Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Data Mining

Course ID: CO3029

Chapter 1: Overview

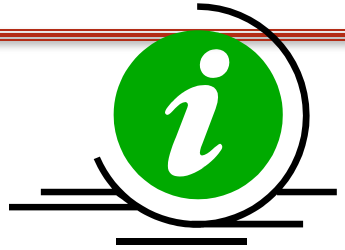
Assoc. Prof. TRAN MINH QUANG

quangtran@hcmut.edu.vn

http://researchmap.jp/quang

2020/4/7

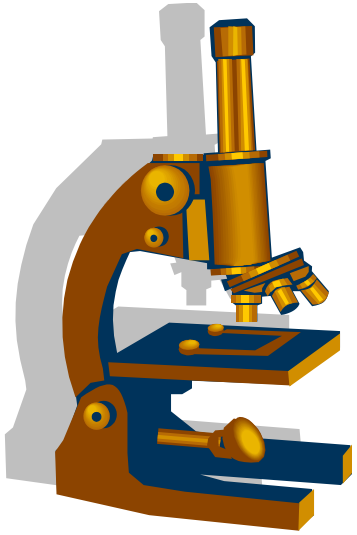
DATA MINING: A QUICK GLANCE

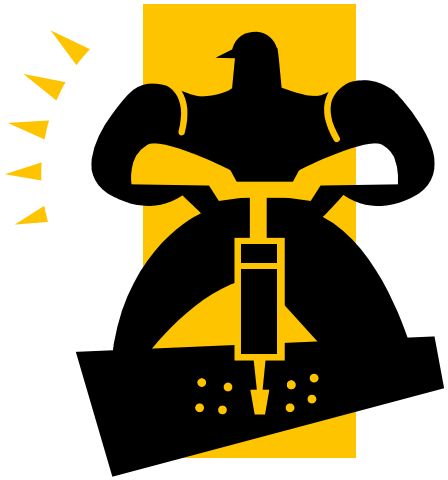


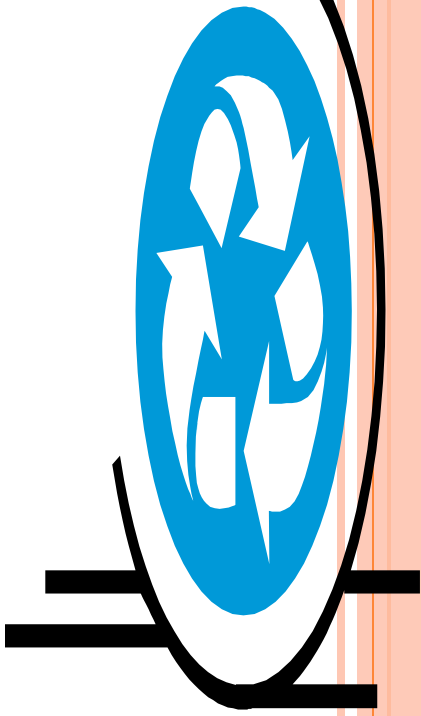
Information/
Knowledge

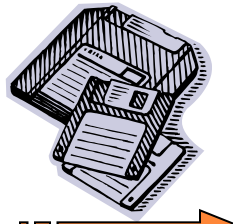
Mining

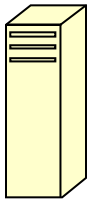
Data

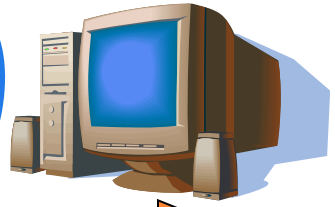


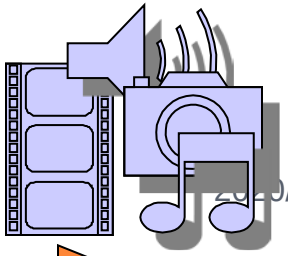












Content

1. Practical situations with Data mining
2. Knowledge discovery
3. Main concepts
4. Roles of data mining
5. Applications
6. Summary

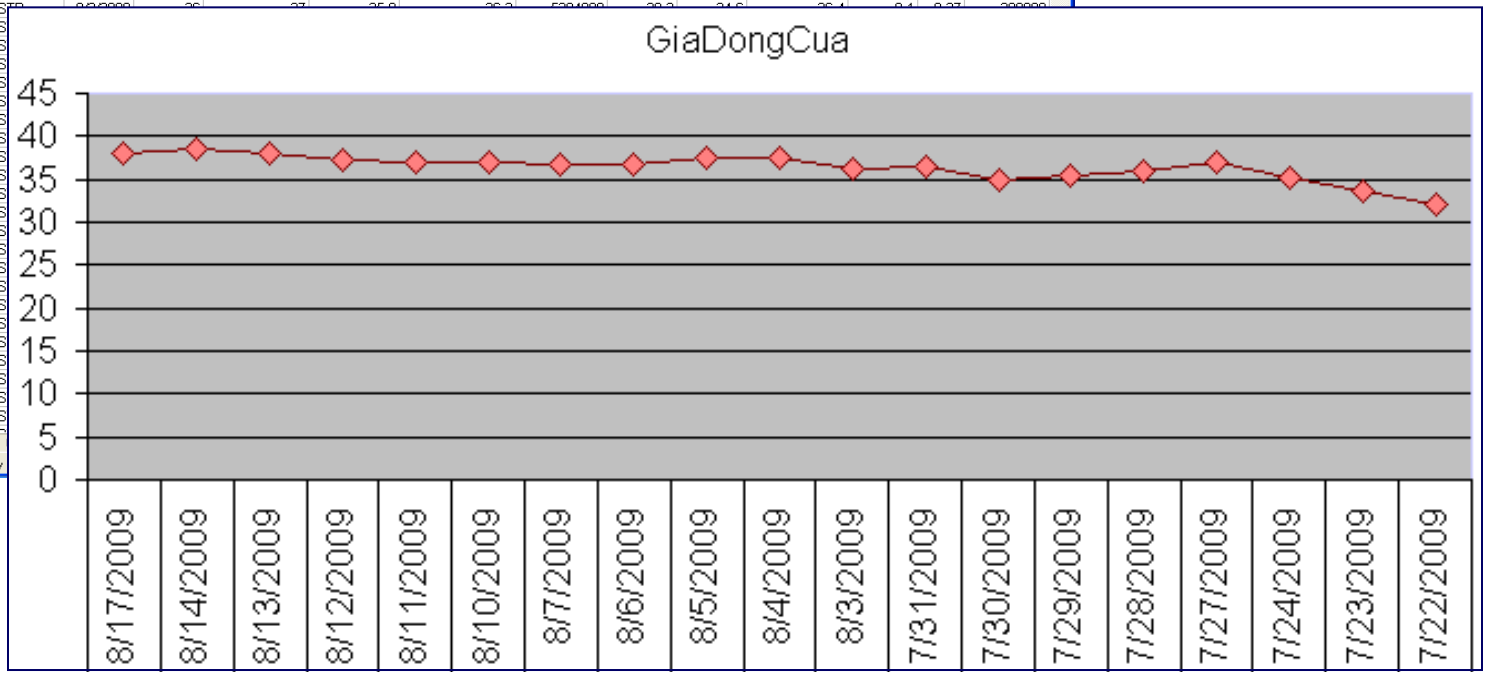
1. Situation 1

Estimating whether the guy who using a credit card with ID = 123456 is its owner of not

