CSCI 5408 Data Analytics: DM and DW Tech (Week 10)

- Ass4 Due: Today Mar 14
- Ass5 Due: 28
  - **Ass5-Tutorial: Mar 15**
- Write answers for review questions
  - Final Exam: Apr 20, 3:30-5:30 PM
- Reading: Lecture 15; Text: Ch1, Ch6 of 3rd edition (or Ch5 of 2nd edition)

Part II Outline

Overview: (Week 8)

1. Introduction: Overview on DM & DW *__Ass4:__ ETL/DW/OLAP*

2. Data preprocessing

DW & OLAP: (Week 9)

3. Data warehousing and OLAP

Basic DM Tasks & Algorithms:

4. Association pattern mining (Week 10-11) *__Ass5:__ Association DM*

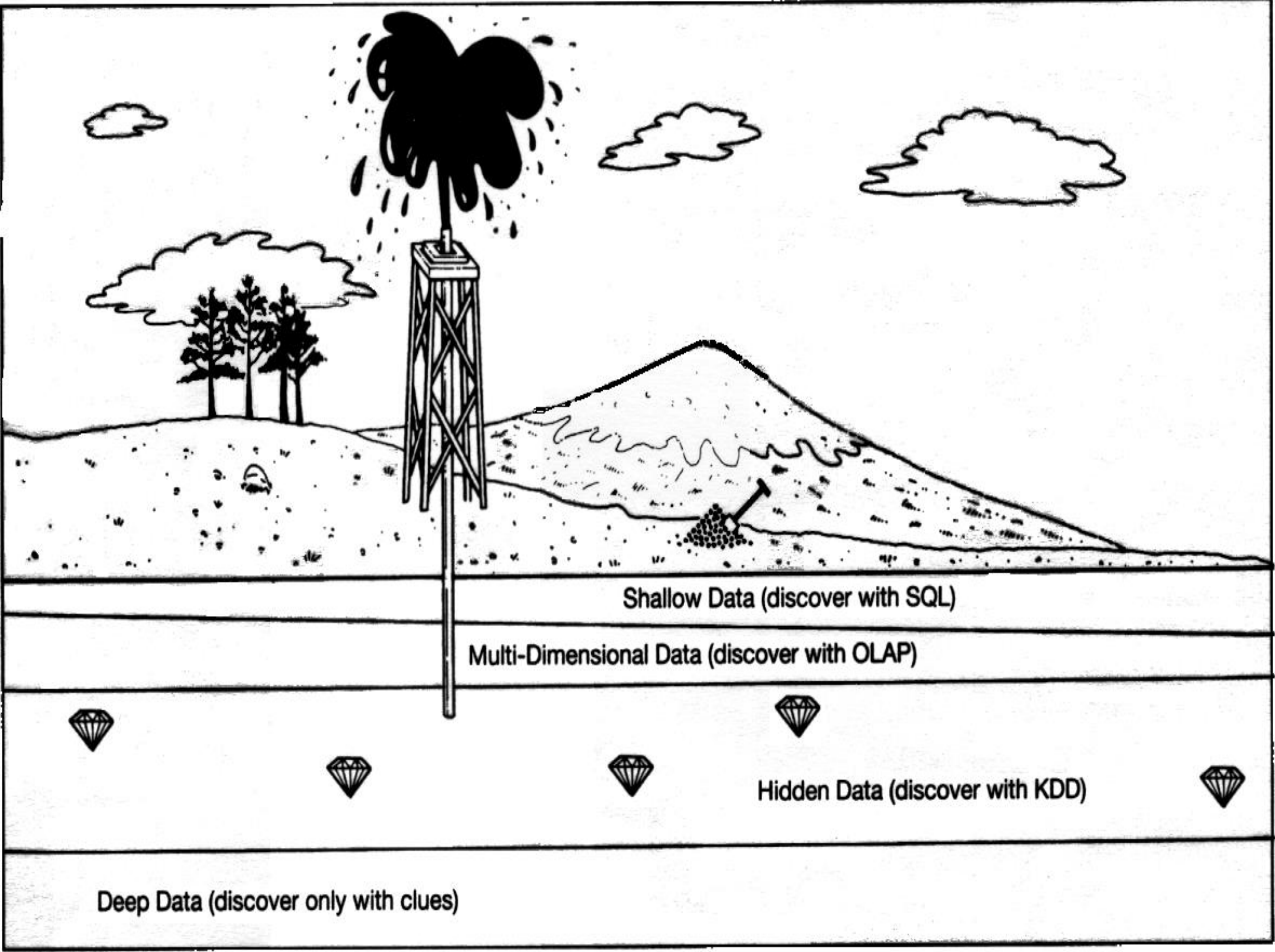
5. Classification/prediction (Week 11-12) *__Ass6:__ Classification DM*

6. Clustering analysis (Week 13)

7. Characterization/Generalization (Week 13)

Recap Types of Information Process for DSS

- **On Line Transactional (information) Process (OLTP)**
  - Track/record/retrieve original data records of every day business operations for answering *“what, when, where”* type of questions: Operational databases (Relational DB and SQL)
- **On Line Analytical (information) Process (OLAP)**
  - Store & manipulate summaries of various groupings of original data records for answering *“what happened to the business”* type of questions - Analytical databases: Data warehouses and OLAP
- **Knowledge discovery from data**
  - Discover/analyze hidden patterns of abstractive information (knowledge) for answering *“why and what to happen next”* type of questions: Data Mining (DM)

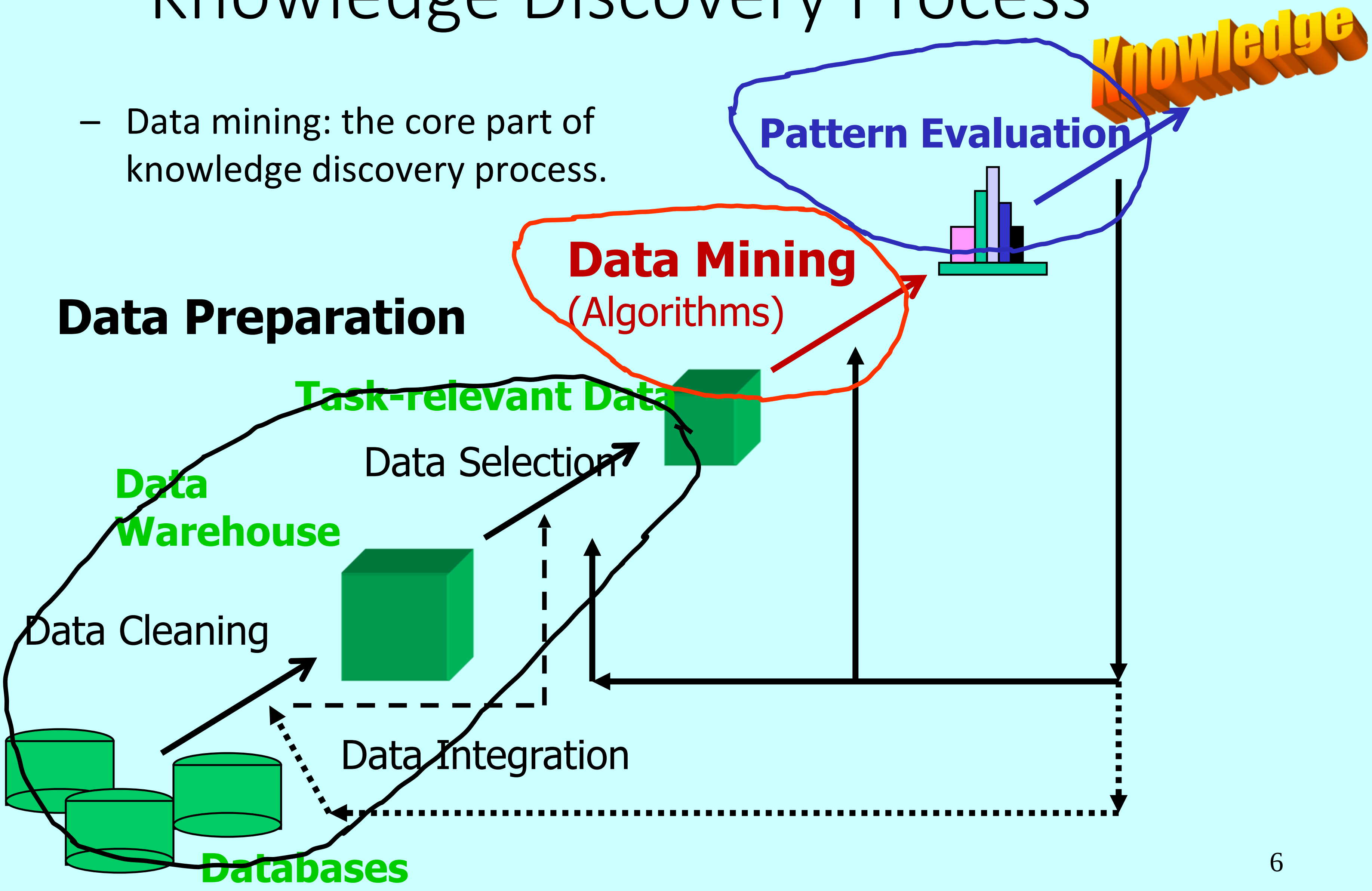

Shallow Data (discover with SQL)
Multi-Dimensional Data (discover with OLAP)
Hidden Data (discover with KDD)
Deep Data (discover only with clues)

Concept of Data Mining

- Data mining (DM) is the process of finding unknown, valid, and actionable knowledge (patterns/regularities) from collected data
- Discovered knowledge is used to assist for decision making in terms of
  - **Explanation:** understanding/explaining about current behaviors
  - **Prediction:** predicting for future outcomes

Knowledge Discovery Process

- Data mining: the core part of knowledge discovery process.

