PRIVACY IN OPEN SEARCH: A REVIEW OF CHALLENGES AND SOLUTIONS

SAMUEL SOUSA, GRAZ UNIVERSITY OF TECHNOLOGY, AUSTRIA
CHRISTIAN GUETL, GRAZ UNIVERSITY OF TECHNOLOGY, AUSTRIA
ROMAN KERN, KNOW-CENTER GMBH, AUSTRIA

3rd International Open Search Symposium
October 13th, 2021

Privacy

- **Privacy** is a concept related to limiting the extend of **information** an individual is willing to share [1].
- **Data privacy** laws have been approved in many countries in recent years, e.g., European Union (EU)'s **General Data Protection Regulation (GDPR)** [2].
- **GDPR** grants EU residents the control over their personal data.

Source: http://gdpr.eu

Motivation

- **Privacy** has been gaining attention over many fields, e.g., Machine Learning, Deep Learning and Information Retrieval (IR).
- Many **privacy-preserving techniques** have been proposed.
  - Encryption.
  - Multi-party computation.
  - Differential privacy.
  - Etc.
- Numerous **attacks** target private data.
  - Eavesdropping.
  - Reverse engineering.
  - Etc.
- **Need for open data**, due to scientific, governmental and press reasons.

Goals

- **This work** aims at:
  - Pointing out open **privacy challenges**.
  - Reviewing **privacy-preserving techniques**.
- Main focus: tasks featuring **text data**.
  - **Explicitly** presented private information, e.g., **names of people**.
  - **Implicitly** presented private information, e.g., **location description**.

Methodology

- Creation of **expressions for search** on Google Scholar.
- Collection of **112 papers** in total.
- **11 papers** were selected for review.

	A	B
1	**Search Expression**	**Number of Selected Papers**
2	"cooperative web crawling" AND "privacy" AND "deep learning"	0
3	"distributed web crawling" AND "privacy" AND "deep learning"	0
4	"web indexing" AND "privacy" AND "deep learning"	0
5	"web search" AND "privacy" AND "deep learning"	12
6	"web science" AND "privacy" AND "deep learning"	10
7	"web mining" AND "privacy" AND "deep learning"	5
8	"content retrieval" AND "privacy" AND "deep learning"	1
9	"content analysis" AND "privacy" AND "deep learning"	0
10	"distributed databases" AND "privacy" AND "deep learning"	0
11	"distributed systems" AND "privacy" AND "deep learning"	0
12	"cloud security" AND "privacy" AND "deep learning"	0
13	"cluster security" AND "privacy" AND "deep learning"	0
14	"distributed systems security" AND "privacy" AND "deep learning"	0
15	"search and retrieval" AND "privacy" AND "deep learning"	2
16	"deep learning for search"	0
17	"geospatial search" AND "privacy" AND "deep learning"	1
18	"geospatial analysis" AND "privacy" AND "deep learning"	0
19	"Innovative search" AND "privacy" AND "deep learning"	0
20	"data search" AND "privacy" AND "deep learning"	0
21	"data retrieval" AND "privacy" AND "deep learning"	0
22	"crisis management" AND "privacy" AND "deep learning"	2

Paper	Citations	Venue	Selected
Collaborative Deep Learning for Recommender Sy	1137	KDD	1
Text summarization using unsupervised deep lea	122	Expert Systems With Applications	1
Scene Text Detection and Recognition: The Deep Learning Era	89	International Journal of Computer Vision	1
Abstractive text summarization using LSTM-CNN based deep learning	79	Multimedia Tools and Applications	1
Twitter Spam Detection based on Deep Learning	74	ACM International Conference Proceeding Ser	1
Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion	51	Information Processing and Management	1
Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation	34	SIGIR	1
From web search to healthcare utilization: privacy-sensitive studies from mobile data	24	Journal of the American Medical Informatics Association	1
Hybrid query expansion using lexical resources and word emb e ddings for sentence retrieval in question answering	23	Information Sciences	1
Multilingual opinion mining on YouTube – A convolutional N-gram BiLSTM word embedding	21	Information Processing and Management	1
Tagvisor: A Privacy Advisor for Sharing Hashtags	21	WWW	1
Speak Up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment	18	NACL	1
Query-oriented text summarization based on hypergraph transversals	18	Information Processing and Management	1
Context-Aware Document Term Weighting for Ad-Hoc Search	9	World Wide Web	1
Improving named entity recognition in noisy user-generated text with local distance neighbor	6	Neurocomputing	1
Linear feature extraction for ranking	5	Information Retrieval Journal	1
DeepHate: Hate Speech Detection via Multi-Faceted Text Representations	4	ACM Conference on Web Science	1
Counterfactual Online Learning to Rank	3	Advances in Information Retrieval	1
Facilitating SQL Query Composition and Analysis	1	Proceedings of the ACM SIGMOD International Conference on Management of	1
Computing Local Sensitivities of Counting Queries with Joins	1	ACM SIGMOD International Conference on Management of Data	1
Learn2Link: Linking the Social and Academic Profiles of Researchers	0	AAAI Conference on Web and Social Media	1
Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter	0	IEEE Access	0
Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifier	132	IEEE Symposium on Security and Privacy Wo	0
A neural algorithm for a fundamental computing			

