS

# Chapter 4: IDENTITY RESOLUTION

## 4.1 INTRODUCTION

The background chapters of this dissertation start in Section 22 with a brief overview of identity, digital identities, attribution of digital identities, relationships of digital identities, and identity attribution in the South African network environment. This is followed by Section 2.3 which focuses on issues of identity ambiguity which limits identity verification to enable digital evidence forensic soundness to attain admissibility in any court of law. This is the concentration of this dissertation. Approaches for identity resolution are discussed in next Section 2.4. Identity resolution in big data digital forensics is discussed in section 2.5 and the chapter concludes with a summary in section 2.6. The backgr

ound chapters continue with Chapter 3 focusing on big data as they apply to resolve victims and attacker identity in the South African network environment for the context of criminal investigations. Chapter 4 concludes the background chapters with a look at digital forensics as they apply to resolve identities in big data, its processes, requirements, software applications, tools, and techniques.

Since this dissertation deals with the identities in a network environment that is used by law enforcement agents (LEAs) and forensic investigators in their quest to correctly identify the identity traits of victims and attackers in a crime, it is necessary to first introduce the concept of identity.

## 4.2 IDENTITY

Although there have been several attempts at defining identity in the literature. Identities can be a person or individual based on distinguishing characteristics such as physical traits, biographical data, personality, beliefs, culture, roles, psychological self-image, or perceptions {Michel, 2020} In this dissertation, however, the focus is on digital sciences, which is referred to as digital identity in a cyber environment or digital platform. (Phillips et al., 2020) defines identity as characteristics that distinguish one identifiable individual from another. Identity is derived from a set of attributes that simply describes an individual identity. Some studies define identity from two or more aspects which are personal identity and social identity (Wang, 2015). Personal identity attributes are identifiers that are commonly used in record management systems to distinguish one person from others. They describe one’s self-perception as an individual having personal identifiers like name, date, place of birth, social security number, height, weight, and biometric information such as fingerprint, and DNA. Whereas a social identity is one’s biographical history that builds up over time. Social identity relates to a social context like being a member of certain social groups example country, culture, and social media platforms.

### Digital Identity

According to {Jenkins, 2021} digital identity is the gateway to the digital world for a person, organization, or device as it uses sets of attributes to identify a particular person, organization, or device in the digital world. Bailur, 2020} on the other hand defines digital identity as an indication of converting human identities into machine-readable digital data. Digital identity goes with three functions of identification, authentication, and authorization; all are performed digitally as noted by {Nyst, 2016}. Identification here is the process of establishing information about an individual, authentication on the other hand is the process of declaring an identity that was previously established during identification using a form of credentials like username, single sign-on, or passwords. While authorization involves the process of determining what actions the individual may perform or services they can access based on an authenticated identity. Digital identity arises organically from the use of personal information on the web and the shadow data created by the individual’s actions online. They are called shadow data because users are always unaware that everything they do digitally online automatically saves and can be reused for so many purposes like investigations and personalization marketing. This information is often used by website owners and advertisers to identify and track users. This information automatically links and relates to the incident evidence and the scene of digital evidence such as the time and location of the incident. A digital identity is the body of information about an individual, organization, or electronic device that exists in cyberspace online.

(Mary C. (Kay) Michel, 2018) observed that some important traits can be referenced from cyber to identify a real-world individual. A situation where one same user is having two identities for one identity brings a lack of integrity. Facebook’s Mark Zuckerberg promotes the ethics of an integrated identity where a single version of selfhood is maintained across diverse contexts and human relationships (Brusseau, 2019). This is mainly achievable with emphasis on the reality of users being authentic by employing real names and having a single account. According to (Phillips et al., 2020) in digital data, obtaining the true identity (or identities) of an attacker or victim can be a challenge for law enforcement agencies due to identity ambiguity as a result of misspelled data, variations in naming order, case sensitive data and inconsistencies which could be intentional or unintentional. This misleads investigations. As such using personal or biological identity attributes alone is seen as contextual since a user gives selective information which can be falsified and deceptive when providing authentication information. So, the question points to the problem of how to tell that multiple identity reference points to the same user identity.

In cases of numerous deduplicates of a particle that refers to the same identity, this can create privacy and security risks like identity theft and hacks. Individuals may also claim multiple digital identities across a variety of communities like opening a new account, accessing existing accounts, and gaining credibility to engage with online services. In a normal situation, each identity structures and attributes correspond to one, and only one, real-world entity. This implies that each identity has attributes and activities that must distinguish them like social media photos, phone numbers, and id numbers but they can be stolen, lost, or damaged. It must be noted that for any real-world object to be represented in any data set, there should be only one record. But this is almost impossible as there are so many causes that bring about the existence of ambiguity and duplicated data about unique entities. This is the main goal of identity resolution. Identity resolution is the collection of algorithms to standardize, normalize, and then compare data values to establish that two different records refer to the same entity or to determine that they don’t relate. (Wang, 2015) asserts that identity resolution is a special type of entity resolution that specializes in identity management. He claims that the use of only personal identity attributes like date of birth, first name, and user id number isn’t enough for identity resolution as personal identity attributes come with data quality problems like unintentional errors, intentional deception, and missing data.

#### Attribution of Digital Identity

So much literature has classified digital identities into different attributes. According to {Shalini Yadav, 2019} attributes can be classified into three classes of profile attributes such as first name, last name, gender, location, education, profession, email, language, and date of birth; network attributes like friendships, group membership, fan page and followers; and content attributes such as an individual’s tweets, video posts, image posts, YouTube links and so on. (King, 2018) claims that before the explosion of the internet, voluminous data and improved new technologies, digital data, and forensics tools existed to identify a person even with the related crime evidence but today, the ease of accessing an identity attribute may be considered an invasion of their privacy. They further classified cyber identity resolution analytics to include the codification of features into the broad categories of biographic, behavioral, relationship, biometric, and physiological data which are discoverable online. (Mary C. (Kay) Michel, 2018) observed that identity includes information provenance and records of how the evidence was processed and communicated which entails people-involved references and devices used in the incident. The author further reiterated that biographical, behavioral, and biometric feature sets have hidden ways of linking traits, aligning ground truth identity core, and detecting fake identities. {Jain, 2015} proposed an automated method for identity resolution on online social media using four attributes profile attributes, content attributes like the writing style, self-mention behavior, and network attributes.

Studies have demonstrated that social information, when incorporated into matching algorithms, can improve the performance of identity resolution much better than personal profiles. This is because social identity attributes seem to be more reliable than personal profiles in that they cannot be easily altered or falsified by an individual. According to {Kahina Gani, 2012} even if a user decides to create several identities for different purposes and tries to hide the relations between the different identities she/he creates, the user's latent habits manifest in different ways like the writing styles, interests, temporal social behaviors, which creates a link established in between which constructs the signature of the actual identity of the user.

A person is always associated with several different identities, having different credentials/features, or having the same identity provider or application (J Elliott, 2011). This is mainly because, under different circumstances, a person may wish to transact with another person as different representations for example, as an individual, as a pseudonymous individual, as an agent of a company, and so on. However, this creates multiple digital identities of the same individual causing identity ambiguity. In general, category of the attributes or features of digital identities data are categorized into demographic/biographic identity data, biographic, biometric, physiological, and behavioral data which are discussed below:

1. Demographic or biographic identity data

This data is a person’s self-perception as an individual. It only considers a person in isolation. (Mary C. (Kay) Michel, 2018) addressed personal identity attributes as temporal characteristics. Hence, due to this limitation, many recent studies are now exploiting adding social context information such as social behaviors and relationships for identity resolution. Which invariably improved the performance of identity resolution. (Edwards, 2018) discussed a method of identity resolution in online social media that addressed data quality measures. It can also be very complex due to the special data characteristics of missing data values, entry errors, data ambiguities, intentional identity fraud, and deception in identity records. Examples of this data are name, date of birth (DOB), addresses, gender, death date, education, career, and physical traits like hair color or race. However, with gender identity data, due to today’s emerging non-binary categories of categorizing gender examples like transgender; challenges can occur in its use for categorizing identity data. Demographic or biographic identity data is also called social history data as it builds over time (Phillips et al., 2020). Considering that in the real world, individuals interact with each other in society both virtually online and in person. The social context of an individual can be a very useful factor when detecting and/or resolving identities along with the user’s data relating to both physical appearance and legal documentation. Although {King, 2018} observed that because biographic identity core data is publicly available on government sites or social network profiles, it is easy to create fake profiles with this data which creates cyber identity challenges in the use of demographic or biographic identity data.

