Knowledge Discovery and Data Mining

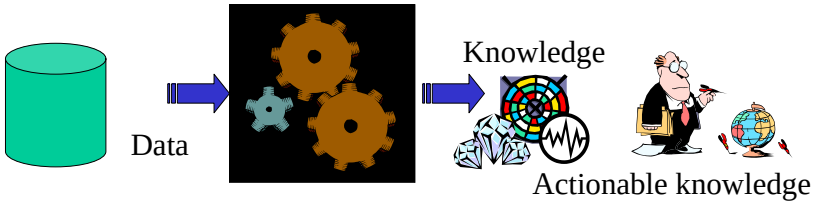
Fall 2007

Dr. Osmar R. Zaïane

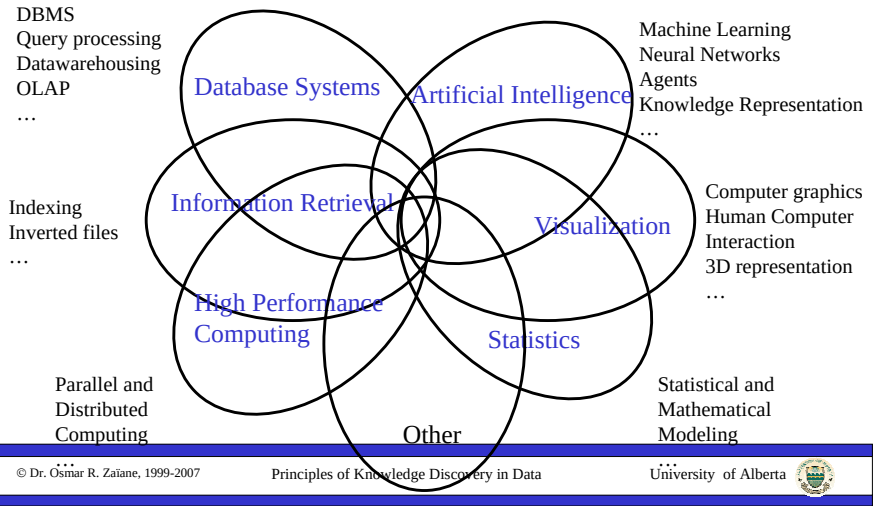
University of Alberta

Data Mining in a nutshell

Extract interesting knowledge
(rules, regularities, patterns, constraints)
from data in large collections.



KDD at the Confluence of Many Disciplines



Who Am I?

Database Laboratory

蔡頤安

زيان

UNIVERSITY OF ALBERTA

Osmar R. Zaïane, Ph.D.
Associate Professor
Department of Computing Science

221 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: zaiane@cs.ualberta.ca
http://www.cs.ualberta.ca/~zaiane/

PhD on *Web Mining* & *Multimedia Mining* With **Dr. Jiawei Han** at Simon Fraser University, Canada

Research Interests:
Data Mining,
Web Mining,
Multimedia Mining,
Data Visualization,
Information Retrieval.

Applications:
Analytic Tools,
Adaptive Systems,
Intelligent Systems,
Diagnostic and Categorization,
Recommender Systems

Achievements:
(in last 6 years):
2 PhD and 18 MSc,
90+ publications, ICDM PC co-chair 2007, ADMA co-Chair 2007, WEBKDD and MDM/KDD co-chair (2000 to 2003)
Currently: 3 PhD + 1 MSc

Who Am I?

Database Laboratory

Research Activities

Pattern Discovery for Intelligent Systems

Since 1999

蔡頤安

زيان

UNIVERSITY OF ALBERTA

Osmar R. Zaïane, Ph.D.
Associate Professor
Department of Computing Science

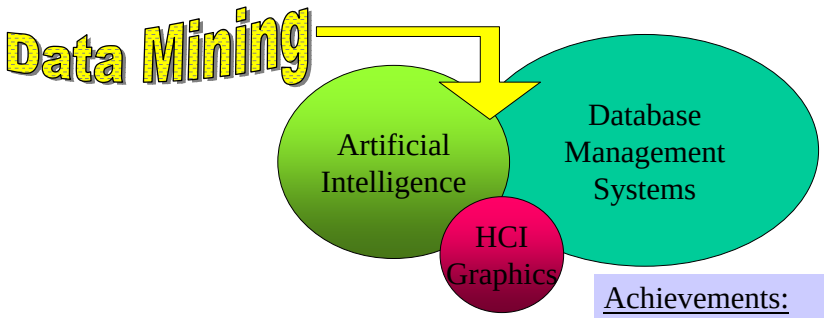
352 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: **zaiane@cs.ualberta.ca**
http://www.cs.ualberta.ca/~zaiane/

Currently:
4 graduate students
(3 Ph.D., 1 M.Sc.)

