FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study program/specialization: Industrial Economics Entrepreneurship and Technology Management	Spring semester, 2017 Open
Author: Marius Engan	..... (signature of author)
Internal supervisor: David Häger	
Title of master's thesis: Big Data & GDPR	
Credits: 30	
Keywords: Big Data, GDPR, Privacy, Smart Meter	Number of pages: 133 + supplemental material/other: 39 Stavanger, June 15th, 2017

Title page for Master's Thesis

Faculty of Science and Technology

Master Thesis

Master of Industrial Economics

Big Data and GDPR

A study of how the the new EU data protection regulation impact the value of data

Universitetet i Stavanger

Marius Engan

University of Stavanger

June 15, 2017

Abstract

The European Commission has implemented the General Data Protection Regulation (GDPR) which will replace the current, but obsolete, Data Protection Directive 95/46/EC. When legally effective, May 25th 2018, it will impose a much stricter regime and sanctions which magnitude may force bankruptcy.

It increases dramatically the scope of what is considered personal data while restricting the processing as such. Thus, curtailing businesses' opportunity to drive value through big data analytics. In an increasingly data-driven economy, where data is drawn in the same breath as competitive advantage, it may seem like the candle is burned at both ends. Pursuant to the issue a question arises to whether the value of data will diminish. Consequently, this work researches how the GDPR will impact the value of data, with an emphasis on value driven trough the big data value chain.

The research is carried out in three phase: A preliminary analysis that identifies a set of value drivers; a primary analysis that identifies influences from the GDPR on said value drivers; and a case study on smart meter data. The results are presented as five assertions which make up the foundation of a discussion.

The research finds that the short-term impact raises concern to limitations put on: realizing value in public interest; harnessing the power of algorithms in automated decision-making; and discovery of new knowledge through data mining. However, the positive long-term impact are expected to overshadow the negatives and to ensure a sustainable data-economy in the future.

The research concerns a legislation that is yet to be enforced. The results are therefore predictions rather than hard facts, but will serve as insight to possible future challenges.

Acknowledgement

The topic of this thesis is a result of my experiences the last two years while studying for my masters degree. Big data has become a passion of mine and I am incredibly happy to be able to finish my education this way.

First and foremost I wish to express my gratitude to my supervisor David Häger for great guidance throughout the process. I am especially gratefully for wise words that made me choose a topic of my own liking. This has made this journey a painfully pleasant.

This page has been left intentionally blank.

Table of Contents

This page has been left intentionally blank.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Increasing processing power and a drastic fall in cost of data storage has combined with a formidable increase in devices connected to the Internet created an explosion of data. The understanding of data as information in a digital format is no longer applicable. Where data used to be the information submitted in a registration form online, it has become so much more. It is location tracking on smartphones, social media activity, health monitoring sensors and credit card transactions. This is data that is continuously generated by consumers around the globe creating massive volumes of a wide variety of formats that needs to be processed in real-time. This phenomenon has been given the name *big data*.[1] Businesses have realize the opportunities therein and analogies such as "data is the new oil" has emerged. Joris Toonders from Wired Magazine suggests that:

> "*Data in the 21st century is like oil in the 18th century: an immensely, untapped valuable asset. Like oil, for those who see data's fundamental value and learn to extract and use it there will be huge rewards.*"

Like oil, data needs to be collected, or drilled for, in order to be obtained, but has little use value in its raw form. Only when processed and analysed, or refined, a potential use value is created. Whereas oil may take the form of fuel or plastic, data may take the form of knowledge, predictions or intelligence. The ultimate value from oil products are realized

when fuel is transformed into into kinetic energy or plastic sold as goods. The value of data on the other hand is realized when a business processes are optimized, a product improved or insights sold for others to benefit. There is, however, one fundamental difference; data does neither suffer from transactional limitations nor depreciate with usage. In fact, data can be used simultaneously across multiple use cases and appreciates with usage,[3] unlike fuel, which is bought once and sold once.

The true value of data is not visible at first sight, but innovative companies with the right tools to aid them are able to extract this value. The value of data must be seen as all the ways it can be employed in the future not the past. After its primary use, data's value still exists, but lies dormant – like a spring – the value is released anew when used for a secondary purpose.[4] The ultimate value obtained can therefore be seen as all the possible ways it can be used and the respective output of each use. This is the *notion behind the value of data*.

Big Data is not one technology, but the combination of a number of traditional and modern technologies that are able to handle the increasingly complex data environment in ways that traditional computing are unable to handle.[1] The technology stack that makes up big data is immense and this thesis has a particular focus on analytics. Big data analytics are the composition of traditional analytics and advanced analytics such as *data mining*, *machine learning* and *artificial intelligence*, which has enabled businesses to gain knowledge and perform tasks that is humanly impossible. The intertwined use of big data analytics transform data from one form to another allowing innovative minds to discover new use and business models, thus enabling companies to realize the full potential.

The potential for data to change how business is conducted and the public is served is indisputable.[4] Setting the stage for a data driven economy are the tech giants Google, Amazon, Facebook and Apple, whom has built an economy equalling the GDP of Denmark by utilizing consumer data to extract knowledge and insight.[5] Data, and consumer generated data in particular, has become a gold mine and is no longer the domain for IT departments but is rather becoming a centerpiece of value.

However, as with the oil industry, the evolving data economy has a dark side. Oil spills from run-aground tankers and platform accidents cause immense damage to the environment. Meanwhile in the data economy, data breaches leave millions of people vulnerable, algo-

rithms discriminate on race and religion and personal information is widely dispersed for companies to exploit. Tech giants such as Google are like oil tankers navigating through a sea of cyber-criminals and human error just waiting for an unavoidable environmental catastrophe.[6]

Just like environmental law aims to safeguard the environment, data protection law aims to safeguard the privacy and the rights and freedoms of individual persons. It has however, become increasingly obvious that, due to recent technological advances, current privacy legislations has become obsolete. The European Union has consequently developed a new regulative reform, namely the General Data Protection Regulative(GDPR), whose primary purpose is to give citizens back control over of their personal data, and to simplify the regulatory environment, making it more comprehensible for companies to safeguard the privacy of individuals without sacrificing profits. A failure to comply, will ultimately lead to potential fines at the magnitude of 20 million Euros or 4% of global revenue. Companies will be required to implement technological and organisational measures in order to ensure compliance and are consequently facing a comprehensive change.

The GDPR sets the stage for an enhanced digital market in the EU by building a single, strong and comprehensive set of data protection rules. It aims to boost innovation in sustainable data services, enhancing legal certainty and strengthening trust in the digital marketplace and not to be a burden to innovation. The GDPR wants to be an enabler for big data services in Europe and provides a framework whose purpose is to balance between the protection of fundamental rights, customer trust and economical growth.[7]

1.2 Problem Definition

When the GDPR becomes legally effective May 25th 2018 it will change the legal conditions in the data economy by imposing requirements and regulations to how personal data is processed and used. It increases dramatically the scope of what is considered personal data while restricting the processing as such. Thus, curtailing businesses' opportunity to drive value through big data analytics. Furthermore, it lays down new ground rules to how businesses can use data after its primary purpose and will therefore restrict businesses' ability to realize the value through new use. Based on the above this thesis has arrived at

the following hypothesis:

> *When the GDPR becomes legally effective May 25th 2018 it will: curtail businesses' opportunity to drive value through big data analytics; restrict businesses' opportunity to realize the value; and will ultimately lead to a diminished value of data.*

The European Commission[8] do, however, recognize the potential of data-driven technologies, services, and in particular, big data as catalysts for economic growth, innovation and digitization. Hence the intention of implementing the GDPR is not to restrict data as a source of value but rather to ensure that the value created is expedient for the greater good and not at the expense of individual members of society. However, by imposing restrictions to generate business value while requiring new investments in security infrastructure, new employments and organisational change, it may seem like the implementation of GDPR is burning the candle at both ends. The purpose of this thesis is therefore to investigate this issues further. Hence the following problem statement shall be answered:

> *How will the GDPR impact the value of data?*

To answer this the problem statement a subset of questions must be answered:

1. How are value created from data?
2. How can this data be used and reused?
3. What changes are imposed by the GDPR?
4. How does the changes apply to big data?

1.2.1 Problem Solution

How are value created from data?

The chosen approach to answer this question is to carry out a preliminary analysis of how value is created within the framework of the big data value chain. This will be used as a foundation for assessing the case. The preliminary analysis is a literature review that

arrives at particular theories that apply to each stage of the analysis. From those a set of value drivers will be identified.

How can this data be used and reused

The value is realized through use. Therefore a survey on potential ways to use data will be carried out. This will require a general understanding of how data can be used as well as particular use cases. A business case will therefore be chosen. This case will also be used as a verification of the value drivers identified.

What changes are imposed by the GDPR?

In order to answer this question the thesis will have to rely heavily on papers and articles published by law firms and the European Union. This is mainly due to the legal language and wording of the GDPR. Any own interpretations is therefore subjected to some degree of uncertainty and will be noted as it occurs throughout.

How does the changes apply to data?

The approach to answer this question constitute of two parts: First the knowledge obtained at this point will be used in combination with supporting literature to assess key characteristics of big data that is particularly subjected to the legal regimen of the GDPR.

Subsequently a case study will be carried out. This involves analyzing how value is created throughout the big data value chain by applying case specific applications and assess them against the value drivers and the findings from the previous analysis. The findings will be the foundation for a discussion.

1.3 Presenting the Case

The context of this thesis can be briefly summarized with three key words: Big data, privacy and European policy. This thesis has therefore chosen a case suitable within this context.

Smart meters has become the symbol of the fusion between IT and energy, as they are capable of performing automated electricity meter readings every 15 to 60 minutes opposed to traditional manual readings every one to two months. This increases the data collection 3000-fold and poses new challenges and opportunities to the utility sector.

The roll-out of smart meters mark the *digitization* of the power sector and the enablement of big data analytics to create a more reliable, efficient and environmentally friendly power grid.[9] However, the roll-out of smart meters has created a discord between consumers and those processing the the data. Namely, they provide a "gateway to the home".[10] On one hand, smart meters present a privacy risk, as they have the capability to monitor and predict behavior of residents that can be used in ways that is in breach of human rights and privacy law.[11] While on the other hand presenting a potential "gold mine" of data for utilities to utilize, both in ethical and unethical ways.

The successful roll-out and effectuation are, however, in the best interest for society as a whole and is an integral part of European policy.[8] Which implies that individuals must share personal information, while those handling the information must consider its sensitive nature.

What makes this case particularly interesting is the fact that the European Union has imposed utilities in all Member States to deploy a environmentally friendly technology that may prove such a threat to the privacy of the end consumer that the full potential may not be realized.

1.4 Limitations

During the course of the literature study it has become evident that intertwining three different fields of research into one easy comprehensible study is challenging. This thesis will therefore address rather intricate fields with an easy understanding, using analogies where fitting and examples rather than descriptions at the technical level.

As far as interpreting the new Regulation, the academic background of the author is a limitation. This research will therefore be supported mainly by papers from law firms and the EU's own publications.

Furthermore, the author recognizes that by assessing a legislation that is not yet legally effective as well as using a case study, which technology is yet to be fully actualized, has its limitations. However, the findings in this thesis will provide insight to the problems to come. The research will therefore serve as proposals for future research rather than presenting hard facts.

1.5 Thesis Outline

The logical structure of this thesis is nontraditional in the sense that it contains three separate fields of research. Each field will be dedicated their own chapters and each of which with their own theory part and analysis. The subsequent chapter and respective analysis will build on the previous. In order to maintain a flow, the thesis has omitted a dedicated part to theory. The literature review of smart metering is provided in the Appendix, whereas the case is initiated with a short summary of the review.

Big Data

Initially this thesis will give an introduction to the world of big data, by describing its key characteristics and some relevant techniques for driving value through analytics. Subsequently a value chain is presented, in which a preliminary analysis is carried out. This analysis arrives at a set of value drivers characteristic to the big data value chain. Conclusively the chapter introduces the risks that has emerged with the trend as well as a theory on the value of trust in this context.

GDPR

The introduction of the GDPR builds on the last section of the previous chapter as the emerging risks are one of the main reasons for the need of a data protection reform. This chapter presents a summary of the current Directive and key changes with the new Regulation. Then a literature supported analysis is carried out to identify influences on the value drivers previously arrived at.

Case study

The case is carried out within the framework of the big data value chain where the findings in the previous chapter are used as a foundation for analysis. The results are presented as challenges to realize the potential of smart metering, which, in a sense, is to answer how the value of smart meter data will be impacted.

Summary of findings

This chapter presents the findings from each analysis, where they are aggregated and presented as a set of assertions which will serve as a foundation for discussion. After the discussion a conclusion is arrived at with a forecast of how the GDPR will impact the value of data.

Chapter 2

Big data

This chapter provides a basic understanding of the term big data and the following aspects are explained: The fundamental characteristics; relevant techniques used to analyse and transform data; a fundamental understanding of how value is realized; and a framework to describe how value is derived throughout its life-cycle. The latter is explained through the value chain of big data. When describing this framework a set of theories are arrived at, from which the value driver are subsequently derived. Concluding this chapter big data will be depicted in a privacy context to provide a frame of reference before the GDPR is explained

2.1 Introduction

Applegate et al.2007 described Information Technology(IT), and the advent of Internet in particular, as a fast moving, still ongoing, global phenomenon that permanently had altered the infrastructure of businesses and industries. Describing how IT created a seismic change in the business environment they stated that IT had become a source of opportunity and uncertainty, advantage and risk and the core enabler, and for some organizations, the only channel through which business is done. Big data is now an emerging field in IT, building on the infrastructure of the Internet, that utilize innovative technology to extract value from extraordinary amounts of information[1] – and not since the advent of the Internet 20 years ago has companies seen higher return on their investments.[13] Companies like Google, Apple, Facebook and Amazon has built their core business around the ability to collect and analyze information to extract business knowledge and insights.[1, 5]Their total revenue has almost equaled the GDP of Denmark with one-tenth of the number of employees. They are also referred to as GAFAnomics. These America based companies has set the stage, where the adoption of big data technology has become an imperative need for organizations to survive and gain competitive advantage. One can even argue that big data has become embedded in the way businesses define and execute strategy as well as defining their unique value proposition.

2.2 Defining Big Data

Plummeting cost of storage, and a tremendous increase in processing power has simultaneously with the rapid emergence of new internet technology, such as the internet of things, led to an exploding speed in which data is generated, processed and consumed. It is creating problems as well as opportunities for individuals, businesses and society as a whole. Consequently, the scientific paradigm named big data has emerged .[14]

The name originated from engineers that had to revamp their analytical tools when the volumes of information was growing so big that it did not fit in the memory of their processing computers. There is no rigorous definition of big data; however, the most prevalent is also the first: Laney [15] defined big data as "high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision

making, insight discovery and process optimization"

Additionally to the classic V's, numerous others has been presented. Whether they should be considered defining characteristics of big data is argued upon, but they carry substance nevertheless. Cartledge counted as many as 19 V's. Four of which, in addition to the original three is depicted in this chapter. *Veracity, validity* and *volatility* is important in operationalizing big data, [17] and is of essence for understanding steps in the value chain of big data. When adding GDPR in the equation, veracity, validity and volatility becomes crucial. Furthermore, the mentioned V's are meaningless unless business *value* can be derived.

2.2.1 Volume

The size and the scale of data collection is increasing at an unmatched rate. Meanwhile, the cost of storage is plummeting. This has created an unprecedented growth in data generation, doubling the the volume every 3 years.[4] By 2020 the amount of useful data is expected to reach 14 zettabytes.[1] It is hard to grasp the magnitude of such amounts of data, but McNamara explained it brilliantly. In 2010 1,2 zettabytes of digital information was generated. This equals to the amount of storage in 75 billion 16 GB Apple iPads which is; enough iPads to fill the entire area of Wembley Stadium 41 times; enough to give every woman, man and child on earth more than 7 iPads; and enough storage to run a full-length episode of the series "24" continuously for 125 million years. As for As of now the number are multiplied by 11,67.

These vast amounts of generated data that needs to be processed in order to provide business value has consequently created the volume problem.[1]

2.2.2 Variety

Data comes in many shapes and formats and come either as structure or unstructured, where numbers and video represents two extremes respectively. Smartphones for example, provide location data, social media information, transactional information, music preferences and browser activity all provide data of different types, formats and for different purposes. This is what characterizes variety.[1]

Data originates from an increasing variety of sources that extends far beyond the scope of the pocket. Much of this is due to the advent of the Internet of Thing (IoT). A forecast by Gartner, Inc predicts that 8,4 billion connected devices will be in use world wide in 2017. The same article says that the 20 billion mark will be crossed in 2020.[19] The potential for IoT is big, so big that the potential number of connected devices theoretically could equal to the amount of atoms on the surface of 100 earths.[20]

The variety of data refers to the range of types and sources.[1] Although, companies won't be interested in processing all data, the variety of data is massive and will only increase. Hence, companies will continue to collect data for processing from an increasingly diverse set of sources. Bringing together these endless streams of diverse data is no small task.[21]

2.2.3 Velocity

Velocity can be explained as the speed of which data is generated, produced, created, or refreshed,[22] and thus a measure of how fast it needs to be be processed.

Industries such as manufacturing and petroleum adopt sensors to monitor their assets and production processes. The more sensors the better situational awareness is achieved, which is driving the adoption of sensors largely. Sensors transmit tiny bits of data at an almost constant rate and as the sensor networks and the IoT grows so will the velocity.

2.2.4 The Additional V's

Veracity

Veracity addresses the trustworthiness of data, and poses some of the big challenges is big data. For insance, in database of customers, individuals oftentimes use a fake email address or fake name, to not be identified. The reasons for submitting wrong or inadequate information can be many. Some don't like target marketing while others has mistrust to the integrity of a company or industry [23] Either way, in order to extract the most value from the data any bias, noise and anomalies must be minimized,[17] and customer trust is, among other things, a prerequisite.

Validity

Imagine, it is January and you are to bet on the next round of Premier League football. The home team has never lost against the opponent in twenty-seven years and currently on a nine game winning streak. Based on all historical data and current form the home team is the obvious favorite and you bet on a home victory. To your surprise, the teams drew and you lost. It turns out that the presumed favorite had three starters traveling with national teams to the Africa Championship, two defenders out with injury and players were generally fatigued due to a rough Christmas schedule.

In the initial stages of analyzing petabyte scale volumes of data it may be quite dirty. It is more important at this stage to reveal the patterns and relationships in the data rather than ensuring its validity. However, after this initial analysis a subset of data may be deemed as important and will thereafter be in need for validation.[17] As for the football example, all historic data was pointing in the direction of home victory, but by omitting the current situation and other factors the presumably accurate prediction was wrong due to lacking validation.

A more critical example would be treating a sick patient just based on observed symptoms. In big data context the complexity is usually higher and information may be noisy. It must therefore be stressed that the derived subsets of data and results from subsequent analysis must be validated and ensured accurate before used in decision making or for other purposes.[17]

Volatility

Traditionally, after data capture, processing and analyzing, data have been stored for later reuse and analysis. However in the age of big data, the volume, variety and velocity has created a need to understand the volatility of data.[17] For instance, continuous streams of data may deem it necessary to reconsider how long data needs to be kept in order to satisfy your need, as these streams may have limited utility for the purpose of the analysis.

One of the challenges with big data is that for some sources the data will always be there, but for other the data will be temporary. It is therefore important to establish the right policies and procedures in defining the requirements for retaining data.[17]

Value

Any of the other V's are basically meaningless unless business value is derived from the data. As Kobielus put it: "Data is only as valuable as the business outcomes it makes possible, though the data itself is usually not the only factor responsible for those outcomes." It is how we use the data, rather than the data it self that allow for recognition of the true value of data.[25] The following describes datafication, which is fundamental in understanding value in terms of big data.

2.2.5 Datafication

To give a perspective of phenomenon of big data Mayer-Schönberger and Cukier told the story about Matthew Fontaine Maury, among the first persons to realize the value of huge corpus of data that smaller amounts lacks. He had experienced issues of omissions and inaccuracies of decades and sometimes centuries old charts and generations old experiences resulting in ships zigzagging the sea and taking courses up to three times longer than necessary.

As newly appointed Superintendent of the Depot of Charts and Instruments and dissatisfied with the current situation, he inventoried barometers, compasses, sextants, and chronographs. Also he would study old logbooks, nautical books and maps as well as seek out knowledge from experienced sea captains. Aggregating all the data he discovered patterns revealing more efficient routes. To improve accuracy he created a standard form for logging every vessel of the U.S navy. Merchants were desperate to get a hold of Maurys' charts. In return he got their logs. Mayer-Schönberger and Cukier refers to this as "an early version of viral social networks", where ships flew a special flag to show their participation in the exchange of information. To fine-tune his chart he sought out random data-points by having captains but throw bottles with information about day, wind, position and currents.

From the gathered data, natural sea-lanes of favorable currents and winds presented themselves. When Maury, the "Pathfinder of the Seas" finally published his work he had plotted 1,2 million data points. The work was essential for laying the first transatlantic telegraph cable and his method was even applied when Neptun was discovered in 1846.

What is so special about this story is how it showcases the value of aggregating, transforming and finding new purpose for data. Maury aggregated knowledge, facts and observations, plotted them and revealed patterns of new efficient routes. He had transformed the data to charts, which provided a new more valuable use. The charts were then distributed to merchants for the transactional fee of new data points. These new data points were then plotted and natural sea-lanes appeared in the data. The data had been transformed anew and took the form of a book. This book went to be used for the purpose of laying the transatlantic telegraph cable and discovering a planet.

The key take away from this story is to understand how information generated for one purpose, by extracting and tabulating, can be transformed into something entirely different that has value for a different purpose.

The notion behind the value of data

The story of Maury shows that data's full value is much bigger than what was realized by initial use. Mayer-Schönberger and Cukier [4] explains the value of data "as an iceberg floating in the ocean. Only a tiny part is visible at first, while much of it is hidden beneath the surface" Innovative companies with the right tools to aid them are able to extract this value – to see new ways in which data can be employed past its initial purpose. After its primary use, data's value still exists, but lies dormant – like a spring – the value is released anew when used for a secondary purpose.[4] The ultimate value obtained from data can therefore be seen as all the possible ways it can be used and the output from each individual use. This is the *notion behind the value of data*, which implies that the ability to discover new uses is fundamental to realize the full potential of data.

Furthermore, when data is collected it's seldom a random action, but for one specific purpose. The immediate value of the data is usually evident to the individual or entity collecting it. The primary uses justify the collection and subsequent processing of data which releases an initial value.[4] This initial value potential for a single data entity has a short half-life and will diminish with time, however when aggregated it will, as in the story of Maury, increase with time and transformation.

Take an online clothing retailer for instance. Looking at ten years old data give little indi-

cation to what shoes the customer wants at that moment, but can however, when seen in relation to historical data of the entire customer base be used to forecast market trends.

Because data is not like material things it does not diminish with use – it is not transactionally limited.[3] It can be processed again and again for a potentially unlimited number of new uses. Hence the data will increase in value over time and when mined new uses may be discovered. Walmart, for instance, "mined" their database of old receipts and discovered that pop-tart sales increased seven-fold ahead of a hurricane.[4]

To create value of big data new technologies and techniques needs to be developed for analysing it.[14] The following section presents a set of techniques relevant for this thesis.

2.3 Advanced Analytics

When handling big data the techniques applied need to have extraordinary capabilities to efficiently process the volumes of data within limited run-times. [14] Big data techniques furthermore involve a large number of scientific disciplines that each of which involve their own techniques. This section puts a particular focus on big data analytics and the use of algorithms. Furthermore, an emphasis is put on analytics on volumes of data rather than the speed, although the speed is briefly depicted in 2.4.1. Based on these limitations the following will depict *data mining*, *machine learning* and *artificial neural networks* in particular.

2.3.1 Data Mining

Data mining is a set of techniques used to extract patterns from data, this include techniques such as *clustering analysis*, *classification*, *regression* and *association analysis*. A brief explanation is given in table 2.1. However, big data mining is more challenging than traditional mining and involves methods from machine learning and statistics to extend existing methods to cope with increased workloads.[14]

Table 2.1: Short description of common data mining techniques [26]

Mining technique	Purpose
Cluster analysis	Divide data into groups that are meaningful or useful or both
Classification	Assigning a objects to one or several predefined categories (used in spam filters)
Association analysis	Discovering interesting relationships hidden in the dataset
Regression	A function that predicts a number

2.3.2 Machine Learning

Machine learning is considered a subjection of artificial intelligence (AI). It's purpose is to design algorithms that allow for computers to evolve behaviors based on empirical data. It is considered one of the most useful techniques in data analysis; as it can the automatically find a simple rule to accurately predict certain unknown characteristics of never before seen data.[27] Furthermore, this rule is wished to generalize; that is, it should not only be able to correctly describe the data at hand, but also correctly describe new random data from the same distribution.[28]

Private data analysis

Where "ordinary" machine learning, as described above, aim to learn and predict without depending specifically on one data point, this is also the aim in*private data analysis*; to reveal information about the private dataset without revealing to much about the single individual. Machine learning and private data analysis are therefore closely linked.[28] This is one reason to why machine learning and algorithms is of particular interest in the GDPR, which will be described later.

In big data analytics the algorithms need to be scaled up, where *deep learning* represents the bleeding edge technique.[14] It is a, as the name implies, a deeper form of learning used for developing autonomous, self-teaching system such as Google's language recognition[29] and AI's with inhuman capabilities.[30]

2.3.3 Artificial Neural Networks

Artificial Neural Networks (ANN) are the parent category of deep learning and is characterized by its ability to "learn like a human", as it's processing system is inspired by the structure of a human brain.[31] There are two subsets off ANN: *supervised learning* and *unsupervised learning*. The former is a two stage process; first the neural network is trained to recognize different classes of data by exposing it to a series of examples. Subsequently it is tested to see how well it has learned by supplying it with unseen sets of data. The latter requires no initial information regarding correct classifications, but rather discover the natural clusters that exist within the data, hence they are able to identify their own classifications and reduce dimensionality. Unsupervised pattern recognition is also referred to as cluster analysis, as the mining technique described above. This shows the interplay between different techniques and technologies.[14]

A general rule is that the more hidden layers and nodes in a neural network the higher accuracy they produce. This is the notion behind deep learning. The complexity increases with learning time and can therefore when applied to big data become very time and memory consuming, however with increasing power. There are two main approaches to this problem: up-scaling or reducing size of dataset[14]

2.4 The Big Data Value Chain

A value chain can be used as an analytical tool to understand the value creation of data technology.[1] A typical value chain categorizes the generic activities of an organization that ads value. A generic value chain is made up of a series of subsystems each with inputs, transformation processes, and outputs. The big data value chain identifies the following key high-level activities: Data acquisition, data analysis, data curation, data storage, data usage. This is also represented in figure 2.1

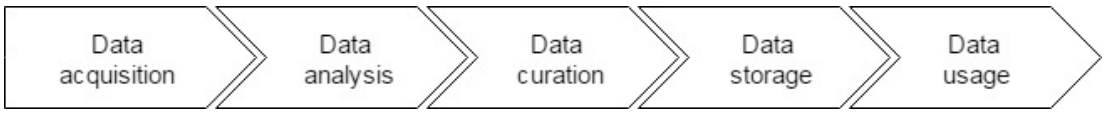

Figure 2.1: The big data value chain (adopted from [1])

The author recognizes the immense technology stack that lay the foundation enabling value through the value chain. However, for the sake of this thesis the value chain needs to be understood at a conceptual level rather than technical. Hence, the following is presented accordingly.

Central to understanding the the big data value chain becomes what is previously depicted in this chapter. The knowledge obtained to this point are used accompanied with supplementing literature to arrive at theories when describing each step of the value chain. The following will serve as a preliminary analysis for later analyses, where a set of value drivers are identified.

2.4.1 Data Acquisition

Previously datafication has been explained, which must be seen in relation to the acquisition of data: A potentially infinite number of data points put in system will eventually reveal a pattern or correlation that generates knowledge. Data acquisition is about collecting and processing so it's interpretable so it can be used in decision-making or stored for analytics. The two theories are:

- The more data collected, the more value can be extracted from it
- The closer to real-time data is processed, the more value it provides decision making and initial purpose.

In big data analytics processing power and storage is no longer economical conundrums which has led to the idea that if feasible, collect everything. [4] This is the notion behind N=all; if you analyse the whole population you will discover what samples fail find. *New knowledge* is obtained and new areas in which data can be used are subsequently revealed.[27] In an unlimited dataset there is potentially unlimited areas in which it can be employed. This is the rationale behind the first theory. The theory can be criticized as more data also means that more irrelevant is captured. However if the data collected is homogeneous the critique don't apply, and the theory will hold.

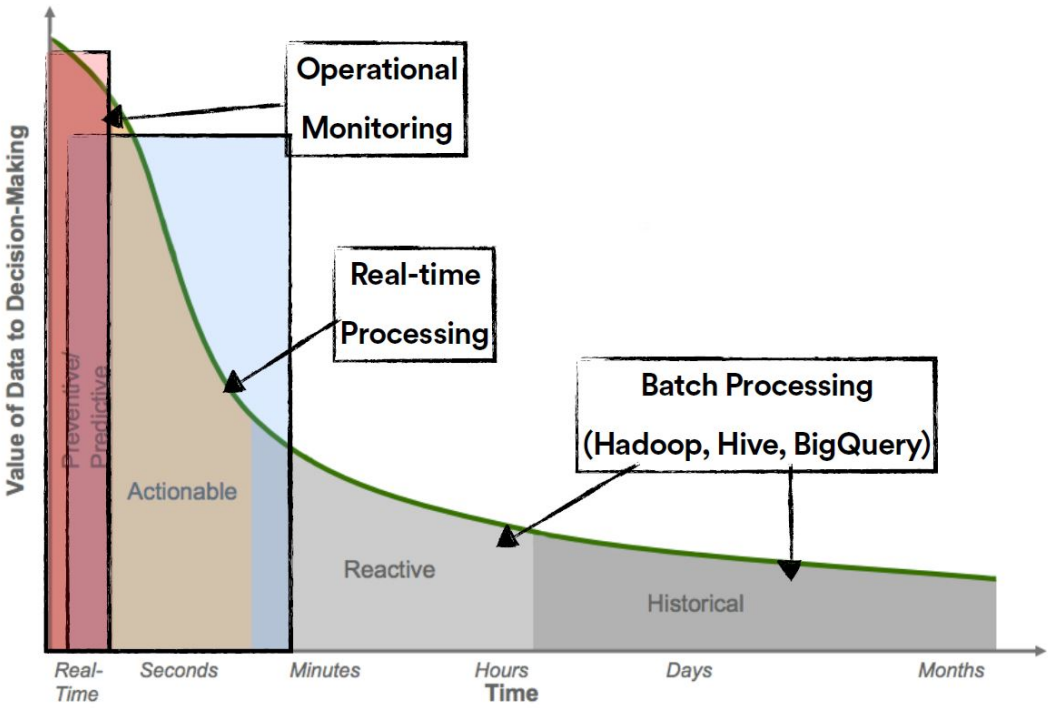

Figure 2.2: Value from speed of processing [32]

The second theory is based on the half-life value of data, illustrated in figure 2.2, where the fundamental idea is that data, like facts, has a half-life. The value of a single data item diminishes with age, which implies the converse: that younger, or closer to real-time data is more valuable.[33] However data can be stored, analysed and used for other purposes. Such as the history with Maury; information about wind and the current at one particular moment will have no value the day after, but when aggregated over time, trends in wind patterns may prove valuable when planning future voyages.

