UMEÅ UNIVERSITY

Privacy-Guardian: The Vital Need in Machine Learning with Big Data

Xuan-Son Vu

Doctoral Thesis, October 2020
Department of Computing Science
Umeå University
Sweden

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

sonvx@cs.umu.se

ISBN 978-91-7855-377-8
ISSN 0348-0542
UMINF 20.11

Cover illustration by Monika Jingar
Printed by Cityprint i Norr AB, Umeå, 2020

Popular Science Abstract

Let us tell you a story of the future, in the year of 2050. The story involves the lives of Alice and Bob. Alice is a netizen (i.e., *the citizen of the Internet*) and a psychologist researcher. Bob is a freelance photographer, who loves to capture unique moments in life. In daily activities, Alice likes to go to a virtual world to interact with her virtual friends and colleagues. Though Alice is an introvert person in the real life, she is very active in the virtual world and loves to share what she thinks and wherever places she has traveled. Bob and Alice never know each other. However one day, when Bob was taking photos of daily activities in a street, Alice was accidentally walking into the scene and was captured. Bob was curious who was the pretty girl in the photo. He started to search for her face using his *secret app*, then he knew the way to see her activities in the virtual world. Based on her online activities, Bob started to know her friends, her hobbies, and her profession. However, what Bob did not know was that, Alice intentionally got into one of the frame of Bob's photos to do research on Bob. Because Alice, as a psychologist researcher, was working on a research in understanding human behaviors titled *'what would be a set of actions a human being will perform when they get curious?'*. By understanding human behaviors, Alice wants to help the law enforcement to predict suicidal tendencies from which, actions could be made before bad things really happen.

Sorry, we lied. The above story is not a story of the future, but a story of our current world. The virtual world is in fact, the social network sites (e.g., Facebook, Twitter, Instagram) that people are using daily. The *secret app* that Bob used to find information of Alice is in fact, a controversial app called ClearView. And what Alice and Bob were doing in the story, they have intruded privacy of people. Given the fact that Alice's research is very valuable for the social good. It helps the decision makers to introduce new policy to better support people's lives. However, there has to have a better solution for such research's needs. In the best scenario, there should have a system to allow researchers to work on private sensitive data while protecting people's privacy. As a step towards this direction, this thesis introduces new set of *Privacy Utilities* and *Privacy-Aware Algorithms* to help researchers and machine learning practitioners work on personal big data. It provides both *system frameworks* and equipped *privacy-aware algorithms* to let researchers work on private sensitive data without worrying about privacy leakages.

Abstract

Social Network Sites (SNS) such as Facebook and Twitter, play a great role in our lives. On one hand, they help to connect people who would not otherwise be connected. Many recent breakthroughs in AI such as facial recognition [Kow+18], were achieved thanks to the amount of available data on the Internet via SNS (hereafter Big Data). On the other hand, many people have tried to avoid SNS to protect their privacy [Sti+13]. However, Machine Learning (ML), as the core of AI, was not designed with privacy in mind. For instance, one of the most popular supervised machine learning algorithms, Support Vector Machines (SVMs), try to solve a quadratic optimization problem in which the data of people involved in the training process is also published within the SVM models. Similarly, many other ML applications (e.g., ClearView) compromise the privacy of individuals presented in the data, especially when the big data era enhances the data federation. Thus, in the context of machine learning with big data, it is important to (1) protect sensitive information (privacy protection) while (2) preserving the quality of the output of algorithms (i.e., data utility).

For the vital need of privacy in machine learning with big data, this thesis studies on: (1) how to construct information infrastructures for data federation with privacy guarantee in the big data era; (2) how to protect privacy while learning ML models with a good trade-off between data utility and privacy. To the first point, we proposed different frameworks empowered by privacy-aware algorithms. Regarding the second point, we proposed different neural architectures to capture the sensitivities of user data, from which, the algorithms themselves decide how much they should learn from user data to protect their privacy while achieving good performances for downstream tasks. The current outcomes of the thesis are: (a) privacy-guarantee data federation infrastructure for data analysis on sensitive data; (b) privacy utilities for privacy-concern analysis; and (c) privacy-aware algorithms for learning on personal data. For each outcome, extensive experimental studies were conducted on real-life social network datasets to evaluate aspects of the proposed approaches.

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

Sammanfattning

Sociala nätverkssajter (SNS) som Facebook och Twitter har spelat en stor roll i våra liv. Å ena sidan hjälper de till att sammankoppla människor som annars aldrig skulle komma i kontakt med varandra. Många av de senaste genombrotten inom AI, såsom ansiktigenkänning [Kow+18], uppnåddes tack vare mängden tillänglig data på internet via SNS (hädanefter Big Data). Å andra sidan har manga försökt att unvika SNS för att skydda deras integritet [Sti+13]. Machine Learning (ML), som kärnan i AI, var emellertid aldrig utformad med integritet i åtanke. Till exempel så försöker ett av de mest populära supervised machine learning algorithmerna, Support Vector Machines (SVMs), att lösa ett kvadratiskt optimeringsproblem där personlig data involverade i träningsprocessen även blir tillänglig inom SVM-modellen.

På liknande sätt äventyrar många andra ML-applikationer (t.ex. ClearView) integriteten för individer som presenteras i data, särskilt när Big Data-eran ökar datafederationen. I kontexten av Machine Learning med Big Data är det alltså viktigt att (1) skydda känslig information (privacy protection) och samtidigt (2) bevara kvaliteten på algoritmernas resultat. (dvs, data utility). För det vitala behovet av integritet vid maskininlärning med stora data studerar denna avhandling: (1) hur man konstruerar informationsinfrastrukturer för datafederation med integritetsgaranti i Big Data-eran (2) hur integritet kan skyddas när man tränar ML-modeller med en bra avvägning mellan data utility och integritet. För den första punkten föreslår vi olika ramverk skapade med integritetsmedvetna algoritmer. Gällande den andra punkten föreslår vi olika neuralarkitekturer för att fånga känsligheten hos användardata, varifrån algoritmerna själva avgör hur mycket de ska lära sig av användardata för att skydda deras integritet samtidigt som de uppnår god prestanda för nedströmsuppgifter. De aktuella resultaten av avhandlingen är: (1) integritetsgaranterad data-federationsinfrastruktur för dataanalys av känsliga data; (2) integritetsverktyg för analys från integritetshänsyn; och (3) integritetsmedvetna algoritmer för inlärning på personuppgifter. För varje resultatet genomfördes omfattande experimentella studier på verkliga sociala nätverksdataset i för att utvärdera aspekter av de föreslagna metoderna.

Insikter och resultat från denna avhandling kan användas av både den akademiska världen och industrin för att tillhandahålla integritetsgaranterad dataanalys och datalärande i Big Data innehållande personlig information. De har också potential att underlätta relevant forskning inom integritetsmedvetet lärande och dess relaterade utvärderingsmetoder.

Preface

This thesis contains a brief description of privacy-aware infrastructures, a discussion on improving privacy protection approaches in natural language processing and machine learning, attachments of 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), **3^{rd} place for best paper awards**, 2019.*

Paper V **Xuan-Son Vu**, Thanh-Son Nguyen, Duc-Trong Le, Lili Jiang. Multimodal Review Generation with Privacy and Fairness Awareness. *Submitted, 2020.*

Paper VI **Xuan-Son Vu**, Duc-Trong Le, Christoffer Edlund, Lili Jiang, Hoang D. Nguyen. Privacy-Preserving Visual Content Tagging using Graph Transformer Networks. *Proceedings of the 28th ACM international conference on Multimedia (ACM MM), 2020.*

Paper VII **Xuan-Son Vu**, Thanh Vu, Son N. Tran, Lili Jiang. ETNLP: a visual-aided systematic approach to select pre-trained embeddings for a downstream task. *Proceedings of the* 13^{rd} *International Conference Recent Advances in Natural Language Processing (RANLP), 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).*

4. Son N. Tran, Dung Nguyen, Tung-Son Ngo, **Xuan-Son Vu**, Long Hoang, Qing Zhang & Mohan Karunanithi, On multi-resident activity recognition in ambient smart-homes, *Artificial Intelligence Review, 2019.*

5. Hoang D. Nguyen, **Xuan-Son Vu**, Quoc-Truong Le, Duc-Trong Le, Reinforced Data Sampling for Model Diversification, *Submitted, 2020.*

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).

Acknowledgments

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 in 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 Johanna, Eddie, Chanh, Hannah, Mahmoud, Monowar, Juan Carlos, Angelika, Birgitte, Ahmad, Mats, Thang, Anne-Lie, Mikael Hansson, Carl Christian, Mirko, Monika, Timotheus, Suna, Michele, Abel, Jakub, Maitreyee, Thomas, Addi, Frank, and many others. Special thanks to Tomas Forsman for being the magical technical guru, for which I am very super 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 closed collaborators, in the arbitrary selection order[*], including Son N. Tran, Thanh Vu, Thanh-Son Nguyen, Duc-Trong Le, and Hoang D. Nguyen, 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 them.

/Xuan-Son Vu

[*] `http://bit.ly/arbitraryselectionprocess`

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 **5Vs** of Big Data [Tsa+15]
3	UGC	User Generated Content [VJ18]
4	DF	Data Federation [Vu+17a; Vu+19a]
5	DA	Data Analysis [Vu+19a]
6	DP	Differential Privacy [Cyn06; DS09]
7	DS	Data Sharing [VTJ19]
8	ML	Machine Learning [Mit97a]
9	MLN	Multi-layer Network [Vu+19b]
10	NA	Network Analysis [Vu+19b]
11	GCN	Graph Convolutional Neural Network [KW17]
12	SVM	Support Vector Machines [CV95]
13	GDPR	General Data Protection Regulation
14	FL	Federated Learning [McM+17]
15	dpUGC	Differentially Private Embeddings for User Generated Contents [VTJ19]
16	SIS	Smart Information System [SW18]

Contents

Chapter 1

Introduction

My PhD studies focus on research in *Privacy-aware Data Federation*, which aims at (1) virtually integrating heterogeneous data from multiple distributed sources and (2) privacy preserving for data learning 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 was generating 25TB of new data every day back in 2016 [Meh+16]. In 2019, the daily amount of data Facebook created was 4PB per day, which was $\sim$ 164 times more than in 2016, according to a statistic of Visual Capitalist[†]. 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 schemata, and using different methodologies. This imposes the requirements on the technology for effective data integration and data sharing across multiple heterogeneous sources. Among this increasing volume of data, individual personal data can be largely collected and analyzed to understand important phenomena, such as early detection of diseases [JDB14] and social service recommendation [DHX14]. 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 for data learning and data analysis with privacy preservation.

1.1 Research Motivation

Most of research in privacy focus in privacy-guaranteed algorithms with two popular approaches including anonymization and sanitization. However, with the complexity of personal data in the social network era, not only is guaranteed privacy needed, understanding privacy concern of users is also important.

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

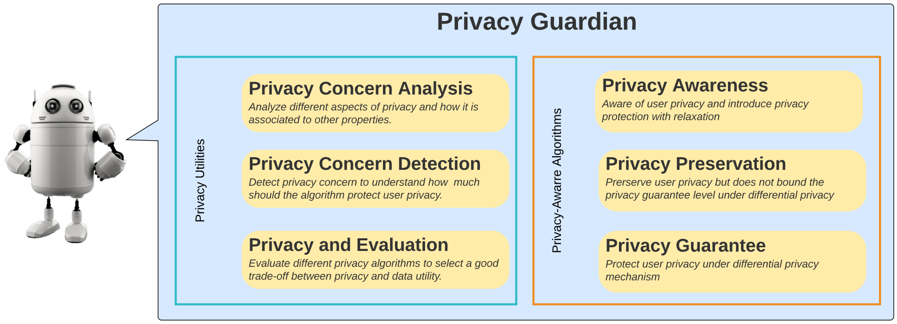

Privacy Guardian
Privacy Utilities
Privacy Concern Analysis
Analyze different aspects of privacy and how it is associated to other properties.
Privacy Concern Detection
Detect privacy concern to understand how much should the algorithm protect user privacy.
Privacy and Evaluation
Evaluate different privacy algorithms to select a good trade-off between privacy and data utility.
Privacy-Aware Algorithms
Privacy Awareness
Aware of user privacy and introduce privacy protection with relaxation
Privacy Preservation
Prerserve user privacy but does not bound the privacy guarantee level under differential privacy
Privacy Guarantee
Protect user privacy under differential privacy mechanism

Figure 1.1: *Privacy-Guardian* addresses vital needs in machine learning with personal big data including: (1) privacy utilities and (2) privacy-aware algorithms. *Privacy utilities* provide insights for privacy related application or algorithms such as (a) privacy concern analysis, (b) privacy concern detection, (c) privacy and evaluation). And *privacy-aware algorithms* employ different mechanism to protect user privacy while learning from their data.

