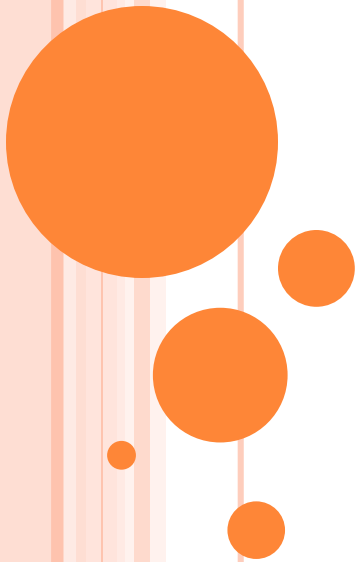


## **MODULE 5:**

# **ADVANCED DATA ANALYTICS FOR HEALTHCARE**



# Introduction

- There are a number of advanced data analytics methods for healthcare.
- These techniques include various data mining and machine learning models that need to be adapted to the healthcare domain.
- The following are the topics that are discussed in this chapter:
  - **Review of Clinical Prediction Models,**
  - **Temporal Data Mining for Healthcare Data**
  - **Visual Analytics for Healthcare Data,**
  - **Information Retrieval for Healthcare**
  - **Data Publishing Methods in Healthcare.**

-



# 1. Review of Clinical Prediction Models

- Clinical prediction forms a critical component of modern-day healthcare.
- Several prediction models have been extensively investigated and have been successfully deployed in clinical practice.
- Such models have made a tremendous impact in terms of diagnosis and treatment of diseases.
- Most successful supervised learning methods that have been employed for clinical prediction tasks fall into three categories:
  - (i) Statistical methods such as linear regression, logistic regression, and Bayesian models;
  - (ii) Sophisticated methods in machine learning and data mining such as decision trees and artificial neural networks; and
  - (iii) Survival models that aim to predict survival outcomes.
- All of these techniques focus on discovering the underlying relationship between covariate variables, which are also known as attributes and features, and a dependent outcome variable.
- The choice of the model to be used for a particular healthcare problem primarily depends on the outcomes to be predicted



- . There are various kinds of prediction models that are proposed in the literature for handling such a diverse variety of outcomes.
- Some of the most common outcomes include binary and continuous forms.
- Other less common forms are categorical and ordinal outcomes.
- In addition, there are also different models proposed to handle survival outcomes where the goal is to predict the time of occurrence of a particular event of interest.
- These survival models are also widely studied in the context of clinical data analysis in terms of predicting the patient's survival time.
- There are different ways of evaluating and validating the performance of these prediction models.
- Different prediction models along with various kinds of evaluation mechanisms in the context of healthcare data analytics required to be studied in detail.
- **(Refer Chapter 10 in Textbook for clinical prediction models)**



## ***2. Temporal Data Mining***


- Healthcare data almost always contain time information and it is inconceivable to reason and mine these data without incorporating the temporal dimension.
- There are two major sources of temporal data generated in the healthcare domain.
- The first is the electronic health records (EHR) data and the second is the sensor data.
- Mining the temporal dimension of EHR data is extremely promising as it may reveal patterns that enable a more precise understanding of disease manifestation, progression and response to therapy.
- Some of the unique characteristics of EHR data (such as of heterogeneous, sparse, high-dimensional, irregular time intervals) makes conventional methods inadequate to handle them.
- Unlike EHR data, sensor data are usually represented as numeric time series that are regularly measured in time at a high frequency. Examples of these data are physiological data obtained by monitoring the patients on a regular basis and other electrical activity recordings such as electrocardiogram (ECG), electroencephalogram (EEG), etc.
- Sensor data for a specific subject are measured over a much shorter period of time (usually several minutes to several days) compared to the longitudinal EHR data (usually collected across the entire lifespan of the patient).



