

HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Introduction to Machine Learning and Data Mining

IT3190

Lecture: Introduction

ONE LOVE. ONE FUTURE.

Contents

- **Lecture 1: Introduction to Machine Learning & Data Mining**
- Lecture 2: Data crawling and pre-processing
- Lecture 3: Linear regression
- Lecture 4+5: Clustering
- Lecture 6: Decision tree and Random forest
- Lecture 7: Neural networks
- Lecture 8: Support vector machines
- Lecture 9: Performance evaluation
- Lecture 10: Probabilistic models
- Lecture 11: Basics of data mining
- Lecture 12: Association rule mining
- Lecture 13: Regularization and advanced topics

Why ML & DM?

- “The most important general-purpose technology of our era is artificial intelligence, particularly **machine learning**” – Harvard Business Review
<https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence>
- A huge demand on Data Science
- “Data scientist: the sexiest job of the 21st century” – Harvard Business Review.
<http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>
- “The Age of Big Data” – The New York Times
http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0



Data Analyst

San Francisco Bay Area
Posted 18 days ago

Data Analyst

Greater New York City Area
Posted 25 days ago

Statistical Analyst - Data...

Greater New York City Area
Posted 9 hours ago

Data Analyst

Greater New York C
Posted 15 days ago

DATA SCIENTIST

Greater New York C
Posted 25 days ago



Data Scientist

Greater New York C
Posted 14 days ago

Marketing Analytics Associate

Greater New York City Area
Posted 24 days ago



Financial Data Analyst

Greater New York City Area
Posted 20 days ago

Data Analyst...

Greater New York City Area
Posted 13 days ago



Home

Profile

Network

Jobs

Interests



Data Analyst

Amazon - Newark, NJ

Posted 24 days ago

[Apply on company website](#)

[Save](#)

Senior Data Analyst - Big Data, Meta Product

TripAdvisor - Newton, MA

Posted 12 days ago

[Apply now](#)

[Save](#)

Home

Profile

Network

Jobs

Interests



Data Analyst

Apple - Daly City - California -US

Posted 18 days ago

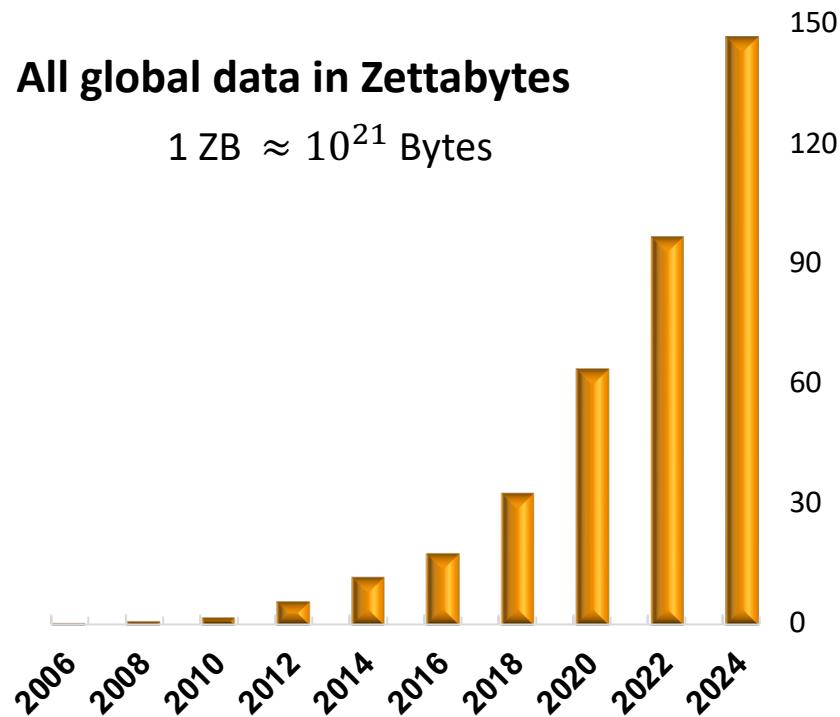
[Apply on company website](#)

[Save](#)

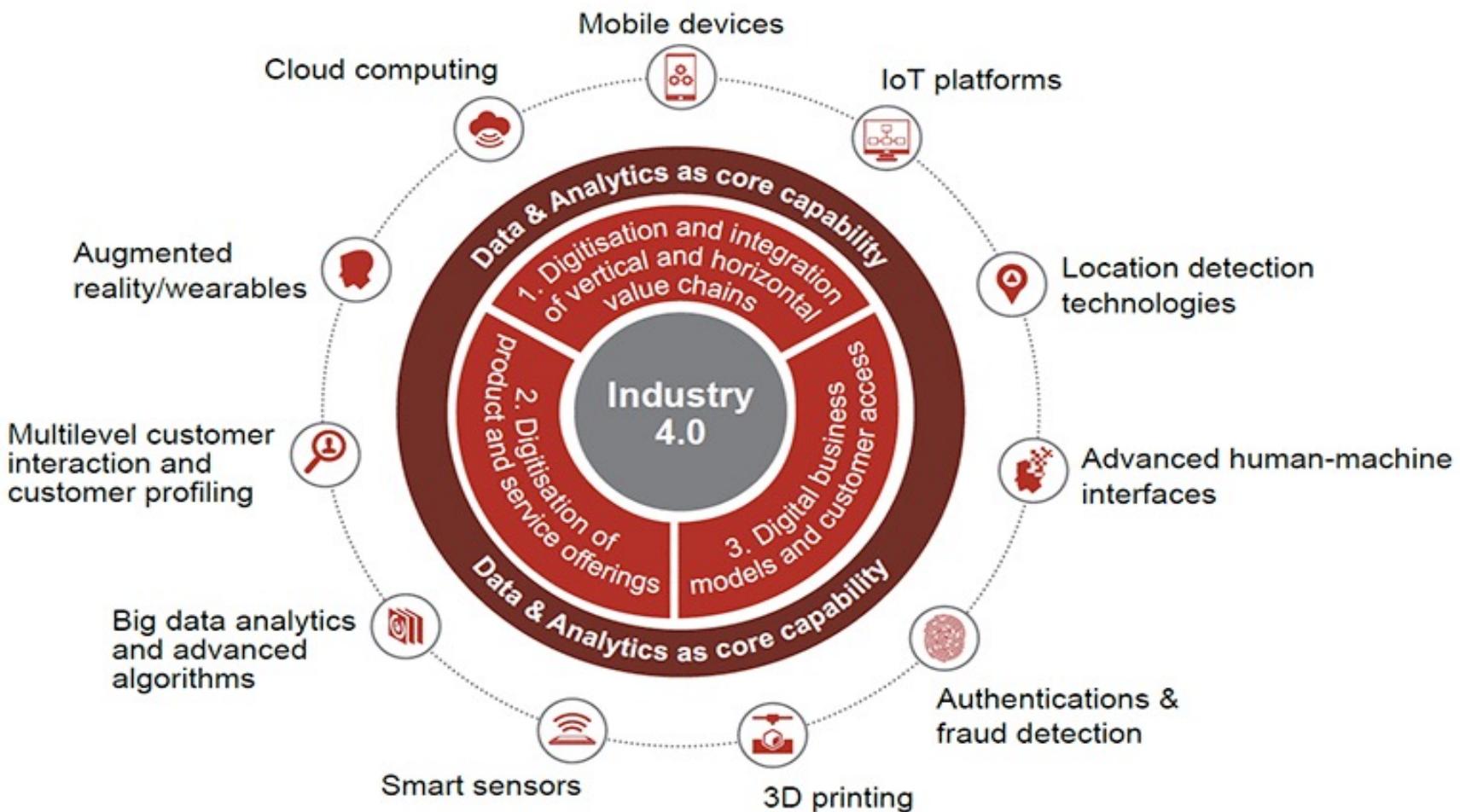


Why ML & DM?

