
Big
Data
&
Privacy
by Emil Granrud Gabrielli

Assignment for the course: “ACIT 4100: Research methods and Ethics” at OsloMet

By Emil Granrud Gabrielli
Student number: s341473

Date: Saturday, 7th of December 2019.

Front-page image composited by Emil Granrud Gabrielli.
For images used in the composition, see reference list.

Table of contents

Introduction

When we close our browser, locks our phone or shuts down our computer, is everything that we just did online just wiped out like it never happened? When visiting an online store for the second time, does that website thinks of us as a brand new visitor? The obvious answer would be “no”, and most of us are aware of this. When we are done using the internet for the day, we might also forget this or not just think about it, but unlike the footprints we leave in the sand, our digital behaviour does not simply disappear.

The amounts of data that we leave behind us and are collected brings issues to the table when it comes to privacy. What can these data be used for, and how big of an impact can it bring? More importantly, what is being done to preserve privacy when collecting these sensitive data? I will attempt to answer these questions in this paper by first introduce you to what Big Data is and then talk about the impact of what Big Data can do by using the Cambridge Analytica scandal as a case study. To go into more details on how Big Data becomes a more wider area, I will talk about the emerging use of Internet of Things (IoT) devices that capture data. I will end this paper by discussing the challenges and laws of preserving privacy in Big Data analytics and talk about how the previously mentioned scenarios might have prompted legislators to find methods to preserve privacy in Big Data.

How is Big Data any different than traditional data sources?

Bill Franks defines some ways in how Big Data is different from traditional data sources in his book *Taming the big data tidal wave*: “While big data certainly involves having a lot of data, big data doesn’t refer to data volume alone. Big data also has increased velocity (i.e., the rate at which data is transmitted and received, complexity, and variety compared to data sources of the past.” (Franks 2012, 5). We can therefore define Big Data with three terms: *Volume*, *Velocity* and *Variety*. Later, I will go into detail in what each of these terms mean, see *The three V’s* in *A brief lesson in Big Data*.

The visible side of Big Data as a consumer, two scenarios

You might have noticed that when you visit your favourite social media platform, an ad for the exact same television or computer that you just checked out on “*SaulsElectronicStore.org"* appears, but this ad is for “*Saul”*’s competitor which sells it much cheaper. Is this the workings of a *Big Brother*? Perhaps a coincidence? The truth is that what you have just experienced is a mixture of Big Data and targeting advertisement in practise.

Have you ever got one of those phone calls from your phone company which offers you a great discount on this new service they are launching, you will just have to bound yourself to that service in a period of six months before you can cancel? We can speculate that your profile showed that you have a high likelihood of changing your mobile phone carrier, so they give you a good offer as an effort to make you stay as their customer. This is called *churn*, and is a word used to describe customers leaving a company, and a model that predicts which customers will have a high likelihood to leave the company is called a *churn model*, or also known as a *churn prediction model*.

Big Data in the industry

As a company, using Big Data analytics to really understand your customers will gain a huge insight in how to advertise more effectively and how to gain new customers. Now that I have covered how a consumer may notice Big Data, I will talk about some examples that the industry may use Big Data.

Churn prediction

As mentioned previously, churn is when a customer switch from one company to an other. A practical examples of churn for Telenor would be that a customer decides to cancel their mobile carrier with Telenor in order to be a customer of Telia, because they might have a special offer.

The intention of a *churn model* is to flag customers which have a risk of churning / cancelling their membership or subscription. This allows the company to try to hold on to their customers by maybe offering discounts or other special perks.

Sentiment analysis

By analysing text data, sentiment analysis tries to know what people are thinking about an organisation or company. Getting data from social media sites is a good source to use with sentiment analysis by analysing the sentences in which the name of the company is mentioned. Examples of questions that might want to get answered with sentiment analysis is:

- Does people like our newly launched product?
- Does people talk about us in a good way, or do we have a bad reputation?
- What are some keywords that are associated with our company?
- What are some words commonly used with our company?

Sentiment analysis can be great for a company to get a valuable insight into their reputation amongst their customers. Also, a company could potentially recognise if some rumours surrounding a new product generates some hype among their customers and what their expectations for that product are like.

Recommendation systems

When you as a customer purchase something online like a book, movie or music, you will probably see something like: "Other customers also purchased…", or at checkout you will see other recommended products that you might also like. Maybe your music or movie streaming service recommends what you should watch or listen to next. Recommendations in these systems can be based on general opinions and trends, or based more on your personal taste after what you have watched, listened to or purchased.

These systems are usually there as an effort to try to keep you on their site or to make you purchase more items from them before you finish your order.

Other areas where Big Data is being used

Big Data might be mostly used in marketing and to give companies a greater insight about their customers and users, and even though we will mostly be focusing on that aspect and on how Big Data is being used in political situations, there are many more areas where Big Data makes a great impact. I will try to give examples of some of these areas to show you how Big Data can be used within health care.

Mental health: suicide prediction

Among people in the ages of 15 to 29 years, suicide is the second leading cause of death. Every year close to 800 000 people loose their life to suicide (World Health Organization). Big Data has found its way into mental health in models that try to predict suicide. In the book *Personalized Psychiatry*, the authors discuss around these machine learning methods used. They say that these models can be improved to be more accurate in their predictions and go on to propose a process that would involve additional machine learning analyses (Passos et al. 2019, 91-92).

Medicine

Within medicine, complex algorithms can try to predict outcomes by learning from huge amounts of data. According to an article by Obermeyer and Emanuel, if we take the example of a chest x-ray, models can go through each pixel of the image and learn how to recognise signs of, for example, fracture lines (2016).

Disease control in Pakistan using Big Data

Researchers from Harvard, Telenor, Oxford, Centre for Tropical Medicine, Centres for Disease Control and University of Peshawar conducted a research on the emergence of dengue viruses in Asia and the Middle East. They wanted to predict the spread of the disease in Pakistan by using data based on climate and mobility data from about 40 million mobile phone subscribers (Wesolowki et al. 2015).

Conclusion: Introduction

In the first leg of our journey, I have presented to you what this paper will be all about. I have barely covered how Big Data is different from other data sources by quoting the definition made by Bill Franks in his book *Taming the big data tidal wave* (Franks 2012, 5). To show the value of Big Data, I have given some examples on how Big Data can be used by the industry to improve their marketing strategy and how these might be visible for us as customers. I have emphasised that Big Data can be used in other ways instead of just as a tool for companies for financial gains. Big Data can be used for mental health prediction, disease control and within medicine to recognise fracture lines in x-rays.

A brief lesson in Big Data

Talking about Big Data requires you to know something about Big Data. To make sure you follow, I will explain some of the terminology and concepts of Big Data. This section is heavily based on what I learned from the web course *INNI3012 Big Data* at NTNU held by Xiaomeng Su and Nils Tesdal (2019). This course gave me a good introduction to the main concepts of Big Data without diving too much into the deep and complicated concepts. By the end of this section, I hope you will have a basic understanding of Big Data and might already see some ways that Big Data can be used that threaten an individuals' privacy.

The term "Big Data"

Simply put, the term *Big Data* is often used when we want to draw a conclusion or to get a valuable insight in a big dataset. What differentiates the term Big Data from other known terms such as *Business Intelligence* is the fact that the data is so huge and complex that we require special algorithms and technologies in order to draw any conclusion from these huge datasets. Traditional methods to work with datasets like within *Business Intelligence* would therefore not be good enough.

