# Guest Editorial Special Issue on: Big Data Analytics in Intelligent Systems

**William Hurst[1*], Chelsea Dobbins[1]**

[1]School of Computing and Mathematical Sciences, Liverpool John Moores University, Liverpool, UK
*Corresponding author: w.hurst@ljmu.ac.uk

*Accepted Month 7, 2015*

**Abstract** The amount of information that is being created, every day, is quickly growing. As such, it is now more common than ever to deal with extremely large datasets. As systems develop and become more intelligent and adaptive, analysing their behaviour is a challenge. The heterogeneity, volume and speed of data generation are increasing rapidly. This is further exacerbated by the use of wireless networks, sensors, smartphones and the Internet. Such systems are capable of generating a phenomenal amount of information and the need to analyse their behaviour, to detect security anomalies or predict future demands for example, is becoming harder. Furthermore, securing such systems is a challenge. As threats evolve, so should security measures develop and adopt increasingly intelligent security techniques. Adaptive systems must be employed and existing methods built upon to provide well-structured defence in depth. Despite the clear need to develop effective protection methods, the task is a difficult one, as there are significant weaknesses in the existing security currently in place. Consequently, this special issue of the Journal of Computer Sciences and Applications discusses big data analytics in intelligent systems. The specific topics of discussion include the Internet of Things, Web Services, Cloud Computing, Security and Interconnected Systems.

***Keywords:*** *Critical Infrastructure Data Analytics, Cyber Security, Data Analysis, Real Time Sensors, Machine Learning, Intrusion Detection, Control System, Cloud Computing, Data fusion, Web Services.*

## 1. Introduction

Within a growing digital world, technological developments, interconnectivity and the use of mechanisation has resulted in a surge in the amount of digital information being produced. Traditional data production services, such as critical infrastructures, which create extensive datasets for forecasting purposes, are growing due to increasing demands. Additionally, their service provision is now more reliant on automation, where analytic techniques play a key role. As a result, system administrators have to deal with significant datasets, constructed from millions of individual components to forecast future service demands and detect system faults. This is also forcing a migration to the cloud computing environment, where advanced data analytic services are supplied.

In addition to traditional information production services, data is now being generated from non-tradition services, including health-care, e-governments and our own personal networks of devices. Each produces distinct data analytic and digital supply chain challenges of their own. The datasets being produced are transforming society, improving the quality of the services provided and offering new solutions to age-old problems. For example, it is now common practice to capture, store and share almost every moment of our lives. Furthermore, the prevalent use of wearable fitness devices also allows us to capture personal biological information, such as heart rate and acceleration, which can be used for reflection. Consequently, with all of this data readily available, people have become interested in storing this type of data and reliving experiences through their collected digital media. As such, this large assortment of information is often referred to as human digital memories, which can be reflected on at a later time.

In general, big data sets tend to be unstructured but when analysed contains significantly useful information. The term big data was created to characterise the emerging trend in the amount of information being produced by traditional and non-tradition service providers. Subsequently, big data analysis techniques, refers to the sorting, processing and uncovering of hidden information in the vast datasets generated. The technological challenge involved in uncovering the concealed data trends, however is significant. The necessities for data storage alone, for example, require substantial storage devices. For that reason, many companies employ the use of big data platforms, such as Apache Hadoop, Microsoft Azure and IBM Big Data Platform. Each employs a cloud computing environment to process the datasets.

The remained of this paper is as follows. Initially, a discussion is put forwards on the Internet of Things and how the growth in smart cities has resulted in the generation of datasets from more sources than ever before. In section 3 the ubiquity of the digital environment is highlighted along with the use of control systems for automation and the inherent security issues being faced. In section 4, big data analytics are discussed, to highlight approaches which are currently being adopted to extract hidden information from significant datasets. The paper is concluded in section 5.

## 2. The Internet of Things

We now live in a mobile and information rich society, where the ability to generate and access a number of different data sources is feasible. The combination of the Internet and emerging technologies, such as near-field communications, real-time localization, and embedded sensors, lets us transform everyday objects into "smart objects" that can understand and react to their environment [1]. Any object, embedded with a sensor, is capable of providing us with information. Through unique addressing schemes, these devices are able to interact with each other and cooperate with their neighbours, to reach common goals [2]. This revolution is known as the Internet of Things (IoT) and can be defined as "a worldwide network of uniquely addressable interconnected objects, based on shared communication protocols" [3]. The proliferation of these devices, within our environment, is becoming more abundant. Currently, according to The NPD Group [4], there are 425 million devices connected to the Internet in U.S. homes; whilst computers were still the primary connected devices, numerous other devices are close behind, such as smartphones, games consoles, Blu-ray Disc players and Internet connected high-definition televisions (HDTVs). Furthermore, by 2016, Cisco predicts that there will be more than 10 billion mobile Internet-connected devices, which exceeds the worlds projected population, at that time, of 7.3 billion [5]. Additionally, global mobile data traffic will increase 11-fold to 15.9 Exabyte's per month, at an annual growth rate of 61%, from 2013 to 2018 [6]. These machines now fit seamlessly into our world and have become so common that we now hardly notice their existence; we now expect everything to be internet-enabled and "smart". This concept was first envisioned by Weiser in the 90s and is now firmly a reality [7]. In order to achieve this instant connectivity and exchange of information between devices, a number of key enabling technologies are required (see Figure 1) [8].

