# Foundations of Privacy

Class 1

# The teachers of the course

**Kostas Chatzikokolakis**
CNRS & Ecole Polytechnique

**Catuscia Palamidessi**
INRIA & Ecole Polytechnique

# Logistic Information

- The course will be in English
- We will put the slides on line before every class
- There will be a written exam at the end of the course (on November 28)
- We will give exercises during the course, leave you some time to solve them, and then show the solution. You should try to solve them, as they will help to prepare for the exam
- Please feel free to ask questions any time. We are very happy when people ask questions, as they help to make the class more interactive and lively

# Plan of the lectures

- Motivations, a bit of history, main problems, research directions (3 hours)
- Differential Privacy and Extensions (6 hours)
- Local Differential Privacy (3 hours)
- Location Privacy (3 hours)
- Quantitative Information Flow (9 hours)

# Motivations

In the "Information Society", each individual constantly leaves digital traces of his actions that may allow to infer a lot of information about himself

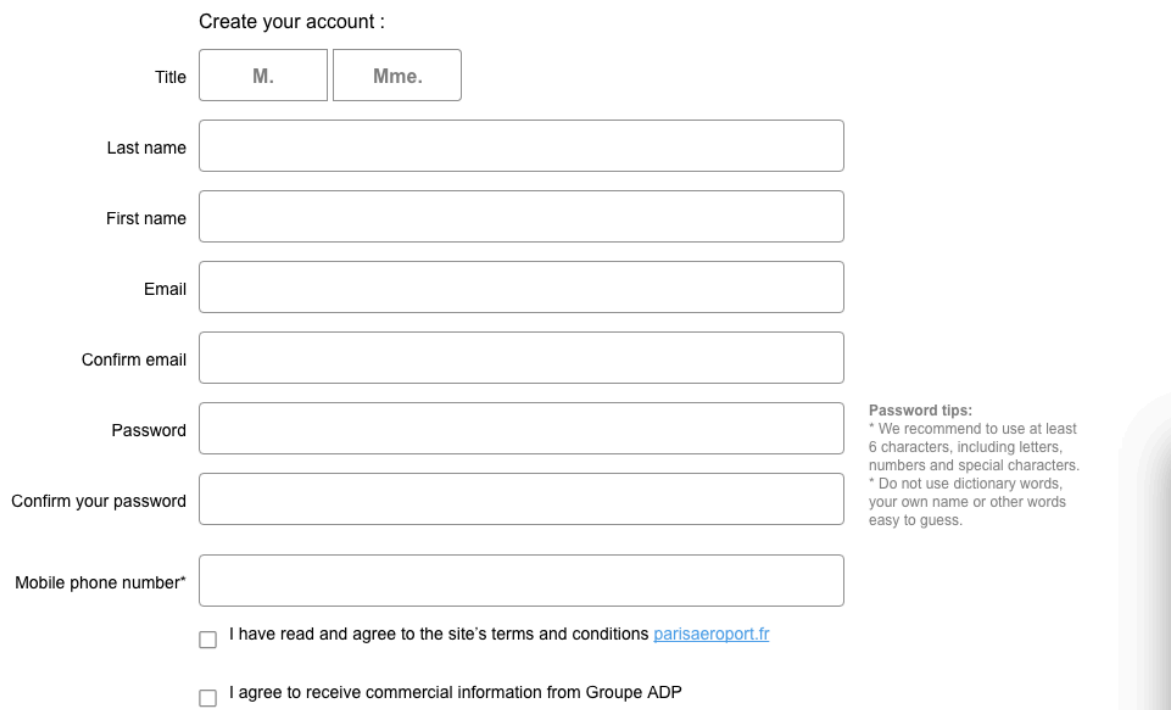
Request to a LBS ⇒ location.

History of requests ⇒ interests.

Activity in social networks ⇒ political opinions, religion, hobbies, . . .

Power consumption (smart meters) ⇒ activities at home.

# Example:
## Personal information in exchange of a service




-We don't know how our information will be used

-The "right to be forgotten" is  very difficult to enforce

# Concerns about privacy

Risk: collect and use of digital traces for fraudulent purposes.

Examples: targeted spam, identity theft, profiling, discrimination, …

The news are full of problems caused by privacy breaches

The need for privacy is intrinsic to the human nature, although it varies a lot from individual to individual, between cultures, and it evolves with time

Privacy is recognized as one of the fundamental right of individuals:

- Universal Declaration of the Human Rights at the assembly of the United Nations (Article 12), 1948.
- European Directive 95/46/EC on the Protection of Personal Data (currently being revised towards a stricter regulation).
- Japanese Act on the Protection of Personal Information from 2003 (current discussions to amend it and make stricter).

# The new European regulation (will be enforced starting from 2018)


## What will be the key changes?

- A **'right to be forgotten'** will help you manage data protection risks online. When you no longer want your data to be processed and there are no legitimate grounds for retaining it, the data will be deleted. The rules are about empowering individuals, not about erasing past events, re-writing history or restricting the freedom of the press.
- **Easier access to your own personal data.**
- **A right to transfer personal data** from one service provider to another.
- When your **consent is required, you must be asked to give it by means of a clear affirmative action.**
- More transparency about how your data is handled, with **easy-to-understand information**, especially for **children**.
- Businesses and organisations will need to **inform you about data breaches** that could adversely affect you **without undue delay**. They will also have to notify the relevant data protection supervisory authority.
- Better enforcement of data protection rights through improved **administrative and judicial remedies** in cases of violations
- Increased **responsibility and accountability** for those processing personal data – through **data protection risk assessments, data protection officers**, and the principles of **'data protection by design'** and **'data protection by default'**.

# Different types of sensitive data

- Sensitive information about an individual :
  - credit card / bank information, home access code, passwords, …
    - sensitive because it can be used to **attack the person or his property**
  - ethnicity, religious beliefs, political opinions, medical status, intimate videos, …
    - Sensitive because it can lead to **discrimination** or **public shame**.
- Identification information : information that can uniquely identify an individual.
  - First and last name, social security number, physical and email address, phone number, biometric data (such as fingerprint and DNA), …
    - Sensitive because it can be used for **identity theft**, to cross-reference databases, or to identify him as the subject of certain actions
- Sensitive information for organizations
  - Industries: production plans, research, strategies, …
  - Governments, police, army, …
- In this course, we will try to encompass the various scenario. We will abstract from the nature of the sensitive information whenever possible, and present the common principles of information protection, but we will also show that the kind of information (and of adversary) induces differences in the approach.

# Why it is difficult to protect privacy

- Traditionally, privacy is protected via:
  - Anonymization
  - Encryption
  - Access control

- However, these methods often fail:
  - encryption and access control cannot protect against the inference of private information from public information
  - anonymization has been proved highly ineffective

# The problem

- In general, the problem of privacy is to protect the disclosure of **sensitive information** of individuals when a collection of data about these individuals (*dataset*) is made **publicly available**
- The process of transforming the dataset in order to avoid such disclosure is called **sanitization**

# Privacy via anonymity

Nowadays, many institutions and companies that collect data use anonymization, i.e., they remove all personal identifiers: name, address, SSN, …



"We don't have any raw data on the identifiable individual. Everything is anonymous"
(CEO of NebuAd, a U.S. company that offers targeted advertising based on browsing histories)

Similar practices are used by Facebook, MySpace, Twitter, …

# Privacy via anonymity

However, anonymity-based sanitization has been shown to be **highly ineffective:** Several de-anonymization attacks have been carried out in the last decade

- The quasi-identifiers allow to retrieve the identity in a large number of cases.
- More sophisticated methods (k-anonymity, $\ell$-diversity, …) take care of the quasi-identifiers, but they are still prone to composition attacks

# Famous deanonymization attacks (I)

- In 2006, AOL Research released a text file containing twenty million search keywords for over 650,000 users, intended for research purposes.
- The file was anonymized (names where substituted by numbers as pseudonyms), but personally identifiable information was present in many of the queries. The NYT was able to locate an individual from the search records by cross referencing them with phonebook listings
- **From the report:** The subject conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 y.o. single men" to "dog that urinates on everything.", "landscapers in Lilburn, Ga", several people with the last name Arnold and "homes sold in shadow lake". It did not take much to identify the subject as Thelma Arnold, a 62-year-old widow with three dogs who lives in Lilburn, Ga.