2.4.2 Data Analysis

Analytics can be understood as the processes of transforming data with the goal of using the information in new ways so that the implicit latent value can be unlocked.[4, 34] There are different perceptions to what determines value through analytics.[33, 35, 36] There is, however, one general idea that recur: the higher level of aptitude that is achieved by the analysis the more valuable the output. This is henceforth referred to as level of transformation and is the determinant for ultimate output value. This thesis adopts the terms descriptive, predictive and prescriptive analytics to represent higher levels of transformation respectively.

This is the rationale for the following theory: A higher level of transformation yield a higher ultimate value output

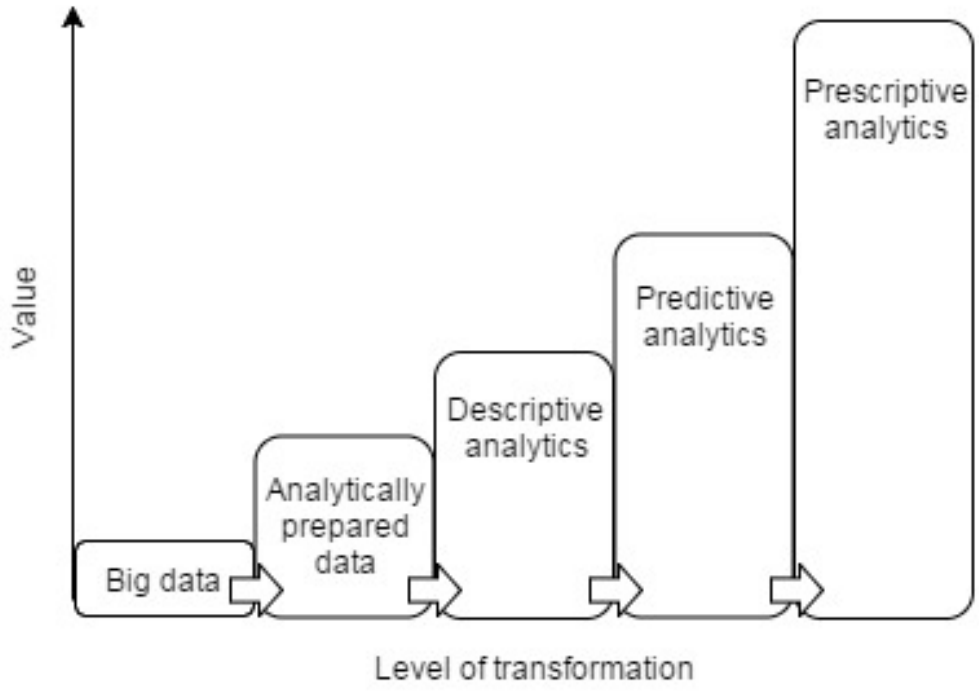

Figure 2.3: The value chain of transformations illustrating the increased value at each transformation (adopted from [33]

- *Descriptive analytics* provide answers to questions like 'what happened' and 'what is happening' by using dashboards, scorecards and reports and is moreover insight to the past.

- *Predictive analytics* discover explanatory and predictive patterns explaining "what will happen" and "why it will happen". Such tools provide businesses the ability to forecast future happenings and the reason they will happen.

- *Prescriptive analytics* provide advice on possible outcomes by determining a set of high-value alternative courses of action. Prescriptive analytics implemented correctly provide the decision maker the best possible information to achieve the best possible outcome. Artificial intelligence can be considered, at the present, as the ultimate prescriptive analytics.[37]

For future reference: when a use of data is the result of prescriptive analytics it is assumed to realize the most possible value from that particular use.

2.4.3 Data Curation

Data curation has been described as the million-dollar word in when talking about big data and has been described as the art of maintaining the value of data.[34] Where it's most important task is to ensure reusability.[23]

In a sense, a dataset can be seen as the soil of a flower: if the soil is not watered and replenished the flower will eventually wither. The same applies to a dataset. The dataset must be accurate and updated in order to avoid a "garbage in garbage out" scenario, meaning tha, computers are only as good as the input: If a machine learning algorithm is trained on biased data the output becomes biased. In fact, in the big data era the acronym has been expanded to "garbage in gospel out", meaning that the general perception of machines as smarter than humans has lead to the understanding that merely using the advanced techniques will lead to insight and improved outcomes.[38] This underpins the importance of data curation in big data analytics.

By maintaining the value, data mining becomes more efficient, and drives more value over time, while the output, being the flower of the analogy, from machine learning algorithms will become more robust over time.[39] From this a theory is arrived at: Curation make algorithms more robust.

In the acquisition stage a critique was addressed to the theory that the more data the more value can be extracted from it. It is obvious that more data also means more to handle in terms of noise, bad data and finding the right purpose for different data. This has created a shift that furthermore has created additional attention to curation. Two subsequent trends has become increasingly prominent in this regard:[34] Data is increasingly considered as corporate assets and consequently a part of companies balance sheets and enterprise valuations. And secondly, valuating data in determining what has present and potential value, or no value at all is becoming increasingly important and has forced companies to reconsider their data management strategies. Based on this one additional theory is arrived at: Curation is like maintaining corporate assets.

2.4.4 Data Storage

The previous steps explained how some data will increase with time while some will decrease. However, as storage cost have plummeted many businesses have strong motivation to keep data, regardless, as new uses may eventually reveal themselves. Additionally, everytime data and analytics are used the value appreciate as they become more complete and accurate.[3] The fact that some data don't depreciate at the same rate has some companies believe that they need to keep data as long as possible.[4] This is the rationale for the following theory: The longer data is stored the more value can be extracted from it. An illustration is provided in figure 2.4 and shows the relation with the theory that the closer to real-time data is processed, the more value it provides decision making and initial purpose.

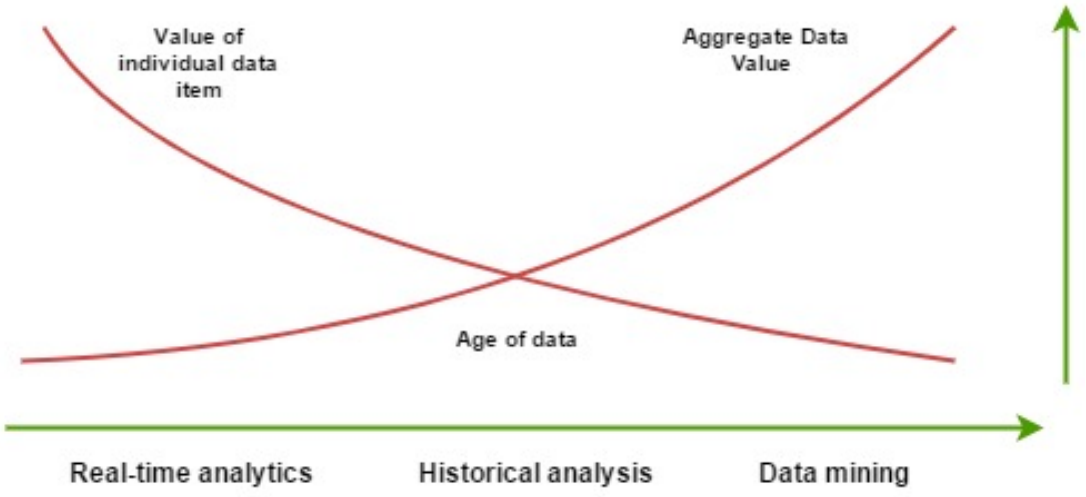

Figure 2.4: Value over time (adopted from[40])

The curation stage described how curation is like maintaining assets, whereas storage on the other hand is like securing your assets. Because the insights buried in the data has transactional value[3] it can even be argued that a safe storage is like secure banking. Which is the rationale behind a second theory: Secure storage is like secure banking

Safe and reliable storage has become of paramount importance and has become a key value driver in big data. Additionally to the need for fast, scalable and cost efficient storage the security aspect has grown in importance of late. [1]

Recent history has shown that the ramifications of a data breach, where personal information has went astray, has had severe consequences for companies and individuals involved.[41]

Hacked companies looses trust leading to churn and customers becoming reluctant to share personal details,[42] which in turn, results in less volume or errenous data. The value of the data decreses nevertheless. However, up to 93% of breaches can be avoided by implementing simple measures.[43]

Hackers methods are getting more sophisticated and even non-experts can wage in an attack where the aftermath of cyber-attacks has left companies bankrupt and employments terminated.[3] At the time of writing this the world has just been hit by one of the largest ransomware attack to date. The New York Times [44] writes that 200 000 computers across 250 countries has been left crippled pending a fee for giving back users access to their data. Some of the targeted were University students working on their thesis. The attackers are estimated to pocket around 1 billion US dollars from the attack.

In light of the presented facts there is little doubt that data has economic value and must be protected accordingly. How much value is created in terms of enhancing the value or potential number of uses or value potential can be discussed. It provides an intangible value nevertheless, as the data will be available for analysis and reuse. Research has been made to whether the information security of a company should be accounted for in enterprise valuations.[45]

2.4.5 Data Usage

So far a general understanding behind what is perceived as the ultimate value of data is provided through the value of reusing data and the value of transforming the data. This subsection will on the other hand provide firstly, a an understanding to **ways** in which data can be reused and secondly, an understanding to **how** data analytics can be used. Mayer-Schönberger and Cukier [4] presents three main ways to release the potential of data value: basic reuse; recombination of data; and designing extensibility into the outset.

Basic reuse

Basic reuse is historically achieved by innovative minds with a vision to identify new purposes, which has left those without forlorn. Often times it is those who are able to identify valuable data "exhaust", the digital trail of a consumer, who thrive.[4] A simple example

is mobile phone operators collecting information on their subscribers' location to route calls. This is a rather narrow technical approach that has a limited value. However, if this information is passed on to, for instance, companies distributing location based advertising and promotions a whole new value is realized. Another example are Google using misspellings in their search query to improve their auto-correction and word suggestion.

Recombination of data

As the notion behind the value of data implies: the sum is more valuable than its parts and when multiple datasets are summed together, that sum is more valuable than the individual dataset. This is the idea behind the methodology called "recombinant data". [4] Sometimes, even, the dormant value of a dataset can only be unleashed by combining it with another. For instance, the combination of two datasets will reveal potential correlations a single dataset would not.

Extensibility by design

Extensibility can be designed into the outset of data collection, or in other words encourage multiple uses from the same dataset.[4] This thesis addresses the strategy as *extensibility by design*. This can furthermore be seen as a particular enabler for transformation. For instance, a surveillance cameras initial purpose is to spot shoplifters, but can additionally be placed so it can track the flow of customers. This extensibility allow for retailers to enhance the layout of the store and judge the effectiveness of marketing campaigns.[4] At an even higher level of transformation, the retailer would eventually predict the flow and, for instance, be advised for suitable sales campaigns.

The former three paragraphs explained ways in which data can be used. Equally important is to understand how to use the data to realize the value. This can be divided into three main categories: applied analytics; operationalized analytics; and monetized analytics. These are in the following presented with examples from the case study.

Applied analytics

Applied analytics is an adapted term to serve the purpose of this thesis; in which, it encompasses both use of traditional analytics and advanced analytics, to drive value through increased performance. Such increased performance can be seen as, among other things, innovation, process optimization and education. In the case study an example is provided where predictions about residential energy consumption is used to optimize pricing models on electricity.

Operationalized analytics

By operationalizing analytics they are made a part of business processes and can be used to drive top-and bottom line revenue.[17] The case study shows that utility companies can use predictive analytics to detect tampering and energy theft in the power grid. Another example is for a call center who uses predictions to identify good targets for upselling and which products they may be interested in. These examples shows how companies can save and make money.

Monetized analytics

Big data analytics can be used to synthesize insights and knowledge that other companies are willing to pay for. This way analytics can be used to drive revenue beyond the insights it provides just for own benefit.[17] The case study provides several examples on this, some more controversial than others. At the controversial side of scale customer profiles that classify behavior can be sold to insurance companies for determining premiums. On the other end of the scale market research made on these profiles can be sold to markers to better target specific demographics.

2.4.6 Big Data Value Drivers

This chapter has depicted the characteristics of big data; provided a fundamental understanding of the phenomenon; presented it in a privacy context; described fundamental techniques to extract value; and at last the big data value chain. The value chain furthermore arrived at a set of theories. The theories are summarized in table 2.2

Table 2.2: Theories about value drivers in big data

Stage	**Theory**
Data acquisition	The more data collected, the more value it can be extracted from it The closer to real-time data is processed, the more value it provides decision making and initial purpose
Data analysis	A higher level of transformation achieves a higher potential use value
Data curation	Curation make algorithms more robust Curation is like maintaining corporate assets
Data storage	The longer data is stored the more value can be extracted from it Secure storage is like secure banking
Data usage	The ultimate value of data is the sum of all the ways in which it can be used and all the respective value outputs

Based on the literature reviewed in this chapter and the theories arrived at a set of concrete value drivers are identified and illustrated in figure 2.5. The one that stands out, as it's not coveres in the theories are the influence of "Algorithmic complexity", which underlying rationale bounds in the presented techniques for extracting value from data in section 2.3, where the general notion is that the more complex an algorithm the more intelligent it is.

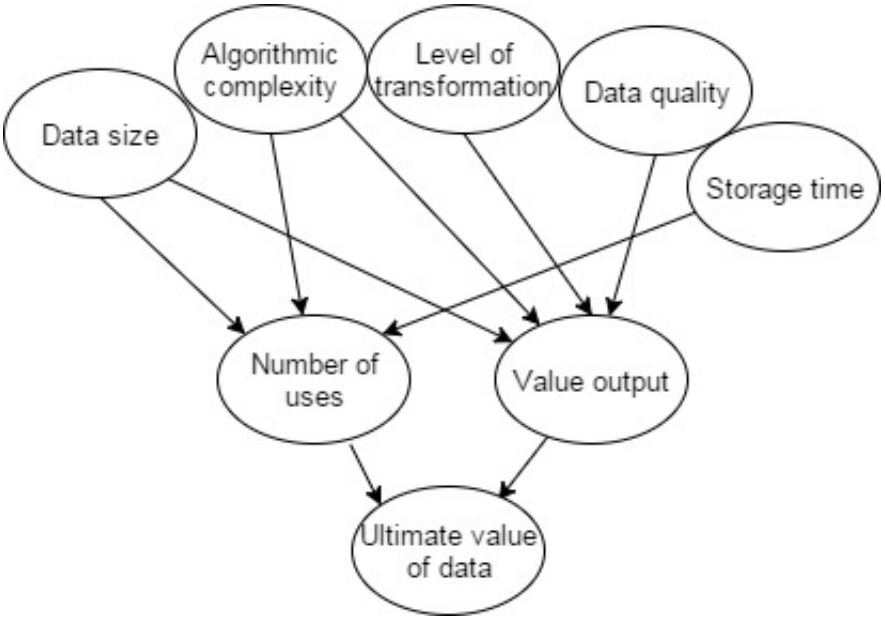

Figure 2.5: Illustration of what drives the value in the big data value chain big data

2.5 Privacy and Big Data

The volume, variety and velocity does not only impose new challenges to processing power, storage and techniques to make sense of the data. The scale of which big data is processed has also taken existing privacy risk to a whole new and unpredictable level.[23] Which also has created the need for a new data protection regulation. This section presents a set of privacy concerns in big data that is relevant for the upcoming chapter.

Lack of control and transparency

As the amount of connected devices increases, so does sources from which personal data is collected. A typical example are health monitoring devices.[23] However, it is not only devices with sensors that produce personal information. Everything individuals do online, from web searches to social media activity to location tracking on mobile devices; the sources becomes more and more unexpected and people seldom know when, why, or how data about them are collected and let alone how it is used.

Data reusability

This chapter has previously described reusability as a characteristic of big data that determine its value, where the scale of storage allow for collection of data to continue indefinite. The data is mined until value is extracted. However, the value extracted may be interesting for parties whose intentions may not be in the best interest of the individual generating the data.[23]

Analysis of data from health monitoring device may provide detailed information about fitness, diseases or risk for diabetes and heart problems. This information may be interesting for an insurance company for determining a health premium or for a marketer to promote dietary supplements.

Re-identificaton

The dormant value of a dataset may only be unleashed when combined with another,[4] but this also triggers a privacy risk. Linking different datasets does not only reveal patterns of value, but can also reveal patterns about individuals, allowing them to be identified or have sensitive information disclosed.[4, 23]

Profiling and automated decision making

Big data analytics can be applied to combined datasets to create profiles about individuals, which can be used in decisions made by automated means. These cases raise a variety of ethical issues where the algorithms making the decisions tend to discriminate based on biased and incomplete data.

2.5.1 The Oxymoron of Big Data and Privacy

The European Agency for Network and Information Security (ENISA)[23] states that "there is no big data without privacy". Based on the risk issues presented above there is little doubt that privacy needs to be a core value in big data, and it needs to be a synergy between the two:"if privacy principles are not respected, big data will fail to meet individuals needs; if privacy enforcement ignores the potential of big data, individuals will not be adequately protected."[23]

ENISA[23] presents a scenario which exacerbates the current situation, to a world of "big data without privacy", in order to emphasize the importance of privacy in big data. In this world, with a massive spread of analytics without data protection, personal data would become commoditized and no longer be the scarce resource that it currently is. With personal information widely dispersed in a digital format, and little to distinguish between each and every personally identifiable data subject, the informational value of data would eventually diminish. That is to say, with no protection and nothing differentiating each other, individuals would become reluctant to provide their data or would give false data in order to obtain the services they want. In this scenario data quality would severely reduce and the value consequently plummet. It is therefore in the best interest of all parties that personal data stays difficult to obtain and a scarce resource, so its value is maintained.

Based on this oxymoron it can be argued that a respect to privacy is essential of the trust between users and service providers in a data-driven economy. And because of a seemingly increasing mistrust to service providers in general a new Regulation could not be more timely.

Chapter 3

The General Data Protection Regulation

This chapter initiates with an introduction to the General Data Protection Regulation (GDPR) and explain the need for a new approach to data protection in the big data era. It also summarizes the relevant provisions from the previous Directive as a foundation for understanding the change to come. Key changes are depicted followed by an analysis of how the Regulation will impact key characteristics of big data. These key characteristics are:

- Algorithmic unfairness and discrimination;
- Opacity of processing
- Tendency to collect all data
- Reuse

The analysis concludes with a set of key findings put in relation to the identified value drivers in the previous chapter.

3.1 About GDPR

Following A Single Digital Market Strategy for Europe, a response to increasing digital transactional activity between Member States, the GDPR is securing consistency around data protection laws and rights crucial to businesses, organisations and individuals.[46] Its purpose is to ensure protection of personal information while exchangeable across borders. The regulation further entails a full harmonization of the privacy Policy in the EU and EEA. Essentially implying that there is no permission to deviate from the the rules or ad supplying ones. However, opening for national rules in special cases[47].

When the regulation becomes legally effective 25 May 2018 it will replace the existing EU Data Protection Directive 95/46/EC (the Directive) and will bring the individual new legal rights, extending the scope of responsibilities for parties handling personal data [48]. The GDPR 2016 states: *" In order to strengthen the enforcement of the rules of this Regulation, penalties including administrative fines should be imposed for any infringement of this Regulation..."*, leading to potential fines up to 4% of total global turnover or 20 million EUR, whichever is higher. The scale of change and the magnitude of legal actions introduced by the new regimen means that companies and organisations across Europe will have to adapt and become consistent and coordinated in their new approach.

3.1.1 Scope of the Regulation

When the regulation comes into effect it will be directly applicable as law to all companies and organisations collecting or processing personal data about EU or EEA citizens, regardless of state, as long as the data collection takes place in the EU. This means that the territorial application of the GDPR covers a much wider scope than the Directive that is being replaced.

To fully grasp the scope of GDPR and the fundamental changes imposed on the business environment the Directive needs to be understood. The following section aims to describe the essentials of the Directive as basis for further elaboration.

3.2 The EU Data Protection Directive

The Directive, like GDPR aims to protect personal data, establishing a regulatory framework which seeks to strike a balance between a high level of individuals privacy protection and the free flow of personal data within the European Union. It is designed to protect the privacy and to protect all personal data collected for and about EU citizens.[50] Mainly related to the processing, utilization, and exchange of personal data.

3.2.1 Background

The history of the basic elements and principles of the Directive go more than forty years back, having remained the same since the introduction of the first data protection act, in the German federal state of Hesse in 1970. Every Member State's data protection acts has subsequently incorporated these elements and principles.[51] Also the Directive encompasses all key elements of Article 8 of the European Convention for the Protection of Human Rights and Fundamental Freedoms,[50] stating its intention to respect the right to privacy whatever the nationality, residence or personal correspondence. Furthermore, the organisation for Economic Cooperation (OECD) introduced in 1980 its "Recommendations of the Council Concerning guidelines Governing the Protection of Privacy and Trans-Border Flows of Personal Date"(OECD Recommendations) introducing seven principles governing the OECD's recommendations for protecting personal data. However, not binding law the principles were;

1. *Notice*: Data subjects should be given notice when their data is being collected
2. *Purpose*: Data should only be used for the purpose stated and not for any other purposes;
3. *Consent*: Data should not be disclosed without the data subject's consent;
4. *Security*: Collected data should be kept secure from any potential abuses;
5. *Disclosure*: Data subjects should be informed as to who is collecting their data;
6. *Access*: Data subjects should be allowed to access their data and make corrections to any inaccurate data; and

7. *Accountability*: Data subjects should have a method available to them to hold data collectors accountable for not following the above principles

At the time data privacy laws varied widely across Europe and subsequently these principle were incorporated into the EU Directive. However, a diverging data protection legislation across EU member states impeded the free flow of data. To amend this growing problem the EU proposed the Directive.

Although the technology in which the legal system was designed on has become obsolete, the system has until now fared relatively well. However, the pace of development in recent years has created a new need. Technologies of processing personal data previously expected finite, traceable and identifiable number computing equipment and processing operations, whereas in this day and age, multitasking, cloud computing and outsourcing is increasingly making it difficult to distinguish between processing actors.[51]

The following section summarizes the Directive. The content is obtained from the Directive it selfand the European Union's own summary of the Directive

3.2.2 Summary of the Directive

There are many similarities between GDPR and the Directive, and large parts of the new regulative builds on principles of the directive. For the purpose of this thesis a description of the previous directive is important to understand the changes imposed by GDPR.

The Directive applies to data processed by automated means, such as digital databases on customer data, and data contained or intended to be a part of non automated filing systems, which in todays day and age can be characterized as non-digital, typically traditional paper files. When processing activity concerns public security, defense, State security Commmunity law, such as operations of concern for public, defense, State security or criminal law the Directive states that it falls outside the scope.

In order to protect the rights and freedom of persons with respect to processing of personal data, the Directive lays down a set of criteria for lawful processing and the principles of data quality. Moreover also giving the data subject certain rights when their data is being processed.

Lawful processing

To ensure lawfulness processing of data the Directive states that data processing is only lawful if:

- the data subject has unambiguously given his consent; or
- processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or
- processing is necessary for compliance with a legal obligation to which the controller is subject; or
- processing is necessary in order to protect the vital interests of the data subject; or
- processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or
- processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed , except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under.

Data quality

Furthermore, stating that for all lawful data processing activities principles must be implemented to ensure data quality, the Directive wants to ensure that certain types of personal data processing is identified and separated from other intrusions into private life [51]. Article 6 in the Directive emphasize that " Member States shall provide that personal data must be:[50]

- processed fairly and lawfully;
- collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical,

statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

- adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;
- accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;
- kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use".

The Directive also lays down a provision for special categories of processing data, prohibiting the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life. However, establishing certain qualifications concerning cases where the protection of vital interests of the data subject or the public or for the purposes of preventive medicine and medical diagnosis.

Right to obtain information

The controller is required to provide the data subject from whom the data are related to with at least the following information:

- the identity of the controller and his representative, if any
- the purpose of the processing for which the data are intended
- further information about recipients of data, questioning as well as the right of access and the right to rectify the data concerning him or her.

These rights are also applicable in cases where the data has not been obtained from the data subject. In this case the above mentioned information shall be given to the data sub-

ject no later than the time of disclosure to a third party. However, when processing is done for statistical purposes or scientific or historical research the provision of above mentioned information is not required. Also if provisioning proves impossible or involves a disproportionate effort. Legal effects to provide safeguards may also overrule this principle.[52]

Right of access

The data subject has the right to obtain from the controller without constraints at reasonable intervals:[50]

- confirmation about whether or not data related to the data subject are being processed and the information of the processing of data as described in Right to obtain information.
- communication in an intelligible form of the data undergoing processing and of available information about their source.
- knowledge of the logic involved in any automatic processing of data at least when the subjected to automated decisions.

If the data processing does not comply with the provisions of the Directive, it is seemed as appropriate for rectification, erasure or blocking. This should also be notified to any third parties to whom the data has been disclosed.

Right to object

The data subject is entitled to a right to object, on request and free of charge, to the processing if the controller anticipates personal data to be processed for the purpose of direct marketing. The data subject should also be informed before personal information is disclosed to a third party intending to use the data for direct marketing and should be explicitly offered the right to object to such disclosures.[50]

Right not to be subjected to automated individual decisions

Article 15 Automated individual decisions is on of the most discussed upon articles of the Directive.[53] It states that every person has the right not to be subjected to decisions that produces legal effects or significantly affects them which is based solely on automated processing of data. This is exemplified with evaluation of performance at work, creditworthiness, reliability and conduct. [50]

Exceptions and restrictions

The scope of the principles relating to the data quality, right to obtain information, right of access and the publicizing of processing operations may be restricted in when it constitutes a necessary measure to safeguard:

- national security;
- defence;
- public security;
- the prosecution of criminal offenses or breaches of ethics in for regulated professions;
- financial and economic interests of importance for a Member State or the EU;
- protection of the data subject or the rights and freedom of others.

Confidentiality and security of processing

The principles of confidentiality and security largely builds on the principles of information security; confidentiality, integrity, availability and non repudiation, when related to the processing for personal data.[52]

The Directive sets the stage for Member States to establish a legal framework for secure and confidential processing of personal data, requiring the controller to implement appropriate technical and organisational measures to protect personal data from accidental or unlawful

destruction, alteration, unauthorized disclosure or access. Furthermore, on the security of processing, the Directive says:

> Having regard to the state of the art and the cost of their implementation, such measures shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected."

In the context of big data and the threat landscape of the digital age, one can argue that this section of the Directive is to ambiguous to enhance any form of harmonization and standardization of neither implementation of measures or enforcement of the Directive.[51]

Notification

The Directive establishes a supervisory authority to whom the controller or any representative need to notify before carrying out any sort of automatic processing operation. For the processing of personal data presenting risk to the rights and freedom of the data subject it is required to conduct a prior checking. It is however up to the Member States to decide what nature of processing activities that carries such risks.[50]

In the notification the controller is required to provide information about:

- name and address of the controller and any representative;
- purpose of the processing;
- type of data to be used and the data subject it is related to;
- recipients of the data and to whom it might be disclosed;
- proposed transfers to third countries

The supervisory authority must keep a register of the processing operations notified. It is also required that measures should be taken to ensure the publication of processing operations to the public.

Judical remedies, liabilities and sanctions

The provisions for juridical remedies and sanctions lays down certain requirements but, leaves it to the Member States to adopt suitable measures to ensure a full implementation. The Directives requires Member States to provide juridic remedies to individuals suffering damage due to unlawful processing of personal data. In case of a breach of their rights, the data subjects are entitled to receive compensation for the damage suffered.[52]

Transfer of personal data to third countries

For transfer of personal data between countries, the Directive is distinguishing between transfer between Member States and third countries. Between Member States free flow is authorized but, when transferring to any third country certain principles, securing adequate level of protection, is required. Transfer from a Member State to a third country with adequate level of protection are authorized. However, transfers may not take place of adequate level of protection is not guaranteed. The Directive lists up a number of exceptions such as explicit consent from the data subject, the conclusion of a contract, necessity for public interest. [52]

To secure a unified practice of the transfer rules The Directive establishes specific procedures. Mainly lying down principles for cooperation between the Commission and Member States when handling third countries with inadequate levels of protection.

3.3 Key Changes

At glance two easily relatable, and undoubtedly intimidating aspects of the GDPR stick out and embodies the change in which the reform imposes the global business environment; the signal effect of tougher sanctioning and a "you can't escape" attitude. There is no doubt that the GDPR wants every organisation to take privacy seriously, and it will hurt if they don't.

Complying with the GDPR becomes a must. This requires organisations to completely transform the ways they handle data throughout its lifecycle; data collection must be

mapped; current processing and analysis must be assessed; contracts reviewed and renegotiated; data security enhanced and additional measures to ensure compliance must be implemented. This section will go more in depth to the fundamental changes the GDPR imposes on the current legislation and how this will impact organisations that process personal data.

