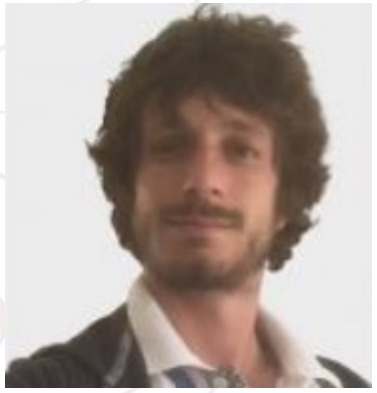
Privacy-Preserving Data Processing

Luca Vassio, Martino Trevisan

Performance 2021

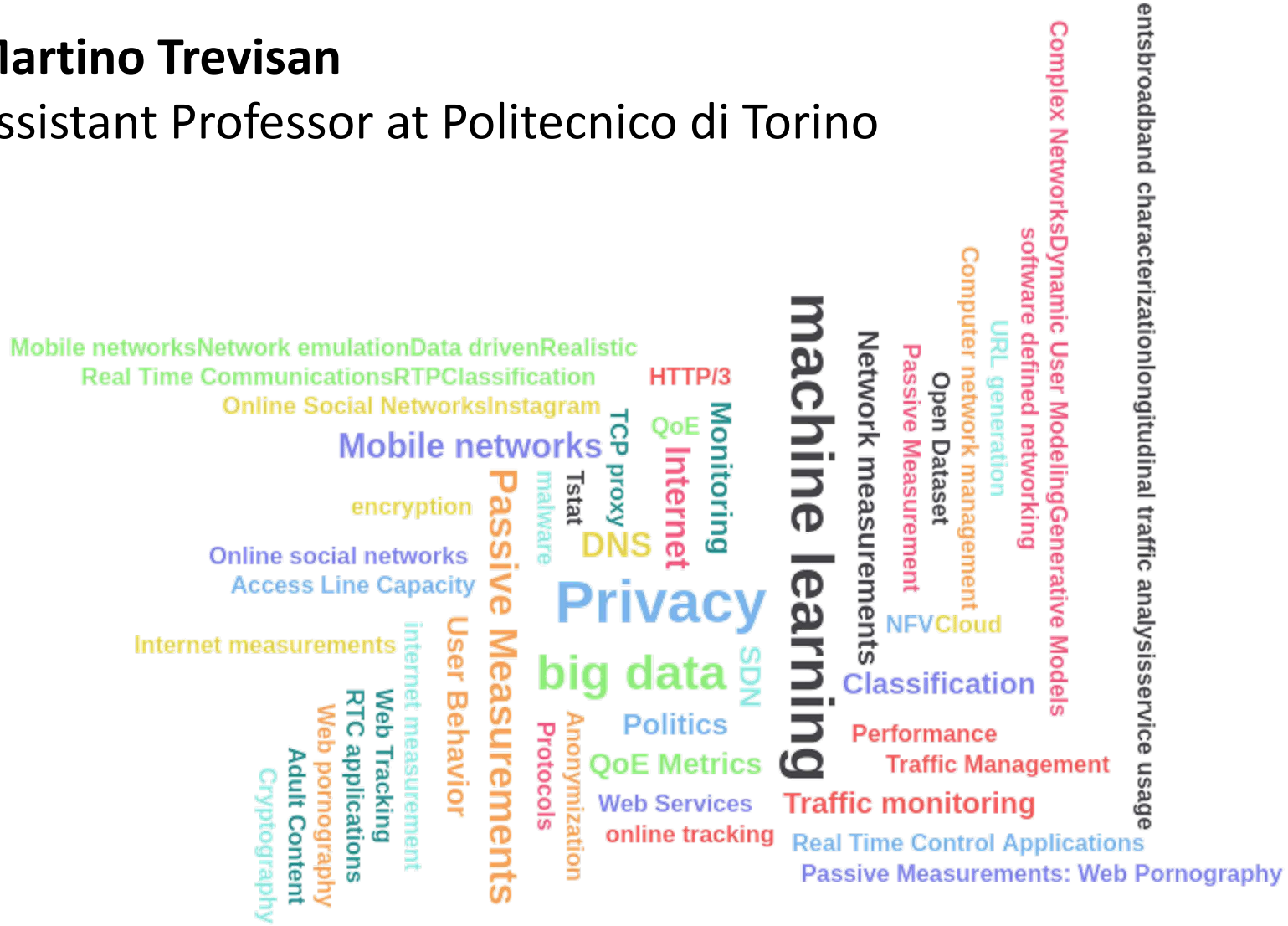
November, 8

The Speakers



Martino Trevisan

Assistant Professor at Politecnico di Torino



The Speakers

Luca Vassio

Assistant Professor at Politecnico di Torino

Goals of this Tutorial

- Understand why Privacy is an issue when processing data regarding people
- Have an overview of the most popular techniques for data anonymization
- Hands on: anonymize a dataset using the state-of-the-art libraries and techniques

Outline

1. The rise of Data-Driven approaches and problematics
2. Why we need Anonymization and why it is difficult
3. Privacy-Preserving Techniques
    - K-anonymity (and variants)
    - Differential Privacy
4. Open-Source tools for anonymization
5. Hands-on on Data

Remember to meet the stars!

The goal of this session is to offer to the PhD students and early career Post-docs attending the Performance 2021 conference (even without presenting a paper) the opportunity to meet each other and get feedback on their work.

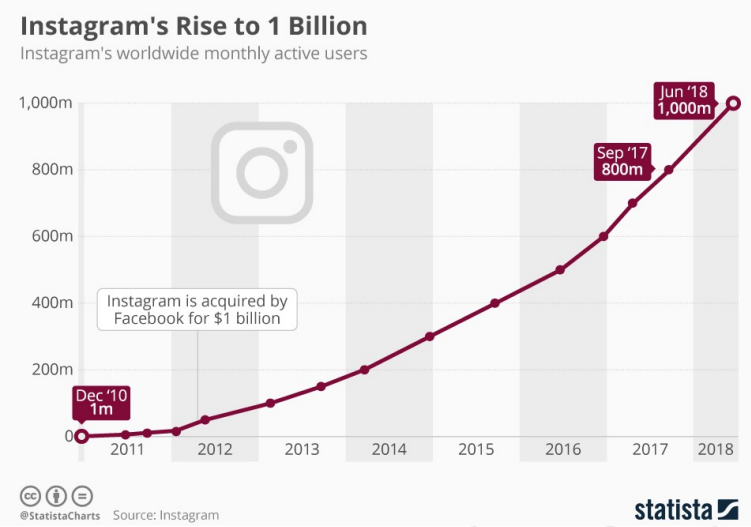
In this session PhD students and Post-docs can meet experienced researchers and get comments on their work.

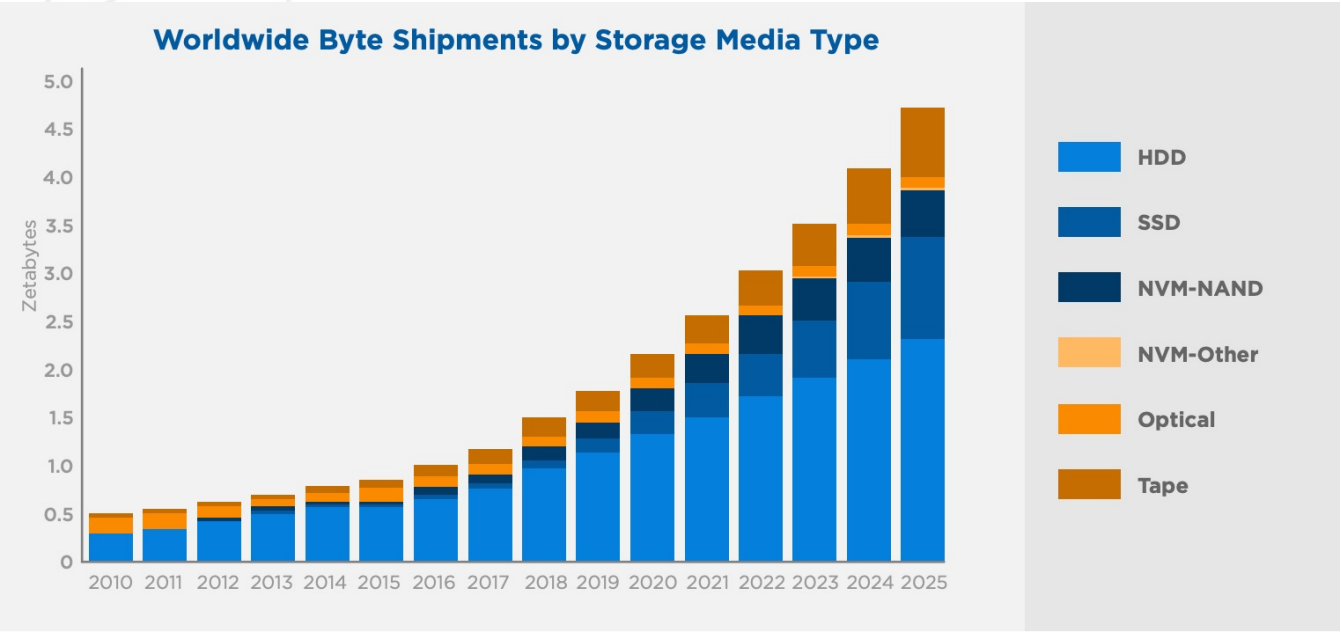
The session is scheduled on Friday November 12 and will last about 1-1.5 hour starting at 11.00 NYC time.
https://www.performance2021.deib.polimi.it/meet-the-star-event/

The rise of Data-Driven approaches and problematics

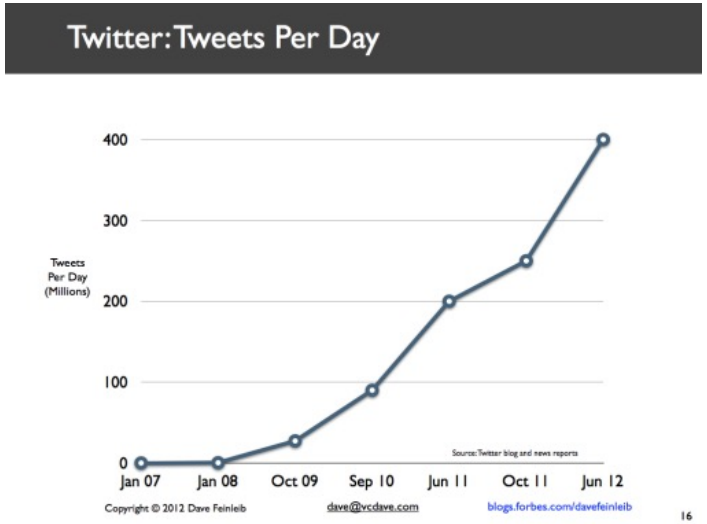
An Unprecedented Amount of Data

- **Continuous Growth**
  - Data volume increases exponentially over time
  - 44x increase from 2009 to 2020
    - Digital data 35 ZB in 2020



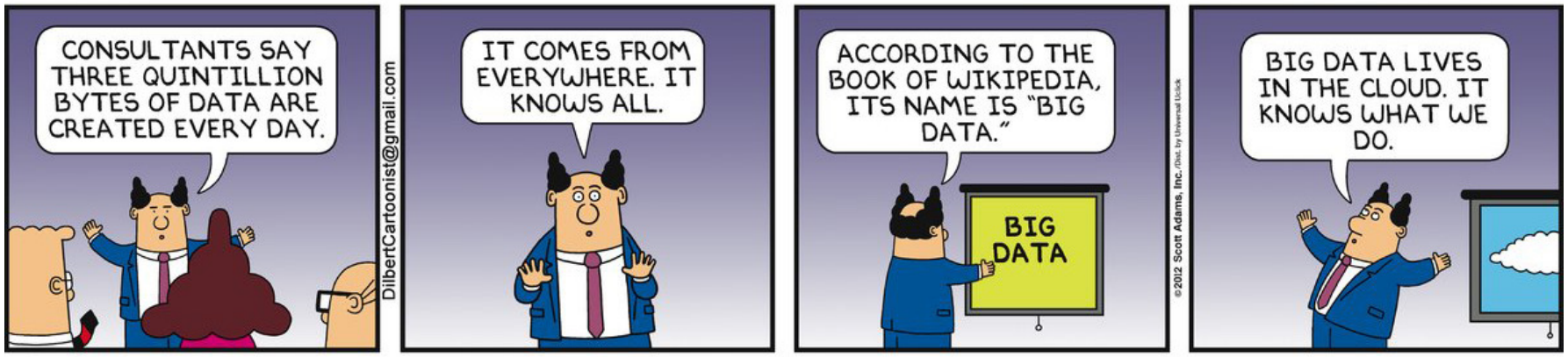


Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018



What are Big data

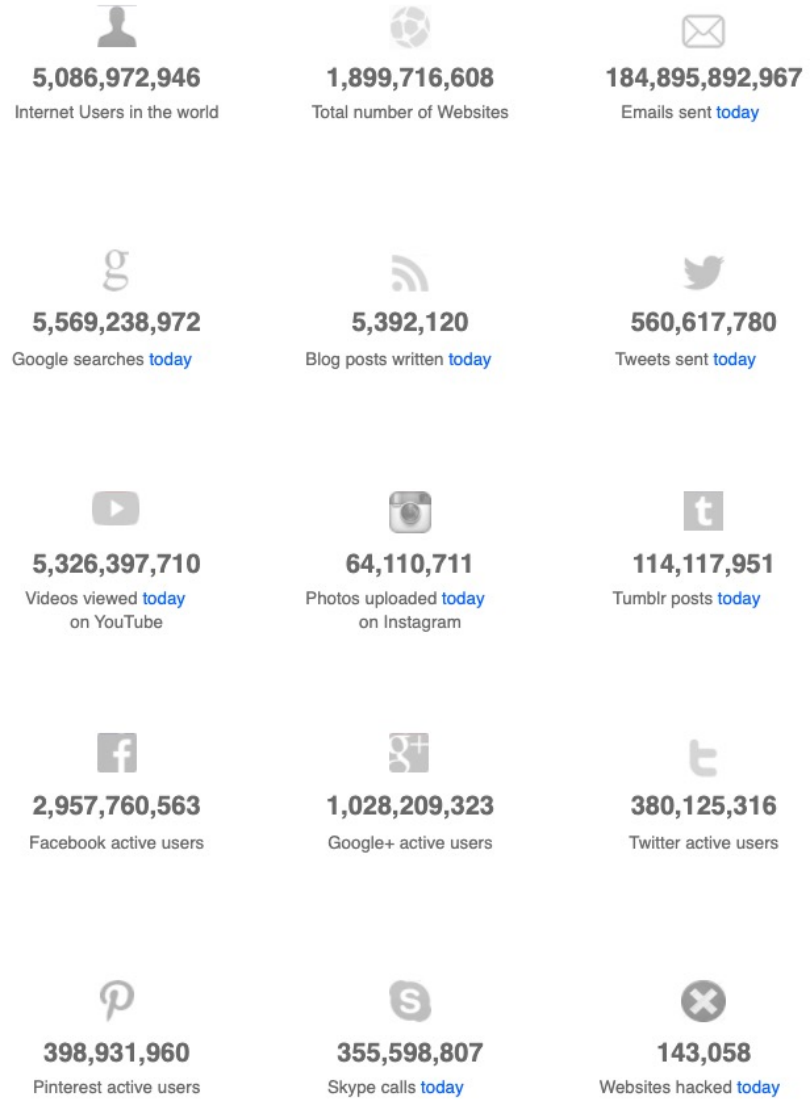
What is big data?



