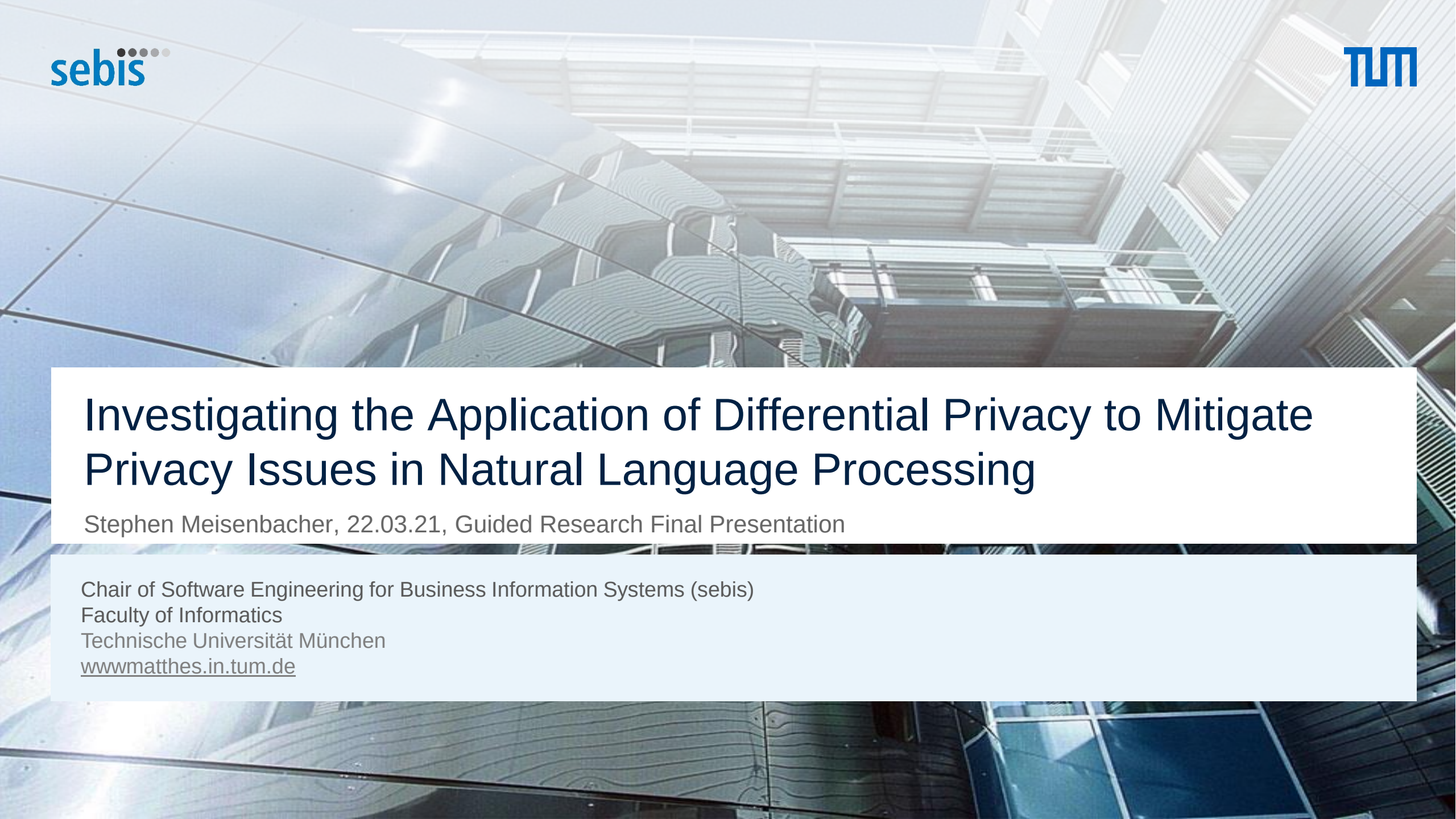


Investigating the Application of Differential Privacy to Mitigate Privacy Issues in Natural Language Processing

Stephen Meisenbacher, 22.03.21, Guided Research Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

Introduction (recap)

Research Questions

Methodology
- Literature Review
- Interviews
- Paper

Findings
- Privacy Vulnerabilities
- Differential Privacy for NLP
- Differential Privacy → d_x-privacy
- Benefits and Limitations

Future Work

DP in NLP Learning Nugget / Web App Examples

Conclusion

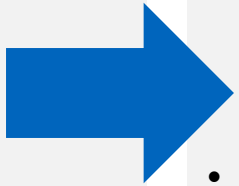
Recap – Background, Motivation, Goals

Background

Main driving points:

1. People's viewpoint on privacy is becoming increasingly transparent
   - Growing awareness of risks → skepticism
   - + perceived lack of control
2. "Big data" on the rise
   - Private sector: booming market cap
   - Academic research follows accordingly
3. Data breaches, their ensuing consequences
   - General upward trend in occurrences – costly!
   - Result: GDPR, CCPA, what's next?
4. Privacy as a topic of interest
   - Privacy research = hot topic
   - "Data Privacy Will Be The Most Important Issue In The Next Decade" *

