STATISTICAL METHODS IN DATA MINING

DR. ALPER VAHAPLAR

Contents

- Database, data warehouse and OLAP
- Data mining process, CRISP-DM
- Data mining process, data preparation
- Unsupervised learning, clustering, hierarchical clustering
- k-means, density based clustering
- Supervised learning, classification methods
- k-nearest neighbor method
- Decision tree algorithms, CART, C4,5, CHAID, QUEST
- Neural networks
- Association rules
- Application of association rules
- Model evaluation
- Application of data mining,
- Presentation of student projects

Knowledge Discovery

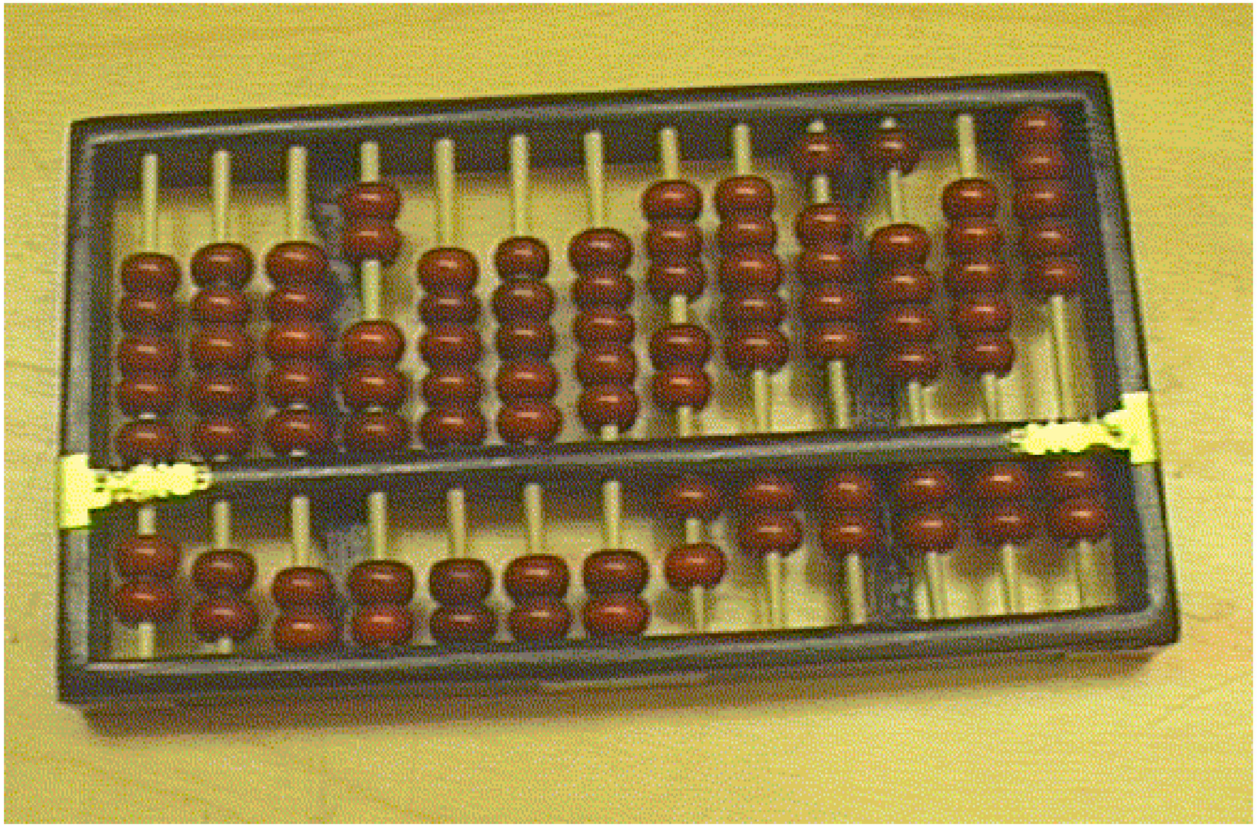
Motivation

Knowledge Discovery

- ✓ "*We are drowning in information, but starved for knowledge.*" (John Naisbitt)
- ✓ Lots of data is being collected and warehoused
  - ◦ NASA Earth observation satellites generate a terabyte (10^9 bytes) of data *every day*
    - ◦ Web data, e-commerce
    - ◦ purchases at department/grocery stores
    - ◦ Bank/Credit Card transactions
- ✓ Improving technology
  - ◦ Computers have become cheaper and more powerful

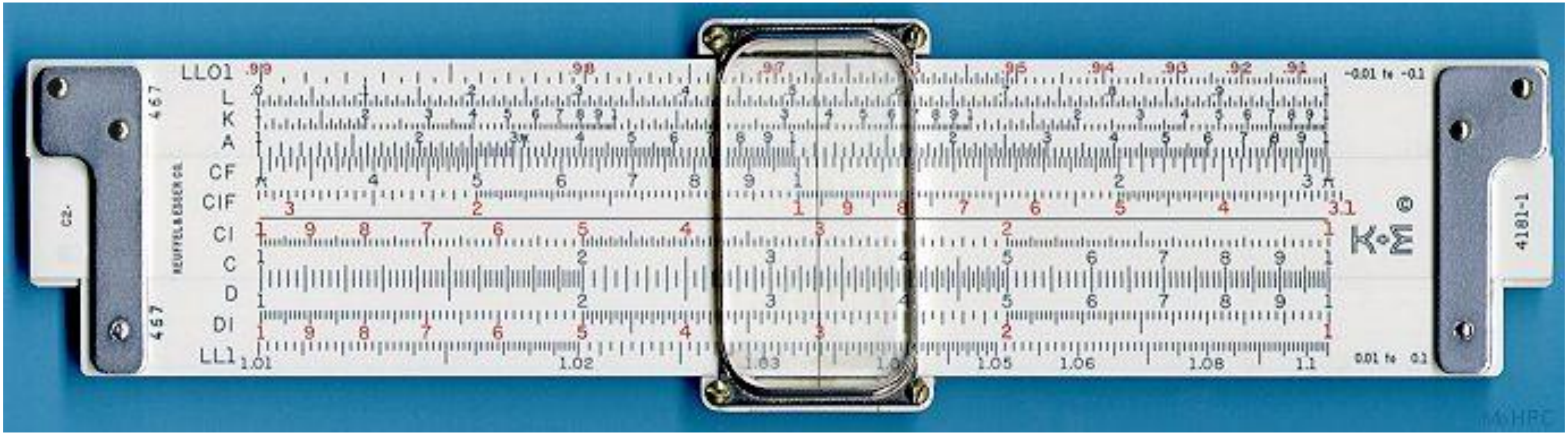
History of Computers

- ✓2600 (BC) – Abacus
  - Simple addition and subtraction



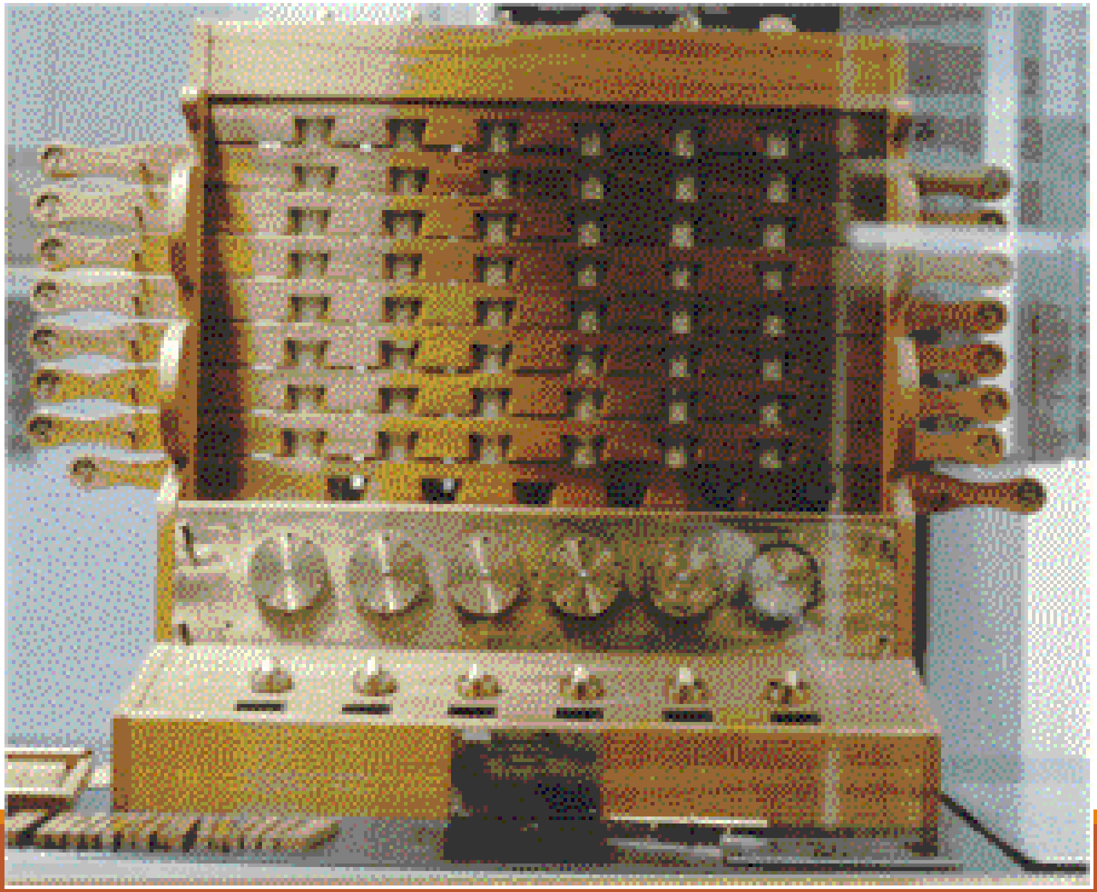
History of Computers

- ✓ 1621– Slide Rule
  - ◦ Addition and subtraction to a constant.