# Naive anonymization

- This is the most obvious solution: remove the identity of individuals from the database, so that the sensitive information cannot be directly linked to the individual

- Example: assume that we have a medical database, where the sensitive information is disease that has been diagnosed
- For instance, Jorah Mormont may not want to reveal that he is affected by greyscale.

| | Name | age | Disease |
|---|---|---|---|
| 1 | Jon Snow | 30 | cold |
| 2 | Jamie Lannister | 39 | amputed hand |
| 3 | Arya Stark | 16 | stomac ache |
| 4 | Bran Stark | 14 | crippled |
| 5 | Sandor Clegane | 45 | ignifobia |
| 6 | Jorah Mormont | 48 | gleyscale |
| 7 | Eddad Stark | 32 | headache |
| 8 | Ramsay Bolton | 32 | psychopath |
| 9 | Daenerys Targaryen | 25 | mania of grandeur |

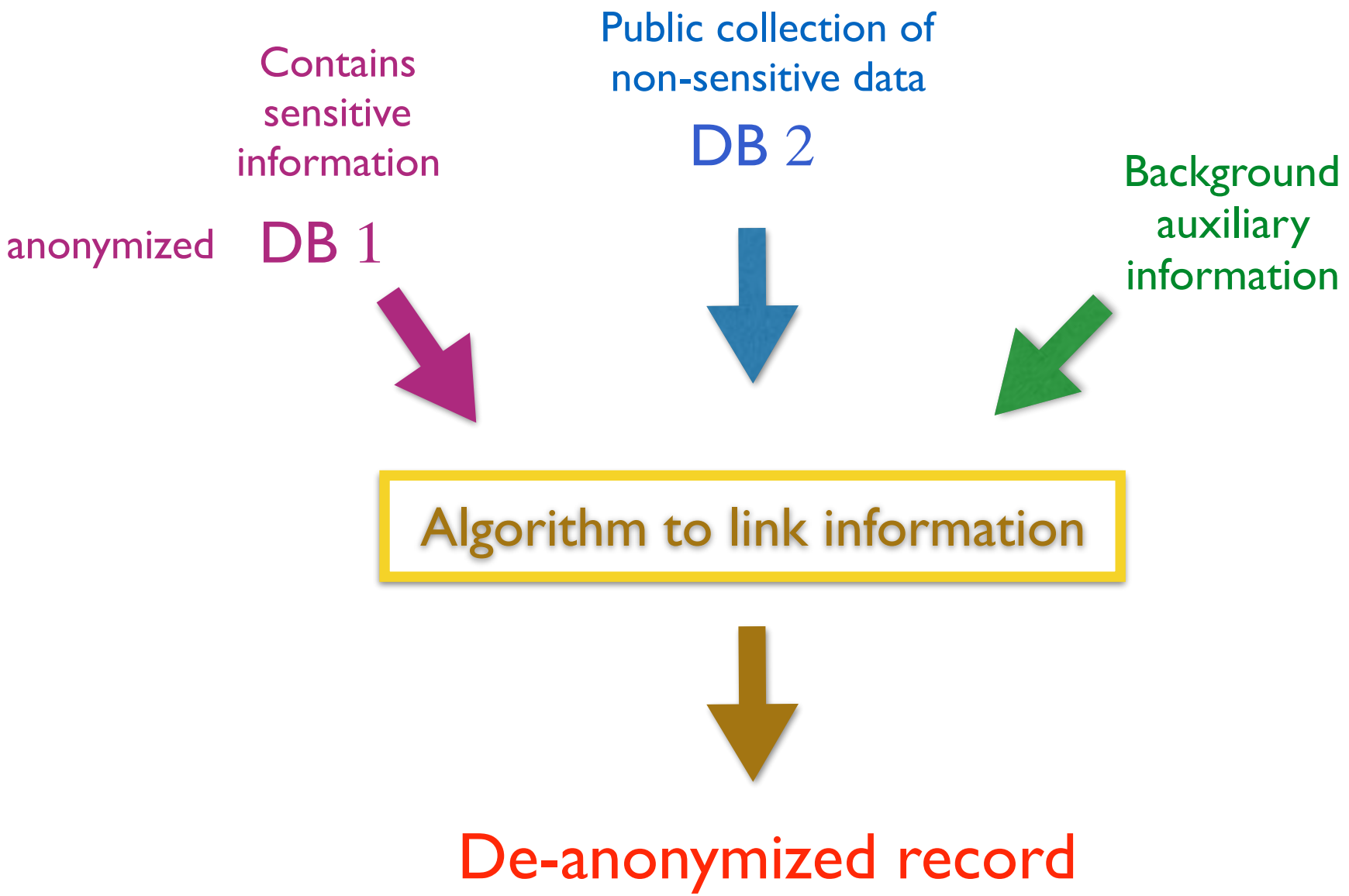
# Naive anonymization

- Anonymization removes the column of the name, so that, for instance, the grayscale disease cannot be directly linked to Jorah Mormont

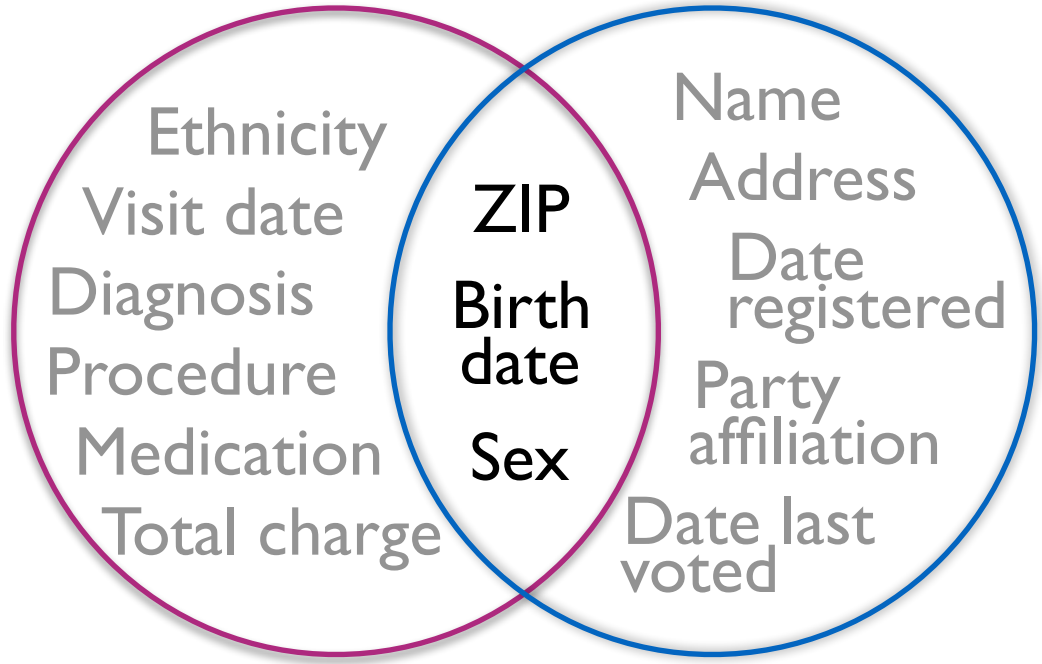
- Hystorically the first method, still used nowadays
- However, this solution has been (already several years ago) shown to be very weak and prone to de-anonymization attacks

| | Name | age | Disease |
|---|---|---|---|
| 1 | - | 30 | cold |
| 2 | - | 39 | amputed hand |
| 3 | - | 16 | stomac ache |
| 4 | - | 14 | crippled |
| 5 | - | 45 | ignifobia |
| 6 | - | 48 | gleyscale |
| 7 | - | 32 | headache |
| 8 | - | 32 | psychopath |
| 9 | - | 25 | mania of grandeur |

