Khoa học phân tích dữ liệu lớn và Học máy thống kê

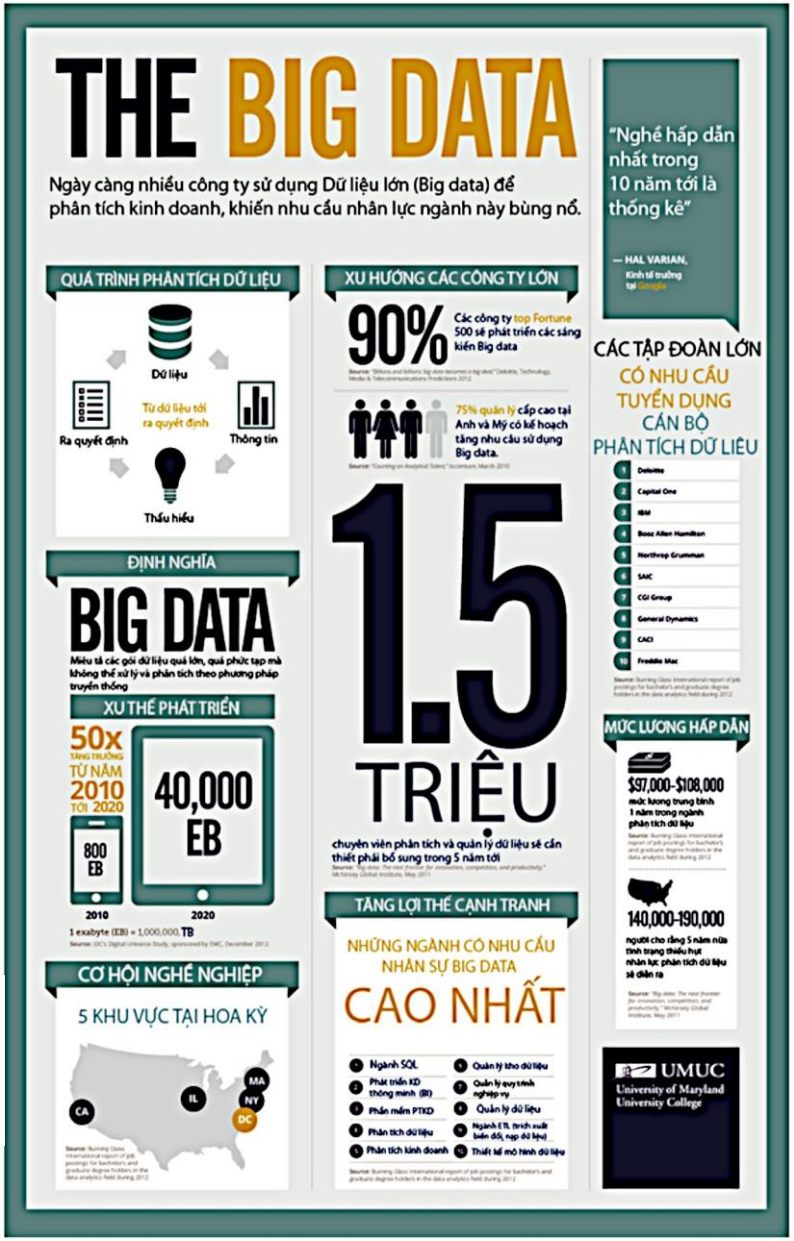
BIG DATA ANALYTICS AND STATISTICAL MACHINE LEARNING

Hồ Tú Bảo

Japan Advanced Institute of Science and Technology

Content

1. Big data analytics
2. Statistical machine learning



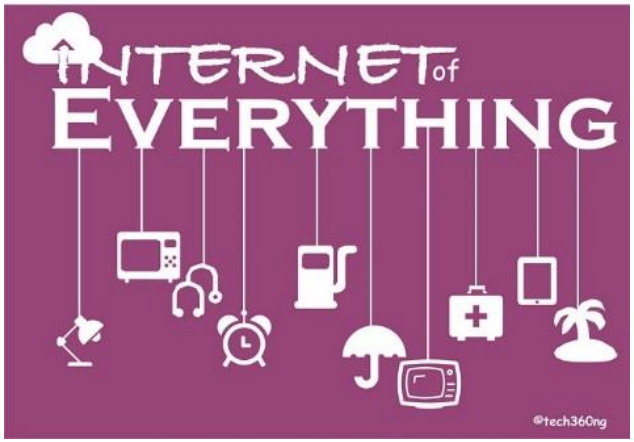
VNEXPRESS
TIN NHANH VIETNAM
Báo tiếng Việt nhiều người xem nhất

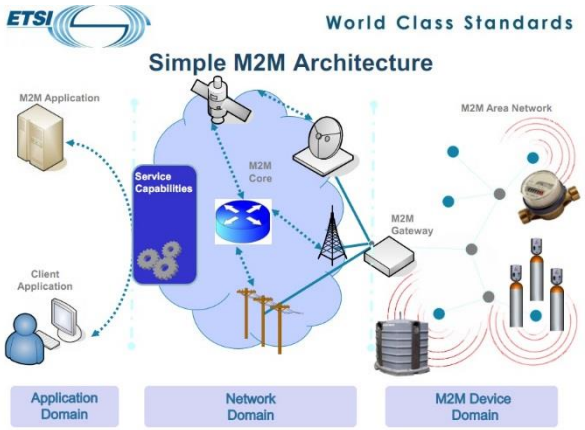
Thứ ba, 7/72015

Nhu cầu nhân lực khổng lồ cho Big Data

Những xu hướng ảnh hưởng của CNTT

Điện toán đám mây

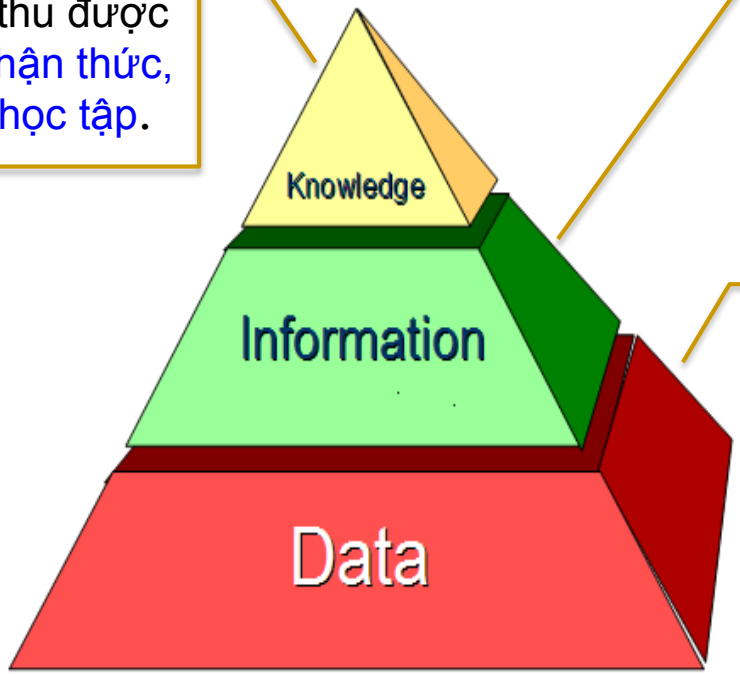




M2M (Machine to Machine)

Data, information, knowledge

Tri thức là **thông tin tích hợp**, như quan hệ giữa các sự kiện, giữa các thông tin... thu được qua quá trình nhận thức, phát hiện hoặc học tập.



Thông tin là **dữ liệu với ý nghĩa** (data equiped with meaning), thu được khi **xử lý dữ liệu** để lọc bỏ đi các phần dư thừa, tìm ra phần cốt lõi đặc trưng cho dữ liệu.

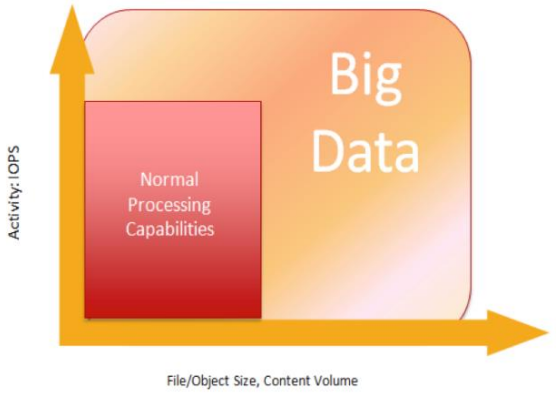
Dữ liệu là **tín hiệu** (signals) thu được do **quan sát, đo đạc, thu thập**... từ các đối tượng. Cụ thể, dữ liệu là **giá trị (values)** của các **thuộc tính (features)** của các đối tượng, được biểu diễn bằng dãy các bits, các con số hay ký hiệu…

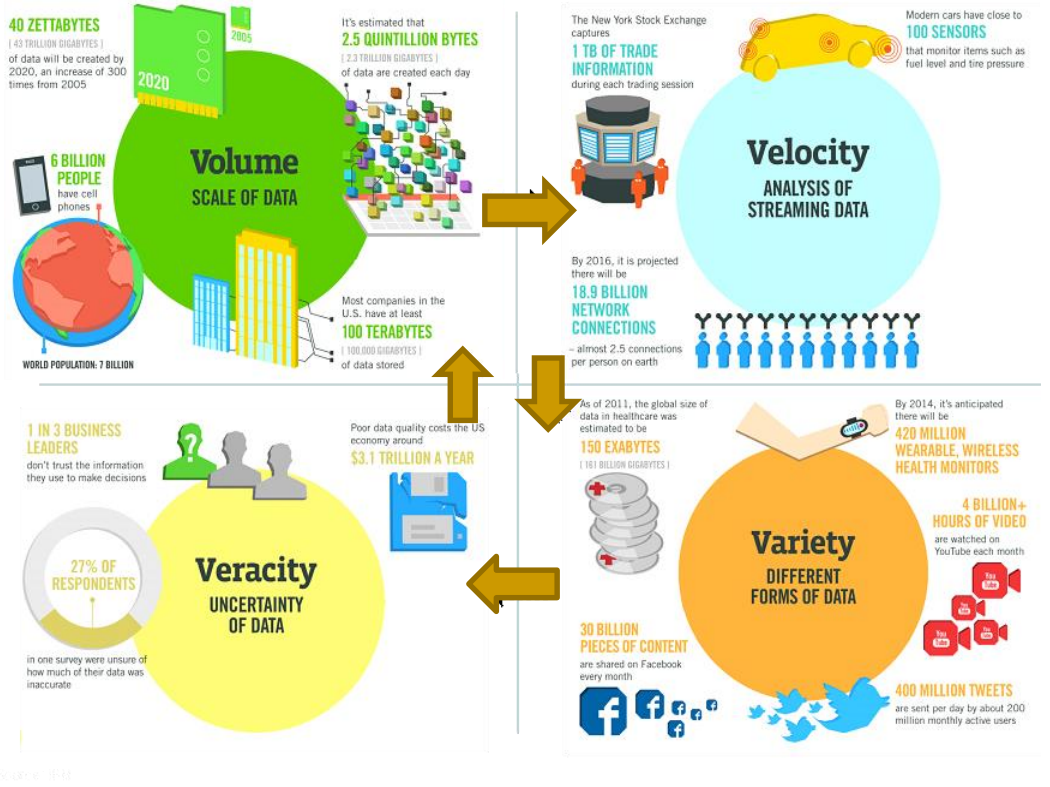
Dữ liệu ở mức độ trừu tượng thấp nhất và cụ thể nhất, thông tin ở mức trên dữ liệu và tri thức ở mức cao nhất.

Big data là gì?

Dữ liệu lớn nói về các **tập dữ liệu rất lớn**

và/hoặc **rất phức tạp**, vượt quá khả năng xử lý của các kỹ thuật IT truyền thống (*View 1*).



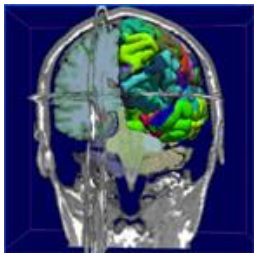


(View 2) Big Data is about technology (tools and processes).

(View 3) Hiện tượng khách quan mà các tổ chức, doanh nghiệp… phải đối đầu để phát triển.

Rất lớn là lớn thế nào?

Kích thước lớn và rất nhiều chiều



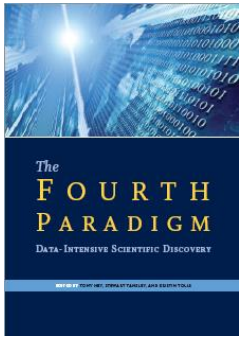
1 human brain at the micron level = 1 PetaByte

Large Hadron Collider, (PetaBytes/day)

Human Genomics = 7000 PetaBytes 1GB / person

Printed materials in the Library of Congress = 10 TeraBytes

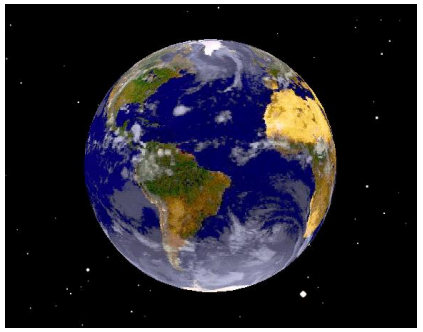
200 of London's Traffic Cams (8TB/day)



1 book = 1 MegaByte

Family photo = 586 KiloBytes

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

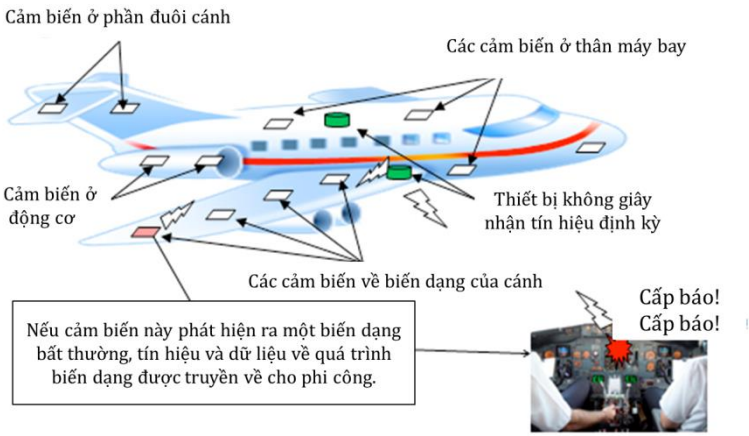


All worldwide information in one year = 2 ExaBytes

Dữ liệu lớn có thể rất nhỏ.
Không phải mọi tập dữ liệu to đều lớn

Big data can be very small. Not all large datasets are big

- **Big** liên quan tới sự **phức tạp** nhiều hơn tới **kích thước lớn**.
- **Dữ liệu lớn** nhưng lại nhỏ
  - Lò hạt nhân, máy bay… có hàng trăm nghìn sensors → sự phức tạp của việc tổ hợp dữ liệu các sensors này tạo ra?
  - Dòng dữ liệu của tất cả các sensors là lớn mặc dù kích thước của tập dữ liệu là không lớn (một giờ bay:
    100,000 sensors x 60 minutes
    x 60 seconds x 8 bytes < 3GB).
- Tập dữ liệu **to nhưng không lớn**
  - Số hệ thống dù tăng lên và tạo ra những lượng khổng lồ dữ liệu nhưng đơn giản.



