Week 2
Engineering Data

Seokho Chi
Professor | Ph.D.
SNU Construction Innovation Lab

Source: Tan, Kumar, Steinback (2006)

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Home Owner	Marital Status	Taxable Income	Defaulted Borrower
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

Data set for predicting borrowers who will default on loan payments

Contingency correlation: two variables are independent or dependent
Rank correlation: relationship b/w rankings of different ordinal variables
Pearson's correlation: linear correlation b/w two variables

Chi-square test: independence comparison b/w two variables
Run test: test if sample data are generated from a random process
Sign test: compare the sizes of two groups (check subjects before and after treatment)

Types of Attributes

Categorical (Qualitative) Discrete Attributes

Numeric (Quantitative) Continuous Attributes

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

Geometric mean: central tendency(3→6 2배, 6→48 8배: 기하평균 4, 3의 4배 후 4배=48)
Harmonic mean: 역수의 산술평균의 역수(전체 거리의 절반을 40km/h로 달리고 나머지를 60km/h로 달렸다면 평균속력은 조화평균인 48km/h(시간을 절반씩 달렸다면 평균속력은 50km/h)
Percent variation: ratio of the absolute variation to the base value (price changes on the stock market)

Types of data sets

- Record
  - Data Matrix (Numeric)
  - Document Data (Count)
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	House Owner	Marital Status	Taxable Income	Defaulted Borrower
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

(1) Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute → Possible 3D Plotting

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

(2) Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component may be the number of times the corresponding term occurs in the document.

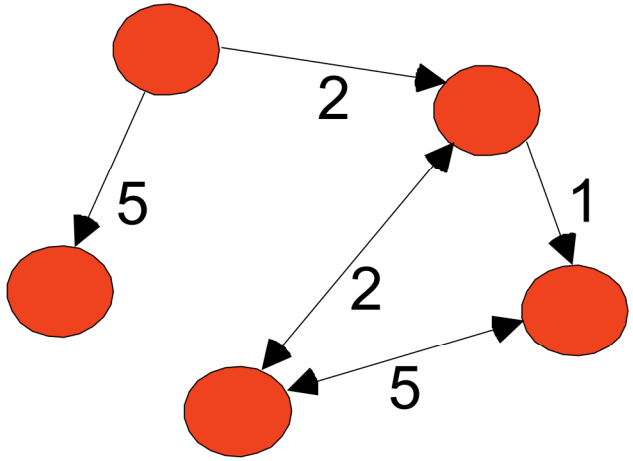
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(3) Transaction Data

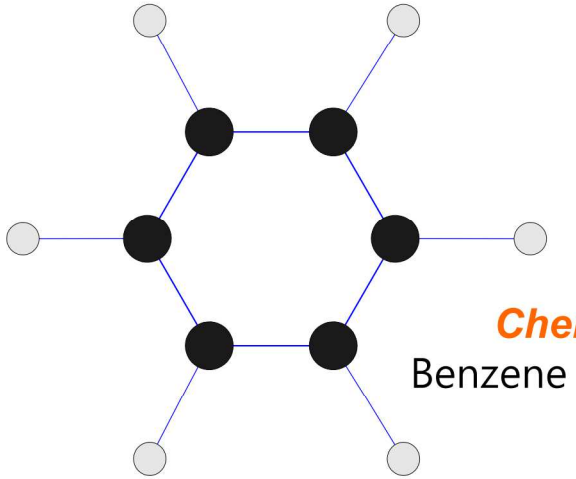
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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



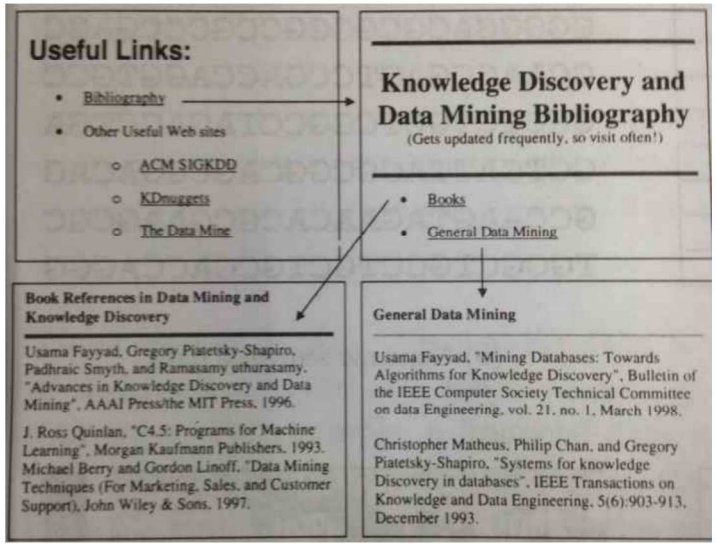
Generic Graph



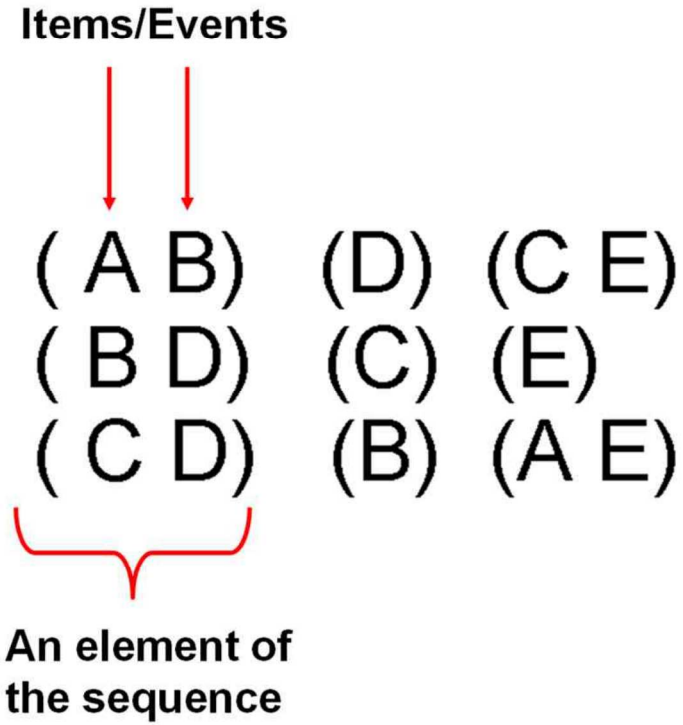
Chemical Data

Benzene Molecule: C_6H_6

HTLM Links



Ordered Data

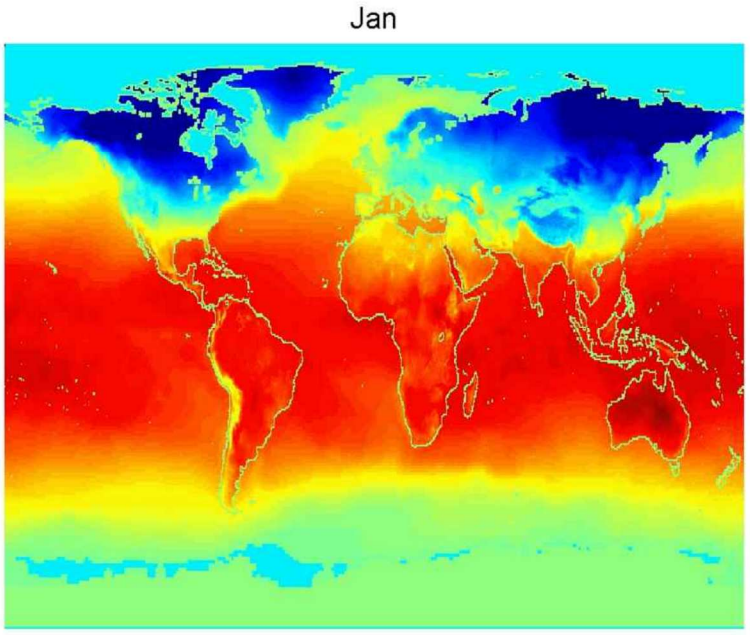


Sequences of transactions

DNA Sequencing

24 Different Human Chromosomes

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



Spatio-Temporal Data

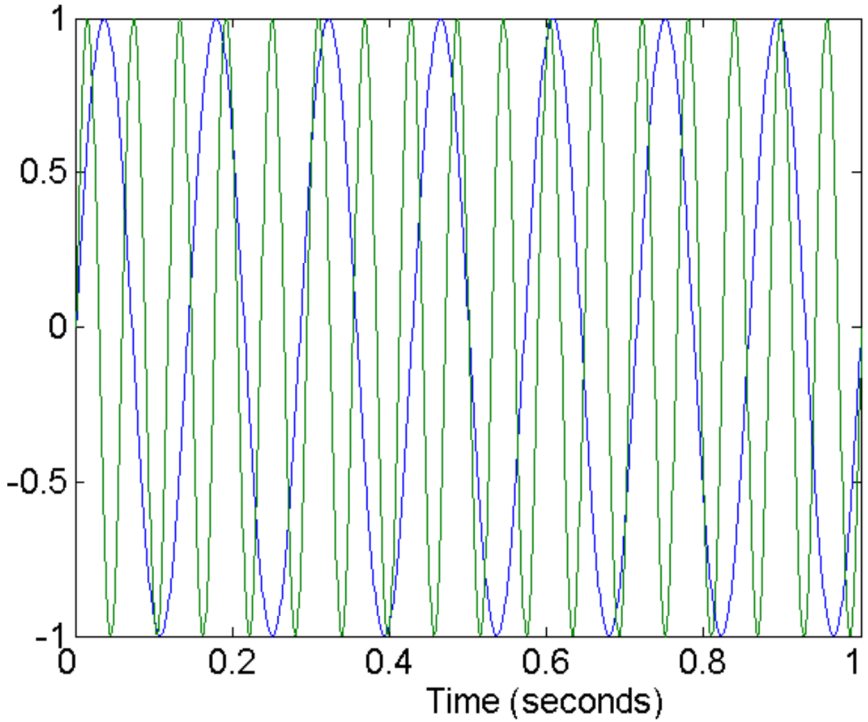
Average Monthly Temperature

Data Quality

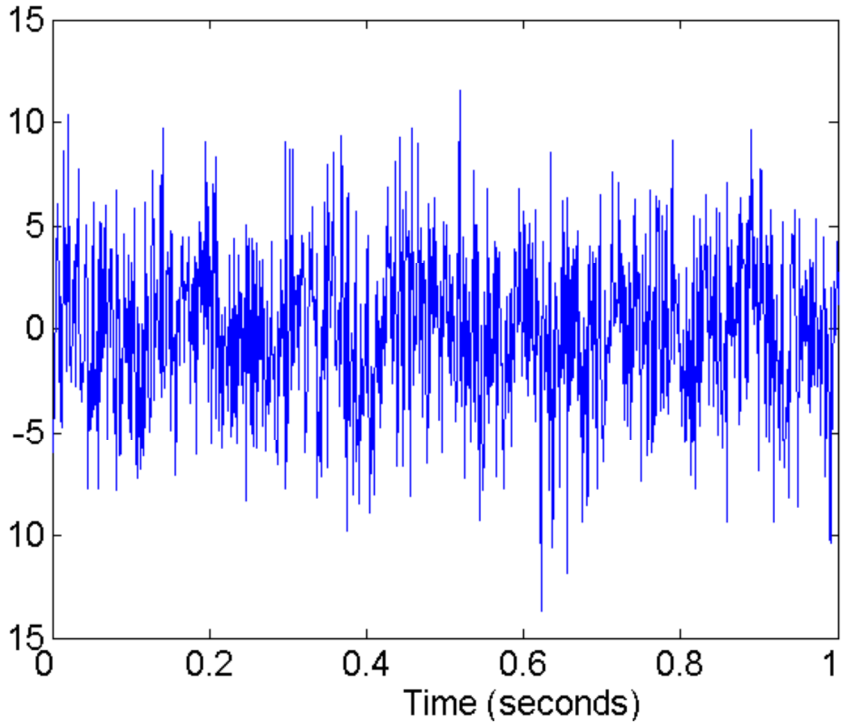
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



