

# Security Information System

## **Chapter 6:** **Anonymization**

# Outline

- Introduction
- K-anonymization
- L-diversity
- T-closeness

# Introduction

- Useful for improving recommendation systems, collaborative research
- Contain personal information
- Mechanisms to protect privacy, e.g. anonymization by removing names
- Yet, private information leaked by attacks on anonymization mechanisms



**m o v i e l e n s**  
helping you find the *right* movies



**amazon.com**



Article [Discussion](#)

---

**AOL search data leak**

---

From Wikipedia, the free encyclopedia

# Introduction

- Huge volume of data is collected
- from a variety of devices and platforms
- Such as Smart Phones, Wearables, Social Networks, Medical systems
- Such data captures human behaviors routines, activities and affiliations
- While this overwhelming data collection
- provides an opportunity to perform data analytics



Data Abuse

## Data Abuse is inevitable:

- It compromises individual's privacy
- Or bridges the security of an institution

# Introduction

- An attacker queries a database for sensitive records
- Targeting of vulnerable or strategic nodes of large networks to
  - Bridge an individual's privacy
  - Spread virus
- Adversary can track
  - Sensitive locations and affiliations
  - Private customer habits
- These attacks pose a threat to privacy

## Database Privacy



## Database Query Outputs



How many people have  
**Hypertension?**

## Network Privacy



## Location & Customer Privacy



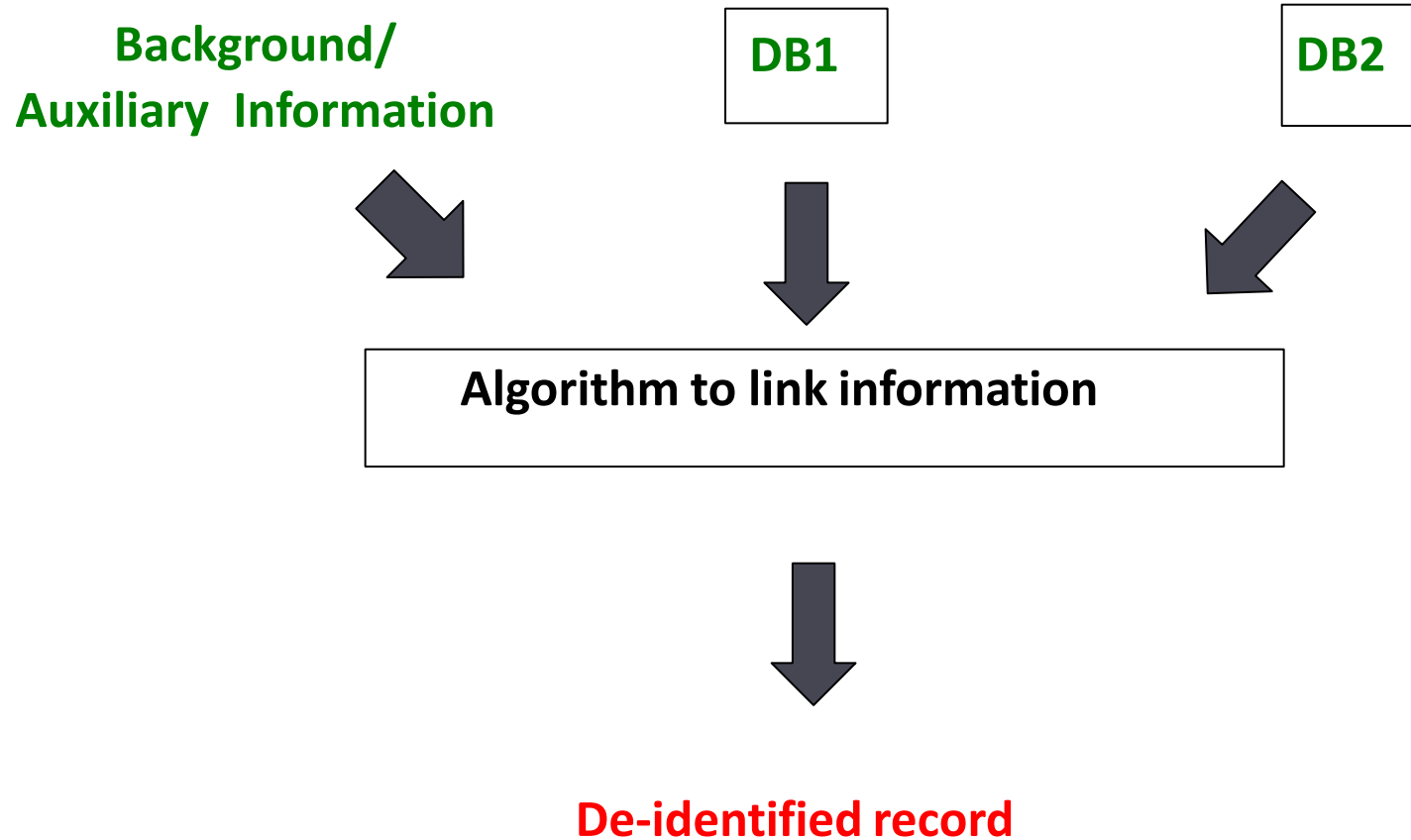
# Introduction

## Privacy Preserving Data Sharing

- It is often necessary to share (publish) data
  - Mandated by laws and regulations
    - E.g., Census
  - For research purposes
    - E.g., social, medical, technological, etc.
  - For security/business decision making
    - E.g., network flow data for Internet-scale alert correlation
  - For system testing before deployment
- Publishing data may result in privacy violations
- **Objective:** Maximize data **utility** while maintaining acceptable level of data **privacy**