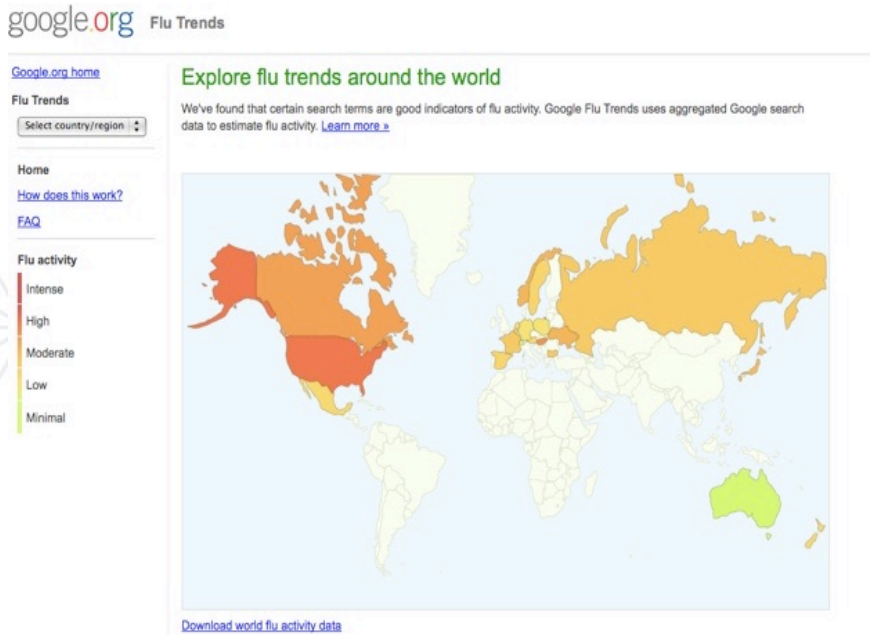
- Many different definitions
  - “Data whose scale, diversity and complexity require new architectures, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it”

Data on the Internet…

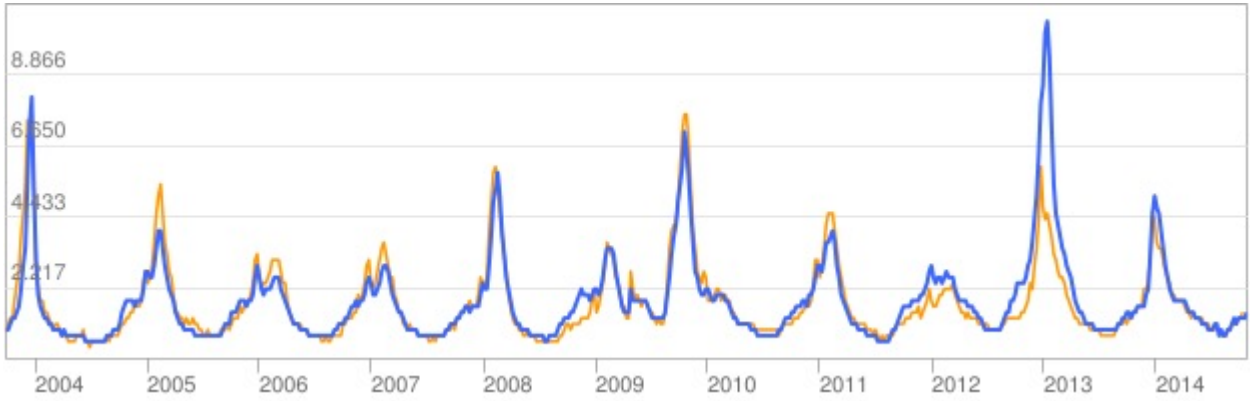
- Internet live stats
  - http://www.internetlivestats.com/



Google Flu trends



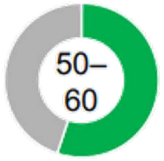
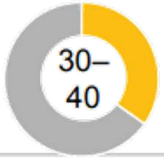
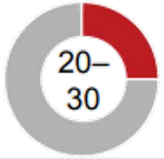
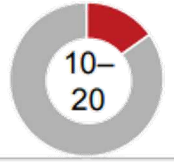
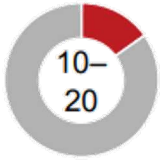
- February 2010
  - Google detected flu outbreak two weeks ahead of CDC data (Centers for Disease Control and Prevention – U.S.A)
  - Based on the analysis of Google search queries



The potential of data-driven economy

- According to McKinsey: «The age of analytics: Competing in a data-driven world»
- *Even under the most conservative estimations, the figures for the potential of data-driven decision making are staggering*
- *Mobility alone could see benefits of close to **2.5 trillion** dollars by 2025, whereas multiple other sectors have annual benefits in the hundreds of billions (e.g., banking with **260 billion** per year).*

The potential of data driven economy

	Potential impact: 2011 research	Value captured %	Major barriers
Location-based data	▪ $100 billion+ revenues for service providers ▪ Up to $700 billion value to end users		▪ Penetration of GPS-enabled smartphones globally
US retail[1]	▪ 60%+ increase in net margin ▪ 0.5–1.0% annual productivity growth		▪ Lack of analytical talent ▪ Siloed data within companies
Manufacturing[2]	▪ Up to 50% lower product development cost ▪ Up to 25% lower operating cost ▪ Up to 30% gross margin increase		▪ Siloed data in legacy IT systems ▪ Leadership skeptical of impact
EU public sector[3]	▪ ~€250 billion value per year ▪ ~0.5% annual productivity growth		▪ Lack of analytical talent ▪ Siloed data within different agencies
US health care	▪ $300 billion value per year ▪ ~0.7% annual productivity growth		▪ Need to demonstrate clinical utility to gain acceptance ▪ Interoperability and data sharing

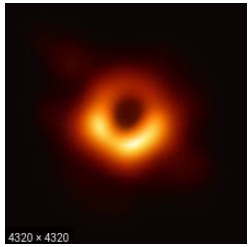
1 Similar observations hold true for the EU retail sector.
2 Manufacturing levers divided by functional application.
3 Similar observations hold true for other high-income country governments.

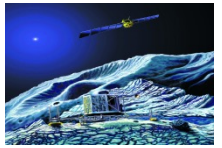
SOURCE: Expert interviews; McKinsey Global Institute analysis

Who generates data?

- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

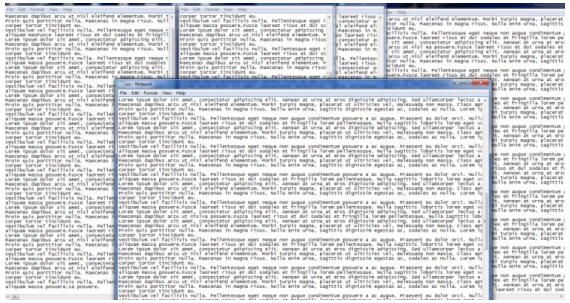
- Health and scientific computing



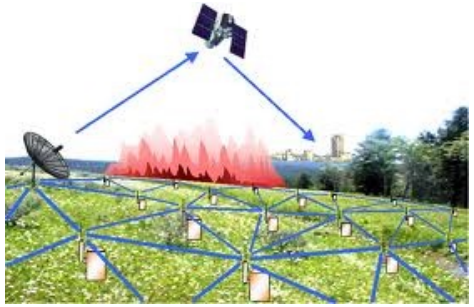


Who generates data?

- Log files
  - Web server log files, machine system log files

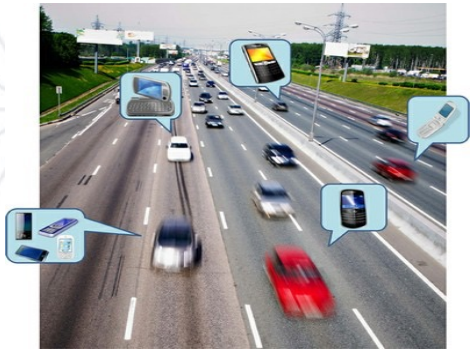


- Internet Of Things (IoT)
  - Sensor networks, RFID, smart meters

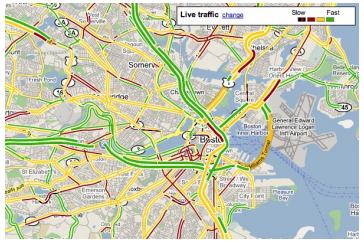


An example of Big data at work

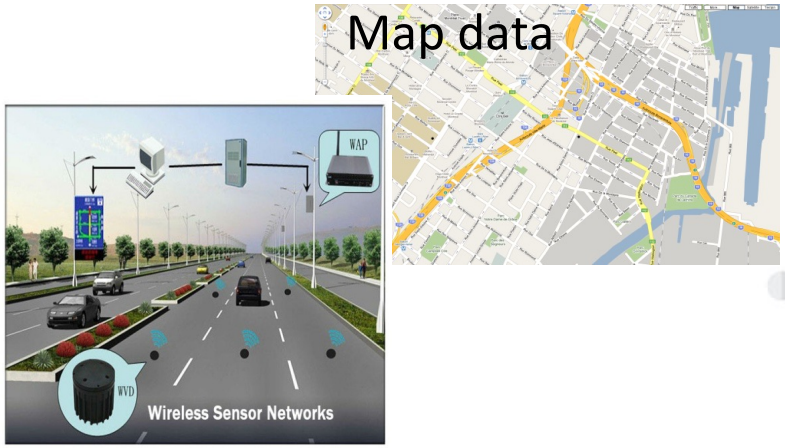
- Crowdsourcing



Computing



Real time traffic info



Sensing

Most of these data are generated by People
This may threaten their Privacy!!!

Who collects data

- Data come from various sources
  - Online shops
  - Search engines
  - Navigation systems and ISPs
  - …
  - **Websites and Ads**

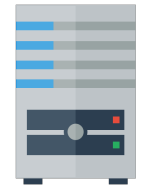
Web tracking

- Collection of information about users' online activity

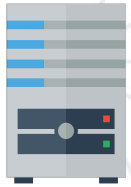
- Fundamental pillar of Web industry
  - Essential to advertisement (Re-targeting, Online Behavioral Advertisement, etc.)

