



IS414: Data Mining

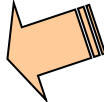
Dr. Waleed M. Ead
waleedead@hotmail.com



IS414: Data Mining

Chapter 2. Getting to Know Your Data

Chapter 2. Getting to Know Your Data

- ❑ Data Objects and Attribute Types 
- ❑ Basic Statistical Descriptions of Data
- ❑ Data Visualization
- ❑ Measuring Data Similarity and Dissimilarity
- ❑ Summary

Types of Data Sets: (1) Record Data

- Relational records
 - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

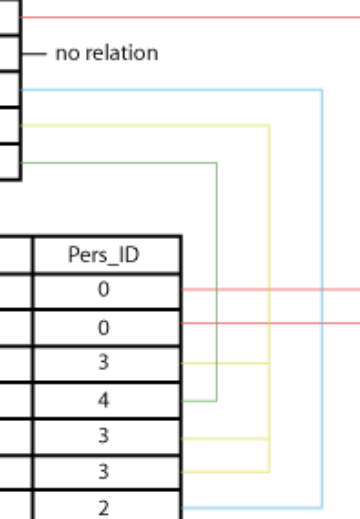
	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2



- Transaction data

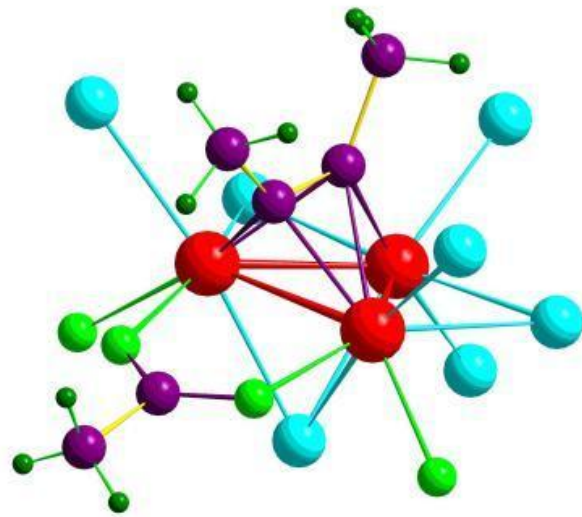
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

	team	coach	pla y	ball	score	game	wi n	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