# Introduction

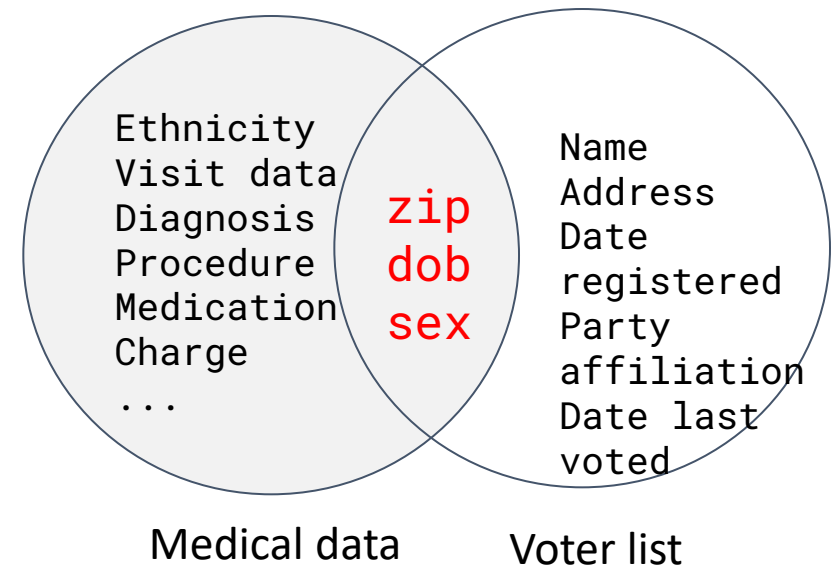
## Non-Interactive Linking



# Introduction

## Data Linkage Attack

- Example: In 1997, a researcher was able to re-identify the information of MA governor by linking released (anonymized) medical data with public voter list.
- Group Insurance Commissions (GIC, Massachusetts)
  - Collected data of ~135k states employees
  - A pseudonymized version was shared with researchers
- Voter data purchased for \$20
- *87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB.*





# Introduction

## Data Linkage Attack

- Fact: 87% of the US citizens can be uniquely linked using only three attributes <Zipcode, DOB, Sex>
- According to MA voter list, only six people share the BoD with state governor
  - 3 men and 3 women
  - One has same Zip code

GIC data

ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	53715	Heart Disease
2	4/13/86	Female	53715	Hepatitis
3	2/28/76	Male	53703	Brochitis
4	1/21/76	Male	53703	Broken Arm
5	4/13/86	Female	53706	Flu
6	2/28/76	Female	53706	Hang Nail

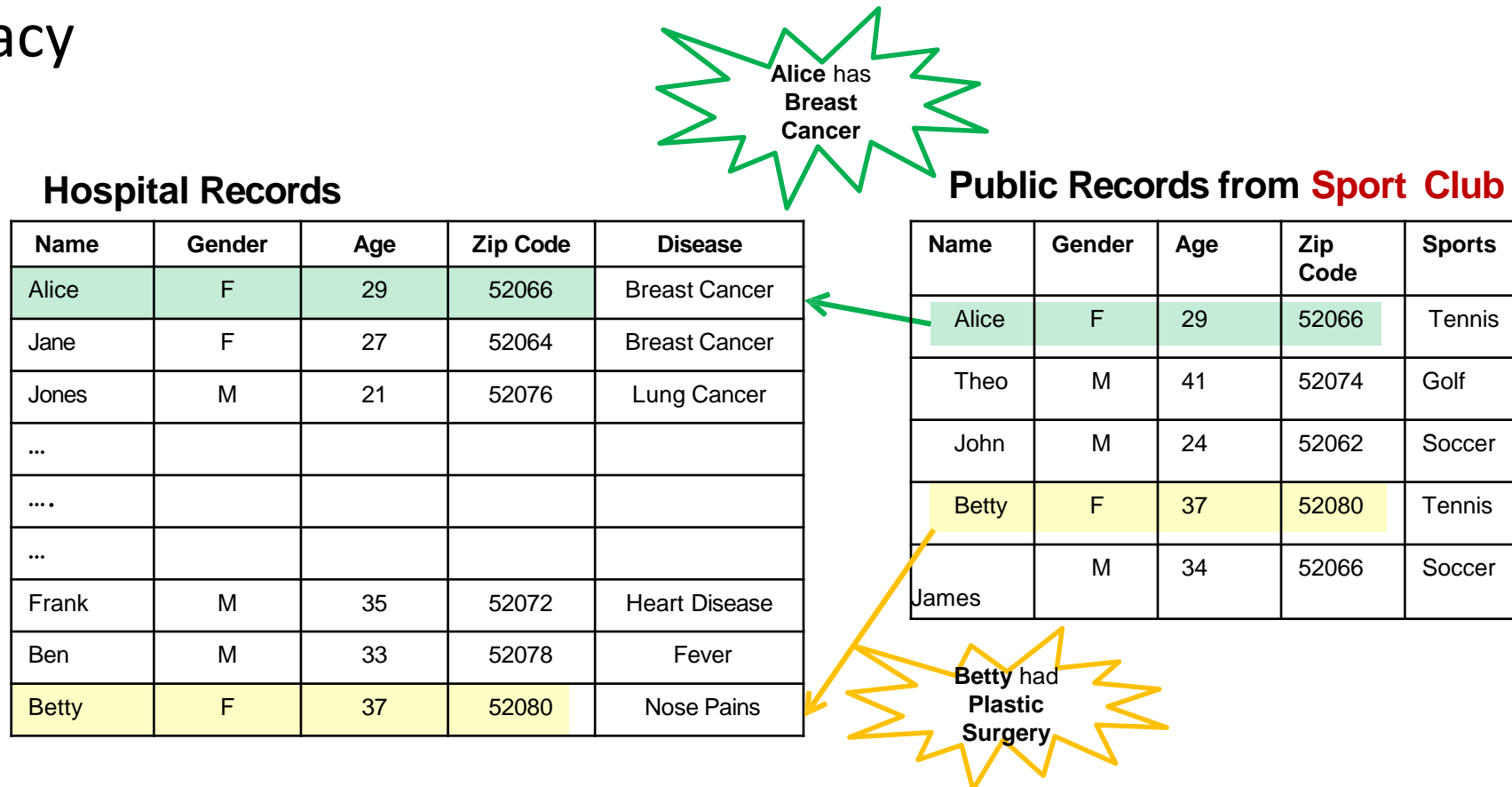
Voter data

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

# Introduction

## Data Linkage Attack

- Linkage Attack: different public records can be linked to it to breach privacy



# Introduction

## Data Linkage Attack

- In 2006, Netflix launched "The Netflix Prize"
- Users names were replaced with random numbers

18k movies

480k  
users

100M  
ratings  
{?, 0,1,...,5}

- The prize was canceled in 2009 after two researchers identified individual users by linking data with IMDB!