# Sweeney's de-anonymization attack by linking



# Sweeney's de-anonymization attack by linking




87 % of US population is uniquely identifiable by 5-digit ZIP, gender, DOB

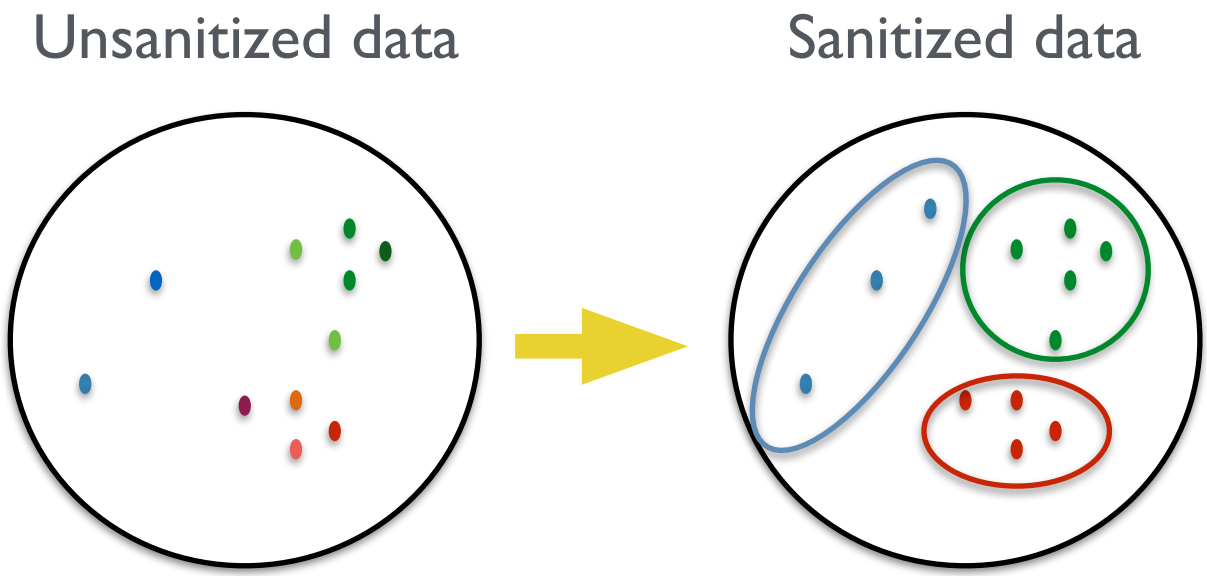
This attack has lead to the proposal of k-anonymity (that I will present later)

# K-anonymity [Sweeney and Samarati, 2000]

- Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals
- Make every record in the table indistinguishable from a least $k-1$ other records with respect to quasi-identifiers. This can be done by:
  - suppression of attributes, and/or
  - generalization of attributes, and/or
  - addition of dummy records
- Linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

# Principle: group anonymity

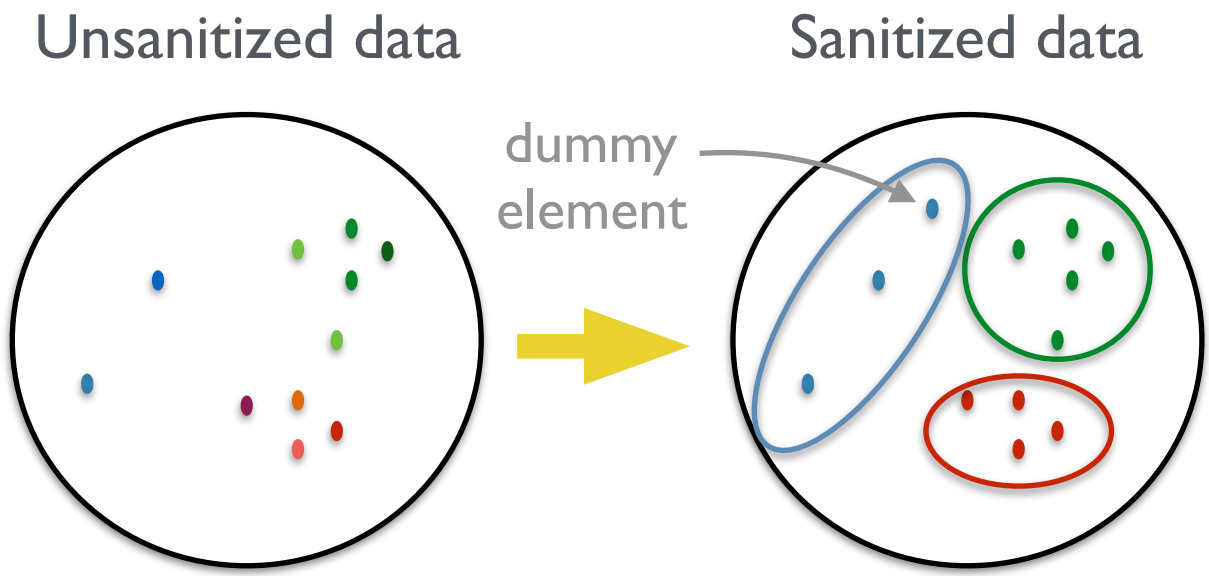
- Ensure that each individual is indistinguishable within a group by removing individual differences




- Of course, the larger are the groups, the better the individuals are protected (within the group)
- k-anonymity ensure that the size of each group is at least k

# Principle: group anonymity

- Ensure that each individual is indistinguishable within a group by removing individual differences




- Of course, the larger are the groups, the better the individuals are protected (within the group)
- k-anonymity ensure that the size of each group is at least k

# K-anonymity

**Example:** 4-anonymity w.r.t. the quasi-identifiers (nationality, ZIP, age)

- achieved by suppressing the nationality and generalizing ZIP and age

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease |
| 3 | 13068 | 21 | Japanese | Viral Infection |
| 4 | 13053 | 23 | American | Viral Infection |
| 5 | 14853 | 50 | Indian | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection |
| 9 | 13053 | 31 | American | Cancer |
| 10 | 13053 | 37 | Indian | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer |
| 12 | 13068 | 35 | American | Cancer |

Figure 1. Inpatient Microdata

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip Code | Age | Nationality | Condition |
| 1 | 130** | < 30 | * | Heart Disease |
| 2 | 130** | < 30 | * | Heart Disease |
| 3 | 130** | < 30 | * | Viral Infection |
| 4 | 130** | < 30 | * | Viral Infection |
| 5 | 1485* | ≥ 40 | * | Cancer |
| 6 | 1485* | ≥ 40 | * | Heart Disease |
| 7 | 1485* | ≥ 40 | * | Viral Infection |
| 8 | 1485* | ≥ 40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

Figure 2. 4-anonymous Inpatient Microdata

# Problems with k-anonymity

- **Problem**: in the sanitized dataset, all the individual in a group may the same value for the sensitive data
- Clearly, the people in that group are not protected from the revelation of their disease
- **Example:** suppose that John's employer knows that John is less than 40, that he lives in a town with ZIP code 12032, and that he visits the hospital. He can learn that John has cancer.

| | Non-Sensitive | | | | Sensitive |
|---|---|---|---|---|---|
| | Rase | Age | Sex | Zip Code | Disease |
| 1 | * | < 40 | * | 120** | Cancer |
| 2 | * | < 40 | * | 120** | Cancer |
| 3 | * | < 40 | * | 120** | Cancer |
| 4 | * | < 40 | * | 120** | Cancer |
| 5 | * | ≥ 50 | * | 151** | Hemophilia |
| 6 | * | ≥ 50 | * | 151** | Cancer |
| 7 | * | ≥ 50 | * | 151** | Virus |
| 8 | * | ≥ 50 | * | 151** | Virus |
| 9 | * | 4* | * | 120** | Hemophilia |
| 10 | * | 4* | * | 120** | Hemophilia |
| 11 | * | 4* | * | 120** | Virus |
| 12 | * | 4* | * | 120** | Virus |

Table 2: 4-anonymous inpatient microdata.

