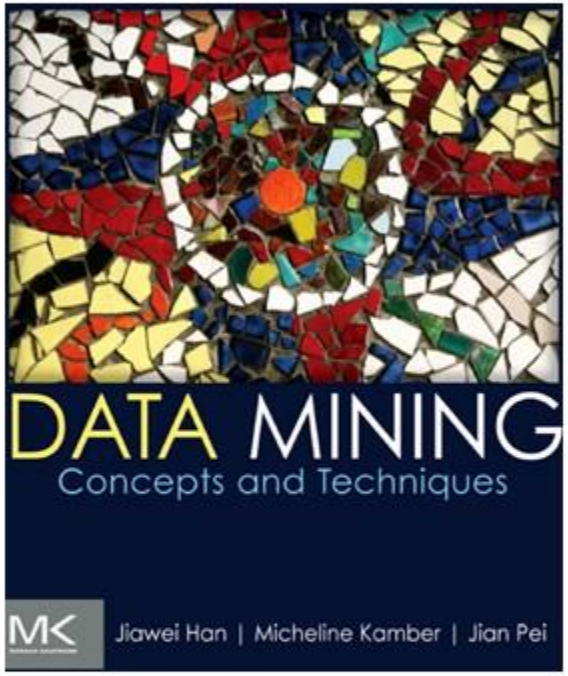
Data Mining Algorithms

Instructor:

Dr. Mohamed H. Farrag



Textbook (s)

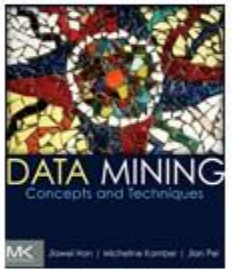
Main textbook,

- ***Data Mining Concepts and Techniques*** **(3rd ed.)**

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University



- ***Introduction to Data Mining,*** **2nd Edition**

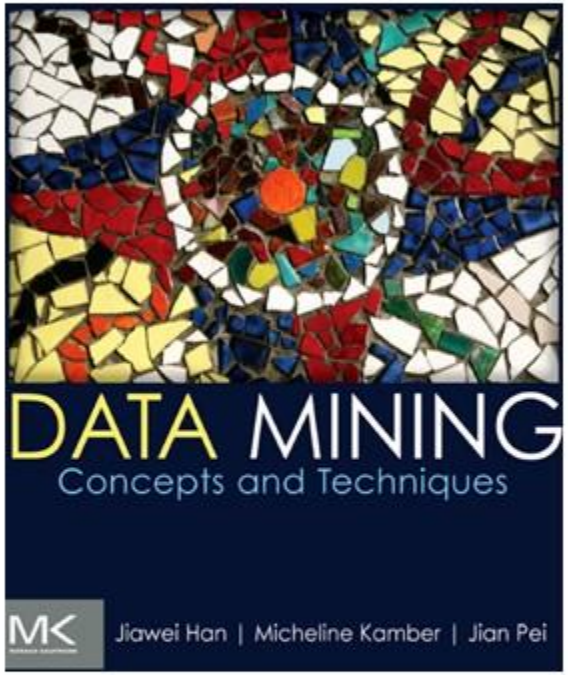
Tan, Steinbach,

Karpatne, Kumar

Modified for Introduction to Data Mining by Dr. Mohamed H. Farrag

Chapter 2

- Getting to Know Your Data



Chapter 2 LEARNING OBJECTIVES

Getting to Know Your Data

- Attributes and Objects
- Types of Data
- Data Quality
- Similarity and Distance
- Data Preprocessing

Getting to Know Your Data

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, dimension, or feature
- **A collection of attributes describe** an ***object***
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Attributes

- **Nominal** Examples: ID numbers, eye color, zip codes, *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}, marital status
- **Ordinal** Examples: grades, height {tall, medium, short}
  - Values have a **meaningful order** (ranking)
  - Size = {small, medium, large}, grades, army rankings
- **Interval** Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Measured on a scale of equal-sized units
  - **Values have order**
    - E.g., temperature in C˚or F˚, calendar dates
  - **No true zero-point**
- **Ratio** Examples: temperature in Kelvin, length, time, counts and monetary quantities
  - **Inherent zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /
  - **Nominal attribute: distinctness**
  - **Ordinal attribute: distinctness & order**
  - **Interval attribute: distinctness, order & meaningful differences**
  - **Ratio attribute: all 4 properties/operations**

Properties of Attribute Values

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Properties of Attribute Values

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

This categorization of attributes is due to S. S. Stevens

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably **infinite set of values**
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: **binary attributes are a special case of discrete attributes**
- Continuous Attribute
  - Has **real numbers** as attribute values
  - Examples: **temperature**, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

