POLITECNICO
MILANO 1863

Public Privacy: A counter archive to expose privacy through visual AI manipulation

Tim Olbrich

Matr. n° 936524
Laurea Magistrale
Design della Comunicazione
A.A. 2020/2021

Supervisor:
Maria de Los Ángeles Briones Rojas

Index of Contents

Index of Figures

Abstract

English

Data is omnipresent in today's societies and influences people's lives in many different ways, yet not many understand its impact and consequently the importance of data privacy. Since data is a big economic factor for companies and business models, new sources of information are constantly utilized to keep the data stream flowing. It leads to monopolies in which Big Data continues to move away from the democratic idea of a liberal internet. Additionally, the user sometimes does not even know about these mechanisms, nor their impacts. This becomes especially notable when customers expose private information in form of photos to the internet. SHEIN, the biggest fashion store in the world, utilizes a reward system to gather images of customers wearing its products. Motivated by gamification strategies, privacy takes a secondary role and loses its importance for the user but benefits the fashion retailer. By designing a counter archive of these images and metadata, a new perspective on the intimacy of private information can be established. This process involves the manipulation of user images and is done with the same tools and algorithms used to extract personal data. The result is a website which aims to reveal part of the data privacy issues and prompts viewers to become more responsible about their own dealing with data.

Italian

I dati sono onnipresenti nelle nostre società e influenzano la vita delle persone in tanti modi diversi, ma non molti comprendono il loro impatto e, di conseguenza, l'importanza della data privacy. Poiché i dati sono un grande fattore economico per le aziende e i modelli di business, nuove fonti di informazione sono costantemente utilizzate per mantenere i flussi di dati. Questo porta a monopoli in cui il Big Data continua ad allontanarsi dall'idea democratica di un internet liberale. Inoltre, l'utente, molte volte, nemmeno conosce questi meccanismi, né il loro impatto. Questo diventa particolarmente notevole quando i clienti espongono informazioni private sotto forma di foto su internet. SHEIN, il più grande negozio di moda del mondo, utilizza un sistema di ricompensa per raccogliere immagini di clienti che indossano i suoi prodotti. Motivata da strategie di gamification, la privacy assume un ruolo secondario e perde la sua importanza per l'utente, mentre beneficia invece il rivenditore di moda. Progettando un contro archivio di queste immagini e metadati, questo lavoro si propone a stabilire una nuova prospettiva sull'intimità delle informazioni private. Il processo implica la manipolazione delle immagini degli utenti e viene fatto con gli stessi strumenti e algoritmi utilizzati per estrarre i suoi dati personali. Il risultato è un sito web che mira a rivelare parte dei problemi di data privacy e invita gli utenti a diventare più responsabili riguardo al proprio trattamento dei dati.

Introduction

A survey in 2018 asked people about their opinion on the use of personal data. In the United States 55% of the participants identified themselves with the statement of not selling their personal data at any price. In Germany, the number was with 67% even higher (Syzygy 2018). It seems that people generally do not want their data to become a sort of currency.
However, this is already the case. No matter if we pay our coffee with Apple Pay or upload a photograph of the same to Instagram. The review on Amazon, the post on Facebook, the match on Tinder, it all reveals parts of ourselves and who we are: Interests, tastes, desires, and fears. Datapoints are accumulated and analyzed, targeting profiles are created and used in the most profitable way. In consequence our digital twin often paints a more precise and honest picture of ourselves than we do in the real world (Rudder 2015). Companies know our porn preferences, mental health issues and our motivations to vote. These information are so private that most people would probably not even tell them their best friends. Private data is frequently created across devices and users, sometimes with the user's agreement, but often unconsciously in the background (Meehan 2019). New legislations such as the General Data Protection Regulation try to initiate a shift in how data is handled. However, data privacy remains to be an abstract concept and people who do not have anything to hide do not need to worry, right? In this context the subsequent project aims to bring light into the darkness, to make data privacy a more concrete and tangible subject by exemplary using real world cases. Moreover, the research focuses on gamification as a strategy to soften users' cautiousness. It takes a closer look at what kind of photographic material users share with the public and how these images expose their privacy. Our society relies on interconnectivity and information exchange. We benefit from Big Data in areas like urban planning, disease research or technological advancement (Mills 2019). While Big Data is an abstract concept, a process of accumulating and analyzing data, it is communication design which can make it tangible. Ultimately, design can help to strengthen the responsible movement within a data driven world, because one thing is sure: We already live in data driven world.

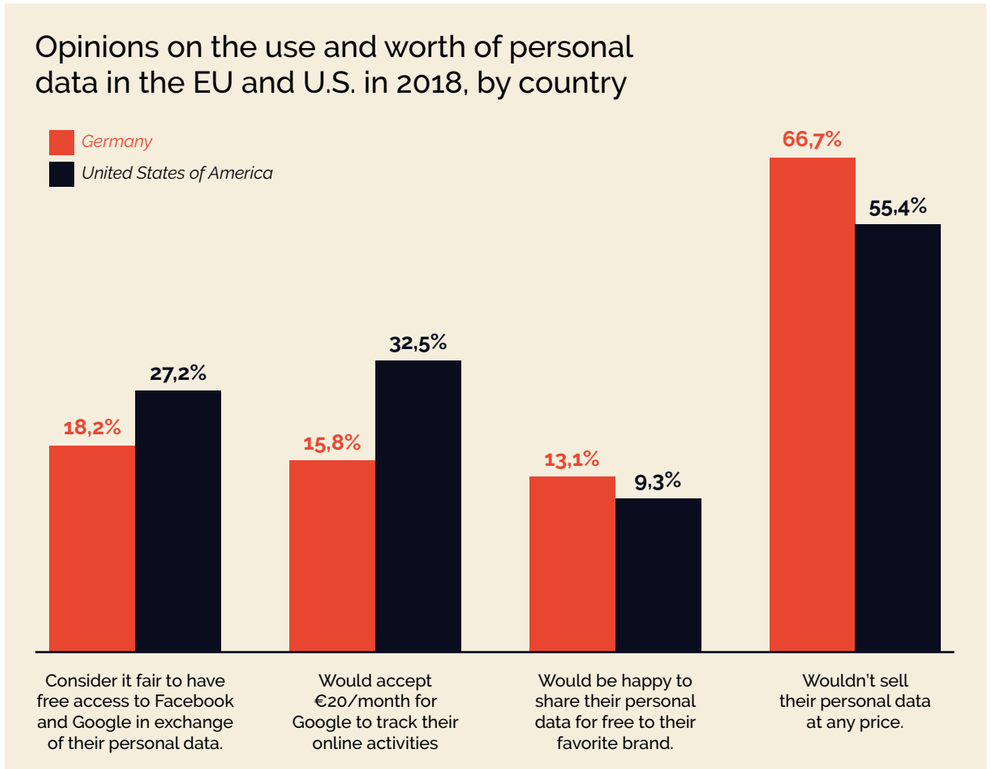

Figure 1: Opinion on selling personal data in selected countries.

A. Big Data, Big Problems?

A.1: An Introduction to Big Data

For our grandparents the local telephone directory must have been the biggest data collection they have ever held in their hands. Today, smartphone users carry thousand times more data with them, every day *(Mayer-Schönberger, Cukier 2013)*. In fact, most of us live with a computer in their pocket that is 100.000 times more powerful than the Apollo Guidance Computer used to land the first men on the moon *(Kendall 2019)*. Additionally, these smartphones create an enormous amount of data every day — for example, the time we set our alarms to. These datapoints make it able to draw conclusions of the person using the smartphone — how tired we were when waking up or how late we partied the other night. These datapoints pile up and the phenomenon is called Big Data. Although there is no universal valid definition for the term, Big Data usually describes datasets larger and more complex than they were a couple of years ago, datasets which can be generated in high-speed and at the same time provide many possibilities of interpretation and usage *(Mayer-Schönberger, Cukier 2013)*.

A big contribution roots in the world wide web and its idea of being a free and open space for everyone. No matter if we want to send an e-mail, post something on Facebook, upload a photo to Instagram or watch a video on YouTube. At first sight, we can do these things without paying for them. In its early days, there was no commercial content in the web; companies who hid themselves behind paywalls quickly had to realize that their customers would switch to alternative, free services. However, free content does not run a business. Consequently, companies developed the strategy to make money out of the users' data *(Schneier 2015: 58)*. What started in the web was later transformed to other busines sectors within the digital and non-digital environment. However, Big Data does not only mean the accumulation of large amounts of data. The term also includes how this data becomes more qualitative, more precise, and how its meaning can influence and transform societies. Suddenly we see coherences we did not know about. In other words, Big Data opens us a new perspective to look at the world *(Mayer-Schönberger, Cukier 2013)*.

Before the existence of computers, it was a tedious job to gather data. One of the most prominent examples is the US-Census, the counting of all citizens in the United States. The constitution required an updated number of the population every ten years. As easy as it sounds, the census from 1880 took eight years. By the time it was done, the number was already outdated because many people died, and babies were born (Ibid.). Today it is obviously much easier, a quick web search returns the result. This is made possible by data that is generated and stored nearly in real time. For example, in 2019 when the first cases of Covid-19 forced many countries into a lockdown, it was possible to predict the epidemic trend by gathering global data. Hospitals and local doctors would not have been able to make these predictions, because they did not have the same data basis *(Dong, Du, Gardner 2020)*. Spot tests in which a small group of people was asked on certain topic are not contemporary anymore. Nevertheless, when sociologists wanted to predict the outcome of an election a couple of years ago, they had to go on the streets or call households to conduct surveys. This group had to be representative, meaning that it should portrait a realistic picture of the group: taking all genders into account, people from different age groups, social classes, and incomes. Although this process worked quite well and lead to more or less accurate predictions, it often could not answer further questions. For example, when someone had asked how high school teachers would vote, no scientist could give an answer even high school teachers were part of the participants. The reason is that the number of teachers within the representative group was not big enough to make a point *(Mayer-Schönberger, Cukier 2013)*. Big Data changes the process: Today we have so many individuals as data points — sometimes even all individuals — that we can zoom into the data without the need of further investigations. At the same time, it leads to the question how reliable Big Data is. In the 80s a group of IBM engineers worked on a new language translation approach. They swapped their existing large and chaotic dataset with a smaller but more accurate one. Until then it was common to use a vocabulary database with grammar rules, but the team decided to use three million sentence pairs from official Canadian parliament translations. The software successfully translated common words but failed at unusual ones. It turns out, the problem was not the quality — the team used legalized translations — but the quantity. A decade later, Google tried a slightly different approach: Instead of caring about quality it used all translations they could find online. A chunk of this data obviously contained poor quality data. However, the overall success rate of the program was much better compared to similar systems because distortions were removed by the pure size of the database. On the contrary, IBM's translations broke, because single data points had a much larger impact on the overall program. A single mistranslated word in the amplitudes of the world wide web do not have the same impact as an error in the Canadian translations *(Ibid.)*.

A.2: Platform Capitalism

Data plays a more important role in businesses than ever before. With the worldwide spreading of smartphones and social media, people quickly got used to share all sorts of information about themselves. The output is a constant data stream which simply did not exist before. The phenomenon is called datafication and describes the process of transforming information into data *(Cukier, Mayer-Schönberger 2013: 29)*. These datapoints are very profitable for companies, but also allow interesting insights of who we are as a society. Thereby it is not surprising that businesses as well as scientists constantly look for new tools which could help them to gain more data. For example, the Japanese Advanced Institute of Industrial Technology developed a set of 360 sensors that could measure the weight distribution on a seat. Based on the distribution it was possible to recognize who sat down, thus the sensors could be placed in cars and function as an additional layer of security — the cars would only start of if it recognized the driver *(Mayer-Schönberger, Cukier 2013)*. Another interesting application is IBM's patent for a floor covering which would be able to detect when and how someone walks over it, or if a person fell and needed help *(Catherine O'Connor 2012)*. While everything, starting with our cars and moving into our houses, seems to receive sensors and connectivity, the direction of this trend is clear. Companies try to access even more sources of data to gain insights and ultimately create products that many people want to buy. Besides that, data can give companies a monopoly position: A company with data, has insights that competitors might not have.

