Data and Data Exploration

Le Ou-Yang

Shenzhen University

Outline

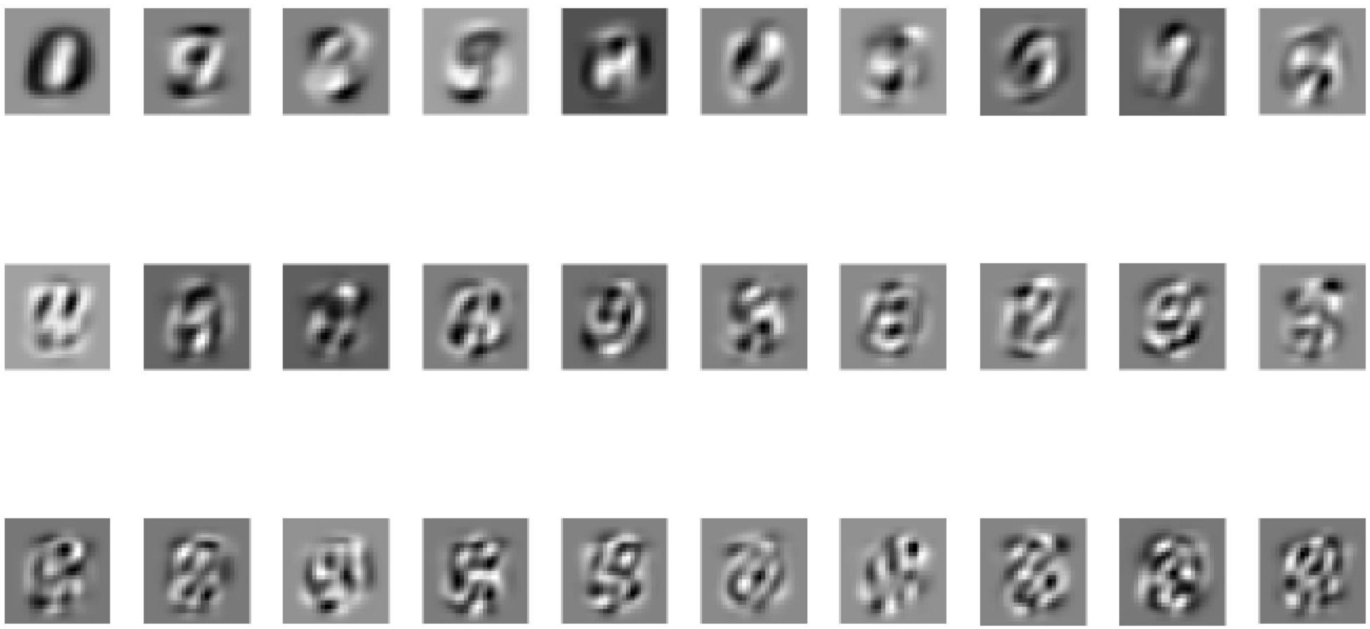
- Additional remarks on Principal component analysis
- Remaining Data Preprocessing
  1) Feature Subset Selection
  2) Attribute Transformation
- What is data exploration?
- Measure of Similarity & Dissimilarity

Principal Component Analysis

MNIST  $= a_1 w^1 + a_2 w^2 + \cdots$

images

30 components:



Eigen-digits

Principal Component Analysis

Face

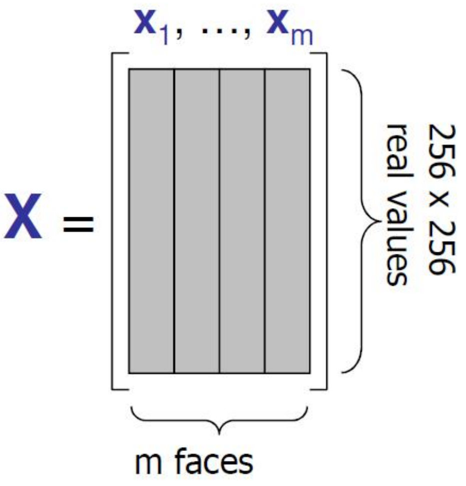
30 components:

http://www.cs.unc.edu/~lazebnik/research/spring08/assignment3.html

Eigen-face

Principal Component Analysis

- Example data set: Images of faces
  - Eigenface approach
    [Turk & Pentland], [Sirovich & Kirby]
- Each face $\mathbf{x}$ is …
  - 256×256 values (luminance at location)
  - $\mathbf{x}$ in $\Re^{256\times256}$ (view as 64K dim vector)
- Form $\mathbf{X} = [\ \mathbf{x}_1 , \ldots, \mathbf{x}_m\ ]$ **centered** data mtx
- Compute $\Sigma = \mathbf{X}\mathbf{X}^T$
- Problem: Σ is $64K \times 64K$ … HUGE!!!



Principal Component Analysis

- Suppose m instances, each of size N
  - Eigenfaces: $m=500$ faces, each of size $N=64K$
- Given $N \times N$ covariance matrix Σ, can compute
  - all N eigenvectors/eigenvalues in $O(N^3)$
  - first k eigenvectors/eigenvalues in $O(k\ N^2)$

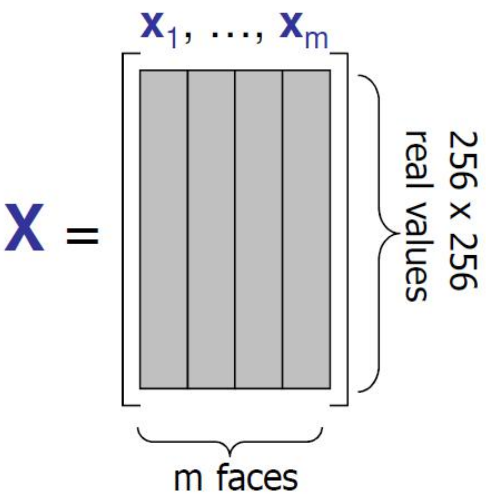
- But if $N=64K$, EXPENSIVE!

Principal Component Analysis

- Note that m<<64K
- Use $\mathbf{L}=\mathbf{X}^T\mathbf{X}$ instead of $\Sigma=\mathbf{X}\mathbf{X}^T$
- If $\mathbf{v}$ is eigenvector of $\mathbf{L}$
  then $\mathbf{Xv}$ is eigenvector of Σ

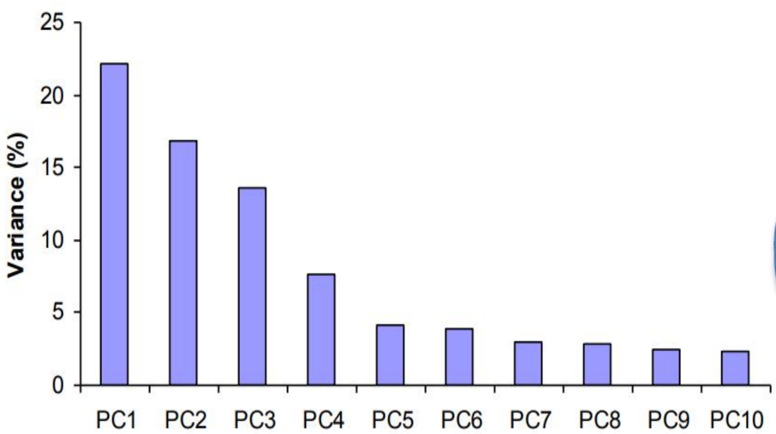
Proof:

$$\mathbf{L}\ \mathbf{v} = \gamma\ \mathbf{v}$$
$$\mathbf{X}^T\mathbf{X}\ \mathbf{v} = \gamma\ \mathbf{v}$$
$$\mathbf{X}\ (\mathbf{X}^T\mathbf{X}\ \mathbf{v}) = \mathbf{X}(\gamma\ \mathbf{v}) = \gamma\ \mathbf{X}\mathbf{v}$$
$$(\mathbf{X}\mathbf{X}^T)\mathbf{X}\ \mathbf{v} = \gamma\ (\mathbf{X}\mathbf{v})$$
$$\Sigma\ (\mathbf{X}\mathbf{v}) = \gamma\ (\mathbf{X}\mathbf{v})$$



Additional remarks

- How many PCs?
  - We want to retain as much information as possible using these components.
  - We can compute each PC explains how much variance and then makes decision (still a parameter)

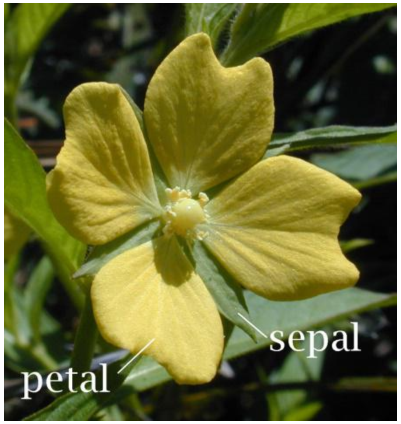


$$\frac{\lambda_k}{\sum_{i=1}^{N} \lambda_i}$$ Propotion of variance

$$\frac{\sum_{k=1}^{d} \lambda_k}{\sum_{i=1}^{N} \lambda_i}$$ Cumulative propotion

Example of a data: Iris Flower Data Set

- Many of the exploratory data techniques are illustrated with the famous ***Iris Flower*** data set (a.k.a. "**Iris**").
  - Available at the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician R.A. Fisher
  - Three flower types (classes):
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
  - Four (non-class) attributes
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length
  - Total number Instances = 150



https://en.wikipedia.org/wiki/Iris_flower_data_set

Iris setosa

Iris versicolor

Iris virginica

R Example using Iris data

- Iris

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- irisPCA<-**princomp**(iris[-5]) # Exclude Species and perform PCA
- summary(irisPCA)

