European Commission

DEDS

European Joint Doctorate in
Data Engineering for Data Science (DEDS)

Doctoral Project Plan
Synopses-Driven Data Integration and Federated Learning
Eros Fabrici

1 Project Summary

The term Big Data refers to data sets that are too large or complex to be dealt with by traditional data processing software. In particular, Big Data capture 4 dimensions: Velocity, Variety, Veracity, and Volume. This large quantity of data available nowadays has a lot of statistical and business value, therefore their analysis is of core importance for business decision-making. Nevertheless, the data involved in those processes contains a lot of sensitive information regarding individuals (e.g. health sector). There is therefore also the need to guarantee the privacy of the individuals while being able to extract insights and patterns from the data.

Data Integration is the set of processes to gather and bridge data from heterogeneous sources together in order to have a unified view. The premise of data integration is to make data more freely available and easier to consume and process by systems and users. Research on Data Integration started more than 50 years ago [17, 34] but as we entered the Big Data era, new challenges arose, which include scaling [13] data integration processes while guaranteeing the privacy [49, 20] of the individuals involved in the datasets. Over the last decade new techniques have been applied for improving computational performance, consisting of the use of parallelizing the computation by using big data processing platforms [13], or algorithmically, namely using summarization techniques [8], used for approximate fast and approximate querying, improving the performance of Machine Learning processes [21, 2, 28] as well as in some data integration scenarios. Despite the progress made, it is difficult to combine efficiency (the integration is completed with no, or very few errors), computational performance and privacy altogether [23].

Machine Learning and Data Integration have really close relationship [12]. In particular, it is possible to leverage the first to improve the performance of the second and vice-versa. An evolving branch of Machine Learning (ML) is Federated Learning (FL), which consists in building a model in a federated setting (when the data is distributed across different data owners) and model in a collaborative way, without moving the data to a central server. This new technique is a good match with the need of privacy in ML nowadays. Despite its advantages, there are a lot of open problems in FL [29]. In particular, data in FL needs to be pre-processed (e.g. align the schemas, record linkage) in order to start the training process. This would require the data to move out of the edge-devices, therefore threatening privacy.

This Ph.D. aims to explore algorithms, data structures and ML for scaling Data Integration tasks while guaranteeing privacy and achieving good performance, tailored to the Federated Learning pre-processing workflow.

2 Scientific Content of the Doctorate Project

2.1 Background

2.1.1 Data Integration

Data Integration (DI) is the practice of consolidating data from disparate sources into a unified view. It has been studied since the birth of relational databases. It is characterized by three main steps:

1. **Schema alignment**, a process that takes as input a set of different schemas on the same domain and outputs a *mediated schema*, an *attribute matching* and a *schema mapping*.
2. **Record linkage**, also referred as entity resolution, computes a partitioning of the set of records from different datasets, such that each partition identifies the records that refer to a distinct entity.
3. **Data fusion**, whose aim is to identify which are the best records to represent a specific entity, when a source provides conflicting values.

2.1.2 Federated Learning

FL is a ML approach where a model is trained across multiple decentralized data owners. Each data owner trains a local model and then, either in a centralized, decentralized or heterogeneous approach, builds a global model. It differs from a distributed machine learning as the data is not expected to be identically distributed. Besides the advantage of having a distributed computation, guaranteeing more efficiency, it gained a lot of popularity due to the fact that data is not exchanged between the parts involved, thus guaranteeing privacy.

It has gained a lot of popularity both in research and industry, in particular in digital health [38].

2.1.3 Differential Privacy & Synopses for Big Data

Differential Privacy (DP) is a technique for sharing datasets' information without compromising the privacy of the individuals. The idea is to add noise to the data such that the new distribution is close to the real one, but not equal.

More specifically, it guarantees that any sequence of outputs (response to a query) is equally likely to occur, independent of the presence or absence of any individual in the dataset. The main advantage of DP is that it is robust against membership and information inference attacks, i.e. when an attacker is able to de-anonymize an anonymized dataset via linkage attacks. A good example of de-anonymization attack is the one proposed by Narayanan et.al. [35] where from an anonymized dataset of Netflix subscribers' viewing history, they de-anonymized it through a linkage attack with the Internet Movies Database, revealing the users' apparent political preferences and other sensitive information.

Synopses or summaries are a set of technique and probabilistic data structures to compute compact description of big datasets. These methods compute lossy, compact summaries of data, from which it is possible to carry out interactive analyses and queries. They have been used extensively for streaming data, now gaining popularity also in ML and FL.

