DATA MINING LECTURE 2

Data Preprocessing

Data Preprocessing

- Data Preprocessing: An Overview
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Why Preprocess Data?

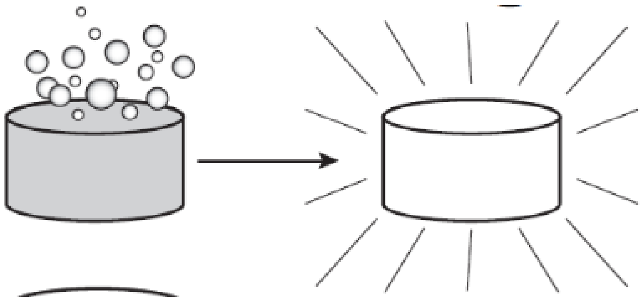
- Welcome to the Real World! *Data is Dirty!*
- But, No quality data, no quality mining results!
- Preprocessing is one of the most critical steps in a data mining process

Why Preprocess Data? Contd.

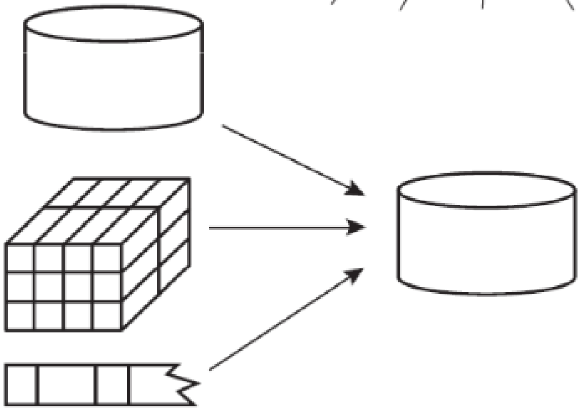
- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Preprocessing: Major Tasks

Data cleaning



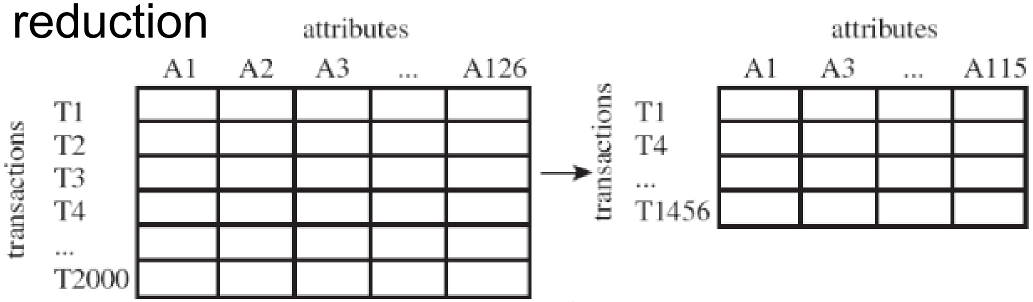
Data Integration



Data transformation

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

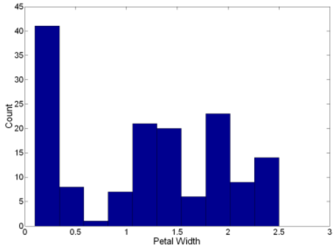
Data reduction



	A1	A2	A3	...	A126
T1					
T2					
T3					
T4					
...					
T2000					

	A1	A3	...	A115
T1				
T4				
...				
T1456				

Data Exploratory Analysis

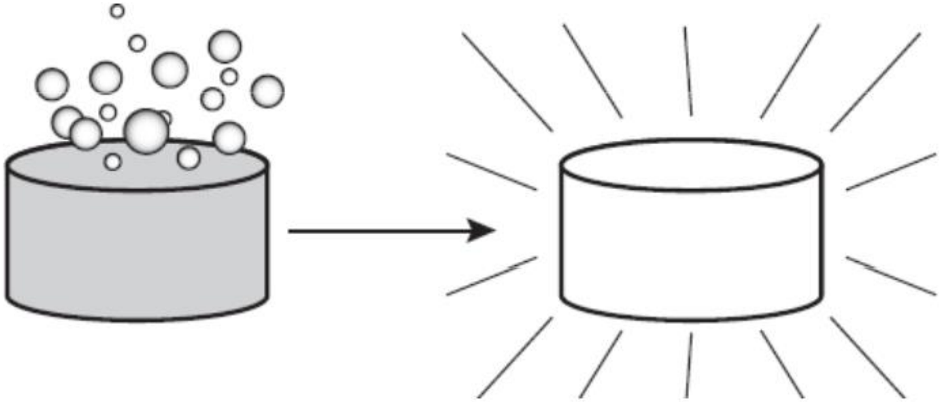


Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Cleaning

Data cleaning



Why Data Cleaning?

- **Data in the real world is dirty. ‘Dirty’ denotes-**
  - Incomplete/missing: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between **duplicate** records
  - Redundant: including everything, some of which are irrelevant to our task
  - Duplicate: Data-entry is duplicated.

Dirty Data

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

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**

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers/noise and smooth out noisy data
  - Correct inconsistent data
  - Remove duplicate/redundant 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
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Noisy Data

- What is noise?

  - Random error in a measured variable.

- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

How to Handle Noisy Data?

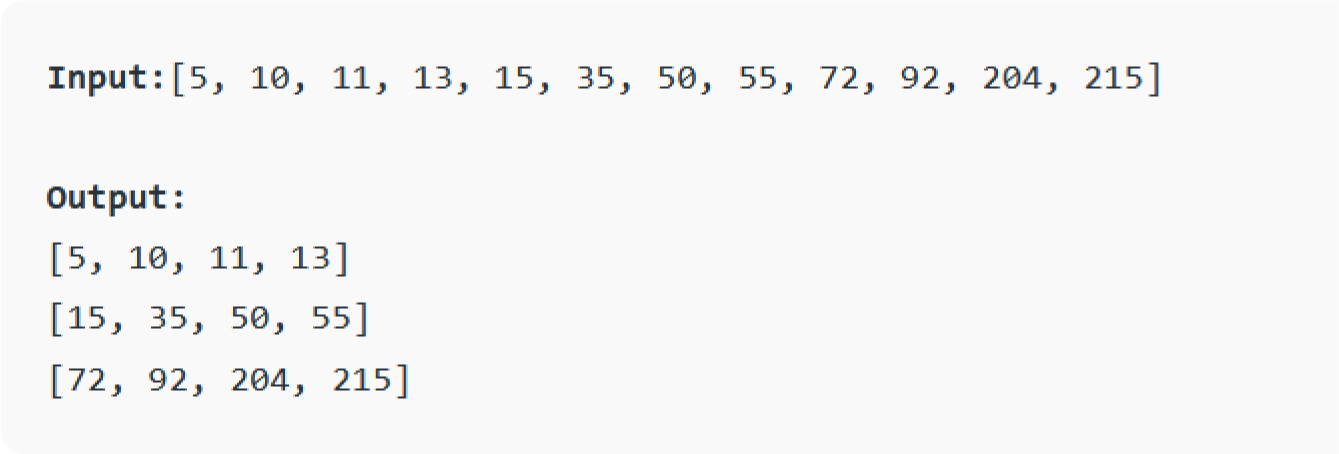
- Binning/bucketing
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - Advantage/disadvantage
    - The most straightforward
    - But outliers may dominate presentation
    - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Advantage/disadvantage
    - Good data scaling
    - Managing categorical attributes can be tricky.

Example

Equal frequency:

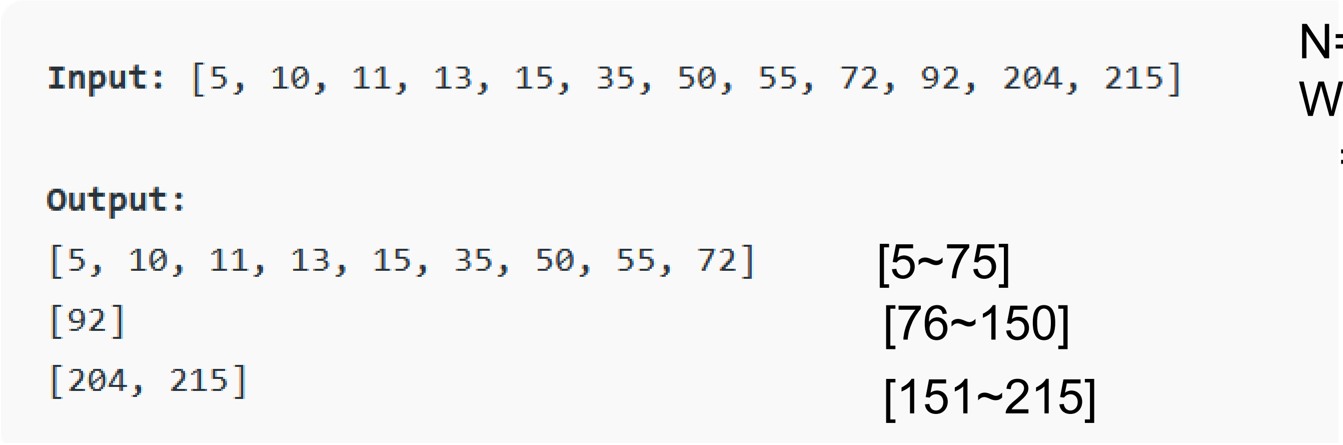


```
Input:[5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:
[5, 10, 11, 13]
[15, 35, 50, 55]
[72, 92, 204, 215]
```

Equal Width:

width of intervals will be: $W = (B-A)/N$



```
Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:
[5, 10, 11, 13, 15, 35, 50, 55, 72]
[92]
[204, 215]
```

N=3

W= (215-5)/3

= 75

[5~75]

