AI and privacy: Safeguarding data in the age of artificial intelligence

Antoine Boutet – Insa-Lyon – Inria Privatics research group

EESN 2024

Data protection regulations

The **data protection** has become a crucial concern in today's digital age

The **GDPR** (General Data Protection Regulation) in 2018 is an example of data protection regulation introduced to cater to the evolving digital landscape

It governs how personal data is collected, stored, and processed and aims to **strengthen the rights of individuals** and enhance transparency in data processing

AI Promises

The **promises of AI** are great:
curing diseases, increasing productivity,
drinving car and flying drones,
composing music or writing poetry,
helping to solve climate change

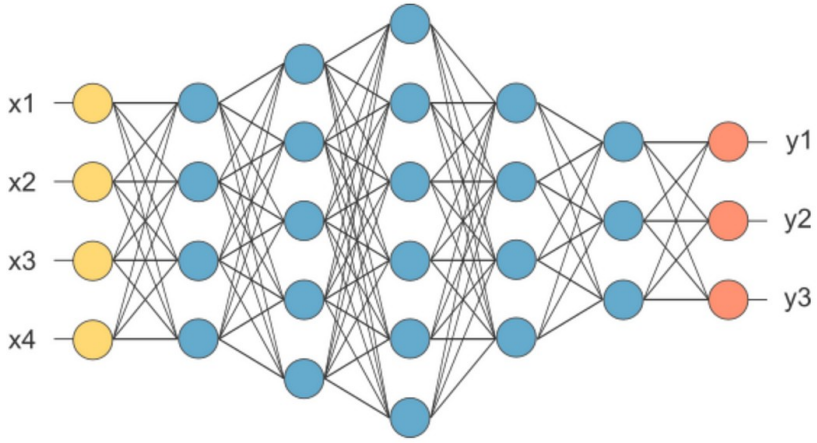
AI can do some things better than humans

AI has affected privacy rights

AI has significantly transformed the way we interact with data

from data

to learning models



The use of AI algorithms to process data has led to the emergence of **new privacy concerns**

AI comes with new risks

- Privacy
- Security
- Fairness
- Explainability

Challenge:
Address globally these risks

AI comes with new risks

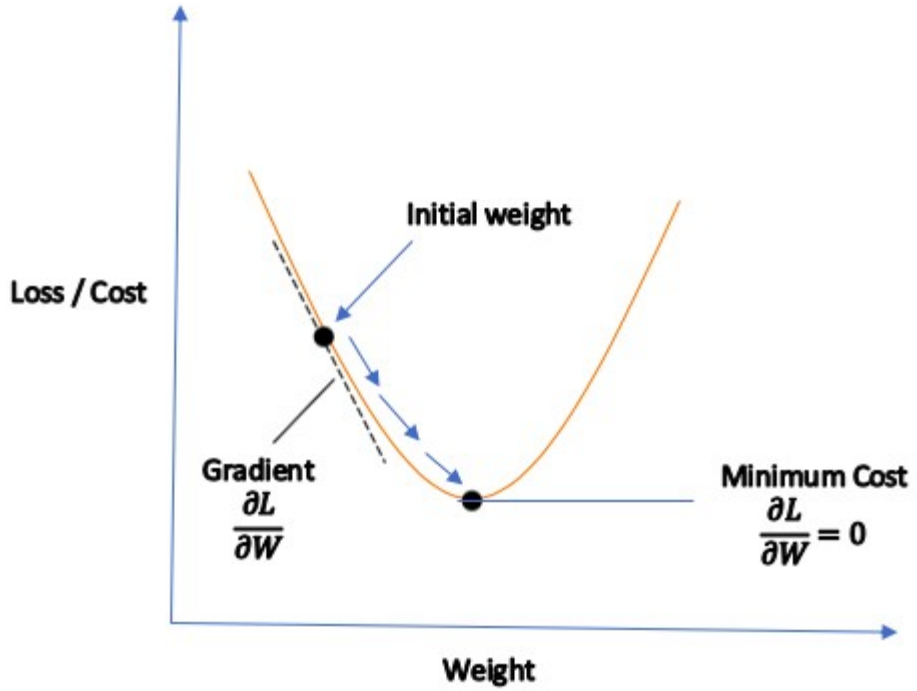
- Privacy
- Security
- Fairness
- Explainability

Challenge:
Address globally these risks

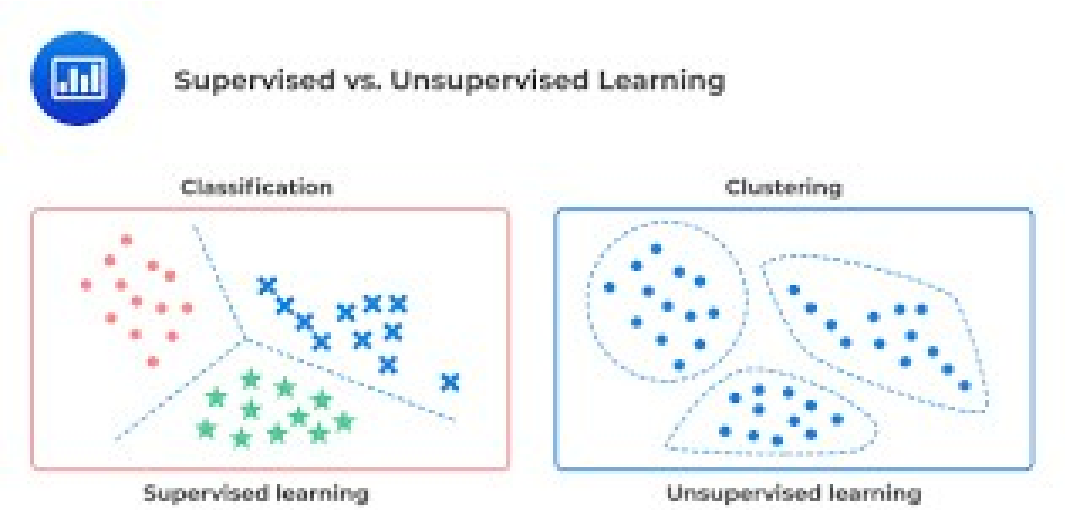
The ever-evolving landscape of AI

Taxonomy

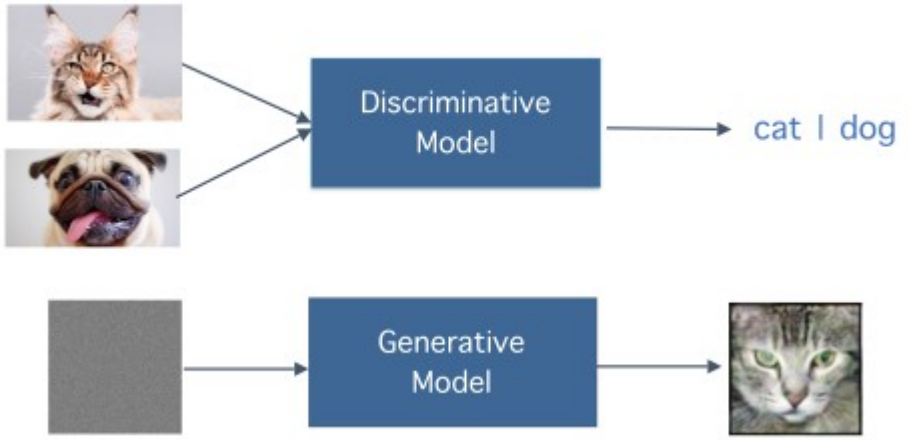
Learning process
- **Training (e.g., SGD)**
- **Validation**
- **Testing**



Learning tasks
- **Supervised**
- **Unsupervised**



Taxonomy

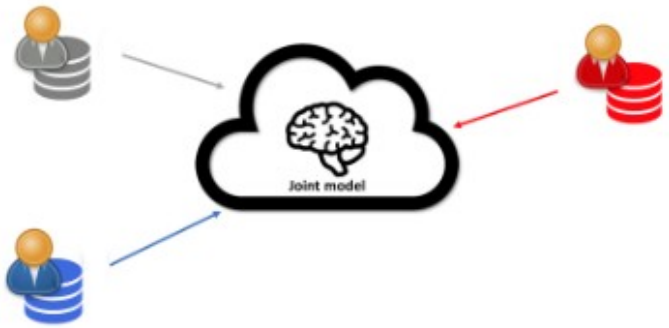


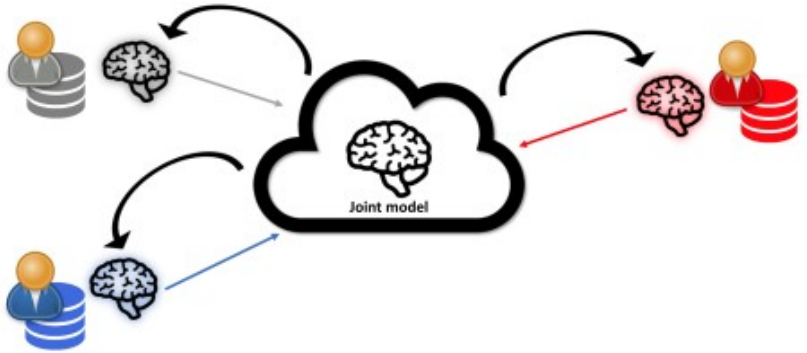
ML approaches

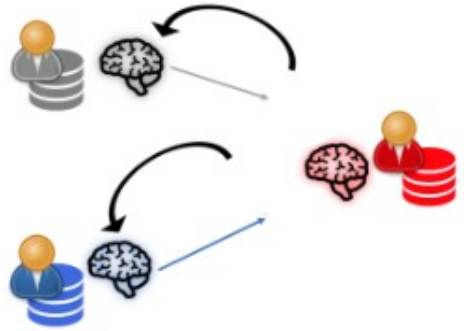
- **Discriminative models**
- **Generative models**

ML Learning

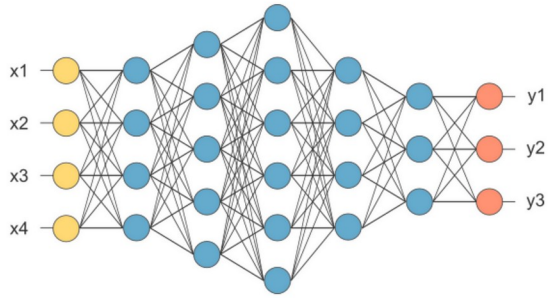
- **Centralized**
- **Collaborative / Federated Learning**
- **Fully Decentralized Learning**





