Type of Data and Data Pre-processing

Dr. Yuzana Win (Nagasaki University, Japan)
Lecturer
Department of Computer Engineering and Information Technology

Lecture Objectives

- To introduce
  - What is Data
  - Types of Data
  - What Data Pre-processing is used for?
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction

What is Data?

- Data is a collection of
  - raw facts
  - no context
  - numbers and text

Types of Data

- Data sets differ in a number of ways. Typically two types of data

(1) Quantitative
  - Numbers, tests, counting measuring

(2) Qualitative
  - Words, images, observations, conversions, photographs

❖This type of data determines which tool and techniques can be used to analyse the data.

Data Set

- A data set is a collection of data objects
  - sometimes called data objects as record, point, vector, pattern, event, case, sample, observation or entity

- Data objects are described by a number of attributes that capture the basic characteristics of an object

Example: Student Information

- each row corresponds to a student and
- each column is an attribute that describes some aspect of a student, such as grade point average (GPA) or identification number (ID)

Attributes

Student data object

Student ID	Year	Grade Point Average (GPA)
1034262	Senior	3.24
1052663	Sophomore	3.51
1082246	Freshman	3.62

What is an attribute?

- An attribute is a property or characteristic of an object that may vary, either from one object to another or from one time to another.
  - Example: eye color of a person, temperature, etc.
- Attribute is also know as **variable**, **field**, **characteristic**, *or* **feature**
- Attribute values are *numbers* or *symbols* assigned to an attribute

Type of Attributes

- There are different types of attributes
- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Example: ranking
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit
- Ratio
  - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order : < >
  - Addition: + -
  - Multiplication: * /

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to 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, the differences between values are meaningful, i.e., a unit of measurement exists. (+, −)	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, electrical current	geometric mean, harmonic mean, percent variation

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes or ID numbers, or numeric, such as counts.
  - Often represented as integer variables
  - Binary attributes (e.g., true/false, yes/no, male/female, or 0/1) are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers
  - Example: temperature, height, or weight
  - Continuous attribute are typically represented as floating-point variables

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

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

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.