- Given the different natures of EHR data and sensor data, the choice of temporal data mining methods for these types of data are often different.
- appropriate
- EHR data are usually mined using temporal pattern mining methods, which represent data instances (e.g., patients' records) as sequences of discrete events (e.g., diagnosis codes, procedures, etc.) and then try to find and enumerate statistically relevant patterns that are embedded in the data.
- On the other hand, sensor data are often analyzed **using signal processing and time-series analysis techniques** (e.g., **wavelet transform, independent component analysis, etc.**)
- **(Refer Chapter 11 in Textbook for Temporal Data Mining Methods for Healthcare Data )**



### 3. Visual Analytics for Healthcare

- The ability to analyze and identify meaningful patterns in multimodal clinical data must be addressed in order to provide a better understanding of diseases and to identify patterns that could be affecting the clinical workflow.
  - Visual analytics provides a way to combine the strengths of human cognition with interactive interfaces and data analytics that can facilitate the exploration of complex datasets.
  - Visual analytics is a science that involves the integration of interactive visual interfaces with analytical techniques to develop systems that facilitate reasoning over, and interpretation of, complex data.
  - Visual analytics is popular in many aspects of healthcare data analysis because of the wide variety of insights that such an analysis provides. Due to the rapid increase of health-related information, it becomes critical to build effective ways of analyzing large amounts of data by leveraging human–computer interaction and graphical interfaces.
  - In general, providing easily understandable summaries of complex healthcare data is useful for a human in gaining novel insights.
  - In the evaluation of many diseases, clinicians are presented with datasets that often contain hundreds of clinical variables.
  - The multimodal, noisy, heterogeneous, and temporal characteristics of the clinical data pose significant challenges to the users while synthesizing the information and obtaining insights from the data.
- 

- The amount of information being produced by healthcare organizations opens up opportunities to design new interactive interfaces to explore large-scale databases, to validate clinical data and coding techniques, and to increase transparency within different departments, hospitals, and organizations.
- While many of the visual methods can be directly adopted from the data mining literature, a number of methods, which are specific to the healthcare domain, have also been designed.
- **(Refer Chapter 12 in Textbook for visual analytics techniques for Healthcare Data )**





## 4. Information Retrieval for Healthcare

- Although most work in healthcare data analytics focuses on mining and analyzing patient-related data, additional information for use in this process includes scientific data and literature.
- The techniques most commonly used to access this data include those from the field of information retrieval (IR).
- IR is the field concerned with the acquisition, organization, and searching of knowledge-based information, which is usually defined as information derived and organized from observational or experimental research.
- The use of IR systems has become essentially ubiquitous. It is estimated that among individuals who use the Internet in the United States, over 80 percent have used it to search for personal health information and virtually all physicians use the Internet.
- Information retrieval models are closely related to the problems of clinical and biomedical text mining.
- The basic objective of using information retrieval is to find the content that a user wanted based on his requirements.




- This typically begins with the posing of a query to the IR system.
- A search engine matches the query to content items through metadata.
- The two key components of IR are:
  - **Indexing**, which is the process of assigning metadata to the content, and
  - **Retrieval**, which is the process of the user entering the query and retrieving relevant content.
- The most well-known data structure used for efficient information retrieval is the inverted index where each document is associated with an identifier. Each word then points to a list of document identifiers.
- This kind of representation is particularly useful for a keyword search.
- Furthermore, once a search has been conducted, mechanisms are required to rank the possibly large number of results, which might have been retrieved.
- A number of user-oriented evaluations have been performed over the years looking at users of biomedical information and measuring the search performance in clinical settings.
- **(Refer Chapter 14 in Textbook for indexing and retrieval approaches of information retrieval for healthcare data )**



## 5. Data Publishing Methods in Healthcare

### Privacy-Preserving Data Publishing

- In the healthcare domain, the definition of privacy is commonly accepted as “a person’s right and desire to control the disclosure of their personal health information” .
  - Patients’ health-related data is highly sensitive because of the potentially compromising information about individual participants.
  - Various forms of data such as disease information or genomic information may be sensitive for different reasons.
  - To enable research in the field of medicine, it is often important for medical organizations to be able to share their data with statistical experts.
  - Sharing personal health information can bring enormous economical benefits.
  - This naturally leads to concerns about the privacy of individuals being compromised.
  - The data privacy problem is one of the most important challenges in the field of healthcare data analytics.
  - Most privacy preservation methods reduce the representation accuracy of the data so that the identification of sensitive attributes of an individual is compromised.
  - This can be achieved by either perturbing the sensitive attribute, perturbing attributes that serve as identification mechanisms, or a combination of the two.
- 