1. Biometric data

This entails automated biometrics recognition that uses anatomical, physiological, and/or behavioral characteristics to identify, authenticate or verify a person’s identity. Biometric identity trait data involves collecting human surface traits like fingerprint, face, or iris into a stored image. Other examples of biometric data could be behavioral like user typing pattern, gait, or Stylometry artifact and writing style {Yiming Zhang, 2019} leveraged both writing and photography styles to develop an intelligent identification system that automatically links multiple accounts of the same individuals. Also, {T. Alesea, 2021} proposed a user identity management system that provided a solution to the problem of Identity theft with the help of privacy-preserving multi-factor authentication to ensure strong security, he integrated a face recognition system into the normal username/password authentication. Recent studies in identity resolution have also stressed using biometric identity data as it is considered high-security identification and identity authentication method deployed in various applications. An example is a cloud computing-based face resolution framework for IoT applications (Chen et al., 2020).

1. Physiological data

Digital identity data can also be collected from physiological devices, for example, software tools and data that can analyze a person’s biological health function metrics such as pulse, heart rate, and blood pressure {King, 2018}. This data includes wearable technology that possesses advanced capabilities for sensing, storing, processing, and communicating various forms of physiological data. They resolve or reveal an identity with location heat maps based on GPS data. For an electronic device identity like IoT ecosystem and disparate technologies attributes example heart rate. {Héctor P. Martínez, et.al., 2013} used deep learning for affective modeling from multiple physiological signals such as skin, conductance, blood volume pulse, and behavioral and bodily responses to stimuli which are collected and used as the affective model input.

1. Behavioral data

According to {Mary C. (Kay) Michel, 2018} behavioral data is temporal and based on a person’s actions, interactions with others, and environment within a situation or scenario. {Wang, 2015} defines it as a social behavior attribute that represents the common characteristics of the social group(s) that an individual belongs to as they reflect the psychological view of personal identity on individuals. As individuals interact with their environment, they leave behind identifications that are traceable example IP-address, Mac addresses although they are likely to be masked. {King, 2018} explained that the unique digital configurations associated with an individual identity are their browser, the application, and device fingerprints which they have used as they are now discoverable through tools like web analytic client-side script requested parameter metrics that link identification to users. According to the report {Ana Beduschi et.al, 2017} behavioral attributes identifier data are important in building digital identities existing in identity systems as they provide new lines of data that can be used for verification.

#### Relationships of Digital Attributes

A digital identity possesses an identifier or attribute that sufficiently identifies this person within any set of persons. Examples are sequence numbers, email addresses, phone numbers, identity card numbers, names, and so on. These attributes or features are very diversified as it is ascribed to a body of information about an individual, organization, or electronic device that exists in cyberspace online. (J Elliott, 2011) asserted that identity providers associate validated attributes, features, or credentials to digital identity. Since these individuals, organizations, or electronic device digital identities all engage with the public, having security and privacy risks at different levels. Any network joined by these identities is subject to their vulnerabilities (Stoyanova, 2020). As such being able to resolve digital identities or identify the attributes/features of a cyber identity is so paramount to mitigating cybercrime and enhancing forensic investigations.

(Ruff, 2018) discussed the model of self-sovereign identity (SSI) as a breakthrough for revolutionizing digital identity relationships between individuals, organizations, and devices or things. He believes that the use of SSI will help eradicate multiple forms of almost the same identity profiles. He modeled three relationships of digital attributes that help in developing the digital identity relationships between individuals, organizations, and devices or things. These are:

1. Siloed or traditional digital identity relationship model approach

Siloed or traditional is the oldest digital identity relationship model where organizations issue digital credentials to their users that they use to assess the organization's services. With this trust is established between the user identity and the organization using some secret personal data attributes like birthdays, usernames, passwords, maiden names, pins, biometrics, or physical tokens. Because this same form of trust relationship repeats for every app, website, and organization; it results in different separate credentials for each relationship established. Although the traditional model reduces the risks of criminals and enhances cyber security as identities are registered and recognized since these credentials are first required before the use of services. With the increase in the diverse use of heterogeneous digital devices, the traditional or siloed approach to authenticating, detecting, and resolving identity is limited and not sustainable.

1. Third-party identity provider approach

The third-party identity provider approach adds a third-party organization that acts as an identity provider (IDP) between the user identity and the organization or service. The identity provider controls the digital credentials by providing single sign-on or single credentials like Google does, which is used everywhere else. This reduces the presence of different separate credentials needed for different websites. The IDP model is mainly used in social media logins on the web such as Facebook, Google, and Twitter serving as the identity provider. Here trust is established between a user identity and the identity provider using protocols like OAuth, SAML, or OpenID Connect.

1. The self-sovereign or peer-to-peer approach

The self-sovereign or peer-to-peer is now the emerging approach for establishing decentralized digital identity and verifiable credentials to prove user possession of attributes mostly amongst internet of things (IoT) devices. With a self-sovereign identity approach user, privacy and IoT security are improved and dependency on third-party organizations is reduced.

### Identity Attribution in the South African Network Environment

Today organizations and people around the world rely on networking environments to share communication, knowledge, opinions, and experience, through computing and digital devices resources especially. With the explosive use of the internet and big data, all networking environment resources communicate very easily within the speed of light through protocols, packets, hubs, bridges, switches, routers, etc. However, the same features that make the networking environment valuable also make them targets for a variety of digital fraud. This is due to the availability of duplicate and false identity records as a result of data ambiguity, and data errors unintentional and/or intentional deceptions. According to (King, 2018) due to the increased cybercrime and ease of creating fraudulent or stolen profiles, a need exists for resolving a digital or cyber identity. As such detecting duplicate identities and taking action to resolve identities of both malicious and legitimate users as quickly as possible is imperative to protect users in the South African networking environment. As this also maintains the trustworthiness of the network.

According to Nickson M. Karie (2016) in all minutes or seconds of our daily lives, with the use of computing digital devices, we leave our digital footprints behind which brings some digital traces that can be monitored and retrieved. Examples of these digital traces are deleted files, registry entries, log files, internet history cache, and automatic application software backup files. This then implies that every action we take as individual identities when using our computers or digital devices has a diverse number of implications that could be both advantageous and detrimental. Hence, whatever we do use our computers or digital systems leads us to the subject of investigating through digital forensics. In the process of digital forensic investigations, it is almost impossible without an ability to resolve individual identities and attributes of the digital footprints that are scattered all around in cyberspace. Massimo Leone (2021) emphasizes that the identification of individual digital identities is a primary necessity in our societies today to unmask impostures, disguising, and forgery identities in digital networks.

{Jain, 2015} explains that users create unlinked diverse identities on platforms for various purposes. However, these unlinked identities raise concerns for enterprises and security practitioners. Identifying and detecting duplicates of attackers and victims amid so many records of data in the network needs to be an automatic mechanism used by law enforcement and investigating services to that prevent cyber digital fraud and abuse. Identity reconciliation is a fundamental problem due to data ambiguity, and intentional and unintentional data errors, especially in the big data network environment. It is the process of uncovering user identity records that are co-referent to the same real-world individual. It involves combining user attributes about a person into one identity from non-digital and Internet-connected digital data. According to (Shalini Yadav, 2019) identity resolution is considered one of the pivotal techniques to reveal redundant identities, which are found co-referent to the same real-world user. The goal of identity resolution is to link these artifacts and attributes to create a suspect profile or a unique identity match if enough data points exist in a networking environment. (Kobayashi and Talburt, 2014) defines identity resolution (IR) as the process of resolving a set of input references against an existing set of managed Entity Identity Structures (EIS) stored in an Identity Knowledgebase (IKB). He explains that the main aim of identity resolution is to locate matches to the references in the identity knowledge base and retrieve their identity information.

