Introduction to Data Analytics

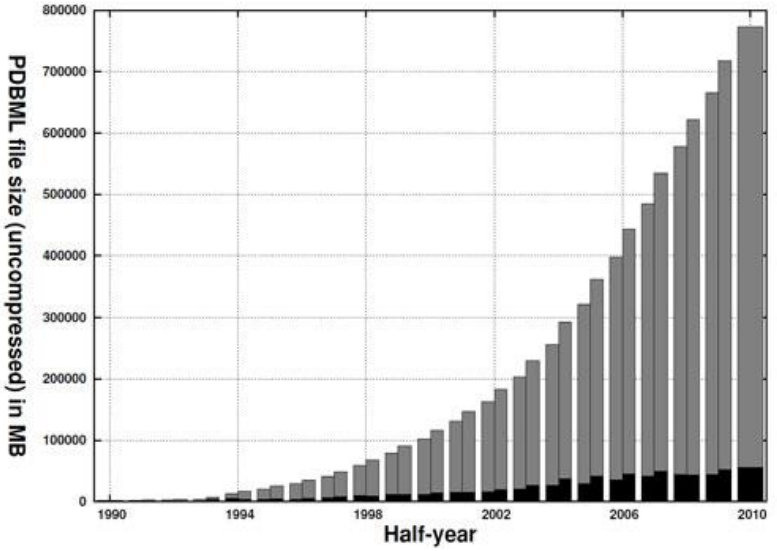
Xin Gao
Xin.gao@kaust.edu.sa
July 27, 2022
SDU

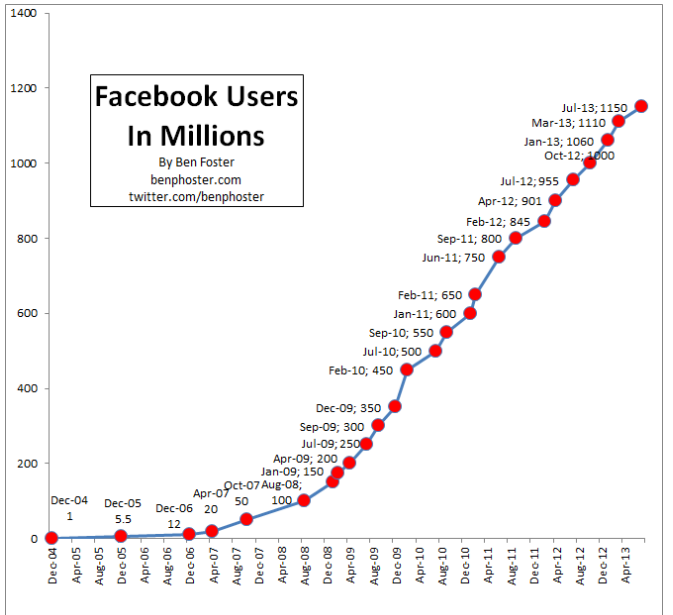
What is Data Mining?

- Data mining is the process of **discovering unknown/new patterns** from large **data** sets involving methods from **statistics** and **artificial intelligence** but also **database management**.
  - Valid: hold on new data with some certainty.
  - Novel: non-obvious to the system.
  - Understandable: humans should be able to interpret the patterns.

Why Data Mining? – Commercially

- Massive of data is being collected and warehoused
  - Web data: facebook, google, amazon, twitter.
  - Biological data: DNA sequences, protein structures.
  - Bank/credit card transaction data: Samba, Paypal.
  - Mobile data: AT&T, Mobily.
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Provide better, customized services



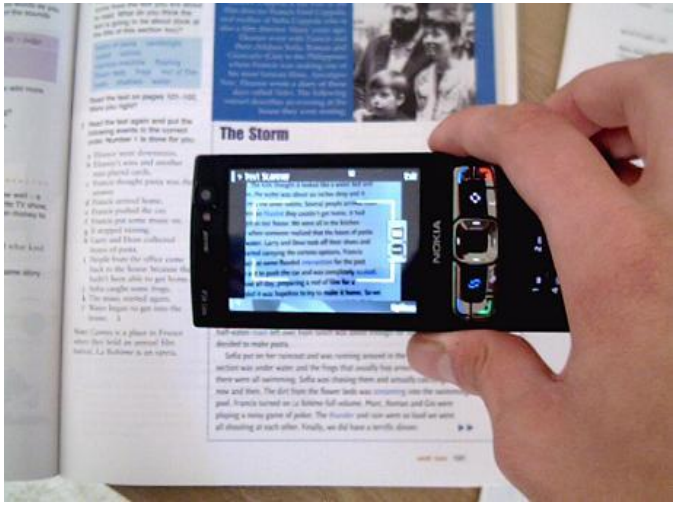


Why Data Mining? – Scientifically

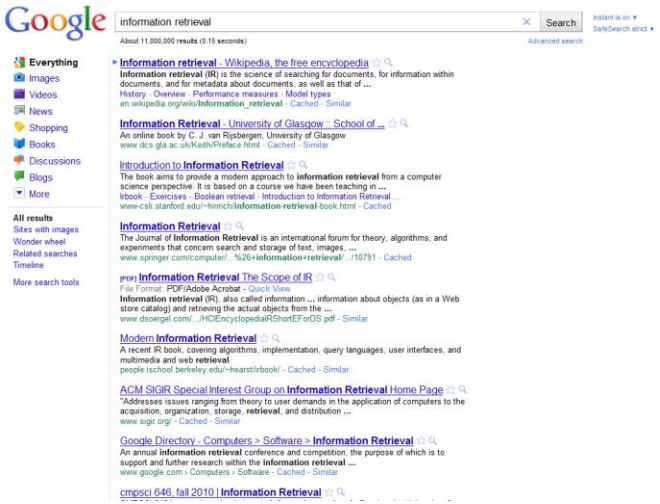
- Data collected and stored at enormous speeds (GB/hour)
  - Remote sensors on a satellite.
  - Microarrays generating gene expression data.
  - NGS generating DNA sequences.
  - Scientific simulations generating terabytes of data.
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in extracting features from data, classifying data, visualizing data, or interpret data patterns