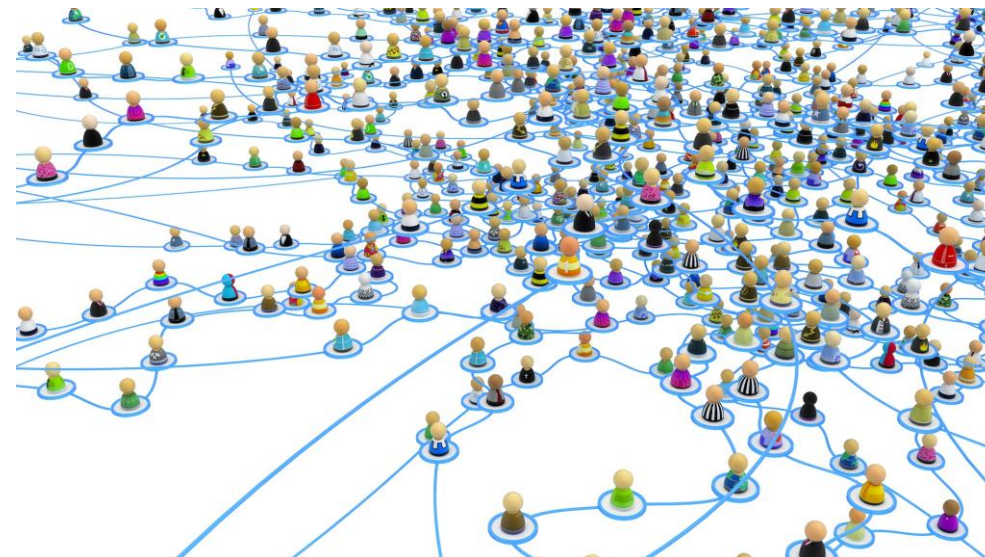
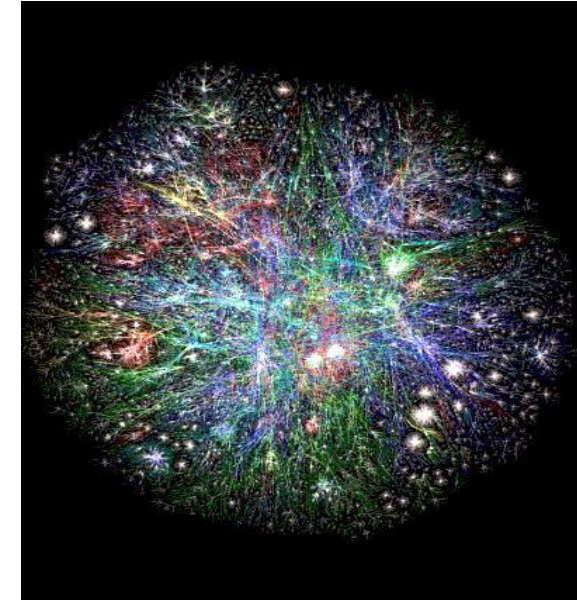
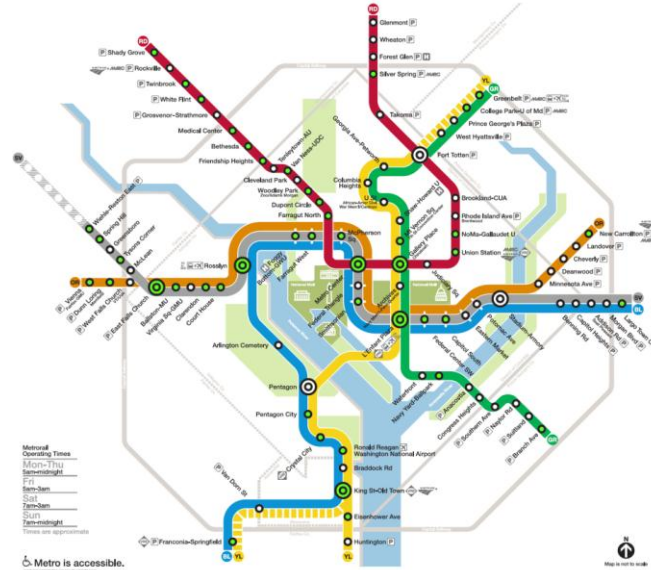
- Document data: Term-frequency vector (matrix) of text documents

Types of Data Sets: (2) Graphs and Networks

- ❑ Transportation network
- ❑ World Wide Web



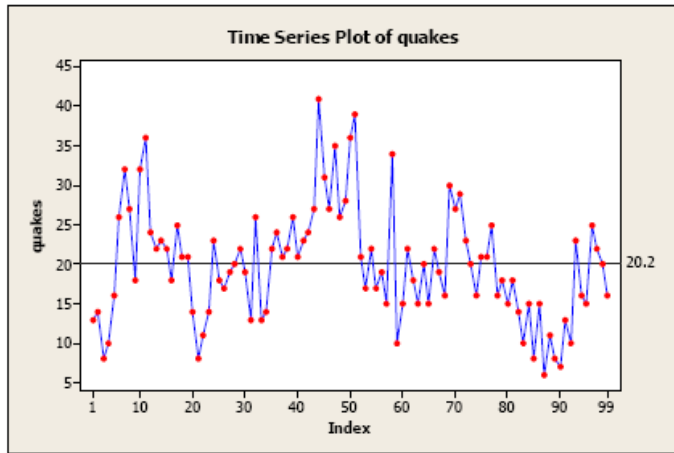
- ❑ Molecular Structures
- ❑ Social or information networks



