

PRIVACY PRESERVING MACHINE LEARNING

LECTURE 1: INTRODUCTION & COURSE OVERVIEW

Aurélien Bellet (Inria)

Master 2 Data Science, University of Lille

- Researcher at Inria Lille
- Member of the [Magnet team](#) (Machine learning in information networks) on ML in/with graphs and applications to NLP
- My current research topics of interest:
 - [Privacy-preserving ML](#)
 - [Decentralized and Federated ML](#)
 - Representation learning for NLP & speech
 - Fairness in ML
- More details and contact info on my [homepage](#)

- 12 sessions of 2 hours, roughly one per week until January 12, 2023
- Lectures and lab sessions
- Evaluation:
 - 1 practical (50%)
 - 1 report & presentation of research paper (50%)
- Course page (with lecture slides, practicals, textbook references, etc):
http://researchers.lille.inria.fr/abellet/teaching/private_machine_learning_course.html

1. Context & motivation
2. Course overview

CONTEXT & MOTIVATION

Ability of an individual
to seclude themselves or to withhold information about themselves

(“right to be let alone”)

- **Massive collection of personal data** by companies and public organizations, driven by the progress of data science and AI



- Data is **increasingly sensitive and detailed**: browsing history, purchase history, social network posts, speech, geolocation, health...
- It is sometimes **shared unknowingly** and **without a clear understanding of the risks**

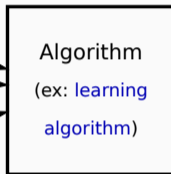
SOME RISKS OF PRIVACY BREACHES

- Improper disclosure of data can have **adverse consequences for individuals**:
 - Credentials
 - Examples: credit card number, home access code, passwords
 - Risks: stealing personal property
 - Identification information
 - Examples: name, bank information, biometric data
 - Risks: identity theft
 - **Information about an individual**
 - Examples: medical status, religious beliefs, political opinions, sexual preferences
 - Risks: discrimination, blackmailing, unsolicited micro-targeting, public shame...
- Some of these risks can affect anyone (even if they think they have “**nothing to hide**”) and without individuals knowing it (cf. Cambridge Analytica scandal)

- There is **increasing regulation to address privacy-related harms** related to the collection, use and release of personal data
 - General regulations (e.g., adoption of GDPR by the EU in 2018)
 - Sector- and context-specific regulations, e.g. in health, education, research, finance...
- **Privacy has a cost on the utility** of the analysis, but ideally it **should not destroy it**
- One of the main goals of privacy research is to **find good trade-offs between utility and privacy** so we can **better protect individuals** and also **unlock new applications**

(Figure inspired from R. Bassily)

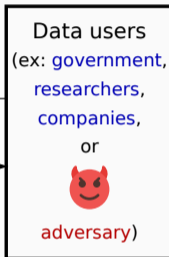
Individuals
(data subjects)



queries

answers

(ex: aggregate statistics,
machine learning model)



- Goal: achieve utility while preserving privacy (conflicting objectives!)
- This is separate from security concerns (e.g., unauthorized access to the system)
- Any ideas on how to do this?


DATA “ANONYMIZATION” IS NOT SAFE

Name	Birth date	Zip code	Gender	Diagnosis	...
Ewen Jordan	1993-09-15	13741	M	Asthma	...
Lea Yang	1999-11-07	13440	F	Type-1 diabetes	...
William Weld	1945-07-31	02110	M	Cancer	...
Clarice Mueller	1950-03-13	02061	F	Cancer	...

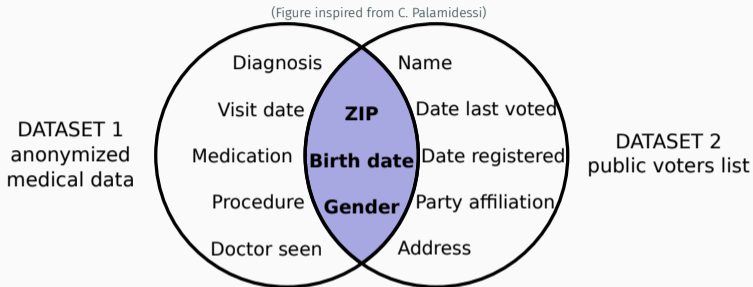
- **Anonymization:** removing **personally identifiable information** before publishing data
- First solution: **strip attributes that uniquely identify an individual** (e.g., name, social security number...)

DATA “ANONYMIZATION” IS NOT SAFE

Name	Birth date	Zip code	Gender	Diagnosis	...
	1993-09-15	13741	M	Asthma	...
	1999-11-07	13440	F	Type-1 diabetes	...
	1945-07-31	02110	M	Cancer	...
	1950-03-13	02061	F	Cancer	...

- **Anonymization:** removing **personally identifiable information** before publishing data
- First solution: **strip attributes that uniquely identify an individual** (e.g., name, social security number...)
- Now we cannot know that William Weld has cancer!
- Or can we? 

DATA “ANONYMIZATION” IS NOT SAFE



- **Problem:** susceptible to **linkage attacks**, i.e. uniquely linking a record in the anonymized dataset to an identified record in a public dataset
- For instance, an estimated 87% of the US population is uniquely identified by the combination of their gender, birthdate and zip code
- In the late 90s, L. Sweeney managed to re-identify the medical record of the governor of Massachusetts using a public voters list

DATA “ANONYMIZATION” IS NOT SAFE

Name	Birth date	Zip code	Gender	Diagnosis	...
	1993-09-15	13741	M	Asthma	...
	1999-11-07	13440	F	Type-1 diabetes	...
	1945-07-31	02110	M	Cancer	...
	1950-03-13	02061	F	Cancer	...

- Second solution: *k*-anonymity [Sweeney, 2002]
 1. Define a set of attributes as quasi-identifiers (QIs)
 2. Suppress/generalize attributes and/or add dummy records to make every record in the dataset indistinguishable from at least $k - 1$ other records with respect to QIs

DATA “ANONYMIZATION” IS NOT SAFE

	Quasi identifiers			Sensitive attribute	
Name	Age	Zip code	Gender	Diagnosis	...
	20-30	13***		Asthma	...
	20-30	13***		Type-1 diabetes	...
	70-80	02***		Cancer	...
	70-80	02***		Cancer	...

- Second solution: *k*-anonymity [Sweeney, 2002]
 1. Define a set of attributes as *quasi-identifiers* (QIs)
 2. Suppress/generalize attributes and/or add dummy records to *make every record in the dataset indistinguishable from at least $k - 1$ other records with respect to QIs*
- Better now?
- No! *Can still infer that W. Weld has cancer* (everyone in the group has cancer)

DATA “ANONYMIZATION” IS NOT SAFE

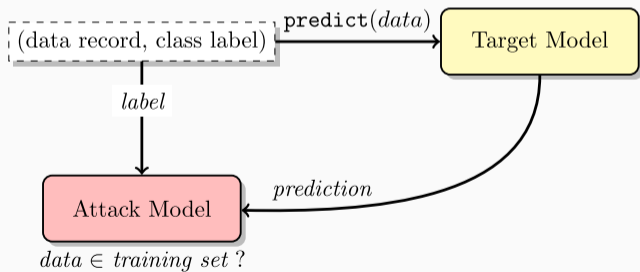
- Variants of k -anonymity (t -closeness, ℓ -diversity) try to address the previous issue but require to modify the original data even more, which often destroys utility
- In high-dimensional and sparse datasets, any combination of attributes is a potential PII that can be exploited using appropriate auxiliary knowledge
 - De-anonymization of Netflix dataset protected with k -anonymity using a few public ratings from IMDB [Narayanan and Shmatikov, 2008]
 - De-anonymization of Twitter graph using Flickr [Narayanan and Shmatikov, 2009]
 - 4 spatio-temporal points uniquely identify most people [de Montjoye et al., 2013]
- **Conclusion:** data cannot be fully anonymized AND remain useful

AGGREGATE STATISTICS ARE NOT SAFE

- How about releasing **aggregate statistics about many individuals**?
- **Problem 1: differencing attacks**, i.e. combining aggregate queries to obtain precise information about specific individuals (note: this can be hard to detect)
 - Average salary in a company before and after an employee joins
- **Problem 2: membership inference attacks**, i.e. inferring presence of known individual in a dataset from (high-dimensional) aggregate statistics
 - Statistics about genomic variants [[Homer et al., 2008](#)]
- **Problem 3: reconstruction attacks**, i.e. inferring (part of) the dataset from the output of many aggregate queries [[Dinur and Nissim, 2003](#)]
 - See [this short video](#) to understand the basic idea of the Dinur-Nissim
 - See [this blog post](#) and [longer video](#) to learn how this was applied by the US Census

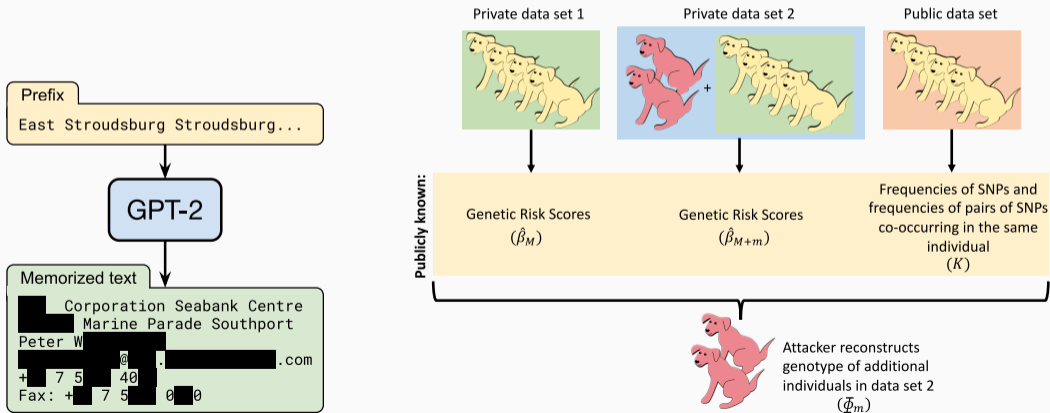
ML MODELS ARE NOT SAFE

- ML models are elaborate kinds of aggregate statistics!
- As such, they are susceptible to **membership inference attacks**, i.e. inferring the presence of a known individual in the training set
- For instance, one can exploit the confidence in model predictions [Shokri et al., 2017] [Carlini et al., 2022]



ML MODELS ARE NOT SAFE

- ML models are also susceptible to **reconstruction attacks**
- For instance, one can **extract sensitive text from large language models** [Carlini et al., 2021] or **run differencing attacks on ML models** [Paige et al., 2020]



- **Revealing ordinary facts to inappropriate parties** may also be problematic, especially if an individual is followed over time
- Example: Alice buys bread every day for 20 years and then stops
 - An analyst might conclude that Alice has been diagnosed with type 2 diabetes
 - This may be wrong, but in any case Alice could be harmed (e.g., charged with higher insurance premiums)

1. **Auxiliary knowledge** (also called **background knowledge** or **side information**): we need to be robust to whatever knowledge the adversary may have, since we cannot predict what an adversary knows or might know in the future
2. **Multiple analyses**: we need to be able to track how much information is leaked when asking several questions about the same data, and avoid catastrophic leaks

COURSE OVERVIEW

IN THIS COURSE YOU WILL LEARN...

1. How to mathematically define “privacy” in a robust manner
2. What are the basic building blocks of a private algorithm
3. How to design algorithms that provide high utility while preserving privacy
4. How to apply these concepts to data analytics and machine learning

First attempt at privacy definition

“An analysis of a dataset is private if **the result reveals no more** about an individual **than what was already known** about him/her before the analysis.”

- Bayesian version: posterior belief same as prior belief
- **Problem 1:** **Impossible to reveal exactly nothing** if the result is to depend at all on the data (otherwise we get zero utility)

First attempt at privacy definition

“An analysis of a dataset is private if **the result reveals no more** about an individual **than what was already known** about him/her before the analysis.”

- **Problem 2:** “Before/after” requirement **unachievable under auxiliary knowledge**
- Think of “stupid priors” (e.g., a person’s height is between 10 and 20 meters)
- Think about whether Bob’s privacy was violated in the following example:
 - Suppose an insurance company knows that Bob is a smoker
 - A medical data analysis reveals that smoking and cancer are correlated
 - The insurance company decides to raise Bob’s rates
- This happens **even if Bob’s data wasn’t included in the analysis!**
- Such correlations are precisely **the kind of things we want to be able to learn**

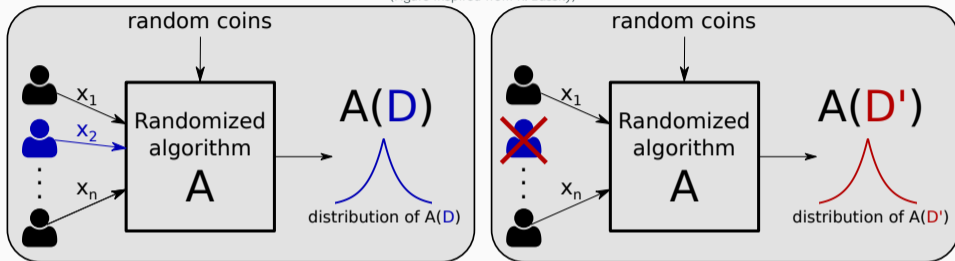
Second attempt at privacy definition

“An analysis of a dataset is private if what can be learned about an individual in the dataset **is not much more** than what would be learned **if the same analysis was conducted without him/her in the dataset.**”

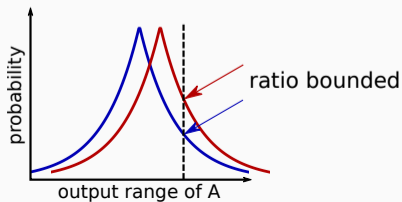
- Intuition: **cannot infer the presence/absence of an individual in the dataset, or anything “specific” about an individual** (here, ‘specific’ refers to information that cannot be inferred unless the individual’s data is used in the analysis)
- Note: to be robust to auxiliary knowledge, **randomization is necessary**
 - Consider a deterministic algorithm which is non-trivial (i.e., there exists a query and two datasets that yield different results)
 - Changing one record at a time, we see there exists a pair of datasets differing only in a single record on which the same query yields different results
 - An adversary knowing that the dataset is one of these two learns the differing record

DIFFERENTIAL PRIVACY

(Figure inspired from R. Bassily)



- **Neighboring** datasets $D = \{x_1, x_2, \dots, x_n\}$ and $D' = \{x_1, x_3, \dots, x_n\}$
- **Requirement:** $\mathcal{A}(D)$ and $\mathcal{A}(D')$ should have “close” distribution



Definition (informal)

\mathcal{A} is (ϵ, δ) -differentially private (DP) if for all neighboring datasets D, D' and all sets S :

$$\Pr[\mathcal{A}(D) \in S] \leq e^\epsilon \Pr[\mathcal{A}(D') \in S] + \delta.$$

- For meaningful privacy guarantees, think of $\epsilon \leq 1$ and $\delta \ll 1/n$
- We refer to “pure” ϵ -DP when $\delta = 0$
- **Key principle:** **privacy is a property of the analysis**, not of a particular output (in contrast to e.g., k -anonymization)
- First proposed in [Dwork et al., 2006] by Dwork, McSherry, Nissim and Smith (who won the Gödel prize in 2017)

- DP is **immune to post-processing**: it is impossible to compute a function of the output of the private algorithm and make it less differentially private
- DP is **robust to arbitrary auxiliary knowledge**: it **bounds the relative advantage** that an adversary gets from observing the output of an algorithm
- DP is **robust under composition**: if multiple analyses are performed on the same data, as long as each one satisfies DP, all the information released taken together will still satisfy DP (albeit with a degradation in the parameters)

- Suppose that $\mathcal{A}(D) = (\mathcal{A}_1(D), \dots, \mathcal{A}_K(D))$ where each \mathcal{A}_k is (ϵ, δ) -DP
- What are the privacy guarantees of \mathcal{A} ?
- **Simple composition**: \mathcal{A} is $(K\epsilon, K\delta)$ -DP (simple proof)
- **Advanced composition**: \mathcal{A} is $(\epsilon', K\delta + \delta')$ -DP with $\epsilon' \approx \sqrt{K \log(1/\delta')} \epsilon$ (more involved)

Private computation of numeric functions via **output perturbation** (noise addition)

- **Laplace mechanism**: $\mathcal{A}(D) = f(D) + Y$ with $Y \sim \text{Lap}(C/\epsilon) \rightarrow \mathcal{A}$ is ϵ -DP
- **Gaussian mechanism**: $\mathcal{A}(D) = f(D) + Y$ with $Y \sim \mathcal{N}(0, C \cdot \frac{\log(1/\delta)}{\epsilon^2}) \rightarrow \mathcal{A}$ is (ϵ, δ) -DP

Private computation of non-numeric functions via **sampling from a distribution**

- **Exponential mechanism**
 - For every possible output r of the function $f(D)$, assign a score $s(r, D)$
 - Sample an output r with probability proportional to $\exp(s(r, D) \cdot \epsilon)$
 - This mechanism is ϵ -DP

