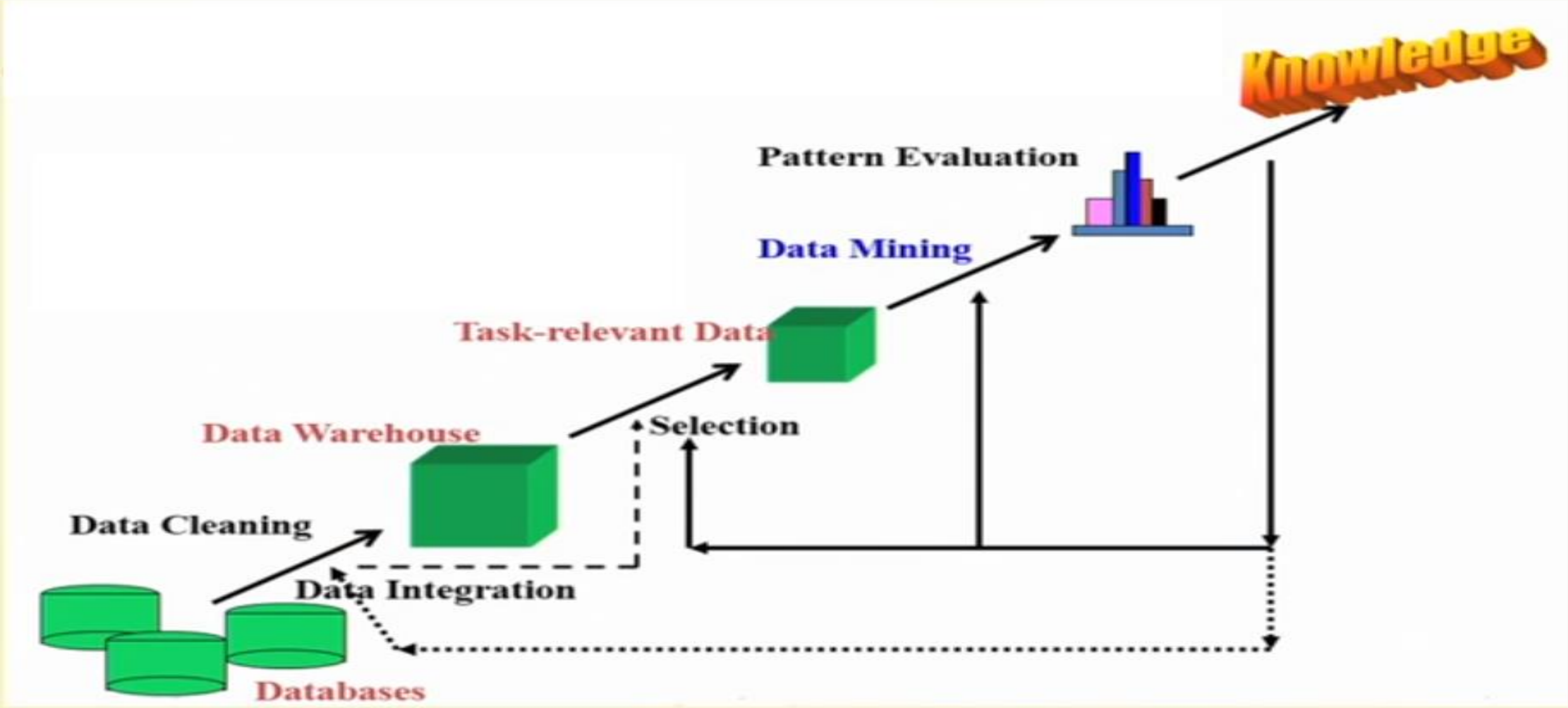
Data Pre-processing

By: Dr. Abhishek Verma

Email-id: abhishek.verma@gla.ac.in

Knowledge Discovery in Data: Process



What is Data?

