

# Khoa học Dữ liệu và Cách mạng Công nghiệp lần thứ Tư

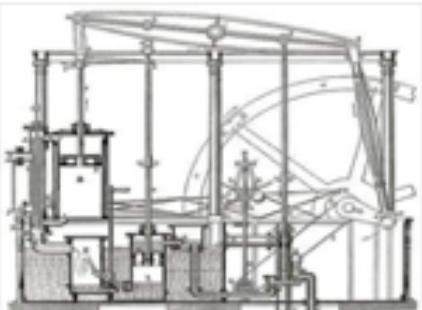
# Outline

- Cách mạng công nghiệp lần thứ tư
- Khoa học dữ liệu là gì?
- Nguyên lý và phương pháp của khoa học dữ liệu

# Cách mạng công nghiệp lần thứ tư?

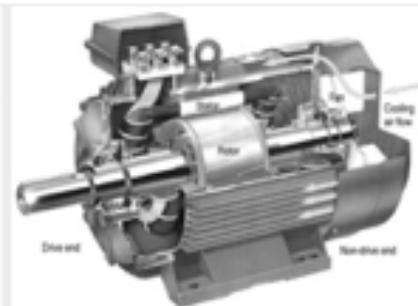
Đặc trưng của một cuộc cách mạng công nghiệp:

- Có **đột phá** của khoa học và công nghệ
- Tạo ra sự thay đổi về **bản chất** của sản xuất



Cách mạng công nghiệp lần thứ nhất về **sản xuất cơ khí** với máy móc dựa vào **động cơ hơi nước**.

Cuối thế kỷ 18 đầu thế kỷ 19



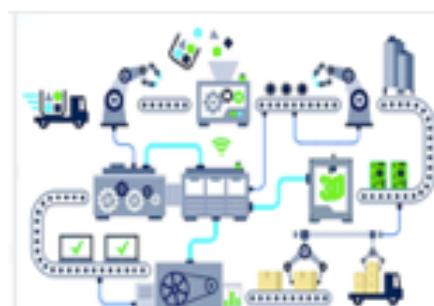
Cách mạng công nghiệp lần thứ hai về **sản xuất hàng loạt** với máy móc dựa vào **năng lượng điện**.

Cuối thế kỷ 19 đầu thế kỷ 20



Cách mạng công nghiệp lần thứ ba về **sản xuất tự động** với **máy tính, điện tử** và **cách mạng số hóa**.

Từ thập kỷ 70 của thế kỷ 20



Cách mạng công nghiệp lần thứ tư về **sản xuất thông minh** nhờ các **đột phá** của **công nghệ số**.

Bắt đầu từ bây giờ

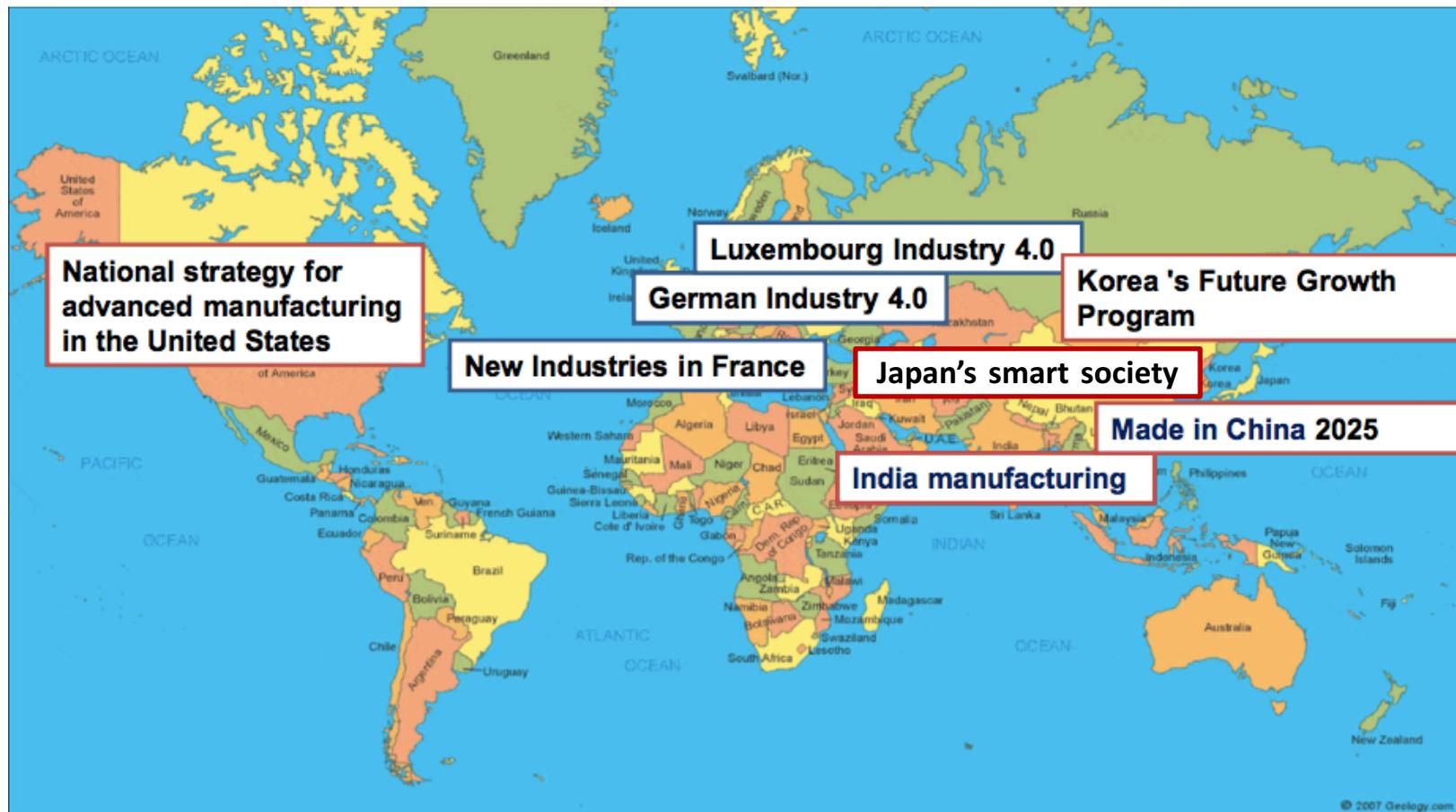
# Cách mạng công nghiệp lần thứ tư?

Đặc trưng của một cuộc cách mạng công nghiệp:

- Có **đột phá** của khoa học và công nghệ
- Tạo ra sự thay đổi về bản chất của sản xuất

... **sản xuất thông minh** dựa trên tiến bộ của công nghệ thông tin, công nghệ sinh học, công nghệ nano... với nền tảng là các **đột phá của công nghệ số** trên **cyber-physical systems**.

# Chiến lược của các nước phát triển



Klaus Schwab (WEF), The Fourth Industrial Revolution

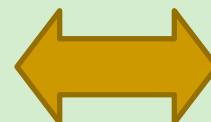
Alistair Nolan (OECD), Enabling the Next Production Revolution: Implications for Policy, Hanoi, 12.2016

# Cách mạng số hoá và cyber-physical systems

- ‘Phiên bản số’ các thực thể: Biểu diễn các thực thể bằng ‘0’ và ‘1’ trên máy tính (digitalization)
- Thí dụ: ô-tô, bệnh án điện tử...
- **Hệ kết nối không gian số-thực thể (cyber-physical system):** hệ kết nối các thực thể và ‘phiên bản số’ của chúng.



Hành động trong  
thế giới các thực thể

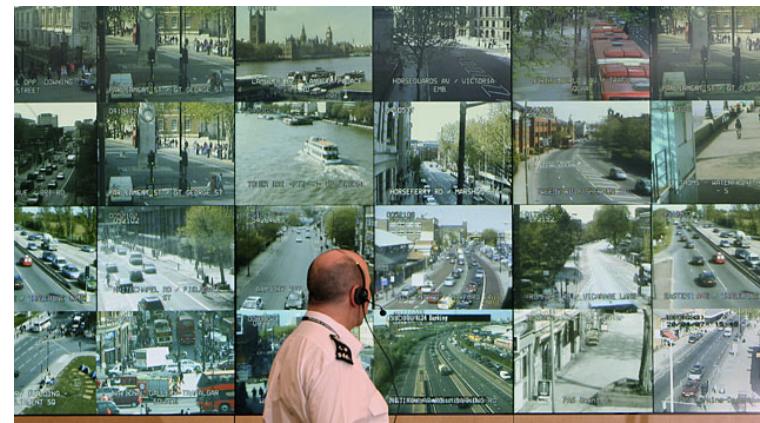
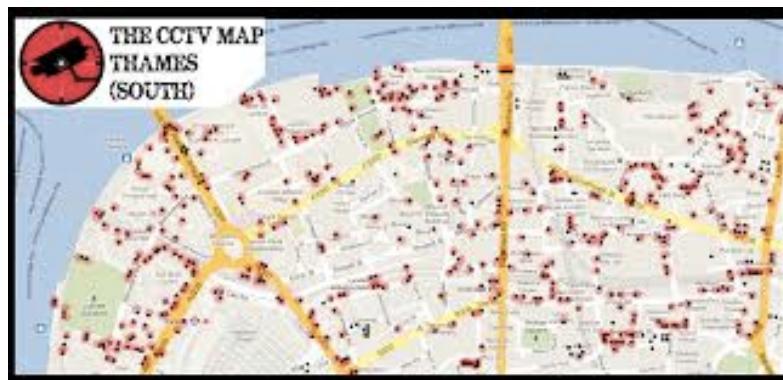


Tính toán, điều khiển  
trên không gian số

**Thay đổi phương thức sản xuất**

# London CCTV (Closed circuit TV)

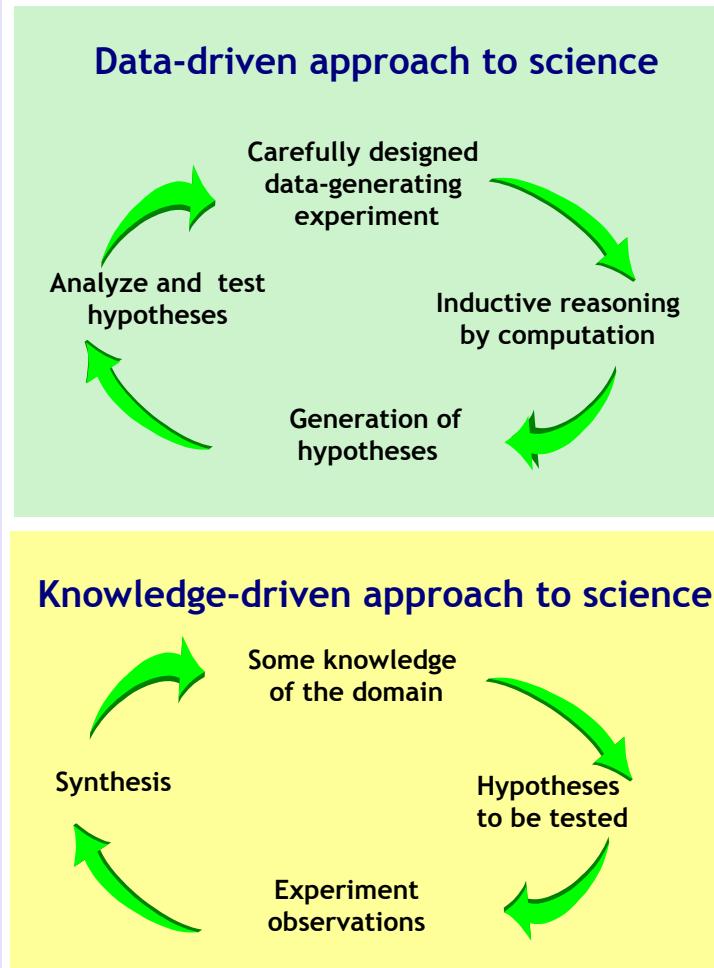
- 500 triệu bảng (video surveillance)
- Cung cấp 95% thông tin về các vụ phạm tội
- Bắt nhầm người → Hàn quốc: từ 60m, 45° nghiêng



# Data-intensive science: a shift in science

Làm khoa học dựa vào dữ liệu, nhằm tìm tri thức từ dữ liệu.

Cách truyền thống nhằm kiểm chứng các giả thiết có được từ trên tri thức đã biết.

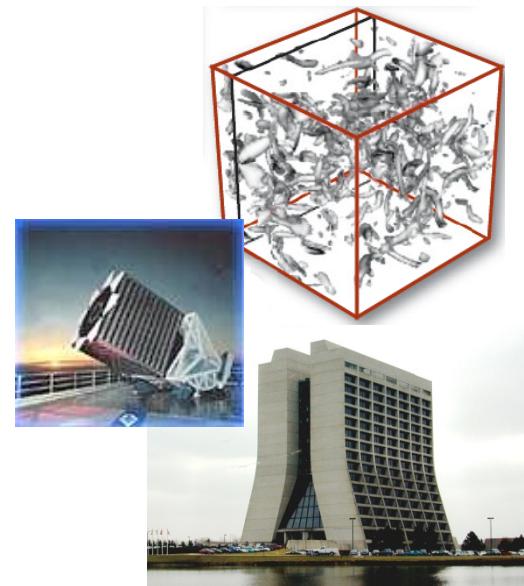


# Science paradigms

- Thousand years ago:  
science was **empirical**  
*Describing natural phenomena*
- Last few hundred years:  
**theoretical branch**  
*Using models, generalizations*
- Last few decades:  
**a computational branch**  
*Simulating complex phenomena*
- Today: **Data exploration** (eScience)  
*Unify theory, experiment, and simulation*
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes databases/files using data management and statistics.

