Big Data and the Digital Self

Stefan Kohl

Abstract

Big Data is a buzz-word that refers to various aspects of the handling of huge amounts of data in order to obtain new knowledge from existing information. Outside the world of science, Big Data is applied in the economic and public sector, by collecting data generated by users of the web and of mobile devices. This paper will give a short introduction to the key concepts behind Big Data and its fields of applications. Since data can be used to gain deep insights into one's private life, it is further necessary to estimate the corresponding risks and to reason the value behind the concept of privacy. Finally, some failed strategies to preserve privacy will be mentioned before the paper concludes with a discussion on the necessary measures for politics and society to keep Big Data under control.

What is Big Data?

Big Data has gained increasing attention in recent years. The buzz-word appears both in scientific and industrial surveys of the IT market as well as in discussions about the social consequences of acquiring every single portion of data left by us when using new data processing technologies. There are many definitions of Big Data, most of them highlighting a specific aspect. The most general definition has been termed by the US market research company Gartner as follows: "Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Gartner 2013). This definition can be split into the following three components:

(1) The first component comprises the three V's; volume, velocity, and variety. Volume refers to the huge amount of data that is generated and needs to be stored on disks. Velocity means both the short period of time used to generate these data and the fast availability of results

from data processing. Variety refers to the different kinds of data that need to be handled. There is, for example, easy-to-process metadata (e.g., sender/recipient identifiers, date and time of transmission, geographical position data, etc.) in contrast to social media messages written in human language, which is much more complicated to analyze by a machine.

(2) Simply put, „cost-effective, innovative forms of information processing“, means the need for high performance and capacity of modern data processing systems.

(3) Finally, „enhanced insight and decision making“ is the goal of Big Data strategies, driven by the knowledge that is the outcome of the process. In a nutshell, one who uses Big Data is trying to deduct predictions on future events based on information about past events.

How does Big Data work?

Before talking about what Big Data is used for in the economy and in public, I will first give a general idea on how Big Data systems work. Discussing technical details would be beyond the scope of this paper. Furthermore, they strongly depend on the actual implementation, e.g. Apache Hadoop and MongoDB. Big Data is mainly about extracting knowledge from huge amounts of information by using statistical methods. Computer systems with powerful storage and computation capacities are used to process data in high quantity rather than quality. This is the process of Knowledge Discovery, which transforms information (i.e., data embedded in context) into knowledge.

Data Acquisition is the first step of the overall process of Knowledge Discovery. Many sources of data exist, with the following being be the most important ones:

(1) Traffic data, also called metadata in a more general perspective, is data for classifying the content of a message or a file and providing general information about it. Examples for metadata are the addresses of sender and recipients of an email or the numbers involved in a telephone con-

versation. When using mobile phones, the cell identifier (CID, i.e., a discrete approximation of the current geographical position of the mobile phone) is also part of the traffic data, as well as date and time of calling. Data retention (also known as „Vorratsdatenspeicherung“ in German) actually is about storing traffic data. No content of telephone conversations or Internet sessions will be stored. This might seem useless, but in fact, metadata is more valuable than content, because, in terms of statistical algorithms, it is easier to analyze for a computer, and, as said before, Big Data is more about quantity than quality (Krempl 2014). Just consider the following fictional example: There are records indicating that you called both your bank advisor and a car salesman several times within two weeks. According to statistical correlation, it is most likely that you apply for a credit to buy a car, because whenever people call both their bank advisors and car salesmen within a short period of time, there is a high chance of being involved in a car purchase. This is sufficient for the assumption made by the algorithm. On the other hand, computationally analyzing natural language is an expensive and complicated task for computer systems and technology currently is far away from producing reliable results. The inaccuracy of automated translation software (e.g., Google Translator) speaks volumes. Metadata of web sessions is used in a similar way. Due to search requests on Google, or social media plugins on several websites, it is possible to track your surfing behaviour and deduct your interests. You may have already noticed that, when searching for product information on Google, you will be confronted with advertisement on similar products over the next few weeks, whenever you visit a website that uses Google AdSense. But also, information posted on your social media profile is very valuable for Big Data applications. Again, it is not necessary to perform natural language analysis of the website content. Analyzing keywords, tags, and subjects is sufficient in most cases.