*https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/



Motivation + Goals

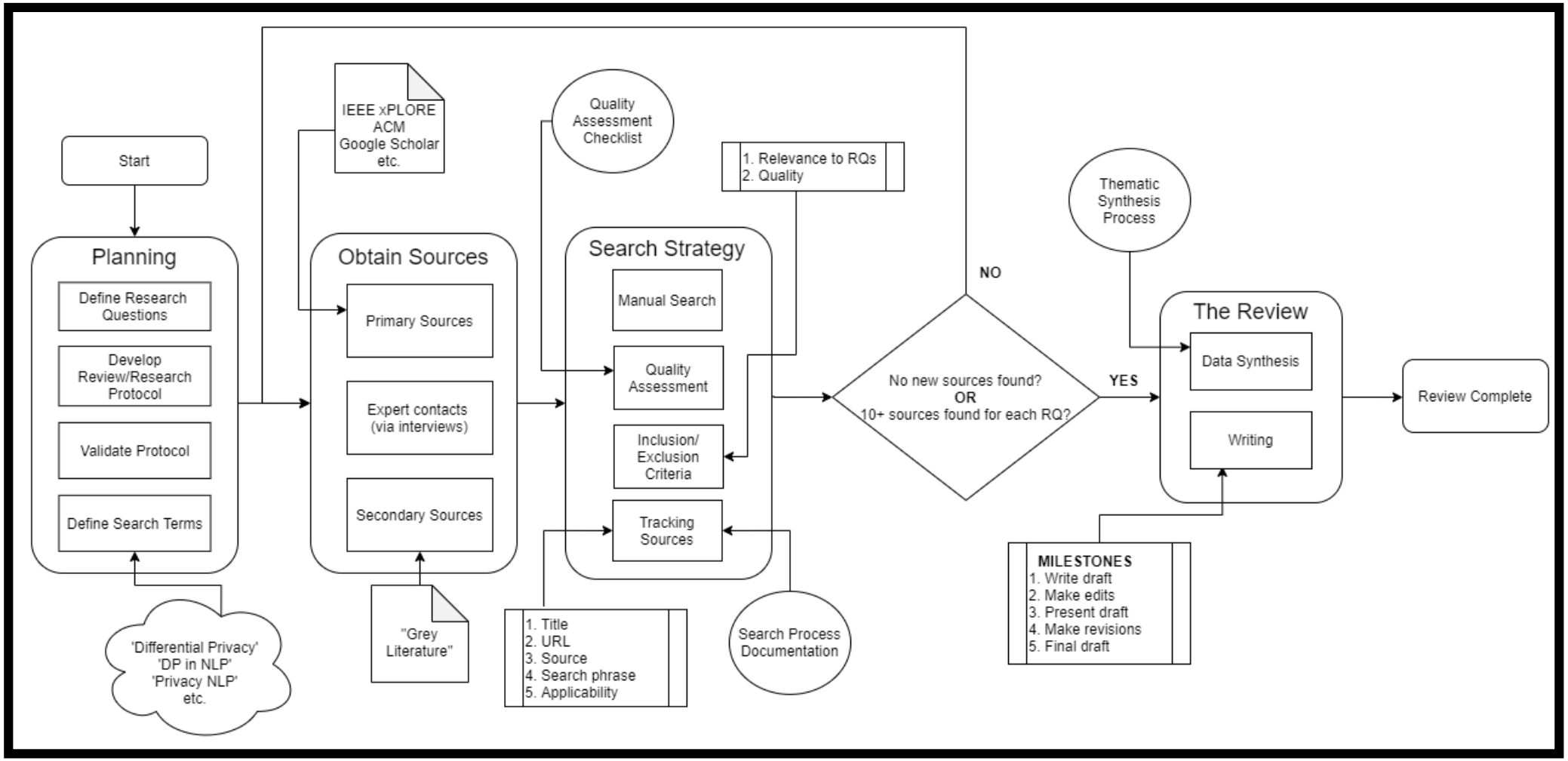
"As a result":

- Protection of privacy is as important as ever
  - Need for privacy-enhancing technologies (PETs)
- Great candidate: Differential Privacy
  - Mathematical foundations, privacy guarantee, useful properties
- Scope: Natural Language Processing
  - Dealing with large-scale, unstructured text data
  - DP + NLP feasible?
- Goal: gain an overview of DP in NLP
  - Privacy vulnerabilities
  - Technical applications and use cases
  - Pros, cons
  - Overall: current work + potential

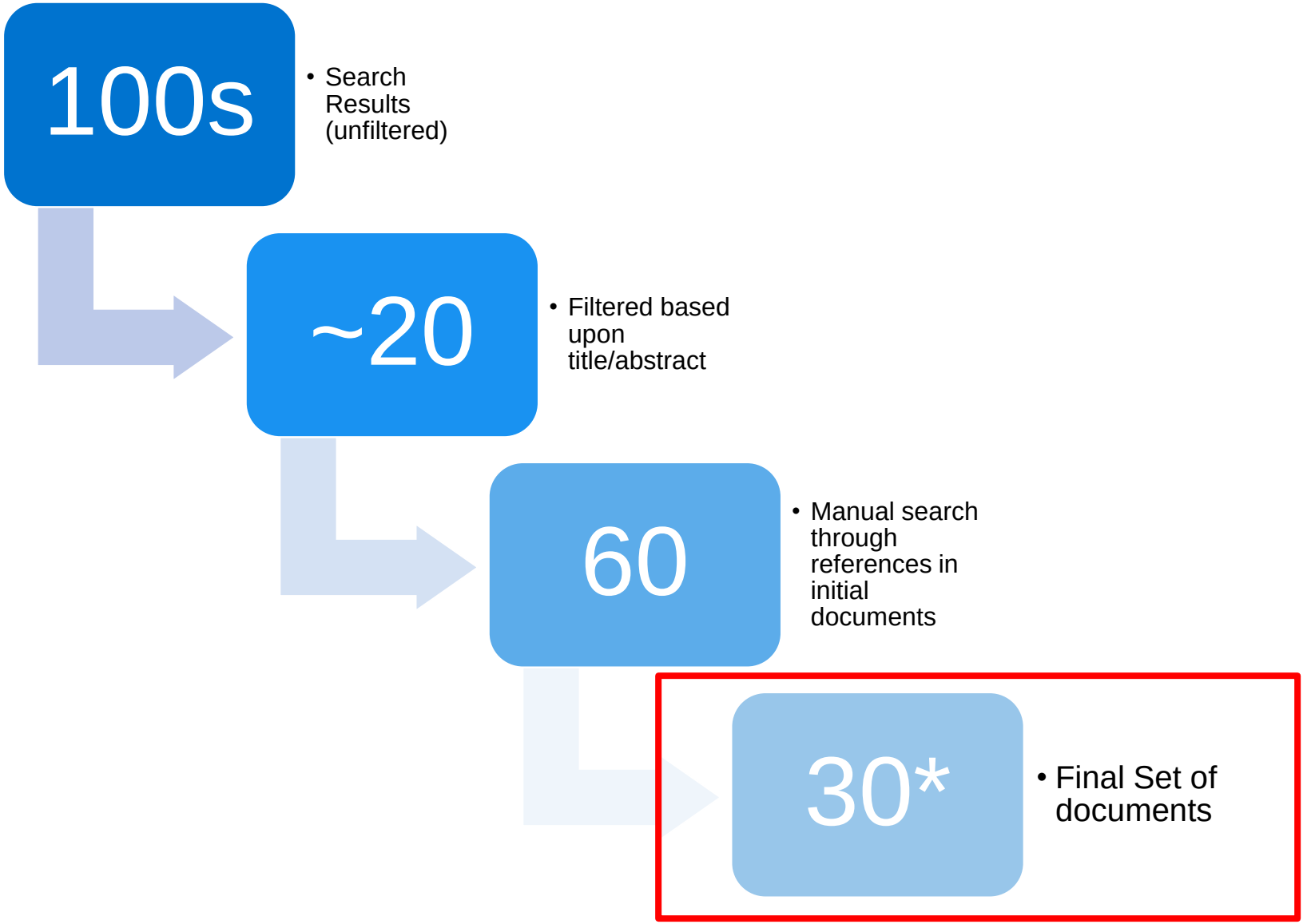
Research Questions

1. What vulnerabilities to current NLP techniques is Differential Privacy capable of preventing?
2. What are the foundations of Differential Privacy, and how can it be applied to NLP tasks?
3. What are the distinct benefits and limitations of applying Differential Privacy to NLP tasks?

