M L
D M

Machine Learning

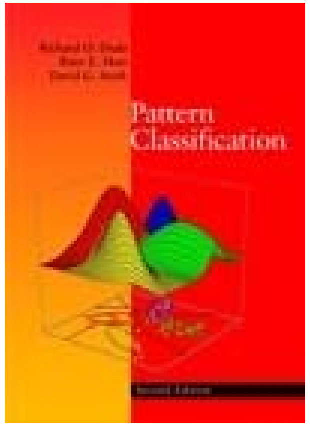
Xin-Shun Xu（许信顺）

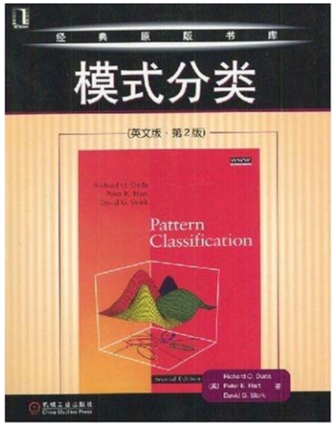
School of Software, Shandong University

Course Information

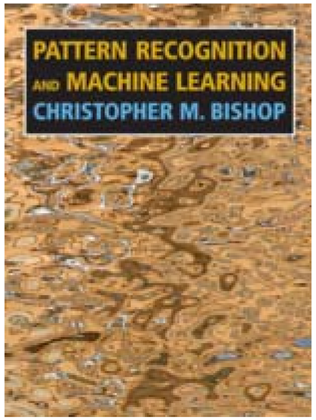
- Credits
  - 3 credits with 32 lecture hours + 32 experiment hours
  - Final Exam—last week
- How to get scores
  - Attendance 5%
  - Projects/experiments 55%
    - Room 303 & 305，
    - 6^{th}-13^{rd} weeks，5:30-9:30/Friday
  - Final Exam 40%

Textbook



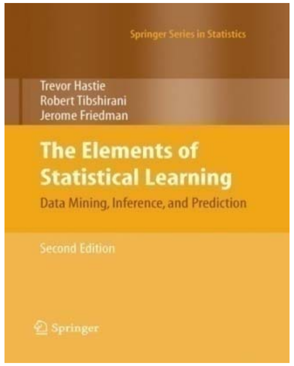


- Richard O. Duda, Peter E. Hart, David G. Stork. Pattern Classification (2nd), John Wiley & Sons, 2000.
- 模式分类（英文版, 第二版），机械工业出版社，2004.

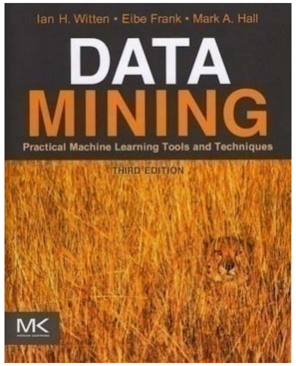


- Christopher M. Bishop. Pattern Recognition and Machine Learning (1st), Springer, 2006.

References

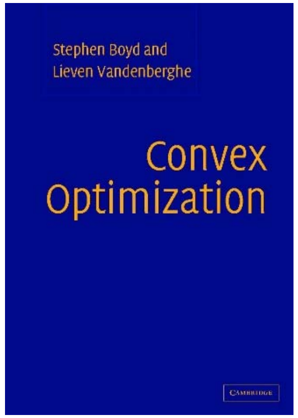


- Hastie, Tibshirani and Friedman, *The Elements of Statistical Learning*: *Data Mining*, *Inference, and Prediction (*2^{nd}*), Springer-Verlag, 2009.*



- Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (3^{rd}), Morgan Kaufmann, 2011.

- Stephen Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.



References

- 周志华，机器学习，清华大学出版社，2016.

- 第一部分　机器学习基础知识
  - 概念，模型评估与选择，线性模型
- 第二部分　经典机器学习方法
  - 决策树，神经网络，支持向量机，贝叶斯、集成学习，聚类模型，降维度量学习
- 第三部分　进阶知识
  - 特征选择与字典学习，计算学习理论，半监督学习，概率图模型，规则学习，强化学习

- Rich Web Resources!

Course Information

MIMA

- Contents
  - I. Introduction
  - II. Bayesian Decision Theory
  - III. Parameter Estimation
  - IV. Non-Parameter Estimation
  - V. Linear Discriminant Functions
  - VI. Neural Networks
  - VII. Decision Trees
  - VIII. Ensemble Learning
  - IX. Support Vector Machine

Course Information

- Mathematical background
  - Linear algebra
  - Probability theory
  - Statistics
  - Information theory

Course Website

http://mima.sdu.edu.cn/members/~xinshunxu

許信順

Xin-Shun Xu

PhD, Professor

MIMA Group
School of Computer Science and Technology,
Shandong University, Jinan, China

SHANDONG UNIVERSITY

MIMA
Machine Intelligence & Media Analysis

Sections:

- >> Biography
- >> Research Interest
- >> Publications
- >> Projects
- >> Courses
- >> Correspondence

Links:

- >> CFP
- >> AI Datasets
- >> IEEE

Biography:

Currently I am a professor of School of Computer Science and Technology at Shandong University and a leader of MIMA (Machine Intelligence and Media Analysis) Group.

I received my M.S. degree in Computer Science from Shandong University, China in 2002, and Ph.D. degree in computer science from University of Toyama, Japan in 2005. In the same year, I joined School of Computer Science and Technology of Shandong University as an associate professor. From 2009 to 2012, I am also a postdoctoral fellow in LAMDA Group, led by professor Zhi-Hua Zhou.

Research Interest:

My research interests include: Machine Learning, Information Retrieval, Computational Intelligence, Bioinformatics, Pattern Recognition, Data mining and Combinatorial Optimization. Now I am working on Multi-Instance Learning, Multi-Label Learning, Semi-supervised Learning, Ensemble Learning and Media Analysis & Retrieval.

Remarks

MIMA

- Lectures are important, but not enough
- You should review what have been taught with more hours than the class hours/week
- You should be familiar with all terminologies related with this course.
- Practice what you have learned

No Pain, No Gain!!!

Chapter 1 Introduction

ACM Turing Award

(2012)
Goldwasser, Shafi
Micali, Silvio

(2011)
Pearl, Judea

(2010)
Valiant, Leslie Gabriel

(2009)
Thacker, Charles P. (Chuck)

(2008)
Liskov, Barbara

(2007)
Clarke, Edmund Melson
Emerson, E. Allen
Sifakis, Joseph

(2006)
Allen, Frances ("Fran") Elizabeth

(2005)
Naur, Peter

(2004)
Cerf, Vinton ("Vint") Gray
Kahn, Robert ("Bob") Elliot

(2003)
Kay, Alan

(2002)
Adleman, Leonard (Len) Max
Rivest, Ronald (Ron) Linn
Shamir, Adi

(2001)
Dahl, Ole-Johan *
Nygaard, Kristen *

(2000)
Yao, Andrew Chi-Chih

(1999)
Brooks, Frederick ("Fred")

(1998)
Gray, James ("Jim") Nicholas *

(1997)
Engelbart, Douglas *

(1996)
Pnueli, Amir *

(1995)
Blum, Manuel

(1994)
Feigenbaum, Edward A ("Ed")
Reddy, Dabbala Rajagopal ("Raj")

(1993)
Hartmanis, Juris
Stearns, Richard ("Dick") Edwin

(1992)
Lampson, Butler W

(1991)
Milner, Arthur John Robin Gorell ("Robin") *

(1990)
Corbato, Fernando J ("Corby")

(1989)
Kahan, William ("Velvel") Morton

(1988)
Sutherland, Ivan

(1987)
Cocke, John *

(1986)
Hopcroft, John E
Tarjan, Robert (Bob) Endre

(1985)
Karp, Richard ("Dick") Manning

(1984)
Wirth, Niklaus E

(1983)
Ritchie, Dennis M.*
Thompson, Kenneth Lane

(1982)
Cook, Stephen Arthur

(1981)
Codd, Edgar F. ("Ted") *

(1980)
Hoare, C. Antony ("Tony") R.

(1979)
Iverson, Kenneth E. ("Ken") *

(1978)
Floyd, Robert (Bob) W *

(1977)
Backus, John *

(1976)
Rabin, Michael O.
Scott, Dana Stewart

(1975)
Newell, Allen *
Simon, Herbert ("Herb") Alexander *

(1974)
Knuth, Donald ("Don") Ervin

(1973)
Bachman, Charles William

(1972)
Dijkstra, Edsger Wybe *

(1971)
McCarthy, John *

(1970)
Wilkinson, James Hardy ("Jim") *

(1969)
Minsky, Marvin

(1968)
Hamming, Richard W*

(1967)
Wilkes, Maurice V.*

(1966)
Perlis, Alan Jay *

What are machine learning and data mining?

What is Data Mining?

MIMA

- Data Mining
  - Data mining is the analysis of (often LARGE) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. [D. Hand et al., Principles of Data Mining]

数据挖掘是通过对(大规模)观测数据集的分析，寻找确信的关系，并将数据以一种可理解的且利于使用的新颖方式概括数据的方法.

What can be mined out?

MIMA

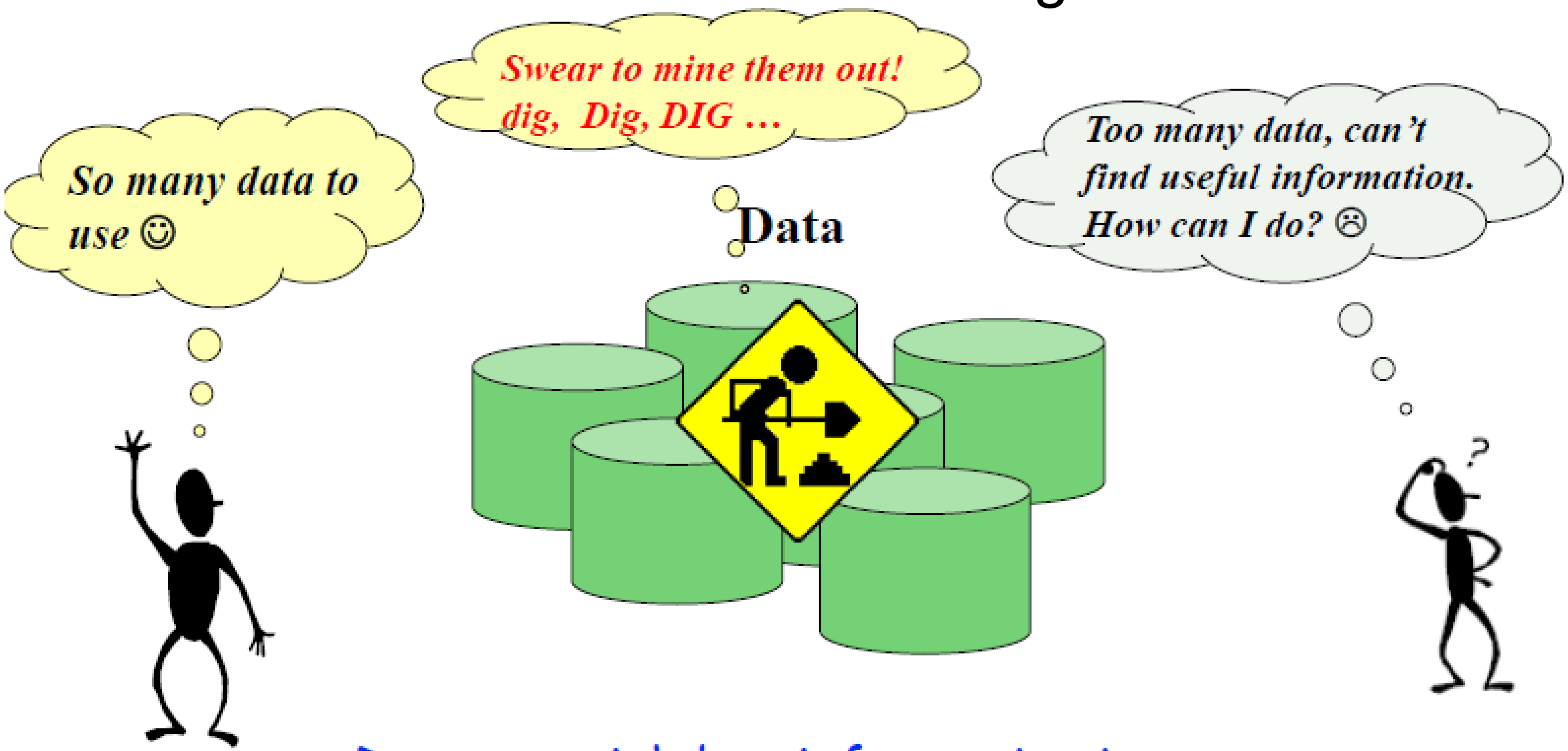
- Knowledge
- Hidden relationships,
- Underlying reality

Information that should be potentially useful for decision making or understanding the nature of the task.