# $\ell$-diversity [Kifer et al., 2007]

- A solution: $\ell$-diversity.
- The idea is to form the groups in such a way that each group contains a variety of values for the sensitive data
- It's computationally heavy: To find the optimal solution is a combinatorial problem with exponential complexity

| | Non-Sensitive | | | | Sensitive |
|---|---|---|---|---|---|
| | Rase | Age | Sex | Zip Code | Disease |
| 1 | * | $\leq 50$ | * | 120** | Cancer |
| 2 | * | $\leq 50$ | * | 120** | Cancer |
| 9 | * | $\leq 50$ | * | 120** | Hemophilia |
| 11 | * | $\leq 50$ | * | 120** | Virus |
| 5 | * | $> 50$ | * | 151** | Hemophilia |
| 6 | * | $> 50$ | * | 151** | Cancer |
| 7 | * | $> 50$ | * | 151** | Virus |
| 8 | * | $> 50$ | * | 151** | Virus |
| 3 | * | $\leq 50$ | * | 120** | Cancer |
| 4 | * | $\leq 50$ | * | 120** | Cancer |
| 10 | * | $\leq 50$ | * | 120** | Hemophilia |
| 12 | * | $\leq 50$ | * | 120** | Virus |

Table 5: 3-diverse table

# t-closeness

- Also the $\ell$-diversity has problems, though:
  - the requirement of $\ell$-diversity may be too strict (for instance, certain values of the disease, like having a cold, may not need to be protected)
  - the requirement of $\ell$-diversity may not be enough. For instance, if **almost all individuals** in a certain group have cancer, the attacker will infer that information (for a given individual in the group) with high probability
- To amend these problems, the t**-closeness** requirement was proposed: the idea is that the grouping is done in such a way that the distribution in each group is close to the general distribution

# Problems with k-anonymity and similar methods

- **Composition attacks**
  - Combination of knowledge coming from different sources (linking attacks)
  - Open world: Even if present data are protected, in the future there may be some new knowledge available

- **Everything can turn out to be a quasi-identifier**
  - Especially in high-dimensional and sparse databases.

Question: suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <30 | * | AIDS |
| 2 | 130** | <30 | * | Heart Disease |
| 3 | 130** | <30 | * | Viral Infection |
| 4 | 130** | <30 | * | Viral Infection |
| 5 | 130** | ≥40 | * | Cancer |
| 6 | 130** | ≥40 | * | Heart Disease |
| 7 | 130** | ≥40 | * | Viral Infection |
| 8 | 130** | ≥40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

(a)

| | Non-Sensitive | | | Sensitive |
|---|---|---|---|---|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <35 | * | AIDS |
| 2 | 130** | <35 | * | Tuberculosis |
| 3 | 130** | <35 | * | Flu |
| 4 | 130** | <35 | * | Tuberculosis |
| 5 | 130** | <35 | * | Cancer |
| 6 | 130** | <35 | * | Cancer |
| 7 | 130** | ≥35 | * | Cancer |
| 8 | 130** | ≥35 | * | Cancer |
| 9 | 130** | ≥35 | * | Cancer |
| 10 | 130** | ≥35 | * | Tuberculosis |
| 11 | 130** | ≥35 | * | Viral Infection |
| 12 | 130** | ≥35 | * | Viral Infection |

# De-anonymization attacks (II)

**Robust De-anonymization of Large Sparse Datasets. Narayanan and Shmatikov, 2008.**

**Showed the limitations of K-anonymity**

De-anonymization of the **Netflix Prize dataset** (500,000 anonymous records of movie ratings), using **IMDB** as the source of background knowledge.

They demonstrated that an adversary who knows just a few preferences about an individual subscriber can identify his record in the dataset.


# De-anonymization attacks (III)

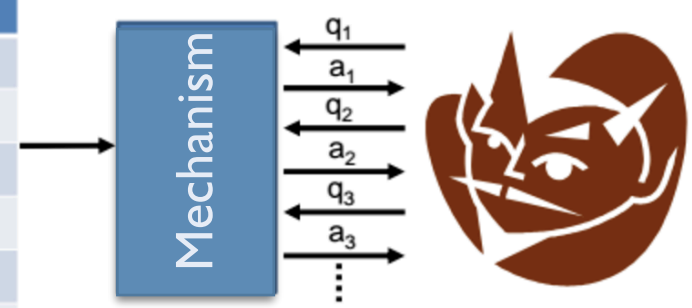
**De-anonymizing Social Networks.**
**Narayanan and Shmatikov, 200**

By using only the network topology, they were able to show that 33% of the users who had accounts on both **Twitter** and **Flickr** could be re-identified in the anonymous Twitter graph with only a 12% error rate.

# Protection of datasets via an interface

- Do not make the microdata available, but only aggregated information, by querying the interface.
- **Example:** Statistical Databases (SDB), often used for research purposes. For example, a medical SDB can be used to study the correlation between certain diseases and other attributes like: age, sex, weight, etc.

| Sex | Blood | ⋯ | HIV |
|---|---|---|---|
| F | B | ⋯ | Y |
| M | A | ⋯ | N |
| M | O | ⋯ | N |
| M | O | ⋯ | Y |
| F | A | ⋯ | N |
| M | B | ⋯ | Y |




- One can only retrieve aggregated information, not personal records
  - *"What is the average weight of people affected by the disease ?"*
  - *"Does Don have the disease ?"*

# There is still the problem of composition attacks Example

- A medical database D1 containing correlation between a certain disease and age.
- Query: "what is the minimal age of a person with the disease"

| name | age | disease |
|---|---|---|
| Alice | 30 | no |
| Bob | 30 | no |
| Carl | 40 | no |
| Don | 40 | yes |
| Ellie | 50 | no |
| Frank | 50 | yes |

D1 is 2**-anonymous with respect to the query**. Namely, every possible answer partitions the records in groups of at least 2 elements

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

- A medical database D2 containing correlation between the disease and weight.
- Query: "what is the minimal weight of a person with the disease"

| name | weight | disease |
|---|---|---|
| Alice | 60 | no |
| Bob | 90 | no |
| Carl | 90 | no |
| Don | 100 | yes |
| Ellie | 60 | no |
| Frank | 100 | yes |