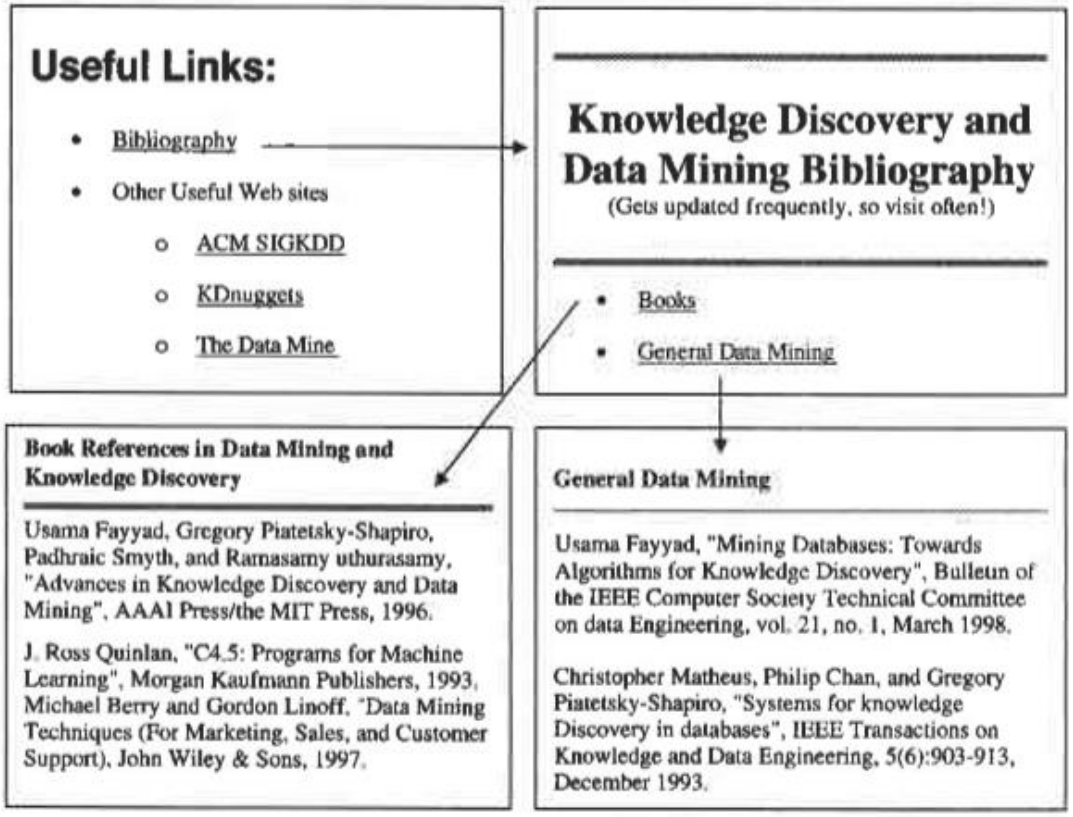
Transactions

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

CEIT
Think smart, be different.

Graph Data

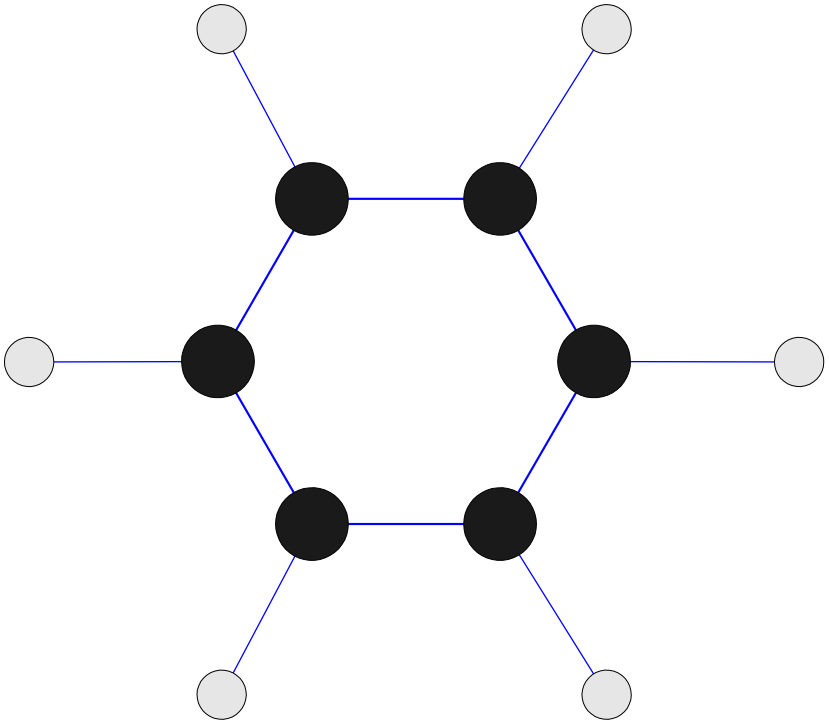
- The relationships among objects frequently convey important information. The data is often represented as a graph.



Molecular Data

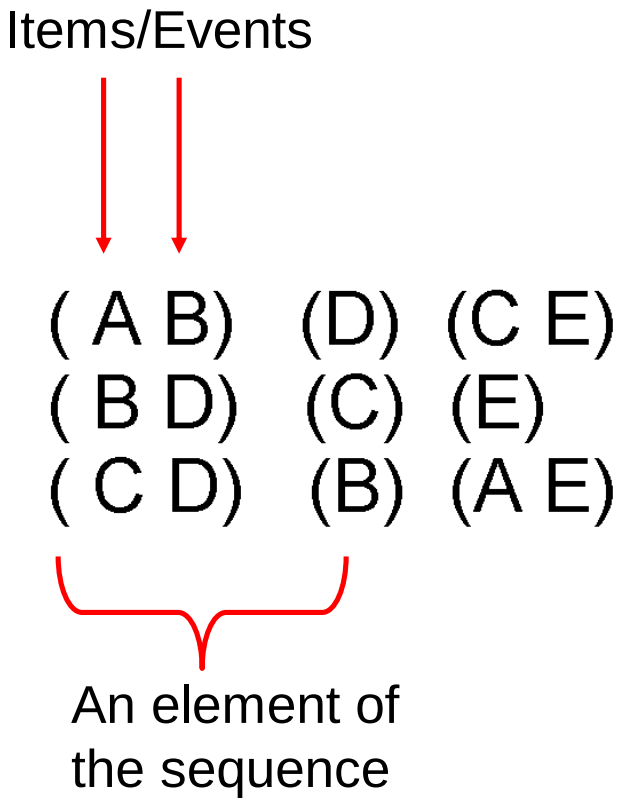
- If objects have structure and that contain subobjects have relationships, then such objects are frequently represented as graphs.