2.2 State of the Art

2.2.1 Privacy-aware Data Integration in the Big Data Era

Privacy in the context of data management has gained a lot of popularity over the last decade, as public awareness about issues in management sensitive data increased. Due to this, privacy

became of central importance in the field of Big Data Management, analytics and processing.

In the particular case of DI, privacy-preserving techniques has been used extensively in literature, especially for record linkage. With regards to schema matching and data fusion, there are fewer works on guaranteeing privacy, most of the works are based on guaranteeing efficiency, by using both rule-based and learning-based approaches [40].

Schema matching. Schema matching aligns attributes and data types. It is one of the oldest problems studied for data integration and the traditional approaches consist in extracting knowledge according to a predefined schema. They can be categorized as it follows:

- **Schema-level matchers**, where only the metadata is considered (e.g. column labels, data-types). Linguistic matching is mostly used here (stemming, tokenization, etc.) [3].
- **Instance-level matchers**, where the content of the columns is used for matching by using probabilistic approaches [10] or rule-based approaches.
- **Hybrid matchers** that combine the two matchers just described.

In schema matching the problem of *volume* has to be taken into account only in few specific cases, for example when we consider millions sources from the web [37], but not in typical DI scenarios, where the number of sources is limited.

For what we know so far, there is only one work in the field of **Privacy-Preserving Schema Alignment** [43]. The idea is that two parties map their schemas to a global one, which is defined by a third party. Subsequently, they send the encrypted mapped schemas to the third party that will run an intersection between the two and then send back the results to the parties.

This is a basic and effective technique, but this implies the schemas to be non-opaque. If that is not the case, the only way to infer the relationships between the columns is to use information theoretic techniques or synopses to estimate the inclusion dependencies between the columns. [30] implements such an algorithm. More precisely, they capture the relationships between the columns of a dataset with a graph containing information theoretic values. In this way they reduce the problem to a graph matching problem.

Another work [10] consist in using synopses to mine the structure of a database. In particular, they use MinHashes, q-grams, and linear sketches to estimate joinability, join directions and composite attributes. It does not use any privacy preserving technique, but it is straightforward the chance to use differential privacy algorithms.

Other works [50] we explored that make the use of Locality Sensitive Hashing (LSH) and MinHashes, called LSH Esemble. MinHash is a technique for quickly estimating how similar two sets are. LSH is a dimensionality reduction method, that makes use of hash functions that guarantees that the hashed elements maintain a similar distance in the reduced space. Briefly, LSH Esemble splits the single MinHash signatures in b bags. Then each bag is hashed with LSH in a bucket. Given a query column, it is first indexed with the algorithm just described, then a hashed column is a candidate where at least one signature maps in the same bucket where the query column is hashed. Here, differential privacy can be applied to guarantee a formal privacy guarantee, as similarly done in [18].

A line of work [41, 42] uses a sketching technique for reducing a column to a small sample, to then compute the correlation between different columns.

As it is possible to deduce from the above, there are very few works on privacy preserving techniques for schema alignment.

Record linkage. Record Linkage, also called Entity Resolution, consists in finding records, among different data sources, that refers to the same real world entity. It is the most important problem in integrating data from different sources.

Generally, it proceeds in three steps:

1. **blocking records** that are likely to be a match;
2. **compare pairs of records** in order to decide if it's a match;
3. **clustering records** according to the previous step's results.

Approaches consisted mostly in rule-based techniques [17, 19] for the first two steps, while for clustering either rule-based or optimizing a particular objective function [22].

Recently, supervised learning approaches (e.g. Support Vector Machines, Decision Trees, Random Forest) showed to obtain high precision and recall [9], at the cost of generating training labels, i.e. to obtain a precision and recall of 99% on linking a pair of datasets, 1.5M training labels are required [11].

Performance and efficiency is not only the main concern of Record Linkage. In a real-world scenario, the data involved in the linkage may be sensitive, and methods to guarantee the privacy of the individuals is a major concern. Privacy-Preserving Record Linkage (PPRL) identifies the set of techniques that aim to link different datasets in a privacy-preserving manner. There are four main approaches for guaranteeing privacy, which may also be used together (hybrid solutions) in RL: cryptographic protocol and Secure Multiparty Computation (SMC), differential privacy, masking and generalization techniques (e.g. k-anonymity).

Initially, Secure Multiparty Computation (SMC) techniques were used, in particular, the Paillier crypto-system [36]. These protocols are reliable and very effective, with the downside of a very prohibitive computational cost.