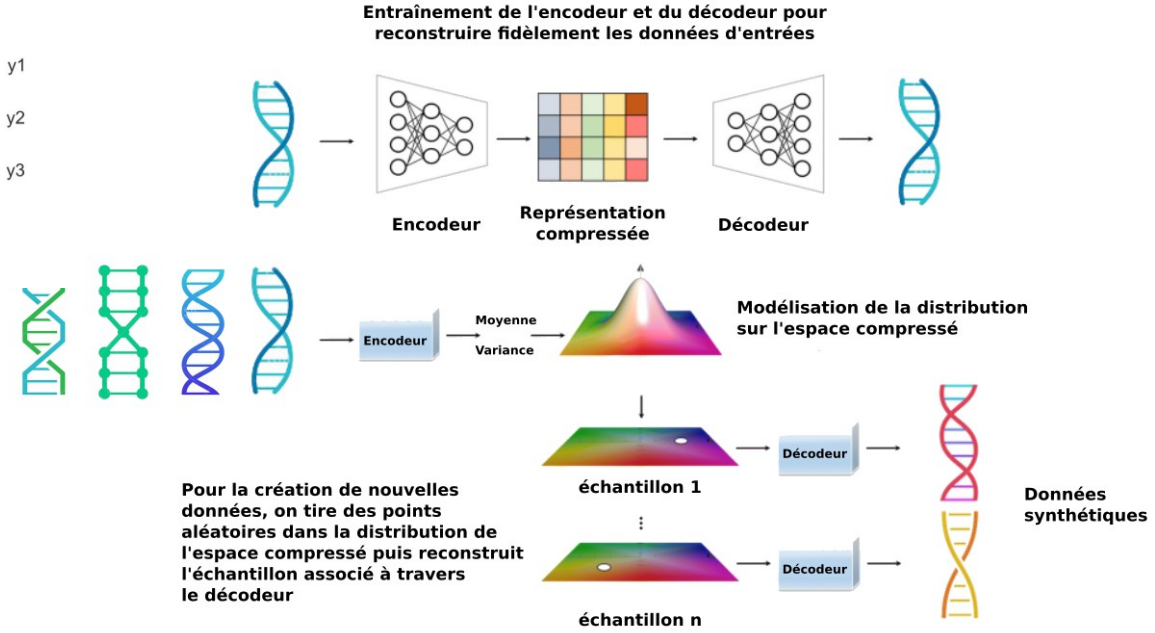


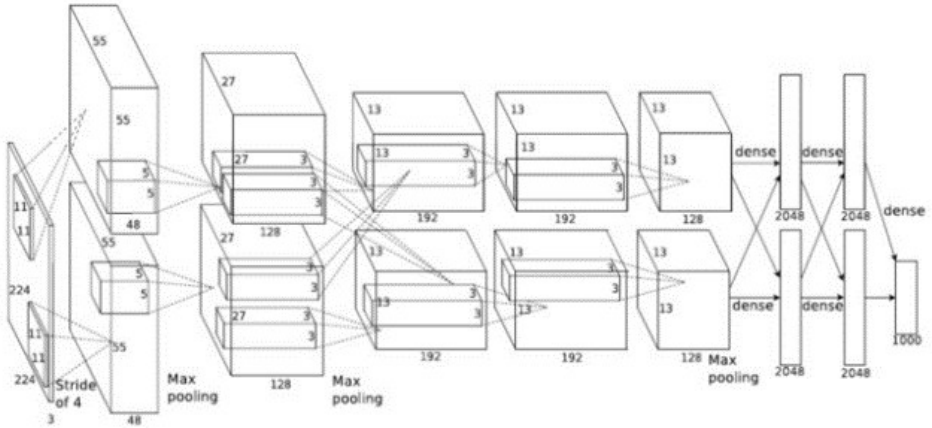
Taxonomy

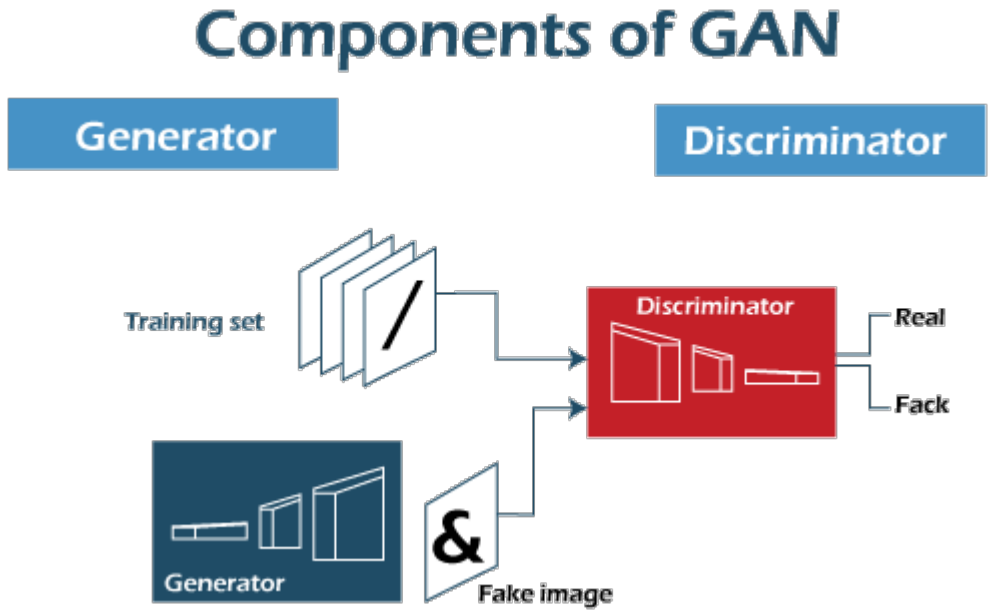


Model Architecture

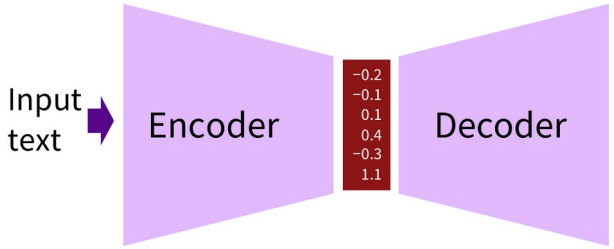
- Fully Connected
- Off-the-shelf Architectures (eg AlexNet)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Generative Adversarial Networks (GAN)
- Variational Auto Encoder (VAE)
- ...