# Introduction

## Main Problems

- How to define privacy and utility in data sharing?
- How to anonymize data to satisfy privacy while providing utility?
- How to measure the utility of published data?

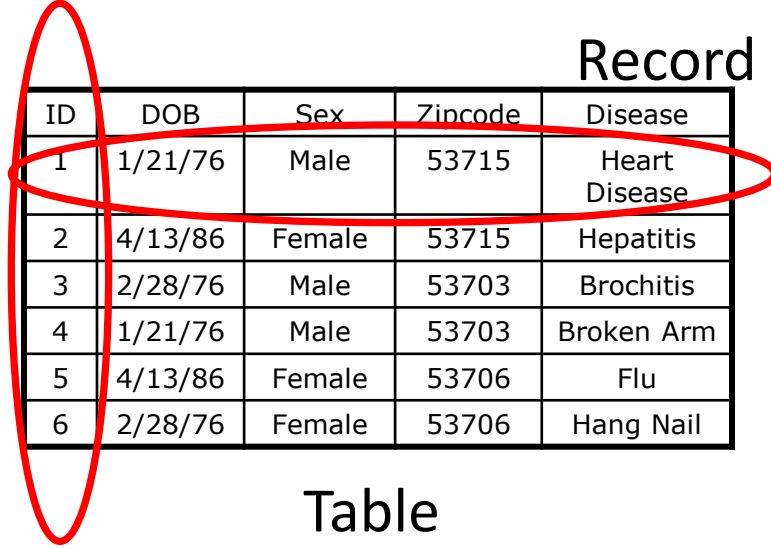
# Introduction

## Terminologies

- **Key Attribute:** uniquely identifies an individual directly
  - Name, Address, Cell Phone
  - Always removed before release!
- **Quasi-Identifier:** A set of attributes that can be potentially linked with external information to re-identify entities
  - ZIP code, Birth date, gender
  - Can be removed, but utility will degrade
- **Sensitive Attribute:** A set of attributes that need to be released to researchers but raises privacy concerns
  - Medical record, wage, etc.

Attribute

Record



ID	DOB	Sex	Zipcode	Disease
1	1/21/76	Male	53715	Heart Disease
2	4/13/86	Female	53715	Hepatitis
3	2/28/76	Male	53703	Brochitis
4	1/21/76	Male	53703	Broken Arm
5	4/13/86	Female	53706	Flu
6	2/28/76	Female	53706	Hang Nail

Table

# Introduction

## Terminologies

Key Attribute		Quasi-identifier			Sensitive attribute
Name		DOB	Gender	Zipcode	Disease
Andre		1/21/76	Male	53715	Heart Disease
Beth		4/13/86	Female	53715	Hepatitis
Carol		2/28/76	Male	53703	Brochitis
Dan		1/21/76	Male	53703	Broken Arm
Ellen		4/13/86	Female	53706	Flu
Eric		2/28/76	Female	53706	Hang Nail

# Introduction

## Privacy Requirements in Data Sharing

- Objective is to preventing the following disclosures

- 1. Membership disclosure:** link a particular individual to a table (E.g. Bob is an engineer -> he is sick)
- 2. Identification disclosure:** link an individual to a particular record (E.g. Alice is 30 year old -> writer)
- 3. Attribute disclosure:** undiscover a new (sensitive) attribute about an individual (E.g. Writer Alice is 30 year old -> flu)

ID	Job	Sex	Age	Disease
1	Engineer	Male	35	Hepatitis
2	Engineer	Male	38	HIV
3	Lawyer	Male	38	Flu
4	Writer	Female	30	Flu
5	Writer	Female	31	Hepatitis
6	Actor	Female	31	Hepatitis
7	Actor	Female	32	HIV

# K-anonymity

- A table is “**k-anonymized**” if each record cannot be identified from other  $k-1$  records when only “quasi-identifiers” are considered
- These  $k$  records form an **Equivalence Class** (recall the anonymity set)
- Alice who is
  - A writer
  - 30 years old
  - How many records can we link him to?

	ID	Job	Sex	Age	Disease
EC1	1	Professional	Male	[35-40)	Hepatitis
	2	Professional	Male	[35-40)	HIV
	3	Professional	Male	[35-40)	Flu
EC2	4	Artist	*	[30-35)	Flu
	5	Artist	*	[30-35)	Hepatitis
	6	Artist	*	[30-35)	Hepatitis
	7	Artist	*	[30-35)	HIV

- k-Anonymity ensures that an individual cannot be linked to a record with probability  $> 1/k$