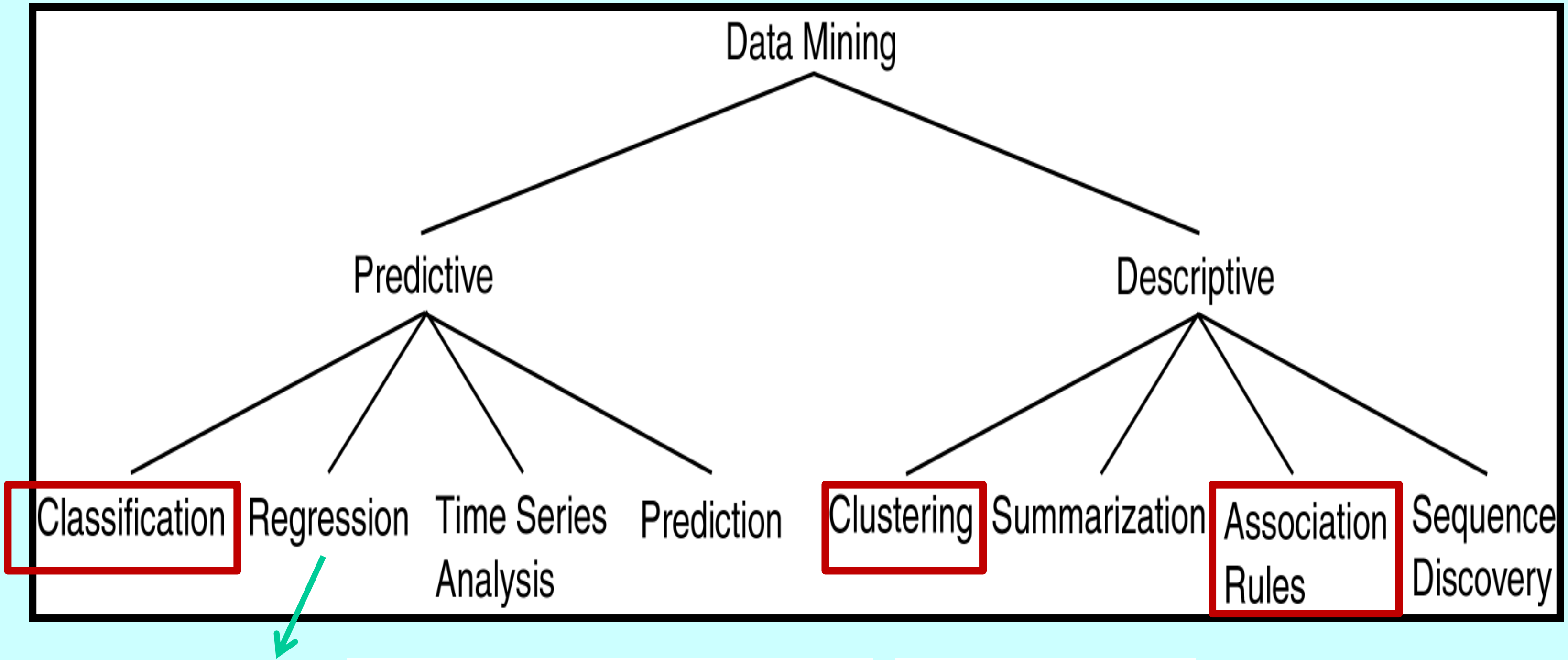


Steps of KD Process

- Understand the application domain:
  Relevant prior knowledge and goals of application.
- Create a target data set: data selection/integration from different data sources.
- Data cleaning.
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, etc.
- Choosing type of knowledge to be mined
  - Classification, regression, association, clustering, summarization, etc.
- Choosing the mining algorithm(s).
- Data mining process: induction and search for patterns of interest.
- Pattern evaluation and knowledge presentation
  - Visualization, transformation, removing redundant patterns, etc.

Apply the discovered knowledge to applications.

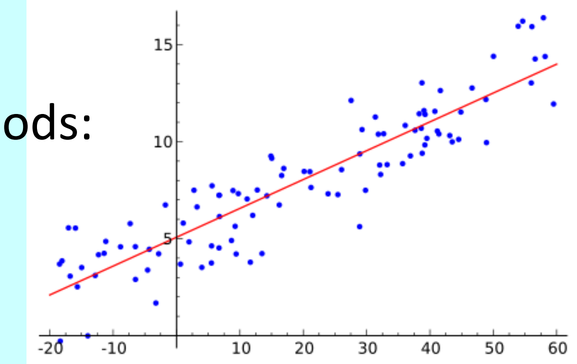
Data Mining Functions (Task Areas)

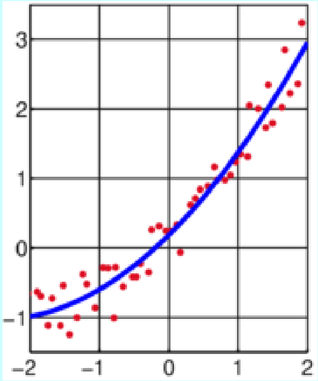


Data Mining

- Predictive
  - Classification
  - Regression
  - Time Series Analysis
  - Prediction
- Descriptive
  - Clustering
  - Summarization
  - Association Rules
  - Sequence Discovery

Statistical Modeling methods:





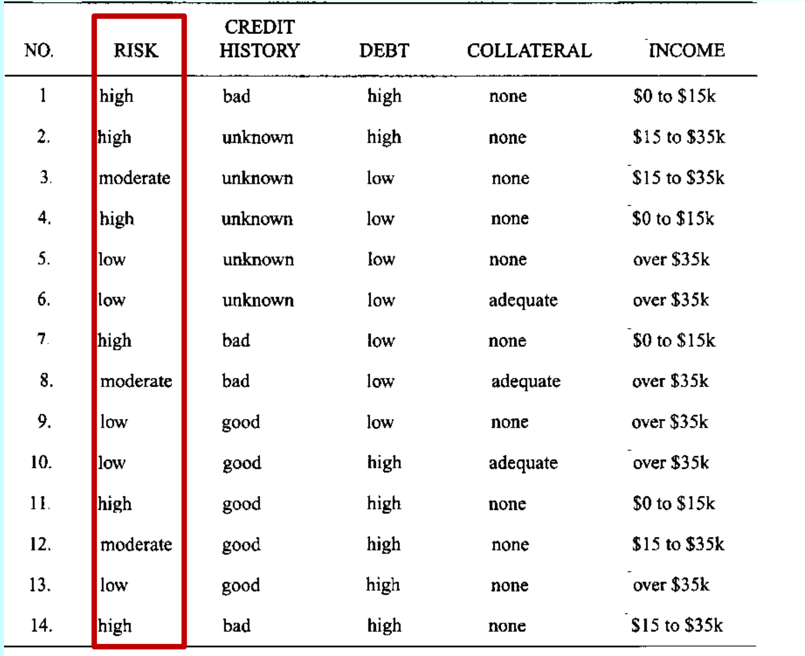
E.g., The results of fitting data points with a linear and quadratic functions.

Basic Knowledge Discovery Tasks

- Finding targeted concept model: ***Classification***
- Finding ***Association*** rules
- ***Clustering*** objects into groups
- ***Generalization*** *(or Characterization/ Summarization)*

Classification Mining

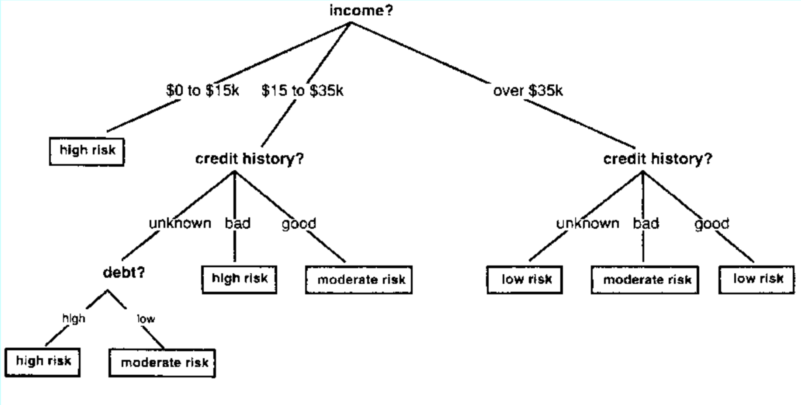
1. Training Data Set:



NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

2. Classification Knowledge Discovery (Model Construction):



A simplified decision tree for credit risk assessment.

