# (numerical) Artificial intelligence & privacy protection

## Approaches and challenges

Aurélien Bellet (Inria, Magnet team)

Journée HumAIn
February 8, 2019

# WHY SHOULD WE CARE?

# CURRENT STATE OF THINGS

- Connected devices (phones, watches, home assistants, sensors) collect data about people everywhere, in all their activities
- New privacy issues raised by artificial intelligence and its use by large tech companies owning massive amounts of personal data
- AI technology can be invasive (e.g., personalized ads and content) and lead to unfair treatment (e.g., Google showing ads for well-paid jobs to men more than women)
- Increasing political and societal awareness
- GDPR came into effect in May 2018
  - "Privacy-by-design" (can be via pseudo-anonymization)
  - Transparency, accountability
  - Fines up to 4% of yearly revenue

# Data re-identification and re-construction

- Data pseudo-anonymization consists in removing personally identifiable information from a dataset
- Some other combination of features (“quasi-identifiers”) can potentially be used to re-identify the person
- This is difficult to assess, and it is impossible to account for all potential additional public data or background knowledge
- Possible to perform data reconstruction only through query access to the database or the trained machine learning model

# DATA RE-IDENTIFICATION: AOL BROWSING LOGS IN 2006

The New York Times

**Technology**

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERA

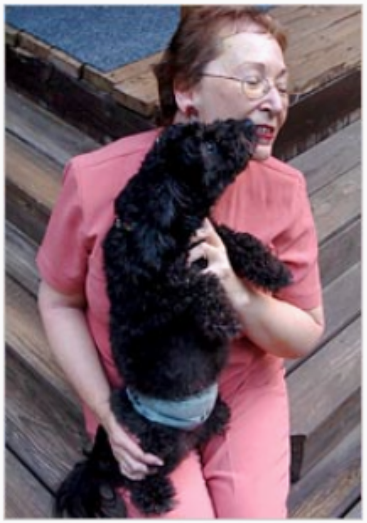
## A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS



Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

# DATA RE-IDENTIFICATION: NYC TAXI DATA IN 2014

The Guardian

## New York taxi details can be extracted from anonymised data, researchers say

**FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income**

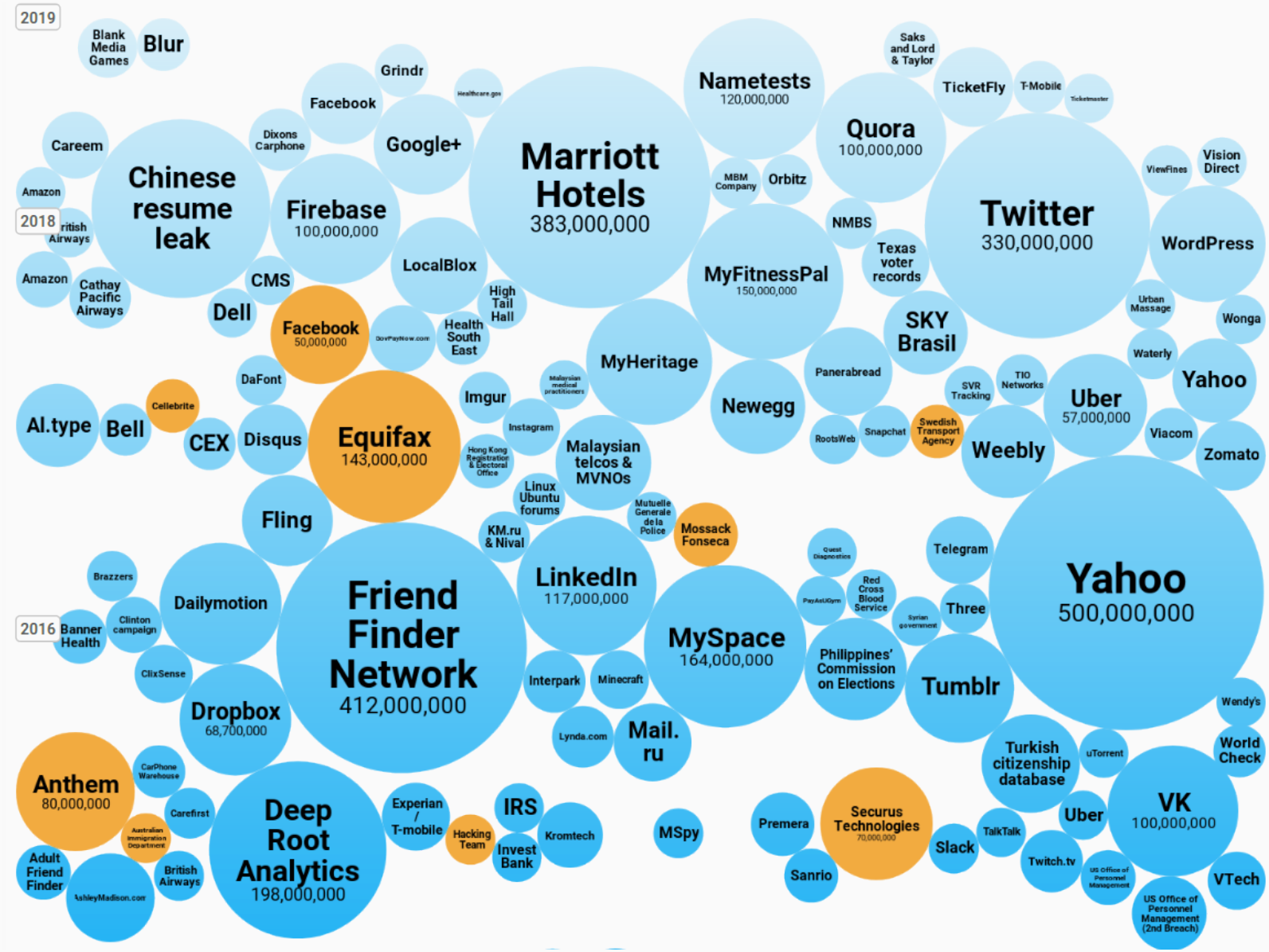

Data about New York city taxi drivers and rides could be de-anonymised, researchers warn. Photograph: Jan Johannessen/Getty Images

**Alex Hern**

Fri 27 Jun 2014 15.57 BST

New York City has released data of 173m individual taxi trips - but inadvertently made it "trivial" to find the personally identifiable information of every driver in the dataset.

# DATA BREACHES IN CENTRALIZED DATABASES




credit: https://informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

# UNLOCKING NEW OPPORTUNITIES

- In some cases, several organizations own data which, if combined, would offer new opportunities
- But they do not want to share it due to legal/IP reasons, or to preserve their competitive advantage
- Example: large-scale medical studies across many hospitals from different countries

# LESSONS LEARNED THE HARD WAY

1. **Anonymization is not enough:** we need more robust privacy definitions
2. **Centralization is risky and costly:** we need a shift to a more decentralized paradigm

# CENTRALIZED SETTING

# CENTRALIZED SETTING

- A trusted party holding a private dataset $D$
- **Goal:** trusted party wants to release useful quantities computed on the data while preserving privacy of the individual records
- Example 1: a hospital wants to study the link between factors and diseases (e.g., smoking and lung cancer)
- Example 2: open data initiatives

# CENTRALIZED DIFFERENTIAL PRIVACY: DEFINITION

- $\mathcal{A}$: a randomized algorithm taking a dataset as input
- Neighboring datasets $D$ and $D'$: differ in a single data point
- For $\epsilon > 0$, $\mathcal{A}$ is $\epsilon$-differentially private ($\epsilon$-DP) [Dwork, 2006] if for all neighboring datasets and sets of possible outputs $\mathcal{O}$:

$$Pr(\mathcal{A}(D) \in \mathcal{O}) \leq e^{\epsilon} Pr(\mathcal{A}(D') \in \mathcal{O})$$

- **Interpretation:** the output of $\mathcal{A}$ does not reveal whether a particular data point was used

# CENTRALIZED DIFFERENTIAL PRIVACY: PROPERTIES

- Robustness to background knowledge: the guarantee holds even if the attacker knows all the dataset except for one record
- Allows to track privacy budget for multiple analyses on the same dataset: if 2 algorithms are $\epsilon$-DP, then their composition is $2\epsilon$-DP
- Invariance to postprocessing: if an output is $\epsilon$-DP, then any processing independent from the data is also $\epsilon$-DP

$\rightarrow$ DP is often seen as a **gold standard for privacy**, and will be used by the US Census starting in 2020

