



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Advanced Database Systems (INFS3200)

Lecture 10: Data Privacy and Quality

Lecturer: A/Prof Sen Wang

School of Electrical Engineering and Computer Science (EECS)

Faculty of Engineering, Architecture and Information Technology

The University of Queensland

Recap

DB Integration:

- Why DB Integration and Related Issues
- Global Information Systems
 - Federated Databases
 - Multidatabases
- Mediator-Wrapper Architecture
- Challenges in DB Integration
- Three Steps for DB Integration
- View-based DB Integration
 - Global-as-view vs Local-as-view
- Limitations of Views

Data Linkage:

- Why Data Linkage & What is NOT Data Linkage
- Data Linkage Applications
- Linking with Different Granularity
 - Field Matching
 - Edit Distance
 - q-Gram and Jaccard Coefficient
 - TF/IDF and Cosine Similarity
 - Numeric Similarity
 - Record Matching
 - Weighted Sum & Rule-based Approaches
 - Group Matching
 - Canopy Cluster

Outline

Privacy issues

- Data privacy – definition & challenges
- Privacy preserving techniques for dataset publishing
 - k -anonymity, l – *diversity*, t -closeness
 - Differential privacy

Data Quality issues

- Data Quality Dimensions
 - Concept of data quality
 - Data life cycle
- Four basic Steps of Data Governance w.r.t. Quality

Privacy and Data Release – A Case Study

NYC taxi and limousine commission released 2013 trip data.

- Start point, end point, timestamps, taxi id, fare, tip amount.
- 173 million trips “anonymized” to remove identifying information.



NYC
Taxi & Limousine
Commission

- **Traffic patterns:** traffic management strategies and infrastructure development.
- **Urban planning:** population density, business districts, and points of interest.
- **Socioeconomic trends:** areas with higher spending or dependence on taxis.
- **Predictive modeling:** resource allocation and optimizing driver availability.

<https://data.cityofnewyork.us/Transportation/2013-Yellow-Taxi-Trip-Data/t7ny-aygi/data>

Privacy and Data Release

Use a simple **MD5 hash** to **anonymize** personally identifiable information (the driver's licence number) → **NOT easily reversed**.

"Alex" -> a08372b70196c21a9229cf04db6b7ceb

The data had been anonymised by hashing, a **cryptographic** function which is supposed to be "**one-way**":

- it's **very easy** to find the hash of a given piece of data, and **very hard** – mathematically impossible, in theory – to find the piece of data which resulted in a given hash.

Privacy and Data Release

Licenses are all **six-digit** (e.g. xxx,xxx) or **seven-digit numbers** (e.g. 5,xxx,xxx) starting with a five. That means that there are only **2m** possible license numbers

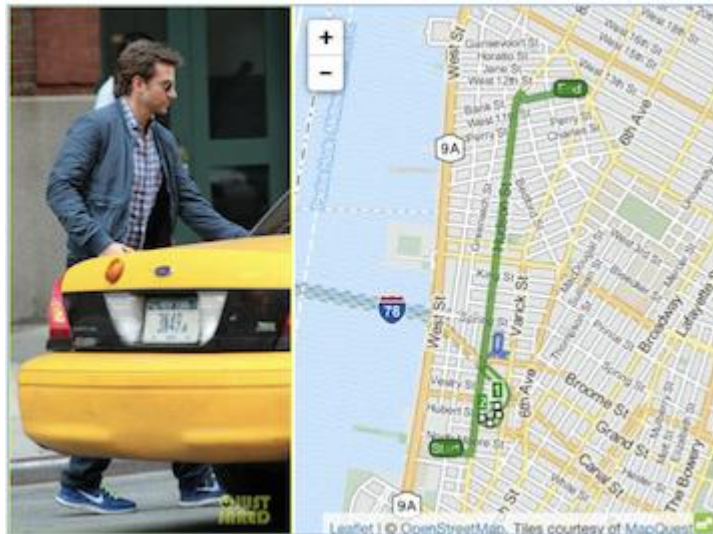
But once the possible entries have been down to **2m** different numbers, it was a matter of only **minutes** to determine which numbers were associated with which pieces of anonymised data.

Could yield personal details, such as drivers' address and income!

Privacy and Data Release

What's worse, with other publicly available data, one can **link** people to taxis and find out where they went

- For example, paparazzi pictures of celebrities.



Bradley Cooper (actor)



Jessica Alba (actress)

Privacy and Data Release

Not just celebrities: can find trips starting at “sensitive” locations.

- – For example, Larry Flynt’s Club (adult club)

Can find more about *venue’s customers*.

- “Examining one of the clusters ... **only one of the five** likely drop-off **addresses** was inhabited; a search ... revealed its **resident’s name**. By examining other drop-offs at this address ... this gentleman also frequented ... “Rick’s Cabaret” and “Flashdancers”. Using websites like Spokeo and Facebook ... able to find out his ... relationship status, court records and even **a profile picture!**”

Data Privacy

Data privacy: empowering users to make their own decisions about **who can process** their data and **for what purpose**

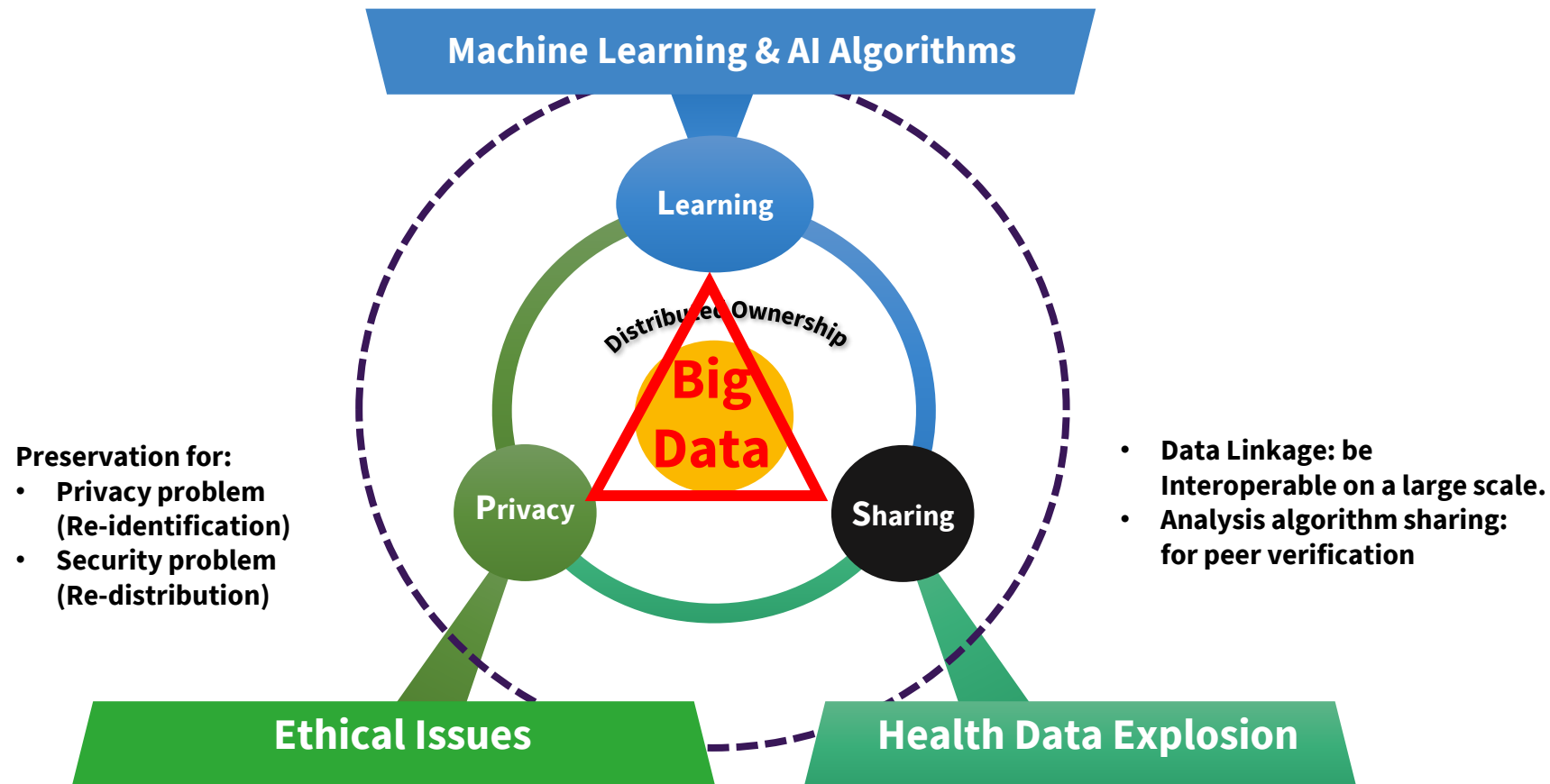
Data privacy is the relationship among:

- (1) the collection & dissemination of data,
- (2) technology,
- (3) the public expectation of privacy, and
- (4) the legal and political issues surrounding them



Satisfying Learning-Privacy-Sharing Altogether Is Impossible

- Machine Learning on individual private patterns (e.g., DNA Sequencing)
- AI with Human-in-the-Loop



Data Utility and Data Privacy

The challenge of **data privacy** is to **utilize data** while **protecting** individual's privacy preferences and their personally identifiable information

Sensitive information

- **Identity**
 - Direct identifiers: attributes that explicitly identify individuals
 - Quasi-identifiers: attributes that in combination with others lead to identification (e.g. address + age + diagnosis)
- **Sensitive attributes**
 - Attributes that individuals are not willing to disclose, such as salary, health, religion
- **Relationship** (e.g. social relationships & family relationships)

An Example of Statistical Attack

Privacy rules

- Cannot query about individual's salary

Attack queries: Statistical attacks exploit aggregate queries like SUM and COUNT to infer sensitive information that might NOT be directly accessible.

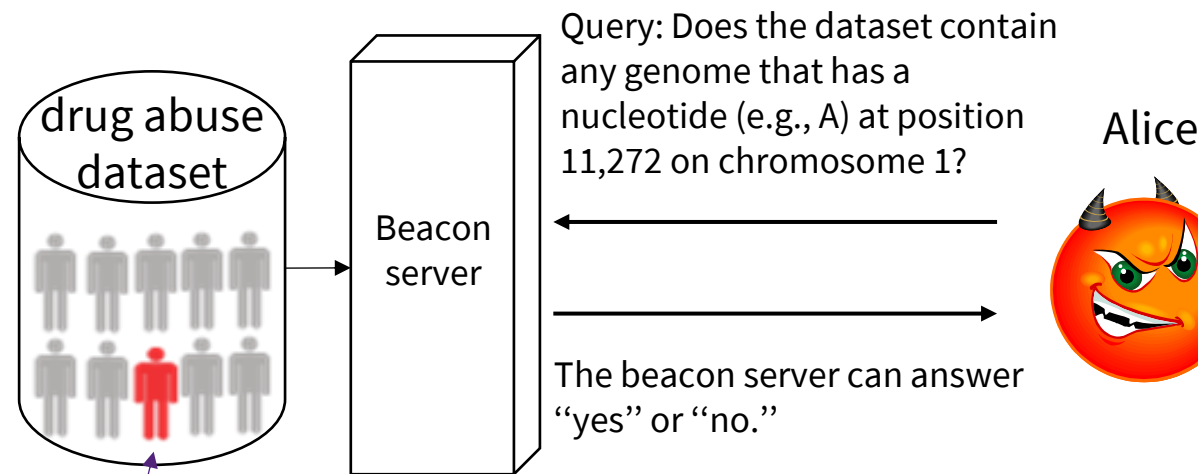
```
select    count (*)  
from      staff  
where     title = "Professor"
```

```
select    sum(salary)  
from      staff  
where     title = "Professor"
```

An Example of Real-world Statistical Attacks

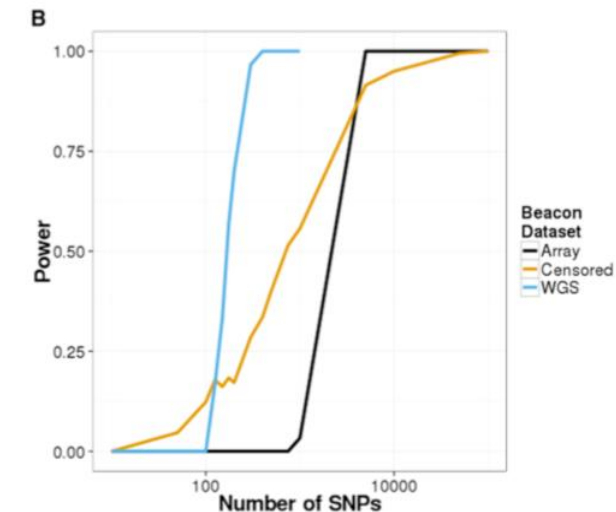
Privacy rule: the drug abuse dataset is not accessible to users, but only allows them to **query** the allele-presence information.

By calculating the likelihood of responses, the attacker can differentiate individuals in the beacon from those not in the beacon.