Two Sine Waves

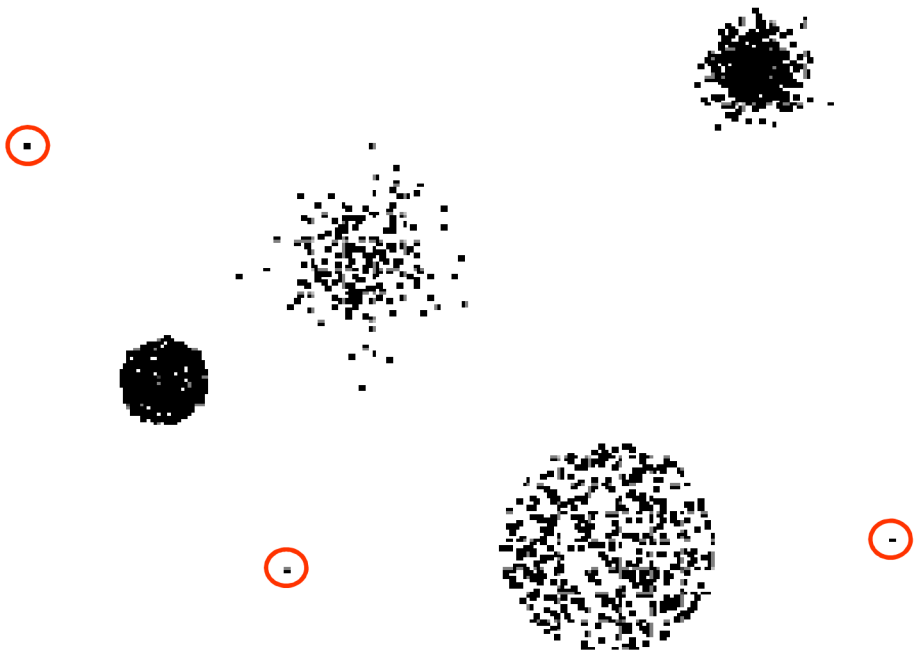


Two Sine Waves + Noise

Randomly generated due to machinery problems or network problems
Should be removed before outlier detection

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values (data is not available)
  - 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)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources: where assign? or different objects?

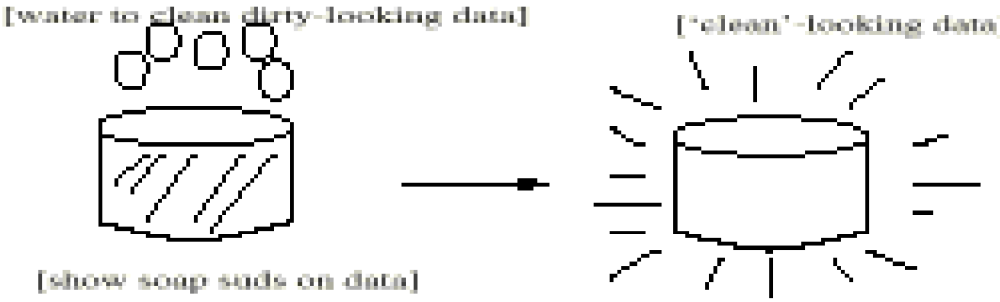
- Examples:
  - Same person with multiple email addresses

Why Data Preprocessing?

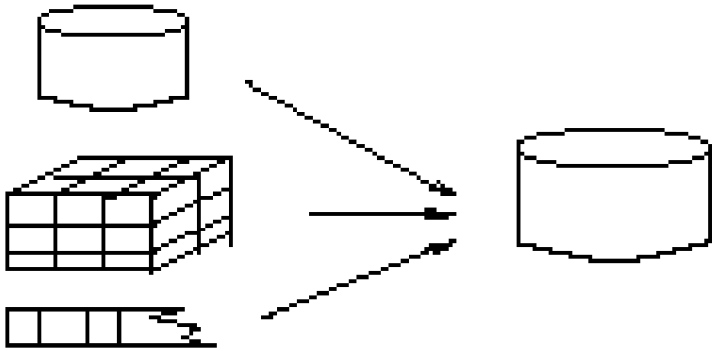
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- Less quality data, less quality mining results!
  - Quality decisions must be based on quality data
  - Garbage in → Garbage out

Major Tasks in Data Preprocessing

Data Cleaning



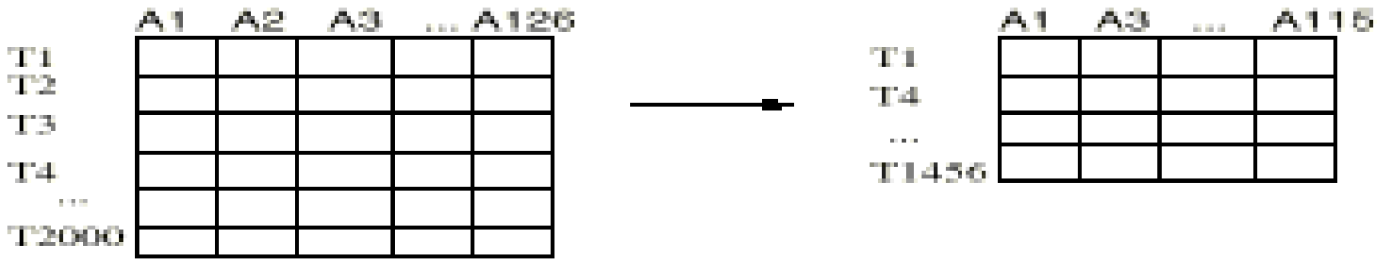
Data Integration



Data Transformation

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

Data Reduction



Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many records have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

How to Handle Missing Data?

- Ignore the record: usually done when class label is missing
  - assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.

- Fill in the missing value manually: tedious + infeasible?
  - Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
  - Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

How to Handle Noisy Data?

- Binning method
  - sorting data → partition into bins (same # or range)
  - then one can smooth by bin means, median, boundaries, etc.
- Clustering
  - detect and remove outliers
- Regression
  - smooth by fitting the data into regression functions
- Combined computer and human inspection
  - detect suspicious values and check by human

Data Transformation

- Simple functions: x^k, log(x), e^x, |x|
- Normalization: scaled to fall within a small, specified range
  - 100→50 vs 4→2
  - min-max normalization (consider the entire scale)

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

  - z-score normalization (consider the distribution)

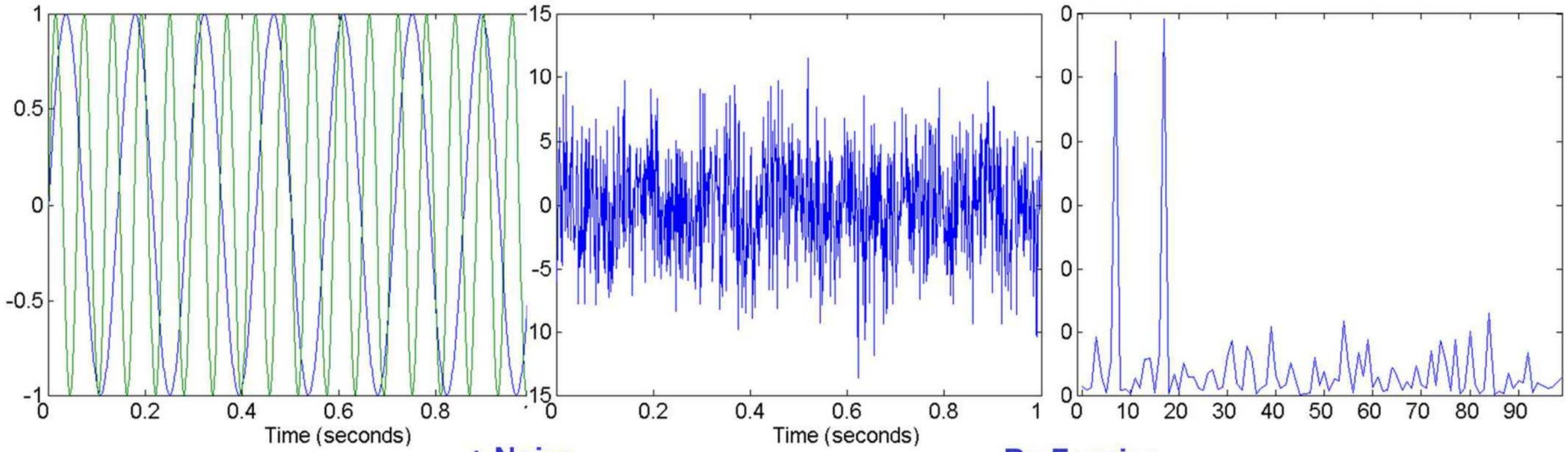
$$v' = \frac{v - mean_A}{stand\_dev_A}$$

  - normalization by decimal scaling

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

Data Transformation

- Feature selection
  - select new attributes that can capture the important information in a data set much more efficiently than the original attributes
  - feature extraction → mapping data into new space (e.g., Fourier, Wavelet) → feature construction



