COSI 104a Introduction to machine learning

Chapter 3 – Data

Instructor: Dr. Hongfu Liu
Email: hongfuliu@brandeis.edu

Main source: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Outline

- **Attributes and Objects**
- **Types of Data**
- **Data Quality**

What is Data?

- **Collection of data objects and their attributes (features)**
- **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

Attribute (Feature) Values

- **Attribute values are numbers or symbols assigned to an attribute for a particular object**

- **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

Types of Attributes

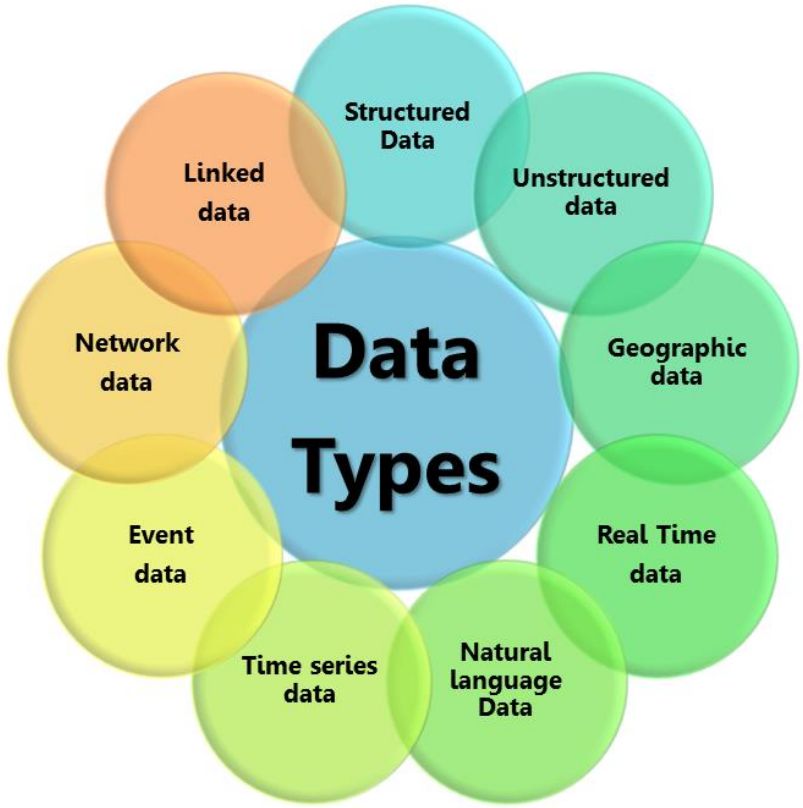
- **There are different types of attributes**
  - **Categorical (Discrete)**
    - Nominal: ID numbers, eye color, zip codes
    - Ordinal: rankings (e.g., taste of potato chips, grades)
  - **Numerical (Continous)**
    - Interval: calendar dates, temperatures in Celsius or Fahrenheit.
    - Ratio: temperature in Kelvin, length, time, counts

	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

- **Difference Between Ratio and Interval**
  - Is it physically meaningful to say that a temperature of 10 ° is twice that of 5°
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?

Types of Data

- **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**
  - 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 n by m matrix, where there are n rows, one for each object, and m columns, one for each attribute
  - What about document and transaction data?

