
ĐẠI HỌC
BÁCH KHOA
25 YEARS ANNIVERSARY
SOICT
ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

1


ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
Introduction to Data Mining

2

Introduction

- Data is growing at a phenomenal rate
- Users expect more sophisticated information
- How?

UNCOVER HIDDEN INFORMATION

DATA MINING

Data Mining Definition

- Finding hidden information in a database
- Fit data to a model
- Similar terms
  - Exploratory data analysis
  - Data driven discovery
  - Deductive learning

Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preference – Technique to choose the best model
- Search – Technique to search the data
  - "Query"

Database Processing vs. Data Mining Processing

- Query
  - Well defined
  - SQL

■ Data
– Operational data

■ Output
– Precise
– Subset of database

- Query
  - Poorly defined
  - No precise query language

■ Data
– Not operational data

■ Output
– Fuzzy
– Not a subset of database

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Basic Data Mining Tasks

- Classification maps data into predefined groups or classes
  - Supervised learning
  - Prediction
  - Regression
- Clustering groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

Basic Data Mining Tasks (cont'd)

- Link Analysis uncovers relationships among data.
  - Affinity Analysis
  - Association Rules
  - Sequential Analysis determines sequential patterns.

CLASSIFICATION

- Assign data into predefined groups or classes.

But it isn't Magic

- You must know what you are looking for
- You must know how to look for you

Suppose you knew that a specific cave had gold:

What would you look for?

How would you look for it?

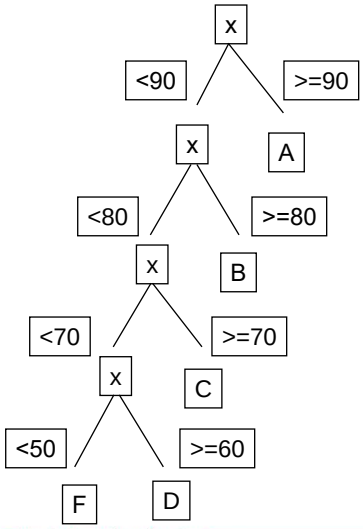
Might need an expert miner

"If it looks like a duck, walks like a duck, and quacks like a duck, then it's a duck."

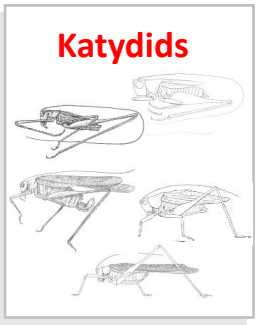
"If it looks like a terrorist, walks like a terrorist, and quacks like a terrorist, then it's a terrorist."

Description	Behavior	Associations
Classification (Profiling)	Clustering (Similarity)	Link Analysis

Classification Ex: Grading



Given a collection of annotated data. (in this case 5 instances of Katydids and five of Grasshoppers), decide what type of insect the unlabeled example is.



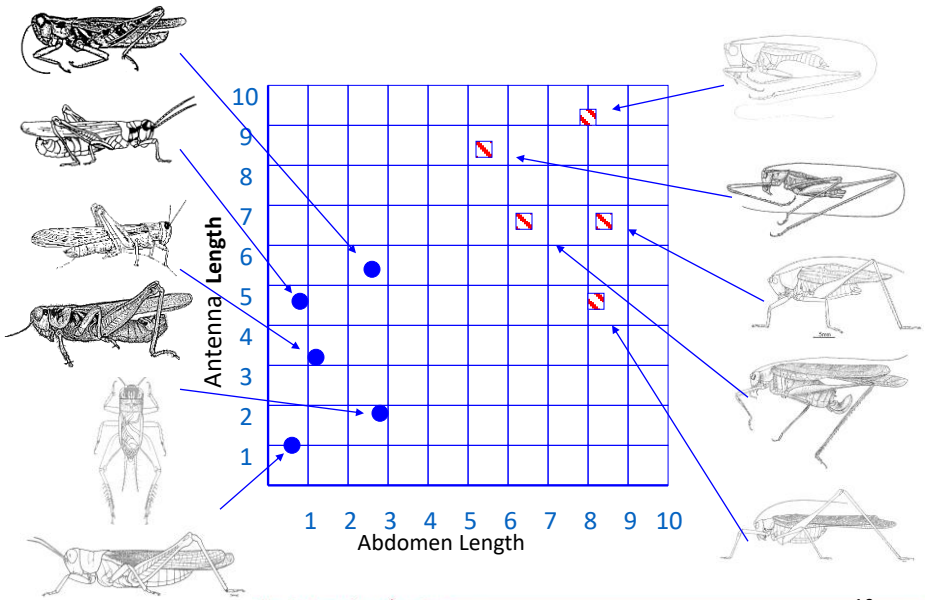
The classification problem can now be expressed as:

Given a training database predict the class label of a previously unseen instance

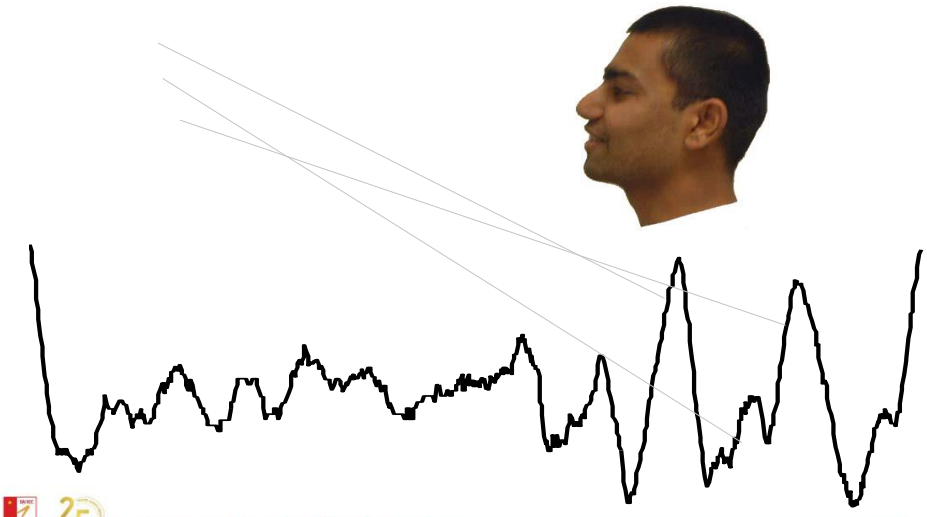
Insect ID	Abdomen Length	Antennae Length	Insect Class
1	2.7	5.5	Grasshopper
2	8.0	9.1	Katydid
3	0.9	4.7	Grasshopper
4	1.1	3.1	Grasshopper
5	5.4	8.5	Katydid
6	2.9	1.9	Grasshopper
7	6.1	6.6	Katydid
8	0.5	1.0	Grasshopper
9	8.3	6.6	Katydid
10	8.1	4.7	Katydid

previously unseen instance =

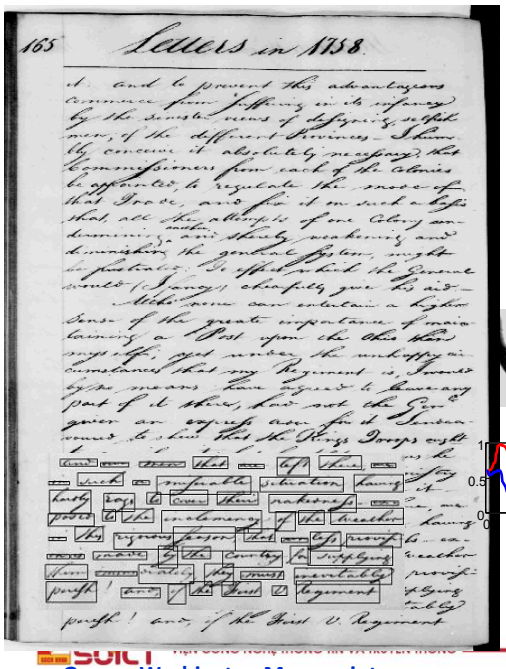
11	5.1	7.0	???????