The three V's

As mentioned previously in *How is Big Data any different than traditional data sources?* we use three V's to explain the data we have to face when working with Big Data. Thinking about these three V's to explain Big Data makes it easy to remember and will really help to paint a picture of what Big Data is all about.

Velocity

When talking about velocity, we talk about data coming at us in high speeds. As a result of high refresh rates we can receive data in real-time from many sources and at huge amounts. This often results in short time to respond in order to process these data effectively. Just think of the refresh rate of one sensor. Having more sensors would result in a lot of data transferred incredibly frequently and quickly. We therefore have to know how we can process these types of data.

Variety

Will our data all be True/False or just contain one integer? No. Variety acknowledges that our data can be in any shape or form. These may be structured data such as a database table, unstructured data like PDFs, photos or chat logs, or maybe even semi-unstructured data like XML, JSON or CSV files. Our Big Data model will have to know how to process these files in order to gain any insight to what we want to know.

Volume

Perhaps what most people are thinking when hearing the words *Big Data*; volume is the huge amounts of datasets that we have access to. These datasets can be as huge as petabytes and require huge storage spaces and technologies to process.

Data sources

The data collected comes from many different sources and, as mentioned above, comes in many different formats and types. Using many different sources of data can create big and complex profiles on any user, which can be used for marketing or to get a huge insight in any users behaviour online or in the real world.

Time and location data

Using the GPS on your phone and knowing the location of your Wi-Fi connection makes it easy to know where you connect from. Sometimes these data can be based on the location of your IP location or to be as accurate as GPS coordinates.

From the perspective of a company, knowing your location gains huge value. A company might recommend that you visit their store located just three blocks away, or that the majority of your customers that live in Kristiansand tends to go through a certain street downtown that might be the ideal place to build a new store.

It should go without saying that this type of data gathering is very sensitive when it comes to privacy. Users are literally being tracked. Using this type of data gathering should therefore be treated very carefully.

Social network data

Digging deep into a company's customers social networking profiles, like Facebook and LinkedIn, will make it aware of the network that you are a part of and what you like. Knowing what interests you, which ads and sites you have clicked *Like* on and who your friends are will decide which ads you see on that site. Furthermore, this will give companies a good insight in where they are the most popular, then they can decide which platforms they should target marketing on the most.

Web data

Knowing which sites a user visits, what that user searches for and what is purchased are tracked to gain an insight in the behaviour of a user on the web. A company can use this to their advantage to recommend products at checkout or to enhance their targeted marketing model.

Companies may use a third-party to do the tracking by placing ads on websites that registers which sites you visit and see the ad. Franziska Roesner did a great presentation on this subject where she talks about how a third-party can track your web traffic by showing you ads on, for example, The New York Times and CNN. This third-party, which is responsible for the ads, now know that you visited these sites. Franziska talks more about this and how to detect and defend against it in her presentation *Detecting and Defending Against Third-Party Tracking on the Web* (Roesner, Kohno and Wetherall 2012).

Text data

This data can be gathered from sources like emails, documents, Facebook feeds and chat logs. The focus on text data is to extract key components and use these as input for an analytic process. This might be a good source for data to be used for sentimental analysis.

Sensor data

Sensors are all around us. You have some sensors in your car and in your phone. These sensors sends out data rapidly in order to capture problems right as they happen, or the sensors might capture the movement of your cellphone and know when it is time to rotate the screen. From this rapid stream of data we can read measurements in real-time.

When writing about the data sources, I thought about a scenario that I think would be interesting to share. If a company plans to launch a new campaign, and they decide to test its effectiveness, they could test this by putting up posters in specific areas in different cities. They could then measure where each new page visit are from and see if the visitors location is close to that of a poster.

Analytical techniques

There are many analytical techniques to use when analysing the data gathered from Big data. Here I will cover two of them.

Summary statistics

We can use *summary statistics* to determine a sum, an average or to show rates of some sort. Questions like “Does our churn differ from the ages 18-24 and 25-32?” and “How many of our customers are likely to purchase our new service?” can be answered using summary statistics.

Database querying

A database is an effective way to keep an overview of data and attributes. If we store our users in a database, each user will have fields related to their age, gender, location and many other fields that will make each entry detailed and informative. By sending a query to that database we can get back data we want to take a closer look at.

We can use tools that will analyse the database and return the data and information we want. Questions like “Where is our most profitable customers that is 18 years old located?” can be answered by running a query to retrieve a list of the most profitable customers at the age of 18 sorted by location or profitability.

Algorithms and models

In order to draw any conclusion of these huge datasets, we create models with the use of algorithms. What algorithms and models we use is entirely up to what you want to know. Maybe you work for an internet service provider and want to know which customers are in danger to cancel their subscription. You therefore create a CHURN-model that uses the K-nearest Neighbour algorithm which will calculate the probability of a customer to switch to an other provider. By knowing what algorithms and models to use with the data we can gain the insight that we want to have. I will not go into too much detail on the algorithms, but I will talk about the one I know the most about: KNN (K-nearest neighbour). I will also mention how a Churn model works, and what Cluster analysis is.

K-nearest neighbour

When getting to know how the algorithm works, the name would seem more clear than what it may seem at first. Simply put, K-nearest neighbour (KNN for short) tries to predict what group a new instance is a member of. The algorithm does this by getting the *nearest neighbours*, which is the instances with the shortest distance from our *new* instance. Our new instance would then be classified as belonging to the class that has the most *nearest neighbours*. This is just a simplified explanation on how KNN works, it can get a lot more complex. To see a visualisation about K-nearest neighbour, see Figure 1.

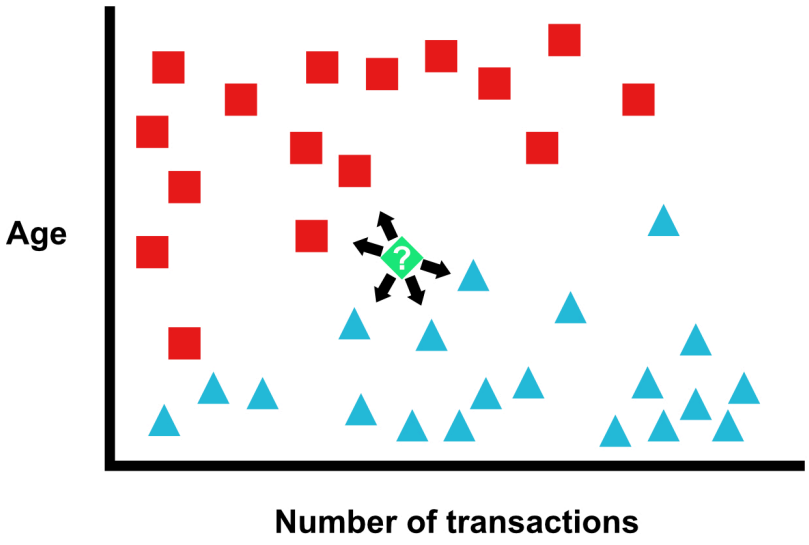


Fig 1: Here we have a very simple model that tries to classify users using two metrics: age and number of transactions. K-nearest neighbour tries to classify the new instance, marked here with a question mark ("?") in a green diamond shape, by first placing it in our model. Then, five closest neighbours are chosen, which in this case is three triangles and two squares. Our new instance will then be classified as belonging to the triangles. ***Disclaimer: This is a fictional figure with no accurate metrics and is meant to only display how KNN works.***

Churn model

A churn model have traditionally been relying on historical data to see what characteristics a person that churns have. Maybe a certain demographic have a higher chance to churn, or perhaps annual income matter. Having a model that can recognise the traits of a person that churns, customers can then be examined against this set of characteristics to see if they have a high likelihood of churning. Let us also imagine that the company notices that some users have been visiting their cancellation policy web page, that they have been visiting stores belonging to a competitor and complained on the company's facebook page. Adding this information to the comparison will really improve their churn prediction.