Methodology – Overview



Methodology – Literature Review



*not including “supplementary” material

Methodology – Interviews

- **Semi-structured interviews** with 4 people
  - Experts in the field working with privacy + NLP
  - e.g. Researcher at Amazon, PhD candidate
- **30-60 minute** video conference interview
- Generally following a pre-defined list of questions:
  - **General**: current work, background with privacy, thoughts on privacy + NLP
  - **RQ1**: NLP privacy vulnerabilities, attack types, preventative work so far
  - **RQ2**: DP foundations, application to NLP, use cases, technical implementations
  - **RQ3**: major advantages, current limitations, future improvement, thoughts on future of private NLP
  - Total: 18 questions + sub-questions + 1-2 tailored specifically to interviewee
- After the interview: transcribe
  - ~3 hours → ~25 pages
  - Helpful for writing the paper

Methodology – Paper

- Overall goal: summary report of research findings
- Main sections:
  1. Introduction of problem
  2. Overview of current/related work
  3. Privacy vulnerabilities to NLP
  4. Foundations of DP + early applications to NLP
  5. Foundations of d_x-privacy + applications
  6. In-depth discussion of applicability, benefits, limitations
  7. Prognosis for future work
- End result:
  - 12 page Guided Research report
  - Separate journal submission (PETS 2021)

https://petsymposium.org/

Privacy Vulnerabilities in NLP – Where's the Risk?

Word Embeddings

- Representations of words in vector space
- Main goal: capture semantic (or contextual) associations between words
  - "Distributional" semantics
- Heavily used in modern NLP
- Models trained on billions of sentences → possibly sensitive
- Issue: when private information is encoded into embeddings

Language Leakage

- Text organized into some representation
  - e.g. syntactical, lexical features
- Inherently a "stylometric profile"
- Good for features
- But: conveniently expose implicit information about authorship style and identity
- "Reverse-engineering" pieces of text is relatively easy

Neural NLP

- Much of NLP today is done using a neural component
- Main goal: learn patterns to make accurate predictions, classifications, etc.
- Problem: almost too powerful
- Memorization of sensitive parts of the input text
- *Exposure* is tied to learning → not ideal

Privacy Vulnerabilities in NLP – Use Cases

Data Release

Setup: data is released in some form to a third party

Goal: utilization in research, creation of a product, etc.

Problem: a malicious user gains access to this data

There are many cases where such (text) data is shared:

- Medical context (doctors' notes, hospital records)
- Online reviews
- Social media / blog posts
- Government records
- Genome sequences

Even problematic in embedding form!

Model Abuse

Setup: centralized or decentralized learning

- Centralized: central model does computations
- Decentralized: local computations, central server (i.e. in cloud)
- Good example: IoT applications

Goal: any shared learning task

Problem: presence of a malicious user

Two ways for malicious user to infer data:

- Black-box access: query outputs
- White-box access: black-box + access to (sample of) original data

In the wrong hands, query outputs can be exploited!

Privacy Vulnerabilities in NLP – Exploitations and Attack Pipeline

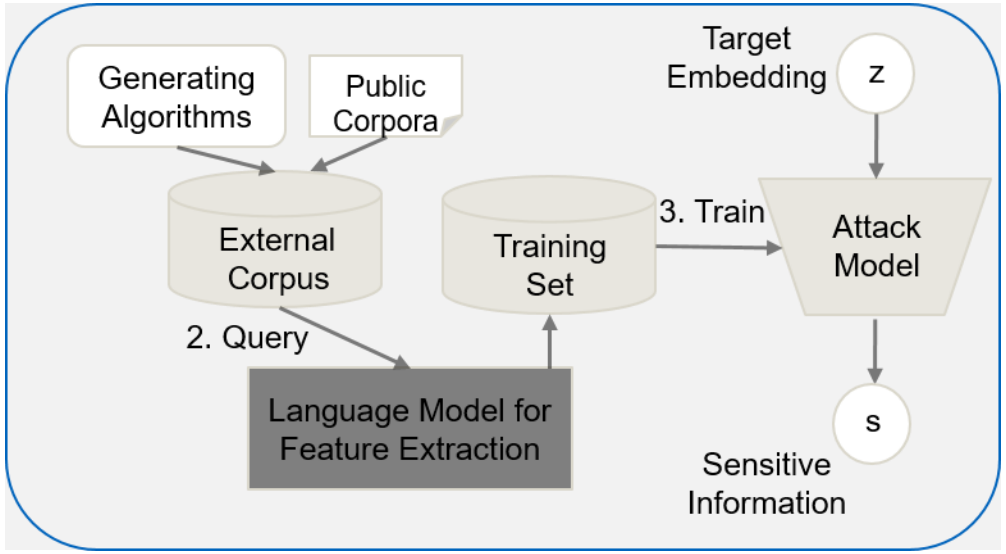
- Focus here: *Inference Attacks*
  - Infer information from obtained data
  - Text: “learn what’s inside”
- Two classes (relevant to DP + NLP)

Pattern Reconstruction

- Target: sensitive information with fixed format
  - e.g. SSNs, zip codes, birth dates, phone #s
- Attacker’s goal: reconstruct these fixed-format strings given some text representation

Keyword Inference

- Target: keywords given some domain knowledge
- Little to no prior knowledge is really necessary



A general attack pipeline that can be used to leverage the mentioned attacks

Based on: X. Pan, M. Zhang, S. Ji and M. Yang, "Privacy Risks of General-Purpose Language Models"

Differential Privacy for NLP – Foundations

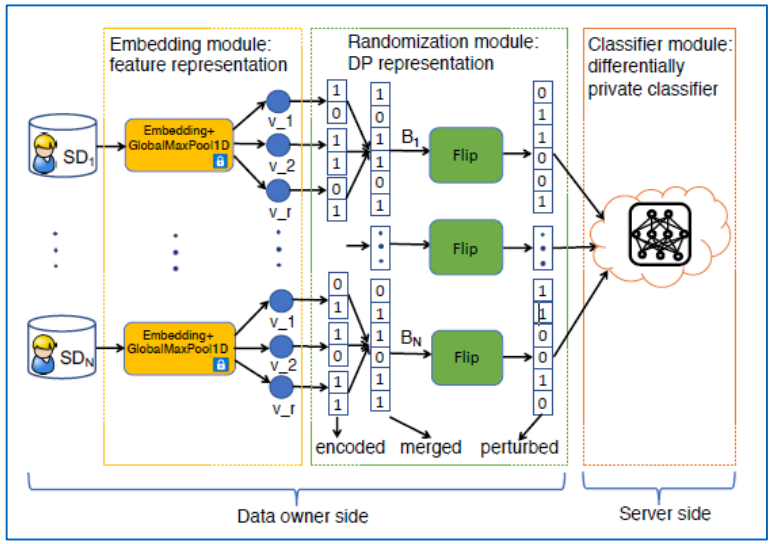
- DP key idea: privacy protection for the *individual*
  - Originally for (structured) database setup
  - Randomized response → “plausible deniability”
  - ε parameter to quantify privacy guarantee