Biến dữ liệu lớn thành giá trị

Turning big data into value

- Dữ liệu lớn nhưng không phân tích được cũng không có giá trị gì.
- Phân tích dữ liệu lớn cho phép các tổ chức giải quyết các bài toán phức tạp trước kia không thể làm được
  → ra quyết định và hành động tốt hơn.
- Các ưu thế cạnh tranh (Competitiveness advantages).
- Cho những hiểu biết sâu (insights) về các hành vi phức tạp của xã hội con người.
- Đột phá (breakthrough) trong khoa học.

"Chỉ Thượng đế là đáng tin, mọi thứ khác đều phải dựa vào dữ liệu"

Data analysis vs. Data analytics

Data Scientist: The Sexiest Job of the 21st Century
(Harvard Business Review, October 2012)

Dữ liệu lớn cơ hội lớn

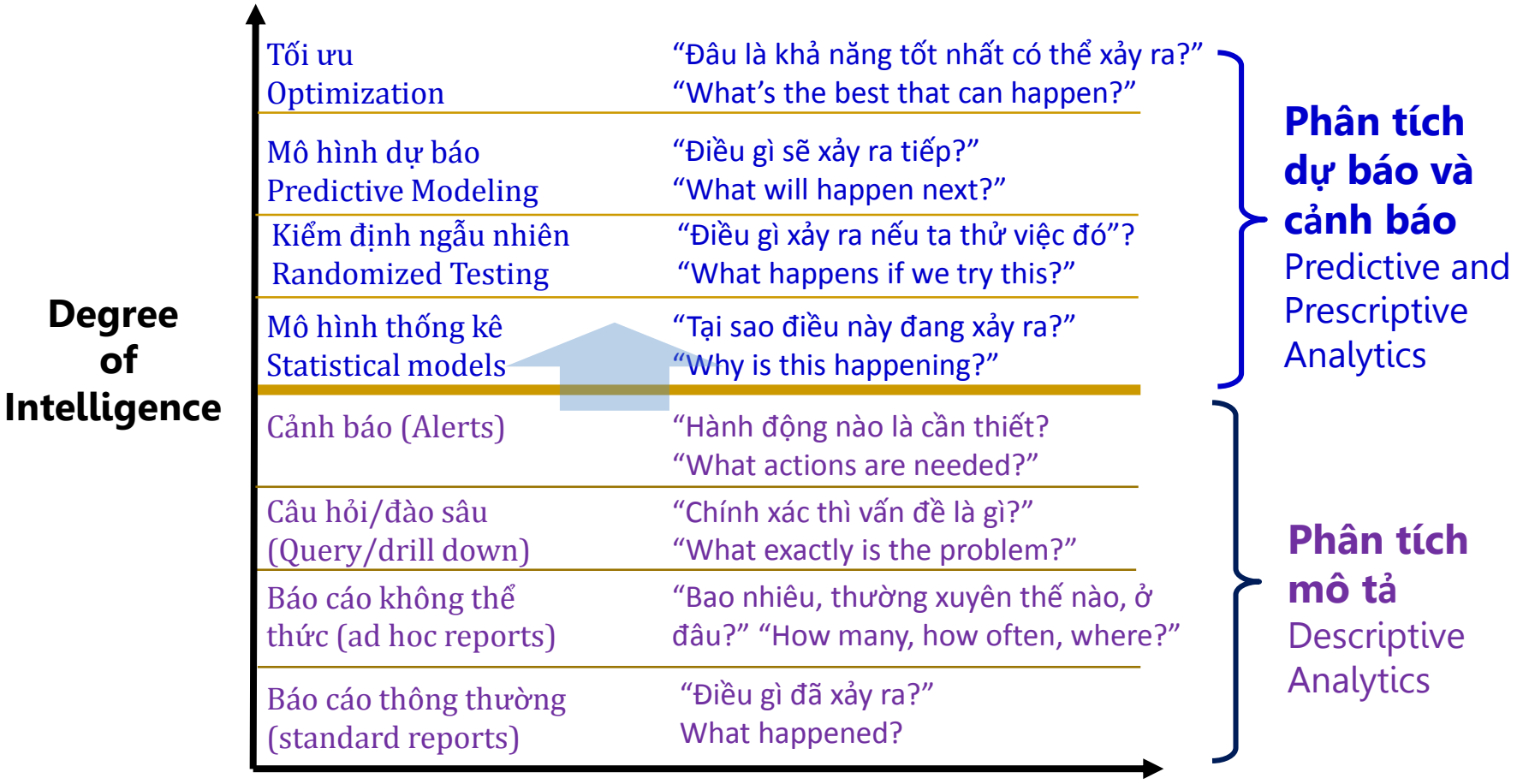
Nhiều công ty lớn chuyển dần từ chế tạo sản phẩm sang cung cấp dịch vụ, chẳng hạn như dịch vụ phân tích kinh doanh (business analytics).

- IBM's past: Chế tạo servers, desktop computers, laptops, và thiết bị cho hạ tầng cơ sở.
- IBM's today: Loại bỏ một số thiết bị phần cứng như laptops, đầu tư hàng tỷ đôla để xây dựng và nhằm tạo dựng vị trí dẫn đầu trong **phân tích kinh doanh**.

http://dawn.com/2012/07/25/big-data-big-analytics-big-opportunity/ (25 July 2012)

Khoa học phân tích dữ liệu là gì?

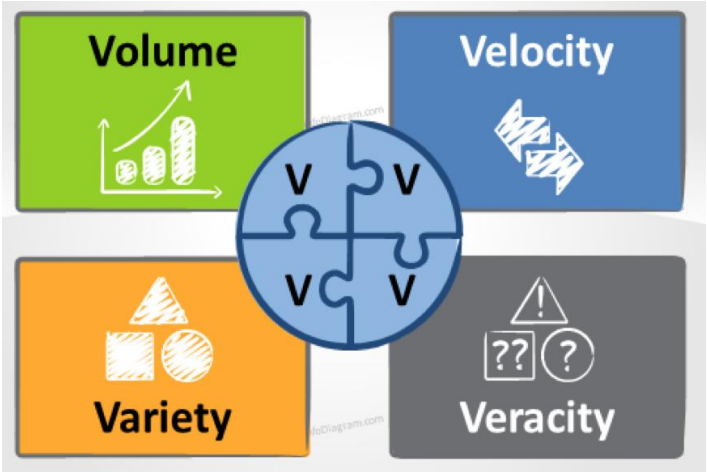
What are Data Analytics?



Tại sao phân tích dữ liệu lớn lại rất khó?

Bốn tính chất của dữ liệu (4V) & hai việc: dự đoán và phân tích quan hệ.

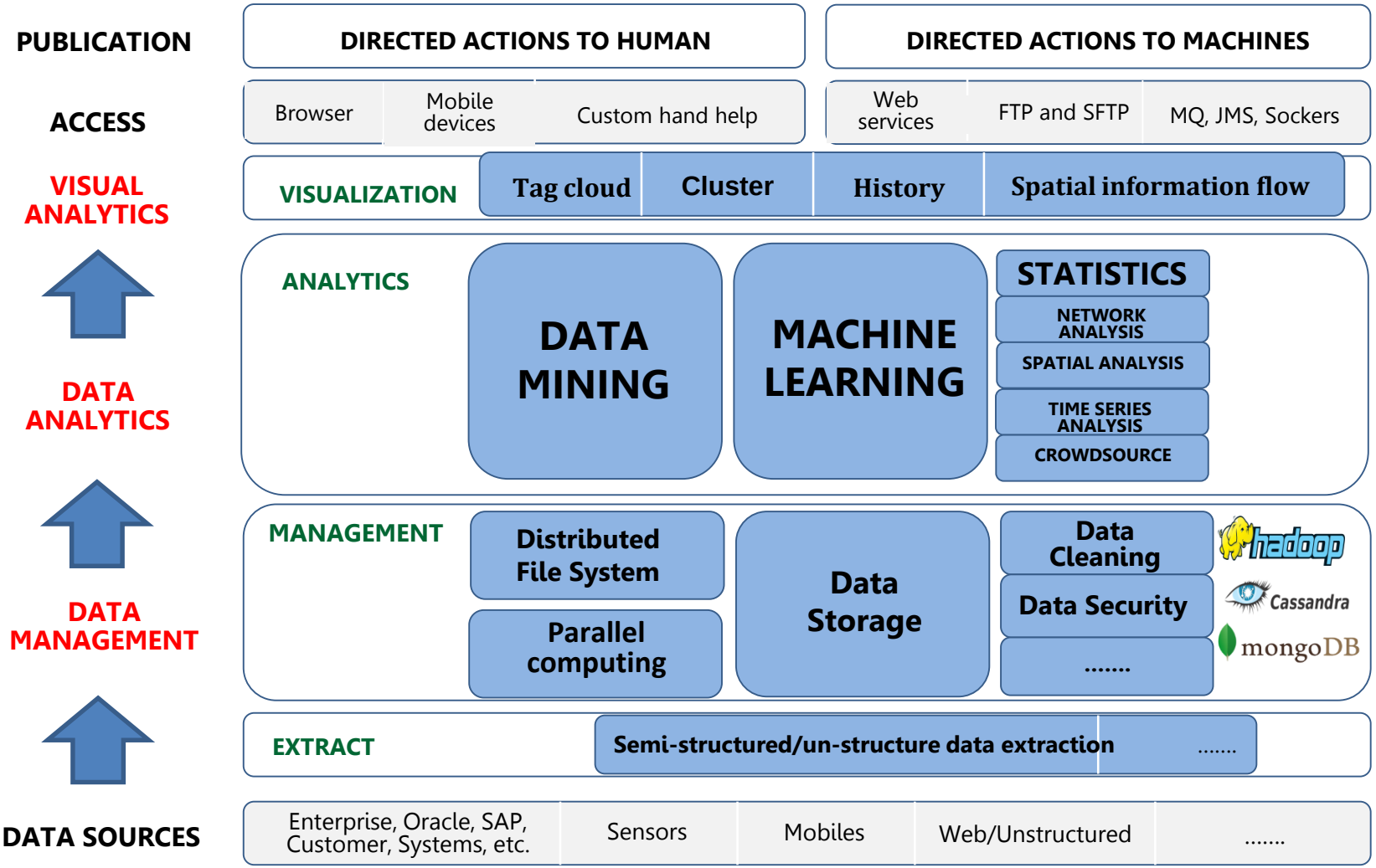
1. Số chiều rất lớn + dữ liệu kiểu khác nhau, chuyển động của dữ liệu, nhiễu trong dữ liệu → kém hiệu quả.
2. Số chiều rất lớn + số đối tượng rất lớn → tính toán nặng nề và thuật toán không khả kích (scalable).
3. Dữ liệu lớn đến từ nhiều nguồn, thu thập ở những thời điểm khác nhau bởi kỹ thuật khác nhau → không thuần nhất, khác biệt và lệch (bias).