Why Data Mining

MIMA

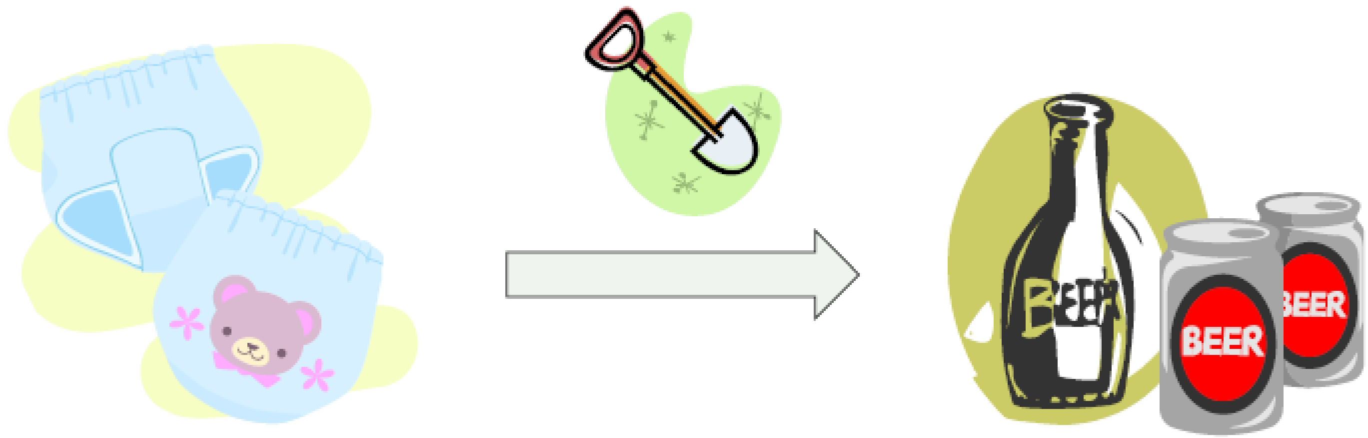
- 2010.9 Flickr has 5 billion images 3000/minute
- 2009.7 YouTube has 0.2 billion-hour videos, 35
- 2009.2 Facebook has 15 billion images



Data are rich but information is poor

Example (I)

- Mining supermarket transactions



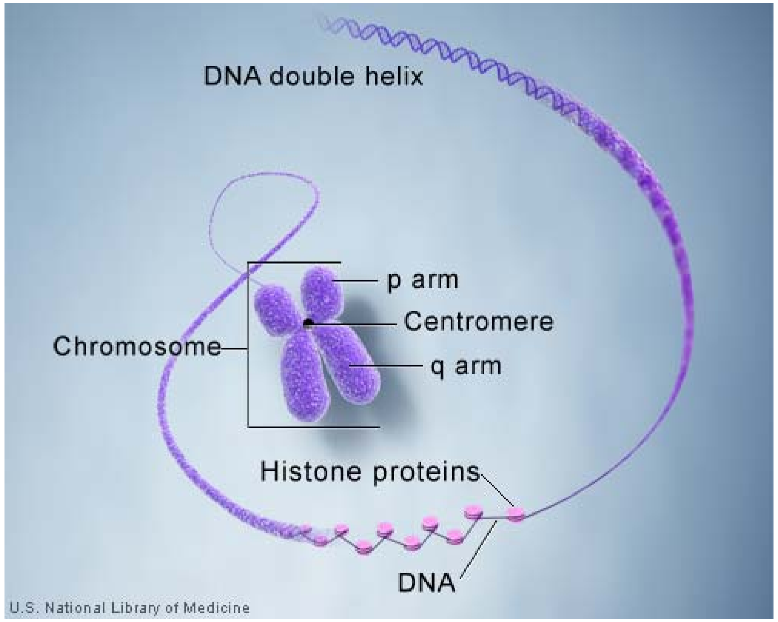
Example (II)

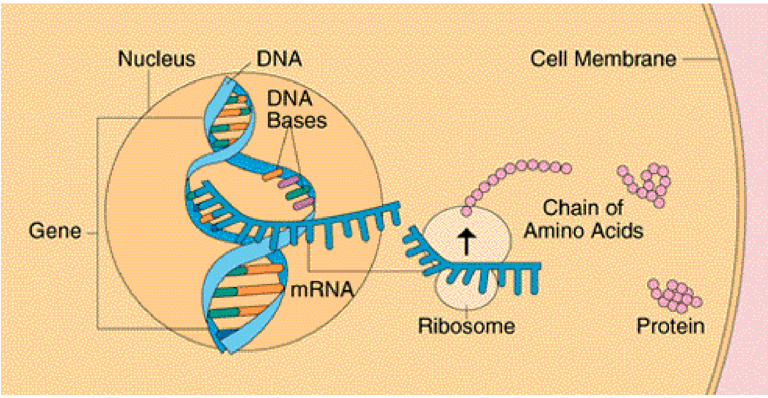
- Mining the valuable customers.

Example (III)

- Decide whether the current access is intrusion.
  - Building models based on historical records, e.g., decision tree, neural networks
  - Anomaly detection

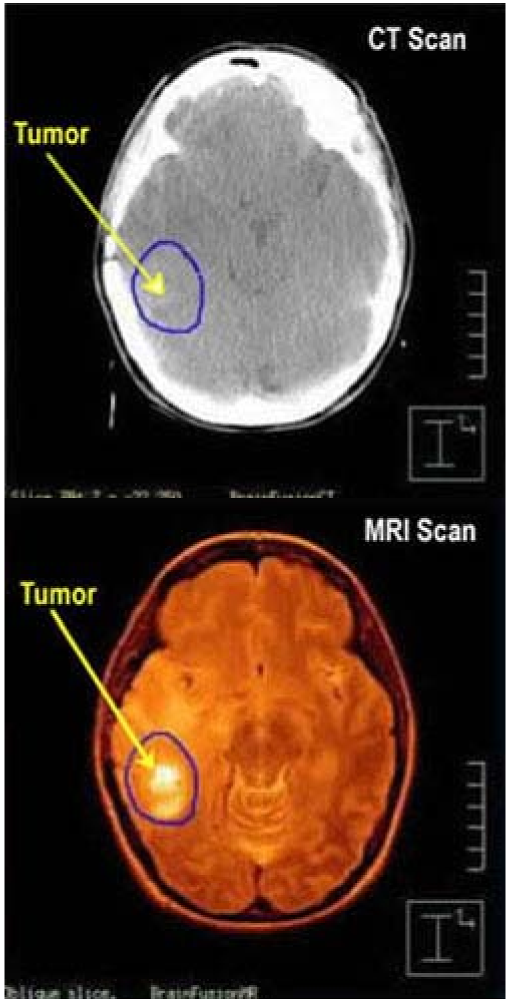
Example (IV)





- Identify gene function
- Finding key genes
- Identifying gene expression patterns
- Identifying gene interactions
- Finding similar genes.

Example (V)



- Cancer Diagnosis
  - Given data on expression levels of genes, classify the type of tumor.
  - Discover categories of tumors having different characteristics.
- CAD(Computer-Aided Diagnosis) systems help improving the diagnosis of doctors based on the historical cases.

Example (VI)

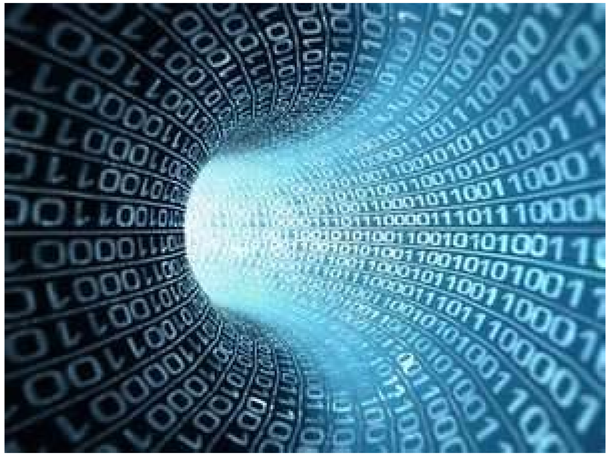
- Ming the web structures, web content, web usage and web logs to identify interesting patterns for web search, user behavior identification, even relationships among people.