The biggest social network — Facebook — is usable free of charge would most people argue. Truth is that users pay with their personal data and Meta, the company behind Facebook is very good in collecting it. The acquisition goes far beyond facebook.com itself because website operators often use Facebook's tracking tool, the Facebook Pixel to monitor their website's performance. It tracks user metrics such as the overall number of page clicks, or how long users stay on the website (Meta Platforms). However, while website operators can only evaluate data for pages they own, Facebook can read data of all websites that use their tool. The more websites use the tracking pixel, the more comprehensive and fine-meshed it can track a user. Consequently, users can be tracked across pages and even without being registered at Facebook. According to Nick Srnicek, a lecturer in Digital Economy at the King's College London, Facebook's need for data collection is not surprising *(Srnicek 2017: 55)*. In fact, Facebook's business model is built around it: In 2020, 97,9% of Meta's revenue came from the advertising sector which conversely means that it heavily relies on private user data to create sophisticated advertising profiles *(Meta Platforms 2022)*. In its commercial essence the goal is to create user profiles that are as precise and predictive as possible. These profiles can then be targeted by advertisers who want to address a very specific group of users. That this model works flawlessly can be seen through the Signal's campaign. Signal is a messenger service which sticks out as an alternative to the Facebook messenger and similar competitors by not collecting private user data. Their campaign ran on Instagram and revealed which targeting options Facebook offers its customers. The ad banner consisted of a simple blue background and a text which for example said: "You got this ad because you're a K-pop-loving chemical engineer. This ad used your location to see you're in Berlin. And you have a new baby. And just moved. And you're really feeling those pregnancy exercises lately." Signal was able to precisely determine their audiences' location, interests, and life situations, made possible by Facebook's targeting profiles.However, Facebook itself was not very much into the idea and banned Signal *(Harada 2021)*.

Big Data and the way Facebook and other social media networks utilize it can be seen very critical. Privacy invasion, fake news and political delusion are the aspects most often heard in news outlets, whereas Srnicek argues that the economic aspects are often neglected. Meta and other data driven businesses are a capitalist companies in a capitalist society, which means they want to maximize their income *(Srnicek 2017: 3 ff.)*. We already know this trend from the Fordist Business Model, a term named after the business magnate Henry Ford, which describes an era of mass production, mass consumption and vertical integration. In other words, as many cars were sold to as many customers as possible while everything from the supplier to the salesmen were part of the company. With the 1980, the Post-Fordist Business Model focused much more on the flexibility and customization of a product. For example, a car had to bring different interior options to please customers with a more personalized experience. Mass production of the exact same product was not enough anymore. Since then, we see a shift from vertical to lean businesses, meaning that all parts of the company not directly necessary for the business are outsourced: The design and management for example could be done in high-income economies while the actual manufacturing could be outsourced to low-income economies with cheaper workforces. It reduces the price of the product and generates higher profits *(Ibid: 9–19)*. Today, businesses shift towards platforms in which they connect multiple groups of interests and provide them technical infrastructure. For example, Facebook connects consumers to advertisers while it provides the tools to build personal profiles, company pages, messaging tools and targeting options. At the same time, it lives of its network: the more users, the more valuable the social network. The network effect is crucial for platforms business to work *(Ibid: 27)*. For example, Airbnb does not own any apartments or hotels, it relies on users who are willing to let theirs. Uber relies on its drivers; without them no one could call a chauffeur. Deliveroo needs restaurants and deliverers to become part of their service, only then it can offer food. Moreover, these platforms do not actively let apartments, drive people, or cook food, they only provide the infrastructure to bring businesses and customers together *(Ibid: 39)*.

Platform Capitalism describes the rise of highly specialized and its competitor's superior tech companies. A key ingredient to their success is the data gained from everyone who uses their services which gives them the ability to form monopolies. For example, people who search something on Google could do this because the search engine returns rapid and accurate results. Additionally, by searching on Google, users provide data that can be used to improve the search algorithms, which conversely makes it even faster and more precise to attract more users. In the end, Google becomes predominant, because no other search engine has the information to develop such a sophisticated algorithm *(Ibid: 26)*. Furthermore, platforms make use of cross- subsidization to attract or keep certain user groups. People using Gmail, Google's email service, usually do not pay for it, so it would be a losing game on itself for the company. However, free e-mail services attract users which can then be used to attract more advertisers at a higher rate, thus the overall model becomes profitable *(Ibid: 26, 55)*. Important in this process is to keep in mind that platforms do not underly the usual market response. A low demand does not automatically lower the price as the example of Uber shows. Someone who wants to call an Uber will see more available cars in the app than there are in reality. On the other side, if there is a local demand for more drivers,

Uber will raise the prices in that area to attract more Uber drivers. Thus, users are not able to view a neutral picture of the market anymore, because Uber dictates what the market is. Another example is Facebook which does not show the same content to every user, but selected content to keep the user on the feed as long as possible *(Ibid: 26ff.)*. Srnicek says:" While often presenting themselves as empty spaces for others to interact on, they in fact embody a politics"*(Srnicek 2017, 26)*, meaning that platforms regulate and rule in their own ways and no outside force directs them. Another problem we should be aware of is a platform's monopoly position, because at one point it will be impossible to catch up with such, simply because it would cost too many resources, too much data to beat them. Taking a closer look at these platform models, they can be divided into five categories being advertising-, cloud-,industrial-, product- and lean platforms.

A.2.1: Advertising

Advertising platforms, primarily Google and Facebook, generate revenue by giving advertisers the option to precisely target their audience with customized ads. Therefore, these platforms are interested to expand its data and knowledge on internet users to attract more advertising clients. From an economic perspective, these companies rely on collecting personal data, thus data privacy is by nature never going to be one of their concerns *(Ibid: 28–32)*.

A.2.2: Cloud

Cloud platforms let their hardware, infrastructure, and development tools to other companies, to perform cloud-based computing. These services reach from simple cloud hosting to image recognition algorithms, machine learning to artificial intelligences which can be rented on demand. With 32% of the market share Amazon Web Services is the biggest of these cloud platforms, followed with 21% by Microsoft Azure *(Canalys, Statista 2021)*. Consequently, their profit depends on other businesses renting cloud infrastructure *(Srnicek 2017: 32ff.)*.

A.2.3: Industrial

IoT-Factories use sensors and dedicated software to react on certain events within the production line. This allows BASF, one of the biggest chemical producers in the world, to be flexible when a machine breaks down, but it also provides greater flexibility because every produced item can be different from the previous one. The factory is not bound to a static system anymore. The industrial platform model is used by companies that provide IoT assets to the factory. The leading ones, General Electric and Siemens focus on renting out hardware as well as software to a variety of business sectors, not only tech-oriented ones *(Ibid: 34–37)*.

A.2.4: Product

Not only technology companies utilize the platform model, more and more industries adapt to it. Interconnected factories produce goods that can collect data which in return are used to improve the product, lower its maintenance costs, or provide more flexibility. For example, Rolls Royce offers its jet engines per hour of usage and takes care of maintenance and replacement parts. Every engine is equipped with multiple sensors which collect data on its usage: from weather conditions to maintenance issues. With these information, Rolls Royce was able increase fuel efficiency and life expectation of its engines *(Ibid: 37ff.)*.

A.2.5: Lean

Lean platforms only provide a platform and infrastructure, but do not own any real assets even though they depend on them. For example, Airbnb does not own any hotels, but at the same time it is the world's largest accommodation platform. By outsourcing as much as possible, these companies minimize cost while being very flexible in their operation. However, they often depend on venture capital welfare to excel competitors purely by their number of resources which at the same time makes them unprofitable *(Ibid: 39–45)*.

A.3: Data Bias

There is a lot we must consider when buying a used car: mileage, year, model, warranty, consumption. Big Data says that color should also be considered, but why? In 2012 a data analysis competition found out that orange vehicles tend to be half as prone to technical defects as other colors. Unfortunately, datasets like these never tell us the reason. Big Data can analyze and draw connection among multiple factors, often within areas we would not have thought of searching for correlations *(Mayer-Schönberger, Cukier 2013)*. However, until today no one figured out what separates orange cars from other ones. This example shows an important aspect of working with data: a table with numbers and letters by itself is not worth much. The data value lies within its interpretation and understanding, only then it can be useful. Furthermore, we should keep in mind that data can be useful and important, but data is not the answer to all our problems. While looking through the data-perspective, it can happen that we misjudge its value. For example, standardized and automatized tests — a common practice in the United States — were introduced to make schools comparable among each other. Moreover, they prevent teachers from giving bad grades to students they dislike. On the contrary, these exams bear the great risk that students and teachers only focus on getting the best results out of these tests. A subject's content, its understanding and quality become a secondary element *(Ibid.)*. Another demonstrative example comes from former United States Secretary of Defense, Robert McNamara. During the Vietnam War he developed a strategy which primary focused on one number: the number of fallen enemy troops. Whenever the number grew, he reckoned the success of his strategy and continued with it. His misconception was that he got lost in one number and forgot about the reality of war. Additionally, it later turned out that the numbers he received were incorrect. Troops reported sugarcoated numbers to please their direct supervisors. Ultimately, the meaningfulness of the data was gone. Decades later we read much more data in much faster pace *(Ibid.)*. Big Data provides meaningful insights and new chances, but no one should run blindly into it. Sometimes it is necessary to take a step back, to question if the data still reflect reality and to make sure that it does not offend our ethical values. If this becomes the case data can influence our behavior in ways of bringing more harm than good.

Police forces rely increasingly on algorithms to identify criminal activities, to track people or to predict risks and in the United States parole boards already use data-based algorithms to estimate how likely it is for someone to fall back into criminal patterns *(Angwin et al 2016)*. Usually, these algorithms make their predictions based on specified factors such as the surrounding, the neighborhood, social status, or drug abuse of a person *(Mayer-Schönberger, Cukier 2013)*. Even medical records are used to evaluate people as the case of Kyle Behm shows. Behm was diagnosed with a bipolar disorder, hence he had to drop out of his university to receive treatment. Later on, he was healthy enough to look for part time jobs and applied at Kroger, the local supermarket. Unfortunately, his job application was rejected, and it turned out that due to the results of a personality test an algorithm has tagged Behm as likely to underperform *(O'Neil 2016: 90)*. This example illustrates how algorithms can favor certain groups while blaming others. This can become especially important when authorities use these kinds of tools, because their justification often argues for more safety. On one hand algorithms and tracking tools help to manage police resources better. On the other hand, it infringes the right for everyone to be treated equally. For example, people could be suspected based on their ethnicity, or circle of friends. A study from 2018, called Gender Shades, reveals that facial recognition technologies do not work universally well for everyone. During tests, lighter-skinned males performed with an error rate of 0,8% while darker-skinned females reached error rates of up to 34,7%. The results are explainable by looking at the dataset used to train the algorithms, it was foremost composed of lighter-skinned faces *(Buolamwini, Gebru 2018)*. It illustrates how algorithms and Big Data are not neutral. Facial recognition is biased, yet it is used by governmental institutions. Moreover, the difficulty is that these algorithms are complex systems and as such difficult to comprehend. Similar to racial bias, an algorithm, used by Amazon in 2018 to simplify the recruiting process, should filter candidates' resumes. It turns out the algorithm was trained on existing applications from the past decade and since the tech industry is still a male dominated field, the system developed a gender bias. It favored men over women *(Dastin 2018)*.
In this case the bias was uncovered relatively quickly and could not cause too much damage. However, when Catherine Taylor applied for a job at the Red Cross in Arkansas, she was rejected due to criminal charges on her record she had never heard of before. Through investigations she found out that another person with her name and birth date caused these inconsistencies. Further research showed that at least ten other data brokers made the same mistake and linked her to the crimes someone else had committed *(Mui 2018)*. While Big Data has an interest in being accurate, errors cannot be completely eliminated. In combination with the fact that most people do not know what data has been collected about them, it makes it very difficult to identify errors. Algorithms are often referred to as black boxes: We give it an input and something else comes out, but we do not know what happens in between. We do not know what happens simply because algorithms become too complex to understand. Moreover, these fragments of software often underly greatest secrecy, making it even difficult for courts to investigate how they are used *(Finn 2017)*. Overall, technology is made by human, thus it is prone to errors. Making these errors and biases visible is almost impossible for non-professionals. A future in which women would not receive a job and workers can be discharged because they are part of a risk group can certainly not be desirable.