- Data mining, inference, prediction
- ML & DM provides an efficient way to make intelligent systems/services.
- ML provides vital methods and a foundation for Big Data.

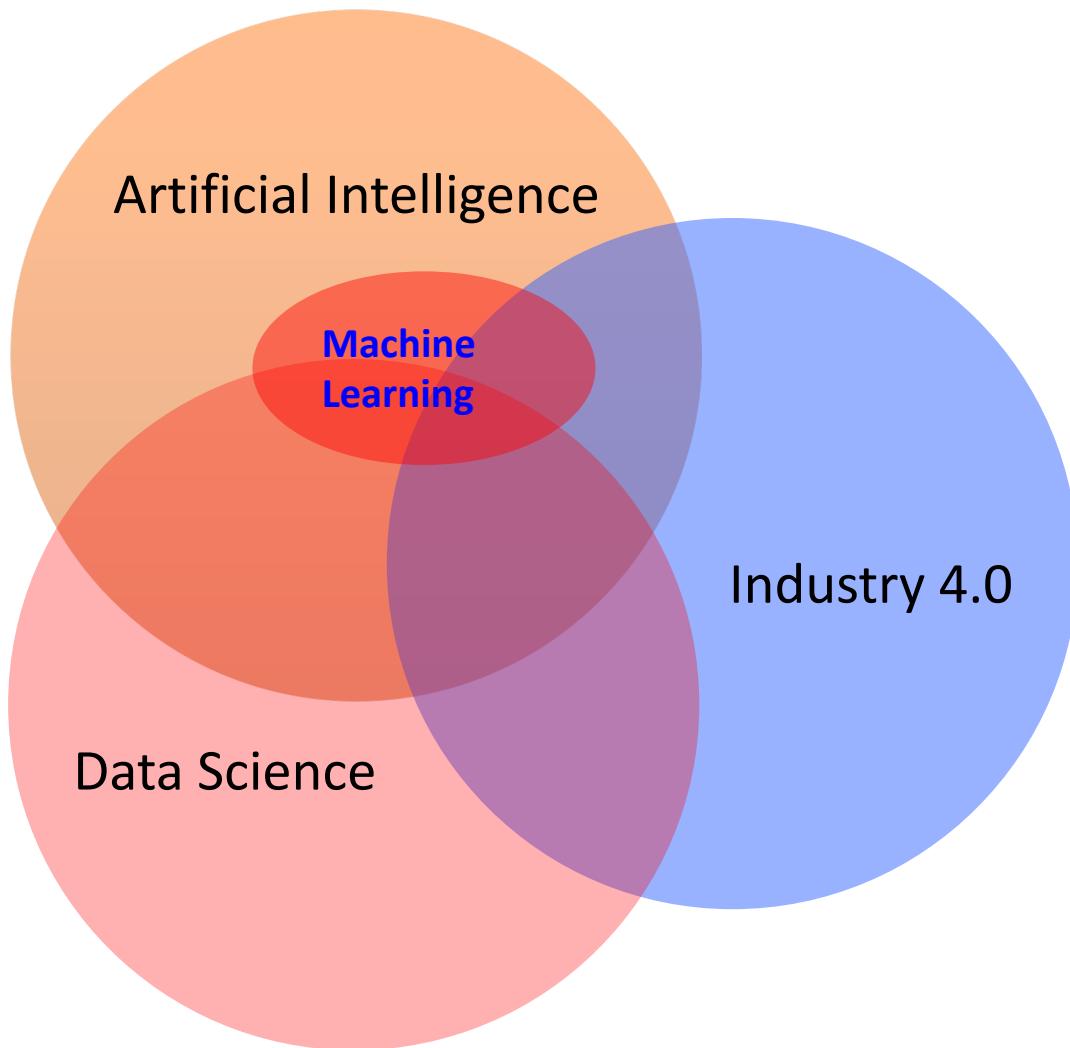


Why? Industry 4.0



<https://www.pwc.com/ca/en/industries/industry-4-0.html>

Why? AI & DS & Industry 4.0



Some successes: IBM's Watson



IBM's Watson Supercomputer Destroys Humans in Jeopardy (2011)

Some successes: Amazon's secret



"The company reported a **29% sales increase** to \$12.83 billion during its second fiscal quarter, up from \$9.9 billion during the same time last year."
– Fortune, July 30, 2012

Lower Priced Items to Consider

	LG 34UM68-P 34-Inch 21:9...		LG 27UD68-P 27-Inch 4
	5 star 164		5 star 54
\$389.89	✓Prime	\$439.00	✓Prime

Is this feature helpful?

LG 34UC98-W 34-Inch 2 UltraWide QHD IPS Mon Thunderbolt
by LG Electronics 131 customer review | 101 answered questions

Available from these sellers.

Style: Thunderbolt

Customers Who Bought This Item Also Bought

	Cable Matters Thunderbolt 2 Cable in White 6.6 Feet / 2m 10		Cable Matters Thunderbolt 2 Cable in Black 6.6 Feet / 2m 38		Cable Matters Thunde 2 Cable in White 3.3 F 1m 38
--	--	--	--	--	--

\$38.99 ✓Prime

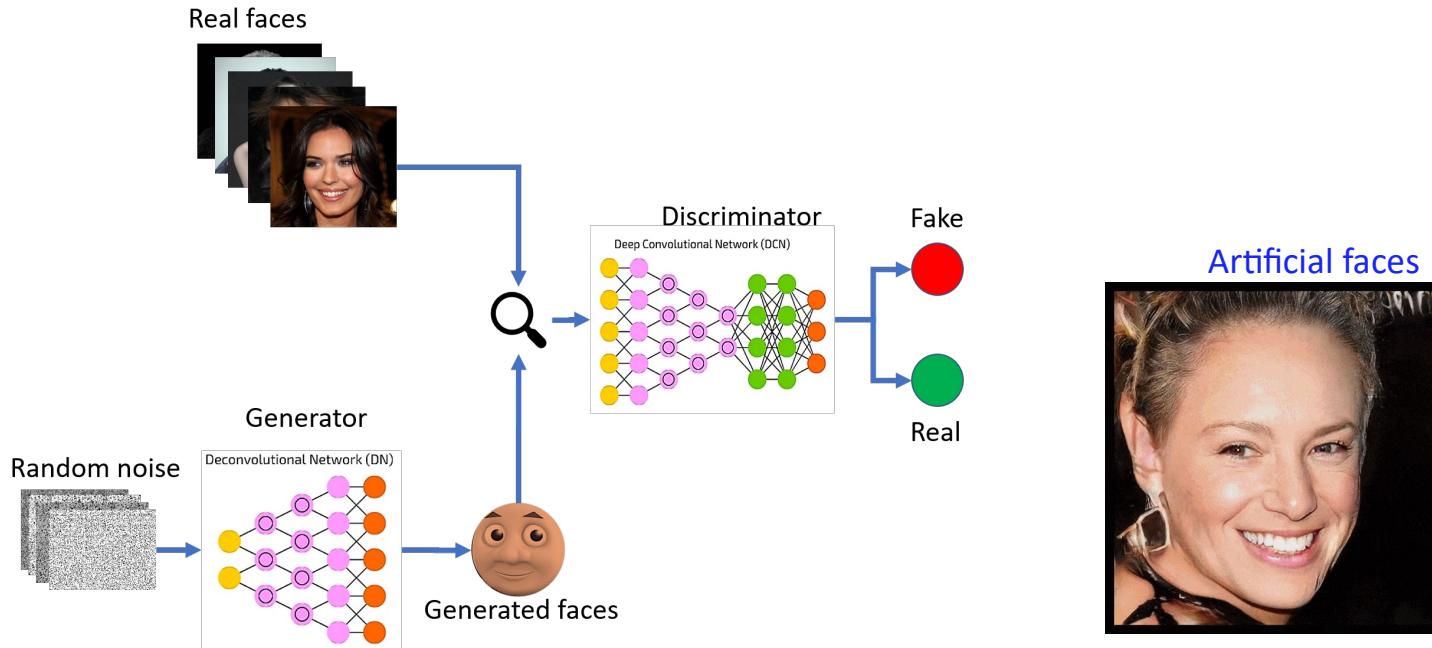
\$31.99 ✓Prime

Some successes: GAN (2014)

❖ Tạo Trí tưởng tượng (Imagination)

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))]$$

Ian Goodfellow



Artificial faces



Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *NIPS*, pp. 2672-2680. 2014.

