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

**Doctoral Project Plan[1]**
**Thesis Title**
**First Name Last Name**

# 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 captures to 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 [16, 30] but as we entered the Big Data era, new challenges arose, which include scaling [14] data integration while guaranteeing the privacy [47, 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 [14], or algorithmically, namely using summarization techniques [6], used for approximate fast and approximate querying, improving the performance of Machine Learning processes [21, 1, 25] 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 [24].

Machine Learning and Data Integration have really close relationship [13]. 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 is Federated Learning, which consists in building a Machine Learning model in a federated setting (when the data is distributed across edge devices) and model in a collaborative way, without moving the data to a central server.

This Ph.D. aims to explore algorithms, data structures and Federated Learning for scaling Data Integration tasks while guaranteeing privacy and achieving good performance.

# 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 relational databases. It is characterized by three main

[1]Choose the appropriate heading. Based on the PhD Study Plan of Aalborg University, available at `http://www.phd.teknat.aau.dk/intranet/phd-study-plan`.

steps:

1. **Schema alignment**, a process that takes as input a set of different schemas on the same domain an 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** aim is to identify which are the best records to represent a specific entity, when a source provide conflicting values.

#### 2.1.2 Federated Learning

Federated Learning is a machine learning approach where a model is trained across multiple decentralized edge-devices, in an edge computing fashion. Each device trains a local model and then, either in a centralized, decentralized or heterogeneous approach, build 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 transportation [15], Industry 4.0 [41] and digital health [36].

#### 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.

Synopses or summaries are a set of technique and probabilistic data structures to compute compact description of big datasets. They gained a lot of popularity over the last decade due to the rise of Big Data.

## 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 two this, privacy became of central importance in the field of Big Data Management, analytics and processing.

In the particular case of Data Integration, privacy-preserving techniques has been used extensively in literature, especially for record linkage. In particular, differentially-private record linkage and cryptography has been used extensively [27, 45, 5, 29, 20, 32, 4, 29]. As regards schema matching and data fusion, there are fewer works on guaranteeing privacy, most of the work is based on guaranteeing efficiency, by using both rule-based and learning-based approaches [39, 38, 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.) [2].
- **Instance-level matchers**, where the content of the columns is used for matching by using probabilistic approaches [42, 9] or rule-based approaches.

- **Hybrid matchers** that combine the two matchers just described.

In schema matching the problem of *volume* have to be taken into account only in few specific cases, for example when we consider millions sources from the web [35], but not in typical DI scenarios, where the number of sources is limited. Universal Schema [38] has revolutionized schema alignment. It consists in extracting (subject, predicate, object) triples, where the predicate can be any word or phrase from texts and instead of outputting mappings between predicates, it adds inferred triples. This is done through matrix factorization [38], while recently it is improved by using Recurrent Neural Networks [7, 31].

For what we know so far, there is no work in literature regarding privacy-aware schema matching. This because most of the techniques are based on the schemas' metadata. Nevertheless, if Instance-level approaches are used, it may be useful to use techniques for guaranteeing the privacy of the individuals, e.g. differential privacy.

**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** to decide if it's a match;
3. **clustering records** according to the previous step's results.

Approaches consisted mostly in rule-based techniques [16, 17] for the first two steps, while for clustering either rule-based or optimizing a particular objective function [23].

Recently, supervised learning approaches (e.g. Support Vector Machines, Decision Trees, Random Forest) showed to obtain high precision and recall [8], 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. Initially, Secure Multiparty Computation (SMC) techniques were used, in particular, the Paillier crypto-system [33]. These protocols are reliable and very effective, with the downside of a very prohibitive computational cost. In other to improve performance, by applying secure transformations to the data [3], such as embedding records to different spaces and then mining them with differential privacy. This comes with the cost of having less accurate results.

Generally, the PPRL protocols proposed for secure two-party private record linkage are not able to meet the following three requirements altogether, without making strict assumptions: (1) **full end-to-end privacy**, (2) **perfect precision and recall** for the matching records and (3) **sub-quadratic computational complexity** [24, 22]. Moreover, multi-party PPRL is a more realistic scenario and only in the last years it has been tackled, with still limited results [44, 45, 43].

**Data Fusion.** Data fusion resolves conflicts between different data sources, by selecting the best record per entity. Access to highly accurate data is critical for industry applications, such as knowledge graph search, so data fusion is often an important step in data integration.