1. SITUATION 2

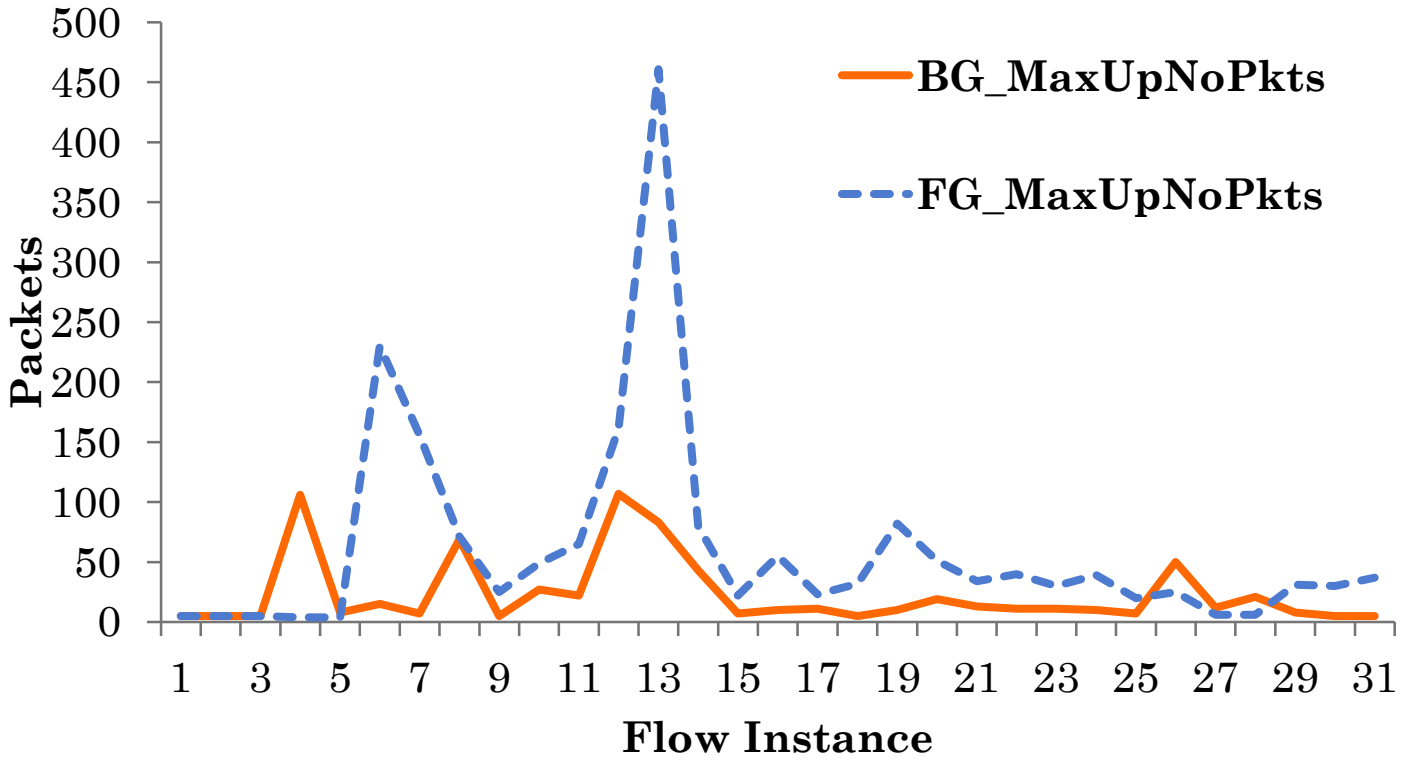
Microsoft Excel - stb.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	MaCK	Ngay	GiaMoCua	GiaCaoNhat	GiaThapNhat	GiaDongCua	KhoiLuongGD	GiaTran	GiaSan	GiaThamChieu	TangGiam	%	GDThoaThua
2	STB	8/17/2009	38.5	38.8	38.1	38.1	5986700	40.4	36.6	38.5	-0.4	-1.04	24343
3	STB	8/14/2009	38	38.7	38	38.5	6886430	39.9	36.1	38	0.5	1.32	340000
4	STB	8/13/2009	38	38.5	37.6	38	8716920	39	35.4	37.2	0.8	2.15	188000
5	STB	8/12/2009	37.3	37.4	37	37.2	5361890	38.7	35.1	36.9	0.3	0.81	200000
6	STB	8/11/2009	37.1	37.3	36.9	36.9	3675610	38.9	35.3	37.1	-0.2	-0.54	0
7	STB	8/10/2009	37.2	37.6	36.8	37.1	6140320	38.5	34.9	36.7	0.4	1.09	0
8	STB	8/7/2009	37	37	36.6	36.7	4525140	38.6	35	36.8	-0.1	-0.27	0
9	STB	8/6/2009	37.4	37.7	36.8	36.8	6647680	39.2	35.6	37.4	-0.6	-1.6	200000
10	STB	8/5/2009	37	37.5	36.9	37.4	5071900	39.2	35.6	37.4	0	0	0
11	STB	8/4/2009	37.8	37.8	36.8	37.4	10313950	38.1	34.5	36.3	1.1	3.03	133000