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

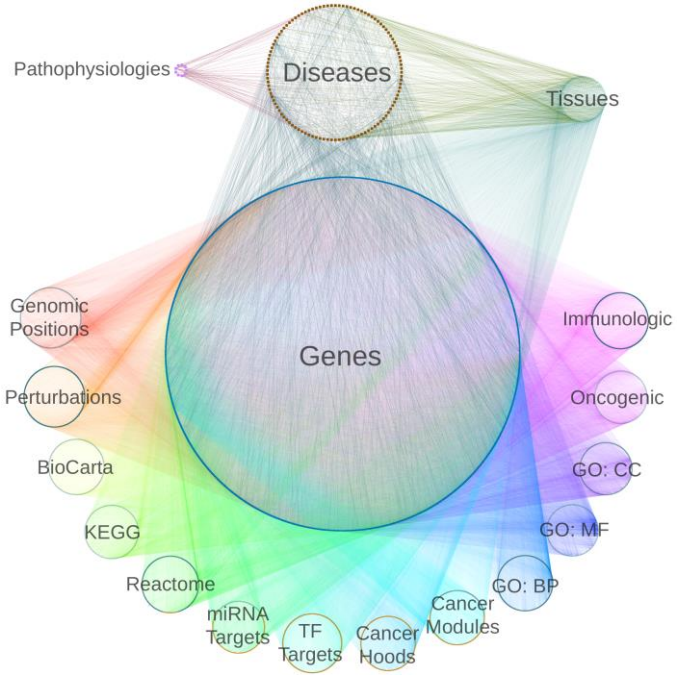
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

