Principles of Knowledge Discovery in Data

Winter 2007

Chapter 1: Introduction to Data Mining

Dr. Osmar R. Zaïane

University of Alberta

Summary of Last Class

- Course requirements and objectives
- Evaluation and grading
- Projects and assignments
- Recommended Textbooks
- Tentative Course schedule
- Course content
- Brief Introduction to some Data Mining Tasks

Course Content

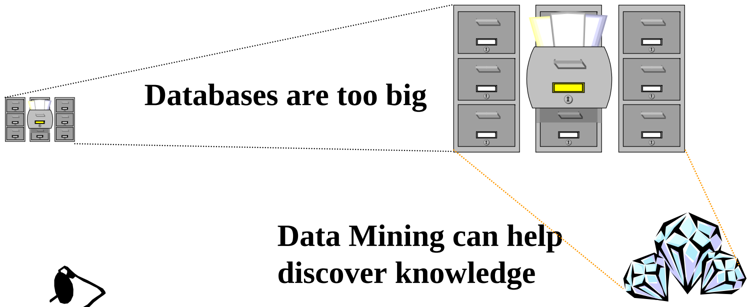
- Introduction to Data Mining
- Association Analysis
- Sequential Pattern Analysis
- Classification and prediction
- Contrast Sets
- Data Clustering
- Outlier Detection
- Web Mining
- Other topics if time permits (spatial data, biomedical data, etc.)

Chapter 1 Objectives

Get a rough initial idea about what knowledge discovery in data and data mining are.

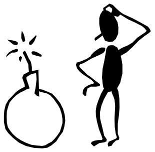
Get an overview about the functionalities and the issues in data mining.

We Are Data Rich but Information Poor



Databases are too big

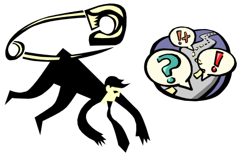
Data Mining can help discover knowledge



Terrorbytes

What Should We Do?

We are not trying to find the needle in the haystack because DBMSs know how to do that.



We are merely trying to understand the consequences of the presence of the needle, if it exists.

What Led Us To This?

Necessity is the Mother of Invention

- Technology is available to help us collect data
  - Bar code, scanners, satellites, cameras, etc.
- Technology is available to help us store data
  - Databases, data warehouses, variety of repositories…
- We are starving for knowledge (competitive edge, research, etc.)

We are swamped by data that continuously pours on us.

1. We do not know what to do with this data
2. We need to interpret this data in search for new knowledge

Evolution of Database Technology

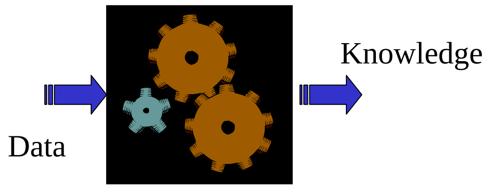
- **1950s**: First computers, use of computers for census
- **1960s**: Data collection, database creation (hierarchical and network models)
- **1970s**: Relational data model, relational DBMS implementation.
- **1980s**: Ubiquitous RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.).
- **1990s**: Data mining and data warehousing, massive media digitization, multimedia databases, and Web technology.

Notice that storage prices have consistently decreased in the last decades

What Is Our Need?

Extract <u>interesting knowledge</u>

(rules, regularities, patterns, constraints) from data in <u>large collections</u>.



A Brief History of Data Mining Research

- <u>1989 IJCAI Workshop on Knowledge Discovery in Databases</u> (Piatetsky-Shapiro)

Knowledge Discovery in Databases
(G. Piatetsky-Shapiro and W. Frawley, 1991)

- <u>1991-1994 Workshops on Knowledge Discovery in Databases</u>

Advances in Knowledge Discovery and Data Mining
(U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)

- <u>1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)</u>

Journal of Data Mining and Knowledge Discovery (1997)

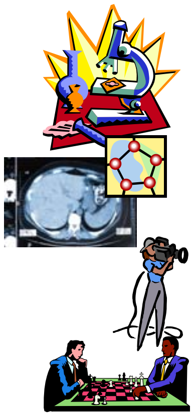
- <u>1998-2006 ACM SIGKDD annual conferences</u>
- <u>2001-2006 IEEE ICDM annual conferences</u>

2001

Introduction - Outline

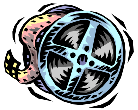
- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Data Collected



- Business transactions
- Scientific data (biology, physics, etc.)
- Medical and personal data
- Surveillance video and pictures
- Satellite sensing
- Games

Data Collected (Con’t)



- Digital media
- CAD and Software engineering
- Virtual worlds
- Text reports and memos
- The World Wide Web

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Knowledge Discovery

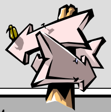
Process of non trivial extraction of implicit, previously unknown and potentially useful information from large collections of data

