

# Data Mining: Data

---

## Lecture Notes for Chapter 2

Introduction to Data Mining , 2<sup>nd</sup> Edition

by

Tan, Steinbach, Kumar

# Outline

---

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

# What is Data?

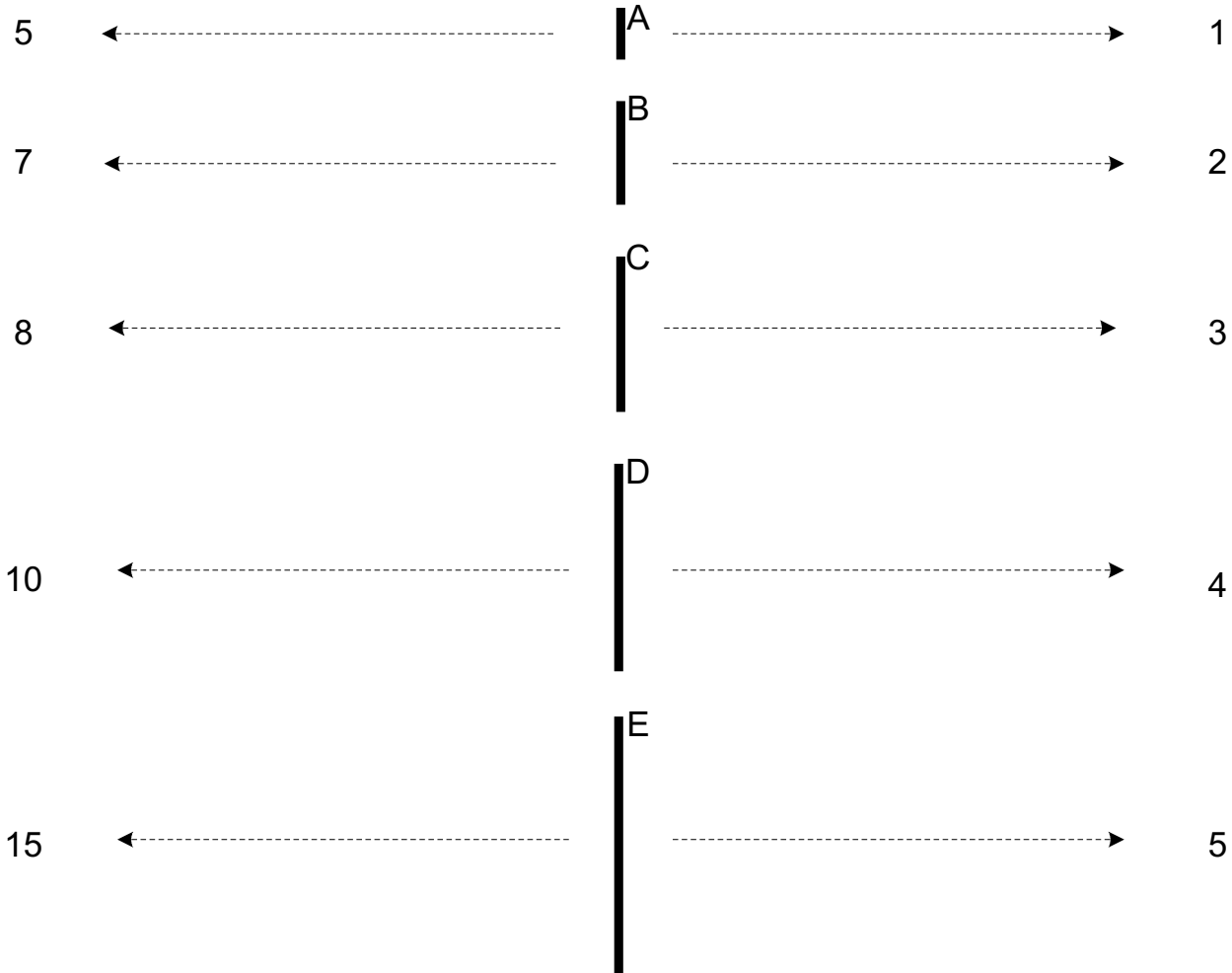
- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe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

**Attributes**

<i>Tid</i>	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

# Measurement of Length

- The way you measure an attribute may not match the attribute's properties.