Attribute	Numerical	Symbolic	
No structure $= \neq$		Places, Color	Nominal (categorical)
Ordinal structure $= \neq \geq$	Age, Temperature, Taste,	Rank, Resemblance	Ordinal
Ring structure $= \neq \geq + \times$	Income, Length		Measurable

Sparse modeling and dimensionality reduction

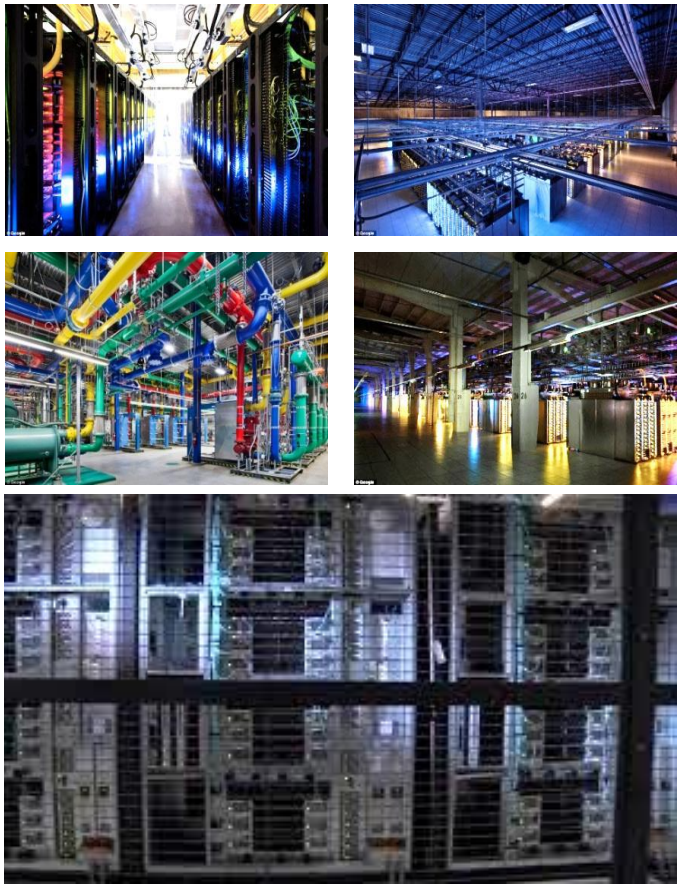
Một lược đồ phân tích dữ liệu lớn



Source: WAMDM, Web group

Cloud Storage và BigQuery của Google

Google Data Center



- Công nghệ: BigQuery (Tableau), Cloud Storage.
- Machine learning core
  - Logistic & linear regression, general convex losses
  - Infusion of L_1 and L_2 regularization
  - On-the-fly curvature estimation
- System infrastructure
  - MapReduce for parallelism
  - Multiple cores and threads per computer
  - Data stored in compressed column-based form

Problem	Number of raw features (M)	Non-zero weights (M)	Fraction of non-zero weights
A	868	20	2.3%
B	333	8	2.4%
C	1762	252	14.3%
D	2172	372	17.1%

Singer Yoram, keynote at ACML'14

Thống kê - Statistics

- **Thống kê** cung cấp các phương pháp và kỹ thuật toán học để phân tích, khái quát và quyết định từ dữ liệu.
- **Nội dung chính**
  - *Thống kê mô tả* (descriptive statistics): phân bố xác suất…
  - *Thống kê suy diễn* (inferential statistics): ước lượng và kiểm định giả thiết thống kê…)
- **Dữ liệu** thí nghiệm và dữ liệu quan sát
  - Dữ liệu thống kê thường được thu thập để *trả lời những câu hỏi được định trước* (experiment design, survey design)
  - Phần lớn là dữ liệu số, ít dữ liệu hình thức (symbolic).
- Nhiều phương pháp phát triển cho tập *dữ liệu nhỏ*, phân tích từng biến ngẫu nhiên riêng lẻ, trước khi có máy tính.

Phân tích dữ liệu nhiều biến

Multivariate analysis

- Phân tích đồng thời quan hệ của nhiều biến ngẫu nhiên
- *Phân tích thăm dò* (EDA, exploratory data analysis) dùng dữ liệu tạo ra các giả thiết vs. việc kiểm định giả thiết trong *Phân tích khẳng định* (CDA, confirmatory data analysis)
  - Factor analysis, PCA, Linear discriminant analysis
  - Regression analysis
  - Cluster analysis
- Thấy gì từ các phương pháp truyền thống?
  - Kết quả nghèo trên dữ liệu lớn và phức tạp
  - Các phương pháp truyền thống chỉ phân tích tập dữ liệu nhỏ.
  - Giá lưu trữ và xử lý dữ liệu giảm nhanh thập kỷ qua.

Phân tích dữ liệu nhiều biến
Multivariate analysis

- Phương pháp phân tích được tạo ra cho các tập dữ liệu có kích thước nhỏ hoặc trung bình, và khi máy tính còn yếu.
- Phân tích thống kê nhiều biến đang thay đổi nhanh do kỹ thuật tính toán nhanh và hiệu quả hơn. Nhiều phương pháp mới được phát triển để giải các bài toán lớn (Pagerank của Google nghịch đảo ma trận kích thước nhiều tỷ chiều)

Nov. 2012: Cray's Titan computer,
17.59 petaflops, 560640 processors.

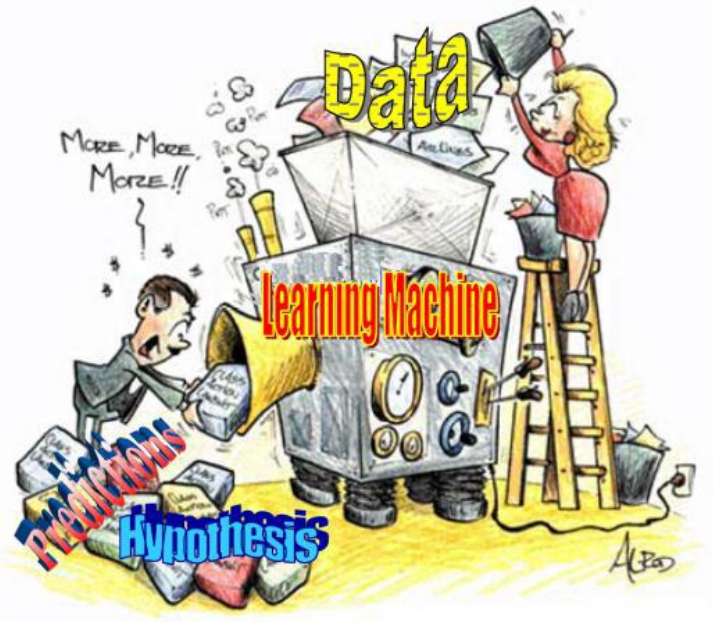
June 2013: China Tianhe-2
33.86 petaflops, 3,120,000 Intel cores

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)



(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)

Khai phá dữ liệu – Data Mining

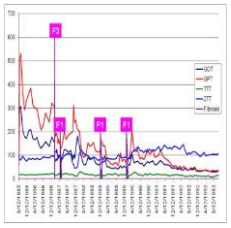
Tự động khám phá, phát hiện các tri thức tiềm ẩn từ các tập dữ liệu lớn và đa dạng.

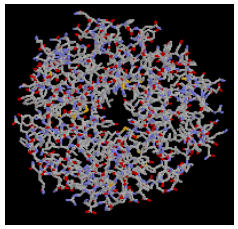
Data mining metaphor: Extracting ore from rock

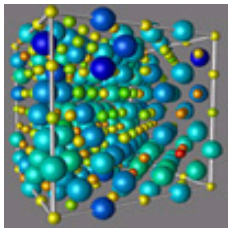
Statistics

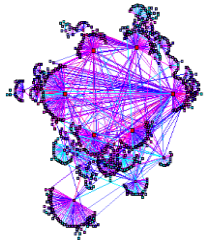
Databases **Machine Learning**

Large and unstructured real-life data









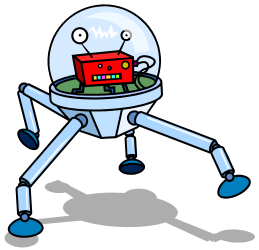
Statistics vs. Machine Learning

Statistics

- Nhấn mạnh suy diễn thống kê hình thức (ước lượng, kiểm định giả thiết).
- Dựa trên các mô hình (models) cho bài toán có số chiều nhỏ, ở dạng số.
- Khoa học đã thiết lập, ít 'văn hóa' thay đổi và thích nghi với môi trường tính toán.
- Có xu hướng mở rộng sang học máy.

Machine learning

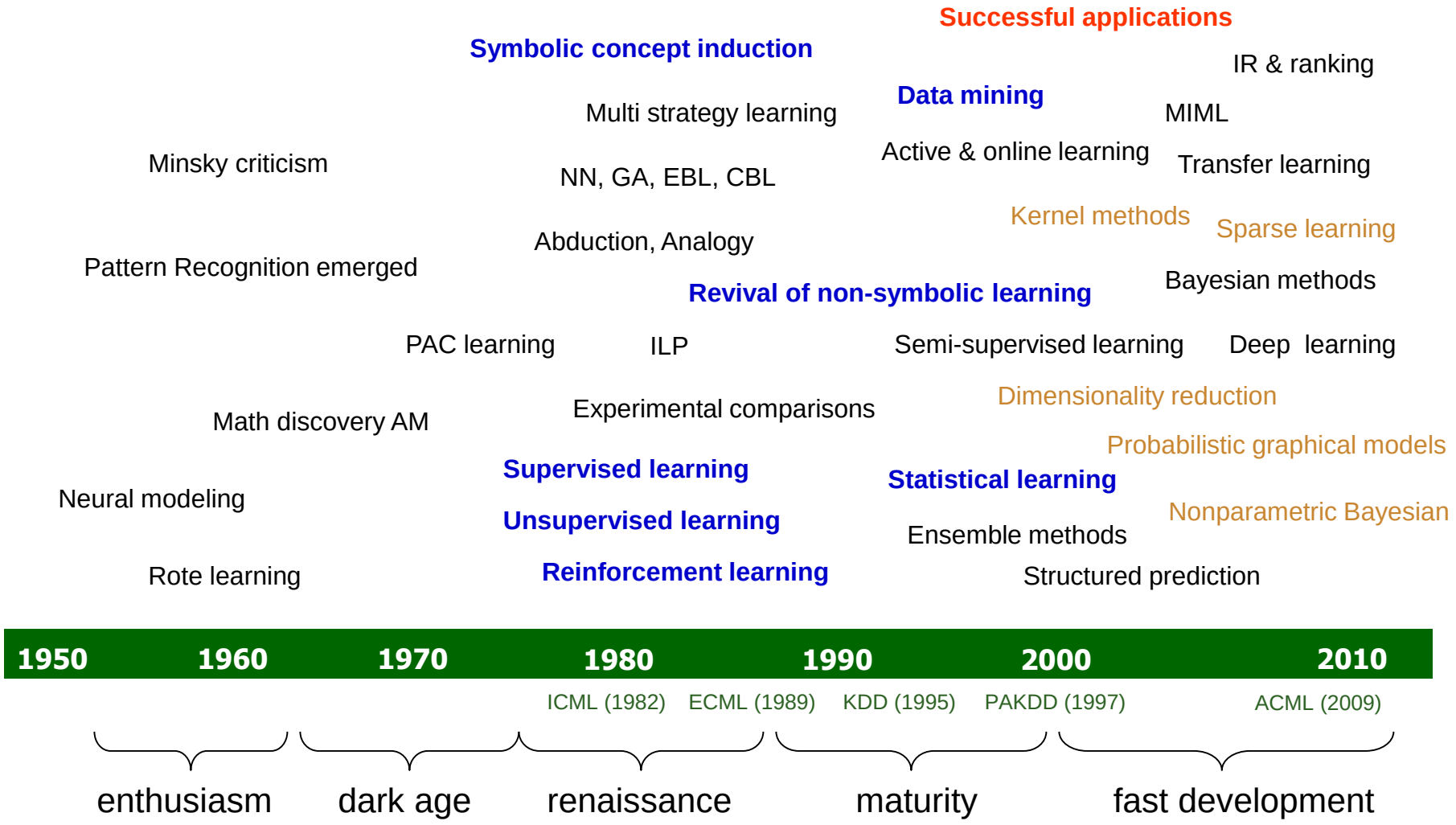
- Nhấn mạnh các bài toán dự đoán, bắt đầu với dữ liệu hình thức.
  - Bước đầu chủ yếu xây dựng và dùng các thuật toán trực cảm (heuristics algorithms).
- Gắn với thống kê nhiều hơn, xây dựng mô hình toán cho các thuật toán (statistical models underlying the algorithms).



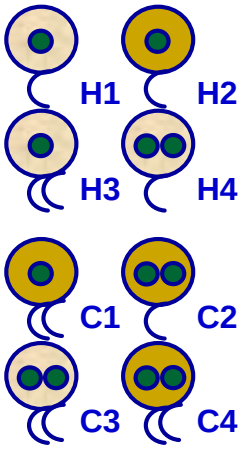
Thống kê vs. Khai phá dữ liệu

Feature	Statistics	Data Mining
Kiểu bài toán & dữ liệu	Có cấu trúc (well structured)	Không cấu trúc/Nửa cấu trúc Unstructured/Semi-structured
Mục đích phân tích và thu thập dữ liệu	Xác định mục tiêu rồi thu thập dữ liệu	Dữ liệu thu thập thường không liên quan đến mục tiêu
Kích thước dữ liệu	Nhỏ và thường thuần nhất	Lớn và thường không thuần nhất.
Mô thức/tiếp cận Paradigm/approach	Dựa trên lý thuyết suy diễn Theory based (deductive)	Phối hợp lý thuyết và trực cảm Theory & heuristic based (inductive)
Kiểu phân tích	Confirmative (khẳng định)	Explorative (thăm dò, khai phá)
Số biến	Nhỏ	Lớn
Giả định về phân bố Distribution assump.	Dựa trên giả định về phân bố	Không giả định phân bố xác suất

Development of machine learning



Supervised vs. unsupervised learning



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

$\boldsymbol{x} = (color, \#nuclei, \#tails)$ $\quad y$

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

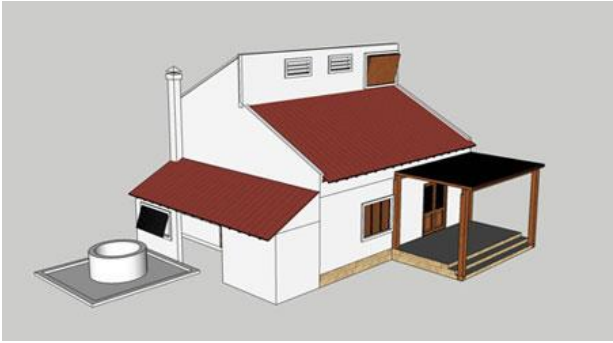
$\boldsymbol{x} = (color, \#nuclei, \#tails)$

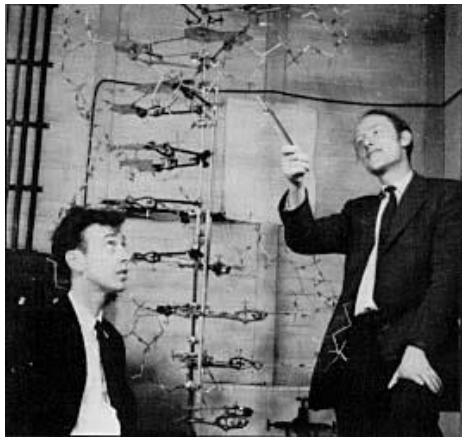
- **Classification** (y is discrete)
  Decision trees, k-NN, SVM, naïve Bayesian, etc.
- **Regression** (y is continuous)
  Linear regression (lasso, ridge), logistic regression …

- **Clustering**
- **Latent variable modeling** (EM, PCA, ICA, NMF, SOM…)
- **Association** learning
- etc.

