Data Mining and Web Algorithms

Course Code: 15B22CI621

Credits: 4 [3+ 1]

Vision and Mission of CSE/IT

VISION

To be a centre of excellence for providing quality education and carrying out cutting edge research to develop future leaders in all aspects of computing, IT and entrepreneurship.

MISSION

MISSION 1: To offer academic programme with state of art curriculum having flexibility for accommodating the latest developments in the areas of computer science and IT

MISSION 2: To conduct research and development activities in contemporary and emerging areas of computer science & engineering and IT.

MISSION 3: To inculcate IT & entrepreneurial skills to produce professionals capable of providing socially relevant and sustainable solutions.

Course Outcome-Cognitive Level

	COURSE OUTCOMES	COGNITIVE LEVELS
CO1	Understand the basics of data mining and pre-processing of data.	Understand Level (Level 2)
CO2	Analyze the transactional data for finding frequent and interesting patterns using association rule mining techniques like Apriori and FP-Growth**.**	Analyse Level (Level 4)
CO3	Apply a wide range of classification techniques like Naïve-bayes, decision tree, and KNN for the numerous application including fraud detection, target marketing, medical diagnosis, etc.	Apply Level (Level 3)
CO4	Cluster the similar/dissimilar objects using different methods like partitioning, hierarchical and density based clustering.	Create Level (Level 6)
CO5	Analyze the link structure of web using page rank and HITS algorithms.	Analyse Level (Level 4)
CO6	Develop recommendation system using collaborative filtering techniques	Create Level (Level 6)

Syllabus & Evaluation

- Course Description
- T1, T2 and T3
- TA :
  - Tutorials /Quiz/Assignments
  - Assignments in PBL Mode
  - Class Performance

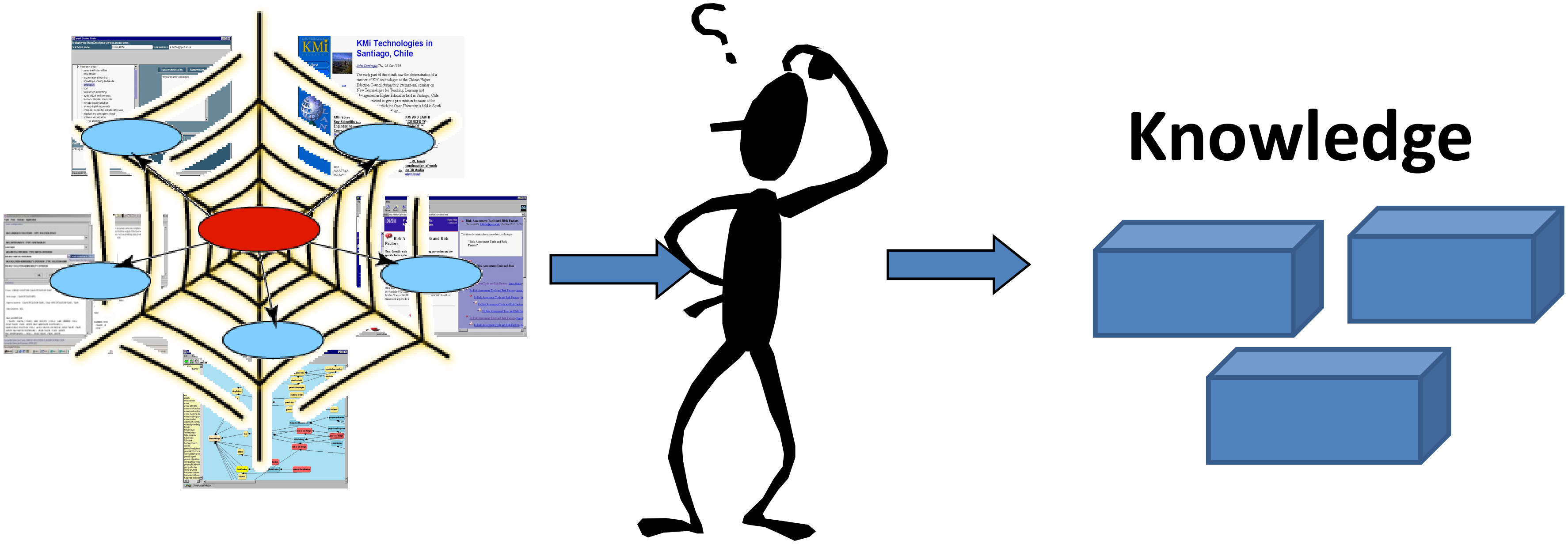
Why Data mining ...

- Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
- Which types of transactions are likely to be fraudulent given the demographics and transactional history of a particular customer?
- If I raise the price of my product by Rs. 2, what is the effect on my ROI(Return on Investment ?
- If I offer only 2,500 airline miles as an incentive to purchase rather than 5,000, how many lost responses will result?
- If I emphasize ease-of-use of the product as opposed to its technical capabilities, what will be the net effect on my revenues?
- Which of my customers are likely to be the most loyal?

Data Mining helps to extract such information

Why Web Algorithms …

WWW



Knowledge

Discovering Knowledge from and about WWW

Data Mining and Web Algorithms

- **Data mining**: turn data into knowledge.

- **Web algorithms**: to apply data mining techniques to extract and uncover knowledge from *web documents* and *services*.

What Is Data Mining?

