Machine learning: What it can do, recent directions and some challenges?

Ho Tu Bao
Japan Advanced Institute of Science and Technology
John von Neumann Institute, VNU-HCM

Content

1. Basis of machine learning
2. Recent directions and some challenges
3. Machine learning in other sciences

Disclaims: This reflects the personal view and most contents are subject of discussion.

About machine learning

How knowledge is created?

Chuồn chuồn bay thấp thì mưa
Bay cao thì nắng bay vừa thì râm

Mùa hè đang nắng, cỏ gà trắng thì mưa.
Cỏ gà mọc lang, cả làng được nước.

Kiến đen tha trứng lên cao
Thế nào cũng có mưa rào rất to

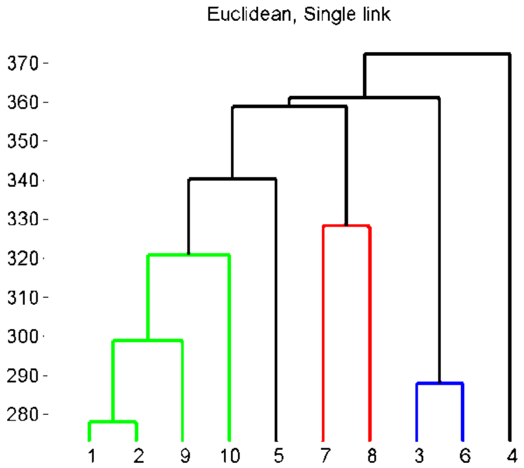
Chuồn chuồn cắn rốn, bốn ngày biết bơi

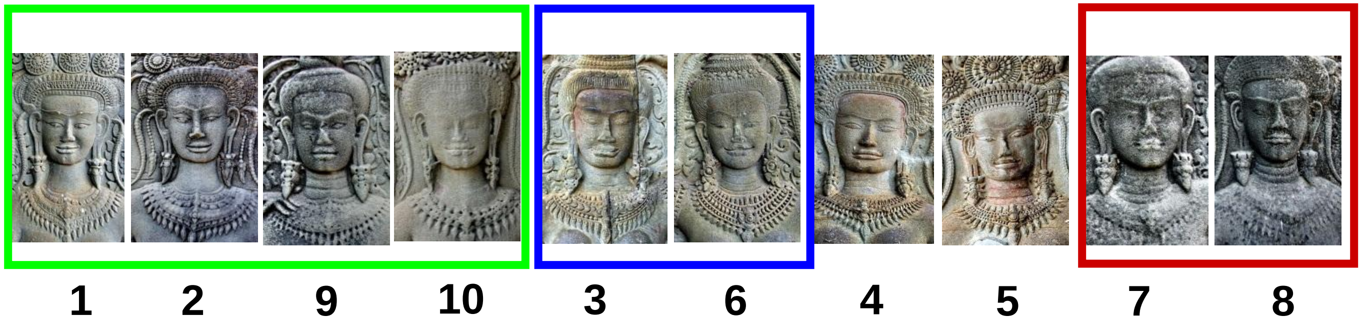
Deduction: $Given\ f(x)\ and\ x_i,\ \ infer\ f(x_i)$
Induction: $Given\ \{x_i\},\ \ infer\ f(x)$

About machine learning

Facial types of Apsaras

- Angkor Wat contains the most unique gallery of ~2,000 women depicted by detailed full body portraits
- What facial types are represented in these portraits?

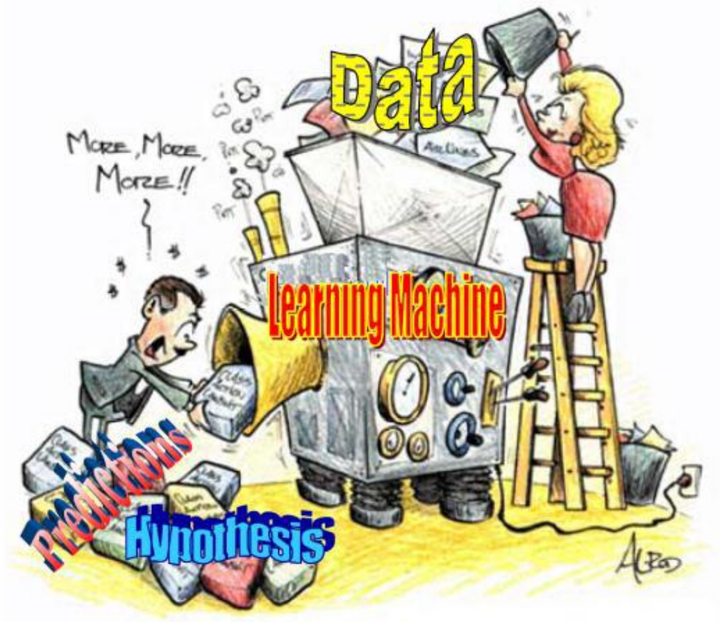




Jain, ECML 2006; Kent Davis, "Biometrics of the Godedess", DatAsia, Aug 2008
S. Marchal, "Costumes et Parures Khmers: D'apres les devata D'Angkor-Vat", 1927

About machine learning
Definition

- Mục đích của học máy là việc xây dựng các hệ máy tính có khả năng thích ứng và học từ kinh nghiệm (Tom Dieterich).
- Một chương trình máy tính được nói là học từ kinh nghiệm **E** cho một lớp các nhiệm vụ **T** với độ đo hiệu suất **P**, nếu hiệu suất của nó với nhiệm vụ **T**, đánh giá bằng **P**, có thể tăng lên cùng kinh nghiệm
(T. Mitchell Machine Learning book)
- Khoa học về việc làm cho máy có khả năng học và tạo ra tri thức từ dữ liệu.



(from Eric Xing lecture notes)

- Three main AI targets: Automatic Reasoning, Language understanding, Learning
- Finding hypothesis f in the hypothesis space F by narrowing the search with constraints (bias)

About machine learning

Improve T with respect to P based on E

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself

T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver.

T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

From Raymond Mooney's talk

About machine learning
Many possible applications

- Disease prediction
- Autonomous driving
- Financial risk analysis
- Speech processing
- Earth disaster prediction
- Knowing your customers
- Drug design
- Information retrieval
- Machine translation
- Water structure
- etc.

Người máy ASIMO đưa đồ uống cho khách theo yêu cầu.

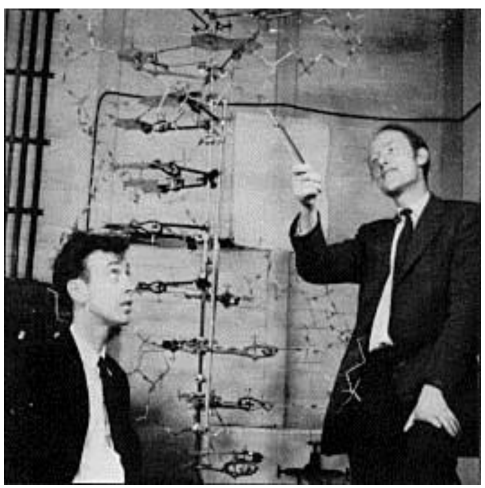
About machine learning

Powerful tool for modeling

Model: Simplified description or abstraction of a reality (mô tả đơn giản hóa hoặc trừu tượng hóa một thực thể).

Modeling: The process of creating models.

Simulation: The imitation of some real thing, state of affairs, or process.

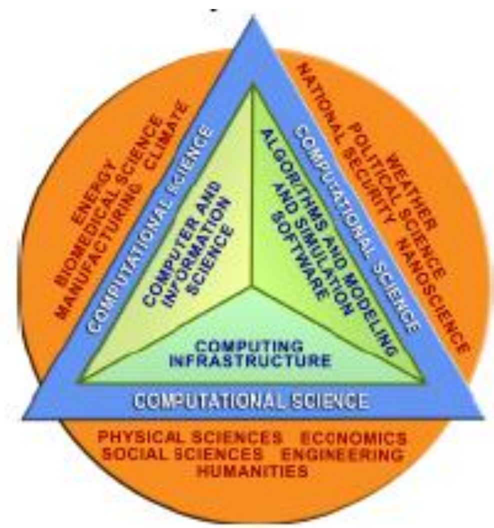


DNA model figured out in 1953 by Watson and Crick

Modeling

Simulation

Data Analysis



Model Selection

Computational science: Using math and computing to solve problems in sciences

About machine learning

Generative model vs. discriminative model

Generative model

- Mô hình xác suất liên quan tất cả các biến, cho việc sinh ra ngẫu nhiên dữ liệu quan sát, đặc biệt khi có các biến ẩn.
- Định ra một phân bố xác suất liên kết trên các quan sát và các dãy nhãn.
- Dùng để
  - Mô hình dữ liệu trực tiếp
  - Bước trung gian để tạo ra một hàm mật độ xác suất có điều kiện.

Discriminative model

- Mô hình chỉ cho các biến mục tiêu phụ thuộc có điều kiện vào các biến được quan sát được.
- Chỉ cho phép lấy mẫu (sampling) các biến mục tiêu, phụ thuộc có điều kiện vào các đại lượng quan sát được.
- Nói chung không cho phép diễn tả các quan hệ phức tạp giữa các biến quan sát được và biến mục tiêu, và không áp dụng được trong học không giám sát.

About machine learning

Generative vs. discriminative methods

Training classifiers involves estimating **f: X → Y**, or **P(Y|X).**
Examples: P(apple | red ∧ round), P(noun | "cá")

Generative classifiers	**Discriminative classifiers**
❑ Assume some functional form for P(X|Y), P(Y) ❑ Estimate parameters of **P(X|Y), P(Y)** directly from training data, and use Bayes rule to calculate **P(Y|X = x_i)** ❑ HMM, Markov random fields, Gaussian mixture models, Naïve Bayes, LDA, etc.	❑ Assume some functional form for P(Y|X) ❑ Estimate parameters of **P(Y|X)** directly from training data ❑ SVM, logistic regression, traditional neural networks, nearest neighbors, boosting, MEMM, conditional random fields, etc.

(cá: fish, to bet)

About machine learning

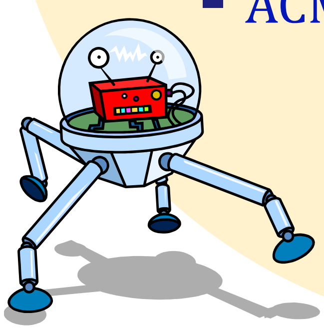
Machine learning and data mining

Machine learning

- To build computer systems that learn as well as human does.
- ICML since 1982 (23th ICML in 2006), ECML since 1989.
- ECML/PKDD since 2001.
- ACML starts Nov. 2009.

Data mining

- To find new and useful knowledge from large datasets .
- ACM SIGKDD since 1995, PKDD and PAKDD since 1997 IEEE ICDM and SIAM DM since 2000, etc.



About machine learning

Some quotes

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Chairman, Microsoft)
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing" (John Hennessy, President, Stanford)
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo)
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, CTO, Sun)
- "Machine learning is today's discontinuity" (Jerry Yang, CEO, Yahoo)

Pedro Domingos' ML slides

About machine learning

Two main views: data and learning tasks

Types and size of data

- Flat data tables
- Relational databases
- Temporal & spatial data
- Transactional databases
- Multimedia data
- Materials science data
- Biological data
- Textual data
- Web data
- etc.

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}

Learning tasks & methods

- Supervised learning
  - Decision trees
  - Neural networks
  - Rule induction
  - Support vector machines
  - etc.
- Unsupervised learning
  - Clustering
  - Modeling and density estimation
  - etc.
- Reinforcement learning
  - Q-learning
  - Adaptive dynamic programming
  - etc.

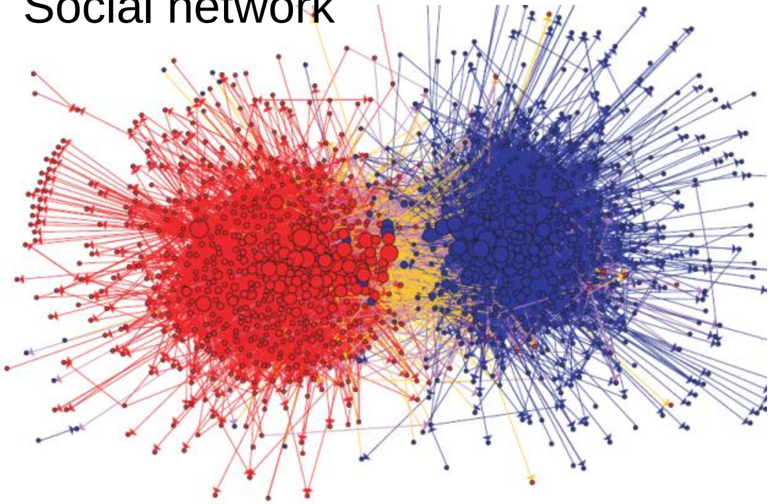
About machine learning

Complexly structured data

A portion of the DNA sequence with
length of 1,6 million characters

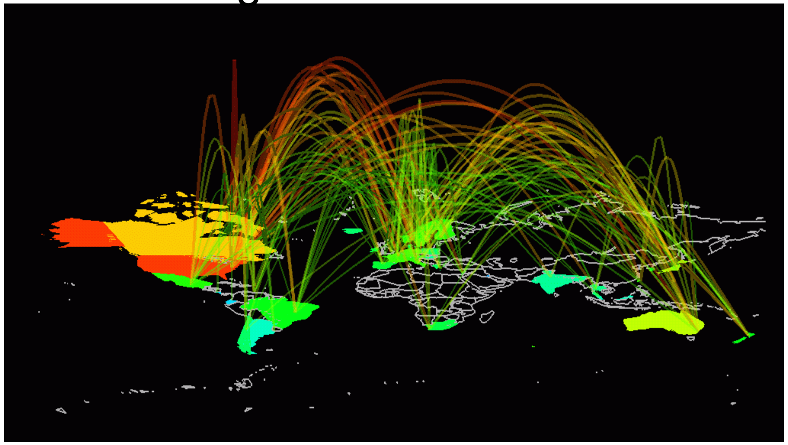
…TACATTAGTTATTACATTGAGAAACTTTATAATTAAA
AAAGATTCATGTAAATTTCTTATTTGTTTATTTAGAGG
TTTTAAATTTAATTTCTAAGGGTTTGCTGGTTTCATT
GTTAGAATATTTAACTTAATCAAATTATTTGAATTAAAT
TAGGATTAATTAGGTAAGCTAACAAATAAGTTAAATTT
TTAAATTTAAGGAGATAAAAATACTACTCTGTTTTATTA
TGGAAAGAAAGATTTAAATACTAAAGGGTTTATATATA
TGAAGTAGTTACCCTTAGAAAAATATGGTATAGAAAGC
TTAAATATTAAGAGTGATGAAGTATATTATGT…