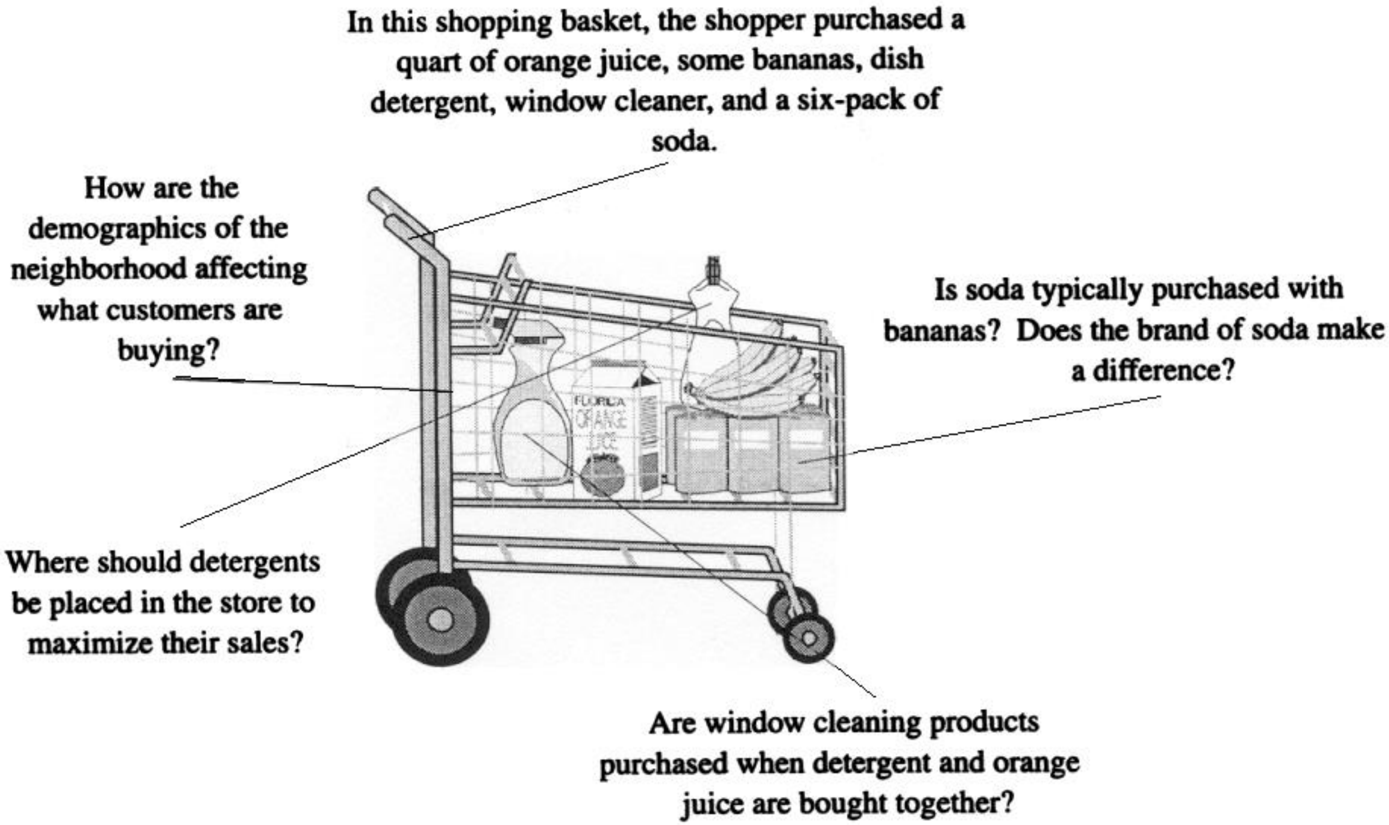
3. Class Predication: (Classify instances by classifier tool)

Input: <RISK=?, CREDIT HIS=good, DEBT=low, COLLATERAL=unknown, INCOME=$30k>

Output: RISK= **moderate risk**

Association Rule Mining

Rule: $X \Rightarrow Y$ *(sup, conf)*



In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, window cleaner, and a six-pack of soda.

How are the demographics of the neighborhood affecting what customers are buying?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

Where should detergents be placed in the store to maximize their sales?

Are window cleaning products purchased when detergent and orange juice are bought together?

- **Support rate of “X⇒Y”**
  - The percentage of transactions in the DB containing both X and Y:

sup (X ⇒Y)
= *(transactions containing both X and Y) / (all transactions)*
= *P(X∪Y)*

*The notion “X ∪ Y” here is that the items in X plus the items in Y by removing any redundant items. E.g., X={1,2,3} and Y={2,4,5}, then X ∪ Y = {1,2,3,4,5}.

- **Confidence rate of “X⇒Y”**
  - The percentage of transactions containing X also contain Y.

conf (X ⇒Y)
= (*transactions containing both X and Y) / (all transactions having X)*
= *P(Y|X)*

  - It indicates a conditional probability that a transaction having X also contains Y. E.g., DB={abc, ab, acd, cde}, *conf(a⇒b) = 66.6%, but sup(*a*⇒b) =2/4 = 50%.*

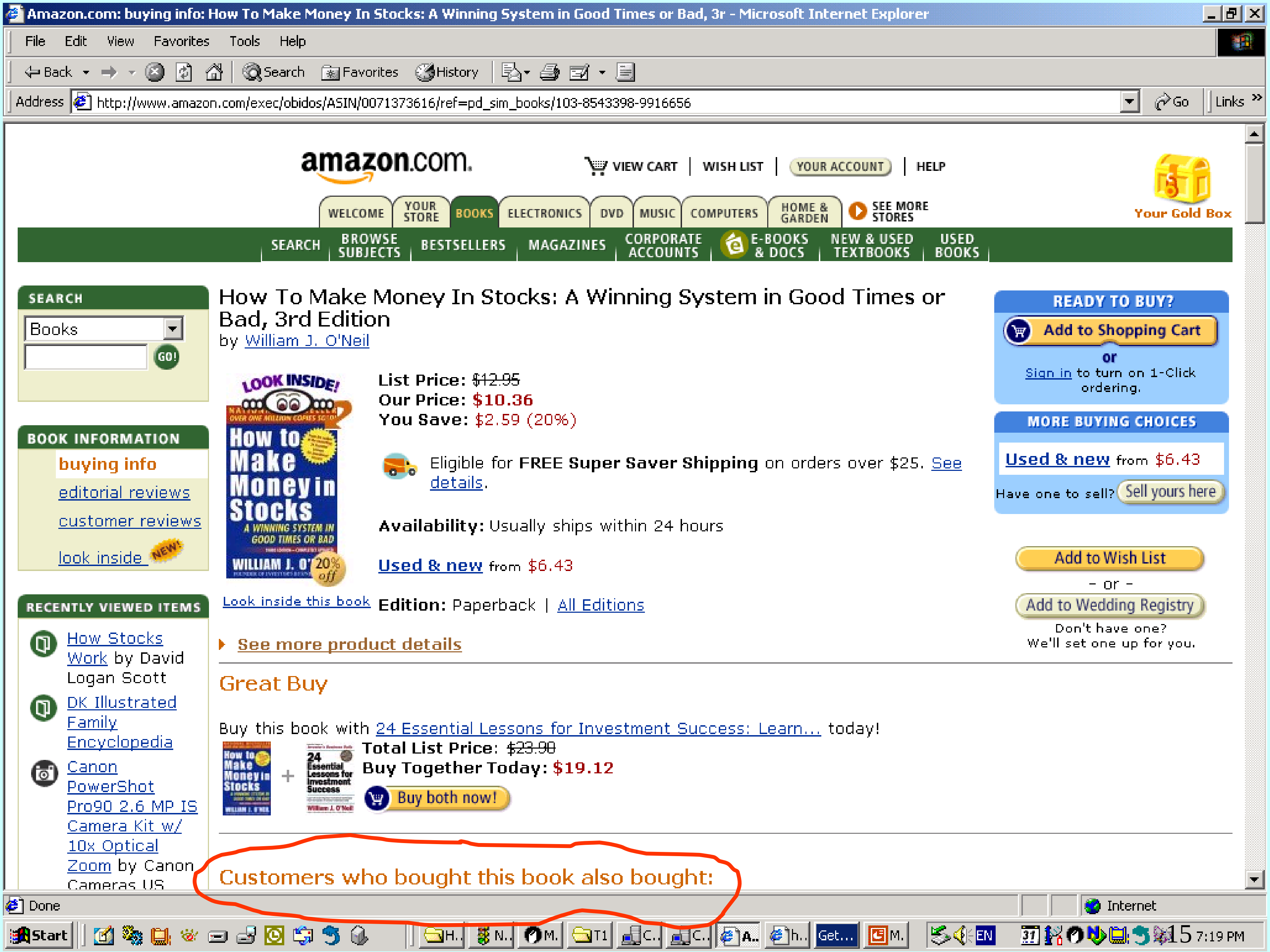
- For market basket analysis:
  - ***{book A}*** $\Rightarrow$ ***{book B, book C}*** *(sup=20%, conf=60%)*

- When adding costumer data into market basket analysis: (for finding group shopping regularities)
  - ***{age(30 ...39), income(42K...48K)}*** $\Rightarrow$ ***{buys(DVD player)}*** *(18%, 70%)*

- The sales promotion analysis (video store database):
  - ***{coupon}*** $\Rightarrow$ ***{new_release}*** *(30%, 52%)*

- Application e.g., Online Recommender System
  - The e-commerce sites collected massive amount of data on customers containing the inforamtion of purchases, browsing transactions (with hidden patterns), usage times, and preferences, etc
    - Get the such information out and stored them properly in supporting individual customer’s need by applying DM and DW technology
  - E.g., The “togetherness patterns” of purchased items require powerful association analysis on the huge transactional data, i.e.

    “Finding frequent togetherness patterns”



How To Make Money In Stocks: A Winning System in Good Times or Bad, 3rd Edition
by William J. O'Neil

List Price: ~~$12.95~~
Our Price: $10.36
You Save: $2.59 (20%)

Eligible for **FREE Super Saver Shipping** on orders over $25. See details.

Availability: Usually ships within 24 hours

Used & new from $6.43

Look inside this book

Edition: Paperback | All Editions

See more product details

Great Buy

Buy this book with 24 Essential Lessons for Investment Success: Learn... today!

Total List Price: ~~$23.90~~
Buy Together Today: $19.12

Customers who bought this book also bought:

or
Sign in to turn on 1-Click ordering.

Used & new from $6.43

Have one to sell?

