DATA AND ATTRIBUTE

BEIYU LIN

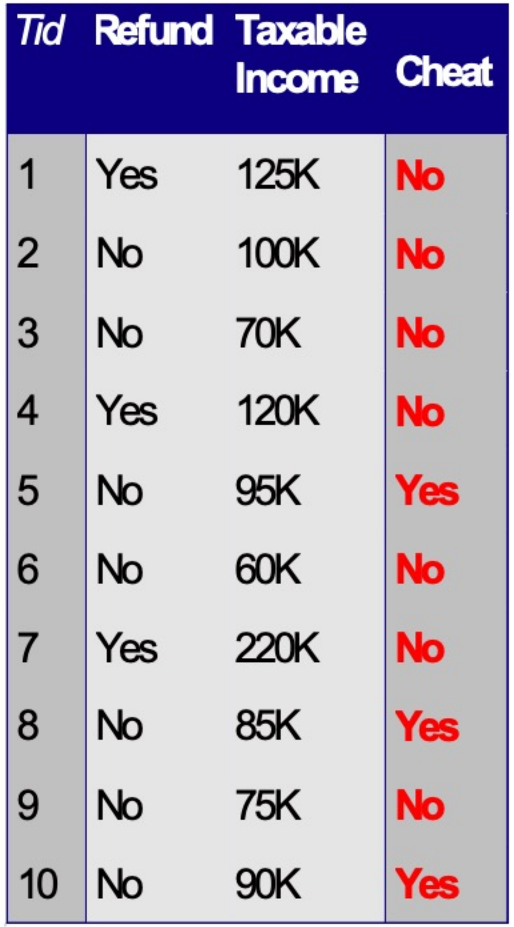
REVIEW

- **What is Data Mining?**
- **Why Data Mining is important?**
- **Data Mining and its Applications**
- **Real Life Examples**

WHAT IS DATA

A collection of data objects and their attributes

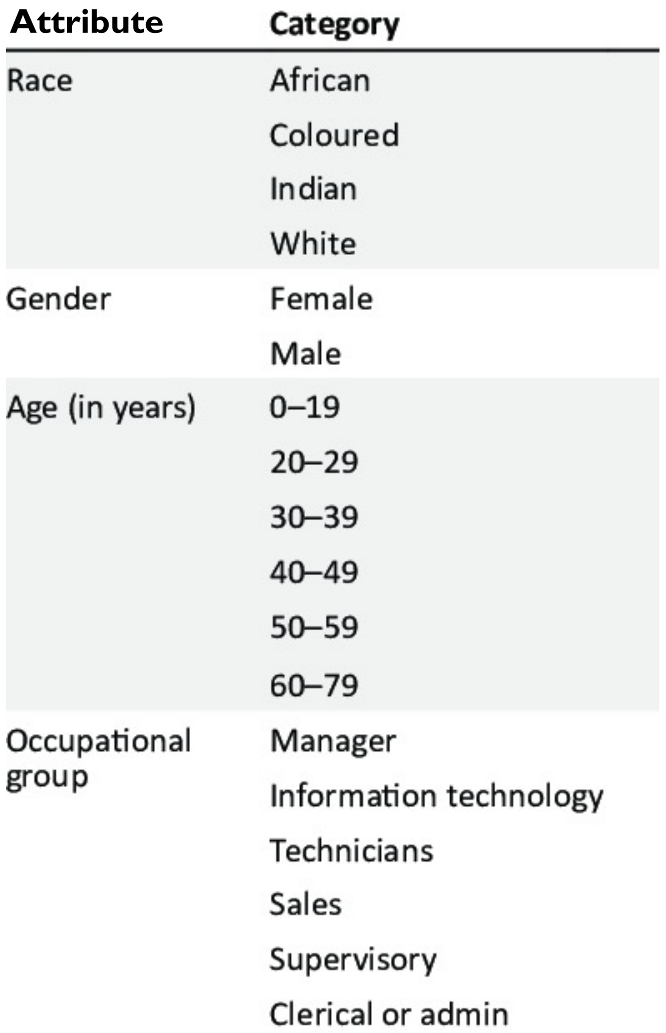
- An attribute is a property or characteristic of an object
  - – Examples: age, gender, income of a person
  - – also known as variable, field, characteristic, dimension, or feature
- An object is described by a collection of attributes
  - – also known as record, point, case, sample, entity, or instance