For the knowledge we have so far, we have two works which use SMC and cryptography. [1] use commutative encryption for matching. No technique for reducing the computational cost (e.g. blocking) is used and they assume that the different tables share a common identifier, which is not true in most DI scenarios. A more recent research article [31] shows that, by adapting a Private Set Intersection algorithm [26], just by using SMC, they were able to improve the performance of the state of the art algorithm that we will describe below. Practically [31] use a scoring function to score the records for then sorting them globally and finally run matching algorithm that slides a window across the sorted records and compare only the records inside the window, all done under the guarantees of SMC.

A line of research that aims to improve performance while guaranteeing privacy consist of a series of work that use embedding techniques with or without differential privacy. More in detail, [43] embed the records by using a technique called SparseMap [25], which is a technique for embedding an object to a Euclidean space, by representing it as a vector whose values are the distances between the object and a subset of reference values. The reference set is defined by one of the parties, thus the security is armed in malicious setting. A similar approach is described in [47] where they embed the records in a complex plane.

To overcome this problem and to have a more formal privacy guarantee, [4] uses a different embedding technique based on frequent grams. Briefly, the two parties mine a common base, by using the counts of the top-k frequent grams and then by using the Laplace mechanism [15] for the counts.

Other works that include masking/embedding techniques make use of Bloom Filters [44, 14]. A Bloom Filter is a probabilistic data strucuture to test whether an element is a member of a set or not. In these works they are used to mask the records, then the matching is done by computing the Dice similarity metric between two BFs. This is not the optimal use of the BFs as their main purpose is to check if an element is not a member of a set, as false negatives are not possibles, while false positives yes. Moreover, BFs does not guarantee any formal privacy guarantee, and these works have been attacked right after their publications [5, 7, 6, 33]. To improve privacy, [46, 45] uses Randomized Response mechanism on Bloom Filters to guarantee Differential Privacy. On the other hand, this lead to poor utility/privacy tradeoff, i.e. to obtain a good linkage quality, a very high privacy parameter is required.

Finally, a line of work focused on the combination of DP and SMC. More in detail, [27]

run a blocking algorithm, then, for each block, a summary of it is computed (the extent of the block for each attribute and the block size) to then add DP noise to it. After that, a filtering phase is performed, to run a SMC matching algorithm between blocks that overlap (based on the extent). A more recent work apply an analogue technique, but the distribution of noise in the blocks is more advanced, by using a private indexing algorithm [39], which also satisfies the new definition called Output Constrained DP, see [24].

2.2.2 Machine Learning and DI

Machine Learning and DI have a very close relationship [11]. For example, it is possible to build a matching function by learning it with labelled examples. The main problem of this approach is that it needs a huge amount of data to obtain high accuracy/recall results. On the other hand, DI is basically a preprocessing step before feeding the data to a ML model.

The main problem of Record Linkage, is the democratization of this process. In particular, with the introduction of distributed representation of words, [16] introduced an approach for tuple embedding. The main advantage wrt the classic embedding approaches is that this is able to capture semantic similarities between words.

2.2.3 Federated Learning

Federated Learning (FL) has been proposed by Google [32]. The idea is to build a global ML model from datasets that are distributed across edge devices, without moving the data. An unbalanced and non-IID (identically and independently distributed) data partitioning across a massive number of unreliable devices with limited communication bandwidth was introduced as the defining set of challenges [29]. Privacy is one of the essential properties of FL. Many techniques exist in literature (e.g. Secure Multiparty Computation, homomorphic encryption), but **Differential Privacy** represents a *de facto* standard for Privacy in many areas (querying, synthetic data generation, etc.) as it guarantees a better computational performance rather than cryptographic approaches.

FL can be categorized as it follows:

- **Horizontal Federated Learning**. Horizontal FL refers when the federated datasets share the same feature space (the column names) but not the sample space (rows). This system assumes that all the participants are honest and secured against an honest-but-curious server [48]. Usually, the learning steps in this system are: (1) each data owner *trains a local model* then the (2) *gradients are sent* to the central server, which applies a (3) *secure aggregation*. Finally, the (4) *model updates* computed by the central server are sent back to the data owners and their local models get updated.
- **Vertical Federated Learning**. Vertical FL is applicable when the datasets share the sample ID space, but the features are different. In this scenario, data pre-processing is required, in particular *schema alignment* and *entity resolution*. These phases require exchanging data with a third party to do the pre-computation, therefore security is more difficult to guarantee in this case.