$$\frac{Pr[\mathcal{M}(x) \in S]}{Pr[\mathcal{M}(x') \in S]} \leq e^{\epsilon}$$

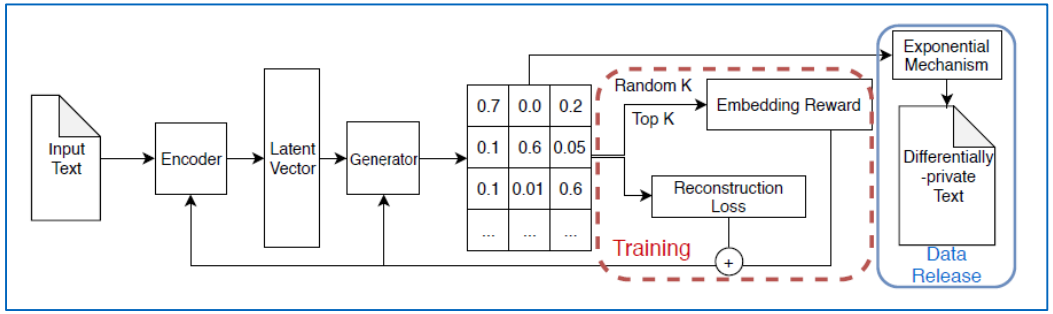
- Why DP for NLP?
  - Not protection against inferences themselves
  - Protection of the *individual* against information gained through inference
  - NLP: protect contributor’s of original text from attacker knowledge gain
  - Plausible deniability w.r.t. whether inferred information is true
- Immediate Challenge with NLP
  - Mainly unstructured data!
  - Who/what is the “individual” in a dataset now?
    - A document, a word?
  - What “databases” are adjacent/neighboring?
    - Standard DP: two databases differing by one entry
    - DP with NLP: unclear

Differential Privacy for NLP

- How to transfer concepts?
  - Two “databases”, each with one document
  - Documents differ by exactly one word (token)
  - Adjacent = can achieve the other document by make a single edit
- Extrapolate: arbitrary number of edits
  - Compositionality of DP important!
  - More edits needed = more distinguishable / less private, and vice versa
- Key concept: *indistinguishability* between documents achieved via composition of edits
- Application examples:
  - **OME**: perturbed binary text vector representations
  - **SYN-TF**: synthetic term frequency vectors via the Exponential Mechanism
  - **ER-AE**: differentially private text generation
  - **DP-SGD**
- (Numerical) text representation required!



Optimized Multiple Encoding (OME)
Lyu, Lingjuan et al. “Towards Differentially Private Text Representations”

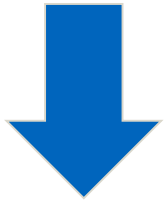


Embedding Reward Autoencoder (ER-AE)
Bo, Haohan et al. “ER-AE: Differentially-private Text Generation for Authorship Anonymization”

From Differential Privacy to d_x-privacy

- Problems with (standard) DP with NLP
  - Made to fit the original inequality
  - Too strict – any two documents are neighboring
    - Result: high ε values required
  - No consideration to semantics of language
    - Good start: Exponential Mechanism
- Need something more tailored to NLP!
- Solution: d_x-privacy
  - a.k.a. metric DP, “generalized DP”
- Key concept: *metric spaces*
  - Data represented as points (think embedding space)
  - (Distance) metric d_X = adjacency measure
- Incorporate into new, modified inequality
- Intuition: required indistinguishability depends on similarity of two documents
  - Smaller distance (greater similarity) = more indistinguishable output must be
  - = “scaling” the noise required

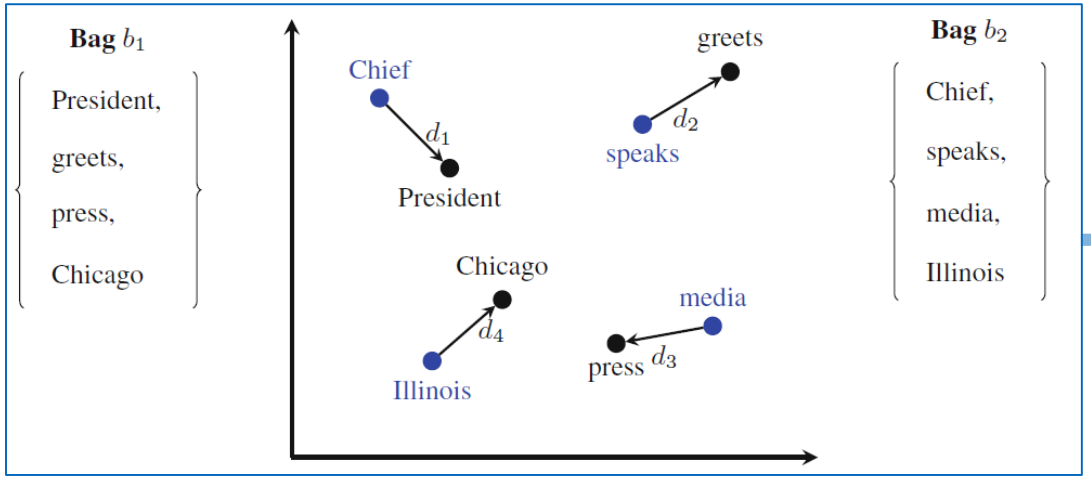
$$K(x)(z) \leq e^{\epsilon} K(x')(z)$$



$$K(x)(z) \leq e^{\epsilon d_{\mathcal{X}}(x,x')} K(x')(z)$$

d_x-privacy for NLP

- Example metric: *Word Mover's Distance (WMD)*
  - i.e. how we "move" one document to another using single word (vector) edits
  - Using the minimal WMD, the DP inequality works
- In the word embedding space, use cosine distance for single words
- Then, required indistinguishability is scaled by semantic closeness
- Key takeaway
  - Standard DP: emphasis is on databases differing by an individual entry
  - Metric DP: emphasis on ***how*** these individuals (i.e. documents) differ



$$d_W = \tfrac{1}{4}\,[d(President, Chief) + d(greets, speaks) + d(press, media) + d(Chicago, Illinois)]$$

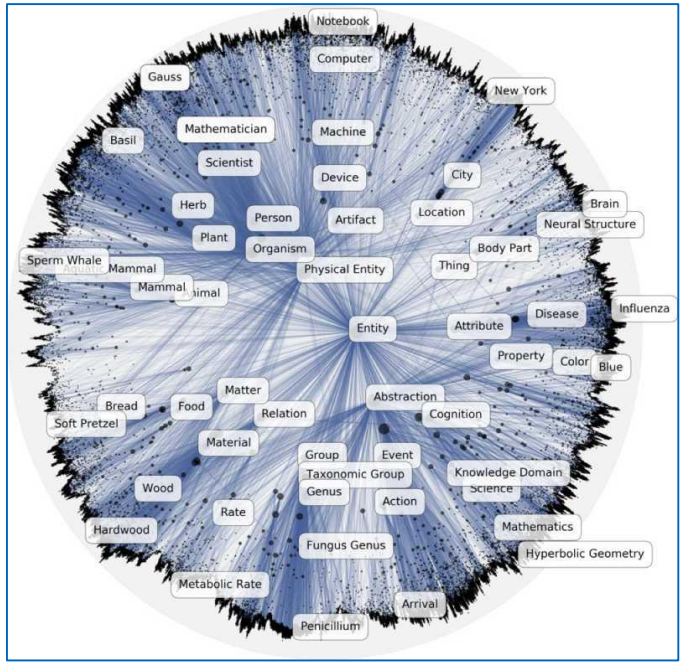
Example from Fernandes, Natasha et al. "Author Obfuscation Using Generalised Differential Privacy."

d_x-privacy for NLP (cont.)