Some successes: AlphaGo (2016)

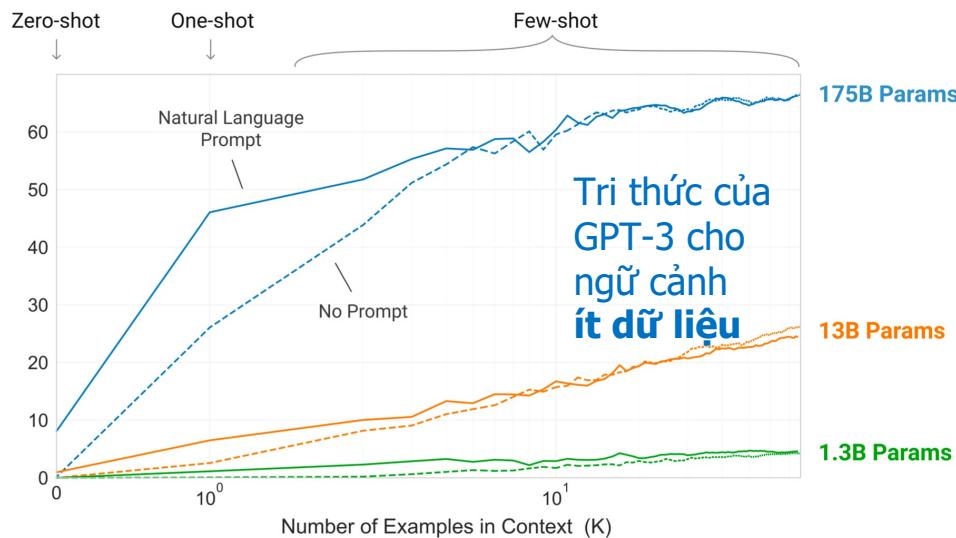
- AlphaGo of Google the world champion at Go (cờ vây), 3/2016
 - Go is a 2500 year-old game.
 - Go is one of the most complex games.
- AlphaGo learns from 30 millions human moves, and plays itself to find new moves.
- It beat Lee Sedol (World champion)
 - <http://www.wired.com/2016/03/two-redefined-future/>
 - <http://www.nature.com/news/google-game-of-go-1.19234>



Some successes: GPT-3 (2020)

■ Language generation (writing ability?)

- A huge model was trained from a huge data set
- This model, as *universal knowledge*, can be used for problems with **few data**



Tri thức của
GPT-3 cho
ngữ cảnh
ít dữ liệu

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

	Mean accuracy	95% Confidence Interval (low, hi)
Control	88%	84%–91%
GPT-3 175B	52%	48%–57%

Con người không
thể nhận diện bài
viết 500 từ là do
máy hay người viết

Some successes: Text-to-image (2022)



Imagen

A bowl of soup

Some pictures generated from short descriptions

Some successes: ChatGPT (2022)

- Human-level Chatting, Writing, QA,...

Why ChatGPT is
about to change
how you work, like
it or not?

- Forbes, Feb. 2, 2023



Audio Live TV Log In

ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business

Updated 1:35 PM EST, Thu January 26, 2023

trời đẹp quá, làm cho tớ bài thơ đi



Tạm biệt mây mù, đón tới ánh sáng

Trời đẹp mới rực, tấm lòng tớ tăng

Sức mạnh của mùa xuân, giữa cánh đồng hoa

Vẻ đẹp tự nhiên, làm cho tớ mãi mê hoặc.

Lấp lánh tràn đầy niềm vui, nụ cười tràn đầy niềm hạnh phúc

Không cần bất kỳ điều gì, chỉ cần trời đẹp để tớ hạnh phúc.

Success of ChatGPT

Machine Learning vs Data Mining

- Machine Learning
(ML - Học máy)

To build computer systems that can improve themselves by learning from data.

(Xây dựng những hệ thống mà có khả năng tự cải thiện bản thân bằng cách học từ dữ liệu.)

- Some venues: NeurIPS, ICML, IJCAI, AAAI, ICLR, ACML, ECML

- Data Mining
(DM - Khai phá dữ liệu)

To find new and useful knowledge from datasets.

(Tìm ra/Khai phá những tri thức mới và hữu dụng từ các tập dữ liệu lớn.)

- Some venues: KDD, PKDD, PAKDD, ICDM, CIKM

Data

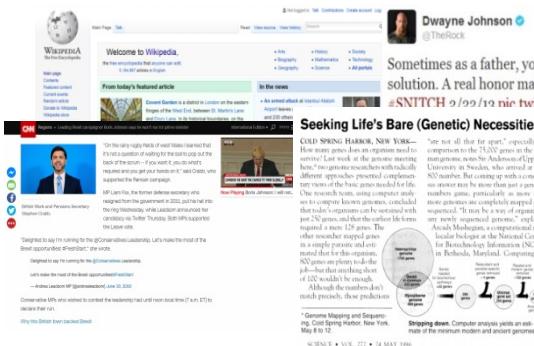
Structured – relational (table-like)

A	B	C	D	E	F	G
1 Country	Region	Population	Under15	Over60	Fertil	LifeExp
2 Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3 Zambia	Africa	14075	46.73	3.95	5.77	55
4 Yemen	Eastern M	23852	40.72	4.54	4.35	64
5 Viet Nam	Western P	90796	22.87	9.32	1.79	75
6 Venezuela (Bo Americas)	29955		28.84	9.17	2.44	75
7 Vanuatu	Western P	247	37.37	6.02	3.46	72
8 Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9 Uruguay	Americas	3395	22.05	18.59	2.07	77

Un-structured

```
{  
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",  
  "title": "[Updating] Câu chuyện xuyên mưa về :  
  "url": "http://techtalk.vn/updating-cau-chuyen-",  
  "labels": "techtalk/Cong nghe",  
  "content": "Vào chiều tối ngày 09/12/2016 vừa  
  "image_url": "",  
  "date": "2016-12-10T03:51:10Z"  
}
```

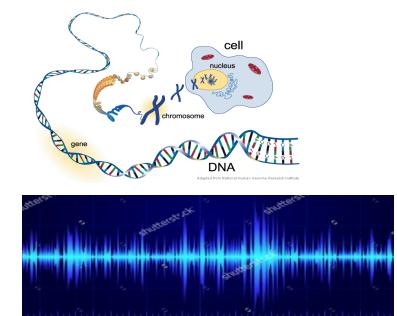
texts in websites, emails, articles, tweets



