DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Supporting Nonexperts in Choosing Appropriate Privacy-Enhancing Technologies

Mohammad Faisal Bin Ahmed

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Supporting Nonexperts in Choosing Appropriate Privacy-Enhancing Technologies

Unterstützung von Nichtexperten bei der Auswahl angemessener Technologien zur Verbesserung der Privatsphäre

Author:	Mohammad Faisal Bin Ahmed
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Oleksandra Klymenko, M.Sc.
Submission Date:	14th of July 2021

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 14th of July 2021 Mohammad Faisal Bin Ahmed

Acknowledgments

Firstly, I would like to thank my parents and wife for supporting me throughout the duration of this thesis. I would also like to thank Prof. Dr. Florian Matthes for providing me the opportunity and guidance to conduct this research. His suggestions and help during every stage of the thesis were extremely thoughtful. I am especially grateful to my advisor Oleksandra Klymenko for taking her time and patience to support and guide me every week. Thanks to her motivation and push, I was able to complete this research work in time despite being sick with COVID-19 for almost three weeks. Finally, I would like to thank the people who have helped me with informal feedback and thesis related queries. Without their help, the thesis may not have reached the stage it is in today.

Abstract

As the world gets more conscious about privacy and data security, stakeholders are relentlessly searching for better and more tailored privacy-enhancing technologies (PETs). As a result, the field is booming and there are so many such technologies available, it is easy to get overwhelmed with the complexities and options. Hence, the main purpose of this thesis is to look at the state of the art in PETs, analyze different methods' pros and cons in various scenarios, and finally create a guidance system that would help nonexperts choose a suitable privacy-enhancing technology for their specific use cases.

Kurzfassung

Da das Bewusstsein für Datenschutz und Datensicherheit weltweit zunimmt, sind die Beteiligten unermüdlich auf der Suche nach besseren und maßgeschneiderten Technologien zum Schutz der Privatsphäre (Privacy-Enhancing Technologies, PETs). Infolgedessen boomt das Feld und es sind so viele solcher Technologien verfügbar, dass man leicht von der Komplexität und den Optionen überwältigt wird. Daher ist der Hauptzweck dieser Arbeit, den Stand der Technik bei PETs zu untersuchen, die Vor- und Nachteile verschiedener Methoden in verschiedenen Szenarien zu analysieren und schließlich ein Leitsystem zu erstellen, das Nicht-Experten bei der Auswahl einer geeigneten Technologie zur Verbesserung der Privatsphäre für ihre spezifischen Anwendungsfälle hilft.

Contents

Acknowledgments **iii**

Abstract **iv**

Kurzfassung **v**

1 Introduction **1**
1.1 Motivation 2
1.2 Research Questions 3
1.3 Goals 4

2 Background **7**
2.1 Privacy 7
2.2 Privacy Enhancing vs. Privacy Preserving 7
2.3 Privacy-enhancing Technology 8
2.3.1 Differential Privacy 9
2.3.2 Federated Learning 9
2.3.3 Homomorphic Encryption 10
2.3.4 Secure Multiparty Computations 10
2.3.5 Zero Knowledge Proof 10
2.3.6 Obfuscation 11
2.3.7 Pseudonymization 11
2.3.8 Synthetic Data Generation 11
2.4 Known Use Cases 12
2.4.1 Cryptocurrency 12
2.4.2 Smart assistance 12
2.4.3 Digital Rights Management 13

3 Research Design **14**
3.1 Protocol 14
3.1.1 Defining objectives 15
3.1.2 Formulating research questions 16
3.1.3 Conducting systematic literature review per research question 17
3.1.4 A Novel way of doing Literature Review 31

3.1.5 Final Findings 33

4 Findings **36**
4.1 Status Quo 36
4.2 Identified Privacy Enhancing Technologies 38
4.2.1 Differential Privacy 39
4.2.2 Federated Learning 40
4.2.3 Zero Knowledge Proof 40
4.2.4 Homomorphic Encryption 41
4.2.5 Secure Multiparty Computation 42
4.2.6 Obfuscation 42
4.2.7 Data Perturbation 43
4.2.8 Anonymization 43
4.2.9 K-anonymity 44
4.2.10 L-diversity 45
4.2.11 T-closeness 46
4.2.12 Pseudonymization 46
4.2.13 Randomization 46
4.2.14 Data Masking 47
4.2.15 Synthetic Data Generation 47
4.3 Identified Criteria 48
4.3.1 Statistical Disclosure 48
4.3.2 Data Publication 48
4.3.3 Distributed Data 48
4.3.4 Edge Devices 49
4.3.5 Data Transportation 49
4.3.6 Need of Real Data 49
4.3.7 Computational Cost 49
4.3.8 Multiple Stakeholders 49
4.3.9 Third Parties 49
4.3.10 Losslessness 50
4.3.11 Accuracy 50
4.3.12 Textual data 50
4.3.13 Data Dimension 50
4.4 Mapping Data 50
4.5 Key Insights 54

5 Related Work **55**
5.1 Categorization of Privacy-enhancing technologies 55
5.2 Privacy metrics identification 56

5.3 Overview of popular privacy-enhancing technologies in the financial sector . . . 56

6 Web App Implementation **58**
6.1 Architecture . . . 59
6.2 Design . . . 60
6.3 Easy Mode . . . 62
6.4 Strict Mode . . . 64
6.5 Data Schema . . . 65
6.5.1 Question Data . . . 65
6.5.2 PET Data . . . 65
6.6 App elements and features . . . 66
6.6.1 The Homepage . . . 66
6.6.2 Question panel . . . 67
6.6.3 Side panel . . . 68
6.6.4 Edit Answer . . . 70
6.6.5 Result section . . . 70
6.6.6 Result Card . . . 71
6.6.7 Quick view information . . . 72

7 Conclusion **74**
7.1 Summary . . . 74
7.2 Limitation . . . 76
7.3 Future Work . . . 76

List of Figures **78**

List of Tables **79**

Listings **80**

Bibliography **81**

1 Introduction

Privacy enhancement or preserving[1] is a necessary procedure any professionals that deal with data must endure. However, the world of privacy is a twisted maze for professionals with untrained eyes. To select a particular technology that suits a specific need, considering complex scenarios and variables that contributes to that need is not an easy task. This difficulty becomes ten fold when there is seemingly no standard or precedence to base your decision onto.

Privacy-enhancing technologies are a relatively new phenomenon in the world of Privacy [1]. As the world today collects more data than ever before, it is evidently more important than ever to protect the privacy of the data subject from potential data theft, unwanted profiling, and leaks [2]. As people are getting more concerned about their privacy, the goal to protect privacy has become a vital business decision for companies, not only a moral obligation [3].

With governments and regulatory bodies forcing new laws that ensure data privacy, companies and professionals are taking necessary steps to meet these criteria [4]. However, these steps are often convoluted and do not follow any standard protocol due to the lack of such [5]. Companies and professionals often have to go through rigorous and time consuming processes to land on a privacy enhancing technology that is suitable for them. However, many who are starting their career in the field rarely possess the necessary insights and experience to follow this route [6]. For non-experts[2], this can be deemed very challenging given their limited background and resources.

While the goal of this thesis is not to create an end-all-be-all general guideline to privacy enhancing technologies, this thesis will however, try to reduce the barrier to entry in the field of privacy enhancing technologies. Throughout this research process, I had a few major goals in my thesis; one of them is: how to simplify the privacy enhancing technology selection process. By doing so, I faced many challenges

[1] Privacy enhancing and privacy preserving terms are used interchangeably. For the literature, privacy enhancing will be used.

[2] A person who is not very experienced or well-versed in the field. In this case, a privacy professional who is new to the world of privacy enhancing technology.

myself and discovered other limitations and discrepancies felt by someone on a similar journey. By learning from these limitations, I would constantly tweak my process and dynamically adjust my research methodology. This paper would share more about this process in detail.

1.1 Motivation

The lack of a proper protocol and guideline to choose any particular privacy enhancing technology was very evident [3]. During my pre-research period, I had to scour through pages after pages, academic literature after literature to find any reference to a general guideline that will help practitioners make informed decisions based on scientific evidence. The lack of a general guideline was hence identified.

The field of privacy enhancing technology is generally considered relatively new in the sea of privacy related research topics [7]. For an original research, finding an even more niche topic brings a lot of challenges forward. Normally to identify a new research topic, one has to have prior knowledge in the field: to get a headstart when searching for potential topics. Before finding the aforementioned research niche, I had to dig through countless research papers: both in the field and other related fields, so that the topic would be as broad as possible.

At first I wanted to know the current status of research in this field. As mentioned earlier, there was not much research done for this particular niche to get an informed position regarding the status quo. The initial pre-research goal was to learn more about the existing privacy enhancing technology from literature of all types. I initially identified a few privacy enhancing technologies that are quite popular and prevalent among practitioners [8]. This gave me a good idea about the types of technologies that exist in the field and their related research scope.

By identifying the initial set of technologies, it became quite evident to me why these particular technologies are used in the cases they are. For example, why should one pick one technology over the other? What metrics dictate this decision, and are there any more unidentified metrics or scenarios available?

Here is a potential scenario where this research can come in handy as a guide.

[3] No literature providing guidance system for privacy-enhancing technologies was identified

Scenario: Felix is a tech lead at a VoD[4] start-up in Munich. He wants some analytics to improve their performance, but he can't send the data to a third-party analytics firm without ensuring proper privacy enhancement first. The data they collect includes name, age, UUID[5], watch history, device info, and payment info. What type of privacy-enhancing technology should Felix use to ensure the privacy of the data subjects is not violated during the analysis? His concerns are: user privacy, potentially malicious analytics firms, data breaches during transit or rest, GDPR violations.

Normally, Felix would have to spend countless hours researching about many different privacy enhancing technologies and their pros and cons, since he's new to the technology, maybe the company would have to hire a third party consultant to find a solution. Both of which are really bad choices for them.

However, using this systematic research as a guidebook and answering a few questions, Felix can get a suitable privacy-enhancing technology for his scenario. In this case, the technology is called Homomorphic Encryption.

1.2 Research Questions

The research will focus on how to create a decision support system for privacy enhancing technology. To traverse through the process, there are three distinct tasks that have to be done. I have formulated these three tasks on the basis of the research as the research questions. These questions were simplified as much as possible while retaining their distinct differences from one another.

Research question 1: What are the state-of-the-art privacy-enhancing technologies and their distinct characteristics?

The question is self-explanatory. This was the basis of how the further research question will play out. In order to create a general guideline to suggest privacy-enhancing technologies, one has to identify the different technologies first. The task here is to formulate a research protocol to conduct this identification process. Finally using the protocol to identify as many privacy-enhancing technologies as possible is a systematic way. The associated distinct characteristics will also need to be identified for a successful mapping and suggestion mechanism.

[4]Video on Demand. Streaming services like Netflix
[5]Universally Unique ID: it's generated and used as an unique ID in many systems

Research question 2: What are the different scenarios where privacy enhancement is required and their corresponding basic requirements?

The next step would be to identify the scenarios that dictate where privacy-enhancing technology will be required. By identifying these scenarios, one can get a picture of how this privacy-enhancing technology works in those situations and potentially why one particular privacy-enhancing technology is preferred over the other. These scenarios or criteria will work as a metric to determine a PETs appropriateness in varying conditions. A nonexperts may use any combination of criteria to determine their preferred technology.

Research question 3: How to guide the decision-making process to allow diverse stakeholders make informed decisions about the use of an appropriate privacy-enhancing technology?

The final task is the marriage of privacy-enhancing technologies identified so far and the criteria one must consider when making an informed decision. However, this is only part of the solution. While mapping gives us the connection between these two entities, it's also important to find an optimum way to present this information to the targeted audience of this guideline. There are multiple ways to do this and more about this will be discussed later in this paper.

1.3 Goals

It is repeated multiple times throughout this document that the main goal of this research is to deliver a decision supporting guideline or system for people who are not expert in privacy-enhancing technologies. However, we have to go into a bit more details on that. How the guideline will look like, its usability and other aspects.

The main deliverable of this research is the mapping data that the decision support system will be based on. Using a systematic literature review, the identified privacy-enhancing technologies and the relevant criteria will be mapped and presented in a way that can be used to select an appropriate technology(s) for the said criteria.

Initially, the plan was to create a tree-like decision making flowchart that will help non-experts make an informed decision. However, as the research went on and more and more data came to light, it became apparent that creating a tree-like structure to visualize the decision support system would be too complicated to be able to articulate by a human mind. More details about this will be discussed in the

limitation portion of this document.

The second goal was to create a prototype for the decision making system. I will be using web based tools to design and develop this prototype. This has a few functional requirements as it will be used mainly by non-experts. These requirements will also be discussed later in the document.

Here are the prototypical mock-ups of how the web app should look:

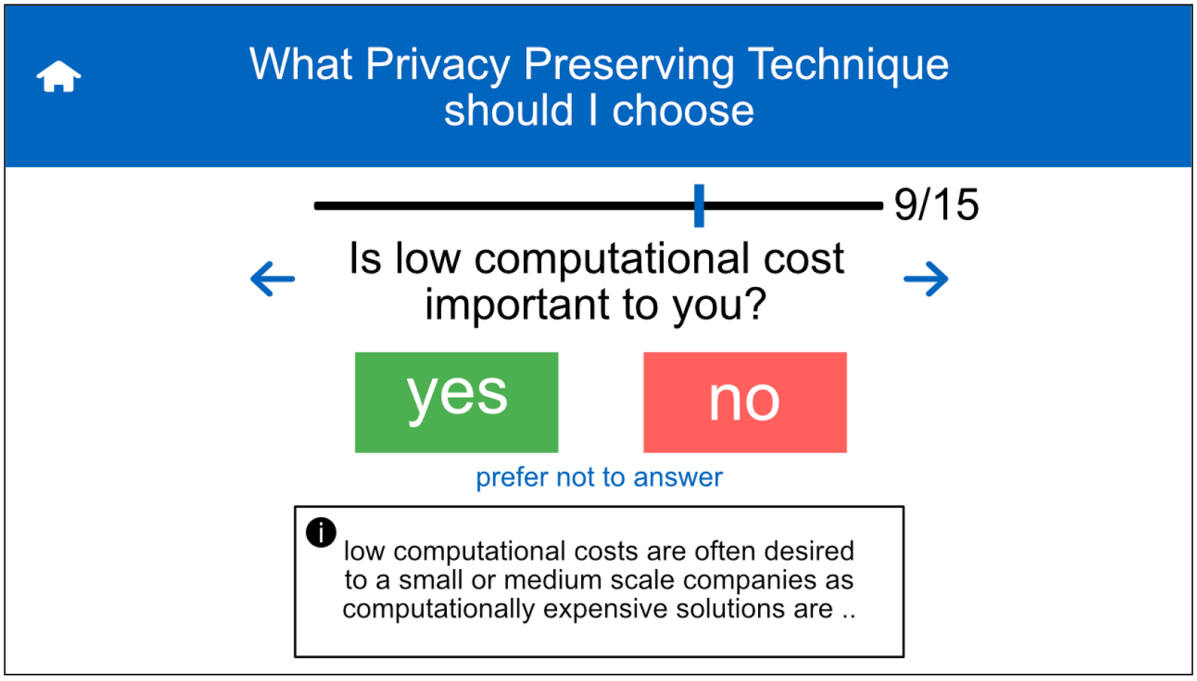

Figure 1.1: Asking users questions

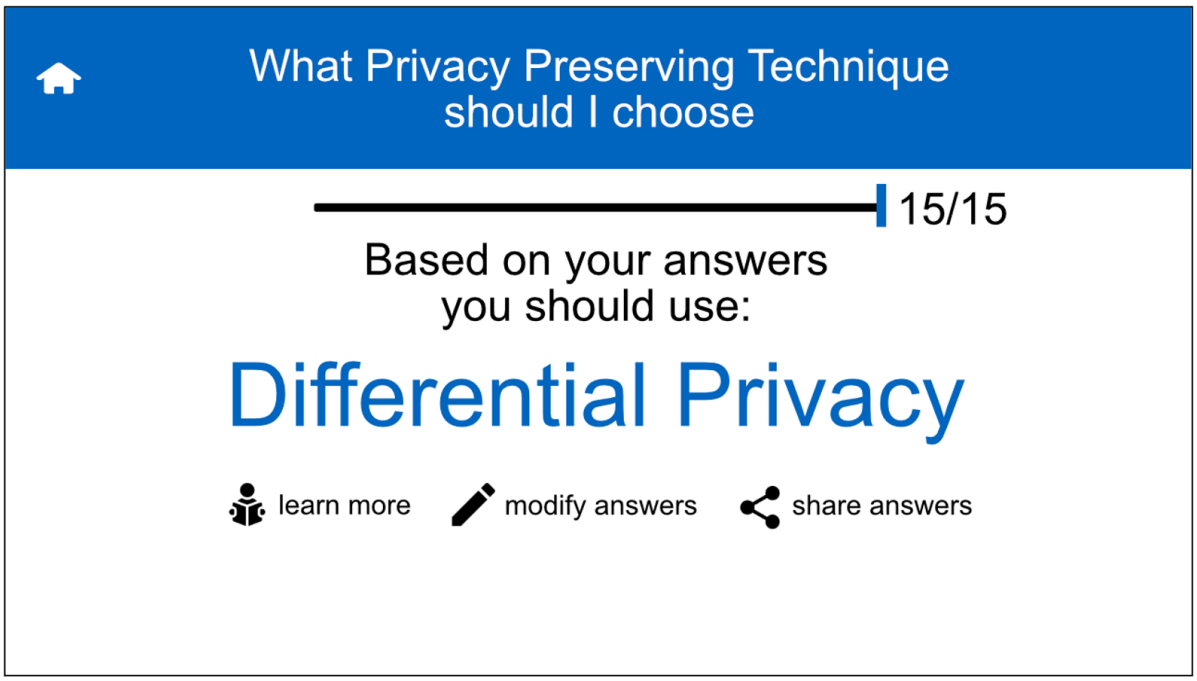

Figure 1.2: Suggesting a privacy enhancing technology at the end of the process

Using the web application and this thesis as a guideline, our goal is to establish a guideline by which people who are a new comer in privacy enhancing technologies can make tailed and informed decisions. The system will be agile enough that if a new technology or criterion is introduced in the future, it can be easily added to the system and kept updated.

2 Background

Any scientific paper needs proper background information to gloss over the terminologies and academic/scientific jargons[1] that are being used in the literature. This is doubly important when the theme of the study is something very niche. Throughout this paper, there are many such keywords that have been used multiple times. While I would love to go over the details as much as possible, to retain brevity and clarity of the literature, the most vital and used terminologies and technologies will be explained below. Depending on the context, these terms can mean different things. Wherever possible, the context will be explained.

2.1 Privacy

To understand how privacy-enhancing technologies work and what they are, we first have to understand what privacy is. Privacy can mean different things depending on the context and its application. However, I have chosen this definition from Professor Tom Gerety that defines it most appropriate for our scenario; according to him, privacy is "the control over or the autonomy of the intimacies of personal identity." [9].