3.3.1 Territorial Scope

Unlike the Directive, GDPR will take direct legal effect in all Member States and there will be no need for transposition into national law. It takes precedence over any conflicting legislation that may exist in the national law of a Member State, this also includes sector-related regulations. By allowing Member States to adopt supplementary laws in defined areas, the GDPR is enabling further regulation to its principles of protection.[49]

The provisions of the GDPR has also has intentions of capturing more overseas organisations and in particular U.S tech companies. Even though the organisation can prove that they are not established in the EU, they will still be caught if the processed personal data originated in, or relates to a good or service provided to a location within the EU.[54]

3.3.2 "All data becomes personal"

The GDPR changes the definition of personal data, now defined as "any information relating to an identified or identifiable natural person". The reform also expanded the definition to include location data and online identifiers.

Recital 26 lowers the bar to what is is regarded as "identifiable". Thus also increasing the scope of what is regarded as personal data. It can be argued that this change is a direct reaction to privacy risks of big data. The recital states that someone is identifiable if anyone can identify a natural person using all means reasonably likely to be used. In other words, even though an organisation is not able identify a natural person, the data may still be personal data.

Recital 30 expands upon the additions, that is, location data and online identifiers in the definition of personal data. It exemplifies online identifiers with IP adresses, location data,

RFID tags and cookies. Hence, the usage of devices, applications and digital tools that may leave traces, which in particular when combined with unique identifiers and other information, may be used to create profiles of natural persons and identify them.

Considering the increasingly sophisticated technologies in big data, and the assessment of the above mentioned recitals, it could be argued that all data generated for or by an individual is personal data. Hence, it could be justifiable to work on the he assumption that all data is personal data given the extremely wide definition of personal data in GDPR.[54]

3.3.3 Data Protection Principles

The core themes of the data protection principles presented by the Directive remain largely the same under the GDPR. Instead of presenting principles relating to *Data quality* the GDPR now presents a set of more relatable principles for processing of personal data. Additionally a new principle has been added

- *The lawfulness, fairness and transparency principle:* Personal data must be processed lawfully, fairly and in a transparent manner.
- *Purpose limitation principle:* Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.
- *Data minimization principle:* Personal data must be adequate, relevant and limited to what is necessary in relation to the purpose(s).
- *Accurracy principle:* Personal data must be accurate and where necessary kept up to date.
- *Storage limitation principle:* Personal data must be kept in a form which permits identification of data subjects for no longer than is necessary for the purpose(s) for which the data are processed.
- *Integrity and confidentiality principle:* Personal data must be processed in a manner that ensures appropriate security of the personal data, using appropriate technical and organisational measures.

- *Accountability principle:* The controller shall be responsible for, and be able to demonstrate compliance with the principles above.

3.3.4 Lawful Processing

The GDPR is significantly rising the bar for what is considered lawful processing. Among the principles mentioned above, and in particular the lawfulness, fairness and transparency principle, requires processing to fall within the legal justification for processing. The Directive provides a structure to this, but the GDPR makes it much harder for organisations to comply.[54]

Consent

The conditions for what is considered valid consent has become much stricter under the GDPR. The Regulation states that a consent to processing of personal data must be freely given, specific, informed and unambiguous.[54] The initial impact of this is that companies will no longer be able to use long illegible terms and conditions written by lawyer for lawyers. Additionally a consent must be as easy to withdraw as it is to give. The new right of the data subject, mainly the right to be forgotten and the right to portability carries extra baggage in this regard if consent is given.

Legitimate interest

The legal justification in which companies can claim legitimate interest for processing is becoming more narrow. The justification existed in the Directive but the enforcement varied significantly among the member states, thus imposing a stricter regime for interpreting what processing is necessary for purposes of legitimate interest.

New purpose processing and reuse

The controller will often want to process data for new purposes to realize more value[4] This potentially conflicts with the purpose limitation principle and the rights of the data

subject.

The GDPR proposes two main solutions to allow new purpose processing. Firstly Article 6(4) sets out a set of considerations to ascertain whether or not the new process is compatible with the purposes for which the personal data were collected in the first place. These include:

1. link between the initial purpose and the intended purpose of further processing
2. the context of the data collection
3. possible consequences to the data subject resulting of further processing for the data subject
4. existence of appropriate safeguards such as encryption and pseudonymisation

Secondly, if the controller concludes that the new purpose is incompatible with the original purpose, then a fresh consent is the most expedient option. Taking into consideration the complex nature of big data, the additional juridic burdens and additional requirement imposed by the GDPR, consent will become a last resort justification for new purpose processing.[54]

3.3.5 Data Subjects Rights

The individuals rights in the Directive is enhanced in the GDPR and the new right of *data portability* is added. The rights are furthermore backed up by provisions that make it easier to claim compensations and generally to enforce the rights of the individual. In the following the different rights and their changes is presented by the article representing them.

Art. 12(1),13,14: Transparency

The enhanced rights for individuals builds largely on greater transparency. Article 12(1) states that information must be provided by controllers to the data subject in a concise, transparent and easily accessible form, using a clear and plain language. Article 13 provides a list of information that must be provided at the time of data acquisition from the data

subject. While Article 14 provides different requirements that apply when information is not acquired from the data subject.[54]

Art. 15 Right to access

The existing regime of the Directive largely holds, but some additional information must be disclosed.[54] One dramatic shift and a big empowerment of the data subject is the right to obtain from the controller confirmation as to whether or not personal data is being processed.

The controller will on such request be required to provide a copy of the personal data in an electric format, free of charge.[55] This means that the barrier for individuals to get access to their personal information almost completely removed and may in the future demand excess resources for organisations to provide information on request.

Art. 17: Right to be forgotten

The right to be forgotten, also known as right to erasure, provides the data subject the right to have his or her personal data erased, disclosure of personal data ceased and processing by third parties halted. These conditions include breach of the purpose limitation principle and lawful processing as well as if consent is withdrawn.

The right is however not absolute as it arises only in a quite narrow set of circumstances where the there is no legal ground for the controller to process the information.[54] In a practical sense this may create situations where third parties may have legal grounds for processing although the party providing them with the information has not.

Also it is worth mentioning that the practical impact of such a decision to erase personal information may not be of public interest.[54] Hence the right requires controllers to compare the subjects' right to the public interest in the availability of the data.

Art. 20: Right to data portability

The current directive has no equivalent to this article and this article brings about an entirely new right. The data subject has the right to transmit from one controller to another their personal data. [54] The purpose of this is to empower the data subject. It will also foster competition between controllers in the EU by supporting the free flow of personal data. Controllers will face new challenges in order to ensure heightened user control The right to data portability will require businesses in a wide range of areas to ensure that they can hand over personal data in a usable and transferable format.

In this process, businesses face new challenges in order to provide better control to users. *Article 12* requires controllers to provide modalities to facilitate the excercise of the data subjects rights. In this context requiring implementation of systems responsive to user requests concerning their data such as interfaces and customer support services.

Article 20 is one of the more controversial topics when discussing the implementation of the Regulation. Particularly applicable to the problem of this thesis is the discussion to whether the right may discourage companies and service provider from creating proprietary information.[56]

Art. 21: Right to object

The right to object to processing of personal data for direct marketing purposes remains as provided in the Directive. A new addition in the Regulation is the right to object to to processing which is legitimized on the grounds of the data controllers legitimate interest or in interest of the public.

If an objection were to be submitted the controller will have to suspend any further processing until the can demonstrate "compelling legitimate grounds" for processing.

Art. 22: The right not to be subject to automated decision taking, including profiling

Article 15 in the Directive came with ambiguity that rendered it inadequate with technological development especially in the department of intelligent systems such as artificial

intelligence.[53] The GDPR expands upon this right and refers explicitly to profiling as an example of automated decision making. Automated decision making and profiling is only permitted where necessary for entering of performing a contract; authorized by EU or Members State Law, or; the data subject has given their explicit consent, such as an opt-in decision.[49]

Still of concern is the scope and ambiguity of this right. Especially when considering legitimate profiling to detect cybercrime and fraud. Also the online advertising industry and website operators are expected to face new challenges with the GDPR requiring them to revisit their mechanics for customer consent. In particular justifying online profiling for behavioral based advertising. [54]

3.3.6 Accountability and Governance

The above mentioned key changes has two reoccurring themes. Firstly companies will be held more accountable for their actions by tougher sanctions and more empowered data subjects. Secondly, data governance does no longer simply imply that companies will have to do the right thing, but will have to prove it even years after a decision has been made. Data subjects other stakeholders and even the media are potential recipients of such information.

Article 5(2) requires the controller to demonstrate compliance with the data protection principles. Some of the enhanced governance obligations that is manifested through these principles are explained in the remaining of this section.

Article 30: Records of processing activities

The requirement imposed on controllers to notify the supervisory authority about any processing operations is abolished. A more general obligation is instead introduced that requires the controller to keep extensive internal records about their data protection activities. organisations employing less than 250 people is exempted from this unless their practice includes high risk processing.

Article 35: Data Protection Impact Assessment

Fundamentally new with the new Regulation is the concept of Data Protection Impact Assessment (DPIA) through Article 35. This is a process designed to describe the processing, assess the necessities and proportionality of a processing and aid in managing the risk to the rights and freedoms of individuals as results of processing.[57]

DPIA will become an essential part of companies ability to showcase that the data protection principles are considered in a practical manner and taken seriously. It also allows companies to measure the level of privacy[58]

The Regulation[49] states that before any processing that, due to its nature, scope, context, purpose and used technologies, is likely to result in high risk to the rights and freedoms of individuals a DPIA must be carried out. The examples of what is *"likely to result in high risk"* given involves a great deal of ambiguity, hence the Article 29 Working Party (WP29)[57] provides a more detailed list, which is presented below:

1. Evaluation or scoring, including profiling and predicting
2. Automated-decision making with legal or similar significant effect
3. Systematic monitoring
4. Sensitive data
5. Data processed on a large scale
6. Datasets that have been matched or combined
7. Data concerning vulnerable data subjects
8. Innovative use or applying technological or organisational solutions
9. Data transfer across borders outside the European Union
10. When the processing in itself *"prevents data subjects from exercising aight or using a service or a contract"*

Sound procedures for carrying out a DPIA is a key part of complying with the GDPR. Successfully doing so should result in greater trust and confidence of data subjects and other

data controllers.[57]

Article 37, 38 & 39: Data protection officers

Another entirely new requirement and a new significant governance burden and cost is the appointment of data protection officers (DPO).The organisations obligated are public authorities and those whose core activities consist of on a large scale *"regular and systematic monitoring of data subjects on a large scale"* or processing of *"special categories of personal data"* Article 39 entrusts the DPO with a minimum of tasks, which means that nothing prevents the DPO to get more responsibilities.[59] The most prominent of these tasks are:

- to inform and advice on compliance with the GDPR and other data protection laws
- to monitor compliance with law and with the internal policies of the organization including assigning responsibilities, awareness raising and training staff
- advice in and monitor DPIA
- cooperate with supervisory authority and act as contact point

The WP29[59] makes certain statements regarding appointment of a DPO worth taking into consideration. Firstly, even when it is not mandatory by means of law to appoint a DPO, organisations will still benefit from designating a DPO on a voluntary basis. Secondly a DPO is a cornerstone of accountability and an appointment can facilitate compliance and become a competitive advantage.

Article 25: Privacy by design and by default

In short, as opposed to treating privacy risks as a reactive action privacy by design requires a proactive approach. Which means taking data protection risks into account throughout the whole life cycle of the system or process development. This entails implementing appropriate technical and orgnisational measures from the outset. Privacy by design is a wide area that has existed for years, but just now enshrined in data protection law. One of the main aspects is its compliance "toolbox" which includes a set of anonymisation techniques such as *pseudonymisation*.

Pseudonymisation is processing personal data in a way that the personal data no longer can be attributed to a specific data subject. This is considered anonymous data and according to WP29[60] suffice for new purpose processing. However, in big data, when combined with another dataset a pseudonymous dataset may be *re-identified*.

Data protection by default means that the strictest possible privacy setting is automatically the default setting. This entails that mechanisms are in place to ensure that the data protection principles are complied with.

3.3.7 Breach Notification and Security of Processing

At the current state of European legislation there exists no universally applicable law requiring notification of data breaches. This is set to change when the GDPR become legally effective through the introduction of the breach notification.

GDPR requires through article 33 the notification of breach to the supervisory authority no later than 72 hours after becoming aware of it. Article 34 on the other hand requires the notification to be given the affected individuals, given that it is a high risk of the rights and freedoms of the individuals.

The obligation to to notify is conditional on awareness. However, Article 32 which undertakes security of processing requires controllers to implement appropriate technical and organisational measures to ensure a level of security that is appropriate to the level of the risk.[49] Thus giving less leeway for companies not be attentive. Such measures include;

- pseudonymisation and encryption of personal data;
- confidentiality, integrity, availability and resilience of processing systems and services;
- ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;
- and processes for testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing

Processors as well as controllers failing to comply with articles relating to security of pro-

cessing and data breach notification faces fines up to 10 million or 2% of annual turnover. However, as data breaches often leads to investigations other areas of non-compliance may be uncovered. Hence there is quite possible the maximum penalty of 20 million Euros or 4% of global annual turnover will be triggered.[54]

3.3.8 Data Transfers

The GDPR will not make any significant material changes to the current rules of the Directive for cross-border transfers of personal data but changes significantly the severity of penalties inflicted with failure to comply. The current regime is limited in its sanctioning for breaching transfer restrictions. This is a big contrast to the GDPR which will inflict the highest category of fines.

3.4 Big Data under the GDPR

The GDPR sets the stage for a new digital market in the EU by building a single, strong and comprehensive set of data protection rules. It aims to boost innovation in sustainable data services, enhancing legal certainty, strengthening the trust in the digital marketplace and not to be a burden to innovation. The GDPR wants to be an enabler for big data services in Europe and provides the framework that strikes a balance between the protection of fundamental right, consumer trust and economical growth.[61]

This section expands upon this statement and explains how the GDPR will affect the use of big data in the EU and for companies located in other countries that want to exstract value from personal information collected in and regarding EU citizens.

This analysis identifies four characteristics of big data particularly subjected to the the GDPR:

1. unfairness and discrimination
2. opacity of processing
3. the tendency to collect all data

4. finding new purpose

This section identifies and analyses key aspects relating to the above mentioned characteristics and how they will be subjected to legislative acts under the upcoming regimen.

Section2.3 provides an overview of relevant technologies and techniques used in big data analysis, thereof explaining data mining and machine learning which is highly relevant for this analysis. Chapter 2.4 explained at a conceptual level how value is created and maintained through the big data value chain, all the way from capture to usage. Algorithms, data mining and machine learning are fundamental building blocks throughout. The following presents challenges and opportunities presented by the GDPR in that context.

Before continuing with the analysis a summary of relevant definitions from *Article 4* are provided below:

- **Personal data** is any information relating to an identified or identifiable natural person, where a natural person refers to an individual human being
- **Data subject** is the natural person to whom the data relates
- **Processing** means any operation or set of operations which is performed on personal data or on sets of personal data whether or not by automated means.
- **Profiling** is any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;
- **Controller** is a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller
- **Processor** means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data
- **Consent** is any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative

action, signifies agreement to the processing of personal data relating to him or her

- **Pseudonymisation** is the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information

- **Sensitive information** is also addressed as special categories of personal information (Recital 10) where sensitive information is racial or ethnic origin, political opinions, religious, philosophical beliefs, health, sex life or sexual orientation.

3.4.1 Unfairness and Discrimination

Big data analytics with machine learning in the forefront is about to render all kinds of consequential decisions about human beings. However, the lack of empathy in computers creates a dark side. Hardt [62] states that "machine learning is not, by default, fair or just in any meaningful way". Whereas, *Article 15* of the Directive was ambiguous on the topic,[53] the GDPR tries to make amends.

Traditional analytics and big data analytics differ in many ways. The general notion is that a traditional analysis involves deciding what you want to find out and constructing a query by identifying relevant entries. Big data analytics takes another approach. It usually involves a *discovery phase* and a *application phase* also regarded as "thinking with data" and "acting with data".[46]

The discovery phase runs algorithms against the data to find correlations. The outcome carries some uncertainty referred to as "unpredictability by design". When relevant correlations are identified new algorithms are created and applied to particular cases in the application phase.[63] Machine learning is widely used in recommendation systems, credit and insurance risk assessments, advertising and social networks.[64]

Automated individual decision making may refer to algorithms that make decisions based on predictions from learning systems. Ultimately this proves challenging to the rights and freedoms of natural persons and in particular unequal treatment of individuals based their membership in particular groups such as gender, race or religion.Goodman and Flaxman [64] describes this as discrimination.

The issue is addressed specifically through *Article 22(4)* and and *Recital 71* The former states that sensitive information shall not be processed irrespective of the necessity for entering or performing a contract, authorization by law or consent. *Article 9(2(a))* makes an exemption to where it is necessary for reasons of substantial public interest. The latter requires data controllers to implement technical and organisational measures to prevent discrimination.

Big data is not neutral with regards to discrimination, which can be explained by machine learning algorithms being programmed by people and are strings of math that rely on data that have been collected from society. As society contains discrimination such as inequality and exclusion, so will the data. Big data is also subjected to a uncertainty bias, which refers to discrimination due to a group being underrepresented in the sample. This meaning that risk averse algorithms causes predictions, other things equal, to favour the groups that are better represented.[64]

Thus, a sole reliance on data mining can lead companies to be non-compliant with the new Regulation. Machine learning algorithms are however capable of singling out patterns of discrimination given that they exist in the training dataset. Thus it is possible to prohibit the use of sensitive information. However, as datasets become increasingly large, correlations become increasingly complex and difficult to detect. This may cause the task of exhaustively identifying and excluding data features correlated with sensitive categories impossible.[64] Consequently, as some algorithms are of essential company value, they may be reluctant to exclude certain covariaties and choose to face the risk of non-compliance despite possible fines.

It is therefore not always possible for companies to safeguard against discrimination in algorithmic decision making and achieving fairness might be to computationally expensive. It may also place additional demands on those that engineer the learning process. Since some of the most interesting applications of AI tend to be at the limit of what's currently computationally and humanly feasible, the additional resources necessary for achieving fairness may be limited.[62]

The transparency principle aims to provide insight to a lack of fairness and discrimination, which shall ensure that the data subject gets informed to why an algorithm disfavoured them. However, this may also not always be possible due to uncertainties to how the

machine reasoned when arriving to a decision. This is also referred to as the "black box".

3.4.2 Opacity of Processing

Where the last section explained how machine learning algorithms may provide discriminating results, this section is a continuation of the issue and focuses on the logic behind automated decisions and profiling, or lack thereof. One of the big differentials between traditional analytics and big data is a lack of humane comprehension to the rationale behind an algorithmic decision. While machines become smarter and are able to make "inhumane" decisions, the ability of humans to provide an explanation to the rationale behind the decisions diminish accordingly[63].

As these algorithms may inevitably make discriminating, erroneous or unjustified decisions, the GDPR requires transparency on the rationale behind the particular decision. *Articles 13 and 14* state that, when profiling takes place, a data subject has the right to "meaningful information about the logic involved." also referred to as *the right to explanation. Article 15* furthermore specifies the *right to access* information as the right of the data subject to obtain confirmation to whether or not personal data is being processed and a right to access that personal data.[49] Put in other words, if one's loan application is denied the rationale behind the decision must be provided along with a disclosure of information to what data was used, given that the decision was automated and not subject to human intervention.

Deep learning is one of the state of the art machine learning techniques which feed outputs into successive layers using the previous layer for input. The complexity of successive layers feeding outputs to the next create a "black box" effect. This is what characterizes the opacity of the processing in big data. The algorithm uses an input and produce an output, opacity occurs as the recipient of the end output rarely understand the the reason behind the particular decision in question.[65] Additionally, the inputs may even be entirely or partially unknown. The inevitable opacity of processing makes it difficult to understand the reasons for decisions made as result of deep learning.[64] Furthermore implying that the more powerful an algorithm becomes, the harder it is to decipher.

This begs the question to whether *Articles 12(1)-15* in the GDPR are complied with under

automated decision making. If this is the case, what are the requirements to explain an algorithm's decision? The issue is two-fold in this sense. Being unable to provide "in a meaningful way" the logic behind a decision is a breach of *the right to explanation* and being unable to know what input goes into the algorithm may cause a breach of *the right to access*.

However, some law professionals[66] argue that the *right to explanation* is flawed due to the word "solely" in *Article 22* and is argued to be easy complied with just by introducing a human in the process. In the opinion of the author there may be some truth to this, but arguably not practicable. Human intervention would require increased manpower doing, quite repetitive tasks, which is contradicting the purpose of automation all together. That said, the issue of transparency in algorithmic decision-making stays relevant. Companies, including the public, is loosing consumer trust due to privacy concerns and the demand for transparency has consequently increased. Furthermore Burrell [65] distinguishes between three barriers to transparency as a result of:

1. corporations and institutions intentionally concealing decision making procedures;
2. current state of affair where reading and writing code is not general knowledge;
3. a mismatch between complexity due to mathematical optimization and the demand for human interpretation

The first barrier is partly amended by *Article 13*, with a right to be informed and to be provided the logic behind.[64] This rises a big questions to what this will mean for protection of company secrets and intellectual property? The opacity of algorithms could therefore be attributed to self-protection in the name of competitive advantage.[64] Companies may therefor become reluctant to disclose details on their algorithms as they may contain valuable information about their business. However, opacity may also cover manipulation of consumers or patterns of discrimination.

When the GDPR become legally effective either scenario will present increased risk for companies, however the latter should be avoided at all cost. The GDPR do nevertheless present certain data-intensive companies with a dilemma with two horns. On one side disclosing valuable information may weaken companies competitiveness whilst on the other side the repercussions of concealing information will not only lead to sanctions, but

also investigations which might expose areas of non-compliance not even known to the controller or processor.

The second barrier is addressed by *Article 12(1)* requiring the information mentioned above, to be "concise, transparent, intelligible and easily accessible form, using clear and plain language".[64] The ability to translate an algorithmic process to common tongue is a scarce expertise. Thus, companies may be forced to invest in competence, which may become very expensive for smaller firms.

As algorithm gets better, the more complex the logic behind decisions become, which in turn, is the third barrier of transparency. This poses especial challenges to whether it will be possible to provide a interpretable explanation of the decision-making rationale. The rise of deep learning makes it even more challenging. There is nevertheless light at the end of the tunnel as it is an emerging field of research. Research are carried out to gain insights to the behaviour of opaque algorithms.[67]. Meanwhile, new data sources emerge with unintelligible semantics. Correspondingly advanced analytics is developed to process the data and complexity increases.

For companies to be able to harness the full power of machine learning, and deep learning in particular, development of means to provide transparency in processing will become crucial. If not, trade-offs between value creation and enforcement of privacy may become a future scenario. Consequentially, reducing the value of the output and potentially the discovery of new uses of the data.

3.4.3 "More data more problems"

So far challenges of fairness and transparency in processing related to machine learning algorithms has been depicted. The efficiency and subsequently the value driven from these techniques relies also on data to train on as well as datasets to analyze. When being able to harness all data, why sample? Mayer-Schönberger and Cukier [4][p.27] put i beautifully "A normal distribution is, alas, normal. Often the really interesting things in life are found in places that samples fail to fully catch." This is also the notion behind N=all: If feasible, all data is collected. An indiscriminate collection and ad hoc retention of data can provide individuals and society many benefits.[27] However, in light of recent technological

advances, with data mining in the forefront, more data means more problems.

The expanded definition of personal data described in 3.3.2, has created the notion that all consumer generated data, or data exhaust, can be considered personal data. This indeed, when tending towards collecting all data possible increase the risk of processing personal data in automated decision-making by a long shot. Companies may ignorantly possess personal data.[63] This may prove it impossible to provide a data subject information to whether personal data is being processed or not, enhancing the "black box" effect additionally. Regardless of this, the introduction of GDPR and the tendency to collect all data has other implications as well.

A big distinction between data mining and previous techniques of processing is the ability extract value from extensive volumes of complex data. Other important features of data mining include creating so called "new knowledge", such as abstract distributions and useful predictions. Additionally, the ability to create their own hypothesis automatically, make them non-reliant on human creativity.[27] So from a purely research point of view the more data the better, but this is not the case when the rights and freedoms of individuals are concerned.

Data mining vs. data minimization and storage limitation

The ability of companies to store data in an easy accessible way allowing for aggregation and mining is essential to extract value of data over time. However, retaining personal data in big data environments require attention and care, particularly concerning the risk of identifying patterns relating to specific individuals. Thus, emphasizing the importance of curation. Moreover, the *data minimisation* and the *storage limitation* principle addresses in particular the risk of identifying individuals or sensitive information concerning them. The provisions require that "personal data must be adequate, relevant, limited to, and kept in a form which permits the identification of individuals stored for no longer than necessary in relation to the purpose of the processing."[49] The rationale behind the respective principles are; if the data does not exist it cannot be abused; and the longer the data stays in a form in which it can be abused or stolen the higher the risk of exactly that to happen. This makes a reconciliation the notion of data mining difficult.[27] Considering that the power and value of data mining comes from large and variable datasets, then a limitation

would diminish the potential outcome. Furthermore, the storage limitations principle states that personal data should be erased after the purpose of the processing.[49] Which is only adding to the restraints.

The GDPR presents pseudonymisation as a potential mend to the issues above.[60] This subsection and the following is closely related, where the following presents the purpose limitation principle as an inhibitor of reuse, whose opportunities can be discovered through data mining.

Reuse vs. purpose limitation

The notion behind the value of data was presented initially in this thesis. It explains how the value of data must be considered in terms of all the ways it can be employed in the future, where the ultimate value can be seen as the sum of all the uses. [4] Furthermore, essential to understanding how these uses are enabled is the value chain of transformation, that a higher value output is achieved at higher levels of transformation. Where a higher level of transformation may as well represent one or multiple new usage areas.

This was portrayed previously in this section as the differences between traditional analytics and the ability of machine learning to discover new knowledge and generate an hypothesis beyond the imagination of humans. This often results in the revelation of new areas of usage. However, the GDPR puts a spanner in the works through the *purpose limitation*. Two main issues arise. Firstly, it presents a barrier to discover new usage areas, and secondly, it curtails the opportunity to realize the value.

The principle states that "personal data should be collected for a specified, explicit and legitimate purposes and not be further processed in a manner that is incompatible with those purposes."[49] If compatibility with the purpose can not be proven a fresh consent must be obtained. But because a fresh consent entails additional juridic baggage this is wished avoided. Consequently, companies may be tempt to create exhaustive contracts, but this is prohibited with the new provision on explicit consent.

However, The Article 29 Working Party(WP29)[60] explains that the ultimate objective of this principle and the word "explicit" is to ensure that the purpose(s), however unrelated, should be without vagueness or ambiguity, leaving the purpose clear to all involved in the

processing, irrespective of cultural background, level of understanding or special needs. This excludes contracts that allow processors leeway in doing what they want with the personal data. Since it is impossible to get consent for a purpose that is yet to be discovered this principle is considered as a barrier to the development of big data analytics.[63]

The GDPR open up for ways in which collected personal data can be processed and used for new purposes. *Article 6(1)(b)* allow for further processing for historical, statistical and scientific research as long as appropriate safeguards are implemented, which means that risk to the data subject should be mitigated or at least minimized. The provision highlights that the data should not support "measures or decisions" regarding particular individuals[60]

Particularly interesting in the context of this thesis is *statistical research*, which encompasses commercial purposes such as market research and public interest such as the environmental research,[60] which is relevant for the case.

The WP29[60] distinguishes between two types of further processing: detecting trends and correlations; and finding out about individuals and make decisions affecting them. For the latter, the WP29 advocates *functional separation* between analytics operations and for the latter suggesting to always obtain a consent.[63] Henceforth referred to as type 1 and type 2 further processing.

The WP29 explains functional separation as the means to secure that "data used for statistical purposes or other research purposes should not be available to "support measures or decisions" that are taken with regard to the individual data subjects concerned (unless specifically authorized by the individuals concerned)." However, they(WP29) suggest that most data can be released for reuse given sufficient aggregation or effective anonymisation with an exception of open data. It can be argued to what extent this statement is valid due to recent technological advancements since the adoption of the Opinion in 2013. The continuation of the thesis assumes that it holds nevertheless.

Privacy risk is not mitigated completely by pseudonymisation. In a big data environment there is an increasingly persistent risk for re-identification,[60] and when two datasets considered non-sensitive is combined the risk for re-identification or discovering sensitive information increases.[4] The data subject should be informed about such risks and may consequently become reluctant to consent.

3.4.4 Seeing Through the Challenges

From a consumer point of view it stands to reason that the unique value proposition must at least be in accordance to the privacy risk exposure. In which they also should be well informed about in a clear, concise and meaningful way.[49] Equally important is building trust between the consumer and the controller. Additionally to consent as a trust builder,[68] the Regulation provides different measures in this regard, particularly through enhance rights of the data subject such as:

- access and correction;
- right to be forgotten;
- right to object;
- right to be informed;

Also worth mentioning in this context is the appointment of data protection officers and privacy by design and by default.

The "oxymoron of big data and privacy", explained in 2.5.1, illustrates through an exacerbation of the current state of affairs the value of privacy in big data. The success of service providers using big data is closely linked to their capacity to build and maintain customer trust,[7] and is in the sense essential for ensuring value over time. Thus, when considering the oxymoron, the GDPR becomes an enabler of value in the long-run.

Compliance for long-term value

GDPR imposes a high bar for compliance, and tough sanction upon those who fail to comply. The most efficient alternative to processing may therefore be to not process personal data at all. This may be done by wiping all personal data or rendering it completely anonymous.[63] However, as previously explained, purging personal data and correlated data may render a dataset worthless. Let alone the fact that, assuming all data as personal data result in no dataset at all.

This leaves rendering the dataset anonymous as the most expedient alternative. The GDPR

introduces the concept of pseudonymisation which is the processing of personal data so that it can not be attributed to a specific data subject.[49] This may allow for companies to keep data for longer enabling more value to be realized through analysis and reuse.