(2) Locating a device by using its sensors and wireless adapters is called Geolocation. It is very common on mobile phones and used by service providers to present information, which might be relevant to the user at his or her current location. By consecutively tracking the locations

of a mobile phone user, a movement profile can be generated. Some athletes are using special services (e.g., Endomondo, Runmeter GPD Pedometer, AllSport GPS, etc.) voluntarily to track their training progress on their smartphones. But not only smartphones are used for that purpose. Special consumer devices to measure certain aspects of the body are sold to so-called „Quantified Sellers". Ubiquitous and Wearable Computing deal with technologies, that are attached to our environment or ourselves and aim to be „invisible", compared to the more intrusive appearance of computers (Weiser 1991). Put together, these technologies form an omnipresent network of instruments to record various characteristics of ourselves and the environment we live in.

Of course, this only covers certain fields of applications, most likely the projection of human individuals into digital profiles. Big Data also plays an important role in science, for example at CERN, where scientists have to deal with vast numbers of data, generated by the Large Hadron Collider (Redmond & Wilson 2012, 147).

After acquisition, data needs to be stored in Data Warehouses. As mentioned above, Big Data deals with amounts on a far different level than we are used to. As a result, Data Warehouses consist of many thousands of hard discs, built into racks, with high-scaled cooling and power supply systems.

(3) The final step of Knowledge Discovery is called Data Mining. This is where computation actually comes in, by filtering data in regard to some aspects of interest and applying statistical methods to extract correlation. Knowledge, in some definitions, is nothing more than putting two or more pieces of information into relation, and to use them as guides for actions (Knowledge Process, Inc. 2009). Let us have a look at our previous example: The two pieces of information are the fact that you called your bank advisor and that you contacted a car salesman. The knowledge is, that you will buy a car and that you may want to contract an automobile insurance. As an insurance agent, you will probably visit me to apply for contract, so I need to find out more about you. This already leads to the next question.

Who profits from Big Data?

When using personal information for Data Mining, it is not the actual content of the piece of information that Big Data actors are interested in. They are not interested in the fact that you are a fan of a certain movie or song. They want to know what this „being a fan of a certain movie or song" tells about you. By using statistical correlations, the goal of Data Mining is to extract knowledge about you, which has not been explicitly revealed by yourself (or the information about you). For example, consider you are a fan of the movie series „The Fast and the Furious". The insurance company then wants to find out what this means for your personal driving behaviour and the estimation of risk, which is an important indicator for calculating insurance rates. This whole process is called Scoring, which has become more and more important to companies in recent years. A score is some kind of personal information that describes a customer, usually a single number quantifying the likelihood of a customer to cause financial disadvantage to the contract partner. In other words, a score indicates how big the risk will be when signing an insurance contract with a certain customer.

Scores are nothing new, they have been used in the financial sector since the mid of the 20th century. The so-called credit score measures the financial solvency of a person. The solvency decides if there will be a contract at all, which payment conditions arise, and, in case of credits, how much the interest rates will be. Of course, the better a customer's score is, the better the conditions for him or her. Traditionally, the score depends on a person's income and financial history.

In the last decade, however, more and more creative methods were introduced to calculate scores, both in the financial and the insurance sector. Big Data plays a major role here. As it is already common to give you a score depending on where you live („Geo-Scoring", see Rosenbach 2008), information on your age, your education, your life-style, etc. will influence the estimation of your financial solvency or risk likelihood, conducted by rating agencies. At the moment, research is going on to improve the respective algorithms (Knowledge@Wharton 2013). The goal of Scoring is simple: Maximizing yields by minimizing risks.