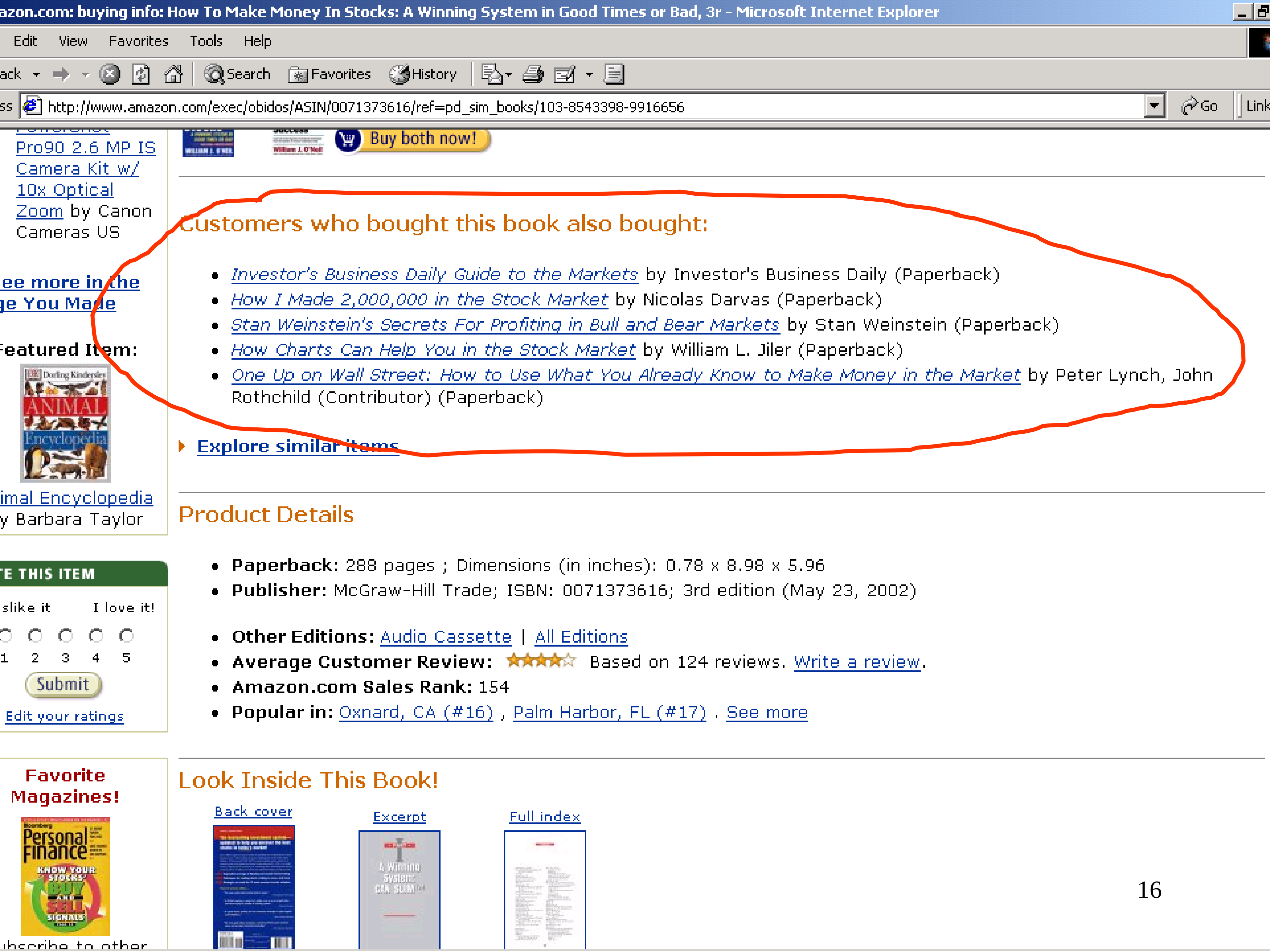
- or -

Don't have one?
We'll set one up for you.

How Stocks Work by David Logan Scott

DK Illustrated Family Encyclopedia

Canon PowerShot Pro90 2.6 MP IS Camera Kit w/ 10x Optical Zoom by Canon Cameras US


azon.com: buying info: How To Make Money In Stocks: A Winning System in Good Times or Bad, 3r - Microsoft Internet Explorer
Edit View Favorites Tools Help
ack
Search
Favorites
History
ss
http://www.amazon.com/exec/obidos/ASIN/0071373616/ref=pd_sim_books/103-8543398-9916656
Go
Link
Buy both now!
Pro90 2.6 MP IS Camera Kit w/ 10x Optical Zoom by Canon Cameras US
ee more in the
ge You Made
Featured Item:
imal Encyclopedia
y Barbara Taylor
E THIS ITEM
slike it
I love it!
○ ○ ○ ○ ○
1 2 3 4 5
Submit
Edit your ratings
Favorite Magazines!
ubscribe to other
Back cover
Excerpt
Full index

Customers who bought this book also bought:

- *Investor's Business Daily Guide to the Markets* by Investor's Business Daily (Paperback)
- *How I Made 2,000,000 in the Stock Market* by Nicolas Darvas (Paperback)
- *Stan Weinstein's Secrets For Profiting in Bull and Bear Markets* by Stan Weinstein (Paperback)
- *How Charts Can Help You in the Stock Market* by William L. Jiler (Paperback)
- *One Up on Wall Street: How to Use What You Already Know to Make Money in the Market* by Peter Lynch, John Rothchild (Contributor) (Paperback)

▸ Explore similar items

Product Details

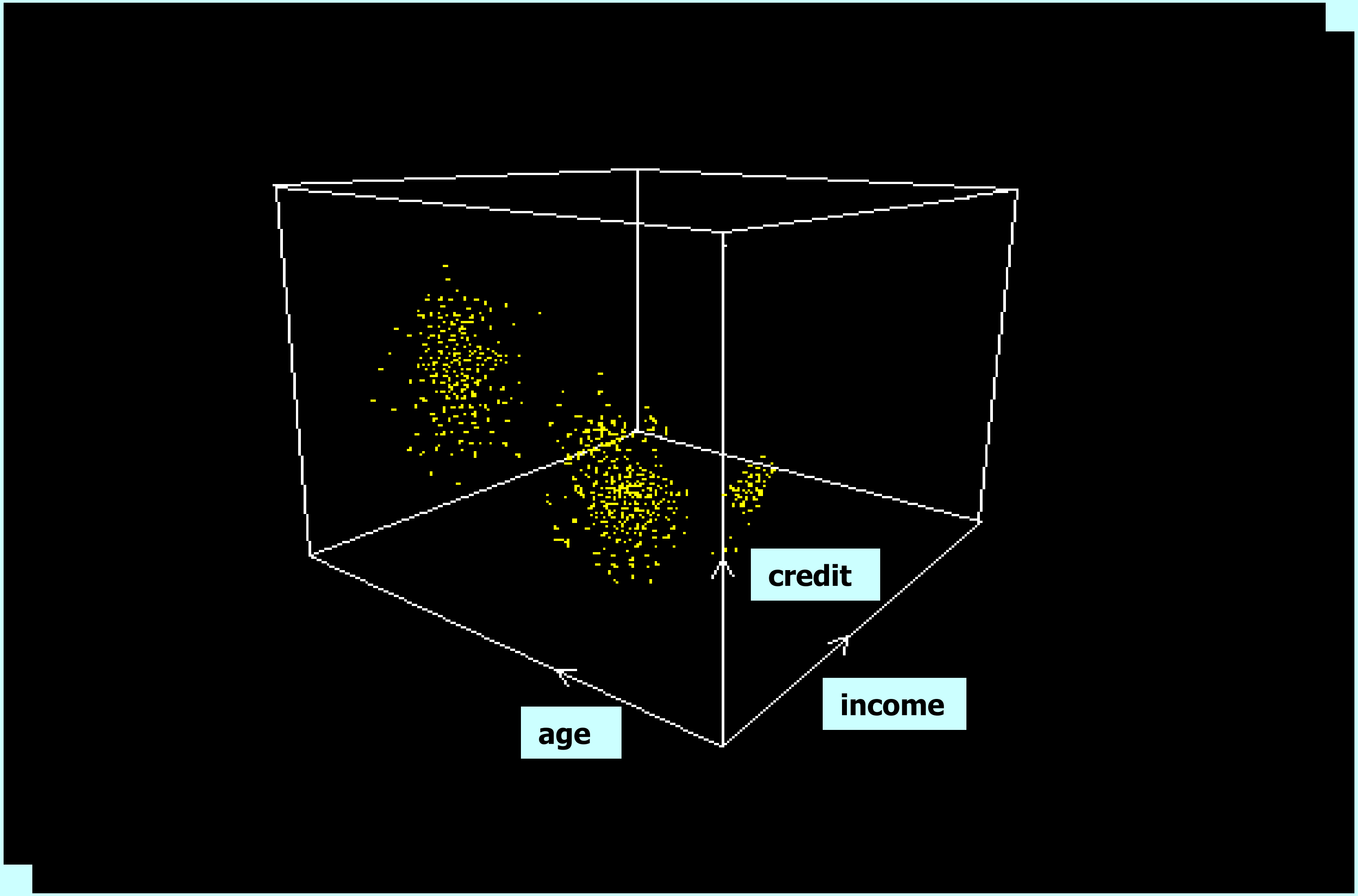
- **Paperback:** 288 pages ; Dimensions (in inches): 0.78 x 8.98 x 5.96
- **Publisher:** McGraw-Hill Trade; ISBN: 0071373616; 3rd edition (May 23, 2002)