Besides being unintentionally biased, algorithms are already used to intentionally control sentiments. Despite the common western belief of a liberal web, which spreads freedom and democracy, the internet can also be used for the opposite. It sometimes distracts from politics, distributes anti-democratic propaganda, and supports intolerant acts. While the extremes of this trend are most visible in authoritarian regimes, western governments are not innocent on manipulating and censoring the web. The reason is that the internet can be easily manipulated: access can be restricted, and content can be controlled *(Morozov 2011: 83 f.)*. Influencers, bloggers, reviewers, they all promote certain values. Just like a Western blogger supports political views, a blogger in an authoritarian regime does the same. Additionally, authoritarian regimes filter opinions that do not correspond to their ideologies while supporting nationalistic or xenophobic values. For example, Maksim Kononenko, started as conservative Russian blogger who supported the government's political course. He later became the co-host of a TV show, because of its positive attitude towards the Kremlin *(Ibid: 128)*. In an online world it is relatively easy to spot positive and negative dispositions. Unlike traditional print media or technologies such as TV and radio, which do not rely on interconnectivity, the internet provides the tools of surveillance by design: A comment under a video, or post and someone's opinion is published. In some countries, such as China, the government can force that monitoring programs to be pre-installed on every sold computer. Users must fear constant cyber surveillance. At the same time, people are not free

to buy products they want, simply because they do not have access to them (Ibid: 98). Consequently, the idea to think the internet would liberate these countries because their citizens could see behind the propaganda curtain is a false assumption. Although authoritarian regimes use these controlling mechanisms much more obvious, Western governments also manipulate the web, but often more indirect. For example, when the United States accused the Iranian government to support terrorists andimposed sanctions against the country, one of them was to blockade website. Iranians suddenly had no access to American websites and internet services anymore. Later when protests against the Iranian regime broke out, the American sanctions were loosened, and Iranians gained access to social networks like Facebook and Twitter again. This move allowed protestors to better organize themselves which aligned with the American interests *(Ibid: 1–4, 209–211)*.

A.4: Surveillance Capitalism

The reason of Google being one of the biggest technology companies today is not rooted in luck or the especially hard work, rather than the social and technological processes that accompanied its genesis. In 1998 Google was founded by Stanford students Larry Page and Sergey Brin as a simple search engine. It already collected user data as a byproduct of searches, but only used them to improve their algorithms. This rapidly changed in the year 2000 as a consequence of the dot-com crash. Google's investors left and the company suddenly had to come up with new ways of making profit, hence the AdWords-Team was found and started to sell behavioral data to the advertising industry — Surveillance Capitalism was born *(Zuboff 2019: 50–54)*. As we know today behavioral data is the foundation for every surveillance system because it allows to control and to even predict the future to a certain level. Since Google's business model is built upon it, the company naturally has to increase its data supplies in order to stay competitive. After joining the data collection business, the company published six declarations in which it claimed human experiences as an unpropertied resource, which the company was allowed to transform into behavioral data. Arising thereby, Google claimed to own these data, the knowledge, and the freedom to decide for which objective it should be used. Finally, the company accredited itself with the interpretational sovereignty of the correct usage. This appropriation builds the base of Google's supremacy today because the knowledge is not shared equally, nor is it democratically legitimated, nor are there any laws that could stop Google *(Ibid: 118–130)*.

The web is an extralegal room ever since and its lack of regulations helped companies like Google and Facebook to grow. That explains their big interest in keeping that way: Tech giants depend on undisturbed data collection. At the same time, these companies have the money and political influence to avoid drastic changes, often with the argument that regulations would limit innovation, hence the internet must stay irregulated. Furthermore, the events on the 11th of September 2001 promoted a governmental legitimation. Data privacy was no longer an issue, everything and everyone focused on public safety. In the United States as well as the European Union laws were established which allowed a comprehensive internet surveillance. Today it is part of the daily business to store all kinds of data, not only from computers and smartphones, but all kinds of IoT devices such as speakers, thermostats, and wearables *(Zuboff 2019: 16–17)*.

A.5: Data Privacy

When people think on the life in former east Germany, their mind quickly makes associations with a surveillance state. No matter with who you had contact with, it was always wise to stay cautious because it could have been someone from the secret police. The Ministry for State Security was interested in everything: who talked about who, wo did not align with the government and who watched West-German TV. Contemporary witnesses reported later how oppressive and constrictive it felt to live in an environment of such mistrust. Although Germany reunified in 1990 one thing did not change: Governments still collect vast amount of data about their citizens. The NSA datacenter in Utah stores 1 billion times more data than the former Ministry of State Security (Mayer-Schönberger, Cukier 2013). The difference is that today people seem to accept it. The world wide web is easy and fast to use, most things are free and open to everyone. Why should one be afraid of it? Bruce Schneider, a computer security expert and cryptographer, states that the "most common retort against privacy advocates" *(Schneier 2006)* is the nothing to hide argument, meaning that a majority of online users think they do not have anything to hide and therefore do not need to care about their data privacy. A decade ago, it would have been unthinkable to equip every room of a house with microphones, today people voluntarily place smart speaker into their bedrooms. Smartphones are omnipresent and track every movement with an accuracy of one meter. Most people have a digital twin, a version of themselves formed by the data traces they constantly leave. Technology provides these possibilities and governments make use of it. For example, in the United States, the FBI maintains a database of more than 52 million faces *(Schneier 2015: 32)*. If necessary, they can track individuals in with cameras at public spaces, busses, public facilities, and stores. By no doubt it is difficult to escape CCTV, but the bigger portion of data is handed in freely by online users. In 2012 a father complained that his daughter would frequently receive advertisements for baby products. He thought it was inappropriate for his teenage daughter. In the end it turned out that the girl was pregnant: She searched for baby products, before she told her family about the pregnancy *(Ibid: 39)*. Google saves every search query ever run. The company does not only know their users better than their families or best friends do, it also knows them better than the user itself. Schneier says: "Google knows more about what I am thinking of than I do, because Google remembers all of it perfectly and forever." *(Schneier, "Data and Goliath", 26)* In other words, people tend to forget that they leave traces in the web which are hardly if not impossible to remove. Their wishes and fears, who they secretly have a crush on, which party they vote for, their illnesses. All these information are not in the hand of the user anymore. After all, what is the problem about it? If someone does not plan to do something illegally, it does not matter to expose private data, right? In many ways this approach is naive because it does not respect the social influence. Mass surveillance always goes along with limitations, often so slowly that they are not noticeable at first sight. In 2012, the Irishman Leigh Van Bryan tweets: "Free this week, for a quick gossip/prep before I go and destroy America." People who know Bryan closer probably knew that he intended to party and drink a lot in the United States. However, the agents who picked him up upon his arrival at the airport and interrogated him for five hours did not know that *(Ibid: 110)*. In 2013, a Hawaiian man was arrested, because he posted a video on Facebook which showed him drinking and driving. In his defense he stated that it was only a joke, and no real alcohol was involved *(Ibid.)*. Regardless of whether the involved parties acted smart or not, it shows that their digital twin was used to convict them for a crime they did not commit. It raises the question to what degree individuals can freely speak online, or if they must expect consequences.

The most extreme example is the drone targeting strategy of the US military. At its peak half of the US drone operations targeted unidentified individuals, who were only targeted based on personal and behavioral characteristics. It is unknown how accurate these datapoints are *(Ibid.)*. At first glance, it might not be problematic to share private information online, but the risk to run into social constraints remains. No one should need to think twice if they want to google a topic. In long terms a society would become conform, individual opinions would be suppressed, and alternative opinions would not be allowed. Edward Snowden, a whistleblower who revealed classified information of several US surveillance programs in 2013, summarizes the situation as follows: "Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say. "*(Snowden 2015)*, meaning that claiming nothing to hide would correspond to giving up the right for privacy.

B. Everything is a Game

B.1: An Introduction to Gamification

Most people remember a moment when they have been asked for their PAYBACK card or frequent flyer number. The idea behind it is easy: The customer can collect points in specified stores — often of the same franchise — and in return receives a discount or reward. These programs aim to motivate customers to buy more frequently from the same brand while not getting attracted to competitors, a practice which is commonly used in many industries from supermarkets, over online shops to global airline networks. A similar concept is used within businesses when the motivation runs dry, or the team cannot come up with new imaginative ideas. In these moments, the team leader or presenter often comes up with a deck of playing cards, seemingly obscure methodologies, or role-playing scenarios to bring the group on track *(Zichermann, Linder 2013)*. All these scenarios fall under the term of gamification because they apply elements and principles of game-design to a non-game context, they gamify something which naturally is not a game like experience *(Deterding et al: 2)*. Moreover, gamification cannot only be found within the business, commerce, and advertising industry, but in many different shapes such as education, health, work environment, or sustainability; both online and offline. Their advantage within businesses is that they act as simulations, nothing must be real, but everything can be explored. Players can escape their usual behavior and take risks, thus new chances or insights can be gained. This strategy is applied since 2007 to United States Army officers who take game-based trainings to improve their negotiation skills. Another example from the bond company PIMCO predicted that the Lehman Brothers would collapse, which ultimately happened in 2008 *(Zichermann, Linder 2013)*. Their predictions used scenario-based planning — usual quantitative data models were not able to foresee a collapse. Other businesses use gamification to sheer their employees up. According to the World Health Organization one of the most common stress factors at work is the lack of control over the own destiny *(Ibid.)*. Especially work that requires monotonous tasks such as cashiers checking out items becomes boring quickly and employees lose the sense of control. The American supermarket chain Target tackled this problem by introducing a checkout game. It calculated the speed of swiping an item over the scanner and rated if the scan was too slow or within the desired speed. Not only did it introduce a sense of control to the repetitive task, but it also gave a feeling of accomplishment *(Ibid.)*.

