Probabilistic Methods for Privacy and Secure Information Flow

Catuscia Palamidessi

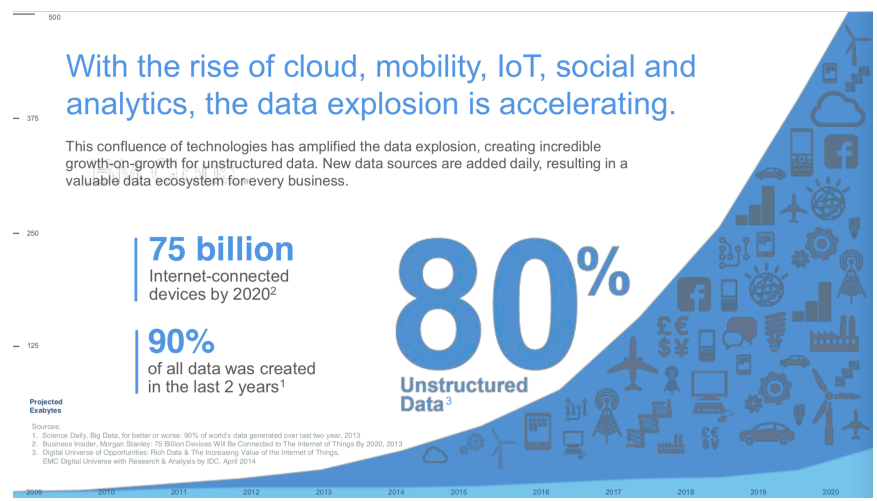
Content of the lectures

- Motivations, a bit of history, main problems
- Differential Privacy
- Local Differential Privacy
- Privacy vs Utility
- Quantitative Information Flow

Motivations

Privacy is not a new issue, but in our times the problem is exacerbated by the Big Data revolution: data are collected and stored in enormous amounts, and there is the computing power to analyse them and extract all sort of sensitive information

Also, data are accumulated at an increasing speed. According to a research made by IMB in 2017, 90% of the world data had been generated in the last 2 years!



Risks about privacy breaches

Sensitive information can be used for fraudulent purposes.

- Credentials

  Examples: credit card numbers, home access code, passwords, …

  Risks: Stealing personal property

- Information about the individual

  Examples: medical status, intimate videos, religious beliefs, political opinions

  Risks: discrimination, blackmailing, public shame

- Identification information, i.e., information that can uniquely identify an individual

  Examples: name, SSN, bank information, biometric data (such as fingerprint and DNA)

  Risks: Identity theft

Issues concerning privacy

Individuals

Enterprises

Research

data

service

Data analyst

Service provider

Data collection

Issue I: Inference attacks

The problem of Privacy is complicated because sensitive information can be derived using **side information**, i.e., correlated information that is necessarily public or anyway available to the attacker (inference attacks).

Example: all voters vote for the same candidate

- The typical countermeasures used in security (e.g., encryption, access control) do not help here
- The side knowledge of the adversary can increase with time

Issue 2: Trade off with utility

The measure to protect privacy should not destroy the utility of the data.

One of the main issues in the research about privacy-protection mechanisms is to find a good trade-off with utility

In general we consider two kinds of utility: the Quality of Service (QoS) and the precision of the analysis

Issues concerning privacy

Individuals

Enterprises

Research

Utility of analysis

data

service

QoS

Data analyst

Service provider

Data collection

Issue 3: Whom can we trust?