- **Other Editions:** Audio Cassette | All Editions
- **Average Customer Review:** ★★★★☆ Based on 124 reviews. Write a review.
- **Amazon.com Sales Rank:** 154
- **Popular in:** Oxnard, CA (#16) , Palm Harbor, FL (#17) . See more

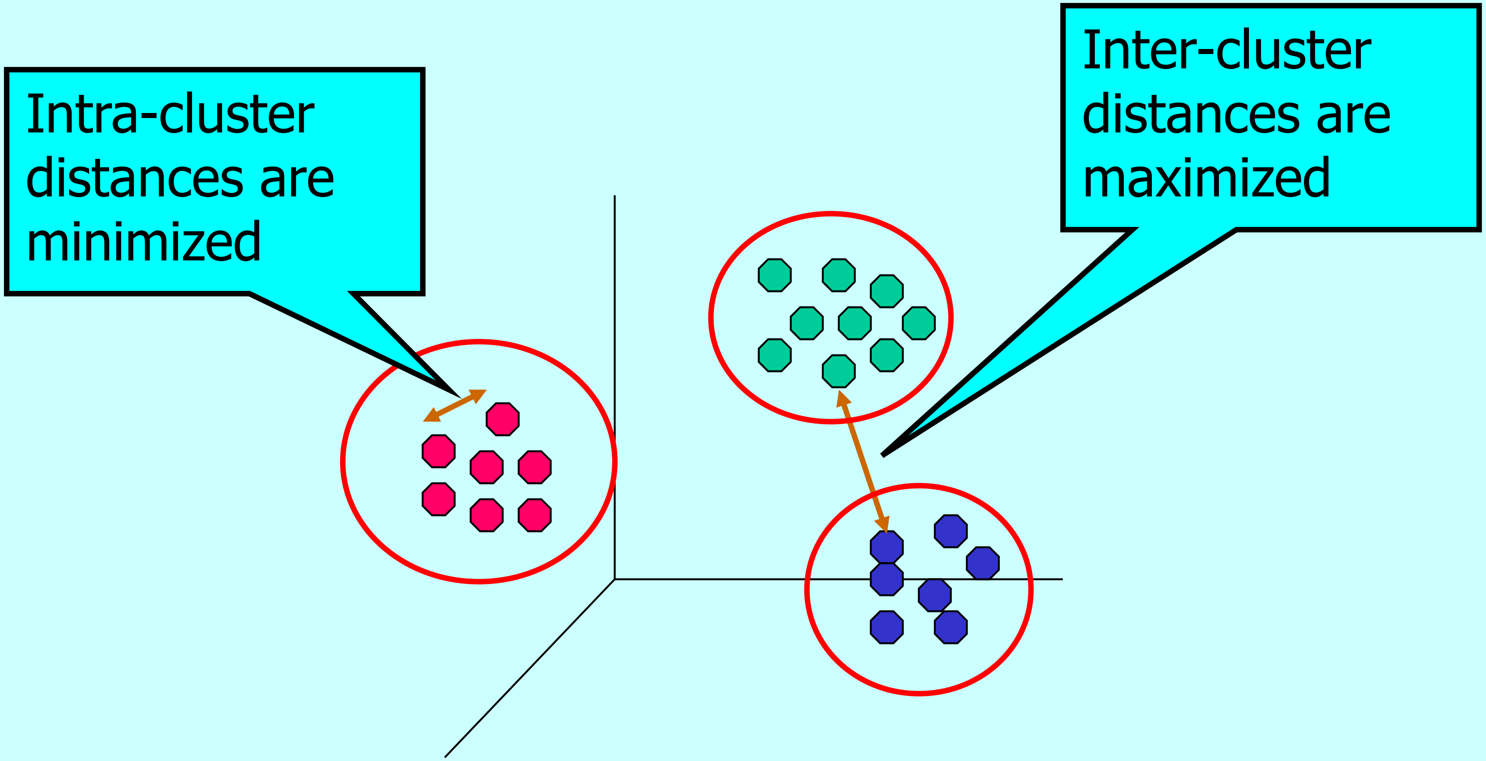
Look Inside This Book!

Clusters Mining

E.g. Finding clusters of customers in a 3D DB:



- Finding groups of objects such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering e.g., Text categorization for information retrieval

- **Application**:

Automatic text categorization and text retrieval for PPML archive (Doc/Theses/MCSthesisGuo00.pdf, MACSprojectQi03.pdf)

- **Data set:**
  - PPML: Pediatric Pain Mailing List (600 worldwide subscribers)
    Pediatric: The branch of medicine that deals with the care of infants and children and the treatment of their diseases.
  - The archive DB: 7129 files (14MB), in 1999.
  - The DB is growing in the rate around 3M per year.
  - A lot of unpublished domain information presented in the data.

- **Problem and Objective:**
  - It takes time to **find relevant Emails** from a large collection of Emails. Some **users may not know what proper key words** to use for forming a proper query.
  - Develop an information retrieval system which provides certain guidance for searching PPML archive.

From raw data to clean and normalized data:

```
Return-path: <Drhbg@aol.com>
Received: from DIRECTORY-DAEMON by SYSWRK.UCIS.DAL.CA
 (PMDF V4.3-13 #6307) id
<01J615F5VHLS00BCUD@SYSWRK.UCIS.DAL.CA>; Fri, 01 Ja
n 1999 15:42:13 -0400
Received: from imo23.mx.aol.com by SYSWRK.UCIS.DAL.CA (PMDF
V4.3-13 #6307)
 id <01J615F12Y6O00CSAS@SYSWRK.UCIS.DAL.CA>; Fri, 01 Jan
 1999 15:42:07 -0400
Received: from Drhbg@aol.com by imo23.mx.aol.com (IMOv18.1)
 id NVXFa07005 for <pediatric-pain@ac.dal.ca>; Fri,
 1 Jan 1999 14:41:54 -0500 (EST)
Date: Fri, 01 Jan 1999 14:41:54 -0500 (EST)
From: Drhbg@aol.com
Subject: Re: Management of nerve injury
To: pediatric-pain@ac.dal.ca
Message-id: <7deeafb8.368d2502@aol.com>
MIME-version: 1.0
x-Mailer: AOL 2.5 for Windows
Content-type: text/plain; charset=US-ASCII
Content-transfer-encoding: 7bit
```

I agree with William Fenton. I think mexiletine should be used as a second line drug. I ordinarily treat patients with chronic neuropathic pain. However, on a number of occasions, I have treated patients with acute neuropathic pains such as sciatica or brachial plexopathies. I have prescribed gabapentin at the outset of the pain, and have found that patients have responded extremely well. They often require lower than anticipated dosages of opioid analgesics. I doubt there is any data on the benefits of early use of anticonvulsants, but a case-control study would be of value.

Return-path: …

……..