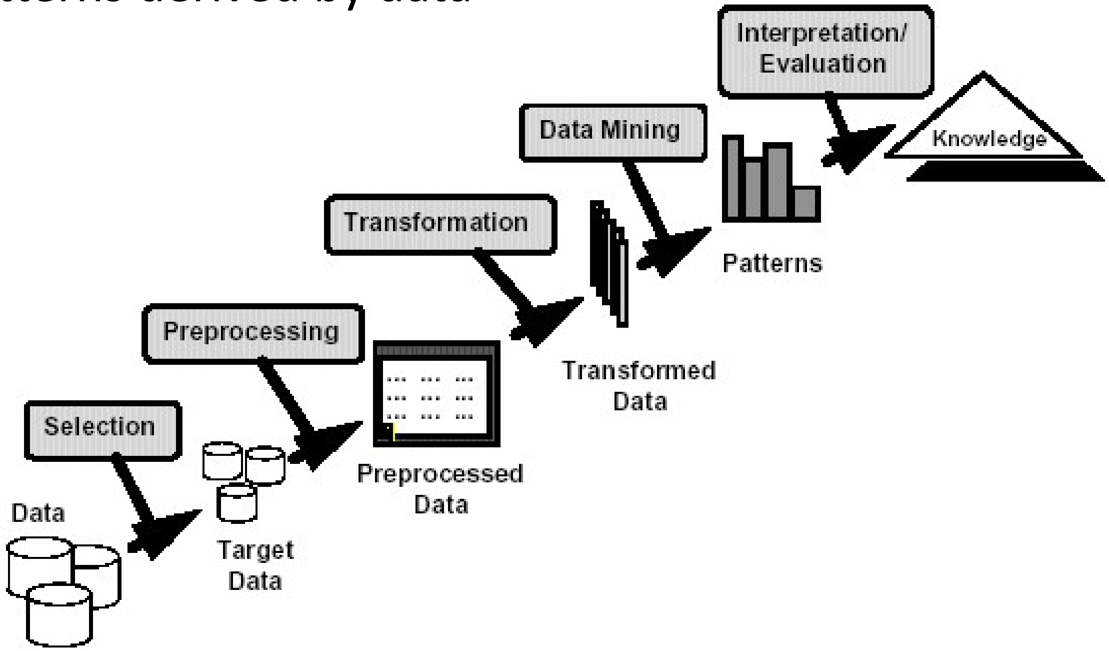
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery in Databases(KDD) Process

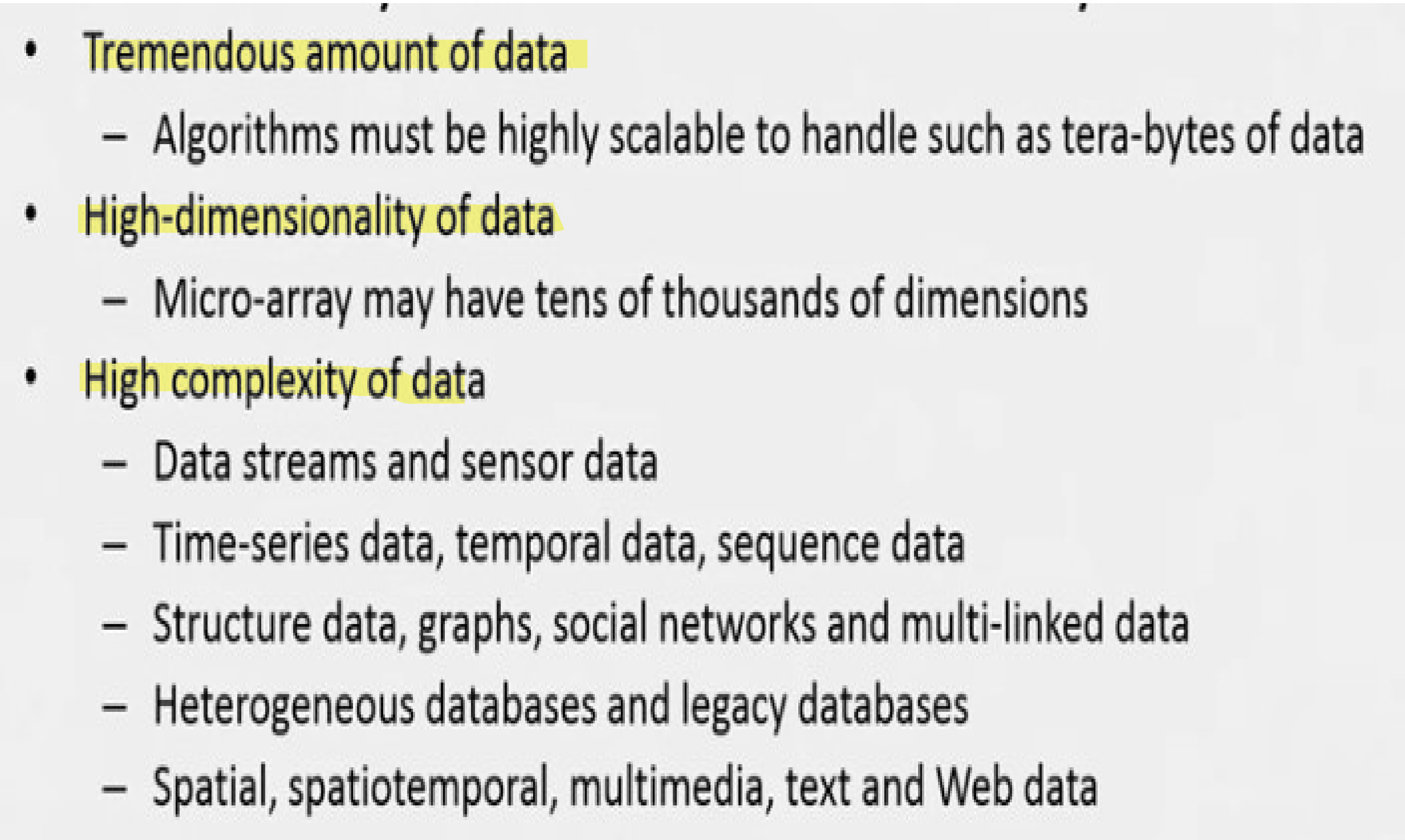
- **KDD** is the overall process of finding useful information from data. It is an integration of multiple technologies for data management such as database management, data warehousing, statistic machine learning, decision support, and others such as visualisation.
- **Data mining(core of KDD process)** is the use of algorithms to extract information and patterns derived by data

Imp Diagram



U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996.
M. De Martino, A. Bertone, R. Albertoni, H. Hauska, U. Demsar, M. Dunkars. Technical Report of Data Mining, INVISIP IST-2000-29640, Information Visualisation for Site Planning, WP No2: Technology Analysis, D2.2, 28.2.2002

