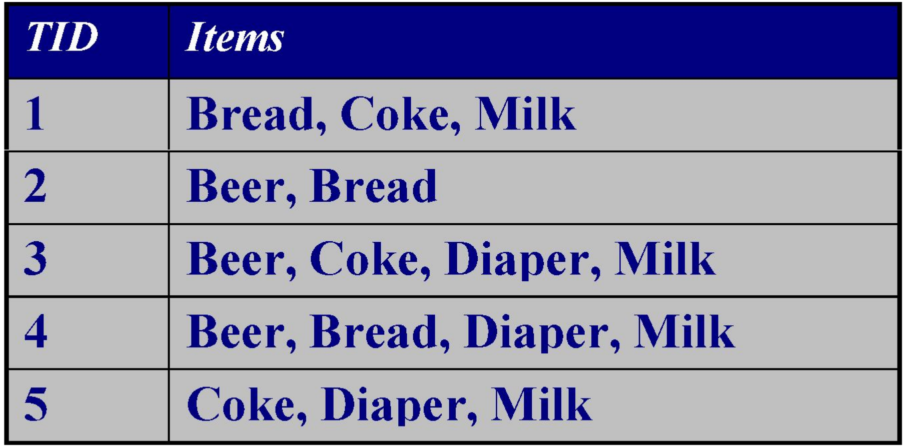
Association Rule Discovery : Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.



TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery : Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Coke, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Coke in the antecedent => Can be used to see which products would be affected if the store discontinues selling coke.
  - Coke in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Coke to promote sale of Potato chips!

Association Rule Discovery : Application 2

- Supermarket shelf management.
  - Goal: To identify items that are purchased together by many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery : Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery : Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.

$$(A \quad B) \quad (C) \longrightarrow (D \quad E)$$

- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

Sequential Pattern Discovery : Example

- In point-of-sale transaction sequences,

    - Computer Bookstore:
        - (Intro_To_Visual_C) (C++_Primer) --> (Perl_for_dummies,Tcl_Tk)
    - Athletic Apparel Store:
        - (Shoes) (Racket, Racketball) --> (Sports_Jacket)

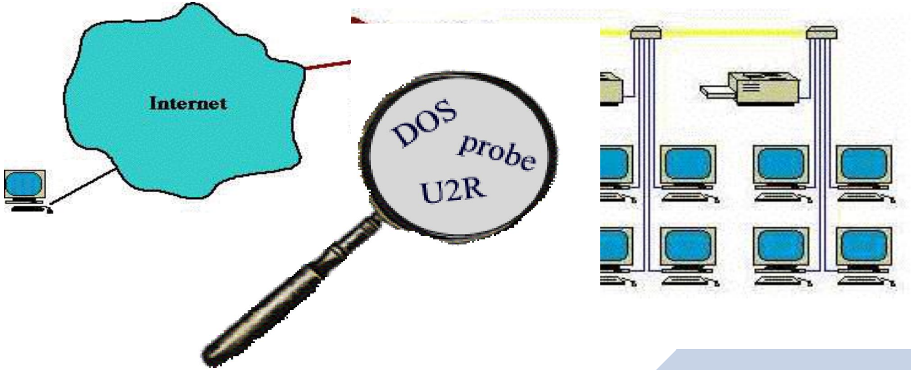
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Time series prediction of stock market indices.
  - Income prediction on basis of qualifications and other characteristics of individuals

Deviation / Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
    - Credit Card Fraud Detection
    - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Open Source Data Mining Tools

- Python
- R
- Weka
- Knime
- Rapidminer
- Matlab
- Tableau

Contribution of Data Mining

- Less expenditures
  - Automated systems instead of manual ones
  - Selection of customers to mail new promotions of the company

- Effective decision making
  - Careful expansion of the business
  - Product selection
  - Pricing

Contribution of Data Mining

- Increased sales
  - Shelf management to increase the sale of certain items
  - What types of products can be sold together?
  - How does one retain profitable customers?