Cluster analysis

Using cluster analysis, we can try to group entries into common groups that we call clusters. Within these clusters, every entry is closely related to each other and this can help to identify traits within these groups. Maybe members of one cluster is more likely to enjoy a product more than other clusters, so targeted marketing to that cluster will be more effective than marketing that product to other clusters.

Conclusion: A brief lesson in Big Data

In this section of the paper, I have given you a lesson in some concepts about Big Data. After reading this section you should have a good understanding on how Big Data is different from other traditional data sources. I also covered some ideal sources to get data which can be used in algorithm based models to gain insight or to draw a conclusion. By now you should see why Big Data is so valuable to companies, maybe even speculate on some major privacy issues.

Cambridge Analytica

To show just how big of an impact Big Data mixed with targeted marketing can make, there might not be a better example than to go into details in the Cambridge Analytica scandal. When you hear about Cambridge Analytica, you might think about the company's impact in the digital marketing campaign for 2016 Presidential Candidate: Donald Trump. The digital marketing campaign was called Project Alamo. I will in this section talk about Cambridge Analytica's contribution to Project Alamo, but also mention its contribution in the election in Trinidad and take into consideration how Big Data was a big part of it. Lastly, I will briefly talk about the repercussions of this scandal.

In this section you will notice that I talk a lot about the Netflix documentary *The Great Hack* by Karim Amer and Jehane Noujaim (2019). The documentary turned out to be a more valuable source of the Cambridge Analytica scandal than I expected. When I tried to research some topics, like what role Cambridge Analytica had in the Trinidad election, not a lot of research was published on the subject, but the documentary inspired some articles to be published on news websites. This could be explained by the fact that the Cambridge Analytica scandal is very recent, or that the documentary revealed new information about Cambridge Analytica by talking directly to former Cambridge Analytica employees.

What was Cambridge Analytica?

Cambridge Analytica was a consulting company who specialised into helping politicians and political parties. They were famous for using techniques in Big Data to get a detailed insight in how to come up with the best marketing strategy to the benefit of who they were working for. How accurate that description might be, I think the CEO of Cambridge Analytica, Alexander Nix, said it better: *"We are a behaviour change agency. The holy grail of communications is when you can start to change behaviour."* (Amer and Noujaim 2019, 1:01:03).

Project Alamo

In the documentary *The Great Hack*, they show a clip of the CEO of Cambridge Analytica, Alexander Nix, talking about their methods. He talks about their model having between four to five thousand data points on every American voter, which they can use to predict each individuals personality (Amer and Noujaim 2019, 14:10). This was for the campaign for Ted Cruz which led Cambridge Analytica to gather huge amounts of data for each American voter, which they transferred over to the digital marketing campaign for Presidential candidate Donald J. Trump, called Project Alamo.

Access to data

In a paper published by Jim Isaak and Mina J. Hanna called *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection* (2018), they write that researchers from Global Science Research (GSR) were working with Cambridge Analytica to develop "O.C.E.A.N." profiles, which stands for:

- **O**penness
- **C**onscientiousness
- **E**xtraversion
- **A**greeableness
- **N**euroticism

They were going to create these profiles through a personality quiz where the quiz required each participant to give Global Science Research (GSR) access to their Facebook profile. By using Facebook Open API, this would also allow GSR access to the data of that participants Facebook friends (Isaak and Hanna 2018).

As mentioned previously in *A brief lesson in Big Data*, one common source of data is through social media. Cambridge Analytica created "O.C.E.A.N." profiles on as many American citizens as

possible and linked this with what that profiles' activity on Facebook was like. Also mentioned in *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection,* using this information and integrate it with information they gathered from many other sources such as voting results and online purchases, Cambridge Analytica were now sitting on more than 5000 data points on 230 million adults in the US (Isaak and Hanna 2019). With so many data points on so many adults in the US, they could predict how they would vote, or as Alexander Nix put it: *"Because it's personality that drives behaviour, and behaviour that obviously influences how you vote."* (Amer and Noujaim 2019, 14:22).

The digital marketing campaign

In *The Great Hack*, a clip is shown where Alexander Nix talks about their strategy when having such a large set of data points: *"We could then start to target people with highly-targeted digital video content."* (Amer and Noujaim 2019, 14:26). In the case of Project Alamo, Cambridge Analytica claimed to be responsible for coming up with the slogan "Stop Crooked Hillary", which tried to turn people from voting for Hillary Clinton and instead vote for Donald J. Trump.

According to a paper published by Jesse Gordon, a senior campaign official said that the digital team that worked on the Trump campaign had unique strategies for three demographics they focused on in order to have them not to vote for Hillary Clinton (2019). The paper uses an article published in Bloomberg by Joshua Green and Sasha Issenberg, which interviewed a senior official (2016). Among the demographics that the campaign focused on was idealistic white liberals, and two more demographics:

African American voters

In order to turn African American voters, ads was made using a remark Hillary Clinton made in 1996 where she suggested that African Americans had violent tendencies. One strategy was to run an ad on some selected African American radio stations.

Through Facebook, an animation in the style of *South Park*, was delivered to certain African American voters that used the same audio from 1996 where Hillary Clinton calls African American "Super Predators" with violent tendencies (Green and Issenberg 2016).

Young women

To make Hillary Clinton be less appealing to young women voters, the digital campaign team focused on women that said they were sexually assaulted by Bill Clinton and harassed or threatened by Hillary Clinton (Green and Issenberg 2016).

An article published in *Anthropology Today* written by Roberto J. González quotes the founder of Bitcurrent, Alistair Croll, saying:

> *"After Eisenhower, you couldn't win an election without radio. After JFK, you couldn't win an election without television. After Obama, you couldn't win an election without social networking. I predict that in 2012, you won't be able to win an election without Big Data."* (González 2017, 9).

He might have been on to something. Those responsible for the digital marketing campaign for Trump spent one million dollars **per day** on Facebook ads when Project Alamo was on its peek, (Amer and Noujaim 2019, 8:30). Additionally, Alexander Nix mentions in his presentation at Concordia Annual Summit in New York that Cambridge Analytica helped Ted Cruz go from the lesser known candidates to the last one standing to challenge Trump for the nomination (Nix 2016). This claim is backed up by the fact that according to an article from the Washington Post by Tom Hamburger, the Cruz campaign thanks Cambridge Analytica for its success (Hamburger 2015).