Social network



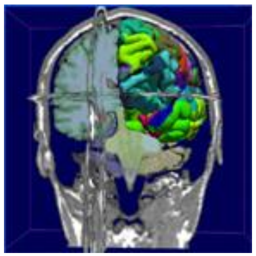
Immense text

Web linkage



About machine learning

Huge volume and high dimensionality



1 human brain at the micron level = 1 PetaByte

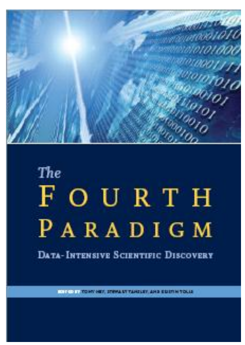
Large Hadron Collider, (PetaBytes/day)

Human Genomics = 7000 PetaBytes 1GB / person

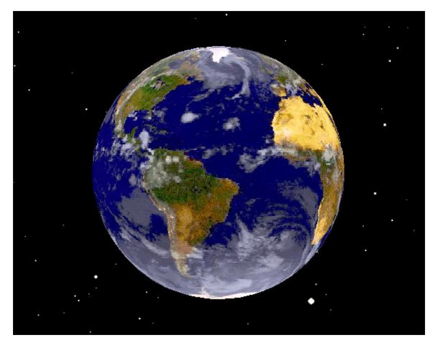
Printed materials in the Library of Congress = 10 TeraBytes

200 of London's Traffic Cams (8TB/day)

1 book = 1 MegaByte



Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}

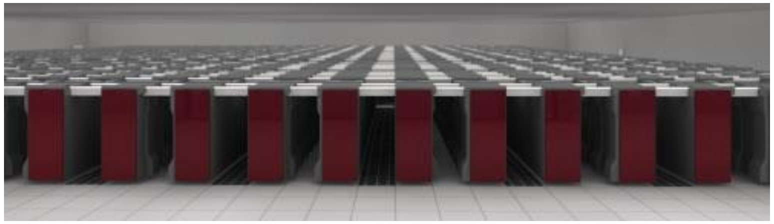


All worldwide information in one year = 2 ExaBytes

Family photo = 586 KiloBytes

Adapted from Berman, San Diego Supercomputer Center (SDSC)

About machine learning
New generation of supercomputers



Japan's K computer

- China's supercomputers Tianhe-1A: 7,168 NVIDIA® Tesla™ M2050 GPUs and 14,336 CPUs, 2,507 peta flops, 2010.
- Japan's "K computer" 800 computer racks ultrafast CPUs, 10 peta flop (2012, RIKEN's Advanced Institute for Computational Science)
- IBM's computers BlueGene and BlueWaters, 20 peta flop (2012, Lawrence Livermore National Laboratory).

IBM BlueGene

http://www.fujitsu.com/global/news/pr/archives/month/2010/20100928-01.html (28.9.2010)
http://www.hightechnewstoday.com/nov-2010-high-tech-news/38-nov-23-2010-high-tech-news.shtml (23 Nov. 2010)

Content

1. Basis of machine learning
2. Recent directions and some challenges
3. Machine learning in other sciences

Development of machine learning

Successful applications

Symbolic concept induction

IR & ranking

Data mining

Multi strategy learning

MIML

Active & online learning

Minsky criticism

Transfer learning

NN, GA, EBL, CBL

Kernel methods

Sparse learning

Abduction, Analogy

Pattern Recognition emerged

Bayesian methods

Revival of non-symbolic learning

PAC learning

ILP

Semi-supervised learning

Deep learning

Dimensionality reduction

Experimental comparisons

Math discovery AM

Probabilistic graphical models

Supervised learning

Statistical learning

Neural modeling

Nonparametric Bayesian

Unsupervised learning

Ensemble methods

Rote learning

Reinforcement learning

Structured prediction

1950 1960 1970 1980 1990 2000 2010

ICML (1982) ECML (1989) KDD (1995) PAKDD (1997) ACML (2009)

enthusiasm | dark age | renaissance | maturity | fast development

From 900 submissions to ICML 2012

66 Reinforcement Learning
52 Supervised Learning
51 Clustering
46 Kernel Methods
40 Optimization Algorithms
39 Feature Selection and Dimensionality Reduction
33 Learning Theory
33 Graphical Models
33 Applications
29 Probabilistic Models
29 NN & Deep Learning
26 Transfer and Multi-Task Learning
25 Online Learning
25 Active Learning
22 Semi-Supervised Learning
20 Statistical Methods
20 Sparsity and Compressed Sensing
19 Ensemble Methods
18 Structured Output Prediction
18 Recommendation and Matrix Factorization
18 Latent-Variable Models and Topic Models
17 Graph-Based Learning Methods
16 Nonparametric Bayesian Inference
15 Unsupervised Learning and Outlier Detection

ne learning

Successful applications

induction

IR & ranking

Data mining

egy learning

MIML

Active & online learning

Transfer learning

L, CBL

Kernel methods

Sparse learning

alogy

Bayesian methods

Revival of non-symbolic learning

Semi-supervised learning

Deep learning

tal comparisons

Dimensionality reduction

Probabilistic graphical models

ning

Statistical learning

earning

Nonparametric Bayesian

Ensemble methods

nt learning

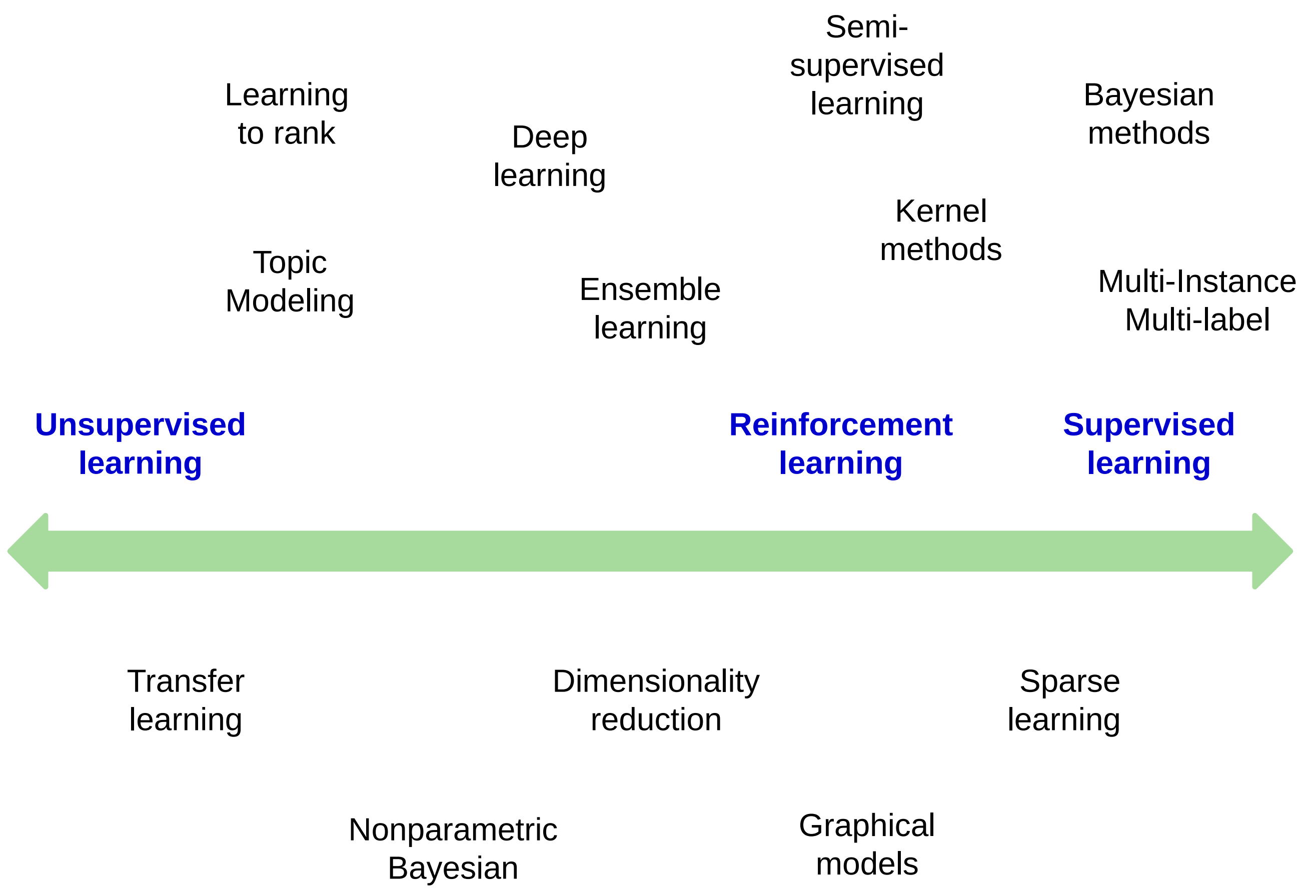
Structured prediction

1990 **2000** **2010**

ECML (1989) KDD (1995) PAKDD (1997) ACML (2009)

ce maturity fast development

Relations among recent directions

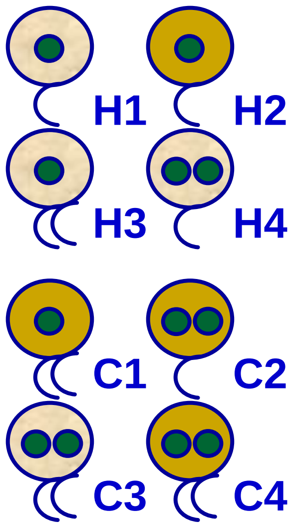


Supervised vs. unsupervised learning

Given: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- x_i is description of an object, phenomenon, etc.
- y_i is some property of x_i, if not available learning is unsupervised

Find: a function $f(x)$ that characterizes $\{x_i\}$ or that $f(x_i) = y_i$



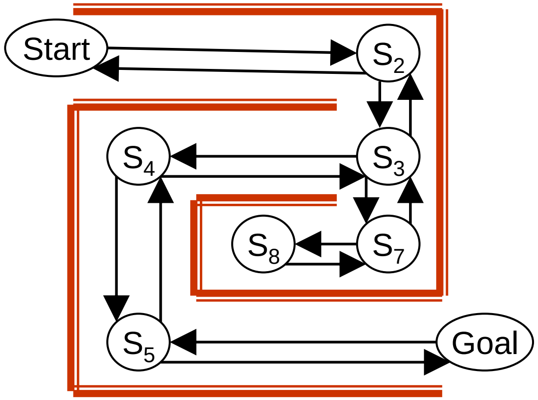
Unsupervised data

	color	#nuclei	#tails
H1	light	1	1
H2	dark	1	1
H3	light	1	2
H4	light	2	1
C1	dark	1	2
C2	dark	2	1
C3	light	2	2
C4	dark	2	2

Supervised data

	color	#nuclei	#tails	class
H1	light	1	1	healthy
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

Reinforcement learning



Concerned with how an agent ought to take actions in an environment so as to maximize some cumulative reward. (… một tác nhân phải thực hiện các hành động trong một môi trường sao cho đạt được cực đại các phần thưởng tích lũy)

- The basic reinforcement learning model consists of:
  - a set of environment states S;
  - a set of actions A;
  - rules of transitioning between states;
  - rules that determine the *scalar immediate reward* of a transition;
  - rules that describe what the agent observes.

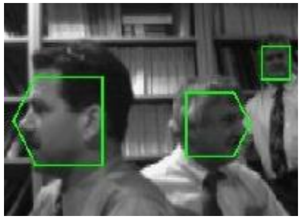
Active learning and online learning

Online active learning

Active learning

A type of supervised learning, samples and selects instances whose labels would prove to be most informative additions to the training set. (… lấy mẫu và chọn phần tử có nhãn với nhiều thông tin cho tập huấn luyện)

- Labeling the training data is not only time-consuming sometimes but also very expensive.
- Learning algorithms can actively query the user/teacher for labels.



Online learning

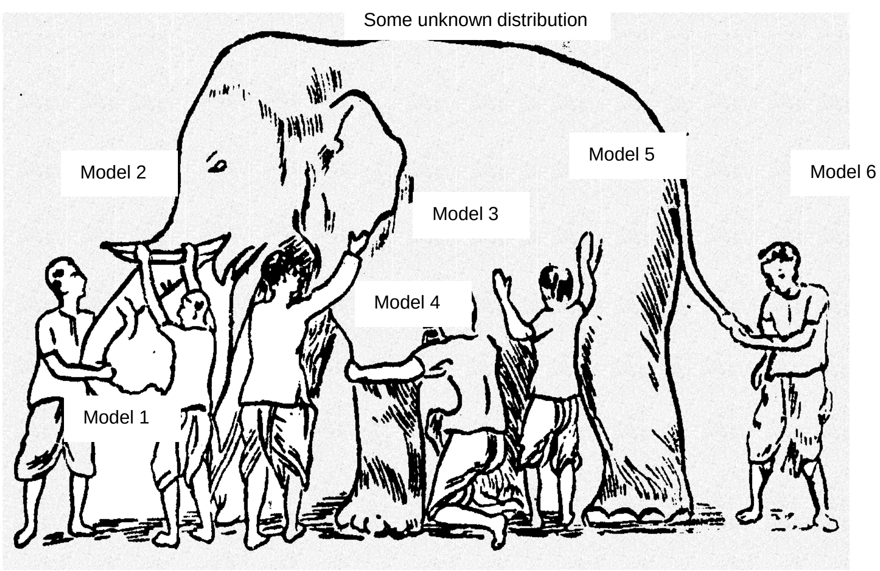
Learns one instance at a time with the goal of predicting labels for instances. (ở mỗi thời điểm chỉ học một phần tử nhằm đoán nhãn các phần tử).