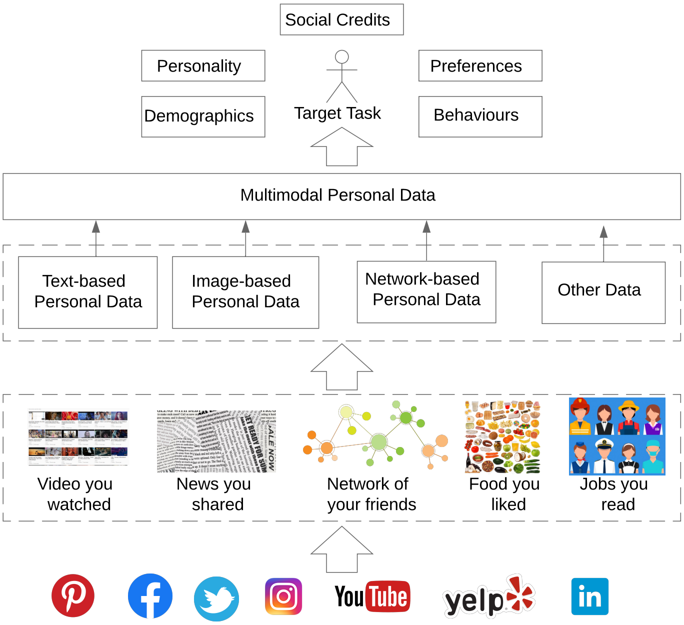


Figure 1.2: Multimodal personal data on SNS helps to improve scientific research and personalized systems, however, it introduces new challenges in privacy protection.

Different *privacy utilities* equipped to analyze user privacy concerns in this thesis is illustrated in the left side column of Figure 1.1. These privacy utilities are needed for two reasons. Firstly, a large amount of personal data was collected in the past, and the data controller has no way to contact the users to ask for their consent. Therefore, there should be an automatic solution to detect privacy concern level (e.g., high, medium, low) of users, to protect them accordingly as if users had given their consent. Secondly, many users are not aware of privacy issue unless they have had privacy breaches in the past, or there is a system reminding them to pay more attention into the sensitivity of their data. Beyond the essential privacy preservation, users' privacy concerns vary in different contexts. Regarding privacy guarantee aspect, the high availability of heterogeneous and multimodal data makes the task even more difficult for protecting user privacy. Figure 1.2 visualizes the collection of different data that contributes to the multimodal data of a user based on popular social network services (e.g., Facebook, Twitter, Linkedin, Instagram etc.). It shows various types of user data (from conventional unimodal data to multimodal data) and potential downstream tasks that can use machine learning algorithms to learn and serve as a core solution of an application. Because of the complexity of the whole learning process from a raw data to an application (e.g., a personalized recommender), it is not always trivial to apply existing privacy-guaranteed approaches into these applications. The main reason of this fact is the current limitation of the field since most of the current learning methods are task and data oriented. It means, with different task and data, one must apply or build a different model to solve the task. There is no *one-size-fits-all* model that can solve all tasks yet. This fact is termed as "no free lunch theorem"*. AutoML [Rea+20] is an interesting line of research towards general AI. Despite having significant progress, it is far from reality to be applied arbitrarily on any tasks. Therefore, there is a high need for each related application, to be designed and equipped specific tools for privacy analysis and privacy guarantee. From these reasons, this thesis introduces a collection of privacy-related algorithms to deal with different needs in machine learning with big data.

Briefly, Figure 1.1 describes **Privacy-Guardian**, the vital need in machine learning with big data in which both (1) *privacy utilities* and (2) *privacy-aware algorithms* are explored to support different purposes related to privacy in research and applications. *Privacy utilities* provides different approaches to detect and analyze privacy concern of users based on their data. And *Privacy-aware algorithms* proposes different learning methods to address privacy of users in the learning process.

1.2 Privacy in a Nutshell

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

Privacy. Many legal systems protect a right to privacy. However, 'privacy' remains an elusive and controversial concept [Bar17]. In the book called *Pri-*

*https://en.wikipedia.org/wiki/No_free_lunch_theorem *vacy* [Bar17], Barendt addressed the fact that "some writers have rejected the idea that there is a 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 an intimate couple, the intruder could only be charged with a *break-in crime* but not for something else (e.g., violated private 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" [Bar17]. 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., as norm, in law, and with 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 discuss all aspects of privacy. We, however, chose to refer privacy as another synonym called 'the right to be let alone' [Bar17], 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 [JGK16]. 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" [JGK16]. Briefly, in this thesis, we focus on two issues of privacy: (1) the right to be forgotten; and (2) re-identification problems running on personal data. To the former, in our proposed frameworks, we can keep track of user data and hence are able to fulfill user's requests to erase their data. With regard to re-identification, we proposed both systematic architectures and privacy-aware algorithms to address re-identification problems. Along with the definition of privacy, it is also necessary 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'. Article 2(a) in the European Union Directive employs the following definition:

> "*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:

[††] gdpr.eu/right-to-be-forgotten/

> "*Personal data* means any information relating to an identified or 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 biometric data, the new GDPR's regulation defines *biometric data* 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 is *privacy* violated? Cynthia Dwork [Cyn06], who introduced *differential privacy* - the current state-of-the-art privacy-guaranteed approach by definition, has re-introduced the desideratum for statistical databases by Dalenius [Dal77]: 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. [KWG13] listed that privacy may be breached under following (but not all) circumstances:

- Personal information is combined with external datasets leading 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.

Table 1.1: Main differences between privacy and security [JGK16]

#	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 [Hu+16]

Up to this point, we have covered the nature of privacy and what issues related to privacy we should pay attention to. We summarize the following main issues among them that this thesis tries to address:

- Firstly, regarding 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) enabling data processing (e.g., data analysis, learning models, etc.) on sensitive data with privacy preservation.
- Thirdly, privacy and security are not always the same. Security can be used to strengthen privacy protection but security solutions (i.e., the classical cryptography methods) are not the direct solution to privacy protection.

1.3 Research Objectives

In the age of social media and expert systems, the fear of missing out (a.k.a FoMO) has attached people to many platforms such as search engines (Google, Yahoo!), or social network sites. The new term was so popular that Oxford Dictionaries add *FoMO* into their vocabulary together with *selfie* and others in 2013[†]. The activities on these platforms contain digital footprints of users. The stored information ranges from explicit information (e.g., demographic information) to implicit ones (e.g., special interest of users). From this, computer algorithms can learn and answer non-trivial questions such as *what kind of movies does a user like?* to *what is the personality of a user?*. In the following parts, we will introduce related privacy issues and the need for a *Privacy-Guardian.*

Privacy-Guardian:
the vital need in machine learning with big data.

The need to have a *Privacy-Guardian* is important because privacy is no longer a "mist", a potential threat but a real issue. Given enough user data, computer algorithms can give more insights about the user than we can imagine. Not in a science fiction book, but in reality, a computer algorithm now can predict personality of users better than those made by their friends or even their spouses [Far+16]. Thanks to the amount of data available on social media (e.g., Twitter), it shows the values of such platforms when they are used to make better choices, decision for us. For example, a better recommender system would save us much time in a process of searching before buying new products. Instead of manually reading all relevant products, a recommender can understand our shopping behaviors and our preferences based on our shopping history, to suggest potential items we want to buy. Another example is a personalized search engine. It knows exactly what results we are looking for based on our personal information. For instance, results of a query *what is python?* for a software engineer would be more about *programming languages* than *animals*. However, the highly sensitive information in the personal data puts users at risk of privacy breaches, e.g., ClearView[‡] or Cambridge Analytica[§]. In both events, a huge amount data of users on SNS was used to (1) understand and manipulate user political views (i.e., Cambridge Analytica) and (2) search profile of a random person using only a picture of their face. ClearView brings privacy concerns from facial recognition into focus since it could end anyone's ability to walk down the street anonymously. Although the tools from Cambridge Analytica or ClearView might bring many profits, privacy should have always be the first priority, to avoid bad influences to people's lives.

To address the privacy problem in machine learning with big data, we associate the problem in the following topics: (1) big data and data federation; (2) machine learning with big data; (3) privacy preservation for algorithms running on personal big data. Generally, these topics aim at answering the main question:

[†] `https://bit.ly/FoMO-Oxford-Dictionary`

[‡] `https://bit.ly/ClearView-Privacy-Issue`

[§] `https://bit.ly/Cambridge-Analytica-Scandal-Fallout`

How to protectively enhance personal data for research?

Drowning in the ocean of different choices making us rely on personalized platforms such as personalized search engines, personalized recommender systems, personalized news feed. Not only personal data is used for some "nice to have systems" that we can opt-out, but it is also invaluable for research in important domains such as research in health care, psychology, or social science, to support people's lives. Without the use of personal data (e.g., medical records of patients) many research and applications such as early disease prediction [HML04], or traffic jam prediction[¶], would not be possible. However, using personal data is undeniably a double-edged sword. One edge allows us to get deep knowledge on every choices the system can make, to increase user experience. But the other side of personal data might be termed as "losing the right to be left alone" - the fundamental need of privacy right. Losing this right means that we cannot choose to be invisible, to have privacy under our control. This is why this thesis focuses to address the vital need in privacy on learning with big data.

To describe the main question, there are important terms which need to be clarified specifically. First, *personal data* refers to the data belongs to a specific individual. It can be used directly or indirectly to re-identify that individual. *Personal data* is in fact, the most available data today in the age of social networks. This explains why protecting user privacy on social network data is a crucial topic. Second, to enhance personal data for research, any algorithms running on user data need to be aware of user privacy. Based on these main topics, we aim at two main research objectives as follows:

- Objective **RO1**: *Research on framework architecture to enable research on personal data.*
- Objective **RO2**: *Research on privacy-aware algorithms to address the privacy issues.*

1.3.1 Privacy-Aware Infrastructure

Regarding the **objective RO1**, we need to investigate two questions: (1) how to work with personal data without accessing the raw data?; (2) how to analyze data and publish results safely, with privacy-guarantee?

Answering the first question is a fundamental requirement since the best way to protect data privacy is to not distribute the data to any third parties. Having all data under the control within the originated infrastructure can reduce the misuse of the data and privacy leakage. However, user data is normally located in different locations and in different formats (e.g., textual data, media data, SQL format etc.). This makes the first question even harder to address. Moreover, allowing researchers to work on a unified framework with different analysis on multiple types of data is another big challenge.

[¶] `https://bit.ly/how-google-maps-knows-traffic`

In **Paper I** and **Paper II**, we worked on these two questions to allow flexible analysis can be performed on sensitive data on the proposed framework without accessing raw data. The proposed framework also supports privacy-guaranteed data analysis, to make sure analytical results are safe to be published. In many cases, privacy breaches can be found in even some simple statistical results such as counting. For example, hypothetically, there is a medical report of a city about the number of Coronavirus cases with a demographic histogam table. If it happens to show that, in an age group of 95 to 100, there is one case that has a positive test to the virus, it is possible that many people in this city will be able to figure out who this person is, since it rarely happens to have a person at that age range. This naive example shows that, even a simple histogram can have privacy leakage. In a complicated way, privacy leakage can be found in a pre-trained model such as the case of a trained facial generation model [FJR15]. The authors could use a hill-climbing algorithm to trace back and find whose face was used to train the generation model. Because of the mentioned issues, *privacy by design*[‖] has been getting lots of attention to be a must-have requirement for any personal data processing system.

In general, this section explains the main topics we focus in dealing with the way researchers can work on personal data. To further support more research to be done within the unified framework that we proposed, we investigate deeper into different privacy-aware algorithms to address privacy issues.

1.3.2 Privacy-Aware Algorithms

The main questions we investigated to discover different aspects of privacy-aware research in personal data are: (1) how to enable private, sensitive data to be used in research with privacy preservation?; (2) how to learn from multimodal data with privacy preservation?; (3) how to evaluate and select hyperparameters to have a good trade-off between data privacy and data utility?

Different terms related to privacy. There has been different research using different terms: privacy concern detection [VJ18], privacy concern analysis [Vu+19b; VTJ19], privacy awareness [Vu+20b], privacy preservation [Vu+20a], privacy guarantee [Cyn06; DS09; Vu+17b; Vu+19c]. However, there were not any articles which define these terms and their differences specifically. To this end, in the scope of this thesis, we define in what contexts these terms are used:

- *Privacy concern analysis* tries to analyze the privacy concerns of users, to understand and associate with other properties (e.g., the correlation between privacy concern, demographic, education level). **Paper II** [Vu+19a] contributes to this aspect. Similarly, [Vu+19b] also analyzes privacy concerns and their relation to politics, demographic using multilayer network analysis.

- *Privacy concern detection* is used to describe the process of detecting privacy concern levels based on *user preference* or *user data*. The former case

[‖] https://en.wikipedia.org/wiki/Privacy_by_design is user-driven privacy concern detection, and the latter is data-driven privacy concern detection. For user-driven, in many cases, it is not trivial to decide privacy concern level of user when we cannot have user consents directly. There is a phenomenon known as *privacy paradox* [Bd17]. The privacy paradox shows even if people are privacy aware in general, they do not behave according to their stated attitudes. Therefore, there are different methods (e.g., using questionnaire and hypothesis testing [YOU09]) used to naturally detect privacy concern level of users, instead of asking user preference directly. Regarding data-driven privacy concern detection, it refers to the sensitivity of different types of data in different domain. For instance, medical data is normally more sensitive, hence, privacy concern is higher than other data types, such as a *like* activity on social networks. To address this problem, we proposed an approach to automatically detect privacy concern level of users based on user data in **Paper III** [VJ18].