The main methods for data fusion are rule-based [12] and also data-mining based [34]. Graphical models are also used in this context [18] as well as semi-unsupervised approaches [37].

Privacy-preserving data fusion has not been studied deeply in literature. There are a few context specific works, for example [10] identifies privacy issues and future research directions for data fusion in Internet-Of-Things and [19] which focuses on Differential Privacy in the context of Cyber-Physical Systems.

#### 2.2.2 Federated Learning

Federated Learning (FL) has been proposed by Google [28]. 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 [26]. 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 *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 security against an honest-but-curious server [46]. Usually, the learning steps in this system are: (1) each data owner *train a local model* then the (2) *gradients are sent* to the central server, which applies a (3) *secure aggregation* on them and received by the federation. 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

## 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.
- After understanding and analyzing the offerings of current solutions, we will propose algorithms that will use Differential Privacy and Synopses that will satisfy all the three requirements mentioned in the state of the art. The goal here is to design and implement the solution 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

# 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. Oscar Romero (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 Proposed Education and Training Programme

The DEDS education and training programme is composed of several activities.

- **Research**, where doctoral candidates work on a novel research problem guided by two supervisors who will advise them to gradually become independent researchers.
- **Research-specific courses**, aimed at providing doctoral candidates with focused state-of-the art technical skills pertaining to their research topic.
- **Innovation and entrepreneurship courses**, aimed at complementing the scientific training of doctoral candidates with business-related aspects such as entrepreneurship, intellectual property rights, etc.
- **Methodological and communication courses**, aimed at introducing the necessary research methods and communication skills.
- **Language courses**, aimed at introducing the local language at each partner university.
- **Summer and winter schools**, wherein candidates will obtain feedback about their research from invited researchers and practitioners, as well as get international contacts in both academia and industry.
- **Tutoring**, whereby candidates will be involved in teaching activities (e.g., supervising student projects and delivering exercises) while being coached by their supervisors or other experienced staff.
- **Knowledge dissemination and participation to scientific events**, aimed at allowing doctoral candidates to present and confront their findings, thereby familiarising themselves with essential practices such as peer-review and public debating.
- **External cooperation and secondments**, aimed at ensuring that the candidate participate actively in another research environment outside his/her home and host universities. These activities are realised typically with DEDS partner organisations.

Please detail in the following subsections your personalised education and training programme, taking into account your previous background and future career prospects. This programme must be approved from both co-supervisors.

Activities adding at least 30 ECTS credits must be outlined. A tabular listing of all activities performed or to be performed during the doctorate project is to be included. Group the activities according to the categories specified above. For each activity, the title, time, location, organiser, and ECTS credits must be included together with an indication of whether the activity has been completed. Please use this table:

| Activity | Place/Organised by | ECTS | General/Project course | Status |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

*TPR: The table of activities must be updated with planned and hitherto completed activities.*
*RPR: The contents and the extent of the completed activities must be reported. It is expected that all training activities have been finalised in order to devote the last year of the project for finalising the Doctoral Dissertation.*

### Section 4.1. Planned Courses

Courses adding at least 20 ECTS credits must be outlined.

Only courses at doctorate level are approved. If a course at master level is deemed to be highly relevant for the doctorate project, the co-supervisors can establish a study group on the topic, which includes the master course and additional reading/discussion to bring it up to doctorate level. A written report on participation in a study group must be completed to get course credit. To ensure the scientific level, the study circle must be headed by a member of the scientific staff, who is Professor or Associate Professor (senior scientist level). A 2-3 ECTS study circle organised by the co-supervisors on the state of the art in the research field of the doctorate study is recommended.

*TPR: The course table should be updated with more specific information for the completed courses, as well as with the rest of planned courses for the rest of the doctorate project.*
*RPR: The table must be updated, reporting the complete set of courses that are completed in the doctorate project.*

### Section 4.2. 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 TODS, 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.

Each participation to scientific events must be accompanied by a written report by the doctorate candidate that relates the specific activity to the doctorate project. This report must be of general value for the project. Activities that relate to workshop and conference participation must not exceed 6 ECTS credits.

*TPR: Updated plan for dissemination of knowledge and findings from the doctorate project other than those listed in Section 2 must be specified.*
*RPR: The final realisation of the knowledge and findings dissemination from the doctorate project other than those listed in Section 2 must be reported.*

### Section 4.3. 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 of 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.