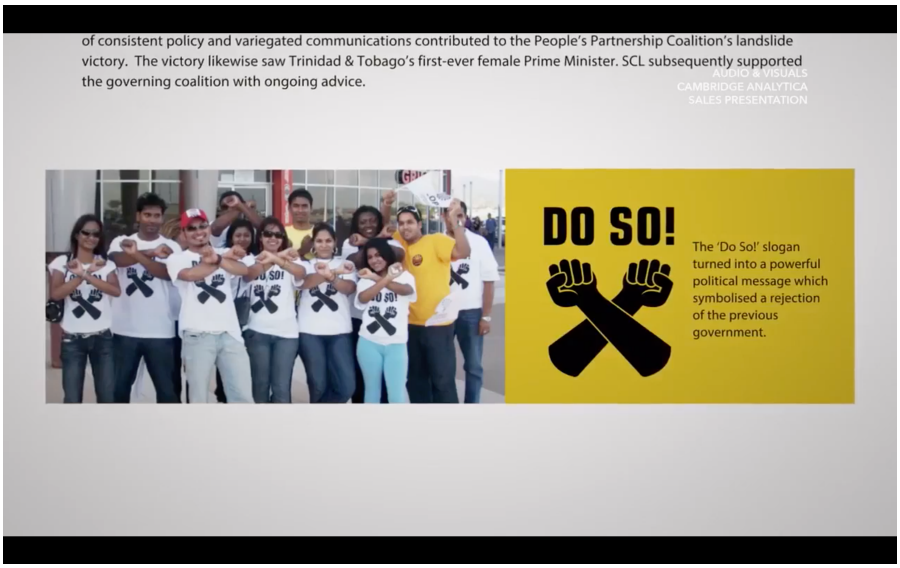
Trinidad

Alexander Nix uses Trinidad as an example when talking about Cambridge Analytica being a “behaviour changing company" in a soundbite shown in *The Great Hack* (Amer and Noujaim 2019, 01:01:15). He explains that in Trinidad there are two main political parties; the United National Congress Party (which Nix refers to as “the Indians”) and its major opponent called the People’s National Movement (which Nix refers to as “the Blacks”). Alexander Nix go on to talk about how they were hired to help the United National Congress Party to win the election.

The “Do So!” campaign

Cambridge Analytica’s strategy in Trinidad was to reduce the interest of voting for the demographic that would vote for the opposing party. They came up with the *Do So!* campaign, which was meant to be a protest against politics and voting by simply saying: “Do so! Don’t vote.”. The campaign was successful in increasing voter apathy according to Nix: “… and the difference in 18 to 35 year-old turnout was like 40%. And that swung the election about 6%, which was all we needed in an election that’s very close.” (Amer and Noujaim 2019, 01:03:10).

This campaign focused on creating a trend specifically among the opposing party’s demographic where you had to “Be part of the gang. Do something cool.” (Amer and Noujaim 2019, 01:01:55). According to Alexander Nix, the campaign had to be non-political and reactive because their targeted demographic was “lazy” and did not care about politics. (Amer and Noujaim 2019, 01:01:45).



Screenshot from the Documentary *The Great Hack* showing some campaign material related to *Do So!*. The screenshot shows a group of people forming a “X” with their fists, which represents the logo for *Do So!*. The screenshot also shows the logo for the *Do So!* campaign (Amer and Noujaim 2019, 01:02:05).

In an article posted on Global Voices Advox by Jada Steuart, the political party denies hiring Cambridge Analytica (2019). Furthermore, the article explains that most of the citizens thought that the *Do So!* campaign was not something put into action by Cambridge Analytica or the political party involved. They thought instead that the campaign was inspired by the actions of a senior citizen called Percy Villafana, who would cross his arms to deny the 2010 prime minister from accessing his property when he was going around visiting people in an effort to collect votes (Steuart 2019).

The scandal

What Cambridge Analytica was doing did not go unnoticed and concerns were raised questioning if what Cambridge Analytica was doing were right. Even though it took a while before the scandal blew up entirely, some articles were published early on which criticised how Cambridge Analytica gathered and used data.

Harry Davies

When Cambridge Analytica gathered data from Facebook to use with the Ted Cruz campaign, Harry Davies wrote an exclusive for the Guardian about the issue. In the article, he writes that the psychological data gathered from "tens of millions" of Facebook users are gathered without their permission (Davies 2015).

Further in the article, he goes on to explaining the ethical and privacy issues related to building these complicated models based on psychological data gathered without the users permission. Harry was also aware that when a user was taking a survey, which they had to log into Facebook to take a part off, all the information about that users' friends on Facebook was also gathered. He calls this "seeding" and goes on to explain the process of harvesting data this way (Davies 2015).

Carole Cadwalladr and Channel 4 News

In *The Great Hack*, Carole talks about how she started looking into how Cambridge Analytica was involved in the Brexit campaign, which tried to convince the people of Britain to leave the European Union (Amer and Noujaim 2019, 16:36). Carole got in contact with Christopher Wylie, an ex-employee of Cambridge Analytica, which confirms what Harry Davies had written about in the Guardian. Wylie got in contact with Professor Aleksandr Kogan at the University of Cambridge, which offered him apps on Facebook that would not only gather the information on that user who gave it permission, but also gather data from that users friends. Wylie says that among the data that was gathered was *status updates, likes* and *private messages* and that this data was used to build psychological profiles (Amer and Noujaim 2019, 00:22:00).

The scandal did not fully escalate before Cadwalladr brought in Channel 4 News and The New York Times that published undercover videos of meetings about the Trump campaign with Alexander Nix. Channel 4 News launched a series called *Data, Democracy and Dirty Tricks* by Channel 4 News Investigation Team (2018). In the undercover video by Channel 4 News, Alexander Nix can be heard talking about their methods which may include faking a corruption video that could be posted online in order to shame any competition (Channel 4 News 2018, 13:42).

This scandal led to the suspension of Alexander Nix and Facebook getting a fine on five billion dollars (Julia Carrie Wong, 2019). It also led to many questions around our online privacy. I will talk more about how the Cambridge Analytica scandal may have inspired politicians to look into ways of keeping privacy in Big Data in the section called: *Challenges and laws in keeping Big Data privacy*.

Conclusion: Cambridge Analytica

In the two scenarios I have talked about; Cambridge Analytica's involvement in *Project Alamo* and *Do So!*, I have drawn attention to two different methods that Cambridge Analytica used in order to swing an election in favour of their client. Cambridge Analytica used voter data gathered from their workings on the Ted Cruz campaign to Project Alamo in order to know which demographics to target in order to get people to vote for Donald Trump instead of Hillary Clinton. They created "O.C.E.A.N." profiles to get a good insight in the behaviour of each American voter and link this information to their Facebook activity. This allowed Cambridge Analytica to create specific digital video content to show specific groups of people in an effort to convince them that it would be better to vote for Trump instead of Clinton.

The *Do So!* campaign is a contrast to Project Alamo. Instead of trying to convince people to vote for the United National Congress Party, they wanted to promote voter apathy in the opposing political party by

starting a campaign that was supposed to be a protest against politics and voting. The campaign encouraged that people could protest simply by abstaining from voting.

A thorough investigation done by Carole Cadwalladr and Channel 4 News, with the help of Cambridge Analytica ex-employee Christopher Wylie, exposed Cambridge Analytica's dangerous methods and how it is a threat to our privacy. This resulted in the downfall of Cambridge Analytica and a huge fine for Facebook (Ballhaus 2018).

I will end this section with a quote from Christopher Wylie, which I think best describes Cambridge Analytica's use of Big Data analytics and why it was so dangerous:

> "Throughout history, you have examples of grossly unethical experiments. […] I think that, yes, it was a grossly unethical experiment. You are playing with the psychology of an entire country without their consent or awareness. And not only are you, like, playing with the psychology of an entire nation, you're playing with the psychology of an entire nation in the context of the democratic process." (Amer and Noujaim, 2019, 23:06).

Internet of Things (IoT) devices

Small devices connected to the Internet has been emerging more and more. These devices collect data in order to provide feedback or to change their behaviour to be more suited the user. In this section I will talk about these types devices. I will start by explaining what a IoT device is and give examples of IoT devices you may have at home.