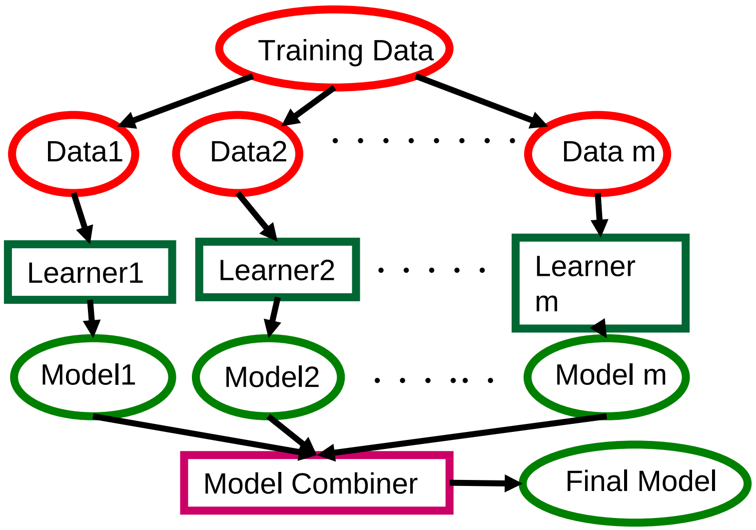
- Instances could describe the current conditions of the stock market, and an online algorithm predicts tomorrow's value of a particular stock.
- Key characteristic is after prediction, the true value of the stock is known and can be used to refine the method.

Ensemble learning

Ensemble methods employ multiple learners and combine their predictions to achieve higher performance than that of a single learner. (… dùng nhiều bộ học để đạt kết quả tốt hơn việc dùng một bộ học)

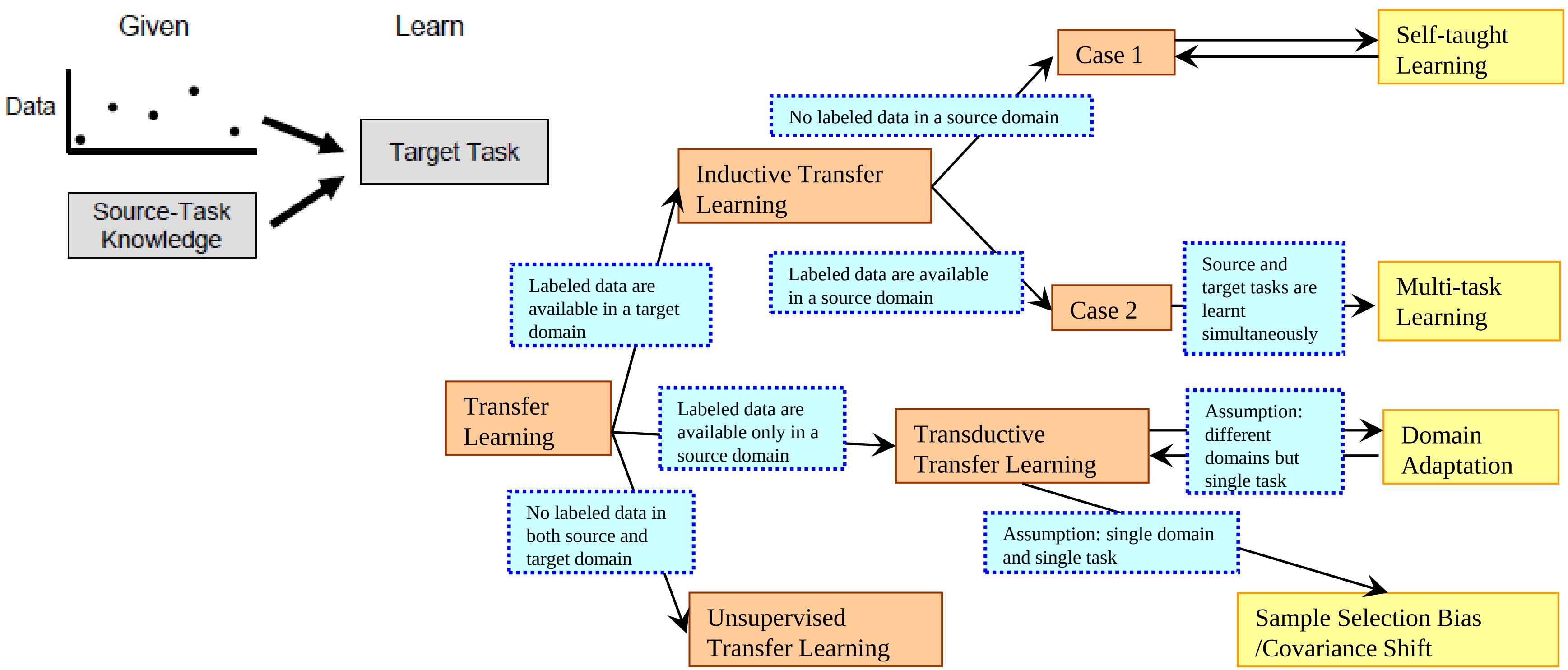
- **Boosting**: Make examples currently misclassified more important
- **Bagging**: Use different subsets of the training data for each model





Transfer learning

Aims to develop methods to transfer knowledge learned in one or more source tasks and use it to improve learning in a related target task. (truyền tri thức đã học được từ nhiều nhiệm vụ khác để học tốt hơn việc đang cần học)



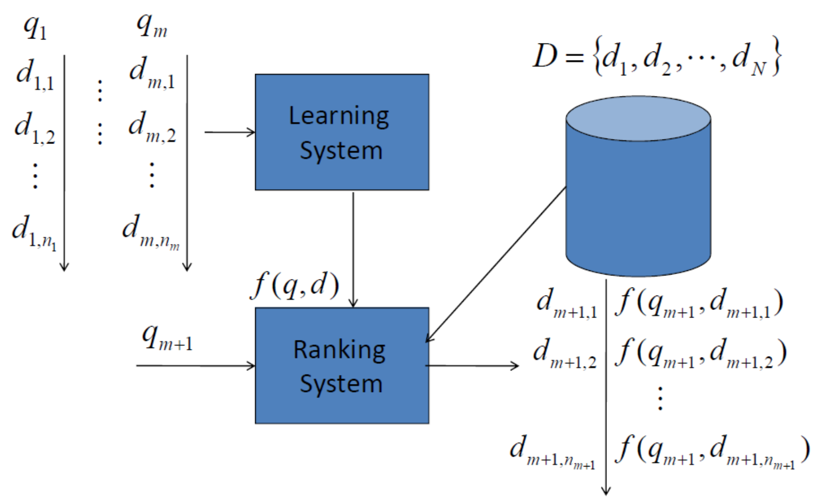
Induction: $Given\ \{x_i\},\ \ infer\ f(x)$
Transduction: $Given\ \{x_k\},\ \ infer\ x_j\ from\ x_i$

Learning to rank

The goal is to automatically rank matching documents according to their relevance to a given search query from training data. (học từ dữ liệu huấn luyện để tự động xếp thứ tự các tài liệu tìm được liên quan tới một câu hỏi cho trước).

- Pointwise approach:
  Transform ranking to regression or classification (score)
- Pairwise approach:
  Transform ranking to pairwise classification (which is better)
- Listwise approach:
  Directly optimize the value of each of the above evaluation measures, averaged over all queries in the training data.

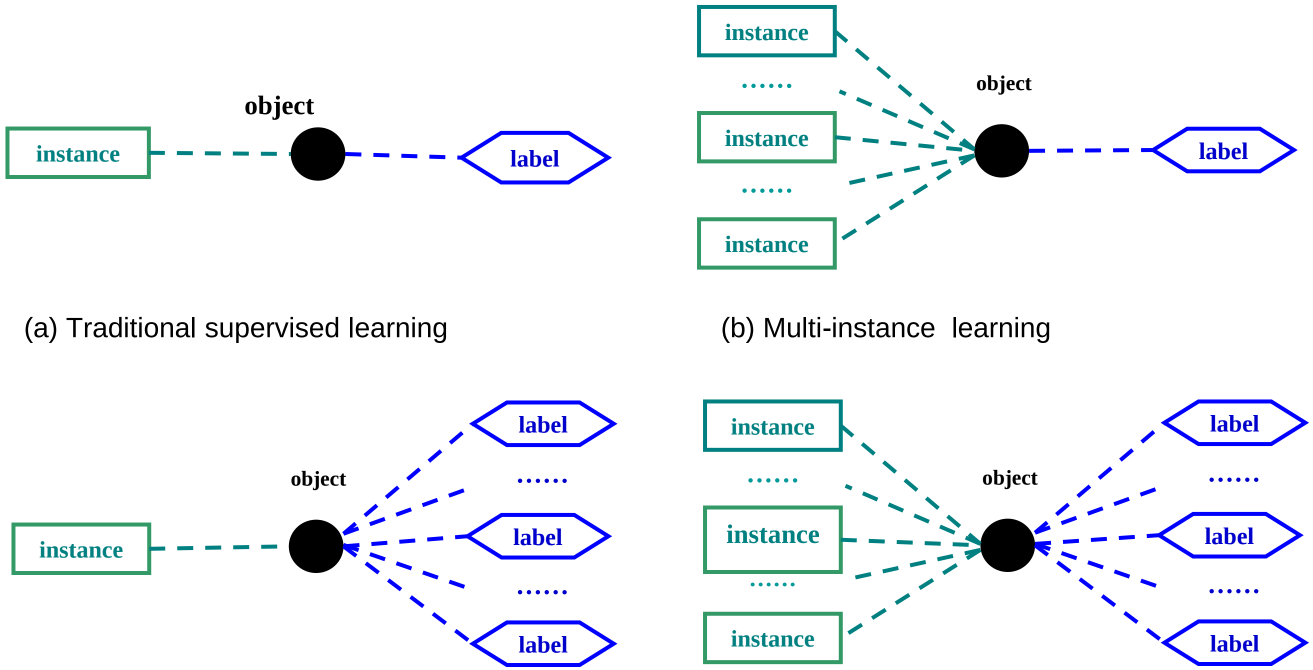
Example	DocID	Query	s_T	s_B	Judgment
Φ_1	37	linux	1	1	Relevant
Φ_2	37	penguin	0	1	Non-relevant
Φ_3	238	system	0	1	Relevant
Φ_4	238	penguin	0	0	Non-relevant
Φ_5	1741	kernel	1	1	Relevant
Φ_6	2094	driver	0	1	Relevant
Φ_7	3191	driver	1	0	Non-relevant



Example from Stanford lectures

Multi-instance multi-label learning

MIML is the framework where an example is described by multiple instances and associated with multiple class labels. (một lược đồ bài toán khi mỗi đối tượng được mô tả bằng nhiều thể hiện và thuộc về nhiều lớp).



Tom Dieterich et al., 1997

Zhi-Hua Zhou et al., 2008

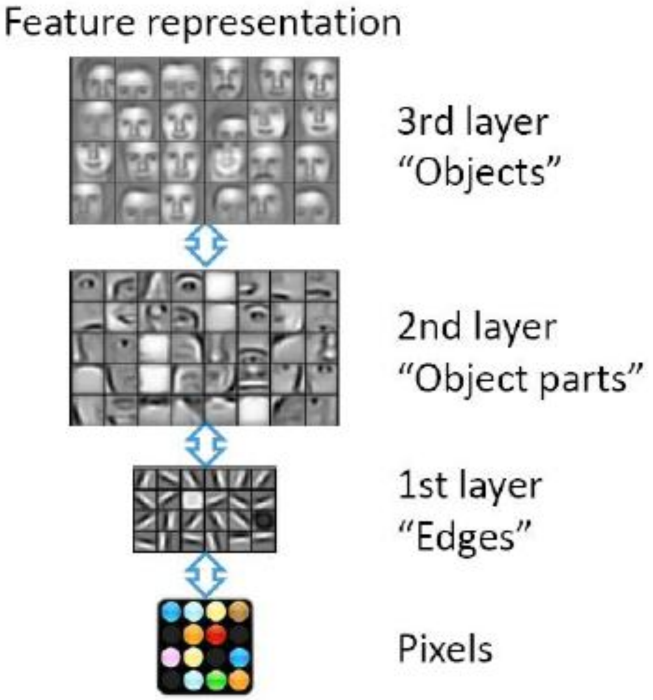
(a) Traditional supervised learning (b) Multi-instance learning

(c) Multi-label learning (d) Multi-instance multi-label l earning

Deep learning

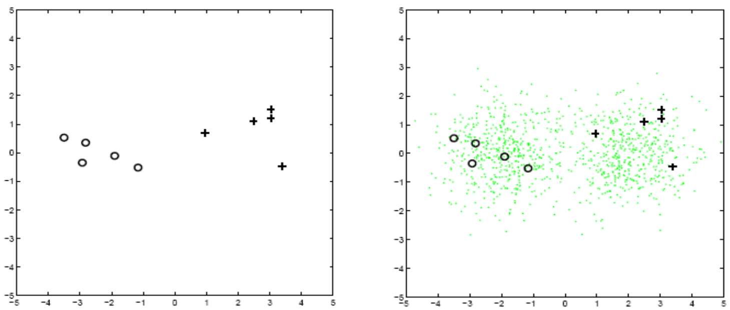
A subfield of machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. (học nhiều cấp độ biểu diễn để mô hình các quan hệ phức tạp trong dữ liệu)

- Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture.
- Key: Deep architecture, deep representation, multi levels of latent variables, etc.



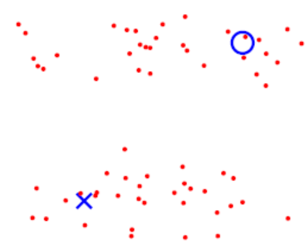
Semi-supervised learning

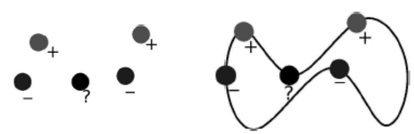
A class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. (dùng cả dữ liệu có nhãn và không nhãn để huấn luyện, tiêu biểu khi ít dữ liệu có nhãn nhưng nhiều dữ liệu không nhãn)

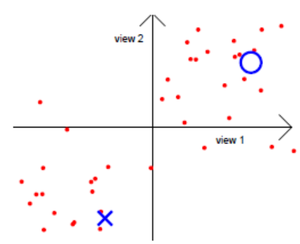


Classes of SSL methods

- Generative models
- Low-density separation
- Graph-based methods
- Change of representation







Assumption	Approach
Cluster Assumption	Low Density Separation, eg, S3VMs
Manifold assumption	Graph-based methods (nearest neighbor graphs)
Independent views	Co-training

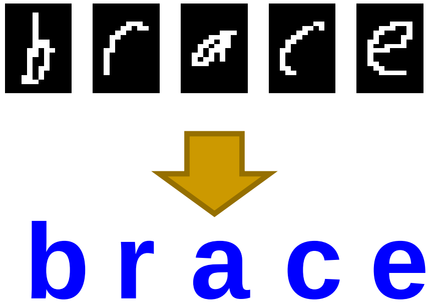
Challenges in semi-supervised learning

- Real SSL tasks: Which tasks can be dramatically improved by SSL?
- New SSL assumptions? E.g., assumptions on unlabeled data: label dissimilarity, order preference
- Efficiency on huge unlabeled datasets
- Safe SSL:
  - no pain, no gain
  - no model assumption, no gain
  - wrong model assumption, no gain, a lot of pain

  → develop SSL techniques that do not make assumptions beyond those implicitly or explicitly made by the classification scheme employed?

