UMEÅ UNIVERSITY

Privacy-Awareness in the Era of Big Data and Machine Learning

Xuan-Son Vu

Licenciate Thesis, September 2019
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

sonvx@cs.umu.se

ISBN 978-91-7855-110-1
ISSN 0348-0542
UMINF 19.06

Printed by Cityprint i Norr AB, Umeå, 2019

Abstract

Social Network Sites (SNS) such as Facebook and Twitter, have been playing a great role in our lives. On the one hand, they help connect people who would not otherwise be connected before. Many recent breakthroughs in AI such as facial recognition [49], were achieved thanks to the amount of available data on the Internet via SNS (hereafter Big Data). On the other hand, due to privacy concerns, many people have tried to avoid SNS to protect their privacy [83]. Similar to the security issue of the Internet protocol, Machine Learning (ML), as the core of AI, was not designed with privacy in mind. For instance, Support Vector Machines (SVMs) try to solve a quadratic optimization problem by deciding which instances of training dataset are support vectors. This means that the data of people involved in the training process will also be published within the SVM models. Thus, privacy guarantees must be applied to the worst-case outliers, and meanwhile data utilities have to be guaranteed.

For the above reasons, this thesis studies on: (1) how to construct data federation infrastructure with privacy guarantee in the big data era; (2) how to protect privacy while learning ML models with a good trade-off between data utilities and privacy. To the first point, we proposed different frameworks empowered by privacy-aware algorithms that satisfied the definition of differential privacy, which is the state-of-the-art privacy-guarantee algorithm by definition. Regarding (2), we proposed different neural network architectures to capture the sensitivities of user data, from which, the algorithm itself decides how much it should learn from user data to protect their privacy while achieves good performance for a downstream task. The current outcomes of the thesis are: (1) privacy-guarantee data federation infrastructure for data analysis on sensitive data; (2) privacy-guarantee algorithms for data sharing; (3) privacy-concern data analysis on social network data. The research methods used in this thesis include experiments on real-life social network dataset to evaluate aspects of proposed approaches.

Insights and outcomes from this thesis can be used by both academia and industry to provide privacy-guarantee data analysis and data sharing in personal data. They also have the potential to facilitate relevant research in privacy-aware representation learning and related evaluation methods.

Preface

This thesis contains a brief description of privacy-aware infrastructures, a discussion on improving privacy protection approaches in natural language processing, machine learning, and the following papers.

Paper I **Xuan-Son Vu**, Lili Jiang, Anders Brändström, Erik Elmroth. Personality-Based Knowledge Extraction for Privacy-preserving Data Analysis. *ACM, Proceedings of the Knowledge Capture Conference (K-CAP), 2017.*

Paper II **Xuan-Son Vu**, Addi Ait-Mlouk, Erik Elmroth, Lili Jiang. Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics. *ACM, Proceeding of WWW'19 - The World Wide Web Conference, 2019.*

Paper III **Xuan-Son Vu**, Lili Jiang. Self-adaptive Privacy Concern Detection for User-generated Content. *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing),* ***best student paper award****, 2018.*

Paper IV **Xuan-Son Vu**, Son N. Tran, Lili Jiang. dpUGC: Learn Differentially Private Representation for User Generated Contents. *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing),* ***3rd place for best paper awards****, 2019.*

Paper V **Xuan-Son Vu**, Abhishek Santra, Sharma Chakravarthy, Lili Jiang. Generic Multilayer Network Data Analysis with the Fusion of Content and Structure. *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2019.*

Other publications

The following publications were published during my PhD studies. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

1. **Xuan-Son Vu**, Lucie Flekova, Lili Jiang, Iryna Gurevych, Lexical-semantic resources: yet powerful resources for automatic personality classification, *In: Proceedings of the 9th Global WordNet Conference, January 2018*.

2. Son N. Tran, Qing Zhang, Anthony Nguyen, **Xuan-Son Vu**, Son Ngo, Improving Recurrent Neural Networks with Predictive Propagation for Sequence Labelling, *In: Proceedings of the 25th International Conference on Neural Information Processing (ICONIP-2018)*.

3. Thanh Vu, Dat Quoc Nguyen, **Xuan-Son Vu**, Dai Quoc Nguyen, Michael Catt, Michael Trenell, NIHRIO at SemEval-2018 Task 3: A Simple and Accurate Neural Network Model for Irony Detection in Twitter, *In: Proceedings of Proceedings of NAACL-HTL'18, at the 12nd International Workshop on Semantic Evaluation (SemEval-2018)*.

Financial support for this work is provided by the Umeå University for the project Privacy-aware Data Federation under project no. 570066000, and the Swedish Foundation for International Cooperation in Research and Higher Education (UHR).

Acknowledgements

I first would like to thank my supervisor, Lili Jiang, for her consistent and immense support throughout my PhD studies. All opportunities for improving our research were only possible because of her advice, patient supervision, and thoughtful insights. To be here, please make sure you remember her name, she is awesome physically (running) and mentally (supervision) :).

Secondly, I would like to thank my co-supervisor Erik Elmroth (a.k.a the Big Boss) for giving me general insights on how to look at different angles of research problems, educational opportunities and directions on how important it is to design a research problem before "getting hands dirty". His wise and aimed view on important aspects of research made our works thrive, going beyond our limits.

One of the neat opportunities offered to me by my supervisors was to work abroad at the ITLab, Texas. And for this, I thank Sharma and Abhishek for their insights and discussions in many graph-based problems.

I also would like to thank the very great researchers, friends, and collaborators at the Department of Computing Science, HPC2N, and UMIT Lab, in arbitrary order, including Eddie, Chanh, Mahmoud, Monowar, Juan Carlos, Angelika, Birgitte, Ahmad, Mats, Thang, Anne-Lie, Mikael Hansson, Carl Christian, Mirko, Monicar, Timotheous, Suna, Michele, Abel, Jakub, Maitreyee, Thomas, Addi, Frank, Jakub, and many others. Special thanks to Tomas Forsman for being the magical technical guru, for which I am very grateful.

Last but not least, to all my friends, family and country (Vietnam), who helped me getting here. Especially, to my lovely wife and my son (Anh and Aaron), who have been always unconditionally supported me in life. To my parents (Thao and Loan), who endlessly support me on the way I go. To my brothers (Quy and Duong) and their families, who always believed in me and supported me spiritually throughout writing this thesis. To my collaborators (Thanh Vu, and Son N. Tran) for their excellent insights in research and how to have a successful PhD life. No words can be enough to say how thankful I am when having supports from these people.

/Xuan-Son Vu

Abbreviations

Table 1: List of terminologies and abbreviations used in the thesis.

#	Term/Abbreviation	Explanation
1	All Data	All available data of the world
2	Big Data	Refers to the "5V's" Big Data of [87].
3	UGC	User Generated Content [92].
4	DF	Data Federation [93, 90]
5	DA	Data Analysis [90]
6	DP	Differential Privacy [19, 26]
7	DS	Data Sharing [94]
8	ML	Machine Learning [63]
9	MLN	Multi-layer Network [91]
10	NA	Network Analysis [91]
11	SVM	Support Vector Machines [18]
12	GDPR	General Data Protection Regulation

Contents

Chapter 1

Introduction

My PhD studies focus on research in "Privacy-aware Data Federation", which aims at virtually integrating heterogeneous data from multiple distributed sources and preserving privacy during data federation and data analysis. According to a recent study, the volume of corporate data doubles each year and the public Web grows by over seven million pages a day. Facebook, a social networking site alone, back in 2016, was generating 25TB of new data every day [59]. And the daily amount of data Facebook is creating now, in 2019, is 4PB per day, which is $\sim$ 166 times more than that in 2016, according to a recent statistic [†]. The vastly increasing volume of data is generated and/or collected by people across organizations (e.g., governments, academic institutions, business corporations, web users) for different purposes, in different schema, and using different methodologies. This imposes the requirements on the technology for effective data integration and data sharing across multiple heterogeneous sources. Among these increasing volume of data, individual personal data can be largely collected and analyzed to understand important phenomena, such as early detection of diseases [40] and social service recommendation [21]. However, user concerns rise from a privacy perspective, with sharing an increasing amount of information regarding their profile information, health, service usage and activities. Thus, it is critical to developing techniques to enable data federation and data sharing without losing privacy.

1.1 What is Privacy?

Before we start to discuss related problems in privacy, it is important to understand privacy and in what scenario privacy is violated.

Privacy. Many legal systems protect a right to privacy. However, 'privacy' remains an elusive and controversial concept [6]. In the Privacy book [6], Barendt addressed that "some writers have rejected the idea that there is a

[†] `www.visualcapitalist.com/how-much-data-is-generated-each-day/`

discrete right to privacy. In their view, it is derivative from well-established rights, such as property rights and personal rights not to be touched or observed without consent, and it would be possible to dispense with it as a distinct right.". According to this view, hypothetically, if a person intruded into a house and took a photo of two intimate couples, the intruder only got a *break-in crime* but not for something else (e.g., violated privacy space, harassment). Also in the book, Barendt mentioned that "other writers do not share this skepticism, but disagree about the value of privacy or, put another way, over the justifications for protecting it by law or under a constitution" [6]. From the above discussions, we understand that privacy is a complex topic and it has been gone through many different generations to have agreements (e.g., GDPR). Generally, privacy can be preserved in three ways (i.e., norm, law, technology). Since this thesis focuses more on the technical solutions to protect privacy of individuals according to the current law (i.e., GDPR), we do not attempt to make a clear definition of privacy. We, however, chose to refer privacy as another synonym called 'the right to be let alone' [6], which is termed in the GDPR regulation as "the right to be forgotten" [††]. It means that the data subject can "obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay". Moreover, privacy is also about "information privacy", which is the privilege to have some control over how personal information is collected and used [42]. It is "the capacity of an individual or group to stop information about themselves from becoming known to people other than those they give the information to" [42]. Briefly, in this thesis, we focus on two problems of privacy: (1) the right to be forgotten; and (2) protect re-identification problems of personal data. To the former, in our proposed frameworks, we can keep track of user data and hence, be able to fulfill user's requests to erase their data. About re-identification, we proposed both systematic architectures and privacy-guarantee algorithms to protect re-identification problems. However, in order to be sure of what kind of data that needs protection, we need to understand 'what are personal data?'.

What are 'personal data'? Any data-protection law will also need to define the concept of 'personal data' or 'personal information'. In the article 2(a) of the European Union Directive employs the following information:

> "*Personal data* means any information relating to an identified or identifiable individual natural person ('data subject'); an identifiable individual is one who can be identified directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity."

Or newly in GDPR's article 4:

> "*Personal data* means any information relating to an identified or

[††] gdpr.eu/right-to-be-forgotten/

> identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

Even though the European Union Directive did not mention about biometric data, however, the new GDPR's regulation defines biometric as:

> "Personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allows or confirms the unique identification of that natural person, such as facial images or dactyloscopic data."