Target Bob information: Bob's DNA

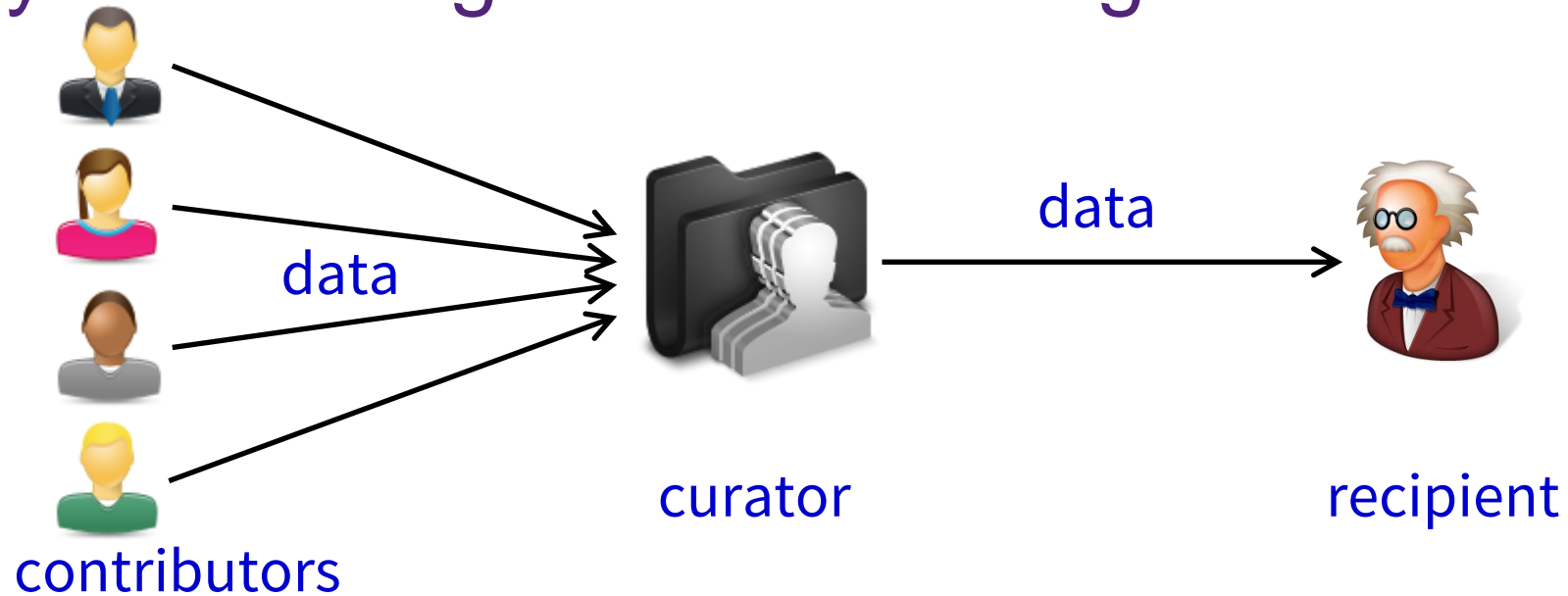
The Beacon Project by the Global Alliance for Genomics & Health (GA4GH) aims to simplify data sharing through a web service ("beacon") that provides only allele-presence information.



Power of Re-identification Attacks on Beacons
Constructed with Real Data: With just **250 queries**, beacon membership could be detected with 95% power and a 5% false-positive rate

(Suyash , . et al, 2015)

Privacy Preserving Data Publishing

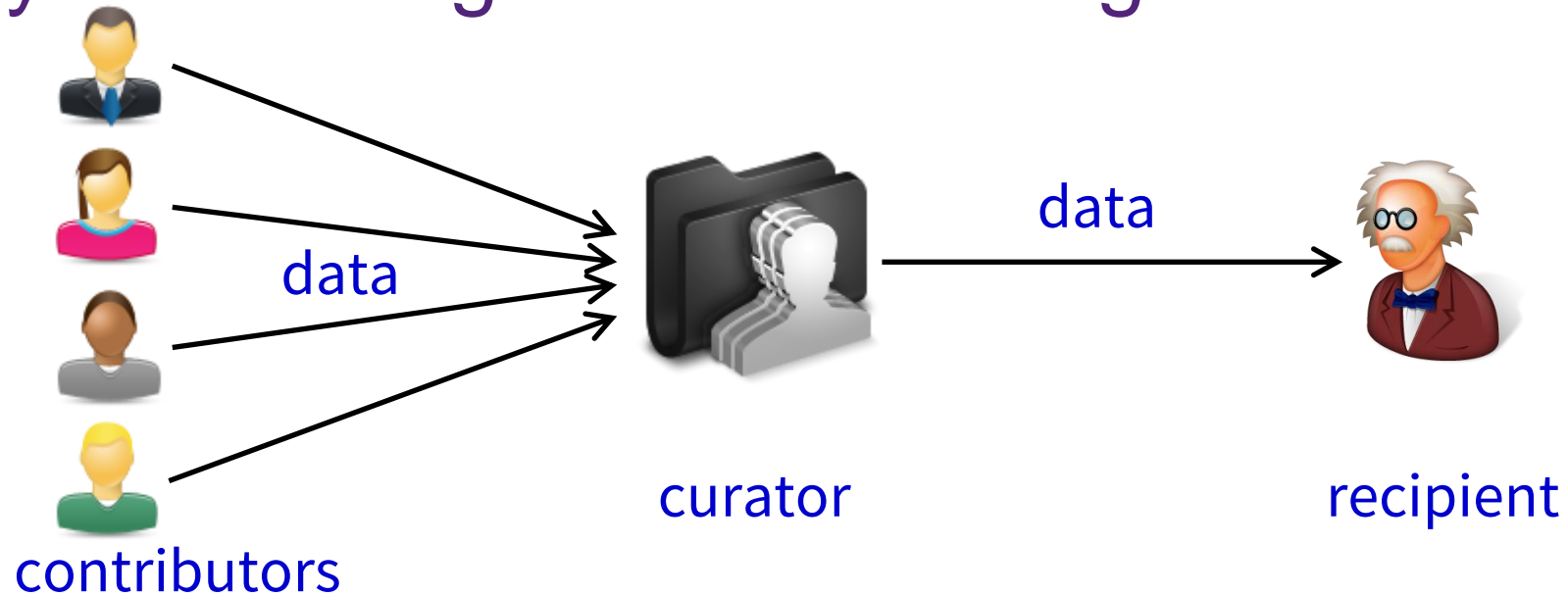


Contributors: provide data about themselves

Curator: collects data and releases them in a certain form

Recipient: uses the released data for analysis

Privacy Preserving Data Publishing



Objectives:


- The privacy of the contributors are protected
- The recipient gets useful data

Privacy Breach: The MGIC Case

Curator: *Massachusetts Group Insurance Commission* (MGIC)

Data released: “anonymized” medical records

Intention: facilitate medical research



Name	Birth Date	Gender	ZIP	Disease
Alice	1960/01/01	F	10000	flu
Bob	1965/02/02	M	20000	dyspepsia
Cathy	1970/03/03	F	30000	pneumonia
David	1975/04/04	M	40000	gastritis

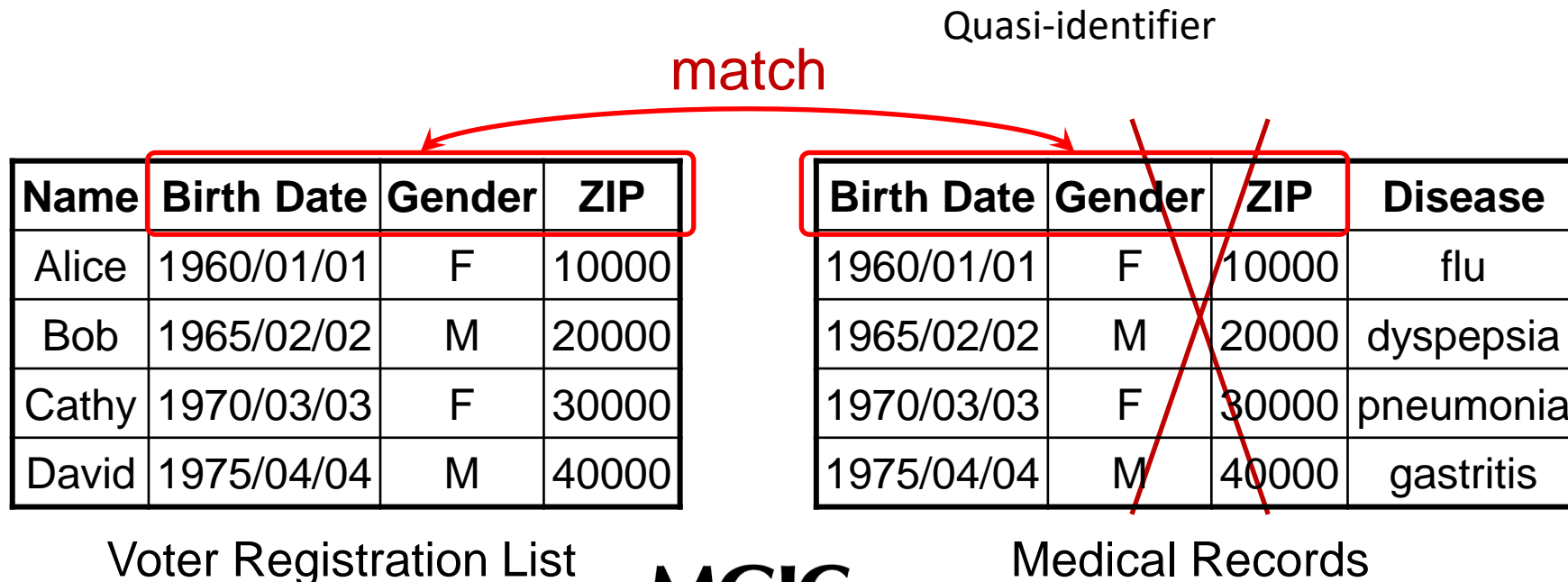
Medical Records

Privacy Breach: The MGIC Case

Curator: *Massachusetts Group Insurance Commission* (MGIC)

Data released: “anonymized” medical records

Intention: facilitate medical research



MGIC

Where Do I Get These Records?

FORBES > INNOVATION > CYBERSECURITY

191 Million US Voter Registration Records Leaked In Mystery Database

Thomas Brewster Forbes Staff

Senior writer at Forbes covering cybercrime, privacy and surveillance.

Follow



Dec 28, 2015, 08:50am EST



This article is more than 8 years old.

A whitehat hacker has uncovered a database sitting on the Web containing various pieces of personal information related to 191 million American citizens registered to vote. On top of the concomitant problems of disclosing such a significant leak to that many people, no one knows who is actually responsible for the misconfiguration that left the data open to anyone.

Lessons Learned

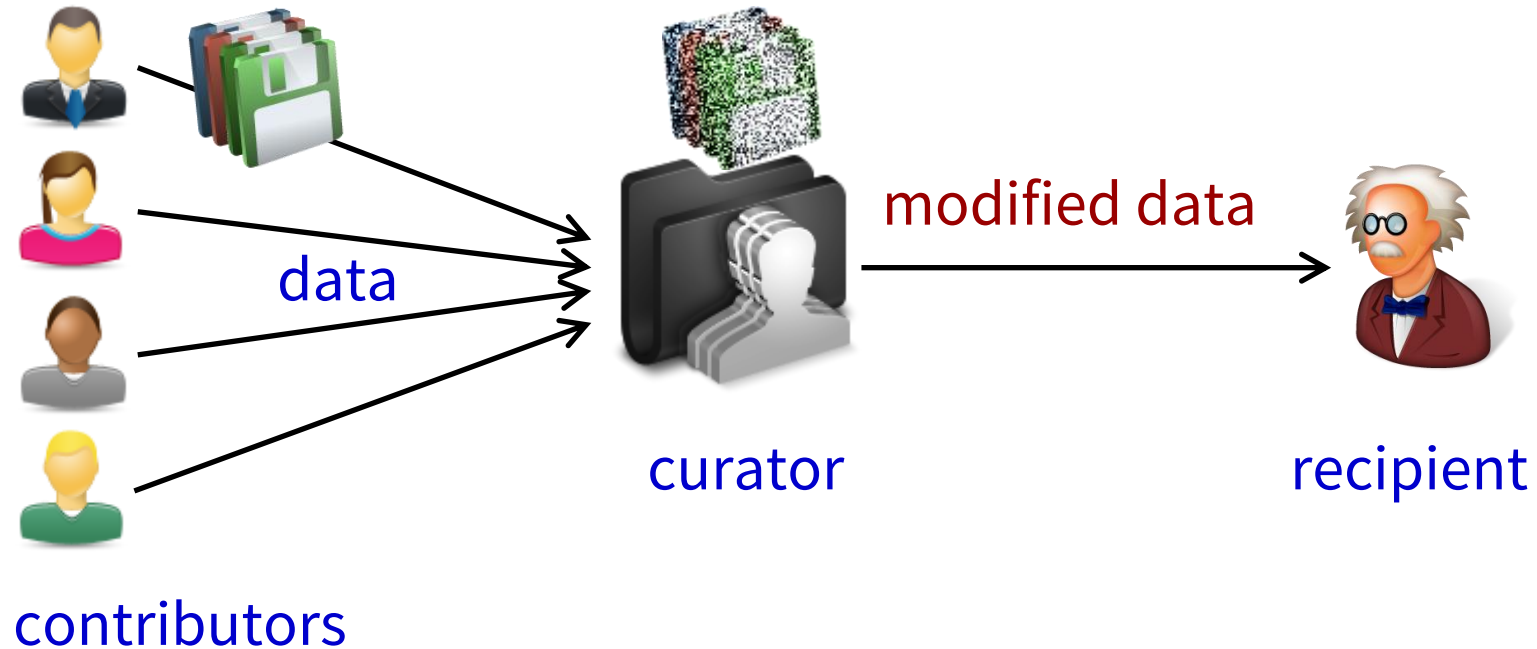
Any information released by the **data curator** can potentially be exploited by the adversary

- Taxi data sharing case: weakly protected driver's name
- MGIC case: genders, birth dates, ZIP codes
- Drug abuse dataset case: search queries
- Voter registration case: security at curator's side

Solution?