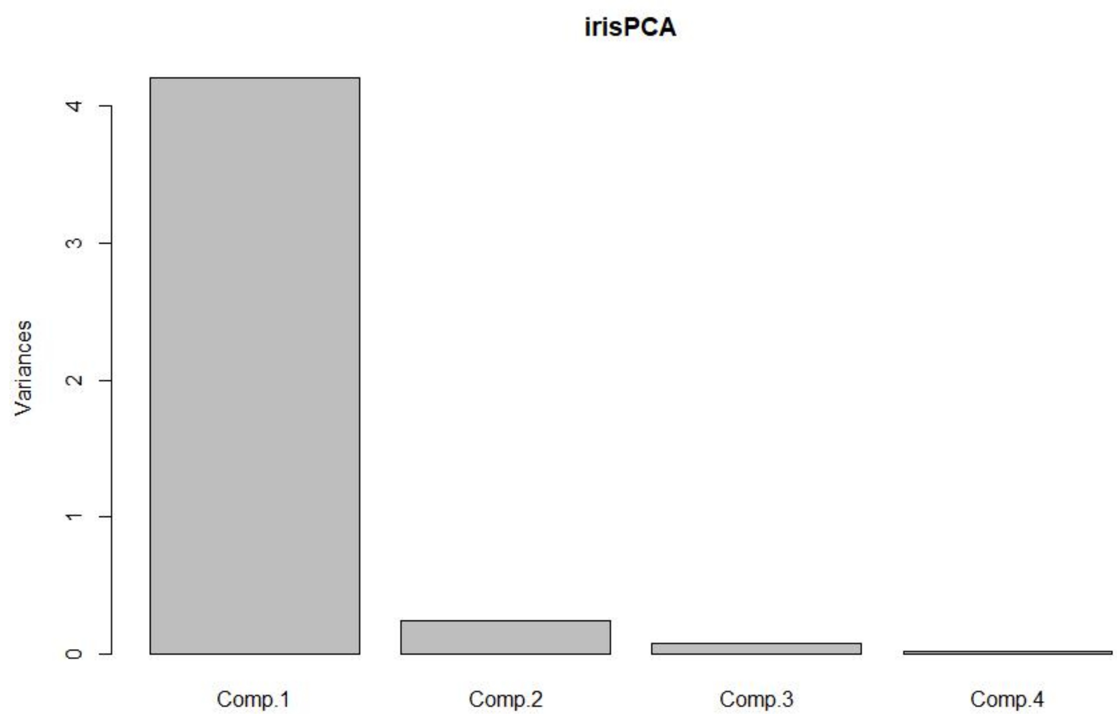
PC1 PC2 PC3 PC4

```
>
> summary(irisPCA)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
```

92.5% of variation is explained by PC1 alone; 97.8% is explained by PC1 and PC2

Screen plot

- It shows the proportion of the total variation that is explained by each of the components. Perhaps 1 or 2 PC2 will be sufficient
- screeplot(irisPCA)



Principal Component Analysis

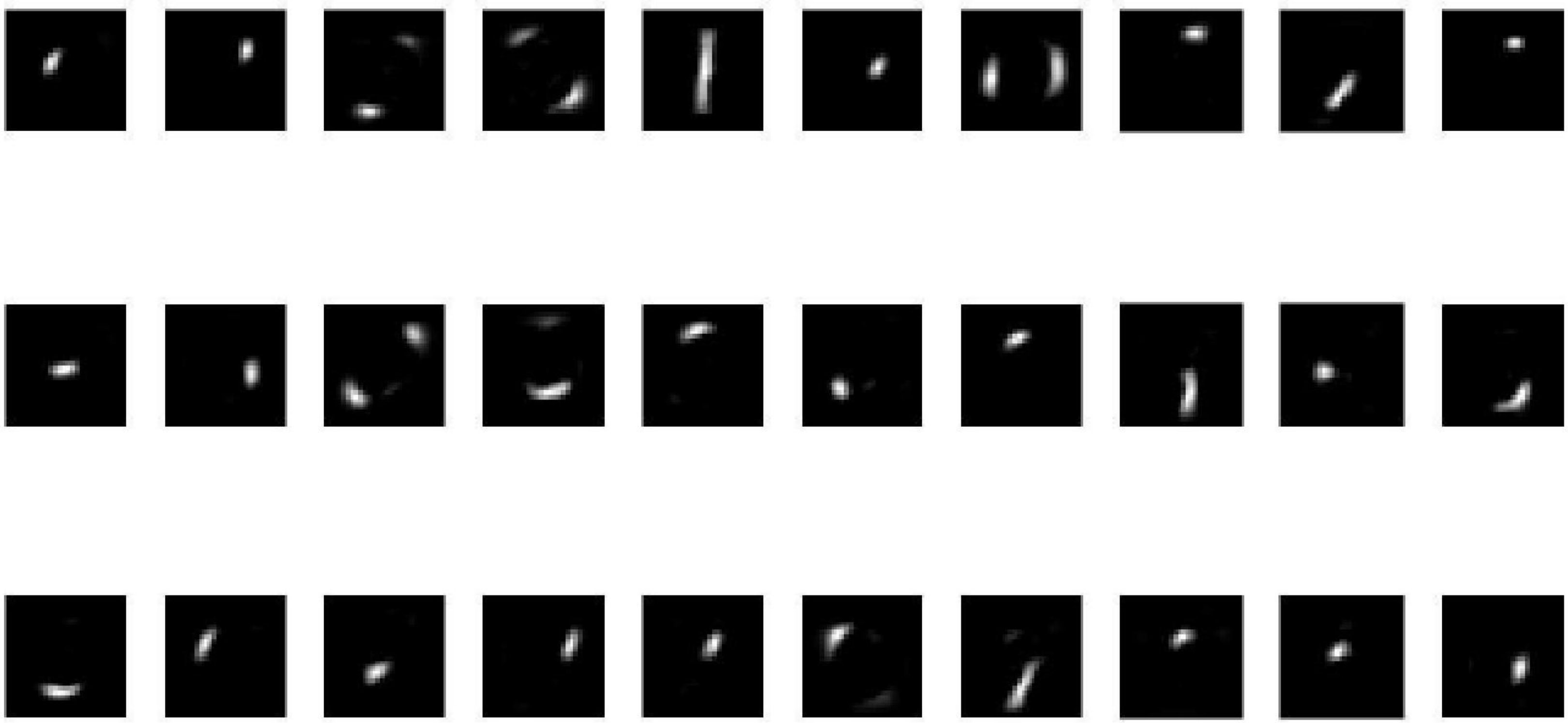
$$= a_1 w^1 + a_2 w^2 + \cdots$$

Can be any real number

- PCA involves adding up and subtracting some components (images)
  - Then the components may not be "parts of digits"
- Non-negative matrix factorization (NMF)
  - Forcing a_1, a_2 ...... be non-negative
    - additive combination
  - Forcing w^1, w^2...... be non-negative
    - More like "parts of digits"
- Ref: Daniel D. Lee and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001.

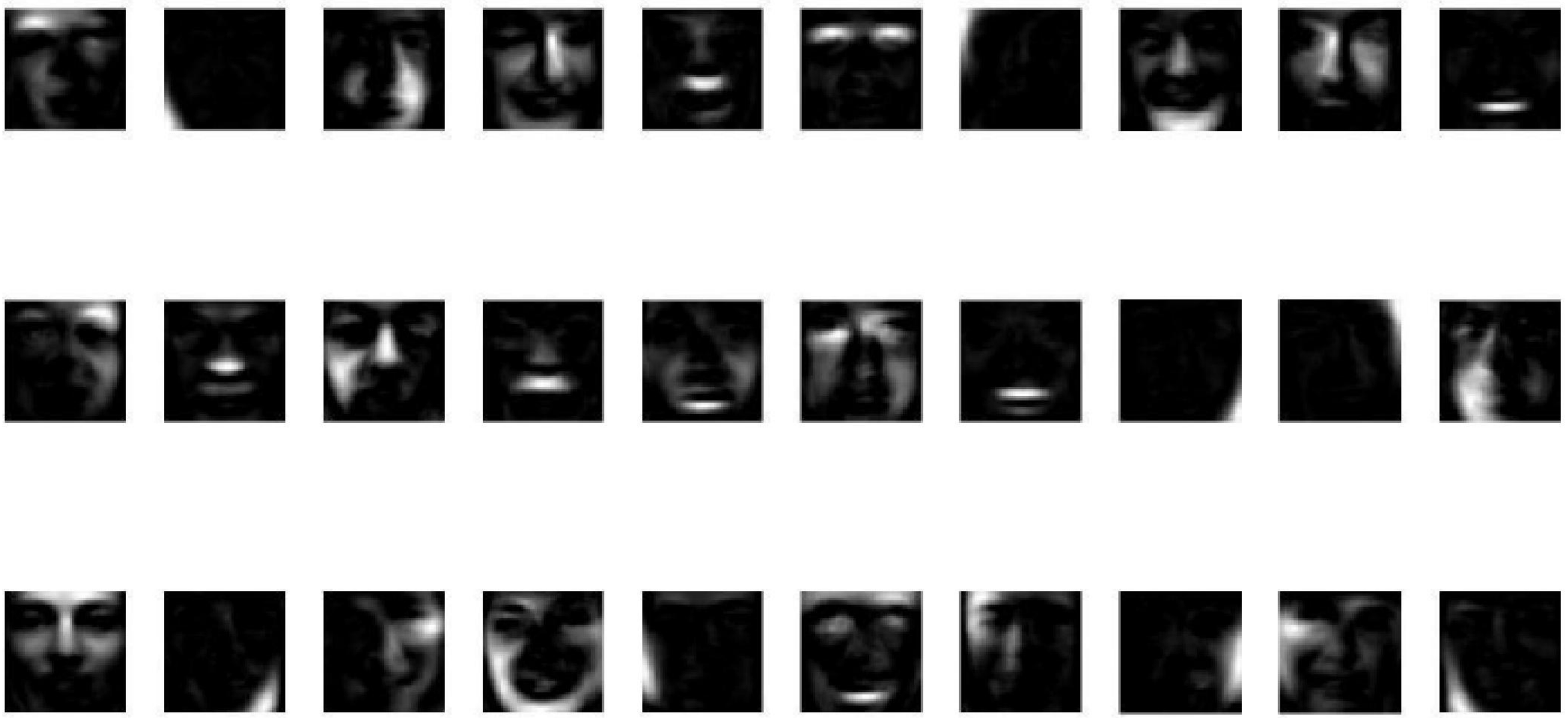
Non-negative Matrix Factorization

NMF on MNIST



Non-negative Matrix Factorization

NMF on Face



Outline

- Additional remarks on Principal component analysis
- Remaining Data Preprocessing
    1) Feature Subset Selection
    2) Attribute Transformation
- What is data exploration?
- Measure of Similarity & Dissimilarity

Feature Subset Selection (FSS)

- PCA maps data into different dimensions which is somewhat hard to explain.
- FSS is another way to reduce dimensionality of data
- Redundant features
  - Example: **purchase price** of a product /services/dinner and the amount of **sales tax paid**
- Irrelevant features
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Selection

- Feature Selection is a process that *chooses an optimal subset of features* according to a certain criterion.
- Why we need FS:
  - To reduce dimensionality, noise and complexity
  - to improve performance.
  - to visualize the data for model selection.
  - to improve the model understandability

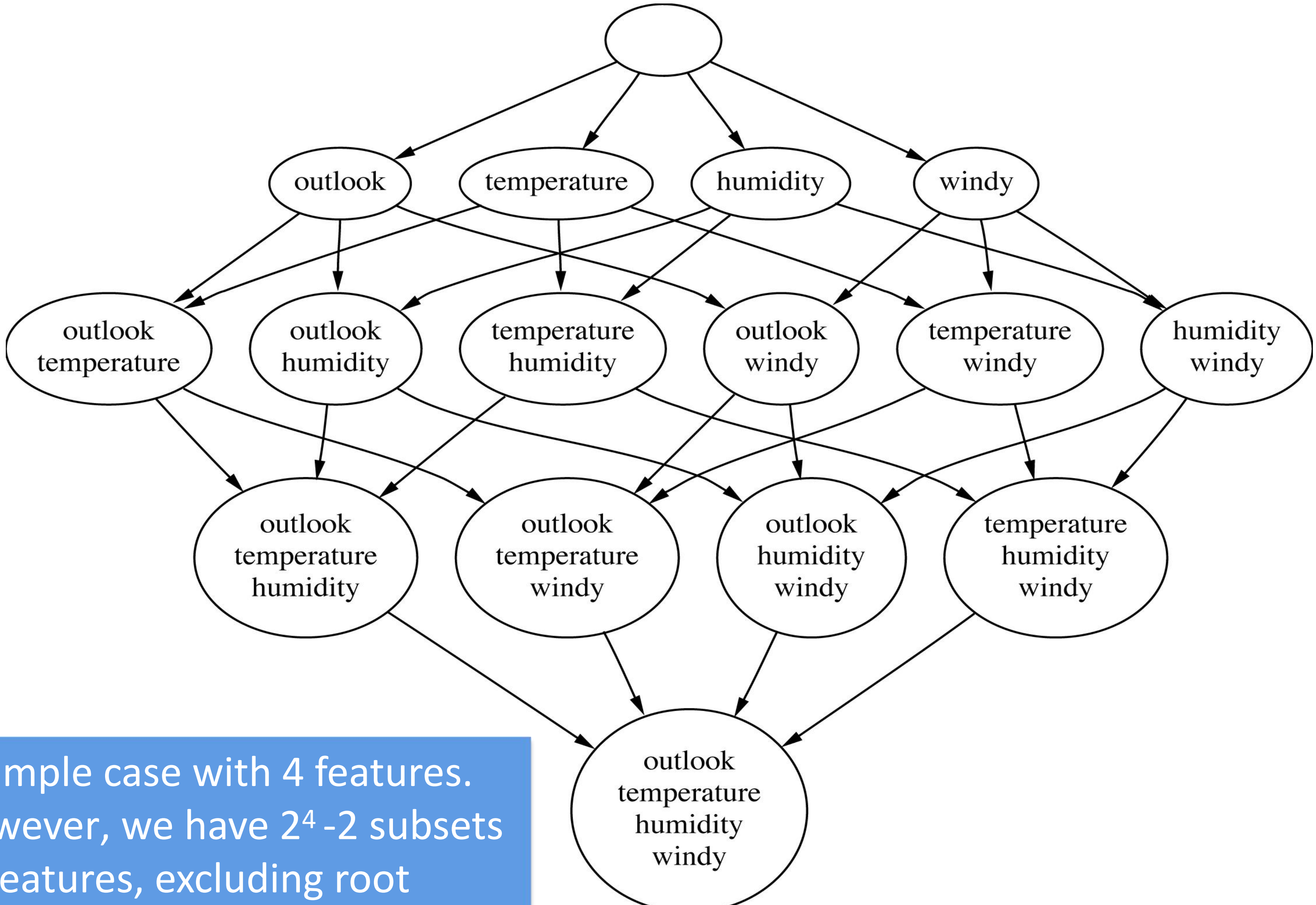
Feature Subset Selection

- Techniques:
  - Brute-force approach:
    - Try all possible feature subsets as input to machine learning algorithm. **Number of features could be huge**!

Weather Data

	outlook	temperature	humidity	windy	play
1	outlook	temperature	humidity	windy	play
2	sunny	hot	high	FALSE	no
3	sunny	hot	high	TRUE	no
4	overcast	hot	high	FALSE	yes
5	rainy	mild	high	FALSE	yes
6	rainy	cool	normal	FALSE	yes

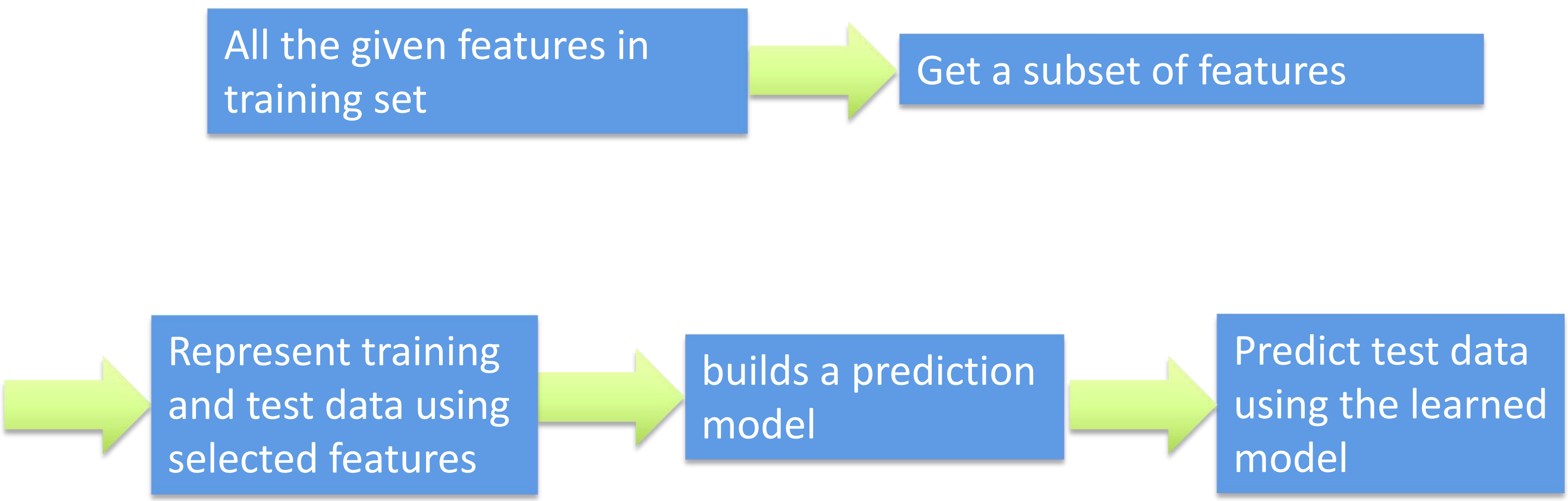
Attribute subsets for weather data



A simple case with 4 features. However, we have 2^4 -2 subsets of features, excluding root (empty set) and leave (full set)

Feature Subset Selection

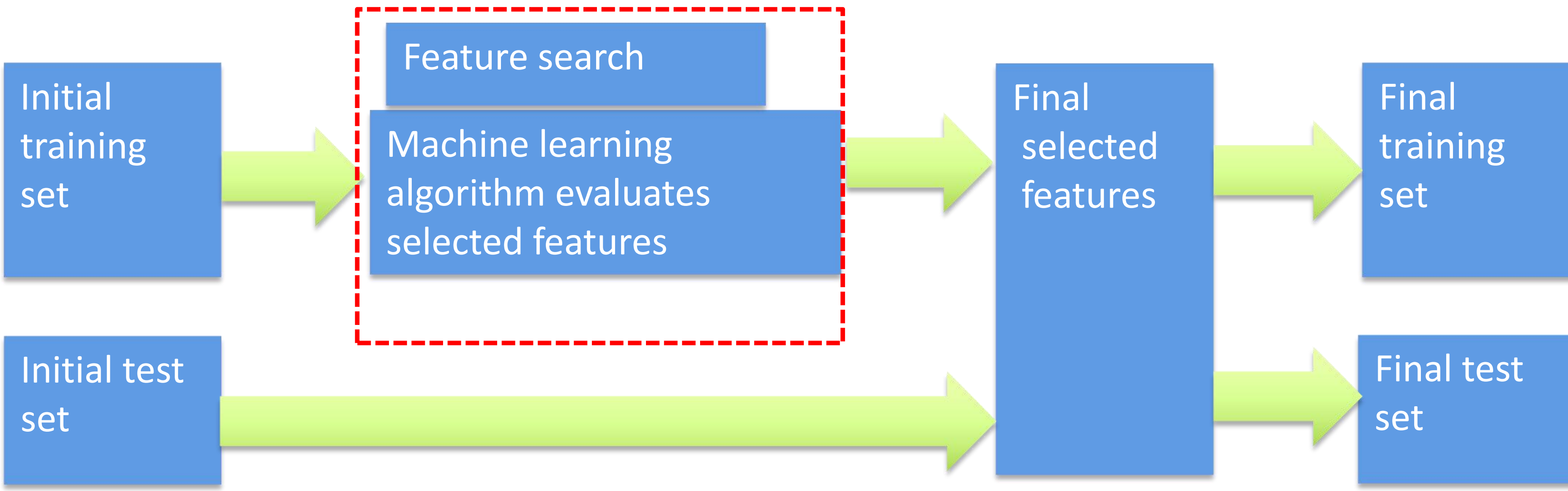
- Techniques:
  - Filter approaches:
    - Features are selected **before** machine learning algorithm is run



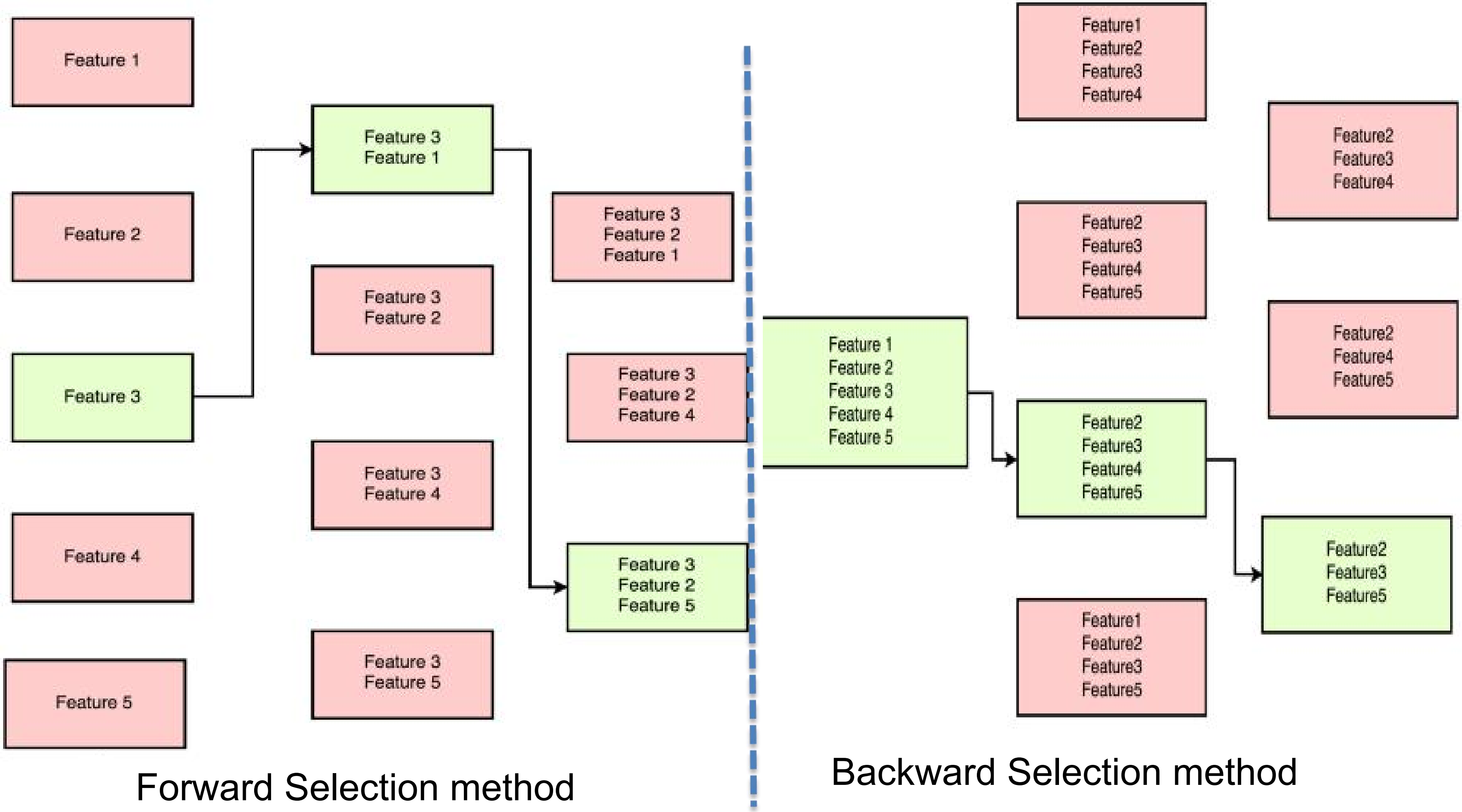
Feature Subset Selection

– Embedded approaches:
  - Feature selection occurs **naturally as part** of the machine learning algorithm, e.g. C4.5. We select best features (e.g. using information gain) to build a tree in top-down fashion

– Wrapper approaches:
  - Use a machine learning algorithm as a **black box** (compute accuracy) to find best subset of attributes



Feature Search
Common greedy approaches



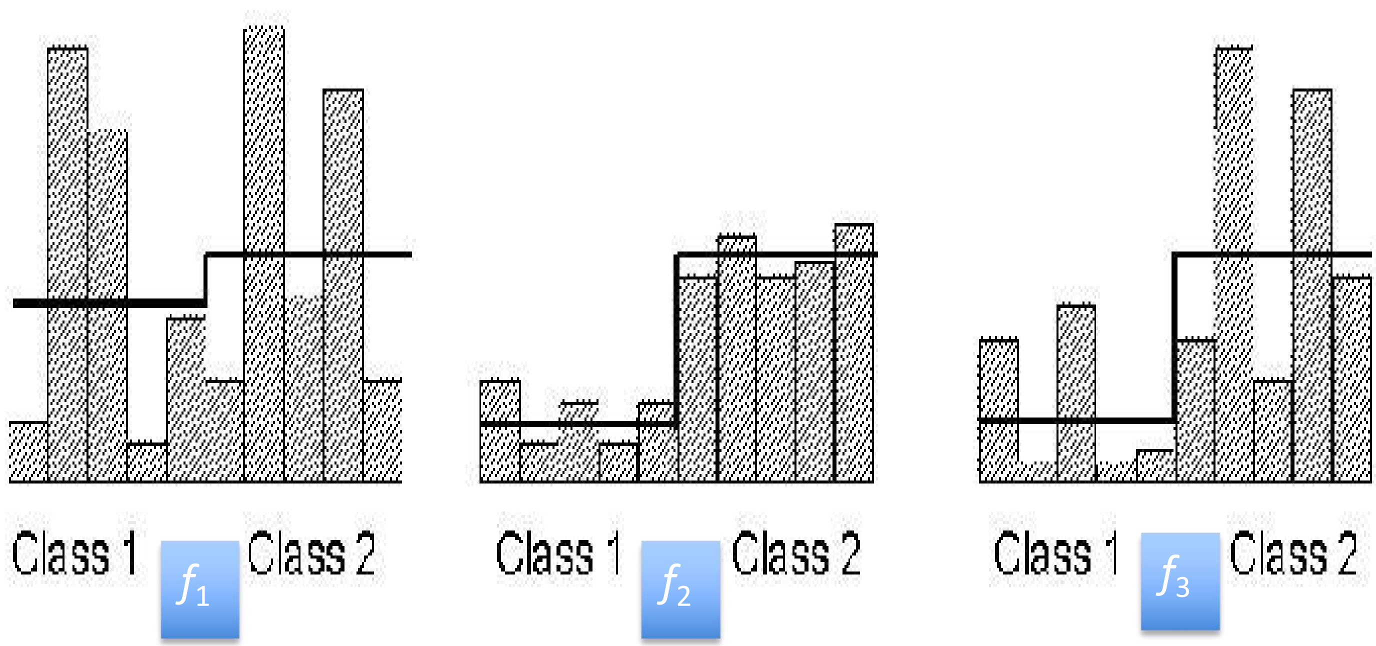
Two figures are from https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc

One Example of Feature/Signal Selection

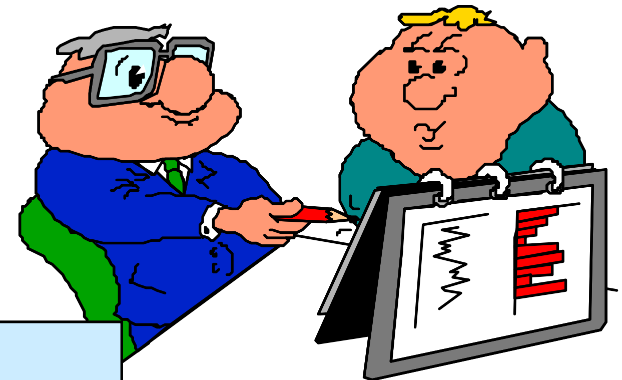
- Given a sample space of p dimensions
- It is possible that some dimensions are irrelevant or less important
- Need to find ways to separate those dimensions that are relevant from those that are irrelevant

Signal Selection (Basic Idea)

- Choose a feature with low *intra-class distance* (*variance* is smaller)
- Choose a feature with high *inter-class distance* (*mean* difference is bigger)
- Given features f_1, f_2 and f_3 for binary classification task (Class 1 and 2), which feature is the best?



Signal Selection (*t*-statistics/ *t*-test)



Formulate H_0 and H_1

Select Appropriate Test

Choose Level of Significance

Calculate Test Statistic

Determine Prob Assoc with Test Stat

Compare with Level of Significance, α

Determine Critical Value of Test Stat

Determine if Test Stat falls into (Non) Rejection Region

Reject/Do not Reject H_0

Draw Conclusion

Signal Selection (*t*-statistics/ *t*-test)

Controlled via sample size (=1-Power of test)

Typically restrict to a 5% Risk = level of significance

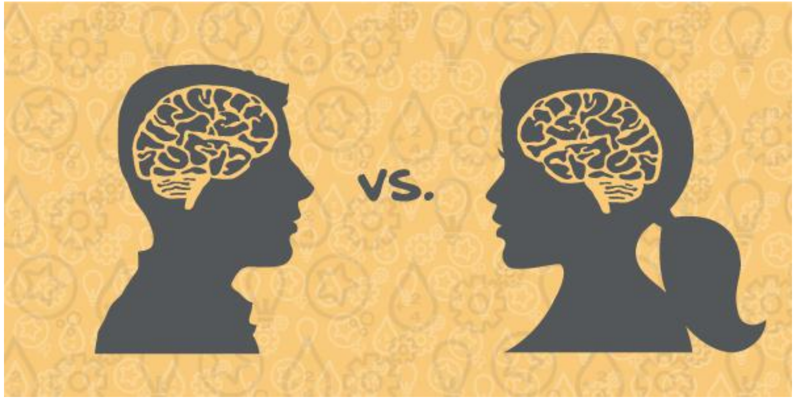
	Study reports **NO** difference (Do not reject H_0)	Study reports **IS** a difference (Reject H_0)
H_0 is true Difference Does **NOT** exist in population	✔	X Type I Error
H_1 is true Difference **DOES** exist in population	X Type II Error	✔

Prob of this = Power of test

Signal Selection (*t*-statistics/ *t*-test)

- The *t* test is one type of inferential, parametric statistic

- Determine whether there is a statistically significant difference between the means of two groups



- **It is also known as independent samples t-test, two sample t-tests, between samples t-test and unpaired samples t-test.**

Signal Selection (*t*-statistics/ *t*-test)

Given $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$

Collect $\{X_{11}, X_{12}, \cdots, X_{1n_1}\}$ $\{X_{21}, X_{22}, \cdots, X_{2n_2}\}$

Calculate $\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2}, \quad s_2 = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}$$

Case 1: for unknown $\sigma^2 = \sigma_1^2 = \sigma_2^2$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, (\frac{1}{n_1} + \frac{1}{n_2})\sigma^2\right)$$

$$\rightarrow \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Signal Selection (*t*-statistics/ *t*-test)

$$\frac{(n_1-1)s_1^2}{\sigma^2}+\frac{(n_2-1)s_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}=\sqrt{\frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}}$$

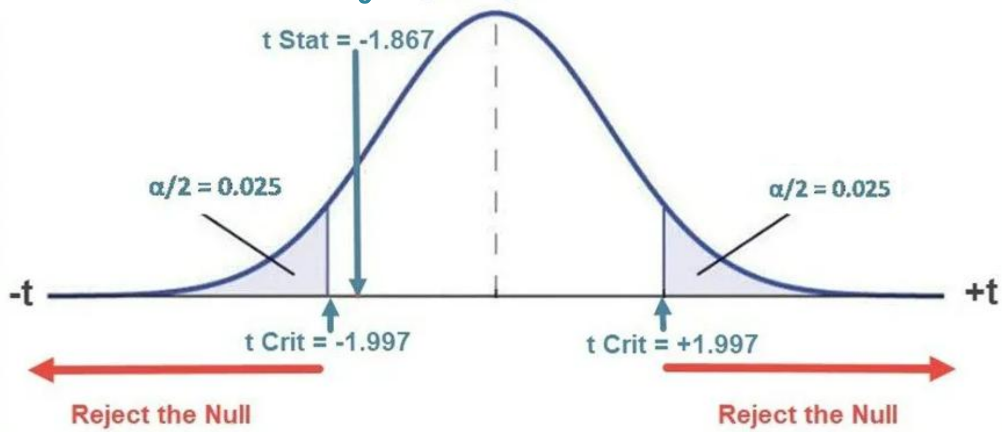
Case 2: for completely unknown

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Signal Selection (*t*-statistics/ *t*-test)

	one-tailed test		two-tailed test
hypothesis	$H_0 : \mu_1 \geqslant \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leqslant \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
test statistic (t distribution)	$t = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$		$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
deg. of freedom	n_1+n_2-2		
rejection	reject H_0 if $t < -t_\alpha$	reject H_0 if $t > t_\alpha$	reject H_0 if $\lvert t \rvert > t_{\alpha/2}$

Two-tailed test H_0: $\mu_1 \neq \mu_2$



Outline

- Additional remarks on Principal component analysis
- Remaining Data Preprocessing
  1) Feature Subset Selection
  2) Attribute Transformation
- What is data exploration?
- Measure of Similarity & Dissimilarity

Attribute/Variable Transformation

- **A function** that maps the *entire set of values* of a given attribute to **a new set** of replacement values via certain math functions (an original value as input to generate a new value)
- Simple math functions: v^k, *log*(v), e^v, |v|, 1/v, *sin* v
  - Could be scale down/up
  - Normalization (or Standardization)

Normalization (frequently used)

- Min-max normalization:
  - $[min_A, max_A]$ ---> $[new\_min_A, new\_max_A]$

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

  - Example:

    Annual income range [12,000, 300,000] normalized to [0.0, 1.0]. Then 73,000 is mapped to

$$\frac{\boxed{73{,}000} - 12{,}000}{300{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.21$$

$$\frac{\boxed{12{,}000} - 12{,}000}{300{,}000 - 12{,}000}(1.0 - 0) + 0 = 0 \qquad \frac{\boxed{300{,}000} - 12{,}000}{300{,}000 - 12{,}000}(1.0 - 0) + 0 = 1$$

Normalization (cont)

- Z-score normalization

  (μ_A: mean, σ_A: standard deviation): $v' = \dfrac{v - \mu_A}{\sigma_A}$

- Example: Consider a value v=73,000,

- Let μ_A = 54,000, σ_A= 16,000. Then $\dfrac{73{,}000 - 54{,}000}{16{,}000} = 1.225$

- Normalization by Decimal Scaling

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

here *j* is the smallest integer such that Max(|v'|) < =1

1, 10, 100, 1000 ->1/10^3, 10/10^3 , 100/10^3 , 1000/10^3 (Here j=3; If we use j=4, then it will not be the smallest integer)

Outline

- Additional remarks on Principal component analysis
- Remaining Data Preprocessing
  1) Feature Subset Selection
  2) Attribute Transformation
- What is data exploration?
- Measure of Similarity & Dissimilarity

What is data exploration?

- **A preliminary exploration of the data to better understand its characteristics.**
- **In our discussion of data exploration, we focus on**
  - Summary statistics
    - Summarize the properties of the data
  - Visualization
    - Making use of humans' abilities to recognize patterns

Example of a data:
Iris Flower Data Set

- Many of the exploratory data techniques are illustrated with the famous ***Iris Flower*** data set (a.k.a. ``**Iris**”).
  - Available at the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician R.A. Fisher
  - Three flower types (classes):
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
  - Four (non-class) attributes
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length
  - Total number Instances = 150

Iris setosa

Iris versicolor

Iris virginica

1. Summary Statistics

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include
    - ***Frequency***, ***location***, and ***spread***
    - Examples: Location – mean / median
Spread – standard deviation
  - Most summary statistics can be calculated in a single pass through the data

Measures of Location: Mean and Median

- Suppose I have data x_1 x_2, …, x_m
- The *mean* is the most common measure of the location of a set of points.

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

- However, the *mean* is very sensitive to outliers (some ppl feel their salaries are not as high as the averaged salary)
- Thus, the *median* or a *trimmed* mean is also used:

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

Trimmed-*n* mean: take out the smallest ones and largest ones, and then calculate mean on the remaining numbers.

Measures of Spread: Range and Variance

- *Range* is the difference between the **max** and **min**
- The *variance* or *standard deviation* is the most common measure of the **spread** of a set of points.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- However, this is also sensitive to outliers, other measures are often used.

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Frequency and Mode

- The ***frequency*** of an attribute value is the ***percentage of time*** the value occurs in the data set, e.g., given m samples, the attribute value is selected from $\{v_1, \cdots, v_i, \cdots, v_k\}$

$$frequency(v_i) = \frac{N(v_i)}{m}$$

  - where $N(v_i)$ denotes the number of samples that have value v_i .

- The ***mode*** of an attribute is the most **frequent attribute *value***
- The notions of ***frequency*** and ***mode*** are typically used with categorical data