Therefore, in the new GDPR's regulation, data subjects are being protected from re-identification problems of not only their direct information (e.g., social security number) but also from biometric data that can re-identify data subjects from specific technical processing (e.g., user behaviors, facial images).

When privacy is violated? Cynthia Dwork [19], who introduced *differential privacy* - the current state-of-the-art privacy-guarantee approach by definition, has re-introduced the desideratum for statistical databases by Dalenius [20]: access to a statistical database should not enable one to learn anything about an individual that could not be learned without access [†]. Intuitively, the definition requires that any algorithms outputting information about an underlying dataset are robust to any change of one sample, thus protecting privacy. We will explore more about *differential privacy* later in next sections. In a discrete way, Katal et al. [46] list that privacy may be breached under following (but not all) circumstances:

- Personal information when combined with external datasets may lead to the inference of new facts about the users. Those facts may be secretive and not supposed to be revealed to others.
- Personal information is sometimes collected and used to add value to business. For example, individual's shopping habits may reveal a lot of personal information.
- The sensitive data are stored and processed in a location not secured properly and data leakage may occur during storage and processing phases.

At the current scope of this thesis, our contributions lie more on the first and the second circumstances to avoid privacy breaches. The third circumstance requires more work in security that we might investigate in future work.

[†]Semantic security against an eavesdropper says that nothing can be learned about a plaintext from the ciphertext that could not be learned without seeing the ciphertext.

Privacy vs. security. In many cases, people got confused between privacy and security since they normally appear together in the main tracks of top journals or conferences in computer science. However, they are not the same. Data privacy is focused on the use and governance of individual data (e.g., setting up policies in place to ensure that consumers' personal information is being collected, shared and utilized in appropriate ways). Security concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit [14]. While security is fundamental for protecting data, it is not sufficient for addressing privacy. Table 1.1 shows some (but not all) differences between privacy and security.

Until here, we have covered the nature of privacy and what issues related to privacy we should pay attention to. We summarize here some main points that this thesis tries to focus:

- First, when talking about privacy, we want to protect data subjects from two problems: (1) re-identification, (2) assure the right to be forgotten.
- Secondly, privacy breaches can happen in many different ways, however, in this thesis, we focus on (1) re-identification problem when user data are being collected and shared; (2) allowing data processing (e.g., data analysis, learning models, etc.) on sensitive data without privacy leakages.
- Thirdly, privacy and security are not always the same. Security can be used to strengthen privacy protection but security is not the direct solution to privacy protection.

Table 1.1: Difference between privacy and security [42]

#	Privacy	Security
1	Privacy is the appropriate use of user's information	Security is the "confidentiality, integrity and availability" of data
2	Privacy is the ability to decide what information of an individual goes where	Security offers the ability to be confident that decisions are respected
3	The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	Security may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency [37]

1.2 Research Problems and Objectives

Concerning the above challenges, my PhD research especially focuses on the following two aspects: (1) data federation across heterogeneous and distributed

data and (2) privacy preservation on data federation and data processing, which includes data analysis, data sharing, etc. Especially, data federation methodologies empower a unified interface over these heterogeneous data sources by virtual integration. This approach provides users a perspective that allows data of multiple heterogeneous sources can be queried transparently and quickly. Privacy preservation methodologies achieve the balance between data utility and individual sensitive data protection to build a privacy-aware web ecosystem.

Here are the main research objectives of this thesis:

RO1 (Data Federation): to address the heterogeneity and distribution of multiple data sources, as well as the sensitivity of register data and web data:

- Objective RO1a: to develop collaborative query-based data federation algorithms that cope with data distribution and provide users a unified interface to quickly access multiple data sources.
- Objective RO1b: to address how the privacy-guarantee frameworks can support robust research topics including privacy-concern data analysis on register data with scalability and high performance by design and test on a real-world dataset.

RO2 (Privacy Preservation): to show the frameworks' effectiveness and cross-disciplinary utility regarding privacy in data processing including:

- Objective RO2a: to apply differential privacy solution on register data for privacy-aware detection (e.g., privacy-concerns detection).
- Objective RO2b: to develop differential privacy algorithms that prevent adversarial attackers run inferences on personal data using big data to find privacy leakage.
- Objective RO2c: to collaboratively apply the privacy-aware data federation solution on social network for privacy-guarantee data sharing and social network analysis.

Generally, objective **RO1** targets at finding different system architectures to re-construct personal data from which, they support the objective **RO2** to develop different privacy-aware algorithms to protect data privacy.

1.3 Methodology and Research Contributions

In this section, we answer two questions including: (1) how we address research objectives? and (2) what is the main contributions of this thesis? Both questions are important since *methodology* mainly shows what is the main research directions we could follow to achieve the research objectives, and *research contributions* summarizes our achievements in this thesis.

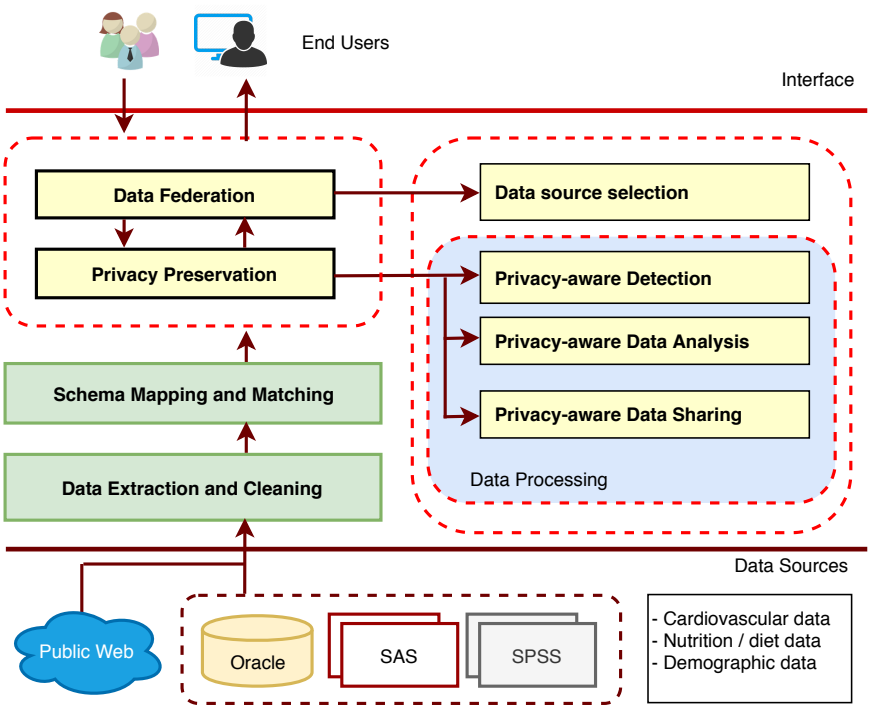

Figure 1.1: Architecture Design of Privacy-aware Data Federation Frameworks, where the red parts are the main focuses of this thesis.

A. Methodology

To meet the research objectives of this thesis, we investigated different research topics in both (1) system architectures regarding data federation and (2) privacy-guarantee algorithms for data processing. Regarding system architectures, we focus on this topic because of two important facts. Firstly, it is because of the big gap between theories and real-life applications of privacy-aware data federation systems. There are some well-established frameworks such as PINQ [58] or GUPT [66], however, they mainly act as out-of-the-box solutions to traditional database systems to achieve privacy protection but not for data federation system. Secondly, it lacks of practical privacy-aware algorithms in real-life system, which can be used by end-users (e.g., a psychology researcher), that can efficiently address privacy issues to protect personal data. In [43], the authors have summarized different works in privacy and the majority of them are privacy-aware algorithms, which we will discuss in more detail in the following parts. Moreover, there exists research work on privacy preservation for register data [29, 1, 75, 72, 97], however, they either try to (1) address privacy issues on image datasets [29, 1, 72, 97] (because images are easier to add noise than heterogeneous data, therefore, it is easier to protect privacy) [94] or (2) address privacy issues on some selected properties [102] but not for centralized data collection. These limitations are addressed in more

detail in **Paper IV** [94].

By analyzing the distributed heterogeneous data across multiple data sources including open web data and register data, we propose privacy-aware data federation framework as shown in Figure 1.1, where the red parts are the focus of my PhD studies. From the aspect of research, we mainly address the academic challenges in data federation and privacy preservation. From the application point of view, this project solves real challenges in privacy issues on social network data (e.g., Facebook).

B. Data Federation

The main challenges in processing federated database queries originate from the data distribution, heterogeneity and autonomy. We construct our federation infrastructure by firstly deploying the well-known data federation framework Teiid [86, 81], based on which, we address the following scientific problem:

- Data source selection: given a natural language query, the system has to figure out which variables from which data sources are involved in the query analysis in order to find the answer. To address this problem, we proposed a rule-based approach to find related variables on a virtual database from which the system selects correct variables of the original data source for data analysis. In the paper I [93] and paper II [90], we applied this approach to build open-access frameworks allowing researchers to work on register data that would otherwise is not easy to access and perform data analysis.

C. Privacy Preservation

The module of privacy preservation focuses on balancing the needs of the researchers to pursue scientific research as well as the privacy of individuals in their datasets.

- Privacy-aware Detection: to detect how much privacy-guarantee should the system protect user data to balance the trade-off between privacy protection and data utility. It is important to mention that in many datasets, we have no way to ask data subjects for their privacy-concerns (e.g., a dataset was collected 100 years ago and most data subjects had died; or similarly, a dataset was collected anonymously and there is no way to contact the data subjects). Additionally, for data analysts, who want to guarantee privacy protection for their analytic results, it is not straightforward for them to define privacy-guarantee level. This happens because the proper distribution of the limited privacy budget across multiple computations require significant mathematical expertise [66].

- Privacy-aware Data Analysis: Any outputs from a data analysis running on personal data should guarantee privacy. Therefore, this module

assures analytic outputs of the system are guaranteed under privacy protection mechanisms (e.g., differential privacy [19]).

- Privacy-aware Data Sharing: to protect privacy of a dataset before sharing to other third-parties. There have been different privacy-guarantee algorithms for data sharing such as K-anonymity [79], L-diversity [56], t-Closeness [55]. However, most of them are vulnerable to privacy attacks which will be discussed further in this section.

Privacy Preservation Methods. Privacy protection can be divided into two methods namely (1) data sanitization and (2) anonymization. In 2008, Narayanan and Shmatikov [68] proposed an effective de-anonymized algorithm to break privacy of Netflix Prize Contest [69]. In 2009, there was a very subtle privacy violation when Wang et al. [96] showed how published GWAS (Genome-Wide Association Study) results [31] revealed whether specific individuals from the study were in cancer group or healthy group. Since then, researchers have been focusing more on data sanitization approach to protect privacy. Data sanitization process commonly can be performed in 4 different ways (see figure 1.2) including (1) input perturbation [10], (2) output perturbation [24], (3) internal/objective perturbation [95, 15], and (4) sample-and-aggregate [71]. In 2006, Dwork firstly introduced differential privacy (DP) in her ICALP paper [19] to capture the increased risk to one's privacy incurred by participating in a database. It seeks to provide rigorous, statistical guarantees against what an adversary can infer from learning the results of some randomized algorithms. Thus, most of DP algorithms are not categorized in the first privacy protection approach (i.e., input perturbation) but in the other three approaches [58, 66]. Input perturbation is more common in data curation [27].