Many Steps in KD Process

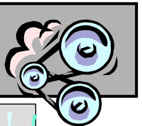
- Gathering the data together

- Cleanse the data and fit it in together

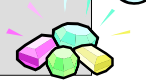
- Select the necessary data



- Crunch and squeeze the data to extract the *essence* of it



- Evaluate the output and use it



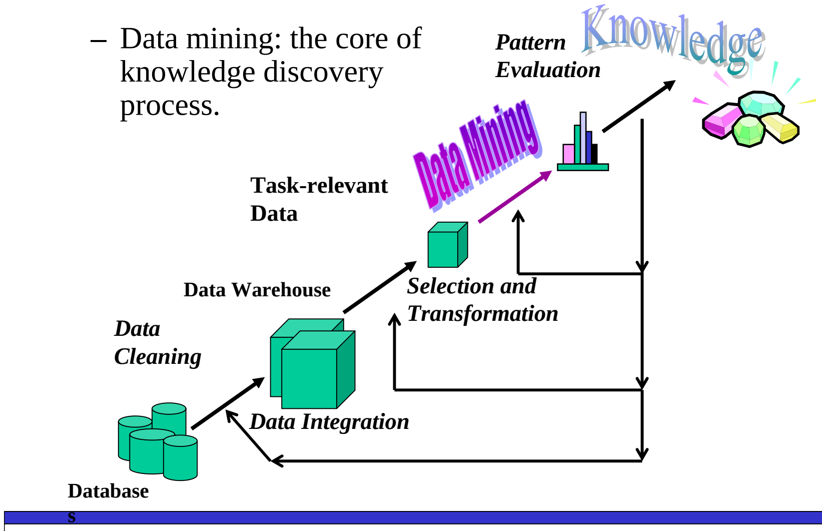
So What Is Data Mining?

- **In theory, *Data Mining* is <u>a step</u> in the knowledge discovery process. It is the extraction of implicit information from a large dataset.**
- In practice, data mining and knowledge discovery are becoming synonyms.
- There are other equivalent terms: KDD, knowledge extraction, discovery of regularities, patterns discovery, data archeology, data dredging, business intelligence, information harvesting…

- Notice the misnomer for data mining. Shouldn't it be knowledge mining?

Data Mining: A KDD Process

– Data mining: the core of knowledge discovery process.



Pattern Evaluation

Knowledge

Data Mining

Task-relevant Data

Data Warehouse

Selection and Transformation

Data Cleaning

Data Integration

Database s

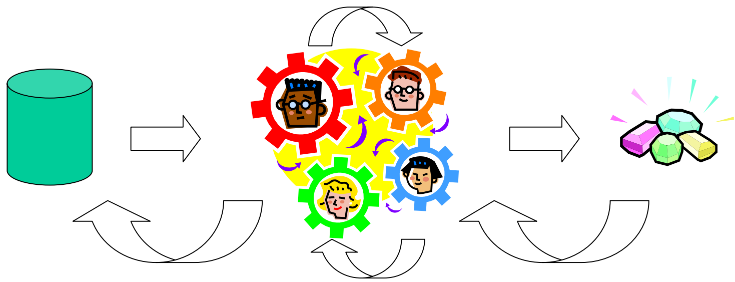
Steps of a KDD Process

- ❑ Learning the application domain
  (relevant prior knowledge and goals of application)
- ❑ Gathering and integrating of data
- ❑ Cleaning and preprocessing data (may take 60% of effort!)
- ❑ Reducing and projecting data
  (Find useful features, dimensionality/variable reduction,…)
- ❑ Choosing functions of data mining
  (summarization, classification, regression, association, clustering,…)
- ❑ Choosing the mining algorithm(s)
- ❑ Data mining: search for patterns of interest
- ❑ Evaluating results
- ❑ Interpretation: analysis of results.
  (visualization, alteration, removing redundant patterns, …)
- ❑ Use of discovered knowledge

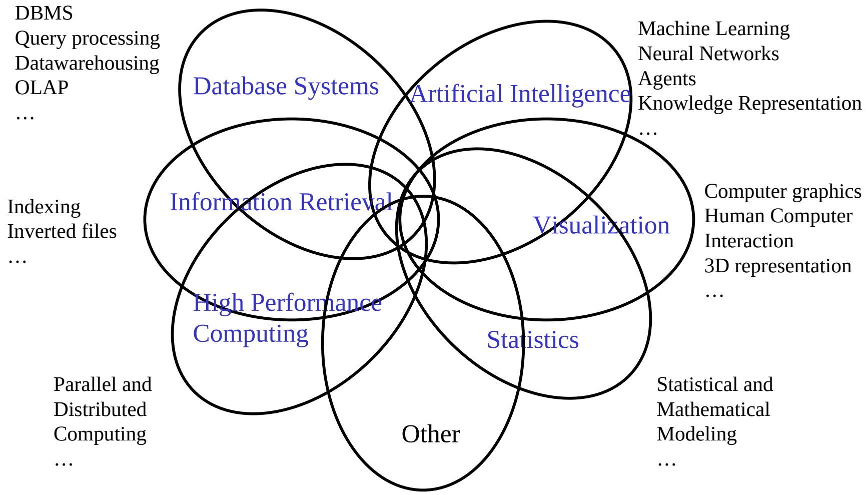
KDD Steps can be Merged

Data cleaning + data integration = data pre-processing
Data selection + data transformation = data consolidation

KDD Is an Iterative Process



KDD at the Confluence of Many Disciplines



DBMS
Query processing
Datawarehousing
OLAP
…

Database Systems

Artificial Intelligence

Machine Learning
Neural Networks
Agents
Knowledge Representation
…

Indexing
Inverted files
…

Information Retrieval

Visualization

Computer graphics
Human Computer Interaction
3D representation
…

High Performance Computing

Statistics

Parallel and
Distributed
Computing
…

Other

Statistical and
Mathematical
Modeling
…

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Data Mining: On What Kind of Data?