Issues concerning privacy

Individuals

Enterprises

Research

data

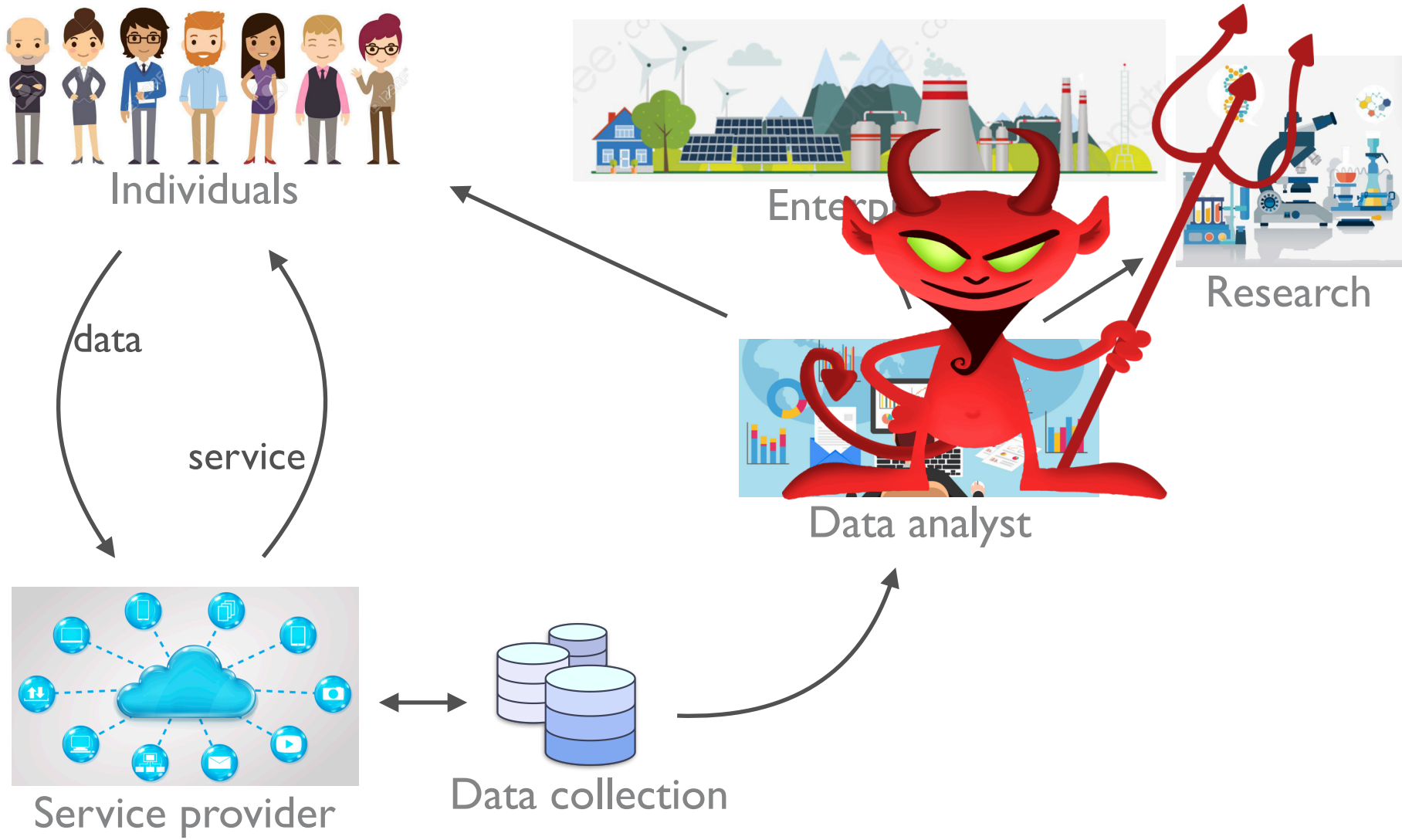
service

Data analyst

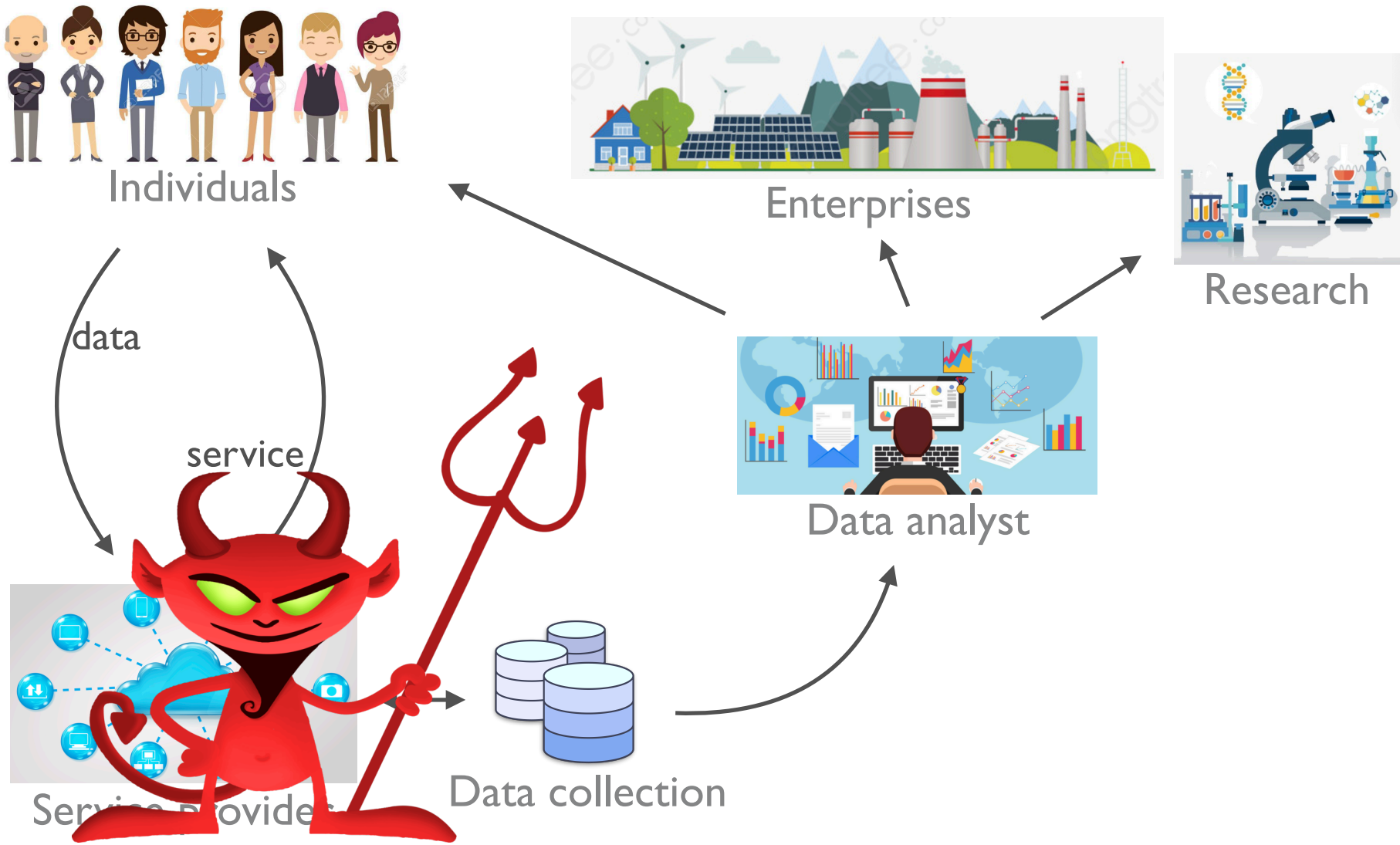
Service provider

Data collection

Issues concerning privacy



Issues concerning privacy



Issue 3: Whom can we trust?