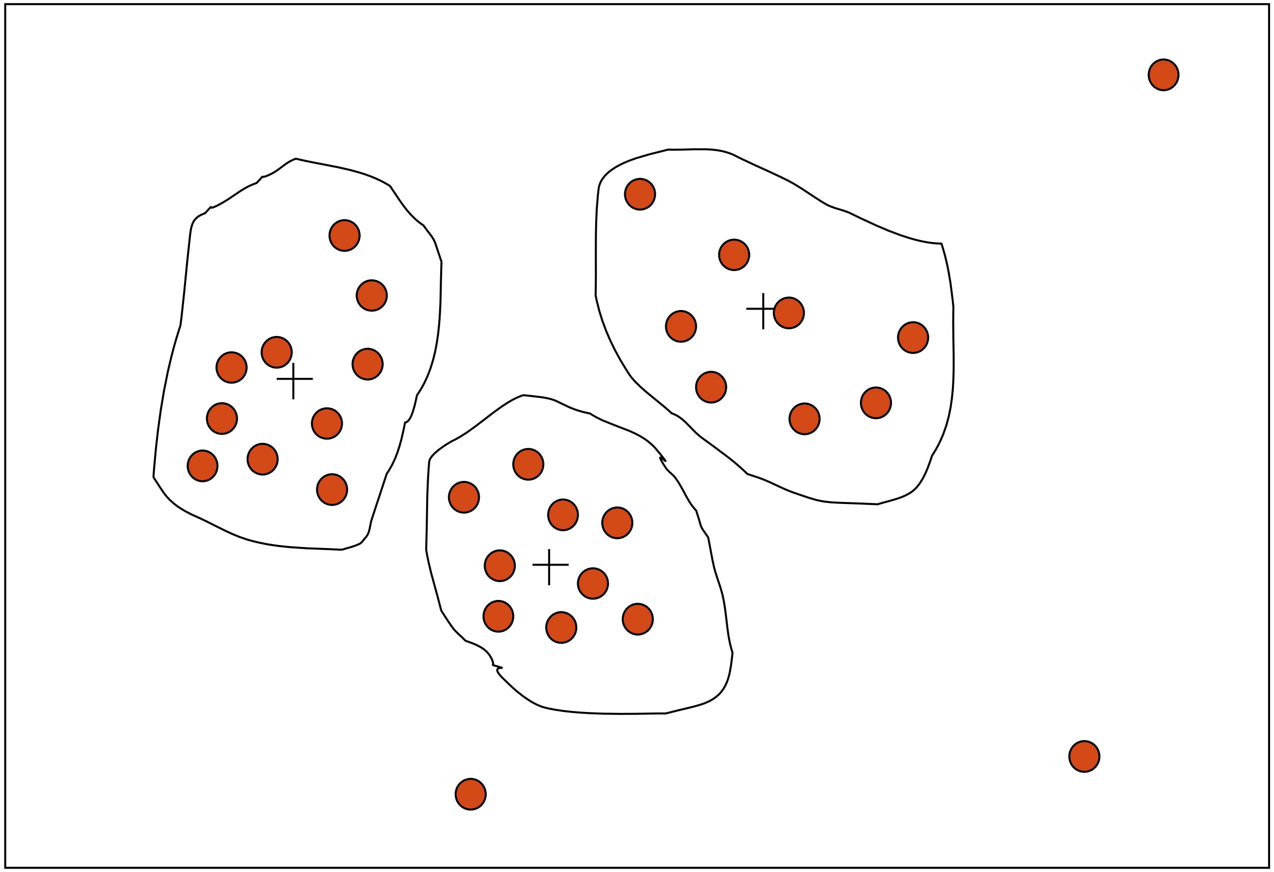
[76~150]

[151~215]

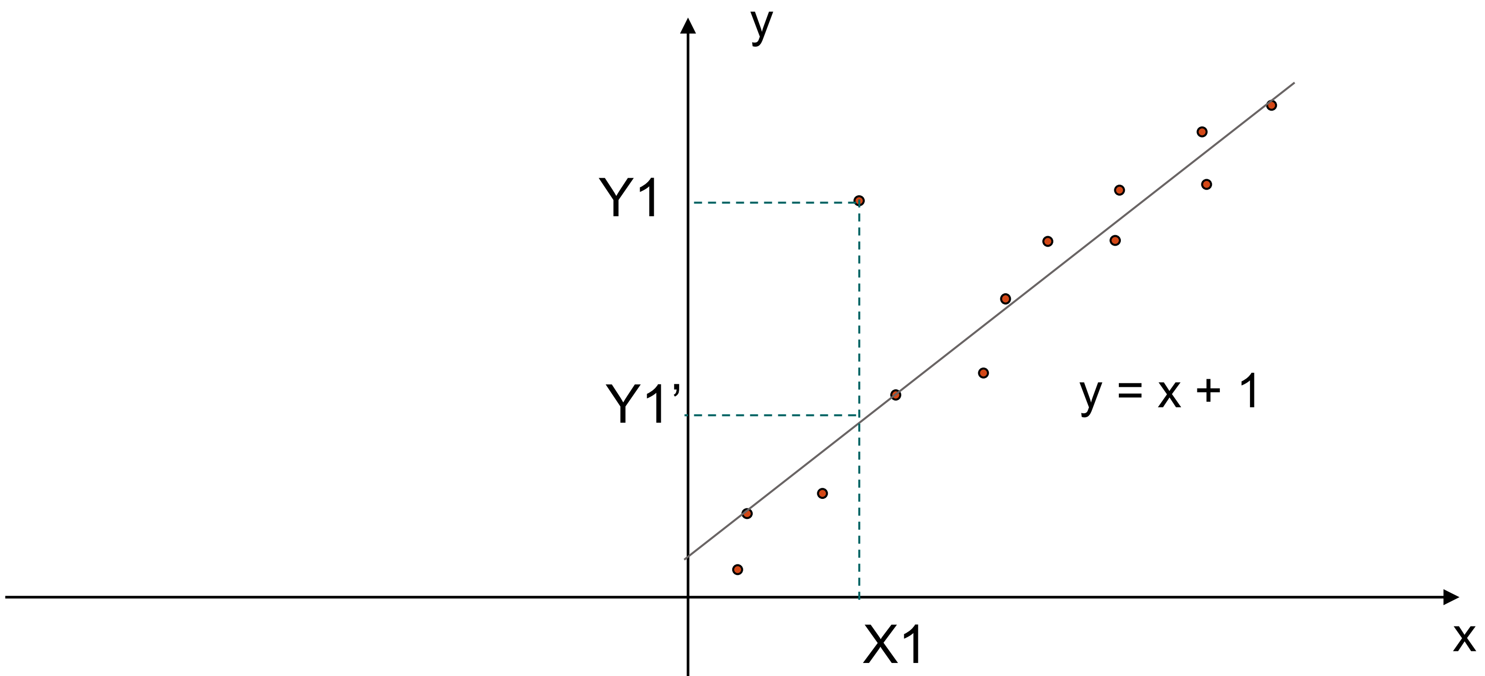
Equal depth/frequency binning: An example

- 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

Cluster Analysis



Regression



Note

- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface

How to Handle Inconsistent Data?

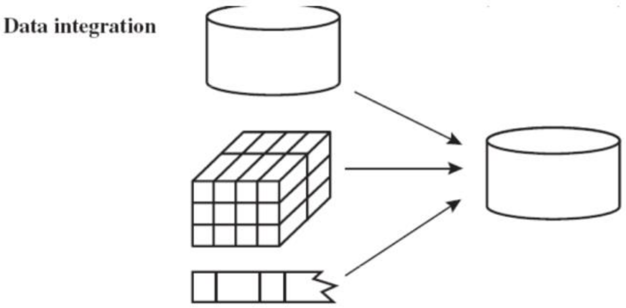
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

- Handling Inconsistent Data-
  - Manual correction using external references
  - Semi-automatic using various tools
    - To detect violation of known functional dependencies and data constraints
    - To correct redundant data

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Integration



Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration and object matching
  - Entity identification problem:
    - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
    - Value conflict, e.g., good height = tall; poor height = short
- Why such problem arises?
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Schema integration and object matching

- *custom_id* and *cust_number*
  - – Schema conflict

- "Yes" and "No", and '1' and '0' for an attribute
  - – Value conflict

- Solutions
  - - meta data (data about data)

Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Detecting Redundancy

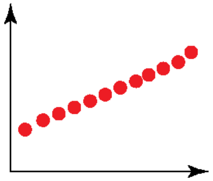
- If an attributed can be “derived” from another attribute or a set of attributes, it may be redundant.
- Some redundancies can be detected by correlation analysis, e.g.,

  – **Correlation analysis**

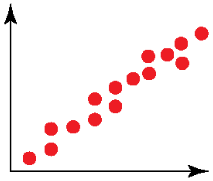
  – Chi-square test

Correlation Analysis

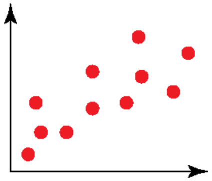
- The correlation coefficient is a value that indicates the strength of the relationship between variables.
- The coefficient can take any values from -1 to 1. The interpretations of the values are:
  - **-1:** Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).
  - **0:** No correlation. The variables do not have a relationship with each other.
  - **1:** Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).