- *Privacy awareness*: this term describes an algorithm aware of privacy issues and introduce an approach to preserve user privacy with a relaxation. In many complex tasks such as text generation, it is not trivial to guarantee user privacy under differential privacy mechanism. The method we proposed in **Paper V** [Vu+20b] is one example. We proposed a learning approach to protect further use of user data by introducing a privacy controller module to learn user/entity embeddings with privacy-guarantee. These layers are also frozen during the training process, to avoid the further use of user data. However, the text generation task on multimodal data has several components including a visual model backbone (e.g., ResNeXt-50 [He+16], InceptionNetV3 [Sze+16]), a language model backbone (e.g., LSTM [HS97]), and other neural layers to learn user/entity representations for personalization. Therefore, it is very difficult for the whole model to be optimized with noise injection at the same time. This is why we consider not using dp-optimizers (e.g., dp-SGD [Aba+16b]), a relaxation in privacy preservation in order to achieve a good data utility for the task.

- *Privacy preservation*: this notion refers to a higher level of privacy protection, in which no sensitive information of user data is used before sending to the algorithms. The user's privacy is then preserved. However, it does not bound the chance of an adversary, who can exploit the output of the model and infer side information under differential privacy mechanism. The learning method we proposed in **Paper VI** [Vu+20a] is an example of this.

- *Privacy guarantee*: this is the highest level of privacy protection in which algorithms running on user data fulfill the requirements of differential privacy. The methods we proposed in **Paper IV** [Vu+19b] and **Paper I** [Vu+17b], are two typical examples of this privacy protection level.

- *Privacy aware algorithms.* Last but not least, this phrase refers to a line of algorithms dealing with privacy related issues. Unless otherwise specified, the phrase *'privacy-aware'* acts as a prefix, to refer to an algorithm or an application works on privacy in general.

To categorize these research questions into research objectives, we put them into the same objective called *privacy-aware algorithms.* This thesis focuses on two type of algorithms: (1) privacy-aware data analysis, and (2) privacy-aware learning algorithms. The former covers analytic algorithms that run on user data, to discover statistical information for reporting and researching purposes (e.g., histogram query in **Paper I** [Vu+17b], privacy concern analysis in **Paper II** [Vu+19a]). The latter includes learning algorithms that process user data to find patterns for a downstream task such as classification (**Paper III** [VJ18]), learning representation (**Paper IV** [Vu+19b]), text generation (**Paper V** [Vu+20b]), multimedia tagging (**Paper VI** [Vu+20a]), evaluation (**Paper VII** [Vu+19c]).

We divide the research objective **RO2** (*Privacy-aware algorithms*) into four main research objectives to better focus on different aspects of privacy in personal data:

RO2-a: *Learning for privacy-guaranteed representation for data sharing.*

RO2-b: *Learning on multimodal data with privacy preservation.*

RO2-c: *Providing privacy utilities for getting deeper insights of user's privacy concern.*

RO2-d: *Providing privacy utilities for evaluating privacy-aware algorithms, to have a good trade-off between privacy and data utility.*

The **objective RO2-a** stays in the research area of learning privacy-guarantee representations for data sharing [BB09] since it is very valuable for research to have information from sensitive-register data**. One example is in the context of medical text data. It is very sensitive, however, if the representation from a sensitive corpus is available, many related works can be improved, such as ICD Coding from Clinical Text [VNN20]. It is noted that, *"ICD coding is a process of assigning the International Classification of Disease diagnosis codes to clinical/medical notes documented by health professionals (e.g. clinicians)"* [VNN20]. The ICD coding process is very costly and it requires significant human resources. Therefore, having the information from sensitive medical texts can significantly improve the performance of downstream tasks on public dataset. There have been different research methods working on this direction such as RAPPOR [EPK14], which tries to add noise into a collected data to protect data privacy of users before sending out to any algorithms.

**Register data is *by-product* of registers held for administrative purposes. They are mostly dedicated for governmental planning (and research), such as population level surveys (e.g., censuses), or cause of cancer data.

McMahan et al. introduced DP-RNN [McM+17], a differential private model for learning user-level language models in a federated manner. By federated user-level model on a local device, it allows to train a global model based on user-level data without uploading sensitive user data to a centralized server. In **Paper IV** [VTJ19], we contributed to this direction by introducing a language model in the context of centralized personal data, in which sensitive data are already located in a central database and there is a need to learn a privacy-guaranteed representation on the data. In the paper, we showed that having privacy-guaranteed information from a sensitive corpus does help a downstream task to gain better performance. Similarly, the same approach could be used to boost performance of related task in health care when having a privacy guaranteed representation from medical records. Because many disease names with personal information would not be seen elsewhere but in the medical text data.

The **objective RO2-b** addresses the big issue in the information age called *multimodal data*. On many social network sites, users often interact and share multimodal content, which include multiple data types such as a post with status and picture. Moreover, user activities on social media and their network connections together form a graph network data. It means that, any algorithms working on user data should also be aware of multimodal data. Working on this objective, we address the privacy of multimodal data with privacy preservation on two problems: (1) privacy in text generation (**Paper V** [Vu+20b]) and (2) privacy in multimedia tagging (**Paper VI** [Vu+20a]). In **Paper VI** [Vu+20a], we introduced a new learning architecture for multimedia tagging problem by integrating the use of global knowledge and avoiding the use of local knowledge with two mechanisms. First, to avoid of sensitive visual information such as faces, plate numbers, ID numbers, these information are censored by a preprocessing step, to avoid privacy leakage. Second, links between entities in the dataset are sometimes very sensitive. Therefore, we introduce a differential privacy graph (dp-graph) construction method to create a privacy guarantee adjacency matrix between target tags of the task. The dp-graph will guide the learning process to avoid the use of sensitive links. Altogether, different components allow our new architecture perform better than previous state of the art on the same benchmark data (i.e., MS-COCO).

The **objective RO2-c** targets at providing *privacy utilities* for understanding privacy concerns of users. In **Paper III** [VJ18], we worked on this problem to automatically detect privacy concern of unknown users, and protect their data privacy as if they were given their consents. Similarity, in **Paper II** [Vu+19a], we provided a set of tools for analyzing privacy concern of users based on association rules. From which decision makers (e.g., service providers, data controllers) can introduce new change in their system, to increase user satisfaction in terms of privacy policy.

Last but not least, the **objective RO2-d** aims at providing *privacy utilities* for the evaluation problem of privacy-aware algorithms. This topic has been raising attentions regarding the aim of comparing between different models on a downstream dataset. However, it is not always trivial to tell one model is better than others without the involvement of human-in-the-loop. For instance, learning a language model is not a direct task, but a middle process to achieve better performance on a task. Similarly, learning to generate visual images (e.g., GAN [Goo+14]) cannot easily be evaluated automatically. Therefore, it is even more difficult to compare between privacy-guarantee models versus conventional models. In **Paper VII** [Vu+19c], we proposed a systematic approach with visual aid to compare between different language models, to find a good setting for a downstream task. It is also proved to be useful to efficiently select hyper-parameters to have a good trade-off between privacy and data utility. The proposed evaluation approach helps to achieve a new state-of-the-art result on a well-known task (i.e., name entity recognition task) at the time of publication.

In summary, this section introduces the notion of privacy, research objectives, and main ideas of the papers equipped in this thesis.

1.4 Methodologies

1.4.1 Privacy-Aware Infrastructure

To work on **the objective RO1** in *Privacy-Aware Infrastructure*, we start our investigation on how to build a framework with data federation to (1) facilitate researchers to work on sensitive data without tedious and lengthy application procedure; and (2) avoid potential data breach during research process and publishing research findings. The problem is associated with two big topics: (1) data federation [Inc16] and (2) privacy guarantee data analysis [JGK16; Meh+16; Cyn06].

To meet the above objectives, we worked on both research topics in (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 [McS09] or GUPT [Moh+12], however, they mainly act as out-of-the-box solutions to traditional database systems to achieve privacy protection but not for data federation systems. Secondly, it lacks of practical privacy-aware algorithms in real-life system, which can easily be used by end-users (e.g., a psychology researcher), that can efficiently address privacy issues to protect personal data. In [JLE14], 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 [FJR15; Aba+16a; Pha+17; Pap+18b; Wu+18], however, they either try to (1) address privacy issues on image datasets [FJR15; Aba+16a; Pap+18b; Wu+18] (because adding noise to protect privacy for images is easier than that for heterogeneous data) [VTJ19] or (2) address privacy issues on some selected properties [ZDS18] but not for centralized data collection. These limitations are addressed in more detail in **Paper IV** [VTJ19].

1.4.2 Privacy-Aware Algorithms

We research on both methods for (1) privacy preservation for single modalities (e.g., text, image, audio) and (2) privacy-awareness in learning multimodal data. Staying on the ground of research to improve existing methods in privacy preservation on single modality data, we could investigate deeper to address new challenges and solutions for privacy-aware learning on multimodal data.

Privacy Preservation

In privacy preservation, we investigated into different methods to show the frameworks' effectiveness and cross-disciplinary utility regarding privacy in data processing.

Privacy protection can be divided into two methods namely (1) data sanitization and (2) anonymization. In 2008, Narayanan and Shmatikov [NS08] proposed an effective de-anonymized algorithm to break privacy of Netflix Prize Contest [Net09]. In 2009, there was a very subtle privacy violation when Wang et al. [Wan+09] showed how published GWAS (Genome-Wide Association Study) results revealed whether specific individuals from the study were in cancer group or healthy group [Vis+12]. 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 1.3) including (1) input perturbation [Blu+05], (2) output perturbation [Dwo+06], (3) internal/objective perturbation [WCX19; CH11], and (4) sample-and-aggregate [NRS07]. Later in 2006, Dwork firstly introduced differential privacy (DP) in her ICALP paper [Cyn06] 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 [McS09; Moh+12]. Input perturbation is more common in data curation [EPK14].

Differential privacy [Cyn06] 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: (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 [I+15] 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 [Moh+15]. However, DataSHIELD only can answer single one-dimensional statistics, which do not always satisfy 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 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. [EPK14]). And this privacy-guarantee modification makes the learning part be more difficult. Ji et al. addressed in [JLE14] that general ideas of privacy-preserving machine learning algorithms are learning a model on clean data, then use exponential mechanism [Vai+13; MW09; Sal+11] or Laplace mechanism [Vin12; CSS12] 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 3.4.1 of Chapter 3 we also address some evaluation problems of privacy-guarantee machine learning models.

Privacy-aware multimodal learning

Concerning the digital footprints of users when they use social network sites, a huge amount of multimodal data is generated at the same time. For instance, a video uploaded to Youtube.com would contain user speech, user images, and their text data in the description. Similarly, for every like/comment on Facebook or Twitter, users are adding information to the graph of a graph network, which can tell which friends are more closed to the user than others. Using these multimodal data can boost performance of personalized systems, to understand user behaviors better such as recommender system [SFC18], or personalized search [Ngu+19]. However, from the privacy-aware perspective, it is more challenging to address privacy-guarantee to these models.

Multimodal data is associated to multimodal features for challenging tasks such as lipreading. Because it is not immediately clear what the appropriate visual features that should be extracted for the task [Ngi+11] are. In fact, it should be a multimodal features with the combination of both speech features and visual features. Figure 1.4 shows a multimodal review on Yelp.Com with two photos, user review (text).

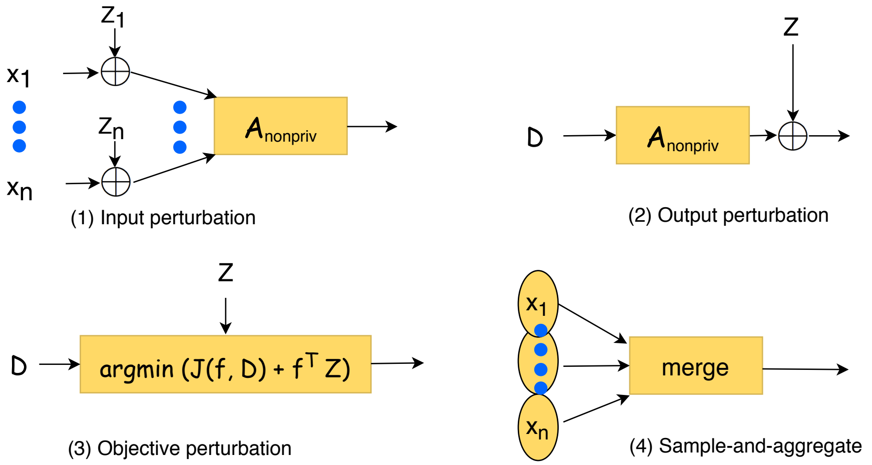
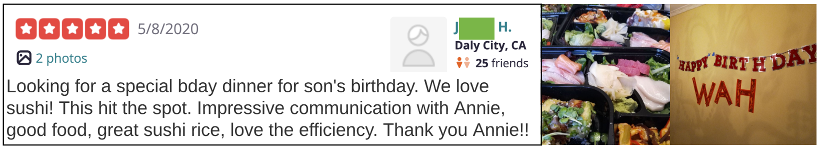


Figure 1.3: Common privacy protection approaches: (1) *input perturbation* adds noise to the input before running algorithm; (2) *output perturbation* algorithm adds noise to the results; (3) the *internal/objective perturbation* randomizes the internals of algorithm, and (4) the *sample-and-aggregate* computes query on disjoint subsets data and then uses differentially private method to select max.