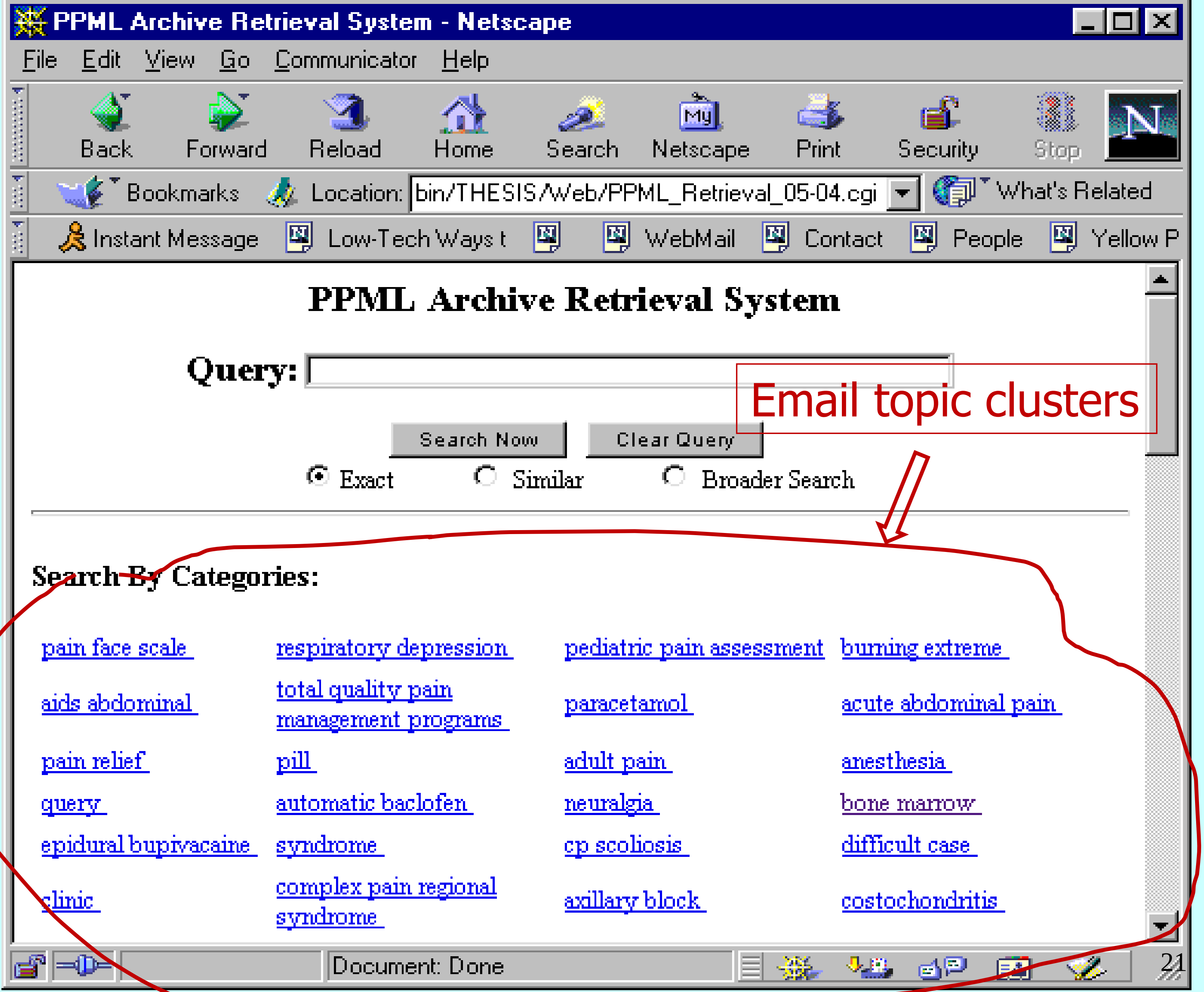
Thread 1: Opioids and Meningitis

Date: Wed, 04 Jan 1995 16:54:48 -0500 (EST) From: posterSubject: opioids and meningitisX is a 13 month (9.8kg) old boy suffering from acute meningitis (pneumocoque) treated with IV cefotaxime; at day three, I have been called as pediatric pain consultant to assess X; I have discovered an extreme painfull state: one could not handle or touch him without producing screaming. The child was unable to move spontaneously he looked paralysed by pain and hypertonia ; he also presented a neurological complication : ptosis at the right side.The pain treatment was IV acetaminophen. The first day I have prescribed IV Nalbuphine (weak opioid u antagonist and agonist) 11mg/24h after a loading dose of 1.4 mg; Pain at rest has been succesfully relieved but not the mobilisation pain; the dose has been increased at 14 mg/day wihout relieving the pain associated with moving; he has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4 days later. Neurological examination and CT scan have been still normal (except ptosis) during this period. No opioid's side effects have been observed.What do you think of this case ?Have you any experience with opioids and acute meningitis ?Dr Poster, Pediatric pain unit, Poster Hospital

Date: Wed, 04 Jan 1995 17:27:25 -0500 (EST) From: first replySubject: re: opioids and meningitisIs there any periosteal involvement? If so an NSAID (ibuprofen or naproxen) may be much more effective than even opioid.

Date: Wed, 04 Jan 1995 19:06:32 -0400 From: second replySubject: Re: opioids and meningitisPoster writes:>X is a 13 month (9.8kg) old boy suffering from acute meningitis...>extreme painfull state: one could not handle or touch him without>producing screaming....>The first day I have prescribed IV Nalbuphine ...>succesfully relieved but not the mobilisation pain;...>has moved spontaneously limbs 2 days later; nalbuphine has been stopped 4>days later. Neurological examination and CT scan have been still normal...I have used IV morphine for similar severe meningitis pain, with success. I wouldn't hesitate to use a pure opioid agonist (in conjunction with acetaminophen, NSAID, and/or tricyclics). However, it sounds like you have the situation under control.Second Reply, Associate Professor, Dept and University

Date: Thu, 05 Jan 1995 18:58:32 -0800 (PST) From: Third ReplySubject: Re: opioids and meningitisI wonder if the problem is not due to severe arachnoiditis that is secondary to the inflammation. I would suggest a trial of steroids in this patient, perhaps in combination with a benzodiazepine to reduce the spasm. Narcotics may reduce the pain but I would not like to keep X on them for too long. Good luck Third Reply



PPML Archive Retrieval System

Query:

Search Now
Clear Query

● Exact ○ Similar ○ Broader Search

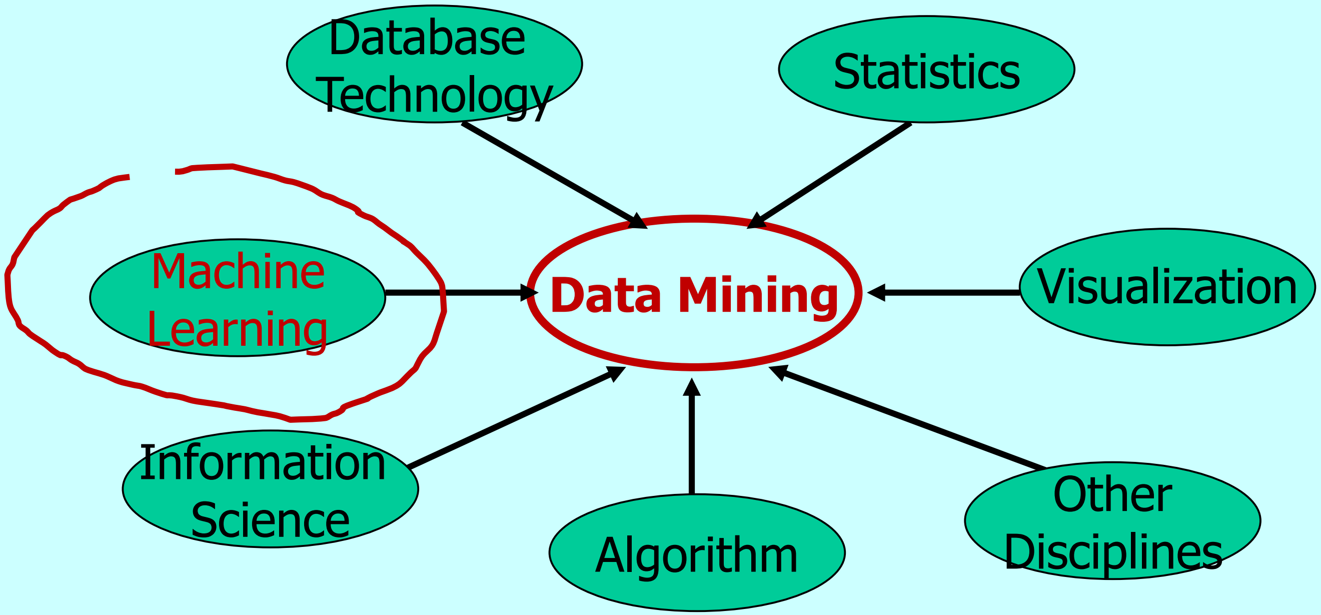
Search By Categories:

pain face scale
respiratory depression
pediatric pain assessment
burning extreme
aids abdominal
total quality pain management programs
paracetamol
acute abdominal pain
pain relief
pill
adult pain
anesthesia
query
automatic baclofen
neuralgia
bone marrow
epidural bupivacaine
syndrome
cp scoliosis
difficult case
clinic
complex pain regional syndrome
axillary block
costochondritis

Email topic clusters

DM vs. Machine Learning (ML)

- Data Mining: Confluence of multiple disciplines



- ML focus on learning mechanism/paradigm/methodology.
- DM focuses on application and extension of ML as core of a solution system for discovering knowledge from large data sets.

What Is Learning?

An operational definition:

A certain 'task' to be carried out either well or badly, and a 'subject' that is to carry out the task; how to determine when someone has learned something.

An individual learns how to carry out a certain task by making a **transition** from a situation in which the task cannot be carried out properly to a situation in which the same task can be carried out properly under the same circumstances.

The process of this transition is called Learning, or Training.

Self-learning Computer System

The goal of **self-learning computer** is to generate programs itself when environment has been changed, so that it is enable to carry out new tasks.

Computer: speed and accuracy, but lack of flexibility/creativity.
Human: rich in flexibility/creativity by ability to learn, imagine.

Machine Learning (ML) methods are developed to obtain knowledge automatically from data for carrying on new unknown tasks.

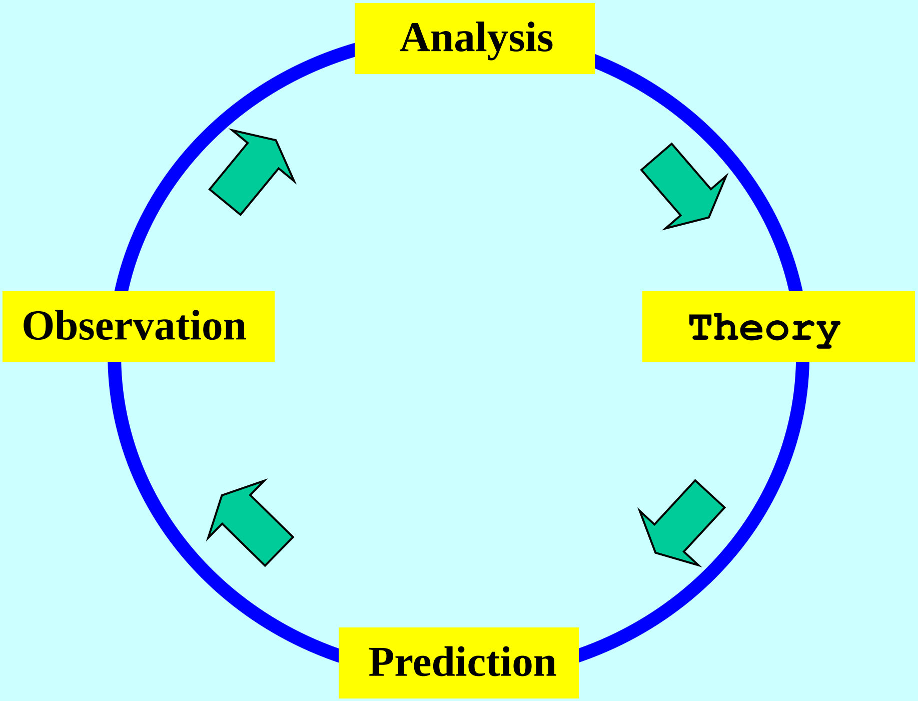
ML has a strong relationship with the methodology of science, as they both share the process of knowledge discovery.