Example (VII)

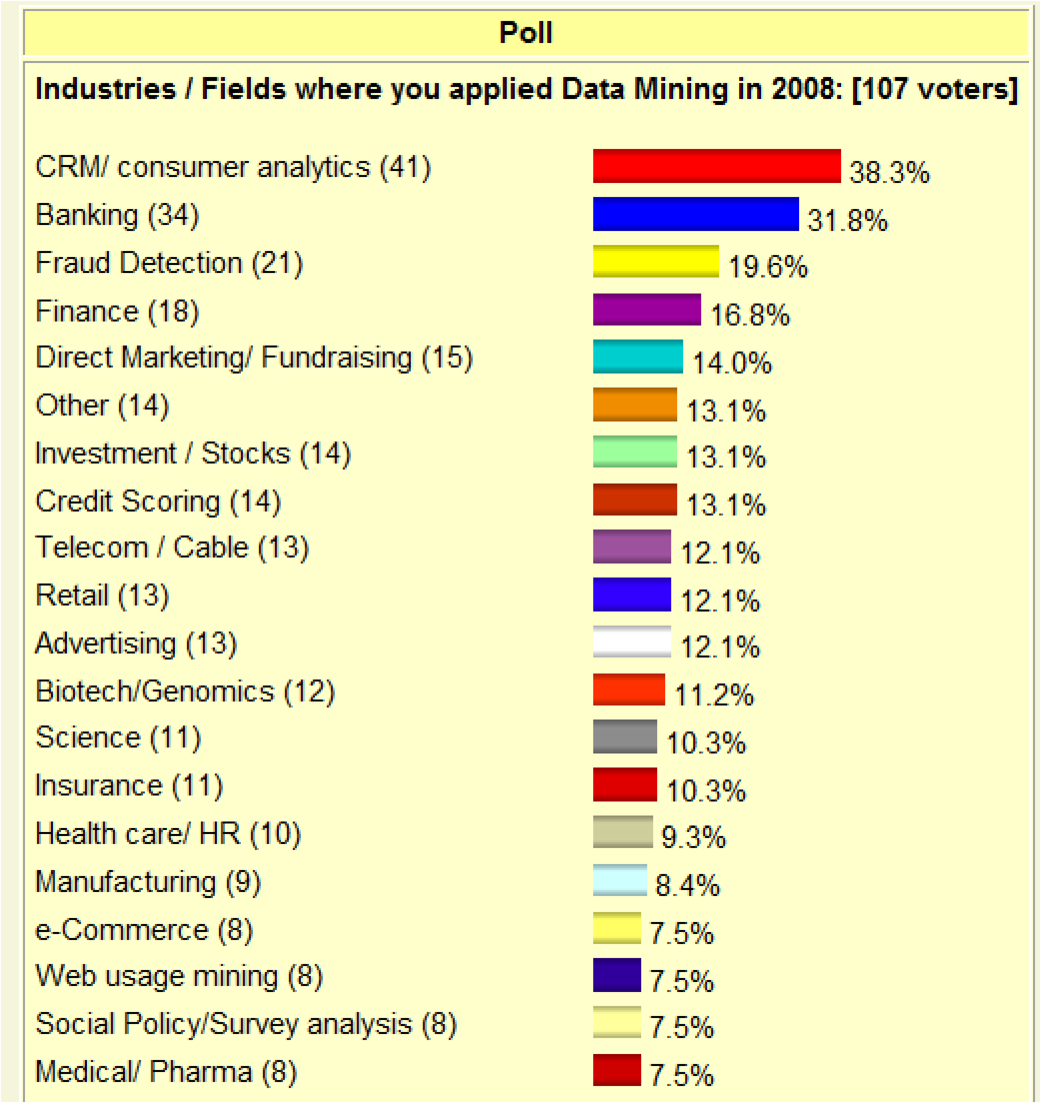
- Mining financial data
  - Fraud detection
  - Stock trends discovery

Example (VIII)

- Marketing
  - Given data on age, income, etc., predict how much each customer spends.
  - Discover how the spending behaviors of customers are related.
  - Fair amount of data on each customer, but messy
  - May have data on a very large number of customers.



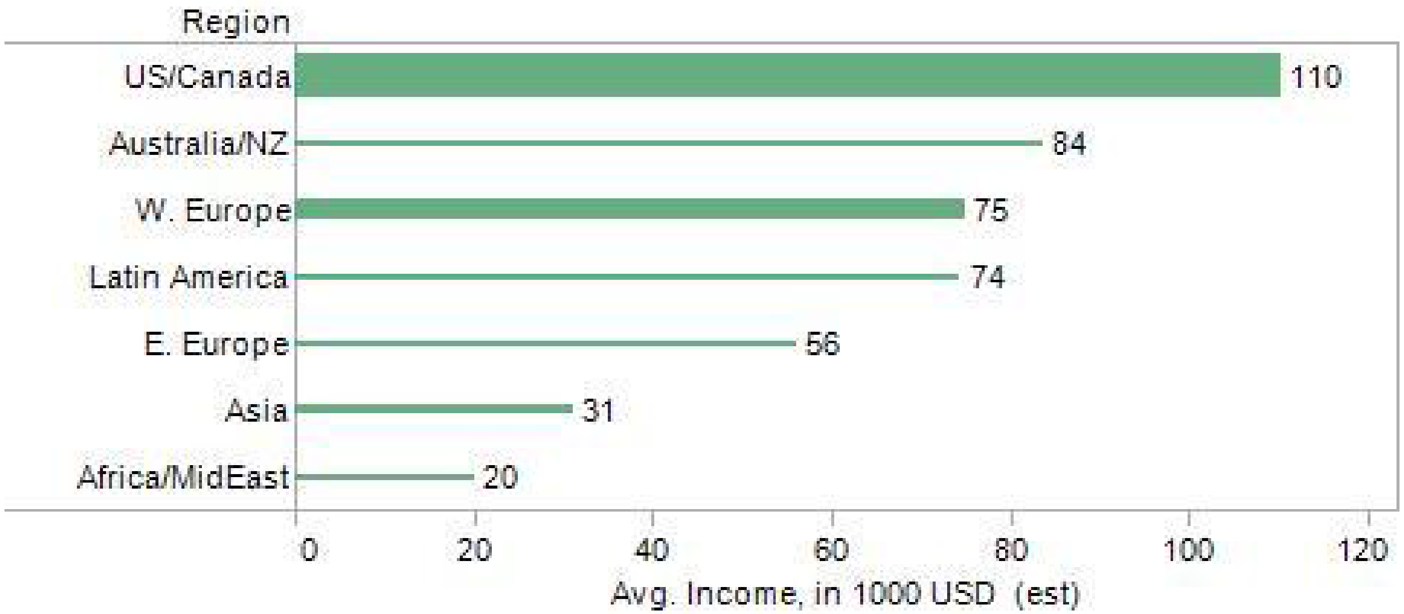
Top DM Fields



- In 20 fields, 9 of them are related to finance/commerce.
- 8 of these fields are among the top 10
- Telecom, bioinform. Medical, health, …, are also top fields.

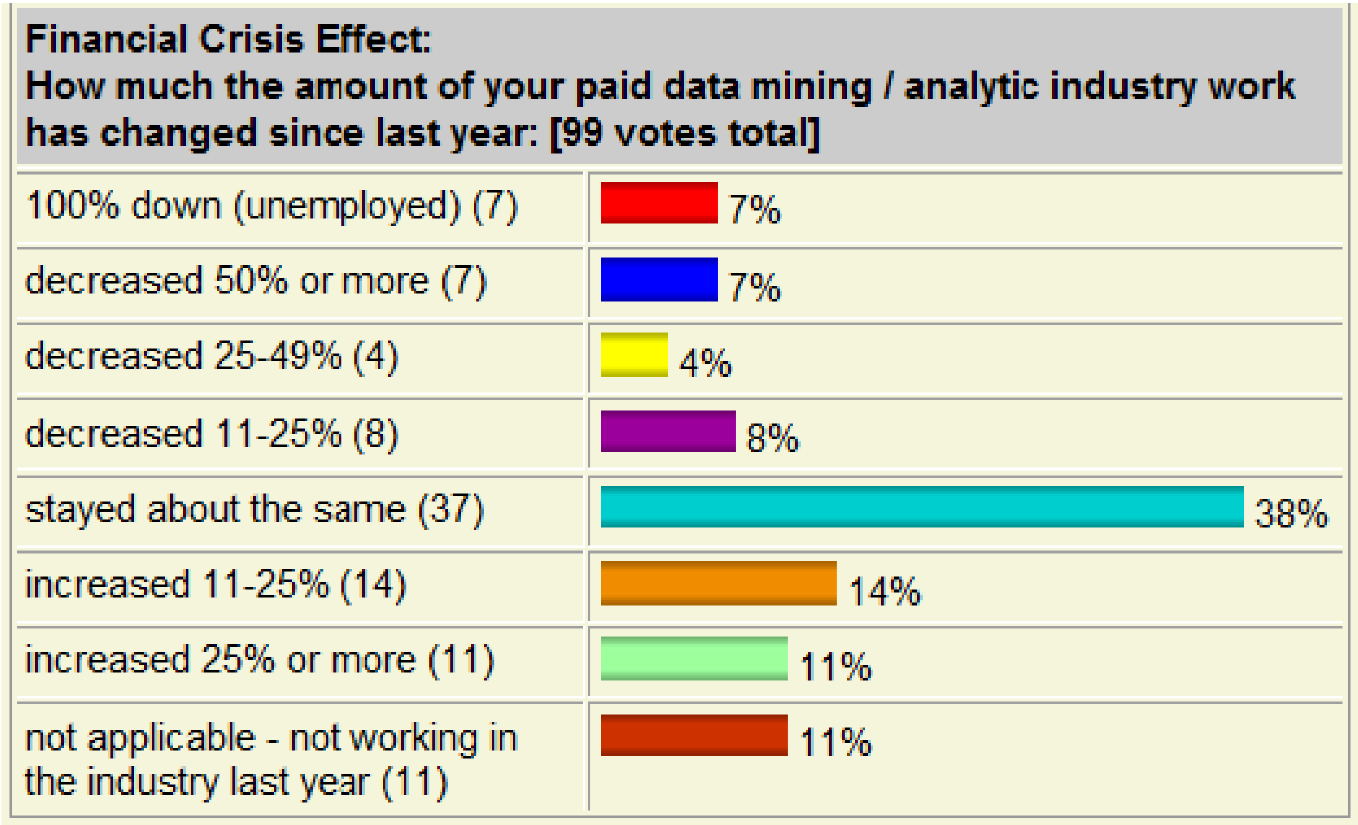
[KDNudges Poll, 2009]

Salaries of data miners



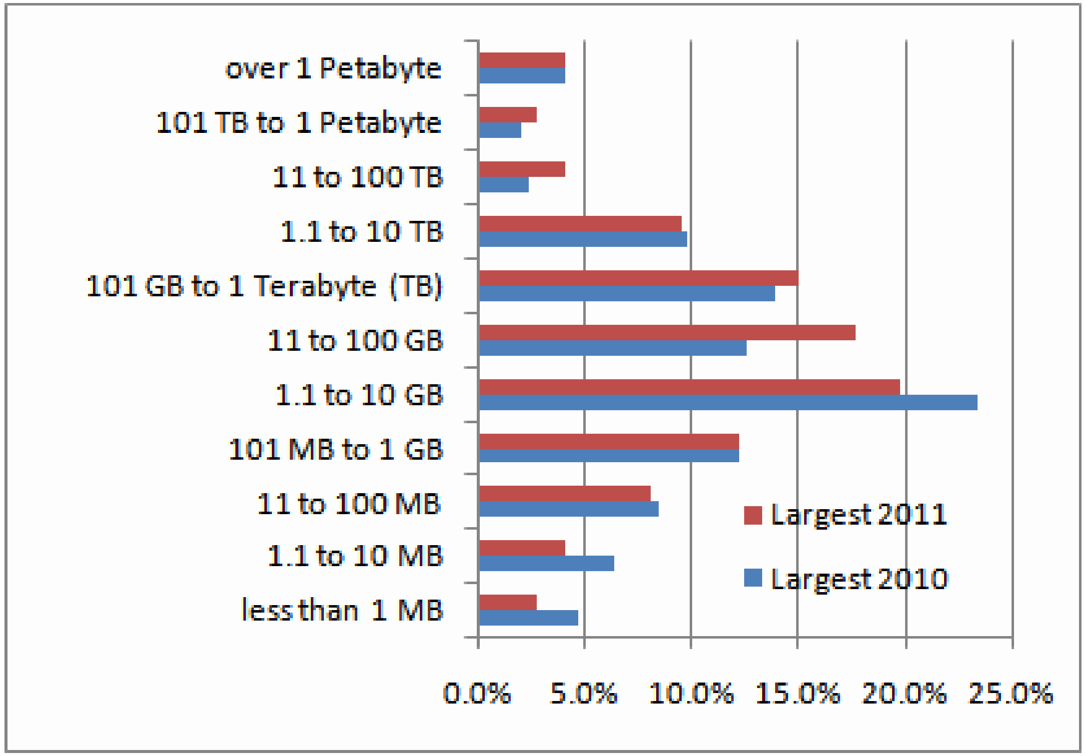
[KDNudges Poll, 2009]