In the 21st century, privacy is defined in many ways. There are legal definitions, there are technological definitions, and also definitions by academics. Legally speaking, privacy is the right a natural person has to be let alone. [10]

In a nutshell, privacy gives us rights and tools to protect our choices and entities in a way that we dictate ourselves. Privacy gives us the freedom to make aspects of our life private; ultimately, it also gives us assurance that proper repercussions shall be taken if our rights to privacy are violated. [11]

2.2 Privacy Enhancing vs. Privacy Preserving

Privacy enhancing, as its name suggests, is a way of enhancing the privacy of a data subject with respect to data collection and processing. The term privacy enhancing

[1] specific words and terms used in a distinct profession

the privacy preserving is used interchangeably. Preserving one's privacy is in a way enhancing the said privacy. [12]

2.3 Privacy-enhancing Technology

Privacy-enhancing technologies are tools that are at our disposal to help us ensure our privacy when our personal data is stored, transferred, and analyzed digitally [13]. There are many ways privacy-enhancing technologies are categorized, however, there are not any standards to categorize them under distinct umbrellas. Judging the academic articles, many have decided to insert them in three major categories: [14]

1. Artificial Intelligence (AI) / Machine Learning (ML) technique: These technologies use artificial intelligence techniques either fully or partially to preserve the privacy of the data. These are relatively recent developments however, they are gaining more popularity over time. [15]

2. Cryptographic techniques: These technologies use mathematics and cryptographic techniques to ensure the privacy of the data and data subject. [16] [17]

3. Data Masking techniques: These techniques randomize/anonymize certain parts of the data to make them indistinguishable from their original form. These techniques are mainly less computationally intense than the other techniques. [18]

While these categories are a way for us to understand and help us choose these technologies better, there are often edge cases that do not satisfy these categories [19]. For example, one privacy-enhancing technology can use any combination of these techniques to achieve its privacy target. However, it's still worth identifying the major categories.

While one of the main goals of this research is to identify the privacy-enhancing technologies out there; even before the research began officially, many of these well-known technologies were identified [8]. While in the pre-research phase, these technologies were identified from various scholarly sources and scientific literature. While these technologies are part of the findings of the study, their internal mechanisms, some characteristics, and uses are worth mentioning. The names of these technologies, where they are used, how they work, and in what scenarios they protect against are mentioned below:

2.3.1 Differential Privacy

Where it is needed: It's used mainly in statistical disclosure control, where it is necessary to limit exposure of individual/personal data. [20]

How it works: Adding substitution of some individual data to make it impossible to deduce the original data. End result will be a little different but the difference is negligible. This results in data that are private and unidentifiable while retaining the statistical significance of the original data. [20]

Protects against: Differencing attack, Linkage attacks, Reconstruction attacks, Re-Identification attacks. [21] This technology protects one's private info, but not the general info about the population/sample data subject is based in.[20]

Implementation: the Coin-toss algorithm, Laplace distribution, Siri (Apple's personal assistance). [22]

2.3.2 Federated Learning

Where it is needed: It is used mainly in decentralized data analysis that requires training model aggregation without combining and storing data centrally to perform the training. This allows multiple edge devices to share data insight without violating the privacy of their subjects. This is used in IoT applications and mobile devices. This has generally low computation requirements. [23]

How it works: Creating analysis of individual datasets and combining the analysis models to create an aggregated model. Then the aggregated (better) model is loaded onto edge devices for better prediction and further training. [23]

Protects against: centralized data breaches, corporate secrets being leaked, and espionage.[24]

Example: Siri, this is why it seems to take forever to mature. One noticeable thing about Android and iOS is how their respective personal assistance is evolving over time. While Siri was introduced along before Google Assistant, the latter soon picked up and now have surpassed Siri on performance and dictation. [22] Contrary to popular belief this is mostly due to how both systems analyze the data and store it. Apple touts its privacy concerns and it shows with Siri. Siri uses Federated learning to train its voice dictation and recognition model on devices. Since this data never leaves the data subject's devices, privacy is protected, however, the aggregated model has time and data size constraints. Compared to a centrally stored data model, it takes longer to reach a similar level of accuracy. [25]

2.3.3 Homomorphic Encryption

Where it is needed: Encrypted data analysis, when the analysis is being outsourced to third-party analysts or cloud providers. Moreover, when the data in transit is vulnerable to breaches.

How it works: performs mathematical operations on encrypted data. There are many different types of homomorphic encryption (HE) depending on the level of computation. For example, simple homomorphic encryption: addition/multiplication, and fully homomorphic encryption: arbitrary circuits with unbounded depth. However, this is computationally expensive, and performing fully HE is often not plausible. [26]

Protects against: Data breaches during analysis. [27]

Implementation: Microsoft SEAL, CryptoNets, PySyft. [28]

2.3.4 Secure Multiparty Computations

Where it is used: when no trusted third party is present or wanted. No tradeoffs are possible between privacy and accuracy.

How it works: one method is called additive sharing. Each party shares x parts of their information with x number of parties. Then each computes the different parts they have. Then the results from each party are combined. It is less computationally expensive than Homomorphic Encryption.[29]

Protects against: malicious third parties, non-cooperation and lack of trust between stakeholders, powerful adversaries like hacker groups or nation states' organizations.

Implementation: FairplayMP [30]

2.3.5 Zero Knowledge Proof

Where it's used: to prove ownership of info without leaking/sharing the/additional info. In other words, you can mathematically prove that you know something without exposing what[31]

How it works: one party (prover) convinces the other party (verifier) that they possess some information without actually revealing said information. It typically uses a function that is applied to the information or knowledge being verified [31]. The person who truly knows the information will be able to predict the outcome of that function without revealing how the function works. And the verifier can use different combinations of inputs and outputs to reassure the prover knows the function.

Protects against: Untrusted computational environment. [32]

Implementation: authentication mechanism, Zcash blockchain. [33] [32]

2.3.6 Obfuscation

Where it is used: where data masking is needed. Software IP protection. [34]

How it works: making the original data difficult to read or infer. Adding misleading data to the mix. Replacing some data with distracting data. This technique is mainly used in software engineering where a binary of software is obfuscated to make it very hard for an attacker to read and identify how the individual code of the binary works, essentially blocking an attacker from exposing the inner workings of software or making it hard to tamper with the data. [34]

Uses: software publishing, deployment, and data sharing. [35]

Protects against: reverse engineering, trade secret leaks. Warez and cracks.

Implementation: Obfuscapk. [35]

2.3.7 Pseudonymization

Where it is used: when the data should be made unlinkable. While this provides some level of anonymization, this is not completely anonymized. [36]

How it works: replacing data identifiers with pseudonyms. The goal of this technique is to make the data anonymized, but in special cases, it can be reversed back to its original form unlike traditional anonymization techniques. Basically, some parts of the data are removed and anonymized and these parts are stored separately. The former is essentially anonymized, while if needed, the latter can be added to it and the original form can be restored. In other words, it's reversible with additional info. The additional info is kept separately. [36]

Uses: GDPR. This is one of the transitional techniques suggested by the EU GDPR committee. [36]

Protects against: statistical disclosure. [37]

2.3.8 Synthetic Data Generation

Where it is used: When real data is not accessible, sufficient, or needed. As the data are fake, the data subject's privacy is not violated.

How it works: creating fake data sets that closely resemble real/expected data. Any type of calculation performed on the data shouldn't vary too dramatically Using the Deepfake technique to test face recognition algorithms. Deepfake generates images and videos that do not resemble any natural person. However, to test whether an image feature recognition or video object recognition algorithm works, you do not necessarily need data or a real natural person. In these use cases, it's highly recommended to use synthetic data. However, this generation takes a lot of time and effort. Therefore, the initial time of entry can be lengthy. [38]

Protects against: Real data breaches.
Implementation: Deepfake generator, SynSys. [38] [39]

2.4 Known Use Cases

There are many instances where privacy-enhancing technologies are currently being used in the industry. Some of the most important ones are described below:

2.4.1 Cryptocurrency

Zcash, a popular cryptocurrency, uses Zero-knowledge proof in its blockchain. The so-called variant of zero-knowledge proof, dubbed as zk-SNARKS [33] is used as the method to retain anonymity in the blockchain network, yet at the same time ensuring the reliability of the information.

According to a blog post [40] released by the ZCash authority, Zero knowledge proof allows ZCash blockchain's one party (the prover) to prove to the other party (the verifier) that a statement is true without revealing any information except the validity of the statement itself. Anonymity is one of the main attractions of cryptocurrency and by using zero knowledge proof, ZCash provides an extra layer of anonymity over the concept of blockchain.

2.4.2 Smart assistance

Privacy-enhancing technology like federated learning is currently being used in digital personal assistants like Siri. Since iOS 13, Apple has been using Federated learning on both Siri and iPhone's keyboards. [41] Apple has its own custom chip that analyzes and stores the analysis in the device itself. Thus the personal data never leaves the device itself. However, to gain important telemetry and personalization information, Apple uses federated learning. Apple has a smart suggestion and voice recognition ML model in the cloud. The on-device chip analyzes this data and removes any personal elements from the data. Then they use an on-device ML model to train on that data. Only the newly trained model leaves the user device and it get aggregated to the primary model hosted in the cloud. Finally, the new centrally aggregated ML model is sent to all users over the air. This way, the users retain their privacy while ensuring an improved and more personal experience with their devices.

Google also uses the same Federated Learning technique in their Android personal

assistant. Like Apple, they use a combination of Federated Learning and Differential Privacy to ensure user privacy while keeping the user satisfied with their improved suggestion engine. [42]

2.4.3 Digital Rights Management

Code Obfuscation is used by software companies and professionals in the industry. This is in effect to protect their intellectual property and distribute their software binaries[2] without sacrificing the right as the original author. [43] There are many known vulnerabilities in software development and distribution tools that let bad actors and hackers reverse engineer the original source code or tamper the code in a way that the original authors have not intended.

Code Obfuscation works by replacing certain parts of the original source code in a way that would become very difficult to reverse engineer. [44] There are many tools and programs that do it automatically. For example, there is a tool called Terser that automatically makes JavaScript code very hard to read upon the building phase of a web application. [45] Developers also have the option to minify[3] the code to make it even more unreadable.

[2] A file format that is used to distribute and install an application. i.e. exe files for Windows OS
[3] Making the code shorter by removing white spaces, line breaks, indentations.

3 Research Design

While research can be a convoluted and time-consuming process, a proper research method can make this process a lot easier and faster [ ]. For example, there are many established research methodologies out there. Some are broad and can be applied to a large spectrum of studies, while others are tailored towards specific fields of research. Even in computer science-related research, there can be a plethora of methods that can be selected for a particular topic.

In other words, the first task was to design a suitable research model from a set of methods.

To identify the best possible methodology for the research, we have to ask ourselves a few questions:

- What type of evidence I have to collect during my research period,
- How this evidence will be collected, and
- On what basis, this information will be used towards the research.

Since I will be largely focusing on existing research to find existing privacy-enhancing technologies and map these technologies with their respective scenarios, it was decided that systematic literature review techniques shall be used for this research.

A systematic literature review can be very broad in its nature. It can be applied to a multitude of fields, namely, medical science, statistics, and also, computer science. While many variations are available, to the core, they all work in the same way.

3.1 Protocol

The protocol I came up with resembles any standard literature review protocol, however many elements have been tailored for the need of a systematic literature review for software engineers. The book by Barbara Kitchenham heavily influenced this process. [46]

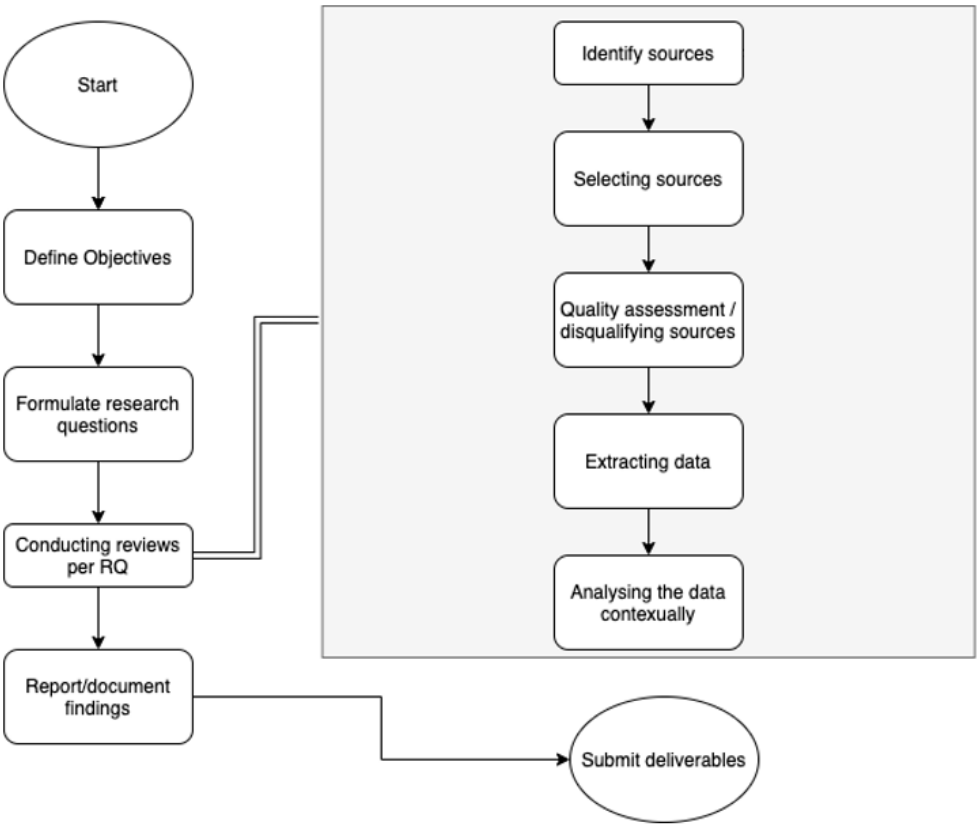

Figure 3.1: Research protocol

–

The research protocol varies for the initial phase of the research and the latter part of the research. Initially, the research protocol is established to identify the literature and deduce information from them regarding the identification of privacy-enhancing technologies. This is done in 4 stages:

3.1.1 Defining objectives

Since all of our research questions will need answers based on existing literature, its sourcing and filtering play a great role in terms of the quality of the research. During the pre-research phase of the thesis, it is very important to set up the research protocol just right to avoid any unwanted consequences. As a result, I set up a moderately modified version of the protocol with clear paths and goals in mind. The initial stage is to clearly define the objective.

The main objective at this stage was to create defined goals and paths for the thesis. As mentioned at the beginning of the article, the primary goal is to find a decision

support system for non-expert professionals. At this stage, we break down that goal into its nitty-gritty details.

First, the research questions needed to be formulated, based on the research questions, source material selection, and quality filtering will take place. Beyond that, useful data will be collected from the source material and analyzed. After the analysis of the first two research questions, we can answer the final question with the help of inference and more source material

When finalizing the findings, it is vital to deliver on the primary goal of the paper. It's been designed in a way that, upon the findings of the first two iterations, the third question will be answered by mapping the two prior results.

3.1.2 Formulating research questions

Since the primary goal of this study is to help nonexperts by mapping existing information from literary sources, the research questions need to answer each step of the major processes.

First, we need to know about the current status of privacy-enhancing technology. To suggest to people what technology they should choose, it goes without saying that we need to identify the current technologies first. But how do you define which technology should make the cut? There are many technologies there that have very shadowy lines between them, they are subclasses of the same technology and often are distinguished by a small niche application. [19] More information on how a distinct technology was selected will be discussed in the findings area.

To help with this decision process, it is important to know about the characteristics of these technologies like their applications, pros, and cons.

Second, we need to identify the scenarios or criteria upon which a privacy-enhancing technology is selected. Once we are done with answering the first research question and we have a list of technologies from it, by researching about these technologies and their application and why they are selected in a scenario where they are, these scenarios can be compiled and be presented as the result in this stage.

And finally, when we have information from both questions, and can map the data to get our desired objectives. However, there are many ways to present that information. Since we are targeting non-expert professionals, it is important to take that into consideration and present the information or the recommender/decision support system in a way that inexperienced users can find easy to use and get the recommendation from.

After all these stages, we will have reached our desired goal of implementing a working solution for the decision support system.

3.1.3 Conducting systematic literature review per research question

For research question 1, this was pretty straightforward. This stage of the process: conducting reviews per research question part consists of 5 stages:

Identifying initial sources

To identify the sources - in this case - scientific literature, we have to start broadly. Initially, 3 very general search terms were picked to cover as many privacy-enhancing technologies as possible. The search terms were:

- Privacy
- Privacy preserving
- Privacy enhancing.

As our source aggregator and search engine, we have used Google Scholar and CrossRef to identify these sources. One of the tools that were used to search the literature is called "Publish and Perish." It's a desktop application that can be used to search for scientific literature from the sources mentioned above. The source materials are fetched with respect to the most cited first. The publications are varying, including IEEE, ACM, and other notable science publishers.

Figure 3.2: Publish or Perish application

Publish or Perish[1] uses its own API to search on these source aggregators in bulk, this reduces the complexity and time needed to compile a large number of sources. The results can be viewed in a tabular format. It fetches literature metrics like title, authors, publication, date, and the abstract. We will mostly deal with the abstract, title, and publication date. These individual data can be exported in a JSON or CSV format for importing in a data analysis tool.

However, Publish and Perish does not retain this information in a way that can be used to manage and locally filter. For this purpose, another tool called Zotero is used. This is used to manage all the literature from a central location, besides management, the data can be tabulated, filtered, and exported at our will. This feature will become beneficial at a later stage of this study.