# K-anonymity

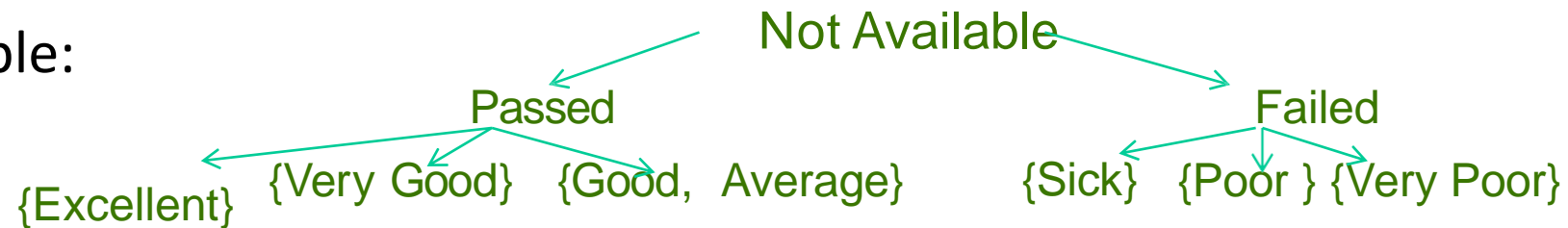
- k-Anonymity ensures privacy by Suppression or Generalization of quasi-identifiers.

- Suppression:

- Accomplished by replacing a part or the entire attribute value by “\*”
- Suppress:           Postal Code : 52057 → 52\*\*\*
- Suppress:           Gender :       i) Male → \*    ii) Female → \*

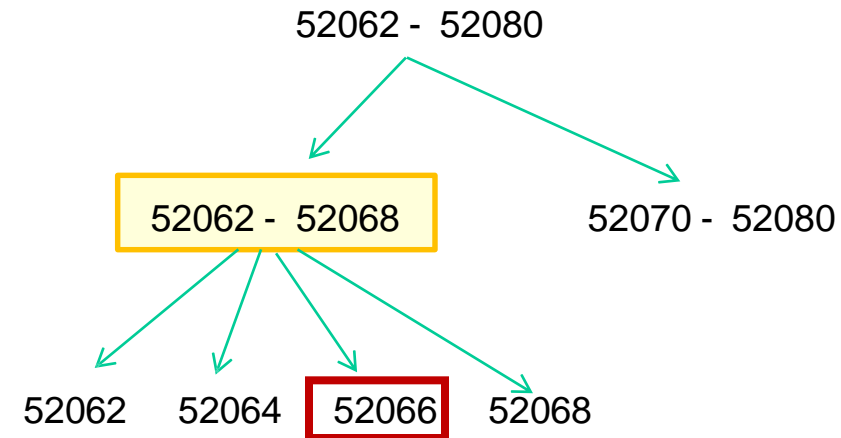
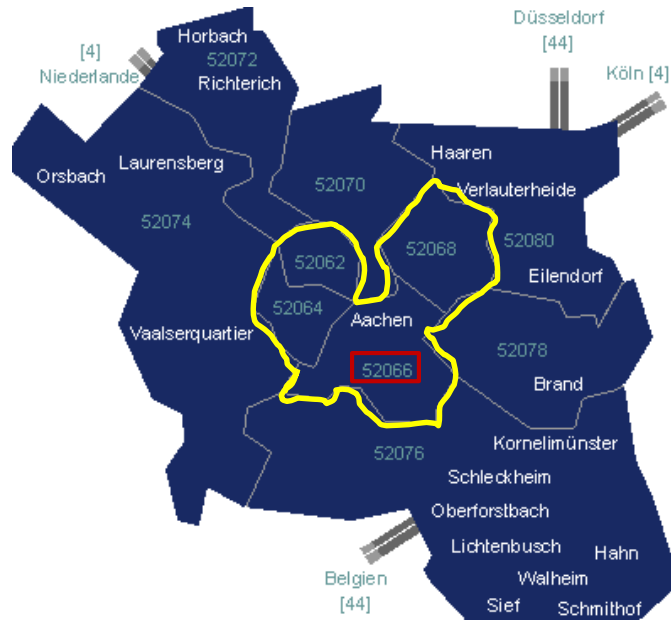
- Generalization:

- Example:



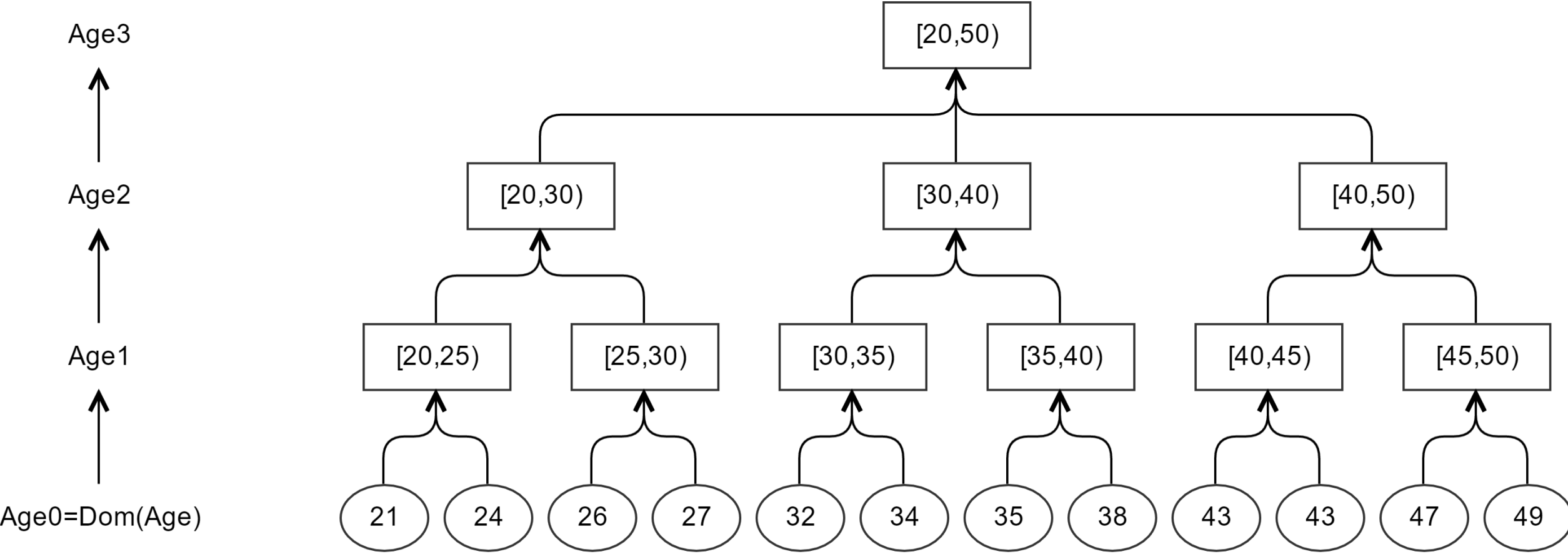
# K-anonymity Generalization

Generalization of Postal Code:

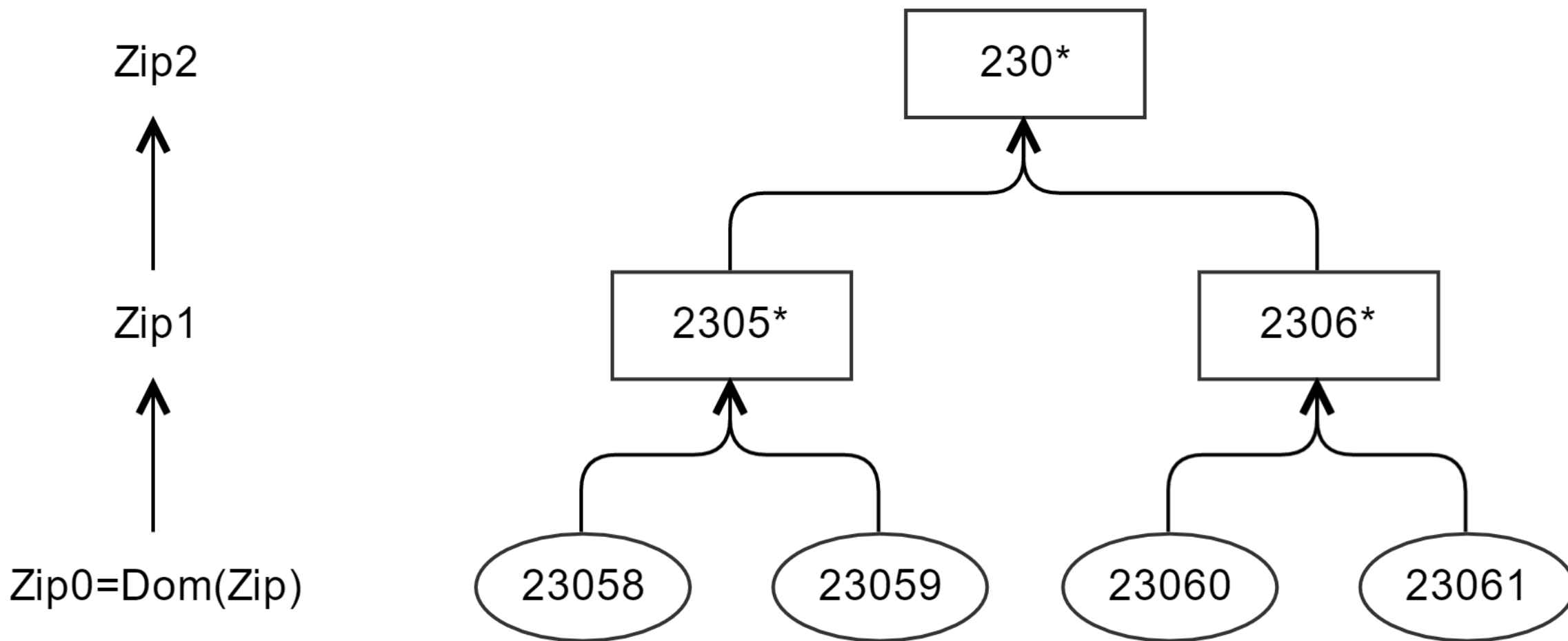


Generalization can be achieved by (Spatial) Clustering

# Example



# Example



# K-anonymity

## Generalization

### **1.Rounding:**

- For continuous numerical values, you can generalize by rounding to a certain number of decimal places or to the nearest integer.

### **2.Binning or Bucketing:**

- Divide the numerical range into bins or buckets and assign values to the corresponding bin. This is often used to create ranges or intervals.

### **3.Scaling:**

- Use scaling to transform numerical values into a different scale. For example, you could normalize values between 0 and 1 or scale them to a specific range.

### **4.Thresholds:**

- Define thresholds to group numerical values into categories based on specific criteria.
- Example: 25, 30, 35, 40, 45, 50, 55, 60, 65, 70

# K-anonymity Suppression

- Example: \$50,000, \$40,000, \$60,000, \$45,000, \$75,000, \$55,000, \$65,000, \$42,000, \$70,000, \$48,000
- Complete Suppression
  - Remove the entire "Income" attribute
- Suppressing Specific Values or Ranges ( $\geq$  \$60,000)
  - **Suppressed Dataset:** \$50,000, \$40,000, \$45,000, \$55,000, \$42,000, \$48,000, \$60,000
- Binning or Bucketing ( $<$ \$50,000, \$50,000-\$60,000,  $>$ \$60,000)
  - **Suppressed Dataset:**  $<$ \$50,000,  $<$ \$50,000,  $>$ \$60,000,  $<$ \$50,000,  $>$ \$60,000,  $<$ \$50,000,  $>$ \$60,000,  $<$ \$50,000,  $>$ \$60,000,  $<$ \$50,000
- Rounding (multiple \$10,000)
  - **Suppressed Dataset:** \$50,000, \$40,000, \$60,000, \$50,000, \$80,000, \$60,000, \$70,000, \$40,000, \$70,000, \$50,000