Data Transformation

- Discretization
  - Continuous data → categorical data (intervals)

    e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9 → low, medium, high

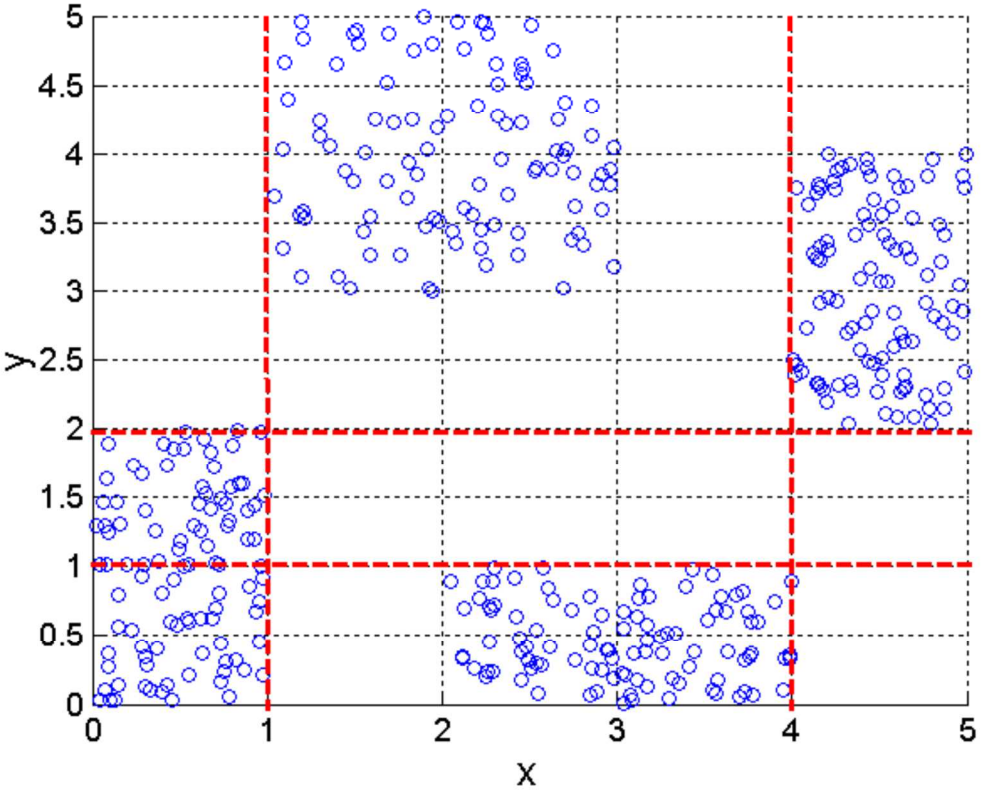
  - Subtasks:

    Decide how many categories to have

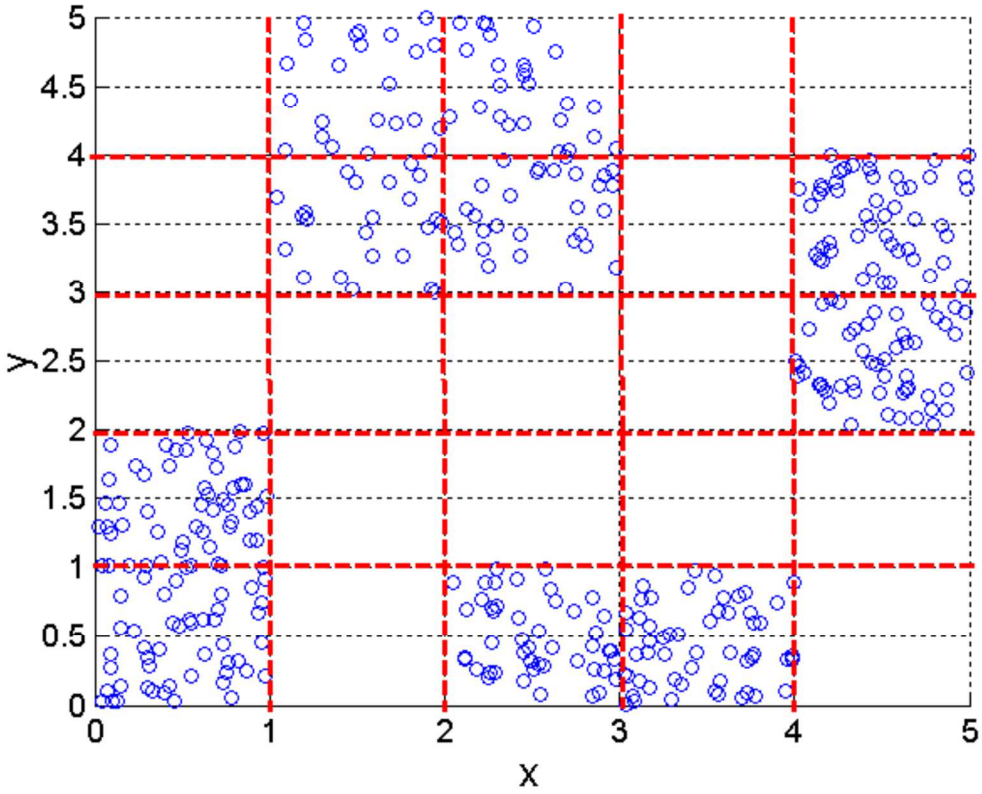
    Determine how to map the continuous values to these categories

Discretization Using Class Labels

(supervised)



3 categories for both x and y

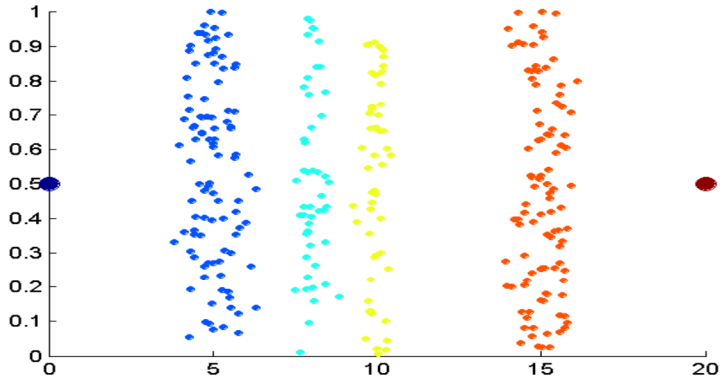


5 categories for both x and y

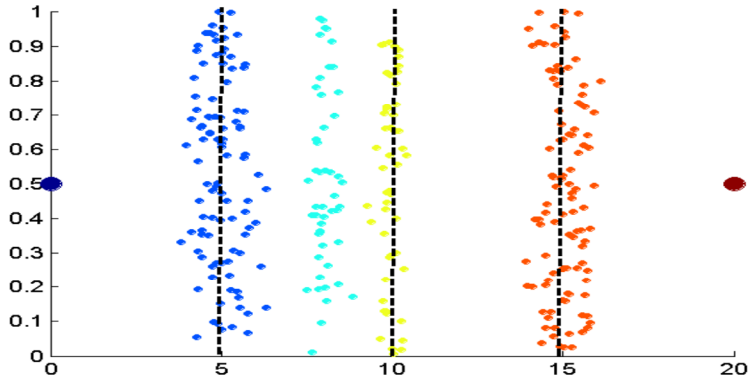
4 Classes

Discretization without Using Class Labels

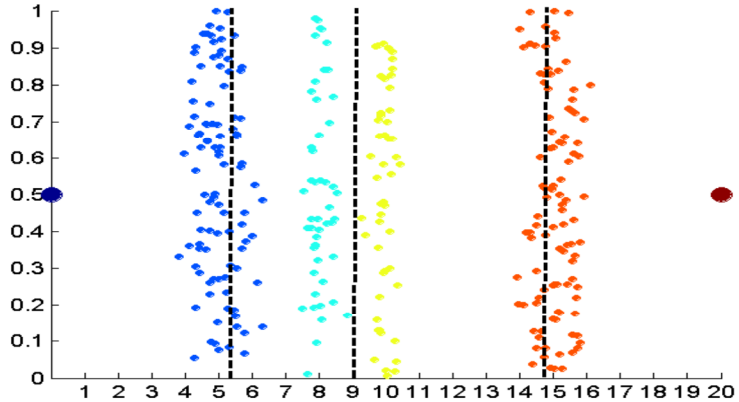
(unsupervised)



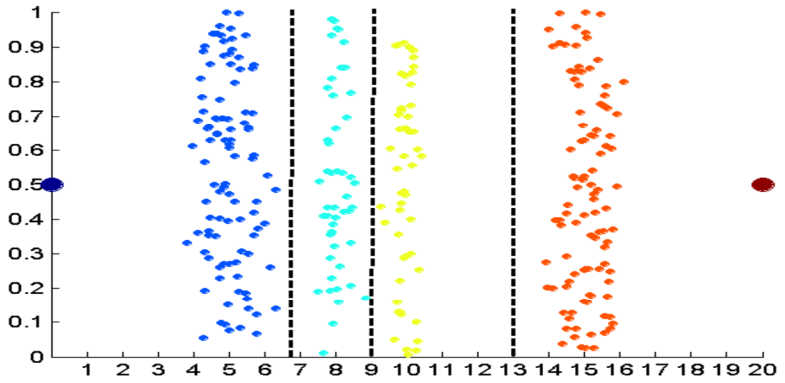
Data



Equal interval width



Equal frequency



K-means

Unknown Classes

Data Reduction

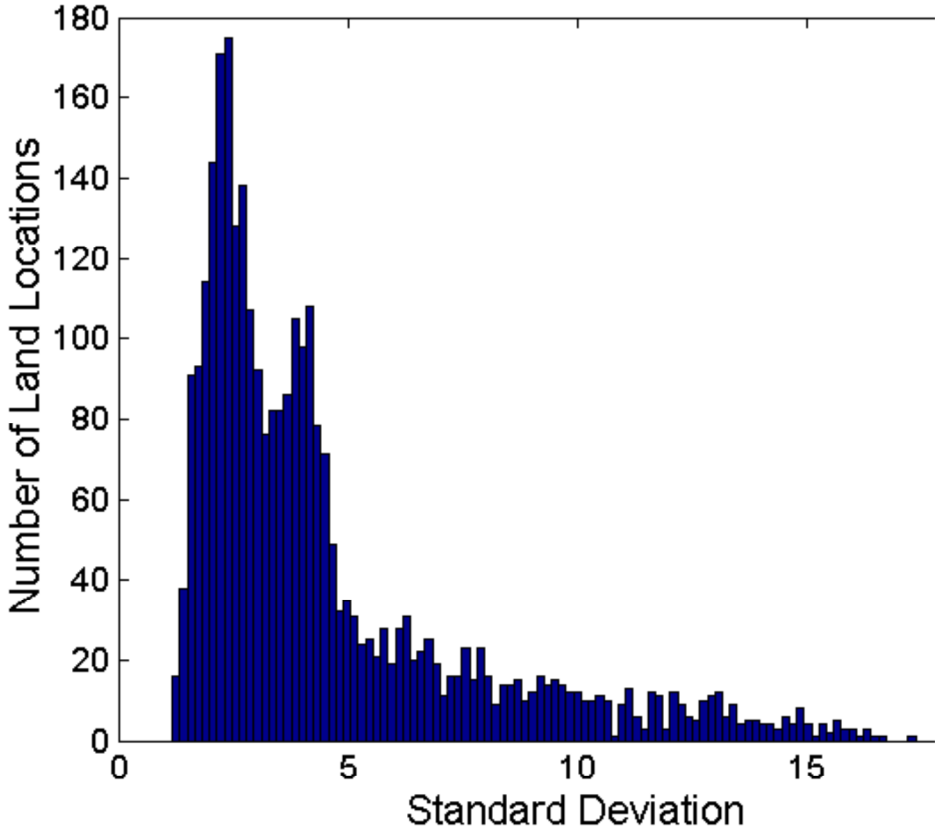
- Combining two or more attributes (or objects) into a single attribute (or object)