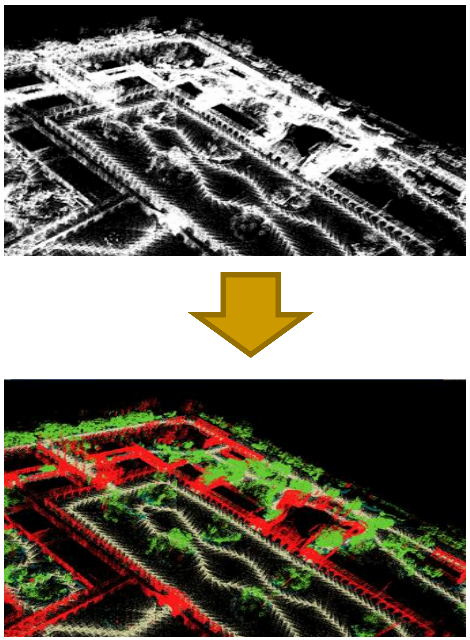
Xiaojin Zhu tutorial

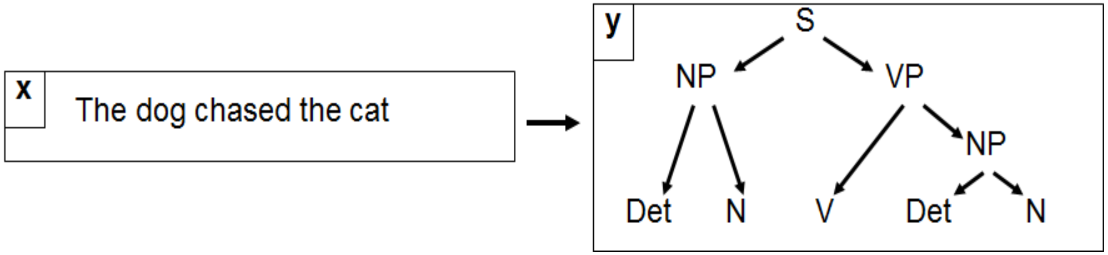
Structured prediction



An umbrella term for machine learning and regression techniques that involve predicting structured objects. (liên quan việc đoán nhận các đối tượng có cấu trúc).

- Examples
  - Multi-class labeling
  - Protein structure prediction
  - Noun phrase co-reference clustering
  - Learning parameters of graphical models

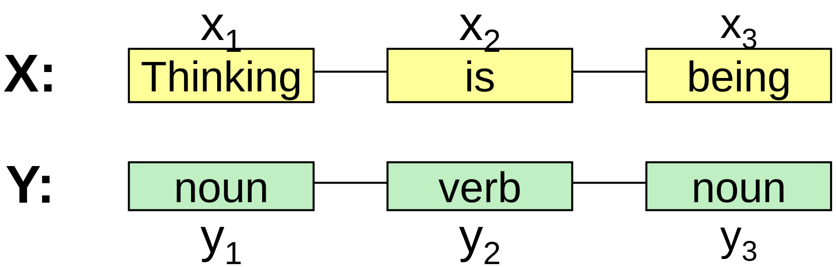


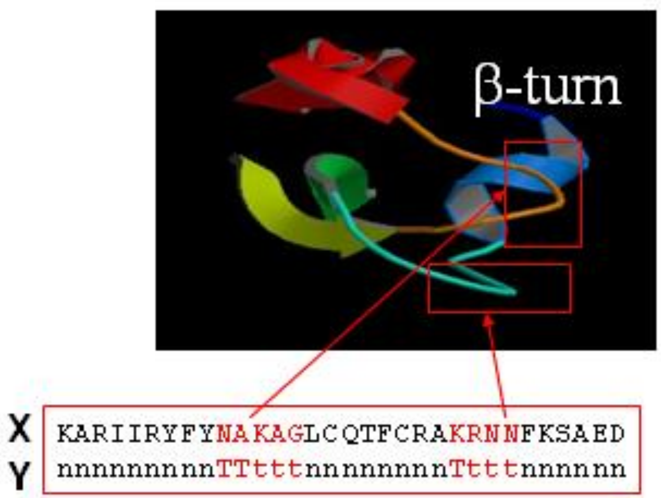


Structured prediction

Example: Labeling sequence data problem

- X is a random variable over data sequences
- Y is a random variable over label sequences whose labels are assumed to range over a finite label alphabet A
- Problem: Learn how to give labels from a closed set Y to a data sequence X





- POS tagging, phrase types, etc. (NLP),
- Named entity recognition (IE)
- Modeling protein sequences (CB)
- Image segmentation, object recognition (PR)
- Recognition of words from continuous acoustic signals.

Pham, T.H., Satou, K., Ho, T.B. (2005). Support vector machines for prediction and analysis of beta and gamma turns in proteins, Journal of Bioinformatics and Computational Biology (JBCB), Vol. 3, No. 2, 343-358

Le, N.T., Ho, T.B., Ho, B.H. (2010). Sequence-dependent histone variant positioning signatures, *BMC Genomics*, Vol. 11 (S4)

Structured prediction

Some challenges

- Given $\{(x_i, y_i\}_{i=1}^{n}$ drawn from an unknown joint probability distribution P on $X \times Y$, we develop an algorithm to generate a scoring function $F: X \times Y \to \mathcal{R}$ which measures how good a label y is for a given input x.

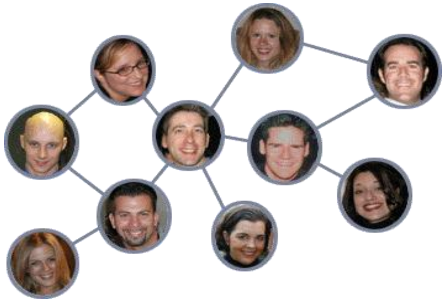
- Given $\hat{x}$, predict the label $\hat{y} = \underset{y \in Y}{\operatorname{argmax}} F(\hat{x}, y)$. F is generally considered are linearized models, thus $F(x, y) = \langle w^*, \phi(x, y) \rangle$, e.g, in POS tagging,

$$\phi(x, y) = \begin{cases} 1 \text{ if suffix}(x_i) = \text{"ing" and } y_i = VBG \\ 0 \text{ otherwise} \end{cases}$$

- A major concern for the implementation of most structured prediction algorithms is *the issue of tractability*. If each y_i can take k possible values i.e. |Yi| = k, the total number of possible labels for a sequence of length L is k^L. Find optimal y is intractable.

VBG = Verb, Auxiliary be, present part

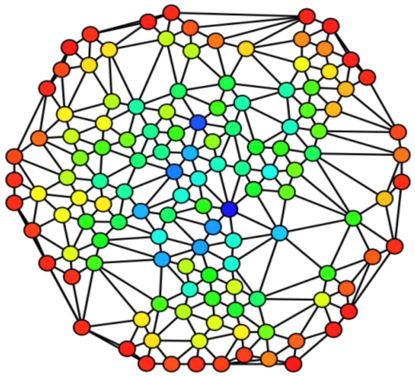
Social network analysis



Picture from Matthew Pirretti's slides

Social media describes the online tools that people use to share content, profiles, opinions, insights, experiences, perspectives and media itself, thus facilitating conversations and interaction online between people. These tools include blogs, microblogs, facebook, bookmarks, networks, communities, wikis, etc.

- Social networks: Platforms providing rich interaction mechanisms, such as Facebook or MySpace, that allow people to collaborate in a manner and scale which was previously impossible (interdisciplinary study).
- Social network study: structure analysis, understanding social phenomenon, information propagation & diffusion, prediction (information, social), general dynamics, modeling (social, business, algorithmic, etc.)



Hue (from red=0 to blue=max) indicates each node's betweenness centrality.

Figures from wikipedia

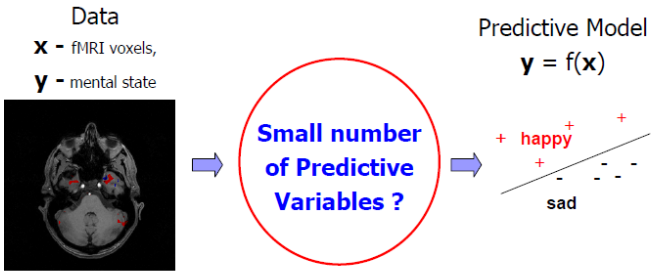
Social network analysis

Some challenges

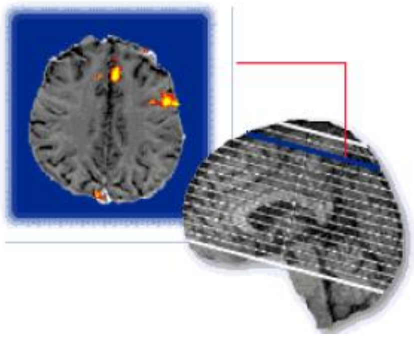
- Structural analysis: Focus on relations and patterns of relations requires methods/concepts different from traditional statistic and data analysis (e.g., graphical model, dependencies?)
- Centrality and prominence: Key issue in social network analysis is the identification of the most important or prominent actors (nodes). Many notions: degree, closeness, betweeness, rank of the actors.
- Influence: The capacity or power of persons or things to be a compelling force on or produce effects on the actions, behaviour, opinions, etc., of others (e.g., author topic models, twiter mining, etc.)
- Knowledge challenge: Enabling users to share knowledge with their community (e.g., cope with spam, privacy and security).
- Collaborative production (e.g., Wikipedia and Free Software): collaborative content creation, decentralized decision making, etc.

Sparse modeling

Selection (and, moreover, construction) of a small set of highly predictive variables in high-dimensional datasets. (chọn và tạo ra một tập nhỏ các biến có khả năng dự đoán cao từ dữ liệu nhiều chiều).



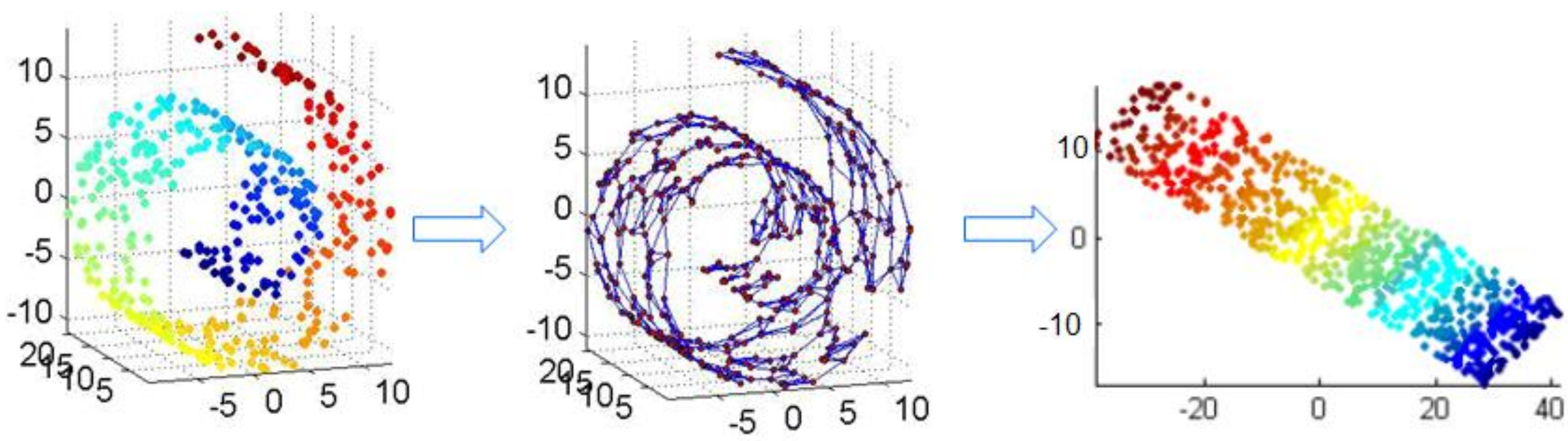
- Rapidly developing area on the intersection of statistics, machine learning and signal processing.
- Typically when data are of high-dimensional, small-sample
  - 10,000-100,000 variables (voxels)
  - 100s of samples(time points)
- Sparse SVMs, sparse Gaussian processes, sparse Bayesian methods, sparse regression, sparse Q-learning, sparse topic models, etc.



Find small number of most relevant voxels (brain areas)?

Dimensionality reduction

The process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. (quá trình rút gọn số biến ngẫu nhiên đang quan tâm, gồm lựa chọn biến và tạo biến mới).