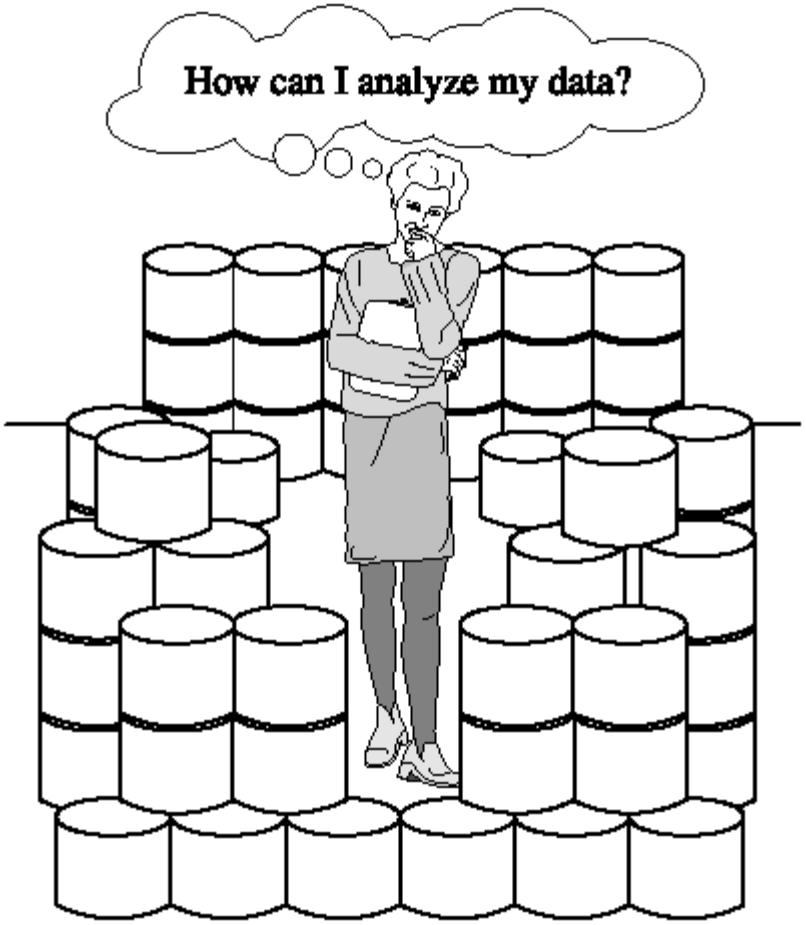
Predict the price of a stock, e.g., STB

1. SITUATION 3



Network attacking detection based on traffic analysis

1. The Fact…



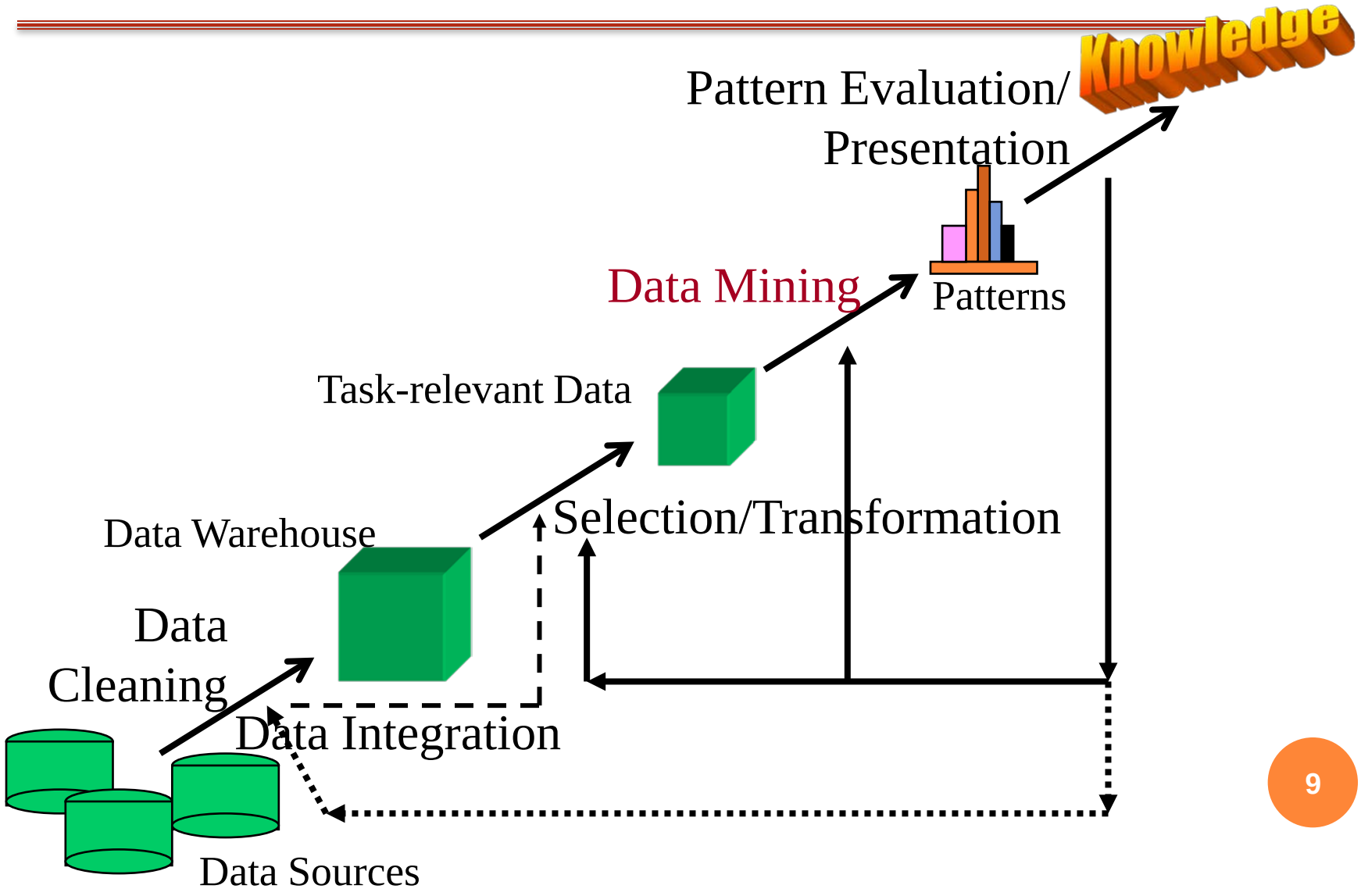
We are data rich, but information poor

“Necessity is the mother of invention” - Plato

2. Knowledge Discovery from Database (KDD)

- "Knowledge discovery in **databases** is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable **patterns**"
  - Frawley, W. J et al. (1991). Knowledge discovery in databases: an overview.
- "Knowledge discovery from **databases** is the process of using the database along with any required selection, preprocessing, sub-sampling, and transformations of it; to apply data mining methods (algorithms) to enumerate **patterns** from it; and to evaluate the products of data mining to identify the subset of the enumerated patterns deemed **knowledge**."
  - Fayyad, U.M et al. (1996). Advances in Knowledge Discovery and Data Mining. MIT Press.

2. KDD…


Knowledge
Pattern Evaluation/
Presentation
Data Mining
Patterns
Task-relevant Data
Data Warehouse
Selection/Transformation
Data
Cleaning
Data Integration
Data Sources

2. KDD…

…is an iterative process with following main steps:

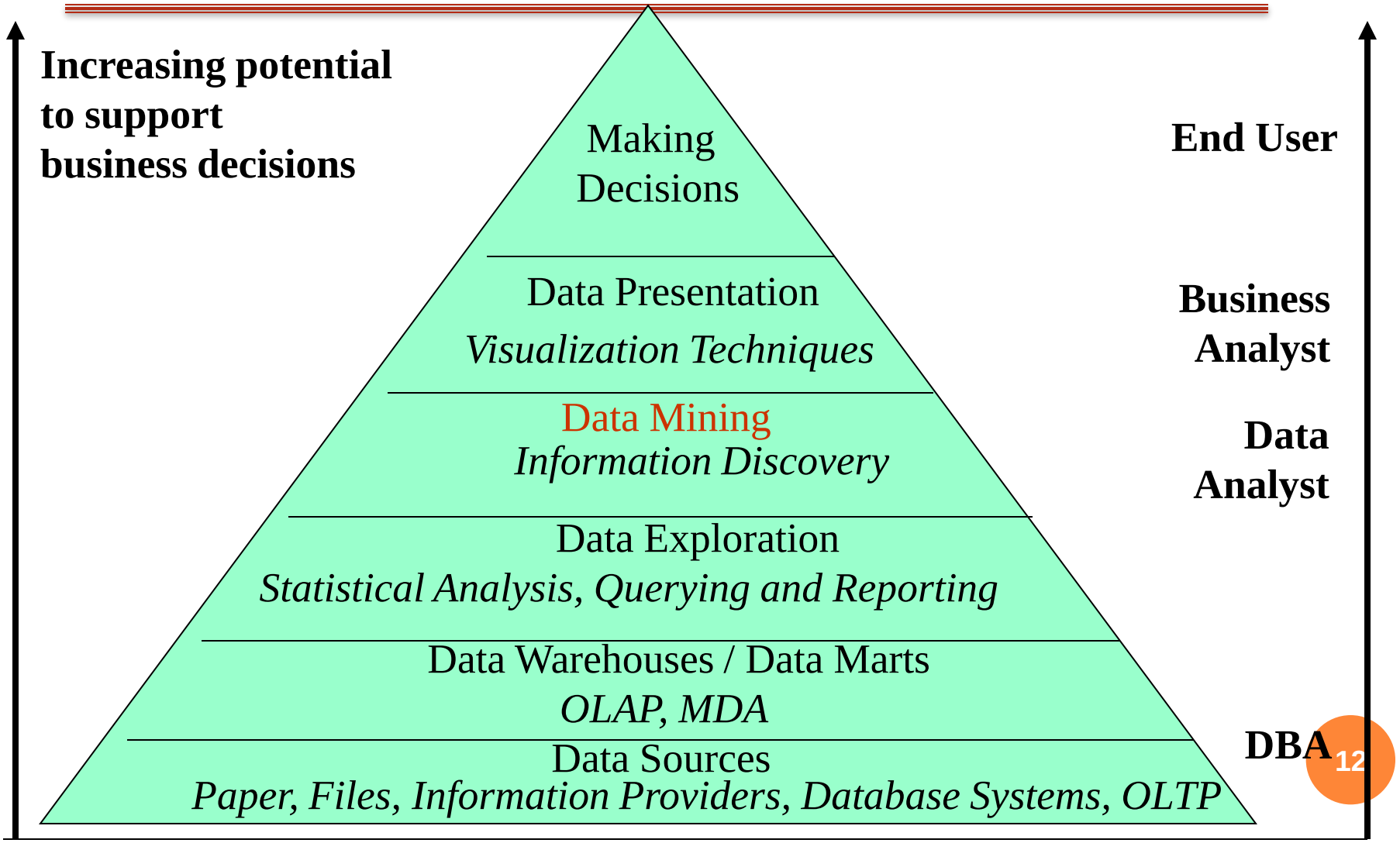
1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

2. KDD…

… each step in KDD process may work with

- Data sources (many types)
- Data warehouse
- Task-relevant data
- Patterns
- Knowledge

2. KDD in the Data Management Pyramid



Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Data Mining

Course ID: CO3029

Chapter 1: Overview

Part 2

Assoc. Prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

Content

1. Practical situations with Data mining
2. Knowledge discovery
3. **Main concepts**
4. Roles of data mining
5. Applications
6. Summary

3. Main Concepts in KDD

- Data mining
- Data mining tasks/functions
- Data mining processes
- Data mining systems

3.1. Data Mining (DM)

DM is a process of …

- "extracting or mining knowledge from large amounts of data"
- "knowledge mining from data"
- "nontrivial extraction of implicit, previously unknown, and potentially useful information from data"

3.1. Data Mining

Similar/common terms

- knowledge discovery/mining in data/databases (KDD)
- knowledge extraction
- data/pattern analysis
- data archaeology, data dredging
- information harvesting
- business intelligence