- Clearly, this process required the reduction in the accuracy of data representation.
- Therefore, privacy preservation almost always incurs the cost of losing some data utility.
- Therefore, the goal of privacy preservation methods is to optimize the trade-off between utility and privacy.
- This ensures that the amount of utility loss at a given level of privacy is as little as possible.
- The major steps in privacy-preserving data publication algorithms are the identification of an appropriate privacy metric and level for a given access setting and data characteristics, application of one or multiple privacy-preserving algorithm(s) to achieve the desired privacy level, and post analyzing the utility of the processed data.
- These three steps are repeated until the desired utility and privacy levels are jointly met.
- Next section focuses on applying privacy-preserving algorithms to healthcare data for secondary-use data publishing and interpretation of the usefulness and implications of the processed data.



# PRIVACY-PRESERVING DATA PUBLISHING

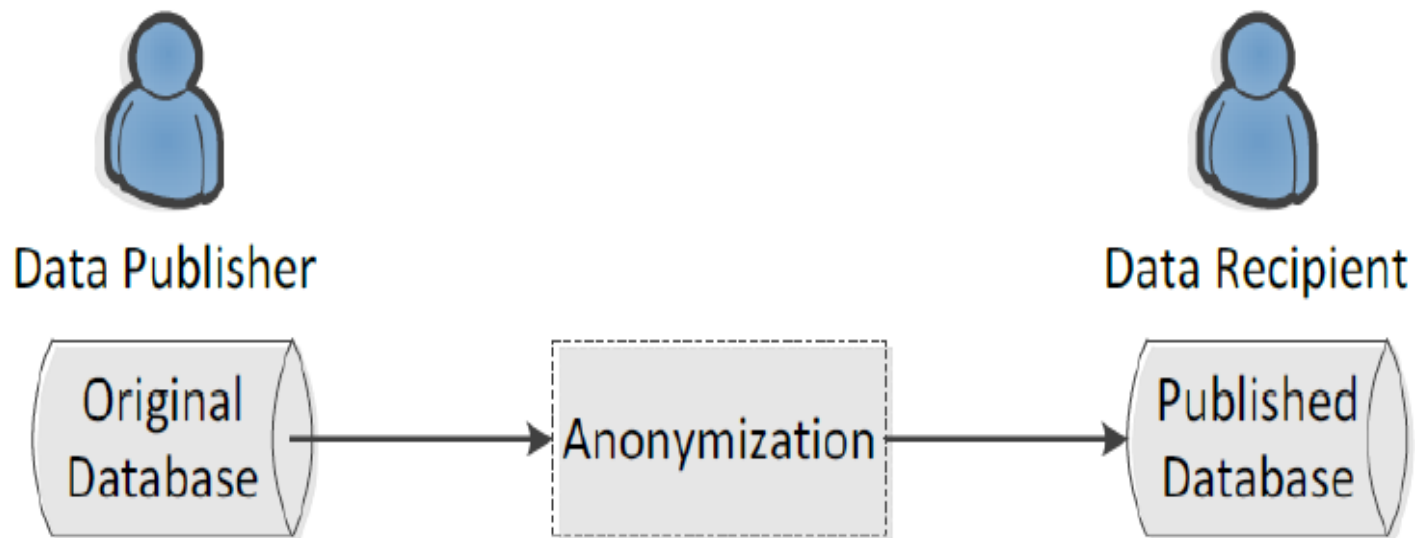
- In the most basic form of privacy-preserving data publishing (PPDP), the data holder has a table of the form:

D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number SSN), containing information that explicitly identifies record owners.

- Quasi Identifier is a set of attributes that could potentially identify record owners, Sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.