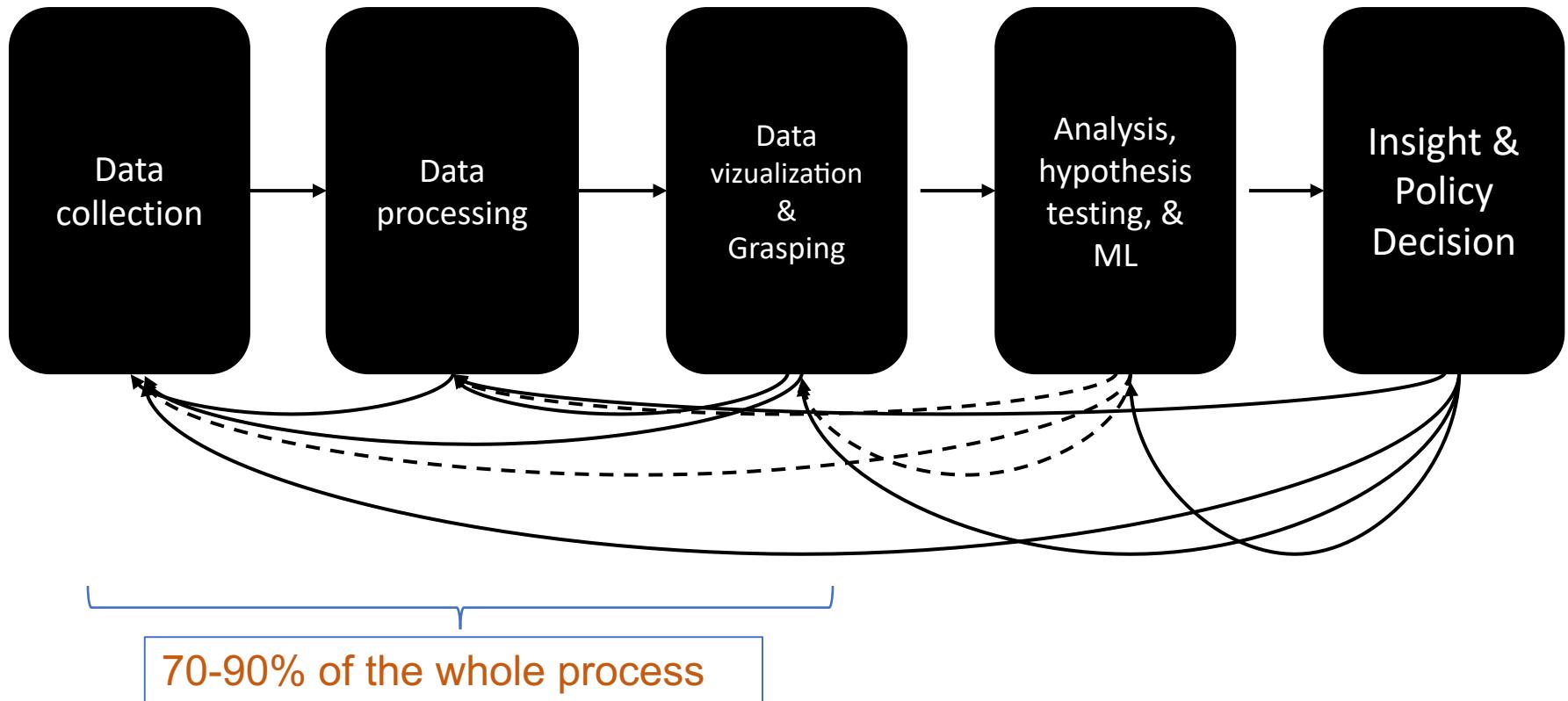
2D/3D images, videos + meta



spectrograms, DNAs, ...

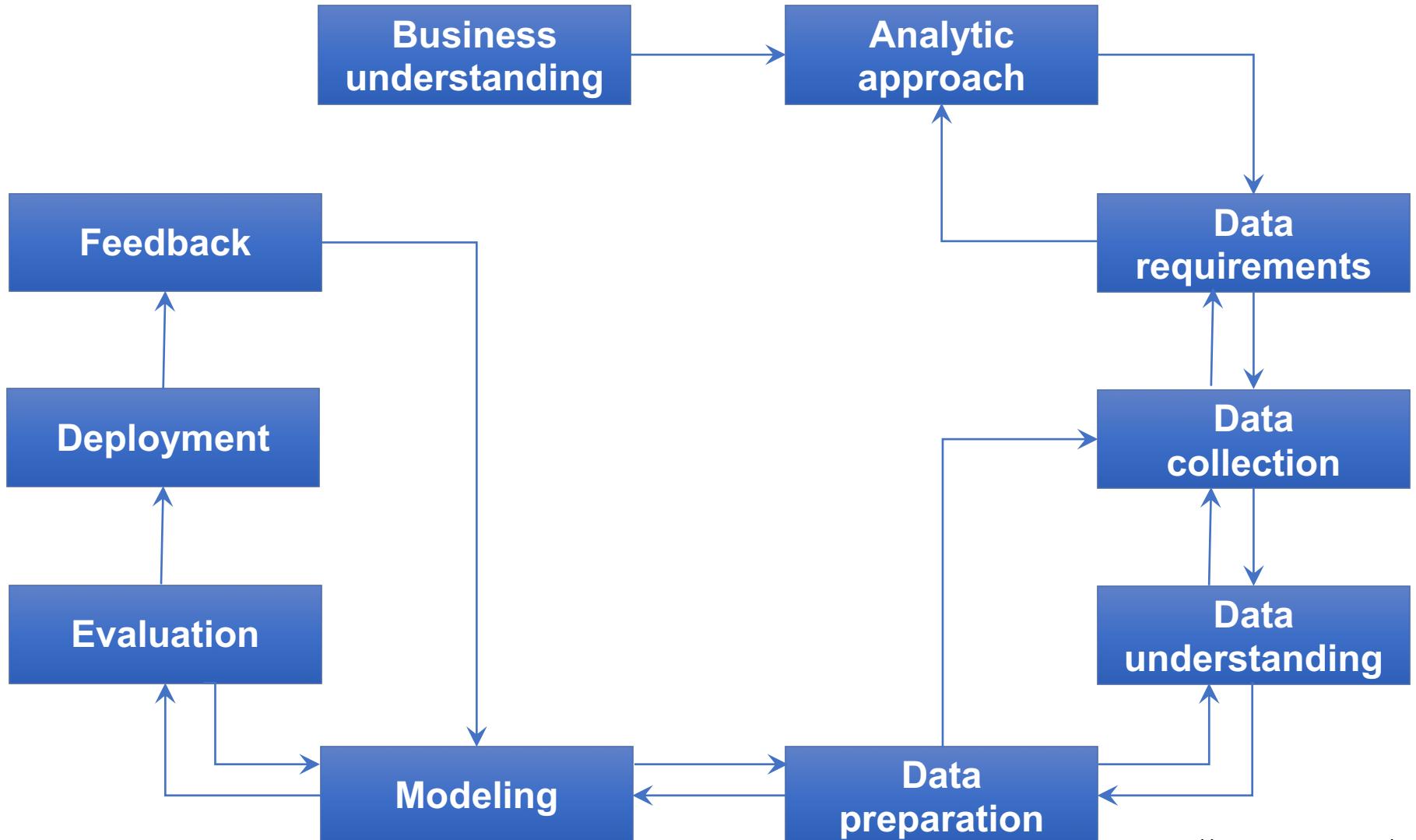


Methodology: insight-driven



(John Dickerson, University of Maryland)

