EPFL

SPRING

SECURITY AND PRIVACY ENGINEERING LABORATORY

# CS-523 Advanced Topics on Privacy Enhancing Technologies

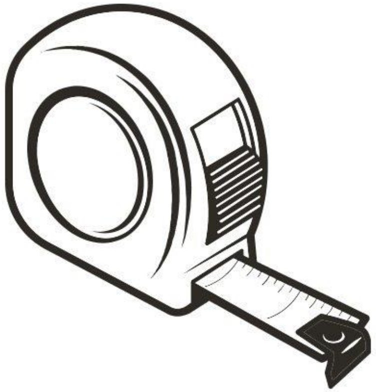
# Privacy-preserving data publishing (Part I)

**Theresa Stadler**
SPRING Lab
theresa.stadler@epfl.ch

Some slides/ideas adapted from: Carmela Troncoso, Jean-Pierre Hubaux, Vitaly Shmatikov

# Introduction
## Anonymization

Course aim: learn **toolbox for privacy engineering**



**Application Layer**

**Network Layer**

*tool*
to eliminate links between data and individuals

*mechanism*
to evaluate privacy

# Goals

## What should you learn today?

- Basic understanding of **anonymization**

- Understand **key pitfalls** of anonymization:
  - Belief that removing personal identifiable information is enough
  - Belief that we can constrain the knowledge of the adversary
  - Ignore that **high-dimensionality** and **sparsity** imply that individuals are **uniquely identifiable**

- Understand **reasoning and metrics** to evaluate anonymization

- Understand **practical issues** when anonymizing high-dimensional datasets

# The promised benefits of data-driven everything…

Better governmental services

Improved health outcomes


A more efficient, greener industrial production

# ...have a flip side

**Use and misuse**

Data can be used for good… and for bad

**Potential harms**

Surveillance, control and manipulation

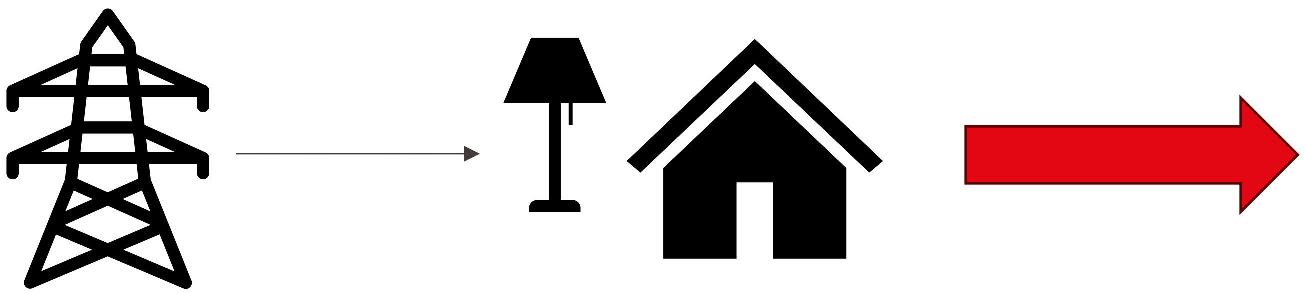
**False conclusions**

Data bias and processing errors can have a strong impact on people’s life

# Utility data used by ICE in the US

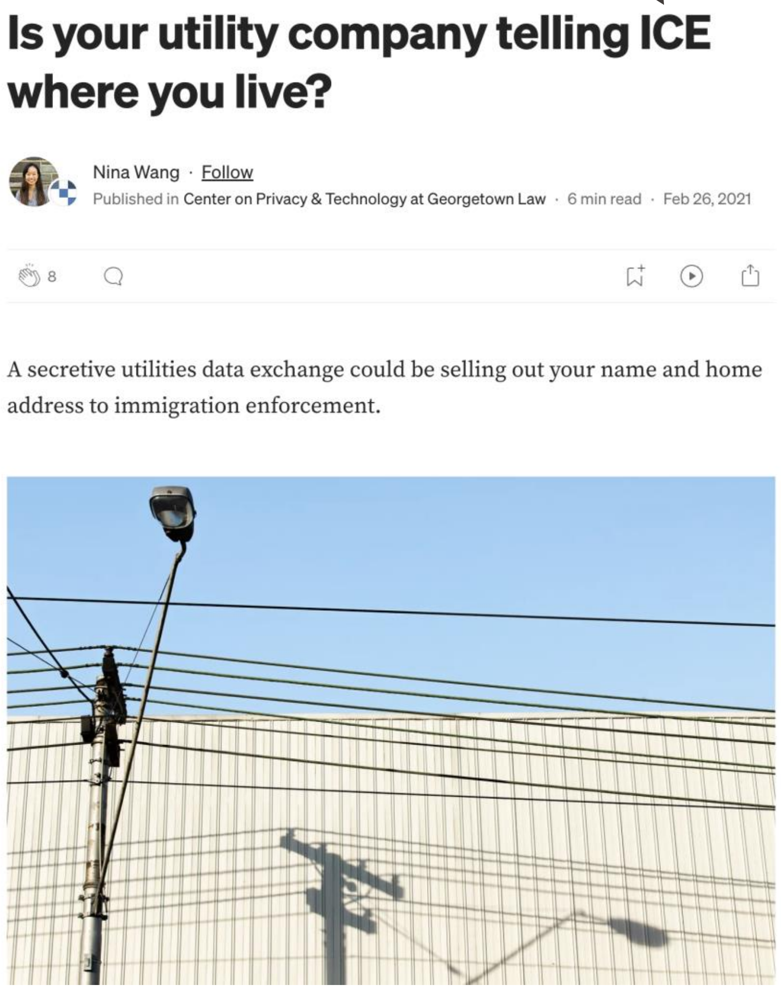
U.S. Immigration and Customs Enforcement

**National Consumer Telecom & Utilities Exchange (NCTUE) collects utilities data for credit assessment**



**NCTUE**

**171M customers (~50% US population)**


## Is your utility company telling ICE where you live?

Nina Wang · Follow
Published in Center on Privacy & Technology at Georgetown Law · 6 min read · Feb 26, 2021

8

A secretive utilities data exchange could be selling out your name and home address to immigration enforcement.

■ https://medium.com/center-on-privacy-technology/is-your-utility-company-telling-ice-where-you-live-ae1c7d187eff