How it works

www.bbc.co.uk

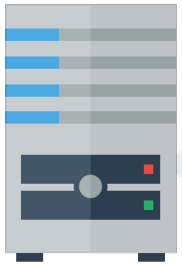


akamaihd.net



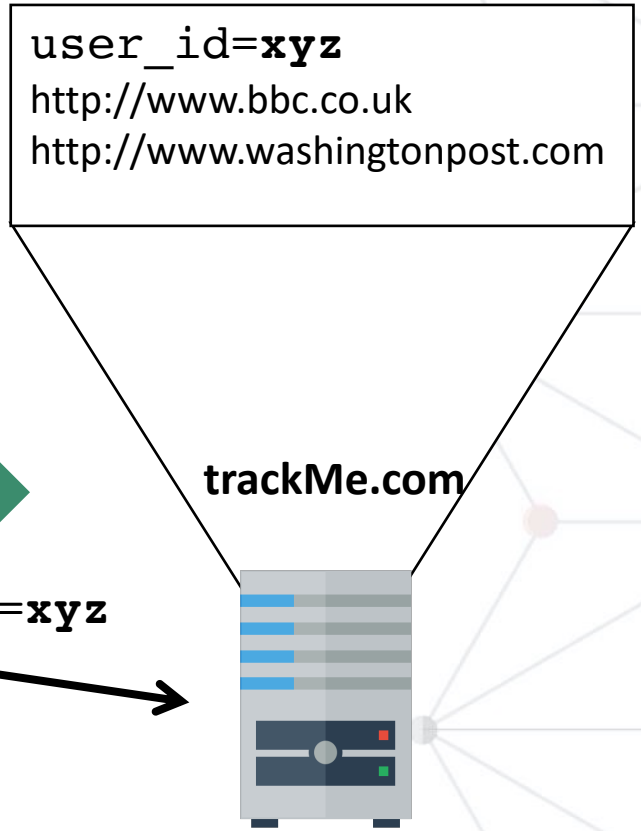
Set-Cookie:
user_id=**xyz**

trackMe.com

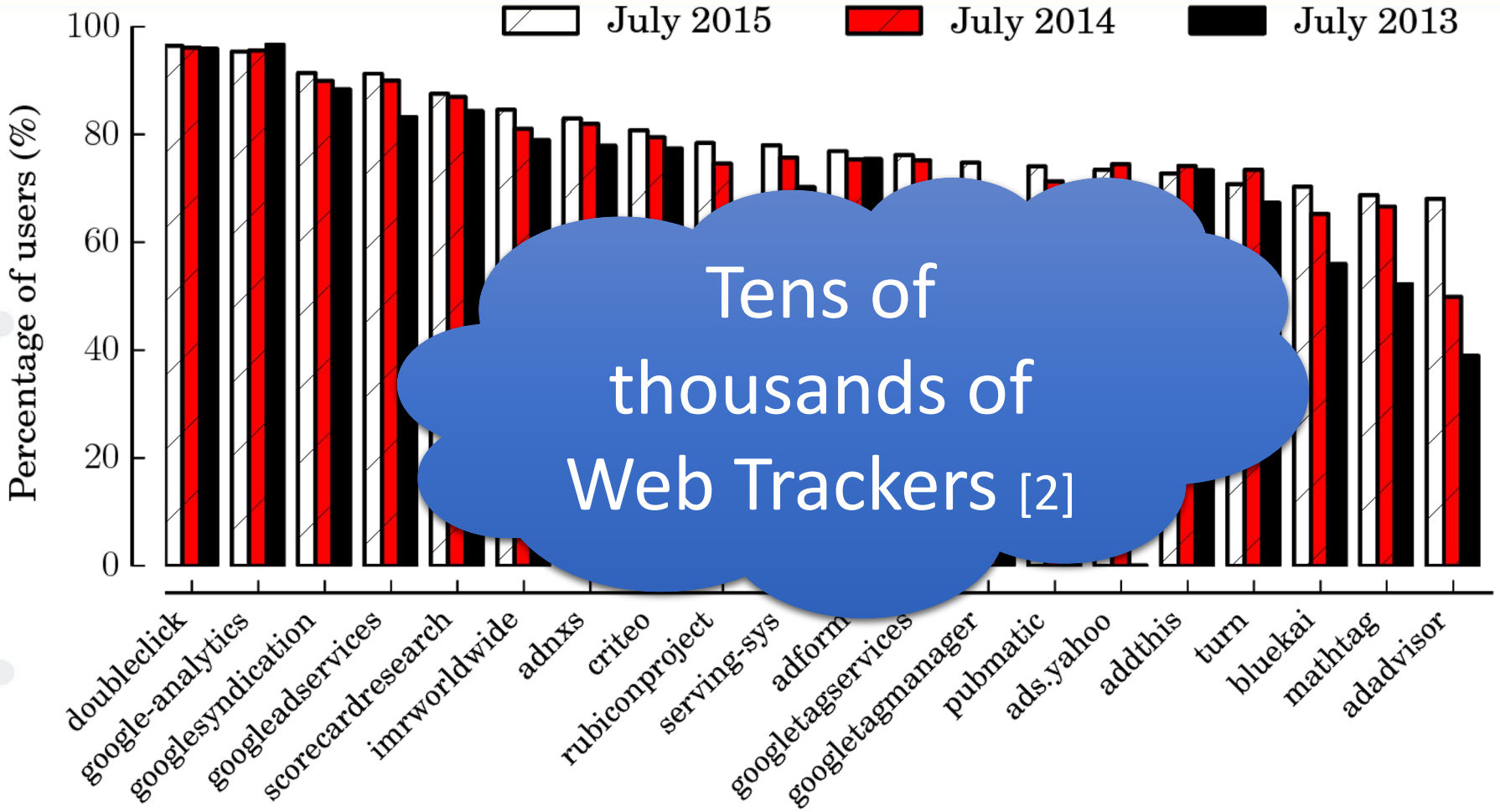


How it works

Cookie:
user_id=**xyz**



How many users are tracked?

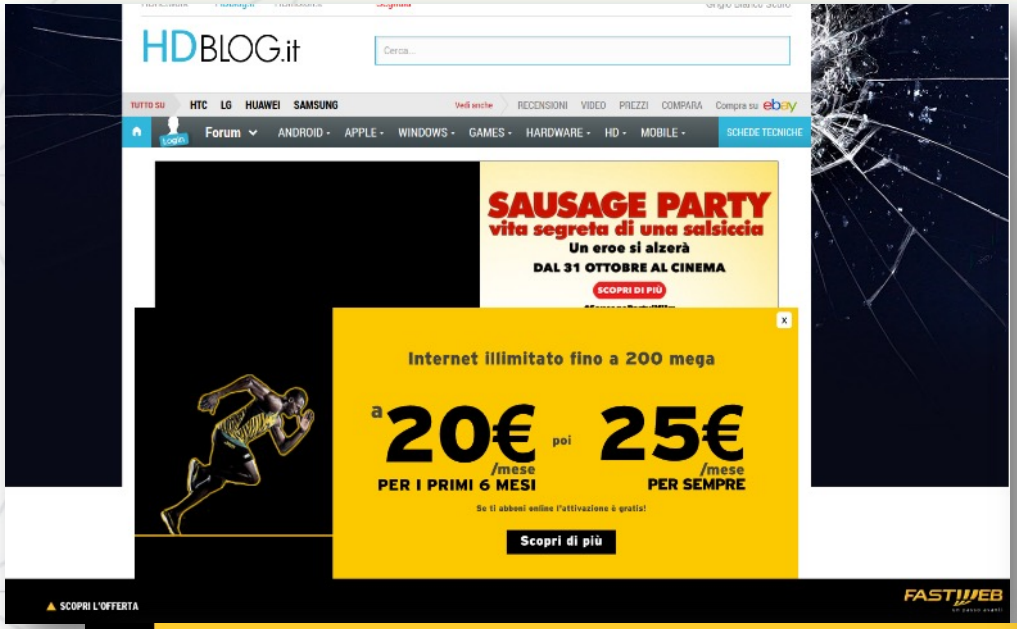


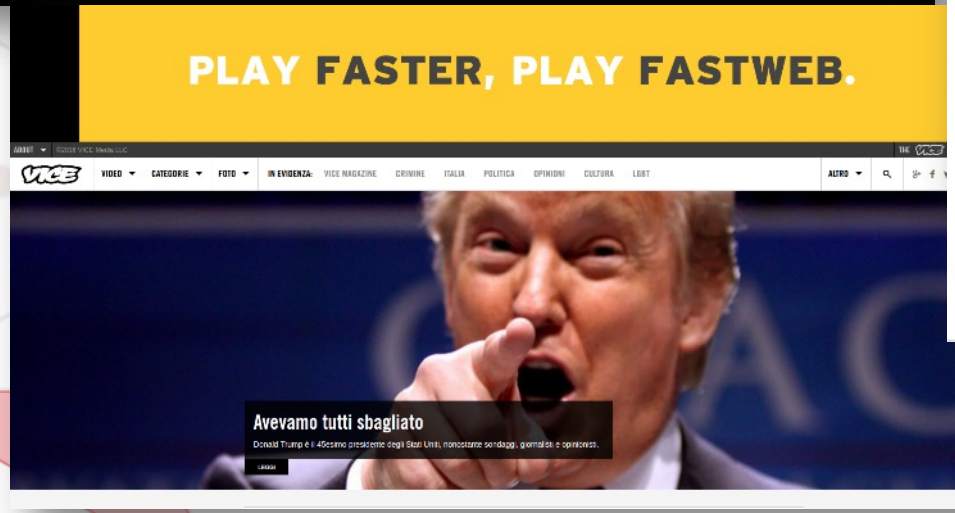
Tens of thousands of Web Trackers [2]

[1] Hassan Metwalley, Stefano Traverso, Marco Mellia, "*Using Passive Measurements to Demystify Online Trackers*", IEEE Computer "Communications and Privacy under Surveillance Issue", 2016

[2] Englehardt, Steven, and Arvind Narayanan. "*Online tracking: A 1-million-site measurement and analysis*", Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security

Tracking is annoying!





Tracking can be dangerous!

Kashmir Hill, Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

+ Follow on Forbes | Subscribe 36k

Forbes

TECH | 2/16/2012 @ 11:02AM | 564,678 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

BUSINESS INSIDER ADVERTISING

The NSA Is Using Google's Advertising Cookies To Track Its Targets

Richard Feloni

Dec. 11, 2013, 11:54 AM 1,917

FACEBOOK LINKEDIN TWITTER EMAIL PRINT

The National Security Agency is using the tracking data intended for Google's advertisers to locate its targets.

According to the Washington Post's new analysis of an internal presentation Edward Snowden leaked earlier in the year, the NSA has been using the numeric identifiers in Google's "PREF" cookies for hacking.

A cookie is a small piece of data sent from a website that is stored in a

NSA Director General Keith Alexander maintains that his agency's tactics are lawful and used against foreign adversaries.

CNN Money International + Markets Economy Companies Tech Autos India Video stock tickers

Data brokers selling lists of rape victims, AIDS patients

by Melanie Hicken @melhicken

December 19, 2013: 12:38 PM ET

Recommend 1.6K

Social Surge - What's Trendin

Ikea gives employee paid pare

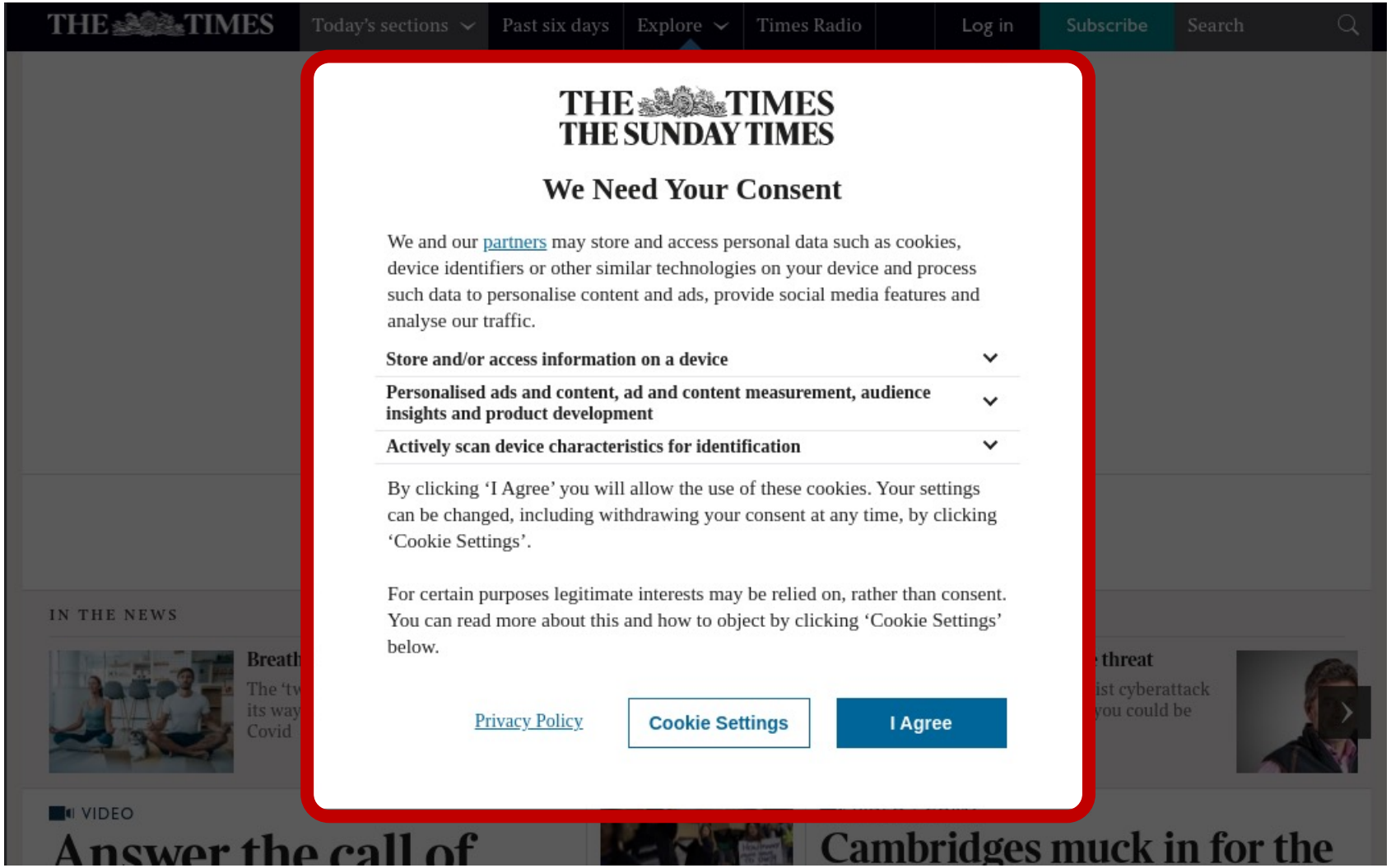
U.S. wom team will champion protest

Newly cor Yellen is a the U.S. e

Laws on Privacy

- Legislators promulgated laws to defend users' privacy
- **Cookie Law** or Directive 2009/136/EC requires:
- Websites to ask *"prior informed consent for storage or for access to information stored on a user's terminal equipment"*
- In other words, a website must ***explicitly*** ask the visitor to authorize the storage and retrieval of data sent through ***cookies*** and similar tracking mechanisms ***before*** delivering and installing them.
- Its implementation has become evident to end-users because of the presence of a "*Cookie Bar*" on most of websites

According to EU, website should ask consent!



General Data Protection Regulation

Regulation 2016/6791 – Entered into Force in May 2018

- Goal: protect users' privacy, and punish violations
- It applies to data of EU citizen, worldwide

What are Personal Data?

Personal data are any information which are related to an identified or identifiable natural person

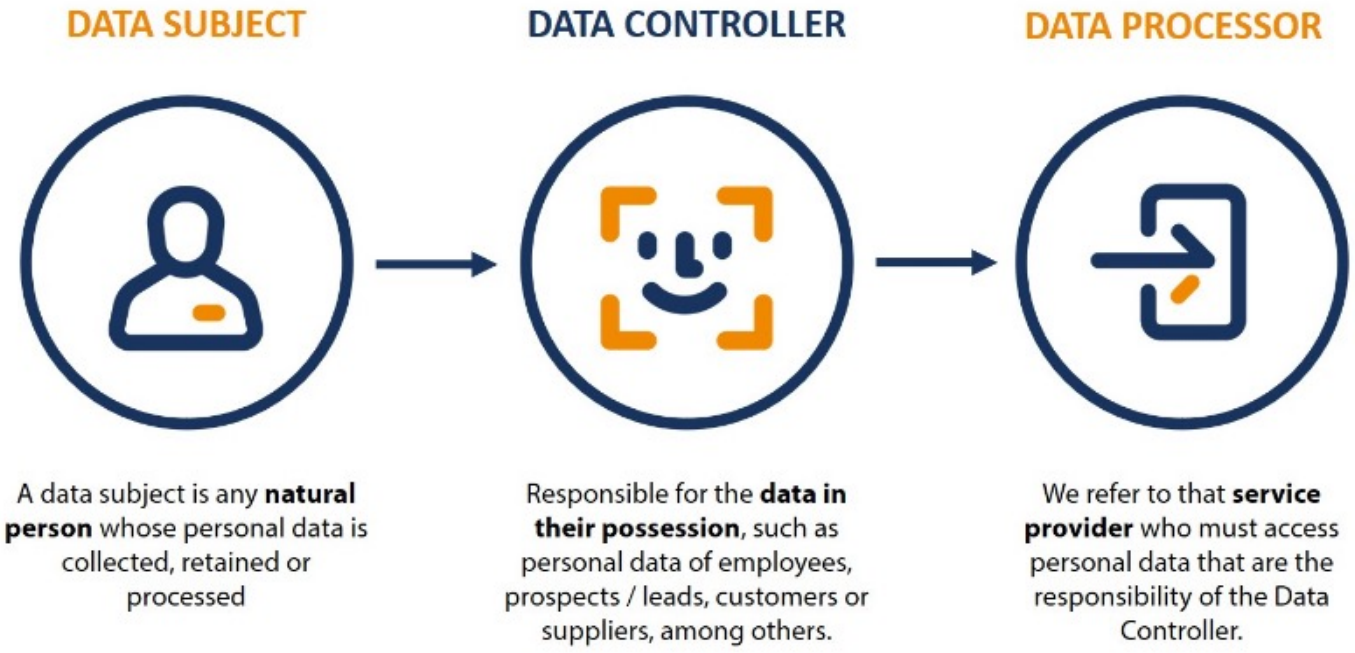
- Can be identifiers (e.g., name, phone number)
- Or special characteristics (e.g., genetic, mental, cultural info)

General Data Protection Regulation

Principles:

1. **Accountability**: be responsible and adopt a safe behavior and document it
2. **Privacy by design:** assess the privacy impact **before** data processing
3. **Privacy By Default:** Cannot accumulate data without an objective
    1. Can accumulate only the data strictly needed for the goal
    2. Companies must always obtain consent of users
4. **Transparency**: simple, clear and complete information
    1. Users must be informed give the **consent freely**
    2. Users should have access to data and be able to "**download**" it and delete it
5. Must inform users and authorities of **data breaches** within **72 hours**

Roles in the GDPR



PIA and DPIA

- Privacy Impact Assessment (**PIA**) is all about analyzing how an entity collects, uses, shares, and maintains personally identifiable information, related to existing risks.
  - used to protect privacy **by design** when an organization starts or acquires a new business, implements a new process, or launches a new product
- Data Protection Impact Assessment (**DPIA**) is all about identifying and minimizing risks associated with the processing of personal data.
  - **regularly applied** to personal data processing, identifying, and mitigating risks.