More Complicated Examples

- ID numbers
  - Nominal, ordinal, or interval?
- Number of cylinders in an automobile engine
  - Nominal, ordinal, or ratio?
- Biased Scale
  - Interval or Ratio

Types of data sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data
- **Graph**
  - World Wide Web
  - Molecular Structures
- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data:
  - Video data:

Important Characteristics of Data

- **Dimensionality**
  - number of attributes for the objects in the data set
  - High dimensional data brings a number of challenges
  - **Curse of dimensionality** (the difficulties associated with analyzing high –dimension data)
- **Sparsity**
  - most attributes of an object have values of 0
  - Fewer than 1% the entries non zero
- **Resolution**
  - it is the frequently possible to obtain data at different levels of resolution.
- **Size**
  - Type of analysis may depend on size of data

Record Data

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

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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**

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

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

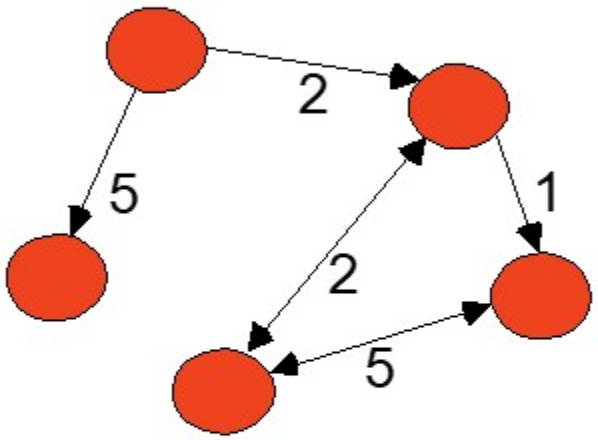
Transaction Data

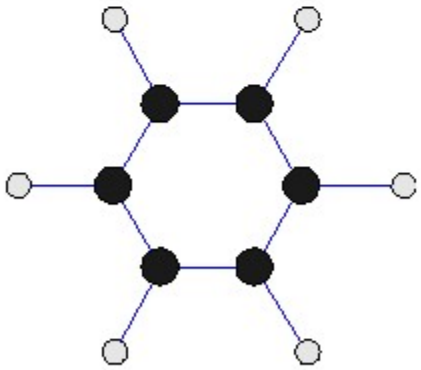
- A **special type of record data**, where
  - Each record (**transaction**) involves a **set of items**.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the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





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

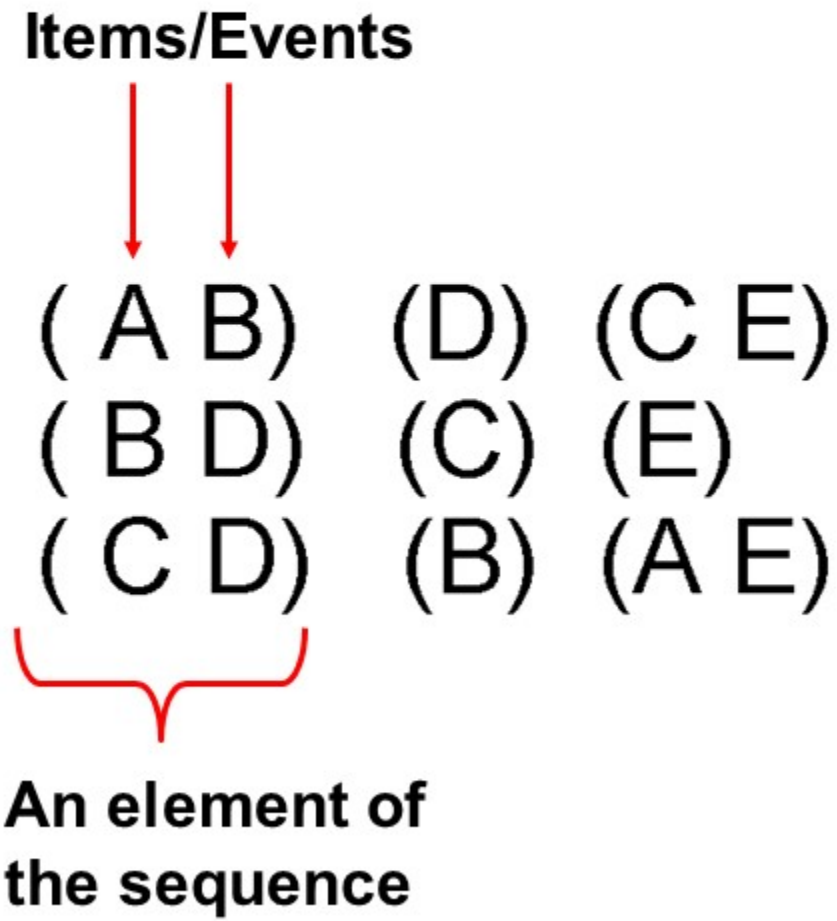
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions



Ordered Data

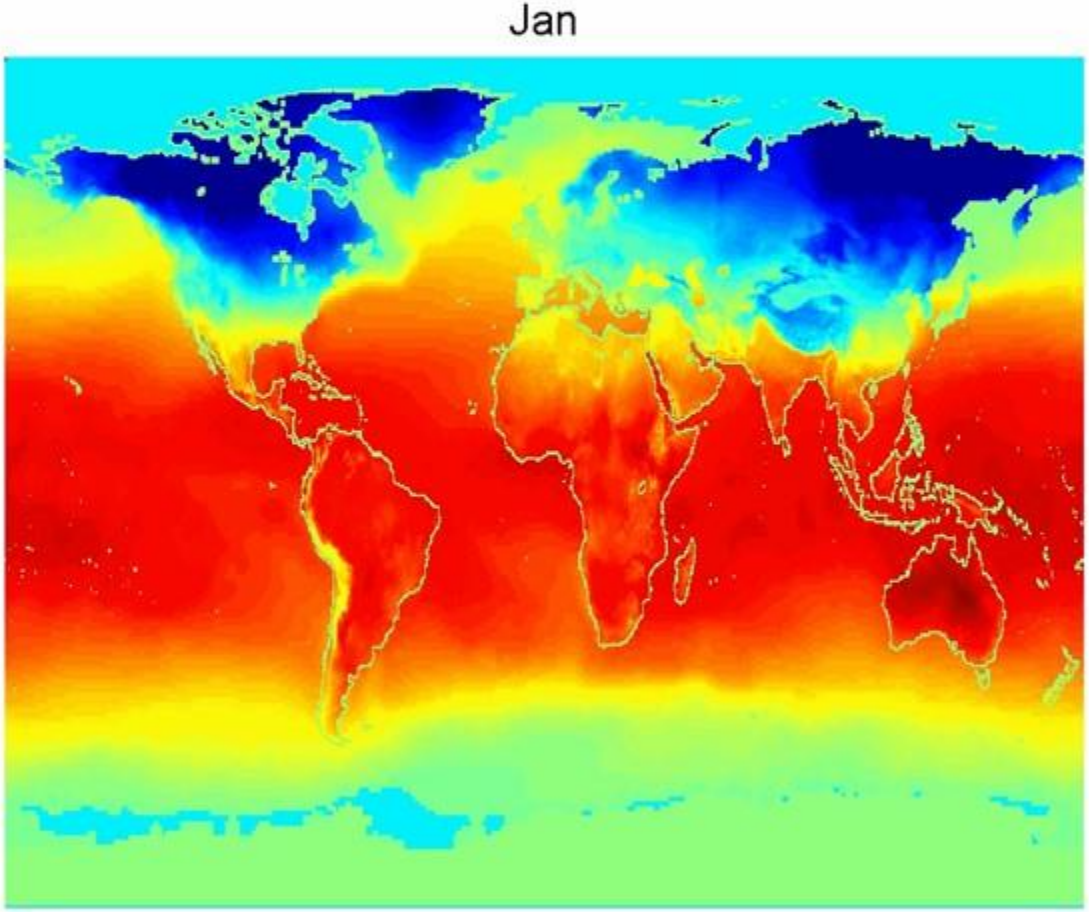
- Genomic sequence data

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

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

- **Poor data quality negatively affects many data processing efforts**
  - Poor data quality costs the typical company at least **ten percent (10%) of revenue**; twenty percent **(20%)** is probably a better estimate."

    Thomas C. Redman, DM Review, August 2004

- **Data mining example**: a **classification model** for detecting people who are loan risks is **built using poor data**
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