Why Not traditional Analysis tool

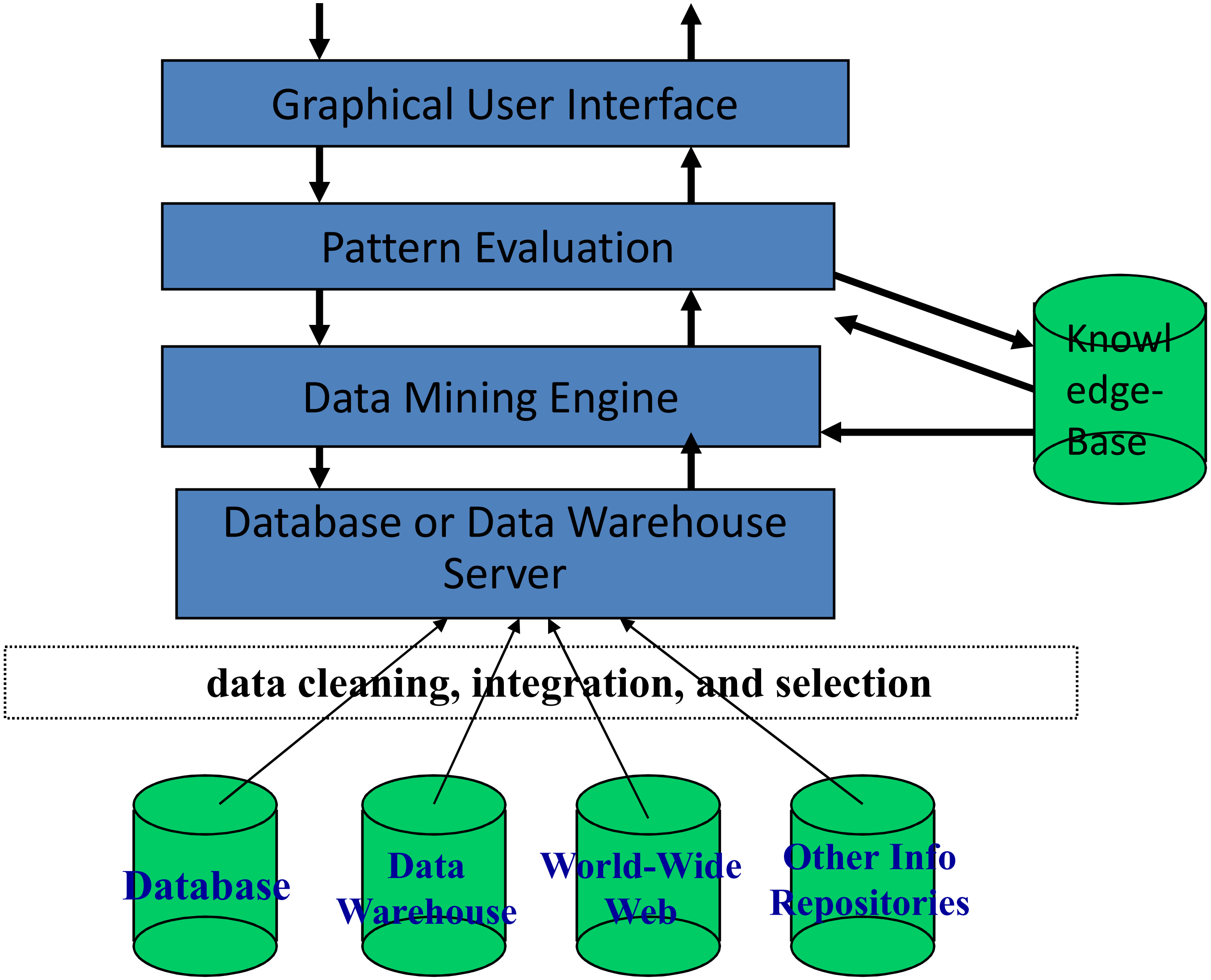


- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data

Why Not traditional Analysis tool: Very Large Data Bases

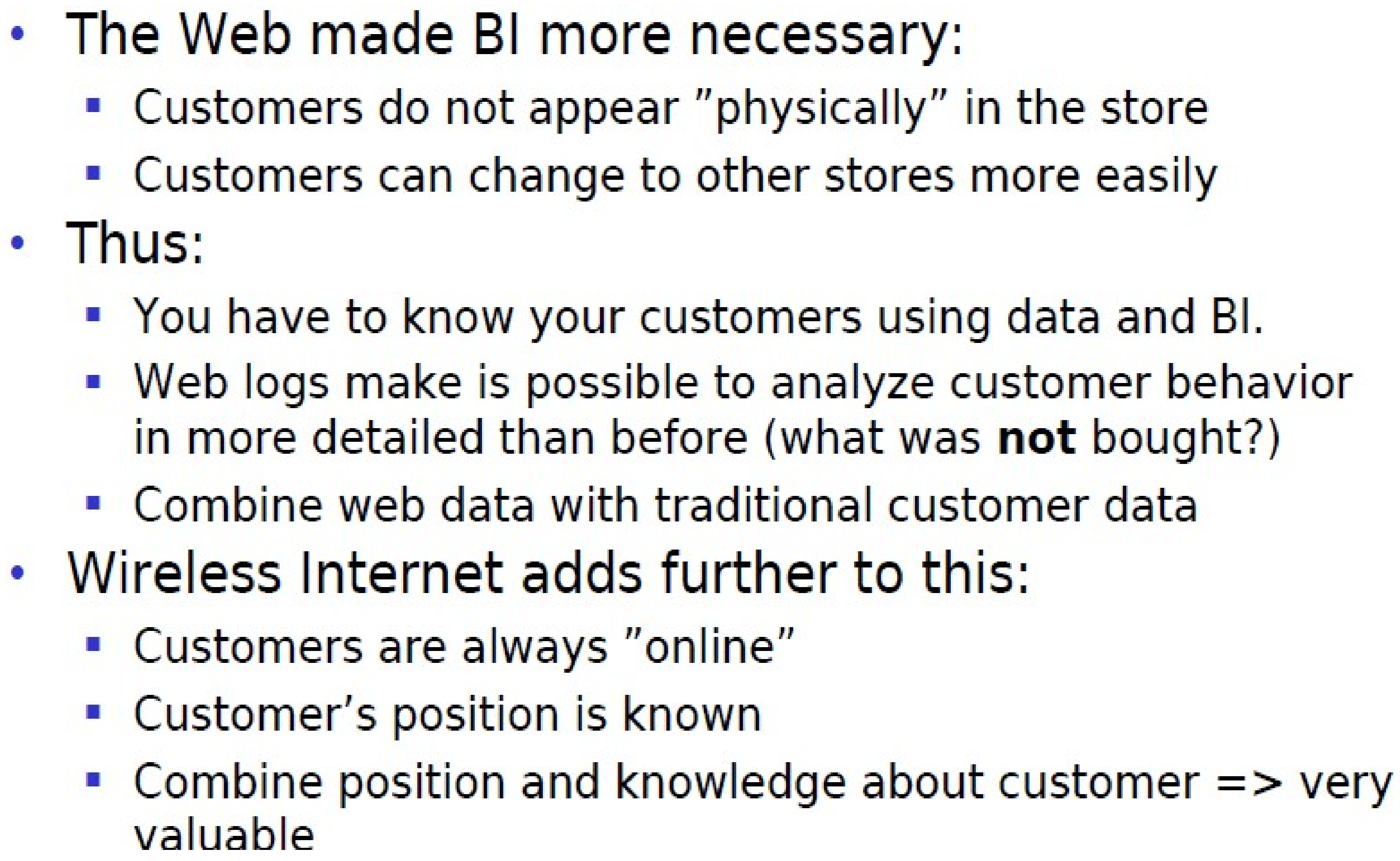
• Terabytes -- 10^12 bytes:	Walmart -- 24 Terabytes
• Petabytes -- 10^15 bytes:	Geographic Information Systems
• Exabytes -- 10^18 bytes:	National Medical Records
• Zettabytes -- 10^21 bytes:	Weather images
• Zottabytes -- 10^24 bytes:	Intelligence Agency Videos

