



BASIC OF DATA ANALYTICS

Zareen Alamgir

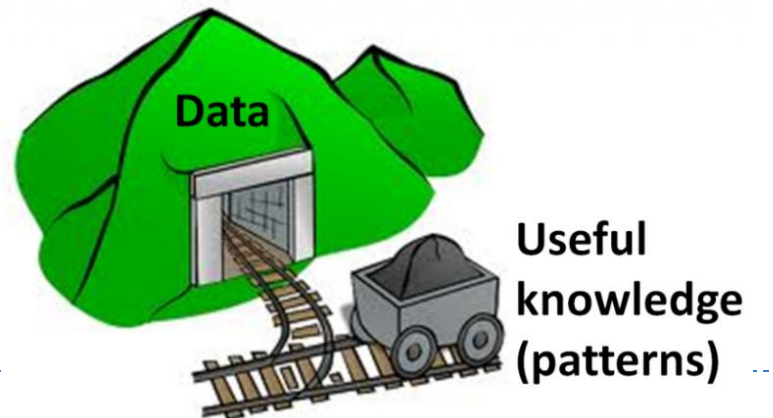
*Content obtained from many sources notably
Introduction to DM by pang and
DM concepts and techniques by Hans*

Data Mining

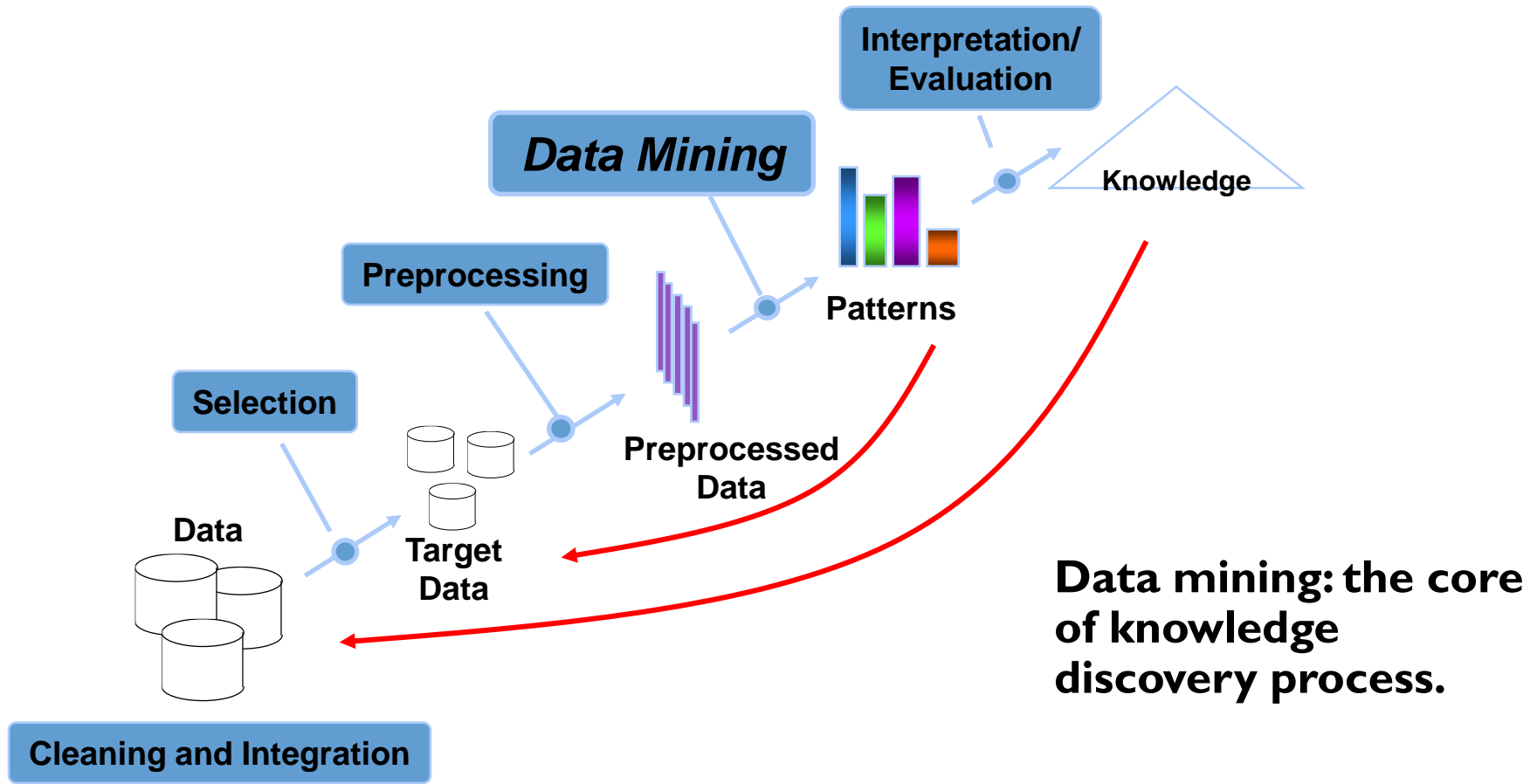


Given lots of data discover patterns that are:

- ▶ **Valid:** hold on new data with some certainty
- ▶ **Useful:** should be possible to act on the item
- ▶ **Unexpected:** non-obvious to the system
- ▶ **Understandable:** humans should be able to interpret the pattern



Data Mining: Process



The Data Analysis pipeline

Mining is not the only step in the analysis process



- ▶ **Preprocessing:** real data is noisy, incomplete & inconsistent
 - ▶ Data cleaning is required to make sense of the data
 - ▶ Techniques: Sampling, Dimensionality Reduction, Feature selection
- ▶ **Post-Processing:** Make the data actionable and useful to the user
 - ▶ Statistical analysis of importance
 - ▶ Visualization

What is Data?

Collection of data objects and their attributes

Attributes

Objects



<i>Tid</i>	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

Transaction Data

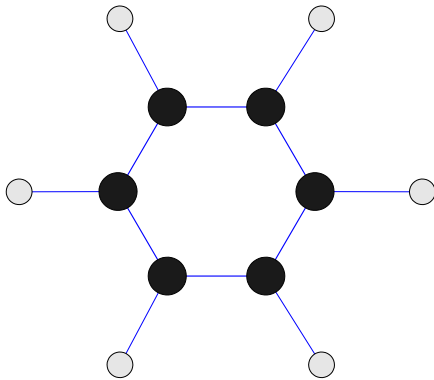
- ▶ A special type of record data, where
 - ▶ each record (transaction) involves a set of items.
 - ▶ For example, a grocery store transactions.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



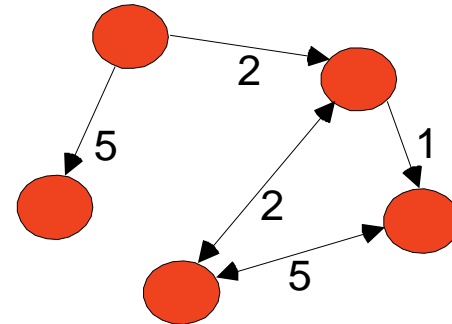
Graph Data

- ▶ World Wide Web
- ▶ Molecular Structures



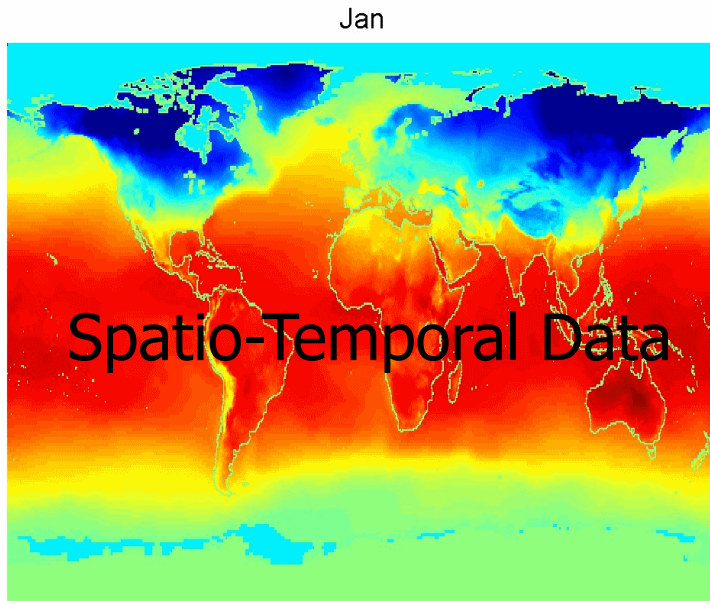
Benzene Molecule: C_6H_6

Generic graph and HTML Links



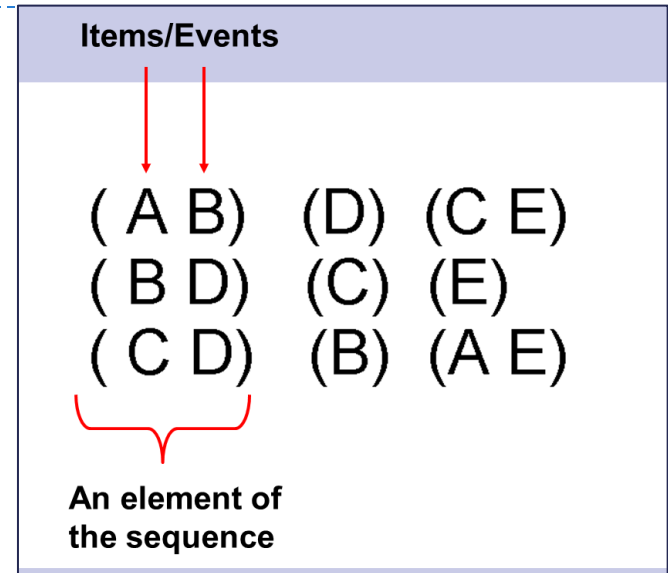
Ordered data

- ▶ **Spatial Data** physical location and shape of objects in a defined space
- ▶ **Temporal Data** changes or events occurring over time
- ▶ **Sequential Data**
- ▶ **Genetic Sequence Data**



Average Monthly Temperature of land and ocean

Sequences of transactions



Genomic sequence data

```
GGTTCGCGCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCCGACCAGG
```


Types of Attributes

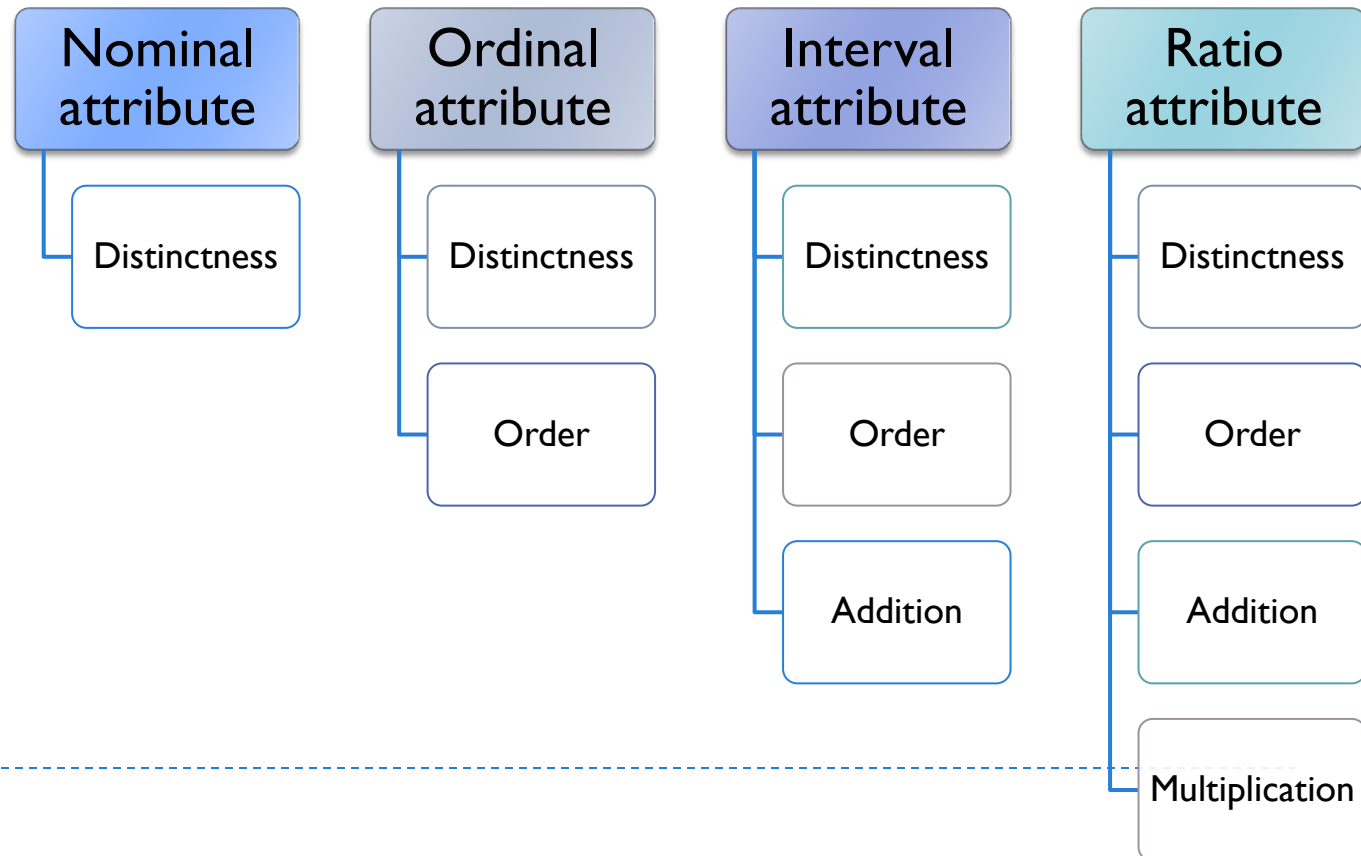
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- ▶ **Nominal:** refers to categorically discrete data
 - ▶ Examples: ID numbers, eye color, zip codes, name
- ▶ **Ordinal:** refers to quantities that have a natural ordering.
 - ▶ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- ▶ **Interval:** data is like ordinal where intervals between each value are equally split
 - ▶ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- ▶ **Ratio:** data is interval data with a natural zero point.
 - ▶ Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- Distinctness: $= \neq$
- Order: $< >$
- Addition: $+ -$
- Multiplication: $* /$

The type of an attribute depends on which of the properties it possesses

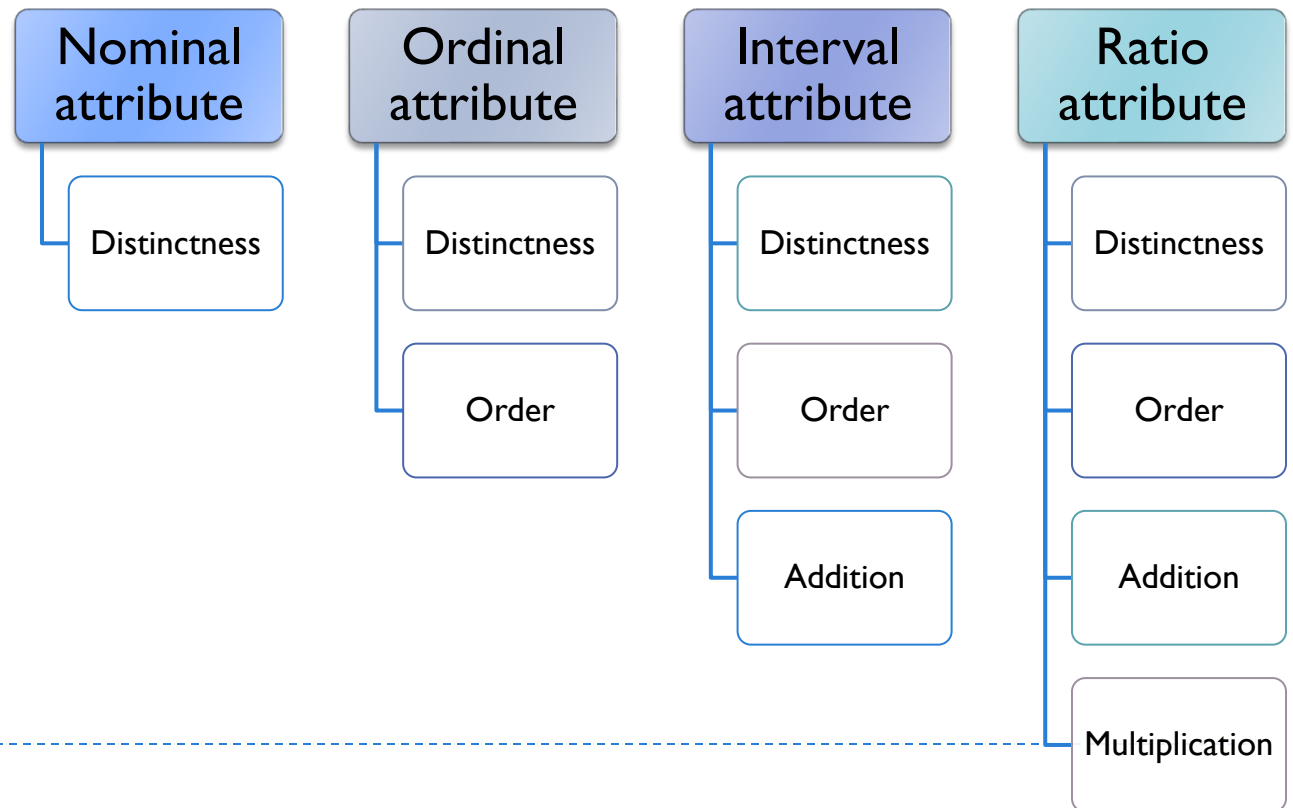


Attribute Type	Description	Examples	Operations
Nominal	Nominal attribute are just different names, i.e., They provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee IDs, eye color, gender	mode, entropy, correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation

Attribute Type	Description	Examples	Operations
Nominal	Nominal attribute are just different names, i.e., They provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee IDs, eye color, gender	mode, entropy, correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	median, percentiles, rank correlation
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, Standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests

Attribute Type	Description	Examples	Operations
Nominal	Nominal attribute are just different names, i.e., They provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee IDs, eye color, gender	mode, entropy, correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, Standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Order Number	Date	Merchant	# Items	Style	Price	Trans Fee
1001	5/11	Walmart	100	High Top	1000	20
1002	5/11	Costco	50	High Top	500	10
1003	5/11	Costco	50	Mid Top	500	10
1004	5/11	Target	100	Low Top	1000	20
1005	5/12	Walmart	50	High Top	500	10
1006	5/12	Walmart	50	Low Top	500	10
1007	5/13	Costco	50	Low Top	500	10
1008	5/13	Target	100	Low Top	1000	20
1009	5/14	Walmart	100	High Top	1000	20
1010	5/15	Walmart	50	Low Top	500	10



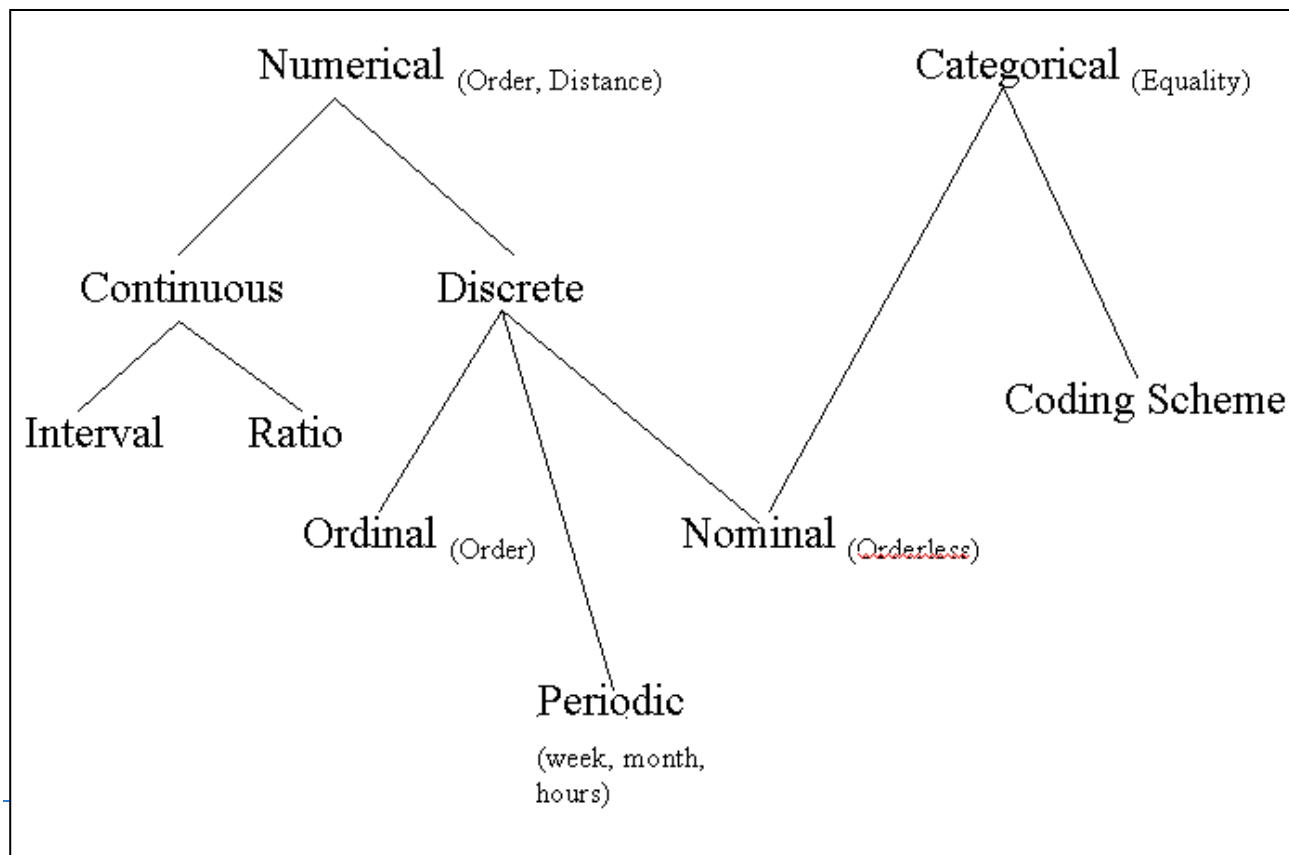
Data Types and Forms

➤ Attribute-value data:

A1	A2	...	An	C

➤ Data types

➤ numeric, categorical (see the hierarchy for its relationship)



Explore Data

- ▶ What do your records represent?
- ▶ What does each attribute mean?
- ▶ What type of attributes?
 - ▶ Categorical
 - ▶ Numerical
 - ▶ Discrete
 - ▶ Continuous
 - ▶ Binary – Asymmetric

<i>Tid</i>	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

Explore Data

- ▶ **Preliminary investigation** of the data to better understand its specific characteristics
 - ▶ It can help to answer some of the data mining questions
 - ▶ To help in selecting pre-processing tools
 - ▶ To help in selecting appropriate data mining algorithms

id	age	sex	region	salary	married	children	car
ID12101	48	FEMALE	INNER_CITY	17546	NO	1	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO
ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO
ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO
ID12107	22	MALE	RURAL	8877.07	NO		
ID12108	58	MALE	TOWN	24946.6	YES		
ID12109	37	FEMALE	SUBURBAN	25304.3	YES		
ID12110	54	MALE	TOWN	24212.1	YES		
ID12111	66	FEMALE	TOWN	59803.9	YES		
ID12112	52	FEMALE	INNER_CITY	26658.8	NO		
ID12113	44	FEMALE	TOWN	15735.8	YES		

**Visualization tools
are important**

Bar graphs, box plots,
scatter plots

Explore Data

- ▶ Things to look at
 - ▶ Class balance
 - ▶ Dispersion of data attribute values
 - ▶ Skewness, outliers, missing values
 - ▶ Attributes that vary together

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

<i>Tid</i>	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

Useful Statistics

Discrete attributes

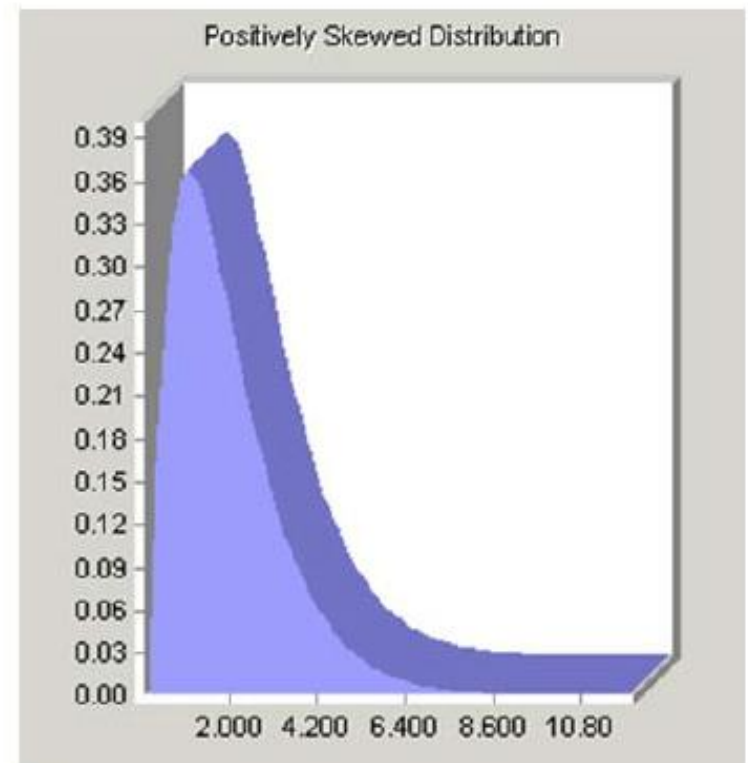
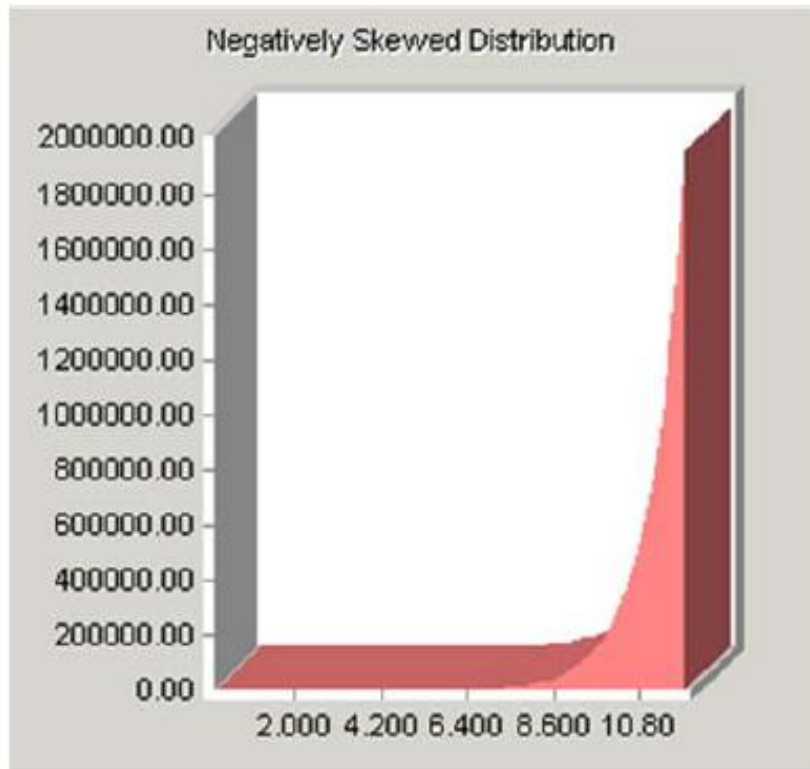
- Frequency of each value
- Mode = value with highest frequency

Continuous attributes

- Range of values, i.e. min and max
- **Mean (average)**
 - Sensitive to outliers
- **Median**
 - Better indication of the "middle" of a set of values in a skewed distribution
- **Skewed distribution**
 - mean and median are quite different



Skewed Distributions of Attribute Values



Dispersion of Data

- ▶ How do the values of an attribute spread?
- ▶ Variance
 - ▶ Variance is sensitive to outliers

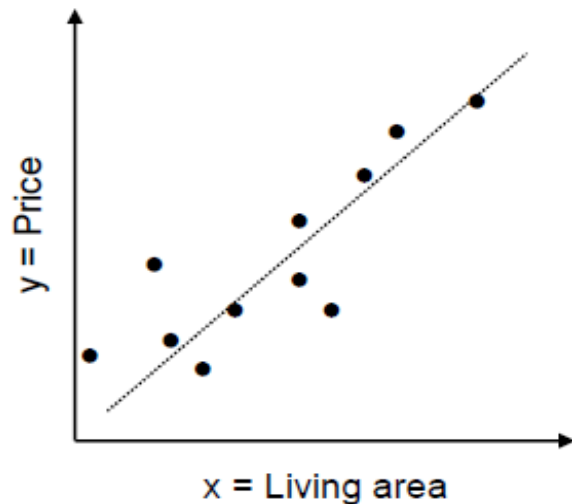
$$variance(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▶ What if the distribution of values is multimodal, i.e. data has several *bumps*?
- ▶ Visualization tools are useful



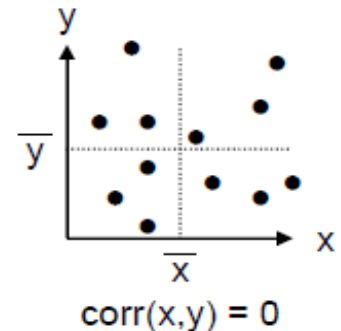
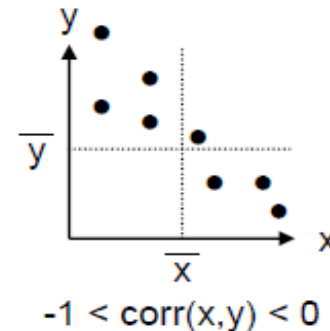
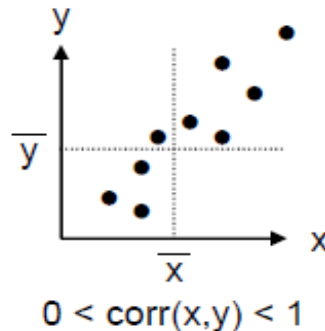
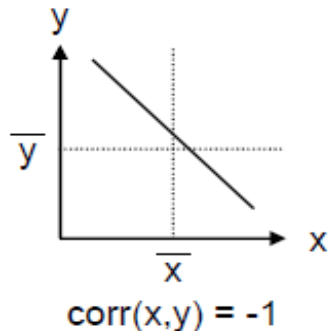
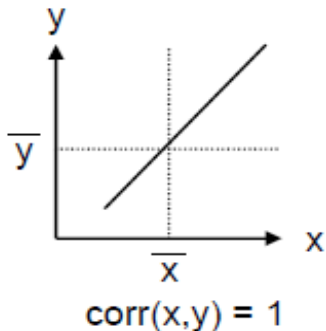
Attributes that Vary Together

There is a **linear correlation** between x and y.



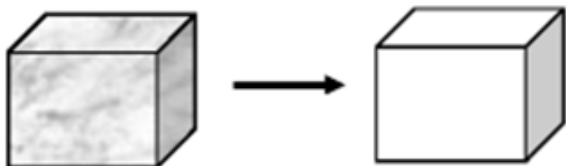
Correlation is a measure that describe how two attributes vary together

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

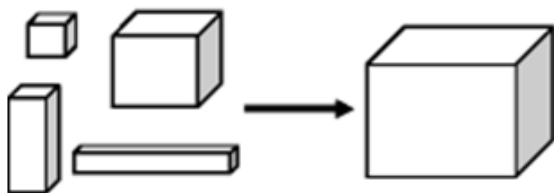


Forms of data preprocessing

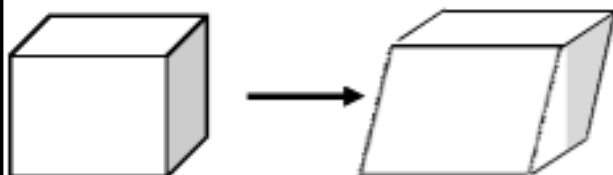
Data cleaning



Data integration

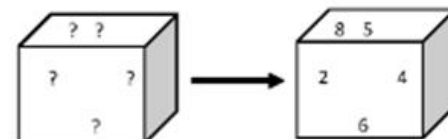


Data transformation

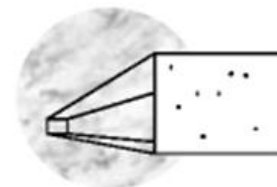


- Fill in missing values
- Smooth noisy data
- Remove outliers
- Resolve inconsistencies

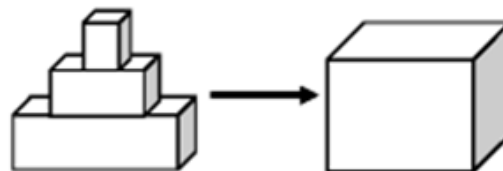
Missing values imputation



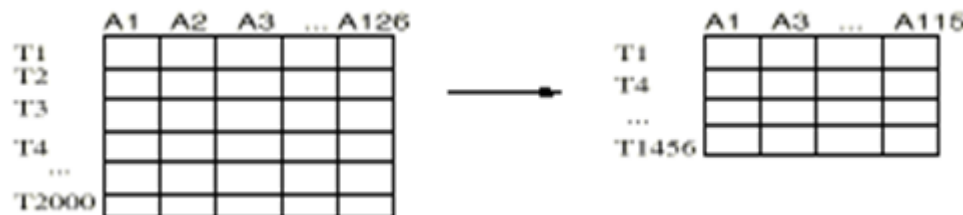
Noise identification



Data normalization



Data Reduction



Missing Values

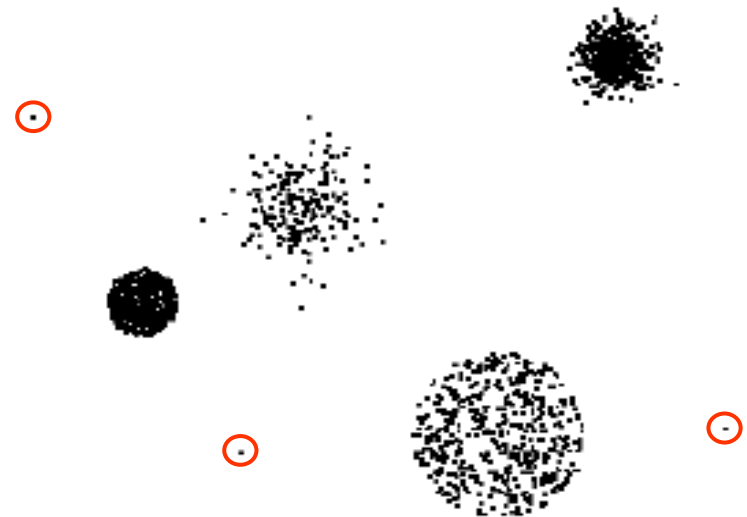
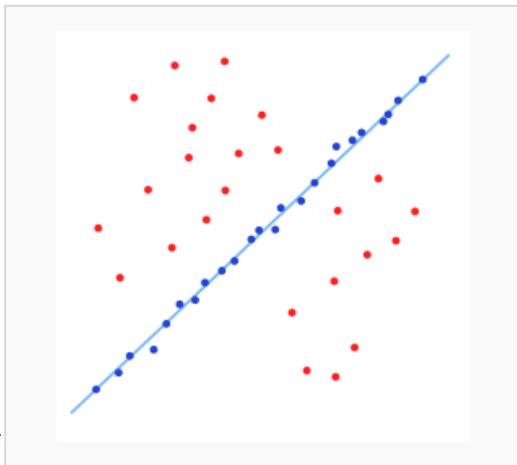
- ▶ **Handling missing values**
 - ▶ Eliminate Data Objects
 - ▶ Estimate Missing Values
 - ▶ Ignore the Missing Value During Analysis
 - ▶ Replace with all possible values (weighted by their probabilities)



Outliers



- ▶ Outliers are data objects with characteristics that are **considerably different** than most of the other data objects in the data set
- ▶ Can help to
 - ▶ detect new phenomenon or
 - ▶ discover unusual behavior in data
 - ▶ detect problems



How to Handle Noisy Data?

▶ Binning method

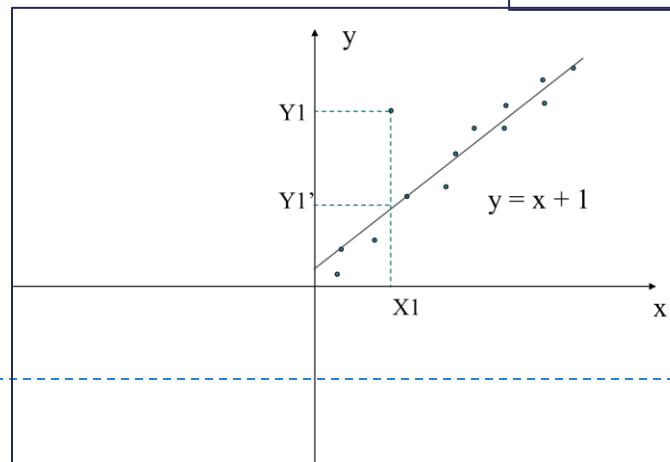
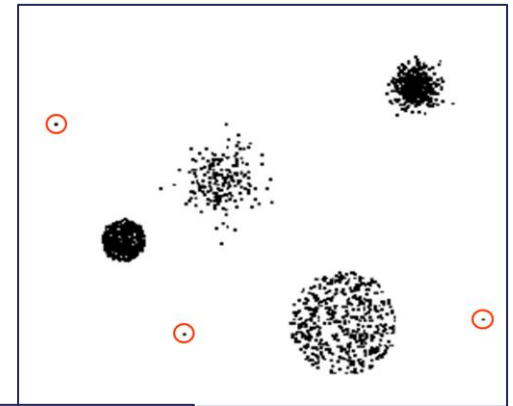
- ▶ first sort data and partition into (equi-depth) bins
- ▶ then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

▶ Clustering

- ▶ detect and remove outliers

▶ Regression

- ▶ smooth by fitting the data into regression functions

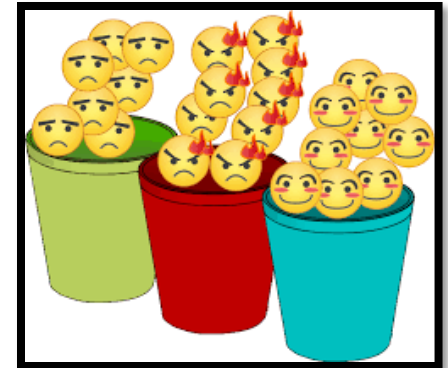


Data Discretization

- ▶ Divide the range of a continuous attribute into intervals
- ▶ Interval labels can be used to replace actual data values.

- ▶ **Advantages**

- ▶ Discretized continuous attribute
 - ▶ Data Reduction – help reduce data size
 - ▶ Data Smoothing (handling noise)
- ▶ Some data mining algorithms only work with discrete attributes
 - ▶ E.g. Apriori for Association Rule Mining



Binning (Equal-width)

- ▶ Equal-width (distance) partitioning
 - ▶ Divide the attribute values x into k equally sized bins
 - ▶ If $x_{\min} \leq x \leq x_{\max}$ then the bin width δ is given by

$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

Attribute values (for an attribute age):

0, 4, 12, 16, 16, 18, 24, 26, 28

Equi-width binning – for bin width of 10:

Bin 1: 0, 4 $[-, 10)$ bin

Bin 2: 12, 16, 16, 18 $[10, 20)$ bin

Bin 3: 24, 26, 28 $[20, +)$ bin

– denote negative infinity, + positive infinity

Binning (Equal-width)

- ▶ Equal-width (distance) partitioning
 - ▶ Divide the attribute values x into k equally sized bins

The best number of bins k is determined experimentally

- ▶ **Disadvantages:**
 - ▶ outliers may dominate presentation
 - ▶ Skewed data is not handled well.



Binning (Equal-frequency)

- ▶ Equal-depth (frequency) partitioning:
 - ▶ An equal number of values are placed in each of the **k** bins.
 - ▶ Good data scaling
- ▶ **Disadvantage:**
 - ▶ Many occurrences of the same continuous value could cause the values to be assigned into different bins
 - ▶ Managing categorical attributes can be tricky.

Attribute values (for an attribute age):

0, 4, 12, 16, 16, 18, 24, 26, 28

Equi-frequency binning – for bin density of 3:

Bin 1: 0, 4, 12

[-, 14) bin

$16-12 = 2 \Rightarrow 12+2=14$

Bin 2: 16, 16, 18

[14, 21) bin

Bin 3: 24, 26, 28

[21, +] bin

Binning Example

- Attribute values (for an attribute age):

► 0, 4, 12, 16, 18, 24, 26, 28 **Sorted**

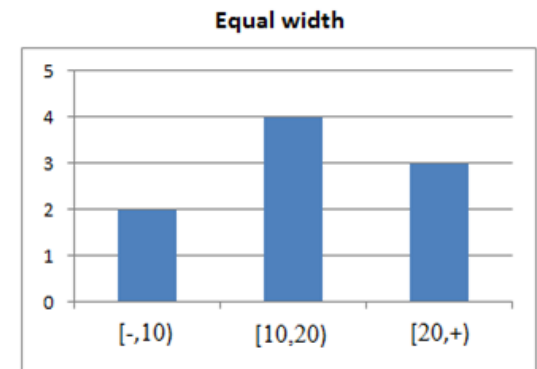
- ▶ **Equi-width binning** – for bin width of 10:

- Bin 1: 0, 4 [-, 10) bin

- Bin 2: 12, 16, 16, 18 [10,20) bin

- Bin 3: 24, 26, 28 [20,+) bin

- ▶ – denote negative infinity, + positive infinity

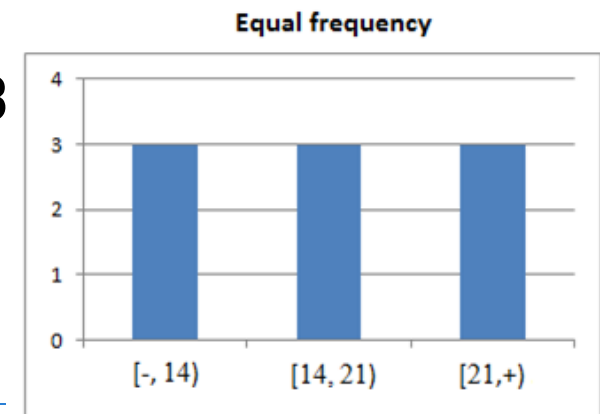


- ▶ **Equi-frequency binning** – for bin density of 3

- Bin 1: 0, 4, 12 [-, 14) bin

- Bin 2: 16, 16, 18 [14, 21) bin

► Bin 3: 24, 26, 28 [21, +] bin



Binning Methods for Data Smoothing

* Sorted data for price: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

* Partition into Equi-depth bins:

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 9

Bin 2: 23, 23, 23, 23 22.75

Bin 3: 29, 29, 29, 29 29.25

Smoothing by bin boundaries:

Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34



Data Transformation

Transform or consolidate data into forms appropriate for mining



person_name	Salary	Year_of_experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

Normalization: scaled to fall within a small, specified range

Data Transformation: Normalization

- An attribute values are scaled to fall within a small, specified range , such as 0.0 to 1.0

- ▶ **Min-Max normalization**

- ▶ performs a linear transformation on the original data.

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- ▶ **Example:** Let min and max values for the attribute *income* are \$12,000 and \$98,000, respectively.
- ▶ Map *income* to the range [0.0;1.0].

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716.$$



Data Transformation: Normalization

- ▶ **z-score normalization(or zero-mean normalization)**

- ▶ An attribute A, values are normalized based on the mean and standard deviation of A.

$$v' = \frac{v - mean_A}{stand_dev_A}$$

- ▶ **Example:** Let mean= 54,000 and standard deviation=16,000 for the attribute *income*
- ▶ With z-score normalization, a value of \$73,600 for *income* is transformed to

$$\frac{73,600 - 54,000}{16,000} = 1.225.$$



Data Reduction

- ▶ 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**
 - ▶ Obtains a reduced representation of the data set that is much smaller in volume
 - ▶ but produces the same (or almost the same) analytical results



Data Reduction Strategies

- ▶ **Dimensionality reduction**
- ▶ **Numerosity reduction**
 - ▶ data is replaced or estimated by alternative smaller data representations
 - ▶ Sampling
 - ▶ Clustering
- ▶ **Discretization and concept hierarchy generation**
 - ▶ replace raw data values for attributes by ranges or higher conceptual levels



Dimensionality Reduction

▶ Purpose

- ▶ Avoid curse of dimensionality
- ▶ Reduce amount of time and memory required by data mining algorithms
- ▶ Allow data to be more easily visualized
- ▶ May help to eliminate irrelevant features or reduce noise

▶ Techniques

- ▶ Principle Component Analysis
- ▶ Singular Value Decomposition
- ▶ Auto encoders
- ▶ Others: supervised and non-linear techniques



Feature selection

- ▶ Another way to reduce dimensionality of data
- ▶ Feature selection (i.e., attribute subset selection):
 - ▶ Select a minimum set of features
 - ▶ such that the **probability distribution** of different classes given the values of the **selected features** is as close to the **original distribution** given the values of all features



Feature Subset Selection

▶ **Redundant features**

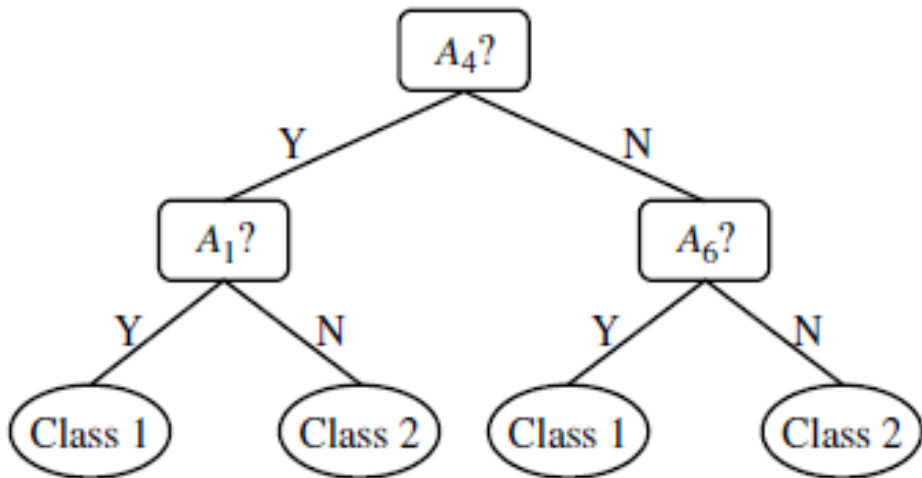
- ▶ duplicate much or all of the information contained in one or more other attributes
- ▶ Example: purchase price of a product and the amount of sales tax paid

▶ **Irrelevant features**

- ▶ contain no information that is useful for the data mining task at hand
- ▶ Example: students' ID is often irrelevant to the task of predicting students' GPA

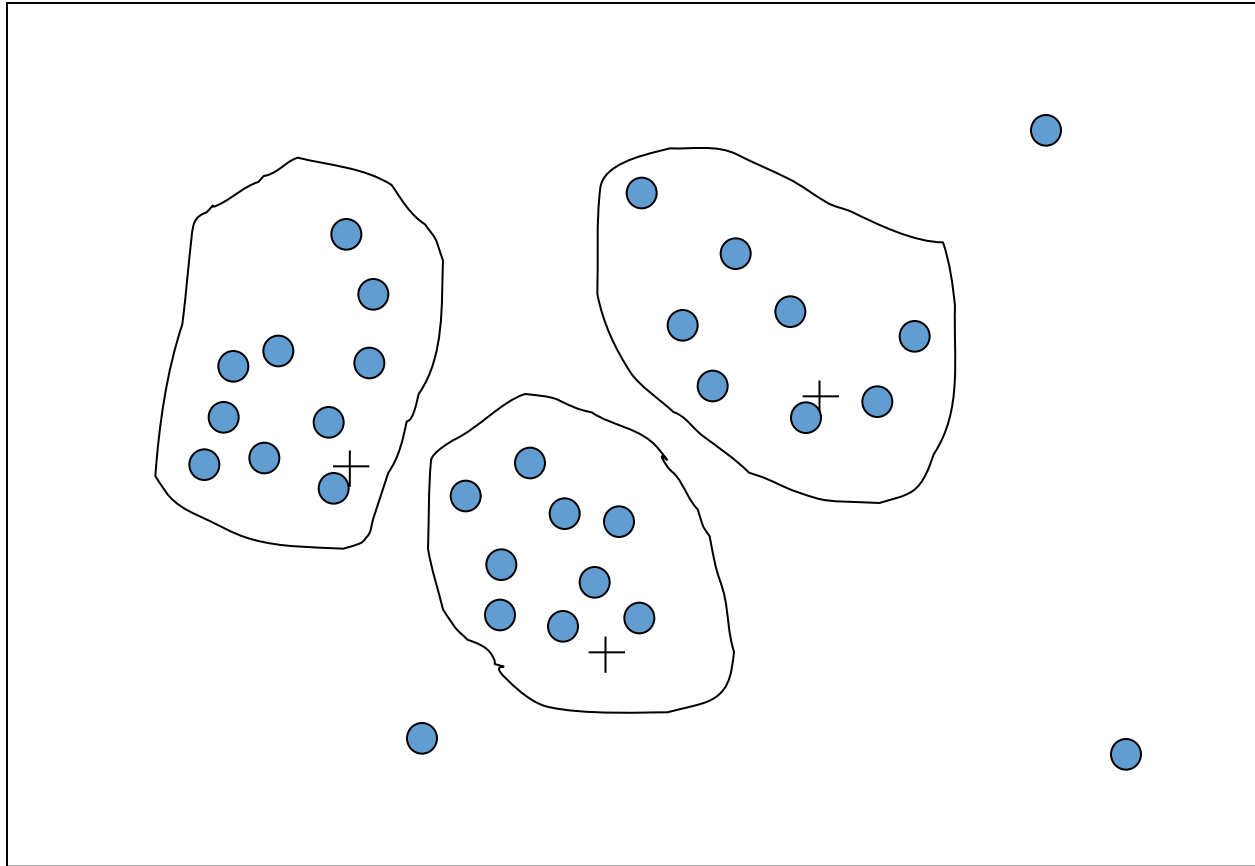


Feature Subset Selection

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1("Class 1") A1 -- N --> C2_1("Class 2") A6 -- Y --> C1_2("Class 1") A6 -- N --> C2_2("Class 2") </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>



Numerosity reduction - Cluster Analysis



**Partition data into
clusters, and store
cluster
representation only**

**Can be very effective
if data is in form of
clusters**



Numerosity reduction - Sampling

Statisticians sample because **obtaining** the entire set of data of interest is too expensive or time consuming.

Example: What is the average height of a person in Pakistan?
We cannot measure the height of everybody

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

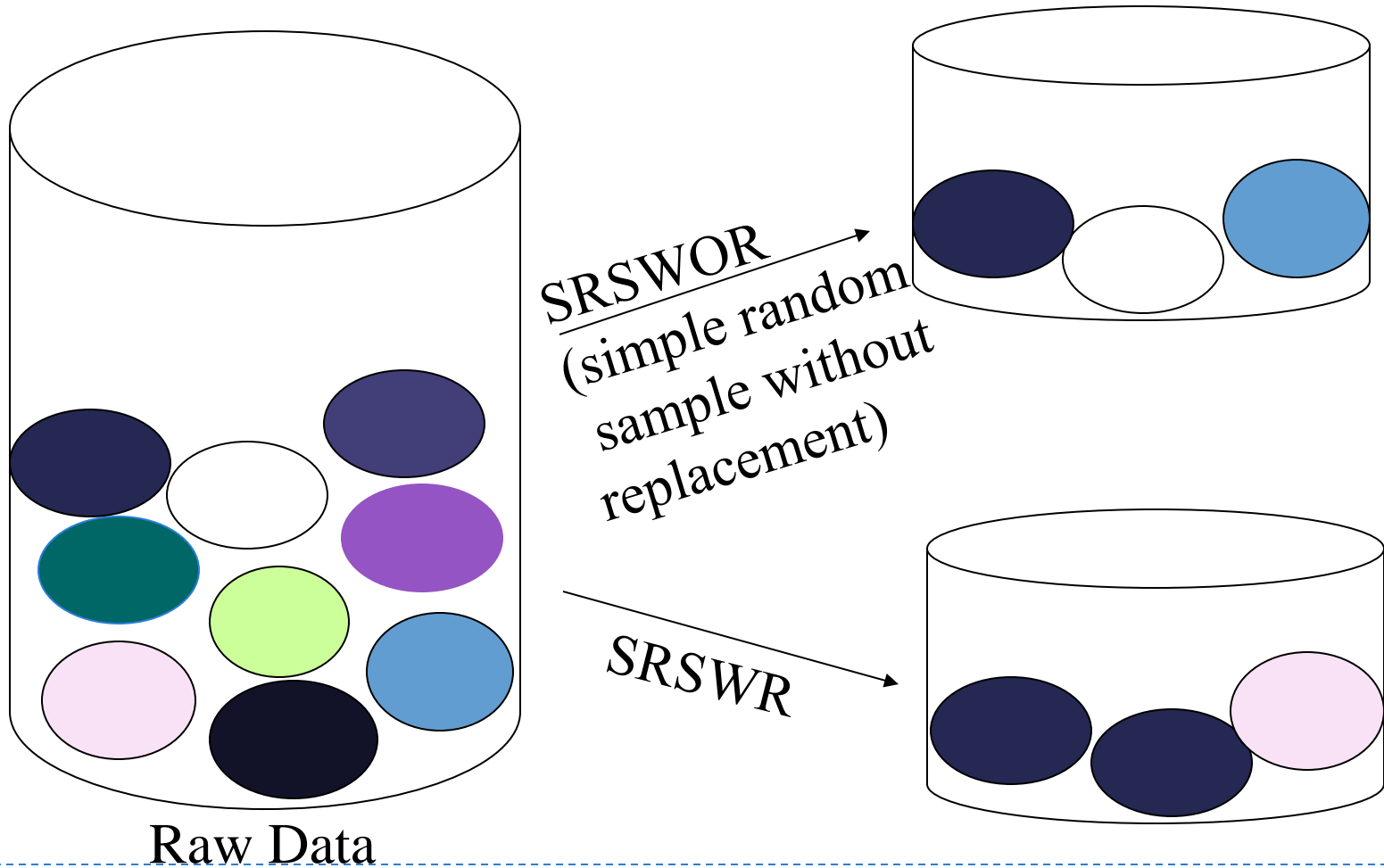
Example: We have 1M documents. How many has at least 100 words in common?

- Computing number of common words for all pairs requires 10^{12} comparisons

Example: What fraction of tweets in a year contain the word "Lahore"?

- 300M tweets per day, if 100 characters on average, 86.5TB to store all tweets

Sampling



Types of Sampling

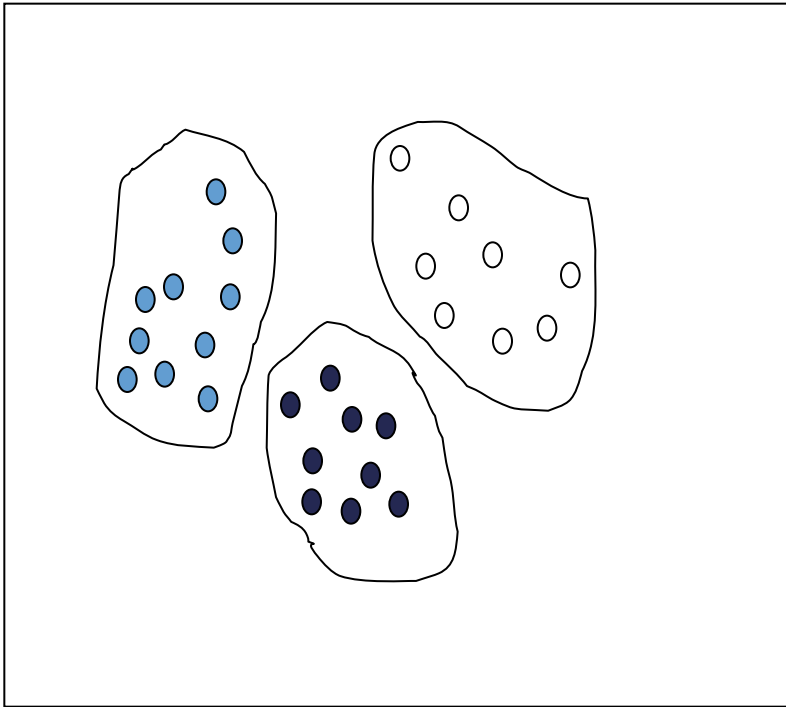
▶ **Stratified sampling**

- ▶ Split the data into several **groups**; then draw random samples from each group.
- ▶ Ensures that both groups are represented.
- ▶ **Example** Find difference between legitimate and fraudulent credit card transactions.
- ▶ **0.1%** of transactions are fraudulent. What happens if we select **1000** transactions at random?
 - ▶ We get **1** fraudulent transaction (in expectation). Not enough to draw any conclusions.
 - ▶ Solution: sample **1000** legitimate and **1000** fraudulent transactions

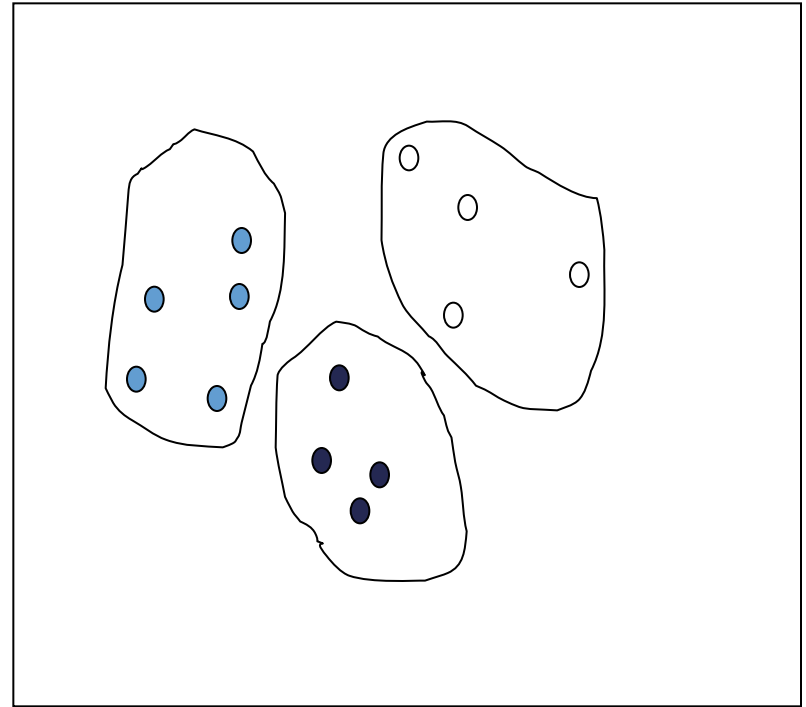


Sampling

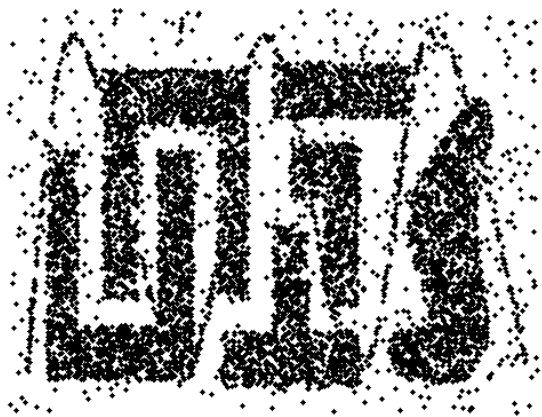
Raw Data



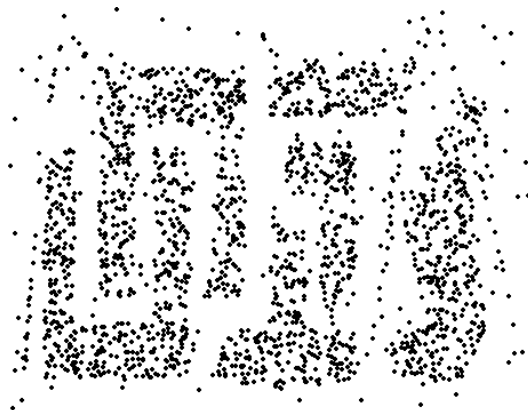
Cluster/Stratified Sample



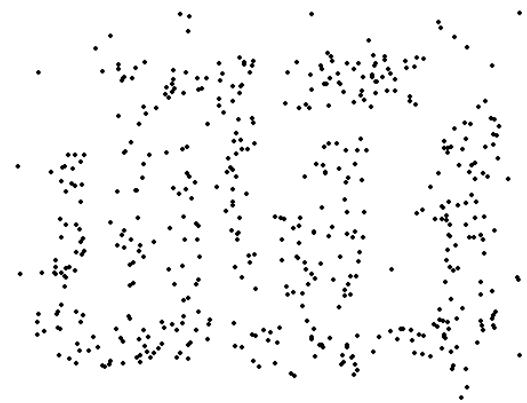
Sample Size



8000 points
Points



2000 Points



500



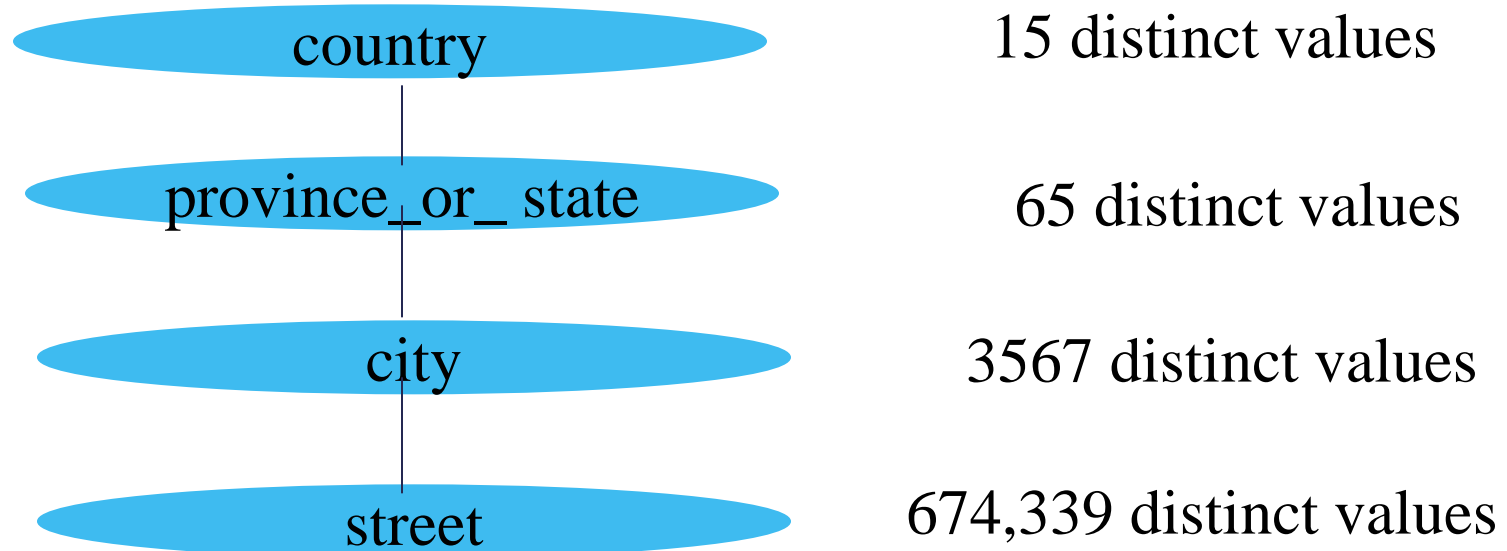
Concept hierarchy

- ▶ **Concept hierarchy**
 - ▶ Reduce the data by replacing low level concepts by higher level concepts
 - ▶ Replace numeric values for the attribute age by higher level concepts such as
 - ▶ **young, middle-aged, or senior**



Automatic Concept Hierarchy Generation

- ▶ Some concept hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the given data set
 - ▶ The attribute with the most distinct values is placed at the lowest level of the hierarchy



What is data exploration?

A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
 - ▶ Humans have a well-developed ability to analyze large amounts of information presented visually
 - ▶ Can help detect general patterns and trends
 - ▶ Can help detect outliers and unusual patterns

Visualization of data is one of the most powerful and appealing techniques for data exploration.

Arrangement

- ▶ Is the placement of visual elements within a display
- ▶ Can make a large difference in how easy it is to understand the data
- ▶ Example:

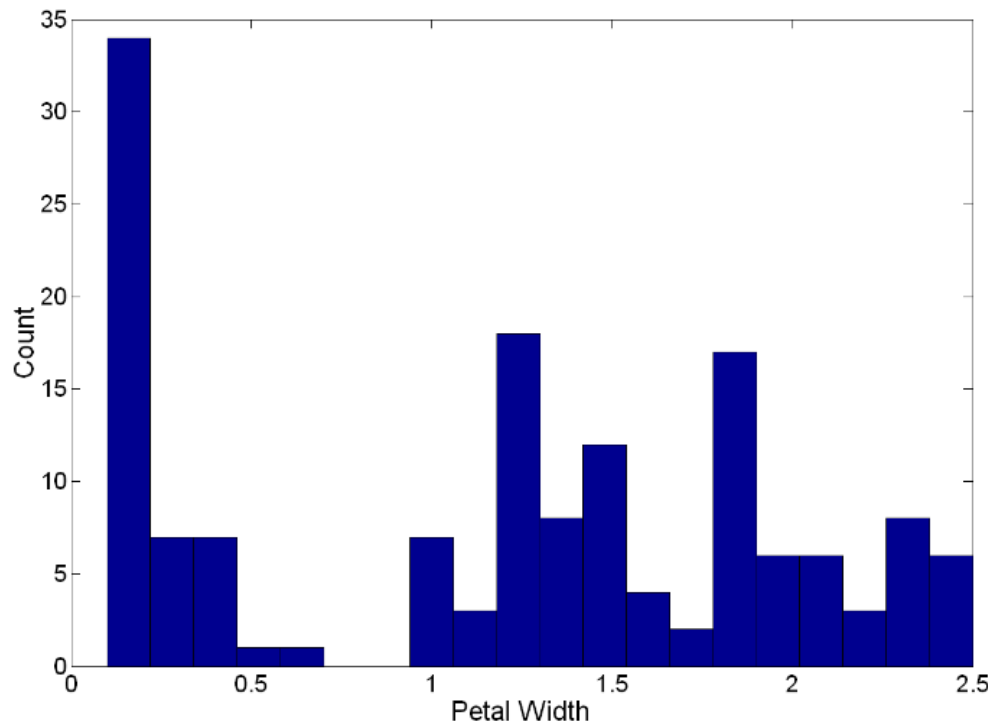
	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0



Visualization Techniques: Histograms

▶ Histogram

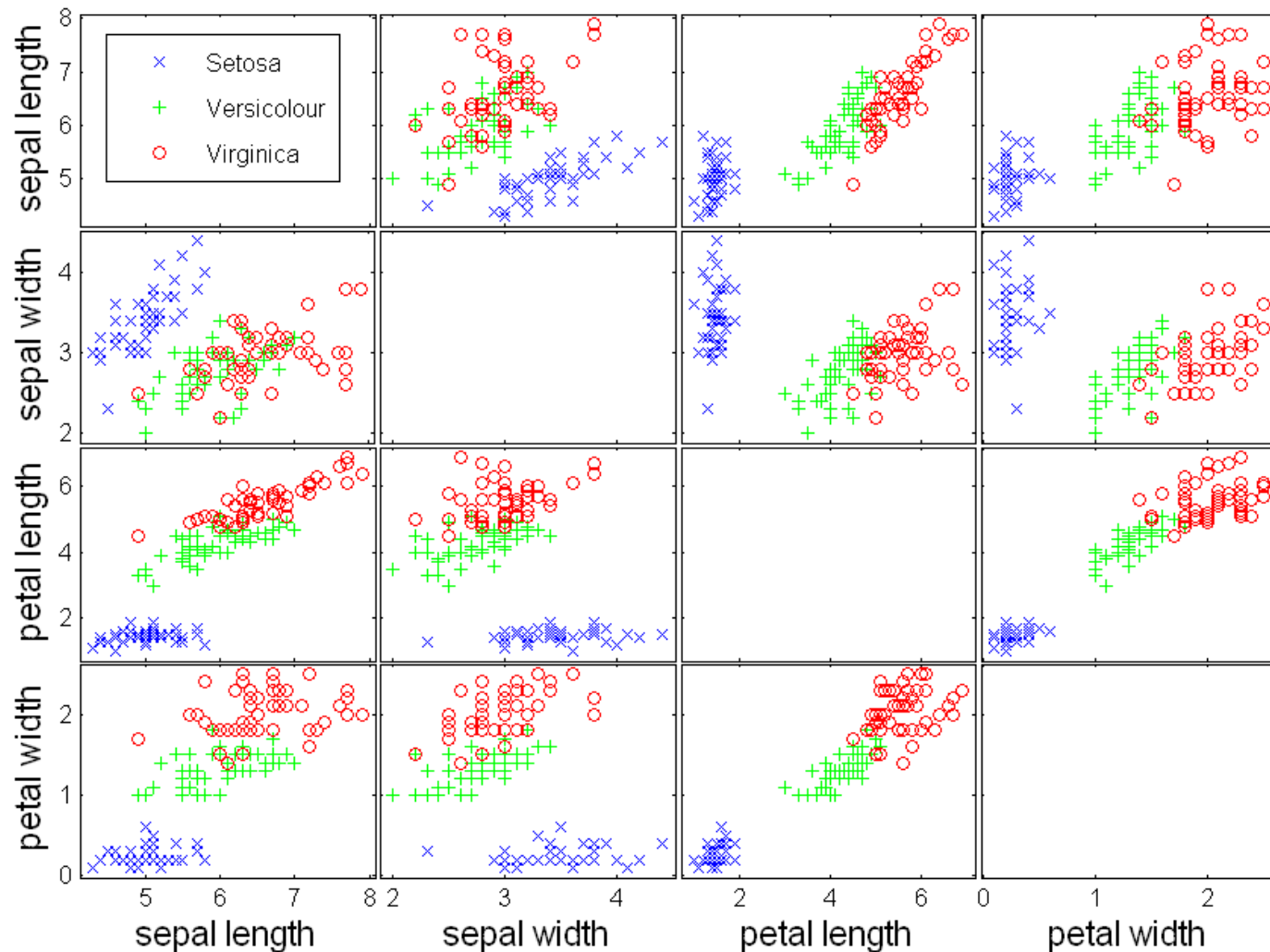
- ▶ Usually shows the distribution of values of a single variable
- ▶ Divide the values into bins and show a bar plot of the number of objects in each bin.
- ▶ The height of each bar indicates the number of objects