[1]Publish or Perish: https://harzing.com/resources/publish-or-perish

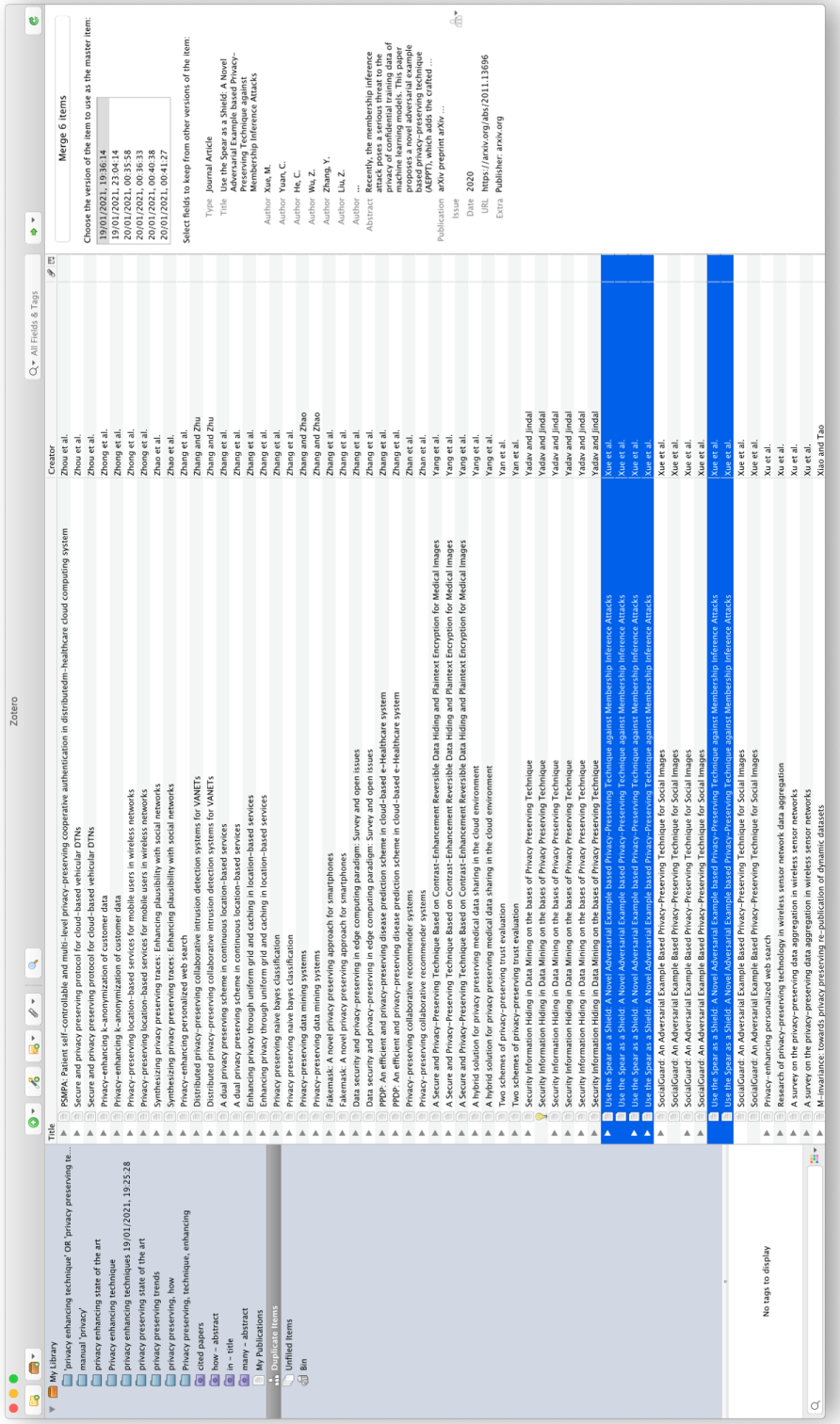

Figure 3.3: Zotero application

Zotero[2] is also used to remove any duplicates from their results, while Publish or Perish does not give us any duplicated results, when we merged the results from the three search results, there were many duplicates that needed to be removed. This is where Zotero came in very handy.

There were some situations where Google would identify abnormal numbers of requests from Publish or Perish. This was a limiting factor in finding an even larger sample size. However, after merging, we are still left with 807 papers. Which are deemed very significant to get a broad picture of the technologies.

Then comes the next major challenge. Finding privacy-enhancing related keywords to conduct the secondary search. As it's implied reading 807 papers manually will not only be time-consuming, given the constraint of a single person, it would be inefficient for the research.

Extracting relevant information

This is where SketchEngine[3] comes into play. SketchEngine is an online tool that is used to analyze textual data using artificial intelligence. SketchEngine creates what it calls a 'corpus' when textual data is fed into it. Using its artificial intelligence model, it will then generate a corpus/list of texts that it thinks are the most important in the collection of the input texts.

[2]Zotero: https://www.zotero.org/
[3]SketchEngine: https://www.sketchengine.eu

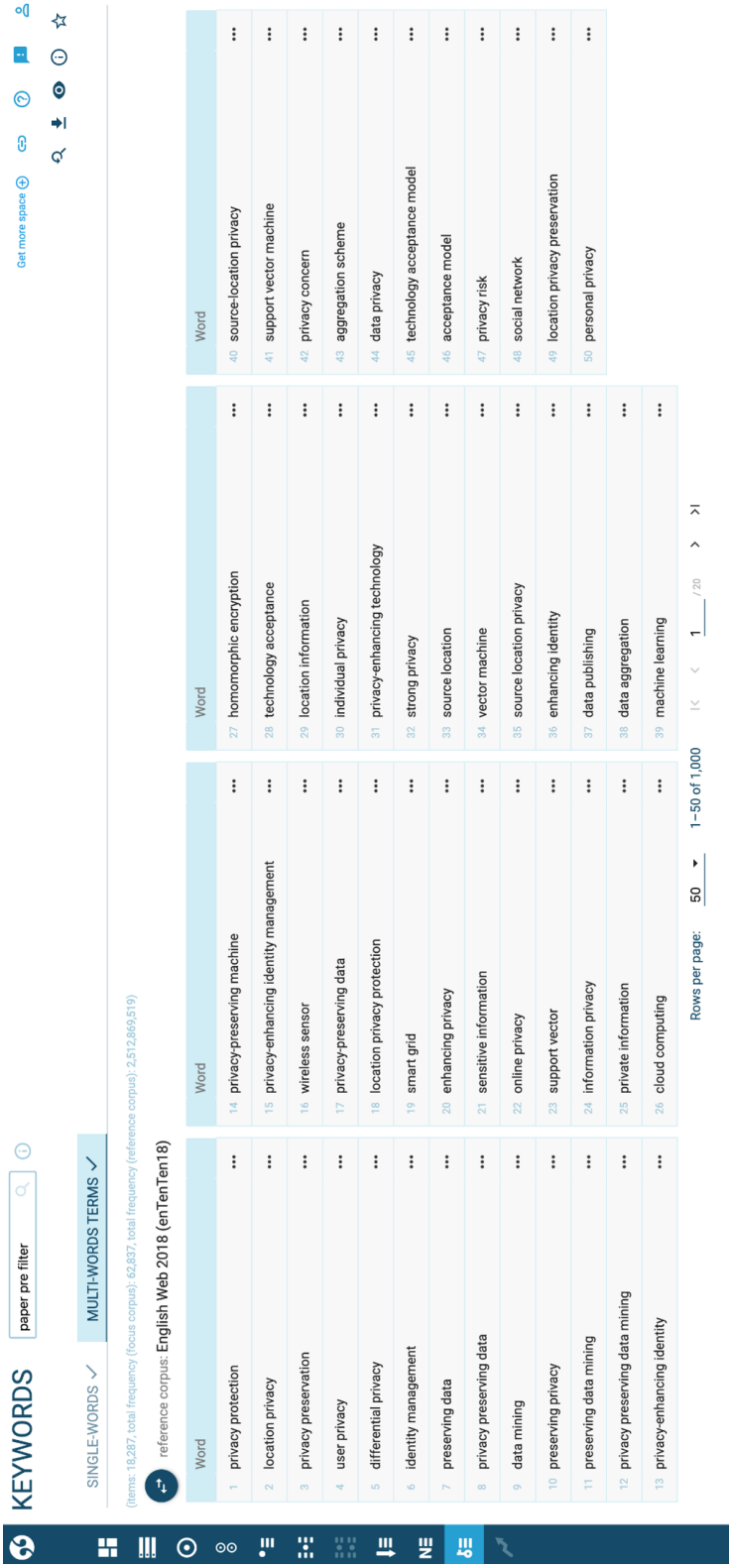

Figure 3.4: SketchEngine corpus

The titles, and the abstracts from these 807 papers are exported as a collection of texts using Microsoft Excel's CSV export feature. Using SketchEngine we create a text corpus that includes single and multi-word keywords found insignificance in the 807 papers we inserted in it. Now the task is to find keywords that are related to privacy-enhancing technologies.

Instead of going through hundreds of papers, we now have to go through only a few keywords. After meticulously going through this corpus and oftentimes looking up individual terms, we have selected the top 30 initial search queries from it. Out of which only 12 were selected for further search. The exclusion criteria are the following:

- Terms that are similar to one another
- Terms that are irrelevant
- Duplicate entries

Now we have our 12 search queries for the next phase of the search. The keywords were:

- privacy preservation
- preserving data mining
- preserving data
- privacy protection
- preserving privacy
- location privacy
- enhancing privacy
- privacy-enhancing identity
- user privacy
- privacy enhancing technology
- enhancing technology
- privacy-preserving machine

Conducting secondary search

Secondary searching should now commence using the keywords found in the prior stage. I again use Publish or Perish to aggregate source materials from these keywords. This time however a different exclusion criterion was chosen.

For this case, I have selected papers from the last 15 years (from 2006). The reason this cutoff criterion was chosen was that there would be too much literature to analyze and identify the privacy-enhancing technologies from. However, 2006 is a good boundary condition for two reasons:

- Privacy-enhancing technologies are a relatively new research field.
- If a technology is important, it should be mentioned in the last 15 years of research.

This condition dramatically reduces the number of materials I have to deal with while retaining quality with negligible, if any, information loss.

Finally, the results are merged and the duplicates are removed to get a unique set of results.

Classification

After merging these results in a single collection using Zotero, I now have 3582 source materials to find privacy-enhancing technologies from. As the number suggests, this is a daunting task. Initially, the papers were read manually to find relevant papers and whitelist them in a separate collection. However, it was later found out that most of the pacers were relevant and it would make more sense to remove irrelevant papers instead.

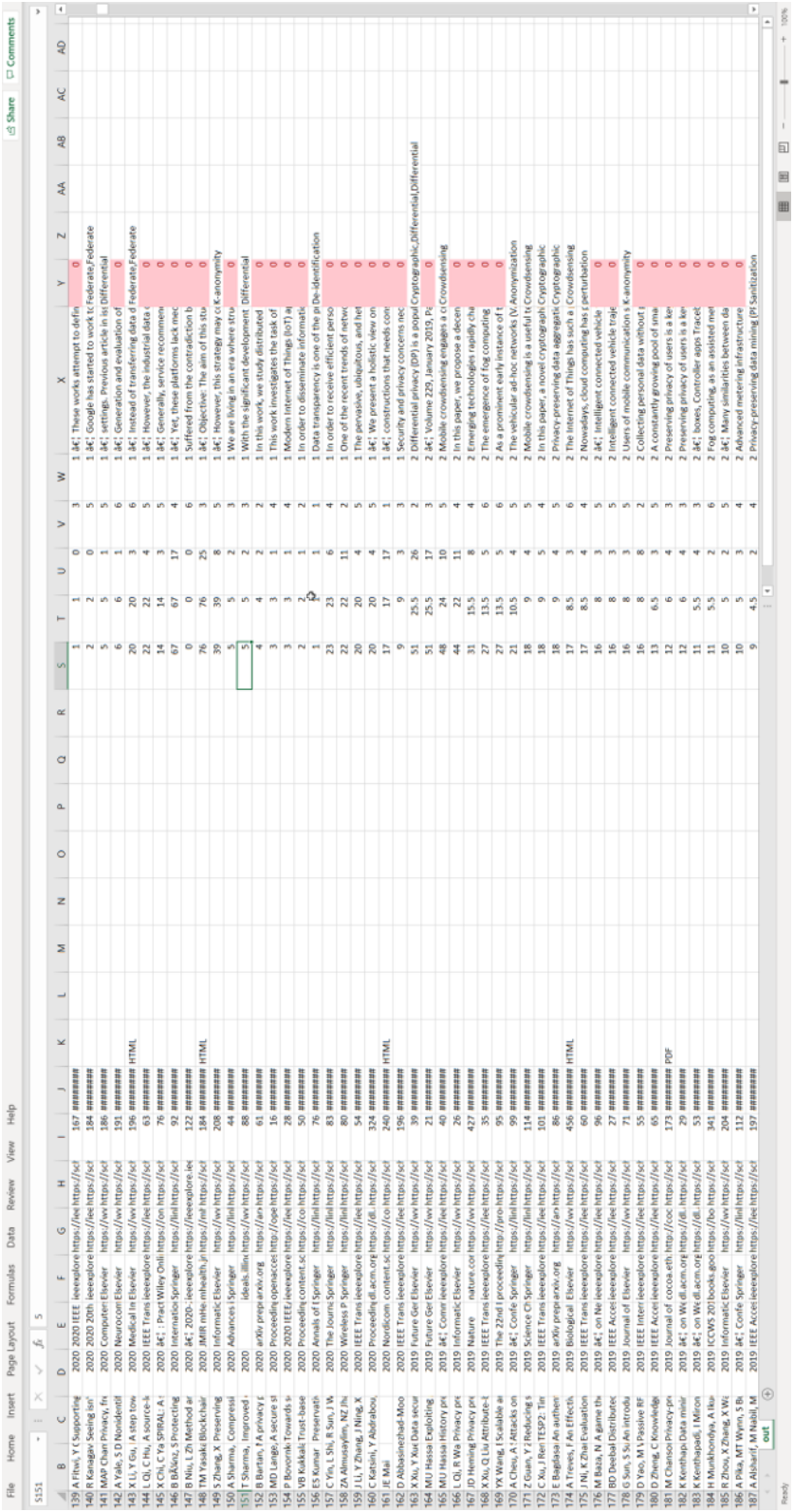

Figure 3.5: The CSV file in Microsoft Excel (assigned PET tags in the last column)

A few days went by, manual filtering is taking its toll on its human seeker. I only managed to filter out a few hundred papers. After a week, the total number of papers still stands at 3176. I have realized, this can't work. I have to find another way to do this more efficiently.

Enter automatic filtering.

Automatic filtering would save us a noticeable amount of time to extract relevant information in a dynamic way. I have used a combination of JavaScript and a CSV file to calculate the final set of papers based on the existing sources.

Program setup I have used NodeJS for running the JavaScript code on the machine. NodeJS provides a lot of tooling and debugging support necessary to run this application. As I already know Javascript, it seemed to be a reasonable decision to use NodeJS.

I have also used JavaScript's ES5 version to make it compatible to run on most machines without needing to transpile it back to an earlier version. Since ES5 is supported in most browser engines as well as in Node's V8 engine, this code is backward compatible.

As to make the program work, the following plugins/packages were used:

- Express: is a library to help run a NodeJs app. Provides a lot of methods to make a NodeJs app simpler and easier to build. Although avoidable. this was used mainly for familiarity and ease.
- FS: a file system module built into NodeJS. This provides methods to read and write from the machine's file system. In our case, this was used to read the CSV file and write to it. In the case of a big file, the file was read asynchronously.
- CSV-Parse: this package was used to parse information from a CSV file into the app. Its parse() method converts CSV data to a native Javascript object.
- Fast-CSV: this package was used to finally write data back to a CSV file for further analysis. CSV file was used to maintain wider compatibility and bring support into the app itself.

The application is currently hosted on GitHub[4]. You can run the application yourself by following the instructions on the repository. All the sources and materials used in this research can be found in the repository.

[4]GitHub repository: https://github.com/FaisalBinAhmed/PET-Guideline

The application is basically structured into four parts. I will go into detail about how each part works. Here's a brief overview of how the program looks.

Figure 3.6: The program running in VS Code

As you can see from the figure, the app uses a list of keywords related to privacy-enhancing technologies, it then scans through the CSV file to match the entries with these keywords, and finally, output is shown on the bottom.

How the program works can be described in the following four steps:

1. Extracting Keywords

As mentioned earlier, we get a set of keywords by feeding textual data into SketchEngine. This same method is applied here. Instead of manually reading thousands of papers - to identify papers with privacy-enhancing technologies in them and then filtering out irrelevant ones - SketchEngine is used to derive keywords found in those papers.

Now, normally someone would look for specific keywords, this would take a lot of time. But with SketchEngine, you get the 2000 most occurring single and multi-word keywords at once. Now the task becomes a lot simpler.

It is important to derive relevant keywords from these two words. That's described in the next step.

2. Creating PET pool

A pool of relevant keywords needs to be created for the automatic classification to function. To do this, each of the 2000 keywords was manually analyzed. Some of the keywords are well known and can be included in the pool right away, while others need to be looked up manually. Some of the keywords were also outright irrelevant. In any case, these were checked against sources and validated. However, this method is manual work and prone to human errors.

3. Assignment of Keywords

The PET pool is the dynamic single source of truth for the keywords in the automated process. The keywords are then matched with the paper programmatically.

The program works as follows:

- First it creates a bag of words containing the titles and the abstracts of the paper.
- Then it checks for unnecessary punctuation and blank spaces in those words and removes them for better matching.
- Now for all the words in the titles and the articles, it checks for a word from the PET pool.
- If it finds a match, the PET words are then assigned to the specific paper.

- These steps are repeated for each word in the PET pool.
- It is also made sure the PET tags are distinct and not repeated.

The pseudo-code for the above steps looks like this:

```
IMPORT csvContents from FileStorage
IMPORT petWords[ ]

PARSE papers[ ] from csvContents
INT counter = 0

FOR each paper in papers[ ]

       VAR matchedWords[ ]

       SANITIZE paper[title] and paper[abstract]
       COMBINE paper[title] and paper[abstract] as paperWords[ ]

       FOR each pet in petWords[ ]

              FOR each word in paperWords[ ]

                     IF word matches pet
AND IF it is not there already
                            ASSIGN the word to the paper
```

Listing 3.1: Pseudo code for the automated classification program

4. Manually adapting keywords pool

While we got many keywords from SketchEngine, a manual inspection may still give us a better set of PET keywords. When the results and the assigned tag were printed from the program, each of the tagged papers was scanned manually. It was thought that paper that matches keywords might possess other useful keywords that went under the radar of SketcEngine, and it turned out to be true.

There were a few keywords that were found manually that were not present in the corpus SketchEngine provided us. Adding those keywords in the PET pool resulted in a larger number of papers being tagged and discovered. This manual process was deemed important in the end.

Here is a brief overview of the numbers:

Table 3.1: Number of sources after manual adaptation

	Before Manual Adaptation	**After manual Adaptation**
keywords	48	53
matched papers	420	623

3.1.4 A Novel way of doing Literature Review

This process is inspired by Barbara Kitchenham's book on Systematic Literature Review for software engineers [46]. The process was tweaked to tailor-fit this thesis' specific needs. As written in this text, many parts of this process are entirely novel approaches. This I think is a byproduct of this research. The tools mentioned here, the way I have conducted the research using SketchEngine, and how I automated the process of filtering, can be used by future researchers in the field. These techniques are an entirely new way of doing systematic literature reviews on a large set of sources - especially where manual inspection is often not plausible.

The time saved here by doing automation, is valuable and can be applied to other fields of research as well. Here is a brief overview of the time saved:

Table 3.2: Time saving of the automated classification process

	Before automation (hours)	**With automation (hours)**
For the first 406 papers	9 (approx)	n/a
for the next 3176 papers	70 (approx)	8 hours

As you can see from the table, the automated process saves approximately 71 hours of man-hour, which is a staggering 90 percent reduction of the time. Not only we are saving a lot of time, but we are also ruling out any potential human errors and bias in the process.