History of Computers

- ✓ 1623 – Calculating Clock
  - ◦ First *gear-driven* calculating machine



History of Computers

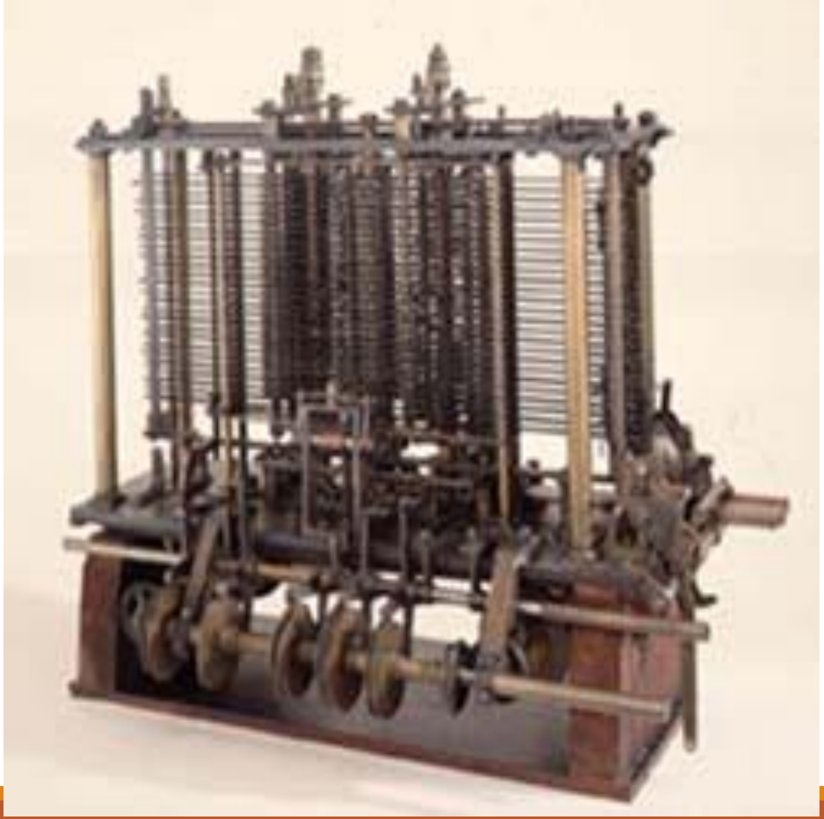
- ✓ *1642 – Pascaline Calculator*
  - Addition with *“carry”*, subtraction with *“borrow”*

History of Computers

- ✓ 1671 – Leibniz Wheel
  - ◦ Multiplication, Division, Square Root operations

History of Computers

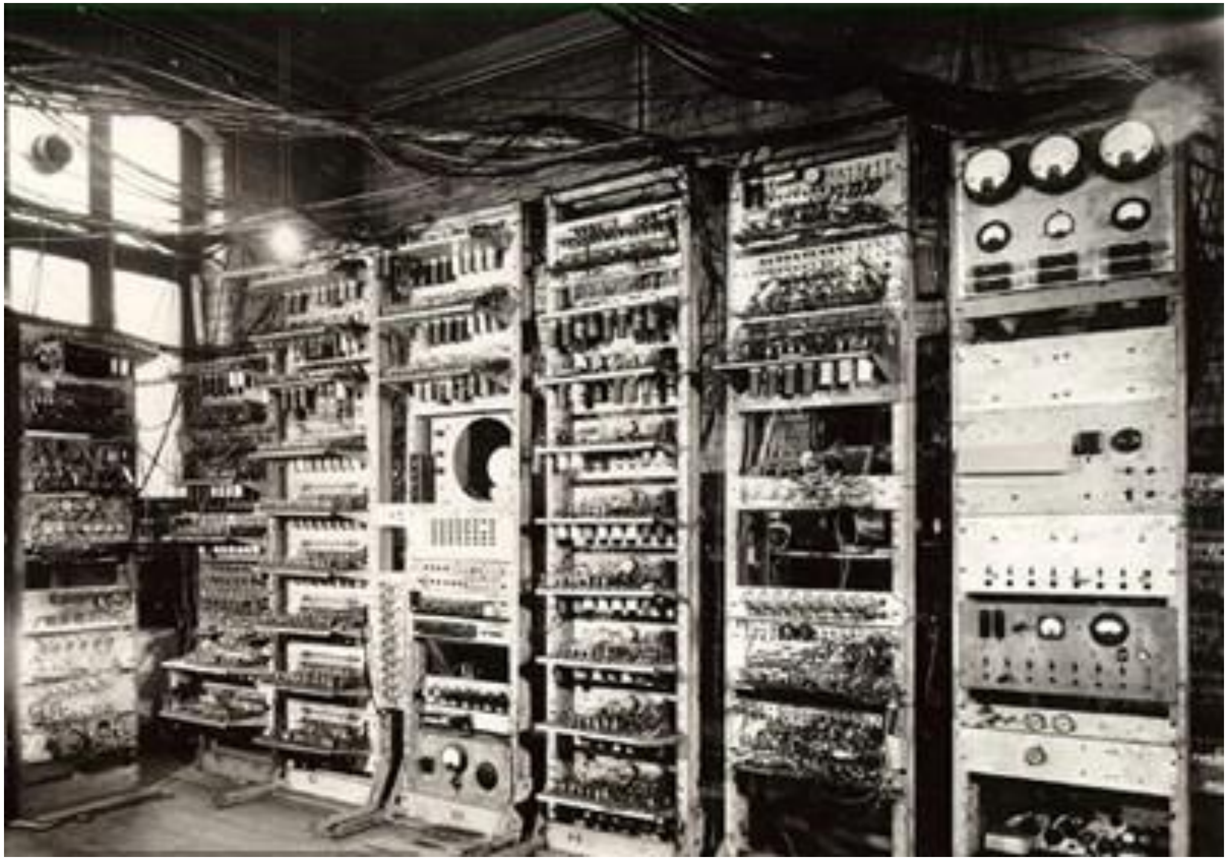
- ✓1801 – Weaving Loom, 1830 – Difference Engine
  - Punched Cards



History of Computers

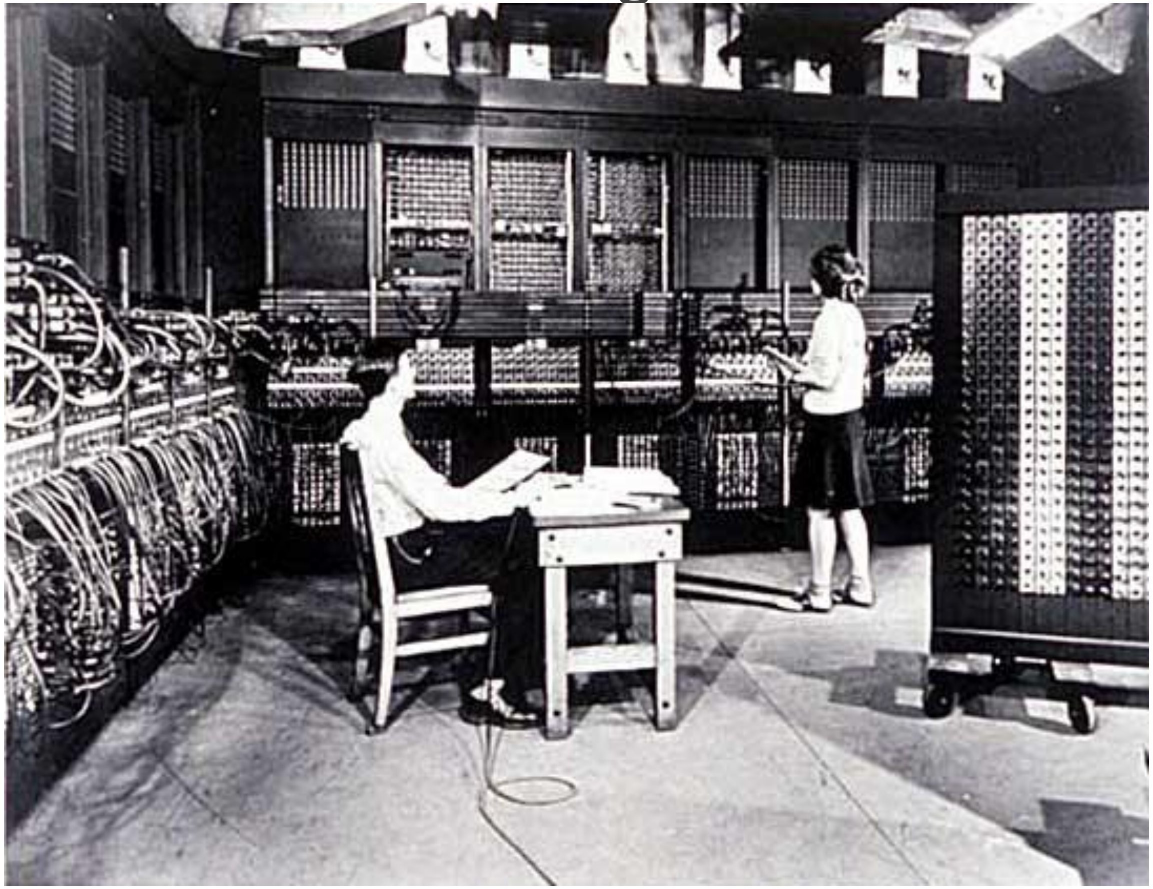
- ✓ 1937 – Mark-1
  - First digital computer
  - Logarithm
  - Trigonometry
  - Slow
  - 1 multiplication

    5 sec.