**This scale preserves only the ordering property of length.**

**This scale preserves the ordering and additivity properties of length.**

# Types of Attributes

---

- There are different types of attributes
  - **Nominal**
    - ◆ Examples: ID numbers, eye color, zip codes
  - **Ordinal**
    - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - **Interval**
    - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - **Ratio**
    - ◆ Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

# 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

# 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.

# Asymmetric Attributes

---

- Only presence (a non-zero attribute value) is regarded as important
  - ◆ Words present in documents
  - ◆ Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

*“I see our purchases are very similar since we didn’t buy most of the same things.”*



# Important Characteristics of Data

---

- Dimensionality (number of attributes)
  - ◆ High dimensional data brings a number of challenges
- Sparsity
  - ◆ Only presence counts
- Resolution
  - ◆ Patterns depend on the scale
- Size
  - ◆ Type of analysis may depend on size of data

# 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

# Record Data

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

<i>Tid</i>	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 a 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

# 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

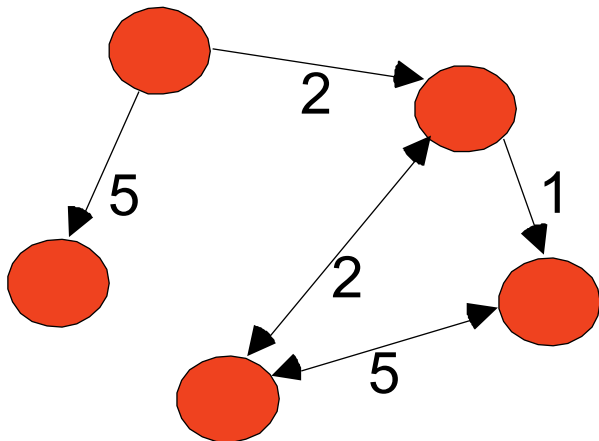
# Transaction Data

- A special type of data, where
  - Each 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.
  - Can represent transaction data as record data

<i>TID</i>	<i>Items</i>
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



## 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

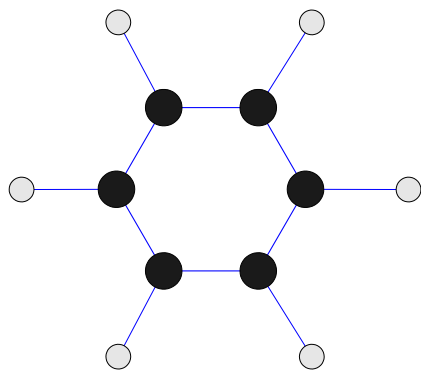
Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

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.