# Regulations



## Laws and regulations require that personal data are protected

→ not leak much about individuals & not used for unforeseen purposes


**Universal Declaration of Human Rights**

**Article 12.** "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks."

10 December 1948, http://www.un.org/en/universal-declaration-human-rights/

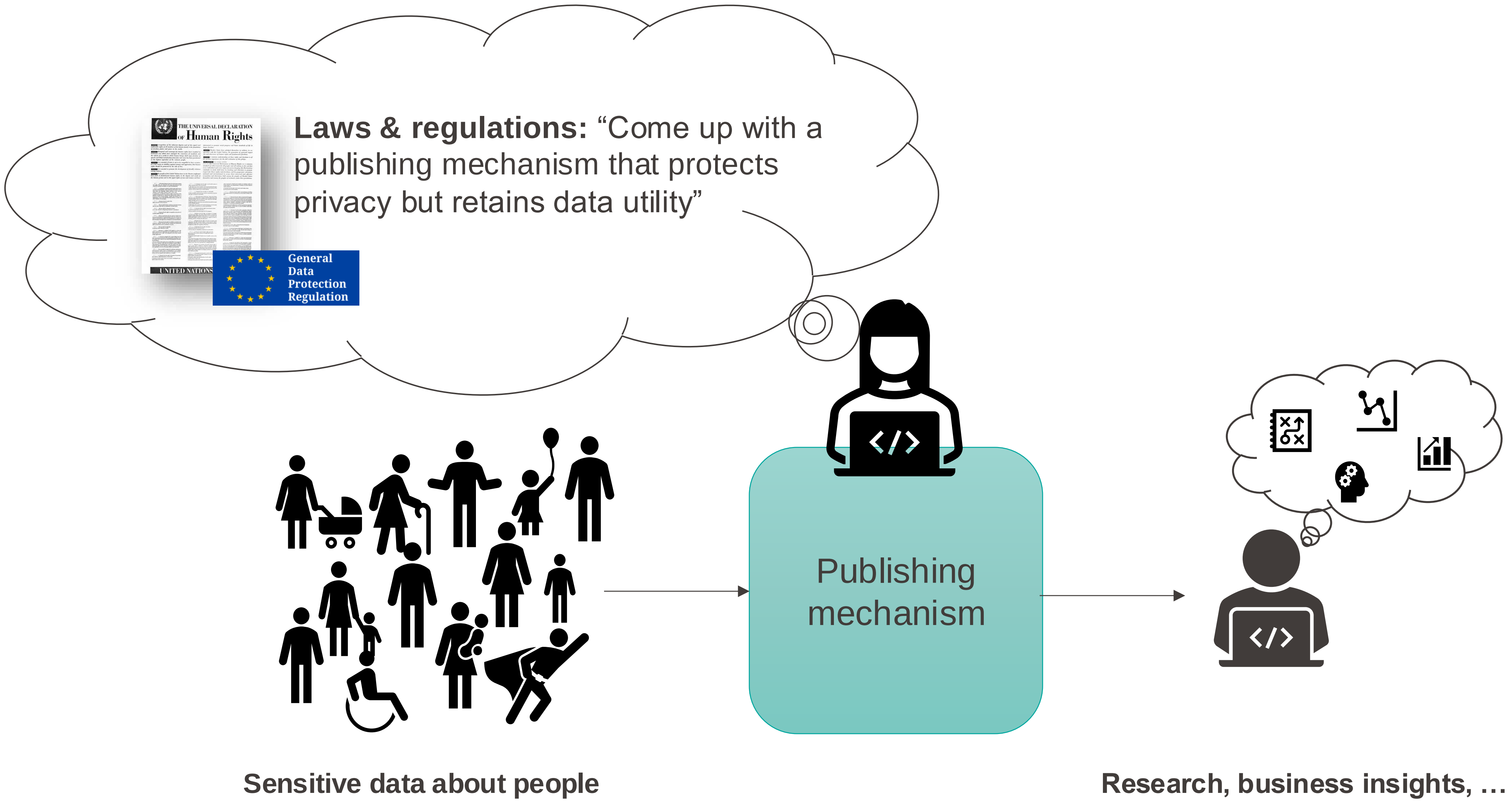
General Data Protection Regulation

**GDPR**

**Article 1.** "**personal data** means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, genetic, …" ;

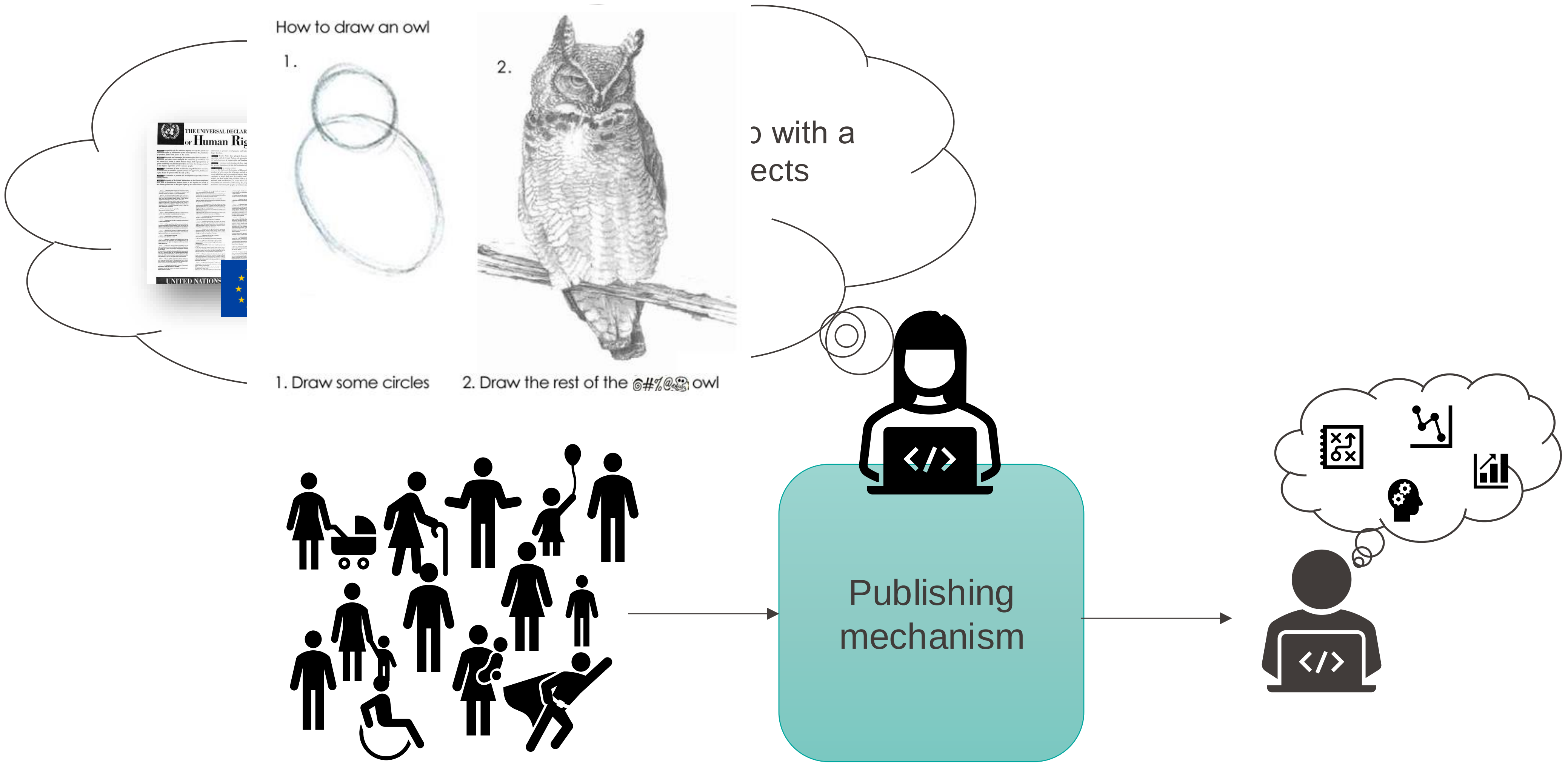
25 May 2018, https://www.eugdpr.org

# And in practice?



# And in practice?!?!?!?





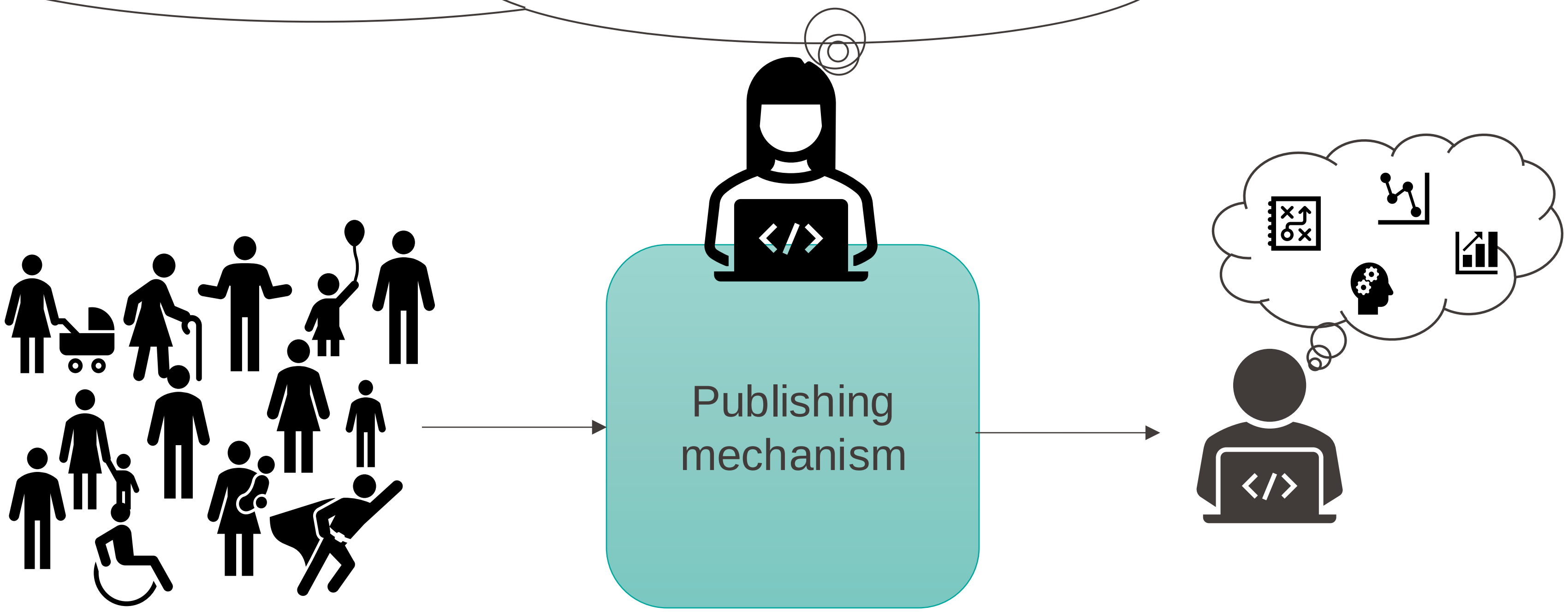

**Sensitive data about people**

**Research, business insights, …**

# The key question

## How to publish useful sensitive data in a privacy-preserving way?

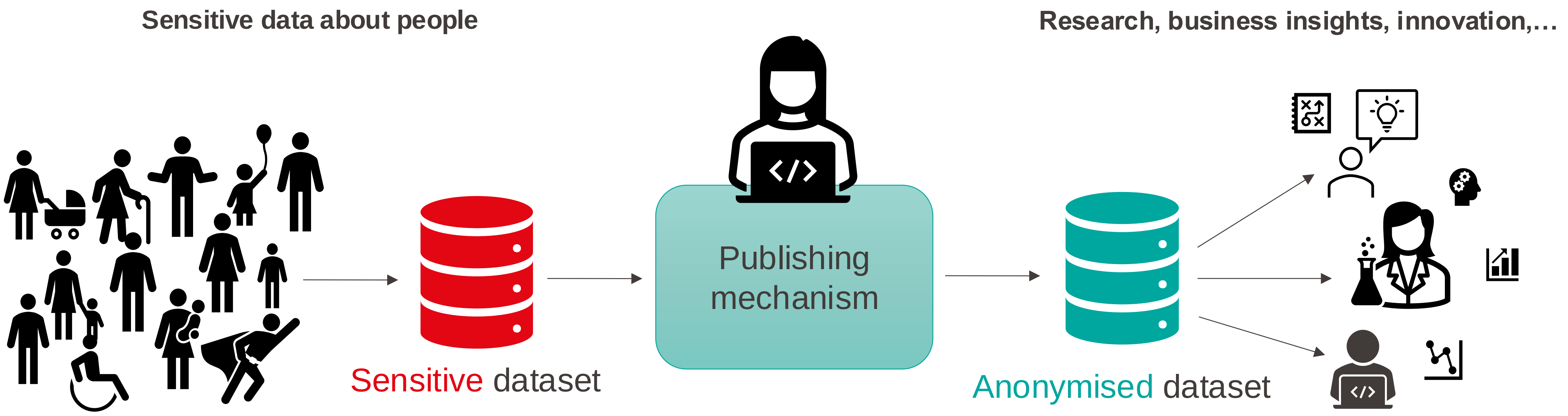
(broad definition of publish: share, publish internally,… anything beyond collection)