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

ATTRIBUTE VALUES

- Attribute values
  - numbers or symbols assigned to an attribute for a particular object
- Attributes and attribute values
  - – Same attribute can be mapped to different attribute values
    - Example: occupational group can be measured in sales or technicians
  - – Different attributes can be mapped to the same set of values



Attribute	Category
Race	African
	Coloured
	Indian
	White
Gender	Female
	Male
Age (in years)	0–19
	20–29
	30–39
	40–49
	50–59
	60–79
Occupational group	Manager
	Information technology
	Technicians
	Sales
	Supervisory
	Clerical or admin

TYPES OF ATTRIBUTES

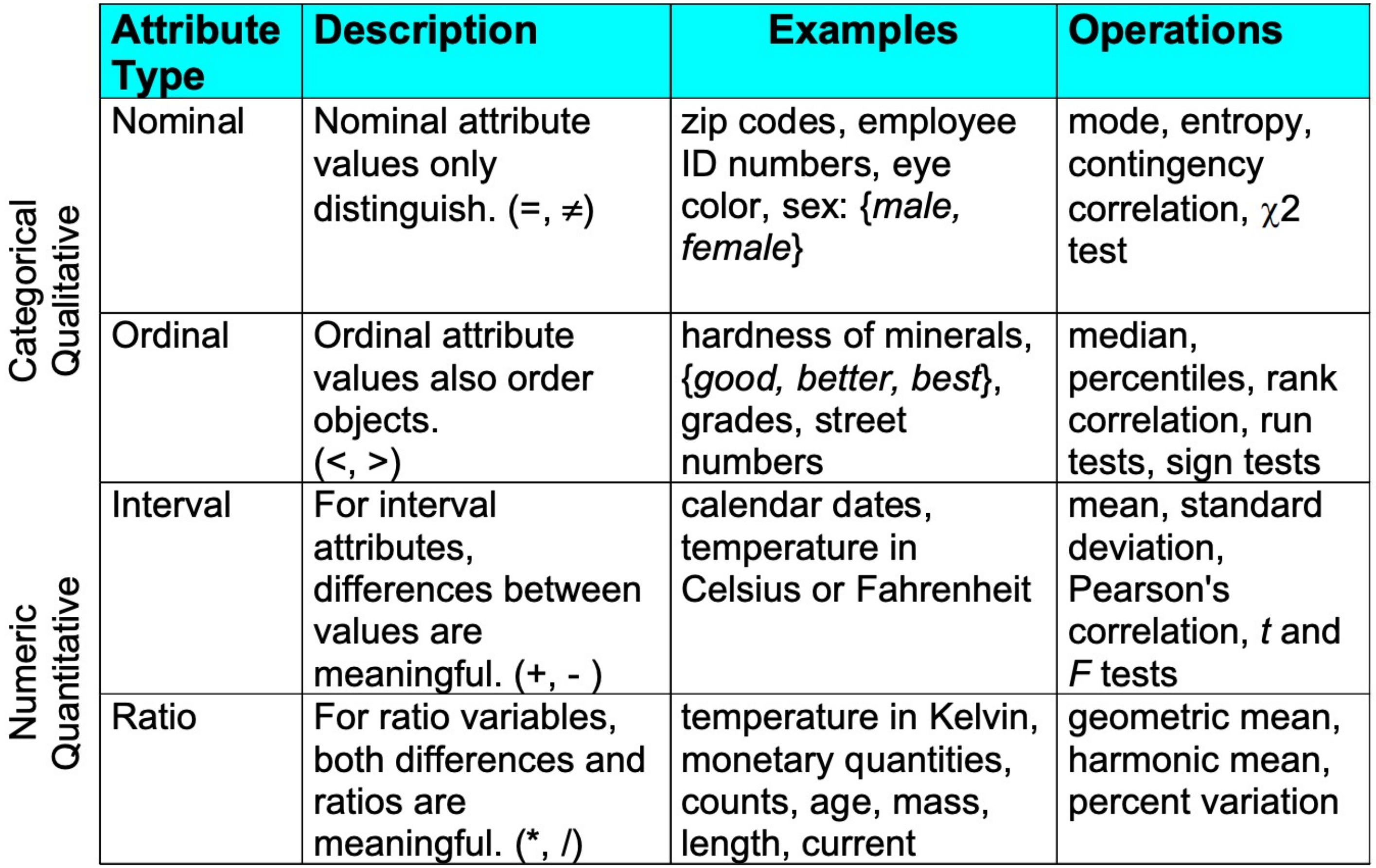
- – Nominal
  - Examples: ID numbers, zip codes
- – Ordinal
  - Examples: rankings
    - taste of red wine on a scale from 1-10; grades (A, A-, B+, B, B-, …);
- – Interval
  - Examples: calendar dates.
- – Ratio
  - Examples: elapsed time (e.g., time to go to school)