Privacy Challenges in Search Tasks

- **Ad-hoc search**:
  - Bag-of-words models may allow data re-identification [3].
- **Query expansion**:
  - Disclosures of query or document content [4].
- **Feature extraction for ranking**:
  - Recovery of original documents by attackers.
- **Online learning for ranking**:
  - Disclosure of user behavior [5].
- **Query composition**:
  - Memorization of query content [6].

Privacy Challenges in Healthcare Tasks

- **Electronic health records**:
  - Queries by malicious users or computation parties, e.g., corrupted servers.

Health Record Health Care Provider's Examination		
Name: Scott Smith	**Gender**: Male	**Date of Birth**: 01/01/1990
Height: 180 cm	**Weight**: 78 KG	**Blood Pressure**: 120/80 mm Hg
Medical History: • COVID-19 infection was confirmed on 02/02/2020, resulting in hospitalization for 14 days with help of mechanical ventilators. • The patient underwent physiotherapy from 2008 until 2010. • The patient underwent a leg operation on 03/03/2008 for correcting a fracture of the right leg femur.		
Allergies: Seafood, nuts, pollen, and dust		
Health Conditions: Diabetes Type II		

Privacy Challenges in Social Media Tasks

- **Opinion mining**:
  - Unintended data memorization.
  - Model inversion attacks.
- **Advisor for hashtag sharing**:
  - Prediction of user location by attackers.
- **Social media profile linking**:
  - Tracking of user behavior.
  - Revealing the identities of anonymous profile owners.

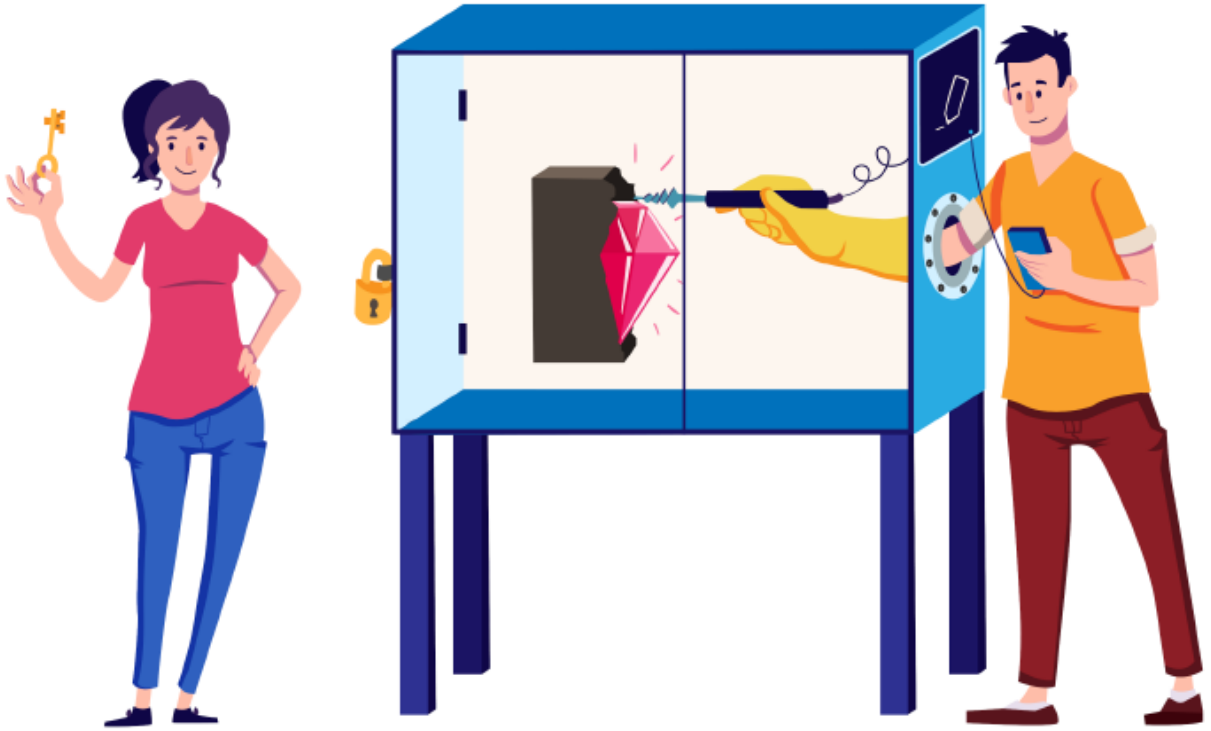
Privacy Challenges in Recommendation Tasks