History of Computers

- ✓1946 – ENIAC (Electrical Numerical Integrator And Computer)
- ✓Military use
- ✓Can perform
  - 5000 addition
  - 385 multiplication
  - 38 sqr-root
- ✓30 tones
- ✓167 m^2



History of Computers

- ✓1970 – IBM mainframes (3090, 7090, 360, 370)
- ✓1971 – First Microprocessor – 4004 – Intel
- ✓1976 – APPLE - Steve Wozniak and Steve Jobs

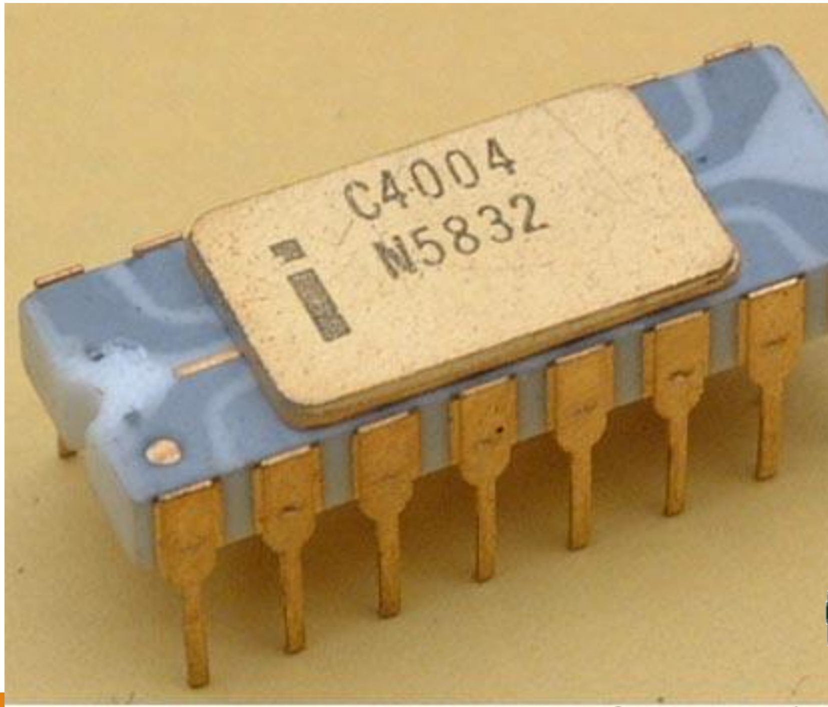
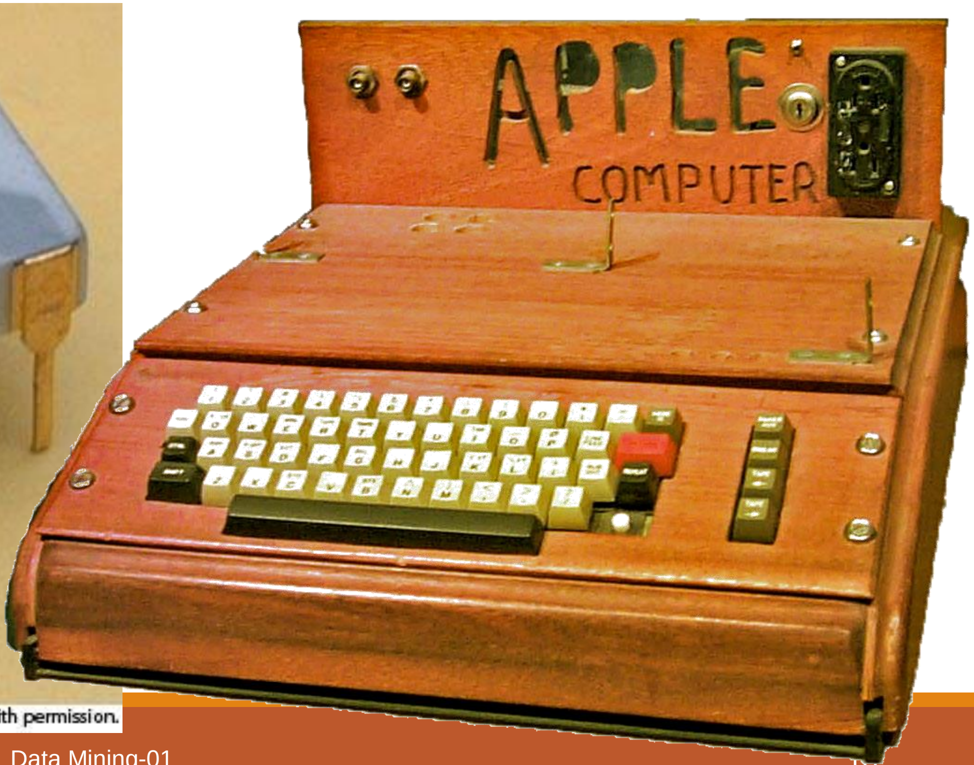


Image courtesy of CPU-Zone.com. Used with permission.



History of Computers

- ✓1980 – IBM PC (Personal Computer)
- ✓Microprocessor
- ✓8086
- ✓80286
- ✓80386
- ✓80486
- ✓80586 (Pentium)
- ✓PII, PIII, PIV…
- ✓Core2 Duo, i5, i7,…

Historical Evaluation

1993…	2020
80386 DX 40 MHz	QuadCore i7 6700 3.4 GHz
512 KByte RAM	16 GByte RAM
100 MByte Hard Disk	4 TByte Hard Disk
14 “ CRT Monitor	22” LED Monitor
32 KByte Display Adapter	NVIDIA GTX 960 2Gb
56 Kbit/sn Modem	Wireless Modem
Floppy Disk Driver	Blu-Ray Disc
MS DOS + Windows 3.1	Windows 10 or Mac OS X Lion

Knowledge Discovery

What is Data Mining?

Data

- ✓ What is ***Data***?
  - ◦ Datum (singular)
  - ◦ Unprocessed (raw) form of information… ☺
  - ◦ the result of a measurement, event or fact.
  - ◦ groups of information that represent the qualitative or quantitative attributes of a variable.
  - ◦ a collection of facts from which conclusions may be drawn; "statistical data”.
  - ◦ known facts, worth to record.
  - ◦ Ex: age, eye color, price, date,
- ✓ Why is it necessary?

Information

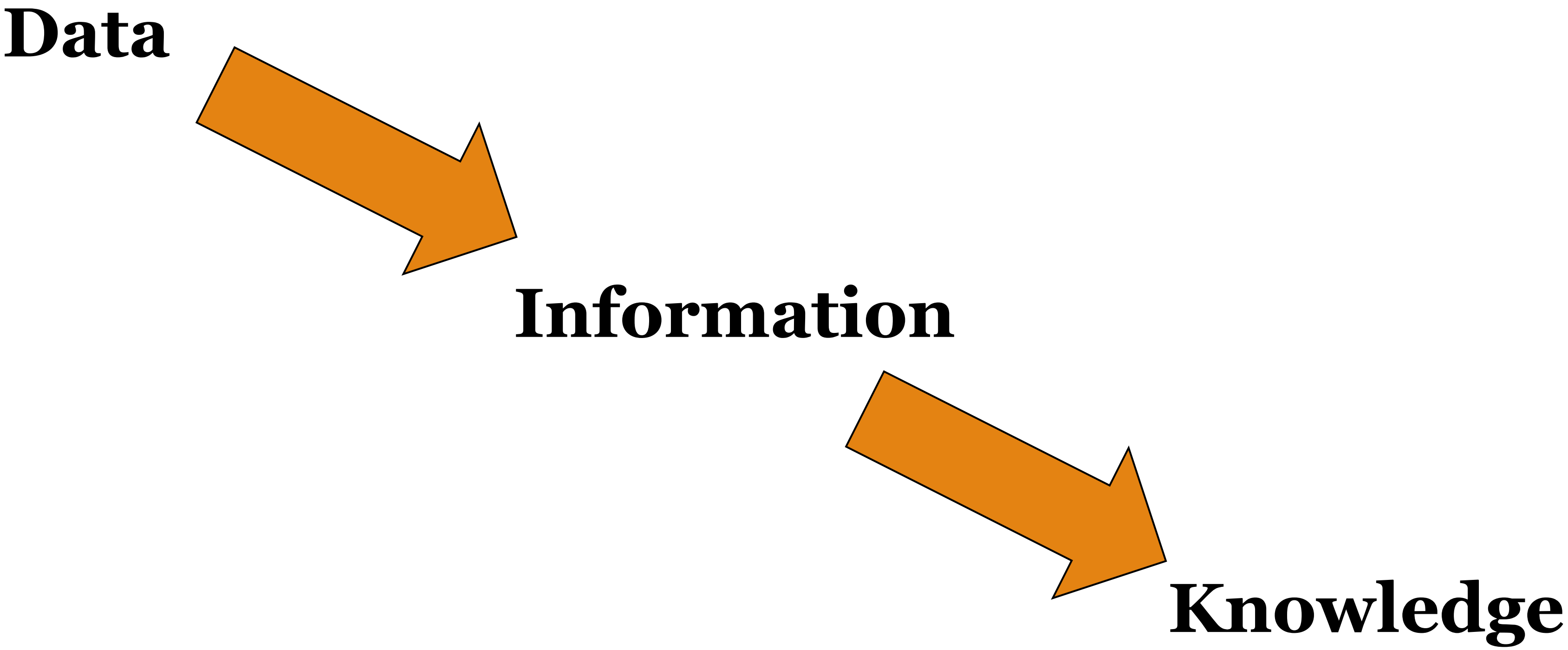
- ✓ What is ***Information***?
  - Processed form of Data … ☺
  - the data that has been processed to be meaningful to the person who receives it.
  - knowledge acquired through study or experience or instruction.
  - Collection of facts that decisions are made on.
  - Statistically analysed data.
  - Ex: increase in the amount of erytrocyte, decrease of sales, etc

Characteristics of “Information”

- ✓ Accurate and Reliable
- ✓ Relevant and Timely
- ✓ Understandable and Transferable
- ✓ “*Expensive*” to collect
- ✓ Provide power and/or advantage

Knowledge

✓ Expertise and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject.