Past 6 years	2001	2002	2003	2004	2005	2006	Total
MSc	5	5	1	4	1	2	**18**
PhD	0	0	0	1	0	1	**2**
Total	**5**	**5**	**1**	**5**	**1**	**3**	**20**

Where I Stand



Research Interests:
Data Mining,
Web Mining,
Multimedia Mining,
Data Visualization,
Information Retrieval.

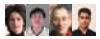
Applications:
Analytic Tools,
Adaptive Systems,
Intelligent Systems,
Diagnostic and Categorization,
Recommender Systems

Achievements:
90+ publications,
IEEE-ICDM **PC-chair** & ADMA **chair -2007**
WEBKDD and MDM/KDD co-chair (2000 to 2003),
WEBKDD'05 co-chair,
Associate Editor for ACM-SIGKDD Explorations

Thanks To

Without my students my research work wouldn't have been possible.

Current: (No particular order)



- Maria-Luiza Antonie
- Jiyang Chen
- Andrew Foss
- Seyed-Vahid Jazayeri

Past: (No particular order)

- Stanley Oliveira
- Yang Wang
- Lisheng Sun
- Jia Li
- Alex Strilets
- William Cheung
- Andrew Foss
- Yue Zhang
- Chi-Hoon Lee
- Weinan Wang
- Ayman Ammoura
- Hang Cui
- Jun Luo
- Jiyang Chen
- Yuan Ji
- Yi Li
- Yaling Pei
- Yan Jin
- Mohammad El-Hajj
- Maria-Luiza Antonie

Knowledge Discovery and Data Mining

Class and Office Hours

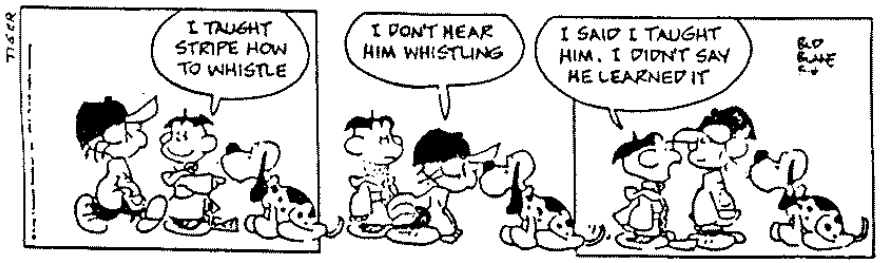
Class:
Tuesday and Thursdays from 9:30 to 10:50

Office Hours:
Tuesdays from 13:00 to 14:00
By mutually agreed upon appointment:
E-mail: zaiane@cs.ualberta.ca
Tel: 492 2860 - Office: ATH 3-52

But I prefer Class:
Once a week from 9:00 to 11:50 (with one break)

Course Requirements

- Understand the basic concepts of database systems
- Understand the basic concepts of artificial intelligence and machine learning
- Be able to develop applications in C/C++ and/or Java



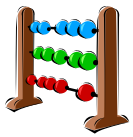
Course Objectives

To provide an introduction to knowledge discovery in databases and complex data repositories, and to present basic concepts relevant to real data mining applications, as well as reveal important research issues germane to the knowledge discovery domain and advanced mining applications.

Students will understand the fundamental concepts underlying knowledge discovery in databases and gain hands-on experience with implementation of some data mining algorithms applied to real world cases.

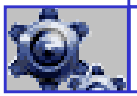
Evaluation and Grading

There is no final exam for this course, but there are assignments, presentations, a midterm and a project.
I will be evaluating all these activities out of 100% and give a final grade based on the evaluation of the activities.
The midterm is either a take-home exam or an oral exam.



