University of Economics in Prague

Faculty of Informatics and Statistics

Department of System Analysis

Big Data and Ethics

Doctoral Thesis

Author:	**Ing. Richard Novák**
Supervisor:	Doc. Ing. Vlasta Svatá, CSc.
Study Program:	Applied Informatics

Prague, 2019

STATEMENT OF ORIGINALITY

I hereby certify that all of the work described within this thesis is the original work of the author. Any published (or unpublished) ideas and/or techniques from the work of others are fully acknowledged in accordance with the standard referencing practices.

Richard Novák, 2019

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Associate Professor Vlasta Svatá, Dr. Antonín Pavlíček and Professor Vojtěch Svátek, my research supervisors, for their patient guidance, enthusiastic encouragement and tolerance of this research work that is interdisciplinary and cover not only the ICT area but also touches on the related fields of Sociology and partially also Philosophy.

I would also like to thank Dr. Tomas Sigmund, for his advice and assistance and useful critiques in keeping our joint publications on schedule. My grateful thanks are also extended to Ms. Jana Cibulková for her help in doing the survey data analysis and to Mr. Ben Koper for our long discussions about the right English formulation of my ideas.

I would also like to extend my thanks to the technicians of T-Mobile Czechia and Slovakia from various departments for their help and sharing their knowledge of Data Science that helped me understand this competitive, sensitive area.

Finally, I wish to thank my wife, Eva, and family for their support and encouragement throughout my study and internship abroad.

ABSTRACT

Big Data is a relatively new term that has so far not been viewed through the lens of applied ethics.

My focus in this thesis is on the awareness of the conflicts arising between Big Data phenomenon and its issues and the relevant ethical principles.

Firstly, I start with the research of other authors and an overview of Big Data and ethics, and the definitions that are generally accepted. Secondly, I continue with the description of data sources and Big Data use cases from the telecommunication industry, demonstrating what is currently feasible, that I will generalize and, furthermore, suggest a comprehensive list of twelve Big Data issues such as Privacy Intrusion, New Barriers, Business Advantage, Power of All data, New Big Brother effect, Missing Transparency, Confusion, Social Pressure, Belief in Legislation, End of Theory, Data Religion and Unawareness of our Data. Thirdly, I describe the existing regulatory framework of the Big Data area with the clarifications and some suggestions for improvement, and I also verify the awareness of the suggested twelve Big Data issues by launching an international survey. Finally, I discuss and conclude the thesis results.

The survey (N=733) of university students, IT professionals and seniors from EU countries, mainly Czechia and Slovakia concluded that Big Data issues are grouped into three different and consistent clusters: hot, cold and warm (suggested by the Ward method that uses the Euclid distance between the mean and standard deviation).

I found, using MANOVA Pillai's statistical test, that clusters are significantly dependent on demography (IT Skills, Occupation and Sex). Warm clusters show interesting dependencies on the demographic category, such as the social pressure perceived important by pensioners and women compared to the underestimated importance reported by men and IT Professionals. The conclusion of the thesis is that the awareness of Big Data issues can be grouped into three consistent clusters that depend on a few demographic variables. I also conclude that there is a need for regulation frameworks to move past Big Data Ethic by Default (Law) to a priori Big Data Ethics by Design approach.

Keywords: awareness, big data issues, cluster analysis, big data ethics by design, demography, digital divide, manova

ABSTRAKT

Termín velká data je relativně nový a neprošel tedy dosud důkladnou diskusí v oblasti aplikované etiky.

V mé disertační práci se zaměřuji hlavně na uvědomění si existence některých problémů velkých dat, které vznikají ze střetu tohoto fenoménu s dosud známými etickými principy.

Disertační práce má následující strukturu. Nejprve je provedena rešerše v oblasti velkých dat a etiky. Potom, pokračuji s popisem datových zdrojů a případových studií i možného komerčního nasazení velkých dat v oboru telekomunikací. Zde se snažím ukázat co všechno je pomocí velkých dat v této oblasti možné, abych následně provedl zobecnění, které mi umožnuje navrhnout souhrnný seznam všech souvisejících problémů a rizik velkých dat. Konkrétně jde o: narušení soukromí, nové bariéry, obchodní výhody, dominance malého počtu datových korporací, efekt velkého bratra, chybějící transparence, zmatení světa, sociální tlak, víra v legislativní řešení, konec obecných teorií, nové datové náboženství a nevědomost o sběru našich vlastních dat. Následně analyzuji současný regulatorní rámec velkých dat a doplňuji svoje vlastní návrhy na zlepšení v této oblasti. Na závěr formuluji shrnutí disertační práce.

Součásti disertační práce je průzkum (N=733) mezi universitními studenty, IT odborníky a seniory v EU, zejména v České republice a na Slovensku. Tento průzkum za použití Wardovy metody, založené na Euklidovy vzdálenosti střední hodnoty a směrodatné odchylky, odhalil existenci tří odlišných shluků problému velkých dat pojmenovaných jako horký, chladný a vlažný shluk.

Průzkum za použití MANOVA statistické metody odhalil, že shluky jsou významně závislé na demografii, konkrétně na IT dovednostech, povolání a pohlaví. Například v horkém shluku je mnohem významnější sociální tlak na seniory a ženy než na ostatní. Závěrem disertační práce tedy je, že uvědomění problémů velkých dat lze rozdělit do tří odlišných shluků, které jsou demograficky závislé. Dalším závěrem je, že v regulatorní rámci, který dopadá na velká data, vzniká poptávka doplnit existující legislativu a právo (mandatorní a dodatečný prvek) o nový regulační prvek etiky spojený s metodikou návrhu IT systémů (Big Data Ethics by Design) aplikovanou v počáteční fázi všech IT datových projektů.

Klíčová slova: uvědomění, rizika a problémy velkých dat, shluková analýza, etika velkých dat založená na metodice návrhu IT datových systémů, demografie, digitální rozdělení, manova

CONTENTS:

1 Introduction

This thesis will explore the costs and benefits of using Big Data in the context of applied ethics. At the heart of my thesis is the following quote from (Sokol, 2016) and (Boyd & Crawford, 2012, p. 671)

"Just because it is possible does not make it ethical."

We can observe that people's insight into complex problems and their attitudes towards life is currently driven by advanced technologies such as Big Data and Artificial Intelligence among others. It means *"Change the instruments, and you will change the entire social theory that goes with them* ". (Latour, 2009, p.155).

Thus, university researchers are formulating new hypotheses about the possible shifts, divides, manipulation and inequalities in society driven by new technologies and are trying to evaluate these changes in society using statistics. This approach to changes in micro-segments of society can be represented, e.g., by the Dutch survey (N=1,356) about digital inequalities in the Internet of Things (Deursen, et al. 2019) or the Taiwan survey (N=775) about problematic internet use among elementary school students (Wang & Cheng, 2019).

The thesis is summarizing the previous research on Big Data ethics and provides a comprehensive view of all the possible Big Data issues that have only been partially discussed up to now (Privacy, Big Brother Effect et al.) or without a broad awareness survey.

This paper suggests a comprehensive list of Big Data issues based on previous research. (Floridi, 2016; Boyd & Crawford, 2012; Cukier & Mayer-Schoenberger, 2013; Cavoukian, 2011; O'Neil, 2016, Spiekermann, 2001, 2018; Anderson, 2008; Andrejovic, 2014; Dijk, 2006; Norris, 2001; DiMaggio& Hargittai, 2001; Allcott, 2017; Haesler, 2013 and Davenport, 2001 among others).

The paper explores the awareness of Big Data issues in a survey (N=733) among different focus groups of university students, IT professionals and seniors from EU countries, mainly Czechia and Slovakia. In the survey, I focus on the demographic differences that follow the work and interesting findings of previous surveys from Australia (Andrejevic, 2014) and the USA (Latonero & Sinnreich, 2014). These case studies showed that demography has an important role in the attitude towards new technologies and their ethics that can later create inequalities and form a digital divide that is based not only on access to new technologies but also on new skills and benefits related to them (DiMaggio & Hargittai, 2001; Dijk, 2006; Deursen & Helsper, 2015).

The thesis also discusses the possible regulatory framework that can assure the compliance of the Big Data phenomenon with data ethics defined, e.g. by Floridi and Taddeo (2016) as interplay of the ethics of data, algorithms and practices.

1.1 Elementary Definitions

There are a few good definitions of Big Data. The most common are from consulting companies like this one from Gartner:

"Big data, in general, is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Gartner, 2018).

A very relevant definition of Big Data for this thesis comes from the Microsoft Research Center (Boyd & Crawford, 2012):

> *"We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:*
> *(1) Technology:*
> *maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets*
> *(2) Analysis:*
> *drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims*
> *(3) Mythology:*
> *the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy".* (Boyd & Crawford, 2012, p. 663)

Awareness of Big Data will be discussed later in our survey; however, I can define the term Awareness now to avoid confusion, because it can differ from situational awareness to contextual awareness. Based on Philip M. Merikle (1984) there are two different definitions of awareness, such as the objective (can be proven) and the subjective (self-reported, cannot be proven). Regarding our survey and questionnaire's design, I follow the subjective definition of awareness that can be described as:

"Awareness is a state wherein a subject is aware of some information when that information is directly available to bring to bear in the direction of a wide range of behavioral processes". (Chalmers, 1997, Chapter 6.3)

Highlighting the results of the survey and its demographics can attract attention to the so far underestimated issues or micro-segments related to Big Data that are now hidden.

The free use of Big Data masks its danger as some respected authors such us Floridi (2016), Sokol (2016) and Boyd & Crawford (2012) warn us.

I should also describe here, in the definition part, the terms of Ethics and Equality to clarify the broader context of the thesis that I will dive deeper into throughout the following chapters.

The term ethics derives from Ancient Greek and was firstly used by Aristotle to name a field of study developed by his predecessors Socrates and Plato.

"Philosophical ethics is the attempt to offer a rational response to the question of how humans should best live." (Wikipedia, Aristotle, 2018)

There are many areas of ethics such as meta-ethics or normative ethics, although for this thesis applied ethics is the most relevant because it can further be adapted to more detailed fields such as bioethics, business ethics or data ethics.

"Applied ethics is a discipline of philosophy that attempts to apply ethical theory to real-life situations." (Ethical World, 2018)

Equality is used in this thesis as a term describing equal opportunities and the rights of all people that are described, for example, in the EU Charter of Fundamental Rights (European Parliament, 2000). These equal opportunities and rights have recently been challenged by new innovations such as the internet and Big Data. These innovations have caused inequalities known as the Digital Divide, which has a special chapter in this thesis. The following definition of equality that is inspired by John Rawls book: *A Theory of Justice*, describes the meaning of the word equality that we will use in this thesis:

"Individuals with similar efforts face the same prospects of success regardless of their initial place in the social system" (Rawls, 1971).

1.2 Scope and Goals of the Thesis

The scope and related goals of the thesis are the following:

- Clarification and sorting of research of other authors relevant to Big Data and Ethics.
- Description of data sources and use cases from telecommunication showing the positive and negative effects of Big Data that can be applicable also to other industries.
- Discussion of regulatory framework and comparing of the different approaches to the ethical assurance of Big Data systems.
- To provide a new proposal and comprehensive list of the all possible Big Data issues that are negatively impacting society and their categorization.

- Execution and evaluation of a large survey about the awareness of the Big Data and Ethics conflict that confirms the theory and hypotheses described in the thesis.

1.3 Structure of the Thesis

The thesis has the following structure. Firstly, I start my research with the research of other authors and overviews of Big Data and ethics, and the definitions that are generally accepted. Secondly, I continue with the description of Big Data use cases from the telecommunication industry, demonstrating what is currently feasible, that I will generalize and, furthermore, suggest a comprehensive Big Data issues list. Thirdly, I describe existing regulatory framework of Big Data area with the clarifications and some suggestions for improvement. Finally, I verify the awareness of the suggested twelve Big Data issues by launching an international survey. The survey research is mainly focused on two European countries (Czechia, Slovakia) and different stakeholders such as university students, IT professionals and seniors. Finally, I discuss and conclude the survey results.

2 Definition of the Research Area

2.1 Research Question

We can formulate the main research question of the thesis as the following:

> **RQ:** *What is the awareness of the stakeholders, such as university students, IT professionals and seniors to our suggested list of Big Data issues arising from the Big Data phenomenon?*

The four hypotheses related to this research question are defined in chapter 6.3. Hypotheses Definition because we need to go first through the research and our description of ethics and use cases before we can formulate the comprehensive list of Big Data issues and related hypotheses.

2.2 Methodology

I plan to use the following **scientific methods** in my thesis.

- **Empirical methods** such as surveys, observation and also interviews for smaller focus groups
- **Logical methods** such as analysis vs synthesis, induction vs deduction, abstraction vs concretization

Data collection method:

- Online questionnaires

Data analytics and statistical evaluation:

- Exploratory data analysis
- Linear regression
- Correlation analysis
- Cluster and Factor analysis
- Hypotheses testing
- MANOVA method

3 Big Data and Ethics Overview

3.1 General Overview

Big Data is a subset of data science[1]; however, there are some aspects that make this topic very unique. The original definitions of Manyika´s team from the McKinsey Institute (2011) and Gartner (2018) focused on the 3Vs: high volume, variety and velocity of data. This definition was later extended to the 5Vs by adding value and veracity described, e.g., by Yuri Demchenko (2013). Nowadays, we discuss Big Data as a socio-technological phenomenon rather than only a technological one (Boyd and Crawford, 2012). Thus, as I review the research of other authors, I prefer to start with the broader view of data science and applied ethics that are umbrellas for the specific topic of Big Data ethics.

Applied ethics covers several areas that are relevant to Big Data such as **computer ethics**, **professional ethics**, **information ethics** and **data ethics**.

Before I discuss, the above mentioned, relevant areas of applied ethics, I will do a short overview of more general terms such as digital ethics, cyber ethics and business ethics.

Digital ethics is an umbrella term covering all issues raising from the conflict of digital technologies and ethics. Robert Capurro defines digital ethics as:

"Digital ethics or information ethics in a broader sense deals with the impact of digital Information and Communication Technologies (ICT) on our societies and the environment at large." (Capurro, 2009).

Cyber ethics covers, in my view, the behavior in a broader area of virtual cyber space of society that is created by ICT. For a definition of cyber ethics, we can use the following:

"Cyber ethics is a set of moral choices individuals make when using internet-capable technologies and digital media." (Chen, 2012)

For a more detailed description of cyber ethics, see Tomas Sigmund´s (2013) article: *Ethics in the Cyberspace*.

Business ethics is, for me, an even more general term covering classical philosophy challenged by business environments and moral problems arising from different issues such as conflicts of interest, social contracts and stakeholder groups. For an overview of

[1] Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. (Wikipedia, 2018)

business ethics, see, e.g., the work of T.Donaldson and T.Dunfee and their article: *Toward a Unified Conception of Business Ethics: Integrative Social Contracts Theory* (Donaldson & Dunfee, 1994).

This thesis is focused on Big Data ethics thus my approach goes from general business ethics to specific professional ethics of data science related to impacted people. Furthermore, my approach goes from general data science to the specifics of Big Data. For a view on the possible intersection of data science with cyber space, which is not my direct focus, I recommend the book by Richard A. Spinello, (2010), *Cyberethics: Morality and Law in Cyberspace*.

In the following text, I discuss the relevant Big Data ethics predecessors such as computer ethics, professional ethics, information ethics and finally also data ethics.

3.2 From Computer to Information Ethics

Computer ethics has been evolving since the invention of computers in the 20th century after the world wars. A deeper look at the moral problems related to information technologies started in the 40s by the work of MIT professor Norbert Wiener that introduced the term "Cybernethics" in his eponymous book (Wiener, 1948). It was followed by his other books. For example, in *The Human Use of Human Beings*, Wiener explored some likely effects of information technology upon key human values like life, health, happiness, abilities, knowledge, freedom, security, and opportunities (Bynum, 2015).

Wiener's findings were followed almost three decades later by many authors using the term computer ethics such as Walter Maner or James Moor. Maner stated in his article *Starter Kit in Computer Ethics*, (Maner, 1980) that, *"Wholly new ethics problems would not have existed if computers had not been invented."* And Moor in his article *What is Computer Ethics?* defined this area with the following:

"*In my view, computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology,* (Moor, 1985).

In a nutshell, computer ethics is focused mainly on new technology innovation, meaning new tools and machines: *"When computers interact with society they are causing new moral problems*." (Friedman and Nissenbaum, 1996). From the point of the involved stakeholders, computer ethics focuses on the relationship of public users and computer professionals and their professional ethics covering: *"Standards of good practice and codes of conduct for computing professionals."* (Gotterbarn, 1991).

On the other hand, the terms of information ethics and data ethics that have been used for the last decade by respected authors like R.Cappuro, L.Floridi and M.Taddeo focus on either more a philosophical approach (Capurro, 2006) or on the content at different levels of abstraction (Floridi, 2008). Content was originally considered different entities such as knowledge, information and data, although the latest developments have moved the focus on raw data that is the new target of our moral actions described as:

"The shift from information ethics to data ethics is probably more semantic than conceptual, but it does highlight the need to concentrate on what is being handled as the true invariant of our concerns. This is why labels such as 'robo-ethics' or 'machine ethics' miss the point, anachronistically stepping back to a time when 'computer ethics' seemed to provide the right perspective." (Floridi & Taddeo, 2016).

3.3 Information and Data Ethics

The foundation of modern information ethics was laid down at the end of the 20th century by Rafael Capurro and Luciano Floridi. Nowadays, this theme is further developed not by individuals but mainly by broader working groups concentrated at university labs with a special focus on data ethics such as Oxford University (Digital Ethics Lab, Floridi´s workplace), Utrecht University (Ethics Institute) and Vienna University (The Privacy and Sustainable Computing Lab).

It seems that the more ICT specific approach of Floridi prevails in the academic and also the business world, although I would like to mention also the more general philosophical approach of R. Capurro.

Capurro's broader concept of information ethics deals with digital ontology and the fundamental question of being. He follows Heidegger's conception of the relation between ontology and metaphysics. Capurro argues that information ethics does not only deal with ethical questions relating to the infosphere but with more general questions of being. This view is contrasted with arguments presented by Luciano Floridi on the foundation of information ethics related mainly to the infosphere. Floridi´s approach is considered by Capurro as a reductionist view of the human body as digital data overlooks the limits of digital ontology and gives up the ethical orientation. (Cappuro, 2006).

It seems that Capurro and Floridi do not understand each other in their approaches; however, both are going the same direction. I will further discuss Floridi´s approach because it is currently the leading view in data science, and at the same time, it is easy to understand, especially for the relevant area of Big Data ethics.

In his article *Information Ethics, Its Nature and Scope* (2006), Floridi suggested the unified approach towards information ethics that he calls macroethics. It consists of three arrows with information as a source, information as a product and information as a target. He also introduced the idea of a moral agent that can generate the information as a product and affect the information environment as a target (Floridi, 2006); (Sigmund, 2015).

The most recent definition of data ethics was done in 2016 by two Oxford academics, Luciano Floridi and Mariarosaria Taddeo, approaching the topic again as macroethics distinguishing the **ethics of data**, **algorithms** and **practices**.

This respected definition of data ethics in Level of Abstraction of data (LoA_D) was done in their article *What is data ethics?* (2016) as:

"In the light of this change of LoA, data ethics can be defined as the branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values). This means that the ethical challenges posed by data science can be mapped within the conceptual space delineated by three axes of research: the ethics of data, the ethics of algorithms and the ethics of practices." (Floridi & Taddeo, 2016).

3.4 Big Data Specifics

I see progress in the relevant data ethics areas described above in the latest work of Floridi and Taddeo; however, in my opinion, there are some Big Data specifics that need to be discussed in more detail.

By these specifics of Big Data, I mean the following:

- **Specific role of stakeholder groups** (organizations, users, state).
- **Use cases of Big Data,** (showing mainly positive benefits)
- **Demand for regulatory framework**
    - this is a reaction of society to Big Data implementation that is increasing information asymmetry. Awareness of this situation is spread mainly between individual users creating pressure through nation states and its citizen associations.
- **Conflicts and issues stemming from the clash between Big Data use cases and ethics**

- Ethics can be described on a different level as I will shortly do in the following paragraphs

The specific role of stakeholder groups means that there is increasing information asymmetry between the individual users that can be considered data poor[2] and big corporations that collect data about the individual users and can be considered data rich. To be data rich means to have data insight into many areas as all of society is getting "datafied"[3] and this data insight leads to many advantages in the form of new business opportunities. A strong role from nation states is expected to regulate this information asymmetry and balance the equal opportunities and basic human rights. However, the power of some corporations derived from their turnover is exceeding; in the cases of the biggest corporations such as Google, Facebook, Microsoft, Apple or Alibaba; the state budgets of many nation states. See the figures below to compare the state budget of European countries and the turnover of leading global corporations that collect data about their users.

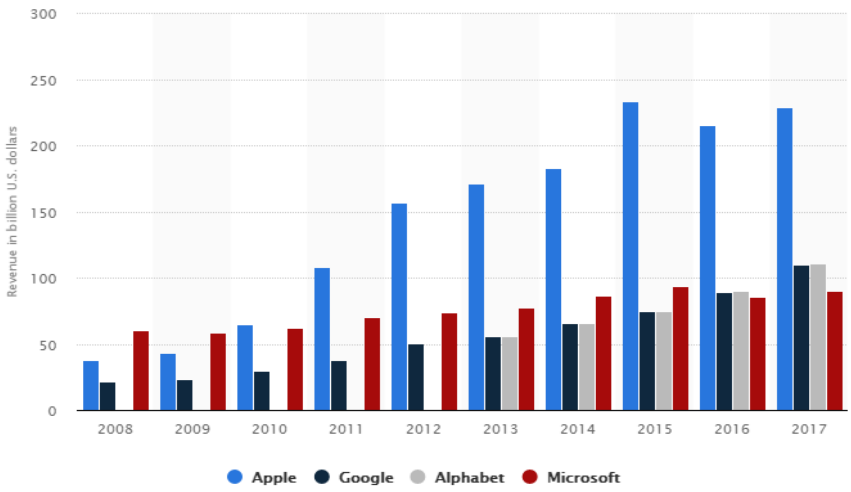


Figure 1 - Revenue comparison of Apple, Google/Alphabet, and Microsoft from 2008 to 2017 (in billion U.S. dollars), source: (https://www.statista.com)

[2] The term information asymmetry and terms data poor and data rich are used by many authors, for example, Boyd & Crawford, (2012)

[3] The term "datafication" means that we create digital data about almost every existing subject and was firstly defined in 2013 by K.N.Cukier and V. Mayer-Schoenberger. I will describe it in more detail in the following chapters.

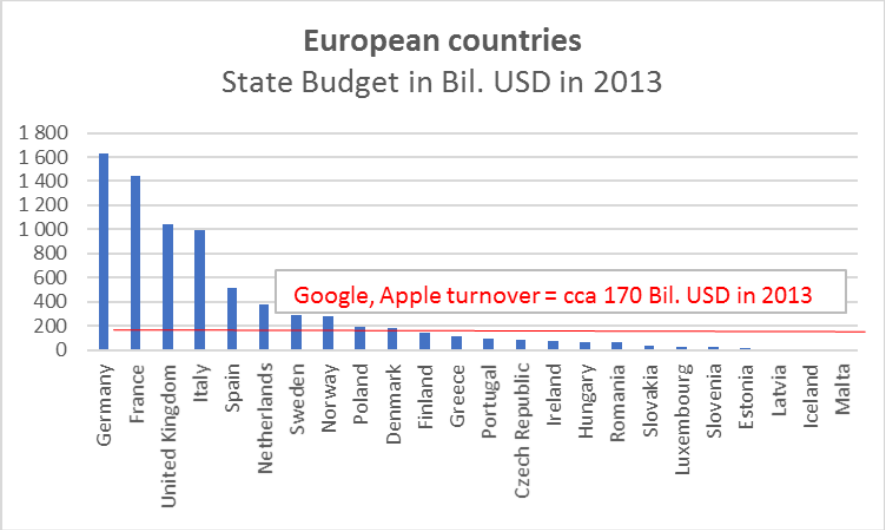


Figure 2 - State budget of European countries in billion USD for year 2013, (Wikipedia, 2019)

From the graphs above, it is clear that the economic power of the biggest corporations is extreme and we have to find new strategies and instruments to govern society where these extremely big corporations operate. This is a rather recent situation related to Big Data and the general digital economic boom accelerated in the 21st century and we have to think about facing this situation. I describe my view on the role of the state and possible ways to govern and regulate society in the era of Big Data in a special chapter, Regulatory Framework of Big Data.

Use cases demonstrate the new possibilities of Big Data technologies that are followed by generated issues (risks and issues) that highlight the specific role of stakeholder groups such as data rich organizations, data poor users that generate the data and the state that should regulate the whole data environment. I dedicate later in this thesis a special chapter describing use cases with a deep dive into specific areas of the telecommunication industry to show what is possible as a preamble to the chapter discussing Big Data issues and conflicts.

Issues and conflicts stemming from the clash of Big Data use cases with ethical values attract attention and create a demand for new regulatory frameworks that can cope with the current digital divide between data rich and data poor stakeholder groups. As noted above, there is a special chapter in this document dedicated to this topic.

I describe demand for regulative framework and possible ways to regulate society in a special chapter, Regulatory Framework of Big Data.

3.4.1 Ethics, Morality and Social Custom

To be able to discuss further the conflict of Big Data and ethics, I will shortly describe a view of the philosophy on this relevant area explaining a basic question: What is the relation of ethics and morality?

In 2016, in his book *Ethics, Life and Institutions*, Czech philosopher Jan Sokol introduces three different sets of rules upon which he explains the way individuals govern themselves:

1. social custom;

2. individual morality, and

3. ethics as a 'search for what is best'.[4]

First, Sokol talks about **social custom**; the everyday habits every society develops and that, on one hand, is the source of xenophobia – to say it more broadly, it is the reason for refusing the different features of other communities in general – but on the other hand, it also makes life in society easier, as it simplifies daily connections between individuals and makes them feel safe.

Second, he introduces **individual morality**, which leads an individual to restrain from certain actions, following religious norms, that may prevail over the multitude. Unlike social custom, individual morality is not automatically dependent on the consent of the majority, and it follows a different authority. A modification of the concept of individual morality comes with the development of law. The law then motivates individuals to act in a certain way, yet it never achieves perfection, leaving certain socially dangerous conduct out of its scope.