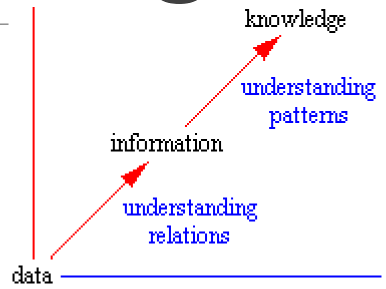


Data

Information

Knowledge

Data – Information – Knowledge



✓ Data

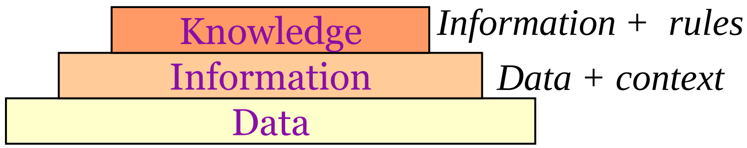
- Facts, numbers, statement of event without relation to other things
- Exp: It is raining.

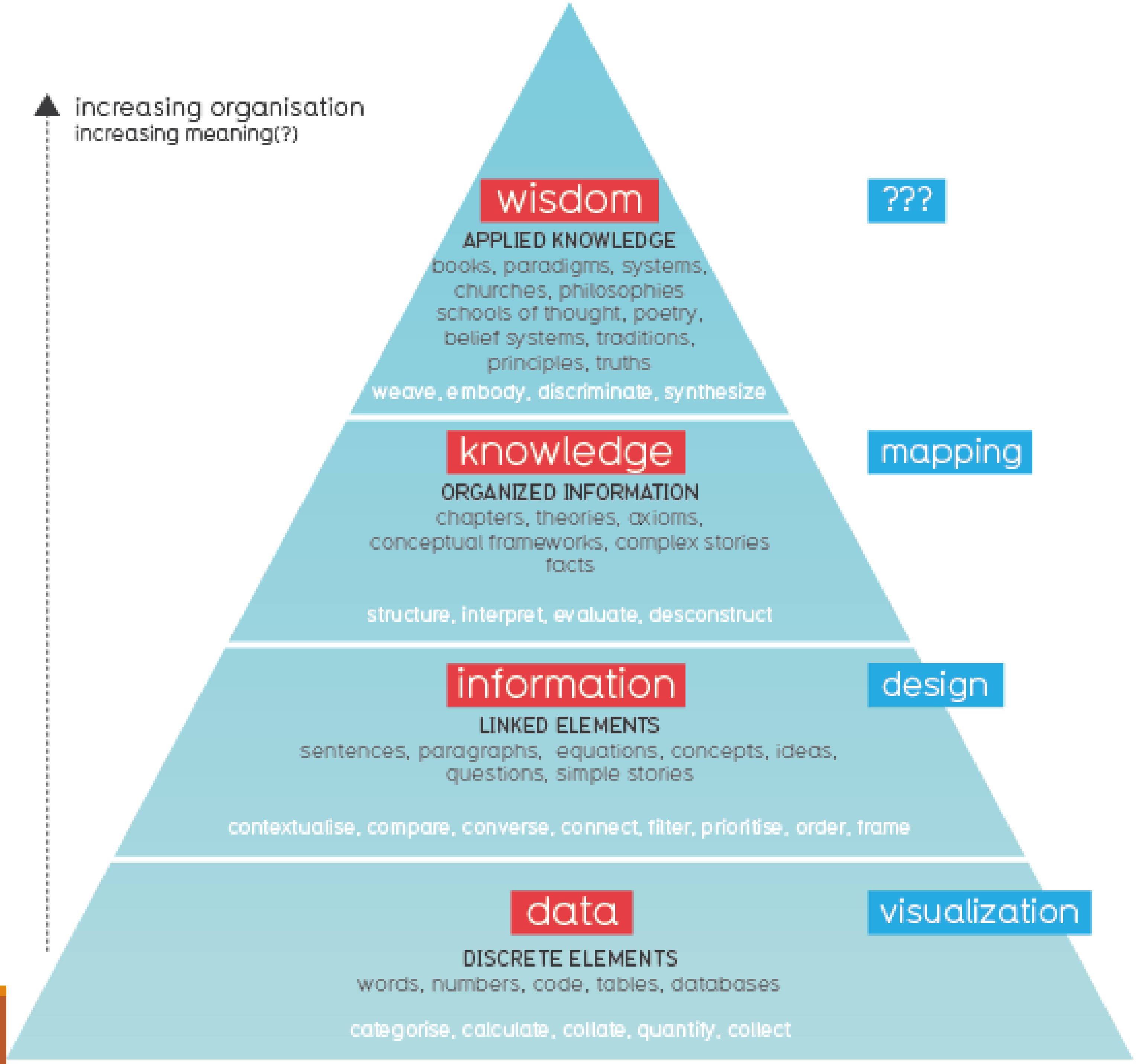
✓ Information

- Data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions
- The understanding of a relationship, possibly cause and effect.
- Exp: The temperature dropped 15 degrees and then it started raining.

✓ Knowledge

- Application of data and information; answers "how" questions
- What is described or What will happen next
- Exp: If the humidity is very high and the temperature drops substantially the atmospheres is often unlikely to be able to hold the moisture so it rains.





increasing organisation
increasing meaning(?)

wisdom

APPLIED KNOWLEDGE

books, paradigms, systems, churches, philosophies schools of thought, poetry, belief systems, traditions, principles, truths

weave, embody, discriminate, synthesize

???

knowledge

ORGANIZED INFORMATION

chapters, theories, axioms, conceptual frameworks, complex stories facts

structure, interpret, evaluate, desconstruct

mapping

information

LINKED ELEMENTS

sentences, paragraphs, equations, concepts, ideas, questions, simple stories

contextualise, compare, converse, connect, filter, prioritise, order, frame

design

data

DISCRETE ELEMENTS

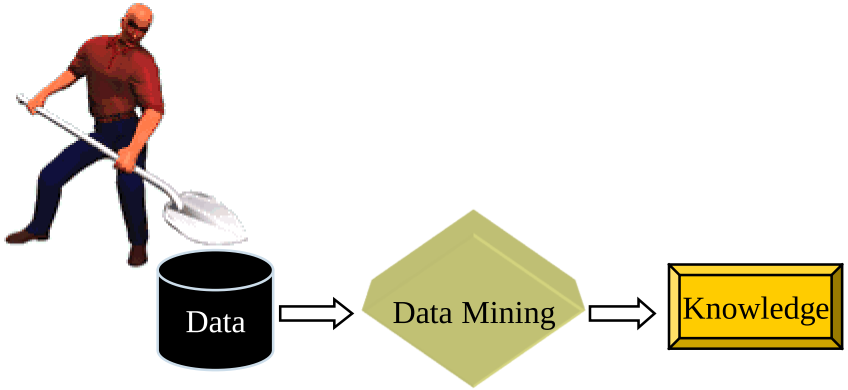
words, numbers, code, tables, databases

categorise, calculate, collate, quantify, collect

visualization

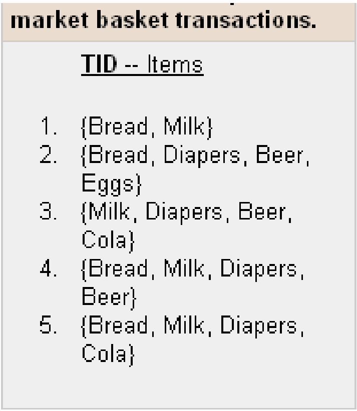
So, What is Data Mining?

- ✓ Data mining refers to extracting or "*mining*" knowledge from large amounts of data.
- ✓ *"Mining"* is a misnomer.
  - *Knowledge Mining,*
  - *Knowledge Extraction,*
  - *Data Archaeology,*
  - *Data Dredging*
- ✓ *Knowledge Discovery in Databases (KDD)*



Data Mining Example

- ✓ Data
  Market sales in ten years

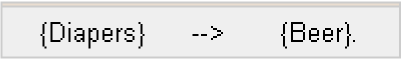


market basket transactions.

TID -- Items

1. {Bread, Milk}
2. {Bread, Diapers, Beer, Eggs}
3. {Milk, Diapers, Beer, Cola}
4. {Bread, Milk, Diapers, Beer}
5. {Bread, Milk, Diapers, Cola}

- ✓ Knowledge
  Many of male customers who buy *diapers* also buy *beer* on every Friday.



{Diapers} --> {Beer}.

Moved the beer and snacks such as peanuts next to the diapers → Increased sales on peanuts by more that 27%

Babies drink beer ???