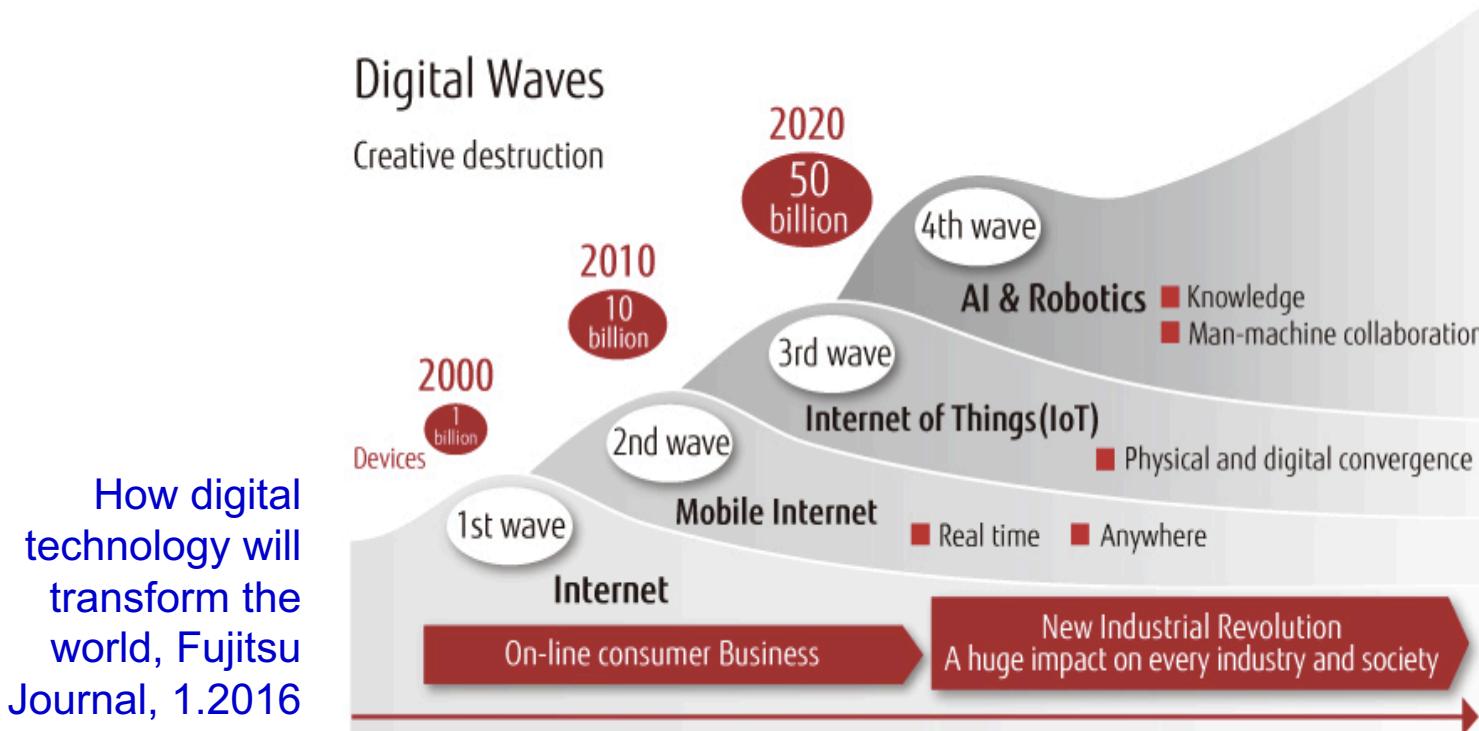


$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K \frac{c^2}{a^2}$$

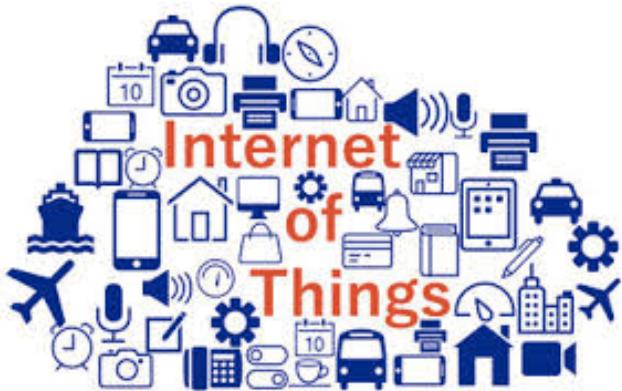


# Công nghệ số (digital technology)

- Số hoá (thí dụ máy ảnh, in ấn, truyền hình...)
- Xử lý dữ liệu được số hoá

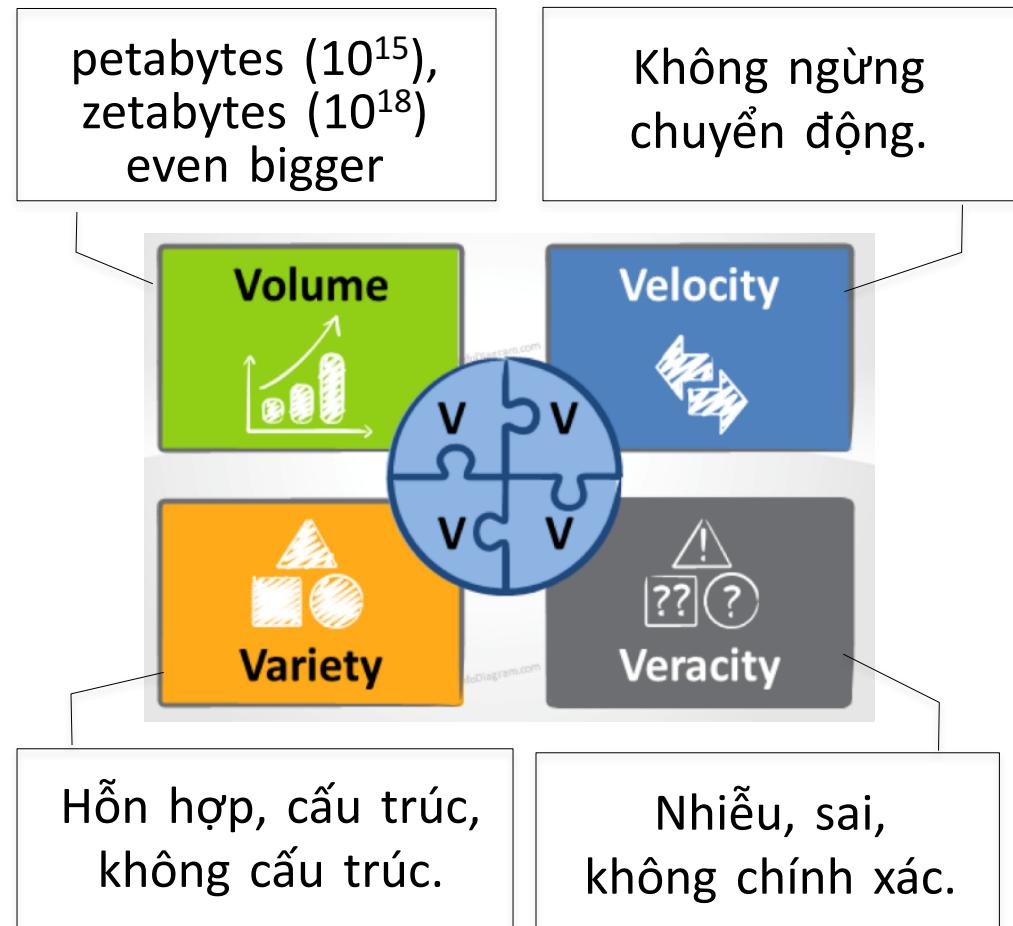


# Đột phá gần đây của công nghệ số



# 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 đổi đầu để phát triển.

# Scale up learning models of Google

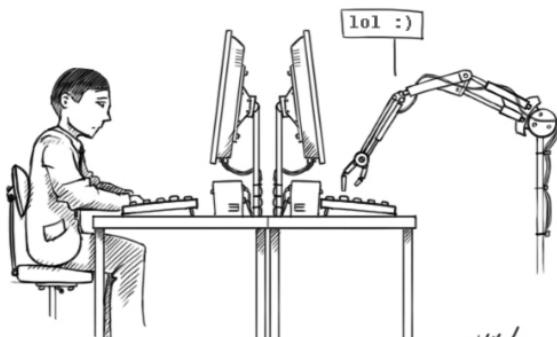
## Google Data Center



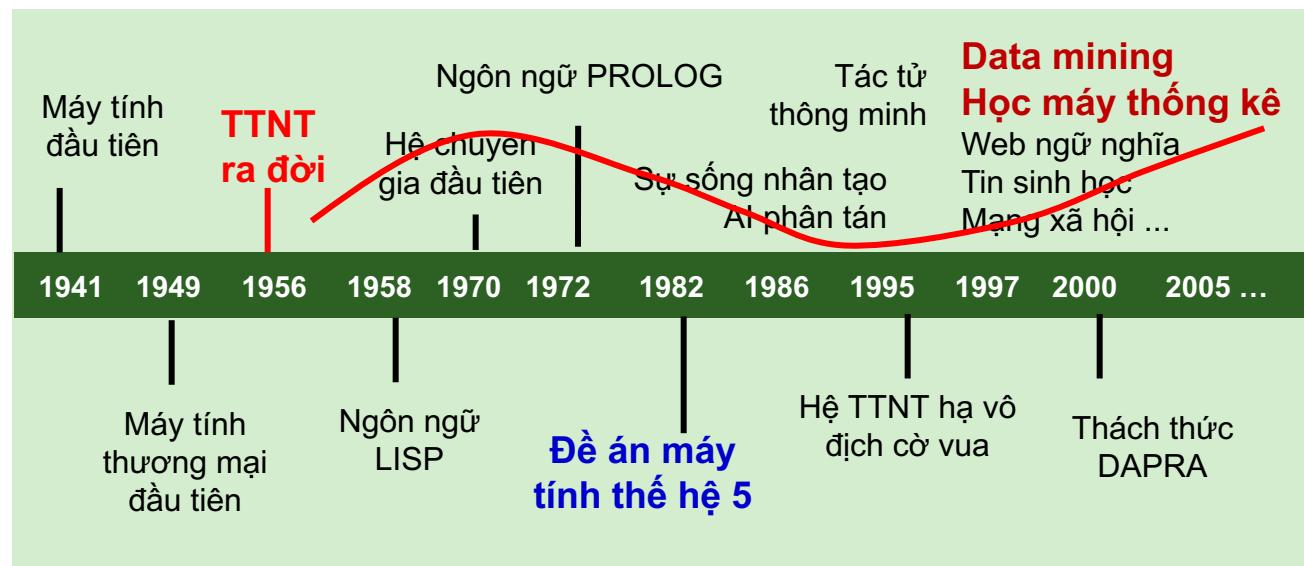
- Công nghệ: BigQuery (Tableau), Cloud Storage.
- Machine learning core
  - Logistic & linear regression, general convex losses
  - Infusion of L<sub>1</sub> and L<sub>2</sub> regularization
  - On-the-fly curvature estimation
- System infrastructure
  - MapReduce for parallelism
  - Multiple cores and 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%

# Artificial Intelligence – Trí tuệ nhân tạo

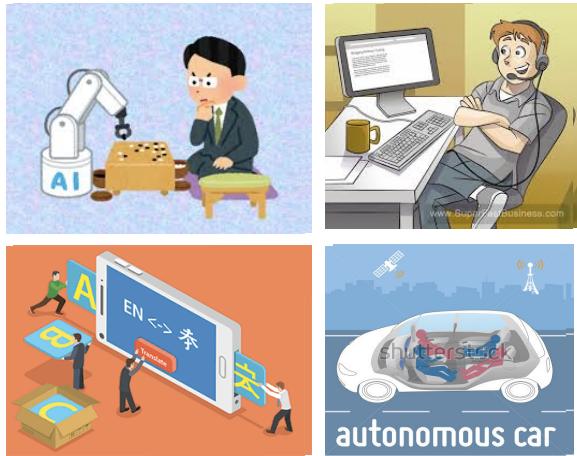


- Làm cho máy có trí thông minh (lập luận, hiểu ngôn ngữ, tự học).
- Phép thử Turing là một cách để trả lời 'máy tính có biết nghĩ không?'

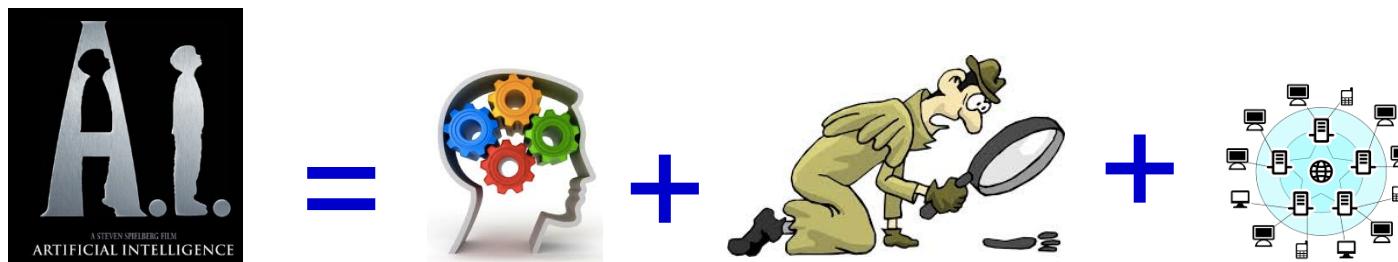


“...nếu có thể Bảo nên chuyển qua làm về trí tuệ nhân tạo vì đây là tương lai của tin học” (thư anh Phan Đình Diệu)

# Artificial Intelligence – Trí tuệ nhân tạo



- Làm cho máy có trí thông minh (lập luận, hiểu ngôn ngữ, tự học).
- AlphaGo, hiểu ngôn ngữ, tiếng nói, chẩn đoán ung thư, ô-tô tự lái...
- Hầu hết thành công gần đây của AI dựa vào học máy (machine learning).



Main conferences: IJCAI, AAAI, ECAI, PRICAI

# Trí tuệ nhân tạo dựa vào học máy

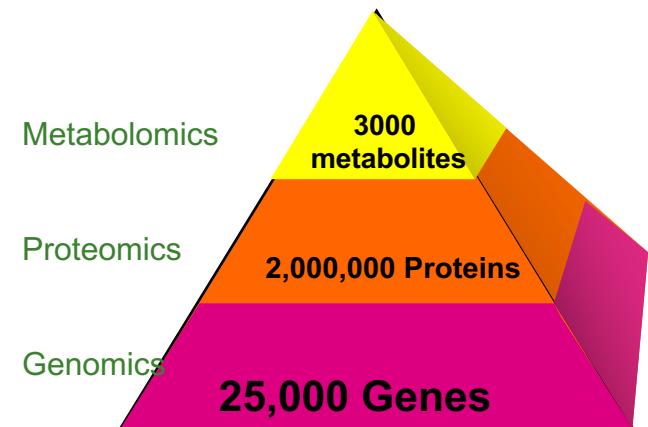
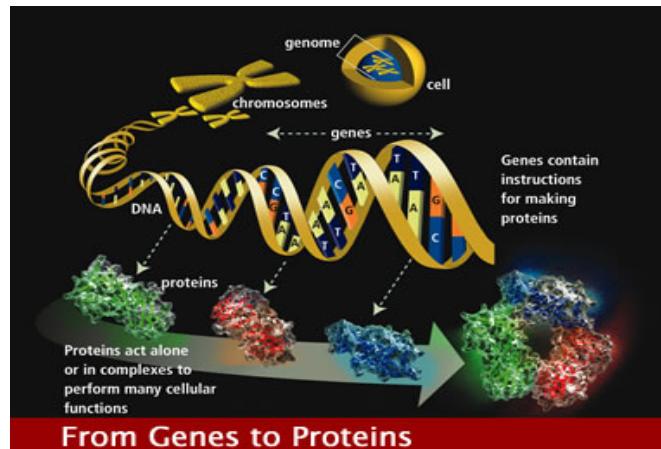
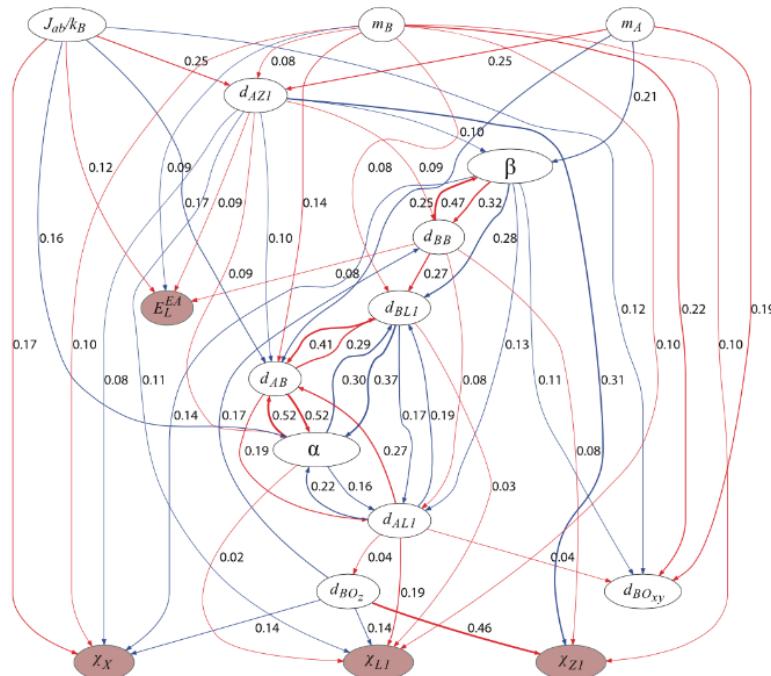
“Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs”

“Rất nhiều người làm các hệ AI nay đã nhận ra rằng, đối với rất nhiều ứng dụng, việc huấn luyện một hệ thống từ các thí dụ đầu vào-đầu ra để có quyết định hành động là dễ hơn rất nhiều việc soạn sẵn các quyết định mong muốn cho mọi tình huống có thể xảy ra.