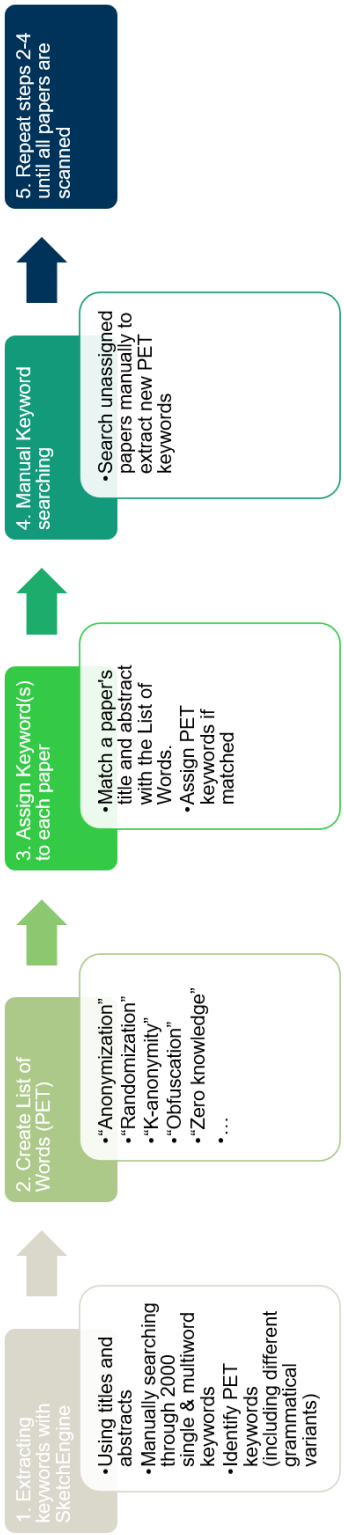

Figure 3.7: Automating the source finding process

3.1.5 Final Findings

After all these processes, we are now left with 623 papers for our analysis. The findings and the analysis will be discussed in the next chapter.

This is how the entire process looks from start to finish. The automated part was executed at the classification stage.

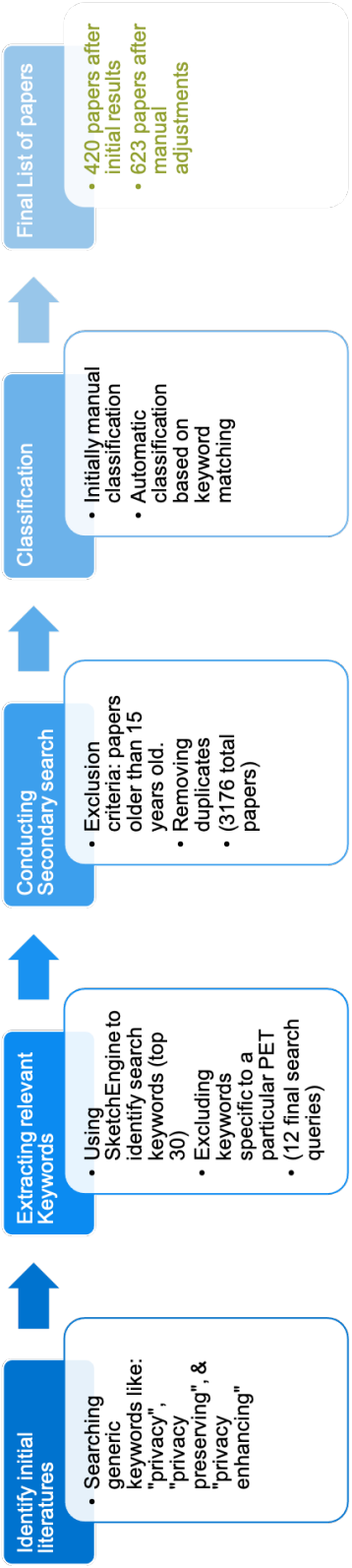

Figure 3.8: Source classification and filtering process

Of course, maintaining and keeping track of this large number of literary sources is not manually possible. Here another tool called 'Scribbr' [5] comes to help. It was very important to keep the sources used in finding the data and sources used during the writing process of the thesis separate. Scribbr helps here by keeping a list of literary articles, white papers, grey papers, internet blog posts, etc. It also helps create a list of all sources as a BibTeX entry. As BibTeX is used to generate the bibliography for this report, Scribbr made it very easy to export those sources.

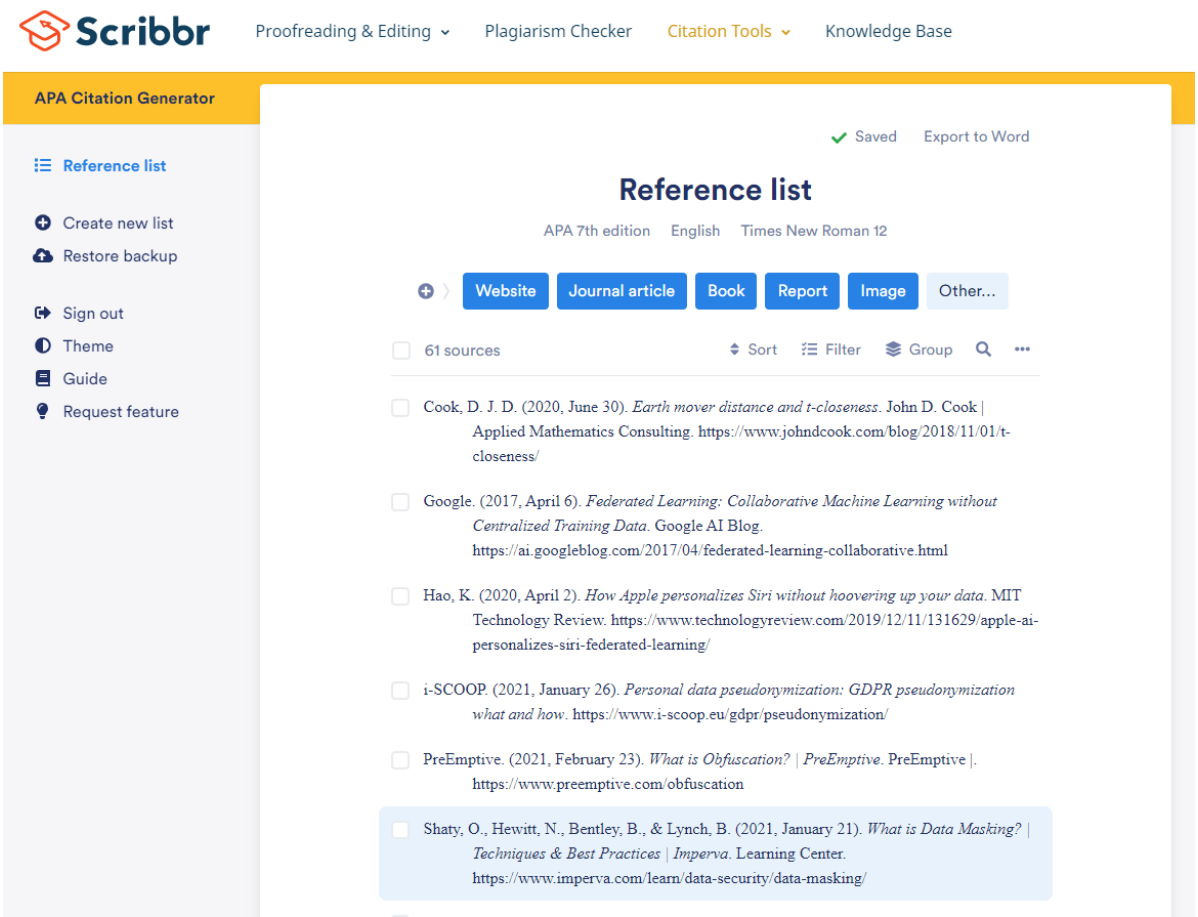

Figure 3.9: Scribbr web application for literature and bibliography management

[5] Scribbr: https://scribbr.com

4 Findings

4.1 Status Quo

The first research goal of this thesis was to identify the current status of the different privacy-enhancing technologies. Without any further ado, here is a chart of the technologies that were identified from the initial research:

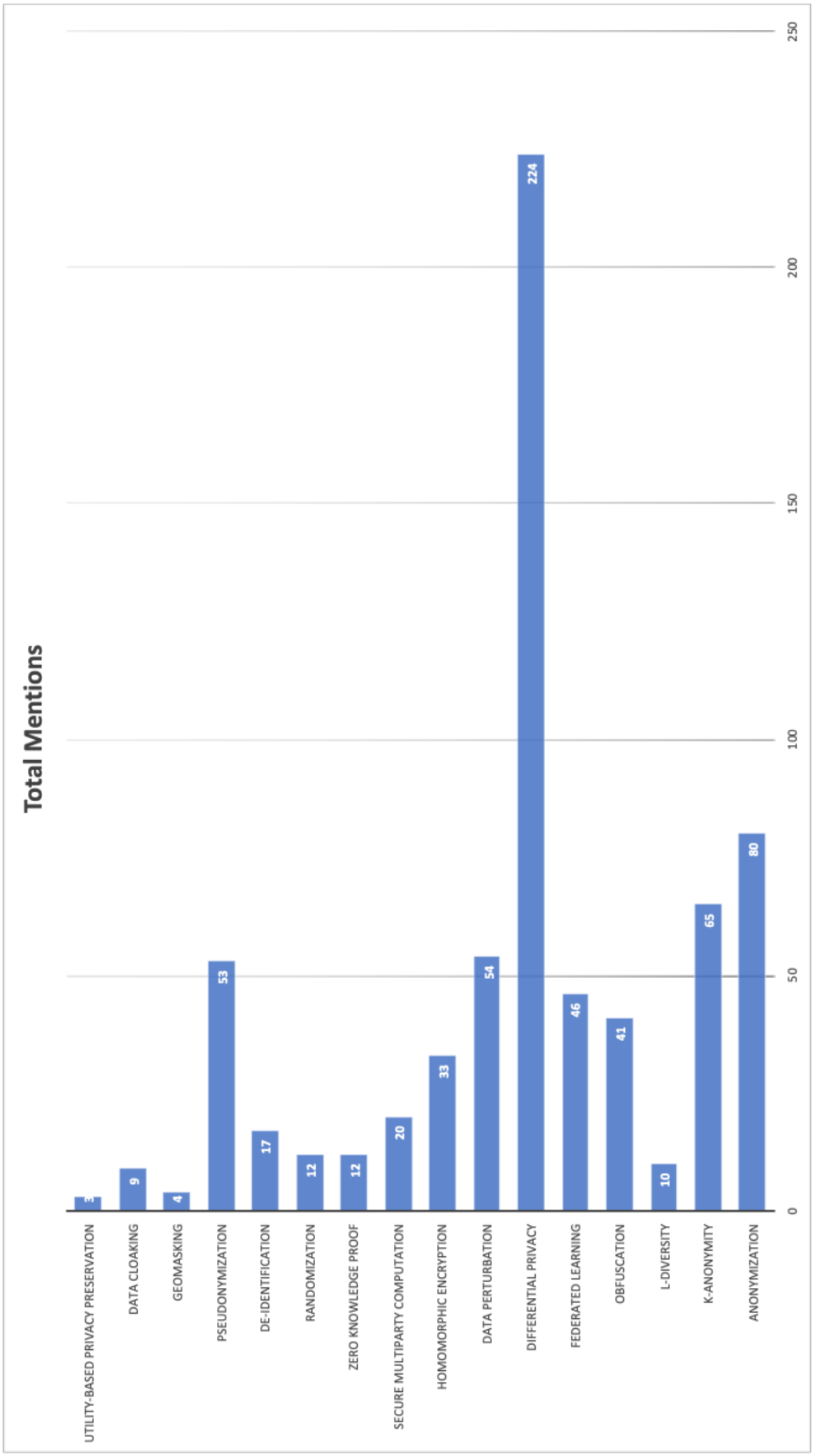

Figure 4.1: Identified PETs after the initial phase

As you can see from the chart, all technologies except for Differential privacy have been mentioned in this paper somewhat uniformly. The large spike in the number of papers mentioning differential privacy is not due to it being the most popular (although that could be a case too) but the name including the term "privacy" in it. One of our initial search terms included these term, henceforth, the discrepancy in the number of papers.

As we mentioned previously, only the papers from the last 15 years were considered for this thesis. As can be seen from the figure below, there are no clear trends in the year by year basis of privacy enhancing technologies. However, this could also be due to our searching process. Publish or Perish has limited API request quota set by Google, thus in tends to fetch the most cited papers by default. Since newer papers tend to be less cited than older ones, it can be inferred the the trend is going slightly upward. However, this claim needs more data to be verified.

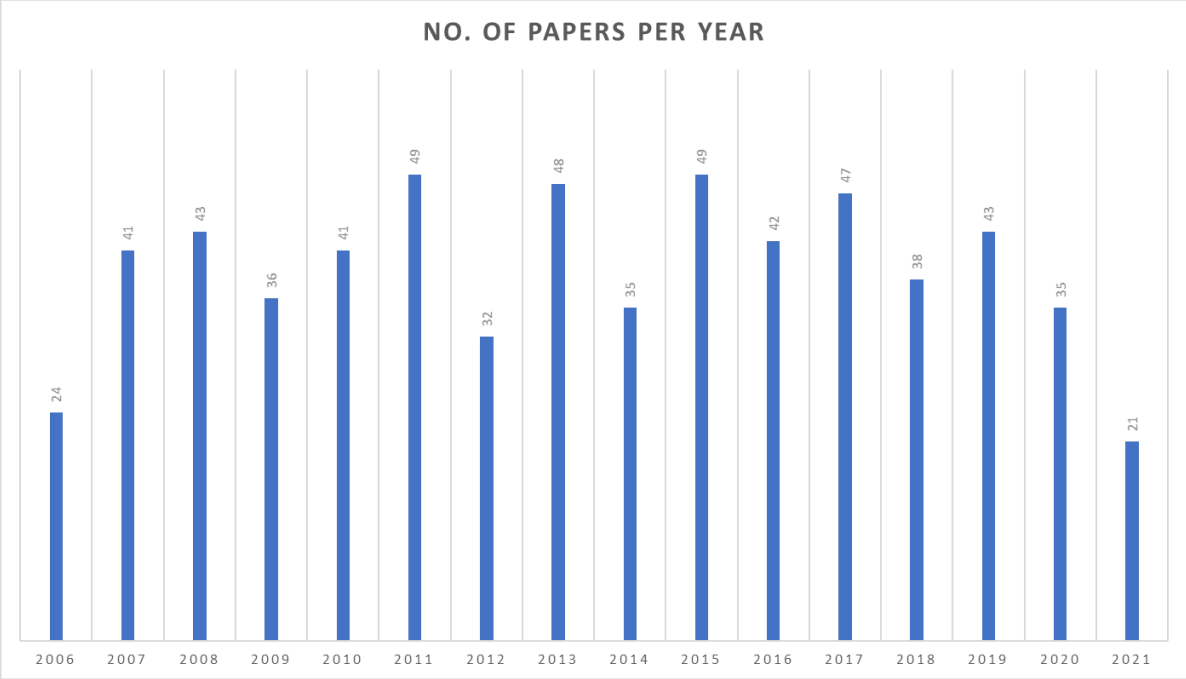

Figure 4.2: Year by year trend in privacy enhancing technology papers

4.2 Identified Privacy Enhancing Technologies

The generated keywords pool was one of the easiest ways to identify the privacy enhancing technologies. Some of the keywords mention the same technology while others needed a little digging to be sure. For example, keywords in the pool like:

k-anon, k-anonymity reference the same privacy enhancing technology. When discovering the technology in this way, all grammatical and other variants of the same keywords were considered.

After the initial sources were identified, the selected papers were read as extensively as possible. While many papers were really easy to deal with. E.g. papers that mention the name of the technology directly in the paper, others were a little tricky to find. Sometimes many technologies were identified in papers where these were not the subject of the study. While secondary situations like this rarely revealed a new technology, this helped the research in later stages to bolster certain hypotheses.

The final list of privacy enhancing technologies is a bit different than the initial version. While the number remained the same, the technology De-identification was removed as it is a subset of anonymization. Some of the very specific and niche technologies like Utility based privacy preservation were removed and others that were identified later from the sources were added in the list. While geomasking was merged with data masking and so on.

The final list of technologies and their characteristics were mentioned below. Many of these technologies have already been explained in the previous chapter, in that case, some additional information will be mentioned below:

4.2.1 Differential Privacy

Differential privacy is a technique of exposing some insights about a specific dataset without really exposing the real value of the dataset. This in turn protects the privacy of the data subjects, and it becomes mathematically infeasible to determine any insights about the original data subjects from the published data.

Differential privacy is mainly used in statistical disclosure. While this loses a little bit of accuracy from the original data, for all intents and purposes, the loss of statistical significance is negligible. [47]

Differential privacy is one of the most popular privacy enhancing technologies in the world today and many companies are thinking about it or already using it in their applications. [48]

How it works:

Differential privacy sometimes uses an intermediary mechanism to control any statistical disclosure of individual data. Let's say there are 50 data subjects that want to know a statistical calculation of their data without actually revealing the individual data. In case any parties know the majority of the data or if the data has been published elsewhere, the data of other unknown subjects can be deduced from it. To prevent this scenario, differential privacy dictates that the data of any random subject be replaced with random data. Hence, adding a "noise" to the overall data. The noise is not statistically significant enough to introduce any discrepancy in the

overall calculation of the data (i.e. the average), but it protects the privacy of those individuals by hiding their real values.

This is particularly useful when some of the data is already known to the outside world. Without adding noise to the other data, the publication of the entire thing makes those other data vulnerable to attacks like linkage.

4.2.2 Federated Learning

Federated learning is mainly used in Artificial Intelligence and Machine Learning scenarios. However, it can be applied to a broad range of use cases. In a nutshell, Federated learning provides a way to share analysis across multiple vendors or locations without actually sharing the data the analysis was based on. [47]

How it works:

Imagine you have 5 business partners that use one of your business software. In order to improve the software experience, you want to know how it works in those systems: its performance and other metrics. However, this information can be private and companies will not agree to share any information that includes its private and sensitive business information. One solution is to build a mechanism that would gain analytical information from these individual companies. However, you might need to duplicate this analytical engine for the specific use cases and in the end, it could become very complicated to manage and aggregate this information.

Federated Learning comes to the rescue in this situation. Instead of building multiple analytical engines that do the same thing, Federated learning provides a mechanism to host the aggregated analytics engine at a central location while also getting information from those individual companies. The individual federated learning model in each company will analyze the data and send only the sanitized analysis to the central server. The server would then aggregate these analyses and create a central model. This new and updated central model will then be pushed to the client side for better analysis. The analysis is being improved while maintaining the privacy and security of these clients.

4.2.3 Zero Knowledge Proof

Zero knowledge proof lets a person (called prover in this case) prove another person (called a verifier) that he knows about something without giving away the knowledge. This is very useful when you can't give away any sensitive information but at the same time have to prove the knowledge of said information. For this reason, zero knowledge proof is very useful to implement authentication mechanisms based on its principles. [31]

How it works:

Imagine you have a skill of identifying the sex of kittens just by looking at them. But how do you prove this skill to a person who can't identify the sex of the kittens? This is where zero knowledge proof comes into play.

To prove the verifier that you indeed know about identifying the sexes of kittens, you don't have to tell the verifier which kitten is what. In other words, you don't need to reveal the information to prove your knowledge. First, two identical kittens but with different sexes will be handed to the verifier and he would assume and call them kittens A and B. By looking at them, you know which kitten is male and which is female but how do you prove it?