- **Do NOT release the exact information from the original data**

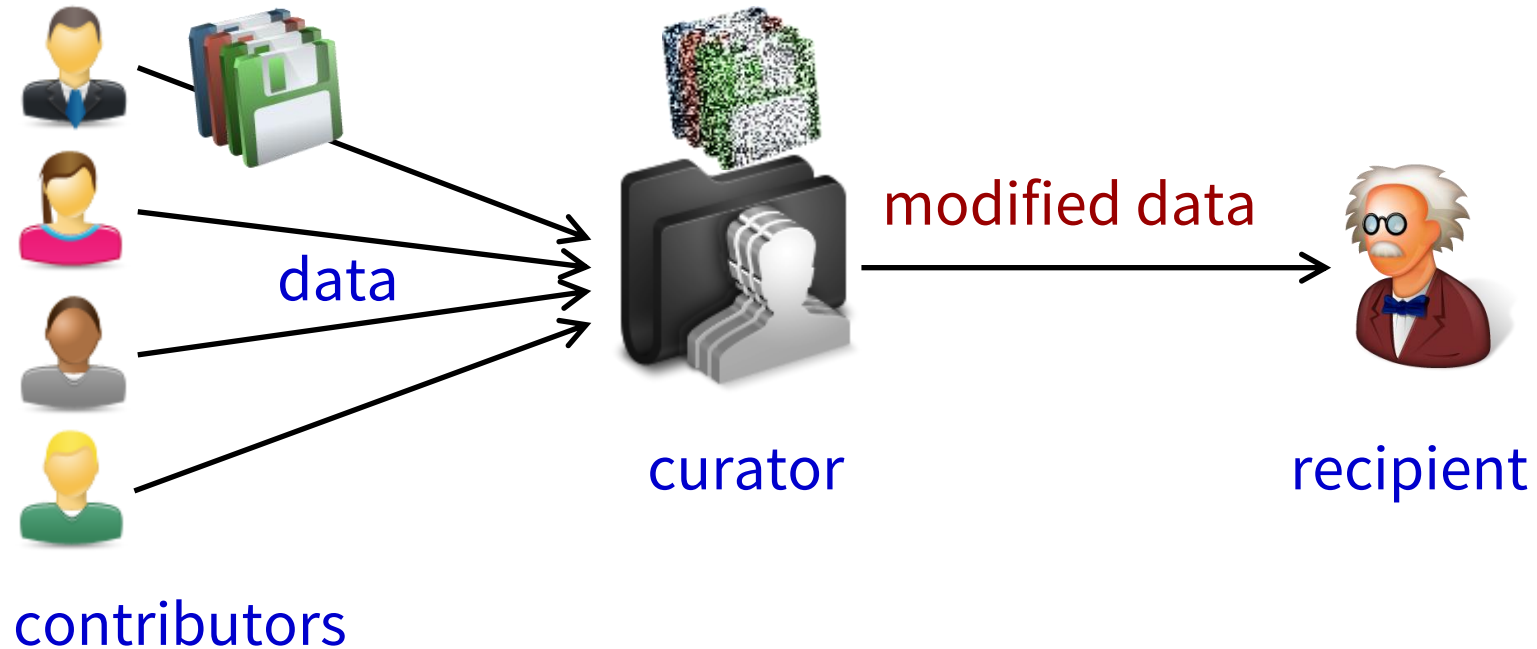
Privacy Preserving Data Publishing



Publish a **modified version** of the data, such that

- the contributors' privacy is “adequately” protected
- the published data is useful for its intended purpose (at least to some degrees)

Privacy Preserving Data Publishing



Two issues

- **privacy principle**: what do we mean by “adequately” protected privacy?
- **modification method**: how should we modify the data to ensure privacy while maximizing utility?

Existing Solutions

Solutions before 2000

- Mostly without a formal privacy model
- Evaluates privacy based on **empirical studies** only

This lecture will focus on solutions with formal privacy models (developed after 2000)

- ***k*-anonymity, *l*-diversity, *t*-closeness**
- **Differential Privacy**

k -Anonymity: Example

Suppose that we want to publish the medical records below

Name	Age	ZIP	Disease
Andy	20	10000	flu
Bob	30	20000	dyspepsia
Cathy	40	30000	pneumonia
Diane	50	40000	gastritis

k -Anonymity: Example

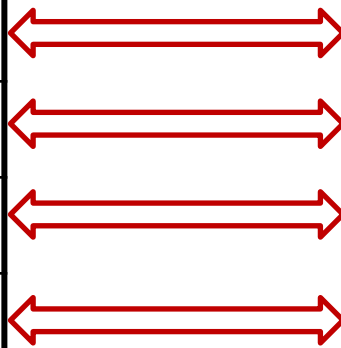
Suppose that we want to publish the medical records below

We know that

- eliminating names is not enough
- because an adversary may identify patients by Age and ZIP

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge



Name	Age	ZIP	Disease
Andy	20	10000	flu
Bob	30	20000	dyspepsia
Cathy	40	30000	pneumonia
Diane	50	40000	gastritis

medical records

k -Anonymity: Example

k -anonymity [Sweeney 2002]

- requires that each individual in a dataset is indistinguishable from $k-1$ others, with respect to their **quasi-identifiers** (Age, ZIP) .

How? - Make Age and ZIP less specific in the medical records

Name	Age	ZIP		Age	ZIP	Disease
Andy	20	10000	↔	20	10000	flu
Bob	30	20000	↔	30	20000	dyspepsia
Cathy	40	30000	↔	40	30000	pneumonia
Diane	50	40000	↔	50	40000	gastritis

adversary's knowledge medical records

k -Anonymity: Example

k -anonymity [Sweeney 2002]

- requires that each individual in a dataset is indistinguishable from $k-1$ others, with respect to their **quasi-identifiers** (Age, ZIP) .

“generalization”

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
-----	-----	---------

20	10000	flu
30	20000	dyspepsia

40	30000	pneumonia
50	40000	gastritis

medical records

$k=2$

k -Anonymity: Example

k -anonymity [Sweeney 2002]

- requires that each individual in a dataset is indistinguishable from $k-1$ others, with respect to their **quasi-identifiers** (Age, ZIP) .

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

$k=2$

k -Anonymity: Example

k -anonymity [Sweeney 2002]

- requires that each (Age, ZIP) combination can be matched to at least k patients

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

k -Anonymity: General Approach

Identify the attributes that the adversary may know

- Referred to as **Quasi-Identifiers (QI)**

Divide tuples in the table into groups of sizes at least k

Generalize the QI values of each group to make them identical

QI

	Age	ZIP	Disease
group 1 {	20	10000	flu
	30	20000	dyspepsia
group 2 {	40	30000	pneumonia
	50	40000	gastritis

medical records

k -Anonymity: General Approach

Identify the attributes that the adversary may know

- Referred to as **Quasi-Identifiers (QI)**

Divide tuples in the table into groups of sizes at least k

Generalize the QI values of each group to make them identical

			QI		
Name	Age	ZIP	Age	ZIP	Disease
Andy	20	10000	[20,30]	[10000,20000]	flu
Bob	30	20000	[20,30]	[10000,20000]	dyspepsia
Cathy	40	30000	[40,50]	[30000,40000]	pneumonia
Diane	50	40000	[40,50]	[30000,40000]	gastritis

adversary's knowledge 2-anonymous table

k -Anonymity: Algorithms

Numerous algorithms for k -anonymity had been proposed

Objective: achieve k -anonymity with the least amount of generalization

This line of research became **obsolete**

Reason: k -anonymity was found to be **vulnerable** [Machanavajjhala et al. 2006]

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

k -Anonymity: Vulnerability

k -anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of size at least k

But it says nothing about the remaining attributes

Result: Disclosure of sensitive attributes is possible

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI		sensitive
Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

k -Anonymity: Vulnerability

k -anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of size at least k

But it says nothing about the remaining attributes

Result: Disclosure of sensitive attributes is possible

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI		sensitive
Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	flu
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

k -Anonymity: Vulnerability

Intuition:

- Hiding in a group of k is not sufficient
- The group should have a **diverse** set of sensitive values

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI		sensitive
Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	flu
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

k -Anonymity – General Idea

- It is a **privacy model** where a dataset is modified to ensure that each record is **indistinguishable** from at least **K-1** other records in the dataset.
- It aims to **prevent re-identification of individuals** by obscuring their identity within a group of similar records.
- Achieving K-Anonymity involves **generalising** or **suppressing** attributes in the dataset to ensure that individuals cannot be uniquely identified based on the released data.

Limitations:

- K-Anonymity is vulnerable when adversaries have knowledge;
- Sensitive values within the anonymised groups could still be revealed.

l -Diversity [Machanavajjhala et al. 2006]

Approach: (similar to k -anonymity)

- Divide tuples into groups, and make the QI of each group identical

Requirement: (different from k -anonymity)

- Each group has at least l “**well-represented**” sensitive values

Several definitions of “well-represented” exist

- Simplest one: in each group, no sensitive value is associated with more than $1/l$ of the tuples

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-diverse table

l-Diversity: Vulnerability

- Suppose that the adversary wants to find out the disease of Bob
- The adversary knows that Bob is **unlikely** to have breast cancer
- So he knows that Bob is **likely** to have diabetes

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	breast cancer
[20,30]	[10000,20000]	diabetes
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-diverse table

l-Diversity: Other Vulnerabilities

l-diversity does not consider overall data distribution (*Skewness Attack*)

- Assume the sensitive attribute is HIV+ or HIV-, and HIV+ is about 1% of the population
- If one class has 25 HIV+ and 25 HIV-, anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.
 - E.g. Bob or Andy has a 50% chance of HIV+

l-diversity does not consider the semantics of sensitive values (*Similarity Attack*)

- All the people in the first group have stomach-related disease

Age	ZIP	Disease
[20,30]	[10000,20000]	HIV+
[20,30]	[10000,20000]	HIV-

2-diverse 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

l-diversity – General Idea

- *l*-Diversity extends the concept of *k*-Anonymity by ensuring that each group of indistinguishable records (as defined by *K*-Anonymity) **contains at least *L* distinct values** for **sensitive attributes**.
- It aims to enhance privacy protection by **preventing attribute disclosure** within anonymised groups.
- By requiring diversity in sensitive attribute values, *L*-Diversity reduces the **risk of inference attacks** where adversaries can exploit the **presence of a single sensitive value** to infer the sensitive information of individuals in the group.

Limitations:

- *l*-diversity may be difficult and unnecessary to achieve (e.g. only two classes).
- *l*-diversity is insufficient to prevent attribute disclosure.

t -Closeness

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

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

What Does 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

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

t -closeness – General Idea

- t -Closeness builds upon k -anonymity and l -diversity by considering the **distribution of sensitive attribute values** within anonymised groups.
 - k -anonymity prevents identity disclosure but **NOT** attribute disclosure
 - l -diversity requires that each eq. class has at least l values for each sensitive attribute, but l -diversity has some limitations
- It ensures that the **distribution of sensitive attribute values** in each group **closely resembles** the overall distribution in the original dataset.
- t -closeness aims to provide stronger privacy guarantees by minimising the discrepancy between the **distributions of sensitive values** in the **anonymised dataset** and the **original dataset**.

Differential Privacy

More general, Support more applications: e.g., *census data*

ANDY GREENBERG SECURITY 06.13.16 07:02 PM

APPLE'S 'DIFFERENTIAL PRIVACY' IS ABOUT COLLECTING YOUR DATA—BUT NOT YOUR DATA

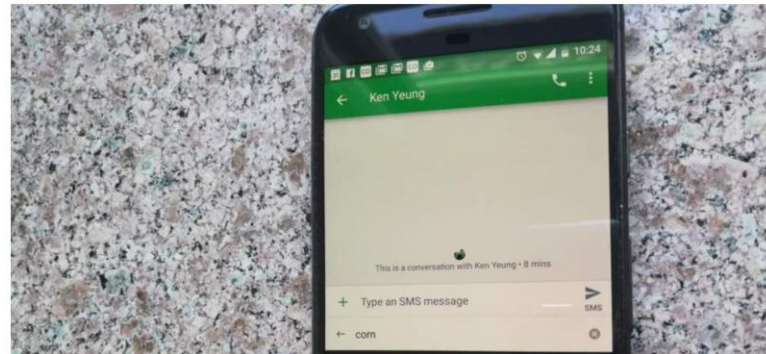


Senior vice president of software engineering Craig Federighi.
JUSTIN KANEPS FOR WIRED

APPLE, LIKE PRACTICALLY every mega-corporation, wants to know as much as possible about its customers. But it's also marketed itself as Silicon Valley's privacy champion, one that—unlike so many of its advertising-driven

Following Apple, Google is exploring differential privacy in Gboard for Android

JORDAN NOVET @JORDANNOVET APRIL 6, 2017 6:50 PM



PUBLISHED MAY 16, 2017 IN RESEARCH

NEW TOOLS SAFEGUARD CENSUS DATA ABOUT WHERE YOU LIVE AND WORK

Algorithms guarantee individual privacy without compromising community insights



Census bureau can still provide accurate information about population demographics without revealing individual-level information.

Differential Privacy: Intuition

Suppose that we have a dataset D that contains the medical record of every individual in Australia

Suppose that Alice is in the dataset

Intuitively, is it OK to publish the following information?

- Whether Alice has diabetes 
- The total number of diabetes patients in D 

Why is it OK to publish the latter but not the former?

Intuition:

- The former completely depends on Alice
- The latter does not depend much on Alice

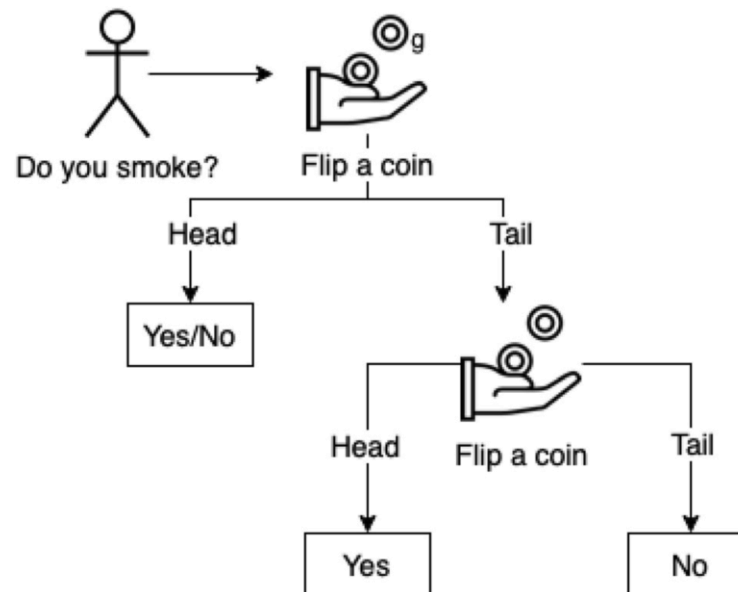
Differential Privacy: Intuition

Goal: nothing about an individual should be learnable from the database without access to the database (?)

