Data

Introduction

- Data is the input information to be mined or visualized
- Different types of data sets possible
- Data mining techniques vary with the type of a data set

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

Important Characteristics of Structured Data

- **Dimensionality**
  - **Curse of Dimensionality**
- **Sparsity**
  - **Only presence counts**
- **Resolution**
  - **Patterns depend on the scale**

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

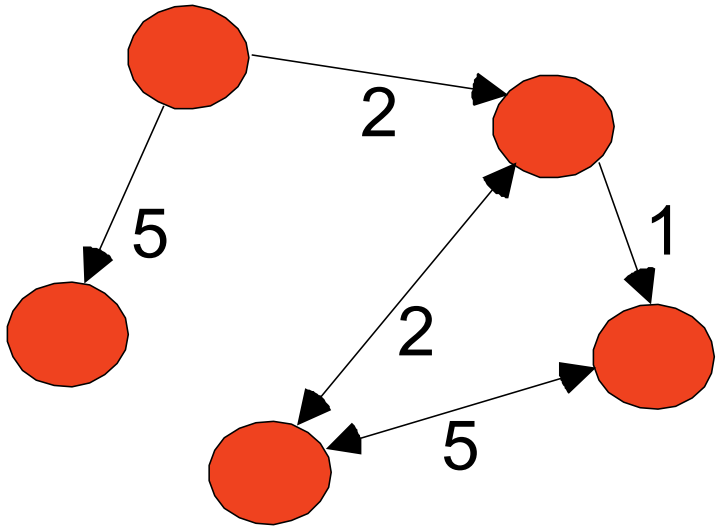
Transaction Data

- 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

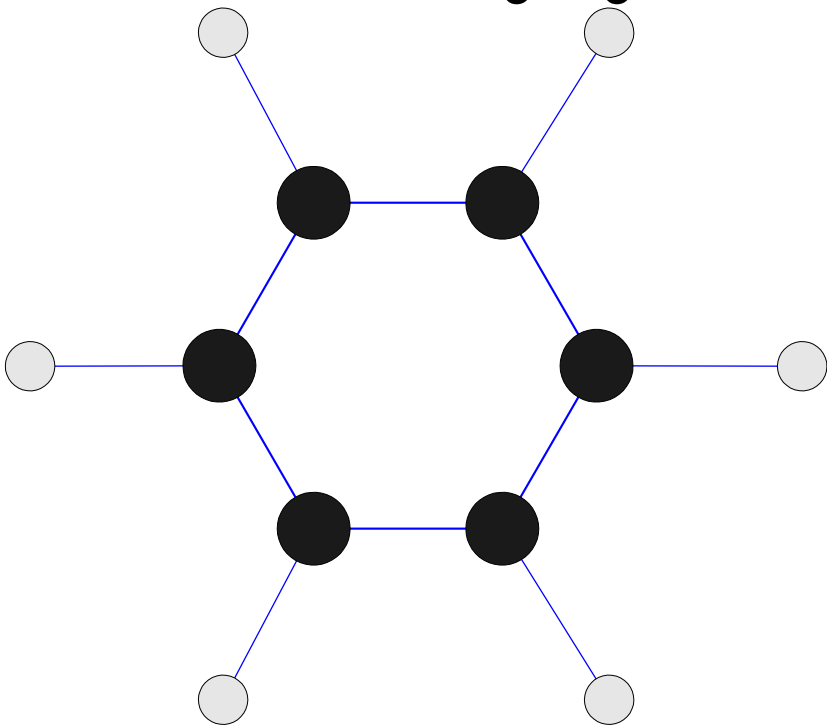
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