Examples

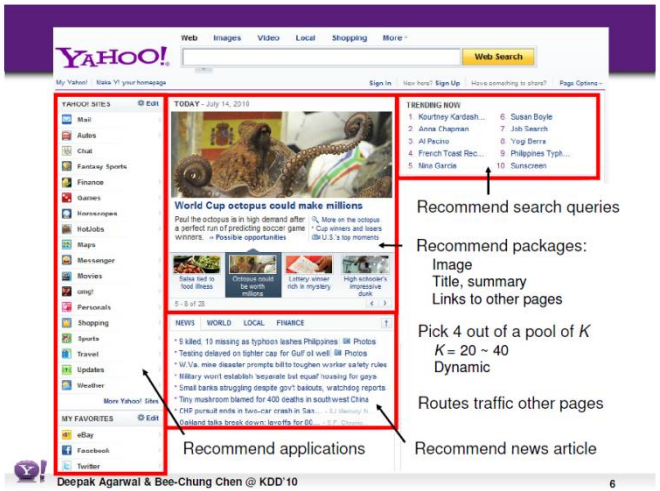
Optical character recognition



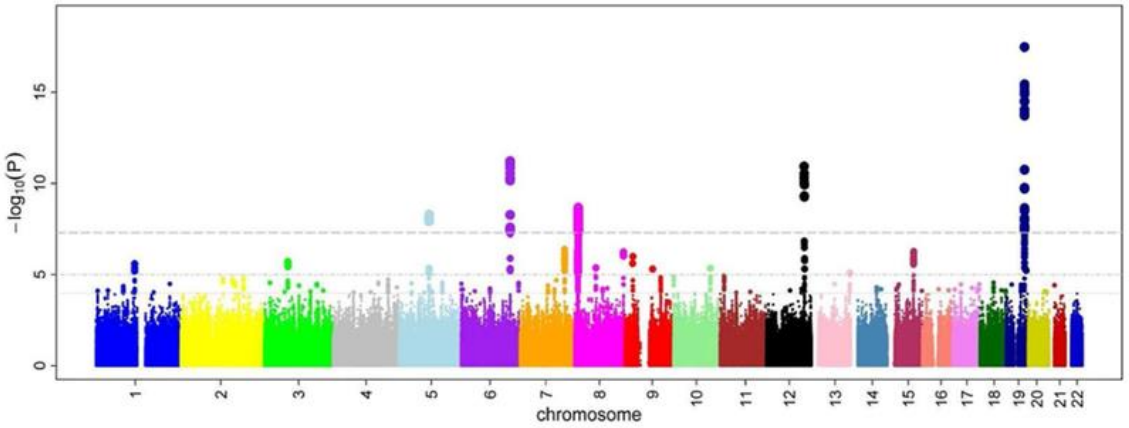
Information retrieval



Recommendation systems

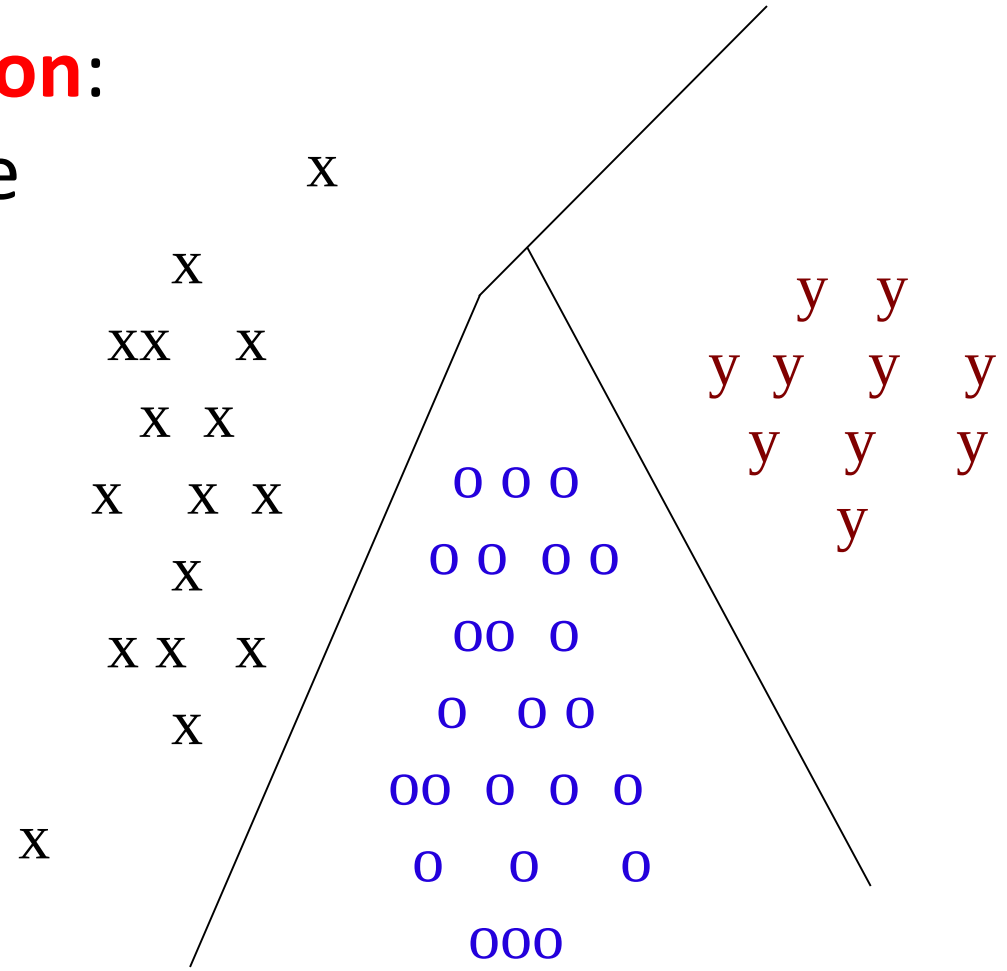


Genome-wide association study



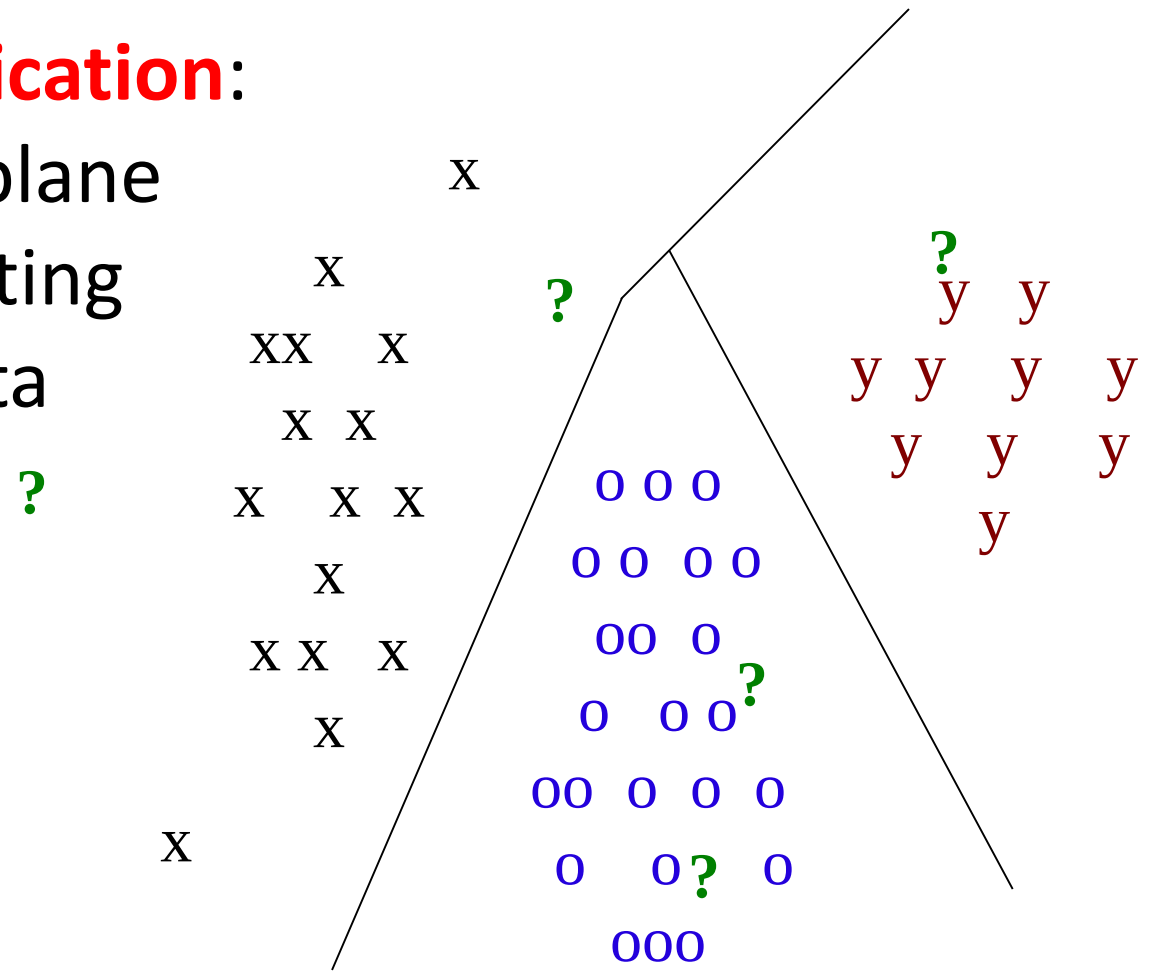
Typical Kinds of Patterns

- **Classification**: Hyperplane separating the data



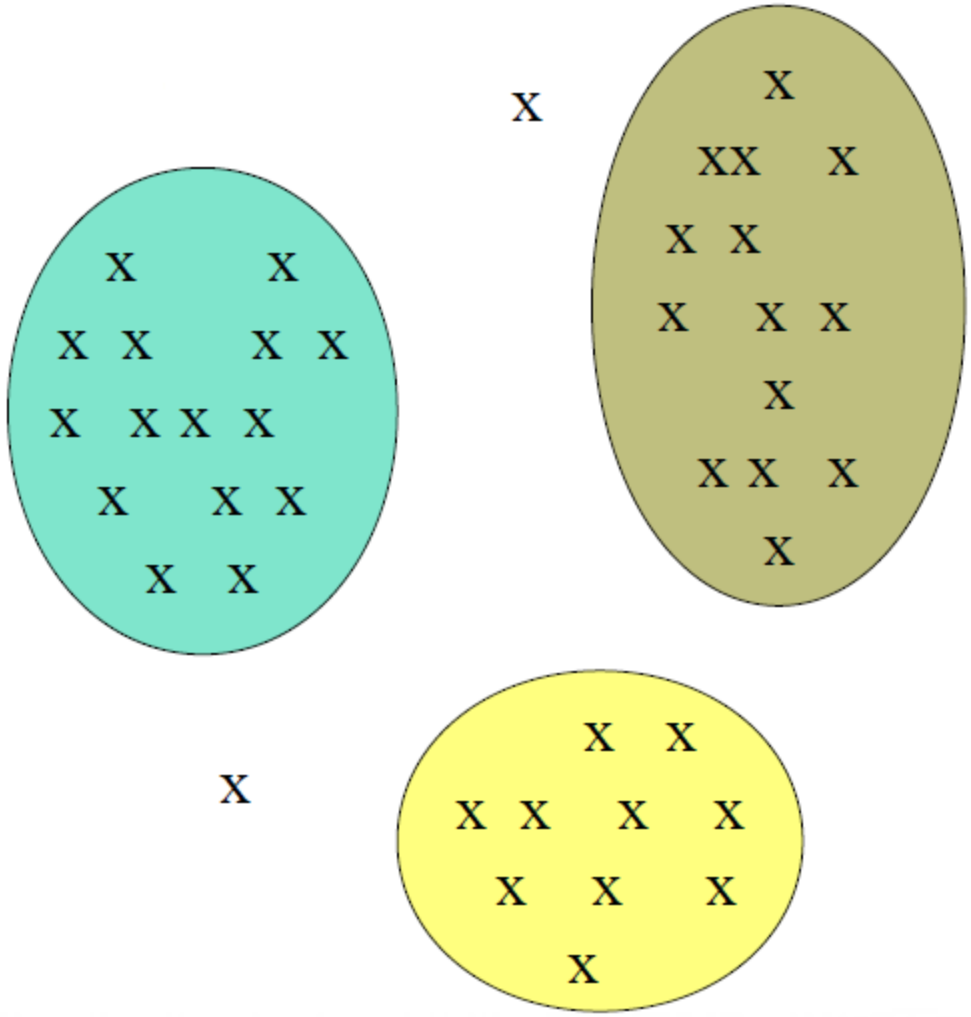
Typical Kinds of Patterns

- **Classification**: Hyperplane separating the data



Typical Kinds of Patterns

- **Clustering**: Groups of similar objects



Typical Kinds of Patterns

- **Frequent itemsets**
  - A common marketing problem: examine what people buy together to discover patterns.
    - What pairs of items are unusually often found together at the supermarket checkout?
      - Answer: milk and cereal.
  - An association problem: examine the co-effects of different factors.
    - Which SNPs usually happens together?