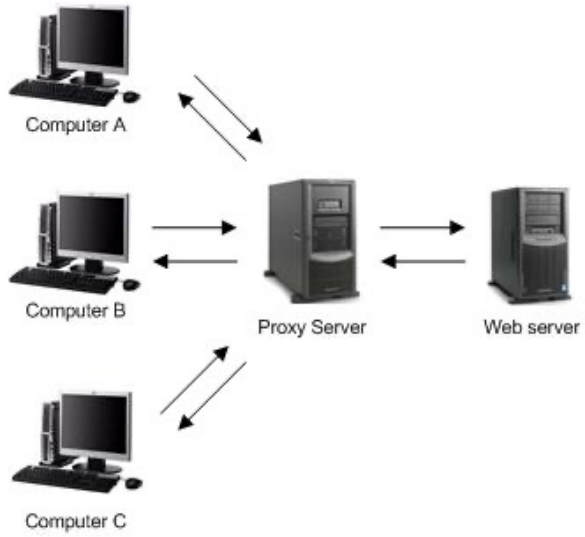
Pitfalls in anonymization

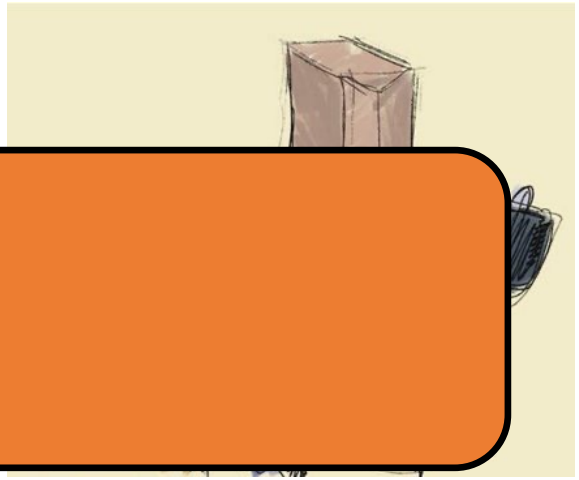
Data Publishing and Lessons Learned

- Companies and public bodies have published data
  - For research, societal benefits challenges
- They involuntarily disclosed personal information
  - They tried to remove sensitive information
  - But this is harder than it appears

Anonymization

- Hide identity, remove identifying info
- **Proxy server / Network data:** remove source IP address
- **Health data** released for research purposes: remove name, address, etc





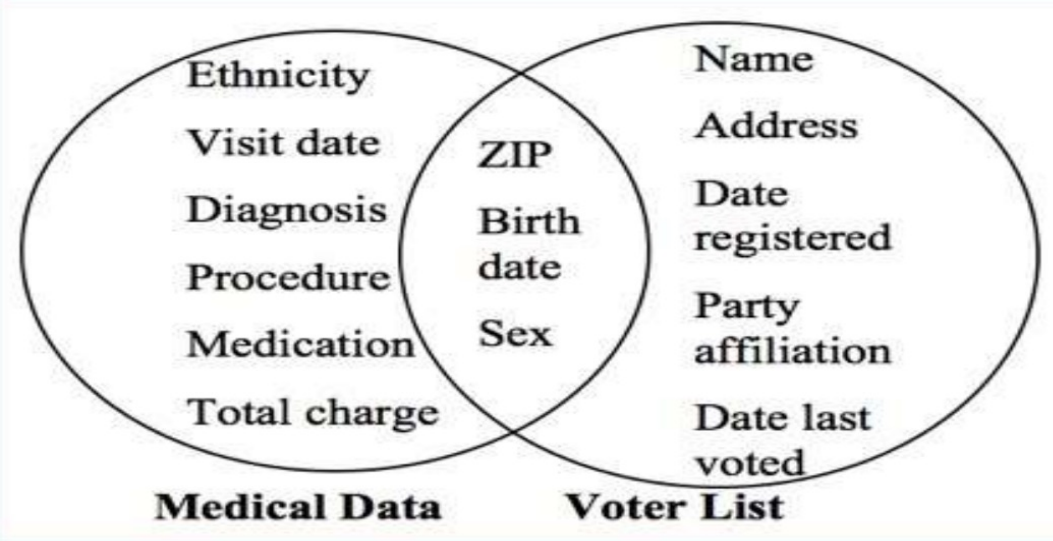
Is this enough?

Popular cases of Data De-Anonymization

- In 1997 the **Massachusetts Group Insurance Commission** decided to release "anonymized" data on state employees that showed every single hospital visit.
- The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number.

Popular cases of Data De-Anonymization

- **Latanya Sweeney**, a computer science researcher purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter.
- By combining data, Sweeney found Governor Weld with ease.
  - Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code.



born July 31, 1945
resident of 02138

Popular cases of Data De-Anonymization

- This demonstrates that removing IDs is not enough for privacy
- Latanya Sweeney in 2000 in “Uniqueness of Simple Demographics in the U.S. Population” stated that
  - *87% of the US population can be uniquely identified by gender, ZIP code and full date of birth.*
- Philippe Golle in 2006 in *Revisiting the Uniqueness of Simple Demographics in the US Population*
  - Recompute this number to 63% of individuals
- Paul Ohm in 2009 in *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*
  - *“Easy reidentification represents a sea change not only in technology but in our understanding of privacy.”*

Popular cases of Data De-Anonymization

- In 2006, AOL researchers released a massive dataset of search queries
- They first "anonymized" the data by scrubbing user IDs and IP addresses.
- User IDs were scrubbed but were replaced with a number that uniquely identified each user.
  - Good for research ☺
  - An individual could be identified and matched to their account and search history by such information ☹

Popular cases of Data De-Anonymization

98280	how does a male's cocaine use affect a fetus	2006-04-10 1
98280	how does a male's cocaine use affect a fetus	2006-04-10 5
98280	birth defects caused by father's cocaine use	2006-04-10 1
98280	birth defects caused by father's cocaine use	2006-04-10 4
98280	are chainletter scams ever successful	2006-04-10 0

Although the searchers were only identified by a numeric ID, some people's search results have become notable for various reasons.

Erik S. Lesser for The New York Times

Thelma Arnold

- “numb fingers”
- “60 single men”
- “dog that urinates on everything.”

Popular cases of Data De-Anonymization

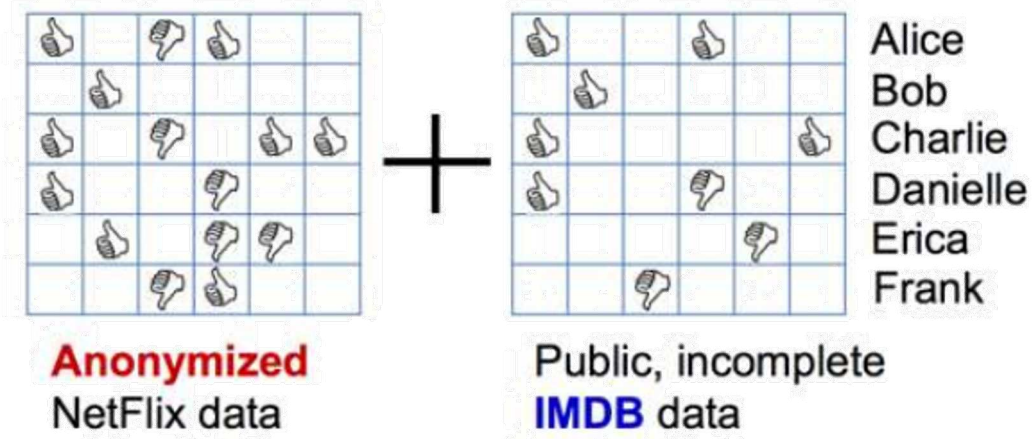
- Netflix Prize dataset, released 2006
  - 100,000,000 (private) ratings from 500,000 users
  - 10% of their users
  - average 200 ratings per user
- Competition to improve recommendations
  - i.e., if user X likes movies A,B,C, will also like D
- Anonymized: usernames replaced by a number

Popular cases of Data De-Anonymization

- Problem: can combine “private” ratings from Netflix with public reviews from IMDB to identify users in dataset
- May expose private info about members…

Popular cases of Data De-Anonymization

Use Public Reviews from IMDB.com



Credit: Arvind Narayanan via Adam Smith

Narayanan, Shmatikov, Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset), 2008

Popular cases of Data De-Anonymization

- Lesson: cannot always anonymize data simply by removing identifiers
- Vulnerable to aggregating data from multiple sources/networks
- Paul Ohm says: *"For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her."*
  - In "Broken promises of privacy: Responding to the surprising failure of anonymization"

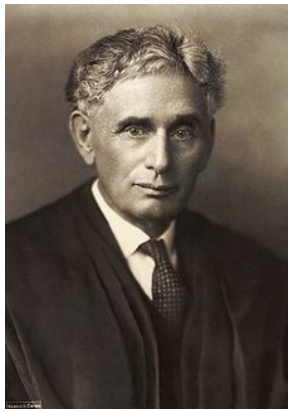
Popular cases of Data De-Anonymization

Other attacks:

- Su et al, De-anonymizing Web Browsing Data with Social Networks , 2017
  - Using links appearing in one's feed is unique
- Korolova, "Privacy Violations Using Microtargeted Ads: A Case Study", PADM
  - Attackers can instrument ad campaigns to identify individual users.
- Calandrino, Kilzer, Narayanan, Felten, Shmatikov, "You Might Also Like: Privacy Risks of Collaborative Filtering"
  - Attackers can infer customers' transaction with a limited amount of auxiliary data

What do we mean by privacy?

- Samuel Warren and Louis Brandeis (1890)
  - "Right to be left alone"
  - In *The Right to Privacy*



- Alan Westin (1967)
  - "Right to control, edit, manage, and delete information about themselves and decide when, how, and to what extent information is communicated to others"

Credits

- Nikhil Jha – Politecnico di Torino
- Eduardo Cuervo - Oculos
- Amre Shakimov - Duke University
- Frank McSherry - Materialize, Inc.
- Krishnaram Kenthapadi - AI @ LinkedIn
- Ilya Mironov - Google AI
- Abhradeep Thakurta - UC Santa Cruz


Perguntas
Fragen
Domande
Galdera
Otázky
Questions
Spørgsmål
Pertanyaan
kysymykset
Frågor
Spørsmål
Cwestiynau
вопросы
Preguntes
Sorular
Въпроси
Vragen
Pytania

Privacy-Preserving Data Processing

Luca Vassio, Martino Trevisan

November 8th, 2021

Outline

1. The rise of Data-Driven approaches and problematics
2. Why we need Anonymization and why it is difficult
3. Privacy-Preserving Techniques
   - k-Anonymity
   - Differential Privacy
4. Open-Source tools for anonymization
5. Hands-on on Data

k-Anonymity

And other cluster-based methods

Classification of attributes

Key attributes

- Anything that identifies the person directly - personally identifying information (PII)
- Name, address, phone number, email
- Always removed before release

Quasi-identifiers

- Not unique, but sufficiently well correlated with a person such that they can be combined with other quasi-identifiers to create a unique identifier
- (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
- Can be used for linking anonymized dataset with other datasets

Classification of attributes

Sensitive attributes

- Medical records, salaries, etc.
- These attributes is what the researchers need, so they are always released directly

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Re-identification by linking

[Latanya Sweeney, 1997]

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Public voter dataset

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

k-Anonymity

- The information for each person contained in the released table **cannot be distinguished from at least k-1 individuals** whose information also appears in the release
  - Example: you try to identify a person in the released table, and the information you have is his birth date, ZIP and gender. However there are k people in the table with the same combination of birth, ZIP and gender
- Any combination of quasi-identifier present in the released table must appear in at least k records

First defined:

[Samarati and Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", 1998]

k-Anonymity

[Samarati et al, 1998]

- **Generalize, modify, or distort quasi-identifier** values so that no individual is uniquely identifiable from a group of k
- In SQL, released table T is k-anonymous if each

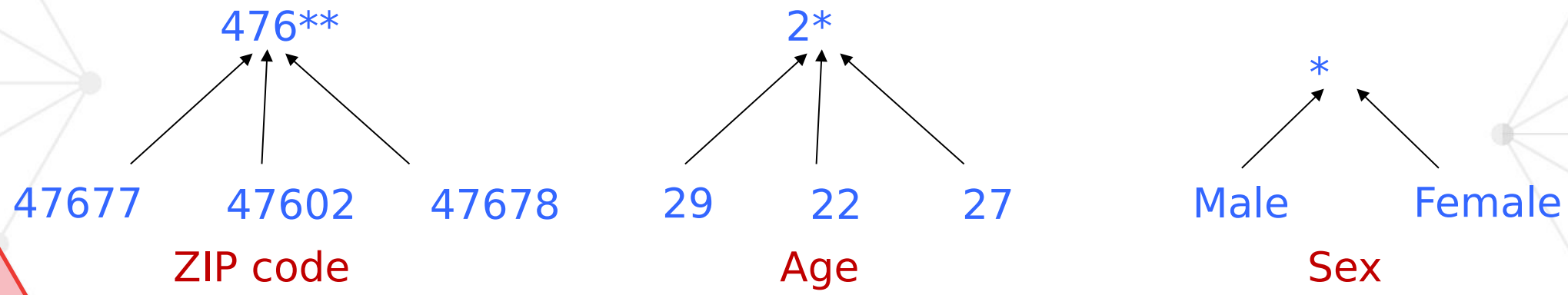
```
SELECT COUNT(*)
FROM T
GROUP BY quasi-identifier combination
```

is **≥ k**

- Parameter k indicates the "degree" of anonymity

Achieving k-Anonymity: Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values