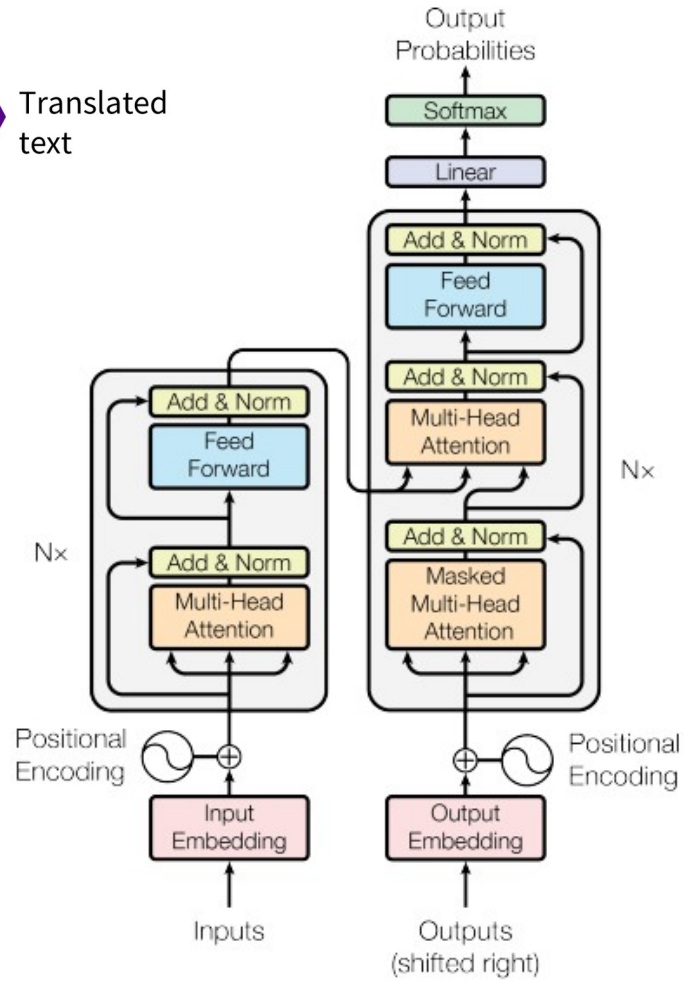






Taxonomy



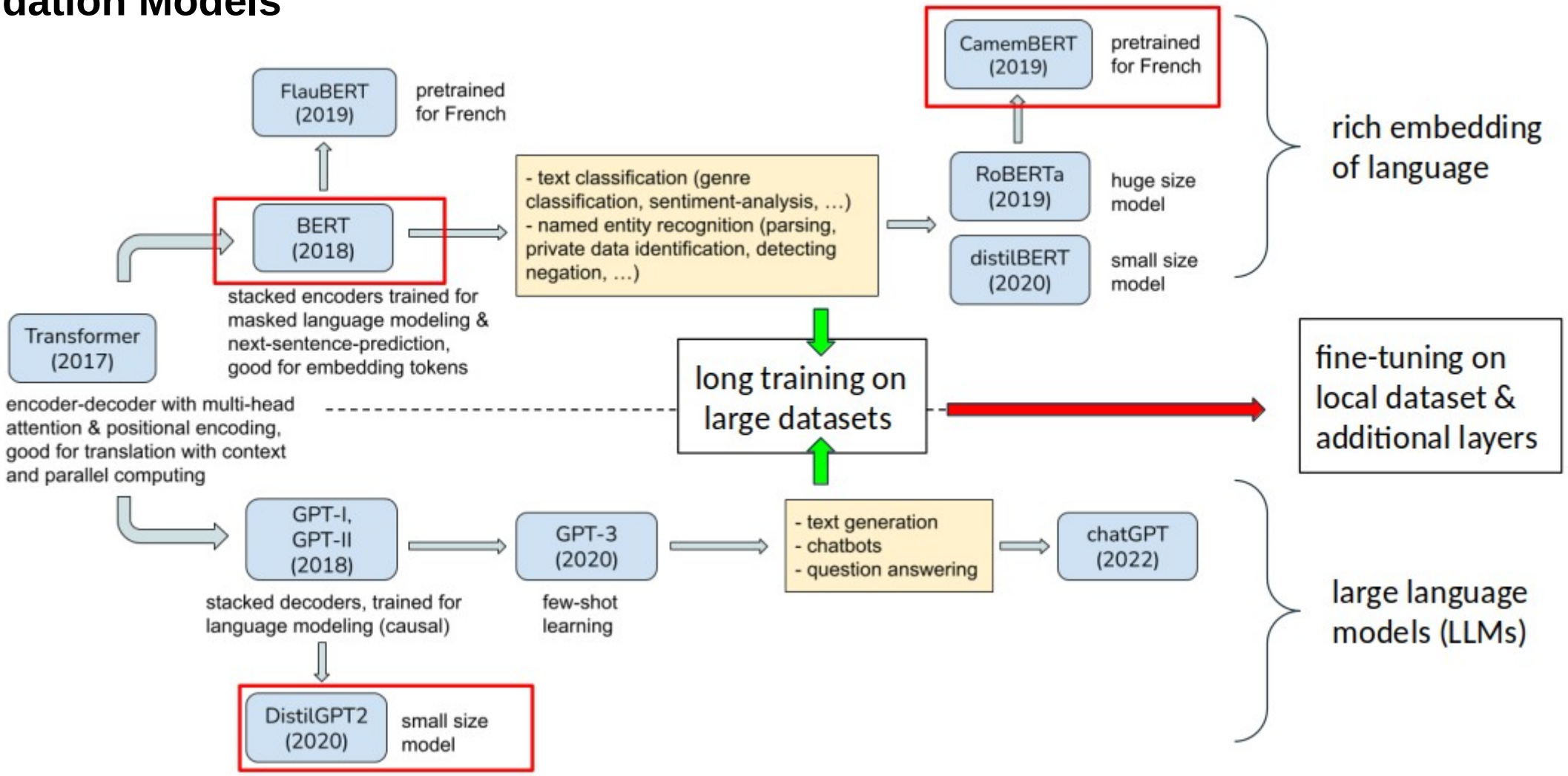


Models for NLP

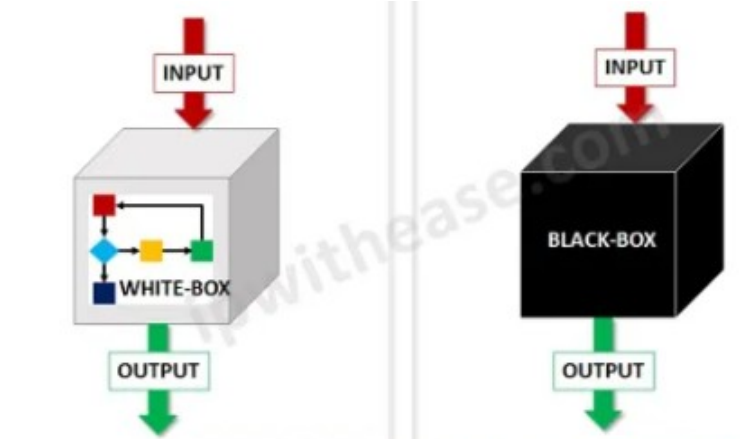
- **Transformer**
- **BERT: stacks on Transformer encoders**
  - **Text classification, entity recognition**
- **GPT: stacks on Transformer decoders**
  - **Question answering, chatbots**

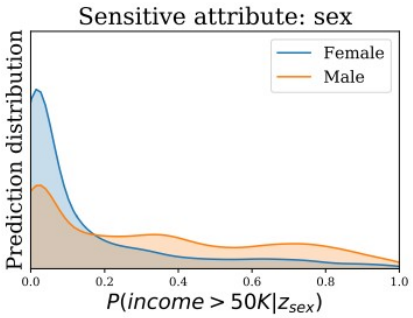
Taxonomy

Fondation Models



Taxonomy



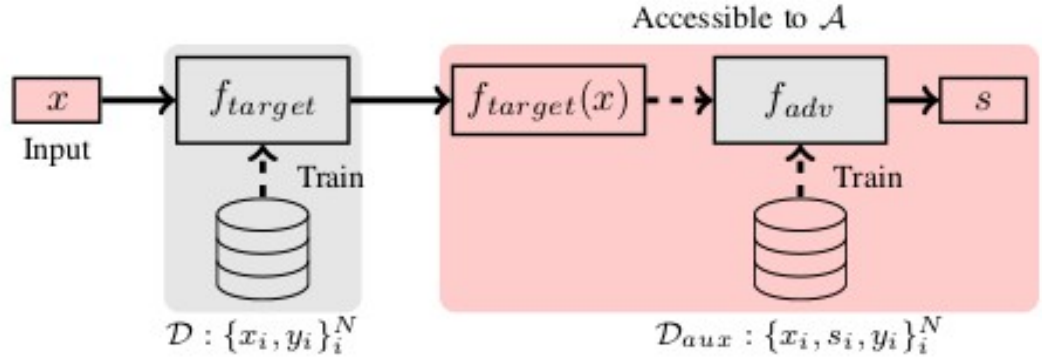


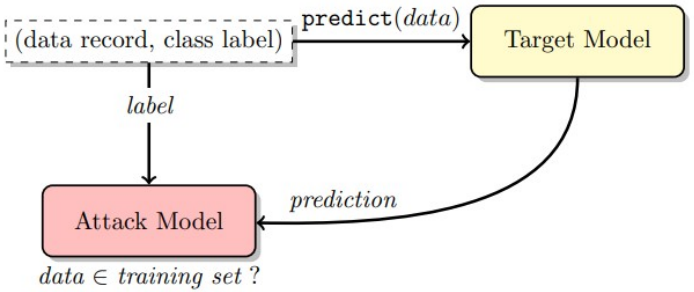
Adversary Models

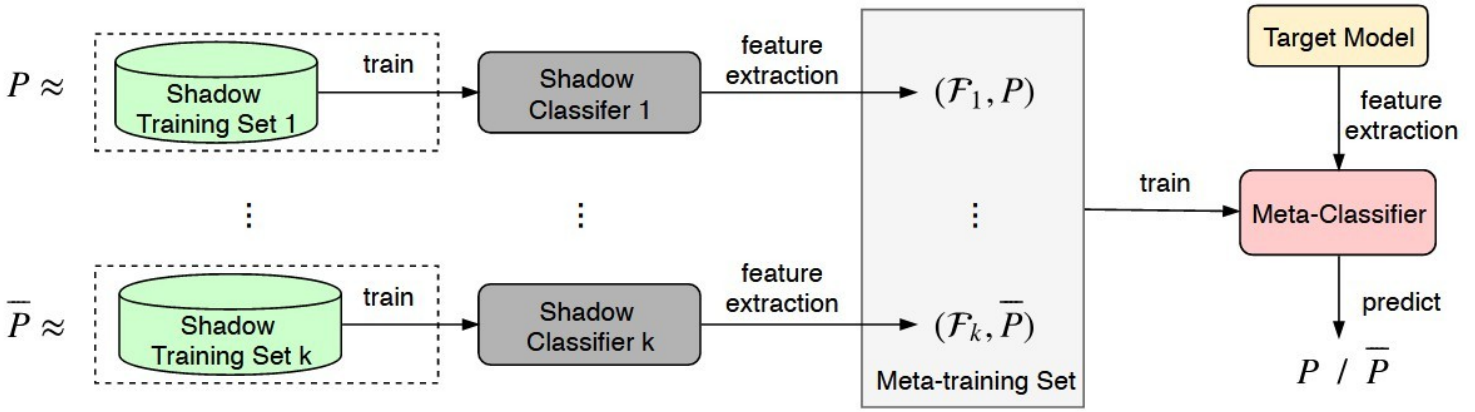
- White-Box vs Black-Box
- Passive vs Active attack

Type of attacks

- Attribute Inference Attack
- Membership Inference Attack (MIA)
- Poisonning Attack
- Backdoor Attack
- Stealing Attack
- ...