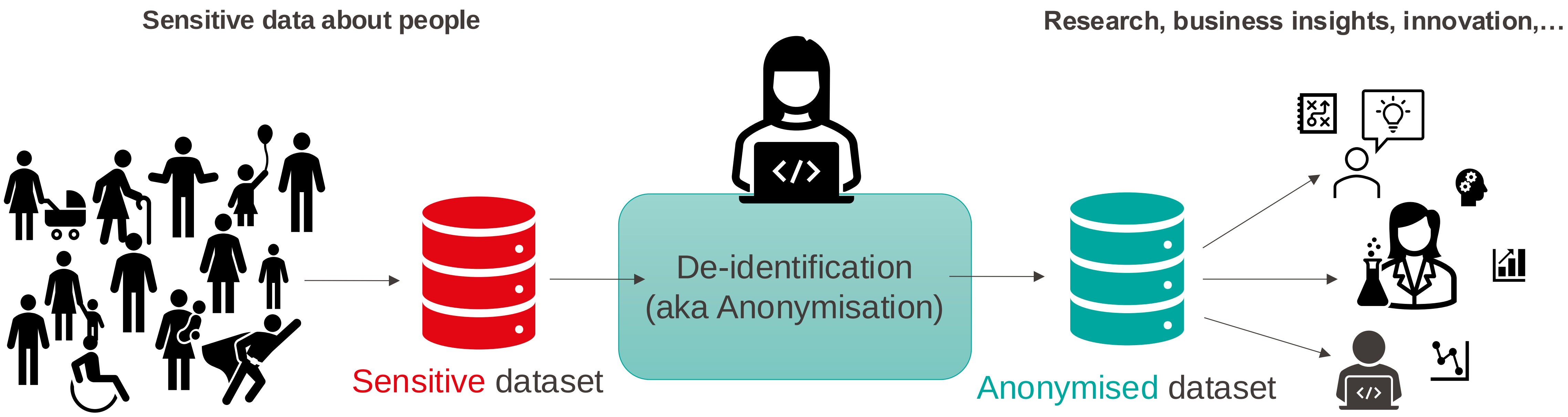

**Sensitive data about people**

**Research, business insights, …**

# Privacy-preserving microdata sharing



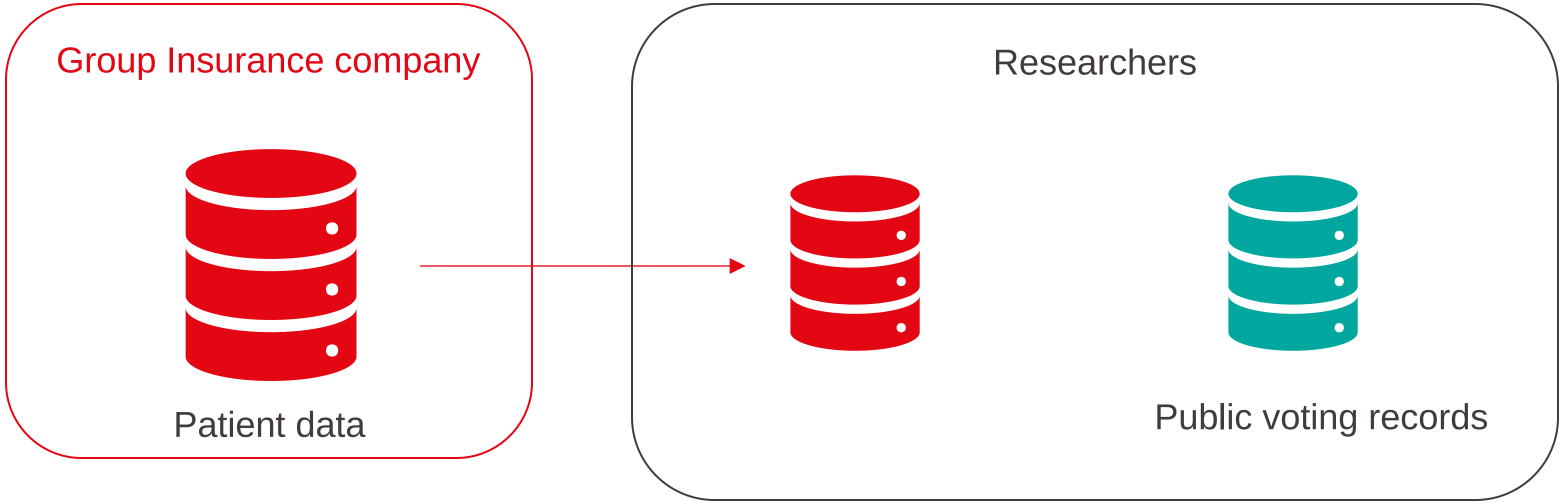
# Privacy-preserving microdata sharing




**Mask or Remove Personally Identifiable Information (PII):**

name, SSN, phone number, address, email, twitter handle,…

# Naïve “de-identification” fails
## Real life example




| Date_of_birth | ZIP code | ethnicity | procedure |
|---|---|---|---|
| 07/07/1960 | 1024 | caucasian | chemotherapy |
| … | … | … | … |

| Date_of_birth | ZIP code | ethnicity |
|---|---|---|
| 07/07/1960 | 1024 | caucasian |
| … | … | … |

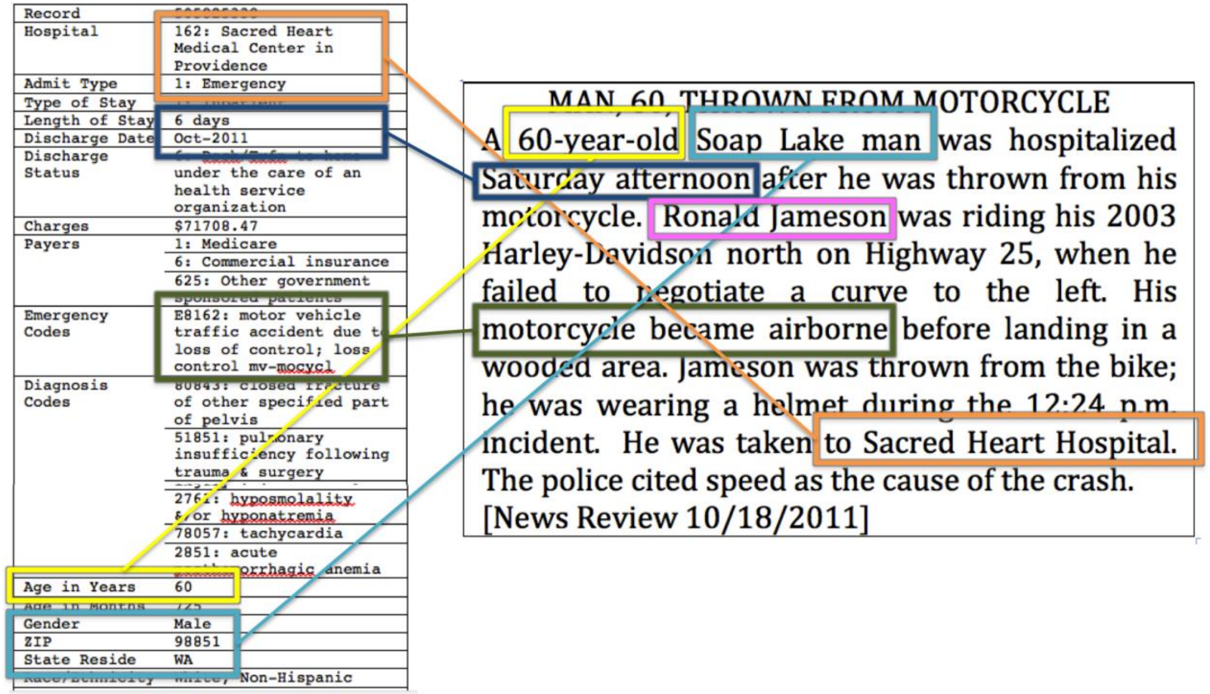
- In Massachusetts, Group Insurance Commission (GIC) collected patient-specific data about ~135K state employees and their families
    - Data contained nearly one hundred attributes: Ethnicity, Visit date, Diagnosis, Procedure, Medication, Total charge, ZIP, Birth date, Sex
- The data had no PII so was believed to be anonymous
- Latanya Sweeney (PhD student) bought voting records in Massachusetts (20$).
    - Voting records included: ZIP, Birth date, Sex, Name, Address, Date registered, Party affiliation, Date last voted
- Partial matching allowed to learn sensitive health information about governor of Massachusetts

- https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf

# 15 years later…

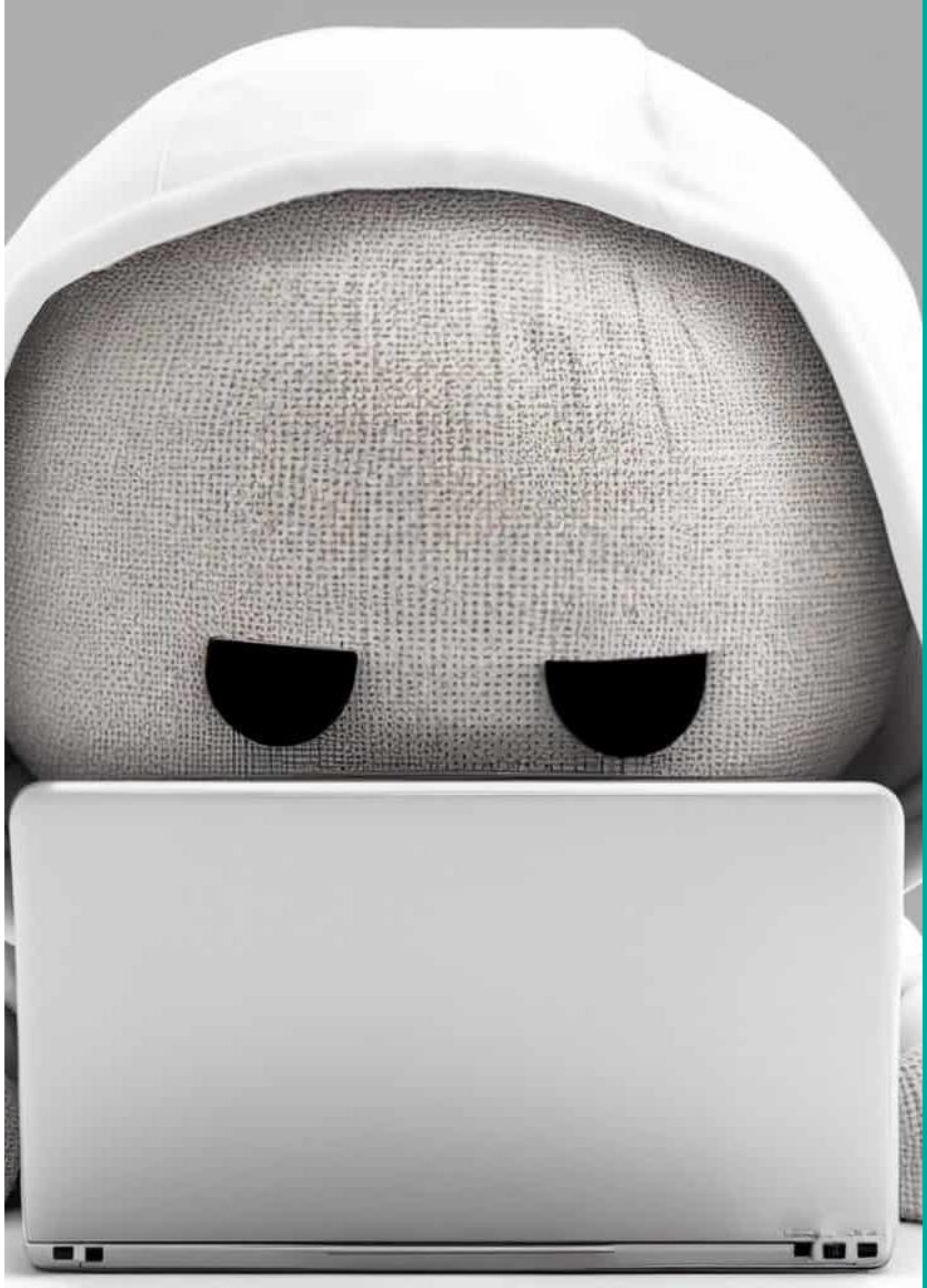


*"newspaper stories about hospital visits in Washington State leads to identifying the matching health record 43% of the time"*




This work from Sweeney prompted Washington state to change their access control policy to health records

- https://techscience.org/a/2015092903/



# How to define privacy threats in data publishing?

# Data publishing privacy threats

What is a **privacy threat** in data publishing?

Must be defined in contrast to the intended purpose of the data publishing

Defined by their capacity, attack strategy, prior knowledge

An unauthorized disclosure occurs when an attacker gains unauthorized access to sensitive data

What new information does the attacker learn about whom?