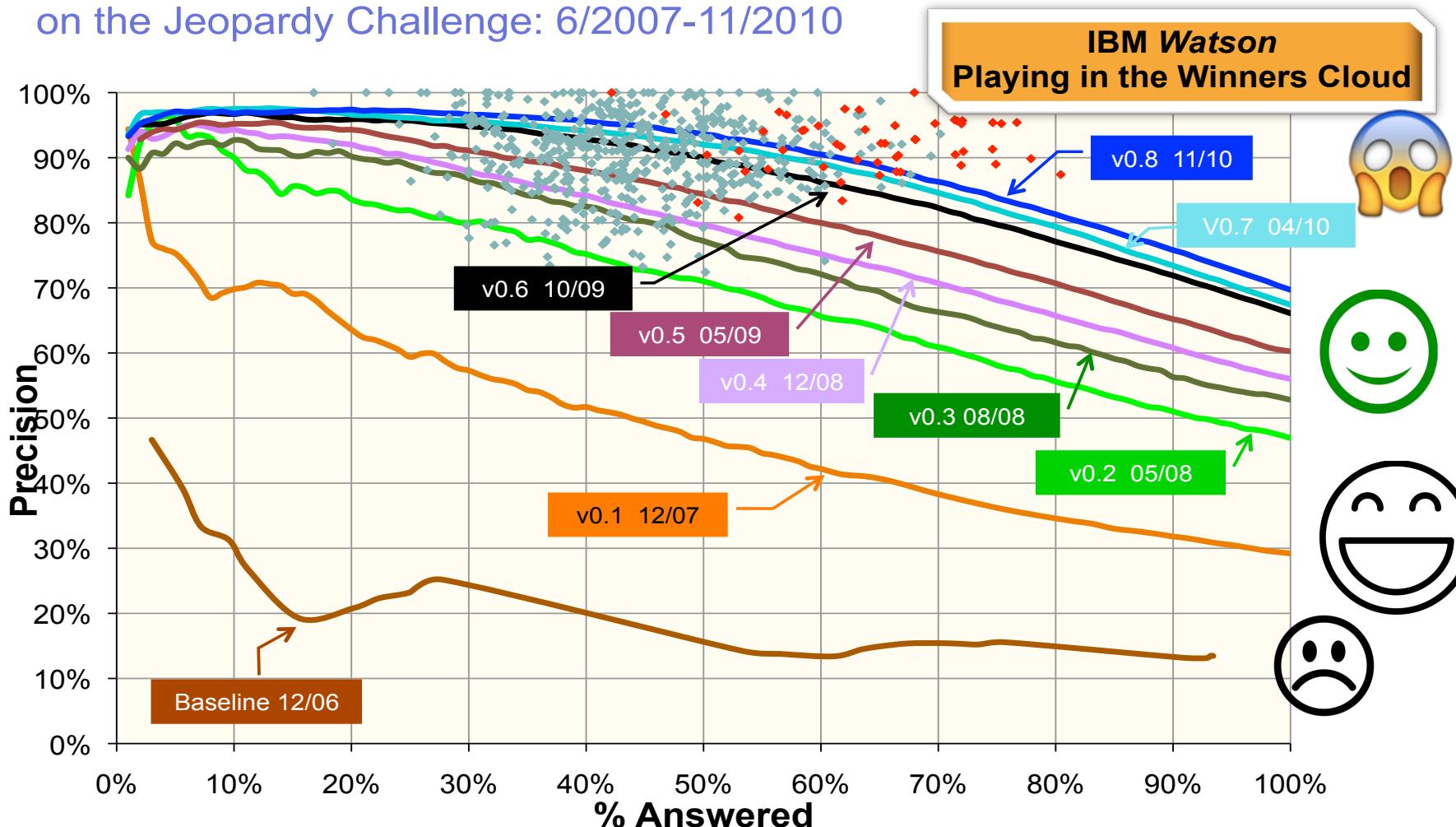
Methodology: product-driven



(<http://www.theta.co.nz/>)

Product development: experience

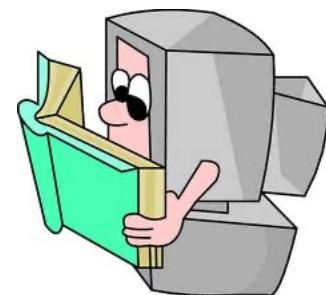
DeepQA: Incremental Progress in Answering Precision
on the Jeopardy Challenge: 6/2007-11/2010



© 2011 IBM Corporation

What is Machine Learning?

- Machine Learning (ML) is an active subfield of Artificial Intelligence.
- ML seeks to answer the question [Mitchell, 2006]
 - *How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?*
- Some other views on ML:
 - Build systems that automatically improve their performance [Simon, 1983].
 - Program computers to optimize a performance objective at some task, based on data and past experience [Alpaydin, 2020]

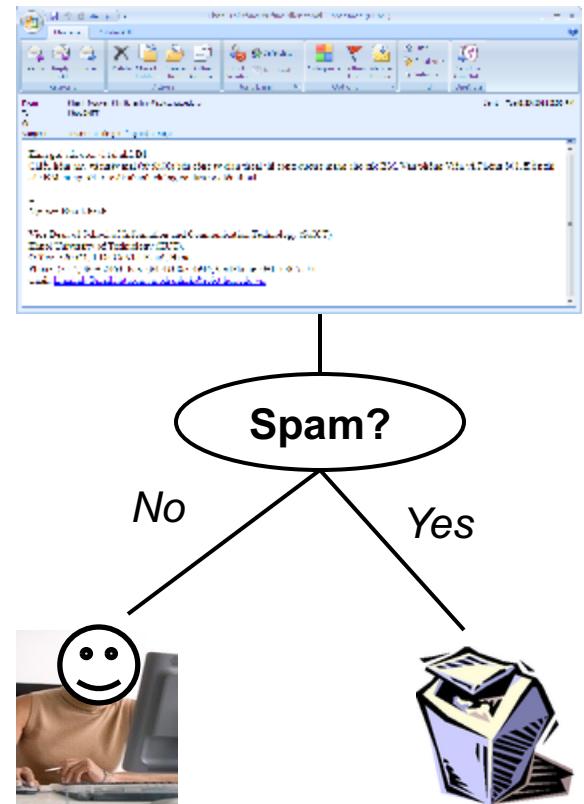


A learning machine

- We say that a machine *learns* if the system reliably improves its performance **P** at task **T**, following experience **E**.
- A *learning problem* can be described as a triple **(T, P, E)**.
- ML is close to and intersects with many areas.
 - Computer Science,
 - Statistics, Probability,
 - Optimization,
 - Psychology, Neuroscience,
 - Computer Vision,
 - Economics, Biology, Bioinformatics, ...

Some real examples (1)

- Spam filtering for emails
 - **T**: filter/predict all emails that are spam.
 - **P**: the accuracy of prediction, that is the percentage of emails that are correctly classified into normal/spam.
 - **E**: set of old emails, each with a label of spam/normal.



Some real examples (2)

- **Image captioning**

- **T:** give some words to describe the meaning of a picture
- **P:** ?
- **E:** set of pictures, each has a short description



lychee-inspired spherical chair



a girl giving cat a gentle hug



a small hedgehog holding a piece of watermelon

A set of images with their labels. [DALL-E 3]

What does a machine learn?

- A **mapping** (function):

$$f : x \mapsto y$$

- x: observations (data), past experience
 - y: prediction, new knowledge, new experience,...

- A **model** (mô hình)

- Data are often supposed to be generated from an unknown model.
(Dữ liệu thường được tạo ra bởi một mô hình nào đó)
 - Learning a model means learning the parameters of that model.
(Học một mô hình có nghĩa là học/tìm những tham số của mô hình đó)

Where does a machine learn from?