Big Data has also become an important instrument in marketing. A company that knows about the interests of some individuals can target potential customers. Or, on the other side, ignore people with little or no chance of affiliation for a certain product. This is nothing new. Rather, we experience this day to day when visiting websites, disposing of junk mails, but, to a smaller degree, also offline, when receiving brochures. Finding out personal preferences and making suggestions for future buying decisions is the major task of Recommender Systems. Recommender Systems analyze people's shopping history and compare it to the history of other persons, which share some similarities. The system will then recommend products, which were bought by those people, but not by yourself. The suggestion uses the following reasoning scheme:

(1) You and person X both like product A.

(2) Person X also likes product B.

Thus, you also might like product B.

This reasoning is highly based on correlation, which is, as already mentioned, a basic condition for Big Data to produce new insights from existing information.

Many profiles used by companies for scoring or marketing were originally not created for their own services. Instead, it is common practice to buy databases containing user profiles from external companies, called information brokers. Anyone who can sell large amounts of data can assume the role of an information broker. Many companies, which do not primarily specialize in information broking, for example mobile phone providers, sell their own databases to other companies in order to obtain an additional income. Sometimes, user profile databases are bought as a first step. While doing business with customers, new information is won and added to the database. Finally, the grown database will be sold again, to other companies (Pepitone 2013). Note that due to the nature of digital data, selling data over and over again will result in an accumulation of information, with the mesh becoming tighter and tighter with each iteration.

Also, politicians show interest in Big Data. Since 9/11, western governments put high efforts into anti-terrorism and crime prosecution strategies by surveilling telecommunication and online traffic. In the US, the legal foundations were set back in 2001 by signing the USA PATRIOT Act. A few years later, the European Union approved the highly criticized Data Retention Directive, which has finally been repealed in March 2014. Finding evidence for crime that has already happened and intercepting communication about future acts of crime or terrorism is the most important motivation for politicians to enforce data retention. But by using public safety as an argument, interest is also shown in the prevention of crime by relying on predictions based on past events. This is called Predictive Policing. The movie „Minority Report" gives a pretty good idea what perfect crime prediction could look like (besides the supernatural powers involved in the movie, which, in reality, would be replaced with Big Data). There will probably never be an exact prediction on which crime will happen, but predictive policing tends to estimate the likelihood of criminal actions within a certain timespan or around a certain place.

Another field of application is Big Medicine, which refers to Big Data applied in medicine. Big Medicine is used to process data from medical treatment reports and patient's files to discover treatments that have already been successful on patients with similar symptoms. This tackles one major problem: there is too much data available for someone to effectively study the advance in medical treatment. It is not possible to see the wood for the trees, where the wood stands for knowledge and the trees for information (Ohlhorst 2012, 87-88).

What are the consequences of using Big Data?

The following discussion refers to the fields of applications mentioned above, all having the involvement of personal information in common. There are also debates around Big Data in the sciences (where no personal information comes into play), guided primarily by questions from the field of the Philosophy of Science (e.g., Will scientific theories be based on correlation rather than on causal explanations?, see Anderson 2013). Wherever individual persons are involved, Big Data will lead to a loss of control over what happens to the affected people's personal data. You will not know who has a digital profile of you, and neither will you know what information this profile contains. This loss of control over personal information is the key to raise concerns regarding privacy. The next section will therefore outline the relationship between privacy and information.

The idea of informational privacy was first expressed by Samuel D. Warren and Louis D. Brandeis in the article „The Right to Privacy", published in the Harvard Law Review in 1890 (Warren & Brandeis 1890). This article was an answer to the technological advance that was going on during the same period. Until then, causing damage to people by violating their privacy has not been an issue. This changed with the advent of photography and newspapers. The disclosure and mass-distribution of details on a person's private life became very easy and gossip could ruin the reputation and wealth of an individual. The right to privacy, which has been deducted from the more general „right to be let alone", should prevent this from happening. In the 20th century, the idea of privacy evolved into the concept of privacy as control over (1) personal information, (2) physical access, and (3) social behaviour and decisions (DeCew 2013). This means that the loss of privacy due to Big Data is equal to the loss of control over information.