- Aggregation
  - Data reduction: Reduce the number of attributes or objects
  - Change of scale: Cities aggregated into regions, states, countries, etc

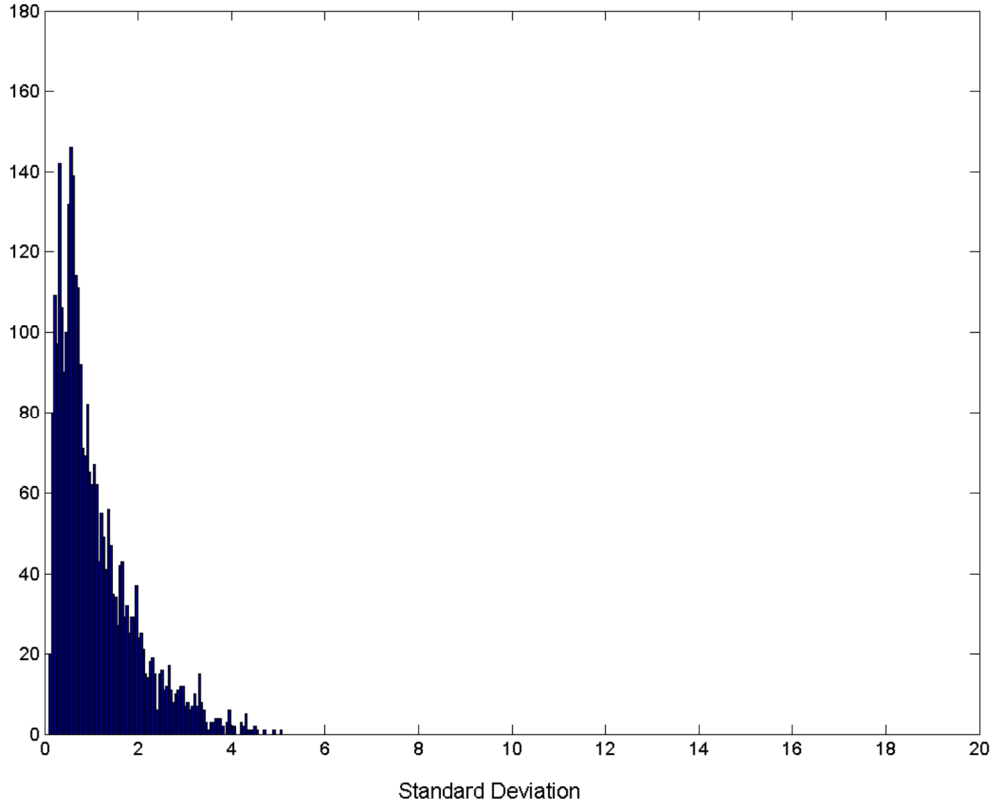
More "stable" data: aggregated data tends to have less variability
With aggregated data, save time and memory for processing + better visualization

Aggregation

Variation of Precipitation in Australia



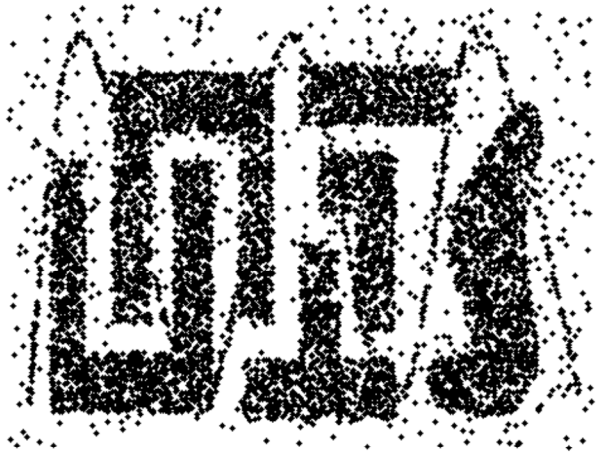
Standard Deviation of Average Monthly Precipitation



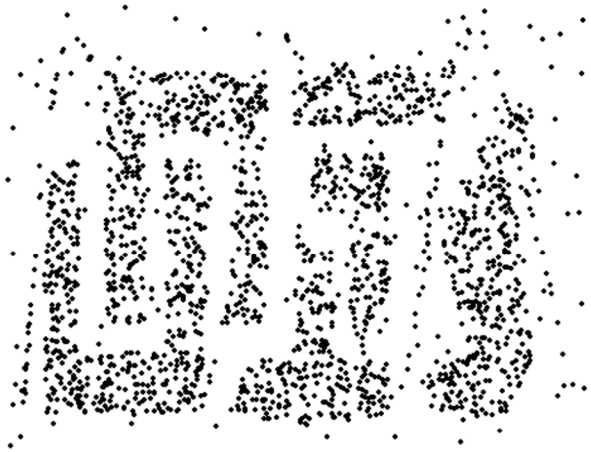
Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
- Sampling is used because obtaining and processing the entire set of data (population) of interest is too expensive or time consuming.
- Sampling will work almost as well as using the entire data sets, if the sample is representative



8000 points



2000 Points

500 Points

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
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

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

Summary Statistics

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

Frequency and Mode

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

Percentiles

- 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 p *th* 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 $x_{50\%}$.

e.g., 90th percentile of exam score: 90% of the students scored less than me.

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 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.

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

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

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

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

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

AAD: Absolute Average Deviation
MAD: Median Absolute Deviation

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike
  - Is higher when objects are more alike.
  - Often falls in the range [0,1] : if p=q, s(p, q) = 1

- Dissimilarity = "Distance"
  - Numerical measure of how different
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0 : if p=q, d(p, q) = 0
  - Upper limit varies

- Proximity: similarity or dissimilarity

Common Properties

- Similarities
  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)

- Distances,
  1. $d(p, q) \geq 0$ for all p and q

     $d(p, q) = 0$ only if $p = q$ (Positive definiteness)
  2. $d(p, q) = d(q, p)$ for all p and q (Symmetry)
  3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r (Triangle Inequality)

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

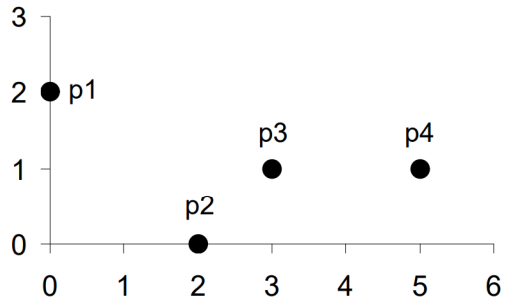
(Dissimilarity)

- Euclidean Distance (distance b/w points)

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

n : the number of dimensions (attributes)

p_k and q_k : the k[th] attribute of data objects p and q



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

Distance Matrix

	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

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

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

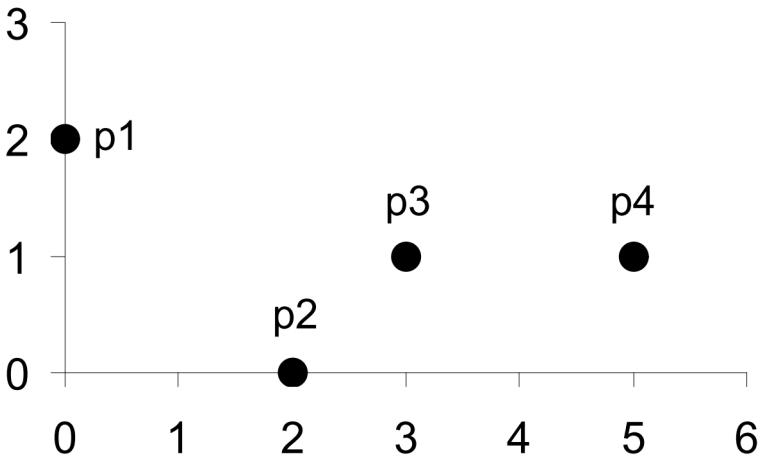
r : a parameter, n : the number of attributes

p_k and q_k : the kth attributes of data objects p and q

- $[r = 1]$ 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
- $[r \to \infty]$ "supremum" (L_{max} norm, L_∞ norm) distance
  - This is the maximum difference between **any component of the vectors**

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 Matrices

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 • indicates vector dot product and $||d||$ is the length of vector d.

Jaccard measure + non-binary vectors

- 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**

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

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

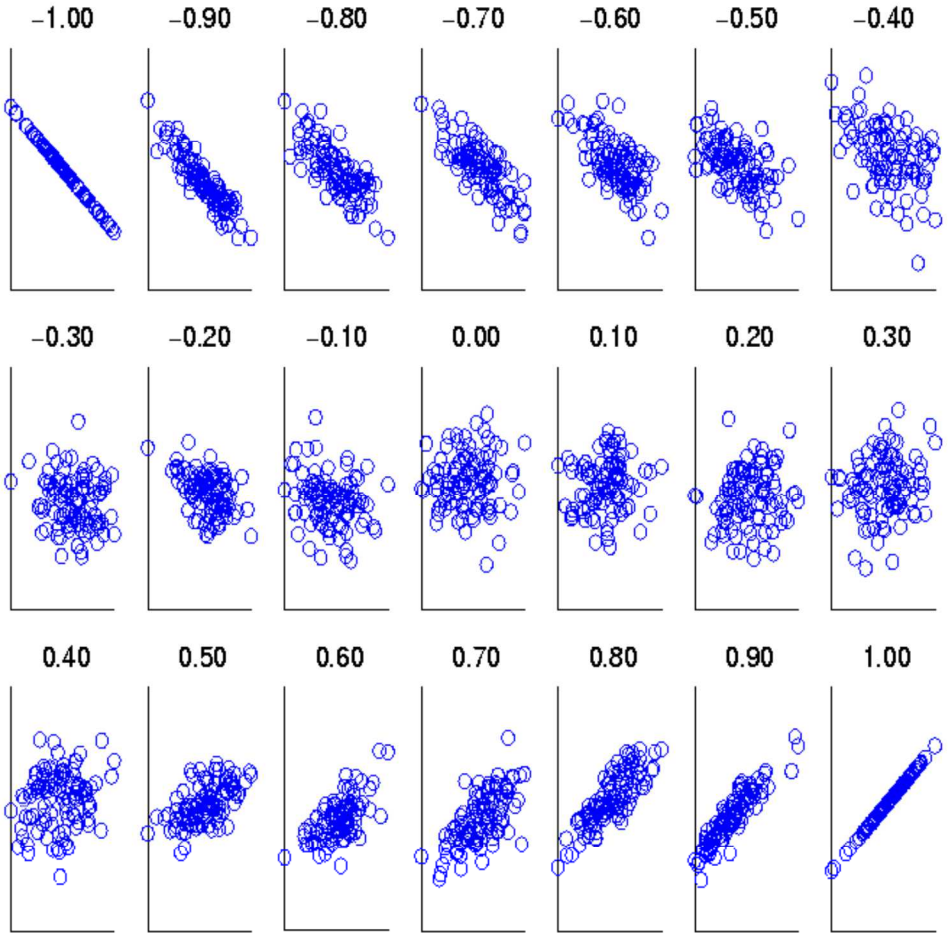
$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = 0.3150$ (1→0° : same direction but different length