# PRIVACY-PRESERVING DATA PUBLISHING



**Figure 1.1 A Simple Model of PPDP**



# DEFINING PRIVACY

- Privacy here means the *logical security* of data
- NOT the traditional security of data e.g. access control, theft, hacking etc.
- Here, adversary uses legitimate methods
- Various databases are published e.g. Census data, Hospital records
  - Allows researchers to effectively study the correlation between various attributes



# NEED FOR PRIVACY

- The data contains:
  - Attribute values which can uniquely identify an individual { zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals { medical condition, salary, location }

	<i>Non-Sensitive Data</i>			<i>Sensitive Data</i>	
<b>#</b>	<b>Zip</b>	<b>Age</b>	<b>Nationality</b>	<b>Name</b>	<b>Condition</b>
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer



# NEED FOR PRIVACY

Published  
Data

	<i>Non-Sensitive Data</i>			<i>Sensitive Data</i>
#	<b>Zip</b>	<b>Age</b>	<b>Nationality</b>	<b>Condition</b>
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

#	<b>Name</b>	<b>Zip</b>	<b>Age</b>	<b>Nationality</b>
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List



# SOURCE OF PROBLEM

- Even if we remove the direct uniquely identifying attributes
  - There are some fields that may still uniquely identify some individual!
  - The attacker can *join* them with other sources and identify individuals

	<i>Non-Sensitive Data</i>			<i>Sensitive Data</i>
<b>#</b>	<b><i>Zip</i></b>	<b><i>Age</i></b>	<b><i>Nationality</i></b>	<b><i>Condition</i></b>
...	...	...	...	...

Quasi-Identifiers



# K-ANONYMITY

- Proposed by Sweeney
- Change data in such a way that for each tuple in the resulting table there are at least  $(k-1)$  other tuples with the same value for the quasi-identifier – **K-anonymized table**

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized



# TECHNIQUES FOR ANONYMIZATION

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but *less* specific value
- Suppression
  - Data not released at all
  - Can be Cell-Level or (more commonly) Tuple-Level



# TECHNIQUES FOR ANONYMIZATION

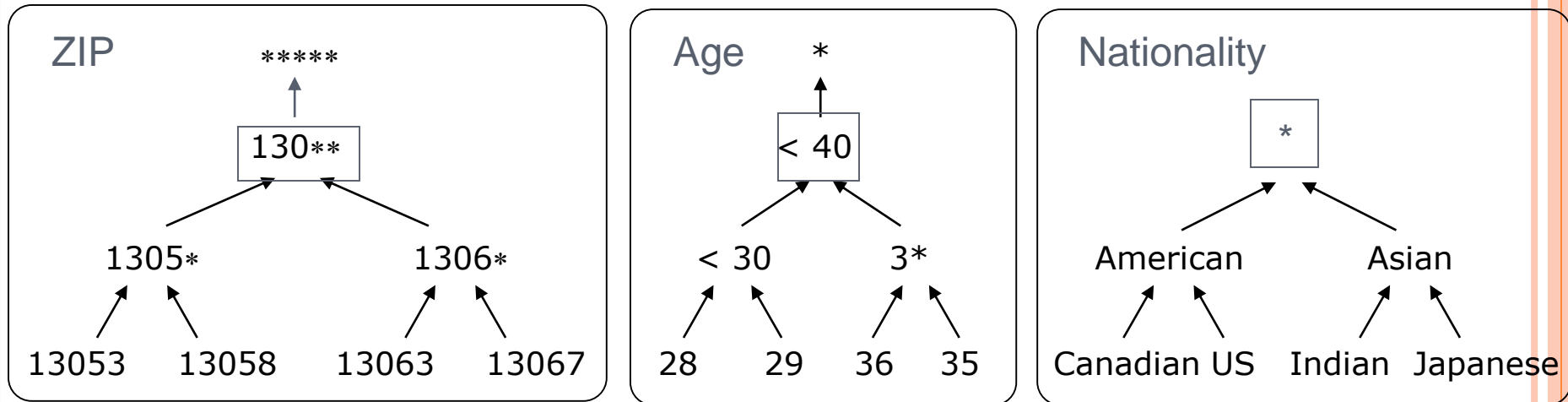
#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Generalization

Suppression (cell-level)



# GENERALIZATION HIERARCHIES



- **Generalization Hierarchies:** Data owner defines how values can be generalized
- **Table Generalization:** A table generalization is created by generalizing all values in a column to a specific level of generalization