# K-anonymity Example

	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

Equivalence class

# K-anonymity

## What is the issue with k-anonymity?

- Example

### Released Hospital Records

Key Attribute	Quasi-Identifier			Sensitive Attribute
Name	Gender	Age	Zip Code	Disease
Alice	F	29	52066	Breast Cancer
Jane	F	27	52064	Breast Cancer
Jones	M	21	52076	Lung Cancer
Frank	M	35	52072	Heart Disease
Ben	M	33	52078	Fever
Betty	F	37	52080	Nose Pains

Alice has  
Breast Cancer

Betty had  
Plastic Surgery

### Public Records from Sport Club

Name	Gender	Age	Zip Code
Alice	F	29	52066
Theo	M	41	52074
John	M	24	52062
Betty	F	37	52080
James	M	34	52066



# K-anonymity

## What is the issue with k-anonymity?

- Remove Key Attributes
- Suppress or Generalize Quasi-Identifiers

Key Attribute	Quasi-Identifier			Sensitive Attribute
Name		Age	Zip Code	Disease
Remove	*	2*	520*	Breast Cancer
	*	2*	520*	Breast Cancer
		2*	520*	Lung Cancer
	*	3*	520*	Heart Disease
	*	3*	520*	Fever
	*	3*	520*	Nose Pains

Name	Gender	Age	Zip Code
Alice	F	29	52066
Theo	M	41	52074
John	M	24	52062
Betty	F	37	52080
James	M	34	52066

- This database table is 3-Anonymous
- Oversuppression leads to stronger privacy but poorer Data Utility

# K-anonymity

## What is the issue with k-anonymity?

- **Advantages:** Provides some protection: linking on ZIP, age, nationality yields 4 records
- **Limitations:** lack of diversity in sensitive attributes, background knowledge, subsequent releases on the same data set

	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

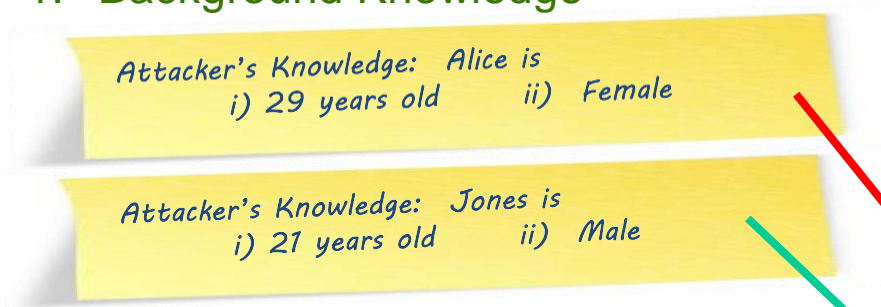
# K-anonymity

## What is the issue with k-anonymity?

Background Knowledge attack

Lack of diversity of the sensitive attribute values (homogeneity)

### 1. Background Knowledge



### 2. Homogeneity

- All Females within 20 years have Breast Cancer. No diversity!!!  
→ Alice has Breast Cancer!
- All 2\*-aged males have lung cancer  
→ Jones has Lung Cancer!

Released Records

Quasi-Identifier			Sensitive Attribute
Gender	Age	Zip Code	Disease
F	2*	520*	Breast Cancer
F	2*	520*	Breast Cancer
M	2*	520*	Lung Cancer
M	2*	520*	Lung Cancer
M	3*	520*	Heart Disease
M	3*	520*	Fever
F	3*	520*	Nose Pains

This led to the creation of a new privacy model called ***l*-diversity**

# L-diversity

- Given a k-anonymized table:
- Ensure that within an equivalence class, there are at least  $l$  “well-represented” values of the sensitive attribute

- $k = 4$
- $l = ?$

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	$\leq 40$	*	Heart Disease
4	1305*	$\leq 40$	*	Viral Infection
9	1305*	$\leq 40$	*	Cancer
10	1305*	$\leq 40$	*	Cancer
5	1485*	$> 40$	*	Cancer
6	1485*	$> 40$	*	Heart Disease
7	1485*	$> 40$	*	Viral Infection
8	1485*	$> 40$	*	Viral Infection
2	1306*	$\leq 40$	*	Heart Disease
3	1306*	$\leq 40$	*	Viral Infection
11	1306*	$\leq 40$	*	Cancer
12	1306*	$\leq 40$	*	Cancer

# L-diversity

- Other variants of l-Diversity
  - Entropy l-Diversity: For each equivalent class, the entropy of the distribution of its sensitive values must be at least  $\log(l)$
  - Probabilistic l-Diversity: The most frequent sensitive value of an equivalent class must be at most  $1/l$
- Limitations of l-Diversity
  - Is not necessary at times
  - Is difficult to achieve: For large record size, many equivalent classes will be needed to satisfy l-Diversity
  - Does not consider the distribution of sensitive attributes

# What is the issue with l-diversity?

- Limitations:
  - Values of the sensitive attribute within one equivalence class may have semantic similarity; can infer some property of the sensitive attribute (i.e., stomach-related disease)
  - Could have high k and low l, resulting in a high occurrence of one value of the sensitive attribute in the equivalence class.