B.2: Motivated Users

B.2.1: Adidas Running

When people think on the life in former east Germany, their mind quickly makes associations with a surveillance state. No matter with who you had contact with, it was always wise to stay cautious because it could have been someone from the secret police. The Ministry for State Security was interested in everything: who talked about who, wo did not align with the government and who watched West-German TV. Contemporary witnesses reported later how oppressive and constrictive it felt to live in an environment of such mistrust. Although Germany reunified in 1990 one thing did not change: Governments still collect vast amount of data about their citizens. The NSA datacenter in Utah stores 1 billion times more data than the former Ministry of State Security *(Mayer-Schönberger, Cukier 2013)*. The difference is that today people seem to accept it. The world wide web is easy and fast to use, most things are free and open to everyone. Why should one be afraid of it? Bruce Schneider, a computer security expert and cryptographer, states that the "most common retort against privacy advocates" *(Schneier 2006)* is the nothing to hide argument, meaning that a majority of online users think they do not have anything to hide and therefore do not need to care about their data privacy. A decade ago, it would have been unthinkable to equip every room of a house with microphones, today people voluntarily place smart speaker into their bedrooms. Smartphones are omnipresent and track every movement with an accuracy of one meter. Most people have a digital twin, a version of themselves formed by the data traces they constantly leave. Technology provides these possibilities and governments make use of it. For example, in the United States, the FBI maintains a database of more than 52 million faces *(Schneier 2015: 32)*. If necessary, they can track individuals in with cameras at public spaces, busses, public facilities, and stores. By no doubt it is difficult to escape CCTV, but the bigger portion of data is handed in freely by online users. In 2012 a father complained that his daughter would frequently receive advertisements for baby products. He thought it was inappropriate for his teenage daughter. In the end it turned out that the girl was pregnant: She searched for baby products, before she told her family about the pregnancy *(Ibid: 39)*. Google saves every search query ever run. The company does not only know their users better than their families or best friends do, it also knows them better than the user itself. Schneier says: "Google knows more about what I am thinking of than I do, because Google remembers all of it perfectly and forever."

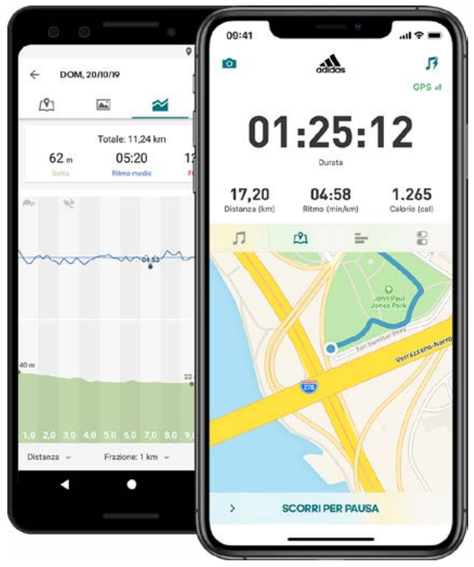

Figure 2: Adida Training.

B.2.2: Duolingo

With over 42 million monthly users, Duolingo is the market leader of language learning apps. A study from 2020 shows that 51,7% of the users feel that Duolingo's gamification strategy affected them in a positive way *(Rodríguez Aranda 2020)*. So, what makes them so successful in the competitive market of language learning apps? Duolingo's gamification already starts outside the app, specifically at its push notifications. The mascot of the company, a green bird called Duo, personalizes notifications and reminders with "Hi! Its Duo.", a simple yet powerful way to engage users as it led to a 5% rise in daily app interactions. Furthermore, the learning progress is tracked with a badge system which allows users to look back at what they have already completed or to show their achievements to friends. During study sessions, the app gives immediate feedback to users by playing a short but distinct ping-sound for every correct answer. Positive reinforcement is known to motivate people and even if the feedback is only a sound, it pushes the user's motivation to continue learning. As in most games, a leaderboard in which player can compete against each other provides a healthy learning competition, thus it motivates students to stay on track. Besides that, long term motivation is achieved by engagement streaks: the user must come back to the app once a day for a certain amount of time in order to receive extra credit. This provides people with another reason to come back and at the same time makes it harder to stop using Duolingo — no one likes to lose the extra credit collected from a streak *(Daniels)*.

B.2.3: Amazon

The online marketplace Amazon shows that gamification does not always has to happen within a virtual environment. In 2021 it reported that it installed video screens close to workstations throughout a warehouse. Employees could play video games such as racing or building a castle during their working times. Progression in the game was made by completing work in the warehouse and works could play individual or in teams. The goal is to take out the monotony of the job and to engage employees. However, critiques mentioned: "If a job is so tedious that it requires a distraction from monotony, maybe it's the job itself that needs re-consideration. ", meaning that Amazon does not solve the problem by covering it up with a game. Moreover, it is feared that in the future the game progression could be used to monitor what a worker accomplishes or does not accomplish. Amazon itself denies these plans at the moment and underlines that it neither monitors game results nor penalizes employees who wish not to make use of the tool *(Anderson 2021)*.

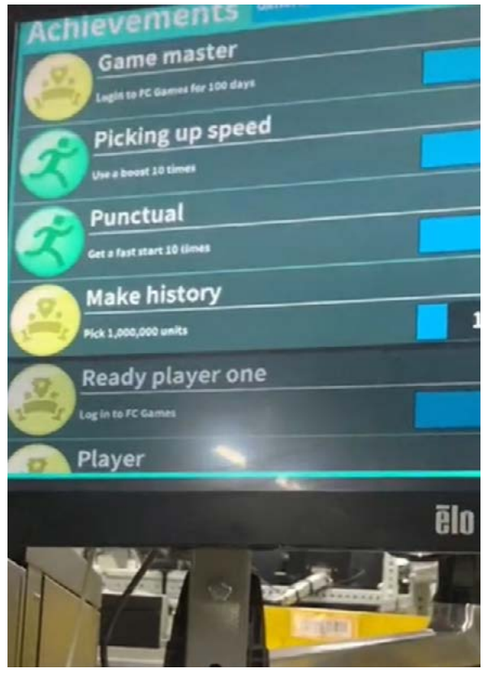

Figure 3: Amazon using gamification at its warehouses to motivate employees.

B.3: SHEIN and the Fashion Game

The common ground of gamification is to motivate a user, which at first sight seems to be a legitimate and good cause. For example, if a child in school receives a stamp for finishing the math exercise in a special manner, it could be extra motivated for the next one. However, the crossing of a red stoplight in favor of driving more ecofriendly could lead to unforeseen side effects of gamification, sometimes with negative consequences *(Zichermann, Linder)*. Besides that, there are several businesses which use a playful approach to cover themselves up, or to silence the user's self-reflection mechanisms, one of these being SHEIN.

The global fashion market went through a drastic disruption within the past decades. A woman in 1950 had to spend around 9$ for a new dress, which nowadays equals to roughly 72$. Today a dress costs around 12$ at Forever 21. Supply chains, materials and labor costs decreased and made it possible to offer clothes for such a low price. Entering the digital era, companies explored ways to better manage their resources, thus a just-in-time model evolved which produces what the consumer wishes and is flexible enough to adjust to ever changing trends *(Nguyen 2020)*. Starting in the 1990s with the fashion label Zara, the New York Times first described the phenomenon of fast turnround times as "fast fashion". Zara was able to take their garments from design into stores in just about 5 weeks, revolutionary at that time *(Hanbury 2021)*. Unsurprisingly competitors caught up in the race and the data driven society gave fashion new ways of production. Automated factories in which robots run around the clock are not uncommon, nor is it unusual to see around 500 new designs per week (Kozlowski 2020). All speed and product volume seem minor when compared to the new market leader, the fashion giant from China with 9000 new products daily: SHEIN, the company whose app was the most downloaded in the United States in 2021 and outrun Amazon *(Matsakis, Tobin, Chen 2021)*. The label likes to put itself into scene with vibrant fashion shows, live streams, musicians, and big portion of drama. Its website shows mostly extravagant models with curvy bodies in lewd. A t-shirt can be bought for as low as 3$, jeans for 6$ and with countless promotions customers can save another 20% on their purchase. According to Euromonitor, SHEIN is the largest fashion online store worldwide and primarily addresses young woman around the globe *(Hanbury 2021)*. It manages multiple websites, ships to 220 countries and operates from its headquarters in Guangzhou, China. Founded in 2008 by Chris Xu, it started as a shop for wedding dresses. In 2012 it expanded to women's wear and started its global breakthrough in 2015 *(Ibid.)*. Members of the 25+ generation most likely never heard of SHEIN as it completely relies on its online stores and social media channels. It utilizes sophisticated algorithms to predict trends and react on users in a fraction of the time its competitors need. For example, once a trend is discovered it is automatically transformed into a sewing pattern and send to one of the many subcontractors in Guangzhou. At the same time local garment facilitators provide the necessary material to start production. Depending on the difficulty of the pattern, local sewers either work on the whole product or only on parts of the item to send it to a more experienced sewer in a later step. All this happens in relatively small batches because a success is not guaranteed for the product yet. Finally, photos of the item in use complete the process and customers around the world can start shopping. Depending on how big the interest is, more can be produced on demand *(Kollbrunner 2021)*. That this approach is not flawless became visible countless times in the past, for example when the company released a swastika necklace or a Muslim praying rug. Moreover, SHEIN has being accused to steal designs in several cases *(Matsakis, Tobin, Chen 2021)*.

Despite being a fashion store, SHEIN's mobile applications are full of gamification elements which causes controversies as a lot of teenagers get addicted to the casino-like experience. Everything in the app is built around a point system, whereas 100 points equal to one US dollar *(SHEIN)*. For every dollar spend on products, customers receive one point back. However, most ways to gain points are free. For example, users receive notifications to open the app on regular intervals, SHEIN calls it checking-in, to browse the product catalogue and receive points for just spending time on the app. Another way to collect points is to participate in the many livestreams and events held regularly to promote the brand. Luck based games like spinning a wheel of fortune inside the app promise even more rewards.
Overall, countless and constantly changing mechanisms try to keep users bound to the app and ultimately to SHEIN's products, but one method stands out:
The option to receive points in return for giving product reviews. Once a product is purchased and delivered, users have the option to review the items. A simple text review brings 5 points, while photos give an additional 10 points. Moreover, users who publish their body measurements and information on how well a product fits are credited with an additional 2 points, which in total sums up to a maximum of 17 points per review (Ibid.). On the official help-page SHEIN writes: "If your post is selected as a featured piece by SHEIN editors, you will earn a bonus of 50 points." *(SHEIN, "About receiving SHEIN points")*, an additional 50 points sounds attractive to many users, unfortunately it is not clear what the criteria are to become picked by an editor. Users seem to think that reviews with many likes — appreciations by other users — are more likely to be selected. However, there is no evidence for this theory, and it is not even validated if editor picks exist at all, because all product reviews are displayed in the same manner. This does not stop people from begging for points or posting photos in the hope of receiving points. For SHEIN it must be a welcome opportunity: The more users buy a product and post a photo, the more new customers and potential buyers are attracted.

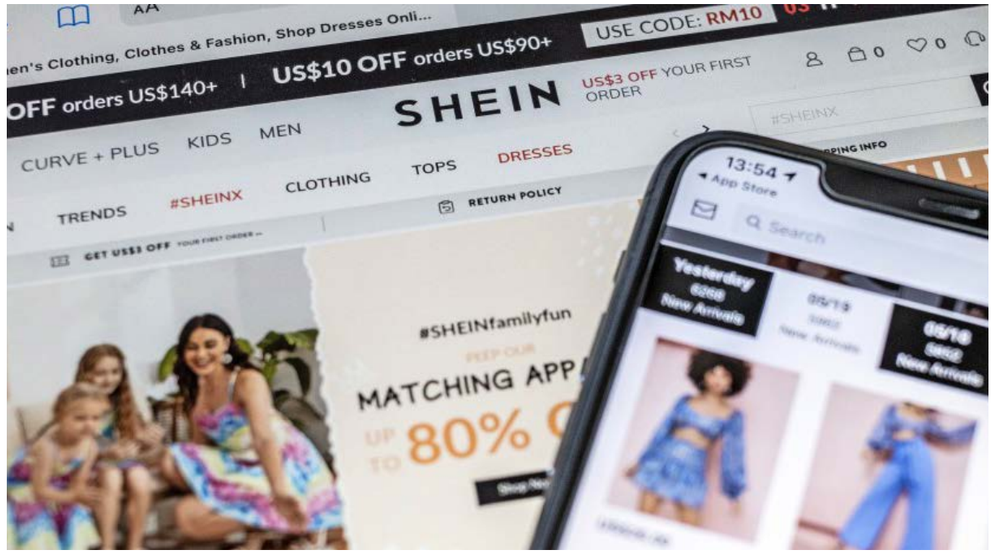

Figure 4: SHEIN website.