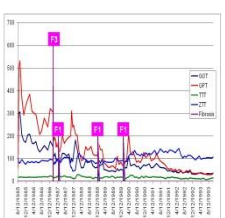
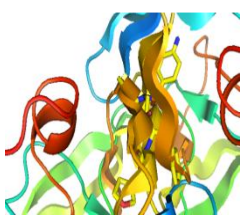
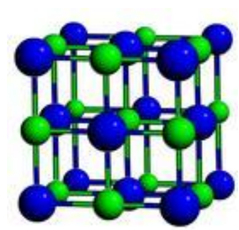


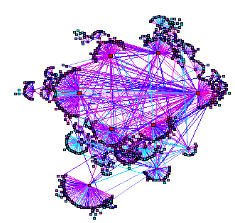
Data Points (Swiss roll) 4-NN Graph 2D Embedding Result

Kernel methods

Learning from non-vectorial data

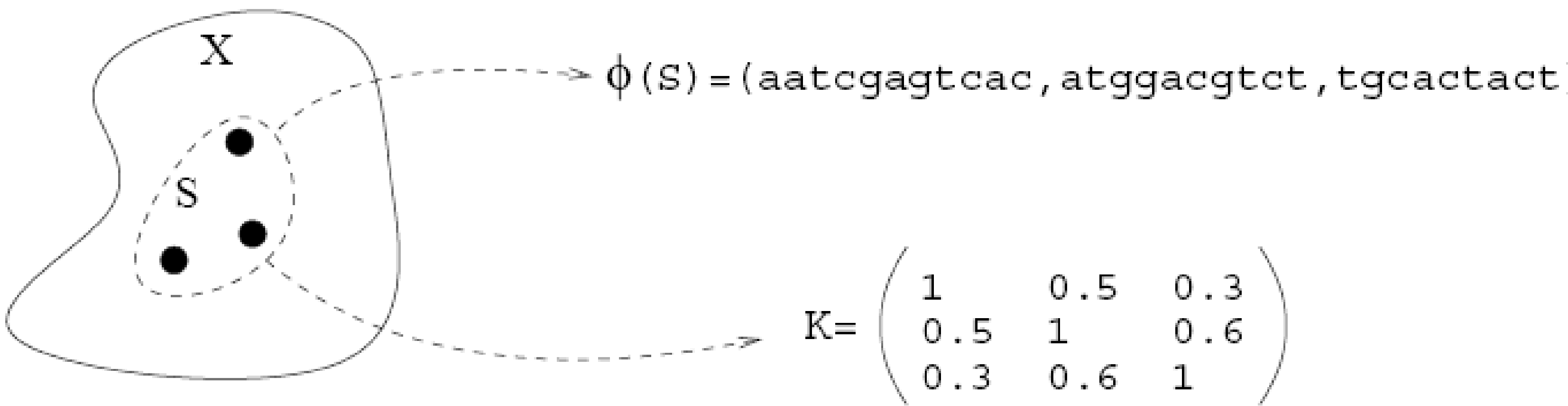
- Current
  - Most learning algorithms work on flat, fixed length feature vectors
  - Each new data type requires a new learning algorithm
  - Difficult to handle strings, gene/protein sequences, natural language parse trees, graph structures, pictures, plots, …
- Key Challenges
  - One data-interface for multiple learning methods
  - One learning method for multiple data types



Kernel methods

Data representations

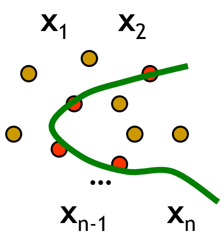


- $\mathcal{X}$ is the set of all oligonucleotides, $\mathbb{S}$ consists of three oligonucleoides.
- Traditionally, each oligonucleotide is represented by a sequence of letters.
- In kernel methods, $\mathbb{S}$ is represented as a matrix of pairwise similarity between its elements.

Kernel methods

The basic ideas

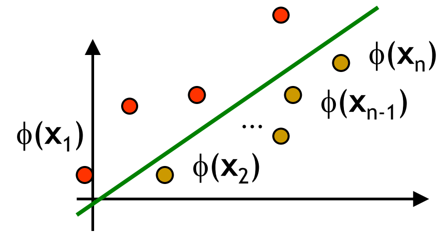
Input space X



inverse map ϕ^{-1}

$\phi(\mathbf{x})$

Feature space F

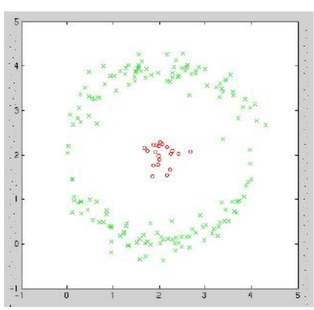


$$\boxed{k(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)}$$

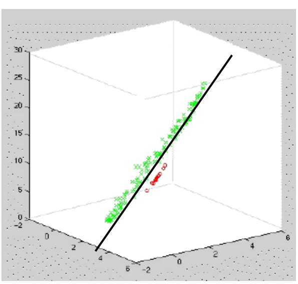
kernel function k: $\mathcal{X}\text{x}\mathcal{X} \rightarrow \mathcal{R}$

Kernel matrix K_{nxn}

kernel-based algorithm on K
(computation done on kernel matrix)



$$\phi : \mathcal{X} = \mathcal{R}^2 \rightarrow \mathcal{H} = \mathcal{R}^3$$
$$(x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2)$$



Các phương pháp dựa trên biến đổi dữ liệu bằng các hàm kernel sang một không gian mới nhiều chiều hơn nhưng ở đó có thể dùng các phương pháp tuyến tính.

Kernel methods

Some challenges

- The choice of kernel function. In general, there is no way of choosing or constructing a kernel that is optimal for a given problem.
- The complexity of kernel algorithms. Kernel methods access the feature space via the input samples and need to store all the relevant input samples.

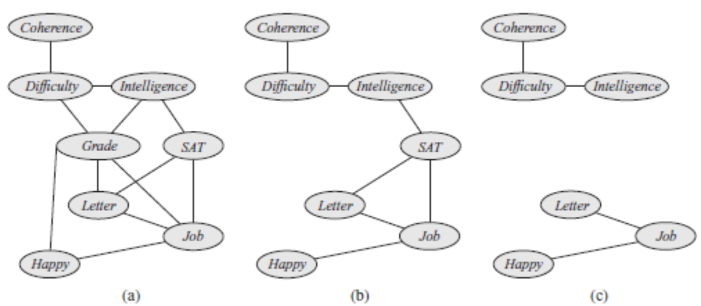
  Examples: Store all support vectors or size of the kernel matrices grows quadratically with sample size → scalability of kernel methods.
- Incorporating priors knowledge and invariances in to kernel functions are some of the challenges in kernel methods.
- L1 regularization may allow some coefficients to be zore → hot topic
- Multiple kernel learning (MKL) is initially (2004, Lanckriet) of high computational cost → Many subsequent work, still ongoing, has not been a practical tool yet.

Probabilistic graphical models

Also called graphical model and is a way of describing/representing a reality by probabilistic relationships between random variables (observed and unobserved ones). (Cách môt tả và biểu diễn các hệ thống phức tạp bằng các quan hệ xác suất giữa các biến ngẫu nhiên (biến hiện và ẩn).

Marriage of graph theory and probability theory in a powerful formalism for multivariate statistical modeling.

- Directed graphical models (Bayesian networks) and undirected graphical models (Markov networks). Fundamental: *modularity* (a complex system = combining simpler parts).
- A general framework of:
  - *Bayesian networks*: HMM, NB, Kalman filters, mixture model…
  - *Markov networks*: CRF, MaxEnt, LDA, Hopfield net, Markov chain...



Initial set of factors   Reduced to context   Reduced to context

Probabilistic graphical models

The main issues

- Representation: How a graphical model models a reality? Which forms?
  - Graph describing realities by nodes representing variables and arcs their relations: directed and undirected graphical models
- Learning: How we build graphical models?
  - The *structure* and *parameters* of each conditional probabilistic dependency (known or unknown structure fully or partially observability)
- Inference: How can we use observed variables on these models to computer the posterior distributions of subsets of other variables?
  - Variable elimination, dynamic programming, approximation, inference in dynamic Bayesian networks.
- Applications: How to use graphical models to model some reality, to learn it from observed data and to infer on it to answer the questions?

Daphne Koller & Nir Friedman, Probabilistic Graphical Models, Principles and Techniques, MIT Press, 2009

Probabilistic graphical models

Graph theory and Probability theory

- A directed graphical model consists of a collection of prob. distributions that factorize as (pa_k = set of parent nodes of x_k):

$$p(x_1, \dots, x_m) = \prod_{k=1..m} p\,(x_k|\mathrm{pa}_k)$$

- A undirected graphical model consists of a collection of probability distributions that factorize as

$$p(x_1, \dots, x_m) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C\,(x_C)$$

$\mathcal{C}$ = {maximal cliques of graph}, ψ_C is the compatibility function.

- Characterize prob. distributions as conditional independencies among subsets of random variables.

- For undirected graphical models, conditional independence is identified with *reachability* notion.

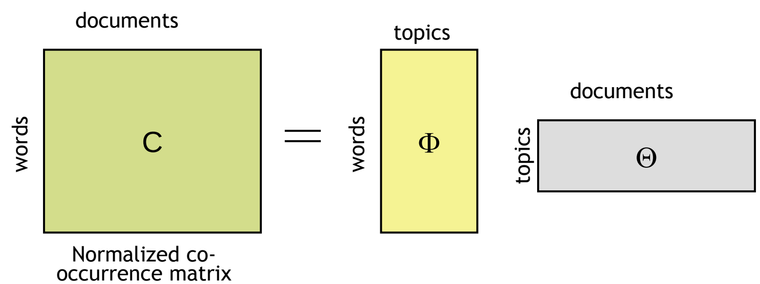
- *A, B, C* = disjoint subsets of vertices.

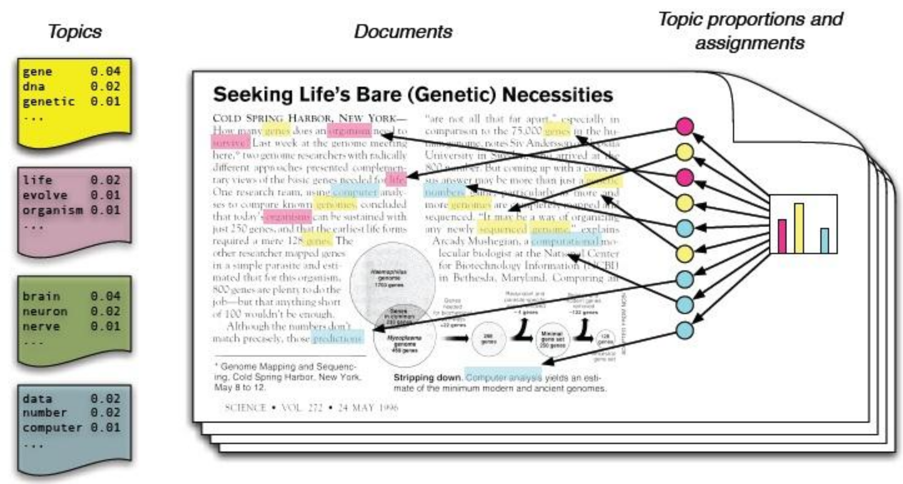
  Say X_A is independent of X_B given X_C if there is no path from a vertex in A to a vertex in B when we remove the vertices C from the graph.

- Consider all A, B, C → all cond. independence assertions.

Probabilistic graphical models

Topic models: Roadmap to text meaning



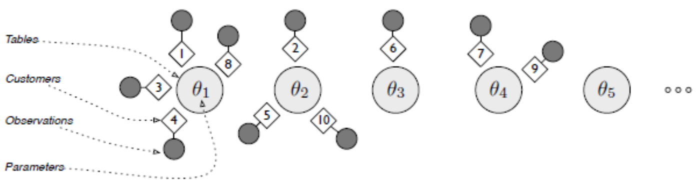


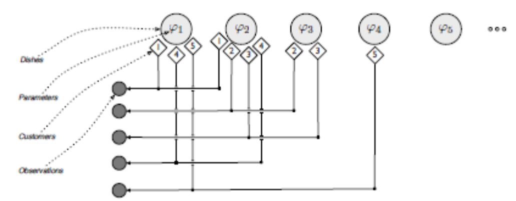
- **Key idea**: *documents are mixtures of latent topics, where a topic is a probability distribution over words.*
- Hidden variables, generative processes, and statistical inference are the foundation of probabilistic modeling of topics.

Blei, D., Ng, A., Jordan, M., Latent Dirichlet Allocation, JMLR, 2003

Non-parametric Bayesian learning

- Traditional model selection: (1) Compare models that vary in complexity by measuring how well they fit the data, (2) Complexity penalty
- *Bayesian nonparametric (BNP) approach is to fit a single model that can adapts its complexity to the data.* Example: Do not fixing the number of clusters but estimates how many clusters are needed to model the observed data.
- Two common models
  - BNP mixture models (Chinese restaurant process mixture) infers the number of clusters from the data.
  - Latent factor models decompose observed data into a linear combination of latent factors (provide dimensionality reduction when # factor < # dimension).





Gershman and Blei, A tutorial on Bayesian nonparametric models, J. of Mathematical Psychology, 2012
Non-parametric statistics: non assumption about probability distribution or non-fixed structure of model.

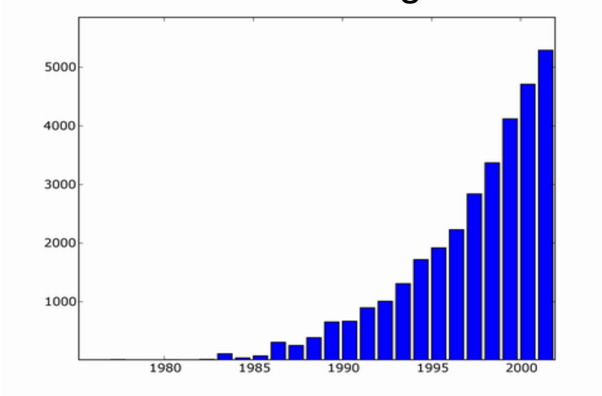
Non-parametric Bayesian learning

- The basic computational problem in BNP modeling (as in most of Bayesian statistics) is computing the posterior.
- The most widely used posterior inference methods in Bayesian nonparametric models are Markov Chain Monte Carlo (MCMC) methods. The idea MCM methods is to define a Markov chain on the hidden variables that has the posterior as its equilibrium distribution (Andrieu et al., 2003).
- An alternative approach to approximating the posterior is variational inference (Jordan et al., 1999), which is based on the idea of approximating the posterior with a simpler family of distributions and searching for the member of that family that is closest to it.
- Limitations: hierarchical structure, time series models, spatial models, supervised learning.

