Data Management

CT051-3-M

Topic 4 – Data Preprocessing – PART 1

Topic & Structure of Lesson

- Need for data preparation
- Multidimensional view of data quality
- Major tasks in data preprocessing
  - Data cleaning

Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation

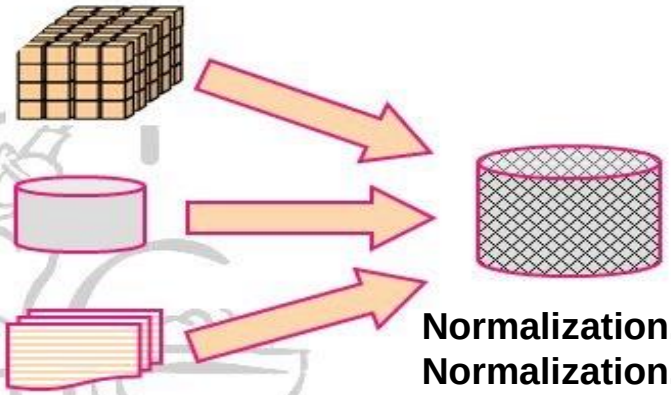
Multi-Dimensional Measure of Data Quality

Major Tasks in Data Preprocessing

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data Cleaning

Integration of multiple databases, data cubes, or files



Data Integration

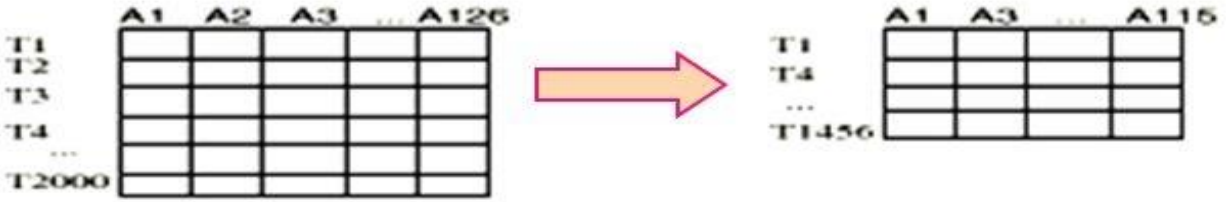
Normalization, Mean Normalization, Standardization

-2,32,100,59,48 → -0.02,0.32,1.00,0.59,0.48

Data Transformation

Duplication

Data Reduction



Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- Summary

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

How to Handle Missing Data?

1. Ignore the tuple (instance): usually done when class label is missing
2. Fill in the missing value manually: boring + infeasible?
3. Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
4. Use the attribute mean to fill in the missing value
5. Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
6. Use the most probable value to fill in the missing value