3.1. Data Mining: Data sources

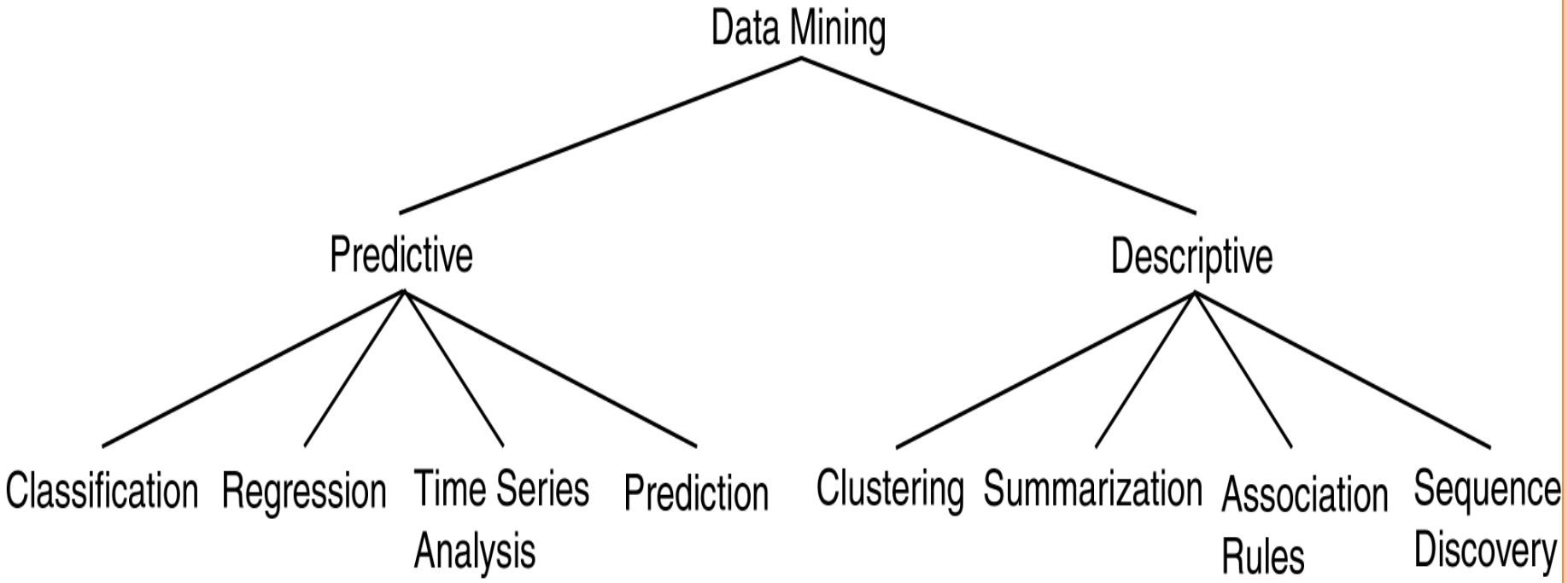
- DM from large amounts of data…
  - Any types: structure, non-structure, semi-structure from various data sources
  - Data sources
    - Flat files
    - Databases: relational databases, object-relational databases, NoSQL,…
    - Transactional databases, data warehouses
    - From various domains: spatial databases, temporal databases, spatio-temporal databases, time series databases, text/docuement databases, multimedia databases, …
    - Data from the web: WWW, Social networks…
  - Pack or streaming data sources: Data warehouse for BI, real-time IoT systems,….

3.1. Data Mining: Knowledge

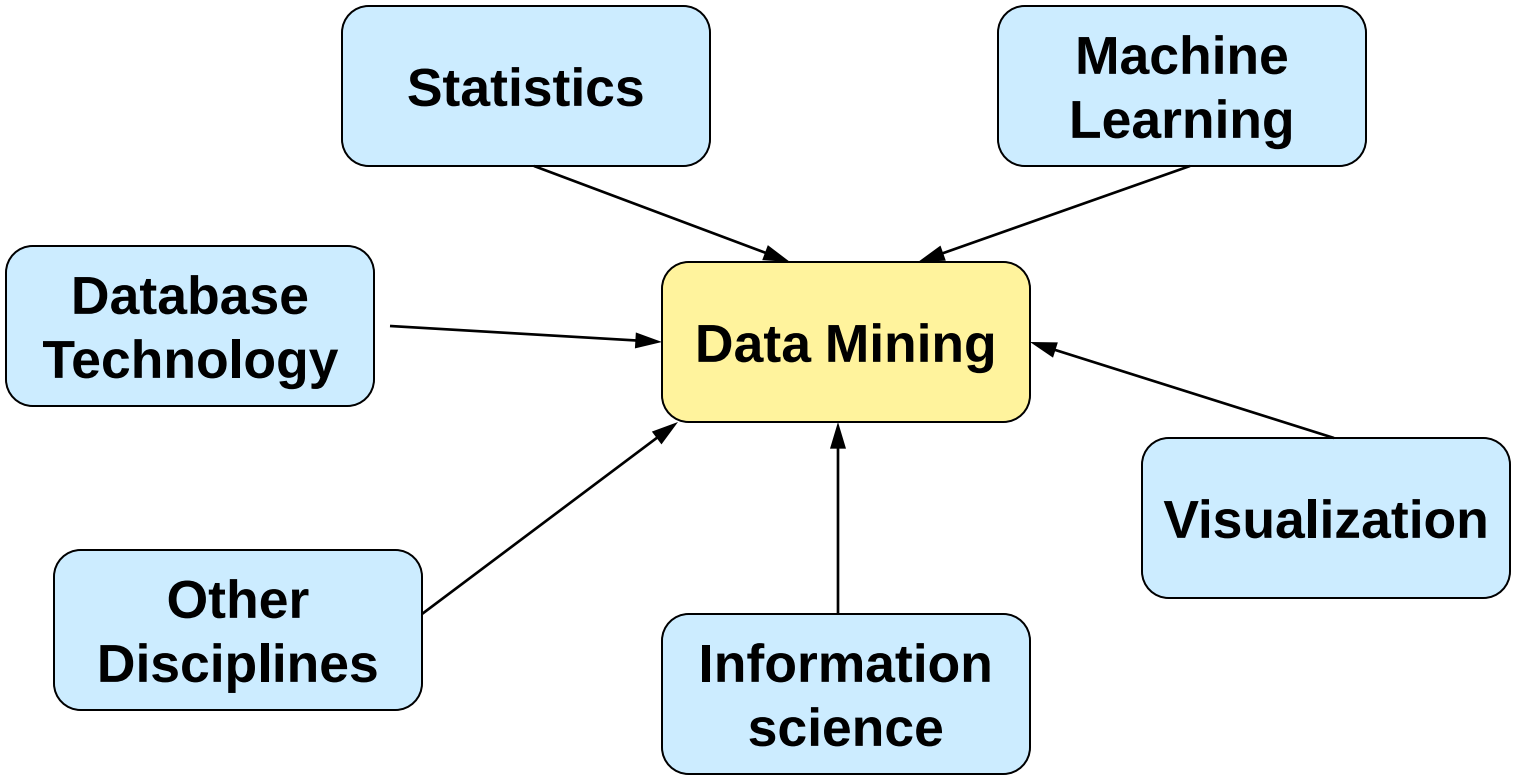
- The mined/analyzed knowledge can be:
  - Description of data classes
  - Frequent patterns, Association patterns
  - Classification and Prediction
  - Clustering Model
  - Outliers
  - Trends of behaviors, data,…
  - …

3.1. Data Mining: Knowledge

- Categories of mined/analyzed knowledge :
  - Descriptive: Describe common characteristics of objects in the dataset (situation 1)
  - Predictive): Capability to infer/predict new information based on current data (situations 2, 3)



3.1. Data Mining

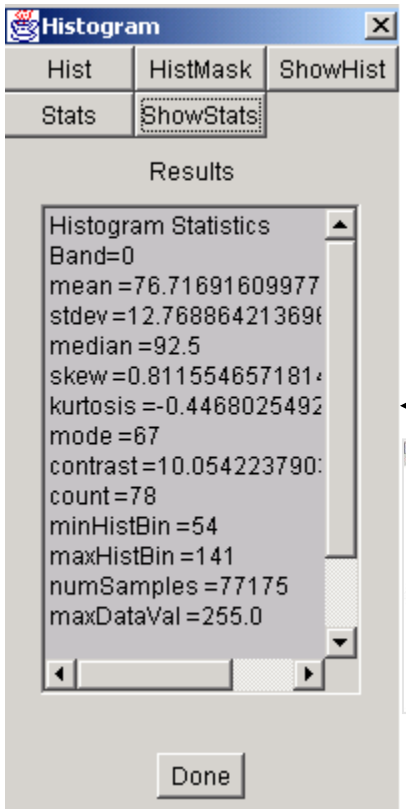


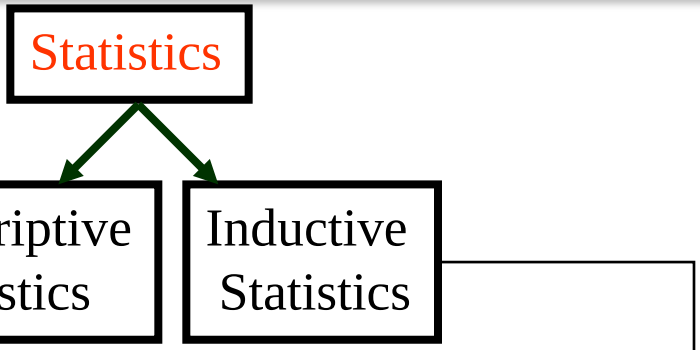
"Data mining as a confluence of multiple disciplines"