Facial Recognition

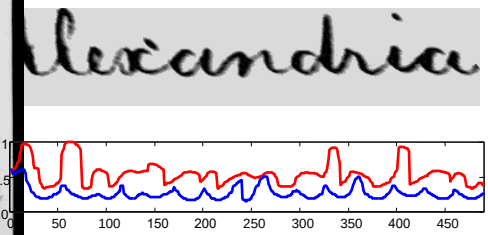


17



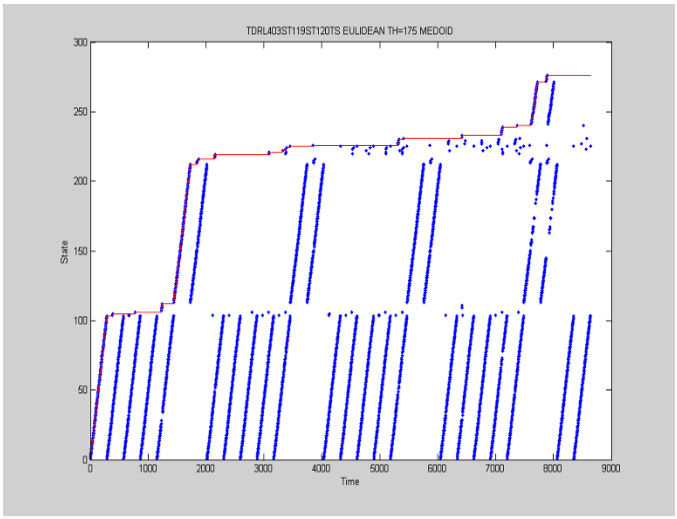
George Washington Manuscript

Handwriting Recognition



18

Anomaly Detection



But if the largest manhunt in the Washington area since Abraham Lincoln's assassination shows anything, it serves as a reminder that criminal profiling is more art than science. And that multiple killers don't always fit neatly into statistical profiling categories.

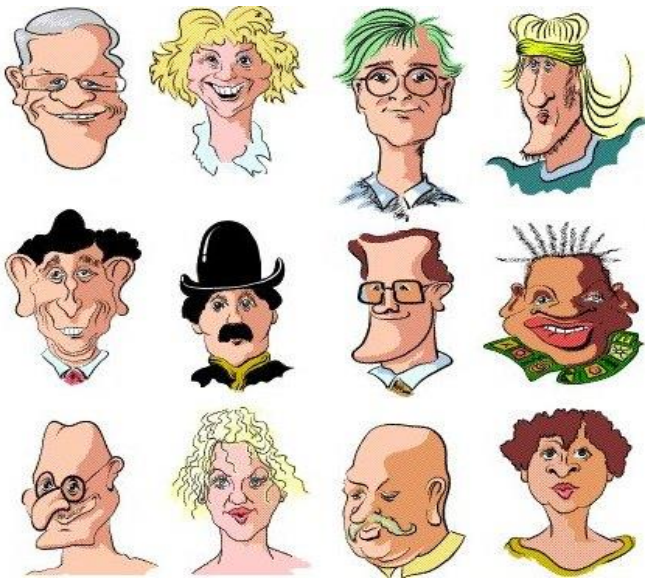
Profiling the characteristics of a criminal, once dismissed as conjecture, is widely used today to help investigators solve hard-to-crack cases.

"We are going to see new

"What people should know bout profiling is that there's o magic to it," says James lan Fox, a criminologist at ortheastern University. "Its asically educated hunches."

CLUSTERING

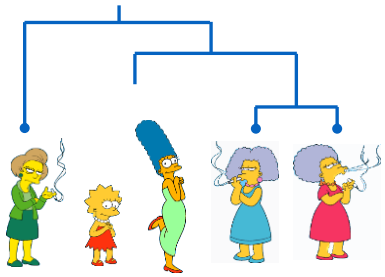
- Partition data into previously undefined groups.



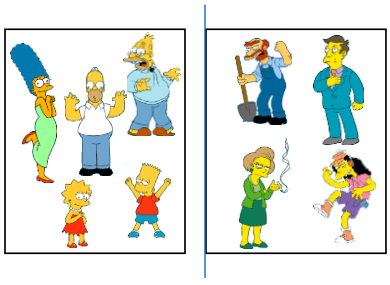
What is Similarity?