Differential privacy [19] aims to provide statistical guarantees against what an adversary can infer from observing the results of some randomized algorithms such as recommendation algorithm or personalized search engine. For any data analysis system, there are two essential modules including (1) data manager, and (2) data analyzer. Practically, these two modules are designed in such a way that allows the analysis module performs locally so all data stays at source, within the governance structure and control of the originating data. As a typical example, DataSHIELD [38] is a system that is implemented following this approach. It can be used to run analysis of individual-level data from multiple studies or sources without physically transferring or sharing data and without providing any direct access to individual-level data [65]. However, DataSHIELD only can answer single one-dimensional statistics, which is not always satisfied researchers' needs. In fact, user query might either range from numeric to non-numeric query or from one-dimensional to multi-dimensional statistic query. Thus, our goal is applying differential privacy to fulfill user needs of flexible query types.

Typically, DP methods reduce the granularity of representation in order to protect confidentiality. There is, however, a natural trade-off between information loss and the confidentiality protection because this reduction in granularity

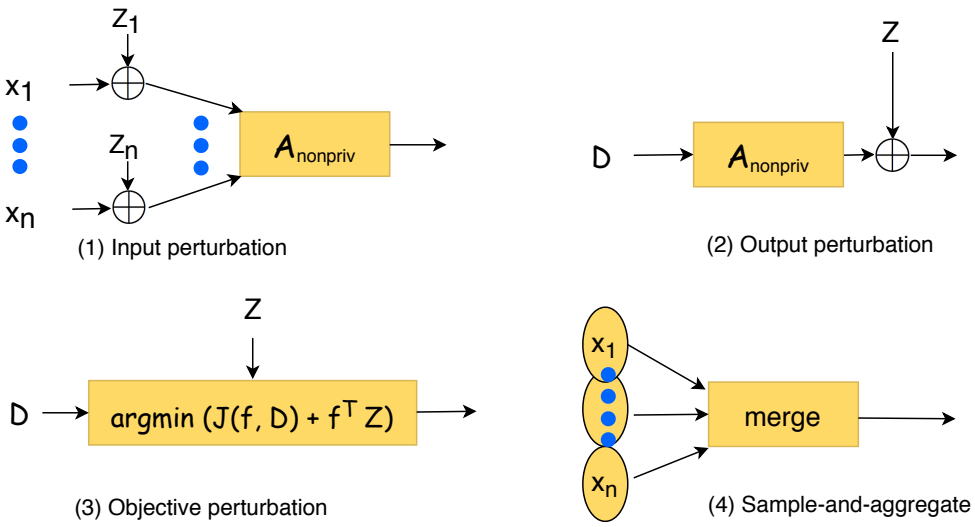

Figure 1.2: Common privacy protection approaches: (1) *input perturbation* adds noise to the input before running algorithm; (2) *output perturbation* runs algorithm then adding noise to the results; (3) the *internal/objective perturbation* randomizes the internals of algorithm, and (4) the *sample-and-aggregate* computes query on disjoint subsets data and then use differentially private method to select max.

results in diminished accuracy and utility of the data, and methods used in their analysis. To measure this trade-off, we often apply learning-algorithms on both of the raw data and privacy-aware data. It is different from regular learning algorithms in the sense that training data is no longer the original data. It has been modified in such a way that there is no trace back to know where is the data come from or who is a particular participant. For instance, putting some random noises on user responses to guarantee user privacy is one of such methods (e.g., Erlingsson et al. [27]). And this modification makes the learning part be more difficult due to the privacy-guarantee modification. Ji et al. addressed in [43] that general ideas of privacy-preserving machine learning algorithms are learning a model on clean data, then use exponential mechanism [88, 62, 78] or Laplace mechanism [89, 16] to generate a noisy model. However, due to privacy issue, raw data is no longer available but sanitized data. Because of this reason, how to evaluate privacy-guarantee models in comparison to no-privacy guarantee models is a big challenge. In section 3.3 of Chapter 3 we also address some evaluation problems of privacy-guarantee machine learning models.

D. Research Contributions

This thesis contributes to knowledge in privacy-aware (1) data federation, (2) data analysis, and (3) data sharing within the context of the research objectives. Figure 1.3 and Figure 1.4 show the overview of research contributions of this thesis.

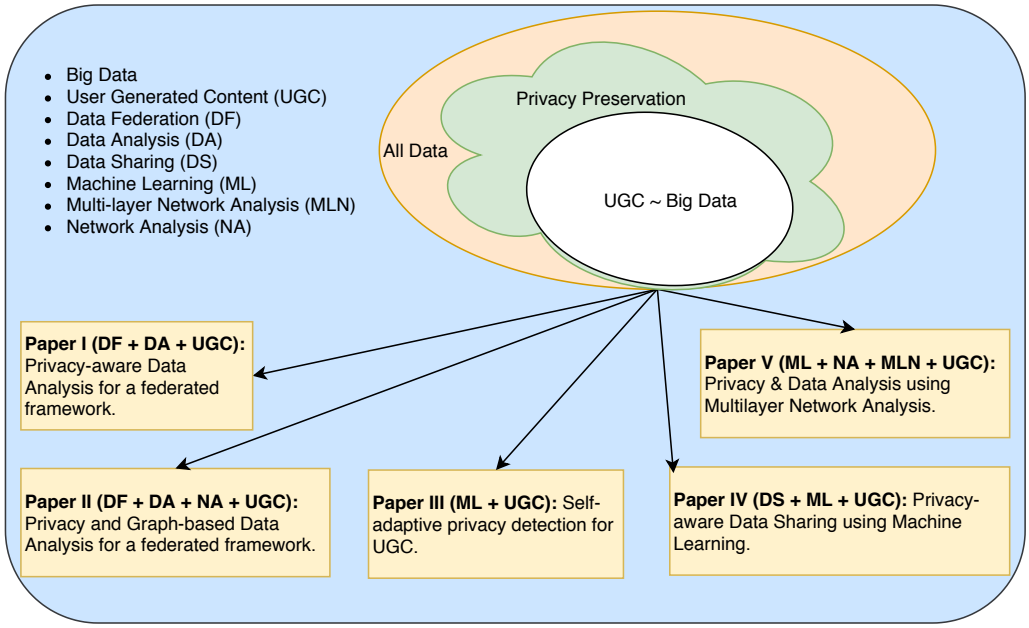

Figure 1.3: Overview of the five papers and its connections to Privacy-Aware Data Federation and UGC (or Big Data).

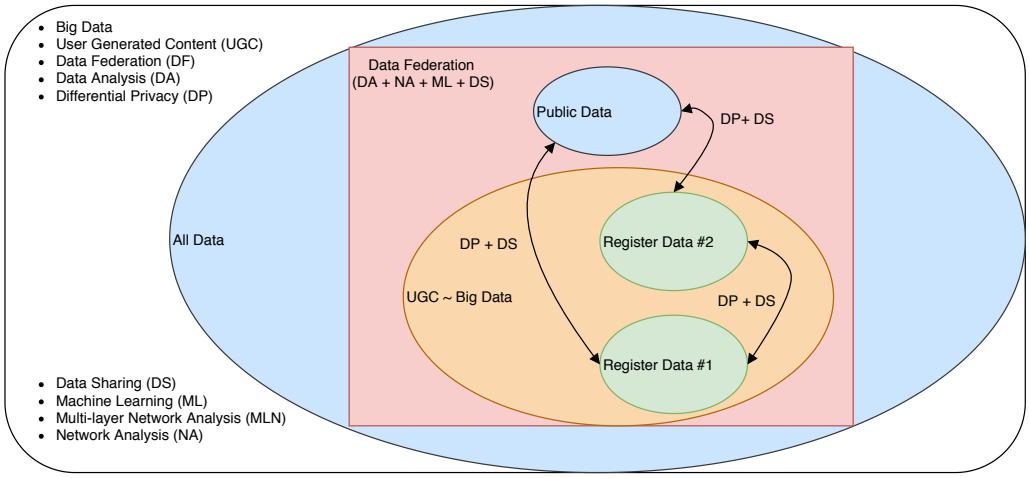

Figure 1.4: Overview of Data Federation Infrastructures with respects to different type of data and features it can support in this thesis.

To address **RO1**: we proposed different system architectures to both guarantee privacy and have a good trade-off between privacy and data utility in **Paper I** [93] (RO1a) and **Paper II** [90] (RO1b).

To address **RO2**: we proposed different privacy-aware algorithms to address different privacy issues in personal data. Firstly, because of the fact that many datasets were collected a long time ago, and we have no way to ask privacy concerns of the data subjects, whose data were collected. Hence, **Paper III** [92] (RO2a) proposed a solution to detect privacy based on the contents of the collected data themselves. Secondly, to prevent running inference attacks on personal data (i.e., RO2b), **Paper III** [92] and **Paper V** [91], proposed methods to limit how much information can a random algorithm learn from the data to satisfy the definition of differential privacy. Lastly, **Paper IV** [94], **Paper V** [91] answer the question of **RO2c** on how to apply the privacy-aware data federation solution on social network for privacy-guarantee data sharing **Paper IV** [94] and social network analysis **Paper V** [91].

In summary, main contributions of this thesis in the five papers are:

- Propose system architectures of open-access frameworks for data federation and data analysis that allowing researchers to work on register data faster with privacy-guarantee analytic results.
- Propose privacy-aware algorithms to balance the trade-off between data privacy and data utility.
- Propose privacy-aware data sharing for learning representations and share them safely for public usage without the necessity to share the raw data.

1.4 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 presents a comprehensive background of privacy-aware infrastructures on different topics including data federation, data sharing and data analysis. Some challenges in privacy-aware infrastructures are also mentioned in this chapter. Chapter 3 discusses privacy issues in machine learning models where privacy must be guaranteed at the worst-case outliers and thus data utilities will also be affected severely. Chapter 4 summarizes research papers included in the thesis; followed by future work beyond this thesis.

Chapter 2

Privacy-Aware Infrastructures

In this chapter we describe the role of privacy-aware data federation framework, which is the software system that manages the multiple data sources and analytic application in a federated manner.