2.3 Project Objectives

The goal of this PhD can be divided in two main objectives. The first is to develop Privacy Preserving DI algorithms to improve the computational performance with respect to state-of-the-art algorithms. The second objective is to experiment those algorithms in the FL process, namely aligning the schemas of the datasets and linking the records. In particular, we plan to experiment how these novel algorithms impact the performance of the FL models, with respect to FL solutions that do not align the data or that use other DI algorithms.

Briefly, our objectives are:

1. **O1.** Provide Privacy Preserving DI solutions (i.e. schema alignment and PPRL) by tackling the computational performance limitations.
2. **O2.** Prototype a DI plugin to integrate to the FL data integration phase, with the aim of improving the learned models' performance.

2.4 Key Methods

We will try to apply the following methods to achieve the project's objectives and ensure the production of high quality results:

- Study the literature review of the current Privacy-Preserving Data Integration techniques and analyze their strengths and weaknesses. This step can be considered continuous over the course of the PhD as more and more scientific content is published continuously at top tier conferences and journals.
- Study Differential Privacy and Synopses in order to understand how to apply them for DI integration algorithms.
- After understanding and analyzing the offerings of current solutions, we will propose algorithms that will use Differential Privacy and Synopses with the aim of obtaining solutions that are computationally performant and differentially private, without losing efficacy (i.e. precision and recall). The goal here is to prototype those solutions in a simulated federated environment.
- Regarding the evaluation of the proposed solutions, appropriate benchmarks will be considered ensuring the correctness of our results.

2.5 Significance and Outcome

Big Data is the core of most businesses nowadays. Data are being generated, analyzed, and used at an unprecedented scale, and data-driven decision-making regards all aspects of society. As the value of data increases when it can be linked with other data, addressing big data integration is critical. Big Data introduced also privacy concerns, that have been dealt with policy regulations (e.g. GDPR).

However, most of the solutions applied now struggle to guarantee good privacy, good efficacy and good computational performance altogether. Moreover, the actual literature made a small use of techniques like Differential Privacy and Synopses, which are de facto standard in other areas (e.g. private data analysis and processing big data streams). The doctoral project will investigate further the application of these techniques to improve the performance of DI algorithms. Finally, these techniques will be applied to FL scenarios, to observe how the Synopses-driven and differentially-private DI algorithms can improve the quality of the learned models.

The expected outcome of this project includes (1) presentation of our work in top tier conferences and journals, (2) collaboration with other research teams and/or industrial partners (e.g. secondment), to exchange ideas and boost the results of our work, and (3) open-sourcing critical components of our work.

3 Co-supervisors/Candidate Co-operation Agreements

The project will be carried out in three years during which the PhD student will stay in one research institution and one university. During the first and the third year, the candidate will work in Athena Research Center (ARC) under the supervision of Prof. Minos Garofalakis

(ARC). During the second year, the program will take place in Universitat Politecnica de Catalunya (UPC) under the supervision of Prof. Josep Lluís Berral García and PhD Besim Bilalli (UPC). The project will be a joint work of all parties, hence close co-operation is expected in the following way.

The progress of the project will be validated through frequent meetings between the candidate and his supervisors. The candidate will meet on a weekly basis with his home supervisor and one or two times per month with his host supervisor (the opposite when he will be hosted at UPC). Following typical business practice, the expectations and tasks planned for each meeting will be clearly communicated in advance, with a reasonable notice, both from the supervisors to the candidate and vice-versa. Standard tools of the trade will be used to boost collaboration, such as a shared repository for documents and code artifacts (e.g., Mendeley Library, GitHub, etc.), communication platforms (e.g. Skype, Teams).

4 Work Plan

4.1 Timetable

The PhD is a 3-year-long study, spanning from May 1st, 2022 to February 28th, 2025.

Time	**Plan**
05/22 - 09/22 (ARC)	Winter School at ARC Preparation of the two months study plan Literature review of Privacy Preserving Data Integration eBISS 2022 Summer School
Milestones	**Submission of two months study plan**
09/22 - 05/23 (ARC)	Participation in 1 PhD courses as shown in the courses' table Preparation of the 11 months study plan Winter School at AAU Tutorial "Privacy Preserving Data Integration" on VLDB 2023
Milestones	**Submission of 11 months study plan**
05/23 - 03/24 (UPC)	Summer School at UPC Participation in 2 PhD courses as shown in the courses' table Synopses-Driven Private Schema Alignment Design and develop an algorithm for PPRL using synopses and differential privacy
03/24 - 06/24 (SPR)	Secondment at Spring Techno
06/2024 - 02/2025 (ARC)	Prototype a plug-in for integrating PPDI to Federated Learning systems Evaluate the plug-in Write the thesis
Milestones	**Submission of the PhD thesis**

4.2 Thesis Outline

The thesis will be constituted by the research papers that will be written during the whole duration of the project. The thesis report will include (1) an introduction chapter providing background knowledge, motivation, research challenges, objectives and contributions, (2) a chapter about the state of the art (survey paper), (3) one chapter for each published research paper that will include technical details of the implementation, evaluation etc., and (4) a last chapter concluding the work, discussing lessons learned and providing future directions.