It is the third, Sokol argues, that best regulates relations in a society**: ethics as a search for what is best**. Individuals govern themselves the best when they have got a goal, an ideal to follow, or when they compare their actions with the conduct of others within the society.

3.5 Summary of Big Data Ethics Overview

I agree with the latest development of data ethics done in 2016 in the work of Floridi and Taddeo that is for me a logical evolutionary step from other authors such as Wiener, Maner, Moor, Gotterbarn that focused on the relation of men and machines (computer and professional ethics). I also believe that Capurro's theoretical work in the area of information ethics with great focus on the broader concept of information ethics and

[4] A similar distinction is introduced by Ricoeur, Oneself as Another, and after him Cornu, La confiance dans tous ses états, e.g. pp. 11, 38. Cf. also Ogien, L'éthique aujourd'hui, p. 16.

digital ontology contributed to the current founding landscape of data ethics done by Floridi and Taddeo.

Their approach is based on different levels of abstractions (LoA). This view of data ethics changes the focus from information to raw data. It offers a flexible and complex approach to the topic. Their ideas of dealing with data ethics as macro-ethics consisting of the **ethics of data**, **algorithms** and **practices** is a solid attempt to solve the complexity of this area avoiding the narrow and ad-hoc approach.

In spite of agreeing with Floridi and Taddeo's work in data ethics, I find Big Data ethics as rather different from general data ethics. Describing these specifics and their implication is, as I believe, one of my contributions to the scientific area of data ethics.

I named the Big Data ethics specifics in the chapters above, such as the specific role of stakeholder groups, opportunities arising from the new use cases, challenges (issues) and conflicts stemming from Big Data and also the new demand for a regulatory framework.

I will dedicate the following chapters to describing in more detail these specifics of Big Data ethics, keeping in mind the originally research question of the thesis that I will try to confirm in my own survey of Big Data and ethics.

4 Use Cases of Big Data

I will briefly describe and categorize the possible data sources and use cases showing what is feasible in the area of Big Data based on my previously published article at SMSIS conference in Ostrava (Pavlicek, Novak, 2015) and I will be very specific in the area of telecommunication. Besides the categorization, I will show examples from the perspective of mobile network operators such as geo-location use cases and will show overview of use cases in financial industry that are based on telco data. The majority of the content of this chapter I have already published together with my colleagues at the CONFENIS conference in Shanghai (Pavlicek, Doucek, Novak, 2017) and in the magazine Computer World (Novak, Kovarnik, 2015).

I will also show how the use cases from telecommunication can be relevant to other industries like the financial industry with the aim of showing what is possible in Big Data implementation that I further discuss and challenge in the following chapter: Digital Divide Conflict and Big Data Issues.

4.1 General Data Sources

When speaking about ethical data collection and data categorization or classification, we should primarily discuss data sources and data types in general and then focus on Big Data specifics.

Data sources, in general, are called **white**, **grey**, **not published** and **black.** However, as we will see later, only white data sources are candidates for ethical data collection when respecting the valid legislation. For deeper insight and further description, I recommend special literature because of the limits of this paper, I summarize new factors only briefly while using the description done by Z. Molnar in 2012.

White Data sources are officially published data, such as press releases, annual company reports, newspapers, magazines, official statistics and others. **Grey Data** sources are data that were not originally intended to be published but could be found in specific institutions or databases, such as SIGLE (System for Information on Grey Literature in Europe). A special category of sources is **Not Published Data,** which are mainly known only to its owner only and can be harvested by primary research via interviews and questionnaires.

There are also **Black Resources** that are outside of the public zone and are considered to be closed data sources and secret information, e.g. medical, financial, and governmental secret data. Harvesting such data is considered unethical and violates the law as well.

Speaking about White Data sources, we should not forget the worldwide initiatives called **Open Data** that is defined as follows:

"Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copy write, patents or other mechanisms of control" (Aberer, 2007).

The term "open data" itself is recent, gaining popularity with the rise of the internet and, especially, with the launch of open-data government initiatives such as *http://www.data.gov/* or in the Czech Republic *http://opendata.cz/*

The Open Data national initiative exists in the Czech Republic and is supported mainly by academics from the University of Economics in Prague and from the Faculty of Mathematics and Physics of Charles University in Prague.

4.1.1 Data Location and Structure

Based on its location, data can be divided into internal and external. Internal data means the internal data of the organization that is stored, managed and often also produced by internal processes, systems and people.

External data is located outside the organization. Organizations often monitor external data sources and try to import them if they find them important for their function and if they are able to manage their import to internal systems. External data is very important for Competitive Intelligence and also for Big Data.

Another view of data types stem from the distinction whether the data are structured or unstructured. Structured data have defined formats and can be stored to specific locations (table, row and column) of databases. With structured data, it is possible to perform a certain set of standard procedures, such as calculation, queries, reports and others.

Unstructured data is without a given format: such as plain text or documents of special format (images, videos, and CADs). The simple examples of unstructured data are documents in MS Office / Open Office. The unstructured data are usually handled on the level of files or objects and are not manageable by the standard Relational Database Systems as for example MS SQL, MySQL, et alia.

Though the table below summarizes it, such a standard view of data from the point of enterprises and their IT does not reflect all the new Big Data challenges that we will focus on in the next chapter.

Data Type	Internal source	External source
Structured	Enterprise databases and information systems as ERP, CRM etc.	Specific databases, catalogs, price lists etc.
Unstructured	Text files, emails, meeting minutes, special format report etc.	Web pages, social networks, emails etc.

Table 1 - Standard categorization of data types, (Molnár, 2012)

4.1.2 Big Data and Different Categorizations

Variety of data structure is one of the key components in the definition of Big Data. Big Data may be structured, unstructured or also semi-structured data that can be understood as a special sub-category in Table above.

Buneman's (1997) definition of semi-structured data: *"A form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data."*

As an example of semi-structured data, we can name human readable formats which are generally all texts in ASCII or special formats such as XML (eXtensible Markup Language) or JSON (JavaScript Object Notation).

They are special databases such as Hadoop or noSQL database engines that are designed to work with semi-structured and unstructured Big Data. For a detailed description of these technologies, see literature related to Big Data technologies, e.g., (Manyika et al. 2011).

The opposite of the human readable formats is machine readable data, which is data in binary formats or more complex structures that is readable by computers and machines in general.

From the point of view of Big Data categorization, which we are going to suggest, it is more important to know how the data is generated than if it is structured. For example, spatial and transactional data are typically produced by machines and equipment. The machines that produce data are linked to the billing systems, telematics systems or other systems that can store and manage the transactional data without any human interaction.

We may recognize the need for a new more scalable categorization of Big Data based on the criteria that combine an innovative Big Data source and the original purpose of its generation.

One possible view on Big Data categorization (classification) is based on the terminology of homogenous and heterogeneous networks as proposed by Jiawei Han, who is considered to be one of leaders of the Knowledge and Data Discovery (KDD) discipline (Han et al. 2012).

Homogenous networks, based on professor Han´s classification, connect subjects and objects with the same class of data as people with people in social networks or machine with machine in telematics applications so it is a projection of real life to the data networks that are easier for humans to work with. On the other hand, the heterogeneous networks are closer to real life as they connect people with their activities, peers, devices, generally everybody with everything.

For the purpose of this paper, I have decided to use my own categorization (classification) specifics for Big Data. Before we approach that categorization, we should summarize who or what generates the data or, in other words, what is the driver for the data explosion. Generally, data can be generated by people, machines or digital processes and data manipulation. New factors in the Big Data age are, for example, smart devices (machines) or social media and the internet in general (people). Hence, thanks to new applications, we face new ways of human interaction and communication. In the area of data processing and manipulation, we are currently in the situation where the costs of storage are so low that we simply may digitally store everything about our personal or business activities without having to erase anything. It means that the more data we have, the more we produce by its manipulation. All these new factors interact together and this has caused the data explosion known as the Big Data phenomenon. The simple summary of the description above is shown in Table 2 below.

Data Drivers Who / What generates data?	**Description of new Big Data factors**
People	Applications such as social media and others on the internet and Web 2.0 environment make people communicate more and interact and share rich content
Machines	Penetration of smart phones and smart devices (Internet of Things, IoT) has increased significantly in the last decade and these machines have produced an almost constant high volume of data with variable data structure
Digital processes & Data manipulation	Costs of storage is so low that we digitally store everything about our activities and erase nothing so the more data we have, the more we produce by data manipulation

Table 2 - Data generators and new Big Data factors (Pavlicek, Novak, 2015)

The more detailed categorization that we propose clusters the variable Big Data generators to the logical categories based on the similarity of the original data source. The suggested categorization has some overlaps between categories, but it was created with the purpose to group generators of data into the logical categories with respect to the Big Data specifics and also with respect to possible further data processing or commercial usage. The suggestions of such Big Data categorization are found in Table 3 below.

ID	Data sources	Examples	Description and Purpose
1	**Smart devices**	Phones, TVs, Fridges, IoT	Devices that are online and can generate data and interact with other users and devices
2	**Social networks**	Facebook, Twitter, LinkedIn, YouTube, Instagram	Connect people with other people to entertain and exchange information and content
3	**Spatial and transactional systems**	GPS, Payment systems, Telemetric systems	Locate subjects and objects, inform about systems and machines status and generate transactions
4	**Corporate systems**	ERP, CRM, CMS, DW	Collecting business data about customers, employees, products, technologies and their attributes
5	**Systems with special securing efforts**	Public authorities, healthcare systems, security systems	Collecting private and sensitive data about persons, assets, finance and other data that are stored and managed with special security efforts
6	**World Wide Web - Internet**	Content of Internet designed for easy access of its users	System of interlinked hypertext documents accessed via the internet. With internet tools such as search engines, it is not only very easy to consume content but also to contribute to it
7	**Media production tools and equipment**	Cameras, Dictaphones	Video, audio, voice records creation, production and manipulation for private and business purposes
8	**Special sources**	CAD, emails, OCR	Special digital activities such as Computer Aided Design (CAD) that, thanks to humans or machines, generate data that can be shared immediately (online) or stored offline and shared later

Table 3 - Big Data categorization, (Pavlicek, Novak, 2015)

For legal reasons, it is essential to know what the content of such data sources are and whether any personal or sensitive data in the Big Data categories can be found.

Personal data in the Czech Republic follows EU regulation and is defined in the General Data Protection Regulation (GDPR). Under the definition of this regulation, a name and surname, birth identification number, address (residential but also IP address) or a telephone number of a natural person and many others fall within the scope of personal data and therefore shall receive protection from the law.

Table 4 below summarizes the Big Data sources and, through the main drivers for data generation, reflects also the possibility of such categories to contain personal and sensitive data regulated in the GDPR.

ID	Data sources	Main driver for Data	Can contain personal or sensitive data?
1	**Smart devices**	Machines	Yes
2	**Social networks**	People	Yes
3	**Spatial and transactional systems**	Machines	Yes
4	**Corporate systems**	Digital processes and Data manipulation	Yes
5	**Systems with special securing efforts**	All	Yes
6	**World Wide Web - Internet**	People	Yes
7	**Media production tools and equipment**	Machines	Yes
8	**Special sources**	All	Yes

Table 4 - Big Data categorization related to its drivers and personal data, (Pavlicek, Novak, 2015)

As visible from the above, **all data categories** can contain personal or sensitive data either directly or indirectly through the context or combination of multiple data sources. This is an important finding for the legal consequences it might have.

4.1.3 Summary of Big Data Sources

The articles above proposed the categorization (classification) specifics for Big Data. It took into consideration the source of the data (Data Drivers / Who / What generates data? – on the scale People /Machines / Digital processes & Data manipulation) and new Big Data factors. My colleague A.Pavlicek and I, presented in our article (Pavlicek & Novak, 2015) a more detailed categorization of clusters of variable Big Data generators to the logical categories based on the similarity of the original data source (Smart devices, Social networks, Spatial and transactional systems, Corporate systems, Systems with special

security efforts, World Wide Web – Internet, Media production tools and equipment, Special sources). It was created with the purpose to group generators of data to the logical categories with respect to Big Data specifics and also with respect to possible further data processing or commercial usage. And finally, we also mapped possible collisions with the GDPR.

4.2 Technologies and Methods of Big Data

There are many technologies classified as Big Data. Some of them have been already verified and are currently used in business, whereas some are promising but not mature enough yet. We can find a good overview of these technologies in the many[5] Gartner Hype Cycles named as emerging technologies such as: Internet of Things, Artificial Intelligence and Data Management among others, (Gartner, 2019).

Among the already mature and currently used technologies belongs, e.g. web analytics, predictive analytics, social media monitoring, speech recognition, MapReduce approach (Hadoop), video search, content analytics and others.

There is also the broad scope of available commercial or open source tools and products focused on special areas of different IT components such as: databases, discovery and visualization tools and special Big Data analytics tools[6] among others. There is visible trend to move the specialized Big Data tools to Cloud environment provided as Software as a Service (SaaS); however, the high data volumes and other specifics of public organizations and enterprises are still a challenge. Although, the offer, scalability and the variability of the Big Data products that meet the different needs is really huge now.

Since I am limited by the range of this section, I will only demonstrate more closely the differences between the structured Relation Database System (RDBS) and the non-structured Hadoop[7] DBs and also shortly describe speech recognition that were

[5] Betsy Burton, analyst from Gartner clarified in 2015. "We've retired the big data hype cycle. I know some clients may be really surprised by that because the big data hype cycle was a really important one for many years. But what's happening is that big data has quickly moved over the Peak of Inflated Expectations," she continues, …and has become prevalent in our lives across many hype cycles. So big data has become a part of many hype cycles." Source: www.gartner.com

[6] E.g. on specialized web https://softwareconnect.com/big-data-analytics/ where Big data analytics tools are defined as the following: "Let end-users analyze huge volumes of transactional data and other sources that may be left otherwise untapped by conventional business intelligence programs…", are listed 17 recommended products such are: Arcadia Enterprise, TIBCO Spotfire, Periscope Data. Qlik Sense, Tableau, Looker, Domo, Sisense, Power BI from Microsoft, Oracle Analytics Cloud, Dundas BI, ClicData, BOARD, CALUMO, Yellowfin BI, Datawrapper and Halo BI.

[7] „Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware using MapReduce algorithms. It provides massive storage for any kind of data,

considered as the leading technological applications that introduced the Big Data phenomenon to the wider audience of IT specialists and also IT users.

4.2.1 Hadoop Technology

For better understanding of the difference between RBDS and Hadoop, see the table below. The table is based on the analysis of companies Teradata and Hortonworks.

RDBS	**Hadoop**
Stable scheme	Evolving Scheme
Leverages Structure Data	Structure Agnostic
ANSI SQL	Flexible Programming
Iterative Analysis	Batch Analysis
Fine Grain Security	N/A
Cleansed data	Raw Data
Seeks	Scans
Updates /Deletes	Ingest
Service Level Agreement	Flexibility
Core Data	All Data
Complex Joins	Complex Processing
Efficient Use of CPU/IO	Low Cost Storage

Table 5 - Comparison of RDBS and Hadoop (Hortonworks, Teradata, 2013).

The possible deployment and benefits of Hadoop can be seen, e.g., in marketing units where companies communicate with their customers through different online channels. Thus, they can benefit from the combination of the open source Hadoop database, where, e.g., social media un-structured data are stored, and CRM (Customer Relation Management) RDBS, where structure data related to their products and customers can be found.

enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.“, (SAS Institute. 2019, approached at: https://www.sas.com/nl_nl/insights/big-data/hadoop.html)

Another opportunity to use Hadoop is to use it in operation units where, e.g., the utility industries have engaged many field equipment generating logs, statuses and geographic coordinates. All that, in combination with the relations databases such as CRM or inventory DBs, can predict fraud or future behavior like outages or other characteristics.

In commercial business, the implementation of Hadoop is expected to be combined with standard RDBS and specific corporate modules such as: CRM, ERP (Enterprise Resource Planning), Inventory DBs and some other.

Industries with the biggest potential and the highest possible benefits from the improved KPIs (Key Performance Indicators), are industries where a big number of customers or customers´ technological data are stored. Among these industries undoubtedly belongs e-Commerce, Retail, Telecommunication, Utility and also Media and Finance Industries.

4.2.2 Speech Recognition

Human communication via speech is the most natural way people communicate. Therefore, in case we want to use computing and digital technologies to work with the human voice, we need to be able to convert human speech into the language of computers and their algorithms. For that, we have to deploy technologies of discrete samples stored in the media memory and moreover, we have to be able to convert speech into the text and vice versa. There are many more features connected to voice communication, such as the possibility to recognize a speaker only by the help of voice biometry, to identify the language that speaker is using, to recognize his emotions and many others (Cenek, 2012).

The technologies related to speech recognition have been developing for the last few decades and lot of progress has been made in the last few years. Nowadays, they seem to be close to their maturity thanks to the help of available computing power and applied mathematics. Nonetheless, we should differentiate between the online speech recognition and offline post-processing of the stored voice data, because there is a big difference in demand for the IT resources in these two different tasks.

To name a few ways of the possible application of digitalized speech technologies in the business environment, I can mention voice-activated personal assistants, call centers and their online IVRs modules or call agent behavior evaluation based on key words or emotion. The detection of key words can be used also for the intelligence service and national security, as we could see recently during the affair related to the PRISM program of National Security Agency in the US, (Marshall, Edward, 2013).

If we use such technologies in marketing units, we can benefit from the correlation of customers emotions analyzed during customer calls and their pairing to CRM data. This helps to segment customers based on their emotion and that can be used later, e.g., for targeted sales campaigns suggesting new products to customers with positive emotions. Another example is to start the churn prevention activities focused on customers with negative emotions identified during their speech with call centers.

4.2.3 Big Data Technologies and Business

The possible deployment of Big Data technologies in business life strongly depends on its commercial benefits that can be measured. The deployment of innovative technologies like Big Data are nowadays related to the improvements of corporate performance of the whole company measured by KGIs (Key Global Indicators) and KPIs (Key Performance Indicators). An improvement in KPIs may be the potential outcome from the introduction of new technologies such as speech recognition, and it should contribute to a positive change of KGIs.

With respect to the above mentioned, we deploy and benefit from these technologies only if smaller elements in the functional unit of IT have an effect on higher levels of the organization and causes the measurable benefits of globally observed KGIs.

When the technologies do not bring the KGIs benefits or any competitive advantage in commercial life, they are left in the lab as prototypes and they are not introduced to mass production in line with the rule above. For detail strategy how to implement Big Data technologies into business environment see my special article in journal of System Integration (Novak, 2014).

4.2.4 Big Data Technologies and Individuals

In our lives, we do not act just because there may be a direct commercial benefit out of it, but sometimes we also act because we want to feel good or entertain ourselves or, more generally, we want to satisfy our needs as they were defined, for example, by Adam Maslow in his theory of human motivation (Maslow, 1943).

The dissemination of the Big Data phenomenon in human life is nonetheless a bit viral. Since we want to satisfy our needs and information can do that plus, information is very easy to get with the help of search engines, we are facing an information flood which surpasses our limit to control it. It can mean privacy intrusion or confusion in world understanding that we will discuss in later chapters when naming all the possible Big Data issues.

4.2.5 Summary of Big Data Technologies

The development of technologies classified as Big Data or Artificial Intelligence is enormous and new inventions are expected. Thus, it can be concluded that Big Data technologies and methods can definitely make our business life more efficient. However, when it comes to the life of the individual, Big Data will not make our world better unless we start applying critical thinking and use the system approach to process them instead of simple data consumption.

The following chapter describes the real examples how to use some Big Data technologies in one specific industry of Telecommunication. It follows previously described Data sources and the Hadoop, Cloud computing and Predictive Analytics technologies are the main enablers that support the use cases mentioned further on.

4.3 Telecommunication and Big Data

The telecommunication industry is very specific in terms of its network assets covering Fixed, Mobile networks and also some other specifics that are described in the figure below.

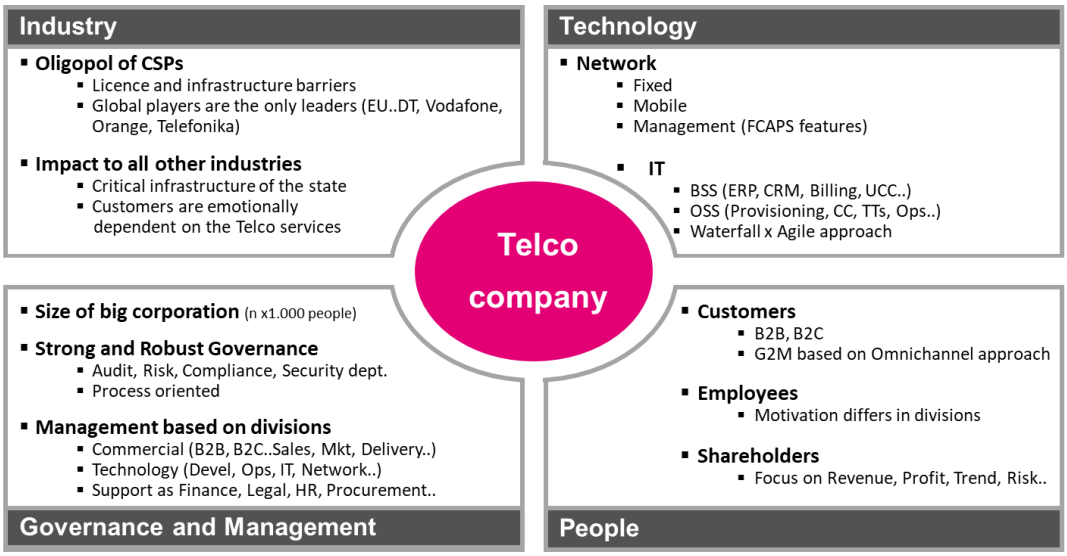


Figure 3 - Description of Telco specifics (author)

For data collection about individual users, I will describe in more depth the very important technologies and processes related to mobile network operators in the following paragraphs with inspiration and examples from T-Mobile Czech Republic that were officially published in the two following sources Pavlicek, Doucek, Novak, Strizova (2017) and Novak, Kovarnik, (2015).

4.3.1 Mobile Network Operator Introduction

Modern smart phones have become ubiquitous communications tools—now used not only for phone calls and text messages but also for accessing the internet, taking pictures and recording videos with integrated camera, navigating with GPS or watching videos and playing games. The proliferation of mobile phones amongst the general population is immense. The percentage of active mobile SIM cards[8] within the population reached 96% in 2014, (ITU, 2014). In developed countries, the number of SIM cards has surpassed the total population, with a penetration rate now reaching 121%, whereas, in developing countries, it surpassed 85% and keeps growing.

Analyzing the spatiotemporal distribution of phones geolocated to the base transmitting towers (BTS) may serve as a great tool for population monitoring. With data being collected by mobile network providers, the prospect of being able to map the changing human population movement and distributions over relatively short intervals (while preserving the anonymity of individual mobile users) paves the way for new applications and a near real-time understanding of patterns and processes in human geography (Deville, P. et al., 2014).

4.3.2 Mobile Phone Location Technology

Mobile network operators (MNO) must be aware of the geographic location of each mobile phone in the network in order to be able to route calls to and from them and to seamlessly transfer a phone conversation from one base station to a closer one as the user is moving. This originally technical necessity was transformed into a commercial opportunity to increase the Average Revenue Per User (ARPU) through what is now known as 'Location Based Services' (LBS). LBS are all services that use the location information of a mobile device to provide a user with location-aware applications and services. Such location information can be provided by the mobile network operator, the mobile phone device, or a combination of both, but this thesis focuses on the former.

The initially proposed LBS applications were very broad, creative and raised quite a lot of expectations. For example, users were offered the possibility to make requests like 'where is the nearest…?' (hospital, gas station, bank, restaurant, etc.), identify friends that walk nearby (Foursquare), ask for navigation instructions when lost (Google maps), locate lost phone (device locator), or receive a promotion from a familiar store when walking past it (location based 'spam') (Mateos & Fisher, 2006). Nevertheless, LBS failed to deliver its

[8] Subscriber Identification Module (SIM), widely known as a SIM card, is an integrated circuit that is intended to securely store the international mobile subscriber identity (IMSI) number and its related key, which are used to identify and authenticate subscribers on mobile telephony devices (such as mobile phones and computers), (Wikipedia, 2019). To simplify it for the purpose of this paper, we can consider one SIM card to be equal to one mobile terminal.

promises at the turn of the century, and its huge forecasted market potential did not come to reality (Zetie, 2004). This is partly because early services have been very restricted due to the poor location accuracy available, and the limited capabilities of both the handheld hardware (screen size and quality, processing power and storage capacity) and the network data transfer speeds and bandwidth (Mountain & Raper, 2001) (Mateos & Fisher, 2006). However, the second wave of geolocation services is coming right now, and this time it seems like it is here to stay.

Mobile networks are composed of cells around a BTS. Each active mobile phone, therefore, can be located by triangulating the geographic coordinates of its BTS. This network-based positioning method is simple to implement, phone and user independent, and its accuracy depends directly upon the network structure; the higher the density of towers, the higher the precision of the mobile communication geolocalization, (Mateos & Fisher, 2006).

Records of the time and associated cell of anonymous mobile phone users are valuable indicators of human presence and offer a promising alternative data source for increasing the spatial and temporal detail of large-scale population datasets, (Deville, P. et al., 2014). Mobile phone geolocation can be therefore used to:

- observe human mobility patterns at the individual level (police and security services only),
- monitor movements and activities of selected population using aggregated data,
- improve responses to disasters and conflicts, (Skrbek & Kvíz, 2010).
- plan epidemics elimination strategies,
- explore traffic flows and prevent traffic jams,
- study intensity of human activities at different times,
- identify seasonality in both domestic and foreign tourist numbers and destinations.

Legislation in the USA and EU also requires mobile network operators to provide an accurate location for calls to emergency services.

4.3.3 Geolocation in T-Mobile Czech Republic

T-Mobile is the largest Czech mobile network operator, which is in regular contact with about six million SIM cards (40% market share) with an aggregate data rate of hundreds of millions of signal records generated daily. T-Mobile had decided to take full advantage of the Big Data and geolocation potential and over the last three years has developed a

series of unique solutions that add value to the customer and provide a competitive edge for the company. In this paper we present a sample of the most interesting solutions. But first, let's look into some definitions.

Data anonymization

Every geolocation project starts with anonymization or pseudonymization[9]. The legislation of the Czech Republic and EU stipulates that it is always necessary to perform data anonymization or pseudonymization before data processing, thus preventing the identification of individual end-users. T-Mobile uses sophisticated encryption algorithms to remove identification and uses aggregated data for processing, so only meta data arise in the calculations, which are the only ones used to interpret the results later.