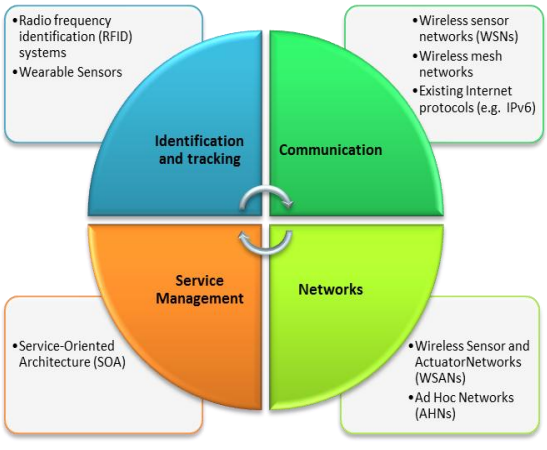


**Figure 1**. Internet of Things Enabling Technologies

As more and more devices become part of our global data space, unquestionably, the main strength of the IoT is the high impact it will have on several aspects of everyday-life and the behaviour of potential users, including mobile social networking, real world search, lifelogging, enterprise computing and groupware and urban mobility systems [2] (see Figure 2).

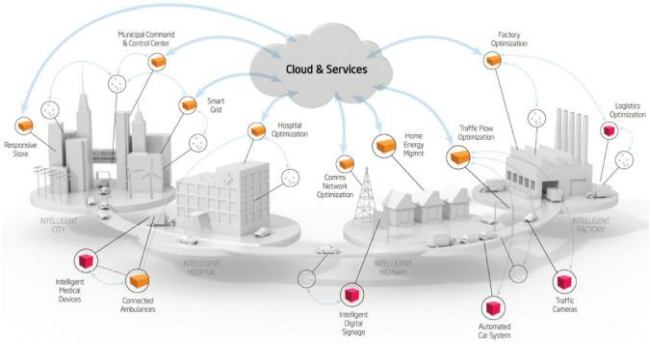


**Figure 2**. Applications of the IoT [9]

For instance, deploying and using smart things in enterprises can facilitate the communication and collaboration among co-located or non-co-located employees, whilst understanding human movement in urban environments has direct implications for the design of future urban public transport systems. However, this instant connectivity of devices has also enabled machine-to-machine connections (M2M) to increase. This has in turn led to the emergence of a new paradigm of the Internet of Everything (IoE), which aims to connect people, processes, data and things [6]. M2M connections, including security and automation, smart metering and utilities, maintenance, building automation, automotive, healthcare and consumer electronics, etc., are being used within a broad range of industries to monitor information [6].

### 2.1. Smart Cities

Around the world, urbanisation and economic growth are increasing; by 2020, the global market for smart urban systems will amount to approximately $400 billion annually and by 2050 it is estimated that 75% of the world's population will live in cities [10]. As part of these developments, the IoT/IoE stands to be the greatest benefit to these metropolitan areas [11]. As such, the term "smart cities" has been coined and often gets used in describing these developments, where the aim is to utilise a communication infrastructure to connect public resources, increase the quality of the services offered to its citizens, whilst reducing operational costs [12]. As depicted in Figure 3, this idea encompasses many aspects of society that "springs from the synergic interconnection of key industry and service sectors" [12].

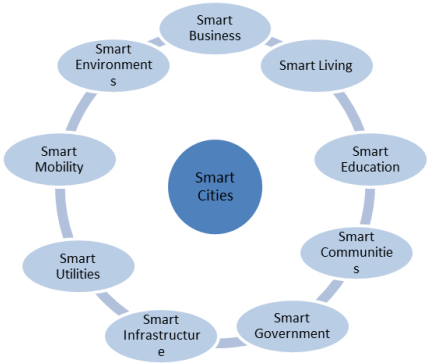


**Figure 3**. Smart City concept

Around the globe, there are many examples of such smart city initiatives. For instance, Masdar (United Arab Emirates) aims to establish a free-carbon zone by utilizing energy from renewable resources, reducing and recycling waste, creating sustainable transport links and developing sustainable and renewable alternatives to fossil fuels [13]. Meanwhile, Rio de Janeiro (Brazil) aims to tackle mudslides and flooding by connecting multiple systems together to improve crisis and transportation management. The system uses data generated from weather sensors, video surveillance, and field personnel, overlaid on a comprehensive geographic information system (GIS), to help the city in preparing and responding to flood-related incidents [14]. Additionally, Barcelona (Spain) utilises this concept to connect its transportation links and monitor the flow of people around the city. Its smart bus stops are connected to the city's fibre network to provide real-time timetables; smart parking spots are connected to the WiFi network to detect the presence of cars, whilst a city-wide network of sensors provides real-time information on the flow of people, noise and other forms of environmental pollution, as well as traffic and weather conditions [15]. As well as improving transport links and disaster relief, smart cities also intend to improve our quality of life by promoting an eco-friendly and sustainable environment and whereby healthcare services are connected to its citizens, regardless of their location [15]. Within the UK, Bristol City Council is developing a "smart city service within the SPHERE project (Sensor Platform for HEalthcare in a Residential Environment, 2013-2018) to monitor the health and well-being of people living at home" [15]. This can be achieved by using smaller and more powerful wearable sensing devices, which can collect a wealth of data about an individual. Additionally, the World Health Organisation's (WHO) Healthy Cities is a global initiative that was introduced in 1988 to "put health high on the social, economic and political agenda of city governments" [16]. Currently, approximately 90 cities are members of the WHO European Healthy Cities Network, and 30 national Healthy Cities networks across the WHO European Region have more than 1400 cities and towns as members [16].

A key contributor to these developments is the Internet of Things, which have enabled us to live a connected society of people and devices. Smart cities are one example that are revolutionising the way in which people interact with their environment and are vital in aiding resource management. Our reliance on fossil fuels is limited and new sustainable initiatives are required to offset our reliance on natural resources.