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

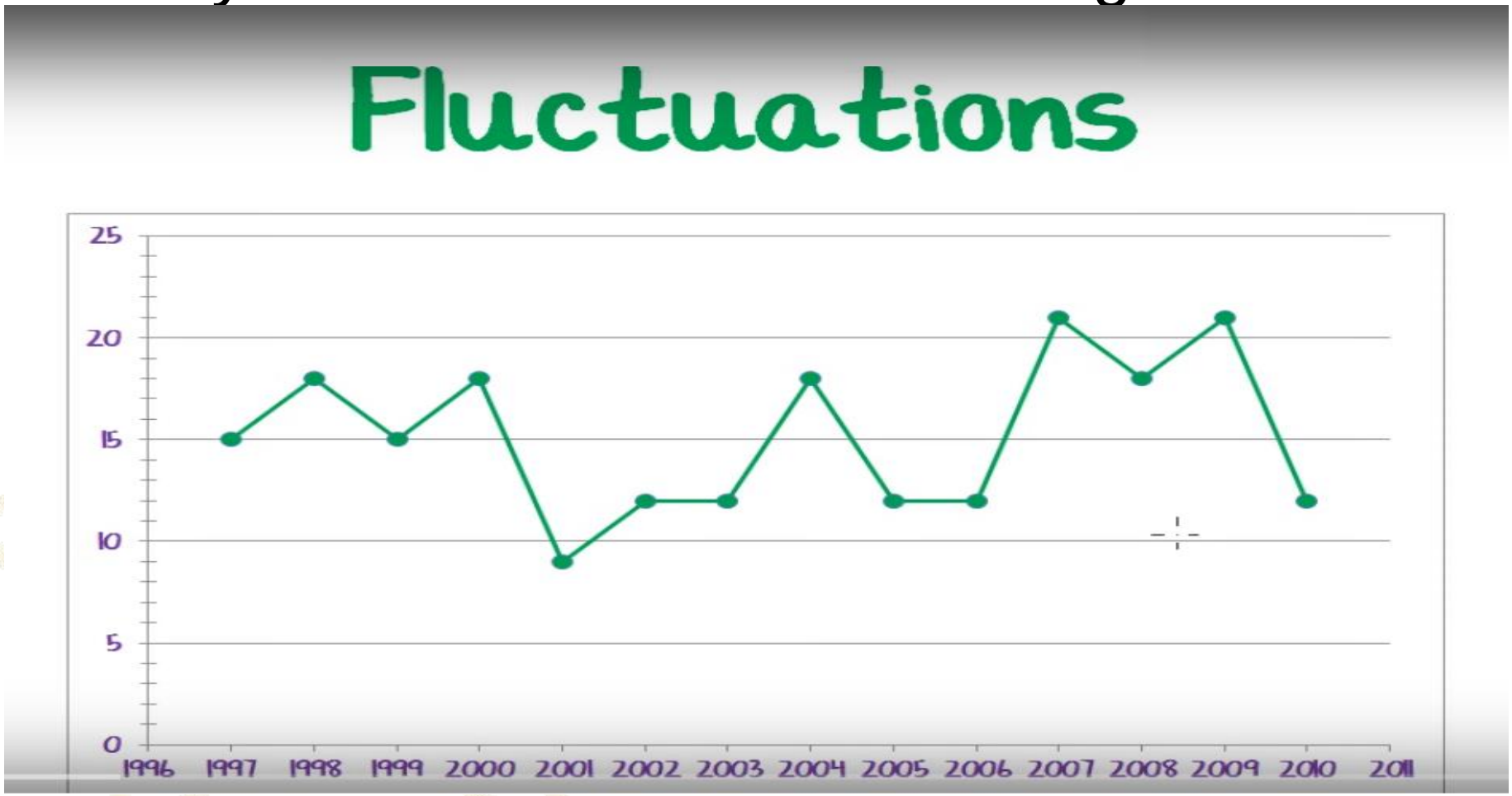
1. Binning method (Quantization):
   - first sort data and partition into (equi-depth) bins
   - then one can smooth by bin means, smooth by bin boundaries, etc.
2. Clustering
   - detect and remove outliers
3. Regression
   - smooth by fitting the data into regression functions
4. Combined computer and human inspection
   - detect suspicious values and check by human

- Just say age : 10-15 -children
- 16-20 - teen
- 21-25 - adults
- 26-30 - adults
- 31-40 - adults
- 41-54 –matured adults
- 55 above =S C

1. Binning Method

Binning – Data Smoothing

- Why do we need data smoothing ?