Technological background

The source of T-Mobile's geo-mobile data is residual signaling data from mobile cell identification, which makes it possible to know the approximate location of the mobile terminal and thus the distribution of the population in space and time. Further refinement of the position can be calculated if needed. Signaling data arises from typical mobile events such as a call, data transmission, SMS message, terminal transfer between individual transmitters, or upon a report call to the infrastructure in the so-called periodic specification when the terminal is periodically called for a signaling response. Data from signaling (after anonymization) can be stored in the data warehouse for further processing using classic business intelligence tools or special IT tools such as Hadoop and others supporting large data.

Continuous online monitoring system

The current distribution of mobile devices can be mapped through residual signaling data. A random but quite representative pattern of the Czech population's mobility can be recalculated in real time into aggregated geodemographic matrix of mobility. Based on both global and local system calibration (according to control check-points), they are recalculated to represent the real number of persons in each area. Specialized software allows displaying the distribution of the population in nearly real time, as well as historic time lapse sequences.

---

[9] „Pseudonymization and Anonymization are different in one key aspect. Anonymization irreversibly destroys any way of identifying the data subject. Pseudonymization substitutes the identity of the data subject in such a way that additional information is required to re-identify the data subject.", see source at: https://www.protegrity.com/blog/pseudonymization-vs-anonymization-help-gdpr.

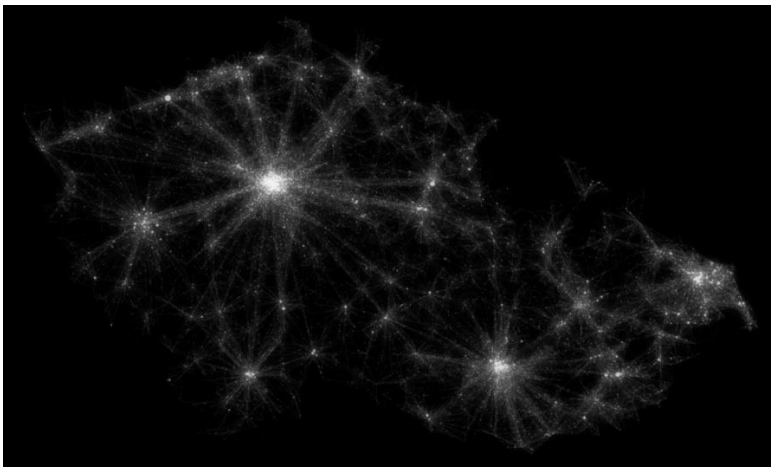


Figure 4 - Online monitoring visualization – movement of population in the Czech Republic (author)

Figure 5 - Example of online monitoring visualization – detail (author)

Business intelligence and big data tasks

Some typical business intelligence and big data processing tasks that need to be handled when working with anonymized data exported from signaling to a data warehouse are as follows:

- Keep, search and archive records of terminals in a given area.
- Position these terminals in required geographic formats such as centroid, square, cadaster, or any given polygon.
- Deal with signal skipping between neighboring BTS.
- Deal with the problems near international border areas (roaming).
- Store the number of people using mobile phones in a given area in a specific time slot, together with time-lapse data.

- Manage algorithms to count unique terminal approaches versus cumulative access to all terminals.
- Identify the origin and destination matrix, which is important for determining the motion vector.
- Compute the whole population to allow other data layers calibration.
- Solve the non-homogeneity of data in some areas.
- Create enhanced models in locations where network topology does not meet the requirements in terms of precision.
- Modal split, that is to distinguish the movement of the population from the point of view of transport, such as public transport (train, boat or bus) or individual transport.

Fields of application

Mobile geolocation is successfully used in a number of cases, the most common uses being:

- Crisis management (lost children, information on people in the area of fire, floods or chemical threats) (Skrbek, 2009).
- Detecting population mobility for state infrastructure and urbanization planning (new roads, P+R areas, public transport, land use plans).
- Commercial statistics (number of visitors to shopping centers, outdoor, festivals, tourism and city and area, replacement or supplementation of Czech statistical office research).
- Optimizing traffic flows.
- Location-based services such as a mobile ad for nearby services.

The list of examples would be unlimited with the possibilities enriched by other external data (weather, social networks, CRM systems, etc.) taken into consideration.

4.4 Use Cases: Big Data geolocation analysis

T-Mobile is very active in the big data field and, together with partners from the academic and commercial sectors, is involved in a wide range of research and commercial projects. The following paragraphs demonstrate the above-mentioned possibilities on the practical examples of implementation by the biggest mobile network operator in the Czech Republic.

4.4.1 Use Case - Pilot Case Study of Šumava National Park

The objective of this pilot tourism-oriented project was to calculate the number of visitors in the Lipno, Kvilda, Modrava and Horská Kvilda regions at the turn of 2013 and 2014, to find out where visitors came from, how long they stayed and which places they visited. Such data is useful for the national park administration, municipalities and local businesses - hoteliers, restaurateurs, sports facilities, and other entities. They get to know and possibly target the tourists of the Šumava Mountains. The National Park Service, in cooperation with T-Mobile and KPMG, prepared the long-term geolocation analysis.

And what are the results of the case study? (TTG, 2014). Most foreign visitors arrived from the Netherlands (36%), Germany (35 %), Austria (6 %) and Russia (5 %). In total, 260,000 visitors arrived in Šumava during the monitored period, of which 24% were from abroad.

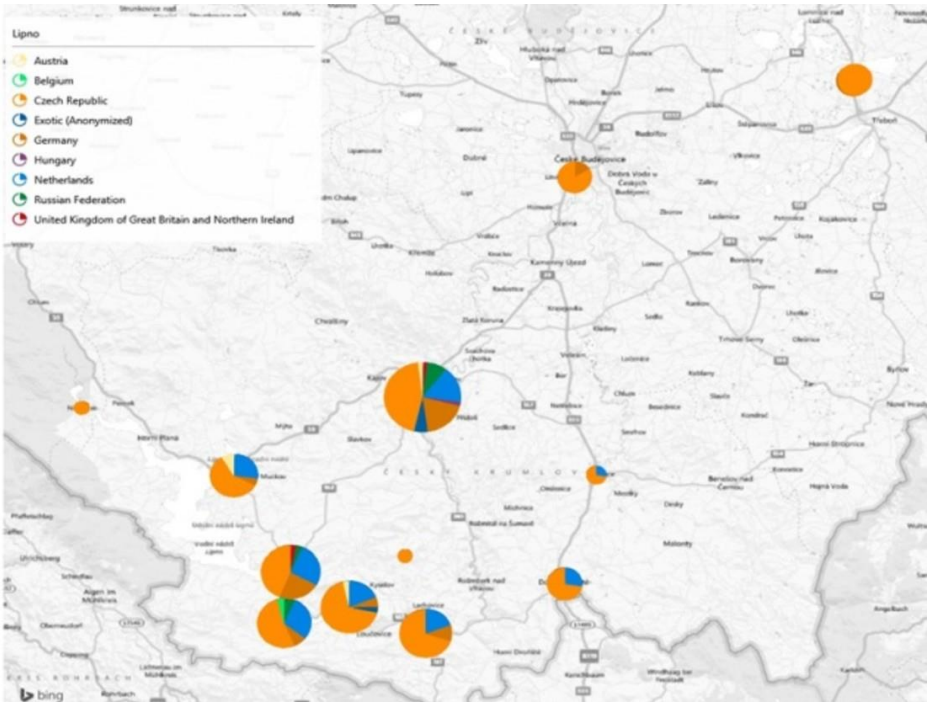


Figure 6 - Distribution of visitors in Šumava National Park (Doucek, Pavlicek, Novak, Strizova, 2017)

The main destinations for both domestic and foreign tourists are usually near (< 10 km) the place where they spent the night. One day trip was made by more than 1,000 (0,5 %) domestic tourists and 700 (1,2 %) foreign tourists. According to KPMG's analysis, tourists spent more than 211 million CZK during the two-month survey period of the region, of which about 70% accounted for domestic and 30% for foreign tourists.

4.4.2 Use Case – Czech Ski Resorts

Based on a very positive response to the pilot study, a very thorough analysis of the behavior of visitors to Czech mountain resorts was conducted in 2015. Six top Czech and Moravian ski resorts were analyzed: Harrachov (Giant mountains), Pec pod Sněžkou (Giant mountains), Rokytnice nad Jizerou (Giant mountains), Špindlerův Mlýn (Giant mountains), Kohútka (Javorníky), Lipno (Šumava)

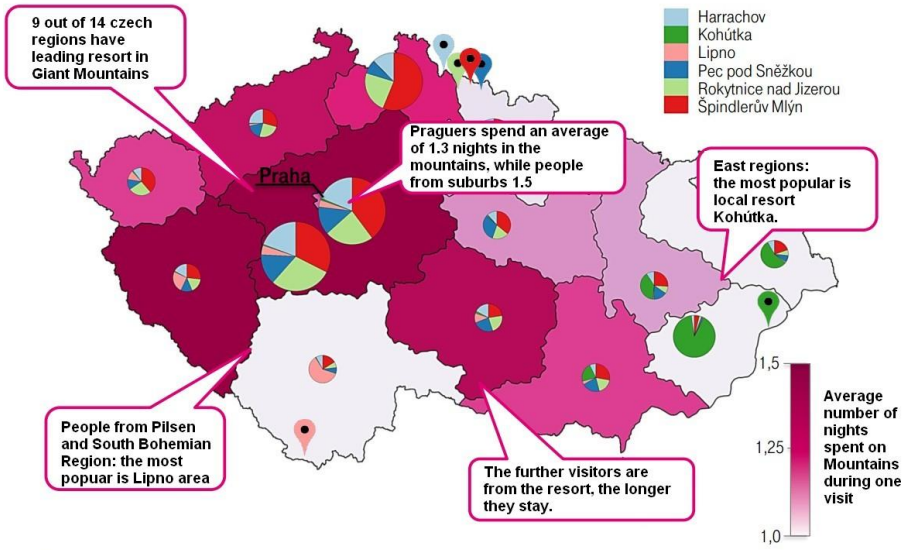


Figure 7 - Czech mountain ski resorts – origin of visitors, (Doucek, Pavlicek, Novak, Strizova, 2017)

The data from the mobile network was only one of the sources - it supplements the information about the profile of the visitors, obtained by a questionnaire survey, and the sample survey of the Czech population.

The results answer some important business related questions such as: What are the visitors of the mountains doing and what do they expect? How many one-day visitors are there and how many tourists sleep in the mountains? Do they differ in behavior? When do the Slovaks begin to travel to the mountains? What services are missing the most numerous visitor groups? That is valuable information, which would be otherwise hard to get.

The most interesting findings are presented in Figures 7,8 and 9.

Figure 8 - Czech mountain ski resorts – What is the most visited resort in The Czech Rep.? (Doucek, Pavlicek, Novak, Strizova, 2017)

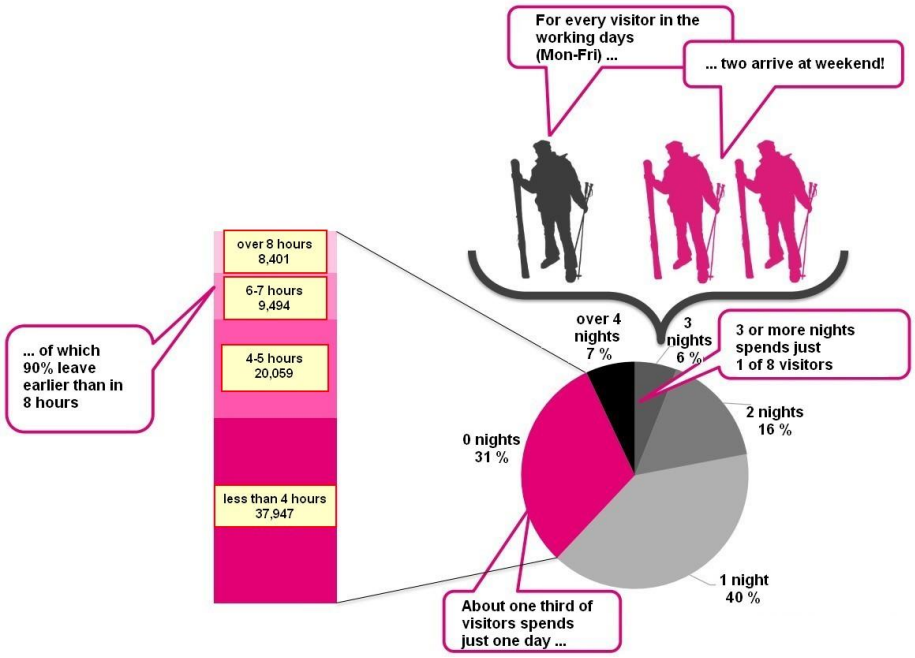


Figure 9 - Czech mountain ski resorts – length of stay, (Doucek, Pavlicek, Novak, Strizova, 2017)

4.4.3 Use Case - Mobile Data and the City Territorial and Development Plan

The use of geolocation data is not limited to the commercial sector. Public administration also has a number of tasks, where geolocation can be useful. The data make it possible to map the mobility of inhabitants in detail and better understand the mobility dynamics. Detailed monitoring of mobility enables the classification of territory in terms of public needs and helps with the sustainable development of communities.

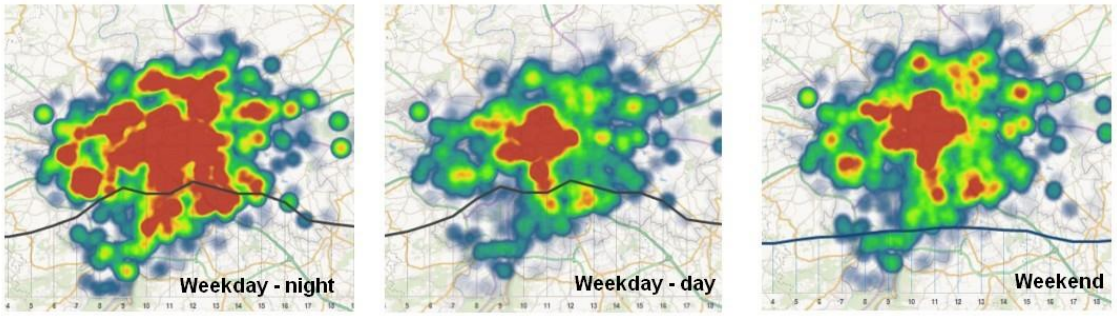


Figure 10 - Use of mobility data for city development coordination – density of population day, night, weekend, (Doucek, Pavlicek, Novak, Strizova, 2017)

Data can help quantify in detail mobility links between the city and its wider surroundings and help, for example, to plan public transport, housing or new schools and amenities accordingly (see Fig. above).

Figure 11 - Average distribution of the inhabitants of Černý Most area (part of Prague) in the daytime (typical weekday). Purple - inhabitants at home. Orange - inhabitants traveling outside of his / her home, (Doucek, Pavlicek, Novak, Strizova, 2017)

4.4.4 Use Case - Václav Havel Airport Prague

Václav Havel Airport Prague, formerly Prague Ruzyně International Airport (IATA: PRG, ICAO: LKPR) is the international airport in the capital of the Czech Republic. It is, with over

13 million passengers in 2016 (over 15 million expected in 2017), the busiest airport in the newer EU member states. It serves as a hub for Czech Airlines, Travel Service, SmartWings, Wizz Air, and Ryanair. Its 2 runways can handle 71,000 t of cargo and 137,000 aircraft movements per year.

The survey was conducted from July 2016 to May 2017 and recorded 12.2 million passengers - 6.218 million arrivals (of which 1.856 million Czech and 4.361 million foreigners) and 5.987 million departures (of which 2.071 million Czech and 3.915 million Foreigners). A significant seasonal component appeared in the airport's operations.

Days	1	2	3	4	5	6	8	9	10	11	12	13	14	more
Percentage	12%	20%	26%	15%	6%	4%	2%	1%	1%	1%	1%	1%	1%	6%

Table 6 - Length of stay in the Czech Republic, (Doucek, Pavlicek, Novak, Strizova, 2017)

Although these basic statistical data can be gathered by counting arrivals and departures and aircraft occupancy capacities, the use of mobile geolocation brings additional, enhanced capabilities. It is possible to monitor the movement of passengers in the Czech Republic or the destination of Czechs abroad. We can, for instance, easily detect the length of stay of foreigners in the Czech Republic (see Table above), or even check, which UNESCO monuments they decided to visit (see Table 6). The analysis can go even deeper – we can identify the day of UNESCO visit or even in what order they have been visited. For each monument, we can calculate its popularity by different nationalities (Russians seem to prefer Cesky Krumlov and Kutna Hora – they constitute 44.1% and 44.8% of foreign visitors there, Americans like to go to Lednice – 40.4%)

UNESCO site	Prague castle	Cesky Krumlov	Kutna Hora	Telc	Olomouc	Tugendhat	Holasovice	Litomysl	Lednice	Kromeriz
%	85,87%	7,92%	3,98%	0,36%	0,35%	0,35%	0,32%	0,30%	0,26%	0,10%

Table 7 - UNESCO sites visited by airport passengers, (Doucek, Pavlicek, Novak, Strizova, 2017)

Last but not least, it is also possible to analyze the movement of the visitors in the defined localities. Figure 11 displays the distribution of tourists from Russia, Germany, and Italy during 24 hours in Prague. Data allow measuring quantity and distribution of foreign visitors in the targeted area. Thanks to traffic data, it is possible to identify the potential for further development.

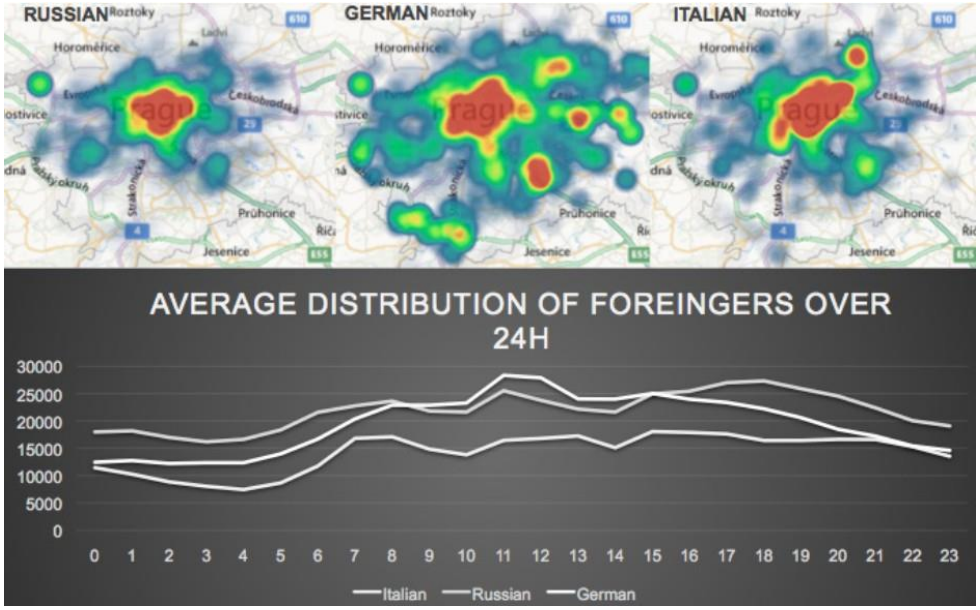


Figure 12 - Distribution of Russian, German and Italian tourists in Prague, (Doucek, Pavlicek, Novak, Strizova, 2017)

Mobile phone geo-monitoring also allows the determining of the catchment area of Prague airport – as shown in Fig. 13, it mainly serves the central Bohemia region. Mobile monitoring can be (to some extent) used even abroad. Figure 14 pictures the destinations of Czech travelers – flying from V. Havel airport – categorized by the length of their trip.

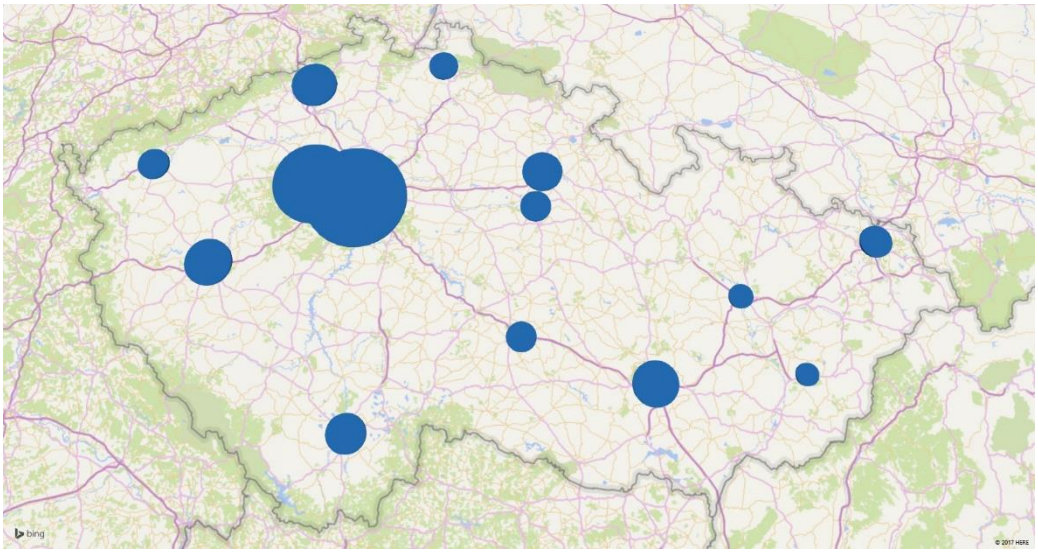


Figure 13 - Origin of Czech travelers – departures from V. Havel airport, (Doucek, Pavlicek, Novak, Strizova, 2017)

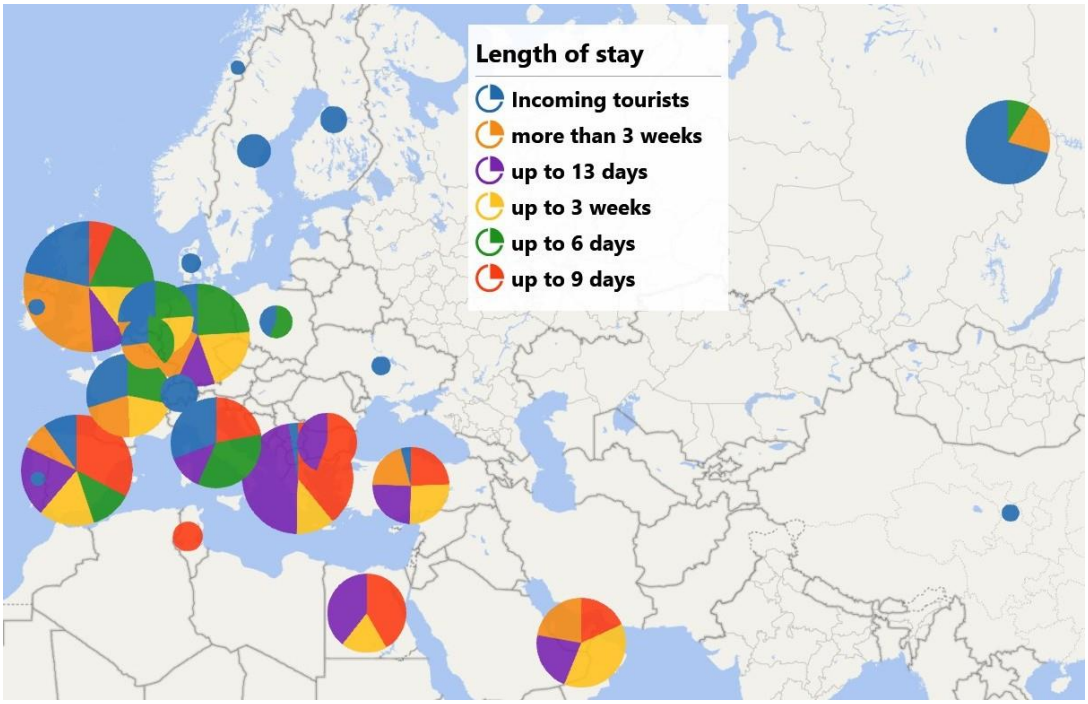


Figure 14 - Destinations of Czech travelers – departures from V. Havel airport and length of the trip (Doucek, Pavlicek, Novak, Strizova, 2017)

4.5 Use Cases in Financial Industry That Are Based on Mobile Data

The figure below shows the real examples of use cases from the financial industry that are based on Big Data from a telecommunication provider. It is visible that there is strong multi industry potential of Big Data coming from telecommunication.

Credit Scoring	Telco provides credit scoring based on behavioral data We covers even New-To-Market and New-To-Bank customers
Churn prediction	To predict churn of the customers and mitigate it early is essential Trigger is the behavior on the internet, e.g. competition web site visit SMS/MMS marketing and retention campaign targeting
Fraud prediction	Customer background check based on Telco data (Billing, CRM..) Suspicious customer internet behavior analysis
Branch network optimization	Optimal branch placement in relation to customer distribution Locality preselection for new branch opening
Customer targeting (One to One marketing goal)	Based on customers current locality (near to branch, airport, etc.) Specific product offers based on customer knowledge e.g. the internet activity and sociodemographic profile

Figure 15 - Examples of use cases in financial industry that are based on telco data (author)

4.6 Use Cases Conclusions

The paragraphs above presented real-world studies conducted in the Czech Republic that combined mobile device geolocation, different data sources and Big Data approaches. Despite the relative youth of this field – the first studies started less than 5 years ago – it is becoming a standard tool for a near real-time understanding of patterns and processes in human geography. It allows the observation of mobility patterns and the study of the intensity of human activities at different times and places. The data thus obtained are very accurate, up-to-date, and can bring previously unpredictable views and facts.

Following the previously mentioned use cases, with examples from the telecommunication industry, I can now derive a more general list of positive and also negative factors related to Big Data implementation.

Among the positive, I can briefly name the following:

Insight into complex problems, new business opportunities, increased security, decreased costs thanks to better analytics, speeding up complex decisions or more targeted and personalized offers, and more.

Among the negative, I can briefly name the following:

Invasion of privacy, decreased civil freedom, increased state and corporate control, opportunity imbalance stemming from information asymmetry, a new digital divide forming a new hierarchy in society, loss of objectivity and more. The above factors will be discussed in more detail in the chapter Big Data issues with a special focus on the negative factors as, in the near future, further rapid development and massive expansion of Big Data implementation can be expected.