Financial Crisis Effect on DM



[KDNudges Poll, 2009]

How large can the data set be



[KDNudges Poll, 2011]

Region (voters)	Largest Dataset Analyzed (median)	% analyzed TB+ data
US/Canada (53)	11-100 GB	30.2%
Europe (49)	11-100 GB	18.4%
Asia (20)	1-10 GB	10%
Latin America (15)	1 GB	6.7%
Africa/Middle East (7)	1-10 GB	28.6%

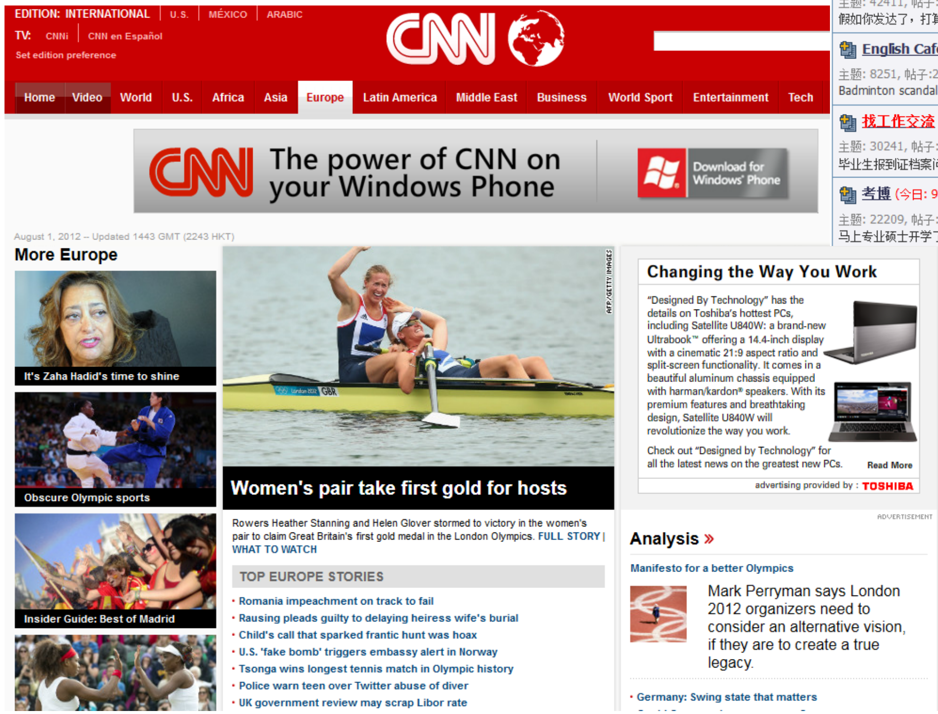
Data types in mining tasks

- Text data

Data types in mining tasks

MIMA

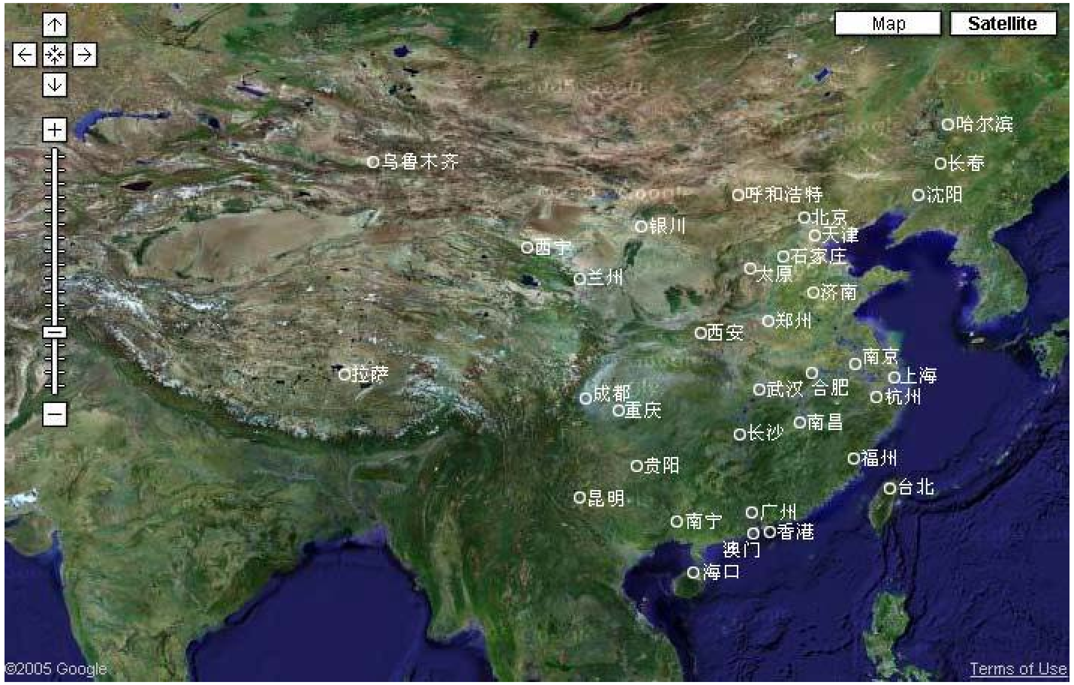
- Web data
  - Weibo
- Social media



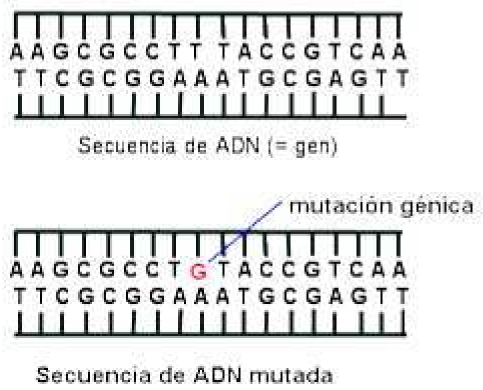
Data types in mining tasks

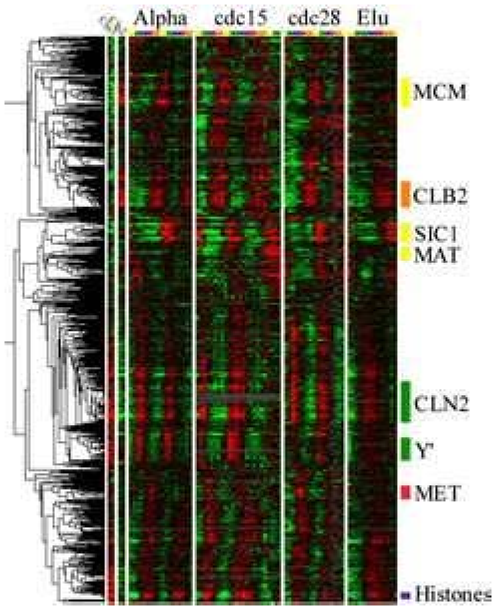
Data types in mining tasks

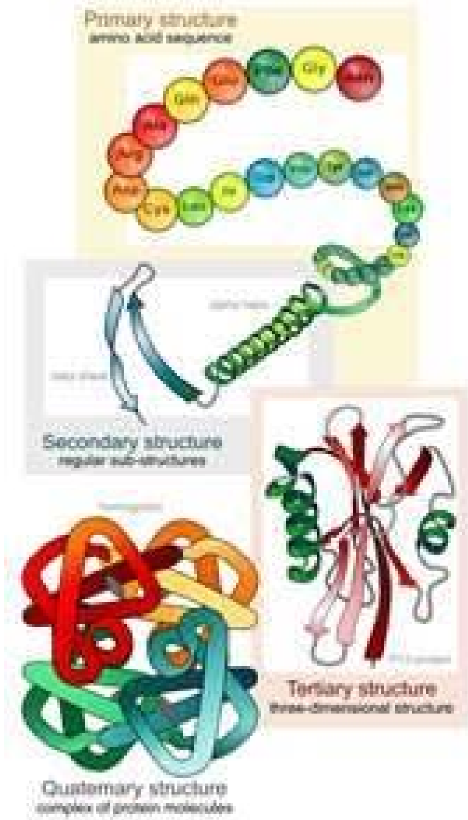
- Temporal and spatial data



Data types in mining tasks







Data types in mining tasks

MIMA

Types of Data Analyzed/Mined in the past 12 months [144 voters]	
table data (fixed # of columns) (102)	70.8%
time series (56)	38.9%
itemsets / transactions (52)	36.1%
text (free-form) (43)	29.9%
anonymized data (38)	26.4%
social network data (28)	19.4%
other (22)	15.3%
web content (19)	13.2%
XML data (17)	11.8%
web clickstream (15)	10.4%
email (15)	10.4%
images / video (11)	7.6%
music / audio (3)	2.1%

- Compared to 09
  - Decline of spatial data/audio/text (free-form)
  - Increase of social network data/web content/XML data.

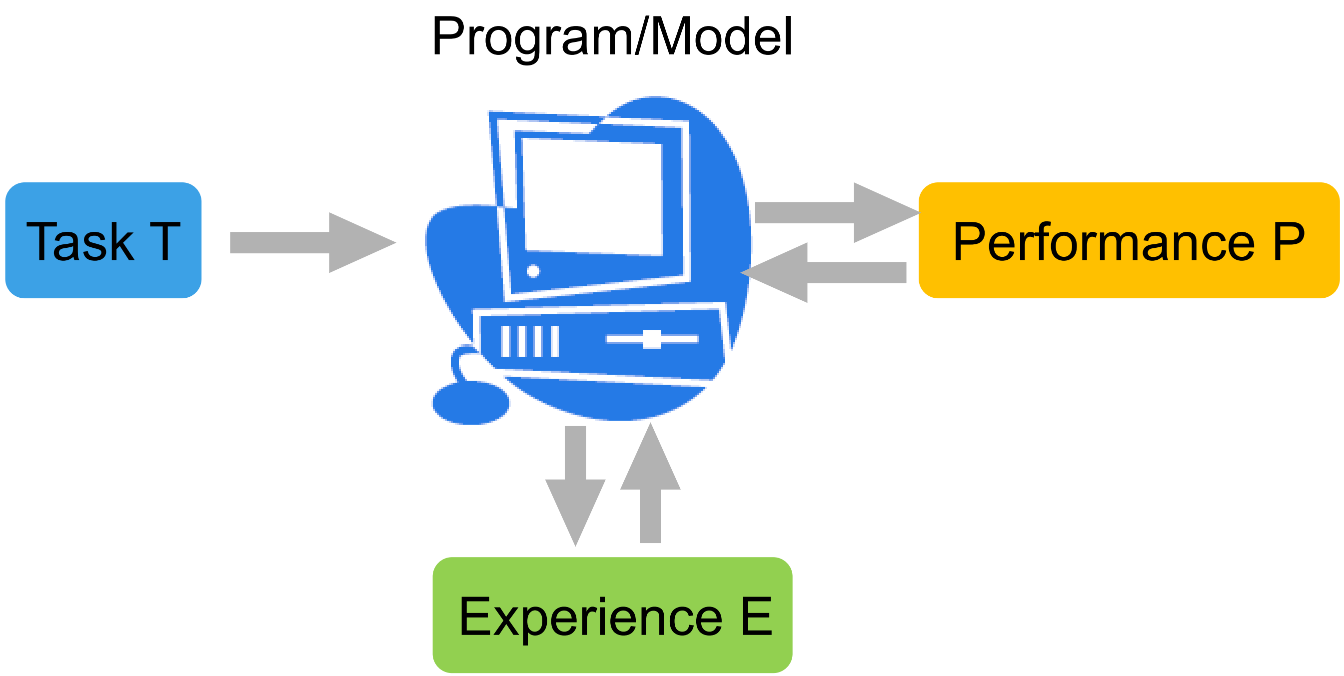
[KDNudges poll, 10]

So, what is DM

- Large
  - Small data does not requires DM.
- Observational data
  - Not experimental data. No control on data collection.
- Unsuspected relationships
  - Relationships should be correct and significance.
- Novel
  - Common sense is useless.
- Understandable
  - The miming results will present to user for making decision.
- Useful
  - The mining results should be useful to users.

