# Tutorial on Privacy-Preserving Natural Language Processing

**Ivan Habernal**[1] **Fatemehsadat Mireshghallah**[2] **Patricia Thaine**[3]
**Sepideh Ghanavati**[4] **Oluwaseyi Feyisetan**[5]

[1]Trustworthy Human Language Technologies, Technical University of Darmstadt
[2]Computer Science and Engineering Department, University of California San Diego
[3]Private AI, Canada
[4]School of Computing and Information Science, University of Maine
[5]Meta, USA

## Abstract

This cutting-edge tutorial will help the NLP community to get familiar with current research in privacy-preserving methods. We will cover topics as diverse as membership inference, differential privacy, homomorphic encryption, or federated learning, all with typical applications to NLP. The goal is not only to draw the interest of the broader community, but also to present some typical use-cases and potential pitfalls in applying privacy-preserving methods to human language technologies.

## 1 Introduction

Human language technologies play an essential role in the modern society. From automatic machine translation to drug discovery, NLP has had an undeniable impact on everyone's life. However, many of the recent achievements of state-of-the-art models come at a price that everyone must pay. In the race for yet better performing systems, the research has completely ignored the fact that within the extreme amounts of data needed for the 'hungry' models, there are private information of actual living persons (Carlini et al., 2020). Our sensitive information – be it explicitly mentioned in the texts we or someone else writes about us, or implicitly in our writing style – is at stake with current NLP models. Privacy matters a lot to society, but has been largely neglected by NLP researchers.

This tutorial aims to close this gap by offering the community insights into state-of-the-art approaches to privacy-preserving NLP. We will cover diverse topics, such as membership inference, differential privacy, homomorphic encryption, or federated learning, all with typical use-cases and applications. The tutorial will try to balance theoretical foundations with practical considerations.

## 2 Tutorial outline

We propose a **half-day** tutorial (3 hours) divided into four following thematic blocks.

### 2.1 Block 1: Attacks (30 minutes)

This block will provide an overview why differential privacy is needed by introducing and discussing reconstruction attacks and examples of difference attacks (Dinur and Nissim, 2003). We will discuss how an algorithm can be blatantly non-private via an example from census data and explain inefficient and efficient attacks. We then discuss reconstruction attacks in practice for several cases (Cohen and Nissim, 2020). We conclude this block by briefly explaining some examples of tracing attacks (Homer et al., 2008) and (Dwork et al., 2015).

### 2.2 Block 2a: Defence with formal guarantees (60 min)

This block will introduce differential privacy, a mathematical framework for privacy protection (Dwork and Roth, 2013). We will explain the typical setup (why this privacy approach has 'differential' in its title) and the formal definitions. Then we will address some basic DP mechanisms and show their NLP applications. This part will involve a few mathematical proofs, but our aim is to make it low-barrier and accessible to a very broad audience.

In the second part, we will introduce some cryptographic tools, namely homomorphic encryption and secure multiparty computation. The main focus will be on introducing the basics of lattice-based cryptography and homomorphic encryption and the most popular schemes (BGV, CKKS). We will go over the available libraries (PALISADE, HElib, SEAL) and dive into an NLP-specific example.



*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*
*Tutorial Abstracts*, pages 27–30
May 2-4, 2023 ©2023 Association for Computational Linguistics

### 2.3 Block 2b: Defences without formal guarantees (20 min)

Apart from privacy-preserving schemes that directly optimize for a given definition of privacy, there are given execution models and environments that help enhance privacy and are not by themselves privacy-preserving in a formal sense. This block will introduce privacy-enhancing methods such as federated learning (McMahan et al., 2017), split learning (Vepakomma et al., 2018) and regularizer-based methods (Coavoux et al., 2018; Mireshghallah et al.; Li et al., 2018).

Federated learning and split learning are both based on distributed learning and are great methods for application in enterprise and clinical setups. Regularizer based and private representation learning methods add extra terms to the loss function to limit the memorization and encoding of sensitive data within the model.

### 2.4 Block 3: Privacy in industry (40 min)

Companies have practical constraints when deploying privacy preserving technologies. Some of these include deployment and computation at scale, or guarantees that solutions meet compliance or regulatory requirements. There is also the trade-off between privacy, utility, bias, fairness, (Farrand et al., 2020) as well as explainability and verifiability of the implemented solutions.

In this section, we will dive deep into different technologies and discuss their trade-offs from an industry perspective. We will also highlight how the community can help accelerate progress along different dimensions.

### 2.5 Block 4: Open problems in privacy in NLP (30 min)

We will talks some further NLP specifics, such as (1) perturbing long-form text with differential privacy without losing the content, and (2) introducing better auditing methods for measuring memorization in discriminative and generative large language models (BERT or GPT based models).

## 3 Reading list


- Joseph P. Near and Chiké Abuah. 2021. *Programming Differential Privacy*. Online, https://uvm-plaid.github.io/programming-dp/
- (Optional) Cynthia Dwork and Aaron Roth. 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407


## 4 Tutorial specifics

The proposed tutorial is considered a **cutting-edge** tutorial that presents recent advances in an emerging area of privacy-preserving techniques for NLP. The topic presented has not been covered in previous ACL*-family tutorials in the last 4 years. We estimate that at least 60% of the papers covered in this tutorial are from researchers other than the instructors. It is also **different** from other tutorials, e.g., on differentially-private machine learning, as we target NLP with all its peculiarities related to human language.

The **preferred venue** for this tutorial would be 1) ACL, 2) EACL, 3) EMNLP. We prefer ACL due to travel arrangements of presenters located in the U.S.