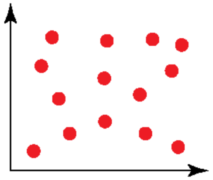
Perfect Positive Correlation



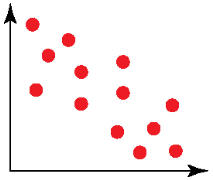
Strong Positive Correlation



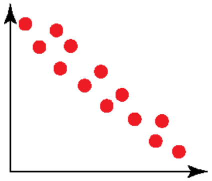
Weak Positive Correlation



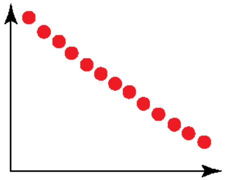
No Correlation



Weak Negative Correlation

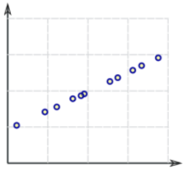


Strong Negative Correlation



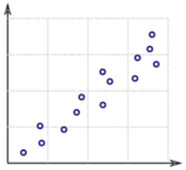
Perfect Negative Correlation

Perfect Positive Correlation



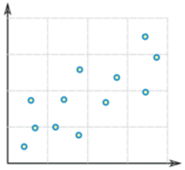
1

High Positive Correlation



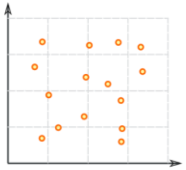
0.9

Low Positive Correlation



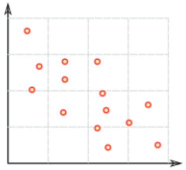
0.5

No Correlation



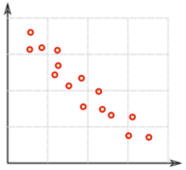
0

Low Negative Correlation



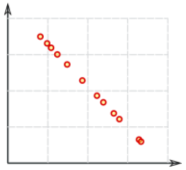
-0.5

High Negative Correlation



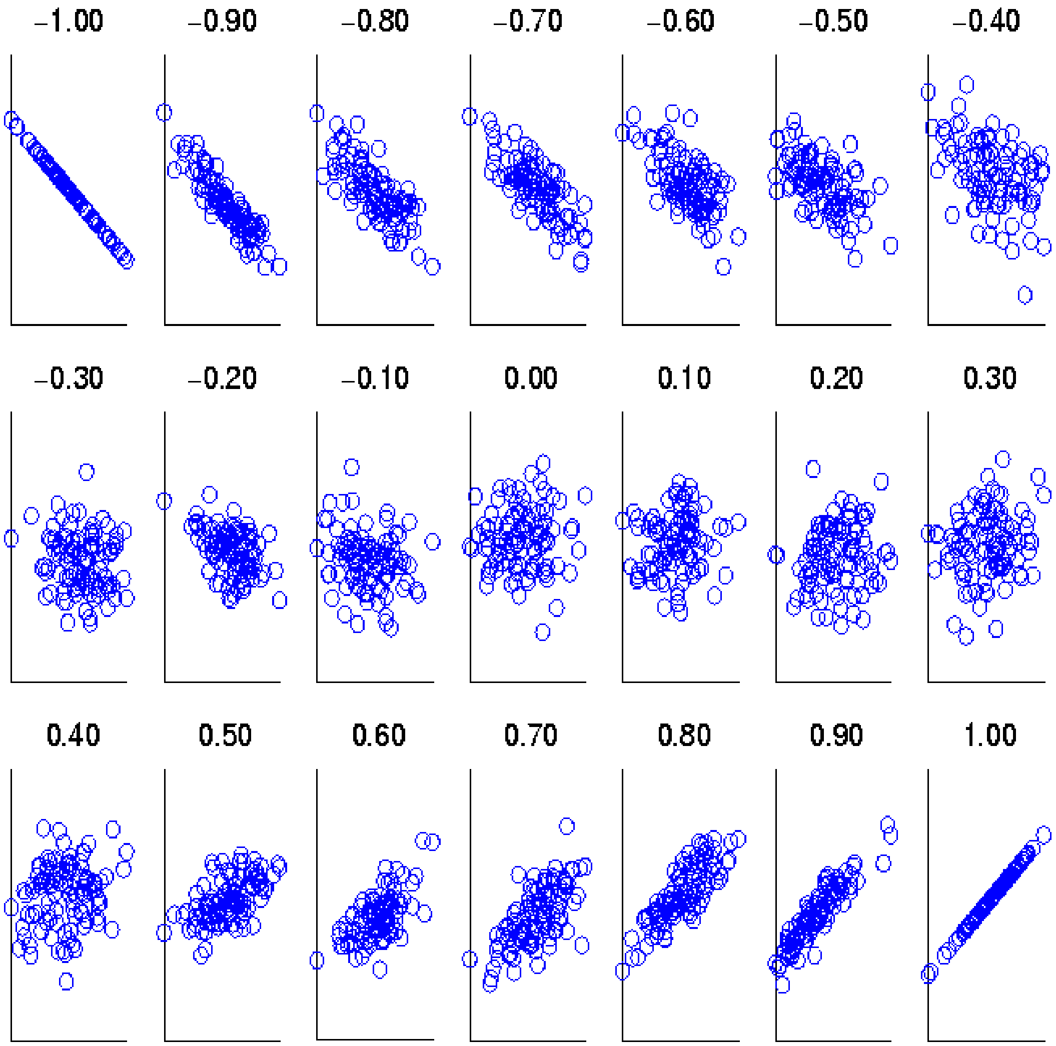
-0.9

Perfect Negative Correlation



-1

Visually Evaluating Correlation

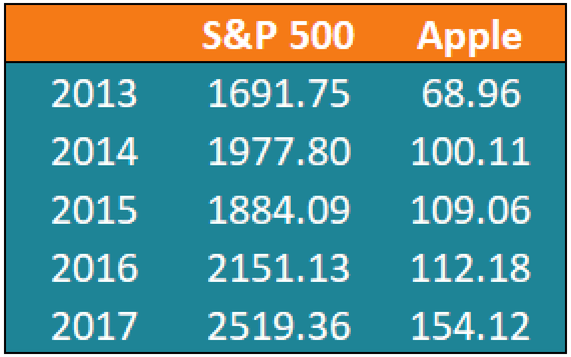


Scatter plots showing the similarity from –1 to 1.

Correlation Coefficient, *r*

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

X Y



	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12

Where:

- **r_{xy}** – the **correlation coefficient** of the linear relationship between the variables x and y. It is also known as ***Pearson's correlation coefficient*** when applied to sample data.

- **x_i** – the values of the x-variable in a sample

- **$\bar{x}$** – the mean of the values of the x-variable

- **y_i** – the values of the y-variable in a sample

- **$\bar{y}$** – the mean of the values of the y-variable

Example Data:

	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12
	X	Y

Task:

Find correlation between two attributes [X and Y]

We will do it by finding the **correlation-coefficient** between them

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Step 3 **Step 4** **Step 5**

	S&P 500	Apple	a $(x-\bar{x})$	b $(y-\bar{y})$	a x b	a^2	b^2
2013	1691.75	68.96	- 353.08	- 39.93	14,096.91	124,662.66	1,594.09
2014	1977.80	100.11	- 67.03	- 8.78	588.22	4,492.48	77.02
2015	1884.09	109.06	- 160.74	0.17	- 27.97	25,836.07	0.03
2016	2151.13	112.18	106.30	3.29	350.16	11,300.52	10.85
2017	2519.36	154.12	474.53	45.23	21,465.08	225,182.62	2,046.11
Mean	**2044.83**	**108.89**	**Sums**		**36,472.40**	**391,474.35**	**3,728.10**

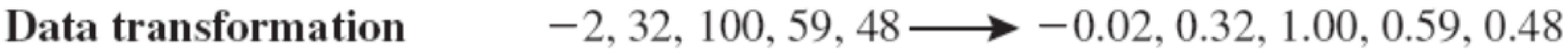
$$r_{xy} = \frac{36,472.40}{\sqrt{391,474.35 \times 3,728.10}} = 0.95$$

The coefficient indicates that the given variables have a high positive correlation.

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation ⬅
- Data Reduction
- Summary

Data Transformation



Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data Transformation

- √ Smoothing: remove noise from data (binning, clustering, regression)
- √ Aggregation: summarization, data-cube generation
- Generalization: concept hierarchy climbing
- √ Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- √ Attribute/feature construction
  - New attributes constructed from the given ones

Data Summarization

- Summary Statistics
  - Summary statistics are numbers that summarize properties of the data

    - Summarized properties include frequency, location and spread
      - Examples: location - mean
        spread - standard deviation

    - Most summary statistics can be calculated in a single pass through the data

Central Tendency: Mean and Median

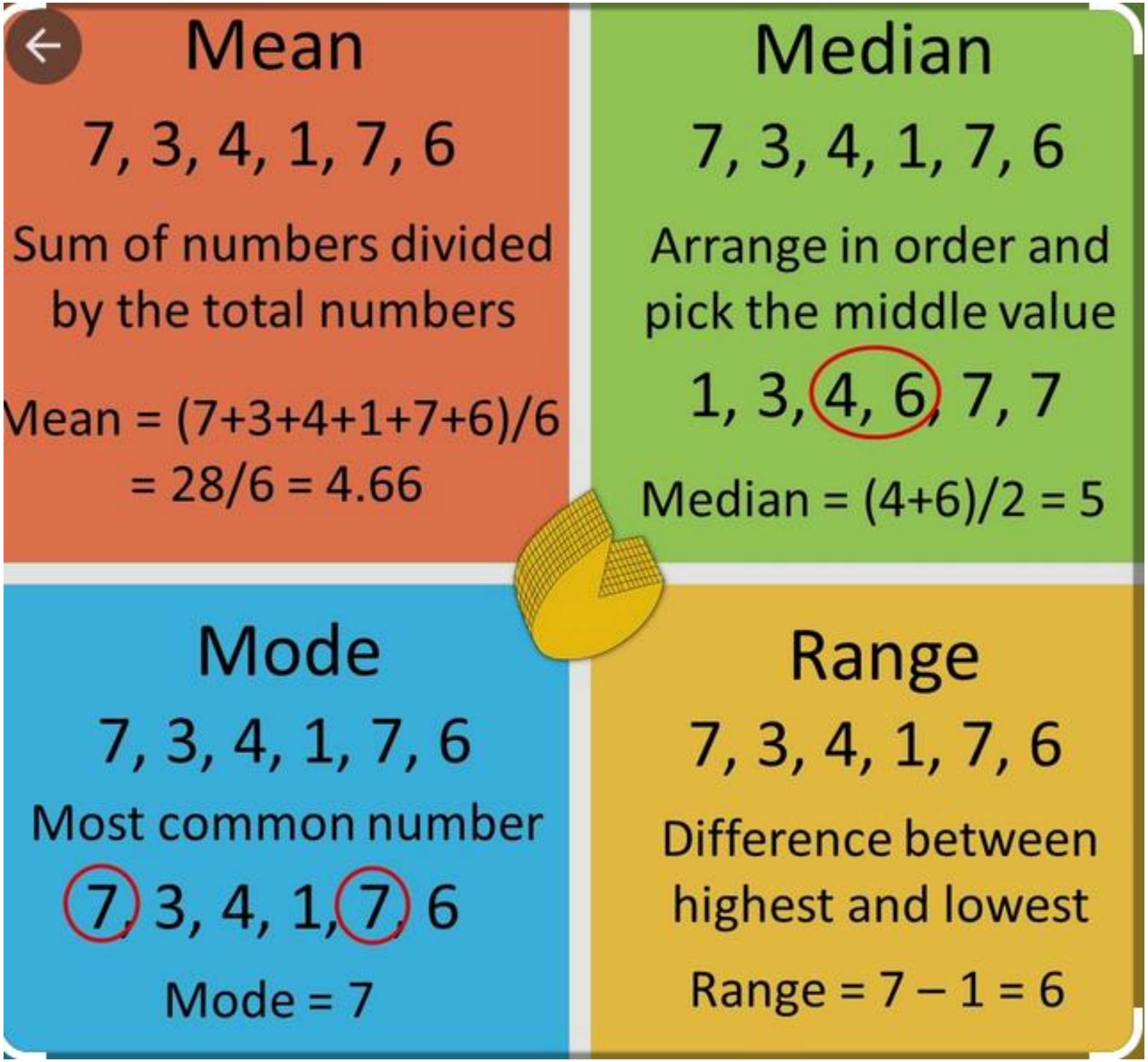
- The mean/average is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\mathrm{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\mathrm{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value



Mean

7, 3, 4, 1, 7, 6

Sum of numbers divided by the total numbers

Mean = (7+3+4+1+7+6)/6
= 28/6 = 4.66

Median

7, 3, 4, 1, 7, 6

Arrange in order and pick the middle value

1, 3, 4, 6, 7, 7

Median = (4+6)/2 = 5

Mode

7, 3, 4, 1, 7, 6

Most common number

7, 3, 4, 1, 7, 6

Mode = 7

Range

7, 3, 4, 1, 7, 6

Difference between highest and lowest

Range = 7 – 1 = 6

Data Transformation: Normalization

Data Transformation: Normalization

Min-max normalization

$$v' = \frac{v \ - \min(A)}{\max(A) - \min(A)}$$

Z-score normalization

$$v' = \frac{v \ - \ \mu}{\sigma}$$

Decimal Scaling

$$v' = \frac{v}{10^j}$$

min-max normalization

Given Dataset, [A: 8, 10, 15, 20]

Task: Perform min-max normalization on the given data, and calculate normalized values.

Equation

$$v' = \frac{v \ - \min(A)}{\max(A) - \min(A)}$$

$\min(A) = 8$
$\max(A) = 20$

For v=8,
Normalized value, $v' = \frac{8 - 8}{20 - 8}$
$= 0$

For v=10,
Normalized value, $v' = \frac{10 - 8}{20 - 8}$
$= 0.16$

For v=15,
Normalized value, $v' = \frac{15 - 8}{20 - 8}$
$= 0.58$

For v=20,
Normalized value, $v' = \frac{20 - 8}{20 - 8}$
$= 1.0$

min-max normalization (contd.)

Given Dataset, [A: 8, 10, 15, 20]

Task: Perform min-max normalization on the given data **within a range [40, 60]**, , and calculat normalized values.

Equation

$$v' = \frac{v \ - \min(A)}{\max(A) - \min(A)} [newmax(A) - newmin(A)] + newmin(A)$$

$\min(A)$ = 8

$\max(A) = 20$

$newmax(A) = 40$

$newmax(A)$ = 60

For v=8,

Normalized value, $v' = \frac{8 - 8}{20 - 8}$ [60-40]+40

$= 40$

For v=10,

Normalized value, $v' = \frac{10 - 8}{20 - 8}$ [60-40]+40

$= \ 43.33$

For v=15,

Normalized value, $v' = \frac{15 - 8}{20 - 8}$ [60-40]+40

$= \ 51.67$

For v=20,

Normalized value, $v' = \frac{20 - 8}{20 - 8}$ [60-40]+40

$= \ 60$

Z-Score Normalization

Given Dataset, [A: 8, 10, 15, 20]

Task: Perform z-score normalization on the given data, and calculate normalized values.

Equation $$v' = \frac{v - \mu}{\sigma}$$

Mean, $\mu = \frac{\Sigma x_i}{N} = \frac{8+10+15+20}{4} = 13.25$

Standard Deviation, $\sigma = \sqrt{\frac{\Sigma\,(x_i - \mu)^2}{N}}$

$= \sqrt{\frac{(8-13.25)^2+(10-13.25)^2+(15-13.25)^2+(20-13.25)^2}{4}} = 4.65$

For v=8,

Normalized value, $v' = \frac{8 - 13.25}{4.65}$

$= -1.129$

For v=10,

Normalized value, $v' = \frac{10 - 13.25}{4.65}$

$= -0.698$

For v=15,

Normalized value, $v' = \frac{15 - 13.25}{4.65}$

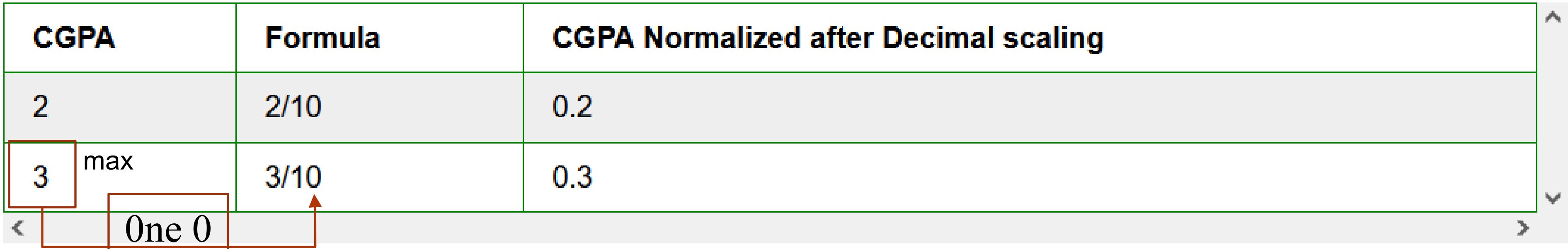
$= 0.376$

For v=20,

Normalized value, $v' = \frac{20 - 13.25}{4.65}$

$= 1.45$

Example of Decimal scaling :



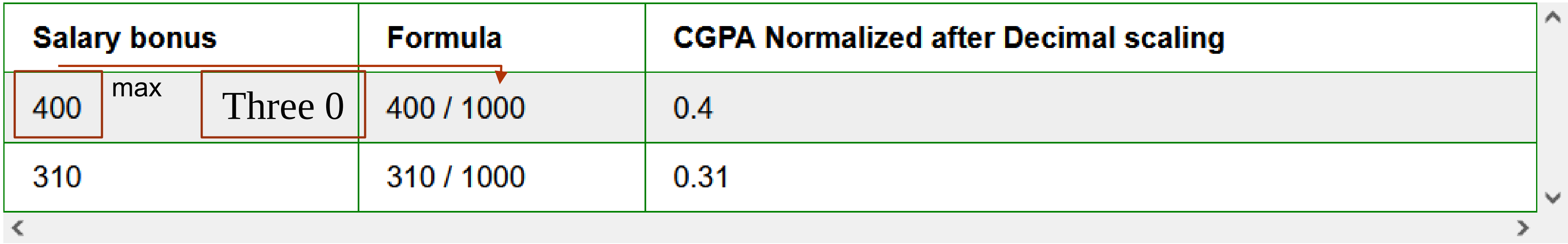
CGPA	Formula	CGPA Normalized after Decimal scaling
2	2/10	0.2
3	3/10	0.3

We will check the maximum value among our attribute CGPA. Here maximum value is 3 so we can convert it to a decimal by dividing by 10. Why 10?

we will count total numbers in our maximum value and then put 1 and after 1 we can put zeros equal to the length of the maximum value.

Here 3 is the maximum value and the total numbers in this value are only 1. so we will put one zero after one.

Example 2:



Salary bonus	Formula	CGPA Normalized after Decimal scaling
400	400 / 1000	0.4
310	310 / 1000	0.31

We will check the maximum value of our attribute “**salary bonus**“. Here maximum value is 400 so we can convert it into a decimal by dividing it by 1000. Why 1000?

400 contains three digits and we so we can put three zeros after 1. So, it looks like 1000.

Attribute/Feature Construction

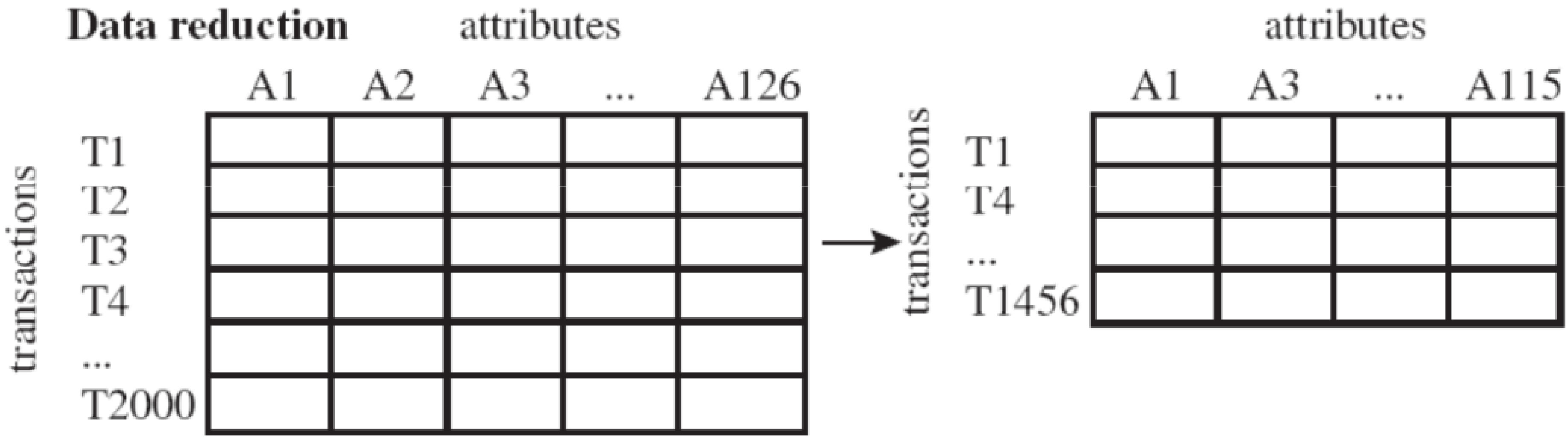
- New attributes are constructed from given attributes and added in order to help improve accuracy and understanding of structure in high-dimension data
- Example
  - Add the attribute *area* based on the attributes *height* and *width*

Area = height X width

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Reduction



- Reduction in number of sample
- Reduction in number of attribute (dimension)

Data Reduction

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set

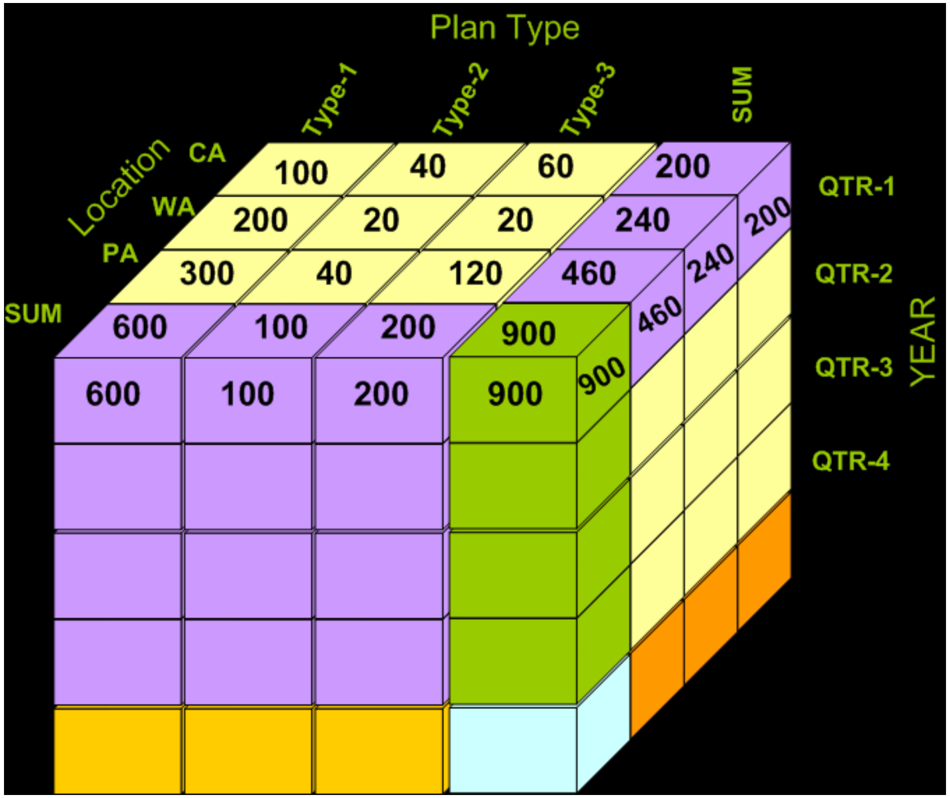
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

Data reduction strategies

- Data cube aggregation-
- Dimensionality reduction — e.g., remove unimportant attributes, Principal Component analysis (PCA), Wavelet transforms, feature-selection techniques
- Numerosity reduction — alternative-brief form of data representation; e.g., Regression and Log-Linear Models, Histograms, clustering, sampling
- Data Discretization – e.g., Binning, entropy-based discretization, cluster analysis, interval merging by chi-square analysis etc.
- Concept hierarchy generation-

Data Reduction 1: Data Cube Aggregation

- Multiple levels of aggregation in data cubes
- Queries regarding aggregated information should be answered using data cube, when possible

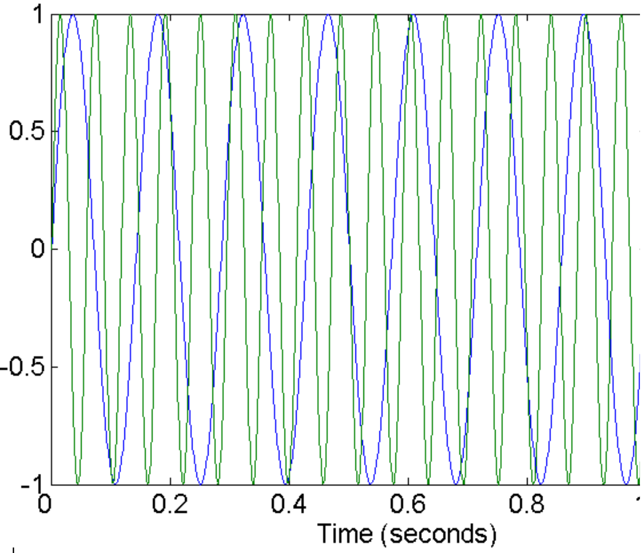


Data Reduction 2: Dimensionality Reduction

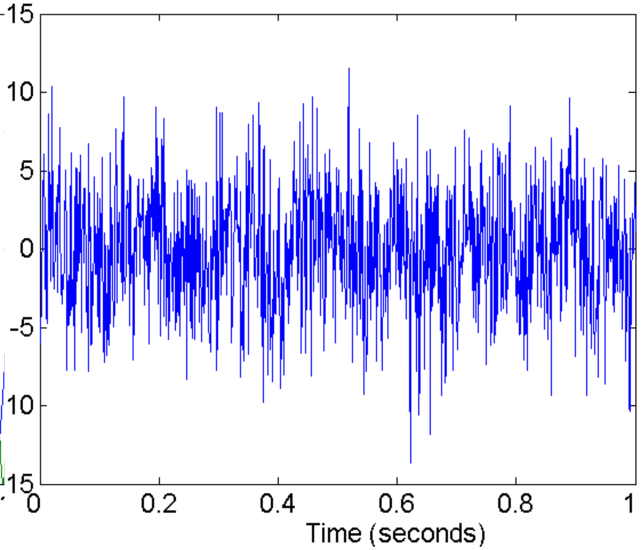
- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

Mapping Data to a New Space

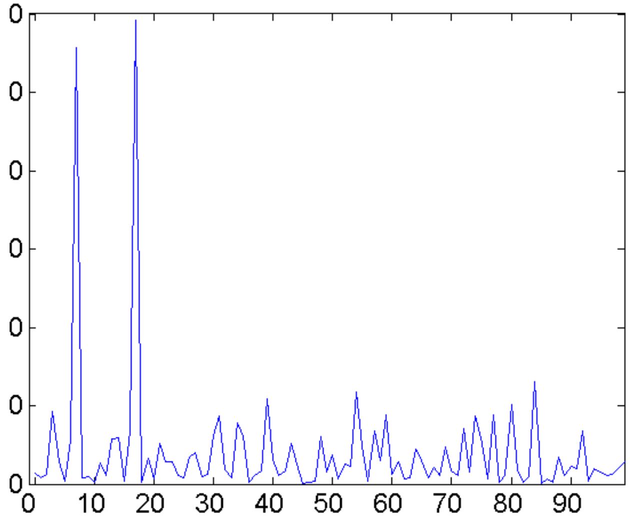
- **Fourier transform**
- **Wavelet transform**



Two Sine Waves



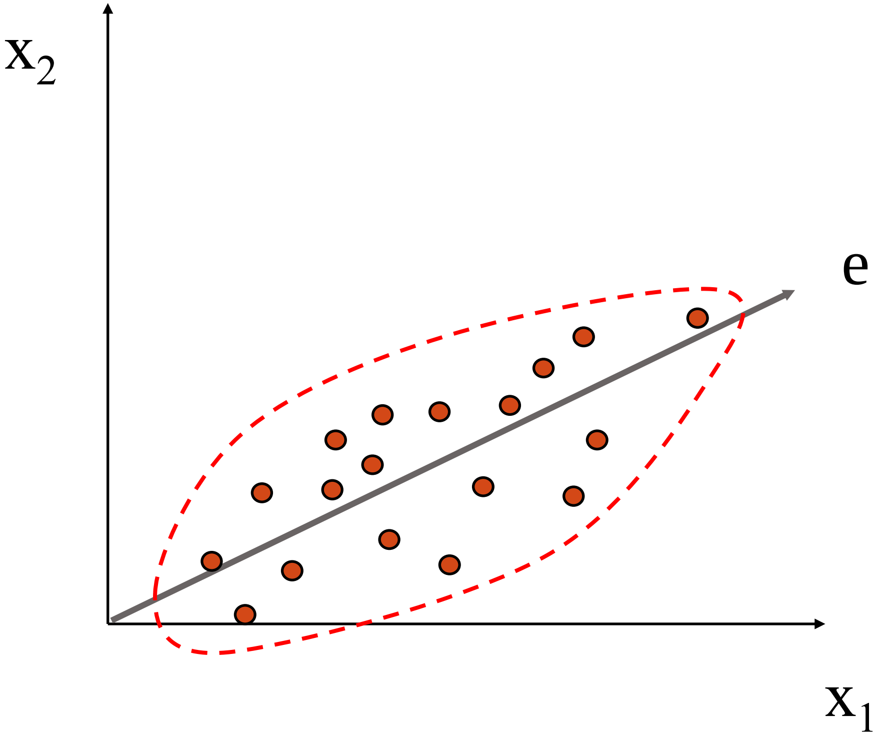
Two Sine Waves + Noise



Frequency

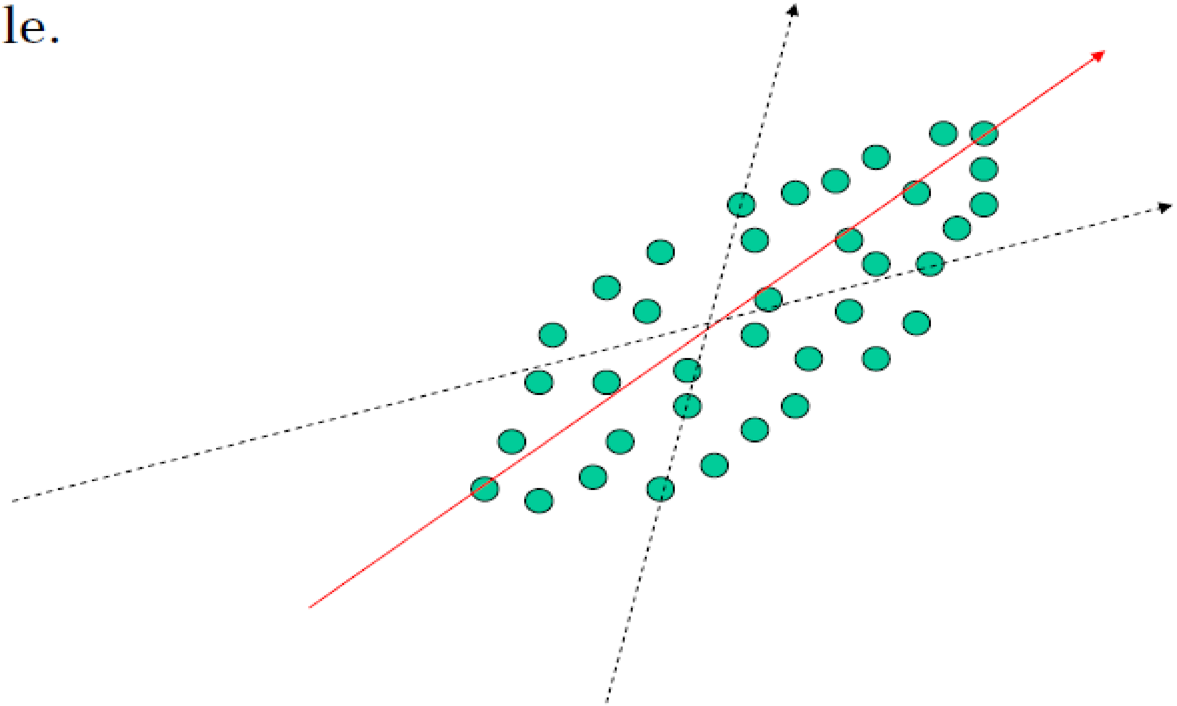
Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space.