What is Machine Learning?

- Machine learning/learning
  - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. [Tom Mitchell, Machine Learning]



Typical ML Problems

MIMA

- Document Search
  - Given counts of words in a document, determine what its topic is.
  - Group documents by topic without a pre-specified list of topics.
  - Many words in a document, many, many documents available on the web.
- Image/Video Understanding
  - Given an image/a video, determine what objects it contains.
  - Determine what semantics it contains
  - Determine what actions it contains.

Typical ML Problems

- Cancer Diagnosis
  - Given data on expression levels of genes, classify the type of tumor.
  - Discover categories of tumors having different characteristics.
- Marketing
  - Given data on age, income, etc., predict how much each customer spends.
  - Discover how the spending behaviors of customers are related.
  - Fair amount of data on each customer, but messy
  - May have data on a very large number of customer.

Machine Learning Framework

MIMA

- Supervised learning
- Unsupervised learning
- Reinforcement Learning
- Semi-supervised learning
- Active learning
- …

Supervised Learning

- In the ML literature, a supervised learning problem has the following characteristics:
  - We are primarily interested in prediction.
  - We are interested in predicting only one thing.
  - The possible values of what we want to predict are specified, and we have some training cases for which its value is known.
- The thing we want to predict is called the target or the response variable.
- Usually, we need training data (test set?)

Supervised Learning

- For classification problem, we want to predict the class of an item.
  - The type of tumor, the topic of a document, the semantics contained in an image, whether a customer will purchase a product.
- For a regression problem, we want to predict a numerical quantity.
  - The amount of customer spends, the blood pressure of a patient, etc.
- To make predictions, we have various inputs,
  - Gene expression levels for predicting tumor type, age and income for predicting amount spent, the features of images with known semantics

How to Make Predictions?

MIMA

- main methods
  - We can train a model by using the training data to estimate parameters of it, then use these parameters to make predictions for the test data.
  - Such approaches save computation when we make predictions for test data. That is we estimate the parameters once, and use them many times.
  - e.g. Linear regression

$$\widehat{y} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_j$$

- other methods
  - Nearest-Neighbor like methods

Nearest-Neighbor Methods

MIMA

- Make predictions for test data based on a subset of training cases, e.g., by approximating the mean, median or mode of P(y|x).

$$\widehat{y}(x) = \frac{1}{K} \sum_{i \in N_K(x)} y_i$$

- Big question: How to choose K?
- If K is too small, we may “overfitting”, but if K is too big, we will average over training cases that aren’t relevant to the test case.

Comparisons

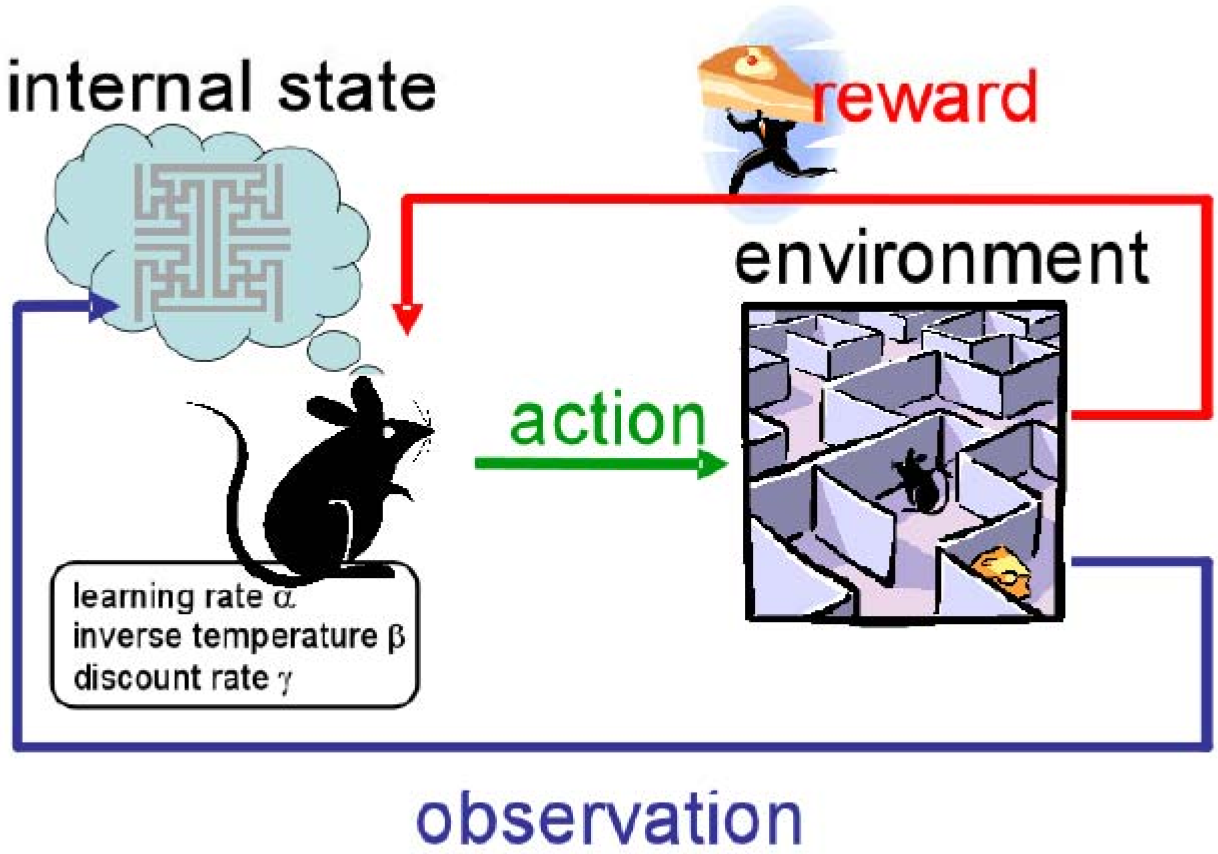
- These two methods are opposite w.r.t. computation.
  - NN like methods are memory-based methods. We need to remember all the training data.
  - Linear regression, after getting parameters, can forget the training data, and just use the parameters.
- They are also opposite w.r.t. to statistical properties.
  - NN makes few assumptions about the data, and has a high potential for overfitting.
  - Linear regression makes strong assumption about the data, and consequently has a high potential for bias.

Unsupervised Learning

MIMA

- For an unsupervised learning problem, we do not focus on prediction of any particular thing, but rather try to find interesting aspects of the data.
- Clustering
- Examples
  - We may find clusters of patients with similar symptoms, which we call diseases.
  - We may find clusters of large number of images.

Reinforcement Learning



Reinforcement Learning

MIMA

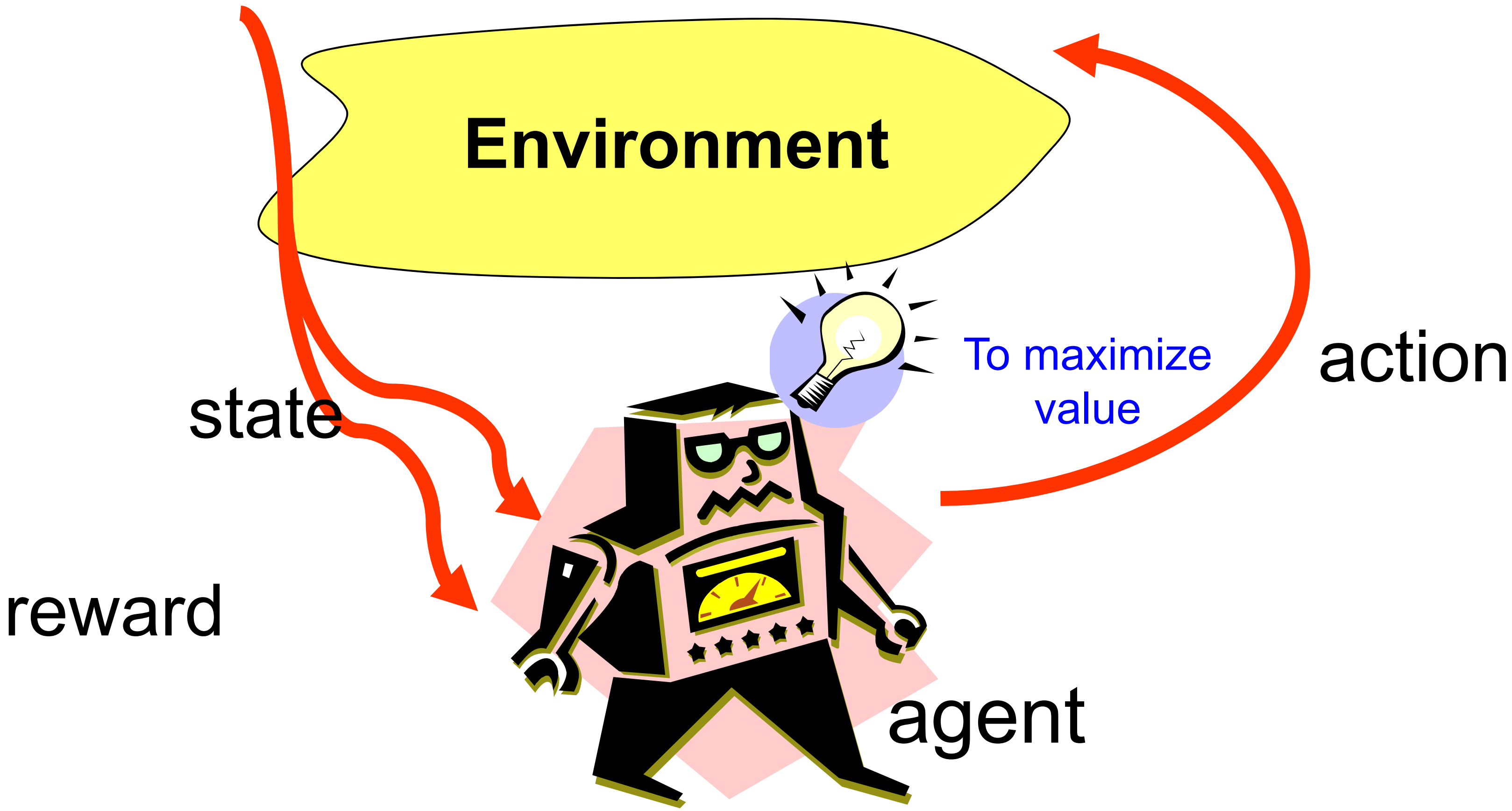
- Learning from interaction (with environment)
- Goal-directed learning
- Learning what to do and its effect
- Trial-and-error search and delayed reward
  - The two most important distinguishing features of reinforcement learning

Reinforcement Learning

- The agent has to *exploit* what it already knows in order to obtain reward, but it also has to *explore* in order to make better action selections in the future.
- Dilemma – neither exploitation nor exploration can be pursued exclusively without failing at the task.