- Learn from a set of training examples (**training set**, tập học, tập huấn luyện) { $\{x_1, x_2, \dots, x_N\}$; $\{y_1, y_2, \dots, y_M\}$ }
 - x_i is an observation (quan sát, mẫu, điểm dữ liệu) of x in the past.
 - y_j is an observation of y in the past, often called *label* (nhãn) or *response* (phản hồi) or *output* (đầu ra).
- After learning:
 - We obtain a model, new knowledge, or new experience (f).
 - We can use that model/function to do **prediction** or **inference** for future observations, e.g.,

$$y = f(x)$$

Two basic learning problems

- There is an *unknown* function y^* that maps each x to a number $y^*(x)$
 - In practice, we can collect some pairs: (x_i, y_i) , where $y_i = y^*(x_i)$
- **Supervised learning (học có giám sát):** find a function $y = f(x)$ from a given training set $\{x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N\}$ so that $y_i \approx f(x_i)$ for every i .
 - **Classification** (categorization, phân loại, phân lớp): if y only belongs to a discrete set, for example {spam, normal}
 - **Regression** (hồi quy): if y is a real number

Supervised learning: classification



BIRDS NEST TREE

Supervised learning: Regression

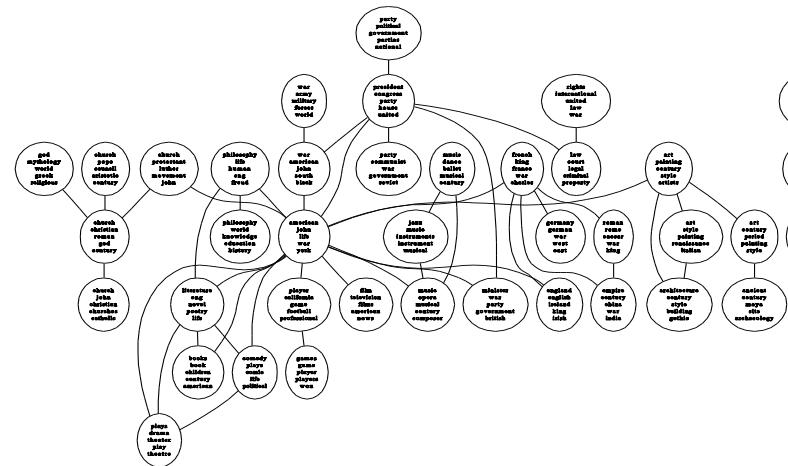
- Prediction of stock indices

1	10.00	9	29.99	29.13	23.99		
2	10.24	7	62.31	62.00	75.64	-7.35	
3	10.52	7	62.31	62.00	75.64	-7.35	
4	10.80	3	34.26	34.75	43.92	-6.75	
5	10.86	34	75.86	75.33	25.09	+8.33	
6	10.86	46.34	12.26	12.25	12.45	-4.25	
7	10.49	88.90	12	435.86	435.63	126.58	+6.63
8	35.63	34.75	1	54.23	54.33	54.18	-8.33
9	21.87	75.33	7	46.32	46.34	23.64	+1.34
10	99.12	12.25	45	88.54	88.90	64.15	+2.98
11	3.43	35.63	6	43.45	43.66	43.62	-1.66
12	25	21.87	45	12.23	12.86	75.21	+4.86
13	10	89.12	7	434.64	434.49	632.55	-7.49
14	7	23.43	34	32.21	32.00	12.21	-3.80
15	6	65.25	5	65.75	65.22	23.46	+8.01
16	5	42.96	12	123.74	123.76	121.51	-8.01



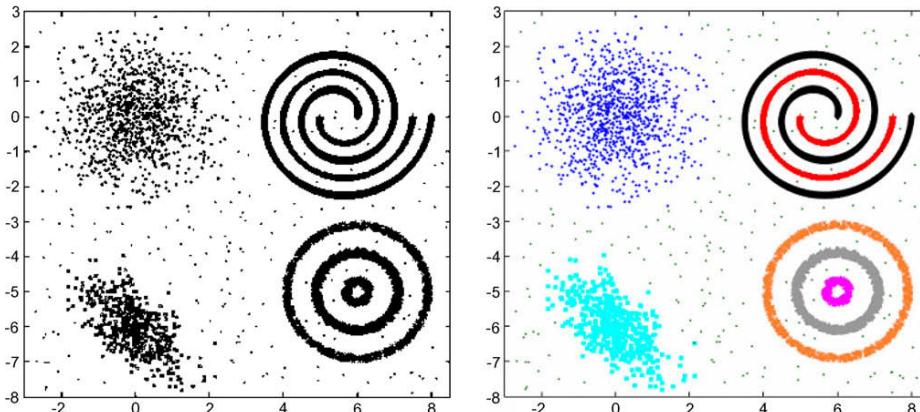
Two basic learning problems

- **Unsupervised learning (học không giám sát):** learn a function $y = f(x)$ from a given training set $\{x_1, x_2, \dots, x_N\}$.
 - y can be a data cluster
 - y can be a hidden structure
 - y can be a trend
 - Other: semi-supervised learning, reinforcement learning, ...

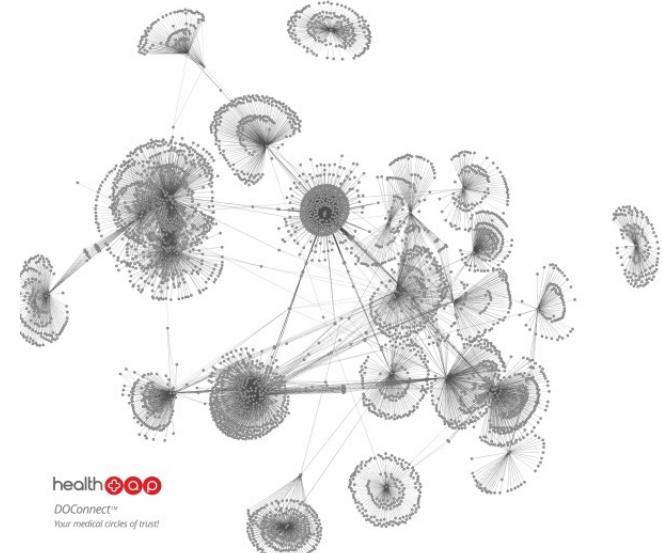


Unsupervised learning: examples (1)

- Clustering data into clusters
 - Discover the data groups/clusters

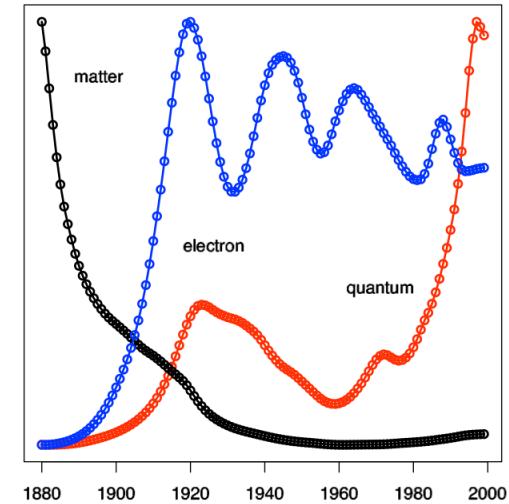


- Community detection
 - Detect communities in online social networks



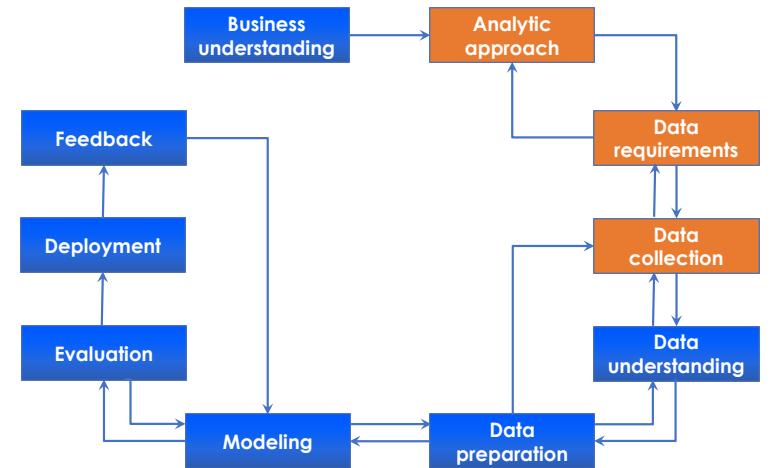
Unsupervised learning: examples (2)