Two Types of Clustering

Hierarchical



Partitional



Hierarchical Clustering Example Iris Data Set

Sentosa

Versicolor

Virginica

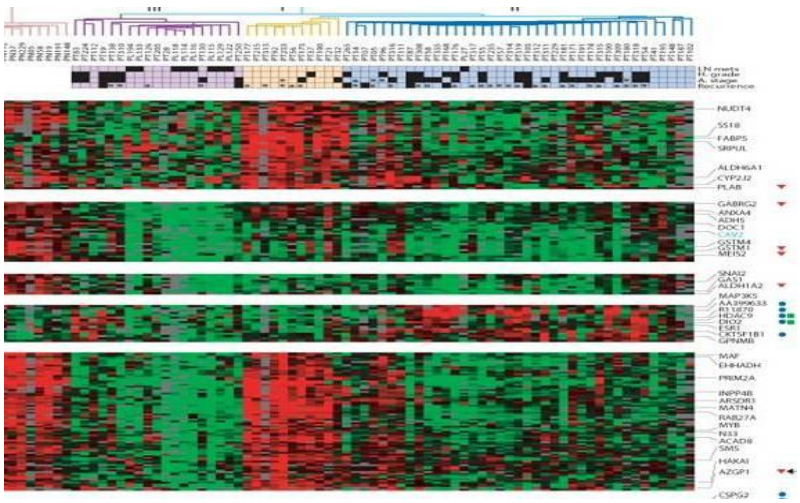
http://www.time.com/time/magazine/article/0,9171,1541283,00.html

Microarray Data Analysis

- Each probe location associated with gene
- Color indicates degree of gene expression
- Compare different samples (normal/disease)
- Track same sample over time
- Questions
  - Which genes are related to this disease?
  - Which genes behave in a similar manner?
  - What is the function of a gene?
- Clustering
  - Hierarchical
  - K-means

Microarray Data - Clustering

"Gene expression profiling identifies clinically relevant subtypes of prostate cancer" *Proc. Natl. Acad. Sci. USA*, Vol. 101, Issue 3, 811-816, January 20, 2004

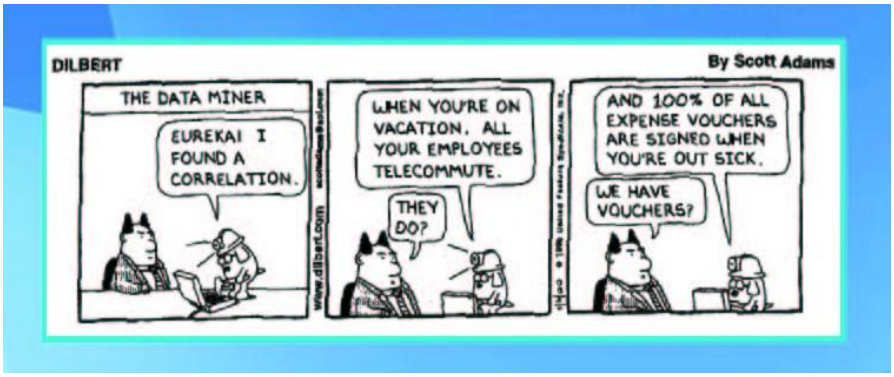


ASSOCIATION RULES/ LINK ANALYSIS

- Find relationships between data

ASSOCIATION RULES EXAMPLES

- People who buy diapers also buy beer
- If gene A is highly expressed in this disease then gene A is also expressed
- Relationships between people
- Book Stores
- Department Stores
- Advertising
- Product Placement