Architecture: Typical Data Mining System



https://www.javatpoint.com/data-mining-architecture

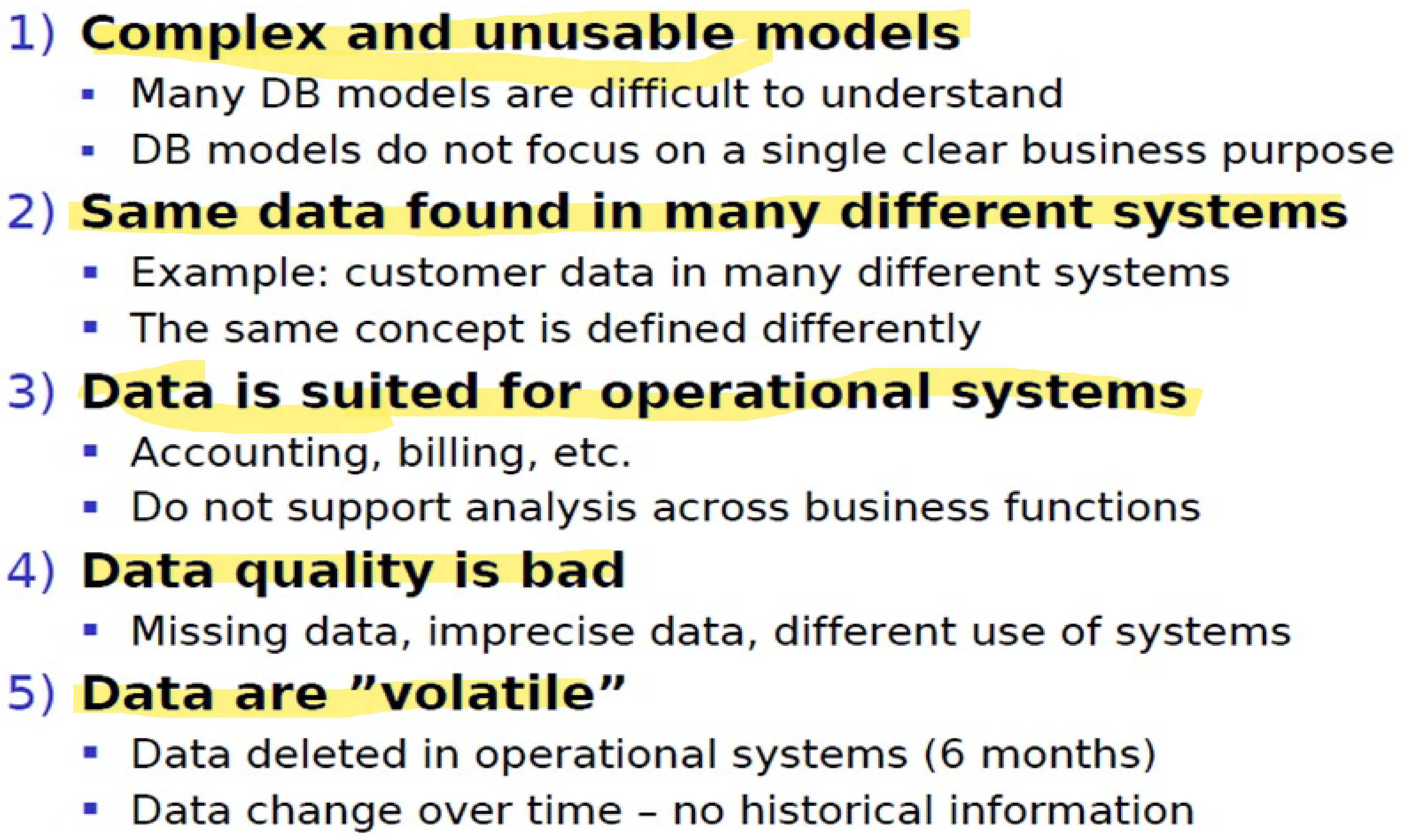
Why Data Mining in Business Intelligence(BI)



- The Web made BI more necessary:
  - Customers do not appear ”physically” in the store
  - Customers can change to other stores more easily
- Thus:
  - You have to know your customers using data and BI.
  - Web logs make is possible to analyze customer behavior in more detailed than before (what was **not** bought?)
  - Combine web data with traditional customer data
- Wireless Internet adds further to this:
  - Customers are always ”online”
  - Customer’s position is known
  - Combine position and knowledge about customer => very valuable

Why Data Mining in Business Intelligence(BI)

BI problems:



1) **Complex and unusable models**
   - Many DB models are difficult to understand
   - DB models do not focus on a single clear business purpose
2) **Same data found in many different systems**
   - Example: customer data in many different systems
   - The same concept is defined differently
3) **Data is suited for operational systems**
   - Accounting, billing, etc.
   - Do not support analysis across business functions
4) **Data quality is bad**
   - Missing data, imprecise data, different use of systems
5) **Data are ”volatile”**
   - Data deleted in operational systems (6 months)
   - Data change over time – no historical information

Business Intelligence(BI) : Solution

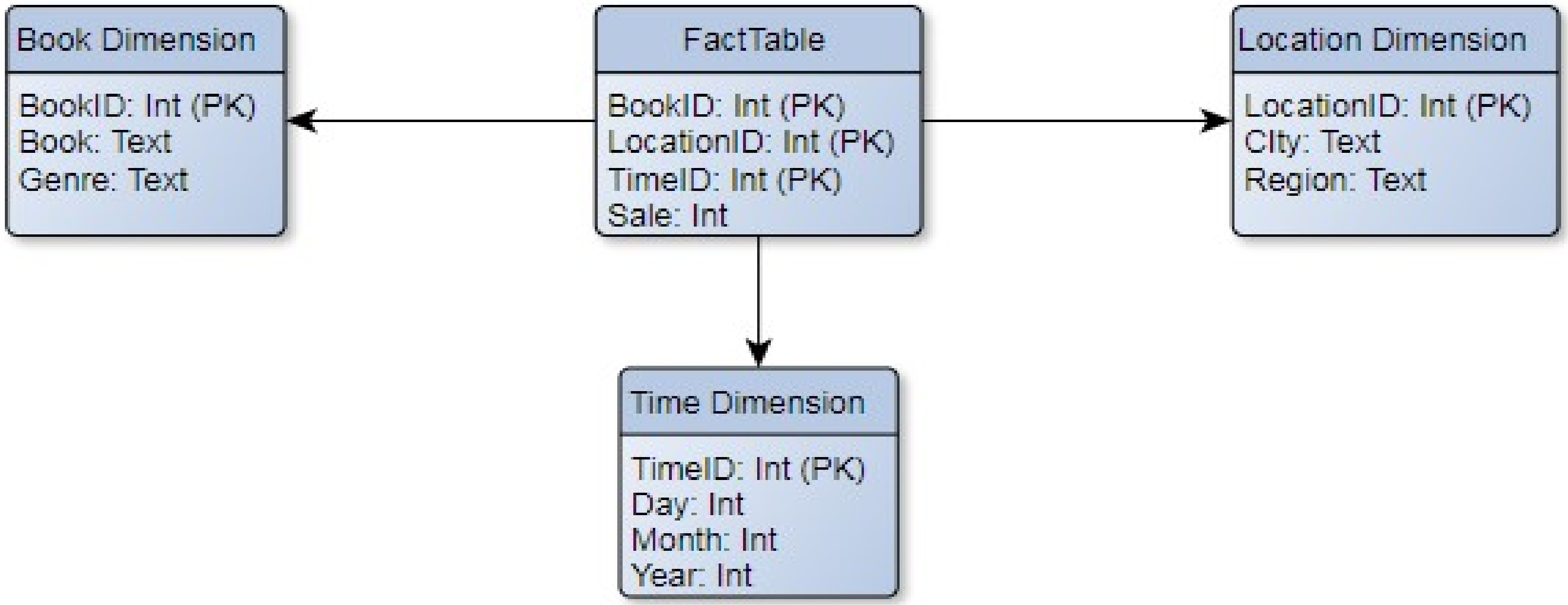
- A new analysis environment (includes Data Mining techniques) with data warehouse at the core, where the data is
  - Integrated
  - Subject oriented
  - Non-volatile
  - Time variant

Barry Delvin, IBM consultant, "A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."
W.H. Inmon, " a data warehouse is a subject oriented, integrated, time-varying, and non-volatile collection of data that is used primarily in organizational decision making."

A data mart is one piece of a data warehouse where all the information is related to specific business area. Therefore it is considered a subset of all the data stored in that particular database, since all data marts together create a data warehouse.

Design of DW

- Modeling of DW is done using dimensional model such as start schema , snowflake etc. Following is an example of Start Schema.



Data Warehouse example

Source: https://chrthomsen.github.io/pygrametl/doc/quickstart/beginner.html

OLTP vs Data Warehouse

- OLTP Online Transaction Processing
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User
  - SQL type query language

- Warehouse (DSS) Decision Support System
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)
  - SQL, OLAP(Online Analytical Processing) tools, Data mining Techniques, Reporting Tools