But what is the value of privacy? Why do we need it at all? Coherentist authors on the concept of privacy who elaborated on the different aspects of privacy as control give various reasons. Adam Moore, for example, refers to empirical studies which have shown that a lack of privacy in terms of physical access leads to stress and aggression against the own fellows, which has been observed on both humans and animals. How the concept of privacy is realised in practice depends on the respective civilisation or social entity. However, each one has strategies for its members to temporarily release themselves from their social role. In the absence of such mechanism, self-reflection and self-development would arguably not be possible (Moore 2003). James Rachels defends privacy as precondition to social inclusion, driven by the control over social behaviour. This is also related to informational privacy, since an individual needs to retain control over the information he or she discloses or conceals to others. Which personal or even intimate information is actually shared between two people depends on their relationship. Their relationship also determines how they behave between each other. Hence, a co-worker is treated differently than a spouse. To get along with each other, one needs to adapt to the different situations that arise from the relationship between two people. A third (probably uninvited) person who interferes with the interaction causes the other two to change their behaviour, e.g., by turning to another (less intimate) topic in the course of a conversation (Rachels 1975). Again, this can be seen as a loss of control since the presence of a third person restricts the interaction between the dialogue partners. This is consistent with coherentist theories on privacy, which state privacy as precondition to other values, for example, dignity or intimacy (DeCew 2013).

The unwanted disclosure of intimate details is the loss of control over personal information. And this is exactly what happens when Big Data discovers personal information that has not been made public explicitly. This is the difference between voluntarily providing information on social network profiles, like age, interests, attitudes, occupation, etc. and the process of Data Mining, which uses statistical algorithms to find out even more about us. In addition, we cannot control these processes; even if we do not use computers and mobile devices, because ultimately, we are still part of a society that uses a digital infrastructure for e.g. social insurance and registration systems.

From the loss of privacy I will move on to the more tangible consequence of statistical discrimination caused by the primacy of statistics. Statistical discrimination refers to discrimination of people as a side-effect of drawing conclusions from statistical correlations. The scheme behind this kind of discrimination is similar to the scheme of statistical reasoning used by Big Data:

(1) Person X features trait A.

(2) People who feature trait A also feature trait B (in statistics).

Thus, Person X also features trait B.

In general, statistical discrimination occurs when personal traits or circumstances, which are not bad in the first place, statistically correlate with facts that are. For example, statistical correlations between solvency and sociographic characteristics will be applied to customers in advance, regardless of the adequacy in particular cases. A higher income or a faultless financial history will be irrelevant if you live in an area with a higher number of socially disadvantaged people. This is something that can already be experienced. For example, some online shops refuse to offer debit-based payments to customers with certain zip codes (Rosenbach 2008). Statistical discrimination may also be the consequence of the example above: What does your enthusiasm with the movie series „The Fast and the Furious“ tell about your driving habits? By providing this information on a social network site, an insurance company will probably raise your insurance rate due to a higher estimated risk based on statistical correlation. This fictional example was inspired by a study that explores the correlation of driving habits and playing street-racing computer games. It has been shown that young men who regularly play games of this genre are more often involved in car accidents than others (Vingilis 2013). Research is currently being conducted on how to create Big Data applications, which use information from social media to create more accurate scores. It is just consequential for companies to consider all kind of available correlation to maximize their profits. That is the reason why the German insurance company Generali Versicherung is introducing discounts for customers who voluntarily wear self-tracking devices in 2015: Significantly more data will be available for a more accurate adaption of the insurance rate. Furthermore, customers will be encouraged to improve their health, which will probably reduce costs for insurance settlements in the long term (Schumacher 2014).

These models are a form of discrimination, meaning that people might get wrongly accused by drawing conclusions from statistical correlations. This problem has already existed for many years (e.g., insurance rates, in fact, depend on age and gender) but will be elevated to new levels.

How shall Big Data be handled?

Of course, many people are aware of the risks that may arise from Big Data and try to defend privacy by enforcing laws. However, the strategies currently available were established years before the advent of Big Data and therefore miss some critical aspects.