4.2.1 Tentative Publication List

The publications that can be considered at the moment of writing the Doctorate Project Plan are the following:

1. A tutorial paper targeting to illustrate the state of the art of Privacy Preserving Data Integration. Our main contributions will be to initiate a new reader to the field of PPDI, by illustrating the current applied techniques through examples. Comparison of the methods, limitations, research challenges, and future work will also be included.
   - Journal: VLDB 2023
2. Two research papers presenting, respectively, a Differentially Private (DP) and Synopses-driven schema alignment algorithm and a DP and Synopses-driven Record Linkage algorithm. Both papers will include an evaluation of the algorithms (e.g. computational complexity, benchmarks).
   - Conferences: respectively VLDB 2024, SIGMOD 2024
3. A demo paper introducing a plug-in for Privacy Preserving DI for FL systems, which will include an evaluation of how the FL model performance is impacted by the proposed DI algorithms.
   - Conference: IEEE ICDE 2025

5 Proposed Education and Training Programme

During the PhD studies, it is necessary to have research activities adding up to at least 30 ECTS credits. The ECTS points should be divided between general and research-related courses. Courses can either be taken in National and Kapodistrian University of Athens or in Universitat Politecnica de Catalunya, with conference attendance and other activities contributing as well.

Activity	At	ECTS	Type	Time	Status
Big Data Management	UPC	6	Project	Autumn'23	Planned
Introduction to Big Data Analytics	UPC/BSC	3	Project	Winter'23	Completed
Research Methods	UPC	6	General	Autumn'23	Planned
Winter School (ARC)	ARC	3	Project	Spring'22	Completed
Summer School (ULB)	ULB	3	Project	Spring'22	Completed
Winter School (AAU)	AAU	3	General	Fall'22	Planned
Summer School (UPC)	UPC	3	General	Spring'23	Planned
Conference Attendance	TBD	6	Project	Multiple	Planned
Greek Language course	NKUA	-	Project	Multiple	Planned
Spanish Language course	UPC	-	Project	Multiple	Planned

6 Knowledge Dissemination and Participation to Scientific Events

We plan to disseminate the product knowledge by publishing papers in top tier conferences, such as ACM SIGMOD, VLDB, IEEE ICDE, EDBT, etc. and journals, such as VLDB J., ACM Surveys, IEEE TKDE, Information Systems, etc. Moreover, we will pursue opportunities to expose our work to additional outlets (e.g., AI Summit, ACM/IEEE local chapters, meet-ups) through presenting talks and tutorials or giving demonstrations, in order to open a communication channel with the big data engineering, and big data management communities. In this way, we will (a) advertise our work and explore collaboration and exploitation opportunities, and (b) collect valuable feedback that will ameliorate and/or redirect our research.

7 External Co-operation

The doctorate candidate will spend time studying both in Greece and Spain. Furthermore, a secondment of three months will take place, where the candidate will join Spring Techno, where he will work on a complex Federated Learning scenario with real data. During the following three years, all ESRs will meet in four different winter/summer schools to present their work, receive feedback, exchange ideas, and get exposed to new challenges. During these schools, candidates will have the opportunity to get in touch with academic and non-academic partners, presenting them their findings, reflecting on new opportunities, and opening the way for further collaboration. Finally, the candidate may co-operate with external researchers or research teams, in case that his work can be combined or merged with similar works of others.

8 Agreements on Immaterial Rights to Patents

Patents and immaterial rights will be handled according to general rules applied by Athena Research Center, National and Kapodistrian University of Athens, and Universitat Politecnica de Catalunya.

9 Financing Budget

As regards the long-term career ambitions and objective of the candidate, pursuing a PhD offers This project is one of the 15 ESRs of Data Engineering for Data Science PhD programme, which is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 955895. The funding covers expenses related to the successful completion of the project, such as work equipment, research experiments, training activities and others that are relevant to the programme.