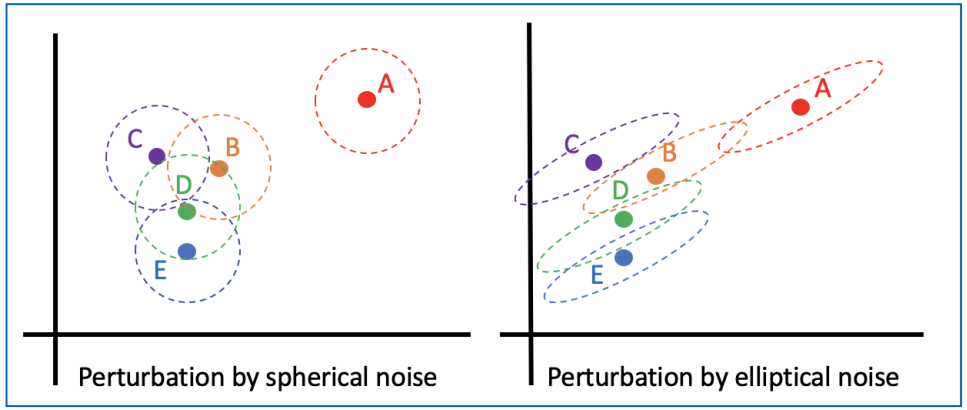
Some applications:

- Original approach: Euclidean space
  - Calibrated noise via Laplace Mechanism
  - Projection to nearest neighbor
- Hyperbolic space
  - Key idea: model hierarchical relationships in language
  - Makes sense, especially with hyperbolic geometry
- Mahalanobis distance
  - Tackle problem of sparseness
  - Take into account the shape of a particular (sub-)space

Important to note: change of metric / metric space quite seamless with metric DP!



Hyperbolic Embeddings example, source: https://www.techleer.com/articles/478-insight-into-hierarchical-representations-through-poincare-embedding/



Xu, Zekun et al. “A Differentially Private Text Perturbation Method Using a Regularized Mahalanobis Metric.”

DP in NLP: Benefits

1. **Reasoning** about privacy in NLP
   - Traditionally hard with NLP / text
   - DP provides the foundation
2. **Flexibility**
   - Today: many different NLP architectures with underlying space, geometries, metrics, etc.
   - DP provides flexibility to work with these
   - “Future-proof”?
3. **Scalability**
   - Much of the unstructured data in question is at scale → big data
   - PET of choice must be able to handle this
   - DP: “schematic” that allows for optimization and tailoring to the problem at hand
4. Promising **Privacy Preservation**
   - Great initial results in *privacy experiments*
   - Best way to exhibit: decreasing performance of attack models
   - Also: privacy statistics, e.g. N_w and S_w

DP in NLP: Limitations

Utility

- Not necessarily negative
- Often though: clear utility hit
  - Lower ε → lower utility
- But: "no free lunch"
  - More privacy must mean less "something else"
- Quantifying and evaluating this tradeoff is crucial
- Ultimately: design choice
  - What do I **value** more?
  - ε: a great indicator for balancing privacy and utility

Structural Limitations

- DP imposes structure to inherently unstructured text
  - i.e. reasoning about text as documents
  - "Transfer" from standard DP
- d_x-privacy addresses this
  - Criticism: how much are we willing to deviate?
- Another limitation: DP assumes static nature
  - But much of NLP data today is streaming / dynamic
  - Time value?

Downstream vs. Natural Language

- Related to utility
- Basically with DP + NLP: add noise to text representations
- Downstream task is end goal:
  - Effects of noise are less important
- However:
  - Projecting back to natural language is non-trivial
  - So far: grammatically shoddy, quite unnatural language
- Bottleneck: choice of embedding model

DP in NLP: Limitations (cont.)

Explainability

- Key issue
- Essentially: how to explain what is going on with DP + NLP
- Some questions:
  - How does DP fit with NLP?
  - What changes does text undergo?
  - When is text *truly* private?
- DP gives a good start
  - ε useful but not complete
  - Still somewhat cryptic
- Ultimately, greater transparency is needed
  - (Hopefully) obtained through time with further research

Not a Fix-all

- Perhaps self-evident
- Good: quantifiable guarantee
- But:
  - At the cost of structural limitations / design requirements
  - Only when using data to make broader generalizations (“learning tasks”)
- Mapping to text representation not always desired or possible
- Needed: better study of DP vs. other PETs w.r.t. NLP

Future Work

1. Further exploration of the **privacy-utility tradeoff** when applying DP to NLP
   - Maximum privacy + minimum utility loss = ultimate goal
   - Important for better explainability
2. Integration of DP into **modern NLP architectures**
   - So far: only rather “simple” NLP pipelines
   - Would be nice to see with advanced sequence models (e.g. transformers), adversarial networks, etc.
3. Compatibility with **more recent text representation models**
   - For example, with contextual word embeddings (e.g. BERT) – not easy!
4. The role of DP in **non-static data settings**
   - DP’s role, applicability, effectiveness
   - Investigating DP with streaming datasets, the “time value” of ε
5. Other **generalizations of standard Differential Privacy**
   - d_x-privacy is a good start, others possible?
6. Explaining DP in NLP
   - Privacy concerns the individual user
   - → maintain transparency with this person!

Results (Deliverables)

List of deliverables from this Guided Research:

1. Guided Research Report
2. Journal Submission to the PET Symposium
3. (8) Learning Nuggets for new “DP in NLP” learning path:
   - Privacy Vulnerabilities in Natural Language Processing
   - Differential Privacy in Natural Language Processing
   - Differential Privacy: Applications to NLP
   - d_x-privacy: Generalized Differential Privacy
   - d_x-privacy: Applications to NLP
   - The Exponential Mechanism
   - Benefits of Differential Privacy in NLP
   - Limitations of Differential Privacy in NLP
4. Web App for corresponding Learning Nuggets (developed by SEBA Lab Course)

DP in NLP Learning Nugget / Web App Examples

DP in NLP Learning Nugget Examples (1)

Differential Privacy in Natural Language Processing

Introduction

- Differential Privacy is a relatively novel concept that boasts a mathematical foundation in quantifying privacy preservation.
- The topic has been largely researched in the context of protection of individuals within a structured dataset.
- When considering Natural Language Processing and its goals, it is necessary to change our view of Differential Privacy.
- This can be accomplished in two ways: one will be discussed here, another way can be found in d_x-privacy: Generalized Differential Privacy.
- It is also useful to view Differential Privacy juxtaposed to other privacy-preserving techniques, especially when coming from the standpoint of privacy in NLP.
- Also important to note is that the study of Differential Privacy's applicability to NLP is an ongoing research topic, and the current state certainly encompasses the infancy stages in this study.