Pseudonymisation provides additional benefits. One of which is protection of the data subject's privacy in case of personal data breach as pseudonymous data are less likely to cause harm to affected individuals. Even though a risk for re-identification exist the risk of sanctions and claims for the relevant organisation is considerably reduced.[63]

There is nevertheless still an elevated focus on security in the GDPR explicitly provided through the "integrity and confidentiality principle" and in particular *Article 32*, which states that "controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk." Some of the required measures include pseudonymisation and encryption of personal data as well as confidentiality, integrity, availability and resilience of processing systems.[49]

As the word "appropriate" insinuates, the GDPR require companies to consider current state of the art technology and are therefore expected to stay ahead of possible threats to privacy imposed by either the analytics technology used or external threats such as cyber attacks. This may eventually lead to a big "win" for achieving consumer trust as well.

3.5 Key Findings and Concluding Remarks

This chapter presented the GDPR at a holistic level, described the need for change, an overview of the former legislation and the coming reform. Key changes was explained and how these will impact big data was depicted. The key findings are presented below and summarized in a influence diagram.

One of the big features of big data analytics and in particular machine learning is the algorithms ability to make decision is beyond human comprehension. This is both in terms of the intelligence of the decision but also in terms of humanity. This challenges the *fairness* and *transparency* principles of the GDPR in particular:

Fairness

- Because of discrimination in society there is a high possibility that the data processed by algorithm provide discriminatory results. Also, risk averse algorithms may disfavour minorities in the population. Addressed *Article 22* in GDPR.
- It may become impossible to exhaustively purge a dataset for discriminatory information without rendering it worthless.
- Some companies rely heavily on their algorithms and the GDPR may require them to discard competitive advantage. Consequently, some companies choose face the risk despite huge fines.

Opacity of processing

- The complexity of algorithms create black boxes". Making it difficult to explain and understand the rationale behind behind a automated decision or profiling. Addressed by *Articles 12(1)-15*.
- Opacity may become unavoidable as analytics become more advanced
- Opacity may be due to: concealment of decision making procedures in the name of competitive advantage; covering manipulations and discrimination; lack of expertise and general knowledge; and mismatch between algorithmic complexity and demand for human interpretation.
- Dilemma with two horns: On one side, disclosure of algorithms logic may weaken competitiveness, on the other side repercussions of concealment may lead to maximum fines.
- Algorithmic transparency become essential to harness the full power of big data analytics under the GDPR. Failure to do so may lead to trade-offs between value and enforcement of privacy.

The second big feature of big data is the tendency to use all data. Because cost of storage and processing has plummeted companies are able and know how to benefit from collecting

all data feasible to obtain. Bigger datasets enable mining for knowledge that a smaller data sets fail to provide. However, this challenges the principles of *data minimization* and *storage limitation.*

Tendency to collect all data

- If the data does not exist it cannot be abused and the risk of such is reduced by erasing it.
- Data increases in value with aggregated levels and analytics over time. Thus, a reconciliation between data mining and the principles will become difficult.
- Generally it is expedient to store data for indefinite periods of time but the GDPR requires personal data to be erased after initial purpose
- Pseudonymisation ensures that companies can store data for longer, but with risk of re-identification
- The GDPR adds additional responsibilities to ensure accuracy of data, ensuring that data quality is maintained

The third big feature of big data is that the ultimate value increases with reuse. One of the purposes of mining the data is knowledge discovery, which may lead to new usage areas. This challenges the *purpose limitation* principle in particular.

Reuse

- The principles of *purpose limitation*, *data minimization* and *storage limitation* presents barriers to discovery of new use areas and consequently curtail companies opportunity to realize value through new use.
- Further processing for detecting trends and correlations requires functional separation(type1)
- Further processing for finding out about individuals and make decisions about them(type 2) requires consent

- Further processing may be carried out for purposes such as market research and the environment, as long as the data does not support measures or decisions regarding particular individuals
- Can not obtain consent for a purpose not yet discovered
- While it may be tempting, exhaustive contracts is ruled out
- May become a barrier to big data analytics
- Sufficient aggregation or effective anonymisation can justify reuse, except open data.

This chapter described the GDPR as a trust builder and an enabler of long-term value on basis of the "oxymoron of big data and privacy". Additionally the importance of accuracy in ensuring data quality was described. Figure3.1 illustrates the key findings of this analysis as influences on the identified value drivers of big data from 2.4.6

The following chapter presents a case where these findings will be used as a frame of reference to analyse how the GDPR influences the ultimate value of smart meter data.

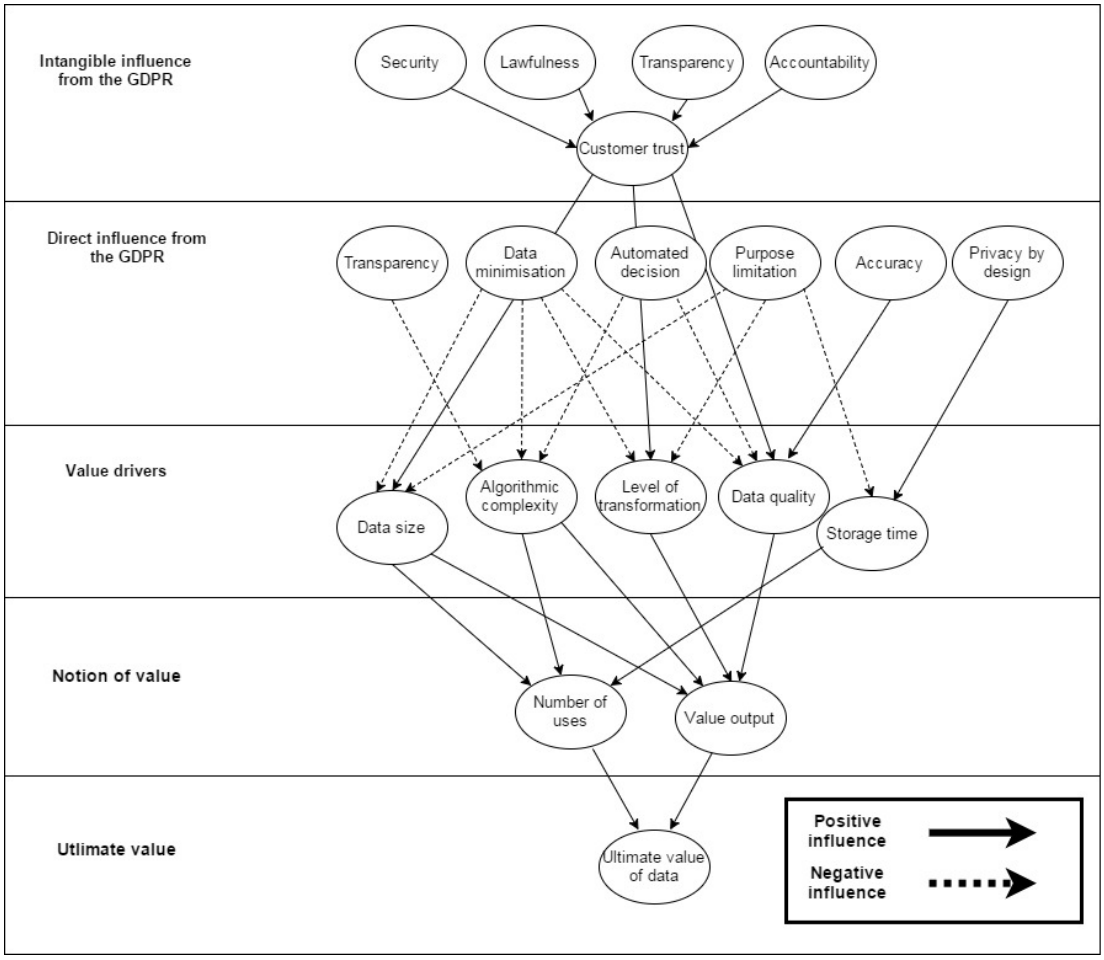

Figure 3.1: Influence diagram: Big data value drivers and GDPR influences

Chapter 4

Case study: Smart Meter Data

So far this work has identified a set of value drivers in big data and identified related influences from the GDPR, which was presented in the conclusion of previous chapter. This chapter exemplifies influences and tries to identify new ones.

The author acknowledges that there are more applications of smart meter data than what is covered, but has chosen those most prominent through a literature review, which can be found in A B and C Furthermore, smart metering has several privacy concerns not covered in this case study. The concerns covered are related to big data.

This chapter initiates with the background of smart metering, before presenting the fundamentals for understanding smart meter data in the context of the thesis. The analysis is carried out through the big data value chain.

Literature review

Prior to the analysis a comprehensive literature review was carried out where particular data-intensive applications was singled out for the case study. This resulted in the influence diagram, presented in figure 4.1, that illustrates the breakdown of transformations and applications of smart meter data. Transformations in bold with respective input and output.

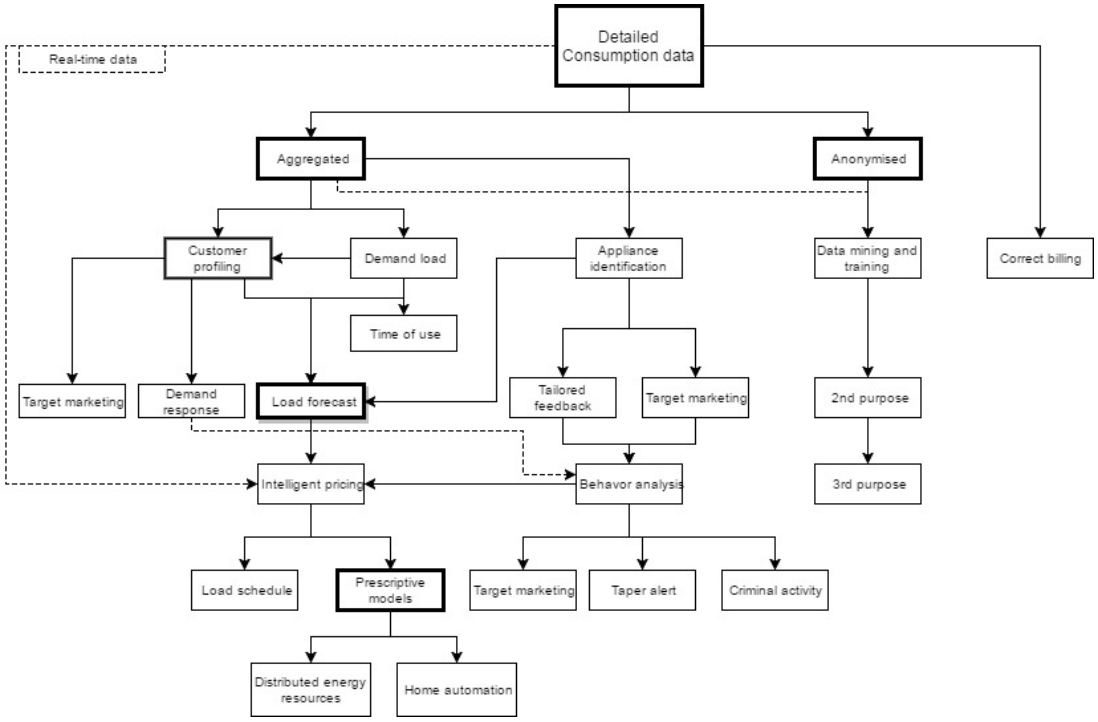

Figure 4.1: Influence diagram of smart meter enabled

4.1 Background

Utilities are facing a growing challenges in ensuring a balance in supply and demand of electric power. The cost of generating and delivering electrical power correlates to the demand in the grid, and as it stand, utilities are facing a challenge of *peak load* at particular hours (peak hours) of the day, which challenges the integrity of the grid.[69] The implications are high cost of electricity supply and correspondingly high prices for the end-user. As consumers continue to purchase new electricity demanding appliances the demand for utilities to generate and deliver increases accordingly. If the development continues the

capacity of the power system will face its limit and new investments will be required. This is neither in the interest of the consumer or the utility.

Adding to the challenge are the push from policy makers and "green" initiatives. The EU 2020 Strategy[70] aims to "reduce greenhouse gas emissions by at least 20% compared to 1990 levels; increase the share of renewable energy sources to 20%; and a 20% increase in energy efficiency." Consequently, the fleet of electrical vehicles (EV) are growing and renewable energy sources (RES) are increasingly penetrating the grid. The challenge remains, however, and is boosted even, as EV-charging happens at peak load and the most efficient power generation occur at *load valleys*. In some countries potentially resulting in overgeneration during valleys and risk to grid integrity during peaks-

Society as a whole face three challenges in the future: shifting consumption away from peak hours; reducing baseline consumption; and increase share of renewable energy. The consensus is furthermore to address the challenge at the demand-side, that is, changing customer energy behavior. Demand-side management (DSM) has an important role in meeting the challenges in the power networks.[71] DSM is the utility activities that influence customer use of electricity and include customer centric approaches such as feedback programs, demand response (DR) programs, and demand-side automation.[72] DSM also include consumer generation, such as solar panels and wind power, and energy storage, which is in the following referred to as distributed energy resources (DER). Furthermore DER is a part of the smart grid that allow for two-way electricity flow and "prosumers"[73] as well as short-term flexibility between the use of centralized generation and distributed generation(DG).[71] Which, in turn, ensures better reliability in the grid and efficient use of RES.

However, current technology and IT solutions limit utilities' ability to communicate and cooperate with the end consumer.[1] The wide spread deployment of smart meters mark the fusion between IT and provide utilities last-mile communication and insight in the grid. This is furthermore acknowledged as the first step in facing the future challenges of the grid.[74]

4.2 Smart Metering

Traditionally, households have submitted their energy consumption by manually reading their electricity meter once a month or by a monthly visit by a meter reader.[75] However, a widespread world wide roll-out of smart meters allow for the majority of customers to have their energy consumption automatically read at set intervals, usually between 15 to 60 minutes and are even able to opt for sub-minute readings for more advanced services services that can be integrated to the *smart home*. In this particular context home energy management systems(HEMS) are relevant, which utilizes the home area network (HAN) provided through the advanced metering infrastructure(AMI). More about these technologies are provided in Appendix C.3, A.2.3 and A.2 respectively.

The smart meter allow for consumers to register own generation, or to become "prosumers" and be compensated thereafter. When opting for higher resolutions, or shorter time intervals of reading or sampling rates, customers are able to recieve better services and utilities are able to harness more data to utilize. When collecting data from smart metering devices, assuming that 1 million collection devices retrieve 5 kB of data per single collection, the potential data volume growth in a year can be up to 2920 TB.[1] Also illustrated in table 4.1. This constitutes the volume characteristic of big data. Smart meter data from a large utility

Table 4.1: The amount of data collected by 1 million smart meters a year[76]

Sampling rate	24h	1 hour	30 min	15 min
Number(billion)	0,37	8,75	17,52	35,04
Volume(terrabyte)	1,82	730	1460	2920

may generate millions, and even billions of events every second, that needs to be processed and analyzed in order to ensure reliability of the grid and to deliver services to consumers. This constitutes the velocity of big data. Utilities can furthermore combine meter data such with multiple sources such weather, distributed generation and home appliances. This constitutes the variety of big data. The possibilities in collecting and analysing smart meter data are seemingly endless. However, the following is restricted to a few, mainly aimed at the consumer.

4.3 Applying big data analytics to smart meter data

Assumptions and limitations

The utility sector of different countries distinguish between roles in the supply side differently. Some common roles are:

- generate power and distribute;
- operate the grid and ensure reliability;
- supplying power to the end consumer;
- buying and selling power;
- processing meter data;
- providing products and services to the end consumer;

These roles are given different names depending on country and market model and their roles oftentimes overlap. Therefore, to avoid any ambiguity and confusion in the following, and because of the context of the utility sector, the parties are addressed as utilities in the following.

The prior assumptions of this thesis hold for this analysis. Where the notion of processing in GDPR for simplicity sake is limited to:

- big data analytics used for knowledge discovery and new usage areas, that is, achieving value over time
- and big data analytics enabling higher level of transformation, that is, achieving a higher value potential.

Processing is what drives value through the value chain while reuse is the realization of the value. Furthermore, the *notion behind the value of data* is the sum of all possible areas in which data can be used.

Further processing is simply processing for a new purpose where this analysis adopts the WP29[60]*The Opinion 03/2013 on purpose limitation* as best practice. This entails in short:

- Analytics for detecting trends and correlations (Type 1) requires functional separation.
- Further Type 1 processing is considered compatible if for the purpose of public interest, historical, statistical and scientific research. Entailing that, in this particular case, utilities are allowed to process personal data at default privacy setting, for purposes related to the smart grid, demand side management, and market research.
- Finding out about individuals and make decisions (Type 2) requires fresh consent
- Further Type 2 processing is considered compatible if fresh consent is obtained, given that the new purpose is incompatible with initial purpose of collection.
- Smart meter data is assumed as not open data, as it's personally identifiable. It can therefore if, sufficiently aggregated or effectively anonymized through means such as pseudonymisation and encryption, be reused.
- The safeguards above is furthermore assumed to be selected based on a sufficient DPIA.

Privacy by default under the GDPR requires companies to provide the highest privacy setting on their services and products as default. In case of smart meters it is assumed that the default privacy setting is a sampling rate at 60 min as this will not give detailed information about occupancy. Any resolution higher than 60 min will require a fresh consent.

Smart meter data, namely consumption data is assumed to fall under the definition of *personal data* in the GDPR. Collection of consumption data and processing as such in the name a sustainable society, and in particular the smart grid and demand-side management (DSM), is assumed to be in public interest. It is however, not considered in the public best interest that utilities increase competitive advantage by marketing new products and services for DSM

Initially, when presenting the case an additional problem statement was presented as: *How will the GDPR impact potential of smart metering?*

Accordingly to the the *how* to use smart meter data explained in 2.4.5. this analysis assumes that utilities either want collect and process data to:

- apply by DSM or smart grid;
- operationalize by target marketing and identification of irregularities;
- monetize by selling consumption data, raw or synthesized

4.3.1 Data acquisition

In a big data context the acquisition of data is seen in relation to the phenomenon of datafication as described in 2.2.5. This applies to acquisition of smart meter data as well, where the aggregation of millions of individual meter readings, when aggregated and put in a system can yield the foundation for new knowledge, intelligent systems and new revenue streams.

Smart meters provide four different types of data: power consumption data, generation data, event data, and power quality data. Power consumption data is the most relevant data type for this thesis. However, it must be noted that the other data types has their relevance for most applications, but the focus through this analysis is on power consumption data, which is further divided into the following:

1. *Detailed consumption data:* 15 to 60 min interval readings of electricity consumption;
2. *Billing interval data:* Readings at the beginning and end of billing intervals;
3. *Aggregate statistical data:* Monthly consumption, comparisons and history;
4. *Broadcast data:* Communication to the use about price chance, critical peaks and reliability;

From the above mentioned, the by far most used measurement data is detailed consumption data, henceforth referred to as consumption data and is the main type of concern.

Smart meters ability to provide detailed information from consumer energy consumption is unprecedented. The higher the resolution the more information can be inferred. Different algorithms[77, 78] can used to identify appliances in the household and predict the behavior of customers. Figure 4.2 illustrate how a higher data resolution yield different richness of information about consumption. At half-hour intervals indicative periods of

consumption can be inferred, giving insight to for example occupancy of the household. At 1 minute intervals on the other hand appliances can be identified due to their load signatures.[11] Taking it to the extreme: by using a 0,5s^{-1} sampling rate Greveler et al. [79] were able to reveal the type of TV-channel and even using the power profile to identify the content of the media that was displayed on TV-screens.

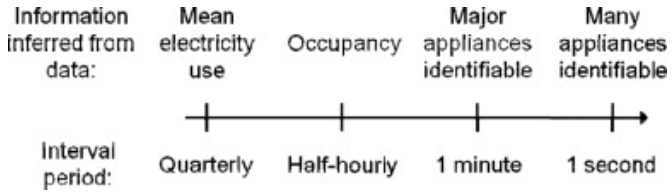

Figure 4.2: Examples of information inferred from different data resolutions[80]

Smart meters are undoubtedly able to provide valuable information about a the household and the resolution is a decisive factor to the richness of the information. Larger datasets and more granular data yield a higher probability for data mining algorithms to discover hidden patters and for learning algorithms to be accurate with their predictions.[14]

However, as more information is inferred the more intrusive smart metering becomes and the potential for revealing personal and sensitive information increases[11, 80]. The global consulting firm refers to smart meters as a "gateway to the home"[10] which pretty much summarizes the controversy of smart metering. Summarized in figure 4.2 are some privacy concerns that arises with a high resolution.

Table 4.2: Examples of interested parties and their intentions (Adopted from[81])

Interested parties	**Purpose**
Insurance companies	Determine health premiums based on behavior indicating illness
Marketers	Target advertising
Creditors	Determining behavior that might indicate creditworthiness
Law enforcements	Identification of suspicious or illegal activity
Criminals	Identification of best times for burglary and valuable appliances
Civil litigators	Identification of property boundaries and activities on the premise
Landlords	Verification of lease compliance
Private investigators	Monitoring of specific events
Press	Get information about famous people

The GDPR addresses this issue and will directly curtail the leeway to take advantage of higher resolutions. Particularly relevant provisions are the data *minimisation principle*. It will directly prohibit excessive collection by resolutions higher than what is necessary to achieve the purposes informed about in the customer consent

Additionally, because of *privacy by default* customers with smart meters will by default provide the least sensitive information necessary for the purpose of the processing. This will most likely mean that hourly to half-hourly sampling rates will be default and an "opt-in" consent will be needed for any higher resolutions. As stated previously it is assumed to be 60 min by default in the following.

This is a classic case of "chicken or the egg". The optimal situation for the utility is to collect as much consumption data as possible, namely at the highest possible resolution, but is restricted by design in the GDPR. The only way to obtain this information will be through a fresh consent, which then will become the only way to maximize the value from the smart meter.

However, if data protection authorities find it necessary to obtain higher resolutions for purposes in "public interest" this may open for additional processing opportunities for utilities. For example letting utilities sample at necessary resolutions for processing in public interest, but only process at agreed upon resolutions for purposes provided in the customer consent. However, the understanding of the author is that the GDPR does not open for such actions.

Processing smart meter data

Utilities processes data in real-time to provide decision-support in demand response and operation of the the grid. The ability to process data in real-time is of great importance to ensure reliability in the grid. Real-time processing provide the ability to monitor the status of the grid, and modern technology know how to utilize this data to ensure efficiency and resilience.[82] Figure 4.3 illustrate the importance of timely data in decision-making where smart meters provide the infrastructure to monitor the grid all the way to the end-user. Enabling proactive pricing and execution of load control during emergencies.[75]

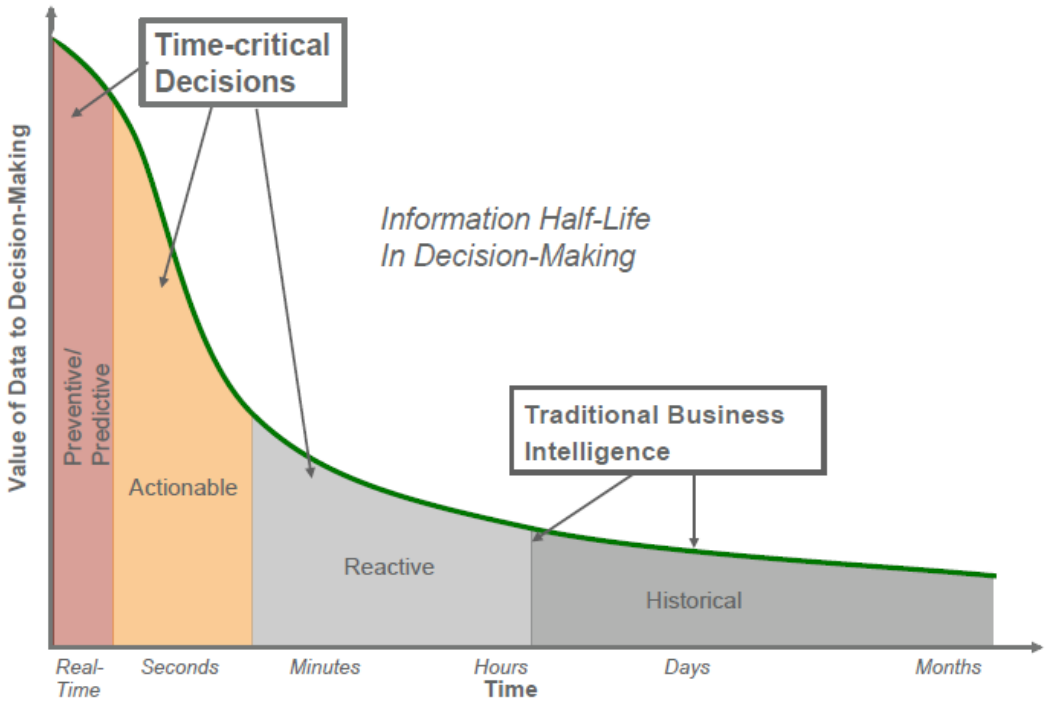

Figure 4.3: Value from speed of processing [32]

In addition to secure grid operations, real-time processing provide customers timely feedback giving them increased flexibility to respond to demand in the market. The value of processing speed is manifested in figure 4.4 where the real-time feedback yields the largest energy savings.

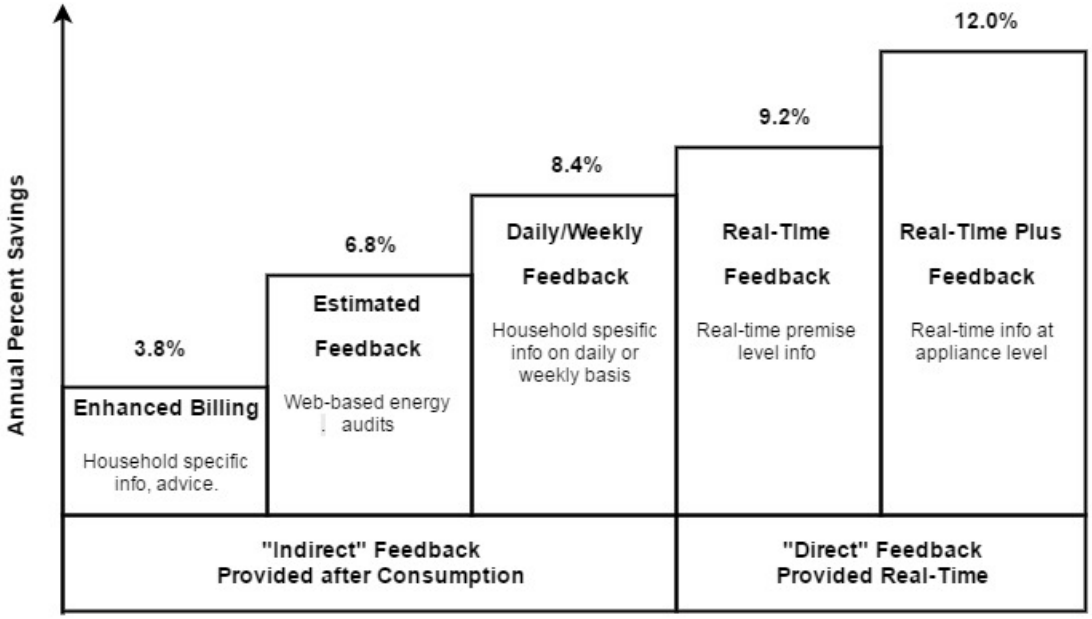

Figure 4.4: Value of timely feedback (adopted from [83])

A higher resolution enable more detailed feedback, for instance about appliances that consume the most power, and at what time of day they should and should not be used. Hence, it comes to show that customers as well as society as a whole can benefit largely from high resolution data. This is, however a two-egged-sword: on one side, high resolution means high risk to privacy, while on the other side, providing detailed information will save customers money and ultimately also benefit the environment. However, consumers can not be forced to opt-in for more detailed feedback. A voluntary explicit consent must therefore be obtained.

Because of the seemingly high savings potential from feedback programs these could eventually become enablers for utilities collect at higher resolutions. This assuming that utilities won't be able to collect at higher resolution without customer consent for public interest purposes. Such a feedback program would arguably be opted for through a user friendly smartphone app.

4.3.2 Data analysis

In the big data era, one of utilities biggest and most important challenges is to use smart meter data beyond its core function, which is measuring consumption for billing purposes. In this context one can argue that smart meters are a prime example of extensibility by design manifested in a new technology. Figure 4.5 illustrate different levels of transformation and applications enabled relative to resolution.

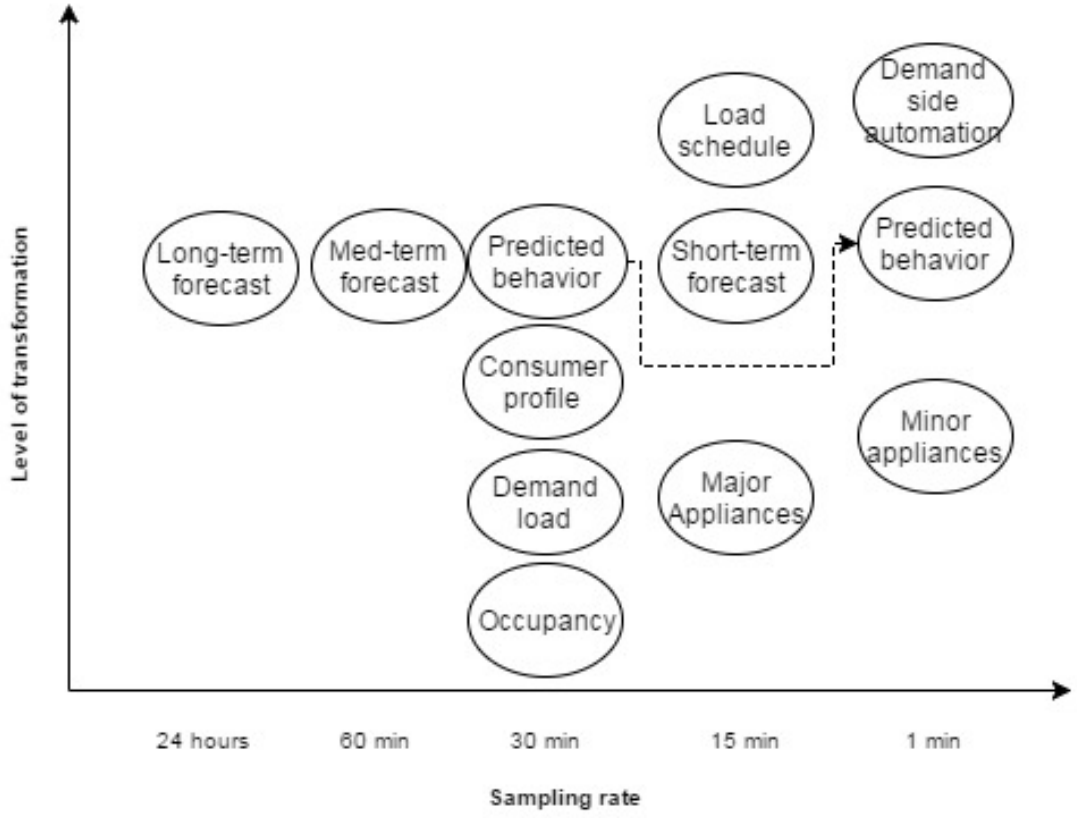

Figure 4.5: Applications enabled with resolution and transformation

This section provides an understanding to how analysis enable new use of smart meter data and at higher levels of transformation. The assumption is that the more advanced the analytical approach becomes, the higher value output. Chapter 2.4.2 explained three different analytics approaches representing different levels of transformation. The list below summarizes how the extensibility of smart meter data is manifested at different levels of transformation.