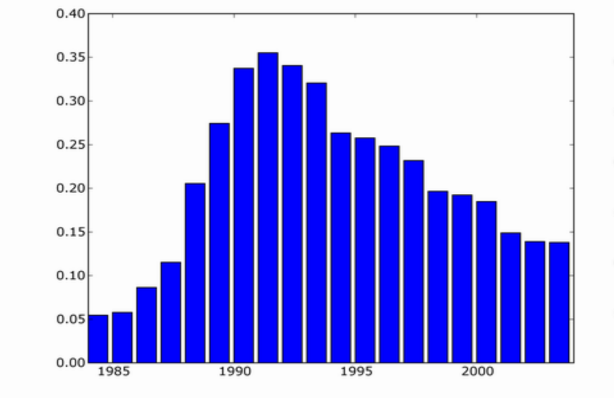
Trends in machine learning (Google scholar)

December 16, 2005

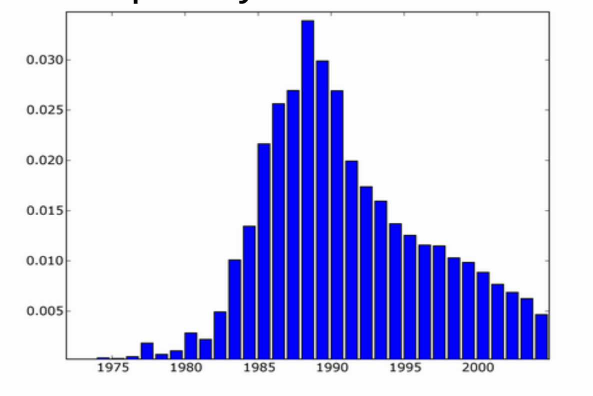
Machine learning



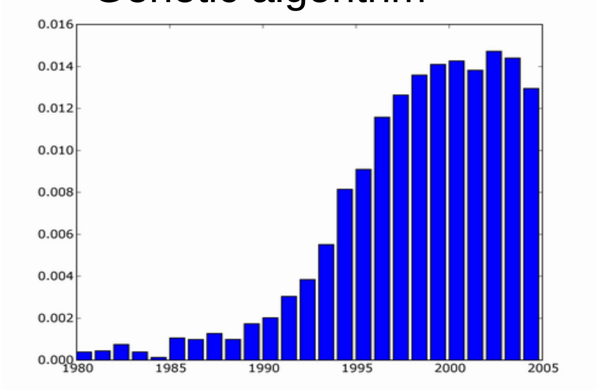
Neural network



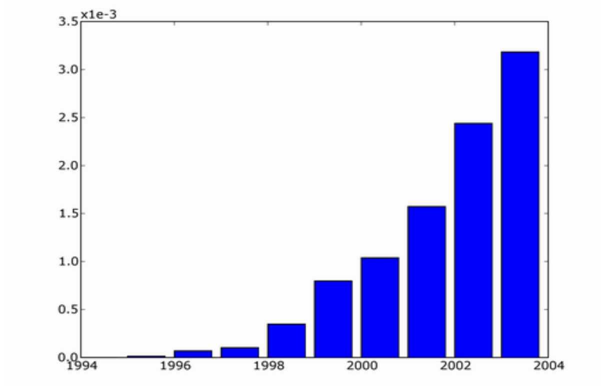
Expert systems



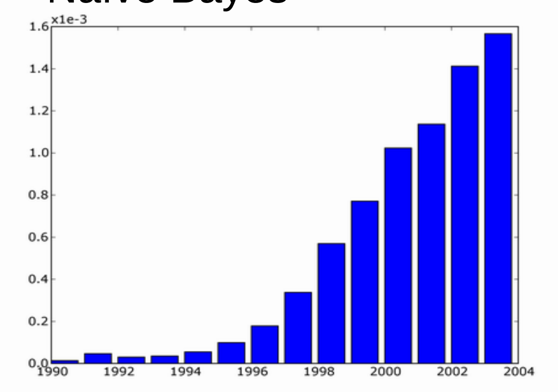
Genetic algorithm



SVM



Naïve Bayes



Content

1. Basis of machine learning
2. Recent directions and some challenges
3. Machine learning in other sciences

"Les attentes le plus vives concernent des secteurs où les mathématiques se frottent aux autres disciplines''. (Rien n'arrête les mathématiques, J. CNRS, 5.2010)

''những mong đợi lớn nhất nằm ở các lĩnh vực có sự thâm nhập của toán học vào khoa học khác".

Cédric Villani (Fields medal 2010)

Machine learning and language processing

Essence of NLP

text

Lexical / Morphological Analysis

Word segmentation

Tagging

Chunking

Syntactic Analysis

Grammatical Relation Finding

Named Entity Recognition

Word Sense Disambiguation

Semantic Analysis

Reference Resolution

Discourse Analysis

meaning

The woman will give Mary a book

POS tagging

The/Det woman/NN will/MD give/VB Mary/NNP a/Det book/NN

chunking

[The/Det woman/NN]$_{NP}$ [will/MD give/VB]$_{VP}$ [Mary/NNP]$_{NP}$ [a/Det book/NN]$_{NP}$

relation finding

subject

[The woman] [will give] [Mary] [a book]

i-object

object

Machine learning and language processing

Archeology of NLP

Trainable parsers

1990s–2000s: Statistical learning

- algorithms, evaluation, corpora

1980s: Standard resources and tasks

- Penn Treebank, WordNet, MUC

Web since 1990

1970s: Kernel (vector) spaces

- clustering, information retrieval (IR)

1960s: Representation Transformation

- Finite state machines (FSM) and Augmented transition networks (ATNs)

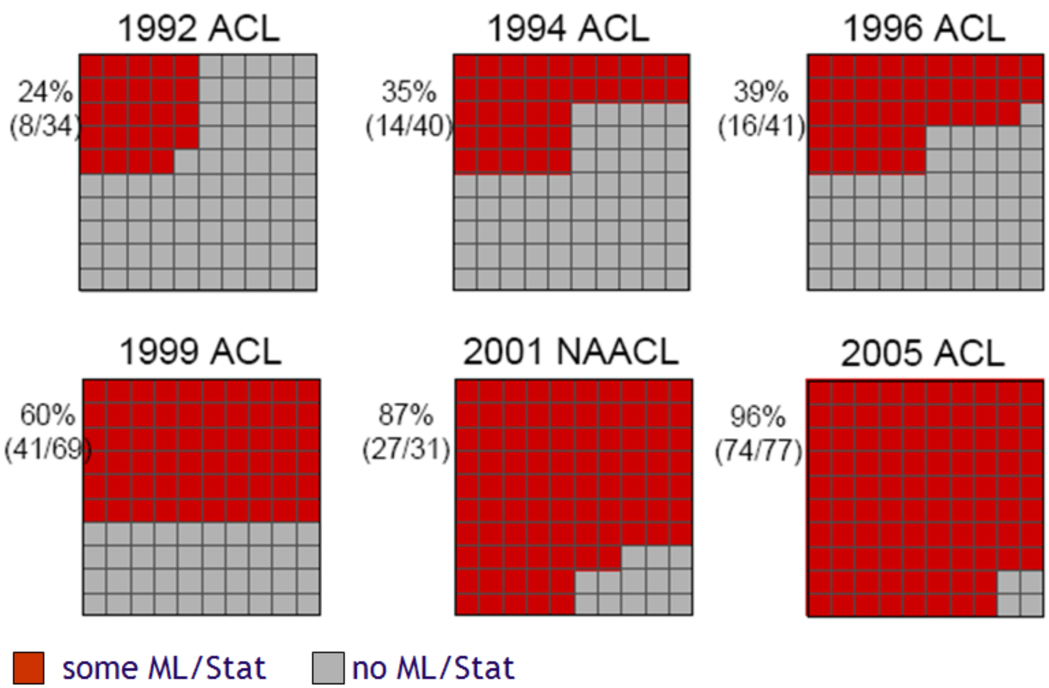
1960s: Representation—beyond the word level

- lexical features, tree structures, networks

From Levy, COLING 2004

Machine learning and language processing

More statistical machine learning in NLP



- Manual software development of robust NLP systems is very difficult and time-consuming.
- Most current state-of-the-art NLP systems are constructed by using machine learning methods trained on large supervised corpora.

From Marie Claire's talk, ECML/PKDD 2005

Machine learning and language processing

Information retrieval (IR)

- **Narrow-sense**: Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within *large collections* (usually on computers).
- **Broad-sense**:
  - General problem: how to manage text information?
  - How to find useful information? (information retrieval), e.g., Google
  - How to organize information? (text classification), e.g., automatically assign email to different folders
  - How to discover knowledge from text? (text mining), e.g., discover correlation of events.

LEARNING TO RANK

MULTI-LABEL CLASSIFICATION

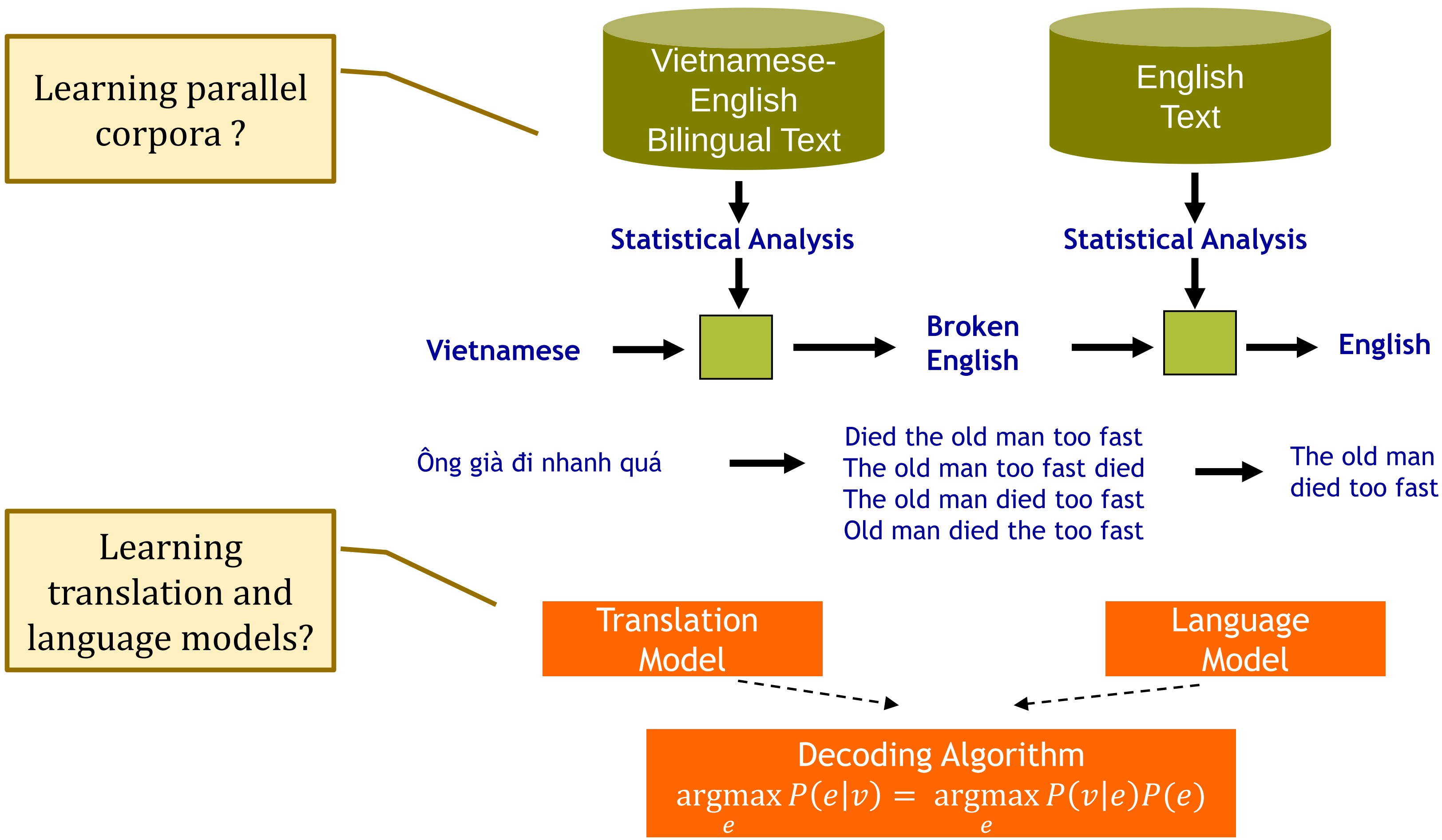
DIMENSIONALITY REDUCTION

TOPIC MODELING

WEB SEARCH

Machine learning and language processing

Statistical machine translation



Machine learning and language processing

Some challenges

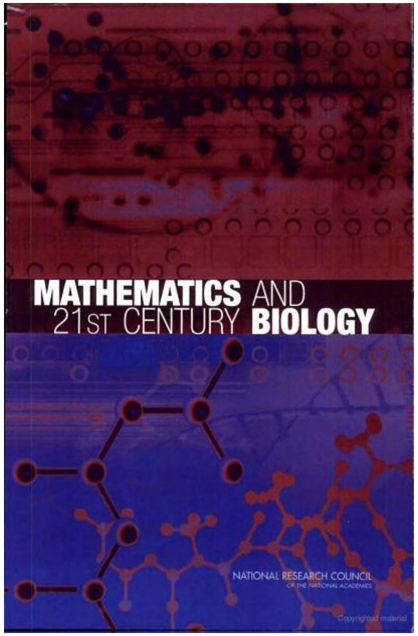
- (Semi)Automate the construction of corpora to be use in statistical algorithms by machine learning.
- Employ and develop advanced statistical machine learning methods to effectively solve problems in language processing: structured prediction, transfer learning, topic modeling, ranking, etc.
- Combine domain knowledge of each language (Vietnamese) into general statistical learning methods.
- Ambiguity, scale, and sparsity are the main challenges for statistical techniques for language processing.
- Usage: Know which methods are appropriate for each task in language processing.