Model and Modeling

- Model: Mô tả hay biểu diễn khái quát của một hiện thực.
- Modeling: Quá trình tạo ra mô hình.





DNA model figured out in 1953 by Watson and Crick

Grande challenges in modeling?

➔ Mô hình giao thông tại Hà Nội?

➔ Mô hình thị trường và giá cả?

➔ Mô hình một dịch bệnh?

Mô hình là tập hợp các phân bố xác suất với tham số

$$M = f(x, y; \theta) | \theta \in \Omega\}$$

Some key concepts in statistical machine learning

1. Mô hình mô tả và mô hình dự đoán
   (Generative models and discriminative models)
2. Mô hình tham số và mô hình không tham số
   (Parametric models vs. non-parametric models)
3. Lựa chọn mô hình (Model selection)
4. Quá khít (Overfitting)
5. Điều chỉnh (Regularization)
6. Mô hình thưa (Sparse modeling)
7. Giảm số chiều (Dimensionality reduction)

Some key concepts in statistical machine learning

Generative model vs. discriminative model

Generative model

- Mô hình về quan hệ của **tất cả các biến**, mô tả việc các dữ liệu được ngẫu nhiên sinh ra trong mối liên quan với một số biến ẩn.
- Học một phân bố xác suất liên hợp (joint probability distribution) của các biến quan sát được và biến đích

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x_1, \dots, x_n, y_1, \dots, y_n)$$

- Tiêu biểu cho bài toán học với dữ liệu không nhãn (unlabeled data).

Discriminative model

- Mô hình về mối quan hệ phụ thuộc có điều kiện của **biến đích** với biến quan sát được (bỏ qua việc mô hình tường minh các biến quan sát được).
- Học một phân bố xác suất có điều kiện của biến đích khi có các biến quan sát

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(y_1, \dots, y_n|x_1, \dots, x_n)$$

- Tiêu biểu cho bài toán học với dữ liệu có nhãn (labelled data).

Some key concepts in statistical machine learning

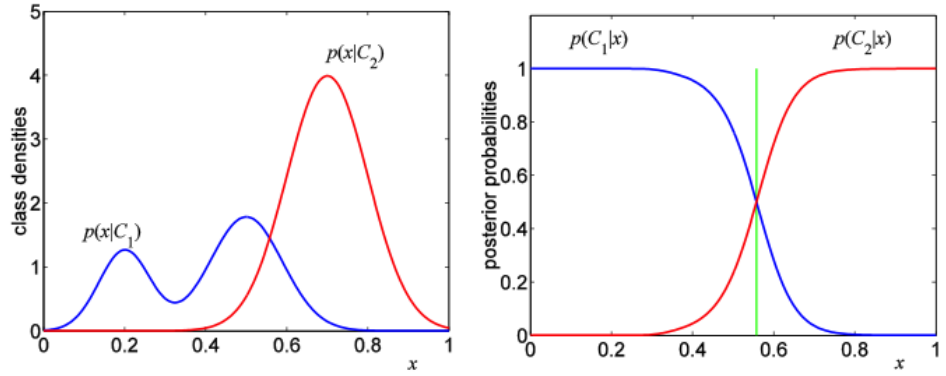
Generative model vs. discriminative model

Generative model

- Học các hàm có dạng $p(\boldsymbol{x}|\boldsymbol{y}), p(\boldsymbol{y})$.
- Ta ước lượng trực tiếp tham số $p(\boldsymbol{x}|\boldsymbol{y}), p(\boldsymbol{y})$ từ dữ liệu huấn luyện, và từ đó dùng luật Bayes để tính $p(\boldsymbol{y}|\boldsymbol{x})$.
- HMM, Markov random fields, Gaussian mixture models, Naïve Bayes, LDA, etc.

Discriminative model

- Học các hàm có dạng $p(\boldsymbol{y}|\boldsymbol{x})$
- Ước lượng tham số của $p(\boldsymbol{y}|\boldsymbol{x})$ trực tiếp từ dữ liệu huấn luyện.
- SVM, logistic regression, neural networks, nearest neighbors, boosting, MEMM, conditional random fields, etc.



Some key concepts in statistical machine learning

Parametric model vs. non-parametric model

Considering probabilistic models of the form $p(x|y)$ or $p(x)$

Parametric models

Có một số cố định các tham số (*a fixed number of parameters)*.

Một họ mô hình tham số của các phân bố có thể được mô tả bởi một số hữu hạn các tham số, dưới dạng một *vector tham số k*-chiều $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.

- Ưu điểm: Thường ước lượng nhanh được các tham số
- Hạn chế: Cần giả thiết nhiều hơn về phân bố của dữ liệu.

Non-parametric models

Có số tham số không cố định. Số tham số tăng dần theo độ lớn của dữ liệu (*number of parameters grow with the amount of training data)*.

- Ưu điểm: Linh hoạt hơn.
- Hạn chế: Không ước lượng được tham số với dữ liệu lớn (Computationally intractable for large datasets).

Some key concepts in statistical machine learning

Parametric model vs. non-parametric model

Parametric regression

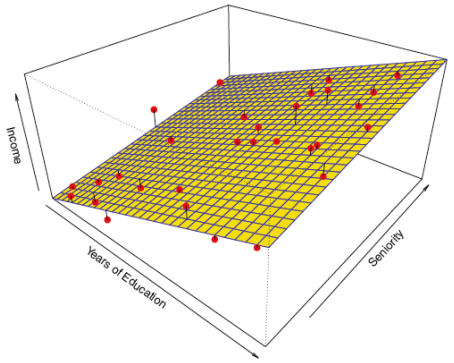
- Chọn mô hình hồi quy tuyến tính

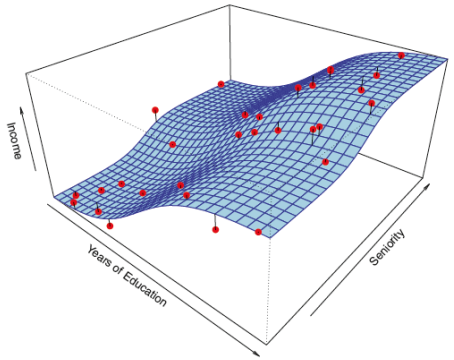
$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$

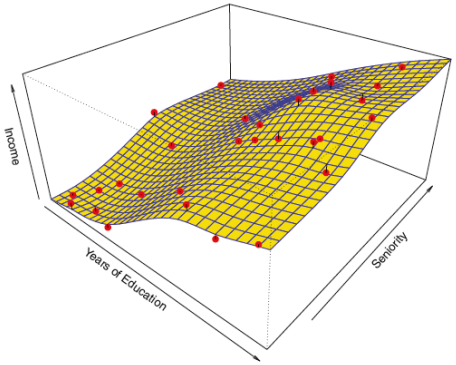
- Dùng dữ liệu huấn luyện để học mô hình.

Non-parametric regression

- Không giả thiết gì về dạng của *f*. Tìm kiếm một ước lượng của *f* gần nhất với các điểm dữ liệu nhưng không quá xù xì hoặc uốn lượn (without being too rough or wiggly).
- Dùng *thin-plate spline* để ước lượng *f* với độ mịn được chọn trước.



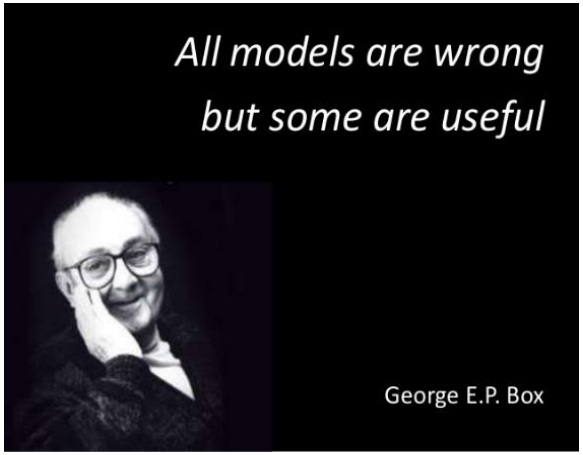




Some key concepts in statistical machine learning

Model selection

- Thí dụ các bài toán chọn lựa mô hình
  - Is it a linear or non-linear regression I should choose?
  - Which neural net architecture gives the best generalization error?
  - How many neighbors should I take in consideration in a nearest-neighbor algorithms?
  - Should I use a linear model, a decision tree, a neural net, a local learning algorithms?
  - Which of the 50 features are relevant for this problem?
- **Problem**: Chọn mô hình thích hợp nhất để phân tích một tập dữ liệu với mục tiêu định trước.
- Liên quan việc lựa chọn
  - Mô hình thích hợp
  - Tham số của mô hình



(1919-2013)

Some key concepts in statistical machine learning

Model selection

Theoretical

- Minimum description length (MDL, 1978, mô hình nén dữ liệu)
- Bayesian information criterion (BIC, 1978, mô hình tương thích khi kích thước tăng).
- Akaike information criterion (AIC, 1973, mô hình dự đoán)
- etc.

AIC

- $AIC = 2k - 2\ln(L)$
- L là giá trị cực đại của hàm likelihood của mô hình (đo sự chưa phù hợp của mô hình)
- k là số tham số cần ước lượng (penalty khi kích thước mô hình tăng, tức nhấn mạnh tính tằn tiện (parsimony)).
- Chọn mô hình với cực tiểu AIC.

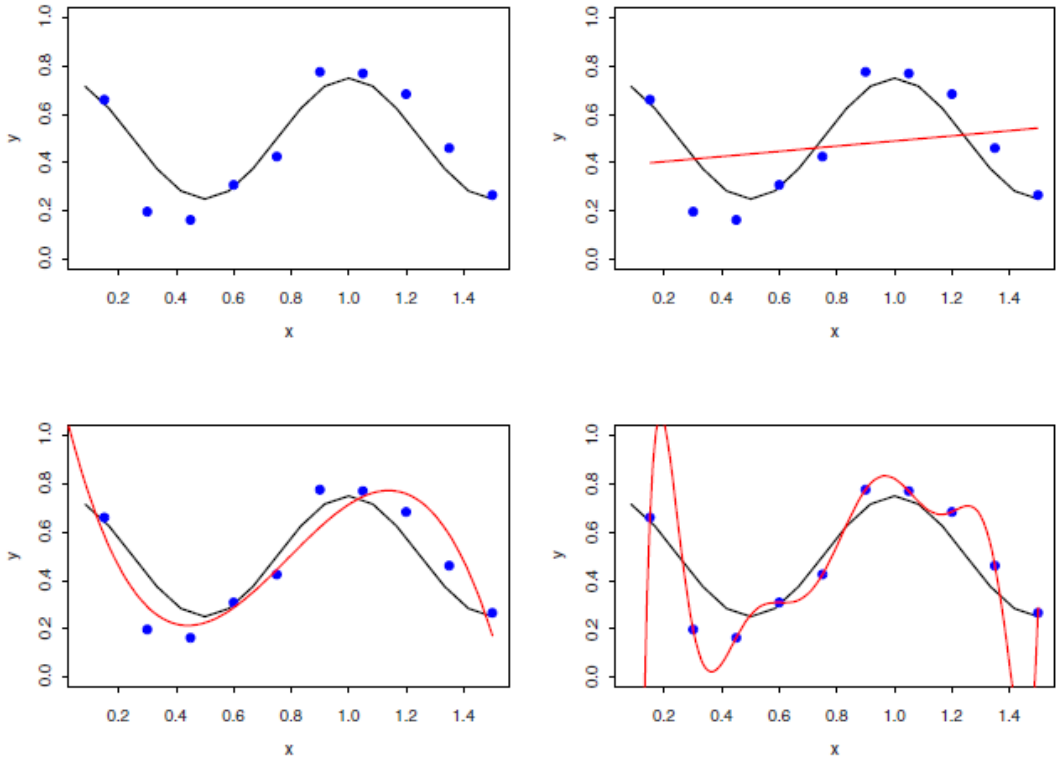
Empirical

- Boostrap (Efron 1979)
- Cross-validation (Stone 1974)
- Test set validation (holdout)
- Jacknife
- Linear regression
- Shibata's model selector (1981)
- etc.

Việc lựa chọn mô hình phải chủ yếu dựa vào tính khái quát, không phải sự phù hợp với dữ liệu.

Some key concepts and issues

Overfitting



Overfitting (quá khít) xảy ra nếu mô hình quá cồng kềnh, phức tạp, hoặc quá nhiều tham số.

"~" = "has the same distribution as"

Some key concepts in statistical machine learning

Approaches to preventing overfitting

- **Phạt (Penalty)**: Đưa vào một *đại lượng điều chỉnh* (regularization term hoặc regularizer) khi đánh giá ϵ_{test}: $\boldsymbol{\epsilon_{test} = \epsilon_{train} + penalty}$. Khi huấn luyện mô hình, ta tìm $\boldsymbol{w}$ để cực tiểu hóa hàm mục tiêu $J(\boldsymbol{w}) = \epsilon_{train}(\boldsymbol{w}) + penelty(\boldsymbol{w})$
  - MAP provides a penalty: $h_{max} = argmax_h\ p(S|h)p(h)$
  - Structural risk minimization (using PAC theory)
  - Generalized cross-validation
  - Akaike's information criterion (AIC)
- **Chia đôi, đánh giá chéo và tạo mẫu ngẫu nhiên (Holdout, cross-validation, bootstrap)** ($S = S_{train} \cup S_{test}$)
  - Xác định bằng thực nghiệm khi nào hiện tượng quá khít xảy ra
- **Hội chẩn (ensembles)**
  - Trung bình các mô hình Bayesian: Lấy mẫu các giả thiết h tương ứng với xác suất hậu nghiệm P(h|S) [e.g., Markov chain Monte Carlo MCMC]
  - Nhiều cách để tạo "hội chẩn": Bagging, random forests, etc.