1. *Descriptive analysis* is mainly analyzing aggregate data to reveal peak loads at different locations in the grid throughout the day. Used to create demand loads

2. *Predictive analysis* create consumer profiles and predict behavior and load forecasting based on typical load patterns and additional data sources such as weather and season

3. *Prescriptive analysis* use predictive models to create intelligent systems that respond to real-time events such as pricing signals and consumer behavior

The outline of this subsection is, however, different. The following is structured to showcase a transformation to the reader – from sampling, to aggregation, to clustering and eventually to profiling. Subsequently demand load is explained as another track of transformation. Moving out in the value chain of transformations, load forecasting is explained as the combination of customer profiling, demand loads and additional data, illustrating the power of combining datasets. Concluding the transformation two models for optimizing demand side energy consumption is presented. The states, which is consumer profiles, demand loads, load forecasts and optimization models are referred to as transformations in the following.

Consumer profiling

Daily meter readings when aggregated provide energy consumption patterns of households. They can be used for a variety of purposes and different levels of granularity provide input to different analysis, where weekly, monthly and yearly analysis of consumption behaviour can help utilities plan for future energy requirements as well as help consumers manage their energy consumption.[9] Information such as occupancy and in home activities can be derived as shown in figure 4.6(a) and (b) respectively.

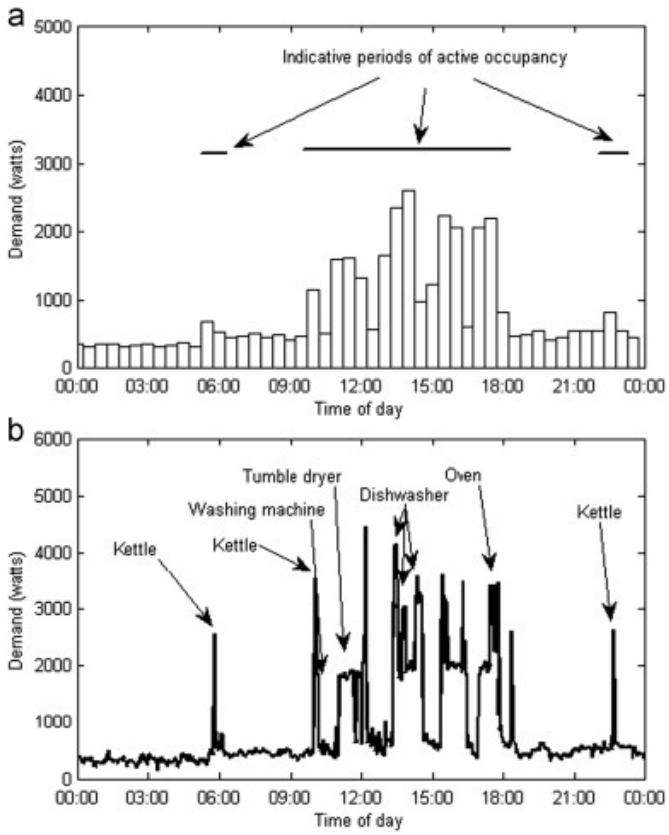

Figure 4.6: Information inferred from half hour(a) and 1 minute(b) readings[80]

Analytics help identify hidden trends in consumers energy behaviour, also called typical load patterns (TLP).[84] Algorithms can identify consumer groups by creating a load profile represented by a classification of each electricity customer on the basis of their behaviour.[85] These customers can be segmented and targeted with information, services and products that is relevant for them. [86]

Figure 4.7 show the use of a clustering algorithm to identify different consumer patterns which generally corresponds to typical energy personalities shown in figure 4.8.

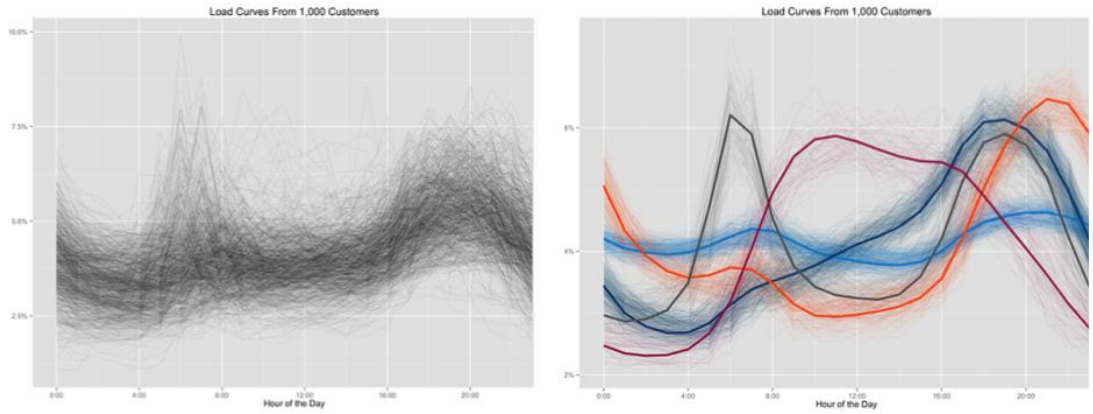

Figure 4.7: Appliance of a clustering technique to discover usage patterns [87]

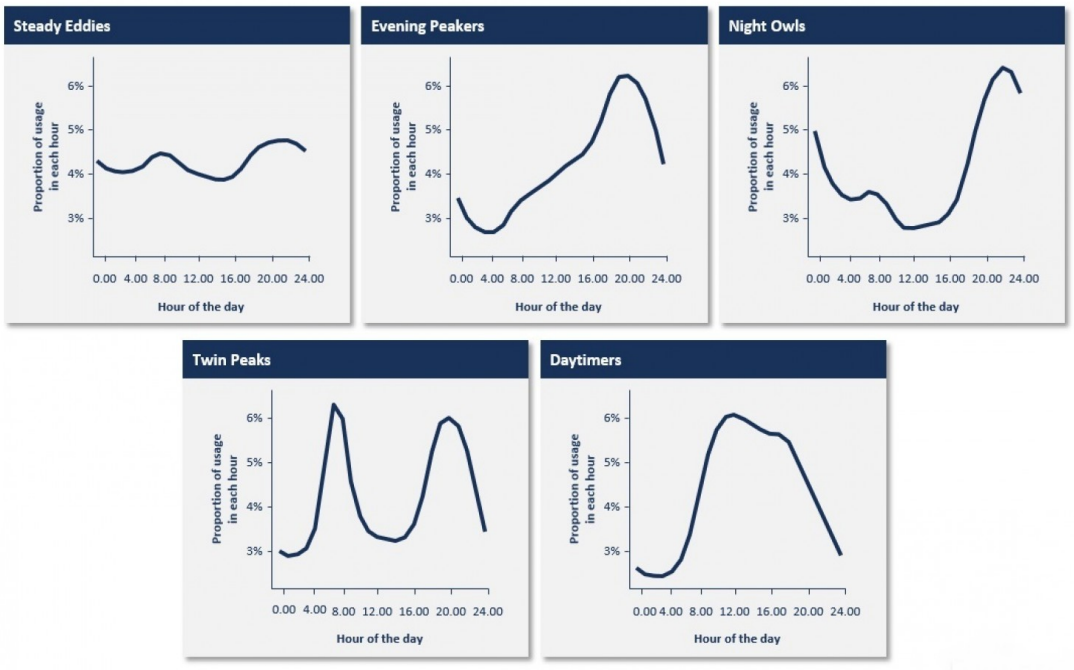

Figure 4.8: Different types of energy personalities [88]

Classifying data subjects based on their energy behaviour is in the following assumed to fall under the definition of *profiling* in the GDPR. Different personalities are associated with different groups of the demographic: The "Steady Eddies" tend to live in condos and have a high level of winter energy usage from electric heating; "Evening Peakers" represent single family homes using a lot of power in the summer on air conditioning; the "Night Owl" is usually young people and apartment owners; "Twin Peaks" are linked to wealthier families in single-family homes with electric heating; and the "Daytimer" is usually old people with few kids.[88]

Depending on the personality, for example, different energy-efficiency programs may be promoted. "Steady Eddie" would benefit from a reduced baseline consumption, whereas an "Evening Peaker" will benefit from reducing peak usage. "Twin Peaks" may benefit from both as they have high peaks and a high average baseline. Considering that they tend to have a high income and maybe own an electric vehicle a Home Energy Management System would be ideal.[89]

The following extrapolates on the previous example with a potential scenario. Gajowniczek and Ząbkowski [90] presented a study using unsupervised machine learning to detect household characteristics at appliance level. This will arguably enable even more detailed personal profiling and possibly stereotypical classifications such as: "The Gamer", derived from use of PS4; "The Workaholic" derived from occupancy hours and load signature of a laptops associated with business people; or "The Athlete" derived from rarely using the shower at home and has a fitness inspired social media profile. In this not too far-fetched scenario, target marketing would become more efficient while automated decisions would become more discriminating. Assuming equal income, a loan application would probably favor "The Workaholic", while a health insurance premium would probably favor the "The Athlete".

Article 22 and the definition of profiling in *Article 4(4)* is understood in a manner that the the data subject has the right not to have their personal data processed for analysis or predictions of performance at work, economic, situation, health, personal preferences, interests, reliability, behaviour, locations or movement. The above mentioned examples will thus fall under mentioned provisions.

Article 21(2,3,4): The right to object give the data subject the right to object to processing of personal data, including profiling, for direct marketing purposes and should be informed about such processing latest at the time of first communication.[49] In other words, the GDPR inhibits customer profiling for target marketing purposes. However, pursuant to *Article 21(6)* the utility will be able to use customer profiling for public interest, which may include market research according to the WP29 [60]. Based on the above it is assumed that the GDPR will become a barrier to profiling for target marketing, but utilities will be able carry out market research based on these profiles in the name of public interest. Furthermore, it can be discussed upon if targeting customers for direct saving advice will be in public interest, while targeting customers for selling new products and services is not.

Hence, it remains unsure whether target marketing may be enabled without customers consent. As long as customers allow it, it will be enabled.

These energy personalities may be used for a multitude of purposes additionally to the above. Profiling play a central part in demand side management such as tailored feedback programs, but may also be used for utilities to monetize such as providing profiles to marketers or decision makers such as insurance companies. However, demographics have stereotypes and prejudices associated with them and the energy personalities are arguably relatable to stereotypes in society. Hence, the use of consumer profiling is not suitable according to the *lawfulness, fairness and transparency* principle. *Article 22* will also give the customers the right not to be processed in this manner. Direct monetization on customers individual profiles can therefore be concluded to be heavily regulated under the GDPR

Customer profiling is on of the most discussed upon applications of analytics in smart metering and big data context, much because of its wide array of usage areas, which may be considered intrusive. The difference to the level of detail and sensitivity in the examples above are clear, where the latter scenario has become a privacy concern,[11, 91] and consequently may be seen as a barrier to obtaining consent all together. However, the GDPR prohibits the release of customer profiles for reuse to third parties, such as insurance companies, and is in the sense an enabler of customer trust.

However, the monetization of consumer profiles is considered one of the big opportunities presented by smart metering, but it will be hard to prove compatibility with the purpose of collection. According to the WP29[60] the data can be released for reuse if safeguarded appropriately with the DPIA. An assumption is made that that the level of sensitivity inflicted by the data resolution in the examples above represents thresholds in which it is considered compatible and incompatible to sell. For example, given safeguards such as sufficient aggregation or efficient anonymisation, it will be possible to monetize on consumption profiles at 60 min sampling rate, but not at 1 min sampling rates. The main grounds for this assumption is that consumption profiles, even if anonymous, is considered personally identifiable. And when combined with another dataset (not functionally separated) – as is the case when sold – the risk of re-identification increases. The upper and lower thresholds represent acceptable and unacceptable levels of risk in this regard.

Demand load

Initially the peak loads was described. Demand loads are the aggregated form of peak loads under a transformer. Flattening these loads are a critical requirement in the grid and is therefore of uttermost importance in serving public interest. Customers aggregated load profiles can be cross-referenced with demand loads to single out households who contributes for better or for worse during peak hours. [84]

The need to reveal load peaks is to know when the strain on the grid is at its highest and thus knowing when to implement strategies for a demand response. Different demand response program utilizes different strategies such as feedback, time-of-use pricing, incentives or load control. For further reading on the topic see appendixC. The instant benefit is reduction of load peaks, thus reducing need for future grid investments, lower electricity prices and reducing potential blackouts in the grid. A more energy savvy consumer will be able to shift their consumption away from load peaks benefiting both the utility and themselves.

The processing to reveal a demand load is therefore assumed not be inhibited by the GDPR in any particular way. The most prominent reason being that it uses aggregated consumption and is therefore not reliant on individually identifiable data. Cross-referencing with under performing households is a common practice, which can be used for targeting those specific households with saving advice and may not comply due to reasons mentioned in the previous section. However, it will arguably serve the public best interest to analyze trends in the demographic that correlates to those under performing households. The impact of the GDPR on this particular transformation is considered negligible.

Load forecasting

Load forecasting has always been important for utilities as accurate load forecasting result in economic, reliable and secure operation and planning of the power system.[92] With the development of the smart grid, however, and in particular the introduction of smart meters, load forecasting has become one of the most valuable analytics applications. The availability of time-interval data, opposed to traditional monthly reading, has made load forecasting more accurate and possible within smaller forecasting horizons. Traditionally

forecasting has been used for long term planning,[84] but the sampling rate enabled by smart meters allow for forecasting over much shorter horizons at higher levels of detail. [93]

Forecasting in different horizons and aggregation allow for a variety of applications ranging from modelling for market electricity prices[84]; creating automatic load operation schedules for household appliances[94]; short-term load forecast for individual households[93];as well as for microgrids[95]. Modelling electricity prices and scheduling household appliances represents two different sides in terms of intrusiveness.

Analytics for load forecasts incorporates different sources of information such as weather data and seasonal characteristics. One approach[93] is to model the residential load using a weather component and a lifestyle component. The lifestyle component is dependent on individual consumption patterns at appliance level. The weather component model how heating, ventilation and air-conditioning is affected by the weather conditions. The interrelation between forecasting and customer profiling is therefore present at appliance level. It can therefore be concluded that load forecasting has the potential to become intrusive at appliance level.

What poses a potential dilemma for the data protective authority in this regard is to consider whether to allow utilities to process at appliance level for forecasting purposes without the consent of the customer. The aggregation of load forecast at appliance level from a considerable portion of the population would probably provide huge opportunities for utilities. On one hand, the utility would be able to use this data to enhance their algorithms, resulting in enhanced operations or new and enhanced products such as a home energy management system (HEMS), that optimizes the use of the weather components in the algorithm mentioned in the previous.

Some[9, 96] argue that forecasting at residential level could reveal or predict behavior irregularities. This could be used to detect fraud and energy theft. This is important for many utilities but one can ask how much privacy across the entire customer database must be sacrificed in order to capture a couple of offences when this requires higher resolutions. This is addressed by *Article 10* and exceeds the scope of this analysis as it requires a deeper understanding at juridic level.

Optimization models

Utilities' and customers' main objective is optimize energy consumption at the demand side to to reduce overall energy expenditure and to shift consumption away from peak loads. Benefiting customers through reduced billings and the system with stability.

The following presents two different approaches to demand side optimization. One for optimizing appliance use in smart homes and one for scheduling residential power generation for optimizing such services.

The previous section presented an algorithm for forecasting loads modelling the residential load with a lifestyle and a weather component. The lifestyle component can be changed using feedback, making consumers more energy savvy, or through incentives for behavioral change mainly focused on reducing peak demand. This may for example result in deliberately washing clothes outside of peak hours, however some washing machines and dryers have become smart, but that is besides the point. There are, nevertheless, home appliances that is not that easily managed such as those under the weather component such as heating, ventilation and air condition, but also electric vehicles (EV) which is becoming increasingly common.

Whereas demand response may be effective in changing consumer behavior, most residential customers are neither proactive enough nor have the time to perform demand response 24 hours a day.[97] Thus, automated demand response becomes the solution.

One algorithm[97] for HEMS has the ability to control prioritized appliances and restrict total power consumption of the household to always stay below a certain limit, being a load or price ceiling, while considering customer preferences in a prioritized manner. If heating is more important than charging the car the charging will subside until consumption is below the limit or the demand response event is over.

Such a system can be further improved by optimizing the use of DER.[98] Distributed generation, (DG) such as microgrid,[99] and energy storage are common examples of DER. In fact, due to increasing DG and consumption there is a risk for overgeneration in load valleys and increasing consumption during load peak.[100] Hence, the optimization of DER will enable end consumers to store energy during valleys and use it during peaks, resulting in an overall flattening of the load curve. In[98] Pedrasa et al. provides such a solution with

an algorithm scheduling residential DER into a HEMS. Such a system will, in other words, provide the optimized use of energy storage, renewable energy and fossil fuel to power a home whose major appliances respond to price signals and comfort of the residents.

Such algorithms are assumed to require 1 min sampling rate as they operate at appliance level. A consumer opting for a HEMS is most likely more energy savvy than risk averse and will probably not have any issues with consenting as long as service is delivered. However, as mentioned, a high resolution is important in the realization of DER for households. In order to realize this it is assumed that utilities need sufficient amounts of unbiased data to train their algorithms, that is, data representing the population and not a sample of high income households, who is more likely to opt for optimized models in initial phases of implementation. It can therefore be argued that the realization of an integrated DER to the HEMS, in the short-term, is limited by the default privacy setting.

4.3.3 Data curation

Smart meter data will have to be curated in order to maintain its value. This means for example that the database must be updated everytime a household gets new residents. This is addressed by the accuracy principle which states that data should be "where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay"

As smart meter data gets fed into algorithms the output requires high quality data to avoid a "garbage in barbage out" scenario. By requiring utilities to ensure *accuracy* of data the GDPR puts an additional incentive for utilities to enhances the quality of the output and will ensure that the value extracted from mining and other means will increase over time rather than diminish.

Also it must be rest assured that all data in the database are sufficiently anonymous in case of data breach. One particular concern in case of data breach is the ability to foresee occupancy and thus plan for robbery.[11]

4.3.4 Data storage

Data storage is a key enabler for advanced analytics and in particular for traditional non-IT-based sectors such as energy.[1]

Utilities want to keep data for as long as possible as they want to mine consumer data for new insights, thus realizing the value of data generated over time and reveal new usage areas. This may become a challenge to the storage limitation principle. It requires that data should not be stored in a "a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed". As consumption data is considered personally identifiable data, even if it is anonymized,[101] data mining may become a challenge under the GDPR.

However, the provision also states that personal data "can be stored for longer insofar the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes."[49] The regulation furthermore refers to *Article 89(1)* that provides certain requirements for safeguards when processing in public interest. The wording of the GDPR on this subject is ambiguous, but it stands to reason that it may or may not be possible depending on who processes the data. The author believes the wording is intentional to make mining for public interest easier while restricting mining for solely own benefit.

4.3.5 Data usage

So far this section has described within the big data value chain framework how value is created throughout. The main finding is that: data resolution is key to drive value from smart meters. Higher sampling rate gives richer information and ability to reveal new usage areas. Also higher resolution enables higher levels of aptitude, exemplified with DER. Because privacy by default requires sampling rate at 60 min as default this inhibits utilities in utilizing the full power of smart metering. However, as the GDPR don't restrict further processing for public interest a key question becomes whether utilities will be allowed to sample at higher resolutions regardless of consent.

A proposed model was to allow utilities to sample at required resolutions to process in

public interest, while processing at agreed upon resolution for the purposes in the consent. If not the utilities must see to other means to obtain consent. The use of feedback programs may be a good channel for this. Providing an app to the end consumer could even be sufficient.

Other key findings is summarized below:

- Customer profiles can be synthesized through analytics, but restricted to archetypes(60 min) rather than stereotypes(1 min) by the GDPR. This is considered a trust builder
- Profiles can not be used for direct marketing purposes but can be used for market research without consent.
- Without consent monetization on customer profiling will be heavily regulated and a threshold in this regard was suggested. At 60 min one can monetize, while at 1 min one can not.
- GDPR ensures data value is maintained over time. Thus, also allowing mining to generate more value over time.
- Consumption data can be stored for longer after initial purpose is achieved for the sake of public interest. Which includes market research.

The following section consider the above key findings in terms of three main ways in which big data analytics can be used. The previous stages of this analysis has touched upon them with examples of use that has been assessed in terms of how the GDPR will affect them:

- applied analytics was described in relation to demand side management
- operationalized analytics was described in relation to target marketing
- monetized analytics was described in relation to selling data to third parties

This is furthermore the limitation to scope of the following. The different use areas represents, in their respective order, customer expectation to products and services delivered as a trade-off to privacy risk exposure. This is illustrated in figure 4.9.

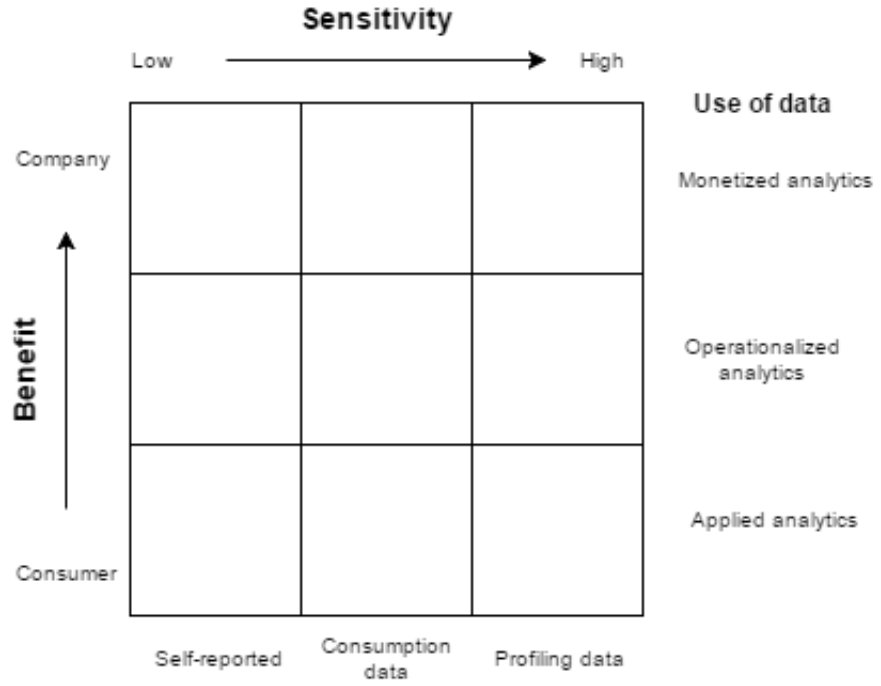

Figure 4.9: Figure showing how customers demand more in return for providing more sensitive data (Adopted from [102]

At a lower level utilities and customers alike benefit from the processing of personal data, but as the utility wants to diverge from the original purpose a conflict of interest emerges, where companies ultimately want to monetize on the expense of the privacy of consumers. The general notion is that the more the company want to use data for their own benefit the harder it becomes to justify further processing and is more likely to need fresh consent.

The previous sections has described how value is driven through collection, transformation and mining as well as maintained and secured thorough curation and storage. This particular part of the value chain describes the value that is realized through use.

Demand side management

Demand side management is an important element of utilities planning approach and the introduction of smart meters has created new opportunities in customer engagement.[72] Whereas household energy consumption increases the price of producing and delivering electricity becomes a growing concern for utilities. Carbon emissions are top agenda in European policy context and there is little doubt that changing household consumption behaviour is of great great interest for society as a whole. Additionally, in deregulated,

and consequently more competitive markets, customer trust and reduction of churn has become an imperative for staying competitive.

Smart meters enable new applications provided to the end consumer that educate, enable greater engagement and make home energy management smart. The following depicts data intensive applications enabled and improved by customer profiling. Appendix C depicts feedback systems, dynamic pricing and demand side automation such as HEMS, while Appendix B depicts customer behavior. These two reviews are the foundation for an assessment of the applications in the following to determine their value. The key take away from the review in the context of this analysis is described.

Feedback to customers is considered an important part of educating the customer to become more energy savvy, and shows how richer and more timely information enables customers to better change their behavior through feedback programs. Key findings are:

- Information technology such as IHD and smartphone apps empowers consumers to change their behavior through education which is closely related to the self-determination theory. It is furthermore considered more efficient than incentives such as punishment.

- Tailored feedback enable consumers to better understand their own behavior. At lower resolutions consumers can compare them selves with neighbours and see how consumption corresponds to demand load. Also billing forecasts are enabled. Higher resolutions enable suggestions to behavior change at appliance level.

- The closer to real-time feedback is provided the easier it is for customers to respond timely to changing demand load and pricing signals.

Utilities use load forecasts based on consumer profiles to create pricing models that reflects the cost of producing and providing energy more accurately. This is previously referred to as dynamic pricing. Different pricing models fit different customer profiles. This increases the efficiency of feedback programs and demand response. However, demand response programs has their drawbacks as shifting load may ultimately create an increase in load valleys. They are efficient in reducing stress in the grid and market price during peak periods, but may not lead to reduced energy consumption.

Pricing models such as real-time-pricing enable more advanced applications for the end-user such as automation. Automation in a HEMS use pricing signals to optimize energy consumption from appliances such as heating, ventilation and air-conditioning accounting for occupant hours, comfort preferences and price sensitivity. Similarly to dynamic pricing, automation reduced consumption during peak hours. It will save the customer on the electricity bill, but may actually increase the overall consumption due to consumers becoming more wreckless.

The above show that different strategies serves different purposes. However, the ultimate value delivered is different and feedback programs are arguably the best long term solution. Key aspects are summarized below.

- Feedback programs reduce baseline consumption and enable demand response, however limited to customer attitude

- Dynamic pricing provide customers suitable pricing programs and an incentive to reduce peak loads, but may not result in overall reduction.

- Demand-side automation will save money and stress in the grid but can increase overall consumption

An interesting take away from the study was that it somewhat contradicts the theory, that value is increased with transformation. Where automation does not yield a higher saving directly. Studies show that implementing feedback before implementing demand-side automation increases savings more that jumping straight to automation.[103] It does not debunk the theory it just show the synergic effect of transformation.

Based on the above it can be argued that the foundation of demand side management should be built on educating the consumer through feedback programs and the most value is realized accordingly. This substantiates the proposed model for utilities to use feedback programs as a channel to obtain higher resolutions. Which furthermore will be an enabler of dynamic pricing and demand side automation and their efficiency as such.

Target marketing

Behavior analytics can be used to optimize decision making and to drive bottom- and top-line revenue[17]. A better understanding of how consumers respond to different demand side strategies will help utilities more effectively develop new products and services to satisfy their customers and increase revenue from selling them.

Detailed descriptions of each customer and their consumption patterns enable utilities to target customers with tailored offers based on their energy personality. A model can, for example, be built to predict customers who are good targets for up-selling when they call into a call center. For instance, a customer with a high baseline usage will probably benefit from a different pricing tariff than a customer with high peaks at morning and evening. Furthermore with IT supported applications utilities can cost-efficiently market their products and services through various channels such as in-home displays and smartphone apps. Additionally, as the market for HEMS is growing the practice of promoting physical products to the home will arguably become more prominent.

Several types of predictive analytics and forecasting applications are based on smart meter data. By analysing households load curves an algorithm can predict customers' heating hardware based on smart meter data alone. Utilities can in the same way estimate how buildings are setting their thermostats. Homes that has supposedly inefficient setpoints may be offered personalized saving advice, through for example in-home displays, or could be targeted for demand response programs as they contribute to higher peak loads.[86]

Target marketing is, however, perceived as more intrusive and customers may become more reluctant to allow for their personal information to be processed for such purposes.

As the key findings state: Profiles can not be used for target marketing without an explicit consent. Many use areas within target marketing will however as discussed fall between what is in public interest, and should be considered. As for the call center example this is a good way to communicate and build trust with the consumer, whereas marketing products to a HEMS on the other hand may prove difficult under the GDPR. It will most likely depend on the value proposition of the utility to whether target marketing will be enabled or not.

Monetization

Datasets generated for one purpose may prove extremely valuable for companies in completely different sectors and they are willing to pay for it – and smart meter data is no different. Companies generating a rich pool of raw data can sell it with little investment or leverage the unique data into for example customer profiles to conduct high value transactions.[36]. Some say that mining smart meter data is like mining for gold.[91] If that prediction were to hold monetizing smart meter data could very well become the new "cash cow"[104] for utilities.

In the emerging market of business analytics companies are now offering utilities to create customer profiles and provide a direct link to appropriate third-party organisations so they can monetize.[105] Table 4.3 provide a list of relevant third parties who could be interested.

Table 4.3: Third parties interested in consumption data(Adopted from[81]

Interested parties	Purpose
Insurance companies	Determine health care premiums based on behaviors indicating illness
Marketers	Target advertising
Creditors	Determining behavior that might indicate creditworthiness
Law enfrocements	Identification of suspicious or illegal activity

Whereas this may become a big money opportunity, customer consent will be additionally hard to obtain.[102] For example allowing insurance companies to decide premium based on indications that the customer may have health issues is not in the interest of the customer, nor the GDPR, where this was concluded as impossible under the GDPR in 4.3.2. However, including human intervention in the decision making process may make such a decision justified.

One of the key findings from the previous finds that monetizing on personal data related to an individual will be heavily regulated under the GDPR and is assumed impossible without a consent. Furthermore a consent for a new purpose at this level would entail more juridic baggage than, for instance, marketing as there is a third party involved as well. This may eventually put an end to such transactions. However, utilities will under the assumptions be able to monetize on market research from data at a 60 min sampling rate by releasing it for reuse.

4.4 Summary and concluding remarks

Smart meter data in its raw form has little value besides correct billing purposes. The ultimate value of smart meter data, as with all big data, must be seen in relation to all the possible ways it can be employed in the future. The first determining factor in this regard is the resolution of consumption data. The higher the resolution, the more data can be mined for new knowledge, and the more advanced and robust the analytics and predictions become. Pursuant to this notion, the GDPR imposes one particular barrier trough *privacy by design and by default*, which is the first main finding:

> ***Privacy by default*** **restricts the value potential of analytics to be realized.**

A model was proposed where utilities are allowed to sample at required resolutions for processing in public interest but only processing at agreed upon resolutions according to the purposes of the consent. However this is not specified as allowed in the GDPR

Because further processing in public interest seemingly won't become a restriction under the GDPR; it will not become a barrier for utilities to develop products and services supporting demand side management programs and the smart grid. An example was given that a successful integration of EV-charging and solar panels as DER may be dependent on higher resolutions than provided by default. This underpins the magnitude of barrier 1.

For utilities to stay competitive they need to attract new customers, reduce churn and deliver electricity at lowest possible cost. Respectively this can be achieved by offering better products, better services and suitable pricing programs. However, where target marketing is a solution to this manner, the GDPR create a barrier through the *right to object* in particular. This is the second main finding:

> ***The right to object*** **restricts the opportunity to drive top-and bottom line revenue**

Target marketing becomes a grey area under the GDPR, where customers are able to object to processing for that purpose. Depending on how the GDPR gets enforced utilities may be allowed to target customers with consultation on energy behavior and pricing programs through suitable channels such as customer support. A possible approach is for utilities to use their market research to target new customers. However, target marketing is only a

consent away.

Smart meters enable utilities to become an integrated part of the data economy; thus, also enabling them generate revenue through the same means as the technology giants. Namely, selling personal data to third parties, which is the idea behind comparisons between data mining with gold mining. However, monetizing on personally identifiable data will be heavily regulated, mainly through the *consent* and *purpose limitation* principle. As a consent can not be obtained on a purpose not yet discovered these provisions largely remove the potential to monetize on individuals consumption data. This is the last main finding:

> **The *consent* and *purpose limitation* principle restricts the opportunity to monetize**

The WP29 suggests that personal data can be released for reuse given appropriate safeguards, but the big money arguably lies in monetization on data concerning particular individuals. However, with the right value proposition maybe customers will allow for utilities to sell their information, yet this entails a lot of juridic baggage and may therefore set a final stop to this practice. Whereas this means lost opportunities, it also means a big win for customer trust that may eventually pay back in increased willingness to provide more personal data.

The three main findings identified through this analysis are the main focus points regarding the case in the remainder of the thesis. They are also presented as additions to the influence diagram from the previous analysis. The following chapter summarizes all key findings and presents them as a basis for the discussion.

Chapter 5

Summary of Findings

This chapter presents the finding of the carried out analyses in their respective order, with comments where necessary. The preliminary analysis identified the value drivers of big data. The analysis of the influences of GDPR on characteristics of big data resulted in a influence diagram of relevant provisions on the value drivers. The case study identifies three more influences, hence a updated influence diagram is presented. Concluding this chapter an assessment is made and a set of main focus points is presented as assertions to be discussed.

Big data value chain theories

Table 5.1 presents a revised version of the theories derived from the preliminary analysis. The theories in *italic* are identified as miscellaneous and will not be directly influenced by the GDPR in terms of contribution to the ultimate value of data, in the sense it is understood in this thesis. They will however be affected positively:

- *accuracy* and *minimisation* will influence the curation positively, and in the sense that data is like corporate assets the GDPR aids in their maintenance.
- The GDPR requires data storage to reflect state of the art technology and risk exposure. In the sense that secure storage is like secure banking, the GDPR will undoubtedly influence this in a positive way.

Table 5.1: Theories about value drivers in the big data value chain

Stage	**Theory**
Data acquisition	The more data collected, the more value can be extracted from it The closer to real-time data is processed, the more value it provides for the initial purpose
Data analysis	The higher level of transformation achieved the higher potential use value
Data curation	Better curation maintains value over time and make algorithms more robust *Curation is like maintaining corporate assets*
Data storage	The longer data is stored the more value can be extracted from it *Secure storage is like secure banking*
Data usage	The ultimate value of data is the sum of all the ways in which it can be employed and the respective value output of each use

Particular challenges of the GDPR

- *Article 22* restricts processing of personal data that may yield a discriminating output. As all consumption data potentially becomes personal this may render datasets and algorithms worthless.

- *Article 12(1) - 15* require algorithmic transparency which becomes more challenging with more powerful algorithms. This may create a trade-off between harnessing the power off algorithms and compliance

- *Storage limitation* restricts companies from storing information long enough to maximize the value of the data set.

- *Data minimisation* restricts companies from collecting data that may prove valuable in the discovery of new use.

- *Purpose limitation* restricts companies further processing because one can not obtain a consent to a use that is not yet discovered. Additionally it may restrict companies ability to achieve higher levels of transformation. Seen as a big challenge to big data analytics

- *Privacy by default* ensures highest possible privacy setting and will restrain, from the outset, collection of data. Which, in turn, has significance for value driven through the value chain, regardless of purpose.

Particular positive effects of the GDPR

- *Accuracy* and *data minimization* require companies to improve their curation process which, in turn, will improve the analytics output, maintenance and ability to reuse data.

- The GDPR will on a general basis increase customer trust, which in turn has positive effects on willingness to share information, data quality, and obtaining consent for further processing.

Other impacts and implications

- *Anonymisation* and *aggregation* becomes essential for companies to retain their data and secure new use, thus fundamental under the GDPR to maximize the value of data

- *Privacy by design* and DPIA provide a "toolbox" of techniques to stay compliant and may prove as valuable competencies to gain customer trust

Big data value influences

Figure 5.1 illustrates the value drivers of big data and what provisions from the GDPR that influences the ultimate value. Dashed lines are negative influences and solid lines are positive influences.

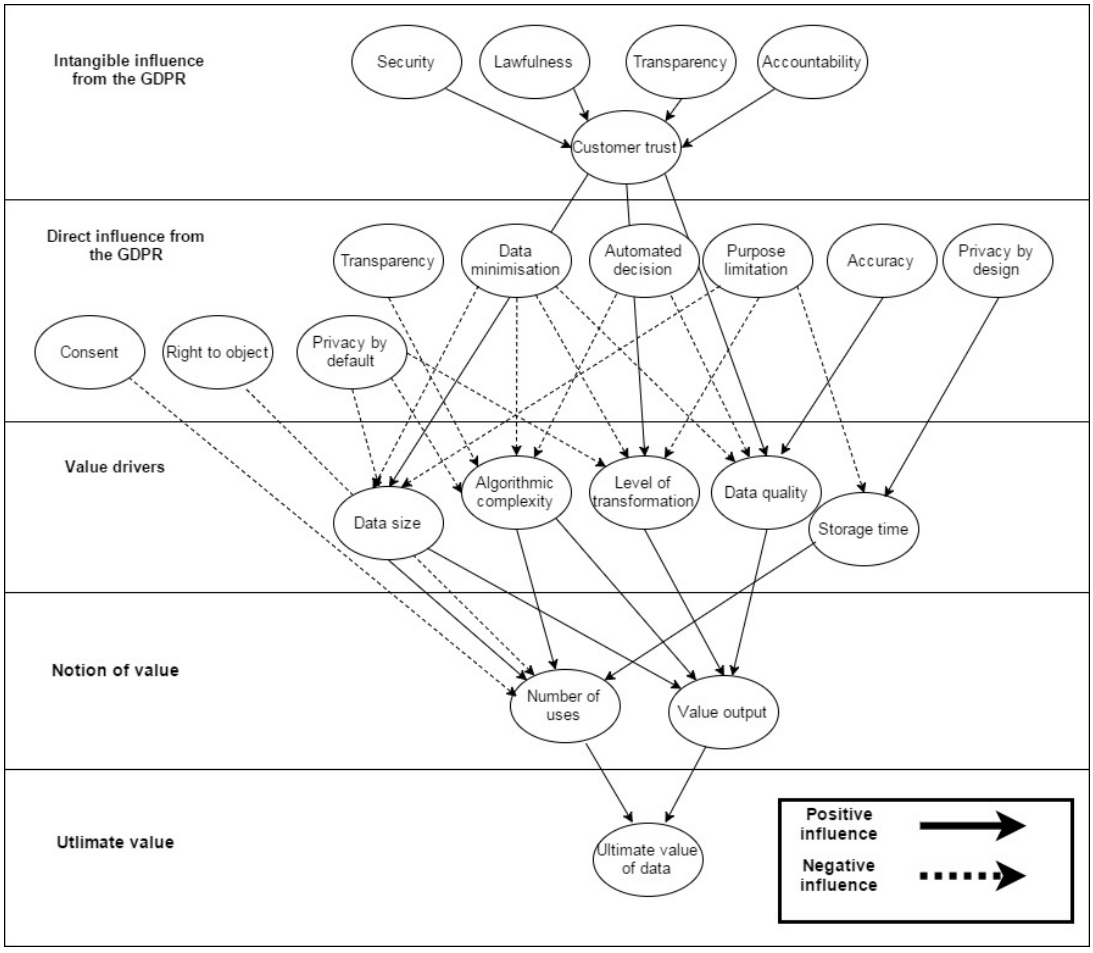

Figure 5.1: Illustration of value drivers in big data and influences from the GDPR

A surprising finding in the case study was the influence of privacy by default. The initial analysis failed to identify privacy by default as an influence, which turned out to potentially have a big impact on the ultimate value of smart meter data. Two other additional influences was found in particular, which restrain number of use areas. These were: The *right to object* restricting ability to drive top- and bottom line revenue from profiling and

the consent and *purpose limitation* restricting opportunity to monetize on new purpose. Whereas the latter two becomes rather speculative, the former raise some important questions regarding how privacy by default is in public interest or not.

The key findings of the analysis can be summarized as five assertions.

- *Privacy by default* is a barrier to drive value throughout the big data value chain.
- *Fairness, transparency* and the expanded definition of *personal data* challenges companies ability harness the full power of algorithmic decision making.
- *Consent, purpose limitation* and *storage limitation* are barriers to discover new use of data.
- The GDPR is a trust builder and an enabler of value in the long-term
- The GDPR aims to strike a balance between value creation for "the greater good" and monetization on the expense of privacy.

This page has been left intentionally blank.

Chapter 6

Discussion

The case study found three new negative influences to the value of data that was not yet discovered prior to the analysis. The most prominent was *privacy by default*, as it was also not expected. The finding raised an important question to whether it is, at least under given assumptions, expedient with privacy by default on consumption devices such as smart meters. This is the starting point of the the discussion. The discussion provides strong and banal examples to amplify the points being made, but are not absolutes. Based on the discussed a forecast is made.

The case showed that privacy by default, by requiring the highest privacy setting on smart meters, will curtail, from the outset, utilities opportunity to realize value in public interest, and in this particular case sacrificing "green" initiatives. A suggestion was made to allow for extensive collection for processing in public interest, but this is not addressed explicitly in then GDPR, and seems to be a omitted aspect. This may be a special case for smart metering, but it can be argued that this may translate to all customer products that has *extensibility by design*, referred to in 4.3.5, such as health and location data from smart phones and smart watches, which is valuable data for, among other things, health research and social science. This may therefore become a major barrier for realizing the ultimate value of data in the name of public interest.

However, privacy by default serve as privacy insurance from the outset. This was demonstrated through its contribution in 4.3.2. With a higher resolution more stereotypical profiles can be derived from the data, where in this case highest privacy setting provide a

safeguard against discrimination. Because society are discriminating, data collected from behavior in society will be as well. Thus, a higher sampling rate yield a higher chance of algorithms to make discriminating decisions. However, utilities are not a decision maker, that can provide "legal effects or significantly effect an individual", but consumption data used in decision making is nevertheless of big interest. Insurance companies, for instance, are interested in such information at individual level. This has been a concern of late, but is prohibited under *Article 22*. This type of use has been predicted as one of the big monetization opportunities and the future "cash cow" of utilities,[101] but gets restricted under the new regime. In this case the GDPR serve as a barrier of value, but more so an important trust builder.

The example also demonstrate how intertwined the provisions are. Where privacy by default easily can be omitted through an opt-in consent: *Article 22* secure that discrimination are avoided; *data minimisation* ensure excessive amount of smart meter data are not collected; and *purpose limitation* ensure that it is not misused. Furthermore, the highlighted provisions are mutually reinforcing, which can be said in general about the provisions in the GDPR, and is yet another reason to have trust in the Regulation.

Innovative utilities and insurance companies may still be able to provide valuable transactions of customer profiles or other smart meter generated data. Cyber-insurance has become one of the hot topics of late, which underpins the monetary value of data additionally. For instance, a consumption profile may reveal traits that relates to computer skills, where a more computer savvy profile probably would get a discounted insurance premium. This would still not be compliant with the GDPR, but will arguably be compliant when applied at corporate level to help determine the risk premium. For example, looking at an aggregate consumption of all employees. In this case, the value of the information would increase with higher resolutions, enabling not only behavior patterns and stereotypes to be decisive factors, but for instance amount of time used on computer, and even brand of computer. This could reveal correlations with trends in the population, which in turn can be used to classify a companies risk exposure. Such a service could prove extremely valuable and can arguably be carried out *lawfully*. Hence, it come to show that privacy by default also restricts companies ability to monetize, although further processing is justified.

Privacy by default will undoubtedly have an immediate impact on the value of data, both for uses in public interest and for uses with reasonably ethical purposes. It can therefore be

argued to whether exemptions should be made for excessive processing in public interest. There is one possibly major drawback to this though; allowing for wiggle room may result in mistrust to those collecting and consequently counteract the sense of insurance and trust with default highest privacy setting. Companies should therefore aim to obtain such levels of detail by other means. In the case study it was suggested that utilities find innovative ways to obtain higher resolutions of consumption data. One proposed solution was to provide a user friendly smartphone app providing timely detailed feedback. Processing and analysing this amount of data definitely would become a challenge for the processor, but it is a possible future solution nevertheless.

That said, customers have a history of trading privacy for little monetary reward,[101] which give reason to believe that customers are willing to consent to just about anything as long as an appropriate value proposition is in place. In this regard an interesting development has occured in Norway, where a utility and a telecommunications company (telecom), namely Fjordkraft and Telenor, has entered a partnership. The utility offer a mobile subscription delivered by the telecom at a discounted price given that the customer subscribes to the energy services.[106] The telecom receive additional revenue from telecommunication equipment and the utility from a growing customer base, while both get the mutual benefit of consumers data. Furthermore, what makes this development especially intriguing is that, prior to the partnership, Telenor bought the company, Tapad whom has has developed an algorithm that identify individuals between the devices they use for hyper-targeted marketing.[107] The level of detail on consumer profiles when combining smart meter data, smartphone data and an algorithm that identifies individuals is unimaginable – and so is the potential privacy concern. This shows how utilities may be moving from being energy providers to service providers.

Consequently to utilities changing their business models to collect more data, some believe that data will become more valuable than the commodity that's being consumed to generate the data.[91] According to the theories arrived at in 2.4 this statement would hold, and particularly the theory that "the more data collected, the more value it can be extracted from it". In such a scenario free electricity is a tempting thought, but would arguably be counterproductive as the consumption would skyrocket. A more viable business model could be free electricity under certain thresholds in the grid. However, this is rather speculative, but illustrates a point nevertheless. If utilities were to provide value propositions

that is "to good to refuse", a scenario such as the one portrayed in the "oxymoron of big data and privacy",(2.5.1) would be relatable. Because the current legislations are obsolete; technology is developing at a fast pace; trends in enforcement of privacy protection are distressing; and the cyber threat is increasing, it can be argued that current state of affairs are not viable. In turn, this give reason to believe that a future diminished data value is present. Thus emphasizing the importance of providing a regulatory framework that require state or the art technical and organisational measures to be implemented appropriate to the risk present. This is provided by the GDPR and particularly reinforced by *privacy by design*, which require a proactive approach, ensuring that companies must keep up with technology trends to stay compliant.[49]

The above illustrates the GDPR as a trust builder, which will emerge as a long-term enabler of value as customers will gain, or at least, maintain trust in controllers and processors. The intangible influences illustrated in figure 5.1 will therefore have substantial impact after the Regulation becomes legally effective. One of which was briefly touched upon in the case, but failed to illustrate; namely, *fairness* of processing, which is particularly enforced through *Article 22: Right not to be subjected to automated individual decision-making, including profiling*. One particular challenge in this regard is how the increased scope of the GDPR through an extended definition of *personal data* may deem it impossible to purge datasets for discriminating data or discriminatory correlations. Whereas this may not be the situation with smart meter data, it is definitely relevant for insurance companies and creditors. Another issue presented was risk averse algorithms disfavouring minorities, which provides another fairness issue in decision processes. However, as the use of "solely" in the wording of the provision implies; by adding human intervention it is possible to go around the provision. In an increasingly automated society, nevertheless, adding human intervention in all automated decisions regarding individuals won't be practicable.

Self-driving cars are undeniably the most talked about automation technology out there, and on the road they will make automated decisions continuously that will significantly affect humans. The reaction time of a passenger will arguably not be sufficient to claim human intervention. This is not in breach of the GDPR, but in case of an accident, how will algorithms make decisions? One scenario is to chose between different cars whose brands may be related to stereotypes in society. Another is to choose between two individuals, which will require a whole other level of processing *sensitive information*. An individual on

the road or strolling won't have the opportunity to neither object or exercise any right not to be made an automated decision about. Another aspect to this is to provide *transparency* to the decision making process, where in this case a rationale must be provided to why the car chose one person over the other. Self driving cars whose programming is extremely complex will, due to "black boxes" in deep neural networks, prove challenging if not impossible to decipher. If so, it's a little paradoxical considering that the EU are promoting self-driving cars as a flagship initiatives enabled under the GDPR.[61] Based on this, it can be argued that *Article 22* and the *transparency* may become barriers to the actualization of automation technology in everyday life.

"Black boxes" are addressed as one of the main barriers to transparency in algorithmic decision making. However, there exists algorithm that provide transparency, but the development lags behind the increasing complexity in processing, nevertheless.[65] In order to stay compliant to the GDPR it may become a scenario that companies won't be able to harness the full potential of their algorithms due to lacking transparency. Although the GDPR's purpose is to not inhibit innovation it can be argued, under these particular circumstances, that it does. Another important aspect to this may be to consider companies whose business rely heavily on their algorithms and the confidentiality of them. This may present two particular challenges. Firstly, *Article 22* may make their algorithms worthless if ensuring fairness proves impossible. Secondly, companies may conceal relevant information about their algorithms in the name of competitive advantage. However, it remains to be seen due to uncertainties revolving enforcement once the GDPR becomes legally effective.

However, fairness and securing that personal data won't be misused in ways that can harm individuals is an important stepping stone for building customer trust.[108] Furthermore, increased transparency, the knowing that means are taken to ensure security of personal data, and that those handling it will be held accountable, amplifies the trust building impact of the GDPR. In section4.3.4 a theory was derived, that ambiguous wording of certain provisions, and in that case *storage limitation* and *Article 89(1)*, are intentional to strike a balance between what can be seen as further processing for "the greater good" and for exclusively the benefit of the company. Hence, a perception of not being exploited will arguably add to the trust building capabilities of the GDPR.

The third key finding in the analysis was the impacts of *consent* , *purpose limitation* and *storage limitation*. One of the main value drivers of big data is to find new ways to reuse

data. This is directly curtailed by these provisions. It was argued that a company can not specify a purpose that is not yet discovered, whereas one of the key features of big data analytics is to discover new use. textitStorage limitation amplifies the effect by limiting the time, and thus the probability to find new uses. What is especially distressing in this regard is that if the purpose is not known, one can not obtain consent for it, even if in public interest. This is furthermore a potential contradiction to the Digital Single Market Strategy for Europe, which prioritize big data as a value driver.[8] It is therefore of paramount importance to gain the customer trust that enable fresh consents. From the EU part, resources are made available to aid companies in the transition[61] as well as compliance and trust building tools such as DPIA, DPO and Privacy by design. From the companies part, they must ensure compliance and communicate it as such. Equally important are offering value propositions justifying the risk exposure of providing more personal information. As privacy risk exposure decreases with compliance the synergy effect of GDPR and innovation, enabling unique value propositions, will hopefully surpass the limitation presented by mentioned provisions, yet this remains to be seen.

Chapter 7

Conclusion

The research in this thesis had its goal to answer the question: *How will the General Data Protection Regulation impact the value of data?* Subsequently a perception of data value was derived, and addressed as the *notion behind the value of data*. This is a general understanding of the value of data stating that the value of data must be seen as the all the possible ways it can be used in the future and the respective value output from each use. This furthermore created two sub questions to be answered: How can one maximize number of uses? And how can one achieve the highest possible value potential for an individual use case? To answer those questions a preliminary analysis was carried out in the big data value chain. Section 2.5 identified five main value drivers: data size, algorithmic complexity, level of transformation, data quality and storage time.

Consecutively an analysis of the GDPR was carried out with respect to four key characteristics of big data that is particularly challenging to privacy and the rights and freedoms of individuals: unfairness and discrimination, opacity of processing, tendency to collect all data and finding new purpose. Section 3.5 presented the results in an influence diagram showing GDPR influences on the identified value drivers.

Following the analysis, a case study on smart meter data was carried out within the the big data value chain framework, where previous findings was exemplified. The impact of the GDPR was assessed throughout, and three additional influences was identified, where the impact of *privacy by default* proved especially prominent. The final influence diagram was presented in 5. From the results of the analysis five assertions was derived and highlighted

in the discussion:

- *Privacy by default* is a barrier to drive value throughout the big data value chain;
- *Fairness, transparency* and the expanded definition of *personal data* challenges companies ability harness the full power of algorithmic decision making;
- *Consent, purpose limitation* and *storage limitation* are barriers to discover new use of data;
- The GDPR aims to strike a balance between value creation for "the greater good" and monetization on the expense of privacy;
- The GDPR is a trust builder and an enabler of value in the long-term

From these it becomes evident that the GDPR, when legally effective, will have an immediate impact on the value driven through the value chain. Particularly noteworthy is the compound impact that *privacy by default, consent, purpose limitation* and *storage limitation* will have on the value created through the value chain both for public interest and for purposes that is in other ways legitimate.

The manifestation of *fairness, transparency* and the expanded definition of *personal data* through enforcement remains to be seen, but will be a challenge nevertheless, and may in particular become a challenge to the actualization of automation in everyday life, where human intervention is not feasible. Whereas these points may oppose the GDPR, the current state of affairs are not viable, and a change must happen. With personal data increasingly dispersed in a digital format, and companies innovating to gather more as such, the scarcity and the integrity of data, and thus its long-term value is at risk of diminishing. The GDPR is a necessary evil, where the intangible impacts of the GDPR will have the most substantial long-term impact. *Transparency, accountability, security* and *lawfulness* will eventually renew customer trust while reducing risk exposure, a synergic effect resulting in:

- more obtainable consents, thus more data transformed and reused;
- more data acquired, thus more data to extract value from;
- more sustained customer relationships, thus more time and aggregated data to mine;

- and more accurate data, thus more valuable output

The GDPR aim to change the power balance between customers and companies handling their personal data. When legally effective this encompasses all data generated by and about individuals. Every company processing personal data will be affected. Initially, capabilities to drive and realize value will be curtailed, forcing a shift from customer exploit to customer satisfaction. Whereas this will impact the short-term value of data in a negative way, the long-term impact will overshadow the negatives and ensure a sustainable data-economy in the future.

This page has been left intentionally blank.

Bibliography

[1] J.M. Cavanillas, E. Curry, and W. Wahlster. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer International Publishing, 2016. ISBN 9783319215686. URL https://books.google.no/books?id=dT3bsgEACAAJ.

[2] Yonego Joris Toonders. Data is the new oil of the digital economy, n.d. URL https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/.

[3] William Schmarzo and Dr. Sidaoui Mouwafac. Applying economic concepts to big data to determine the financial value of the organization's data and analytics research paper. Technical report, University of San Francisco, 2017.

[4] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. An Eamon Dolan book. Houghton Mifflin Harcourt, 2013. ISBN 9780544002692. URL https://books.google.no/books?id=uy4lh-WEhhIC.

[5] Fabernovel. Gafanomics: New economy, new rules, 2014. URL https://innovate.fabernovel.com/work/study-gafanomics-new-economy-new-rules/.

[6] Dennis Hirsch. The glass house effect: Why big data is the new oil, and what to do about it. In *Future of Privacy Forum.< http://www. futureofprivacy. org/wp-content/uploads/Hirsch-Glass-House-Effect1. pdf>. Last accessed September*, volume 13, 2015.

[7] European Commission. Eu data protection reform what benefits for businesses in europe?, 2016. URL hhttp://ec.europa.eu/justice/data-protection/document/factsheets_2016/data-protection-factsheet_01a_en.pdf.

[8] European Commission. A digital single market strategy for europe. Communication COM(2015) 192 final, European Union, 2015.

[9] I. S. Jha, S. Sen, and V. Agarwal. Advanced metering infrastructure analytics; a case study. In *2014 Eighteenth National Power Systems Conference (NPSC)*, pages 1–6, Dec 2014. doi: 10.1109/NPSC.2014.7103882.

[10] EY. Smart energy meters provide"gateway" to the home, 2015. URL https://skyvisionsolutions.files.wordpress.com/2015/12/ey-smart-meter-gold-mine.pdf.

[11] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private memoirs of a smart meter. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, pages 61–66. ACM, 2010.

[12] L.M. Applegate, R.D. Austin, and D.L. Soule. *Corporate Information Strategy and Management: Text and Cases*. McGraw Hill higher education. McGraw-Hill, 2007. ISBN 007-124419-0. URL http://www.hbs.edu/faculty/Pages/item.aspx?num=11565.

[13] Laks Srinivasan. Analysis paralysis — how to turn big data into profits, 2017. URL http://insidebigdata.com/2017/05/03/analysis-paralysis-turn-big-data-profits/.

[14] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[15] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.

[16] Chuck Cartledge. How many vs are there in big data? *n.a*, 2016.

[17] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman. *Big Data For Dummies*. –For dummies. Wiley, 2013. ISBN 9781118644171. URL https://books.google.no/books?id=XPkAEFXo7VgC.

[18] Paul McNamara. How big is a zettabyte?, 2010. URL http://www.techworld.com/storage/how-big-is-a-zettabyte-3222999/.

[19] Gartner. Gartner says 8.4 billion connected "things" will be in use in 2017, up 31 percent from 2016, 2017. URL http://www.gartner.com/newsroom/id/3598917.

[20] Steve Leibson. Ipv6: How many ip addresses can dance on the head of a pin?, 2008. URL http://www.edn.com/electronics-blogs/other/4306822/IPV6-How-Many-IP-Addresses-Can-Dance-on-the-Head-of-a-Pin-.

[21] Neil Beihn. The missing v's in big data: Viability and value, 2013. URL https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/.

[22] George Firican. The 10 vs of big data, 2017. URL https://upside.tdwi.org/Articles/2017/02/08/10-Vs-of-Big-Data.aspx?Page=1.

[23] European Union Agency For Network and Information Security(ENISA). Privacy by design in big data, 2015.

[24] James Kobielus. Measuring the business value of big data, 2013. URL http://www.ibmbigdatahub.com/blog/measuring-business-value-big-data.

[25] Ben Rossi. How to measure the value of big data, 2015. URL http://www.information-age.com/how-measure-value-big-data-123460041/.

[26] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining: Pearson New International Edition*. Pearson Education Limited, 2013. ISBN 9781292038551. URL https://books.google.no/books?id=jC6pBwAAQBAJ.

[27] Liane Colonna. Mo' data, mo' problems? personal data mining and the challenge to the data minimization principle. In *Making Ends Meet hosted by Stanford Law School and The Center for Internet and Society*, 2013.

[28] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

[29] Bernhard Marr. What is the difference between deep learning, machine learning and ai?, 2017. URL https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/#7dad8b0c26cf.

[30] Sam Byford. Google's alphago ai defeats world go number one ke jie, 2017. URL https://www.theverge.com/2017/5/23/15679110/go-alphago-ke-jie-match-google-deepmind-ai-2017.

[31] Michael Lloyd-Williams. Discovering the hidden secrets in your data-the data mining approach to information. *Information research*, 3(2), 1997.

[32] Josh Baer. Shortening the feedback loop: How spotify's big data ecosystem has evolved to produce real-time insights, 2016. URL https://www.youtube.com/watch?v=PXSVDqRVFSU.

[33] Andrew Stein. Big data and analytics, the analytics value chain - part 3, 2013. URL http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/.

[34] Mary Schacklett. Data curation takes the value of big data to a new level, 2016. URL http://www.techrepublic.com/article/data-curation-takes-the-value-of-big-data-to-a-new-level/.

[35] Vitria. Iot analytics - a new vision is needed, 2015. URL http://www.vitria.com/wp-content/uploads/2015/08/advanced-analytics-for-iot-wp.pdf.

[36] Accenture. What's your data worth?, 2013. URL https://www.accenture.com/gb-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_16/Accenture-Whats-Your-Data-Worth.pdf#zoom=50.