1. **Global model**: we trust the server / data curator.
   - The sanitization is done by the curator.
   - Utility is precision of analysis.
   - Two cases:
     1. the (sanitized) micro data are made available, or
     2. they are not available, we can only query the database

2. **Local model**: the server / curator may be corrupted or unable to protect the data.
   - The sanitisation is done at the user's side
   - Both kinds of utility should be taken into account
   - The sanitised micro data are made publicly accessible.

The local model has become more popular recently since people tend to trust less and less the service providers and curators (also due to recent scandals). Some big companies (e.g., Google and Apple) have developed their own LDP systems.

Scenario 1.1:

Global model

The micro data are made available

First solution: **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

	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

First solution: **anonymization**

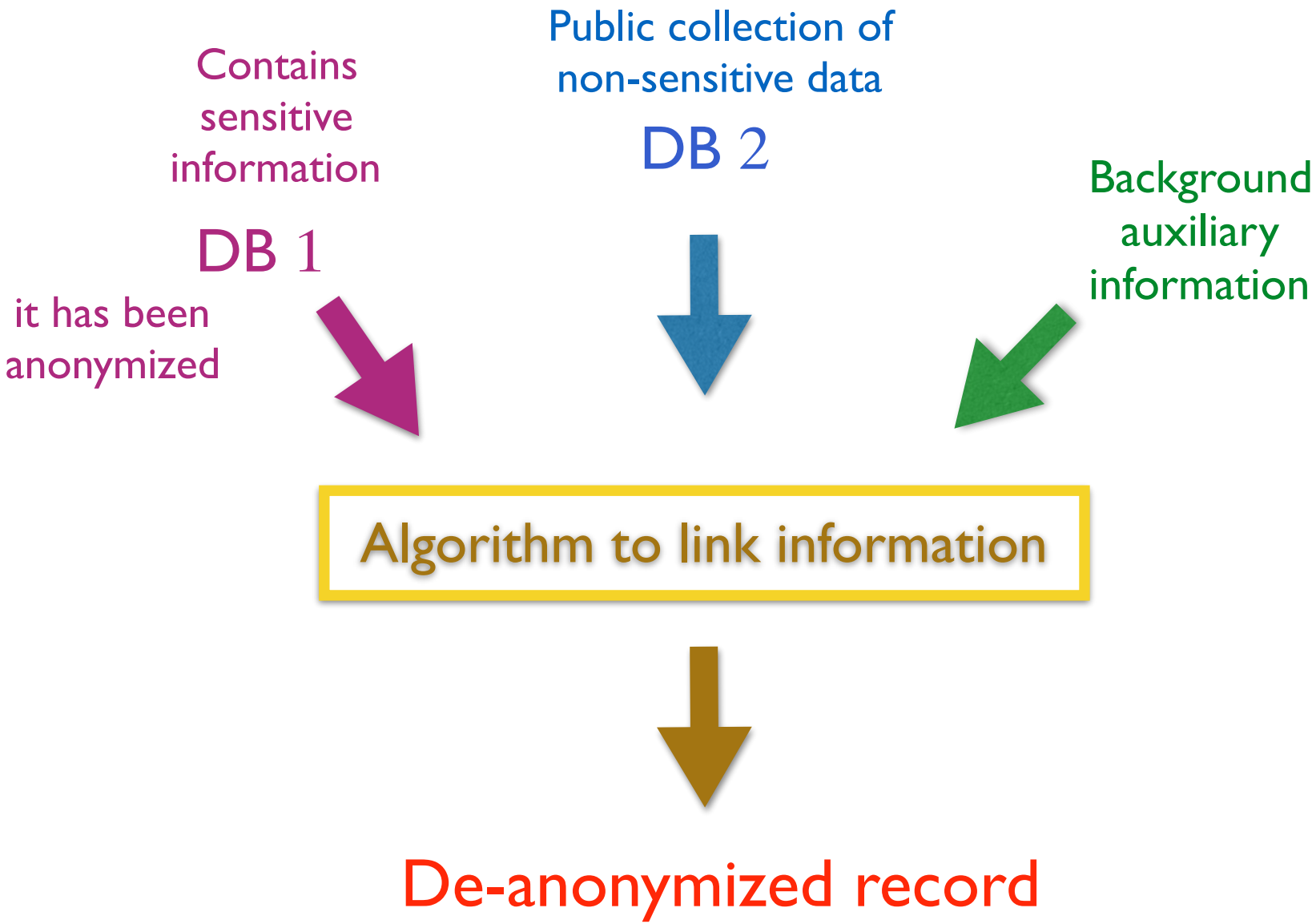
- Anonymization removes the column of the name, so that, for instance, the grayscale disease cannot be directly linked to Jorah Mormont