M.I. Jordan, T. Mitchell. Machine Learning: Trends, perspectives, and prospects. *Science*, 349 (6245), 255–260, 2015.

# Công nghệ số và sinh học, công nghệ nano

- Bioinformatics
- Materials genomics initiatives



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

# Dưa hấu và thịt lợn rớt giá?

- 4/5/2017, Thủ Tướng nói về dưa hấu và thịt lợn
- Bộ trưởng Nguyễn Xuân Cường: “Cung lớn hơn cầu và từ sản xuất, chế biến đến tìm kiếm thị trường còn yếu kém, dẫn đến dư thừa và bế tắc đầu ra”.
- Dưa hấu và thịt lợn có số hoá được không? Có thể làm cho sản xuất này thông minh?
- CMCN4 là cách mạng không chỉ của ... công nghiệp  
→ Ta cần làm nông nghiệp và du lịch thông minh? Giáo dục, môi trường và y tế thông minh? Các lĩnh vực khác?



**Có thể thực hiện đến đâu sự thay đổi phương thức sản xuất mới trong việc ta muốn và cần làm?**

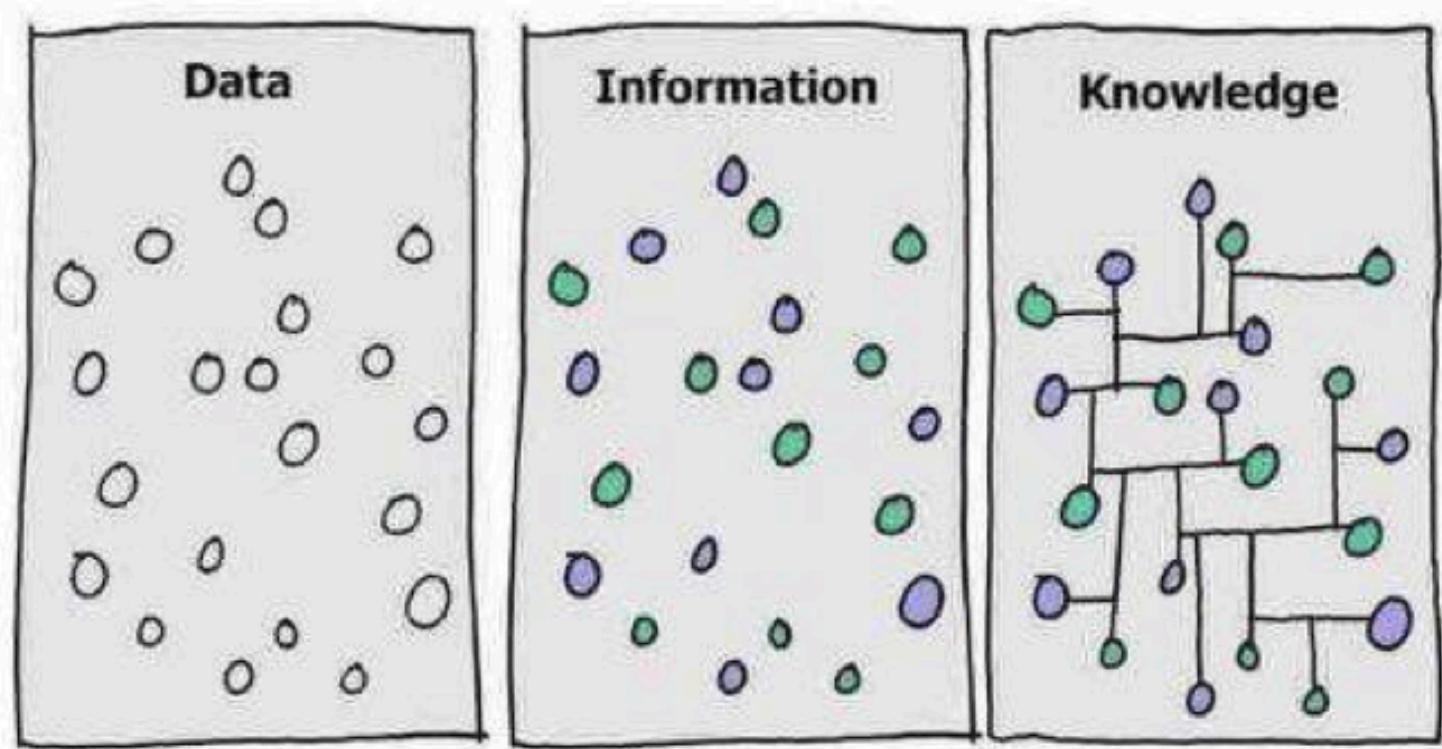
# Ta nên và có thể đi trong CMCN4 thế nào?

- Nông nghiệp và du lịch thông minh? Giáo dục, môi trường và y tế thông minh? Lựa chọn và làm chủ những công nghệ số và các công nghệ cao cần cho mình?
- Ai nuôi trồng những 'cây và con' như ta? Sản lượng bao nhiêu? Nhu cầu thị trường? Dịch chuyển trồng lúa sang 'cây con' khác ở đâu? Bao nhiêu? Giá trị hơn bao nhiêu?
- Số hoá được sông ngòi, tính toán và mô phỏng được các tình huống lũ lụt? Làm e-health thế nào?
- Chiến lược và chính sách quốc gia, thay đổi của các doanh nghiệp, lực lượng tinh hoa của KH&CN (CMCN4 không thể làm chỉ bởi ý chí mà phải bằng tri thức).
- Vai trò to lớn của toán học.

# Outline

- Cách mạng công nghiệp lần thứ tư
  - **Khoa học dữ liệu là gì?**
  - Nguyên lý và phương pháp của khoa học dữ liệu
-

# Data, information, knowledge

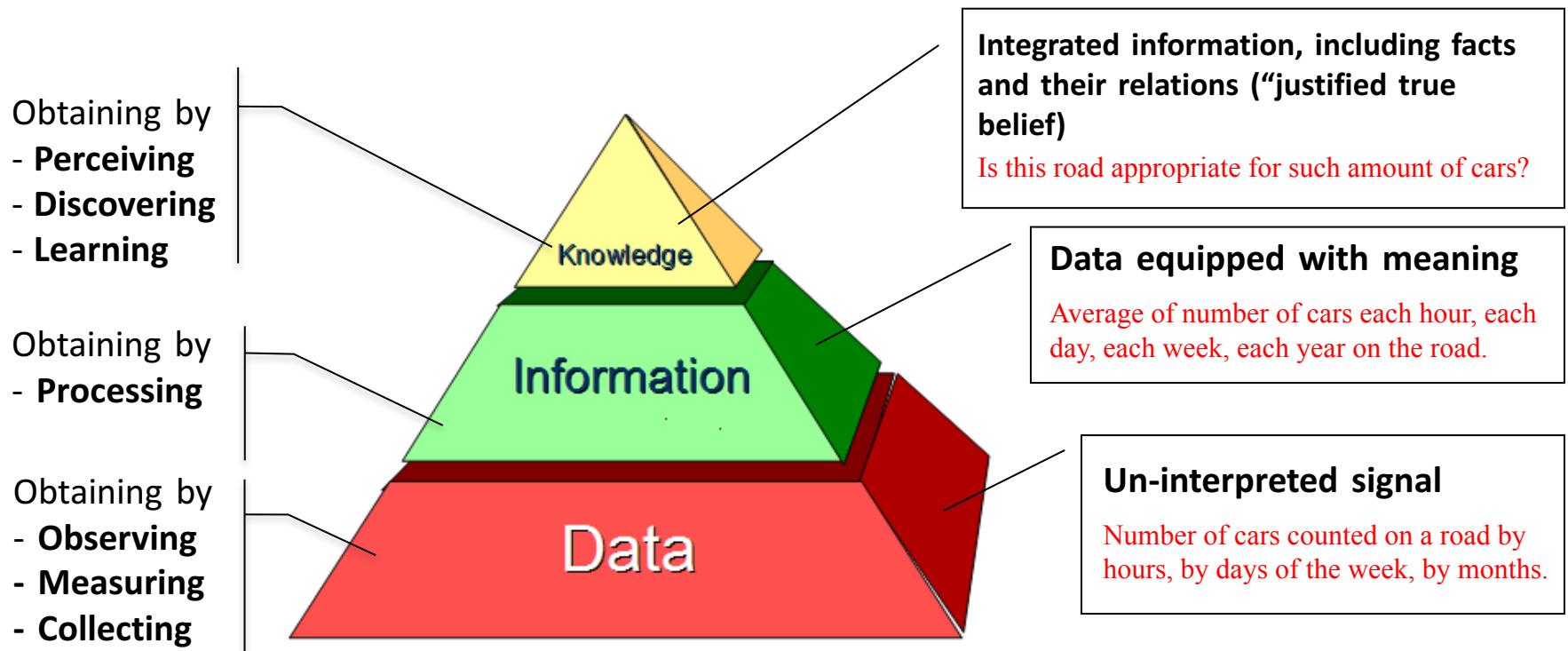


---

From Julien Blin

# Data, information, and knowledge

Knowledge can be considered data at a high level of abstraction and generalization.

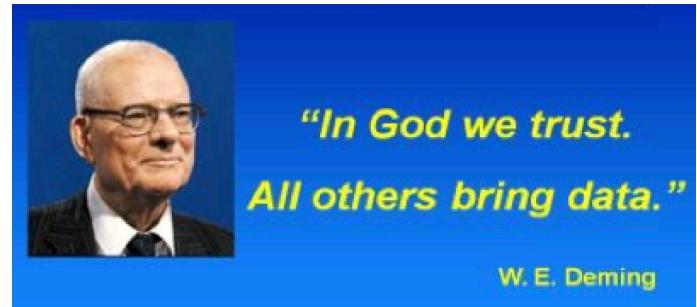
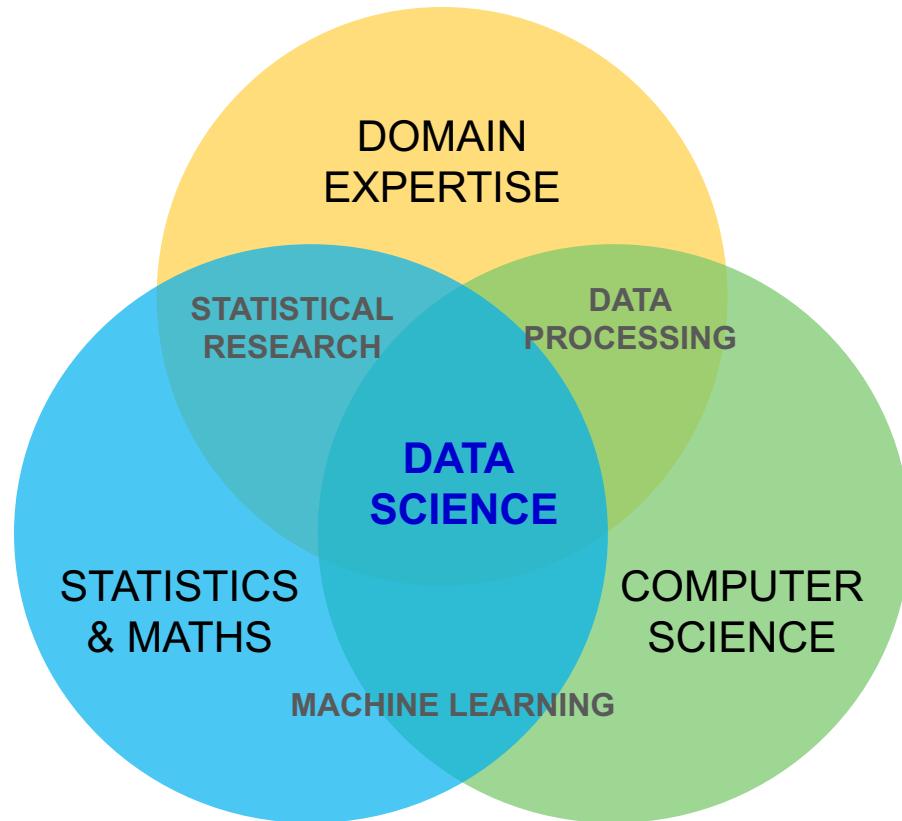


# Vài định nghĩa về Khoa học dữ liệu?

- There is not yet a definition agreed by all.
- Some examples

NIST (National Institute of Standards and Technology)	<p><b>Data science is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing</b></p> <p><i>Trực tiếp trích rút tri thức hành động từ dữ liệu qua quá trình phát hiện, thiết lập và kiểm nghiệm các giả thiết.</i></p>
Microsoft	<p><b>Data science is about using data to make decisions that drive actions.</b></p> <p><i>Dùng dữ liệu tạo quyết định dẫn dắt hành động</i></p>

# Khoa học dữ liệu

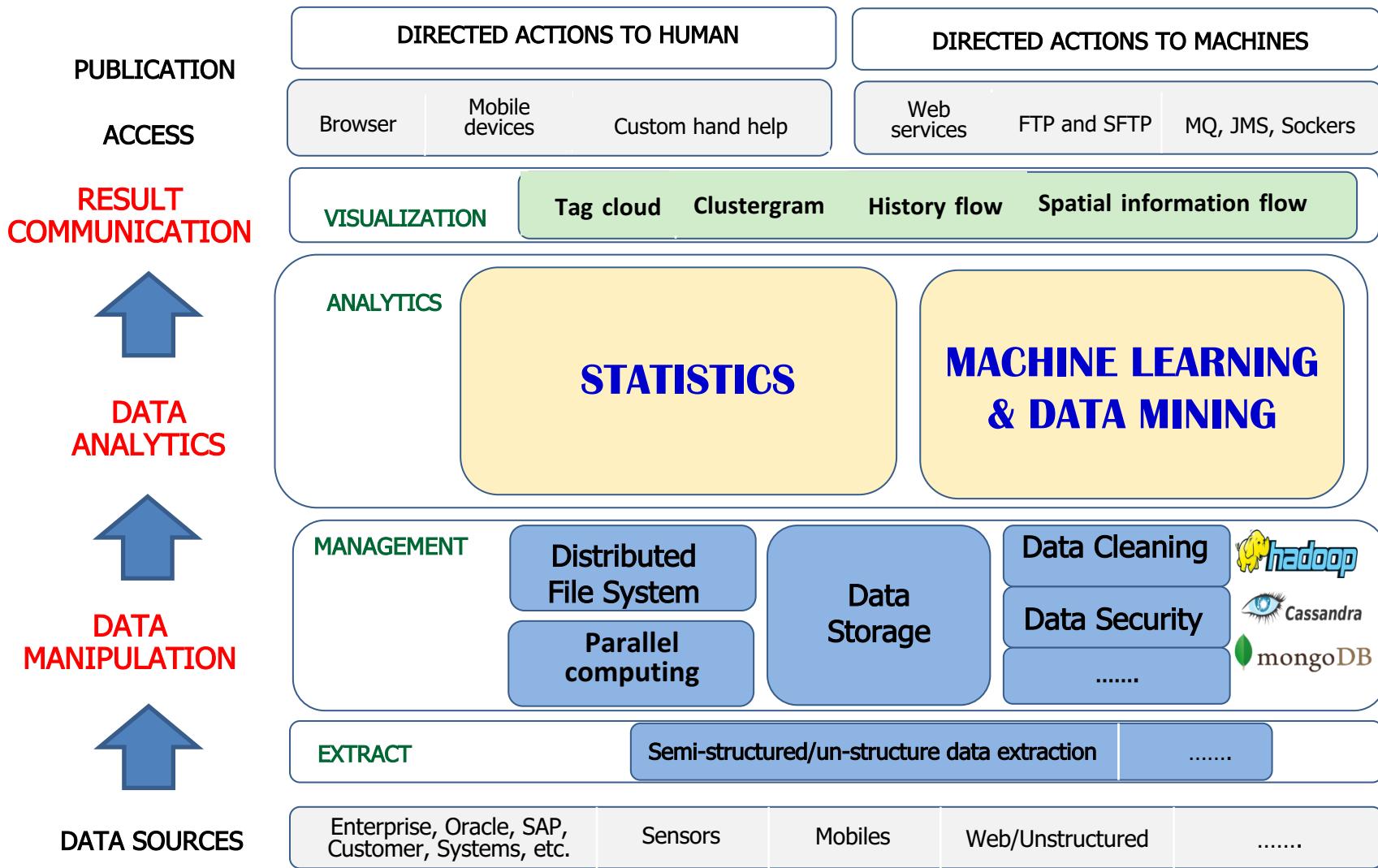


“Ta chỉ tin vào Thượng đế.  
Mọi thứ khác phải dựa vào  
dữ liệu”