- Assignments 20% (2 assignments)
- Midterm 25%
- Project 39%
  - Quality of presentation + quality of report and proposal + quality of demos
  - Preliminary project demo (week 11) and final project demo (week 15) have the same weight (could be week 16)
- Class presentations 16%
  - Quality of presentation + quality of slides + peer evaluation
- A+ will be given only for outstanding achievement.

Projects



Choice	Deliverables
Implement data mining project	Project proposal + project pre-demo + final demo + project report

Examples and details of data mining projects will be posted on the course web site.

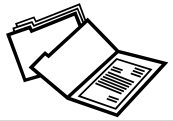
Assignments

1- Competition in one algorithm implementation (in C/C^{++})
2- Devising Exercises with solutions

More About Projects

Students should write a project proposal (1 or 2 pages).

- project topic;
- implementation choices;
- Approach, references;
- schedule.



All projects are demonstrated at the end of the semester.
December 11-12 to the whole class.
Preliminary project demos are private demos given to the instructor on **week November 19**.

Implementations: C/C^{++} or Java,
OS: Linux, Window XP/2000 , or other systems.

More About Evaluation

Re-examination.

None, except as per regulation.

Collaboration.

Collaborate on assignments and projects, etc; do not merely copy.

Plagiarism.

Work submitted by a student that is the work of another student or any other person is considered plagiarism. Read **Sections 26.1.4** and **26.1.5** of the University of Alberta calendar. Cases of plagiarism are immediately referred to the Dean of Science, who determines what course of action is appropriate.

About Plagiarism

Plagiarism, cheating, misrepresentation of facts and participation in such offences are viewed as serious academic offences by the University and by the Campus Law Review Committee (CLRC) of General Faculties Council.
Sanctions for such offences range from a reprimand to suspension or expulsion from the University.

Notes and Textbook

Course home page:
http://www.cs.ualberta.ca/~zaiane/courses/cmput695/

We will also have a mailing list and newsgroup for the course.

No Textbook but recommended books:

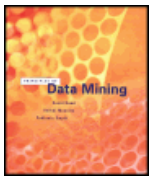
Data Mining: Concepts and Techniques
Jiawei Han and Micheline Kamber
Morgan Kaufmann Publisher

2006
ISBN 1-55860-901-6
800 pages

2001
ISBN 1-55860-489-8
550 pages

http://www-faculty.cs.uiuc.edu/~hanj/bk2/

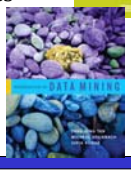
Other Books



- Principles of Data Mining
  - David Hand, Heikki Mannila, Padhraic Smyth,
    MIT Press, 2001, ISBN 0-262-08290-X, 546 pages

- Data Mining: Introductory and Advanced Topics
  - Margaret H. Dunham,
    Prentice Hall, 2003, ISBN 0-13-088892-3, 315 pages

- Dealing with the data flood: Mining data, text and multimedia
  - Edited by Jeroen Meij,
    SST Publications, 2002, ISBN 90-804496-6-0, 896 pages



- Introduction to Data Mining
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar
    Addison Wesley, ISBN: 0-321-32136-7, 769 pages

Course Schedule

(**Tentative**, subject to changes)

There are 13 weeks from Sept 6th to December 4th.

Week 1:	Sept 6	: Introduction to Data Mining
Week 2:	Sept 11-13	: Association Rules
Week 3:	Sept 18-20	: Association Rules (advanced topics)
Week 4:	Sept 25-27	: Sequential Pattern Analysis
Week 5:	Oct 2-4	: Classification (Neural Networks)
Week 6:	Oct 9-11	: Classification (Decision Trees and +)
Week 7:	Oct 16-18	: Data Clustering
Week 8:	Oct 23-25	: Outlier Detection
Week 9:	Oct 30-Nov 1	: Data Clustering in subspaces
Week 10:	Nov 6-8	: Contrast sets + Web Mining
Week 11:	Nov 13-15	: Web Mining + Class Presentations
Week 12:	Nov 20-22	: Class Presentations
Week 12:	Nov 27-29	: Class Presentations
Week 13:	Dec 4	: Class Presentations
Week 15:	Dec 11	: Project Demos

Due dates
-Midterm
week 8
-Assignment 1
week 6
-Assignment 2
variable dates

Course Content

- Introduction to Data Mining
- Association analysis
- Sequential Pattern Analysis
- Classification and prediction
- Contrast Sets
- Data Clustering
- Outlier Detection
- Web Mining
- Other topics if time permits (spatial data, biomedical data, etc.)