Binning Method

- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling

Binning Methods for Data Smoothing

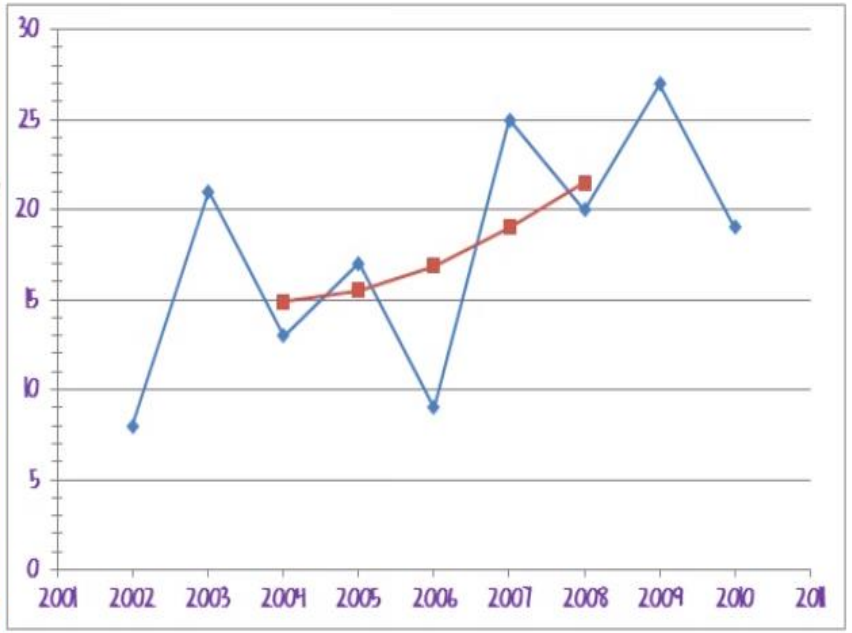
* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Example: "3 Mean Smoothing"

Year	Sales ('000)	3 Mean Smoothed
2002	8	
2003	21	$\frac{8+21+13}{3} = 14$
2004	13	$\frac{21+13+17}{3} = 17$
2005	17	$\frac{13+17+9}{3} = 13$
2006	9	$\frac{17+9+25}{3} = 17$
2007	25	$\frac{9+25+20}{3} = 18$
2008	20	$\frac{25+20+27}{3} = 24$
2009	27	$\frac{20+27+19}{3} = 22$