Internet of Things

In the paper *Big Data Privacy in the Internet of Things Era*, the authors explain that IoT defines a massive amounts of objects, sensors or devices connected through a network of networks: *"The IoT connects people and things anytime, anyplace, with anything and anyone, […]"* (Perera et al. 2015).

Internet everywhere around you!

The amounts of devices that you can purchase that needs an internet connection is huge. We can take a quick look around us and see how devices are connected to each other with the use of internet. If you have two or more devices from Apple you might think of Apple's ecosystem, or the Philips Hue smart lighting system. Some more examples on IoT devices:

- Smart speakers
- Smart doorbells
- Smart watches
- Smart plug adapter
- Smart fridges
- Wi-Fi smart led light bulbs / systems
- Home surveillance systems
- Temperature and humidity measurers
- Media streaming devices (like Google Chromecast and Apple TV)

So, we got some devices around us that are connected to the Internet. The issue here is that these devices are very personal and might be harbouring very private data on us. Well, knowing the temperature and humidity in my house or how many times I open and close my fridge might not be the biggest issue if that data was leaked, but if someone were given access to a home surveillance systems that is more serious. I will discuss these issues more in the *Challenges and laws in keeping Big Data privacy* section.

Your own IoT with Arduino and Raspberry Pi

Even though I have mentioned IoT devices you can purchase, I would like to briefly talk about how anyone can create their own IoT devices at home as a creative developer using Arduino and/or Raspberry Pi.

Using Arduino, being a microcontroller, you can create interactive devices where only the sky is the limit. Maybe you would like to create your very own automatic cat feeder (Rundhall 2019), or maybe control access to something using a RFID Reader (Mukherjee 2016)? Since Arduino is highly customisable, programming it to collect data on every use or to collect GPS data with its NEO-6m module is very doable.

Raspberry Pi is a small but powerful computer with a OS based off Debian (a Linux distro) called Raspbian. Its area of uses are, like Arduino, very wide. Since Raspberry Pi is a computer we can also treat it like a server, which opens up more possibilities within Big Data. A Raspberry Pi can also be used to monitor a network (Cawley 2015).

If you are interested in collecting data with the devices you create on your own, you are controlling what data is being gathered and how they are treated. Therefore, no privacy is in violation unless someone would get access to your data, but you are in charge of your very own security as well.

Conclusion: Internet of Things

In this section I have explained what Internet of Things are and where we can find these devices; they can be found everywhere. These devices has become a part of our modern times and are meant to make our technological life easier and more interesting. Going home and telling our smart speaker: “Turn the lights on in the living room” will make it send a message to our smart bulbs in the living room to turn the lights on. As we sit down in our couch, we might get a notice on our phone telling us that our home surveillance system registered motion by our front door as we hear a ring from our smart doorbell. IoT can be a lot of fun, and we can suspect more and more devices to appear.

Challenges and laws in keeping Big Data Privacy

Up until this point I have presented some cases where we can find Big Data that might affect us. With Cambridge Analytica, huge amounts of data was collected and used to create models that would find personality traits and linked this to that individuals Facebook account in order to predict how that person would vote. With Internet of Things devices emerging and has become a part of our daily lives, we might find that data on us are being gathered and used in ways we have no control over.

There are some issues that need to be addressed in order to maintain security and preserve privacy when it comes to Big Data. For those working with IT, these privacy issues are clear and there might not be any doubt that data should be kept secure. In this section I will first briefly talk about the peoples' opinion on their personal data and how Cambridge Analytica may have inspired politicians to address the issues surrounding privacy. Then I will talk a little about privacy issues related to emerging IoT devices and then about general privacy challenges related to Big Data. I will within these topics talk about laws that may have emerged as a result.

Do we value our data?

Charith Perera and Arkady Zaslavsky conducted two surveys when it comes to privacy in IoT devices. In the first survey they asked what peoples opinion was about a model called *"*Sensing as a service*"*. This is a fictional model taken as an example to see how people view the value of private data captured through sensors in IoT devices. As they explain: "Sensing as a service envisions a marketplace in which contextually enriched sensor data can be exchanged between different parities for financial or social benefits" (Perera et al. 2015).

In the first survey, the participants were to assume that the data was 100 percent secure and that each individuals' privacy was guaranteed. 137 people participated in the survey, and the majority would be in favour of such a trading-based model. The survey went into more detail as to how much money one would expect the owners of the IoT solutions to sell your data for. 67 percent said that they would expect less than 500 dollars of value each year. They also discovered that such a "marketplace" would motivate 65 percent of the participants to purchase IoT devices in order to be a part of this (Perera et al. 2015).

What was interesting was that in the second survey, without mentioning that privacy is assured, they asked 1000 people if they themselves would be interested in exchanging personal data for a financial return. 79 percent of the participants was negative to that idea (Perera et al. 2015).

We can say that people are aware of privacy when it comes to private data. This fact can be backed up by a survey done by TRUSTe, which says that about 60 percent of internet users in the U.S. and 47 percent of British users are aware the smart devices collect data on their personal activities (TRUSTe 2014).

Cambridge Analytica

An article was published in the Nature Science Journal that addressed the fact that the Cambridge Analytica scandal reveals a need for research to be done on the subject of personal data usage (Nature Science Journal 2018). In the article, they write that Facebook has restricted the harvesting of data by third parties after the Cambridge Analytica scandal. However, in academic research these types of data can be very useful, especially in medical and psychological studies. Therefore, in the Nature article, they suggest that we should redefine what data counts as *public* data. We can indeed see that after the scandal of Cambridge Analytica, action to protect individuals privacy online is made.

Principles in Privacy

In the article *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection*, they suggests what a legislation of privacy and data protection should include in order to protect and ensure privacy for the user (Isaak and Hanna 2018). This includes the following topics, which is

based on the position statement from IEEE-USA Digital *Personal Privacy, Awareness and Control* (IEEE-USA 2018):

Public Transparency
This section talks about how information about the data gathered from you and the mechanisms used to collect these data should be informed to the user. Furthermore, information on how that data is kept and shared with third-parties should also be disclosed. If the website or application places any files or content (like *Cookies*), the users are to be informed that content has been placed on that users' device and be explained to what the use of that content is.

Complete disclosure
Every user should have the possibility to obtain the information stored on them by the website itself and any third-parties that have access to that information.

Control
Users must have the ability to request not to be tracked, which must be respected. Users must have the ability to delete any personally identifiable data and easily delete and uninstall any content that is placed on that users' device. If a user gives consent to that website to collect data from that user, then that consent can not be interpreted to include the collecting of any other users.

Notification
The company or organisation have to inform directly to all of the users if their data had been lost or misused and the users have to get as much information on the misuse or violation as possible, which includes the parties that were responsible for violating the privacy.

For paid advertisement and content shown on any site, a notification should be seen that informs the user that this is paid content. This includes clear link to the source of the ad which have more information on parties involved in the advertisement.

S.2728 - Social Media Privacy Protection and Consumer Right Act of 2018
A bill introduced to the senate by Senator Amy Klobuchar in 2018 says that online platform operators are obligated to inform users of what personal data is created on them and what it is used for by the website or any third parties. This information should be presented to the users before they create an account on that website (Klobuchar 2018).

Furthermore, the bill says that the operators of the online platform will have to offer a copy of any users' personal data, free of charge. If the personal data has been somehow leaked or misused, the operators should within 72 hours inform the users that were affected by this (Klobuchar 2018).

This bill takes into account many of the topics that are considered vital in keeping our personal data private, and we can see similarities in what is found in this bill and the position statement from IEEE-USA.