The verifier would then swap the kitten unbeknownst to you whether he swapped it or not. If he swaps, kitten A will take the place of kitten B and vice versa. However, you have no way of knowing if he swapped or not, unless of course you identify them by their sex and tell the verifier whether he swapped them or not. This swapping takes place multiple times. With each correct answer, you eliminate the chance of randomly guessing and after a certain number of successful guesses, the verifier can be assured that you indeed know to identify the sex if kittens.

4.2.4 Homomorphic Encryption

Homomorphic encryption is a cryptographic technique that lets an entity perform computational operations on an encrypted data without needing to decrypt the data with plain text first. While this might sound like an antithesis of the idea of cryptography, it lets any participating entity perform operations without sharing of the original encryption keys. [49]

How it works:

Imagine you want to implement a voting system where voters can identify their vote has been counted but without telling anyone publicly who they voted for. If someone votes for Person A, they receive a unique encrypted code that resembles their vote in the system. These codes are publicly available and anyone can verify whether their vote was valid. However, if the vote itself was encrypted, how do you verify who got the most votes? Homomorphic encryption comes into play here. Instead of decrypting everyone's votes and then counting the tally, all votes for person A will be counted by multiplying the cipher text of the individual votes. Then the big cipher text will be publicly decrypted, revealing and validating the number of votes for Person A. This system ensures the individual votes of the voters are never decrypted. Hence, ensuring cryptographic integrity while adding extra computations on top.

4.2.5 Secure Multiparty Computation

Secure multiparty computation is useful when multiple parties want to do computation on a shared data together but they don't trust each other. In normal cases an intermediary can solve this issue by being trusted by everyone. However there's no guarantee that the intermediary is not corrupted and is colluding with one or multiple parties. This is where secure multiparty computation comes into play. Instead of relying on an intermediary, an algorithm is used to play the same role. [47]

How it works:

Imagine three big technology companies want to calculate the average wage of their employees salaries without revealing the respective average salary of each company. Normally, an intermediary can calculate the average and distribute the information. However, secure multiparty computation does away with intermediaries and uses a smart algorithm.

Instead of giving away the salary information, each company randomly generates two numbers between 0 and 3. Each company now multiplies their own average salary with these numbers. Each company now provides one of these two numbers to another company and the other number to the last company. Now one company has a different version of the same information about the other companies.

This way the other two companies can't share the information about the first company and derive its original value, as only the first company knows the numbers it was multiplied with.

Now the first company would add the two numbers it was provided by the other companies and subtract the numbers it provided to the other companies.

After this process, all the sums will be added together to derive the actual sum of the three companies. Finding the average is now just the matter of dividing it by 3. This way, the companies are able to compute their average salary without actually revealing the private information. This way, the necessity of trust and intermediary is eradicated.

4.2.6 Obfuscation

Obfuscation is the process of obscuring the meaning of communication deliberately to make the data being transmitted harder for other parties to understand. It is used in natural language and is also used in software engineering and data privacy. Although it should not be confused with ciphertext.

Unlike ciphertext, obfuscated code can be made sense of without needing to know any secret keys. However, that does not mean it is easy to decipher. In software engineering, programmers use many techniques to make their source code as unreadable as possible. Therefore, in the case of a leak, it would be quite hard for

adversaries to understand. [50]

Obfuscation is very useful to ensure intellectual protection and digital rights management. An obfuscated source code is harder to reverse engineer than otherwise. It also makes tampering with the code and its redistribution harder.

How it works:

There are many techniques to apply obfuscation. For software engineering, a code can be obfuscated by obscuring its control flows, string encryption, renaming, obfuscation, minification, and so on. Normally any combination of these techniques are used. [51]

To use string encryption, all the string values in the software binary are encrypted and replaced with ciphertext, the secret key is usually hidden in the binary and is unintelligible to the human reader. The original strings are only restored when the software process is in the runtime. This makes it extremely hard for reverser engineers to understand critical parts of the source code and their operations.

4.2.7 Data Perturbation

Data perturbation is the technique of adding noise to a dataset that makes it harder for any unauthorized entity to derive individual records of a data subject.

There are many ways to achieve data perturbation. Differential privacy, data masking, randomization, etc. are different techniques to achieve data perturbation. These techniques can be applied to a dataset in any combination.

Data perturbation is of two types: 1. Probability distribution approach and 2. Value distortion approach. [52]

In probability distribution approach, the data is replaced from the same data sample it's a part of. On the other hand, in the value distortion approach, the noise can be added to it in an additive or randomized way. It uses a decision tree classification method to decide which type of noise should be added to a data depending on various criteria - much like this guideline.

How it works: Data perturbation is an umbrella term for various techniques. These techniques are explained already in the same chapter. Please refer to their individual implementations.

4.2.8 Anonymization

Anonymization is the process of removing amany information from a dataset that can be used to link back to a data subject, violating their privacy. Anonymization removes personally identifiable information to retain its subject's anonymity.

Anonymization is an umbrella term for many techniques. Different properties of anonymization can be achieved by implementing any combination of anonymiza-

tion techniques. Such properties are k-anonymity, l-diversity, t-closeness etc. An anonymized data set is said to be k-anonymous if the properties of k-anonymity have been achieved in the data set. The techniques to achieve anonymization are: generalization, swapping, classification, suppression etc. [53]

How it works:

Data anonymization is irreversible. Meaning, it is no longer possible to retrieve personal information from the data set after anonymization has been applied. In classification technique: any value containing personal information is removed and replaced with the type of the information. For example: if the data contains: Faisal has bought a house in Munich in 2021," it will be replaced with anonymized data as "[NAME] has bought a house in [LOCATION] on [DATE]."

4.2.9 K-anonymity

K-anonymity is a concept of adapting anonymization in a way that it becomes harder to do re-identification attacks on a dataset by linking it with another dataset published elsewhere. It is a property of an anonymized dataset and there are many ways to achieve k-anonymity.

It is said that in the US 87 percent of the population can be uniquely identified with just 3 information: gender, date of birth, and zip code [54]. Considering how much data is collected from us by various data collectors and processors, it is scary that so little information is needed to link any data back to us. This is especially scary when our sensitive and critical information like: medical records and pension data are susceptible to breaches if no privacy enhancing is applied everywhere.

How it works: If a data set contains re-identifiable attributes such as name, age, gender, postcode etc, it could be used to narrow the data down to an individual by linking relevant data from other sources. Here is an example dataset before applying k-anonymity:

Table 4.1: A non anonymized data containing sensitive information

Name	**Postcode**	**Age**	**Gender**	**Dietary Preference**
Kyle	80803	34	Male	Vegetarian
Suman	91245	23	Male	Vegetarian
Masud	60654	26	Male	Halal
Muller	56541	26	Male	No preference
Susan	40865	20	Female	Vegan
Terry	40112	18	Diverse	Vegan
Vanessa	10034	45	Female	Kosher

To achieve k=3 k-anonymity on the above data, some attributes are redacted, some were generalized and some parts were replaced with quasi identifiers[1]. This data is still significantly similar from a statistical perspective but since it is k-anonymized, it's not easy to re-identify back to the data subjects. [54]

Table 4.2: The anonymized version of the data with k-3 anonymity achieved

Name	Postcode	Age	Gender	Dietary Preference
*	80***	34	Male	Vegetarian
*	91***	23	Male	Vegetarian
*	60***	26	Male	Halal
*	56***	26	Male	No preference
*	40***	20	*	Vegan
*	40***	18	*	vegan
*	10***	45	Female	Kosher

K-anonymity must follow these conditions:

The other columns that are intact must not reveal information about the column that has been generalized.

The sensitive columns should have varying values. It must not be the same throughout the dataset.

K-anonymity works when the dimension of data is low. Higher dimensional data has more probability to contain other columns that may reduce the effectiveness of generalized columns. [55]

4.2.10 L-diversity

L-diversity is an extension of k-anonymity in a way that reduces the granularity of the data representation. L-diversity addresses a few weaknesses of k-anonymity and protects against attacks such as Homogeneity attack and Background Knowledge attack. It also addresses the shortcoming of k-anonymity which doesn't provide group anonymity. [56]

L-diversity is computationally more expensive to achieve and even after the property is achieved it is still susceptible to many other attacks like skewness and similarity attacks. [55]

How it works:

A dataset is said to be l-diversified if for each group of records that share a combination of key attributes, there must be an l well represented value for each

[1] An identifier that is linked to the data subject but can not be used to uniquely identify the said subject. i.e. date of birth

sensitive attribute. In other words there must be 'l' number of distinct values in the sensitive attributes/columns.

4.2.11 T-closeness

T-closeness is an extension of l-diversity and k-anonymity as it addresses some of the shortcomings of both anonymization properties. [57]

T-closeness also addresses the two vulnerabilities of k-anonymity such as homogeneity attack and background knowledge attack.

T-closeness defines that the distribution of sensitive data in a group should not be too far from its distribution in the full population. It's very effective regarding higher dimensional data and has less utility loss compared to k-anonymity. [55] [58]

4.2.12 Pseudonymization

Pseudonymization as its name suggests, is a process and sort of achieving anonymization but with an asterisk. In other words, pseudonymization is the process in which any personally identifiable information in a data set is replaced by pseudonyms. [59]

Pseudonymization is currently promoted by European Union because it lets companies easily implement pseudo-anonymization while complying with its GDPR[2] regulations.

How it works:

The pseudonyms can be any code, randomly generated token, or data strings that are unique to the data set and are linked to the original data subject. The mapping of pseudonyms to their respective data is stored in a different database or location to reduce the chance of breach and linkage. This is a reversible process as it lets companies or data processors restore its original content by linking the pseudonyms from the mapping data. However, since the data is pseudonymized and stored separately, it reduces the chance of any adversaries to re-identify the information back to the original subjects. [37]

Pseudonymization is one of the fastest and computationally least expensive ways to achieve privacy enhancement. This is one of the major reasons European Union is advocating its adoption.

4.2.13 Randomization

Randomization is a technique where in a large dataset, the individual cells or fields or the data in those fields will be randomly arranged or changed. This is done in a

[2]General Data Protection Regulation: a set or privacy regulation provided by the European Union

way it doesn't affect the entire data set in a meaningful way. However, this added randomization makes deriving any information regarding a particular data subject less probable. [60]

How it works:

Randomization is considered a data perturbation technique. Therefore it follows the principles of data perturbation.

For example, in a Microsoft proposed data randomization technique the data is randomized by storing the XORed data in another field instead of the original data. This provides the dataset probabilistic protection and the individual data become harder to infer and link. [61]

4.2.14 Data Masking

A obfuscation technique specific to data. The term has been used interchangeably with obfuscation. [62][63]

4.2.15 Synthetic Data Generation

Synthetic data generation or random data generation is a technique to computationally generate random/fake data that would closely resemble the real data in a particular use case.

For example, you want to train your fancy new artificially intelligent video filter app that recognizes faces in a video and apply funny animal filters to it on the fly. To teach your algorithm, you would need tons of facial data of humans of different races, color, ethnicities, and features. However, managing a lot of data for this situation can be hard. Not to mention the privacy implications. However, you don't really need facial data of actual human beings to train your data. You can use AI generated image sets to train your algorithm and it would work as if it is fed real data. [64]

How it works:

There are many ways to generate synthetic data depending on the use cases. For example, if you need to train a machine learning model that recognizes hand waving in a video, you can use Deepfake to generate data sets for this training purpose.

To start with you'd need some actual video footage of people waving hands and some images of artificially generated people. Then Deepfake will use these two data to create a doctored version of the video where those persons in the artificially created images will be seen waving their hand. Deepfake will make sure their bodily features are reflected in the new video. Now can use these synthetic videos to train your application; without sacrificing anyone's privacy. [65]

4.3 Identified Criteria

The second major objective of this research project is to find the different scenarios or criteria under which conditions a certain technology is favored over the other. In this process, Barbara Kitchenham's Systematic Literature Review for Software Engineers book also came to rescue. [46] The process of finding a particular criterion is briefly described as follows:

- target a specific technology first.
- find relevant literature for this technology
- from the literature, derive a specific condition or criteria where the particular technology is favored.
- create a pool with all identified criteria
- repeat above process for all technologies
- remove duplicates, and merge similar criteria under any umbrella term

Initially the papers from the pool were prioritized for this process, however, it later became difficult to manage so many sources. The sources used for each identified criteria will be mentioned in their dedicated section.

4.3.1 Statistical Disclosure

Statistical disclosure control is useful when you want to control the statistical confidentiality of data. These are a set of measures to reduce or eliminate the risk of data being used to link back to the original subject or to expose any sensitive or personal information. [66]

4.3.2 Data Publication

If you want to publish the data you're currently dealing with, you have to take additional measures before you publish the data. [67] The data need to be processed, checked against private info and local law before it can be published. The purpose of the publication should also need to be taken into account.

4.3.3 Distributed Data

If the data is stored in multiple locations, then managing the date in all of those locations is deemed necessary. Since the data will be transported, synced and stored, it should also be made sure the data and privacy are protected at those stages. [68]

4.3.4 Edge Devices

The data stored in edge devices can be handled in two ways. In one way, the raw private data does not need to leave user devices for processing, the other relies on direct transfer of data between the client and a server. The tradeoff here is the accuracy of the insights and telemetry received from the end users. [68]

4.3.5 Data Transportation

The data that will be transferred has to be protected against certain threats, mainly the data in rest and transit are vulnerable to many attacks. Moreover if the data crossed international boundaries, the local legislations and regulations take precedence. For many reasons the data can not be transferred to another country or region. [69]

4.3.6 Need of Real Data

In some cases, the desired goal can be achieved without using real data at all. For example, if you want to generate a fake dataset, or train your ML model with similar data for purposes like image recognition, deep fake creation, and populating websites with dummy data. In these cases, the real data has no benefits over synthetically generated data, especially as the purpose is to learn the type of data, not the data itself. [39]

4.3.7 Computational Cost

The amount of processing power and memory taken to implement a specific technology can be a limiting factor in many scenarios. If the data is being processed in an IoT device or device with limited computational power, certain technologies may not be eligible to be used. [70]

4.3.8 Multiple Stakeholders

If there are more than one stakeholder involved, it can raise multiple concerns. The stakeholders may not agree on sharing their data between them. It's also possible the stakeholders do not have the same rights to the data. [8]

4.3.9 Third Parties

If a third party is involved during any phase of the data life cycle, a certain policy can restrict sharing of data with that party. If you need to share the data with a third

party, certain technologies can help alleviate this issue while providing minimal loss had it been shared without such restrictions. [8, 19]

4.3.10 Losslessness

The losslessness of a privacy enhancing technology is important when you're dealing with mission critical data and the insights derived from the data should reproduce the same output for the same input. In other words, the process of applying any privacy enhancing technology should not change the data in a way that it can affect the performance of the useful metrics. [71]

4.3.11 Accuracy

Certain privacy enhancing technology provides better accuracy than others. In other words, for the same data, certain technologies will produce consistent results. Accuracy is important when you have a huge amount of data and a little discrepancy in accuracy can lead to large change in the final output. [71]

4.3.12 Textual data

Text based data needs different processing than non-textual media-based data. Some technologies work better with text-based data. Since nontextual data are stored and formatted in specific ways, even storing the same data in different formats can lead to varying privacy enhancements by the same technologies. [68]

4.3.13 Data Dimension

This is an important metric when comparing different anonymization techniques like k-anonymity, l-diversity and t-closeness. Although there is an improvement over the former, the higher dimension a dataset is, the former tends to provide better anonymity. [57]

4.4 Mapping Data

The identified privacy enhancing technologies and the criteria are now mapped together. As mentioned before the data here is not conclusive. There are many instances where mapping one criterion to a specific PET does not make sense. For example: Edge devices and Synthetic Data Generation. There were also other instances where I couldn't find sufficient evidence to link a PET with a criterion.

The data is open source and can be found in the project's GitHub repository. Anyone can contribute to improve this data with proper sources, they just have to submit a pull request to the repository. The sources for this data are also hosted in the repository.

The tables below have 4 legends:

- Zero - 0: means the criterion does not match with the PET.
- One - 1: means the criterion matches with the PET.
- Not Applicable - N/A: means the criterion does not make sense for the PET.
- Unsure - ?: means there are not enough evidence to link the two.

Table 4.3: Mapping PETs with criteria (Part-1)

	Statistical Disclosure	**Data publication**	**Distributed data**	**Edge devices**	**Data transported**	**No need of real data**
Differential Privacy	1	1	1	1	N/A	1
Federated Learning	?	?	1	1	1	?
Zero Knowledge proof	?	1	?	?	1	0
Homomorphic Encryption	?	N/A	?	1	1	?
Secure Multiparty Computation	?	?	1	N/A	1	?
Obfuscation	?	1	N/A	N/A	N/A	?
Data Perturbations	1	1	N/A	N/A	N/A	?
Anoymization	1	1	?	N/A	N/A	?
K-anonymity	1	1	?	N/A	N/A	?
L-diversity	1	1	?	N/A	N/A	?
T-closeneess	1	1	?	N/A	N/A	?
Psuedonymization	0	1	?	N/A	N/A	?
Randomization	1	1	?	N/A	N/A	N/A
Synthetic Data Generation	1	?	?	N/A	N/A	1

Table 4.4: Mapping PETs with criteria (Part-2)

	High Computational cost	**Multi Stakeholders**	**Third party involved**	**losslessness**	**Accuracy not important**	**Textual data**	**Higher Dimension**
Differential Privacy	0	1	1	0	1	?	N/A
Federated Learning	1	1	1	?	1	?	N/A
Zero Knowledge proof	?	1	1	1	?	N/A	N/A
Homomorphic Encryption	1	?	1	1	?	N/A	N/A
Secure Multiparty Computation	1	1	1	1	?	N/A	N/A
Obfuscation	1	?	N/A	1	1	?	N/A
Data Perturbations	0	?	N/A	0	1	?	N/A
Anoymization	N/A	1	1	0	1	1	N/A
K-anonymity	0	1	1	0	1	1	0
L-diversity	1	1	1	0	1	1	0
T-closeneess	?	1	1	0	1	1	1
Psuedonymization	0	1	1	0	1	1	N/A
Randomization	?	N/A	N/A	0	1	?	N/A
Synthetic Data Generation	1	N/A	N/A	0	1	1	N/A

4.5 Key Insights

Throughout the timeline of this thesis, as I was doing research, I found out many interesting insights that are not only relevant for this thesis but also will help future researchers in the field. They can also be called the lessons learned potion of this thesis.