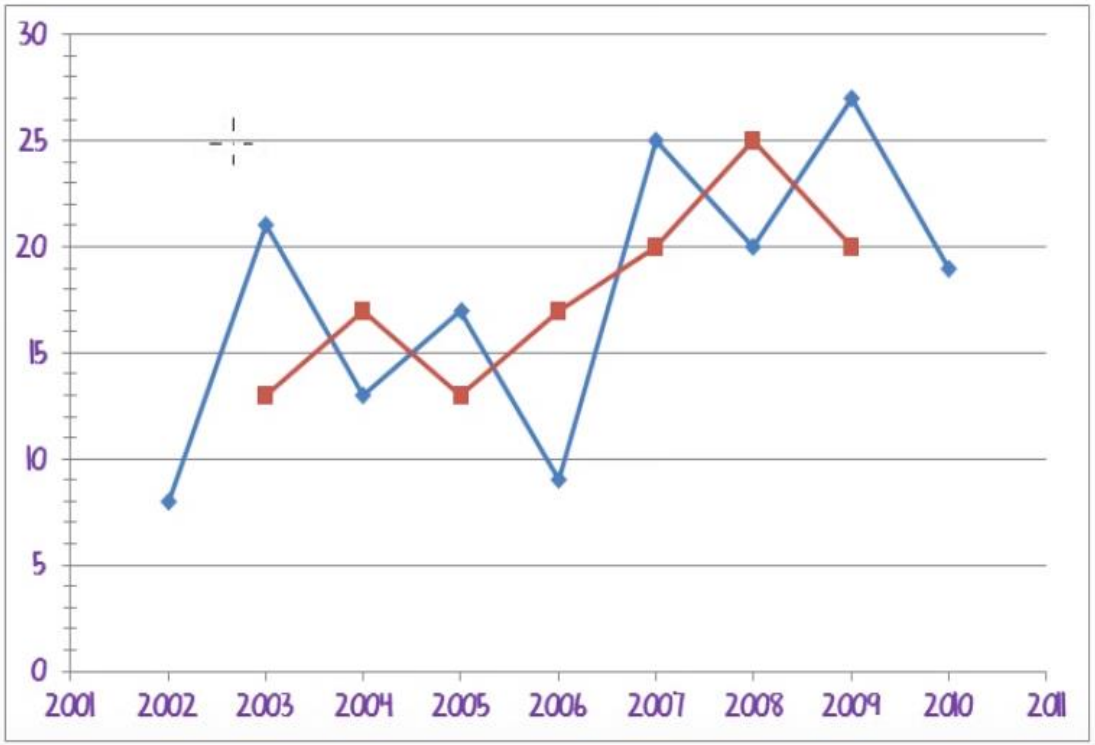
Example: Mean Smoothing - Centering

Year	Sales ('000)	4 Mean Smoothed	Centring
2002	8		
2003	21	$\frac{8+21+13+17}{4}=14.75$	$\frac{14.75+15}{2}=14.875$
2004	13	$\frac{21+13+17+9}{4}=15$	$\frac{15+16}{2}=15.5$
2005	17	$\frac{13+17+9+25}{4}=16$	$\frac{16+17.75}{2}=16.875$
2006	9	$\frac{17+9+25+20}{4}=17.75$	$\frac{17.75+20.25}{2}=19$
2007	25	$\frac{9+25+20+27}{4}=20.25$	$\frac{20.25+22.75}{2}=21.5$
2008	20	$\frac{25+20+27+19}{4}=22.75$	
2009	27		