2.1 Privacy-aware in Big Data

Big data [48] is a term used for very large data sets that have more varied and complex structure. It specifically refers to data sets that are so large or complex that traditional data processing applications are not sufficient. Big data is compared to a double-edged sword. Because of Big Data, people are not easy to be "forgotten" as one of the fundamental policy stated in the GDPR regulations [*]. However, taking the advantages of Big Data, it can help businesses and organizations to improve internal decision making power and can create new opportunities through data analysis [59]. It can also solve big problems of society like in healthcare (e.g., disease forecasts [40], quantifying mental healths [17]). It can also help to promote the scientific research and economy [59]. Despite the benefits we can achieve from using big data to understand the world in various aspects of human endeavors, it faces many risks regarding privacy such as the incidents of Cambridge Analytical [†], AOL search data leak [‡], or Netflix Prize Contest [§]. Therefore, to balance the trade-off of both data privacy and the benefit of Big Data, many studies are focusing on this direction to address this new challenge [19, 26, 93, 91, 94]. To name a few, in order to ensure big data privacy, various mechanisms have been developed in recent years including K-

[*] gdpr.eu/right-to-be-forgotten/

[†] en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

[‡] en.wikipedia.org/wiki/AOL_search_data_leak

[§] en.wikipedia.org/wiki/Netflix_Prize

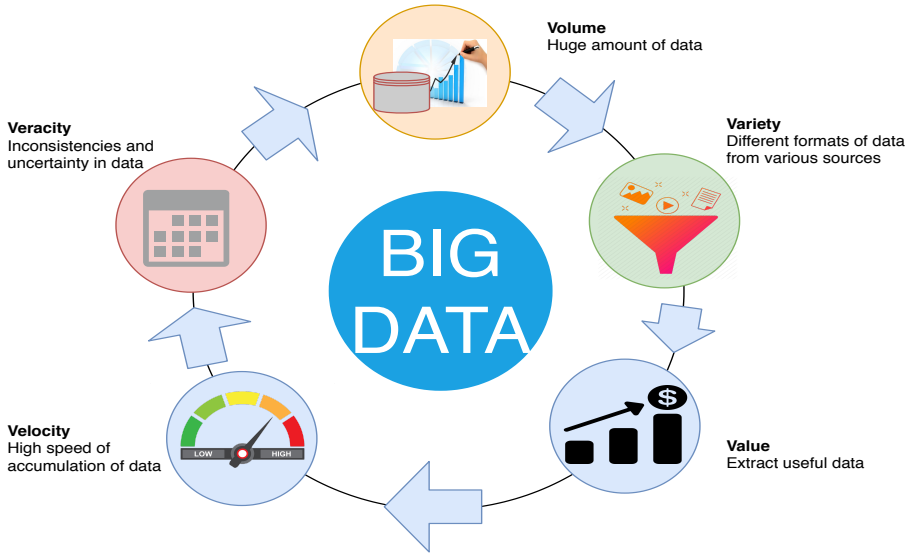

Figure 2.1: The properties of big data are reflected by 5V's, which are veracity, validity, value, variability, venue, vocabulary, and vagueness [87].

Anonymity [79], L-diversity [56], t-Closeness [55], and differential privacy [19, 26]. In general, these mechanisms can be grouped based on the stages of big data life cycle [59], i.e., data generation, storage, and processing.

1. **Data Generation**: Data can be generated from various distributed sources [59]. Privacy research topics in relation to this process are access restriction [100] and falsifying data [100].

2. **Data Storage**: storing big data securely is very challenging since it involves many parties during the process (e.g., data provider, data warehouse manager). Therefore, we need to ensure that the stored data are protected against threats such as direct attack to data centers, misconduct of the direct data manager etc. Among conventional mechanisms to protect data security [13] and privacy [82], one promising technology to address these requirements is storage virtualization, enabled by the emerging cloud computing paradigm [60].

3. **Data Processing**: it refers to any processes running on data including data transformation, data analysis, data sharing, etc. Since privacy regarding the data processing part is the main topic of this thesis, they were being reviewed in detail in the subsection 1.3.

2.1.1 Privacy-aware Data Federation

In order to analyze harmonized data across different sources, there are three general approaches: pooled data analysis, summary data meta-analysis, and federated data analysis [39]. The first two approaches, pooling individual-level

data in a central location and meta-analyzing summary data from participating studies, are commonly used in multi-center research projects. However, these two approaches require data to be transferred to central servers which is the main risk of privacy leakage. The third approach is the focus of my PhD studies, which co-analyzes harmonized data across multiple sources by performing federated analysis of geographically-dispersed datasets.

Data federation, a form of data virtualization, is a process whereby data is collected from distinct databases without ever copying or transferring the original data itself [†]. Data federation creates a single repository that does not contain the data itself, rather its metadata. A widely mentioned technology is data integration, where the data could be copied from each individual data sources. Therefore data integration contains data federation.

2.1.2 Privacy-aware Data Sharing

The purpose of privacy-aware data sharing is to avoid privacy leakage after publishing data for third parties. On the one hand, it must hide information about data subjects. On the other hand, for the released data to be useful, it should be possible to learn something significant. Several research areas are related to this problem. Each makes different assumptions and has different constraints. In most cases, it involves research in micro-data anonymization since this type of data contains much identifiable personal information. This area focuses on efficiently and effectively anonymizing data in a very small (micro) dataset by altering the content of the dataset to make it impossible to identify a specific individual in the dataset. K-anonymity [79] was one of the most popular methods and various different algorithms implement this technique such as [85, 56]. Some anonymization algorithms perform well on any given micro-dataset regardless of the content or use of that micro-dataset. The techniques use generalization and suppression [85]. Some studies (e.g., LeFevre et al. [53]) propose algorithms that support the generation of anonymous views based on a specific work-load focus. The others (e.g., Xiong [98]) proposed a top-down priority scheme for anonymization; this allows a priority to be assigned to some set of Quasi-Identifiers to minimize the perturbation on those specific fields. Bhumiratana and Bishop [9] proposed a different method, which they called an "orthogonal approach" to these two directions. They proposed a framework to balance privacy and data utility. It provides a formal, automatic communication between a data collector and a data user to negotiate on what level should they agree on privacy protection while maintaining good data utility. It is really a huge amount of work in data anonymization that this thesis cannot cover all of them. However, it is worth to mention that, micro-data is not only the sensitive data (e.g., user text data) because in the big data era, data can be linked to many side datasets that make anonymization methods be vulnerable to privacy leakages.

[†] `https://en.wikipedia.org/wiki/Federated_database_system`

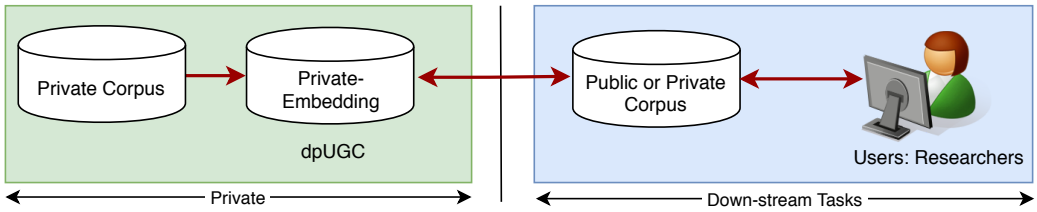

Figure 2.2: Overview of our safe-to-share embedding model that can be used to facilitate research on sensitive data with privacy-guarantee.

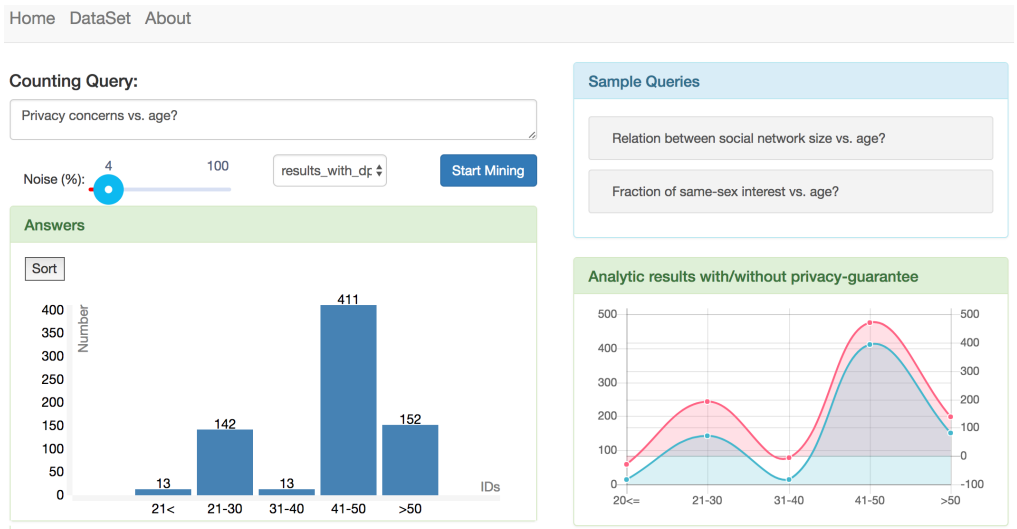

Figure 2.3: An example of privacy-guarantee histogram (the red line) in **Paper I** [93].

With the recent advancements in Deep Learning, privacy-aware data sharing now can be much more different. Since deep learning is about learning representations, it can be used for data publishing by sharing the data representations instead of the raw data. Figure 2.2 shows a high-level overview of data sharing with privacy-guarantee in the **Paper IV**.

2.1.3 Privacy-aware Data Analysis

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. The target of *privacy-aware data analysis* is to protect privacy of individuals in the analyzed data. It means that any analytic results from the process cannot be used to re-identify any individuals from the data. Figure 2.3 presents an example showing the difference between privacy-guarantee histogram versus raw histogram. While the main trend of the statistic is the same, the privacy-guarantee histogram protects the chance

to re-identify any individual in the result. Naturally, you might have this question in your head: "A histogram only shows statistics of a population, how come can it reveal privacy breaches?". However, that is not always the case. If it happens to show only one person in the category "<21" (less than 21 years old), and by side information, an adversary knows that there is only one boy in the dataset is less than 21 years old. Then the information that the adversary knows about an individual before and after seeing the histogram is different. This means, according to the privacy definition of [19], the histogram causes a privacy breach of an individual (i.e., the boy). To avoid this situation, a privacy-guaranteed histogram is already considered the most sensitive case (e.g., only one individual in a category), then it will add noise to the histogram to mask the chance to re-identify any other information.

According to Dwork [19], there are two natural models for privacy mechanisms in data analysis: interactive and non-interactive. In the non-interactive setting the data collector, a trusted entity, publishes a "sanitized" version of the collected data; the literature uses terms such as "anonymization" and "de-identification". Traditionally, sanitization employs techniques such as data perturbation and sub-sampling, as well as re-moving well-known identifiers such as names, birth dates, and social security numbers. It may also include releasing various types of synopses and statistics. In the interactive setting the data collector, again trusted, provides an interface through which users may pose queries about the data, and get (possibly noisy) answers. For the non-interactive setting, it might be easier to protect privacy since all queries are given in advanced and there is time for calculating privacy-guarantee results. However, this setting is very time consuming for both *data processor* (i.e., the party has the control over data) and researchers, who want to analyze the data. Therefore, the second one - i.e, interactive setting, is more favored but it is more challenging. Because the analytic process is interactive, it is difficult to prevent an adversary from running inferences based on outputs to find internal settings (e.g., amount of noise) of the system. From knowing the internal settings, the adversary can reverse the noisy outputs to get the original results.