➢**OLTP Systems are used to *"run"* a business**

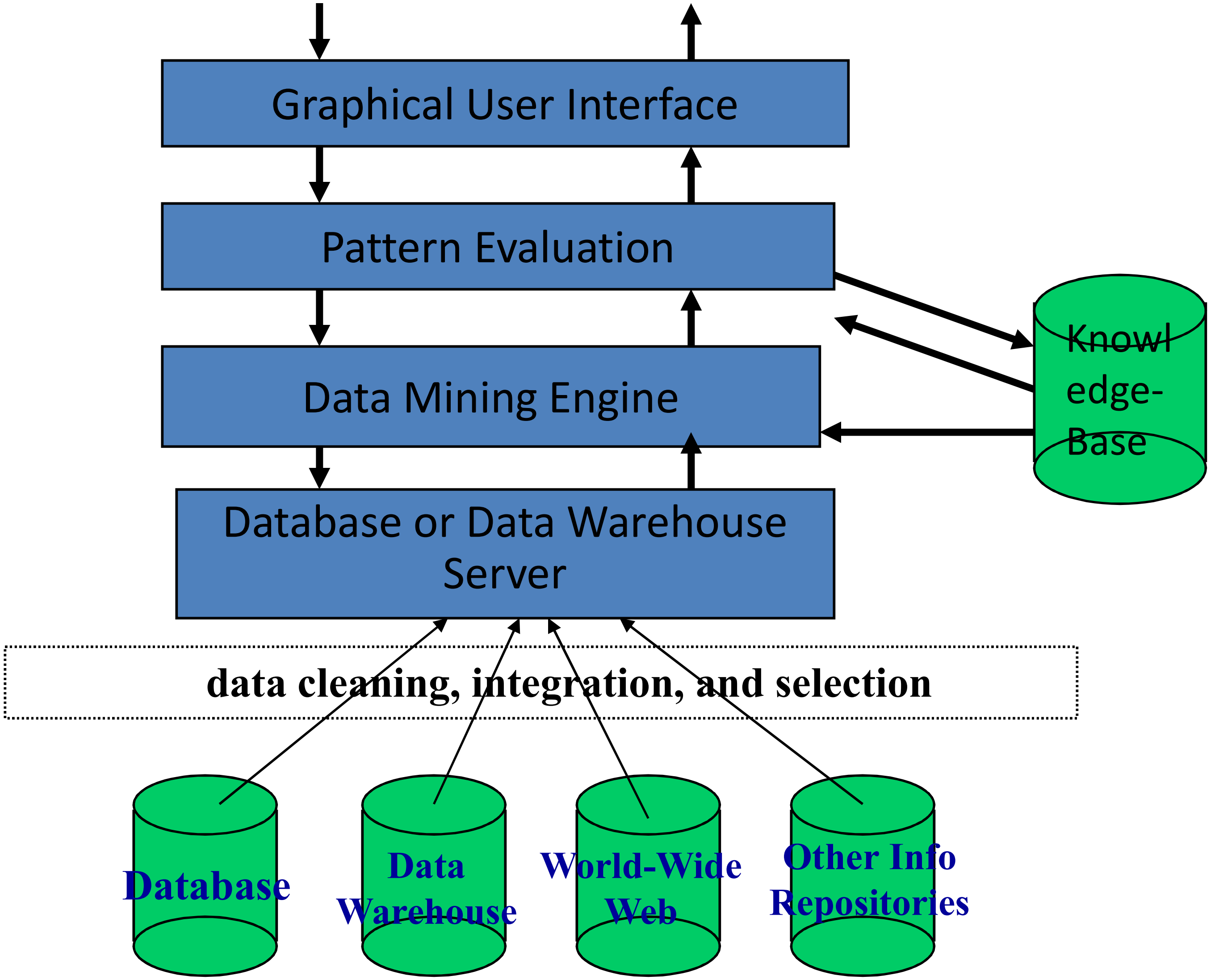
➢**The Data Warehouse helps to *"optimize"* the business**

Data Warehouses in market today

- Teradata
- Amazon Web services(AWS)
- Cloudera
- Oracle
- MarkLogic

- **Data mining: turn data into knowledge from data warehouse .**

Revisit : Architecture of Typical Data Mining System



https://www.javatpoint.com/data-mining-architecture

Data mining engine

- ***Prediction Methods*** :use some variables to predict unknown or future values of other variables.
  - **Classification & Prediction**

  (Pattern Recognition is type of classification)
  - **Regression (for continuous variable)**

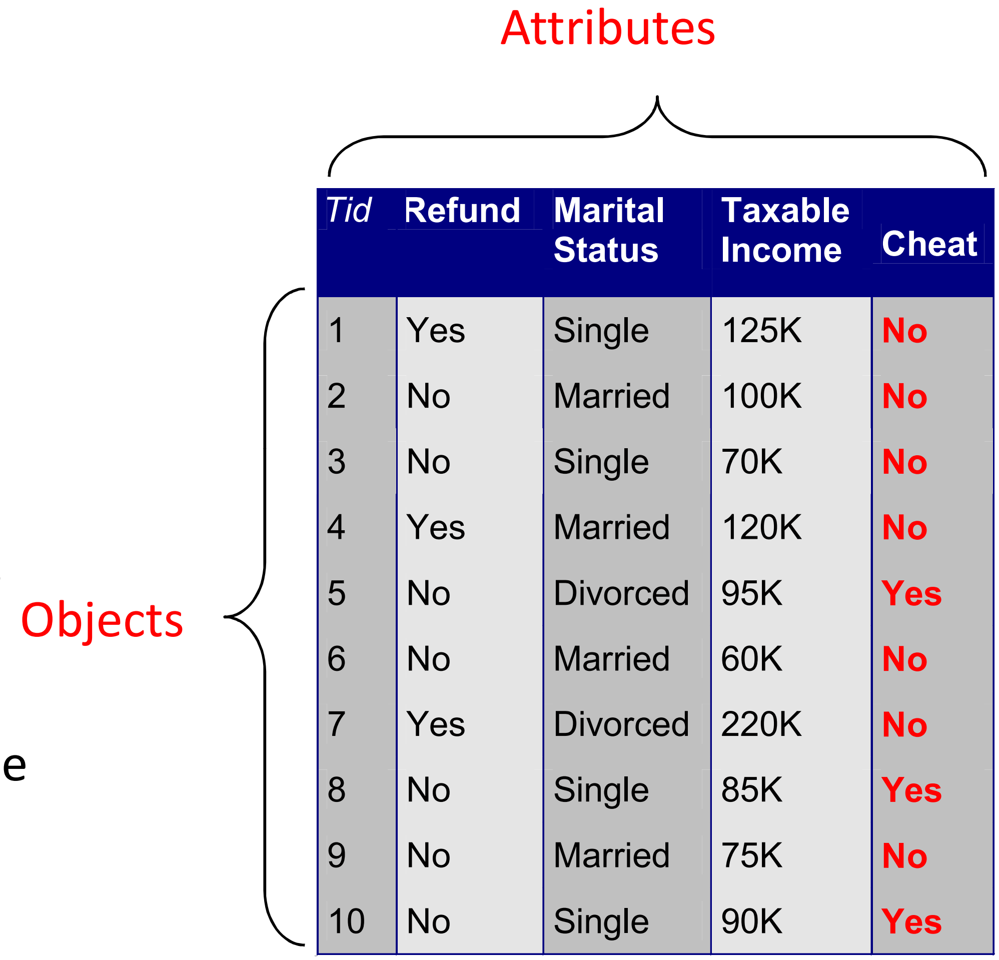
- ***Description Methods***: find human-interpretable patterns that describe the data.
  - **Clustering**
  - **Association Rule Discovery**
  - **Summarization(Characterization)**

What is Data Set?