Machine learning and molecular medicine

Mathematics for biology in the 21st century

- Understanding molecules (phân tử)
- Understanding cells (tế bào)
- Understanding organisms (vật sống)
- Understanding populations (quần thể)
- Understanding communities and ecosystems (cộng đồng, hệ sinh thái)

As math for physics in the 20th century

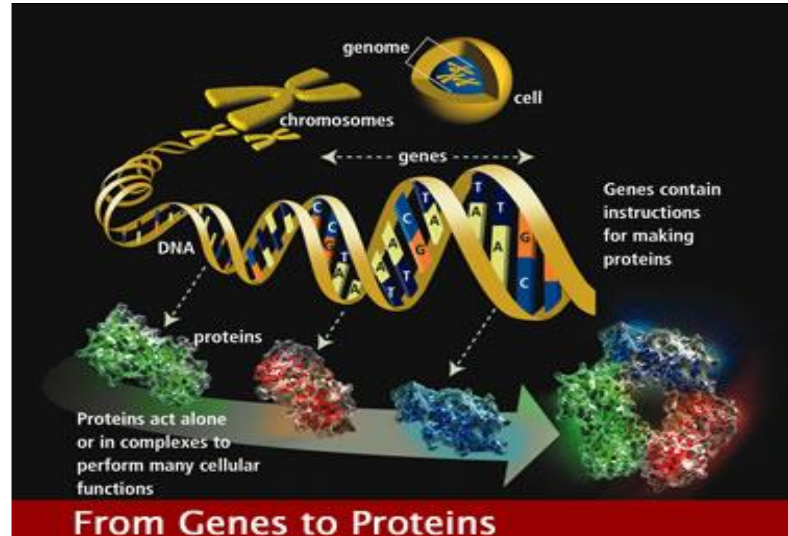


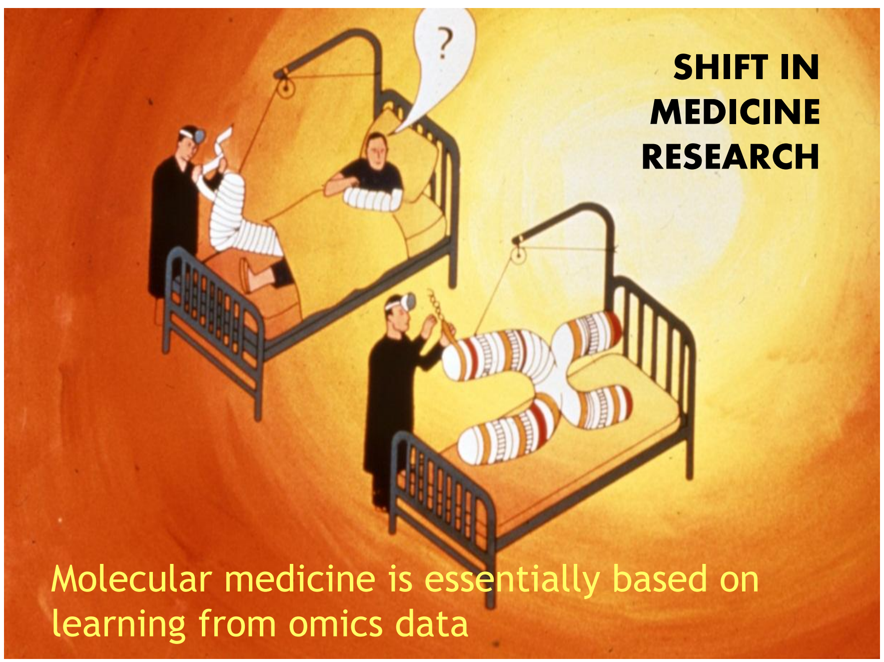
National Academy of Sciences. The National Academies Press, 2005
http://www.nap.edu/catalog.php?record_id=11315

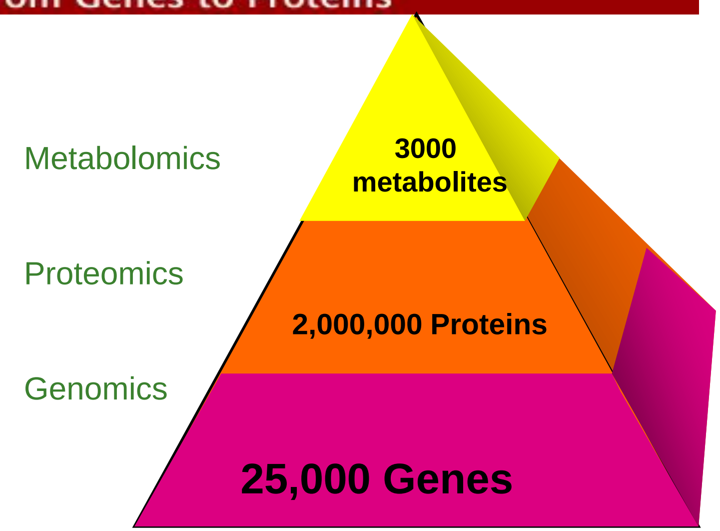
Toán học trong khoa học máy tính và khoa học về sự sống (Tia Sáng, 9.2010)
http://www.tiasang.com.vn/Default.aspx?tabid=111&CategoryID=2&News=3434

Machine learning and molecular medicine

Molecular medicine



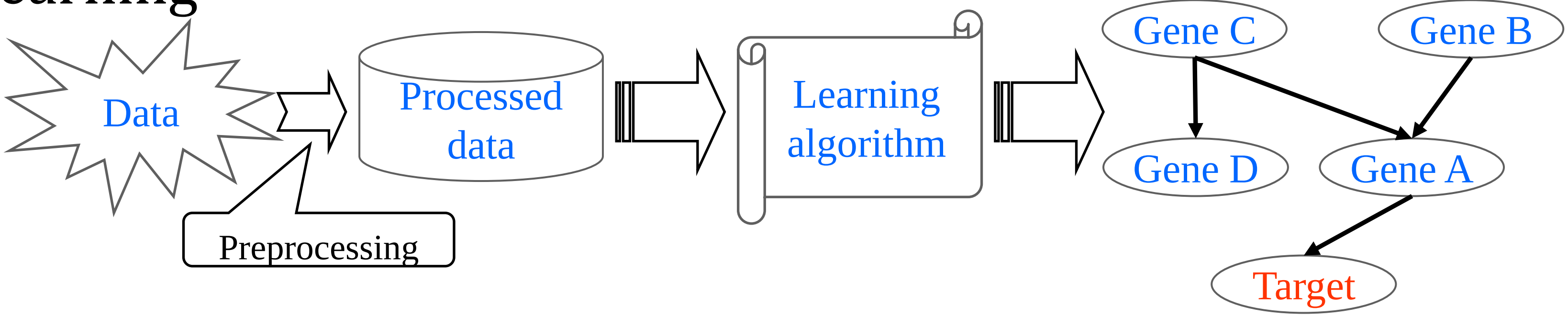




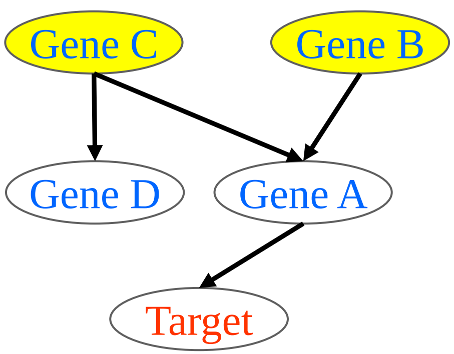
Machine learning and molecular medicine

Relations between disease and symptoms

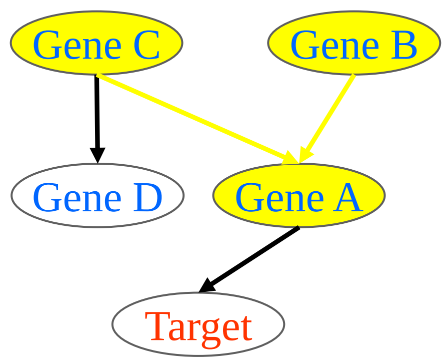
Learning



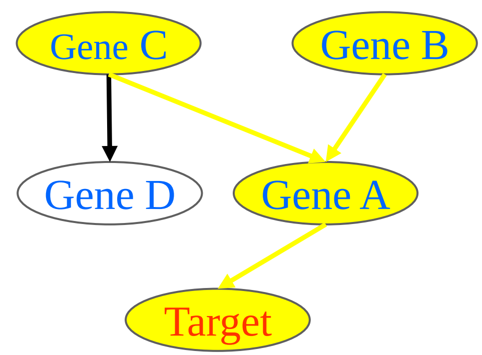
Inference



The values of Gene C and Gene B are given.



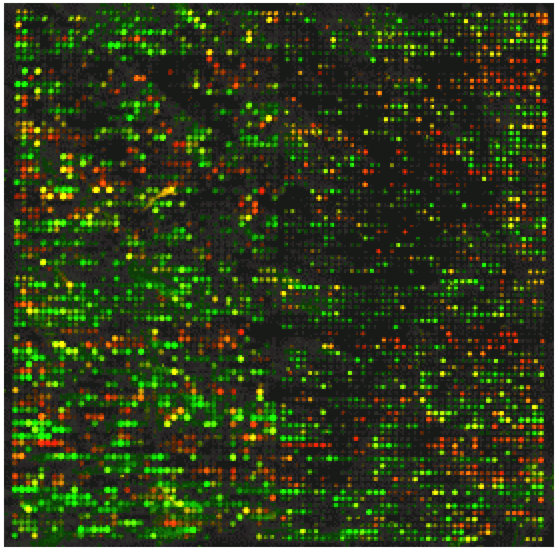
Belief propagation

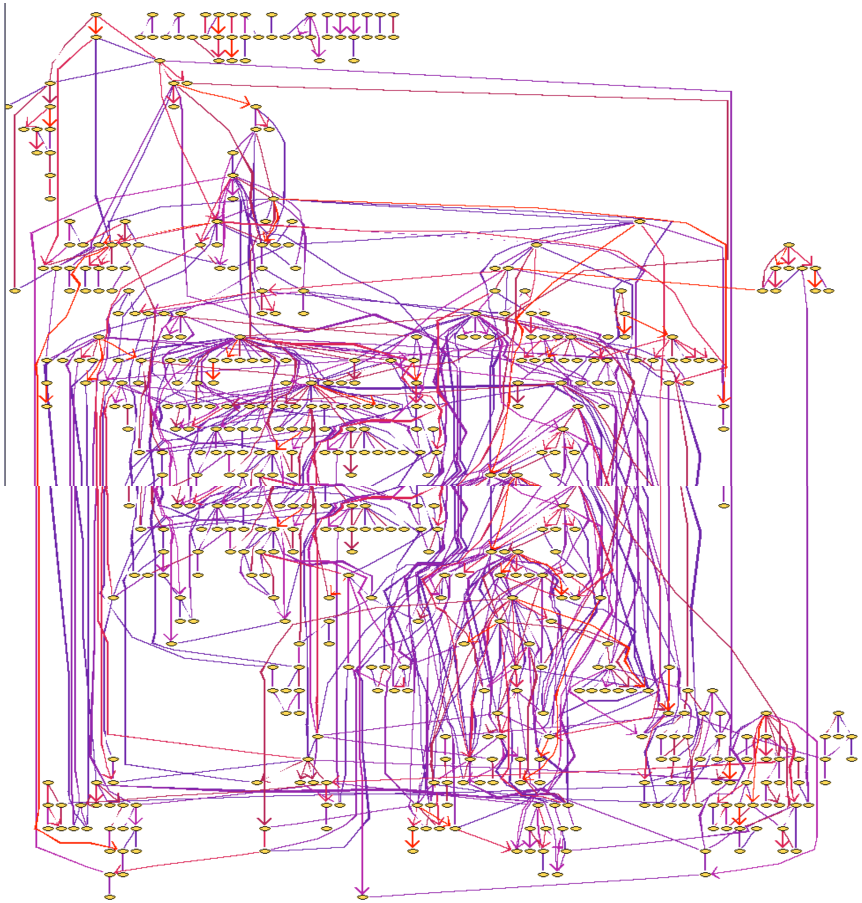


Probability for the target is computed.

Machine learning and molecular medicine

Discovering biological network (reconstruction)



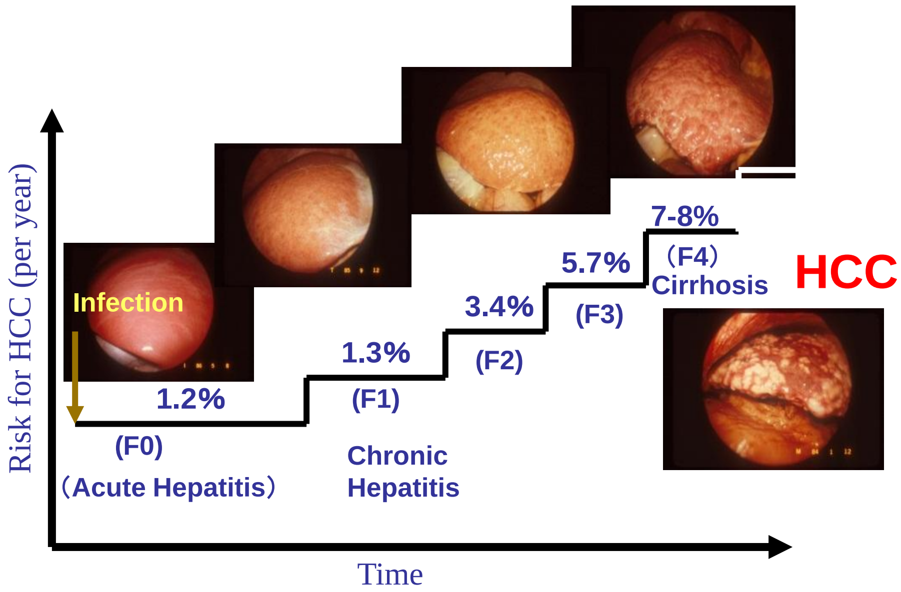


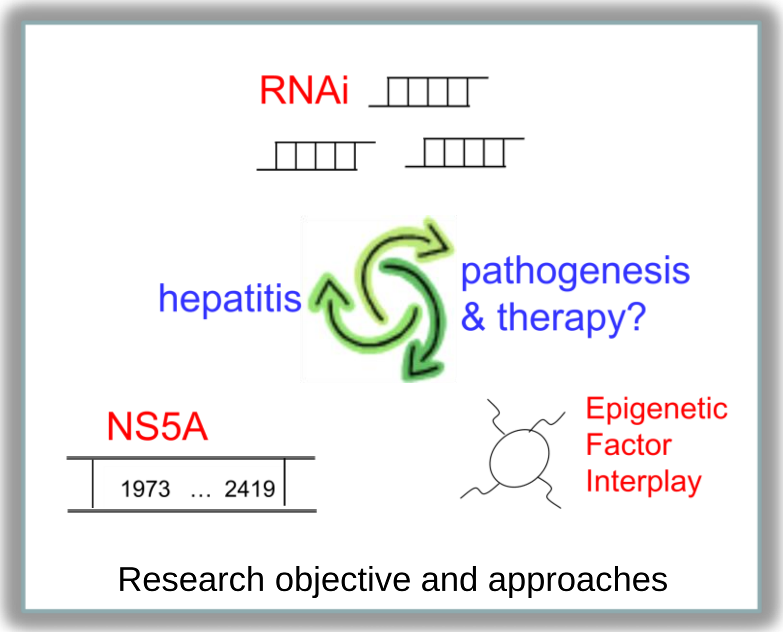
Machine learning and molecular medicine

Liver disease study

Project's goal (2010-2013)

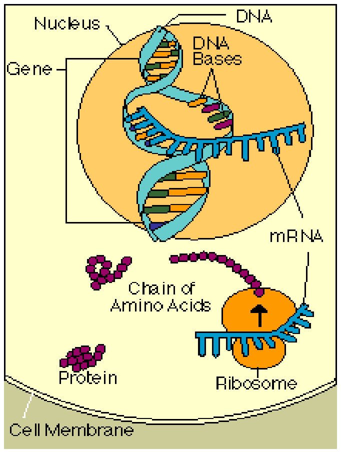
Develop methods to exploit omics data for creating new and significant knowledge on pathology and therapy of liver diseases.