The major insights are listed below:

- Manual vs automatic classification: Manually searching or filtering through thousands of sources for classification is time consuming, prone to human errors and biases.
- Similarity with other fields: Privacy enhancing technology related keywords sometimes result in getting papers from mathematics and medical backgrounds.
- Grammar variants: While creating the list of words in the classification phase, one must include all the grammatical variants for the keywords for optimal results.
- Spelling matters: Some search terms provide more results than other. For example: 'privacy technique' has less number of results compared to 'privacy technology."
- Dominant term: 'Privacy enhancing' is more dominantly used in academia than 'privacy preserving.' Although these two terms are used interchangeably.
- Subset technology: Many privacy enhancing technologies are a subset of a larger privacy enhancing technology. For example: k-anonymity and l-diversity for anonymization.
- Anonymous: Many novel privacy enhancing technologies do not mention their name, or do not propose a new name for their technology in the title/abstract of the paper. They mostly just go by the attack they protect against or the super set technology they are inheriting. For example: 'a privacy enhancing technique for k-means clustering in data mining.'

5 Related Work

In this chapter, other literary work similar to this research is discussed briefly. Although there are many good pieces of literature that touch on the topic of privacy-enhancing technologies, the lack of research on properly classifying existing privacy-enhancing technologies is very evident. The chapter will hence focus on three key categories where existing literature resembles parts of this thesis directly or indirectly.

5.1 Categorization of Privacy-enhancing technologies

One interesting approach to categorize existing privacy-enhancing technologies was to tailor their applications regarding smart cities. In this paper, James Curzon et al. tried to categorize privacy-enhancing technologies by identifying privacy risks associated with smart cities. [72] In order to sustain our ever-increasing urbanization, smart cities are inevitable. With anything smart and digitalization comes a lot of technical challenges. As cities become smarter, the amount of data collected will also increase. With more data comes more privacy risk. People all over the world are also getting more conscious about their data security. With stricter regulation and consumer concerns, companies are looking towards preserving the privacy of these data.

There are many new and upcoming technologies that will enable new types of cyber-physical interactions and data transfer. In order to protect the data and preserve its privacy in every state, good knowledge of privacy-enhancing technologies is required. Technologies like big data, cloud computing, Edge Computing, and wireless sensor networks will facilitate larger amounts of data being transferred and stored.

The paper tried to categorize and map privacy-enhancing technologies in terms of three different focuses. One categorizes the types of privacy-enhancing technologies by the method they are implemented. Another categorizes the technologies in terms of their application in a different layers of smart cities. The final categorization is based on the type of data they work on. The paper also presents the privacy technologies' advantages and disadvantages. Overall their effort to streamline the classification and present the privacy-enhancing technologies in terms of smart city

applications was very helpful.

5.2 Privacy metrics identification

Finding a proper way to measure different metrics for any quantifiable or qualitative entity is the backbone is scientific study. It is one thing to measure the metric as an experimental scientist, it is another thing to propose new methods and metrics. In this paper, Isabel Wagner et al. tried to come up with many technical metrics to quantifiably measure the effectiveness of different privacy-enhancing tools and technologies. [19] It's a high-level technical overview of over 80 privacy metrics. The authors present that there are mainly four ways to classify privacy metrics. These privacy metrics will be used in turn to decide the performance of different privacy-enhancing technologies. The characteristics are 1. Adversary goals, 2. Adversary capabilities, 3. Data Sources, and 4. Inputs for computation of metrics.

There are over 80 privacy metrics detailed in this paper that are grouped by the outputs they measure. These metrics are mostly low-level and useful if someone wants to measure the effectiveness of privacy-enhancing technologies with greater granularity. Some of the metrics that were relevant for this thesis and were identified in the paper are information leakage, disclosure, k-anonymity, differential privacy, etc.

5.3 Overview of popular privacy-enhancing technologies in the financial sector

Before starting this thesis it was important to get an overview of the current privacy-enhancing technologies and some of their use cases. It is also important to understand how each of these technologies works and how they can adapt to different contexts if need be. This white paper released by the World Economic Forum [8] provides a very intuitive overview of 5 privacy-enhancing technologies. The descriptions are easily understandable and written in a way anyway without any technical background can get the gist of it.

The paper acknowledges the benefits of data sharing in the banking and financial sectors and understands to keep up with the ever-progressive technologies, the financial sector has to adapt to survive. To get a better overview of the collected data, these financial organizations often have to collaborate and share data between them. With more shared data comes the potential of better analytics.

However, the paper also argues financial organizations must not neglect the drawbacks of this data sharing. As regulations are getting stricter, maintaining user privacy is no longer a feature, it's a requirement. The benefits and drawbacks of data sharing among financial organizations are three folds: 1. To institutions, 2. To the regulators, and 3. Consumer-facing.

The paper talks about five prominent privacy-enhancing technologies that could be useful in the financial sector. It goes over differential privacy, homomorphic encryption, zero-knowledge proof, federated learning, and secure multiparty computation. In each explanation, the author reiterates where these technologies could have been used by describing a prior financial mishap. It also gives some ideas about the potential subsectors and applications where these privacy-enhancing technologies can be employed.

6 Web App Implementation

The recommender system will be implemented as a web application. The initial primary goal of this thesis was to create a decision support tree that would help the user reach their desirable privacy-enhancing technology by traversing the tree. However as the data get more complex and multifold, it became clear that it would be very unintuitive for the users to traverse the tree in a meaningful way. Additionally, creating such a tree from multi-dimensional open-ended data is also very challenging.

Here's a visual representation of the original idea:

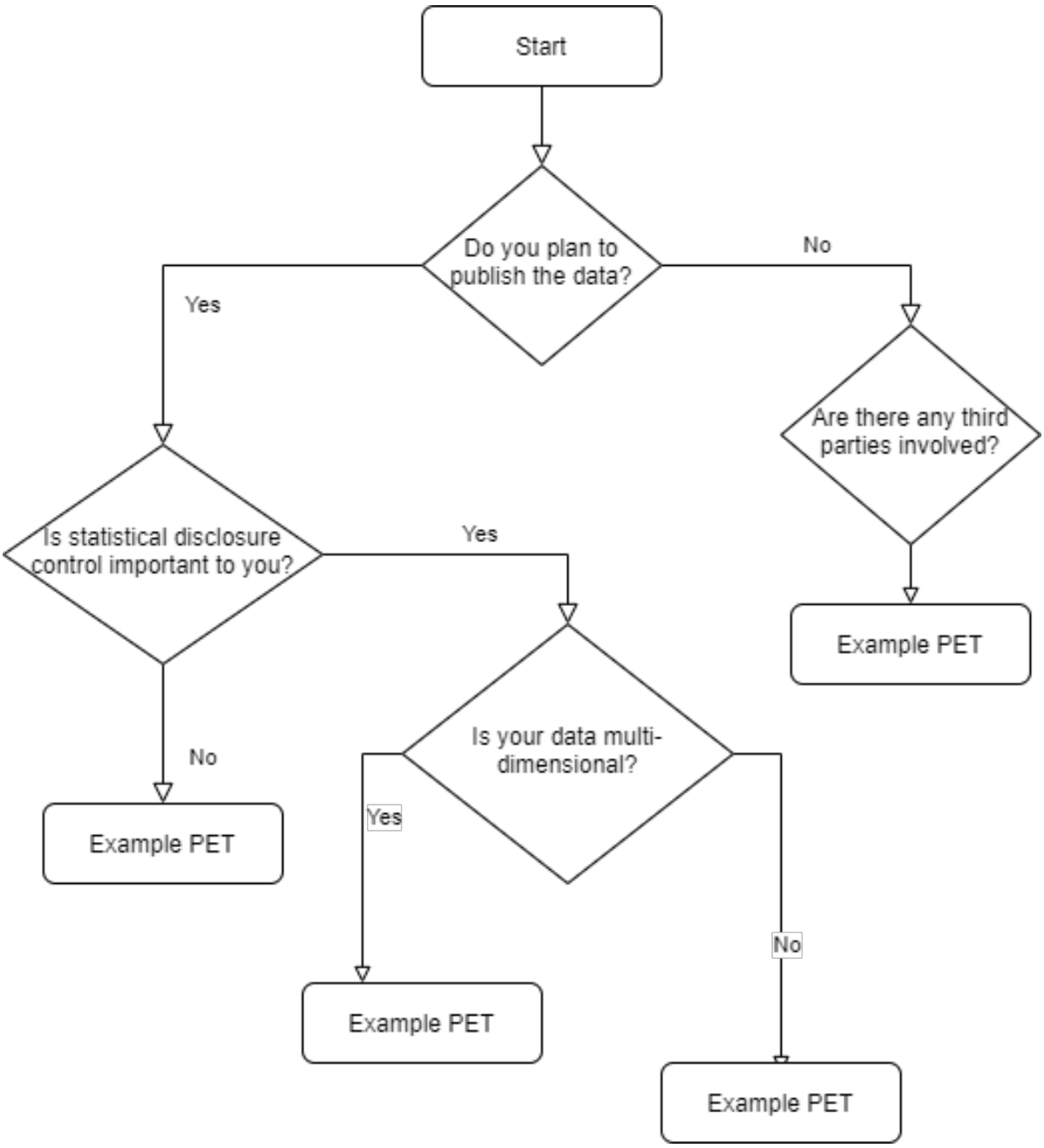

Figure 6.1: An example of how the PET decision chart would look

6.1 Architecture

The application will be based on ReactJS. The app uses many reactive elements to work and React's one-way data binding is a necessary aspect of the application. The

app is a front-end single page application[1]. All data and the business logic of the application are processed at the client side. Hence the app uses no server or database. As a result, the app can work offline as a Progressive Web App as well.

The technologies and tools used to develop the web application are mentioned below:

- React: A web framework for the front end. Functional components were used with the help of React Hooks.
- NextJS: React-based front end framework suitable for static websites. The web application is bootstrapped from create-next-app. NextJS provides built-in support for CSS modules and easier routing. Deploying the application is also fairly simple thanks to NextJS' parent Vercel.
- Ant Design/UI: This React-based UI library was used to quickly prototype the design element. This also provides consistent UI elements and many built-in functionalities. This is an alternative for Google's Material UI.
- JavaScript: Plain version of JavaScript 2015 (ES6) will be used as the language of the application. Typescript was considered for better debugging and tooling support. However, plain JavaScript was chosen in the end for flexibility.

The information provided in the app should be always kept up to date. The source code for the application and data is hosted on GitHub[2]. Everyone is encouraged to contribute the application, the mapping data, and revise the existing data with newer information. As the current data is not conclusive and lacks much mapping information, this open-source approach would solve this issue and keep non-experts updated regarding privacy-enhancing technologies.

6.2 Design

The web application is divided into a few pages. The navigation is managed by NextJS' implementation of React Router.

When a user first runs the application, he is greeted with a page that consists a list of questions at the sidebar and the question itself. This is called the Hybrid mode. How it works and why it is designed this way will be mentioned later in their dedicated sections.

[1] A web application that works as a mobile application in a sense that it doesn't require the browser to reload a page to fetch and present users with new data

[2] GitHub repository: https://github.com/FaisalBinAhmed/PET-Guideline

As previously shown in the mockups section, the app will consist of three main user stories:

- As a user, I want to answer a few questions from the homepage to get my appropriate privacy-enhancing technology(s)
- When provided with an answer, I want to see the justifications for the answers. I can also choose an alternative if I please.
- As a non-expert user, I want to learn more about each technology and criteria right from the website. Additional sources may be supplied as well.

To fulfill these requirements as well as additional options. These features were selected to be implemented:

- Homepage with a clear structure
- Navigation element to show users their current progress
- Information regarding technologies and the criteria
- Ability to view multiple alternative privacy-enhancing technologies
- Option to edit the answer
- Option to skip an answer
- ability to gather more knowledge via internal and external sources.

These options give users great flexibility and are tailored toward their expertise level. The application is built as a tool that will help professionals get their job done.

The app uses two modes for the users. Since we are targeting nonexperts professionals, there will be one mode of the application that is specifically tailored towards them. It is aptly named the "Easy Mode." There are mainly two reasonings for it:

- The data this guideline is based on isn't conclusive. There are many instances where it was not possible to map a criterion with a specific privacy-enhancing technology. Furthermore in many instances, some of the criteria do not make sense or are not applicable for a specific privacy-enhancing technology.
- Many criteria aren't really a deal-breaker to many professionals; hence, asking them to answer it or expecting them to know about each criterion is unrealistic. The strict answering rules might lead them to pick an answer wrongly.

Hence the "easy mode" was introduced. This mode is the default behavior of the application. However, it's also understood that some users might be more expert than the others or sometimes they might need a definitive answer from the guideline. As a result, there is also another mode called the "Strict Mode."

The Strict Mode will strictly follow the mapping data and adapt the questions based on the previous answers. More information regarding this mode is in the following section.

6.3 Easy Mode

Since the application is tailored towards non-expert people, it was decided to include what's called the Easy mode. As mentioned previously, the easy mode lets the non-experts people decide about the privacy-enhancing technology in a less strict way.

How it works The basic idea behind the easy mode is simple. The user is asked all the questions regarding each criterion/scenario. After each answer, the suitable Privacy-enhancing technologies are put in a pool. When every answer is answered, the best answers from the pool are picked as the suggested privacy-enhancing technology. The best answer is picked on the basis of two conditions. The technology that satisfies the most conditions. If multiple candidates are there, then the candidate that mismatches with the least number of choices from the user is pushed to the top. This happens due to there not being enough mapping data as the source - as in there are some mapping where the data is unsure or irrelevant.

If there are fewer than two PETs that match both in terms of matched and missed answers, the next best PET from the pool is selected. The pseudo-code for this process is depicted below:

```
IMPORT questions[ ], mappingData[ ]

VAR inputs[ ], pool [ ], sortedPool[ ]

FOR each question in questions[ ]
       INPUT answer
       STORE answer to inputs[ ]

FOR i in all inputs[ ]
       FOR j in all pets[ ]
```

```
                IF i matches j
                STORE the i value in the pool[ ]

FOR each item in pool[ ]
        COUNT the number of occurrences per each item
        SORT the value from highest to lowest
        STORE the new key-value pair in sortedPool[ ]

INT highestScore, secondHighestScore
VAR bestPet[ ], secondBestPet[ ]

CALCULATE highestScore and secondHighestScore from sortedPool[ ]
FOR each item in sortedPool
        IF item matches highestScore
                STORE item in bestPet[ ]
        ELSE IF item matches secondHighestScore
                STORE item in secondBestPet[ ]
VAR suggestedPet[ ], alternativePet[ ]

FOR each item in bestPet ] and secondBestPet[ ]
        FOR each mapping in petData[ ]
                IF item does not match mappingData[ ]
                COUNT additional missedItem for the item

SORT bestPet[ ] and secondBestPet[ ] ascending by missedItem count

IF bestPET has more than 2 items
IF bestBet[0] has the least missedItem
        STORE bestPet[0] in suggestedPet[ ]
        STORE other items to alternativePet[ ]
ELSE IF there are multiple items with the same missedItems
        STORE all items to suggestedPet[ ]

ELSE STORE bestBet[ ] in suggestedPet[ ]
        AND IF secondBest[ ] has more than 2 items
                STORE the first two in the alternativePet[ ]
                ELSE STORE all to the alternativePet[ ]
```

```
OUTPUT suggestedPet[ ] and alternativePet[ ]
```

Listing 6.1: Pseudo-code for the easy mode PET suggestion

There are many instructions in this pseudo-code that are implemented as a separate function. For example the SORT, COUNT, and CALCULATE instructions are actually custom functions written in vanilla JavaScript. Moreover, the questions and mappedData arrays are complex JavaScript object arrays; each time these arrays were accessed, it required special attention to code as writing code that manipulates or read multi-dimensional data is very hard. It is even harder to match two multi-dimensional arrays as such. Nonetheless, this algorithm actually works. The actual implementation of this algorithm can be found in the thesis's GitHub repository.

6.4 Strict Mode

The Strict mode as its name suggests, follows the mapping data strictly. Since according to the mapping data, it does not make sense to ask the user about all the criteria. The application chooses one of the criteria from his best suggested privacy-enhancing technology from the Non-expert mode, and dynamically adjusts the rest of the question based on the following answers.

This mode helps people who have stricter requirements about specific technologies and criteria, but also someone who is a bit more knowledgeable in the field. It does not make sense for our main target users to use this mode. However, for some edge cases, the functionality is still there.

– The pseudo-code for this process is depicted below:

```
IMPORT questions[ ], mappingData[ ]

VAR inputs[ ], pool [ ], counter, index
VAR dynamicQuestions[ ]
SET dynamicQuestions[0] to questions[0]
SET counter to 1
SET index to 0

WHILE counter > 0

INPUT answer for dynamicQuestions[index]
STORE answer to pool[index]
FOR each data in mappingData[ ]
       IF answer matches pet in data
              PUSH pet to pool[ ]
```

```
            FOR each pet in pool[ ]
            FIND questions from questions[ ] that matches ALL pet in pool[ ]
            IF no question matches BREAK
            ELSE IF the questions are not already in dynamicQuestions[ ]
                   PUSH questions to dynamicQuestions[ ]
                   SET counter to the number of questions
            ELSE REMOVE pet from pool[ ]
      counter--


OUTPUT pool[ ]
```

Listing 6.2: Pseudo-code for the Strict mode PET suggestion

6.5 Data Schema

The application uses Javascript objects to store raw data. There are mainly two kinds of data in the application: 1. Question Data: containing the criteria and respective questions, and 2. PET Data: containing privacy-enhancing technologies and the mapping information.

6.5.1 Question Data

The Question Data is an array of objects that contains the questions and their relevant information in the application. The schema for the objects looks like this:

Table 6.1: Question Data Schema

Key	**Explanation**
id	A unique ID for the question
name	The text of the question
importance	The description shown below the question
yes	Small description why one should choose yes
no	Small description why one should choose no
step	The header for the sidebar and pop-over

6.5.2 PET Data

PET data stores information regarding different privacy-enhancing technologies and their corresponding mapping relation to identified criteria. The mapping information

is stored in an array with respect to the cell position in the mapping data table. Here is how the schema looks like:

Table 6.2

Key	Explanation
id	A unique ID for the technology
name	The name of the technology
arr	Mapping info stored as an array of numbers
info	Basic information about the technology
how	Information regarding how the technology works
links	The useful links. It is an array of objects

6.6 App elements and features

6.6.1 The Homepage

The homepage is the starting point of the application. By default, the user is greeted in the easy mode. On the homepage, they are presented with a set of questions. The homepage is divided into two main sections:

- Side panel: consists of the criteria that denote each question. The sidebar also acts as an answer record and navigational element.
- Body: it consists of the question card and the logo of the application.