2.2 Challenges

This part discusses challenges in protecting privacy for heterogeneous and distributed data and also how to effectively scale the federated system with privacy-guarantee is a big research topic.

2.2.1 Heterogeneous and Distributed Data

As the volume of data is increasing and more open data is promoted, it is extremely difficult to predict the potential risk for individual privacy leakage. Also, there are various different types of data (e.g., text, speech, video, network etc.) located differently in many locations, therefore, effectively protect privacy

of individuals is a big challenge. Here we list some main challenges regarding privacy for heterogeneous and distributed data:

- **Privacy-aware for edge computing**: to avoid latency between a user action and a server response, many service providers have deployed edge computing to distribute jobs to edge nodes. Thanks to this architecture, it reduces the computation pressure of the data center. As a result, user data now is distributed in many edge nodes. However, some edge nodes with poor security preserving may become the fuse of the intruder's malicious attack [23].

- **Privacy-aware for learning representations**: because of heterogeneous and distributed data, it is a big challenge to effectively learn a good representation for a given user data. For example, user A has text data distributed at a data center D_1, image data at D_2, audio data at D_3. And due to privacy issues, for any user representation coming out from a data center, it has to be a privacy-guarantee representation. Due to this reason, it is a big challenge to compute a single representation for user A given noisy representations from different data centers D_1, D_2, D_3.

- **Privacy-aware for federated learning**: similarly to the above challenge, in federated learning [57], there is a local model stays at the same location to learn from user data. However, because the data located in each location is incomplete, the model has a very little information to contribute to the global model for improving at some tasks at user level (e.g., user profiling task for recommendation). Thus, how to effectively monitor the noise in federated learning so that when they are being aggregated with each other at the global model, information from the same user can be aggregated with less noise. At that point, the aggregated information at the global model will be more valuable to be used for other tasks (e..g, recommendation).

2.2.2 Scalability Problems

Given the fact that federated system allows data to be located differently in many locations, however, how to perform high-performance data analysis on Big Data is a big question. In **Paper II** [90] we already proposed to use Elastic Search system to perform high performance data analytics. However, the Indexing system was not federated since it requires more work to federate all indexing systems in different locations and aggregate analytic results across all indexing systems. In future work, we also plan to address this issue to fulfill the requirement of high-performance data analysis.

Chapter 3

Privacy-Aware Machine Learning

In this chapter we talk briefly about Machine Learning, from which, we address related problems in learning privacy-guarantee representations for UGC (i.e., the representation of Big Data).

3.1 A Brief Introduction to Machine Learning

Machine Learning itself is a big topic and this thesis cannot go too much in details. However, we want to gently go over some of main ideas in Machine Learning that lead to privacy issues.

What is Machine Learning? The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest. There is no formal definition of machine learning, however, the most widely used definition is that of CMU † Professor Tom Mitchell [64]:

> "A computer program is said to learn from experience E, with respect to some class of tasks T and performance measure P if its performance at tasks in T as measured by P improves with experience E."

Intuitively, the definition means that a computer program can learn to improve performance measured by P at some tasks T through experience E. For example, if we say, **Pepper** - a robot, has the ability to learn how to clean a house. Then we need to show that **Pepper** can perform a task $t \subset T$ (i.e., clean the house) by exploring all corners in the house after some times (i.e., experience E). If the performance P in this task is the cleaning time, then

†Carnegie Mellon University

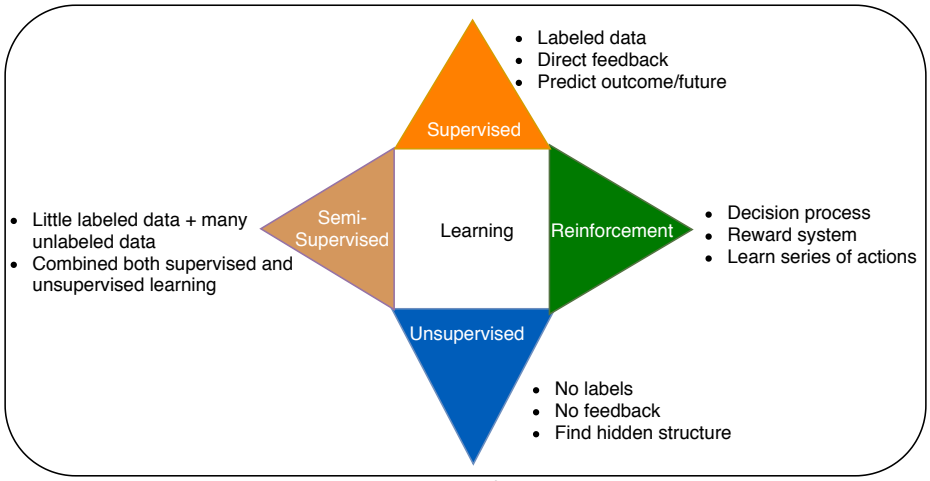

Figure 3.1: Four popular types of Machine Learning paradigms.

p_{i+1} has to be smaller than p_i, where $\{p_i, p_{i+1}\} \subset P$ are the cleaning time of **Pepper** at experience $\{e_i, e_{i+1}\} \subset E$, respectively. In other words, **Pepper** learned how to clean the house more efficient after some experiences. From the definition and the example, we understand that a machine learning model has to improve its performance through experiences. There are different learning paradigms in ML and among them, there are four basic paradigms shown in Figure 3.1 are briefly summarized as follows:

1. **Unsupervised learning models**: experience a dataset containing many features, then learn useful properties of the structure of this dataset. In the context of deep learning, which is the subset of ML, we usually want to learn the entire probability distribution that generated a dataset. Some other unsupervised learning algorithms perform other roles, like clustering, which consists of dividing the dataset into clusters of similar examples.

2. **Supervised learning models**: experience a dataset containing features, but each example is also associated with a label or target. For example, we can teach **Pepper** to differentiate between obstacles and empty space inside a house by training point-and-shoot cameras to classify millions of images labeled with 0 (for empty space) and 1 (obstacles). Based on the trained models, at the deployment phase, **Pepper** will be able to avoid obstacles by classifying surrounding images.

3. **Semi-supervised learning models**: combine from both supervised and unsupervised models to perform better than a singular paradigm at specific tasks. Semi-supervised learning may refer to either transductive learning or inductive learning [103].

4. **Reinforcement learning models**: interact with an environment, so there is a feedback loop between the learning system and its experiences.

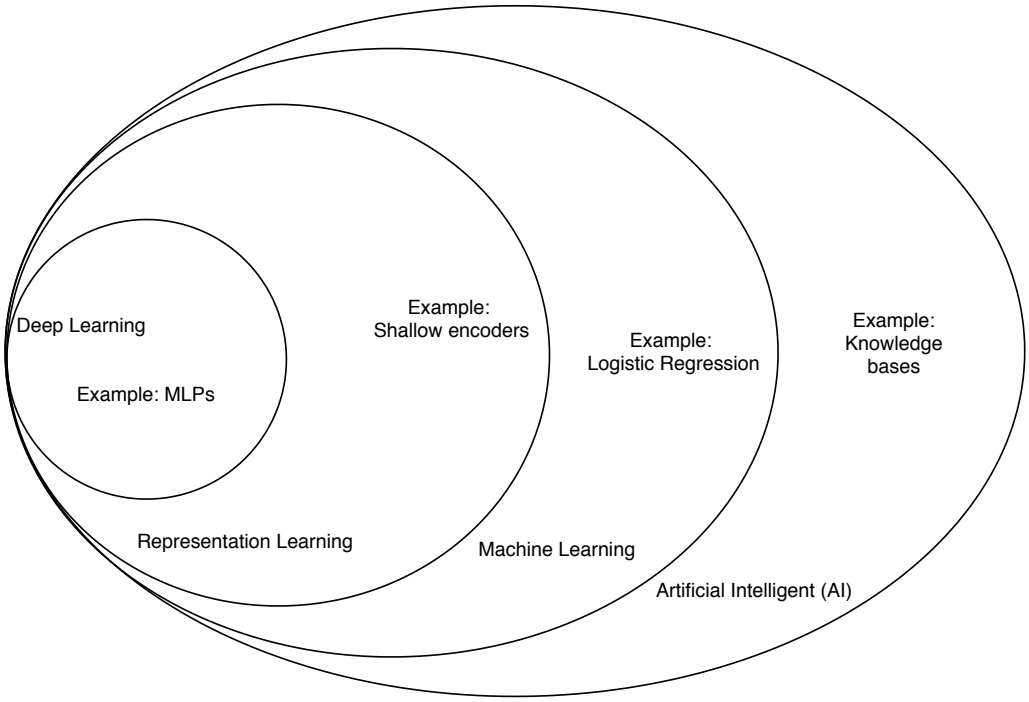

Figure 3.2: A Venn diagram showing how deep learning is a kind of representation learning by Goodfellow et al. [32].

> This line of algorithms are not the focus of this thesis, therefore, please see Sutton and Barto [84] for information.

Based on these paradigms, we will discover how privacy is related to each of them. In general, for supervised and unsupervised paradigms, the training experience E is normally given through a training data (or a training environment for reinforcement learning), which can be used interactively to improve the performance P on some task T. And normally, training data is collected from human generated data (e.g., news articles, Youtube's videos, Tweets etc.), they might contain sensitive information. Because of this characteristic, machine learning models might reveal sensitive information of individuals in training data.

Deep Learning (DL): is a specific kind of machine learning that is achieving many successes recently. Figure 3.2 shows how DL and ML are correlated, in which, DL is a subset of machine learning and focuses more on representation learning - the key factor that leads to recent advancements in Deep Learning [35, 36].

3.2 Privacy-Aware in Machine Learning

From the traditional machine learning point of views, most of machine learning models will need to learn from a training data generated by human. Therefore,

many researchers have been working on improving existing machine learning models to protect privacy of individuals containing in training data. Table 3.1 shows list of traditional algorithms which already have privacy-aware models.

Table 3.1: List of differentially private models. * denotes that DL was separated into a different paradigm since its architecture is flexible and can be used to train supervised or unsupervised models.

Paradigm	#	Privacy-aware models
Supervised	1	DP-Naive Bayes [88]
	2	DP-Linear Regression [101]
	3	DP-Linear SVM [95]
	4	DP-Logistic Regression [99]
	5	DP-Kernel SVM [77]
	6	DP-Decision Tree Learning [30]
	7	DP-Online Convex Programming [41]
	8	DP-K-nearest neighbours (KNN) [33]
Unsupervised	9	DP-K-means [71]
	10	DP-Feature Selection [89]
	11	DP-Principle Component Analysis (PCA) [34, 45]
Deep Learning*	12	DP-Differential Private Stochastic Gradient Descent(dpSGD) [2]
	13	DP-Convolutional Neural Network with differential privacy [50]
	14	DP-recurrent language models [57]
	15	DP-Word2Vec (dpUGC) [94]
	16	Private Aggregation of Teacher Ensembles (PATE) [72]
	16	And many others [75, 29, 1, 97, 70, 102, 76]

Differential privacy in Machine Learning

As in Chapter I, subsection 1.3 mentioned, differential privacy is the currently state-of-the-art approach to protect privacy for data analysis, data sharing, or machine learning models. Therefore, we now discuss more in detail how DP can protect privacy in training machine learning models, hereafter called DP-Models.