OWASP ML Security top 10

OWASP PROJECTS CHAPTERS EVENTS ABOUT

OWASP Machine Learning Security Top Ten

Main | Charter | Related | Glossary

owasp incubator License CC BY-SA 4.0

📌 ***Important Information***

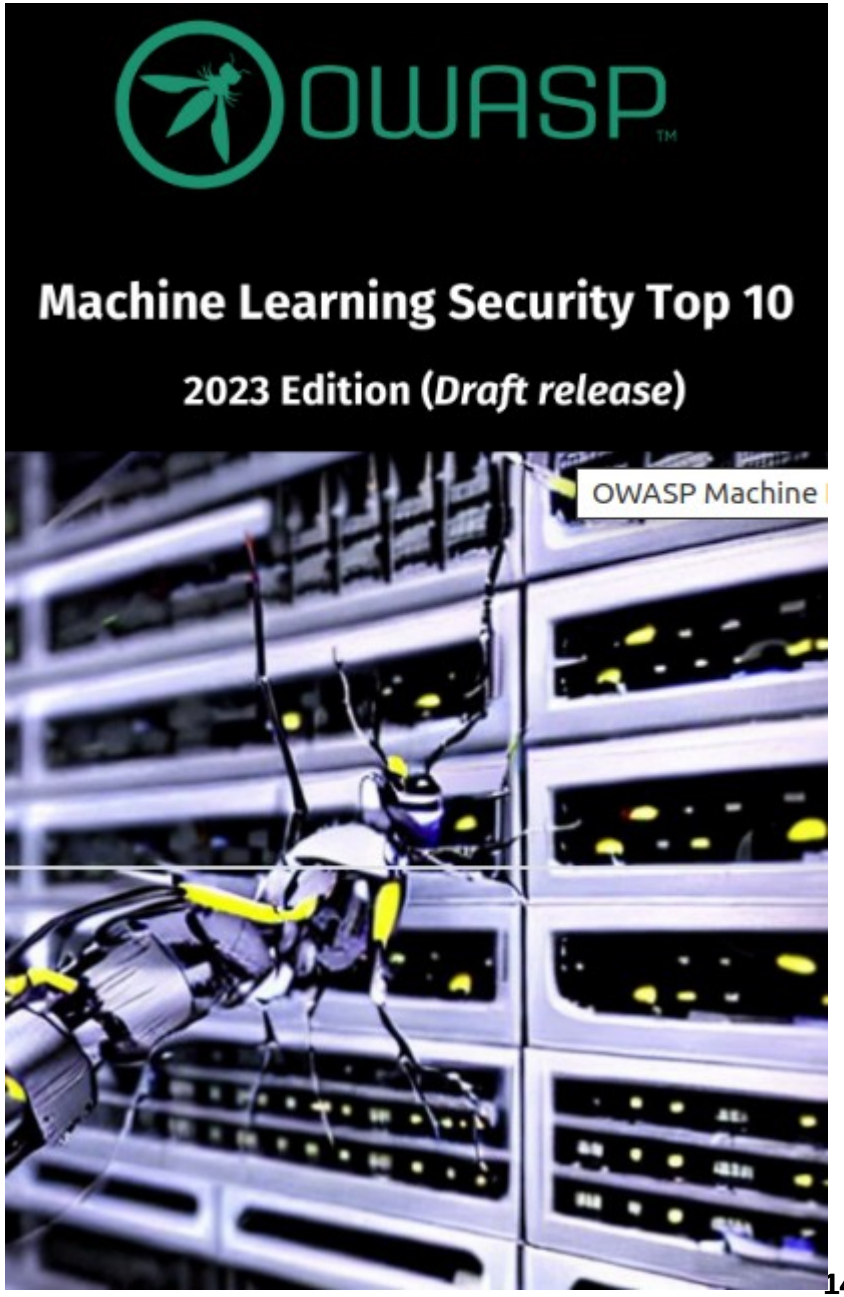
The current version of this work is in draft and is being modified frequently. Please refer to the project wiki for information on how to contribute and project release timelines.

Overview

Welcome to the repository for the OWASP Machine Learning Security Top 10 project! The primary aim of the OWASP Machine Learning Security Top 10 project is to deliver an overview of the top 10 security issues of machine learning systems. More information on the project scope and target audience is available in our project working group charter

Top 10 Machine Learning Security Risks

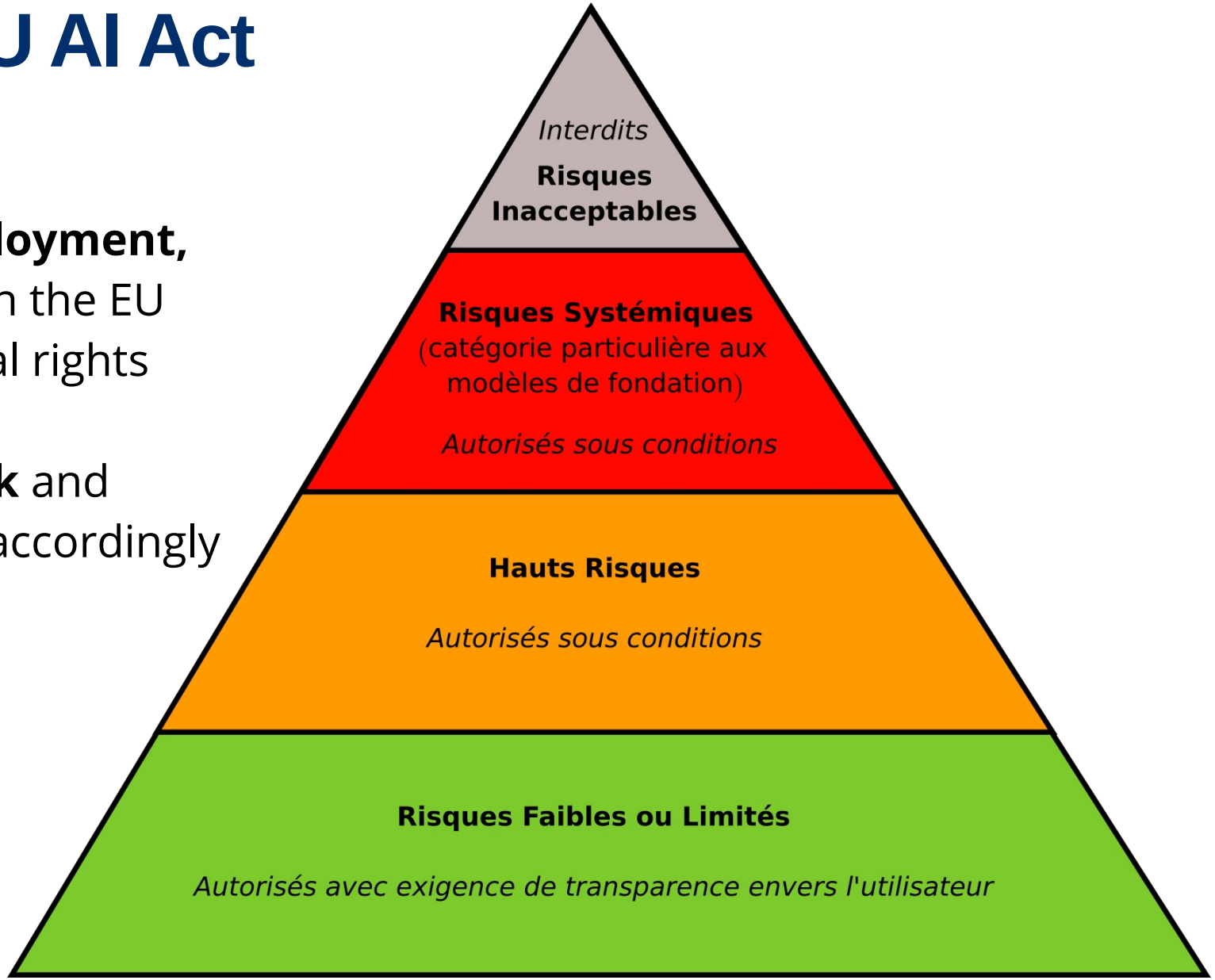
- ML01:2023 Input Manipulation Attack
- ML02:2023 Data Poisoning Attack
- ML03:2023 Model Inversion Attack
- ML04:2023 Membership Inference Attack
- ML05:2023 Model Theft
- ML06:2023 AI Supply Chain Attacks
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning



Regulation on AI: EU AI Act

Regulate the **development, deployment, and use** of artificial intelligence in the EU to ensure safety and fundamental rights

Categorize different **levels of risk** and prescribes regulatory measures accordingly



Risques Inacceptables

Risques Inacceptables

Sont interdits les systèmes les plus intrusifs et/ou usages pouvant avoir des conséquences néfastes démesurées sur les droits des individus ou la société (systèmes d'IA à des fins de manipulation, crédit social, ...).

Risques Systémiques

Risques systémiques

(catégorie particulière aux modèles de fondation)

Sont autorisés sous conditions les systèmes d'IA dit « de fondation » entraînés sur de grands ensembles de données et conçus pour produire des outputs pouvant être utilisées pour effectuer des tâches variées.

Pour ces systèmes : obligation de documentation et respect des règles en matière de copyright. Si capacités FLOPS > 10^{25} = considérés comme posant un « risque systémique » alors obligations renforcées incluant des obligations de notification d'incidents auprès de la Commission, de gestion et de documentation du risque, et des obligations de cybersécurité pour les modèles ainsi que les infrastructures soutenant les modèles.

Hauts Risques

Hauts risques

Sont autorisés sous conditions les systèmes susceptibles d'engendrer des conséquences néfastes importantes sur la santé, la sécurité et les droits fondamentaux des individus.

Pour ces systèmes : dispositif de gestion des risques, évaluation de conformité, documentation technique, registre des logs, exigences en matière de gouvernance des données, supervision humaine, et de cybersecurité.

Risques Faibles ou Limités

Risques faibles ou limités

Systèmes d'IA relevant originellement de la catégorie supérieure mais dont l'incidence sur la prise de décision et l'impact restent minimes, et systèmes d'IA qui, d'emblée, ne relèvent pas de la catégorie à « haut risque ».

Pour ces systèmes: obligation de transparence dès lors que le système est destiné à interagir avec des personnes physiques.

Obligations et Encouragements Transversals

Les obligations et encouragements : transversal

Devoir de transparence transversal concernant tout système destiné à interagir directement avec l'humain (information claire des usagers, outputs marqués électroniquement).

Encouragé pour tout fournisseur et déployeur de système d'IA : Adoption de codes de conduite construits sur la base d'indicateurs de performance clés tels que le respect des recommandations européennes en matière d'IA « digne de confiance », la minimisation de l'impact environnemental des systèmes d'IA, la promotion d'une « culture IA, » en interne…

Exclus du champs d’aplication du AI Act :

Exclus du champs d’application du AI Act :

Systèmes d’IA a des fins militaires, de défense, de sécurité nationale,
ou utilisés spécifiquement pour la recherche scientifique
ou open-source non mis sur le marché ni déployés.