Data Mining Real World Success Stories

- **Bank of America identified savings of $4.8 million in 2 years by using a credit risk management system, i.e., examination of only borderline applicants.**

- **BBC's data mining based program scheduler determines the timing to show programs as good as the best planner but at much less cost.**

Data Mining Real World Success Stories

- Bell Atlantic developed telephone technician dispatch system. They must decide what type of technician to dispatch to resolve the reported complain.

- Bell Atlantic save more than 10 million dollars per year by using data mining rule based system because they make fewer erroneous decisions.

Data Mining Real World Success Stories

- Safeway (UK)'s data mining system found that the top - spending 25% customers often purchase a particular cheese product ranked below 200 in sales.

- Normally, without the data - mining results, the product would have been discontinued and would disappoint the best customers.

- Safeway continues to order this cheese, although it is ranked low in sales.

Lecture 4

Introduction To Data Science

Dr. Faisal Kamiran

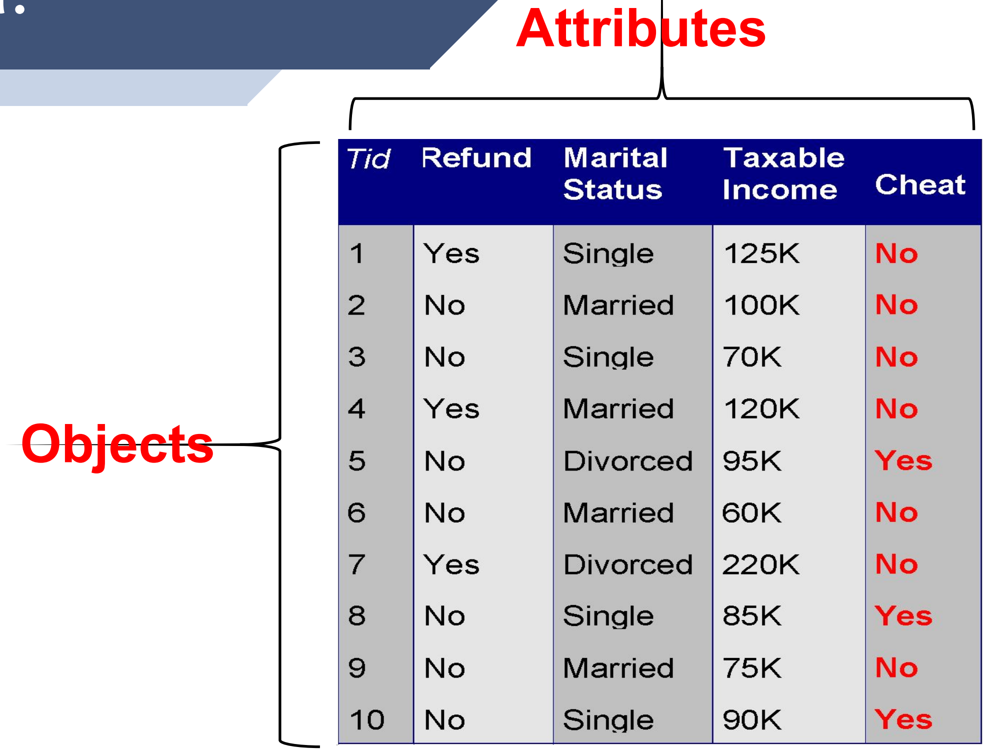
What is today’s agenda?

Today we are going to learn following things :

- Data Understanding
- 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, 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

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - ◆ Example: height can be measured in feet or meters

  - Different attributes can be mapped to the same set of values
    - ◆ Example: Attribute values for ID and age are integers
    - ◆ But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

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
  - 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 with limited precision.
  - Continuous attributes are typically represented as floating-point variables.