MAP: Maximum a posterior

Some key concepts and issues

Regularization

- Giả sử dữ liệu huấn luyện $(\boldsymbol{x}_i, y_i)$, $i = 1, \dots, m$ theo một phân bố $p(\boldsymbol{x}, y)$, $\boldsymbol{x}_i \in \mathcal{X}$, $y_i \in \mathcal{C} = \{C_1, \dots, C_k\}$. Dự đoán y khi có các $\boldsymbol{x}$ mới nhằm tìm hàm $f: \mathcal{X} \to \mathcal{C}$ sao cho sai số nhỏ nhất.
- *Lỗi huấn luyện* (training error): Trung bình của *hàm mất mát* (loss function) trên dữ liệu huấn luyện, thí dụ

$$\epsilon_{train}[f] = \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i)), \text{ e.g., } c(x_i, y_i, f(x_i) = \begin{cases} 0, y_i = f(x_i) \\ 1, y_i \neq f(x_i) \end{cases}$$

- Bài toán dự đoán nhằm tìm hàm f đạt cực tiểu *lỗi kiểm tra* (test error)

$$\epsilon[f] = E\big[\epsilon_{test}[f]\big] = E[c\big(x, y, f(x)\big)]$$

- $\epsilon_{train}[f]$ nhỏ không đảm bảo $\epsilon[f]$ nhỏ

Some key concepts and issues

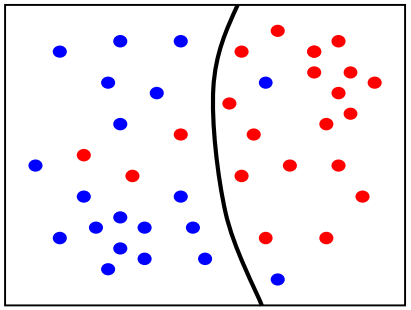
Regularization-Điều chỉnh

- *Regularization* là việc đưa một đại lượng điều chỉnh (regularizator or regularization term) vào quá trình học để ngăn cản hiện tượng quá khít

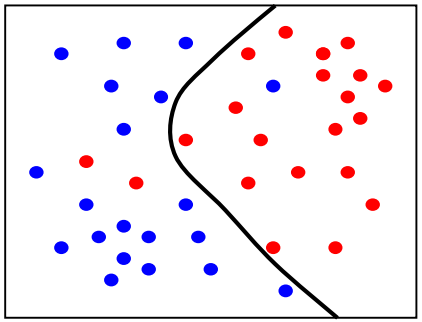
$$\epsilon[f] = \epsilon_{train}[f] + \lambda \times \text{regularizer}[f]$$

- Đại lượng điều chỉnh thường liên quan đến độ phức tạp của lời giải
  - Hạn chế về độ mịn của hàm (smoothness)
  - Giới hạn chuẩn của không gian vector.

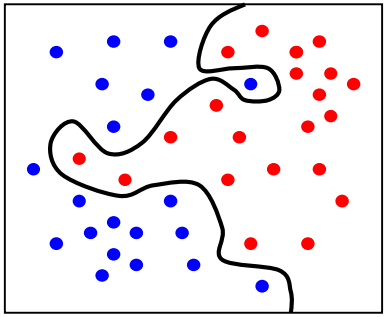
$\epsilon_{train}[f1] = 5/40$



$\epsilon_{train}[f2] = 3/40$



$\epsilon_{train}[f3] = 0$



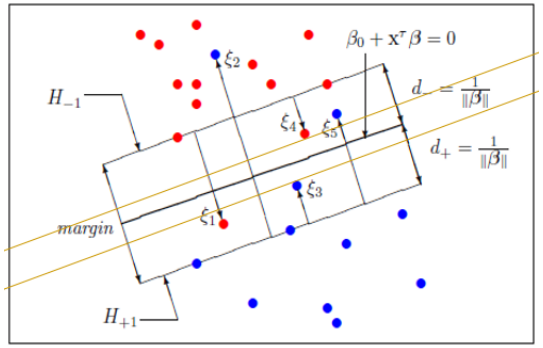
Some key concepts and issues

Regularization

- Điều chỉnh trong SVM (with slack variables)

$$\text{minimize } \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi^i$$

$$\textit{subject to } \xi_i \geq 0, y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) \geq 1 - \xi_i, i = 1..n$$



- Điều chỉnh trong k-means clustering

$$\min_{\mathcal{A}_k C_k} \sum_{k=1}^{K}\sum_{X_i \in \mathcal{A}_k} \|X_i - C_k\|^2 + \sum_{j=1}^{p} J(C_{(j)})$$

$C_{(j)} = (C_{1j}, \ldots, C_{Kj})^T$ and C_{kj} is the jth element of C_k, $J(C_{(j)}) = \lambda_j \|C_{(j)}\|$

Sun, W. (2012) Regularized k-means clustering of high-dimensional data, Journal of Statistics, Vol. 6.

Mô hình thưa

Sparse modeling

- **Mô hình thưa**: *Có một số tham số hay trọng số (weights) khác zero.*
  Less is more: *ước lượng và giải thích dễ hơn mô hình dày (dense model).*

- Cho N mẫu $\{(x_i, y_i)\}_{i=1}^{N}$, với các biến mô tả $x_i = (x_{i1}, \dots, x_{ip})$ và biến đích $y_i \in \mathbb{R}$. Ta xấp xỉ y_i bởi hồi quy tuyến tính với $\beta = (\beta_1, \dots, \beta_p)$ và $\beta_0 \in \mathbb{R}$

$$f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

- Dùng bình phương tối thiểu, ta muốn cực tiểu hàm mất mát bình phương

$$\underset{\beta_0, \beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \right\}$$

- Có nhiều lý do để xem xét một cách khác: Tăng độ chính xác bởi co (shrinking) các hệ số về zero, và giảm bớt hệ số để dễ giải thích hơn. Một cách làm phổ biến là hạn chế p-norm của β.

Sparse learning

Lasso regression

- *Lasso regression* dùng L1-norm tìm $(\hat{\beta}_0, \hat{\beta})$, với ràng buộc $\|\beta\|_1 \leq t$ có thể được viết như $\sum_{j=1}^{p}|\beta_j| \leq t$

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right\}$$
$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \leq t$$

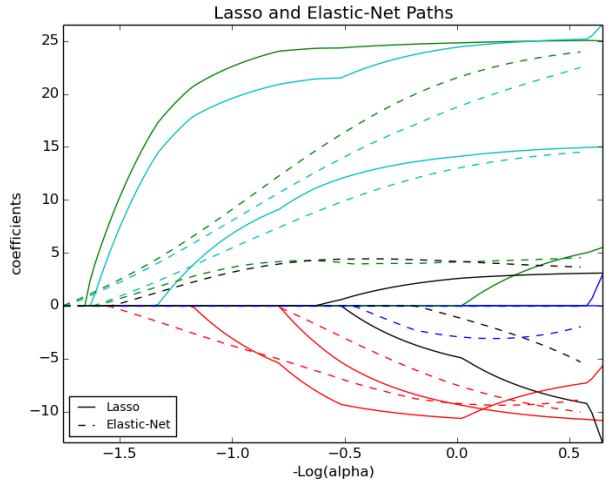
- *Ridge regression* dùng L2-norm

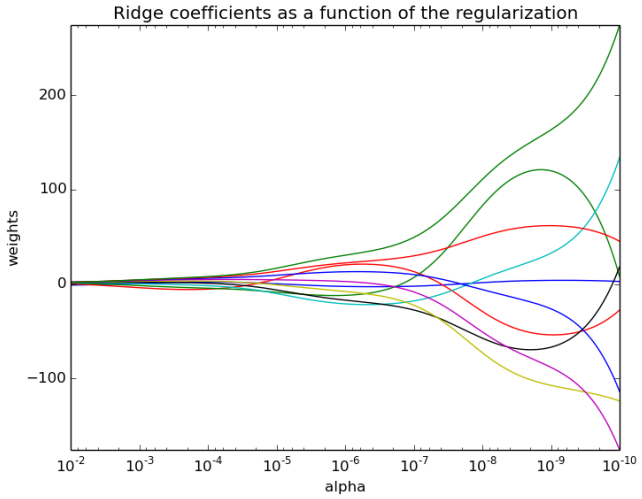
$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right\}$$
$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t^2$$

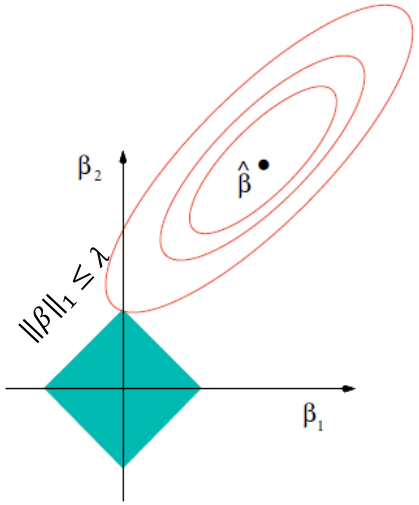
(2) can be solved by a simple coordinate decent algorithm

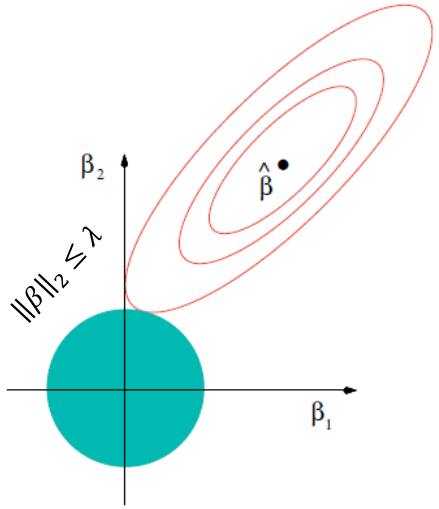
Sparse learning

Lasso regression









Hình ảnh về ước lượng của hồi quy lasso (trái) và ridge (phải). Vùng màu xanh và vùng ràng buộc $|\beta_1| + |\beta_2| \leq t$ và $\beta_1^2 + \beta_2^2 \leq t^2$, và các đường ellipse màu đỏ là đường viền của hàm residual-sum-of-squares. Điểm $\hat{\beta}$ biểu diễn ước lượng bình phương tối thiểu thông thường unconstrained).

Sparse learning

L_q penalties

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- Với q < 1, có rất nhiều hệ số zero nhưng miền ràng buộc là nonconvex. Với q > 1, miền ràng buộc là convex nhưng mọi không có hệ số zero. Chỉ riêng tại q = 1, tính lồi đã gặp tính thưa (sparvexity). The Lasso for $q = 1$ and ridge regression for $q = 2$.

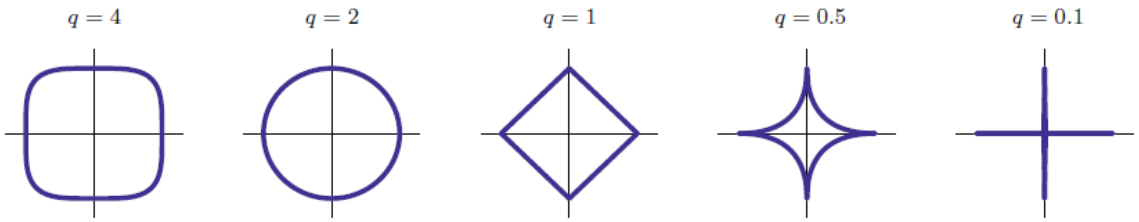
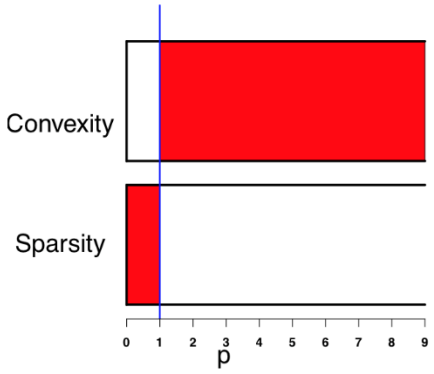
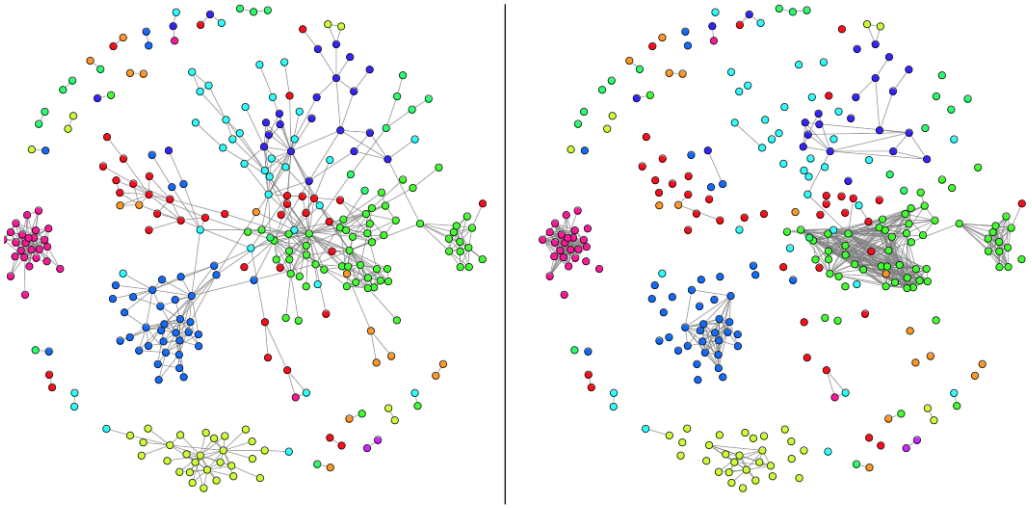


Figure 2.6 *Constraint regions* $\sum_{j=1}^{p} |\beta_j|^q \leq 1$ *for different values of* q. *For* $q < 1$, *the constraint region is nonconvex.*

