Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

# Knowledge Discovery and Data Mining I

Winter Semester 2018/19

# Agenda

# Data Privacy

## Situation

- Huge volume of data is collected
- From a variety of devices and platforms (e.g. Smartphones, Wearables, Social Networks, Medical systems)
- Capturing human behaviors, locations, routines, activities and affiliations
- Providing an opportunity to perform data analytics

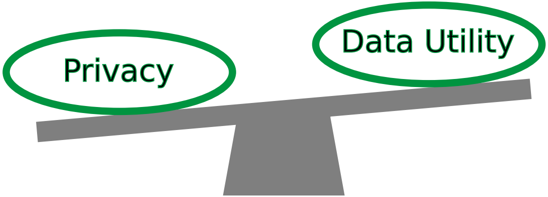
## Data Abuse is inevitable

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

# Data Privacy

- These privacy concerns need to be mitigated
- They have prompted huge research interest to *protect data*
- But,

$$\text{Strong Privacy Protection} \implies \text{Poor Data Utility}$$
$$\text{Good Data Utility} \implies \text{Weak Privacy Protection}$$




**Challenge**

Find a good trade-off between Data Utility and Privacy

# Data Privacy

## Objectives of Privacy Preserving Data Mining

- Ensure data privacy
- Maintain a good trade-off between data utility and privacy

## Paradigms

- $k$-Anonymity
- $l$-Diversity
- Differential Privacy

# Linkage Attack

## Method

Different public records can be linked to it to breach privacy

**Hospital Records**

| *Private* | *Public* | | | |
|---|---|---|---|---|
| Name | Sex | Age | Zip | Disease |
| Alice | F | 29 | 52062 | Breast Cancer |
| Janes | F | 27 | 52064 | Breast Cancer |
| Jones | M | 21 | 52066 | Lung Cancer |
| Frank | M | 35 | 52072 | Heart Disease |
| Ben | M | 33 | 52078 | Fever |
| Betty | F | 37 | 52080 | Nose Pains |

**Public Records from Sport Club**

| *Public* | | | | |
|---|---|---|---|---|
| Name | Sex | Age | Zip | Sport |
| Alice | F | 29 | 52062 | Tennis |
| Theo | M | 41 | 52066 | Golf |
| John | M | 24 | 52062 | Soccer |
| Betty | F | 37 | 52080 | Tennis |
| James | M | 34 | 82066 | Soccer |

# $k$-Anonymity

## $k$-Anonymity

Privacy paradigm for protecting data records before *publication*

Three kinds of attributes:

1. *Key Attributes*: Uniquely identifiable attributes (e.g., Name, Social Security Number, Telephone Number)
2. *Quasi-identifier*: Groups of attributes that can be combined with external data to uniquely re-identify an individual (e.g. (Date of Birth, Zip Code, Gender))
3. *Sensitive Attributes*: An attacker should not be able to combine these with the key attributes. (e.g. Disease, Salary, Habit, Location etc.)

# $k$-Anonymity

**Attention**

Hiding key attributes alone does not guarantee privacy.

An attacker may be able to break the privacy by combining the quasi-identifiers from the data with those from publicly available information.

**Definition: $k$-Anonymity**

Given a set of quasi-identifiers in a database table, the database table is said to be *$k$-Anonymous*, if the sequence of records in each quasi-identifier exists at least $k$ times.

Ensure privacy by *Suppression* or *Generalization* of quasi-identifiers.

# $k$-Anonymity: Suppression

## Suppression

Accomplished by replacing a part or the entire attribute value by placeholder, e.g. "?" (= generalization)

## Example

- Suppress Postal Code: 52062 $\mapsto$ 52???
- Suppress Gender: Male $\mapsto$ ?; Female $\mapsto$ ?