Agenda

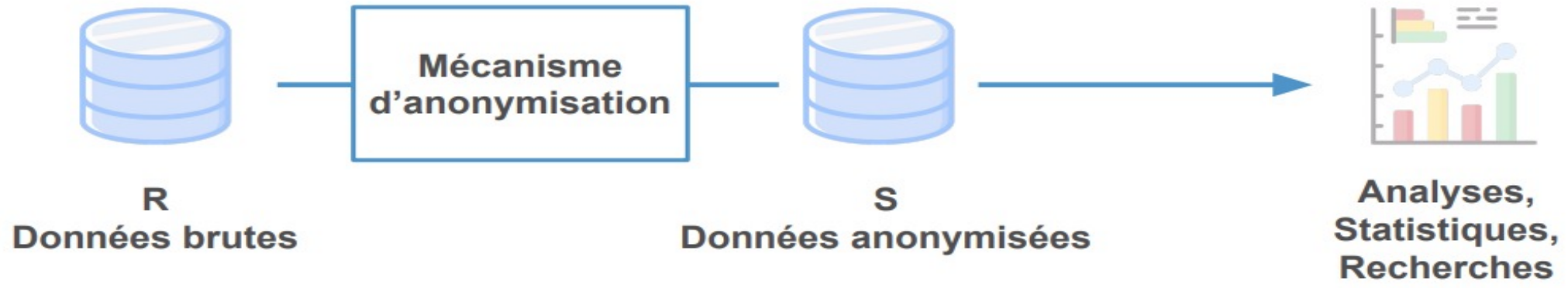
- **Generative models: data sharing via data anonymisation vs synthetic data**

- **Federated Learning (FL) → OT3 Sécurité et Vie Privée**
  - **FL using personalized and private layers**
  - **MixNN: Protection of FL Against Inference Attacks by Mixing Neural Network Layers**
  - **Quantifying the learning gains and the vulnerabilities to adversarial attacks**
  - **Organisation of challenges**

- **NLP Models: privacy risks and countermeasures → OT3 Sécurité et Vie Privée**

Data Sharing via Anonymisation

Anonymisation is imposed by the GDPR to keep the personal information confidential



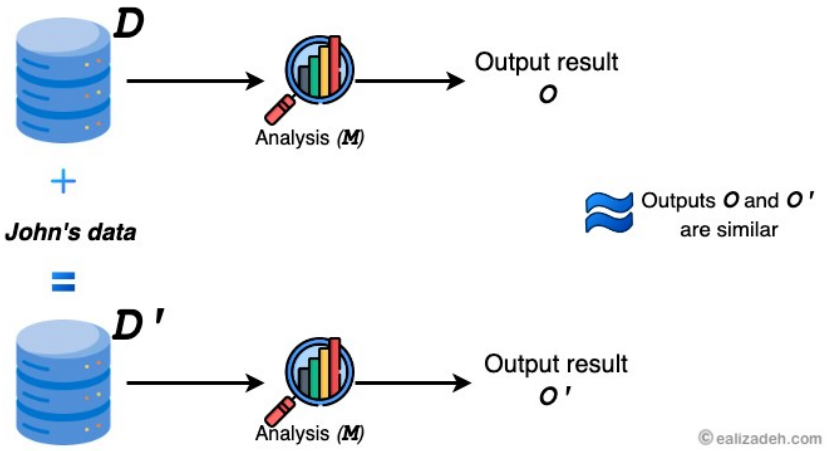
An anonymization solution must respect three creteria:

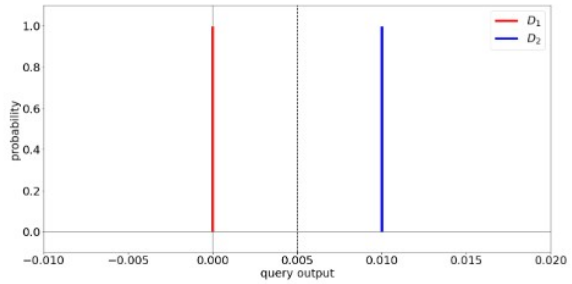
- **Individualization: is it always possible to distinguish an individual?**
- **Correlation: is it possible to link separate datasets about the same individual?**
- **Inference: can we infer information about an individual?**

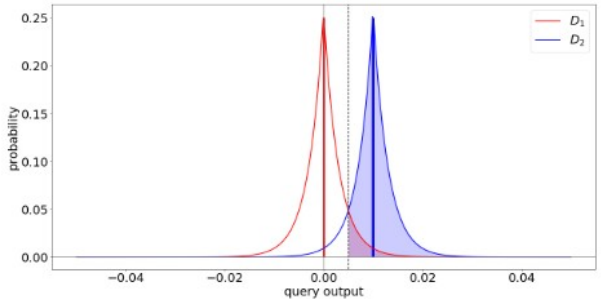
Data Sharing via Anonymisation

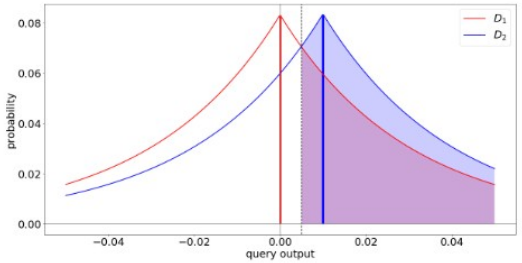
Anonymisation techniques:
k-anonymity, l-diversity, t-closeness, …

Differential Privacy



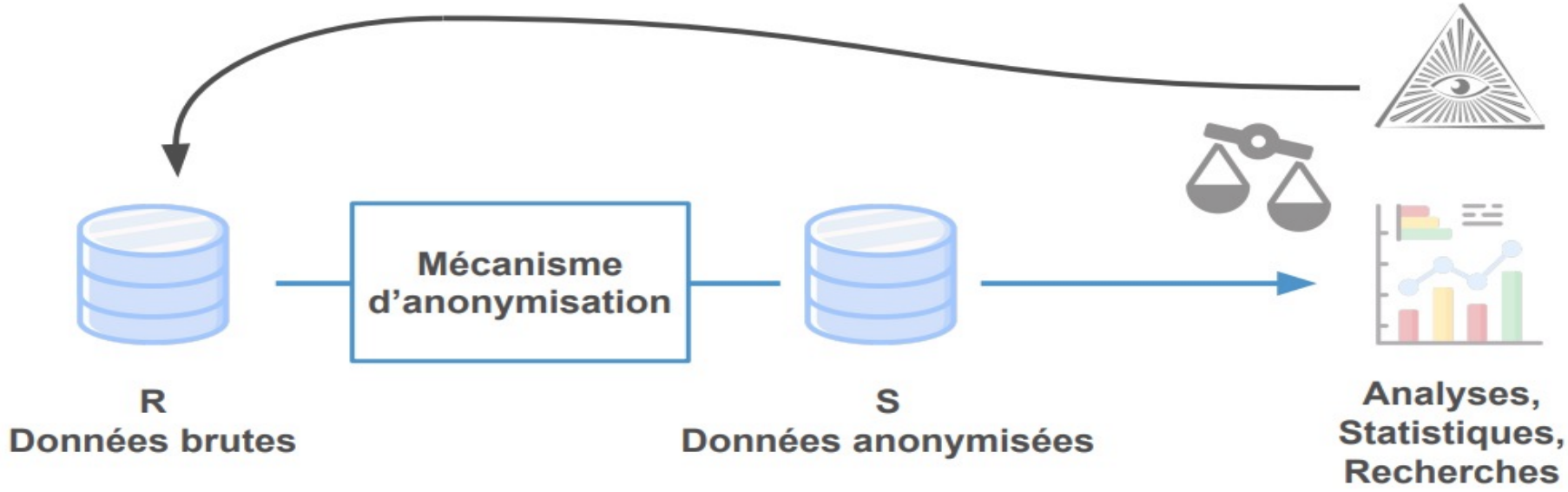






Data Sharing via Anonymisation

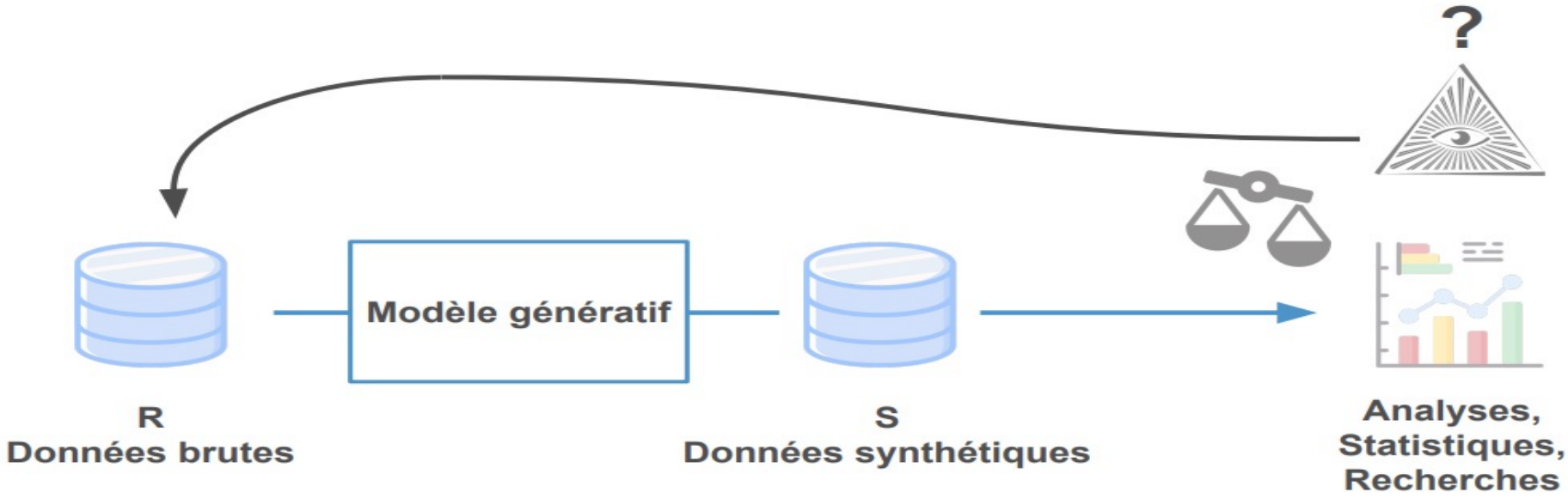
Anonymisation is a difficult task (there are no free food)