0→90° : do not share, low similarity)

Correlation

- Measure the linear relationship between objects
- Standardization, then compute into [-1, 1]



Density

- Euclidean density: # of points per unit volume
  - Cell-based: Divide region into a number of rectangular cells of equal volume and count
  - Center-based: Count the number of points within a specified radius of the center point

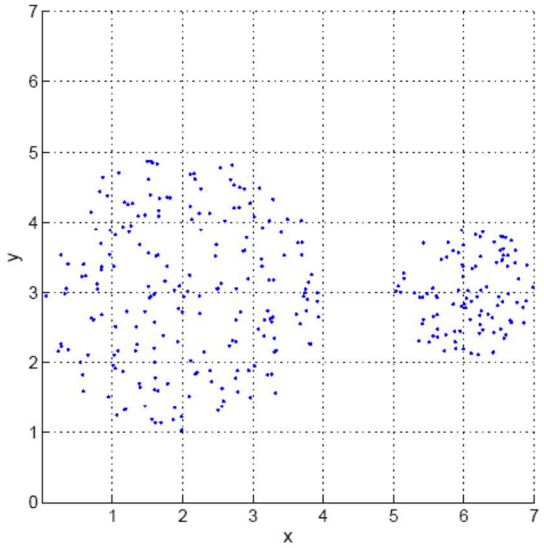


Figure 7.13. Cell-based density.

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.

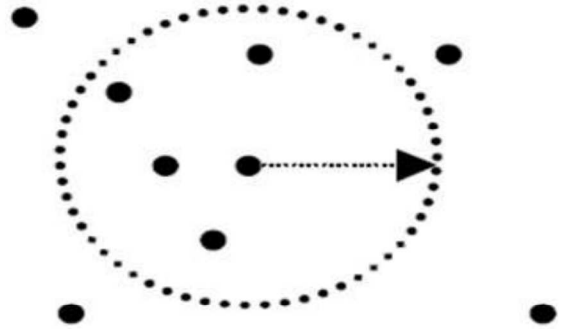


Figure 7.14. Illustration of center-based density.

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)
  - 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

What is data exploration?

- Exploratory Data Analysis (EDA): An approach for data analysis that employs a variety of techniques (mostly graphical) to
  - maximize insight into a data set
  - uncover underlying structure
  - extract important variables
  - detect outliers and anomalies
  - test underlying assumptions
  - develop parsimonious models
  - determine optimal factor settings
- Techniques
  - Plotting the raw data (e.g. data traces, histograms, probability plots)
  - Plotting simple statistics (mean and standard deviation plots, box plots)
  - Positioning such plots so as to maximize our natural pattern-recognition abilities

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://archive.ics.uci.edu/ml/
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal(꽃받침) width and length
    - Petal(꽃잎) width and length

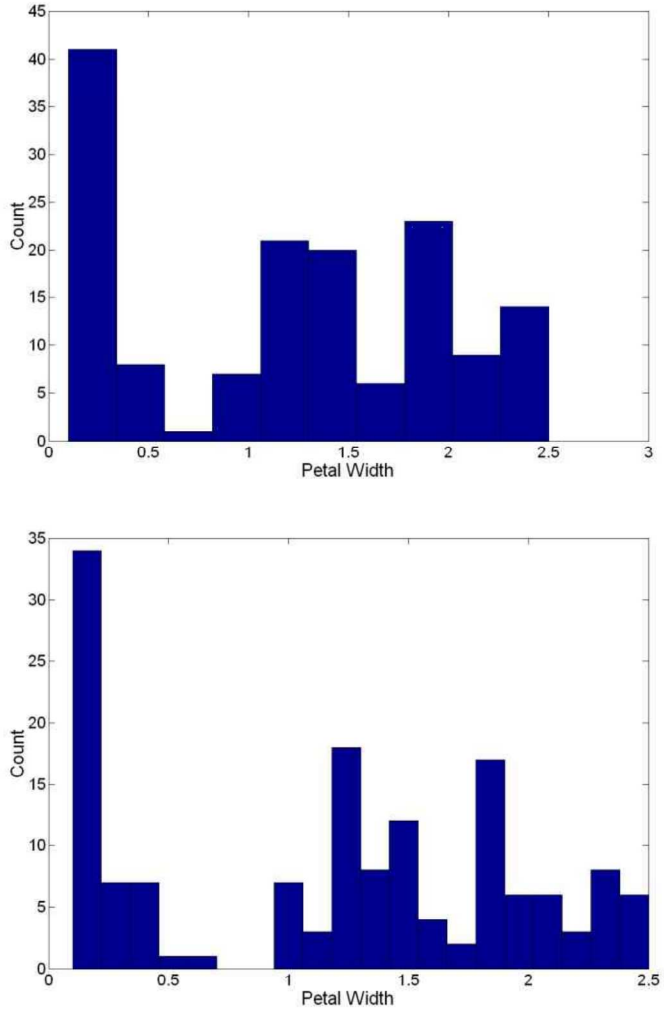
Visualization

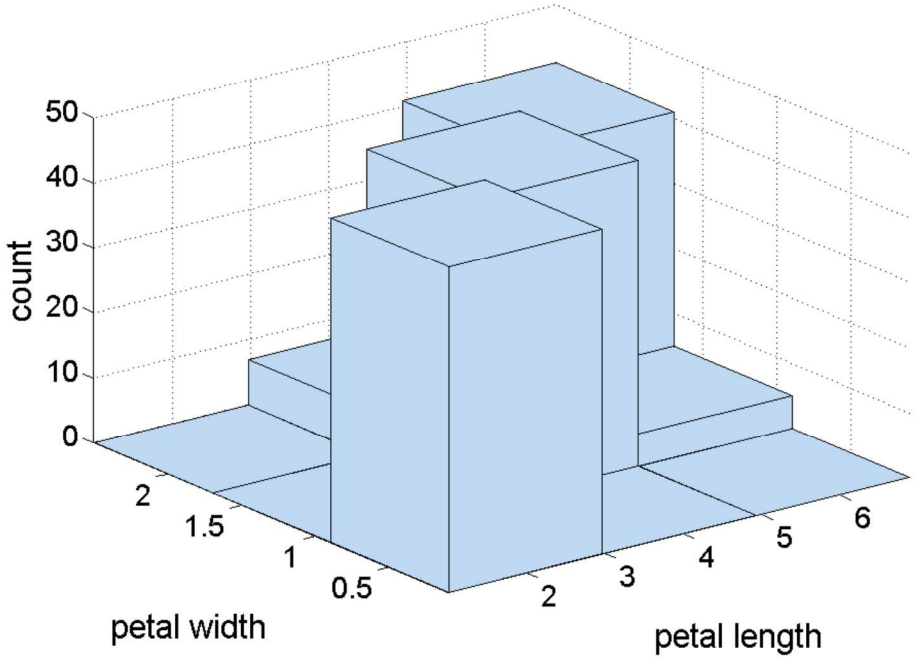
- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- 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

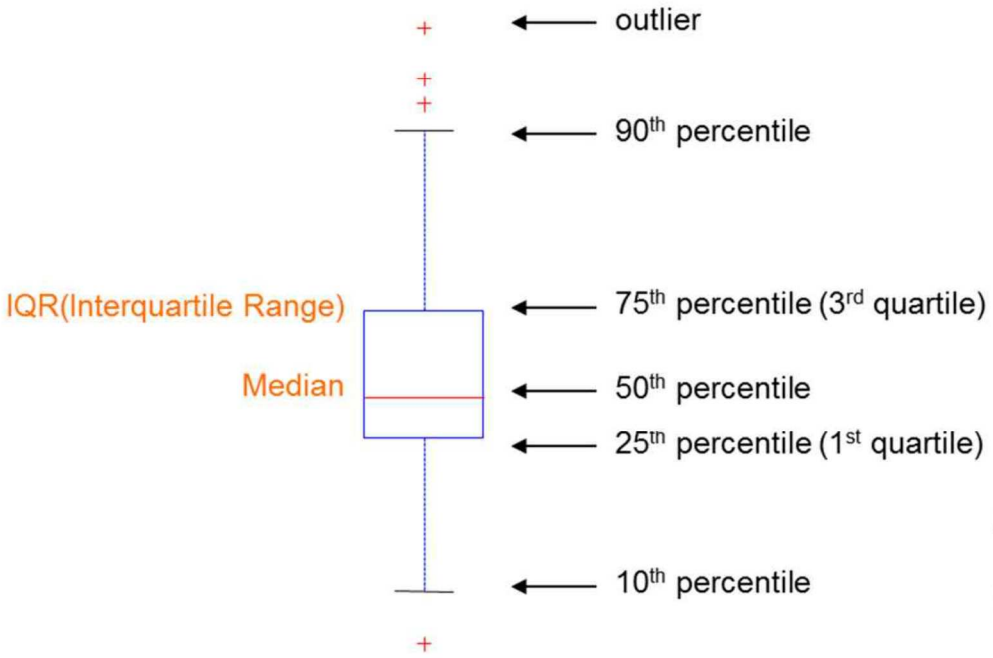
Visualization Techniques: Histograms

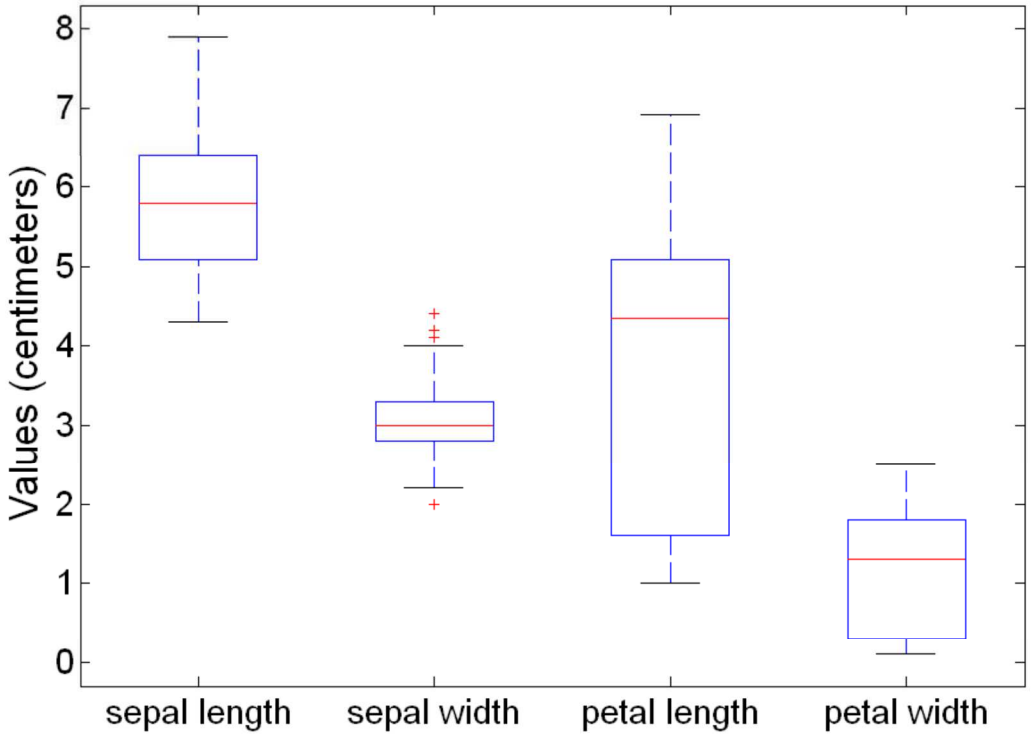
- Example: Petal Width (10 and 20 bins, respectively)