Example: Petal Width
(10 and 20 bins, respectively)

Visualization Techniques: Scatter Plots

Scatter Plot Array of Iris Attributes



Iris Sample Data Set

- ▶ Many of the exploratory data techniques are illustrated with the Iris Plant data set.
 - ▶ Can be obtained from the UCI Machine Learning Repository
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
- ▶ Three flower types (classes):
 - ▶ Setosa
 - ▶ Virginica
 - ▶ Versicolour
- ▶ Four (non-class) attributes
 - ▶ Sepal width and length
 - ▶ Petal width and length

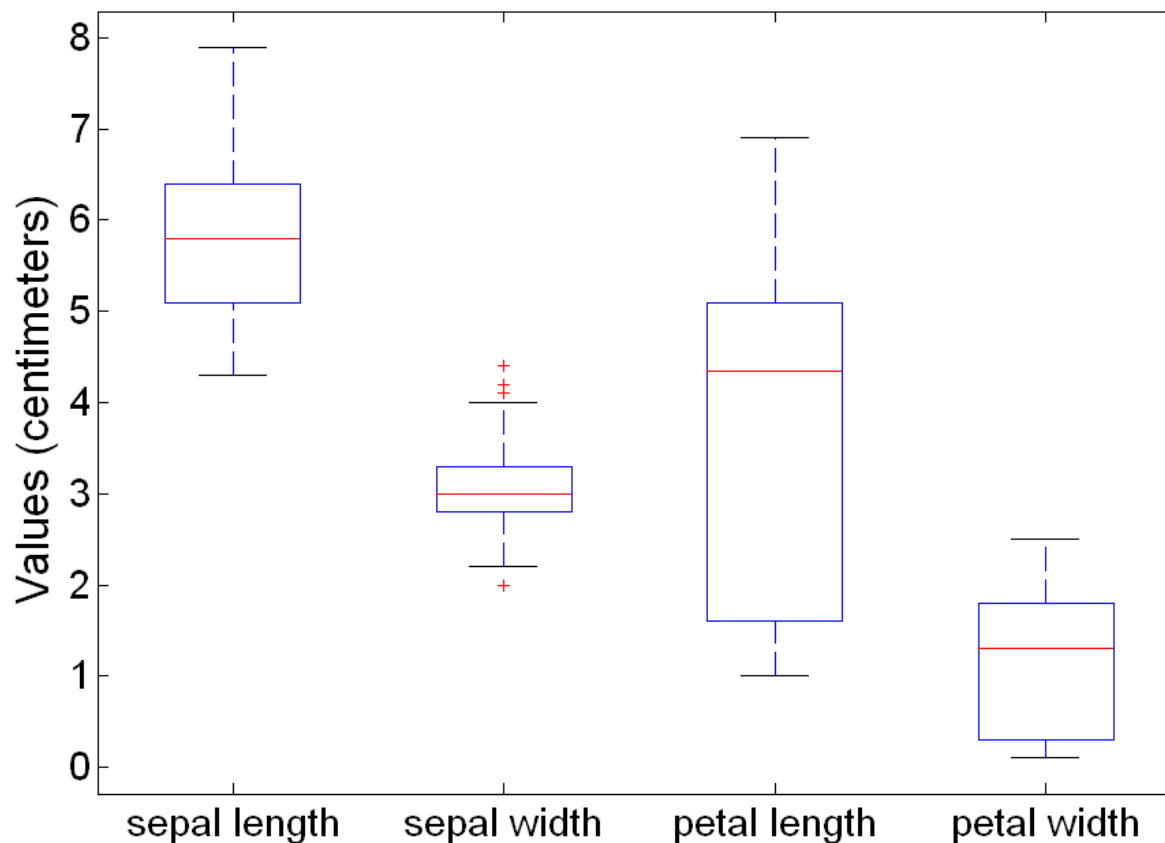


Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.



Example of Box Plots

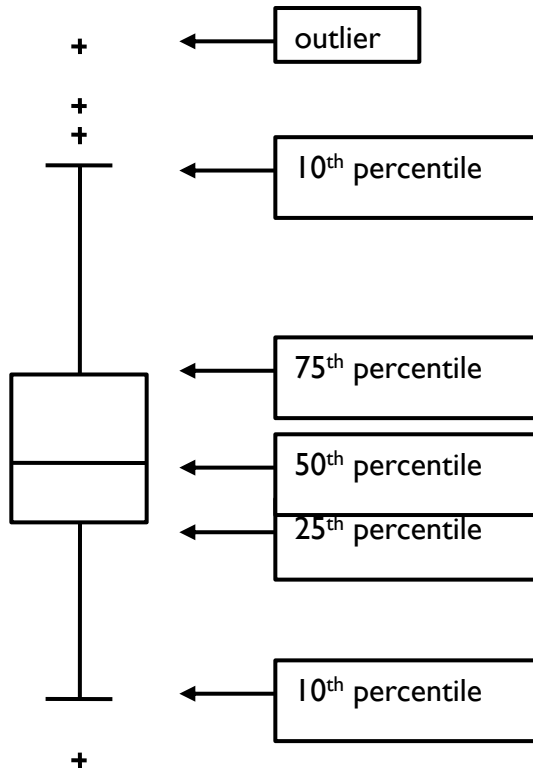
- ▶ Box plots can be used to compare attributes



Visualization Techniques: Box Plots

Box Plots

- ▶ Boxplots are a popular way of visualizing a distribution.
- ▶ Following figure shows the basic part of a box plot



A box plot provides information about an attribute

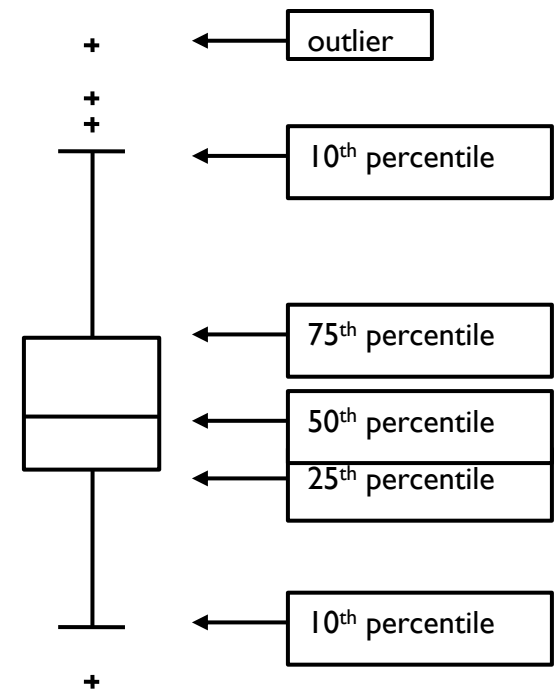
- range
- median
- normality of the distribution
- skew of the distribution
- plot extreme cases within the sample

For continuous data, the notion of a percentile is more useful.

For instance, the 50th percentile is the value such that 50% of all values of x are less than it .

Box Plots Example

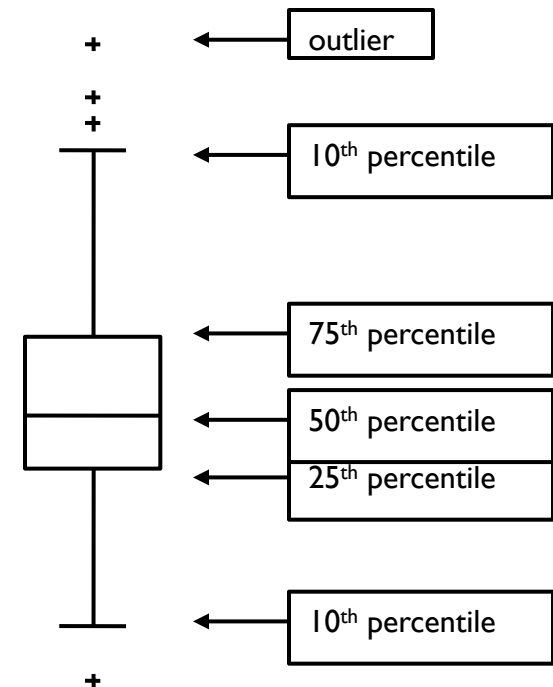
- ▶ A boxplot incorporates the five-number
(*Minimum, Q1, Median, Q3, Maximum*)
- ▶ Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, *IQR*.
- ▶ The **median** is marked by a line within the box
- ▶ **Two lines (called *whiskers*)** outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.



Box Plots Example

When dealing with a moderate number of observations, it is worthwhile to plot **potential outliers individually**.

- ▶ To do this in a boxplot, *the whiskers are extended to the extreme low and high observations only if these values are less than $1.5 \times IQR$ beyond the quartiles.*
- ▶ Otherwise, the whiskers terminate at the most extreme observations occurring within $1.5 \times IQR$ of the quartiles. The remaining cases are plotted individually.



Box Plots Example

Attribute values: 6 47 49 15 42 41 7 39 43 40 36

Sorted: 6 7 15 36 39 40 41 42 43 47 49



Box Plots Example

Attribute values: 6 47 49 15 42 41 7 39 43 40 36

Sorted: 6 7 15 36 39 40 41 42 43 47 49

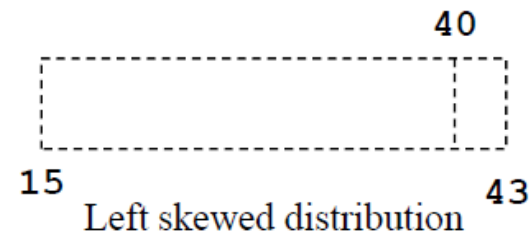
$Q_1 = 15$ lower quartile

$Q_2 = \text{median} = 40$

(*mean* = 33.18)

$Q_3 = 43$ upper quartile

$Q_3 - Q_1 = 28$ interquartile range



Available in **WEKA**

- **Filters**

InterquartileRange