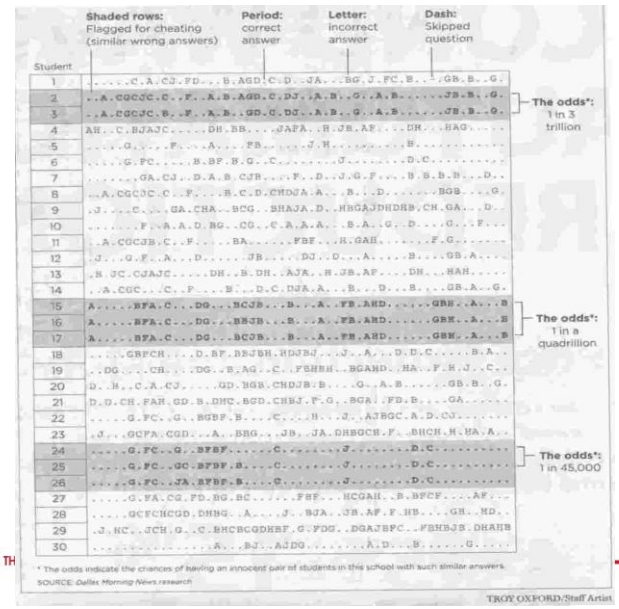
Data Mining Introductory and Advanced Topics, by Margaret H. Dunham, Prentice Hall, 2003.
DILBERT reprinted by permission of United Feature Syndicate, Inc.

31

The Dallas Morning News

ding Newspaper $1.50 Dallas, Texas, Sunday, June 3, 2007 dallasnews.com

Joshua Benton and Holly K. Hacker, "At Charters, Cheating's off the Charts:, *Dallas Morning News*, June 4, 2007.