# Data publishing privacy threats

**Membership disclosure**: an individual's data is **in** a dataset of sensitive nature

Think: Dataset of criminal records, dataset of highly contagious diseases, dataset about harassed victims

| Date_of_birth | ZIP code | gender | sensitive |
|---|---|---|---|
| 07/07/1960 | 1024 | female | **value1** |
| 01/09/1976 | 1015 | male | **value1** |
| 01/08/1987 | 1024 | male | **value1** |
| 12/09/1976 | 1025 | female | **value1** |
| 01/08/1999 | 1023 | male | **value1** |
| … | … | … | … |

Target

Also sometimes called table linkage

# Data publishing privacy threats

**Attribute disclosure**: an individual's data is in a dataset, and this individual's anonymity set has **a unique sensitive attribute**

Think: Individual's anonymity set only contains sexual assaults, only contains patients with AIDS, only contains transgender victims

| Date_of_birth | ZIP code | gender | sensitive |
|---|---|---|---|
| 07/07/1960 | 1024 | **female** | **value2** |
| 01/09/1976 | 1015 | male | value1 |
| 01/08/1987 | 1024 | male | value2 |
| 12/09/1976 | 1025 | **female** | **value2** |
| 01/08/1999 | 1023 | male | value1 |
| … | … | … | … |

Target

Also sometimes called attribute inference

# Data publishing privacy threats

**Record disclosure**: an individual's data is in a dataset, and this individual's anonymity set contains **only one record**

Think: Individual assault's date and place, date of contracting AIDS and reason, date of harassment and place

| Date_of_birth | ZIP code | gender | sensitive |
|---|---|---|---|
| 07/07/1960 | 1024 | female | value2 |
| 01/09/1976 | 1015 | male | value1 |
| 01/08/1987 | 1024 | male | value2 |
| 12/09/1976 | **1025** | **female** | **value2** |
| 01/08/1999 | 1023 | male | value1 |
| … | … | … | … |

Target

Also sometimes called singling out, re-identification, unique record linkage

# Data publishing privacy threats

**Disclosure can be probabilistic or certain**

Prior belief

40% value1

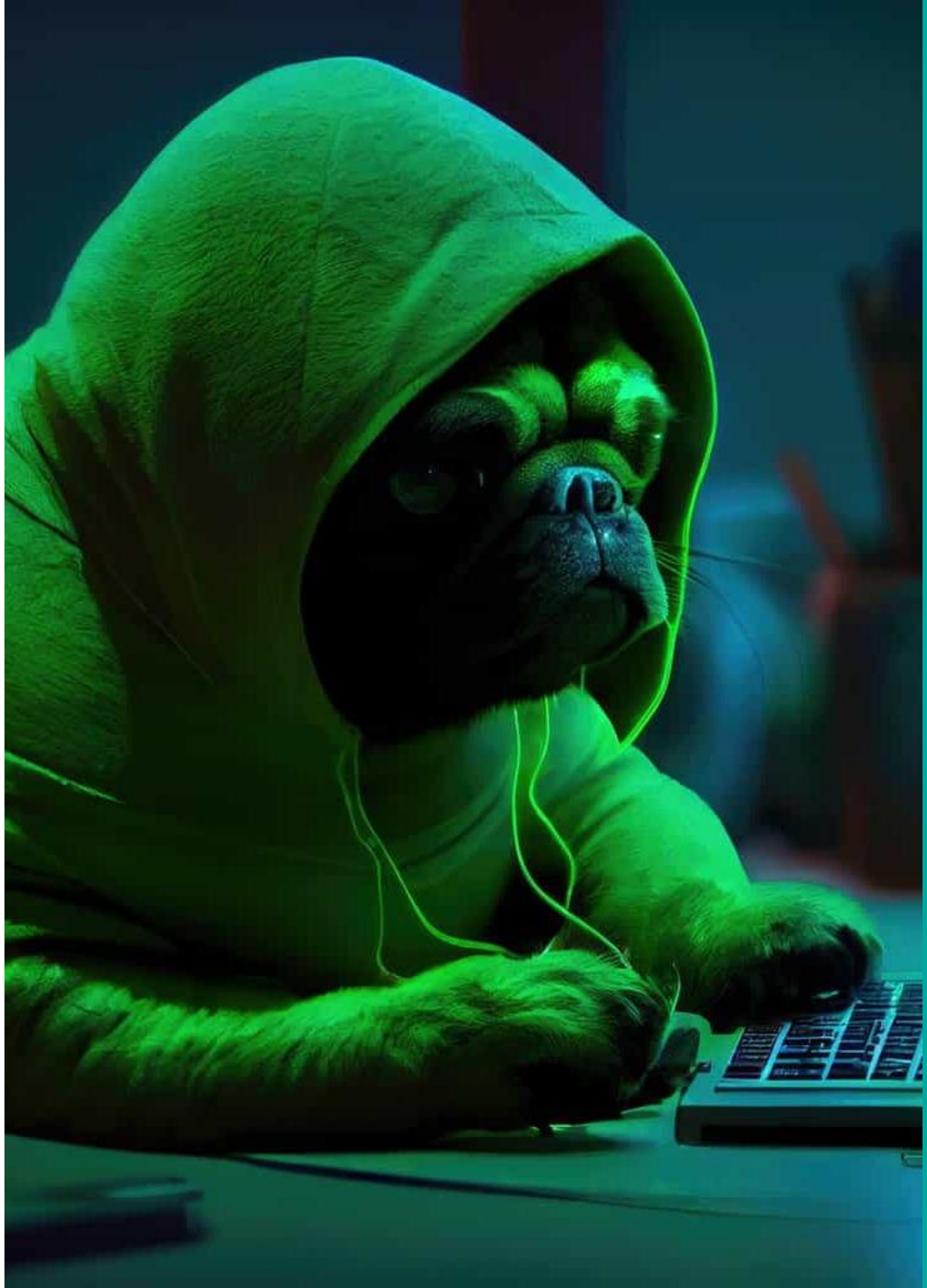
60% value2

| ZIP code | gender | sensitive |
|---|---|---|
| **1024** | **female** | **value1** |
| 1015 | male | value1 |
| 1024 | male | value2 |
| **1025** | **female** | **value2** |
| 1023 | male | value1 |

Posterior belief

10%

90%





# Case study: The Airbnb Lighthouse project

# Case study:
# Airbnb Lighthouse project

- **Airbnb has a problem:** Gap in booking acceptance rates based on users' perceived race
  - See #AirbnbWhileBlack

- **Intended purpose**: Measure discrepancies in Airbnb guest acceptance rates to tackle discrimination

- **Privacy concern**: An internal attacker might learn perceived race of users (primary concern is attribute disclosure)

- **Key question**: How to tag users' profiles with perceived race and measure gap in acceptance rates while preventing privacy violations?

**Disclaimer:** for the purpose of the lecture some examples may not be super faithful to reality. Whole account by Airbnb here: https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

# Case study: Airbnb Lighthouse project

## AirBnB land

| UserId | name | photo | hometown | education | hobbies | n_accept | n_reject |
|---|---|---|---|---|---|---|---|
| 1 | John | URL1 | Athens/GA | None | Basketball | 6 | 1 |
| 2 | Carla | URL2 | Boston/MA | PhD | Running | 4 | 2 |
| 3 | Nathan | URL3 | Seattle/WA | BSc | Basketball | 10 | 2 |
| 4 | Darnell | URL4 | Atlanta/GA | High School | Basketball | 2 | 4 |

# Airbnb privacy risks
# Problem 1: Direct identifiers

Direct identifier

Sensitive attribute

| name | hometown | education | hobbies | n_accept | n_reject | race |
|---|---|---|---|---|---|---|
| John | Athens/GA | None | Basketball | 6 | 1 | White |
| Carla | Boston/MA | PhD | Running | 4 | 2 | Latino |
| Nathan | Seattle/WA | BSc | Basketball | 10 | 2 | White |
| Darnell | Atlanta/GA | High School | Basketball | 2 | 4 | Black |
| Emma | Cambridge/MA | MSc | Running | 6 | 0 | White |
| Jamal | Redmond/WA | BSc | Basketball | 3 | 3 | Black |
| Raven | Seattle/WA | BSc | Basketball | 2 | 4 | Black |
| Ben | Macon/GA | High School | Basketball | 2 | 2 | Asian |
| Molly | Salem/MA | MSc | Running | 4 | 1 | White |
| Markus | Spokane/WA | BSc | Basketball | 3 | 1 | White |

# Case study: Airbnb Lighthouse project

## AirBnB land

| UserId | name | photo | hometown | education | hobbies | n_accept | n_reject |
|---|---|---|---|---|---|---|---|
| 1 | John | URL1 | Athens/GA | None | Basketball | 6 | 1 |
| 2 | Carla | URL2 | Boston/MA | PhD | Running | 4 | 2 |
| 3 | Nathan | URL3 | Seattle/WA | BSc | Basketball | 10 | 2 |
| 4 | Darnell | URL4 | Atlanta/GA | High School | Basketball | 2 | 4 |

**1. Map userId->nid**

## AirBnB discrimination team

| nid | n_accept | n_reject |
|---|---|---|
| 8 | 6 | 1 |
| 75 | 4 | 2 |
| 435 | 10 | 2 |
| 23 | 2 | 4 |

| name | photo | nid |
|---|---|---|
| John | URL1 | 8 |
| Carla | URL2 | 75 |
| Nathan | URL3 | 435 |
| Darnell | URL4 | 23 |

# Case study: Airbnb Lighthouse project

AirBnB land

| UserId | name | photo | hometown | education | hobbies | n_accept | n_reject |
|---|---|---|---|---|---|---|---|
| 1 | John | URL1 | Athens/GA | None | Basketball | 6 | 1 |
| 2 | Carla | URL2 | Boston/MA | PhD | Running | 4 | 2 |
| 3 | Nathan | URL3 | Seattle/WA | BSc | Basketball | 10 | 2 |
| 4 | Darnell | URL4 | Atlanta/GA | High School | Basketball | 2 | 4 |

1. Map userId->nid

AirBnB discrimination team

| nid | n_accept | n_reject |
|---|---|---|
| 8 | 6 | 1 |
| 75 | 4 | 2 |
| 435 | 10 | 2 |
| 23 | 2 | 4 |

2. Encrypt to PK partner & delete mapping

Research partner land

# Case study: Airbnb Lighthouse project

## AirBnB land

| UserId | name | photo | hometown | education | hobbies | n_accept | n_reject |
|---|---|---|---|---|---|---|---|
| 1 | John | URL1 | Athens/GA | None | Basketball | 6 | 1 |
| 2 | Carla | URL2 | Boston/MA | PhD | Running | 4 | 2 |
| 3 | Nathan | URL3 | Seattle/WA | BSc | Basketball | 10 | 2 |
| 4 | Darnell | URL4 | Atlanta/GA | High School | Basketball | 2 | 4 |

**1. Map userId->nid**

### AirBnB discrimination team

| nid | n_accept | n_reject |
|---|---|---|
| 8 | 6 | 1 |
| 75 | 4 | 2 |
| 435 | 10 | 2 |
| 23 | 2 | 4 |