- a collection of number assigned as value to quantitative variable and/ or characters assigned as value to qualitative variables, or
- collection of records and their attributes
- An attribute is a characteristic of an object
  - Example: Colours of yes, temperature, etc.
  - Attribute is also known as variable, feature, characteristics, fields, etc.
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity or instances

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Attributes

- Nominal
  - Used to assign individual cases to categories
  - Example: eye colour, ID number, Zip code, etc
- Ordinal
  - Used to rank order cases
  - Example: ranking (eg. movie on scale of 1-10), height (tall, medium, short), grades
- Interval
  - Example: Calendar dates, longitude, latitude
- Ratio
  - Same as interval variable but they have a “true zero”
  - Example: time, length, population, age

Properties of Attribute values

- The type of an attributes depends on which of the following properties it possess:
  - Distinctness: $= \neq$
  - Order: $< >$
  - Addition: $+ -$
  - Multiplication: $* /$

- Nominal: Distinctness
- Ordinal: Distinctness, Order
- Interval: Distinctness, Order, Addition
- Ratio: all 4 properties

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countable infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: Binary attributes are special cases of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variable

Type of data sets

- Record Data
  - Data Matrix
  - Transaction data
- Graph Data
  - World wide web
  - Molecular structure
- Ordered
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic sequence data

Record Data

- Data that consists of a collection of records, each of which consists of fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multidimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Data Matrix Example for Documents

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

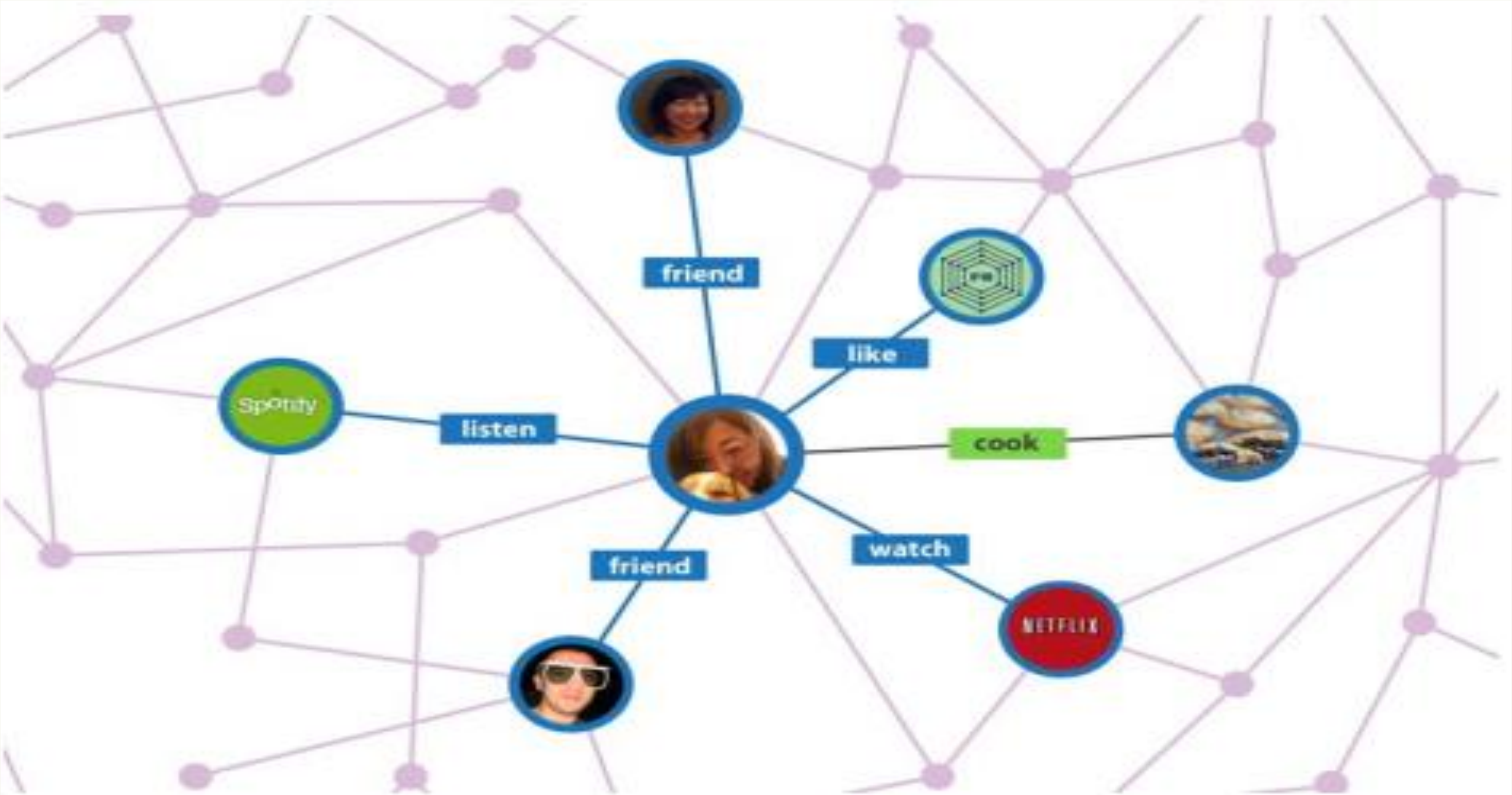
- A typical type of record data, then
  - Each record (transaction) involves a set of items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Market-Basket Dataset