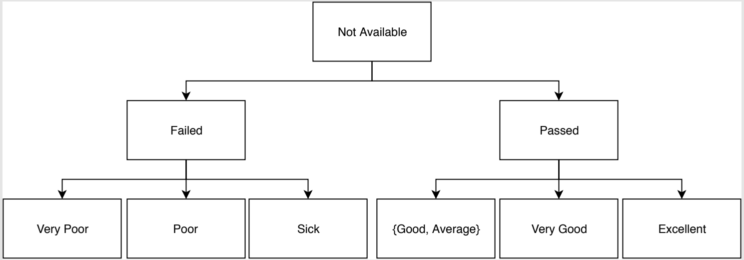
# *k*-Anonymity: Generalization

## Generalization

Accomplished by aggregating values from fine levels to coarser resolution using generalisation hierarchy.

## Example

Generalize exam grades:



# Shortcomings: Background Knowledge Attack

## Background Knowledge Attack

Lack of diversity of the sensitive attribute values (homogeneity)

## Example

- *Background Knowledge*: Alice is (i) 29 years old and (ii) female
- *Homogeneity*: All 2*-aged females have Breast Cancer.
  $\implies$ Alice has BC!

**Release**

| *Quasi Identifier* | | | *Sensitive* |
|---|---|---|---|
| Sex | Age | Zip | Disease |
| F | 2? | 520?? | Breast Cancer |
| F | 2? | 520?? | Breast 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

## Distinct $l$-Diversity

An quasi-identifier is $l$-diverse, if there are at least $l$ different values. A dataset is $l$-diverse, if all QIs are $l$-diverse.

## Example

*Not "diverse"*

| **Quasi Identifier** | **Sensitive** |
|---|---|
| QI 1 | Headache |
| QI 1 | Headache |
| QI 1 | Headache |
| QI 2 | Cancer |
| QI 2 | Cancer |

*2-diverse*

| **Quasi Identifier** | **Sensitive** |
|---|---|
| QI 1 | Headache |
| QI 1 | Cancer |
| QI 1 | Headache |
| QI 2 | Headache |
| QI 2 | Cancer |

# $l$-Diversity

## Other Variants

- *Entropy $l$-Diversity*: For each equivalent class, the entropy of the distribution of its sensitive values must be at least $l$
- *Probabilistic $l$-Diversity*: The most frequent sensitive value of an equivalent class must be at most $1/l$

## Limitations

- Not necessary at times
- 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

# Background Attack Assumption

- *k*-Anonymity and *l*-Diversity 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, called *Differential Privacy*

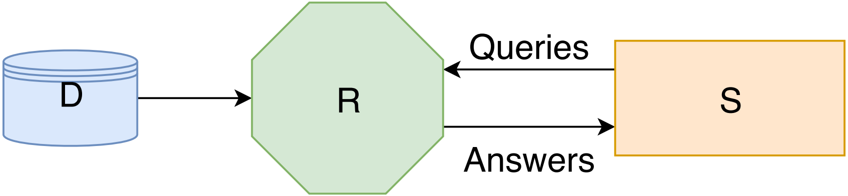
# Differential Privacy

**Core Idea**

Privacy through data perturbation.

- The addition or removal of one record from a database should not reveal any information to an adversary, i.e. your *presence* or *absence* does not reveal or leak any information.
- Use a randomization mechanism to perturb query results of `count`, `sum`, `mean` functions, as well as other statistical query functions.

# Differential Privacy




## Definition

A randomized mechanism $R(x)$ provides *$\epsilon$-differential privacy* if for any two databases $D_1$ and $D_2$ that differ on at most one element, and all outputs $S \subseteq Range(R)$

$$\frac{Pr[R(D_1) \in S]}{Pr[R(D_2) \in S]} \leq \exp(\epsilon)$$

$\epsilon$ is a parameter called *privacy budget/level*.

# Data Perturbation

Data perturbation is achieved by noise addition.

## Some Kinds of Noise

- Laplace noise
- Gaussian noise
- Exponential Mechanism