Limit: utility and privacy tradeoff difficult to calibrate

Still a risk of re-identification and inference + reduction of the information

Sharing via Synthetic data

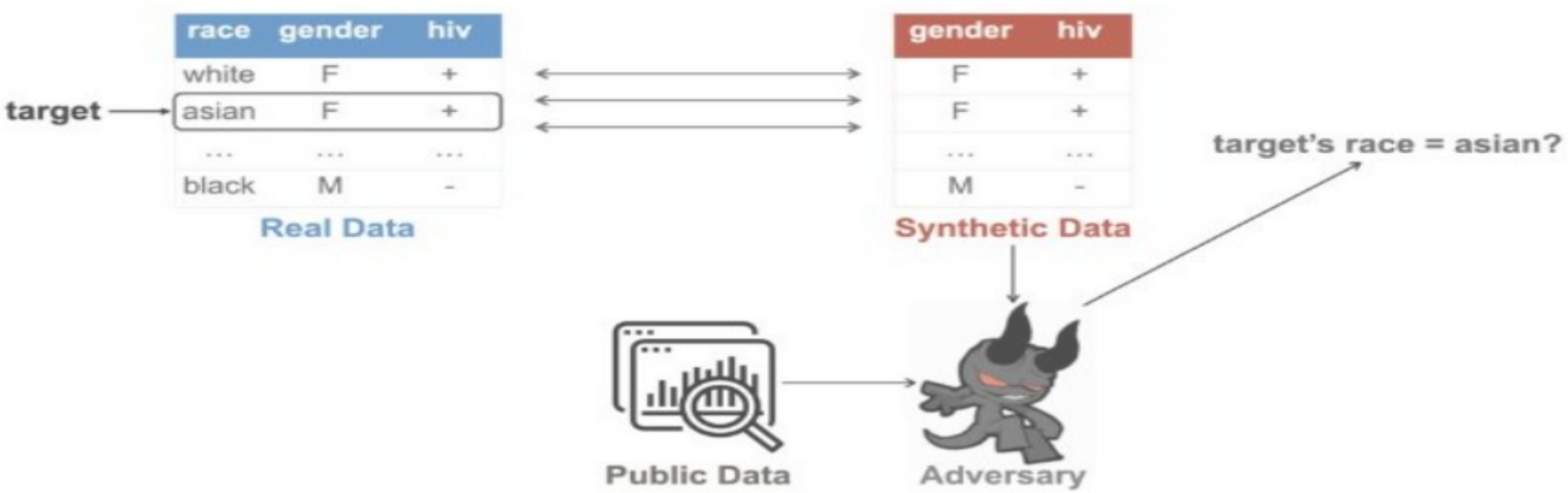


Evaluation of the confidentiality:

- **Linking**
- **Attribute inference**
- **Membership inference**

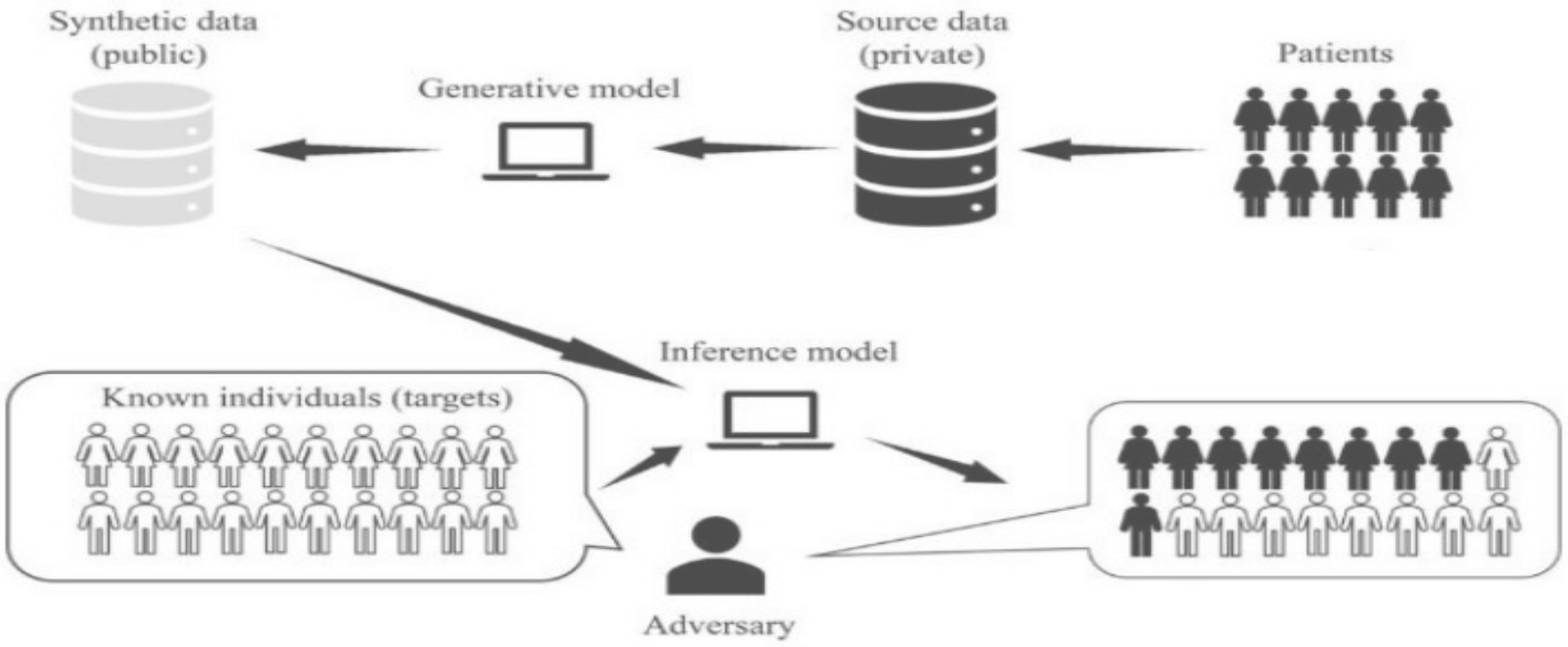
Sharing via Synthetic data

Attribute inference



Sharing via Synthetic data

Membership inference attack:

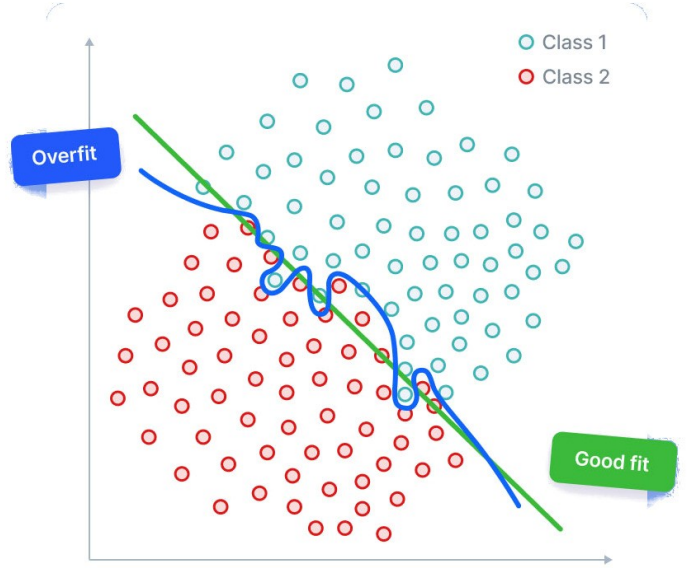


Sharing via Synthetic data

Membership inference attack: the inference model allows the adversary to distinguish a data point that has been already seen during the training (e.g., shadow models [1,2], Monte Carlo Attack [3], Data copying detection [4])

These attacks are based on the overfitting of the generative model

→ **the influence of a data point used during the training can be detected**



[1] Oprisanu et al. Measuring Utility and Privacy of Synthetic Genomic Data. ArXiv:2102.03314, 2021
[2] Stadler et al. Synthetic Data - A Privacy Mirage. ArXiv:2011.07018, 2020
[3] Hilprecht et al. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. PETS 2019.
[4] Meehan et al. A Non-Parametric Test to Detect Data-Copying in Generative Models. AISTATS 2020.
[5] Hayes et al. LOGAN: Membership Inference Attacks Against Generative Models. ArXiv: 1705.07663, 2019

It is not a binary risk

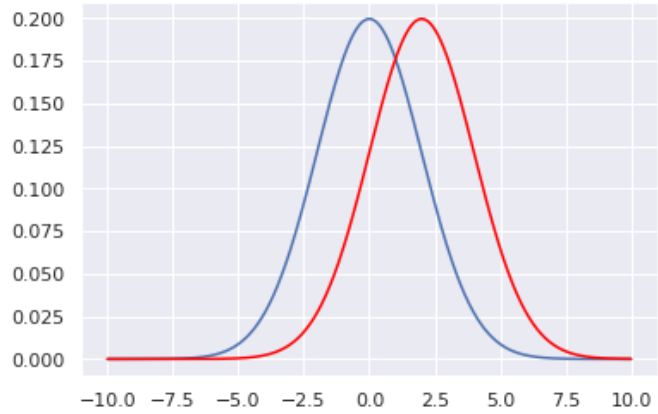
Measure a probability of risk

The overfitting is not similar on all the data points, outliers are more vulnerable

→ Taken into account / focus on outliers [1, 2]

[1] Stadler et al. Synthetic Data - Anonymisation Groundhog Day. ArXiv:2011.07018, 2020
[2] Carlini et al. Membership Inference Attacks From First Principles. ArXiv:2112.03570

Differential Privacy



Entrainement du modèle génératif avec de la DP [1, 2, 3]

→ Coût important en terme d'utilité / performance de la génération [4]

[1] Jordon et al. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. ArXiv:2011.07018, 2020

[2] Xie et al. Differentially Private Generative Adversarial Network.. ArXiv:1802.06739, 2018
[3] Li et al. DPSyn: Experiences in the NIST Differential Privacy Data Synthesis Challenges. ArXiv:2106.12949, 2021

[4] Stadler et al. Synthetic Data - Anonymisation Groundhog Day. ArXiv:2011.07018, 2020