5 Regulatory Framework of Big Data

The American lawyer and professor of law at Harvard university, Lawrence Lessig, described several causes of possible regulations of cyberspace such as the market, legislation, social norms and architecture in his book *Code and Other Laws of Cyberspace,* (1999).

"Our choice is not between "regulation" and "no regulation." The code regulates. It implements values, or not. It enables freedoms, or disables them. It protects privacy, or promotes monitoring. People choose how the code does these things. People write the code. Thus, the choice is not whether people will decide how cyberspace regulates. People--coders--will." (Lessig, 1999).

Although in computer science, "code" typically refers to the source code of a computer program, in law, "code" usually refer to the texts of the valid legislation. In his work Lawrence Lessig explores the ways in which code in both senses serve as an instrument for social control, leading to his dictum that "Code is law."[10]

I was inspired by the general structuring of the possible regulative framework and I have updated this approach for Big Data and defined my specific structure listed below that is inspired by Lawrence Lessig's view.

- **Market**
    - General principals how to govern society
- **Social norms and human values**
    - Human values and rights, e.g., (European parliament, 2000) and (Schwartz, 2012)
    - Professional group ethics defined and valid for its members
- **Law** (Big Data Ethics by Default)
    - General and specific legislation such as GDPR in the EU.
- **Architecture[11]** (Big Data Ethics by Design)

[10] Lessig later updated his work in order to keep up with the prevailing views of the time and released the book as Code: Version 2.0 in December 2006; although, the original formulations are more explicit.

[11] The architecture term was used originally by Lessing as one of the four general powers regulating social systems thus it is not equivalent to the commonly used term for IT architecture, e.g. in TOGAF framework. However, Lessing's term architecture that is similar to the above defined term of "code" for him, meaning although software code covers general social systems to the technological systems and different approaches to the ethical assurance of Big Data systems that will be described further.

- Different approaches to the ethical assurance of Big Data systems
    - Methods applied either a priory or ex post of Big Data systems implementation
- **Big Data Ethics by Design**
    - General approach and DEDA methodology as an example of the method applied a priory to the Big Data systems design and its implementation.

5.1 Market

Worldwide, the level of information accessible about society has increased significantly over the period of the last century, particularly after World War II. This process rapidly continues today and has accelerated with the development of Big Data. The power behind its development stems from the market, not the state. Big Data, as we know it, is a market driven phenomenon, and as such, it is much more suitable to be regulated through the market, rather than by the state. This, indeed, does not mean that states do not stand a chance against Big Data or that they should not strive to regulate it. However, as I argue later in this chapter, the matter of Big Data is not best regulated by the state in a restrictive manner, but proactively, using market tools.

In his concept of governmentality, Michel Foucalt addresses the question of how to govern in the emerging global society (Foucalt, 1978). He describes the contemporary population to be governed by a government by means of apparatuses of security, using a political economy. The security apparatuses are to provide the population with a general feeling of well-being. In doing so, Foucalt suggests the state`s actions be rather restricted and subtle in their nature, yet consequently very influential in their outcome; he offers a depiction of a multicentered society, which, even from the position of a state, is best regulated by the market mechanisms and by injecting the individuals with ideas, rather than forcing on them the government`s will by the means of straight discipline and law. The individuals then become autoregulated, auto-disciplined.

In order to be able to govern in this manner, in order to conduct such governmentality, an extensive amount of information about the population is required. With the development of Big Data, this becomes increasingly accessible; the problem is that it is not primarily the state gathering the information, it is the technology companies and, subsequently, other business sectors in the market.

Faced with the new phenomenon of the global technology companies, it may be rather difficult for the states to control them. In any case, the answer to these questions is not to open a vast conflict of sovereignty and discipline between the “private” Big Data and

the government. On the contrary, for states, it should prove more effective to engage with the technology companies using the instruments of Foucalt`s governmentality, as described above. This is not so as to indicate that the state should give up forcing the technology companies to bear their fair share of taxes or to acknowledge their accountability; only that in doing so, states could more effectively achieve these aims not by the sheer tools of legislative regulation, but through the governmentality and the construction of an autoregulated society.

5.2 Social Norms and Human values

5.2.1 Social Norms

The philosopher Jan Sokol introduces three different sets of rules on how individuals govern themselves where, besides social custom and individual morality, he argues that the best rules regulate relations in a society: ethics as a search for what is best (Sokol, 2016). This accounts for a positive action on the side of an individual, a creation.

These forces correspond closely to the two forces, imitation and innovation, which were the ones that the French sociologist Gabriel Tarde identified at the very beginning of the 20th century, substantially ahead of his time. Tarde saw no difference between natural and social sciences; he understood society to function on the same sets of networks as a cell would (Latour, 2009, p. 160). As Bruno Latour points out in his celebration of Tarde`s early genius, the sociology focusing on the holistic approach, on the whole, has lost its privileged status: “because everything is a society, there is no clear divide between the biological and the social”. With the development of technology, it shows that society in its nature is truly no different than a cell. Only a lack of tools caused Tarde to be pushed aside by sociologists, who argued a difference between the individualistic and the holistic.

Faced with the fast developments of the unregulated cyberspace, Lawrence Lessig (Lessig, 2000) warned against a regulatory attack on the internet, which he saw coming back in the beginning of millennium. Instead, he proposed for the common values and shared consensus to guide cyberspace`s regulation by positive means, infusing ideas into individuals through positive action and example. To borrow a biological term, using the connection first anticipated by Tarde, the way to regulate cyberspace is not to build an artificial castle of rules for it, but to observe it as a developing cell, subtly and naturally influencing it from the inside. Based on my opinion that is why Big Data Ethics by Default (Law) must be enriched by Big Data Ethics by Default that is more about self-regulation principles of data specialist working with Big Data in the field of all possible ICT projects.

5.2.2 Relationship Between Human Values and Sociology

Leading sociologists from the previous century such as Emile Durkheim (1897-1964) and Max Weber (1905-1958) said human values are a central concept for explaining social behavior of groups and individuals. As stated by the Israeli sociologist Shalom H.Schwartz:

"Values have played an important role not only in sociology, but in psychology, anthropology, and related disciplines as well. Values are used to characterize cultural groups, societies, and individuals, to trace change over time, and to explain the motivational bases of attitudes and behavior." (Schwartz, 2012).

Before I discuss in detail, in a special chapter, how Big Data use cases and society represented by the shared human values are in conflict, I will briefly introduce two complex concepts of human values, namely the Schwartz theory of human values and the European Charter of fundamental rights.

5.2.3 Human Values Based on Schwartz Theory

The Schwartz theory covers ten values that are recognized in all cultures worldwide that has been confirmed by a comprehensive survey done on more than 60,000 individuals in 64 nations (Wikipedia, 2019).

Although the ten values are recognized by all nations, they differ is their importance among different nations and individuals, thus the values can be ordered by their importance relative to one another. It is also important to mention that these values can be considered as beliefs that are impossible to separate from actions.

Based on Schwartz, we can differentiate the motivation of individuals and explanations of their behavior following attributes such as attitude, belief, values, traits and norms. The differentiation between attitudes and values that are essential for me while we will ask for them in our own Big Data survey. We can describe the difference between attitudes and values by the following quote from Schwartz:

"Attitudes are evaluations of objects as good or bad, desirable or undesirable. Attitudes can evaluate people, behaviors, events, or any object, whether specific (ice cream) or abstract (progress). They vary on a positive/negative scale. Values underlie our attitudes; they are the basis for our evaluations. We evaluate people, behaviors, events, etc. positively if they promote or protect attainment of the goals we value. We evaluate them negatively if they hinder or threaten attainment of these valued goals. If we value stimulation highly and attribute little importance to security values, for example, we are likely to have a positive attitude toward bungee jumping." (Schwartz, 2012).

The list of ten basic human values below shows that we can group values into four categories based on their similarity such as openness to change, self-transcendence,

conservation, self-enhancement. A short description of these ten values done below is an extract from the original Schwartz theory (Schwartz, 2012).

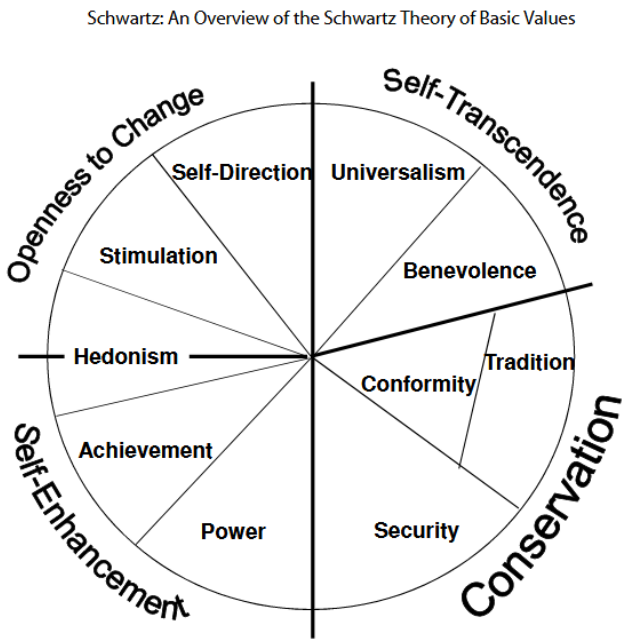


Figure 16 - The list of basic human values is shown below (2012, Schwartz).

Self-Direction
Defining goal: independent thought and action-choosing, creating, exploring.

Stimulation
Defining goal: excitement, novelty, and challenge in life.

Hedonism
Defining goal: pleasure or sensuous gratification for oneself.

Achievement
Defining goal: personal success through demonstrating competence according to social standards.

Power
Defining goal: social status and prestige, control or dominance over people and resources.

Security
Defining goal: safety, harmony, and stability of society, of relationships, and of self.

Conformity

Defining goal: restraint of actions, inclinations, and impulses likely to upset or harm others
and violate social expectations or norms.

Tradition
Defining goal: respect, commitment, and acceptance of the customs and ideas that one's
culture or religion provides.

Benevolence
Defining goal: preserving and enhancing the welfare of those with whom one is in frequent
personal contact (the 'in-group').

Universalism
Defining goal: understanding, appreciation, tolerance, and protection for the welfare of the 'in-group'.

It is important to know that these ten values interact together and we can also group them into the two big groups based on whether they have either a personal or social focus.

Values with personal focus are: achievements, power, hedonism, stimulation, self-direction.

Values with social focus are: security, conformity, tradition, universalism, benevolence.

A short evaluation of these Schwartz values that were also observed in our Big Data survey is in a special chapter: 7.5.1 Exploratory Data Analysis.

5.2.4 Human Values Based on EU Charter of Fundamental Rights

There is a long history in the declaration of human rights worldwide and especially in the US and Europe after the second world war. For a deeper overview of the international norms in the area of human rights, I recommend the book by D.C. Thomas, *The Helsinki Effect* (2001).

In 2000, the Council and the Commission of the European Parliament solemnly proclaimed the Charter of Fundamental Rights of the European Union.

In the preamble of this document it is written:

"The peoples of Europe, in creating an ever closer union among them, are resolved to share a peaceful future based on common values. Conscious of its spiritual and moral heritage, the Union is founded on the indivisible, universal values of human dignity, freedom,

equality and solidarity; it is based on the principles of democracy and the rule of law. It places the individual at the heart of its activities, by establishing the citizenship of the Union and by creating an area of freedom, security and justice." (European parliament, 2000).

The following list of values can be derived from this document:

- **Dignity**

Human dignity, right to life, integrity of person …etc.

- **Freedoms**

Right to receive and impart information, Liberty and security, respect for private life, protection of personal data …etc.

- **Equality**

Equality before the law, culture, religious and linguistic diversity, equality between women and men …etc.

- **Solidarity**

Workers right, prohibition of child labor, family protection, social security, healthcare …etc.

- **Citizens Right**

Right to vote and run for political office, right to good administration, freedom of movement and residence …etc.

- **Justice**

Presumption of innocence and right of defense …etc.

A short evaluation of these European values that were also observed in our Big Data survey is in a special chapter: 7.5.1 Exploratory Data Analysis.

5.2.5 Professional Group Ethics

The professional group ethics is usually defined by professional organizations that group people based on a shared professional identity.

"The common identity is produced and reproduced through occupational and professional socialization by means of shared educational backgrounds, professional training and vocational experiences, and by membership of professional associations (local, regional, national and international) and institutes where practitioners develop and maintain a shared work culture.", (Evetts, 2006).

Every industry wants to keep certain standards and best practice procedures to avoid the negative effects that could harm their customers or members of their professional organizations. Membership in such a professional organization related to specific industries or occupations is usually not mandatory; however, there are some benefits such as personal certification, available training, access to knowledge bases, possible participation in conferences that are available only to members of these organizations. And in some cases, if you are not a member of such a professional organization, you practically cannot do your job, like doctors of medicine that are not members of Camera Medica.

It is very typical that professional organizations declare their ethical standards valid for its members that are not legally binding but their breaking can be the official reason for a member's expulsion from the professional organization that has a similar impact to being punished from the court. Professional group ethics belong to applied ethics, because different problems are created and solved by different professional groups, e.g. medical doctors or by farmers. Philosopher Jan Sokol said:

"The achievements of modern science, technology, economics and organisation have enormously broadened the scope of human possibilities; and millions of people around the world are dedicated to the continued expansion of these possibilities. However, there are also a growing number of people who are troubled by the use we make of these incredible possibilities. Among the first of these were the physicists who, after the explosion of the first atomic bomb, were genuinely horrified by the forces they had unleashed. And the expansion of such possibilities has only gathered pace since then, giving the ancient question – 'how ought we to live?' – a new meaning and a new urgency, as attested to by the rich literature, the plethora of ethical codices and commissions and even our everyday public debate."[12] (Sokol, 2016).

The topic of professional ethics that is defined and valid for its members and belongs to regulative framework as a logical regulative tool that fits between general social norms and specific legislation.

5.3 Law (Big Data Ethics by Default)

"*Legislation, or statutory law, can have many purposes: to regulate, to authorize, to outlaw, to provide (funds), to sanction, to grant, to declare or to restrict*." (Voermans, 2009).

[12] The Illinois Institute of Technology database of ethical codes for various professions (http://ethics.iit.edu.research/codes-ethics-collection) lists over 850 of them.

The law usually works ex post as a reaction to the social and economic situation of society and the adoption of an appropriate legal regulation is always a matter of long-term social development. A persisting conflict between the right to receive and impart information, and the right to protection of personal data, arising, e.g., from the EU Charter of Fundamental Rights, is a small example of the complicated balancing act among the rights and duties of the involved stakeholders.

For a long time, opinion that granting the right to access to information and simple protection of a person´s privacy through private measures, such as the Civil Code, are sufficient has been prevailing. However, with rapidly developing technologies, and increasing power of state and corporations supported by growing production of high amounts of electronic data carrying information about individuals, the public interest on protection of consumers[13] and individuals against risks connected to possible misuse of such data was growing and legislators had to take an action. The action should be not only in the area of "intensity" if there is a law breach, where the Criminal Code should take place but there is a long-term visible need for actions reflecting the general impact of technologies on public interest and to defend human values stated in, e.g., the EU Charter of Fundamental Rights.

The law and legislation are different per each country all around the globe, and even in the European Union there are significant differences in local legislation implementation. We cannot omit also the national sector specifics of e.g. government, banking, insurance, healthcare, utilities, telecommunication and some others that can have the special legislative acts imposing special duties relating to clients or other sensitive data, such as secrecy obligation of the Police, medical doctors, attorney of law and others. Beside the legislation defending the basic human values (the EU Charter of Fundamental Rights, the Constitutions[14], the Consumer Law, Civil and Criminal Codes among others) and Corporate law[15] followed by the sectors specific regulation we can see that there are also some

[13] Consumer protection and right for the fair treatment that is strong focus of EU can be defined as the following: „In regulatory jurisdictions that provide for it (comprising most or all developed countries with free market economies), consumer protection is a group of laws and organizations designed to ensure the rights of consumers as well as fair trade, competition and accurate information in the marketplace." (Wikipedia, 2019).

[14] "A constitution is an aggregate of fundamental principles or established precedents that constitute the legal basis of a polity, organization or other type of entity, and commonly determine how that entity is to be governed. When these principles are written down into a single document or set of legal documents, those documents may be said to embody a written constitution; if they are written down in a single comprehensive document, it is said to embody a codified constitution." at Wikipedia (Wikipedia, 2019).

[15] „Corporate law (also known as business law or enterprise law or sometimes company law) is the body of law governing the rights, relations, and conduct of persons, companies, organizations and

specific legal norms (e.g. eIDAS[16]) or sub-norms of industries and professional organization[17] that should be in general now covered, or later integrated, by recently implemented General Data Protection Regulation (EU) 2016/679 (GDPR)[18], that has the ambition to solve majority of data related situations that can occur.

Although the use and processing of Big Data are not yet specifically regulated in the European Union, in particular, aspects of the volume and diversity of data predetermine and set up the scope of legal regulation in the area of personal data protection. As described in chapter 4.1 about Data sources, Big Data will in most cases contain personal data[19] that are protected by the special legal regime. Following the GDPR and the Convention No. 108 ETS[20], automated processing of personal data is only possible if the obligations arising from this legislation have been fulfilled.

The basic obligation arising from personal data protection regulations is to process personal data only based on a legitimate legal title. In many cases, the processing of Big Data in the commercial sphere is based on a relatively flexible legal title of a legitimate interest of the controller or third party. However, the flexibility of this legal title does not relieve the controller of his further obligations. In particular, the controller is obliged to subject its legitimate interest in the processing of the data in question to a balance test before the processing, which measures its legitimate interest concerning the rights and freedoms of the data subject. A valid legal title must be available to the controller from

businesses. The term refers to the legal practice of law relating to corporations, or to the theory of corporations. Corporate law often describes the law relating to matters which derive directly from the life-cycle of a corporation." (Hansmann & Kraakman, 2004)

[16] „eIDAS (Electronic Identification, Authentication and Trust Services) is an EU regulation on electronic identification and trust services for electronic transactions in the European Single Market. It was established in EU Regulation 910/2014 of 23 July 2014 on electronic identification and repeals directive 1999/93/EC from 13 December 1999", (Wikipedia, 2019).

[17] E.g. Camera Medica

[18] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=cs

[19] „Personal data are any information which are related to an identified or identifiable natural person.The data subjects are identifiable if they can be directly or indirectly identified, especially by reference to an identifier such as a name, an identification number, location data, an online identifier or one of several special characteristics, which expresses the physical, physiological, genetic, mental, commercial, cultural or social identity of these natural persons. In practice, these also include all data which are or can be assigned to a person in any kind of way. For example, the telephone, credit card or personnel number of a person, account data, number plate, appearance, customer number or address are all personal data." See art. 4 (1) at GDPR.

[20] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data https://rm.coe.int/1680078b37

the start of processing, i.e. from the initial storage of Big Data on the data carrier or server administrator.

Furthermore, the controller[21] should clearly define the purpose of the intended processing of personal data before commencing processing following the purpose limitation principle. However, the obligation to define the purpose before the start of processing and the possibility of future processing for other purposes considerably complicates the use of Big Data. This is both because the controller is often unable to specify the processing purposes in advance, not least because of the use of machine learning and artificial intelligence tools, but also because the controller is only authorized to process personal data for other incompatible purposes if such processing is based on the data subject's consent to processing for other purposes or is legitimately required by law.

The proper fulfilment of the above obligations is necessary to fulfil the principle of transparency since the controller is obliged to inform the data subject not only about his identity and the identity of the recipients of personal data but also about the legal basis of personal data processing and its purpose. Given the nature of Big Data, as in an unstructured set of information, it can be expected that compliance with the transparency principle will be similarly complex or virtually impossible for administrators. In such a case, however, the controller is obliged to adequately inform the data subjects, usually by providing information concerning the processing of personal data on its website. The processing information provided by the controller should be intelligible to the data subject, which in turn places increased demands on the controller especially when Big Data is processed by the tools of machine learning and artificial intelligence.

Although Big Data generally contains personal data, large volumes of non-personal information may also be encountered. Such non-personal information may be purely technical information or anonymized 'personal' data. The use of non-personal data in the European Union is regulated by Regulation (EU) 2018/1807 on the framework of the free flow of non-personal data in the European Union ("Regulation 2018/1807")[22]. Regulation 2018/1807 does not, in principle, limit the processing of non-personal Big Data but, on the contrary, makes it easier to move around the European Union. However, following

[21] Article 35 of the GDPR also covers: „Data Protection Impact Assessments (DPIA). The DPIA is a part of the "protection by design" principle. Examples when examples DPIA is required are: If you're using new technologies and If you're tracking people's location or behavior, among others". GDPR

[22] REGULATION (EU) 2018/1807 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. https://eur-lex.europa.eu/eli/reg/2018/1807/oj

the Commission's informative guidance[23], Regulation 2018/1807 only applies to data sets of a purely non-personal nature, and for files containing personal data, it is necessary to follow the GDPR directly.

Last but not least, it is important to note that when processing Big Data, both those containing personal data and those without personal data, the controller has to comply with the obligations set out by other legal regulations whose examples we mentioned at the beginning of the chapter. These obligations may, on the one hand, arise from sectoral regulation, here we draw attention especially to the processing of personal data in the telecommunication, healthcare or insurance industries (not only), but also rights and obligations relating to the protection of intellectual property and trade and professional secrecy. Under the principle of legality, the fulfilment of these obligations is a necessary condition for the legitimacy of any processing of personal data. Similarly, any processing of Big Data for profiling a data subject is subject to a specific regime, and the controller will have to obtain data subject's consent for profiling purposes.

In conclusion, the Big Data analysis represents a major opportunity for entities operating (not only) in the private sector. However, the development and technological possibilities for the processing of personal data are strongly limited by legal regulation, in particular by legal regulation concerning personal data. Obligations arising from the legal regulation of personal data protection can be difficult to fulfil, as is proved by the factual impossibility of fulfilling the information obligation in the case of using machine learning tools and artificial intelligence.

The legal regulation of Big Data is as described above currently mainly unified with the regulation regarding the handling of personal data. This constitutes an issue, as the regulation oftentimes subsequently fails to distinguish between the big technology companies on the one hand and small or medium businesses on the other, making the law somewhat ineffective and the imposed burden on the medium and smaller businesses disproportionate. Indeed, the idea of tightly regulating Big Data primarily by the law, even in the case of special Big Data legislation, might still be inefficient, as I argue above using Foucault`s governmentality.

The current and most significant legislation in the EU regarding personal data is as mentioned several times above the GDPR directly applicable since the 25th of May 2018. Its main object is to enable individuals (data subjects) to have control over their personal data, while unifying the legal framework in the EU and the EEA countries (Iceland,

[23] Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN

Lichtenstein, Norway) and posing duties on controllers and processors of personal data, in order to prevent their misuse.

Although GDPR is by nature a directly applicable regulation, there was still national implementation needed on behalf of the member states, for instance to accommodate the foundation of the state-based individual supervisory authorities.

Beside the general legal principles described above it is also useful to name a few significant examples where GDPR enhanced earlier rules or even introduced new ones, I choose the four following areas examined below in more detail: 1) consent, 2) the right to access, 3) penalties and 4) extraterritorial applicability.

In line with other areas of consumer protection within the EU, GDPR introduces a more consumer-friendly way for individuals to give consent with the use of their personal data, making it compulsory for the request of consent to be in an intelligible and accessible form, using plain language. It is also required that consent be as easy to withdraw as it is to give.

The right to access makes it possible for individuals to request information as to whether their personal data is being processed and for what purpose, and the individuals are also entitled to obtain a copy of their personal data in an electronic format.

In order to enforce the rules, enterprises may be fined in amounts up to €20 million or 4% of their annual worldwide turnover. The existence of a fine is often seen as an important measure that should safeguard the compliance of even the biggest companies with the rules. Also, discretion applied in the imposing of fines may contribute to the balancing of the conditions between medium and small businesses and big ones.

GDPR ends the ambiguity of the application of the rules of personal data processing when it establishes the rule under which the residence of the data subject is decisive, regardless the location of the enterprise processing his/her personal data.

In the future, the room for an additional regulation[24] may lie within each individual industry sectors, mainly to level the playing field and promote fair competition. For example, in finance, the European supervisory authorities together published the Joint Committee Final Report on Big Data, dated the 15th of March 2018, which addresses,

[24] E.g. European Commission in June 2019 approved: Open Data and Public Sector Information Directive (EU) 2019/1024 (shortcut: open data directive) replacing the Public Sector Information Directive 2003/98/EC. The implementation deadline for national states is July 16th 2021. See: https://ec.europa.eu/digital-single-market/en/legislative-measures and Regulation on the Free Flow of Non-Personal Data (EU) 2018/1807 was extended by its guidance see: https://ec.europa.eu/commission/presscorner/detail/en/MEMO_19_2750

among others, the inequalities associated with Big Data, potential shortcomings in the transparency of Big Data tools, or potential non-regulatory barriers to the use of Big Data. Other regulation in specific industries are in EU legislative pipeline in Insurance industry[25] but also in other sectors.

Because of the possible sanctions and restrictions enabled by state power, we can call Big Data legislation "**Big Data Ethics by Default".** It means that legislation on how to deal with data, (e.g. GDPR) must be an inherent and default part of every ICT project, and it is guaranteed by the state power and its enforcement.

5.4 Architecture (Big Data Ethics by Design)

Following the previous clarification of Lessig's term architecture that differs from the standard IT term and his bridge from social systems to the technological systems, which are produced by people and can impact society, we will focus here mainly on the regulation of the Big Data systems that is the theme of this paper. Basically, in order to meet the defined[26] ethical values, we can regulate the Big Data systems either before or after their implementation. We will discuss these different approaches to the assurance of ethics in Big Data systems more closely in the next chapters that are diving deeper into the architecture elements of regulation framework.

5.4.1 Different Approaches to Ethical Assurance of Big Data Systems

Nowadays, we can see the rise of solutions that could minimize the negative impacts on society caused by technologies such as Big Data and Artificial Intelligence (AI)[27]. Basically, the two different approaches can be applied when discussing the compliance of complex Big Data systems with defined ethical principles. We can talk about approaches applied a priori, meaning at the beginning of the Big Data system design, (e.g. the DEDA method described further) and approaches applied ex post, meaning in addition to original implementation (e.g. the Z-Inspection method described further). The a priori methods represent a more philosophical approach based on guided discussion and the ex post methods represent the more IT approach based on processes, tools and metrics. It is probably a good idea to combine both methods together, meaning to have guided

[25] E.g. EIOPA, The European Insurance and Occupational Pensions Authority (EIOPA) that was initiated by the European Commission, established in October 2019 Consultative Expert Group on Digital Ethics in Insurance, that should support new data ethics regulation in Insurance industry, See: https://eiopa.europa.eu/Pages/News/EIOPA-establishes-Consultative-Expert-Group-on-Digital-Ethics-in-Insurance.aspx

[26] Definition of ethical values can follow either the generally accepted values in society described, e.g., in EU Charter of Fundamental Rights or can be more specific as we can describe them more specifically before the start of Big Data system design and its implementation

[27] I will use in this chapter a general term Big Data covering also AI and all related technologies.

discussion about ethical design before the Big Data system is implemented and then apply some kind of assurance or sanity check method.

