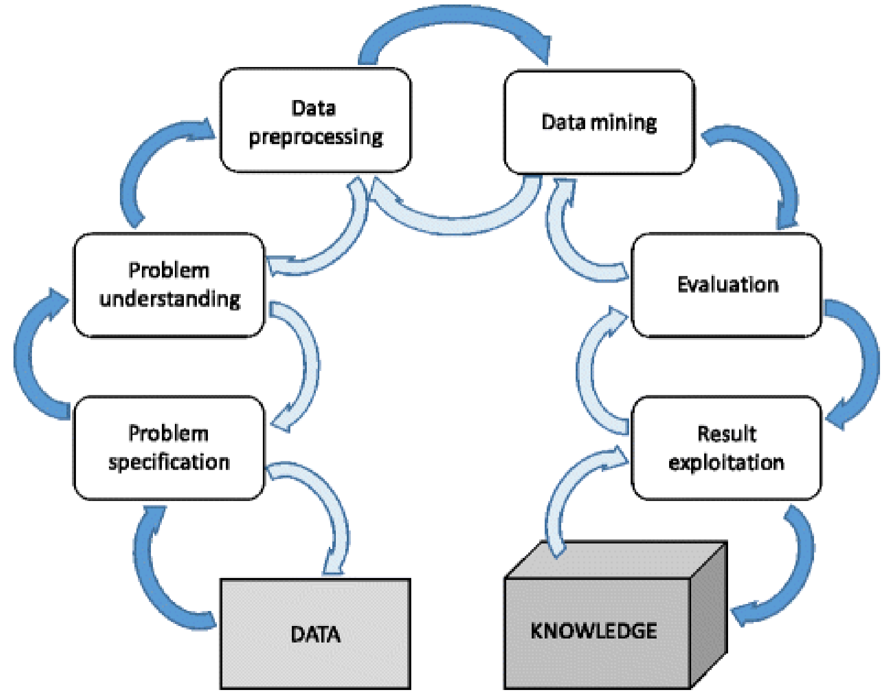


Data Pre-processing

Zareen Alamgir

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

The Data Analysis pipeline

Mining is not the only step in the analysis process

Data Pre-processing → Data Mining → Result Post-processing

- 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

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**

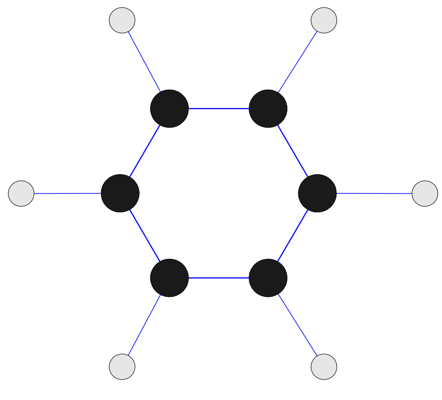
Transaction Data

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

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

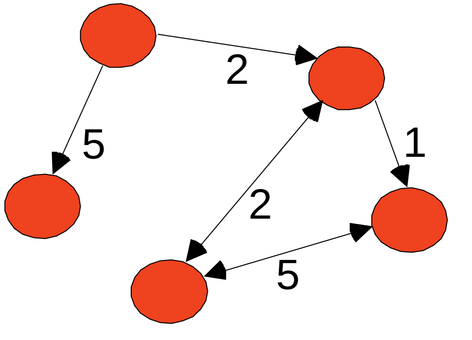
Graph Data

- World Wide Web
- Molecular Structures



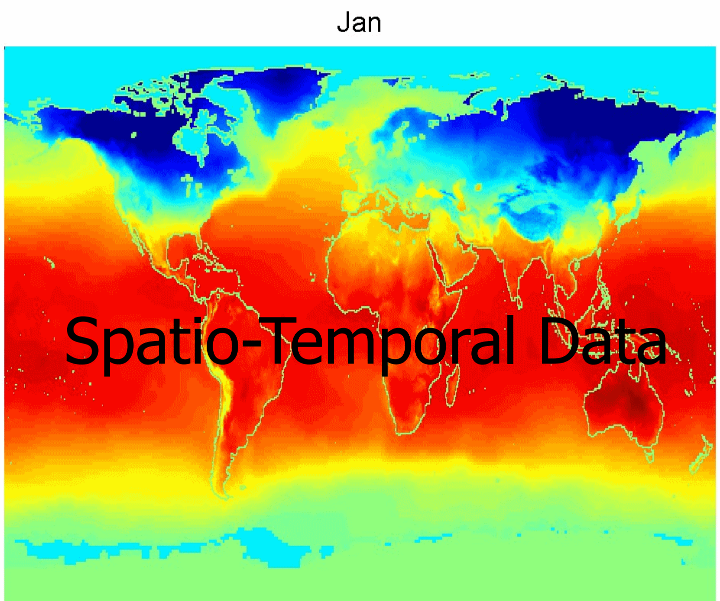
Benzene Molecule: C_6H_6

Generic graph and HTML Links



Ordered data

- **Spatial Data**
- **Temporal Data**
- **Sequential Data**
- **Genetic Sequence Data**



Average Monthly Temperature of land and ocean

Sequences of transactions

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

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: * /

- Nominal attribute
  - Distinctness
- Ordinal attribute
  - Distinctness
  - Order
- Interval attribute
  - Distinctness
  - Order
  - Addition
- Ratio attribute
  - Distinctness
  - 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

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, {*good, better, best*}, 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, {*good, better, best*}, 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, *t* and *F* 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, {*good, better, best*}, 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, *t* and *F* 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

Interval vs Ratio

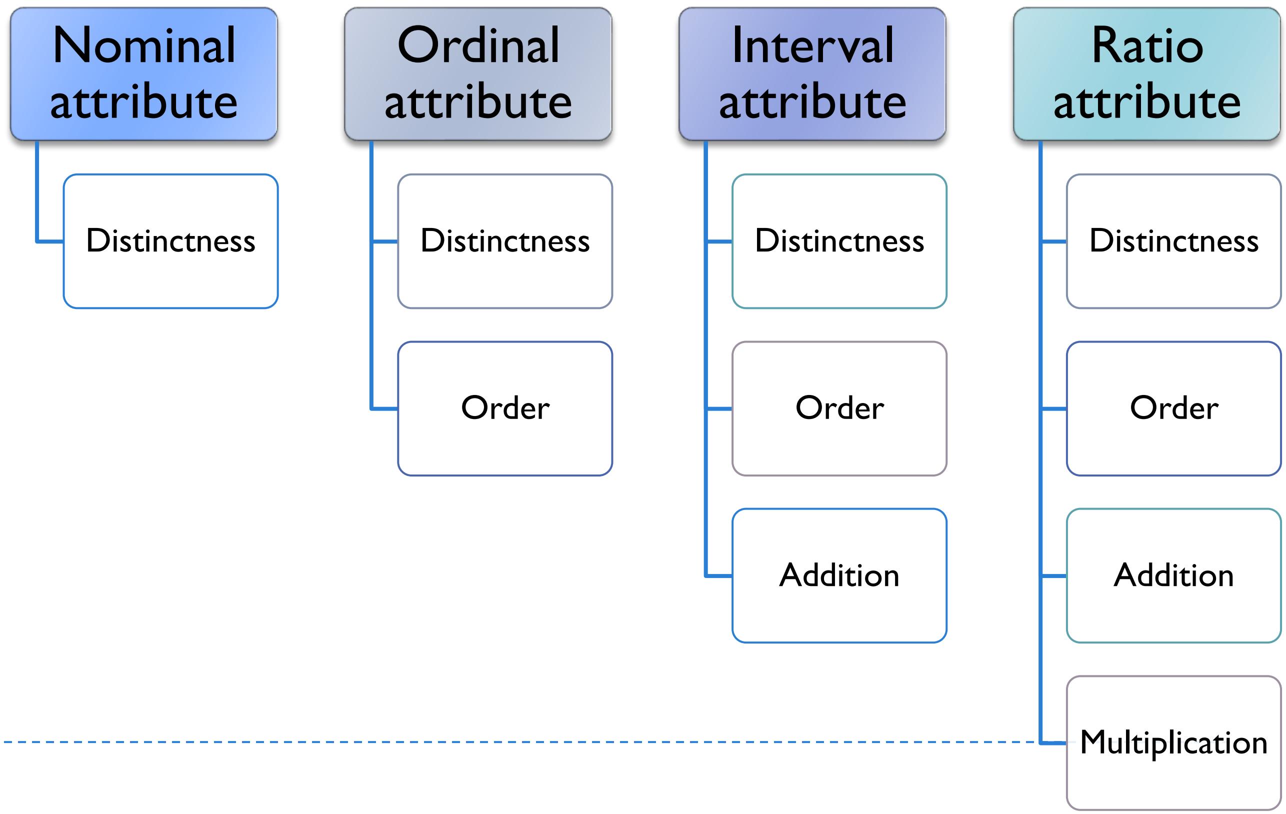
Interval data is like ordinal except we can say the **intervals** between each value are equally split.
Example: temperature in degrees Fahrenheit. The difference between 29 and 30 degrees is the same magnitude as the difference between 78 and 79

Ratio data is interval data with a natural zero point.
Examples: time is ratio since 0 time is meaningful.
Degrees Kelvin has a 0 point (absolute 0)

When measured on Kelvin scale the temperature of 100 is in physically meaningful way double of 50. As kelvin has fixed zero value.

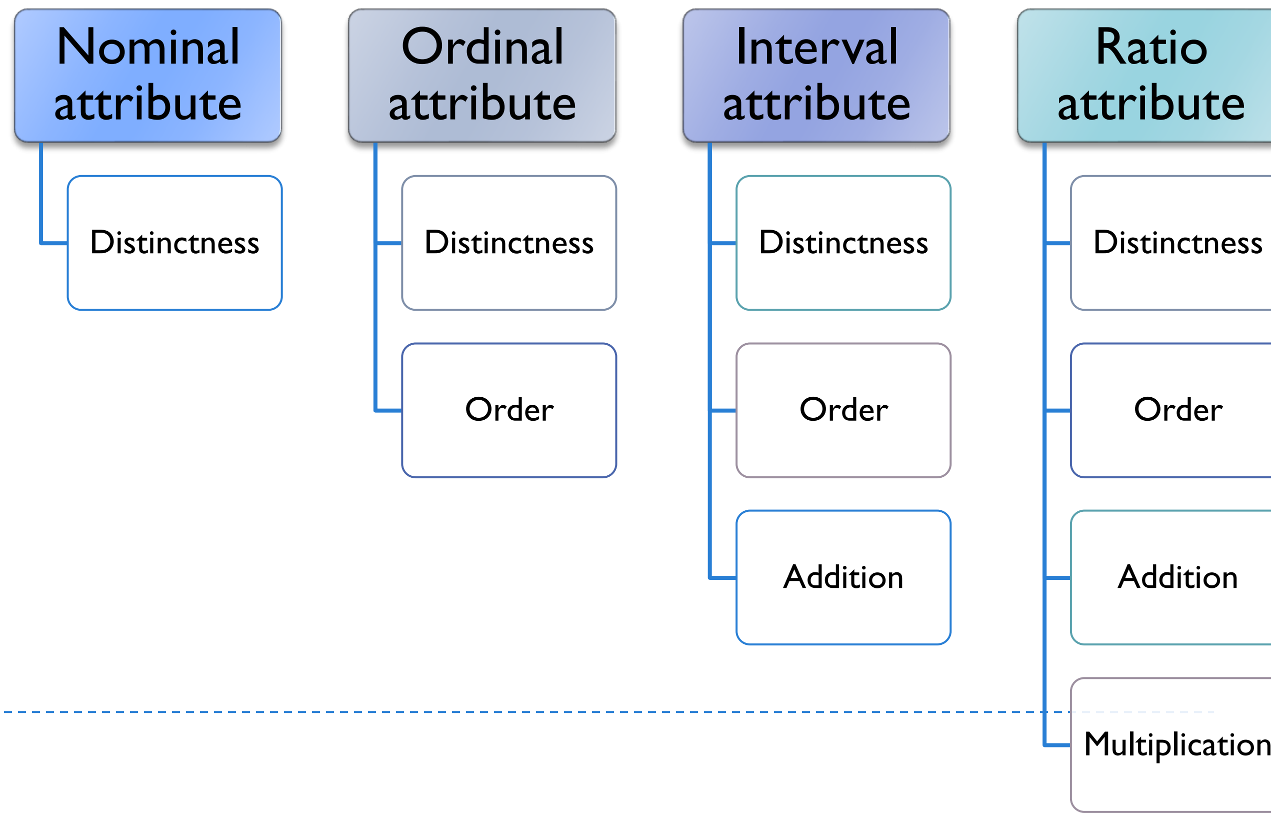
This is not true for Celcius or Farenheit scale. It does not have fixed zero and 50celcius is not half of 100celcius in physical term.

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



Titanic Dataset

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN



Discrete, Continuous, Asymmetric Attributes

Discrete Attribute

- Has finite or countably infinite set of values *(integer)*
- Nominal, ordinal, binary attributes
- Ex: zip codes, words in a collection of documents

Continuous Attribute

- Has real numbers as attribute values
- Interval and ratio attributes
- Ex: temperature, height, or weight

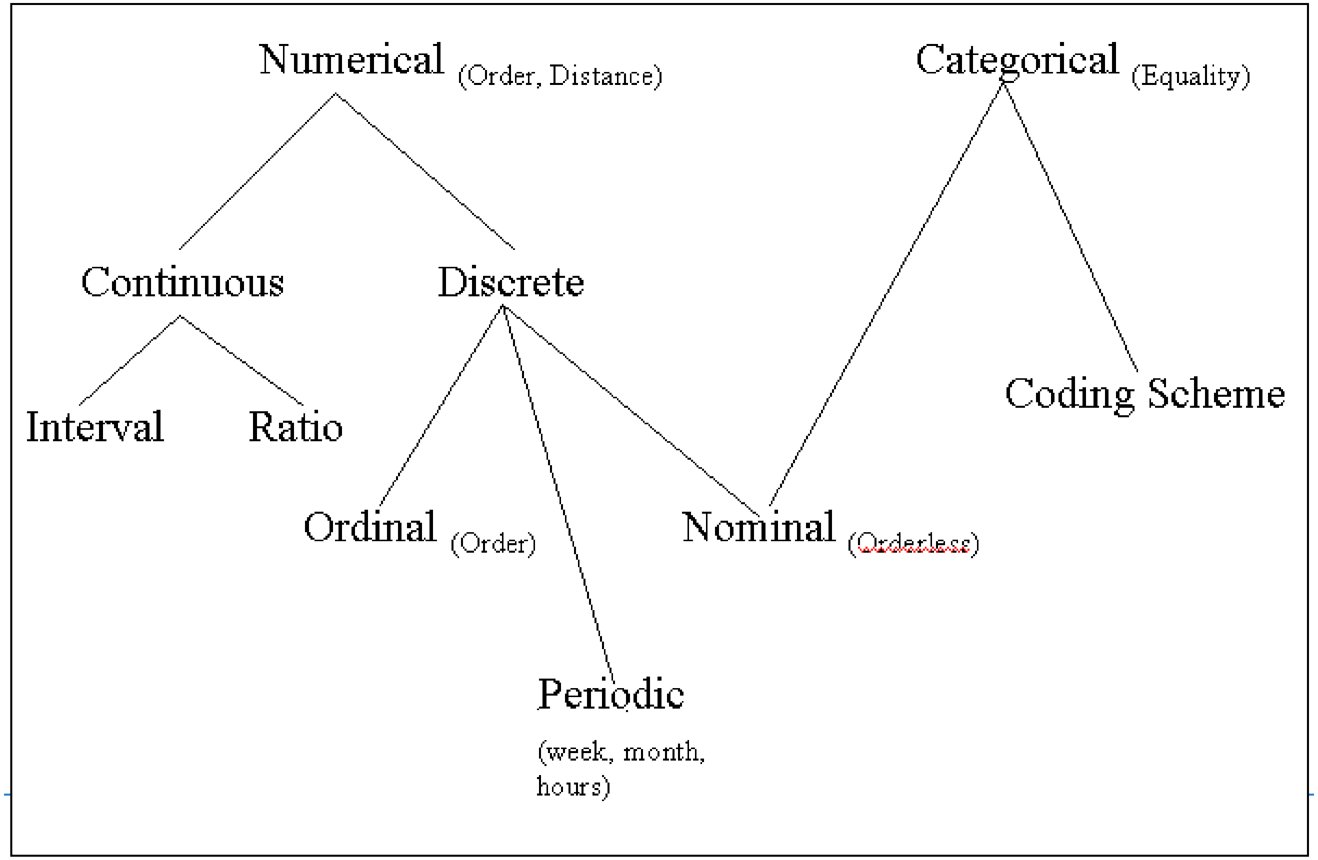
Asymmetric Attribute

- Only presence is regarded as important
- Ex: HIV positive

Data Types and Forms

- Attribute-value data:
- Data types
  - numeric, categorical (see the hierarchy for its relationship)

A1	A2	…	An	C



Step 1: To describe the dataset

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

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**

Step 2: To explore the dataset

- 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

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

Visualization tools are important

Histograms, 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

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

Useful Statistics

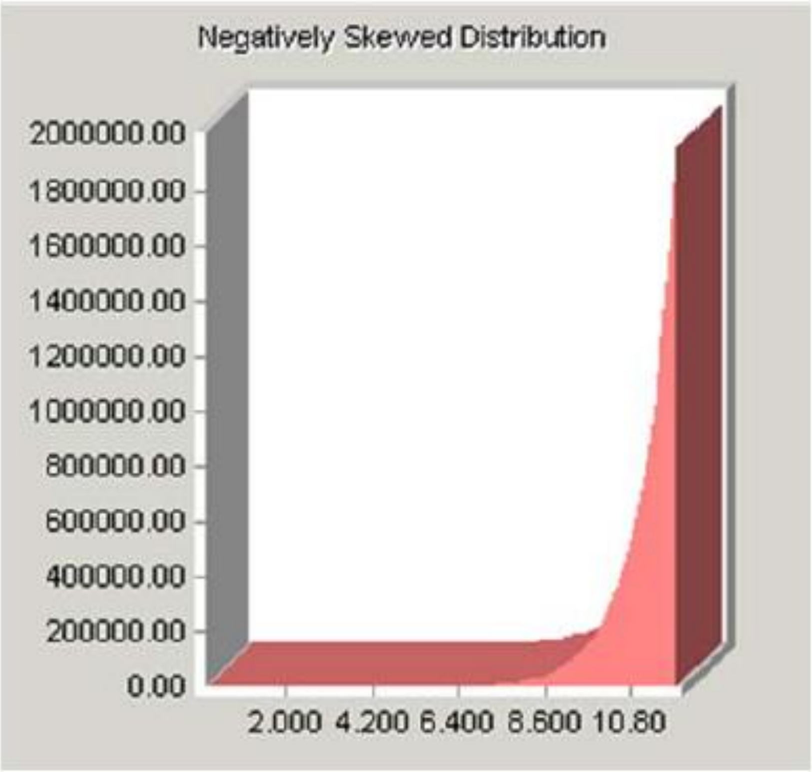
Discrete attributes

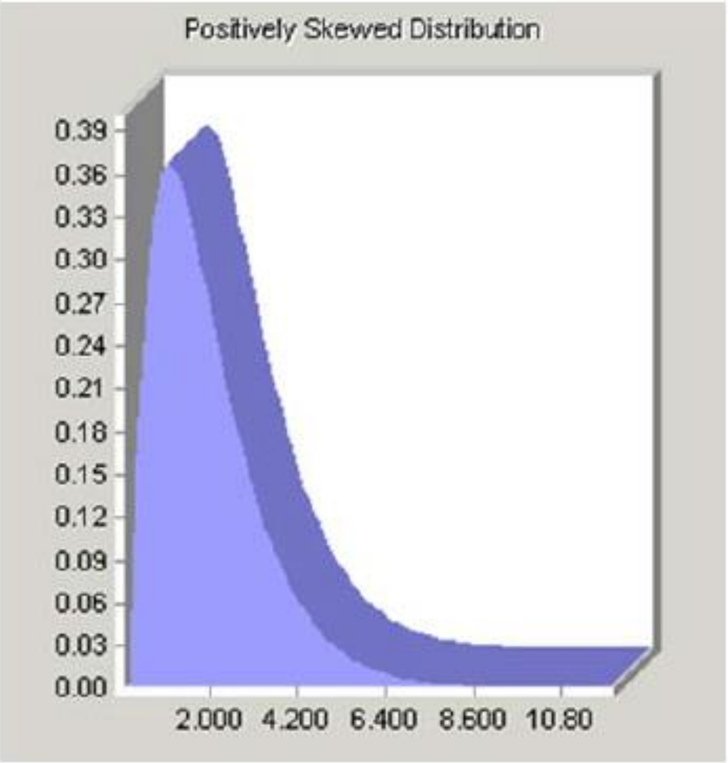
- Frequency of each value
- Mode = value with highest frequency

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

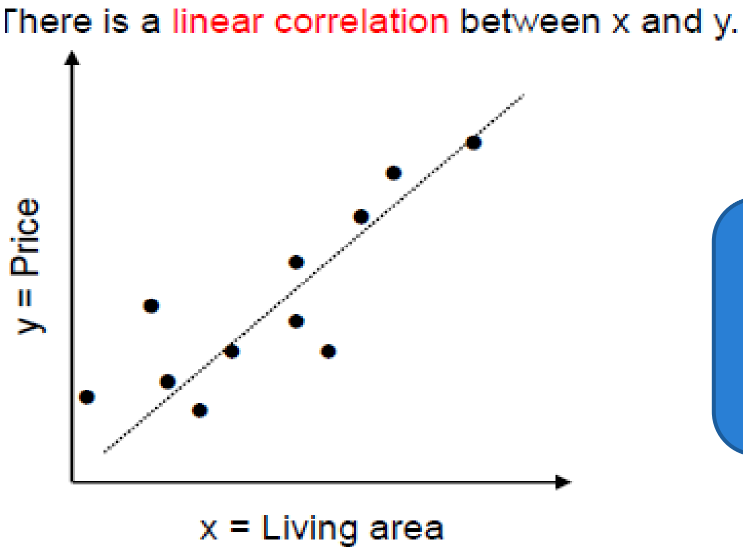
- How do the values of an attribute spread?
- Variance
  - Variance is sensitive to outliers

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

- What if the distribution of values is multimodal, i.e. data has several *bumps*?

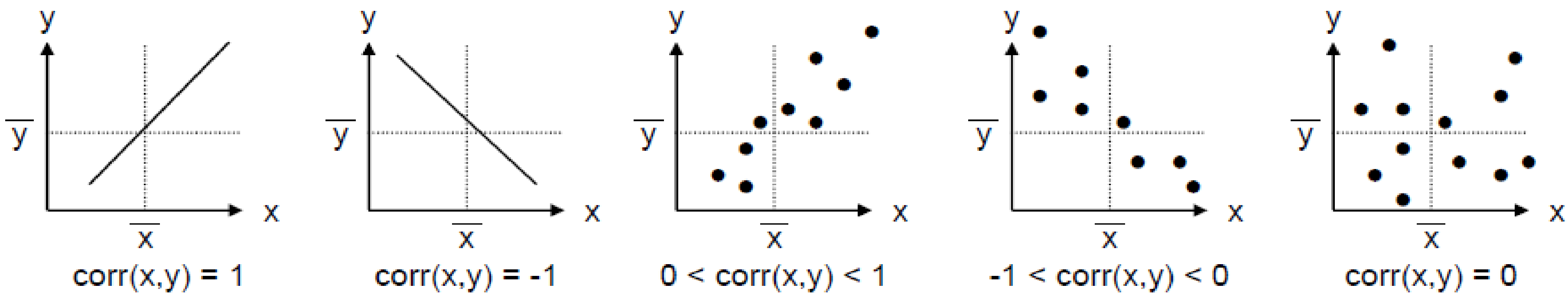
- Vizualization tools are useful

Attributes that Vary Together

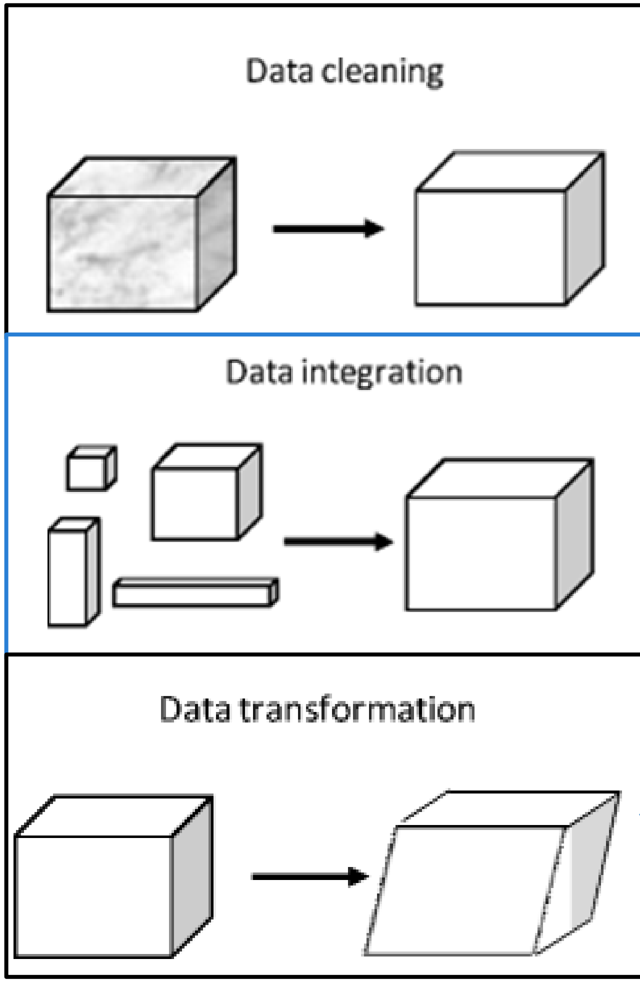


Correlation is a measure that describe how two attributes vary together

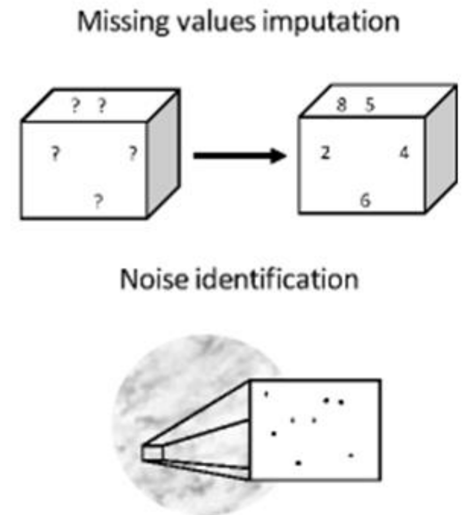
$$corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

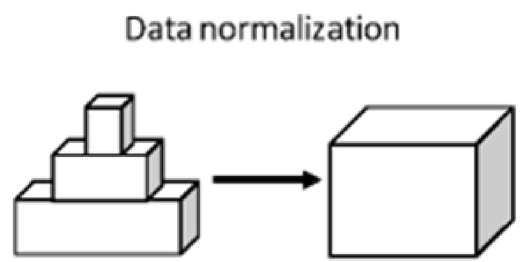


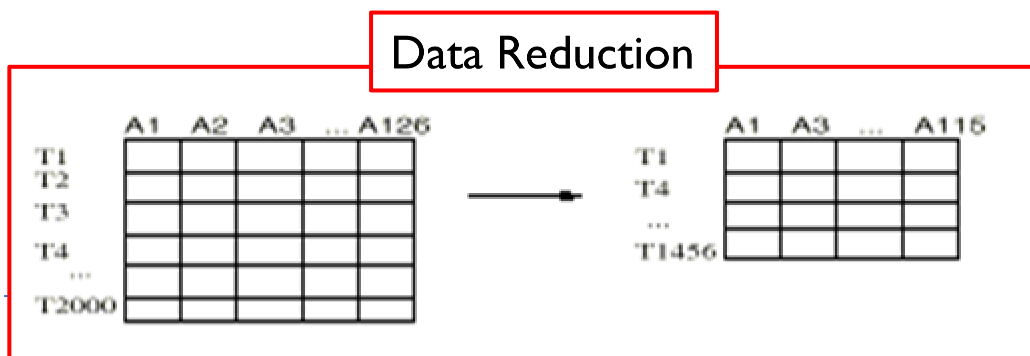
Forms of data preprocessing



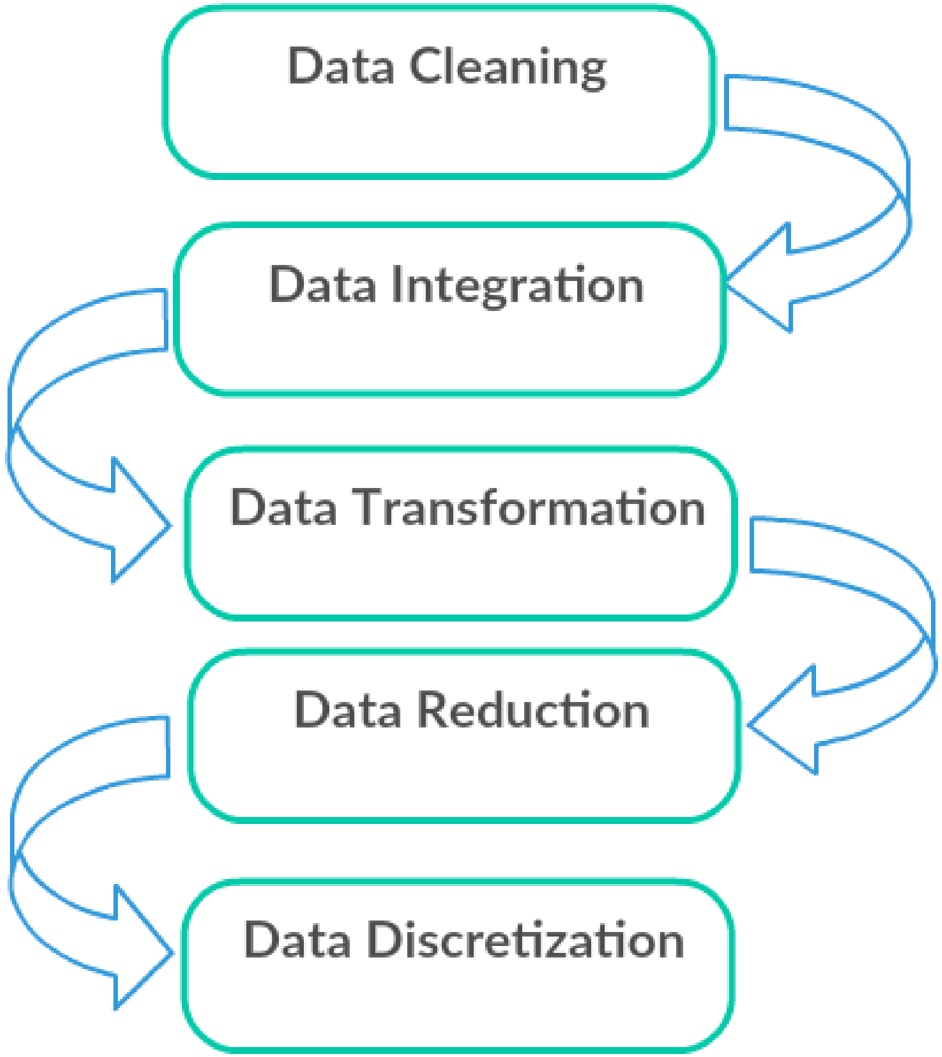
- Fill in missing values
- Smooth noisy data
- Remove outliers
- Resolve inconsistencies







Forms of data preprocessing



Data Cleaning -> Data Quality

- 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

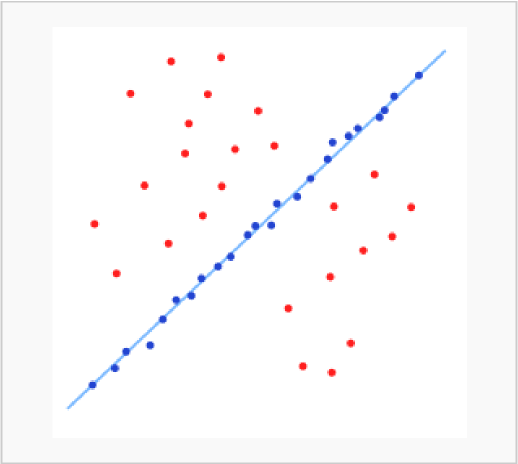
Missing Values

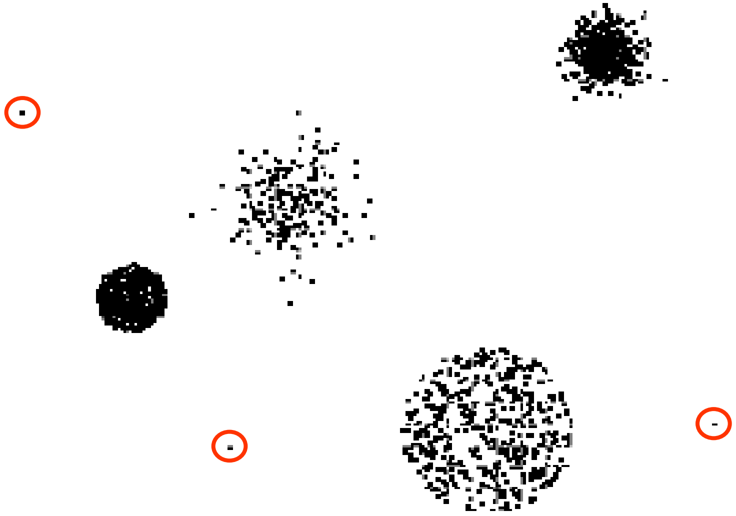
- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

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

Outliers

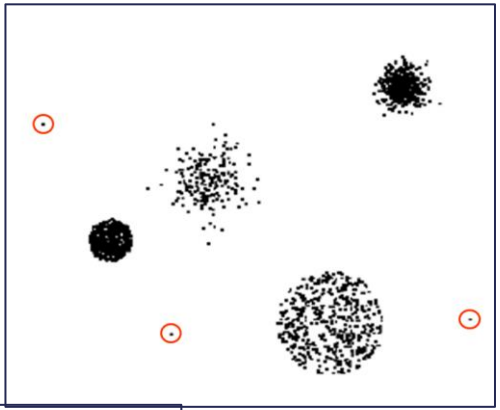
- Outliers are data objects with characteristics that are ***considerably different*** than most of the other data objects in the data set
- Can help to
  - detect new phenomenon or
  - discover unusual behavior in data
  - detect problems



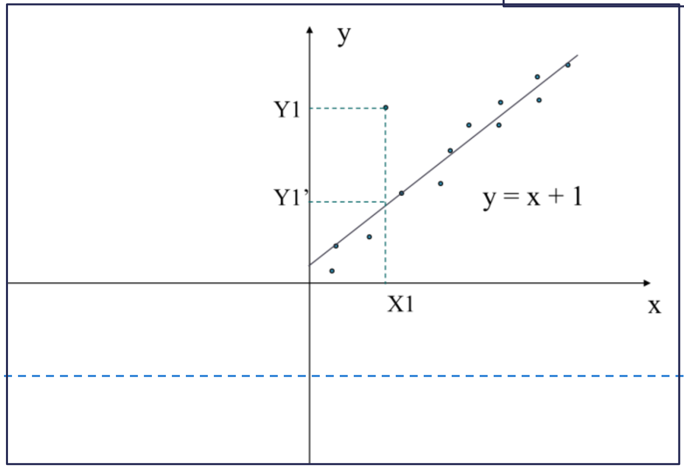


How to Handle Noisy Data?

- 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



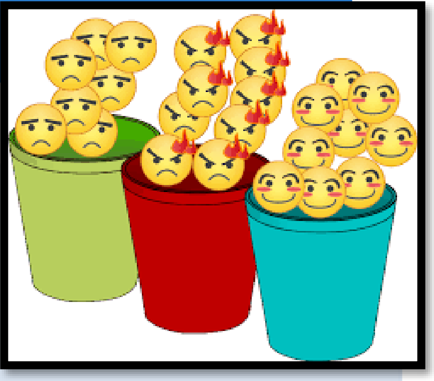
Data Discretization

- 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

Data Discretization



Unsupervised discretization

(Class labels are ignored)

- Equal-interval binning
- Equal-frequency binning

Supervised discretization

- Entropy-based discretization
- It tries to maximize the "purity" of the intervals
  - That is to contain as less as possible mixture of class labels

Binning (Equal-width)

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

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

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):

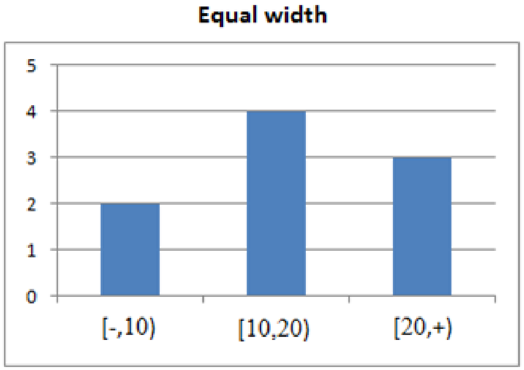
0, 4, 12, 16, 16, 18, 24, 26, 28

Equi-frequency binning – for bin density of 3:

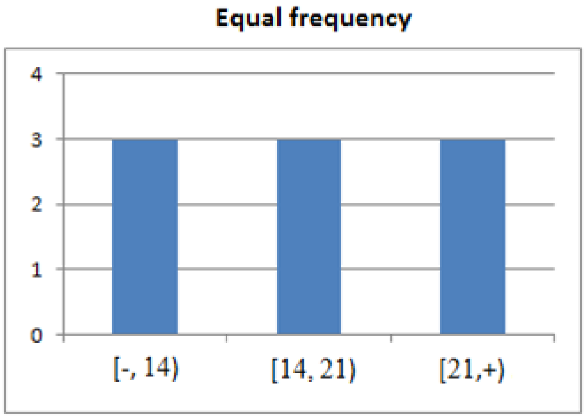
Bin 1: 0, 4, 12	[-, 14) bin
Bin 2: 16, 16, 18	[14, 21) bin
Bin 3: 24, 26, 28	[21,+] bin

Binning Example

- Attribute values (for an attribute age):
  - 0, 4, 12, 16, 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

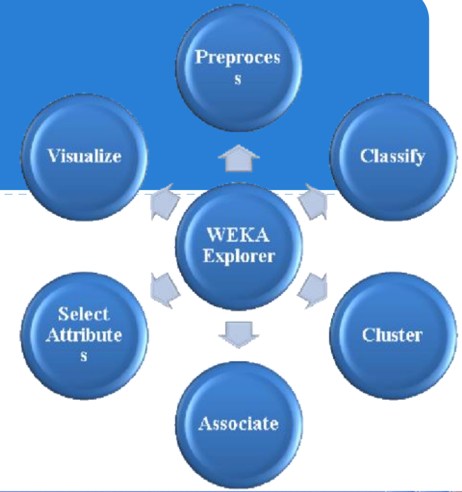
* Sorted data for price: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
* Partition into Equi-depth bins:

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: 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

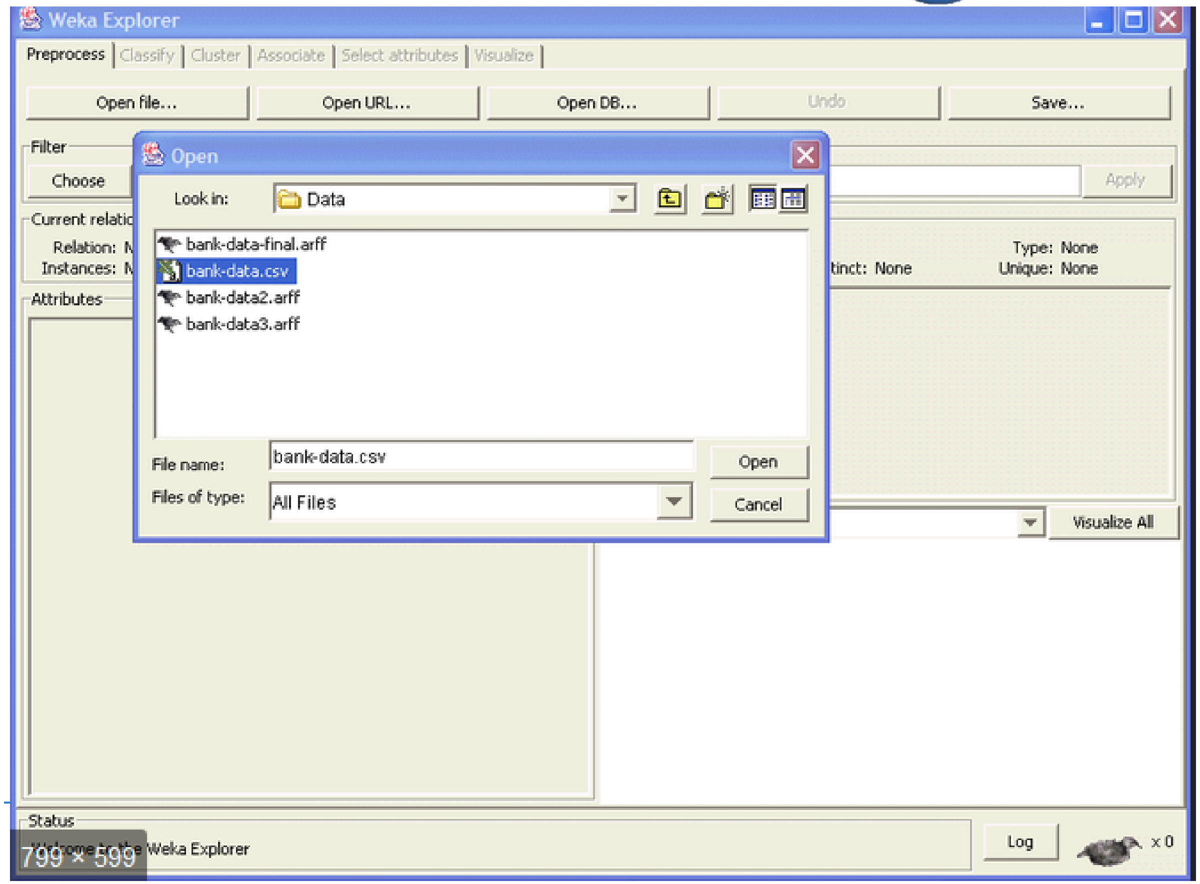
WEKA

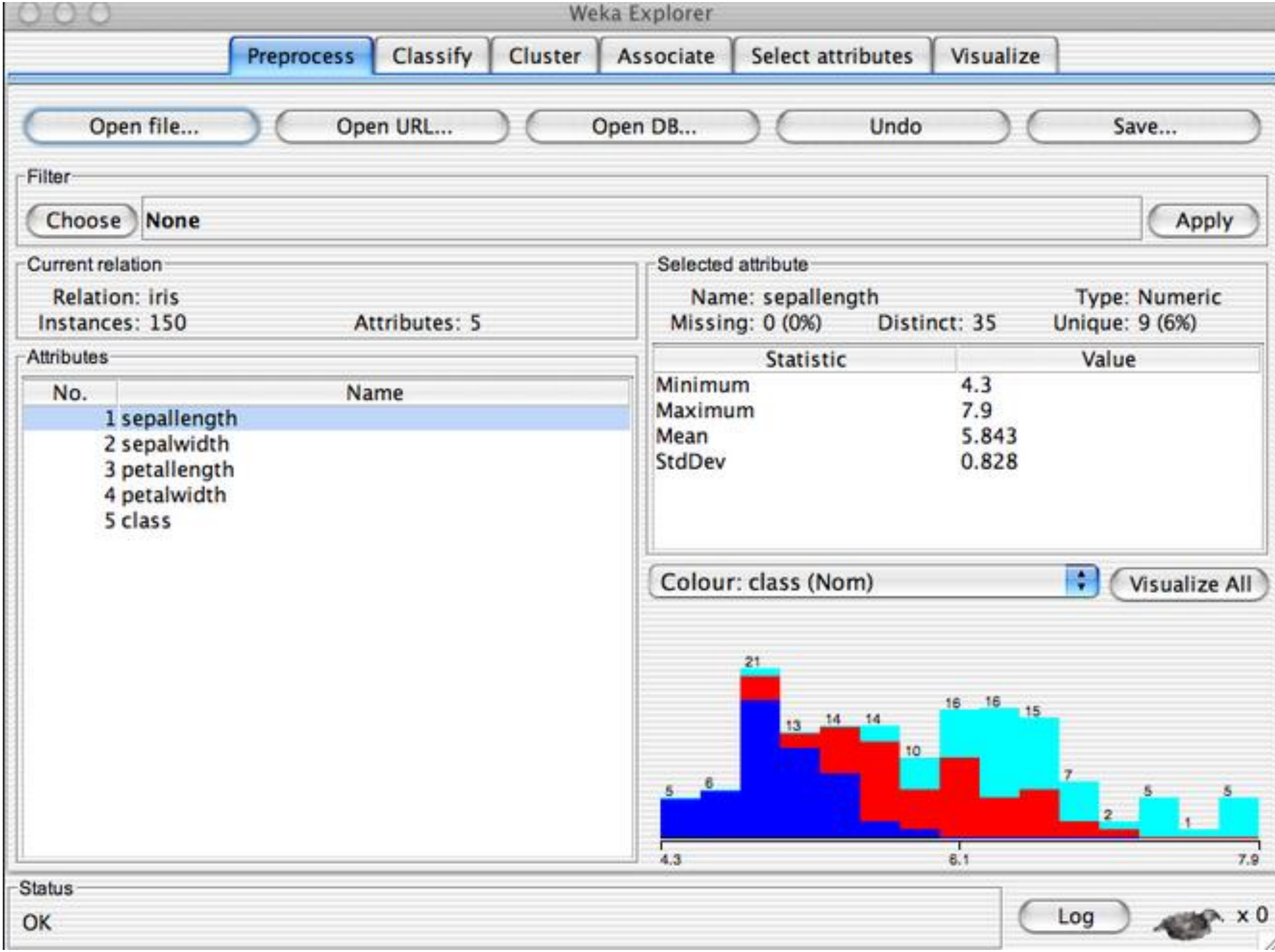


- An Automated Tool for Data Mining

- Download
- Install
- Explore Weka
- Read Tutorial

- Open an existing Data Set (IRIS)
  - Explore dataset





Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Statistic
Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Colour: class (Nom)
Visualize All
21
13
14
14
10
16
16
15
7
5
6
2
5
1
5
4.3
6.1
7.9
Status
OK
Log
x 0

Filters in Weka

- Filters – algorithms that transform the input dataset in some way

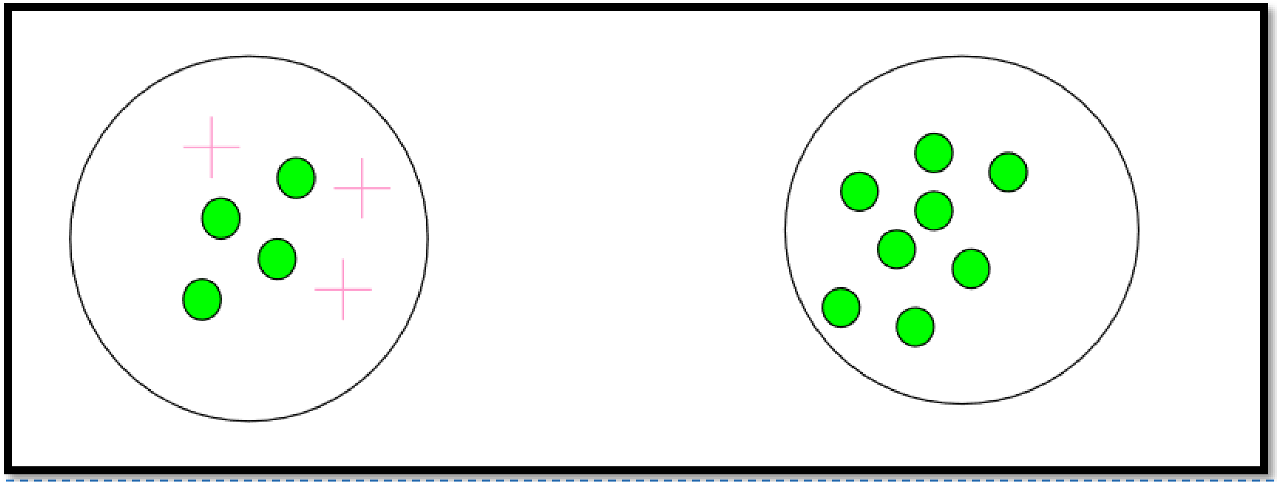
Filters		
Unsupervised	Attribute filter	`ReplaceMissingValues` `NumericTransform`
	Instance filter	`Resample`
Supervised		
	Attribute filter	`AttributeSelection` `Discretize`
	Instance filter	`Resample` `SpreadSubsample`

Discretization in Weka

Attribute Filter		Options
Unsupervised	`Discretize`	`bins`
		`useEqualFrequency`
Supervised	`Discretize`	

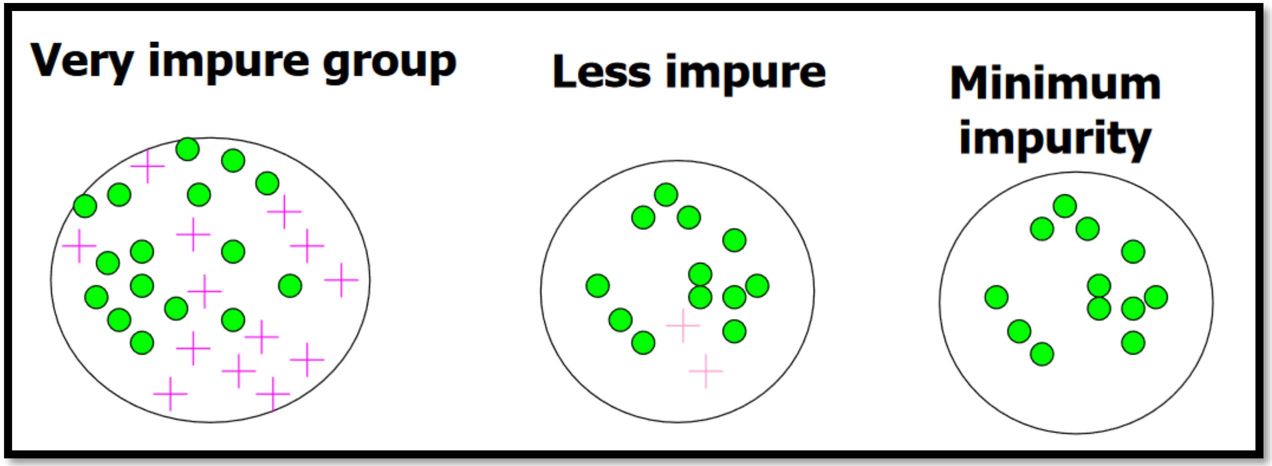
Supervised Discretization

- **Entropy-based discretization:**
  - The main idea is to split the attribute's value in a way that generates bins as "pure" as possible
- **Entropy:** Measures the level of impurity in a group of examples



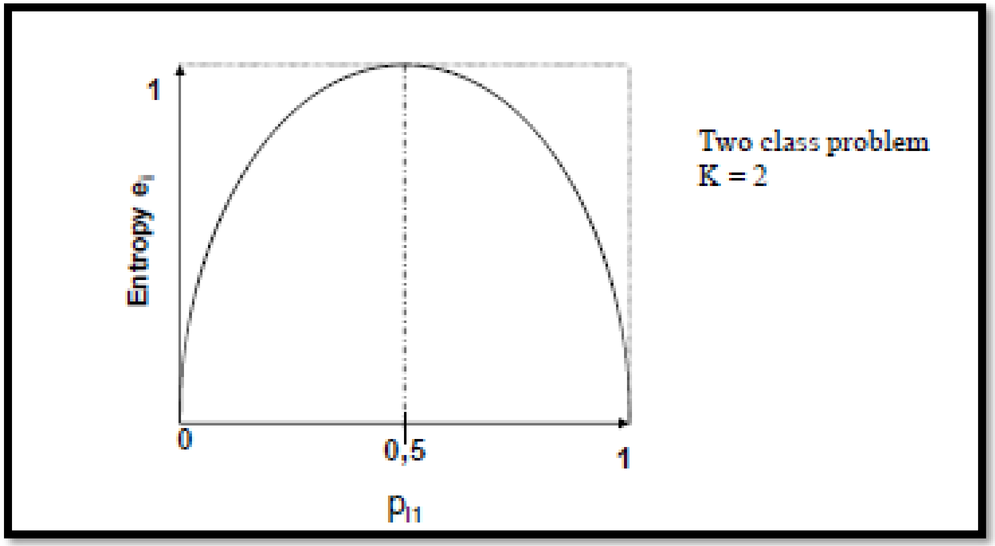
Entropy

- We need a measure of "**impurity of a bin**" such that
  - A bin with uniform class distribution has the highest impurity
  - A bin with all items belonging to the same class has zero impurity
  - The more skewed is the class distribution in the bin the smaller is the impurity



Entropy

$$e_i = -\sum_{j=1}^{k} p_{ij}\, log_2\, p_{ij}$$

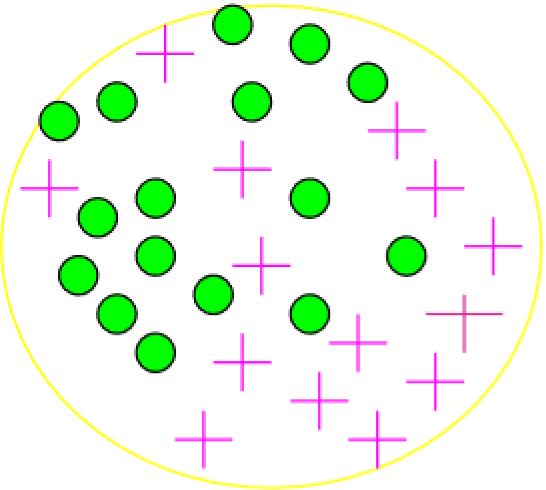


- $\boldsymbol{k}$ is the number of different class labels,
- $\boldsymbol{m_i}$ is the no. of values in ***i**^th^* interval of a partition
- $\boldsymbol{m_{ij}}$ is the no. of values of ***class j*** in ***i**^th^* interval
- $\boldsymbol{pij = \frac{m_{ij}}{m_i}}$, is the probability of ***class j*** in **i**th interval (relative frequency of class j)

Entropy

$$e_i = -\sum_{j=1}^{k} p_{ij} \, log_2 \, p_{ij}$$

- $\boldsymbol{K = 2}$ is the number of different class labels,
- $\boldsymbol{m_i = 30}$ is the no. of values in $\boldsymbol{i^{th}}$ interval of a partition
- $\boldsymbol{m_{ij}}$ is the no. of values of ***class j*** in $\boldsymbol{i^{th}}$ interval
- $\boldsymbol{pij = \frac{m_{ij}}{m_i}}$, is the probability of ***class j*** in **i**$^{\mathbf{th}}$ interval



16/30 are green circles;
14/30 are pink crosses
$\log_2(16/30) = -.9$
$\log_2(14/30) = -1.1$
Entropy = -(16/30)(-.9)- (14/30)(-1.1) = .99

Discretize using Entropy

- Place splits in a way that maximize the purity of the interval
- **A simple approach**
  - start by ***bisecting a continuous interval*** so that resulting interval gives min entropy.
  - This technique need to consider individual points only as we assume we have ***ordered list*** of points.
  - The splitting process is repeated with the interval with the ***worst entropy***,
    - until user specified number of intervals are reached.

A	C
4	N
5	Y
8	N
12	Y
15	Y

Entropy

- Total entropy of a partition is the weighted average of the individual entropies

$$e = \sum_{i=1}^{n} w_i \, e_i$$

- $\boldsymbol{n}$ is the number of intervals
- $\boldsymbol{m}$ is the number of values
- $\boldsymbol{w_i = \frac{m_i}{m}}$ is the fraction of values in i^{th} interval

$$e_i = -\sum_{j=1}^{k} p_{ij} \, log_2 \, p_{ij}$$

Entropy-based discretization

- **Algorithm**
  - Sort the sequence
  - Calculate Entropy for your data.
  - **For each potential split in your data…**
    - Calculate Entropy in each potential bin
    - Find the net entropy for your split
    - Calculate entropy gain
  - **Select the split with the highest entropy gain**
  - Recursively (or iteratively) perform the partition on each split until a termination criteria is met
    - Terminate once you reach a specified number of bins
    - Terminate once entropy gain falls below a certain threshold.

$$e_i = -\sum_{j=1}^{k} p_{ij} \, log_2 \, p_{ij}$$

$$e = \sum_{i=1}^{n} w_i \, e_i$$

A	C
4	N
5	Y
8	N
12	Y
15	Y

http://kevinmeurer.com/a-simple-guide-to-entropy-based-discretization/

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- Let us calculate the Entropy of the above dataset

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Calculate the Entropy of this dataset**

- Two class Label (Y, N)
- **Interval 1(entire data)**

$$e_i = -\sum_{j=1}^{k} p_{ij}\, log_2\, p_{ij}$$

Y	N
3	2

$$Entropy(D) = -(\tfrac{3}{5}log_2(\tfrac{3}{5}) + \tfrac{2}{5}log_2(\tfrac{2}{5})) = .529 + .442 = .971$$

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 1 at 4.5**
  - Two class Label (Y, N)
  - **Interval 2**

	Y	N
<=4.5	0	1
>4.5	3	1

$$e_i = -\sum_{j=1}^{k} p_{ij}\, log_2\, p_{ij}$$

Entropy for each bin

$$Entropy(D_{<=4.5}) = -(\tfrac{1}{1}log_2(1) + 0log_2(0)) = 0 + 0 = 0$$

$$Entropy(D_{>4.5}) = -(\tfrac{3}{4}log_2(\tfrac{3}{4}) + \tfrac{1}{4}log_2(\tfrac{1}{4})) = .311 + .5 = .811$$

Remember !! Total entropy of a partition is the weighted average of the individual entropies

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 1 at 4.5**
  - Two class Label (Y, N)
  - **Interval 2**

	Y	N
<=4.5	0	1
>4.5	3	1

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$

Entropy for each bin

$$Entropy(D_{<=4.5}) = -(\tfrac{1}{1} log_2(1) + 0 log_2(0)) = 0 + 0 = 0$$

$$Entropy(D_{>4.5}) = -(\tfrac{3}{4} log_2(\tfrac{3}{4}) + \tfrac{1}{4} log_2(\tfrac{1}{4})) = .311 + .5 = .811$$

Entropy (split1) $= \frac{1}{5}(0) + \frac{4}{5}(.811) = .6488$

$$e = \sum_{i=1}^{n} w_i \, e_i$$

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 1 at 4.5**
  - Two class Label (Y, N)
  - **Interval 2**

	Y	N
<=4.5	0	1
>4.5	3	1

$$e_i = -\sum_{j=1}^{k} p_{ij}\, log_2\, p_{ij}$$

Entropy (split1) $= \frac{1}{5}(0) + \frac{4}{5}(.811) = .6488$

$$Entropy(D) = -(\tfrac{3}{5}log_2(\tfrac{3}{5}) + \tfrac{2}{5}log_2(\tfrac{2}{5})) = .529 + .442 = .971$$

$$\boxed{Gain(D_{new}) = .971 - .6488 = .322}$$

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 2 at 6.5**
  - Two class Label (Y, N)
  - Interval 2

	Y	N
<=6.5	1	1
>6.5	2	1

$$e_i = -\sum_{j=1}^{k} p_{ij} \, log_2 \, p_{ij}$$

$$e = \sum_{i=1}^{n} w_i \, e_i$$

Entropy for each bin

$$Entropy(D_1) = -(.5log_2(.5) + .5log_2(.5)) = 1$$

$$Entropy(D_1) = -(\tfrac{2}{3}log_2(\tfrac{2}{3}) + \tfrac{1}{3}log_2(\tfrac{1}{3})) = .389 + .528 = .917$$

Entropy (split2) $= \frac{1}{3}(1) + \frac{2}{3}(.917) = .944$

$$\boxed{Gain(D_{new}) = .971 - .944 = .027}$$

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 3: 10**
  - Two class Label (Y, N)
  - Interval 2

	Y	N
<=10	1	2
>10	2	0

$$e_i = -\sum_{j=1}^{k} p_{ij} \, log_2 \, p_{ij}$$

$$e = \sum_{i=1}^{n} w_i \, e_i$$

Entropy for each bin

$$Entropy(D_{<=10}) = \frac{1}{3} log_2(\frac{1}{3}) + \frac{2}{3} log_2(\frac{2}{3}) = .917$$

$$Entropy(D_{>10}) = -(1 log_2(1) + 0 log_2(0)) = 0$$

Entropy (split 3) $= \frac{2}{5}(0) + \frac{3}{5}(.917) = .55$

$$\boxed{Gain(D_{new}) = .971 - .55 = .421}$$

Entropy Example

Hours Studied	A on Test
4	N
5	Y
8	N
12	Y
15	Y

- **Split 4: 13.5**
  - Two class Label (A, not A)
  - Interval 2

	A	Not A
<=13.5	2	2
>13.5	1	0

Entropy Example

- **Choose the split**
  - The best split is split 3, which gives us the max gain of 0.421.
  - We will partition the data there!

- According to the algorithm, we can further bin our attributes in the bins we just created.

- This process will continue until we satisfy a termination criteria.

Entropy-based discretization

- **Algorithm**
  - Sort the sequence
  - Calculate Entropy for your data.
  - **For each potential split in your data…**
    - Calculate Entropy in each potential bin
    - Find the net entropy for your split
    - Calculate entropy gain
  - **Select the split with the highest entropy gain**
  - **Recursively (or iteratively) perform the partition on each split until a termination criteria is met**
    - Terminate once you reach a specified number of bins
    - Terminate once entropy gain falls below a certain threshold.

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$

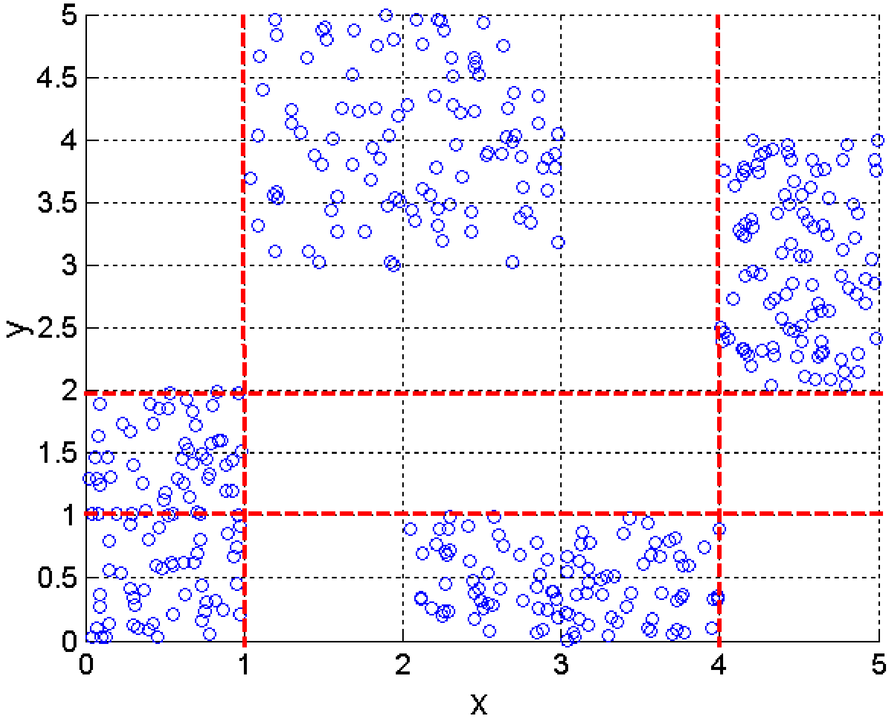
$$e = \sum_{i=1}^{n} w_i \, e_i$$

A	C
4	N
5	Y
8	N
12	Y
15	Y

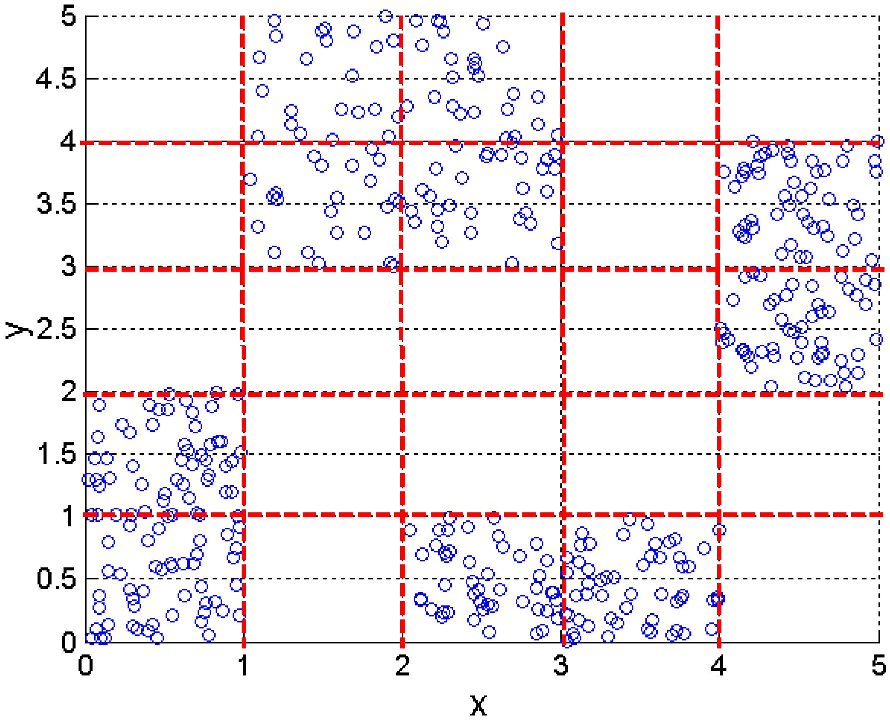
http://kevinmeurer.com/a-simple-guide-to-entropy-based-discretization/

Discretization Using Class Labels

- **Entropy based approach**



3 categories for both x and y



5 categories for both x and y

Data Transformation

Transform or consolidate data into forms appropriate for mining

Smoothing: remove noise from data

Aggregation: summarization, data cube construction

- Daily sales data aggregated to compute monthly or annual amount

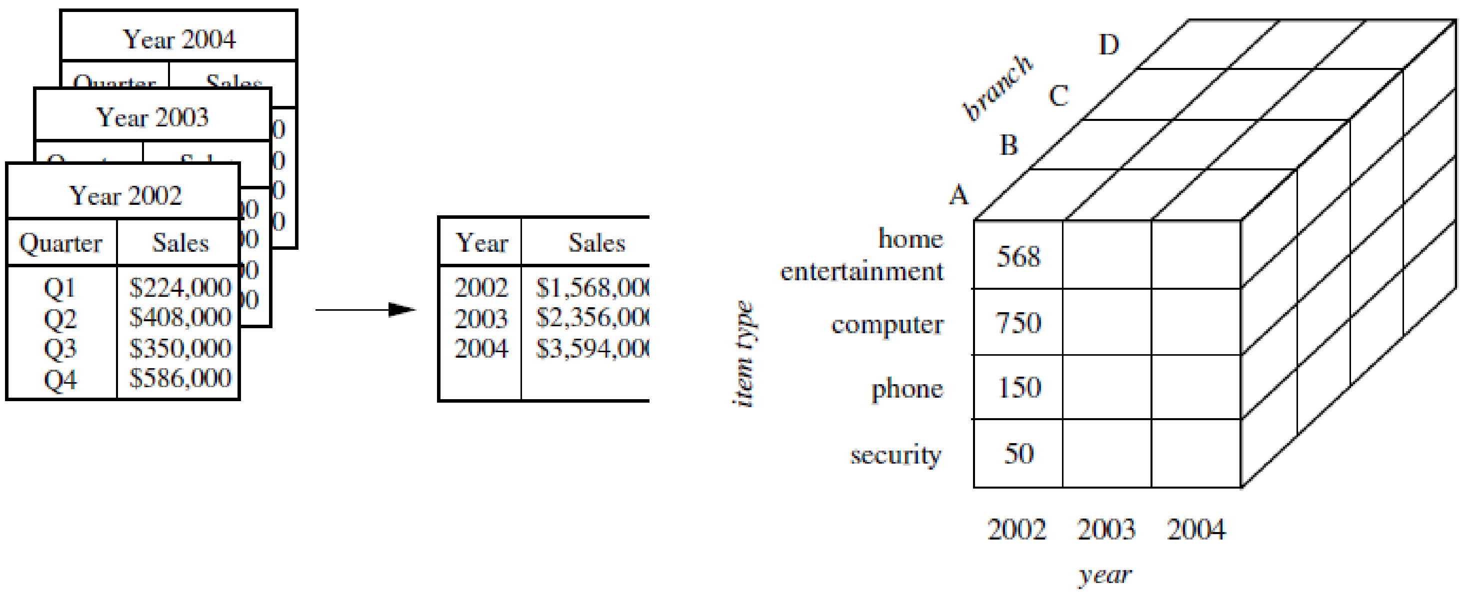
Generalization: concept hierarchy

Normalization: scaled to fall within a small, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Data Transformation

- **Aggregation:** summarization, data cube construction
  - Daily sales data aggregated to compute monthly or annual amount



A data cube for sales

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.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **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 Transformation: Normalization

- **Decimal scaling**
  - normalizes by moving the decimal point of values of attribute *A*.
  - The number of decimal points moved depends on the maximum absolute value of *A*.

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \mathrm{Max}(|v'|) < 1$$

- Example: Suppose that the recorded values of *A* range from -986 to 917.
  - The maximum absolute value of *A* is 986.
  - To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$)
  - -986 normalizes to -0.986 and 917 normalizes to 0.917.

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
    - Histograms
    - Clustering
- **Discretization and concept hierarchy generation**
  - replace raw data values for attributes by ranges or higher conceptual levels
- **Data compression**
  - use encoding schemes to reduce the data set size

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

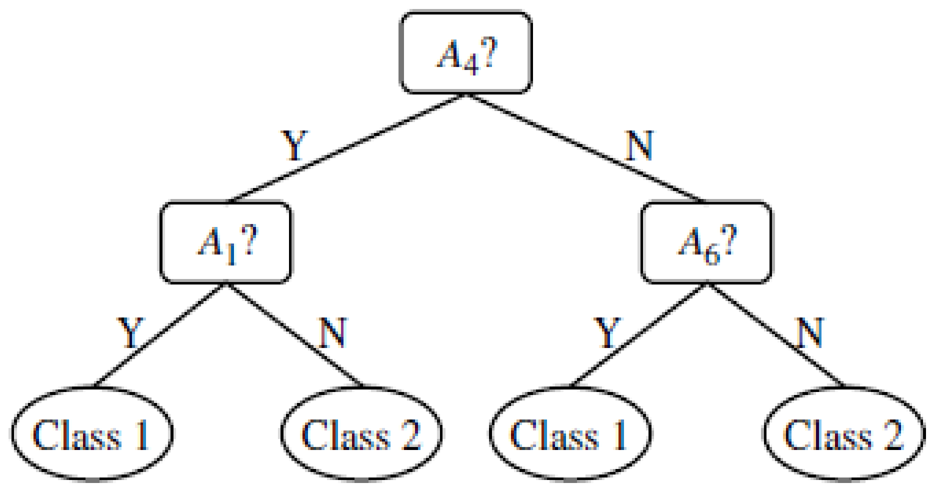
Feature Subset Selection

Brute-force approch	• Try all possible feature subsets as input to data mining algorithm
Embedded approaches	• Feature selection occurs naturally as part of the data mining algorithm
Filter approaches	• Features are selected before data mining algorithm is run
Wrapper approaches	• Use the data mining and machine learning algorithm as a black box to find best subset of attributes

Feature Subset Selection

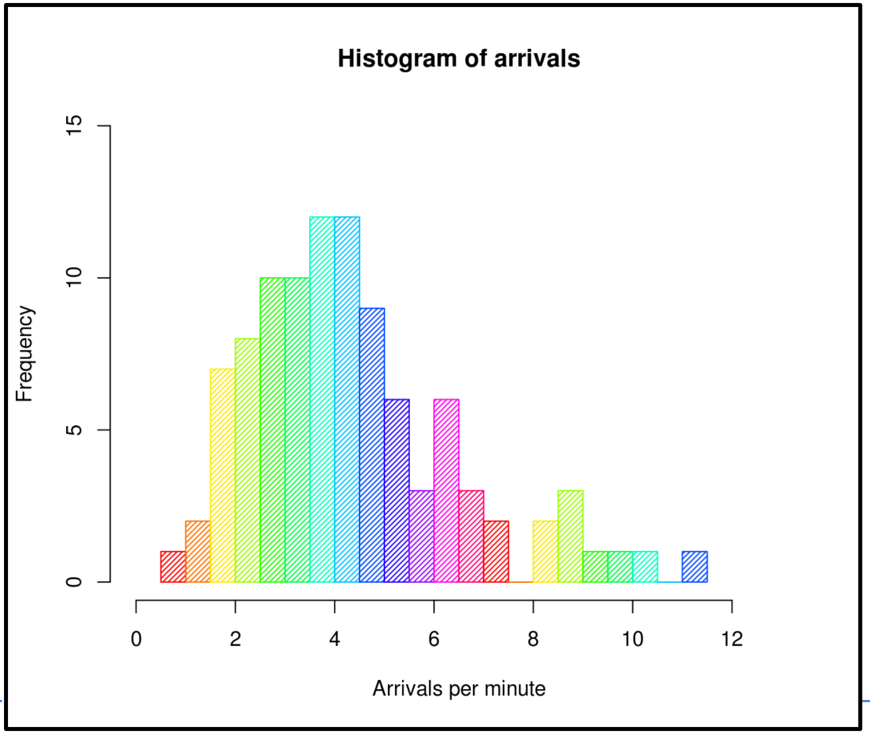
- Wrapper approaches … Heuristic methods (due to exponential # of choices):
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction

Feature Subset Selection

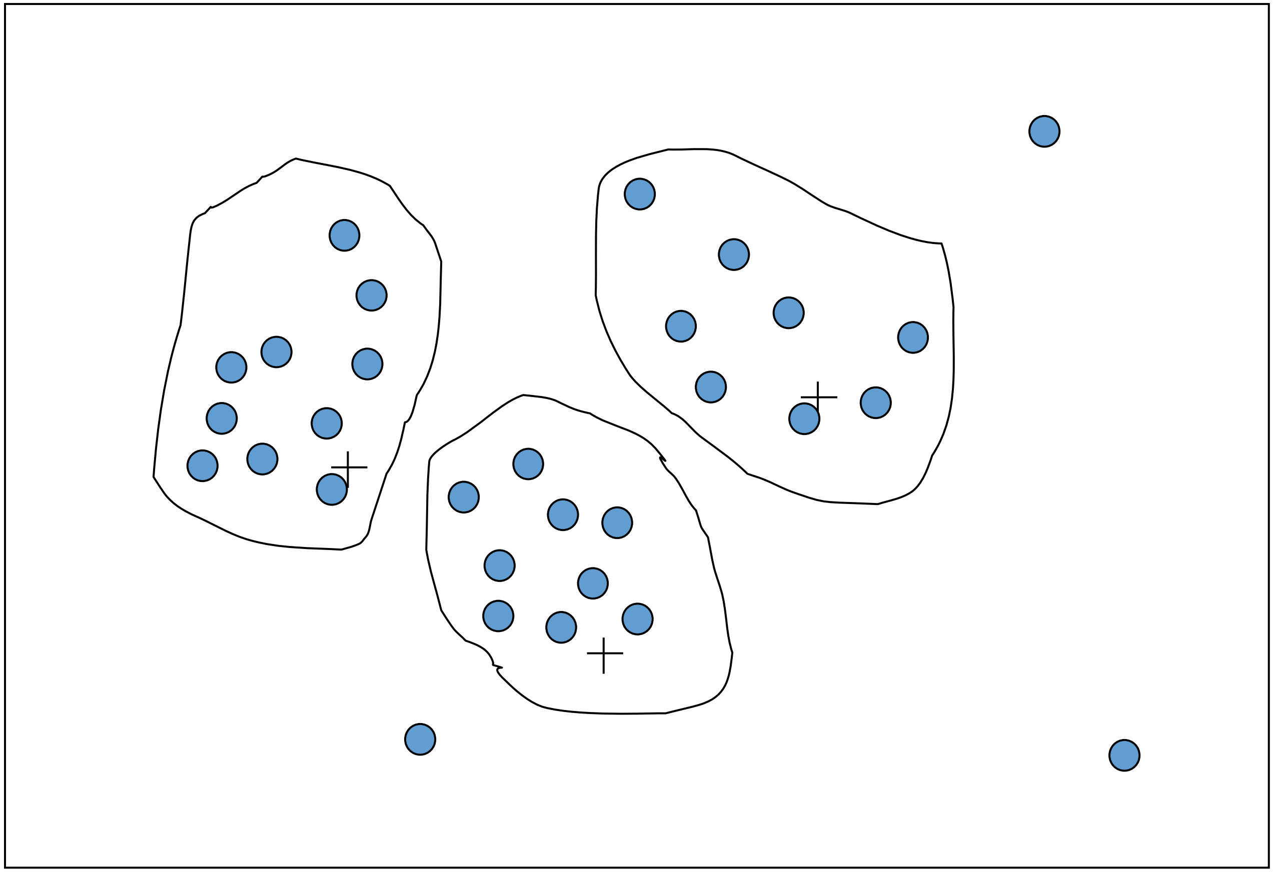
Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: {} => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Numerosity reduction -Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Same as Binning
- Can be constructed optimally in one dimension using dynamic programming



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 …

The key principle for effective sampling is the following:

using a sample will work almost as well as using the entire data sets, if the sample is representative

A sample is representative if it has approximately the same property (of interest) as the original set of data

Otherwise we say that the sample introduces some bias

Types of Sampling

Simple Random Sampling

- There is an equal probability of selecting any particular item

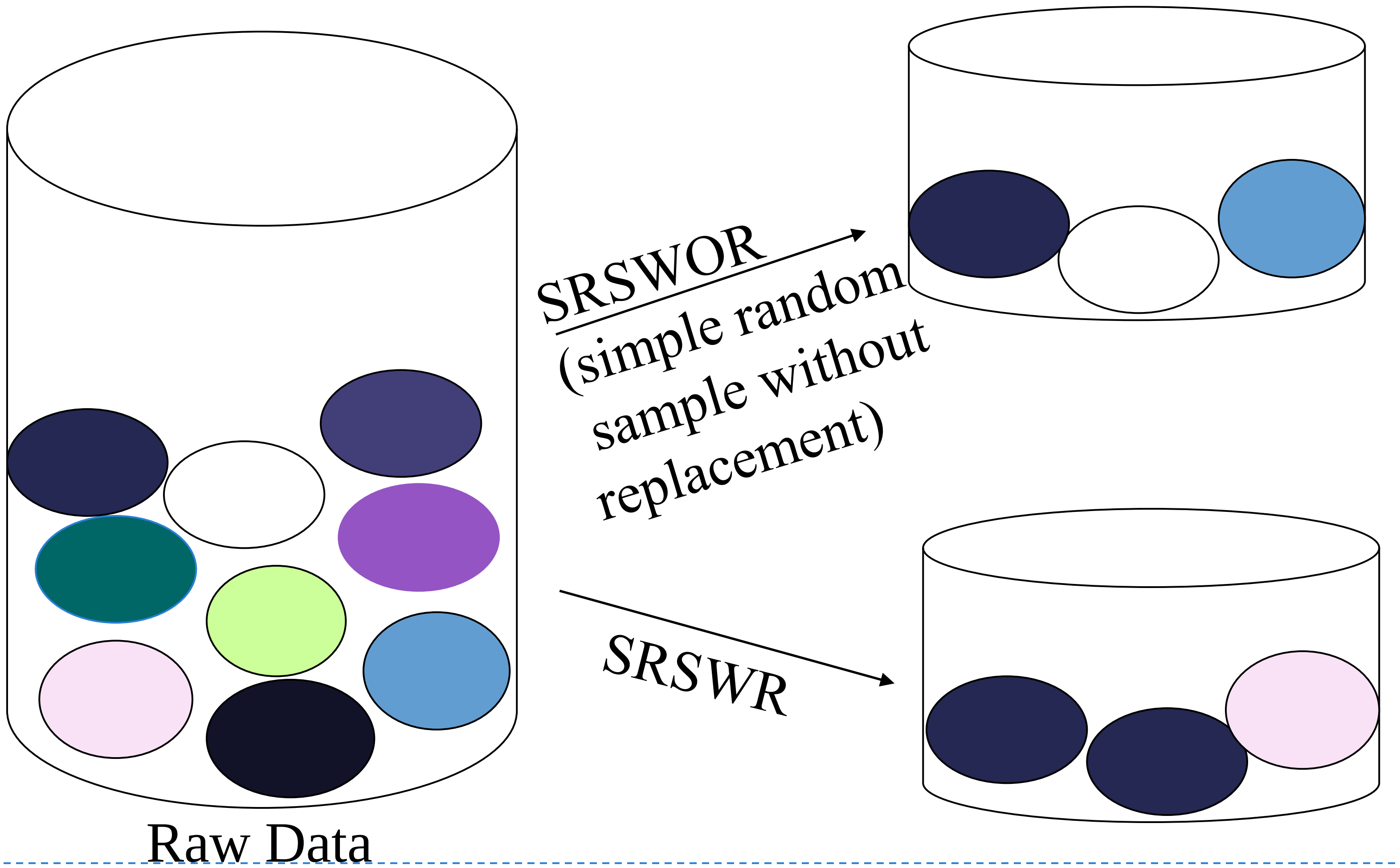
Sampling without replacement

- As each item is selected, it is removed from the population

Sampling with replacement

- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once.
- This makes analytical computation of probabilities easier

Sampling


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data

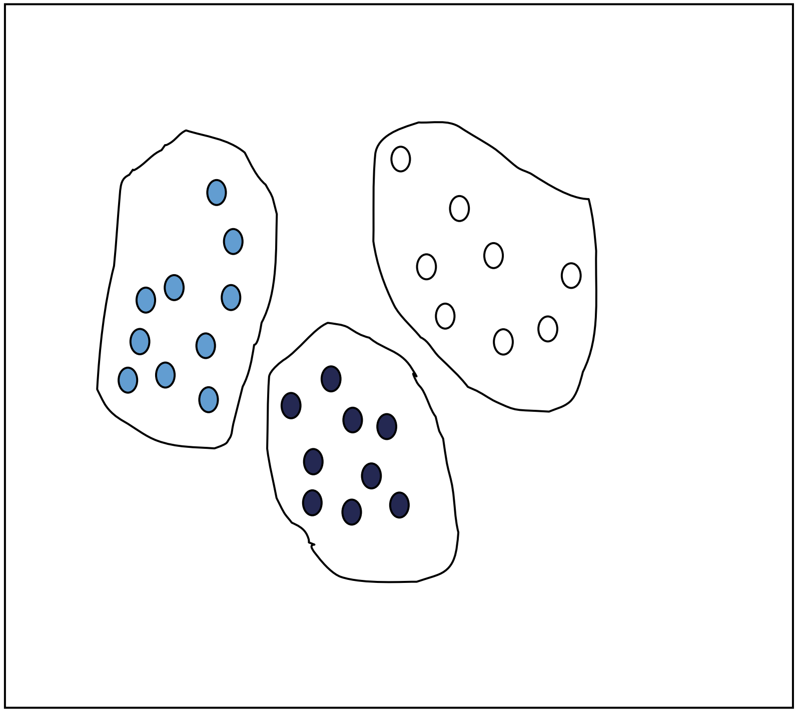
Types of Sampling

- **Stratified sampling**
  - Split the data into several **groups**; then draw random samples from each group.
  - Ensures that both groups are represented.

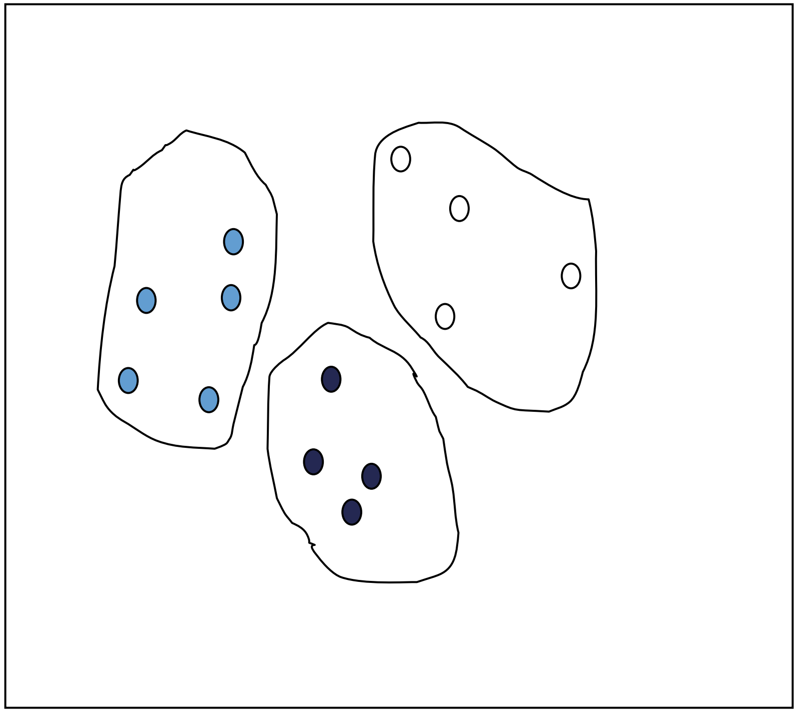
  - **Example** Find difference between legitimate and fraudulent credit card transactions.
  - 0.1% of transactions are fraudulent. What happens if we select 1000 transactions at random?
    - We get 1 fraudulent transaction (in expectation). Not enough to draw any conclusions.
    - Solution: sample 1000 legitimate and 1000 fraudulent transactions

Sampling

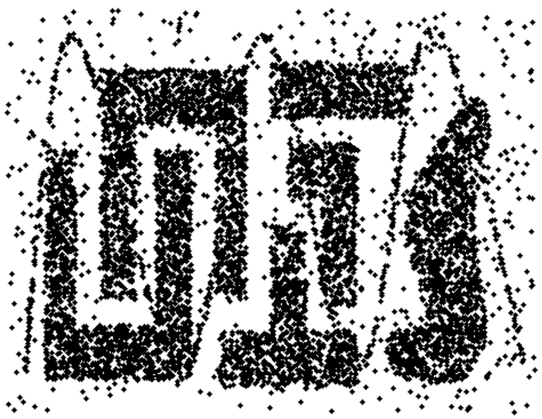
Raw Data



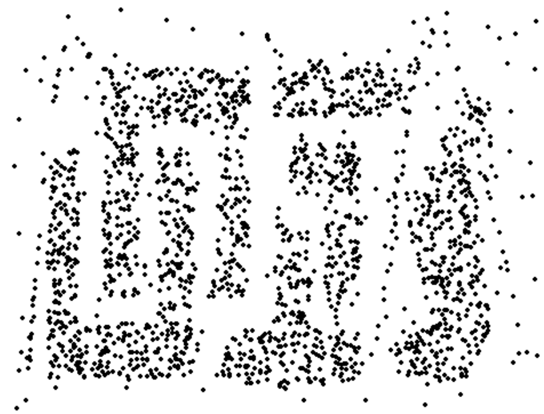
Cluster/Stratified Sample



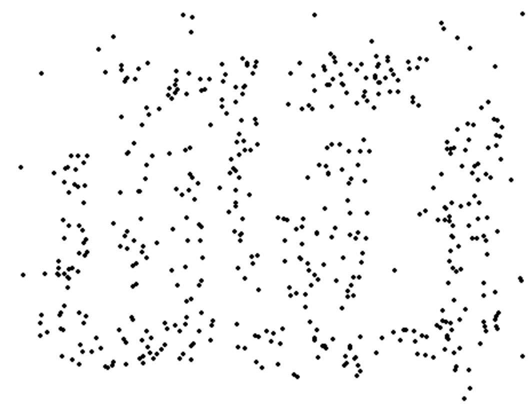
Sample Size



8000 points Points



2000 Points



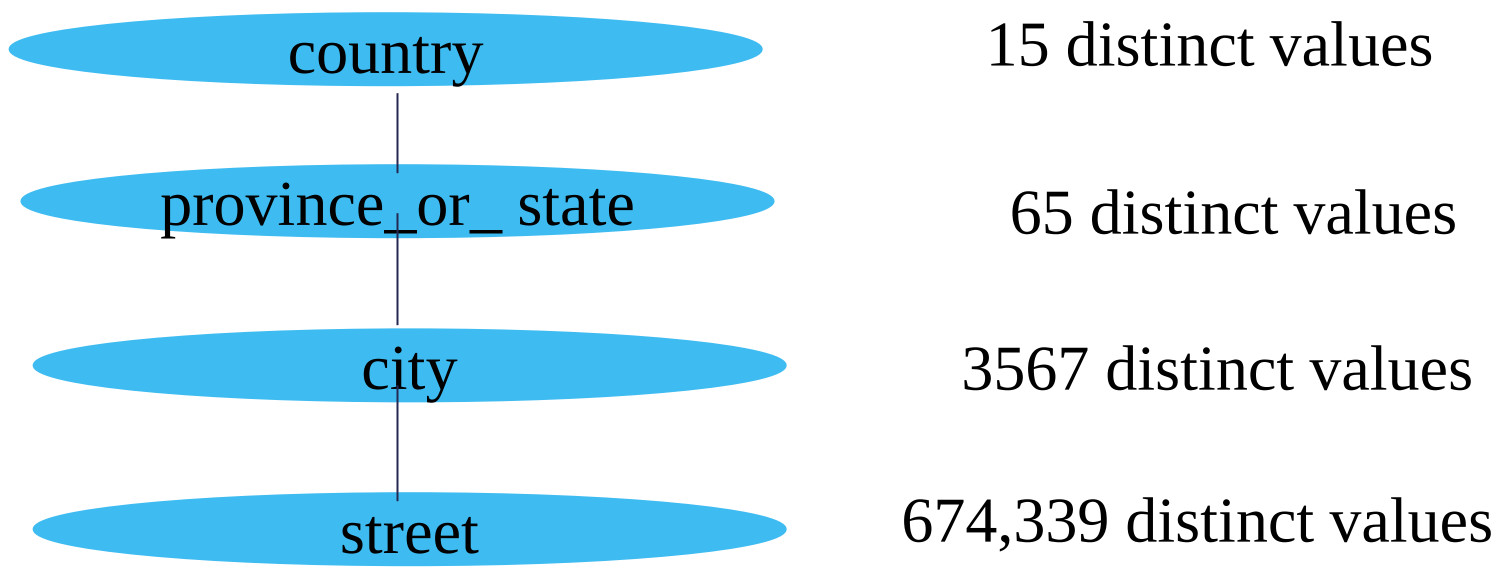
500

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
  - Help select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

Visualization

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

  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

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

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

Iris Sample Data Set

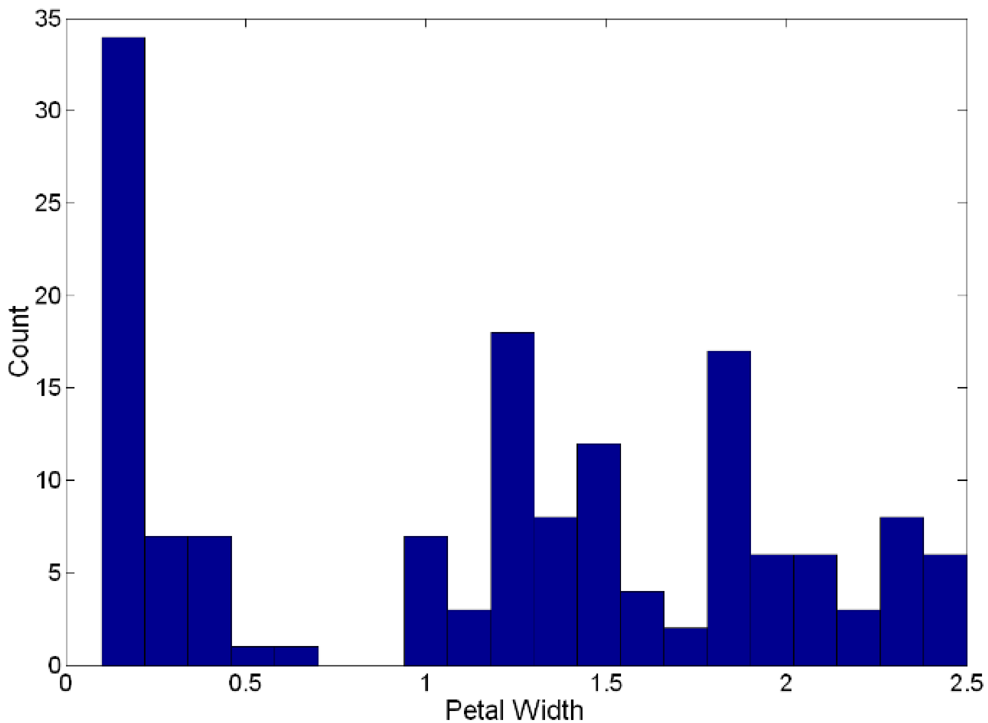
- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

- 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.

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

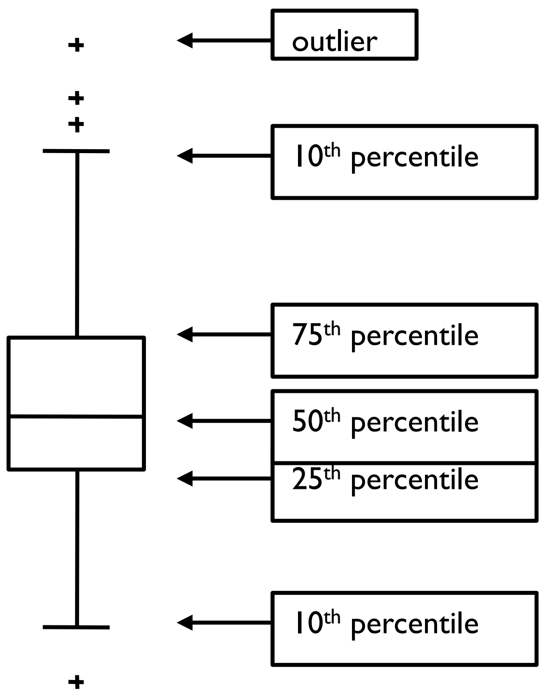


Example: Petal Width

(10 and 20 bins, respectively)

Visualization Techniques: Box Plots

- Box Plots
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



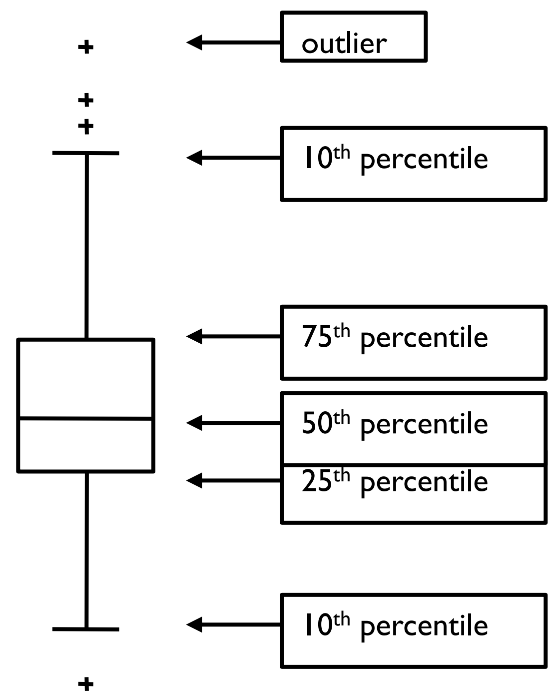
A box plot provides information about an attribute

- 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 .

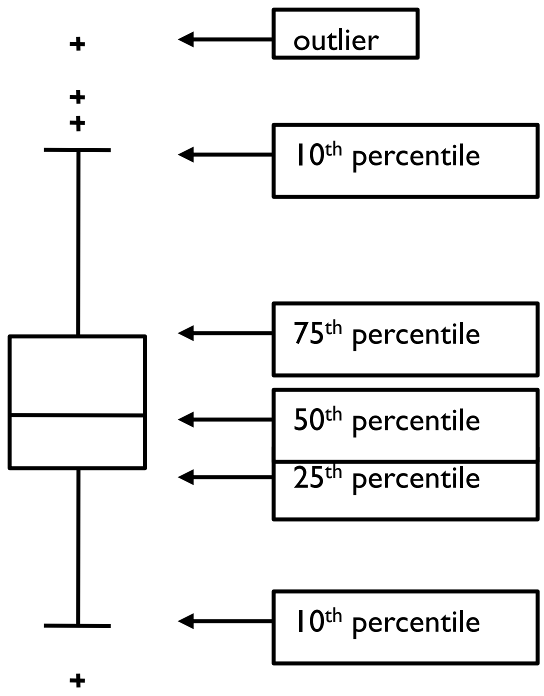
Box Plots Example

- Boxplots are a popular way of visualizing a distribution.
- A boxplot incorporates the five-number(*Minimum*,*Q*1, *Median*, *Q*3, *Maximum*) summary as follows:
  - 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×IQR beyond the quartiles.*
  - Otherwise, the whiskers terminate at the most extreme observations occurring within 1.5×*IQR* of the quartiles. The remaining cases are plotted individually.
- Boxplots can be used in the comparisons of several sets of compatible data.



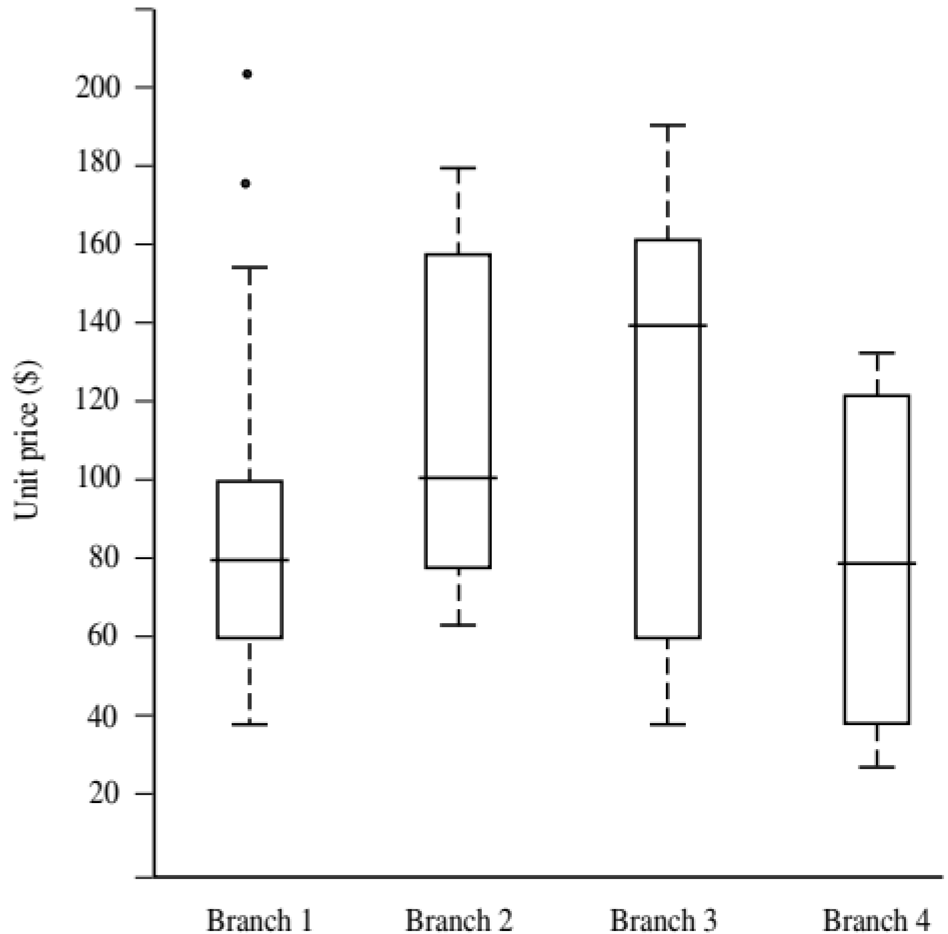


Figure 2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

- For details on box plot see page 54 of the book "Datamining Concepts and techniques".

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`

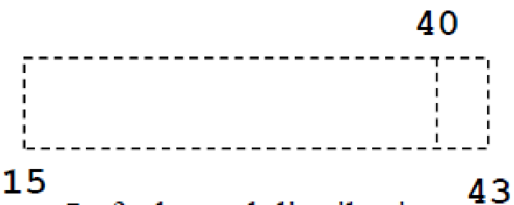
Sorted: `6 7 15 36 39 40 41 42 43 47 49`

$Q_1 = 15$ lower quartile

Q_2 = median = 40 (*mean = 33.18*)

$Q_3 = 43$ upper quartile

$Q_3 - Q_1 = 28$ interquartile range



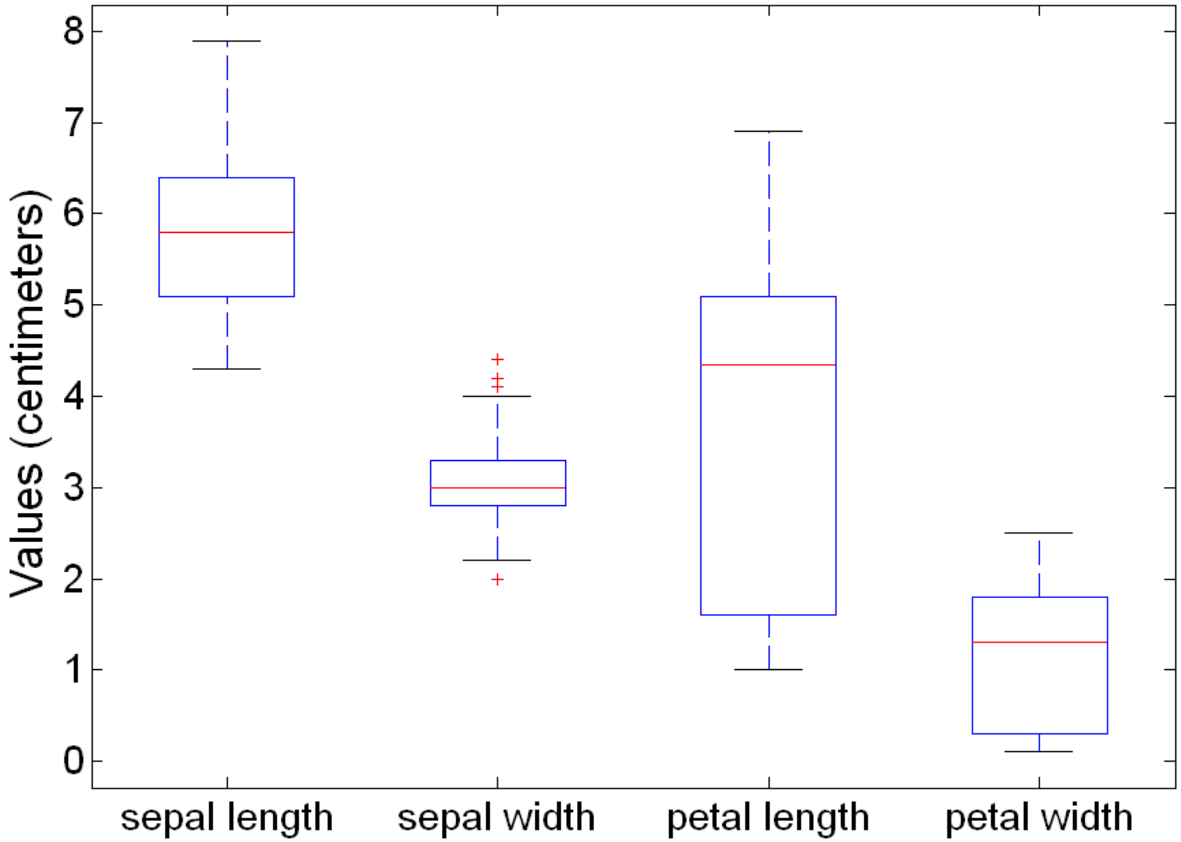
Left skewed distribution

Available in **WEKA**

- **`Filters`** **`InterquartileRange`**

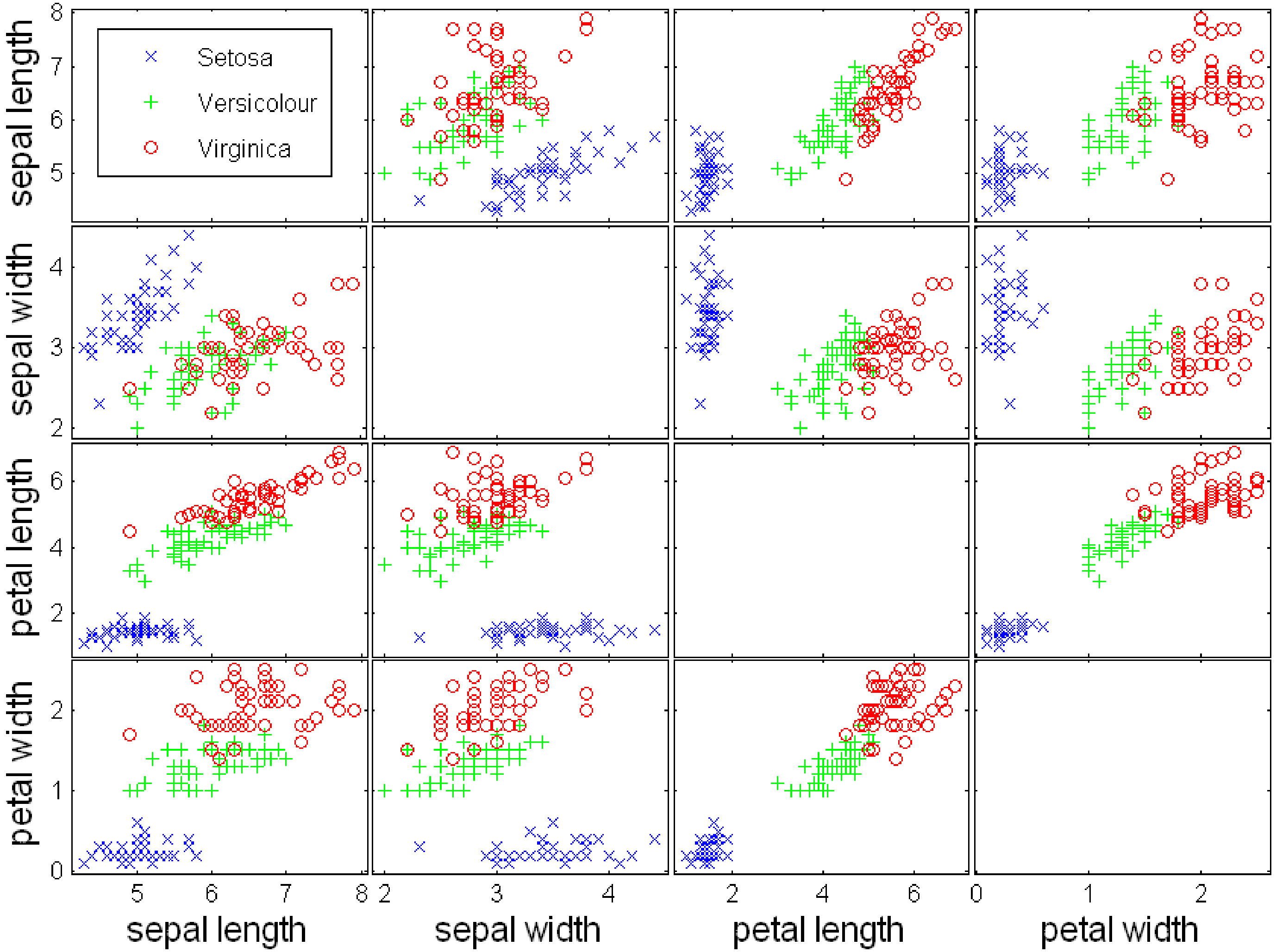
Example of Box Plots

- Box plots can be used to compare attributes



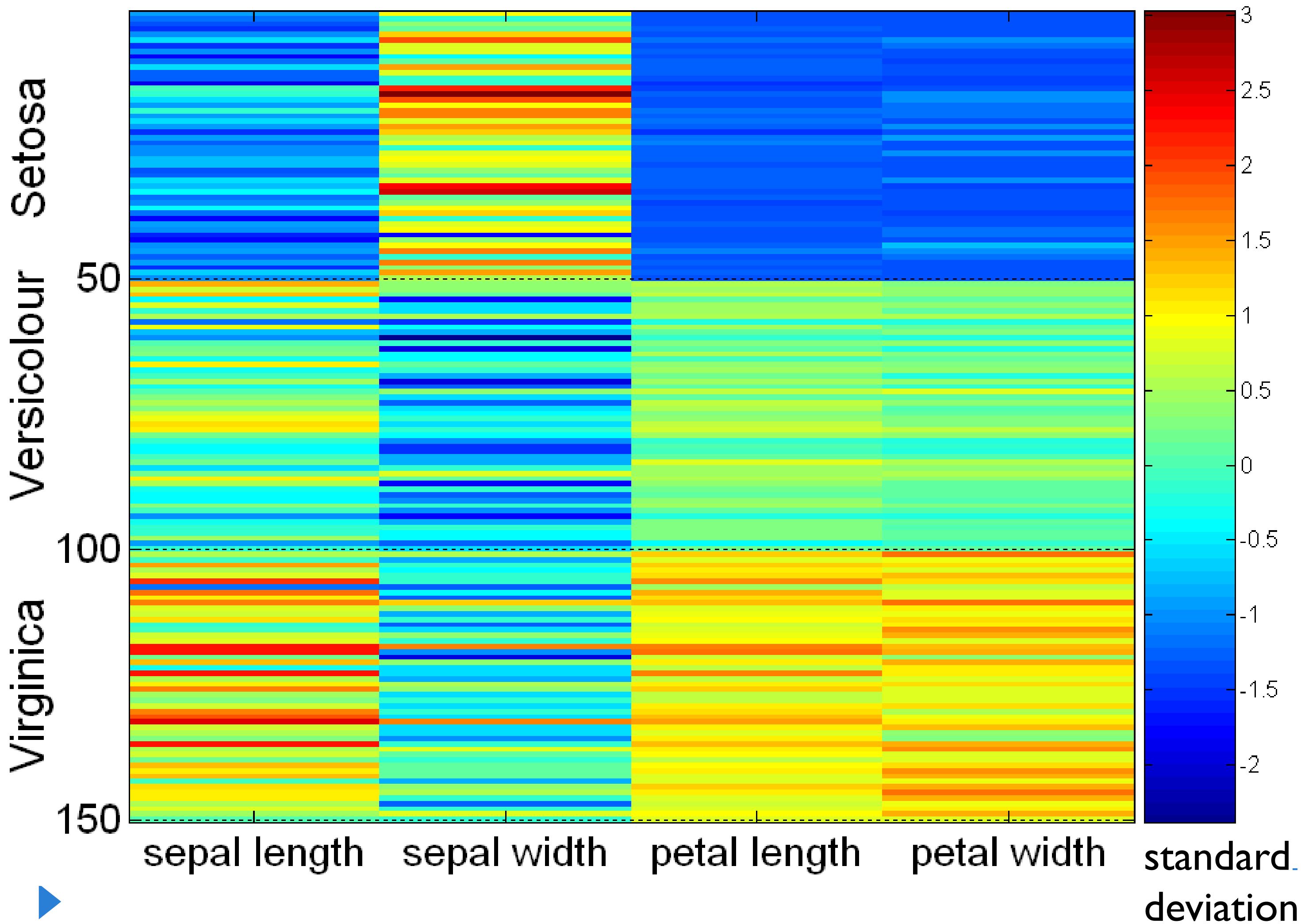
Visualization Techniques: Scatter Plots

Scatter Plot Array of Iris Attributes



Visualization Techniques: Matrix Plots

Visualization of the Iris Data Matrix



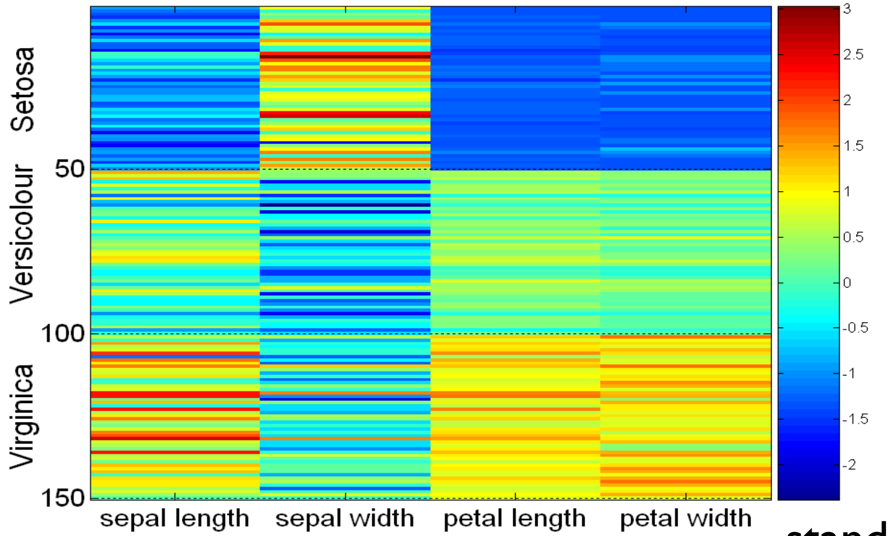
Sort objects **according to class**.

Help to detect if all objects in a same class have similar attribute value for some attribute

Column data are standardize to have a mean of 0 and standard deviation of 1

Visualization Techniques: Matrix Plots

- Matrix plots
  - Useful when objects are sorted according to class
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects



standard deviation

Visualization Techniques: Matrix Plots

Visualization of the Iris Correlation Matrix