Let's do some Data Mining!

For those of you who watch what you eat...
Here's the final word on nutrition and health. It's a relief to know the truth after all those conflicting medical studies.

- **The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.**
- **The Mexicans eat a lot of fat and suffer fewer heart attacks than the British or Americans.**
- **The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans**
- **The Italians drink excessive amounts of red wine and suffer fewer heart attacks than the British or Americans.**
- **The Germans drink a lot of beer and eat lots of sausages and fats and suffer fewer heart attacks than the British or Americans.**

CONCLUSION:
Eat and drink what you like. Speaking English is apparently what kills you.

Quick Overview of some Data Mining Operations

Association Rules
Clustering
Classification
Outlier Detection

What Is Association Mining?

- **Association rule mining searches for relationships between items in a dataset**:
  - Finding association, correlation, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.
  - Rule form: "**Body → Head [support, confidence]**".
- **Examples:**
  - buys(x, "bread") → buys(x, "milk") [0.6%, 65%]
  - major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [1%, 75%]

Basic Concepts

A transaction is a set of items: $T=\{i_a, i_b,\ldots i_t\}$

$T \subset I$, where I is the set of all possible items $\{i_1, i_2,\ldots i_n\}$

D, the task relevant data, is a set of transactions.

An association rule is of the form:
$P \rightarrow Q$, where $P \subset I$, $Q \subset I$, and $P \cap Q = \varnothing$

$P \rightarrow Q$ holds in D with support s
and
$P \rightarrow Q$ has a confidence c in the transaction set D.

Support($P \rightarrow Q$) = Probability($P \cup Q$)
Confidence($P \rightarrow Q$)=Probability(Q/P)

Association Rule Mining

Frequent Itemset Mining → FIM → 1 → abc

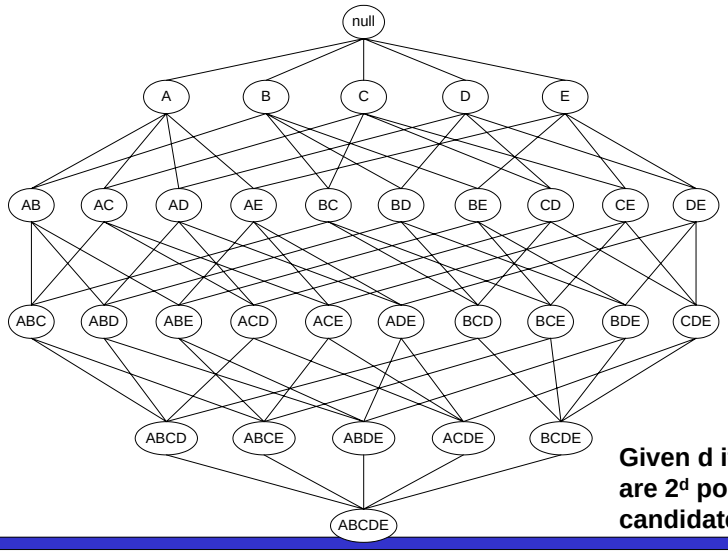
Association Rules Generation → 2 → ab→c, b→ac

Bound by a **support** threshold

Bound by a **confidence** threshold

- Frequent itemset generation is still computationally expensive

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Frequent Itemset Generation

- Brute-force approach (Basic approach):
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database

Transactions

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

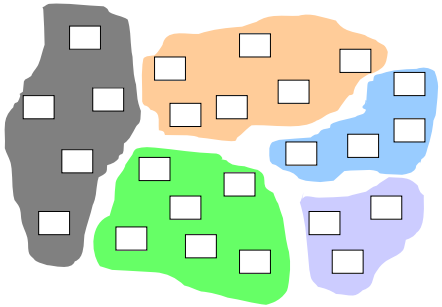
N (transactions), w (items per transaction)

List of Candidates — M

  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => **Expensive since M = 2^d !!!**

Obviously not the right way to do it.