## 4.3 IDENTITY AMBIGUITY

According to {Hakak, 2015} identity ambiguity is a situation where there is a lack of clarity or no sufficiently strong identity surrounding the meaning assigned to existing claims to an identity which leads to confusion as one’s identification could be abruptly lost. He believes that identification is critical and identity ambiguity is never a viable option, individuals in a state of identity ambiguity will always attempt to regain identification one way or the other. In digital forensics, computer-related crime evidence is extracted and then presented to a court of law for criminal or civil proceedings. However, for the processes to be accepted, certain criteria that satisfy comprehensiveness, authenticity, identity ownership or authorship, and objectivity of evidence have to be followed.

Identity ambiguity is defined as a situation where two or more profiles describe the same physical person. It is when one identity is having multiple forms of belonging as variation occurs in identity profiles (Charlotte Maene, 2021). Criminals and attackers mostly create multiple identities to hide their consistent involvement in cybercrimes and illegal activities. Due to these multi-dimensional identity configurations, creating more than one identity profile for each identity results in the ambiguity of identities in databases, electronic systems, or a network. As such trying to decode or interpret the identities relating to the same person can be very challenging. Identity ambiguity even gets more complicated in big data.

Some of the root challenges of identity ambiguity can be traced to the freedom of individuals to define who they are; however way they want without any standards, prescribed patterns, benchmarks, or restrictions whatsoever. As stated by Slater (1997) individuals are no longer dictated an identity position by governing institutions. Instead, they have become free in terms of defining who they are, lacking prescribed patterns or benchmarks. However, in a sea of possible dissimilarities, variations, and meanings of identities where some are for the same individuals; this causes field overlap and conflict creating multiple identities with a sense of identities being lost, ambiguous, disconnected, and adrift {Schouten, 2017}. This is the situation in cyberspace today.

According to Isaac Constans (2021), the real problem with resolving identity is not mainly person related. This is so as we see today where people are forced to access an application or data, using a form of identification either usernames, passwords, access cards, fingerprints, or other means to identify who they are. With the use of devices on the network that exchanges data; for example, servers, workstations, virtual machines, applications, and even cloud workloads and containers, organizations must ensure that machine identity management is properly managed. Especially with cyber security, collecting metadata on device identity types, configuration, location, credentials, keys, and certificates used must be of utmost importance. This is paramount as {Carole Tansley, 2013} noted that ambiguity about identity is uncomfortable, at all levels both individual and organizational, and must be resolved quickly to achieve some renewed appearance of clarity about the identities. This is the aim of this dissertation.

## APPROACHES FOR IDENTITY RESOLUTION

The problem of identity resolution has been approached from many different perspectives, including traditional, machine learning, matching algorithms, graph theory, and system design. The identity resolution approach typically uses one or more combinations of techniques including rule or score-based comparisons, distance measures, graph-based deterministic, and probabilistic. In aid of identifying suitable approaches, we survey some of the existing literature on identity resolution across various network environments and platforms.

{Fan Zhou, et.al 2018} broadly classified approaches of identity resolution into two main classifications which are Feature-based approaches and Network-based methods. The feature-based approach requires the extraction of a set of independent features from user profiles or behavioral activities to map across identities on different sites leveraging the user’s profile information. Features like username, writing style, posts, Spatio-temporal, and timestamped location data; to mention but a few for identification of the user. The network-based methods, on the other hand, use an end-to-end network alignment approach and requires structural information to align networks based on anchor nodes, embedding structures of the nodes from their local context. The authors describe the network-based approach as focusing on utilizing network features to link user identities. According to {Rachana Y. Patila, 2022} the existing network protocols are now insufficient for pinpointing the exact device identification of a user device and collecting the required digital evidence of cybercrime, which eventually makes the process of identity resolution for forensic investigation very difficult. The authors proposed a protocol that uses fingerprinting techniques deploying hash tree methods to ensure nonrepudiation and tracking users' identities to their true source by collecting target data from the device in the form of a device fingerprint with the help of an agent process. This protocol has no dependency on any third party or ISP hence nonrepudiation here means that no party can deny the authenticity of digital signatures on the network. As such true identifications of the source identity and origin of attacker or victim device of cyber-crime are made possible from the server-side mirroring the client-side of the network of devices.

Resolving or detecting a user identity within a network can either be text-based {Deepesh Kumar Srivastava, 2020}, image-based {Pengfei Hua et.al, 2018, Pengfei Hu et.al, 2017}, audio-based, and/or video-based. Conflict resolution amongst identities is done using unique identifiers, attributes, or features of the user identity which are available across the network environment of use. These attributes/features are very diverse depending on the environment of the user identity. According to Matthew Edwards et.al (2016) the profiles of a person or identity are matched together based on the similarity of their features. These features could be simple biographical attributes to inferred characteristics like the identity writing style. He further explains that identity resolution aims to allow different sets of information about a person to be connected. Text-based identity resolution uses name-based feature classifiers and techniques (Matthew Edwards, 2016). Image-based identity resolution uses a perceptual hashing technique to identify the key features of all profile images and then calculate the Hamming distance between the two hashes (Starkweather, 2016).

Identity resolution has been effectively used on identity management systems and databases to evaluate customers, and their interests, in the marketing brands and the advertisement sector to create a unified interface mainly using graph data structure framework to link user identities. So much literature has been done on user identity linkage across online networks that is, trying to match one identity attribute to all their online identities (Shalini Yadav, 2019), (Jain, 2015), (Neha Talokar, 2017), (Deepesh Kumar Srivastava, 2020). This has led to several methodologies being used including neural networks, natural language processing, text mining techniques, similarity analysis, and graph analysis; to mention but a few. Identity resolution also exists among the internet of things (IoT) and electronic devices (Pengfei Hu, 2017). Some other researchers provided a comprehensive survey on the potential identity resolution systems that may be used in the industrial internet of things (IIoT) to automate control amongst sensors, industrial equipment, products, and staff in the factory to enable the ability of a system or system component to gather information about its environment at any given time and adapt behaviors accordingly for intelligent production {Yuzheng Ren et.al., 2021}.

Most research literature categorized existing identity resolution methods and approaches into four groups - rule-based, statistical, machine learning, and deep learning methods, which are:

### The rule-based approach

The rule-based approach is simply labeling an identity pair as match or non-match using a set of rules. This method uses specific rule sets to match records of identities to distinguish similar records. This rule-based methodology is always combined with (Phillips et al., 2020) combined the rule-based scoring system and graph-based analysis for identity resolution of policing data. According to {Wang and Li, 2015} several rule-based identity resolution approaches are mainly based on matching rules encoded by domain experts. For example, IBM’s InfoSphere. Identity Insight provides an identity analytics solution with a set of rules predefined by human experts as well as sophisticated algorithms. An example of such a rule is – a situation where two identity records have identical dates of birth and last names, the system will resolve them into one only if the matching score of their first names is above a threshold. They used a supervised learning method to determine a threshold for match decisions. However, with the existence of data quality problems that are very common in big data such as missing values, entry errors, data ambiguity, and deceptions, exact-match heuristics tend to have high specificity but low sensitivity in detecting true matches. Also, they explained the challenge of the rule-based approach, that creating rule sets can be highly time-consuming and expensive. Another limitation of the set of rules could be not portable across different contexts and diverse domains. This makes the rule-based approach not applicable to resolving identities in a big data networking environment as it can be limited due to (Phillips et al., 2020) asserting that it has a low sensitivity in detecting true matches because of missing or incorrect data. Although this can be improved by creating an effective and robust rule set that can work in a variety of different contexts, this can be time-consuming which is unacceptable in big data due to its characteristics of velocity, variety, speed, and voluminosity.