Knowledge Discovery – Data Mining

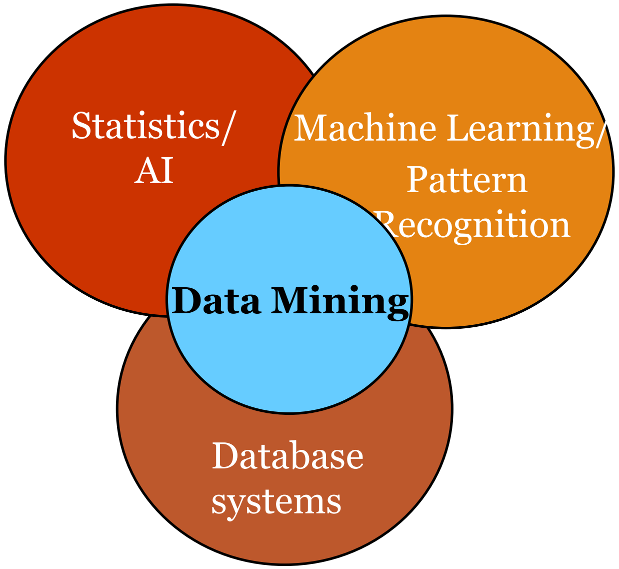
- ✓ Knowledge Discovery is the ***non-trivial*** extraction of ***valid***, ***novel***, ***previously unknown*** and ***potentially useful*** knowledge from large databases.

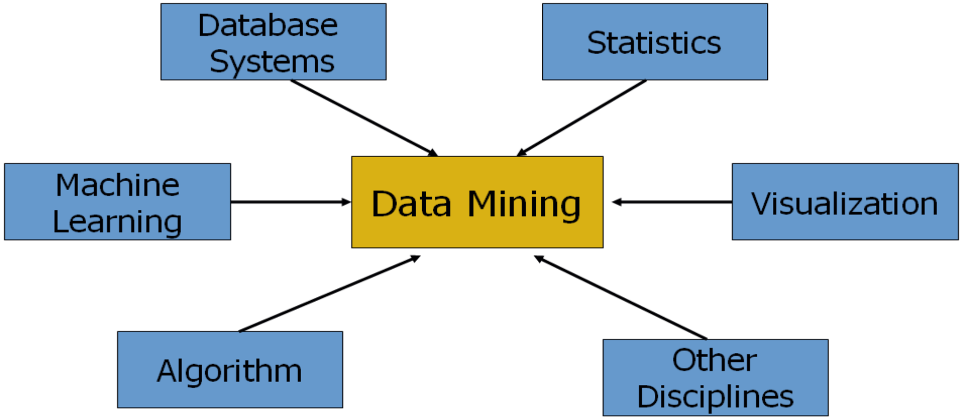
- ✓ Knowledge Discovery is the process of automatically discovering useful information in large data repositories. [Tan, Steinbach, Kumar, 2006]

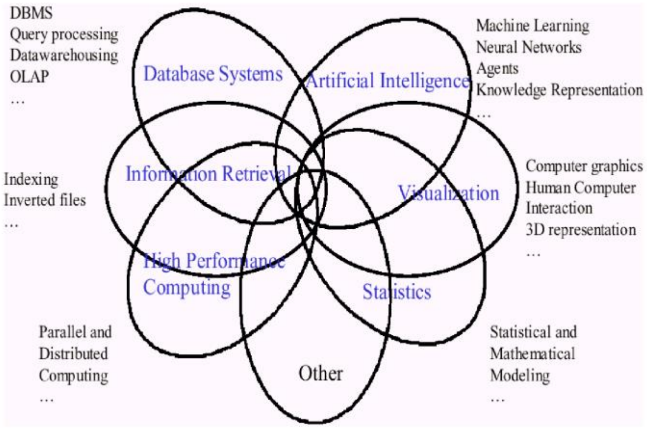
- ✓ Data mining is an interdisciplinary field bringing together techniques from *machine learning*, *pattern recognition*, *statistics*, *databases*, and *visualization* to address the issue of information extraction from large databases.

Intersection of Multiple Disciplines

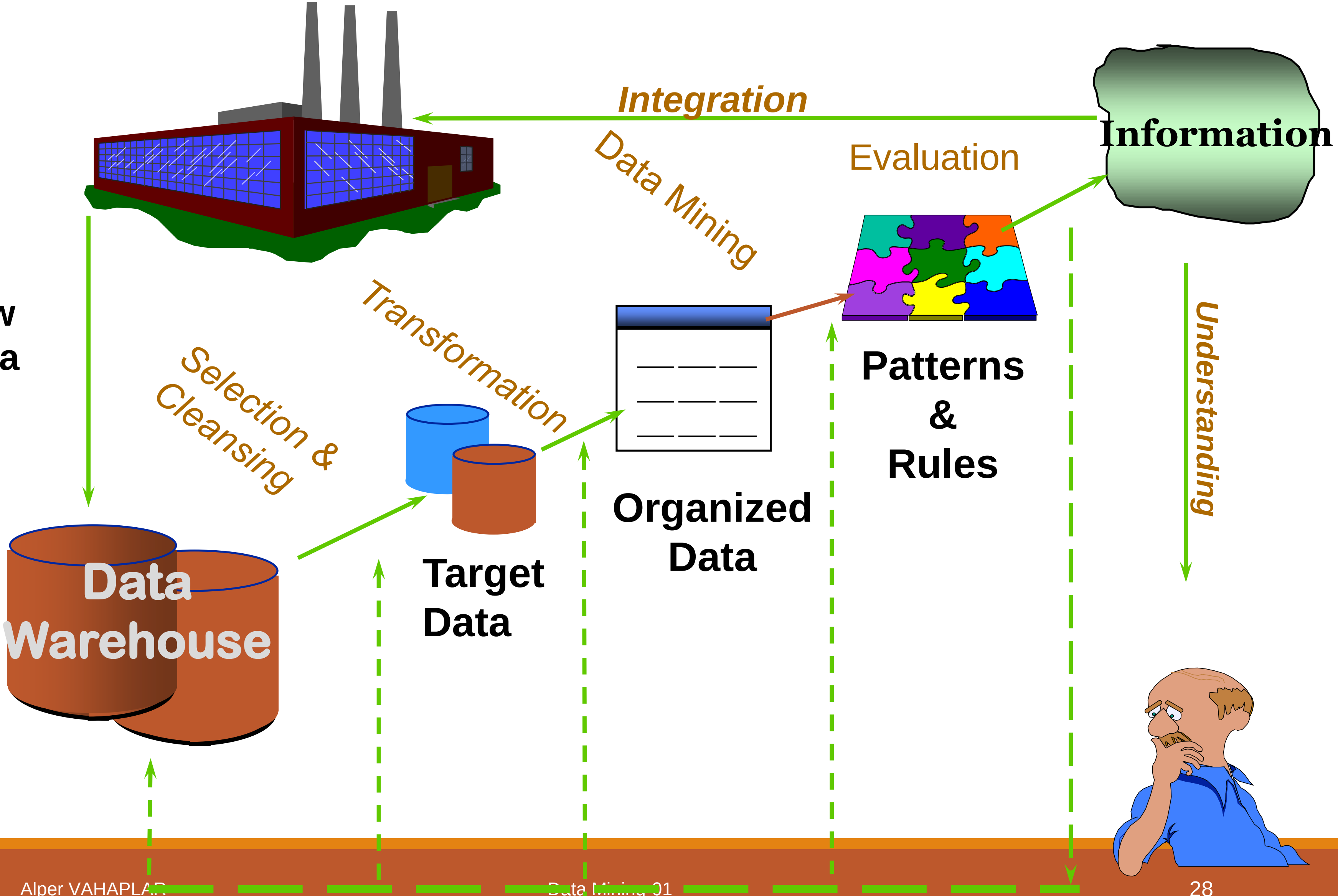
- Database Systems, Data Warehouse and OLAP
- Statistics
- Machine Learning / AI
- Visualization
- Information science
- High Performance Computing
- Other disciplines:
    - Neural Networks, Mathematical Modeling, Information Retrieval, Natural Language Processing …