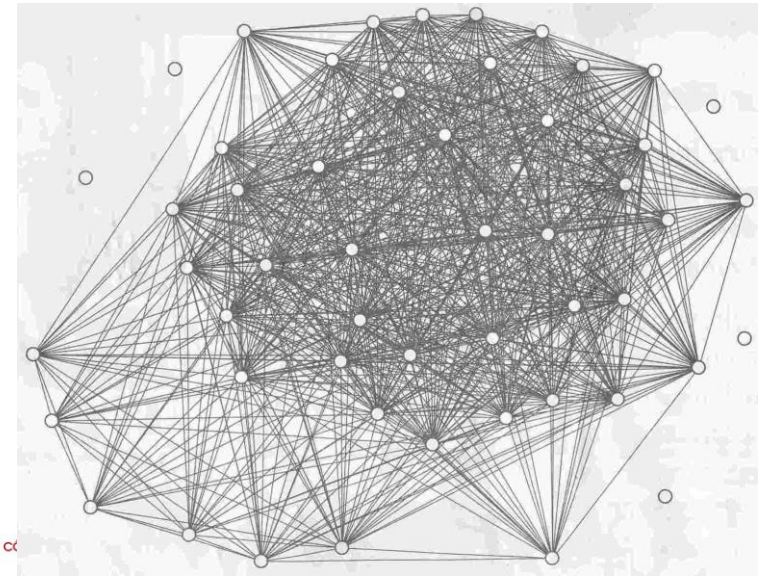
32

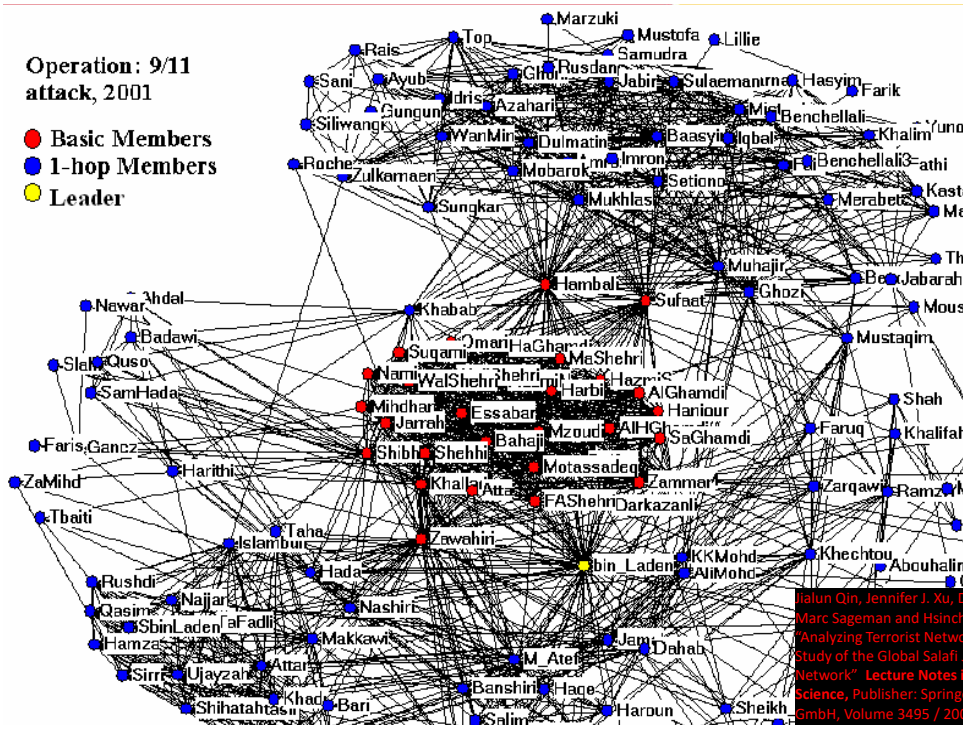
No/Little Cheating

Joshua Benton and Holly K. Hacker, “At Charters, Cheating’s off the Charts:, *Dallas Morning News*, June 4, 2007.

Rampant Cheating

Joshua Benton and Holly K. Hacker, “At Charters, Cheating’s off the Charts:, *Dallas Morning News*, June 4, 2007.





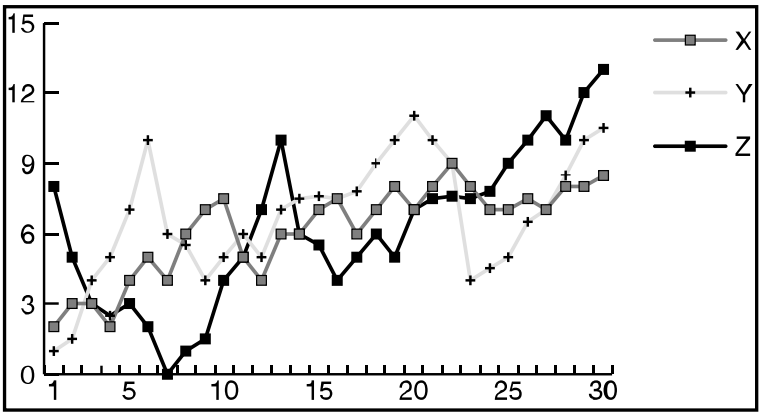
35

Ex: Stock Market Analysis

- Example: Stock Market
- Predict future values
- Determine similar patterns over time
- Classify behavior

36

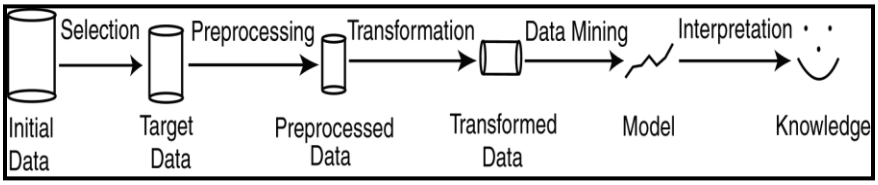
Ex: Stock Market Analysis



Data Mining vs. KDD

- Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data.
- Data Mining: Use of algorithms to extract the information and patterns derived by the KDD process.

KDD Process



Modified from [FPSS96C]

- Selection: Obtain data from various sources.
- Preprocessing: Cleanse data.
- Transformation: Convert to common format. Transform to new format.
- Data Mining: Obtain desired results.
- Interpretation/Evaluation: Present results to user in meaningful manner.

KDD Process Ex: Web Log

- Selection:
  - Select log data (dates and locations) to use
- Preprocessing:
  - Remove identifying URLs; Remove error logs
- Transformation:
  - Sessionize (sort and group)
- Data Mining:
  - Identify and count patterns; Construct data structure
- Interpretation/Evaluation:
  - Identify and display frequently accessed sequences.
- Potential User Applications:
  - Cache prediction
  - Personalization

