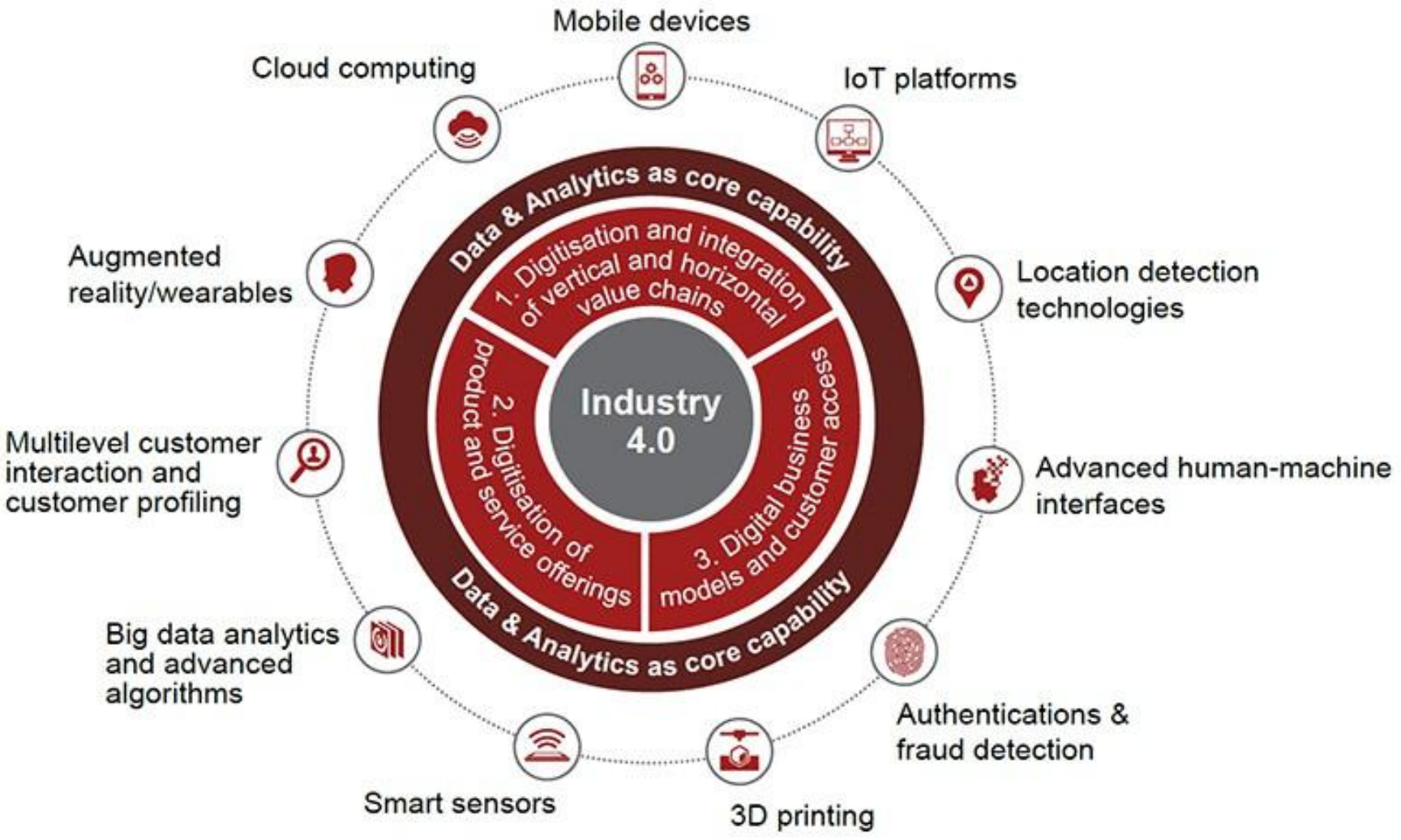
Nội dung môn học

- **Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu**
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- Lecture 9: Học dựa trên xác suất
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp
- Lecture 13: Thảo luận ứng dụng học máy và khai phá dữ liệu trong thực tế

Tại sao nên biết Học Máy & Khai phá dữ liệu?