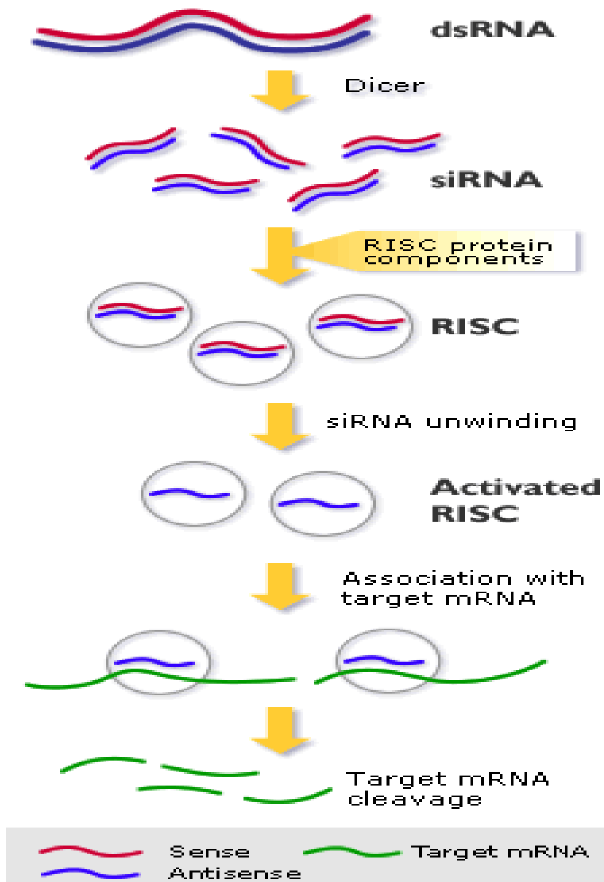


Machine learning and molecular medicine

RNA interference (RNAi) and hepatitis



DNA → mRNA → Protein



Fire, A., Mello, C., Nobel Prize 2006

- RNAi (siRNA and miRNA) is post-transcriptional gene silencing (PTGS) mechanism.
- Chemically synthesized siRNAs can mimic the native siRNAs produced by RNAi but having different ability.
- **Problem**: Selection of potent siRNAs for silencing hepatitis viruses?

Machine learning and molecular medicine

RNA interference (RNAi) and hepatitis

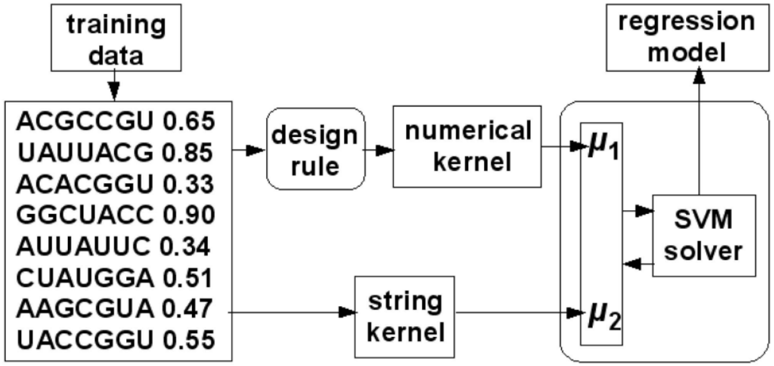
Which siRNA have high knockdown efficacy from 274.877.906.994 siRNA sequences of 19 characters from {A, C, G, U}?

Empirical siRNA design rules

Position/Nucleotide	A	C	G	U
17	C> A> G	A >U> C		U> C> G
12	A>C=G	A>U>C	A >U >G	C>G>U
…	…	…	…	…

Machine learning approach

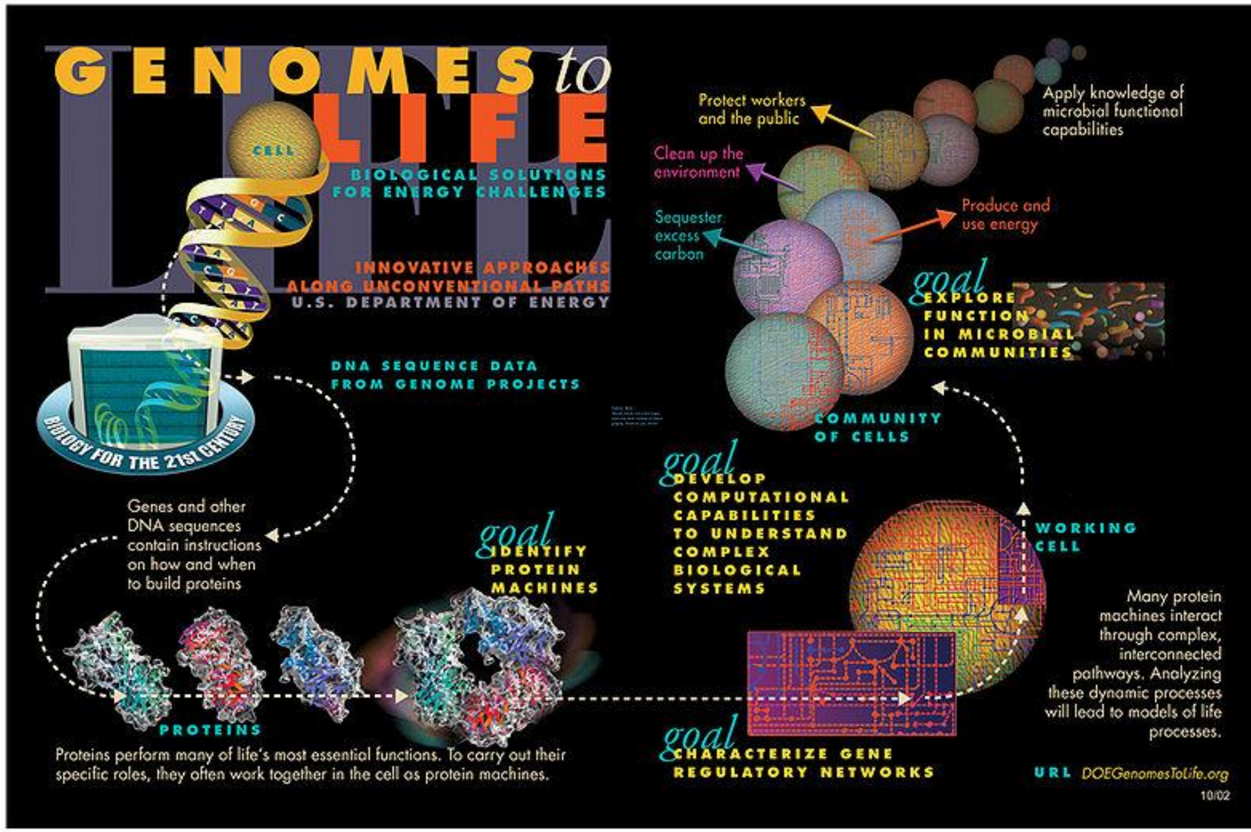
(Qiu, 2009; Takasaki 2009; Alistair 2008, etc.)



- Learn a function f(.) that scores the knockdown efficacy of given siRNAs?
- Generate siRNA with highest knockdown efficacy?

Machine learning and molecular medicine
Graphical models in bioinformatics

- **Genomics**: Modeling of DNA sequences: gene finding by HMM, splice site prediction by BN.
- **Preteomics**: Protein contact maps prediction or protein fold recognition by BN.
- **Systems biology**: Complex interactions in biological systems

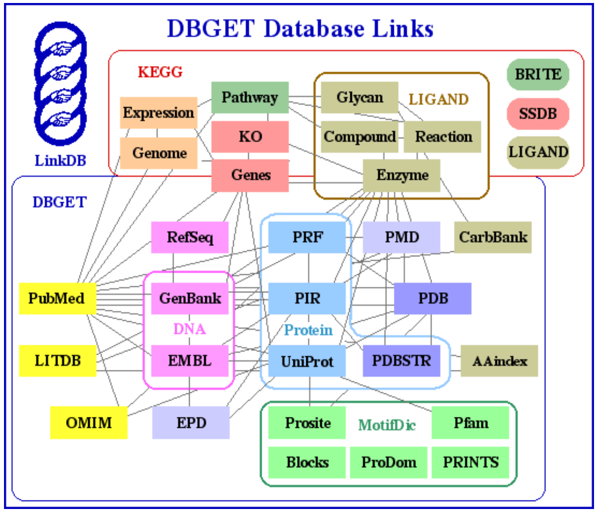


Pedro Larranaga et al., Machine learning in bioinformatics, Briefing in Bioinformatics, 2006
Tran, D.H., Pham, T.H., Satou, K., Ho, T.B. (2006). Conditional Random Fields for Predicting and Analyzing Histone Occupancy, Acetylation and Methylation Areas in DNA Sequences.

Machine learning and molecular medicine

Some challenges

- New problems raise new questions
- Large scale problems especially so
  - Biological data mining, such as HIV vaccine design
  - DNA, chemical properties, 3D structures, and functional properties → need to be fused
  - Environmental data mining
  - Mining for solving the energy crisis
- Network reconstruction (graphical models, Bayesian nonparametric models, etc.)



Nguyen, T.P., Ho, T.B. (2011). Detecting Disease Genes Based on Semi-Supervised Learning and Protein-Protein Interaction Networks, Artificial Intelligence in Medicine, Vol. 54, 63-71

Nguyen, T.P., Ho, T.B. (2008). An Integrative Domain-Based Approach to Predicting PPI, *Bioinformatics and Comput.Biology*, Vol. 6, Issue 6

Take home message

- Statistical machine learning has greatly changed machine learning.
- It opened opportunities to solve complicated learning problems.
- However it is difficult and need big effort to learn.
- Machine learning systems can always get better, learn more, work faster and in ever more ways.

Program of Statistical Machine Learning

Hồ Tú Bảo,
18-22 June

1. An overview of machine learning, recent directions
2. Regression
3. Kernel methods and SVM
4. Dimensionality reduction
5. Graphical model and topic modeling

Nguyễn Xuân Long,
30 July-3 August

1. Finite and hierarchical mixture models
2. Dirichlet, stick-breaking and Chinese restaurant processes
3. Infinite mixture models
4. Nonparametric Bayes: Hierarchical methods
5. Nonparametric Bayes: Asymptotic theory

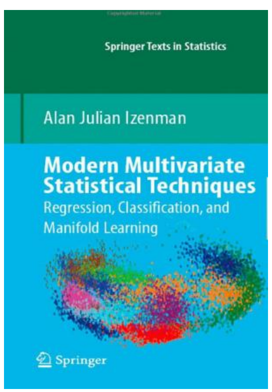
John Lafferty,
6-10 June 2012

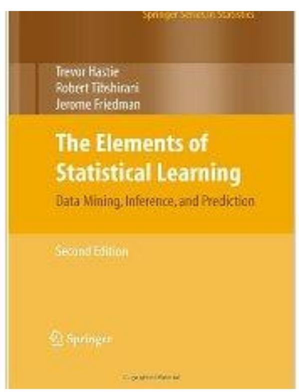
1. Sparsity in regression
2. Graphical model structure learning
3. Nonparametric inference
4. Topic models

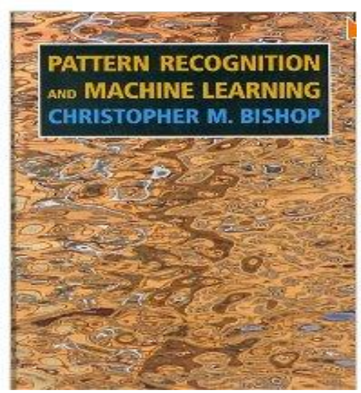
Discussion through the project period, especially 12-18 August 2012

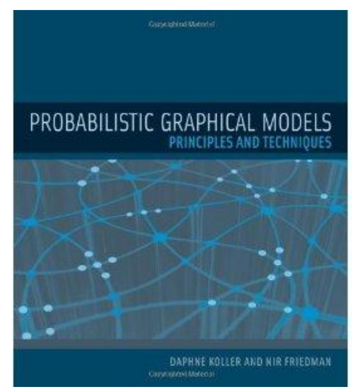
Lecture schedule

Day	Lecture	Content
18/6	L1	Machine learning: Recent directions, some challenges and what it can do for other sciences
19/6	L2	Model assessment and selection in regresion
20/6	L3	Kernel methods and support vector machines
21/6	L4	Dimensionality reduction and manifold learning
22/6	L5	Graphical models and topic models









Michael I. Jordan's students & postdoc (58)

- Francis Bach, Prof., ENS: graphical models, sparse methods, kernel-based learning
- Yoshua Bengio, Prof., U. Montréal: Deep learning, ML for understanding AI
- David Blei, A. Prof., Princeton U.: PGM, topic models, BNM
- Zoubin Ghahramani, Prof., U. Cambridge: Gaussian, BNM, inference, PGM, SSL,...
- Gert Lanckriet, A. Prof., U. San Diego: Computer music, Opt & ML, MKL, bioinfo.
- XuanLong, Ass Prof., U. Michigan: SML & Opt., BNM, distributed stat. inference,...
- Andrew Ng, A.Prof., Stanford U.: Unsup. Learning, Deep Learning, Robitics,...
- Lawrence Saul, Prof, U San Diego: App. of ML to computer systems & security
- Ben Taskar, Ass Prof, U Penn.: Determinantal point processes, Structured Pred.
- Yee-Whye Teh, Lect, U. Col. London: HDP (919), BNM, Bayesian tech, Appro. Infer.
- Martin Wainwright, Prof., U. Berkeley: PGM, stat. signal & image, coding & compres.
- Yair Weiss, Professor, Hebrew University
- Daniel Wolpert, Prof, U. Cambridge: Motor neuroscience
- Eric Xing, A. Prof, CMU: ML and biology, PGM,...