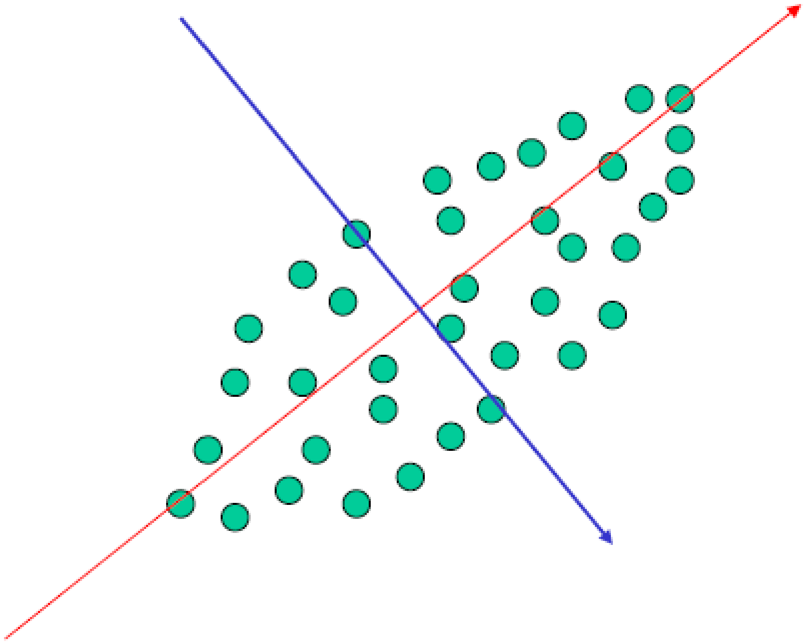
Basic Idea of PCA

Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



Basic Idea of PCA

Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



Attribute Subset Selection [Feature Selection]

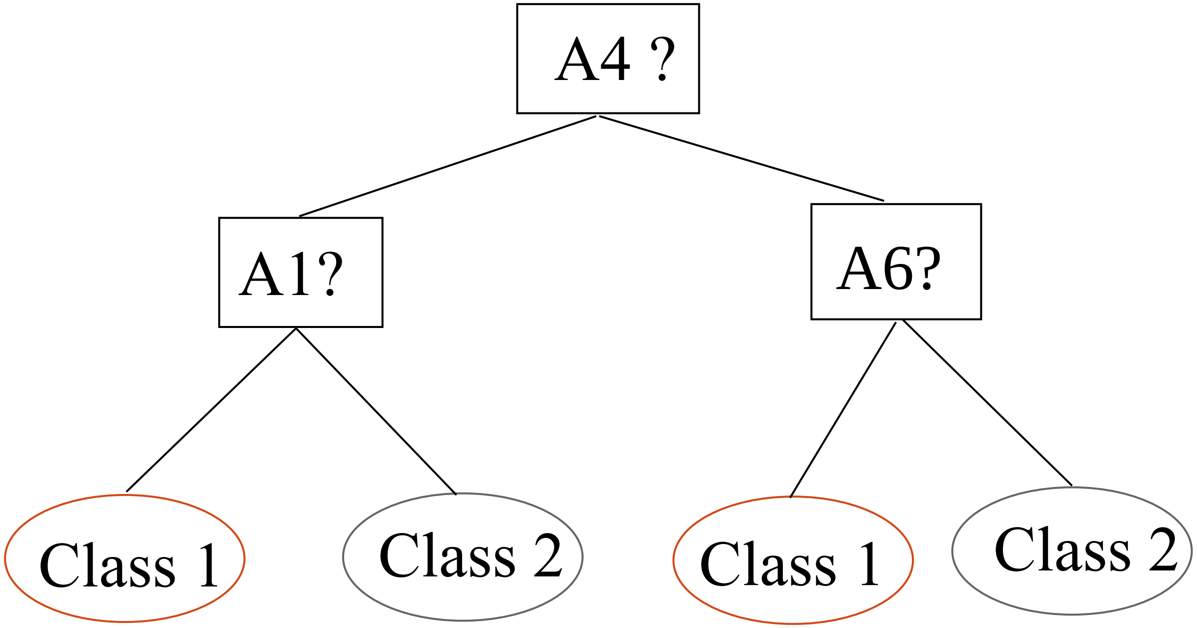
- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Heuristic Search in Attribute Selection

- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests (statistical approaches)
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

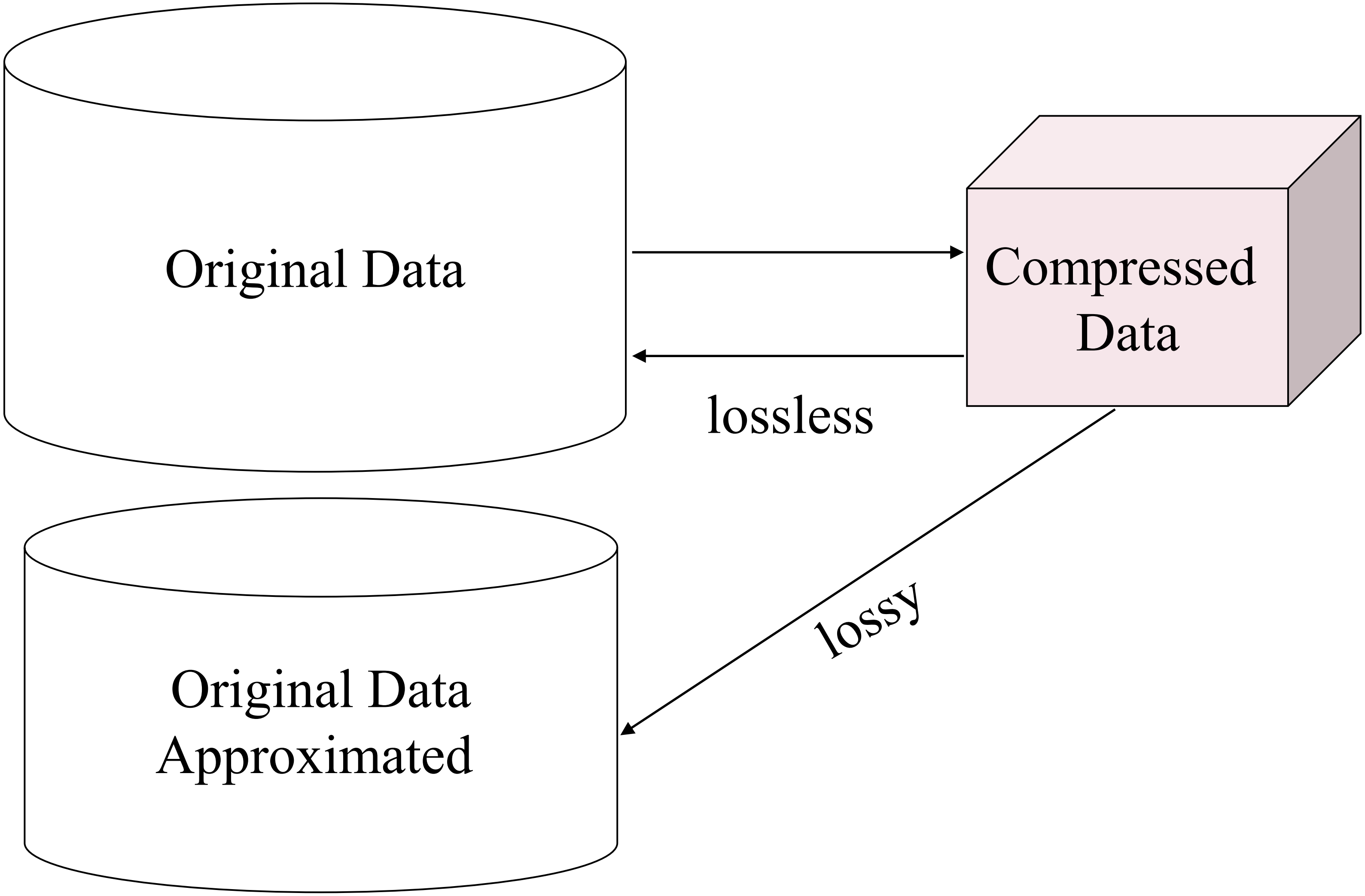
Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



------> Reduced attribute set: {A1, A4, A6}

Data Compression



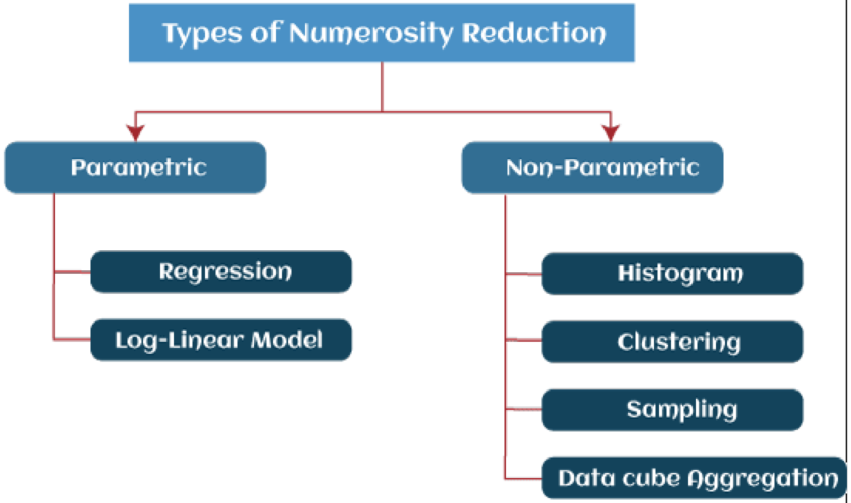
Data Reduction 3: Numerosity Reduction

- In the Numerosity reduction, the data volume is reduced by choosing an alternative, smaller form of data representation.

- These techniques may be ***parametric*** or ***nonparametric***.

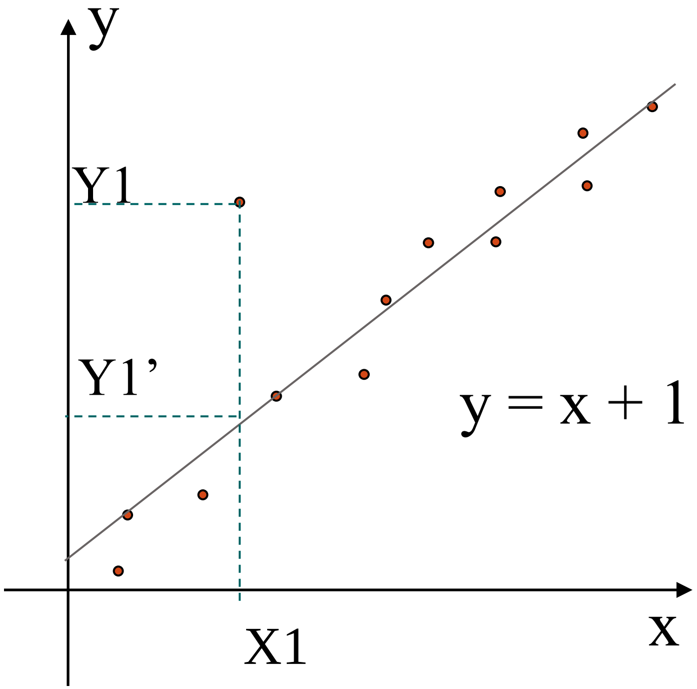
- ***For parametric methods***, a model is used to estimate the data, so that only the data parameters need to be stored, instead of the actual data, for example, Log-linear models.

- ***Non-parametric methods*** are used for storing a reduced representation of the data which include histograms, clustering, and sampling.



Parametric Method: Regression Analysis

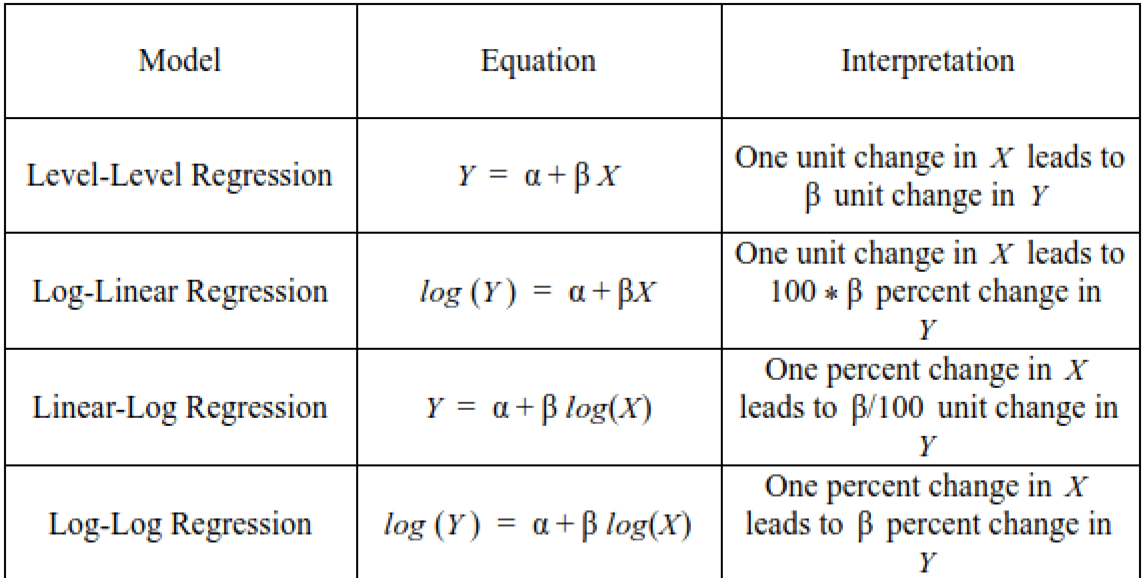
- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regression Analysis and Log-Linear Models

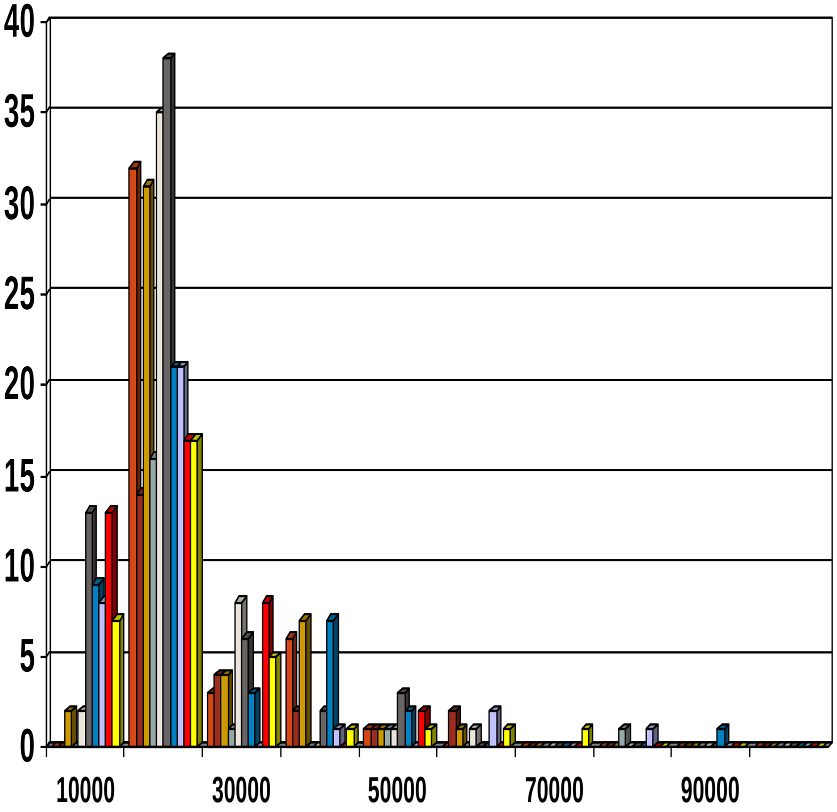
- Linear regression: $Y = w X + b$
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots.$
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - Approximate discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduct



Model	Equation	Interpretation
Level-Level Regression	$Y = \alpha + \beta X$	One unit change in X leads to β unit change in Y
Log-Linear Regression	$log\,(Y) = \alpha + \beta X$	One unit change in X leads to $100 * \beta$ percent change in Y
Linear-Log Regression	$Y = \alpha + \beta\, log(X)$	One percent change in X leads to $\beta/100$ unit change in Y
Log-Log Regression	$log\,(Y) = \alpha + \beta\, log(X)$	One percent change in X leads to β percent change in Y

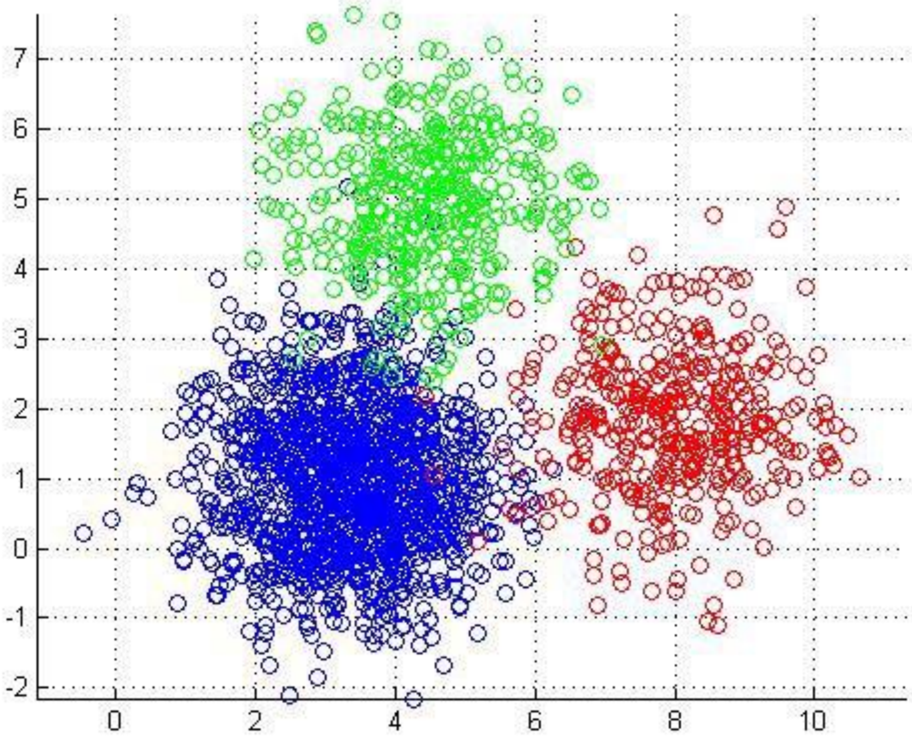
Non-parametric Method: Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Non-parametric Method: Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only



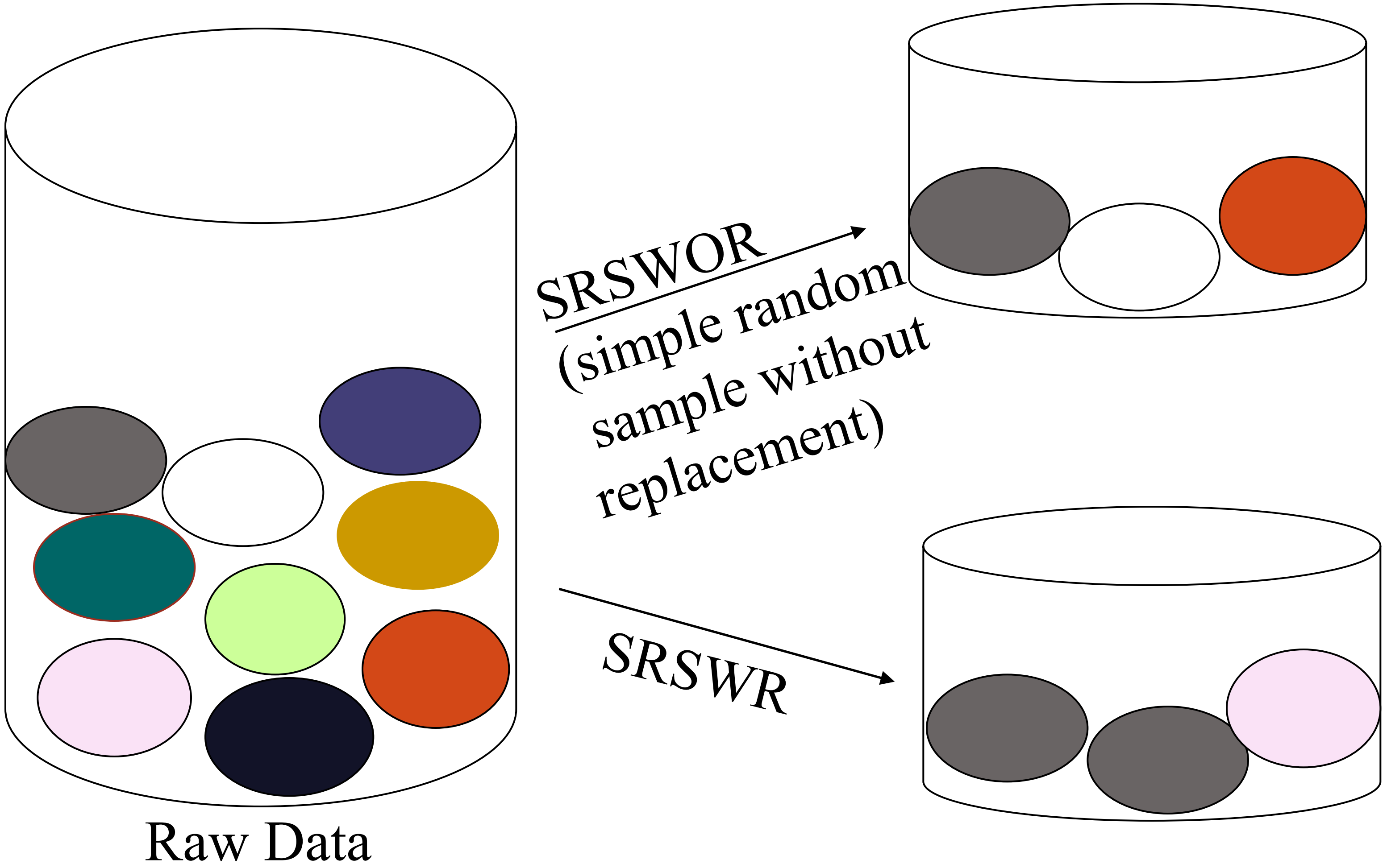
Non-parametric Method: Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:

Types of Sampling

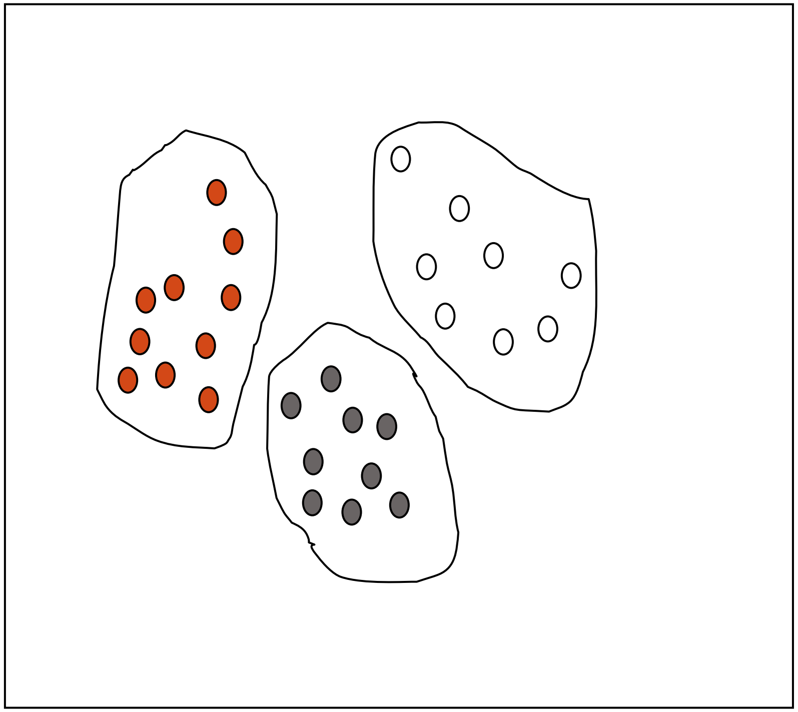
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)

Sampling: With or without Replacement

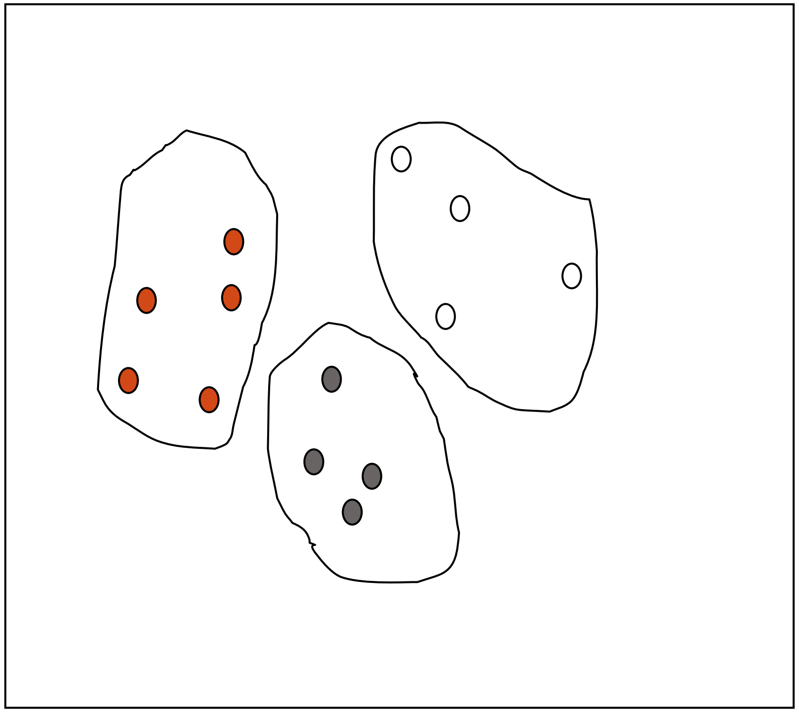


Sampling: Cluster or Stratified Sampling

Raw Data

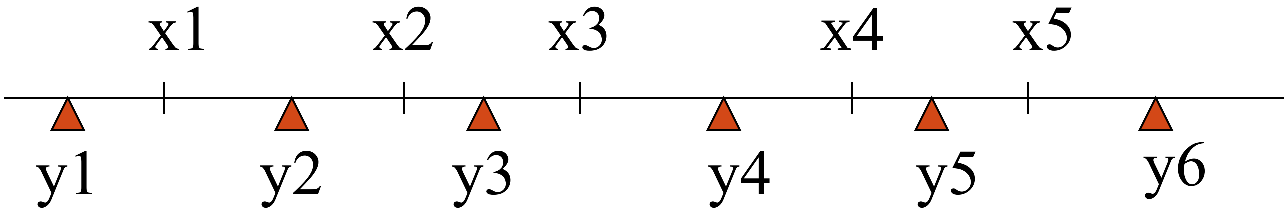


Cluster/Stratified Sample



Data Reduction 4: Discretization/Quantization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization/Quantization:
  - * divide the range of a continuous attribute into intervals



  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- Concept Hierarchies
  - reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Discretization and concept hierarchy generation for numeric data

- Hierarchical and recursive decomposition using:
  - Binning (data smoothing)
  - Histogram analysis (numerosity reduction)
  - Clustering analysis (numerosity reduction)
- Entropy-based discretization
- Segmentation by natural partitioning

Concept Hierarchy Generation

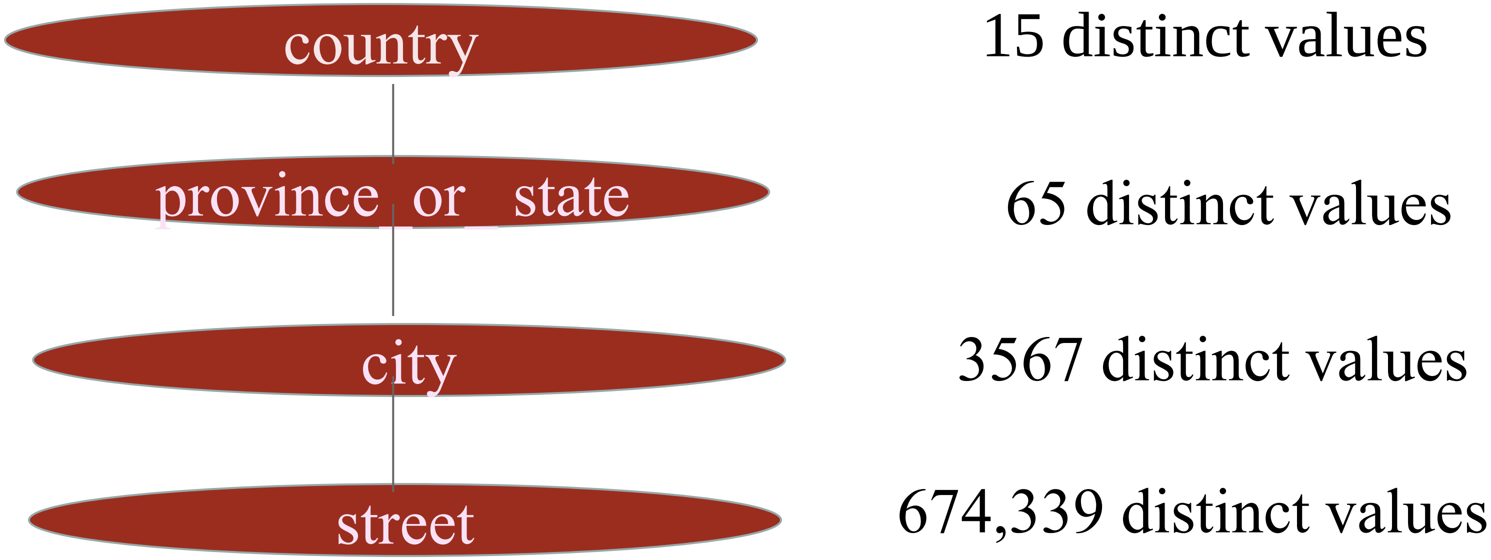
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - $street < city < state < country$
- Specification of a hierarchy for a set of values by explicit data grouping
  - **{Urbana, Champaign, Chicago} < Illinois**
- Specification of only a partial set of attributes
  - E.g., only $street < city$, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: $\{street, city, state, country\}$

Concept hierarchy generation w/o data semantics - Specification of a set of attributes

Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level of the hierarchy (limitations?)



country — 15 distinct values

province_or_ state — 65 distinct values

city — 3567 distinct values

street — 674,339 distinct values

SUMMARY

Data preprocessing

- **Data cleaning**
  - **Missing values**
    - Use the most probable value to fill in the missing value (and five other methods)
  - **Noisy data**
    - Binning; Regression; Clusttering
- **Data integration**
  - **Entity ID problem**
    - Metadata
  - **Redundancy**
    - Correlation analysis (Correlation coefficient, chi-square test)
- **Data trasnformation**
  - **Smoothing**
    - **Data cleaning**
  - **Aggregation**
    - **Data reduction**
  - **Generailization**
    - **Data reduction**
  - **Normalization**
    - Min-max; z-score; decimal scaling
  - **Attribute Construction**
- **Data reduction**
  - **Data cube aggregation**
    - Data cube store multidimensional aggregated information
  - **Attribute subset selection**
    - Stepwise forward selection; stepwise backward selection; combination; decision tree induction
  - **Dimensionality reduction**
    - Discrete wavelet trasnforms (DWT); Principle components analysis (PCA);
  - **Numerosity Reduction**
    - Regression and log-linear models; histograms; clustering; sampling
  - **Data discretization**
    - Binning; historgram analysis; entropy-based discretization;
Interval merging by chi-square analysis; cluster analysis; intuitive partitioning
  - **Concept hierarchy**
    - Concept hierarchy generation