General Methodology of Science

- The fundamental tasks of scientists are to **explain** and to **predict,** based on observations and the existing knowledge, and to discover new knowledge (e.g., the laws of gravity and forces, etc)
- A general methodology of science:

The process of scientific research ideally takes the form so called **Empirical Cycle.**

Empirical Cycle Model:



1. **Observation:** we start with a number of observations.

2. **Analysis:** we try to find patterns in these observations.

3. **Theory:** if we have found some regularities, we formulate a theory (hypothesis) explaining the data.

4. **Prediction:** our theory will predict new phenomena that can be verified by new observations.

Empirical cycle: An on-going process

In stage 4 of the cycle there are two possibilities:

a) **Predictions are correct,** in which case our theory is corrected, or
b) **Predictions are wrong.**

If b), we have to analyze the new observations and try to come up with a new theory. So the whole process starts again.

- This why we speak of an empirical cycle:

The process goes on and on forever, and we can refine our theories indefinitely.

The same holds, apart from changes of detail, for a manager who tries to analyze a market to develop new products or optimize production.

- We can formulate hypotheses to explain empirical observations but that we can never prove that they are true.

E.g., The evolution of the basic physics theories: the laws of gravity and forces

- **The gravity theory**: *Isaac Newton* (1642-1727)

***Newton's Law of gravitation** is very accurate only <u>when gravity is weak</u> – and must be replaced by **Einstein's general theory of relativity** in <u>strong gravitational field</u>.

- **The general theory of relativity**: *Albert Einstein* (1879-1955)

*Similarly, **relativity theory** must be replaced by **quantum mechanics** when examining <u>interactions on microscope scale</u>, such as the big bang singularity, or at the edge and center of a black hole.

- **The quantum gravity theory**: *Stephen Hawking* (1941-), etc.

A Case Study

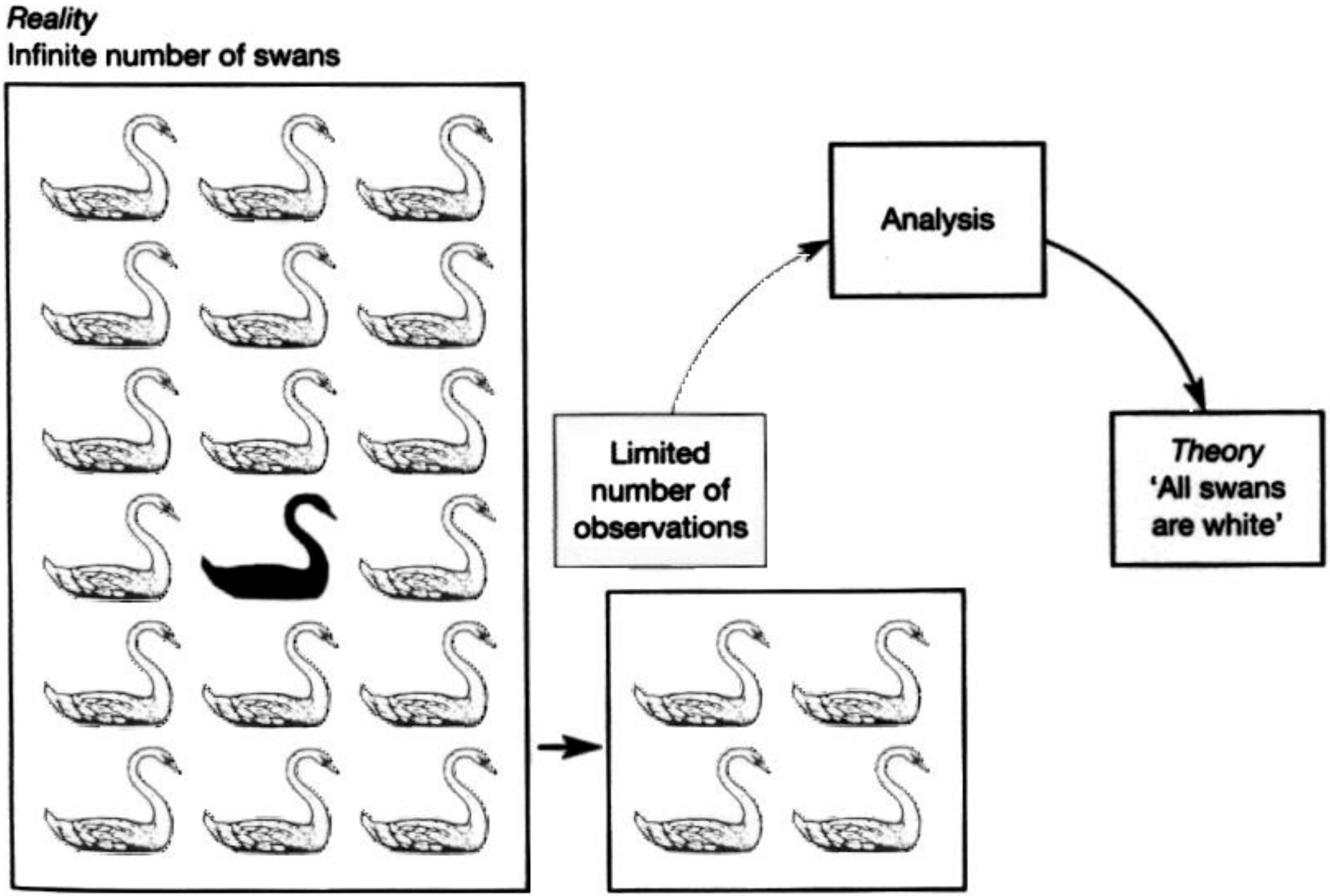
Application: formulate a hypothesis concerning the color of swans.

Observation: We observe a number of swans that all white.

Analysis: Swan – White (regularity)

Theory: “All swans are white’’ (knowledge)

Prediction: Swan -> White (performing prediction)


Reality
Infinite number of swans
Analysis
Limited number of observations
Theory
'All swans are white'

Forming hypothesis by induction

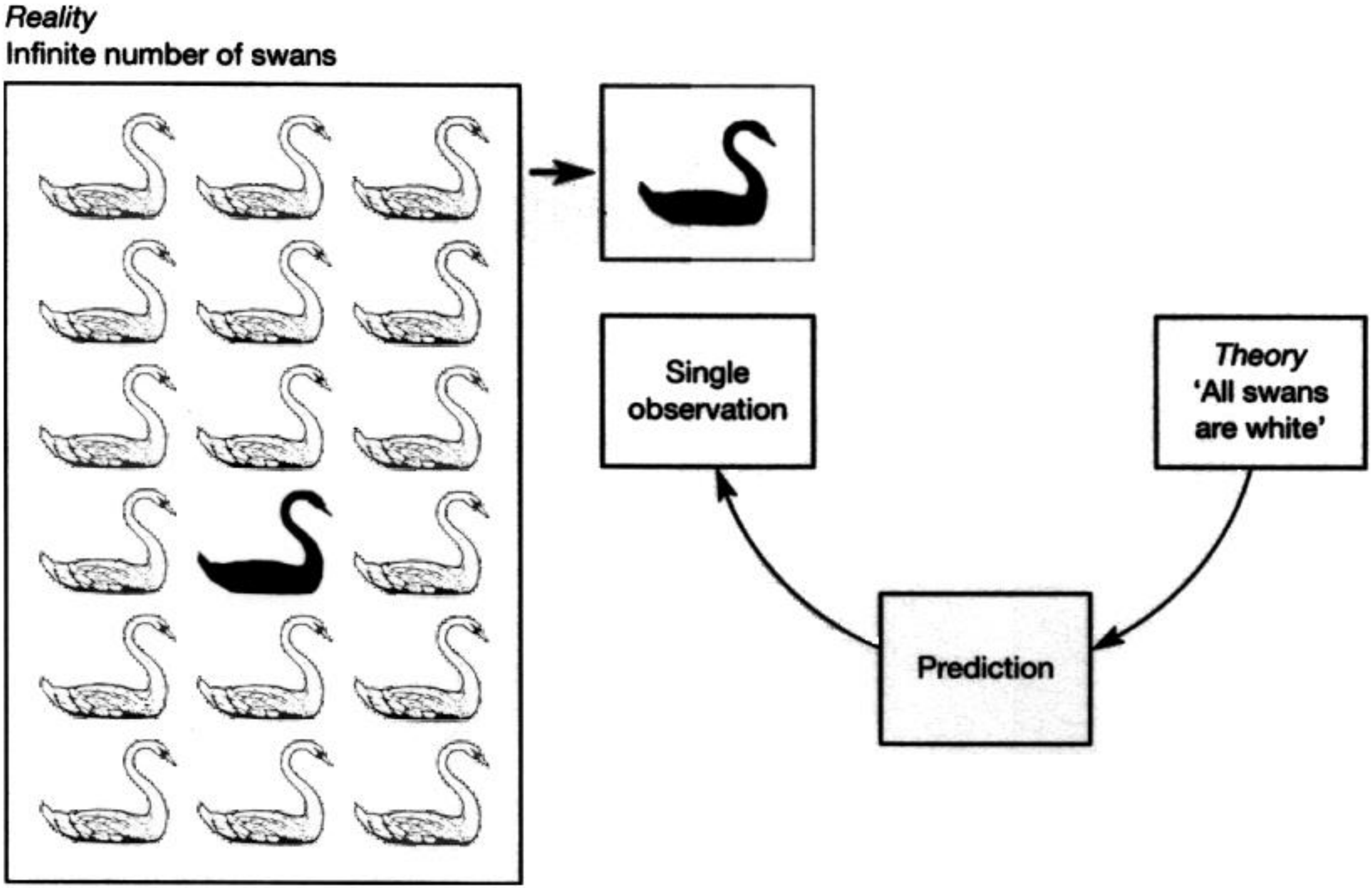
- **A hypothesis can be automatically formed by inductive learning**

Induction is a general strategy of inference process widely used in machine learning. In this method, a model can be formed by drawing inductive inferences from a set of example facts

The process induces new information through generalization from specific set of data which are teacher-supplied or environment-supplied examples.