## Section 5. 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.

## Section 6. Financing Budget

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.

## Section 7. Career Development Plan

In this section, the candidate's career plan and development are described. When the thesis is handed in (M4), this section is revisited in a self-contained document called "Career Development Plan" (CDP) to be signed by the candidate and the supervisors.

### Section 7.1. Long-Term Career Objectives

*DPP: Describe long-term career goals (over 5 years) and how to become able to reach those goals.*
*TPR, RPR, CDP: Update as needed if the career plans have evolved.*

### Section 7.2. Objectives Covered in Project

Describe which development objectives will be/have been achieved in the project with respect to

1. Research skills and techniques
2. Research management and co-operation
3. Communication skills

4. Other professional training
5. Networking activities and opportunities
6. Other activities with professional relevance

*CDP: Include also published and accepted papers as well as completed course activities such that the document is self-contained.*

## Section 8. References

# References

[1] 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.

[2] 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.

[3] Luca Bonomi, Li Xiong, and James J. Lu. "LinkIT: Privacy preserving record linkage and integration via transformations". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2013), pp. 1029–1032. ISSN: 07308078. DOI: 10.1145/2463676.2465259.

[4] Luca Bonomi et al. "Frequent grams based embedding for privacy preserving record linkage". In: *ACM International Conference Proceeding Series* (2012), pp. 1597–1601. DOI: 10.1145/2396761.2398480.

[5] Chris Clifton and Gunther Schadow. "Privacy Preserving Data Integration and Sharing". In: (), pp. 1–10.

[6] 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.

[7] Rajarshi Das et al. *Chains of Reasoning over Entities, Relations, and Text using Recurrent Neural Networks*. 2016. DOI: 10.48550/ARXIV.1607.01426. URL: https://arxiv.org/abs/1607.01426.

[8] 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, 1431â1446. ISBN: 9781450341974. DOI: 10.1145/3035918.3035960. URL: https://doi.org/10.1145/3035918.3035960.

[9] Tamraparni Dasu et al. "Mining database structure; or, how to build a data quality browser". In: (2002), p. 240. DOI: 10.1145/564691.564719.

[10] Wenxiu Ding et al. "A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion". In: *Information Fusion* 51 (2019), pp. 129–144. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2018.12.001. URL: https://www.sciencedirect.com/science/article/pii/S1566253518304731.

[11] Xin Luna Dong. "Challenges and Innovations in Building a Product Knowledge Graph". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, p. 2869. ISBN: 9781450355520. DOI: 10.1145/3219819.3219938. URL: https://doi.org/10.1145/3219819.3219938.

[12] Xin Luna Dong and Felix Naumann. "Data fusion: resolving data conflicts for integration". In: *Proceedings of the VLDB Endowment* 2.2 (2009), pp. 1654–1655.

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

[14] 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.

[15] Ahmet M. Elbir, Burak Soner, and Sinem Coleri. *Federated Learning in Vehicular Networks*. 2020. DOI: 10.48550/ARXIV.2006.01412. URL: https://arxiv.org/abs/2006.01412.

[16] 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.

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

[18] Jing Gao et al. "Mining Reliable Information from Passively and Actively Crowdsourced Data". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, 2121â2122. ISBN: 9781450342322. DOI: 10.1145/2939672.2945389. URL: https://doi.org/10.1145/2939672.2945389.

[19] Nicholaus J. Gati et al. "Differentially private data fusion and deep learning Framework for CyberâPhysicalâSocial Systems: State-of-the-art and perspectives". In: *Information Fusion* 76.April (2021), pp. 298–314. ISSN: 15662535. DOI: 10.1016/j.inffus.2021.04.017. URL: https://doi.org/10.1016/j.inffus.2021.04.017.

[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] Adam Groce, Peter Rindal, and Mike Rosulek. "Cheaper Private Set Intersection via Differentially Private Leakage". In: *Proceedings on Privacy Enhancing Technologies* 2019.3 (2019), pp. 6–25. DOI: 10.2478/popets-2019-0034.

[23] Oktie Hassanzadeh et al. "Framework for Evaluating Clustering Algorithms in Duplicate Detection". In: *Proc. VLDB Endow.* 2.1 (Aug. 2009), 1282â1293. ISSN: 2150-8097. DOI: 10.14778/1687627.1687771. URL: https://doi.org/10.14778/1687627.1687771.

[24] 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.