**2. Encrypt to PK partner & delete mapping**

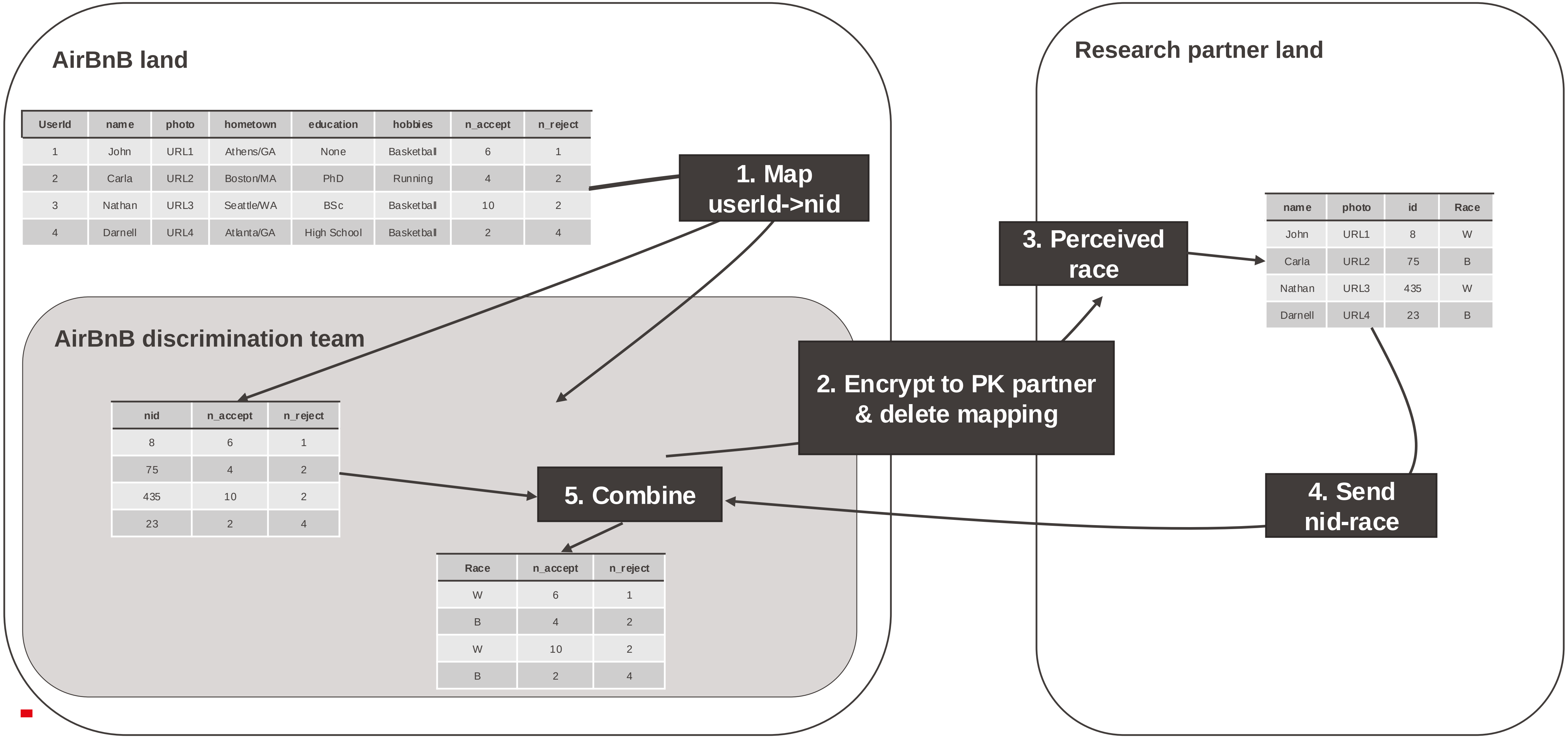
## Research partner land

**3. Perceived race**

| name | photo | id | Race |
|---|---|---|---|
| John | URL1 | 8 | W |
| Carla | URL2 | 75 | B |
| Nathan | URL3 | 435 | W |
| Darnell | URL4 | 23 | B |

**4. Send nid-race**

# Case study: Airbnb Lighthouse project




| UserId | name | photo | hometown | education | hobbies | n_accept | n_reject |
|---|---|---|---|---|---|---|---|
| 1 | John | URL1 | Athens/GA | None | Basketball | 6 | 1 |
| 2 | Carla | URL2 | Boston/MA | PhD | Running | 4 | 2 |
| 3 | Nathan | URL3 | Seattle/WA | BSc | Basketball | 10 | 2 |
| 4 | Darnell | URL4 | Atlanta/GA | High School | Basketball | 2 | 4 |

| name | photo | id | Race |
|---|---|---|---|
| John | URL1 | 8 | W |
| Carla | URL2 | 75 | B |
| Nathan | URL3 | 435 | W |
| Darnell | URL4 | 23 | B |

| nid | n_accept | n_reject |
|---|---|---|
| 8 | 6 | 1 |
| 75 | 4 | 2 |
| 435 | 10 | 2 |
| 23 | 2 | 4 |

| Race | n_accept | n_reject |
|---|---|---|
| W | 6 | 1 |
| B | 4 | 2 |
| W | 10 | 2 |
| B | 2 | 4 |

# Airbnb privacy risks
# Problem 2: Quasi-identifiers

| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| nid | hometown | education | hobbies | n_accept | n_reject | race |
| 45 | Athens/GA | None | Basketball | 6 | 1 | White |
| 245 | Boston/MA | PhD | Running | 4 | 2 | Latino |
| 23 | Seattle/WA | BSc | Basketball | 10 | 2 | White |
| 78 | Atlanta/GA | High School | Basketball | 2 | 4 | Black |
| 92 | Cambridge/MA | | Running | 6 | 0 | White |
| 12 | Redmond/WA | BSc | Basketball | 3 | 3 | Black |
| 99 | Seattle/WA | BSc | Basketball | 2 | 4 | Black |
| 128 | Macon/GA | High School | Basketball | 2 | 2 | Asian |
| 67 | Salem/MA | MSc | Running | 4 | 1 | White |
| 43 | Spokane/WA | BSc | Basketball | 3 | 1 | White |

# k-anonymity, l-diversity, t-closeness, and the likes…

# $k$-anonymity

Each person contained in the database **cannot be distinguished from at least $k$-1 other individuals** whose information also appears in the released database.

# k-anonymity Privacy

- Given a table D, find a table D' such that
  - D' satisfies the k-anonymity condition

| name | gender | zipcode | problem |
|---|---|---|---|
| John | male | 1012 | Cancer |
| Zoey | female | 1003 | Flu |
| Nathan | male | 1004 | Heart Disease |
| Lucas | male | 1005 | Heart Disease |
| Sam | male | 1004 | Flu |
| Max | male | 1012 | Cancer |
| Mathias | male | 1005 | HIV+ |
| Sarah | female | 1012 | Herpes |
| Julia | female | 1012 | Flu |

# k-anonymity

- To ensure anonymity, quasi-identifying attributes can be:
  - *generalized*
  - *suppressed*

- The process of making the database k-anonymous is called database sanitization.

| name | gender | zipcode | problem |
|---|---|---|---|
| John | * | 1012 | Cancer |
| Zoey | * | 100* | Flu |
| Nathan | * | 100* | Heart Disease |
| Lucas | * | 100* | Heart Disease |
| Sam | * | 100* | Flu |
| Max | * | 1012 | Cancer |
| Mathias | * | 100* | HIV+ |
| Sarah | * | 1012 | Herpes |
| Julia | * | 1012 | Flu |

$k$=4

# k-anonymity through generalisation

| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| nid | hometown | education | hobbies | n_accept | n_reject | race |
| 45 | Athens/GA | None | Basketball | 6 | 1 | White |
| 245 | Boston/MA | PhD | Running | 4 | 2 | Latino |
| 23 | Seattle/WA | BSc | Running | 10 | 2 | White |
| 78 | Atlanta/GA | High School | Basketball | 2 | 4 | Black |
| 92 | Cambridge/MA | MSc | Running | 6 | 0 | White |
| 12 | Redmond/WA | BSc | Basketball | 3 | 3 | Black |
| 99 | Seattle/WA | BSc | Running | 2 | 4 | Black |
| 128 | Macon/GA | High School | Basketball | 2 | 2 | Asian |
| 67 | Salem/MA | MSc | Running | 4 | 1 | White |
| 43 | Spokane/WA | BSc | Basketball | 3 | 1 | White |

# k-anonymity through generalisation

| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| nid | gen(hometown) | gen(education) | hobbies | n_accept | n_reject | race |
| 45 | GA | Low | Basketball | 6 | 1 | White |
| 245 | MA | High | Running | 4 | 2 | Latino |
| 23 | WA | Mid | Running | 10 | 2 | White |
| 78 | GA | Low | Basketball | 2 | 4 | Black |
| 92 | MA | High | Running | 6 | 0 | White |
| 12 | WA | Mid | Basketball | 3 | 3 | Black |
| 99 | WA | Mid | Running | 2 | 4 | Black |
| 128 | GA | Low | Basketball | 2 | 2 | Asian |
| 67 | MA | High | Running | 4 | 1 | White |
| 43 | WA | Mid | Basketball | 3 | 1 | White |

# k-anonymity through generalisation



| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| **nid** | **gen(hometown)** | **gen(education)** | **hobbies** | **n_accept** | **n_reject** | **race** |
| 45 | GA | Low | Basketball | 6 | 1 | White |
| 245 | MA | High | Running | 4 | 2 | Latino |
| 23 | WA | Mid | Running | 10 | 2 | White |
| 78 | GA | Low | Basketball | 2 | 4 | Black |
| 92 | MA | High | Running | 6 | 0 | White |
| 12 | WA | Mid | Basketball | 3 | 3 | Black |
| 99 | WA | Mid | Running | 2 | 4 | Black |
| 128 | GA | Low | Basketball | 2 | 2 | Asian |
| 67 | MA | High | Running | 4 | 1 | White |
| 43 | WA | Mid | Basketball | 3 | 1 | White |

$k$=2

# k-anonymity through suppression

| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| **nid** | **gen(hometown)** | **gen(education)** | **hobbies** | **n_accept** | **n_reject** | **race** |
| 45 | GA | Low | * | 6 | 1 | White |
| 245 | MA | High | * | 4 | 2 | Latino |
| 23 | WA | Mid | * | 10 | 2 | White |
| 78 | GA | Low | * | 2 | 4 | Black |
| 92 | MA | High | * | 6 | 0 | White |
| 12 | WA | Mid | * | 3 | 3 | Black |
| 99 | WA | Mid | * | 2 | 4 | Black |
| 128 | GA | Low | * | 2 | 2 | Asian |
| 67 | MA | High | * | 4 | 1 | White |
| 43 | WA | Mid | * | 3 | 1 | White |

EPFL

# k-anonymity through suppression



| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
| --- | --- | --- | --- | --- | --- | --- |
| nid | gen(hometown) | gen(education) | hobbies | n_accept | n_reject | race |
| 45 | GA | Low | * | 6 | 1 | White |
| 245 | MA | High | * | 4 | 2 | Latino |
| 23 | WA | Mid | * | 10 | 2 | White |
| 78 | GA | Low | * | 2 | 4 | Black |
| 92 | MA | High | * | 6 | 0 | White |
| 12 | WA | Mid | * | 3 | 3 | Black |
| 99 | WA | Mid | * | 2 | 4 | Black |
| 128 | GA | Low | * | 2 | 2 | Asian |
| 67 | MA | High | * | 4 | 1 | White |
| 43 | WA | Mid | * | 3 | 1 | White |

$k$=3

EPFL

# $k$-anonymity
# Privacy… And Utility?



- Given a table D, find a table D' such that
  - D' satisfies the $k$-anonymity condition
  - D' has the maximum utility (minimum information loss)

- NP-hard problem.
- Some heuristics exist for some utility metrics.