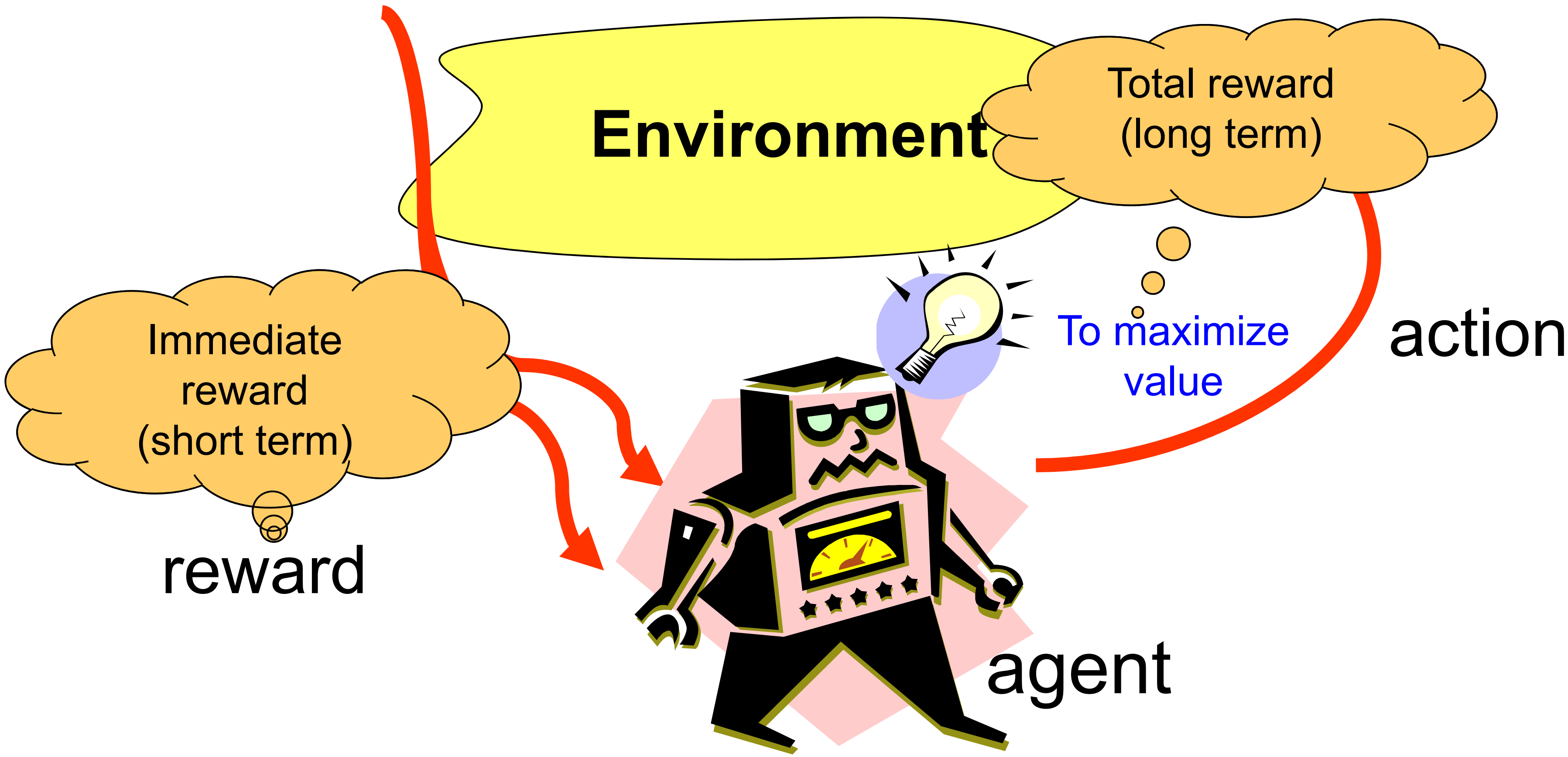
Reinforcement Learning

- Main Elements



Reinforcement Learning

- Main Elements



Reinforcement Learning

MIMA

- Example (Bioreactor)
- State
  - current temperature and other sensory readings, composition, target chemical
- Actions
  - how much heating, stirring are required?
  - what ingredients need to be added?
- Reward
  - moment-by-moment production of desired chemical

Reinforcement Learning

- Example (Pick-and-Place Robot)
- State
  - current positions and velocities of joints
- Actions
  - voltages to apply to motors
- Reward
  - reach end-position successfully, speed, smoothness of trajectory

Reinforcement Learning

MIMA

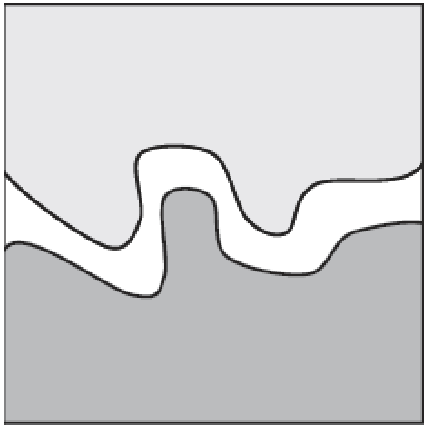
- Example (Recycling Robot)
- State
  - charge level of battery
- Actions
  - look for cans, wait for can, go recharge
- Reward
  - positive for finding cans, negative for running out of battery

Semi-supervised Learning

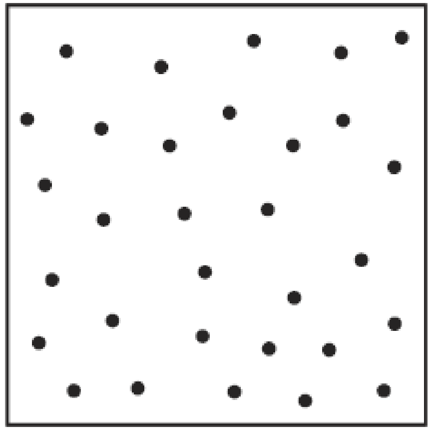
MIMA

- As the name suggests, it is in between Supervised and Unsupervised learning techniques w.r.t the amount of labelled and unlabelled data required for training.
- With the goal of reducing the amount of supervision required compared to supervised learning.
- At the same time improving the results of unsupervised clustering to the expectations of the user.

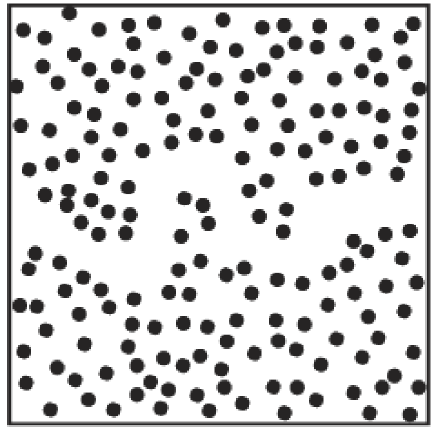
Semi-supervised Learning



Original decision boundary



When only labeled data is Given.



With unlabeled data along with labeled data

With lots of unlabeled data the decision boundary becomes apparent.

Overview of SSL

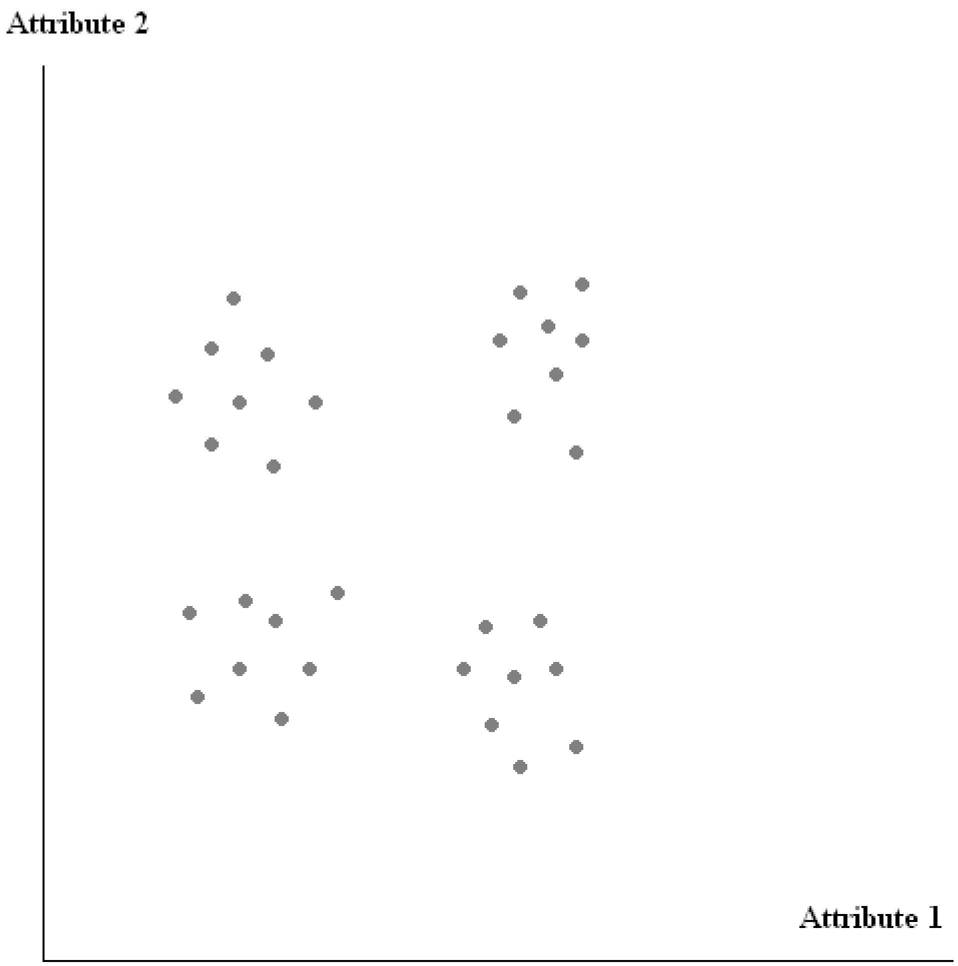
MIMA

- Constrained Clustering
- Distance Metric Learning
- Manifold based Learning
- Sparsity based Learning (Compressed Sensing).
- Active Learning

