# Data privacy: Motivation

Vicenç Torra

February, 2018

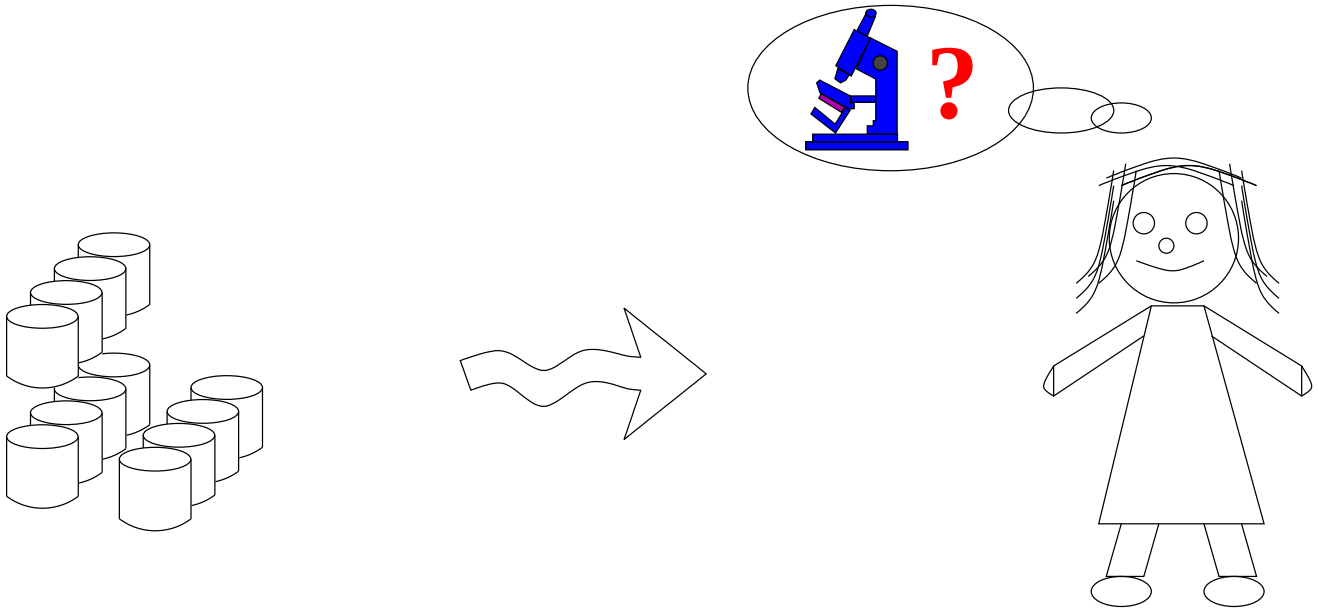
SAIL + PICS, School of Informatics, University of Skövde, Sweden

# Outline

# Motivation

# Introduction

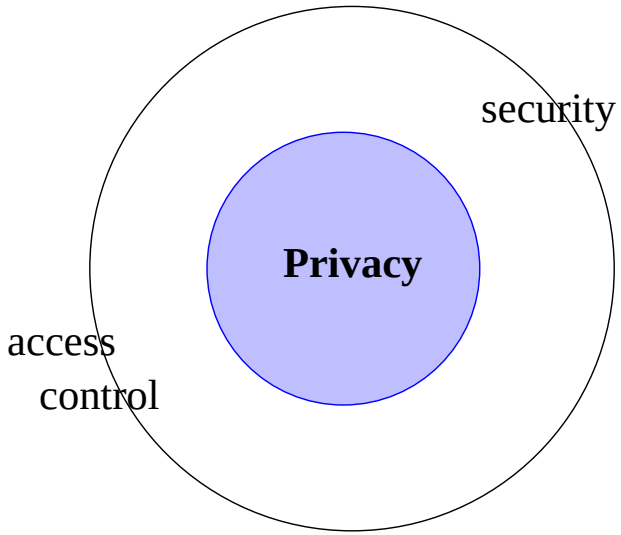
- Data privacy: core
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.




E.g., you are authorized to compute the average stay in a hospital, but maybe you are not authorized to see the length of stay of your neighbor.

# Introduction

- Data privacy: boundaries
  - Database in a computer or in a removable device
    $\Rightarrow$ access control to avoid unauthorized access
    $\Longrightarrow$ Access to address (admissions), Access to blood test (admissions?)
  - Data is transmitted
    $\Rightarrow$ security technology to avoid unauthorized access
    $\Longrightarrow$ Data from blood glucose meter sent to hospital. Network sniffers
    Transmission is sensitive: Near miss/hit report to car manufacturers



# Motivation

- Legislation. Privacy a fundamental right. (Ch. 1.1)
  - Universal Declaration of Human Rights (UN). European Convention on Human Rights (Council of Europe). General Data Protection Regulation (EU). National regulations.
- Companies own interest.
  - Competitors can take advantage of information.
- Avoiding privacy breach. Several well known cases (see later).

# Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[1]

  Names, Age, Sex, Occupation, Place of birth, Last place of residence, Yes/No, condition (healthy?)

LIST OF PASSENGERS.

REPORT AND LIST OF PASSENGERS taken on board the S.S. Hooper at London.

| NAMES | AGE | SEX | OCCUPATION | PLACE OF BIRTH | Last Place of Residence | CONDITION |
|---|---|---|---|---|---|---|
| [illegible] | 11 | Male | Gentleman | Aberdeen Eng | Chelmsford Eng | Healthy |
| [illegible] | 24 | | Lawyer's Clerk | Madras India | Saltburn Eng | |
| [illegible] | 21 | | Tailor | London | London | |
| [illegible] | 26 | | Surgeon | London | London | |
| [illegible] | 28 | | Gentleman | Boston U.S. | Boston U.S. | |
| [illegible] | 30 | | | Scotland | Boston U.S. | |
| [illegible] | 16 | | | London (?) | London | |
| [illegible] | 44 | | | Essex Eng | Essex Eng | |
| [illegible] | 28 | | | Alnwick Northd | Boston U.S. | |
| [illegible] | 28 | | Cattle Dealer | Boston U.S. | Boston U.S. | |

[1] https://www.sec.state.ma.us/arc/arcgen/genidx.htm

# Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - ⋆ 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

# Difficulties

- Difficulties: highly identifiable data
  - Data from mobile devices:
    - ⋆ two positions can make you unique (home and working place)
  - AOL[2] and Netflix cases (search logs and movie ratings)
    ⇒ User No. 4417749, hundreds of searches over a three-noth period including queries 'landscapers in Lilburn, Ga' ⇒ Thelma Arnold identified!
    ⇒ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...
    (i.e., high dimensional data)

[2] http://www.nytimes.com/2006/08/09/technology/09aol.html

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?
  - Example #2:
    - ⋆ Car company goal: Study driving behaviour in the morning
    - ⋆ Data: First drive (GPS origin + destination, time) $\times$ 30 days
    - ⋆ No "personal data", is this ok?

# Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?
  - Example #2:
    - ⋆ Car company goal: Study driving behaviour in the morning
    - ⋆ Data: First drive (GPS origin + destination, time) $\times$ 30 days
    - ⋆ No "personal data", is this ok?
    - ⋆ NO!!!: How many (cars) go from your parking to your university everymorning ? Are you exceeding the speed limit ? Are you visiting a psychiatrisc every tuesday ?

# Difficulties

- Data privacy is "impossible", or not ?
  - Privacy vs. utility
  - Privacy vs. security
  - Computationally feasible