PROPERTIES OF ATTRIBUTE VALUES

- The type of an attribute depends on the properties/operations it possesses:
  - – Distinctness: = ≠
  - – Order: < >
  - – Differences: + -
  - – Ratios: * /

- Nominal attribute: distinctness (e.g., ID number)
- Ordinal attribute: distinctness & order (e.g. ranking)
- Interval attribute: distinctness, order & meaningful differences (e.g., calendar dates)
- Ratio attribute: all 4 properties/operations

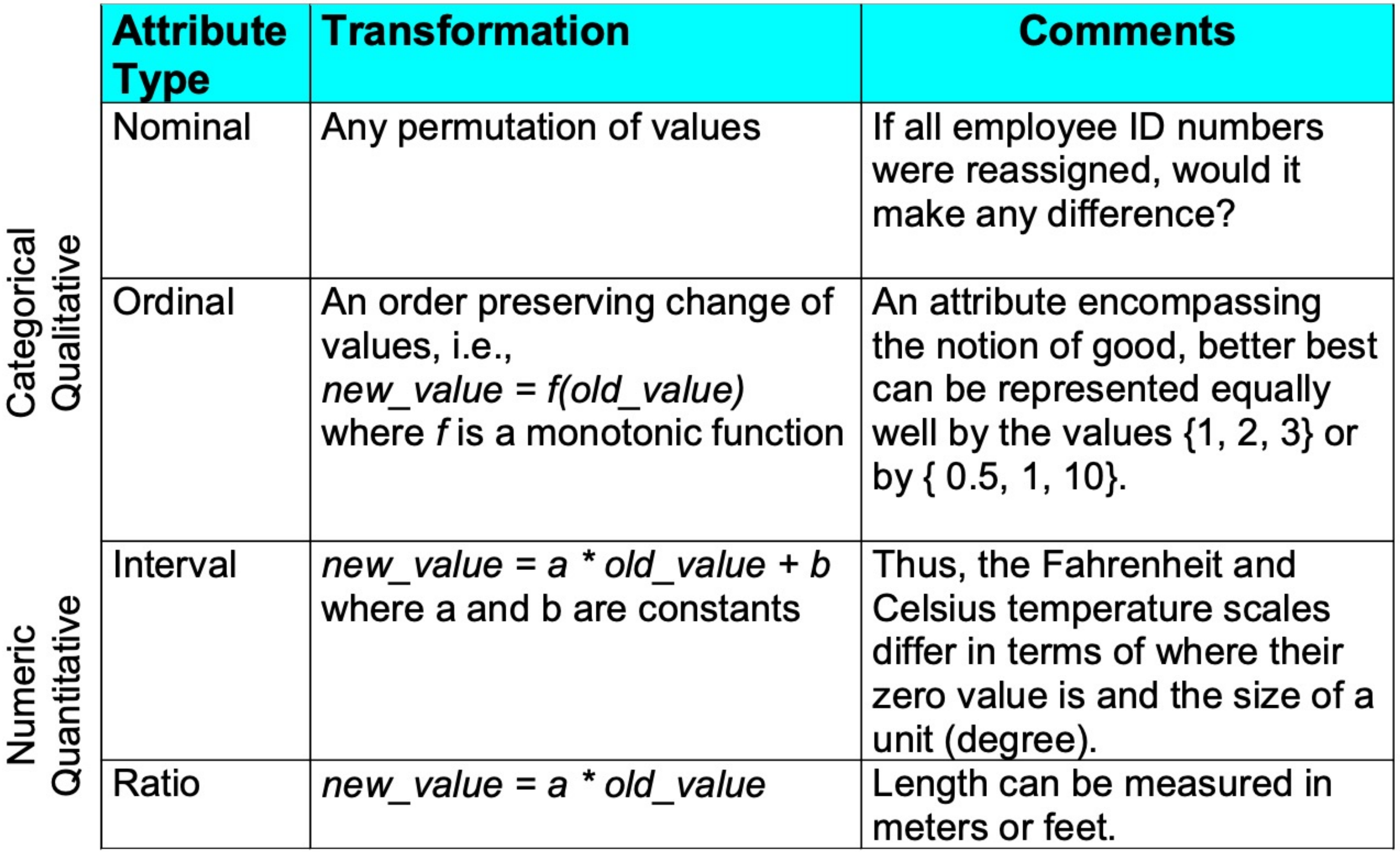
ATTRIBUTE



	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

This categorization of attributes is due to S. S. Stevens

ATTRIBUTE



	Attribute Type	Transformation	Comments
Categorical Qualitative	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}.
Numeric Quantitative	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.

This categorization of attributes is due to S. S. Stevens

DISCRETE AND CONTINUOUS ATTRIBUTES

- Discrete Attribute
  - – a finite or countably infinite set of values
  - – Examples: a set of eye colors (blue, green, black, brown…)
  - – Often represented as integer variables (e.g., 0, 1, 2, ..).
  - – Note: binary attributes are a special case of discrete attributes (e.g., 1 or 0)
- Continuous Attribute
  - – Has real numbers as attribute values (e.g., heights: 5.4, 6.3, etc)
  - – Examples: temperature, height, or weight.
  - – Continuous attributes are typically represented as floating point variables (weight: 111.5 pounds)

CRITIQUES OF THE ATTRIBUTE CATEGORIZATION

- Incomplete

– Partially ordered (e.g. missing values)

– Partial membership

– Asymmetric binary

– Cyclical (e.g., daily routines)

– Multivariate

– Relationships between the data (auto-correlations; not independent; not from an identical distribution)

- Real data is approximate and noisy

– may not recognize the proper attribute types

– approximate one attribute type by another

KEY MESSAGES FOR ATTRIBUTE TYPES

Choose the operations that are "meaningful" for the type of data

– distinctness, order, intervals, and ratios are only four properties of data

– the data type you see

– may not capture the full information or may suggest hidden properties

– analysis may depend on other properties of the data (e.g., many statistical analyses depend only on the distribution)

– it may vary in different domains.

IMPORTANT CHARACTERISTICS OF DATA

- Dimensionality (number of attributes) -- high dimensional data brings some challenges

  (e.g., using genes as attribute while studying health related problems)

- Sparsity -- only presence counts

  (e.g., the GPS based traffic trajectory data)