To address the challenge of revealing information about an individual in the training data, **differential privacy** [19, 26, 52, 51] essentially hides any individual by ensuring that the resulting model is nearly indistinguishable from the one without that individual. Differential privacy provides a strong guarantee of privacy even when the adversary has arbitrary external knowledge. The basic idea is to add enough noise to the outcome (e.g., the model resulting from training) to hide the contribution of any single individual to that outcome. Let

D be a collection of data records, and one record corresponds to an individual. A mechanism $\mathcal{M} : D \to \mathbb{R}^d$ is a randomized function mapping database D to a probability distribution over some range. $\mathcal{M}$ is said to be differentially private if adding or removing a single data record in D only affects the probability of any outcome within a small multiplicative factor. The formal definition of (ϵ, δ) differential privacy is:

Definition 1. ***[(ϵ-δ)-differential privacy]*** A randomized mechanism $\mathcal{M}$ is $(\epsilon,\ \delta)$-differential privacy where $\epsilon \geq 0, \delta \geq 0$, if for all data records in D and D' differing on at most one record, and $\forall \mathcal{S} \subseteq \text{Range}(\mathcal{M})$:

$$\Pr\left[\mathcal{M}(D) \in \mathcal{S}\right] \leq e^{\epsilon} \times \Pr\left[\mathcal{M}(D') \in \mathcal{S}\right] + \delta$$

The values of (ϵ, δ) here are called **privacy-budget**. They control the level of the privacy, i.e., smaller values of (ϵ, δ) guarantee better privacy but lower data utility. Since the introduction of differential privacy, there have been many other privacy-guarantee algorithms invented to fulfill the definition as shown in Table 3.1.

How differential privacy is applied in ML? the short answer to this question is to inject noise to the learning models following the distribution of the privacy-guarantee mechanisms (e.g., laplace mechanism [25]). It sounds easy to introduce noise into the machine learning models, however, how to control the amount of noise as well as how to control the noise will severely affect the learning models. For instance, if one simply injects noise into the resultant pre-trained models (e.g., word embedding models), the pre-trained models will no longer posses any useful information (e.g., the similarity between words in the model), therefore, will completely destroy the data utility. Phan et al. [75] introduced adaptive laplace noise to "smartly" distribute the noise to different features from which, their models can achieve both privacy and good data utility. Intuitively, most research in privacy-preservation ML models will try to use the same (or even less) level of noise but achieve better performance on some tasks in comparison to other models.

Privacy-Aware in Deep Learning

Deep learning is a kind of representation learning [32]. Therefore, it is not surprising when many researchers are trying to guarantee privacy for DL models since they are being applied in many sensitive tasks such as face recognition [49], genome prediction [8]. In this part, we mainly discuss on how differential privacy is added to deep learning models in order to achieve privacy-guarantee representations.

Loss function. (or *Loss* shortly) is one of the main terminologies using in deep learning to measure the penalty for mismatching between predicted outputs and the ground-truth outputs in the training data [1]. The loss $\mathcal{L}(\theta)$ on parameters θ is the average of the loss over training example $\{x_1, \ldots, x_N\}$ of a dataset D, so $\mathcal{L}(\theta) = \frac{1}{N}\Sigma_{i=1}^{N}\mathcal{L}(\theta, x_i)$. The training process is actually

a process of optimizing the set of parameters θ to find the acceptable small loss, that hopefully can reach an exact global minimum. From this loss, in the following parts, we will discuss in details how it can be hooked to provide privacy-guarantee DL models.

How to achieve DP-Models in deep learning? There have been different ways to provide privacy-guarantee DL models. Here we list two major approaches for training DP-Models in deep learning as follows:

1. Abadi et al. [1]: introduced DP-SGD (differential privacy for stochastic gradient descent (SGD)) - one of the main building block for achieving differential privacy in deep learning. In DP-SGD, constructed noise, that satisfied the definition of differential privacy [19], is injected to DL models during the optimization process:

$$\mathcal{M}(D) = \Sigma_{i \in B} \tilde{\nabla}(f(x_i)) + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

   where $\tilde{\nabla}(f(x_i))$ denotes the gradients clipped with a constant $C > 0$ for a minibatch $B \subset N$. $\mathcal{N}(0, S_f^2 \cdot \sigma^2)$ is the noise from the Gaussian noise mechanism [25] to function f of sensitivity S_f with mean 0 and noise scale σ.

2. PATE (Private Aggregation of Teacher Ensembles): introduced by Papernot et al. [72], in which they used multiple teachers to learn representations from sensitive data. Afterwards, the representations are shared differentially private to student models. Then the student models can use the DP-representations to improve tasks in public data. Following this mechanism, private data can be used to improve tasks in public data.

There are different directions to achieve privacy guarantee in training deep learning as well, however, most likely, they will follow the four different ways of injecting noise as shown in Figure 1.2.

3.3 Evaluation Problems of DP-Models

Following the problem of heterogeneous data, evaluating the effectiveness of privacy-guarantee algorithms is not trivial. The naive way to evaluate any privacy-guarantee models is to compare the performance between privacy-guarantee (DP) and non privacy-guarantee (Non-DP) models. The naive evaluation approach only works for well-established problems with well-established evaluation metrics, such as precision, recall, F1, accuracy (for classification), mean-average-error (for regression), and the like. However, for some learning tasks such as learning representations (e.g., Word2Vec [61], Elmo [74], Bert [22], Caffe [44]), there are no specific evaluation metrics for these models since they are pre-trained models that can be used for other down-stream tasks. Thus, there is no standard way to evaluate and compare between DP and Non-DP representations. Some works tried to compare the performances by using the

pre-trained models on down-stream tasks, then using the performances of the down-stream tasks to compare them [92, 75]. This is one way to show the difference in performances between DP and Non-DP algorithms, however, it is not a direct strategy to evaluate the models. We actually expect to have some evaluation metrics that directly evaluate the representation space inside those pre-trained models, from which, we know what models are performed better than others. In **Paper I** [93] and **Paper IV** [94], we showed different ways to evaluate performance of DP and Non-DP algorithms, however, they are preliminary works toward this direction. Thus, much work needs to be done to address this problem.

Chapter 4

Summary of Contributions

This chapter shows an overview of the thesis contribution by giving a summary of equipped research articles. First and foremost, it is important to show what are my contributions to each paper in Table 4.1. The list only shows that I contributed to the big part of each paper, however, it is never one-man's work.

Table 4.1: List of my contributions on each paper equipped in this thesis.

Paper	My contributions
Paper I [93]	- (1) Formulated research questions and solutions; (2) implemented the whole framework; (3) run experiments and evaluations; (4) wrote-up the paper together with other co-authors.
Paper II [90]	- (1) Formulated research questions and solutions; (2) implemented more than 60% of the whole framework; (3) investigated into case-studies to show in the paper; (4) wrote-up the paper together with other co-authors.
Paper III [92]	- (1) Formulated research questions and solutions; (2) implemented the neural network models; (3) run experiments and evaluations; (4) wrote-up the paper together with other co-authors.
Paper IV [94]	- (1) Formulated research questions and solutions; (2) implemented the neural network models; (3) run experiments and evaluations; (4) wrote-up the paper together with other co-authors.
Paper V [91]	- (1) Formulated research questions and solutions; (2) implemented the neural network models and run related experiments & evaluations; (3) wrote-up the paper together with other co-authors.

In the following sections, each paper is summarily described with reference

to the research objectives in Section 1.2 in Chapter 1. Lili Jiang acted as the main supervisor and Erik Elmroth had the role of the second supervisor. Thus, in most papers, advisers had advisory roles that include discussions about problem formulation, methodologies, experiments, evaluations, and how to present results. They also provided valuable feedback and suggestions during the writing process of all papers as well as this thesis.

4.1 Paper I [†] & II [††]

The existing infrastructures have many limitations of addressing privacy-guarantee methods on data analysis of federated databases. Some systems (e.g., PINQ [58], GUPT [66] provide the way to control user queries to satisfy differential privacy definition. However, they are more about a library that can be used by other system developers to integrate into their system, not for random researchers who want to access register data and have privacy-guarantee research results. Therefore, our proposed frameworks (called KaPPA [93] and INFRA [90]) fulfill this requirement by providing unified open-access frameworks that let researchers can flexibly discover register datasets and run data analysis within the frameworks.

KaPPA. Data-sharing is a good and fastest way to facilitate cross-disciplinary studies, to have larger sample sizes. It reduces the effort of making new data for many problems and makes optimal use of available data. However, sharing personal data between research parties raise a big problem in terms of privacy and data's confidentiality. To this end, we introduce KaPPA as a solution to the data-sharing and data analysis problem. Using KaPPA, the raw data will never leave the original data holder infrastructure and it is easier to control the use of the data and protect data-privacy for data analysis.

Cross-disciplinary studies have been conducted with the need for integrating these personal data from multiple sources. This data integration, however, dramatically increases the risk of privacy leakage [93]. Therefore, KaPPA was introduced to protect privacy of personal data using differential privacy for interactive privacy-preserving data analysis. Table 4.2 compares the differences between the traditional process in research on register data versus the process using KaPPA and INFRA, which is another proposed system of this thesis.

INFRA. Different from KaPPA that can focus on answering analytic queries in a form of privacy-guarantee histogram, INFRA [90] allows researchers to analyze register data in many different ways. In the paper II, using INFRA system, researchers can run data mining algorithms (e.g., association rule mining [3]) to find hidden patterns between multiple variables, from which, they narrow down

[†] **Personality-Based Knowledge Extraction for Privacy-preserving Data Analysis**, Xuan-Son Vu, Lili Jiang and Anders Brändström and Erik Elmroth, *ACM, Proceedings of the Knowledge Capture Conference (K-CAP), 2017.*

[††] **Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics**, Xuan-Son Vu, Addi Ait-Mlouk, Erik Elmroth, Lili Jiang, *ACM, Proceeding of WWW'19 - The World Wide Web Conference*, 2019.

Table 4.2: Procedure to research on sensitive data in a comparison between regular research process (i.e., * refers to [4]) and our proposed frameworks (i.e., ** refers to [93, 90]).