Also D2 is 2-**anonymous**

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

# k-anonymity is not compositional

Combine with the two queries:

minimal weight and the minimal age of a person with the disease

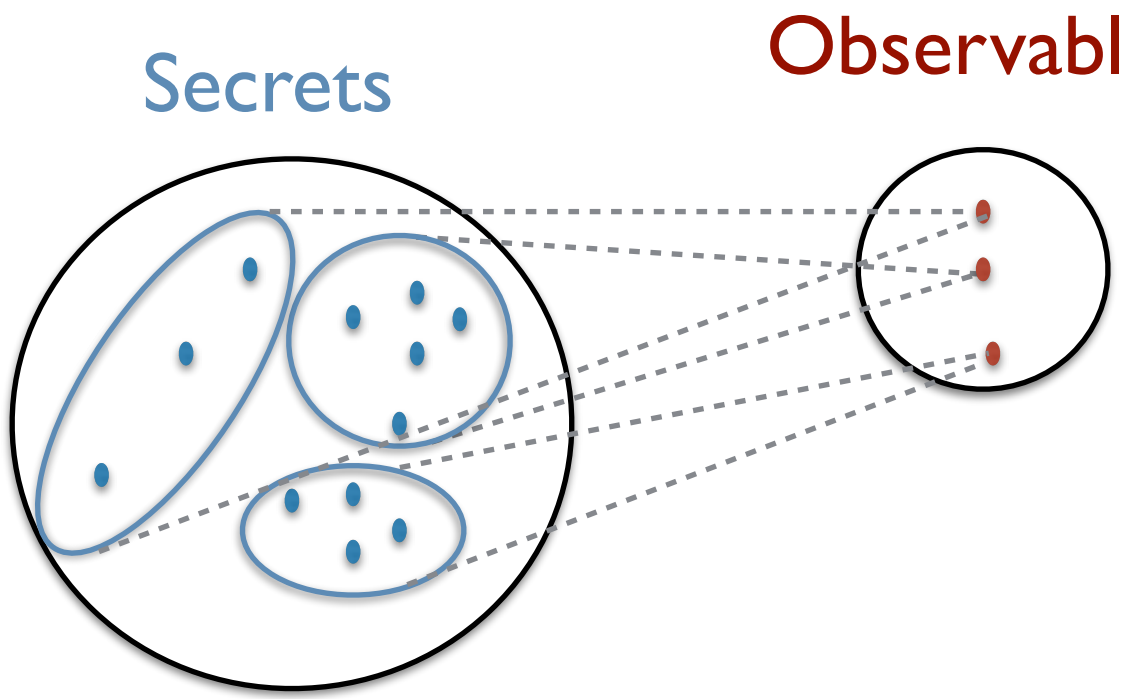
Answers: 40, 100. **Unique!**

| name | age | disease |
|---|---|---|
| Alice | 30 | no |
| Bob | 30 | no |
| Carl | 40 | no |
| Don | 40 | yes |
| Ellie | 50 | no |
| Frank | 50 | yes |

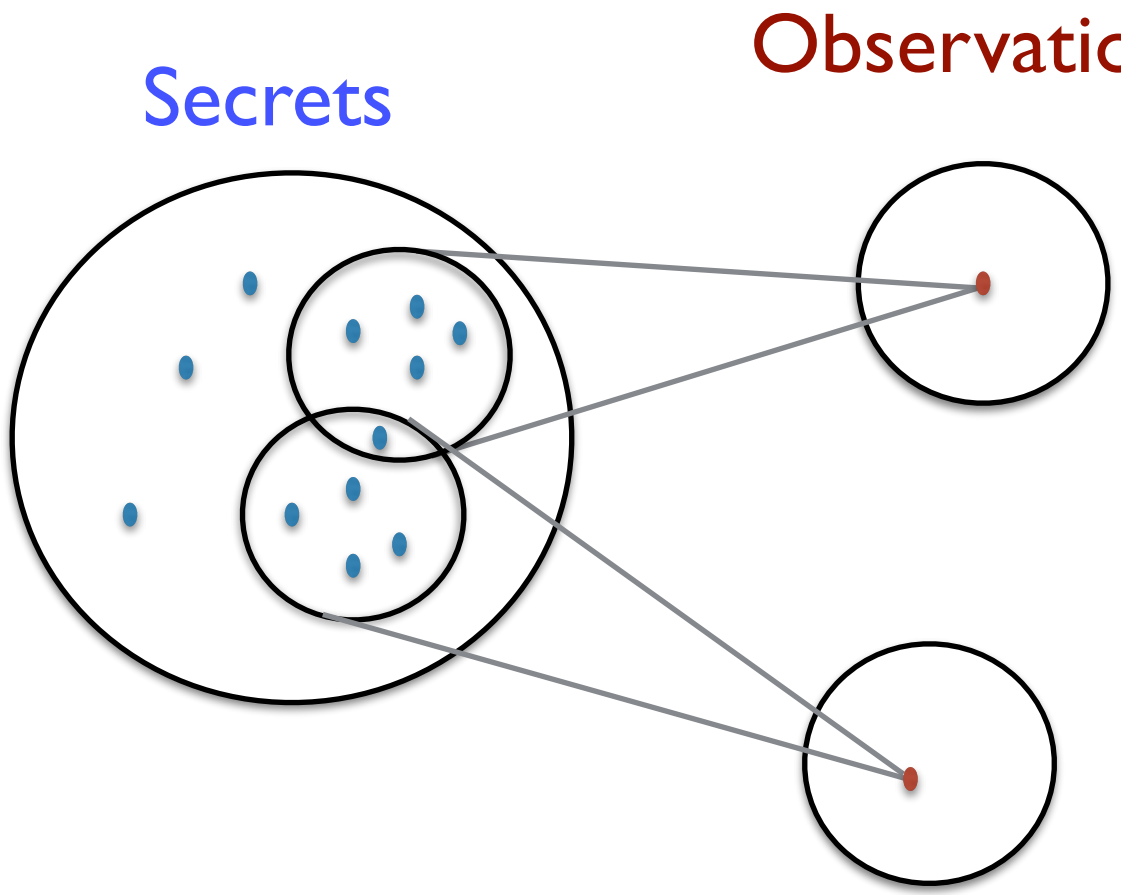
| name | weight | disease |
|---|---|---|
| Alice | 60 | no |
| Bob | 90 | no |
| Carl | 90 | no |
| Don | 100 | yes |
| Ellie | 60 | no |
| Frank | 100 | yes |

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

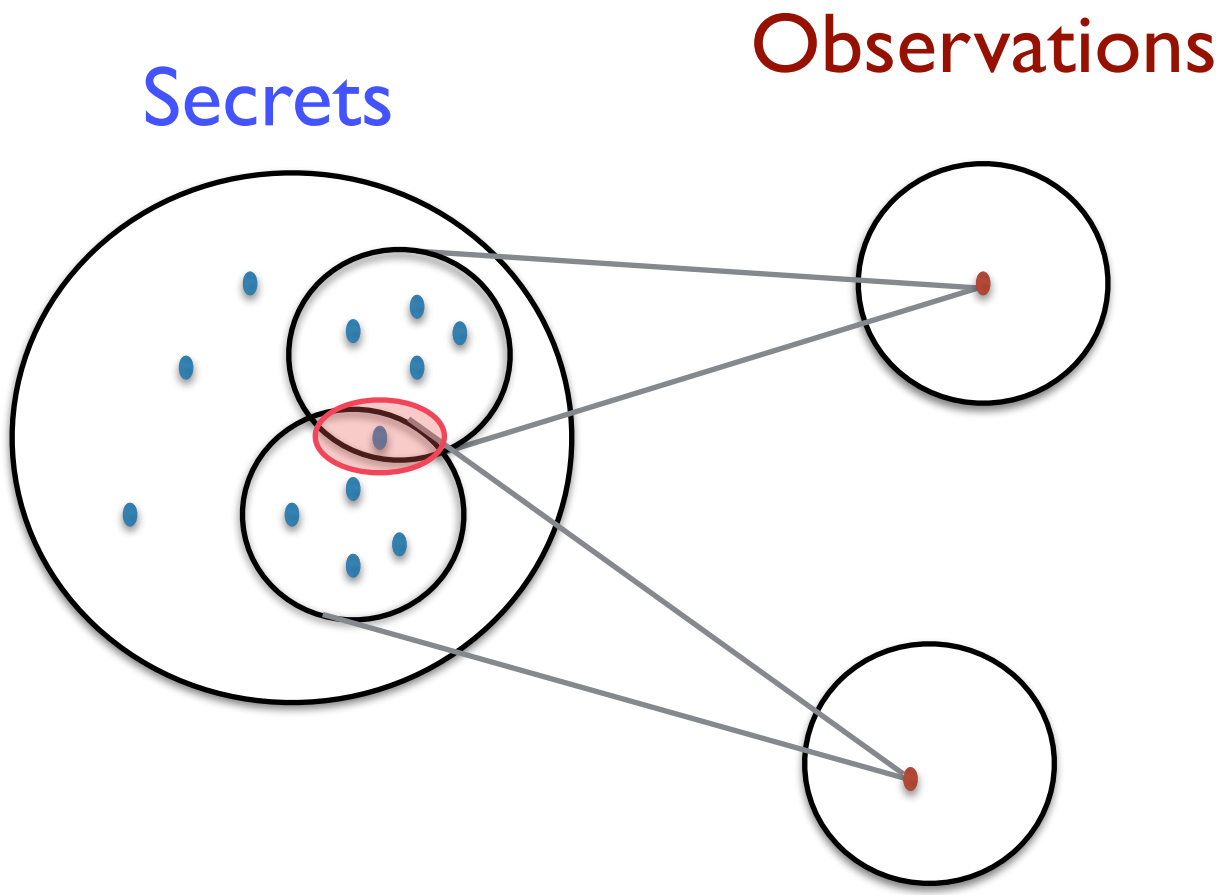
Composition attacks are a general problem of **Deterministic approaches :** They are all based on the principle that one observation corresponds to many possible values of the secret (group anonymity)