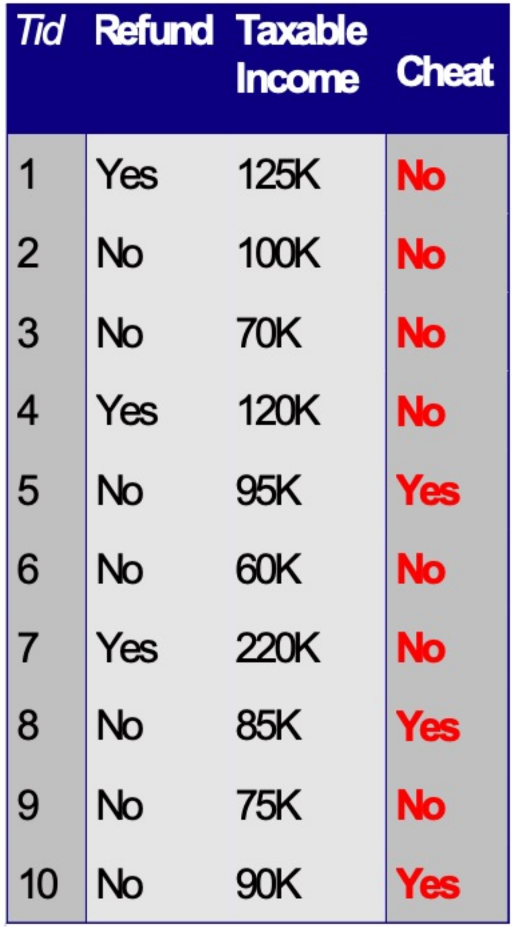
- Resolution – patterns depend on the scale
- Size of data

TYPES OF DATA SETS

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data (e.g., IoT data)
  - Genetic Sequence Data

RECORD DATA

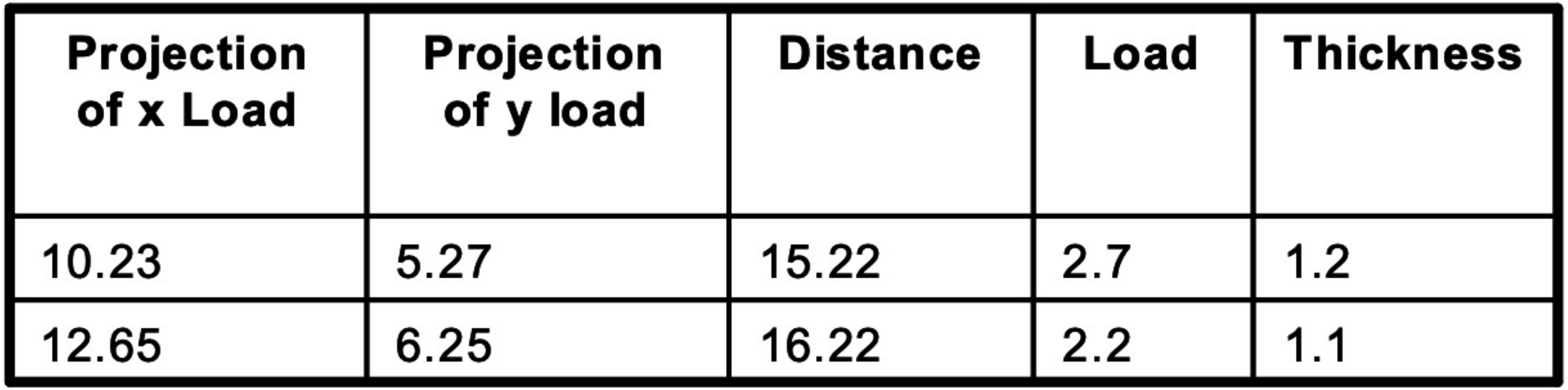
- a collection of records, each of which consists of a fixed set of attributes