# Actually… For what Airbnb wants

| Masked identifier | Quasi-identifier | | | | | Sensitive attribute |
|---|---|---|---|---|---|---|
| nid | hometown | education | hobbies | n_accept | n_reject | race |
| 45 | Athens/GA | None | Basketball | 6 | 1 | White |
| 245 | Boston/MA | PhD | Running | 4 | 2 | Latino |
| 23 | Seattle/WA | BSc | Running | 10 | 2 | White |
| 78 | Atlanta/GA | High School | Basketball | 2 | 4 | Black |
| 92 | Cambridge/MA | MSc | Running | 6 | 0 | White |
| 12 | Redmond/WA | BSc | Basketball | 3 | 3 | Black |
| 99 | Seattle/WA | BSc | Running | 2 | 4 | Black |
| 128 | Macon/GA | High School | Basketball | 2 | 2 | Asian |
| 67 | Salem/MA | MSc | Running | 4 | 1 | White |
| 43 | Spokane/WA | BSc | Basketball | 3 | 1 | White |

# Actually… For what Airbnb wants

| Masked identifier | Quasi-identifier | | | Quasi-identifier | | Sensitive attribute |
|---|---|---|---|---|---|---|
| **nid** | [redacted] | [redacted] | [redacted] | **n_accept** | **n_reject** | **race** |
| 45 | [redacted] | [redacted] | [redacted] | 6 | 1 | White |
| 245 | [redacted] | [redacted] | [redacted] | 4 | 2 | Latino |
| 23 | [redacted] | [redacted] | [redacted] | 10 | 2 | White |
| 78 | [redacted] | [redacted] | [redacted] | 2 | 4 | Black |
| 92 | [redacted] | [redacted] | [redacted] | 6 | 0 | White |
| 12 | [redacted] | [redacted] | [redacted] | 3 | 3 | Black |
| 99 | [redacted] | [redacted] | [redacted] | 2 | 4 | Black |
| 128 | [redacted] | [redacted] | [redacted] | 2 | 2 | Asian |
| 67 | [redacted] | [redacted] | [redacted] | 4 | 1 | White |
| 43 | [redacted] | [redacted] | [redacted] | 3 | 1 | White |

# k-anonymise

| nid | n_accept | n_reject |
|---|---|---|
| 45 | 6 | 1 |
| 245 | 4 | 2 |
| 23 | 10 | 2 |
| 78 | 2 | 4 |
| 92 | 6 | 0 |
| 12 | 3 | 3 |
| 99 | 2 | 4 |
| 128 | 2 | 2 |
| 67 | 4 | 1 |
| 43 | 3 | 1 |

**Group similar entries**

| nid | n_accept | n_reject |
|---|---|---|
| 45, 92 | 6 | [0,1] |
| 245, 67 | 4 | [1,2] |
| 23 | 10 | 2 |
| 78,99,128 | 2 | [2,4] |
| 12,43 | 3 | [1,3] |

# k-anonymise

| nid | n_accept | n_reject |
|---|---|---|
| 45 | 6 | 1 |
| 245 | 4 | 2 |
| 23 | 10 | 2 |
| 78 | 2 | 4 |
| 92 | 6 | 0 |
| 12 | 3 | 3 |
| 99 | 2 | 4 |
| 128 | 2 | 2 |
| 67 | 4 | 1 |
| 43 | 3 | 1 |

**Group similar entries**

| nid | n_accept | n_reject |
|---|---|---|
| 45, 92 | 6 | [0,1] |
| 245, 67 | 4 | [1,2] |
| 23 | 10 | 2 |
| 78,99,128 | 2 | [2,4] |
| 12,43 | 3 | [1,3] |

**Suppress the outlier**
**Take mean for rest**

| nid | n_accept | n_reject |
|---|---|---|
| 45, 92 | 6 | 0.5 |
| 245, 67 | 4 | 1.5 |
| 78,99,128 | 2 | 2.66 |
| 12,43 | 3 | 2 |

# k-anonymise

$k$=2

| nid | n_accept | n_reject |
|---|---|---|
| 45 | 6 | 0.5 |
| 245 | 4 | 1.5 |
| | | |
| 78 | 2 | 2.66 |
| 92 | 6 | 0.5 |
| 12 | 3 | 2 |
| 99 | 2 | 2.66 |
| 128 | 2 | 2.66 |
| 67 | 4 | 1.5 |
| 43 | 3 | 2 |

| nid | n_accept | n_reject |
|---|---|---|
| 45, 92 | 6 | [0,1] |
| 245, 67 | 4 | [1,2] |
| 23 | 10 | 2 |
| 78,99,128 | 2 | [2,4] |
| 12,43 | 3 | [1,3] |

**Suppress the outlier**
**Take mean for rest**

| nid | n_accept | n_reject |
|---|---|---|
| 45, 92 | 6 | 0.5 |
| 245, 67 | 4 | 1.5 |
| 78,99,128 | 2 | 2.66 |
| 12,43 | 3 | 2 |

# k-anonymise

$k$=2

**Sensitive attribute**

| nid | n_accept | n_reject | race |
|---|---|---|---|
| 45 | 6 | 0.5 | White |
| 245 | 4 | 1.5 | Latino |
| | | | White |
| 78 | 2 | 2.66 | Black |
| 92 | 6 | 0.5 | White |
| 12 | 3 | 2 | Black |
| 99 | 2 | 2.66 | Black |
| 128 | 2 | 2.66 | Asian |
| 67 | 4 | 1.5 | White |
| 43 | 3 | 2 | White |

We still learn that:

45 and 92 (users with 6 accepts) are **White**

78, 99, and 128 (users with 2 accepts) **aren't White**

# k-anonymity
# Privacy… Not guaranteed

Equivalence class

| gender | zipcode | problem |
|---|---|---|
| * | 1012 | Cancer |
| * | 100* | Heart Disease |
| * | 100* | Heart Disease |
| * | 100* | Heart Disease |
| * | 100* | Heart Disease |
| * | 100* | Heart Disease |
| * | 1012 | Cancer |
| * | 1012 | Herpes |
| * | 1012 | Flu |

Does not provide privacy when sensitive values lack **diversity !**

Example: anyone in the database with zipcode 100* is known to have a heart disease

# $\ell$-diversity



- An equivalence class has $\ell$-diversity if there are at least $\ell$ **well-represented values for the sensitive attribute.**
- A dataset has $\ell$-diversity if every equivalence class has $\ell$–diversity.

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476** | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach cancer |
| 4 | 4790* | ≥ 40 | 6K | gastritis |
| 5 | 4790* | ≥ 40 | 11K | flu |
| 6 | 4790* | ≥ 40 | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | Stomach cancer |

A 3-diverse hospital records **dataset**

**Disclaimer:** Airbnb does not use l-diversity. It uses p-sensitivity, a related relaxed notion.
https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

# $\ell$-diversity - Limitations

$\ell$-diversity does **not consider semantics** of sensitive values

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476** | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach cancer |
| 4 | 4790* | ≥ 40 | 6K | gastritis |
| 5 | 4790* | ≥ 40 | 11K | flu |
| 6 | 4790* | ≥ 40 | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | Stomach cancer |

All patients in this equivalence class have stomach issues

# $\ell$-diversity - Limitations

$\ell$ -diversity does **not consider distribution** of sensitive values

## Original dataset

| | |
|---|---|
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Flu |

99% have cancer

## Anonymization A

| | |
|---|---|
| Q1 | Flu |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |

Q1: 423**, >60
Q2: 423**, <60

## Anonymization B

| | |
|---|---|
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Flu |
| Q2 | Flu |

# $\ell$-diversity - Limitations

$\ell$-diversity does **not consider distribution** of sensitive values

## Original dataset

| | |
|---|---|
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Flu |

99% have cancer

## Anonymization A

| | |
|---|---|
| Q1 | Flu |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Ca[illegible] |
| [illegible] | [illegible] |
| [illegible] | [illegible] |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |

50% cancer $\Rightarrow$ quasi-identifier group is "diverse"

**BUT: Leaks a ton of information about Q1**

Q1: 423**, >60
Q2: 423**, <60

## Anonymization B

| | |
|---|---|
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Flu |
| Q2 | Flu |

# $\ell$-diversity - Limitations

$\ell$-diversity does **not consider distribution** of sensitive values

## Original dataset

| | |
|---|---|
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Flu |

99% have cancer

## Anonymization A

| | |
|---|---|
| Q1 | Flu |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Ca |
| [illegible] | [illegible] |
| [illegible] | [illegible] |
| Q2 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |

Q1: 423**, >60
Q2: 423**, <60

50% cancer ⇒ quasi-identifier group is "diverse"
**BUT: Leaks a ton of information about Q1**

## Anonymization B

| | |
|---|---|
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | C…cer |
| Q2 | Ca |
| Q2 | Ca |
| Q2 | Car |
| Q2 | Flu |

99% cancer ⇒ quasi-identifier group is <u>not</u> "diverse"
…yet anonymized database does not leak anything

# t-closeness

- An equivalence class has **t-closeness** if the **distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.**
- A dataset has t-closeness if all equivalence classes have t-closeness.

# So now we have privacy… Right?!

| Quasi-identifiers | | Sensitive | |
|---|---|---|---|
| Ethnicity | ZIP | HIV | Diagnosis |
| Caucasian | 787XX | HIV+ | Flu |
| Asian | 787XX | HIV- | Flu |
| Asian | 787XX | HIV+ | Herpes |
| Caucasian | 787XX | HIV- | Acne |
| Caucasian | 787XX | HIV- | Herpes |
| Caucasian | 787XX | HIV- | Acne |

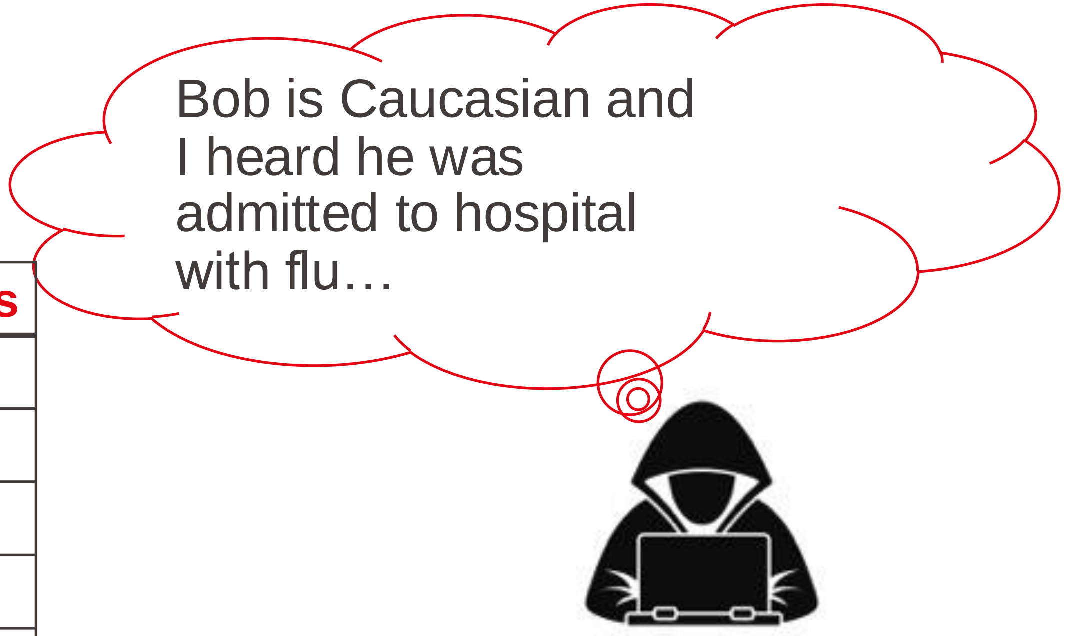
This table is k-anonymous, l-diverse and t-close…

…does it provide privacy?