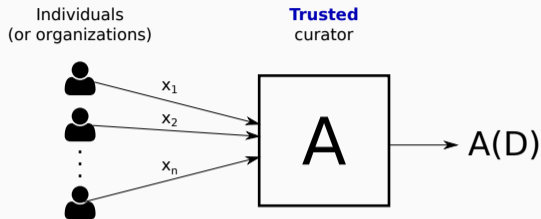
- We will mainly consider **Empirical Risk Minimization** (ERM):

$$\hat{\theta} \in \arg \min_{\theta} \left\{ F(\theta) := \frac{1}{n} \sum_{i=1}^n L(\theta; x_i, y_i) \right\}$$

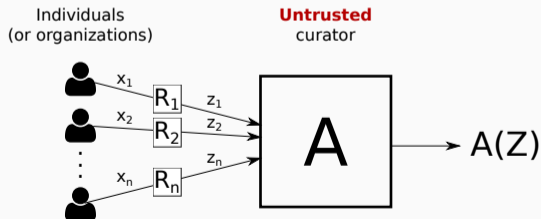
- **Output perturbation**: adding noise to $\hat{\theta}$
- **Objective perturbation**: adding noise to $F(\theta)$
- **Gradient perturbation**: adding noise to $\nabla F(\theta^t)$ + composition over iterations
- **Privacy amplification** by subsampling and by iteration
- Results on the utility/privacy trade-offs in ERM

UNTRUSTED CURATOR SETTING

Trusted curator model (also called global model or centralized model):
 \mathcal{A} is differentially private wrt dataset D



Untrusted curator model (also called local model or distributed model):
Each \mathcal{R}_i is differentially private wrt record (or local dataset) x_i



- **Local differential privacy** (LDP): extreme case where each participant holds a dataset of size 1 (e.g., his/her own personal record)
- **Large utility gap** between global and local models
 - Example: for averaging, error is $\Theta_\epsilon(1/n)$ in global model and $\Theta_\epsilon(1/\sqrt{n})$ in local model
 - LDP only useful for very large n (e.g., large-scale industrial applications)
- Can consider **intermediate trust models** and/or use **cryptographic primitives** to obtain better utility
 - **Secure aggregation**: curator only observes average of messages
 - **Shuffling**: curator cannot tie a message to a particular participant
- Applications to **federated learning** [Kairouz et al., 2021]

OTHER TOPICS (IF TIME PERMITS)

- Theoretical limits in the trade-off between utility and privacy (lower bounds)
- Variants of differential privacy
- Inference attacks on ML models
- Robustness to malicious participants
- Learning anonymized representations

Many opportunities to work on privacy-preserving ML and federated learning in **Magnet**:

- Master internships
- PhD positions
- Engineer positions
- ...

Talk to me if you're interested!

- [Carlini et al., 2022] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022).
Membership inference attacks from first principles.
In *S&P*.
- [Carlini et al., 2021] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021).
Extracting training data from large language models.
In *USENIX Security*.
- [de Montjoye et al., 2013] de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013).
Unique in the crowd: The privacy bounds of human mobility.
Scientific Reports, 3.
- [Dinur and Nissim, 2003] Dinur, I. and Nissim, K. (2003).
Revealing information while preserving privacy.
In *PODS*.
- [Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.
In *Theory of Cryptography (TCC)*.

- [Homer et al., 2008] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008).
Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays.
PLOS Genetics, 4(8):1–9.
- [Kairouz et al., 2021] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D’Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021).
Advances and Open Problems in Federated Learning.
Foundations and Trends® in Machine Learning, 14(1–2):1–210.
- [Narayanan and Shmatikov, 2008] Narayanan, A. and Shmatikov, V. (2008).
Robust de-anonymization of large sparse datasets.
In *IEEE Symposium on Security and Privacy (S&P)*.

- [Narayanan and Shmatikov, 2009] Narayanan, A. and Shmatikov, V. (2009).
De-anonymizing social networks.
In *IEEE Symposium on Security and Privacy (S&P)*.
- [Paige et al., 2020] Paige, B., Bell, J., Bellet, A., Gascón, A., and Ezer, D. (2020).
Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores.
In *International Conference on Research in Computational Molecular Biology RECOMB*.
- [Shokri et al., 2017] Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017).
Membership inference attacks against machine learning models.
In *IEEE Symposium on Security and Privacy (S&P)*.
- [Sweeney, 2002] Sweeney, L. (2002).
k-anonymity: A model for protecting privacy.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570.