Privacy preserving techniques for data science

Judith Sáinz-Pardo Díaz
Director: Dr. Álvaro López García
Instituto de Física de Cantabria (IFCA), CSIC-UC

Summary

Data-driven technologies have undergone rapid growth over the last years, especially thanks to the availability of large volumes of data available to be processed and analyzed (i.e. "**big data**"). **Machine learning**, **artificial intelligence** and **deep learning** empower a wide range of applications (like artificial vision, natural language processing, speech recognition, anomaly detection, etc.).
Because of this, the study and further development of different **privacy techniques in a data science and data analysis context are a key area for scientific research.**

Main objectives

Throughout this thesis proposal we will explore different techniques for the analysis, transmission and **secure management of data**, with special focus on **data privacy** (through anonymity and differential privacy techniques among others), decentralized machine learning, deep learning applications, etc.
The main objectives of this project include the research and improvement of different techniques for analyzing, processing, publishing and sharing sensitive data while maintaining their privacy, as well as the study of new paradigms in the field of **data security and anonymization.**

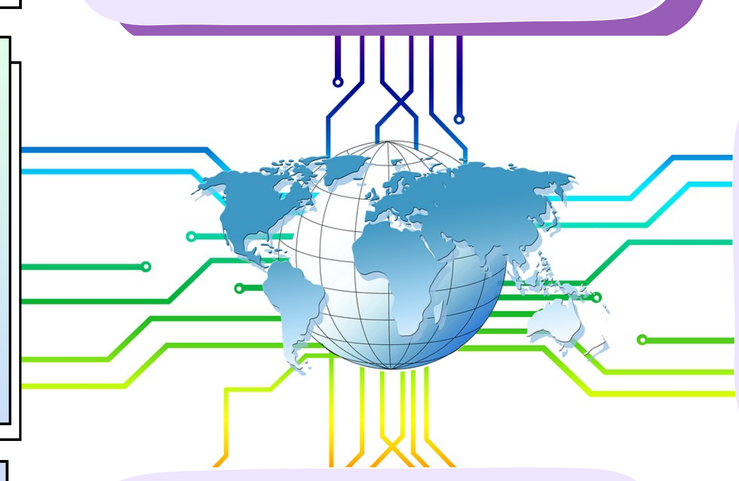
Deployment of the research

- **PART 1:** Secure access to data. Classical **anonymity** and **pseudonymity** tools and **differential privacy**, both local and global. Goal: facilitate the scientific community to process, analyze, share and publish data with security guarantees.
- **PART 2: Encryption schemes** and **Homomorphic Encryption** techniques. Incorporation into a data analysis pipelene. Goal: allow applying machine/deep learning models with sensitive data in secure environments.
- **PART 3:** Privacy preserving techniques for training and inference in **artificial intelligence models**. **Distributed learning** architectures, such as **Federated Learning** among others. Goal: deployment and application of such arquitectures in different real application cases.

Data science and big data

Data science: methods and techniques used for data processing and analysis.

Big data: extremely high-volume datasets that require complex computational techniques to extract information, trends and patterns.



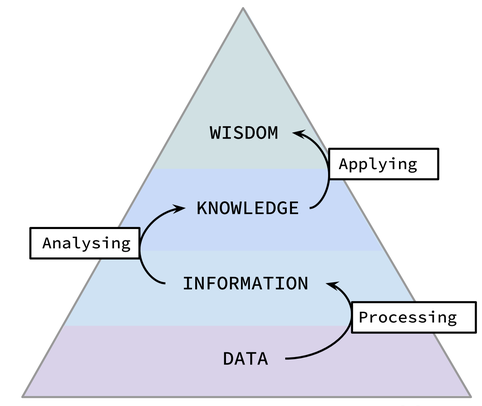


Figure 1. Knowledge pyramid. Adapted from [1].

Machine and deep learning

Machine learning is a branch of **artificial intelligence** that studies how to endow machines with learning capabilities.
Deep learning is a type of machine learning based on the use of artificial neural network architectures in order to extract higher-level characteristics from the data.

Data privacy

When handling data that includes **sensitive information**, it is important to focus on data privacy to **avoid potential security breaches**.
Different regulations exist in this context (e.g. **GDPR**) in order to protect users.

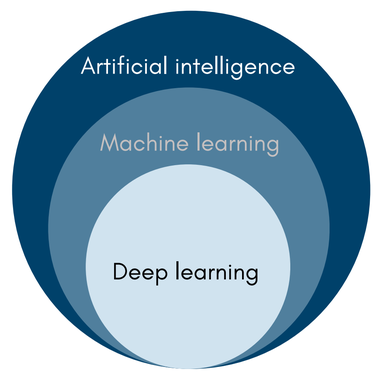


Figure 2. Venn diagram: artificial intelligence, machine learning and deep learning.