Types of Data Sets: (3) Ordered Data

□ Video data: sequence of images

□ Temporal data: time-series



□ Sequential Data: transaction sequences

□ Genetic sequence data

	Start
Human	GTTTGGAGG --- ATGTTCAACAAATGCTCCTTTCATTCCCTATTTACAGACCTGCCGCA
Chimpanzee	GTTTGGAGG --- ATGTTCAATAAATGCTGCTTTCACCTCCCTATTTACAGACCTGCCGCA
Macaque	GTTTGGAGG --- ATGTTCAATAAATGCTCCTTTCATTCCCTATTTACAAACTTGCCGCA
Human	GACAATTCGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee	GACAATTCGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque	GACAATTCGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human	GATCTGGAGACTAA - CTC TGAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Chimpanzee	GATCTGGAGACTAAACCTGTAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Macaque	TATCTGGAGACTAAACCTGTAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Human	CAGAATACGATTTAGCAAATTAATCTCTTAAGATATATTTTACATTTCTATATTTCTCCTA
Chimpanzee	CAGAATACGATTTAGCAAATTAATCTCTTAAGATATATTTTACATTTCTATATTTCTCCTA
Macaque	CAGAATATGATTTAGCAAATTAATCTCTTAAGATATATTTTGCACTTCTATATTTCTCCTA
Human	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTTCATAAAGCCAGGTATACA --- TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGCCGATATGTCACCTTTCATAAAGCCAGGTATACA --- TTATG
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTTCACAAAGCCAGGTATATATACATTACG
Human	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTGTCCAAAAATTTAAATTTCT
Chimpanzee	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTGTCCAAAAATTTAAATTTCT
Macaque	GACAGGTAAGTAAAAA - CATATTATTTATTCTAGBTTTTGTCCAAAGATTTTAAATTTCT
Human	AAC TGT TGC CGT GTG T TGG TAA --- TGT AAAAC AAAC TCAGTACA
Chimpanzee	AAC TGT TGC CGT GTG T TGG TAA --- TGT AAAAC AAAC TCAGTACA
Macaque	AAC TGT TGT GC ATGTG T TGG TAA --- CBT AAAAC AAAA TTCAGTACG

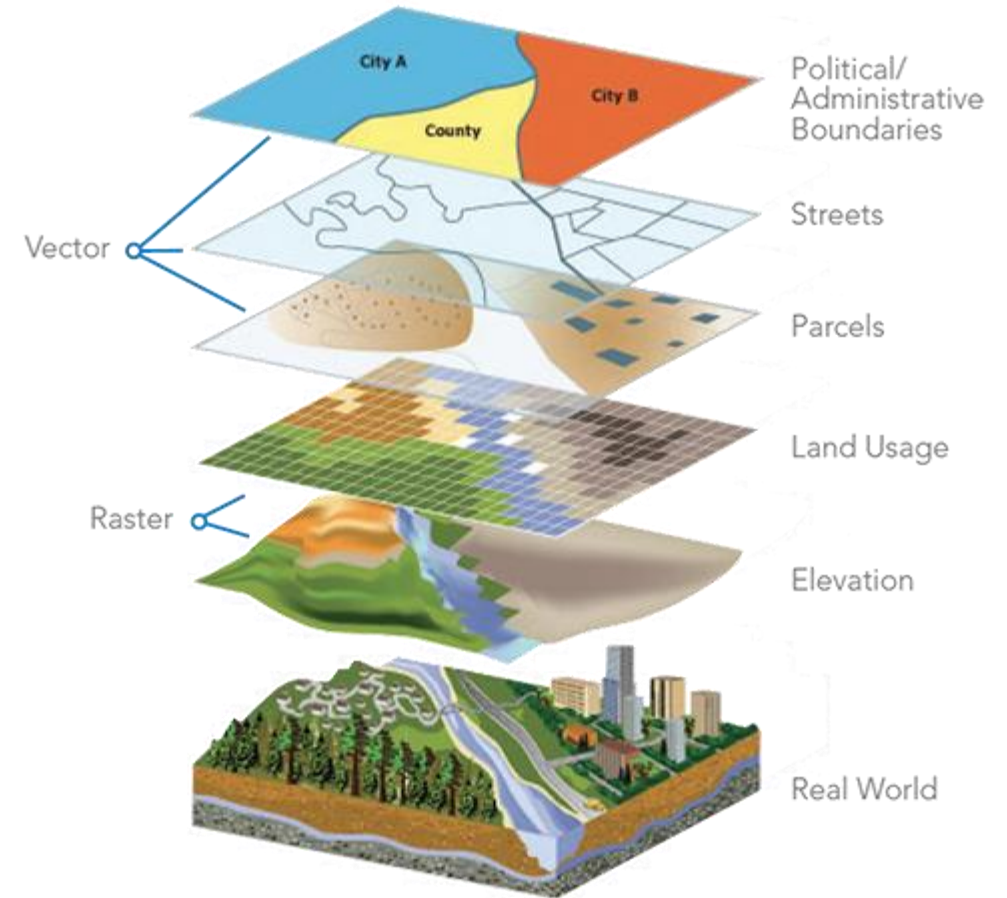
Types of Data Sets: (4) Spatial, image and multimedia Data