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

# T-closeness

- Given a k-anonymized and l-diverse table:
  - Ensure that the distance between the distribution of each sensitive attribute in the eq. class and the distribution of the attribute value in the whole table is  $\leq t$
  - Salary:  $t = 0.167$
  - Disease:  $t = 0.278$

	ZIP Code	Age	Salary	Disease
1	4767*	$\leq 40$	3K	gastric ulcer
3	4767*	$\leq 40$	5K	stomach cancer
8	4767*	$\leq 40$	9K	pneumonia
4	4790*	$\geq 40$	6K	gastritis
5	4790*	$\geq 40$	11K	flu
6	4790*	$\geq 40$	8K	bronchitis
2	4760*	$\leq 40$	4K	gastritis
7	4760*	$\leq 40$	7K	bronchitis
9	4760*	$\leq 40$	10K	stomach cancer

# Re-identification Attacks in Practice

- Examples:
  - Netflix-IMDB
  - Movielens attack
  - Twitter-Flicker
  - Recommendation systems – Amazon, Hunch,...
- Goal of De-anonymization: To find information about a record in the released dataset



# Background Attack Assumptions

- k-Anonymity, l-Diversity, t-Closeness make assumptions about the adversary
- They at times fall short of their goal to prevent data disclosure
- There is another privacy paradigm which does not rely on background knowledge
- It is called Differential Privacy

# Bài tập 1 – k-anonymity

- Cho tập dataset về thông tin bệnh nhân.
  1. Xác định thuộc tính bán định danh, thuộc tính nhạy cảm.
  2. Xác định số lần lặp lại của các dòng dữ liệu trong tập dữ liệu chỉ có các thuộc tính bán định danh.
  3. Xác định giá trị  $k$  nhỏ nhất để tập dữ liệu thỏa mãn tiêu chí k-anonymity
  4. Với giá trị  $k$  tìm được trong câu 3 tập dữ liệu có đảm bảo tính riêng tư chưa?
  5. Với  $k = 5$  thực hiện tổng quát hóa trên thuộc tính Gender & Medical Condition. Xây dựng tập dữ liệu ẩn danh đạt  $k = 5$ .

Name	Age	Gender	Medical Condition
Alice	25	Female	Diabetes
Bob	30	Male	Hypertension
Charlie	22	Male	Asthma
David	35	Male	Diabetes
Emma	28	Female	Asthma
Frank	40	Male	Hypertension
Grace	45	Female	Diabetes
Henry	32	Male	Hypertension
Irene	27	Female	Asthma
Jack	38	Male	Diabetes
Kate	26	Female	Hypertension
Luke	23	Male	Asthma
Mary	42	Female	Hypertension
Neil	29	Male	Asthma
Olivia	33	Female	Diabetes
Paul	36	Male	Asthma
Quinn	31	Female	Hypertension
Robert	39	Male	Diabetes
Sarah	24	Female	Asthma
Tom	34	Male	Hypertension

# Bài tập 1 – k-anonymity

- Cho tập dataset về thông tin bệnh nhân.
  1. Xây dựng cây ẩn danh cho thuộc tính Age.
  2. Xây dựng cây ẩn danh cho thuộc tính Gender & Age.
  3. Xây dựng cây ẩn danh cho thuộc tính Age & Gender.
  4. Xây dựng cây thuộc tính ẩn danh cho Age, Gender & Medical Condition

Patient ID	Age	Gender	Medical Condition
1	25	Female	Diabetes
2	30	Male	Hypertension
3	22	Male	Asthma
4	35	Male	Diabetes
5	28	Female	Asthma
6	40	Male	Hypertension
7	45	Female	Diabetes
8	32	Male	Hypertension
9	27	Female	Asthma
10	38	Male	Diabetes

# Bài tập 1 – k-anonymity

- Cho tập dataset về nhân sự IT
  1. Phân tích tập dataset, xác định các thuộc tính bán định danh và thuộc tính nhạy cảm trong tập dataset
  2. Chọn và giải thích các tổ hợp 2 thuộc tính để xây dựng cây ẩn danh k-anonymity.

Employee ID	Age	Salary	Department	Years of Experience
1	28	50000	IT	3
2	35	60000	Finance	7
3	22	45000	HR	1
4	40	70000	IT	10
5	28	55000	Marketing	5
6	38	65000	Finance	8
7	45	75000	IT	15
8	32	58000	Marketing	4
9	27	48000	HR	2
10	35	62000	Finance	6

# Bài tập 1 – k-anonymity

Patient ID	Age	Blood Pressure	Cholesterol Level	Diagnosis
1	40	120/80	Normal	Hypertension
2	55	140/90	High	Diabetes
3	30	110/70	Normal	Asthma
4	45	130/85	High	Diabetes
5	35	125/78	Normal	Asthma
6	50	150/95	High	Hypertension
7	60	130/80	Normal	Diabetes
8	32	115/75	Normal	Asthma
9	42	135/88	High	Hypertension
10	48	125/82	Normal	Hypertension

- Cho tập dataset về thuộc tính bệnh nhân
  1. Phân tích tập dataset, xác định các thuộc tính bán định danh và thuộc tính nhạy cảm trong tập dataset
  2. Chọn và giải thích các tổ hợp 2 thuộc tính để xây dựng cây ẩn danh k-anonymity.

# Bài tập 1 – k-anonymity

Student ID	Age	Gender	GPA	Major
1	20	Female	3.5	Computer Sci.
2	22	Male	3.2	Biology
3	21	Female	3.8	Physics
4	23	Male	3.0	History
5	20	Female	3.6	Chemistry
6	22	Male	3.5	Computer Sci.
7	21	Female	3.7	Physics
8	23	Male	3.1	History
9	20	Female	3.9	Chemistry
10	22	Male	3.4	Biology

- Cho tập dataset về sinh viên

1. Xác định những vi phạm rủi ro trong ẩn danh
2. Nêu những rủi ro trong việc tái định trong tập dữ liệu.
3. Lựa chọn các thuộc tính để thực hiện ẩn danh đảm bảo tính riêng tư cho tập dữ liệu

# Bài tập 1 – k-anonymity

Patient ID	Age	Blood Pressure	Cholesterol Level	Diagnosis
1	40	120/80	Normal	Hypertension
2	55	140/90	High	Diabetes
3	30	110/70	Normal	Asthma
4	45	130/85	High	Diabetes
5	35	125/78	Normal	Asthma
6	50	150/95	High	Hypertension
7	60	130/80	Normal	Diabetes
8	32	115/75	Normal	Asthma
9	42	135/88	High	Hypertension
10	48	125/82	Normal	Hypertension

- Cho tập dataset về tình hình sức khỏe bệnh nhân:
  1. Thực hiện tổng quát hóa cho thuộc tính Age, Blood Pressure.
  2. Thực hiện nén (suppression) cho thuộc tính nhạy cảm Diagnosis.