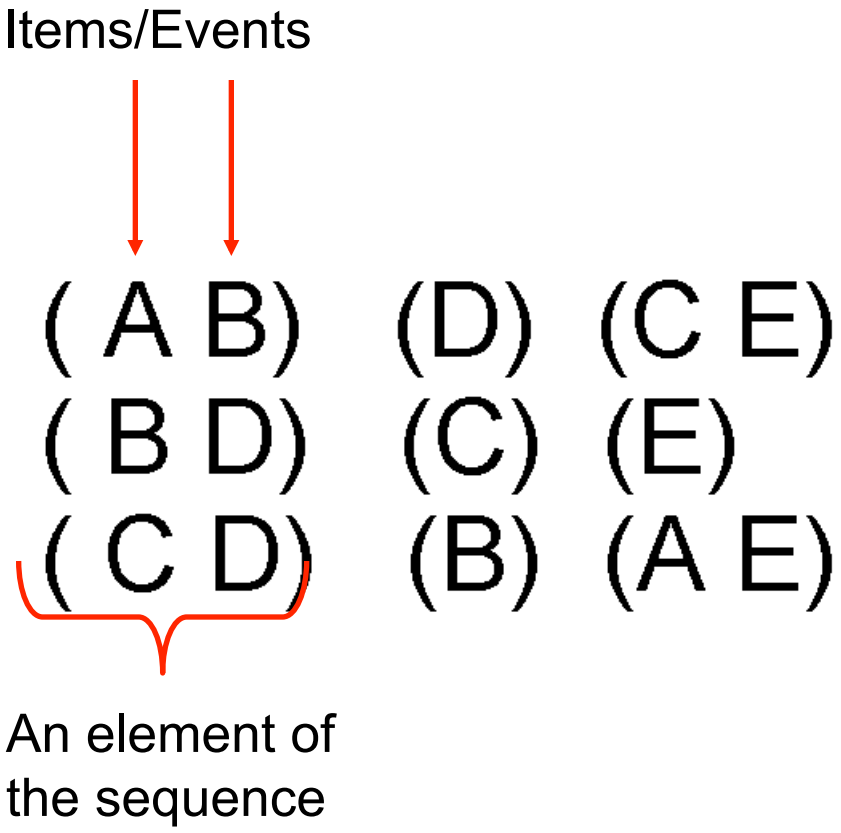
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- Sequences of transactions



Ordered Data

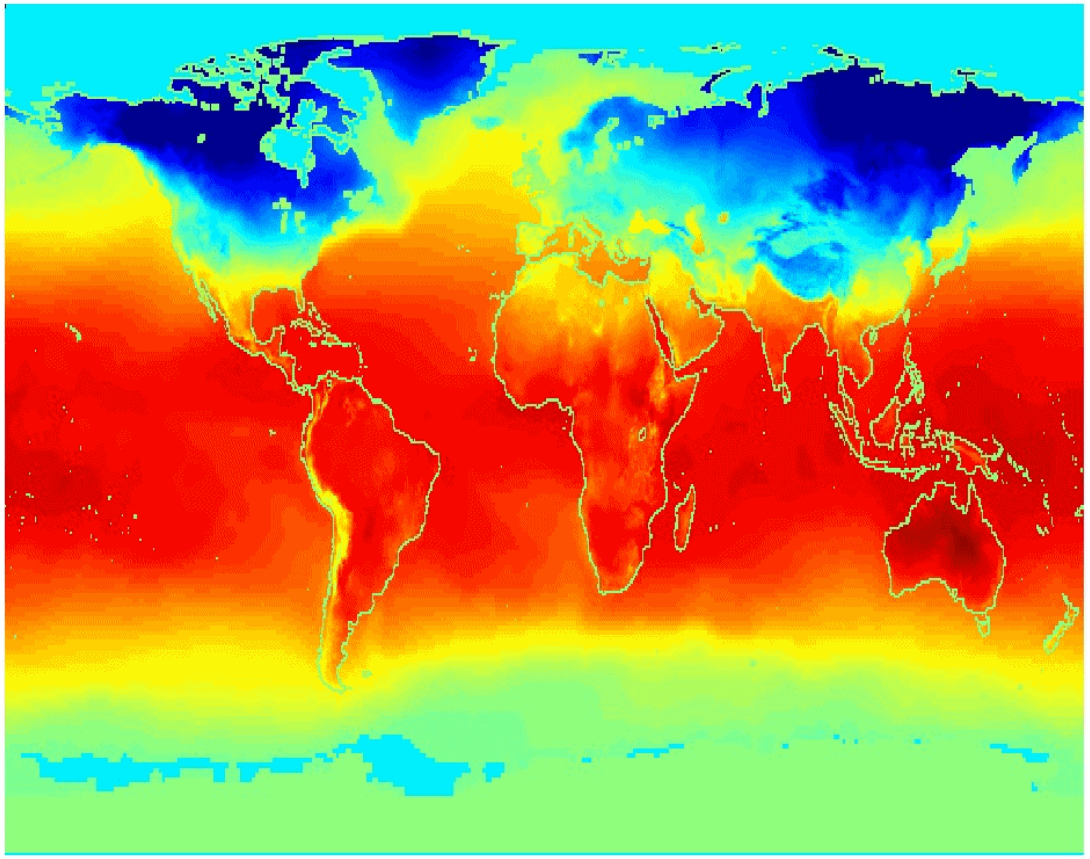
- Genomic sequence data

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

Ordered Data

- Spatio-Temporal Data

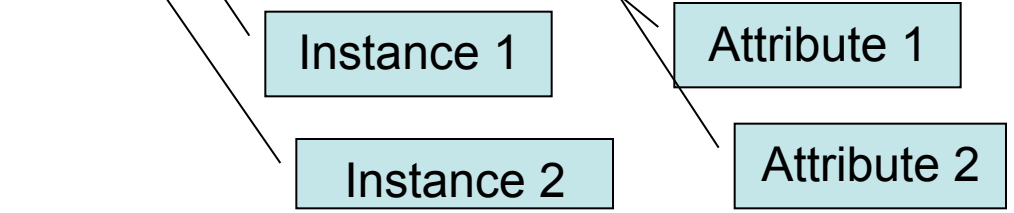
Average Monthly Temperature of land and ocean



Flat File Data (Record Data)

- Collection of instances of something
- Each instance is described using a finite set of attribute
- Example weather data

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes



Majority of data mining techniques work on flat file data

At the moment we focus on this type of data set

Data Mining Tasks on Flat File Data

- Classification
  - Predicting the value of the class attribute (e.g. Play in the weather data) based on the values of the other attributes of an input instance
- Clustering
  - Grouping together input instances with 'similar' attribute values
- Association Rule Mining
  - Predicting the value of any of the attributes based on the values of one or more remaining attributes of an input instance
  - Similar to classification
- Summary of Data Mining
  - Instances are either classified or clustered.
    - Using attribute values

Relational Data

- Relational data is distributed among several relations (tables) linked by relational keys.
- Each relation stores attribute data corresponding to a real world entity (identified in Entity-Relationship modelling)
- Data management is easier with relational database but data mining is harder
- Flat files can be created by denormalizing two or more relations
  - a reverse of normalization learned in database courses
- Such flat files may contain spurious regularities
  - E.g, supplier address predicted from supplier

Attribute Values

- Attribute values represent a measurement of that attribute's quantity
- Attribute values can be
  - Discrete – come from a finite or countably infinite set of values
  - Continuous – real numbers
- Statisticians define four levels of measurement
  - Nominal – labels or names – e.g. {rainy, overcast, sunny}
  - Ordinal – orderable labels – e.g. {hot>mild>cold}
  - Interval – equidistant and orderable numbers – e.g. {85 in Fahrenheit}
  - Ratio – equidistant and orderable numbers with a defined zero – e.g. length measurements

Operations allowed on Levels of Measurement

- Nominal
  - = and ≠
- Ordinal
  - Operations allowed for nominal and
  - < and >
- Interval
  - Operations allowed for Ordinal and
  - + and –
- Ratio
  - Operations allowed for interval and
  - * and /

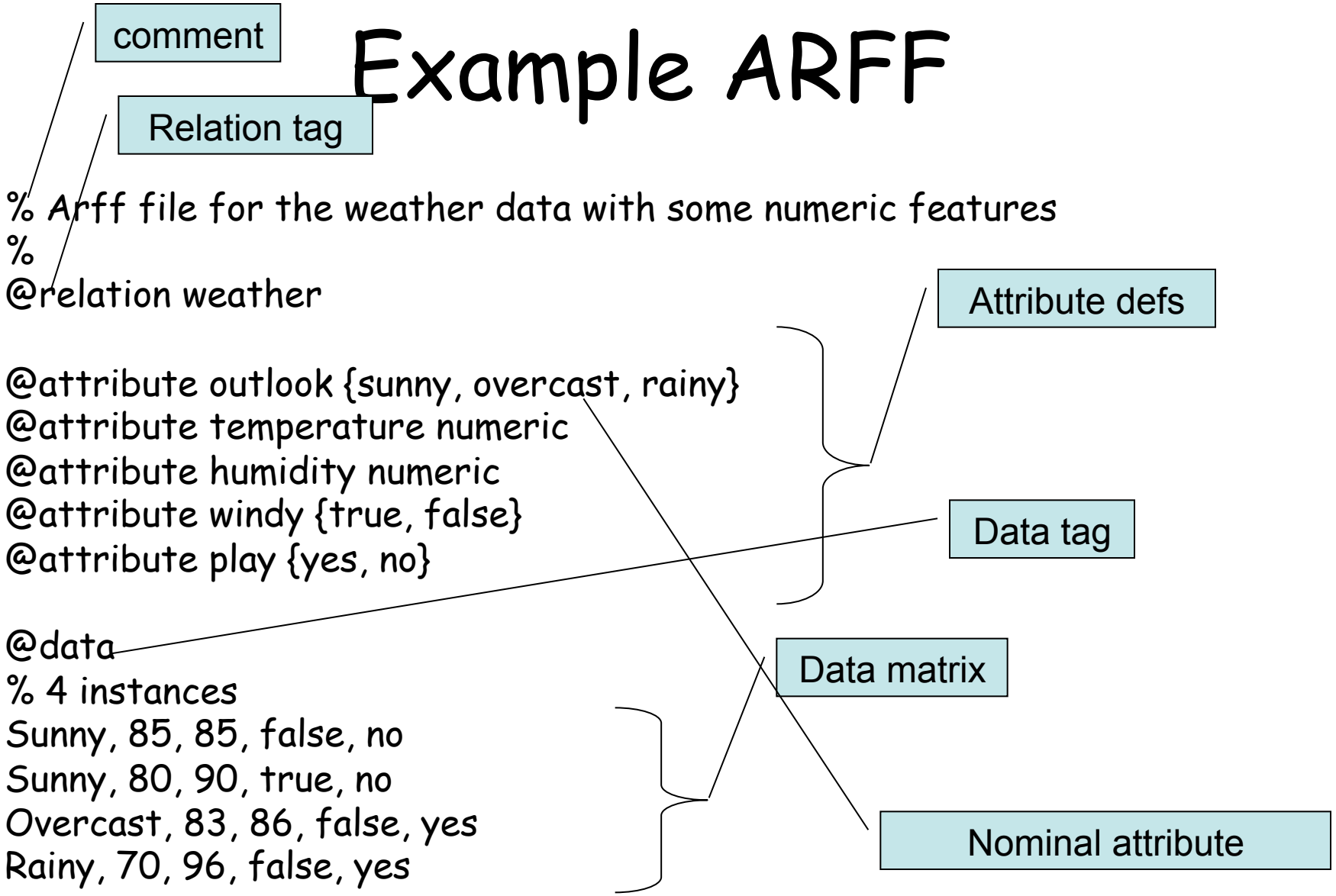
Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency 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
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

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

ARFF file format

- Attribute relation file format
- Important information about data is present in metadata
- Metadata needs to be packed into the data set
  - Arff achieves this using special file formatting
- Arff has four components
  - Comments – % as the first character of a line – rest of the line is treated as the comment
  - Name of the relation – @relation tag followed by the relation name
  - Block of attribute definitions - @attribute tag followed by name and type of attribute
  - Actual data - @data tag followed by the data matrix on a new line with attribute values ordered similar to the order used in the attribute definitions

Example ARFF



Pre-processing

- Most important step in data mining
  - Input data is never readily available for data mining
- Several iterations required to get it right
- Involve
  - Data warehousing
  - Sparse Data
  - Attribute types
  - Missing values
  - Inaccurate values

Data Warehousing

- Each department of an organization manages data in its own way
  - Record keeping style
  - Conventions
  - Degrees of data aggregation
  - Different primary keys
- University has student record database, staff pay role database etc.
- Data warehousing is the integration of departmental data
  - In some cases may require overlay data to be integrated as well
  - Overlay data refers to data not usually collected by an organization
    - E.g demographic data
  - In some cases may require appropriate aggregation of data
    - E.g. number of hours spent on research to be added to the number of hours spent on teaching

Sparse Data

- Input data instances may contain 'zero' as the value for most of the attributes
  - E.g. market basket data matrix with customers as instances (rows) and shopping items as attributes (columns) contains zero purchases for most shop items
- Such sparse data in arff wastes lot of file space with zeros
- Arff allows alternative data specification for sparse data
- Example
  - Sparse data in normal arff format

  0,26,0,0,0,0,63,0,0,0, "class A"

  0,0,0,42,0,0,0,0,0,0, "class B"
  - Sparse data in special arff format

  {1 26, 6 63, 10 "class A"}

  {3 42, 10 "class B"}
- Sparse data has lot of zeros, not missing values – which are discussed later

Attribute Types

- Arff files use mainly two data types, nominal and numeric.
- Numeric measurements can be interpreted differently by different data mining techniques
- Knowledge of inner workings of data mining technique required to define attribute values
  - Only then the operations performed by data mining technique are meaningful
- E.g. when a data mining algorithm performs operations allowed for ratio scales, numeric data is normalized
- Standard way of normalizing data is
  - Subtract the mean of the attribute from each value and divide the deviation with the standard deviation of the attribute
  - The resulting standardized data has a mean of zero and standard deviation of one.
- Distances between attributes with ordinal scales need to be defined meaningfully
  - Zero if the values are different and one if they are the same
- Some nominal attribute values might naturally map onto some numeric values
- On the other hand, some numeric values might be simply numerically coded nominal values

Missing Values

- Similar to the 'null' values in databases
- Semantics of null values are not well defined
  - Unknown or unrecorded or don't cares
- Default assumption is
  - Missing values are irrelevant (don't care) for data mining
- The exact semantics of missing values useful for data mining
  - Knowledge of the domain context required to define the exact semantics
  - E.g. medical diagnosis possible based on the tests doctor decided to make, rather than the results of the tests
- Arff files use '?' to denote missing values
- If the meaning of the missing value is known an additional value 'test not done' can be added to the attribute values

Inaccurate Values

- Data for data mining is collected from several sources
  - Each source collects data for a purpose other than data mining
  - This means, input data always contains attributes that are suitable for the original purpose, but lack generality
    - Tolerable errors and omissions in the original data set assume significance for data mining tasks
- Several sources of errors
  - Typographic errors
    - Show up as outliers
  - Duplicates
  - Systematic Errors
    - Supermarket checkout operator using his own loyalty card when customer does not supply his loyalty card
  - Stale data
    - Addresses and telephone numbers change all the time

Data Inspection

- Summary: know thy data before thou apply data mining!!
  - Types of Data
  - Types of Attributes
  - ARFF
  - Data Pre-processing

Next Lecture

- Exploratory Data Analysis (EDA)

- Core Reading:
  - Lecture Slides in MyAberdeen
- Recommended Reading:
  - Kumar Book Chapter 3.1~3.3
- Further Reading:
  - Chapter 3.4, Exercises of Chapter 3

Acknowledgement

- Some of the slides are based on the course slides provided by
  - Tan, Steinbach and Kumar (Introduction to Data Mining)

- Some pictures are taken from various online resources.