[37] Mike Forstieri. What exactly the heck are prescriptive analytics?, 2017. URL http://blogs.forrester.com/mike_gualtieri/17-02-20-what_exactly_the_heck_are_prescriptive_analytics.

[38] Todd Winey. Garbage in, gospel out?, 2017. URL https://www.intersystems.com/intersystems-blog/pulse/garbage-in-gospel-out/.

[39] Eirik Aflekt. Tba. Master's thesis, University of Stavanger, 2017.

[40] Gil Press. The big data landscape revisited, 2013. URL https://www.forbes.com/sites/gilpress/2013/04/23/the-big-data-landscape-revisited/#435345671c30.

[41] J.E. Dunn. 21 of the most infamous data breaches affecting the uk, 2017. URL http://www.techworld.com/security/uks-most-infamous-data-breaches-3604586/.

[42] Darkreading.com. Survey: Customers lose trust in brands after a data breach, 2016. URL http://www.darkreading.com/vulnerabilities---threats/survey-customers-lose-trust-in-brands-after-a-data-breach/d/d-id/1325570.

[43] Giuseppe Macri. Consumers are losing trust in hacked companies, 2016. URL http://www.insidesources.com/consumers-are-losing-trust-in-hacked-companies/.

[44] Russel Goldman. What we know and don't know about the international cyber-attack, 2017. URL https://www.nytimes.com/2017/05/12/world/europe/international-cyberattack-ransomware.html?_r=0.

[45] Eric Poole. Quantifying business value of information security. Technical report, SANS Institute, 2009.

[46] ICO. Overview of the general data protection regulation (gdpr), apr 2017. URL https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/.

[47] Regjeringen. Personvernforordning, apr 2017. URL https://www.regjeringen.no/no/sub/eos-notatbasen/notatene/2014/aug/forslag-til-personvernforordning/id2433856/. Posisjon note.

[48] DLA Piper. A guide to the gener al data protection re gulation, nov 2016. URL https://www.dlapiper.com/~/media/Files/Insights/Publications/2016/12/General%20Data%20Protection%20Regulation%20Brochure.PDF.

[49] European Commission. General data protection regulation, 2016. URL http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

[50] European Commision. Directive 95/46/ec of the european parliament and of the council, 1995.

[51] Paul De Hert and Vagelis Papakonstantinou. The proposed data protection regulation replacing directive 95/46/ec: A sound system for the protection of individuals. *Computer Law & Security Review*, 28(2):130–142, 2012.

[52] European Union. Protection of personal data, apr 1995. URL http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Al14012.

[53] Lee Bygrave. Minding the machine: art 15 of the ec data protection directive and automated profiling"(2000). *Privacy Law and Policy Reporter*, 7:67, 2010. URL http://www5.austlii.edu.au/au/journals/PrivLawPRpr/2000/40.html.

[54] DLA Piper. Eu data protection regulation - key changes, n.d. URL https://www.dlapiper.com/en/uk/focus/eu-data-protection-regulation/key-changes/.

[55] Trunomi. Gdpr key changes, n.d. URL http://www.eugdpr.org/key-changes.html.

[56] Aysem Diker Vanberg and Mehmet Bilal Ünver. The right to data portability in the gdpr and eu competition law: odd couple or dynamic duo? *European Journal of Law and Technology*, 8(1), 2017.

[57] Article 29 Working Party. Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in a high risk" for the purposes of regulation 2016/679, 2017. URL http://ec.europa.eu/newsroom/document.cfm?doc_id=44137. Adopted on 4 April 2017).

[58] Lukasz Olejnik. Data protection impact assessment. first guidelines, 2017. URL https://blog.lukaszolejnik.com/data-protection-impact-assessment-first-guidelines/.

[59] Article 29 Working Party. Guidelines on data protection officers ('dpos'), 2016. URL http://ec.europa.eu/newsroom/document.cfm?doc_id=44100. Adopted on 13 December 2016.

[60] Article 29 Working Party. Opinion 03/2013 on purpose limitation, 2013. URL http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf. Adopted on 2 April 2016.

[61] European Commission. The eu data protection reform and big data, 2016. URL http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf.

[62] Moritz Hardt. How big data is unfair, 2014. URL https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

[63] ICO. Big data, artificial intelligence, machine learning and data protection, apr 2017. URL https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

[64] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[65] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[66] Ethan Chiel. Eu citizens might get a 'right to explanation' about the decisions algorithms make, 2016. URL http://fusion.kinja.com/eu-citizens-might-get-a-right-to-explanation-about-the-1793859992.

[67] Anupam Datta, Shayak Sen, and Yair Zick. Algoritahmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617. IEEE, 2016.

[68] Lukasz Olejnik. Gdpr consent requirements. first ico guideliness, 2017. URL https://blog.lukaszolejnik.com/gdpr-consent-requirements-first-ico-guidelines/.

[69] Q Qdr. Benefits of demand response in electricity markets and recommendations for achieving them. *US department of energy*, 2006.

[70] EUROPEAN COMMISSION. Europe 2020 a strategy for smart, sustainable and inclusive growth, 2010. URL http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF.

[71] Zahra Baharlouei and Massoud Hashemi. Demand side management challenges in smart grid: A review. In *Smart Grid Conference (SGC), 2013*, pages 96–101. IEEE, 2013.

[72] D.Y. Goswami and F. Kreith. *Energy Efficiency and Renewable Energy Handbook, Second Edition*. Mechanical and Aerospace Engineering Series. CRC Press, 2015. ISBN 9781466585096. URL https://books.google.no/books?id=GtaYCgAAQBAJ.

[73] Chris Beard. *Smart metering for dummies*. –For dummies. Wiley Publishing, Inc, 2012. ISBN 978-0-470-74164-1. URL https://www.cgi-group.co.uk/smart-metering-for-dummies.

[74] IS Group et al. Managing big data for smart grids and smart meters. *IBM Corporation, whitepaper (May 2012)*, 2012.

[75] Ramyar Rashed Mohassel, Alan Fung, Farah Mohammadi, and Kaamran Raahemifar. A survey on advanced metering infrastructure. *International Journal of Electrical Power & Energy Systems*, 63:473–484, 2014.

[76] Kaile Zhou, Chao Fu, and Shanlin Yang. Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, 56:215–225, 2016.

[77] Kofi Afrifa Agyeman, Sekyung Han, and Soohee Han. Real-time recognition non-intrusive electrical appliance monitoring algorithm for a residential building energy management system. *Energies*, 8(9):9029–9048, 2015.

[78] Kaustav Basu, Vincent Debusschere, and Seddik Bacha. Residential appliance identification and future usage prediction from smart meter. In *Industrial Electronics Society, IECON 2013-39th Annual Conference of the IEEE*, pages 4994–4999. IEEE, 2013.

[79] Ulrich Greveler, Peter Glösekötterz, Benjamin Justusy, and Dennis Loehr. Multimedia content identification through smart meter power usage profiles. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[80] Eoghan McKenna, Ian Richardson, and Murray Thomson. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41:807–814, 2012.

[81] KT Weaver. A perspective on how smart meters invade individual privacy, 2014.

[82] Massoud Amin. Challenges in reliability, security, efficiency, and resilience of energy infrastructure: Toward smart self-healing electric power grid. In *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, pages 1–5. IEEE, 2008.

[83] Karen Ehrhardt-Martinez, Kat A Donnelly, Skip Laitner, et al. Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities. Technical report, American Council for an Energy-Efficient Economy Washington, DC, 2010.

[84] D. and X. Yu. Smart electricity meter data intelligence for future energy systems: A survey. *IEEE Transactions on Industrial Informatics*, 12(1):425–436, Feb 2016. ISSN 1551-3203. doi: 10.1109/TII.2015.2414355.

[85] Gianfranco Chicco, Roberto Napoli, Petru Postolache, Mircea Scutariu, and Cornel Toader. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, 18(1):381–387, 2003.

[86] Ben Packer. 7 reasons why utilities should be using machine learning, 2015. URL https://blogs.oracle.com/utilities/utilities-machine-learning.

[87] Barry Ficher. We plotted 812,000 energy usage curves on top of each other. this is the powerful insight we discovered., 2013. URL https://blogs.oracle.com/utilities/load-curve-archetypes.

[88] Chris Mooney. Knowing your "energy personality" can save you a lot of money, 2015. URL hhttps://www.washingtonpost.com/news/energy-environment/wp/2015/03/03/why-knowing-your-energy-personality-could-help-save-you-a-lot-of-money/?utm_term=.7ebb495f76cc.

[89] M. Pipattanasomporn, M. Kuzlu, and S. Rahman. An algorithm for intelligent home energy management and demand response analysis. *IEEE Transactions on Smart Grid*, 3(4):2166–2173, Dec 2012. ISSN 1949-3053. doi: 10.1109/TSG.2012.2201182.

[90] Krzysztof Gajowniczek and Tomasz Ząbkowski. Data mining techniques for detecting household characteristics based on smart meter data. *Energies*, 8(7):7407–7427, 2015.

[91] Weaver K.T. 'smart' meters generate a 'gold mine of data' for utilities, 2015. URL https://smartgridawareness.org/2015/12/31/smart-meters-generate-gold-mine-of-data/.

[92] Kwang-Ho Kim, Jong-Keun Park, Kab-Ju Hwang, and Sung-Hak Kim. Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert systems. *IEEE Transactions on Power Systems*, 10(3):1534–1539, 1995.

[93] M Ghofrani, M Hassanzadeh, M Etezadi-Amoli, and MS Fadali. Smart meter based short-term load forecasting for residential customers. In *North American Power Symposium (NAPS), 2011*, pages 1–5. IEEE, 2011.

[94] Pengwei Du and Ning Lu. Appliance commitment for household load scheduling. *IEEE transactions on Smart Grid*, 2(2):411–419, 2011.

[95] Nima Amjady, Farshid Keynia, and Hamidreza Zareipour. Short-term load forecast of microgrids by a new bilevel prediction strategy. *IEEE Transactions on Smart Grid*, 1(3):286–294, 2010.

[96] D. Alahakoon and X. Yu. Advanced analytics for harnessing the power of smart meter big data. In *2013 IEEE International Workshop on Inteligent Energy Systems (IWIES)*, pages 40–45, Nov 2013. doi: 10.1109/IWIES.2013.6698559.

[97] Manisa Pipattanasomporn, Murat Kuzlu, and Saifur Rahman. An algorithm for intelligent home energy management and demand response analysis. *IEEE Transactions on Smart Grid*, 3(4):2166–2173, 2012.

[98] Michael Angelo A Pedrasa, Ted D Spooner, and Iain F MacGill. Coordinated scheduling of residential distributed energy resources to optimize smart home energy services. *IEEE Transactions on Smart Grid*, 1(2):134–143, 2010.

[99] Robert H Lasseter and Paolo Paigi. Microgrid: A conceptual solution. In *Power Electronics Specialists Conference, 2004. PESC 04. 2004 IEEE 35th Annual*, volume 6, pages 4285–4290. IEEE, 2004.

[100] Jim Lazar. Teaching the "duck" to fly. *Montpellier, VT: Regulatory Assistance Project*, 2014.

[101] David Perera. Smart grid powers up privacy worries, 2015. URL http://www.politico.com/story/2015/01/energy-electricity-data-use-113901.

[102] Timothy Morey, Forbath Theodore, and Achoop Allison. Customer data: Designing for transparency and trust, 2015. URL ttps://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust.

[103] VaasaETT. Smarte målere (ams) og feedback. Technical Report 30, Norwegian water resources and energy directorate (NVE), 2014.

[104] R.K. Wysocki. *Effective Project Management: Traditional, Agile, Extreme*. Wiley, 2013. ISBN 9781118729311. URL https://books.google.no/books?id=RTBIAgAAQBAJ.

[105] ONZO. About onzo, 2016. URL https://www.youtube.com/watch?v=uluKjzqHDz0. Youtube advertisement.

[106] S.J. Olsen. Nå blir også fjordkraft mobiloperatør, 2017. URL https://www.tek.no/artikler/fjordkraft-blir-mobiloperator-tilbyr-ekstra-gode-priser-til-egne-kunder/382045.

[107] D.R. Jerijervi. Telenor tar snarvei i konkurransen med facebook og googler, 2016. URL http://kampanje.com/tech/2016/02/--telenor-har-tatt-snarvei-konkurransen-med-facebook-og-google/.

[108] Cognizant. The business value of trust, 2016. URL https://www.cognizant.com/whitepapers/the-business-value-of-trust-codex1951.pdf.

[109] Norwegian water resources and energy directorate (NVE). Smart metering (ams), mar 2017. URL https://www.nve.no/energy-market-and-regulation/retail-market/smart-metering-ams/.

[110] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo. Big data analytics in power distribution systems. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, Feb 2015. doi: 10.1109/ISGT.2015.7131868.

[111] Ye Yan, Yi Qian, Hamid Sharif, and David Tipper. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *IEEE communications surveys & tutorials*, 15(1):5–20, 2013.

[112] European Union. Smart metering deployment in the european union, n.d. URL http://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union.

[113] Claude R. Olsen. Venter kvantesprang i teknologiutnyttelse, feb 2017. URL http://smartgrids.no/venter-kvantesprang-i-teknologiutnyttelse/.

[114] Networked Energy Services Corporation. Smart meter to han connectivity, n.d. URL https://www.networkedenergy.com/uploads/whitepapers/SmartMeter_HAN_WP_v21.pdf.

[115] Meera Balakrishnan. Smart energy solutions for home area networks and grid-end applications. *Proc. Smart Energy*, pages 67–73, 2012.

[116] Lauren Callaway and Neil Strother. Market data: Meter data management. Technical report, Navigant Research, 2015.

[117] Rob Young, John McCue, and Christian Grant. The power is on: How iot technology is driving energy innovation, 2016. URL https://dupress.deloitte.com/dup-us-en/focus/internet-of-things/iot-in-electric-power-industry.html.

[118] Tim Fairchild. The soft grid 2013-2020: big data & utility analytics for smart grid. *Research Excerpt, SAS in association GTM Research*, 2013.

[119] Jukka V. Paatero and Peter D. Lund. A model for generating household electricity load profiles. *International Journal of Energy Research*, 30(5):273–290, 2006. ISSN 1099-114X. doi: 10.1002/er.1136. URL http://dx.doi.org/10.1002/er.1136.

[120] Kaile Zhou and Shanlin Yang. Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustainable Energy Reviews*, 56:810 – 819, 2016. ISSN 1364-0321. doi: http://doi.org/10.1016/j.rser.2015.12.001. URL http://www.sciencedirect.com/science/article/pii/S1364032115013817.

[121] Dave Webb, Geoffrey N Soutar, Tim Mazzarol, and Patricia Saldaris. Self-determination theory and consumer behavioural change: Evidence from a household energy-saving behaviour study. *Journal of Environmental Psychology*, 35:59–66, 2013.

[122] Jessica Stromback, Christophe Dromacque, Mazin H Yassin, and Global Energy Think Tank VaasaETT. The potential of smart meter enabled programs to increase energy and systems efficiency: a mass pilot comparison short name: Empower demand. *Vaasa ETT*, 2011.

[123] Philip Lewis, Rafaila Grigoriou, Christophe Dromacque, Anna Bogacka, and Steve Xu. Assessing the potential of energy consumption feedback in norway. Technical report, Vaasa ETT, 2015.

[124] Øystein Meland. Norske vindmøller produserte 2,5 twh i fjor, 2016. URL https://www.ge.no/geavisa/norske-vindmoller-produserte-25-twh-fjor/.

[125] THEMA Consulting Group. Teoretisk tilnærming til en markedsløsning for lokal fleksibilitet. Technical Report 30, Norwegian water resources and energy directorate (NVE), 2015.

[126] Severin Borenstein, Michael Jaske, and Arthur Rosenfeld. Dynamic pricing, advanced metering, and demand response in electricity markets. *Center for the Study of Energy Markets*, 2002.

[127] NEST. Nest labs introduces world's first learning thermostat, 2011. URL https://nest.com/fr/press/nest-labs-introduces-worlds-first-learning-thermostat/.

[128] C. Vivekananthan, Y. Mishra, and F. Li. Real-time price based home energy management scheduler. *IEEE Transactions on Power Systems*, 30(4):2149–2159, July 2015. ISSN 0885-8950. doi: 10.1109/TPWRS.2014.2358684.

[129] A. H. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2): 120–133, Sept 2010. ISSN 1949-3053. doi: 10.1109/TSG.2010.2055903.

Appendix A

Smart Metering

A.1 Introduction

Traditionally, households have submitted their energy consumption by manually reading their electricity meter once a month or by a monthly visit by a meter reader.[75] However, a widespread world wide roll-out of smart meters allow for the majority of customers to have their energy consumption automatically read, and other benefits that comes along with it. Smart meters allow for communication from water and gas meters, this aspect of smart meters is however, not considered in the thesis.

Initially the big difference between meter readers is that smart meters provide detailed Time of Use (ToU) consumption, typically automatic readings at 15 to 60 minutes intervals. ToU refers to the meter's ability to record consumption in terms of *when*, rather than *how much* is consumed.[73] ToU creates immense possibilities for consumers to gain knowledge about their consumption creating a behavioral change, that will not only provide customers with correct cheaper billings, but also flatten consumption peaks which, in turn, will benefit the environment and reducing the need for expensive network investments.[109]

Two-way communication also allow smart meters to be remotely instructed and re-configured either by the consumer itself or from a control center on half of the system operator. Such remote actions include:[73, 75]

- **Meter readings on-demand**: Near-real time meter readings is possible.

- **Change of tariff**: Customers can change rate and structure of tariff in response to a price change instigated by the supplier.
- **Change of payment method**: For example, switching between credit or pre-payment options.
- **Change in read frequency**: The consumption read interval can be changed on demand.
- **Load limiting/shedding**: Remote or control over a customers consumption at a level agreed upon. Mainly to balance generation and demand, but also preventing customer from receiving unnaturally high billings
- **Tamper alerts**: Automatic detection, notification and response to tampering attempts
- **Disablement/enablement**: Ability to remotely turning on or off the supply as a response to for example, a tamper alert or risk for network blackout.
- **Messaging**: Communicating directly with the customer through channels such as In-Home Display (IHD), smart phone apps or web-portals.
- **Firmware updates**: Updating the smart meter software in order to provide new functionalities and fix bugs.

Different literature provide different numbers in terms of order of magnitude when comparing smart and conventional readers. Nevertheless the difference is paramount. Compared to non-interval data, Yu et al. [110] estimates that smart meters will generate up to 3000 times more data while Beard [73] estimates 4000 times the amount of the past. Yu et al. [110] furthermore estimates that by 2022 the amount of data generated annually by smart meters alone will reach 2 petabytes worldwide. This is definitely a considerable amount, which in the context of big data opens up a new world of possibilities in terms of analytics.

Consumers will be provided near real-time information about their electricity consumption and prices. This is furthermore facilitating for new energy related services, such as demand response programs enabling consumers to shape their consumption in accordance to market price while reducing the need for future investments in the grid.[109]

Below is a brief summarization of features enabled by smart metering

- Dynamic pricing or time-based pricing
- Feedback programs
- Demand response programs
- Supports smart home and automation
- Net energy metering and power purchase agreements
- Promoting efficient power consumption through for example rewarding microgeneration
- Power quality monitoring
- Failure and outage notification and predictions of such
- Energy theft detection

A.2 Advanced Metering Infrastructure

An Advanced Metering Infrastructure is an integration of different technologies in a configured infrastructure of different levels in a hierarchy.[111] The building blocks of the AMI, as shown in figure A.1, consists mainly of smart meters, communication networks, Meter Data Management Systems (MDMS) and means to collect data into software applications and interfaces.[75]

The collected data can be transmitted through a wide variety of fixed and public network standards. The AMI host system receive the consumption data where it, subsequently, is sent to a MDMS that manages data storage and analysis and provides the utility service provider with useful information. [75]

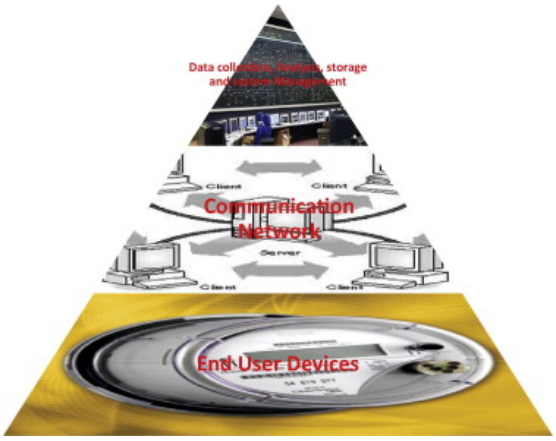

Figure A.1: Schematic view of the AMI building blocks[75]

A.2.1 Smart meters and smart devices

The end user is equipped with hardware and software that is comprised with state-of-the-art technologies for data collection and measurement. These smart devices at consumer level are typically smart meters that communicates consumption data to both the user and the service provider. However, as the Internet of Things i increasingly becoming a part of everyday life there is reason to believe that more devices at the user end, able to collect consumer data and communicate with the AMI, will become prominent. Smart thermostats is already an established technology sensors are increasingly integrated within the the home network.

A.2.2 Communication

The European Union expects 72% of European consumers will have a smart meter by 2020.[112] Hence, the data communication, also referred to as *head-end* system collects data from from a large and disparate set of smart meters. The head-end system handles the two-way communication as well as sending service request and receives messages, such as scheduled meter readings.[73]

Standard communication is key for the smart meters to be able to send and receive collected information. When considering the number of smart meters and correspondingly number of users a highly reliable reliable communications network is essential for transferring the ever growing volume of data. The process of designing and selecting a communi-

cation network must be painstaking and requires the careful consideration of the following factors:[75]

- Huge amounts of data transfer
- Restriction in data access
- Confidentiality of sensitive data
- Representing complete information of consumer's consumption
- Showing grid status
- Authenticity of data and precision in communication with target device
- Cost effectiveness
- Ability to host modern features beyond AMI requirements
- Supporting future expansion

There are many different architectures and networks available for realizing an AMI thus, different the mediums and communication technologies is used as well. Everyone has their advantages and disadvantages Some of which are:[73, 75]

- Power Line Carrier (PLC):
- Broadband over Power Line (BPL)
- General Packet Radio Service (GPRS):
- Short Message Service (SMS)
- Radio:
- Wi-Fi:
- WiMax
- Bluetooth
- Peer-to-Peer

- ZigBee:

By many, AMI in it self is, is not considered to give enough value through meter readings for billing purposes to justify the initial investment[113]. One proposed solution is smart meter to Home Area Network connectivity, that will drive customer engagement and allow for deriving more value from the smart meter investment.[114]

A.2.3 Home Area Network

The Home Area Network(HAN) is a dedicated network, connecting, usually via ZigBee, devices in the home to the overall smart metering system. Smart devices within the premise of a building can therefore communicate directly to smart meters and to each other. [115]

The typical first device in the HAN is In-Home Displays(IHD), which provide consumers with near real-time information on energy usage, cost and greenhouse gas generation. The *smart home* is no longer theme of the future and many appliance manufacturers are becoming more keen on the idea of home networking.[73] The next chapter, in subsection

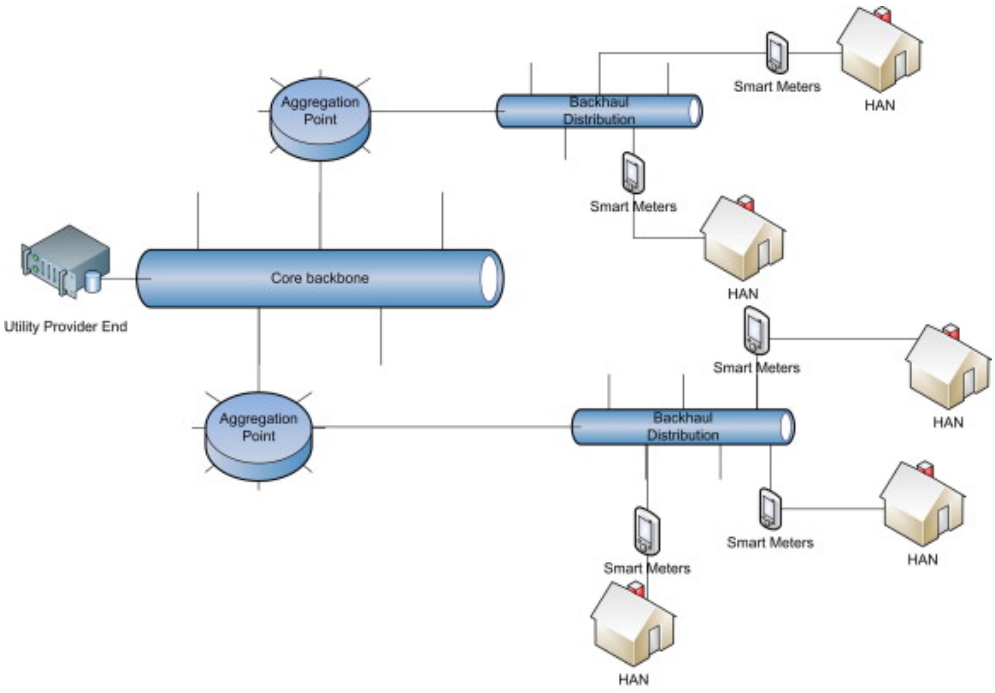

Figure A.2: Overview of utility network[75]

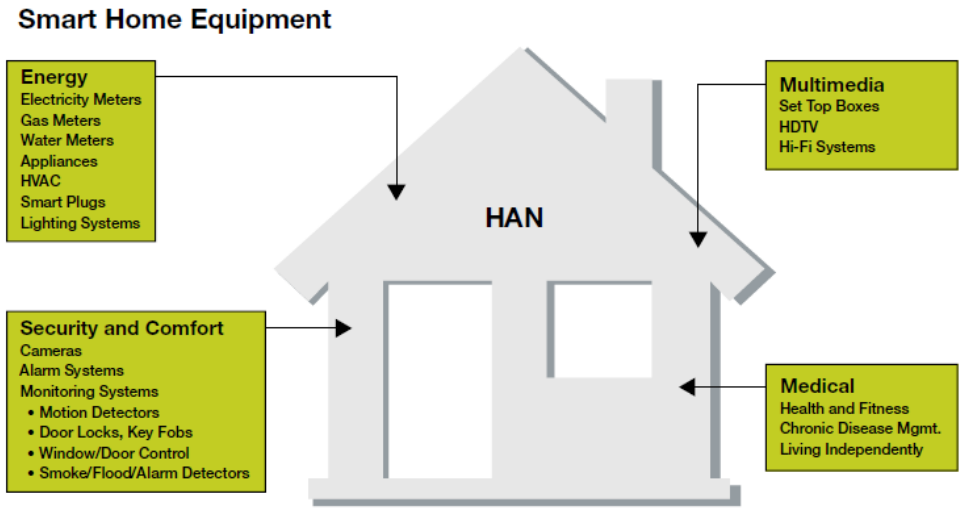

Figure A.3: Forecasted smart networked home[115]

A.2.4 Meter data management systems

The amounts of data collection from smart meters is so huge that it is considered as big data. Unless the data can be transformed into valuable information, there won't be much benefit for those who wants to utilize it. Thus, the tools for managing and analyzing the big data is required.[75]

Meter Data Management Systems (MDMS), a system that collects, processes and stores meter data to help utilities turn information into valuable insights and improve the operation of the grid.[116] Mohassel et al. [75] stresses that MDMS should be able to address the three demands: improvement and optimization of utility grids, improvement and optimization of utility management and enabling customer engagement. This regardless of the features and complexity of the system.

MDMS constitute a central part of the overall data management system at the utility provider end and the objective of this system, additionally to storing and processing, is analyzing data for billing purposes, handling demand response, consumption profiles and real time reactions to emergencies in the grid. This system is a multi modular structure consisting of the following modules:[75]

- Meter Data Management System (MDMS)
- Consumer Information System (CIS), billing system and utility website

- Outage Management Systems (OMS)
- Enterprise Resource Planning (ERP)
- Mobile Workforce Management
- Geographic Information Systems
- Transformer Load Management

Additionally, MDMS has several responsibilities regarding the AMI data. Mainly ensuring a accurate and complete flow of information from consumers to the other management modules of the data management systems

Whereas MDMS has been a part of utilities IT suite for years, the introduction of smart metering has rendered much of the traditional capabilities obsolete. In order to handle the vast amounts of data from smart meters, and data from new applications associated with them, the need for solutions providing analytics, scalability and flexibility has become prominent in order to satisfy the demands of MDMS. This presupposes predictive and prescriptive analytics as methods for gaining insights and taking action in an increasingly complex data environment. Hence, meter data analytics(MDA) is more or less becoming an integrated part of MDMS.

The capabilities of MDA to use smart meter data enable applications such as:

- Outage management
- Distribution management
- Demand response
- Time-of-use rates
- Power quality monitoring
- Behind-the-meter distributed energy resources integration
- Home energy management

In terms of flexibility and scalability, local IT-infrastructures gets outdated quickly as the

unpredictability of the data environment demands continuous upgrade of storage and processing power. Many companies, even the big ones, does not possess the financial and physical resources to handle this in the long run, driving the need for alternative solutions. Cloud computing provide the means in which MSMS always will have access to state-of-the art technology that will satisfy all technical requirements along the value chain. Moreover, meter data contains critical personal information and business critical information. This requires a disaster proof storage facility and top end data security. Not to mention, data back up and contingency plans must also be in place. Such provision can be very capital intensive, where cloud computing and visualization will provide the means to which this is manageable. However, the security of data may still be of concern[75]

A.2.5 Big data and utility analytics

Simultaneously as more and more devices is connected to the Internet of Things, the electric utility industry knows how to leverage new technology to improve efficiency and performance of the power grid. Gathering data from sensors improve the resilience of the grid, this data is subsequently used to actively manage resources and finally used to provide stakeholders with informed decision making about power usage and decision making.[117]

The more connected devices, the more apparent becomes the need for improved meter data management technology. The use of multivariate data management systems will allow utilities to collect, organize, manage and analyze data from distributed sensors and smart appliance alongside with smart meter data. Analytics is therefore no longer typical reporting tools and descriptive analytics using historical data. The analytical tools available provide utilities with real-time predictive analytics enabling them to be more proactive in decision making. This will make utilities capable of managing grid conditions such as intermittent loads, renewable energy sources, changing weather patterns which, in turn, represent the ultimate goal for smart grid capabilities. Figure A.4 demonstrate the three primary domains of grid analytics and how they sit relative to the physical infrastructure that will rely heavily on analytics the enterprise, grid operations (transmission and distribution), consumer-oriented offerings. Additionally energy portfolio management and trading is considered as a fourth domain that will become increasingly important as the supplier market becomes more competitive. All of the mentioned domains are ripe

with opportunity.[118] Vendors of services are therefore competing like never before on analytics delivered to the smart grid.