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

# A scheme of data science



# How does people collect data?

- Dữ liệu chính là **giá trị của các thuộc tính** (features, attributes, properties, variables) của các đối tượng, thu được do quan sát, đo đạc và thu thập (số hoá).
- Hai cách thu thập dữ liệu

Lấy mẫu  
ngẫu nhiên

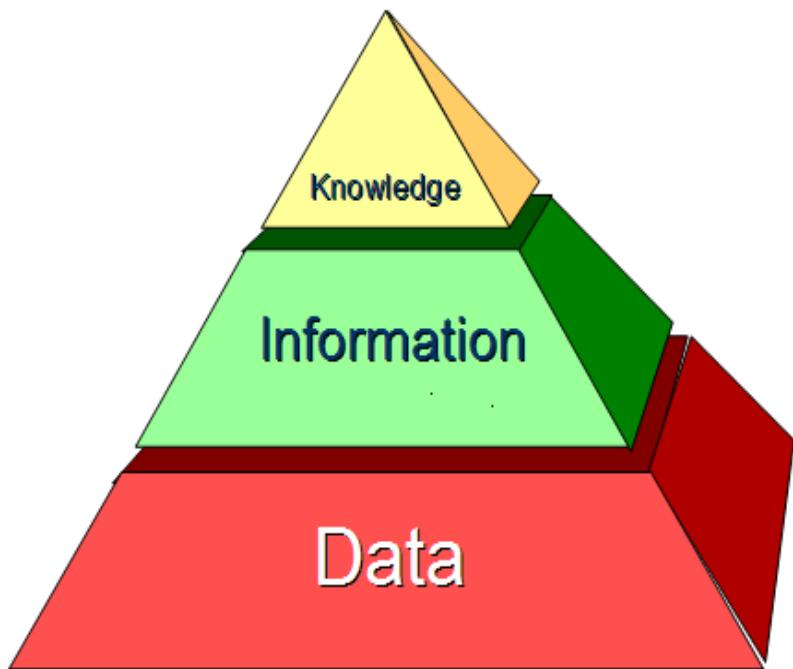
Thu mọi dữ liệu  
có được

**Thống kê truyền thống:** Có mục tiêu rồi  
lấy mẫu và phân tích.

**Khai phá dữ liệu:** Dữ  
liệu thu thập không liên  
quan đến mục tiêu nào.

# From data to knowledge?

Có thể xem tri thức là dữ liệu ở mức tổng quát hóa cao (generalization).



Nhiều khoa học liên  
quan việc đi **từ** dữ liệu  
đến tri thức

- Statistics
- Machine Learning
- Data Mining
- Data Science

# Thống kê - Statistics

- **Thống kê:** Phân tích, khái quát, ra quyết định từ dữ liệu.
- **Nội dung chính**
  - Thống kê mô tả (descriptive statistics)
  - Thống kê suy diễn (inferential statistics)
- **Dữ liệu**
  - Thu thập để trả lời những *câu hỏi định trước*
  - Phần lớn là dữ liệu số, ít dữ liệu hình thức.
- 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*

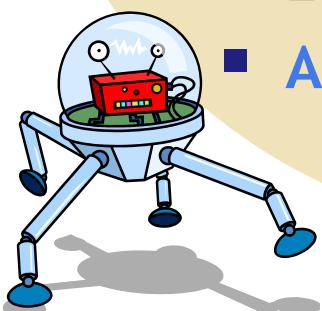
- Phân tích quan hệ của nhiều biến trên máy tính
- Phân tích khẳng định (CDA, confirmatory data analysis) kiểm định giả thiết
- Phân tích khám phá (EDA, exploratory data analysis) dùng dữ liệu tạo ra các giả thiết. Nhiều phương pháp: Factor analysis, PCA, Linear discriminant analysis, Regression analysis, Cluster analysis
- Đặc điểm
  - Có nền tảng toán học sâu sắc và vững chắc
  - Thay đổi với CNTT, học máy/khai phá dữ liệu

# Machine learning and data mining



## Machine learning

- To build computer systems that learn as human does.
- ICML since 1982 (33th ICML in 2016), ECML since 1989.
- ECML/PKDD since 2001.
- **ACML** starts Nov. 2009.



## Data mining

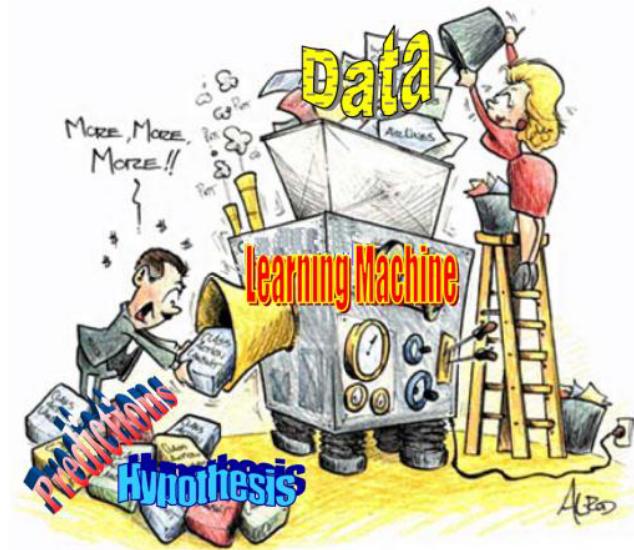
- To find new and useful knowledge from large datasets.
- ACM SIGKDD (1995), PKDD and **PAKDD** (1997) IEEE ICDM and SIAM DM (2000), etc.

ACML: Asia Conference on Machine Learning

PAKDD: Pacific Asia Knowledge Discovery and Data Mining

# Machine learning

- “Field of study that gives computers the ability to learn without being explicitly programmed”  
(Arthur Samuel, 1959).
- “Machine learning sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with **automatics improvement over time**, and inference and **decision-making under uncertainty**.” (Jordan & Michell, 2015)



(from Eric Xing lecture notes)

# Statistics vs. Machine Learning

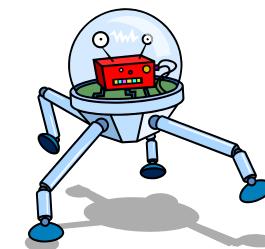
## Statistics

- Suy diễn thống kê (ước lượng, kiểm định giả thiết).
- Ban đầu **mô hình** cho bài toán có số chiều nhỏ, ở dạng số.
- Khó thay đổi ‘văn hóa’, thích nghi với môi trường tính toán.
- Mở rộng sang học máy.



## Machine learning

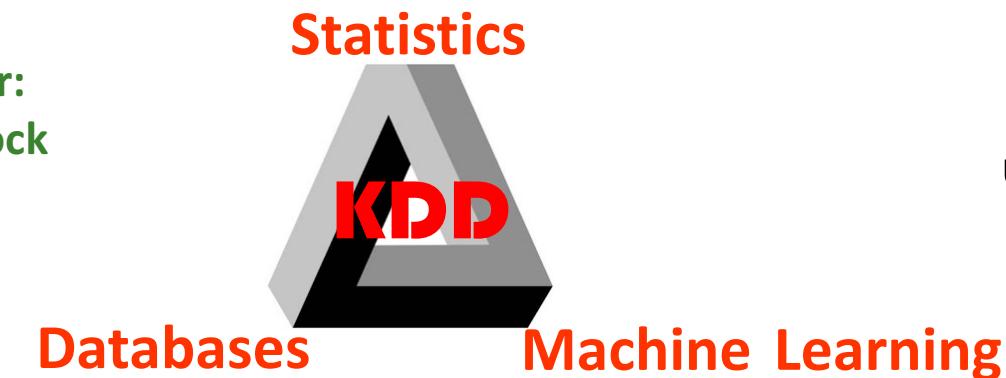
- Dự đoán với dữ liệu hình thức.
- Bắt đầu với các **thuật toán trực cảm** (heuristics).
- Dần gắn với thống kê, có **mô hình toán học** cho các thuật toán.



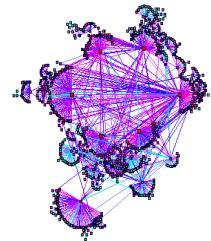
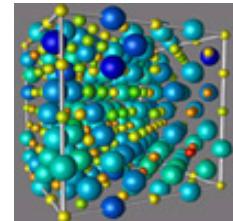
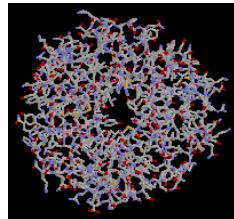
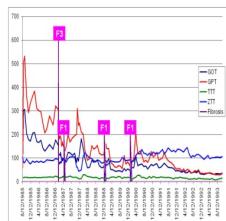
# Khai phá dữ liệu – Data Mining

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

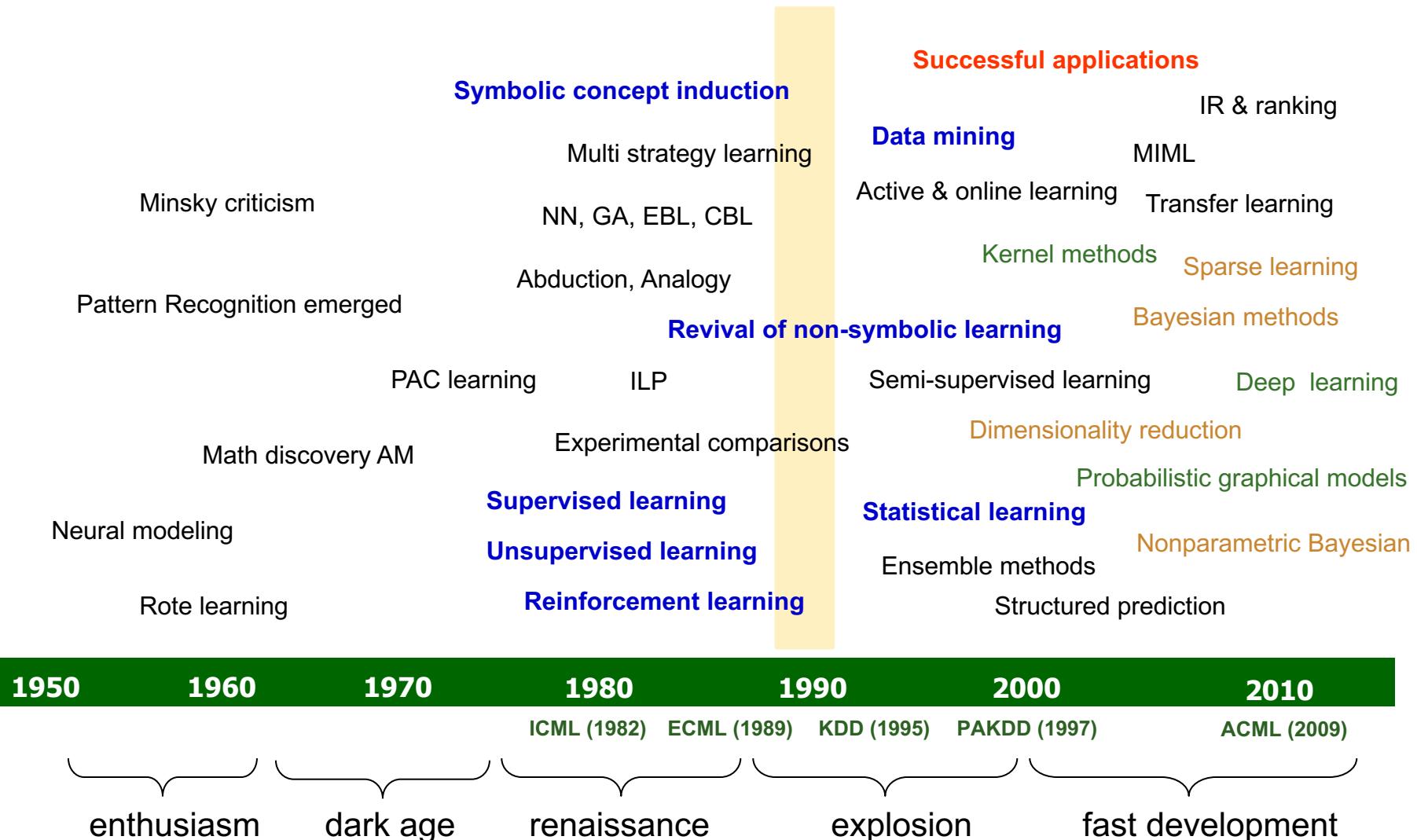
Data mining metaphor:  
Extracting ore from rock



Large and  
unstructured  
real-life data



# Development of machine learning



# Outline

- Cách mạng công nghiệp lần thứ tư
- Khoa học dữ liệu là gì?
- **Nguyên lý và phương pháp của khoa học dữ liệu**

---

Một số slides chưa chuyển qua tiếng Việt nhưng sẽ được trình bày bằng tiếng Việt

# Nguyên lý của Khoa học dữ liệu?

**Principle** = a basic idea or rule that explains or controls how something happens or works (Cambridge Dict.)

Ý tưởng cơ bản hoặc các nguyên tắc để giải thích tại sao mọi sự lại xảy ra hoặc để điều khiển sự vận hành.



1. Data type and structure
2. Process
3. Methods
4. Model selection

# Data types and structure vs. methods

## Data types and structures

- Flat data tables
- Relational
- Temporal
- Transactional
- Multidimensional
- Genomic
- Materialized
- Textual data
- Web data
- etc.

## Mining tasks and methods

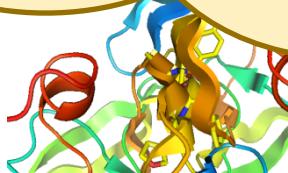
Hầu hết các phương pháp  
được phát triển cho dữ liệu  
ở dạng bảng. Nếu không cần  
chuyển dữ liệu về dạng bảng  
hoặc cải tiến/thích nghi  
phương pháp.