### Statistical approach

This approach covers two main approaches for accomplishing identity resolution which are - the probabilistic and deterministic approaches which are used for matching identity pairs. The probabilistic statistical approach resolves identities based on predictions of the likelihood of matching identity pairs based on the agreement among attributes. Probabilistic algorithms work with a given confidence interval. Studies have shown that the probabilistic models achieve good performance for identity matching. (Charlotte Maene, 2021) used the Latent Profile Analysis (LPA) which is a statistical technique to identify identity patterns and hidden groups.

Deterministic on the other hand resolve identities based on information that is known to be certain and true. It is always based on first-party data. Most times statistical approach to identity resolution uses an identity graph that houses all the known identifiers that correlate with known user identities. The graph uses nodes to represent and link the attributes, identifiers, or edges to show a similarity between nodes. {Hassan Kazemian, 2022} proposed a graph-based method of identity resolution that combined four categories of attributes which are physical identity, an official identity which is a person's legal status, virtual identity seen on a computer screen, and social identity which is set of behavioral characteristics that is recognizable as a member of a group. This new method used community detection and centrality measurement as graph analysis techniques and defined a Louvain algorithm that combined graph-based relational similarity with corresponding attribute detection based on the selected attributes.

Statistics techniques use record linkage algorithm, deduplication, and pairwise comparison in the areas of statistics. {Wang, 2015} used matching algorithms with pair-wise comparison, transitive closure, and collective clustering using three different attributes - personal, social behavior, and social relationship for identity resolution. Also {Tanner Fry, 2020} proposed a method that finds all author IDs belonging to a single developer in this entire dataset using blocking, matching models, pairwise, and random forest models.Although studies show that statistical models achieve good performance for identity matching, however, the statistical parameters may not be accurately estimated in the absence of sufficient training data. This makes the statistical approach not applicable to resolving identities in a big data networking environment

### 4.4.3 The machine learning approach

Machine learning describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks. {Christian Janiesch, 2021}. It is a sub-field of artificial intelligence where machines and devices can perceive their environment, learn, and imitate becoming intelligent like the human brain in solving problems and performing functions {Shah, 2018}. Some machine learning algorithms which are widely used are Linear Classifier, Logistic Regression, Naïve Bayes (NB), Bayesian Network, Support Vector Machines (SVM), Decision Tree, Random Forest, AdaBoost, Bootstrapped Aggregation (Bagging), k-Nearest Neighbour (k-NN) and Artificial Neural Network (ANN). Machine learning can be applied to various application domains and sub-domains like prediction, computer vision, semantic analysis, natural language processing, and information retrieval. This is made possible as it uses a wide range of open-source frameworks available that enable machine learning engineers to create, implement and maintain machine learning systems, generate projects and create new impactful machine learning systems. Such as Apache, Mahout, Spark MLib, TensorFlow, Oryx 2, Accord.NET, and Amazon Machine Learning.

For identity resolution, the machine learning approach automatically extracts patterns from annotated training data with annotated matching pairs and builds resolution models for new identity records. Big organizations like Amazon web services (AWS) perform identity resolution with machine learning to review users' personally identifiable information (PII) attributes in each user profile and automatic profile matching by automatically merging similar profiles. Recently a lot of significant studies have been done on field images, biometrics systems, and machine learning for identity resolution. While surveying the literature, a lot of these studies are done using machine learning models like neural networks, vector machines, and random forests to recognize facial images. Some studies claim that biometric-based technology in identification applications for analysis is inherently more reliable and distinctive. Some other studies proposed a blockchain-based identity management system for a chain of industrial electrical equipment (Jianye Cui, 2020). Machine learning combines information technology, statistics, probability, artificial intelligence, psychology, neurobiology, and so many other disciplines. {Nasteski, 2017} observed that machine learning solves problems by building a model with algorithms that allow the computer to learn and mimic the human brain as a good representation of a selected dataset. The researcher also reiterated that learning is done to recognize patterns, learn from experience, and abstract new information from data or optimize the accuracy and efficiency of its processing and output. Incorporating machine learning algorithms to train data enhances decision-making and analysis. As such, so many studies today apply machine learning algorithms either from supervised learning, unsupervised learning, semi-supervised and hybrid or reinforcement learning approach which combines both supervised and unsupervised data which are discussed below:

#### Supervised learning

Supervised learning uses a training dataset that covers examples for the input as well as labeled answers or target values for the output. With this algorithm, the classes are predetermined and model a set of inputs on labeled or ground truth (GT) value datasets used to train models. Once efficient training is complete, models can identify such labels automatically in new, unseen samples in a generalizable fashion and used to predict the target variable y given new or unseen data points of the input features x. Supervised learning models are sub-divided into classification models which are called classifiers that map the input into predefined classes and regression models. Examples of algorithms for classification models are support vector machines, decision trees, and probabilistic summaries. {Cao Xiao, 2015} presented a scalable approach of a supervised machine learning pipeline for detecting and classifying clusters of online social network accounts as malicious or legitimate with key features like name, email address, company, or university. The methods used are logistic regression, support vector machine, and random forest. Similar accounts were placed in clusters and feature engineering was used to profile accounts for feature modeling. (Manasa Priya Koduri, 2021) proposed a cross-device graph solution of identity resolution using supervised machine learning algorithms using IP-related features and behavioral features to better improve linkages between devices. In a structured set of data, supervised learning is used with the python dedupe library using machine learning to effectively deduplicate records to create matches, not matches, and possible matches since data records and attributes are labeled. Then the training data are used to predict matching records with the unseen data.

#### Unsupervised learning:

According to {Christian Janiesch, et.al, 2021} unsupervised learning takes place when the learning system is supposed to detect patterns without having any pre-existing labels or specifications. As such only one variable x data are trained with the goal of finding structural information of interest, such as groups of elements that share common properties. This is known as clustering where similar groups are placed in segments. With this algorithm, the classifications label is automatically developed and modeled as a set of inputs on unlabelled datasets. As such, there is no previous knowledge from these raw records or datasets about the nature of the relationship between users or data. Autoencoders (AEs) are a well-known class of algorithms that can be applied to unsupervised representation learning for example and has been used in many applications, such as pattern identification and dimensionality reduction. {Karhunen J, 2015} proposed a multi-view learning approach to unsupervised person Reidentification. With unstructured data that we see in heterogeneous big data, attributes are unlabelled and data sources are heterogeneous. This unlabelled data uses unsupervised learning.

#### Semi-supervised learning:

This algorithm combines both a small amount of labeled data and a large number of unlabelled data examples during training to generate an appropriate function or classifier. According to {M. BalaAnand, et.al., 2019} semi-supervised learning has an advantage in its capacity to develop patterns with great generalizability from extremely limited labeled datasets then the unlabelled datasets follow for training purposes. The authors further proposed an enhanced graph-based semi-supervised learning algorithm to detect fake users from a large volume of Twitter data by examining the activity of users over a broadened period.

#### Reinforcement learning:

According to {Christian Janiesch, et.al, 2021} reinforcement learning system doesn’t provide input and output pairs, but rather describes the current state of the system, specifies a goal, provides a list of allowable actions and their environmental constraints for their outcomes, letting the ML model experience the process of achieving the goal by itself using the principle of trial and error to maximize a reward. Reinforcement learning algorithms learn from rules and policies of how to act, given an observation of the real world without specifying how a target task is to be implemented. The main purpose of reinforcement learning is to learn good action sequences through interaction with the environment which is typically referred to as a policy. {Deqiang Ouyang, 2018} employed deep reinforcement learning to discard misleading and confounding frames to find the most representative frames from video pairs for video-based person re-identification and {Fan Zhou, et.al, 2018} proposed a novel deep reinforcement learning-based algorithm to study linking user identity by leveraging the duality of mapping between any two networks.