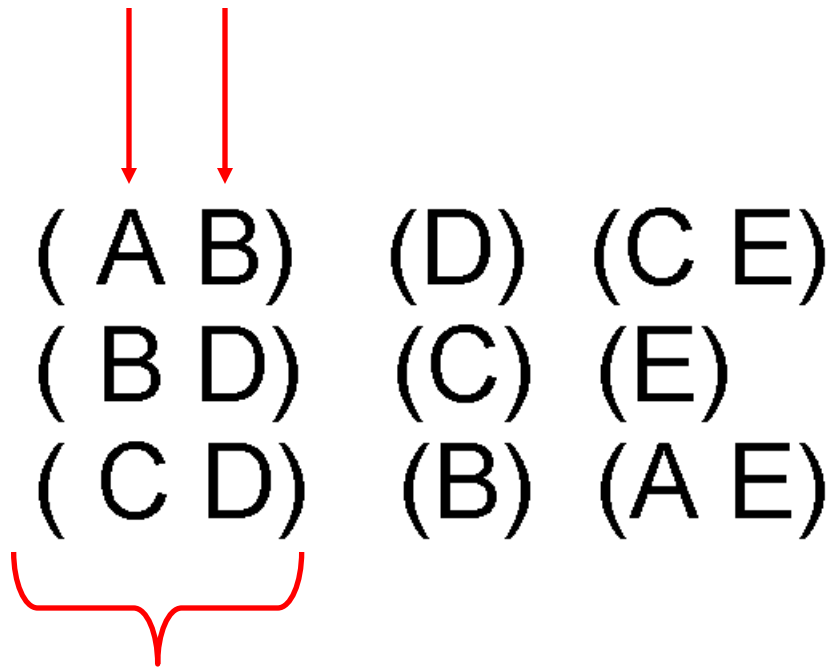
Benzene Molecule: C6H6

# Ordered Data

---

- Sequences of transactions

**Items/Events**



**An element of  
the sequence**



# Ordered Data

---

- Genomic sequence data

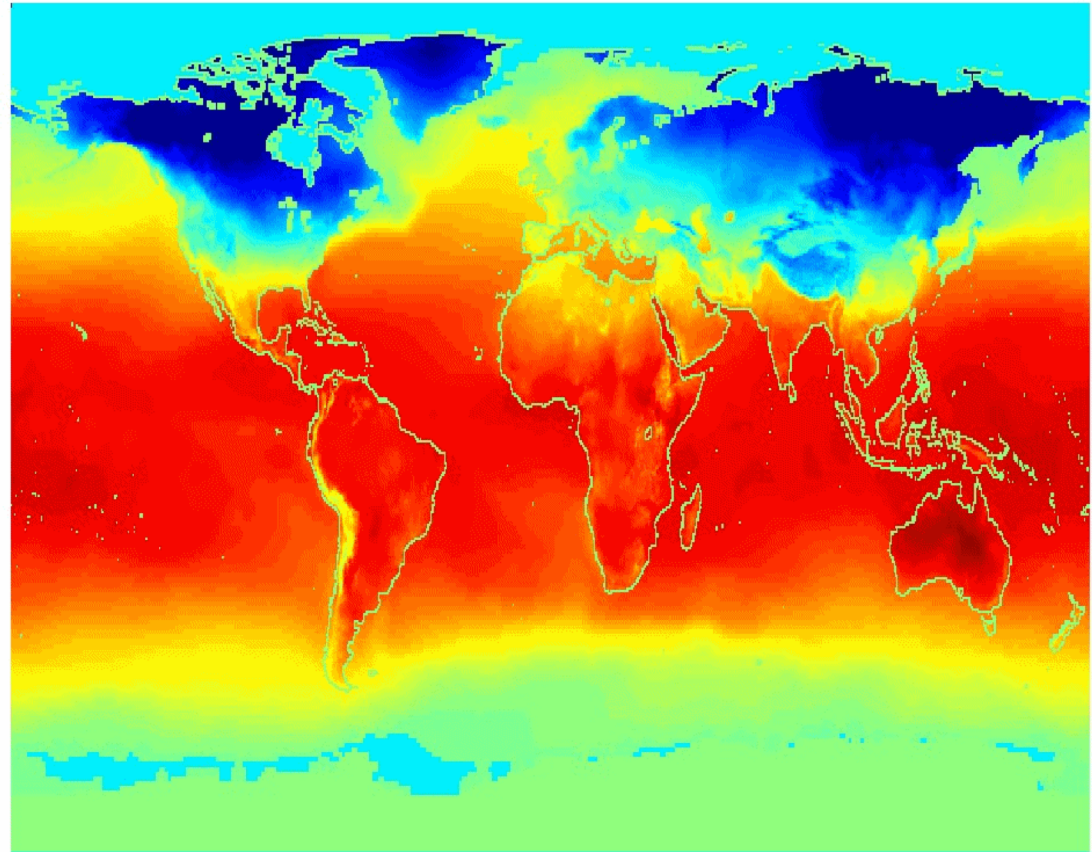
**GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG**

# Ordered Data

- Spatio-Temporal Data

**Average Monthly  
Temperature of  
land and ocean**

Jan



# Data Quality

---

- Poor data quality negatively affects many data processing efforts
- 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