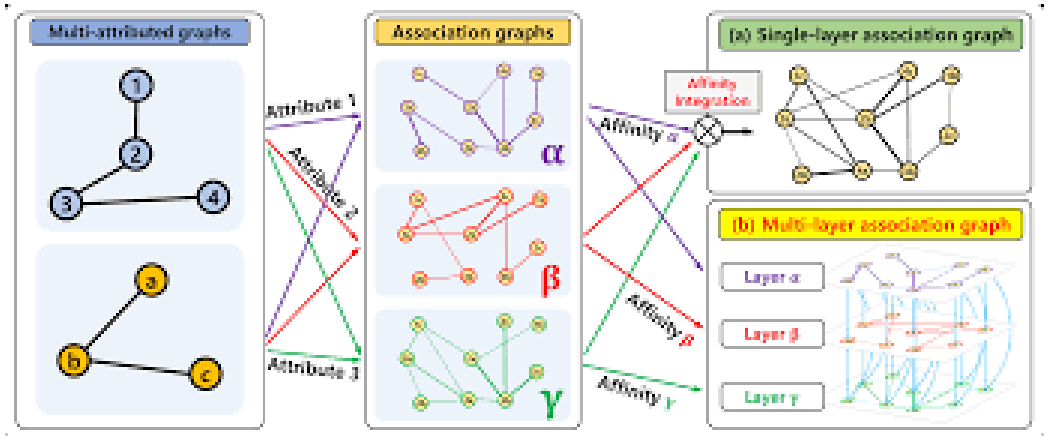
Graph Data

- **Graph data consists of nodes and edges**

Homogeneous graph



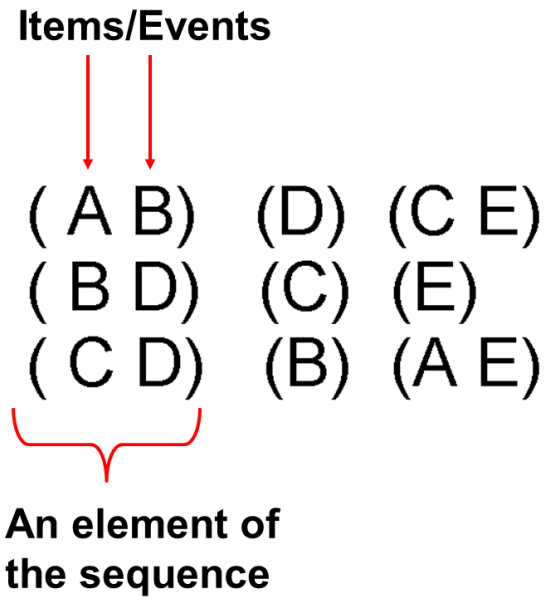
Heterogenous graph



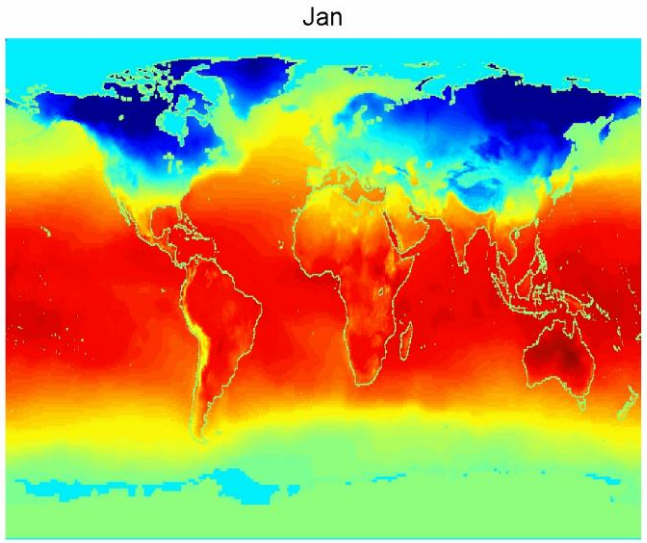
Multi-attributes graph

Ordered Data

- **Sequences of transactions, Genomic sequence data, Spatio-Temporal Data**



GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG



Data Quality

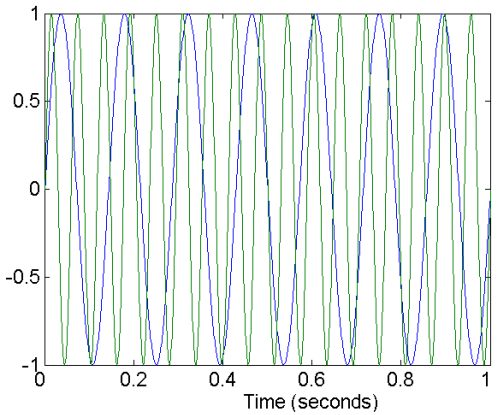
- **Problems**
  - What kinds of data quality problems?
  - How can we detect problems with the data?
  - What can we do about these problems?

- **Examples**
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

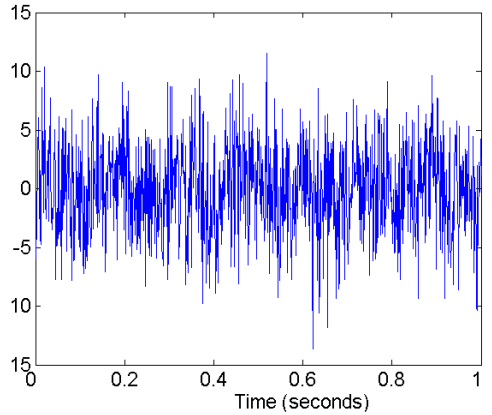
- **Solutions**
  - **Data level**
  - **Algorithm level**

Data Quality

- **Noise**
  - For objects, noise is an extraneous object (outlier)
  - 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

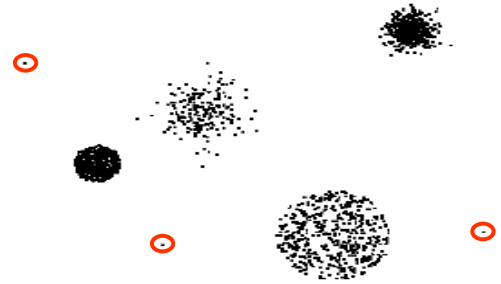


Two Sine Waves



Two Sine Waves + Noise

- **Outlier are data objects with characteristics that are considerably different than most of the other data objects in the data set**
  - Outliers are noise that interferes with data analysis
  - Outliers are the goal of our analysis (Credit card fraud, Intrusion detection)



- **Why 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)

- **Handle missing values**
  - Eliminate data objects or variables
  - Estimate missing values
    - Statistics: average value, mode value
    - Model: nearest neighbor, prediction, centroid
  - Ignore the missing value during analysis


?
Q
&
&
A
!