### 4.4.4 Deep learning

Deep learning is a machine learning concept that is based on artificial neural networks {Christian Janiesch, 2021}. Deep learning is derived from machine learning. {Shah, 2018} refers to deep learning as deep neural networks because neural networks are inspired by how human brains works, having a large number of multiple hidden layers and parameters which are used for feature extraction, activation, and transformation. Neurons start by holding a bunch of numbers as pixels showing the dimension width, length, and depths of the input image having a total of 784 neurons. These 784 neurons make up the first layer of the network while the grayscale value numbers of the pixels range from one to zero where one ‘s are for white pixels and zeros are for black pixels. Each neuron in the individual layers of the network breaks down problems into various smaller components to solve. Deep neural networks operate as activation in one layer determines the activations of the next layer where each layer learns from the previous layer (unsupervised), that is, the output of one-layer acts as an input to the next layer considering that each layer as an individual one layered artificial neural network; continually for all the layers {Sreenivas Sremath Tirumala}. Deep learning can be supervised learning- labeled, unsupervised learning - unlabelled, and hybrid learning – having both labeled and unlabelled data just like machine learning.

According to {M. Mohammadi, 2018} deep learning is a state-of-the-art approach that can deliver many accurate inferences, which has also changed the way intelligent decisions are made by computers. Deep learning differs from machine learning in the terms of its proficiency and effectiveness when applied to voluminous data which is paramount today. Also with deep learning, feature extraction is automated. Deep learning has become very popular today, especially in big data environments for analyzing and solving problems due to the voluminous increase in data. Just like machine learning, learning tasks with deep learning can be supervised - using labeled training data, unsupervised - analyzing unlabelled datasets, deep networks for hybrid learning. Some examples of deep learning techniques are Multi‑layer Perceptron (MLP), Recurrent Neural Network (RNN) effective for analyzing natural language text, speech, and handwriting recognition, Convolutional Neural Network (CNN or ConvNet) effective for analyzing images and objects, Deep Autoencoders, Restricted Boltzmann Machines, Generative Adversarial Networks, Kohonen Map or Self‑Organizing Map (SOM), Deep Belief Network (DBN), Recursive Neural Networks, Deep Reinforcement Learning (DRL) and many more hybrid approaches. Deep learning uses multiple layers or stages that are used to represent the abstractions of data through which data is processed for building computational and data-driven models {Sarker, 2021, Daniel S. Berman, 2019}.

Deep learning is mostly applied for machine learning, artificial intelligence as well as data science and analytics. In our world today, deep learning has been successfully applied behind self-driving cars, automated web services like recommendation engines (for example Netflix), and smart voice assistants like Google virtual assistants. While deep learning has brought us so many exciting developments, new deep learning evolution combines neural networks with symbolic AI forming Neuro-Symbolic AI for logical reasoning (MIT-IBM Watson AI Lab). This mainly establishes not only the “what” but also the “why”, processing the cause-effect relationships. For example, deep learning neural networks are used to identify what kind of shape, feature, or color a particular object has. But applying symbolic reasoning to this deep learning can take it a step further to communicate even more exciting properties about the object such as the area of the object, its volume, and so on. Although it is still an evolving space of research, researchers believe that symbolic AI algorithms will help integrate common sense reasoning and domain knowledge into deep learning. A good example is while using deep learning algorithms to detect an image, a neuro-symbolic system would use a neural network’s pattern recognition capabilities to identify objects and a symbolic artificial intelligence (AI) logic is used to understand the image or objects better.

Deep learning has been applied in many application areas such as natural language processing, sentiment analysis, cybersecurity, business, automated marketing practices, virtual and/or voice assistants, recommendation engines, visual recognition, healthcare, robotics, and many more. Big organizations like Google, Microsoft, Apple, IBM, Facebook, Yahoo, etc. establish deep learning research groups for implementation in their products and services. A practical example is MIT-IBM Watson artificial intelligence (AI) lab's new research of combining deep learning neural networks to symbolic AI now forming Neuro-Symbolic AI for enabling not only identifying objects, shapes, features, or colors of particular objects but also communicating object properties. Hence giving the data even more meaning. Also, Google uses deep learning for building powerful voice and image search services, recognition algorithms, image retrieval as well as image indexing. The study of {Ee Hung Chang et.al, 2013} proposed an anti-phishing method that used Google Image Search to retrieve the real identity of a website. Also, Microsoft uses deep learning for speech recognition (MAVIS), searching for audio and video files through human voices and speeches. Deep learning has automation power and excellent learning capabilities from historical data {Sarker, 2021}. He further noted some key properties and dependencies that differentiate deep learning and is essentially needed for using deep learning techniques which are:

* Data Dependencies: Deep learning is characteristically dependent on a large amount of data to build data-driven models as they perform poorly with small data types.
* Hardware Dependencies: As deep learning trains models with large datasets, they require large computational operations to optimize operations efficiently; for example, high-performance machines with sometimes multiple graphical processing units (GPUs).
* Feature Engineering Process: With deep learning, the process of extracting features such as characteristics, properties, and attributes takes reduced time and effort.
* Model Training and Execution time: Training a deep learning algorithm takes a long time due to a large number of parameters and datasets that are used.
* Black-box Perception and Interpretability: Deep learning is considered quite difficult to explain how a deep learning result was obtained. This is termed “black-box” due to its poor reasoning and interpretability of results.

{Daniel S. Berman, 2019} discussed how deep learning (DL) methods and techniques are used in cyber security applications like building generalizable models for detecting {Yuan, 2016}, classifying malware {Yousefi-Azar, 2017}, detecting botnets {Anderson, 2016}, {Woodbridge, 2016} and attacks {Shibahara, 2017}. The authors noted that learning can be shallow (having just one hidden layer) and deep (which is attempting to learn on multiple levels). Feature extraction is performed by the first few layers of the deep network. He reiterated that layers with a DL model can be trained independently of each other allowing parameters to be optimized in small, manageable chunks, that require significantly fewer resources. Deep learning can also be used to identify and classify file types using DBNs by taking a signal processing approach {Cox, 2015} and network traffic identification.

Deep learning approaches have been interestingly successfully applied in so many domains ranging from natural language, image recognition or identification, writing styles recognition, and computer visioning to mention just a few. One main benefit of deep learning is its ability to analyze and learn massive amounts of unsupervised data where the raw data is largely unlabelled and uncategorized {Maryam M. Najafabadi, et.al., 2016}. Noting that today’s digital content is inherently multimedia, such as text, images, audio, and video, according to {Wei Zhang, et.al., 2019}, deep-learning techniques have greatly boosted the intelligence of multimedia analysis significantly, especially in classification, detection, regression, semantic segmentation, captioning, and content generation.

Deep learning algorithms have recently been greatly used for identity resolution and digital forensics and has been significantly successful. CNN is mainly applied to images, videos or object identification {Kimberley D. Orsten-Hooge, et.al., 2019; Sheng Hea, 2021} used deep learning algorithm for recognizing, identifying and verifying faces of drivers in masked videos and writer identification, it was very successful with over 98% performance; validating the efficiency of deep learning. Amongst so many researchers, recent works on image and video identification are almost focused on convolution neural networks-based approaches {Vasileios Balafas and Nikolaos Ploskas, 2021}. The authors noted that deep learning has led to a significant improvement in object detection mostly with algorithms based on CNNs showing impressive gains in face recognition accuracy on problems and image classification. They used convolutional neural networks to detect objects in images by taking an input image, assigning meaning (which are weights and biases) to different objects of the image, and distinguish between them. {Esra Ataer-Cansizoglu, 2019} proposed a deep learning method for face verification on very low-resolution face images for identity-preserving face super-resolution. The author specifically used deep neural network to improve face recognition accuracy for image visual quality. This is to ensure that a given low-resolution face image is the same person as in a high-resolution gallery image. The author noted that of recent, deep neural networks which is a deep learning now dominate the face recognition field because of their high accuracy. {Fan Zhou, et.al 2018} proposed a deep neural network-based algorithm for user identity linkage which aims to find users across different social platforms that refer to the same individual/entity. It samples the networks and learns to encode network nodes into vector representation to capture local and global network structures.