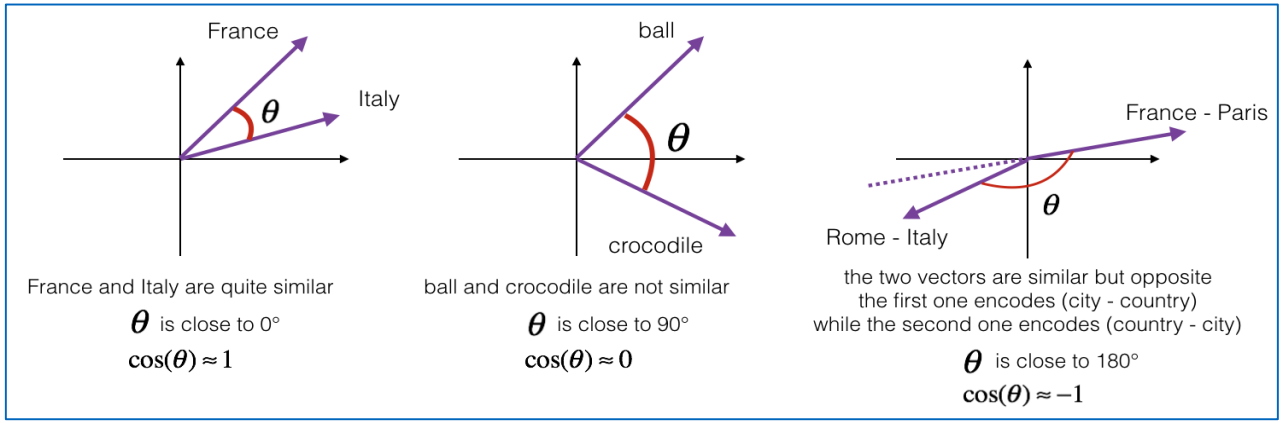
Learning Objectives

- You will (re-)learn what Differential Privacy is, and what its core concepts and guarantees are
- You will encounter some other privacy-preserving techniques, and realize how they might not be the best option when dealing with unstructured text data
- You will see how this notion of Differential Privacy can be transferred to Natural Language Processing

DP in NLP Learning Nugget Examples (2)

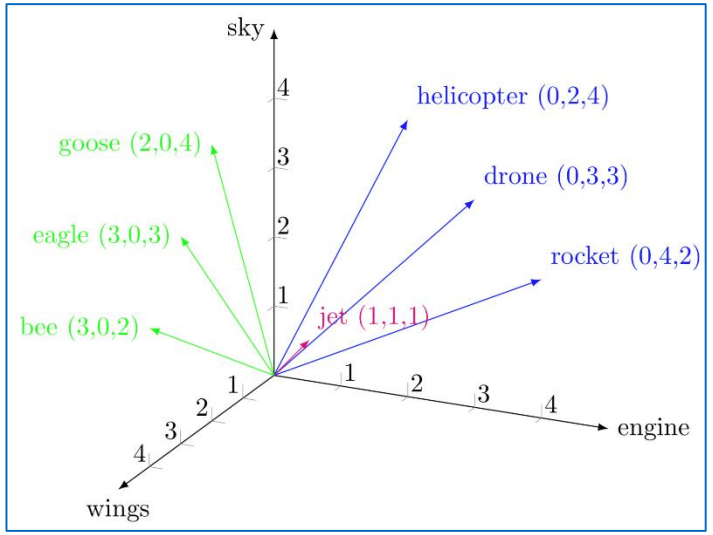
Word Embeddings and their Risks

- The growing success of word embeddings for use as *general purpose language models* has increased their attention and utilization.
- To recap, word embeddings:
  - Are a representation of words in vector space, have a set dimension, and are real-valued
  - Provide a convenient basis for computation due to its numerical nature
  - Can be easily used to define semantic similarity (i.e. cosine similarity between vectors)
  - Can be generalized even further to represent sentences or documents
- Recently, word embeddings have been used in high-profile application such as Google's search engine.
- To achieve an accurate language model, these embedding models are usually pre-trained on **billions** of sentences.
  - Naturally, one can imagine that these sentence might contain private or sensitive information.
  - To further exacerbate the issue, embeddings are designed to capture information and relations within text.
- As a result, language models based upon these word embeddings can be at risk.
  - This can especially be true when considering NLP tasks in which sensitive data is used.
  - The real-valued representation of embeddings might be misconstrued as safe, but:
  - We will see in a few slides how these vulnerabilities might be exploited.



Right: a simplified example of word embedding vectors (dim=3), https://corpling.hypotheses.org/495

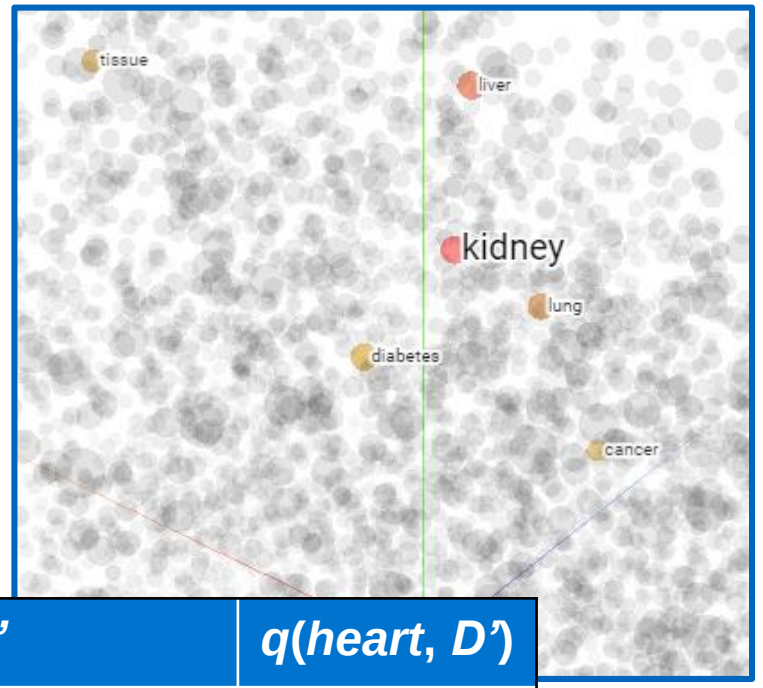
Left: an explanation of the cosine similarity measure used with word embeddings, https://datascience-enthusiast.com/DL/Operations_on_word_vectors.html



DP in NLP Learning Nugget Examples (3)

The Exponential Mechanism for NLP: an Example

- Let's see a simplified example showing how the Exponential Mechanism can be useful to NLP applications
  - Imagine the following simplified embedding space* (see right):
    - *heart* and its 5 nearest neighbors
    - Following the notation:
      - *D = heart*
      - $\mathcal{R}$ = {*liver*, *lung*, *tissue*, *diabetes*, *cancer*}
      - The corresponding scores are given in the table (1 – cosine distance)
- If we apply the Exponential Mechanism, we obtain the following probabilities:
  - Here, ε is chosen to be 2
  - Sensitivity would be 0.59 – 0.38 = 0.21
  - Probabilities are obtained by normalization