Traditional sensitive data analysis process*		Our proposed data analysis process**	
1. Research on requirements of data usage. 2. Working on research proposal 3. Send application to access data 4. Waiting for Approvals Panel's decision. 5. Negotiating for data and set-up. 6. Starting the analysis. 7. Repeat from 1 to 6 with new variables.		1. Register for accessing the system (online approval). 2. Research on the data. 3. Release research results. 4. No need to re-register for new variables, just change queries.	
Waiting time	in months	**Waiting time**	Less than a day
Privacy-guarantee	Regulation-constraint	**Privacy-guarantee**	Statistical guarantee

the interested variables to dig deeper for their research. Similar to KaPPA, the INFRA system is an open-access system and it does not require any special application procedure such as [4] for analyzing register data since all analytic processes are being done within the system, and no raw information will be shown to the researchers.

4.2 Paper III†

Paper III works on objective **RO2a** to solve privacy protection on any random datasets that were collected before and had no way to trace back to the data subjects. Thus, the main goal of this paper is to present a self-adaptive approach for privacy concern detection, which automatically detects the privacy need of individuals based on personality information extracted from their UGC data. In this way, we provide trade-off of sufficient privacy protection and data utility. The **main contributions** of this paper include:

- Introducing a neural network model that can learn and automatically predict the privacy-concern degree of individuals based on their personalities.
- Evaluating the effectiveness of personality based privacy-guarantee through extensive experimental studies on a real UGC dataset.
- Solving an imbalanced data distribution issue in privacy-concern detection raised by Vu et al. [93] using an over-sampling approach.

†**Self-adaptive Privacy Concern Detection for User-generated Content**, Xuan-Son Vu, Lili Jiang, *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2018.*

4.3 Paper IV [†]

Paper IV targets objective **RO2b** and **RO2c** since it introduces differential privacy algorithms for text data sharing. In this paper, we propose to use word embedding to share text distribution from a sensitive text corpus to facilitate similar tasks in public data. Word embedding, also known as word representation, represents a word as a vector capturing both syntactic and semantic information, so that the words with similar meanings should have similar vectors [54]. This representation has two important advantages: efficient representation due to dimensionality reduction, and semantic contextual similarity due to a more expressive representation.

Thanks for these advantages, word embedding is widely used to learn text representation for text analysis tasks. Some commonly used word embedding models include Word2Vec [61], GloVe [73], and FastText [12] and successfully applied in a variety of tasks like parsing [5], topic modeling [7]. However, since word embedding models preserve pretty much semantic relations between words, the shared pre-trained models may lead to privacy breaches especially when they were trained from UGC data such as tweets and Facebook posts. For instance, user *first name* (e.g., "John"), *last name* ("Smith") and *disease* (e.g., "prostatitis") may be represented as similar vectors in word embedding model. Even user real name is absent from the pre-trained models, other available information such as *username, address, city name, occupation*, could be represented with similar vectors, with/without auxiliary data, leading to re-identification risk to discover the individual to which the data belongs to, by using some approaches like author identification [67], age and gender prediction [28]. One might argue that the sensitive information likes *user, password* should not be leaked out and should have been removed from the embedding model. However, the purpose of learning from sensitive data is to learn the model without privacy leakage for facilitating research on sensitive data. To protect privacy, we statistically guarantee the chance to re-identify individuals by using output from the pre-trained models. Thanks to that, further research on the sensitive data **at large scale** can be possible such as "what is the common patterns between users when they configure their passwords?" (to analyze security risks) or "what diseases are normally unspeakable but get shared online?" (to analyze user behaviours on social networks).

As discussed above, it is critical to protecting privacy when learning embedding model for UGC data sharing. To address the challenge of revealing information about an individual in the training data, this paper proposed to use differential privacy [19] in a neural network architecture to learn privacy-guarantee word embedding models.

[†]**dpUGC: Learn Differentially Private Representation for User Generated Contents**, Xuan-Son Vu, Son N. Tran, Lili Jiang, *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2019.*

4.4 Paper V [†]

Paper V targets **RO2c**, i.e., collaboratively apply the privacy-aware data federation solution on social network for social network analysis. The approach used in this paper, termed **M**ulti**L**ayer **N**etwork (MLN) analysis with decoupling, is in its early stages and being researched actively. The MLN approach does not change the analysis, *except how* datasets are modeled and analyzed. It has been receiving a lot of attention in the last decade due to its advantages: i) *allows modeling of a complex dataset using a set of user-definable simple, single graphs termed layers)* , ii) *allows the same analysis as the traditional approach on this model without loss of accuracy*, iii) is amenable to parallelism (for scalability) and has been shown to be better in storage requirements and efficiency. There are other advantages as well [80, 47]. This paper is the first one, to the best of our knowledge, to apply this approach for the analysis of *one of the largest/densest real-world social network data collection*, although it has been used in several experimental studies on smaller/sparser datasets [14, 11].

The contributions of this **Paper V** include: **(1)** using a novel, emerging MLN approach for flexible analysis of a large complex real-world dataset (e.g., to understand how privacy-concerns vary in different age groups), **(2)** establishing its modeling benefits, flexibility of analysis, and efficiency of computation, **(3)** integrating content analysis seamlessly with structural network analysis, and **(4)** extensive analysis and result validation for the social network work datasets.

4.5 Future Work

The presented studies in this thesis are possible to be extended in many directions. First, the federated infrastructure designs are limited to some in-of-the-box features. At the current state, they can support much different analysis, however, they do not support any analytic programming languages such as R or Python. Thus, this extension might be very valuable for researchers. Secondly, to the privacy-guarantee algorithms, as mentioned before in previous sections, it is not straightforward to evaluate the performance of DP versus Non-DP algorithms. Therefore, more works in this direction have to be done to find good evaluation metrics for relevant problems. Lastly, in the near future, we are targeting to explore different privacy-guarantee mechanisms to support privacy-guarantee data sharing tasks since this line of tasks are very important to facilitate data sharing and hence, improve research performances of other topics.

[†]**Generic Multilayer Network Data Analysis with the Fusion of Content and Structure**, Xuan-Son Vu, Abhishek Santra, Sharma Chakravarthy, Lili Jiang, *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2019.*

Bibliography

[1] M. Abadi, A. Chu, I. Goodfellow, H. Brendan McMahan, I. Mironov, K. Talwar, and L. Zhang. "Deep Learning with Differential Privacy". In: *ArXiv e-prints* (July 2016).

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: ACM, 2016, pp. 308–318.

[3] Rakesh Agrawal and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In: *Proceedings of the 20th International Conference on Very Large Data Bases*. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[4] Australian National Data Service (ANDS). *Application process to research on sensitive data with Ethics and Consent*. `https://www.adrn.ac.uk/get-data/application-process/`ANDS's application process and `https://utas.libguides.com/researchdatamanagement/ethics_sensitivedata`ANDS's ethics and consent. 2017. (Visited on 06/30/2017).

[5] Mohit Bansal, Kevin Gimpel, and Karen Livescu. "Tailoring Continuous Word Representations for Dependency Parsing". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 809–815.

[6] E. Barendt. *Privacy*. The International Library of Essays in Law and Legal Theory (Second Series). Taylor & Francis, 2017. ISBN: 9781351908801.

[7] Kayhan N. Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Samuel Gershman. "Nonparametric Spherical Topic Modeling with Word Embeddings". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2016), pp. 537–542.

[8] Pau Bellot, Gustavo de los Campos, and Miguel Pérez-Enciso. "Can Deep Learning Improve Genomic Prediction of Complex Human Traits?" In: *Genetics* 210.3 (2018), pp. 809–819.

[9] Bhume Bhumiratana and Matt Bishop. "Privacy Aware Data Sharing: Balancing the Usability and Privacy of Datasets". In: *Proceedings of the 2Nd International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '09. Corfu, Greece: ACM, 2009, 73:1–73:8.

[10] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. "Practical Privacy: The SuLQ Framework". In: *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '05. Baltimore, Maryland: ACM, 2005, pp. 128–138.

[11] Brigitte Boden, Stephan Günnemann, Holger Hoffmann, and Thomas Seidl. "Mining Coherent Subgraphs in Multi-Layer Graphs with Edge Labels". In: *Proc. of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2012, pp. 1258–1266.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[13] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou. "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data". In: *IEEE Transactions on Parallel and Distributed Systems* 25.1 (2014), pp. 222–233.

[14] Alessio Cardillo, Jesús Gómez-Gardenes, Massimiliano Zanin, Miguel Romance, David Papo, Francisco Del Pozo, and Stefano Boccaletti. "Emergence of network features from multiplexity". In: *Scientific reports* 3 (2013).

[15] Kamalika Chaudhuri and Daniel Hsu. "Sample Complexity Bounds for Differentially Private Learning". In: *Proceedings of the 24th Annual Conference on Learning Theory*. Ed. by Sham M. Kakade and Ulrike von Luxburg. Vol. 19. Proceedings of Machine Learning Research. Budapest, Hungary: PMLR, Sept. 2011, pp. 155–186.

[16] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. "Near-optimal Differentially Private Principal Components". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 989–997.

[17] Glen Coppersmith, Mark Dredze, and Craig Harman. "Quantifying Mental Health Signals in Twitter". In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 51–60.

[18] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Machine Learning*. 1995, pp. 273–297.

[19] Dwork Cynthia. "Differential Privacy". In: ICALP. 2006, pp. 1–12.

[20] Tore Dalenius. "Towards a methodology for statistical disclosure control". In: *statistik Tidskrift* 15.429-444 (1977), pp. 2–1.

[21] Shuiguang Deng, Longtao Huang, and Guandong Xu. "Social network-based service recommendation with trust enhancement". In: *Expert Systems with Applications* 41.18 (2014), pp. 8075–8084.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT 2019* (2019), pp. 4171–4186.

[23] M. Du, K. Wang, Y. Chen, X. Wang, and Y. Sun. "Big Data Privacy Preserving in Multi-Access Edge Computing for Heterogeneous Internet of Things". In: *IEEE Communications Magazine* (2018), pp. 62–67.

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284. ISBN: 978-3-540-32732-5.

[25] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3–4 (Aug. 2014), pp. 211–407.

[26] Cynthia Dwork and Adam Smithy. "Differential privacy for statistics:What we know and what we want to learn". In: (2009).

[27] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. Scottsdale, Arizona, USA: ACM, 2014, pp. 1054–1067.

[28] Lucie Flekova and Iryna Gurevych. "Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media Notebook for PAN at CLEF 2013". In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. 2013.

[29] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. 2015, pp. 1322–1333.

[30] Arik Friedman and Assaf Schuster. "Data Mining with Differential Privacy". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: ACM, 2010, pp. 493–502.

[31] *Genome-wide association study (GWAS)*. (Visited on 06/30/2017).

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[33] Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz, and Yucel Saygin. "Differentially Private Nearest Neighbor Classification". In: *Data Min. Knowl. Discov.* 31.5 (Sept. 2017), pp. 1544–1575.

[34] Moritz Hardt and Aaron Roth. "Beyond Worst-case Analysis in Private Singular Vector Computation". In: *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*. STOC '13. Palo Alto, California, USA: ACM, 2013, pp. 331–340.