# Bài 2 – L-diversity

- Cho tập dataset về bệnh nhân:
  - Xác định thuộc tính bán định danh, thuộc tính nhạy cảm.
  - Xác định k bằng giá trị bao nhiêu?
  - Thực hiện ẩn danh với k vừa tìm được?
  - Xác định l bằng bao nhiêu đối với thuộc tính nhạy cảm Medical Condition với tổ hợp thuộc tính bán định danh (Age, Gender)
  - Thực hiện ẩn danh hóa để đạt L vừa tìm được

Name	Age	Gender	Medical Condition
Alice	25	Female	Diabetes
Bob	30	Male	Hypertension
Charlie	22	Male	Asthma
David	35	Male	Diabetes
Emma	28	Female	Asthma
Frank	40	Male	Hypertension
Grace	45	Female	Diabetes
Henry	32	Male	Hypertension
Irene	27	Female	Asthma
Jack	38	Male	Diabetes
Kate	26	Female	Hypertension
Luke	23	Male	Asthma
Mary	42	Female	Hypertension
Neil	29	Male	Asthma
Olivia	33	Female	Diabetes
Paul	36	Male	Asthma
Quinn	31	Female	Hypertension
Robert	39	Male	Diabetes
Sarah	24	Female	Asthma
Tom	34	Male	Hypertension



# Bài 3 – tấn công liên kết cho 2 tập dataset ẩn danh

SSN	Name	Race	DateOfBirth	Sex	ZIP	Marital Status	HealthProblem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

[illegible]

# Bài 3 – tấn công liên kết cho 2 tập dataset ẩn danh

- Cho 2 tập dataset, thực hiện tấn công liên kết để biết tình trạng hôn nhân và sức khỏe của Carlson.
- Tập dữ liệu trên mới thực hiện ẩn danh bằng cách xóa thuộc tính định danh. Nhưng các thuộc tính nhạy cảm vẫn còn rủi ro tính riêng tư.
  - Thực hiện ẩn danh với thuộc tính Health Problem là thuộc tính nhạy cảm và thuộc tính bán định danh (Race, DOB, và Sex ). Thực hiện ẩn danh với  $K = 2$ , có bao nhiêu lớp tương đương. Kết quả bảng dữ liệu sau khi ẩn danh với  $K = 2$ .
  - Thực hiện ẩn danh với thuộc tính Health Problem là thuộc tính nhạy cảm và thuộc tính bán định danh (ZIP, Marital Status và Sex ). Thực hiện ẩn danh với  $K = 3$ , có bao nhiêu lớp tương đương. Kết quả bảng dữ liệu sau khi ẩn danh với  $K = 3$ . Xác định  $L$  bằng bao nhiêu để thỏa mãn  $l$ -diversity.

# Bài 04

- Cho tập dataset
  - Xây dựng k-anonymity. Tìm giá trị k lớn nhất
  - Xây dựng bảng dữ liệu thô đạt 3-diverse
  - Xác định bao nhiêu lớp tương đương

	Postal code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	14853	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

# Bài 05

	zip code	age	salary	disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

	zip code	age	salary	disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	>=40	6K	gastritis
5	4790*	>=40	11K	flu
6	4790*	>=40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

# Bài 05

- Cho bảng dữ liệu thô và bảng thỏa 3-diversity. Phân tích bảng 3-diversity
  - Xét có thỏa mãn 3-anonymization
  - Xét có thỏa 3-diversity
  - Tìm các lớp tương đương
  - Vấn đề của bảng 3-diversity là gì? Và cách giải quyết

# Bài 6 – tấn công liên kết

- Cho 2 bảng dữ liệu công khai sau:

id	Name	Address	Zip Code	Sex	Year of birth	PV#	PV Date
1	Paelix	Schmidtweg 4	1321JE	M	1975	1234-01	6-Jun-2001
2	Jans	Wagenstraat 9	1212ZK	F	1960	3453-97	1-May-1997

id	Name	Address	City	Sex	Date of birth	Case#	Crime type	Crime date
3	Paelix	Schmidtweg 4	Almere	M	4-May-1975	2535-01	1	5-6-2001
3	Paelix	Schmidtweg 4	Almere	M	4-May-1975	2535-01	2	6-6-2001
4	Burg	Knuthstraat 48	Tiel	F	6-Oct-1975	2342-01	1	6-6-2001

# Bài 07

- Cho bảng dữ liệu thô. Xác định:
  - Thuộc tính định danh, thuộc tính bán định danh
  - Xây dựng cây tổng quát hóa cho thuộc tính Age cho trường hợp  $k = 4$
  - Xây dựng cây tổng quát hóa cho thuộc tính ZipCode cho trường hợp  $k = 4$
  - Thực hiện ẩn danh cho  $k = 4$ . Sau đó xét l-diversity, đánh giá l bằng bao nhiêu?

SSNumber	Age	ZipCode	Condition
1234-12-1234	21	23058	heart disease
2345-23-2345	24	23059	heart disease
3456-34-3456	26	23060	viral infection
4567-45-4567	27	23061	viral infection
9012-90-9012	32	23058	kidney stone
0123-12-0123	34	23059	kidney stone
4321-43-4321	35	23060	aids
5432-54-5432	38	23061	aids
5678-56-5678	43	23058	kidney stone
6789-67-6789	43	23059	heart disease
7890-78-7890	47	23060	viral infection
8901-89-8901	49	23061	viral infection