# K-MINIMAL GENERALIZATIONS

- There are many k-anonymizations – which *one* to pick?
  - Intuition: The one that does not generalize the data more than needed (decrease in utility of the published dataset!)
- **K-minimal generalization:** A k-anonymized table that is not a generalization of another k-anonymized table



#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13053	< 40	*	Viral Infection
3	13067	< 40	*	Heart Disease
4	13067	< 40	*	Cancer

2-minimal  
Generalizations

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Viral Infection
3	130**	3*	Asian	Heart Disease
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Viral Infection
3	130**	< 40	*	Heart Disease
4	130**	< 40	*	Cancer

NOT a  
2-minimal  
Generalization



## K-MINIMAL GENERALIZATIONS

- Now, there are many k-minimal generalizations! – which one is *preferred* then?
- No clear and “correct” answer. It can be
  - The one that creates min. *distortion* to data, where distortion
  - The one with min. *supression* i.e. which contains a greater number of tuples *and so on*

$$D = \frac{\sum_{\text{attrib } i} \frac{\text{Current level of generalization for attribute } i}{\text{Max level of generalization for attribute } i}}{\text{Number of attributes}}$$



# COMPLEXITY & ALGORITHMS

- If we allow for generalization to a different level for each value of an attribute, the search space is exponential
- More often than not, the problem is NP-Hard!
- Many algorithms have been proposed
  - Incognito
  - Multi-dimensional algorithms (Mondrian)



## K-ANONYMITY DRAWBACKS

- K-anonymity alone *does not* provide full privacy!
- Suppose attacker knows the non-sensitive attributes of
- And the fact that Japanese have very low incidence of heart disease

	<b><i>Zip</i></b>	<b><i>Age</i></b>	<b><i>National</i></b>
Bob →	13053	31	American
Umeko →	13068	21	Japanese



# K-ANONYMITY ATTACK

Original Data →

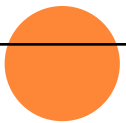
	<i>Non-Sensitive Data</i>			<i>Sensitive Data</i>
<b>#</b>	<b>ZIP</b>	<b>Age</b>	<b>Nationality</b>	<b>Condition</b>
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart 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

## 4-anonymized Table

	<i>Non-Sensitive Data</i>			<i>Sensitive Data</i>
#	<b>ZIP</b>	<b>Age</b>	<b>Nationality</b>	<b>Condition</b>
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	Bob has Cancer!			Cancer
11				Cancer
12	130**	3*	*	Cancer

Umeko  
Matches  
here

Bob  
Matches  
here



# K-ANONYMITY DRAWBACKS

- Basic Reasons for leak –
  - Sensitive attributes lack *diversity* in values
    - Homogeneity Attack
  - Attacker has additional *background knowledge*
    - Background knowledge Attack
- Hence a new solution has been proposed *in-addition* to k-anonymity – *l-diversity*



## Addressing the Problems of Simple Anonymization Techniques

- Provide guarantees that re-identification will not be possible within some bounds
  - Eg: can only map a given individual to a set of 50 individuals

1. k-anonymization
2. l-diversity
3. t-closeness
4. Differential privacy

# ADDRESSING ANONYMIZATION

## PROBLEMS:

### K-ANONYMITY

- A dataset has k-anonymity if at least k individuals share the same identifying values

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

k=2





# ADDRESSING ANONYMIZATION

## PROBLEMS:

### l-DIVERSITY

- A dataset has l-diversity if the individuals that share the same identifying values have at least l distinct values for the sensitive attribute

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

l=1



# ADDRESSING ANONYMIZATION

## PROBLEMS:

### T-CLOSENESS

- A dataset has  $t$ -closeness if the individuals that share the same identifying values have values for the sensitive attribute that are within a threshold  $t$  of diversity
- Threshold is mathematically defined for the data