Differential privacy: the risk to one's privacy should not substantially increase as a result of participating in a statistical database

Coin Toss Algorithm

Deniability by Randomization



Using Randomized Algorithms

Differential confidentiality is a process that introduces **randomness** into the data

Example: *Are you over 35 years old?*

- Throw a coin
- If head, then answer honestly
- If tail, then throw the coin again and answer "Yes" if head, "No" if tail

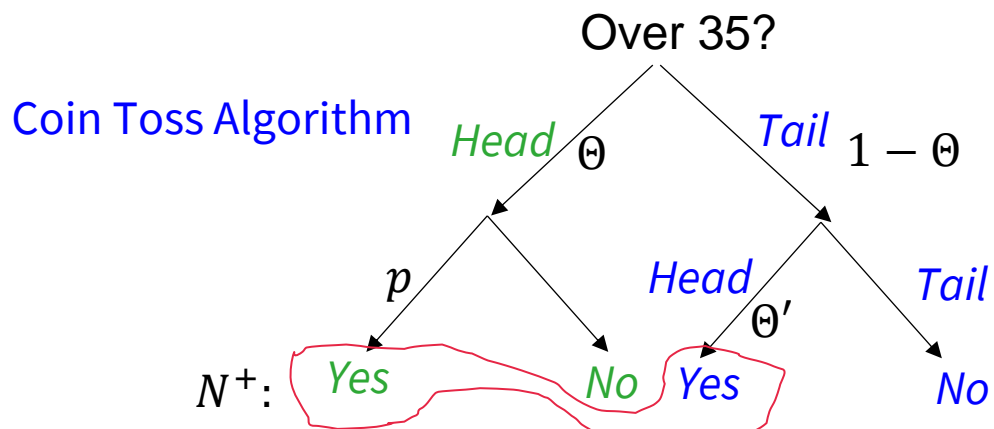
The confidentiality arises from the **refutability** of the individual responses

Individual's **deniability** is provided via the randomization.

Data Utility

Data with many responses are significant

- Positive responses are given to a quarter of people who are under 35 and three-quarters by people who are over 35
- Given a sufficiently large number of responses, can we estimate the true proportion of people over 35 years old (denoted as p) from the observed proportion of people answering “yes” (denoted as q)?



N : the entire response number

N^+ : the entire positive response number

N_t^+ : the true positive response number (true over 35)

N_f^+ : the false positive response number (false over 35)

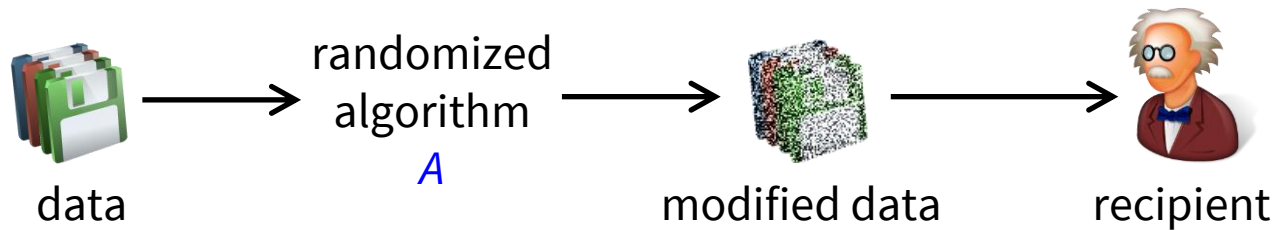
$$N^+ = N_t^+ + N_f^+$$

$$N * q = N * \theta * p + N(1 - \theta) * \theta'$$

$$\theta, \theta' \rightarrow 1/2$$

We expect to obtain $q = (1/4) + p/2$ positive responses

Differential Privacy: Definition

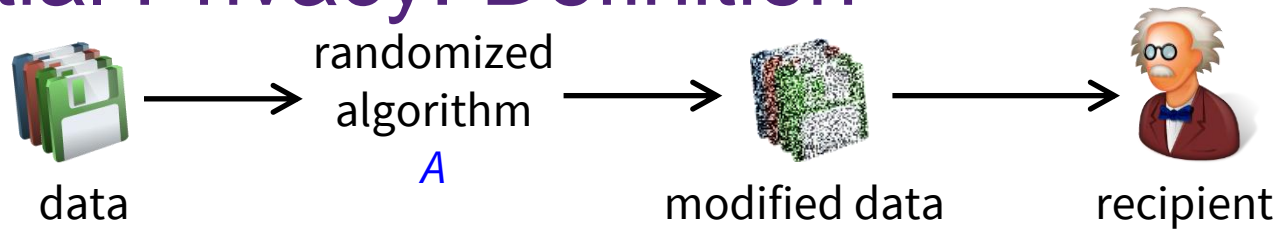


Neighboring datasets

- Two datasets D and D' , such that D' can be obtained by changing one single tuple in D
- If the mechanism behaves **nearly identically** for D and D'
- Attacker can't tell which one was used



Differential Privacy: Definition



- **Definition:** A randomized algorithm A satisfies ϵ -differential privacy, iff for any two neighboring datasets D and D' and for any output O of A ,

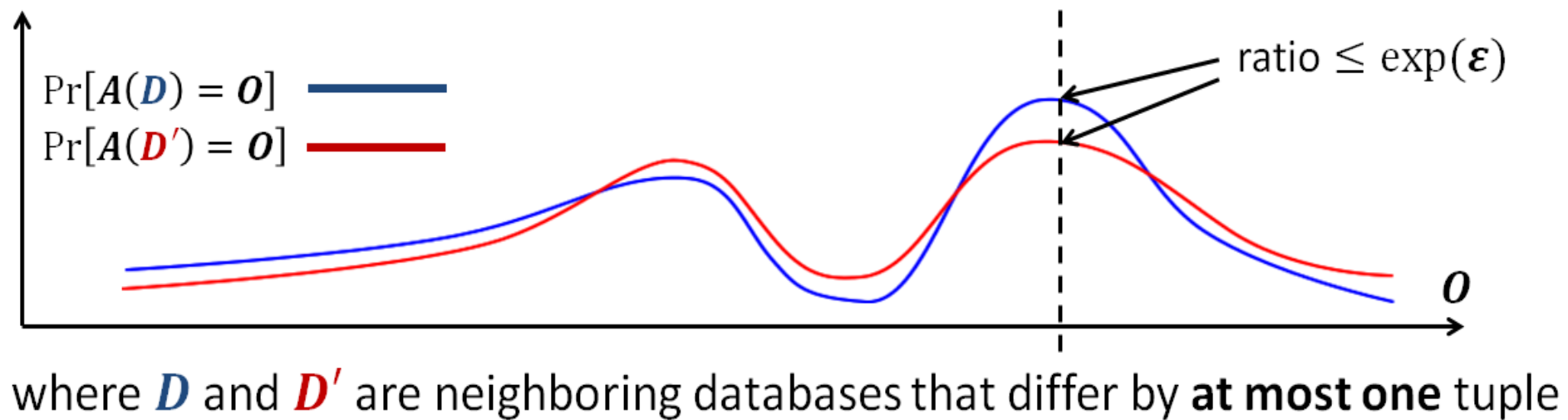
$$\Pr[A(D) = O] \leq \exp(\epsilon) \cdot \Pr[A(D') = O]$$

Control the strength of privacy protection

- **Rationale:** The output of the algorithm does not highly depend on any particular tuple in the input

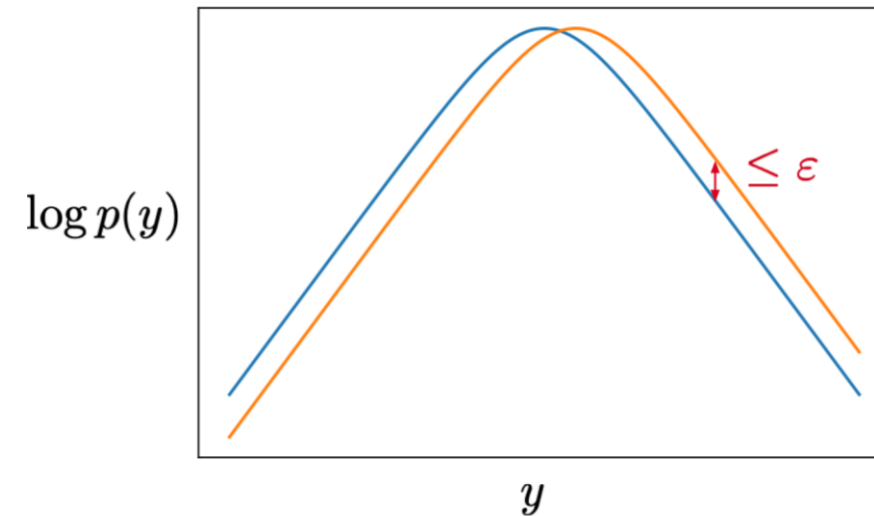
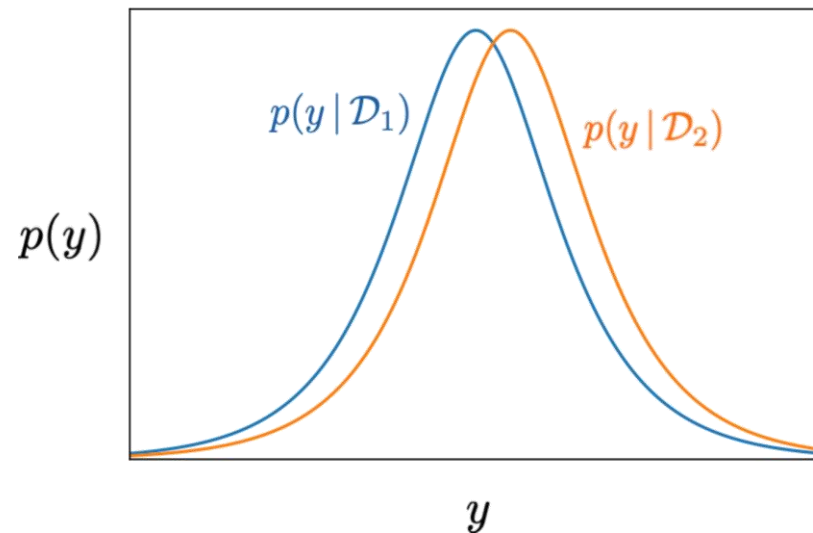
Differential Privacy: Illustration

Illustration of ϵ -differential privacy



Differential Privacy: Illustration

Illustration of ϵ -differential privacy



An Example in Electronic Health Records

Suppose that we have a set D of medical records

We want to release statistical information, e.g., the number of diabetes patients in D :

$f(D) = \text{select count(*) from } D \text{ where disease = "diabetes";}$
(say we have 1,000 diabetes patients in the dataset)

Example: How to Release Data

Non-private solution: Release $f(\mathbf{D})$ directly

But **it violates differential privacy**, since

$$\Pr[f(\mathbf{D}) = \mathbf{1000}] \leq \exp(\epsilon) \cdot \Pr[f(\mathbf{D}') = \mathbf{1001}]$$

does not hold

How to do it in a differentially private manner?

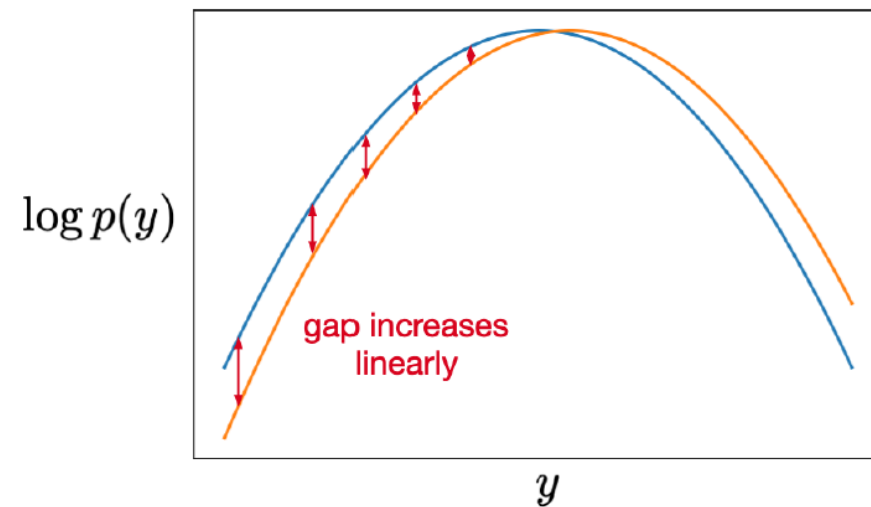
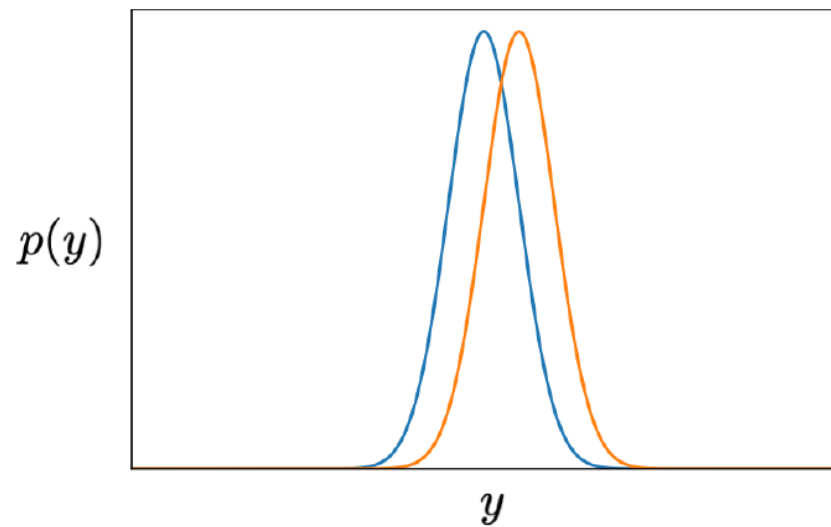
Injecting noise into every count before releasing it

- $A(\mathbf{D}, f) = f(\mathbf{D}) + \text{Noise}$

Question: what kind of **noise** should we add?