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

DATA MATRIX

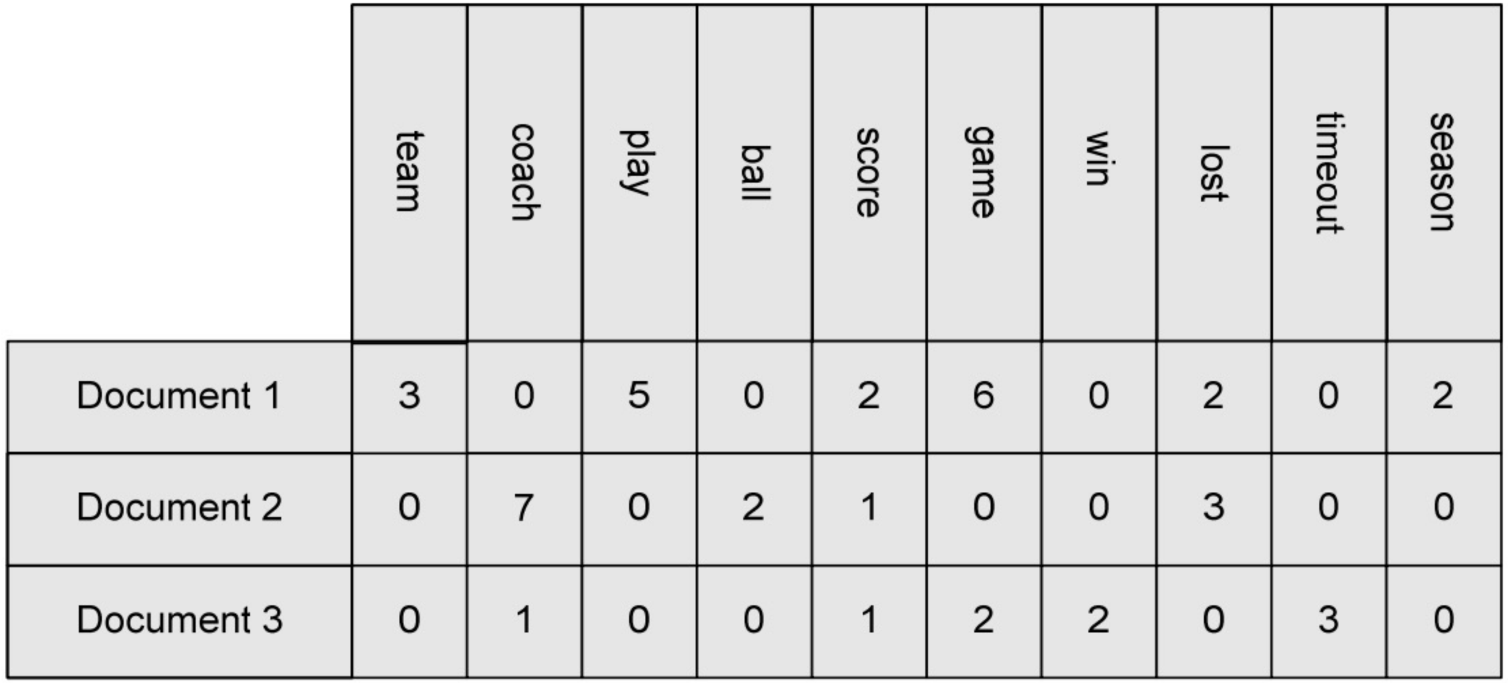
- with same fixed set of attributes and multiple objects.
- data set can be represented by an m (rows) by n (columns) matrix



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

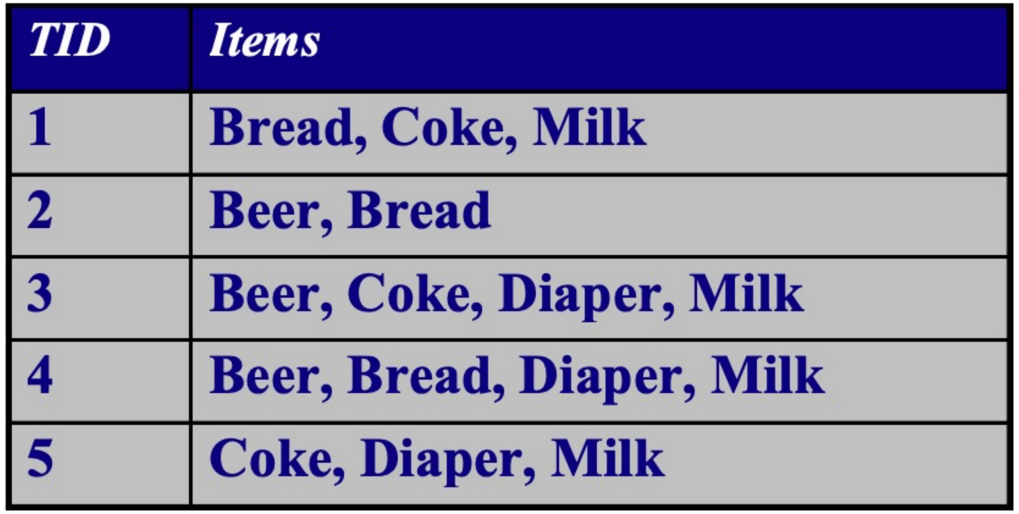
- Each document becomes a ‘term’ vector (or word-based)
  – each term/word is an attribute of the vector
  – the value of each attribute is the number of times this term / word occurs in the document.