Prediction

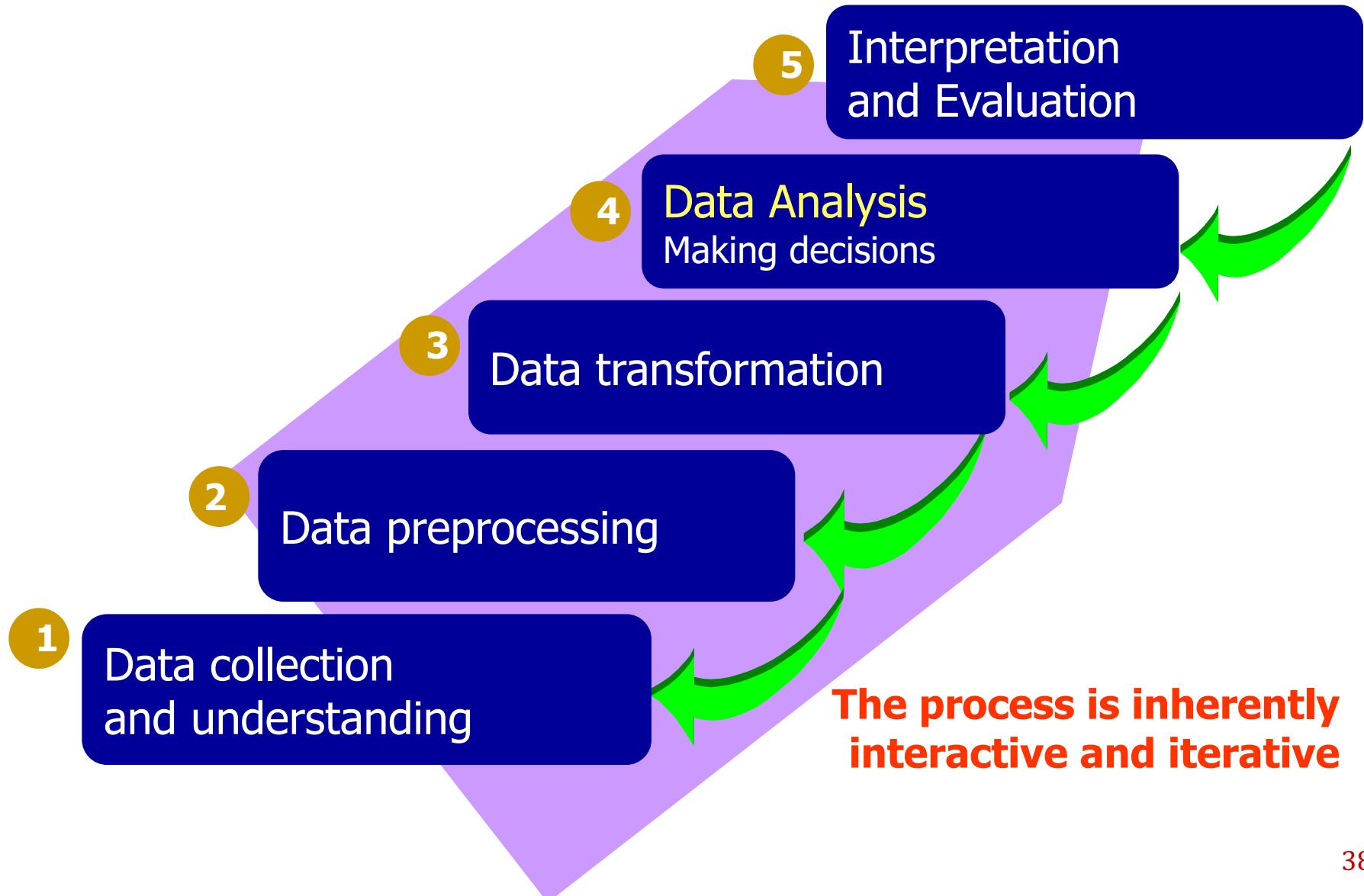
analysis

Clustering

- Summarization
- etc.



# The data analysis process



# Why we should care about data types?

Combinatorial search in hypothesis spaces (machine learning)



Attribute	Numerical	Symbolic	
No structure $= \neq$		Places, Color	<b>Nominal or categorical</b> (Binary, Boolean)
Ordinal structure $= \neq \geq$	<b>Integer:</b> Age, Temperature	Rank, Resemblance	Ordinal
Ring structure $= \neq \geq + \times$	<b>Continuous:</b> Income, Length		Measurable

Possible analysis operations (thus methods, algorithms) depend on data types

Often matrix-based computation (multivariate data analysis)

# Structured or unstructured data?

## ■ Structured data

- ❑ Can be stored in database SQL in table with rows and columns.
- ❑ Only about 5-10% of all available data.

## ■ Semi-structured data

- ❑ Doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze.
- ❑ XML documents and NoSQL databases documents are semi structured

	swims	has fins	flies	has lung	is a fish
Herring	yes	yes	no	no	yes
Cat	no	no	no	yes	no
Pigeon	no	no	yes	yes	no
Flying fish	yes	yes	yes	no	yes
Otter	yes	no	no	yes	no
Cod	yes	yes	no	no	yes
Whale	yes	yes	no	yes	no



```
@BOOK{Maz91,  
author = "J. Mazzeo",  
year = "1991",  
title = "Comparability of Computer and Paper-and-Pencil Scores",  
address = "(College Board Rep. No. 91). Princeton, NJ",  
publisher = "Educational Testing Service",}
```

```
@BOOK{Mil93,  
author = "M. E. Miller",  
year = "1993",  
title = "The Interactive Tester (Version 4.0) [Computer software]",  
publisher = "Psystek Services",  
address = "Westminster, CA",}
```

Articles in a Latex database

# Structured or unstructured data?

## ■ Unstructured data

- Unstructured data represent around 80% of data. It often include text and multimedia content.

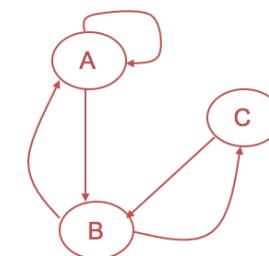
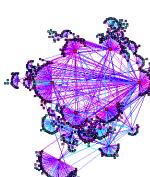
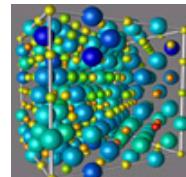
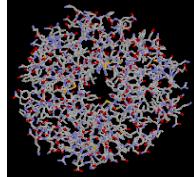
*Example:* e-mail messages, word documents, videos, photos, audio files, webpages and many other kinds of business documents.

- A key issue in data science is **representing unstructured data**

*Example:* The DNA sequence

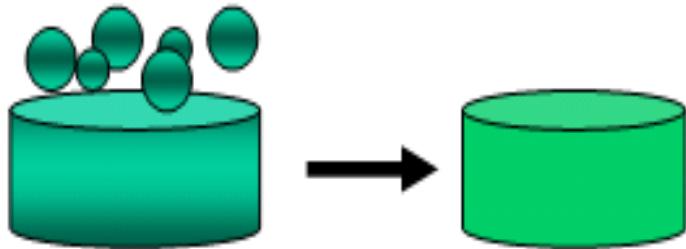
“...TACATTAGTTATTACATTGAGAACTTATAATTAAAAAAAGATTC...”

can be represented by different ways for computation such as sliding windows, motifs, kernel function, web link... representation

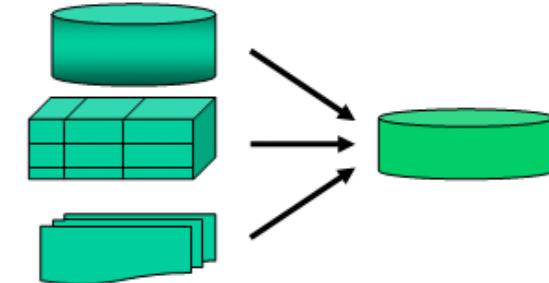


	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

# Major tasks in data preprocessing

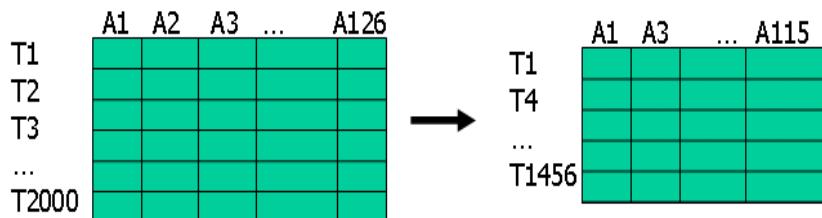


1 Data cleaning

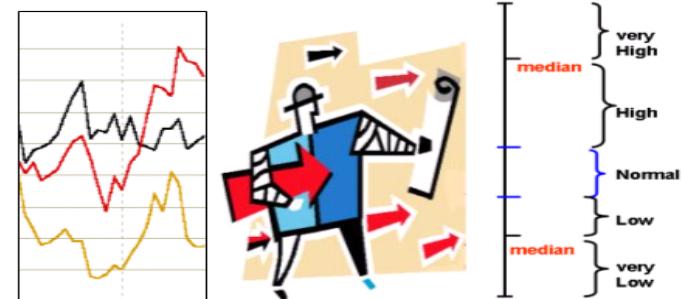


-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

2 Data integration and transformation

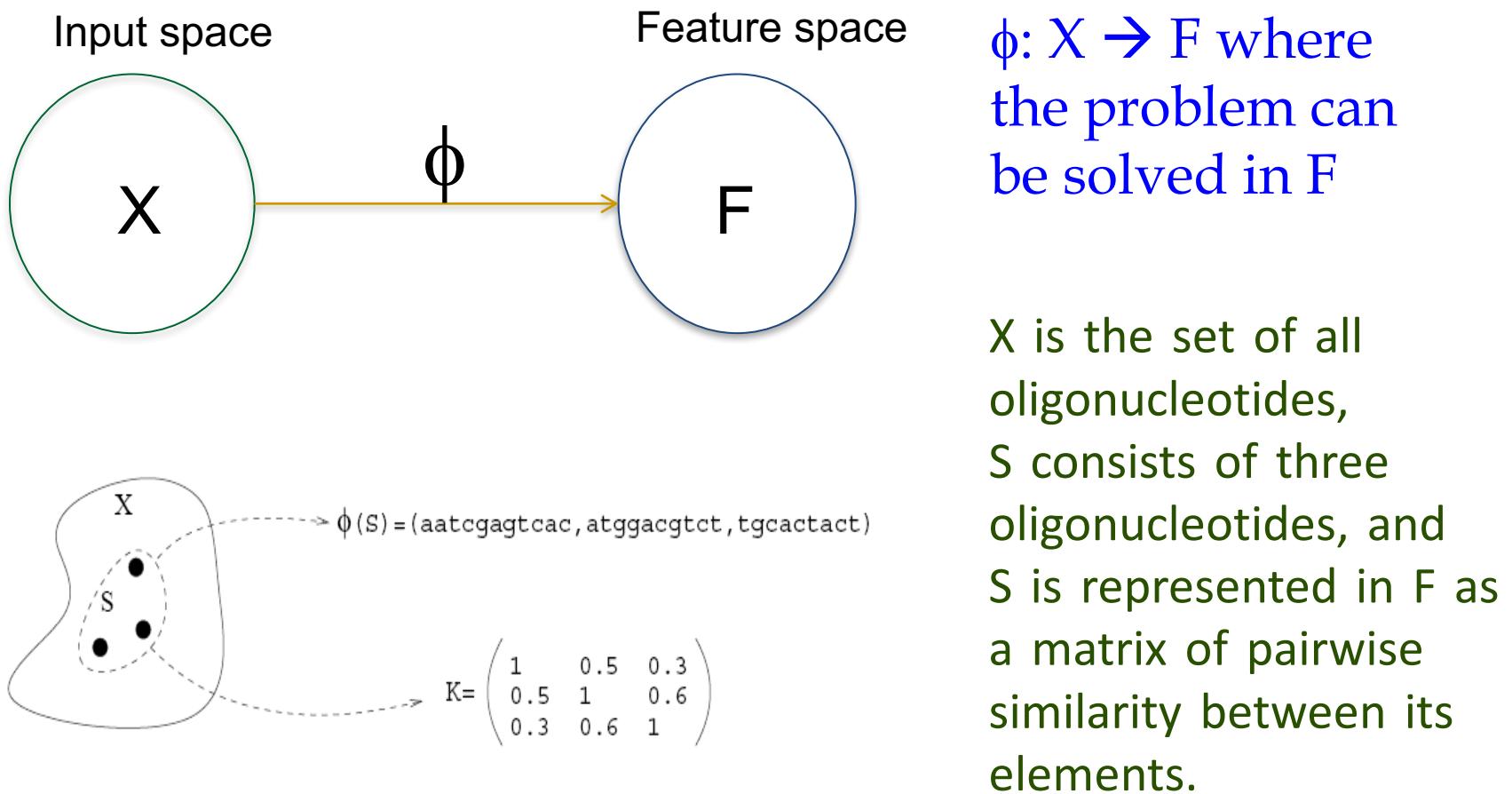


3 Data reduction  
(instances and dimensions)



4 Data discretization

# Data transformation



# Data Transformation

facebook's profits have jumped in the first three months of the year as network closes in on two billion users according to its latest results the people using facebook each month increased to 194 billion of which nearly use it daily the company said the us tech giant reported profits of just over \$3bn £24bn in the first quarter a 76% rise year-on-year however it warned that growth in ad revenues would slow down the company has also come under sustained pressure in recent weeks over its handling of hate speech child abuse and self-harm on the social network on wednesday facebook chief executive mark zuckerberg announced it was hiring 3000 extra people to moderate content on the site facebook bolsters moderating team zuckerberg addresses facebook killing a quarter of the worlds population now uses facebook every month with most of the new users coming from outside of europe and north america speaking after the results mr zuckerberg said the size of its user base gave facebook an opportunity to expand the sites role moving into tv health care and politics with that foundation our next focus will be building community he said theres a lot to do there ad slowdown the company grew its revenue from advertising which accounts for almost all of facebook's income by 51% to \$79bn in the period however chief financial officer

facebook has denied it is targeting insecure young people in order to advertising amid a row over a leaked document a research paper reportedly published by the australian newspaper was said to go into detail about how users post about self-image weight loss and other issues facebook claims the research was shared with advertisers but said the article was misleading facebook does not offer tools to target people based on their emotional state the network said the analysis done by an australian researcher was intended to help marketers understand how people express themselves on facebook it was never used to target ads and was based on data that was anonymous and aggregated facebook has an established process to review the research we perform this research did not follow that process and we are reviewing the details to correct the oversight stressed and stupid according to the australian the report was seen by marketers working for several major australian banks and was written by facebook executives david fernandez and andy sinn the document said facebook had the ability to monitor photos and other posts for users who may be feeling stressed defeated anxious nervous stupid overwhelmed silly useless or a failure the research only covered facebook users in australia and new zealand the statement on monday appeared to soften an earlier comment which mooted the possibility of disciplinary action over

.....  
8 documents

D1

D2

N-gram

social media 4  
fake account 4  
on facebook 5

....  
TF-IDF

Vocab\_size = 1066

D1 = (0.0, 0.006, 0.005, ...., 0.009, 0.006, 0.0)

D2 = (0.012, 0.0, 0.0, ...., 0.0, 0.0, 0.006)

Doc2Vec

Hidden layer size = 5

D1 = (0.257, -0.745, 1.332, -0.598, 0.013)

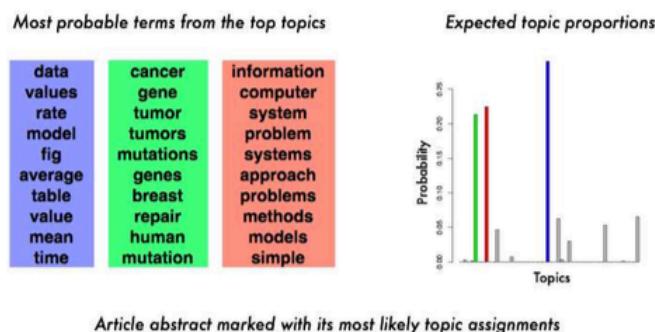
D2 = (0.246, -0.485, 1.113, -0.37, -0.06)

# 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

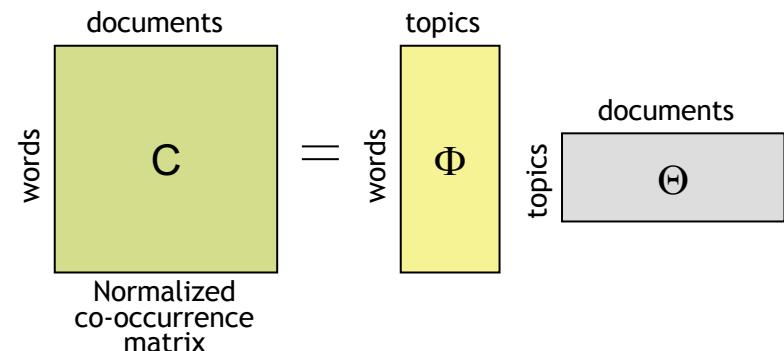


Article abstract marked with its most likely topic assignments

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.