Based on the raising interest in this topic, we expect around 30 participants. The tutorial will be **self-contained**, however attendees should have solid **background** in basic deep learning technologies in NLP (representations, architectures, optimization)[1] and to brush up knowledge of probability and statistics (Laplace or Gaussian distributions and probability bounds).[2]

We are committed to **open-source** all teaching materials under permissible license.

## 5 Tutorial presenters

**Diversity** considerations:

- 4 academia and 2 industry affiliations
- 3 female instructors
- Participation of senior (up to Assistant Professor) and junior (PhD candidate) instructors

Details of the organizing committee are included below in alphabetical order.

### Oluwaseyi Feyisetan (Meta, USA)

Seyi is a Staff Research Scientist at Facebook. Prior to Facebook, he was a Senior Applied Scientist at Amazon where he worked on Differential Privacy in the context of NLP. He holds 4 pending patents

[1] For example (Goldberg, 2017)

[2] For example Chapter 1–4 and 8–9 from (Mitzenmacher and Upfal, 2017)

with Amazon on preserving privacy in NLP systems. He completed his PhD at the University of Southampton in the UK and has published in top tier conferences and journals on crowdsourcing, homomorphic encryption, and privacy. He has served as a reviewer at top NLP conferences including ACL and EMNLP. Prior to Amazon, he spent 7 years in the UK where he worked at different startups and institutions focusing on regulatory compliance, machine learning and NLP within the finance sector. He also sits on the research advisory board of the IAPP.

**Sepideh Ghanavati (University of Maine, USA)**

Assistant professor in Computer Science at the University of Maine. She is the director of Privacy Engineering - Regulatory Compliance Lab (PERC_Lab). Her research interests are in the areas of information privacy and security, software engineering, machine learning and the Internet of Things (IoT). Previously, she worked as an assistant professor at Texas Tech University, visiting assistant professor at Radboud University, the Netherlands and as a visiting faculty at Carnegie Mellon University. She is the recipient of Google Faculty Research award in 2018. She has more than 10 years of academic and industry experience in the area of privacy and regulatory compliance and has published more than 30 peer-reviewed publications. She was a co-organizer of the 'Privacy and Language Technologies' at the 2019 AAAI Spring Symposium and has been part of the organizing committee of several workshops and conferences in the past.

**Ivan Habernal (Technische Universität Darmstadt, Germany)**

Ivan Habernal is currently leading a junior independent research group at the Technical University of Darmstadt, Germany, funded ad-personam by the state of Hessen. His group entitled "Trustworthy Human Language Technologies" focuses on privacy-preserving NLP and legal argument mining, among others. He has a track of top NLP publications (h-index 19), chairing workshops and tutorials, area chairing, organizing SemEval competition, giving invited talks, and also some recent industrial experience in areas where privacy matters a lot but the tools are not ready yet (healthcare and online personalization).

**Fatemeh Mireshghallah (University of California, USA)**

Fatemehsadat Mireshghallah is a Ph.D. student at the CSE department of UC San Diego. Her research interests are Trustworthy Machine Learning and Natural Language Processing. She received her B.S. from Sharif university of technology in Iran. She is a recipient of the National Center for Women & IT (NCWIT) Collegiate award in 2020 for her work on privacy-preserving inference, and a finalist of the Qualcomm Innovation Fellowship in 2021. She has interned twice at Microsoft Research's Language and Intelligent Assistance group, where she worked on private training of large language models. She is also serving as a NAACL 2022 D&I co-chair and WiNLP committee member.

**Patricia Thaine (University of Toronto, Canada)**

Patricia Thaine is the Co-Founder and CEO of Private AI, a Computer Science PhD Candidate at the University of Toronto and a Postgraduate Affiliate at the Vector Institute doing research on privacy-preserving natural language processing, with a focus on applied cryptography. She also does research on computational methods for lost language decipherment. Patricia is a recipient of the NSERC Postgraduate Scholarship, the RBC Graduate Fellowship, the Beatrice 'Trixie' Worsley Graduate Scholarship in Computer Science, and the Ontario Graduate Scholarship. She has eight years of research and software development experience, including at the McGill Language Development Lab, the University of Toronto's Computational Linguistics Lab, the University of Toronto's Department of Linguistics, and the Public Health Agency of Canada. She is the Co-Founder and CEO of Private AI, the former President of the Computer Science Graduate Student Union at the University of Toronto, and a member of the Board of Directors of Equity Showcase, one of Canada's oldest not-for-profit charitable organizations.

## References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. *arXiv preprint*.

Maximin Coavoux, Shashi Narayan, and Shay B. Co-

hen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Aloni Cohen and Kobbi Nissim. 2020. Linear program reconstruction in practice. *J. Priv. Confidentiality*, 10.

Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 202–210, New York, NY, USA. Association for Computing Machinery.

Cynthia Dwork and Aaron Roth. 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407.

Cynthia Dwork, Adam D. Smith, Thomas Steinke, Jonathan R. Ullman, and Salil P. Vadhan. 2015. Robust traceability from trace amounts. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 650–669. IEEE Computer Society.

Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.

N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *ACL*.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning differentially private language models without losing accuracy. *ArXiv*, abs/1710.06963.

Fatemehsadat Mireshghallah, Huseyin A Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. Privacy regularization: Joint privacy-utility optimization in language models.

Michael Mitzenmacher and Eli Upfal. 2017. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*, 2nd edition. Cambridge University Press.

Joseph P. Near and Chiké Abuah. 2021. *Programming Differential Privacy*. Online, https://uvm-plaid.github.io/programming-dp/.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. *ArXiv*, abs/1812.00564.