★★★★★ 5/8/2020

2 photos

J[redacted] H.
Daly City, CA
25 friends

Looking for a special bday dinner for son's birthday. We love sushi! This hit the spot. Impressive communication with Annie, good food, great sushi rice, love the efficiency. Thank you Annie!!

Figure 1.4: An example of multimodal data (text, images): an online review on Yelp.Com with personal information in **Paper V** [Vu+20b].

A naive method is to extract a unique representation vector from multimodal features and then apply existing privacy-guarantee algorithm for single modality feature. However, because of the noise added to the learning model during the training process, with the complexity of multimodal data, it would make the whole model be very difficult to be optimized. To address this, in **Paper VI** [Vu+20a] and **Paper V** [Vu+20b], we propose new neural architectures to deal with privacy for multimodal data. **Paper V** [Vu+20b] introduces different modules for privacy-aware learning, while **Paper VI** [Vu+20a] incorporates global knowledge into the learning process, to reduce the use of sensitive information in local data.

1.5 Research Contributions

Generally with the two objectives in this thesis, objective **RO1** targets at finding different system architectures to re-construct personal data from which, they support the objective **RO2** to enable research in sensitive data by providing (1) privacy utilities and (2) privacy-aware algorithms to protect user privacy.

1.5.1 Contributions to Privacy-Aware Infrastructure

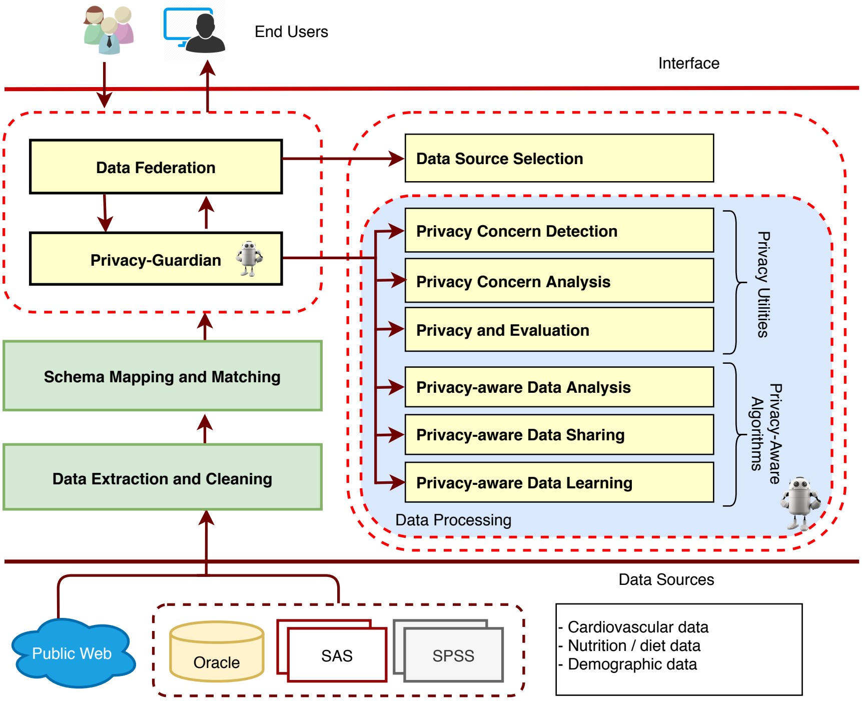


Figure 1.5: Architecture Design of Privacy-Aware Data Federation Frameworks, where the red-dash-line parts are the main focuses of this thesis.

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 1.5, in which 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).

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 [Inc16; SL90], based on which, the main scientific problem we address here is the data source selection:

- 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, we proposed a rule-based approach to find related variables on a virtual database from which the system selects correct variables of interests out of the original data sources for data analysis. In **Paper I** [Vu+17a] and **Paper II** [Vu+19a], we applied this approach to build two open-access frameworks. These frameworks allow researchers to work on register data that would otherwise is not easy to access and perform privacy-guaranteed data analysis.

1.5.2 Contributions to Privacy Research

The *Privacy-Guardian* module provides *Privacy Utilities* and *Privacy-Aware Algorithms* to enable research on personal sensitive data. It focuses on balancing the needs of researchers (who want to pursue scientific research) as well as data donors (who want to protect privacy of their data).

- Privacy Concern 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 [Moh+12]. We contributed to this aspect in **Paper II** [Vu+19a] and **Paper III** [VJ18].

- 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 [Cyn06]). We contributed to this topic in **Paper I** [Vu+17b].

- Privacy-Aware Data Sharing: to protect privacy of a dataset before sharing to third-parties. There have been different privacy-guarantee algorithms for data sharing such as K-anonymity [Sam01; SS98], L-diversity [Mac+06], t-Closeness [LLV07]. However, most of them are vulnerable

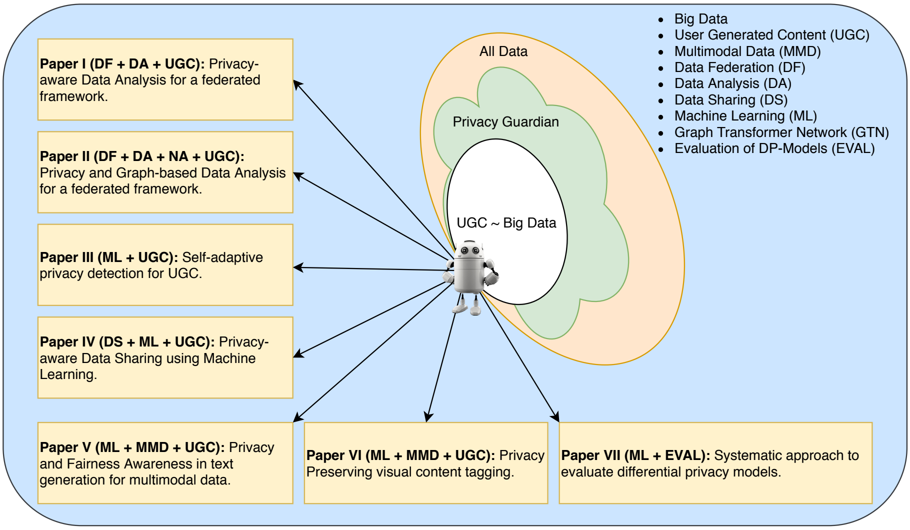

Big Data
User Generated Content (UGC)
Multimodal Data (MMD)
Data Federation (DF)
Data Analysis (DA)
Data Sharing (DS)
Machine Learning (ML)
Graph Transformer Network (GTN)
Evaluation of DP-Models (EVAL)
All Data
Privacy Guardian
UGC ~ Big Data
Paper I (DF + DA + UGC): Privacy-aware Data Analysis for a federated framework.
Paper II (DF + DA + NA + UGC): Privacy and Graph-based Data Analysis for a federated framework.
Paper III (ML + UGC): Self-adaptive privacy detection for UGC.
Paper IV (DS + ML + UGC): Privacy-aware Data Sharing using Machine Learning.
Paper V (ML + MMD + UGC): Privacy and Fairness Awareness in text generation for multimodal data.
Paper VI (ML + MMD + UGC): Privacy Preserving visual content tagging.
Paper VII (ML + EVAL): Systematic approach to evaluate differential privacy models.

Figure 1.6: (Algorithm Perspective). Overview of the seven papers and their connections to Privacy-Aware Data Federation and UGC (or Big Data). Each paper is detailed in a square box with the main objective and relevant research topics that were addressed in it.

to privacy attacks which will be discussed further in this section. In **Paper IV** [Vu+19b], we introduced a privacy-guarantee algorithm to learn representation for data sharing.

- Privacy-Aware Machine Learning: to preserve user privacy while learning models for downstream tasks [Vu+20b]. In details, we contributed to this topic in visual tagging (**Paper V** [Vu+20b]) and text generation (**Paper VI** [Vu+20a]).

1.5.3 Summary of Contributions

This thesis contributes to knowledge in (1) privacy-aware data federation, (2) privacy-aware algorithms within the context of the research objectives. Figure 1.6 and Figure 1.7 show the overview of research contributions of this thesis. Figure 1.6 represents the seven papers and their main research topics in relation to *big data*. Figure 1.7 shows the contributions of all seven papers in the context of different data - i.e., from sensitive data to public data. For instance, to use register data for improving downstream tasks on public data, one can apply the contributions in **Paper IV** [Vu+19b] to obtain privacy-guarantee embeddings, which can be used to improve the tasks.

Here we discuss the contributions of each paper with respect to different research objectives. For the objective **RO1**, we proposed two different system

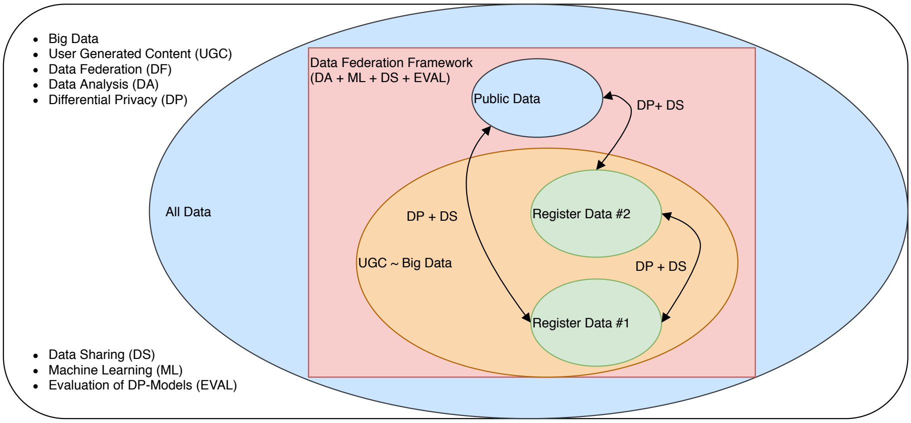


Figure 1.7: (Data Perspective). Overview of Data Federation Infrastructures with respects to different types of data and features that are supported in this thesis. Each oval shape represents a type of data, and the connection denotes related algorithms are required for privacy-guarantee.

architectures to both guarantee privacy and have a good trade-off between privacy and data utility in **Paper I** [Vu+17b] and **Paper II** [Vu+19a]. Regarding **RO2**, we proposed different privacy-aware algorithms to address different privacy issues in personal data. Firstly, to prevent inference attacks on personal data (i.e., RO2-a), we proposed methods to limit how much information can a random algorithm learn from the data to satisfy the definition of differential privacy in **Paper IV** [Vu+19b]. Secondly, to address *privacy-aware learning* on multimodal data (i.e., RO2-b), in **Paper V** [Vu+20b] and **Paper VI** [Vu+20a], we introduced new learning architectures to limit the use of sensitive data, to better protect user privacy. Regarding *privacy utilities* of research objective RO2-c, in **Paper II** [Vu+19a] and **Paper III** [VJ18], we contributed a set of tools and algorithms to analyze and recognize privacy concern of users based on their data. Lastly, the research objective of *privacy utilities* for evaluating privacy-aware algorithms (RO2-d), in **Paper VII** [Vu+19c], we proposed a systematic approach to evaluate and select a good set of hyper-parameters, to balance between privacy and data utility. In summary, the main contributions of this thesis in the seven 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.

- Propose *privacy utilities* for detecting and understanding privacy concern of users, to better protect their data.

Generally, this section describes the contributions of the seven equipped papers in two main research areas namely *privacy-aware infrastructure* and *privacy related research.* For each area, we detailed how different paper contributed to the relevant research topic. Especially, we visualized the contributions of all papers with *algorithm perspective* (see Figure 1.6) and *data perspective* (see Figure 1.7). These figures help to summarize the key algorithms we contributed and also, how they can be applied in different type of data.

1.6 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 presents privacy-aware infrastructures. It describes comprehensive background of three different topics including privacy-aware data federation, privacy-aware data sharing, and privacy-aware data analysis. These three topics are highly correlated to the objective **RO1**. In each topic, we highlight relevant contributions that this thesis contributed. Afterwards, two emerging challenges in privacy-aware infrastructures are discussed in details. The first challenge will be *privacy-aware for heterogeneous and distributed data.* And the second challenge will be *the relation between privacy preservation and scalability.* To summarize this chapter, we briefly describe the main contents and what will be discussed in the following chapters.