- a topic is a probability distribution over words.
- documents are mixtures of latent topics.

### Topic models



Two problems of learning and inference

# Machine Learning: View by data

## Labelled vs. Unlabelled data

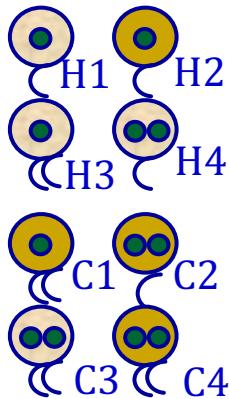
**Given:**  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

- $x_i$  is description of an object, phenomenon, etc.

- $y_i$  is some property of  $x_i$ , if  $y_i$  is not available data is **unlabelled**, otherwise **labelled**.

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

(**supervised learning**) [in between: **reinforcement learning**]



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	label
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

The problem is usually called **classification** if “label” is categorical, and **prediction** if “label” is continuous (in this case, if the descriptive attribute is numerical the problem is **regression**)

# Machine learning: View by method nature

Tribes	Origins	Master Algorithms
Symbolists	Logic, philosophy	Inverse deduction
Evolutionaries	Evolutionary biology	Genetic programming
Connectionists	Neuroscience	Backpropagation
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

---

The five tribes of machine learning, Pedro Domingos

# Symbolists



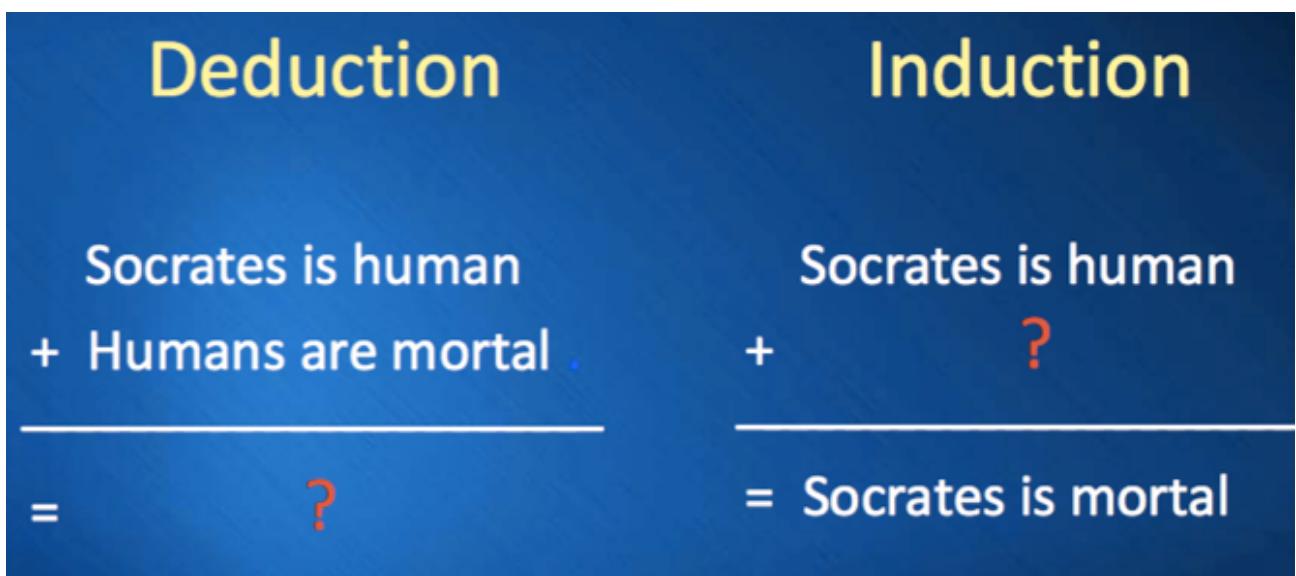
Tom Mitchell



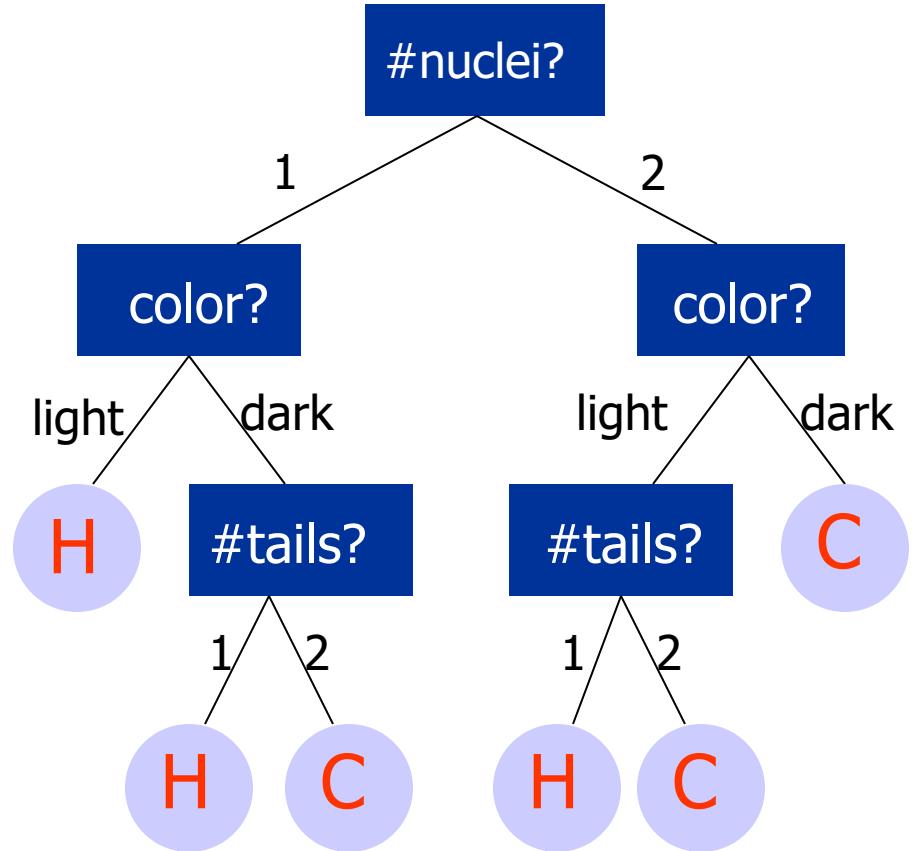
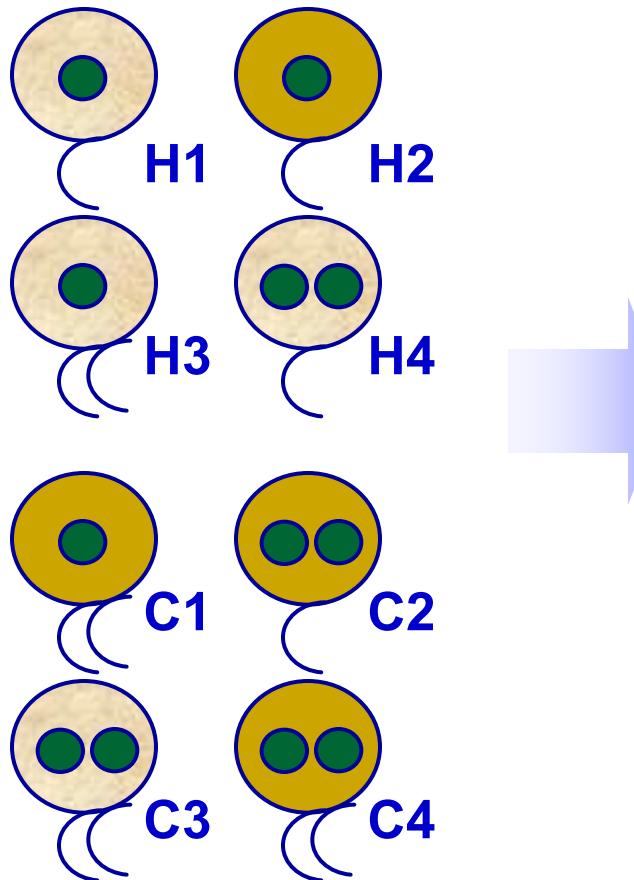
Steve Muggleton



Ross Quinlan



# Classification with decision trees



# Evolutionaries



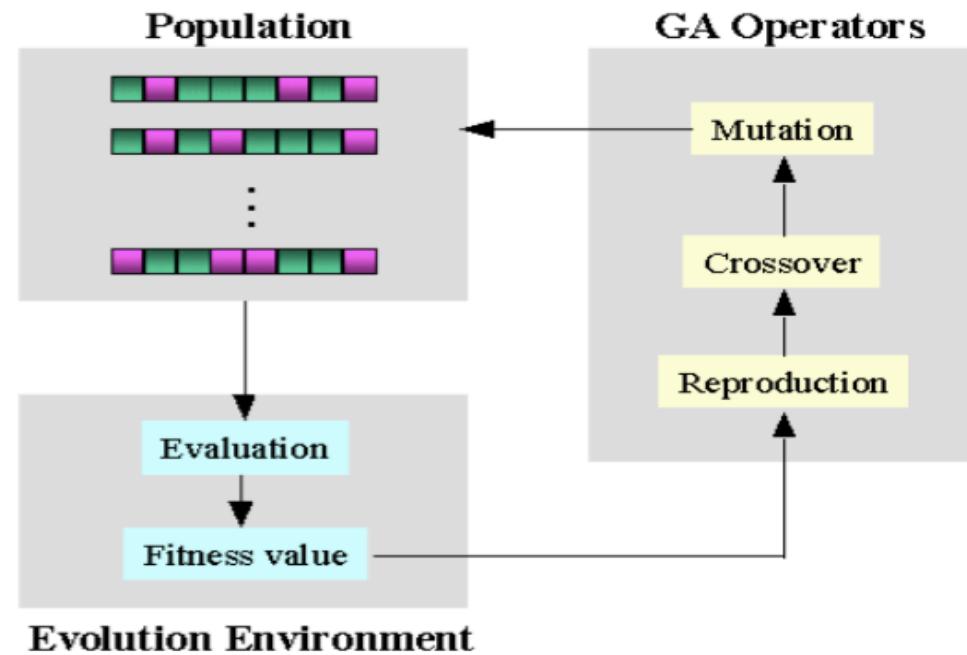
John Koza



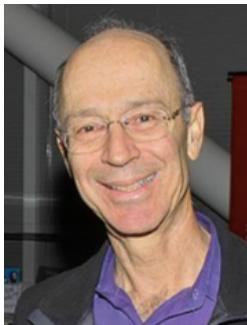
John Holland



Hod Lipson



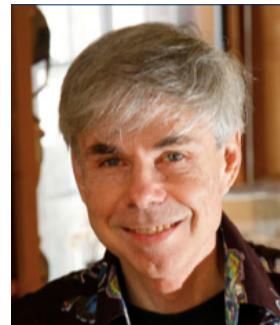
# Analoziger



Peter Hart



Vladimir Vapnik



Douglas Hofstadter

## ■ Instance-based classification

- Using most similar individual instances known in the past to classify a new instance

## ■ Typical approaches

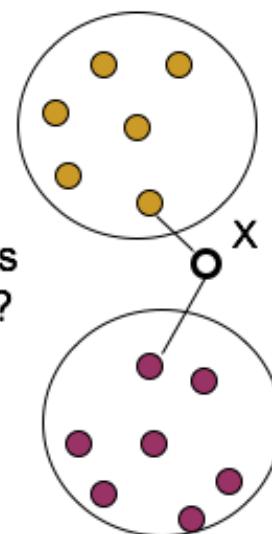
- **k-nearest neighbor approach**

- Instances represented as points in a Euclidean space

Class A

X belongs  
to A or B?