Related Topics

- Databases
- OLTP
- OLAP
- Information Retrieval

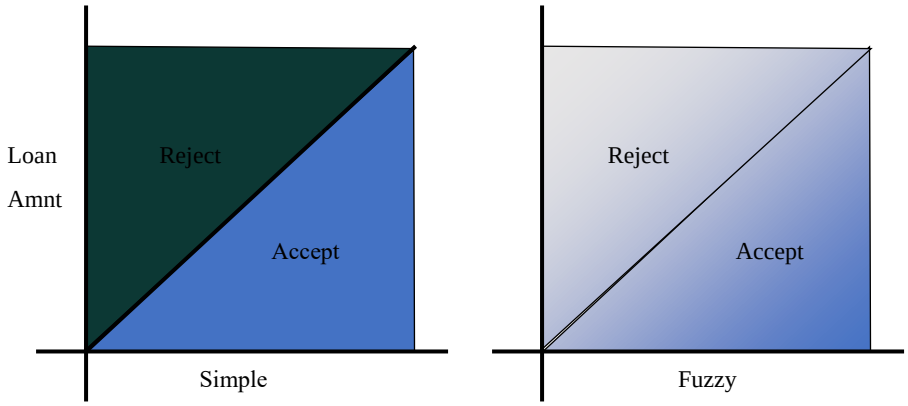
DB & OLTP Systems

- Schema
  - (ID,Name,Address,Salary,JobNo)
- Data Model
  - ER
  - Relational
- Transaction
- Query:

```
SELECT Name
FROM T
WHERE Salary > 100000
```

DM: Only imprecise queries

Classification/Prediction is Fuzzy



Information Retrieval

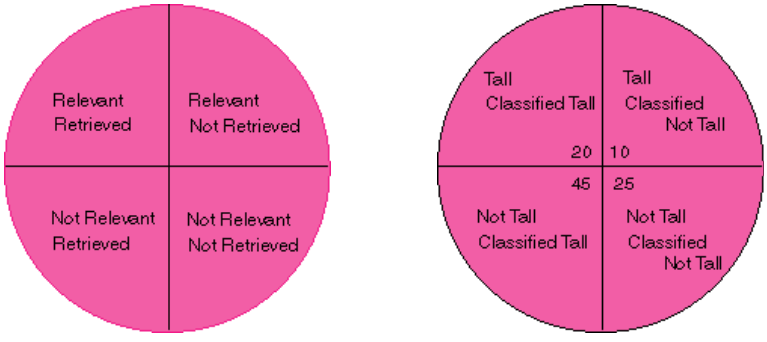
- ***Information Retrieval (IR):*** retrieving desired information from textual data.
- Library Science
- Digital Libraries
- Web Search Engines
- Traditionally keyword based
- Sample query:
  Find all documents about "data mining".

DM: Similarity measures;
Mine text/Web data.

Information Retrieval (cont'd)

- ***Similarity:*** measure of how close a query is to a document.
- Documents which are "close enough" are retrieved.
- Metrics:
  - ***Precision*** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$
  - ***Recall*** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$

IR Query Result Measures and Classification



IR Classification

OLAP

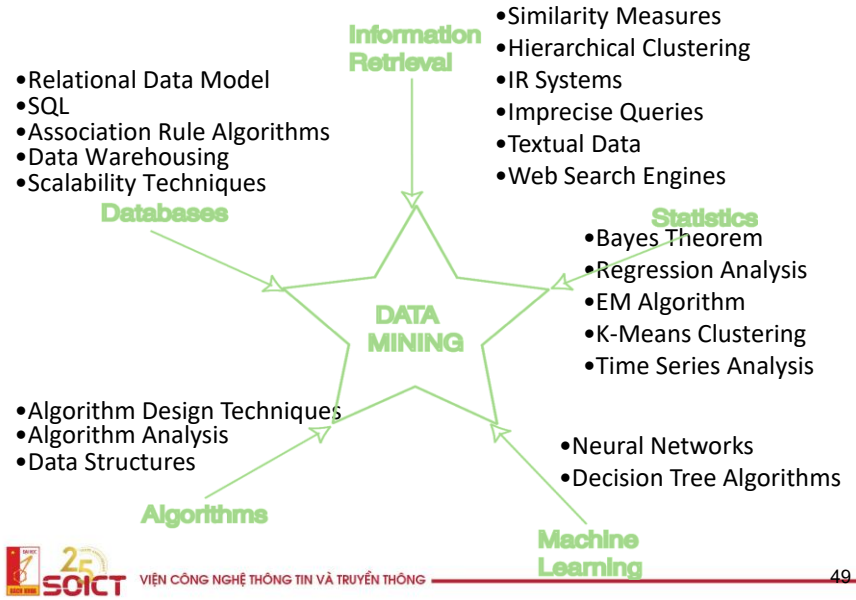
- ***Online Analytic Processing (OLAP):*** provides more complex queries than OLTP.
- ***OnLine Transaction Processing (OLTP):*** traditional database/transaction processing.
- Dimensional data; cube view
- Visualization of operations:
  - ***Slice:*** examine sub-cube.
  - ***Dice:*** rotate cube to look at another dimension.
  - ***Roll Up/Drill Down***