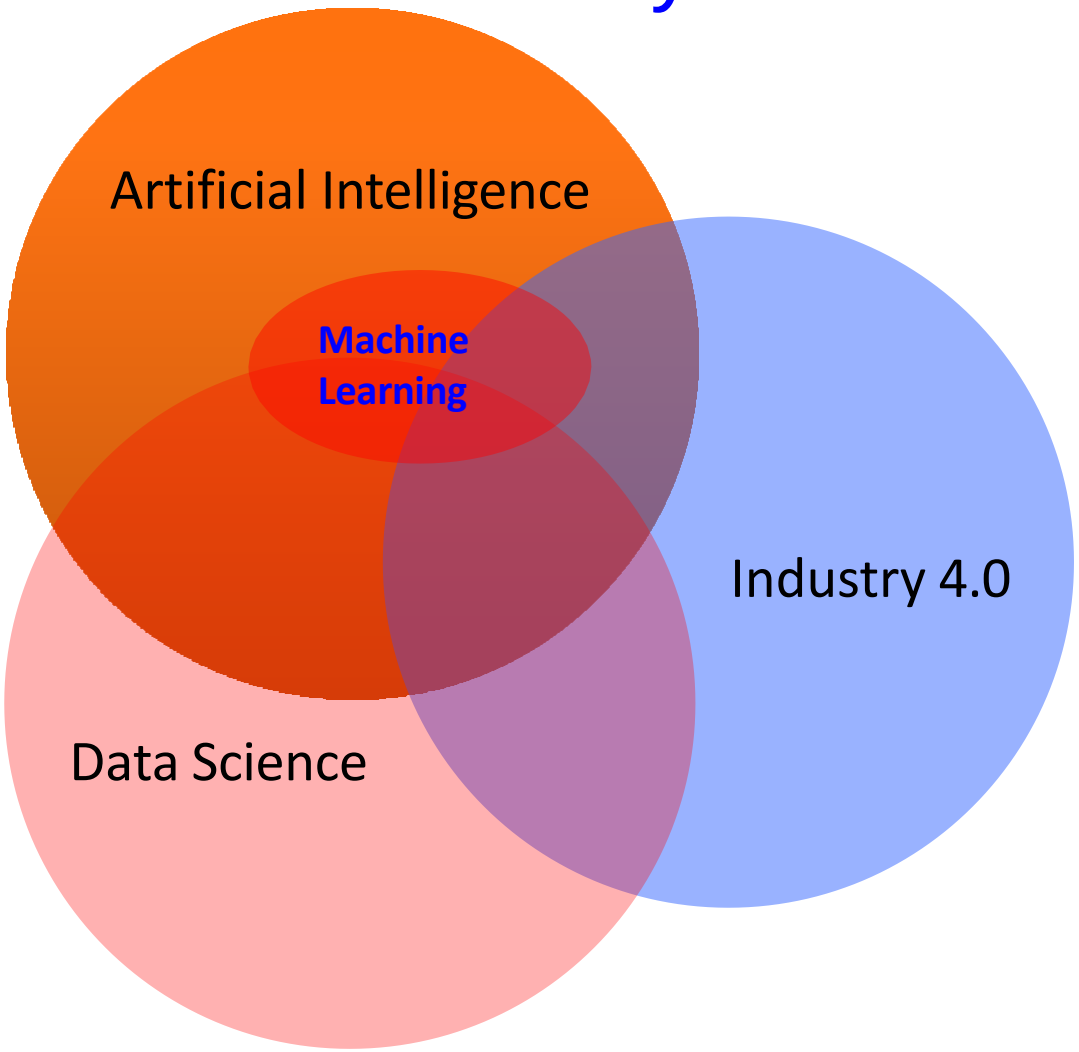
- "The most important general-purpose technology of our era is artificial intelligence, particularly **machine learning**" - Harvard Business Review
  https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence
- Nhu cầu lớn về Data Science
- "Data scientist: the sexiest job of the 21st 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