### 2.2. Smart Devices for Healthcare

Another sector that has benefited from the IoT paradigm is healthcare, where it plays an important role in many applications, including clinical care, remote monitoring and early intervention/prevention [17]. Features of the IoT, including global connectivity, ubiquitous identification, sensing, and communication capacities, have greatly benefited this field. For instance, all objects in healthcare systems (e.g. people, equipment, medicine, etc.) can be tracked and monitored constantly. Furthermore, all healthcare-related information (e.g. logistics, diagnosis, therapy, recovery, medication, management, finance, and even daily activity) can be collected, managed, and shared efficiently [8]. For example, patients can have sensors attached to their body to monitor their physiological signals (heart rate, breathing, etc.), whilst static sensors attached to the bed can monitor the pressure that is being exerted on particular pressure point hotspots. Using mobile internet access (e.g. WiFi, 3G, 4G, etc.), this data can then be wirelessly transmitted to the doctor's devices (e.g. laptop, mobile phone, tablet, etc.) so that they can have a real-time view of their patient [8], [18]. Furthermore, smart appliances such as fridges, can be used to monitor nutrition, provide dietary control and analyse eating habits [19].

Additionally, wearable health-monitoring systems (WHMS), composed of several on-body and intra-body biosensors, are capable of providing a real-time and unobtrusive outlet to monitor a patients' physiological parameters [20]. Data, including electrocardiogram (ECG), blood pressure, respiration, body and/or skin temperature, etc. can be obtained and transmitted using short-range communication technologies, such as ultra-wideband radio technology [21], Bluetooth [22] and ZigBee [23], to remotely monitor an individual once they have left hospital. Taking this concept further, stretchable electronic tattoos are flexible patches that can provide continuous tracking by wirelessly transmitting information such as heart rate, brain activity, body temperature, and hydration levels to a mobile device [24].

As it can be seen, the IoT is changing the way we interact with the world and our devices. As more and more "things" join this interconnected network, the information that we have access to is increasing. This flow of data is essential for allowing us to make informed decisions about ourselves and to aid in the development of a sustainable environment. However, although there are many initiatives that are utilising this idea, there are still many technological challenges to overcome. Such issues include, scalability (data transfer, processing, and management) of the network, communication between heterogeneous networks, integration with existing systems, big data analytics of information within the network and security and privacy of data [8].

## 3. A Digital World

One such technical challenge is the emergence of cyber-attacks, which has changes the security focus of information systems. Protecting infrastructures, wireless devices and the smart grid from cyber-threats, in an increasingly digital age, is a matter of growing urgency for governments and private industries across the globe. The consequences of a successful attack would be disastrous, ranging from potential loss of life and the compromise of military defence to economic damage or a devastating effect on the operation of government services. The need for improved security is evident.

### 3.1. Critical Information Infrastructures

Critical infrastructures include sectors such as energy resources, finance, food and water distribution, health, manufacturing and e-government services [25]. Their service provision is often dispersed over large geographic areas [26]. In recent years, however, critical infrastructures have become increasingly dependent on ICT to facilitate communication. Consequently, they have

become more vulnerable and face a new threat from the digital domain.

The surge in the use of ICT in critical infrastructures for automation has led to the fact that researching critical infrastructure protection and simulation environments naturally results towards a focus on digital industrial control systems.

The current threat levels facing critical infrastructures are higher than ever before. The volume of sophisticated cyber-attacks is starting to put a strain on defences currently in place. This increasing level of cyber-attacks demonstrates how important it is to ensure that the protection measures being used are continually evolving to keep up to date with new and emerging threat levels.

The topic of cyber-defence has become a key issue for debate in many governments and in growing frequency by CEOs of global corporations [27], [28] as well as being well documented and at the forefront of many news articles. The USA defence secretary Leon Panetta, for example, has highlighted the effects that cyber-attacks could have on a nation, comparing the potential impact of a successful attack to that of the terrorist attacks of 9/11. The United Kingdom, in particular, has also been very vocal on the large volume of cyber-attacks that occur daily, which are aimed at government services and global corporations. Secretary of State for Foreign Affairs, William Hague, has highlighted the volume and variety of cyber-threats being encountered. Whilst many of the attempted attacks remain small, for example, malicious emails [29], [30] containing Trojan horses [31], the sheer volume of the attacks occur-ring regularly poses a cause for concern.

### 3.1.1 Digital Control Systems

The growing cyber threat is a particularly worrying factor for industrial networks. Disasters have the potential to escalate with interconnectivity bringing the risk of cascading failure [32]. The cost of physical consequences reflects the ever growing need for effective critical infrastructure protection for the future safeguarding of the services which are heavily relied upon by the population.

An industrial network can be broken down into multiple layers consisting of a business layer, a supervisory layer and a control system layer. Each layer has an important role in the running of a critical infrastructure [33]. The business layer, as its name suggests, is associated with the entire commercial and trade aspects surrounding the infrastructure. The supervisory layer, however, oversees system operation. It is in this layer that the Supervisory Control and Data Acquisition system (SCADA) would have a key role. As Knapp *et al.*, identify, often the mistake is made to refer to all control systems as SCADA [33]. SCADA, however, is just one component of what makes up a critical infrastructure control network. The control system layer consists of process and control networks where automation is conducted through the use of Programmable Logic Controllers (PLC) [33].

Due to the variety of services critical infrastructures have to provide, the different types of technology currently used in each layer differs depending on the type of services provided by the infrastructure. This makes generic security systems extremely difficult to create. Despite that fact, there is a common dependence on certain types of technologies such as nodes, sensors and control systems, as well as the use of off-the-shelf components.

Given these layers inside critical infrastructures, it is clear that security can become an issue when the access points into the system are more numerous than ever before. At this point, resilience becomes an important factor, given the fact cyber-attacks are increasing at an alarming rate. The need to remain one step ahead of the attacker is becoming more and more important. Dependence on digital industrial networks means that the consequences of failure can produce unexpected results and must be planned for.

### 3.1.2 Security Measures

Currently, critical infrastructures are protected through the use of Unified Threat Management systems (UTM) and Intrusion Detection Systems (IDS). UTM's are considered to be the most important network security device in critical infrastructures and are a combination of software, hardware and network technologies [34]. The network technologies used by UTM systems generally include firewalls, pattern recognition, IDS and embedded analysis middleware. Critical infrastructures generally use different types of UTMs in the same infrastructure for defence in depth purposes. In addition the fact that UTMs operate behind a firewall means that the number of false positive alarms is reduced as the firewall acts as a filter for malicious activity.

UTMs tend to be divided into two groups including loosely-coupled and tightly coupled:

- Loosely-coupled UTMs integrate security products from various manufacturers meaning that interoperability between the components is an issue.
- Tightly-coupled UTMs are when the UTM has been developed by a single manufacturer meaning all the security functions have been developed by a single vendor meaning there are no interoperability issues [35].

A UTM is effectively a security gateway which filters and monitors network flows between Internet and Enterprise Network [36]. UTMs integrate multiple security technologies such as control interfaces, message formats, communication protocols and security policies and for that reason the management of security technologies in UTMs is a big challenge [35]. In UTMs high-efficiency and reliability are important factors as 80% of IT budgets are spent on maintaining the status-quo as well and averting downtown that is the result of human error [35].

Effective UTMs and IDS provide a sense of security for computers and network data by identifying, in real time, misuse or unauthorised use whilst allowing the system to continue functioning. IDSs typically use techniques such as Protocol Analysis, Signature-based detection, Anomaly Detection and intrusion detection sensors to protect the infrastructure [37]. Protocol Analysis is used to analyse data from the network and compare it with the model for expected behaviour to detect anomalies. Signature-based detection relies on patterns to identify data identified as being an intrusion by comparing the attack with known signatures. Signature-based detection however is non-adaptive and cannot detect attacks, which do not have a signature. Anomaly Detection, for example, a sudden

increase in data flow, is used to detect anomalies. Whilst, Intrusion Detection is a technique used to identify Trojans or viruses that are sent in emails or documents.

Defence in Depth is an important aspect to a critical infrastructure. Multiple IDS are often used alongside each other to combat multiple threats and different technology is used on each layer of the infrastructure to ensure that if an attacked penetrates one layer they are not automatically able to access the next one.

One additional defence technique currently implemented by critical infrastructures is the Honeypot. The Honeypot is used to fool attackers into thinking they have penetrated the system when in fact they have been diverted to a fake environment in which the attack is kept as long as possible and studied. This defence approach also aids with developing a picture of how an attack takes place.

For an effective defence in depth approach, critical infrastructures are divided into various levels of security. The Low Level security, for example, would be accessible by low level employees who only require basic access to the infrastructure to perform their functions and have access to only a small amount of data. Whereas the High Level would only be accessible by management and system administrators and tends to contain more sensitive data then the low levels. It is well documented that a large number of attacks are caused by unhappy staff members, or old employees who wish to carry out some sort of revenge of the system for mistreatment general unhappiness. For that reason the multilevel security is ideal as the disgruntled employee using their username and password to access the system will only be able to breach the Low Level.

Figure 4 displays an architectural view of these various layers and details a simplified view of the defence in depth technique of a critical infrastructure. Despite this use of defence in depth for securing critical infrastructures there are still numerous examples of attacks occurring daily and whilst the majority are unsuccessful the threat continues to grow.

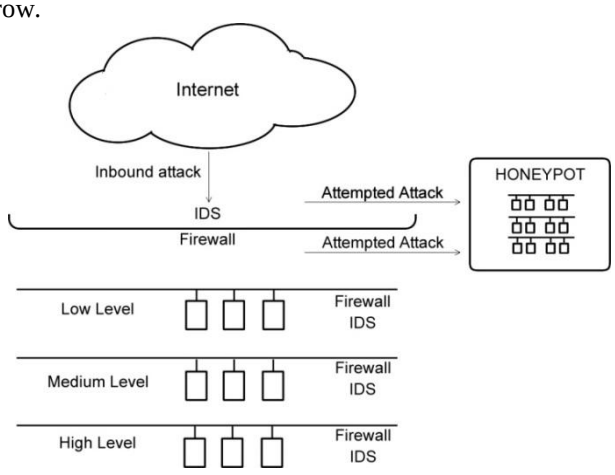


**Figure 4**. Defence in Depth

This mounting threat is exacerbated further, as new technologies integrate into existing infrastructure networks in order to fit the increasing demands of customers. Issues, such as, safe interaction and the threat of cascading failure further enforce the need to build on defence in depth measures. This is particularly the case for the implementation of the smart grid, as discussed in the following section.

## 3.2. Smart Grid

The smart grid is a clear example of the movement towards introducing new innovations to enhance service provision; where security and reliability is becoming a growing concern. Its implementation is a drive to update the provision of electricity services, with electricity industries being progressively transformed [38]. Functionality is intertwined with critical infrastructures, such as nuclear power plants; as the smart grid offers essential enhancements to the traditional power grid infrastructures.

The way in which electricity is generated and distributed is revolutionised by its introduction [39]. For example, the dynamic pricing for customers, distribution management advancements and demand management are all features which are brought about by this innovation. This is such a key advancement, as electricity is the central utility for the functioning of other critical infrastructure networks and society as a whole. Consequently, in this section, the focus is on smart grid components and enhancements and the implications this has on big data analytics.

Specifically the application of the Advanced Metering Infrastructure (AMI) is one of the most significant aspects of the smart grid infrastructure. It is used to perform the collection, storing, analysis, power consumption data collection and management process [40], [41]. The AMI facilitates two way communications between the customer and the utility company, with the aim of allowing both parties to view real-time data. The end user systems include home displays which allow the customer to have a greater insight into their energy usage and precise costings. However, with the introduction of technologies such as the AMI, big data analytics are required to assess the vast volumes of data produced. The potential insights into the benefits of the data produced can range from new innovative health care opportunities, such as dementia and depression monitoring through behavioural analysis [42]; to general user profiling for security purposes [43].

## 3.3. Data Breach and Cyber-Threat Sources

To conclude this section, we identify some of the most common sources of attacks and threats to critical infrastructure networks and other intelligent systems. Attacks occurring from inside an infrastructure are a problem, which infrastructures are becoming increasingly aware of and preparing for. An attack, which originates from the inside of an infrastructure, has the advantage that security measures can be bypassed and damage can be done before security has a chance to respond [44]. Often such attacks would occur as a result of an employee being disgruntled or upset with the organisation.

There is always the threat to insider attack where a disgruntled employee may choose to inflict damage on the system, or steal valuable information for personal reasons [45]. Additionally, the vulnerability caused using wireless networks means that a carefully placed laptop or smart phone may provide a way into a system that is considered secure[46]. The damage a thoughtfully place a USB stick or CD is evident from the well-known Wikileaks insider attack where an employee stole millions of confidential diplomatic communications sent from United States embassies and American foreign policy strategies, which were then published on the Internet. These diplomatic

communications, referred to as cables, were stolen through the use of a CD or USB stick which was used to store the insider information without detection.

The increased variety and sophistication of cyber-attacks is, in part, due to the variety of attack sources. Ranging from insider-threats to Hacktivists the sources of attacks can vary depending on the situation [44]. Nicholson A, *et al.*, highlight some of the various culprits of cyber-attacks, each of are discussed below:

States or Governments: With the ability to disable or severely cripple a country's ability on the other side of the globe through use of a remote computer, it is clear to see why governments are currently investing heavily in cyber-warfare technologies. As news agencies frequently highlight, state created viruses can potentially be the major threat to SCADA systems due to the level of sophistication and financial investment, which has gone into its development [44].

Organised Crime: Any attacks, which originate as part of organised crime are usually motivated by money. As Nicholson *et al.*, discuss, attackers often have access to substantial amounts of money and target banks or large companies, which can be held at ransom.

Hobbyists: An unusual threat critical infrastructures face comes from individuals who see a cyber-attack on a system as a challenge or thrill or something, which is simply curiosity motivated [44].

Script Kiddies: Similarly to hobbyists, script kiddies tend to be individuals who have limited access to sophisticated technologies and perform their attacks through use of limited scripts.

Hacktivists: Attacks which originate from Hacktivists tend to be individuals or groups who have political reasons for the their attacks [44]. For example, if a group wish to protest over the implementation of a new law, or make a political statement, then often cyber-attacks are conducted as a means to gain attention. One group known as Anonymous, have successfully conducted several high profile attacks, such as, targeting the UK police web forum in order to make a political statement.

The problems current control systems experience, enforce the need for an improvement in the technology used. However, critical infrastructures do not only have to cope with multiple control system vulnerabilities, they are also facing various type of SCADA system perpetrators, as Nicholson et al. discuss [44].

The result of the variety of threats control systems face in the immediate future signifies that there is a clear need for a high level multifaceted security system to safeguard critical infrastructures.

# 4. Big Data Analytics

Due to the increasing demands placed on critical infrastructures, technological enhancements (such as the smart grid) and the growing Internet of Things; the amount of data that is being produced, every day, is rising. As such, it is now more common than ever to deal with extremely large datasets.

As systems develop and become more intelligent and adaptive, analysing their behaviour and adapting to growing security demands is a challenge. The speed of data generation is increasing rapidly. Intelligent systems are capable of generating a phenomenal amount of information and the need to analyse their behaviour, to detect security anomalies or predict future demands for example, is becoming harder. In this section, the focus is on big data analysis techniques and processes involved in the evaluation of big data sets.

## 4.1. Data Pre-Processing

Raw data has out of range values, and the collection process is often unrestrained, meaning datasets contain missing values or coefficients which produce ambiguous results. For that reason, initially, data requires pre-processing. This acts as a filter to remove unwanted values and clean the data prior to feature extraction. Pre-filtering the data extracts any elements that are not required by the feature extraction stage. Redundant values, which do not conform to the filter parameters and irrelevant aspects of the data, are removed.

This process includes various stages, such as: cleaning and normalisation of raw data. Cleaning involves verifying that there are no missing values and smoothing data. Noisy data, which refers to corrupt and meaningless values, are also removed. The cleaning process also removes duplicated values; otherwise, the results of the data classification would be compromised. This could also include specifying a value range to cut out coefficients, which are outside the scope of our requirement. This process is displayed in Figure 5.

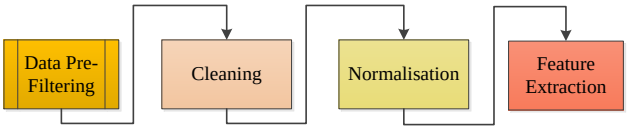


**Figure 5**. Data Pre-Filtering

Normalisation is used to allow the classifiers to treat the data equally. In other words, the data is manipulated so that coefficients in the dataset are standardised. This prevents raw data values from over contributing to the classification process and affecting the results.

Once processed, features can be extracted from the dataset for the machine learning and data analysis procedure. Features are aspects of the data, which allow for a representation of overall system behaviour. In the training mode, extracted features form feature vectors for both normal and abnormal behaviour. A feature vector would contain information about system and individual component behaviour.

## 4.2. Data Processing Techniques

Data processing is an essential part of big data analytics. In this sub-section the focus is on two distinct approaches, which are used to uncover information from large datasets.

### 4.2.1 Unsupervised Machine Learning

Unsupervised learning methods, such as clustering, aim to discover the natural grouping(s) of a set of patterns, points, or objects and is based on a proximity relationship; data that are similar tend to share an external relationship, which can be established to assemble the data into clusters [47], [48]. Due to its ease of implementation, simplicity, efficiency, and empirical success, the most popular algorithm for clustering is K-means [47]. It is a simple iterative method that is used to partition n observations

into a user-specified number of clusters, k [49],[50]. The data objects are grouped together into "compact" clusters with the assumption that all objects, within one group, are either mutually similar to each other or they are similar with respect to a common representative or centroid [51]. A requirement of the algorithm is that the user needs to specify the number of clusters (K) that they require. This is the most critical user-specified parameter, with no perfect mathematical criteria [47]. Therefore, defining K can be challenging and may be seen as a drawback [50], as the best number of clusters can be difficult to distinguish. However, a silhouette plot can overcome this. This graphical display illustrates which objects lie well within their cluster, and which ones are in the incorrect clusters [52]. It is very useful for selecting the 'appropriate' number of clusters, as it gives an idea of how well separated the clusters are [52], [53]. In conjunction with the algorithm, the appropriate number of clusters can easily be defined. There have been many implementations of the algorithm, across multiple domains. Other algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Ordering Points to Identify the Clustering Structure (OPTICS), have proven to be useful for analysing spatial data [54], [55]. In one such approach, using DBSCAN together with OPTICS, location points have been clustered into points of interest, with associated photographs being linked the locations [50]. The advantages of using DBSCAN, over K-means, is that DBSCAN is less sensitive to noise and allows clusters of arbitrary shape, whilst providing deterministic results [50]. However, a drawback of density-based clustering algorithms, such as DBSCAN and OPTICS, is that when clusters of different densities exist only particular kinds of noise points are captured. Furthermore, they don't perform well when clusters are close and border each other [54].

#### 4.2.2 Supervised Machine Learning

Supervised machine learning refers to the training of a data set, based on two sets of data. The process involves providing the classifiers with the 'right answer', so that they are trained to identify differing data sets autonomously. Ideally, the supervised learning stage requires the evenly sized data sets of both the normal behaviour and attack data

The technique can include some of the following data classifiers: Uncorrelated Normal Density based Classifier (UDC), Quadratic Discriminant Classifier (QDC), Linear Discriminant Classifier (LDC), Polynomial Classifier (PLOYC), k-Nearest Neighbour (KNNC), Decision Tree (TREEC), Parzen Classifier (PARZENC), Support Vector Classifier (SVC) and Naïve Bayes Classifier (NAIVEBC). A brief description of each of these techniques is provided in the following subsection. Each of these classifiers have the ability to learn how to recognise abnormal values in a dataset.

Linear Discriminant Classifier (LDC), is a technique which works by sorting or dividing data into groups based on characteristics to create a classification [56]. A discriminant function is obtained by monotonic transformation of posterior probabilities [57]. In other words, it performs an ordered transformation of unknown quantities, which are separated by a linear vector.

Quadratic Discriminant Classifier (QDC) works in a similar way to LDC by dividing the data into groups based on given characteristics. However, by using QDC the data is divided using a quadratic surface rather than a one-dimensional one. QDC makes no assumptions that covariance are alike. In other words, it assumes that the changing of two random variables will not be the same [58].

Uncorrelated Normal Density based Classifier (UDC) also operates comparably to the QDC classifier but computation of a quadratic classifier, between the classes in the dataset, is done by assuming normal densities with uncorrelated features. Quadratic Bayes takes decisions by assuming different normal distribution of data [59]. LDC, QDC and UDC are density based classifiers.

Polynomial Classifier (POLYC) is also a linear based classifier and it is used to sort data by evaluating the weighting, using a linear combination of features and considering the variables of the objects [57]. In detail, it functions by adding polynomial features (which are constant coefficients) into the data which supports the training of the classifier.

Decision Tree (TREEC) is a classifier which uses decision rules to divide the classes of data [57]. It operates by using criterion functions (the sum of squared errors), stopping rules (criteria for appropriate number of splits in a decision tree) or pruning techniques (the removal of unwanted tree sections). Using decision tree is a particularly ideal choice of classifier because it is well-known as one of the most effective supervised classification techniques [58].

Parzen Classifier (PARZENC) functions by including aspects of the training data when the classifier is built up. It is a non-linear classifier and it has the benefit that its parameters can be user supplied or optimised [57], [58].

k-Nearest Neighbour (KNNC) is similar to the Parzen Classifier in that it includes training data when building up the classifier. KNNC however, predicts values based on the 'k-closest' values from the training set. In other words data is classified by a majority decision by identifying 'k-objects' which are nearest to its neighbours [57].

Support Vector Classifier (SVC) functions by predicting two possible outputs from a given training feature. It uses quadratic programming for optimisation and its non-linearity is determined by the kernel, which maps data into a set. Naïve Bayes Classifier (NAIVEBC) functions by applying Bayes' theorem to the dataset with independent suppositions. NAIVEBC is able to function with missing values and has the ability to learn incrementally [60].

## 5. Conclusion

In this paper, a discussion was put forward on the challenges of big data analytics in the growing digital world. As the amount of data being created every day increases, uncovering information in significantly large datasets is becoming more of a challenge. Factors, such as information security, digital threats and information sharing, require the use of big data analytics to uncover hidden information and enhance the services provided.

In terms of security, improved support can be provided, as well as cost efficiency. Processing large datasets, using big data evaluation techniques, to uncover anomalous behaviours in a system, can enhance existing security methods. In the IoT, big data analytics has benefits for the

well-being of people and helps with the evolution of integrated digital devices. This is particularly the case in healthcare, where it plays an important role in the early detection of degenerative illnesses.

# References


[1] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy, "Smart objects as building blocks for the Internet of things," IEEE Internet Comput, vol. 14, no. 1, pp. 44–51, 2010.

[2] L. Atzori, A. Iera, and G. Morabit, "L. Atzori, A. Iera, and G. Morabit," Comput. Networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] L. Mainetti, L. Patrono, and A. Vilei, "Evolution of wireless sensor networks towards the Internet of Things: A survey," in In 19th International Conference on Software, Telecommunications and Computer Networks, 2011, pp. 1–6.

[4] N. Group, "More than 400 Million Devices are Connected in U.S. Homes, According to The NPD Group," https://www.npd.com/wps/portal/npd/us/news/press-releases/more-than-400-million-devices-are-connected-in-us-homes-according-to-the-npd-group/. [Accessed: 08-Jan-2013]. .

[5] Cisco, "Cisco Visual Networking Index Forecast Projects 18-Fold Growth in Global Mobile Internet Data Traffic From 2011 to 2016," Available: http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=668380. [Accessed: 01-Aug-2012]. .

[6] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018," pp. 1–40, 2014. .

[7] M. Weiser, "The Computer for the 21st Century," Sci. Am, vol. 265, no. 3, pp. 94–104, 1991.

[8] L. Da Xu, W. He, and S. Li, "Internet of Things in Industries: A Survey," IEEE Trans. Ind. Informatics, vol. 10, no. 4, pp. 2233–2243, 2014.

[9] Intel, "Intel," Mark. Appl. 2012, vol. Available:

[10] D. for B. I. & S. UK, "Smart Cities: Background paper," vol. 2013, pp. 3–11.

[11] S. Mitchell, N. Villa, M. Stewart-Weeks, and A. Lange, "S. Mitchell, N. Villa, M. Stewart-Weeks, and A. Lange," 2013, pp. 1–21.

[12] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," IEEE Internet Things J, vol. 1, no. 1, pp. 22–32, 2014.

[13] E. Tok, F. Al Mohammad, and M. Al Merekhi, "Crafting Smart Cities in the Gulf Region: A Comparison of Masdar and Lusail," Eur. Sci. J, vol. 2, pp. 130–140, 2014.

[14] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter Cities and Their Innovation Challenges," Comput. (Long. Beach. Calif), vol. 44, no. 6, pp. 32–39, 2011.

[15] M. N. K. Boulos and N. M. Al-Shorbaji, "On the Internet of Things, smart cities and the WHO Healthy Cities," Int. J. Heal. Geogr., vol. 13, no. 10, pp. 1–6, 2014.

[16] W. H. O. (WHO), "Healthy Cities," [Online]. Available: http://www.euro.who.int/en/health-topics/environment-and-health/urban-health/activities/healthy-cities. [Accessed: 11-Jun-2015]., 2015.

[17] D. Niewolny, "How the Internet of Things Is Revolutionizing Healthcare," 2013.

[18] C. Dobbins, P. Fergus, G. Stratton, M. Rosenberg, and M. Merabti, "Monitoring and reducing sedentary behavior in the elderly with the aid of human digital memories," Telemed. e-Health, vol. 19, no. 3, pp. 173–185, 2013.

[19] S. Luo, J. S. Jin, and J. Li, "A Smart Fridge with an Ability to Enhance Health and Enable Better Nutrition," Int. J. Multimed. Ubiquitous Eng, vol. 4, no. 2, pp. 69–80, 2009.

[20] A. Pantelopoulos and N. G. Bourbakis, "Prognosis - A Wearable Health-Monitoring System for People at Risk: Methodology and Modeling," IEEE Trans. Inf. Technol. Biomed, vol. 14, no. 3, pp. 613–621, 2010.

[21] W. Hirt, "Ultra-wideband radio technology: overview and future research," Comput. Commun, vol. 26, no. 1, pp. 46–52, 2003.

[22] J. Haartsen, "Bluetooth - The universal radio interface for ad hoc, wireless connectivity," Ericsson Rev, vol. 1, no. 3, pp. 110–117, 1998.

[23] J.-J. Wang and S. Wang, "Wireless sensor networks for Home Appliance Energy Management based on ZigBee technology," in in 2010 International Conference on Machine Learning and Cybernetics, 2010, pp. 1041–1046.

[24] M. Swan, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," J. Sens. Actuator Networks, vol. 1, no. 3, pp. 217–253, 2012.

[25] M. Merabti, M. Kennedy, and W. Hurst, "Critical Infrastructure Protection: A 21st Century Challenge," in 2011 International Conference on Communications and Information Technology (ICCIT) , 2011, pp. 1–6.

[26] Á. MacDermott, Q. Shi, M. Merabti, and K. Kifiyat, "Considering an elastic scaling model for cloud security," in International Conference for Internet Technology and Secured Transactions (ICITST), 2013.

[27] S. Sridhar and G. Manimaran, "Data integrity attacks and their impacts on SCADA control system," in IEEE PES General Meeting, 2010, pp. 1–6.

[28] J. Walker, B. J. Williams, and G. W. Skelton, "Cyber security for emergency management," in 2010 IEEE International Conference on Technologies for Homeland Security (HST), 2010, pp. 476–480.

[29] E. J. Kartaltepe, "Towards Blocking Outgoing Malicious Impostor Emails," in 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM'06), 2006, pp. 657–661.

[30] R. Amin, J. Ryan, and J. van Dorp, "Detecting Targeted Malicious Email Using Persistent Threat and Recipient Oriented Features," IEEE Secur. Priv. Mag., no. 99, pp. 1–1, 2011.

[31] S. Tang, "The Detection of Trojan Horse Based on the Data Mining," in 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp. 311–314.

[32] C. Barrett, R. Beckman, K. Channakeshava, F. Huang, V. Kumar, A. Marathe, and M. Marathe, "Cascading Failures in Multiple Infrastructures: From Transportation to Communication Network," in IEEE International Conference on Critical Infrastructures, 2010, pp. 1–8.

[33] E. Knapp and J. Broad, "Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA and Other Industrial Control Systems," Syngress, Elsevier, 2011.

[34] Y. Chao, C. Bingyao, D. Jiaying, and G. Wei, "The research and implementation of UTM," Wirel. Mob. Comput. (CCWMC 2009), IET Int. Commun. Conf., pp. 389–392, 2009.

[35] Y. Zhang, F. Deng, Z. Chen, Y. Xue, and C. Lin, UTM-CM: A Practical Control Mechanism Solution for UTM System. IEEE, 2010.

[36] F. Deng, A. Luo, Y. Zhang, Z. Chen, X. Peng, X. Jiang, and D. Peng, TNC-UTM: A Holistic Solution to Secure Enterprise Networks. IEEE, 2008.

[37] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," IEEE Netw., vol. 8, no. 3, pp. 26–41, 1994.

[38] E. Brezhnev and V. Kharchenko, "BBN-based approach for assessment of Smart Grid and nuclear power plant interaction," in East-West Design & Test Symposium (EWDTS 2013), 2013, pp. 1–7.

[39] M. Erol-Kantarci and H. T. Mouftah, "Energy-Efficient Information and Communication Infrastructures in the Smart Grid: A Survey on Interactions and Open Issues," IEEE Commun. Surv. Tutorials, vol. 17, no. 1, pp. 179–197, 2015.

[40] R. Berthier, J. G. Jetcheva, D. Mashima, J. H. Huh, D. Grochocki, R. B. Bobba, A. A. Cardenas, and W. H. Sanders, "Reconciling security protection and monitoring requirements in Advanced Metering Infrastructures," in 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2013, pp. 450–455.

[41] D. Li and B. Hu, "Advanced metering standard infrastructure for smart grid," in 2012 China International Conference on Electricity Distribution, 2012, pp. 1–4.

[42] C. Chalmers, W. Hurst, M. Mackay, and P. Fergus, "Smart Meter Profiling For Health Applications," in The Internal Joint Conference on Neural Networks, 2015.

[43] C. Chalmers, W. Hurst, M. Mackay, and P. Fergus, "Profiling Users in the Smart Grid," in The Seventh International Conference on Emerging Networks and Systems Intelligence, 2015.

[44] N. Nicholson, "SCADA Security in the light of Cyber-Warfare," Elsevier Comput. Secur. J., vol. 31, no. 4, pp. 418–436, 2012.

[45] H. Zhang, J. Ma, Y. Wang, and Q. Pei, "An Active Defense Model and Framework of Insider Threats Detection and Sense," in 2009 Fifth International Conference on Information Assurance and Security, 2009, pp. 258–261.

[46] L. Buttyan, D. Gessner, A. Hessler, and P. Langendoerfer, "Application of wireless sensor networks in critical infrastructure protection: challenges and design options [Security and Privacy in Emerging Wireless Networks," IEEE Wirel. Commun., vol. 17, no. 5, pp. 44–49, Oct. 2010.

[47] A. K. Jain, "Data Clustering: 50 years beyond K-means," Pattern Recognit. Lett, vol. 31, no. 8, pp. 651–666, 2010.

[48] G. a. Wilkin and X. Huang, "K-Means Clustering Algorithms: Implementation and Comparison," in in Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007), 2007, pp. 133–136.

[49] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl. Inf. Syst, vol. 14, no. 1, pp. 1–37, 2007.

[50] [50] B. Kikhia, A. Boytsov, J. Hallberg, Z. ul H. Sani, H. Jonsson, and K. Synnes, "Structuring and Presenting Lifelogs based on Location Data," Image (IN), vol. 4, pp. 5–24, 2011.

[51] B. Fischer and J. M. Buhmann, "Bagging for path-based clustering," IEEE Trans. Pattern Anal. Mach. Intell, vol. 25, no. 11, pp. 1411–1415, 2003.

[52] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math, vol. 20, pp. 53–65, 1987.

[53] M. G. Forero, F. Sroubek, and G. Cristóbal, "Identification of tuberculosis bacteria based on shape and color," Real-Time Imaging, vol. 10, no. 4, pp. 251–262, 2004.

[54] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," Data Knowl. Eng, vol. 60, no. 1, pp. 208–221, 2007.

[55] G. McArdle, A. Tahir, and M. Bertolotto, "Spatio-temporal clustering of movement data: An application to trajectories generated by human-computer interaction," ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci, vol. 1, no. 2, pp. 147–152, 2012.

[56] E. Kuncheva, L, Combining Pattern Classifiers: Methods and Algorithms, IEEE Transactions Neural Networks, vol. 18, issue: 3, 2004.

[57] P. Fergus, P. Cheung, A. Hussain, D. Al-Jumeily, C. Dobbins, and S. Iram, Prediction of Preterm Deliveries from EHG Signals Using Machine Learning, PLoS One, vol. 8, no. 10, p. e77154, Oct. 2013.

[58] R. P. . Duin, P. Juszczak, P. Paclik, P. Pakalska, D. De Ridder, D. M. . Tax, and S. Verzakov, A Matlab Toolbox for Pattern Recognition, Version 4. Delft Pattern Recognition Research, 2007.

[59] F. Lotte, Study of Electroencephalographic Signal Processing and Classification Techniques towards the use of Brain-Computer Interfaces in Virtual Reality Applications, 2009.

[60] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249–268, 2007