Chapter 3 discusses privacy issues in machine learning models, which are the main topics of the objective **RO2**. It describes a normal machine learning process and then, showing different privacy attacks against the machine learning process. In a subsection called *Privacy-Aware Machine Learning*, we will summarize a comprehensive list of privacy-aware algorithms which will consist of both traditional machine learning algorithms and deep learning algorithms. Related to challenges, we will introduce two emerging challenges in privacy-aware machine learning including *evaluation* and *ethical issues.* Along each subsection, we will highlight related contributions that this thesis made in relevant topics. The summary part of this chapter will conclude what have been discussed and open some future topics for discussion.

In chapter 4, we will summarize all research papers included in the thesis, and the contributions of the PhD candidate in each work. Moreover, we will briefly introduce each paper and what contributions were made in each paper in relation to the research objective of the thesis. Lastly, we will address 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 applications in a federated manner.

2.1 Privacy-Aware Big Data

Big data [KAH15] 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 [Meh+16]. It can also solve more challenging problems of society like in healthcare (e.g., disease forecasts [JDB14], quantifying mental heaths [CDH14]). It can also help to promote the scientific research and economy [Meh+16]. Despite the benefits we can achieve from using big data to understand the world in various aspects of human endeavors, it raises 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 data privacy and the benefit from big data, many studies are focusing on this direction to address this new challenge [Cyn06; DS09; Vu+17a; Vu+19b; VTJ19]. To name a few, in order to ensure big data

[*] 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`

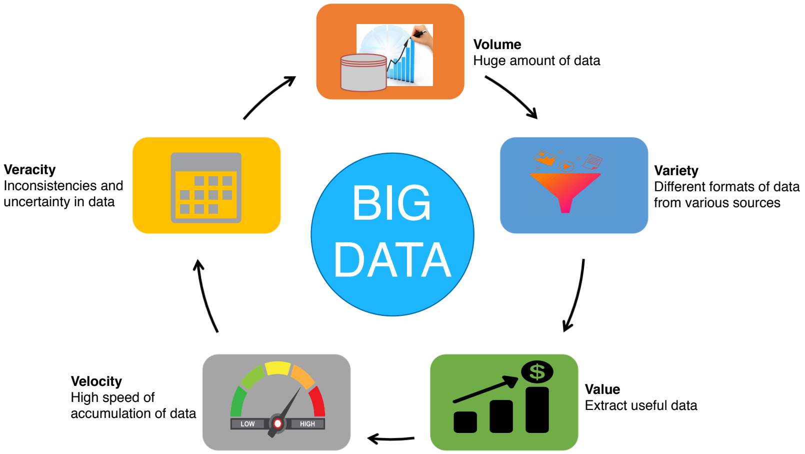

Volume
Huge amount of data
Variety
Different formats of data from various sources
Value
Extract useful data
Velocity
High speed of accumulation of data
LOW
HIGH
Veracity
Inconsistencies and uncertainty in data
BIG DATA

Figure 2.1: The properties of big data are reflected by 5V's, which are veracity, volume, variety, value, velocity [Tsa+15]. These 5Vs are very important for a privacy-aware infrastructure that supports data analysis on both public and sensitive data.

privacy, various mechanisms have been developed in recent years including K-Anonymity [SS98], L-diversity [Mac+06], t-Closeness [LLV07], and differential privacy [Cyn06; DS09]. In general, these mechanisms can be grouped based on the stages of big data life cycle [Meh+16], i.e., data generation, storage, and processing.

1. **Data Generation**: Data can be generated from various distributed sources [Meh+16]. Privacy research topics in relation to this process are access restriction [Xu+14] and falsifying data [Xu+14].

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 [Cao+14] and privacy [Sou+14], one promising technology to address these requirements is storage virtualization, enabled by the emerging cloud computing paradigm [MG11].

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 1.5.2.

Going into details of each V in the $5Vs$ Big Data, Table 2.1 correlates characteristics of each V with privacy issues. In generally, each of $5Vs$ introduces new challenges in protecting user privacy. Among them, *volume*, *variety*, *value* were addressed in the seven papers equipped in this thesis.

Table 2.1: The correlation between *5Vs* in Big Data and Privacy. For each V, there is a description of it and also how it can be an issue regarding privacy.

5V's	Description	Privacy issue
Volume	Huge amount of data	User data is huge and it requires more resources and human involvements to process, which might be the sources of privacy leakages.
Variety	Different formats of data from various sources	User data is stored in all different formats and located everywhere.
Value	Extract useful data	The extraction process can potential have privacy leakages.
Velocity	High speed of accumulation of data	Privacy-aware algorithms also need to adapt to this speed.
Veracity	Inconsistencies and uncertainty in data	Privacy-aware algorithms also need to handle inconsistencies and uncertainty in data.

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 [I+16]. 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, they both 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 source. 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 one hand, it must hide information about

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

data subjects. On the other hand, for the released data to be useful, it should be used to learn knowledge on specific domains without worrying about privacy leakages. 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 [SS98] was one of the most popular methods and various different algorithms implement this technique such as [Swe02; Mac+06]. 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 [Swe02]. Some studies (e.g., LeFevre et al. [LDR08]) propose algorithms that support the generation of anonymous views based on a specific work-load focus. The others (e.g., Xiong [XR08]) 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 [BB09] 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.

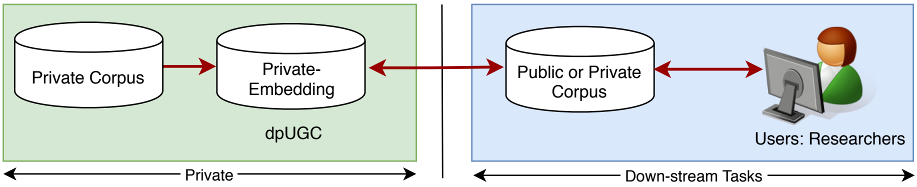
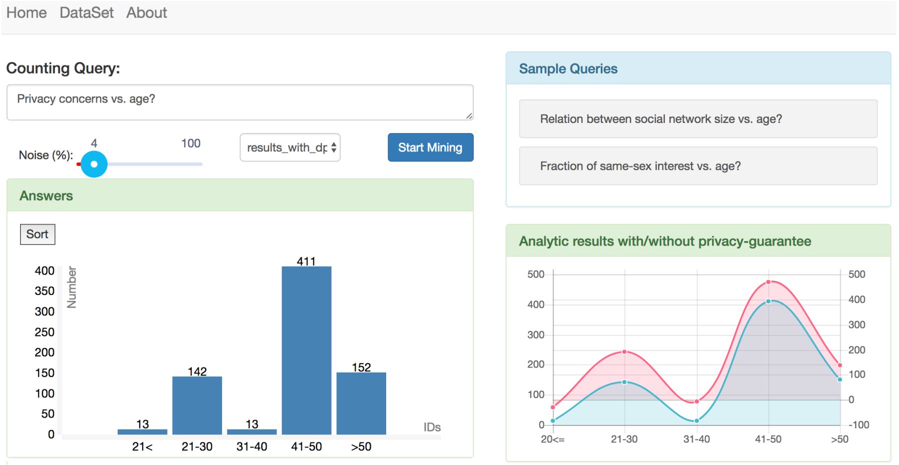


Figure 2.2: Overview of our proposed *safe-to-share* embedding model in **Paper IV** [Vu+19b]. Using the proposed approach, the pre-trained word embedding set can be shared to facilitate research on sensitive data with privacy-guarantee.

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 **Paper IV** [Vu+19b].



Counting Query: Privacy concerns vs. age?

Noise (%): 4 100

results_with_dp

Sample Queries

Relation between social network size vs. age?

Fraction of same-sex interest vs. age?

Figure 2.3: An example of privacy-guarantee histogram (the red line) in **Paper I** [Vu+17a]. In general, privacy-guarantee histogram keeps the overall information of the population while it masks the chance to re-identify individual information.

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 individual from the data. Figure 2.3 presents an example showing privacy-guarantee histogram versus raw histogram, from which we can see their differences. While the main trend of the statistic is the same, the privacy-guarantee histogram prevents the chance to re-identify any individual in the result. Naturally, one might have this question in mind: "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 [Cyn06], 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 reduce the chance to re-identify any other information.

According to Dwork [Cyn06], 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 for calculating privacy-guarantee results. However, this setting is very 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. We also discuss here on how to effectively scale the federated system with privacy-guarantee since it is an emerging challenge.

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

- **Privacy-aware 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 [Du+18].

- **Privacy-aware representation learning**: 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 federated learning**: similarly to the above challenge, in federated learning [McM+17], 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** [Vu+19a] 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.

In summary, this chapter addresses privacy-aware infrastructures for (1) big data, (2) data federation, (3) data sharing, and (4) data analysis. Since big data and data federation are strongly connected, there should be more research in both algorithms and systems to establish new standards for privacy-aware infrastructures running on big (and or federated) data. For *data sharing*, it is undeniably important for future research in which it could open many possibilities to facilitate cross-domain studies thanks to the combined information from public and register data. Regarding the *data analysis* process, it is a gateway between researcher and data. When the data analysis process is designed to protect privacy (e.g., embedded in a privacy-aware infrastructure), it can open up a lot of new research directions that researchers could not have done before due to lack of access to the register data. Afterwards, we also described two main challenges for privacy-aware infrastructures in order to adapt to the complexity of big data.

Chapter 3

Privacy-Aware Machine Learning

In this chapter, we talk briefly about Machine Learning (ML) from which, we address related problems in learning privacy-guaranteed representations. From application's perspective, ML and Big Data are used to enable new technologies in smart information systems (SIS) (i.e., information systems with the use of ML as the core solutions) [SW18]. However, the traditional machine learning algorithms were not designed with privacy in mind. Therefore, we start with describing a standard machine learning process. Afterwards, different privacy attacks and privacy-guaranteed algorithms are discussed.

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 may cause 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 from CMU[†] Professor Tom Mitchell [Mit97b]:

> "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.

[†] Carnegie Mellon University

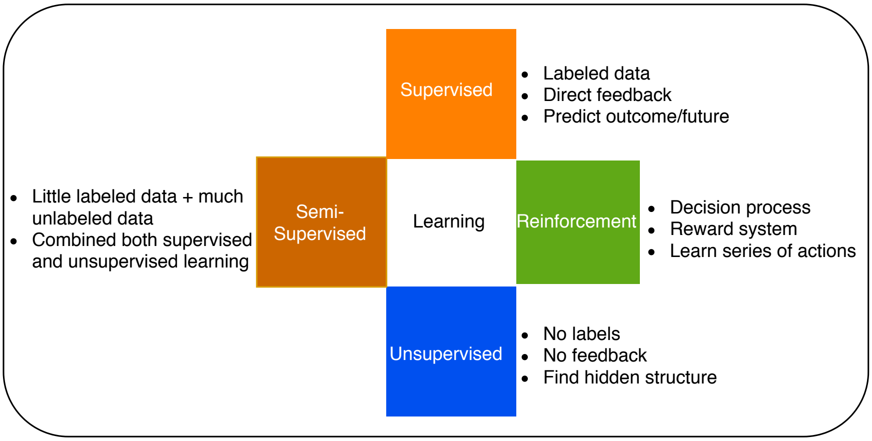


Figure 3.1: Four popular types of Machine Learning paradigms including (1) supervised ML, (2) unsupervised ML, (3) semi-supervised ML, and (4) Reinforcement learning.

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 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 samples, each of which has a list of attributes (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 samples, each of which has a list of attributes (features), and is 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 [Zhu08].

4. **Reinforcement learning models**: interact with an environment, so there is a feedback loop between the learning system and its experiences. This line of algorithms are not the focus of this thesis. However, interested audiences are recommended to see Sutton and Barto [SB18] for detailed 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 type 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 [HS06; HOT06].