- Hystorically the first method, still used nowadays

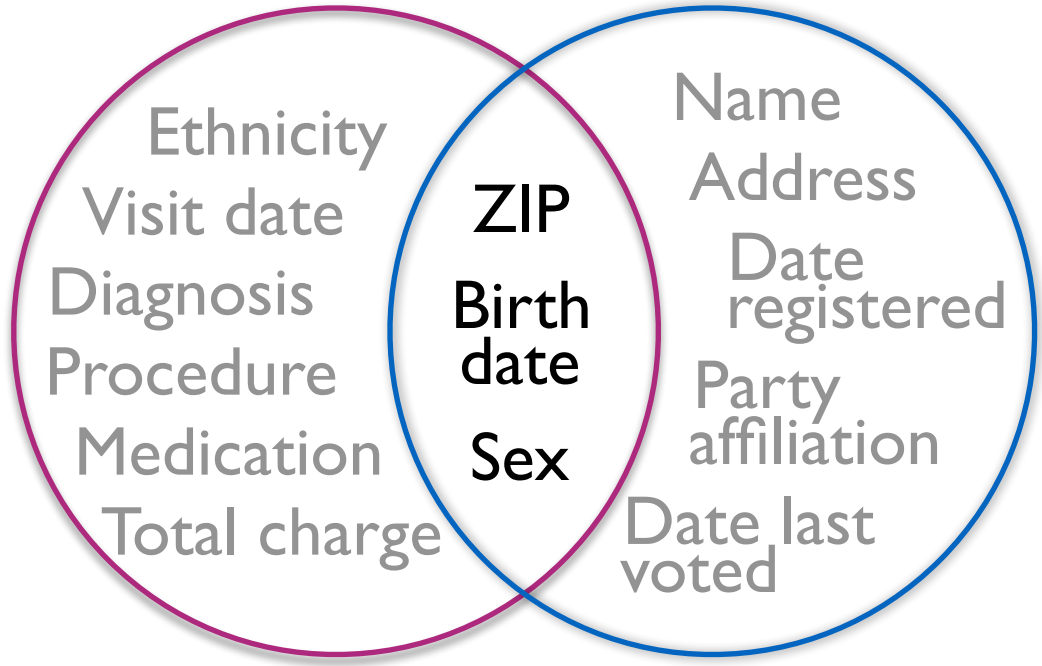
- However, this solution has been (already several years ago) shown to be ineffective, i.e., vulnerable to de-anonymization attacks

	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

De-anonymization attack (I). Sweeney'98



De-anonymization attack (I). Sweeney'98



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

This attack has lead to the proposal of k-anonymity

K-anonymity [Samarati & Sweeney]

- 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

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

- Obvious problem: in the sanitized dataset, all the individual in a group may the same value for the sensitive data, like in this table
- Clearly, the people in that group are not protected from the revelation of their disease

	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.

ℓ-diversity [Kifer et al.]

- A solution to this problem was proposed under the name of ℓ-diversity.
- The idea is to form the groups in such a way that each group contains a variety of values for the sensitive data

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

Table 5: 3-diverse table

Problems with k-anonymity and similar methods

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

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

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, 2009

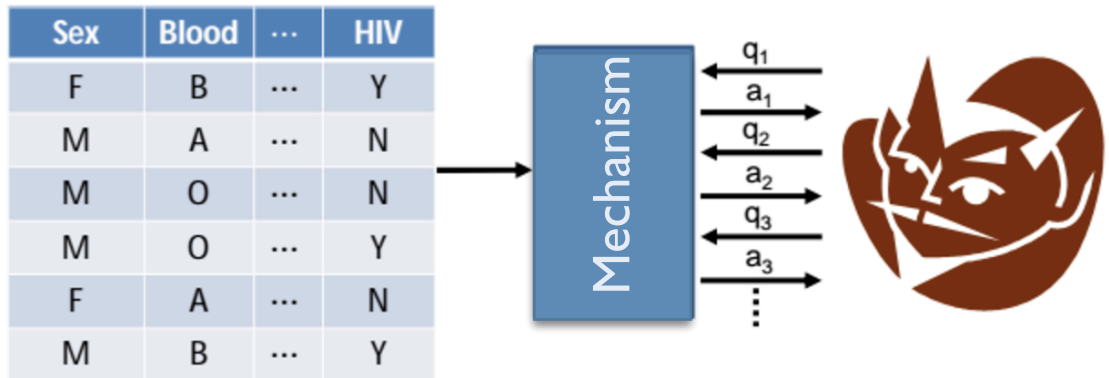
By using only the network topology, they were able to show that 33% of the users who had accounts on both **Twitter** and **Flickr** could be re-identified in the anonymous Twitter graph with only a 12% error rate.

Scenario 1.2:

Global model

Micro data not accessible, we can only query the DB

Protection of datasets via an interface



- 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

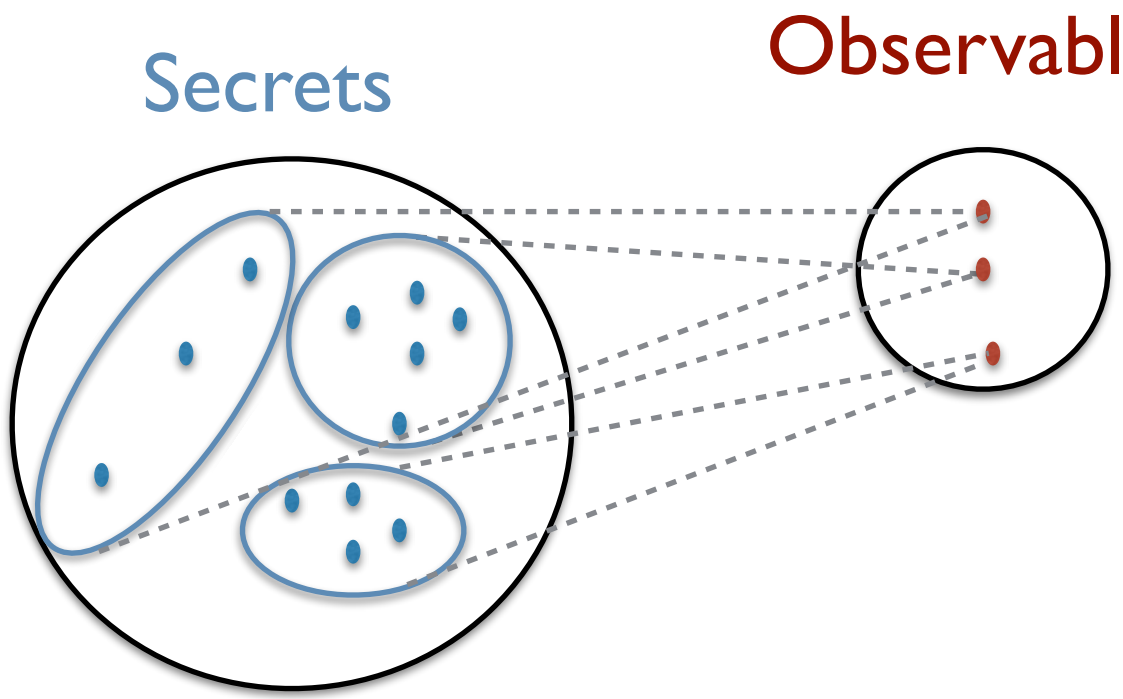
Answers: 40, 100. **Unique!**

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

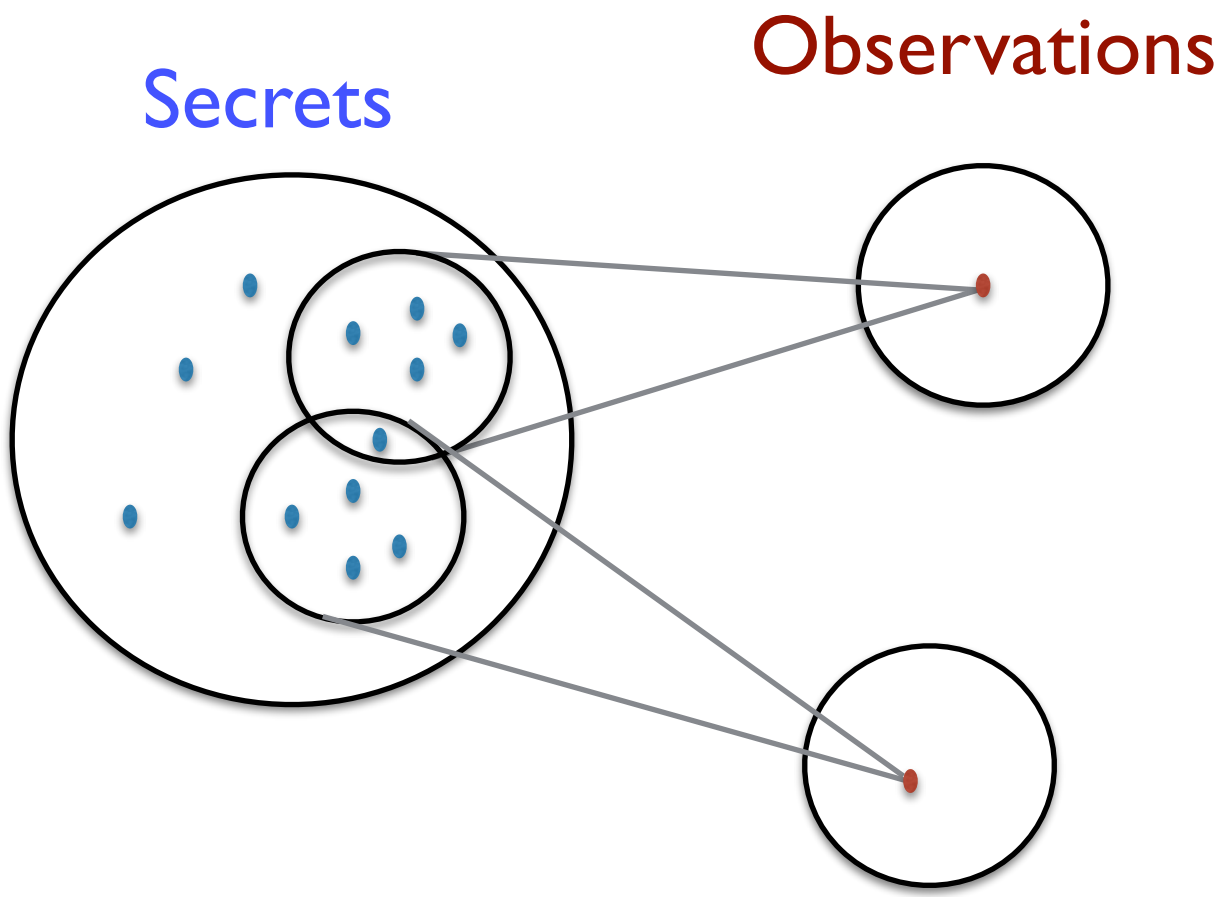
name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