Example of generalization

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Example of generalization

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

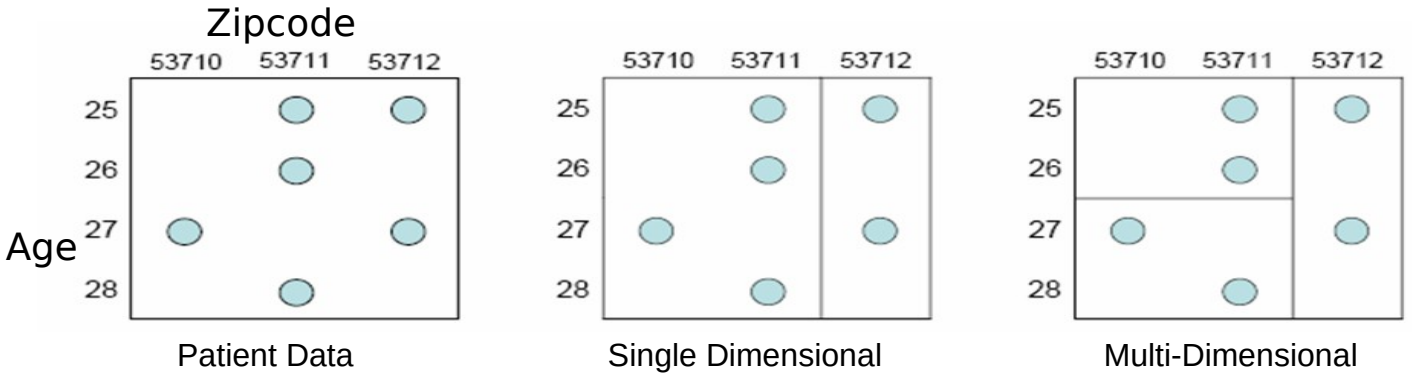
External data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don’t learn Andre’s problem

Generalization algorithms

- There are tens of k-anonymization algorithms
- Example: *Greedy Partitioning Algorithm, Bucketization, ...*
- **Problem:** find multi-dimensional partitions of the data, where each partition has two or more data points (i.e. k=2)



- Optimal k-anonymous strict multi-dimensional partitioning is NP-hard
- Optimal = minimum information loss = maximum utility

Achieving k-Anonymity

Generalization

- Replace specific quasi-identifiers with less specific values

Suppression

- When generalization causes too much information loss
  - This is common with “outliers”

Determining the value of k

The value of k depends on

- Number of records in the table
- Number of quasi-identifiers
- The distribution of each quasi-identifier
- The relationship between quasi-identifier

Rule of thumb:

- k increases → privacy increases
- k increases → utility decreases

What is the best value of k?

2-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-26]	Male	53711	Flu
3	[25-26]	Male	53711	Brochities
2	[25-27]	Female	53712	Hepatitis
5	[25-27]	Female	53712	HIV
4	[27-28]	Male	[53710-53711]	Broken Arm
6	[27-28]	Male	[53710-53711]	Hang Nail

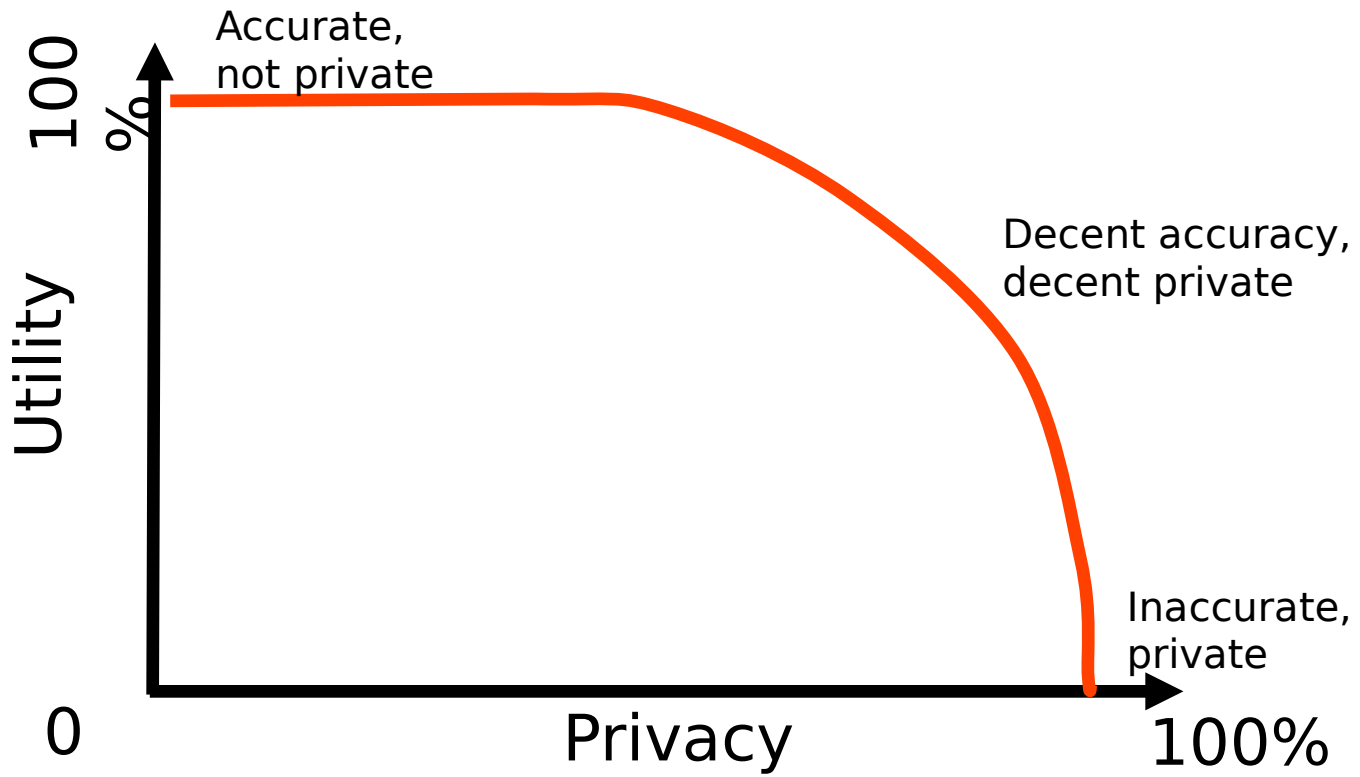
3-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-26]	*	5371*	Flu
3	[25-26]	*	5371*	Brochities
2	[25-26]	*	5371*	Hepatitis
5	[27-28]	*	5371*	HIV
4	[27-28]	*	5371*	Broken Arm
6	[27-28]	*	5371*	Hang Nail

4-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-28]	*	5371*	Flu
3	[25-28]	*	5371*	Brochities
2	[25-28]	*	5371*	Hepatitis
5	[25-28]	*	5371*	HIV
4	[25-28]	*	5371*	Broken Arm
6	[25-28]	*	5371*	Hang Nail

Utility vs. Privacy



Curse of dimensionality

[Aggarwal, VLDB '05]

Generalization fundamentally relies on spatial locality

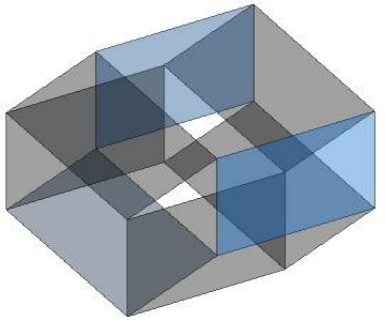
- Each record must have k close neighbors

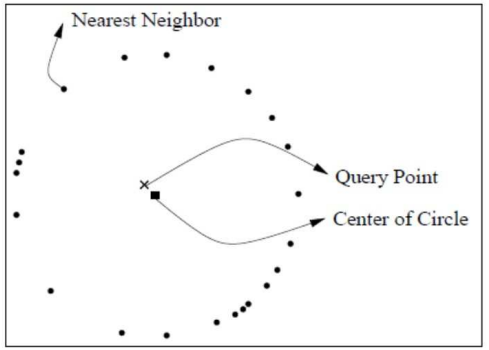
Real-world datasets are very sparse

- Many attributes (dimensions)
  - Netflix Prize dataset: 17000 dimensions
  - Amazon customer records: several million dimensions
- "Nearest neighbor" is very far

Projection to low dimensions loses all info

- k-anonymized datasets results useless





Attacks on k-Anonymity

k-Anonymity does not provide privacy if:

- Sensitive values in an equivalence class lack diversity
- The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

Carl does not have heart disease

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

A solution: l-Diversity

[Machanavajjhala et al., ICDE '06]

Principle

- Each equi-class contains at least l well-represented sensitive values

Zipcode	Age	Disease
476**	2*	Acne
476**	2*	Heart disease
476**	2*	Flu
476**	2*	Heart disease
476**	2*	Flu
476**	2*	Flu
476**	3*	Flu
476**	3*	Acne
476**	3*	Cancer
476**	3*	Acne
476**	3*	Flu
476**	3*	Cancer

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

l-Diversity

Distinct l-Diversity

- Each equivalence class contains l distinct sensitive values

Probabilistic l-Diversity

- The frequency of the most frequent value in an equivalence class is bounded by 1/l

Entropy l-Diversity

- The entropy of the distribution of sensitive values in each equivalence class is at least log(l)

…

Neither sufficient, nor necessary

Original dataset

…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-

99% are HIV-

Anonymization A

Q1	HIV+
Q1	HIV-
Q1	HIV+
Q1	HIV+
Q1	HIV-
Q1	HIV-
Q2	HIV-
Q2	HIV+

Anonymization B

Q1	HIV-
Q1	HIV-
Q1	HIV-
Q1	HIV+
Q1	HIV-
Q1	HIV-
Q2	HIV-
Q2	HIV+

99% HIV- quasi-identifier group is not “diverse”
…yet anonymized database does not leak anything

50% HIV+ quasi-identifier group is “diverse”
This leaks a ton of information

Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Skewness attack

- ~ probabilistic inference attacks
- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

l-diversity does not consider overall distribution of sensitive values!

Sensitive attribute disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

Further privacy: t-Closeness

k-anonymity prevents identity disclosure but not attribute disclosure

- To solve that problem **l-Diversity** requires that each eq. class has at least l values for each sensitive attribute
- **t-Closeness** requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table

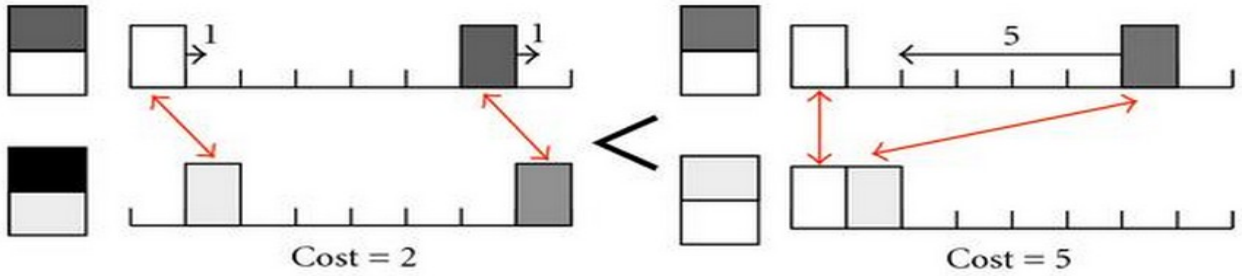
t-Closeness protects against attribute disclosure but not identity disclosure

t-Closeness

Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (within a distance t)

- Earth Mover Distance to capture semantic relationship among sensitive attribute values
- Privacy is measured by the information gain of an observer regarding distribution of sensitive attributes

Information Gain = Posterior Belief – Prior Belief



[Li, Li, Venkatasubramanian, ICDE'2007]

Similarity attack example

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

Trick question: Why publish quasi-identifiers at all??

k-Anonymous, l-Diverse t-Close dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous,
l-diverse and t-close…

It is secure, right?

What does an attacker know?

Bob is Caucasian and I heard he was admitted to hospital with flu…

This is against the rules!
"flu" is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
[illegible]	787XX	HIV+	Shingles
[illegible]	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

k-Anonymity can be harmful

- Focuses on data transformation, not on what can be learned from the anonymized dataset
  - “k-anonymous” dataset can leak sensitive information
- “Quasi-identifier” fallacy
  - Assumes a priori that attacker will not know certain information about the target
- Relies on locality
  - Destroys utility of many real-world datasets

Differential Privacy

A rigorous framework for privacy-preserving analysis of datasets

Why is anonymization hard?

It’s hard to guess what capabilities attackers will have, especially in the future

- Future datasets
- Future techniques
- Future computational power

Analogy with crypto: Cryptosystems today are designed based on what quantum computers might be able to do in 30 years

Why Differential Privacy?

Strong, quantifiable, composable mathematical privacy **guarantee**

It is (by design) resilient to known and unknown attack modes!

DP enables many computations with personal data while preserving personal privacy

First defined:

[Dwork, McSherry, Nissim, and Smith, Calibrating Noise to Sensitivity in Private Data Analysis, in Third Theory of Cryptography Conference, TCC 2006.]