Gaussian Noise

Attempt 1: Gaussian Noise

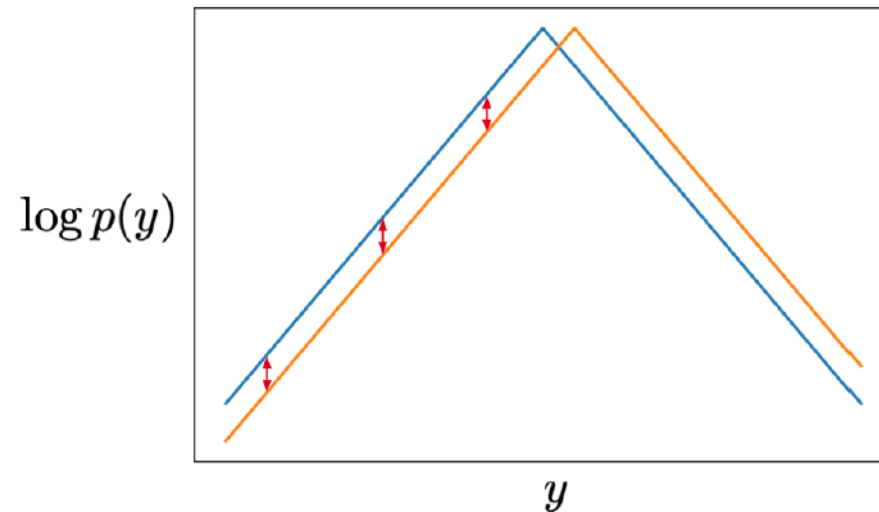
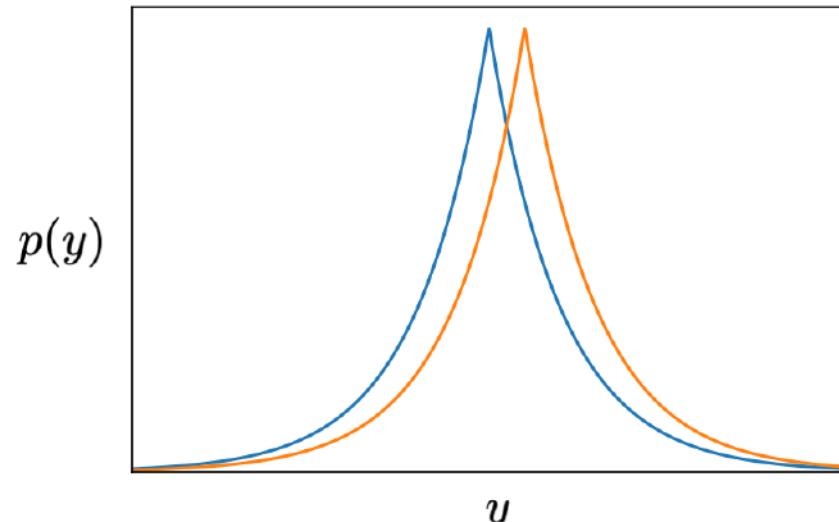


- Gaussian noise violates our definition, but only because of the tails.
- Handles High-Dimensional Data Well
- Easier to Implement in Some Systems

Laplace mechanism

Attempt 2: Laplace Distribution

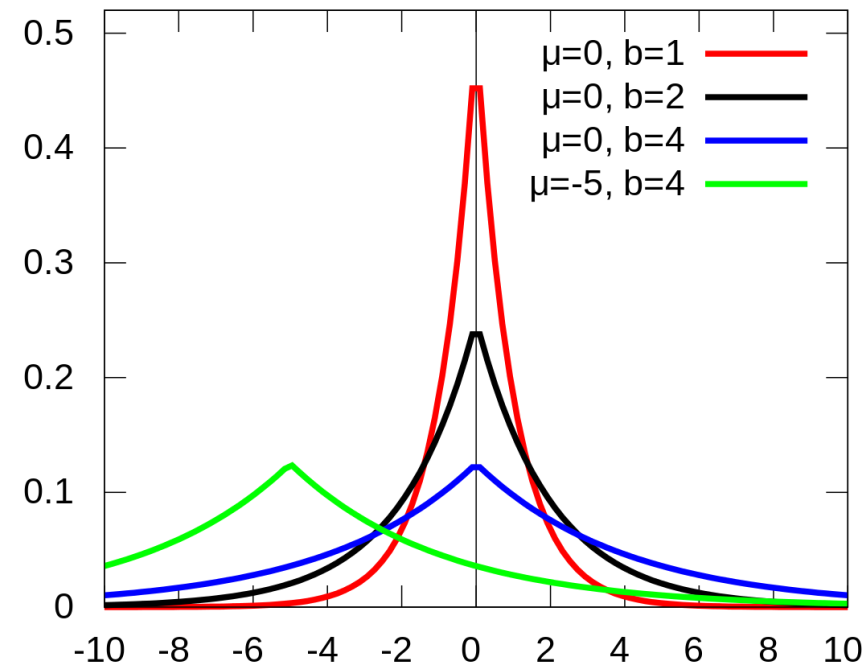
$$p(y; \mu, b) = \frac{1}{2b} \exp \left(-\frac{|y - \mu|}{b} \right)$$



- b is a parameter that determines the scale of the distribution.
- **Strongest Privacy:** Laplacian ensures that the presence or absence of a single individual has a bounded influence on output.
- But it is **not always practical**, especially when the output needs to be bounded (% in 0 and 100, age between 0 and 120). **Very large or very small noise** leads to no sense for the bounded data.

Laplace mechanism

Noise $\sim \text{Lap}(b)$, i.e., the noise are i.i.d. drawn from the Laplace distribution with b



$\text{Lap}(b)$ b is referred as the *scale*

Sensitivity

Sensitivity of a function f

$$\Delta f = \max_{D, D'} |f(D) - f(D')|$$

Sensitivity captures **how much one person's data can affect the output**

What is sensitivity for counting query?

$$\Delta f = 1$$

Adding Laplace Noise

Add Laplace noise with $b = \Delta f / \epsilon$ before releasing the number of diabetes patients in D

- Noise depends on f and ϵ , not on the dataset

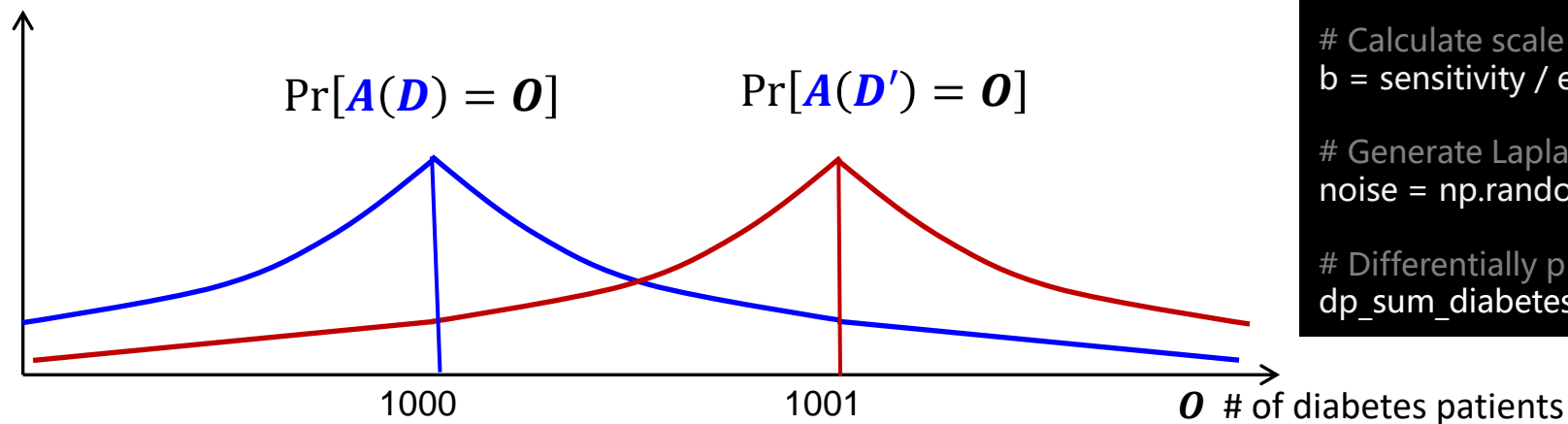
$$y_i \sim \text{Laplace} \left(f(\mathcal{D})_i, \frac{\Delta f}{\epsilon} \right)$$

The noise is **calibrated** to the privacy requirement: **higher sensitivity queries** and **tighter privacy constraints** imply more noise.

Adding Laplace Noise

Add Laplace noise with $b = \Delta f / \epsilon$ before releasing the number of diabetes patients in D

➤ Noise depends on f and ϵ



```
import numpy as np

# Original query result (the true sum pts)
true_sum_diabetes_pts = np.sum(pts['diabetes']) # 1000 in this case

# Sensitivity for removing one pt
sensitivity = 1

# Privacy budget
epsilon = 0.1 # A small epsilon represents stronger privacy

# Calculate scale parameter for Laplace distribution
b = sensitivity / epsilon

# Generate Laplace noise
noise = np.random.laplace(0, b, 1)

# Differentially private mean salary
dp_sum_diabetes_pts = true_sum_diabetes_pts + noise
```

array([1008.39428328])

Statistical Attack Revisited

Attack queries

```
select count(*)  
from staff  
where title = "Professor"
```

```
select sum(salary)  
from staff  
where title = "Professor"
```

Plausible deniability

- an individual's presence or absence in a dataset cannot be confidently inferred, even if the attackers know the results of a query.

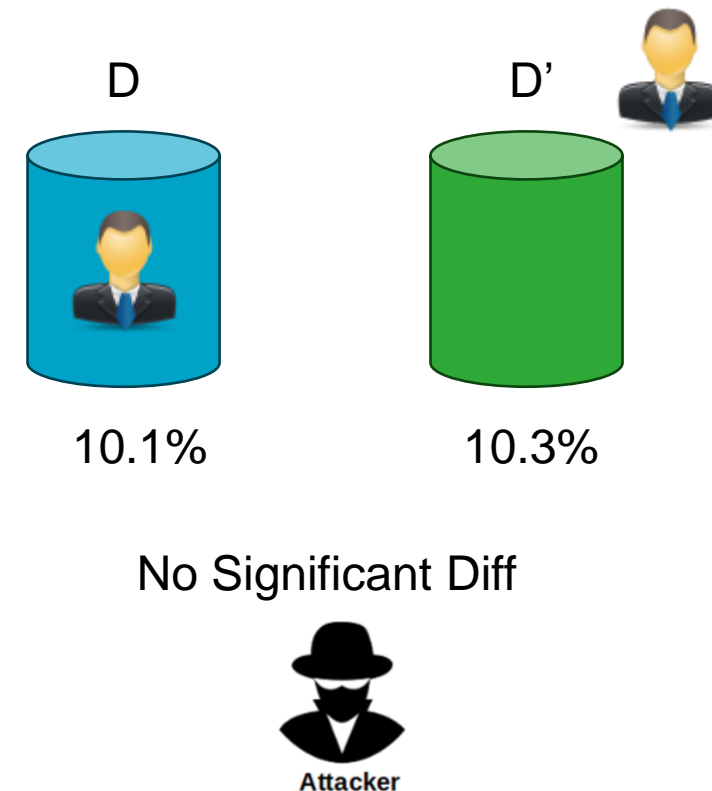
Comparison with k -anonymity, l -diversity and t -closeness

Differential privacy does not directly model the adversary's knowledge

- Can achieve ϵ -DP by adding a **random noise** value
- Uncertainty due to noise \rightarrow **plausible deniability**

It is more general

- There is no restriction on the type of the output O
- It can be a number, a table, a set of frequent itemsets, a regression model, etc.



Limitations of DP

Utility loss

- Differential privacy works by adding noise to the output of queries on the dataset, which can lead to a loss of accuracy or utility (especially when dataset is small).

Computational complexity

- Differential privacy requires computationally expensive operations, particularly when working with large datasets.
- impractical in real-time or high-throughput scenarios.

Beyond Data Privacy...

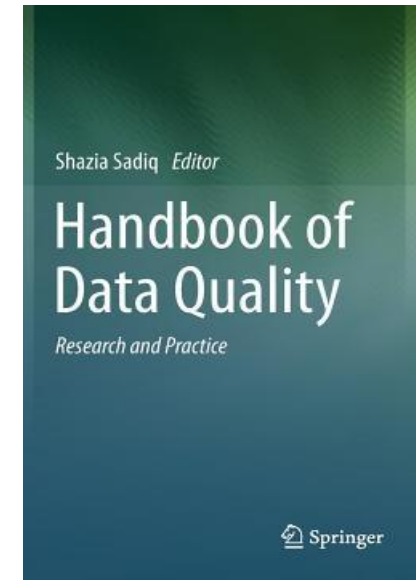
Data Quality Dimensions

- Concept of data quality
- Data life cycle

Management of Data quality

- Measurement of data quality costs and improvements
- Strategies for data quality improvement
- Maturity models and data governance
- Computational approaches

Four Basic Steps of Data Governance



Data Quality Issues: Example

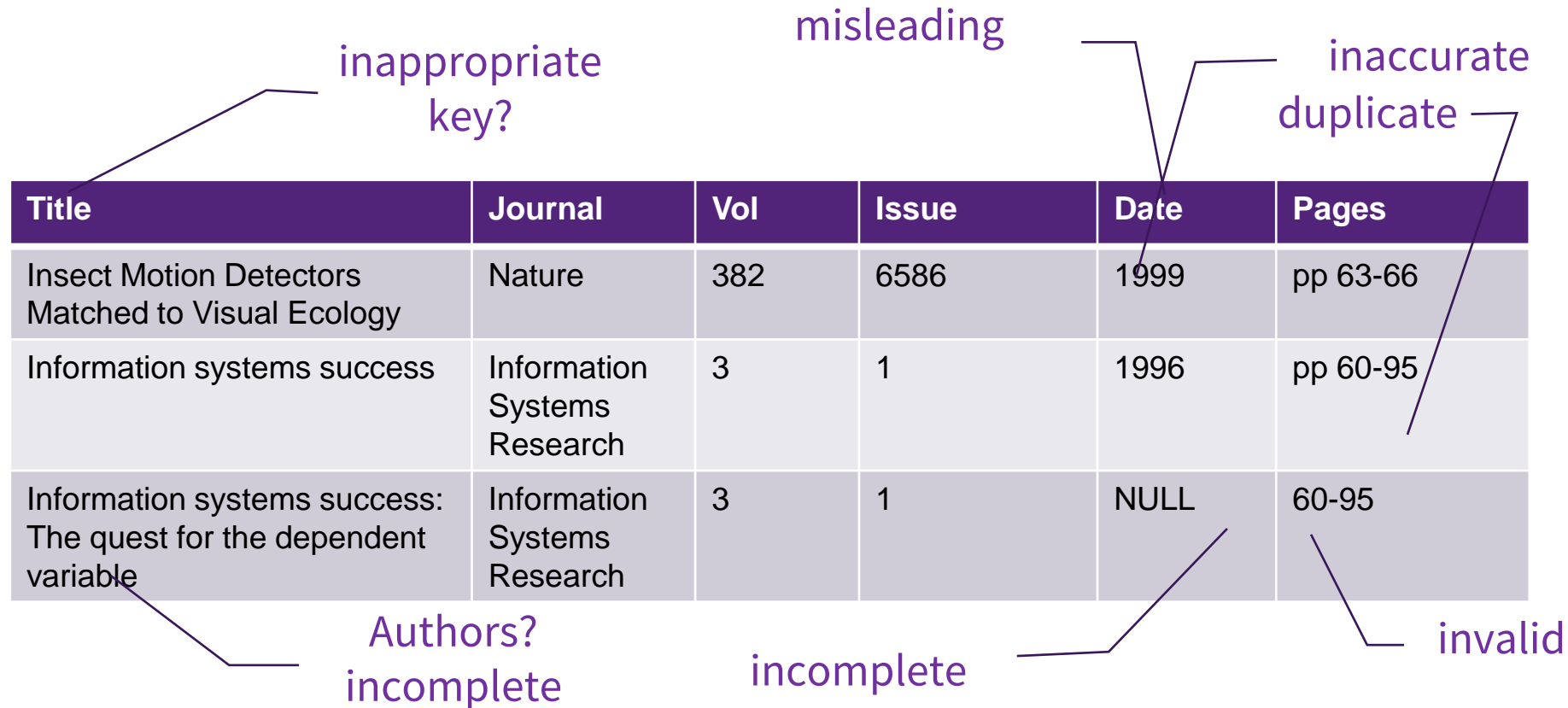


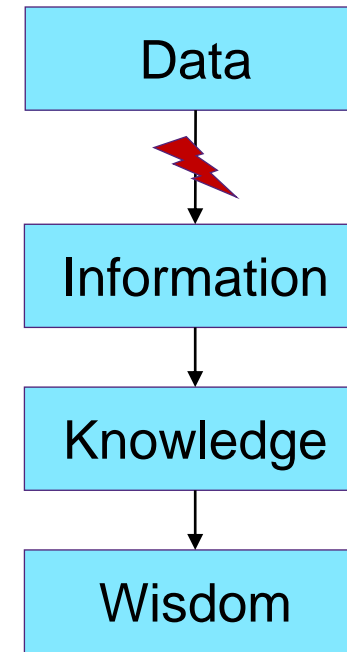
Diagram illustrating data quality issues in a table:

- inappropriate key?**: Points to the **Title** column header.
- misleading**: Points to the **Date** column header.
- inaccurate duplicate**: Points to the **Pages** column header.
- Authors? incomplete**: Points to the **Title** cell of the third row.
- incomplete**: Points to the **Date** cell of the third row (NULL).
- invalid**: Points to the **Pages** cell of the third row (60-95).

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	pp 63-66
Information systems success	Information Systems Research	3	1	1996	pp 60-95
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

Impact of Data Quality

- Customer Dissatisfaction
- High Costs
- Low Job Satisfaction
- Organizational Mistrust
- Poor Decision Making
- Impedes Re-engineering
- Hinders Long-term Business Strategy



→ Breaking the Information Chain

D. Tapscott (1996) The Digital Economy: Promise and Peril in the Age of Networked Intelligence, New York: McGraw-Hill, 1996.

Redman (1996) Data Quality for the Information age. Artech House, Norwood, MA., USA. 1996.

Data Quality Dimensions

Integrity (Meaningless)

Accuracy (Erroneous)

- Postcode “4109” is typed “4019”

Completeness (Missing)

- Students don’t have to declare a major till graduation, so major is missing in most enrolments

Currency (Obsolete)

- Old phone numbers

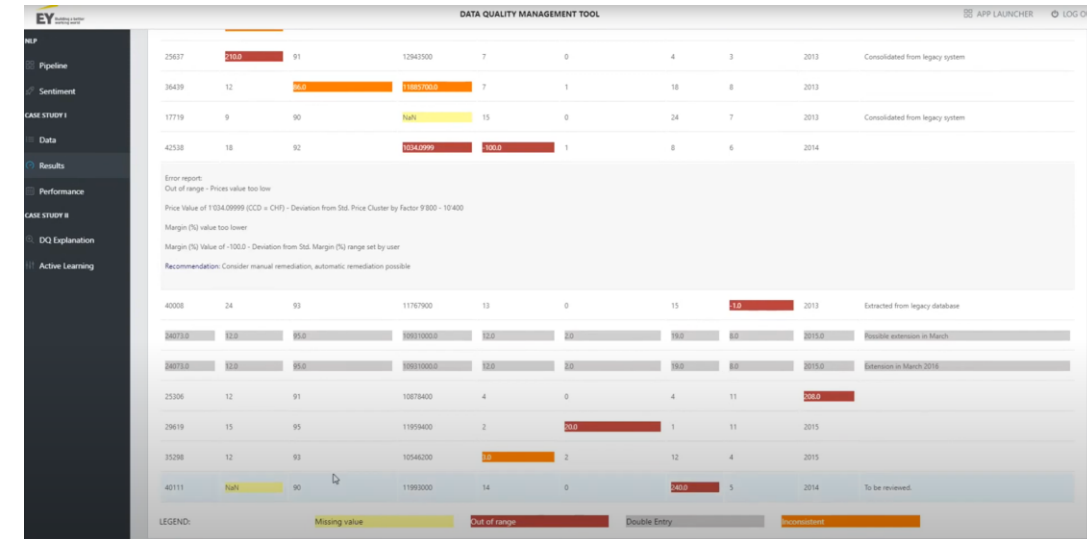
Representational Consistency (Inconsistent)

- EECS Vs Electrical Engineering and Computer Science

Accessibility (Unavailable)

- Server down, privacy concerns

Reliability & Trust (Uncertainty)



Patient_ID	Name	DOB	Blood_Type	Diagnosis
1	John Doe	21-07-1985	O+	Diabetes
1	Jane Doe	15-01-1990	AB-	Hypertension
3	NULL	31-12-3000	XZ	Unknown

Maintaining Data Integrity

- Static Constraints

- True for the **lifetime of the database**
- Control the meaning of data
- Specify the relationships between data

- Dynamic Constraints

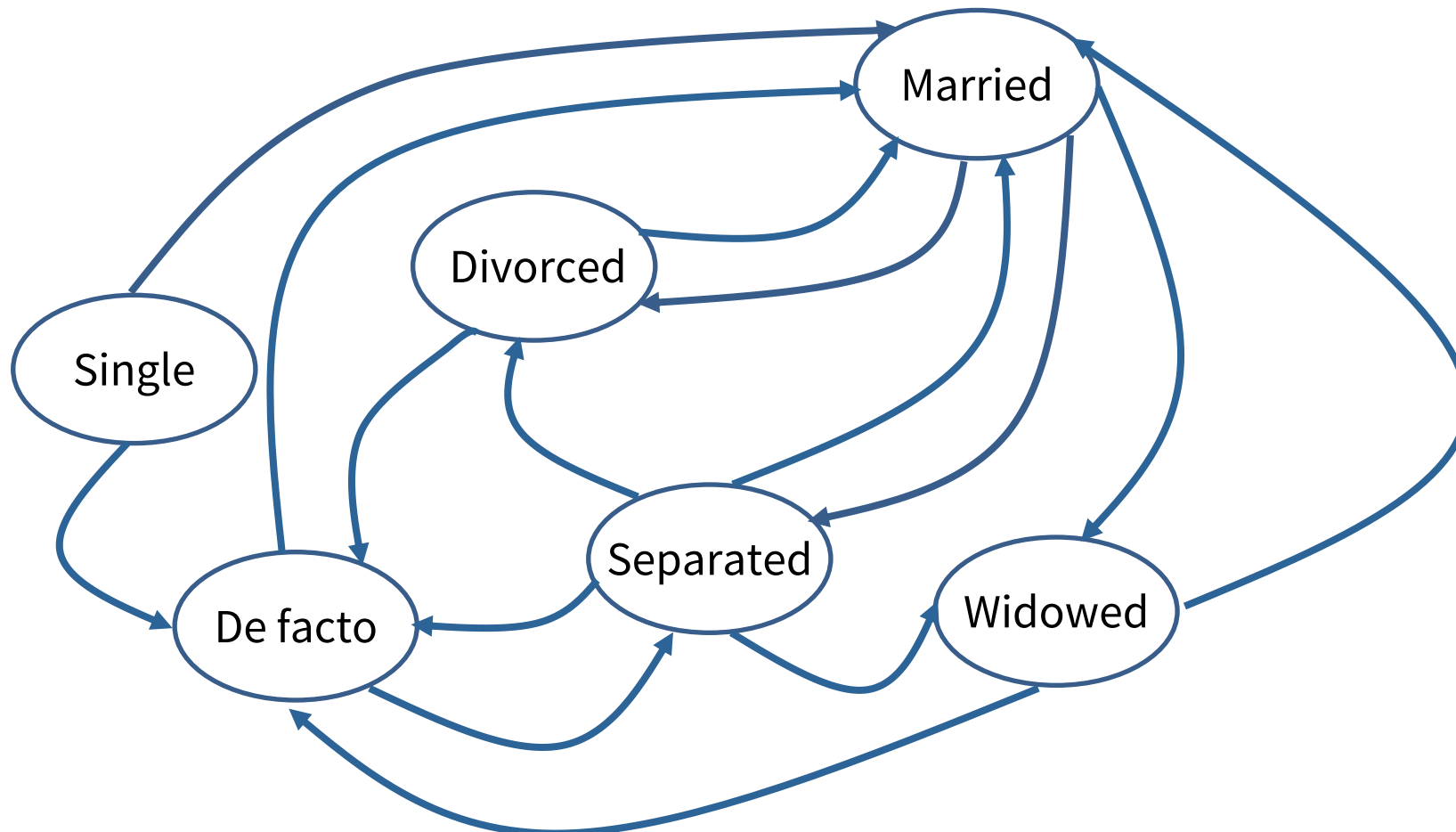
- True for the **database operations**
- Control the transitions between “database snapshots” – **states** of database
- Specify the legal usage of data

Patient_ID	Name	DOB	Blood_Type	Diagnosis
1	John Doe	21-07-1985	O+	Diabetes
1	Jane Doe	15-01-1990	AB-	Hypertension
3	NULL	31-12-3000	XZ	Unknown

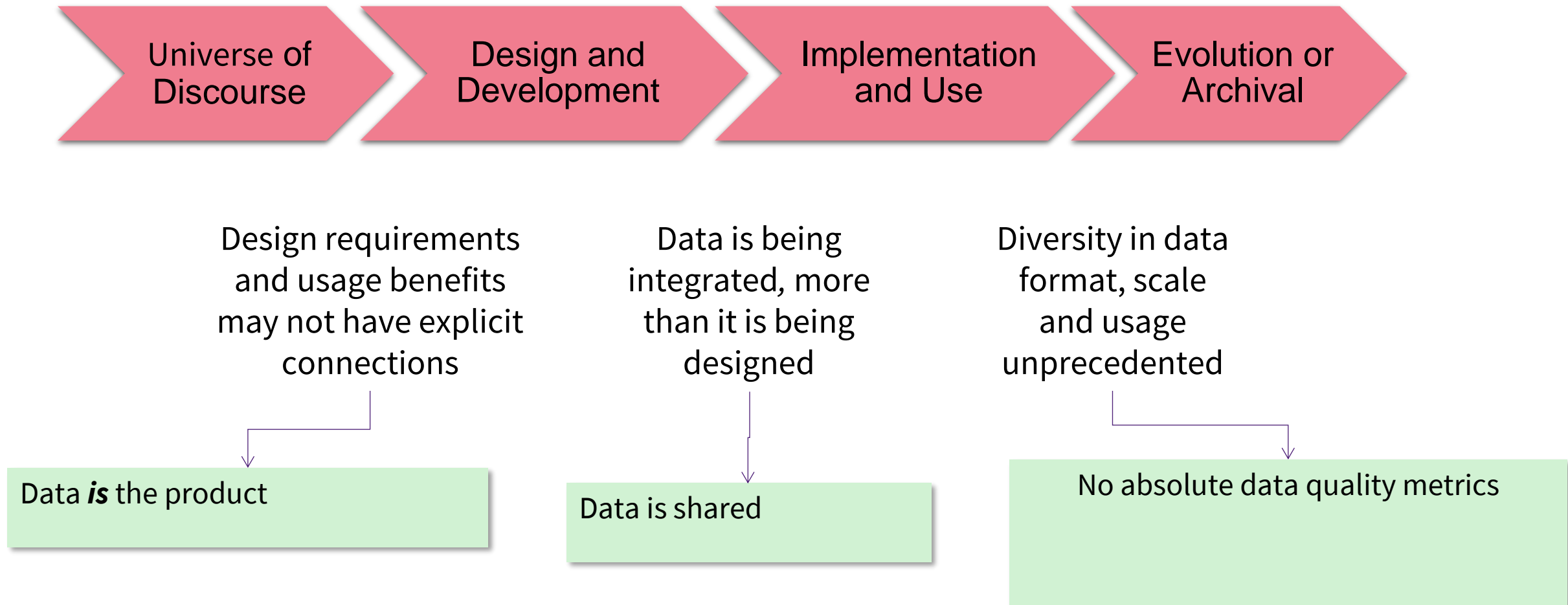
e.g. Discharge date must be “ \geq ” admission date

Dynamic Data Integrity

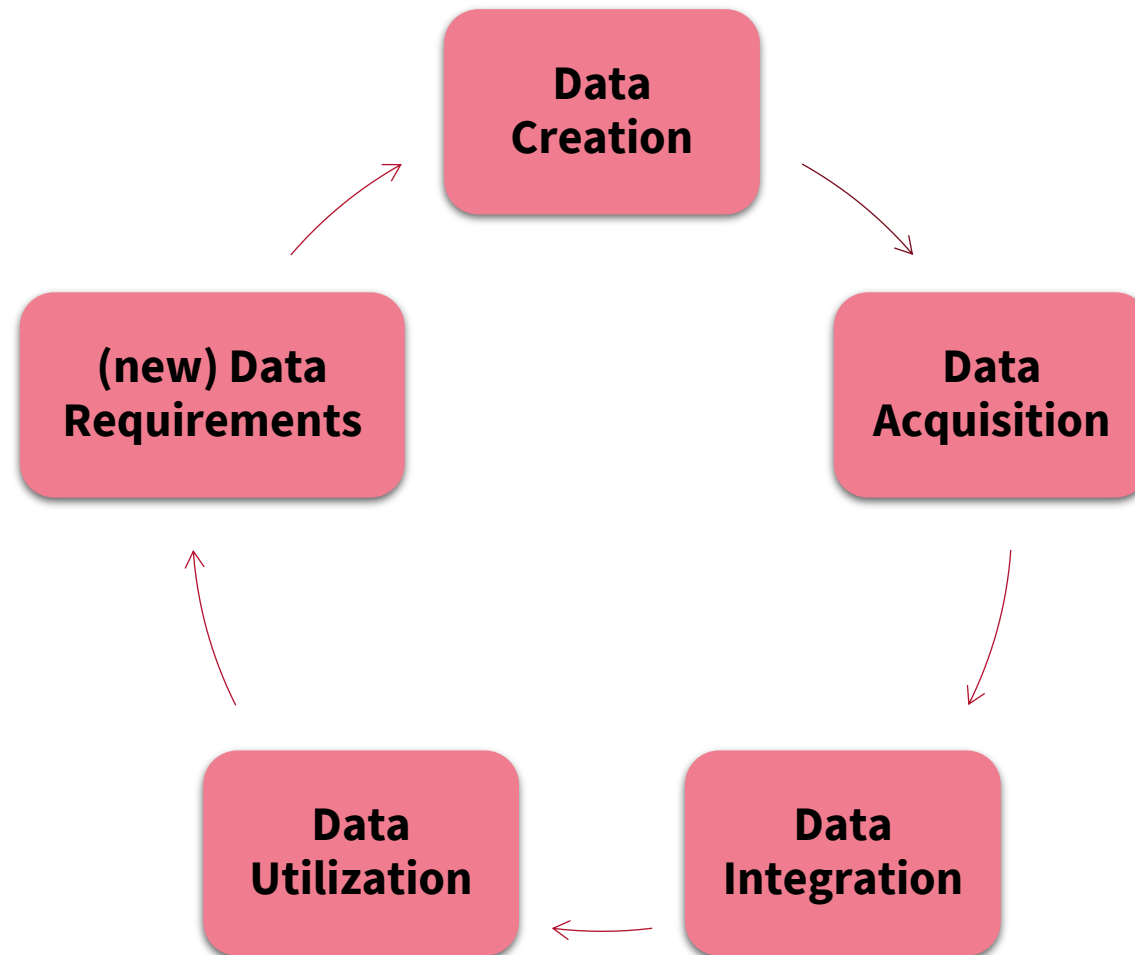
Question: Can we modify one's marriage status in a database without considering this pattern?