- Spatial data: maps



- Image data:

- Video data:



Important Characteristics of Structured Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

Data Objects

- ❑ Data sets are made up of data objects
- ❑ A **data object** represents an entity
- ❑ Examples:
 - ❑ sales database: customers, store items, sales
 - ❑ medical database: patients, treatments
 - ❑ university database: students, professors, courses
- ❑ Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*
- ❑ Data objects are described by **attributes**
- ❑ Database rows → data objects; columns → attributes

Attributes

- ❑ **Attribute (or dimensions, features, variables)**
 - ❑ A data field, representing a characteristic or feature of a data object.
 - ❑ *E.g., customer_ID, name, address*
- ❑ **Types:**
 - ❑ Nominal (e.g., red, blue)
 - ❑ Binary (e.g., {true, false})
 - ❑ Ordinal (e.g., {freshman, sophomore, junior, senior})
 - ❑ Numeric: quantitative
 - ❑ Interval-scaled: 100°C is interval scales
 - ❑ Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50 °K
- ❑ Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- ❑ Q2: What about eye color? Or color in the color spectrum of physics?

Attribute Types

- ❑ **Nominal:** categories, states, or “names of things”
 - ❑ *Hair_color = {auburn, black, blond, brown, grey, red, white}*
 - ❑ marital status, occupation, ID numbers, zip codes
- ❑ **Binary**
 - ❑ Nominal attribute with only 2 states (0 and 1)
 - ❑ Symmetric binary: both outcomes equally important
 - ❑ e.g., gender
 - ❑ Asymmetric binary: outcomes not equally important.
 - ❑ e.g., medical test (positive vs. negative)
 - ❑ Convention: assign 1 to most important outcome (e.g., HIV positive)
- ❑ **Ordinal**
 - ❑ Values have a meaningful order (ranking) but magnitude between successive values is not known
 - ❑ *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)

- **Interval**

- Measured on a scale of **equal-sized units**

- Values have order

- E.g., *temperature in C° or F°, calendar dates*

- No true zero-point

- **Ratio**

- Inherent **zero-point**

- We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).

- e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

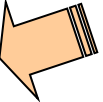
□ Discrete Attribute

- Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

□ Continuous Attribute

- Has real numbers as attribute values
 - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

Chapter 2. Getting to Know Your Data

- ❑ Data Objects and Attribute Types
- ❑ Basic Statistical Descriptions of Data 
- ❑ Data Visualization
- ❑ Measuring Data Similarity and Dissimilarity
- ❑ Summary