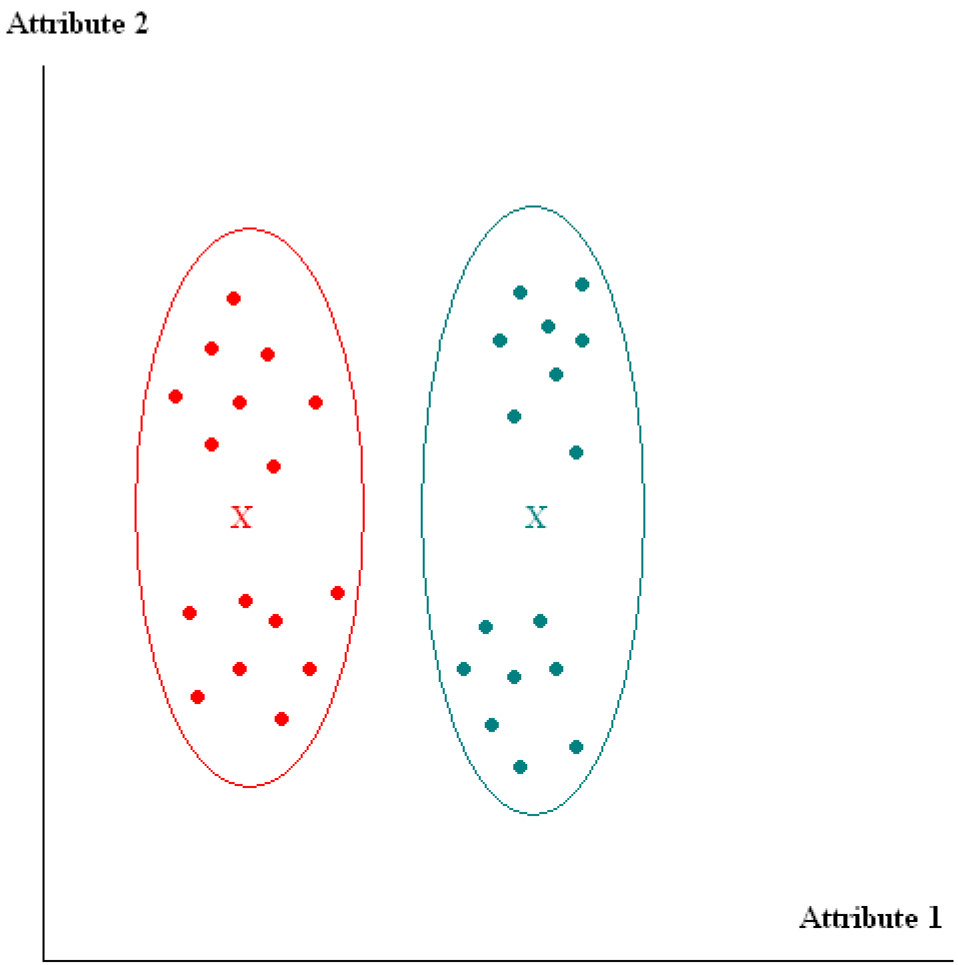
Constrained Clustering

- When we have any of the following:
  - Class labels for a subset of the data.
  - Domain knowledge about the clusters.
  - Information about the 'similarity' between objects.
  - User preferences.
- May be pairwise constraints or a labeled subset.
  - ***Must-link*** or ***cannot-link*** constraints.
  - Labels can always be converted to pairwise relations.
- Can be clustered by searching for partitionings that respect the constraints.
- Recently the trend is toward similarity-based approaches.

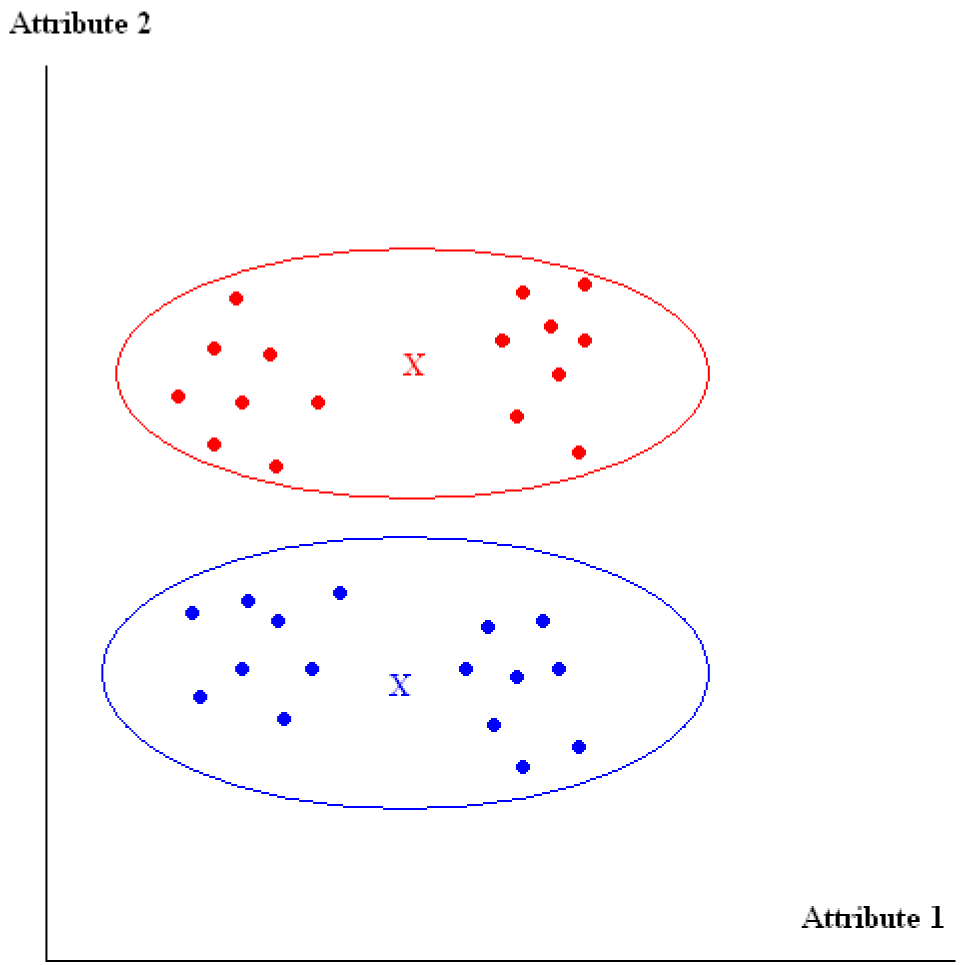
Constrained Clustering



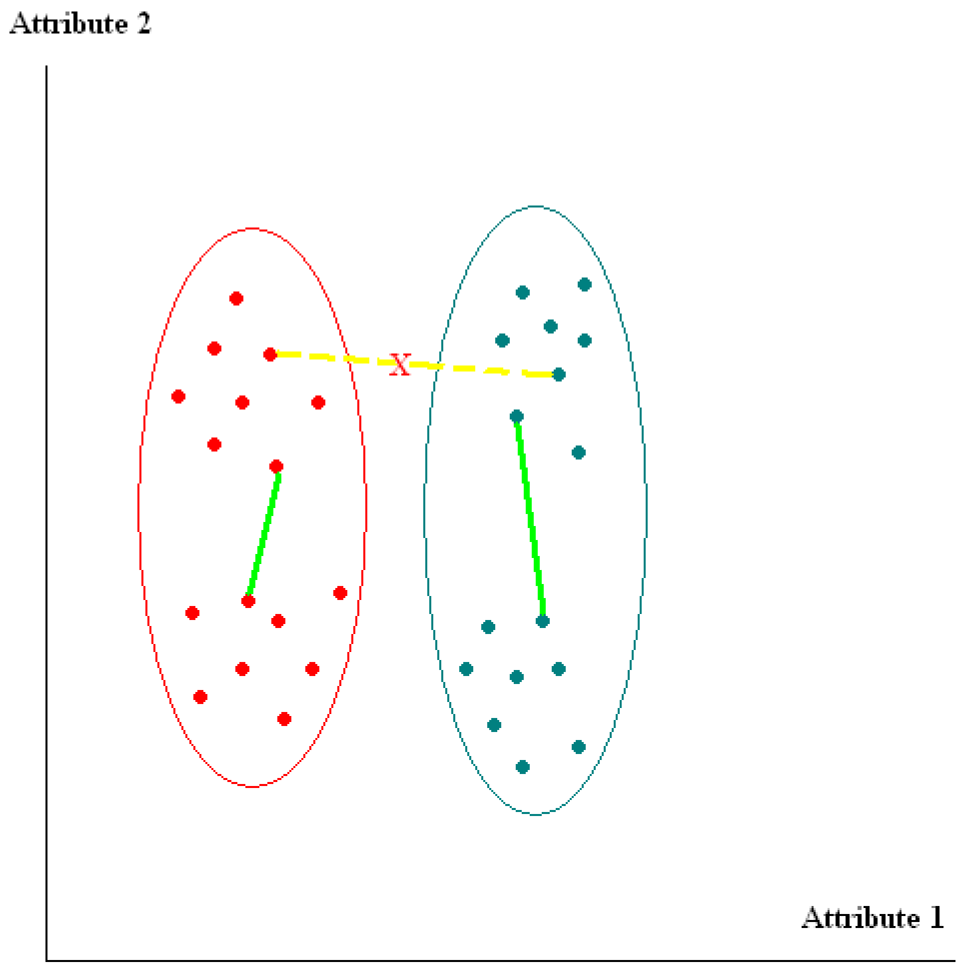
Constrained Clustering



Constrained Clustering



Constrained Clustering



Active Learning

MIMA

- Basic idea:
  - Traditional supervised learning algorithms passively accept training data.
  - Instead, query for annotations on informative images from the unlabeled data.
  - Theoretical results show that large reductions in training sizes can be obtained with active learning!

But how to find images that are the most informative ?

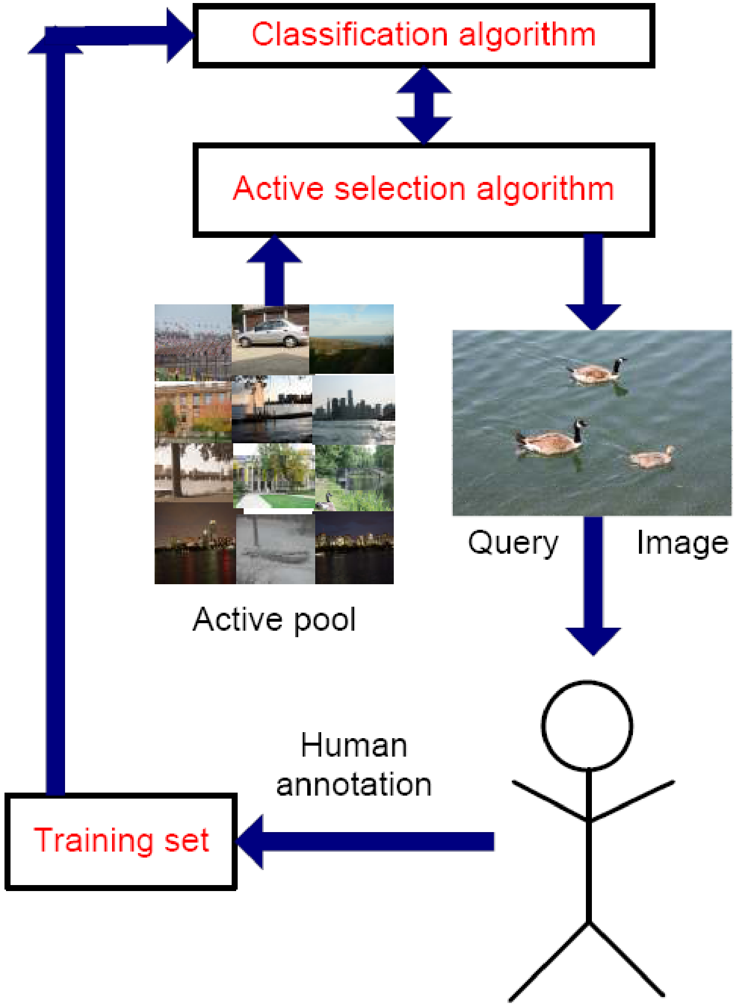
Active Learning

MIMA

- One idea uses uncertainty sampling.
- Images on which you are uncertain about classification might be informative!