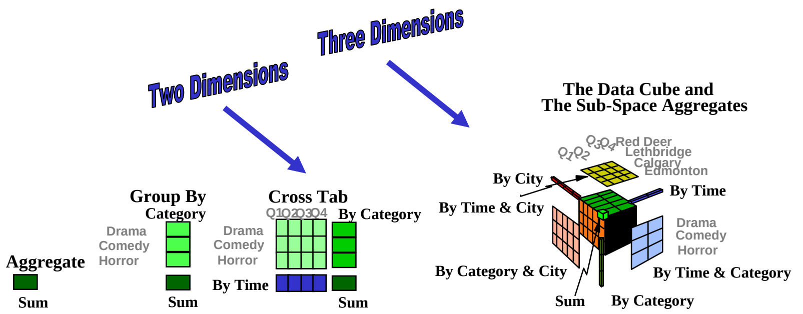
- Flat Files
- Heterogeneous and legacy databases
- Relational databases

  and other DB: Object-oriented and object-relational databases
- Transactional databases

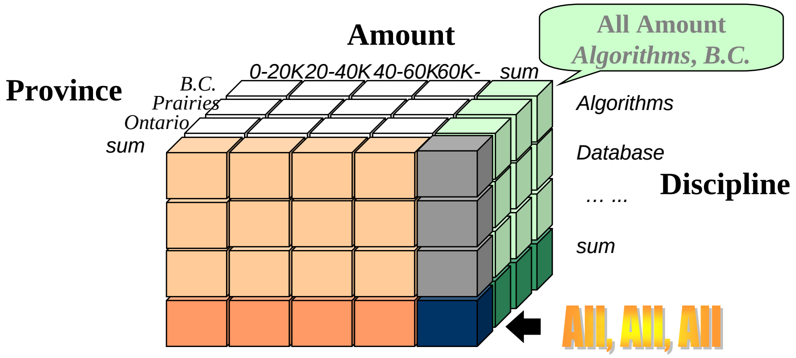
  Transaction(<u>TID</u>, Timestamp, UID, {item1, item2,…})

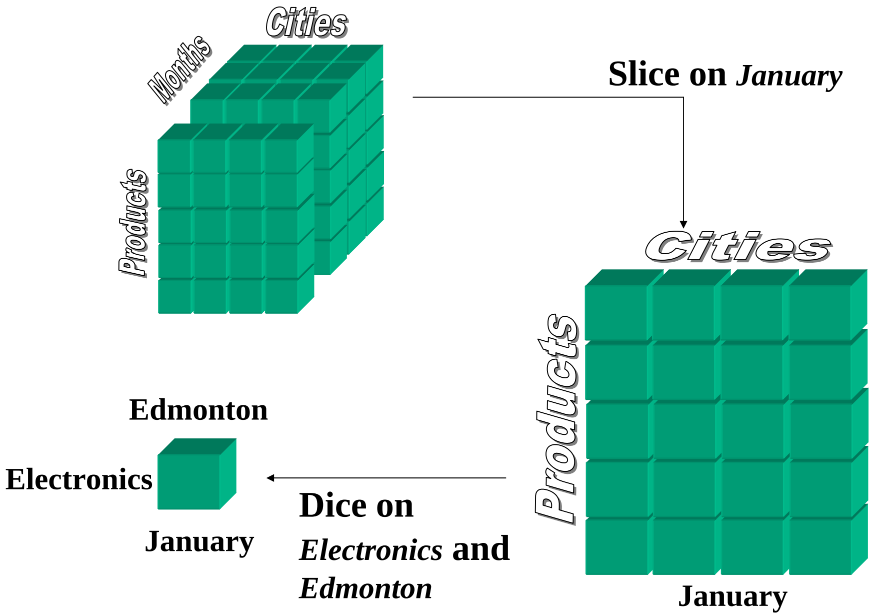
Data Mining: On What Kind of Data?

- Data warehouses



Construction of Multi-dimensional Data Cube



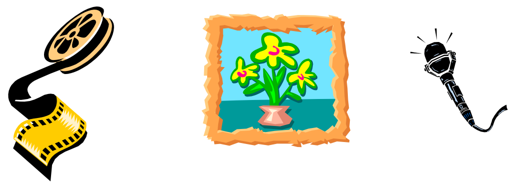


Slice on *January*

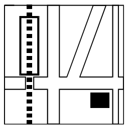
Dice on *Electronics* and *Edmonton*

Data Mining: On What Kind of Data?

- Multimedia databases

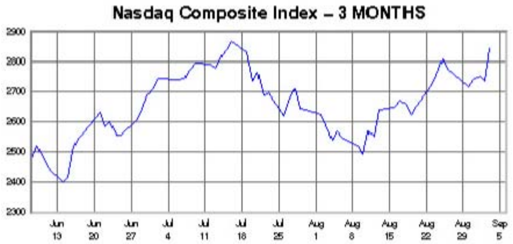


- Spatial Databases



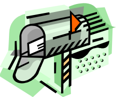
Data Mining: On What Kind of Data?

- Time Series Data and Temporal Data



Data Mining: On What Kind of Data?

- Text Documents



- The World Wide Web
  - ➢The content of the Web
  - ➢The structure of the Web
  - ➢The usage of the Web

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

What Can Be Discovered?

What can be discovered depends
upon the data mining task employed.

- Descriptive DM tasks
  Describe general properties

- Predictive DM tasks
  Infer on available data

Data Mining Functionality

- Characterization:

Summarization of general features of objects in a target class. (Concept description)

Ex: Characterize grad students in Science

- Discrimination (also Contrasting):

Comparison of general features of objects between a target class and a contrasting class. (Concept comparison)

Ex: Compare students in Science and students in Arts

Data Mining Functionality (Con’t)

- Association:

  Studies the frequency of items occurring together in transactional databases.