I will talk about the DEDA method example later in special chapter because I believe that it is very well balancing the legislation framework (Big Data Ethics by Default) that acts ex post; however, it is probably useful to describe here also an example of possible post implementation assurance method.

Professor Roberto Zicary and his colleagues working at Frankfurt Big Data Lab introduced methodology called Z-Inspection described in their presentation *Z-inspection: Towards a process to assess Ethical AI,* presented recently at cognitive science talks (Zicari, 2019).

This methodology is an attempt to define an assurance process that can be used by ICT experts to evaluate the complex Big Data systems. The difference of Z-inspection method compared to the proprietary improved approaches used currently by auditing[28] and consulting companies such are Deloitte, KPMG, E&Y or PwC is that the Z-inspection is an academic work that is focused on a broader discussion of the ethical assurance of Big Data systems rather than on structured audit reports dedicated to company shareholders or executives.

The Z-inspection methodology described by Zicari (2019) consist of the following steps:

"1. Define a holistic Methodology
   a. *Extend Existing Validation Frameworks and Practices to assess and mitigate risks and undesired "un-ethical side effects", support Ethical best practices.*
   b. *Define Scenarios (Data/ Process/ People / Ecosystems),*
   c. *Use/ Develop new Tools, Use/ Extend existing Toolkits,*
   d. *Use/Define new ML Metrics,*
   e. *Define Ethics AI benchmarks*
2. *Create a Team of inspectors*
3. *Involve relevant Stakeholders*
4. *Apply/Test /Refine the Methodology to Real Use Cases (in different domains)*
5. *Manage Risks/ Remedies (when possible)*
6. *Feedback: Learn from the experience*

[28] The Global Technology Audit Guide (GTAG) was recently extended with material called: "Understanding and Auditing Big Data" that is rather general and will probably go through further development. Material is available at: https://na.theiia.org/standards-guidance/recommended-guidance/practice-guides/Pages/GTAG-Understanding-and-Auditing-Big-Data.aspx

7. *Iterate: Refine Methodology / Develop Tools", (*Zicari, 2019).

The focus of the Z-Inspection is on the ethical, legal and also technical aspects of complex Big Data System resulting in a score showing the different grades of ethics and transparency of evaluated ICT systems. The score ranks from the worst level (Black Box) to the best level (fully ethical and transparent).

Z-Inspection uses different techniques to investigate the ICT systems at different layers (data, process, people), distinguishing at the macro or micro level where, e.g., at the micro level, there is a deep dive into datasets and software code.

The iterative process of Z-inspection uses the so called "path approach" that describes the dynamic of the inspection and depends usually on a team of inspectors. The path is a rather intuitive approach and can start with the predefined set of steps or run just randomly trying to discover missing parts of Big Data systems that are not visible at the beginning of the inspection.

The Z-Inspection methodology and the inspectors are ready to use the set of already existing metrics and open source tools with the purpose of mapping the Big Data systems. Some examples of such tools are listed here:

- What if Tool, Facets, Model and Data Cards (Google),
- AI Fairness 360 AI Explainability 360 Open Source Toolkit (IBM),
- FairML, https://github.com/adebayoj/fairml,
- Aequitas (Univ. Chicago) https dsapp.uchicago.edu/aequitas,
- Lime (Univ. Washington) https ://github.com/marcotcr/lime.

Using the tools above can be useful but can also open a new paradox of transparency like: "*What if transparency of AI is controlled by another AI and if so then who validates the AI controller?",* (Zicari, 2019).

The above described Z-Inspection methodology is in an early stage and is expected to be developed further by the Frankfurter Big Data Lab. I think that we can expect that other methodologies will be introduced and popularized in the academic and also business environments. I believe that these new findings about ethical assurance of Big Data systems will later move to the professional standards and ICT frameworks such are DAMA-

DMBOK[29], COSO[30], ITIL[31], COBIT[32] or ISO/IEC 27 0xx[33] and maybe also became part of them. I will further focus on the DEDA approach that is important to this thesis because of its a priori deployment that balances the post legal regulation. It also follows the ethical approach formulated in previous chapters such as "ethics is a search for what is best" (Sokol, 2016) rather than post evaluation of completed Big Data systems done by the above described Z-Inspection or similar assurance methods.

5.4.2 Big Data Ethics by Design

Ethics by design, in its essence, is an approach that has existed since the introduction of computers, namely computer ethics. It was described by Friedman & Kahn (2007) as part of the Human-Computer-Interaction in their book *Human values, ethics, and design;* however, specifically data ethics by design is a very recent trend and there are not many publications related to this area.

In my view, Big Data Ethics by Design should follow a similar activity called Privacy by Design[34] (PbD) that was developed by Dr. Ann Cavoukian in the 1990s as a response to the growing threats to online privacy.

---

[29] „Data Management Association (DAMA) is a non-profit organization of data-management professionals focused on data-management disciplines that, as result of the work of its members, produced best-practice standards in data management known as the DAMA Data Management Body of Knowledge (DAMA-DMBOK)." See at (http://www.dama.org/)

[30] „COSO (ERM) COSO's Enterprise Risk Management is an integrated framework defined by the Committee of Sponsoring Organizations of the Treadway Commission (COSO) that is expanding Internal Control and Integrated Framework to provide guidance for a comprehensive enterprise-wide approach to managing risk." See at (www.isaca.com)

[31] „ITIL (Information Technology Infrastructure Library) is a set of practices for IT service management (ITSM) that focuses on aligning IT services with the needs of business. ITIL describes processes, procedures, tasks, and checklists which are not organization-specific. ITIL underpins ISO/IEC 20000 – however, these two frameworks do have some differences." See at (http://www.wikipedia.org/)

[32] „COBIT 5's key components are: Principles, Policies and Frameworks. Processes. Organizational Structures. Culture, Ethics and Behavior. Information. Services, Infrastructure and Applications. People, Skills and Competencies." at: https://www.isaca.org. Especially, the COBIT part of Ethics and Behavior is the one where the new findings about Big Data and its impact to society and ethics can fit. There is ongoing discussion in ISACA organization about future development of the COBIT standard.

[33] „ISO/IEC 27000 provides the overview of information security management systems (ISMS), and the terms and definitions commonly used in the ISMS family of standards. It is applicable to all types and sizes of organization (e.g. commercial enterprises, government agencies, not-for-profit organizations). There are also some industry specific extensions, e.g. ISO/IEC 27011, (Telecommunication), ISO/IEC 27019, (Energy), ISO/IEC 27799, (Health) among others." See at (www.iso.org)

[34] The general term Privacy by Design is also used in GDPR where is described as the following: „The term "Privacy by Design" means nothing more than "data protection through technology design." Behind this is the thought that data protection in data processing procedures is best adhered to when it is already integrated in the technology when created. Nevertheless, there is still uncertainty about what "Privacy by Design" means, and how one can implement it.", See at https://gdpr-info.eu/issues/privacy-by-design/

"Privacy by Design (PbD) is an approach to protecting privacy by embedding it into the design specifications of information technologies, accountable business practices, and networked infrastructures, right from the outset." (Cavoukian, 2011).

From the quote above, it is clear that Privacy by Design is moving attention from reactive legal regulation (Data Ethics by Default) to a proactive moment when processes and code are designed by their authors. That is, in my opinion, the only way how to manage the complexity and speed of the digital world and the related data ethics topics.

In the words of Professor Lessig:

"For citizens of cyberspace, . . . code . . . is becoming a crucial focus of political contest. Who shall write that software that increasingly structures our daily lives? As the world is now, code writers are increasingly lawmakers. They determine what the defaults of the Internet will be; whether privacy will be protected; the degree to which anonymity will be allowed; the extent to which access will be guaranteed. They are the ones who set its nature. Their decisions, now made in the interstices of how the Net is coded, define what the Net is." (Lessig, 1999).

I see that with the development of Big Data and the following Artificial Intelligence that many new ethical questions that go beyond the essential rules of Asimov's laws of robotics and their implications for current information technology will be opened. (Clarke, 1994). There are some well-known dilemmas already, but also new questions which stem from the Big Data implementation into specific industries as described below, for example, the automotive industry and the area of self-driving cars.

> *"Autonomous vehicles will also have a direct impact on our society that today we can barely imagine. Numerous critical questions arise: What are the prospects concerning data security? How shall we deal with wide-ranging interventions in our own mobile autonomy? What problems result when an autonomous vehicle crosses national borders? In what form will insurance companies assume liability for autonomous vehicles involved in accidents in the future? Or, vice versa:* ***Can we continue to leave humans at the wheel at all, and may driving robots prove to increase road safety?"*** (Maurer, at el, 2016).

This fast development of data science in many industries is creating demand for a more flexible regulative approach. In my opinion, the most relevant and flexible regulative approach is the Big Data Ethics by Design methodology covering a set of pre-defined data ethics relevant questions and related processes that the designers should go through when designing the complex data science solutions. It seems to me to be a more relevant approach than the strict rules of legislation that support a formal "check box" mentality of involving specialists to classify actions as either legal or illegal.

As the area of Big Data Ethics by Design is in early stages, see in next chapter, I appreciate very much the recent work done at Utrecht University and their DEDA methodology.

5.4.3 DEDA Methodology

DEDA is an abbreviation for Data Ethics Decision Aid, which is a methodology developed at Utrecht University as a result of the cooperation between the special focus groups of the Utrecht Data school and the Data Ethics Lab[35].

DEDA is an ethical assurance approach focused on the Big Data projects that are based on guided discussion that should include all the people relevant to a project and take place before the Big Data system is designed and therefore should trigger the changes to improve the ethical part of the project in the early stages. Because of the timing, open questions and guidance that limit "a check-box mentality" this approach should be more effective than ex post methods and corrections applied to already implemented Big Data systems.

DEDA can be described as a tool that has three steps that each project related to data manipulation needs to go through.

Step one is to learn the principles of the DEDA methodology from the DEDA manual that provides additional background information.

Step two is to call for a project meeting with all the involved people in the new data project. The team should include not only IT specialists but all other relevant departments such as sales, marketing, risk, compliance, legal …etc. In step two, the team goes through a pre-defined set of questions that should help the project team realize whether their decision to start such a project has any ethical issues. Step three is to take a decision about the project (Go, No-Go, Change or Modify project).

There are 29 questions in the DEDA methodology described by its authors (Schäfer, Franzke, Utrecht & Fransen, 2012) that are grouped into the following categories:

> ***"Algorithms***
> *1a, Is there someone in the team who can explain how the algorithms in use work?*
>
> *1b, Can you communicate how the algorithms work?*
>
> ***Source***
> *2, Where do the data(set) come from?*

[35] For more details about this DEDA methodology, visit link https://dataschool.nl/deda/?lang=en to web site describing the work done at the Utrecht University and Utrecht Data School.

3, Have you checked the quality of the data(set)?

4, Is there ´best before´ data for this specific dataset?

Anonymization

5, Are the data anonymized or pseudo-anonymized?

6a, Have you tested the anonymization?

6b, Who is in possession of the encryption key?

Visualisation

7, Are the data or the produced results suitable for visualization?

8, In what way could the data be visualized?

9, Can the visualization be interpreted in a different (misleading) way?

***Access** (to Data)*

10a, How is the access monitored?

10b, Who has access to the dataset?

Open access & reuse of the dataset,

11a, Are (parts of) the data suitable for reuse? If so, under which conditions could they be reused?

11b, What are the possibilities of reuse?

12, What dangers do you see with reusing data?

Responsibility

13, Which laws and regulations apply to your project?

14a, Can you name a person responsible for handling

14b, Who is responsible if something goes wrong?

15, Is there a danger that particular people or groups could be discriminated against by your project?

16, Would there be any suitable (commercial) partners for your project?

Transparency / Accountability

18a, How transparent can you be with the public about your project?

18b, Is there a danger of public outrage?

Privacy

19, Are sensitive data actively used in your project? (If 'NO', go to question 23)

20, Do you have insights into privacy sphere of citizens?

21, Does the dataset allow insight into personal communication of citizens?

22a Have you checked the PIA (Privacy Impact Assessment?)

22b, Have you had contact with Data Privacy Officer?

BIAS

23a, What outcomes are you expecting personally?

23b, What do other team member and colleagues expect?

24a, Do you have vague feelings about the project?

24b, Do you have a good feeling about the project?

25a, Is the sample a truthful representation of population?

25b, Who is missing or invisible in your dataset?

26, Are you gathering the right information for your goal?

27a, Does your decision change if you think about long terms effects? Why?

27b, Can you imagine a future scenario in which your current decision might matter?

Informed consent

28, How do you inform people that the data are used?

29, Do citizens have a choice to opt-out?", (Schäfer, Franzke, Utrecht & Fransen, 2012).

Step three is about taking a decision and taking into account the answers to the questions of step two and also the general consideration for the questions below.

- *"Does the project meet the standards of good governance and responsibility?*
- *Which outcome is the best for the majority of involved subjects, the city and its residents? (utilitarism)*
- *Can you imagine yourself on the other side? (virtue ethics)*
- *Does your approach respect the autonomy of all subjects who are involved? (Kantianism)*
- *What are the problems particular to this project? (moral particularism)*
- *How can you make sure that the following values are respected?*
    - *Freedom of choice*
    - *Freedom of speech*
    - *Mutual respect*

- *Trust*
- *Diversity*
- *Creativity*
- *Peace and good life."*

(Schäfer, Franzke, Utrecht & Fransen, 2012).

5.4.4 Authors Proposal for DEDA Improvements

DEDA designed at Utrecht University and its Ethical Lab is a very complex and philosophical model, maybe even too complex for implementation in commercial organizations.

To make the implementation for organizations easier, I suggested to authors[36] a Go to Market (G2M), which is based on a multi-stage approach wherein the first stage, called a Health Check, can be done very fast.

Although the Health Check should only have 10 questions max and can be implemented as a self-care tool (App), it should attract the attention of various commercial stakeholder groups (Compliance, Legal, Risk, Audit, Sale, Finance, Technicians...) for the following workshop. The workshop that will follow after the Health Check should cover the full complexity of DEDA and is expected to be more time demanding.

The expected effect of taking the self-care Health Check first is that the commercial stakeholders will attend the workshop with some insight already and will also be more motivated and better prepared. Being well prepared is essential as they will know the complexity of the DEDA model already and can think ahead about the soft areas, such as the values of the company. This should help avoid disasters at workshops that we have had some experience with from previous implementations.

<u>Go to Market detail:</u>

- The selection of the roughly ten questions in the Health Check tool should lean towards hard questions rather than soft questions.
- As hard questions, I consider the ones that relate to algorithms, source data, anonymization …etc. In my opinion, the soft questions are, e.g., value related questions.
- The selection of Health Check questions should open the complexity of DEDA and respect the commercial governance model. In other words, we should select the questions in a way that there is no chance that only one department such as compliance

[36] I was in November 2018 on short internship in Utrecht university and had a chance to discuss it with the authors of DEDA.

or technicians can cover all the questions alone without other units being involved. This should motivate internal discussion before the workshop.

- The results of the Health Check can be categorized by a, e.g., red, orange or green status.

- Reaching the green status can be followed by awarding the Green label with a strong disclaimer that even if it looks good at first glance this is only a preliminary and self-care assessment that needs to be followed by a deep dive workshop.

- The red or orange result of the Health check should attract the attention of stakeholders to the specific areas of DEDA (e.g. better understanding of the algorithms).

- I do expect that some commercial stakeholders will end up with the Health Check only and will not go for the full version of DEDA. However, the ones that do follow will be better prepared and motivated for the workshop that requires more time and effort of all participants.

I have spent time on a short internship at Utrecht University and handed over my proposal to improve DEDA to its authors, so hopefully this will help to enlarge this interesting approach outside the Netherlands and Utrecht University. For the purpose of this thesis, I found DEDA as a very interesting and specific example of how the Big Data Ethics by Design approach could look.

5.5 Conclusion of Regulatory Framework

Based on my research I propose to move from ex post Big Data Ethic by Default (Law) to a priori Big Data Ethics by Design approach that should be inherent part of every ICT project that process data. I strongly believe that autoregulation element of ethics of ICT experts is the only way how to face the fast-paced evolution of ICT and especially the data science. The Big Data Ethics by Design approach should be supported by professional organizations of ICT experts and data scientists where the membership should be essential part of their qualification and ability to work in this area. I suggest to focus on development of methods applied a priori to implementation of Big Data systems as above described DEDA. I believe that the research and suggestions done in previous chapters will enable the Big Data Ethics by Design to become a specific discipline of data ethics. This new discipline should follow the Big Data specifics described in chapter 3.4 such as: the role of stakeholders (organizations, users, states), new use cases, growing demand for regulations and arising conflicts and issues in society caused by Big Data.

6 Digital Divide Conflict and Big Data Issues

Following the chapters Use cases and Regulative Frameworks and the herein described human values, we can see that Big Data issues and human values are powers that are inherently in conflict. I will use the outputs of the chapters above to name a few typical conflicts shown below.

A few basic examples of the possible conflicts[37] between human values and Big Data specific factors are shown below:

- Privacy vs New business opportunities
- Privacy vs Insight into complex problems
- **Equality vs Digital divide**

The scope of this thesis does not allow for the discussion of all these conflicts in detail. That is why I am selecting only the conflict of equality versus the digital divide to discuss in detail, taking into account the Big Data specifics.

For sure, it would be interesting to further discuss other conflicts in more detail; however, the issues of Big Data related to privacy have already been thoroughly described by other authors such as Sarah Spiekermann, Ann Cavoukian, Tomas Sigmund and more.

I have selected, as a special focus of this thesis from hereinafter, the conflict of equality versus the digital divide. I will also touch on the invasion of privacy and other significant negative factors, which, in the following paragraph, are referred to as Big Data issues.

6.1 Digital Divide Introduction

The term Digital Divide in relation to IT technology is not new. The term Digital Divide was coined in the eighties with the introduction of the internet in America and its nonlinear spread between its first users. We call this internet issue the First Digital Divide. The author of the current term Digital Divide is not precisely known; however, based on Gunkel, (2003) and Dijk (2006), the first official publication about that topic was by the US Department of Commerce´s National Telecommunication and Information Administration (NTIA, 1999).

The definition of the First Digital Divide is based on the following van Dijk quote:

[37] There are special papers discussing the all possible conflicts stemming from the implementation of Big Data and other related technologies is. For more details see, e.g. (Whittlestone at el, 2019).

"The First Digital Divide commonly refers to the gap between who do and those who do not have access to the new forms of information technology." (Dijk, 2006).

During the time when sociologist sand IT experts analyzed this phenomenon in more detail, the term of the Second Digital Divide was also introduced. This term is more complex and is less dependent on physical access to the hardware infrastructure and is more focused on other relevant aspects.

A good definition of the Second Digital Divide comes, for example, from DiMaggio and Hargittai (2001) that suggested five dimensions along which divides may exist:

- *"Technical means (software, hardware, connectivity quality);*
- *autonomy of use (location of access, freedom to use the medium for one's preferred activities);*
- *use patterns (types of uses of the internet);*
- *social support networks (availability of others one can turn to for assistance with use, size of networks to encourage use); and,*
- *skill (one's ability to use the medium effectively)."* (DiMaggio & Hargittai, 2001)

It is important that we can view the topic of the Digital Divide from different perspectives and focus more either on the technological aspects (Kling, 1998) or social aspects (Katz & Rice, 2002) (Norris, 2001).

Pipa Norris in her book, *Digital Divide: Civic Engagement, Information Poverty*, and the Internet Worldwide pointed to divides at three levels:

- *"The global divide which encompasses differences among industrialized and lesser developed nations;*
- *the social divide which points to inequalities among the population within one nation;*
- *and, a democratic divide which refers to the differences among those who do and do not use digital technologies to engage and participate in public life."* (2001, Norris).

In the most recent publication, there is an attempt to introduce the term the Third Digital Divide as it was done, for example, by Alexander J. A. M. van Deursen and Ellen J. Helsper in their article: THE THIRD-LEVEL DIGITAL DIVIDE: WHO BENEFITS MOST FROM BEING ONLINE? They are shifting the focus from technologies and skills towards the focus on who benefits from outcomes of new technology introduction and how interconnected an offline and online world is (Deursen & Helsper, 2015). Their approach is shown in the following figure.

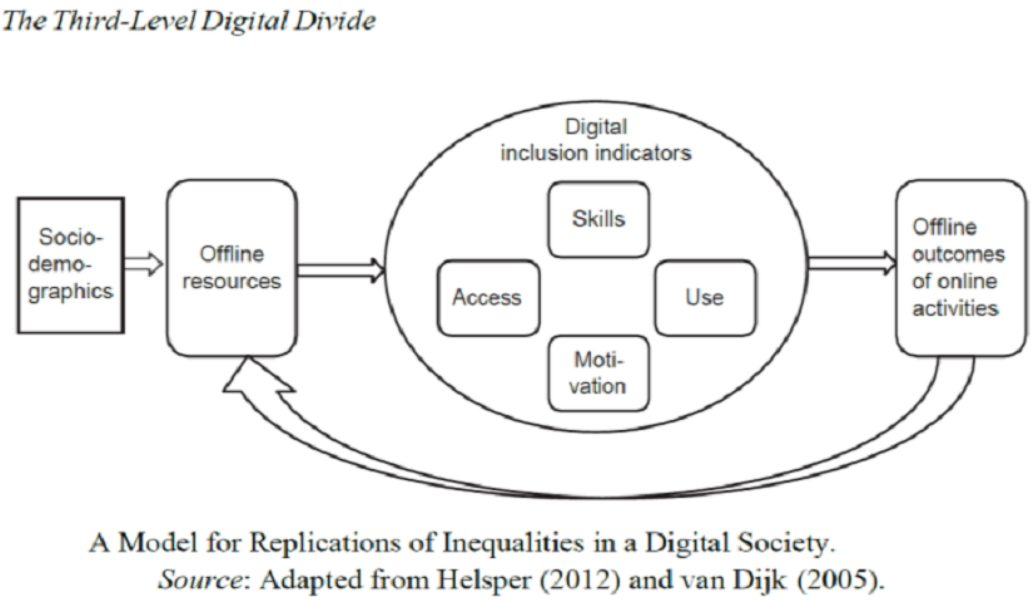


Figure 17 - The Third Level Digital Divide (Helsper, 2012)

6.1.1 Big Data and Digital Divide

Big Data can be considered a new technology causing a similar divide issue as the internet and other technological innovations well described above. Thus, it is a good question: what is really new in the phenomenon of social divide? This phenomenon, related mainly to the internet, has been described by many authors, e.g. by Katz & Rice (2002) or Norris (2001) and the problem of inclusion of segregated groups is also known (NTIA, 2002).

I think that to answer that question, we must come back to the specific issues of Big Data and look for their consequences. Furthermore, we should discover whether the population is at all aware of that divide caused by data rich and data poor groups (2012, Boyd & Crawford).

We have to come back to the research of Big Data issues and also the research of data ethics pioneers such as Floridi and Taddeo, Boyd and Crawford, Cukier and Mayer-Schoenberger, Anderson, Dijk, Norris, Allcott, Haesler, Creemers and others. We should also not exclude the specific area of Big Data surveys such as, for example, Mark Andrejevic's. In 2014 he published a very relevant article named: *THE BIG DATA DIVIDE*, supported also by his own survey done between respondents in Australia.

We can start the list of issues with the reputable work of Boyd and Crawford named: *CRITICAL QUESTIONS FOR BIG DATA*, which can be summarized in the so called Six Provocations of Big Data listed below:

- *"Big Data changes the definition of knowledge*
- *Claims to objectivity and accuracy are misleading*

- *Bigger data is not always better data*
- *Taken out of context, Big Data loses its meaning*
- *Just because its accessible does not make it ethical*
- *Limited access to Big Data creates a new digital divide."* (Boyd and Crawford, 2012).

Although the six provocations create a good foundation for the list of Big Data issues, we recognize that more of them exist. We researched comprehensively Big Data issues to create a list of twelve issues compiling the work of relevant authors and our own proposals. This list is later translated into our survey questions mapping Big Data issues awareness among stakeholders.

We understand the following twelve issues as unique; however, we are aware that some of them can be sorted into groups based on their focuses. This is especially the case of issues 2-4 in the list below, which we can call the Digital Divide group.

6.2 List of Big Data Issues

(1) **Privacy Intrusion** is currently a very hot issue because users are freely giving access to their personal data to get better digital service without control of its future usage, e.g. (Floridi, 2014) and (Taylor, Floridi & Van der Sloot, 2016).

- *"In this context, key issues concern possible re-identification of individuals through data-mining, -linking, -merging and re-using of large datasets, as well as risks for so-called 'group privacy', when the identification of types of individuals, independently of the de-identification of each of them, may lead to serious ethical problems, from group discrimination (e.g. ageism, ethnicism, sexism) to group-targeted forms of violence".* (Floridi & Taddeo, 2016, p 3)

- The Privacy by Design, suggested by authors such as Ann Cavoukian (2001) and Sarah Spiekermann (2012), is good inspiration for solving the privacy issue. This approach should guarantee the proactive design of code, respecting the privacy of digital customers before a legal penalty might be applied.

(2) **New Barriers** of socio-technological capabilities were described firstly in the eighties by Sir Tim Berners-Lee, inventor of the worldwide web. Berners-Lee had in mind a form of data divide not simply between those who generate the data and those who collect, store, and sort it, but also between the capabilities available to those two groups (Andrejevic, 2014). However, this elementary challenge is well

known from the different Digital Divides described by many authors, e.g. Norris (2001), DiMaggio & Hargittai, (2001), Dijk, (2006) and Deursen & Helsper, (2015).

(3) **Business Advantages** are available to a limited group of companies. We can make a list of roughly TOP 1,000 companies which are all big multi-national data collectors such as Telecommunication (Telco) operators, financial institutions, energy and utility companies, and big on-line corporations. These corporations are able to collect, store and manipulate large data sets about almost everything as the world is becoming "datafied". Datafication is a new term for a modern technological trend turning many aspects of our life into computerized data (Cukier and Mayer-Schoenberger, 2013) and transforming this information into new forms of value (O'Neil and Schutt, 2013). This challenge is partially related to competition disruption.