Problem of the deterministic approaches: the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



Problem of the deterministic approaches: the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



Too bad!!! What can we do?

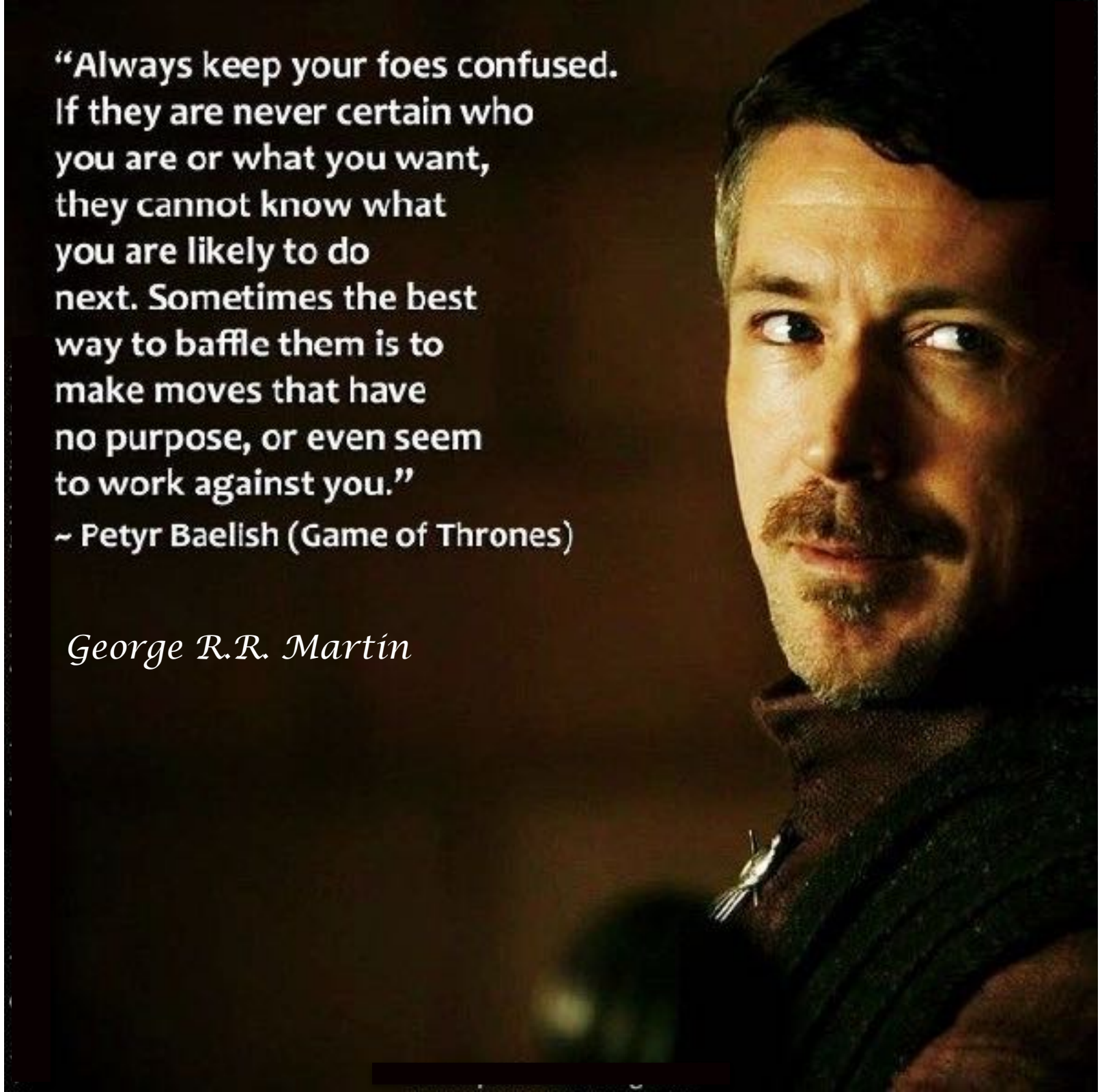
This is a job for...

# Random man!

# Probabilistic approaches

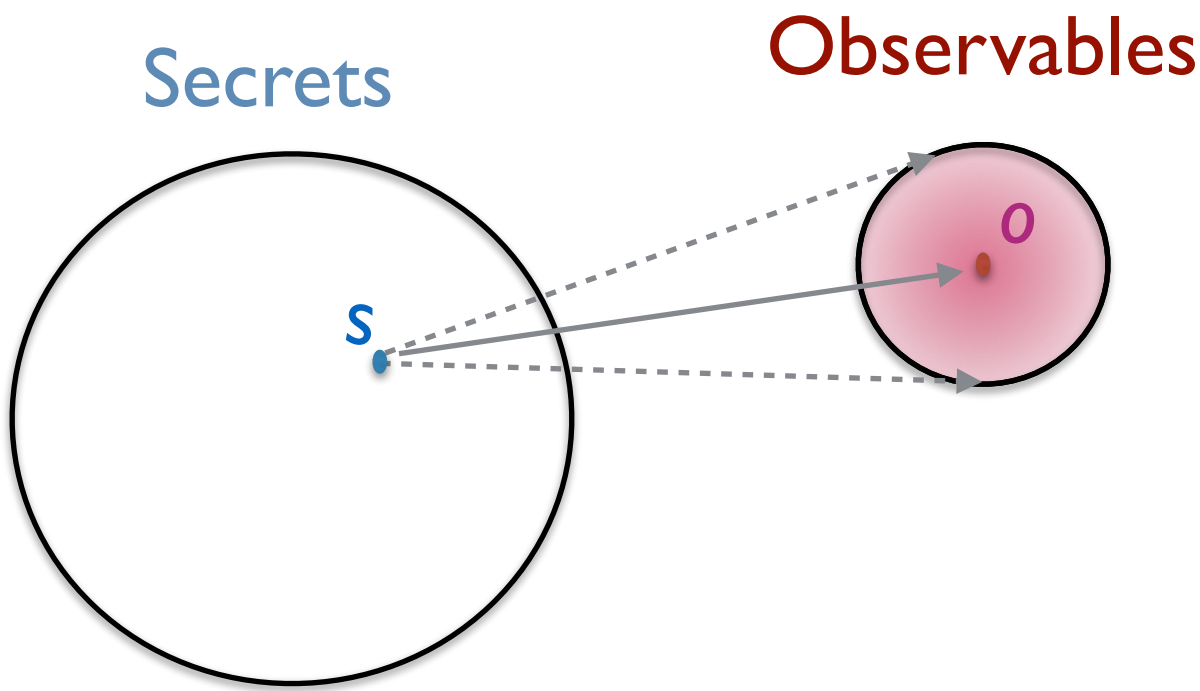
Modern techniques are based on randomization:

**probabilistic approaches.**


"Always keep your foes confused.
If they are never certain who
you are or what you want,
they cannot know what
you are likely to do
next. Sometimes the best
way to baffle them is to
make moves that have
no purpose, or even seem
to work against you."
~ Petyr Baelish (Game of Thrones)
George R.R. Martin

# Probabilistic approaches

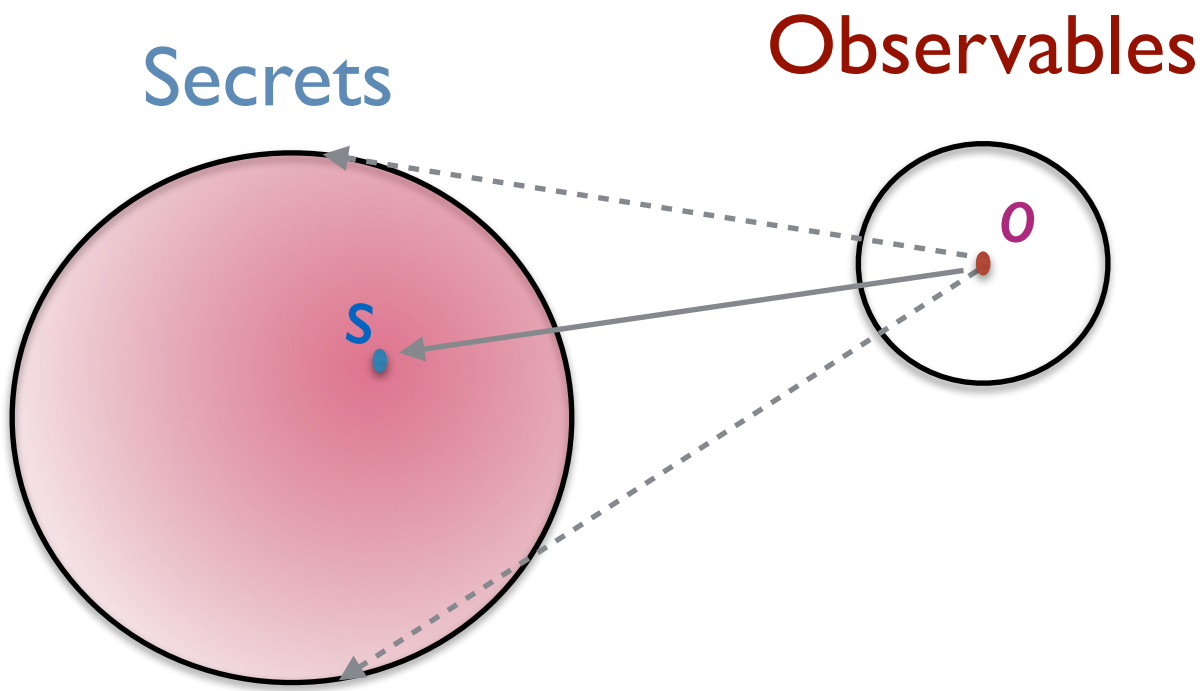
Every secret can generate any observable, according to a certain probability distribution.



