# Data Preprocessing

## UE 141 Spring 2013

Jing Gao
SUNY Buffalo

# Outline

- **Data**
- **Data Preprocessing**
  - Improve data quality
  - Prepare data for analysis
- **Exploring Data**
  - Statistics
  - Visualization

# Document Data

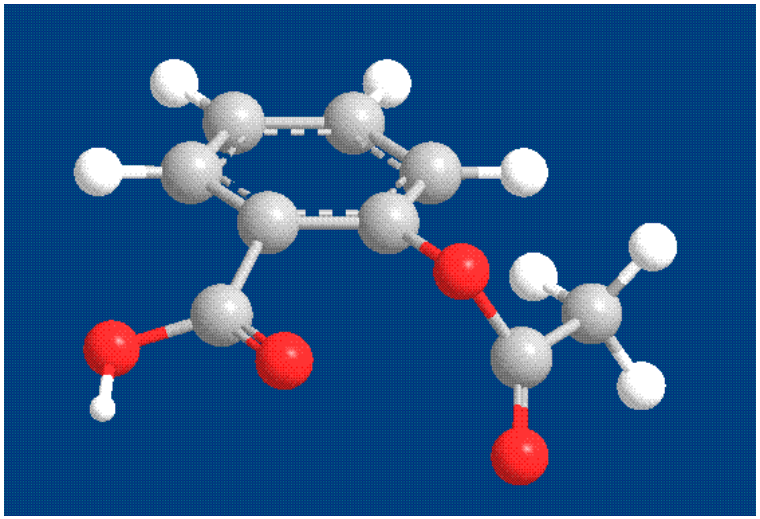
- Each document is a collection of words

# Transaction Data

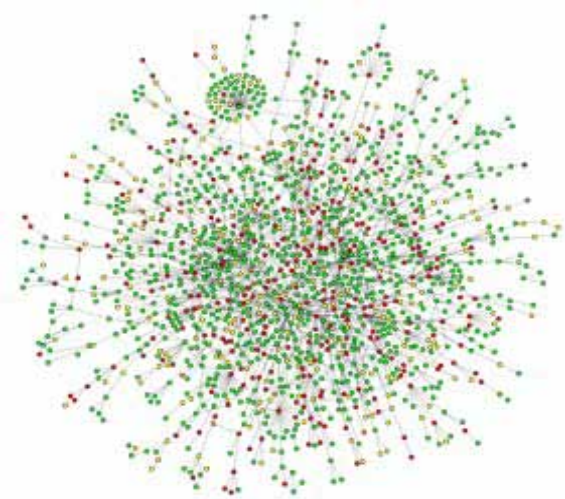
- A collection of transactions
  - Each transaction involves a set of items