Graph data

Example: Facebook graph and HTML links



Ordered data

- Genetic sequence data

Species	Alignment of Amino Acid Sequences of β-globin
Human	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Monkey	1 VHLTPEEKNA VTTLWGKVNV DEVGGEALGR LLLVYPWTQR FFESFGDLSS
Gibbon	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Human	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
Monkey	51 PDAVMGNPKV KAHGKKVLGA FSDGLNHLDN LKGTFAQLSE LHCDKLHVDP
Gibbon	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFAQLSE LHCDKLHVDP
Human	101 ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH
Monkey	101 ENFKLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH
Gibbon	101 ENFRLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH

Data Quality

- What kind of data quality problems?
- How can we detect the problem with the data?
- What can we do about these problem?
- Examples of data quality problems:
  - Missing values
  - Noise and outliers
  - Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

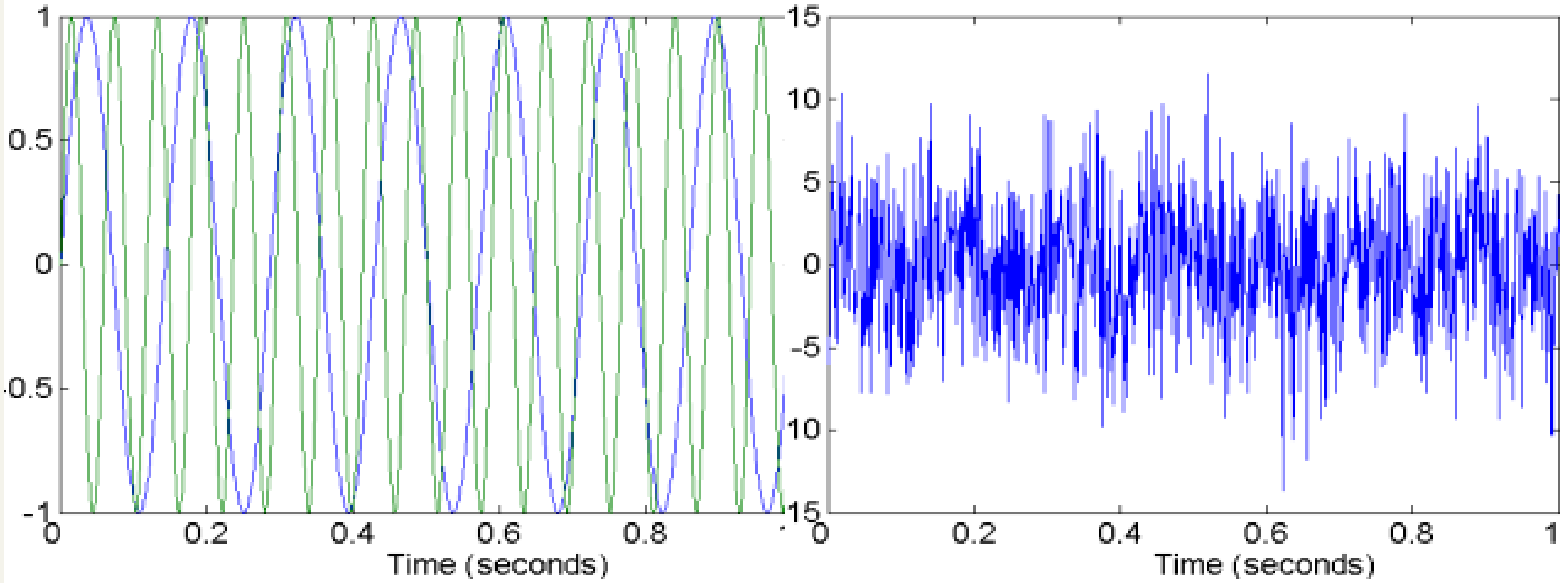
Data Quality: Missing Values

- Reasons for missing values
- Information is not collected
  (e.g., people decline to give their age and weight)
- Attributes may not be applicable to all cases
  (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

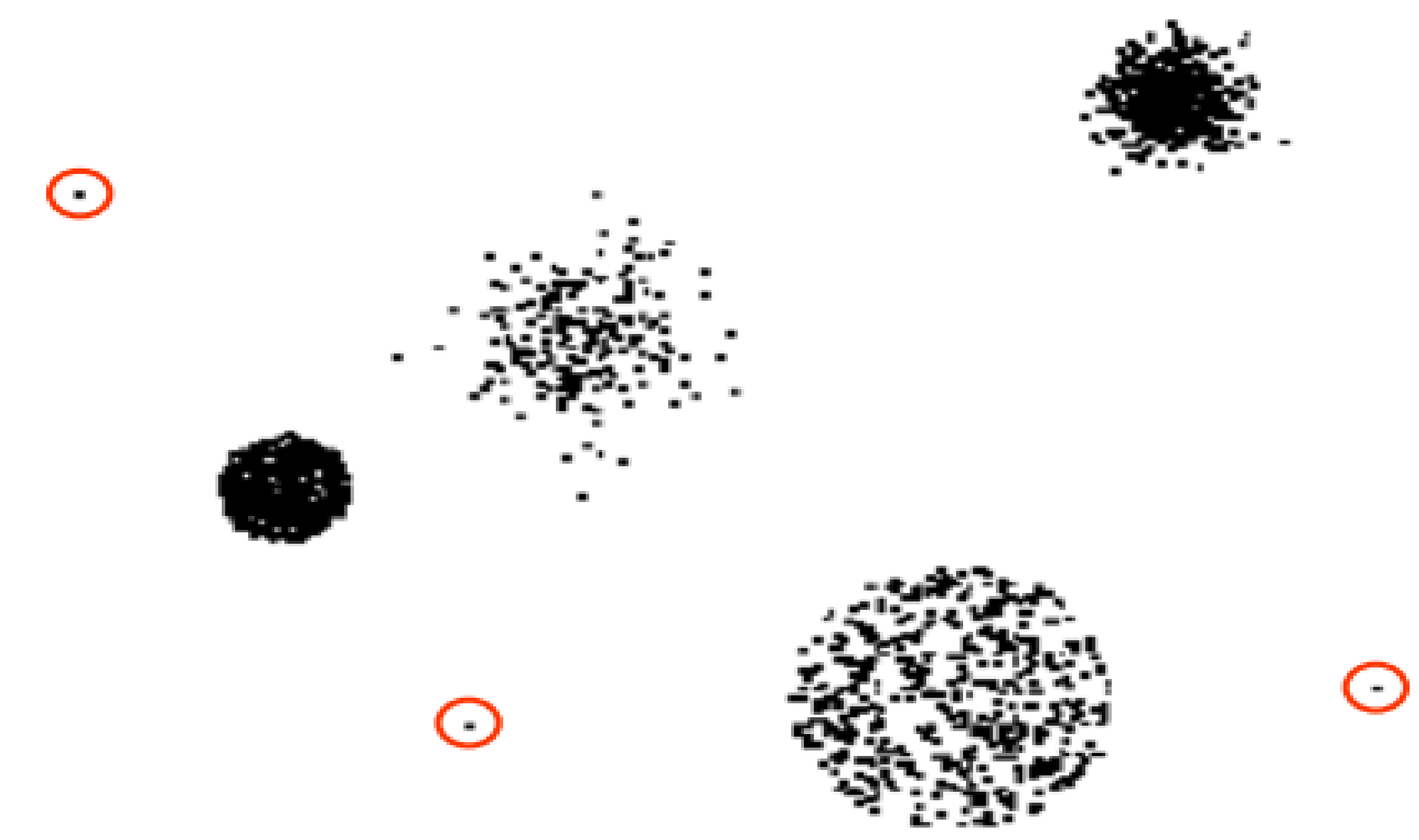
Data Quality: Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on