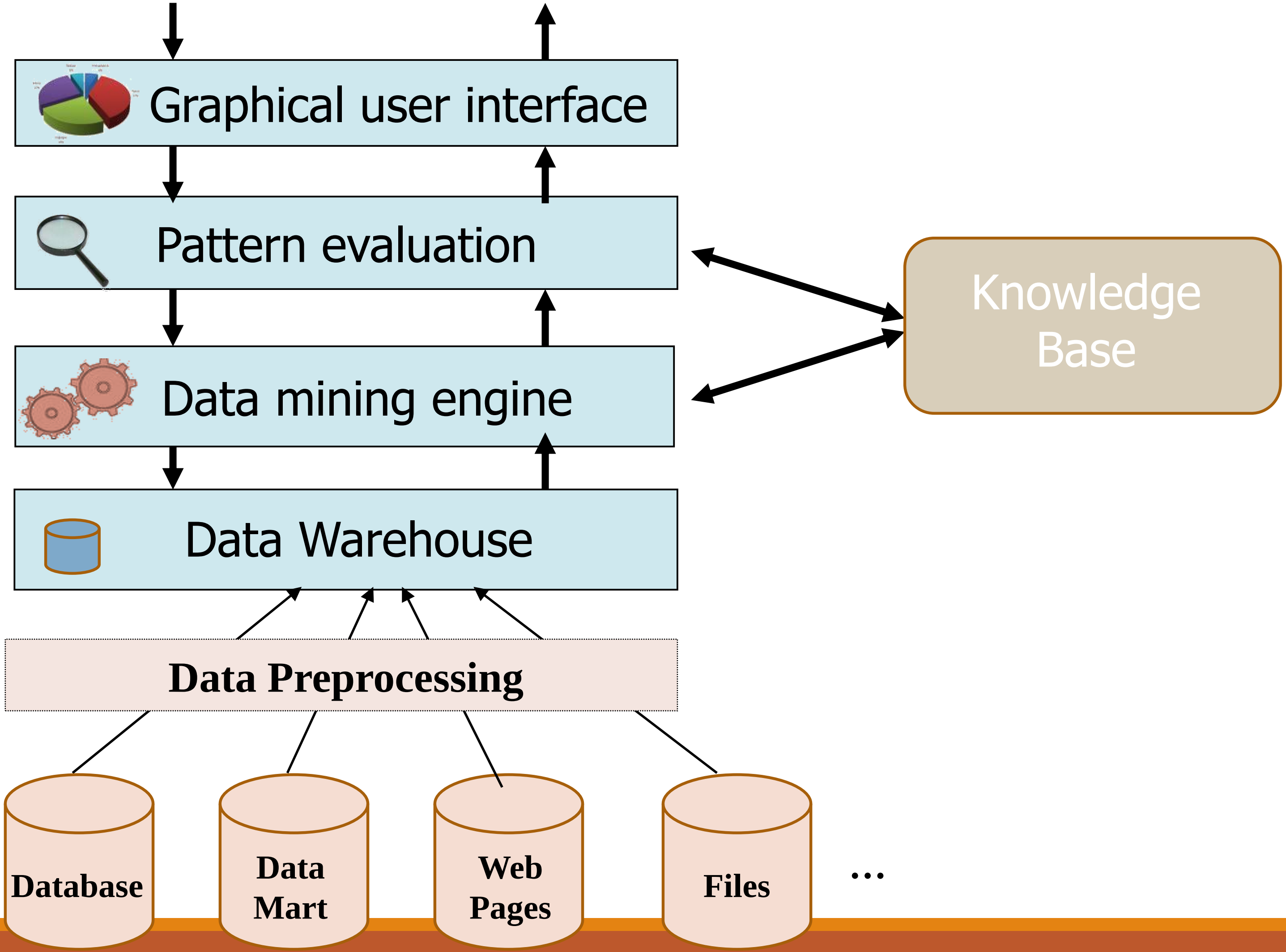




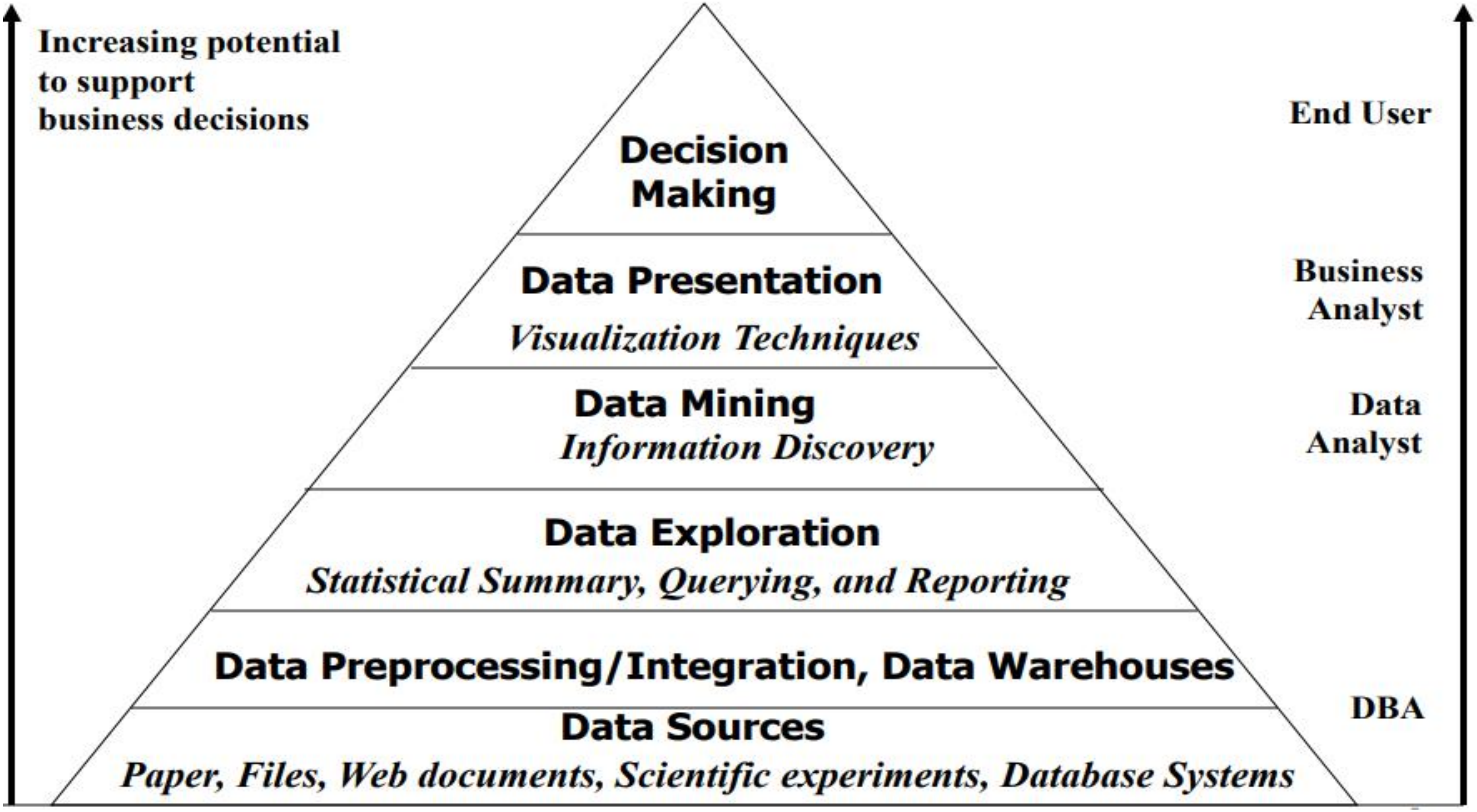
KDD Process


Integration
Information
Data Mining
Evaluation
Raw Data
Transformation
Selection & Cleansing
Patterns & Rules
Understanding
Organized Data
Target Data
Data Warehouse

Architecture: Typical Data Mining System

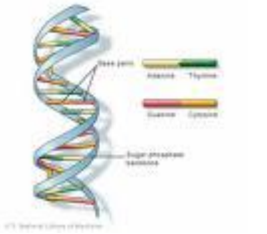


Data Mining Procedure



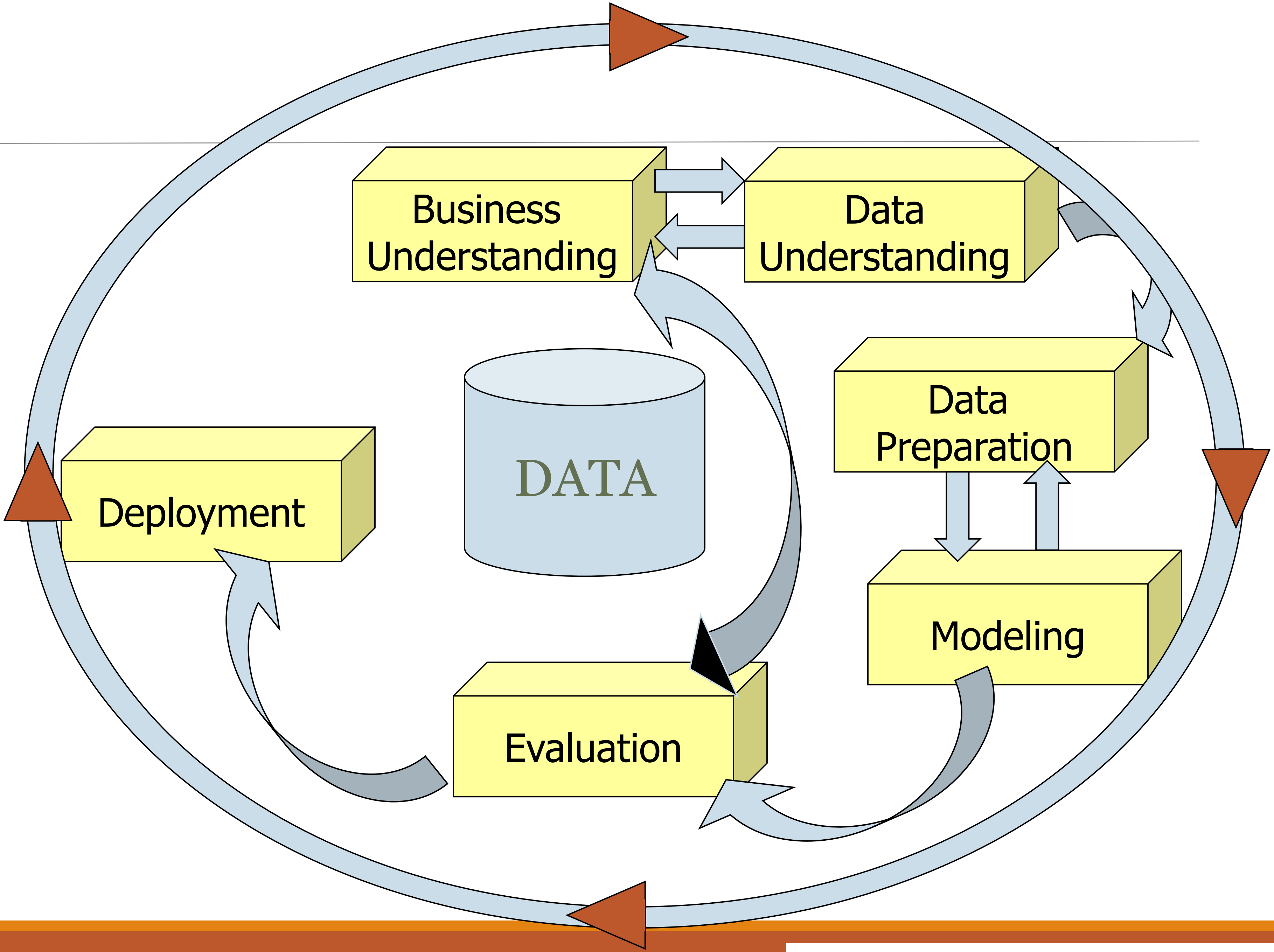
Data Mining Applications

- ✓ Marketing
- ✓ Banking, Insurance and Finance
- ✓ Telecommunication
- ✓ Health, Drug Industry and Bioinformatics
- ✓ Outlier and Fraud Detection
- ✓ Science and Engineering
  - Astronomy,
  - Industry,
  - Chemistry,
  - Sports,
  - Network
  - …



Approaches for KDD

- ✓ 5A by SPSS
  - Assess, Access, Analyze, Act and Automate
- ✓ SEMMA by SAS
  - Sample, Explore, Modify, Model, Assess
- ✓ CRISP-DM by DaimlerChrysler, SPSS (IBM), NCR (1996)
  - Cross Industry Standard Process for Data Mining



Phases of CRISP-DM Life Cycle

CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM

1. Business understanding phase

✓ The project objectives and requirements understanding

✓ Data mining problem definition.