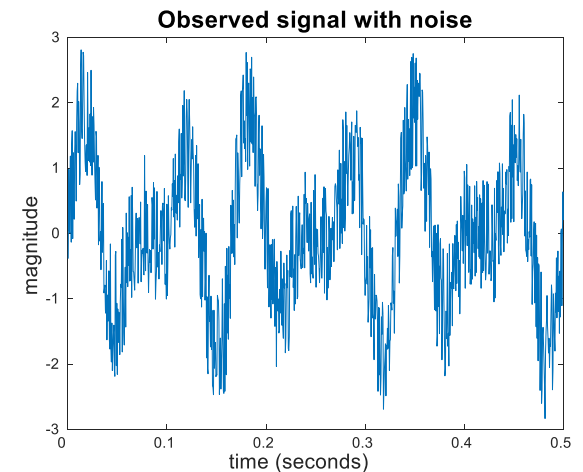
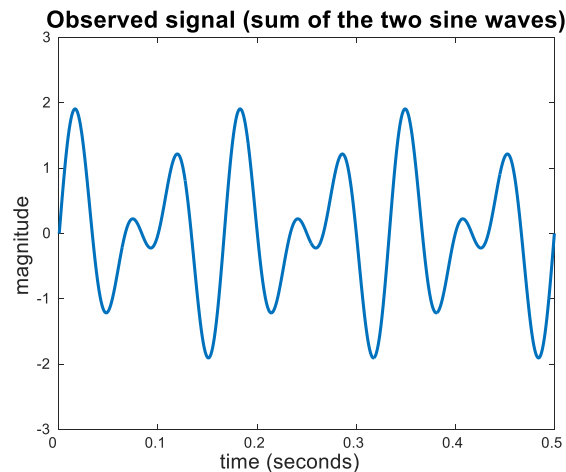
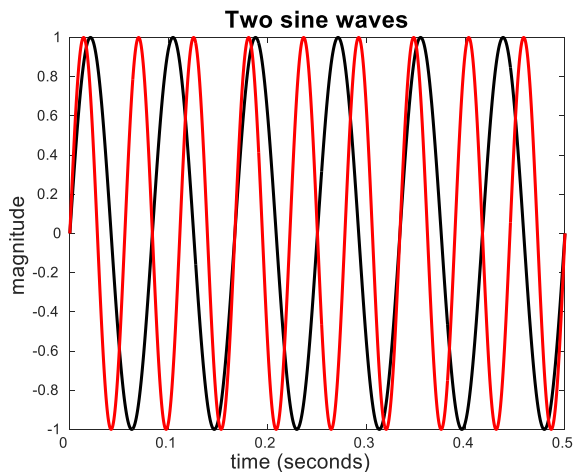
# Data Quality ...

---

- 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
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

# Noise

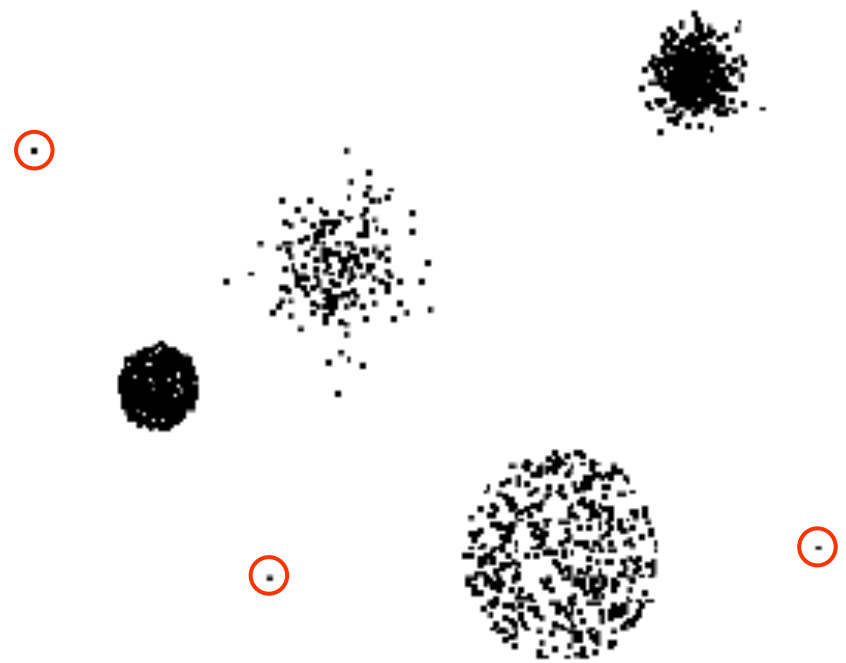
- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
  - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - ◆ The magnitude and shape of the original signal is distorted



# 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
- When should duplicate data not be removed?