[35] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507.

[36] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets". In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[37] Fei Hu, Yu Lu, Athanasios V. Vasilakos, Qi Hao, Rui Ma, Yogendra Patil, Ting Zhang, Jiang Lu, Xin Li, and Neal N. Xiong. "Robust Cyber–Physical Systems: Concept, models, and implementation". In: *Future Generation Computer Systems* 56 (2016), pp. 449–475.

[38] Budin-Ljøsne I, Burton PR, Isaeva J, Gaye A, Turner A, Murtagh MJ, Wallace S, Ferretti V, and Harris JR. "DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis". In: *Public Health Genomics* (2015), pp. 87–96.

[39] Fortier I, Raina P, Van den Heuvel E R, Griffith LE, Craig C, Saliba M, Doiron D, Stolk RP, Knoppers BM, Ferretti V, and Granda P. "Maelstrom Research guidelines for rigorous retrospective data harmonization". In: *International journal of epidemiology* (2016).

[40] Michael J Paul, Mark Dredze, and David Broniatowski. "Twitter Improves Influenza Forecasting". In: *PLoS currents* 6 (Oct. 2014).

[41] Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. "Differentially Private Online Learning". In: *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by Shie Mannor, Nathan Srebro, and Robert C. Williamson. Vol. 23. Proceedings of Machine Learning Research. Edinburgh, Scotland: PMLR, 25–27 Jun 2012, pp. 24.1–24.34.

[42] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. "Big data privacy: a technological perspective and review". In: *Journal of Big Data* (2016).

[43] Zhanglong Ji, Zachary Chase Lipton, and Charles Elkan. "Differential Privacy and Machine Learning: a Survey and Review". In: *CoRR* abs/1412.7584 (2014). arXiv: 1412.7584.

[44] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, 2014, pp. 675–678.

[45] Michael Kapralov and Kunal Talwar. "On Differentially Private Low Rank Approximation". In: *Proceedings of the Twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '13. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2013, pp. 1395–1414.

[46] A. Katal, M. Wazid, and R. H. Goudar. "Big data: Issues, challenges, tools and Good practices". In: *2013 Sixth International Conference on Contemporary Computing (IC3)*. 2013, pp. 404–409.

[47] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. "Multilayer Networks". In: *CoRR* abs/1309.7233 (2013).

[48] Kostas Kolomvatsos, Christos Anagnostopoulos, and Stathes Hadjiefthymiades. "An Efficient Time Optimized Scheme for Progressive Analytics in Big Data". In: *Big Data Research* 2.4 (2015), pp. 155–165.

[49] Kamran Kowsari, Mojtaba Heidarysafa, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. "RMDL: Random Multimodel Deep Learning for Classification". In: *CoRR* abs/1805.01890 (2018). arXiv: 1805.01890.

[50] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. "Certified robustness to adversarial examples with differential privacy". In: *arXiv preprint arXiv:1802.03471* (2018).

[51] Jaewoo Lee and Chris Clifton. "Differential Identifiability*". In: *Proceedings of KDD*. 2012.

[52] Jaewoo Lee and Chris Clifton. "How much is enough? choosing ϙ for differential privacy". In: (2011), pp. 325–340.

[53] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. "Workload-aware Anonymization Techniques for Large-scale Datasets". In: *ACM Trans. Database Syst.* 33.3 (Sept. 2008), 17:1–17:47.

[54] Omer Levy and Yoav Goldberg. "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, 2014, pp. 171–180.

[55] 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*. 2007, pp. 106–115.

[56] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. "L-diversity: privacy beyond k-anonymity". In: *22nd International Conference on Data Engineering (ICDE'06)*. 2006, pp. 24–24.

[57] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. "Learning Differentially Private Language Models Without Losing Accuracy". In: *CoRR* abs/1710.06963 (2017). arXiv: `1710.06963`.

[58] Frank D McSherry. "Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis". In: *SIGMOD*. 2009.

[59] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo. "Protection of Big Data Privacy". In: *IEEE Access* 4 (2016), pp. 1821–1834.

[60] Peter M. Mell and Timothy Grance. *SP 800-145. The NIST Definition of Cloud Computing*. Tech. rep. Gaithersburg, MD, United States, 2011.

[61] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* (2013). arXiv: `1301.3781`.

[62] D. J. Mir and R. N. Wright. "A Differentially Private Graph Estimator". In: *2009 IEEE International Conference on Data Mining Workshops*. 2009, pp. 122–129.

[63] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[64] T. Mitchell. "Machine Learning". In: McGraw-Hill, 1997, p. 2. ISBN: 978-0-07-042807-2.

[65] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. "DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis". In: (2015), pp. 87–96.

[66] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. "GUPT: Privacy Preserving Data Analysis Made Easy". In: *SIGMOD*. 2012.

[67] A. M. Mohsen, N. M. El-Makky, and N. Ghanem. "Author Identification Using Deep Learning". In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2016, pp. 898–903.

[68] A. Narayanan and V. Shmatikov. "Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset)". In: (2008).

[69] Netflix. *Netflix Prize Contest*. 2009. URL: `https://en.wikipedia.org/wiki/Netflix_Prize` (visited on 06/30/2017).

[70] Hiep H. Nguyen, Abdessamad Imine, and Michaël Rusinowitch. "Detecting Communities Under Differential Privacy". In: *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. WPES '16. Vienna, Austria: ACM, 2016, pp. 83–93.

[71] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. "Smooth Sensitivity and Sampling in Private Data Analysis". In: *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*. STOC '07. San Diego, California, USA: ACM, 2007, pp. 75–84.

[72] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. "Scalable Private Learning with PATE". In: *Sixth International Conference on Learning Representation (ICLR 2018)* (2018).

[73] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[74] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations". In: *Proc. of NAACL*. 2018.

[75] NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. "Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning". In: *CoRR* abs/1709.05750 (2017).

[76] Vadim Popov, Mikhail Kudinov, Irina Piontkovskaya, Petr Vytovtov, and Alex Nevidomsky. "Distributed Fine-tuning of Language Models on Private Data". In: *International Conference on Learning Representations*. 2018.

[77] Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. "Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning". In: *Journal of Privacy and Confidentiality* abs/0911.5708 (2009).

[78] Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Y. Zhao. "Sharing Graphs Using Differentially Private Graph Models". In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. IMC '11. Berlin, Germany: ACM, 2011, pp. 81–98.

[79] Pierangela Samarati and Latanya Sweeney. *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*. Tech. rep. 1998.

[80] Abhishek Santra, Sanjukta Bhowmick, and Sharma Chakravarthy. "Efficient Community Re-creation in Multilayer Networks Using Boolean Operations". In: *International Conference on Computational Science, ICCS 2017*. 2017, pp. 58–67.

[81] Amit P. Sheth and James A. Larson. “Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases”. In: *ACM Comput. Surv.* (1990), pp. 183–236.

[82] O.M. Soundararajan, Y. Jenifer, S. Dhivya, and T.K.P. Rajagopal. “Data Security and Privacy in Cloud Using RC6 and SHA Algorithms”. In: *Networking and Communication Engineering* 6.5 (2014).

[83] Stefan Stieger, Christoph Burger, Manuel Bohn, and Martin Voracek. “Who Commits Virtual Identity Suicide? Differences in Privacy Concerns, Internet Addiction, and Personality Between Facebook Users and Quitters”. In: *Cyberpsychology, Behavior, and Social Networking* 16.9 (2013). PMID: 23374170, pp. 629–634.

[84] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[85] Latanya Sweeney. “Achieving K-anonymity Privacy Protection Using Generalization and Suppression”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10.5 (Oct. 2002), pp. 571–588.

[86] *Teiid: a data virtualization system that allows applications to use data from multiple, heterogeneous data stores*. `http://teiid.jboss.org/`. commit xxxxxxx. 2016.

[87] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos. “Big data analytics: a survey”. In: *Journal of Big Data* 2.1 (2015), p. 21.

[88] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong. “Differentially Private Naive Bayes Classification”. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 1. 2013, pp. 571–576.

[89] Staal A. Vinterbo. “Differentially Private Projected Histograms: Construction and Use for Prediction”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 19–34. ISBN: 978-3-642-33486-3.

[90] Xuan-Son Vu, Addi Ait-Mlouk, Erik Elmroth, and Lili Jiang. “Graph-based Interactive Data Federation System for Heterogeneous Data Retrieval and Analytics”. In: *Demo Track, In: Proceedings of the The Web Conference 2019*. TheWebConf ’19 - formerly WWW. International World Wide Web Conferences Steering Committee, 2019.

[91] Xuan-Son Vu and Lili Jiang. “Generic Multilayer Network Data Analysis with the Fusion of Content and Structure”. In: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, April, 2019*. La Rochelle, France, 2019.

[92] Xuan-Son Vu and Lili Jiang. "Self-adaptive Privacy Concern Detection for User-generated Content". In: *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Vol. Volume 1: Long papers, p., March 2018*. Hanoi, Vietnam, 2018.

[93] Xuan-Son Vu, Lili Jiang, Anders Brändström, and Erik Elmroth. "Personality -based Knowledge Extraction for Privacy-preserving Data Analysis". In: *Proceedings of the Knowledge Capture Conference*. K-CAP 2017. ACM, 2017, 45:1–45:4.

[94] Xuan-Son Vu, Son N. Tran, and Lili Jiang. "dpUGC: Learn Differentially Private Representationfor User Generated Contents". In: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, April, 2019*. La Rochelle, France, 2019.

[95] Di Wang, Changyou Chen, and Jinhui Xu. "Differentially Private Empirical Risk Minimization with Non-convex Loss Functions". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, Sept. 2019, pp. 6526–6535.

[96] Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K. Reiter, and Zheng Dong. "Privacy-preserving Genomic Computation Through Program Specialization". In: CCS. 2009, pp. 338–347.

[97] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. "Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Cham: Springer International Publishing, 2018, pp. 627–645.

[98] Li Xiong and Kumudhavalli Rangachari. "Towards Application-Oriented Data Anonymization". In: *nternational Workshop onPractical Privacy-Preserving Data Mining* (2008).

[99] Depeng Xu, Shuhan Yuan, and Xintao Wu. "Achieving Differential Privacy and Fairness in Logistic Regression". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: ACM, 2019, pp. 594–599.

[100] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren. "Information Security in Big Data: Privacy and Data Mining". In: *IEEE Access* 2 (2014), pp. 1149–1176.

[101] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. "Functional Mechanism: Regression Analysis Under Differential Privacy". In: *Proc. VLDB Endow.* 5.11 (July 2012), pp. 1364–1375.

[102] Ye Zhang, Nan Ding, and Radu Soricut. "SHAPED: Shared-Private Encoder-Decoder for Text Style Adaptation". In: *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)* (2018).

[103] Xiaojin Zhu. "Semi-Supervised Learning Literature Survey". In: *Comput Sci, University of Wisconsin-Madison* 2 (July 2008).