Machine Learning Process

Figure 3.3 details different modules of a machine learning process inspired by *schematic of a typical deep learning workflow* of Raghu and Schmidt [RS20]. For clarity, we call each of big *red box* is a module, hence, we have there modules including (1) Data Preparation Module, (2) Learning Module, and (3) Inference Module. Inside each module, there are different steps in *blue boxes.* For instance, the *Inference Module* has *Infer* as the main step for analyzing model's performance and serving for downstream applications. It is noted that, different from Raghu and Schmidt [RS20], *Infer* and *Serving* steps are added to describe a standard machine learning process in practice because of two reasons. First, any steps in the *Inference Module* needs to base on the *Infer* step in which, pre-trained models from *Learning Module* are used for the validation, analysis, and serving steps. Second, *Serving* step is typically mandatory for any AI-base solutions to be used in real applications. For example, a breast cancer prediction application needs to serve a downstream application (e.g., a doctors' aid system) by processing a breast X-ray image as an input to classify it to either *malignant* or *benign.*

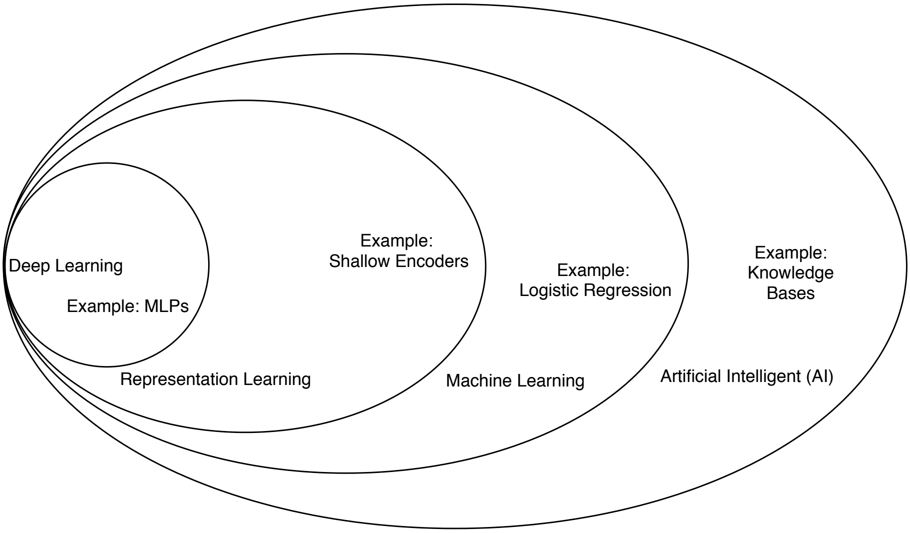


Figure 3.2: A Venn diagram showing how deep learning is a kind of representation learning by Goodfellow et al. [GBC16].

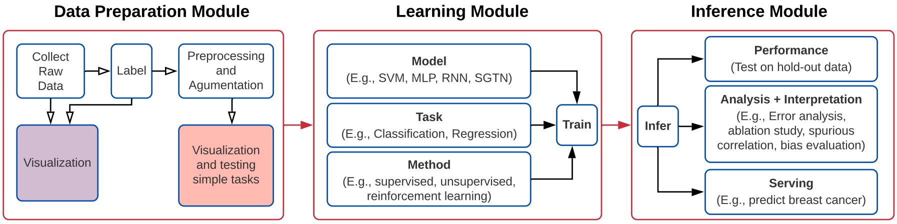

Data Preparation Module
Collect Raw Data
Label
Preprocessing and Agumentation
Visualization
Visualization and testing simple tasks
Learning Module
Model
(E.g., SVM, MLP, RNN, SGTN)
Task
(E.g., Classification, Regression)
Method
(E.g., supervised, unsupervised, reinforcement learning)
Train
Inference Module
Infer
Performance
(Test on hold-out data)
Analysis + Interpretation
(E.g., Error analysis, ablation study, spurious correlation, bias evaluation)
Serving
(E.g., predict breast cancer)

Figure 3.3: A machine learning process inspired by *schematic of a typical deep learning workflow* of Raghu and Schmidt [RS20]. Here we added *Infer* and *Serving* steps to demonstrate a standard machine learning pipeline in practice.

To understand better the ML process, here we go into details of training and inference steps formulated by Papernot et al. [Pap+18a].

- **Training step**. Most* ML models can be seen as parametric function $h_\theta(x)$ taking an input x and a parameter vector θ. A learning algorithm learns from training data to find the values of parameters θ. For supervised learning, the parameters are adjusted in such a way to reduce the gap between model predictions $\hat{y} = h_\theta(x)$ and expected outputs y indicated by the dataset. In deep learning, the gap is measured by loss functions, such as *mean-square-error* (MSE) for a regression problem. Then $\mathcal{L}_{MSE}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2)$. Here N is the total number of training samples.

*Though some models are non-parametric such as the nearest neighbor.

- **Inference step**. Once training completes, the trained model is used to infer on unseen data, hence, the values of parameters h_θ are fixed. Given a new input x, the model computes $h_\theta(x)$. If the task is a regression problem, $h_\theta(x)$ returns a real value.

In general, this part describes a standard ML process. Next, we will discover different types of privacy attacks can be performed to exploit privacy leakages out of machine learning based applications in SIS. Accordingly, different types of privacy guarantee algorithms are also introduced with respect to different process of the pipeline.

3.2 Privacy Attacks against the ML Process

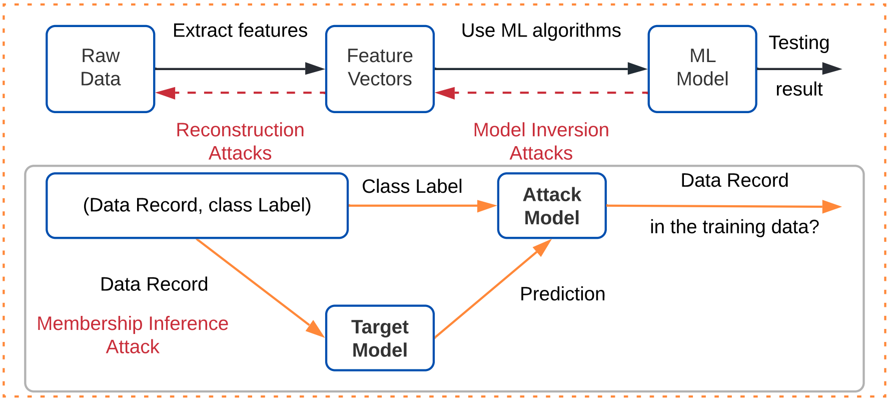


Figure 3.4: Different privacy attacks by Al-Rubaie and Chang [AC19].

Privacy attacks and privacy-guarantee approaches are normally inline to the machine learning process. From an adversary's perspective, understanding a typical learning process, one could exploit loopholes from which, he/she can design different attacking methods to exploit privacy leakages. Similarly, in order to guarantee user privacy, researchers need to understand both how ML process works and how privacy attacks are performed. To target the privacy of a SIS, adversaries commonly interested in recovering information about the training data or the learned model [Pap+18a]. Typically, there are four popular privacy attacks, in which three of them are shown in Figure 3.4, against a ML process addressed in [AC19] as follows.

- Reconstruction Attacks: "the adversary's goal is reconstructing the raw private data by using its knowledge of the feature vectors".
- Model Inversion Attacks: here, "the adversary's target is creating feature vectors that resemble those used to create an ML model by utilizing the

responses received from that ML model. Such attacks utilize the confidence information (e.g., probability or SVM decision value) that is sent back as a response for testing samples submitted by the results party."

- Membership Inference Attacks "aim to determine if the sample was a member of the training set used to build this ML model (adversary's target)". Commonly, a *meta-classifier* (or sometime is called a *shadow model*) is used to observe outputs of a model of interest h_θ to check if a certain individual's record was used to train h_θ based on a certain statistical property. This type of attacks is categorized as *Black-box attack* [Pap+18a].

- De-anonymization (Re-identification): "anonymization by removing personal identifiers before releasing the data to the public may seem like a natural approach for protecting the privacy of individuals.". Hence, re-identification attacks aim at de-anonymous user information to identify what individual was involved in the pretrained model or a data collection.

Projecting our equipped papers to the ML process and privacy attacks, we can summarize how different papers contribute to these topics. In fact, this thesis tries to address privacy-guarantee issues covering all three main processes (shown in Figure 3.3) of a ML process as follows.

- **Data**: We introduce different approaches to protect the user data. In **Paper I** [Vu+17b] and **Paper II** [Vu+19a], we proposed different privacy-aware infrastructures to work on sensitive data. Paper IV introduces an approach to guarantee privacy for data sharing.

- **Learning**: In **Paper V** [Vu+20b] and **Paper VI** [Vu+20a], we designed new learning models to preserve user privacy while optimizing for downstream tasks. More importantly, these papers address an emerging direction in protecting user privacy in multimodal data.

- **Validation & Analysis**: In **Paper VII** [Vu+19c], we worked on the evaluation approach to propose a systemaic approach for selecting good hyper-parameters to balance between privacy and data utility. Moreover, in **Paper I** [Vu+17b], **Paper II** [Vu+19a], and **Paper III** [VJ18], we showed different *privacy utilities* to analyze privacy concerns of users.

Regarding privacy, it is noted that all seven papers focus on de-anonymization (re-identification) aspect of privacy. Additionally, in **Paper VI** [Vu+20a], beside the contribution to re-identification aspect, we proposed the use of global knowledge and local privacy-guaranteed knowledge to minimize the possibility of executing membership attacks against the proposed model.

3.3 Privacy-Aware 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 contained in training data. Table 3.1 shows a list of traditional algorithms which already have privacy-aware versions, to adapt to the urgent needs in privacy preservation.

Table 3.1: List of differentially private models. Here, *Deep Learning* was put into a separated paradigm since its architecture is flexible and can be used to train learning models in supervised or unsupervised manners.

Paradigm	#	Privacy-aware models
Supervised	1	DP-Naive Bayes [Vai+13]
	2	DP-Linear Regression [Zha+12]
	3	DP-Linear SVM [WCX19]
	4	DP-Logistic Regression [XYW19]
	5	DP-Kernel SVM [Rub+09]
	6	DP-Decision Tree Learning [FS10]
	7	DP-Online Convex Programming [JKT12]
	8	DP-K-nearest neighbours (KNN) [Gur+17]
Unsupervised	9	DP-K-means [NRS07]
	10	DP-Feature Selection [Vin12]
	11	DP-Principle Component Analysis (PCA) [HR13; KT13]
Deep Learning	12	DP-Differential Private Stochastic Gradient Descent(dpSGD) [Aba+16b]
	13	DP-Convolutional Neural Network with differential privacy [Lec+18]
	14	DP-recurrent language models [McM+17]
	15	DP-Word2Vec (dpUGC) [VTJ19], dpSENTI [Vu+20b]
	16	Private Aggregation of Teacher Ensembles (PATE) [Pap+18b]
	17	And many others [Pha+17; FJR15; Aba+16a; Wu+18; NIR16; ZDS18; Pop+18]

Differential privacy in Machine Learning

As mentioned in Chapter I, subsection 1.5.2, differential privacy (DP) is currently the 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** [Cyn06; DS09; LC11; LC12] 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 given the assumption that 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 $\mathcal{M} : D \rightarrow \mathbb{R}^d$ 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 (ϵ, δ)-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 $\forall \mathcal{S} \subseteq$ 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 as shown in Table 3.1 invented to fulfill the definition.

How to apply *differential privacy* 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 [DR14]). 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. [Pha+17] 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 Deep Learning

Deep learning is a kind of representation learning [GBC16]. 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 [Kow+18], genome prediction [BCP18]. 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 [Aba+16a]. 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. [Aba+16a]: 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 [Cyn06], 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 [DR14] 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. [Pap+18b], 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 1.3 earlier. It is noted that, there is an emerging approach called Secure Multiparty Computation (SMPC) for training deep learning models using differential privacy, federated learning, and encrypted computation (e.g., Homomorphic Encryption (HE)[†]). For future work, we would like to explore more on this direction to incorporate more privacy-preservation algorithms into our proposed frameworks.

[†] `en.wikipedia.org/wiki/Homomorphic_encryption`

3.4 Challenges

3.4.1 Evaluation Problems of DP-Models

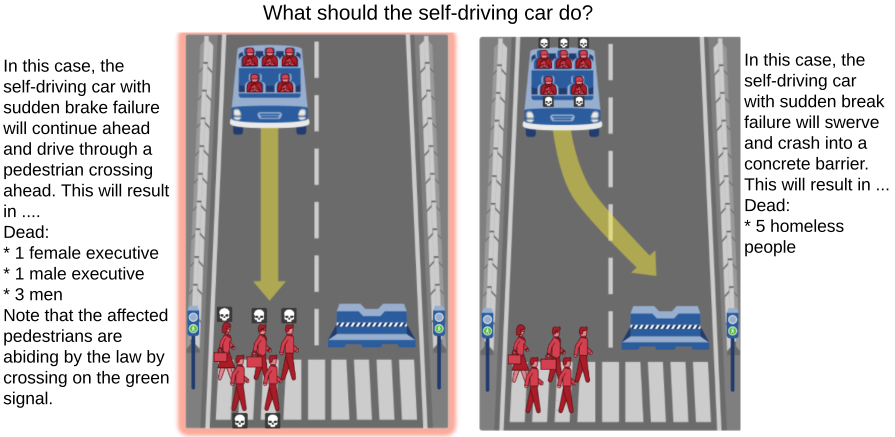
Following the problem of heterogeneous data, evaluating the effectiveness of privacy-guarantee algorithms is not trivial. The naive way to evaluate any privacy-guarantee models is to compare the performance between privacy-guarantee (DP) and non privacy-guarantee (Non-DP) models. The naive evaluation 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 [Mik+13], Elmo [Pet+18], Bert [Dev+19], Caffe [Jia+14]), 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 [VJ18; Pha+17]. 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** [Vu+17a] and **Paper IV** [VTJ19], 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.