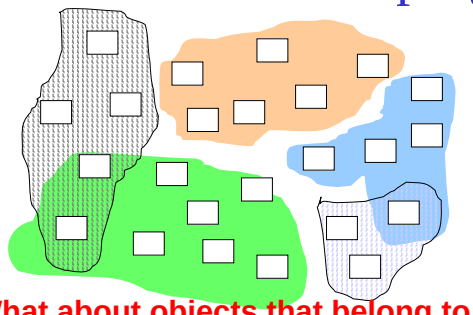
Grouping



Grouping
Clustering
Partitioning

- **We need a notion of similarity or closeness (what features?)**
- **Should we know apriori how many clusters exist?**
- **How do we characterize members of groups?**
- **How do we label groups?**

Grouping

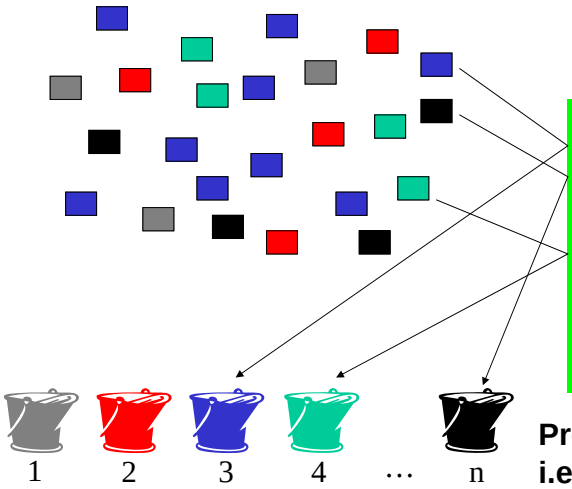


Grouping
Clustering
Partitioning

What about objects that belong to different groups?

- **We need a notion of similarity or closeness (what features?)**
- **Should we know apriori how many clusters exist?**
- **How do we characterize members of groups?**
- **How do we label groups?**

Classification



Classification
Categorization

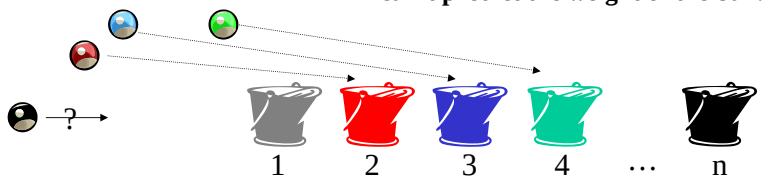
Predefined buckets
i.e. known labels

What is Classification?

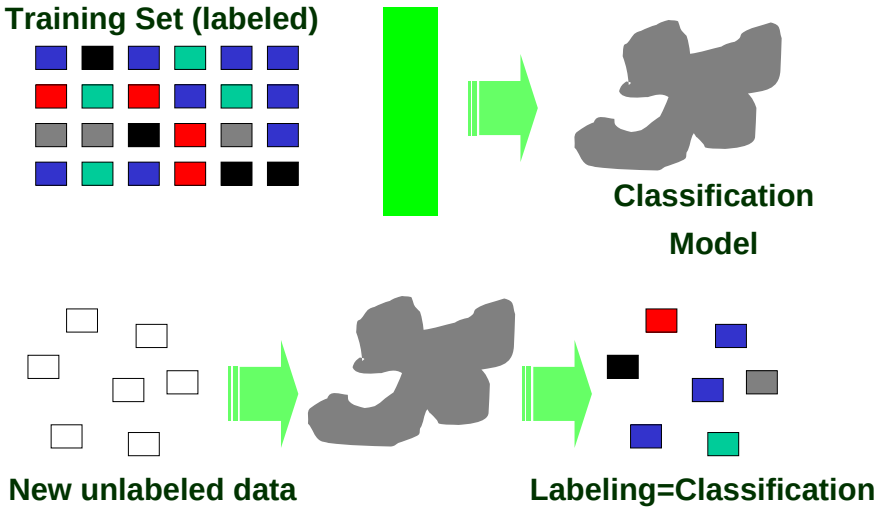
The goal of data classification is to organize and categorize data in distinct classes.