	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

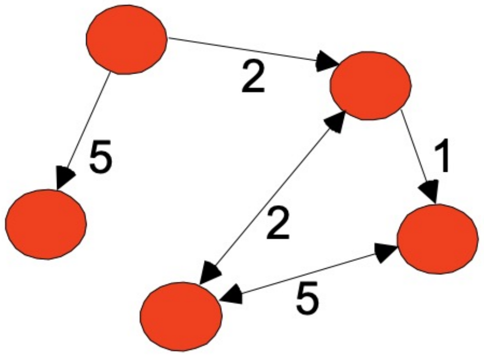
- A special type of data, where
  – each transaction involves a set of items.
  – for example, a person went for a grocery shopping.
    a. the set of products purchased is a transaction,
    b. the individual products that were purchased are the items.
  – can also represented as record data

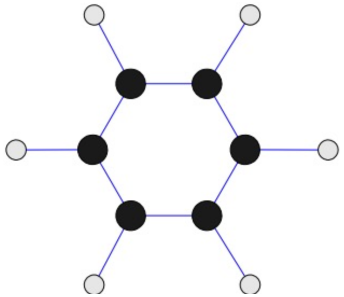


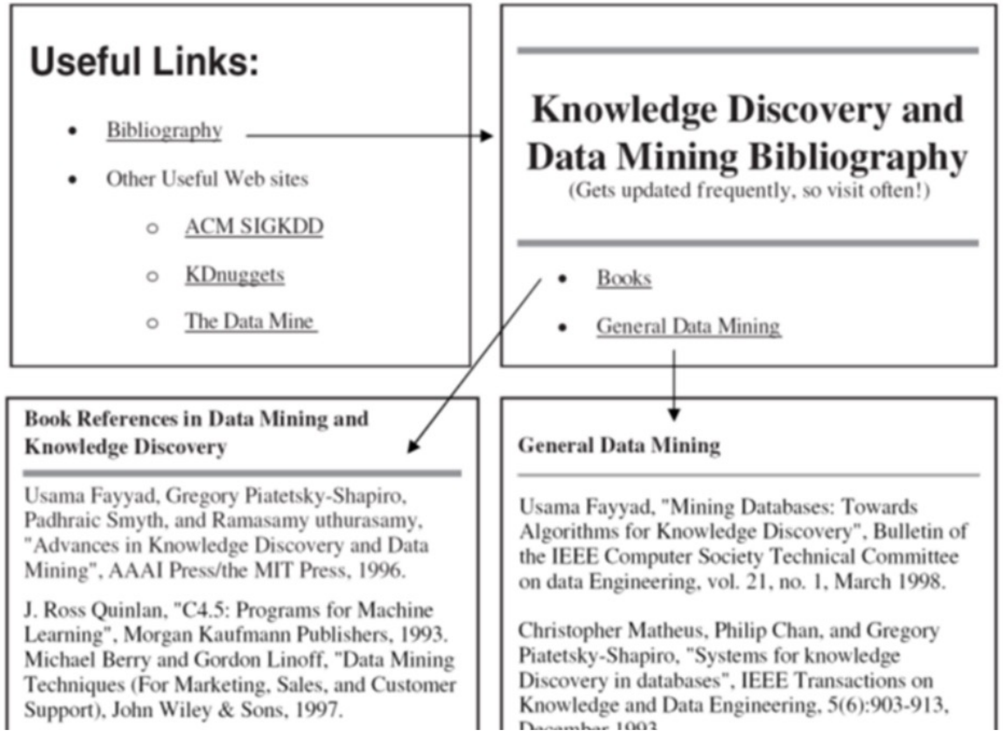
TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