Data Quality: Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Data Quality: Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

- Imputation
- Outlier management
- One hot encoding
- Feature selection
- Filter and Wrapper approach

Imputation (filling in) of missing data

- Imputation is performed using a number of different algorithms, which can be subdivided into single and multiple imputation methods.

- Single imputation methods
  - a missing value is imputed by a single value

- Multiple-imputation methods
  - several likelihood- ordered choices for imputing the missing value are computed and one “best” value is selected.

Imputation Contd…

- **Single imputation**
  - Mean imputation
  - Hot deck imputation

- **Multiple imputation**

Single imputation Contd…

Mean imputation

- Mean imputation, also called unconditional mean imputation, is a widely used imputation method
- Mean imputation assumes that the mean of a variable is the best estimate for any case that has missing information on this variable
- For **continuous variable**, each missing value is imputed with the mean of known values for the same variable
- For **categorical variable**, the missing values of are the mode of the observed values of same variable

Case	Var1	Var2	Var3
1	9	8	8
2	7.44	7	6
3	8	5	6
4	7	4	5
5	9	5	7
6	8	8	9
7	6	7	6
8	5	9	7
9	7	8	?
10	8	8	7

- *Advantages*
  - fast,
  - simple,
  - ease to implement, and
  - no cases are excluded
- *Limitations*
  - underestimation of the population variance
  - thus a small standard error
  - possibility of Type I error.

Single imputation Contd…

Hot deck imputation

- Hot-deck imputation is a procedure where the imputed values come from other cases In the same data set
- for each object that contains missing values, the most similar object is found, and the missing values are imputed from that object

Case	Var1	Sex	Var2	Var3
1	9	F	8	8
2	8.25	F	7	6
3	8	F	5	6
4	7	F	4	5
5	9	F	5	7
6	8	M	8	9
7	6	M	7	6
8	5	M	9	7
9	7	M	8	?
10	8	M	8	7

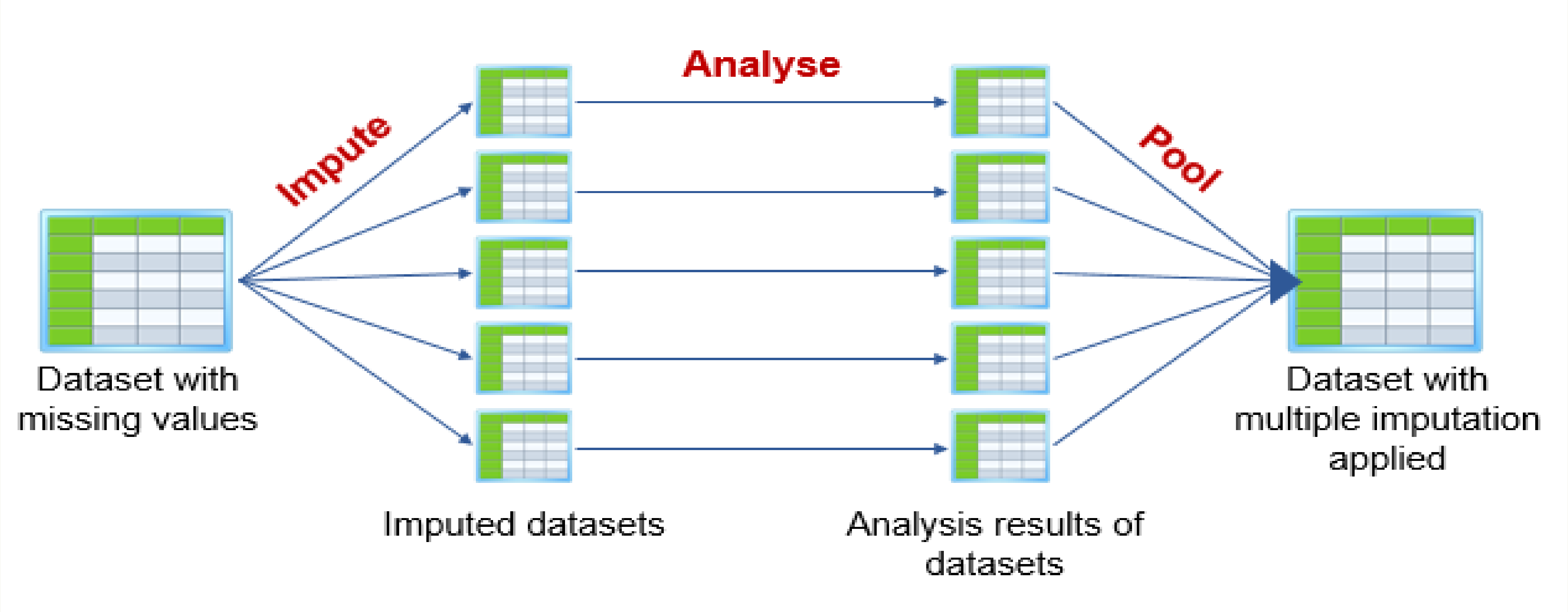
- *Advantages*
  - preserves the population distribution
  - it is better than mean imputation
- *Limitations*
  - distort correlations and covariances

Other type of Single imputation

- Regression imputation
- Cold-deck imputation
- Expectation Maximisation (EM)
- Sequential imputation
- Last observation carried forward
- Worst case and Best case imputation

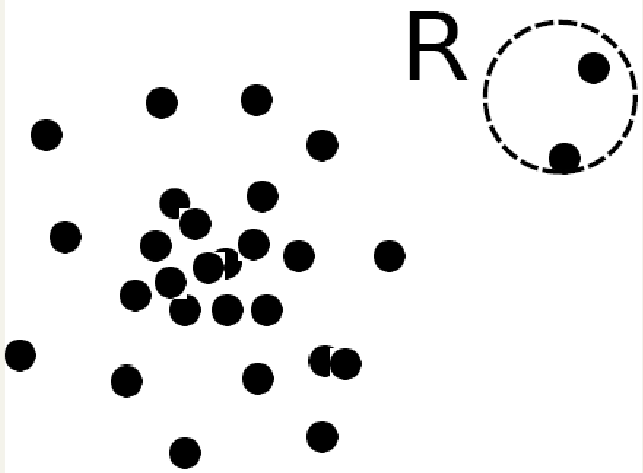
Multiple imputation

- The idea of Multiple Imputation is to replace each missing value with multiple acceptable values that represent a distribution of possibilities.
- This results in a number of complete datasets (usually 3-10):



Outlier management

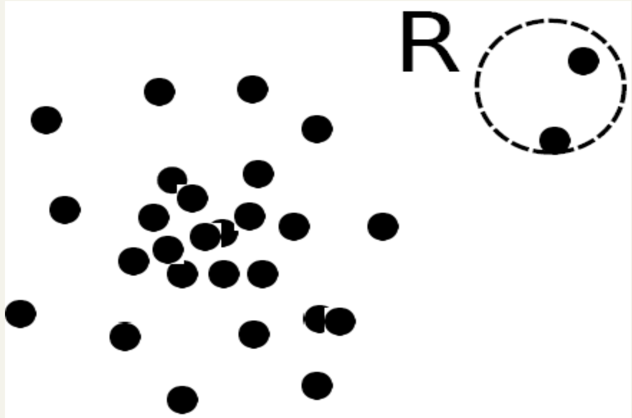
- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection



Types of Outliers

- Three kinds:
  - *Global,*
  - *Contextual*
  - *Collective*

- **Global outlier** (or point anomaly)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation

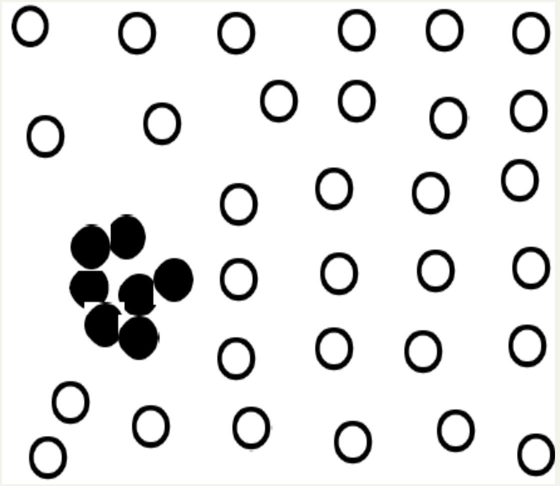


Types of Outliers Contd…

- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a selected context
  - Ex. 40^o C in Mathura: outlier? (depending on summer or winter?)
  - Attributes of data objects should be divided into two groups to detect O_c
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature, pressure, humidity
  - Issue: How to define or formulate meaningful context?

Types of Outliers Contd…

- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages to each other
  - Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.



Outlier Detection

- Two ways to categorize outlier detection methods:
  - Based on whether user-*labeled* examples of outliers can be obtained:
    - Supervised,
    - Unsupervised, and
    - Semi-supervised methods
  - Based on *assumptions about normal data and outliers*:
    - Statistical,
    - proximity-based, and
    - clustering-based methods

Supervised Methods

- Modeling outlier detection as a classification problem
- Methods for Learning a classifier for outlier detection effectively:
  - Model normal objects & report those not matching the model as outliers, or
  - Model outliers and treat those not matching the model as normal
- Challenges
  - Imbalanced classes, i.e., outliers are rare

Unsupervised Methods

- Assume the normal objects are somewhat ``clustered'‘ into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area

Semi-Supervised Methods

- In many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting the model of normal objects are detected as outliers

Statistical Methods

- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model
  - The data not following the model are outliers.
- Methods are divided into two categories: *parametric* vs. *non-parametric*
- **Parametric method**
  - Assumes that the normal data is generated by a parametric distribution with parameter θ
  - The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution
- **Non-parametric method**
  - Not assume an a-priori statistical model and determine the model from the input data
  - Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance
  - Examples: histogram

Parametric Methods I: Univariate Outliers Based on Normal Distribution

- Often assume that data are generated from a normal distribution, learn the parameters from the input data, and identify the points with low probability as outliers

- Ex: Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
  - Use the maximum likelihood method to estimate μ and σ

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - For the above data with n = 10, we have $\hat{\mu} = 28.61$ $\hat{\sigma} = \sqrt{2.29} = 1.51$

  $\mu \pm 3\sigma$ region contains 99.7% data

  - Consider the value 24
  - 28.61 – 3*1.51 = 24.08
  - So, 24 is an outlier

Statistical Methods – Box Plot

- Values less than Q1-1.5*IQR and greater than Q3+1.5*IQR are outliers
- Consider the following dataset:
- 10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