- A model is first created based on the data distribution.
- The model is then used to classify new data.
- Given the model, a class can be predicted for new data.

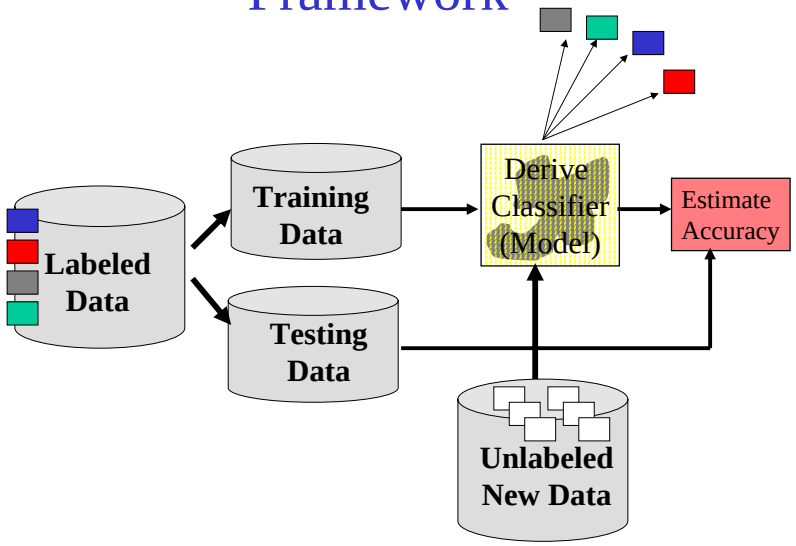
With classification, I can predict in which bucket to put the ball, but I can't predict the weight of the ball.



Classification = Learning a Model



Framework



Classification Methods

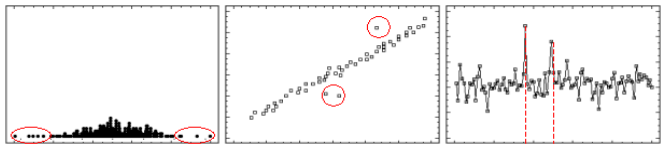
- Decision Tree Induction
- Neural Networks
- Bayesian Classification
- K-Nearest Neighbour
- Support Vector Machines
- Associative Classifiers
- Case-Based Reasoning
- Genetic Algorithms
- Rough Set Theory
- Fuzzy Sets
- Etc.

Outlier Detection

- To find exceptional data in various datasets and uncover the implicit patterns of rare cases
- Inherent variability - reflects the natural variation
- Measurement error (inaccuracy and mistakes)
- Long been studied in statistics
- An active area in data mining in the last decade
- Many applications
  - Detecting credit card fraud
  - Discovering criminal activities in E-commerce
  - Identifying network intrusion
  - Monitoring video surveillance
  - …

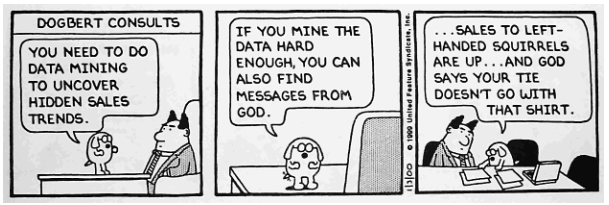
Outliers are Everywhere

- Data values that appear inconsistent with the rest of the data.
- Some types of outliers



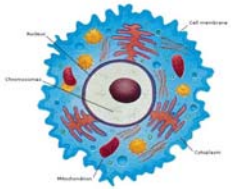
(a) Outliers w.r.t. a distribution (b) Outliers w.r.t. a pattern (c) time series outliers