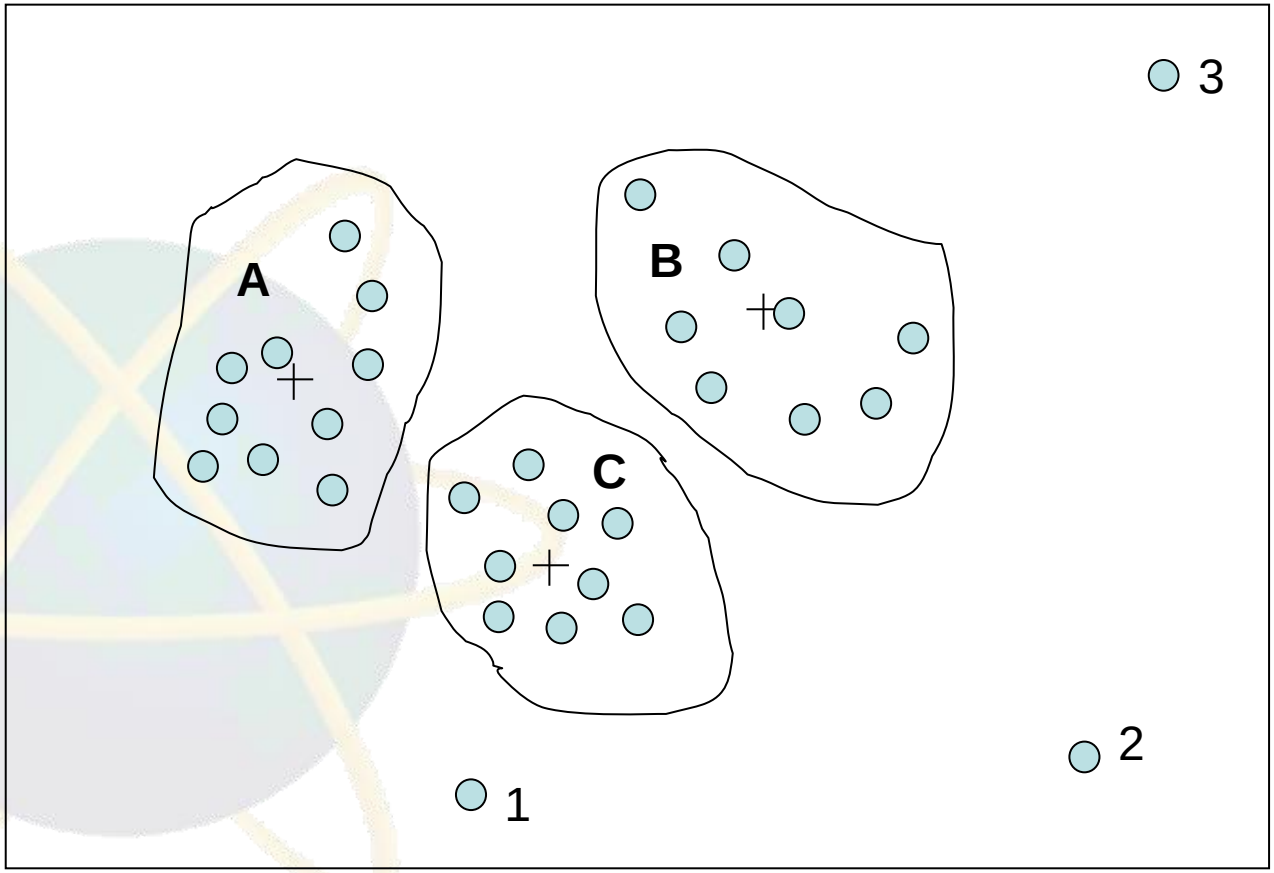
Example: Median Smoothing

Year	Sales ('000)	3 Median Smoothed
2002	8	
2003	21	13
2004	13	17
2005	17	13
2006	9	17
2007	25	20
2008	20	25
2009	27	20



2. Clustering

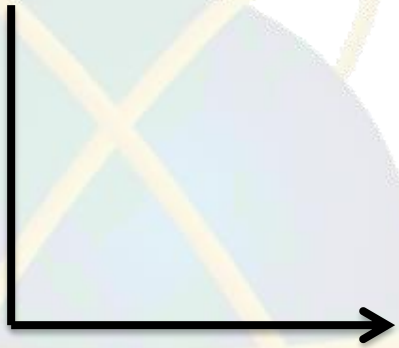
Cluster Analysis



What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION



General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - low inter-class similarity (between 2 classes)
  - high intra-class similarity (within a class)
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

Typical Requirements of Clustering in Data Mining

- Scalability : work good on small sets only
- Ability to deal with different types of attributes
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- High dimensionality
- Interpretability and usability

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning condition.
  - *k-means* : Each cluster is represented by the center of the cluster.