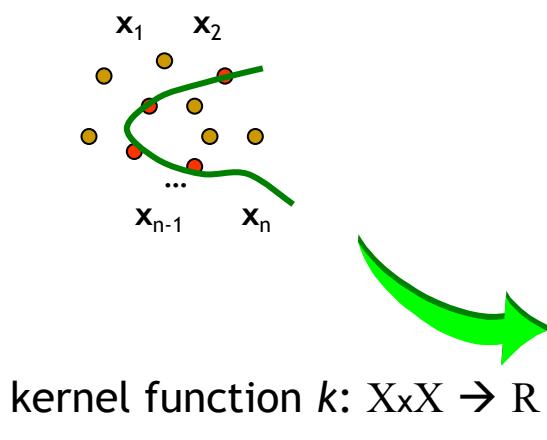
Class B



# Kernel methods

*The basic ideas*

Input space X

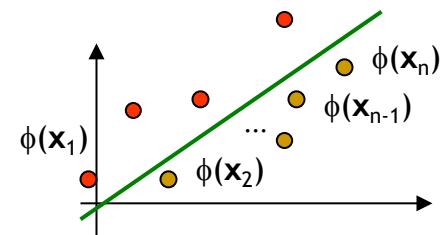


inverse map  $\phi^{-1}$   
 $\phi(x)$

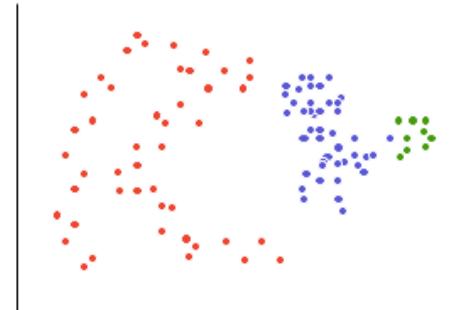
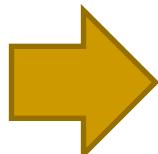
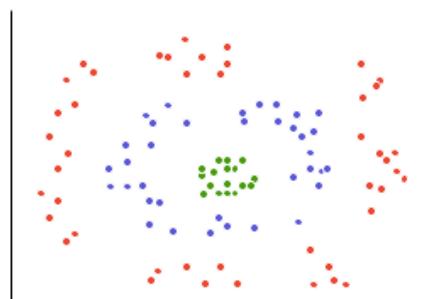
$$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

Kernel matrix  $K_{n \times n}$

Feature space F



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



# Connectionists



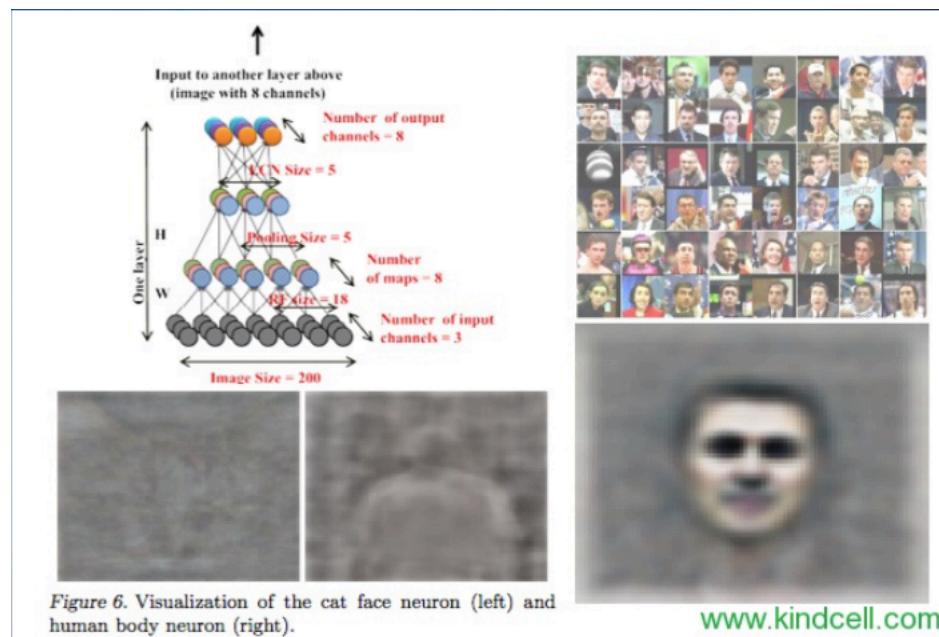
Yann LeCun



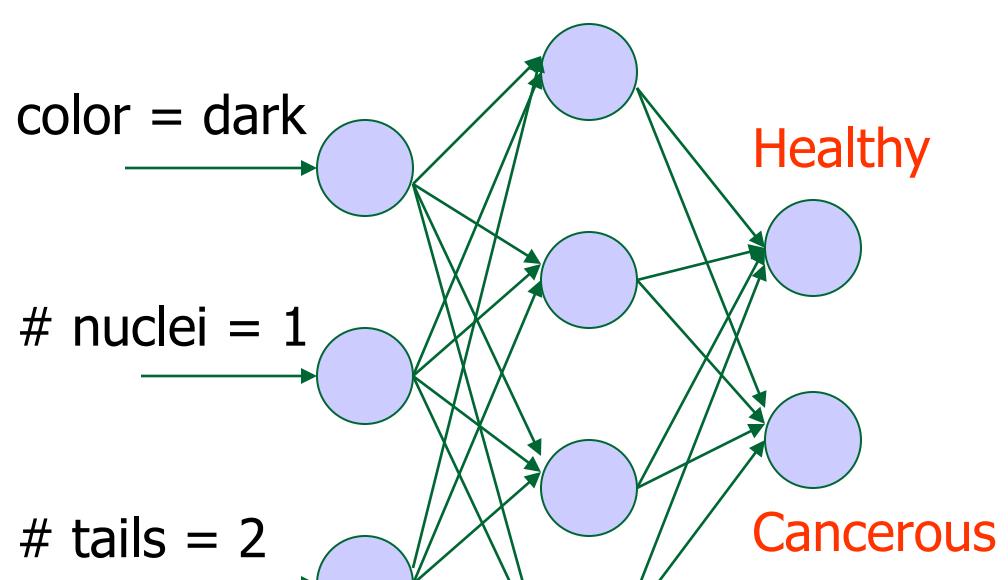
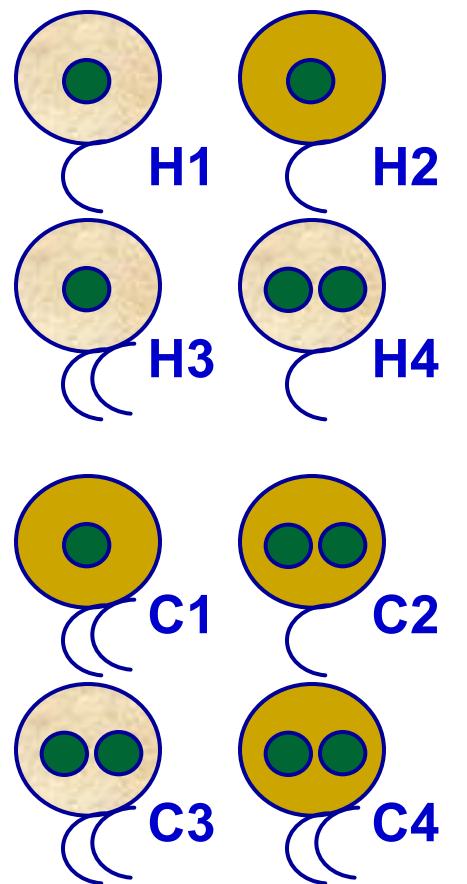
Geoff Hinton



Yoshua Bengio

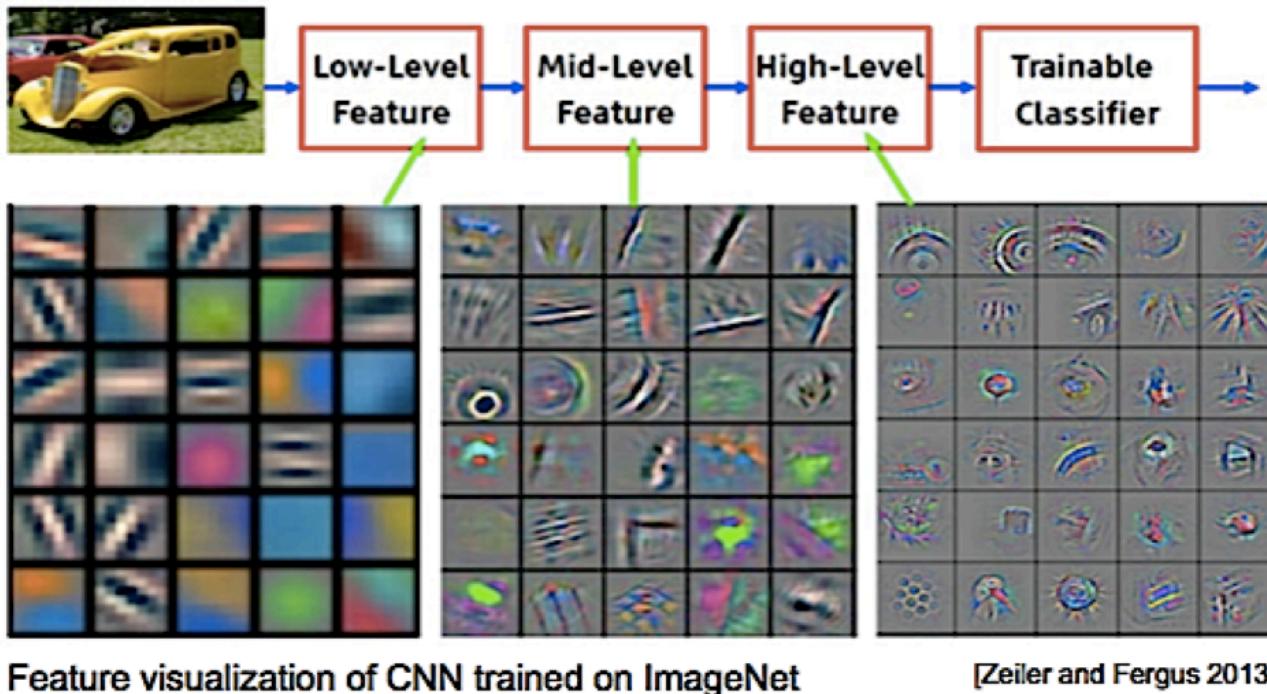


# Classification with neural networks



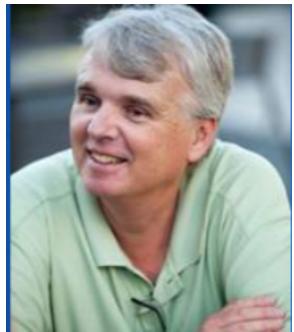
# Deep Learning

“Deep Learning: machine learning algorithms based on learning **multiple levels** of representation and abstraction” Joshua Bengio



GS Phùng Quốc Định sẽ nói về các mô hình của deep learning (học nhiều tầng), chia sẻ các kinh nghiệm, bài học, hạn chế và xu hướng trong lĩnh vực này.

# Bayesians in machine learning



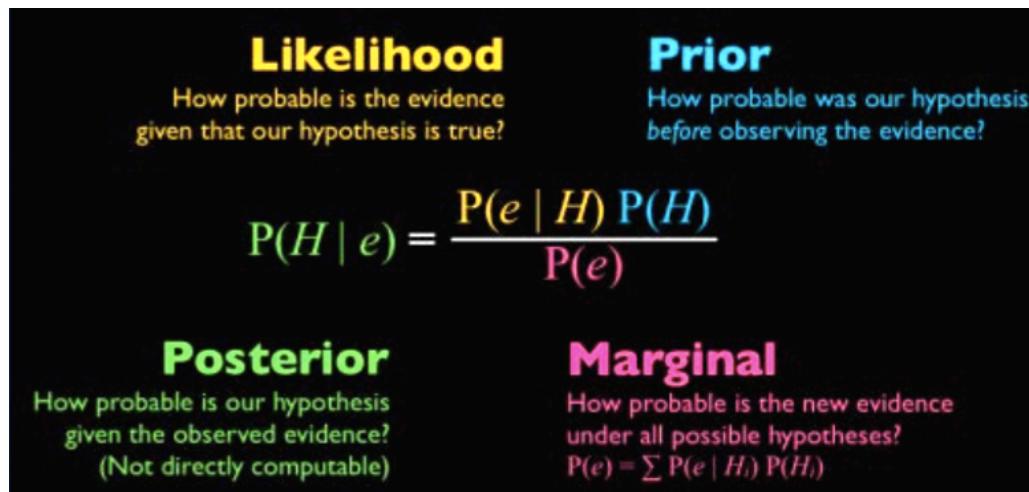
David Heckerman



Judea Pearl



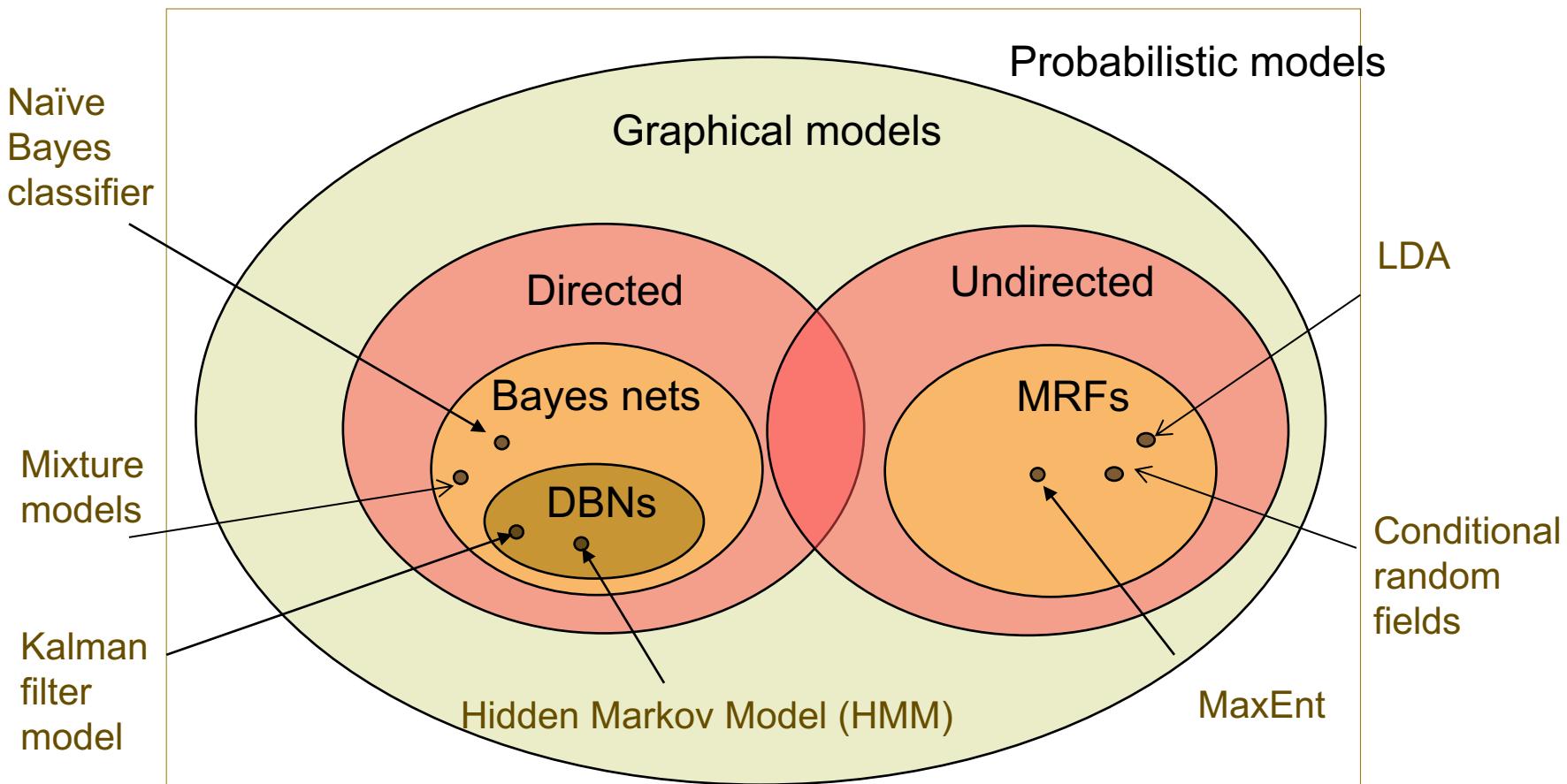
Michael Jordan



GS Nguyễn Xuân Long sẽ chia sẻ một số nền tảng thống kê của khoa học dữ liệu

# Probabilistic graphical models

## *Instances of graphical models*



# Probabilistic graphical models

## *Approximate inference*

### **Sampling inference** (stochastic methods)

**Markov Chain Monte Carlo (MCMC)** cho kết quả dưới dạng một tập các mẫu (samples) tìm được từ phân bố hậu nghiệm (posterior distribution).

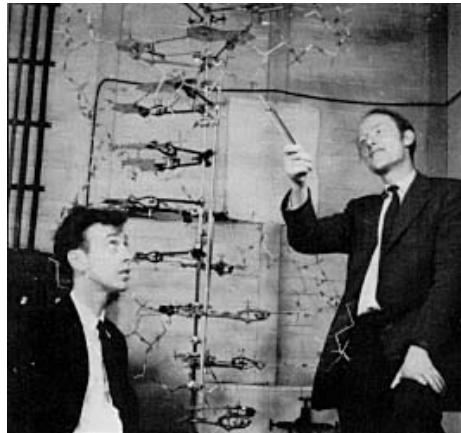
### **Variational inference** (deterministic methods)

Tìm một phần tử tối ưu của một họ các phân bố xấp xỉ bởi cực tiểu hóa một tiêu chuẩn thích hợp đo sự khác nhau giữa các phân bố xấp xỉ và phân bố hậu nghiệm chính xác.