Data Life Cycle: an Old & Static View



Data Life Cycle: a New & Dynamic View

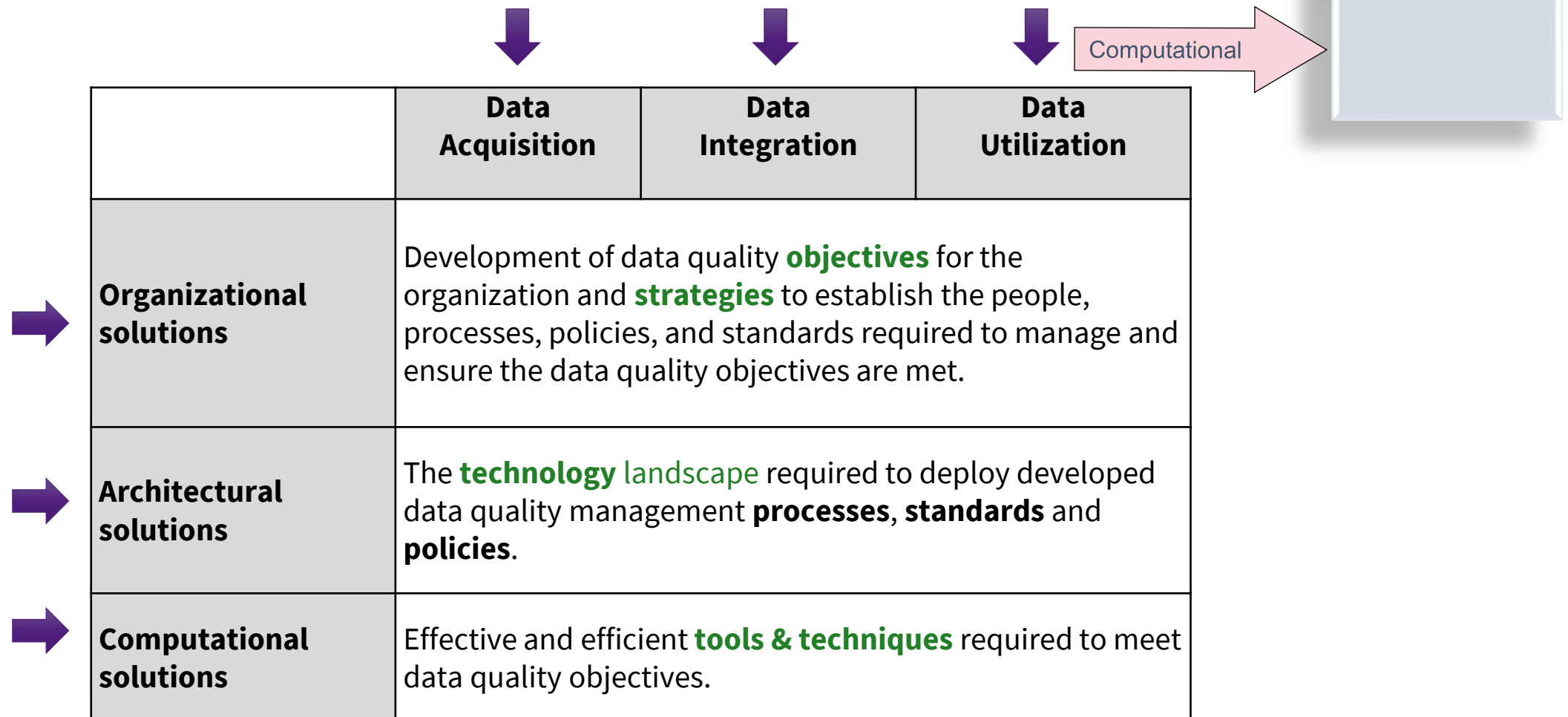


Data Quality Problems

Data Quality w.r.t. three aspects of its life cycle:

Data Acquisition	Data Integration	Data Utilization
<p>Data entry errors, poor input validation, lack of user training</p> <p>Character recognition in document management systems</p> <p>Performance of RFID readers</p> <p>Dealing with ambient noise in sensor network communication...</p>	<p>Differences in format</p> <p>Structural (schema) and semantic (constraints, values) mismatches</p> <p>Privacy preservation, access control</p> <p>Error propagation</p> <p>Attribution, audit, tracking, lineage...</p>	<p>Informational overload/ relevance/usability</p> <p>Real time analytics, business intelligence</p> <p>Data mining & statistical (un)truths</p> <p>Privacy & trust...</p> <p>Tracing and Tracking, ...</p>

Problems → Solutions

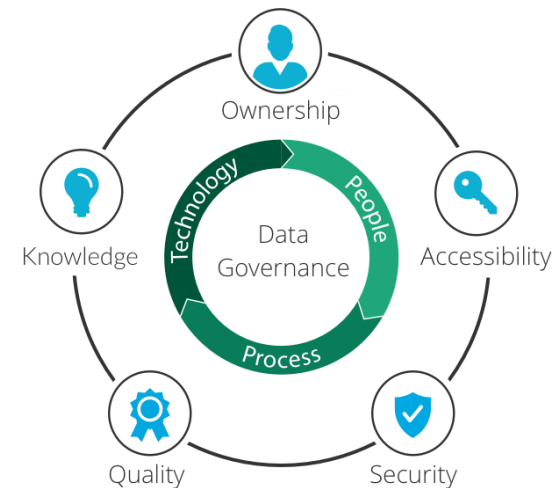


Data Governance

Data Governance: is the practice of identifying important data across an organization, ensuring it is of high quality, and improving its value to the business.

Data Government Policy: is a document that formally outlines how organizational data will be managed and controlled. A few common areas covered by data governance policies are:

- **Data quality** – ensuring data is correct, consistent and free of “noise” that might impeded usage and analysis.
- **Data availability** – ensuring that data is available and easy to consume by the business functions that require it.
- **Data usability** – ensuring data is clearly structured, documented and labeled, enables easy search and retrieval, and is compatible with tools used by business users.
- **Data integrity** – ensuring data retains its essential qualities even as it is stored, converted, transferred and viewed across different platforms.
- **Data security** – ensuring data is classified according to its sensitivity, and defining processes for safeguarding information and preventing data loss and leakage.



Governance vs. Management

"Governance" is the **strategic task** of setting the organisation's goals, direction, limitations and accountability frameworks.

"Management" is the **allocation of resources** and overseeing the day-to-day operations of the organisation.

One way to think about this is that

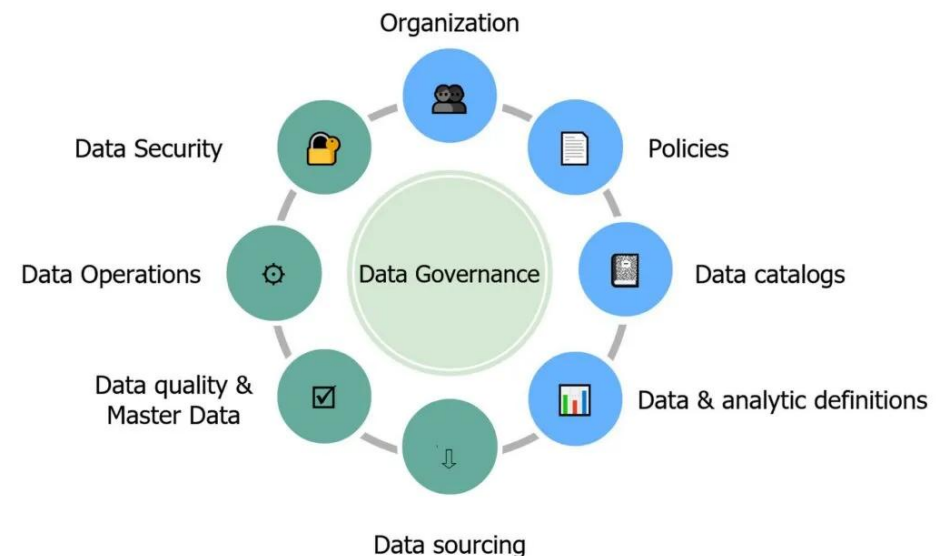
- **Governance** determines the **"What?"** - what the organisation does and what it should become in the future.
- **Management** determines the **"How?"** - how the organisation will reach those goals and aspirations.

Optional

Four basic Steps of Data Governance w.r.t. Quality (1/4)

There are four steps in data governance w.r.t. quality:

- **Recognize the problem**
- Measure its costs
- Devise strategy for improvement
- Aim towards data governance maturity



Optional

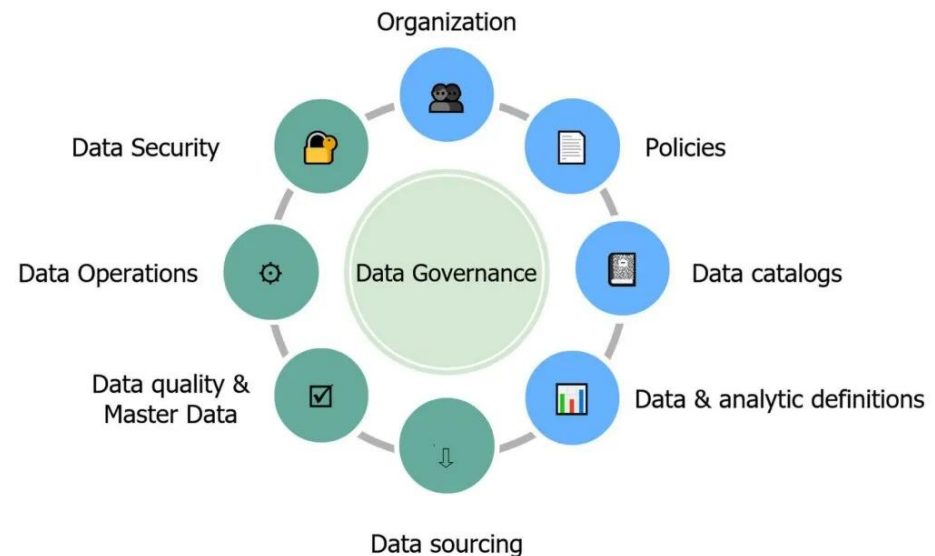
Four basic Steps of Data Governance w.r.t. Quality (2/4)

Recognize the problem

Measure its costs

Devise strategy for improvement

Aim towards data governance maturity



Optional

Each Error is a Cost

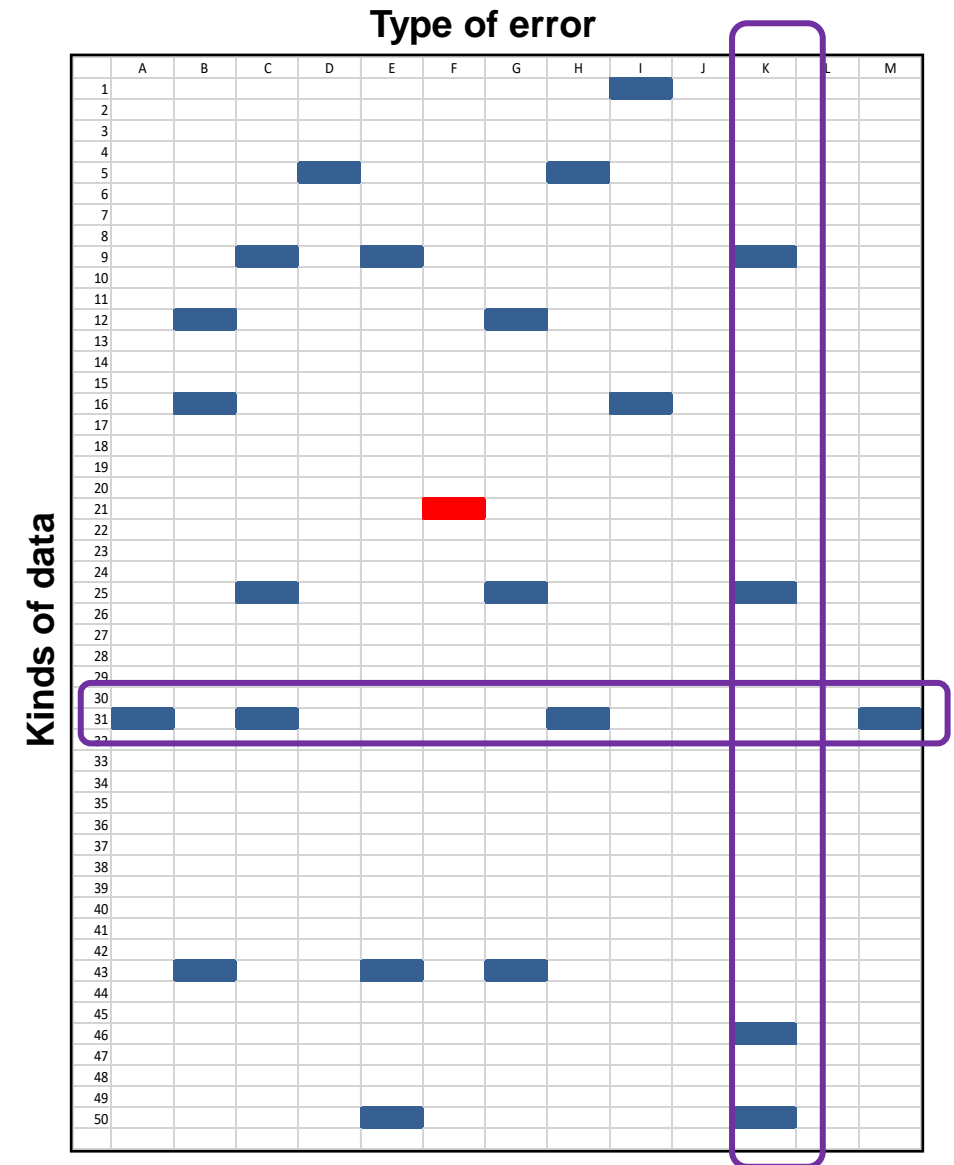
Quantitative analysis:

How much is each error costing the organization?