C.

The Public Privacy Project

Data privacy in its current form does not receive the relevancy it deserves. It is an abstract term, something a lot of people might find to complicated or too far away from their day-to-day routine. Although it moved into the focus of public news channels and recent legislations, people most often think on social networks as places where data must be protected. However, this cannot be enough in a data driven society. Algorithms are involved in many parts of our routine, sometimes more and sometimes less obvious. Therefore, it is important to develop competences on how to live responsibly with our digital twin. The Public Privacy project consists of real-world data, conducted from real users around the world. Moreover, its goal is to illustrate what can be read into private data, more specifically into photos. The final output will be a website to provide access to as many people as possible. In the end, a visitor should be able to draw a connection from the abstract practice of data protection to their own use with data. Thus, the communication artefact functions as a synthesizer between data and user. It provides meaning and guides through the process of self-reflecting the user's own behavior with personal information. On the other hand, it can only display parts of a broad spectrum. It is not a complete, nor definitive representation of issues concerning data privacy. Furthermore, the project does not aim to value, or to compare positive and negative aspects of our data driven society. It foremost addresses the possible risks of publishing personal photos to the internet.

C.1: Data Collection

The foundation of this project consists of data that is publicly accessible by visiting SHEIN's website. However, the key is to accumulate large amount of data to be able to search, categorize and reference within them, which by nature is not possible. In other words, the website does not provide an official way to collect its data, nor does it allow the user to search for something other than the products it offers. Furthermore, SHEIN — like many other webservices — implements safety measures to prevent automated data collection. This is a common practice as part of their business model heavily relies on the information which are exposed on the website. On one hand, making it difficult to protect business data is an understandable measurement. On the other hand, it is not impossible to collect data from protected websites. First, before collecting anything, it is important to know what data must be collected. Overall, the project is concerned about data provided by users, mostly pictures and texts users upload publicly. This information differs from private content — for example, a user's password — as they are usually not available without breaking into the system. Second, in this phase of the project it makes sense to keep a relatively broad mindset. Since it is not yet defined how the data will be further synthesized, or what data might be needed in later phases, everything that can be collected should be collected. Limitations can result from cost reduction which is often

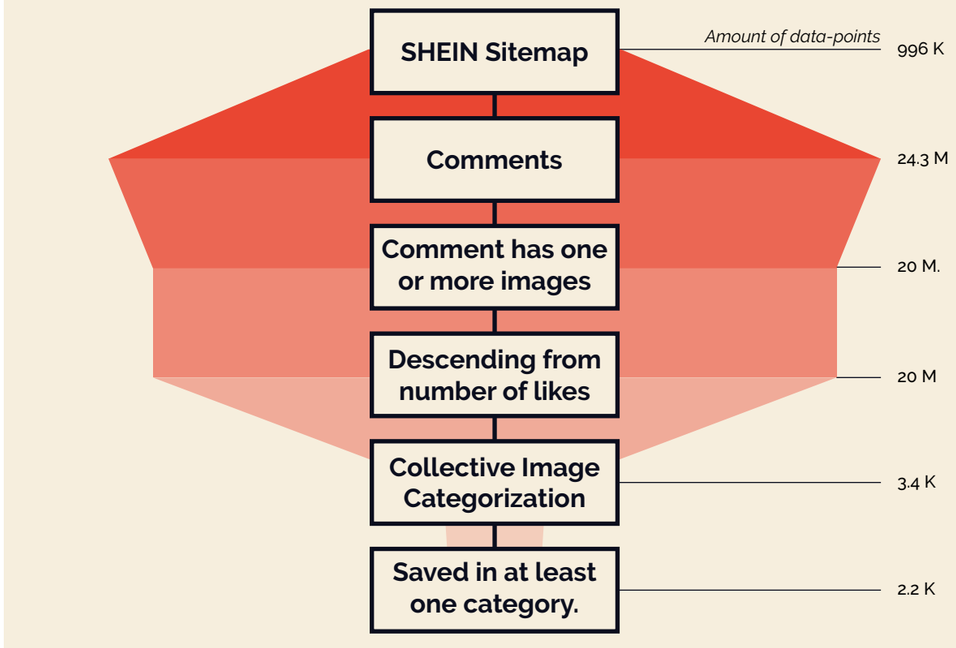

Figure 5: Methodology framework showing the data scraping pipeline.

comprised of two factors: the computing time it takes to gather the data and the cost of storing the data. A responsible planning from the beginning can reduce costs and become valuable during the later analysis.

C.1.1: Overarching Methodology

SHEIN orders their products by category, but also provides various filters to narrow down the search. In addition to that, it uses a pagination system which shows 120 products per page before the user must click a button to load the next page. When investigating how a website structures its data, it can be interesting to compare both desktop and mobile version. In the case of SHEIN, the products are listed in similar ways, however, the mobile version does not use pagination. Instead, it infinitely loads new items once the user scrolls to a certain threshold. For each product SHEIN has a detail list of attributes, these contain for example the materials, sizes, product images and information about the model seen on the images. An important attribute is the product id. It allows to uniquely identify the product across the catalogue which becomes an important aspect during the data analysis. Furthermore, the most interesting section is built by the reviews users can give to each product. Only users who bought a product can comment under it. The comment mainly contains text, optionally up to five photos and a size component which allows users to give feedback on how good the item fits. Now a common approach could be the creation of an automation that runs through every product page, saves the product attributes, and then crawls through the comment section. However, a difficulty on SHEIN is that when a user browses the catalogue and the web browser requests a page, it only loads the layout in the first step. In a second step, a script initiates the request of the product itself. In other words, when crawling the website, the automation would need to wait first for the initial layout and second for the complete loading of the content, which in the end turns out to be slow. A different approach makes use of an undocumented API, an application programming interface. Usually, an API provides easy access to developers when they want to connect their application with a another one. A request is sent to the server which then sends an answer. SHEIN is using the same technology for its mobile site, which can be detected by listening to the data traffic between the browser and the SHEIN server. By imitating this request, the server's answer can be processed directly and archived. It saves time, because the page does not need to load and instead of requesting one comment at a time, 20 comments can be loaded at once. Before starting with the data scraping, it is vital to determine which products should be scrapped. For example, it could be a random list of products, the first n products per category or all products of one category. For this project, the sitemap of SHEIN's international website was loaded. A sitemap can be seen as a directory with all pages of a website. Its function is to serve search engines as a lookup table, so they can index the entries and provide more relevant search results to users. SHEIN's sitemap contains around 900.000 items which were loaded into a database. From there each product was accessed one by one to in a random manner to scrape the product itself and its comments. In order to avoid blocking, a rotating proxy server redirected the traffic from the host to the SHEIN server. In other words, every request looked like it would come from a different source, to make it more difficult to block a specific origin of requests.

C.1.2: Data Storage Strategy

Scraping data from a website is only one side of the workflow. Depending on the amount of information being collected, it becomes important to develop a strategy of saving the data. While smaller collections can be managed by simple spreadsheet tools, such as Microsoft Excel or Google Tables, it makes sense to explore different solutions once the project reaches a bigger scale. The key is to act proactive and if possible, decide on a tool early in the beginning, as switching at a later point can be difficult and resource intensive. Several options include to work with a dedicated folder structure or multiple smaller documents instead of a single spreadsheet file. However, for this project a dedicated database software has been used. At the beginning of the decision process, the requirements to the project were listed:

- Accessibility: The data should be accessible not only from one device. This would open the possibility to use multiple instances of the scraper from different machines. For example, the scraper could run around the clock on a raspberry pi and during night times additionally from an office computer.

- Backups: Nothing would have been more disappoint than scraping data and losing it due to an error. Therefore, the storage strategy had to incorporate some sort of backup which would save the data in regular intervals and provide an easy way to restore it. Moreover, when owning a backup smaller mistake such as a wrong line of code in a newly implemented feature could be easily reverted.

- Scalability: By the start of scraping, the number of products was not yet known. It made it difficult to assimilate the potential disk space it would take to store the data. To not limit the project by choosing a database that could only handle a certain amount of datapoints, scalability became an important factor.

- Speed: Depending on how big the collection becomes, it can be a tedious process to filter, analyze and extract the relevant datapoints. It certainly can be useful to question if a system's hardware is fast enough for a pleasant workflow.

- SQL vs. NoSQL: While SQL databases are organized in fixed rows and columns like spreadsheets, the NoSQL approach uses json-objects, key-value pairs or nodes and edges. Since the scraping script returns its data in the form of objects, the database should support a NoSQL document technology. For example, a scraped skirt could contain 90% polyester and 10% cotton while a purse is made of 100% leather. In this case the document database provides a more flexible structure.

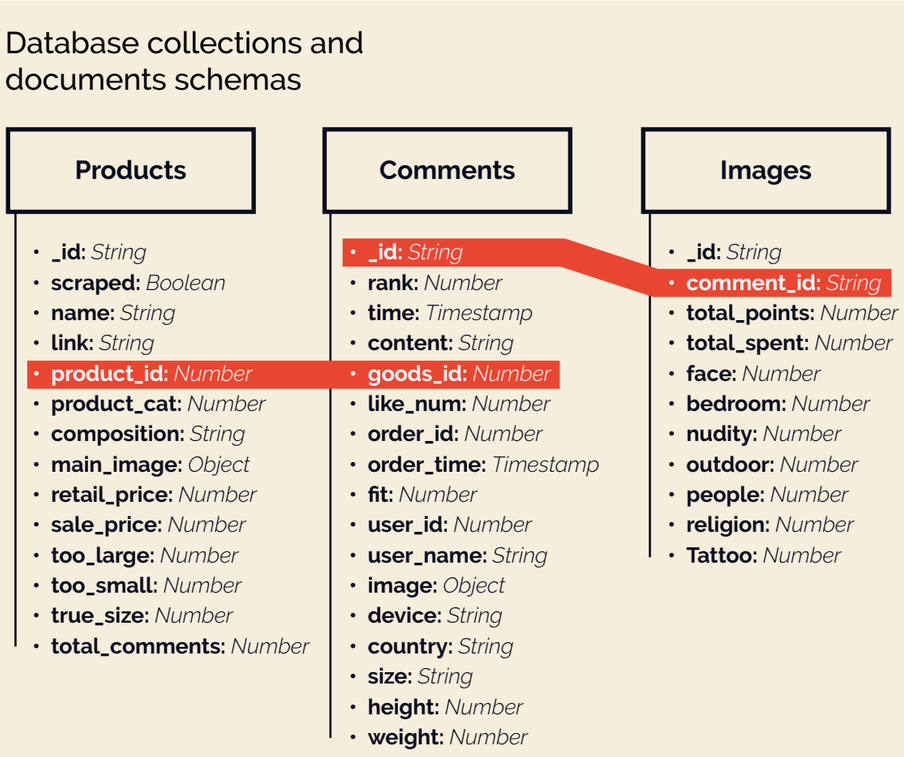

Figure 6: Database schemas.