DM: May use OLAP queries.

DM vs. Related Topics

Area	Query	Data	Results	Output
DB/OLTP	Precise	Database	Precise	DB Objects or Aggregation
IR	Precise	Documents	Vague	Documents
OLAP	Analysis	Multidimensional	Precise	DB Objects or Aggregation
DM	Vague	Preprocessed	Vague	KDD Objects

Data Mining Development



KDD Issues

- Human Interaction
- Overfitting
- Outliers
- Interpretation
- Visualization
- Large Datasets
- High Dimensionality

Overfitting

- Suppose we want to predict whether an individual is short, medium, or tall. What is wrong with this data?

Name	Gender	Height	Output
Mary	F	1.6	Short
Maggie	F	1.9	Medium
Martha	F	1.88	Medium
Stephanie	F	1.7	Short
Bob	M	1.85	Medium
Kathy	F	1.6	Short
George	M	1.7	Short
Debbie	F	1.8	Medium
Todd	M	1.95	Medium
Kim	F	1.9	Medium
Amy	F	1.8	Medium
Wynette	F	1.75	Medium

KDD Issues (cont'd)

- Multimedia Data
- Missing Data
- Irrelevant Data
- Noisy Data
- Changing Data
- Integration
- Application

WARNING

- With data mining you don't always know what you are looking for.
- There is not one right answer.
- The data you are using is noisy
- Data Mining is a very applied discipline.
- A data mining course provides you tools to use to analyze data.
- Experience provides you knowledge of how to use these tools.

http://ieeexplore.ieee.org/iel5/6/32236/01502526.pdf?tp=&arnumber=1502526&isnumber=32236

Social Implications of DM

- Privacy
- Profiling
- Unauthorized use
- Invalid results and claims

Data Mining Metrics

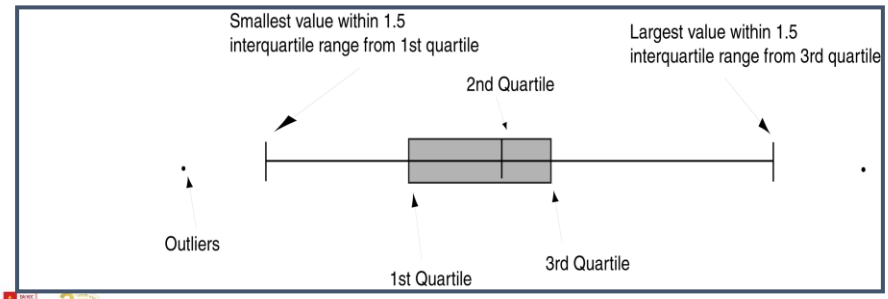
- Usefulness
- Return on Investment (ROI)
- Accuracy
- …
- Space/Time

Visualization Techniques

- Graphical
- Geometric
- Icon-based
- Pixel-based
- Hierarchical
- Hybrid

Models Based on Summarization

- Visualization: Frequency distribution, mean, variance, median, mode, etc.
- Box Plot:



DM Tools

- XLMiner – Easy addin to Excel

http://www.solver.com/xlminer/index.html

- Weka – Open Source; Visualization, Functionality, Interface

http://www.cs.waikato.ac.nz/ml/weka/

- SAS (JMP) – Commercial Product
- SPSS – Commercial Product
- MATLAB – Statistical/Math Applications
- R – Programming

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

62

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict

63