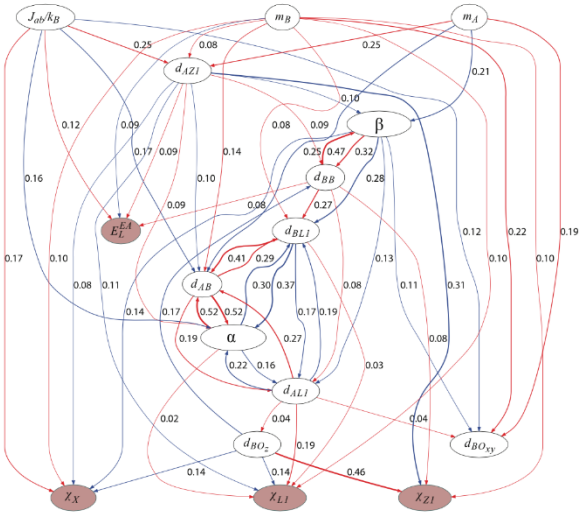


Sparse learning

Graphical Lasso and Parallel Lasso



S&P 500



Using Lasso in study of new material design

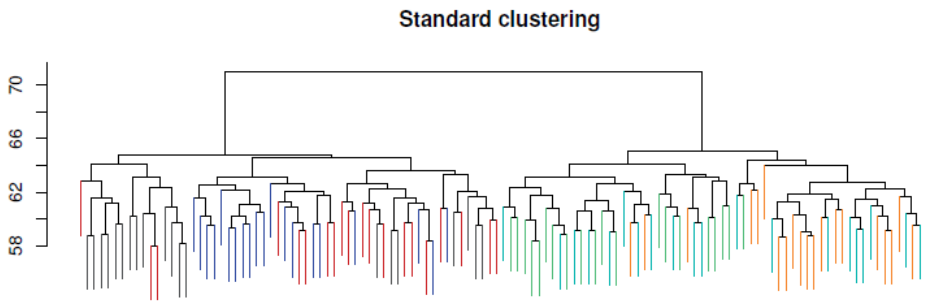
Dam, H.C., Pham, T.L., Ho, T.B., Nguyen, T.A., Nguyen, V.C. (2014). Data mining for materials design: A computational study of single molecule magnet, *The journal of Chemical Physics* Vol. 140, Issue 4, 28 January 2014

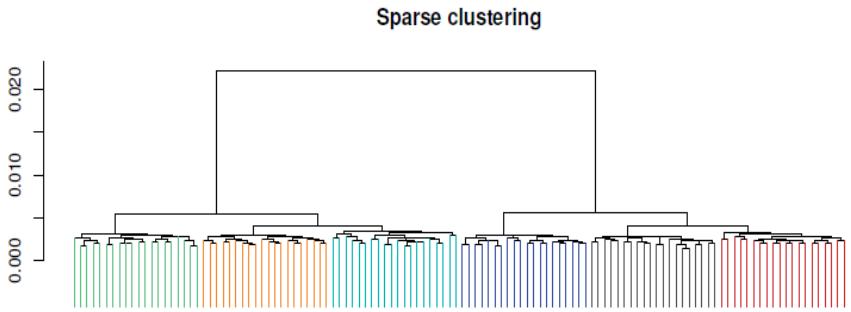
Phân tích thưa dữ liệu nhiều biến

Sparse multivariate methods

- Ma trận dữ liệu **X** với số chiều $N \times p$. Các thành phần chính của **X** nhận được từ phân tích giá trị đặc biệt (singular value decomposition) $\mathbf{X} = \mathbf{UDV}^T$.
- Ta có thể rút ra các thành phần chính *thưa* (sparse principal components) khi áp dụng phân tích ma trận có phạt cho **X** với ép buộc (enforced) tính thưa trên các biến.

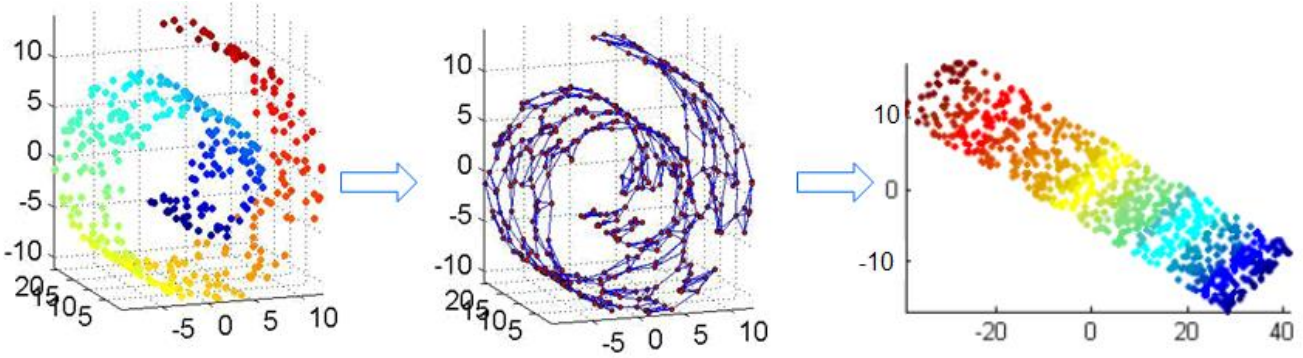
Input Matrix	Result
Data matrix	sparse SVD and principal components
Variance-covariance	sparse principal components
Cross-products	sparse canonical variates
Dissimilarity	sparse clustering
Between-class covariance	sparse linear discriminants





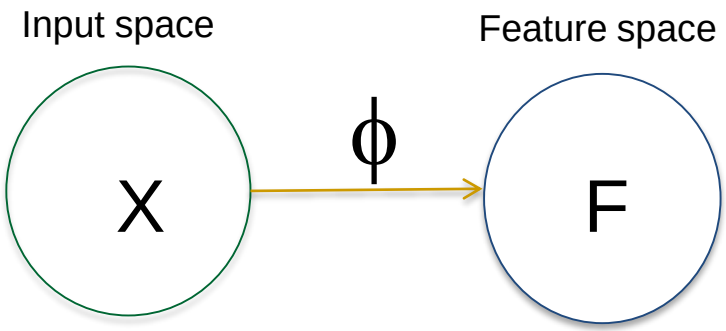
Dimensionality reduction

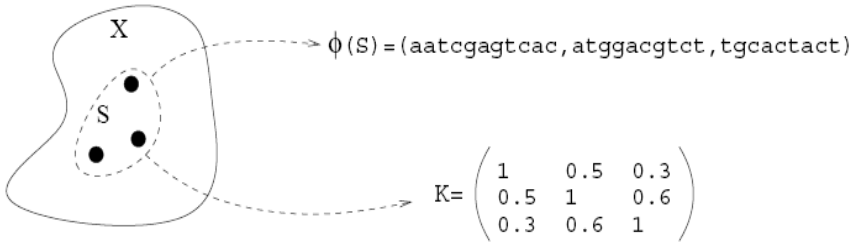
- Mặc dù dữ liệu được thu thập với nhiều chiều (biến), *số chiều thật sự* (intrinsic dimension) của dữ liệu ở nhiều ứng dụng có thể nhỏ hơn nhiều.
- Tập dữ liệu $\boldsymbol{X} \subset \mathbb{R}^m$ có *số chiều thật sự* là $p \leq m$, nếu $\boldsymbol{X}$ có thể được biểu diễn (xấp xỉ) bởi m tham số tự do.
- Rút gọn số chiều (dimensionality reduction) là việc tìm số chiều thật sự của một tập dữ liệu $\boldsymbol{X}$, gồm các phương pháp *lựa chọn* biến (feature selection) *và tạo biến mới* (feature extraction).



Dimensionality reduction

Transformation



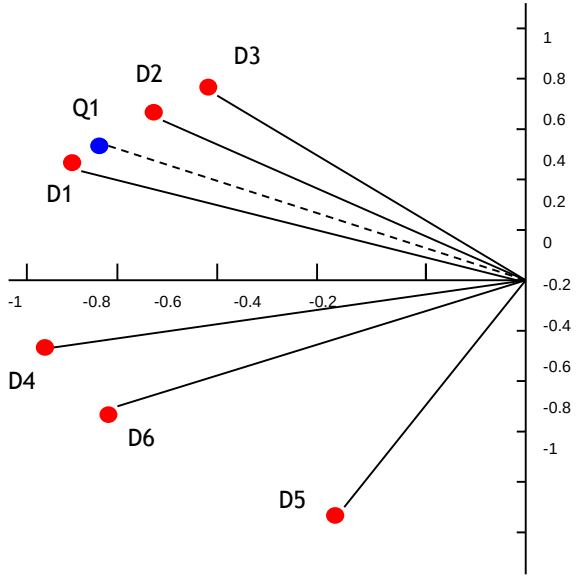


ϕ: X → F where the problem can be solved in F

- PCA: Principle component analysis
- CCA: Canonical correlation analysis
- ICA: Independent component analysis
- NMF: Nonnegative matrix factorization
- Nonlinear dimensionality reduction: Kernel PCA, ISOMAP, LLE, etc.

Phân tích ngữ nghĩa ẩn

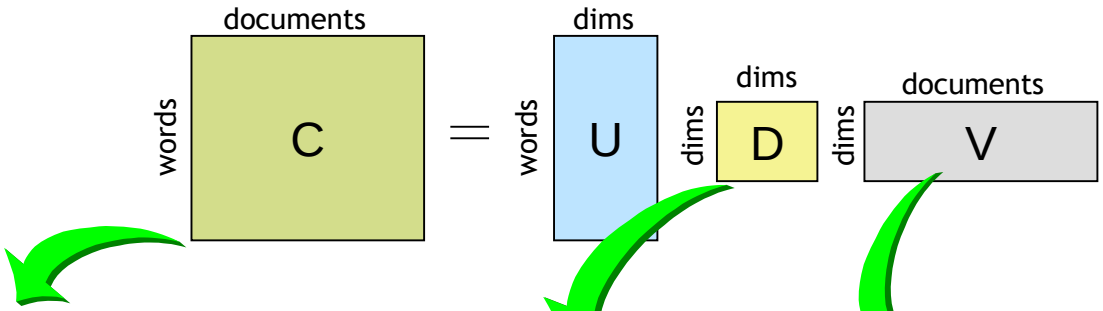
Latent semantic analysis (LSA)



LSA (Deerwester, 1990) phân nhóm các văn bản trong một không gian ngữ nghĩa thu gọn tương ứng với các mẫu dạng từ (word patterns) cùng xuất hiện.

$$\cos(x, y) = \frac{x.y}{|x||y|}$$

cos(d3, q1) = 0
cos(d5, q1) = 0
cos(d4, q1) ≠ 0
cos(d6, q1) ≠ 0



	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

	D1	D2	D3	D4	D5	D6	Q1
Dim. 1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
Dim. 2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

Topic model

Key idea

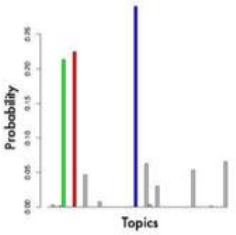
Molecular Classification of Cancer Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M .L. Loh, J .R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander

Most probable terms from the top topics

data	cancer	information
values	gene	computer
rate	tumor	system
model	tumors	problem
fig	mutations	systems
average	genes	approach
table	breast	problems
value	repair	methods
mean	human	models
time	mutation	simple

Expected topic proportions

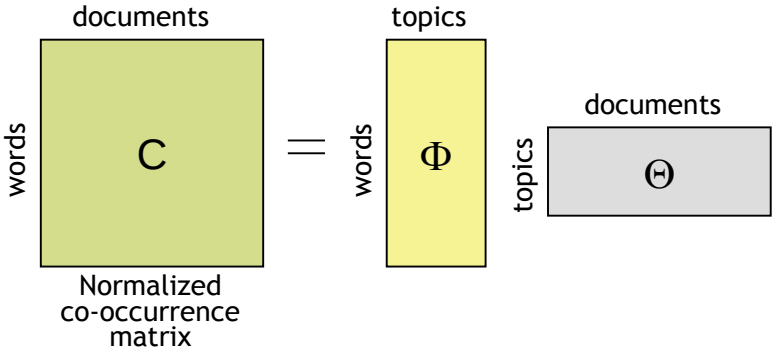


Article abstract marked with its most likely topic assignments

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

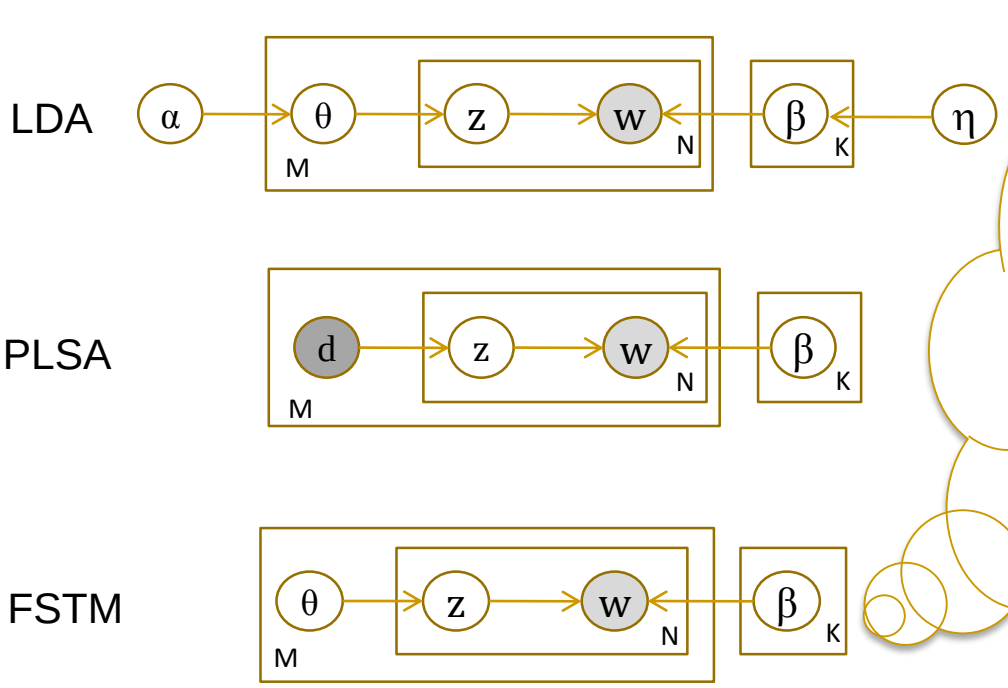
- Một chủ đề là một phân bố của các từ cùng xuất hiện (a topic is a probability distribution over words.)
- Một văn bản là một pha trộn của các chủ đề ẩn (a documents are mixtures of latent topics).