D'	$\exp(\frac{\varepsilon q(D,r)}{2\Delta q})$	Pr[D']
liver	16.602	0.367
lung	9.647	0.213
tissue	6.560	0.145
diabetes	6.315	0.140
cancer	6.108	0.135

D'	*q(heart, D')*
liver	0.590
lung	0.476
tissue	0.395
diabetes	0.387
cancer	0.380

*Created from https://projector.tensorflow.org/

DP in NLP Learning Nugget Examples (4)

Key Takeaways

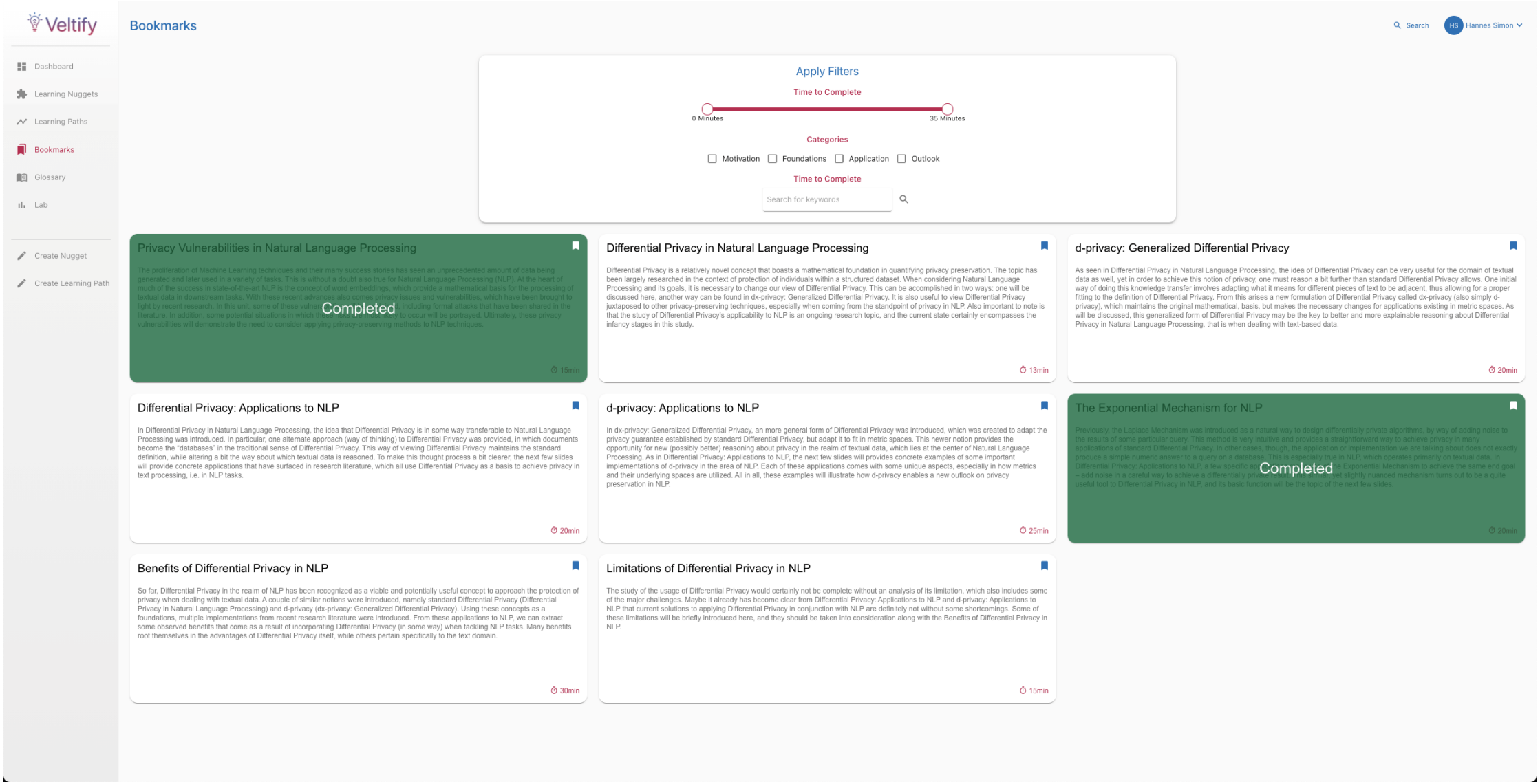
- The application of Differential Privacy to NLP does not come without its limitations.
- Chief among these limitations are an often-observed utility hit, a general challenge for better explainability about privacy in the textual domain, and the fact that Differential Privacy is only a good answer in specific use cases.
- An important takeaway here is that much like the privacy-utility tradeoff, the need for improved privacy cannot come without sacrifices in other areas – “there is no free lunch”.
- While it is true that some of these limitations can cause concern, one must not forget that it comes as the price of improved privacy preservation in an age where this is becoming more and more crucial*.
- Furthermore, these limitations serve as a great basis for future work in pursuit of better privacy-preserving techniques (for NLP).