(4) **The Power of "All Data"** is available only to a few and roughly the TOP 10 list of monopolies (e.g. Google, Facebook, Alibaba, Amazon, Microsoft, Apple, etc.) that can predict almost anything. That was mentioned within the datafication phenomenon by Cukier and Mayer-Schoenberger (2013) or well described also in the following quote from Mark Andrejevic:

- *"It is about finding new ways to use data to make predictions, and thus decisions, about everything from health care to policing, urban planning, financial planning, job screening, and educational admissions. At a deeper level, the big data paradigm challenges the empowering promise of the Internet by proposing the superiority of post-explanatory pragmatics (available only to the few)"* (Andrejevic, 2014, p. 1673-75).

o and

- *"Even if users had access to their own data, they would not have the pattern recognition or predictive capabilities of those who can mine aggregated databases. Moreover, even if individuals were provided with everyone else's data (a purely hypothetical conditional), they would lack the storage capacity and processing power to make sense of the data and put it to use" (Andrejevic, 2014, p. 1674).*

This challenge relates to the disruption of competition and creates a world of oligopoly that is not transparent and companies often more powerful than the countries in which they operate.

(5) **A New Big Brother Effect** is the current renewed phenomenon of Big Brother in Big Data, highlighting that we consider a state and all other data collecting corporations to be, by their nature, good and never bad. For example, Google declared in their famous mission statement: “Don´t be evil.” We, as a population, are step by step giving up our decision-making power and control over our lives to anonymous corporations and nation states like in Orwell’s novel that originally invented the term "Big Brother" (Orwell, 1961).

- The term Big Brother, now renewed with Big Data, is related to commercial organizations but also at a state level, like in China, where the Social Credit Score system has been implemented. This system, which is in a pilot phase now and should be in full operation in 2020, is based not only on the payment history of individuals but also on the monitored behavior of individuals (NPR, 2018) (Creemers, 2018).

- *Besides the possible misuse of Big Data by a state or for political purposes as was the case with Cambridge Analytica and Facebook, Big Data in relation to social media opened a totally new pandora's box of fake news and manipulation following the 2016 US presidential election and the role of social media* (Allcott, 2017).

(6) **Missing Transparency** results from unclear algorithms during the decision-making process analyzing Big Data because the related algorithms start to get extremely complex. A consequence is that people's insight is replaced by the so-called black box approach (Rosicky, 2011). The problem is well described in the work of Cathy O'Neil (2016): *WEAPONS OF MATH DESTRUCTION.*

- The situation became more complicated not just because the algorithms are extremely hard to understand from a mathematical point of view, but because the execution and access to the algorithm usually belong to the state or large corporations. These organizations such as Google, Facebook, Amazon and others apply very strong and hierarchical security procedures, considering this algorithmic know-how as their intellectual property that needs to be defended.

- *“There are machines that learn, that are able to make connections that are much, much finer than you can see and they can calibrate connections between tons and tons of different facets of information, so that there is no way you as a human can understand fully what is going*

on there." (J. Haesler, personal communication, February 26, 2013) in (Andrejevic, 2014, p. 1681).

(7) **Confusion,** meaning the loss of clarity stemming from Big Data, creates the perception of the real world interpreted via "datafication" as opaque and unclear. It is partially not just because of datafication but also, thanks to the media and the effect of the Attention Economy (Davenport 2001). This effect needs still more and more shocking and negative news to attract the attention of its customers.

Besides the effect of the Attention Economy and its exaggeration, the datafication by itself is causing an extreme data flood. It creates a situation wherein people can lose faith in a better world because Big Data is confusing what is right and wrong.

(8) **Social Pressure** on users and their peers to use new services (Andrejevic, 2014). To keep one's place in the community or to become competitive, one is forced to use the current Big Data services. For the user, that exposes their personal data to Big Data analysis; bigger players extend the use of applications and services processing Big Data.

(9) Some people express a **Belief in Legislation**. They claim that proper and effective legislation can solve the majority of Big Data problems. However, it is widely agreed that legal regulation is mandatory but not sufficient, partially because it works mainly reactively in governing society. Lessig, Latour, Foucault, and Sokol among others focus on this theme.

(10) **End of theory:** while Big Data changes the definition of knowledge (Boyd and Crawford, 2012) and as described in 2008 by Chris Anderson in his famous article published in Wired magazine, it creates the so called "End of Theory":

- *"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves "* (Anderson, 2008, p 2).

(11) **Data Religion** adoration means more and more data can be traced by many professionals, e.g., (McAfee et al., 2012) although:

"Bigger data are not always better data" and "Big Data claims to objectivity and accuracy are misleading." (Boyd and Crawford, 2012).

(12) **Unawareness of Our Data** is the customers' unawareness of the data insight consolidated on the side of oligopolies (e.g. Mobile operators). For details, see our survey done as part of this article that follows the previous publication about possible use cases in telecommunication (Doucek, Pavlicek, Novak, Strizova, 2017).

The summary of the Big Data issues research is shown in the table below.

ID	Issue	**Issue description** (questions in our survey)	Source reference
1	**Privacy Intrusion**	Whether Big Data has an important impact on the privacy of individuals.	(Cavoukian, 2011), (Spiekermann, 2012), (Floridi, 2016)
2	**New Barriers**	Whether a respondent feels like Big Data creates barriers among society depending on Big Data availability.	(Norris, 2001), (DiMaggio& Hargittai, 2001), (Dijk, 2006)
3	**Business Advantage**	Whether a respondent feels like specific business advantages are available to corporations that actively collect and use Big Data.	(Cukier and Mayer-Schoenberger, 2013)
4	**Power of All Data**	Whether a respondent feels like there are only a few data monopolies (such as Google or Facebook) that can see the global view and predict the future.	(Andrejovic, 2016)
5	**New Big Brother Effect**	Whether a respondent feels like people are being observed by technologies all the time and their life can be manipulated without their knowledge.	(Orwell, 1961), (NPR, 2018), (Allcott, 2017), (Creemers, 2018)
6	**Missing Transparency**	Whether a respondent feels that due to complicated Big Data technologies, they lose transparency.	(O'Neil, 2016), (Haesler, 2013)
7	**Confusion**	Whether a respondent feels like Big Data cause confusion in determining what is right and wrong.	(Davenport, 2001), (Floridi, 2016)
8	**Social Pressure**	Whether a respondent feels that there is pressure on people to use new services that are used by others.	(Andrejovic, 2016)
9	**Belief in Legislation**	Whether a respondent believes that proper legal regulation can solve all Big Data problems.	Lessig, Latour, Foucault, Sokol among others.
10	**End of Theory**	Whether a respondent feels like it is not important to understand underlying principals but to be able to get results.	(Boyd and Crawford, 2012), (Anderson, 2008)
11	**Data Religion**	Whether a respondent feels like the quality of decisions depends only on how much data one is able to collect.	(McAfee et al., 2012). (Boyd and Crawford, 2012)
12	**Unawareness of Our Data**	Whether we are unaware of our data that are collected about us by service providers, e.g. Telco operator.	Authors, (Doucek, Pavlicek, Novak, Strizova, 2017)

Table 8 - Big Data issues list with references, (Author)

6.2.1 Categorization of Big Data Issues

The survey presented in the paper studies the awareness and importance of the issues reported by the respondents[38]. Based on the previous research, we can see that current literature largely focuses on issues such as Privacy Intrusion and the Big Brother Effect, although some issues are just occasionally mentioned but not attracting much attention (Belief in Legislation, et al) and some are noted by authors almost like hidden factors that do not deserve special attention (Missing Transparency, et al). Thus, we suggest differentiating and categorizing the issues into the following groups:

- **Hot issues**
  - There is almost unanimous agreement that the issue is important among the respondents.
- **Cold issues**
  - A strong majority of the respondents do not consider this issue to be important.
- **Warm issues**
  - The respondents are expected to be unaware or have no consensus about the importance of this issue.

6.3 Hypotheses Definitions

Following the previous definition of the twelve Big Data issues and inspired by similar surveys by Mark Andrejevic done across Australia in 2011 and Latonero & Sinnreich in the US in 2014, which showed interesting results differentiated based on demography (gender, age, education, and state), we formulated the following hypotheses that we would like to test in our paper:

H 1. It is possible to divide the Big Data issues into three consistent groups of cold, hot and warm issues based on their awareness among the population.

[38] We expect that awareness has a strong correlation to importance and understanding of complex problems and issues such as Big Data that was confirmed on other complex issues, e.g. Stock Markets (Guiso & Jappelli, 2005)

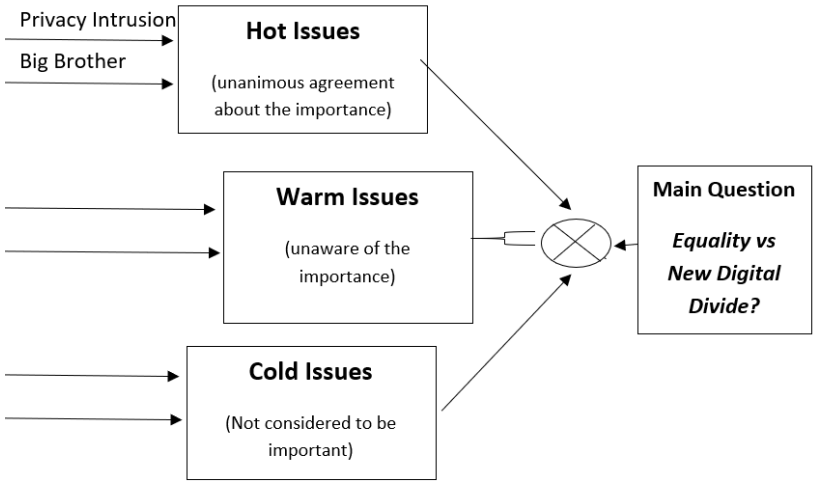


Figure 18 - General model of suggested categorization of issues into groups, *(Author)*

H 2. Privacy and Big Brother issues are the Big Data problems considered to be most critical by respondents and other issues are evaluated as much less important.

H 3. The awareness of personal data collected by Telco operators about people is not high and for some specific Telco practices the awareness is even below 50%.

H 4. There is very different awareness of Big Data clusters (hot, cold, warm) and of some specific issues among the population based on demography, especially:

a. IT Skills
b. Occupation
c. Sex

6.4 Methodology of Survey Evaluation and Hypotheses Testing

This subsection presents the methodology of the survey evaluation and the means of the hypotheses testing. For this purpose, various statistical methods were used. The initial insight was obtained via Exploratory Data Analysis (EDA).

In order to evaluate hypothesis H1, I used a Cluster Analysis based on the Ward method and Silhouette coefficient for measuring cluster consistency.

Hypothesis H2 and hypothesis H3 were tested by a t-test.

For evaluation of H4, I chose MANOVA Pillai's test and Welch t-test, F-test be relevant because we fulfilled all the assumptions of this statistical method.

7 Big Data and Ethics Survey

7.1 Methodology of Survey

The survey is considered as experimental and quasi-representative because of the sample selection. The goal of the survey is mapping the attitude of the respondents to the previously defined list of Big Data issues. For answers, we used the format of a typical five grade Likert scale (Strongly agree, Agree, Neutral, Disagree Strongly disagree) where we transformed originally ordinal variables to quantitative[39].

Because of the sample selection, we cannot apply the findings from the survey backwards easily to the whole population; however, we can describe and discuss different attitudes of different respondent groups based on the available dataset, which is the core benefit of our survey.

7.2 Data Set Description

The data were collected via an online questionnaire from the 26th of November 2018 till the 30th of March 2019. IT professionals were invited to the survey by direct email and IT students were required to participate during their lessons while seniors (pensioners) were approached mainly personally by the authors. The questionnaire was sent to respondents by "an authority" (a teacher/manager/family member), hence the response rate was high, 53%.

The **total number of observations** obtained from the survey was **903**. We removed 13 observations with low variability, indicating non-serious approaches to the survey. There were also 157 observations with incomplete demographic answers that were excluded from the dataset as well. Thus, the **final number** of observations discussed in the survey is **733**.

7.3 Questionnaire Structure

The structure of the questionnaire is the following:

A. Main question of Big Data survey

Q1: Do you think that new Big Data technologies impact equality among people?

B. Big Data analytical questions / Issues

[39] For possible transformation of ordinal data see e.g., (Rezankova, Novak, 2019).

The analytical questions are based on the list of Big Data Issues previously described and we ask for their awareness and importance.

C. Human values questions

These questions do investigate the attitude of respondents to the basic human values that are relevant to ask for in relation to previous Big Data analytical questions.

The values questions were inspired by research done in chapter 5, Regulatory Framework of Big Data, and namely the paragraphs discussing the human values that are based on Schwartz Theory of Basic Values (2012, Schwartz) and also EU Charter of Fundamental Rights (2007, EU 303/01).

I do differentiate the values that are recognized on the personal level of individuals and on the social level of the whole society.

Namely, I will ask for the following personal values: Human Dignity, Self-direction and Achievement, Pleasure and Security.

Among social values that we investigate in the survey are: Benevolence, Universalism, Equality and Solidarism.

D. Socio-demographic questions

This is a standard part of each questionnaire that enables statistical evaluation and later views on analytical questions based on the belonging of respondents to the relevant demographic categories.

We ask respondents about the following categories: age, sex, country of origin, occupation, place of living and level of IT experience.

7.4 Questionnaire of Big Data Ethics

At the beginning of the online questionnaire, the Survey Introduction is as follows:

"The following survey is focused on the Big Data phenomenon and awareness of its impact on society. The term Big Data is used in the survey as a general term covering applications, processes and infrastructure that are used for the collection and processing of data in big volumes, variety and velocity.

Big Data is herein used to cover both the technologies and also general trends in society where data are analysed with the purpose to influence and predict people´s behavior. Google, Facebook or mobile operators and some others are good examples of companies

operating in the field of Big Data. Based on their approach to Big Data, these companies can predict many aspects of society, for example, weather forecasts, share prices, voting preferences, relationship breakups and many others." (Author, Online Questionnaire at www.1ka.si)

Declaration and GDPR consent.

This survey is done only for academic purposes and there is no commercial interest.

By continuing, you agree with the processing of your answers for the purpose of their evaluation.

The main question has the following wording:

Main Question
Do you think that new Big Data technologies impact equality among people?

Table 9 – Main question wording (Author)

7.4.1 Big Data Analytical Questions / Issues

Below is shown the wording of the analytic question.

Issue	**How important are the following Big Data issues for you?**
Privacy Intrusion	Privacy intrusion.
New Barriers	Creation of barriers by Big Data dividing people with advantages from the rest.
Power of All Data	Only a few data monopolies such as Google, Facebook ...etc. are able to see the global view and are able to predict the future.
Business Advantage	Specific business advantages are available to a limited group of big corporations such as banks, mobile operators …etc.
New Big Brother Effect	It creates a situation in which we are observed by technologies all the time and our life can be influenced without our knowledge.
Issue	**How much do you agree with the following opinion?**
Missing Transparency	The world is managed by complicated mathematical algorithms that I increasingly do not understand.
Confusion	I am losing faith in a better world and future because Big Data is confusing what is right and wrong.

Social Pressure	Big Data creates social pressure on me to use new services that are used by others.
Belief in Legislation	Proper legal regulation can solve the majority of Big Data problems.
End of Theory	Today it is not essential to understand underlying principals deeply but to be able to show results in a lot of figures and graphs.
Data Religion	I think that the quality of my decisions is dependent only on how much data I am able to collect.

Table 10 - Big Data Analytical Questions wording / Issues 2-11, (Author)

A special area of the issues are the questions on the awareness of respondents about the reality of data collection by Telecommunication operators in the Czech and Slovak Republics investigated by the questions below.

Data Unawareness Issues.	**What data do you think are available to your mobile operator based on your contract?**
Location Tracing	Data about my **location** at almost every moment of the day.
Timing of Activities	**Record** of every call and text that I make.
List of Calls	Data about **phone numbers** with which I am calling or texting.
Web History	A list of **web pages** that I have **viewed** from my mobile.
Message Content	**The content** of my text messages.
Calls Recording	**Recordings** of my phone calls.

Table 11 - Big Data Analytical Questions wording / Issues 12, (Author)

7.4.2 Human values questions

The wording of values is based on questions that were inspired by the Schwartz theory and EU Charter of Fundamental right is shown below.

Value	**How important are the following values for you?**
Personal	**Human Dignity** *Respecting the physical and mental integrity of individuals.*
Personal	**Self-direction and Achievement** To be able to independently decide about my life and success.
Personal	**Pleasure** *To have a lot of pleasure in my life and to enjoy it.*

Personal	**Security** *Feel safe in my life.*
Social	**Benevolence** *Respecting needs of the closest group such as family and friends.*
Social	**Universalism** *Respecting needs of all people and the environment.*
Social	**Equality** *Equality of all people with no differences.*
Social	**Solidarism** *Helping members of society in their hardships.*

Table 13. Human Values Questions

Table 12 - Human Values Questions wording (Author)

7.4.3 Socio-demographic questions

The wording of socio-demographic questions with the list of possible answers is shown below.

1. **Where were you born?**
    - **a.** The Czech Republic
    - **b.** The Slovak Republic
    - **c.** The Netherlands
    - **d.** another European country / *please enter ……..*
    - **e.** outside Europe / *please enter ……..*
2. **Sex?**
    - **a.** Male
    - **b.** Female
3. **Region?**
    - **a.** Big city (more than 100,000 people)
    - **b.** Small city (between 5,000 and 100,000 people)
    - **c.** Rural (less than 5,000 people)
4. **Occupation?**
    - **a.** Student
    - **b.** IT professional
    - **c.** Other / Please write ……..
5. **IT experience?**

a. None - *I do not use computers regularly*

b. Low - *I am able to use the web and to email*

c. Medium - *I do well with tools such as Word, Excel and Powerpoint*

d. Advanced - *I can program and use command line, such as SQL query in databases*

e. Expert - *I regularly do object programming and also machine learning*

6. Age?

a. 20 years old or younger

b. 21-25 years old

c. 26-35 years old

d. 36-50 years old

e. over 51 years old

7.5 Survey Results

7.5.1 Exploratory Data Analysis

The two tables below show the basic descriptive data.

Sex		**IT Skills**		**Occupation**		**Working** experience (Age)		**Country** (of origin)	
Men	430	**Low** - **n**one to average	506	**Students**	608	**Trainees**	377	**Czechia**	**526**
Women	303			**IT Professionals**	100	**Juniors**	240	**Slovakia**	**102**
		High - advance to expert	227	**Pensioners**	25	**Seniors**	116	**other EU**	**43**
								outside EU	**62**
TOTAL									**733**

Table 13 - Demographic data, (Author)

Sex note: the ratio of 59% male and 41% female indicate that both genders are balanced.

IT Skill note: Although it was self-assessment, the questions were supported by specific examples, such as: **Advanced =** *I can program and use command line, such as SQL query in databases* or **Expert** = *I regularly do object programming and machine learning.*

Occupation note: The respondents were mainly students of the Faculty of Informatics at the University of Economics in Prague and employees of telecommunication companies such as T-Mobile Czech and Slovak Telekom.

Working Experience (Age) note: It is based on an age range: **Trainee** (17-20 years), **Junior**

(21-35 years) and **Senior** (> 35 years).

The variability of the dataset can be described by the average score of the issue (M) and its standard deviation (STD). These values can be used to detect the hot, cold and warm clusters of issues that were defined in the chapter: Categorization of Big Data Issues.

Big Data Issues	Main	Privacy.Intrusio	New.Barriers	Power.of.All.Data	Business.Advantage	New.Big.Brother	Missing.Transparency	Confusion	Social.Pressure	Belief.in.Legislation	End.of.Theory	Data.Religion
M	55,0	81,8	63,0	67,2	66,0	80,6	54,7	46,0	64,7	56,9	45,4	53,3
STD	22,5	24,8	25,3	25,0	23,1	24,7	28,1	28,4	27,3	26,1	28,3	28,6

Table 14 - Big Data Issues Awareness (M, STD), (Author)

The highest (M) is "Privacy" and "New Big Brother Effect", which corresponds to hypothesis H2. This hypothesis is tested by a two-sample t-test of the (M) result of Privacy, Big Brother (M>70) compared to the (M) of the remaining nine issues (M=57.5). The p-value < 2.2e-16, is less than the significance level 0.05. Hence, we can conclude H2 to be confirmed and that Privacy and Big Brother issues are considered to be more critical by respondents than other issues.

The List of HumN Values	Human.Dignity	Self.direction.and.Achievement	Pleasure	Security	Benevolence	Universalism	Equality	Solidarism
M	89,8	87,6	82,8	92,6	88,0	75,0	76,1	81,2
STD	16,0	17,5	19,1	14,7	17,7	21,5	26,1	20,5

Table 15 – Human Values Awareness[40] (M, STD), (Author)

[40] The Human Values based on Schwartz (2012) and EU Charter of Fundamental Rights are summarized in this table here and are not discussed further because my main focus was on Big Data issues and I did not discover any relevant relation among Human Values and Big Data Issues. I think it is caused by the questionnaire design where the original focus was put only on Big Data issues and extension (wording) of the values was not smoothly connected to the original Big Data issues.

The average score of human values is steadily high with the maximum for Security (92.6) followed by Human Dignity, Benevolence and Self-direction, which makes sense since people tend to present themselves as supporting basic human values in surveys.

The lowest score of values belongs to Universalism (75) and Equality (76.1), which correspond to our survey participants who were mostly young people that in general prefer individualism. The important findings are that the difference between the highest scored value of Security and lowest scored value of Universalism is pretty high (17.6) in a scale of 1-100 points.

The standard deviation score is consistently higher for issues opposite human values. The human values standard deviation is lower, but it varies more comparing to the standard deviation score of issues.

The standard deviation of the "Equality" value is the highest of all human values followed by Universalism and Solidarism. That implies the validity of the research whose main question investigates the conflict of Big Data and digital divide.

7.6 Cluster Analysis

We can rewrite the definition of hot, cold, and warm groups of issues from previous subsection using only the (M) and (STD) of the awareness from the survey as follows:

- *Hot issues*
    - The "unanimous agreement that the issue is important among the respondents" means that the (M) of questions regarding a hot issue is high, and its (STD) is low.
- *Cold issues*
    - The definition of cold issues "majority of the respondents do not consider this issue to be important" assumes that the (M) of questions regarding a cold factor is low. There is no assumption on (STD) since the definition considers the "majority of the respondents" and several outliers can cause a significant (STD) increase.
- *Warm issues*
    - This group of issues were defined as "respondents are expected to be unaware of the importance of warm issues", hence these issues are expected to have a medium (M). The "no consensus" can be statistically represented by a high variability of answers measured by the (STD).

By reformulating the definitions this way, a cluster analysis using Ward's Euclid distance between (M) and (STD) can be applied on the aggregated data collected in our survey about Big Data issues defined in Table 8. Note, other methods (e.g. average-linkage or

complete linkage) lead to identical clustering solutions. The results are visualized in the following figure.

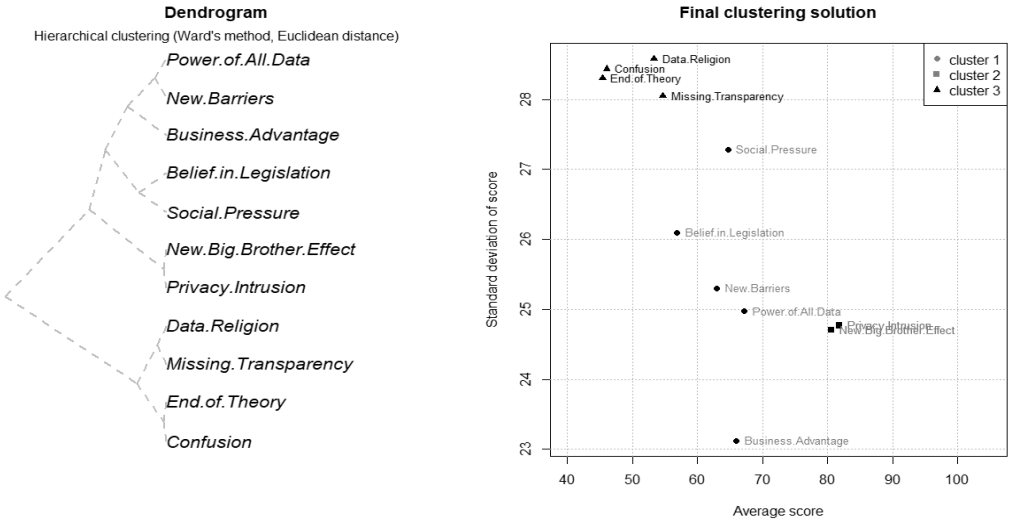


Figure 19 - Results of cluster analysis (Ward´s method), (Author)

The silhouette coefficient is used for the evaluation of the clustering solution consistence and performance. The silhouette coefficient is the average of each object's silhouette width which measures the degree of confidence in the clustering assignment of a particular object.[41]

The value of the average silhouette coefficient is 0.64, which means that the degree of confidence in the clustering assignment is high on average.

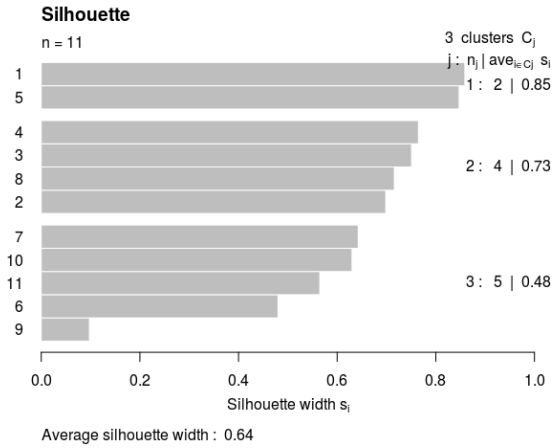


Figure 20 - Silhouette coefficient (average and clusters), (Author)

[41] (Rousseeuw, 1987)

Cluster 1 contains "Privacy Intrusion" and "New Big Brother Effect" (square-shaped points in Figure 2) has a high (M) and (relatively) low (STD). This cluster of issues fulfills the definition of a *cluster of hot issues*. The fact that these two issues were assigned to the cluster of hot issues supports the assumptions about ***hot issues***.