Deep learning approaches has also shown success in speech recognition and speaker identification {Sreenivas Sremath Tirumala and Seyed Reza Shahamiri, 2016 & 2017}. The research study was audio-based as it aimed to identify a speaker based on her voice prints by comparing the voice profile of the speaker against existing profiles of various speakers. This was done by extracting and identifying unique characteristics of speech features from a group or set of speakers and deep learning algorithms used for feature extraction analysis and comparison. Although deep learning for speech and audio is quite new, unsupervised deep learning deep autoencoders (DAE) has been proven effective in speech or audio recognition as authors exploit DAEs ability for speaker identification especially its capacity to identify speakers with minimum number of layers. It was significantly successful achieving superior identification accuracy of 98.8% over the traditional neural networks.

{Danielle L. Ferreira, 2020} applied deep convolutional autoencoders to automatically extract latent spatio-temporal mobility features that characterize individual user mobility from raw mobility datasets to identify user communities. This was used as input to cluster algorithms for identifying communities that group users according to similar spatial and temporal mobility patterns. The result of the framework yielded significant increase in contact time amongst users belonging to the same community, by up to 150% when compared to the baseline.

Recurrent neural network (RNN) on the other hand is mainly applied to natural language text example {Shriya TP Gupta, et.al 2019} used variants of recurrent neural networks deep learning technique (Long Short-Term Memory and Gated Recurrent Unit Network) on texts for identifying authors by defining a suitable characterization of texts to capture the distinctive style of an author. The authors claim that deep learning has very high performance across various text related task as it enables multilevel automatic feature extraction mainly in the use of word embedding. The performance measure had an accuracy of up to 96.7%. Also, the framework used was Google Collaboratory which is Google’s free cloud service for developing deep learning applications on its GPU facilities; implemented in Python, and the Tensorflow library was used for creating the neural networks.

Deep learning algorithms has also been proposed to be used in digital forensics to mitigate cybercrimes. {Nickson M. Karie, et.al 2019} recommended a deep learning cyber forensic generic framework, observing that deep learning uses neural networks to simulate human decision-making as it has the ability to unearth relevant potential digital evidence from big data, and reducing bias in forensic investigations by challenging evidence considered admissible in a court of law or any civil hearing and much more. From the research paper framework, resolving identity is carried out in the evidence analysis process to provide easiness for digital forensic experts as well as helps law investigators in making accurate decisions on the right attacker or victim unique identification.

In summary, so many researchers similar to this dissertation has used deep learning techniques to successfully resolve identification in different multimedia analytics such as text, audio, video or images and also in digital forensics {Shriya TP Gupta, et.al 2019, Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, 2016 & 2017, Kimberley D. Orsten-Hooge, et.al., 2019; Sheng Hea, 2021, Vasileios Balafas and Nikolaos Ploskas, 2021, Esra Ataer-Cansizoglu, 2019}. Deep learning is suited for analysing and extracting useful knowledge from big amounts of data that is collected from different sources {Lei Zhang, 2017}. This makes deep learning suitable and applicable for this dissertation as big data is collected from different sources using big data tools and techniques that are relevant and is analysed using deep learning approach. {Karhunen J, 2015} emphasised that deep learning has produced significant successes in a variety of classification and regression challenges when properly trained. Therefore, this dissertation draws from these explanations (among others) and understands deep learning to be most relevant and applicable use for voluminous data scientifically derived, mathematically proven methods and algorithms in the acquisition. As such deep learning is used in this dissertation since uniquely identifying individual task can be viewed as a multiclass classification problem of a high dimensional feature space.

An overview of identity resolution approaches:

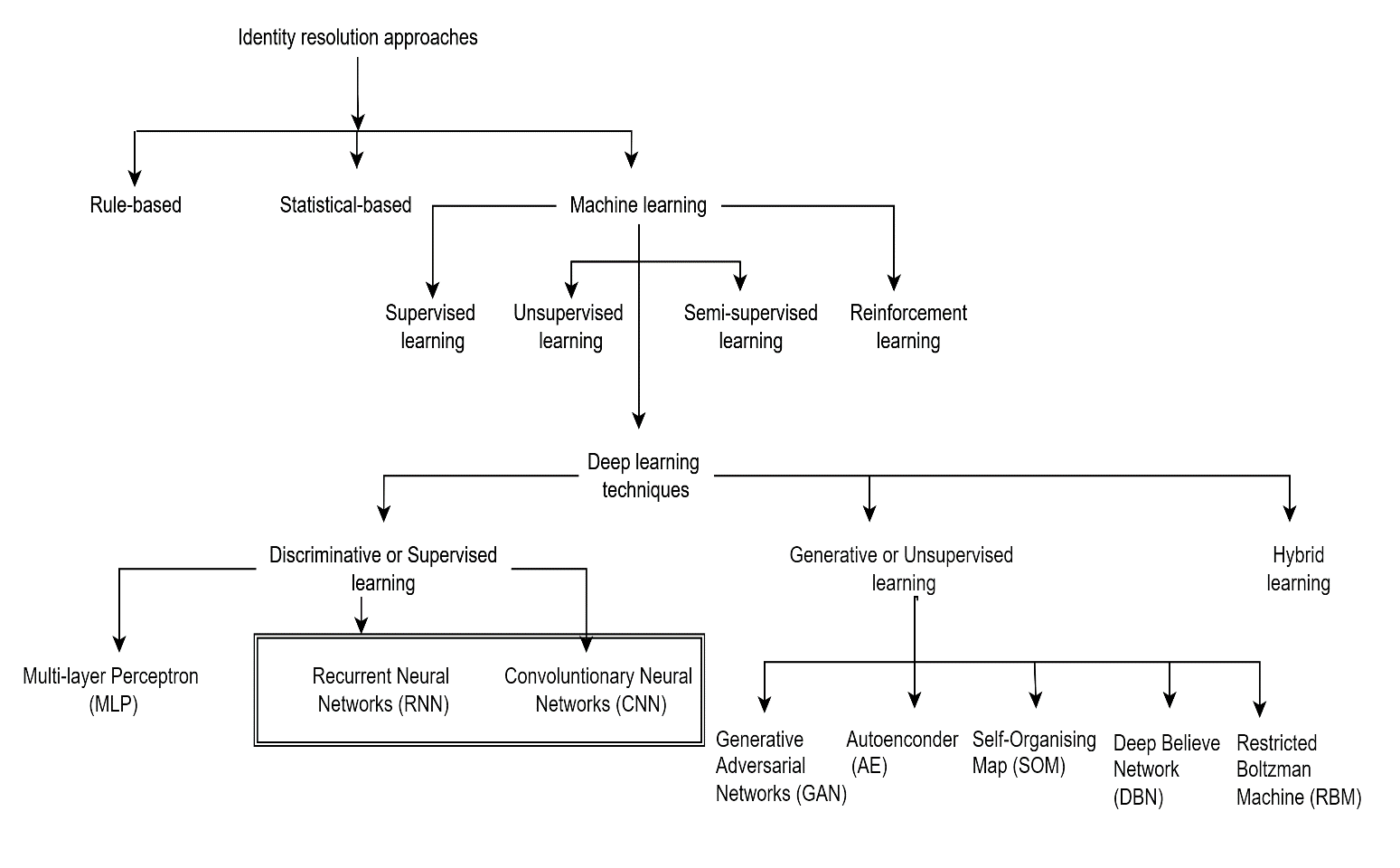
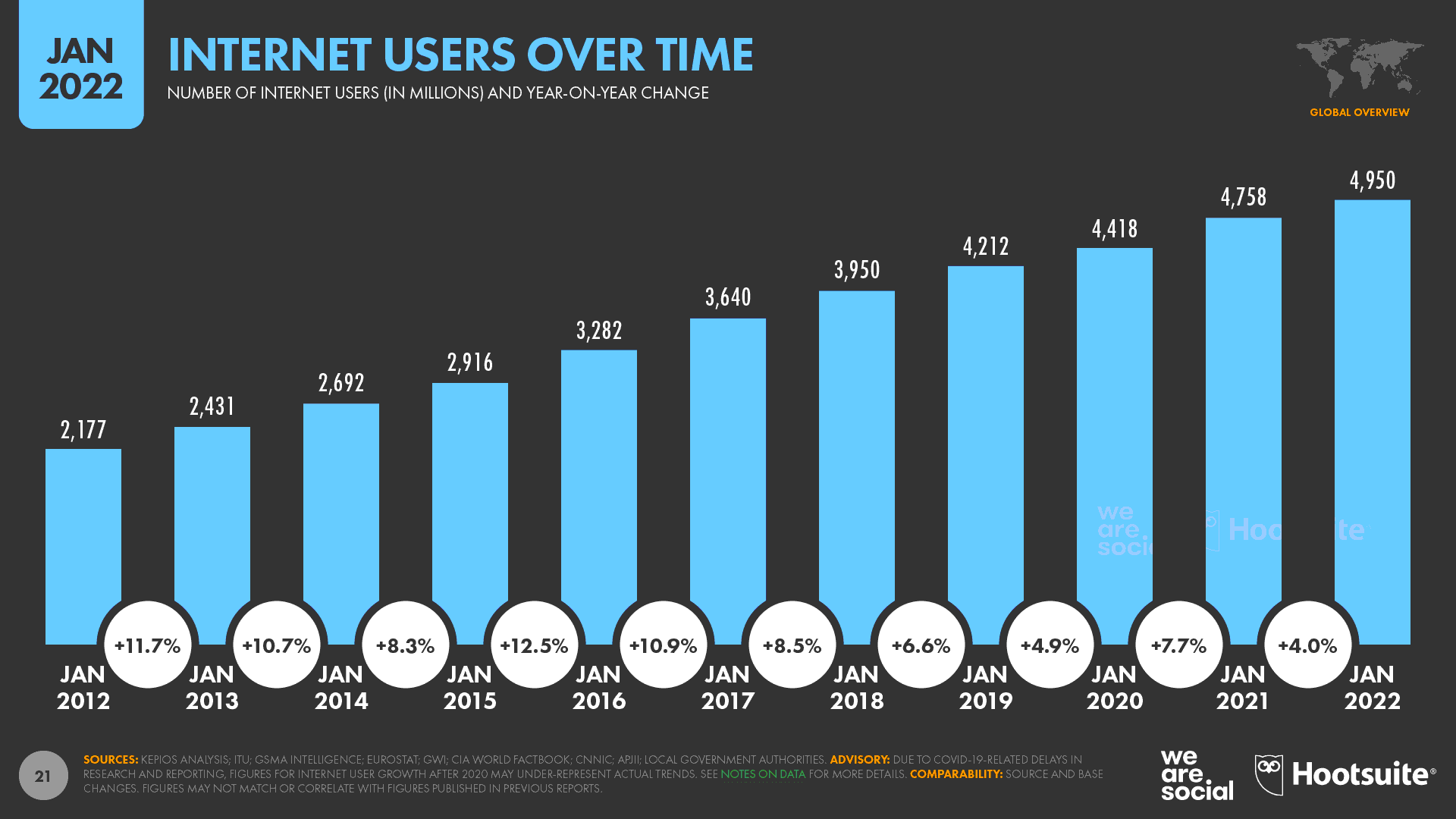