3.1. Data Mining: DB Technologies

- Database technologies help efficiently manage data for mining
  - Big data: paging, swapping from disk to memory, distributed data managements, support for data streaming,…
  - Integrated, categorized in based on different dimensions (e.g., data warehouse)
  - Support various data types: spatial, temporal, spatiotemporal, multimedia, text, Web, …
  - Concurrency control, security, query optimization,…
  - Provide data mining functions:
    - Oracle Data Mining (Oracle 9i, 10g, 11g)
    - SQL Server analyzers, Azure machine learning,…
    - Intelligent Miner (IBM)
    - Standard SWL/MM 6: Data Mining by ISO/IEC 13249-6:2006

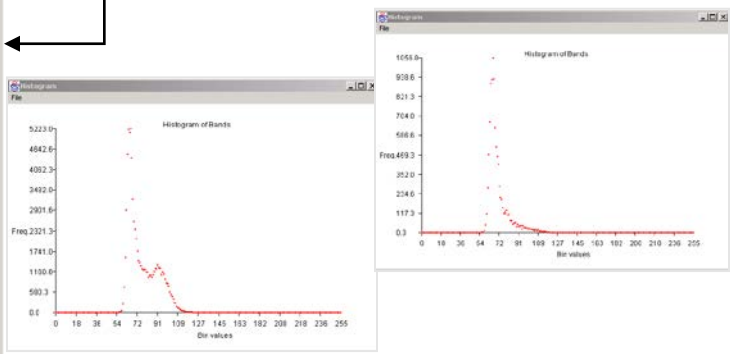
3.1. DATA MINING: STATISTICS





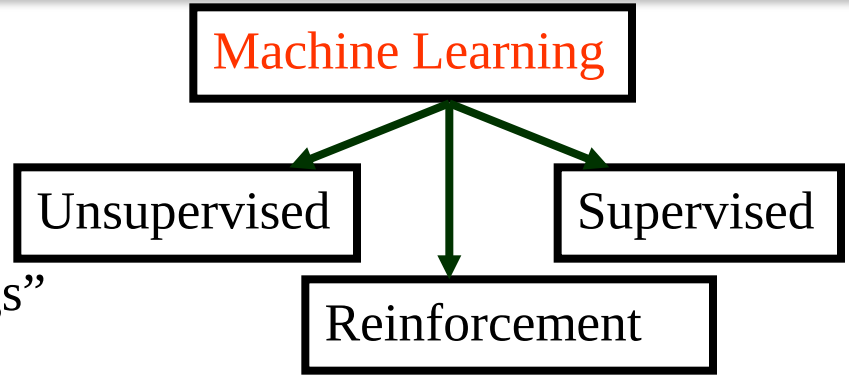
Data description

Prediction and Induction

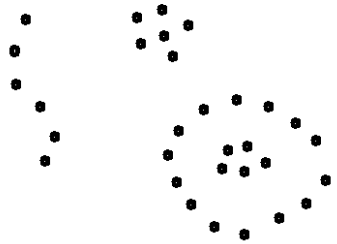


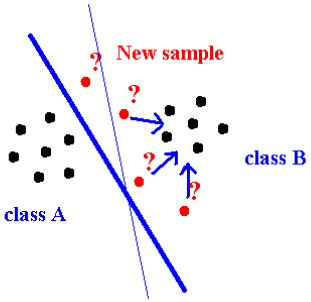
→ **Distributions of data in the two are similar?**

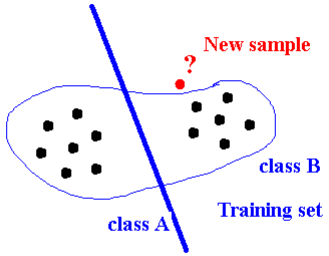
3.1. Data Mining: Machine Learning

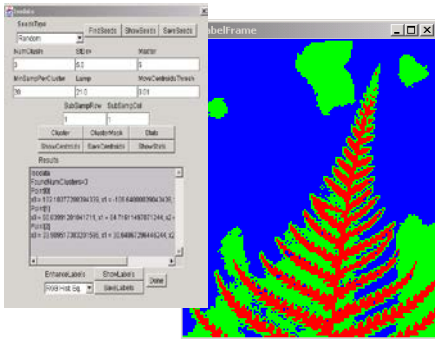


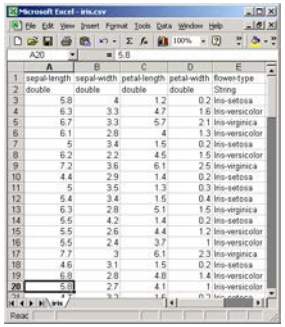
"Natural groupings"

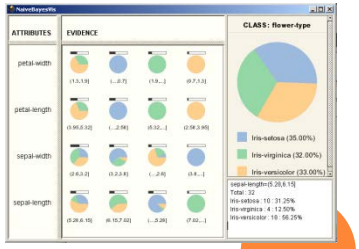






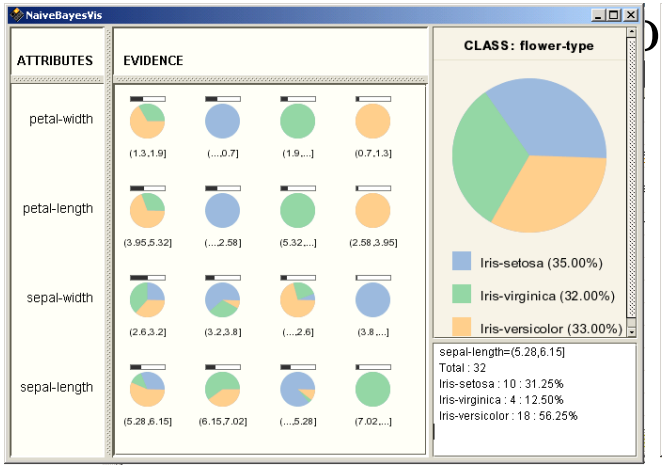




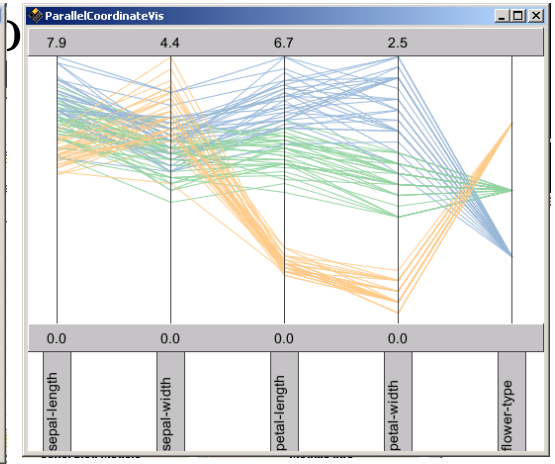


3.1. Data Mining: Visualization

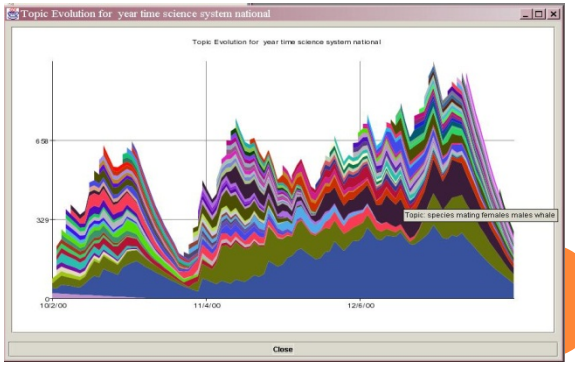
- Improve the meaning of knowledge to users
  - Data: 3D cubes, distribution charts, curves, surfaces, link graphs, image frames and movies, parallel coordinates
  - Knowledge (mining results): pie charts, scatter plots, box plots, association rules, parallel coordinates,