# DIFFERENTIAL PRIVACY

- Only method that provides mathematical guarantees of anonymity
- Main problem addressed: Taking an individual  $I$  off a dataset reveals their sensitive attribute information
  - Eg: retrieving aggregate data before removal, then retrieving aggregate data after removal, and then comparing the difference will give us the sensitive attribute of  $I$
- Main idea: Differential privacy adds “noise” to the retrieval process so that such comparisons do not give us the actual sensitive attribute information
  - “noise” is mathematically defined for the data



# DIFFERENTIAL PRIVACY

- It adds random noise to each query result
- The parties can inject noise from a Laplace or Gaussian distribution
- Producing a result that's not quite exact but that masks the contents of any given row



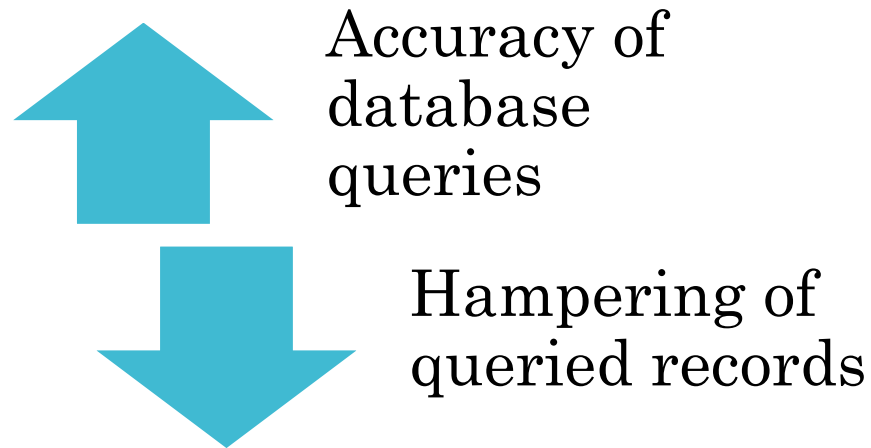
# DIFFERENTIAL PRIVACY

- $\Pr [A (D1) \in S] \leq e^{\epsilon} \times \Pr [A (D2) \in S]$
- where
  - $\epsilon$  is a positive real number
  - $A$  is a randomized algorithm that takes a dataset as input
  - $imA$  denote the image of  $A$
  - $A$  is  $\epsilon$ -differentially private if for all datasets  $D1$  and  $D2$  that differ on a single element, and all subsets  $S$  of  $imA$
  - probability is taken over the randomness used by the algorithm



# DIFFERENTIAL PRIVACY

- It is an emerging data transformation technique and gain a good amount of attention from several years



# ANONYMIZATION

- Anonymization means masking. That is identifying information is removed from the original data to protect personal or private information.



# DATA ANONYMIZATION

- Data Anonymization is a technology that convert clear text into a non-human readable form.
- Data anonymization is the process of de-identifying sensitive data while preserving its format and data type.
- Data anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization.





# ANONYMIZATION OPERATION

- Several anonymization techniques, like generalization and bucketization, have been intended for privacy preserving microdata publishing.
  - Suppression
  - Generalization
  - *Bucketization*
  - *Perturbation*
  - *Slicing*
- *Some of them are discussed in next section:*



# Table 1: Original Table

Name	Gender	Age	Zip code	Disease
Reena	F	29	47677	Cancer
Madhuri	F	22	47602	Cancer
Kapil	M	27	47678	HIV
Ajay	M	43	47905	Flu
Aarti	F	52	47909	Heart problem
Jay	M	45	47906	Heart problem



# SUPPRESSION

- Suppression means removing an entire tuple or attribute value. Replaces tuple or attribute values with special symbol „\*“ that means any value can be there.
- This suppression technique is used in the quasi identifier fields to preserve the individual data .
- **Advantage :**
  - It convert a normal table into anonymous table.
- **Disadvantage :**
  - Quality of the data drastically reduces.