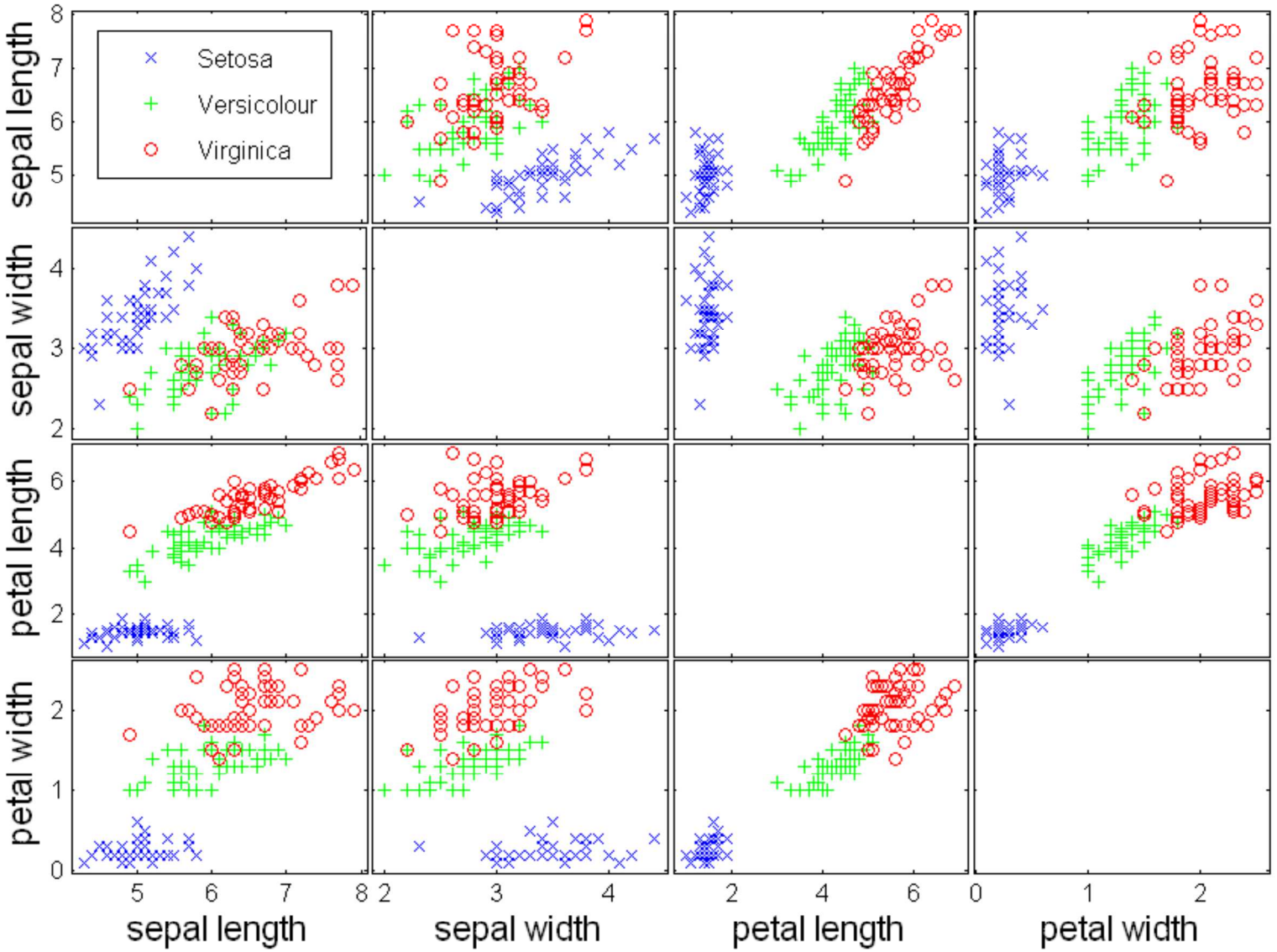


Visualization Techniques: Box Plots



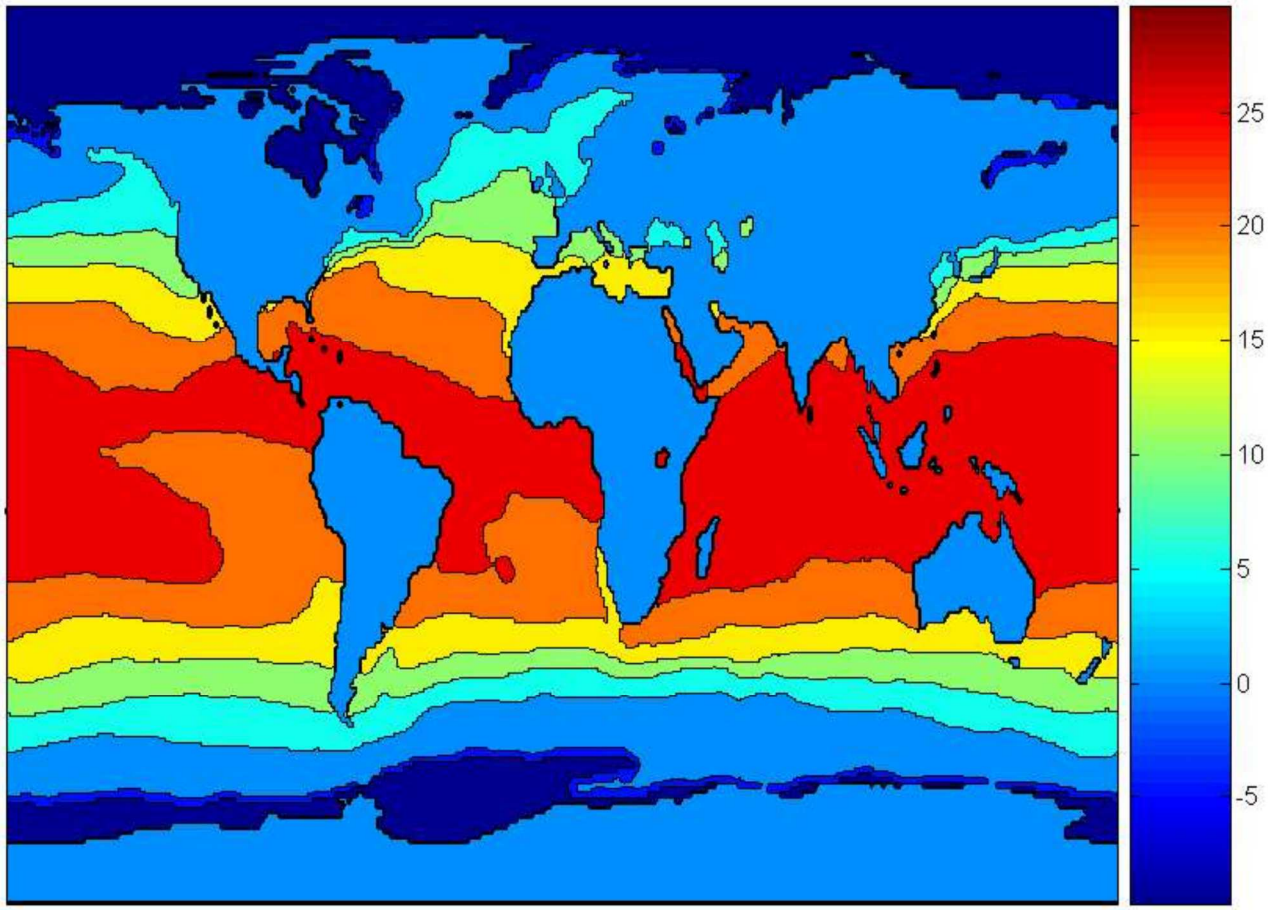


Visualization Techniques: Scatter Plots



Visualization Techniques: Contour Plots

- When a continuous attribute is measured on a spatial grid
- They partition the plane into regions of similar values