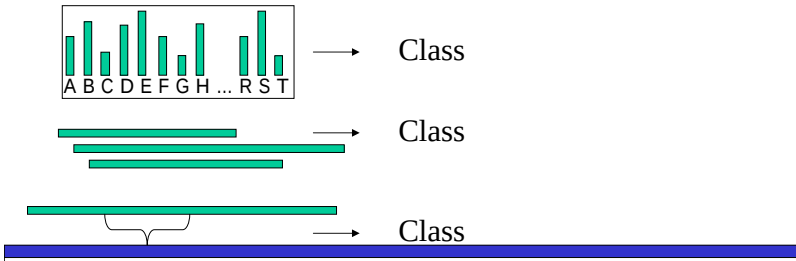
- We will see:
  - Statistical methods;
  - distance-based methods;
  - density-based methods;
  - resolution-based methods, etc.



Protein Localization

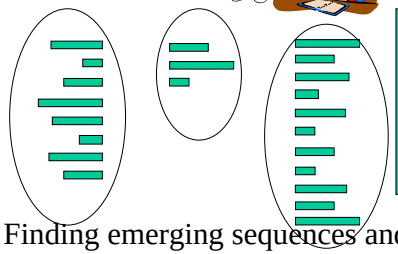
Brassica Napus (Canola)



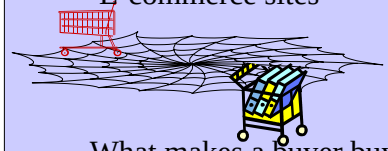


Contrasting Sequence Sets

Collaborative e-learning



E-commerce sites



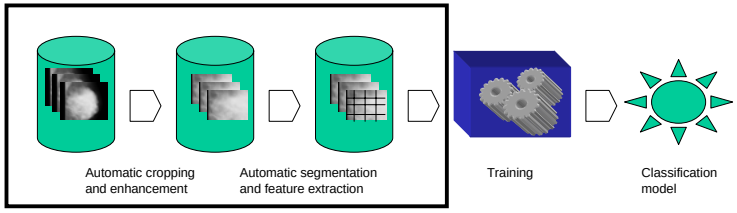
What makes a buyer buy & a non-buyer leave empty handed?

What contrasts genes or proteins?

Finding emerging sequences and contrasting the sets of sequences would give insight about what behaviour leads to success and what doesn't.

Breast Cancer

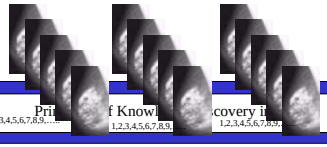
Mammography



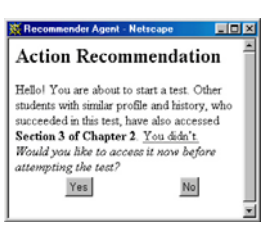
Transaction (IID, **class**, F_1, F_2, F_3, … F_f)

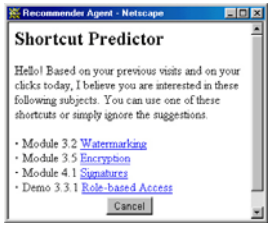
$F_\alpha \wedge F_\beta \wedge F_\gamma \wedge \ldots \wedge F_\delta \rightarrow$ **class**