10 Career Development Plan

After completing the three years programme, the candidate will be an independent researcher, ready to join the academia or pursue a competitive professional career in the industry. During the PhD programme, the candidate will be initiated into research by acquiring competitive research skills, will gain expertise in state-of-the-art technologies, cultivate soft skills, such as communication, presentation and co-operation, and work within a professional environment.

References

[1] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant. "Information Sharing Across Private Databases". en. In: ().

[2] Jesus Antonanzas, Marta Arias, and Albert Bifet. "Sketches for Time-Dependent Machine Learning". In: 1 (2021), pp. 1–9. arXiv: 2108.11923. URL: http://arxiv.org/abs/2108.11923.

[3] Philip Bernstein, Jayant Madhavan, and Erhard Rahm. "Generic Schema Matching, Ten Years Later". In: *PVLDB* 4 (Aug. 2011), pp. 695–701. DOI: 10.14778/3402707.3402710.

[4] Luca Bonomi et al. "Frequent grams based embedding for privacy preserving record linkage". en. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. Maui, Hawaii, USA: ACM Press, 2012, p. 1597. ISBN: 978-1-4503-1156-4. DOI: 10.1145/2396761.2398480. URL: http://dl.acm.org/citation.cfm?doid=2396761.2398480 (visited on 11/14/2022).

[5] Peter Christen et al. "Efficient cryptanalysis of bloom filters for privacy-preserving record linkage". In: *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I 21*. Springer. 2017, pp. 628–640.

[6] Peter Christen et al. "Pattern-mining based cryptanalysis of Bloom filters for privacy-preserving record linkage". In: *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III*. Springer. 2018, pp. 530–542.

[7] Peter Christen et al. "Precise and fast cryptanalysis for Bloom filter based privacy-preserving record linkage". In: *IEEE Transactions on Knowledge and Data Engineering* 31.11 (2018), pp. 2164–2177.

[8] Graham Cormode et al. "Synopses for massive data: Samples, histograms, wavelets, sketches". In: *Foundations and Trends in Databases* 4.1-3 (2011), pp. 1–294. ISSN: 19317883. DOI: 10.1561/1900000004.

[9] Sanjib Das et al. "Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. Chicago, Illinois, USA: Association for Computing Machinery, 2017, pp. 1431–1446. ISBN: 9781450341974. DOI: 10.1145/3035918.3035960. URL: https://doi.org/10.1145/3035918.3035960.

[10] Tamraparni Dasu et al. "Mining Database Structure; or, How to Build a Data Quality Browser". In: *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*. SIGMOD '02. Madison, Wisconsin: Association for Computing Machinery, 2002, pp. 240–251. ISBN: 1581134975. DOI: 10.1145/564691.564719. URL: https://doi.org/10.1145/564691.564719.

[11] Xin Luna Dong and Theodoros Rekatsinas. "Data Integration and Machine Learning: A Natural Synergy". en. In: *Proceedings of the 2018 International Conference on Management of Data*. Houston TX USA: ACM, May 2018, pp. 1645–1650. ISBN: 978-1-4503-4703-7. DOI: 10.1145/3183713.3197387. URL: https://dl.acm.org/doi/10.1145/3183713.3197387 (visited on 10/09/2022).

[12] Xin Luna Dong and Theodoros Rekatsinas. "Data integration and machine learning: A natural synergy". In: *Proceedings of the ACM SIGKDD* (2019), pp. 3193–3194. DOI: 10.1145/3292500.3332296.

[13] Xin Luna Dong and Divesh Srivastava. "Big data integration". In: *Proceedings - International Conference on Data Engineering* September (2013), pp. 1245–1248. ISSN: 10844627. DOI: 10.1109/ICDE.2013.6544914.

[14] Elizabeth A Durham et al. "Composite bloom filters for secure record linkage". In: *IEEE transactions on knowledge and data engineering* 26.12 (2013), pp. 2956–2968.

[15] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". en. In: *FNT in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407. ISSN: 1551-305X, 1551-3068. DOI: 10.1561/0400000042. URL: http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042 (visited on 12/08/2022).

[16] Muhammad Ebraheem et al. "Distributed Representations of Tuples for Entity Resolution". In: *Proc. VLDB Endow.* 11.11 (July 2018), pp. 1454–1467. ISSN: 2150-8097. DOI: 10.14778/3236187.3236198. URL: https://doi.org/10.14778/3236187.3236198.

[17] Ivan P. Fellegi and Alan B. Sunter. "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210. DOI: 10.1080/01621459.1969.10501049.