Questions to Answer

- **What is data?**
- **What kinds of attributes can be used to describe objects?**
- How data are different in types?
- How can we improve data quality?
- How to measure similarities between objects?

What is Data?

- Collection of data objects and their attributes
- An attribute (feature) 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 describes an object (sample)
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Types of Attributes

- Nominal: differentiate objects based on names
  - Examples: ID numbers, eye color, zip codes
- Ordinal: allow rank order, but not degree of difference between them
  - Examples: sick v.s. health, guilty v.s. innocent, completely agree v.s. agree v.s. do not agree, tall v.s. medium v.s. short
- Interval: allow degree of difference, but not the ratio between them
  - Examples: temperature, calendar dates
- Ratio: measurement of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind
  - Examples: length, time, counts

Name	ID	Eye color	Height	Grade of CS220	Arrival date	How many courses taken
Tom	123456	brown	tall	A-	Sep 1, 2012	5

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Questions to Answer

- What is data?
- What kinds of attributes can be used to describe objects?
- **How data are different in types?**
- How can we improve data quality?
- How to measure similarities between objects?

Types of Data Sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - Social network
  - Protein-protein interaction network
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes
- Data Matrix
  - The data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
  - Data set can be represented by an *n* by *m* matrix, where there are *n* rows, one for each object, and m columns, one for each attribute

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Document Data

- Each document becomes a `term' vector,
  - Each term is a component (attribute) of the vector,
  - The value of each component is 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