Fig 2-1 Taxonomy of Identity resolution approaches

## 4.5 IDENTITY RESOLUTION IN BIG DATA DIGITAL FORENSICS

The growing use of big data has also changed the crime scene investigation and is becoming an increasingly valuable source of digital forensic evidence. As a result, the dynamic nature and characteristics of big data, the importance of identity resolution in digital forensics for the law enforcement community is growing daily. The situation is even more aggravated by the fact that internet users continually rise each year and digital crime techniques are also becoming more predominant, sophisticated and better coordinated as the day goes by. (Mary C. (Kay) Michel, 2018) explains that cybercrime continues to grow as humans conduct more online digital activities that generate sensitive data while connected to anyone around the world. Although identity resolution has become prevalent among computer professionals and law enforcement agencies, it is still considered as a somewhat a new field of forensic science in big data digital forensics. The most successful digital strategy of big data companies still relies on knowing their potential users. Figure 2.2 shows the number of internet users continually rise from one year to another.

Fig 2.2 – Overview of continuous increase in internet users over the years



Source: Digital 2022: Global Overview Report – DataReportal – Global Digital Insights <https://datareportal.com/reports/digital-2022-global-overview-report>

Report also shows a continuous upsurge in cybercrime per year which creates even more importance for identity resolution in big data so as to easily identify identities in digital networks. (Mary C. (Kay) Michel, 2018) highlights that the legal system has challenges in keeping pace with today dynamic global networked technological environment and emerging crimes, identifying criminals in digital environment as it is characterized by anonymity, complex hardware and software, and exponential data growth. Fig 2.3 showing the percentage of organisations compromised by at least one successful cyberattack over the years.