- 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 Attributes

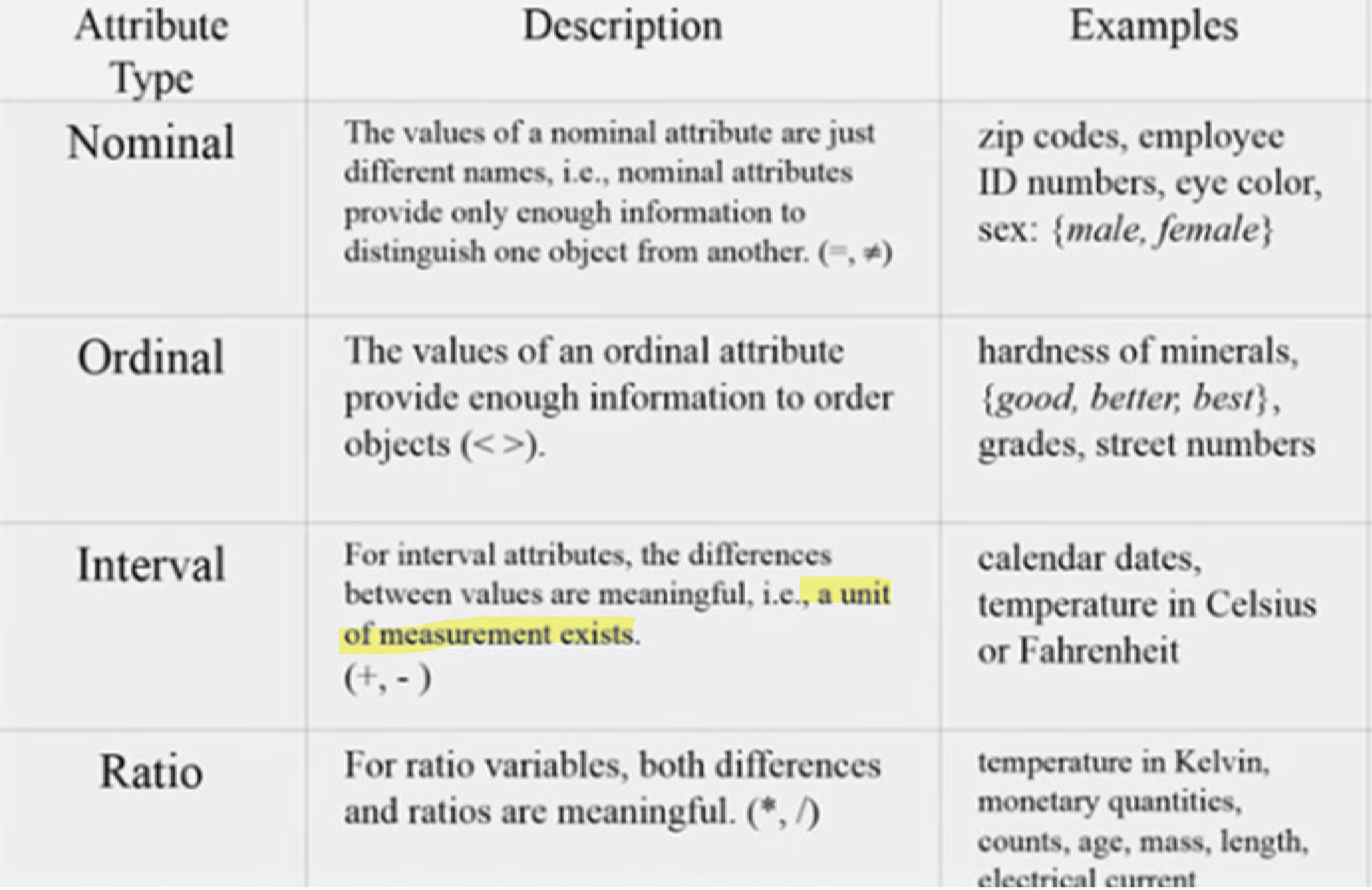
- There are different types of attributes
  - Nominal (Categorical/Qualitative)
    - Examples: ID numbers, eye color, zip codes
  - Ordinal (Categorical/Qualitative)
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval(Numerical/Quantitative)
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio (Numerical/Quantitative)
    - Examples: 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: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Summary of Attribute Values



Attribute Type	Description	Examples
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*}
Ordinal	The values of an ordinal attribute provide enough information to order objects (< >).	hardness of minerals, {*good, better, best*}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current

Quantitative Attributes can be integer-valued(Discrete) or continuous.

Source [2]

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes (two values 0/1) are a special case of discrete attributes .

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.
  -

- Asymmetric Attribute (may be discrete/continuous)
  - Only presence (a no-zero attribute value is regarded as important)
  - Example:
    - A data set of students has an attribute who has value 1 if he took a particular course else 0.
    - As student take only a small fraction of all available courses. It is more important to process only non-zero values.

Distance Measures for Numerical attributes

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: ***Minkowski distance***:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

where *i* = (x_{i1}, x_{i2}, …, x_{ip}) and *j* = (x_{j1}, x_{j2}, …, x_{jp}) are two *p*-dimensional data objects, and *q* ≥ 1 is a positive integer.

- Such a distance is also called L_q *norm in* literature
- If *q = 1, d* is **Manhattan distance (City Block Distance)**

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

Distance Measures for Numerical attributes

- *If q = 2, d* is **Euclidean distance**:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$

- Manhattan distance is also known as L_1 *norm and Euclidean distance as* L_2 *norm.*

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 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