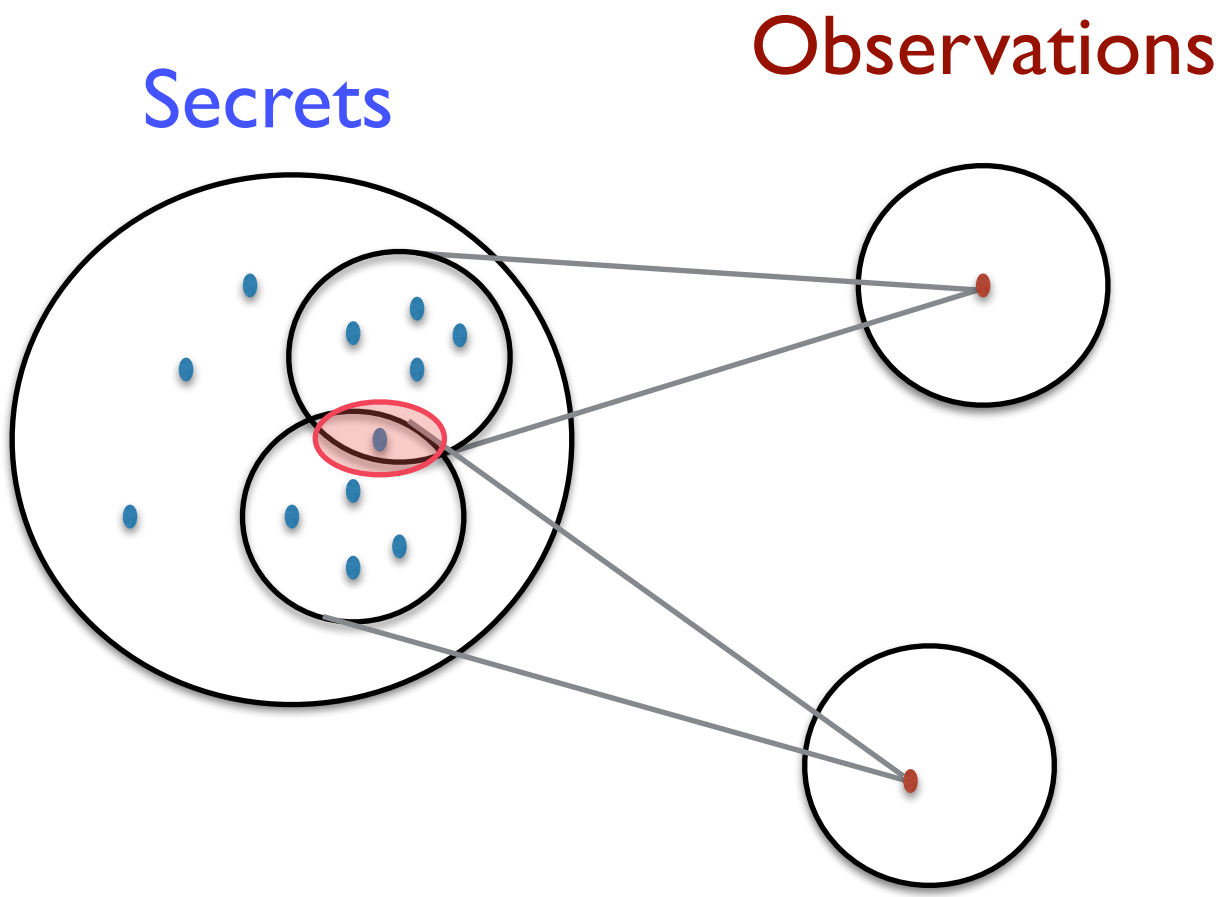
Composition attacks are a general problem of **Deterministic approaches :** They are all based on the principle that one observation corresponds to many possible values of the secret (group anonymity)



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?

Too bad!!! What can we do?

Use probabilistic approaches!

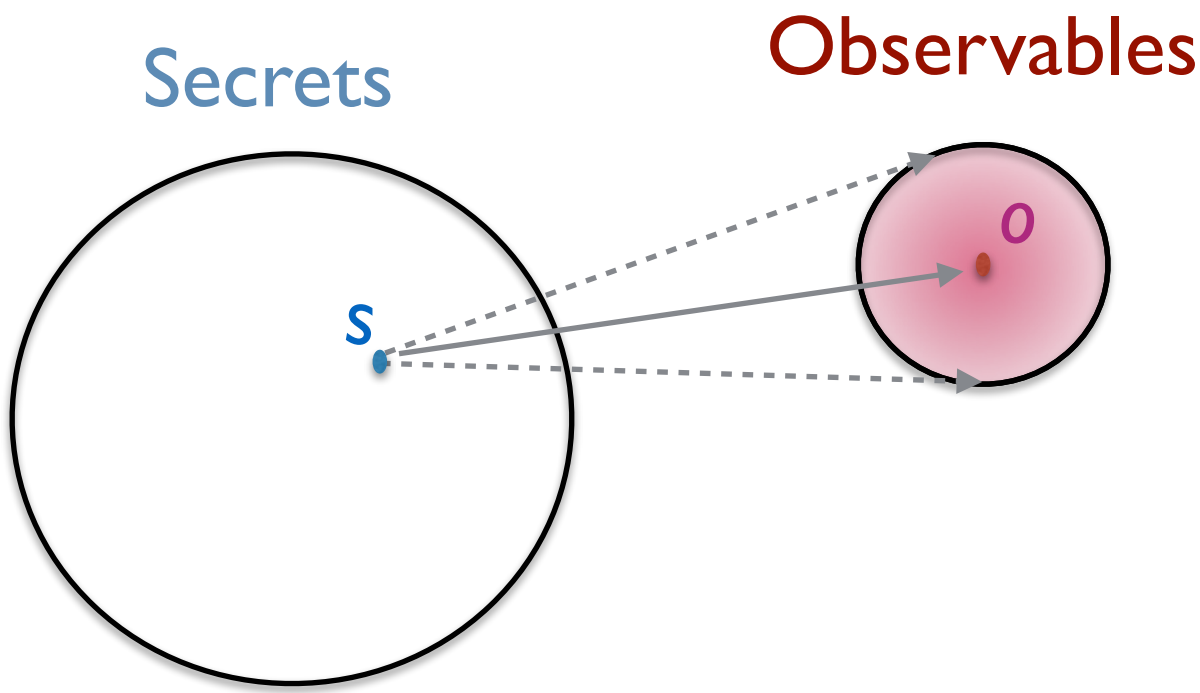
Too bad!!! What can we do?