  Ex: buys(x, bread) → buys(x, milk).

- Prediction:

  Predicts some unknown or missing attribute values based on other information.

  Ex: Forecast the sale value for next week based on available data.

Data Mining Functionality (Con’t)

- Classification:

  Organizes data in given classes based on attribute values. (supervised classification)

  Ex: classify students based on final result.

- Clustering:

  Organizes data in classes based on attribute values. (unsupervised classification)

  Ex: group crime locations to find distribution patterns.

  Minimize inter-class similarity and maximize intra-class similarity

Data Mining Functionality (Con’t)

- Outlier analysis:

  Identifies and explains exceptions (surprises)

- Time-series analysis:

  Analyzes trends and deviations; regression, sequential pattern, similar sequences…

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Is all that is Discovered Interesting?

A data mining operation may generate thousands of patterns, not all of them are interesting.

- Suggested approach: Human-centered, query-based, focused mining

Data Mining results are sometimes so large that we may need to mine it too (Meta-Mining?)

How to measure? ➔ *Interestingness*

Interestingness

- Objective vs. subjective interestingness measures:
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, lift, correlation coefficient etc.
  - Subjective: based on user's beliefs in the data, e.g., unexpectedness, novelty, etc.

Interestingness measures: A pattern is interesting if it is

- ➢easily understood by humans
- ➢valid on new or test data with some degree of certainty.
- ➢potentially useful
- ➢novel, or validates some hypothesis that a user seeks to confirm

Can we Find All and Only the Interesting Patterns?

- Find all the interesting patterns: Completeness.
  - Can a data mining system find all the interesting patterns?
- Search for only interesting patterns: Optimization.
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First find all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns --- mining query optimization (defining and pushing constraints)

Like the concept of *precision* and *recall* in information retrieval

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Data Mining: Classification Schemes

- There are many data mining systems.

  Some are specialized and some are comprehensive

- Different views, different classifications:
  - Kinds of knowledge to be discovered,
  - Kinds of databases to be mined, and
  - Kinds of techniques adopted.

Four Schemes in Classification

- **Knowledge to be mined**:
  - Summarization (characterization), comparison, association, classification, clustering, trend, deviation and pattern analysis, etc.
  - Mining knowledge at different abstraction levels:

    primitive level, high level, multiple-level, etc.

- **Techniques adopted**:
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

Four Schemes in Classification (con't)

- **Data source to be mined**: (application data)
  - Transaction data, time-series data, spatial data, multimedia data, text data, legacy data, heterogeneous/distributed data, World Wide Web, etc.

- **Data model on which the data to be mined is drawn**:
  - Relational database, extended/object-relational database, object-oriented database, deductive database, data warehouse, flat files, etc.

Designations for Mining Complex Types of Data

- **Text Mining:**
  - Library database, e-mails, book stores, Web pages.
- **Spatial Mining:**
  - Geographic information systems, medical image database.
- **Multimedia Mining:**
  - Image and video/audio databases.
- **Web Mining:**
  - Unstructured and semi-structured data
  - Web access pattern analysis

OLAP Mining: An Integration of Data Mining and Data Warehousing

- On-line analytical mining of data warehouse data: integration of mining and OLAP technologies.
- Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- Interactive characterization, comparison, association, classification, clustering, prediction.
- Integration of different data mining functions, e.g., characterized classification, first clustering and then association, etc.

(Source JH)

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Requirements and Challenges in Data Mining

- Security and social issues
- User interface issues
- Mining methodology issues
- Performance issues
- Data source issues

Requirements/Challenges in Data Mining (Con’t)

- Security and social issues:
  - ❖ Social impact
    - Private and sensitive data is gathered and mined without individual’s knowledge and/or consent.
    - New implicit knowledge is disclosed (confidentiality, integrity)
    - Appropriate use and distribution of discovered knowledge (sharing)
  - ❖ Regulations
    - Need for privacy and DM policies

Requirements/Challenges in Data Mining (Con’t)

- User Interface Issues:
  - ❖ Data visualization.
    - Understandability and interpretation of results
    - Information representation and rendering
    - Screen real-estate
  - ❖ Interactivity
    - Manipulation of mined knowledge
    - Focus and refine mining tasks
    - Focus and refine mining results
    - Visual Data Mining (Discovering Interactively)

Requirements/Challenges in Data Mining (Con’t)

- Mining methodology issues
  - – Mining different kinds of knowledge in databases.
  - – Interactive mining of knowledge at multiple levels of abstraction.
  - – Incorporation of background knowledge
  - – Data mining query languages and ad-hoc data mining.
  - – Expression and visualization of data mining results.
  - – Handling noise and incomplete data
  - – Pattern evaluation: the interestingness problem.

(Source JH)

Requirements/Challenges in Data Mining (Con’t)

- Performance issues:
  - ❖ Efficiency and scalability of data mining algorithms.
    - Linear algorithms are needed: no medium-order polynomial complexity, and certainly no exponential algorithms.
    - Sampling
  - ❖ Parallel and distributed methods
    - Incremental mining
    - Can we divide and conquer?

Requirements/Challenges in Data Mining (Con’t)

- Data source issues:
  - ❖ Diversity of data types
    - Handling complex types of data
    - Mining information from heterogeneous databases and global information systems.
    - Is it possible to expect a DM system to perform well on all kinds of data? (distinct algorithms for distinct data sources)
  - ❖ Data glut
    - Are we collecting the right data with the right amount?
    - Distinguish between the data that is important and the data that is not.