# So now we have privacy… Right?!

~~Quasi-identifiers~~ | ~~Sensitive~~

| Ethnicity | ZIP | HIV | Diagnosis |
|---|---|---|---|
| Caucasian | 787XX | HIV+ | Flu |
| Asian | 787XX | HIV- | Flu |
| Asian | 787XX | HIV+ | Herpes |
| Caucasian | 787XX | HIV- | Acne |
| Caucasian | 787XX | HIV- | Herpes |
| Caucasian | 787XX | HIV- | Acne |

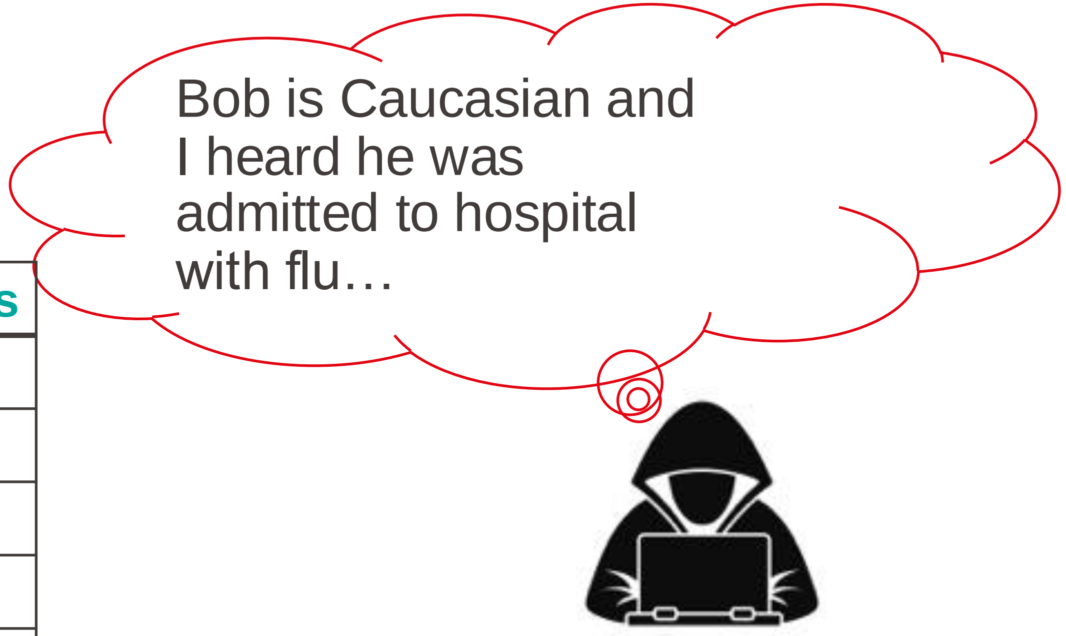


# So now we have privacy… Right?!

Quasi-identifiers

Sensitive

| Ethnicity | ZIP | HIV | Diagnosis |
|---|---|---|---|
| Caucasian | 787XX | HIV+ | Flu |
| Asian | 787XX | HIV- | Flu |
| Asian | 787XX | HIV+ | Herpes |
| Caucasian | 787XX | HIV- | Acne |
| Caucasian | 787XX | HIV- | Herpes |
| Caucasian | 787XX | HIV- | Acne |



# “De-identification” shall fail


**Adversary’s knowledge:** We cannot predict what **auxiliary data** may be available to the adversary

**+**

**The curse of dimensionality**: High-dimensional data is sparse. The more you know about individuals, the less likely it is that two individuals will look alike

**=**

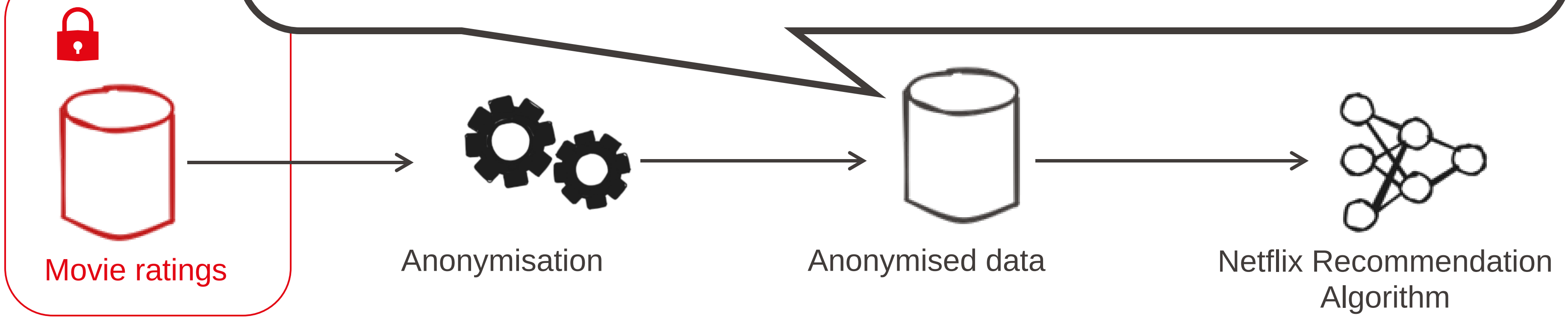
Supposedly anonymized  data can be re-identified with a **linkage attack**

# The curse of dimensionality

# “De-identification” shall fail
# Another real-life example



“***All customer identifying information has been removed****; **all that remains are ratings and dates.** This follows our privacy policy, which you can review here. Even if, for example, you knew all your own ratings and their dates you probably couldn’t identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. **Of course, since you know all your own ratings that really isn’t a privacy problem is it**?*”




| user | pulp_fiction | fight_club | the_minions |
|---|---|---|---|
| theresa | 3/5 | 3/5 | 5/5 |
| carmela | … | … | … |

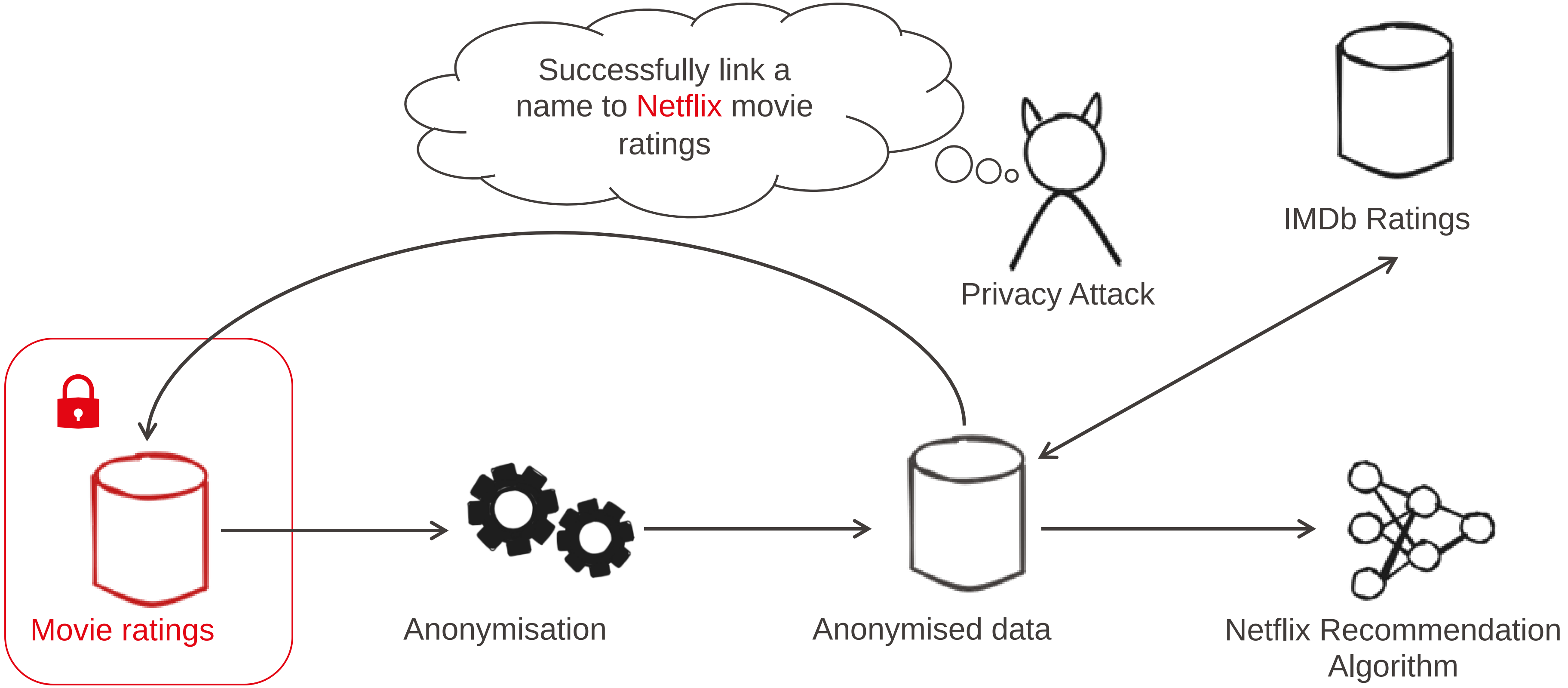
| pulp_fiction | fight_club | the_minions |
|---|---|---|
| 3/5 | 3/5 | 5/5 |
| … | … | … |

- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Narayanan and Shmatikov, 2008

# "De-identification" shall fail
## Another real-life example

| user | pulp_fiction | fight_club | the_minions |
|---|---|---|---|
| theresa | 3/5 | 3/5 | 5/5 |




| user | pulp_fiction | fight_club | the_minions |
|---|---|---|---|
| theresa | 3/5 | 3/5 | 5/5 |
| carmela | … | … | … |

| pulp_fiction | fight_club | the_minions |
|---|---|---|
| 3/5 | 3/5 | 5/5 |
| … | … | … |

- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

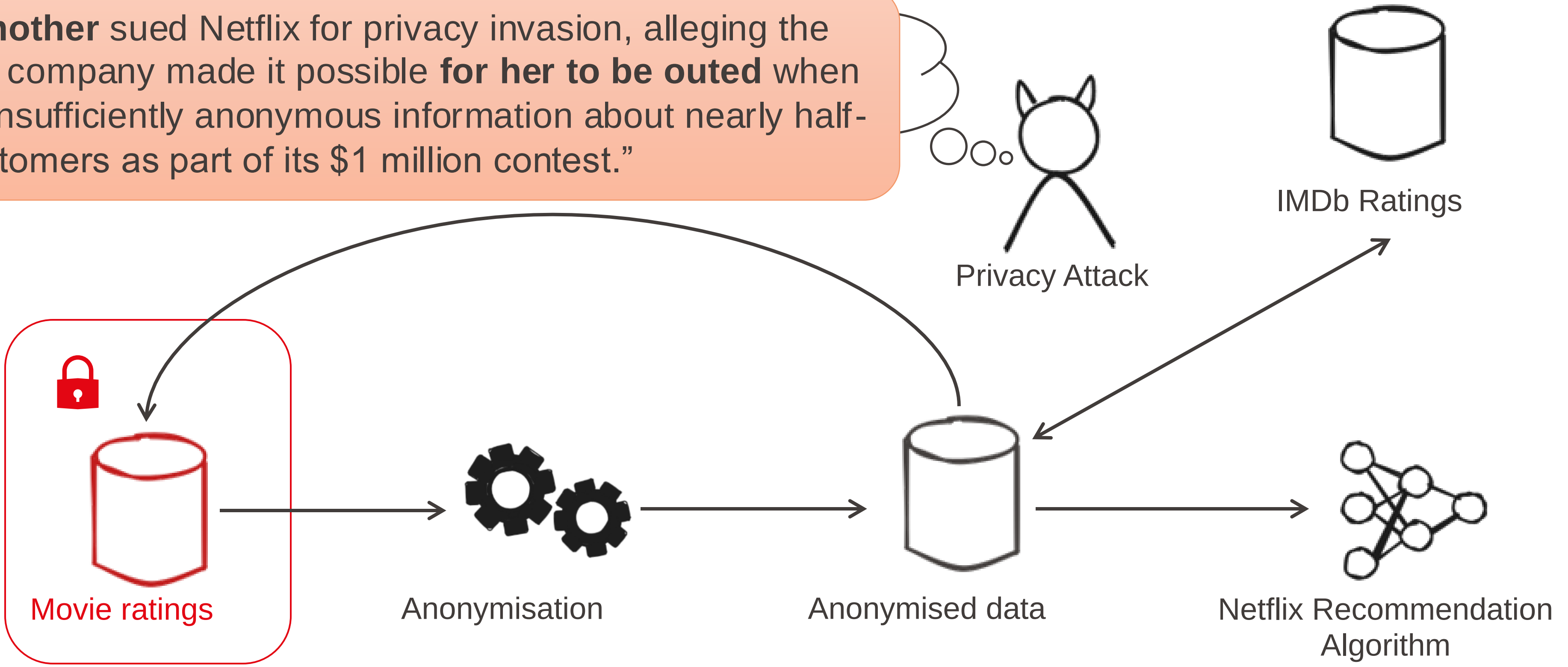
Narayanan and Shmatikov, 2008

# “De-identification” shall fail
# Another real-life example



| user | pulp_fiction | fight_club | the_minions |
|---|---|---|---|
| theresa | 3/5 | 3/5 | 5/5 |

“a **lesbian mother** sued Netflix for privacy invasion, alleging the movie-rental company made it possible **for her to be outed** when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest.”




RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

WIRED

**NetFlix Cancels Recommendation Contest After Privacy Lawsuit**

- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

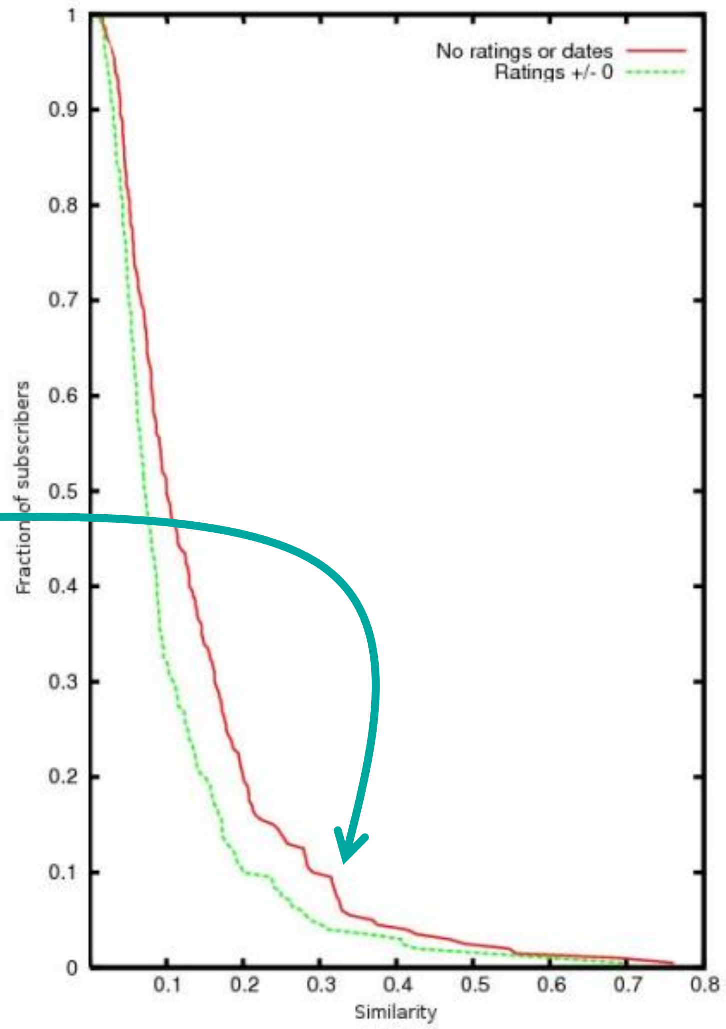
Narayanan and Shmatikov, 2008

# "De-identification" shall fail
# Another real-life example

The average record, has **NO** similar records

**Netflix prize dataset**: for 90% of the records there is no other record that is more than 30% similar (in the spirit of the cosine similarity)

Netflix applied "Perturbation": but utility must be preserved!




Narayanan and Shmatikov, 2008

# "De-identification" shall fail
# Another real-life example

*"With 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For 68%, two ratings and dates (with a 3-day error) are sufficient"*

- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

# “De-identification” shall fail
# Another real-life example



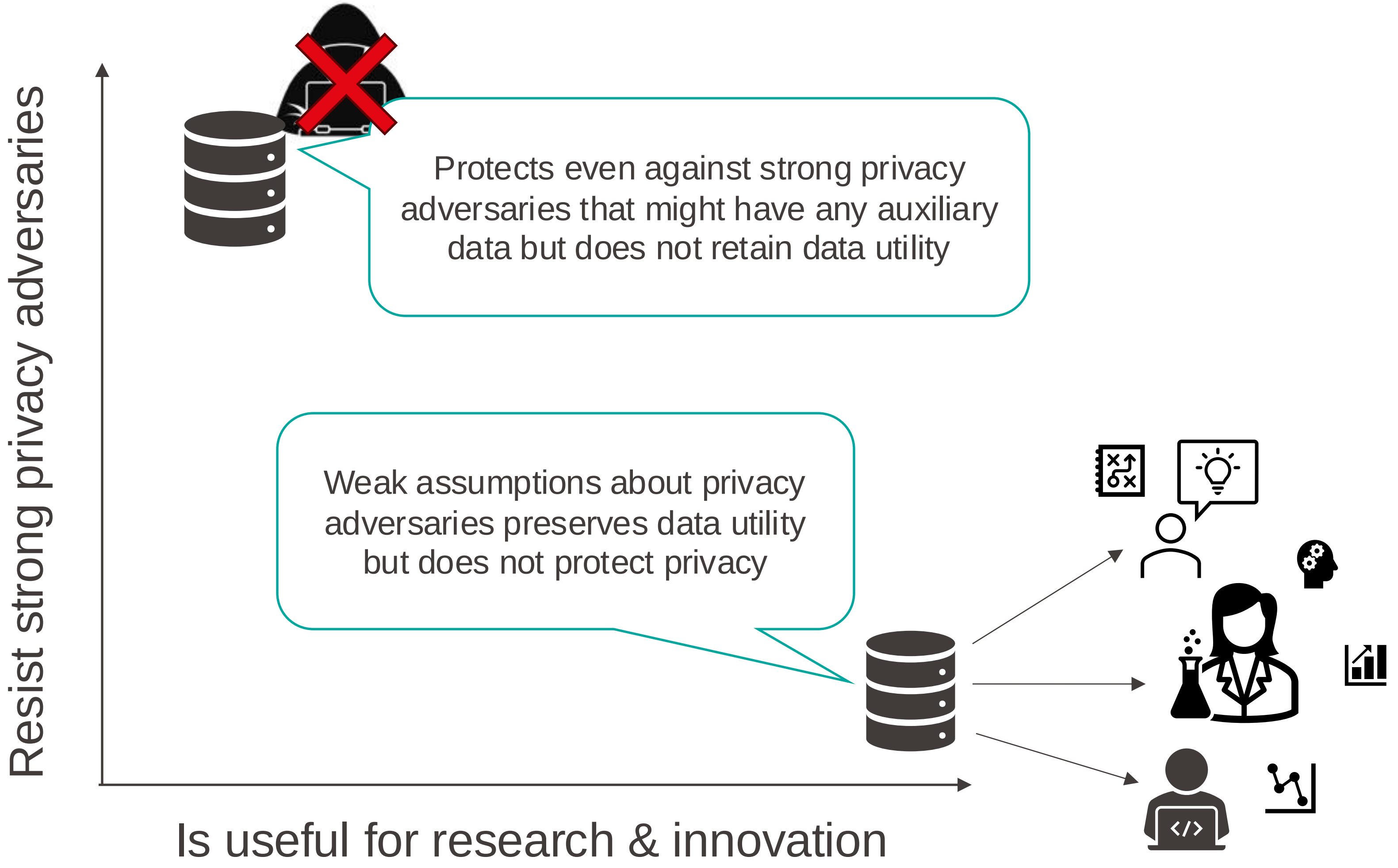
*“With 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For 68%, two ratings and dates (with a 3-day error) are sufficient”*

**Completely removing PII is not possible. PII has no technical definition, we do not know what will make someone identifiable. It all depends on the adversary’s knowledge**

- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

# Conclusions

# The privacy-utility trade-off
## Microdata publishing



EPFL

# So what about Airbnb…



- Airbnb has a **very** concrete goal
  - Needs very few columns, not so sparse – lightly hit by curse of dimensionality
  - Can handle quite some noise

- Airbnb not concerned about public adversaries (only internal)

- Airbnb left hard problems unsolved
  - e.g., removing identifying information in the photos they send to the research partner

    they call this de-identification of photos (what does this even mean?)

# Takeaways

- Data is a valuable asset but also contains a lot of sensitive information
  - When published or shared widely, it can lead to **significant harm for individuals**

- Privacy-preserving data publishing is an extremely hard problem
  - Whenever we remove information to prevent privacy attacks, we also loose this information for utility purposes
  - Best chance we have at solving the problem is for small datasets with very well defined utility function

- Primary challenge is that we cannot predict an adversary’s background knowledge
  - The more high-dimensional the data is the harder this problem becomes