After evaluating the requirements, it was decided to use a managed database from Digital Ocean. This solution ensured accessibility from various hosts and already included a backup solution at no additional costs. The minimum storage size of 15 gigabyte could be easily extended if needed but would also require a higher monthly charge. Since it is a managed service, the package includes maintenance and security updates. Moreover, it comes with a hardware configuration which could be scaled up if needed. The best part of this setup is that Digital Ocean offers a start credit of 100$ to new customers, enough to run the database for some months without the need to pay. This should be enough time to collect and analyze the data, because for the final website, the database will no longer be needed. However, one constraint should be noted before diving into the usage of a database service. In a configuration like the one mentioned before, it is necessary to integrate a backend server to communicate between a frontend and the database itself. It can be a simple Node.js server using express.js for API calls. Nevertheless, it is an additional cost and time factor which should be taken into account.

C.2: Data Analysis

Once the collection process has finished, the next step is to analyze the datapoints. Depending on how big the collection is it can be difficult to identify schemes and patterns. At this point the Public Privacy project collected a total of 90.000 products and 24.3 million corresponding comments. On one hand this was useful, because it gave great flexibility to search across the product catalogue and users. On the other hand, it made it difficult to filter relevant from irrelevant information. To overcome this challenge, a way to view, filter and browse had to be established. In other words, an explorative tool to navigate through each datapoint in an efficient manner.

C.2.1: Database Exploration

SHEIN Atlas serves as a tool to make each data point humanly readable and was created out of the need to map the current state of the project. The underlying technology consists of three parts: the database which contains all products and comments, a web interface which lets the user interactively navigate through the data, and a backend server which connects the database with the frontend through an API interface. Visitors are now able to access the whole catalogue of scraped information by entering search terms to search across all comments, or by selecting dedicated filters to include, respectively exclude comments with certain parameters. Each filter option is converted into a query-string, which is readable by the backend server and passed as a request to the database. The response is a json-object which by default limits to 1000 entries. Afterwards the object is presented to the user in form of a table, or optionally as a grid. Moreover, by scrolling through the page new items utilize the same search parameters to automatically extend the limit of entries. Once clicked on an entry, the customer's page opens. This page chronologically lists all comments of that specific user. Furthermore, it takes into account what the purchased items initially. This information is compared to the reward that was gained by writing a comment, uploading photos and publishing sizing information. Aesthetically SHEIN Atlas is kept relatively simple and neutral, both color and layout wise. This is simply because the content itself must be the primary focus of the tool. Furthermore, it serves as an internal tool to streamline the design process and therefore does not have the same visual requirements as a public catalogue. In the end it is a compromise between usability and time investment.

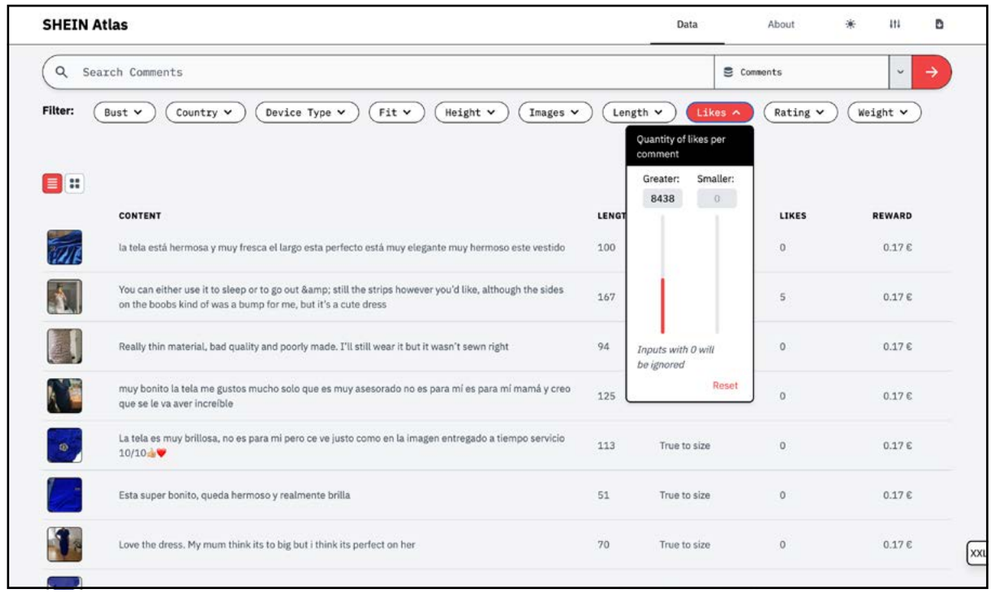

Figure 7: SHEIN Atlas website used to explore the dataset.

A first overview of the data conveys an interesting picture and proves the assumption that SHEIN competes the market by offering extreme low-priced products. Almost 56000 items are within the price range of 10 – 20 euros, which equals to almost 70% of all collected items. Only 5% of the products cost more than 30 euros. The top five categories are Women Dresses (82,9%), Women Jumpsuits (6,3%), Women Sexy Lingerie (3,4%), Women Blouses (0,2%) and Women T-Shirts (0,2%), with Women Dresses being by far the most popular category among the connection. This distribution does not allow estimations on the whole palette of products at SHEIN because the items were randomly loaded, however, it is important to keep this number in mind when evaluating the photographs in a further step. Another interesting insight is SHEIN's use of materials, more than 60% of the items utilize polyester as their foundation which is certainly on the cheaper side, but not sustainable. Furthermore, the most shopped country in the dataset is Mexico with around 18%, followed by the United States with 13% and Saudi Arab with 12%. European countries themselves have much less volume, mostly between 3% and 5%, but they appear frequently making Europe as a whole an important market for SHEIN.

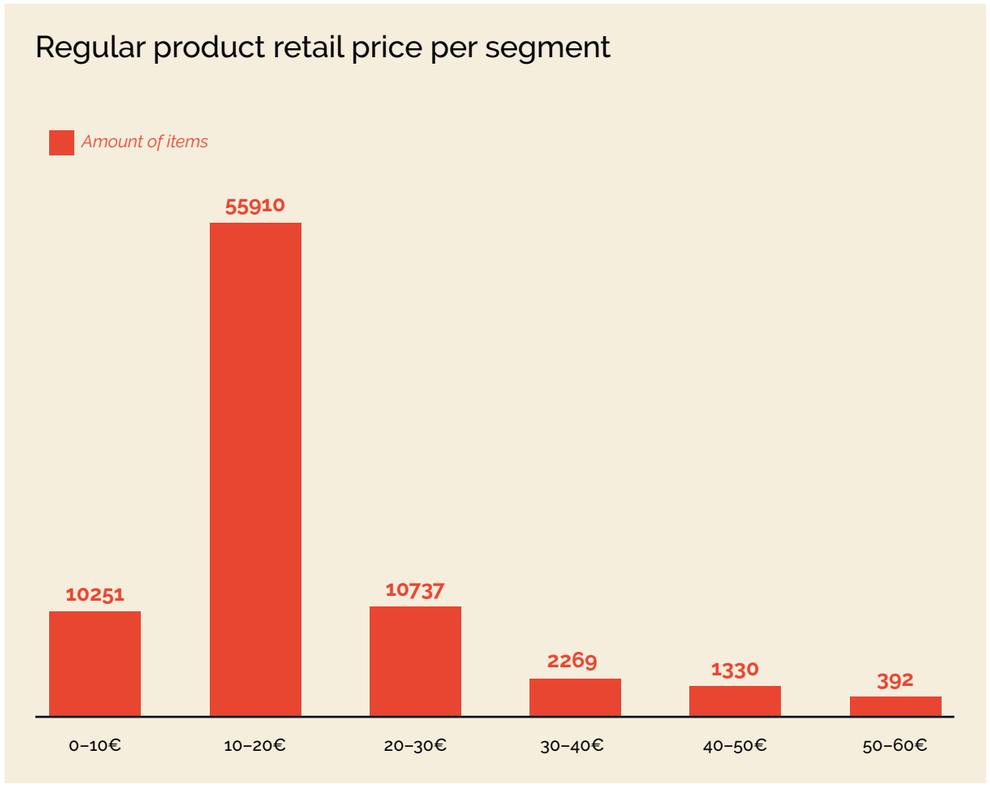

Figure 8: Price exploration per segment.

C.2.2: Collective Image Classification

Gaining a broad overview of the collected dataset is often a helpful first step but narrowing down millions of comments to the once relevant for the project requires some sort of filtering mechanism. A reproducible framework which is comprehensible for the end user had to be developed. After initial thoughts on the dataset, the most interesting elements turned out to be the photos users upload to their reviews. These photos cover a broad group of users, countries, ages, and cultures while not being depended on language. Moreover, they stood in direct connection to the user who uploaded it. In other words, a written comment cannot be validated for its truthiness, nor does it reveal much about its author. A photo instead could show the author and take them into the focus. However, while a lot of the images contain personal information in some sort, it is difficult to empirically measure the level or kind of information, especially regarding an abstract mater like data privacy. The challenge is to gain an objective look at the content of each image. Therefore, six categories were determined to sort the photographic content. Each category dealt with another potential privacy issues as they reveal information about the subject on the photo. Multiple categories could be applied to one image, because the photo could reveal more than one aspect of data privacy. The following categories became part of the framework:

1. Face: Photographs which show at least half of a person's face make it possible to access biometric data. A person's face biometry contains multiple measurements which can be used to recognize that individuum. In terms of privacy, it is relevant in several aspects, one being its criminal misuse, another one being unaware surveillance.

2. Place: The location of where a photo was taken can reveal a lot about subject's habits. This category contains outdoor locations which makes it possible to identify where a person was at the time of the photo being taken, but it also contains images that shows personal spaces, such as the subject's bedroom.

3. Religion: Most countries in the world have a liberal understanding of religion. However, the religious orientation is used to justify persecution and violence ever since. Revealing it reveals part of one's identity. Therefore, images in this category contain at least one religious' symbol, either directly worn by the person or seen in the near environment.

4. Tattoo: Most often tattoos serve an aesthetic purpose or symbolize the belonging to a group. Due to its permanent character, they are also used to uniquely identify a person. For example, tattoos are one aspect of image recognition used by law enforcement offices.

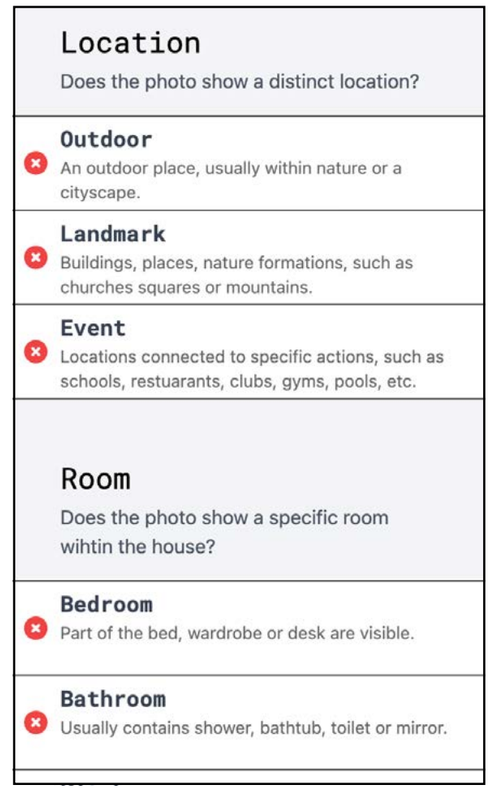

Figure 9: Tool to categorize images.

5. Nudity: Images that show naked or partly naked individuals belong into this category. This includes photos in lingerie, underwear, and bikinis. Being naked reveals some of the most vulnerable and intimate parts of the human body. Publishing intimate photos to the internet always bears the risk of misuse.