Requirements/Challenges in Data Mining (Con't)

- Other issues
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem.

Introduction - Outline

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Potential and/or Successful Applications

- Business data analysis and decision support
  - Marketing focalization
    - Recognizing specific market segments that respond to particular characteristics
    - Return on mailing campaign (target marketing)
  - Customer Profiling
    - Segmentation of customer for marketing strategies and/or product offerings
    - Customer behaviour understanding
    - Customer retention and loyalty

Potential and/or Successful Applications (con't)

- Business data analysis and decision support (con't)
  - Market analysis and management
    - Provide summary information for decision-making
    - Market basket analysis, cross selling, market segmentation.
    - Resource planning
  - Risk analysis and management
    - "What if" analysis
    - Forecasting
    - Pricing analysis, competitive analysis.
    - Time-series analysis (Ex. stock market)

Potential and/or Successful Applications (con't)

- Fraud detection
  - Detecting telephone fraud:
    - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.

      British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
  - Detecting automotive and health insurance fraud
  - Detection of credit-card fraud
  - Detecting suspicious money transactions (money laundering)

Potential and/or Successful Applications (con't)

- Text mining:
  - Message filtering (e-mail, newsgroups, etc.)
  - Newspaper articles analysis
- Medicine
  - Association pathology - symptoms
  - DNA
  - Medical imaging

Potential and/or Successful Applications (con't)

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage.

    Spin-off ➔ VirtualGold Inc. for NBA, NHL, etc.
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining.
  - Identifying volcanoes on Jupiter.

Potential and/or Successful Applications (con't)

- Surveillance cameras
  - Use of stereo cameras and outlier analysis to detect suspicious activities or individuals.
- Web surfing and mining
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages (e-commerce)
  - Adaptive web sites / improving Web site organization, etc.
  - Pre-fetching and caching web pages
  - Jungo: discovering best sales

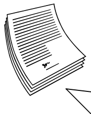
Warning: Data Mining Should Not be Used Blindly!

- Data mining approaches find regularities from history, but history is not the same as the future.
- Association does not dictate trend nor causality!?
  - Drinking diet drinks leads to obesity!
  - David Heckerman's counter-example (1997):
    - buy **hamburgers** 33% of the time, buy **hot dogs** 33% of the time, and buy both **hamburgers** and **hot dogs** 33% of the time; moreover, they buy **barbecue sauce** if and only if they buy **hamburgers**.
    - **hot dogs ➔ barbecue-sauce** has both high support and confidence.(Of course, the rule **hamburgers➔ barbecue-sauce** even higher confidence, but that is an obvious association.)
    - A manager who has a deal on **hot dogs** may choose to sell them at a large discount, hoping to increase profit by simultaneously raising the price of **barbecue**
    - **HOT-DOGS** causes **BARBECUE-SAUCE** is not part of any possible causal model, could avoid a pricing fiasco.

Conference Ranking

Could lead to a MSc co-supervised With Dr. Schaeffer

Researchers write papers



Published in

Journals

Conference Proceedings

Papers could be ranked By citation (impact)

Journals could be ranked By their paper citation (impact) Example: Thomson ISI index

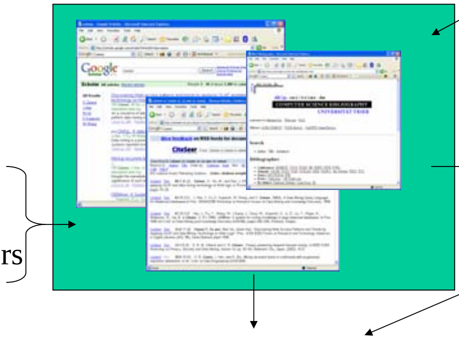
Conferences ranked only subjectively by "reputation"

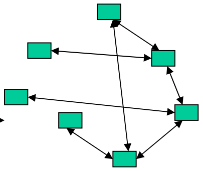
Can it be done more objectively?

- Citation of their papers
- Impact of their authors, PC members…

Conference Ranking

Google scholar
Citeseer
DBLP





Conference

- •Papers
- •Authors
- •PC members
- •Age

Social Network

Reputation, Impact, etc.

Ranking Measure

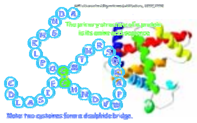
Feature Selection
Classification
Web content mining
Social Network mining

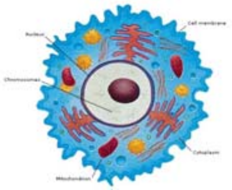
Plant Protein Localization

Identifying proteins and understanding their functions

Could lead to a MSc co-supervised With Dr. Goebel

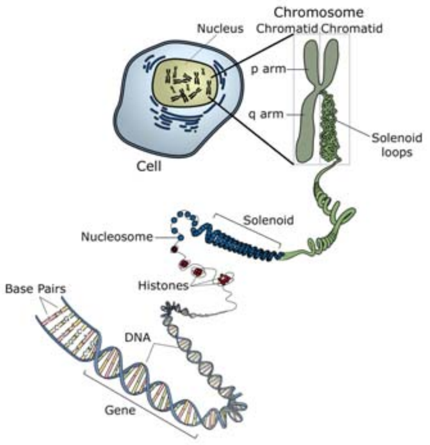
Brassica Napus (Canola)





Identifying the location of the protein helps in understanding its potential function.
Location could be cytoplasm, membrane, etc.
Interested in: Intra-cellular versus extra-cellular.