Pie chart



Parallel coordinates

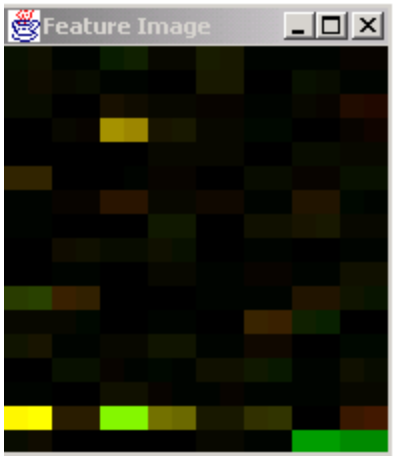


Temporal evolution

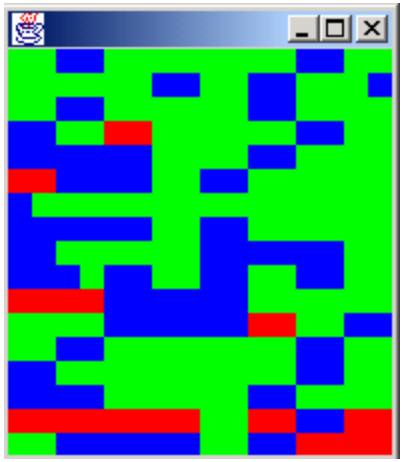
3.1. Data Mining: Visualization

- Labeling mined classes/clusters

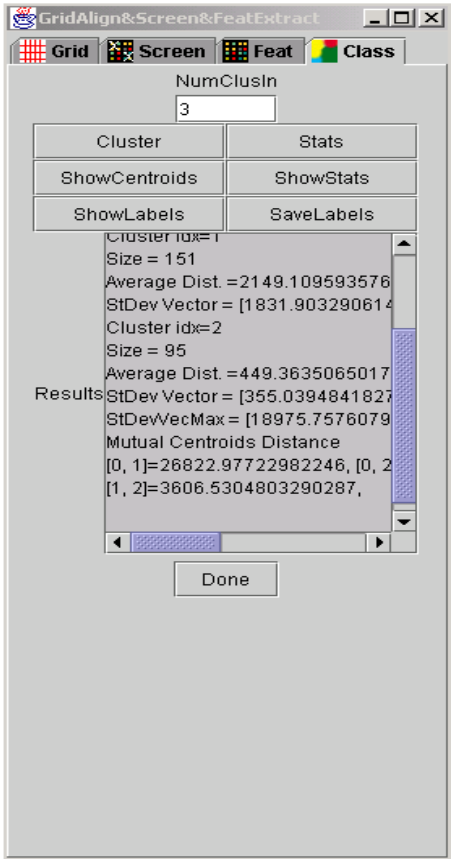
Isodata (K-means) Clustering



Mean Feature Image



Label Image



Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Data Mining

Course ID: CO3029

Chapter 1: Overview

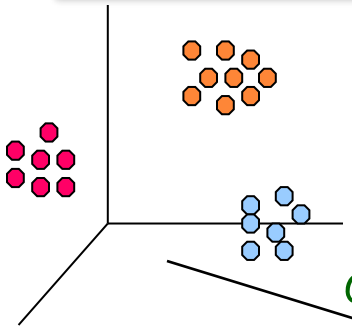
Part 3

Assoc. Prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

3.2. Data Mining Tasks

- Data description
  - Classification
  - Prediction
  - Clustering
  - Association rule mining
  - Trend analysis
  - Outlier
  - Similarity analysis
- …

3.2. Data Mining Tasks

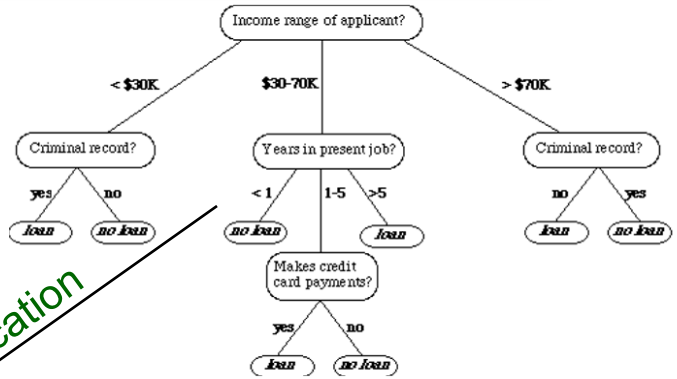


Clustering

Data

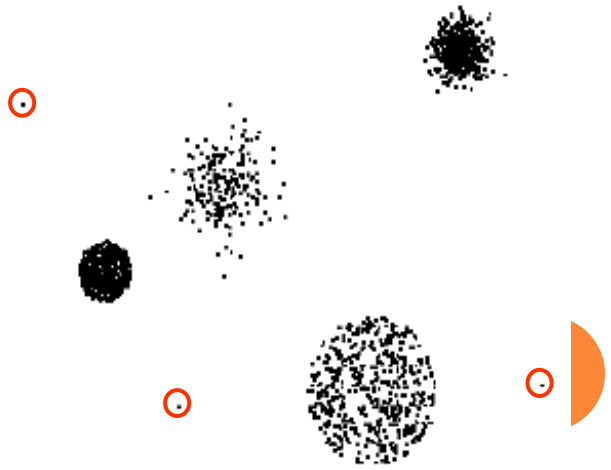
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Classification



Association Rules

Anomaly Detection



others

3.2. Data Mining Tasks: Main factors

- 5 main factors describe a data mining task
  1. Task-relevant data
  2. Expected knowledge
  3. Background knowledge
  4. Interestingness measures
  5. Pattern evaluations and knowledge presentation

3.2. Data Mining Tasks: Main factors

- Task-relevant data: data sources, data types, selected features/dimensions, name of DBs, data warehouse, data tables or objects or documents, criteria for selection data,…
- Expected knowledge: corresponds to a specific mining task which will be executed: classification, clustering, association rules, prediction,….

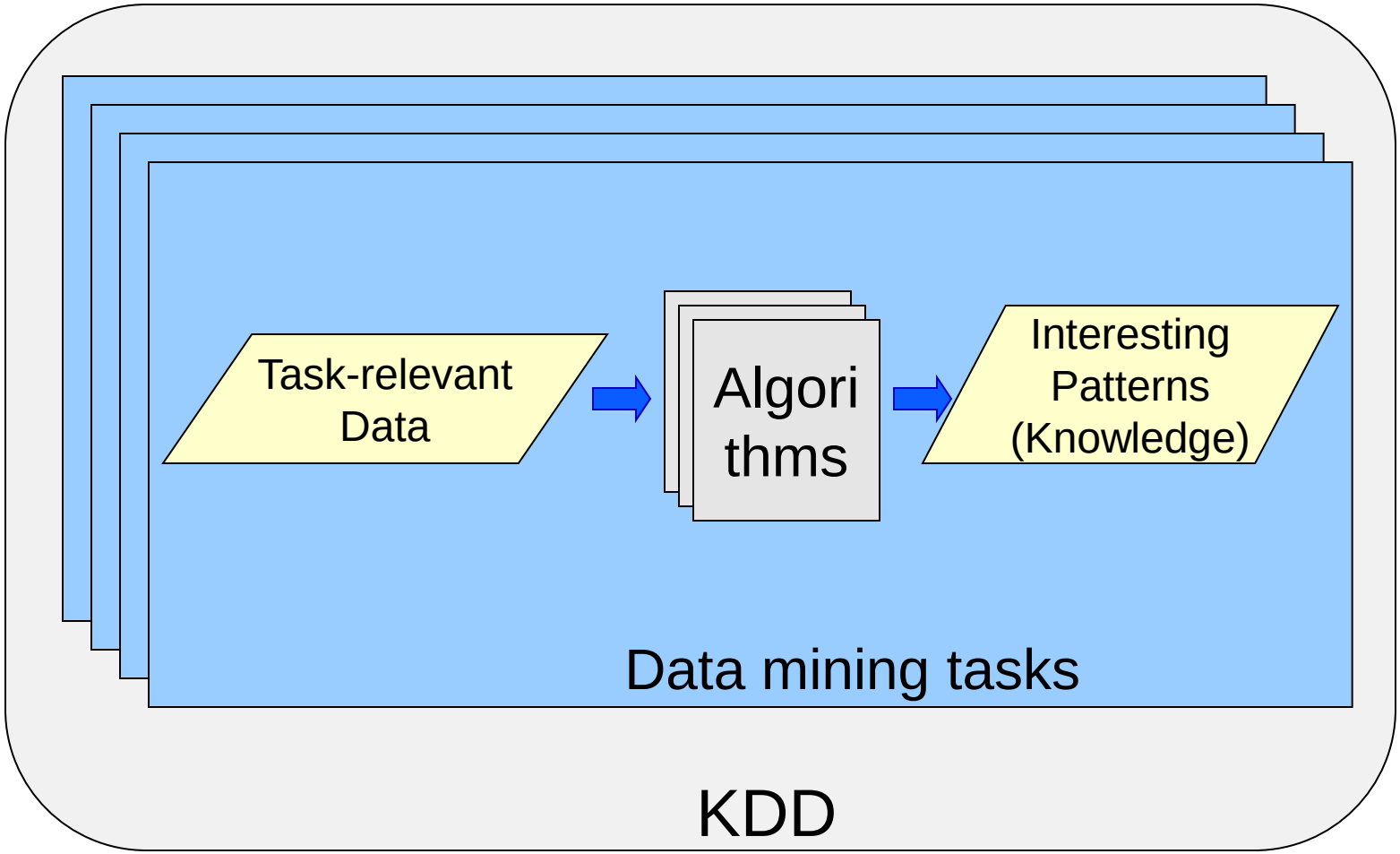
3.2. Data Mining Tasks: Main factors

- Background knowledge:
  - Domain knowledge: finance, education, healthcare,…
  - Supports DM processes: training, evaluating models
- Interestingness measures:
  - With a score/measure, and has a threshold
  - Use for train the model and evaluate the results
  - Different tasks use different measure
  - Needs to be simple, certain, useful and novel