Use probabilistic approaches!

Most of the state-of-the-art techniques are indeed based on randomization

Probabilistic approaches

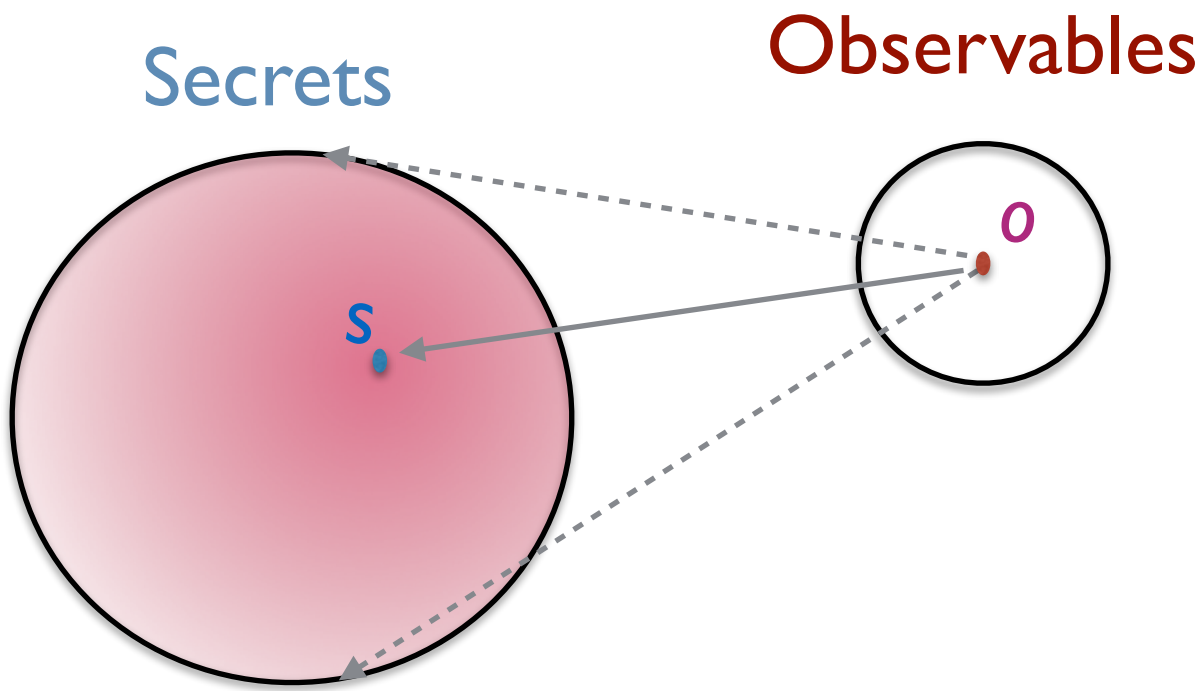
Every secret can generate any observable, according to a certain probability distribution.