For example => Benzene Molecule: C_6H_6



Ordered Data

- The attributes have relationships that involve order in time or space
- Sequential data can be thought of as an extension of record data



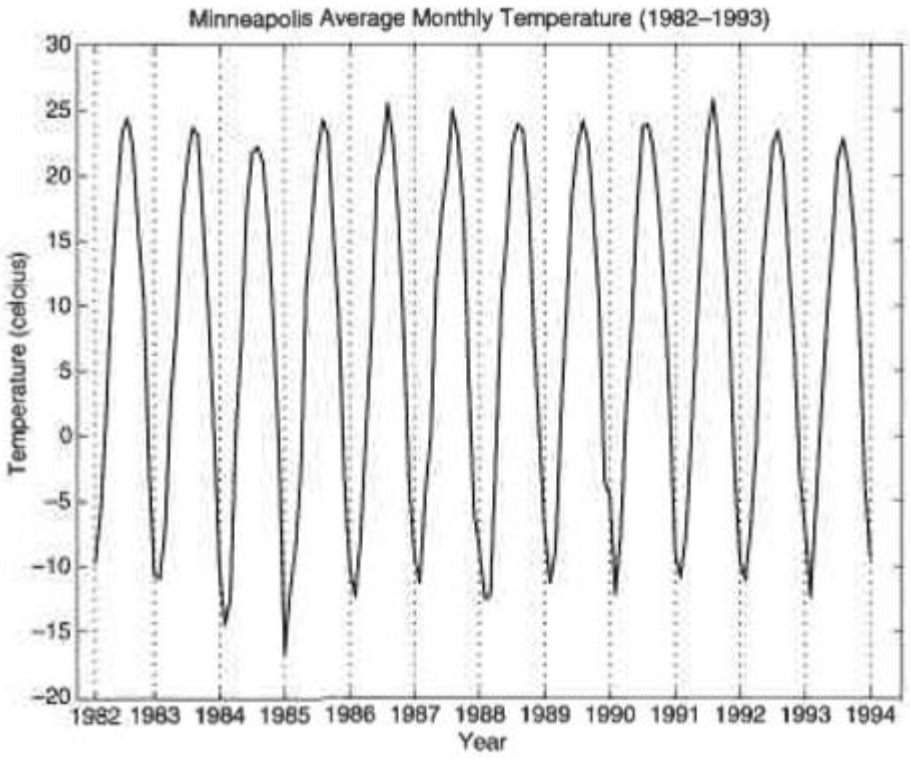
Ordered Data

- Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters.
- Eg. => Genomic sequence data

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

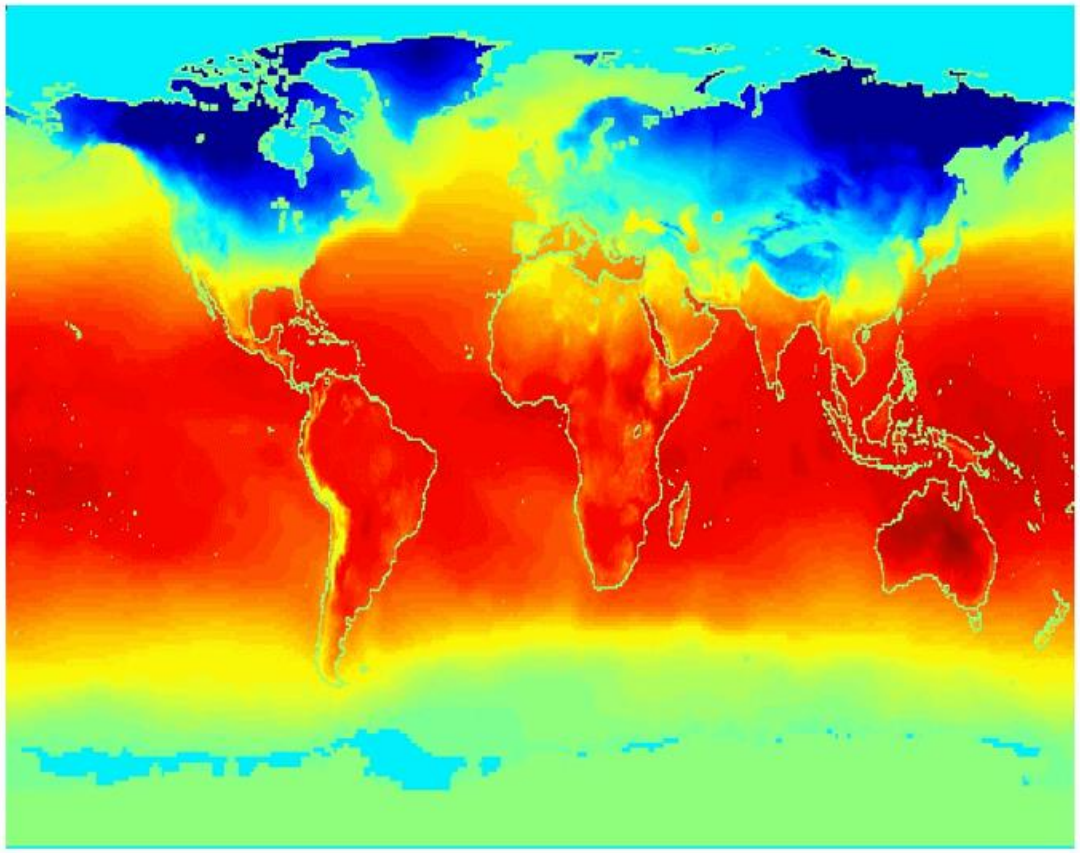
Ordered Data

- Time series data is a special type of sequential data in which each record is a time series, i.e., a series of measurements taken over time.
- Eg. => Temperature time series



Ordered Data

- Some objects have spatial attributes, such as positions or areas, as well as other types of attributes.
- Eg. = > average monthly spatial temperature data of land and ocean



What Data Pre-processing is used for?

- Data is in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors, duplicate or outliers
  - inconsistent: containing discrepancies in codes or names

- No quality data, no quality mining results

Therefore, to overcome the above problems, data pre-processing step is applied to make the data more suitable for analysing the data

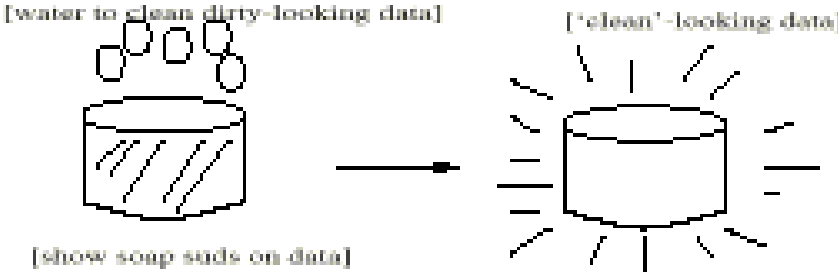
Major Tasks in Data Pre-processing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific range)
  - Aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

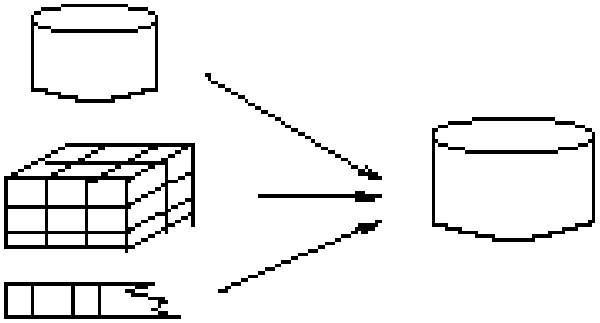
CEIT
Think smart, be different.

Data Pre-processing Tasks

Data Cleaning



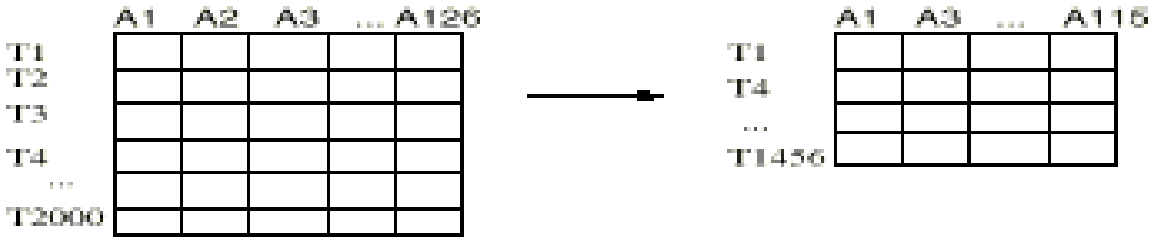
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

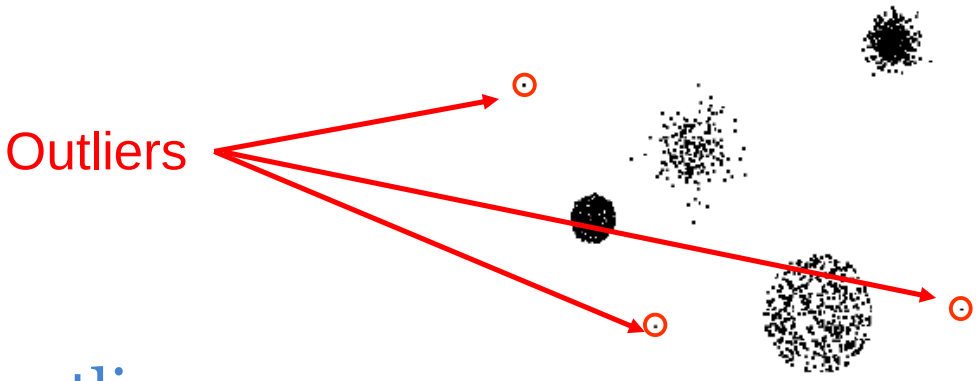
Missing Values

- 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)

Outliers

- Outliers are data objects that are considerably values of an attribute that are unusual with respect to the typical values for that attribute.



- Handle by Clustering
  - detect and remove outliers
- Handle by using Semi-automated method (combined computer and human inspection)
  - detect suspicious values and check manually
- Handle by Regression
  - smooth by fitting the data into regression functions

Correct Inconsistent Data

- Handling Correct Data
  - Manual correction using external references
- Semi-automatic using various tools
  - detect violation of known functional dependencies and data constraints
  - correct redundant data

Data Integration

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - **Data reduction**
    - Reduce the number of attributes or objects
  - **Change of scale**
    - Cities aggregated into regions, states, countries, etc
  - **More "stable" data**
    - Aggregated data tends to have less variability

Data Transformation

- **Smoothing**: remove noise from data
- **Aggregation**: summarization, data cube construction
- **Generalization**: concept hierarchy climbing
- **Normalization**: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

Data Reduction

- Obtains a reduced representation of the data set that is much smaller in volume but produces the same (or almost the same) analytical results.

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

Quiz

- Classify the following attributes as *binary*, *discrete*, or *continuous*. Also classify them as qualitative (*nominal or ordinal*) or quantitative (*interval or ratio*).

(1) Time in terms of AM or PM

(2) Angles as measured in degrees between 0 and 360

(3) Bronze, Silver, and Gold medals as awarded at the Olympics

(4) ISBN numbers for books

(5) Distance from the center of campus

Summary

- Data pre-processing is a big issue for data mining
- Data pre-processing steps includes
  - Data cleaning
  - Data integration and transformation
  - Data reduction
- A lot a methods have been developed but still an active area of research

Next Week Lecture

- Supervised Learning: Classification with Nearest Neighbor Classifier (KNN)

Thank You