- **Recommender systems**:
  - Memorization of user behavior.
  - Memorization of user preferences.
  - Memorization of a user's search history.

Privacy-Preserving Methods

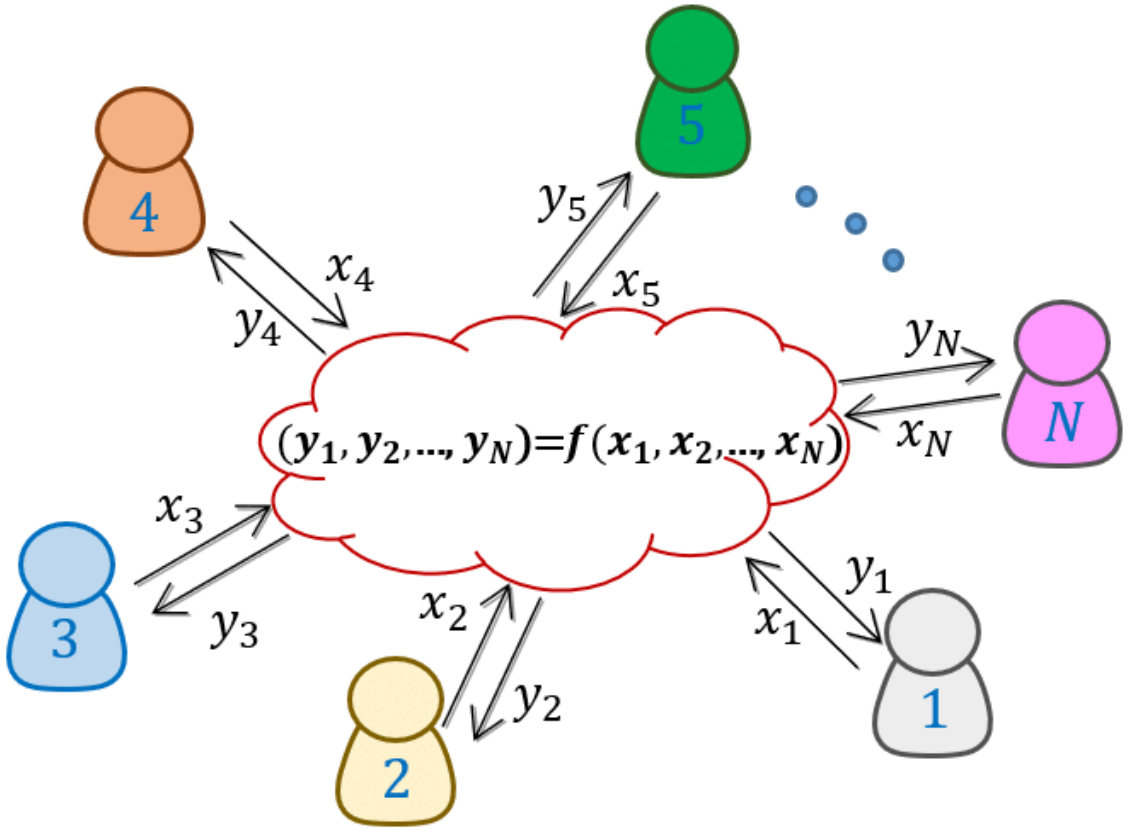
- HOMOMORPHIC ENCRYPTION
- MULTI-PARTY COMPUTATION
- DIFFERENTIAL PRIVACY
- FEDERATED LEARNING