Tại sao? Industry 4.0



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

Tại sao? AI & DS & Industry 4.0



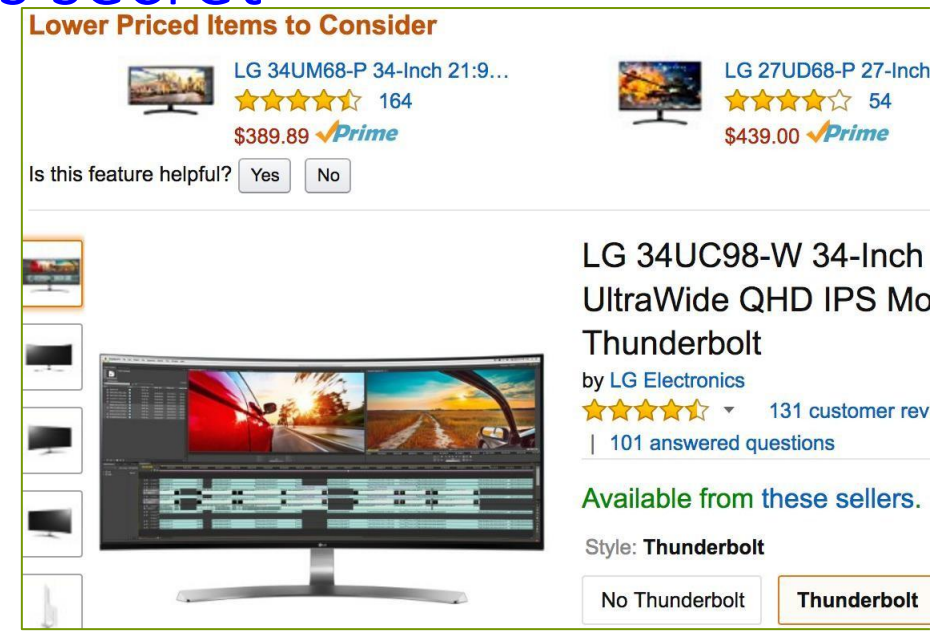
Vài thành công: IBM’s Watson

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

Vài thành công: Amazon's secret

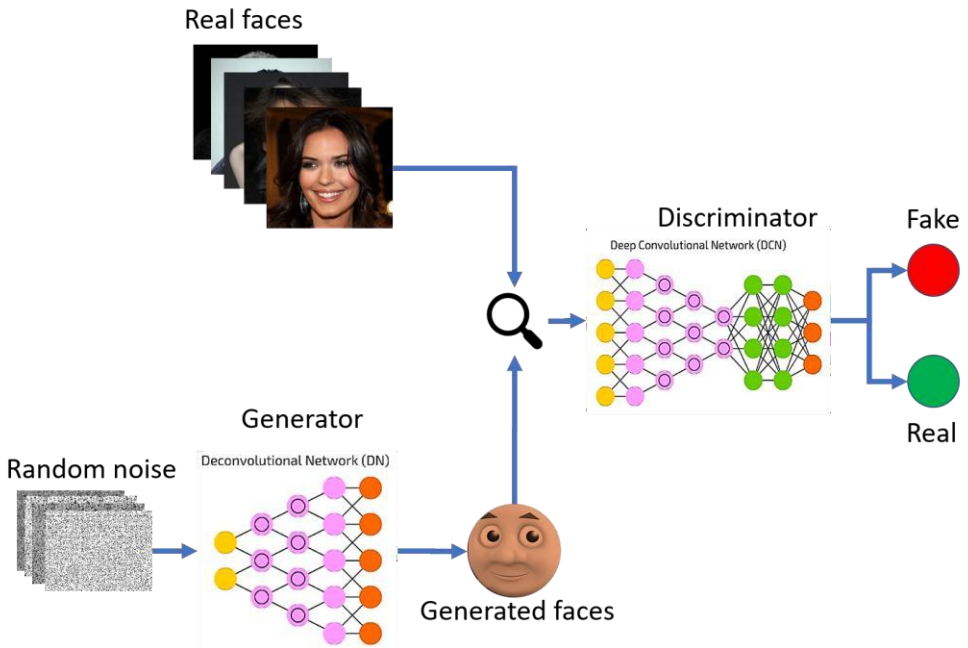
amazon

"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



Vài thành công: GAN (2014)