- What is the notion of uncertainty?
  - Idea: Train a classifier like SVM on the training set.
  - For each unlabeled image, output probabilities indicating class membership.
  - Estimate probabilities can be used to infer uncertainty.
  - A one-vs-one SVM approach can be used to tackle multiple classes.

Active Learning



Other frameworks

MIMA

- Multi-label learning
- Multi-instance learning
- Multi-instance multi-label learning
- Deep learning

ML&DM Problems Versus Traditional Statistics

- Motivations:
  - Prediction----Understanding----Causality
- Much traditional statistics is motivated primarily by showing that one factor causes another. Understanding comes next, prediction last. Such methods often assume linear relationships (perhaps after simple transformations), simple distributions(e.g. normal).
- In machine learning and data mining, the order is usually reversed: prediction is most important. Such problems have a large number of variables

Attitudes in ML&DM and Traditional Statistics

MIMA

ML&DM

1. No settled philosophy or widely accepted theoretical framework.
2. Willing to use ad hoc methods if they seem to work well (though appearances may be misleading).
3. Emphasis on automatic methods with little or no human intervention.
4. Methods suitable for many problems.
5. Heavy use of computing.

Trad. Statis.

1. Classical (frequentist) and Bayesian philosophies compete.
2. Reluctant to use methods without some theoretical justification (even meaningless)
3. Emphasis on use of human judgement assisted by plots and diagnostics.
4. Models based on scientific knowledge.
5. Originally designed for hand calculation, but computing is now very important.

Challenges for ML

MIMA

- Handling complexity
  - Involve many variables, how can we handle this complexity without getting into trouble.
- Optimization and Integration
  - Usually involve finding the best values for some parameters (an optimization problem), or average over many plausible values (an integration problem). How can we do this efficiently when there are many parameters.
- Visualization
  - Understanding what's happening is hard, 2D? 3D?

Challenges for ML

MIMA

- All these challenges become greater when there are many variables or parameters ----the so-called “curse of dimensionality”.
- But more variables also provide more information
- A blessing? A curse?

How to handle complexity --I

- Properly dealing with complexity is a crucial issue for machine learning.
- Limiting complexity is one approach
  - Use a model that is complex enough to represent the essential aspects of the problem, but that is not so complex that overfitting occurs.
  - Overfitting happens when we choose parameters of a model that fit the data we have very well, but do poorly on new data (poor generalization ability).
  - Cross-validation, regularization,

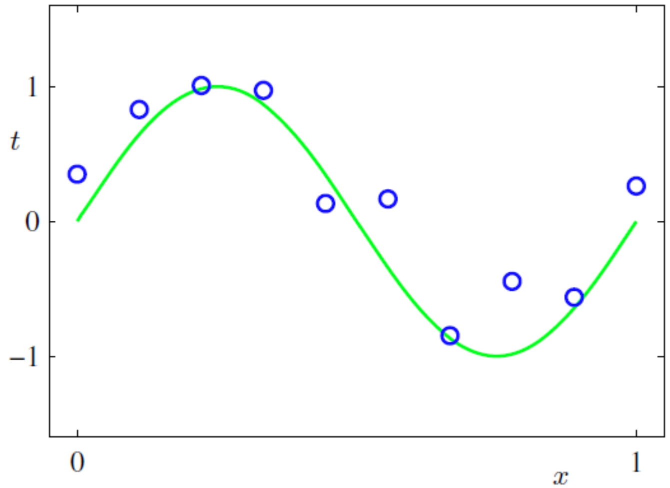
How to handle complexity --II

MIMA

- Reducing dimensionality is another possibility.
  - It is apparent that things become simpler if can find out how to reduce the large number of variables to a small number.
- Averaging over complexity is the Bayesian approach.
  - Use as complex a model might be needed, but don't choose a single parameter values. Instead, average the predictions found using all the parameter values that fit the data reasonably well, and which are plausible for the problem

Example of complexity

- Points are generated for $\sin(2\pi x)$+noise.

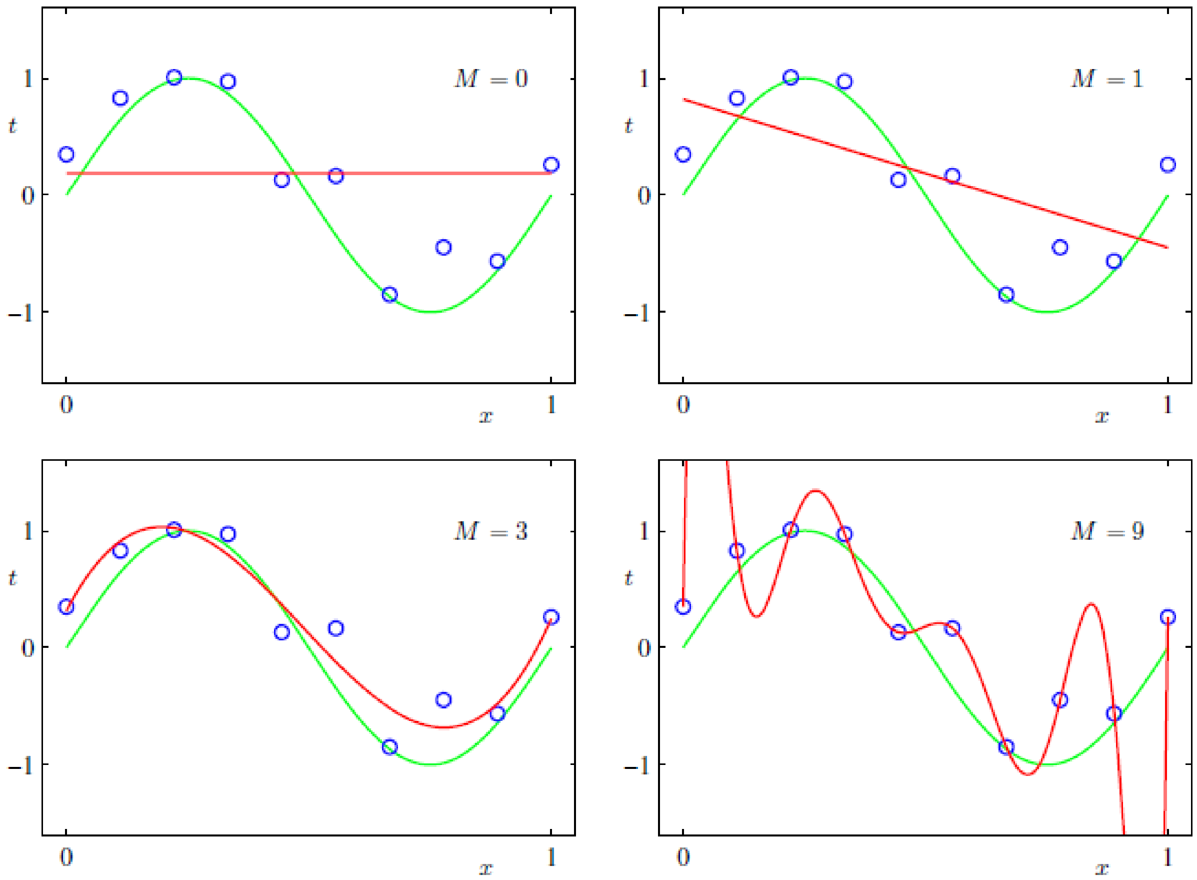


- We further suppose that we fit these data with a polynomial function as follows:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

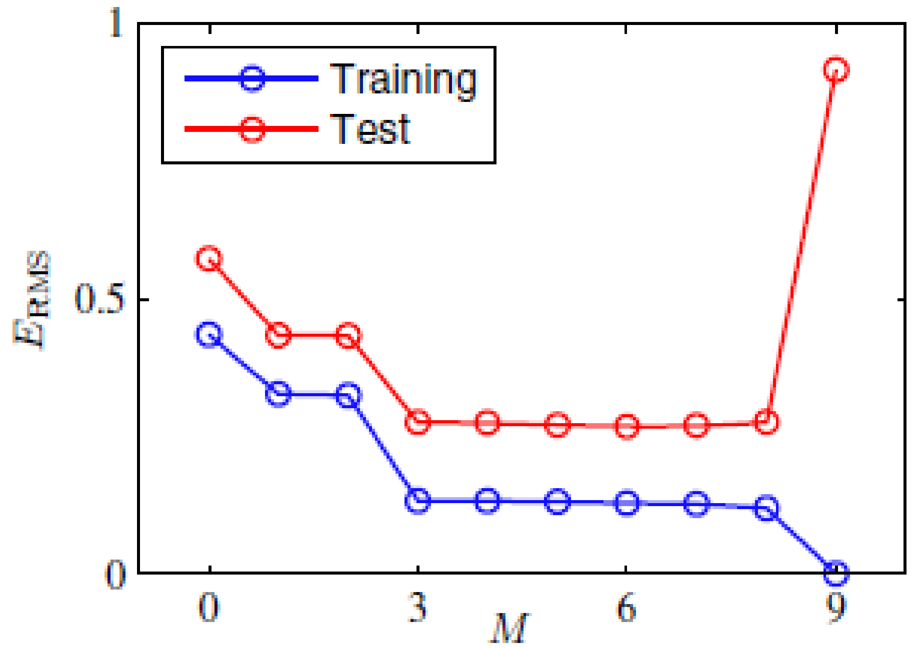
Example of complexity

- Plots of polynomials having various of M, shown as red curves.



Example of complexity

- Graphs of the root-square error, evaluated on the training set and on an independent test set for various values of M



Does Complexity should be limited?

MIMA

- If we make predictions using the “best fitting” parameters of a model, we have to limit the number of parameters to avoid overfitting.
- For this example, the model with M=3 seems good. We might be able to choose a good value for M using the method of “cross validation”, which looks for the value that does best at prediction one part of the data from the rest of the data.
- But we know $\sin(2\pi x)$ is not a polynomial function, it has an infinite series representation with terms of arbitrarily high order.

Does Complexity should be limited?

MIMA

- How can it be good to use a model that we know is false?
  - The Bayesian answer: It is not good. We should abandon the idea of using the best parameters and instead average over all plausible values for the parameters. Then we can use a model (perhaps a very complex one) that is as close to being correct as we can manage.

Reducing Dimensionality

MIMA

- Suppose dimension of input data is 1000, can we replace these with fewer ones, without loss of information.
- On simple way is to use PCA (Principal Component Analysis)
  - Suppose that all data are in a space, we first find the direction of highest variance of these data points, then the direction of second-highest variance that is orthogonal to the first one, so on and so forth…
  - Replace each training sample by the projections of the inputs on some directions.
- Might discard useful info., but keep most of it.

Relationship between ML & DM

- Data-Mining (DM) is the process of extracting patterns from data. The main goal is to understand relationships, validate models or identify unexpected relationships.
- Machine Learning (ML) algorithms allows computer to learn from data. The learning process consist of extracting the patterns but the end goal is to use the knowledge to do prediction on new data.

Differences between ML & DM

MIMA

- They are often interchangeably, but
- Data mining is more often used for problems with very large amounts of data, where computational efficiency is more important than statistical sophistication----often business applications.
- Machine learning is more often used for problems with a flavor of artificial intelligence---such as recognition of objects in visual scenes.

【 Thank You ! 】

Any Question?