GRAPH DATA

- Examples: Generic graph, a molecule, and webpages







Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

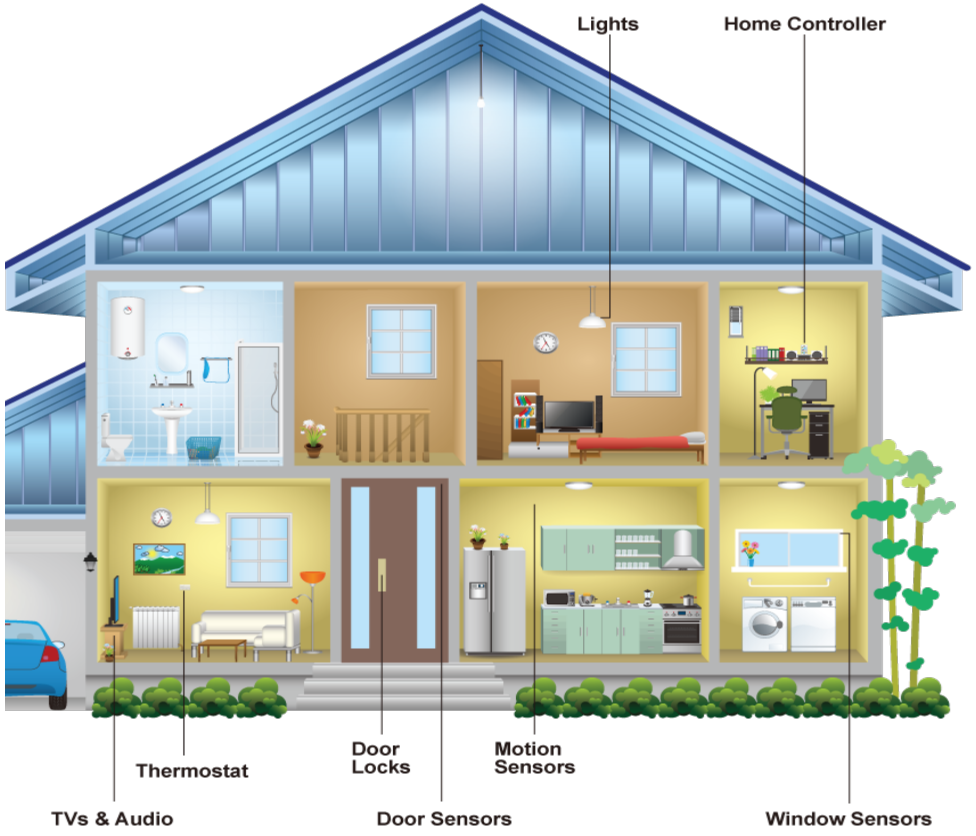
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

IOT DATA

Sensor types: infrared motion(narrow/wide-area), ambient light, magnetic, and temperature sensors.



imgbin.com

```
2011-06-13   21:48:43   Bathroom   ON    Personal_Hygiene
2011-06-13   21:48:44   Bathroom   OFF   Personal_Hygiene
2011-06-13   22:47:02   Bedroom    ON    Personal_Hygiene
2011-06-13   22:47:04   Bedroom    OFF   Sleep
2011-06-13   22:47:06   Bedroom    ON    Sleep
...............
…
2011-06-14   10:11:24   Kitchen    ON    Wash_Dishes
2011-06-14   10:11:25   Kitchen    OFF   Wash_Dishes
2011-06-14   10:11:40   Kitchen    ON    Cook
2011-06-14   10:11:41   Kitchen    OFF   Wash_Dishes
```