bag-of-words

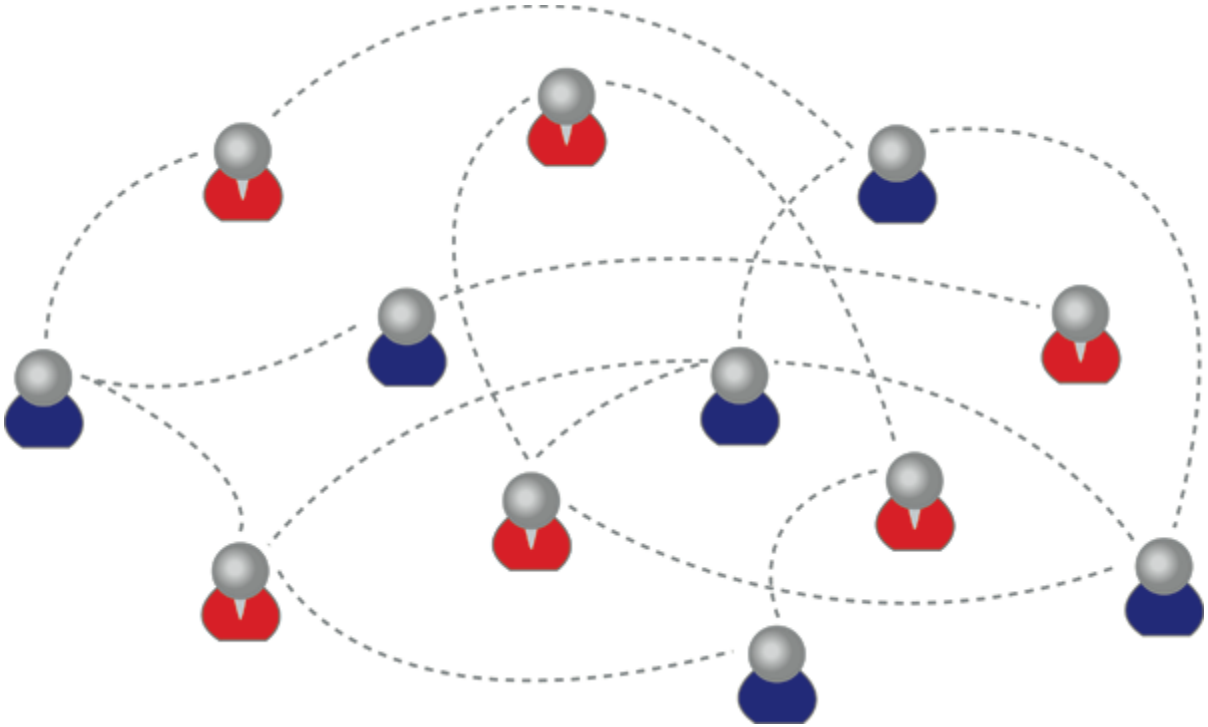
Transaction Data

- 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

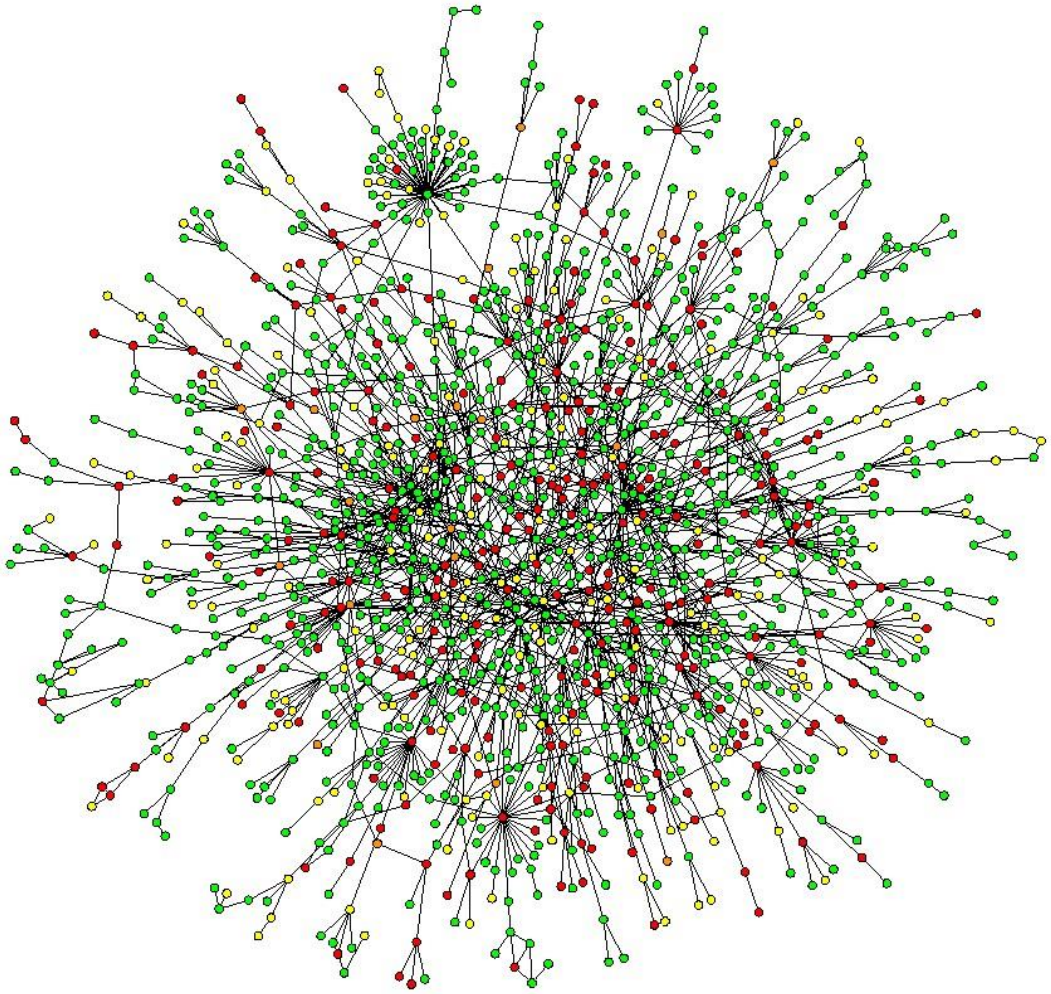
Graph Data

- Examples: Social network



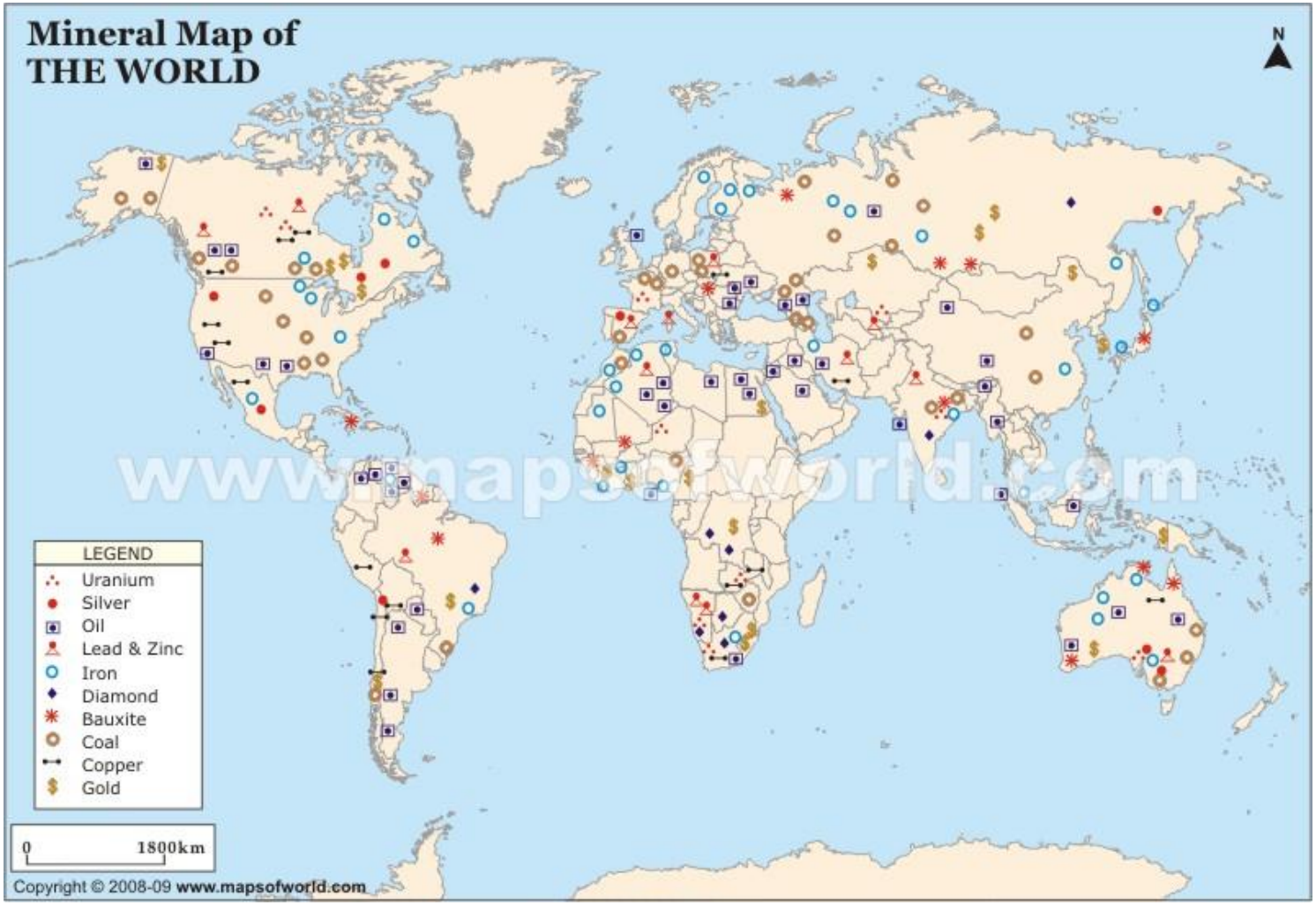
Graph Data

- Examples: PPI network



Ordered Data

- Spatial data: with geographic locations

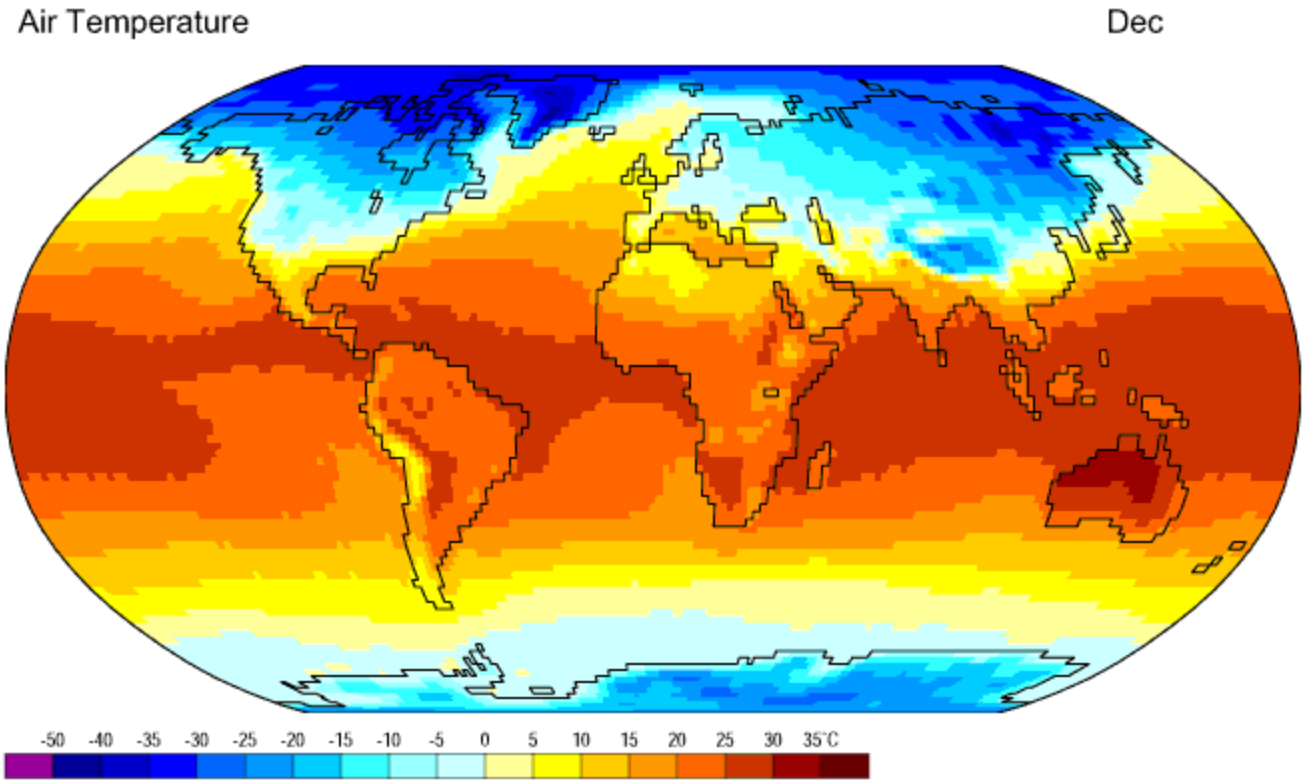


Ordered Data

- Temporal data: with built-in support for handling data involving time

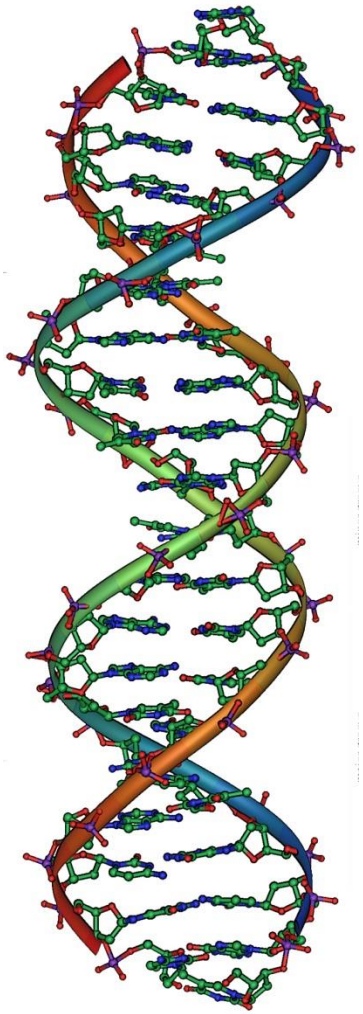
Ordered Data

- Spatial and temporal data:



Ordered Data

- Sequential data



ATGAAAAAGACAGCTATCGCGATTGCAGTGGCACTGGCTGGTTTCGCTACCGTGGCC
CAGGCGGCCTCTGAGGGAAACAGTGACTGCTACTTTGGGAATGGGTCAGCCTACCG
TGGCACGCACAGCCTCACCGAGTCGGGTGCCTCCTGCCTCCCGTGGAATTCCATGAT
CCTGATAGGCAAGGTTTACACAGCACAGAACCCCAGTGCCCAGGCACTGGGCCTGG
GCAAACATAATTACTGCCGGAATCCTGATGGGGATGCCAAGCCCTGGTGCCACGTG
CTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGATGTGCCCTCCTGCTCCACCTGC
GGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAAGGAGGGCTCTTCGCCGA
CATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCACAGGAGGTCGCCCGG
AGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATTCTCTCTGCCGC
CCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTGGGCAGAAC
ATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATACATTG
TCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTGA
AATCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTC
CCCCGGCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGC
AAGCATGAGGCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGA
CTGTACCCATCCAGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGAC
AACATGCTGTGTGCTGGAGACACTCGGAGCGGCGGGCCCCAGGCAAACTTGCACGA
CGCCTGCCAGGGCGATTCGGGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGA
CTTTGGTGGGCATCATCAGCTGGGGCCTGGGCTGTGGACAGAAGGATGTCCCGGGT
GTGTACACAAAGGTTACCAACTACCTAGACTGGATTCGTGACAACATGCGACCG
(SEO ID NO:2)

Questions to answer

- What is data?
- What kinds of attributes can be used to describe objects?
- How data are different in types?
- **How can we improve data quality?**
- How to measure similarities between objects?