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

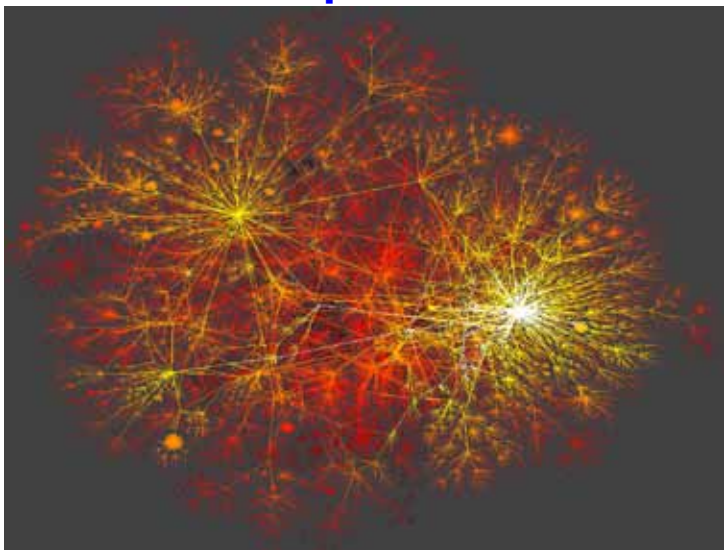
# Graph Data



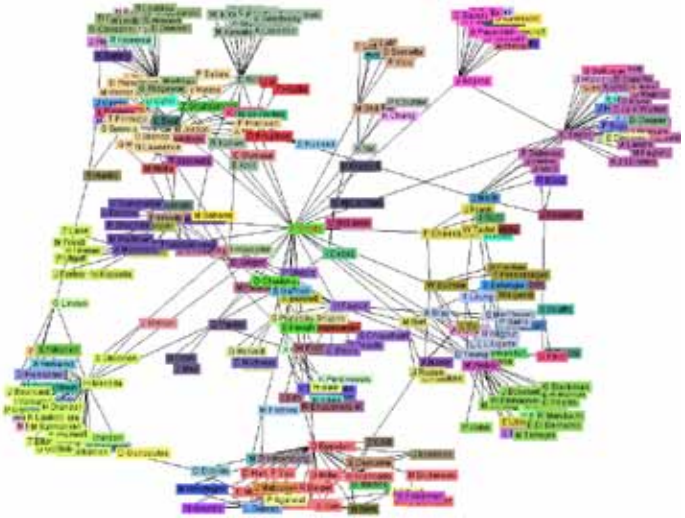
Aspirin



Yeast protein interaction network



Internet



Co-author network

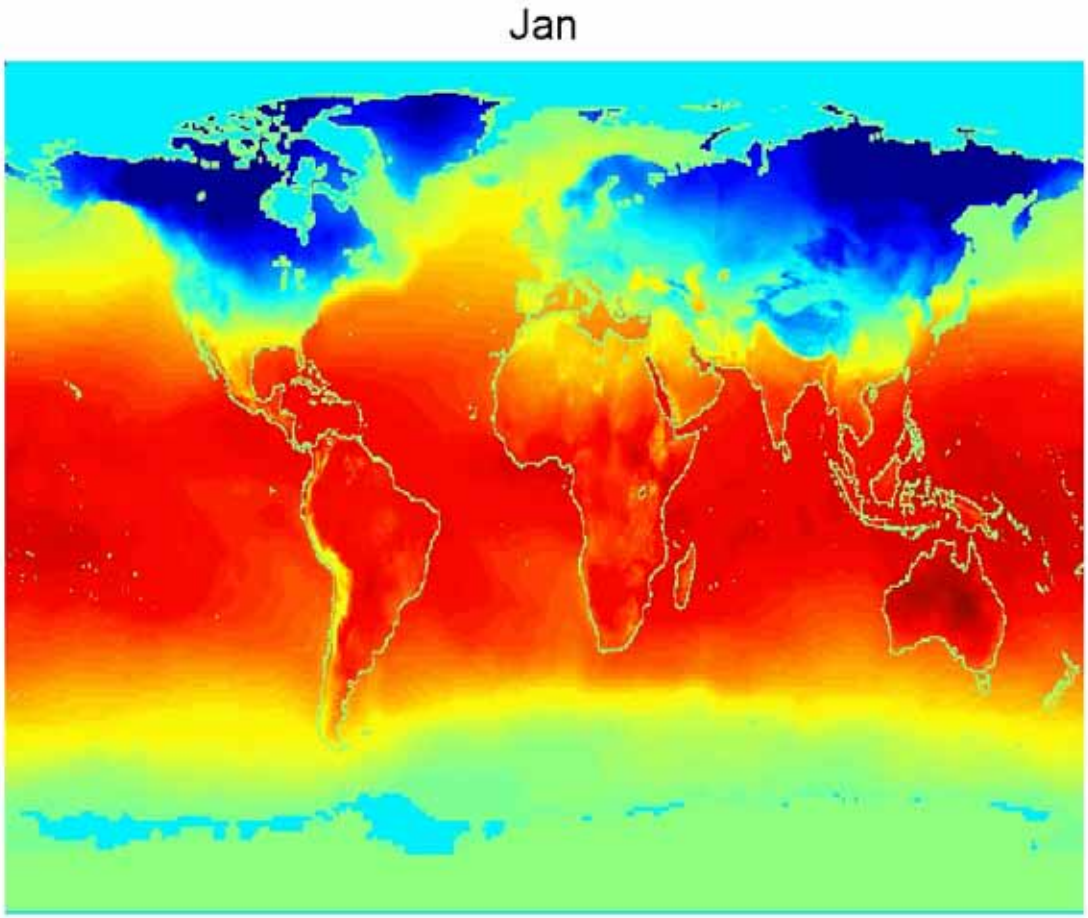
# Sequence Data

- DNA Sequence

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

# Spatial-Temporal Data

Average Monthly Temperature of land and ocean



# 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 |

# Record 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 |

# Types of Attribute

- **Categorical Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, weather conditions, or the set of words in a collection of documents
- **Numerical Attribute**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight

# Outline

- **Data**
- **Data Preprocessing**
  - Improve data quality
  - Prepare data for analysis
- **Exploring Data**
  - Statistics
  - Visualization

# Data Quality Issue

- **Data in the real world is dirty**
  - incomplete: lacking attribute values, lacking certain attributes of interest
    - e.g., occupation=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., Salary=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - Age=“42” Birthday=“03/07/1997”
    - Was rating “1,2,3”, now rating “A, B, C”
    - discrepancy between duplicate records

# Preprocessing

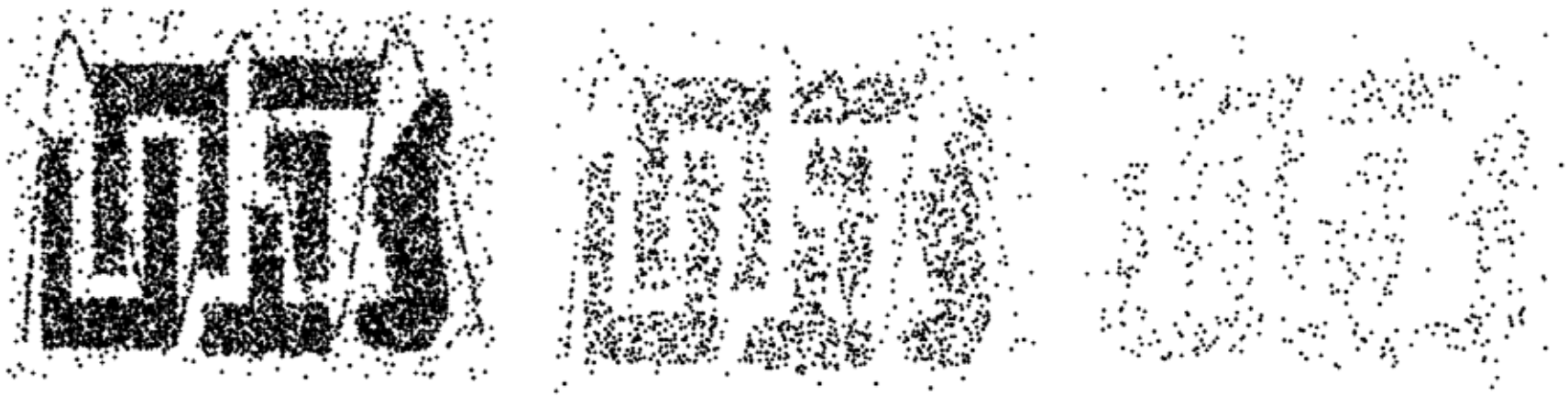
- **Handle missing values**
  - Ignore the records with missing values
  - Estimate missing values
- **Remove outliers**
  - Find and remove those values that are significantly different from the others
- **Resolve conflicts**
  - Merge information from different data sources
  - Find duplicate records and identify the correct information

# Prepare Data for Analysis

- Sampling
- Feature selection
- Dimensionality reduction
- Discretization

# Sampling

- Goal
  - Extract a subset of records so that the selected records are representative of original data



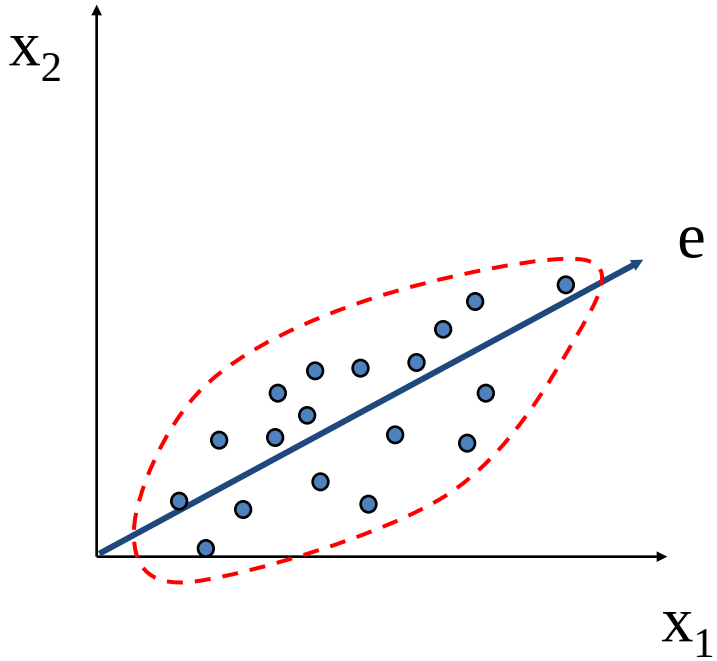
8000 points 2000 Points 500 Points

# Attribute Selection

- **Redundant attributes**
  - 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 attributes**
  - 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

# Dimensionality Reduction

- Goal is to reduce the number of attributes by creating a new set of attributes



# Discretization

- Binning
  - Convert numerical data into categorical data
  - Divides the range into *N* intervals

q Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into 3 bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

# Outline

- **Data**
- **Data Preprocessing**
  - Improve data quality
  - Prepare data for analysis
- **Exploring Data**
  - Statistics
  - Visualization

# Statistics: Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Notation : Let $x_1, x_{2,} \ldots x_n$ are $n$ observations of a variable $x$. Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

# Statistics: Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2, \ldots x_n$ is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}$$

Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

# Statistics: Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

| Age | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Frequency | 5 | 3 | 7 | 5 | 4 | 2 |

Grouped Frequency Distribution of Age:

| Age Group | 1-2 | 3-4 | 5-6 |
|---|---|---|---|
| Frequency | 8 | 12 | 6 |

# Question

- **Data Visualization**
  - Choose a few creative and fascinating examples of data visualization to show to the class

    http://selection.datavisualization.ch/