- How many observations we need in order to validate this theory?

* Everything that science discovers has only temporary value.


Reality
Infinite number of swans
Single observation
Theory
'All swans are white'
Prediction

How many swans are needed?

This number is infinite, since we are speak of all swans.

* No matter how many swans we have seen, we can only be sure of the definitive truth of our hypothesis if we have seen them all.

* It is impossible that we can observe all swans.

<u>How do we corroborate our hypotheses?</u>

How many swans are needed?

This number is infinite, since we are speak of all swans.

* No matter how many swans we have seen, we can only be sure of the definitive truth of our hypothesis if we have seen them all.

* It is impossible that we can observe all swans.

<u>How do we corroborate our hypotheses?</u>

* Uncertainty measure based on statistics.

Falsification

- "A general law can never be verified by a finite number of observations. It can, however, be falsified by only one observation." - Karl Proper

- We need only one observation in order to falsify the theory

  E.g. Falsify the theory “Swans are white” by finding a single black swan.

- We want to develop theories that fit the data but the rules of good science demand that we also need to formulate the exact circumstances under which these theories can be falsified

- Falsification is much more important for scientific work than verification. Why?

Main issues of machine learning

- **Knowledge accuracy:** statistical significance
- **Knowledge representation:** transparency, readability, content richness
- **Supervised vs. unsupervised:** training data & the need of human control
- **Complexity of the search space:** hypothesis space (how many hypothesis there are and how they are related)
- **Bias:** is any mechanism employed by a learning system to constrain the search of a hypothesis

Information Retrieval (IR) vs. DM

- **IR:** To retrieve desired information from an information store, such as a Database, or the Web
- **SQL** is the conventional tool for retrieving information from a relational database
- **Internet** is a new type of information store/database dominated by unstructured textual data which require new information retrieval technology: search engine, including Elasticsearch (scalable & distributed search solution, "Elastic" in 2015)
- IR emphases on finding the original stored information, i.e. not involved in deriving/discovering new information

IR vs. DM

Search engine technology – Index & Selectivity

- Efficient search (indexing tech) - Crawling and indexing: to quickly fetch a large number of web pages into a local repository and to index them based on key words

  Elasticsearch is a search engine based on Lucene. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. It is developed in Java and is released as open source.

- Similarity measure (for user's stated query)

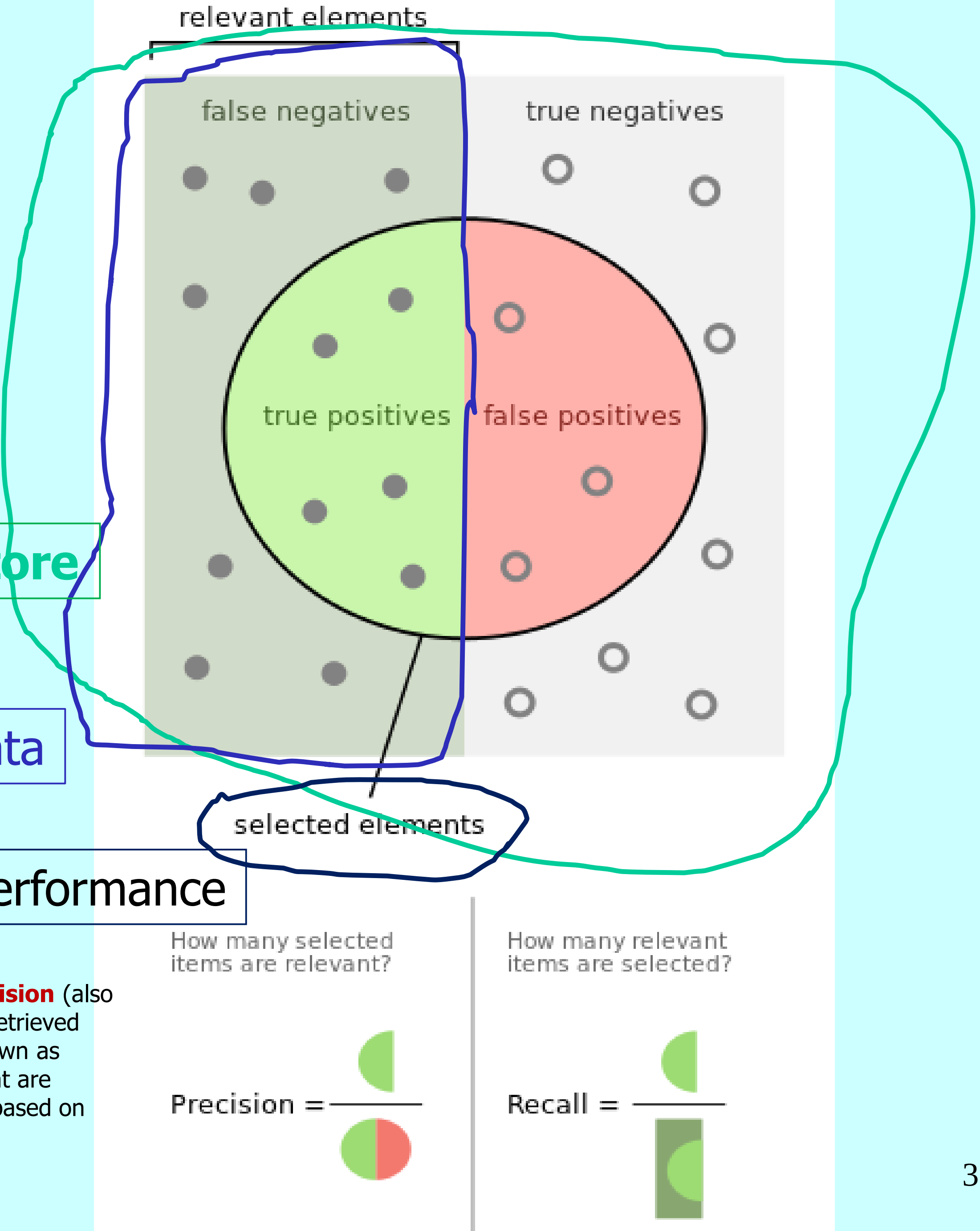
- Evaluation:

  Precision = Retrieved relevant / Retrieved

  Recall = Retrieved relevant / Relevant

IR performance measures:

- **Precision**
- **Recall**



Information store

Relevant data

IR Performance

In both IR and DM with binary classification, **precision** (also called positive predictive value) is the fraction of retrieved instances that are relevant, while **recall** (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

IR vs. DM

- IR’s impact on development of DM
  - Similarity (used in clustering), precision and recall measures (for describing accuracy of predictive modeling technique)
  - E.g., In a classification task, the precision for a class is the *number of true positives.*

- DM supports for establishing index for IR
  - E.g., Text clustering for categorization, etc.

Stats Methods vs. DM

- **Hypothesis, Assumption, and Data Input for Analysis**

- **Hypothesis**
  - **What is a hypothesis?**
    - A **hypothesis** is a proposed explanation for a phenomenon.
    - For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it.

Statistics: - a hypothesis testing based approach, - i.e. have a hypothesis first, then validate it by testing.

DM: - an observation based regularity discovery approach, - which search/discovery/derive a hypothesis that may explain the observations, - i.e. have observations first then based upon to generate a hypothesis which best explain the observation. - It is relative ease with which new insight can be gained.

Stats Methods vs. DM (cont)

- **Assumption on probability distribution**

  Statistics: need strong assumption.

  DM: fewer assumptions or no assumptions at all.

- **Data input (data types and sampling options)**

  Statistics: mainly constrained to numerical data, and require sampling.

  DM: can be any data types, and applied to missive data.

- It is fair to say that statistics traditionally has been used for many of the analyses that are now done with data mining, such as building predictive models or discovering associations in databases.

4. Mining Association Patterns

(Ch6 of 3rd edition, or Ch5 of 2nd edition)

- Association rule mining concepts
- Apriori property and Apriori algorithm
- Mining various kinds of association rules
- Constraint-based association mining

Association Rule (AR) Mining:

“Finding frequent togetherness patterns, then derive ARs”

- AR: Finding interesting association regularities between data items, attributes, concepts, causal structures, sequences, …
- ARs hidden in large databases
  - Each association rule is a piece of knowledge describing a statistics sound relationship between two particular item sets.
  - First proposed by Agrawal, Imielinski and Swami [AIS93]

AR Query Examples:

1. What product items were often purchased together?
2. What are the key factors determining income ≥ \$50,000 for the age group of 25 to 35?
3. What are the subsequent purchases after buying a LED TV?
4. What kinds of DNA are sensitive to a particular new drug?
5. What are the frequent click stream (session) patterns of an E-commerce website?

Review Questions

1. What are the simple rules for choosing solution tools for getting different types of business query information?
2. What are the two general purposes of DM (0r any scientific research)?
3. How the DM technologies are categorized?
4. Can you name three major DM tasks & what is each task about?
5. What is the Empirical Cycle Model (ECM) of scientific research, describe each stage of the process?
6. How to map a DM task to ECM?
7. Why it says "a discovered knowledge only has temporary value"?
8. Why a discovered knowledge needs to be corroborated by statistics?
9. What is the main difference and relationship between IR and DM?
10. What are the main differences between conventional statistical methods and DM techniques?