✓ Prepare strategy for achieving these objectives.

2. Data understanding phase

✓ Initial data collection.

✓ Exploratory data analysis

✓ Identification of the data quality problems.

CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM

3. Data preparation phase

- ✓ Prepare the final data set
- ✓ Select the records and variables you want to analyze
- ✓ Perform transformations on certain variables
- ✓ Clean the raw data

4. Modeling phase

- ✓ Select and apply appropriate modeling techniques.
- ✓ Calibrate parameters to optimize results.
- ✓ Several different techniques may be used for the same problem.
- ✓ If necessary, loop back to the data preparation phase

CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM

5. Evaluation phase

- ✓ Evaluate the one or more models for quality and effectiveness.
- ✓ Determine whether the model in fact achieves the objectives set
- ✓ Come to a decision regarding use of the data mining results.

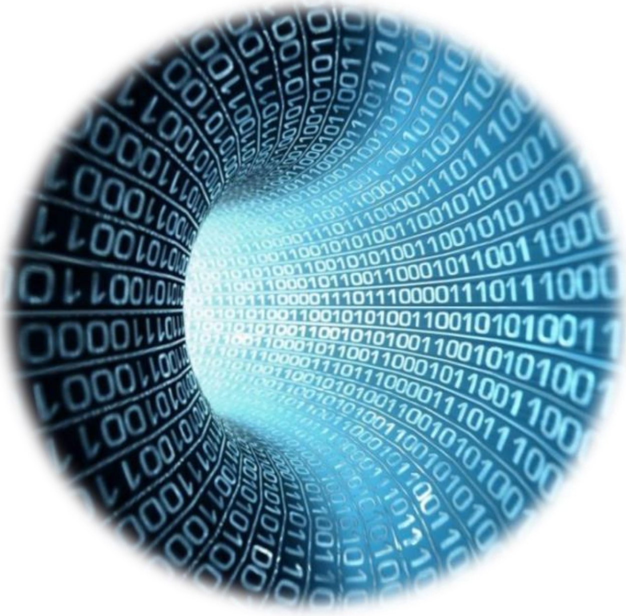
6. Deployment phase

- ✓ Make use of the models created.

Data Mining Tasks

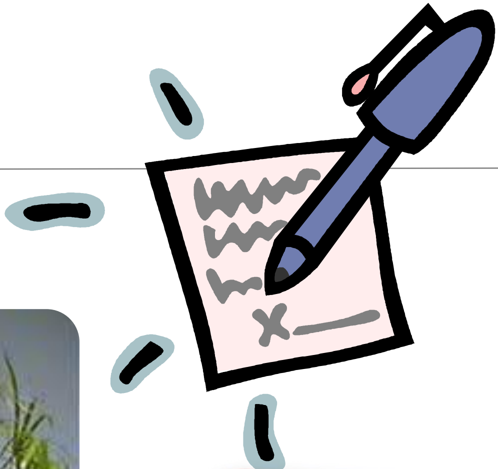
- ✓ Description
- ✓ Clustering
- ✓ Estimation
- ✓ Prediction
- ✓ Classification
- ✓ Association

Data Mining Introduction



Data, Database, Data Warehouse, OLAP, etc.

Where to store "Data"



- ✓Stones,
- ✓Wall of caves
- ✓Animal Skins
- ✓Papyrus Leaves
- ✓Paper
- ✓Computers
  - Flat Files - Sequential, Random Files
  - Formatted Files (Excel, Minitab,...)
  - Databases
  - Data Marts, Data Warehouses

Evaluation of Database Technology

- ✓ 1960s:
  Data collection and database creation – primitive file processing
- ✓ 1970s:
  Relational data model, relational DBMS implementation
- ✓ 1980s:
  Advanced data models (extended-relational, OO, deductive, etc.) Application-oriented DBMS (spatial, scientific, engineering, etc.)
- ✓ 1990s
  Data mining and data warehousing, multimedia databases, and web databases
- ✓ 2000s

  Stream data management and mining

  Data mining and its applications

  Web technology (XML, data integration) and global information systems

Data Types

- ✓Characters (alphanumeric)
- ✓Numerical (integers, floating point, real…)
- ✓Date
- ✓Image
- ✓Voice
- ✓Image + Voice (multimedia data)

Warming up

- ✓ We have a ***universe of objects*** that are of interest.
  - *All the people in the world,*
  - *All the patients in the hospitals of Turkey,*
  - *All dogs in England,*
  - *All web pages on the internet,*
- ✓ The universe of objects is normally ***very large*** and we have only a small part of it.
- ✓ Usually we want to extract ***information*** from the data available to us that we hope is applicable to the large volume of data that we have not yet seen.

Warming up

- ✓ We have a ***universe of objects*** that are of interest.
- ✓ Each object is described by a number of ***variables/attributes*** that correspond to its properties or characteristics that may vary, either from one object to another or from one time to another.
  - ◦ *Ex: eye color, age, temperature, number of children, etc.*
- ✓ The set of variable values corresponding to each of the objects is called a ***record*** or (more commonly) an ***instance***.
- ✓ The complete set of data available to us for an application is called a ***dataset***.
  - ◦ Depicted as tables (instances in *rows*, attributes in *columns*)

Types of Variables

- ✓ Categorical
  - Qualitative
- ✓ Numeric
  - Quantitative

Types of Variables – Scale

- ✓ **Nominal Variables**
  - A variable used to put objects into categories,
  - Ex: color of an object, ID number (1, 2, 3, 4..)
- ✓ **Ordinal Variables**
  - similar to nominal variables, except that having values which can be arranged in a meaningful order,
  - Ex: small, medium, large.

Types of Variables – Scale

- ✓ **Interval Scaled Variables**
  - Interval-scaled variables are variables that take numerical values which are measured at equal intervals from a zero point or origin.
  - A unit of measurement exists.
  - However the origin does not imply a true absence of the measured characteristic.
  - Ex: temperature in Celcius,
- ✓ **Ratio Scaled Variables**
  - similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic.
  - Ex: molecular weight, price in dollars.
  - Differences and ratios is meaningful.

Attribute Type		Description	Examples	Operations
Categorical (Qualiatative)	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
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

Exercise

- ✓Try to investigate and introduce the following data set:
- ✓http://alpervahaplar.com – IST4138
- ✓CarData.xls

Data Mining – On What Kind of Data?

- ✓ Relational Databases
- ✓ Data Warehouses
- ✓ Transactional Databases
- ✓ Object Oriented Databases
- ✓ Spatial Databases
- ✓ Time Series Databases
- ✓ Text and Multimedia Databases
- ✓ World Wide Web