Topic models



Two problems of learning and inference

Fully Sparse Topic Model



- FSTM assumes no explicit prior over topics (β).
- FSTM assumes no explicit prior over topic mixtures (θ) (MAP inference ~ FW inference).
- It assumes a corpus to be composed of K topics, $\beta_1, \ldots, \beta_K$

FSTM = LDA without Dirichlet prior + PLSA with sparsity enforced + Frank-Wolfe algorithm for inference.

Khoat Than & Tu Bao Ho (ECML 2012; IDA 2014, Neurocomputing 2014).
Inference in topic models I: sparsity and trade-off (2015)

FSTM: Large-scale learning

- A machine with 128 CPUs is used, each with 2.9GHz, grouped into 32 clusters each having 4 CPUs
- Webspam with 350,000 documents, 16 millions of dimensions.
- Number of topics: 2000
- #Latent variables for dense models: > 33 billions (>130 Gb in memory)

#Topics	1000	2000
Time per EM iteration	28 minutes	65 minutes
#EM iterations to reach convergence	17	16
Topic sparsity (compared with dense models)	0.0165 (60 times smaller)	0.0114 (87 times smaller)
Document sparsity (compared with dense models)	0.0054 (185 times smaller)	0.0028 (357 times smaller)
Storage for the new representation (compared with the original corpus)	31.5 Mb (757 times smaller)	33.2 Mb (718 times smaller)

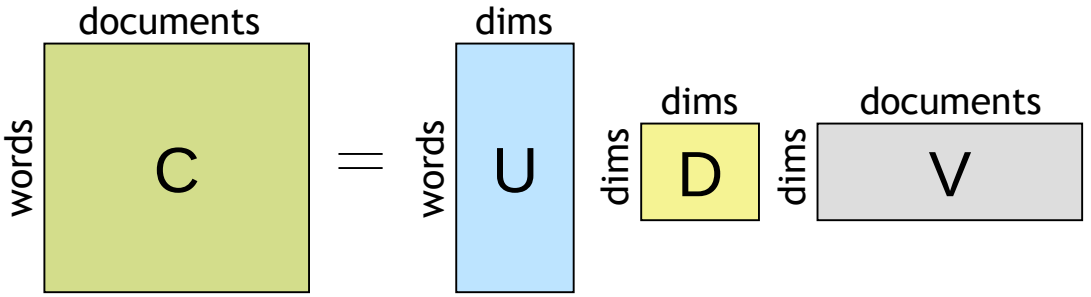
Data	#documents	#dimensions	Storage	Best known Accuracy	Classified by	Repetitions
Original Webspam	350,000	16,609,143	23.3 Gb	99.15%	BMD [Yu et al. 2012]	1
Represented by FSTM						
1000 topics	350,000	1000	31.5 Mb	98.877%	FSTM + Liblinear	5
2000 topics	350,000	2000	33.2 Mb	99.146%	FSTM + Liblinear	5

Các phép biến đổi dưới dạng tích ma trận

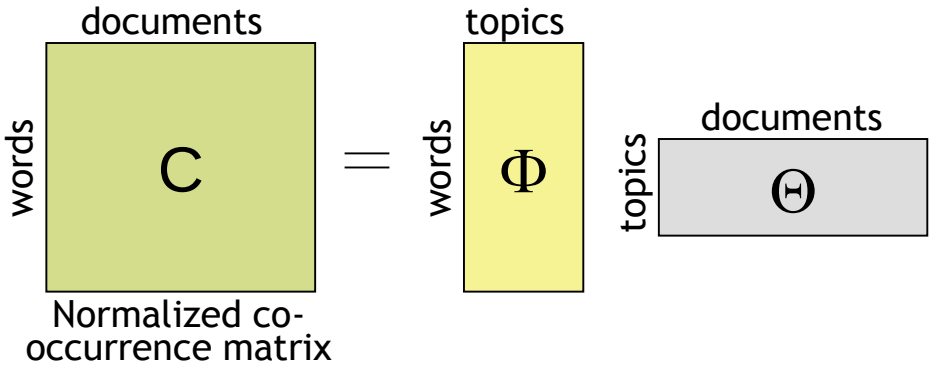
Transformations in form of matrix factorization

Most well known transformation methods can be represented as matrix factorization

Latent semantic analysis



Topic models



Phân tích ma trận không âm

Nonnegative matrix factorization (NMF)

- Ma trận dữ liệu $\boldsymbol{X} = \{x_1, x_2, \dots, x_m\}$ biểu diễn m đối tượng trong không gian $\mathbb{R}^n$. NMF phân tích $\boldsymbol{X}$ thành tích các ma trận không âm

$$\boldsymbol{X \approx FG}$$

  - $F, G \geq 0$ h
  - $F = \{f_1, f_2, \dots, f_m\} \subseteq \mathbb{R}^k$ là biểu diễn mới (dữ liệu) của m đối tượng
  - $G = \{g_1, g_2, \dots, g_k\} \subseteq \mathbb{R}^n$ là các thành phần ẩn (cơ sở) của không gian mới

- NMF thực hiện một phép biến đổi $\boldsymbol{\mathbb{R}^n \longrightarrow \mathbb{R}^k, k \ll n}$.

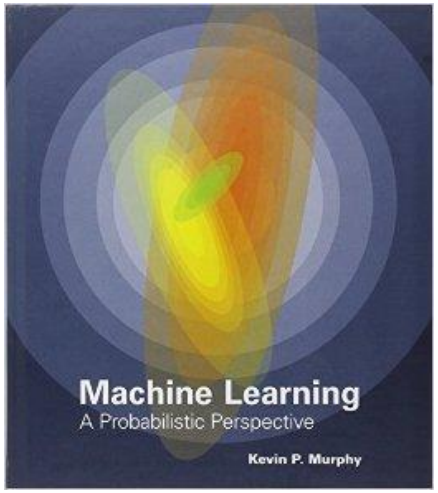
- Chất lượng NMF được đánh giá bởi lượng thông tin không bị mất sau biến đổi, thường đo bằng hàm mục tiêu dưới dạng Frobenius $D = (X \parallel FG) = \|X - FG\|_2^2$.

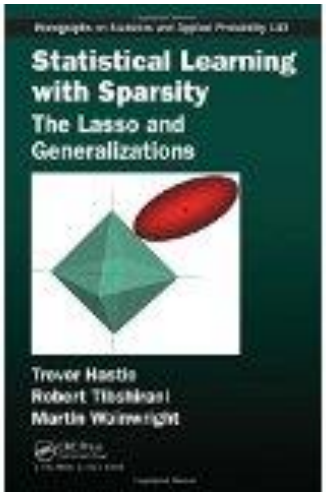
Take home message

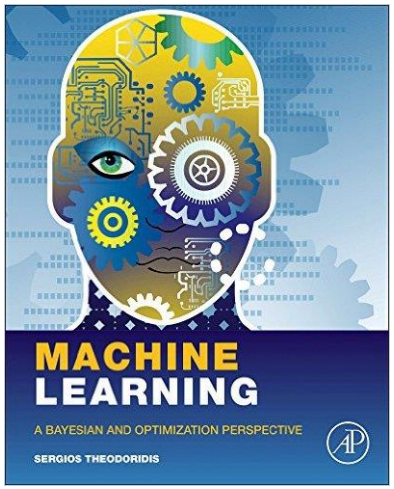
- Khoa học phân tích dữ liệu (lớn) dựa trên phương pháp và công cụ của thống kê, khai phá dữ liệu và học máy.
- Học máy thống kê đang thay đổi rất nhanh, đòi hỏi việc học tập phải theo đuổi liên tục và kiên trì, dựa trên những nền tảng cơ bản tốt.
- Nghĩ đến các bài toán và thách thức của học máy thống kê trong lĩnh vực chuyên biệt.
- Mô hình thưa và rút gọn số chiều là những thách thức cơ bản để đối đầu với phân tích dữ liệu lớn.

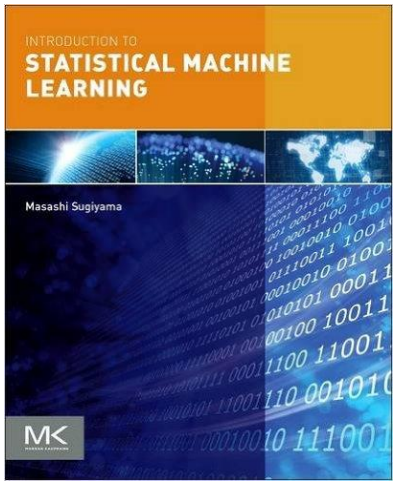
Additional slides

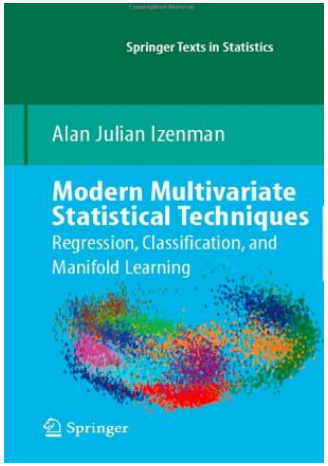
Some typical books

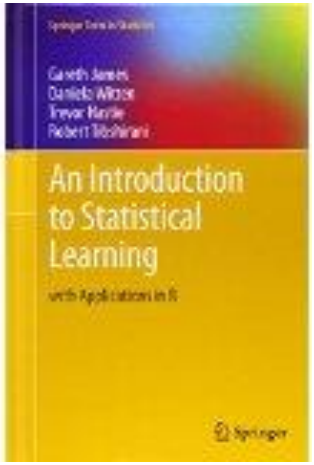


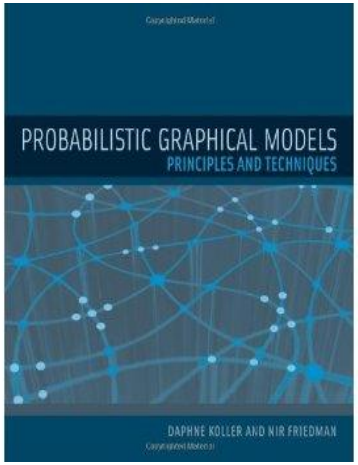


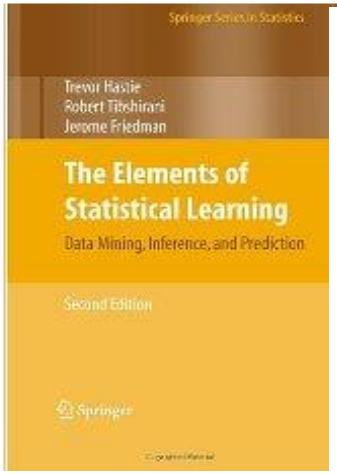


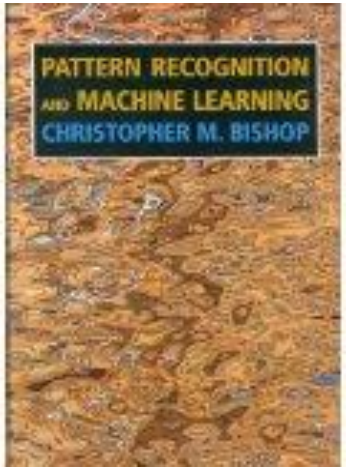






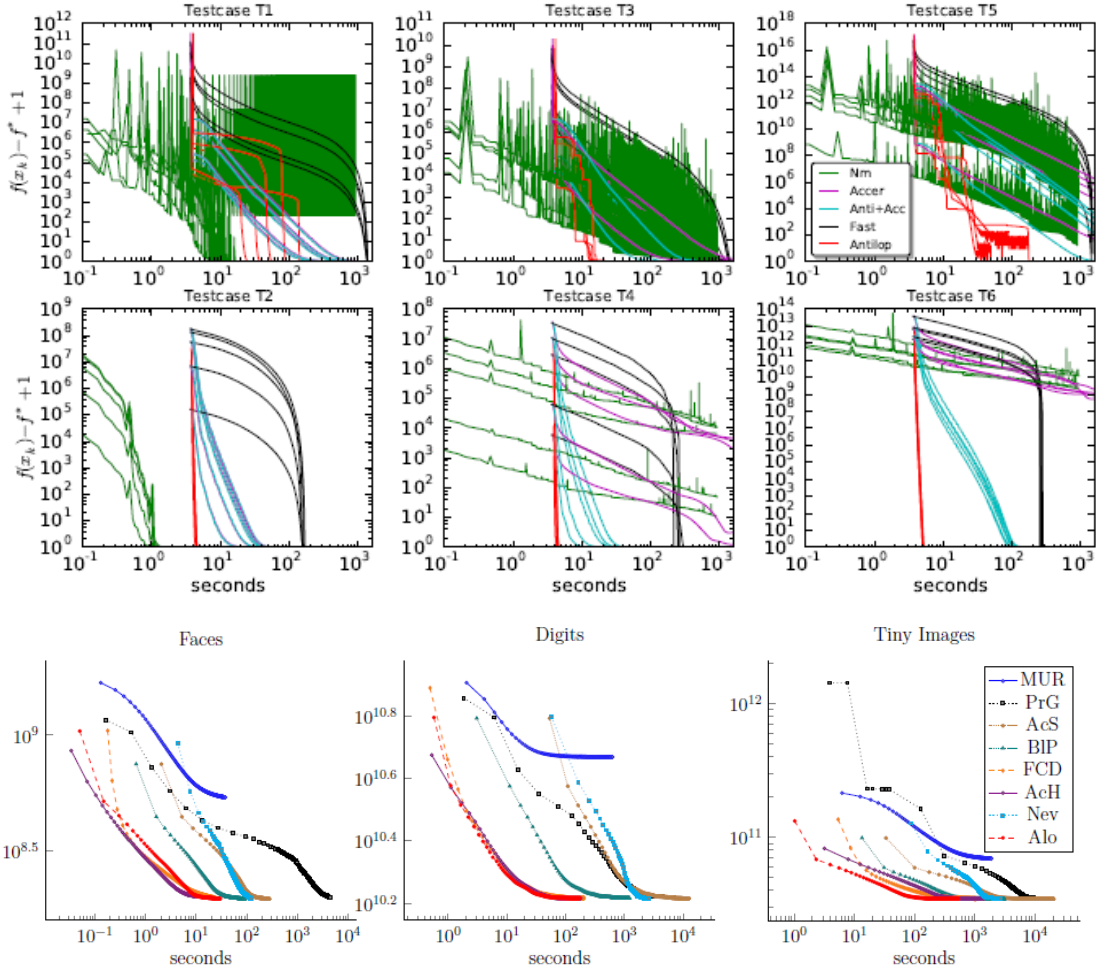






Simplicial nonnegative matrix factorization

- **Problem 1**: Anti-lopsided algorithm for large-scale non-negative least squares.
- **Problem 2**: Fast accelerated parallel and distributed algorithm using limited internal memory for NMF.
- **Problem 3**: Simplicial NMF: Model and accelerated parallel algorithm.