General Data Protection Regulation (GDPR)
The European Union passed a regulation put in effect on May 25th of 2018 that will ensure the protection of personal data online for the people living in the EU (GDPR.eu). The regulation itself is long and an intensive piece of reading, but is available at *EUR-Lex.europa.eu* [1].

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

Mark Kaelin wrote an article in TechRepublic that tries to summarise the key concept of the GDPR. It draws a lot of similarity to the *Social Media Privacy Protection and Consumer Right Act of 2018* and the topics discussed in *Principles in Privacy*. A noticeable different is that the GDPR do not allow usage of long and complex terms and conditions statements. Every request of consent to retrieve private data from any user must be clear and concise (Kaelin 2019).

After Cambridge Analytica, we have seen laws emerging that is based on some principles in personal data privacy. Key concepts of these laws does not base themselves on the mechanics on getting or using the data but instead focusing on informing the users about what data is being gathered and used and informing when violations occur. Basically informing the user what they are agreeing to when visiting that website. These legislations also focuses on the importance of each users ability to gain an insight into their own data and be able to delete the data that the company has on them.

"COPPA"

Something that seems very peculiar is that we can see these laws about online data privacy emerging in the last few years, but as early as in 1998 privacy online was considered an important issue. The Children's Online Privacy Protection Act ("COPPA", for short) is a set of requirements that must be followed by websites and services aimed for children under 13 years in order to keep their information secure online (Federal Trade Commission 1998). This act was first enacted at October 21, 1998 (Federal Trade Commission 1999).

VTech data breech

The Federal Trade Commission filed a lawsuit against VTech for not following "COPPA" when it was made clear that sensitive information regarding children were not kept secure. After a data breech in 2015, the hacker was easily given access to data on more than 4,8 million users and sensitive information like passwords, IP addresses, physical addresses and ZIP codes (Hunt 2015).

VTech was given a penalty of 650,000 dollars and were required to have "a comprehensive data security program subject to independent every-other-year audits for the next 20 years." (Federal Trade Commission 2018).

IoT devices

Users online might be aware that when they sign up for free services like social networking and newsletters, they will become a valuable source for data by that company or application for analysis of your online behaviour. We might also know that these data can potentially be sold to third parties (Perera et al. 2015). In the paper *Big Data Privacy in the Internet of Things Era*, the authors predict that service providers will adopt one of two models (Perera et al. 2015):

- Consumers might pay for preserving privacy when they consume services.
- Consumers might offer to give away data in order to consume services free of charge.

Furthermore, the authors explain some specific challenges when it comes to preserving privacy in IoT devices:

User Consent

In the same paper by Perera, Ranjan, Wang, Khan and Zomaya, they talk about user consent. User consent is about getting permission from the users, and nonusers who are affected by the device or services, to analyse the usage of the IoT device in order to analyse user behaviour (2015). The authors talk about how, on the web, users are presented with the privacy terms and policies that applies to the usage of that website. They go on to mention that one major privacy challenge when it comes to IoT is to ask for permission from the user to collect data in a way that is efficient. The challenge lies heavily on terminology; using the right words to describe what the users gives

permission to, because as the authors put it: “[…] every user has limited time and technical knowledge to engage in the process.” (Perera et al. 2015).

Control

Users should have the ability to have full control of their own personal data that is gathered from the IoT devices. This control includes capability to delete, review or move data. Unfortunately, the IoT solutions that exists in the marketplace provide limited access to users and do not grant any other insight (Perera et al. 2015).

Customisation and Freedom of Choice

Remember earlier when I talked about Apple’s ecosystem? This ecosystem allows Apple devices to talk easily with each other, and if you were to purchase a new Apple device then that device would automatically be a part of that ecosystem as long as you log in with the same AppleID. This might sound like you don’t have any saying in what you want to purchase, because it is easy to think that since all your stuff is from Apple and talk to each other, you would like to purchase your next smart speaker from Apple so that it can easily talk to all your other devices. This limits your Freedom of Choice, and is a privacy problem in IoT. “Moreover, users should be able to choose hardware devices and software components from different vendors to build their smart environments (for example, a smart home). This gives users full control and freedom of choice.” (Perera et al. 2015).

“… on a second though”

Users should have the possibility to change their consent, or remove it entirely. However, users should have an understanding that if the service don’t have access to necessary data, this could affect the performance of the service (Perera et al. 2015). For instance, a smart speaker is not able to process requests you say to it if you don’t give it access to use the microphone. But, the company that provide the service should not treat consumers unfairly by limiting functionality or changing subscription fees in order to motivate a user to provide consent that they would not otherwise give (Perera et al. 2015).

The recommended “onion” in IoT

The authors of *Big Data Privacy in the Internet of Things Era* recognise that there is a lack of technology within IoT that anonymise the users’ data communication paths (Perera et al. 2015). They say that user location can easily be tracked. They go on to say that newer IoT platforms will require technologies such as The Onion Router (The Tor Project) in an effort to conceal the users’ location.

General Big Data Privacy challenges

Until now I have covered some specific cases where challenges in gathering data occur, these challenges were mainly focusing on the lack of transparency from the company or organisation collecting the data from users. In this part I will talk about challenges we see in the general use of using Big Data.

Higher demand

In a feature published by Colin Tankard, he talks about the increasing demand in Big Data (Tankard 2012). He briefly talks about a research done by IDC in 2011 which discovered that 1.8 trillion gigabytes of data was created and replicated in 2011 alone. An issue that the research uncovered is that the information will continue to grow within the next decade (from 2011), by 50 times while the number of IT professionals will grow by just 1.5 times (Tankard 2012). “Data volumes continue to expand as they take in an ever-wider range of sources, much of which is in unstructured form.” (Tankard 2012). If we remember, unstructured forms of data can be the most difficult to process, these are data types like PDFs, photos, audio files or video.

Keeping the data secure

This issue is not exclusive to Big Data, but it is definitely applicable. The data being gathered on users are considered sensitive. Companies store data such as location information, passwords, physical address, e-mail, IP address and much more. These data is not being used one time and then deleted “beyond oblivion”, they are stored and remembered in databases. There are many ways that a hacker can access databases and systems, which VTech got to experience.

VTech data breech

As previously talked about, VTech suffered a data breech in 2015. Shortly put: VTech is a company that creates learning tablets for children, among other tech. A network hacker discovered that these tablets were talking to a website called *Planet VTech*. When he took a look at that website, he discovered that the site was vulnerable to SQL Injection (Rhysider, 2017).

SQL Injection

Considered by the *Open Web Application Security Project* as the number one security risk in 2017, SQL Injection exploits weak SQL code to inject malicious code to manipulate the application to do something it was not designed to do, like returning all information stored in the database (Ruiz 2018).

What made matters worse was that the encryption for the data kept by VTech was encrypted using a simple hash called MD5 with no attempt at *salting* or other complexity to the encryption, which made it easy to decrypt the information gathered by the hacker (Hunt 2015).

Keeping sensitive data secure is very important and there are many layers in order to secure the information being securely kept. Amongst other things you will have to limit access, check for code injection vulnerabilities and have a solid network architecture that will limit the access of a hacker if one would enter the network.

Re-identification attacks

Meiko Jensen talks about issues in Big Data privacy protection in his paper *Challenges of Privacy Protection in Big Data Analytics* (2013). He talks about an ability in big data analytics where scanning for correlations in a huge dataset can lead to a unique *fingerprint* of a single individual. He calls this *re-identification attacks*. He goes on to explain three sub-categories within *re-identification attacks*:

Correlation Attacks

A correlation attack is about linking datasets up and identifying a similar factor, like a user ID. If two datasets has two user IDs which are identical, we can find a correlation in these two datasets with one user and get to know more about that user (Jensen 2013).