- Trends detection
 - Discover the trends, demands, future needs of online users



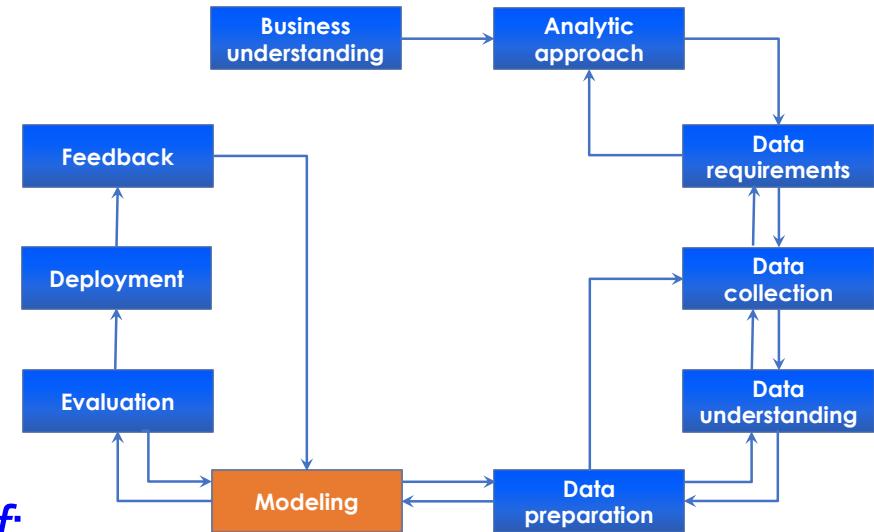
Design a learning system (1)

- Some issues should be carefully considered when designing a learning system.
- Determine the type of the function to be learned
(Determine the learning problem)
 - $y^*: X \rightarrow \{0,1\}$
 - $y^*: X \rightarrow \text{set of labels/tags}$
 - $y^*: X \rightarrow \mathbb{R}$
- Select a training set:
 - The training set plays the key role in the effectiveness of the system.
 - Do the observations have any label?
 - The training observations should characterize the whole data space
→good for future predictions.



Design a learning system (2)

- Select a representation or approximation (model) f for the unknown function y^* (Lựa chọn dạng hàm f để đi xấp xỉ hàm y^* chưa biết)
 - Linear model?
 - A neural network?
 - A decision tree? ...
- Select a good algorithm to learn f :
 - Ordinary least square? Ridge regression?
 - Back-propagation?
 - ID3?



ML: some issues (1)

- **Learning algorithm**
 - Under what conditions the chosen algorithm will (asymptotically) converge? (với điều kiện nào thì thuật toán học sẽ hội tụ?)
 - For a given application/domain and a given objective function, what algorithm performs best? (Đối với một ứng dụng và mục tiêu cho trước, thuật toán nào sẽ tốt nhất?)
- *No-free-lunch theorem* [Wolpert and Macready, 1997]: if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.
 - *No algorithm can beat another on all domains.*
(không có thuật toán nào luôn hiệu quả nhất trên mọi miền ứng dụng)

ML: some issues (2)

- **Training data**
 - *How many observations* are enough for learning?
 - Whether or not does the *size of the training set* affect performance of an ML system?
 - What is the effect of the *disrupted* or *noisy* observations?

ML: some issues (3)

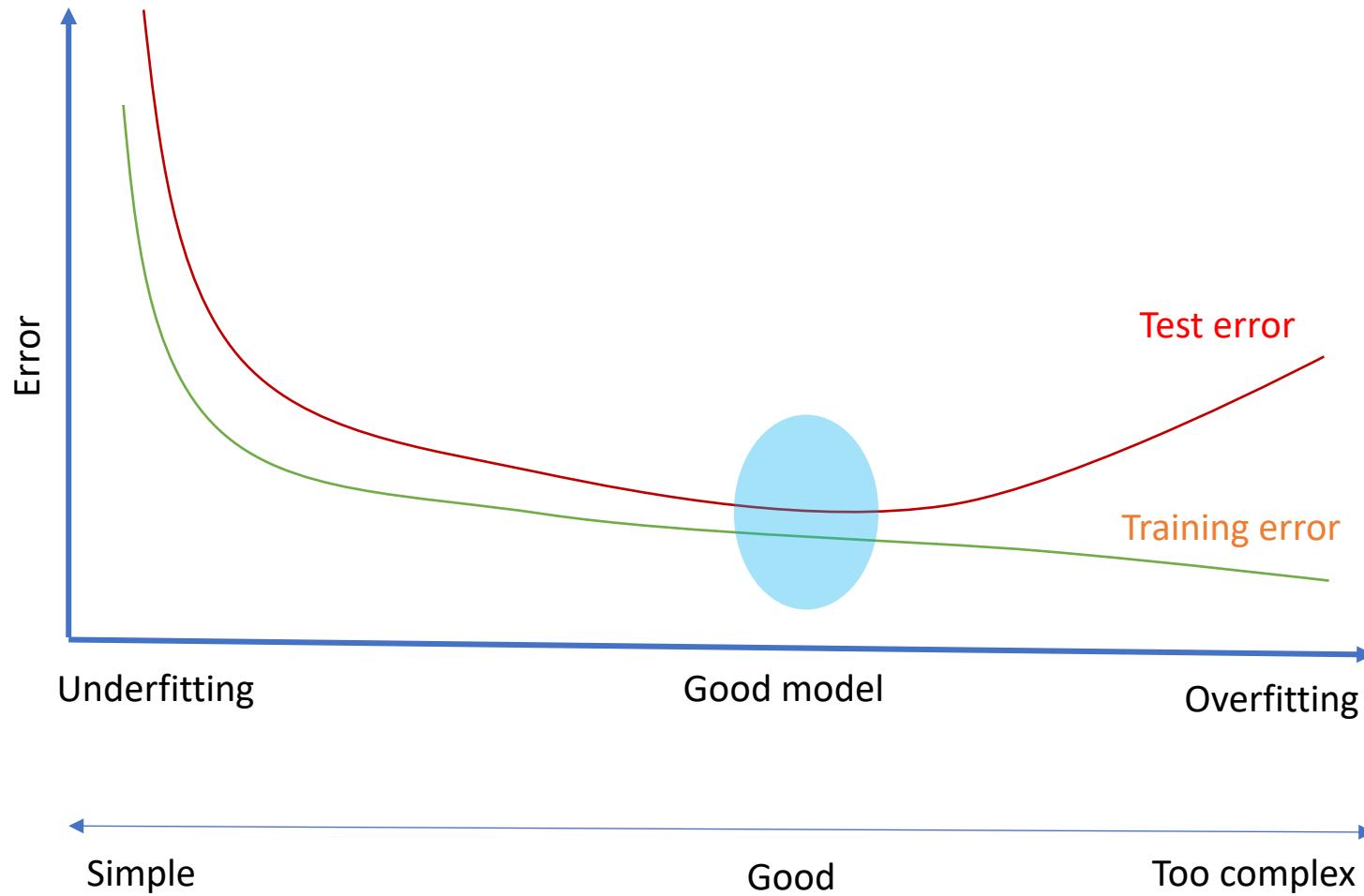
- **Learnability:**

- The goodness/limit of the learning algorithm?
- What is the **generalization** (tổng quát hoá) of the system?
 - ✧ Predict well new observations, not only the training data.
 - ✧ Avoid overfitting.