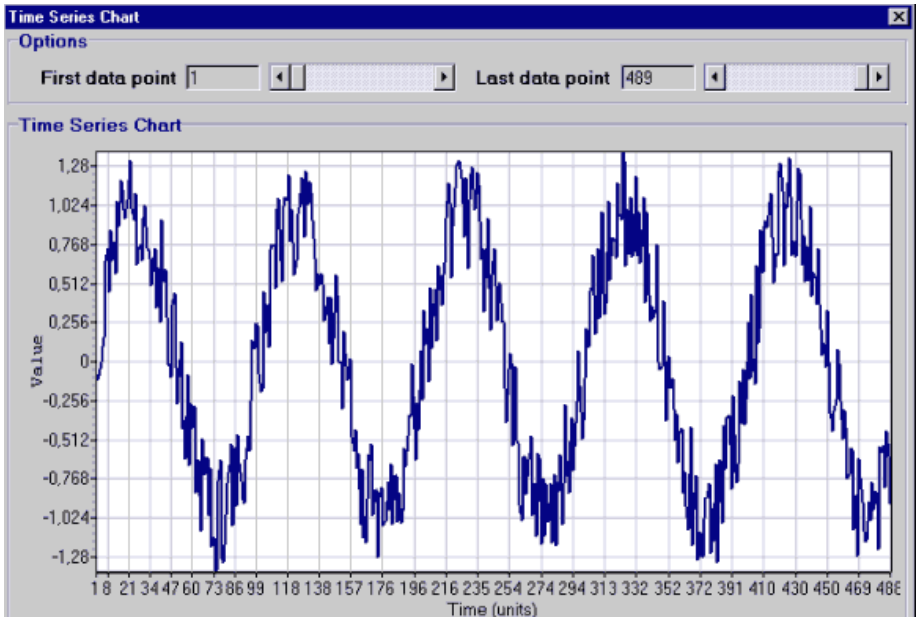
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

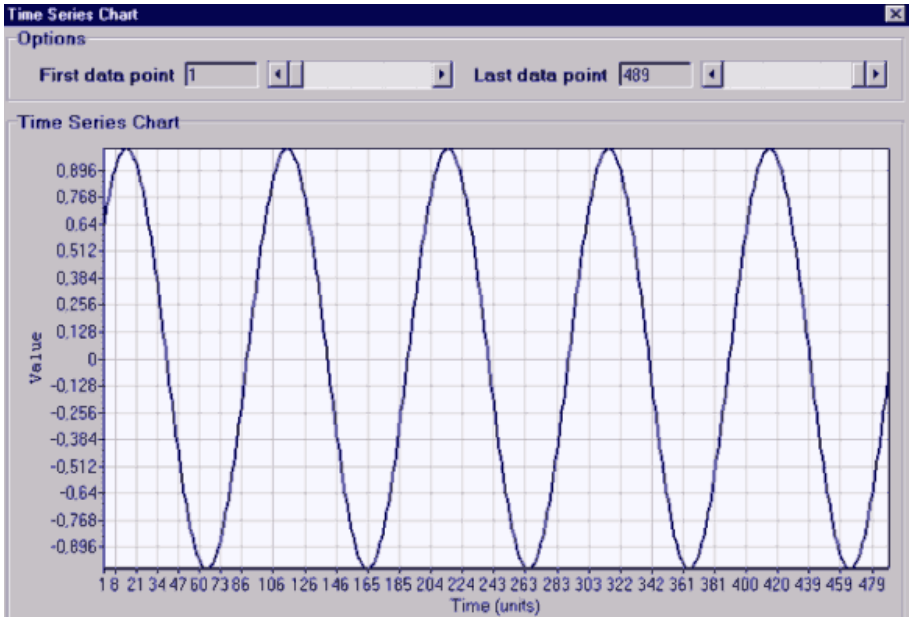
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

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



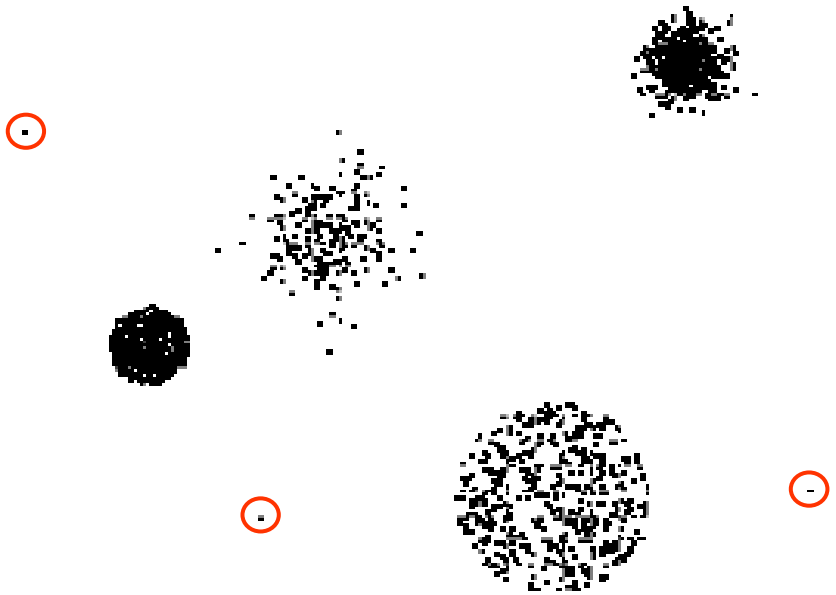
A sine wave with noise



The denoised sine wave

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
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

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
  - Example: the same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

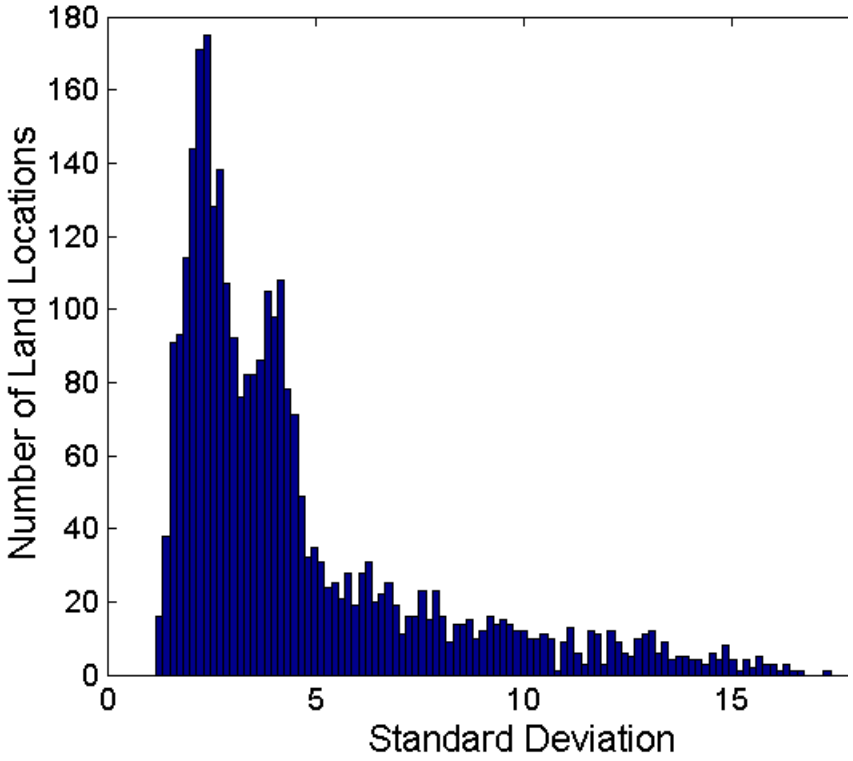
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Attribute Transformation