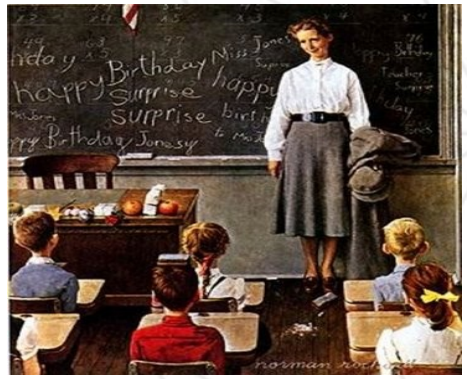
Earlier roots:

[*Warner*, Randomized Response 1965]

Learn from datasets

A research dataset shows that smoking causes cancer

- Smoker S's insurance premiums rise
- However, this is true **even if S not in database!**



Learning that smoking causes cancer is the whole point

- Smoker S enrolls in a smoking cessation program…

Differential privacy: no harm in **participation**

- Outcome of any analysis is essentially equally likely, **independent of whether any individual joins, or refrains from joining, the dataset**

Differential Privacy

- **Worst-case scenario**: an attacker that knows everything, but not if the user is in the dataset
  - The attacker knows the value of every record, **even the target user's one**
- The attacker can perform **queries** to the dataset
  - The answer of the dataset should not hint the presence of the user

"**Blatant non-privacy**" [Dinur and Nissim, 2003]

Overly accurate answers to too many questions destroys privacy

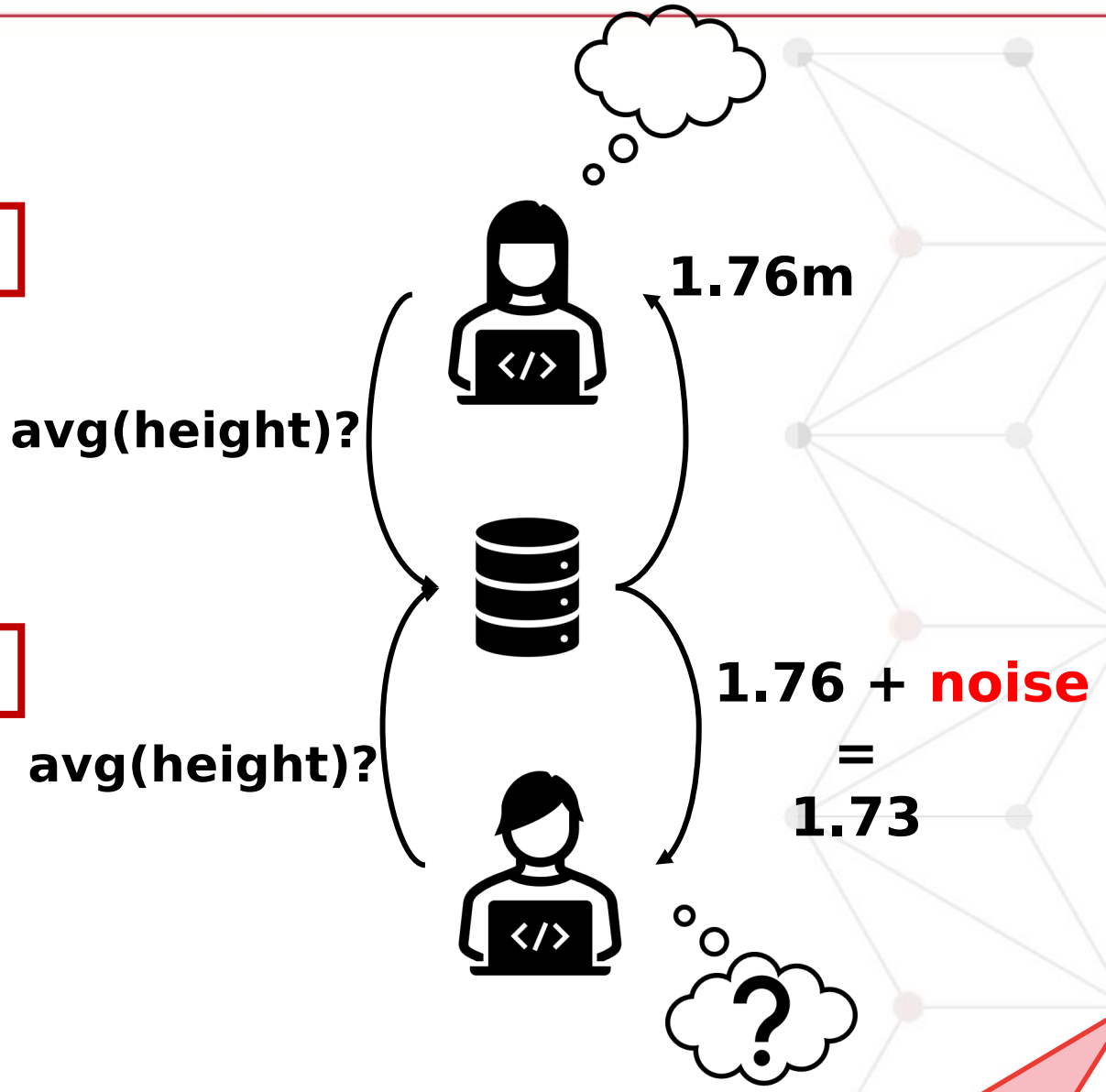
Example: is Luca in the dataset?

Name	Height [m]
Amir	1.65
Esteban	1.73
Xin	1.76

avg(height) → 1.71m

Name	Height [m]
Amir	1.65
Esteban	1.73
Xin	1.76
Luca	**1.90**

avg(height) → 1.76m

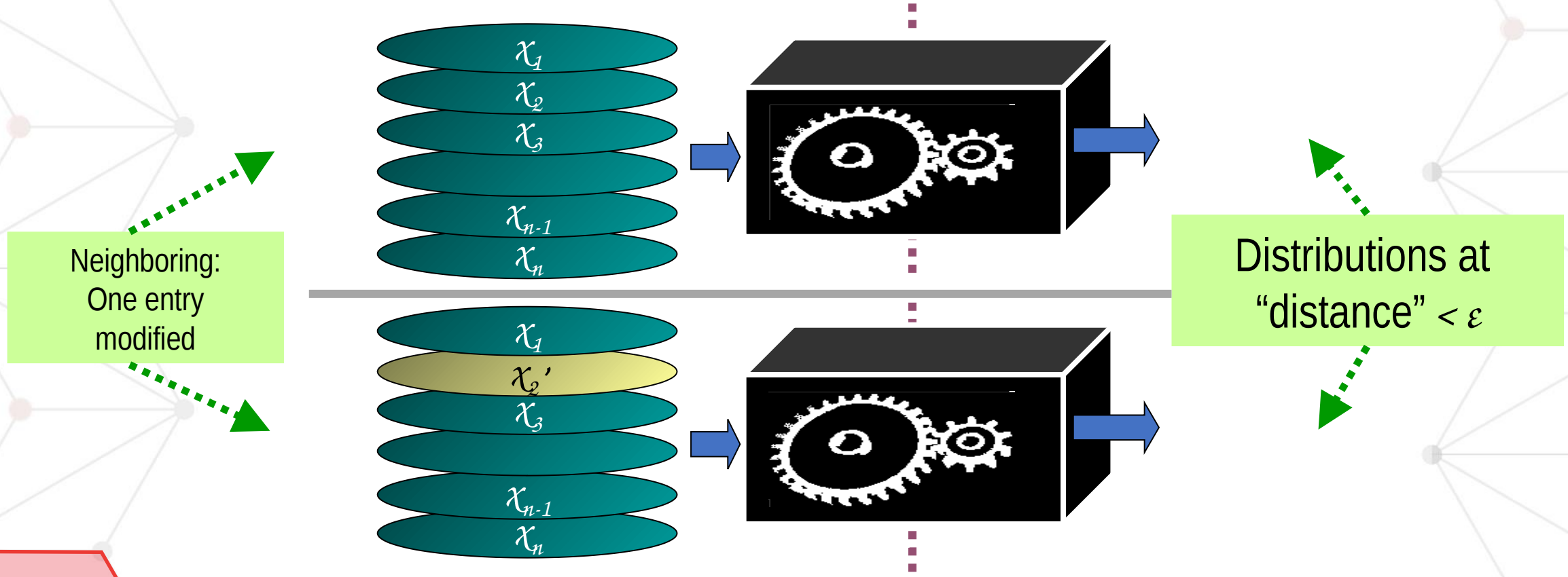


Differential privacy basic setup

- There is a database D which potentially contains sensitive information about individuals
- The **database curator** has access to the full database. We assume the curator is trusted
- The **data analyst** wants to analyze the data. She asks a series of **queries** to the curator, and the curator provides a response to each query
- The way in which the curator responds to queries is called the **mechanism**
- Two databases D and D’ are **neighbouring** if they agree except for a single entry

Differential Privacy

[Dwork et al., 2006]

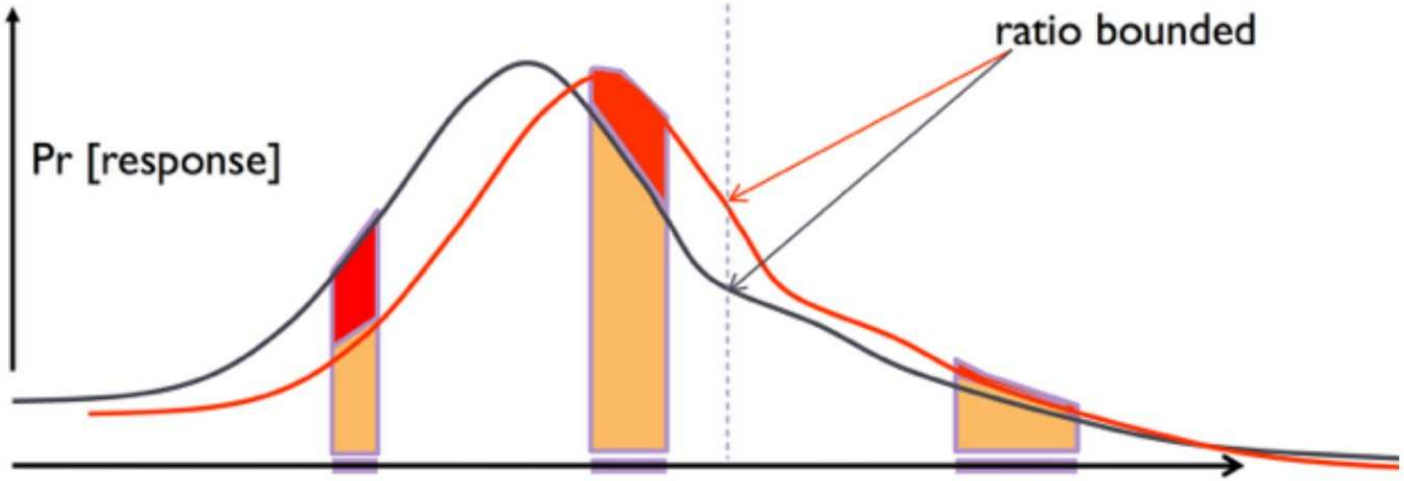


Formally...

[Dwork et al. 2006]

A query mechanism M is ϵ-differentially private if, for any two adjacent databases D and D' (differing in just one entry) and $C \subseteq range(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon} \cdot Pr(M(D') \in C)$$



- Output does **not overly depend on any single tuple**
- Participation in the dataset poses **no additional risk**

Privacy parameter ε

A query mechanism M is ϵ-differentially private if, for any two adjacent databases D and D' (differing in just one entry) and $C \subseteq range(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon} \cdot Pr(M(D') \in C)$$

ε is an arbitrary parameter and controls the privacy of the system

- Low ε, the two quantities are forced to be similar
- High ε, the two quantities are allowed to diverge

However, all queries are not the same

- Some of them may be more intrusive
- What differentiate them is ***sensitivity***

The sensitivity

- A query can be more or less intrusive on the privacy of the users
- Each query f has a **sensitivity**
  - Depends on the **difference** in output between f(D) and f(D')
  - **Global sensitivity** $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$
- The added noise depends on the query sensitivity
  - Sensitive query → more noise
  - Insensitive query → less noise

Implications

Global sensitivity: $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$

- Example: Query f is the average for sets of numbers between 0 and 1

  $GS_{average} = 1/n$

Some queries, such as counting queries, can be answered relatively accurately

Since one tuple affects the result by at most 1 (GS=1)

A small amount of noise can be added to achieve DP

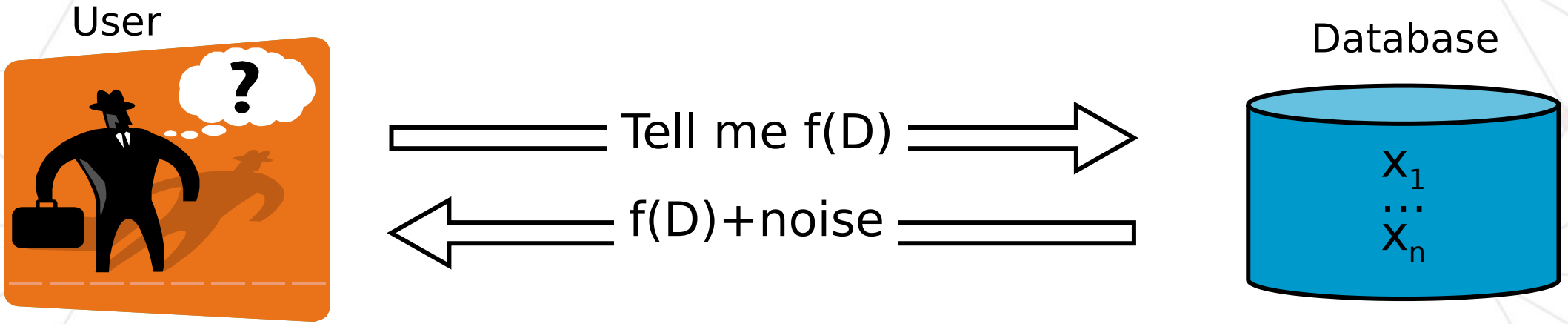
Some queries are hard to answer

E.g., max, since it can be greatly affected by a single tuple (GS unbounded)

Challenge in using it

Find suitable queries to ask so that noisy answers provide most utility

A DP mechanism: output perturbation



Intuition: f(D) can be released accurately when f is insensitive to individual entries x_1, … x_n

We want f(D) + noise to be ε-indistinguishable

How this noise should be generated depending on ε and GS_f ?