Some Biology



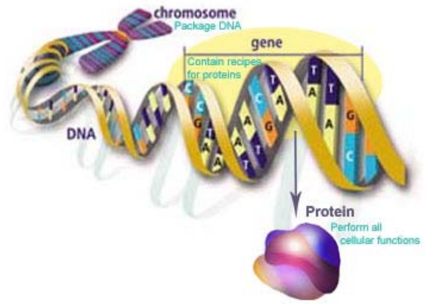
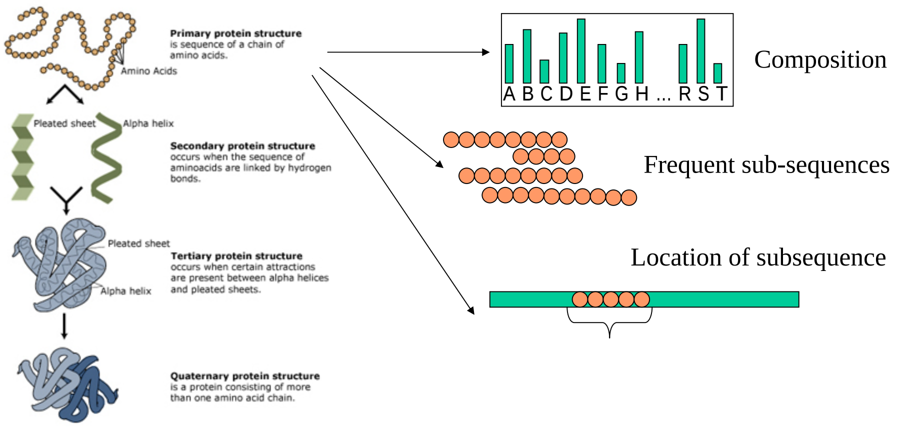


Image adapted from: National Human Genome Research Institute.

Proteins



Primary protein structure is sequence of a chain of amino acids.

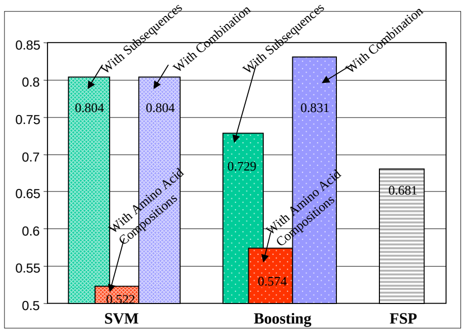
Secondary protein structure occurs when the sequence of aminoacids are linked by hydrogen bonds.

Tertiary protein structure occurs when certain attractions are present between alpha helices and pleated sheets.

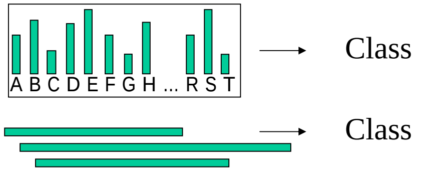
Quaternary protein structure is a protein consisting of more than one amino acid chain.

Image adapted from: National Human Genome Research Institute.

Plant Protein Localization



Project:
Using Associative classifier
- Composition
- Frequent sequence
- Combination
- Location in protein

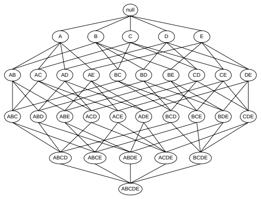


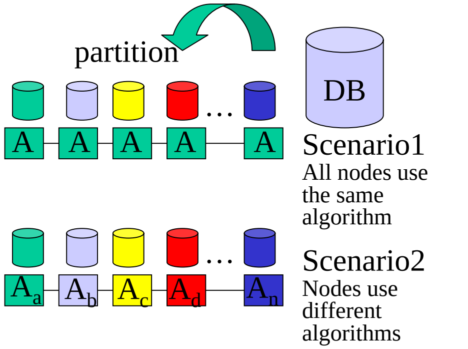
SVM and Boosting + FSP

Feature Selection
Classification
Sequence Analysis
Bioinformatics

Parallel Data Mining

Could lead to an MSc


null
A B C D E
AB AC AD AE BC BD BE CD CE DE
ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
ABCD ABCE ABDE ACDE BCDE
ABCDE



Scenario1
All nodes use the same algorithm

Scenario2
Nodes use different algorithms

Frequent itemset mining:
- Many algorithms exist
- No real winner
- Depends on dataset…

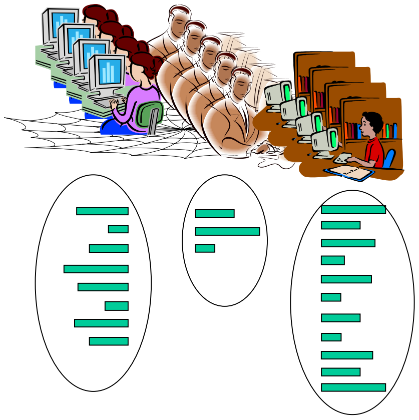
What determines the best algorithm to use given a dataset?