[25] 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.

[26] 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.

[27] Basit Khurram and Florian Kerschbaum. "SFour: A protocol for cryptographically secure record linkage at scale". In: *Proceedings - International Conference on Data Engineering* 2020-April (2020), pp. 277–288. ISSN: 10844627. DOI: 10.1109/ICDE48307.2020.00031.

[28] Jakub KoneÄnÃ½ 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.

[29] Mehmet Kuzu et al. "Efficient privacy-aware record integration". In: *ACM International Conference Proceeding Series* (2013), pp. 167–178. DOI: 10.1145/2452376.2452398.

[30] 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, 191â202. ISBN: 089791001X. DOI: 10.1145/582095.582127. URL: https://doi.org/10.1145/582095.582127.

[31] Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. *Compositional Vector Space Models for Knowledge Base Completion*. 2015. DOI: 10.48550/ARXIV.1504.06662. URL: https://arxiv.org/abs/1504.06662.

[32] Thiago Nóbrega, Carlos Eduardo S. Pires, and Dimas Cassimiro Nascimento. "Blockchain-based Privacy-Preserving Record Linkage: enhancing data privacy in an untrusted environment". In: *Information Systems* 102 (2021), p. 101826. ISSN: 03064379. DOI: 10.1016/j.is.2021.101826. URL: https://doi.org/10.1016/j.is.2021.101826.

[33] 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.

[34] Jeff Pasternack and Dan Roth. "Knowing What to Believe (When You Already Know Something)". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, 2010, 877â885.

[35] 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.

[36] 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.

[37] Theodoros Rekatsinas et al. "SLiMFast: Guaranteed Results for Data Fusion and Source Reliability". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. Chicago, Illinois, USA: Association for Computing Machinery, 2017, 1399â1414. ISBN: 9781450341974. DOI: 10.1145/3035918.3035951. URL: https://doi.org/10.1145/3035918.3035951.

[38] 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.

[39] Diego Rodrigues and Altigran da Silva. "A study on machine learning techniques for the schema matching network problem". In: *Journal of the Brazilian Computer Society* 27.1 (2021). ISSN: 16784804. DOI: 10.1186/s13173-021-00119-5.

[40] Khalid Saleem. "Schema Matching and Integration in Large Scale Scenarios". In: (2009).

[41] Bohdan Shubyn et al. "Federated Learning forÂ Anomaly Detection inÂ Industrial IoT-enabled Production Environment Supported byÂ Autonomous Guided Vehicles". In: *Computational Science – ICCS 2022*. Ed. by Derek Groen et al. Cham: Springer International Publishing, 2022, pp. 409–421. ISBN: 978-3-031-08760-8.

[42] Marcin Szymczak et al. "Content Data Based Schema Matching". In: *Challenging Problems and Solutions in Intelligent Systems*. Ed. by Guy de Trė et al. Cham: Springer International Publishing, 2016, pp. 281–322. ISBN: 978-3-319-30165-5. DOI: 10.1007/978-3-319-30165-5_14. URL: https://doi.org/10.1007/978-3-319-30165-5_14.

[43] Dinusha Vatsalan, Peter Christen, and Erhard Rahm. "Incremental clustering techniques for multi-party Privacy-Preserving Record Linkage". In: *Data and Knowledge Engineering* 128.November 2019 (2020), p. 101809. ISSN: 0169023X. DOI: 10.1016/j.datak.2020.101809. arXiv: 1911.12930. URL: https://doi.org/10.1016/j.datak.2020.101809.

[44] Dinusha Vatsalan, Peter Christen, and Erhard Rahm. "Scalable Privacy-Preserving Linking of Multiple Databases Using Counting Bloom Filters". In: *IEEE International Conference on Data Mining Workshops, ICDMW* 0 (2016), pp. 882–889. ISSN: 23759259. DOI: 10.1109/ICDMW.2016.0130.

[45] Dinusha Vatsalan et al. "Privacy-preserving record linkage for big data: Current approaches and research challenges". In: *Handbook of Big Data Technologies* (2017), pp. 851–895. DOI: 10.1007/978-3-319-49340-4_25.

[46] 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.

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

**European Joint Doctorate in**
**Data Engineering for Data Science (DEDS)**
**Doctorate Project Plan[2]**
**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:

[2]Choose the appropriate heading among the three

The Chair of the Candidate Progress Committee

Professor ....................................

Date: