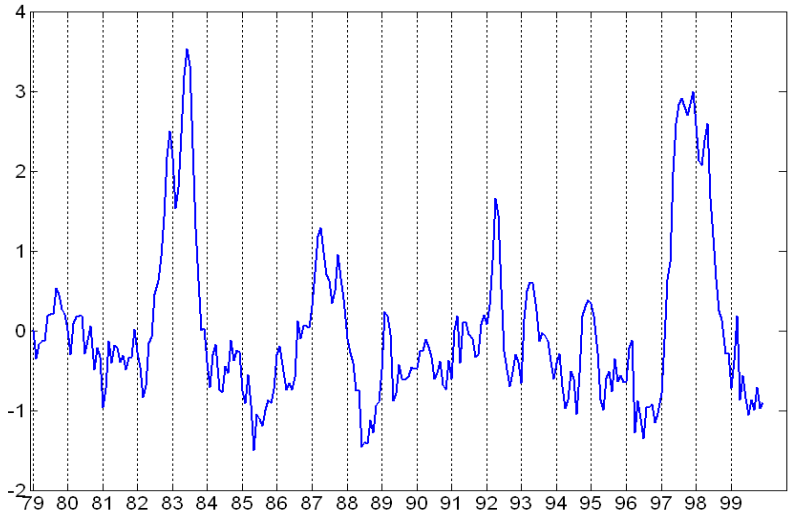
Data Mining: Data and Exploring Data (Chapter 2&3)

郑子彬　副教授
中山大学 计算机学院
zibin.gil@gmail.com
2015年

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Standardization and Normalization



Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

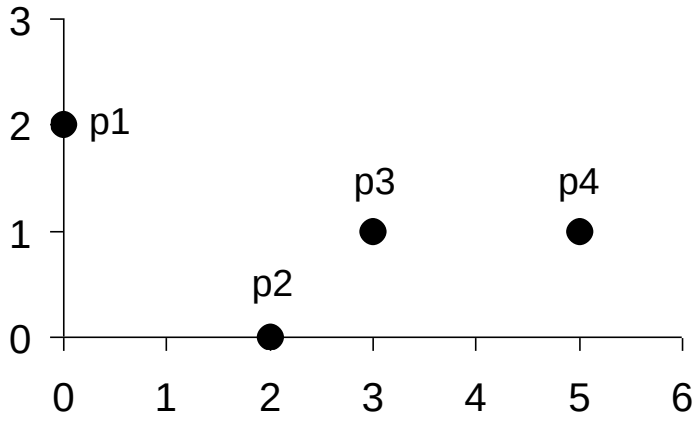
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Minkowski Distance: Examples

- r = 1. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- r = 2. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness，非负性)
2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry，对称性)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality，三角不等式)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric （度量）

Example: Non-metric dissimilarities

- A={1,2,3,4} B={2,3,4}
- A-B = {1} B-A=∅
- dis(A,B) = size(A - B) = 1
- dis(B,A) = size(B - A) = 0
- dis(A,B) = size(A-B) + size (B-A)

Example: Non-metric dissimilarities

- Distance between time of the day:
- d(t1,t2)=t2-t1 if t1<=t2
- d(t1,t2)=24+(t2-t1) if t1>=t2
- d(1PM,2PM) = 1 d(2PM, 1PM) = 23

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values

  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

M_{00} = 7 (the number of attributes where p was 0 and q was 0)

M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

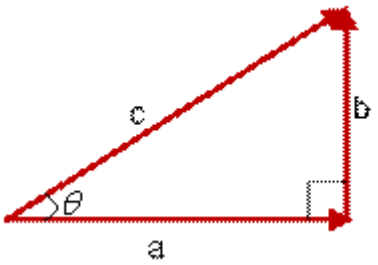
$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

两个向量的交集

两个向量的并集

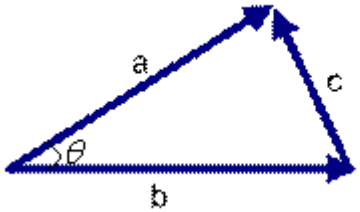
$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$

Cosine Similarity

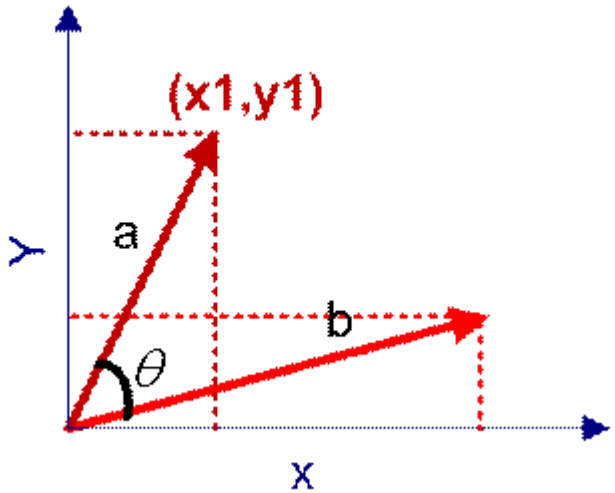


$$\cos(\theta) = \frac{a}{c}$$

公式(1)



$$\cos(\theta) = \frac{a^2 + b^2 - c^2}{2ab}$$



$$\cos(\theta) = \frac{\mathrm{a} \bullet \mathrm{b}}{||\,\mathrm{a}\,|| \times ||\,\mathrm{b}\,||}$$

$$= \frac{(x_1, y_1) \bullet (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$= \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

公式(3)

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where $\bullet$ indicates vector dot product and $\| d \|$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$$\cos(d_1, d_2) = .3150$$

Correlation（PCC皮尔森相关性）

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

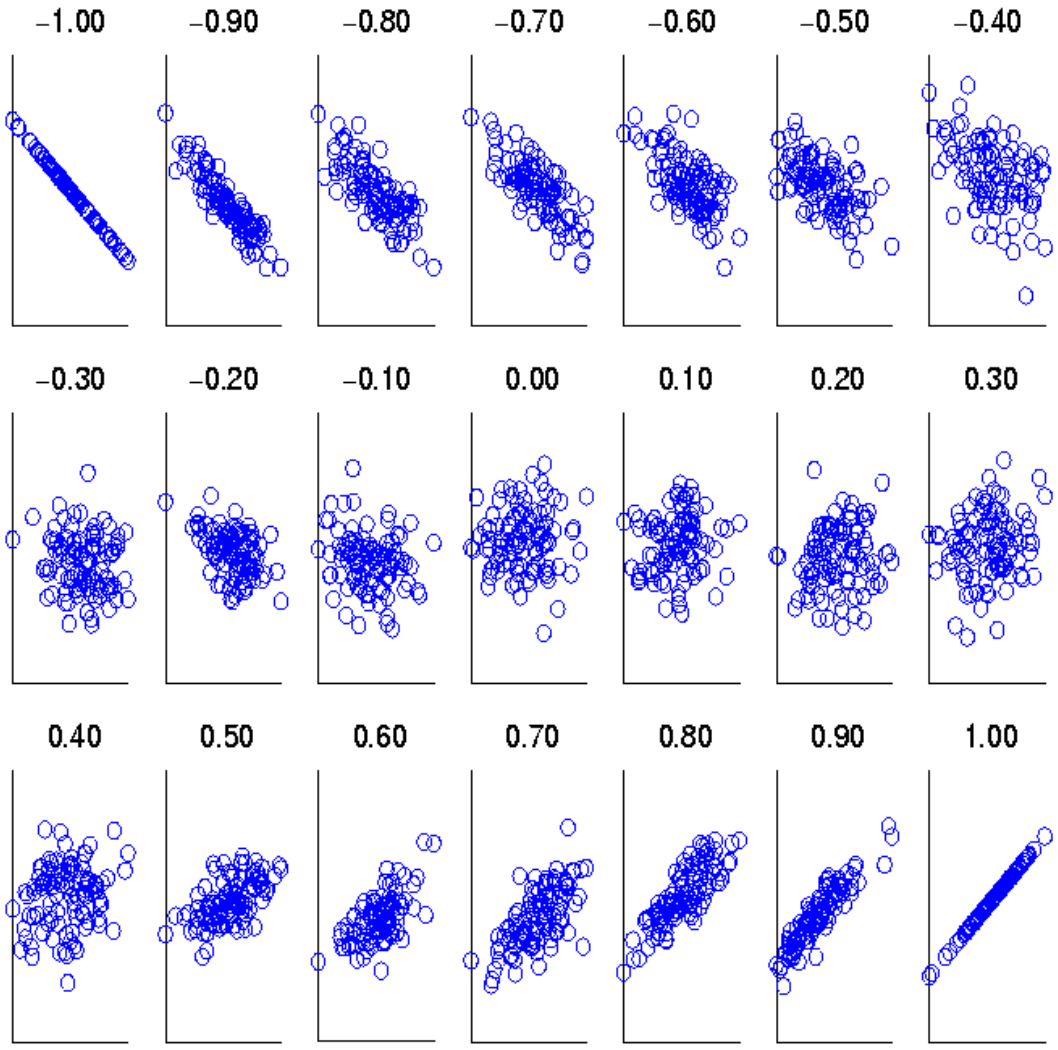
$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

先标准化，再内积；内积代表相似性

Cosine vs PCC：标准化的过程不同

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Density

- Density-based clustering require a notion of density

- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
  - Graph-based density

Euclidean Density – Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains

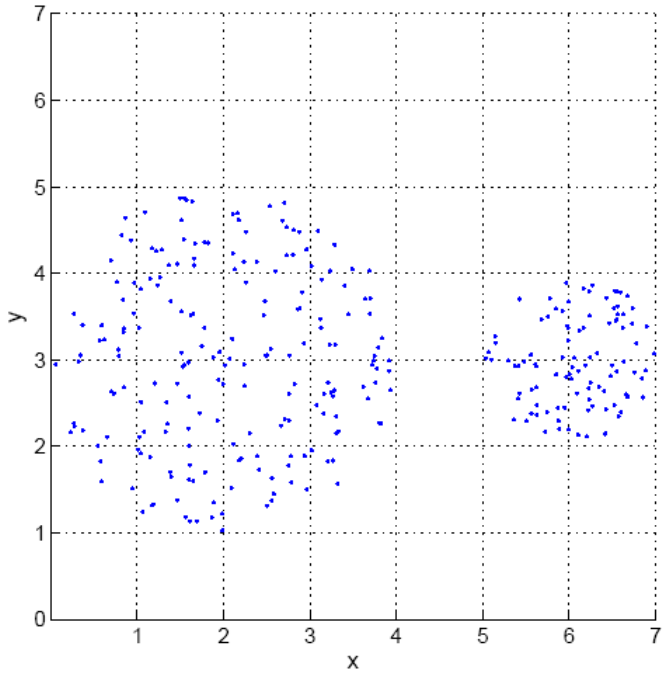


Figure 7.13. Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius of the point

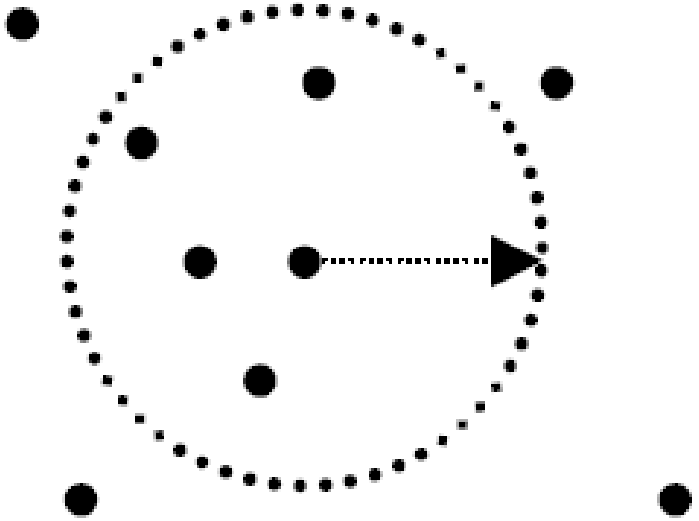


Figure 7.14. Illustration of center-based density.

Chapter 3: Exploring Data

What is data exploration?

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

- Key motivations of data exploration include
  - Helping to 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

- Related to the area of Exploratory Data Analysis (EDA探测性数据分析)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook

  http://www.itl.nist.gov/div898/handbook/index.htm

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- In our discussion of data exploration, we focus on
  - Summary statistics （汇总统计）
  - Visualization
  - Online Analytical Processing (OLAP)

Iris （鸢尾花）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
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal（萼片）width and length
    - Petal （花瓣）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.

Summary Statistics

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

Frequency and Mode（众数）

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

Percentiles（百分位数）

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

- Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value x_p of x such that p% of the observed values of x are less than x_p.

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean（截断均值） is also commonly used.

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

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

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.

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

为什么m-1?

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

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

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

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

Why m-1?

首先，我们假定随机变量 X 的数学期望 μ 是已知的，然而方差 σ^2 未知。在这个条件下，根据方差的定义我们有

$$\mathbb{E}\left[(X_i-\mu)^2\right]=\sigma^2,\quad \forall i=1,\ldots,n,$$

因此 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$ 是方差 σ^2 的一个无偏估计，注意式中的分母不偏不倚正好是 n ！

现在，我们考虑随机变量 X 的数学期望 μ 是未知的情形。这时，我们会倾向于无脑直接用样本均值 $\bar{X}$ 替换掉上面式子中的 μ 。这样做有什么后果呢？后果就是，

如果直接使用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 作为估计，那么你会倾向于低估方差！

Why m-1?

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)+(\mu-\bar{X})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\bar{X}-\mu)(\mu-\bar{X})+(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2
\end{aligned}$$

换言之，除非正好 $\bar{X}=\mu$，否则我们一定有

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 < \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2,$$

而不等式右边的那位才是的对方差的“正确”估计!

这个不等式说明了，为什么直接使用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 会导致对方差的低估。

- 课外阅读：http://www.zhihu.com/question/20099757

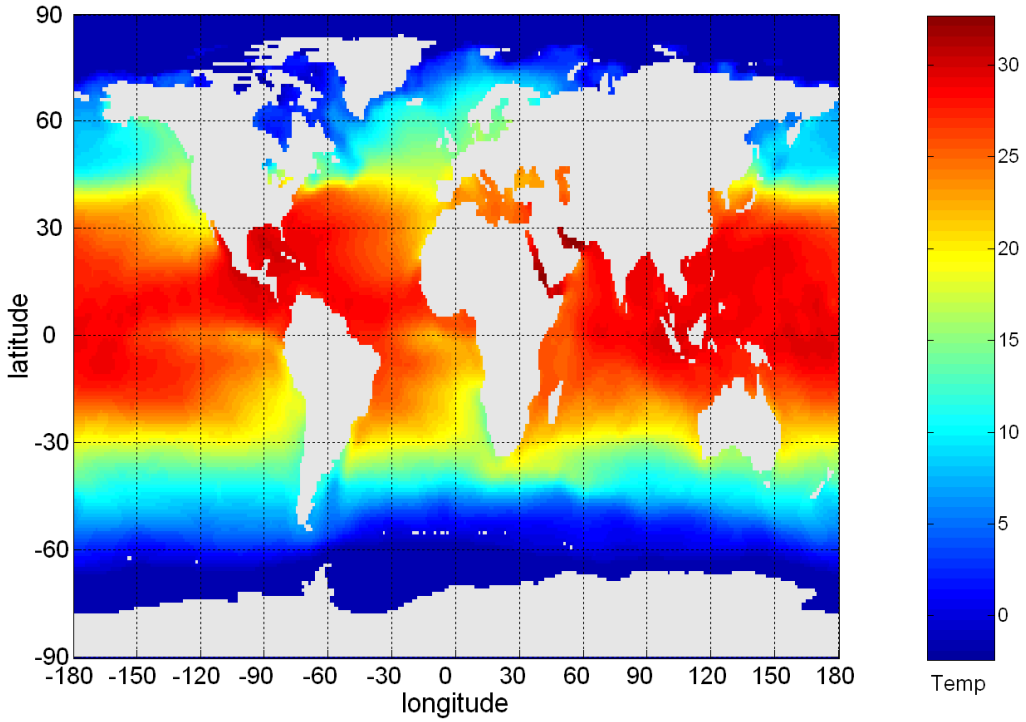
Visualization

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.
  - 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

Example: Sea Surface Temperature

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



Representation

- Is 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

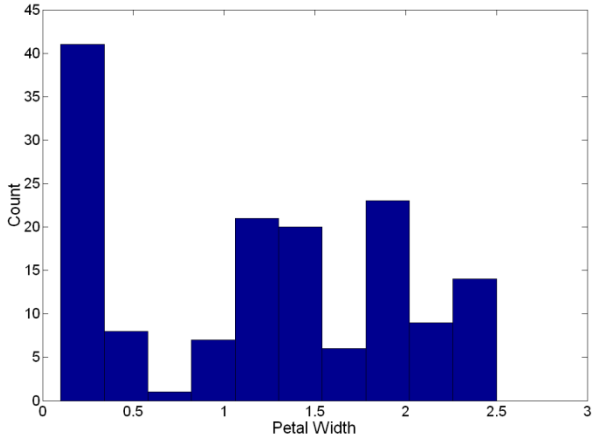
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

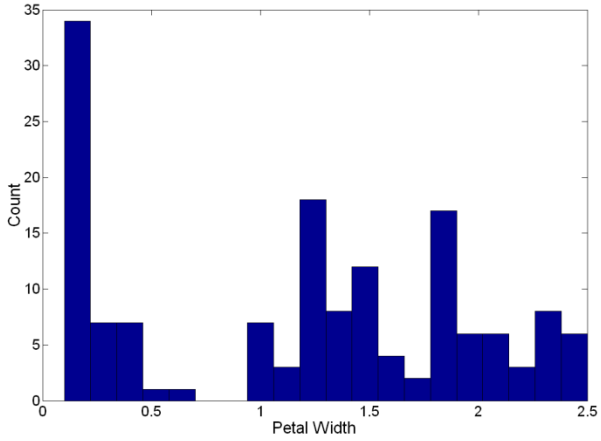
Selection

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve the chossing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

Visualization Techniques: Histograms

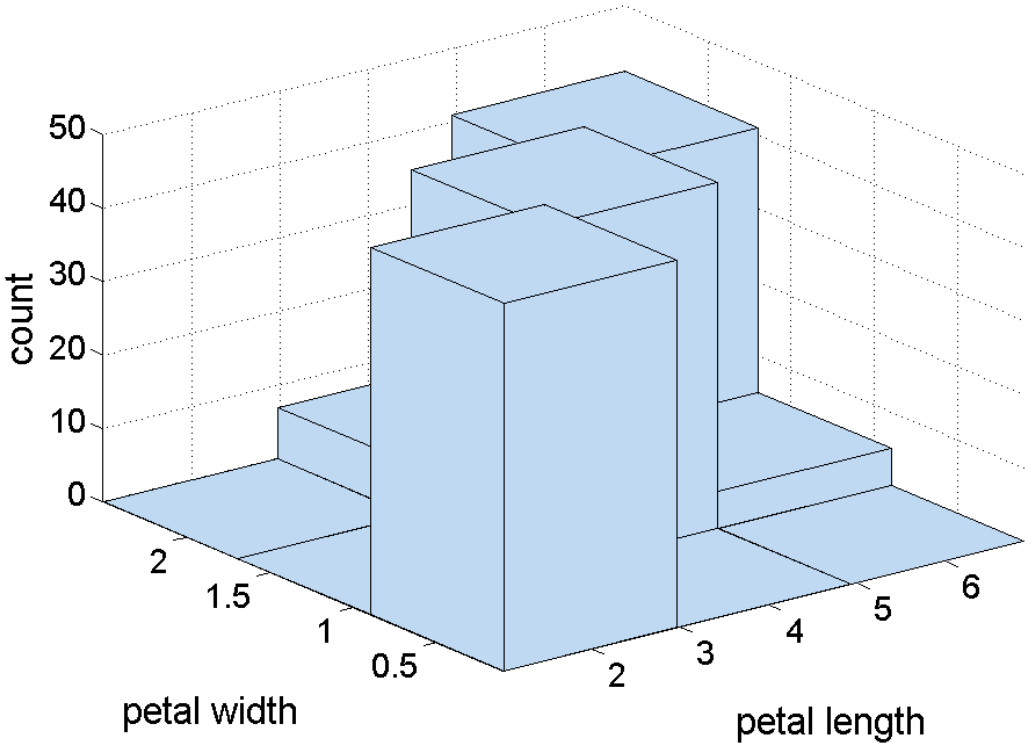
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





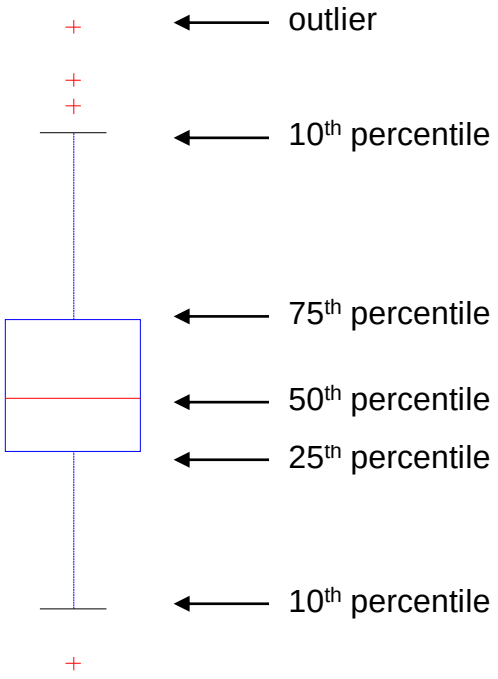
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



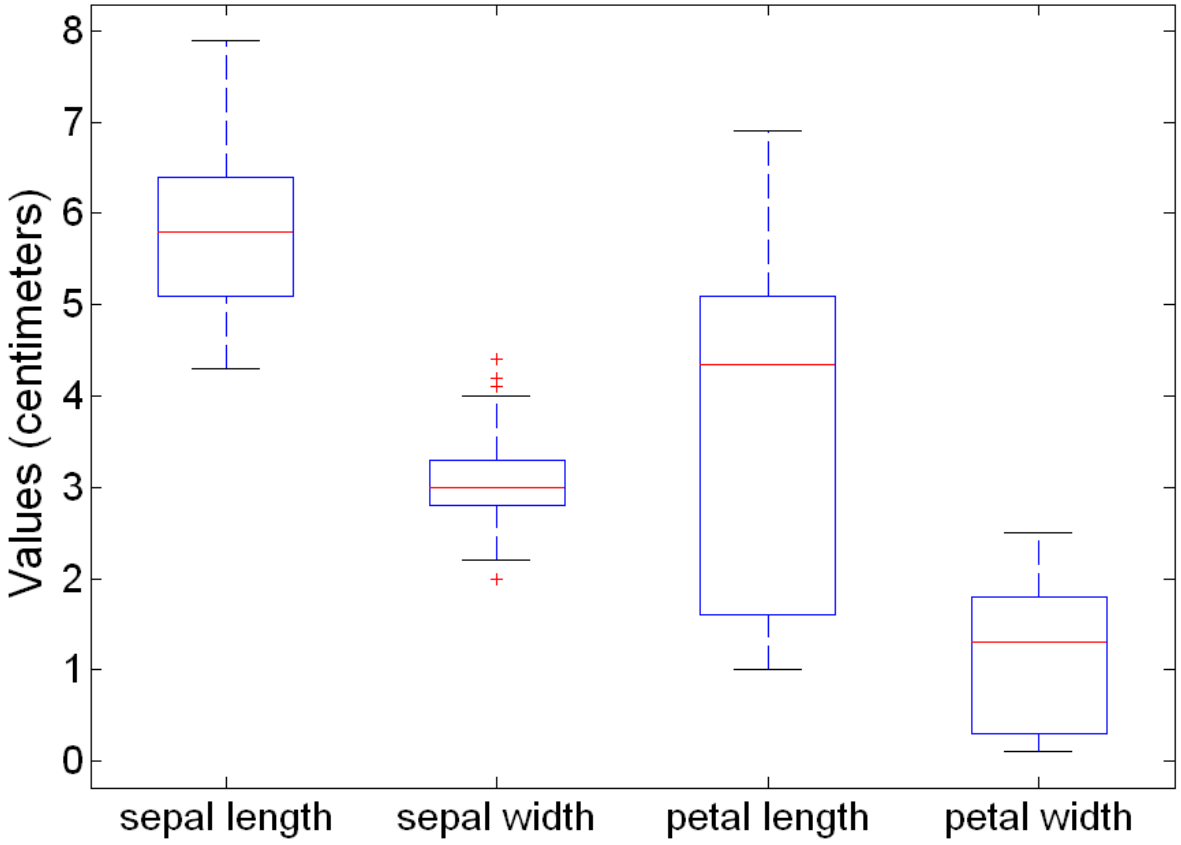
Visualization Techniques: Box Plots

- Box Plots (盒状图)
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

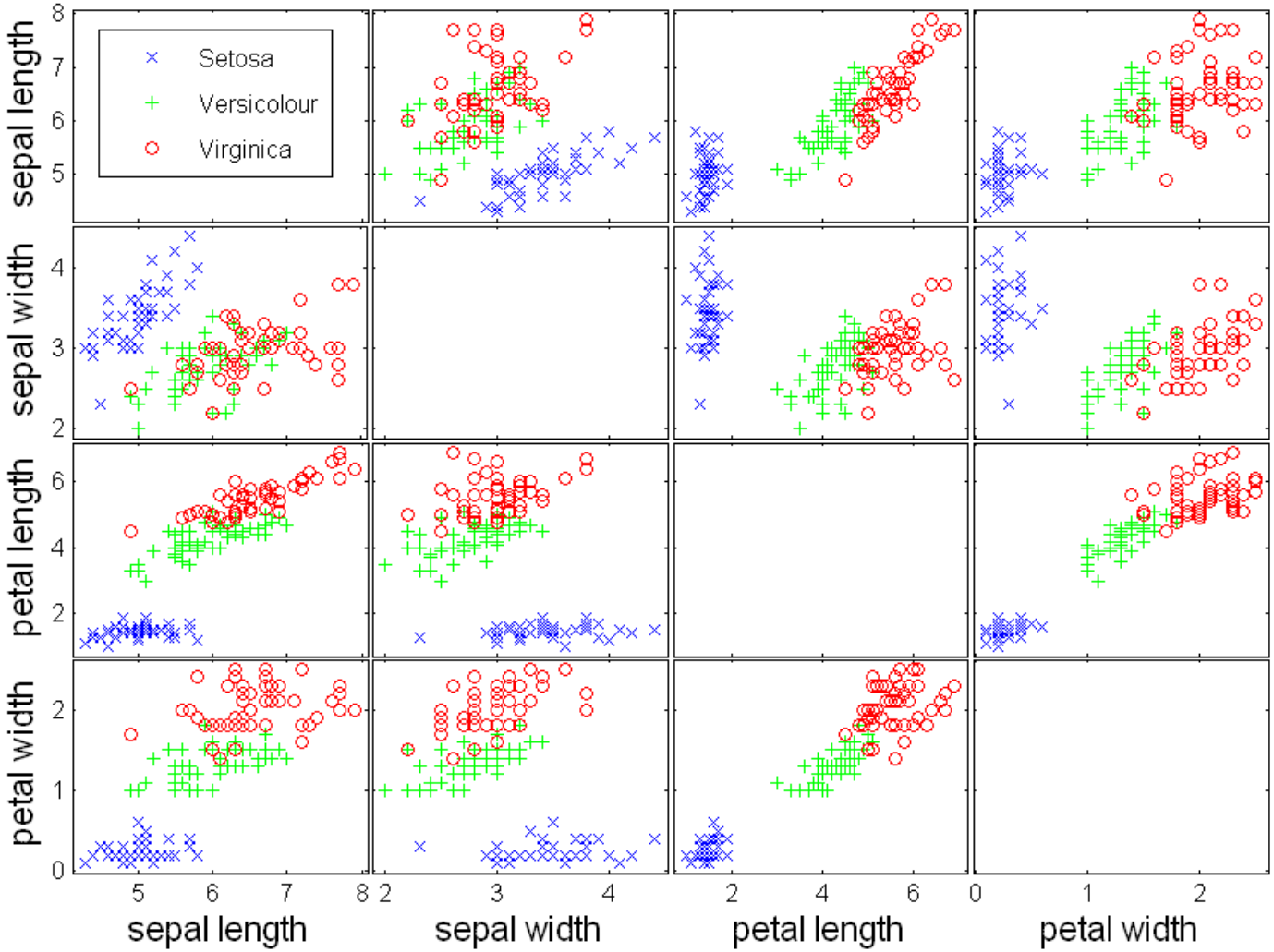
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- Scatter plots （散布图）
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes 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 can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide

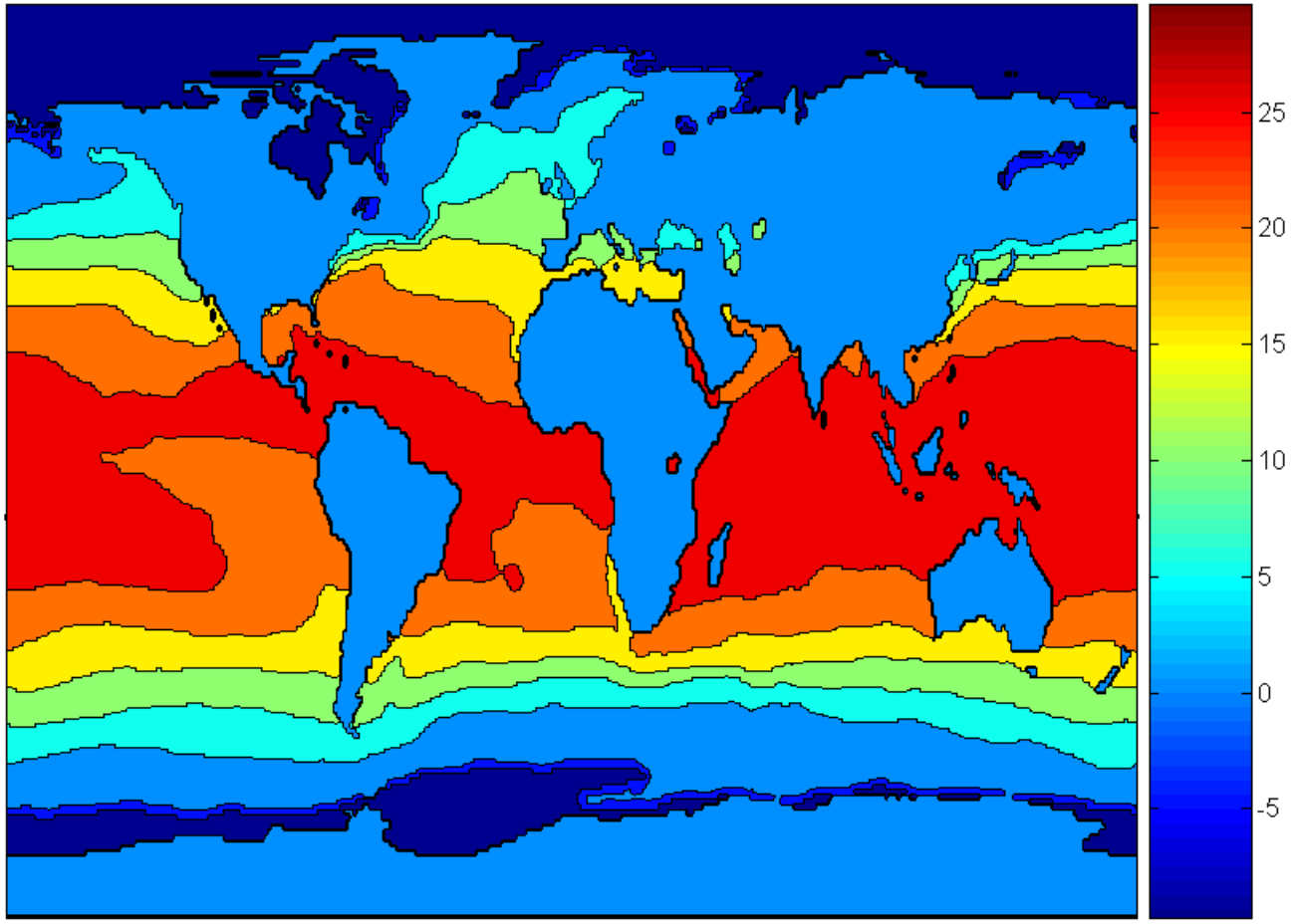
Scatter Plot Array of Iris Attributes



Visualization Techniques: Contour Plots

- Contour plots（等高线图）
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values
  - The most common example is contour maps of elevation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide

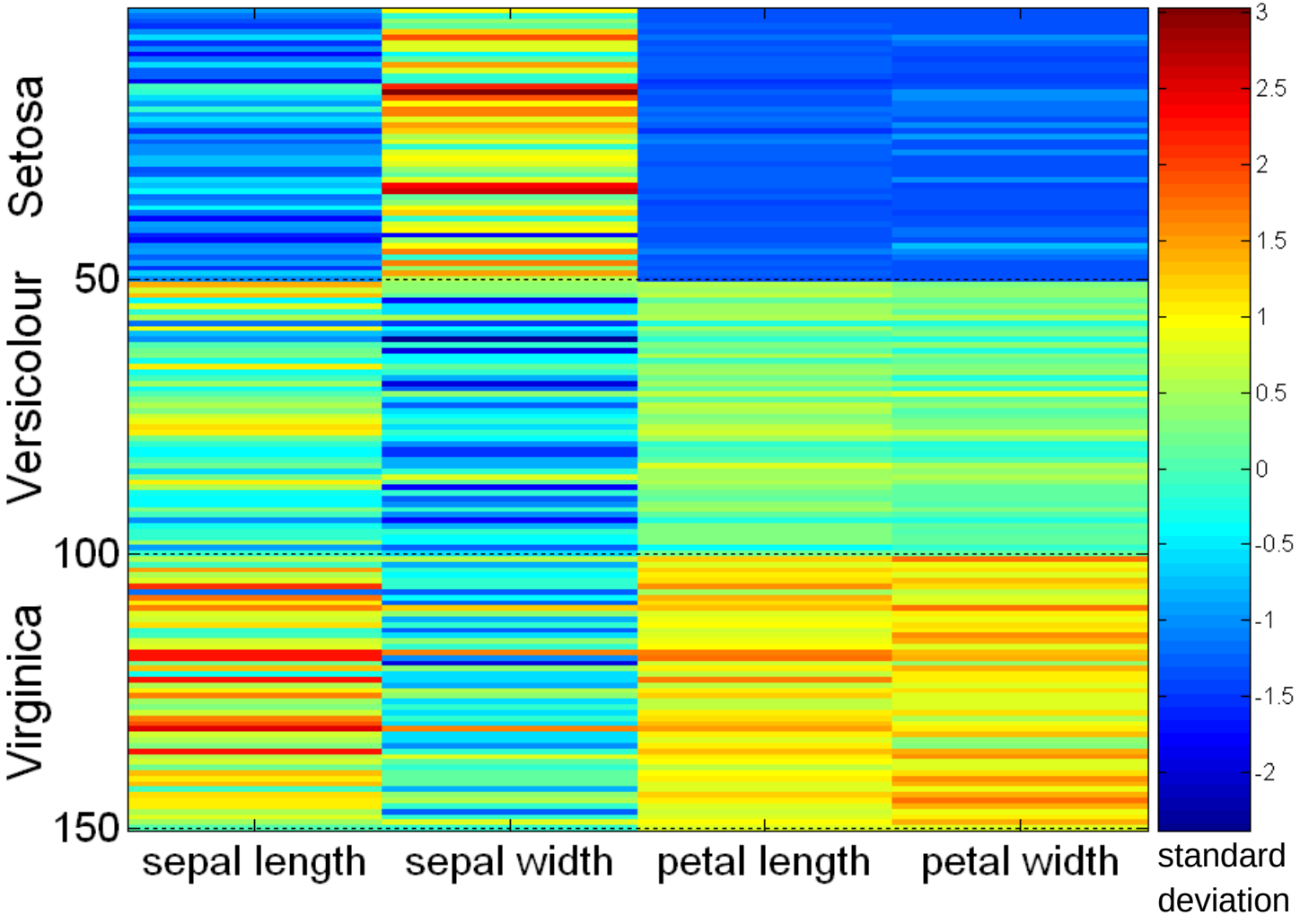
Contour Plot Example: SST Dec, 1998



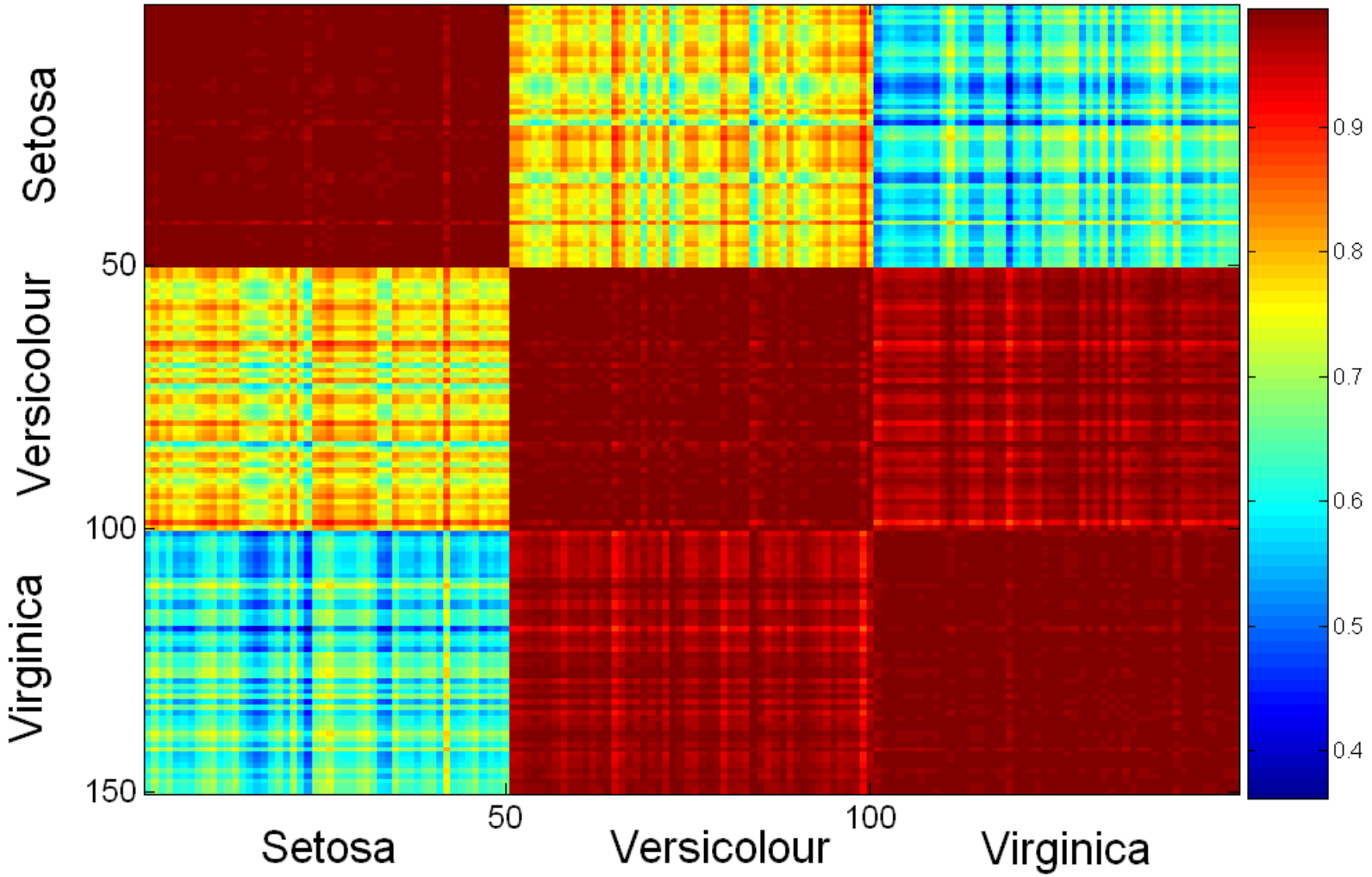
Visualization Techniques: Matrix Plots

- Matrix plots
  - Can plot the data matrix
  - 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 two slides

Visualization of the Iris Data Matrix



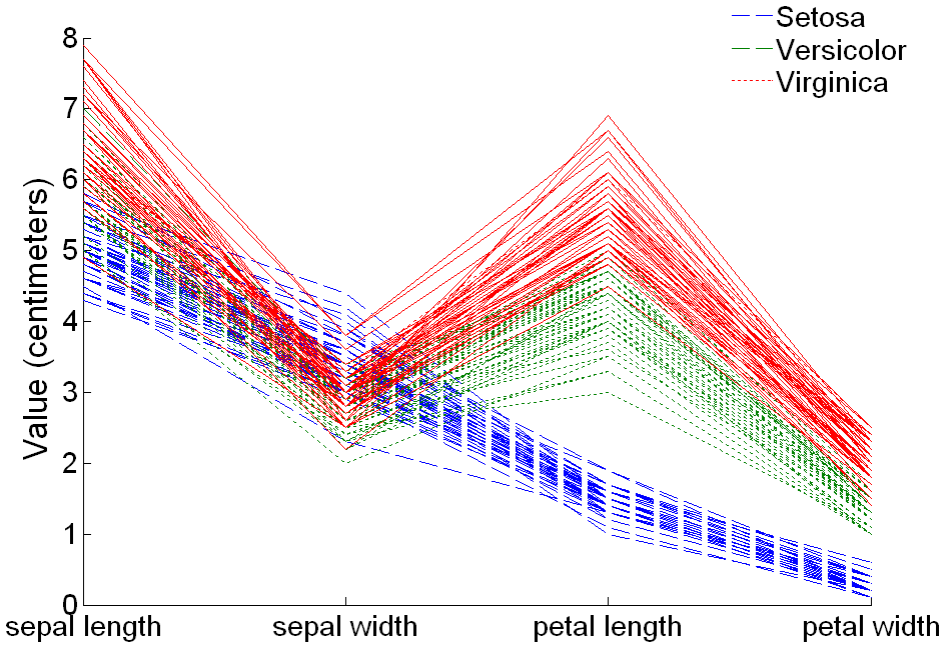
Visualization of the Iris Correlation Matrix

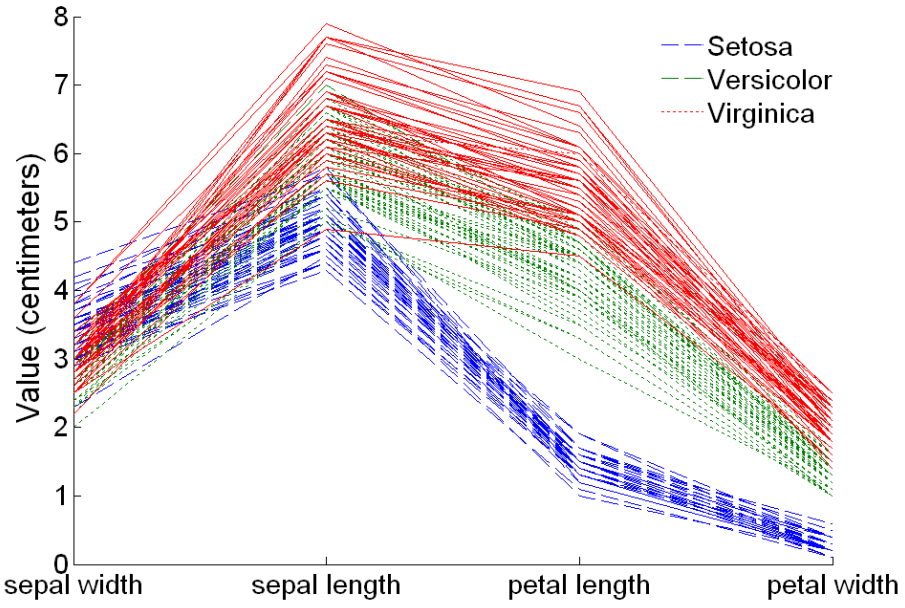


Visualization Techniques: Parallel Coordinates

- 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



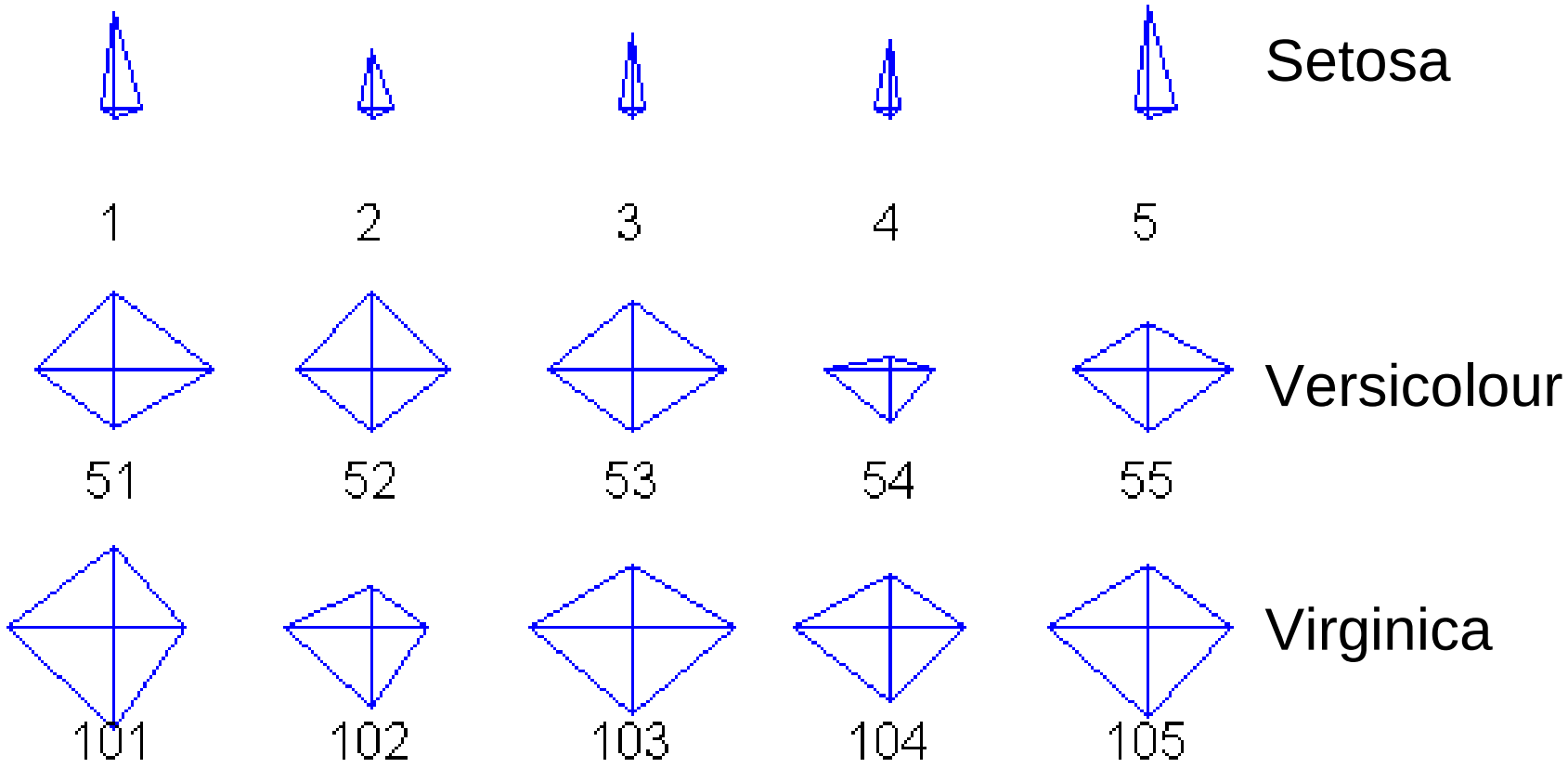


缺点：线交叉

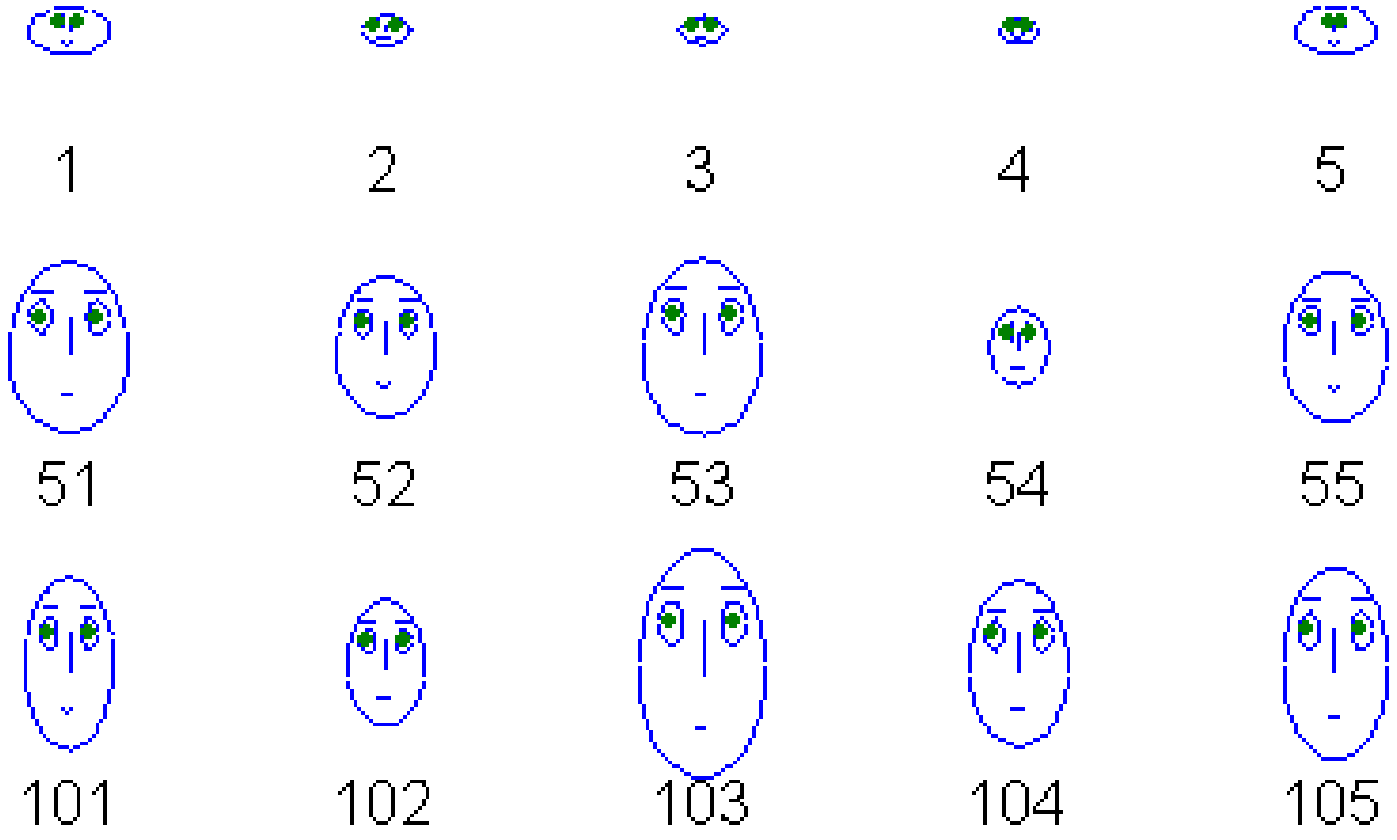
Other Visualization Techniques

- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Star Plots for Iris Data



Chernoff Faces for Iris Data



Setosa

Versicolour

Virginica

数据特征	面部特征
萼片长度	脸部大小
萼片宽度	前额/颚相对弧长
花瓣长度	前额形状
花瓣宽度	颚的形状

OLAP

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database.
- Relational databases put data into tables, while OLAP uses a multidimensional array representation.
  - Such representations of data previously existed in statistics and other fields
- There are a number of data analysis and data exploration operations that are easier with such a data representation.

Creating a Multidimensional Array

- Two key steps in converting tabular data into a multidimensional array.
  - First, identify which attributes are to be the dimensions and which attribute is to be the target attribute whose values appear as entries in the multidimensional array.
    - The attributes used as dimensions must have discrete values
    - The target value is typically a count or continuous value, e.g., the cost of an item
  - Second, find the value of each entry in the multidimensional array by summing the values (of the target attribute) or count of all objects that have the attribute values corresponding to that entry.

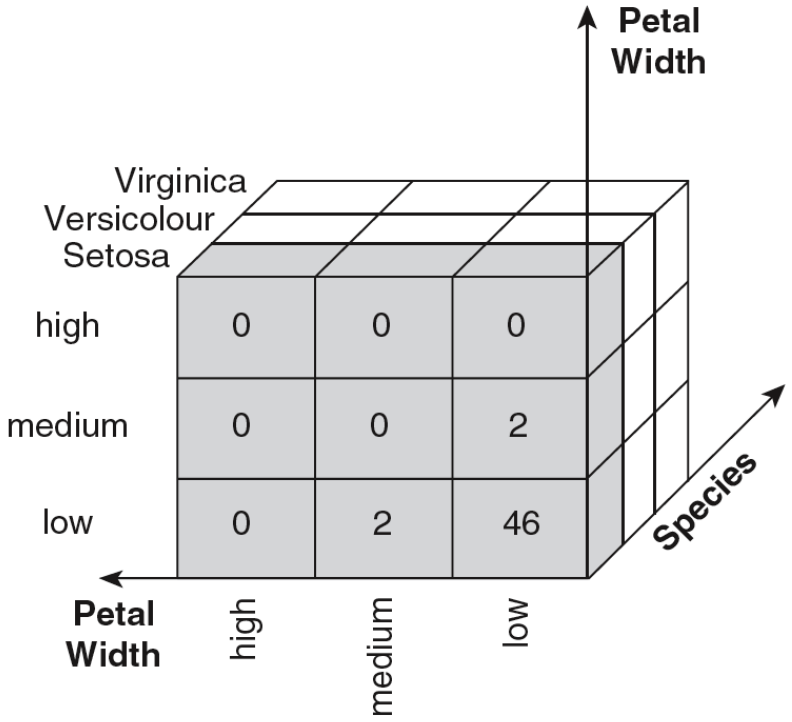
Example: Iris data

- We show how the attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical values: *low*, *medium*, and *high*
  - We get the following table - note the count attribute

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- The figure illustrates the result.
- All non-specified tuples are 0.



Example: Iris data (continued)

- Slices of the multidimensional array are shown by the following cross-tabulations
- What do these tables tell us?

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

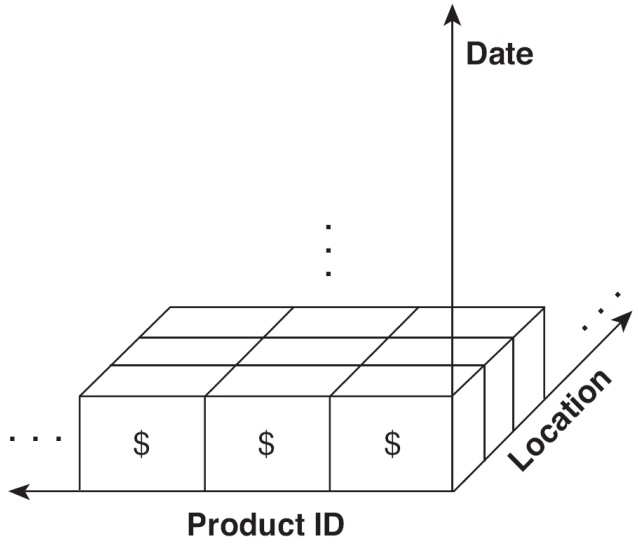
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube
- A data cube is a multidimensional representation of data, together with all possible aggregates.
- By all possible aggregates, we mean the aggregates that result by selecting a proper subset of the dimensions and summing over all remaining dimensions.
- For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.

Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represented as a 3 dimensional array
- There are 3 two-dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, and 1 zero-dimensional aggregate (the overall total)



Data Cube Example (continued)

- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the overall total

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

OLAP Operations: Slicing and Dicing

- Slicing（切片） is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.
- Dicing（切块） involves selecting a subset of cells by specifying a range of attribute values.
  - This is equivalent to defining a subarray from the complete array.
- In practice, both operations can also be accompanied by aggregation over some dimensions.

OLAP Operations: Roll-up and Drill-down

- Attribute values often have a hierarchical structure.
  - Each date is associated with a year, month, and week.
  - A location is associated with a continent, country, state (province, etc.), and city.
  - Products can be divided into various categories, such as clothing, electronics, and furniture.
- Note that these categories often nest and form a tree or lattice
  - A year contains months which contains day
  - A country contains a state which contains a city

OLAP Operations: Roll-up and Drill-down

- This hierarchical structure gives rise to the roll-up and drill-down operations.
  - For sales data, we can aggregate (roll up上卷) the sales across all the dates in a month.
  - Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down下钻) into daily sales totals.
  - Likewise, we can drill down or roll up on the location or product ID attributes.