Association rule mining
Frequent itemset mining
Classification
Parallel computing

Could lead to an MSc

Contrasting Sequence Sets

Collaborative e-learning

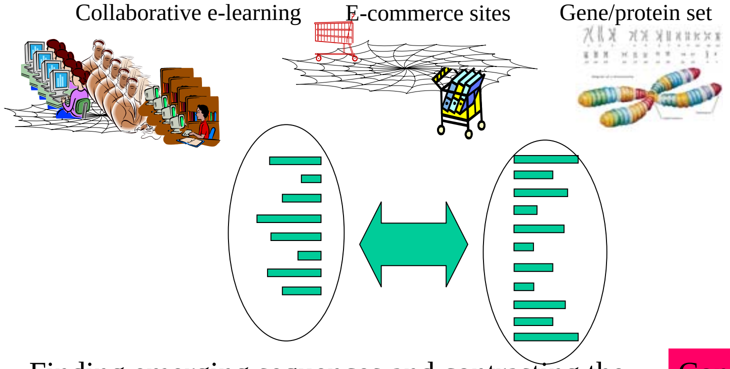


E-commerce sites

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

What contrasts genes or proteins?

Contrasting Sequence Sets



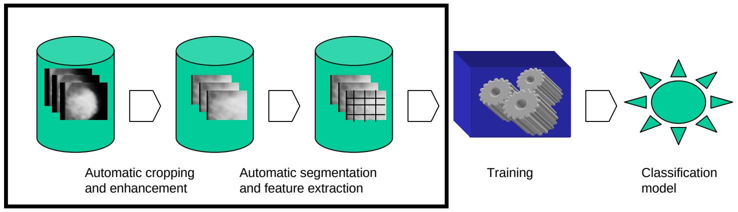
Finding emerging sequences and contrasting the sets of sequences would give insight about what behaviour leads to success and what doesn't.

Contrast sets
Sequence analysis
Outlier detection

Breast Cancer

Could lead to an MSc

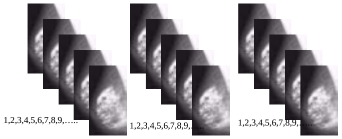
Mammography


Automatic cropping and enhancement
Automatic segmentation and feature extraction
Training
Classification model

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

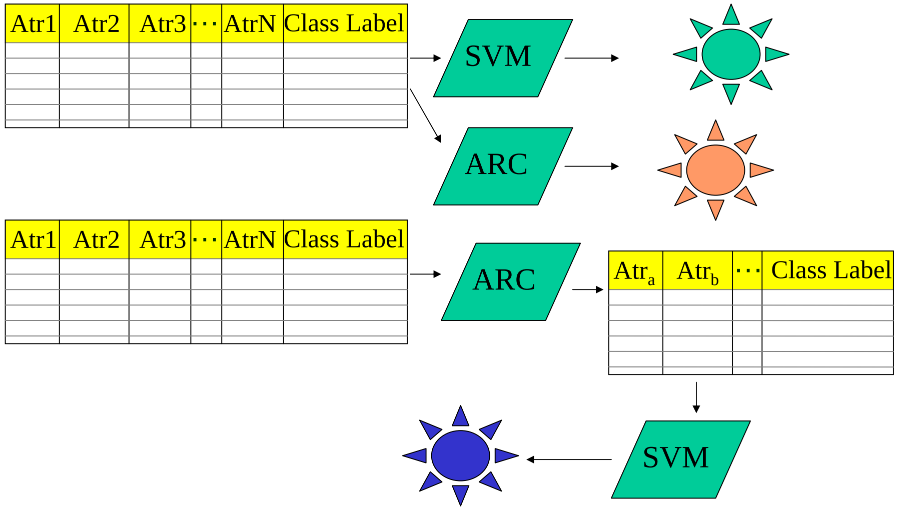
$F_\alpha \wedge F_\beta \wedge F_\gamma \wedge \ldots \wedge F_\delta \rightarrow$ **class**

- Normal
- Malignant
- benign



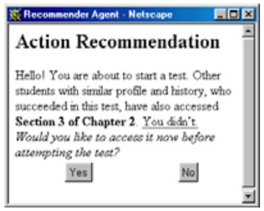
Feature Selection
Classification
Image Processing

Feature space for SVM


Atr1 Atr2 Atr3 … AtrN Class Label
SVM
ARC
Atr1 Atr2 Atr3 … AtrN Class Label
ARC
Atr$_a$ Atr$_b$ … Class Label
SVM

Recommender Systems

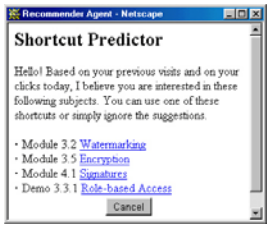
Could lead to a MSc co-supervised With Dr. Basu



Recommender Agent - Netscape

Action Recommendation

Hello! You are about to start a test. Other students with similar profile and history, who succeeded in this test, have also accessed **Section 3 of Chapter 2**. You didn't. *Would you like to access it now before attempting the test?*



Recommender Agent - Netscape

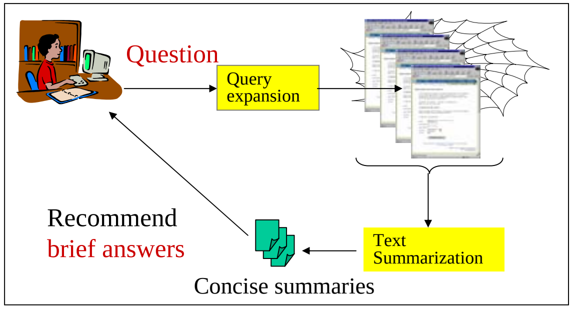
Shortcut Predictor

Hello! Based on your previous visits and on your clicks today, I believe you are interested in these following subjects. You can use one of these shortcuts or simply ignore the suggestions.