Outlook

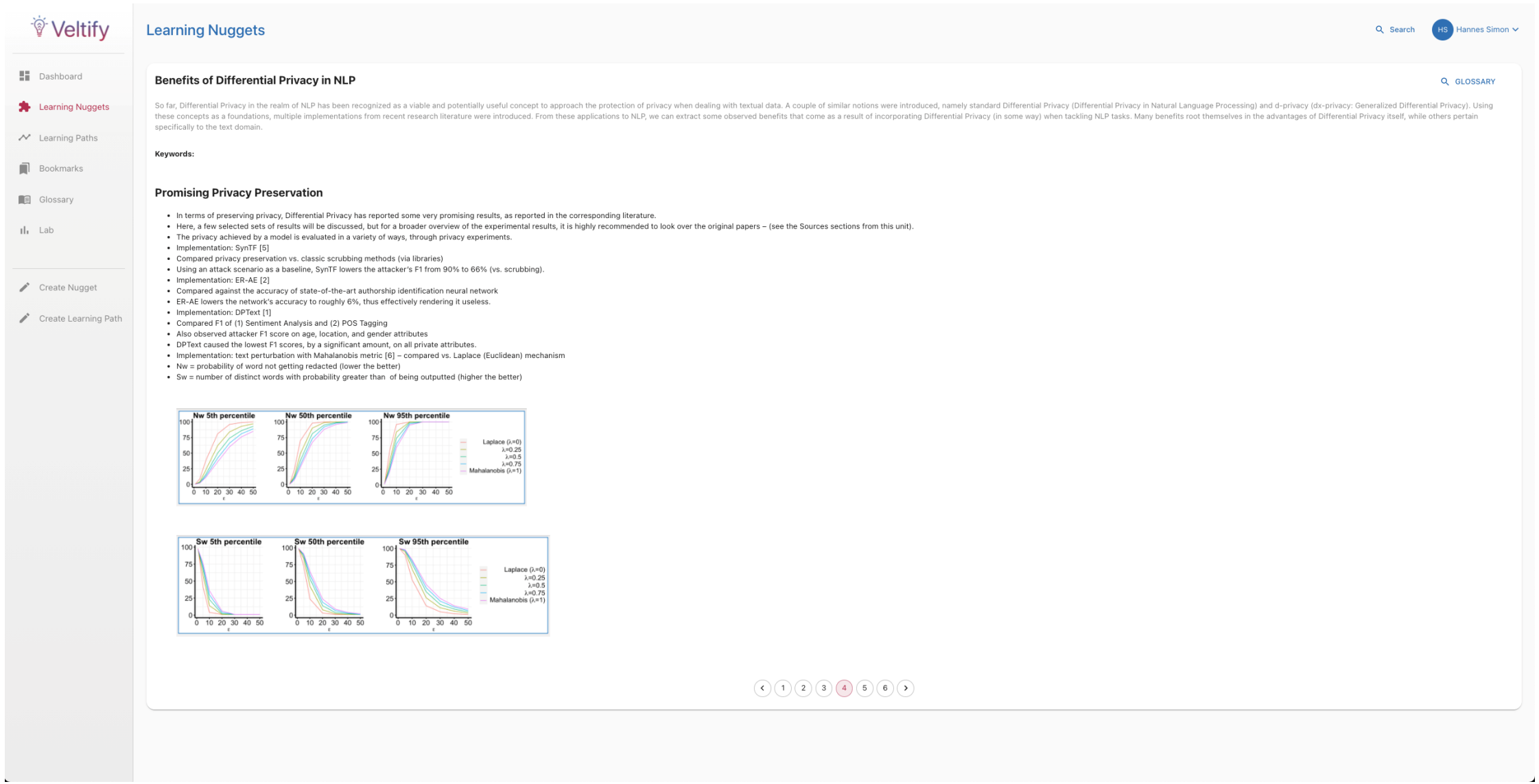
- If not done already, check out Benefits of Differential Privacy in NLP for an overview of some of the best reasons for using Differential Privacy in conjunction with NLP techniques
- The limitations discussed here take root in many of the ideas discussed within this unit, particularly Privacy Vulnerabilities in Natural Language Processing, Differential Privacy in Natural Language Processing, d-privacy: Generalized Differential Privacy, Differential Privacy: Applications to NLP and d-privacy: Applications to NLP

*https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/

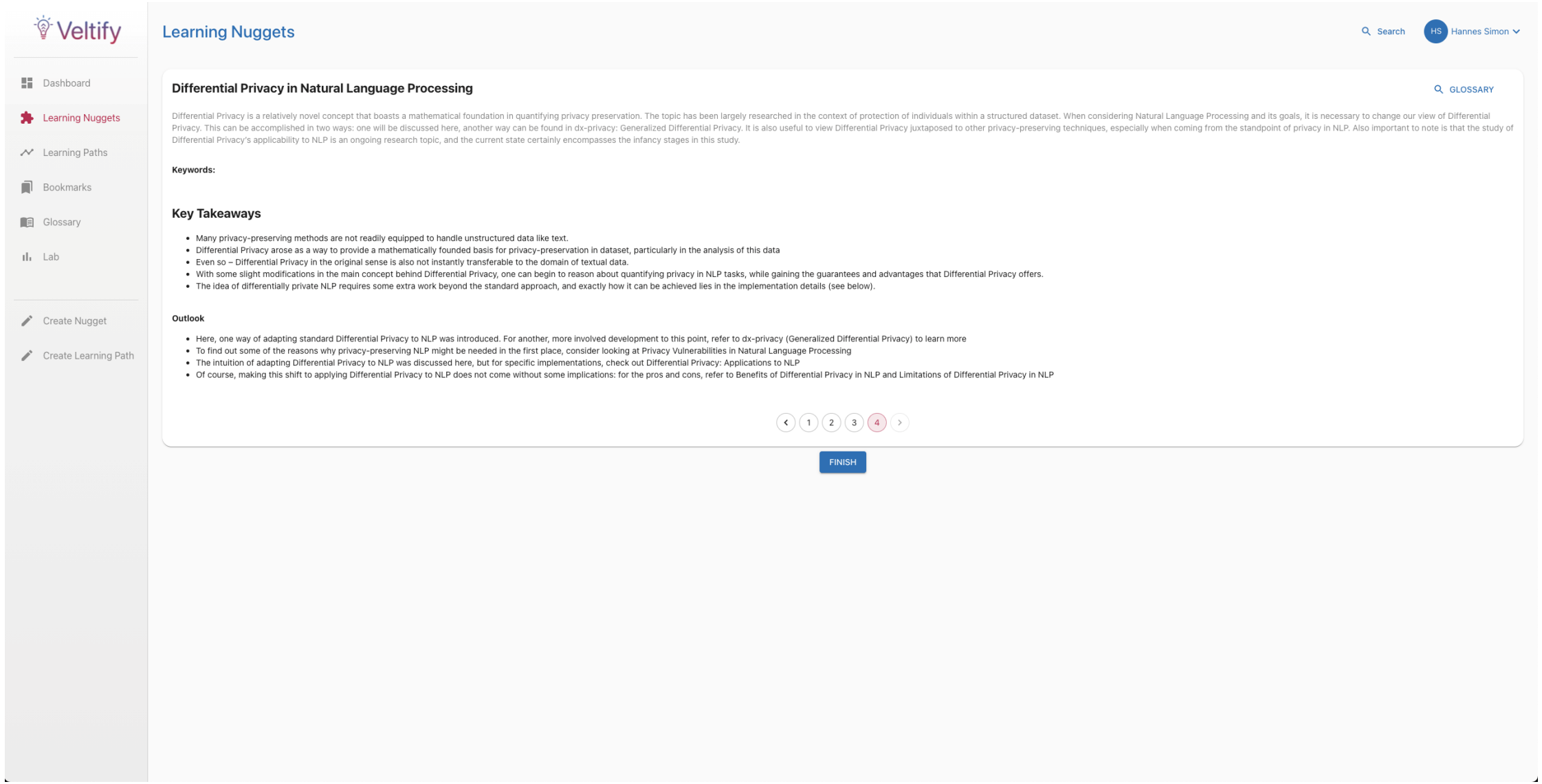
DP in NLP Web App Examples (1)



DP in NLP Web App Examples (2)



DP in NLP Web App Examples (3)



Conclusion

- With the amount and nature of textual data nowadays, there are bound to be some risks in NLP
- DP + NLP is initially a somewhat challenging match, but:
  - When reasoned about, serves a great candidate for privacy preservation
  - In some respects, the “best” to date
- Several very interesting aspects:
  - How to represent text
  - How to define certain concepts, i.e. the “individual”, adjacency
  - In what ways can noise be efficiently added (to the pipeline)
- Many possible future directions
  - Still a young, budding, somewhat theoretical field
- Will DP + NLP become commonplace?
  - Experts: hopefully!
  - My opinion: only a matter of time
- In the end: much learned, new appreciation for the field (+ new take on NLP in general)

Stephen Meisenbacher
stephen.meisenbacher@tum.de

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de