First of all, data privacy acts only focus on the primary use of data (which is, the original intent for storing data), while Big Data adds a secondary use cycle, in some cases even years after the data has been stored. When the user agrees to the data privacy statement of a service provider, one may only guess what will happen to the data in the future (Mayer-Schönberger & Cukier 2013, 193-194). Even though providers are required to ask the user to refresh the agreement, a problem as to the definition of personal data remains: Not all data related to a user represents personal information. But with Big Data, it is possible to reveal personal information hidden in non-personal data by applying statistical algorithms. For example, a person may explicitly state where she lives by entering the home address. This is personal information. Now consider a person is not disclosing any information about herself, but she uses a smartphone to track her daily distance covered by foot. The geographical information sent to the service provider is nothing personal, but it can be used to retrieve the person's identity by including data from external databases (like a telephone book). And that's what Big Data is about: High quantities of behaviour tracking data can be used to deduct high quality personal information. The same holds true for other services like Google or Amazon. It is possible to create a pretty accurate personal profile based on search terms or page/product views.

Furthermore, data privacy acts assume that individuals have full autonomy on the decisions they make when it comes to subscribing for an online service. This might be true from a general point of view. However, decisions regarding social media usage are influenced by social mechanisms like peer pressure. This was observed by Knoll et al. (2013) in a study with adolescents by analyzing their social media usage. This means, when signing up for a WhatsApp account, the question one has to answer individually is: „Will I risk the loss of privacy or the loss of association?“

Accordingly, in many situations a person's autonomy does not refer to decisions on whether to use a certain service or not, but to decisions on risking social exclusion or not.

However, if social exclusion is not an issue, a person might have different reasons not to use a particular service. This action of refusal is called „opting out“ and is an option required by law. In Austria, for example, it is possible to opt out from the recently introduced „Elektronische Gesundheitsakte“ (ELGA). However, for some people or institutions, an empty dataset is not semantically empty at all. Opting out raises suspicion because of the presumption that one who opts out might want to hide something (Schirrmacher 2013, 276-277).

Finally, there are also some technical strategies to preserve privacy, namely anonymization and pseudonomization. Anonymization removes personal information from datasets, while pseudonymization distorts data. Both strategies are used to prevent identification and were established years before Big Data came into action. Big Data tends to be robust against noise and errors. It is also capable of combining datasets from different sources, hence filling the gaps of information (Mayer-Schönberger & Cukier 2013, 196). Supposed anonymization and pseudonymization retain a certain amount of entropy, both methods fail to preserve privacy accordingly.

It is therefore necessary to introduce a completely new approach on how to deal with Big Data. First of all, we must accept that Big Data is an inevitable consequence of the advances of information technology. It also opens new possibilities in sciences and healthcare, which is why reverting the process is not an option. However, it is necessary to change the paradigm of responsibility. At the moment, there are hardly any restrictions for organizations on how they use the technology, while individual persons must deal with full responsibility and risks. In many cases, people do not have the insight on what is going on when they use digital technology and which consequences will follow. Thus, the first step must be a shift of responsibility from individuals to Big Data organizations. They will have to follow both a moral code of conduct and legal rules, which take all the aspects discussed so far into account.

A set of legal acts, however, will not be sufficient to control the huge impact Big Data has on society. Similar to any other technological milestone in modern history, new institutions must be created to control the use of technology and protect society from damage. Society has always been adapting to technological advances. After the pioneer years, people learned about the negative consequences of technology and thus introduced rules and institutions to put them under control. Automobiles are one example: The complex traffic systems, road traffic regulations, the certification of manufacturers, driving licenses etc. Instead of prohibiting the use of cars, we introduced these institutions to increase safety. When it comes to Big Data, we must respond the same way.

References

Anderson, Chris (2013), 'Das Ende der Theorie – Die Datenschwemme macht wissenschaftliche Methoden obsolet', in Geiselberger, Heinrich and Tobias Moorstedt (Eds.), *Big Data: Das neue Versprechen der Allwissenheit*, Berlin: Suhrkamp Verlag GmbH, 124–130.