- Module 3.2 Watermarking
- Module 3.5 Encryption
- Module 4.1 Signatures
- Demo 3.3.1 Role-based Access

Designer's Guide to Color 4

Customers who bought this book also bought:



Question

Query expansion

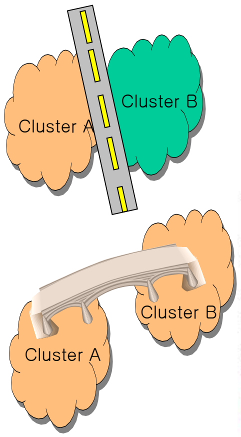
Text Summarization

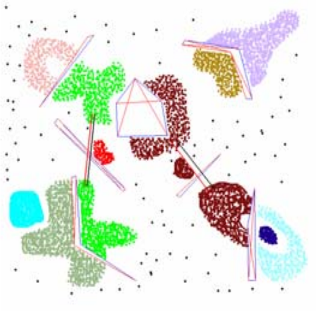
Recommend brief answers

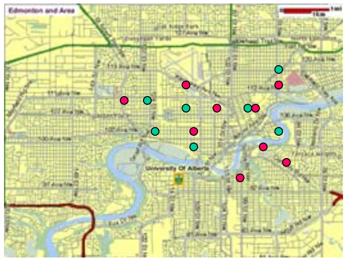
Concise summaries

Web Mining
Clustering
Machine Learning
Text Summarization

Spatial Clustering With Constraints

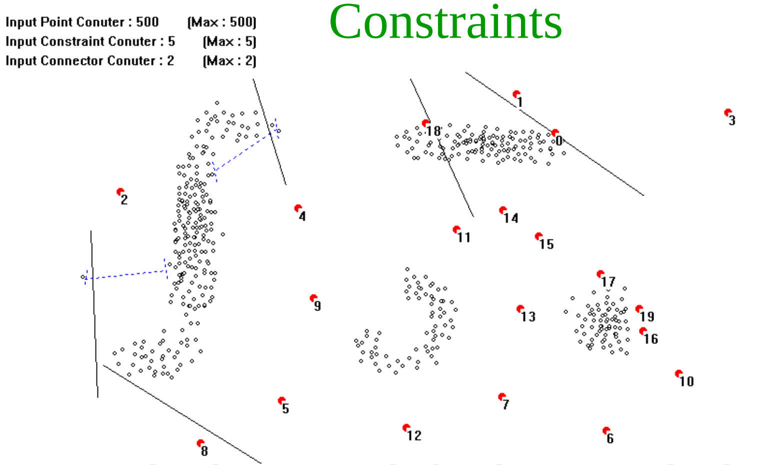






DBCluC: Based on DBScan, density-based clustering with physical constraints

Spatial Outliers (LOF) with Constraints



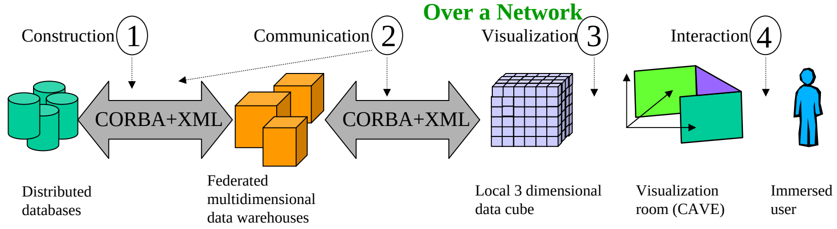
```
Input Point Conuter : 500       (Max : 500)
Input Constraint Conuter : 5      (Max : 5)
Input Connector Conuter : 2       (Max : 2)
```

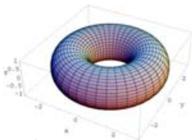
LOF: Local Outlier Factor is also based on DBScan, but does not consider constraints. Change definitions in LOF to consider constraints.

Projects on the back burner

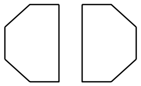
DIVE-ON Project

Data mining in an Immersed Virtual Environment Over a Network





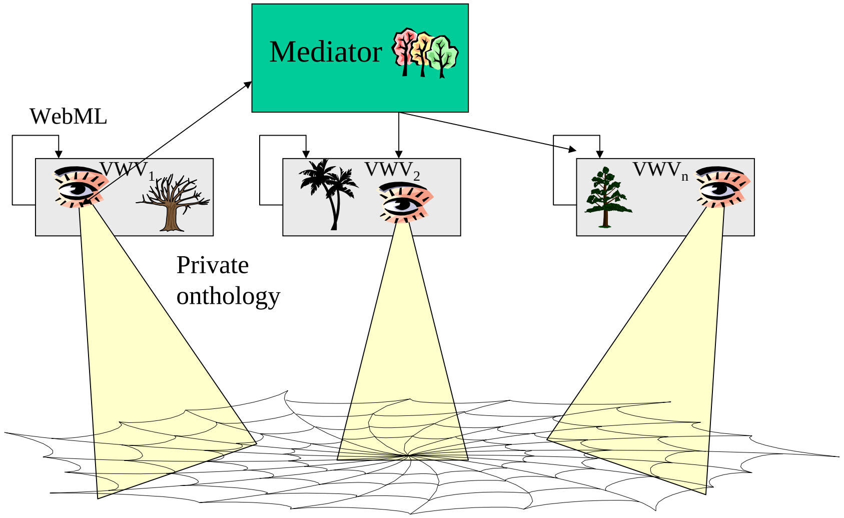
3D rotating torus menu



Hand signs to operate the CAVE with OLAP operations.

VWV Project

Virtual Web View


Mediator
WebML
VWV$_1$
VWV$_2$
VWV$_n$
Private onthology