Aggregation

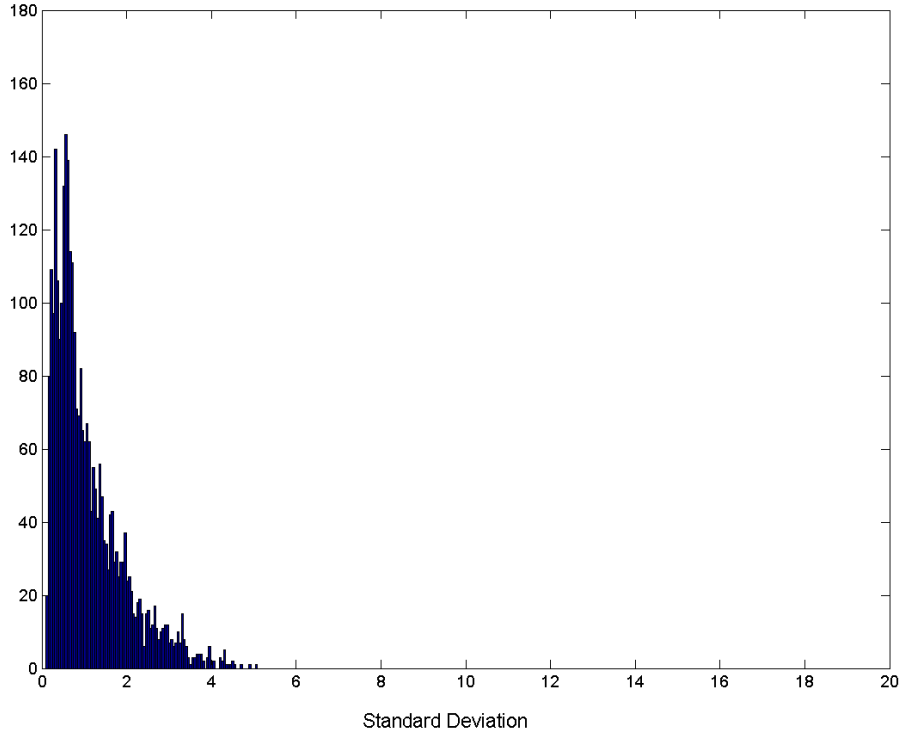
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

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.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Sampling is used because processing the entire set of data of interest is too expensive or time consuming.
- Effective sampling: using the representative samples will work almost as well as using the entire data sets

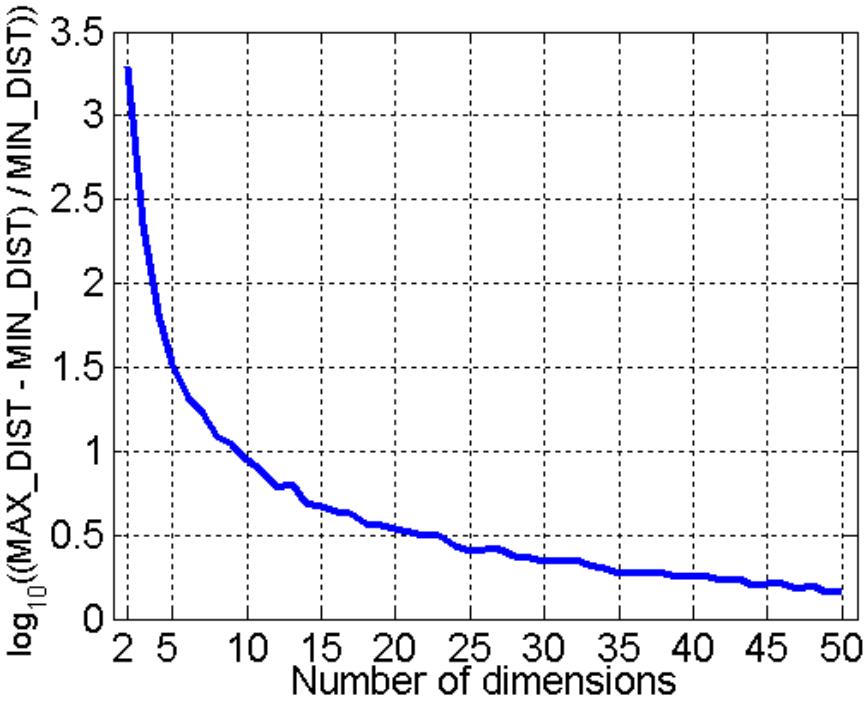
Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement (dependent)
  - As each item is selected, it is removed from the population
- Sampling with replacement (independent)
  - Objects are not removed from the population as they are selected for the sample.
    - 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
  - Ensure an adequate number of samples gained for each subgroup

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

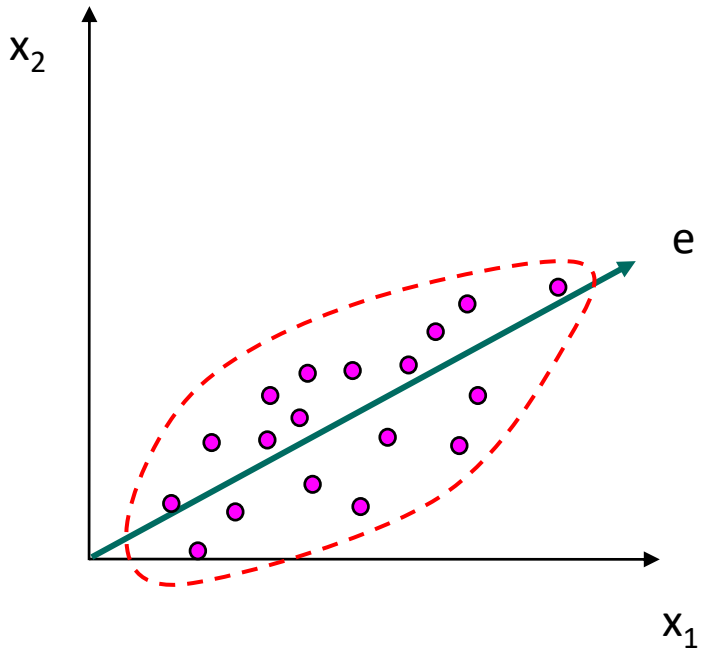
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - Help to eliminate irrelevant features or reduce noise
- Techniques
  - Principle Component Analysis (PCA)
  - Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data

  - Find the eigenvectors of the covariance matrix
  - The eigenvectors define the new space



Feature Subset Selection

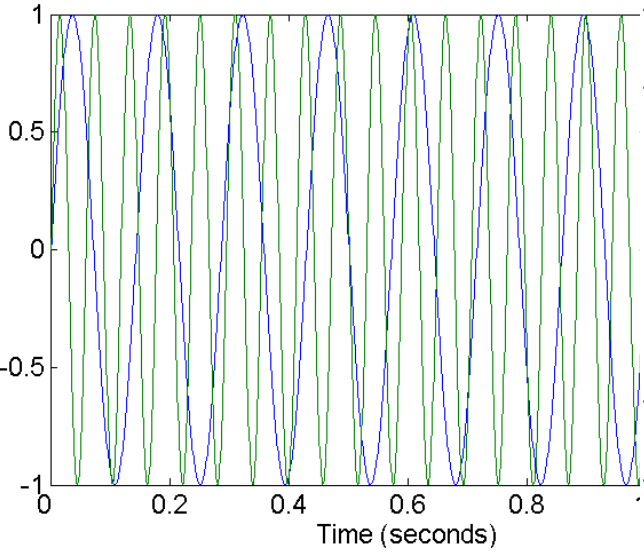
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Creation

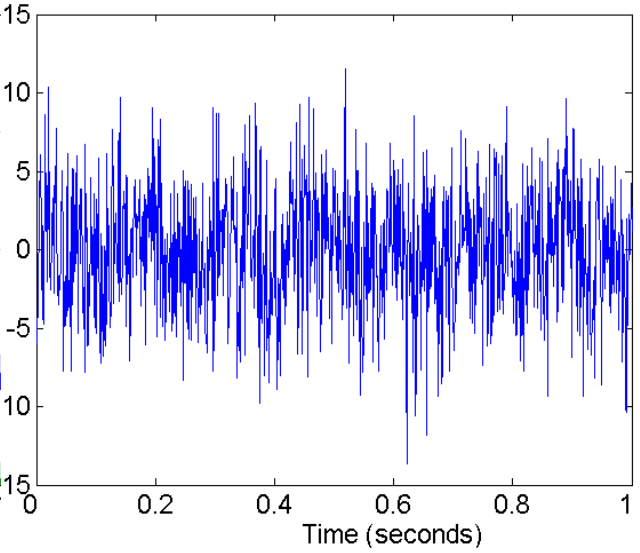
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Mapping Data to a New Space