3.2. Data Mining Tasks: Main factors

- Pattern evaluation and knowledge presentation (ref. slide 25, 26 in Sect. 3.1): Rules, tables, reports, charts, graphs, trees, cubes,…

3.2. DATA MINING TASK



3.2. Data Mining Algorithm: Main elements

- 4 main elements constitute a data mining algorithm
  - Model or pattern structures
  - Score function
  - Optimization and search method
  - Data management strategy

3.2. Data Mining Algorithm: Main elements

- **Model or pattern structures**
  - Model: Presents the dataset in a global view
  - Pattern: Presents characteristics of a subset of the dataset (local view), e.g., for some records /objects or satisfy with some variables
  - Structure: a general function where parameters' values are not defined to describe a model or a pattern

⇒ Model structure: a global summary of the dataset

Ex. **Y = aX + b** is a model structure and **Y = 3X + 2** is a specific model defined from the above model structure

=> Pattern structure: a summary of a sub-dataset

Ex. **p(Y>y1 | X>x1) = p1** is a pattern structure and

p(Y>5 | X>10) = 0.5 is a particular pattern

3.2. Data Mining Algorithm: Main elements

- Score function
  - used to examine how effective/good/relevant a model/pattern present a dataset (by score)
  - used to compare different data mining models, methods,…
  - should not be depended on the dataset and easy to be computed. Ex. likelihood, sum of squared errors, misclassification rate,…

3.2. Data Mining Algorithm: Main elements

- Optimization and search methods
  - Objective: To identify the **structures** and **models**, **patterns** (with specific parameters' values) from the datasets that **fit with the expected score function**.
    - State space: A set of discrete states
    - Searching: begin at a particular state (e.g., at a node in the space), searches in the state space until finding a specific state that is **"best" fit with the score function**
    - Methods: Various approaches: greedy strategy, heuristics, revolution algorithms,…

3.2. Data Mining Algorithm: Main elements

- Data management strategy
  - Depending on data size, types,…
    - Small to medium: Load all to the main memory to process
    - Large/big: Stored in disks/distributed systems. Parts are concurrently processed in memory
  - Support for storage, indexing, retrieving
    - Improve the efficiency, scalability,… of the data mining approaches
    - Database technologies can help

Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Data Mining

Course ID: CO3029

Chapter 1: Overview

Part 4

Assoc. Prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

3.3. Data Mining Processes (DMP)

- DMP is iterative and interactive steps starting with raw data and completing with knowledge of interest. It presents…
  - A systematically way to conduct (plan and manage) a KDD project
  - Assure that the KDD project is optimized

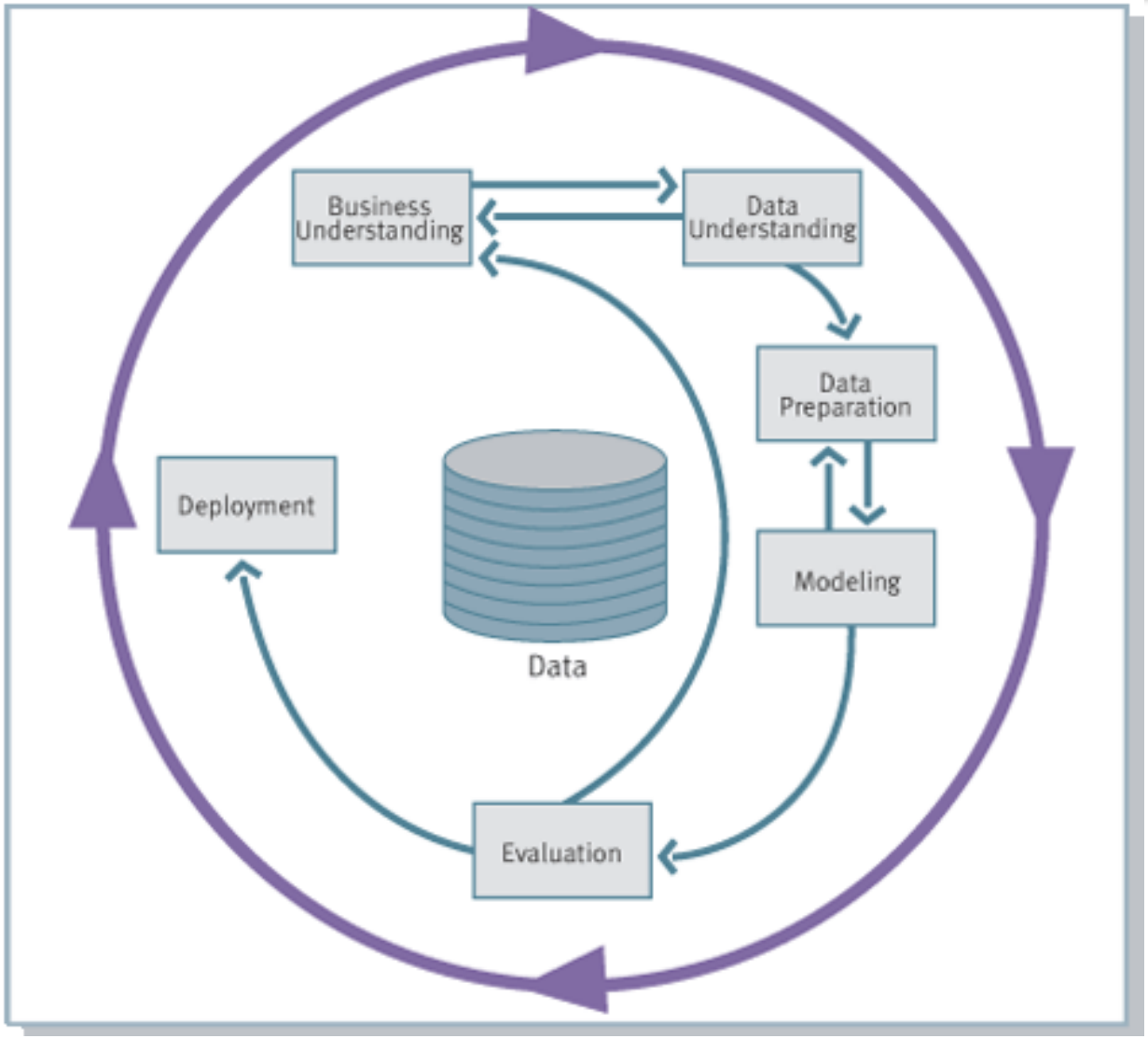
Some standard processes:

- Cross Industry Standard Process for Data Mining (CRISP-DM at www.crisp-dm.org)
- SEMMA (***S***ample, ***E***xplore, ***M***odify, ***M***odel, ***A***ssess) at the SAS Institute