3.4.2 Privacy, Machine Learning, and Ethical Issues

Advances in deep learning in general and in computer vision, natural language processing in particular, have enabled many potential capabilities in both industry and academia. Among many successful cases, neural translation is a typical example. With a new level of human-like translation, it reduces the language boundary and helps people be able to communicate and exchange information much easier than before. Similarly, many advances in self-driving cars, autonomous robots have enabled new world of applications to better support human lives. However, *great powers come great responsibility.* The high concerns in abusing the advances in AI come not only from human safety or human privacy aspects, but also from other ethical issues.

The Moral Machine [Awa+18] is a platform for collecting a human perspective on moral decisions made by machine intelligence, such as self-driving cars. Figure 3.5 shows a hypothetical scenario asking a human perspective for choosing to let the car crash into *five pedestrians* or *a concrete barrier.* The first choice results in the deaths of five pedestrians (i.e., three men and two

[§] www.moralmachine.net



What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ....
Dead:
* 1 female executive
* 1 male executive
* 3 men
Note that the affected pedestrians are abiding by the law by crossing on the green signal.

In this case, the self-driving car with sudden break failure will swerve and crash into a concrete barrier. This will result in ...
Dead:
* 5 homeless people

Figure 3.5: A hypothetical scenario asked on TheMoralMachine[§]. In this case, online visitors are asked to decide whether to let the car crash into abiding-by-the-law pedestrians or a concrete barrier.

educated people) and the second choice results in the deaths of five homeless people in the self-driving car. The moral machine experiment is designed to quantifying societal expectations, from which it can be used to guide machine behavior. This experiment is an important research towards addressing ethical issues of machine intelligence. Next, we will discuss on an overview of different privacy and ethical constraints for a Smart Information System, in which self-driving car is one of its instance.

Stahl and Wright [SW18] successfully show how AI and Big Data are used to enable new technologies in smart information systems (SIS) (i.e., information systems with the use of AI as the core solutions). Figure 3.6 shows the constraints between many factors to a smart information system. SIS is powered by two main technical drivers which are artificial intelligence (AI) and Big Data. They enable functionality for operating key technologies such as Social Media (e.g., Facebook, Twitter). However, the development of SIS are constrained by ethical and human right concerns (the central oval area). These constraints require a lot of effort from both the research community and the industry to successfully apply AI to technology products. Among many issues, the authors found that *privacy and data protection* is the most prominent issue since among 809 papers, there were “177 papers addressed the issue of privacy and data protection” [SW18]. Regarding ethical issues in SIS, they are normally complex and the severity of bad consequences is based on different domains and applications. Not only for the case of self-driving car, similar experiment as *the moral machine* but for other domains is very important. For instance, predictive policing algorithms are racist[¶] because of biases towards Black communities. In forestry, if an AI-driven wood cutting machine fails at

[¶] http://bit.ly/predictive-policing-algorithms-racist

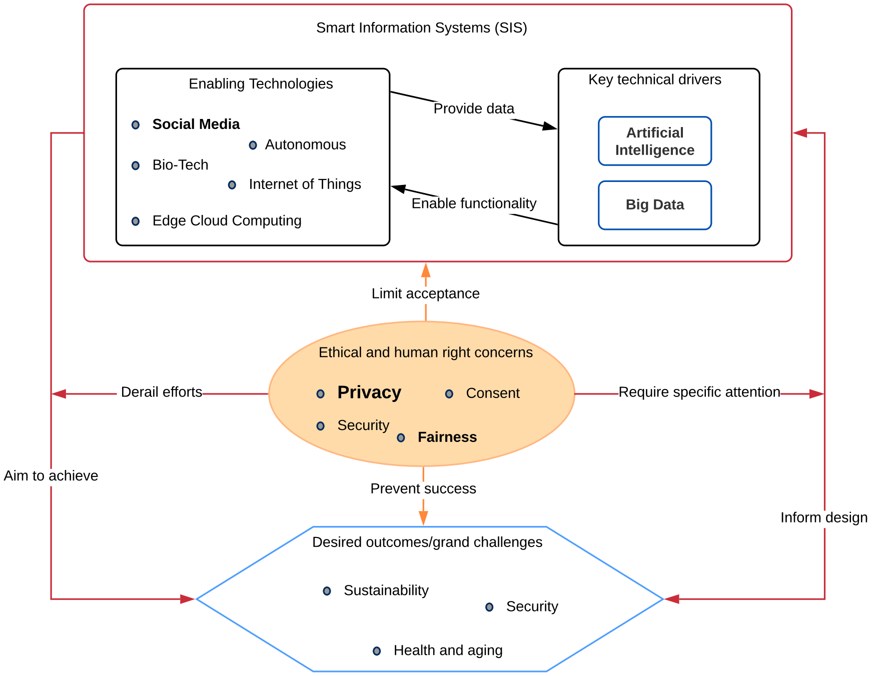


Figure 3.6: A *simplified* ecosystem of smart information systems of Stahl and Wright [SW18]. The ecosystem shows the relations between AI, Big Data and how they are constrained by privacy, security, and ethics in order to achieve desired outcomes in smart information systems.

discriminating a person from a tree, it could end human lives. Regarding this topic, in **Paper V** [Vu+20b], we introduced a step towards the direction to control both privacy and fairness in one SIS. However, because of human safety is always the first, much work has to be done towards the aim of having both *Privacy by Design* and *Ethics by Design* [dAq+18] in SIS.

In summary, this chapter introduces a basic pipeline of a machine learning process, from which different privacy attacks and privacy guarantee methods can be designed. Regarding privacy guarantee, in most of equipped papers, we contributed to the re-identification issue. It is because of both *data analysis* and *data learning process*, they need to use personal data of individuals for downstream applications, which might be the source of re-identification risks. Also, in this chapter, we show two main challenges in privacy-aware machine learning, which are evaluation approaches and ethical issues. They are important topics and require more research works to address constraints of SIS. Addressing these constraints is the key solution to the development of SIS, to be accepted by the society.

Chapter 4

Summary of Contributions

This chapter shows an overview of the thesis contribution by giving a summary of equipped research papers. 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.

In the following sections, each paper is summarily described with reference to the research objectives in Section 1.3. Lili Jiang acted as the main supervisor and Erik Elmroth had the role of the second supervisor. Thus, in most papers, supervisors 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 data analysis infrastructures have limitations in addressing privacy-guarantee methods on data analysis of federated databases. Some systems (e.g., PINQ [McS09], GUPT [Moh+12] 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 [Vu+17a] and INFRA [Vu+19a]) fulfill this requirement by providing unified open-access frameworks that let researchers can flexibly discover register datasets and run data analysis within the frameworks.

[†] **Personality-Based Knowledge Extraction for Privacy-preserving Data Analysis**, Xuan-Son Vu, Lili Jiang, Anders Brändström, 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 4.1: List of my contributions on each paper equipped in this thesis.

Paper	My contributions
Paper I [Vu+17a]	- (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 [Vu+19a]	- (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 [VJ18]	- (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 [VTJ19]	- (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 [Vu+20b]	- (1) Formulated research questions and solutions; (2) implemented privacy and fairness related models and run related experiments & evaluations; (3) wrote-up the paper together with other co-authors.
Paper VI [Vu+20a]	- (1) Formulated research questions and solutions; (2) implemented privacy related models and run related experiments & evaluations; (3) wrote-up the paper together with other co-authors.
Paper VII [Vu+19c]	- (1) Formulated research questions and solutions; (2) implemented most parts of the framework and run related experiments & evaluations; (3) wrote-up the paper together with other co-authors.

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 other problems and makes optimal use of available data. However, sharing personal data between research parties raises a big problem in terms of privacy and data 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 [Vu+17a]. Therefore, KaPPA

Table 4.2: Procedure to research on sensitive data in a comparison between regular research process (i.e., * refers to [AND17]) and our proposed frameworks (i.e., ** refers to [Vu+17a; Vu+19a]).

Traditional sensitive data analysis process*		Our proposed data analysis process **	
1. Research on requirements of data usage. 2. Work on research proposal. 3. Send application to access data. 4. Wait for Approvals Panel's decision. 5. Negotiate for data and setup. 6. Start 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 the queries.	
Waiting time	in months	**Waiting time**	Less than a day
Privacy-guarantee	Regulation-constraint	**Privacy-guarantee**	Statistical guarantee

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 [Vu+19a] 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 [AS94]) to find hidden patterns between multiple variables, from which, they narrow down 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 [AND17] 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:

[†] **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.*

- 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. [Vu+17a] using an over-sampling approach.

Outcomes of this work can be applied for an automatic detection of privacy concern based on user data. With the huge number of data collection having no contact information of data subjects, there should be more research towards this direction to protect user privacy while maintaining useful information for research.

4.3 Paper IV[†]

Paper IV targets at objective **RO2-a** since it introduces differential privacy algorithms for text data sharing. In this paper, we proposed 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 [LG14]. 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 [Mik+13], GloVe [PSM14], FastText [Boj+17], Elmo [Pet+18], Bert [Dev+19], and the like. These pre-trained models have been successfully applied in a variety of tasks like parsing [BGL14], topic modeling [Bat+16]. 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

[†] **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.*

[MEG16], age and gender prediction [FG13]. 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 a model without privacy leakage for facilitating research on sensitive data. To protect privacy, we statistically prevent 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 kind of 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 [Cyn06] in a neural network architecture to learn privacy-guarantee word embedding models. The main contributions of this paper are:

- Introducing a simple yet efficient generalized approach of applying differential privacy on text data to learn embedding model for UGC data sharing.
- Applying user-level privacy-guarantee to differentially private word embedding model to maintain better data utility.
- Conducting extensive experiments to evaluate the effectiveness of our proposed approach to preserve data utility, especially we test the approaches on text analysis task (i.e., regression).

In general, outcomes of this research work can be directly applied to protect privacy for data sharing. For instance, medical text data are very sensitive, however, they are very valuable for research in healthcare. Therefore, this work can be directly applied to learn and share privacy-guaranteed representation for relevant research.

4.4 Paper V[†]

Paper V [Vu+20b] targets at **RO2-b** to generate privacy-aware reviews for multimodal data. The contributions of this **Paper V** are threefold.

- Firstly, we propose a new dp-embedding (dpSENTI) approach for training privacy guarantee embeedings for personalized review generation.
- Secondly, we propose an evaluation approach for sentiment fairness in review domain. We also run the evaluation across multiple language models to evaluate their sentiment fairness in reviews.

[†] **Multimodal Review Generation with Privacy and Fairness Awareness**, Xuan-Son Vu, Thanh-Son Nguyen, Duc-Trong Le, Lili Jiang, *Submitted, 2020.*

- Thirdly, to the best of our knowledge, we are the first to introduce the notions of user privacy and sentiment fairness for the task of review generation. We also evaluate extensively and present insights on multiple tasks ranging from dp-embeddings, sentiment fairness, to review generation. Additionally, the novel dataset will be publicly available with initial benchmark results for the task.

Outcomes of this work can be used to enable various research topics. For example, it could be extended to address the fairness issue of not only food review domain, but also book review or online review in generally. Moreover, we hypothesize that the use of automatic review generation with privacy and fairness awareness can potentially help to improve depression's condition. Fundamentally, *loneliness* is one of the main cause of depression. And *loneliness* exits because there is a virtual barrier stopping patients from connecting to their friends. In this way, generated reviews (e.g., via a personal assistant) could encourage depressed patients to brainstorm and express their emotion to connect with their friends via social network. Nevertheless, it requires more research to address this potential application.

4.5 Paper VI†

Paper VI [Vu+20a] contributes to the **RO2-b** to learn a visual tagging model with privacy preservation on multimodal data. It uses both visual features and graph features to better perform a visual tagging task. This paper has the following contributions:

- We propose **SGTN**, a privacy-preserving visual tagging framework that leverages global knowledge to perform the visual tagging task with new state-of-the-art performances. Meanwhile, it uses less local information of the task to preserve user privacy by avoiding the use of sensitive information (e.g., faces, passport numbers, vehicle license plates).
- We introduce two approaches to construct graph information from label embeddings with privacy guarantee under differential privacy theorem. These constructed graphs help **SGTN** avoid to use private sensitive information from local data.
- We evaluate the effectiveness of **SGTN** with comprehensive experiments on a public bench-marking dataset - i.e., **MS-COCO**, and a real-world education dataset with personal sensitive information.

Outcomes from this work can be applied in practice and research. For practice, the proposed architecture can be applied seamlessly on any visual

†**Privacy-Preserving Visual Content Tagging using Graph Transformer Networks**, Xuan-Son Vu, Duc-Trong Le, Christoffer Edlund, Lili Jiang, Hoang D. Nguyen, *Proceedings of the 28th ACM international conference on Multimedia (ACM MM), 2020.*

tagging task without the need to modify its architecture. Regarding research, we open a new way to incorporate multiple knowledge into a unified neural architecture, which can potentially be adapted to other specific domains such as medical imaging.

4.6 Paper VII[†]

Paper VII [Vu+19c] works on the research objective **RO2-d**, in which we propose a systematic evaluation approach to select a good hyper-parameters for a pre-trained representations. From a set of automatic evaluation metrics, a suitable model is selected to balance between privacy and data utility.