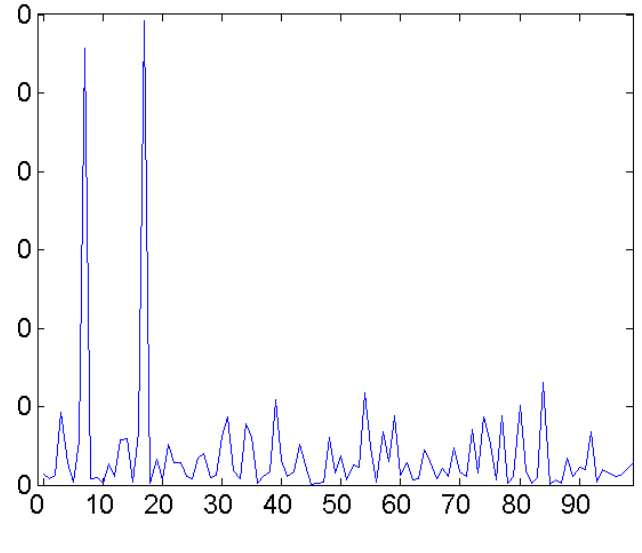
- **Fourier transform**
- **Wavelet transform**



Two Sine Waves



Two Sine Waves + Noise



Frequency

Attribute Transformation

- Maps the entire set of values of a given attribute to a new set of replacement values
  - Simple functions: x^k, log(x), e^x, |x|
  - Normalization → attributes on the similar level of measurement :
    - in range [0,1] :

$$v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

    - with 0-mean and 1-std

$$v' = \frac{v - \mathbf{mean}(v)}{\mathbf{std}(v)}$$

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Questions to answer

- What is data?
- What kinds of attributes can be used to describe objects?
- How data are different in types?
- How can we improve data quality?
- **How to measure similarities between objects?**

Similarity and Dissimilarity

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

Euclidean Distance

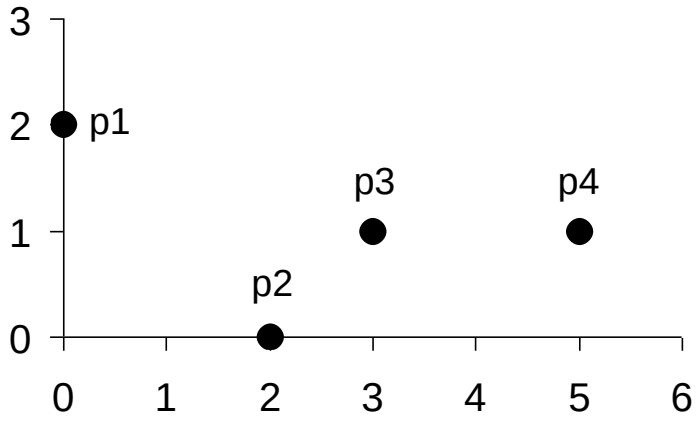
- Euclidean Distance

$$Ed(\boldsymbol{p},\boldsymbol{q}) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

Where m is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects $\boldsymbol{p}$ and $\boldsymbol{q}$.

- Normalization 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

$$dist(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{k=1}^{m} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{p}$ and $\boldsymbol{q}$.

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
  - Weighted distance

$$dist(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{k=1}^{m} w_k |p_k - q_k|^r\right)^{\frac{1}{r}}$$

Minkowski Distance: Examples

-

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

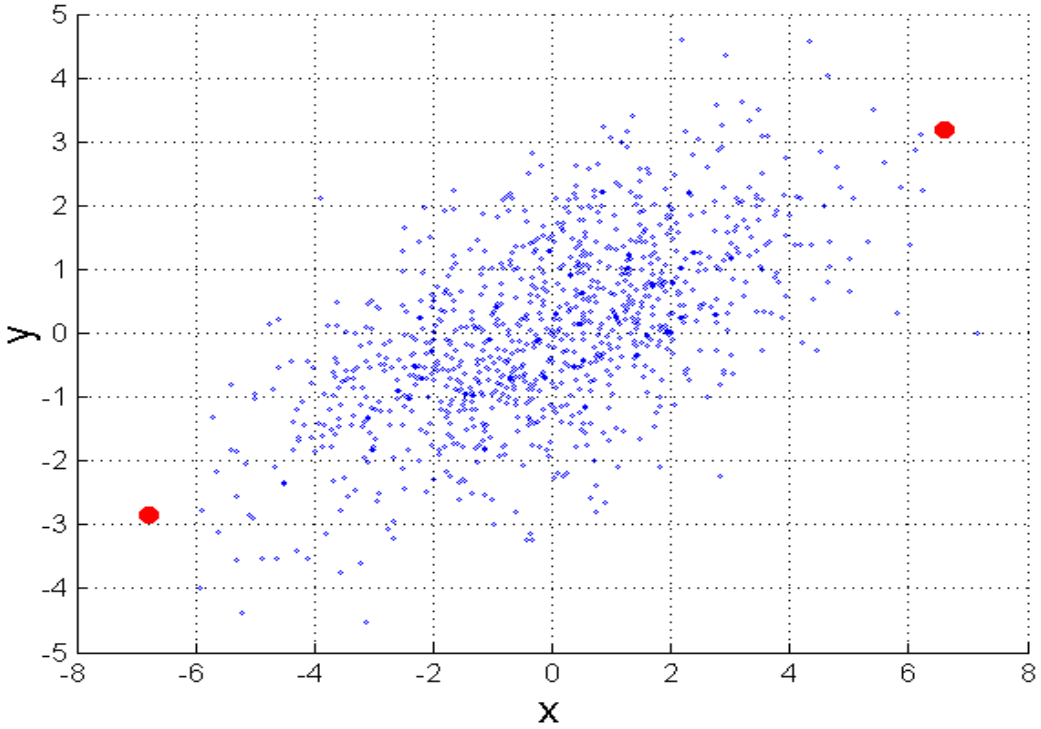
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Mahalanobis Distance

- $$mahalanobis(\boldsymbol{p},\boldsymbol{q}) = (\boldsymbol{p}-\boldsymbol{q})^T \Sigma^{-1} (\boldsymbol{p}-\boldsymbol{q})$$

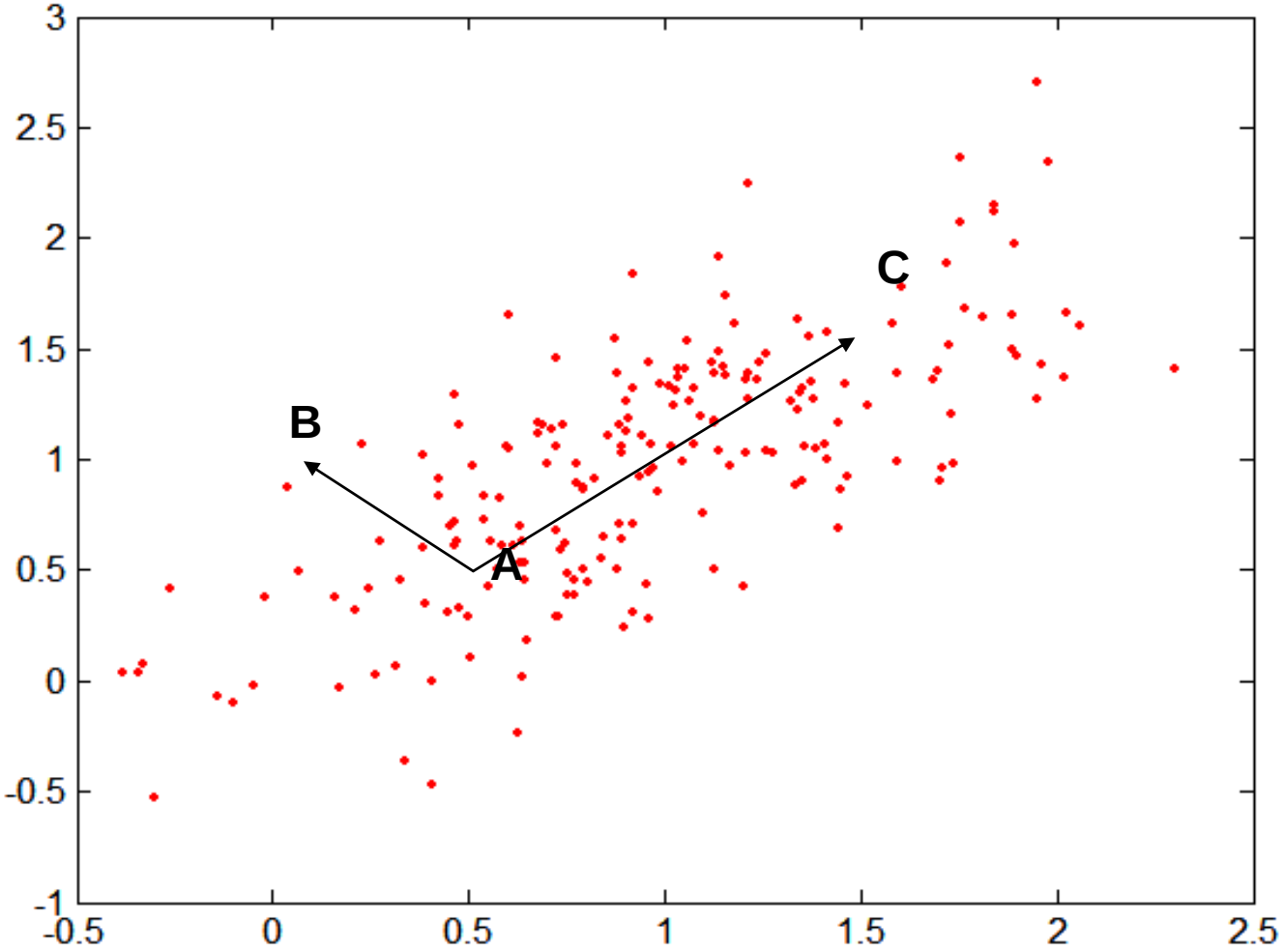


∑ is the covariance matrix of the input data X=[...$\mathbf{x_i}$...$\boldsymbol{p}$…$\mathbf{x_j}$...$\boldsymbol{q}$ ...]

$$\Sigma_{i,j} = \frac{1}{m-1} \sum_{k=1}^{m} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Techniques in Data Exploration

- **Summary statistics**
- Visualization

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

Measures of Location: Mean and Median

- 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.
- The median or a trimmed mean is thus also commonly used.