Type of Attributes

- Categorical (Qualitative)
  - 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 in {tall, medium, short}
- Numeric (Quantitative)
  - Interval
    - Examples: temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Type of Attributes

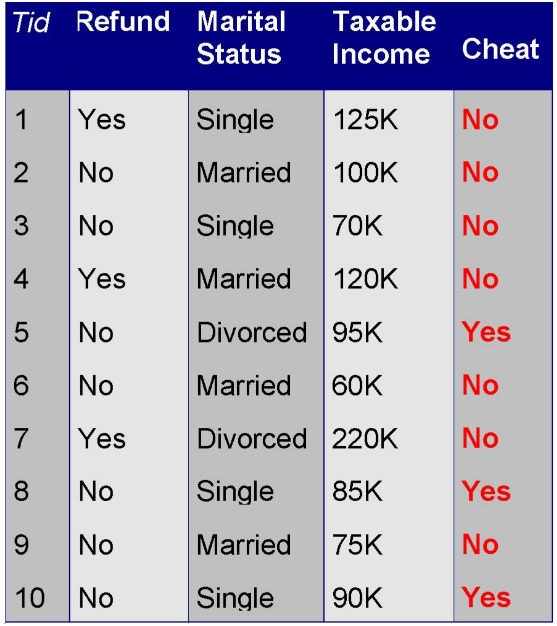
OK to compute....	Nominal	Ordinal	Interval	Ratio
Mode, Entropy	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

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

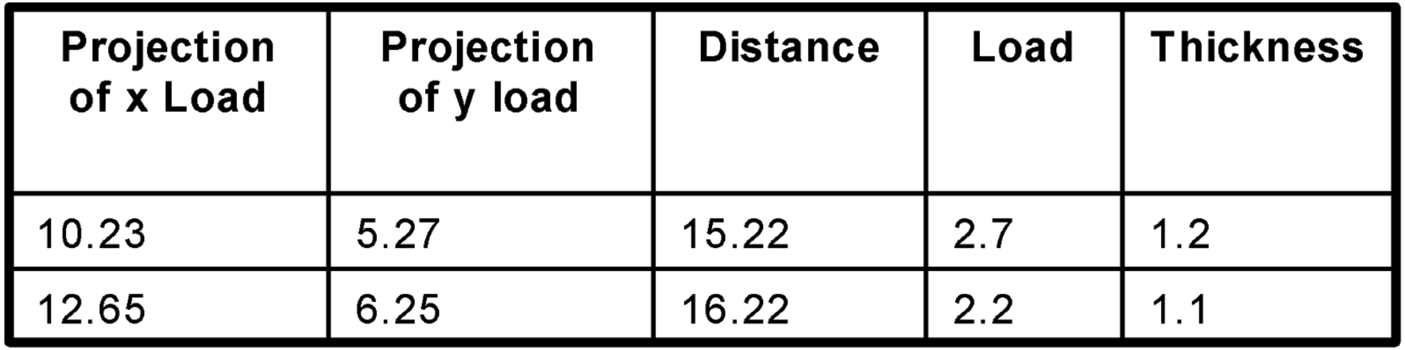


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

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, 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