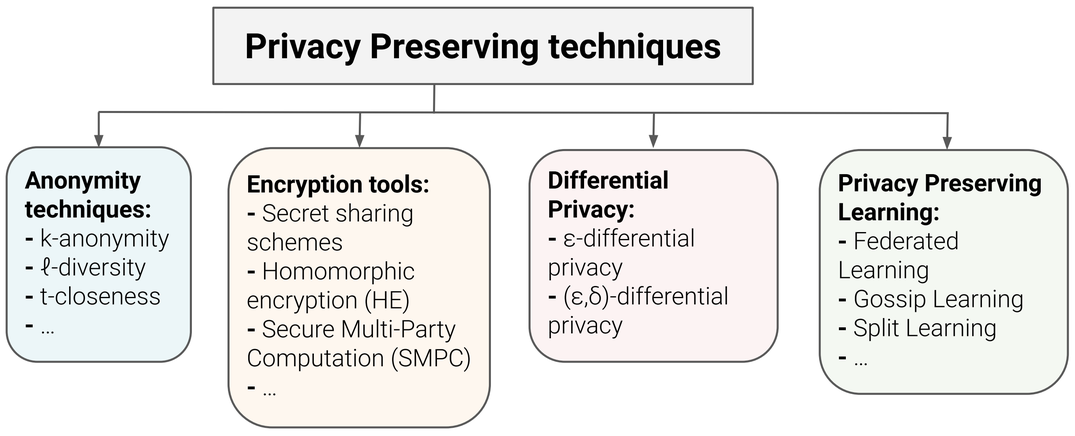


Figure 3. Privacy Preserving techniques summary.

Results: data anonymity

Implementation of *pyCANON*, a Python library and CLI that can be used to know the level of anonymity of a dataset (and thus **publish or share it while being aware of the risks involved**). Nine different techniques will be used for this purpose.
This tool is **open source**, and no prior knowledge of Python or anonymity techniques is required for the user.

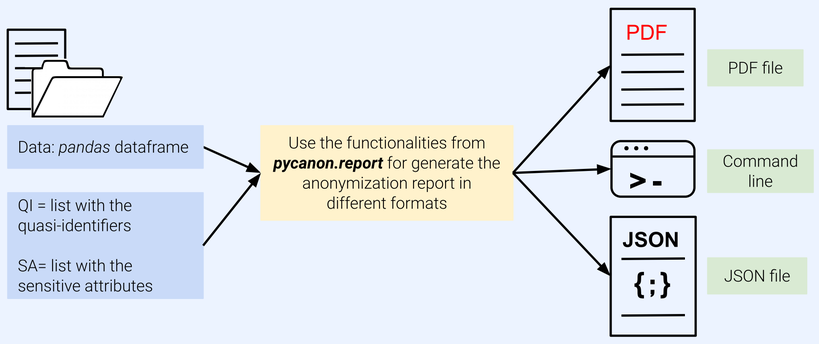


Figure 4. Code example, use of pyCANON report packege for cheking the level of anonymity of a dataset. [2]

Technique	Principal attack which prevents						
	Linkage	Re-identif cation	Homogeneity	Background	Skewness	Similarity	Inference
k-anonymity	✓	✓					
(α,k)-anonymity	✓	✓	✓				
ℓ-diversity			✓	✓			
Entropy ℓ-diversity			✓	✓			
Recursive (c,ℓ)-diversity			✓	✓			
t-closeness					✓	✓	
Basic β-likeness					✓		
Enhanced β-likeness					✓		
δ-disclosure privacy					✓		✓

Table 1. Anonymity tools implemented and attacks on the databases which mainly prevent. [2]

Publication: *Sáinz-Pardo Díaz, Judith, and Álvaro López García. "A Python library to check the level of anonymity of a dataset." Scientific Data 9.1 (2022): 785.*

Acknowledgements:
I would like to thank the funding through the European Commission - NextGenerationEU (Regulation EU 2020/2094) through CSIC's Global Health Platform (PTI+ Salud Global) and the support from the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud" that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101058593.