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

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.

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

- sensitive to outliers, other measures:

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

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

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

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in data set
  - e.g., data: a representative population of people,
    attribute: 'gender'
    frequency of '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 (nominal attributes)

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 .
  - ➔ Sort *N* values of attribute *x* in decreasing order, x_p is the N*(1-p/100)-th one.
  - ➔ p = 50 ---> x_p is close to the median value

Techniques in Data Exploration

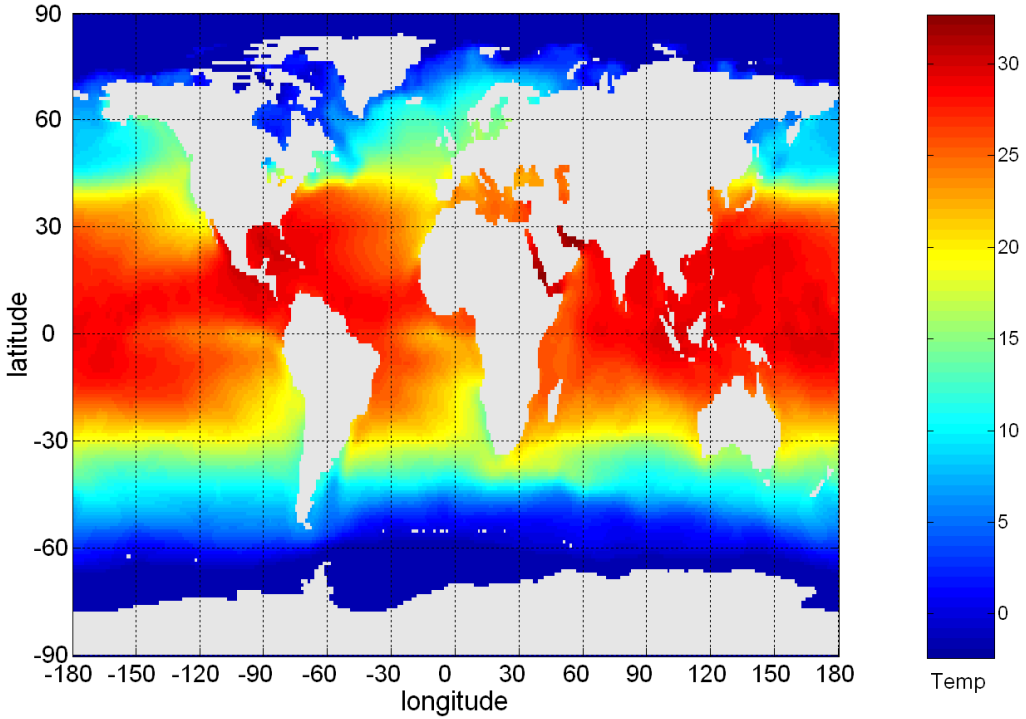
- Summary statistics
- **Visualization**

Visualization

- Definition: 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

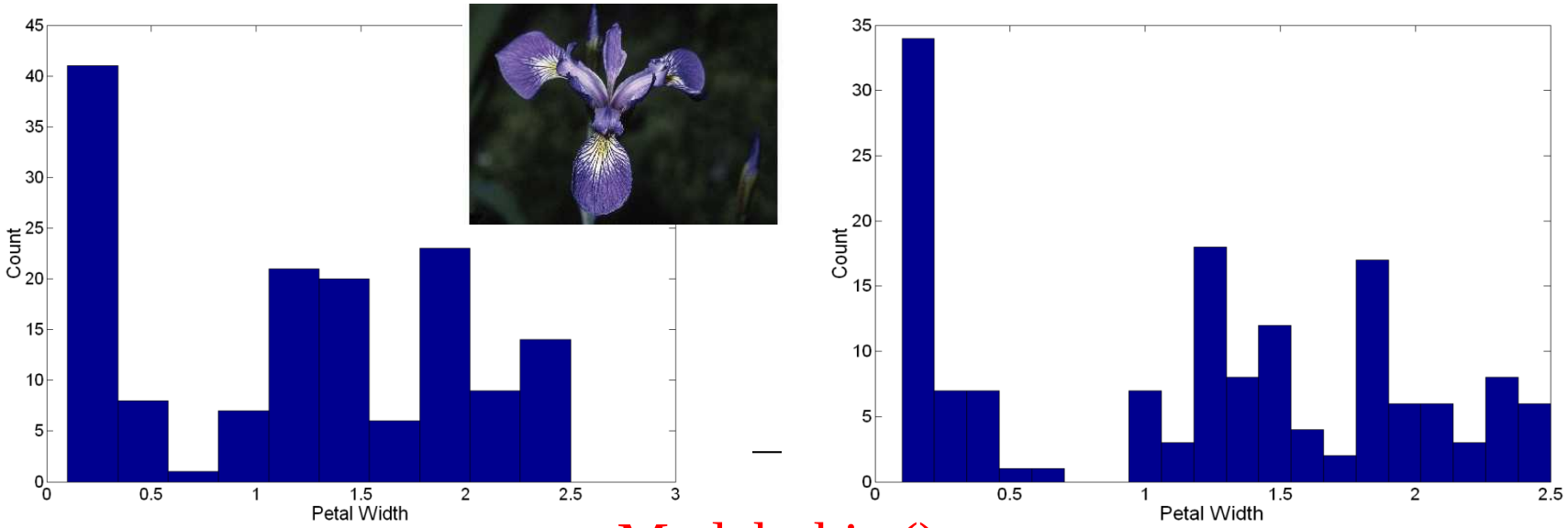
Example: Sea Surface Temperature

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



Visualization Techniques: Histograms

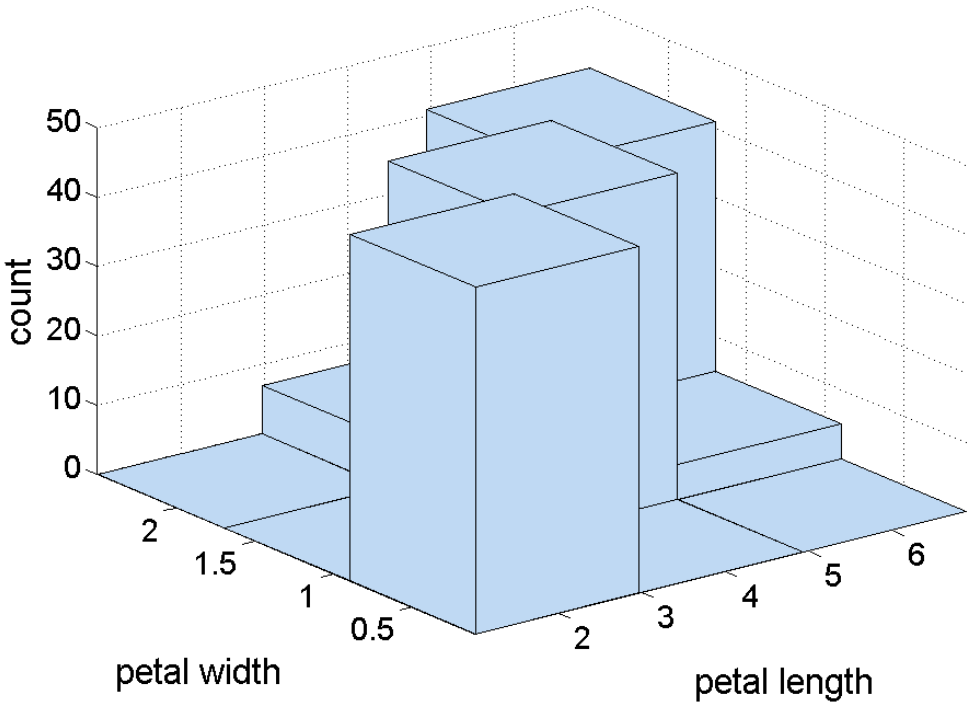
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins, show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively) of Iris Plant data set (UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html)



Matlab: hist()

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 (Invented by John Tukey)
  - Another way of displaying and comparing the distribution of data