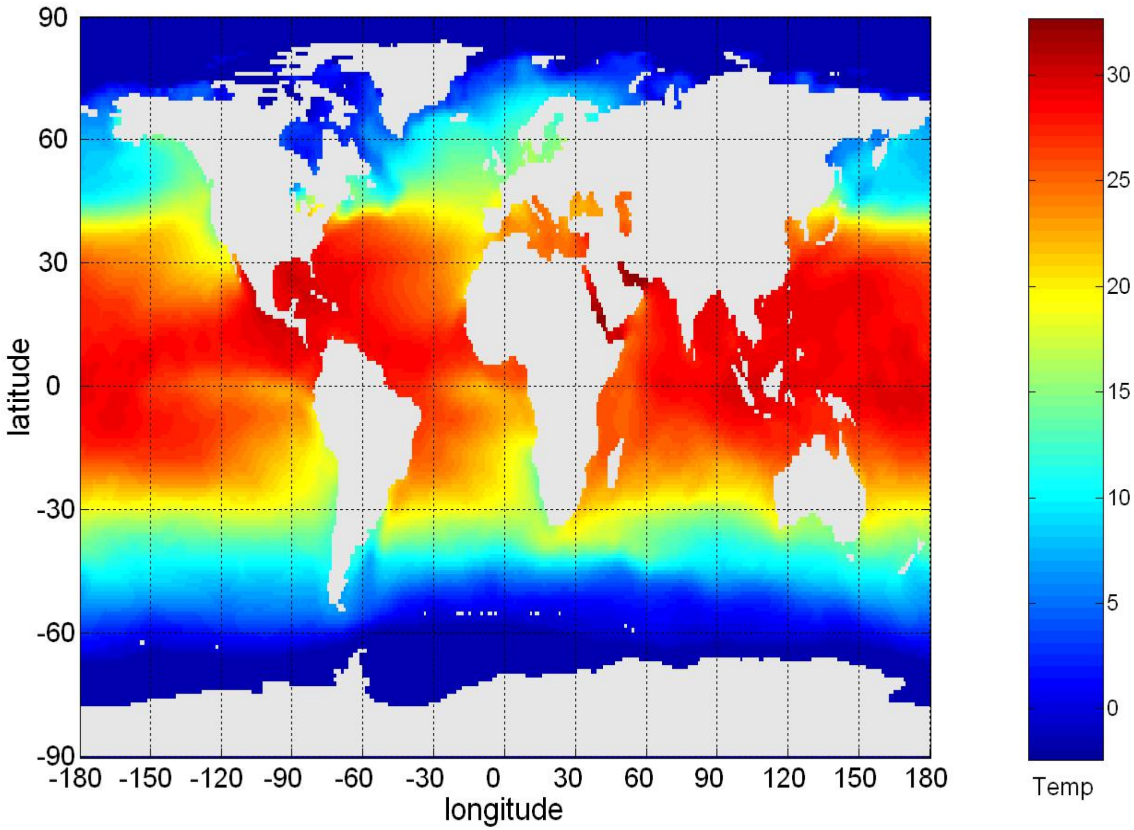
2 Visualization

- **Visualization** is the conversion of data into a visual or tabular **format** so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

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

Example: Sea Surface Temperature Data->picture->story

- Below shows the Sea Surface Temperature (SST) for July 1982. Tens of thousands of data points are summarized in a single figure



Summarizes information from approximately 250,000 numbers and is readily interpreted in a few seconds.

Representation

- The first step of visualization: the mapping of **information** to a **visual** format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

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

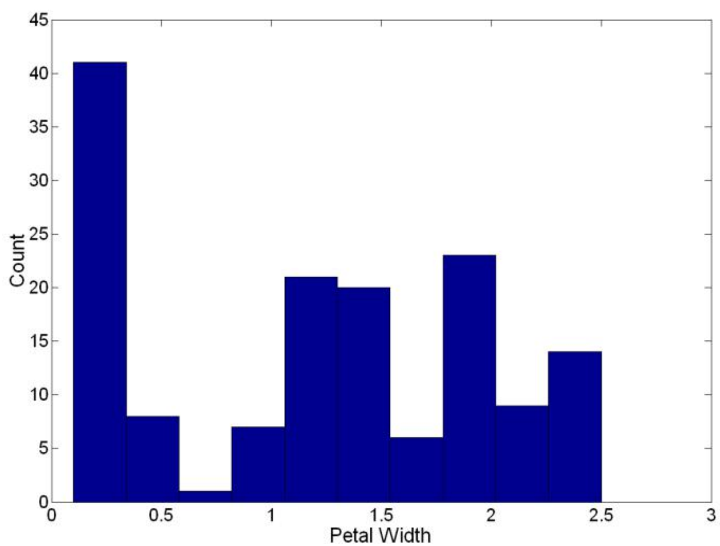
Same data
Re-arrange the sequence of rows and columns

Selection

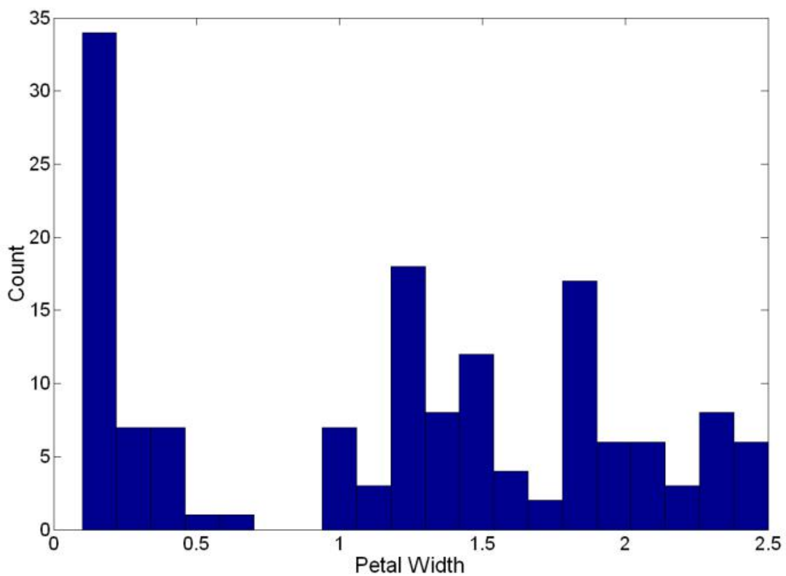
- Is the **elimination** or the **de-emphasis** of certain objects and attributes
- Selection may involve choosing a **subset of attributes**
  - Commonly, pairs of attributes are considered
  - Sophisticatedly, **dimensionality reduction** is often used to reduce the number of dimensions to *two or three*
- Selection may also involve choosing a **subset of objects**
  - A region of the screen can only show so many points

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
  - Shape of histogram depends on the number of bins
- **Example:** Iris data set - **Petal Width** (10 and 20 bins, respectively)



large bin (10 bins)

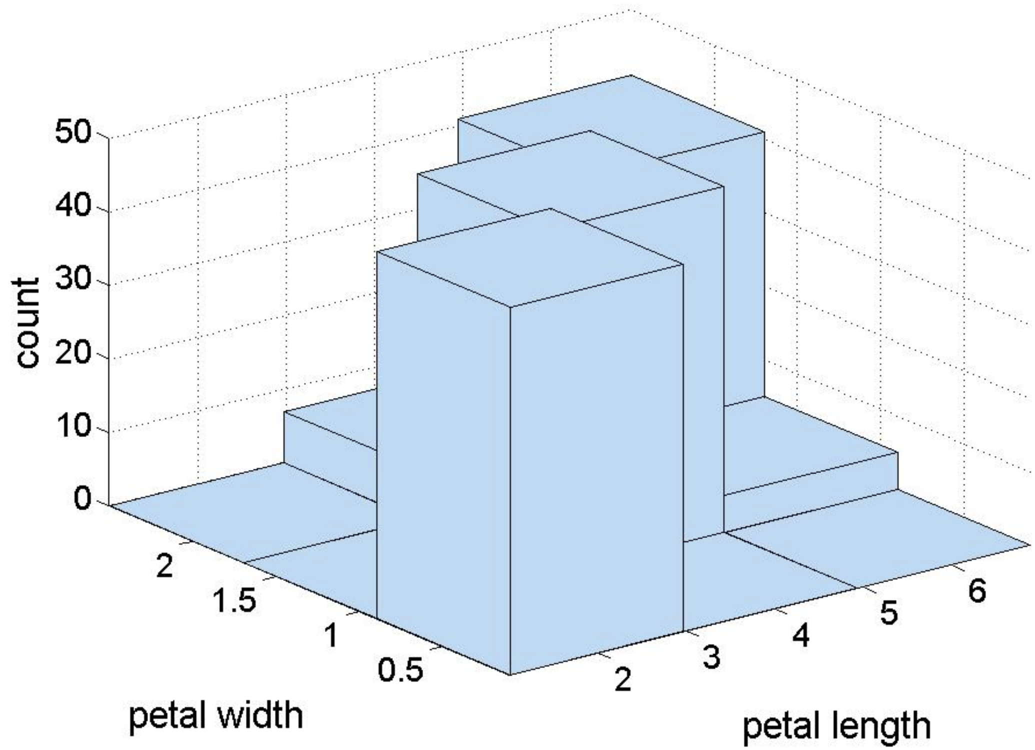


Small bin (20 bins)

Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example:
  - Petal width and Petal length

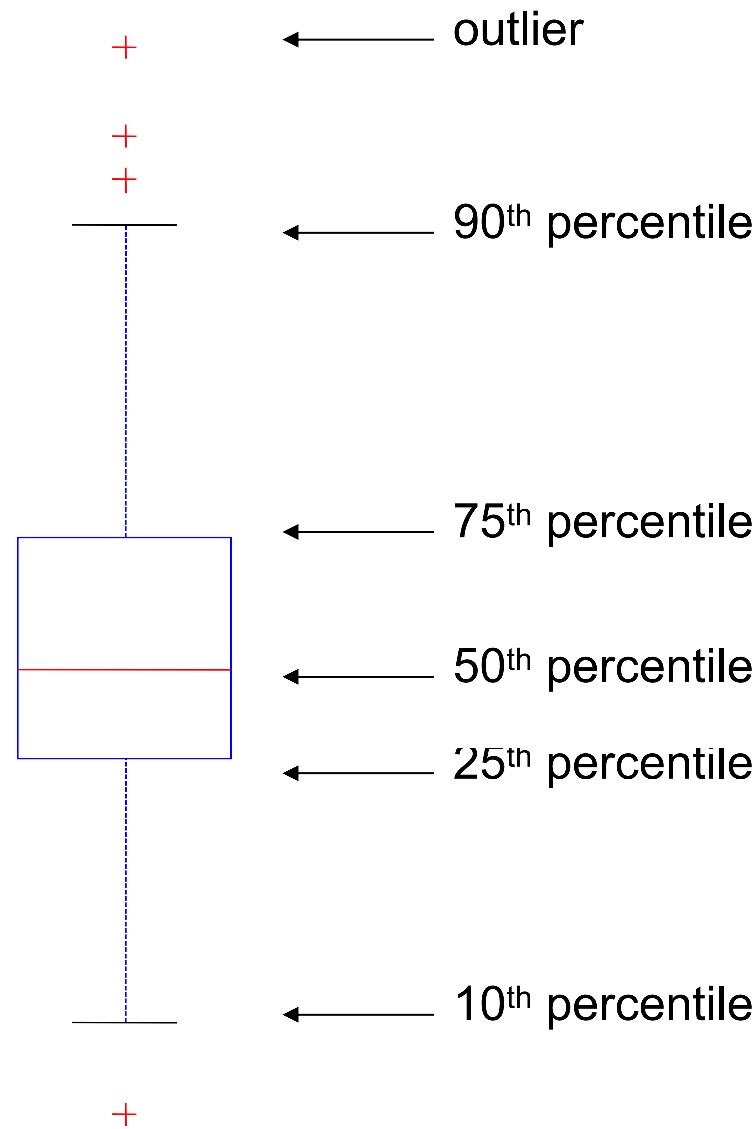
- What does this tell us?