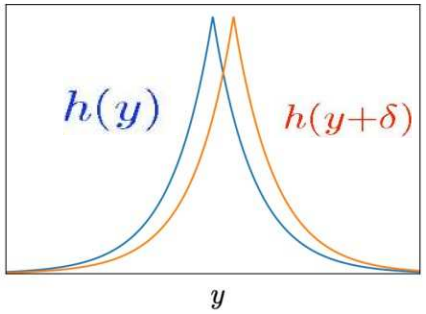
Laplace noise

Theorem

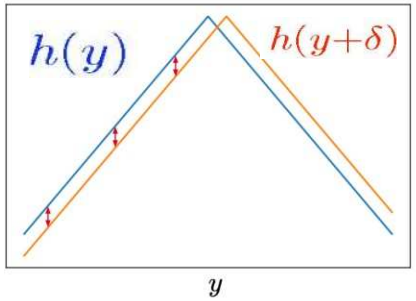
If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$

h(y)



log(h(y))



Laplace noise

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$

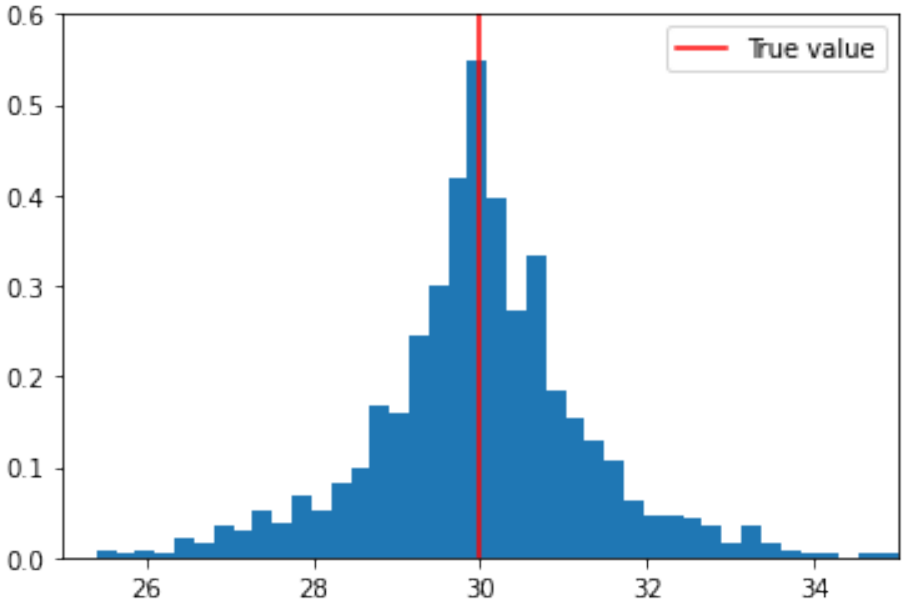
It quantifies:

Noise to add when function is more sensitive (higher GS_f)

Noise to add if we want more privacy (lower ε)

Multiple queries

- What if we allow people to perform the same query over and over again?
- Eventually, the noise will cancel out and the true value will emerge



Properties of Differential Privacy

Group privacy

- Group privacy: εk privacy for a group of size k

Composability

- Applying the sanitization several time yields a graceful degradation
- If A_1 satisfies ε_1-DP, and A_2 satisfies ε_2-DP, then outputting both A_1 and A_2 satisfies $(\varepsilon_1+\varepsilon_2)$-DP
- even just proportional to square root of number of applications

Robustness to side information

- No need to specify **exactly** what the adversary knows
- Any **post-processing cannot improve** the attacker's knowledge

Privacy budget

- To avoid noise cancelling, ε becomes a ***privacy budget*** **to query** the dataset
- The privacy consumption is <u>additive</u>
  - With budget 10, I can choose
    - 10x queries with ε =1
    - 5x queries with ε=2
    - 1x query with ε=10

How to increase number of queries?

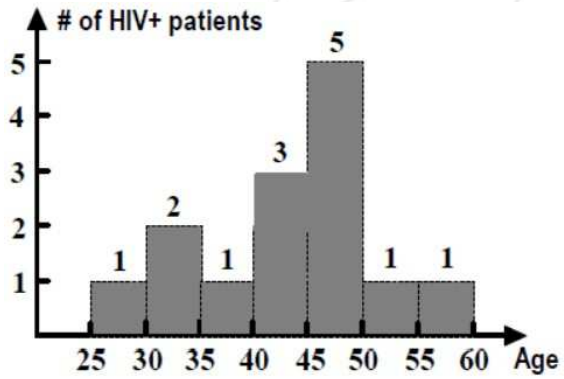
[Blum-Ligget-Roth, 2008]

Use **Coordinated Noise**

If noise is added in with careful coordination, **rather than independently,** DP can save on the budget

Example: histogram queries with d bins:

- You can treat query independently: use d times the mechanism, hence noise Lap(d/ε)
- But actually only need Lap(1/ε), since sensitivity generalizes as max l_1 distance



A glimpse inside the rabbit hole

- What if we don’t trust the data curator?
  - **Local** differential privacy, where data are anonymized before being uploaded to a platform
- What if we want to publish the dataset at once, without having to continuously answer to the queries?
  - **Non-interactive** differential privacy
- What if we have categorical values?
  - Other mechanisms exist: the **exponential mechanism** allows to choose the most-suited value inside a categorical set
- What if we assume the attacker does not know everything?
  - **Relaxed differential privacy**: only statistical knowledge of D and D’, keeping precise knowledge on the element changing

DP: Pros & Cons

Pros 

- **Rigorous** mathematical definition of privacy
- **Flexible**: several mechanisms are available
- **Robust** to postprocessing
- The level of privacy ε can be **chosen** by the system administrator

Cons 

- **Precision** of the queries is affected
- Hard to **explain**
- Many non-trivial DP algorithms require really **large datasets** to be practically useful
- What ε and what privacy budget is **reasonable** for a dataset?

Do I need DP if I don’t care about privacy?

- Almost any usage of the data that is not carefully crafted will **leak something about it**
- Statistics to generalize well should **not be dependent on single instances**
- As for machine learning: it should **not overfit** on training
- We can use DP to **ensure statistical validity of exploratory data**

Better Privacy = Better Data

Credits

Nikhil Jha – Politecnico di Torino
Ashwin Machanavajjhala - Duke University
Vitaly Shmatikov - Cornell Tech
Chris Clifton – Purdue
Muhamad Felemban – KFUPM
Moni Naor- Weizmann Institute of Science
Adam Smith - Penn State
Roger Grosse - University of Toronto
Katrina Ligett - California Institute of Technology
Cynthia Dwork - Harvard University
Matteo Maffei – TU Wien


Savage Chickens
by Doug Savage
I'M TIRED OF THE ANONYMITY OF LIVING IN THE CITY
WHERE DO YOU LIVE?
AROUND
©2006 BY DOUG SAVAGE
www.savagechickens.com


Perguntas
Fragen
Domande
Galdera
Otázky
Questions
Spørgsmål
Pertanyaan
kysymykset
Frågor
Spørsmål
Cwestiynau
вопросы
Preguntes
Sorular
Въпроси
Vragen
Pytania

Privacy-Preserving Data Processing

Luca Vassio, Martino Trevisan

Performance 2021

November, 8

Outline

1. The rise of Data-Driven approaches and problematics
2. Why we need Anonymization and why it is difficult
3. Privacy-Preserving Techniques
    - K-anonymity
    - Differential Privacy
4. Open-Source tools for anonymization
5. Hands-on on Data

Useful Tools

1. Arx
2. DiffPrivLib
3. Google Differential Privacy
4. P-PPA
5. [CNIL]

ARX Data Anonymization Tool

ARX is a comprehensive open-source software for anonymizing sensitive personal data.
It supports a wide variety of

- privacy (and risk) models
- methods for transforming data
- methods for analyzing the usefulness of output data

Has a graphical interface and APIs
Can read data in:

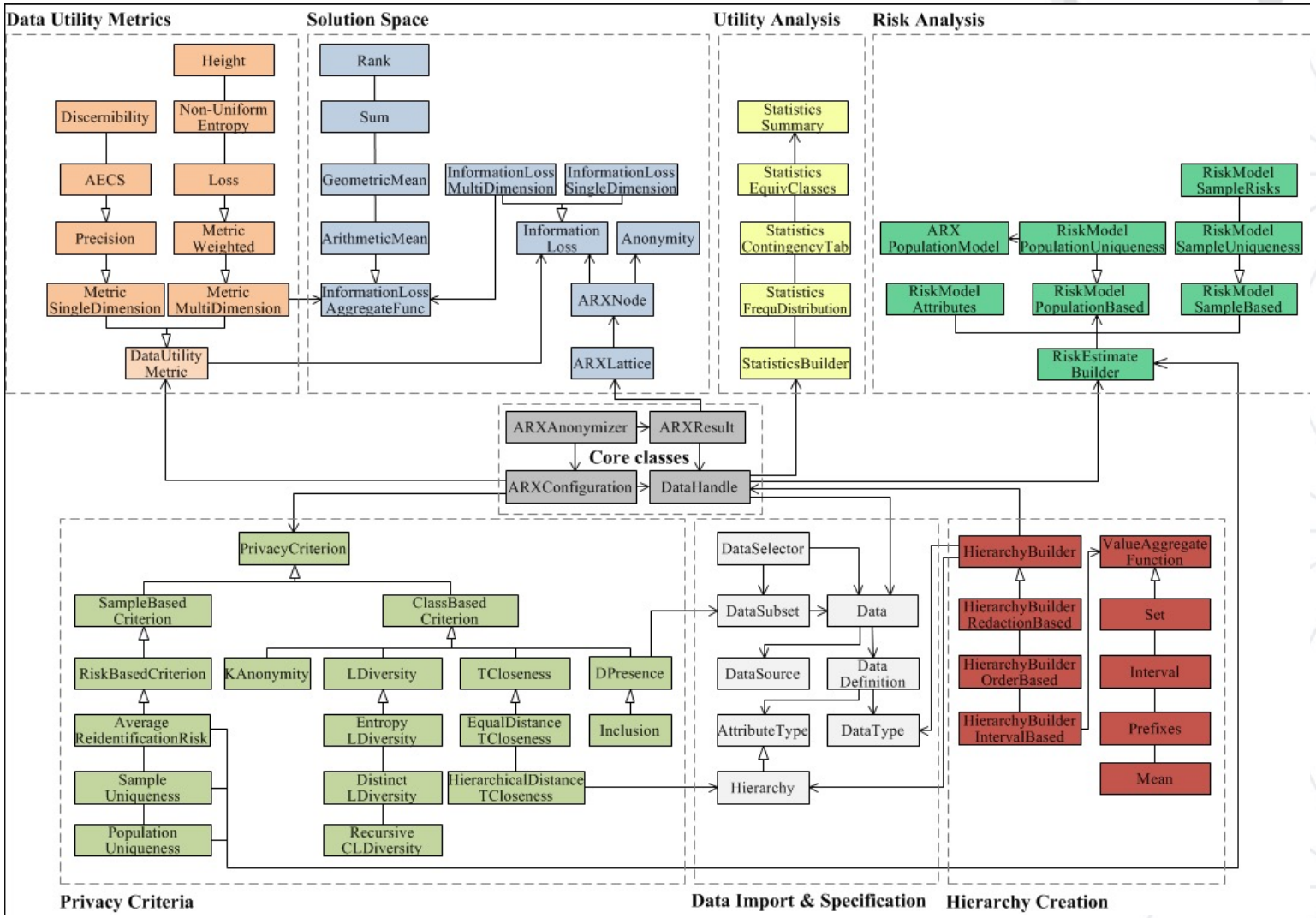
- CSV
- Excel
- SQL DBs (MySQL, PostgreSQL)

ARX Data Anonymization Tool

Privacy models:

- K-anonymity
- l-diversity
- t-closeness
- Differential Privacy (initial)
- Some others

Supports various metrics for measuring information loss



ARX Data Anonymization Tool

Can be used as a library

- Written in Java ☹
- All functionalities (privacy models and metrics) available to developers
- Documentation: https://arx.deidentifier.org/wp-content/uploads/javadoc/current/api/index.html

There is a project called Arx as a Service that offers Arx API as Web API:

- https://navikt.github.io/arxaas/

Limitations:

- Only Java API (but there are web api)
- Arx as a Service allows using Python, but the package is still a bit preliminary
  - **We will use it in the lab**

ARX Data Anonymization Tool

```
curl 'http://localhost:8080/api/anonymize' -i -X POST \
   -H 'Content-Type: application/json' \
   -d '{
  "data" : [ [ "age", "gender", "zipcode" ], [ "34", "male", "81667" ], [ "35", "female", "81668" ], [ "36", "male",
"81669" ], [ "37", "female", "81670" ], [ "38", "male", "81671" ], [ "39", "female", "81672" ], [ "40", "male", "81673"
], [ "41", "female", "81674" ], [ "42", "male", "81675" ], [ "43", "female", "81676" ], [ "44", "male", "81677" ] ],
  "attributes" : [ {
    "field" : "age",
    "attributeTypeModel" : "IDENTIFYING",
    "hierarchy" : null
  }, {
    "field" : "gender",
    "attributeTypeModel" : "SENSITIVE",
    "hierarchy" : null
  }, {
    "field" : "zipcode",
    "attributeTypeModel" : "QUASIIDENTIFYING",
    "hierarchy" : [ [ "81667", "8166*", "816**", "81***", "8****", "*****" ], [ "81668", "8166*", "816**", "81***",
"8****", "*****" ], [ "81669", "8166*", "816**", "81***", "8****", "*****" ], [ "81670", "8167*", "816**",
"81***", "8****", "*****" ], [ "81671", "8167*", "816**", "81***", "8****", "*****" ], [ "81672", "8167*",
"816**", "81***", "8****", "*****" ], [ "81673", "8167*", "816**", "81***", "8****", "*****" ], [ "81674",
"8167*", "816**", "81***", "8****", "*****" ], [ "81675", "8167*", "816**", "81***", "8****", "*****" ], [
"81676", "8167*", "816**", "81***", "8****", "*****" ], [ "81677", "8167*", "816**", "81***", "8****", "*****" ]
]
  } ],
  "privacyModels" : [ {
    "privacyModel" : "KANONYMITY",
    "params" : {
      "k" : "5"
    }
  }, {
    "privacyModel" : "LDIVERSITY_DISTINCT",
    "params" : {
      "column_name" : "gender",
      "l" : "2"
    }
  } ],
}'
```

```
HTTP/1.1 200 OK
Vary: Origin
Vary: Access-Control-Request-Method
Vary: Access-Control-Request-Headers
Content-Type: application/json
Content-Length: 7813

{
  "anonymizeResult" : {
    "data" : [ [ "age", "gender", "zipcode" ], [ "*", "male", "816**" ], [ "*", "female", "816**" ], [
"*", "male", "816**" ], [ "*", "female", "816**" ], [ "*", "male", "816**" ], [ "*", "female",
"816**" ], [ "*", "male", "816**" ], [ "*", "female", "816**" ], [ "*", "male", "816**" ], [ "*",
"female", "816**" ], [ "*", "male", "816**" ] ],

....

  "riskProfile" : {
    "reIdentificationRisk" : {
      "measures" : {
        "estimated_journalist_risk" : 0.09090909090909091,
        "records_affected_by_highest_prosecutor_risk" : 1.0,
        "sample_uniques" : 0.0,
        "lowest_risk" : 0.09090909090909091,
....
}
```

ARX Data Anonymization Tool

How to use it in Python:

1. Launch an ARX Server, using the Dockerized version

```
docker run -p 8080:8080 navikt/arxaas
```

2. Install pyarxaas

```
pip install pyarxaas
```

3. Use it in Python

```
arxaas = ARXaaS(url)
…
anon_result = arxaas.anonymize(dataset, [kanon])
```

IBM DiffPrivLib

Developed by IBM

"This is a library dedicated to **differential privacy** *and machine learning. Its purpose is to allow experimentation, simulation, and implementation of differentially private models using a common codebase and building blocks"*

Written in Python

Documentation at: https://diffprivlib.readthedocs.io/en/latest/

IBM DiffPrivLib

Supports:

- Basic Mechanisms for Differential Privacy
  - Laplacian mechanism
  - Geometric Mechanism
  - Exponential Mechanism
- Tools for Differential Privacy functions
  - Mean, Sum, Histogram

Differentially-private machine learning models

- Classification models
  - Gaussian Naive Bayes
  - Logistic Regression
  - Random Forest
- Regression models
  - Linear Regression
- Clustering models
  - K-Means
- Dimensionality reduction models
  - PCA
- Preprocessing
  - Standard Scaler

IBM DiffPrivLib

Differentially Private Histograms

Plotting the distribution of ages in `Adult`

In [1]:
```
import numpy as np
from diffprivlib import tools as dp
import matplotlib.pyplot as plt
```

We first read in the list of ages in the Adult UCI dataset (the first column).