Nguyen DK., Ho TB. (2015) Anti-lopsided Algorithm for Large-scale Nonnegative Least Square Problems, *arXiv*

Nguyen DK., Ho TB. (2015) Accelerated Parallel and Distributed Algorithm Using Limited Internal Memory for Nonnegative Matrix Factorization

Dữ liệu lớn đến từ đâu?

- Từ các phương tiện xã hội
  Nhìn thấu (insights) được hành vi và ý kiến của khách hàng của công ty.
- Từ máy móc
  Thiết bị công nghiệp, các sensors và dụng cụ giám sát, web logs…
- Từ giao dịch kinh doanh
  ID và giá cả sản phẩm, thanh toán, dữ liệu chế tạo và phân bố, … ,
- Nhiều loại khác

Each day:
230M tweets,
2.7B comments to FB,
86400 hours of video to YouTube

Large Hadron Collider generates 40 terabytes/sec

Amazon.com: $10B in sales in Q3 2011, US pizza chain Domino's: 1 million customers per day

Khoa học phân tích dữ liệu là gì?

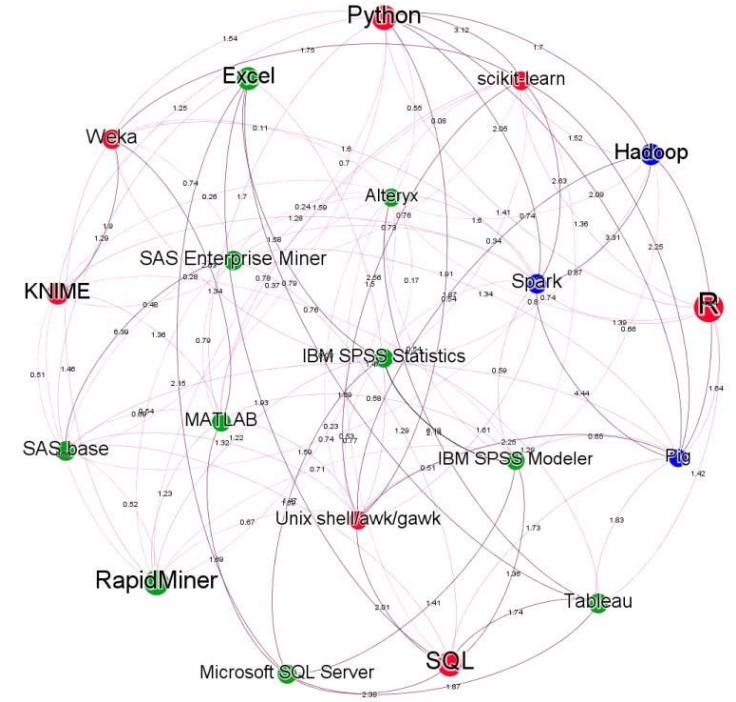
What are Data Analytics?

- … Khoa học về phân tích dữ liệu thô nhằm rút ra các kết luận

  data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information.
  http://searchdatamanagement.techtarget.com/definition/data-analytics

- **Big data analytics** là khoa học về quá trình phân tích dữ liệu lớn để phát hiện ra các mẫu dạng, các quan hệ và thông tin hữu ích để ra quyết định tốt hơn…

  is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions.
  http://www.sas.com/en_us/insights/analytics/big-data-analytics.html



Tóm tắt

- Thống kê, học máy và khai phá dữ liệu ra đời từ những thời điểm khác nhau, có động lực và nội dung nhiều phần khác nhau.
- Ba lĩnh vực này đang xích lại gần nhau, và các phương pháp của ba lĩnh vực này cho phép ta nhiều lựa chọn hơn cho các giải pháp thích hợp trong data analytics.

Which algorithms perform best at which tasks?

Algorithm	Pros	Cons	Good at
Linear regression	- Very fast (runs in constant time) - Easy to understand the model - Less prone to overfitting	- Unable to model complex relationships - Unable to capture nonlinear relationships without first transforming the inputs	- The first look at a dataset - Numerical data with lots of features
Decision trees	- Fast - Robust to noise and missing values - Accurate	- Complex trees are hard to interpret - Duplication within the same sub-tree is possible	- Star classification - Medical diagnosis - Credit risk analysis
Neural networks	- Extremely powerful - Can model even very complex relationships - No need to understand the underlying data - Almost works by “magic”	- Prone to overfitting - Long training time - Requires significant computing power for large datasets - Model is essentially unreadable	- Images - Video - “Human-intelligence” type tasks like driving or flying - Robotics
Support Vector Machines	- Can model complex, nonlinear relationships - Robust to noise (because they maximize margins)	- Need to select a good kernel function - Model parameters are difficult to interpret - Sometimes numerical stability problems - Requires significant memory and processing power	- Classifying proteins - Text classification - Image classification - Handwriting recognition
K-Nearest Neighbors	- Simple - Powerful - No training involved (“lazy”) - Naturally handles multiclass classification and regression	- Expensive and slow to predict new instances - Must define a meaningful distance function - Performs poorly on high-dimensionality datasets	- Low-dimensional datasets - Computer security: intrusion detection - Fault detection in semi-conducter manufacturing - Video content retrieval - Gene expression - Protein-protein interaction

http://www.lauradhamilton.com/machine-learning-algorithm-cheat-sheet

Key concepts in statistical machine learning

Non-parametric density estimation (NPDE)

- Ước lượng hàm mật độ xác suất pdf p nhưng không định sẵn dạng hàm, thỏa mãn

$$p(x) \geq 0, \int_{\mathbb{R}} p(x)dx = 1$$

- Xem R là một vùng nhỏ chứa x và cho N điểm dữ liệu. Mỗi điểm có xác suất P rơi vào R, và toàn bộ K điểm nằm trong R tuân theo luật phân phối nhị phân, ta có $P \approx p(x)V$ với V là kích thước của R. Ước lượng của hàm mật độ có dạng

$$\boldsymbol{p(x) = \frac{K}{NV}}$$ (*) → Dẫn đến hai phương pháp

kernel density estimation (V)

$$\hat{p}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), x \in R^k, \mathrm{h} > 0;$$

k-nearest neighbor (k = K)

$$\hat{p}(x) = \frac{k}{2Nd_k(x)}$$

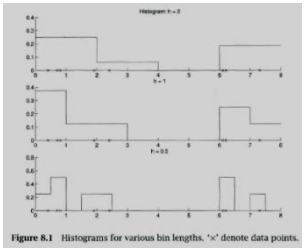


Figure 8.1 Histograms for various bin lengths. '×' denote data points.

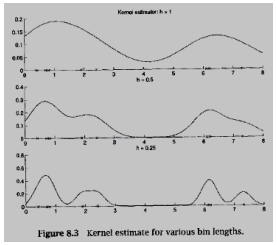


Figure 8.3 Kernel estimate for various bin lengths.

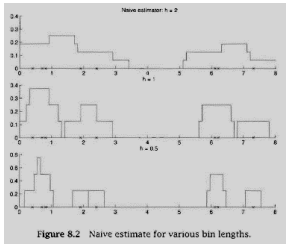


Figure 8.2 Naive estimate for various bin lengths.

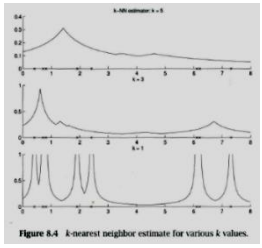
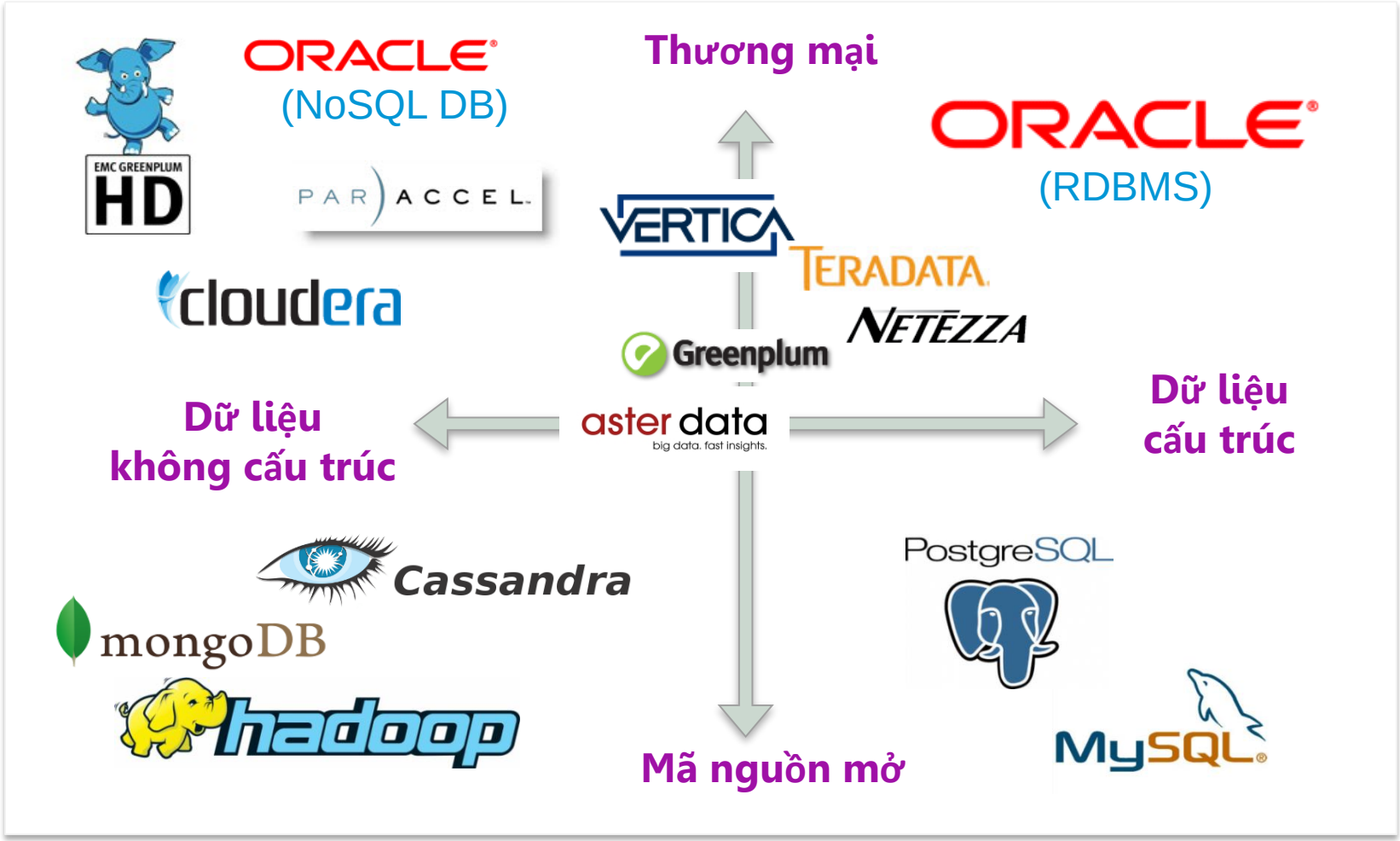


Figure 8.4 *k*-nearest neighbor estimate for various *k* values.

Quản lý dữ liệu lớn

Big data management



Source: Cisco

A dataset often used in machine learning courses

Days	Outlook	Temperature	Humidity	Wind	Class
D1	sunny	hot	high	weak	N
D2	sunny	hot	high	strong	N
D3	overcast	hot	high	weak	Y
D4	rain	mild	high	weak	Y
D5	rain	cool	normal	weak	Y
D6	rain	cool	normal	strong	N
D7	overcast	cool	normal	strong	Y
D8	sunny	mild	high	weak	N
D9	sunny	cool	normal	weak	Y
D10	rain	mild	normal	weak	Y
D11	sunny	mild	normal	strong	Y
D12	overcast	mild	high	strong	Y
D13	overcast	hot	normal	weak	Y
D14	rain	mild	high	strong	N

P(Y) = 9/14
P(N) = 5/14

outlook	
P(sunny | Y) = 2/9	P(sunny | N) = 3/5
P(overcast | Y) = 4/9	P(overcast | N) = 0
P(rain | Y) = 3/9	P(rain | N) = 2/5
temperature	
P(hot | Y) = 2/9	P(hot | N) = 2/5
P(mild | Y) = 4/9	P(mild | N) = 2/5
P(cool | Y) = 3/9	P(cool | N) = 1/5
humidity	
P(high | Y) = 3/9	P(high | N) = 4/5
P(normal | Y) = 6/9	P(normal | N) = 1/5
windy	
P(strong | Y) = 3/9	P(strong | N) = 3/5
P(weak | Y) = 6/9	P(weak | N) = 2/5