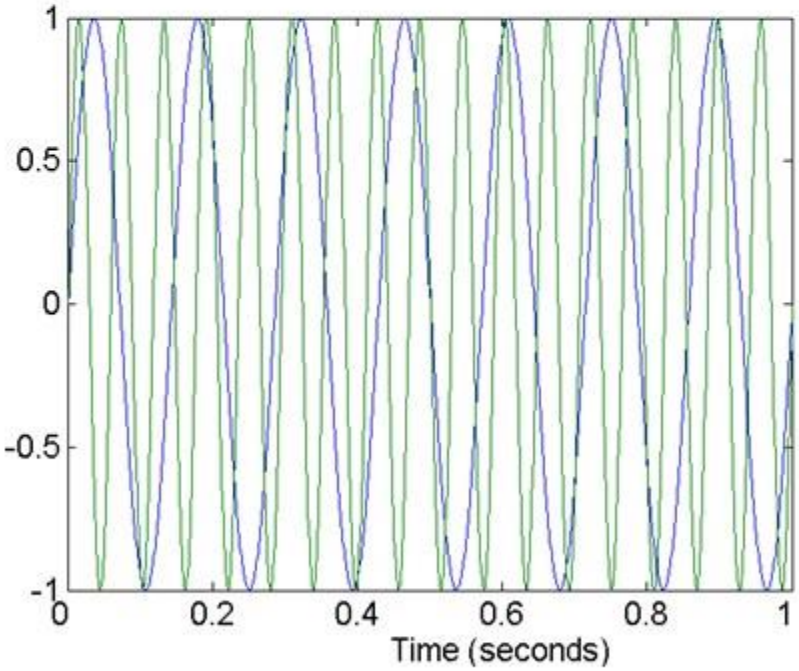
Data Quality

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

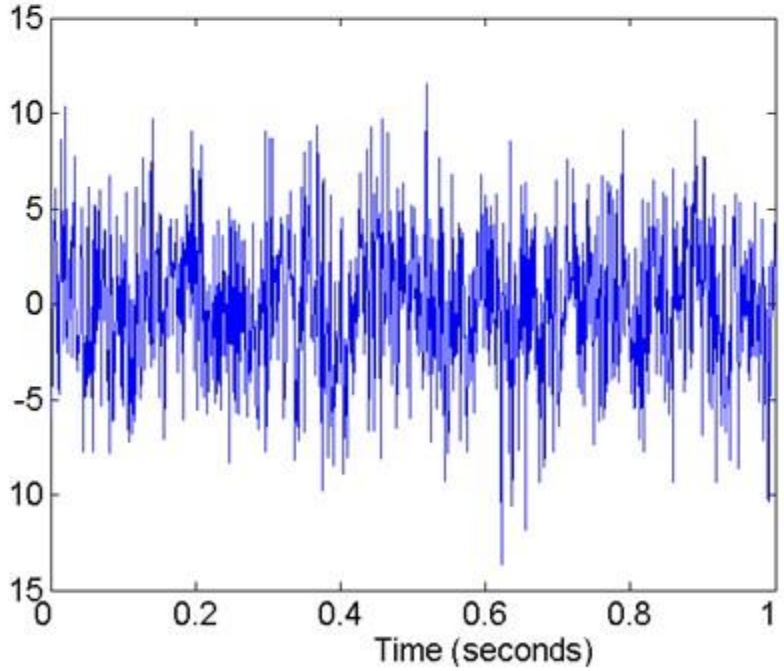
- Examples of data quality problems:
  - **Noise and outliers**
  - **Missing values**
  - **Duplicate data**
  - **Wrong data**

Noise

- **For objects**, noise is an **extraneous object**
- **For attributes**, **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



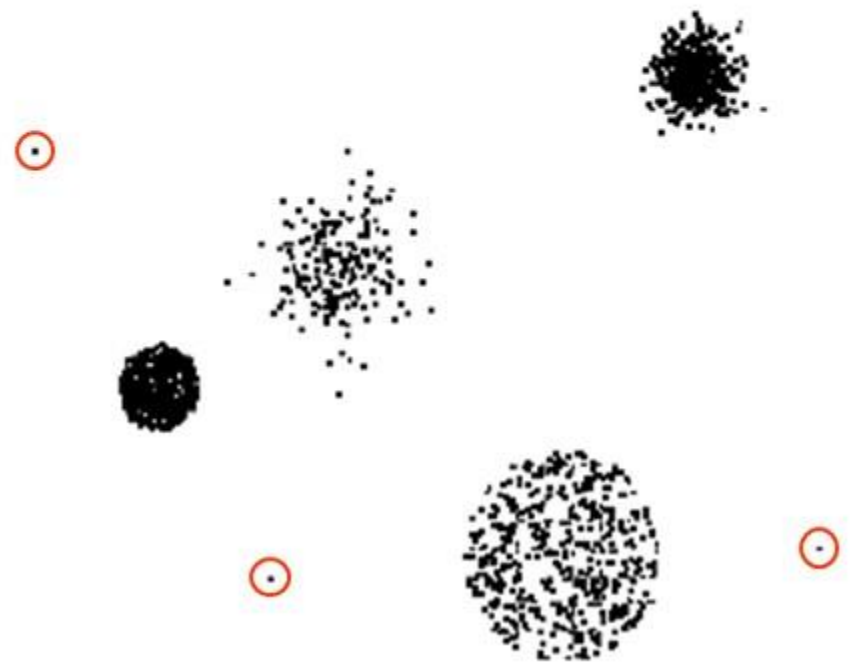
Two Sine Waves



Two Sine Waves + Noise

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection
- Causes?



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 or variables**
  - **Estimate missing values**
    - Example: time series of temperature
    - Example: census results
  - **Ignore** the **missing value** during analysis

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**
- Examples:
  - **Same person with multiple email addresses**
- Data cleaning
  - **Process of dealing with duplicate data issues**

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009