6. People: Uploaded photos which show more than one person certainly represent a special case, because it is not clear if all subjects are aware of the upload. In other words, it could be that the uploading person unconsciously infringed someone else's privacy.

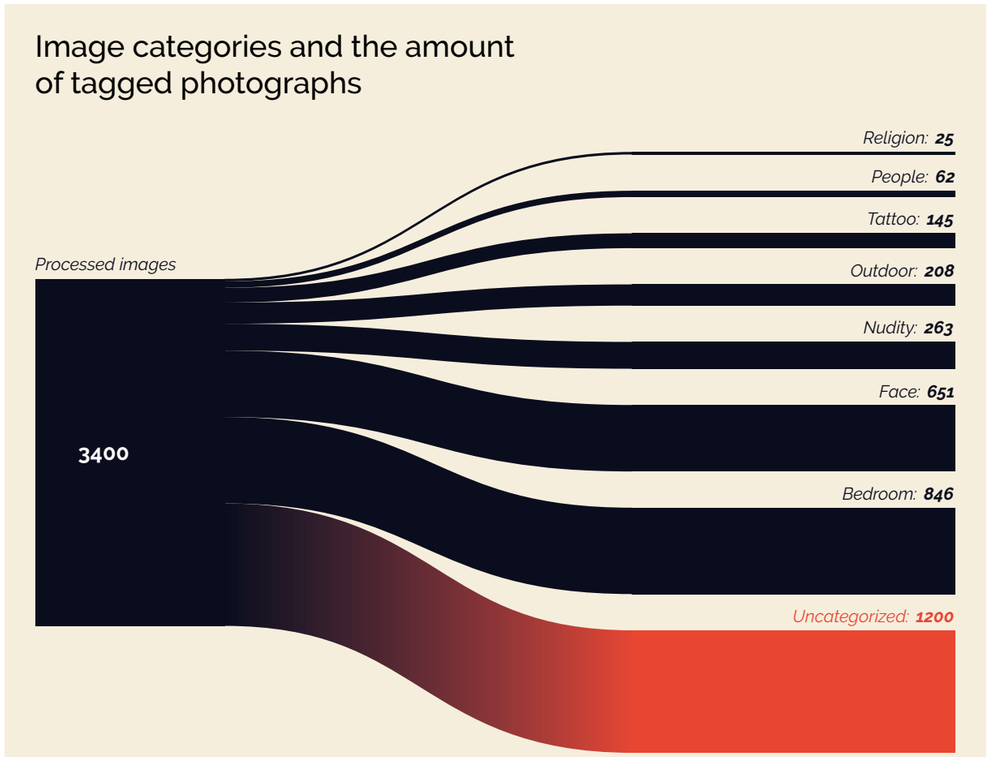

Figure 10: Tagged images per category.

Establishing these categories on paper is not that difficult. The challenging part is to take real data points and sort them accordingly. Image recognition tools were eliminated early on, because they would rely on pretrained data — an effort that would not be in proportion to the outcome. Rather than automating the process, the project utilized a semi-automated workflow: a web interface loads the image, the viewer must observe it and answer predefined questions about it. Based on the answer to these questions, the image will be tagged with any of the six categories. To reduce false tagging, each image will be locked for 48 hours after its first tagging. Afterwards it goes back to the pool of images, to be tagged again. Once an image is tagged at least twice in the same category, it is considered valid. Although this scheme could run on a random pool of images, it descends from the data point with the highest number of likes, as these are assumed to be more relevant for SHEIN users. Applying this method 3400 comments and their images could be extracted and tagged.

C.3: Data Publication

The concepts of Platform- and Surveillance Capitalism already expose what kind of power a few big technology companies hold in their hands. Sociologist Dr. Anat Ben-David advances on this phenomenon and warns that future generations will have less documentation on the early 21st century than on the beginning of the 20th century *(Thorsen 2020)*. This is due to the fact that knowledge is not democratized, rather than kept by a few profit-oriented businesses with a tendency for buying more and more data relevant resources *(Ibid.)*. Once we post something on Facebook, we give away control over it. Similarity, SHEIN declares in its terms of use:

> *By providing, submitting or posting a Submission, you agree to irrevocably license the Submission and all IP rights related thereto to the Company without charge and we shall have the royalty-free, worldwide, perpetual, irrevocable, transferable and sub-licensable right to use, reproduce, distribute, display, publish, present, recite, show, perform, sell, lease, transfer, transform, adapt, edit, shorten, delete, translate, arrange or otherwise change the Submission in its entirety or in part […] All Submissions shall automatically become the sole and exclusive property of us and shall not be returned to you and you agree not to raise any dispute in connection with any use of the Submission by us in the future. (SHEIN: B9)*

In other words, every comment, photo, or other content sent to SHEIN can be stored and used by the company in ways users have no control over. The public does not know what happens to the data and what insights are synthesized. Ben-David explains: "We think of archives as places where we can access knowledge, but in fact, archives act as gatekeepers that restrict the publics' access to records until they are no longer sensitive." *(Sofie Thorsen, "Counter-Archiving: Combating Data Colonialism")* Hence, the gap of knowledge and therefore power between companies and the general public steadily grows. A possible approach to work against this trend is the methodology of counter archiving. It collects, reorganizes and republished data in ways not possible before. In opposite to other data mining strategies in which datasets are gathered and then published, counter archiving consciously incorporates appraisal and discloses how the data was collected *(Ben-David 2020: 256)*. In a proof of concept, Ben-David developed Polibook as an archive of the Israelian parliament on Facebook. It contains every Facebook post of parliament members over a time span of four years. Users can enter a search query and are presented with a network of politicians: the closer politicians are to the search query, the more relevant they are to the topic. Different colors differentiate the parties and upon a click users can read the posts' texts. This reading allows users to escape Facebook's personalized news feed constraints and to draw connections that Facebook itself would not make visible. It exemplifies how most male Iranian politicians avoid topics of gender related issues, or how right-wing politicians refer to infiltrators when talking about to asylum seekers *(Ibid: 258)*.

C.3.1: Project Narrative

By utilizing the counter archiving framework, the Public Privacy project enables user specific reading to personal data. It includes how much users spend on the shopping website and compares it to the value of rewards gained through the purchase and the review of products. The goal is to communicate how sensitive data is published by users to receive a relatively small reward — in form of SHEIN points. The project is realized as a website to provide a broad international access to the information.

Additionally, it starts with English and German language support and the capabilities to extend these through future updates.

Figure 11: Mobile view of the data grid.

C.3.2: Visualizing Privacy

Publishing private content is a delicate task. On one side it must be anonymized to protect personal information, on the other side this project aims to visualize privacy issues. The designer's role is to not only combine both of these aspects, but also to deliver a clear and memorable message. Hence, the challenge of this project is to isolate the privacy issues for each of the six image categories. During first tests the idea of measuring privacy by calculating the percentage of pixels revealing such data was discarded, because it could not provide the expected reading into the photographs. For example, a photograph taken in public reveals information about the location of where it was taken. By counting how many pixels refer to the location, the percentage in relation to the whole image could be calculated. However, the result is still too abstract for the purpose of this project. Moreover, a single image would never be as powerful in its message as multiple images, just like single data points are nearly useless without knowing about the whole collection. Therefore, another iteration on the problem lead to the choice of applying each image category to an individual filter. For illustrating how biometric data is extracted from faces, photographs of the face category were fed into facial recognition software. The result consisted of images with biometric data. Indoor and outdoor photographs both put the subject itself into a secondary position and primarily focused on what the background looks like. In order to strengthen background elements, machine learning tools were used to remove the models themselves from the picture. Furthermore, the nudity category made use of a machine learning algorithm which turned realistic photographs into the style of classic oil paintings. This method allows to display nudity without interfering with the model's privacy. The people category, which contained images that contained two or more people in them, utilized a technique in which the silhouette of each figure was extracted resulting in images that highlight the character count. Finally, the last two categories being religion and tattoos only anonymized potential faces but did not use any further tools. Instead, their thumbnails contain a zoomed in version of their corresponding religious symbols, respectively tattoos.

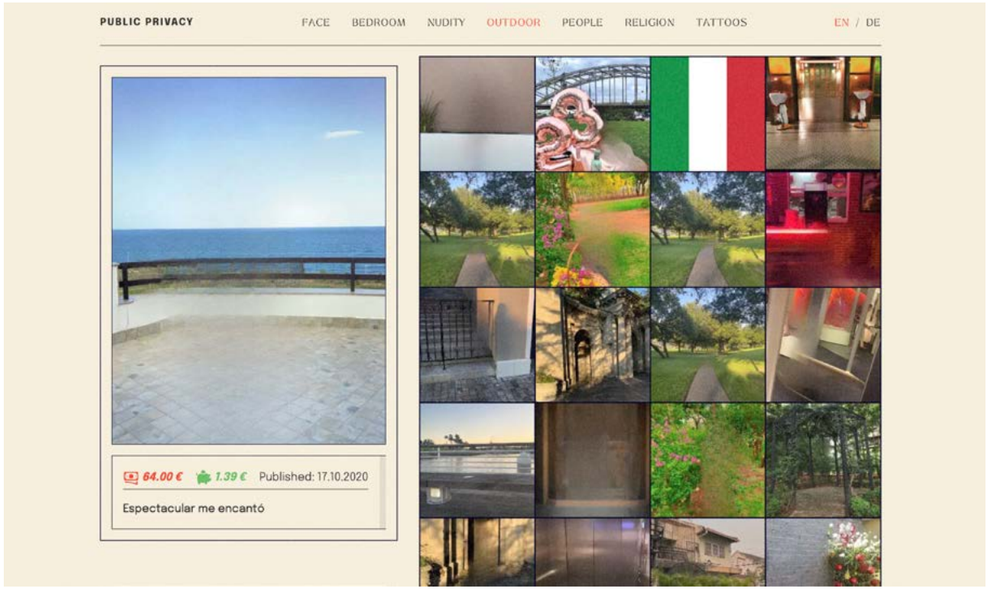

Figure 12: Desktop view of the data grid.

C.3.3: The Data Grid

While designing the layout for each category, two options were considered. The first one focused on a central enlarged image without any further reference to the rest of the dataset. The idea was to avoid any possible distraction and unnecessary side information. The photograph as the central item of meaning should be presented as such. However, this approach would leave out what makes Big Data so powerful: Seeing the whole picture with the ability to zoom into the information. Therefore, the second option containing an image grid was used in the final version. The grid provides a large first overview, but also enables the display of more detailed information. Whenever users hover over an item in the grid, they are presented with the total amount a user spent on SHEIN. Additionally, the thumbnail disappears, and the original image shows up in an enlarged version. Further context can be gained through the review's text below the image frame.

D. Conclusion

D.1: A Data Driven Society

In the 21st Century, data is no longer just a string of ones and zeros, it is a currency. It comes along with power and gives those who own it control. Big Data, the phenomenon of large data accumulations with extended reading capabilities, enables new and innovative perspectives on a variety of problems. It helps researchers and scientists to better understand the world we live in, but also gives advertising companies the tools to create personalized profiles of individuals. At least since then Big Data is of economic interest and continuously attracts more attention. Thereof, the platform model as a specialized form of business evolved and established monopolies in many areas of our live. It bears the risk of becoming depended to certain services when there are no alternatives to it anymore. Furthermore, we have to keep in mind that Big Data cannot be the answer to all our problems: While it often helps, it can also cause more harm than good especially when it does not follow liberal ethics anymore. Algorithms can be bias, and companies often follow their own rules.

