Offer #2024-07810

PhD Position F/M Doctorant F/H: Private synthetic data generation with diversity constraints for healthcare

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria University of Lille centre, created in 2008, employs 360 people including 305 scientists in 15 research teams. Recognised for its strong involvement in the socio-economic development of the Hauts-De-France region, the Inria University of Lille centre pursues a close relationship with large companies and SMEs. By promoting synergies between researchers and industrialists, Inria participates in the transfer of skills and expertise in digital technologies and provides access to the best European and international research for the benefit of innovation and companies, particularly in the region.For more than 10 years, the Inria University of Lille centre has been located at the heart of Lille's university and scientific ecosystem, as well as at the heart of Frenchtech, with a technology showroom based on Avenue de Bretagne in Lille, on the EuraTechnologies site of economic excellence dedicated to information and communication technologies (ICT)

Context

This PhD position will be supported by the CAPS'UL project. The position will be based in the MAGNET team in Lille.

The CAPS'UL project objective is to promote digital health culture for current and future healthcare professionals. Part of the project concerns the design of a high-performance tool for practical situations, enabling concrete and effective collaboration between the various training, socio-economic and medico-social actors in the implementation of training courses. It will provide a credible immersive environment (real software and simulated healthcare data) and teaching scenarios for the entire teaching community.

The INRIA MAGNET team (and hence the recruited collaborators) will contribute to this project by researching machine learning algorithms for synthetic data generation with privacy constraints and dedicated to training.

Assignment

Synthetic data generation is an approach to publishing (high-dimensional) data with the hope to respect some privacy principles. A major expected advantage over traditional privacy-preservation methods is to produce data with high utility for general-purpose applications, rather than protecting results to a given, specific set of queries. This topic has recently been receiving a lot of attention in the Machine Learning and trustworthiness community [7], as well as in the health domain [1].

In this PhD, we will focus on designing privacy-preserving algorithms for synthetic data generation in the context of training applications dedicated to students in the healthcare domain. The specificities of this target context are numerous.
Notably, data can have specific structure, such as being multimodal and/or longitudinal. Data generation should therefore preserve some notions of coherence, in terms of correlations, reflected by domain knowledge or extracted from training data.
Furthermore, health data is often very diverse, with some (possibly very) rare cases being of particular interest for trainees, that are expected to be exposed to a rich variety of situations and patients. The ability to produce synthetic datasets that satisfy some diversity constraints is therefore a major objective. A key issue is then to resolve the tension between diversity and privacy, caused by the fact that some examples making up for the desirable data diversity could be associated with rare patterns, thus bein more easily-identifiable to attacks on data privacy.

The general objective of this PhD is therefore to design algorithms to build generative models for synthetic data under privacy, diversity and coherence constraints. We will follow a formal approach and apply differential privacy principles [3] to get well-founded privacy guarantees. Some studies have investigated unfairness issues in data generation equalities [2]. The combination with differential privacy has notably been studied in [8], that could be the basis for an adaptation to the context described in this project.
A natural direction for diversity constraints is to consider (structured) determinantal point processes in data generation [6]. In order to handle coherence constraints, one direction can be the definition of logical system, given by a priori domain knowledge or graph dependencies similar to Bayesian Networks [9].

Finally, to complement formal privacy guarantees of synthetic generation algorithms, we may use some empirical evaluation tools, some of which have been compiled into ready-for-use code libraries such as TAPAS [5] or Anonymeter [4]. Whereas a number of state-of-the-art attacks rely on worst-case-scenario assumptions about the amount of information at the disposition of the attacker, a research direction will be to propose more realistic settings, informed by our target downstream applications. Selection, adaptation or enhancement of existing attacks to these settings may then provide with additional criteria to evaluate the readiness of our designed algorithms for real-world application.

References

[1] Anmol Arora. “Synthetic data: the future of open-access health-care datasets?” English. In: The Lancet 401.10381 (Mar. 2023). Publisher: Elsevier, p. 997.

[2] Karan Bhanot et al. “The Problem of Fairness in Synthetic Healthcare Data.” In: Entropy 23.9 (Sept. 2021), p. 1165.

[3] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy.” en. In: Foundations and Trends® in Theoretical Computer Science 9.3-4 (2013), pp. 211–407.

[4] Matteo Giomi et al. “A Unified Framework for Quantifying Privacy Risk in Synthetic Data.” In: (2022). Publisher: arXiv Version Number: 1.

[5] Florimond Houssiau et al. TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. arXiv:2211.06550 [cs]. Nov. 2022.

[6]Alex Kulesza and Ben Taskar. “Structured Determinantal Point Processes.” In: Advances in Neural Information Processing Systems. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc., 2010.

[7] Yingzhou Lu et al. Machine Learning for Synthetic Data Generation: A Review. en. Feb. 2023.

[8] David Pujol, Amir Gilad, and Ashwin Machanavajjhala. “PreFair: Privately Generating Justifiably Fair Synthetic Data.” In: Proc. VLDB Endow. 16 (2023), pp. 1573–1586.

[9] Jun Zhang et al. “PrivBayes: Private Data Release via Bayesian Networks.” In: ACM Transactions on Database Systems 42.4 (Oct. 2017), 25:1–25:41.

Main activities

Research

Skills

The applicant is expected to have studied machine learning, statistics and/or optimization, and to have good mathematical skills. Some broad interest for the topic of trustworthy AI and healthcare information systems is a plus.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

1st and 2nd year : 2100 € (gross monthly salarye)

3rd year : 2190 € (gross monthly salary)

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Statistics (Big data) (BAP E)
- **Town/city :** Villeneuve d'Ascq
- **Inria Center :** Centre Inria de l'Université de Lille
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-07-10

Contacts

- **Inria Team :** MAGNET
- **PhD Supervisor :**
  Tommasi Marc / Marc.Tommasi@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.