[18] Natasha Fernandes, Yusuke Kawamoto, and Takao Murakami. "Locality Sensitive Hashing with Extended Differential Privacy". en. In: vol. 12973. arXiv:2010.09393 [cs, math]. 2021, pp. 563–583. DOI: 10.1007/978-3-030-88428-4_28. URL: http://arxiv.org/abs/2010.09393 (visited on 11/25/2022).

[19] Helena Galhardas et al. "Declarative Data Cleaning: Language, Model, and Algorithms". In: *VLDB* (July 2001).

[20] Aris Gkoulalas-Divanis et al. "Modern Privacy-Preserving Record Linkage Techniques: An Overview". In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 4966–4987. ISSN: 15566021. DOI: 10.1109/TIFS.2021.3114026.

[21] Rémi Gribonval et al. "Sketching Datasets for Large-Scale Learning (long version)". In: (2020), pp. 1–35. arXiv: 2008.01839. URL: http://arxiv.org/abs/2008.01839.

[22] Oktie Hassanzadeh et al. "Framework for Evaluating Clustering Algorithms in Duplicate Detection". In: *Proc. VLDB Endow.* 2.1 (Aug. 2009), pp. 1282–1293. ISSN: 2150-8097. DOI: 10.14778/1687627.1687771. URL: https://doi.org/10.14778/1687627.1687771.

[23] Xi He et al. "Composing Differential Privacy and Secure Computation: A case study on scaling private record linkage". In: *Proceedings of the ACM Conference on Computer and Communications Security* (2017), pp. 1389–1406. ISSN: 15437221. DOI: 10.1145/3133956.3134030. arXiv: 1702.00535.

[24] Xi He et al. *Composing Differential Privacy and Secure Computation: A case study on scaling private record linkage*. en. arXiv:1702.00535 [cs]. Sept. 2017. URL: http://arxiv.org/abs/1702.00535 (visited on 10/10/2022).

[25] G.R. Hjaltason and H. Samet. "Properties of embedding methods for similarity searching in metric spaces". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.5 (2003), pp. 530–549. DOI: 10.1109/TPAMI.2003.1195989.

[26] Yan Huang, David Evans, and Jonathan Katz. "Private set intersection: Are garbled circuits better than custom protocols?" In: *NDSS*. 2012.

[27] Ali Inan et al. "Private record matching using differential privacy". In: *International Conference on Extending Database Technology*. 2010.

[28] Jiawei Jiang et al. "SketchML: Accelerating distributed machine learning with data sketches". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2018), pp. 1269–1284. ISSN: 07308078. DOI: 10.1145/3183713.3196894.

[29] Peter Kairouz et al. "Advances and open problems in federated learning". In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210. ISSN: 19358245. DOI: 10.1561/2200000083. arXiv: 1912.04977.

[30] Jaewoo Kang and Jeffrey F Naughton. "On Schema Matching with Opaque Column Names and Data Values". en. In: (), p. 12.

[31] Basit Khurram and Florian Kerschbaum. "SFour: A Protocol for Cryptographically Secure Record Linkage at Scale". en. In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. Dallas, TX, USA: IEEE, Apr. 2020, pp. 277–288. ISBN: 978-1-72812-903-7. DOI: 10.1109/ICDE48307.2020.00031. URL: https://ieeexplore.ieee.org/document/9101375/ (visited on 11/02/2022).

[32] Jakub Konečný et al. *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*. 2016. DOI: 10.48550/ARXIV.1610.02527. URL: https://arxiv.org/abs/1610.02527.

[33] Martin Kroll and Simone Steinmetzer. "Automated cryptanalysis of bloom filter encryptions of health records". In: *arXiv preprint arXiv:1410.6739* (2014).

[34] Ramez El-Masri and Gio Wiederhold. "Data Model Integration Using the Structural Model". In: *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*. SIGMOD '79. Boston, Massachusetts: Association for Computing Machinery, 1979, pp. 191–202. ISBN: 089791001X. DOI: 10.1145/582095.582127. URL: https://doi.org/10.1145/582095.582127.

[35] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[36] Pascal Paillier. "Public-key cryptosystems based on composite degree residuosity classes". In: *International conference on the theory and applications of cryptographic techniques*. Springer. 1999, pp. 223–238.

[37] Rakesh Pimplikar and Sunita Sarawagi. "Answering table queries on the web using column keywords". In: *Proceedings of the VLDB Endowment* 5.10 (June 2012), pp. 908–919. DOI: 10.14778/2336664.2336665. URL: https://doi.org/10.14778%2F2336664.2336665.

[38] Prayitno et al. "A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications". In: *Applied Sciences (Switzerland)* 11.23 (2021). ISSN: 20763417. DOI: 10.3390/app112311191.

[39] Fang-Yu Rao et al. "Hybrid Private Record Linkage: Separating Differentially Private Synopses from Matching Records". en. In: *ACM Trans. Priv. Secur.* 22.3 (July 2019), pp. 1–36. ISSN: 2471-2566, 2471-2574. DOI: 10.1145/3318462. URL: https://dl.acm.org/doi/10.1145/3318462 (visited on 10/05/2022).

[40] Sebastian Riedel et al. "Relation extraction with matrix factorization and universal schemas". In: *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference* June (2013), pp. 74–84.

[41] Aecio Santos et al. "A Sketch-based Index for Correlated Dataset Search". en. In: *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. Kuala Lumpur, Malaysia: IEEE, May 2022, pp. 2928–2941. ISBN: 978-1-66540-883-7. DOI: 10.1109/ICDE53745.2022.00264. URL: https://ieeexplore.ieee.org/document/9835690/ (visited on 10/21/2022).

[42] Aécio Santos et al. "Correlation Sketches for Approximate Join-Correlation Queries". en. In: *Proceedings of the 2021 International Conference on Management of Data*. arXiv:2104.03353 [cs]. June 2021, pp. 1531–1544. DOI: 10.1145/3448016.3458456. URL: http://arxiv.org/abs/2104.03353 (visited on 10/21/2022).

[43] Monica Scannapieco et al. "Privacy preserving schema and data matching". en. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*. Beijing, China: ACM Press, 2007, p. 653. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247553. URL: http://portal.acm.org/citation.cfm?doid=1247480.1247553 (visited on 11/14/2022).

[44] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters". In: *BMC medical informatics and decision making* 9.1 (2009), pp. 1–11.

[45] Rainer Schnell and Christian Borgs. "Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage". en. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. Barcelona, Spain: IEEE, Dec. 2016, pp. 218–224. ISBN: 978-1-5090-5910-2. DOI: 10.1109/ICDMW.2016.0038. URL: http://ieeexplore.ieee.org/document/7836669/ (visited on 01/18/2023).

[46] Wanli Xue et al. "Sequence Data Matching and Beyond: New Privacy-Preserving Primitives Based on Bloom Filters". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 2973–2987. DOI: 10.1109/TIFS.2020.2980835.

[47] Mohamed Yakout, Mikhail J. Atallah, and Ahmed Elmagarmid. "Efficient Private Record Linkage". en. In: *2009 IEEE 25th International Conference on Data Engineering*. ISSN: 1084-4627. Shanghai, China: IEEE, Mar. 2009, pp. 1283–1286. ISBN: 978-1-4244-3422-0. DOI: 10.1109/ICDE.2009.221. URL: http://ieeexplore.ieee.org/document/4812521/ (visited on 11/14/2022).

[48] Qiang Yang et al. "Federated machine learning: Concept and applications". In: *ACM Transactions on Intelligent Systems and Technology* 10.2 (2019), pp. 1–19. ISSN: 21576912. DOI: 10.1145/3298981. arXiv: 1902.04885.

[49] Shui Yu. "Big privacy: Challenges and opportunities of privacy study in the age of big data". In: *IEEE access* 4 (2016), pp. 2751–2763.

[50] Erkang Zhu et al. "LSH Ensemble: Internet-Scale Domain Search". en. In: arXiv:1603.07410 [cs]. arXiv, July 2016. URL: http://arxiv.org/abs/1603.07410 (visited on 01/06/2023).

European Commission

DEDS

European Joint Doctorate in Data Engineering for Data Science (DEDS) Doctorate Project Plan[1] Thesis Title First Name Last Name

This page must be completed and sent together with the project plan/report in a pdf file to the chair of the Candidate Progress Committee.

Project title: .....
Name of doctorate candidate: .....
Email: .....
Supervisor: .....
Home University: .....
Co-supervisor: .....
Host University: .....
Secondment supervisor: .....
Partner organisation: .....
Date of enrolment: .....
Expected date of completion: .....

Signatures

The Doctorate Candidate

.....
Date:

The Supervisor from the Home University

Professor .....
Date:

The Supervisor from the Host University

Professor .....
Date:

The Secondment Supervisor

.....
Date:

[1] Choose the appropriate heading among the three

The Chair of the Candidate Progress Committee

Professor .....

Date: