Privacy

Introduction

Guillaume Raschia — Nantes Université

Last update: January 23, 2025

original slides from J. Near (Univ. of Vermont, CS211: Data Privacy)
and A. Machanavajjhala, M. Hay, X. He; Differential Privacy in the Wild, VLDB'16 & SIGMOD'17 Tutorial

Outline

Administrative

What is data privacy, and how is it violated?

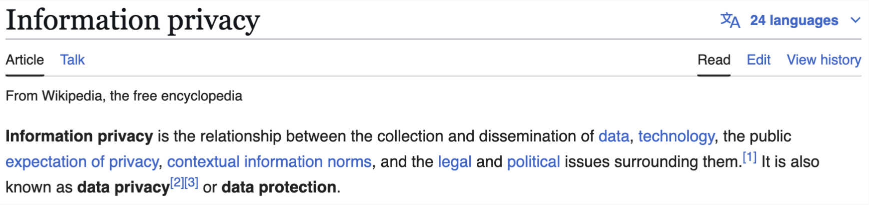
How do data privacy breaches affect us?

Course Information

- Course website on Madoc:
  ipolytech_info5_donneesperso (Données personnelles)
- Instructor: Guillaume Raschia,
  `mailto:guillaume.raschia@univ-nantes.fr`
- Lectures: three 1h15 sessions
- Practice: one 1h30 lab session
- Grading: 1 Homework Assignment (40%) ● 1 Final Exam (60%)

Questions ?

A Non-Definition



Information privacy

From Wikipedia, the free encyclopedia

Information privacy is the relationship between the collection and dissemination of data, technology, the public expectation of privacy, contextual information norms, and the legal and political issues surrounding them.[1] It is also known as **data privacy**[2][3] or **data protection**.

A First Attempt Towards a Definition

Analysis of data preserves data privacy if:

- You learn something useful from the analysis
- The analysis does not violate the privacy of any individual

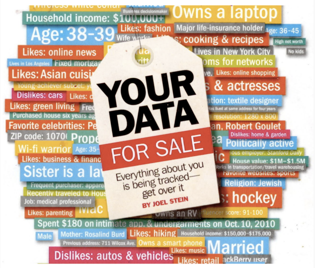
An individual's privacy is violated if:

- The analyst learns something about the individual that s/he did not know before the analysis took place

Danger

This is a very strong statement

Unleash Personal Data!



YOUR DATA
FOR SALE
Everything about you is being tracked—get over it
BY JOEL STEIN
Owns a laptop
Age: 38-39
Household income: $100,000+
Married
hockey
Dislikes: autos & vehicles

Source: TIME

- Personal data is invaluable
  - Advertising, advertising, advertising
  - Genome wide association studies
  - Human mobility analysis
- Personal data is ...personal!
  - age, income, address, likes/dislikes, political orientation, sexual orientation, medical history
  - Redlining, discrimination
  - Foreign interference in democratic processes
  - Physical or financial harm

Personal data is protected

GDPR in Europe (2018), Privacy Act (1974) and HIPAA (1996) in the USA (as well as State laws like CCPA in California, 2020), PIPL in China (2021)

Aside: Privacy is not Security

Data privacy is distinct from data security

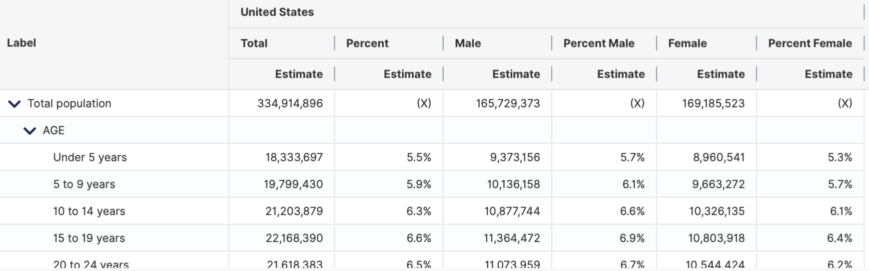
Data security is concerned with who can touch the data:

- **Confidentiality**: ensuring that only the appropriate people can view the data
- **Integrity**: ensuring that only the appropriate people can modify the data

Data privacy is concerned with what can be learned from the data (i.e. its information content)

Example: Census Data

Census protect data privacy via **aggregation**



Label	United States					
	Total	Percent	Male	Percent Male	Female	Percent Female
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Total population	334,914,896	(X)	165,729,373	(X)	169,185,523	(X)
AGE						
Under 5 years	18,333,697	5.5%	9,373,156	5.7%	8,960,541	5.3%
5 to 9 years	19,799,430	5.9%	10,136,158	6.1%	9,663,272	5.7%
10 to 14 years	21,203,879	6.3%	10,877,744	6.6%	10,326,135	6.1%
15 to 19 years	22,168,390	6.6%	11,364,472	6.9%	10,803,918	6.4%
20 to 24 years	21,618,383	6.5%	11,073,959	6.7%	10,544,424	6.2%

S0101 Age and Sex (2023: ACS 1-Year Estimates Subject Tables)

Grouping participants makes it difficult to learn something specific to any individual

Example: Violating Privacy under Aggregation

A company releases the average salary of its employees each year:

Year	Average salary
2023	73 568 €
2024	74 872 €

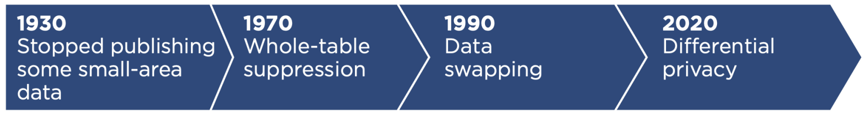
Auxiliary information:

- 58 employees in 2023
- Your friend Bob was hired between the two releases

$$\frac{\sum e_i}{58} = 73\,568 \cdot \frac{\sum e_i + B}{59} = 74\,872$$

Bob's salary: 150 504 €

2020 Census uses Differential Privacy!



1930 Stopped publishing some small-area data

1970 Whole-table suppression

1990 Data swapping

2020 Differential privacy

2017: Announcement [Dajani et al., 2017]

2018: The Disclosure Avoidance System TopDown Algorithm (TDA) [Abowd, 2018].

2021–: Impact and Technical, Social, Ethical Analyses [Dwork, 2019, United States. Bureau of the Census, 2021, Gong et al., 2022, Garfinkel, 2022]

Privacy Violations that Aren't

[Reminder] An individual's privacy is violated if:

- The analyst learns something about the individual that s/he did not know before the analysis took place

Deductive Reasoning

- A study states that coffee drinkers have 100% chance of being mean to pets
- **Auxiliary information**: Joe drinks coffee
- **Conclusion**: Joe is probably mean to his pets

Is this a privacy breach?

Consider:
The breach happens **whether or not Joe participates in the study!**

A Revised Definition

A data analysis violates an individual's privacy if:

- The analyst learns something about the individual that s/he would not have learned if the individual had not participated in the analysis

In other words:

- A privacy-preserving analysis should have the same outcome, **regardless of the participation** of any particular individual

A Series of Disclosures

Identity disclosure or attribute disclosure of individuals

- AOL Web Search Logs
- Netflix Prize
- NYC Taxi Data
- James Comey's X Account
- The Governor of Massachussetts
- GSM Mobility Trace
- (Strava's Heatmap)

AOL Web Search Logs

Michael Barbaro, Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749, The New York Times, 2006.

- 20M de-identified search queries over a 3-month span
- Searches from random user ranging from "*numb fingers*" to "*60 single men*" to "*dog that urinates on everything*"
- **User N° 4417749 is Thelma Arnold**, a 62-year-old widow who lives in Lilburn, Georgia

Ms Arnold commented: "My goodness, it's my whole personal life...I had no idea somebody was looking over my shoulder."

Auxiliary data:

biographical information (dog ownership, location)

Netflix Prize Dataset

Challenge: improve Netflix recommender system

- 100M ratings (1999–2005) over 10M movies from 500 000 users
- **How To Break Anonymity of the Netflix Prize Dataset?** [Narayanan and Shmatikov, 2006]
- 8 ratings over 3-day span: 96% de-anonymized users!

Auxiliary data:

Internet Movie Database (IMDb) ratings

NYC Taxi Data

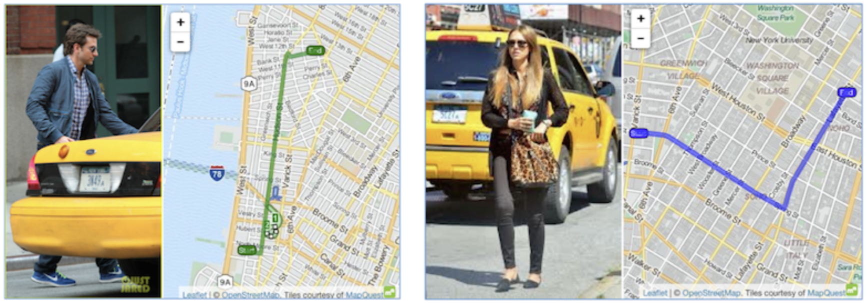
- 20GB, 179M individual trips
- pickup and drop-off location and time, income, anonymized hack license number and medallion number

  ```
  6B111958A39B24140C973B262EA9FEA5,D3B035A03C8A34DA17488129DA581EE7,
  VTS,5,,2013-12-03 15:46:00,2013-12-03 16:47:00,1,3660,22.71,
  -73.813927,40.698135,-74.093307,40.829346
  ```
- 19M possible medallion numbers and 3M license numbers: brute-force MD5 reveals taxi driver's ID

  Vijay Pandurangan. On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs. Medium, June 22, 2014.
- Side concern: where was XXX dropped-off that day? Where does YYY live? Who are those that frequent place ZZZ?

NYC Taxi Data (cont'd)

Tracking Bradley Cooper and Jessica Alba



Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset, September 15, 2014 by ATOCKAR.

NYC Taxi Data (cont'd)

Bradley Cooper's Query

```
SELECT D.dropoff_latitude,
       D.dropoff_longitude,
       F.total_amount,
       F.tip_amount
FROM tripData AS D JOIN tripFare AS F
    ON D.hack_license = F.hack_license AND
       D.pickup_datetime = F.pickup_datetime
WHERE D.pickup_datetime BETWEEN "2013-07-08 19:33:00" AND
                                "2013-07-08 19:37:00"
  AND D.pickup_latitude BETWEEN 40.719 AND 40.7204
  AND D.pickup_longitude BETWEEN -74.0106 AND -74.01
```

Auxiliary data:

geo-tagged celebrity gossip photos

James Comey's X (formerly Twitter) Account

Former FBI director James Comey was behind the account of Reinhold Niebuhr

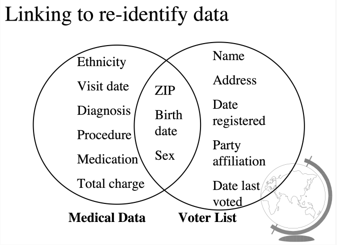
> *Goodbye Iowa. On the road home. Gotta get back to writing. Will try to tweet in useful ways.* pic.twitter.com/DCbu3Yvqt3
>
> *— James Comey (@Comey) October 23, 2017*

This Is Almost Certainly James Comey's Twitter Account, March 30, 2017 by Ashley Feinberg, Gizmodo.

Auxiliary data:

social network (Comey's son)

Latanya Sweeney & Medical Records


Linking to re-identify data

- The Group Insurance Commission (GIC) published 100-field medical records of 135 000 state employees and their families
- William Weld, Governor of Massachussetts, was re-identified [Sweeney, 1997]

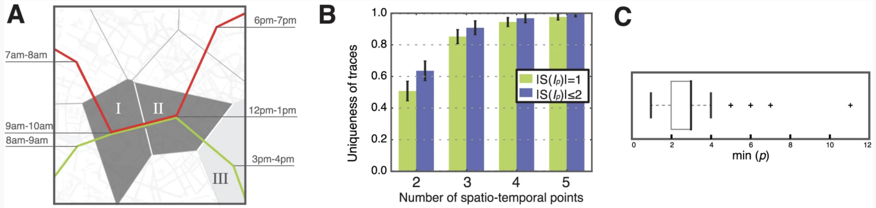
DOB, gender, zip code uniquely identify 87% of people in US

Auxiliary data: Voter rolls

Same technique, used more than a decade later, against the Personal Genome Project [Sweeney et al., 2013] and Health Data as well [Sweeney, 2015].

GSM Mobility Trace

Back in time, 12 points uniquely identify a fingerprint [Locard, 1931]



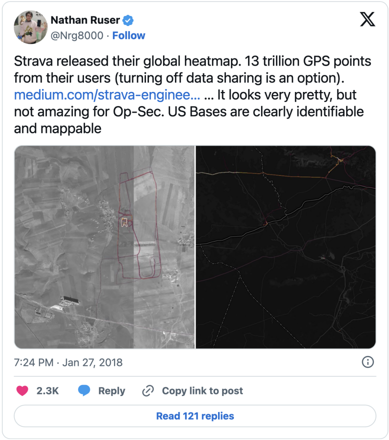
- 15 months recording of spatio-temporal points of 1.5M users
- 114 interactions per user per month for nearly 6 500 antennas

(A) green and red traces are common under I_2 but distinct with I_3

(B) **4 points disclose 95% unique traces** [De Montjoye et al., 2013]

(C) At most 11 points to re-identify all users

Strava's Heatmap


Nathan Ruser
@Nrg8000 · Follow

Strava released their global heatmap. 13 trillion GPS points from their users (turning off data sharing is an option). medium.com/strava-enginee... … It looks very pretty, but not amazing for Op-Sec. US Bases are clearly identifiable and mappable

7:24 PM · Jan 27, 2018

Questions

Were any individuals harmed?

Is this a privacy violation at all?

Lessons Learned

From the above, also partly reviewed in [Ohm, 2010]

- **Aggregation** doesn't necessarily protect individual privacy
- **Anonymization** doesn't necessarily protect individual privacy
- **Large population** don't necessarily protect individual privacy

Also, **machine learning** doesn't necessarily protect individual privacy

- 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 [Shokri et al., 2017][Carlini et al., 2022]

As a rule of thumb, **defining privacy is hard**

Principles for Protecting Privacy

- Privacy threats are counter-intuitive
- We must do something “extra” to ensure privacy
- We should define privacy carefully and precisely
- Challenge: tension between **accuracy and privacy**

References i

- Abowd, J. M. (2018).
  The u.s. census bureau adopts differential privacy.
  In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’18, page 2867, New York, NY, USA. Association for Computing Machinery.
- Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. (2022).
  Membership inference attacks from first principles.
  In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

References ii

- Dajani, A. N., Lauger, A. D., Singer, P. E., Kifer, D., Reiter, J. P., Machanavajjhala, A., Garfinkel, S. L., Dahl, S. A., Graham, M., Karwa, V., et al. (2017).
  The modernization of statistical disclosure limitation at the us census bureau.
  In *September 2017 meeting of the Census Scientific Advisory Committee*.
- 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, 3:1376.

References iii

- Dwork, C. (2019).
  Differential privacy and the us census.
  In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS ’19, page 1, New York, NY, USA. Association for Computing Machinery.
- Garfinkel, S. (2022).
  Differential Privacy and the 2020 US Census.
  MIT Case Studies in Social and Ethical Responsibilities of Computing, (Winter 2022).
  https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census.

References iv

- Gong, R., Groshen, E. L., and Vadhan, S. (2022).
  Harnessing the Known Unknowns: Differential Privacy and the 2020 Census.
  Harvard Data Science Review, (Special Issue 2).
  https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

- Locard, E. (1931).
  ***Traité de Criminalistique*.**
  Desvigne, Lyon.

- Narayanan, A. and Shmatikov, V. (2006).
  How to break anonymity of the netflix prize dataset.
  CoRR, abs/cs/0610105.

References v

- Ohm, P. (2010).
  Broken promises of privacy: Responding to the surprising failure of anonymization.
  UCLA Law Review, 57:1701.

- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017).
  Membership inference attacks against machine learning models.
  In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

References vi

- Sweeney, L. (1997).
  Weaving technology and policy together to maintain confidentiality.
  Journal of Law, Medicine & Ethics, 25(2–3):98–110.

- Sweeney, L. (2015).
  Only you, your doctor, and many others may know.

- Sweeney, L., Abu, A., and Winn, J. (2013).
  Identifying participants in the personal genome project by name.

- United States. Bureau of the Census (2021).
  ***Disclosure Avoidance for the 2020 Census: An Introduction*.**
  Department of Commerce, U.S. Census Bureau.