# CENTRALIZED DIFFERENTIAL PRIVACY: LAPLACE MECHANISM

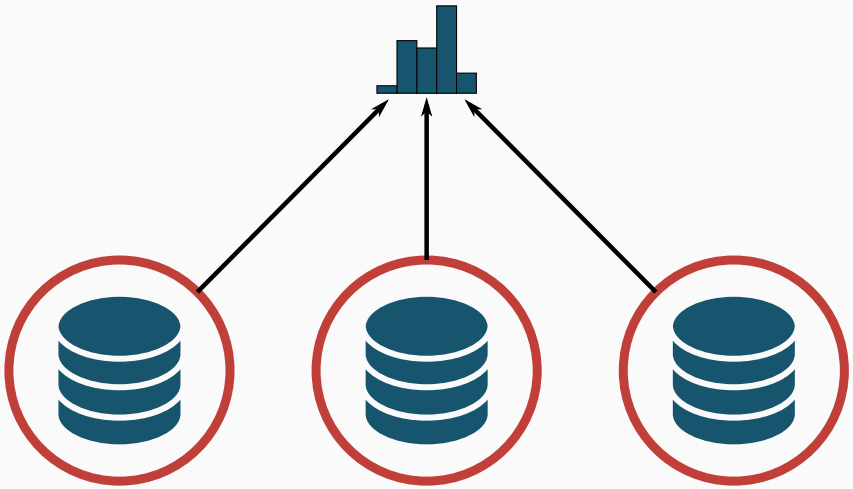
- Assume we want to compute a function $f : \mathcal{D} \to \mathbb{R}$
- Sensitivity $s_f$ of $f$: maximum possible difference $|f(D) - f(D')|$ for two neighboring datasets
- Example: a counting query has sensitivity 1
- Consider $\mathcal{A}_L(D, f, \epsilon) = f(D) + \eta$ where $\eta$ is a random perturbation drawn from the centered Laplace distribution of scale $s_f/\epsilon$
- $\mathcal{A}_L(D, f, \epsilon)$ satisfies $\epsilon$-DP: this is known as the Laplace mechanism
- $\mathcal{A}_L(D, f, \epsilon)$ has expected error $s_f/\epsilon$ (this is worst-case optimal)
- This illustrates the idea of a privacy-utility trade-off

# CENTRALIZED DIFFERENTIAL PRIVACY: APPLICATION TO ML

- In machine learning, the function $f(D)$ outputs a model trained on the dataset $D$
- Examples: a classifier, some cluster centers...
- We can sometimes bound the sensitivity of $f$ directly (often very loosely) and perturb the output model [Chaudhuri et al., 2011]
- It is often better to bound the sensitivity of each step of the learning algorithm (e.g., gradient descent) and perturb each intermediate quantities [Bassily et al., 2014]

# MULTI-PARTY SETTING

# MULTI-PARTY SETTING



- Several parties with data they do not want to share
- The parties, or an untrusted third party (e.g., service provider), want to compute useful quantities on the joint data
- Honest-but-curious parties: follow the protocol, but will infer information about others if they can
- Example 1: hospitals with patient data want to conduct a joint study on a rare disease
- Example 2: identify most popular websites among browser users

# LOCAL DIFFERENTIAL PRIVACY: DEFINITION

- For simplicity, assume each party $i$ has a single data point $x_i$
- $\mathcal{A}$: a randomized algorithm taking a single data point as input
- $\mathcal{A}$ is $\epsilon$-locally differentially private ($\epsilon$-LDP) [Duchi et al., 2012] if for all points $x, x'$ and sets of possible outputs $\mathcal{O}$:

$$Pr(\mathcal{A}(x) \in \mathcal{O}) \leq e^{\epsilon} Pr(\mathcal{A}(x') \in \mathcal{O})$$

- **Interpretation:** plausible deniability (owner can deny having value $x$ on basis of lack of evidence)
- Stronger requirement than centralized DP
- Early instance of local DP used for conducting sensitive surveys: randomized response [Warner, 1965]

# LOCAL DIFFERENTIAL PRIVACY: LAPLACE MECHANISM

- Assume we want to compute a function $f(D) = \sum_{i=1}^{n} g(x_i)$ where each $g(x_i) \in \mathbb{R}$
- We can apply the Laplace mechanism $\mathcal{A}_L(x_i, g, \epsilon)$ locally to each $g(x_i)$, using the sensitivity of $g$ instead of $f$
- The aggregate $\sum_{i=1}^{n} \mathcal{A}_L(x_i, g, \epsilon)$ has expected error $\sqrt{N} s_g / \epsilon$ (this is again worst-case optimal)
- Large gap between centralized and local models: need many parties to obtain useful values with local DP
- Combination with secure multi-party computation approaches can reduce or remove this gap

# LOCAL DIFFERENTIAL PRIVACY: APPLICATION TO ML

- Parties can learn a model locally, perturb it and share: the models can then be aggregated [Pathak et al., 2010]
- Parties can engage in a decentralized algorithm where they share perturbed intermediate quantities

  [Shokri and Shmatikov, 2015, Bellet et al., 2018]

# OPEN CHALLENGES

# SOME OPEN CHALLENGES

- Scalability of algorithms to large number of parties (e.g., mobile phones, IoT) [Bellet et al., 2018]
- Taking full advantage of privacy amplification schemes ( [Úlfar Erlingsson et al., 2018]: anonymity strikes back?)
- Design more tailored, possibly relaxed notions of privacy, with generic mechanisms [Kifer and Machanavajjhala, 2014]
- Match formal privacy guarantees with intuitive notion of privacy for a given application, and relation with legal definitions
- Dealing with collusion and malicious behavior [Dellenbach et al., 2018]
- User-friendly, generic implementations of privacy-preserving machine learning in the multi-party setting (cf OpenMined)

Thank you for your attention!

Questions?

# REFERENCES I


[Bassily et al., 2014] Bassily, R., Smith, A. D., and Thakurta, A. (2014).
**Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds.**
In *FOCS*.

[Bellet et al., 2018] Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. (2018).
**Personalized and Private Peer-to-Peer Machine Learning.**
In *AISTATS*.

[Chaudhuri et al., 2011] Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011).
**Differentially Private Empirical Risk Minimization.**
*Journal of Machine Learning Research*, 12:1069–1109.

[Dellenbach et al., 2018] Dellenbach, P., Bellet, A., and Ramon, J. (2018).
**Hiding in the Crowd: A Massively Distributed Algorithm for Private Averaging with Malicious Adversaries.**
Technical report, arXiv:1803.09984.

[Duchi et al., 2012] Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2012).
**Privacy Aware Learning.**
In *NIPS*.

[Dwork, 2006] Dwork, C. (2006).
**Differential Privacy.**
In *ICALP*, volume 2.

# References II


[Jayaraman et al., 2018] Jayaraman, B., Wang, L., Evans, D., and Gu, Q. (2018).
**Distributed learning without distress: Privacy-preserving empirical risk minimization.**
In *NeurIPS*.

[Kifer and Machanavajjhala, 2014] Kifer, D. and Machanavajjhala, A. (2014).
**Pufferfish: A framework for mathematical privacy definitions.**
*ACM Transactions on Database Systems*, 39(1):3:1–3:36.

[Pathak et al., 2010] Pathak, M. A., Rane, S., and Raj, B. (2010).
**Multiparty Differential Privacy via Aggregation of Locally Trained Classifiers.**
In *NIPS*.

[Shokri and Shmatikov, 2015] Shokri, R. and Shmatikov, V. (2015).
**Privacy-Preserving Deep Learning.**
In *CCS*.

[Warner, 1965] Warner, S. L. (1965).
**Randomized response: A survey technique for eliminating evasive answer bias.**
*Journal of the American Statistical Association*, 60:63–69.

[Úlfar Erlingsson et al., 2018] Úlfar Erlingsson, Feldman, V., Mironov, I., and abd Kunal Talwar, A. R. (2018).
**Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity.**
Technical report, arXiv:1811.12469.

# Secure multi-party computation (MPC)

- **Goal:** parties jointly compute function $f(D)$ without learning anything else than the output
- Based on cryptographic primitives such as homomorphic encryption, oblivious transfer, garbled circuits
- Complementary to DP:
  - Relies on a computational assumption
  - No utility loss, but high computational cost
  - Does not protect against information leaked by $f(D)$

# DP + MPC

- DP and MPC can be efficiently combined in some special cases
- Consider again count queries, which only require to sum quantities from each party (called secure aggregation in MPC)
- MPC protocols for computing noisy output $f'(D) = f(D) + \eta$ are fairly efficient when the number of parties is not too large
- No individual values are observed by any party $\rightarrow$ we recover the utility of centralized DP
- Direct application to ML: privately aggregate locally trained models or iterative updates from different parties with better utility than pure DP approach [Pathak et al., 2010, Jayaraman et al., 2018]