DeCew, Judith (2013), 'Privacy', in Zalta, Edward N. (Eds.), *The Stanford Encyclopedia of Philosophy*, Fall 2013, http://plato.stanford.edu/archives/fall2013/entries/privacy [19 December 2014].

Gartner, Inc. (2013), 'Big Data', *IT Glossary*, http://www.gartner.com/it-glossary/big-data [22 March 2015].

Knoll, Bente, Bernadette Fitz, Patrick Posch, and Lukas Sattlegger (2013), *Ich im Netz – Selbstdarstellung von weiblichen und männlichen Jugendlichen in Sozialen Netzwerken*, http://www.saferinternet.at/fileadmin/files/imaGE_2.0/Ich_im_Netz_Bericht_09012014_FINAL.pdf [19 December 2014].

Knowledge Process, Inc. (2009), 'Data, Information, Knowledge, and Wisdom', *eDocumentation Process*, http://www.documentationprocess.com/Data_Information_ Knowledge_and_Wisdom [23 March 2014].

Knowledge@Wharton (2013), *The 'Social' Credit Score – Separating the Data from the Noise*, http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise [8 May 2014].

Krempl, Stefan (2014), *Studie: Was auf Vorrat gespeicherte Verbindungsdaten verraten*, http://heise.de/-2146213 [13 April 2014].

Mayer-Schönberger, Viktor und Kenneth Cukier (2013), *Big Data – Die Revolution, die unser Leben verändern wird*, Munich: Münchner Verlagsgruppe GmbH.

Moore, Adam D. (2003), 'Privacy – Its Meaning and Value', *American Philosophical Quarterly* 40.3, 215–227.

Ohlhorst, Frank J. (2012), 'Big Data Analytics – Turning Big Data into Big Money', *Wiley and SAS Business Series*, Hoboken, New Jersey: John Wiley & Sons.

Pepitone, Julianne (2013), 'What your wireless carrier knows about you', *CNN Money*, http://money.cnn.com/2013/12/16/technology/mobile/wireless-carrier-sell-data/ [19 December 2014].

Rachels, James (1975), 'Why Privacy is Important', *Philosophy and Public Affairs* 4.4, 323–333.

Redmond, Eric und Jim R. Wilson (2012), 'Sieben Wochen, sieben Datenbanken – Moderne Datenbanken und die NoSQL-Bewegung', *The Pragmatic Programmers*, Cologne: O'Reilly Verlag GmbH & Co.KG.

Rosenbach, Marcel (2008), 'Verbraucherschutz – Die neue Klassengesellschaft', *Spiegel Online Wirtschaft*, http://www.spiegel.de/spiegel/a-548454.html [10 May 2014].

Schirrmacher, Frank (2013), 'Der verwettete Mensch', in Geiselberger, Heinrich and Tobias Moorstedt (Eds.), *Big Data: Das neue Versprechen der Allwissenheit*, Berlin: Suhrkamp Verlag GmbH, 273–280.

Schumacher, Florian (2014), 'Versicherungen und Self-Tracking – Die Versicherung Generali möchte Boni auf Basis von Gesundheitsdaten gewähren', *igrowdigital*, http://igrowdigital.com/de/2014/11/versicherungen-und-self-tracking/ [19 December 2014].

Vingilis, Evelyn, Jane Seeleya, David L. Wiesenthalb, Christine M. Wickensc, Peter Fischerd, and Robert E. Manne (2013), 'Street racing video games and risk-taking driving – An Internet survey of automobile enthusiasts', *Accident Analysis and Prevention* 50, 1–7.

Warren, Samuel D., and Louis D. Brandeis (1890), 'The Right to Privacy', *Harvard Law Review* IV.5, http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm [19 December 2014].

Weiser, Mark (1991), 'The Computer for the 21st Century', *Scientific American* 265.3, 66–75.