Here,

Q2(median) = 14.6

Q1 = 14.4

Q3 = 14.9

IQR = Q3 – Q1 = 14.9 - 14.4 = 0.5

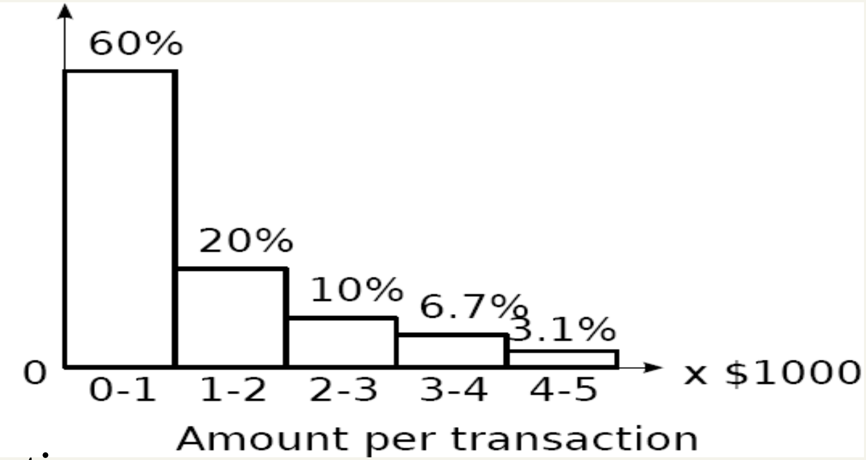
Outliers will be any points:

below Q1 – 1.5×IQR = 14.4 – 0.75 = 13.65 or

above Q3 + 1.5×IQR = 14.9 + 0.75 = 15.65

So, the outliers are at 10.2, 15.9, and 16.4.

Non-Parametric Methods: Detection Using Histogram



- The model of normal data is learned from the input data without any *a priori* structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram:
  - Figure shows the histogram of purchase amounts in transactions
  - A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000
- Problem: Hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, false positive
  - Too big bin size → outliers in some frequent bins, false negative

Statistical Methods – Other Methods

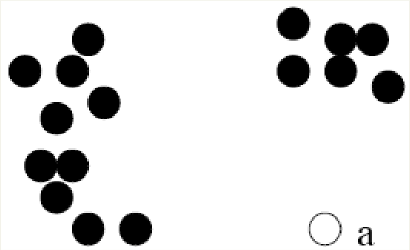
- Grubbs test
- Mahalanobis distance
- Chi Square test

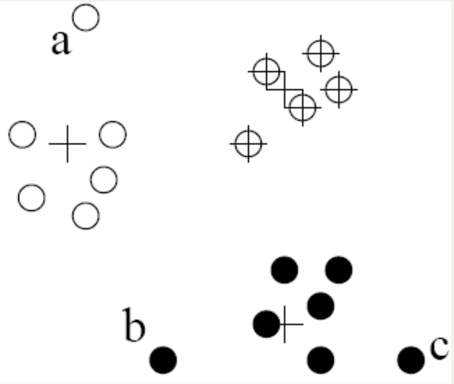
Proximity-Based Methods

- An object is an outlier if the nearest neighbors of the object are far away, i.e., the **proximity** of the object **significantly deviates** from the proximity of most of the other objects in the same data set
- Two types of proximity-based outlier detection methods
  - **Distance-based outlier detection**: An object o is an outlier if its neighborhood does not have enough other points
  - **Density-based outlier detection**: An object o is an outlier if its density is relatively much lower than that of its neighbors

Clustering-Based Outlier Detection

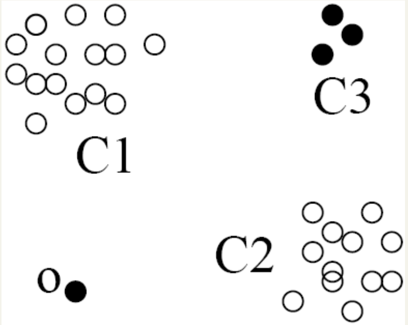
- An object is an outlier if
  - it does not belong to any cluster,





  - there is a large distance between the object and its closest cluster , or

  - it belongs to a small or sparse cluster



Clustering-Based Method: Strength and Weakness

- **Strength**
  - Detect outliers without requiring any labeled data
  - Work for many types of data
  - Clusters can be regarded as summaries of the data
  - Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)
- **Weakness**
  - Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection
  - High computational cost: Need to first find clusters
  - A method to reduce the cost: Fixed-width clustering
    - A point is assigned to a cluster if the center of the cluster is within a pre-defined distance threshold from the point
    - If a point cannot be assigned to any existing cluster, a new cluster is created and the distance threshold may be learned from the training data under certain conditions