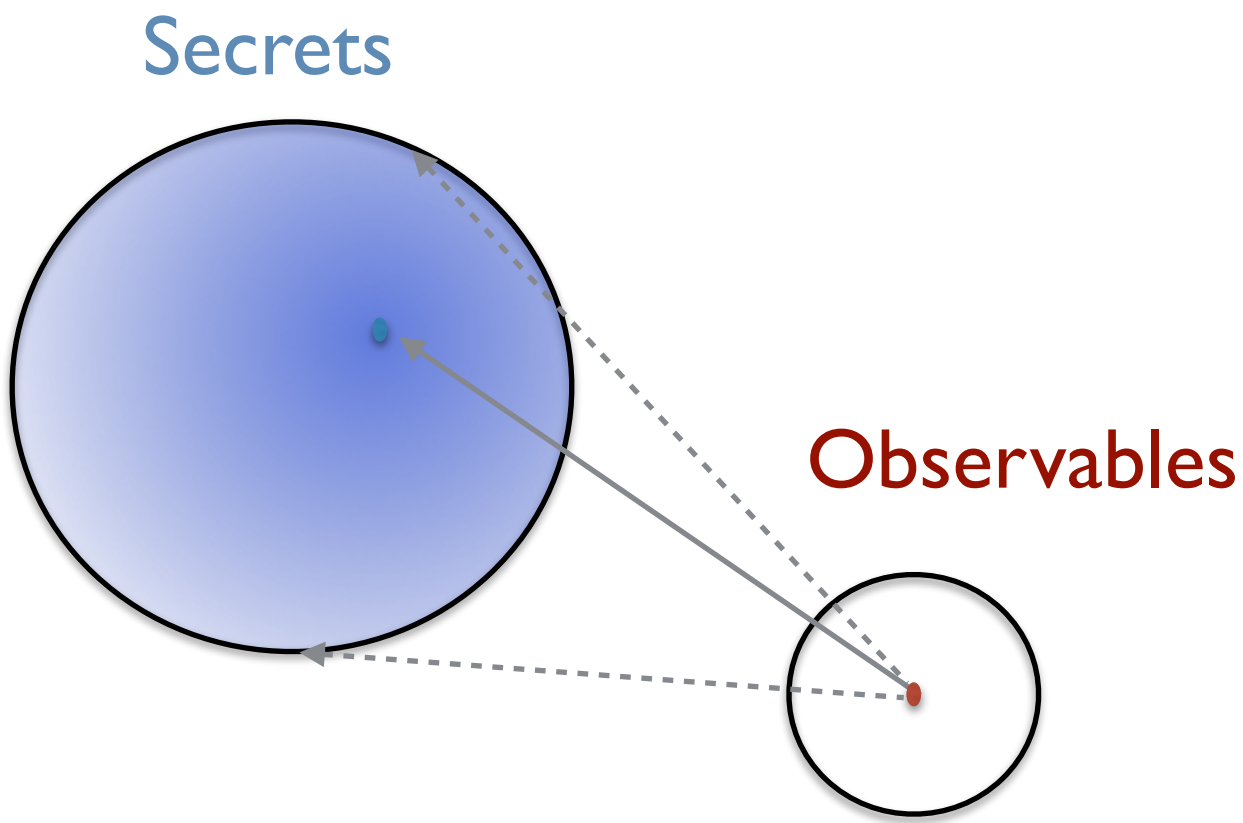
# Probabilistic approaches

By the Bayes law

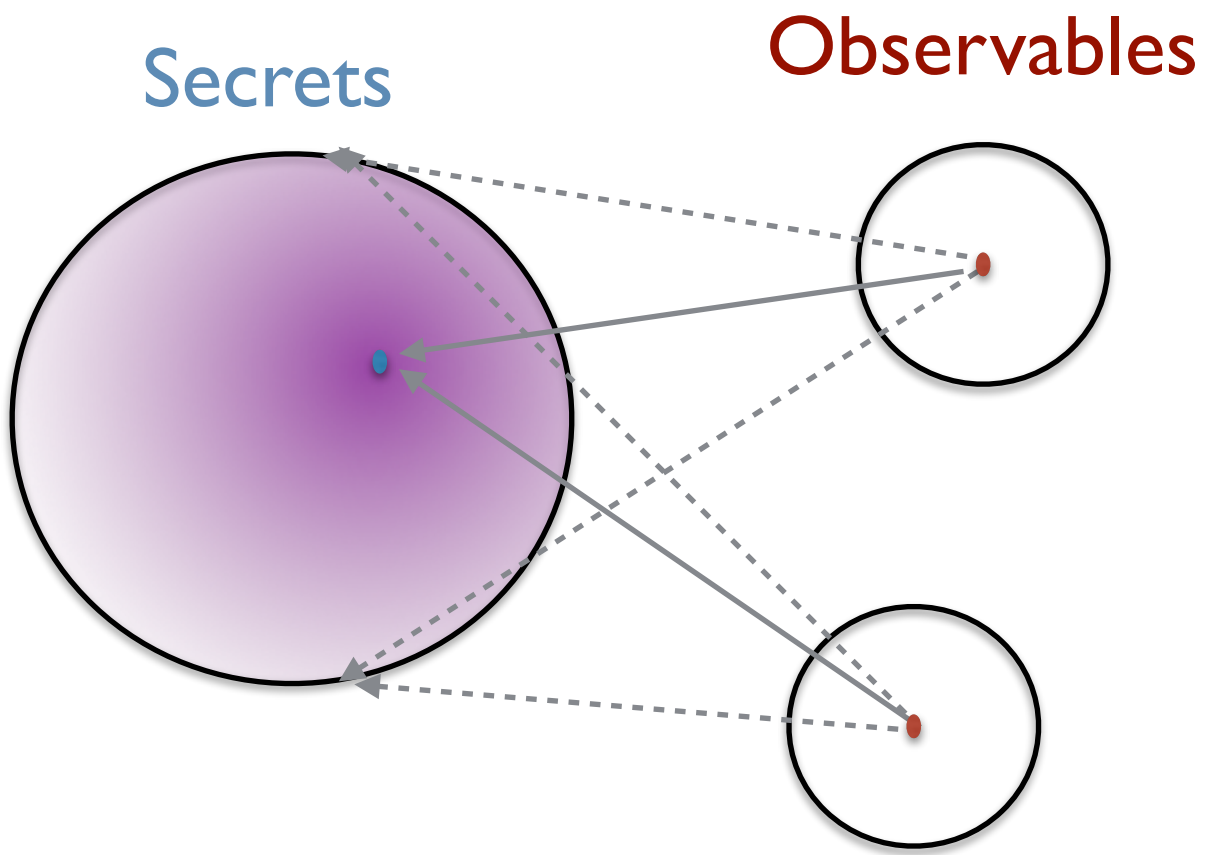
$$p(s|o) \propto p(o|s)$$



# Probabilistic approaches



# Probabilistic approaches



# Randomized approach for statistical databases

Introduce some probabilistic noise on the answer so to obfuscate the link with any particular individual

# Noisy answers

minimal age:

40 with probability 1/2

30 with probability 1/4

50 with probability 1/4

| name | age | disease |
|---|---|---|
| Alice | 30 | no |
| Bob | 30 | no |
| Carl | 40 | no |
| Don | 40 | yes |
| Ellie | 50 | no |
| Frank | 50 | yes |

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

# Noisy answers

minimal weight:

100 with prob. 4/7

90 with prob. 2/7

60 with prob. 1/7

| name | weight | disease |
|---|---|---|
| Alice | 60 | no |
| Bob | 90 | no |
| Carl | 90 | no |
| Don | 100 | yes |
| Ellie | 60 | no |
| Frank | 100 | yes |

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

# Noisy answers

Even if he combines the answers, the adversary cannot tell for sure whether a certain person has the disease

| name | age | disease |
|---|---|---|
| Alice | 30 | no |
| Bob | 30 | no |
| Carl | 40 | no |
| Don | 40 | yes |
| Ellie | 50 | no |
| Frank | 50 | yes |

| name | weight | disease |
|---|---|---|
| Alice | 60 | no |
| Bob | 90 | no |
| Carl | 90 | no |
| Don | 100 | yes |
| Ellie | 60 | no |
| Frank | 100 | yes |

| | |
|---|---|
| Alice | Bob |
| Carl | Don |
| Ellie | Frank |

# Noisy mechanisms

- The mechanisms reports an approximate answer, typically generated randomly on the basis of the true answer and of some probability distribution
- The probability distribution must be chosen carefully, in order to not destroy the utility of the answer
- A good mechanism should provide a good trade-off between privacy and utility. Note that, for the same level of privacy, different mechanisms may provide different levels of utility.

# Differential Privacy

**Definition** A randomized mechanism $\mathcal{K}$ is ε-differentially private if for all databases $x, x'$ **which are adjacent (**i.e., differ for only one record), and for all $z \in \mathcal{Z}$, we have

$$\frac{p(K = z|X = x)}{p(K = z|X = x')} \leq e^{\epsilon}$$

By the Bayes theorem, this definition corresponds to say that the answer given by $K$ does not change significantly the knowledge about an individual (prior and posterior are close)

**Important properties:**

- DP is robust with respect to composition of queries: the level of privacy ε decreases linearly with the number of queries
- The definition of DP is independent from the prior

# Differential Privacy at Google

## RAPPOR

**ABSTRACT**

Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, is a technology for crowdsourcing statistics from end-user client software, anonymously, with strong privacy guarantees. In short, RAPPORs allow the forest of client data to be studied, without permitting the possibility of looking at individual trees. By applying randomized response in a novel manner, RAPPOR provides the mechanisms for such collection as well as for efficient, high-utility analysis of the collected data. In particular, RAPPOR permits statistics to be collected on the population of client-side strings with strong privacy guarantees for each client, and without linkability of their reports.

This paper describes and motivates RAPPOR, details its differential-privacy and utility guarantees, discusses its practical deployment and properties in the face of different attack models, and, finally, gives results of its application to both synthetic and real-world data.

Úlfar Erlingsson

Head of the team
on data security
and privacy at Google

# Differential Privacy at Apple

Differential privacy is the statistical science of trying to learn as much as possible about a group while learning as little as possible about any individual in it.

**Apple has been doing some important work in this area to enable differential privacy to be deployed at scale."**

Craig Federighi,
Vice president of
Software Engineering @Apple

Keynote speech
Annual conference 2016
Apple software developers

# Content of the course

- We will focus on **probabilistic** methods for privacy and security
- Privacy:
  - Differential privacy
  - Local differential privacy (this is what Google does)
  - Location Privacy
- Security  (Kostas will illustrate it next):
  - (Quantitative) Information Flow
    - Leakage of information and inference attacks

# Exercise for next time

Bob wants to find out whether Don is affected by a certain disease d. He knows Don's age and weight, and that Don is going to check in a hospital that maintains an anonymized database of all patients, and that can be queried with queries of the form:

- How many patients are affected by the disease d ?

- What is the average age and weight of the patients affected by the disease d?

Discuss whether Bob can determine, with high probability, whether Don has the disease. What kind of background information Don needs? What kind of queries should he ask?

# Research internships

We have various internship (stage) subjects, ranging from rather theoretical to rather practical

- Privacy and Machine Learning
  - Machine learning attacks to Privacy
- Local Differential Privacy
- Location Privacy

# Research internships

- Location of the internship : LIX, Ecole Polytechnique, within an Equipe INRIA
- The internships will be “gratifié”
- It will be possible to continue the research as a PhD student