3.3. Data Mining Processes (DMP)

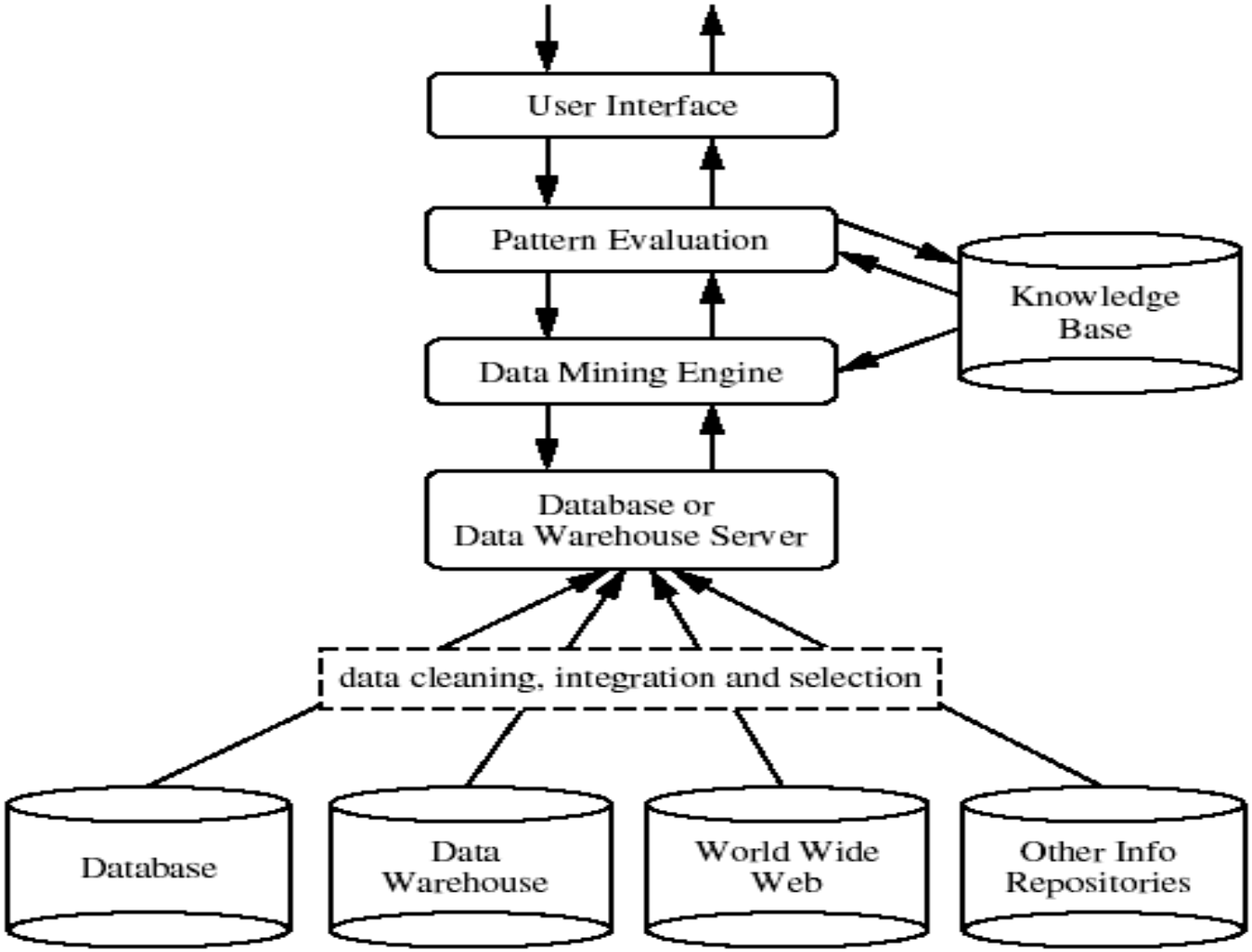
- CRISP-DM
  - Initiated since 09/1996 with more than 200 members
  - Open standard
  - Support industry, applications and tools for KDD
  - Focus on business requirement and technical analysis
  - Provide framework for data mining processes
  - Rich experiences from various application domains and industries

3.3. Data Mining Processes (DMP)



3.4. Data Mining Systems

- A common structure of a DM system



3.4. Data Mining Systems

- **Database, data warehouse**, World Wide Web, and information repositories: Data/information sources used for DM
- **Database hay data warehouse server**: Physical data sources that prepare integrated and relevant data for DM
- **Knowledge base**: Domain/background knowledge
- **Data mining engine**: Conducts DM tasks
- **Pattern evaluation module**: Use interestingness measure (score functions), threshold which can be integrated in the Data mining engine

3.4. Data Mining Systems

- User interface: Support user interaction with the system:
  - To specify data mining tasks, query,…
  - Search for and conduct sophisticated data mining tasks using temporary mining results
  - Verify input data from databases, data warehouse
  - Evaluate mining results
  - Visualize the mined knowledge

3.4. Data Mining Systems

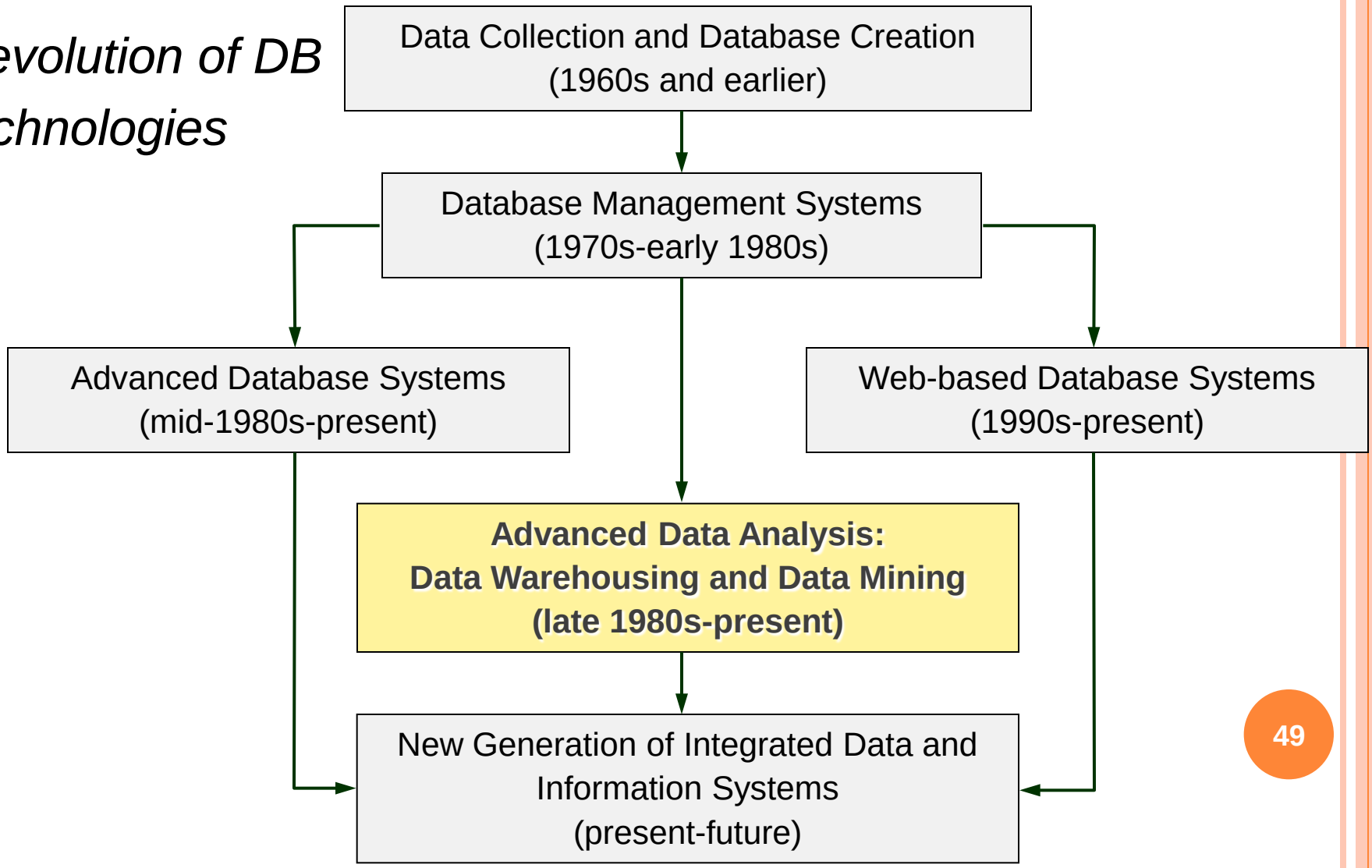
- Features used to examine a DM system
  - Data types
  - Data sources
  - Tasks/Functions and Methods
  - Connecting with data sources: DBs, Data warehouse, WWW, spreadsheets,…
  - Scalabilities, robustness,…
  - Visualization capability

3.4. Data Mining Systems

- Related systems to DM ones
  - Statistical data analysis systems
  - Machine learning systems
  - Information retrieval systems
  - Deductive database systems
  - Database systems
  - Data warehouses
  - …

4. Role of DM

Revolution of DB technologies



4. Roles of DM

- A modern technology in CS, ICT and information management
  - Available everywhere (ubiquitous) and invisible in our life
  - Vast application domains : sales, commerce, bank, finance, insurance, entertainments, healthcares, educations, economics, supply chain management, production, science, tele-comunications, controls,…

5. SUMMARY

- DM/KDD: Extract interest patterns from large DB
- Discovered knowledge must be understandable, useful, nontrivial, valid and evaluable
- Data sources: Various
- DM Tasks: Description, prediction, classification, clustering, association rule mining, co-relation, outlier, trends,…
- 5 factors: Relevant data, expected knowledge, background KN, measures, KN visualization
- 4 elements: model/pattern structure, score function, optimization methods, data management

5. SUMMARY

- 7 steps in KDD: Data cleansing, integration, data selection, data transformation, DM, pattern evaluation, and KN presentation
- DM is the main component in KDD (some time interchangably used)
- Related fieds: DB technologies, statistics, machine learning, computer science, visualization,…

REFENCES

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

Q&A

quangtran@hcmut.edu.vn