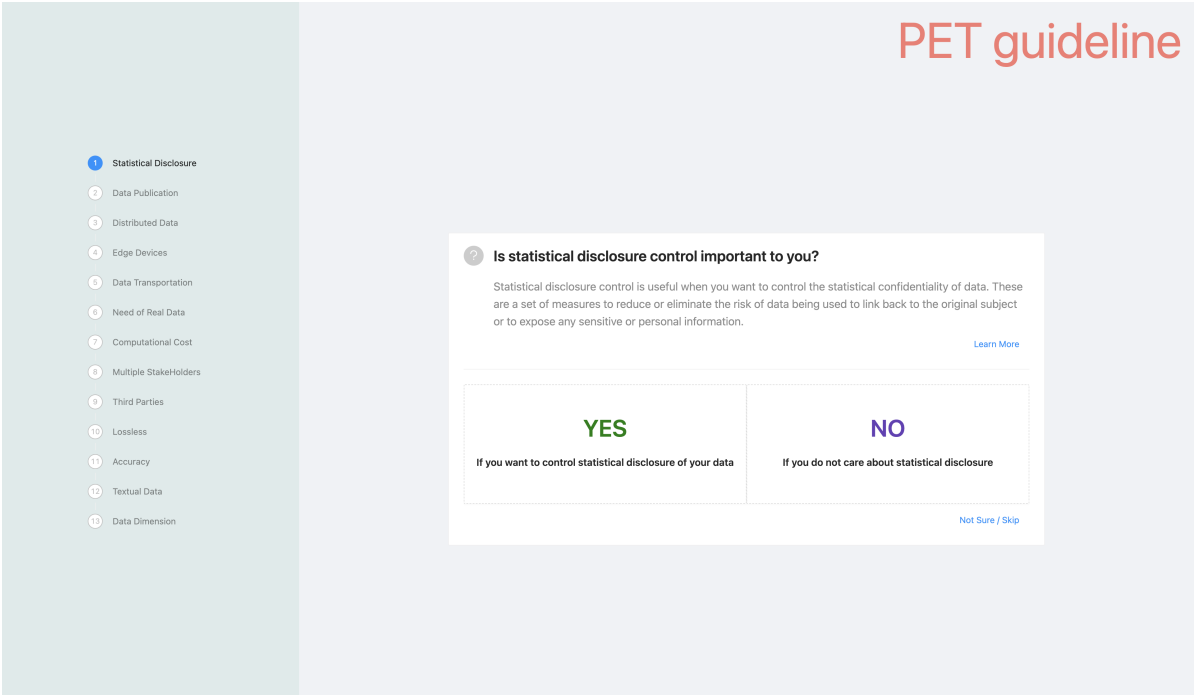

Figure 6.2: The homepage of the application

6.6.2 Question panel

The question panel has two main sections. The first section consists of the question regarding the criteria/scenario and its short explanation. It also includes a 'more information' panel that gives a brief explanation of the criteria in question.

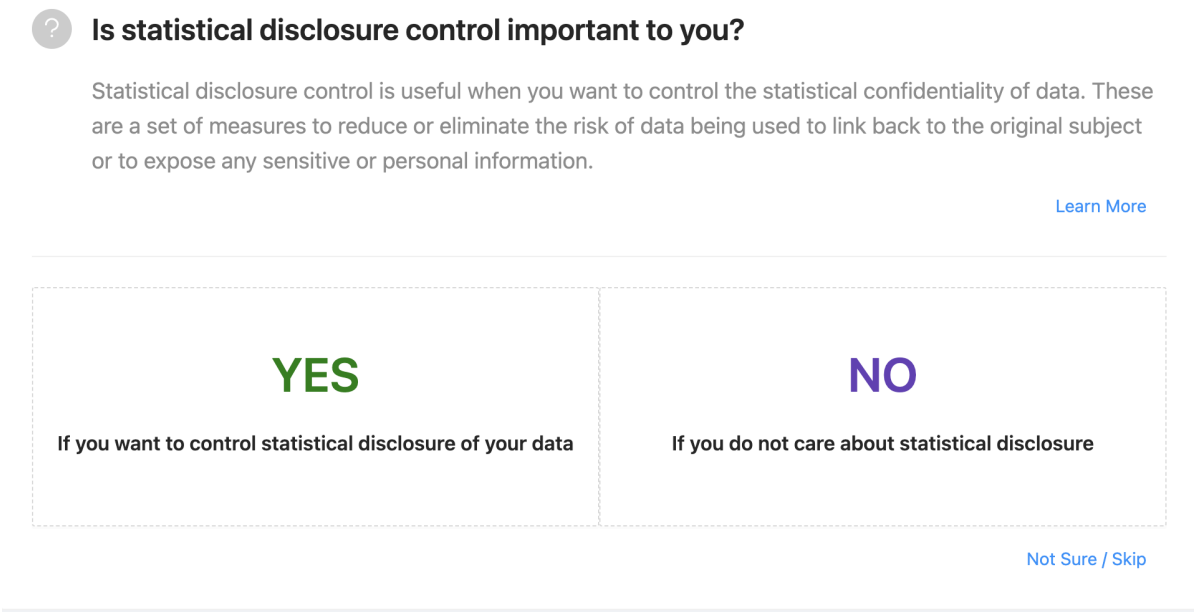

Figure 6.3: A card featuring the question

In the first section, there is a brief explanation about the criteria mean. This is there to help novice users gain some quick information regarding what they are answering for. Of course, this is not sufficient to gain full knowledge. That is why there is a more info panel below.

When the user clicks on the more info button. a fly-out comes out. In the dialogue box, there is more information regarding the specific criteria. The application also provides useful scientific sources and a button to quickly search about the topic.

6.6.3 Side panel

The side panel includes the steps in answering the questions the user is currently in. This is included to help the user quickly identify their progress in answering the question. The current step is marked in the side panel.

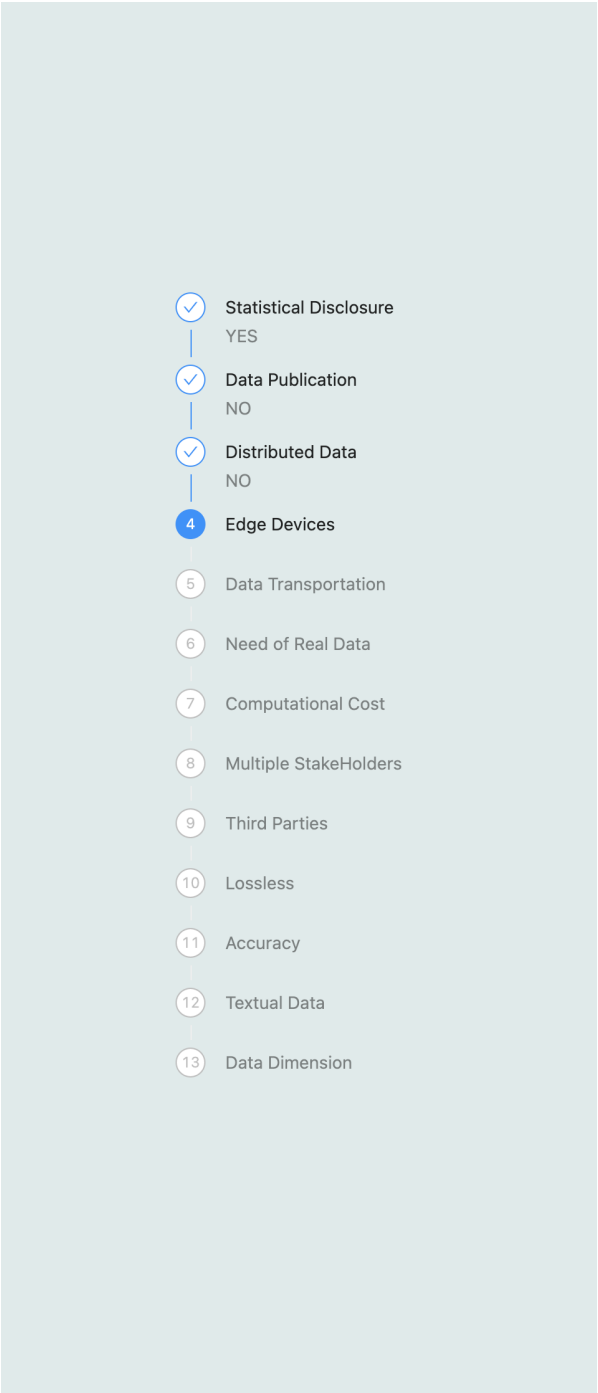

Figure 6.4: The sidebar containing current progress and previous choices

The side panel also has additional features. After each question is answered, the side panel visually shows the answered question in the respective title. Moreover, the side panel items are clickable; users can click on this item and go back to the

respective question.

6.6.4 Edit Answer

The users also have the ability to changes their mind after answering the question. Since the users are no experts, it is expected that many of them will understand something later in the process and will want to change their previous answers.

6.6.5 Result section

The result panel is again divided into two sections. The left side shows the question and the user's selected answer. While the right side shows the user the desired results. The results are categorized into two sections:

- The Best PET(s)
- Your Alternative PET(s)

Each section can have more than one technology. Usually, the technologies in the same section will have the same number of matched criteria for the user's selection. If two Privacy-enhancing technologies match the same number of criteria, then they are sorted based on the least number of missed criteria.

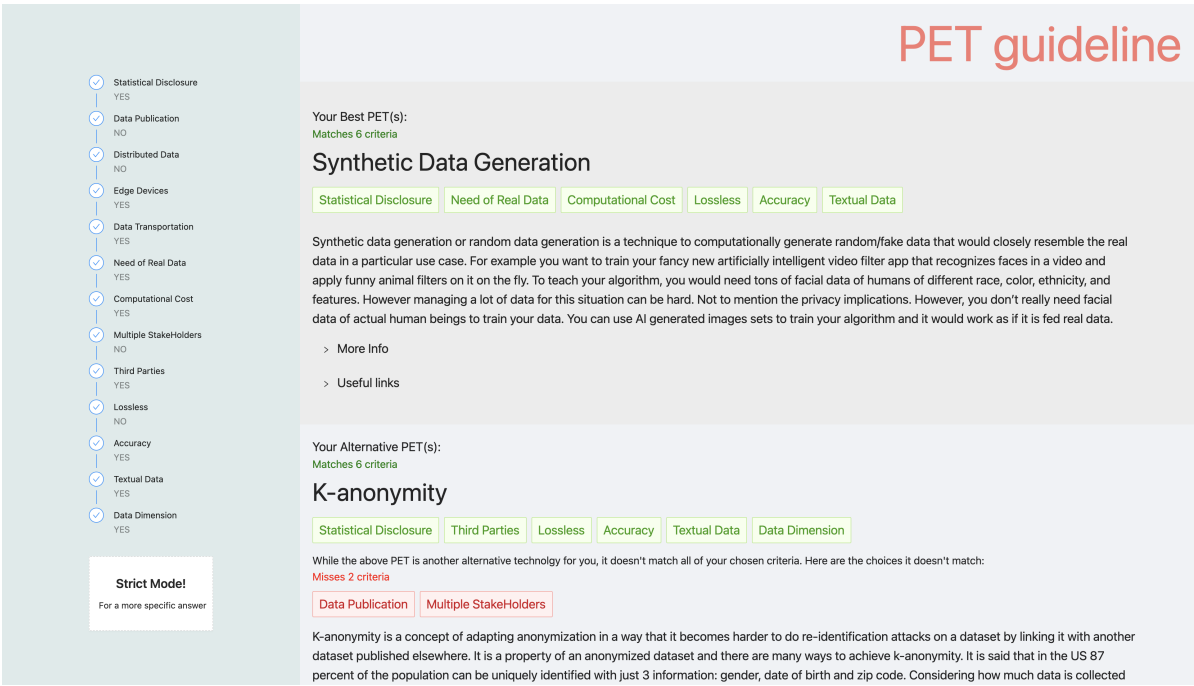

Figure 6.5: The result page is presented when the users answers the questions

6.6.6 Result Card

The results in terms of suggested privacy-enhancing technologies are presented in a way that the user can get maximum information at a glance. During the design process of this component, it was also considered that the viewer should not be overwhelmed with too much information.

The result will show the most important PETs in a distinct color. The matched criteria with the user's answer and the missed criteria are shown in identifiable green and red colors. Users can easily check their prior answers in the sidebar to get an overview of the PETs' compatibility.

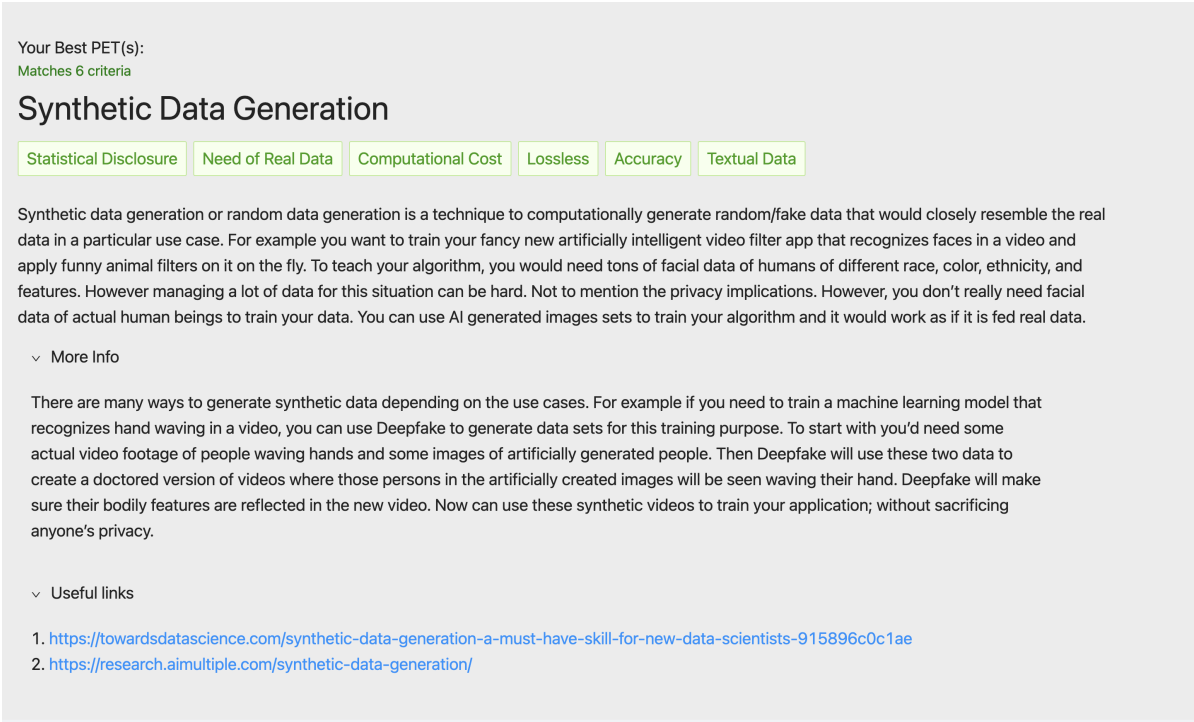

Figure 6.6: The suggested privacy enhancing technology in a result card

As can be seen from the above picture, the user is presented with 3 kinds of textual information. There is a short overview of the technology that is always visible. By clicking on the more info button, users can expand to learn more info about how the technology works. Finally, in the useful section, some internet resources are provided that will guide the user to further information on the technology, its application, and implementation.

6.6.7 Quick view information

In a previous design of the result card, the missed criteria will have their own cards containing excerpts about their core idea. However, this resulted in overwhelming information density. So, I decided to use a pop-over flyout to provide users the same glanceable information but on a mouse hover.

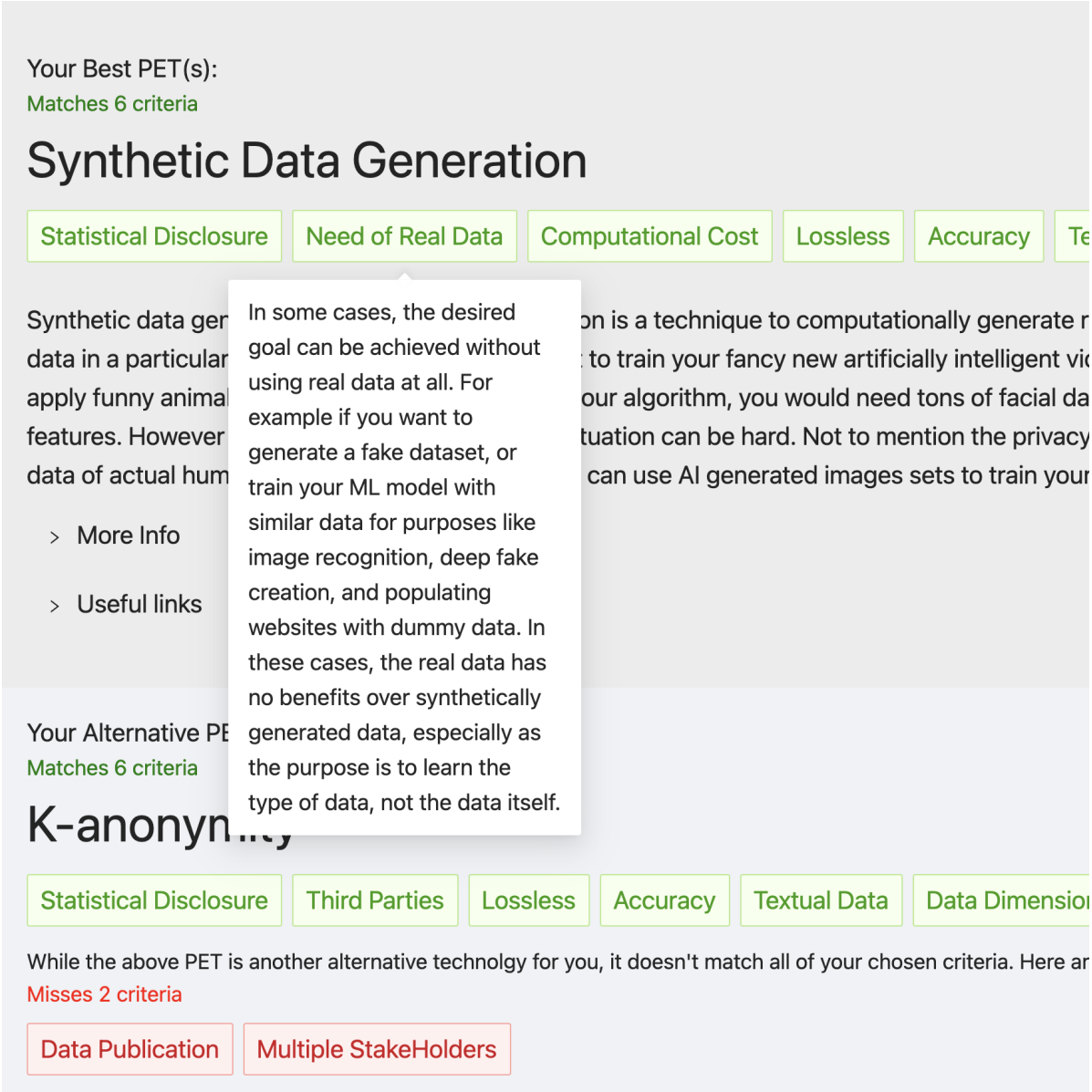

Figure 6.7: Putting mouse on the criteria shows quick excerpt about it

After following these steps, I hope the user will get a good overview of their

appropriate privacy-enhancing technologies. If the user is still not satisfied with the answer and wants a more definitive answer, they can use the strict mode.

7 Conclusion

This thesis set out to find a novel solution to help non-experts find a suitable privacy enhancing technology for their use case. During the thesis, I have faced many hurdles to come up with a viable solution that works in the present and has the necessary foundation to adapt in the future. With the help of a novel systematic literature review protocol, this literature shows how to target a broad topic and narrow it down to propose a solution for a niche audience. Finally, the main research goals of this thesis have been achieved and they are summarized below in terms of the research questions.