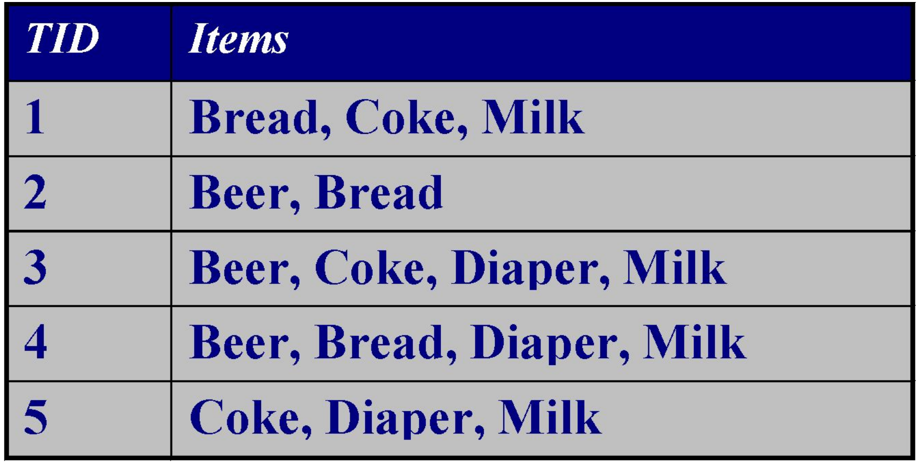
Document Data

- 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

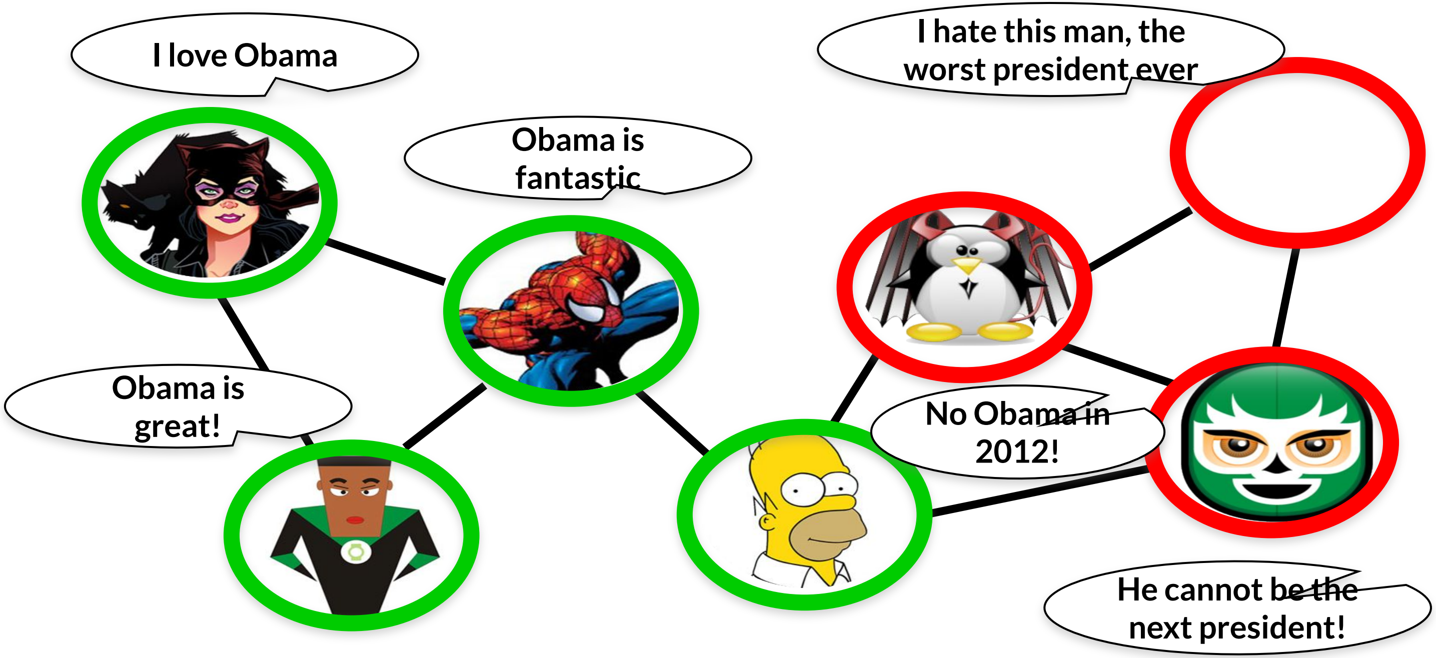
Transaction Data

- 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 : “Love Obama”