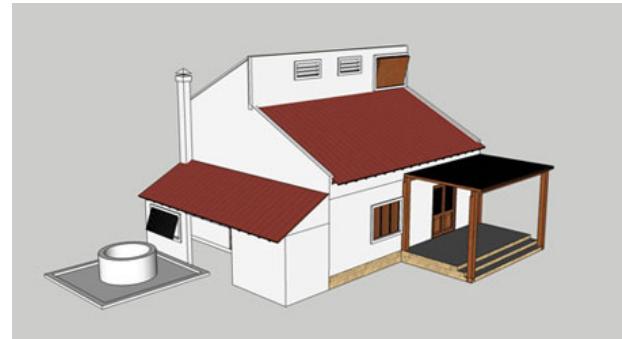
GS Phùng Quốc Định sẽ chia sẻ kinh nghiệm phần này khi nói về big data.

# Model selection

**Model:** Abstract description or representation of a reality.



DNA model figured out in 1953 by Watson and Crick



A model is defined as a parametric collection of probability distributions, indexed by model parameters

$$M = \{f(y|\theta) | \theta \in \Omega\}$$

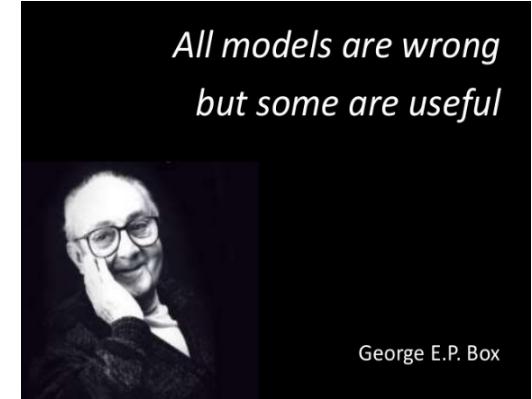
---

**Pignet index (body build index) = Stature in cm - (weight in kg + chest circumference in cm)**

Very sturdy: <10, Sturdy: 10-15, Good: 16-20, Average: 21-25, Weak: 26-30, Very weak: 31-35, Poor: >36

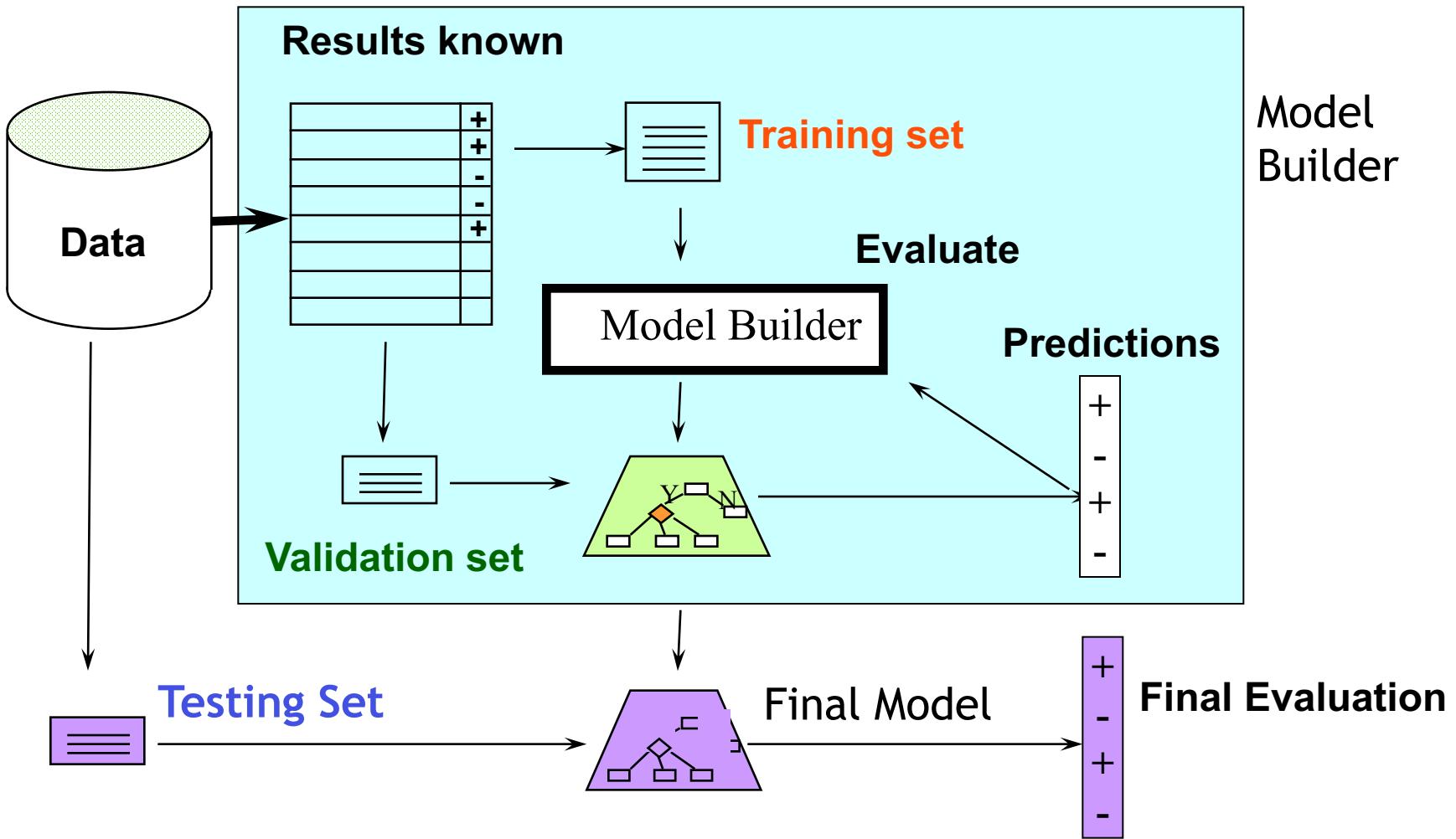
# Model selection

- **Problem:** Choosing *the most appropriate* model(s) given a dataset and the task.
- Relating to selecting
  - Models that can be appropriated
    - Parameters of those models
- Examples of model selection problems
  - 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 k-NN?
  - 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?

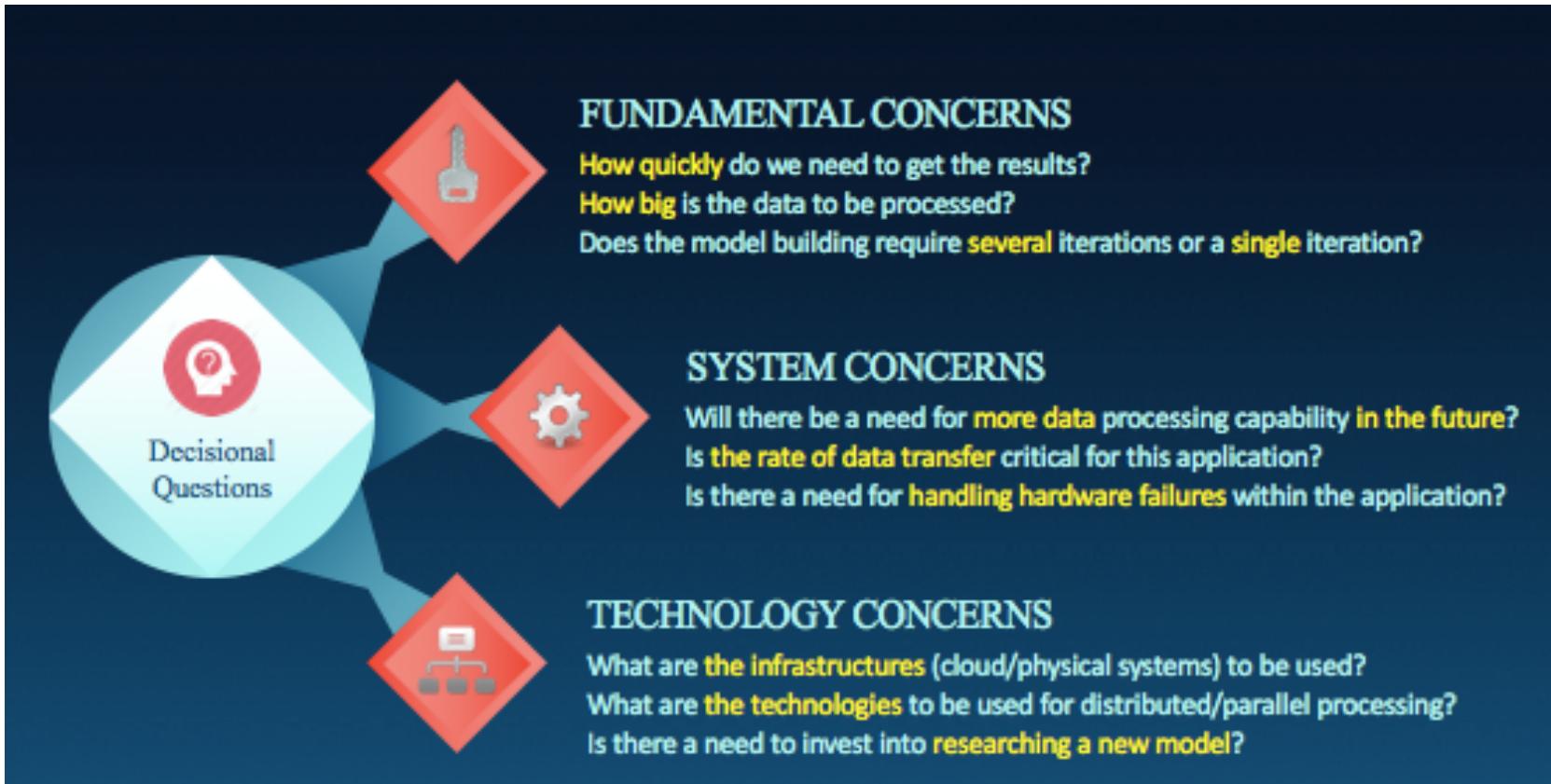


(1919-2013)

# Classification: Train, Validation, Test



# Khía cạnh công nghệ và hệ thống?



Sẽ được chia sẻ trong bài giảng của TS Bùi Hải Hưng và GS Phùng Quốc Định

# Take home message

- Bức tranh tổng thể về khoa học dữ liệu, các khái niệm và nguyên lý cơ bản.
- Khoa học dữ liệu nằm ở trung tâm của công nghệ số, của các đột phá của trí tuệ nhân tạo → vai trò trung tâm trong CMCN4.
- Với từng người: tinh thần cách tân (innovation) đặt ra những việc làm mới và có ý nghĩa cao cần lời giải của phân tích dữ liệu.
- Cơ hội và thách thức của toán học và CNTT của Việt Nam, của giới KH&CN. Cơ hội góp phần vào sự phát triển của đất nước?

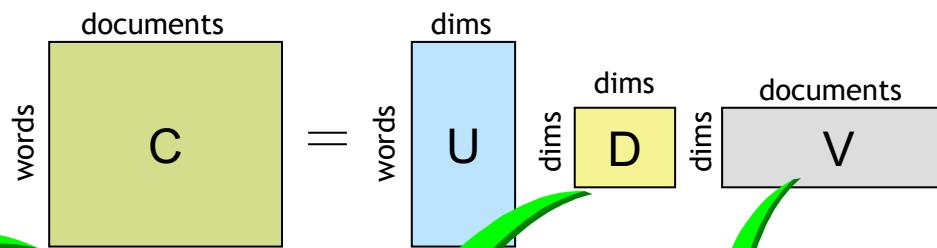
# Latent semantic indexing (LSI)

LSI (Deerwester, 1990) clusters documents in the reduced-dimension semantic space according to word co-occurrence patterns.

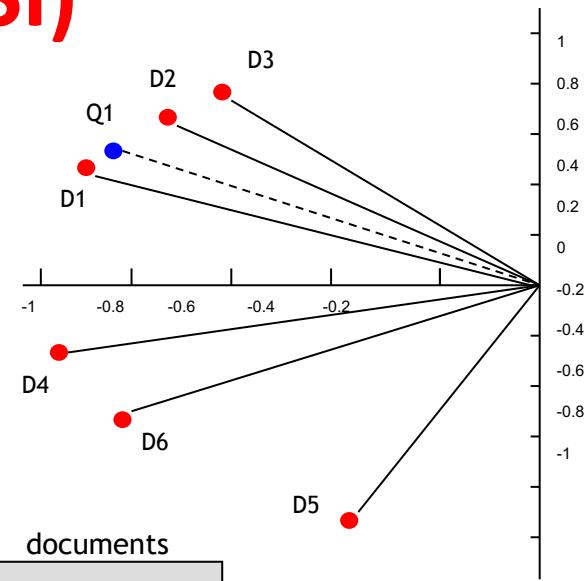
$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$\begin{aligned}\cos(d_3, q_1) &= 0 \\ \cos(d_5, q_1) &= 0 \\ \cos(d_4, q_1) &\neq 0 \\ \cos(d_6, q_1) &\neq 0\end{aligned}$$

	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



# KDD nuggets

*Nguồn thông tin lớn nhất về khai phá dữ liệu*

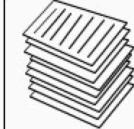
[www.kdnuggets.com](http://www.kdnuggets.com) is website of the data mining community

**KDnuggets™ Data Mining, Analytics, Big Data, and Data Science**

Subscribe to [KDnuggets News](#) | Follow [@KDnuggets](#) [f](#) [in](#) [Contact](#)

search KDnuggets  [Search](#)

Data Mining Software | News | Jobs | Academic | Companies | Courses | Datasets | Data Mining Course | Education | Meetings | Polls | Webcasts

**Simplify your analysis**

**PolyAnalyst 6**  


Swamped with TEXT Data ? Simplify your analysis! PolyAnalyst from Megaputer

**Data Mining, Analytics, Big Data, Data Science**

- Software , Suites , Text , Classification , Visualization
- Jobs in Data Mining, Data Science, Analytics
- Academic / Research positions
- Submit an item for KDnuggets

- Latest KDnuggets News  
Twitter | Facebook | LinkedIn | RSS
- NEW Top stories
- KDnuggets 2015 News
- KDnuggets News Schedule
- Subscribe to KDnuggets News (free bi-weekly newsletter)

**Latest News**

- NYC Data Science Academy Bootcamps, Classes on R, Python, and Machine Learning
- CourseBuffet: Organizing MOOC Courses on Big Data, Data Science, Statistics
- Top stories for Apr 5-11: 10 things statistics taught us about big data analysis; Awesome Public Datasets on GitHub
- Wikibon Big Data Vendor Revenue and Market Forecast, 2020
- NREL: Senior Scientist Computational Statistics

**sas**  
THE POWER TO KNOW.

**Data Management**

What you need to know - and how to use it to get ahead.

[Read the report](#)

  
Get ready TODAY for the data ecosystem of TOMORROW  
REGISTER BY APRIL 9 AND SAVE!  
TDWI Chicago, May 3-8

  
NYC Data Science Bootcamp

**Online Master's Program**  
**PREDICTIVE ANALYTICS**  
NORTHWESTERN UNIVERSITY  
Online MS in Predictive 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-conductor manufacturing - Video content retrieval - Gene expression - Protein-protein interaction

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