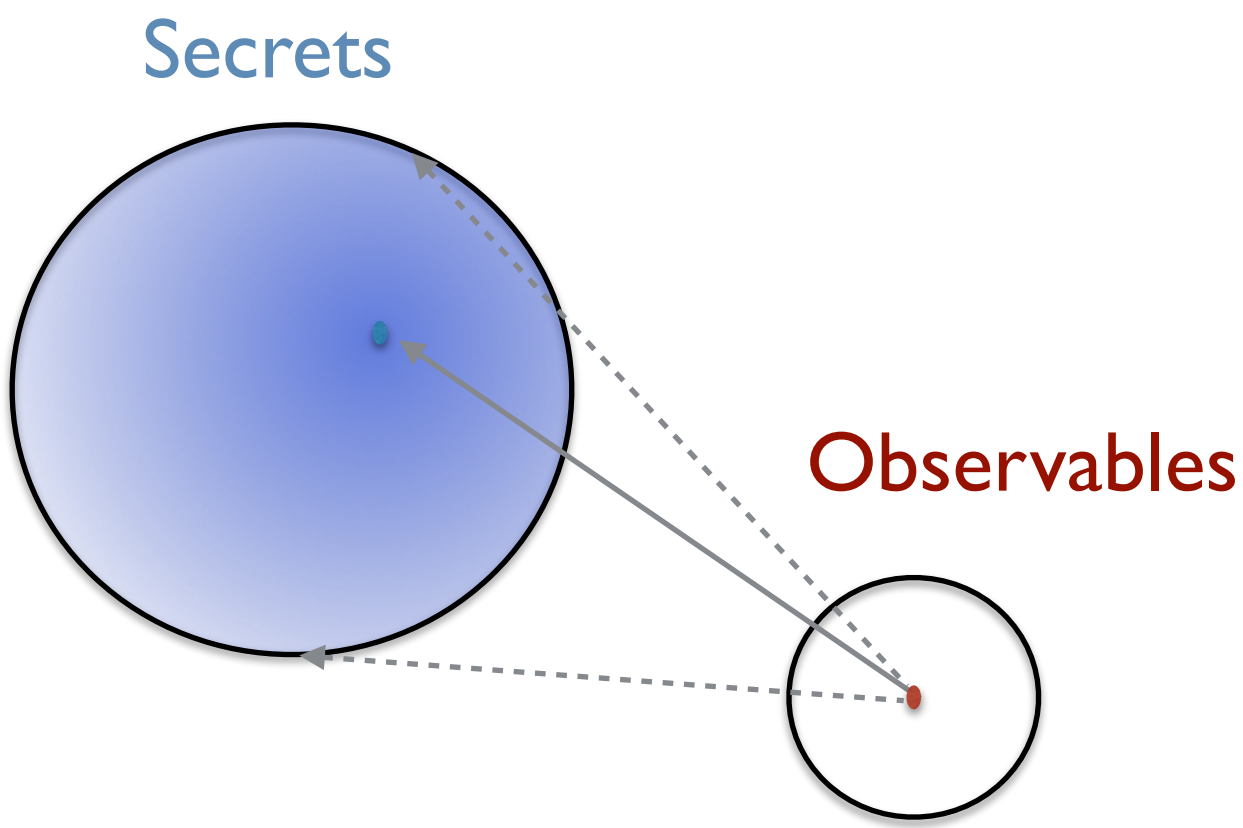
Probabilistic approaches

By the Bayes law

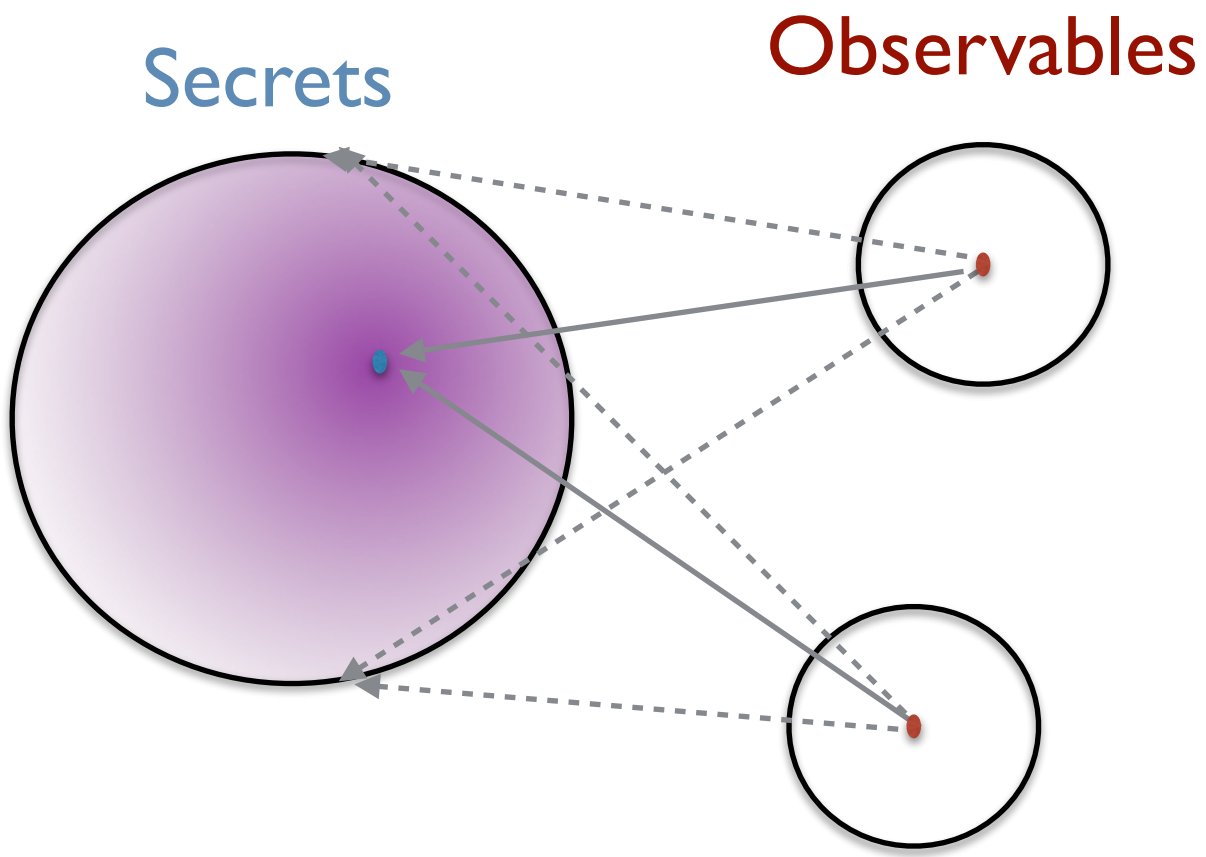
$$p(s|o) \propto p(o|s)$$



Probabilistic approaches



Probabilistic approaches



Randomized approach for DB sanitisation

- Allow accessing the DB only by queries
- 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

This idea is at the basis of differential privacy, which will be the topic of next lecture

Thanks for the attention

Questions?