7.1 Summary

In this section each research question and how this thesis achieved to answer them are described. A short overview of the process, sources and solutions are provided.

Research question 1: What are the state-of-the-art privacy-enhancing technologies and their distinct characteristics?

The primary goal to achieve a solution for this question was to identify the current status of privacy enhancing technologies and their distinct characteristics. As there were no standardized specifications that identify and list these technologies, I had to start with a broader perspective. While many of the technologies found in this step were already known beforehand, a literature must describe the scientific process of how to identify the individual technology.

To identify the technologies, I started by identifying a set of keywords that would include the maximum number of technologies when queried for them. As expected, I discovered many related keywords from the initial set that helped me find a larger set of technologies in the latter stage of the process. The scientific protocol designed for this task is deemed adequate.

Once the privacy enhancing technology related keywords were identified, the major task then was to query them and filter the relevant sources from the results. As I started from a broad perspective, the number of results that needed to be dealt with became implausible to filter manually. Fortunately, I came up with an automated

solution that helped reduce the amount of time needed substantially. This is a major bi product of this research and the protocol should help future researchers who want to do a similar type of systematic literature review.

Finally, the PETs were identified by merging and classifying those keywords per paper basis. Using a standard literature review protocol, their distinct characteristics were also identified from the relevant sources.

Research question 2: What are the different scenarios where privacy enhancement is required and their corresponding basic requirements?

Once the privacy enhancing technologies and their characteristics were identified in the previous stage, the next step was to identify the different scenarios or the so-called criteria where a particular technology makes sense. As the technologies and their characteristics were identified by the systematic literature review, oftentimes the very papers that mention these technologies would describe their applications and scenarios where they make sense. Other times, the scenarios were identified from existing mapping studies and privacy enhancing technology focused literature.

Many of the criteria or scenarios are based on the literature found in the first stage. After identifying the initial list of criteria, decisions were made to merge them and filter them based on their similarities and relevance. Many initially identified scenarios were later discarded as they deemed too closely related to another.

The final set of criteria were then selected for the eventual mapping. While many of the criteria do not make sense for some of the privacy enhancing technologies, they are aptly marked to reduce any sort of anomaly at the latter stage. There is also an option to revise and update this mapping data by independent researchers in the future.

Research question 3: How to guide the decision-making process to allow diverse stakeholders make informed decisions about the use of an appropriate privacy-enhancing technology?

The ultimate goal of this thesis was to combine the information identified in the first two phases and provide a meaningful way to make it useful for nonexperts. Initially the target was to create a decision support tree which can be used to find an appropriate privacy enhancing technology. As mentioned before, the nonexpert would traverse the tree and answer a set of criteria-based questions. Eventually, they would reach a leaf node and be provided with an appropriate PET(s).

Due to technical constraints and usability factors, the solution to present the information was changed. It was decided that a web application would provide a good solution for the targeted audience: people who are nonexperts. As nonexperts

tend to have less clear knowledge about their specific need and criteria, the web application was adopted to keep that in mind. The application provides two ways to suggest appropriate privacy enhancing technologies to its users. By default, it works in the so-called easy mode. And in strict mode, it closely resembles the tree traversing method, albeit in a visual, web based way.

As nonexperts might need clearer context and information to make an informed decision, plenty of knowledge and sources are presented in the application to help them achieve just that. The application was designed with one goal in mind: provide usability and clear flow to the users. Since nonexperts may change their decision midway, the application provides a way to revise their decision at every stage.

Finally, the mapping data and the web application provides a sound and future proof way for nonexperts to learn and make informed decisions about privacy enhancing technologies. The application should cover most of their basic needs.

7.2 Limitation

This research work is not perfect as during its conducting period myriads of drawbacks were faced. However, the limiting factors can be summarized in two ways.

Firstly, the lack of concrete data linking different criteria and privacy enhancing technologies. There were no prior mapping studies that linked standardized criteria or scenarios with PETs that fulfill them; Let alone one for nonexperts professionals. There were many instances where identified criteria could not be mapped to a PET due to there being no solid evidence of their relation.

Secondly, the lack of instruction or guidance on how to create a guideline based on systematic literature review in the perspective of software engineering. While there are many protocols that provide a way to conduct a systematic literature review, the lack of scientific literature describing a way to guide decision makers based on that information was evident. Hence, the protocol and the web based approach were introduced in this research.

7.3 Future Work

Due to time constraints and the nature of a master thesis, there are still many steps left to do to improve the primary goal of this research. Firstly, the web application needs user testing and formal feedback. Interviews with practitioners can be arranged to gain their feedback on the application and how to improve it. Secondly, potential nonexperts can try this application and evaluate it based on their experience and expectations. No application is bug free in its initial release, so more intense testing

with diverse audiences and inputs is needed.

Thirdly, the data used by the application is not conclusive. As more literary work is done in the future, the data will need to be updated and revised. There are many instances where mapping was insufficient. More research work needs to be done to provide a more robust source of truth.

Finally, the application can be integrated with a general privacy recommendation platform. Machine learning can be applied to the PET selection algorithm for a better and more sophisticated output.

List of Figures

1.1 Asking users questions 5
1.2 Suggesting a privacy enhancing technology at the end of the process . 6

3.1 Research protocol 15
3.2 Publish or Perish application 18
3.3 Zotero application 20
3.4 SketchEngine corpus 22
3.5 The CSV file in Microsoft Excel (assigned PET tags in the last column) 25
3.6 The program running in VS Code 28
3.7 Automating the source finding process 32
3.8 Source classification and filtering process 34
3.9 Scribbr web application for literature and bibliography management . 35

4.1 Identified PETs after the initial phase 37
4.2 Year by year trend in privacy enhancing technology papers 38

6.1 An example of how the PET decision chart would look 59
6.2 The homepage of the application 67
6.3 A card featuring the question 68
6.4 The sidebar containing current progress and previous choices 69
6.5 The result page is presented when the users answers the questions . . 70
6.6 The suggested privacy enhancing technology in a result card 71
6.7 Putting mouse on the criteria shows quick excerpt about it 72

List of Tables

3.1 Number of sources after manual adaptation 31
3.2 Time saving of the automated classification process 31

4.1 A non anonymized data containing sensitive information 44
4.2 The anonymized version of the data with k-3 anonymity achieved . . 45
4.3 Mapping PETs with criteria (Part-1) 52
4.4 Mapping PETs with criteria (Part-2) 53

6.1 Question Data Schema 65
6.2 66

Listings

3.1 Pseudo code for the automated classification program 30

6.1 Pseudo-code for the easy mode PET suggestion 62
6.2 Pseudo-code for the Strict mode PET suggestion 64

Bibliography

[1] I. Goldberg, D. Wagner, and E. Brewer. "Privacy-enhancing technologies for the internet". In: *Proceedings IEEE COMPCON 97. Digest of Papers*. IEEE. 1997, pp. 103–109.

[2] M. Madden and L. Rainie. "Americans' views about data collection and security". In: *Pew Research Center* (2015).

[3] K.-W. Wu, S. Y. Huang, D. C. Yen, and I. Popova. "The effect of online privacy policy on consumer privacy concern and trust". In: *Computers in human behavior* 28.3 (2012), pp. 889–897.

[4] P. Ferrara and F. Spoto. "Static Analysis for GDPR Compliance." In: *ITASEC*. 2018.

[5] Y.-S. Martin and A. Kung. "Methods and tools for GDPR compliance through privacy and data protection engineering". In: *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. 2018, pp. 108–111.

[6] L. F. Cranor and N. Sadeh. "Privacy engineering emerges as a hot new career". In: *IEEE Potentials* 32.6 (2013), pp. 7–9.

[7] V. Seničar, B. Jerman-Blažič, and T. Klobučar. "Privacy-enhancing technologies—approaches and development". In: *Computer Standards & Interfaces* 25.2 (2003), pp. 147–158.

[8] World Economic Forum". *The Next Generation of Data-Sharing in Financial Services: Using Privacy Enhancing Techniques to Unlock New Value*. Tech. rep. Sept. 2019. URL: `https://www.weforum.org/whitepapers/the-next-generation-of-data-sharing-in-financial-services-using-privacy-enhancing-techniques-to-unlock-new-value`.

[9] T. Gerety. "Redefining privacy". In: *Harv. CR-CLL Rev.* 12 (1977), p. 233.

[10] J. Rubenfeld. "The right of privacy". In: *Harvard Law Review* (1989), pp. 737–807.

[11] S. Greengard. "Privacy matters". In: *Communications of the ACM* 51.9 (2008), pp. 17–18.

[12] J. Domingo-Ferrer and A. Blanco-Justicia. "Privacy-Preserving Technologies". In: *The Ethics of Cybersecurity*. Ed. by M. Christen, B. Gordijn, and M. Loi. Cham: Springer International Publishing, 2020, pp. 279–297. ISBN: 978-3-030-29053-5. DOI: 10.1007/978-3-030-29053-5_14. URL: https://doi.org/10.1007/978-3-030-29053-5_14.

[13] G. Van Blarkom, J. J. Borking, and J. E. Olk. "Handbook of privacy and privacy-enhancing technologies". In: *Privacy Incorporated Software Agent (PISA) Consortium, The Hague* 198 (2003), p. 14.

[14] C. Dilmegani. *Top 10 Privacy Enhancing Technologies (PETs) in 2021*. July 2021. URL: https://research.aimultiple.com/privacy-enhancing-technologies/.

[15] J. W. Anderson, K. E. Kennedy, L. B. Ngo, A. Luckow, and A. W. Apon. "Synthetic data generation for the internet of things". In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE. 2014, pp. 171–176.

[16] J. Heurix, P. Zimmermann, T. Neubauer, and S. Fenz. "A taxonomy for privacy enhancing technologies". In: *Computers & Security* 53 (2015), pp. 1–17.

[17] C. Gentry. *A fully homomorphic encryption scheme*. Stanford university, 2009.

[18] S. Stalla-Bourdillon and A. Knight. "Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data". In: *Wis. Int'l LJ* 34 (2016), p. 284.

[19] I. Wagner and D. Eckhoff. "Technical privacy metrics: a systematic survey". In: *ACM Computing Surveys (CSUR)* 51.3 (2018), pp. 1–38.

[20] C. Dwork and J. Lei. "Differential privacy and robust statistics". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, pp. 371–380.

[21] A. Chin and A. Klinefelter. "Differential privacy as a response to the reidentification threat: The Facebook advertiser case study". In: *NCL Rev.* 90 (2011), p. 1417.

[22] Apple Inc." *Learning with Privacy at Scale*. Dec. 2017. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[23] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečn, S. Mazzocchi, H. B. McMahan, et al. "Towards federated learning at scale: System design". In: *arXiv preprint arXiv:1902.01046* (2019).

[24] G. Long, Y. Tan, J. Jiang, and C. Zhang. "Federated learning for open banking". In: *Federated learning*. Springer, 2020, pp. 240–254.

[25] Tom Simonite, wired.com". *Apple and Google's AI wizardry promises privacy—at a cost*. June 2021. URL: https://arstechnica.com/gadgets/2021/06/apple-and-googles-ai-wizardry-promises-privacy-at-a-cost/?comments=1.

[26] J. Fan and F. Vercauteren. "Somewhat practical fully homomorphic encryption." In: *IACR Cryptol. ePrint Arch.* 2012 (2012), p. 144.

[27] X. Yi, R. Paulet, and E. Bertino. "Homomorphic encryption". In: *Homomorphic Encryption and Applications*. Springer, 2014, pp. 27–46.

[28] *Microsoft SEAL (release 3.6)*. `https://github.com/Microsoft/SEAL`. Microsoft Research, Redmond, WA. Nov. 2020.

[29] H. P. Byun. *Homomorphic Encryption and Multiparty Computation*. Nov. 2020. URL: `https://baffle.io/blog/homomorphic-and-multiparty-computation/`.

[30] A. Ben-David, N. Nisan, and B. Pinkas. "FairplayMP: a system for secure multi-party computation". In: *Proceedings of the 15th ACM conference on Computer and communications security*. 2008, pp. 257–266.

[31] O. Goldreich and Y. Oren. "Definitions and properties of zero-knowledge proof systems". In: *Journal of Cryptology* 7.1 (1994), pp. 1–32.

[32] S. Grzonkowski, W. Zaremba, M. Zaremba, and B. McDaniel. "Extending web applications with a lightweight zero knowledge proof authentication". In: *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. 2008, pp. 65–70.

[33] D. Hopwood, S. Bowe, T. Hornby, and N. Wilcox. "Zcash protocol specification". In: *GitHub: San Francisco, CA, USA* (2016).

[34] C. S. Collberg and C. Thomborson. "Watermarking, tamper-proofing, and obfuscation-tools for software protection". In: *IEEE Transactions on software engineering* 28.8 (2002), pp. 735–746.

[35] S. Aonzo, G. C. Georgiu, L. Verderame, and A. Merlo. "Obfuscapk: An open-source black-box obfuscation tool for Android apps". In: *SoftwareX* 11 (2020), p. 100403.

[36] A. Bourka, P. Drogkaris, I. Agrafiotis, and European Network and Information Security Agency". *Pseudonymisation Techniques and Best Practices*. Stuttgart, Germany: UTB, 2019.

[37] M. Hintze and K. El Emam. "Comparing the benefits of pseudonymisation and anonymisation under the GDPR". In: *Journal of Data Protection & Privacy* 2.2 (2018), pp. 145–158.

[38] J. Dahmen and D. Cook. "SynSys: A synthetic data generation system for healthcare applications". In: *Sensors* 19.5 (2019), p. 1181.

[39] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, et al. "Deepfacelab: A simple, flexible and extensible face swapping framework". In: *arXiv preprint arXiv:2005.05535* (2020).

[40] Zcash". *What are zk-SNARKs?* Sept. 2019. URL: https://z.cash/technology/zksnarks/.

[41] K. Hao. *How Apple personalizes Siri without hoovering up your data*. Apr. 2020. URL: https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/.

[42] Google". *Federated Learning: Collaborative Machine Learning without Centralized Training Data*. Apr. 2017. URL: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

[43] A. Sethi. "Digital rights management and code obfuscation". MA thesis. University of Waterloo, 2004.

[44] S. Schrittwieser and S. Katzenbeisser. "Code obfuscation against static and dynamic reverse engineering". In: *International workshop on information hiding*. Springer. 2011, pp. 270–284.

[45] S. Laurila. "Comparison of JavaScript Bundlers". In: (2020).

[46] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. "Systematic literature reviews in software engineering–a systematic literature review". In: *Information and software technology* 51.1 (2009), pp. 7–15.

[47] World Economic Forum". *The Next Generation of Data-Sharing in Financial Services: Using Privacy Enhancing Techniques to Unlock New Value*. Sept. 2019. URL: https://www.weforum.org/whitepapers/the-next-generation-of-data-sharing-in-financial-services-using-privacy-enhancing-techniques-to-unlock-new-value.

[48] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. "Privacy loss in apple's implementation of differential privacy on macos 10.12". In: *arXiv preprint arXiv:1709.02753* (2017).

[49] O. V. Foundation. *The Future of Voting: End-to-End Verifiable Internet Voting — Specification and Feasibility Study*. (I was on the Advisory Council for this report.) July 2015.

[50] J.-T. Chan and W. Yang. "Advanced obfuscation techniques for Java bytecode". In: *Journal of systems and software* 71.1-2 (2004), pp. 1–10.

[51] PreEmptive". *What is Obfuscation? | PreEmptive*. Feb. 2021. URL: https://www.preemptive.com/obfuscation.

[52] Techslang". *What is Data Perturbation?* Nov. 2020. URL: https://www.techslang.com/definition/what-is-data-perturbation/.

[53] I. S. Rubinstein and W. Hartzog. "Anonymization and risk". In: *Wash. L. Rev.* 91 (2016), p. 703.

[54] P. Walker. *K – anonymity: An Introduction*. Jan. 2021. URL: https://www.privitar.com/blog/k-anonymity-an-introduction/.

[55] K. Rajendran, M. Jayabalan, and M. E. Rana. "A study on k-anonymity, l-diversity, and t-closeness techniques". In: *IJCSNS* 17.12 (2017), p. 172.

[56] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es.

[57] N. Li, T. Li, and S. Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. 2007, pp. 106–115.

[58] D. J. D. Cook. *Earth mover distance and t-closeness*. June 2020. URL: https://www.johndcook.com/blog/2018/11/01/t-closeness/.

[59] i-SCOOP". *Personal data pseudonymization: GDPR pseudonymization what and how*. Jan. 2021. URL: https://www.i-scoop.eu/gdpr/pseudonymization/.

[60] C. C. Aggarwal and S. Y. Philip. "A survey of randomization methods for privacy-preserving data mining". In: *Privacy-Preserving Data Mining*. Springer, 2008, pp. 137–156.

[61] C. Cadar, P. Akritidis, M. Costa, J.-P. Martin, and M. Castro. *Data randomization*. Tech. rep. Technical Report TR-2008-120, Microsoft Research, 2008. Cited on, 2008.

[62] O. Shaty, N. Hewitt, B. Bentley, and B. Lynch. *What is Data Masking? | Techniques Best Practices | Imperva*. Jan. 2021. URL: https://www.imperva.com/learn/data-security/data-masking/.

[63] G. Ravikumar, T. Manjunath, R. S. Hegadi, and I. Umesh. "A survey on recent trends, process and development in data masking for testing". In: *International Journal of Computer Science Issues (IJCSI)* 8.2 (2011), p. 535.

[64] T. A. Le, A. G. Baydin, R. Zinkov, and F. Wood. "Using synthetic data to train neural networks is model-based reasoning". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 3514–3521.

[65] M. Westerlund. "The emergence of deepfake technology: A review". In: *Technology Innovation Management Review* 9.11 (2019).

[66] A. Nguyen. *Understanding Differential Privacy - Towards Data Science*. July 2021. URL: https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a.

[67] A. Majeed and S. Lee. "Anonymization techniques for privacy preserving data publishing: A comprehensive survey". In: *IEEE Access* (2020).

[68] J.-M. Bohli, N. Gruschka, M. Jensen, L. L. Iacono, and N. Marnau. "Security and privacy-enhancing multicloud architectures". In: *IEEE Transactions on Dependable and Secure Computing* 10.4 (2013), pp. 212–224.

[69] J.-H. Hoepman. "Privacy design strategies". In: *IFIP International Information Security Conference*. Springer. 2014, pp. 446–459.

[70] J. H. Cheon, D. Kim, and D. Kim. "Efficient homomorphic comparison methods with optimal complexity". In: *International Conference on the Theory and Application of Cryptology and Information Security*. Springer. 2020, pp. 221–256.

[71] G. G. Fuster. "Inaccuracy as a privacy-enhancing tool". In: *Ethics and information technology* 12.1 (2010), pp. 87–95.

[72] J. Curzon, A. Almehmadi, and K. El-Khatib. "A survey of privacy enhancing technologies for smart cities". In: *Pervasive and Mobile Computing* 55 (2019), pp. 76–95.