In [2]:
```
ages_adult = np.loadtxt("https://archive.ics.uci.edu/ml/machine-learning
-databases/adult/adult.data",
                        usecols=0, delimiter=", ")
```

From:
https://github.com/IBM/differential-privacy-library/blob/main/notebooks/histograms.ipynb

Differentially private histograms

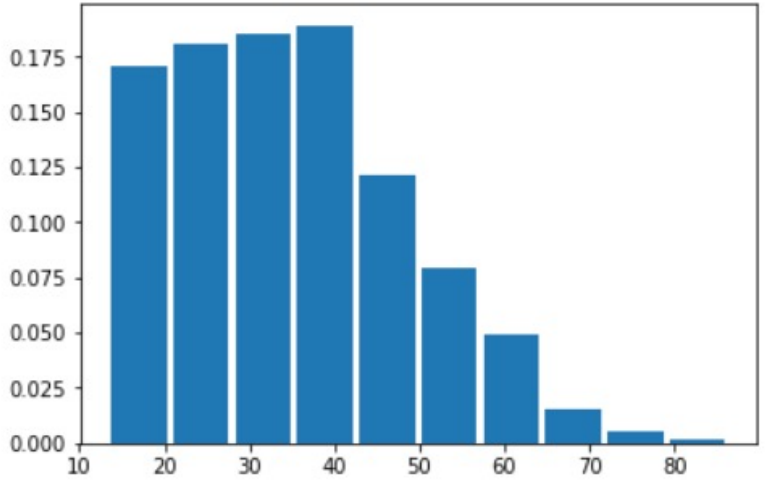
Using `diffprivlib`, we can calculate a differentially private version of the histogram. For this example, we use the default settings:

- `epsilon` is 1.0
- `range` is not specified, so is calculated by the function on-the-fly. This throws a warning, as it leaks privacy about the data (from `dp_bins`, we know that there are people in the dataset aged 17 and 90).

```
dp_hist, dp_bins = dp.histogram(ages_adult)
dp_hist = dp_hist / dp_hist.sum()

plt.bar(dp_bins[:-1], dp_hist, width=(dp_bins[1] - dp_bins[0]) * 0.9)
plt.show()
```

```
.../site-packages/diffprivlib/tools/histograms.py:131: PrivacyLeakWarnin
g: Range parameter has not been specified. Falling back to taking range
from the data.
To ensure differential privacy, and no additional privacy leakage, the r
ange must be specified independently of the data (i.e., using domain kno
wledge).
  "specified independently of the data (i.e., using domain knowledge).",
PrivacyLeakWarning)
```



IBM DiffPrivLib

Differentially Private Histograms

Plotting the distribution of ages in `Adult`

In [1]:
```
import numpy as np
from diffprivlib import tools as dp
import matplotlib.pyplot as plt
```

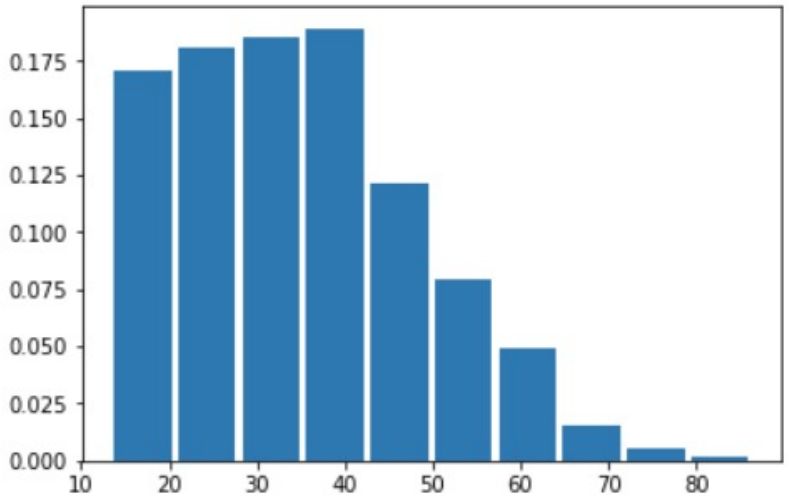
We first read in the list of ages in the Adult UCI dataset (the first column).

In [2]:
```
ages_adult = np.loadtxt("https://archive.ics.uci.edu/ml/machine-learning
-databases/adult/adult.data",
                        usecols=0, delimiter=", ")
```

```
hist, bins = np.histogram(ages_adult)
hist = hist / hist.sum()
```

Using `matplotlib.pyplot`, we can plot a barchart of the histogram distribution.

```
plt.bar(bins[:-1], hist, width=(bins[1]-bins[0]) * 0.9)
plt.show()
```



Google Differential Privacy

Library by Google to generate ε- and (ε, δ)-differentially private statistics over datasets
Similar to IBM DiffPrivLib: https://github.com/google/differential-privacy

- Written in C, less mature

Implements various mechanisms

- Laplace, Gaussian, etc.

Various Statistics:

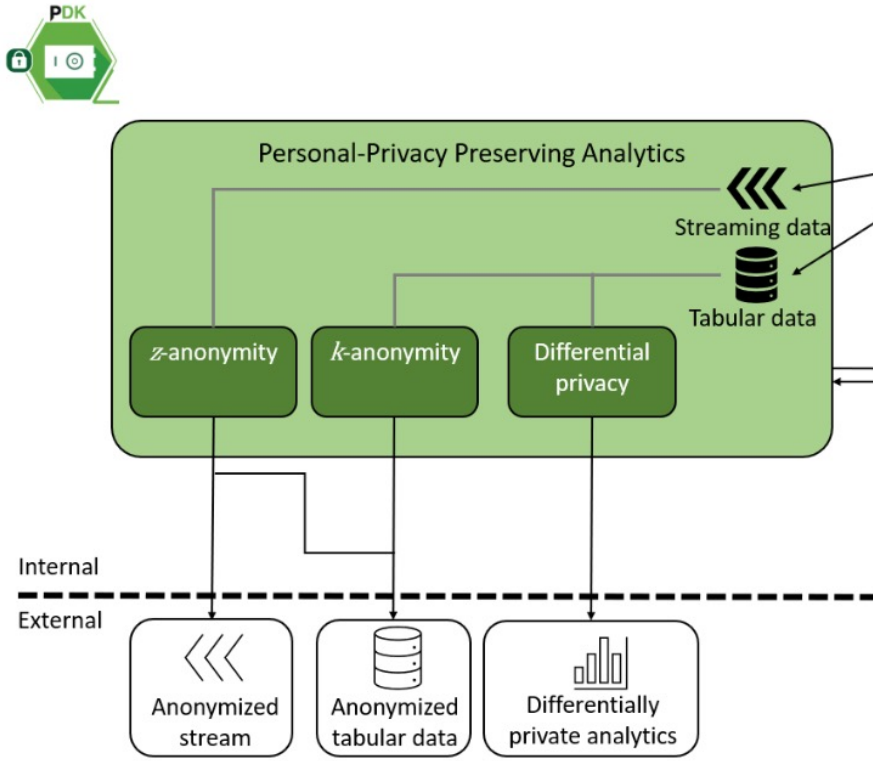
- Count, Sum, Mean, Variance, Quantiles…

There exist Python bindings:

- https://github.com/OpenMined/PyDP
- Still preliminary

Personal Privacy Preserving Analytics

Developed by Politecnico di Torino
In the context of the EU Project PIMCity (https://www.pimcity-h2020.eu/)
It is a simple Python module that offers easy-to-use privacy models



Personal Privacy Preserving Analytics

- The P-PPA can aggregate data from different sources and formats
  - Traditional RDBMS
  - MongoDB data lake
  - Data in CSV files

	A	B	C
1	Transaction Description	Expense Type	Amount
2	Chester Diner	Restaurant	$ 24.22
3	Income Tax Payment	Taxes	$ 535.00
4	Ole Tymes Cafe	Restaurant	$ 12.58
5	Plane ticket to Melbourne	Travel	$ 654.32
6	Odessa's	Restaurant	$ 13.36
7	Car Rental in Australia	Travel	$ 185.55
8	K Crew	Clothing	$ 86.99
9	Ruby's Famous Bbq Joint	Restaurant	$ 5.66
10	Street Corner Market	Restaurant	$ 9.85
11	Airport Parking	Travel	$ 22.55
12	The Friendly Chef	Restaurant	$ 67.85
13	Floorgreen's	Personal Items	$ 24.55
14	Orange Democracy	Clothing	$ 86.99
15	Car Care	Auto Expense	$ 24.22
16	The Narrow Lantern	Restaurant	$ 101.33

Personal Privacy Preserving Analytics

- Written in Python
- Simple installation, only a few Python `requirements`
- Usage a Python module

```
from algorithms.kanonymity.mondrian.mondrian import Mondrian
mondrian = Mondrian(3, user_choice_index=[2,3,4,5])
k_anonymized_dataframe = mondrian.perform(input_dataframe)
```

Personal Privacy Preserving Analytics

- The P-PPA offers Web APIs to allow remote use

Personal-Pivacy Preserving Analytics

1.0.0 OAS3

The PIMCity P-PPA, it's a tool to allow data analysts and stakeholders to retrieve useful information from the data, while preserving the privacy of the users whose data are in the studied datasets. It follows some query example.

k-anonymity Provides k-anonymity privacy guarantees.

GET `/kanon` According to the passed parameters, it permits internal PPPA modules to collect data from the requested source and provide k-anonymized data.

differential privacy Provides data statistics in a differentially private flavour.

GET `/diffpriv` According to the passed parameters, it permits internal PPPA modules to collect data from the requested source and provide differentially private data statistics.

CNIL-PIA software

The PIA software aims to help data controllers build and demonstrate compliance to the GDPR.

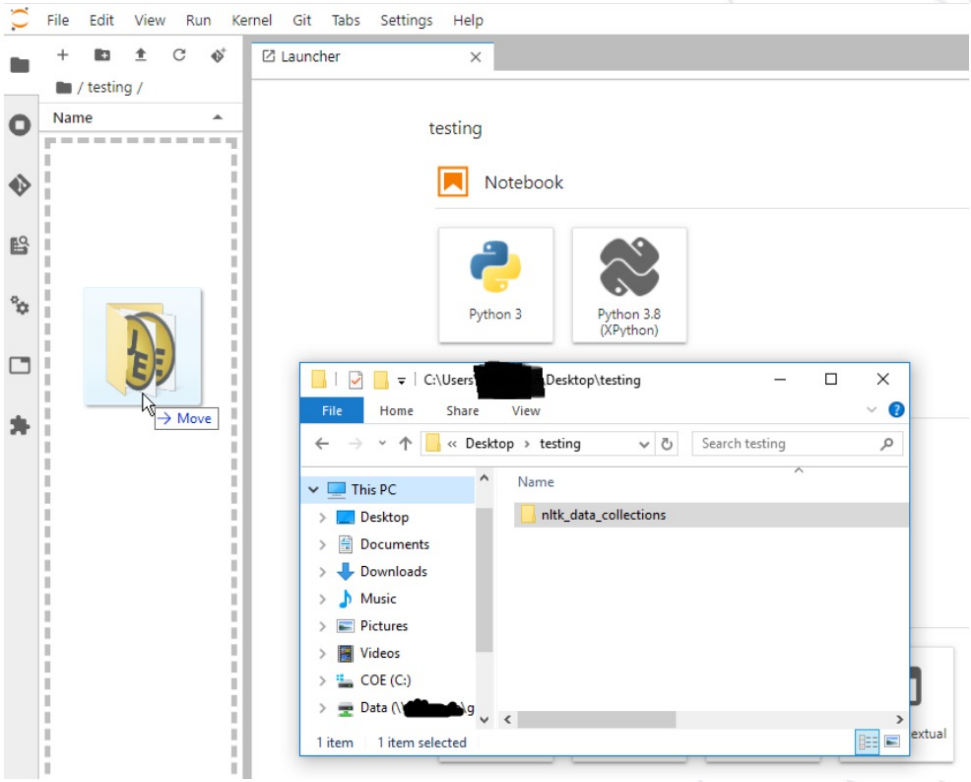
- The tools is available in French and in English.
- It facilitates carrying out a data protection impact assessment
- This tool also intends to ease the use of the PIA guides published by the CNIL.
- Portable and Web Version
- Website: https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment

Summary

TOOL	Language	K-anon & similar	Diff Priv	Web API	Input
ARX	Java	Y	Partially	Y	CSV, Excel, SQL, JSON, Java API
DiffPrivLib	Python		Y		Numpy
Google Differential Privacy	C/C++		Y		C vars
P-PPA	Python	Y	Y	Y	Pandas, CSV, SQL

Access to the Lab

- Go with your browser on: https://jupyter-dev.polito.it/
- Login with:
  - Username: your mail
  - Password: your mail
- Download the notebook from: https://www.dropbox.com/s/xqq66yjwejmuo2o/practice.ipynb?dl=1
  - Can copy the link from chat
- Upload it on Jupyter (drag and drop works ☺)
- Complete the exercises!




Perguntas
Fragen
Domande
Galdera
Otázky
Questions
Spørgsmål
Pertanyaan
kysymykset
Frågor
Spørsmål
Cwestiynau
вопросы
Preguntes
Sorular
Въпроси
Vragen
Pytania