One way to open new data sources and keep existing ones is rooted in gamification, the practice to add game like elements to non-game contexts. The strategy is widely used by many companies and motivates users to stay close to the brand. With reference to SHEIN, the biggest fashion retailer in the world, gamification attracts users to upload personal content in return for rewards. People publish pictures of themselves in private settings and accept that their content becomes property of SHEIN. By creating a dataset of user reviews, it was possible to collectively classify customer photos into six categories. Each category exposed a different aspect of data privacy and was published as part of a counter archive, allowing to browse the data in new ways.

Data is omnipresent and we often cannot imagine living without it anymore: from the tiniest houses to the biggest global companies, IoT devices, smartphones, social networks and search engines, data has become our daily companion. What is missing is a general understanding for the impact of technology, algorithms, and data in general, because the trend of transformation will rather speed up than stop. As a consequence, the gap between a few privileged companies and the idea of democracy threatens to become insuperable. Apart from that, data gives us the power to make our life more pleasant in many ways. All it requires is more responsibility.

D.1.1: Reflection

The following chapter does not contribute to the project itself rather than presenting an explaining of my motivation to pursue it. I am going to mention the challenges and how I overcame them, the design decisions and compromises I had to take and the aspects that I would like to invest further work into.

At first, this project started from the interest in exploring large datasets and the curiosity of what knowledge they might hold. I had always an interest in how technology affects society and vice versa, I wanted to understand why some people cover their apartments with smart home devices while others do not even want to deal with online banking. I wondered why newspapers only refer to social media when they talk about data privacy and why people delete certain apps from their phones but keep others. To use SHEIN as a case study happened more coincidentally than planned and was rooted in an advertisement I came across in the vastness of the web. I quickly investigated SHEIN's website and read articles about their relevancy within the fashion market. Two aspects on the website finally settled my decision: SHEIN's extremely cheap products in combination with the international fan base and the customers themselves who posted numerous comments below each product.

The first task was to collect and save data from SHEIN and it already yielded my first challenge. At that time my basic coding understanding was limited to interfaces and did not go much beyond little web applications. Hence, I had to research how to technically retrieve scrape data from the web. One option turned out to be services which attracted customers with pretty user interfaces and promised a hassle-free method to retrieve the needed data. In my opinion, they did not provide enough flexibility for my project. For example, with these services I had to stick to certain data limits, formats, and strategies to store the results. Moreover, most of these tools came at a cost which I would rather spend on something else. So, I went with the second option and started multiple attempts to code such a tool myself. The coding itself turned out to not be very difficult, however, to find the best strategy was quite a challenge for me. With my limited experience I had to go with the attempt of trial and error, but what helped me was to look at other people's code to see what strategy they followed. All of this happened early in the project, so it was no problem for me to experiment, although I would recommend developing a timeline and stick to it in order to not get lost in the process.

Once I had all data together the next problem was to interpret it. It took me several iterations to determine which aspects of the data are interesting and worth to focus on. Looking back to it, it would have helped me a lot to create an overview for myself of the data I had access to. Even a simple mind map can work wonders because it sorts the floating ideas and brings them to paper. At this point in the project, I started to take a closer look at the photos that users had uploaded. I set on the aspects of data privacy but was not able to give properly name potential issues in the images. In this situation it helped me to categorize what was visible in each of the images; to sort the data I collected made it readable in ways not possible before. For instance, a photo showing a person in the foreground and a bedroom in the background felt important, but I could not name why. Only after categorizing the images and looking at all photos with bedrooms in the background made me realize that people share one of their most private places in the world publicly on the internet. This brings me to a next important step within my process: How do I categorize thousands of images in an effective and short time?

The idea to create a tool that allows multiple users to label images emerged from the lack of alternatives. My objective was to sort as many images as possible into six categories. The categories and the photograph's content were sometimes so exceptional that I was not able to make use of image recognition software or machine learning models within an appropriate timeframe. On the other hand, it was also no option to manually open every image and write its category onto a spreadsheet. To code myself a tool that could be used by me and others offered two benefits: I would not have to do the work of tagging

all by myself and it speeded up the process by packaging everything within a usable interface. I weighted the extra effort of creating such a tool against the benefits and started coding. In the end, I spend one weekend on the coding and within two weeks of use more than 3000 images were reviewed by me and others.

Works Cited

Anderson, George. 2021. “Amazon Gamifies Warehouse Worker Tasks, Raising Experts’ Concerns.” Forbes, March 23.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “ProPublica.” Machine Bias — There’s software used across the country to predict future criminals. And it’s biased against blacks. May 23. Accessed March 23, 2022. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ben-David, Anat. 2020. “Counter-archiving Facebook.” European Journal of Communication, 249–264.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Conference on Fairness, Accountability, and Transparency. New York. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Canalys, & Statista. 2021. “Cloud infrastructure services vendor market share worldwide from 4th quarter 2017 to 3rd quarter 2021 .” Statista. October 28. Accessed March 23, 2022. https://www.statista.com/statistics/967365/worldwide-cloud-infrastructure-services-market-share-vendor/.

Catherine O’Connor, Mary. 2012. “IBM patent sees sensors as high-tech floor boards.” ZDNet. April 23. Accessed March 23, 2022. https://www.zdnet.com/article/ibm-patent-sees-sensors-as-high-tech-floor-boards/.

Cukier, Kenneth, and Viktor Mayer-Schönberger. 2013. “The Rise of Big Data: How It’s Changing the Way We Think About the World.” Foreign Affairs, May/June: 28–40.

Daniels, Benjamin. n.d. “5 examples of gamification that make Duolingo the best at user retention.” StriveCloud. Accessed March 23, 2022. https://strivecloud.io/blog/gamification-examples-boost-user-retention-duolingo/#1f094.

—. n.d. “5 examples of gamification that make Duolingo the best at user retention.” StriveCloud. Accessed March 23, 2022. https://strivecloud.io/blog/gamification-examples-boost-user-retention-duolingo/#1f094.

Dastin, Jeffrey. 2018. “Amazon scraps secret AI recruiting tool that showed bias against women.” Reuters. October 11. Accessed March 23, 2022. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Deterding, Sebastian, Rilla Khaled, Lennart Nacke, and Dan Dixon. 2011. “Gamification: Toward a definition.” CHI 2011 Gamification Workshop Proceedings. 12–15.

Ensheng, Dong, Du Hongru, and Lauren Gardner. 2020. “An interactive web-based dashboard to track COVID-19 in real time.” The Lancet Infectious Diseases. February 19. Accessed March 23, 2022. https://doi.org/10.1016/S1473-3099(20)30120-1.

Finn, Ed. 2017. What Algorithms Want — Imagination in the Age of Computing. The MIT Press.

Hanbury, Mary. 2021. “How China’s most mysterious billion-dollar company, Shein, won over US teens and became TikTok’s most-hyped fashion brand.” Insider. October 5. Accessed March 23, 2022. https://www.businessinsider.com/shein-china-billion-dollar-company-fast-fashion-brand-2021-8?r=US&IR=T.

Harada, Jun. 2021. The Instagram ads Facebook won’t show you. May 4. Accessed March 23, 2022. https://signal.org/blog/the-instagram-ads-you-will-never-see/.

Infinite Styles Ecommerce Co., Ltd. n.d. SHEIN. Accessed March 23, 2022. https://eur.shein.com/bonus-point-program-a-371.html.

Kendall, Graham. 2019. “Your Mobile Phone vs. Apollo 11’s Guidance Computer.” RealClear Science, July 2.

Knozlowski, Anika. 2020. “Fashion’s “Sustainability” Endeavors Need to Be About More than Fabrics, Recycling.” The Fashion Law. August 14. Accessed March 23, 2022. https://www.thefashionlaw.com/fast-fashion-sustainability-is-about-more-than-the-fabrics/.

Kollbrunner, Timo. 2021. “Toiling away for Shein Looking behind the shiny façade of the Chinese “ultra-fast fashion” giant.” Public Eye. November. Accessed March 23, 2022. https://stories.publiceye.ch/en/shein/index.html.

Matsakis, Louise, Meaghan Tobin, and Wency Chen. 2021. “How Shein beat Amazon at its own game — and reinvented fast fashion.” rest of world. December 14. Accessed March 23, 2022. https://restofworld.org/2021/how-shein-beat-amazon-and-reinvented-fast-fashion/.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. Big Data: Die Revolution, die unser Leben verändern wird. Redline Verlag.

Meehan, Mary. 2019. “Data Privacy Will Be The Most Important Issue In The Next Decade.” Forbes, November 26.

Meta Platforms. 2022. “Meta’s (formerly Facebook Inc.) annual revenue from 2009 to 2021, by segment (in million U.S. dollars).” Statista. February 2. Accessed March 23, 2022. https://www.statista.com/statistics/267031/facebooks-annual-revenue-by-segment/.

Meta Platforms, Ink. n.d. “The Facebook pixel: A piece of code for your website that lets you measure, optimise and build audiences for your advertising campaigns.” Meta for Business. Accessed March 23, 2022. https://www.facebook.com/business/learn/facebook-ads-pixel.

Mills, Terence. 2019. “Five Benefits Of Big Data Analytics And How Companies Can Get Started.” Forbes, November 6.

Morozov, Evgeny. 2011. The Net Delusion. New York: PublicAffairs.

Mui, Ylan Q. 2011. “ Business Little-known firms tracking data used in credit scores.” The Washington Post — Democracy Dies in Darkness. July 16. Accessed March 23, 2022. https://www.washingtonpost.com/business/economy/little-known-firms-tracking-data-used-in-credit-scores/2011/05/24/gIQAXHcWII_story.html.

Nguyen, Terry. 2020. “Fast fashion, explained.” Vox. February 3. Accessed March 23, 2022. https://www.vox.com/the-goods/2020/2/3/21080364/fast-fashion-h-and-m-zara.

O’Neil, Cathy. 2016. Weapons of Math Destruction — How Big Data Increases Inequality and Threatens Democracy. New York: Crown New York.

Rodríguez Aranda, Catalina. 2020. “The potential of Duolingo to increase students’ learning and motivation.” Universidad de Cádiz. September 2. Accessed March 23, 2022. http://hdl.handle.net/10498/23812.

Rudder, Christian. 2015. Dataclysm: Love, Sex, Race, and Identity — What Our Online Lives Tell Us about Our Offline Selves. Broadway Books.

Schneier, Bruce. 2015. Data and Goliath — The Hidden Battle to Collect Your Data and Control Your World. W. W. Norton.

Snowden, Edward. 2015. Reddit. Accessed March 23, 2022. https://www.reddit.com/r/IAmA/comments/36ru89/just_days_left_to_kill_mass_surveillance_under/crglgh2/.

Srnicek, Nick. 2017. Platform Capitalism. Great Britain: polity.

Syzygy. 2018. “Opinions on the use and the worth of personal data in the EU and U.S. in 2018, by country.” Statista. Accessed March 20, 2022. https://www.statista.com/statistics/1009192/opinion-on-use-of-personal-data-online-by-country-in-eu-and-us/.

Thorsen, Sofie. 2020. “Counter-Archiving: Combating Data Colonialism.” Medium. November 16. Accessed March 23, 2022. https://medium.com/copenhagen-institute-for-futures-studies/counter-archiving-combating-data-colonialism-be17ffead4.

Zichermann, Gabe, and Joselin Linder. 2013. The Gamification Revolution: How Leaders Leverage Game Mechanics to Crush the Competition. McGraw Hill.

—. 2013. The Gamification Revolution: How Leaders Leverage Game Mechanics to Crush the Competition. McGraw Hill.

Zuboff, Shoshana. 2019. Surveillance Capitalism and the Challenge of Collective Action. New Labor Forum.

—. 2019. The Age of Surveillance Capitalism. New York: PublicAiifairs.