Record : 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

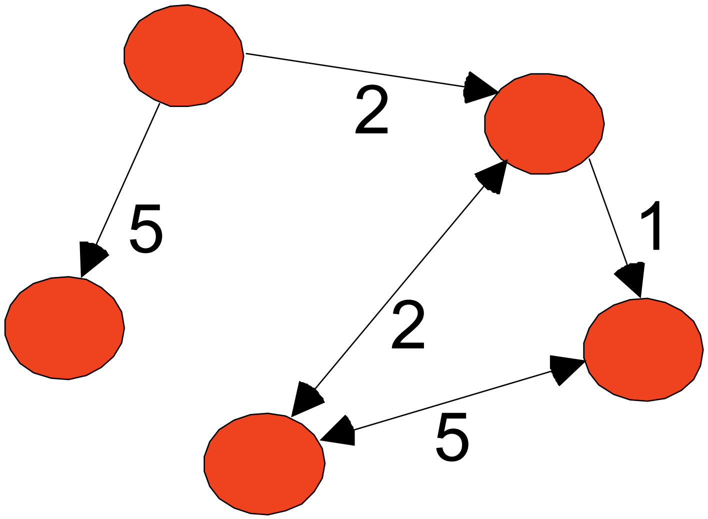
Record : Transaction Data

- 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: Web Data

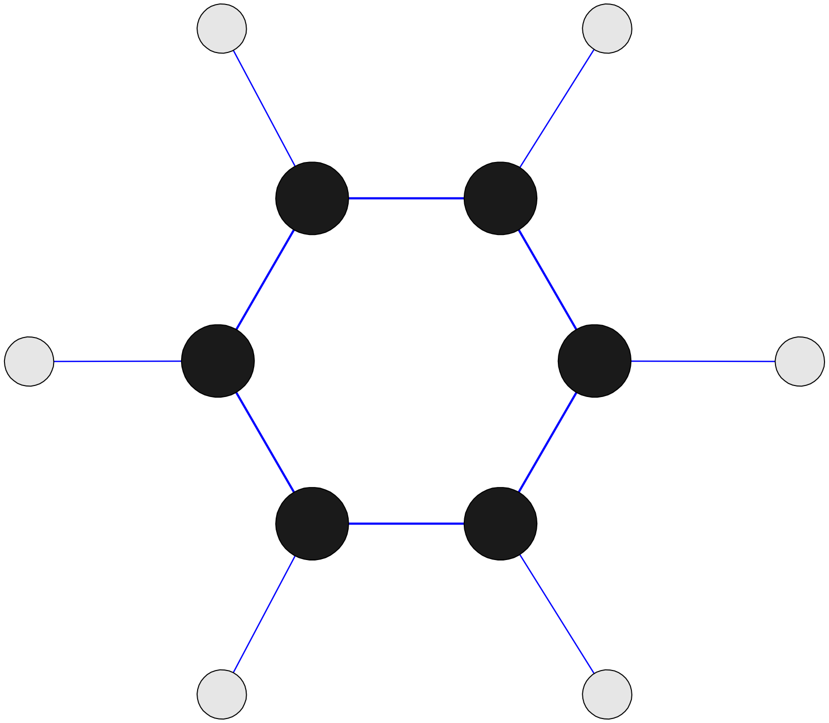
- Objects are represented as nodes and arcs are represented as relationships b/w nodes
- Links conveys information about relevance of a web page to a query.
- Examples: Generic graph and HTML Links



```
<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
```

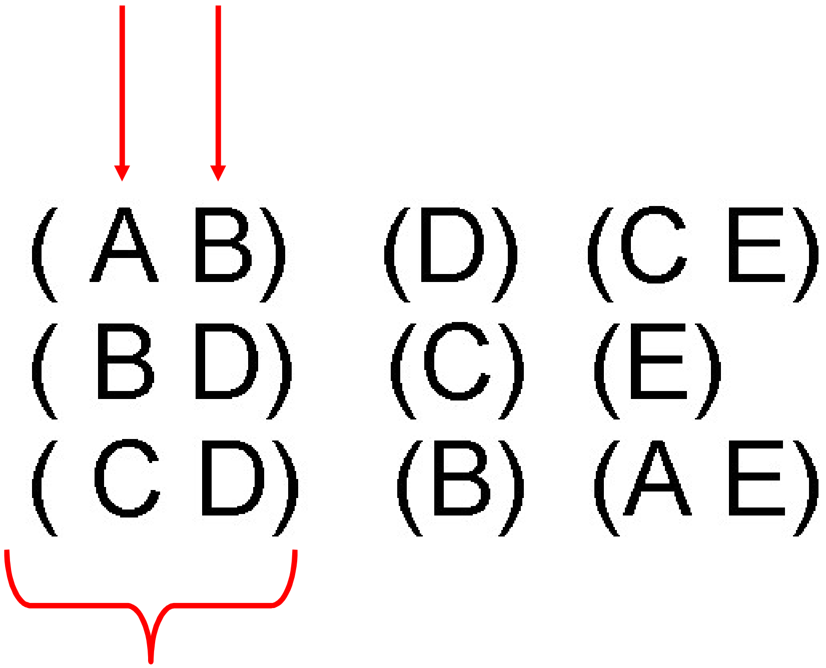
Graph : Chemical Data

- Objects themselves have substructure that can be represented as graph.
  - Nodes(Atoms) and links(chemical bonds)
- E.g. Benzene Molecule: C_6H_6
- task is to find the substructure of chemical compounds and determine whether it is associated with some physical properties such as melting point /heat of formation



Ordered Data : Sequential Data

- Time is attached with each transaction
- Sequences of transactions



Items/Events

(A B) (D) (C E)

(B D) (C) (E)

(C D) (B) (A E)

An element of the sequence

Ordered Data : Sequence Data

- A sequence of individual entities such as words/letters. Eg. Genomic sequence data
- No Timestamp
- But entities follow certain order.
- Mining task is to find similarity measures among genes or protein structure

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

Spatio-temporal Data

- Spatio-Temporal Data

1. Average Monthly Temperature of land and ocean
2. Identify spatiotemporal cascade patterns from crime event datasets can help police department to understand crime generators in a city, and thus take effective measures to reduce crime events .
3. Typical attributes : Location ,Area, Perimeter, shape, occurrence time, duration, time before the event, time after the event

Publicly Available data sets

- http://cs.joensuu.fi/sipu/datasets/

- https://archive.ics.uci.edu/ml/datasets.html

- http://www.kdnuggets.com/2011/02/free-public-datasets.

- http://www.rdatamining.com/resources/data

References

[1] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2012(Third Edition).

[2] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.

[3] Dunham, Margaret H. Data mining: Introductory and advanced topics. Pearson Education India, 2006.

[4] https://www.javatpoint.com/data-mining-architecture

[5] www.nptel.ac.in