How much will the organization save if the errors are eliminated (reduced)

Qualitative Analysis:

- Any low frequency but high impact errors?
- Errors to be dealt with passively or proactively?

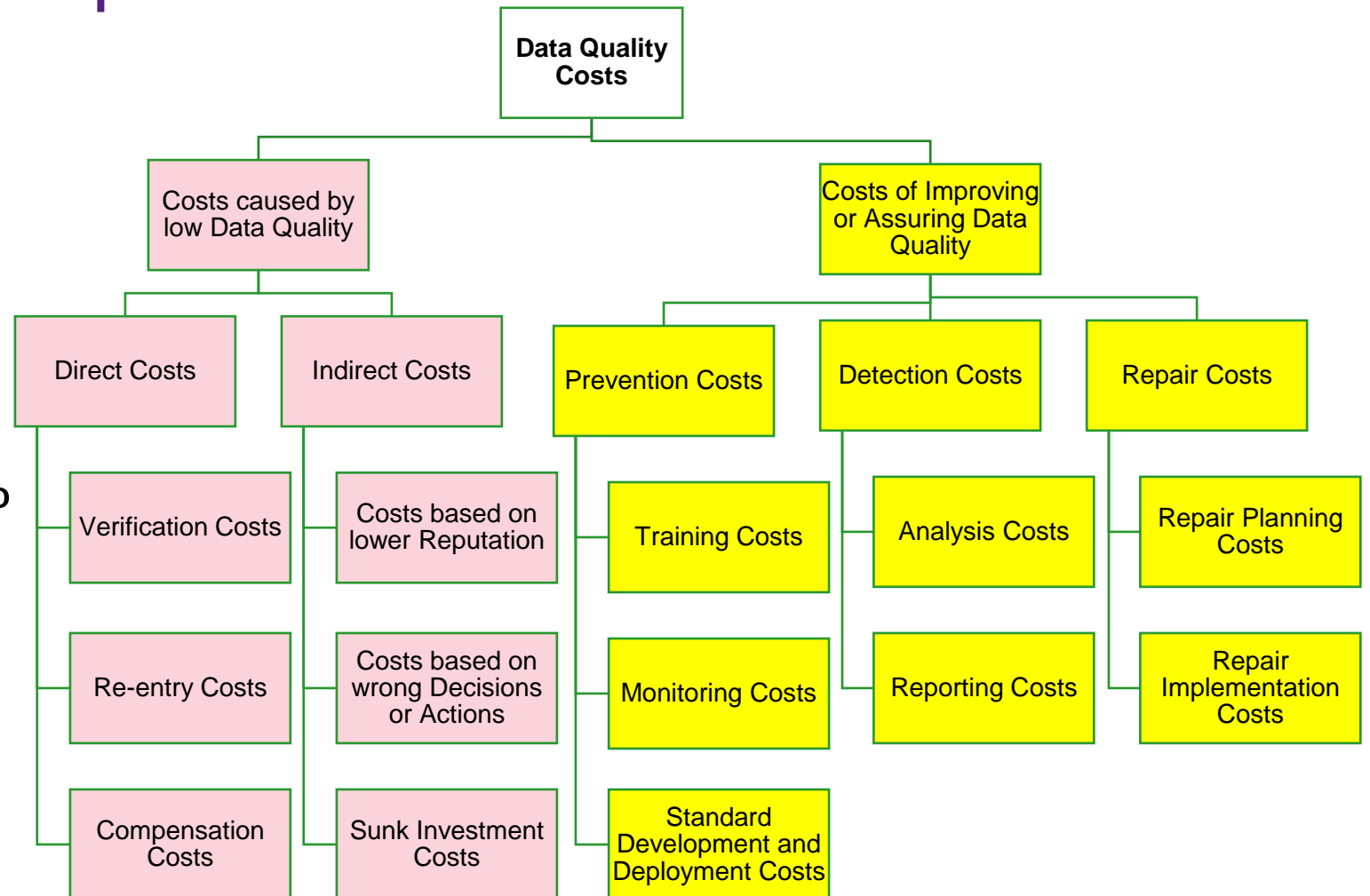


Optional Measuring Cost: Example

Metric: Number of duplicates

Cost (Benefit): Cost of
returned mail e.g. in a
marketing campaign

Success threshold: Reduce
number of duplicates by 15%



Optional

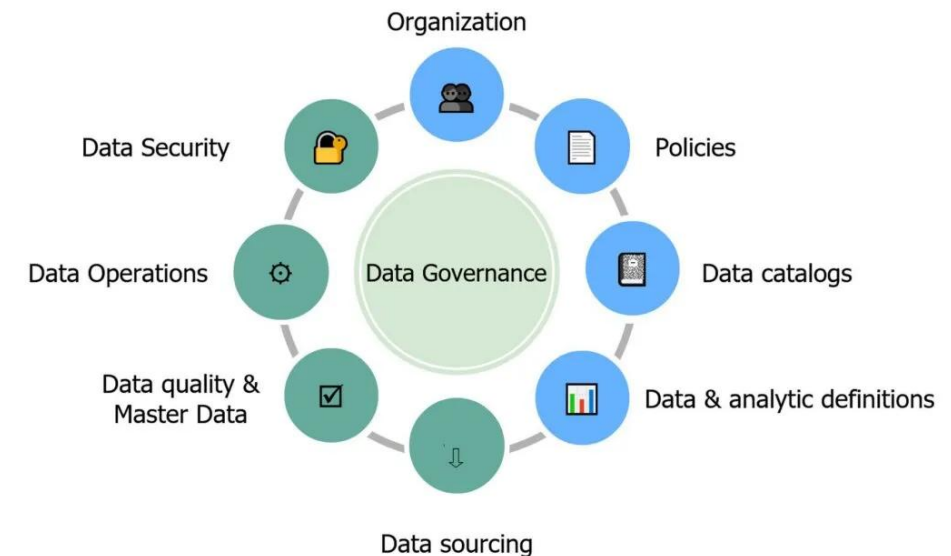
Four basic Steps of Data Governance w.r.t. Quality (3/4)

Recognize the problem

Measure its costs

Devise strategy for improvement

Aim towards data governance maturity



Optional Strategy for Improvement – Short Term

Strategy for improvement (short term)

- Define **metrics** and its relationship to business impact to identify which data to improve
- Produce **baseline**
- Use same metric (periodically/real time) to measure **change from baseline**
- Sustain improvements through **ongoing monitoring**

Optional Strategy for Improvement – Long Term

Strategy for improvement (long term)

- Establish process owner and management **team**
- Describe process **qualitatively** and understand requirements
- Establish **measurement system**
- Establish **process control** and conformance to requirements
- Identify **improvement opportunities**
- Select opportunities and set objectives regarding each
- Make and **sustain** improvements

Optional

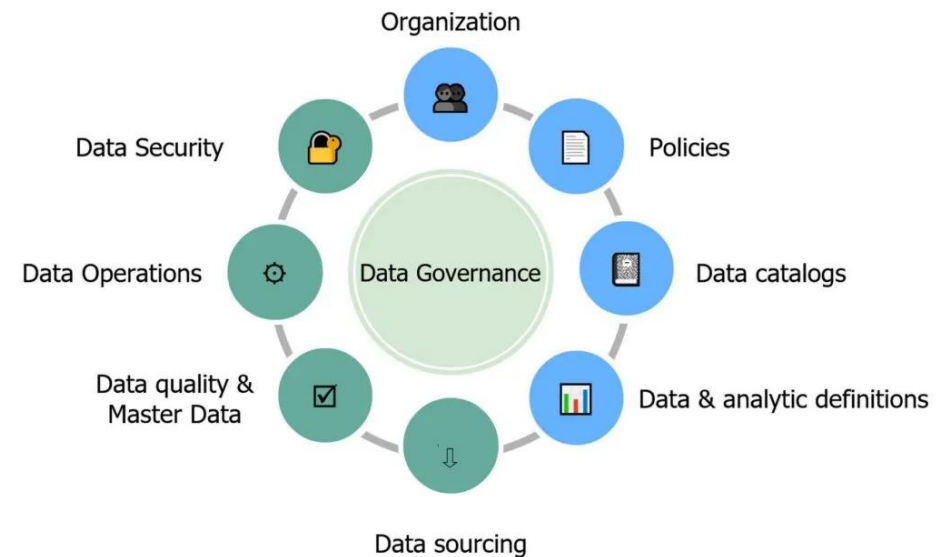
Four basic Steps of Data Governance w.r.t. Quality (4/4)

Recognize the problem

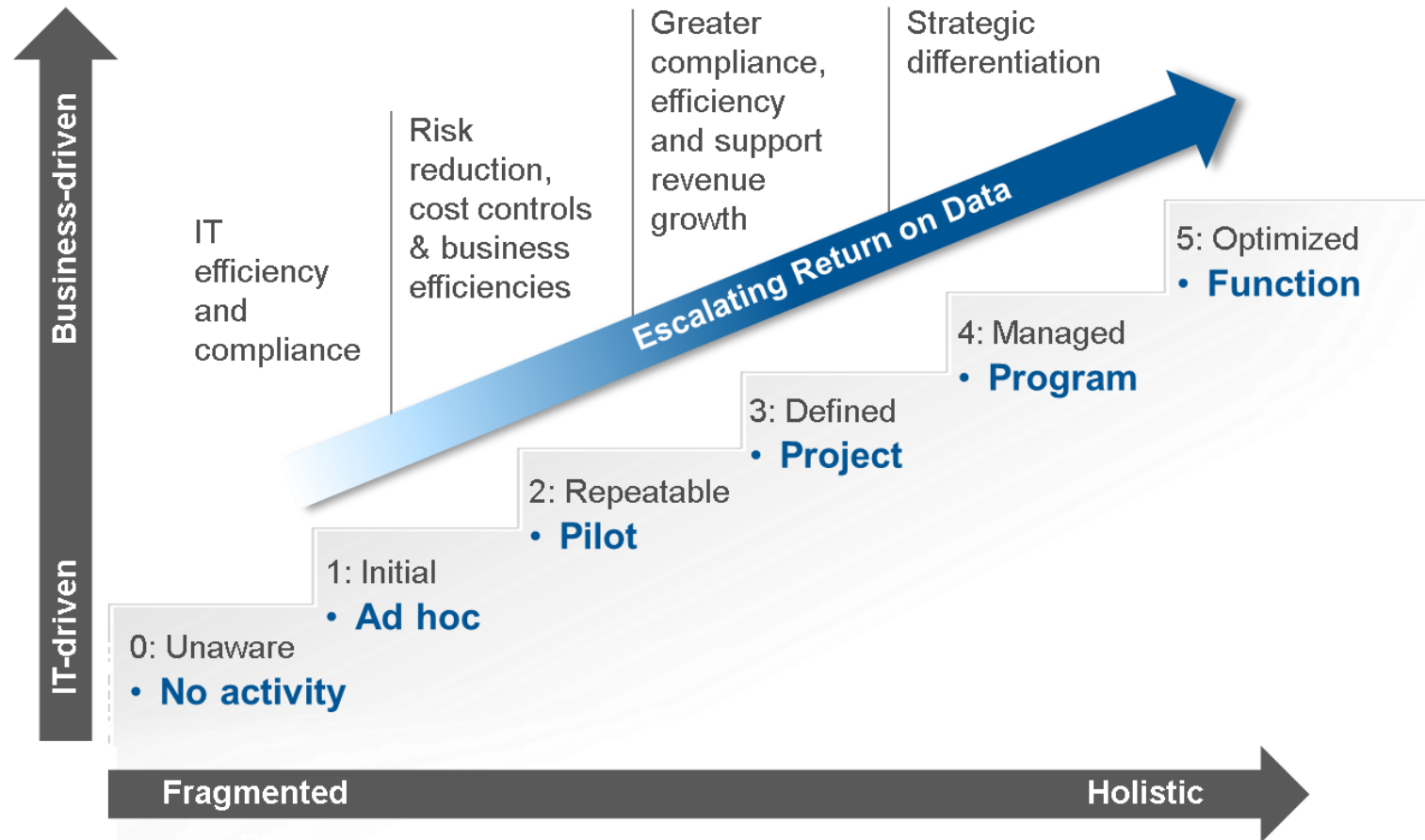
Measure its costs

Devise strategy for improvement

Aim towards data governance maturity



Data Governance Maturity Model (DGM Model)



Summary of Data Privacy

Data privacy protection methods and limitations

- k -anonymity
- l -diversity
- t -closeness
- Differential privacy
 - Deniability by randomization
 - Laplace mechanism

Data Privacy Attacking Methods

- Statistical attacks
- Skewness attacks, Similarity attacks

Next Week

- Lecture: Data Lakes for Big Data Management
- Tutorial: Security, Privacy and Quality
- Practical: Data warehouse