http://en.wikipedia.org/wiki/Iris_flower_data_set

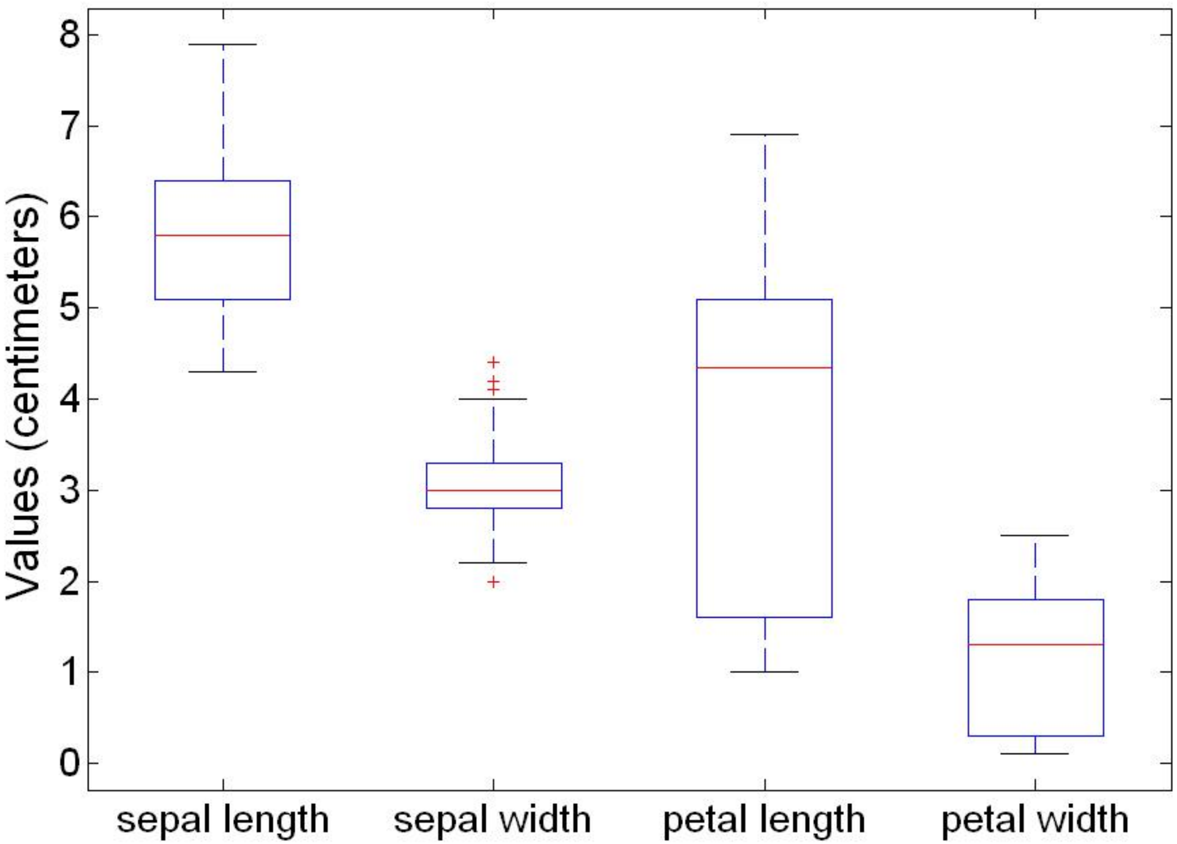
Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - The figure shows the basic part of a box plot



Example of Box Plots

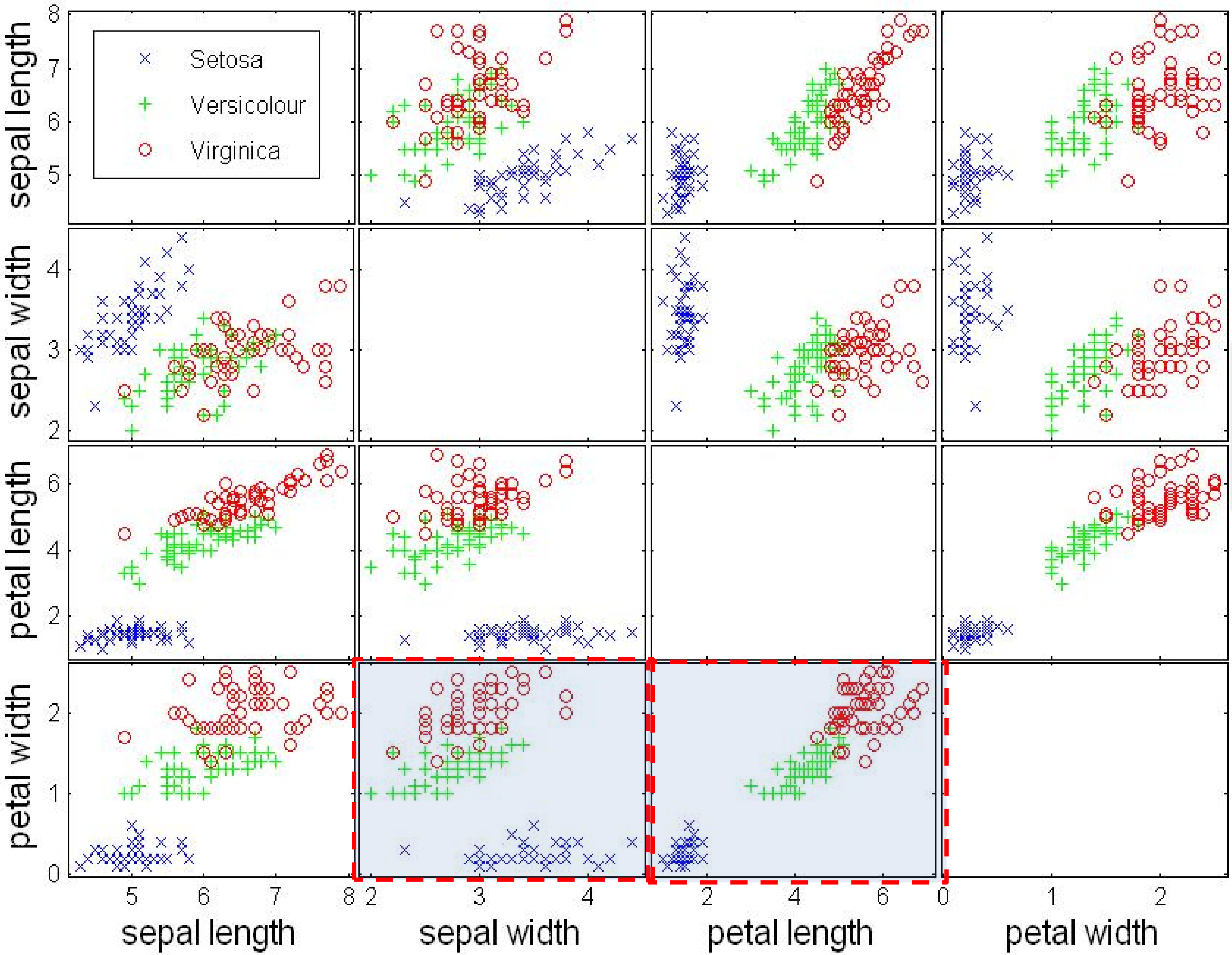
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots are most common, but we can have three-dimensional scatter plots
  - Additional attributes often can be displayed by using the *size*, *shape*, and *color* of the markers that represent the objects
  - It is useful to have arrays of scatter plots that can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide

Scatter Plot Array of Iris Attributes

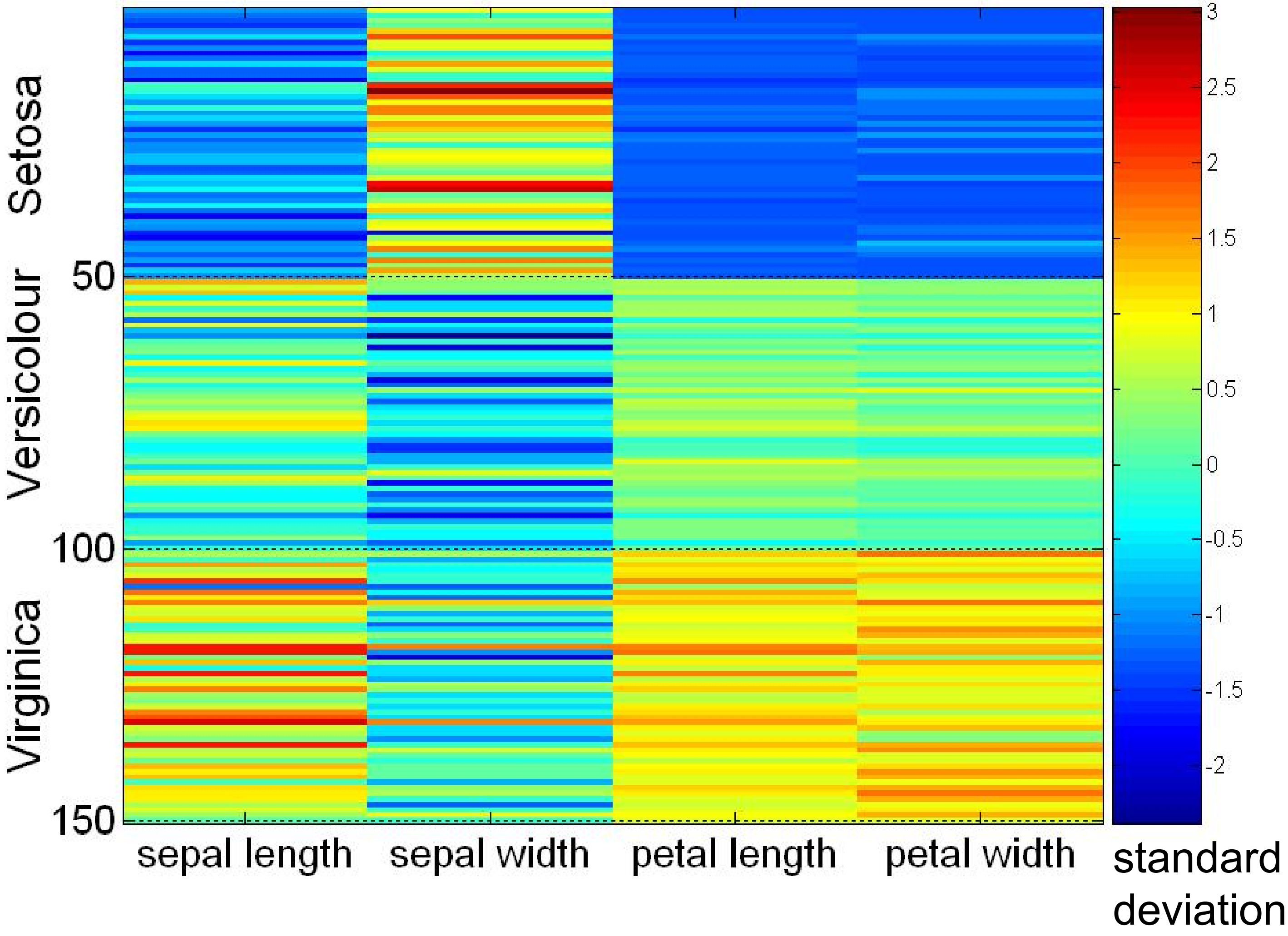


1. Correlations
2. Class distribution

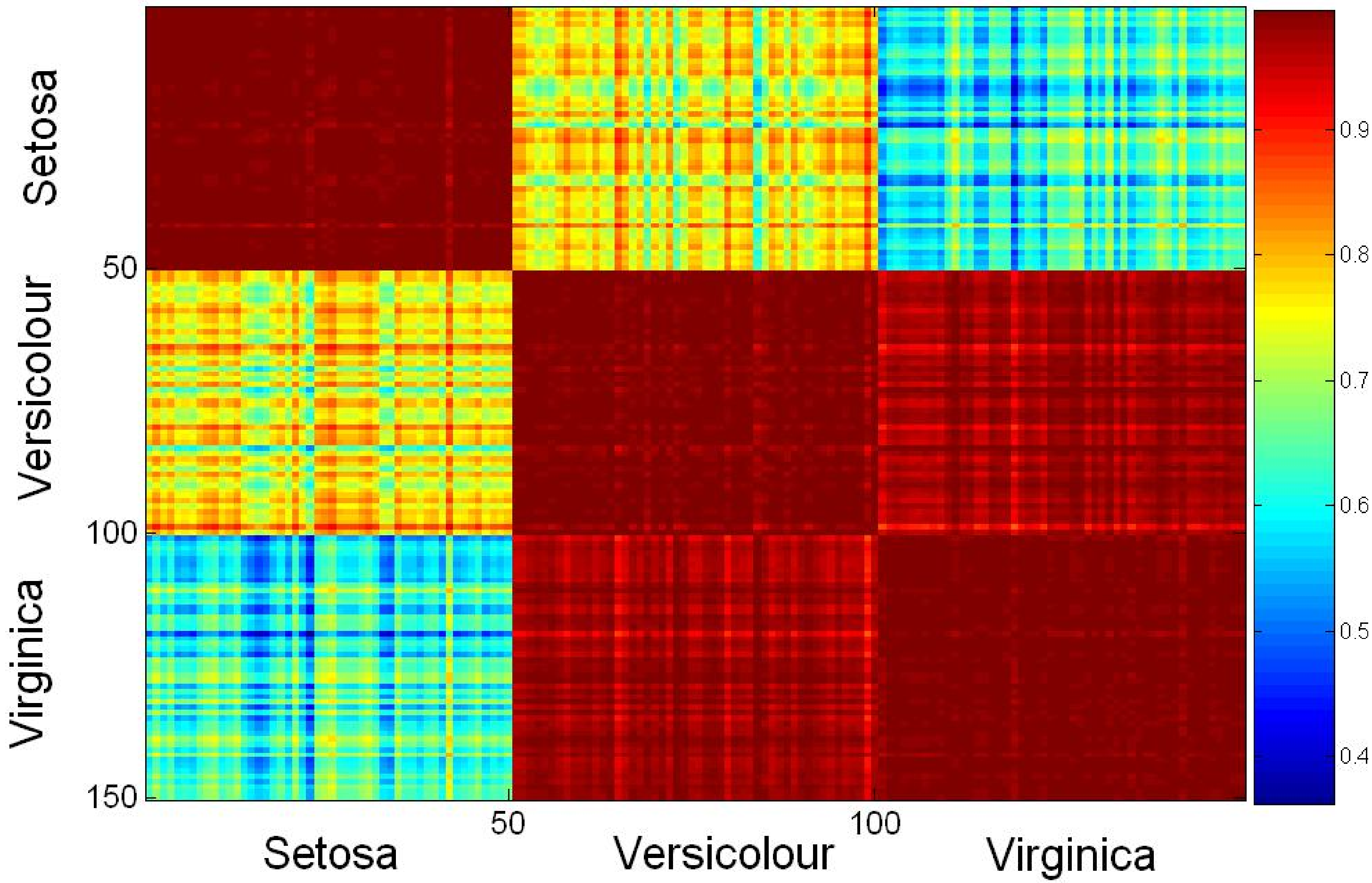
Visualization Techniques: Matrix Plots

- Matrix plots
  - Can plot the data matrix (all the data)
  - This can be 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
  - Examples of matrix plots are presented on the next slide

Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix

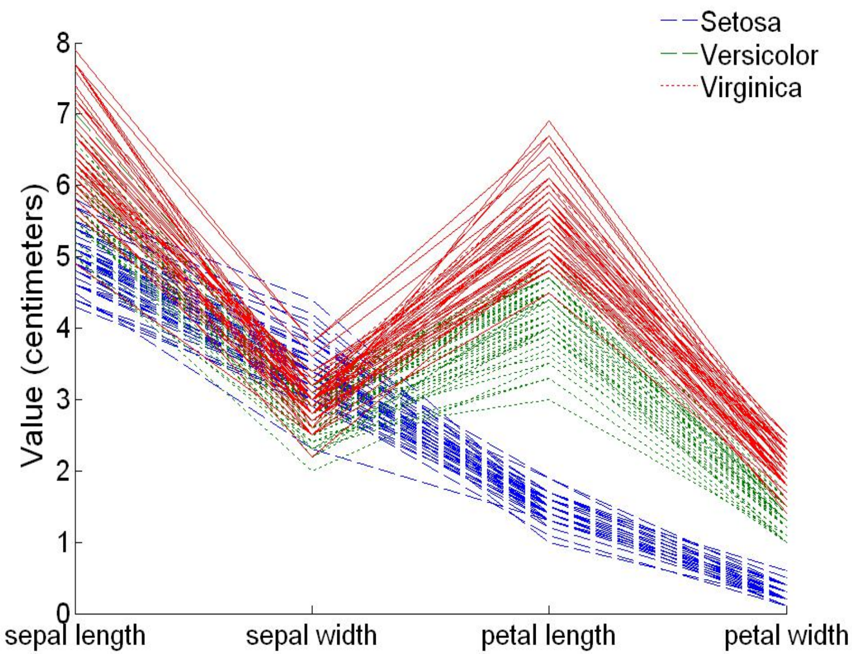


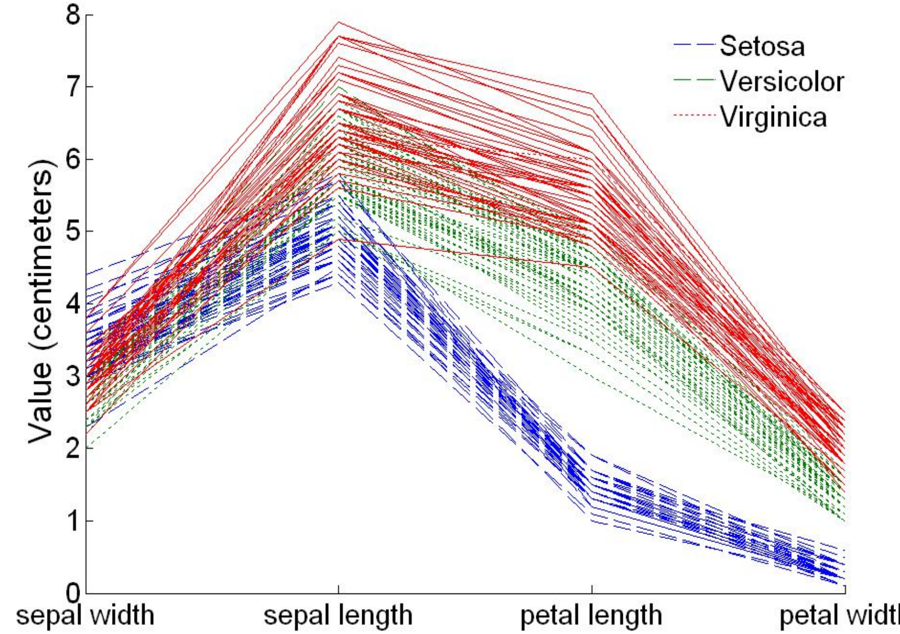
Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data

Visualize all the150 data records





Change the sequence of the first two features