Homomorphic Encryption



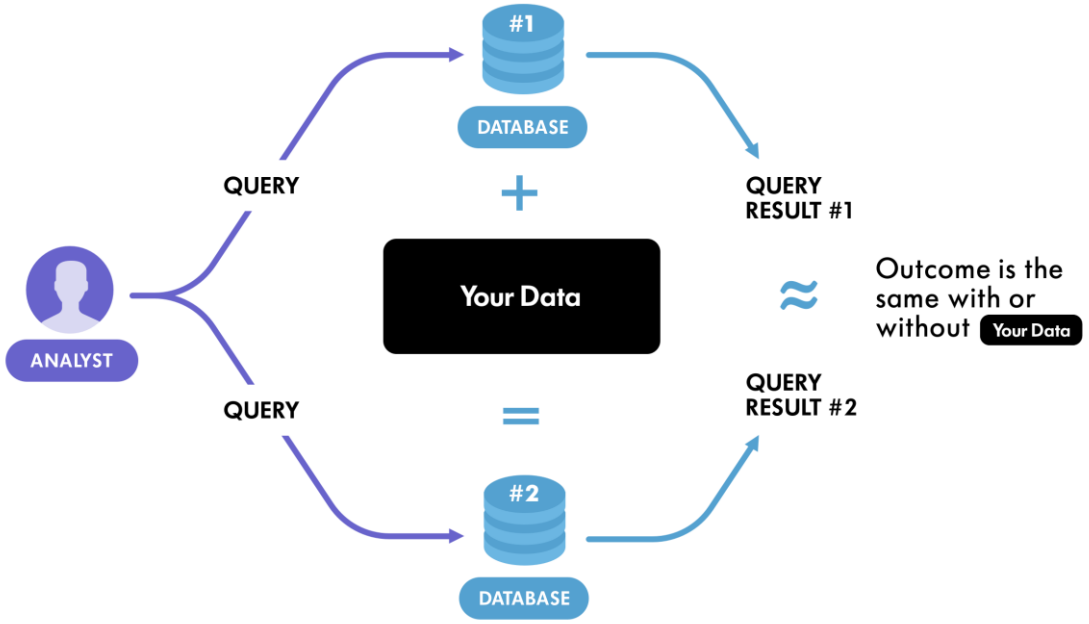
Source: Know-Center GmbH

Multi-Party Computation



Source: Reference [7]

Differential Privacy



Source: Reference [8]

Federated Learning

Source: Know-Center GmbH

Discussion

- The **choice** for a convenient privacy-preserving method: balance between privacy protection and computational costs.
- **Encryption**: Applicable for non-trusted computation parties, e.g., servers.
- **Differential privacy**: It brings formal guarantees against attacks.
- **Federated learning**: It prevents data exchanges in distributed scenarios, e.g., recommender systems.

Conclusion and Future Work

- Privacy is a **critical point** for the development of Open Search systems.
- **Compliance** with **data protection regulations** is a must.
- **Future works**:
  - Addressing privacy challenges for **Open Search use cases**.
  - Studying and discussing the **compliance** with legal requirements, e.g., **EU's GDPR**.

Contact

Thank you!

Questions?

Contact: ssousa@know-center.at

cguetl@tugraz.at

rkern@know-center.at

This work is supported by the EU's Horizon 2020 project TRUSTS under grant agreement No. 871481

References

[1] D. Brickley, M. Burgess, and N. Noy, “Google dataset search: Building a search engine for datasets in an open web ecosystem,” in *The World Wide Web Conference*, 2019, pp. 1365–1375.

[2] E. Commission, “2018 reform of EU data protection rules,” https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf, 2018, date: 2018-05-25, URL Date: 2019-06-17.

[3] Z. Dai and J. Callan, “Context-aware document term weighting for ad-hoc search,” in Proceedings of TheWeb Conference 2020, 2020, pp. 1897–1907.

[4] C. Carpineto and G. Romano, “A survey of automatic query expansion in information retrieval,” ACM Computing Surveys (CSUR), vol. 44, no. 1, pp. 1–50, 2012.

[5] S. Zhuang and G. Zuccon, “Counterfactual online learning to rank,” in European Conference on Information Retrieval. Springer, 2020, pp. 415–430.

References

[6] Z. Zolaktaf, M. Milani, and R. Pottinger, “Facilitating sql query composition and analysis,” in Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020, pp. 209–224.

[7] Lemus, Mariano, et al. "Generation and distribution of quantum oblivious keys for secure multiparty computation." *Applied Sciences 10.12* (2020): 4080.

[8] Winton Group, Ltd., “Using Differential Privacy to Protect Personal Data”, https://www.winton.com/research/using-differential-privacy-to-protect-personal-data, 2018, date: 2018-09-04, URL Date: 2021-10-11.