Results: privacy preserving machine learning

FEDERATED LEARNING: The concept of Federated Learning first emerged in 2017, introduced by McMahan [3]. It can be defined as a paradigm of **distributed machine learning** which allows collaboration between different data owners in order to **train ML models without the need to centralize or share the raw data.**

- Build data driven models exploiting distributed data without the need to centrally store it.
- Collaborative and decentralized approach to machine and deep learning
- Server-client architecture.
- Client data are never uploaded to the server.
- More complete models ensuring user privacy.

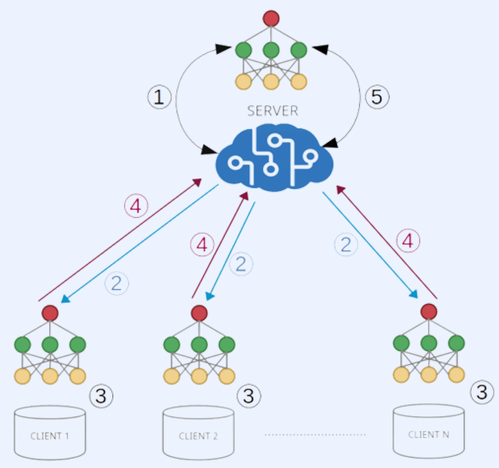


Figure 5. Schema of the federated learning architecture. [4]

Use case: Chest X-Ray image classification [4]

Objective: classify chest X-Ray images according to whether or not the patient has pneumonia.

Data: images of patients with pneumonia: 3875 for training and 390 for testing. Images of patients without pneumonia: 1341 (train) and y 234 (test). Obtained from [5].

Federated learning scenario: split the training set into 3 and 10 clients.

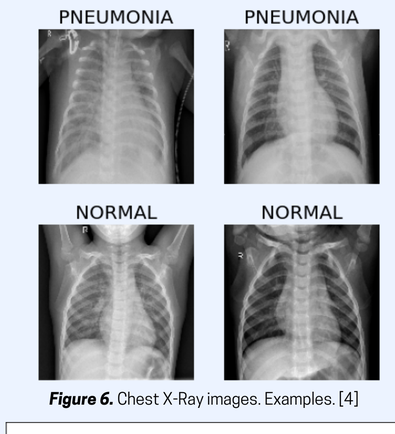


Figure 6. Chest X-Ray images. Examples. [4]

	Test Acc.	Exc. time (s)	Time reduction vs cent. apr. (%)
Centralized approach	0.6619	1386	—
Decentralized approach			
3 clients	0.8029	401	71.07
	0.7308	320.8	76.85
10 clients	0.7212	110	92.06
	0.7340	99	92.85

Table 2. Results for the loss, accuracy and execution time. Federated Learning varying the number of rounds and clients, and centralized approach. Extracted from [4].

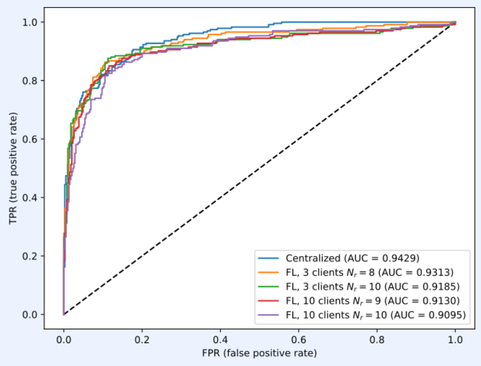


Figure 7. ROC curves in each scenario: centralized and federated learning. with different number of clients. [4]

Publication: *Sáinz-Pardo Díaz, Judith, and Álvaro López García. "Study of the performance and scalability of federated learning for medical imaging with intermittent clients." Neurocomputing 518 (2023): 142-154.*

Conclusions

This poster presents part of the work carried out during the first year of the PhD program, with special emphasis on the work that has resulted in **two publications in high impact journals**.
The initial goal of this thesis is to address the issue of **privacy protection in data science environments** along three complementary lines, as outlined when presenting the deployment of the research. The future work to be done is still very extensive, but will focus on these fundamental pillars for the achievement of **secure computing environments** in terms of **data processing, publishing, sharing and analysis**.

References

[1] Kitchin, Rob. The data revolution: Big data, open data, data infrastructures and their consequences. Sage, 2014.
[2] Sáinz-Pardo Díaz, Judith, and Álvaro López García. "A Python library to check the level of anonymity of a dataset." Scientific Data 9.1 (2022): 785.
[3] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.
[4] Sáinz-Pardo Díaz, Judith, and Álvaro López García. "Study of the performance and scalability of federated learning for medical imaging with intermittent clients." Neurocomputing 518 (2023): 142-154.
[5] Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." cell 172.5 (2018): 1122-1131.