Enabling Big Data Applications for Security

Responsible by Design

The Hague Security Delta

Enabling Big Data Applications for Security

Responsible by Design

The Hague Security Delta

Highlights Big Data Applications for Security

Big Data is the next great opportunity for security and safety organisations and individuals. A computer's ability to process large amounts of data will allow us to perform faster, more accurate analyses and decision making and will create a real-time, situational understanding of security situations.

Big Data, High Performance Computing and Artificial Intelligence have a large set of multifaceted capabilities to contribute to security. In addition of its benefit, the use of these technologies trigger new issues, such as: trust in the technology, the risk of exclusion or discrimination, new vulnerabilities because of the omnipresence of data and the misuse of personal data. Algorithms based on the use of Big Data and Artificial Intelligence are increasingly used for decisions that have significant impact on individuals and societies.

This report introduces a framework for the responsible development and use of Big Data applications in the security domain. The framework's goals are to make programmers, project managers, policy makers and board directors aware of the issues, and to support the development of responsible Big Data applications for security.

How to apply Big Data in security?

It is important to take a responsible approach to prevent and manage the risks for ethics, privacy, security, safety and integrity that may arise in applications that make use of Big Data. The aspects for responsible design that have to be taken into account are referred to as 'invariants'.

Framework of Invariants

Invariants	Observations by…	Measurements with…
Ethics	Ethical committee	Responsibility
Privacy	Privacy Barometer	Reciprocity
Security	Intrusion Detection Systems	Adaptability
Safety	Self-healing	Autonomy
Integrity	Scrubbing	Curation

The framework is explained and examples of its use are shown throughout the report. In addition to technological examples, the report also stresses the need for strategy and organisation. More capable organisations and individuals will be better enabled to act upon security risks.

In time, successful application of big data technologies will rely on the completeness of the collected data, the quality of the relevant analysis and the capability to effectively act upon it. You'd better be prepared.

Foreword

There are all kinds of new applications in development today that make increasingly smarter use of Big Data. Take for example the recent announcements that judicial decisions can be predicted by computers. In the security domain there are numerous possibilities to use these technologies too. At the same time the security domain bears the risk of exclusion and inducing mistrust among the civilian population regarding the limitless use of personal data. This report outlines how to develop technologies responsibly for this challenging domain.

Big Data is the next great opportunity for security organisations and individuals alike. The idea is to feed a computer with large amounts of data in order to perform faster and more accurate analyses and create a real-time, situational understanding of security situations. At this point however, the technological developments remain in a state of flux.

The authors hope to inspire the reader and increase awareness as well as provide insights into how to develop responsible Big Data applications for security.

This report provides a description of technological developments and explains how these developments can be used to support more responsible, effective and efficient security.

Three of the most prevailing observations are:

1. Big Data and High Performance Computing have a very large set of multifaceted capabilities to contribute to security;
2. For the acceptance of the applications, society, public organisations and the corporate business world should establish a framework on the eligibility of these developments;
3. Decisions affecting humanity are increasingly being taken by computers, which raises tremendous responsibility regarding the correctness and ethical values represented in Artificial Intelligence.

The observations evoke, among others, the following five questions that will be addressed in this report:

- What are computers able to do with Big Data?
- What do we ask from the computers?
- What is desirable to ask from the computers?
- What happens in practice at this moment?
- What should be the Call for Action?’

Contents

1. Introduction

1. Introduction

Big data[i] is a trending topic these days. It is not however a completely new phenomenon, and the start goes back years ago when large corporations, government agencies and intelligence services started using ever larger datasets and creating new combinations of different datasets. Back then this was simply large data. What makes it interesting today, and what partially explains the current attention given towards the topic, is that the possibilities are now within reach of everybody, both organisations and individuals.

As a result of the availability of computer power in the cloud, the omnipresence of affordable mobile computers, as well as digitalised datasets, what used to be called large data we now call Big Data. The total volume of usable Big Data grows by the minute as data are uploaded on the Internet, and produced by its users or machines. In combination with smart powerful analytical tools and the fact that the costs of these capabilities are becoming very low, this makes it possible for everyone to produce and exploit data in radically different ways.

Regardless of the application domain, Big Data seems to be a promising terrain where smart combinations of data can not only enhance situational awareness but also support an enhanced understanding of the consequences and options people and organisations have. Big Data can be used with aim to improve a combination of both awareness and understanding. In the security domain this might lead to better preparedness and resilience.

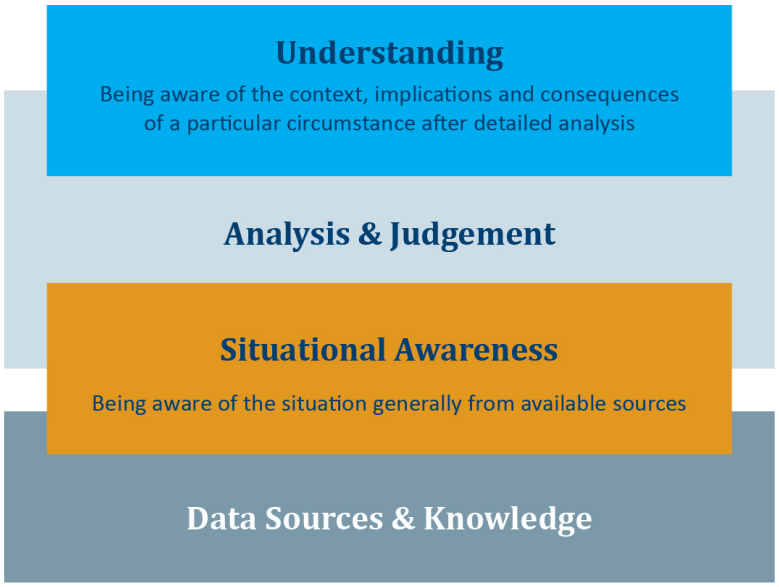


Figure 1 Simplified model sketching the difference between Situational Awareness and Understanding. For a detailed model see British Military Joint Doctrine Publication #4 "Understanding" [ii]

The capability to collect data is becoming of subordinate importance to the analysis and use of these data themselves.[iii] Some describe the developments by V's: volume (the scale of data is expanding), velocity (the analysis of streaming data is speeding up), variety (different usable forms of data are expanding), veracity (the uncertainty of data is expanding too, sometimes called variability) and value (either business value or value for society). In combination, the V's for Big Data enable new, and possibly even disruptive forms of data exploitation.

In Figure 2, we provide an overview of the relation between Dig Data, Data Science and Data Discovery (Source: Gartner).

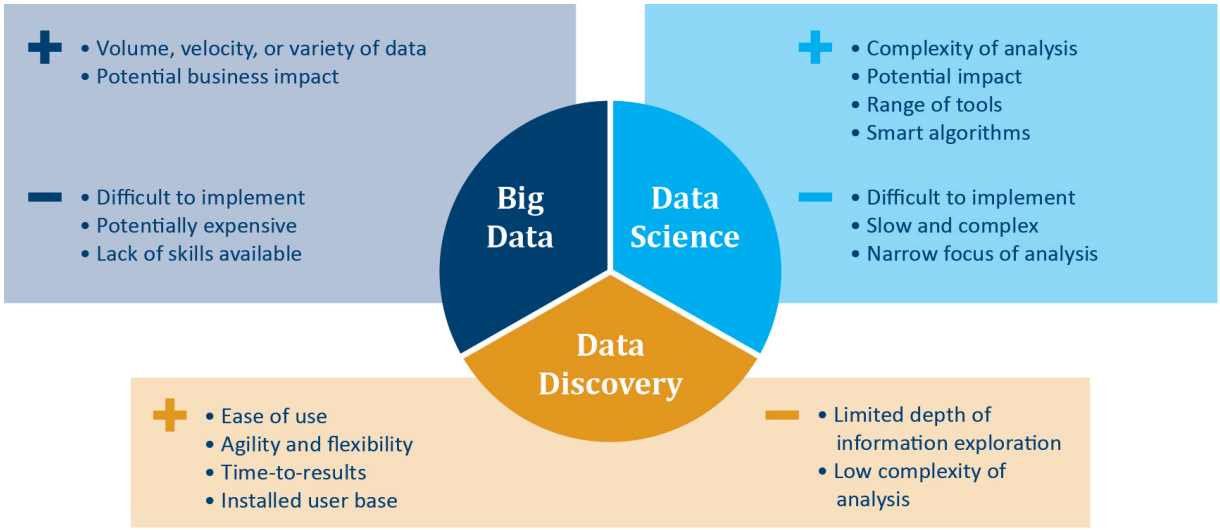


Figure 2 Big Data Discovery is the combination of Big Data, Data Science and Data discovery[iv], *Source* Gartner

The data can come from many sides, from users to applications, from systems to sensors. The involved parties generate more data every day. The maturity of the Internet of Things (see Figure 3, our source is Gartner) as the overarching platform connecting all of these parts is also rapidly growing.

Figure 3 The Internet of Things and Big Data[v]

The current explosion of developments in the digital domain has a large impact on science and society as well as on business and government. We see more data produced every day, more devices connected to the Internet, more applications developed and more people connecting to the Internet, but also more financial losses as a result of cyber-attacks, more cybercrimes and a large decline of public trust in these systems. Many devices connected to the Internet are never updated or cannot receive new firmware, making newly discovered vulnerabilities more challenging to repair easily. As a result, these devices become more susceptible to cyber threats. This process became apparent in October 2016, when a large scale DDoS attack was launched against Dyn by mobilizing a large number of IoT devices, such as printers, security cameras, home routers and baby monitors.

In the security realm, intelligence services were traditionally, and sometimes still are, frontrunners in the way they collect and process data. The emerging Internet of Things, as a system of systems platform with all its users combined, is now much better at collecting and processing data. As a result, alternative media and systems today produce and cover data from almost all walks of life.

Hence, it is clear that Big Data provides several opportunities for the security domain. Large amounts of data in particular, sometimes real time and unstructured, offer opportunities to perform faster and more accurate analyses and create a real-time, situational understanding of security situations. The big question is of course how individuals and organisations concerned with or involved in security issues can benefit from these developments. In order to provide anchor points, this report provides a description of technological developments and explains how these developments can be used to support more responsible, effective and efficient security.

The report focuses on an audience of public and private organisations in the domain of public order and security. In particular, the report will address the way the audience will or could be using Big Data for security in the future. Furthermore, this report addresses what individuals can do to handle security risks. We will not deal with the question as to how these organisations and individuals should use different datasets and analyse them, but will instead sketch an overview of Big Data developments by presenting a framework with which we may better understand these developments and to formulate the necessary technological requirements that aim to ensure that computers act responsibly. We also look towards the future to anticipate a) what possible developments of the phenomenon of Big Data might look, b) what it would require for both public and private organisations and for individuals and c) what opportunities and threats we would then have to take into account.

Last but not least, we hope to inspire readers in exploring the possibilities of Big Data analysis for security in a way that both security as well as privacy is dealt with in a responsible manner.

Readers guide

Chapter 1 introduces the concept of Big Data and it is possible applications for security. Chapter 2 identifies five invariants for computers to take into account. They are: ethical considerations, privacy, security, safety, and integrity. Chapter 3 discusses technical opportunities. Chapter 4 provides a perspective on information strategies for handling Big Data. Chapter 5 provides some ideas on how to best organise the use of Big Data and examples of applications that may change one's opinion on what is desirable. Finally, Chapter 6 summarizes conclusions, recommendations and provides some other observations.

2. The Five invariants in a continuously changing ICT society

2. The Five invariants in a continuously changing ICT society

Our current way of thinking emphasizes (1) the performances of computers and (2) the storage of data (Big Data). In the early stages of the Digital Era (1950-1980) emphasis was on performance, then in later years (1980-2015) there ensued a fierce competition between computing and storage. Computing followed the well-known Moore's law (i.e., every eighteen months the computing capabilities of the computers double), while storage followed the lesser-known, but more aggressive, Kryder's law (i.e., every fourteen months the price of storage capacity of the computers halves).[vi] This essentially meant that our data storage capacity is overtaking our ability to compute the data supplied at an exponential rate. At the same time, we observe effects of the fourth paradigm by Jim Gray[vii] pointing at the transition from data-poor to data-rich science. The first signs of this transition may come from science, but similar transitions can be observed in business and society. Below we discuss the question: what can we ask from computers? We do so by following a path consisting of three sub questions:

1. What are computers able to do with Big Data?
2. What do we ask from the computers?
3. What is desirable to ask from the computers?

The first question: "What are computers able to do with Big Data?" can be answered rather straightforwardly by noting the clear shift of actions that were previously performed by humans and are now performed by machines. Computers have mastered games such as chess and Go, they are instrumental in self-driving cars, and they guide drones in areas of war. Moreover, we see block chains acting as storage capabilities for banking accounts, in principle enabling financial transactions with Bitcoins. Furthermore, computers continually observe many aspects of humans in society. Here, we only have to point to the NSA papers disclosed by Snowden, or to the Data usage policies of Google and Facebook to support the above statement. We see the computer activities being present in all layers of business and commerce. However, we may also see them used in the world of "criminals" and "terrorists". So, much like the data itself, so too are computers everywhere. But we also see that the sensible computation of an ever increasing data tsunami is a lost cause even for these computers. Consequently, we are waiting for more clever boys and girls to help program new systems that can better address these trends. As a result we promote Big Data education and increased High Performance Computing research, as disruptive developments such as quantum computing may bring a solution.

The second question: "What do we ask from the computers?" brings us back to the collection of small and big problems that are in need of being addressed. The issues include several ethical considerations related to privacy, security, safety and integrity. Later on they are referred to as the *invariants.* We discuss all five, with regard to the security domain. For security three distinct roadmaps should be designed, viz. for policy makers, for executors, and for users. The strengthening of security is of utmost importance for the prosperity of the Netherlands.

The first person to direct attention towards these questions in the Netherlands was Bob Herschberg, who at the time, around 1980, was a professor at the Delft University of Technology. He gave his students the task to hack into computers and the wider mainframe of the university. In those early days computers, data, and filesystems where poorly protected, and the students found numerous

weak spots with ease. At the same time however they gained access to sensitive information, immediately raising privacy and ethical concerns.

The delicate balance between technological and societal developments can only be understood when knowledge from both developments are available, and even then it is difficult to answer the question whether the world is technology-driven or socially-driven. Before we address the question in full we will first sketch the "scientific" development of "disruptive" technology in the past, present, and future. Here we remark that Brexit and Donald Trump's election as President of the United States have opened the eyes of many humans, and demonstrated the obsolescence of many old statistical theories. Stratified sampling for example, is doomed to fail. In these times of data science and Big Data, new statistical theories are in need of being developed. All in all we will likely see a development as listed in Table 1.

Table 1 Developments in disruptive technologies

Artificial Intelligence	1950-1990
Machine Learning	1990-2000
Adaptivity	2000-2005
Dimension Reduction	2005-2010
Deep Learning	2010-2020
Big Data & HPC	2012-2025
New Statistics Theories	2014-2025
Quantum Computing	2016-2035

A second view at the developments is as follows. After World War II we saw an explosion of computer science activities, within which Artificial Intelligence (AI) paved its own path. Until 1980 the technological developments remained in the lead. From 1980 to 2010, we saw attention being drawn towards the threats and pitfalls that computers pose to society. The period of 2010 to 2017 has brought about a new phase: the arrival of disruptive technologies.

By 2010, the scientific world was aware of the threats that interconnected databases posed to privacy. Social and legal research emphasised the gravity of the threats. One of the first research contributions in this interesting area is found in Matthijs Koot's (2012) Ph.D. thesis titled "Measuring and Predicting Anonymity."[viii] Koot formulated a mathematical framework with which he could measure and predict the impact on privacy when handling multiple data sources. After his Ph.D. defence at the University of Amsterdam, he and his supervisor stumbled into a case with a surprising combination of four issues, viz. ethical considerations, privacy, security, and integrity.

The case is as follows. A banking company, willing to help capable entrepreneurial youngsters, published a psychological test entitled "how fit am I to be an entrepreneur?" The test itself was meant for unemployed people who sought to start an enterprise of their own. The test was developed for the bank by a small company. Apparently, the test was a big success, given that the website mentioned that 45.000 people already have taken the test. The test was taken online and had a price tag of ten euros.
Upon a participant's answering an extensive list of personal questions and paying, a link was provided where their personal report (30 pages) could be downloaded. The link had the form

<URL>?id=45420. A straightforward combination of the ID number with the website statement of 45.000 aroused suspicion. What would happen if the number 45420 would be changed into 45419 (the first ethical decision)? The results of the previous client would then appear. Further testing on 3.000 earlier IDs showed (the second ethical decision) that the website had no security measures or rate limiters installed. It was immediately understood how fragile the IT system was and the heads of Computer Science, Faculty and University Authorities were notified. The responsible disclosure approach was used to inform the bank, and the personal data obtained in the process was encrypted and later destroyed.

The story of the psychology test is a typical example of a series of data leakage incidents that show the vulnerabilities of IT systems. From the example we learn that computer science is facing at least the following five issues: ethical decisions (curiosity is a clear opponent of ethical behaviour), privacy (to what extent is it allowed to read other people's answers?), security (which security measures should be enforced from IT companies and their employers?), safety (no harm to other people) and integrity (which "unwritten" rules should still be obeyed?)

In this report the question of security will be emphasized as the most prevailing one (What should be done to achieve optimal security?). If we take security as the central point of our discussion, then we see that it is positioned by us as the midpoint of a list of five invariants (in which safety is inserted after security). We have identified these five as the most important ones for a vital, vivid and transparent society. All five have their own dynamics and we will therefore give them individual treatment. However they will also be discussed in connection with one another. In our opinion, so far they have not received the attention they deserve, and certainly not in connection with each other. Up to now the five invariants are dealt with in retrospect (as a reaction on an event), but we would emphasise the prospect of their development.

For this purpose we first have to introduce the concept of safety. We define safety as "protecting the system as well as possible (preferably in a perfect way) against *unintended* damage". Examples of sources of unintended damage are: damage caused by fire, a deluge of water, nuclear material, and earthquakes. The definition of safety is different from that of security in only one word: unintended (safety) versus intended (security). Of course, there is a small grey area between these two notions. However, in this report we focus on security. Safety is considered as a close neighbour of security.[1] This is the answer to our second question.

The third question is: "What is desirable to ask from the computers?" We take the run-off in the 1970s by rule-based systems. Then we see a development to heuristics (1980s), cases (1990s) and data (1995). Collecting, cleaning and using data obviously comprises ethical issues. In the upcoming world of Internet of Things we understand that "data" will play their own role in the three processes of collecting, cleaning and using. For this idea we have coined the concept of responsible data. Below we discuss the questions (a) how can we observe that data is responsible? And (b) how can we

[1] The identification of these five invariants may stimulate colleague researchers to broaden the list by other issues, e.g., transparency. That is quite possible, but by these five invariants, we claim, are the proper characterized the needs of the current and future co-habitation and cooperation of computers and human beings in our society.

measure the extent to which data is responsible? These two important questions will also be discussed in the context of privacy, security, safety, and integrity.

For the University of Amsterdam (UvA), the entrepreneurial case, one case in what would become a series of incidents, was a clear warning that computer science had left its stage of infancy. As soon as ethical issues are left intrinsically unsolved in the contents of a study, that study should be restructured, e.g., along the lines of medical sciences. There, plans on experiments with human beings must first be submitted to and approved by an ethical committee. Analogously, this should happen to computer scientists who would like to perform experiments, tests, or research with IT systems and responsible data.

2.1 Brief overview of the five invariants

2.1.2 Ethical issues and Responsibility

Nowadays, the UvA has an ethical board consisting of several senior staff members and a legal advisor. Research proposals and student projects that may involve ethical issues must be submitted to the board. Of course, there are subcommittees, websites with frequently asked questions, and online questionnaires. Within a period of four years the ECIS (Ethical Committee for Information Sciences) had achieved a front ranked position in the area of ethical consideration. Their line of reasoning is observable and their decisions are measurable. It "measures" to what extent somebody or something is responsible (here we refer to responsible data).
The working definition of responsible data comes from the responsible Data Forum (2014). It states that responsible data is "the duty to ensure people's right to consent, privacy, security and ownership around the information processes of collection, analysis, presentation and reuse of data, while respecting the values of transparency and openness."

2.1.3 Privacy and Reciprocity

Obviously, the complete protection of privacy is impossible and, in our opinion, also undesirable. In particular, the last statement opens the door for discussion on the extent to which privacy should be protected. Some authors even state that privacy nowadays is dead (see, e.g., Mangon, 2014). Disregarding the extreme opinions, there is an issue next to protection, namely reciprocity. At least in the public domain reciprocity should be enforced. For example, citizens have the right to know that they passed 400 cameras during their car drive from Breda to Eindhoven. Similar to responsibility for data, the right on reciprocity (informing the observed side from being observed, e.g. by the authorities) is also a new item for privacy. It can be observed by a privacy barometer[ix] and by a privacy authority. Moreover, it can be measured by a variety of tools that measure awareness, and by mathematical testing models, such as the framework developed by Kort (2012).

2.1.4 Security and Adaptability

Security is the main issue of this report. In the Netherlands there are many groups dealing with security issues. The observable is the success and failure rate of intrusion detection systems (IDS). In case of a failure, the intriguing question is whether observations, recordings and machine learning techniques can effectively support the repair tasks to adapt the system to become more secure.

Adaptability of prevention is important, but even more important is proactive behaviour. The branches of the sciences involved are diverse, and range from psychology to artificial intelligence to high performance computing.

2.1.5 Safety and Autonomy

As stated above, safety and security are closely related. We refer to the intended cause (security) and unintended cause (safety). Next to robustness, the notion of RealTime Foresight (RTF, Weber, 2017) is important for anticipating disasters. After a disaster, the choice of the type of disaster management plays a key in recovering the system. Many of the adaptability operations have been anticipated upon to a large extent. This implies that their incorporation may take place autonomously by processing cooperatively and dynamically essential procedures in intelligent networks. The question here is: to what extent do we give the system autonomy to implement the suggested adaptations in order to let the system survive? The observables are in the area of self healing. The measurements are in the relation between the selfhealing procedure and the autonomy given to them.

2.1.6 Integrity and Curation

The execution of scientific research by human beings should be performed in a way that ensures full integrity. The same holds for machines that replace human beings in some parts of the research. Currently, we have in both areas (humans and machines) a plethora of examples that are exceptions to the requirements of being integrous. A key process that ensures that data fulfils integrity conditions is data scrubbing (an error correction technique). A data scrubbing procedure periodically inspects the storage for errors. It corrects any found errors by using redundant data and different checksum techniques. We use the term 'data curation' since it is a wider technique that also includes the management activities related to the organisation and integration of the data processes.

2.2 Observation and Measurement

In Table 2 we provide an overview of the way of observations and measurements of the five invariants. As stated in Table 2, security can be observed by intrusion detection systems and the measurement of the impact of being not secure can be given by the degree of adaptability. However, an actual implementation of these concepts will immediately reveal different opinions.

Table 2 Five Invariants, Examples of observations and Measurements

Five Invariants	Observations by	Measurements in 2016 with the help of
Ethical issues	Ethical committee	Responsibility
Privacy	Privacy Barometer	Reciprocity
Security	Intrusion Detection Systems	Adaptability
Safety	Self-healing	Autonomy
Integrity	Scrubbing	Curation

Here we distinguish four groups of computer users who all have different interests in the security issue. They are:

1. Security and intelligence services such as the NSA;
2. Google, Facebook and other ICT and digital technology companies;

3. Commerce and public administrations stakeholders;
4. Criminals and terrorists.

In this report we focus on group 3, viz. Commerce and Public Administration.

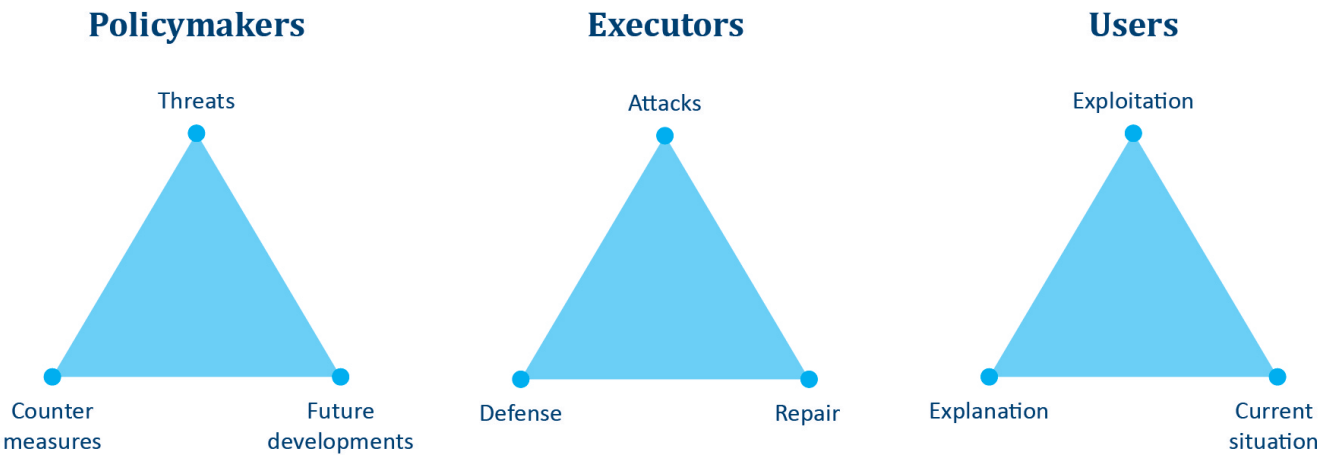
If we seek participation in the world of group 3 on the security issue we will likely receive answers that are heavily dependent on the objective of the group. We distinguish three groups and for all three groups it holds that the goal of security is alike, namely to strengthen the security of the Netherlands and to support an effective commerce and public administration. Here again several distinctions are possible, we mention the following three contrasting pairs:

1 Individual vs. government;
2 Private vs. public;
3 Individual vs. collective.

However, that is not what we aim to investigate; much more interesting is the tripartition into:

1 Policy makers;
2 Executors;
3 Users.

We return to this topic in Section 4. Below we only show the differences in approach. We do so by visualizing them as triangles, whose objectives are placed in the corners, effectively showing their relationships to one another.



The answer to the third question suggests that in order for computers to adequately address the five invariants, they should have sufficient knowledge of the context within which they operate. Indeed there is a large variety of contexts possible here.

3. Technological opportunities

3. Technological opportunities

At the beginning of the electronic computer era Thomas Watson, the then-president of IBM, stated "I think there is a world market for maybe five computers."[x] Indeed, nothing is more difficult than predicting, especially when it is about the far off future. In this chapter we show the trends that demonstrate where technology is leading us. We begin with a short background of the historical developments of computer hardware, software, and artificial intelligence. We then explore the fundamental principles needed to develop applications that are responsible by design - along the lines of the proposed five invariants.

3.1 A short history of Computers, Storage, Networks, and Software

3.1.1. Computers

The early technological developments in computer science and industry were purely hardware oriented. It was about putting a machine together that could automatically process algorithms using inputted data. The essence of such a machine is a digital switch that uses two inputs to produce an outcome. The first computers used relays as switches. The first electronic computers were based on vacuum tubes as switches. The invention of the transistor enabled the revolution in computer industry. Transistors could be miniaturized in integrated circuits (IC) on silicon chips. In the current state of technology, the transistors can become as small as a few nanometres. Figure 4 shows the developments in the miniaturization of transistors on integrated circuits. For decades the number of transistors per square mm has been following Moore's Law, which tells that after approximately eighteen months the number of transistors on IC's has doubled. As a result, more computing power and memory could be put on a piece of silicon of the same size. The increase of computing power was originally achieved by increasing the clock cycles. Physics turned out to end that avenue between 2 and 4 GHz.[xi] Placing more processor units on a chip initially enabled vector and deep pipeline processing of data. This was later followed by the thread and multicore handling, allowing for parallelization and thus, if software can cope with it, a further increase in processing power. Special single task processors often have much more computing power than generalized processors. For example Graphics Processing Units (GPUs) are optimized for a few simple tasks in image generation for display drivers, mostly vector and matrix operations on a massive parallel scale, but calculated in number of multiplications per time unit they beat any processor by orders of magnitude. So many computational scientists transformed their critical code to run on the GPU instead of the CPU and many of the aspects of GPU and field-programmable gate array (FPGAs) are now found in the mainstream processors. These disruptive changes initially enabled Moore's Law to continue (see Figure 4), however new boundaries are already in sight. The most important of those are the difficulty to get the data in time from memory to the processor units (the great memory wall) and the power consumption. New disruptive technology changes are necessary to overcome these barriers and are predominantly explored in research efforts exploring, for instance, new materials, 3D chip design and Quantum Computing.[xii]

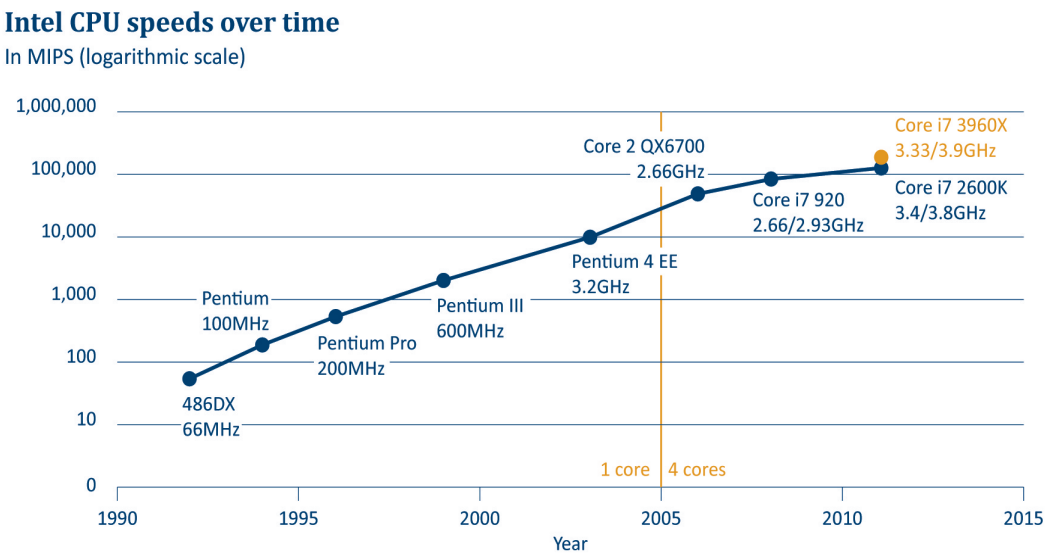


Figure 4 Intel CPU speed development in time

3.1.2 Storage

Similar to the continuous growth in computing power, Mark Kryder proposed a more debated law describing developments in storage. This Kryder law states that storage per unit of price roughly doubles in density every two years. Whether the projections of this law actually emerges or not, the actual worldwide growth in stored data has followed a similar pattern (see figure 5).

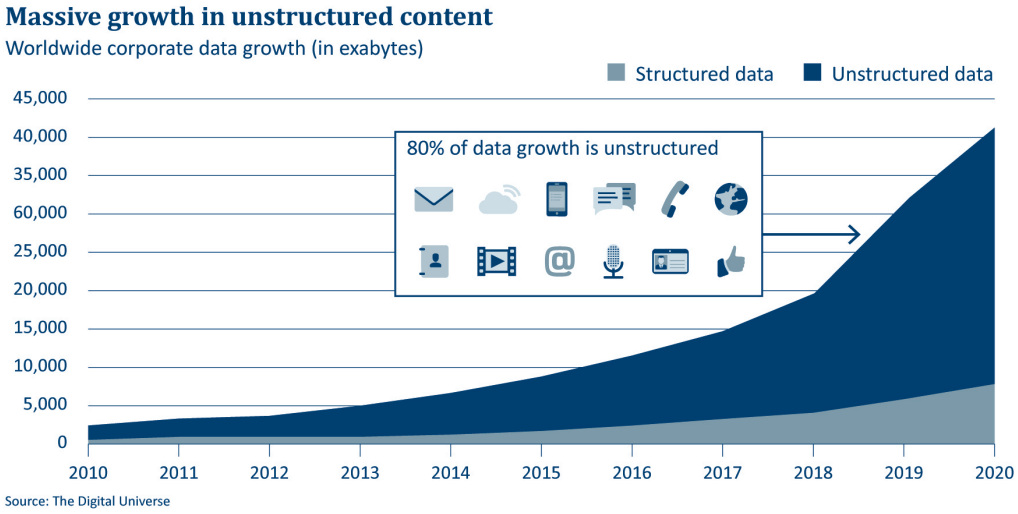


Figure 5 Growth in worldwide stored data

3.1.3 Networks

An important phenomenon that started the early 1960s is the development of computer networks. It started with wired local area networks, and the beginning of the wide area network, the ARPANET projects and the NSF network.[xiii] Over the years this resulted in the current optical fibre based worldwide Internet with connections speeds of 100s of Gigabit/s per colour wave and 10 – 20 Terabit/s per fibre. In technical terms, the different transport mediums are rapidly approaching the Shannon limit, which describes the maximum amount of information that can be transported in a channel of certain spectrum width and noise characteristics. Later, several long and short distance

wireless connection types were developed and standardized, such are Wi-Fi (started in Nieuwegein (NL) at NCR as a cash register radio communication system and protocol < http://www.wilcorpinc.com/wifi_history.htm>), GSM, 2G-5G, Bluetooth, and satellites. In addition, cable television providers and telephone network providers opened their networks for internet data traffic. In one way or the other, all these networks form one global network: the Internet.

3.1.4 Software

Putting a lot of hardware together does not magically turn it into a supercomputer. What makes a system work is the software, which enables a single application to utilize all the capabilities and power for its own benefit to complete the task at hand. This also has seen a phenomenal development. In the late 1970s, remote procedure calls enabled one computer to call a routine running on another computer to do something for the first. In the 1980s and 90s we saw the development of Message Passing Interface (MPI), allowing a stack of physical computers to work on a single problem. At the same time, compilers came into existence and allowed efficient programming of multi-tread and multi-core shared memory processors. The grid developments allowed to combine many MPI stack computer clusters to work together on a single data processing problem, see for example the Large Hadron Collider data processing grid. The cloud developments allowed for the above mentioned technologies to be scaled up to industrial levels that now completely dwarf the demand coming for the science communities.[xiv]

3.1.5 The future: Towards Internet of Things

With these developments, computing devices grew from isolated mainframes and into highly connected sensors within the so-called Internet of Things (IoT, see figure 6). It was when the first mainframes arrived that Thomas Watson made his statement about the limited needs of computing resources for humanity. In the late seventies, the IBM PC and its derivatives entered the offices and homes of many people. The home use of PCs was limited. It was only once the Internet entered the household domain that the PC became a central unit in house. The arrival of the World Wide Web (WWW) especially contributed to this change. At the end of the last century, mobile phones became truly portable handsets. This was enabled by developments in smaller rechargeable batteries, and the reduction of power consumption in the electronic parts. With the advent of smartphones the Internet became available at peoples handsets. Today, everybody is constantly connected. Beyond that, we see all kinds of wearables, sensors, and attenuators linked to smartphones and other computing systems. This development is visible in industry, such as in Supervisory control and data acquisition (SCADA) systems,[xv] in homes in the form of domotics systems,[xvi] and in personal systems as wearables for sports and health monitoring. In some cases the computing systems consist of components that can themselves be considered independent computing systems, while such a System-of-Systems (SoS) as a whole form a more complex system.[xvii]

Figure 6 Computer revolutions in time (from www.vensi.com)

Currently, reducing the power dissipation of computer processors is a major design criterion. When batteries are the main power source, computers must be energy efficient to be truly mobile.

3.2 A Short history of Artificial Intelligence

In 1970, Marvin Minsky predicted: "In from three to eight years we will have a machine with the general intelligence of an average human being." The first AI successes are indeed from the early 1980s, but the abilities of these so-called expert systems were far from those of human beings. AI developments proceeded gradually with several milestones. The first milestone was the defeat of the World Chess Champion Garry Kasparov by the IBM computer Deep Blue in 1997. Another major achievement is the winning of the TV quiz show *Jeopardy!* by IBM's Watson against the show's greatest former human champions in 2011. For this quiz show it was required to "understand" natural language, both in the posed questions and in the numerous sources to find answers. This year, almost 20 years after the defeat of the World Chess Champion, the World Go Champion Lee Sedol was defeated by Google's AlphaGo. Go was considered far more difficult to 'learn' for AI systems, because it was assumed that a specific human capacity, intuition, was needed. Given this, Sedol's defeat was ascribed huge importance within the AI development community.

The reported milestones rely to a great extent on developments in *machine learning*, which is an important branch of artificial intelligence. Machine learning provides the tools required in Big Data applications to obtain information and knowledge from the integrated structured and unstructured data sources. The latest tools are based on deep neural networks, also called deep learning. Much remains to come.

Though the performance of machine learning algorithms is impressive in many cases, for any serious problem there are no systems that work without *errors*. Errors are intrinsic to recognition tasks[xviii] and dealing with their consequences is of the utmost importance in application design. For instance, although men are (statistically) significantly taller than women, differentiating men from women based on length only will result in many errors. Importantly, without additional measures these errors will be repeated over and over. This is especially important to realize within the security domain, as it may lead to loss of trust in the system or the government.

3.3 The invariants in technological security domain

Within the scope of the development of computer systems, in the 1980s we witnessed a change from central mainframe computers to personal computers, where every user controlled his or her own computer and computation. The advantage of this change of control model was the ease of exploiting the increasing computing power for every user. In the last decade we've seen a reverse development, where centralized data centres and clouds store user data and provide computing services. With complex distributed hard and software infrastructures, the users obtain a centralized view of their service provider again.

With the enormous increase of IoT units, this tendency will naturally stop for reasons that we link to the invariants proposed in Chapter 2. In a distributed computing environment with IoTs and Systems of Systems (SoS), the view on their (virtual) world will be local. There is no need to share information with an omniscient being, nor is knowing of all other systems and their behaviour required. Indeed for maximizing the information and knowledge extraction in Big Data research, the slogan is the more data the better. For IoTs and SoSs however, when delegating control and sharing information, the trade-off between a central system at the one extreme and only oneself at the other extreme needs to be established into a new balance.

Below we discuss the invariants with respect to information sharing using underlying data protection principles such as proportionality, purpose limitation and data minimisation at the fundamental sharing levels that we call: *central*, *peers*, and *self*. We take an autonomous car (SoS) as an example,[xix] because all invariants impact on the security aspects of the cars and their environments.

3.3.1 Ethical issues (responsibility)

An automated car could disclose car speed and other potential incriminating information to the authorities. The question whether this is of mutual interest can be argued. Indeed, road safety improves when cars obey the traffic rules. Whether this is ethically desirable is to be considered. We can expect road users to be responsible without maximizing their own goals and challenging road safety. However, with optimally controlled sensors and motors, accidents could be avoided by timely giving selfish drivers way.

1. *Central*: Responsibility towards the authorities could mean that the autonomous cars are designed to strictly adhere to the law, or, alternatively, to enable the authorities to enforce the law maximally.
2. *Peers*: A car can manipulate its peers with untruthful information or by not ignoring collaboration attempts. Indeed, it is responsible behaviour to also let peers enable to reach their goals.
3. *Self*: Finally, in the long run, ethical and responsible behaviour is in the interest of self. The contrary will lead to being an unreliable partner on the road.

3.3.2 Privacy (reciprocity)

An autonomously driving car could share all kinds of information about the state of the car, the driving characteristics, the driver and its passengers. To model other cars on the road, it is required for each car to have a unique identifier. Information stored in relation to a car can be considered personal information. To enable other cars to anticipate, negotiate, and collaborate on their driving actions, there is, however, no need to share information with remotely driving cars or other SoSs. Information sharing considerations at the sharing levels are as follows:

1. *Central*: Disclosing personal information to a central authority has limited utility for the car driver/user and is therefore hardly reciprocal. Indeed, for instance with toll roads, the usability increases with automatic billing. Then, identifying information must be obtained and stored. Other information that the car SoS could share with a central system are observed accidents and traffic jams, or being involved in an accident itself. For the last, the interest is clearly reciprocal.
2. *Peers*: Sharing information with other cars is useful, if the other cars are within scope. Moreover, it is of mutual interest to collaboratively maximize the speed or convenience in reaching their destination. In such cases, the invasion of privacy is limited and reciprocal.
3. *Self*: The car itself can log personal information to learn from past activities and improve for the future. Storing this information permanently introduces a security risk, because by stealing this information the car/driver becomes predictable or otherwise vulnerable.

3.3.3 Security (adaptivity)

To be robust against malicious activities during collaboration and information gathering, the cars must monitor and diagnose their own state, as well as that of their peers, of the central servers, and of the infrastructure. The difference with the Safety invariant is that monitoring and diagnosing for security is more difficult. The intentional change in the system components will be harder to recognize, because the signals of the intrusion may be intentionally hidden.

1. *Central*: The measures to take with respect to the security invariant are largely the same as those for safety. Also for security it is important to monitor and diagnose whether the central systems can be relied upon when it comes to the information they provide and the information provided to them.
2. *Peers*: Since the set of peers of the car changes while moving, the monitoring must be a continuous process. The appropriate actions must be taken adaptively.
3. *Self*: Since the own sensors and subsystems may be compromised the system must diagnose itself for being safe and sane. Inconsistent states of the sensors are signals of security issues or failures. Having sensors for many modalities enables the car to adaptively use those sensor modalities that it trusts.

3.3.4 Safety (autonomy)

The first reason an autonomous car drives autonomously is to release the driver, now mostly passenger, from the burden of controlling the car, finding an optimal route, avoiding obstacles and collaborating with other cars in satisfying their goals. To enable the autonomous car to achieve this, it is provided with a large number of sensors to measure its own state and its surroundings and network connections to allow for information exchange with central servers and peers on the road. With respect to safety, reliance on information at different sharing levels impacts autonomy in the following way:

1. *Central*: The cars must be robust against failure of central servers and of the digital infrastructure. Inevitably, the cars must be robust against network interruptions or delays for their information services. They should be able to at least temporarily operate independent of central servers for route planning, weather forecasts, expected traffic and accidents. The car should monitor its state with respect to its trust on central servers. The car should switch serviced to autonomous mode or even stop driving.

2. *Peers*: Also peers may fail to report their state and intentions. Indeed sharing information helps in better achieving mutual or collective goals. However, in case of network interruptions or peers with failures, the car should be able to autonomously continue driving. The car should monitor its state with respect to its trust on peers. The car should switch from collaborative to autonomous mode or even stop driving, i.e. park.
3. *Self*: Finally, the car must be robust against failure of self, such as its sensors and subsystems. This can be achieved through redundancy based decision making.

3.3.5 Integrity (curation)

Integrity of the data strongly impacts the decision processes on which they are based. Integrity problems include physical issues, such as sensor or power failures, and logical issues, such as software bugs, human input errors, and conversion errors. In the context of the car example, the data can become purposely untruthful due to irresponsible peers or security issues as mentioned in previous sections. Integrity issues can potentially be detected at all levels.

1. *Central*: By combining the information from all reporting cars, redundancy can help in detecting integrity issues that arise as inconsistencies. For instance, a more precise location of an accident can be derived.
2. *Peers*: Also peers could help in detecting and curating data integrity issues. Without prior suspicion, the required overhead can hardly be justified. Clearly, this depends on the type of application and the imposed safety and security levels.
3. *Self*: For an autonomous SoS such as a car, curating its collected data must be accomplished by its own measures. For safety reasons, redundancy of data sources is a robust measure.

Clearly, the developments in hardware, software, and AI have shown the trends and opportunities ahead. Especially for the security domain, applications should be ‘responsible by design’ in order not to jeopardize their trust and acceptability. Sharing information should therefore be limited to well-considered sharing levels along the lines of the proposed five invariants.

4. Information Strategies

4. Information Strategies

Large databases are available everywhere. Both public and private organisations are collecting enormous amounts of data. The market value of many companies is even assessed by the amount of data they have collected. However, as was explained in the previous chapters, there is an important difference between large databases and Big Data. Furthermore, collecting more data does not always yield more value. For instance, in the security domain, more data does not automatically result in more security. The big question of Big Data is how to leverage on the Big Data phenomenon.
Big Data offer a plethora of opportunities, but realizing these opportunities is not that straightforward. Leveraging on the Big Data phenomenon calls for the right information strategies. Only with the right approach and by making the right choices will Big Data deliver its promises.
Extracting added value from Big Data requires meeting at least three conditions. First, a thorough understanding of the Big Data phenomenon is required, enabling a proper assessment of what Big Data is and what results can be expected from Big Data analyses. Second, it requires adequate data management, combining the right data and using the right tools for analyses, enabling meaningful analyses and yielding beneficial results. Third, it is important to take a responsible approach in which the risks that may arise in Big Data processing are properly managed, and preventing issues regarding aspects such as discrimination, privacy, integrity and justice.

In this chapter, these three conditions (understanding, management and approach) are dealt with in more detail, thus providing the necessary elements for an information strategy to leverage on the promises of Big Data. The section on new perspectives explains that the developments concerning Big Data call for new perspectives and new approaches. Classical approaches of posing a question and collecting data to answer that question are no longer the main approach. Rather, in the era of Big Data this is the other way around: interesting and useful knowledge may be hidden in the data, but the main issue is to disclose such knowledge. The section on data floods explains how to deal with the data flood with new types of data management and data analyses. The section on responsible innovation focuses on mapping several risks related to Big Data analyses and offering ways to deal with these risks.

4.1 New perspectives

Collecting large amounts of data is common practice in most public and private organisations. Companies collect personal data on customers, suppliers and employees. Furthermore, they collect data on their business processes, including product information, payment information and shipping information. Public organisations collect similar data on their target groups, employees and business processes. People are generating large amounts of data about themselves (for instance, on social media), but are also increasingly being recorded via sensors (such as cameras and microphones), trackers (such as RFID tags and web surfing behaviour) and other devices (like their mobile phones and wearables for self-surveillance/quantified self). With the introduction of the `Internet of Things', the datafication of our society is increasing exponentially. Before, when a user was communicating (exchanging data) with three separate devices, this resulted in three data flows, but nowadays, when these devices are also communicating with each other, this results in six data flows.

The vast volumes of these data are the main reason why they are called Big Data. However, it is not only size that defines Big Data. Also the velocity of the data (most of these data are generated in real-time and sometimes it is streaming data that is not recorded or stored anywhere) and the variety of the data (Big Data may be unstructured and in different formats like text, numbers, images, sound, etc.) are often-mentioned defining aspects of Big Data (Laney, 2001).[xx]

These large volumes of fast and unstructured data provide three new perspectives. The first is that of a datadriven approach rather than a hypothesisdriven approach. The second is that of a new role for human intuition, a role that is more focused on guiding the process than performing data analyses. The third is that of looking for correlations rather than causal relationships. Big Data can only provide statistical relationships, not causal relationships. However, in many decisionmaking situations, underlying causality (if any) does not have to be known.

Data are a set of facts and Big Data are large sets of facts. The trick is to transform data into knowledge, i.e., to transform sets of facts into patterns that are interesting, reliable and meaningful for users to base their decision making on. The classic approach to distil knowledge from data is to formulate a hypothesis and to collect data to test the hypothesis, in order to accept or reject it.[xxi] In the era of Big Data, in which data are relatively easily available, this approach may be less timeconsuming than it was in the past. However, the exponential growth of data also enables another, completely different approach, which is usually referred to as a data-driven approach. The large amounts of data that are available may hide many interesting patterns and relationships. The issue is no longer tocarefully draft single hypotheses, but rather to pose the right general questions and to discover the knowledge already hidden in the data.[xxii]

Hence, Big Data calls for a different approach for which it is important to look at what the data are telling rather than looking for specific answers to specific questions. However, that is easier said than done, as Big Data can be large, fast and unstructured. Therefore, it is difficult for human beings to use their intuition to get an overview of the data available and to distil useful knowledge from Big Data. It is for this reason that tools for Big Data analytics are needed (see the next section).

Although human intuition may be difficult to use for analysing Big Data, this does not mean that human intuition should be replaced by machine learning and automated analyses altogether. Since Big Data allows for endless amounts of different approaches, it is important to include human intuition to guide the process of knowledge discovery. This may seem contradictory to the data-driven approach, but it is not: it is important to strike the balance between analyses that are neither too narrow nor too broad. Too narrow would be focusing too much on specific questions. Although this may yield specific answers, it also implies no longer listening to what the data are telling and may, therefore, not yield completely novel insights (van den Herik, 2016). Too broad would be using all kinds of data analyses tools to look for patterns, which may indeed yield many patterns, but many of them would not be interesting or novel. In essence, this is comparable to formulating search strings on your web browser: if Google, Yahoo or any other browser shows a million results this is often as useless as zero search results. The art is to produce something between five and ten relevant results.
In empirical statistical research, the focus is often on causal relationships, but Big Data analyses cannot do this – Big Data analyses look for statistical relationships. These relationships are not necessarily causal or may be causal without being understood, for instance, because they are indirect (see Figure 7). Statistical relationships may be used as a starting point to further investigate and perhaps discover underlying causality (Mayer-Schönberger and Cukier, 2013), but it is important to note that merely statistical relations may already be sufficient to act upon. For instance, when there is a statistical relation between crime and zip codes, this may be a reason to increase police surveillance in these areas, even when the underlying causality of this pattern (if any) is known.

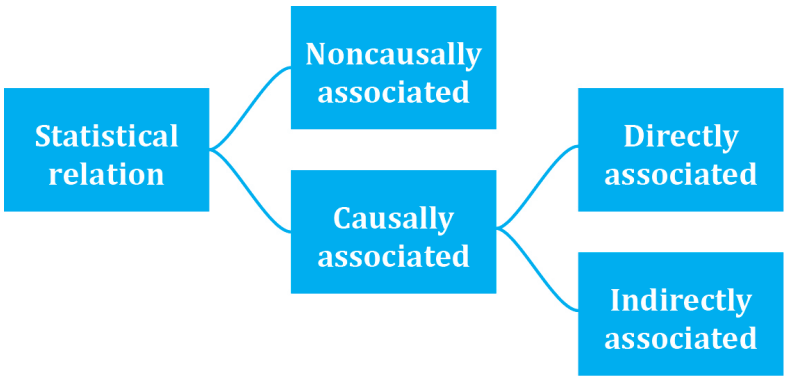


Figure 7 *Different types of relations*[xxiii]

4.2 Dealing with data floods

The potential value of Big Data is unlocked only when leveraged to drive decision making (Gandomi and Haider, 2015). To enable such evidence-based decision making, efficient processes are needed to turn high volumes of data, which are often real-time and very diverse, into meaningful insights. This requires both adequate data management (dealing with technologies and processes to collect, store and prepare data for analyses) and technologies for analyses (dealing with technologies to analyse data and extract added value from Big Data). It is beyond the scope of this report to discuss Big Data management and analysis technologies,[xxiv] but this section examines several relevant aspects to take into account when considering information strategies.

Data reuse
When it comes to data collection and aggregation, the focus is often on data generated on social media (Facebook, Twitter, etc.) and data generated by the Internet of Things, including mobile phones, sensors, trackers and wearables. However it is important to note that the large volumes of data needed to leverage on the benefits of Big Data can to some extent also be established by the reuse of existing data, a source that is sometimes overlooked. From the perspective of organisations processing data, three different types of data reuse can be distinguished, depending on whether the purposes or the context or both differ from the original purpose and context for which data were collected and used (Custers and Ursic, 2016):

- *Data recycling* - using data several times for the same purpose in the same context;
- *Data repurposing* - using data for different purposes than for which they were initially collected, but still in the same context as the original purpose;
- *Data recontextualisation* - using data in another context than in which they were initially collected.

For proper data management, it is essential to use technologies that can deal with large amounts of real-time data. This involves large processing capabilities and, in case the data need to be stored, large data warehouses. However, large parts of Big Data are streaming data that are no longer static but are rather dynamic in nature. These data are not always recorded and stored.

When data need to be stored, this can be done either in centralized databases or in decentralized databases. Note that this distinction refers to the accessibility of the data rather than the locations in which the data are stored. For instance, data can easily be stored in the cloud. This can be a cheap solution for organisations to outsource their data storage. When storing the data in the cloud, it may be stored on different servers in many locations. However, from the perspective of a user, this technical process does not hinder the accessibility of the data and or affect options for data analyses and on the screens of users it will appear as a centralized database.

If the goal is to discover novelties in the data, it may be useful to combine different sources of data. When combining unexpected sources of data, this may yield unexpected results. Therefore, it may be interesting to cooperate with organisations that have completely different datasets. A typical example is the Netherlands Ministry of Security and Justice that began a cooperation with the Royal Netherlands Meteorological Institute (KNMI), to see if there are any relations between crime and weather. Other combinations, like crime and neuro sciences, crime and gaming, and crime and nutrition may also yield promising new insights.

Technologies for Big Data analyses are a group of technologies that enable extracting knowledge from large, fast and unstructured data. Usually these technologies try to reveal patterns in the data. Often the techniques are mathematical algorithms that try to search for patterns in an automated way, in which case they are referred to as data mining technologies. Sometimes the term machine learning is used as well. The technologies can be distinguished according to their workings and the type of patterns they yield accordingly. The most often used are regression, classification and clustering techniques. Regression techniques look for linear and non-linear relations between data.

Classification techniques (most notably decision trees) try to categorize data or subjects in several classes. Clustering techniques (such as hierarchical clustering, k-means clustering and neural network clustering) try to identify different clusters in the data. The main difference between clustering and classification is that clusters may show some extent of overlap, whereas classification uses mutually exclusive classes.

The technologies for analysing Big Data can also be used to create risk profiles. In the security domain, risk profiling may be interesting to assess risks, for instance, regarding which people are likely to commit particular crimes, where and when incidents may occur, and which people and objects are most likely to be victims. This is sometimes referred to as predictive policing. Risk profiling may be used to prevent incidents from happening and to redirect law enforcement capacity to times and places where it is most needed.

4.3 Responsible Innovation

The use of Big Data, particularly the use of data mining and profiling technologies, may yield results that are undesirable from an ethical or societal perspective or illegitimate from a legal perspective. This section discusses some of the risks to take into account when analysing Big Data with automated data analytics. These risks can be categorized according to the five invariants mentioned in Chapter 2 and should be incorporated in any and all information strategies employed in the future.

Starting with the first invariant (ethical issues/responsibility), there are several concerns that can be identified. A major concern is that of discrimination. Discrimination may arise when data collection is biased, but even when data are not biased by those who collect it, it may still be the case that discriminating patterns arise from the data analyses (Custers et al, 2013). For instance, particular attributes may appear in risk profiles that are not acceptable or even violating antidiscrimination laws, particularly when these criteria are used for decision making. This may concern particularly sensitive attributes like religion, political preferences, sexual preferences, criminal records and gender. Research has shown that even when these sensitive attributes are not included in the datasets, they may appear by proxy (Calders et al, 2013). A typical example of such indirectly discriminating profiling is socalled redlining, in which characteristics are ascribed to people on the basis of their zip codes, whereas zip codes may be a strong indicator for someone's ethnic background. There are discrimination aware data mining tools that can be used to avoid these discrimination issues (Zliobaite and Custers, 2016).

Even when analysing Big Data that does not result in illegitimate discrimination, it may still result in patterns that become 'public knowledge' and may result in stigmatization of particular groups. For instance, when specific minorities are overrepresented in criminal records (assume they make up 2 % in criminal records and only 1 % in society), this may lead to stigmatization in which some people might conclude that members of these minorities are much more prone to criminal behavior, even though only a very small percentage of the members of these groups actually have a criminal record.

The second invariant (privacy/reciprocity) also involves risks to take into account. Big Data analyses may predict attributes of people that they may not want to disclose (Custers, 2012). For instance, Kosinski et al. (2013) show that, based on Facebook likes for movies, music, games, comments, etc., reliable predictions can be made about a person's gender, ethnic background, sexual orientation, religion, happiness, substance abuse, parental divorce, intelligence, etc. Furthermore, Big Data analyses may even predict attributes of people that they do not even know, such as their life expectancy, their risk to attract cancer, etc. People may not want to know this information.

From a legal perspective, in some cases consent may be required for data processing. Privacy-preserving data mining tools exist that can address several of these issues. Also methods like privacy impact assessments (Wright and De Hert, 2012) and privacy by design (Cavoukian, 2009) may be helpful in preventing or addressing privacy issues in early stages.

Compliance with privacy and data protection legislation has also become increasingly important for organisations. With the introduction of the EU General Data Protection Regulation (GDPR), which will apply throughout the European Union as of May 2018, organisations that do not comply with this legislation can expect heavy fines, of up to 20 million Euro or, in the case of an undertaking, up to 4% of the total worldwide annual turnover (whichever is higher). This legislation transforms privacy risks for data subjects into financial risks for corporations and governments.

The third invariant (security/adaptivity) regards data breaches. These may be internal or external and they may be intentional or non-intentional. Obviously security safeguards are required not only for the Big Data itself, but also for the tools to analyze the data and for the knowledge resulting from Big Data analytics. Anonymization and pseudonymisation may be useful tools to protect security and privacy. Furthermore, access controls and the use of cryptography may be useful measures to incorporate in this respect.

The fourth invariant (safety/autonomy) can also give rise to pressing concerns. The use of large amounts of data and advanced data analyses technologies may significantly reduce the transparency of decision making. Solove (2004) has indicated that this may lead to Kafka-like problems, in which people are being confronted with decisions that are hard to challenge and with a lack of transparency on the decision making process. It is recommended (and to some extent legally mandatory) to provide transparency about which data is collected, for which purposes and how the data is processed. This will provide data subjects with more autonomy to decide to which forms of processing of data they consent (so-called informational self-determination) and may improve trust in data controllers and processors.

The fifth invariant (integrity/curation) relates to both the Big Data itself and to the way in which the results of Big Data analytics are dealt with. A risk is that of self-fulfilling prophecies. Datasets may be biased and contain hidden prejudices. For instance, when police surveillance would take place in neighbourhoods with ethnic minorities only, it is not surprising that police databases get filled with data on ethnic minorities. When profiling is based on such datasets, the result may be that these ethnic minorities would constitute a higher risk, i.e., according to the data they would be more prone to committing crimes. When the police would use such risk profiles to increase the focus of their surveillance in these neighbourhoods, this would complete the circle. It is recommended to check for any bias in the data before starting the analyses.

Another integrity risk is related to the fact that it is not allowed (from both an ethical and a legal perspective) to subject human beings to fully automated decision-making.[xxvxxvi] Therefore, it is important to include in all Big Data processing and analyses one or more stages in which human beings are involved to check for integrity and other issues. This will also avoid inaccuracies and may increase trust.

5. How to organise?

5. How to organise?

The possibilities of using data for both organisations in the public domain responsible for security and for individuals taking care of their own security[xxvii] are growing rapidly. More and more data becomes available every day, and sophisticated tools that can analyse them are developed at a similarly-rapid pace. These often come in the form of free versions, downloadable for whoever shows an interest in them. Even the platforms, integrated in the Internet of Things on which these apps function, are for free or are available at low cost. It is interesting that organisations barely need to collect data anymore because many data are produced anyway, on a (semi)automated basis.[xxviii] Data is collected because it is considered to be valuable per se (or qualitate qua). It can answer thousands of questions as a result. As discussed in the previous chapters, the collection of data is driving the formulation of questions and hypotheses, and allows for new knowledge to be discovered. So the capabilities for both organisations and individuals to act upon a perceived risk are depending on three fronts: the available data, the quality of the analysis of the data, and the capability to act on it. Depending on the interest of the organisation or the individual, data become valuable in a specific context. Connecting them for that purpose makes them useful and potentially actionable.

Table 3 Data availability, Power of analysis and decision making capability[xxix]

Maturity	Data availability on a situation	Situational awareness	Situational Understanding Capability to informed action
Low	**Poor**	**Poor**	**Poor**
Medium	**Average**	**Average**	**Average**
High	**Full**	**Full**	**Full**

The higher the maturity level in each of the columns in Table 3 are, the better organisations and individuals will be capable of taking care of the risks they might encounter. To reach full situational understanding requires both full data availability and full situational awareness. This is where the developments of Big Data should focus on.

It will be challenging to enable poorly educated people to make use of these possibilities (simple and easy to use with practical action perspectives) to overcome misperceptions regarding risks (relative seriousness of the risks for the organisation and/or individual). More awareness of the risks does not necessarily enhance the individual's feeling that he or she is safe. So full situational understanding should include both actual information placed within its historical context, and how it can be projected on or applied to the actual case of the individual. This would hopefully yield more realistic predictions, but should also be combined with the current capabilities of the individual to avoid the situation, or minimize its consequences.

Big data against Subversive crime

Local and national governments have found it challenging to deal with organised crime because of the smart mix of legal and illegal activities that have insidiously corrupted the networks of local shop owners, car rentals, and other businesses. Criminals use their laundered profits for investments in legal businesses too. The complexity of combining financial, social and technical and/or administrative signals in modern subversive crime makes it an excellent case for Big Data analysis. However a purpose limitation should be used for the data to be usable for this case.

A simplified example to explain in practice what full situational understanding could mean is the app 'Buienalarm'.[xxx] This is a rain forecasting application. The app is geolocation specific, up to some hundred meters, assessing the millimetres of rain per hour with a precision of five minutes. It uses the location information of the telephone, combines that with weather forecasting models and data from weather stations and forecasts up to two hours in advance.

The only feature that is missing is telling the user to take an umbrella or stay inside when even an umbrella won't help you to stay dry. And it could maybe be enhanced using both your current location as well as the place you want to go, telling you where, when and how much rain you might expect and take precautionary measures for.

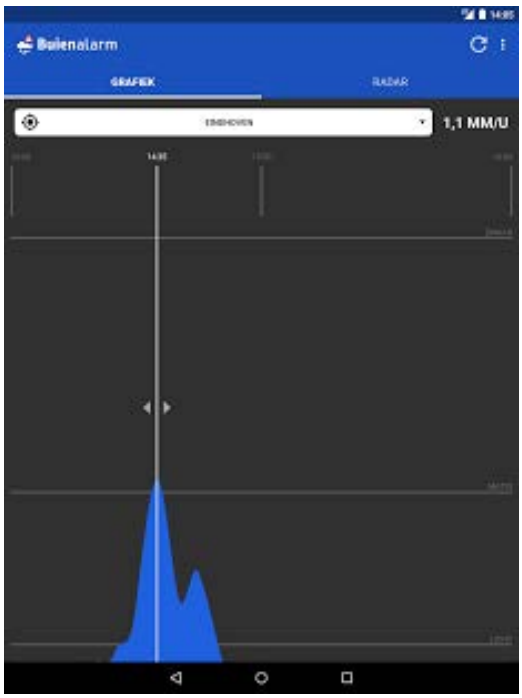


Figure 8 Screenshot of 'Buienalarm'

Security of Public Space Events

These technologies can be potentially interesting and useful for public security authorities policing events such as festivals. By combining datasets of car and train traffic approaching the event location, weather conditions forecasting public behaviour during rain and heat, social media sentiment and geolocation data analysis and other datasets, it is possible to create better situational awareness. For public security services, crowd control options, and information services for the public at the event. However a purpose limitation should be used for the data to be usable for this case.

5.1 Possible consequences for Roles, Responsibilities and Required means

The above general analytical framework classifying full data availability, awareness and ultimately full situational understanding helps in defining the roles, responsibilities and required means of organisations as well as those of individuals. Depending on the legal status of the user of the data (roles) or the purpose of the data usage (personal, commercial, governmental, security, forensics, research etc.) the responsibilities can differ. However they may also differ due to economic reasons regarding efficiency or efficacy. Some (combinations of) data are not allowed to be owned or operated by public organisations or individuals, other data are only permitted to be processed under strict conditions (e.g. purpose limitation, mandates, stakeholder involvement) by these same organisations or individuals. Some data are being produced and analysed already and usage of those data is more efficient and effective than developing one's own data sets. When looking at the total amount of available data, we can see that the usage of data is restricted by the legal status of the entity that has an interest in them. Furthermore, all organisations have limitations with regard to the resources available to them for their analyses (means).

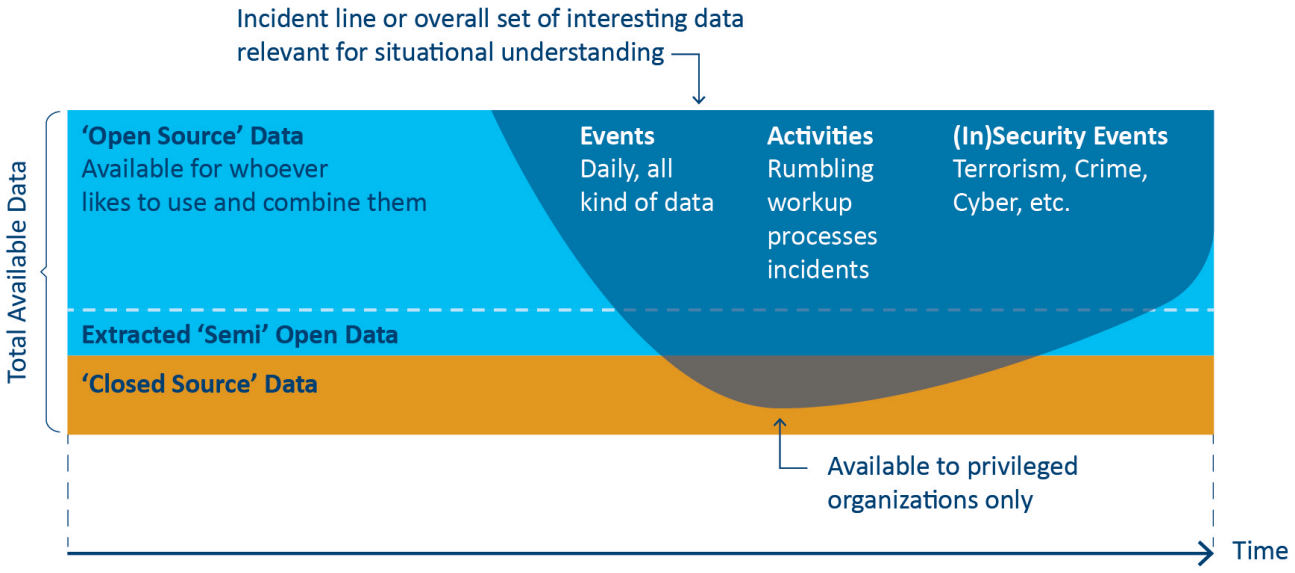


Figure 9 Inspired by the report *Denk- en Datamodel Suspicious Signs*[xxxi]

Figure 9 shows this concept. While Big Data are largely publically-available, some data are only available for privileged organisations or individuals. The combination of all these data (the shaded space in Figure 9) can create full data availability and awareness. It would help organisation efforts to focus on those capabilities that would make a difference by being more efficient and effective.

So in combining this concept for the most effective and efficient data gathering, analysis (awareness) and usage (understanding) we should be able to derive organisational requirements that are needed to optimize the roles, responsibilities and means for security for both organisations and individuals.

5.2 An example of criminal behaviour

An important case showing the subordinate position of the weights of the invariants, ethical decisions and privacy, is in investigating criminal behaviour. We provide an example from the marathon bombing in Boston (see figure 10). On April 15, 2013 an unexpected bomb exploded at the end of the Boston Marathon and only 12 seconds later a second blast took place. There were three dead people and 264 injured persons. Of course, sensor images and recorded pictures were available. However, the crowd was so large that no criminals involved in the bombing could be identified upon initial inspection. After two days the police handed over the case to the FBI, who quarantined the city and attempted to find the criminals in their computer databases. They could not identify them, since they missed clues that could have been used as search heuristics. When two criminals tried to escape from the quarantined area, one was killed by the police force (Tamerlan Tsjarnajev) in a gunfight, while the other managed to escape. One day later Dzhokhar Tsjarnajev was arrested albeit by a coincidence through a blood trace observed by an alert citizen. What could have happened is that Dzhokhar had been arrested with the help of narrative science.

A proper investigation of such a criminal bombing behaviour with modern technological means would have required a combination of Big Data, High Performance Computing, and Deep Learning. These three components form the basis of *narrative science* (cf. LeCun, Bengio, and Hinton, 2015). In narrative science, an intelligent program constructs a possible/real story and points to the criminals.

This approach is also known under the name of storytelling. For lawyers it is called argumentation theory.

Table 4: Future Measurements (A shift from concepts to techniques)

Five invariants	Measurements in 2016 with the help of…	Measurements in 2016 with the help of…
Ethical Issues	Responsibility	Deep Learning
Privacy	Reciprocity	Deep Network Understanding
Security	Adaptivity	Quantum Computing
Safety	Autonomy	Independency
Integrity	Curation	Transparency

For our specific case the facts are as follows. The FBI was looking for suspects from abroad. They intended to compare the foreign persons who had entered the United States in the last three years with the recorded sensor observations. Obviously, that was an impossible task considering the number that had entered the US was not manageable. The prevailing question was then: how can this number be constrained? The bombing itself turned out to be caused by a pressure cooker with needles and nails. Later on the FBI also found the lid of the pressure cooker. These developments gave potential clues on the origin of the culprits. Bombing with pressure cookers was known since 2004. The idea was developed in Chechenya and Ingushetia, and applied in Afghanistan and Pakistan. The Israeli journalist Elizabeth Tsurkov described the working as follows: "The pressure cooker increases the pressure of the explosion [...] and because the metal casings become deadly fragmentation upon detonation." With this information the FBI would have been able to search more effectively in the databases for potential culprits. In this instance, taking this narrowed range of possible culprits into consideration when inspecting the sensor observations would have made it more possible to identify the perpetrators. The real story has many interesting sub-story branches, such as the fact that the FBI had a co-operation with the Hinton-team. Obviously, when searching intensively for criminals ethical decisions and privacy aspects of other persons are fully neglected.

Figure 10 The Boston bombers

5.3 Governmental and Private Security Organisations

Increasingly large amounts of data that are of interest to public organisations for security can be found in open sources. The free flow of data will show exponential growth in the near and more distant future. The most important data for these organisations will be those data that are not publicly available, as it is likely that the sources used to collect them are reserved for their usage only. This means they would have a monopoly on them. The fact is however, that the availability of these specific datasets will grow much slower than the freely available data. The quality might be satisfactory, but in the long term the analysis teams will not be able to compete with the worldwide available data, and analytical capacities and quality and might lose their significance as a result. Given the foreseen future, the possibilities of these developments in data and analytics for these organisations are still growing. Future applications of data analytics for these organisations will however need to focus more on the quality of their privileged data and analytical tooling on the one hand and in combining these results with freely available open source data on the other. Because organisations have by definition limited analytical resources, it would be efficient if they would be capable of collecting data from others almost instantaneously and, when necessary, in a specific context. That will expand their analytical capabilities both in terms of quantity as well as in quality because of the richness and diversity of sources. At the same time analysts would be more capable of making effective and efficient usage of the resources required. The analysts at these organisations have to become more generalists, capable of combining data instead of being experts only on a few topics.

We already see examples of this approach that could be classified as crowd-based sourcing and analytics. Police forces in the Netherlands use applications and communication strategies to have data collected, analysed and interpreted using the general public. 'Opsporing verzocht' is a Dutch television show and a web platform[xxxii] in which the Police asks the general public (the viewership) to help deliver extra information, identify people, give context to a situation etc. in order to more easily solve crimes. Europol is another example of how the general public can support in catching criminals.[xxxiii] In other occasions, like festivals, social media based data produced by the crowd attending e.g. a festival, can be analysed for crowd control or other related processes.

The fact that these developments are currently taking place will most likely also trigger the formation of new organisations that will find business cases, e.g. business analytics as a service, that are more efficient than the traditional ones we currently see.

5.4 Security by and for Individuals

Nowadays people are better informed than ever before. Data can be accessed almost instantaneously. We have smartphones, computers and more and more wearables producing and interpreting data for us in actionable ways. Applications are developed for specific contexts and make it possible for us to be more aware and provide more options for our security than ever before. We have apps for bad weather forecasts like rain or hurricanes, transportation advice avoiding traffic jams and incidents, and dangerous or risky places, etc. We are also capable of producing our own data and sharing them with others to enhance security in new ways, using specific applications for personal surveillance or alarm systems integrated in the Internet of Things, combining different data to enhance the awareness and precision of alarms for ourselves or close-by environments with less false positives and negatives.

We also see people set up their own analytical platforms for specific contexts. A new and important example is the platform of individual journalists (Bellingcat[xxxiv]) that have carried out research on the context, and possible offenders of the shooting-down of passenger airplane MH-17 above the Russian occupied part of the Ukraine in 2014. Another example is the data analysis produced by a 20 years old Dutch citizen[xxxv] on the developments in the Middle East and more specifically on ISIS. Similarly, Burgernet[xxxvi] is a crowdbased solution for collecting and distributing data on incidents and robberies, missing persons etc., that currently has 1.6M users in the Netherlands.

The empowerment of individuals will grow even further. In the future, developers and researchers will not only rely on data provided by official governmental organisations. It may in fact be possible to be more quickly, and maybe even better, informed via the data that crowds have gathered, interpreted and falsified or enhanced. And in that way the situational understanding support by Big Data will become mass individualised stimulating the growth of new opportunities for organising security and safety and strongly enhancing selfreliance of citizens. We should however be cautious with those individuals that will not be able to cope with these developments causing the risk of having a split society where people, depending on their level of education will be better or worse off.

5.5 Consequences for Society and Knowledge Institutes

In general, it was traditionally governmental organisations that were handed the roles, responsibilities and means to take care of the safety and security of its citizens and the state as a whole. However the concept that the state will take care of all aspects of the safety and security of its citizens has gradually eroded. This trend is strengthened by the fact that society is growing more complex, intertwined and more and more in a constant flux, for which anticipating on the vulnerabilities is difficult. This is also due to economic reasons, as it has become too expensive to guarantee full safety and security. Some parts of the security arena are - and in the short term will of course stay - the sole responsibility of governments (defence, external security) but at the same time the notion that citizens should be more self-reliant is developing. A government cannot guarantee full safety and security in all its aspects for the state, its organisations and citizens. So people are encouraged to develop their own capabilities to be self-reliant.

Regarding Big Data analysis (awareness) and the possibilities for organisations and individuals to act in a given situation (understanding) future developments are necessary to fully make use of its possibilities. Knowledge organisations should play a role in supporting the developments mapped in Table 3 , especially focusing on situational awareness and situational understanding enhancement regarding data use under conditions as explained in chapter 2.

5.6 Pokémon Go...for Security, year 2020

Developments related to the usage of Big Data for security are proceeding at a rapid pace. Yet it remains challenging to imagine the full possibilities they might bring. But like in all future-oriented analysis it often helps to use a scenario to describe a plausible set of developments.

New York Times, May 15, 2020

SecPok introduces Gamification of Security

After its first commercial success to apply virtual reality in an entertaining game on a smartphone by Niantic[1] in 2017 the company SecPok started applying the insights gained with *Pokémon Go* for security applications. It is a new and revolutionizing application of Big Data for security.

The company SecPok accelerated the new opportunities of Big Data and computer power with the capacities of end-users with local knowledge and expertise in a game, enhancing local security. Applying computer power for Big Data collection and analytics, the company SecPok developed and released *Secmon Gone* in 2019. This game is an application people can play to search for crime spots and solve crimes by using their presence to prevent criminals from acting. Gamification of real-life security has become a fact since its release. Instead of applying the game to search for Pokémon this new game challenges people to distribute their knowledge of locations they know or suspect to be areas in which criminal activities often take place. In combining this knowledge with constant updated information of people looking for arriving to these places where possible criminal activities take place, the communities have been helping authorities tremendously in the understanding of crime hotspots. At the same time police and security organizations use this application too. The uploading of new areas by both security organizations and individuals alike has resulted in a continual, mutually-beneficial process for both parties involved. Unexpected developments have been observed on the basis of the data produced and gathered worldwide, showing combinations of criminal behavior in networks and places that were previously unknown. The new crime rate numbers in early 2020 in downtown New York City show an amazing drop. Rumor has it that this game *Secmon Gone* is just the first application in a long series of new security games that make it possible for people like you and me to help out on security and safety in cooperation with police to make our world a better place. It is an exciting development turning us all into crime fighters as we have fun playing the game.

Your reporter Stephan Ecure

In this so-called *Secmon Gone* game, many different Big Data applications are combined. Using geolocations of users and having them take pictures and upload them produces a very detailed situational understanding of possible crime hotspots. Asking the users to classify what they see and articulate what they think enhances intelligent machine learning, and is a useful way of preventing and overcoming traditional security barriers. Using these data, municipalities are able to adjust their infrastructure plans, local police can better allocate capacities, crime analysis teams can better understand the phenomena involved, and prosecutors can better prosecute criminal networks. Having the characteristics of the crime spots, the criminals and the solutions analysed by computers, it is now even possible to do forecasting on criminal developments. It might possibly have its downside too. Users have been uploading pictures of people, falsely stating that they are involved in a criminal activity at the crime spot captured in *Secmon Gone*, and making false accusations. Had the company Secpok used the testing framework with the five invariants and its 'central', 'peers' and 'self' scope, they could have prevented that and possibly overcome other flaws as well (see table 5).

[2] https://en.wikipedia.org/wiki/Pokémon_Go

Table 5 The invariants applied to Secmon Gone

Invariants	Measured	*Secmon Gone*
Ethical Issues	Responsibility	From the ethical point of view the responsibility of the users (self) could by enforced by having end-users be able to rate the quality of players (peers) and having an ethical board maintaining (central) oversight of the results of the game and the behaviour of its users in both the front end (gamers) and the back end (police, municipalities and others).
Privacy	Reciprocity	The possibilities for end-users (self) to play or facilitate the game came with the perk of having a meaningful contribution to the society, taking care of your community's safety and security yourself (peer) by helping out responsible authorities (central). For all stakeholders this was seen as a balanced situation between gains (responsibility and fun) and privacy.
Security	Adaptivity	Security of the game software and Big Data storage was taken into account from the beginning onward (central) and ensured to protect the system as well as possible against intended damage. And for users (self and peers) the quality of service is guaranteed.
Safety	Autonomy	Because of continuity of the game availability and data integrity (self and peers), autonomy of the platform was guaranteed by the development (central) of fall-back scenarios for the software and automated software repair facilities built into the different ICT-servers on which the game application ran.
Integrity	Curation	Because of the enormous amounts of data gathered and combined, the possibilities for curation of the data quality were well thought out (central). The possibilities using new forms of curation of data integrity were applied, producing an unparalleled level of quality of data (self and peers) and as a consequence high quality of results obtained by the game (self and peers) as well as the analytical results (central).

5.6 The Future

The future can only be predicted with adequate accuracy when we have a thorough knowledge of the history (see, e.g., Van den Herik, 1991; Hamburg, 2007; Christensen, 2009, Adriaanse, 2015) and the current scientific development (cf. Van de Voort, Pieters, and Consoli, 2015). In broad lines we see two developments: (1) the dependency on intelligent computer programs is increasing; this holds for government, industry, companies, banking and ordinary people; (2) the trust in the outcome of computers is diminishing; causes are hacking, fishing, internal interest setting (libor interest), and pollution setting (automobiles). The two curves are shown in Figure 11.

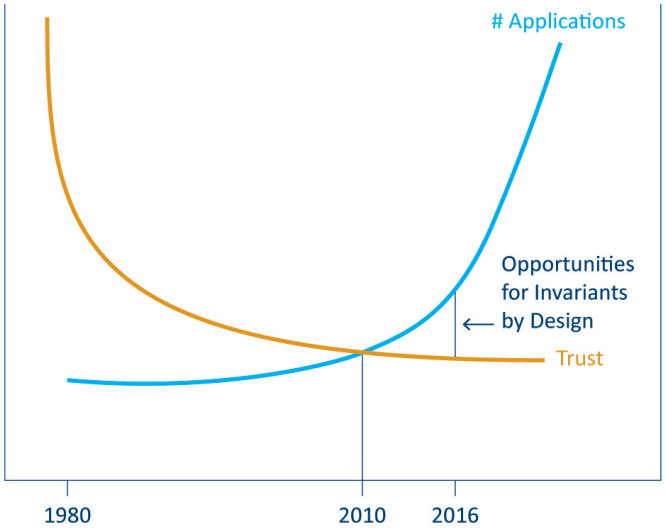


Figure 11 Dependency of applications versus trust in the outcome of applications.

As we see from Figure 11 the curves are divergent after 2010. We marked the distance between them for the year 2016 and argue that the discrepancy can be filled with techniques that incorporate their "goal" by design, such as privacy by design, security by design, safety by design and integrity by design.

Here contention will emerge between the legal worlds of different nations (by their different norms) and their situational awareness (by their different cultures).
A clear point of research is the development of the five invariants in relation to the trust in the outcome of the applications.

Let us take privacy as an example. Is it possible to redirect privacy by design in such a way that the results in a convergence between trust in privacy and the development of the application converge (see Figure 12)?

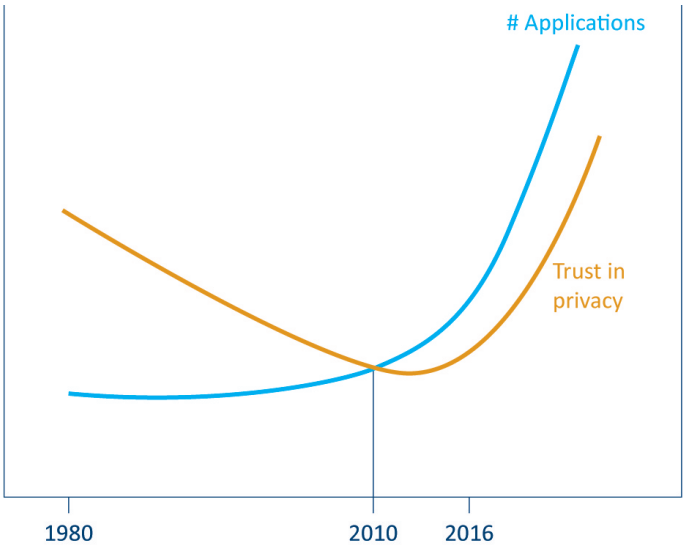


Figure 12 Is convergence between privacy and the development of applications possible?

Similar questions will hold for the other four invariants. Thorough analysis will reveal that all five invariants have a different relation to the development of the applications. Future research will show which invariant has the toughest relation with the new applications. Our conjecture will be that security is high on the list.

6. Conclusions, Recommendations and Observations

6. Conclusions, Recommendations and Observations

6.1 Conclusions

- Big Data for security will undoubtedly be of high interest and value for society to use, both now and in the near future. Applications for event security and subversive crime as examples seem to be very promising. The usability will increase when Big Data is not only used for enhanced situational awareness but also for better situational understanding.
- The governance of the different roles, responsibilities and resource distribution in security organisations needs rethinking regarding data gathering and how analytical processes are organised.
- Big Data in security applications should adhere to standards and principles for responsible innovation.
- Handling Big Data with AI brings forward many ethical issues. This needs to be balanced by an expert ethical board. It should be considered for smaller organisations to have a centralised board to pull resources and have good quality capacity available.
- Knowledge institutions should focus part of their research on developing standards for responsible design, applicable to the domain of Big Data and security.
- It is important to embed ethical behaviour material in the earliest level of education that students receive.
- The importance of accuracy of machine learning and AI can not be underestimated while the impact of false positives is disastrous to those suffering from the consequences. Understanding the associated statistics is a top priority for the results that are to be applied in decision support systems.

6.2 Recommendations

Recent developments in Big Data and AI technology show that many tasks that have traditionally been considered the province of humans can be competitively accomplished by machines. So, be ambitious. To keep the delicate balance between Big Data and AI , two safeguards are possible: to *diminish* attention and research efforts for Big Data and Deep Learning; or to *increase* attention and research efforts for Big Data and Deep Learning. From these two safeguards our preference goes to the second safeguard, provided that we are allowed to introduce three further, more specific safeguards:

- Increase research on AI systems for Big Data and Deep Learning with emphasis on moral constraints.
- Increase research on AI systems for Big Data and Deep Learning with emphasis on the prevention of AI systems to be hacked.
- Establish (a) a committee of Data Authorities and (b) an ethical committee.

When developing products and applications that use Big Data and AI, use the proposed framework of invariants.
To develop applications responsible by design, sharing information should be limited between peer applications and peer users where possible.

6.3 Observations

After the implementation of these recommendations we may conclude that within two waves of disruptive developments (each taking, say, 25 years) computers will be at a par with, or even better at taking ethical decisions than human beings.

Big Data supplements rather than replaces human intuition.

References

Adriaans, P. and Zantinge, D. (1996). *Data mining*, Boston, MA: Addison Wesley.

Adriaanse, J. (2015). *Ethical challenges in turnaround management.* Lecture at the Workshop Leadership Challenges with Big Data. Turning Data into Business, Erasmus University Rotterdam, June 30.

Calders, T., and Custers, B.H.M. (2013). What is data mining and how does it work?, in: B.H.M. Custers et al. (eds.), *Discrimination and privacy in the information society*, Heidelberg: Springer 2013, p. 27 42.

Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X. (2013). Controlling attribute effect in linear regression. In: *Proceedings of 13th IEEE ICDM*, p. 71–80.

Campbell-Kelly, M. (1987). Data communications at the National Physical Laboratory (1965-1975). Annals of the History of Computing, 9(3), 221–247. https://doi.org/10.1109/MAHC.1987.10023

Candan, K.S., and Sapino, M.L. (2010). *Data management for multimedia retrieval*, Cambridge: Cambridge University Press.

Chinchalkar, S. (1996). An Upper Bound for the Number of Reachable Positions. *ICCA Journal*, Vol. 19, No. 3, pp. 181183.

Christensen, B. (2009). Can Robots Make Ethical Decisions?

Custers, B.H.M. (2012). Predicting Data that People Refuse to Disclose; How Data Mining Predictions Challenge Informational Self-Determination, *Privacy Observatory Magazine*, Issue 3. See http://www.privacyobservatory.org/

Custers, B.H.M., and Ursic, H. (2016). Big Data and Data Reuse; A Taxonomy of Data Reuse for Balancing Big Data Benefits and Personal Data Protection, *International Data Privacy Law*, pp. 1 12. doi: 10.1093/idpl/ipv028.

Custers, B.H.M., Calders, T., Schermer, B., and Zarsky, T. (eds.) (2013). *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Heidelberg: Springer

ECIS, ethical board at UvA: http://ivi.uva.nl/research/ethicalcode/ethicalcode.html

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification (2nd Edition). WileyInterscience.

Fayyad, U.M., PiatetskyShapiro, G., and Smyth, P. (1996). The KDD process for extracting useful know ledge from volumes of data, *Communications of the ACM* (39) 1996, No. 11, p. 2734.

Galloway, B. and Hancke G. P.. Introduction to industrial control networks. IEEE Communications Surveys & Tutorials, 15(2):860880, 2013.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big Data concepts, methods and analytics, *International Journal of Information Management* (35) 2015, No. 2, p. 137144.

Gibney, E. (2014). Physics: Quantum computer quest. *Nature*, 516, 24–26. https://doi.org/10.1038/516024a

Google (2016). "Google SelfDriving Car Project Monthly Report March 2016".

Groot, A.D. de (1946). *Het denken van den schaker, een experimenteel-psychologische studie*. Ph.D. thesis, University of Amsterdam.

Groot, A.D. de (1965). *Thought and Choice in Chess* (ed. G.W. Baylor) (translation, with additions, of the Dutch version of 1946). Second edition 1978. Mouton Publishers, The HagueParisNew York.

Hamburg, F. (2007). Kunnen Kennissystemen Professionals Helpen MedischEthisch Beter te Beslissen? *Liber Amicorum in honour of the Sixtieth Birthday of H.Jaap van den Herik*, pp. 187199.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of data mining*, Cambridge, MA: MIT Press.

Hassan, Qusay (2011). "Demystifying Cloud Computing" (PDF). The Journal of Defense Software Engineering. CrossTalk. 2011 (Jan/Feb): 16–21.

Herik, H.J. van den (1991). *Kunnen Computers* Recht*spreken*. Inaugural Address, Leiden University, Leiden, the Netherlands.

Herik, H.J. van den (2016). *Intuition is Programmable*. Valedictory Address, Tilburg University.

Herik, H.J. van den (2016). *The future of ethical decisions made by computers*, in the Art of Ethics in the Information Society, AUP, Amsterdam.

Herschberg, I.S. (1978). *In de ban van de fout.* Inaugural Address, Delft University of Technology, Delft, the Netherlands.

Koot, M.R. (2012). Measuring and Predicting Anonymity. Ph.D. thesis, University of Amsterdam.

Kosinski, M., Stillwell, D. & Graepel, T. (2012). Private traits and attributes are predictable from digital records of human behaviour, *Proceedings of the National Academy of Sciences* (PNAS), www.pnas.org/content/early/2013/03/06/1218772110.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety*, Stamford, CT: META Group Inc.

LeCun, Y., Bengio, Y., and Hinton, G.E. (2015). Deep Learning. *Nature*, Vol. 521, pp. 436444.

Li, Rita Yi Man and Li, Herru Ching Yu and Mak, Cho Kei and Tang, Tony Beiqi, Sustainable Smart Home and Home Automation: Big Data Analytics Approach (September 4, 2016). International Journal of Smart Home, Vol. 10(8), p. 177187, 2016.

MacMahon, B. and Pugh, T.F. (1970). *Epidemiology; Principles and Methods*, Boston: Little, Brown.

Markov, I. L. (2014). Limits on Fundamental Limits to Computation. *CoRR*, *abs/1408.3821*., https://arxiv.org/pdf/1408.3821.pdf

MayerSchönberger, V., and Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work and think*, New York: Houghton, Mifflin, Harcourt Publishing Company.

Morgan, J. (2014). Privacy is Completely and Utterly Dead, And We Killed It. *Forbes*, August 19.

Mosleh, M., Ludlow, P., & Heydari, B. (2016). Distributed Resource Management in Systems of Systems: An Architecture Perspective. CoRR, abs/1604.02114, 362–374.

Seirawan, Y. (1997). The Kasparov DEEP BLUE Match. *ICCA Journal*, Vol. 19, No. 1, pp. 3841.

Solove, D. (2004). *The digital person, technology and privacy in the information age*, New York: New York University Press.

Thomas J. Watson (n.d.). In *Wikipedia*. Retrieved November 2016, from https://en.wikipedia.org/wiki/Thomas_J._Watson

Van de Voort, M., Pieters, W., and Consoli, L. (2015). Refining the ethics of computermade decisions: a classification of moral prediction by ubiquitous machines. *Ethics and Information Technology.* DOI 10.007/11067601593602.

Weber, C.R.M. (2017). *Real-time Foresight – Preparedness for Dynamic Innovation Networks*. Ph.D. thesis. Leiden University, Leiden, the Netherlands (forthcoming).

Zliobaite I. & Custers B. (2016), Using sensitive personal data may be necessary for avoiding discrimination in datadriven decision models, *Artificial Intelligence and Law* (24): 183201.

Endnotes

[i] Big Data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. Source: New Horizons for a Data-Driven Economy, A Roadmap for Usage and Exploitation of Big Data in Europe, Editors: José María Cavanillas, Edward Curry, Wolfgang Wahlster.

[ii] Joint Doctrine Publication #4 (JDP 04) 'Understanding', p 2-5. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/33701/JDP04Webfinal.pdf

[iii] *Big Data in een vrije en veilige samenleving,* Wetenschappelijk Raad voor het Regeringsbeleid, April 2016.

[iv] *Big Data Discovery Is The Next Big Trend In Analytics*, Timo Elliott for ÜberTech, March 23, 2015.

[v] *The Internet of Things and Big Data: A Marriage Made in the Cloud*, Nate Philip, October 2014.

[vi] For completeness: recently there is a debate that this law is slowing down. See http://www.theregister.co.uk/2014/11/10/kryders_law_of_ever_cheaper_storage_disproven/

[vii] https://www.amazon.com/Fourth-Paradigm-Data-Intensive-Scientific-Discovery/dp/0982544200

[viii] Van den Herik and De Laat (Koot's supervisor) have also used this example in their publication The Future of Ethical Decisions made by Computers in Amsterdam University press.

[ix] https://www.privacybarometer.nl/pagina/45/Actuele_stand_van_de_privacy_barometer

[x] See Thomas J. Watson (n.d.). In *Wikipedia*.

[xi] See Markov (2014).

[xii] See Gibney (2014).

[xiii] See Campbell-Kelly (1987).

[xiv] See Hassan (2011).

[xv] See Galloway and Hancke (2013).

[xvi] See Li et al. (2016).

[xvii] See Mosleh et al. (2016).

[xviii] See Duda et al. (2000).

[xix] See Google (2016).

[xx] Gandomi and Haider (2015) estimate that 95 % of Big Data is unstructured.

[xxi] A similar approach is to pose a question and to collect data to answer the question.

[xxii] The proces of automated extraction of knowledge from datasets is called Knowledge Discovery in Databases (KDD), see Fayyad et al. (1996).

[xxiii] Classification from MacMahon and Pugh (1970, p. 18).

[xxiv] For more details on data management technologies, see Candan and Sapino (2010). For an overview of Big Data analyses technologies, see Calders and Custers (2013), Adriaans and Zantinge (1996) and Hand et al. (2001).

[xxv] Preparing for the future of Artificial Intelligence, Executive Office of the President, National Science and Technology Council, Committee on Technology, October 2016 https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[xxvi] *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, Executive Office of the President May 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[xxvii] *Big Society, Big Data. The radicalisation of the network society,* The Hague Centre for Strategic Studies and TNO, Prof. dr Valerie Frissen, TNO/Erasmus University Rotterdam, 2011.

[xxviii] *Big Data in een vrije en veilige samenleving,* Wetenschappelijk Raad voor het Regeringsbeleid, April 2016.

[xxix] Based on ongoing research and testing on automated data analysis and understanding at www.hcss.nl

[xxx] http://www.buienalarm.nl/

[xxxi] *Denk- en datamodel Suspicious Signs*, CCSS, J.G.M. Rademaker MTL, Prof. dr. R. de Wijk, dr. E. Bakker, J. Jansen, Mr. J. Stad, ir. C.J. den Hollander, drs. A.A. de Jong, 2006, http://www.hcss.nl/reports/denk-en-datamodel-suspicious-signs/53/

[xxxii] http://opsporingverzocht.avrotros.nl/

[xxxiii] http://www.interpol.int/notice/search/wanted

[xxxiv] https://www.bellingcat.com/tag/mh17/

[xxxv] Thomas van Linge, https://twitter.com/arabthomness

[xxxvi] https://www.burgernet.nl

Colophon

Enabling Big Data Applications for Security
Responsible by Design

A publication of

The Hague Security Delta
Wilhelmina van Pruisenweg 104
2595 AN Den Haag
T +31(0)70 2045180
info@thehaguesecuritydelta.com
www.thehaguesecuritydelta.com

@HSD_NL

Authors

Dr. Bart Custers, Centre for Law and Digital Technologies, Leiden University
Prof. dr. Jaap van den Herik, Centre for Law and Digital Technology (eLaw) and Director Leiden Centre of Data Science (LCDS) Leiden University
Prof. dr. ir. Cees T.A.M. de Laat, System and Network Engineering, University of Amsterdam
Michel Rademaker MTL, deputy Director The Hague Centre for Strategic Studies, Project leader
Dr. Cor Veenman, Senior Researcher Forensic Big Data Science, Leiden Institute of Advanced Computer Science (LIACS), Leiden University

Printing

The Communication Company

We would like to express our gratitude to everyone involved in the production of this report and the associated project initiatives against Subversive Crime and for Event Security.