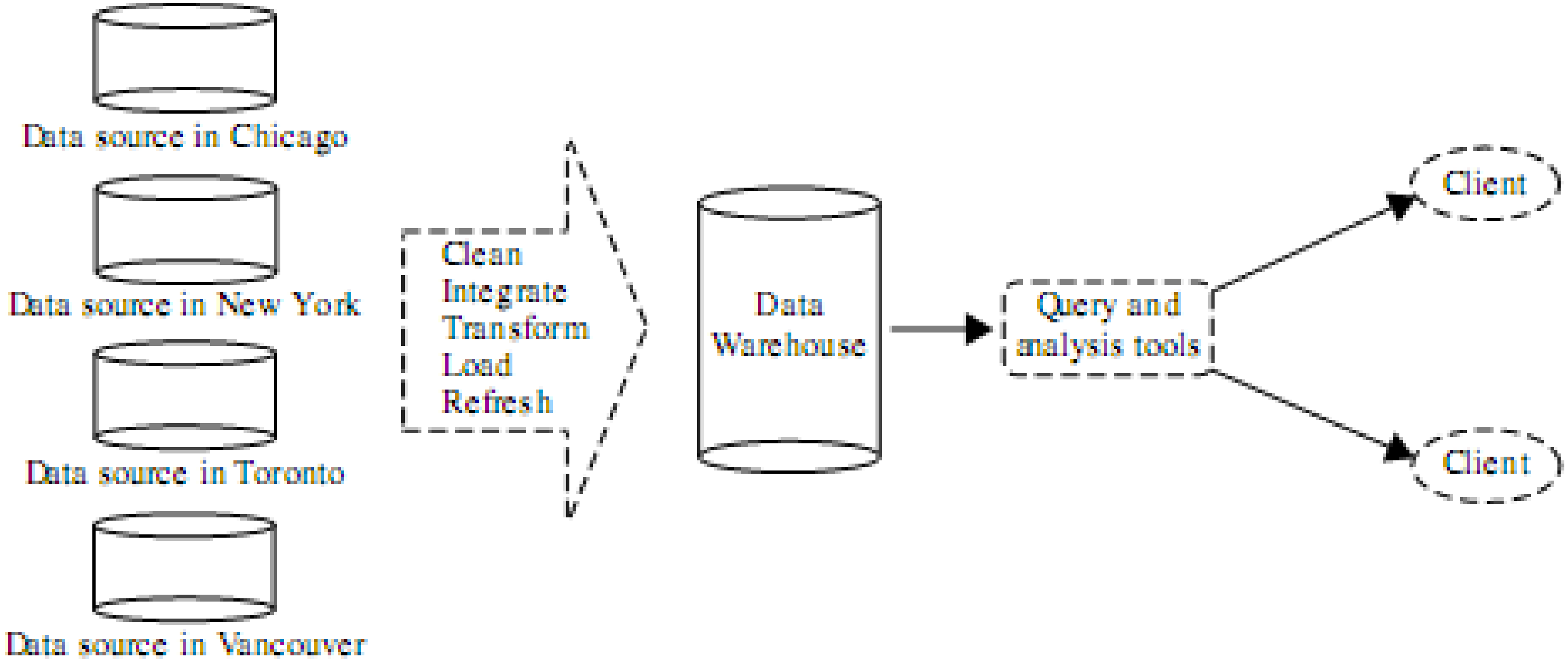
Relational Databases

- ✓ A Relational Database is a collection of tables, consisting a set of attributes (columns) and storing a large set of tuples (records, rows).
- ✓ Each tuple in a table represents an object identified by a unique key.
- ✓ Relational Data can be accessed by database queries, such as SQL.
- ✓ Some operations: join, selection, projection.

Data Warehouse

✓ A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site.

✓ A data warehouse is a subject-oriented, integrated, time- of

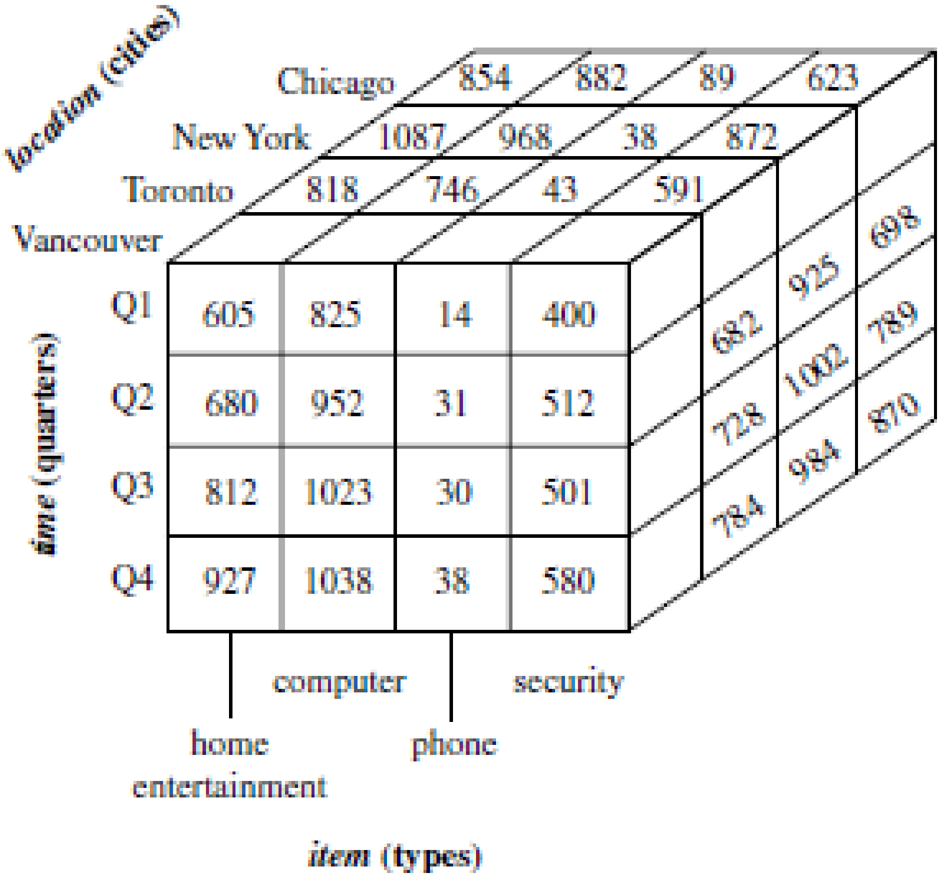


Data Warehouse – OLAP

- ✓ On Line Analytical Processing
- ✓ computer processing that enables a user to easily and selectively extract and view data from different points of view.
- ✓ Traditional query and report tools describe *what is in a database*.
- ✓ OLAP goes further; it's used to answer *why certain things are true.*
- ✓ The user forms a hypothesis about a relationship and verifies it with a series of queries against the data.
- ✓ OLAP Operations
  - Drill Down
  - Roll Up

Table 4.3 3-D View of Sales Data for *AllElectronics* According to *time*, *item*, and *location*

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

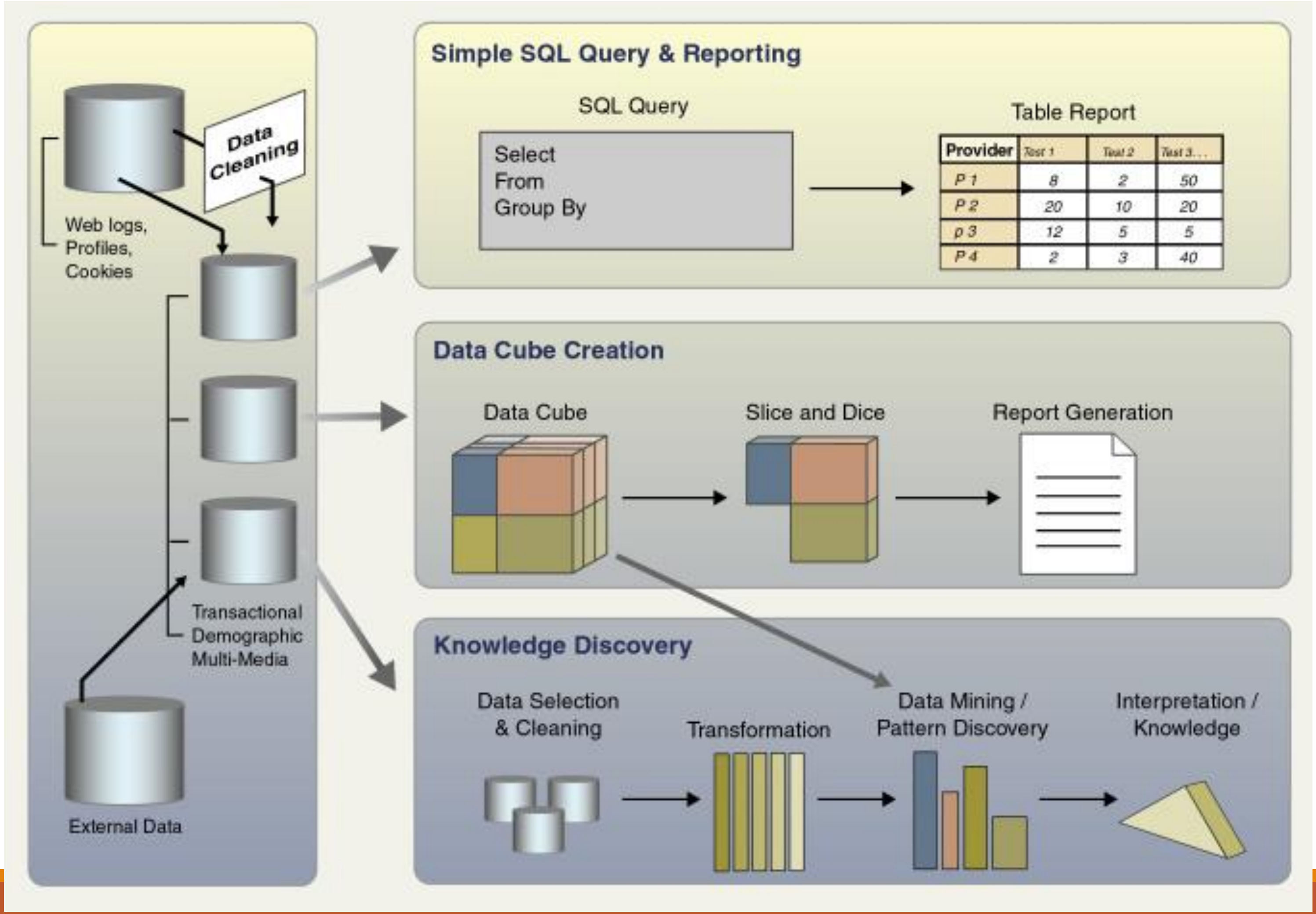


OLAP vs. Data Mining

- ✓The OLAP analyst **generates a series of hypothetical patterns** and relationships and uses queries against the database to verify them or disprove them.
- ✓OLAP analysis is essentially a **deductive process**.

- ✓Data mining is different from OLAP because **rather than verify hypothetical patterns**, it uses the data itself to uncover such patterns.
- ✓It is essentially **an inductive process**.

DBMS, OLAP, and Data Mining



Provider	Test 1	Test 2	Test 3. . .
P 1	8	2	50
P 2	20	10	20
p 3	12	5	5
P 4	2	3	40

Next Week

- ✓Data Understanding,
- ✓Data Visualization,
- ✓Data Preprocessing,
- ✓Data Cleaning…