Arbitrary Identification Attacks

This type of attack has the main goal to link one entry in a dataset to a human individuals’ identity (Jensen 2013). An example of this would be to see if a user ID has a real name associated with it.

Targeted Identification Attacks

This type of attack tries to find more information on a given human being. The attack is successful when an entry in a database is successfully linked to that human being (Jensen

2013). The starting point of the attack would be to first know who you want to find more information on and then see if there is an entry in the database that can be linked to that person.

Conclusion: Challenges and laws in keeping Big Data Privacy

I have presented to you my findings about issues and laws surrounding Big Data Privacy. Based on a study done by TRUSTe, people are aware of the data gathered by the devices we use (TRUSTe 2014). And that people most value their private data, even though some would like to sell their private data in order to gain some financial benefits (Perera et al. 2015).

After Cambridge Analytica, we have seen an increase in laws and regulations that requires transparency and informing the user about the data collecting done on them. Laws and regulations also require the application operators to allow the users to control what data is being gathered and the ability to delete their data from the application and third-parties that use that users' data.

Most privacy laws are new. One law is actually old. The COPPA protects the sensitive information being gathered on children below the age of 13 online, and VTech was charged by not following the regulations mentioned in COPPA when a data breech occurred in 2015. The hacker managed to gain access to a database containing sensitive information on VTech's users.

About IoT devices, the authors of the paper *Big Data Privacy in the Internet of Things Era* comes with an interesting prediction about future privacy models. They predict that consumers must either pay for keeping the data secure when consuming a service, or that consumers will give data freely in order to use a service for free (Perera et al. 2015). Much of the same issues arise around IoT devices as of Big Data; *Transparency*, *User consent* and *Change of consent*. However, IoT devices should also include technology like *Tor* (The Onion Router) in order to hide location and data communication paths (Perera et al. 2015).

We have issues surrounding the general use of Big Data. One of these issues is the increase in demand for competence within the field of data gathering (Tankard 2012). The data have to be kept secure from data breeches, and steps have to be implemented in order to prevent the data from being used if it should be leaked (like complex hashing algorithms and salting of passwords). One hard issue that have to be faced is the case of re-identification attacks, where datasets can scanned in order to find correlations in order to identify one single human identity (Jensen 2013).

Conclusion: Big Data & Privacy

I started off this paper by first talking about the many uses of Big Data and how we as consumers can notice the results of Big Data models in use. Big Data can be used by the industry to create models that would predict which customers have a high likelihood of cancelling their subscription with that company (called *churning*). The owners of web stores can use Big Data to set up a recommendation system that would recommend which products the customer would like to purchase. Big Data also has good uses within medicine, mental health and a diseases' spread prediction.

Big data

We can define Big Data with three words: *Velocity*, *Variety* and *Volume*. Velocity is the speed in which we receive data, we will receive data in high speed as a result of high refresh rates from, for example, sensors which again will result in short time to process these types of data. Variety acknowledges that we have to deal with a lot of different data formats such as structured database tables or unstructured PDF documents. Volume is the huge datasets that we have to process.

There are many sources we can gather data from, like GPS coordinates from a users' mobile phone or what people like on Facebook. We can also track the behaviour of users that visit our website.

Algorithms exists to draw conclusions from data being gathered. One of these algorithms is called K-nearest neighbour, which looks at the closest neighbours for a new entry and tries to classify what that new entry is based on the neighbours. This algorithm can be used in a *Churn Prediction Model* to classify if a user is likely churn.

Cambridge Analytica and IoT Devices

In order to show how Big Data can be a threat to our privacy, I presented two cases in which Big Data was/is in use: the Cambridge Analytica Scandal and the increase in Internet of Things devices.

Cambridge Analytica

Working with the digital marketing campaign for Trump, Cambridge Analytica collected data through Facebook on each American voter that would predict which political candidate they would vote for. Based on this, they created political campaign videos that they would use on a specific group in order to convince them to vote for Trump instead of Clinton.

In Trinidad, Cambridge Analytica convinced voters of the opposing political party to abstain from voting by launching a campaign that encouraged apathy amongst voters. This campaign was called the *Do So!* campaign.

IoT Devices

Internet of Things devices have been emerging and been integrated as a part of our lives. Your smart light bulbs, smart speaker and smart watch collect data about you and your habits. I have been talking mostly about IoT devices you purchase, but you can create your very own IoT devices by using Arduino or Raspberry Pi. The benefits, and arguably the downside, of creating your own IoT devices is that you are in control of security and what data is gathered and used.

Laws about privacy

After Cambridge Analytica, we have seen an increase in laws and regulations about our privacy online. They all want to make sure that users are informed on what data is being gathered from them and how it is being used. Furthermore, the laws require that users can easily change their mind on the consent they originally

gave. We have the Children's Online Privacy Protection Act, which protect the sensitive information surrounding children below the age of 13 and the GDPR that protect the people living in the EU.

IoT devices have some privacy issues related to the use of them. One of them is the fact that the users are not thoroughly informed about the data collected before they give consent. Recommendations are also made that IoT devices should include some sort of technology that hides the users location and communication paths like The Onion Router.

General Big Data issues

The general issues surrounding Big Data is mostly based on how the data is stored and how it is accessed. There are some attacks that are based on correlation, where several databases store separate information that can create a unique identifier if a correlation is found. The data must be kept secure and be able to prevent someone from the outside to be able to use the data if an intrusion should occur.

The demand for competence within Big Data is increasing and challenges may keep appearing when the collecting of data increases and the use of new Big Data techniques is used.

Reference list

Sources

Amer, Karim, and Jehane Noujaim. 2019. *The Great Hack*. Documentary. https://www.imdb.com/title/tt9358204/?ref_=nv_sr_1?ref_=nv_sr_1.

Ballhaus, Rebecca. 2018. “Cambridge Analytica Closing Operations Following Facebook Data Controversy.” *The Wall Street Journal*, May 2, 2018. https://www.wsj.com/articles/cambridge-analytica-closing-operations-following-facebook-data-controversy-1525284140.

Cawley, Christian. 2015. “Turn Your Raspberry Pi into a Network Monitoring Tool.” *MakeUseOf*, June 18, 2015. https://www.makeuseof.com/tag/turn-raspberry-pi-network-monitoring-tool/.

Channel 4 News. 2018. *Cambridge Analytica Uncovered: Secret Filming Reveals Election Tricks*. Channel 4 News. https://www.youtube.com/watch?v=mpbeOCKZFfQ&feature=emb_title.

Channel 4 News Investigation Team. 2018. “Data, Democracy and Dirty Tricks.” *Channel 4 News*, March. https://www.channel4.com/news/data-democracy-and-dirty-tricks-cambridge-analytica-uncovered-investigation-expose.

Davies, Harry. 2015. “Ted Cruz Using Firm That Harvested Data on Millions of Unwitting Facebook Users.” *The Guardian*, December 11, 2015. https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

Federal Trade Commission. 1998. *Children’s Online Privacy Protection Rule (“COPPA”)*. https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

———. 1999. “Children’s Online Privacy Proposed Rule Issued by FTC.” In . https://www.ftc.gov/news-events/press-releases/1999/04/childrens-online-privacy-proposed-rule-issued-ftc.

———. 2018. “VTech Settlement Cautions Companies to Keep COPPA-Covered Data Secure.” January 8, 2018. https://www.ftc.gov/news-events/blogs/business-blog/2018/01/vtech-settlement-cautions-companies-keep-coppa-covered-data.