Particularly, in this paper, we introduce *ETNLP* - i.e., a systematic pipeline to extract, evaluate, and visualize pre-trained embeddings on a specific downstream NLP task (hereafter ETNLP pipeline). The ETNLP pipeline consists of three main components which are *extractor*, *evaluator*, and *visualizer*. Based on the vocabulary set within a downstream task, the extractor will extract a subset of word embeddings for the set to run evaluation and visualization. The results from both *evaluator* and *visualizer* will help researchers quickly select which embedding models should be used for the downstream NLP task. On one hand, the *evaluator* gives a concrete comparison between multiple sets of word embeddings. While, on the other hand, the *visualizer* will give the sense on what type of information each set of embeddings preserves given the constraint of the vocabulary size of the downstream task. We detail the three main components as follows.

- **Extractor** extracts a subset of pre-trained embeddings based on the vocabulary size of a downstream task. Moreover, given multiple sets of pre-trained embeddings, it can combine to get the advantage from a few or all of them. For instance, if one wants to use the character embedding to handle the out-of-vocabulary (OOV) problem in a Word2Vec model, the extractor can combine two sets of embeddings and evaluate their performances seamlessly.

- **Evaluator** evaluates the pre-trained embeddings for a downstream task. Given multiple sets of pre-trained embeddings, it runs a word analogy task proposed by Mikolov et al. [Mik+13]. It is noted that, the word analogy set was available only for English and there was not any publicly available *large* benchmark for low resource languages like Vietnamese. Therefore, we proposed a new analogy set for Vietnamese as well as a new evaluation metric to fulfill new characteristics of Vietnamese.

- **Visualizer** visualizes the embedding space of multiple sets of word embeddings. Given a new set of word embeddings, the visualizer helps to get a

[†] **ETNLP: a visual-aided systematic approach to select pre-trained embeddings for a downstream task**, Xuan-Son Vu, Thanh Vu, Son N. Tran, Lili Jiang, *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), 2019.*

sense of what kinds of information (e.g., syntactic or semantic) do the word-embeddings preserve. It allows us to get word samples from the embedding set to see what is the semantic similarity between different words. We designed two different visualization strategies to explore the embedding space: (1) side-by-side visualization and (2) interactive visualization. The side-by-side ("zoom-out") visualization helps users compare the qualities of the word similarity list between multiple embeddings. For the interactive visualization, it helps researchers "zoom-in" each embedding space to explore how each word is similar to the others.

Outcomes of this work can be used in different ways. For instance, the visual-aided exploration can certainly be adopted for exploring multilingual word embeddings. To the systematic evaluation approach, it also can be used to facilitate new research in automatic evaluation of non-trivial tasks - e.g., natural language understanding.

4.7 Future Work

The presented studies in this thesis are possible to be extended in many directions. First, the federated infrastructure's designs are limited to *off the shelf* 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. This extension might be very valuable for researchers, who want to explore and analyze data in many different ways to fulfill their research's needs. 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.

Bibliography

[Aba+16a] 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).

[Aba+16b] 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.

[AC19] M. Al-Rubaie and J. M. Chang. “Privacy-Preserving Machine Learning: Threats and Solutions”. In: *IEEE Security Privacy* 17.2 (2019), pp. 49–58.

[AND17] 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 June 30, 2017).

[AS94] 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.

[Awa+18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. “The moral machine experiment”. In: *Nature* 563.7729 (2018), pp. 59–64.

[Bar17] E. Barendt. *Privacy.* The International Library of Essays in Law and Legal Theory (Second Series). Taylor & Francis, 2017. ISBN: 9781351908801.

[Bat+16] 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.

[BB09] 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.

[BCP18] 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.

[Bd17] Susanne Barth and Menno D.T. de Jong. “The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review”. In: *Telematics and Informatics* 34.7 (2017), pp. 1038–1058. ISSN: 0736-5853. DOI: `https://doi.org/10.1016/j.tele.2017.04.013`. URL: `http://www.sciencedirect.com/science/article/pii/S0736585317302022`.

[BGL14] 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.

[Blu+05] 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.

[Boj+17] 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.

[Cao+14] 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.

[CDH14] 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.

[CH11] 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.

[CSS12] 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.

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-Vector Networks”. In: *Machine Learning*. 1995, pp. 273–297.

[Cyn06] Dwork Cynthia. “Differential Privacy”. In: ICALP. 2006, pp. 1–12.

[Dal77] Tore Dalenius. “Towards a methodology for statistical disclosure control”. In: *statistik Tidskrift* 15.429-444 (1977), pp. 2–1.

[dAq+18] Mathieu d’Aquin, Pinelopi Troullinou, Noel E O’Connor, Aindrias Cullen, Gráinne Faller, and Louise Holden. “Towards an Ethics by Design Methodology for AI Research Projects”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 54–59.

[Dev+19] 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.

[DHX14] 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.

[DR14] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy”. In: *Found. Trends Theor. Comput. Sci.* 9.3&#8211;4 (Aug. 2014), pp. 211–407.

[DS09] Cynthia Dwork and Adam Smithy. “Differential privacy for statistics: What we know and what we want to learn”. In: (2009).

[Du+18] 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.

[Dwo+06] 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.

[EPK14] 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.

[Far+16] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine Cock. "Computational Personality Recognition in Social Media". In: *User Modeling and User-Adapted Interaction* (2016), pp. 109–142.

[FG13] 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.

[FJR15] 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.

[FS10] 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.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* `http://www.deeplearningbook.org`. MIT Press, 2016.

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27.* Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680. URL: `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[Gur+17] 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.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778.

[HML04] Wayne D Hall, Katherine I Morley, and Jayne C Lucke. “The prediction of disease risk in genomic medicine: Scientific prospects and implications for public policy and ethics”. In: *EMBO reports* 5.S1 (2004), S22–S26.

[HOT06] 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.

[HR13] 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.

[HS06] G. E. Hinton and R. R. Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786 (2006), pp. 504–507.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[Hu+16] 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.

[I+15] 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.

[I+16] 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).

[Inc16] Red Hat Inc. *Teiid: a data virtualization system that allows applications to use data from multiple, heterogeneous data stores.* `http://teiid.io/`. 2016.

[JDB14] Michael J Paul, Mark Dredze, and David Broniatowski. “Twitter Improves Influenza Forecasting”. In: *PLoS currents* 6 (Oct. 2014).

[JGK16] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. “Big data privacy: a technological perspective and review”. In: *Journal of Big Data* (2016).

[Jia+14] 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.

[JKT12] 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.

[JLE14] 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`.

[KAH15] 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.

[Kow+18] 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`.

[KT13] 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.

[KW17] Thomas N. Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *Proceedings of the 5th International Conference on Learning Representations.* 2017, pp. 1–14.

[KWG13] 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.

[LC11] Jaewoo Lee and Chris Clifton. “How much is enough? choosing ϱ for differential privacy”. In: (2011), pp. 325–340.

[LC12] Jaewoo Lee and Chris Clifton. “Differential Identifiability*”. In: *Proceedings of KDD.* 2012.

[LDR08] 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.

[Lec+18] 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).

[LG14] 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.

[LLV07] 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.

[Mac+06] 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.

[McM+17] 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`.

[McS09] Frank D McSherry. “Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis”. In: *SIGMOD.* 2009.

[MEG16] 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.

[Meh+16] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo. “Protection of Big Data Privacy”. In: *IEEE Access* 4 (2016), pp. 1821–1834.

[MG11] Peter M. Mell and Timothy Grance. *SP 800-145. The NIST Definition of Cloud Computing.* Tech. rep. Gaithersburg, MD, United States, 2011.

[Mik+13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space”. In: *CoRR* (2013). arXiv: `1301.3781`.

[Mit97a] T. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[Mit97b] T. Mitchell. “Machine Learning”. In: McGraw-Hill, 1997, p. 2. ISBN: 978-0-07-042807-2.

[Moh+12] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. “GUPT: Privacy Preserving Data Analysis Made Easy”. In: *SIGMOD*. 2012.

[Moh+15] 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.

[MW09] 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.

[Net09] Netflix. *Netflix Prize Contest*. 2009. URL: `https://en.wikipedia.org/wiki/Netflix_Prize` (visited on June 30, 2017).

[Ngi+11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. “Multimodal deep learning”. In: *ICML*. 2011.

[Ngu+19] Dai Quoc Nguyen, Thanh Vu, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. “A Capsule Network-based Embedding Model for Knowledge Graph Completion and Search Personalization”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2180–2189. DOI: `10.18653/v1/N19-1226`. URL: `https://www.aclweb.org/anthology/N19-1226`.

[NIR16] 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.

[NRS07] 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.

[NS08] A. Narayanan and V. Shmatikov. “Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset)”. In: (2008).

[Pap+18a] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman. “SoK: Security and Privacy in Machine Learning”. In: *2018 IEEE European Symposium on Security and Privacy (EuroS P)*. 2018, pp. 399–414.

[Pap+18b] 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).

[Pet+18] 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.

[Pha+17] NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. “Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning”. In: *CoRR* abs/1709.05750 (2017).

[Pop+18] 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.

[PSM14] 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.

[Rea+20] Esteban Real, Chen Liang, David R. So, and Quoc V. Le. “AutoML-Zero: Evolving Machine Learning Algorithms From Scratch”. In: *Proceedings of the 37th, International Conference on Machine Learning (PMLR)*. 2020.

[RS20] Maithra Raghu and Eric Schmidt. *A Survey of Deep Learning for Scientific Discovery*. 2020. arXiv: `2003.11755 [cs.LG]`.

[Rub+09] 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).

[Sal+11] 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.

[Sam01] P. Samarati. “Protecting respondents identities in microdata release”. In: *IEEE Transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SFC18] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. "News Session-Based Recommendations Using Deep Neural Networks". In: *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*. DLRS 2018. Vancouver, BC, Canada: ACM, 2018, pp. 15–23.

[SL90] 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.

[Sou+14] 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).

[SS98] Pierangela Samarati and Latanya Sweeney. *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.* Tech. rep. 1998.

[Sti+13] 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.

[SW18] B. C. Stahl and D. Wright. "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation". In: *IEEE Security Privacy* 16.3 (2018), pp. 26–33.

[Swe02] 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.

[Sze+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision". In: *CVPR*. 2016, pp. 2818–2826.

[Tsa+15] 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.

[Vai+13] 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.

[Vin12] 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.

[Vis+12] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. “Five Years of GWAS Discovery”. In: *The American Journal of Human Genetics* 90.1 (2012), pp. 7–24. ISSN: 0002-9297.

[VJ18] 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.

[VNN20] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. “A Label Attention Model for ICD Coding from Clinical Text”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (July 2020). DOI: `10.24963/ijcai.2020/461`. URL: `http://dx.doi.org/10.24963/ijcai.2020/461`.

[VTJ19] Xuan-Son Vu, Son N. Tran, and Lili Jiang. “dpUGC: Learn Differentially Private Representation for User Generated Contents”. In: *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, April, 2019*. La Rochelle, France, 2019.

[Vu+17a] 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.

[Vu+17b] 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. Austin, TX, USA: ACM, 2017, 45:1–45:4. ISBN: 978-1-4503-5553-7. DOI: `10.1145/3148011.3154479`. URL: `http://doi.acm.org/10.1145/3148011.3154479`.

[Vu+19a] 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.

[Vu+19b] Xuan-Son Vu, Abhishek Santra, Sharma Chakravarthy, 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.

[Vu+19c] Xuan-Son Vu, Thanh Vu Vu, Son Tran N., and Lili Jiang. “ETNLP: A Toolkit for Extraction, Evaluation and Visualization of Pretrained Word Embeddings”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*. 2019, pp. 1286–1295.

[Vu+20a] Xuan-Son Vu, Duc-Trong Le, Christoffer Edlund, Lili Jiang, and Hoang D. Nguyen. “Privacy-Preserving Visual Content Tagging using Graph Transformer Networks”. In: *Proceedings of the 28th ACM international conference on Multimedia (ACM MM)*. 2020.

[Vu+20b] Xuan-Son Vu, Son N. Nguyen, Duc-Trong Le, and Lili Jiang. “WRIST: a Multimodal Writing Review Assistant with Privacy and Fairness Awareness”. In: *Submitted*. 2020.

[Wan+09] 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.

[WCX19] 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.

[Wu+18] 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.

[XR08] Li Xiong and Kumudhavalli Rangachari. “Towards Application-Oriented Data Anonymization”. In: *nternational Workshop on-Practical Privacy-Preserving Data Mining* (2008).

[Xu+14] 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.

[XYW19] 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.

[YOU09] SEOUNMI YOUN. “Determinants of Online Privacy Concern and Its Influence on Privacy Protection Behaviors Among Young Adolescents”. In: *Journal of Consumer Affairs* 43.3 (2009), pp. 389–418. DOI: 10.1111/j.1745-6606.2009.01146.x.

[ZDS18] 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).

[Zha+12] 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.

[Zhu08] Xiaojin Zhu. “Semi-Supervised Learning Literature Survey”. In: *Comput Sci, University of Wisconsin-Madison* 2 (July 2008).