- "Privacy Intrusion" is a well-known issue thanks to the media (GDPR, wiki leaks, etc.) and its relevance to everyday-life.
- "New Big Brother Effect" is a widely popular issue and people, in general, are very familiar with this topic (Orwell's 1984, dictatorships, Czech-Slovakian experience with communism, etc.).

Cluster 3 *is a cluster of* ***cold issues:*** "Data Religion", "Missing Transparency", "End of Theory" and "Confusion" (triangle-shaped points in Figure 2). These issues are very specific and almost science-oriented. Hence the general public may not be aware of them.

- "Data Religion" investigates the role of data quantity in the decision-making process in the Big Data phenomenon, which might be unknown to a wide audience and therefore underestimated.
- "Missing Transparency" deals with mathematical algorithms that are distant from respondents' everyday-life problems.
- "End of Theory" observes the underlining principles of how the world is managed. That may be too philosophical for the respondents.
- "Confusion" represents the fact that Big Data makes the world too complicated. This issue is probably too abstract for the respondents.

Cluster 2 *is a cluster of* ***warm issues*** (round-shaped points in Figure 2). It contains the issues "Belief in Legislation", "New Barriers", "Power of All Data", "Social Pressure", "Business Advantage". The (M) in this cluster is medium, the (STD) varies from very low to high and that is why we will dive deeper in the MANOVA chapter.

Cluster analysis confirmed hypothesis H1: assigning issues into distinct and consistent groups based on their awareness.

7.7 MANOVA / Testing CA and Demography Impact

In this section we will evaluate hypothesis H4.

MANOVA (Multivariate Analysis of Variance) typically[42] helps to answer:

- Do changes in the independent variables (Sex, IT Skills) have significant effects on the dependent variables (hot, cold and warm clusters)?

We chose MANOVA to test H4 because we fulfilled all the assumptions of this statistical method such as normality, equality of variance and univariate outliers among others.

[42] Stevens (2002).

There are a few possible testing methods, such as: Samuel Stanley Wilks' or the Pillai-M. S. Bartlett trace, the Lawley-Hotelling trace, or Roy's greatest root. We chose the Pillai's test because the Pillai-Bartlett criterion = pooled effect variances, which are often considered the most robust and powerful test statistic and gives the most conservative F-statistic (French, 2008).

We have evaluated hypotheses below using the MANOVA test for all demography variables performing the test in R software with the following results:

- H0: Awareness of Hot, Cold, Warm clusters / issues are not dependent on demography.
- H1: Awareness of Hot, Cold, Warm clusters / issues are dependent on demography.

Dependent variables	Independent variable	Outcome/ Significance
Awareness of hot, cold, warm clusters	Work Experience	H0 not rejected
	Age	H0 not rejected
	Sex	**H0 rejected /***
	IT Skills	**H0 rejected /*****
	Region	H0 not rejected
	Occupation	**H0 rejected /****
	Telco awareness	**H0 not rejected** / Note 2
Note 1: *p-value 0.05, **p-value 0.01, ***p-value 0.001 Note 2: Partially see detail in chapter: Test of Unawareness of our Data		

Table 16 - MANOVA test results, (Author)

The conclusions of the MANOVA test above are that the Pillai test confirmed the dependence of clusters on independent variables of demography, IT Skills being the most significant (p-value 0.0001258), followed by Occupation (p-value 0.001365), and also Sex (p-value 0.01232).

Warm cluster "Outliers" and Demography	**Sex**			**IT Skills**			**Occupation**			
	Men	*Women*	*ABS Diff (M)*	*Low*	*High*	*ABS Diff (M)*	*student*	*IT Prof.*	*pension*	*ABS Diff (M)*
Social Pressure (M)	60,9	70,0	9,1***	67,2	59,0	8,2***	64,6	62,3	77,0	14,8*/**
Business Advantage (M)	66,6	65,1	1,5	64,9	68,3	3,4*	65,2	68,0	75,0	9,8*/*
Power of All Data (M)	67,25	67,15	0,1	65,4	71,4	6,0**	65,8	72,3	78,0	12,2-/**
Note: *p-value 0.05, **p-value 0.01, ***p-value 0.001										

Table 17 - Warm cluster "Outlier Issues" and demography, (Author)

The table above summarizes the hypotheses testing[43] about dependencies of warm issues and demography done in R software. We focused on this cluster because it has the lowest consistency (silhouette coefficient =0,48) and the selected issues are likely to change their cluster belongings (almost "Outliers").

We can conclude at p-value 0.05 that there are the following dependencies:

Occupation influences all three issues:

- where especially pensioners are a bit scared about the possible impact of all three Big Data issues and they evaluate the importance of all of them highly.

IT Skills influences the Social Pressure score and also Power of All Data:

- where we can see that high IT Skills (advanced to expert) lead to the belief that we can underestimate the importance of Big Data issues because of our expert knowledge.

Sex influences only Social Pressure:

- where women evaluate much higher importance of Social Pressure caused by Big Data than men.

7.8 Test of Unawareness of our Data

In our survey, we did a test (covered by issue 12) of the respondent's "unawareness" of the real situation evaluating Telco data practices focused mainly on Privacy and Big Brother questions that belong to the hot cluster. The exact questions and the test evaluation are described below.

Issue 12, *Unawareness our data.*	***Question:*** *What data do you think are available to your mobile operator based on your contract?*	**Positive Awareness SCORE**
Timing of activities	**Record** of every call and text that I make.	88%
Location Tracing	Data about my **location** at almost every moment of the day.	66%
Web history	A list of **web pages** that I have **viewed** from my mobile.	53%
Call Recording	**Recordings** of my phone calls.	38%
Message Content	**The content** of my text messages.	28%
	(M)	**55%**

Table 18 - Unawareness of our Data, Issue 12, (Author)

[43] Welch t-test, F-test applied to hypotheses of e.g., H0/Alt: Importance of Social Pressure for women is significantly higher than for men.

In average, the respondent correctly answered 55% of questions about Telco practices. We applied one sample t-test to a percentage of correctly answered questions to evaluate H3. The test has the null hypothesis that the population mean is equal to 0.5 (a respondent correctly answered 50% of Telco Operator questions) and the alternative hypothesis that it is less than 0.5. The p-value of 0.85 > than the significance level of 0.05 means we cannot reject the null hypothesis. However, when testing a percentage of correctly answered questions regarding only Call Recording and Message Content the p-value < 2.2e-16 means that for these two questions we cannot reject the null hypothesis. Hence regarding Hypothesis H3, we claim that even though the awareness of personal data collected by Telco operators is on average above 50%, there are some specific practices of Call Recording (M=28%) and Message Content availability (M=38%), where the awareness (positive knowledge of the practice) is below 50 %.

Following the findings from the cluster analysis, we also investigated the impact of demography on the Telco test score with the conclusions (at p-value 0,05) described below.

Working Experience impacts the Telco test score. An average difference between Working Experience (Seniors to Trainees) is about 17 % where the biggest difference is for Web History questions (34% better score of Seniors vs Trainees) followed by Location Tracing questions (31% better score of Seniors vs Trainees).

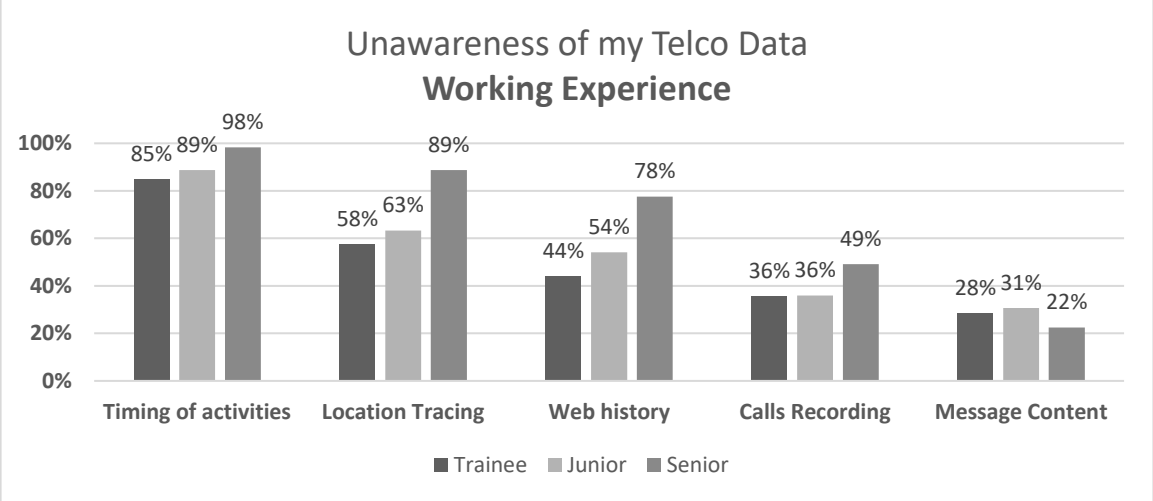


Figure 21 - Working Experience and Unawareness of our Data, (Author)

Sex impacts the Telco score in a way that Males are more aware than Females of their personal data collected by their Telco operator. The average difference is about 7% where the biggest difference is for the Web History data collection question (16% better score of Males) followed by Location Tracing (11% better score of males).

IT Skills impacts the Telco score with the average difference between high IT Skilled compared to low IT Skilled being about 10% while the biggest difference for the Location Tracing question (16% better score of high IT Skilled) followed by Web History and Call recording questions (11% better score of high IT Skilled).

The conclusion of the possible impact of demography on the results of the Telco test (at p-value 0.05) is the following: Work Experience (Age), IT Skills and Sex has an impact on the Telco test results (Issue 12 – Unawareness of our Data).

7.9 Correlation

	Main	Privacy Intrusion	New Barriers	Power of All Data	Business Advantage	New Big Brother Effect	Missing Transpar-ency	Confusion	Social Pressure
Main	1	,234**	,319**	,215**	,203**	,259**	,190**	,341**	,284**
Privacy Intrusion	,234**	1	,403**	,306**	,300**	,563**	,177**	,186**	,262**
New Barriers	,319**	,403**	1	,371**	,358**	,394**	,192**	,270**	,271**
Power of All Data	,215**	,306**	,371**	1	,532**	,362**	,192**	,154**	,244**
Business Advantage	,203**	,300**	,358**	,532**	1	,351**	,181**	,132**	,183**
New Big Brother Effect	,259**	,563**	,394**	,362**	,351**	1	,151**	,153**	,282**
Missing Transparency	,190**	,177**	,192**	,192**	,181**	,151**	1	,416**	,300**
Confusion	,341**	,186**	,270**	,154**	,132**	,153**	,416**	1	,396**
Social Pressure	,284**	,262**	,271**	,244**	,183**	,282**	,300**	,396**	1
Note: **p-value 0.01 *p-value 0.05									

Table 19 - Correlation among Issues (Spearman rho´s), (Author)

- Not surprisingly, the New Big Brother Effect and Privacy Intrusion are positively correlated (0.563). We assume that the respondents who value their privacy are afraid of losing their privacy due to the New Big Brother Effect.

- The survey showed a strong correlation between Business Advantage and Power of All Data (0.532). We assume that these issues are close because both observe the business advantage differentiating only between advantages available to data corporations belonging either to the TOP 1,000 or TOP 10.

	Main	Privacy Intrusion	New Barriers	Power of All Data	Business Advantage	New Big Brother	Missing Transparency	Confusion	Social Pressure	Belief in Legislation	End of Theory	Data Religion
Sex	-,004	-,102**	-,085*	-,001	,032	-,058	-,124**	-,199**	-,167**	-,072	,010	,044
Origin	,023	-,066	-,028	-,003	-,038	-,072	-,031	,078*	-,085*	,051	,122**	,090*
Region	,066	-,030	-,004	,016	,007	,005	-,035	-,084*	-,014	-,026	,043	-,039
Occupation	,037	,014	-,011	,122**	,076*	,047	,036	-,005	,034	,046	,182**	-,036
IT Skills	-,105**	-,013	-,068	,087*	,072	,026	-,115**	-,270**	-,121**	-,145**	-,093*	-,013
Age	,029	,027	-,032	,177**	,093*	,068	,016	-,074	,017	,028	,150**	-,048
Note: *p-value 0,1 and **p-value 0,05												

Table 20 - Correlation among Issues and Demography (Spearman rho´s), (Author)

The table above does not describe a strong direct correlation among issues and demography and that is why using MANOVA and testing hypotheses were so useful to uncover hidden patterns in our dataset.

7.10 Linear Regression

Linear regression analysis[44] was used to describe and explain the relationship between the answer of the main survey question (Big Data vs. Ethics) and the identified Big Data issues.

R-squared 0.232 explains around 23% of total data variability, which is not enough for a predictive model. However, it should be able to fulfill its descriptive purpose (e.g. Falk et. al, 1992 or Cohen, 1988) and confirms that in social sciences, the R-squared is usually not very high due to the random behavior of people. At p-value 0.05, only the issues of New Barriers, Confusion, Social Pressure and Belief in Legislation were significant. That shows that the regression model is not very suitable for the description of the cumulative effects of the issues on the main question and supports our approach stressing the differences among demographic groups.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,482[a]	,232	,219	19,809

a. Dependent Variable: Main,

[44] Done in SPSS and R software

b. Predictors: (Constant), Data Religion, New Big Brother Effect, Belief in Legislation, Missing Transparency, End of Theory, Social Pressure, Business Advantage, New Barriers, Confusion, Power of All Data, Privacy Intrusion

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	Sig.
1	(Constant)	25,577	3,632		,000
	Privacy Intrusion	,028	,040	,031	,482
	New Barriers	,166	,036	,186	,000
	Power of All Data	,049	,039	,054	,208
	Business Advantage	,042	,042	,043	,317
	New Big Brother Effect	,062	,041	,068	,135
	Missing Transparency	,001	,031	,001	,971
	Confusion	,205	,033	,260	,000
	Social Pressure	,075	,032	,091	,021
	Belief in Legislation	-,118	,030	-,137	,000
	End of Theory	,015	,030	,019	,612
	Data Religion	-,048	,028	-,061	,091

Table 21 – Linear Regression Model Summary (Author)

Table 22 – Linear Regression Coefficients, (Author)

7.11 Factor Analysis

Exploratory factor analysis with various rotations was applied to the data. The interpretable outcomes were obtained using varimax rotation. The figure below visualizes the results of the analysis where only the loadings higher than 0.3 are considered as relevant.

The analysis discovered factors of issues related to the personal life of individuals: harm to individuals (F1), unfair business advantages (F3) and a distorted world understanding (F2). But according to our interpretation, it did not uncover a hidden structure of a respondent's perception of given issues.

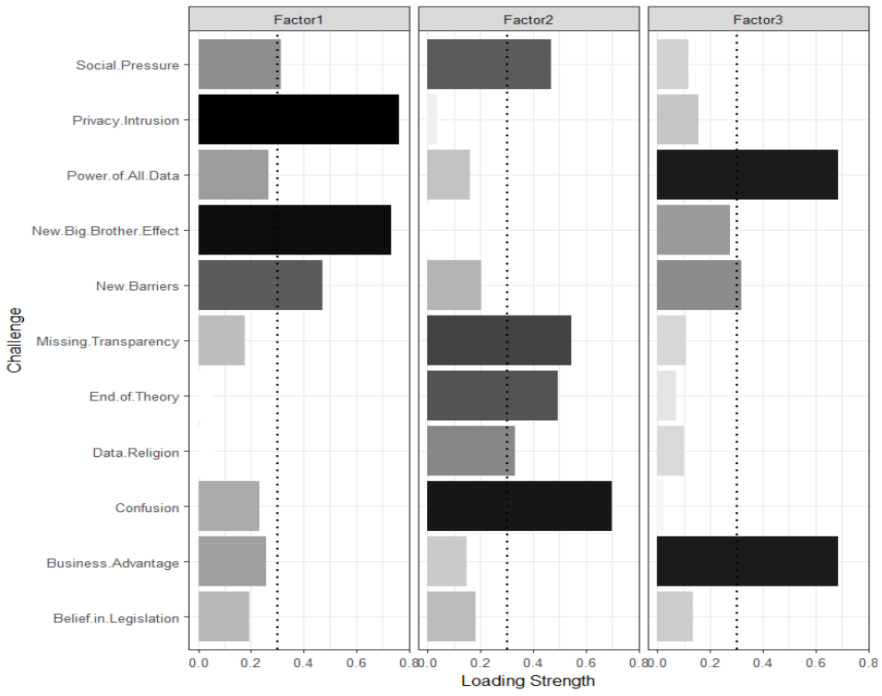


Figure 22 - Results of factors analysis, (Author)

8 Summary of Thesis

8.1 Summary of Theoretical Research

The research done by other authors is summarized in chapter 3, Big Data and Ethics Overview, and shows that there has been a long-term discussion about the impacts of new technologies on society and their ethical norms and values. Computer ethics have been evolving since the invention of computers in the 20th century after the world wars and has been described by many authors such as Norbert Wiener, Walter Maner or James Moor. However, the foundation of modern information ethics was laid down at the end of the 20th century by Rafael Capurro and Luciano Floridi.

During the last few years, there has been a visible shift from information ethics to data ethics based on the idea of two Oxfords academics, Luciano Floridi and Mariarosaria Taddeo, formulating as the following: *"We should concentrate on what is being handled (data) as the true invariant of our concerns and that is why labels such as 'robo-ethics' or 'machine ethics' miss the point.."* (Floridi & Taddeo, 2016).

The most recent definition of data ethics is from 2016 and was done by Floridi and Taddeo, approaching the topic on different levels of abstraction (LoA), such as macroethics, distinguishing the **ethics of data**, **algorithms** and **practices**.

I do agree with the findings outlined in the latest work of Floridi and Taddeo; however, in my opinion, there are still some yet untold Big Data specifics, which I discuss in my thesis in more detail.

By these specifics of Big Data, I mean the following:

- **Specific role of stakeholder groups** (organizations, users, state).
- **Use cases of Big Data,** (showing mainly positive benefits)
- **Demand for regulatory framework**
- **Conflicts and issues stemming from the clash between Big Data use cases and ethics**

I have dedicated a special chapter to each of these specifics of Big Data: chapter 3 discusses Stakeholder Groups, chapter 4 discusses Use Cases of Big Data, chapter 5 discusses Regulatory Framework of Big Data and chapter 6 discusses Digital Divide Conflict and Big Data Issues.

Based on the research done in chapter 6, I have defined my own list of twelve issues specific to the conflict of Big Data and ethics. These issues, whose importance I have

verified by an online survey. are the following: Privacy Intrusion, New Barriers, Business Advantage, Power of All data, New Big Brother effect, Missing Transparency, Confusion, Social Pressure, Belief in Legislation, End of Theory, Data Religion, Un-awareness of our Data.

8.2 Summary of Survey Results

The survey of the twelve newly defined Big Data issues confirmed that Big Data issues form three consistent clusters based on levels of awareness: hot, cold and warm clusters that confirm hypothesis H1.

Privacy and Big Brother are the issues deemed most critical by respondents compared to other issues, we assume due to media attention. Thus, hypothesis H2 was also proved.

A small test of data unawareness, focused mainly on Privacy and Big Brother issues (hot cluster), shows that only 55% of respondents are aware of what kind of data their Telco operator is collecting about them; however, in some practices the awareness is really low, e.g. call recording (38%) and text message content availability (28%). Overall hypothesis H3 was rejected; however, when testing a percentage of correctly answered questions regarding only Call Recording and Message Content availability the result is different. Hence regarding hypothesis H3, I claim that even though the awareness of personal data collected by a Telco operator is on average above 50 % (55% is still considered by us to be low but is above our testing level), there are some specific practices of Call Recording (M=28%) and Message Content availability (M=38%) where the awareness (positive knowledge of the practice) is below 50 %. The test also showed the impact of demography on results.

Hypothesis H4 about the impact of demography on clusters and issues awareness was proved using MANOVA test and Pillai function that confirmed the dependency of clusters on demography such as IT Skills, Occupation, and Sex.

I chose the warm cluster issues of the Social Pressure, Business Advantage and Power of All Data to investigate them deeper from the demographic perspective (silhouette coefficient =0,48 attracted my attention to these “outliers”) with the following findings: Occupation influences all three issues, where especially seniors are a bit scared about the possible impact of all three Big Data issues and they evaluate the importance of all of them highly. IT Skills influence the Social Pressure issue score and partially also Power of All Data, where we can see that high IT Skills lead to the belief that we can underestimate the importance of Big Data issues because of our expert knowledge. And finally, Sex influences Social Pressure, where women evaluate the Social Pressure issue caused by Big Data more importantly than men.

8.3 Summary of Regulatory Framework (BDEbD)

Chapter 5, Regulatory Framework of Big Data, describes four different areas how society can be regulated and governed, based on the inspiration of Harvard´s Lawrence Lessig. I have updated Lessig´s approach in relation to Big Data phenomenon. These general structures for possible regulative framework are described as the following: Market (general principals how to govern society), Law (Big Data Ethics by Default), Social norms and human values (e.g. EU Charter of Fundamental Right), Architecture (Big Data Ethics by Design, BDEbD, and description of other approaches to ethical assurance of Big Data systems).

I see the formation and clarification of the concept of Big Data Ethics by Design (BDEbD) discussed in the context of Big Data and Big Data Ethics by Default (Law), as my new contribution to data science. Although, the general Ethics by Design in its essence is an approach that has been known since the introduction of computers, namely computer ethics, the BDEbD is a very recent development that was not even using this name up to now. It could follow the Privacy by Design approach, that was in general already described in GDPR legislation; however, I have not found many relevant publications related to the theme BDEbD yet.

I believe that because of the fast-paced development of ICT and especially data science (data generation, algorithms, practices), there is no chance to manage or control these areas of ICT industry by Big Data Ethics by Default (Law) and the only way to guarantee the ethical principles in practice is the Big Data Ethics by Design approach, see e.g. the DEDA proposal. This should be supported by professional organizations of ICT experts and Data scientists where the membership should be so important that without it or in case of expulsion from such organizations, you practically cannot do your profession such as is the case of doctors of medicine and Camera Medica.

As the idea of Big Data Ethics by Design is in early stages, I appreciate, and suggest to follow, the recent work done at the Utrecht University and their DEDA (Data Ethics Decision Aid) methodology that is focused on the improvement of ethics mainly in commercial data projects. I have made the analysis of different BDEbD approaches and also my own suggestion for the DEDA methodology improvement described in chapter 5.

Based on my research I propose to move from ex post Big Data Ethic by Default (Law) to a priori Big Data Ethics by Design approach that should be inherent part of every ICT project that is related to data processing respecting the Big Data specifics described in chapter 3.4 as the following: the role of stakeholders (organizations, users, states), new use cases, growing demand for regulations and arising conflicts and issues in society caused by Big Data.

8.4 Benefits of the Thesis

In this chapter, I would like to conclude my contribution to the science described in this thesis. I see the contribution of the thesis as the following:

- Clarification and sorting of research of other authors relevant to Big Data and Ethics.
- Description of data sources and use cases from telecommunication showing the positive and negative effects of Big Data that can be applicable also to other industries.
- Discussion of the currently fragmented regulatory framework of Big Data and comparing two different approaches to ethical assurance applied either a priori or ex post to Big Data implementation. And formation and clarification of the concept Big Data Ethics by Design (BDEbD) that can be a new regulatory mean to practically govern and manage ethical conflicts in the Big Data era focusing on the early stages of IT and Data projects.
- A new proposal of twelve Big Data issues that are negatively impacting society and their categorization into clusters (Hot, Cold, Warm).
- Execution and evaluation of a large survey about the awareness of the Big Data and Ethics conflict that confirms the theory and hypotheses described in the thesis.

8.5 Discussion and Further Research

The previously published papers pertained to different Big Data issues, such as the Six Provocations of Big Data (Boyd and Crawford, 2012) or Digital Divide issues (Norris, 2001), (DiMaggio & Hargittai, 2001), (Dijk, 2006) (Deursen & Helsper, 2015) among others. In my thesis, I have summarized the previous research and suggested a comprehensive list of twelve issues that I categorized into three groups (Hot, Cold, Warm) based on a criterion of different awareness among people that we were able to evaluate with the help of statistical means in the survey. Although the survey was focused on different questions (newly defined twelve Big Data issues), it follows a similar approach to, e.g., the USA (Latonero & Sinnreich, 2014) or Australia (Andrejevic, 2014) where interesting findings were discovered showing the dependency of Big Data and ethics related issues on demography.

The conclusion of the thesis is that the awareness of Big Data issues can be grouped into three consistent clusters (Hot, Cold, Warm) that depend on a few demographic variables such as IT Skills, Occupation and Sex. I also conclude that there is a need in regulation

framework to move from ex post Data Ethic by Default (Law) to a priori Big Data Ethics by Design approach.

As a subject for further research, I suggest studying more deeply the difference between the approach of Big Data Ethics by Design vs by Default also including case studies and surveys of similar sizes that I did for my main focus area of Awareness of Big Data Issues. I also suggest studying Warm and Cold clusters of Big Data issues deeper and to analyze their demographic dependencies and related human values closely to be able to evaluate these currently underestimated issues related to Big Data, because I believe that their awareness and importance could rise, similar to the currently widely discussed issues of privacy and the Big Brother Effect.

9 References

9.1 Literature

Aberer, K., (2007). The semantic Web. In 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 ASWC 2007. Busan, Korea.

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281.

Allcott, Hunt, and Matthew Gentzkow. (2017). "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives, 31 (2): 211-36. DOI: 10.1257/jep.31.2.211

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07.

Andrejevic, M. (2014). Big data, big questions| the big data divide. International Journal of Communication, 8, 17.

Biltgen, Patrick; Ryan, Stephen (2016). Activity-Based Intelligence: Principles and Applications (1 ed.). Norwood, MA: Artech House. p. 151. ISBN 978-1-60807-876-9. Retrieved 6 May 2017.

Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. In: Information, Communication & Society. (pp. 662-679). Volume 15: Issue 5.

Bruce, Peter, and Andrew Bruce., (2017). Practical Statistics for Data Scientists. O'Reilly Media.

Buneman, P., (1997). Semistructured Data. Tutorial. In PODS '97, Symposium on Principles of Database Systems.

Bynum T. (2015). Computer and information ethics. In The Stanford encyclopedia of philosophy (ed. EN Zalta), Winter 2015. See http://plato.stanford.edu/archives/win2015/entries/ethicscomputer/.EN Zalta), Winter 2015. See http://plato.stanford.edu/archives/win2015/entries/ethicscomputer/.

Cadwalladr, C., & Graham-Harrison, E. (2018). The Cambridge Analytica Files. I made Steve Bannon's psychological warfare tool': meet the data war whistleblower.