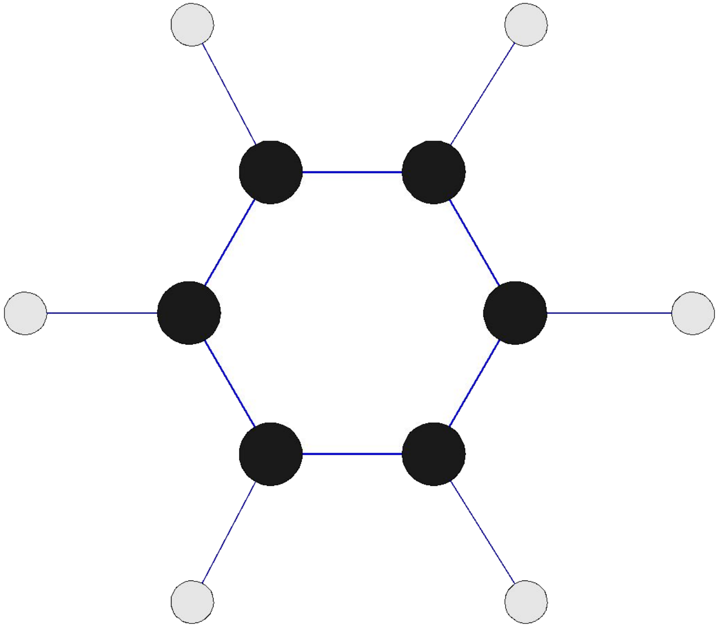


I love Obama
Obama is fantastic
I hate this man, the worst president ever
Obama is great!
No Obama in 2012!
He cannot be the next president!

○ Positive ○ Negative

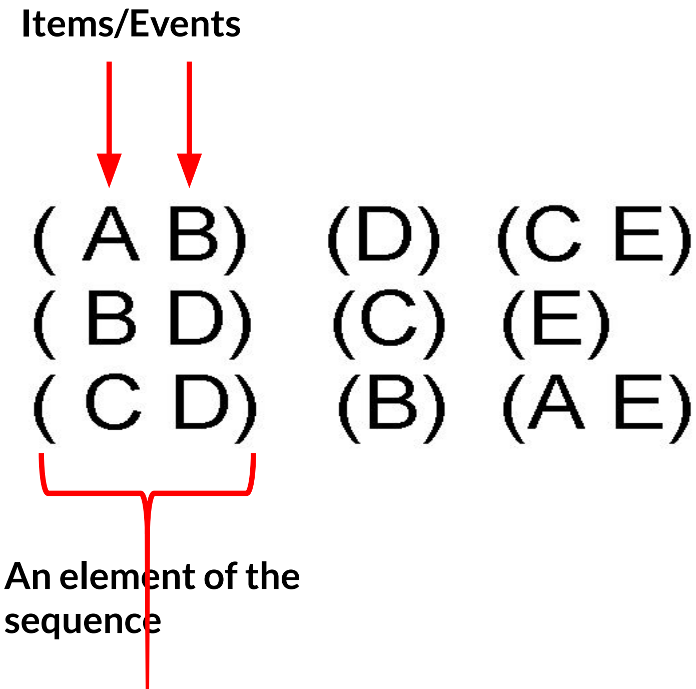
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- Sequences of transactions



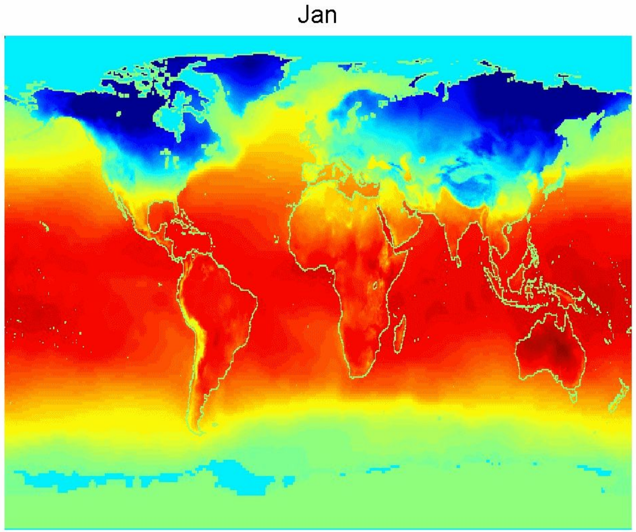
Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

- Spatio-Temporal Data



Average Monthly Temperature of Land and Ocean

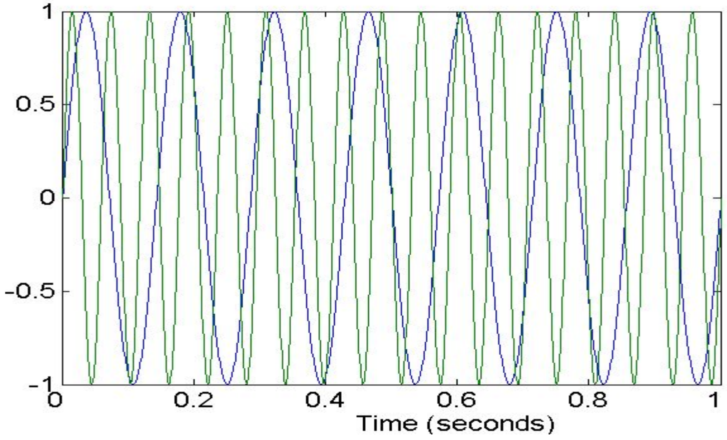
Data Quality

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

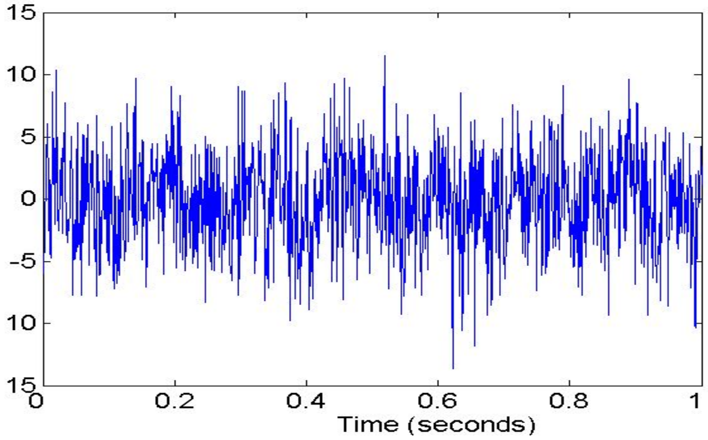
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

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



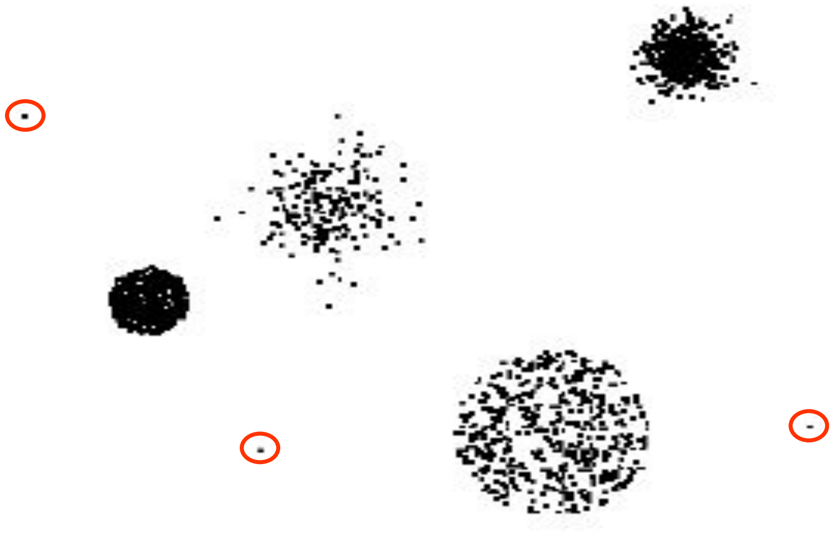
Two Sine Waves



Two Sine Waves + Noise

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)