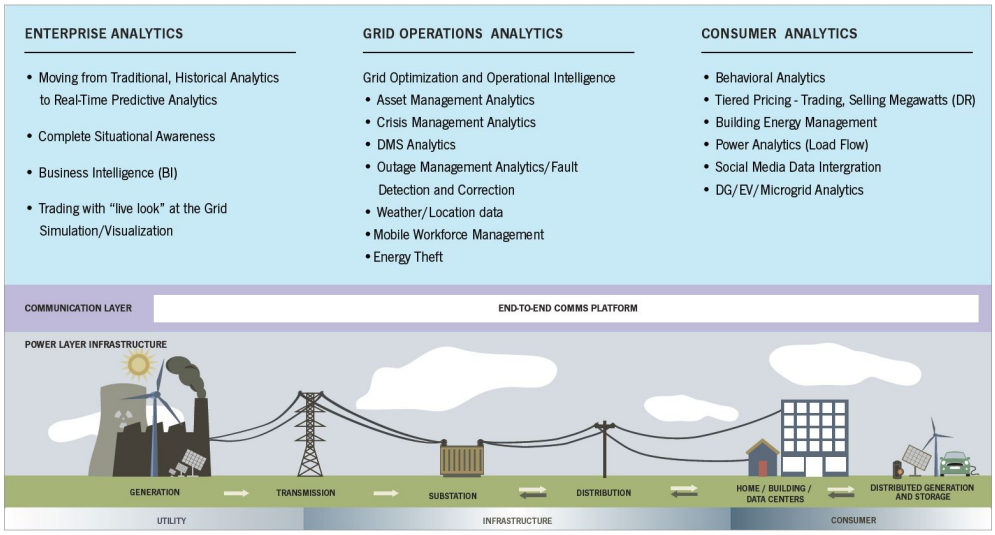

Figure A.4: Three primary domains of smart grid analytics[118]

The smart grid analytics market is divided in different software based analytics solutions including AMI analytics, demand response analytics, energy forecasting analytics, grid optimization analytics and analytics for advanced dashboard visualizations and reporting. All of which has a variety of applications servicing both end users and utilities.

- Geospatial and visual analytics that offer a centralized view of multiple technologies
- Peak load management (via demand-side management analytics) and energy portfolio management analytics
- Consumer behavioral analytics (including comparison to neighbors/peers)
- Home signature and thermostat control analytics
- Time-of-use pricing analytics
- Renewable energy and storage analytics
- Asset protection analytics and predictive asset maintenance
- Service quality analytics
- Revenue protection (including theft and nontechnical loss analytics)

- Analytics to correct legacy system errors (such as CIS and MDMS)

The deployment of smart meters is considered as the first step for ensuring a reliable energy supply, incorporation of distributed generation resources, development of innovative storage solutions, reducing need to invest in infrastructure and generation facilities and to give customers more control over their energy consumptionuse.[74] Smart meter data in the combination with other evolving technologies can generate remarkable volumes of data of high speed and complexity. The opportunity is now, for existing companies as well as start-ups to find ways in which they can transform this big data into business value. Amongst themes to be highlighted is capabilities of forecasting demand, influencing customer usage patterns, optimizing operational performance, preventing power outage as well as counteraction theft and fraud.

This page has been left intentionally blank.

Appendix B

Consumption behavior

B.1 Dimensions of consumption data

Household energy consumption behavior can be described in the three dimensions, time, user and the spatial as illustrated in figure In the time dimension with the introduction of smart meters, electricity consumption may be collected in near real-time. The granularity of the behavior can differ from an 15 minutes to a year. Consumption during an hour is subjected to great randomness, on daily basis patterns appear on usage by time of day, typically creating peak demands for energy at morning time and after work or evenings. Monthly and annual behavior patterns are on the other hand more subjected to external factors such as season, weather conditions.[119]

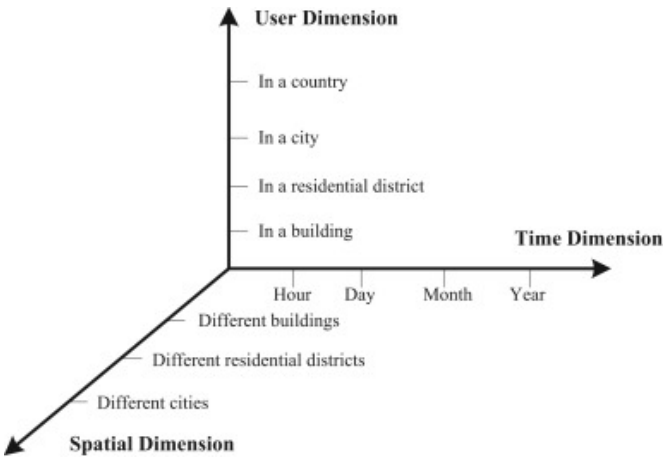

Figure B.1: Figure showing the different dimensions of household energy consumption[120]

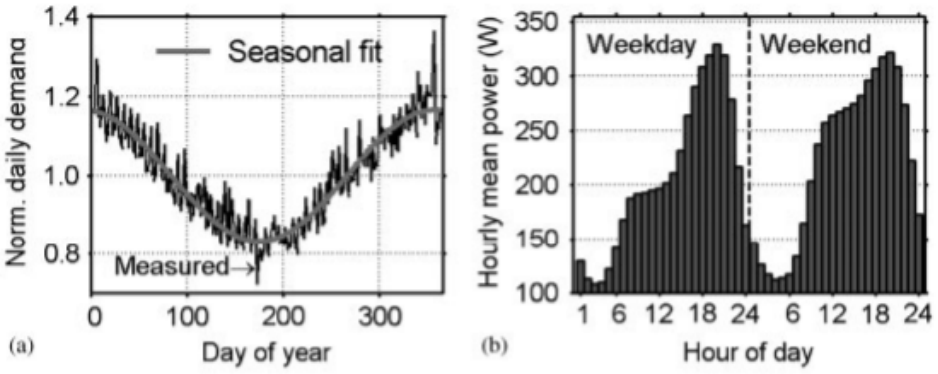

Figure B.2: Graph a) shows a typical seasonal change in demand. Graph b) shows typical demand in a day[119]

The user dimension include both internal and external factors. The internal factors are the subjective intentions of the consumer, such as habits and environmental awareness. External factors include housing characteristics such as building type and size, demographic factors and typical working day. Spatial dimension on the other hand shows the differences in consumer behavior due to geographical environment, level of economic development and climate. Also residential districts, different buildings or even apartments may show significant differences.[120]

An increase in autonomous motivation is associated with an increase in energy saving behavior. In this context self-determination theory provides the evidence of a positive relationship between meeting and supporting a person's basic psychological needs for autonomy, competence and relatedness and motivations. Utilities and government needs to empower consumers with knowledge and information about how to achieve a change and focus attention to identifying ways in which autonomy, competence and relatedness can be satisfied and supported[120]

More controlling approaches such as financial incentives and threat of punishment for non-compliance are seemingly non-significant with intentions and behavior. Moreover, they may even hinder the extent to which a person feels their need for autonomy, competence and relatedness are being supported. Such controlling approaches can furthermore prove more costly and less successful in creating sustainable change.

Webb et al. [121] explains how autonomously motivated behaviors are more likely to be sustained in the long term and suggests ICT supported initiatives to meet and support per-

sons needs for self-determination in creating a sustained behavioral change. Furthermore, Webb et al. [121] states that factors like socio-economic status, type of housing and existing knowledge, motivation to conserve energy and ability to engage in energy conservation must be considered when identifying potential long-term energy reduction strategies.

Energy consumption behavior can be divided in to two major research categories, namely the behavior-oriented paradigm and economic paradigm. This subsection describes the two paradigms, how they relate and why they are important to different strategies in demand side management.

The behavior-oriented paradigm assumes that energy consumption behavior are determined by the complex interplay of intrapersonal factors. Interpersonal factors and external factors presented in table B.1. Understanding these factors by big data analytics may create a better understanding of how to change the behavior.[120]

Table B.1: Examples of influencing factors on energy consumption behavior[120]

Factors		
Intrapersonal	**Interpersonal**	**External**
Habits	Norms	Incentives
Attitudes	Social comparison	Rewards
Values		Punishment

The understanding of these factors has been the subject of numerous research initiatives and several intervention strategies, such as goal-setting, feedback, demonstrations and general information, has been developed throughout with the objective of promoting energy conservation. Zhou and Yang [120] suggests to provide rewards or targeting the individuals perceptions, preferences or abilities in order to induce eco-friendly behavior.

The introduction of smart meters has the potential of providing the consumer with continuous feedback,[122] which is considered the most effective.[120] Feedback programs has proven efficient in terms of energy conservation and different feedback strategies such as those presented by Stromback et al. surpasses the scope of feedback presented by Zhou and Yang.

Depending on choice of technology and channel in which feedback is communicated through a feedback program can implement several intervention strategies, such as those mentioned above. Feedback programs are therefore considered in this thesis as the main

technology enabled by smart meters that best supports autonomously motivated energy behavior.

The rational choice theory is the basic underlying principle of the economic paradigm, suggesting that people with rationality seek to obtain the maximum benefit of with minimum cost in order to maximize their expected utility. In the context of energy consumption behavior consumers tend to make decisions based on the cost, benefits and the available information. From the suggested perspective, users will take action if sufficient information is given. From this perspective DSM is considered as an effective way to promote energy consumption behavioral change. The main objectives is to create a change in the time pattern of energy consumption and the magnitude of network load, ensuring more sustainable load shapes. DSM has six major objectives and tasks, namely peak clipping, valley filling, load shifting, strategic conservation, strategic load growth, and flexible load shape. In DSM several actions can be taken to achieve mentioned objectives, these include;

- energy efficient appliances
- reduction of energy consumption
- shifting of time when energy is consumed
- implementing dynamic pricing

Demand response programs(DR) are subsets of DSM and provide many benefits in regards to the actions mentioned above, such as: The collection of consumption data is important to AMI's provide the infrastructure while smart meters provide the data enabling the above mentioned actions. Also home energy management systems(HEMS) are new more innovative approaches in DSM.

Different initiatives has been implemented, and many of them with the key focus of shifting consumption away from peak periods and to shed consumption at peak periods where the stability and health of the grid is at risk. However, with varying degree of success. Webb et al. explains how autonomously motivated behaviors are more likely to be sustained in the long term and suggests ICT supported initiatives to meet and support persons needs for self-determination in creating a sustained behavioral change.

Furthermore, Webb et al. states that factors like socio-economic status, type of housing

and existing knowledge, motivation to conserve energy and ability to engage in energy conservation must be considered when identifying potential long-term energy reduction strategies.

Active participation of the demand side is considered as a core element of the smart grid and an implementation of smart meters is furthermore viewed as a key building block of the smart grid and the most cost efficient method for increasing demand side involvement and engagement.[122]

This page has been left intentionally blank.

Appendix C

Demand Side Management

Despite uncertainties such as future demand, energy resources, asset availability and grid conditions, the load serving entity should be able to, in real time, meet the changing system demands. This is what makes demand response so valuable. At a relatively low cost, the consumption behavior of the demand side can be altered in order to create more flexibility.

Hence, a solid system for managing the electrical power system must be in place. [72] argues that DSM is a wide term that cover more than its many definitions in literature. However, the most widely accepted definition is:

> Demand-side management is the planning, implementation, and monitoring of those utility activities designed to influence customer use of electricity in ways that will produced desired changes in the utilities load shape,i.e., changes in the time pattern and magnitude of a utility load. Utility programs falling under the umbrella of demand-side management include load management, new uses, strategic conservation, electrification, consumer generation, and adjustment in market share (Gellings 1984-1988)

Describing the following 5 critical components of energy planning as embraced by demand side management:

1. *Influence of customer use:* Any program intended to influence the customers use of energy. Feedback programs is a typical strategy to achieve this characteristic.

2. *Achievement of selected objectives:* The program achieves a load shape change as a result of reduction in average rates, improvements in customer satisfaction, achievement of reliability target, etc.

3. *Evaluation against non-demand-side management alternatives:* A selected demand-side management program must be evaluated to the extent it is possible to supply side alternatives such as generating units, purchasing power, or supply-side storage. This must be seen in relevance to micro grids and distributed generation in the smart grid.

4. *Identify customer response:* Encompasses a process that identifies how customers will respond and not how they should respond. This could possibly include big data analysis of customer response to different demand response programs

5. *The value is influenced by the load shape:* The value of a program is examined by how they influence the costs and benefits throughout the day, week, month, and year.

There are 6 different types of load-shape objectives; peak clipping; load shifting; valley filling; strategic conservation, strategic load growth and flexible load shape. The main objectives of of DSM can however be summarized in 4 points:

- Replace existing appliances with energy efficient once
- Create a reduction of energy consumption
- Shifting of time when energy is consumed
- Implementing dynamic pricing

C.1 Feedback programs

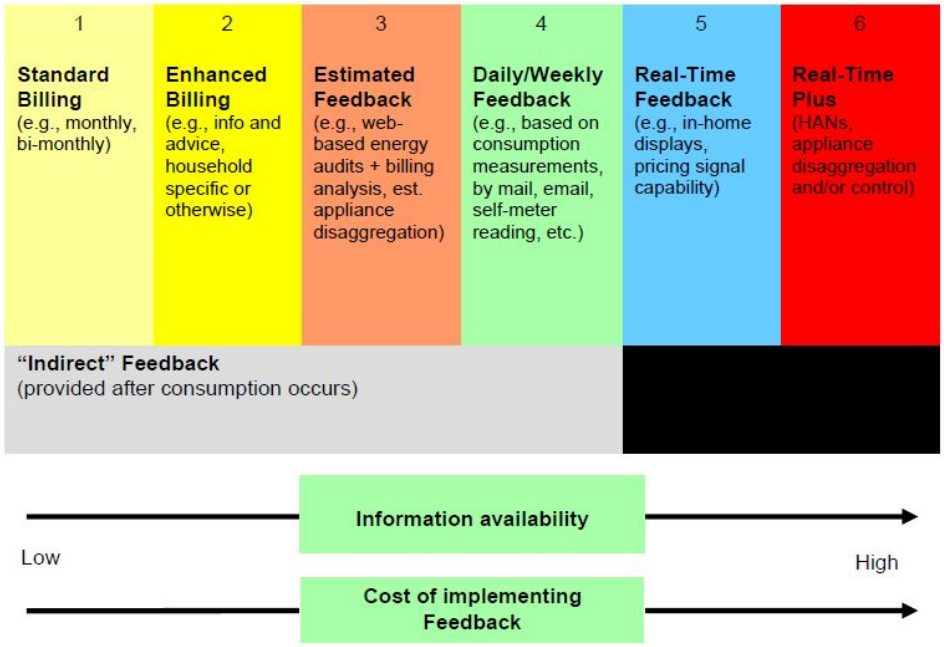

Figure C.1: Types of feedback

In a world where households in general lack knowledge of their energy consumption, the need for information is present. Information about energy costs, saving measures, and environmental impact through feedback programmes will provide households the tools to use energy more efficiently. The channels in which feedback normally is communicated is in-home displays (IHD), informative bills, web portals and mobile applications, each of which has their advantages and limitations. While IHD may be favorable in early stages of a program, mobile applications and web portals are undoubtedly the future of communication channels. As processing speeds and availability of data through the cloud is increasingly becoming an enabler of such features, IHD will arguably become more of a symbol rather than value driving device.[103]

Demand side behavioral change has been proven through feedback programs around the world, yielding substantial results of reductions between 4% to 11%. Even when the feedback programmes are applied to an entire customer base, with an opt-out option, research shows that long term saving around 2%. For a market such as Norway, 2% savings means nearly 2,5 TWh over a three year period,[123] equal to the total production of Norwegian

wind power plants in 2015. [124]

Although, energy savings may seem as the main objective of feedback, additional benefits to consumers and to the utilities and third parties offering services may be achieved. From the perspective of the utility, consumption feedback services leads to improvements of customer loyalty, which in turn has the potential to massively increase the value of customers. Feedback channels can also benefit providers of feedback services as an opportunity to market additional services and products to the customer.Stromback et al. [122] This market may increase as smart homes and home energy management systems become increasingly commercialized. From the perspective of the consumer, a feedback program provides new insight, awareness, achievement and empowerment regarding their electricity consumption. Whether knowing they are not being ripped of, being able to manage their consumption or simply receiving advice or help – transparency and service is generally much appreciated. [103]

Stromback et al. [122] emphasizes the importance of meeting the customers needs in designing a demand side program. The appropriate communication channel and type of feedback is therefore important. C.2 shows different types of feedback divided by the 4 sub-domains, situation, exploration, empowerment and compete(SEEC)

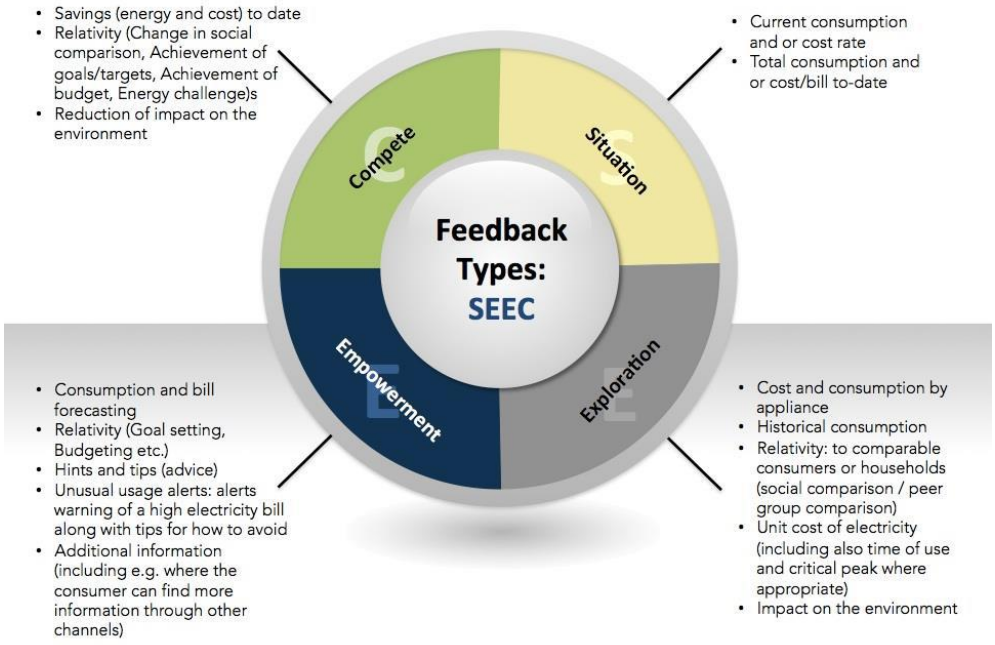

Figure C.2: Different types of feedback from demand side programs[103]

Research has proved the significance of feedback, regardless of economic background and family size. One effect of feedback programmes is the power of self-comparison, which has proven to be a big driver in changing the behavior of consumers. Lewis et al. states that it is important to find out and understand what makes the different customer change their behavior. Big data could be an enabler of individual feedback customization, optimizing the type of feedback and how it is communicated to the individual consumer. An intelligent use of big data would also provide increasing opportunities for real-time recommendations and ad targeting, potentially enabling utilities and service providers new streams of revenue.

The thought is to use big data technology to infer data from sources such as consumption from home appliances, social media activity and shopping habits to create a profile of the single consumer to better provide a tailored experience from the feedback program. The experience would be two-fold. Firstly the feedback is customized so the communication used, will provide the customer the appropriate motivation, based on their profile, derived from analytics on the above mentioned data sources. Secondly, by utilizing higher data resolutions, on the reading intervals on smart meters, near real-time data streams is available. This provides consumers the opportunity to continuously monitor and adapt their behavior through real-time feedback on their in-home or smart devices. From the utility or service providers perspective, such data resolutions provide valuable insights on energy consumption habits from the customer database, insights that can yield additional business value, and competitive advantage.(needs citations and more background knowledge, which is to be included in and introduction section above)

C.2 Demand Response Programs and Dynamic Pricing

Seeing that a feedback system is in place, consumers will be able to adapt to changing electricity prices, as these will be commonly available regardless of communication channel, this is also known as demand response and can be defined as: *Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized.*[69]

Demand response offers a variety of financial and operational benefits across the value chain. closer alignment between customers' electricity prices and and their valuation of electricity increases the resource efficiency.

- Cropped peak periods reduces required generation and transmission and potentially reducing need for future grid investments.[125]
- Lower demand during peak hours reduces the price of electricity production and holds down prices in electricity spot markets
- Reduced demand as response to system reliability problems enhances operators' ability to manage the grid and reducing the potential for outages or blackouts.

The most important benefit of demand response is improved resource efficiency of electricity production due to closer alignment between customers' electricity prices and the value they place on electricity. This increased efficiency creates a variety of benefits, which fall into four groups:

- Participant financial benefits are the bill savings and incentive payments earned by customers that adjust their electricity demand in response to time varying electricity rates or incentive-based programs.
- Market-wide financial benefits are the lower wholesale market prices that result because demand response averts the need to use the most costly-to-run power plants during periods of otherwise high demand, driving production costs and prices down for all wholesale electricity purchasers. Over the longer term, sustained demand response lowers aggregate system capacity requirements, allowing load-serving entities (utilities and other retail suppliers) to purchase or build less new capacity. Eventually these savings may be passed onto most retail customers as bill savings.
- Reliability benefits are the operational security and adequacy savings that result because demand response lowers the likelihood and consequences of forced outages that impose financial costs and inconvenience on customers.
- Market performance benefits refer to demand response's value in mitigating suppliers' ability to exercise market power by raising power prices significantly above production costs.

The rationale behind dynamic pricing is to shift consumption away from peak consumption periods in order to lower consumption periods, lowering cost related to distribution and supply, [122] as well as potentially reducing the need for future grid investments.[125] In short, dynamic pricing means that the price of electricity increases with increased demand.

Demand response is can be classified as price-based or incentive-based. However, only price-based is considered in this thesis. Common pricing program types are:[122]

- *Time-of-Use (TOU):* Aims to induce people into using electricity during times when demand is lower. Prices are therefore higher during high demand periods. These prices are known in advance by the customer, but may be subject to seasonal change.

- *Real-Time Pricing (RTP):* The price paid is tied to the price in the wholesale market. Prices changes only slightly during the day and the consumer can get notified when wholesale prices reach a certain threshold.

- *Critical Peak Pricing (CPP):* Involve substantially increased prices during heightened wholesale prices caused by heightened consumption, such as need for AC on very hot days or stability of the system is at risk.

- *Critical Peak Rebate (CPR):* In many ways the reverse of CPP as consumers are paid for the reduced consumption below predicted levels during peak hours.

In a research conducted by [122] significant reductions in peak demand was achieved. CPP and CPR yielded the highest peak clipping, but as they only occur rarely, the reduction from RTP, which occurs daily, will provide the long term clipping. Consequently, an effect of RTP will reduce the total payment to generators in the wholesale market. In the long run this means less need for new investment in power plants, which is a cost often born by the customer.Borenstein et al. [126] The research also concluded that RTP is the best alternative regarding financial savings in terms of billings to the customer. With an increasing number of smart appliances, electrical vehicles as well as more and cheaper energy storage, RTP will be able to provide even bigger savings when subjected to automated systems.

Borenstein et al. explains that the value of dynamic pricing will be greatest if the sutility can anticipate the customer responses to price changes. Considering this fact, an interplay

between two-way communication from consumption data from smart meters and RTP from the whole sale market will arguably create the best insights in customer response.

Evidently, the interplay of RTP in smart metering systems delivers the best value in terms of consumer insight, financial saving and demand response, compared to other dynamic pricing programs. Therefore, in the continuation of this literature study, RTP is of main interest. The next section will cover automated systems for demand side energy management, which is a combination between dynamic pricing and feedback systems when subjected to analytical tools such as big data.

It is economically optimal to make investments in new grid capacity when the willingness to pay for the increased capacity is higher than the costs of making the expansion. According to economic theory, the price of a good should be set above short-run marginal cost when demand exceeds capacity. Such scarcity pricing in the grid means that network customers with the lowest willingness to pay reduce their load first. In cases where the grid approaches its capacity limit, it may therefore be an alternative to introduce scarcity pricing in order to provide an optimal utilization of the grid capacity. When the revenues from scarcity pricing approaches the long term marginal cost of grid expansion, the grid capacity should be expanded.[125]

C.3 Demand side automation

Manual participation in demand response has its limitations. Much due to the fact that commercial customers has a limited ability to react to price signals. Also critical situations may occur when customers may not be able to react, such as during sleep or when traveling without reception on mobile devices . Stromback et al. [122] stresses that a utility have to notify residential customers one day in advance about shifting load, which consequently may reduce value of the load and profitability of the pricing program. In order to increase the responsiveness of a household, automation is a proposed solution. Automation can be explained as remote controllers in appliances with the ability to communicate with each other. The key features and benefits of demand side automation can be summarized as ;[122, 123]

- fast reactions to peak demand, price signals and system emergencies;

- controllable and sufficient levels of consumption reduction and financial savings and;

- 24/7 availability

There are several different ways to use automation to trigger a demand response. One way is for the utility to remotely control some of the consumers appliances, not requiring any manual involvement from the demand side. Another option is for the consumer to choose to what extent appliances should respond to price signals, this has previously been done through a web portal and yielded great results back in 2002.[122] However, this practice is becoming more obsolete as the home automation market is growing. One example of this is smart or learning thermostats, that learns the behaviors and preferences of the consumer, which allows for tailored heating cycles and temperatures. This is enabled by utilizing data from sensors, algorithms, machine learning techniques and cloud computing.[127] [89] presents an algorithm for intelligent home energy management(HEM) and demand response analytics for automating the management of high power consumption appliances. The algorithm manages loads according to the consumers preset load priority and comfort level in order to limit the power consumption of the total household.

In residential homes three main types of demand response automation exists; manual, semi-automated and fully automated with the fully automated being the most popular. A HEM system monitors and manages the operation of appliances in the home enabling load shifting and shedding. For a demand response to happen an external signal is received by the smart meter typically through a demand curtailment request, for example a CPP. When the signal is received the algorithm in the HEM will reduce household power consumption to a desired level.[89] HEM systems are also enabled to respond to a pricing signal such as RTP. This enables scheduling optimization of home appliances to minimize he cost of energy consumption.[128] The combination of an energy scheduler and a price predictor will lead to significant reduction in users' payments and furthermore reducing the peak-to-average ratio(must be defined), which may enhance utilities willingness to support energy schedulers in their smart meters.[129]

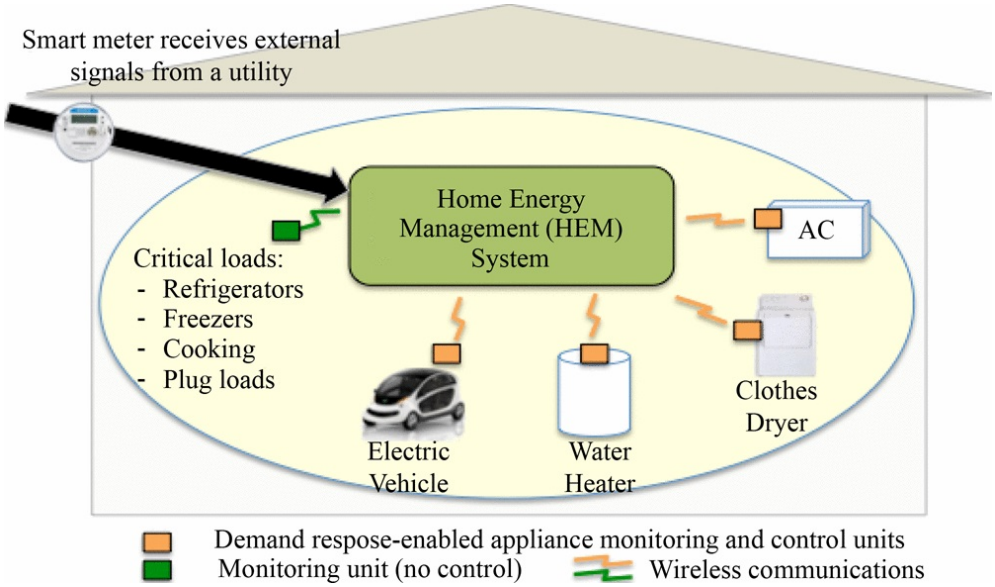

Figure C.3: Overview of a potential HEM system[89]

The growth in electrical vehicles is a massive opportunity for extra savings and convenience by coordinating EV charging with regards to consumption patterns and whole sale market prices.[123] Also demand response in residential markets avoid potential distribution transformer overload problems.[89] Subsequently as energy storage and micro-generation becomes even more cost effective, equipped homes will be provided the ability to become more self sufficient, enabling them to store energy when it is available and cheap(further discussed in (ref something).[123]

The market for home automation is growing rapidly predicted to reach a revenue of $22,5 billion in 2018. This is much due to the Internet of Things lead by smart thermostats. Increased connectivity and interplay between devices and the cloud creates new pricing models, leads to alliances. There is also an increasing degree of mergers and acquisitions between IT, energy companies and Internet of Things start ups, with Nest being the biggest acquisition so far.

C.3.1 Technology critique

There are some drawbacks to demand side automation that Pipattanasomporn et al. [89] points out:

- Customers may need to sacrifice comfort level

- High off-peak demand due to load compensation
- May not reduce baseline consumption