CAFARELLA, M. & CUTTING, D. (2004). Building Nutch: Open Source Search. Magazine Queue - Search engines [online], 2004 Article. Available from: http://dl.acm.org/citation.cfm?id=988408 [cit. 2013-10-04]

Capurro, R. (2006). Towards an ontological foundation of information ethics. Ethics and information technology, 8(4), 175-186.

Cavoukian, A. (2011). Privacy by design in law, policy and practice. A white paper for regulators, decision-makers and policy-makers.

Cenek, P. (2012). Technologie počítačového zpracování řeči: Seminář GTS Czech: Telefonie a řečové technologie: GTS Czech seminar. Prague. OPTIMSYS, s.r.o.

Cibulkova, H., & Novak, R., & Sulc, Z. (2019). Multivariate Methods for Survey Evaluation: A Case Study of Big Data and New Digital Divide. In 22th International AMSE Conference, Nizna, Slovak Republic, in proceedings.

Clarke, R. (1994). Asimov's laws of robotics: Implications for information technology. 2. Computer, 27(1), 57-66.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Creemers, R (2018) China's social credit system: an evolving practice of control. Available at: https://ssrn.com/abstract=3175792;

Cukier, Kenneth; Mayer-Schoenberger, Viktor (2013). "The Rise of Big Data". Foreign Affairs (May/June): 28-40. Retrieved 24 January 2014.

Davenport, T. H., & Beck, J. C. (2001). The attention economy: Understanding the new currency of business. Harvard Business Press.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 48-55). IEEE.

Deursen, A. J., & Helsper, E. J. (2015). The third-level digital divide: Who benefits most from being online? In Communication and information technologies annual (pp. 29-52). Emerald Group Publishing Limited.

Deville, P. et al., (2014). Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. U. S. A. 111, 45, 15888–15893

Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. Poetics, 34(4-5), 221-235.

DiMaggio, P. and Hargittai, E. (2001). "From the 'digital divide' to 'digital inequality': Studying Internet use as penetration increases," Princeton University Center for Arts and Cultural Policy Studies, Working Paper Series number 15.

Donaldson, T., & Dunfee, T. W. (1994). Toward a unified conception of business ethics: Integrative social contracts theory. Academy of management review, 19(2), 252-284.

European Parliament. (2000). Charter of fundamental rights of the European Union. Office for Official Publications of the European Communities.

Falk, R & Miller, Nancy. (1992). A Primer for Soft Modeling. The University of Akron Press: Akron, OH.

Floridi L. (2008). The method of levels of abstraction. Minds Mach. 18, 303–329. (doi:10.1007/s11023-008-9113-7)

Floridi L. (2014). Open data, data protection, and group privacy. Philos. Technol. 27, 1-3. (doi:10.1007/s13347-014-0157-8)

Floridi, L. (2006). Information ethics, its nature and scope, ACM SIGCAS Computers and Society, 36 (3)

Floridi, L., & Taddeo, M. (2016). What is data ethics?

Foucault, M. (1991). The Foucault effect: Studies in governmentality. University of Chicago Press.

French, A., Macedo, M., Poulsen, J., Waterson, T., & Yu, A. (2008). Multivariate analysis of variance (MANOVA). San Francisco State University.

Friedman, B. and H. Nissenbaum (1996), “Bias in Computer Systems,” ACM Transactions on Information Systems, 14(3): 330–347.

Friedman, B., & Kahn Jr, P. H. (2007). Human values, ethics, and design. In The human-computer interaction handbook (pp. 1223-1248). CRC Press.

Gartner Group. (2018). Retrieved from: www.gartner.com

Gerloch, A. (2001). Teorie práva. Dobrá voda: Aleš Čeněk

Gilbert, E. & Karahalios, K. (2009). Widespread Worry and the Stock Market. Department of Computer Science, University of Illinois at Urbana-Champaign. Available from: www.aaai.org/ocs/index.php/ICWSM/

Guiso, L., & Jappelli, T. (2005). Awareness and stock market participation. Review of Finance, 9(4), 537-567.

Han, J., Kamber, M. & Pei, J., (2012). Data mining: concepts and techniques 3rd ed., Waltham: Morgan Kaufmann.

Hansmann, H., & Kraakman, R. (2004). What is corporate law?. Yale Law & Economics Research Paper, (300).

HARGITTAI, Eszter. (2002) Second-Level Digital Divide: Differences in People's Online Skills. First Monday, [S.l.], apr. 2002. ISSN 13960466. Available at: https://firstmonday.org/ojs/index.php/fm/article/view/942/864>. Date accessed: 06 nov. 2018. doi:https://doi.org/10.5210/fm.v7i4.942.

Hilbert, M. (2013). Big data for development: From information-to knowledge societies. Available at SSRN 2205145.

Hortonworks & Teradata. (2013). Hadoop & Data Warehouse: When to use which?. IN: Baldesch, E. & Brobst, S.. Webinar [online]. Available from: http://hortonworks.com/webinars/#library [Accessed 2013-10-04]

Chalmers, David, (1997). The Conscious Mind: In Search of a Fundamental Theory. Oxford: Oxford University Press. p. 225. ISBN 978-0195105537.

Chen, H., Chiang, R., & Storey., V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. In: MIS Quarterly 36(4). (pp. 25). Germany: Inivesitat Trier.

Chen, J. (2012), Cyberethics. University of British Columbia. January 2012. http://etec.ctlt.ubc.ca/510wiki/Cyberethics.

International Telecommunication Union ITU, (2014). World Telecommunication Development Conference (WTDC-14): Final Report, http://handle.itu.int/11.1002/pub/809f5219-en.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani., (2014). An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

Katz J.E. and Rice R.E., (2002). Social Consequences of Internet Use: Access, Involvement and Interaction. Cambridge, Mass.: MIT Press.

Kling, R. (1998). "Technological and Social Access on Computing, Information and Communication Technologies," White Paper for Presidential Advisory Committee on High-Performance Computing and Communications, Information Technology, and the Next Generation Internet, at http://www.slis.indiana.edu/kling/pubs/NGI.htm, accessed 26 March 2002.

Kokolakis, S. (2017) ‘Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon’, Computers & Security, Vol. 64, no. 1, pp. 122-134. http://dx.doi.org/10.1016/0047-2352(87)90075-4

KUMAR, A. & Et. (2009). Predicting the Outcome of Events Based on Related Internet Activity: [patent]. USA. 706/46, 709/224. IN: YAHOO! [online]. Available from: http://www.google.com/patents/US20100205131 [Accessed 2013-10-04]

Latour, B. (2009). 'Tarde's idea of quantification', in The Social After Gabriel Tarde: Debates and Assessments, ed M. Candea, London: Routledge, pp. 145-162. [online] Available at: http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf (19 June 2011)

Lessig. L., (1999). Code and Other Laws of Cyberspace. Basic Books, Inc., New York, NY, USA.

Lessig, L. (1999). Code is law. The Industry Standard, 18.

Maner, W. (1980). Starter kit in computer ethics. Hyde Park, NY: Helvetia Press and the National Information and Resource Center for Teaching Philosophy.

Manyika, J. et al., (2011). Big Data: The next frontier for innovation, competition, and productivity. US: McKinsey Global Institute

Marshall, C. & Edward, C. (2013). NSA Surveillance Leaks: Background and. Congressional Research Service [online], July. Available from: http://webcache.googleusercontent.com/search?q=cache:gLrJVs6JuxIJ:www.fas.org/sgp/crs/intel/R43134.pdf+&cd=1&hl=cs&ct=clnk&gl=cz [cit. 2013-10-04]

Maslow, A. (1943). A theory of human motivation. Psychological Review, 50(4), 370–96. Available from: http://psychclassics.yorku.ca/Maslow/motivation.htm

Mateos, P., Fisher, P.F. (2006). Spatiotemporal accuracy in mobile phone location: Assessing the new cellular geography. In: Drummond, J. et al. (eds.) Dynamic and Mobile GIS: Investigating Changes in Space and Time. pp. 188–211 Taylor & Francis

Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (2016). Autonomous driving. Springer Berlin Heidelberg, Berlin, Germany.

McAfee, Andrew, et al., (2012). Big data: the management revolution. Harvard business review, 90(10), pp.60-68.

Merikle, P.M., (1984). Toward a definition of awareness. Bulletin of the Psychonomic Society, 22(5), pp.449-450.

Molnar, Z., (2012). Competitive intelligence, aneb, jak získat konkurenční výhodu, Praha: Oeconomica, ISBN: 978-80-245-1908-1.

Mountain David, Raper Jonathan. (2001). Positioning techniques for location-based services (LBS): characteristics and limitations of proposed solutions. Aslib Proc. 53, 10, 404–412

National Telecommunications and Information Administration, (1999). "Falling through the net: Defining the digital divide," at http://www.ntia.doc.gov/ntiahome/fttn99/contents.html, accessed 25 March 2002.

Norris, P. (2001). Digital Divide: Civic Engagement, Information Poverty and the Internet in Democratic Societies. New York: Cambridge University Press.

Novák, R. (2014). Big data a možné přínosy v úloze strategického řízení podnikové informatiky. Systémová Integrace, 21.

Novak, R. (2014). Big data and legal regulation. In IDIMT 2014: (Networking Societies - Cooperation and Conflict). (p. p. 153--162). Linz: Trauner: ISBN 978-3-99033-340-2.

NPR, National Public Radio, (2018), China´s Brave New Wo rld, On 11 October 2018 at https://www.npr.org/templates/transcript/transcript.php?storyId=656627240

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

O'Neil, Cathy; Schutt, Rachel (2013). Doing Data Science. O'Reilly Media. p. 406. ISBN 978-1-4493-5865-5.

Orwell, G. 1984 (New York: Signet, 1961)

Pavlicek, A., Doucek, P., Novák, R., & Strizova, V. (2017). Big data analytics–geolocation from the perspective of mobile network operator. In International Conference on Research and Practical Issues of Enterprise Information Systems (pp. 119-131). Springer, Cham.Ethical world. (2018). Retrieved from: https://medium.com/the-ethical-world/ethics-defined-33a1a6cc3064

Pavlicek, A. Novak, R. (2015), Big Data from the perspective of data sources, In SMSIS conference, Ostrava 2015.

Procházka, J. (2012). Prediktivní analýza v prostředí internetu. Bakalářská práce. Brno: MASARYKOVA UNIVERZITA - Filosofická fakulta. Available from: http://is.muni.cz/th/362031/ff_b_a2/?lang=en

Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.

Rezankova, H., & Novak, R. (2019). Effect of ordinal variable transformations on hierarchical clustering results – case study on big data phenomenon. In 22th International AMSE Conference, Nizna, Slovak Republic, in proceedings.

Rosický, A. (2011). Konceptuální myšlení a změna paradigmatu:. IN: Systémové přístupy ‘11. Praha: Vysoká škola ekonomická v Praze, pp. 108-129.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Schäfer, M. T., Franzke, A., Utrecht, G., & Fransen, R. (2012). DEDA.

Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. Online Readings in Psychology and Culture, 2(1). https://doi.org/10.9707/2307-0919.1116

Sigmund, T. (2013). Ethics in the Cyberspace. IDIMT-2013 Information Technology Human Values, Innovation and Economy, 42, 269-279.

Sigmund, T. (2015). Do we need information ethics? In: IDIMT-2015: Information Technology and Society Interaction and Interdependence. (pp. 289-294). Linz: Trauner Verlag Universität.

Skrbek, J. (2009). New Possibilities of Information Services for Special Situations. IDIMT-2009 Syst. Hum. Complex Relatsh. 29, 123–130.

Skrbek, J., Kvíz, J. (2010). Critical Areas of Early Warning System. In: Doucek, P. et al. (eds.) IDIMT-2010: Information Technology - Human Values, Innovation and Economy. pp. 193–202 Universitatsverlag Rudolf Trauner, Linz

Sokol, J. (2016). Ethic, Life and Institutions, Praha: Vyšehrad

Spiekermann-Hoff, S. (2012). The challenges of privacy by design. Communications of the ACM (CACM), 55(7), 34-37.

Stevens, J. P. (2002). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erblaum.

Stowell, S. (2014). Hypothesis Testing. In Using R for Statistics (pp. 143-162). Apress, Berkeley, CA.

Taylor, L., Floridi, L., & Van der Sloot, B. (Eds.). (2016). Group privacy: New challenges of data technologies (Vol. 126). Springer.

TTG. (2014). Big data: Pomohou cestovnímu ruchu na Šumavě? | TTG - vše o cestovním ruchu, http://www.ttg.cz/big-data-pomohou-cestovnimu-ruchu-na-sumave/

Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.

van Deursen, A. J., van der Zeeuw, A., de Boer, P., Jansen, G., & van Rompay, T. (2019). Digital inequalities in the Internet of Things: differences in attitudes, material access, skills, and usage. Information, Communication & Society, 1-19.

Wang, T. H., & Cheng, H. Y. (2019). Problematic Internet use among elementary school students: prevalence and risk factors. Information, Communication & Society, 1-22.

Wiener, N. (1948). Cybernetics. Scientific American, 179(5), 14-19.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). The role and limits of principles in AI ethics: towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195-200). ACM.

Zetie, C. (2004). Location Services Find Their Way To The Enterprise, http://www.informationweek.com/location-services-find-their-way-to-the/26100784.

Zicari, R. (2019). Z-inspection: Towards a process to assess Ethical AI. [Online]. In Cognitive Science. US: Cognitive Systems Institute. Retrieved from http://cognitive-science.info/wp-content/uploads/2019/10/CSIGTalkZicari.20191031.pdf

9.2 List of Tables

9.3 List of Figures

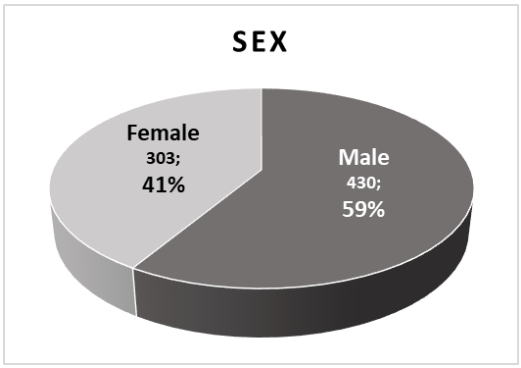
9.4 List of Abbreviations

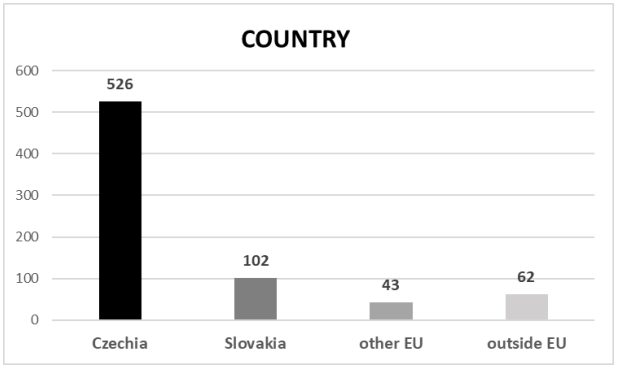
AI	Artificial Intelligence
ARPU	Average Revenue Per User
BI	Business Intelligence
BTS	Base Transmitting Towers
CA	Cluster Analysis
COBIT	Control Objectives for Information and Related Technologies
COSO	Committee of Sponsoring Organizations of the Treadway Commission
CRM	Customer Relationship Management
DAMA	Data Management Association
DB	DataBase
BDEbD	Big Data Ethics by Design
DEDA	Data Ethics Decision Aid
DMBOK	Data Management Body of Knowledge
ERP	Enterprise Resource Planning
EU	European Union
GDPR	General Data Protection Regulation
GTAG	Global Technology Audit Guide
H0	Null Hypothesis
Hx as H1	Alternative Hypotheses
ICT	Information Communication Technologies
ISACA	Information Systems Audit and Control Association
ISMS	Information Security Management Systems

ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
IT	Iinformation Technology
ITIL	Information Technology Infrastructure Library
ITSM	IT Service Management
KGI	Key Global Indicators
KPI	Key Performance Indicators
LBS	Location Based Services
LoA	Levels of Abstraction
M as (M)	Mean Value
MANOVA	Multivariate analysis of variance
MNO	Mobile network operators
NTIA	National Telecommunication and Information Administration
PbD	Privacy by Design
p-value	In statistics, the p-value is the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct. It is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.
RDBS	Relational Data Base System
SaaS	Software as a Service
SIM	Subscriber Identification Module
STD	Standard Deviation
SW	Software
Telco	Telecommunication

10 Annex – Exploratory Data Analysis (Graphs)

All Graphs below are produced as result of survey related to author thesis.

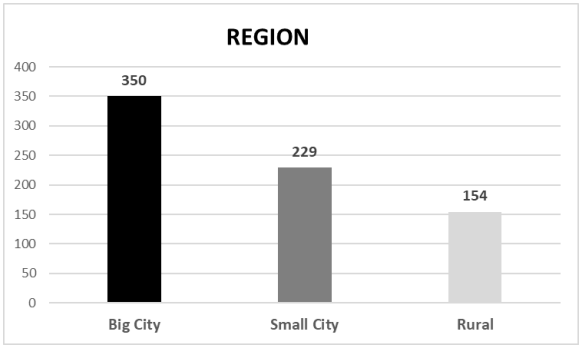


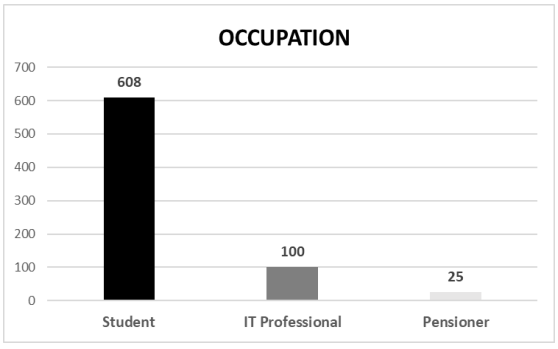


Graph 1. Gender of respondents

Graph 2. Country of Origin

Ratio of 59% males and 41% females means that both genders have enough respondents which is fine for further analysis. Country means place of origin where the respondents were born.



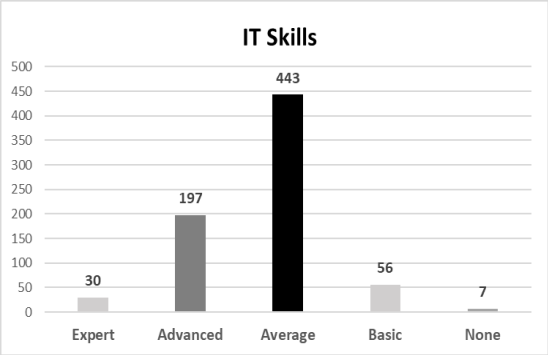


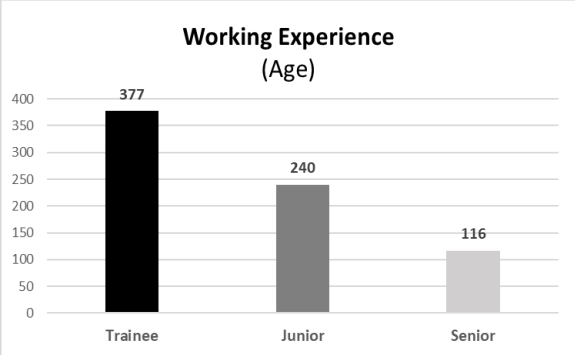
Graph 3. Region of residence

Graph 4. Occupation

Big city means population above 100 000 people, Small city means population from 5 000 to 100 000 people and rural means population below 5 000 people.

The respondents were mainly students of Faculty of Informatics at University of Economics in Prague and also employees of telecommunication companies such as T-Mobile Czech and Slovak Telekom.



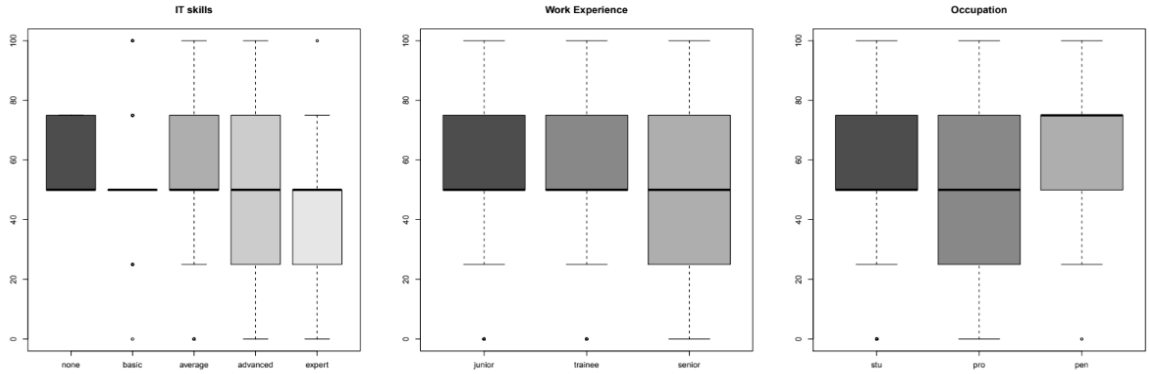


Graph 5, IT Skills

Graph 6, Working Experience (Age)

IT Skill was self-assessment; however, the questions were formulated with very guided examples such as: Advanced = I can program and use command line, such as SQL query in databases or Expert = I regularly do object programming and also machine learning.

Working Experience (Age) was based on age range where Trainee means between 17-20 years, Junior means between 21-35 years and Senior means above 35 years.

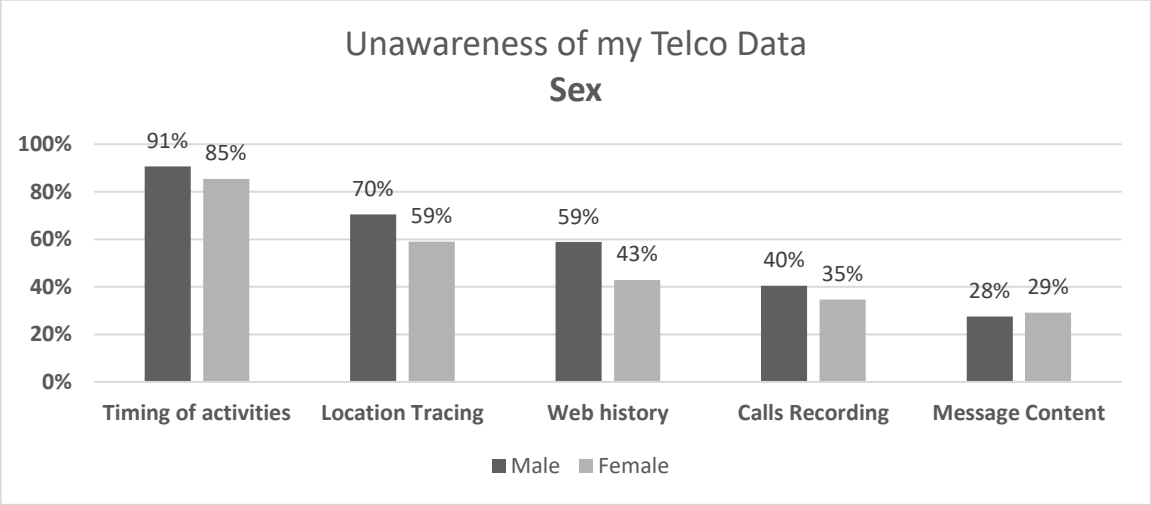


Graph 7. Main (Big Data/ Equality) and IT Skills, Working Experience, Occupation

The Boxplots Graphs above relating the main question of Big Data and Equality (Main) and IT Skills shows that more IT Skilled people such us advanced and expert do not take Big Data and equality conflict so seriously and surprisingly low experienced people (none to average IT Skills) do concern about equality and digital divide issues more.

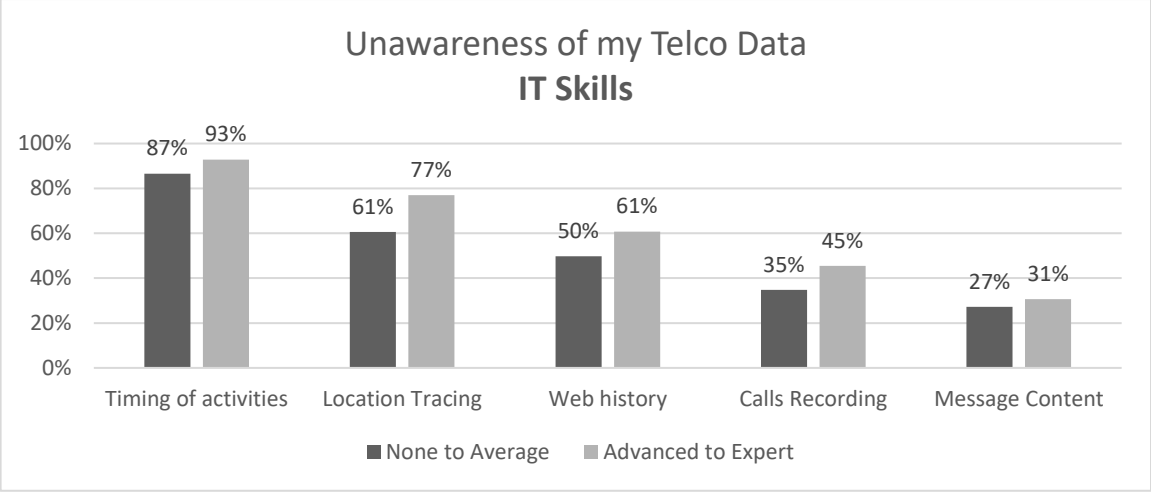
The Boxplot of the Main and Working Experience shows that younger and less work experienced demographic (trainee and juniors) are more aware of equality and digital divide than senior people.

The Boxplot of the Main and Occupation shows that IT Professional underestimate the conflict of Big Data and Equality/ Digital divide compared to university students and pensioners.



Graph 8. Sex and Unawareness of our Data (Telco testing example)

From the graph above it is visible that Males are more aware than Females of their personal data collected by their Telco operator. The average difference is about 7% where the biggest difference is for Web history data collection question (16% better score of Males vs Females) followed by Location tracing (11% better score of Males vs Females).



Graph 9. IT Skills and Unawareness of our Data (Telco testing example)

The average difference between IT Skilled (Advanced to Expert) compared to IT Un-Skilled (None to Average) is about 10 % where the biggest difference if for Location tracing question (16% better score of IT Skilled) followed by Web history and Calls recording questions (11% better score of IT Skilled vs Un-Skilled).