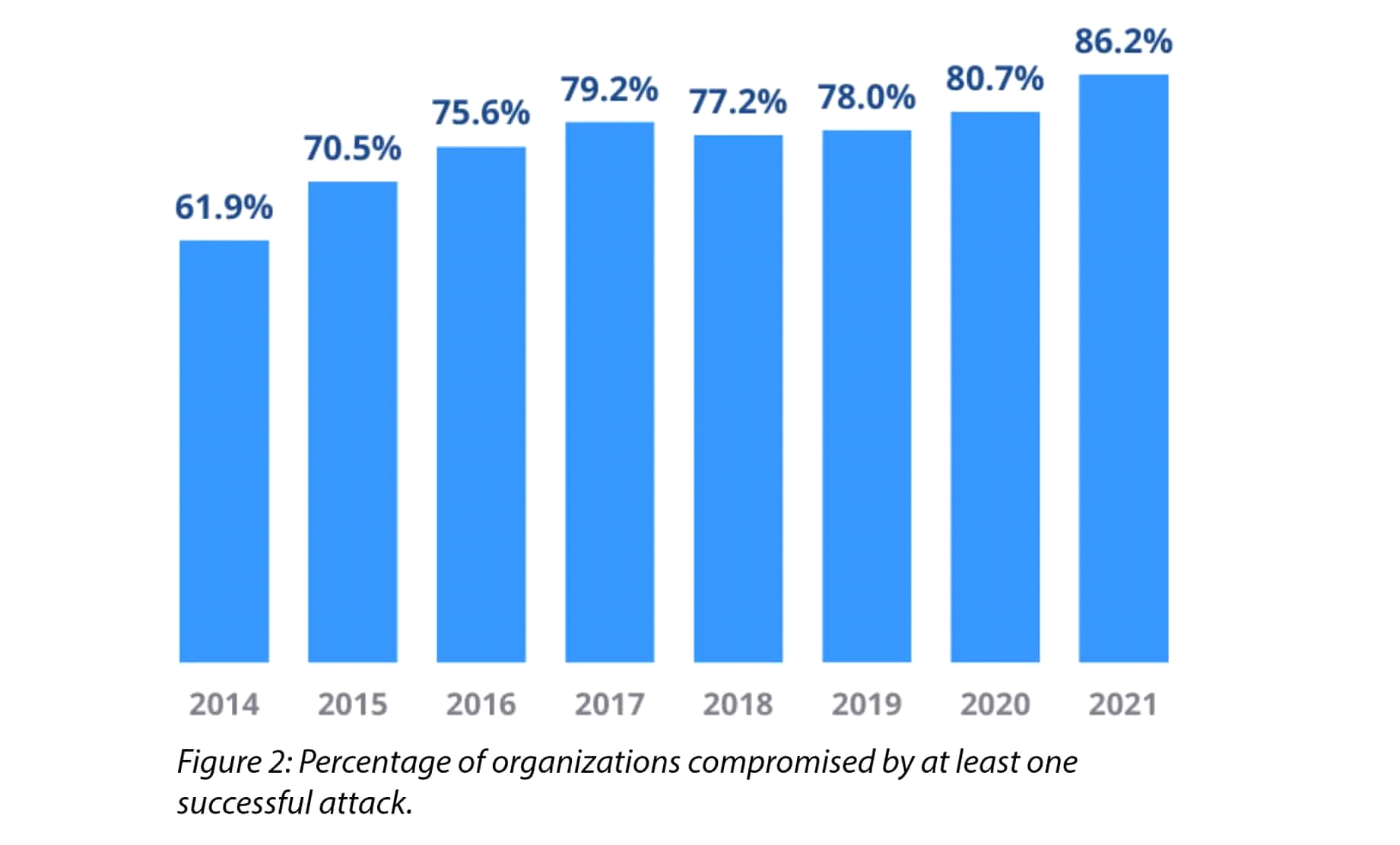


Fig 2.3 Percentage of cybercrime statistics for 2014-2021 in South Africa

Source: CyberEdge Group 2021 Cyberthreat Defense Report

<https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/>

As a way to restrain the growth of digital crimes emancipating from the rise of internet users, identity resolution is forming a significant part in big data digital forensics. Identity resolution tries to recognise an entity by connecting the growing volume of user identifiers and relationships to one individual identity as he or she interacts across different channels, big data sources and devices. Identity resolution attempts to get answers regarding digital crimes incidents through encouraging a unified view that differentiates specific user from another individual identity of digital users. Identity resolution brings together identifiers across online and offline touchpoints. This is the only way to truly understand individual internet users, their footprints, and attributes.

With the aggregation of multiple data sets brought together from different heterogenous sources, the data tends to have duplicates, inconsistencies and inaccuracies when consolidated together into a system. The diversity and voluminosity of big data today, the identity resolution approach of simply matching with rule-based and statistical methods is insufficient and ineffective during an aggregation phase. There is the automated need to link and consolidate identity information with a high level of confidence in the big data environment so as to easily detect cyber attacker and victims.

## 4.6 CHAPTER CONCLUSION

This chapter provided a brief overview of identity, digital identity, its attributes, and identity ambiguity. A description of several approaches of identity resolution followed. These were rule-based, statistical, machine learning and deep learning. Finally, identity resolution in big data digital forensics is discussed.

Having presented a brief overview of identity resolution, big data and digital forensics are necessary to buttress the concepts used in this dissertation, the three background chapters conclude with a descriptive overview of resolving identities in big data digital forensics South African network environment.

References

Brusseau, J. (2019). Ethics of identity in the time of big data.

Casey, E. (2012). Automation and artificial intelligence in digital forensics.

Casey, E. (2019). The chequered past and risky future of digital forensics.

Danda B Rawat, R. D. (2021). Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security.

Deepesh Kumar Srivastava, a. B. (2020). Words are important: A textual content-based identity resolution scheme across multiple online social networks.

Hans Henseler, J. H. (2013). Digital forensics-based pattern recognition for discovering identities in electronic evidence.

Howard Munkhondya, A. I. (2020). Digital Forensic Readiness Approach for Potential Evidence Preservation in Software-Defined Networks.

J.G. Enríquez, F. D.-M. (2017). Entity reconciliation in big data sources: A systematic mapping study.

Jain, P. (2015). Automated Methods for Identity Resolution across Heterogeneous Social Platforms.

Jianye Cui, H. W. (2021). Application of identity resolution and blockchain technology in the whole industrial chain management of electrical equipment.

Jonghyun Kim, G. L. (2021). Edge and identity preserving network for face super-resolution. . .

Leone, M. (2020). From Fingers to Faces: Visual Semiotics and Digital Forensics.

Leone, M. (2021). From Fingers to Faces: Visual Semiotics and Digital Forensics.

Li, J. S. (2020). A Model for Digital Forensic Investigation of Big Data. 2020 3rd International Conference on Artificial Intelligence and Big Data.

Lyer, C. (2021). What Is Identity Resolution? Definition, Process, Advantages with Examples.

Mallick, S. S. (2019). Big Data Computing Application in Digital Forensics Investigation and Cyber Security.

Manasa Priya Koduri, P. X. (2021). Cross-Device Identity Resolution using Machine Learning: A Scalable Device Graph Approach. .

Mayuresh Pandey, a. R. (2020). Identity Resolution in Social Network Using Recommender System.

Michael Phillips, M. H. (2020). A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data.

Michel, M. C. (2020). A Bio-inspired Classification System for Cyber-Physical-Human Identity Resolution.

Nihel Kooli, R. A. (2018). Deep learning-based approach for entity resolution in databases resolution in databases.

Renchao XIE, Z. W. (2021). A novel identity resolution system design based on Dual-Chord algorithm for industrial Internet of Things.

S Bartunov, A. K. (2012). Joint Link-Attribute User Identity Resolution in Online Social Networks. .

S.Venter, A. R. (2019). DIGITAL BEHAVIORAL-FINGERPRINT FOR USER ATTRIBUTION IN DIGITAL FORENSICS: ARE WE THERE YET?

BRUSSEAU, J. 2019. Ethics of identity in the time of big data.

CAO XIAO, D. M. F., THEODORE HWA 2015. Detecting Clusters of Fake Accounts in Online Social Networks.

PHILLIPS, M., AMIRHOSSEINI, M. H. & KAZEMIAN, H. B. A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data. 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 1-4 Dec. 2020 2020. 2077-2083.

RUFF, T. 2018. The three models of digital identity relationships. .

SHALINI YADAV, A. S., PAWAN KUMAR 2019. Multi‑attribute identity resolution for online social network. *SN Applied Sciences,* 1653.

Sun, H. W. (2021). Research on multi decision making security performance of IoT identity resolution server based on AHP.

T. Ruff. (2018) The three models of digital identity relationships. [Online]. Available: <https://medium.com/evernym/the-three-models-of-digital-identity-relationships-ca0727cb5186>

Wang, J. L. (2015). A model of identity resolution: evaluating identity attributes and matching algorithms.

Wang, Z., Zhou, C., Sun, J., Wang, S., Zhan, H., Yu, Y., . . . Cui, X. (2015). Key Technology Research on User Identity Resolution across multi-social media.

Weilong Ren, X. L. (2021). Online Topic-Aware Entity Resolution Over Incomplete Data Streams (Technical Report).

Yuzheng Ren, R. X. (2020). Potential Identity Resolution Systems for the Industrial Internet of Things: A Survey.

While there has been a good deal of research on approaches of identity resolution in diverse environments, including some using machine learning algorithms, the literature still suffers from no standard common frame of reference to measure. These various approaches and results are quite difficult to compare. Although identity resolution tools and techniques allow for comparability, it will be best to have many more measures of comparing approaches, models and frameworks, probably in comparing same datasets or same features used. (Matthew Edwards, 2016) observed that no standardisation can hinder identification of the best-performing methods and the direction of future research. They further reiterated that identity resolution is limited by the inadequate access to a standard of ground-truth data linking of identities of user profiles from different networks. Although some studies base their ground-truth mappings between user profiles on the exact matches in the attributes, features or name fields, attempting to verify such matches with a score generated from a small number of heuristics like the length and infrequency of the name attributes.

(J Elliott, 2011) discussed some major issue of managing multiple identities as legal implications of ownership and revocation of digital identities online. As with identities online, identity ownership can be implied to be owned by third-party identity organisations as well as owned by the individuals that has identity. This can be quite confusing in some instances or domains.

* Google identity resolution for cybercrime
* Cybercrime identity resolve or detect
* Resolving identity ambiguity
* How to identify an attacker or a victim in cyber bullying
* How does organisations identify or detect an attacker or a victim in cyber bullying
* How to automate identity in social media (big data environment)
* What are big data environment

Topics to publish paper:

* Data compression with losing important data for evidence
* **Note: data collection and annotation methods**

**Data annotation, e.g., categorization, tagging, or labeling of a large amount of raw data, is important for building discriminative deep learning models or supervised tasks, which is challenging. more effective**