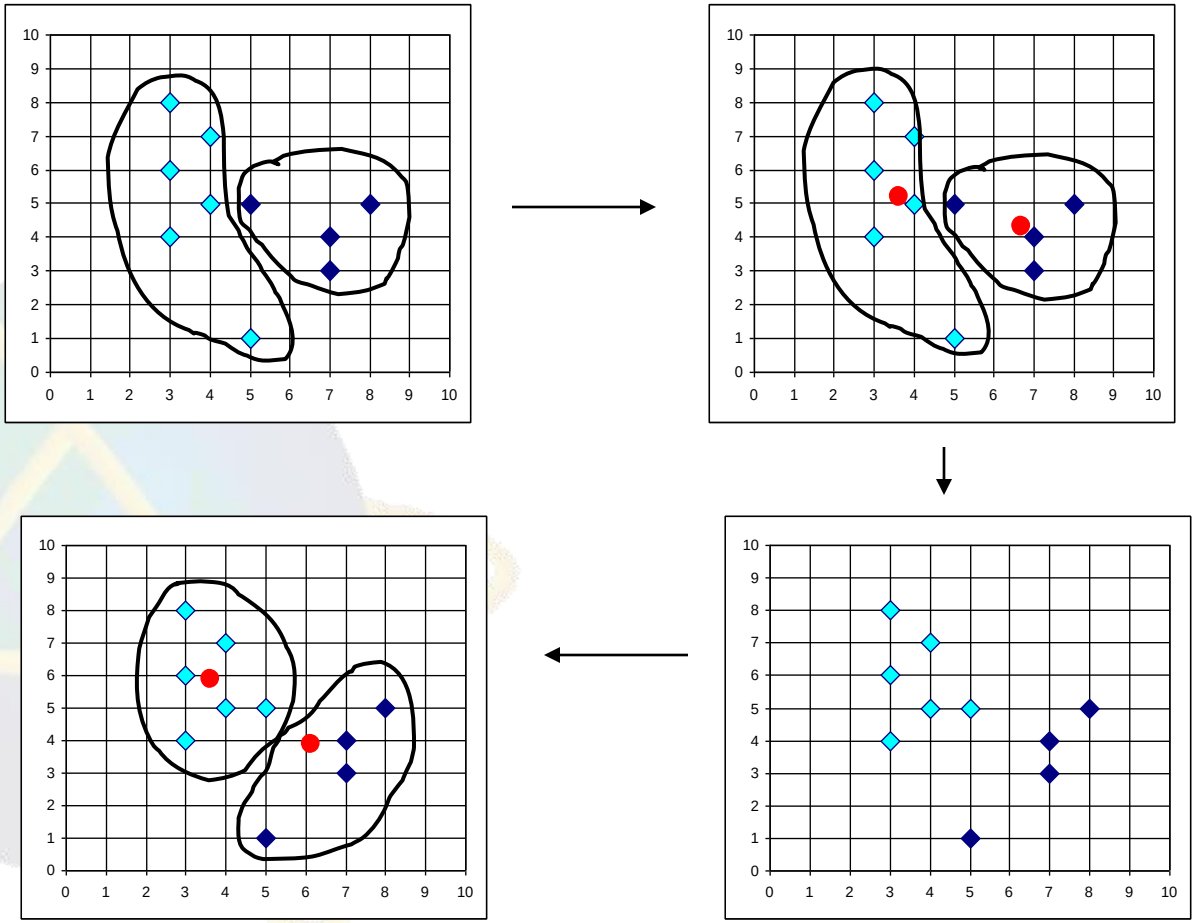
The *K-Means* Clustering Method

k-means algorithm is implemented in 5 steps:

- *Step 1: Ask the user how many clusters k the data set should be partitioned into.*
- *Step 2: Randomly assign k records to be the initial cluster center locations.*
- *Step 3: For each record, find the nearest cluster center. Thus, in a sense, each* cluster center "owns" a subset of the records, thereby representing a partition of the data set. We therefore have *k clusters, C1,C2, . . . ,Ck .*
- *Step 4: For each of the k clusters, find the cluster centroid, and update the* location of each cluster center to the new value of the centroid.
- *Step 5: Repeat steps 3 to 5 until convergence or termination.*