Overfitting (quá khớp, quá khít)

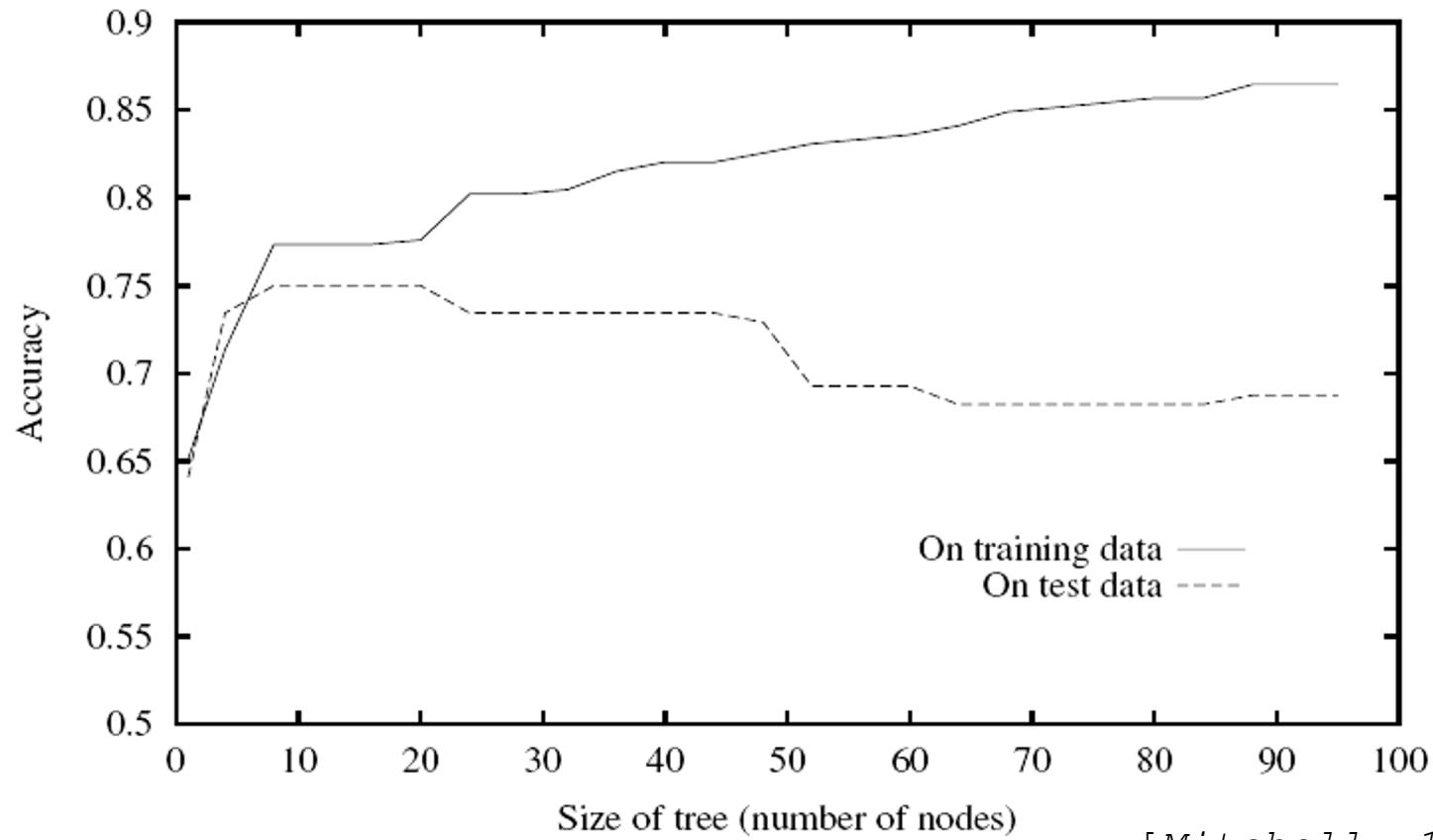
- Function h is called *overfitting* [Mitchell, 1997] if there exists another function g such that:
 - g might be worse than h for the training data, but
 - g is better than h for future data.
- A learning algorithm is said to overfit relative to another one if it is *more accurate in fitting* known data, but *less accurate in predicting* unseen data.
- Overfitting is caused by many factors:
 - The function/model is **too complex** or have too much parameters.
 - **Noises or errors** are present in the training data.
 - The training size is **too small**, not characterizing the whole space.

Overfitting



Overfitting: example

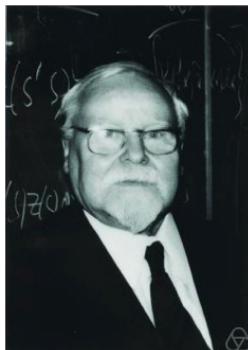
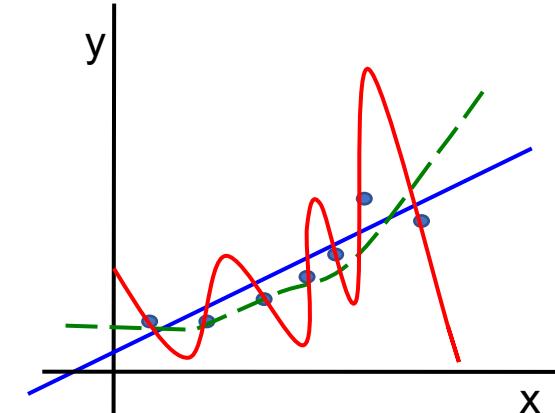
- Increasing the size of a decision tree can degrade prediction on unseen data, even though increasing the accuracy for the training data.



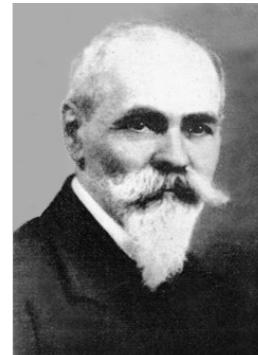
[Mitchell, 1997]

Overfitting: Regularization

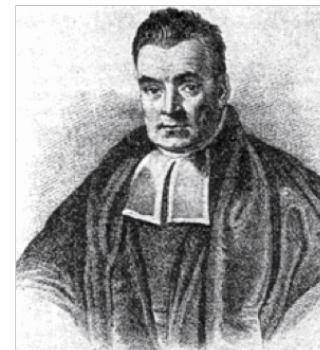
- Among many functions, which one can generalize best from the given training data?
 - *Generalization is the main target of ML.*
 - Predict unseen data well.
- **Regularization:** a popular choice



Tikhonov,
smoothing an ill-
posed problem



Zaremba, model
complexity
minimization



Bayes: priors
over parameters



Andrew Ng: need no
maths, but it prevents
overfitting!

(Picture from <http://towardsdatascience.com/multitask-learning-teach-your-ai-more-to-make-it-better-dde116c2cd40>)

References

- Alpaydin E. (2020). Introduction to Machine Learning. The MIT Press.
- Mitchell, T. M. (1997). Machine learning. *McGraw Hill*.
- Mitchell, T. M. (2006). *The discipline of machine learning*. Carnegie Mellon University, School of Computer Science, Machine Learning Department.
- Simon H.A. (1983). Why Should Machines Learn? In R. S. Michalski, J. Carbonell, and T. M. Mitchell (Eds.): Machine learning: An artificial intelligence approach, chapter 2, pp. 25-38. Morgan Kaufmann.
- Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134-1142.
- Wolpert, D.H., Macready, W.G. (1997), "[No Free Lunch Theorems for Optimization](#)", *IEEE Transactions on Evolutionary Computation* 1, 67.

A large, semi-transparent watermark of the HUST logo is positioned in the background of the slide. The logo consists of the letters "HUST" in a white, bold, sans-serif font, with a red circular arrow graphic to its right.

HUST

THANK YOU !