- Tạo **Trí tưởng tượng** (Imagination)

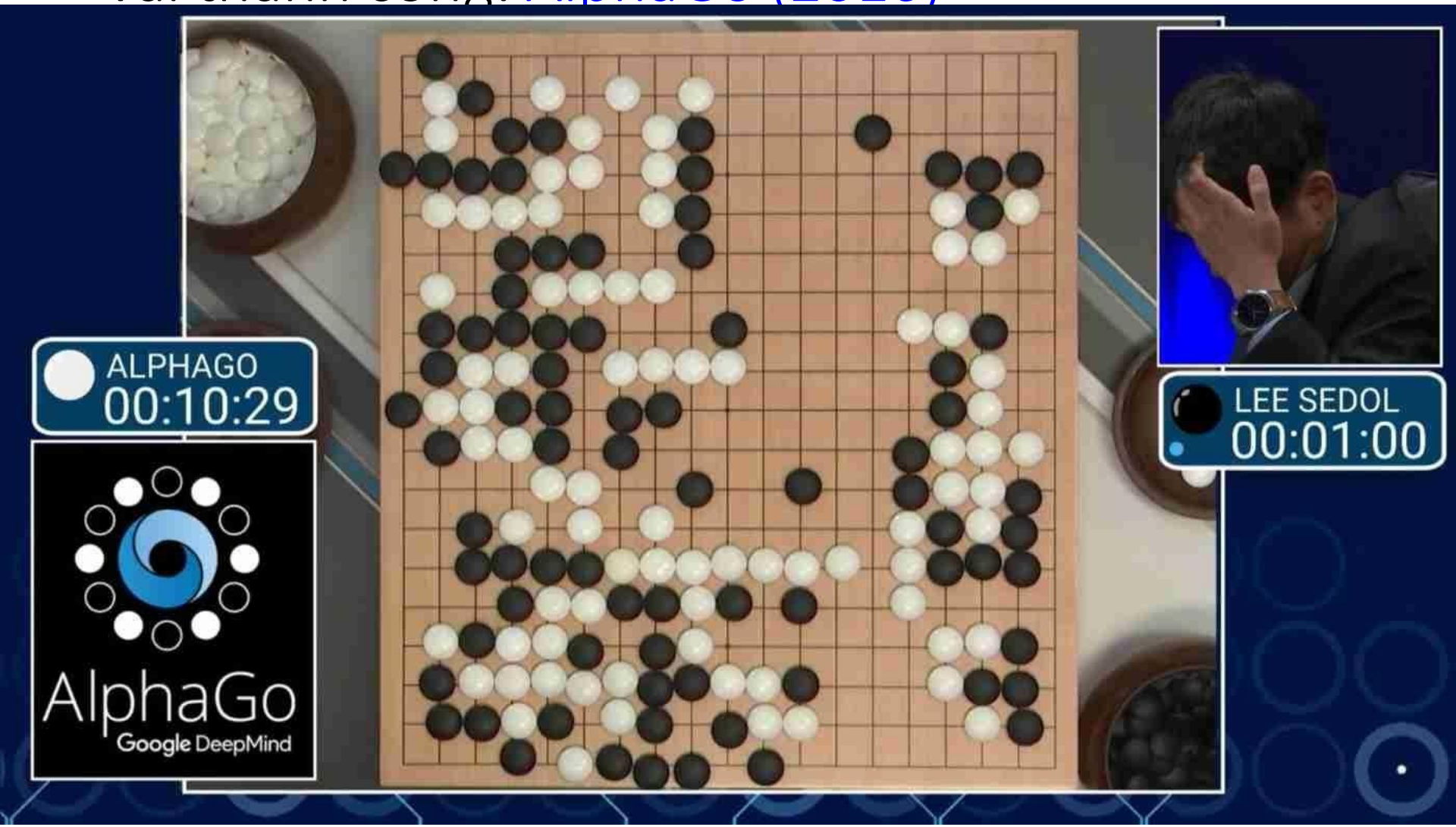


Ian Goodfellow

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.