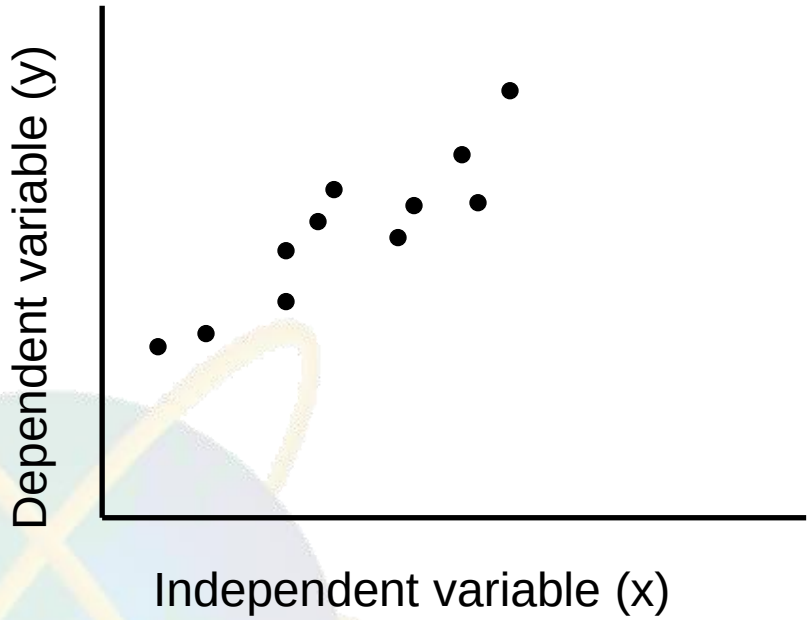
The *K-Means* Clustering Method

- Example



3. Regression

Regression

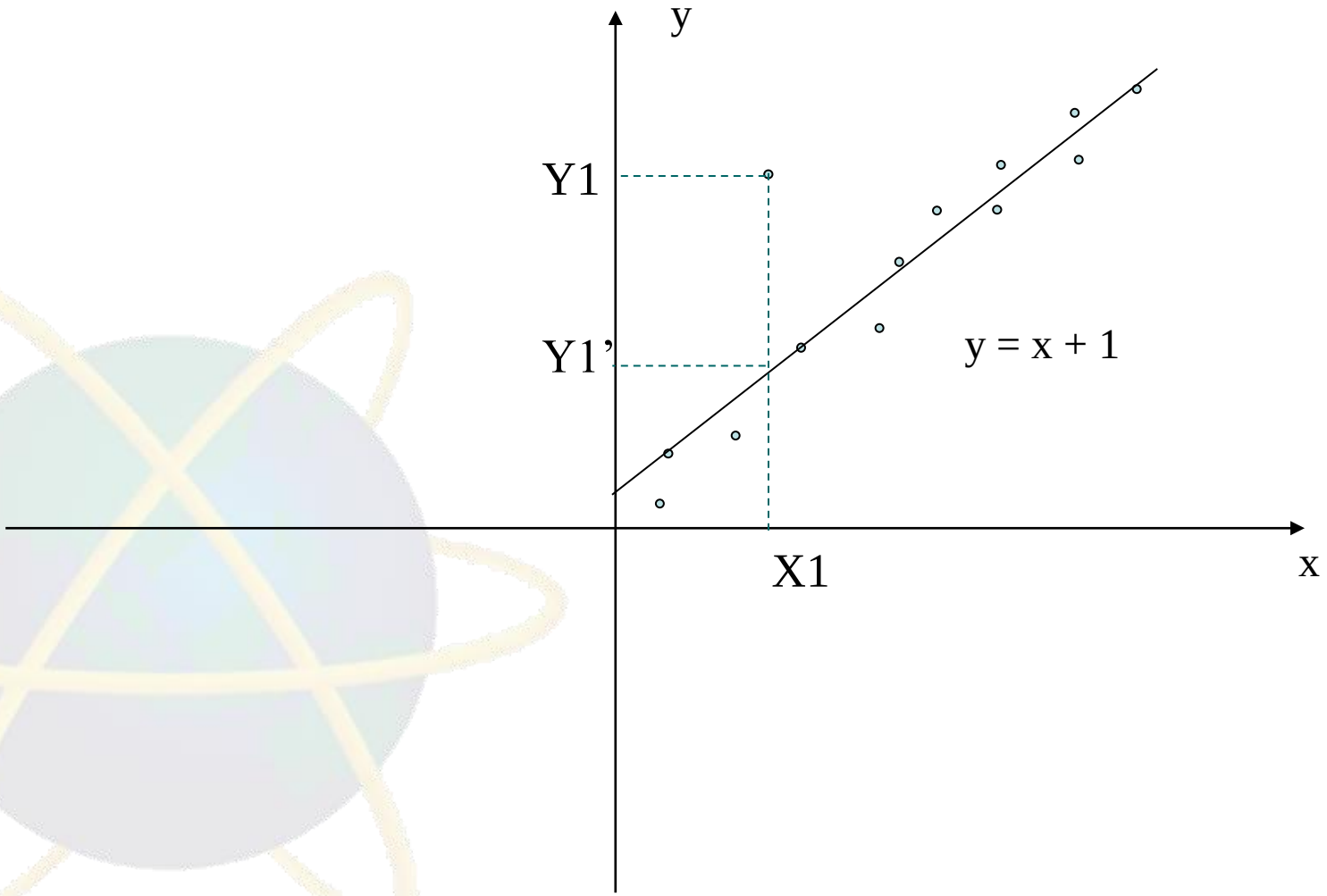


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

Regression



Data Integration and Transformation Next Topic

Question & Answer Session

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Q & A

Question & Answer Session

A · P · U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Q & A