Franks, Bill. 2012. *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. 1st ed. Vol. 56. Wiley and SAS Business Ser. John Wiley & Son, Incorporated. https://ebookcentral-proquest-com.ezproxy.hioa.no/lib/hioa/detail.action?docID=821898.

GDPR.eu. n.d. “General Data Protection Regulation (GDPR).” https://gdpr.eu/tag/gdpr/.

González, Roberto J. 2017. “Hacking the Citizenry?: Personality Profiling, ´big Data´ and the Election of Donald Trump.” *Anthropology Today*, June 1, 2017. https://rai.onlinelibrary.wiley.com/doi/full/10.1111/1467-8322.12348.

Gordon, Jesse. 2016. “When Data Crimes Are Real Crimes: Voter Surveillance and the Cambridge Analytica COnflict.” Thesis paper, University of Saskatchewan. http://dspace.library.uvic.ca/bitstream/handle/1828/11075/Gordon_Jesse_MA_2019.pdf?sequence=1&isAllowed=y.

Green, Joshua, and Sasha Issenberg. 2016. “Inside the Trump Bunker, With Days to Go.” *Bloomberg*, October 27, 2016. https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go.

Hamburger, Tom. 2015. “Cruz Campaign Credits Psychological Data and Analytics for Its Rising Success.” *The Washington Post*, December 13, 2015. https://www.washingtonpost.com/politics/cruz-campaign-credits-psychological-data-and-analytics-for-its-rising-success/2015/12/13/4cb0baf8-9dc5-11e5-bce4-708fe33e3288_story.html.

Hunt, Troy. 2015. “When Children Are Breached - inside the Massive VTech Hack.” November 28, 2015. https://www.troyhunt.com/when-children-are-breached-inside/.

IEEE-USA. 2018. “Digital Personal Privacy, Awareness and Control.” https://ieeeusa.org/wp-content/uploads/2018/08/DigitalPrivacy0618.pdf.

Isaak, Jim, and Mina J. Hanna. 2018. “User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection.” *Computer* 51 (8): 56–59. https://ieeexplore.ieee.org/document/8436400.

Jensen, Meiko. 2013. “Challenges of Privacy Protection in Big Data Analytics.” *2013 IEEE International Congress on Big Data*, September 16, 2013. https://ieeexplore-ieee-org.ezproxy.hioa.no/abstract/document/6597142.

Kaelin, Mark. 2019. “GDPR: A Cheat Sheet.” May 23, 2019. https://www.techrepublic.com/article/the-eu-general-data-protection-regulation-gdpr-the-smart-persons-guide/.

Klobuchar, Amy. 2018. *S.2728 - Social Media Privacy Protection and Consumer Right Act of 2018*. https://www.congress.gov/bill/115th-congress/senate-bill/2728.

Mukherjee, Aritro. 2016. “Security Access Using RFID Reader.” *Arduino Project Hub* (blog). May 7, 2016. https://create.arduino.cc/projecthub/Aritro/security-access-using-rfid-reader-f7c746?ref=platform&ref_id=424_respected__beginner_&offset=1.

Nature Science Journal. 2018. “Cambridge Analytica Controversy Must Spur Researchers to Update Data Ethics.” *Nature*, March. https://www.nature.com/articles/d41586-018-03856-4.

Nix, Alexander. 2016. “Cambridge Analytica - The Power of Big Data and Psychographics.” Video presented at the Concordia Annual Summit, New York. https://www.youtube.com/watch?v=n8Dd5aVXLCc.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. “Predicting the Future - Big Data, Machine Learning, and Clinical Medicine.” *The New England Journal of Medicine*, September 29, 2016. https://search-proquest-com.ezproxy.hioa.no/docview/1824654037/fulltext/ADA508B43EB44DD6PQ/1?accountid=26439.

Passos, Ives Cavalcante, Benson Mwangi, and Flávio Kapczinski. n.d. *Personalized Psychiatry*. Springer, Cham. https://link-springer-com.ezproxy.hioa.no/book/10.1007%2F978-3-030-03553-2#toc.

Perera, Charith, Rajiv Ranjan, Lizhe Wang, Samee U. Khan, and Albert Y. Zomaya. 2015. “Big Data Privacy in the Internet of Things Era.” *IT Professional*, June 2, 2015. https://ieeexplore.ieee.org/document/7116422.

Rhysider, Jack. 2017. “The Peculiar Case of The VTech Hacker.” *Darknet Diaries*. https://darknetdiaries.com/episode/2/.

Roesner, Franziska, Tadayoshi Kohno, and David Wetherall. 2012. “Detecting and Defending Against Third-Party Tracking on the Web.” presented at the 9th USENIX Symposium on Networked Systems Design and Implementation, San Jose, CA. https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/roesner.

Ruiz, Gerson. 2018. “OWASP Top 10 Security Risks - Part 1.” *Sucuri* (blog). October 3, 2018. https://blog.sucuri.net/2018/10/owasp-top-10-security-risks-part-i.html.

Rundhall. 2019. “DIY Automatic Cat Feeder.” *Arduino Project Hub* (blog). July 29, 2019. https://create.arduino.cc/projecthub/rundhall/diy-automatic-cat-feeder-6fb886.

Steuart, Jada. 2019. “Netflix’s ‘The Great Hack’ Highlights Cambridge Analytia’s Role in Trinidad & Tobago Election.” *GlobalVoices Advox*, August 6, 2019. https://advox.globalvoices.org/2019/08/06/netflixs-the-great-hack-highlights-cambridge-analyticas-role-in-trinidad-tobago-elections/.

Su, Xiaomeng, and Nils Tesdal. 2019. “INNI3012 Big Data.” Webcourse, NTNU. https://www.ntnu.no/studier/emner/IINI3012.

Tankard, Colin. 2012. “Big Data Security.” *Network Security* 2012 (7): 5–8. https://www.sciencedirect.com/science/article/pii/S1353485812700636.

TRUSTe. 2014. “Internet of Things Industry Brings Data Explosion, but Growth Could Be Impacted by Consumer Privacy Concerns.” May 29, 2014. https://www.trustarc.com/blog/2014/05/29/internet-of-things-industry-brings-data-explosion-but-growth-could-be-impacted-by-consumer-privacy-concerns.

Wesolowski, Amy, Taimur Qureshi, Maciej F. Boni, Pål Roe Sundsøy, Michael A Johansson, Syed Basit Rasheed, Kenth Engø-Monsen, and Caroline O. Buckee. 2015. “Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan.” *PNAS* 112 (September). https://www.pnas.org/content/pnas/112/38/11887.full.pdf.

Wong, Julia Carrie. 2019. “Facebook to Be Fined $5bn for Cambridge Analytica Privacy Violations - Reports.” *The Guardian*, July 12, 2019. https://amp.theguardian.com/technology/2019/jul/12/facebook-fine-ftc-privacy-violations?__twitter_impression=true.

World Health Organization. n.d. “Suicide Data.” *Mental Health* (blog). Accessed November 24, 2019. https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/.

Image sources

(images used in the front-page composition)

Benz, Andre. 2017. *Group of People Gathered on Street*. Photo. https://unsplash.com/photos/qz7KZgeDmjU.

Li, Vino. 2019. *Untitled*. Photo. https://unsplash.com/photos/pZB4DGuKrYw.

Wang, Naian. 2017. *Photograph of Aurora Lights*. Photo. https://unsplash.com/photos/F9wrh2miJLA.