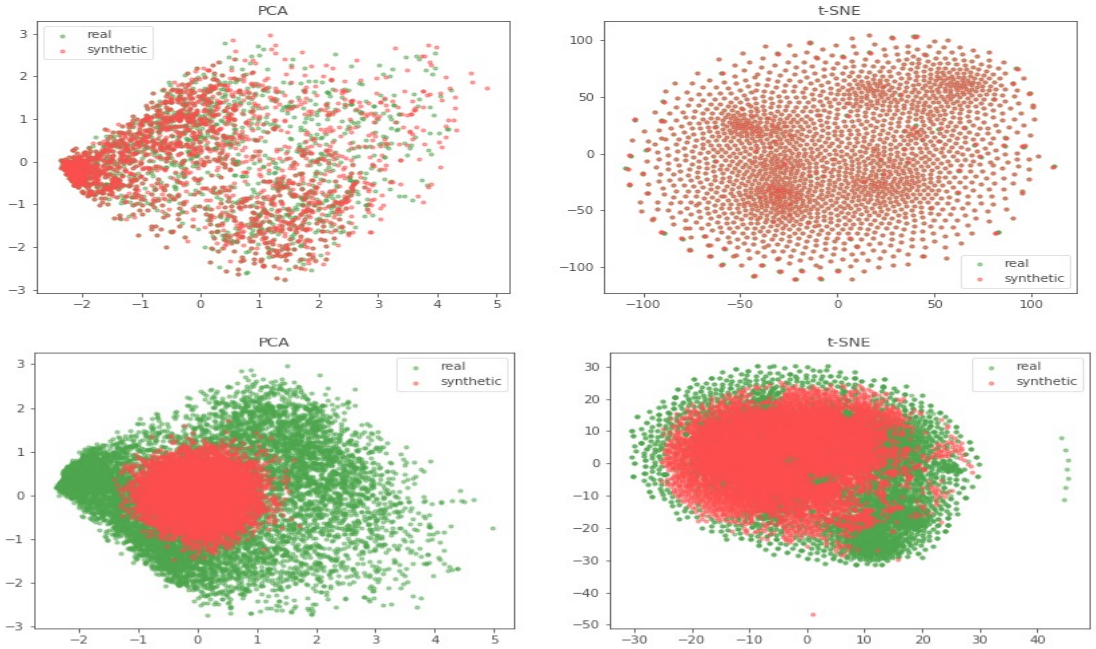
Sharing via Synthetic data

Synthetic data suffer from limitations similar to anonymization (no free lunch)

→ Utility and privacy tradeoff difficult to predict [1]

→ Only the provider of the model is aware of the actual tradeoff

May reproduce, generate or mitigate bias: impact on minority groups of the real data set amplifying fairness issues



[1] Stadler et al. Synthetic Data - Anonymisation Groundhog Day. ArXiv:2011.07018, 2020

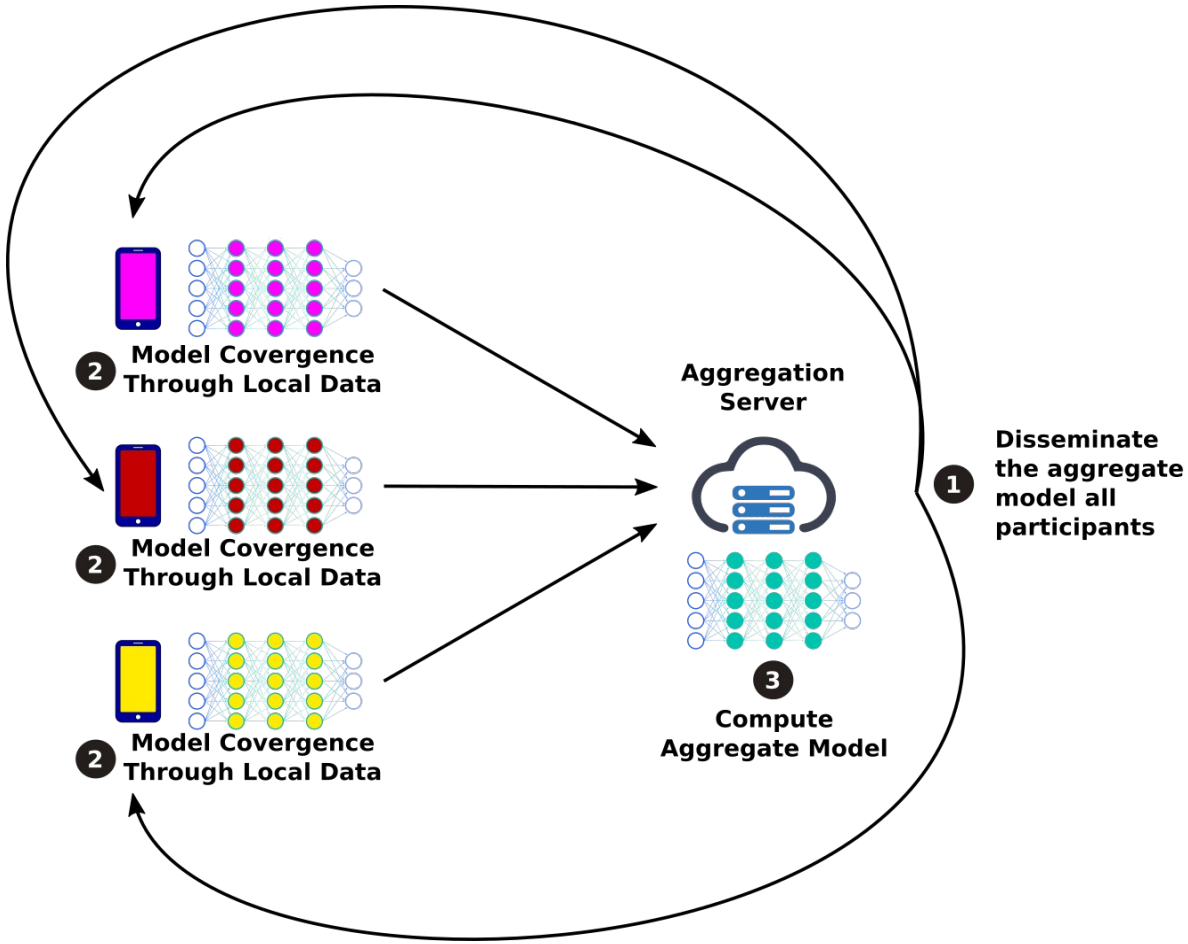
Agenda

- **Generative models: data sharing via data anonymisation vs synthetic data**

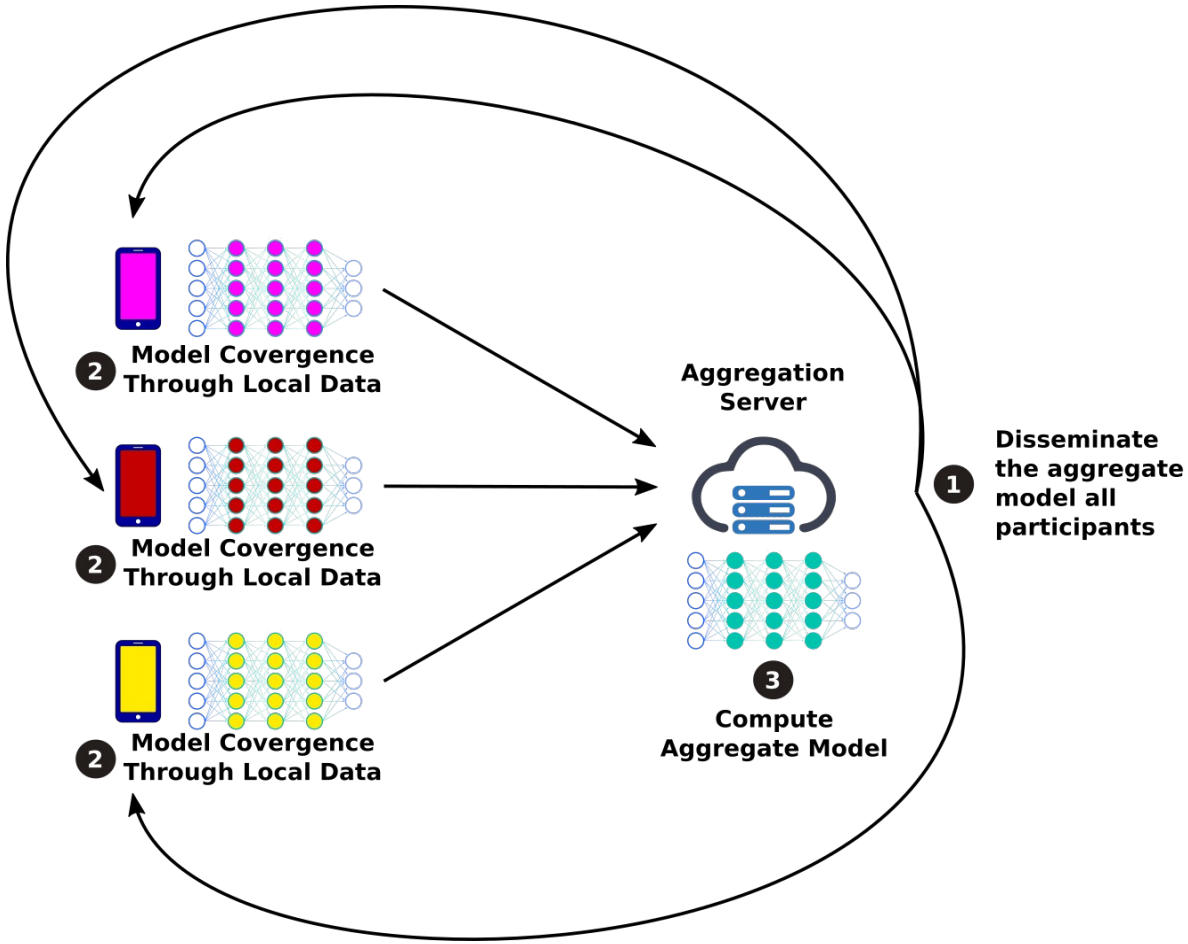
- **Federated Learning (FL) → OT3 Sécurité et Vie Privée**
  - **FL using personalized and private layers**
  - **MixNN: Protection of FL Against Inference Attacks by Mixing Neural Network Layers**
  - **Quantifying the learning gains and the vulnerabilities to adversarial attacks**
  - **Organisation of challenges**

- **NLP Models: privacy risks and countermeasures → OT3 Sécurité et Vie Privée**

Federated Learning



Federated Learning



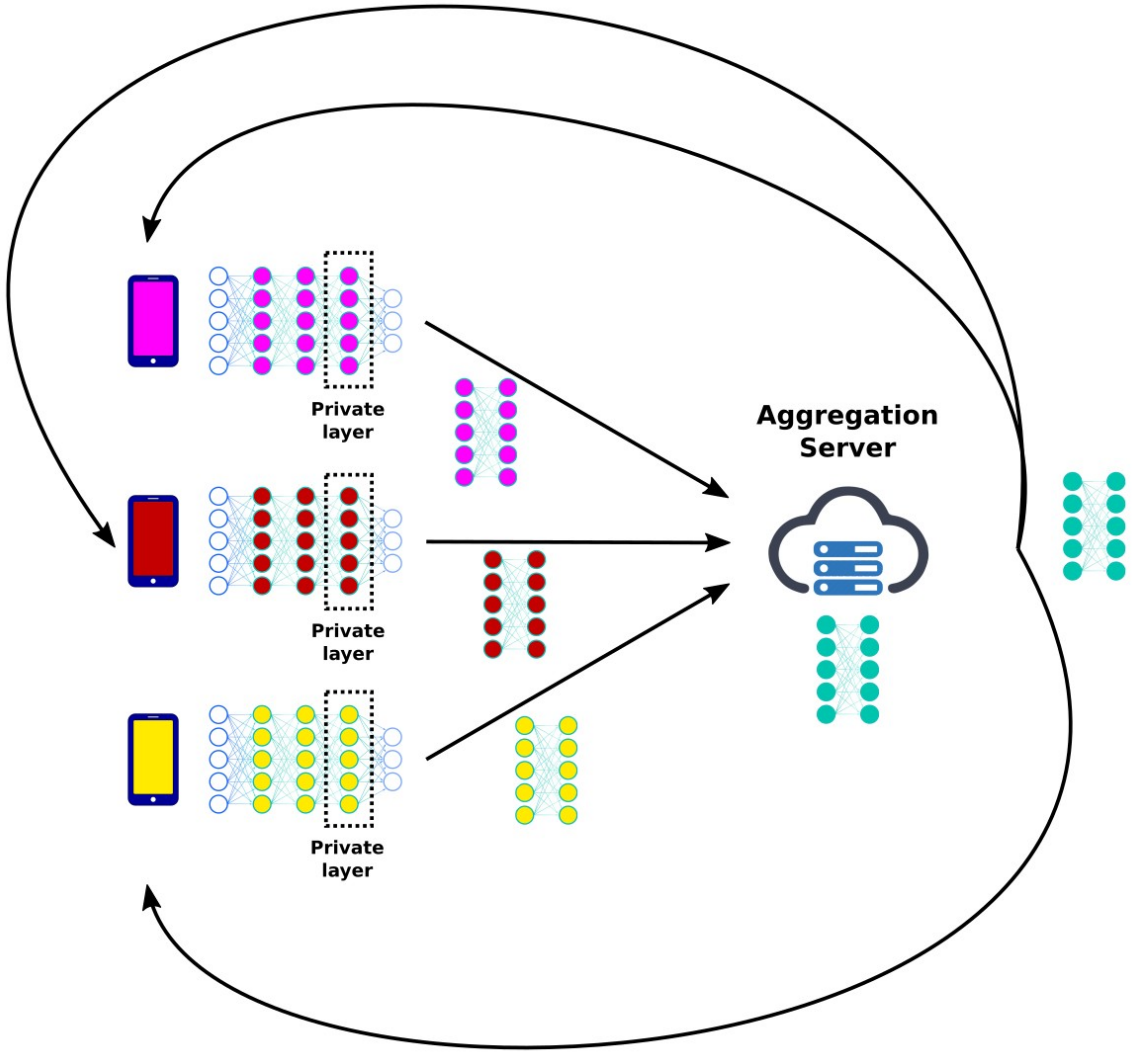
Cross device vs cross silo

Avantages
- **Confidentiality**
- **Sovereignty**

Limitations
- **Data heterogeneity**
- **Fairness**
- **Poisoning**
- **Privacy leakage**

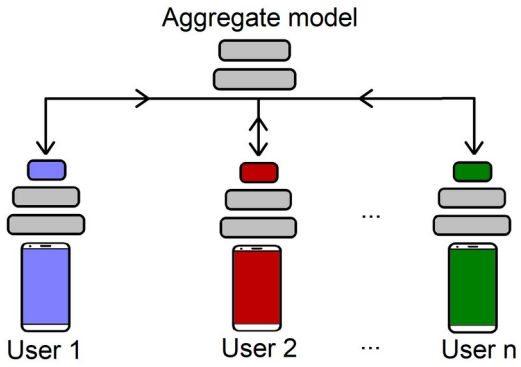
Countermeasures
- **Perturbation (e.g., differential privacy)**
  - **Drastically reduces accuracy**
- **Crypto (e.g., FHE, secure aggregation)**
  - **Important overhead**

FL using personalized and private layers *[MLSP' 21]*

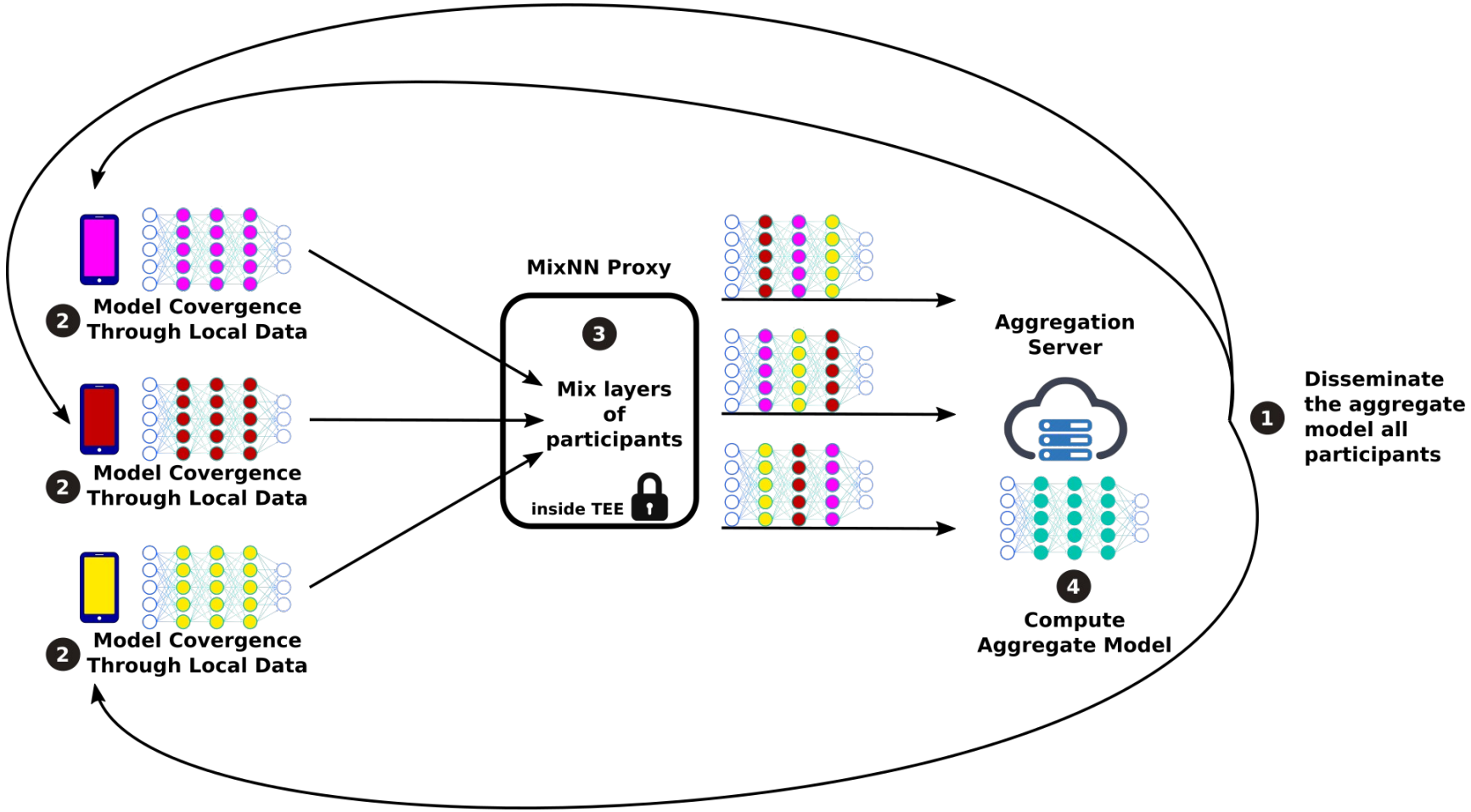


Goal: minimize the information shared with the aggregation server

The private layers can be refined for personalized tasks



MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers *[Middleware' 22]*



Motivations:

- Protection against a curious aggregation server

- Improving the privacy without utility loss

- Can be transparent to the users

Agenda

- **Generative models: data sharing via data anonymisation vs synthetic data**

- **Federated Learning (FL) → OT3 Sécurité et Vie Privée**
  - **FL using personalized and private layers**
  - **MixNN: Protection of FL Against Inference Attacks by Mixing Neural Network Layers**
  - **Quantifying the learning gains and the vulnerabilities to adversarial attacks**
  - **Organisation of challenges**

- **NLP Models: privacy risks and countermeasures → OT3 Sécurité et Vie Privée**

Organisation of challenges

- AI systems need to be audited to ensure well behavior and avoid bias
- The attack surface is not yet well known
- To stimulate new work and improve the state of the art in the field in a fun way: organization of competitions

Privacy Challenge on Federated Learning

- Implementation of poisoning attacks and defense schemes
- Membership Inference Attack and contremeasures
- Introduction of backdoors

Agenda

- **Generative models: data sharing via data anonymisation vs synthetic data**

- **Federated Learning (FL) → OT3 Sécurité et Vie Privée**
  - **FL using personalized and private layers**
  - **MixNN: Protection of FL Against Inference Attacks by Mixing Neural Network Layers**
  - **Quantifying the learning gains and the vulnerabilities to adversarial attacks**
  - **Organisation of challenges**

- **NLP Models: privacy risks and countermeasures → OT3 Sécurité et Vie Privée**

Workshop on Auditing AI

Auditing AI: what issues? what prospects?

May 13th 2024, Campus Cyber

Call for poster

Thanks

antoine.boutet@insa-lyon.fr