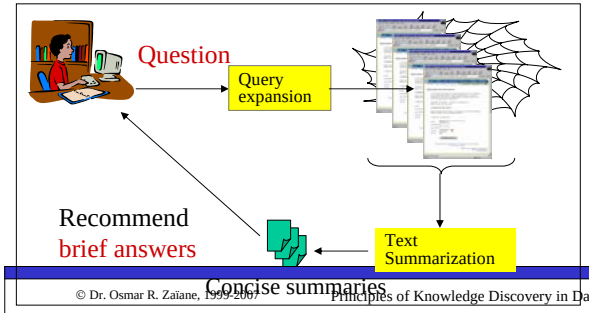
- Normal
- Malignant
- benign



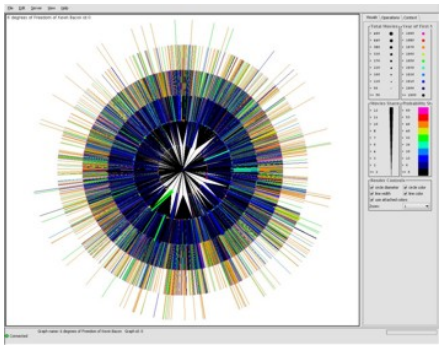
Recommender Systems

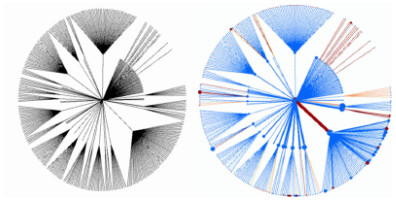


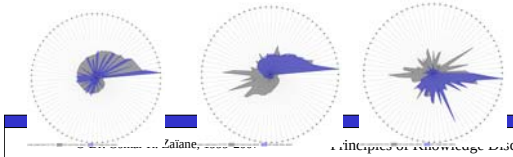


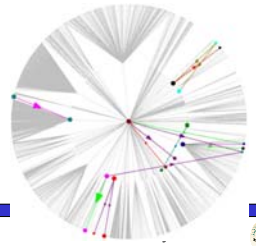


WebViz

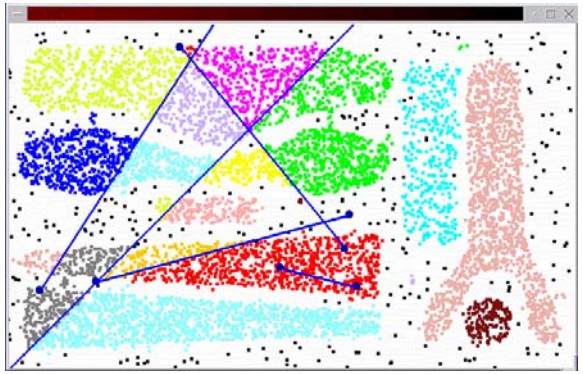








Clustering with Constraints

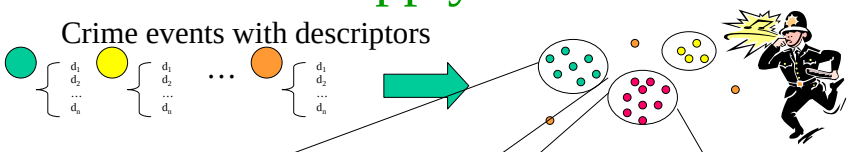


- How to model constraints? Polygons, lines?
- How to include the constraint verification in the clustering phase?

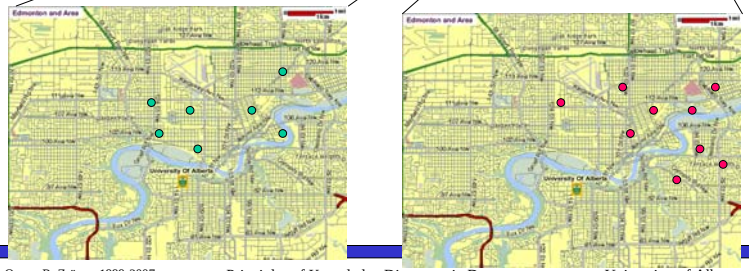
Identification of Serial crimes

How to Apply This?

Crime events with descriptors

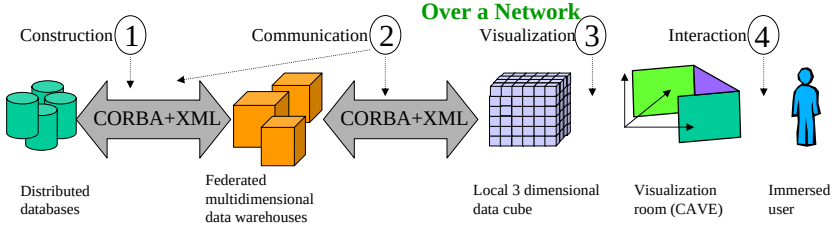


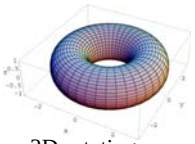
We need to define the notion of *similarity*



DIVE-ON Project

Data mining in an Immersed Virtual Environment Over a Network





3D rotating torus menu

Hand signs to operate the CAVE with OLAP operations

VWV Project

Virtual Web View