Table 1 :  
Original Table

Name	Gender	Zip code	Age	Disease
Reena	F	47677	29	Cancer
Madhuri	F	47675	22	Cancer
Kapil	M	47678	27	HIV
Ajay	M	47905	43	Flu
Aarti	F	47909	52	Heart problem
Jay	M	47906	45	Heart problem

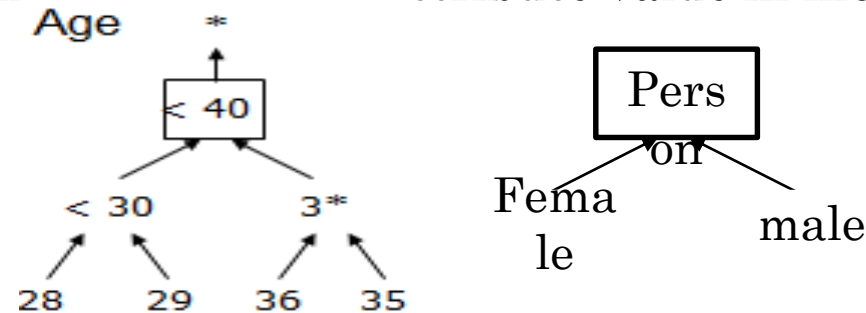
Table 2 : A published table  
by Suppression

Gender	Zip code	Age	Disease
**	<del>47677</del>	**	<del>Cancer</del>
**	<del>47675</del>	**	<del>Cancer</del>
**	<del>47678</del>	**	<del>HIV</del>
**	<del>47905</del>	**	<del>Flu</del>
**	<del>47909</del>	**	<del>Heart problem</del>
**	<del>47906</del>	**	<del>Heart problem</del>



# GENERALIZATION

- Generalization is an operation that changes a value to a more generalized one. If the value is numeric, this value may be changed to a range of values.
- For example, value 26 can be changed to range 26-30. If the value is a categorical value, it may be changed to another categorical value denoting a broader concept of the original categorical value. For instance, city San Francisco can be changed to state California.
- It uses taxonomy to replace attribute value in more



# GENERALIZATION

- **Advantage of generalization:**
  - Sensitive values are replaced with a general none revealing value.
- **Disadvantage of generalization:**
  - For high-dimensional data, generalization loses significant amount of information.
  - Generalization affected from the curse of dimensionality.
  - Generalized data reduces the data utility
  - Correlations between different attributes are lost.



Table 1 :  
Original Table

Name	Gender	Age	Zip code	Disease
Reena	F	29	47677	Cancer
Madhuri	F	22	47675	Cancer
Kapil	M	27	47678	HIV
Ajay	M	43	47905	Flu
Aarti	F	52	47909	Heart problem
Jay	M	45	47906	Heart problem

Table 3 : A published table  
by generalization

Gender	Age	Zip code	Disease
Person	<30	47675- 47678	Cancer
Person	<30	47675- 47678	Cancer
Person	<30	47675- 47678	HIV
Person	43 – 52	47905= 47909	Flu
Person	43 – 52	47905= 47909	Heart problem
Person	43 – 52	47905= 47909	Heart problem



# PERTURBATION

- The idea behind this operation is to de-associate the relationship between quasi identifier and sensitive attribute by partitioning a set of data record into groups and shuffling their sensitive value between each group.
- Under perturbation a value can be changed to any arbitrary value. For example, Male can be changed to Female and vice versa.
- **Disadvantage:**
  - Reduces data utility.





# PERTURBATION


- The following are three commonly used perturbation methods, namely adding noise, value swapping and model-fitting-and-regenerating.
  - *Adding noise* is used for hiding sensitive numerical data. It replace the original sensitive value  $s$  with  $s+r$  is random value drawn from some distribution.
  - *Value swapping* exchange values of sensitive attribute with any individual record.
  - *Model-fitting-and regenerating* defines the model then estimate the parameter and generate the another set of data which follow the defined model.
- 

Table 1 : Original Table

Name	Gender	Age	Zip code	Disease
Reena	F	29	47677	Cancer
Madhuri	F	22	47675	Cancer
Kapil	M	27	47678	HIV
Ajay	M	43	47905	Flu
Aarti	F	52	47909	Heart problem
Jay	M	45	47906	Heart problem

Table 5: A published data by perturbation

Gender	Age	Zip code	Disease
M	29	47677	Cancer
M	22	47602	Cancer
F	27	47678	HIV
F	43	47905	Flu
M	52	47909	Heart problem
F	45	47906	Heart problem