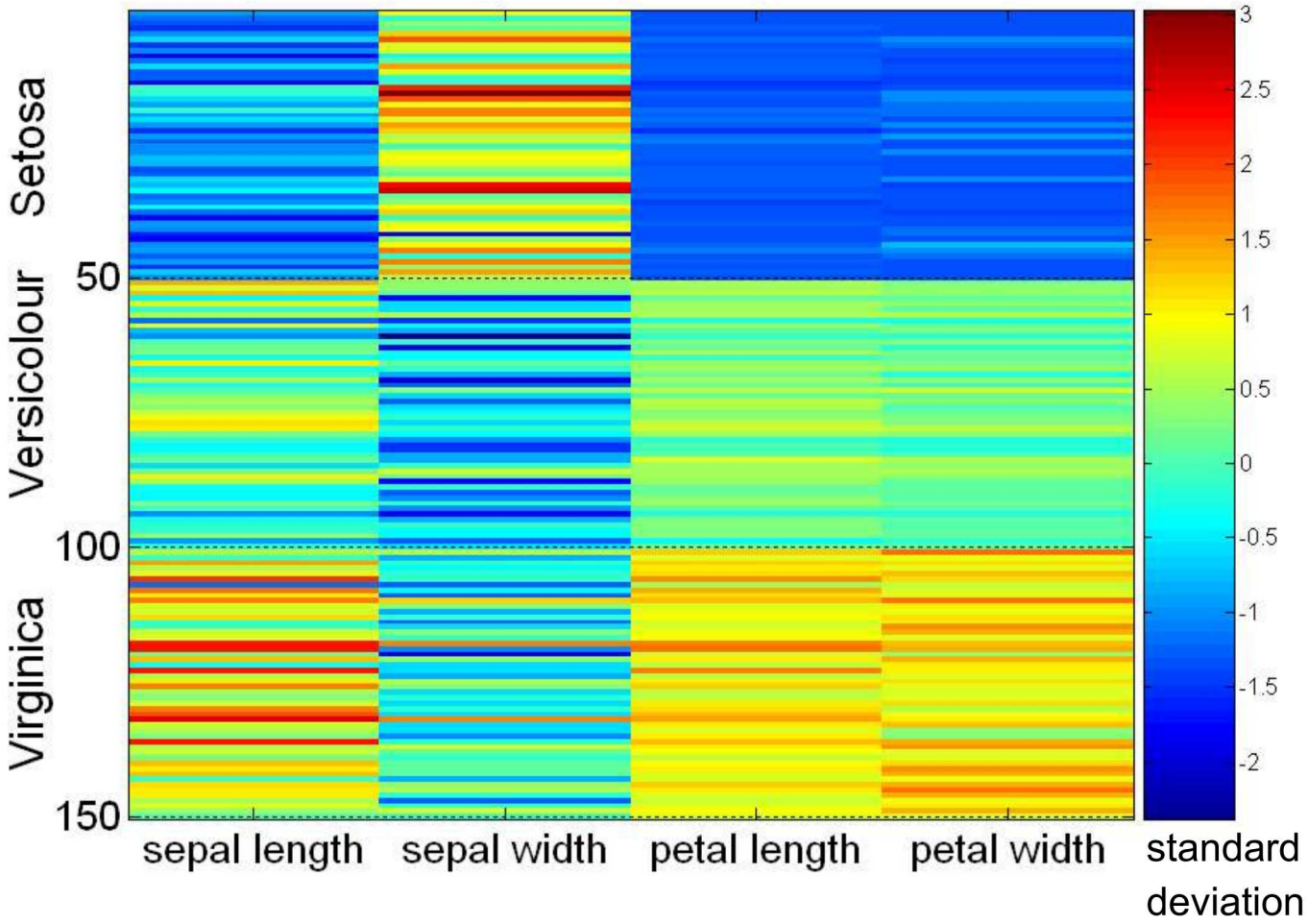


Visualization Techniques: Matrix Plots

- 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

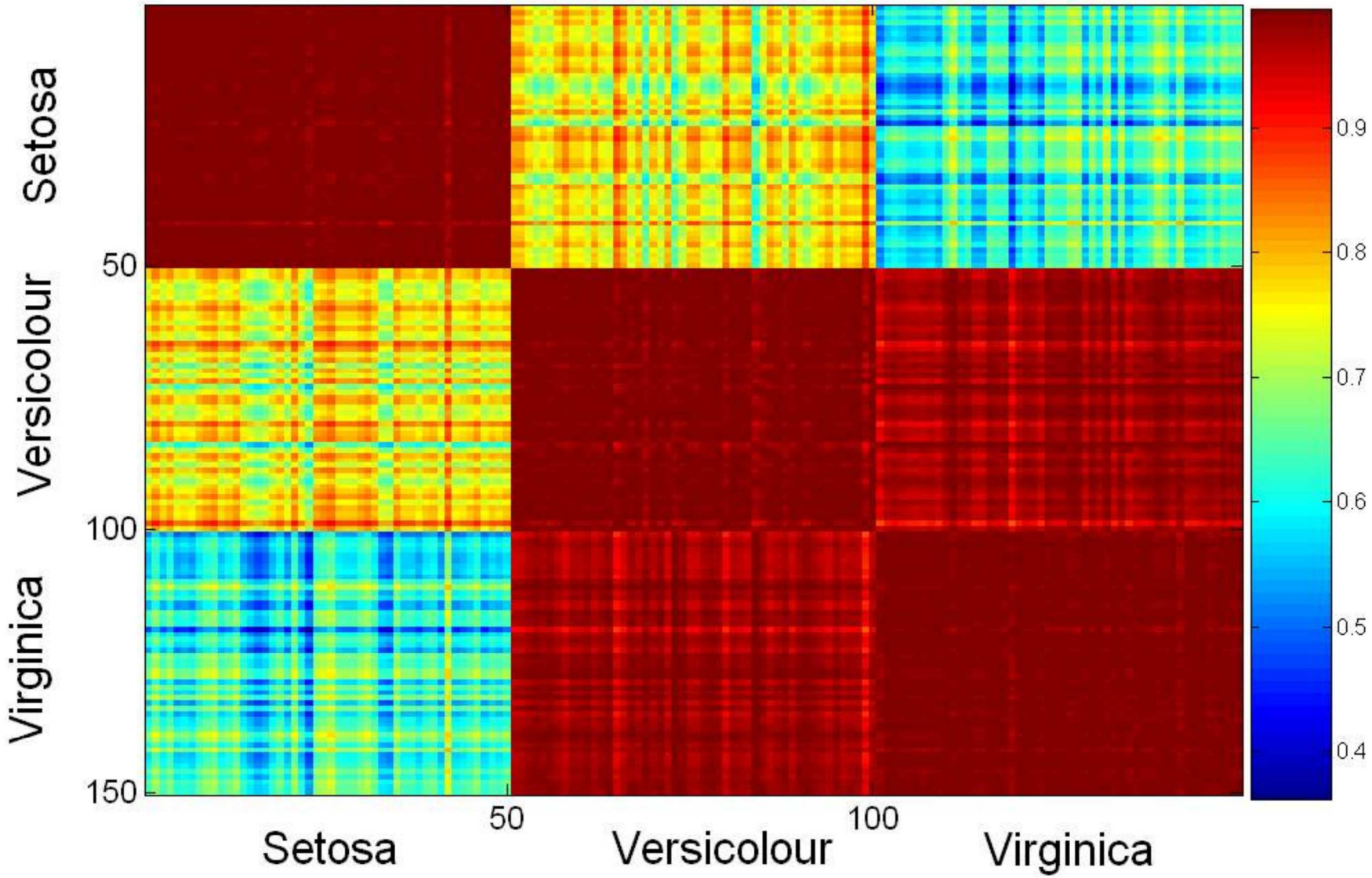
- Simple data matrix & Correlation matrix

Visualization of the Iris Data Matrix

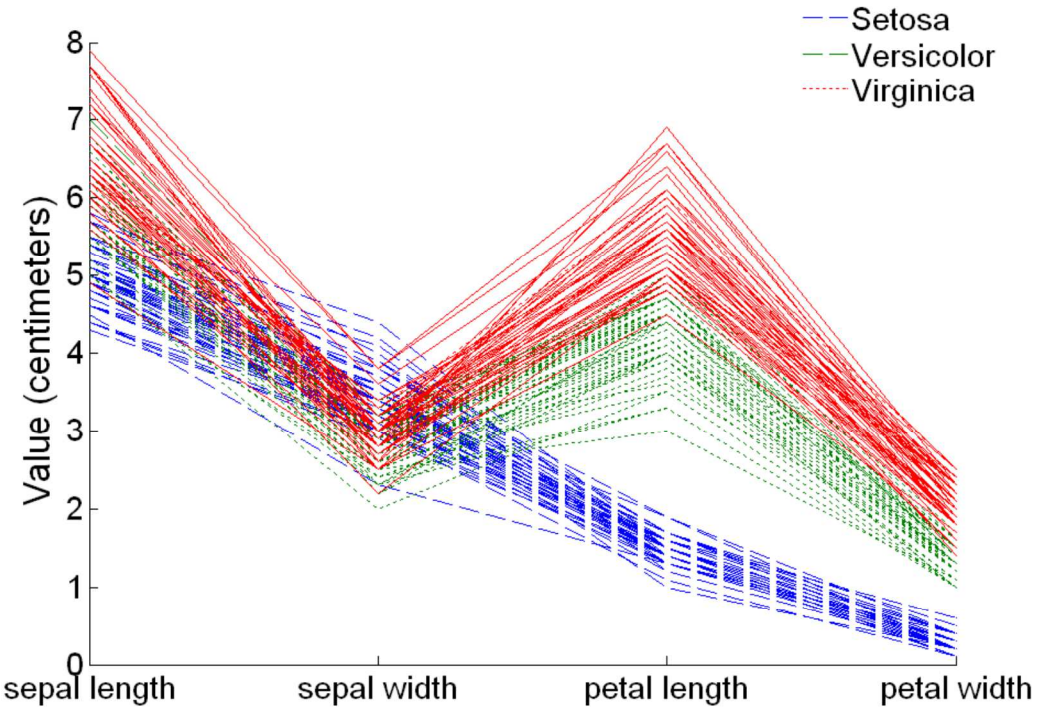


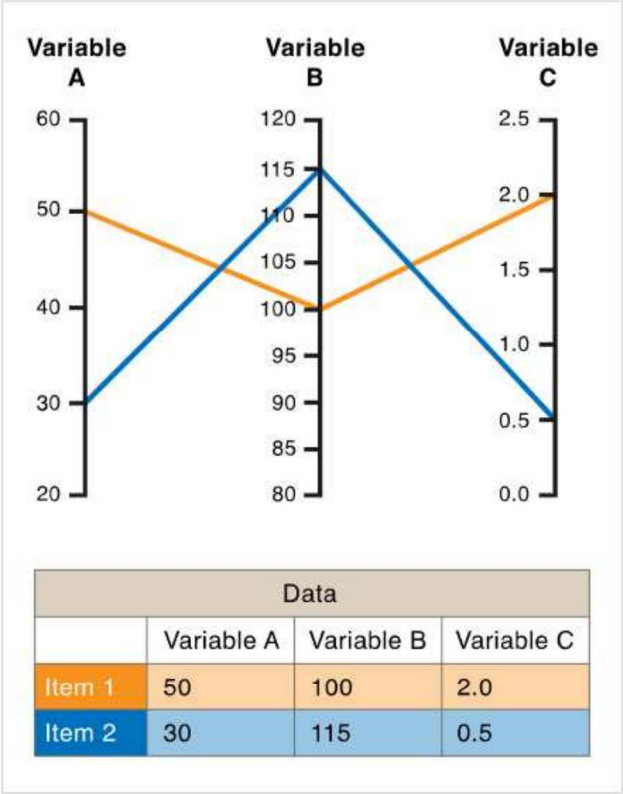
Setosa flowers have petal width and length well below the average.
Versicolour flowers have petal width and length around average.
Virginica flowers have petal width and length above average.

Visualization of the Iris Correlation Matrix



Parallel Coordinates Plots for Iris Data





Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

If lines cross a lot, the picture can become confusion, and thus, it can be desirable to order the coordinate axes to obtain sequences of axes with less crossover in order to identify the patterns better.

Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data(multivariate, numerical 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
  - e.g. comparing computer or car specs across different models

Visualization Techniques: Word Clouds

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

Other Visualization Techniques

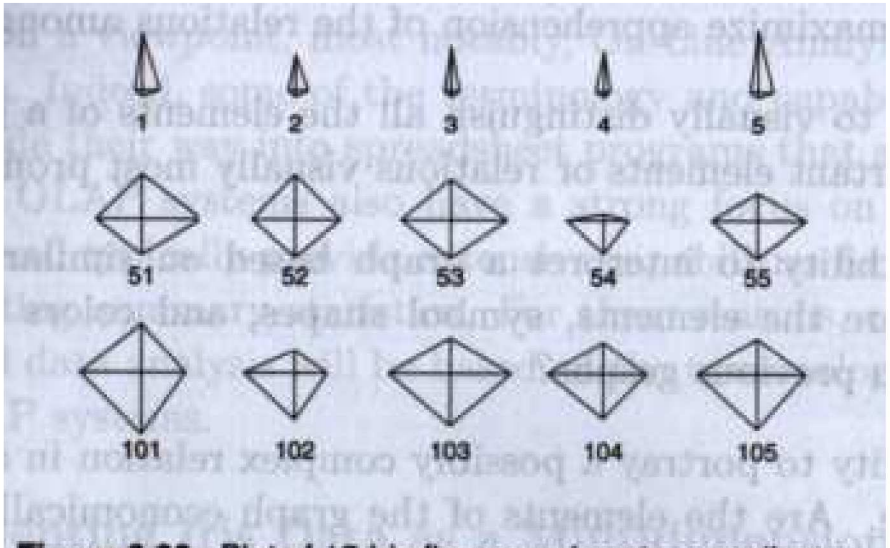
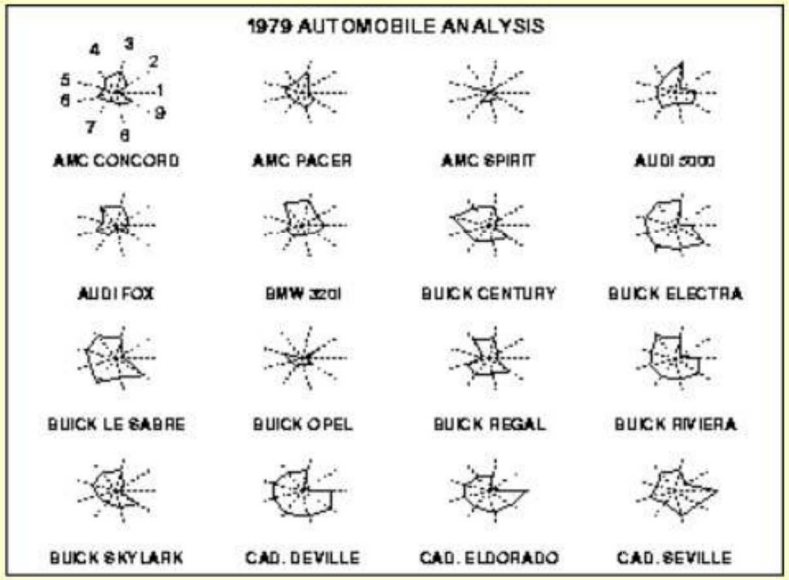


Figure 3.28. Plot of 15 Iris flowers using star coordinates.

Figure 3.29. A plot of 15 Iris flowers using Chernoff faces.

Each star represents a single observation.
(similar to radar chart, spider diagram)

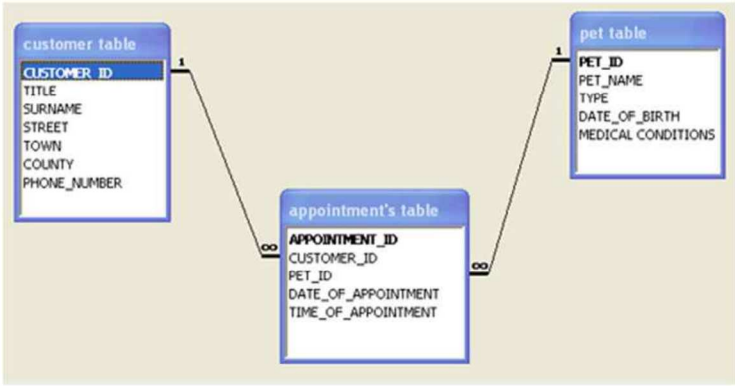
1. Price
2. Mileage (MPG)
3. 1978 Repair Record (1 = Worst, 5 = Best)
4. 1977 Repair Record (1 = Worst, 5 = Best)
5. Headroom
6. Rear Seat Room
7. Trunk Space
8. Weight
9. 9
10. Length

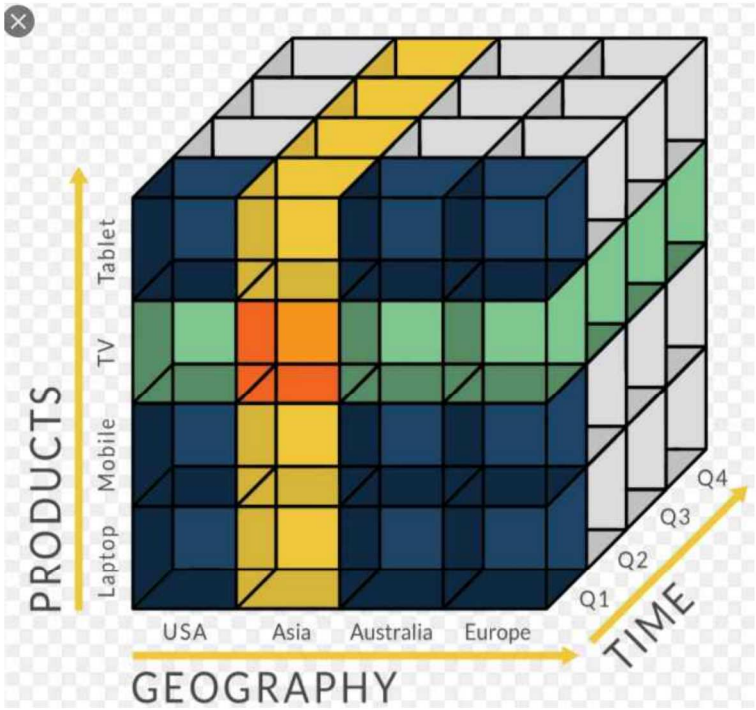


e.g., sepal length = size of face, sepal width = forehead relative arc length
petal length = shape of forehead, petal width = shape of jaw

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 typically uses a multidimensional array representation.
- There are a number of data analysis and data exploration operations that are easier with such a data representation.





Creating a Multidimensional Array

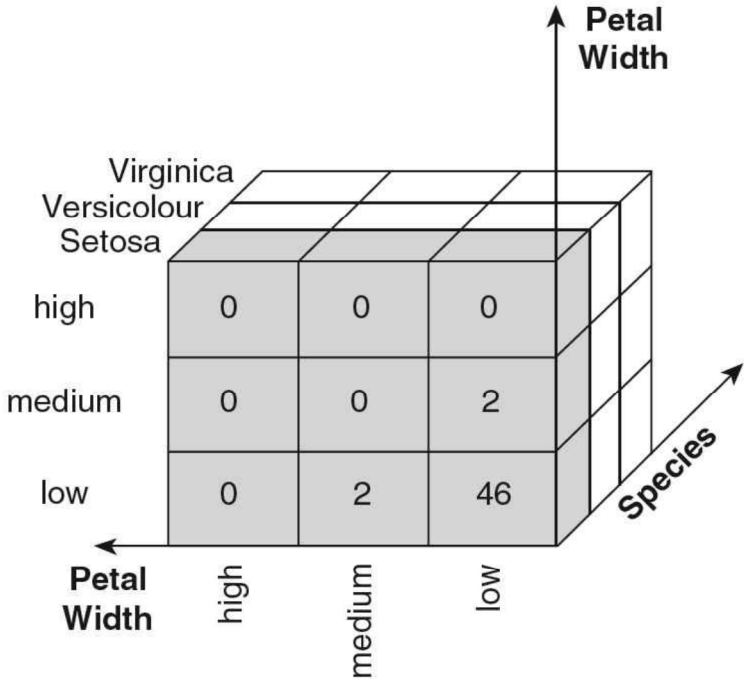
- Two key steps in converting tabular data into a multidimensional array.
  - First, identify which attributes are to be the dimensions and which attribute is to be the target attribute whose values appear as entries in the multidimensional array.
    - The attributes used as dimensions must have discrete 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

dimension target



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

Setosa

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

Versicolour

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

Virginica

Table 3.11. Sales revenue of products (in dollars) for various locations and times.

Product ID	Location	Date	Revenue
⋮	⋮	⋮	⋮
1	Minneapolis	Oct. 18, 2004	$250
1	Chicago	Oct. 18, 2004	$79
⋮	⋮	⋮	
1	Paris	Oct. 18, 2004	301
⋮	⋮	⋮	⋮
27	Minneapolis	Oct. 18, 2004	$2,321
27	Chicago	Oct. 18, 2004	$3,278
⋮	⋮	⋮	
27	Paris	Oct. 18, 2004	$1,325
⋮	⋮	⋮	⋮

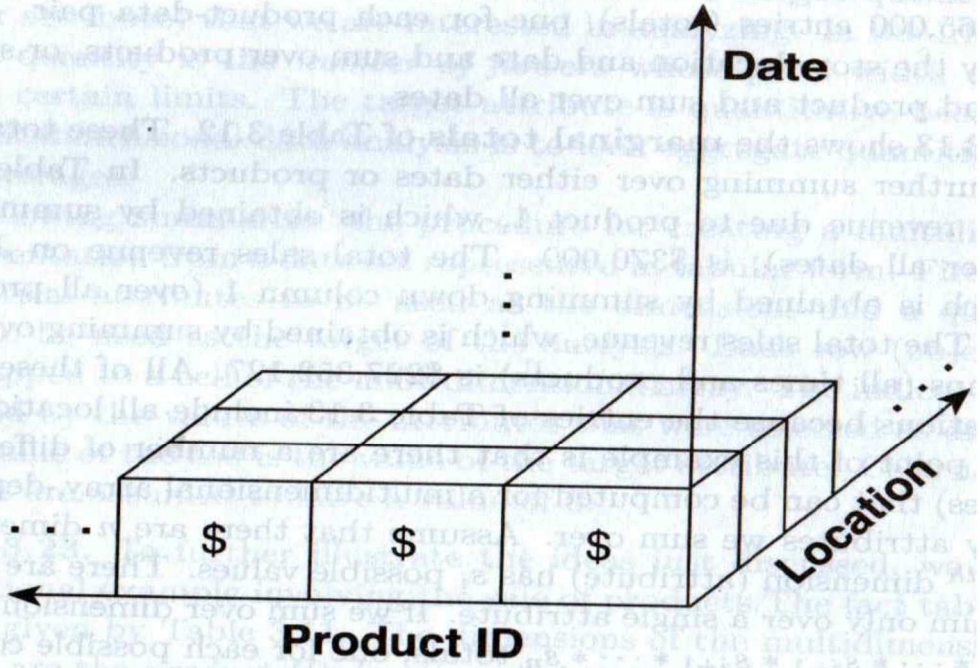


Figure 3.31. Multidimensional data representation for sales data.

Table 3.12. Totals that result from summing over all locations for a fixed time and product.

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004
1	$1,001	$987	...	$891
⋮	⋮			⋮
27	$10,265	$10,225	...	$9,325
⋮	⋮			⋮

Table 3.13. Table 3.12 with marginal totals.

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