Vài thành công: AlphaGo (2016)



http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/

Học máy -- Khai phá dữ liệu

- 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

Dữ liệu

Có cấu trúc – 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

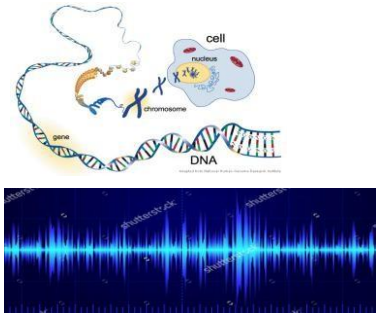
Phi cấu trúc


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

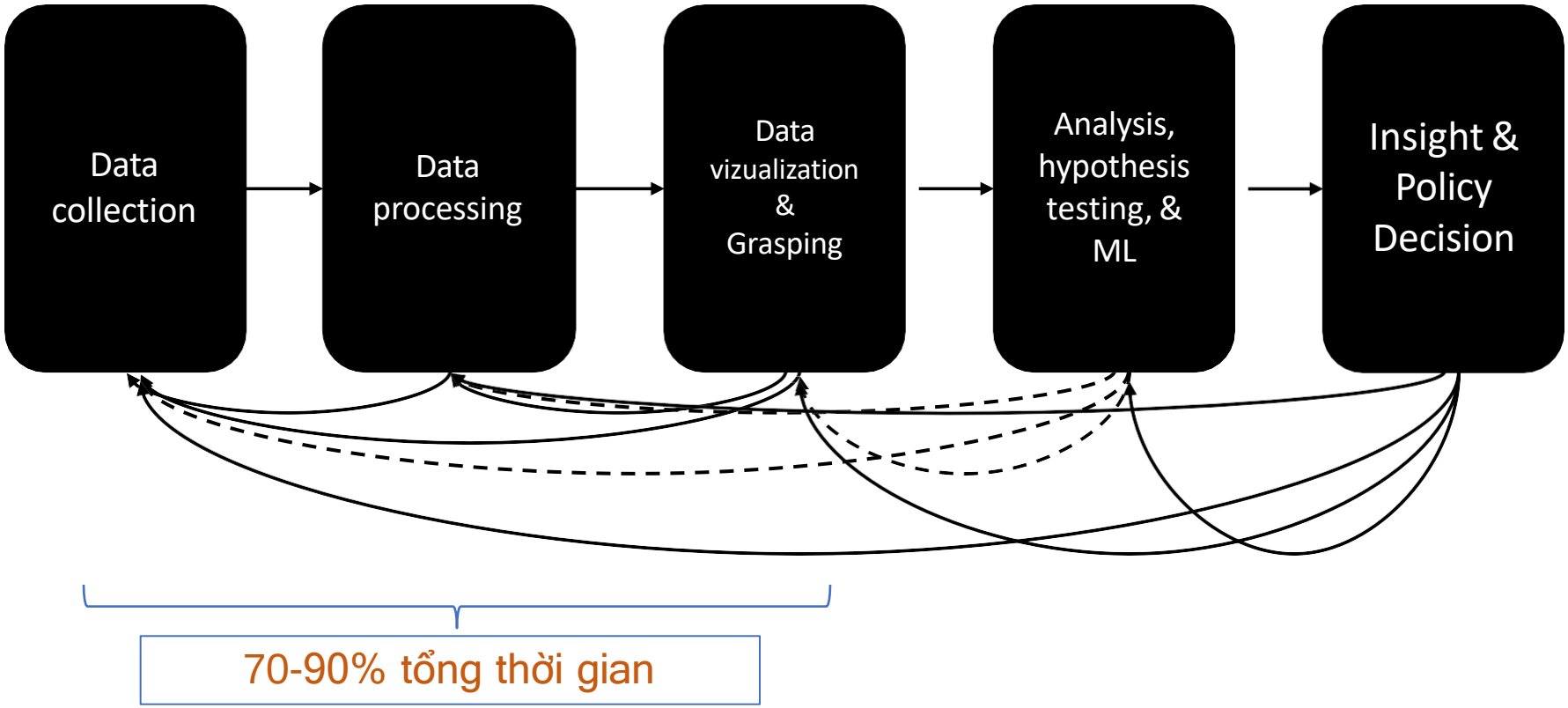

texts in websites, emails, articles, tweets

2D/3D images, videos + meta

spectrograms, DNAs, …

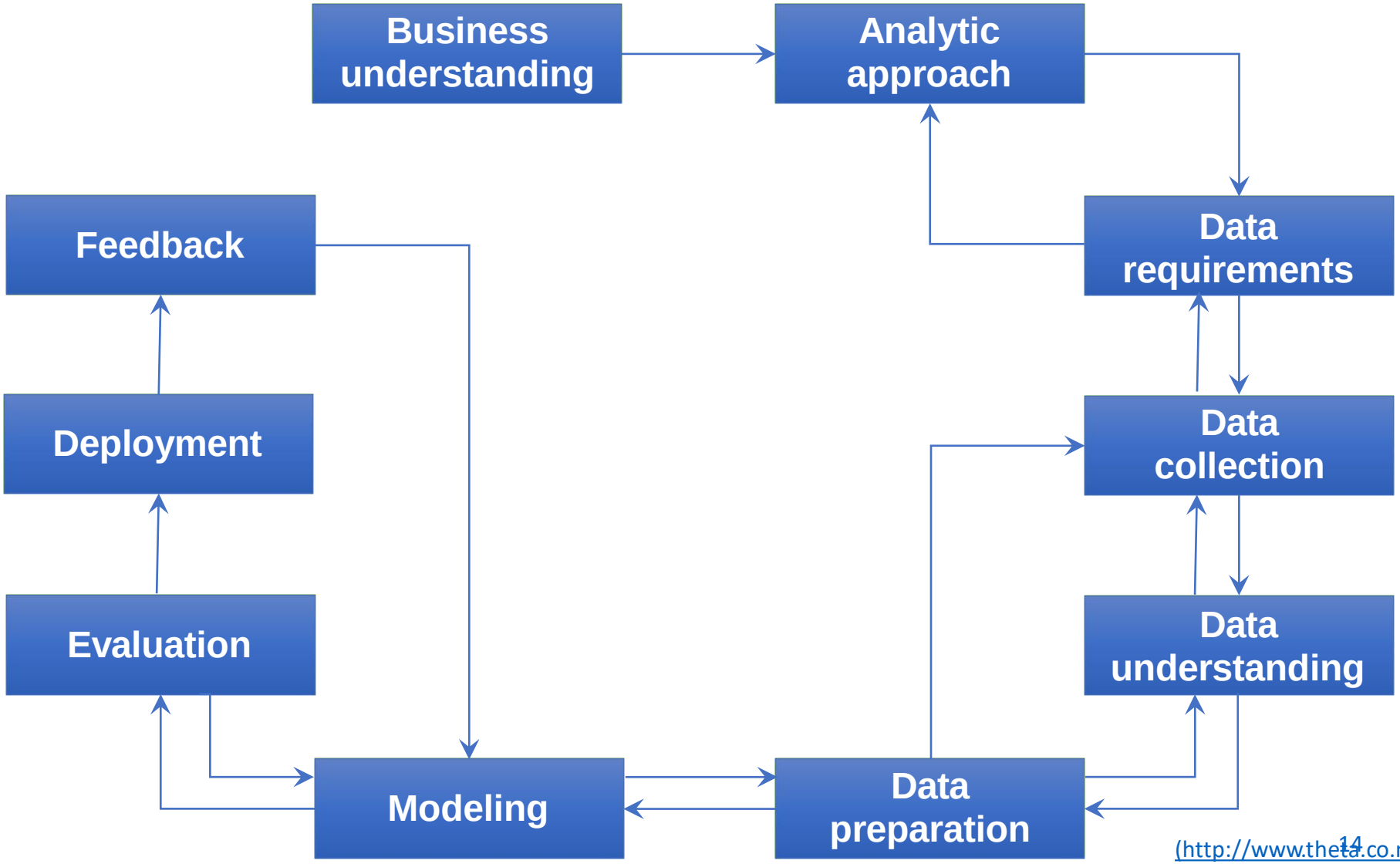


Quy trình thực hiện: hướng tìm tri thức



(John Dickerson, University of Maryland)

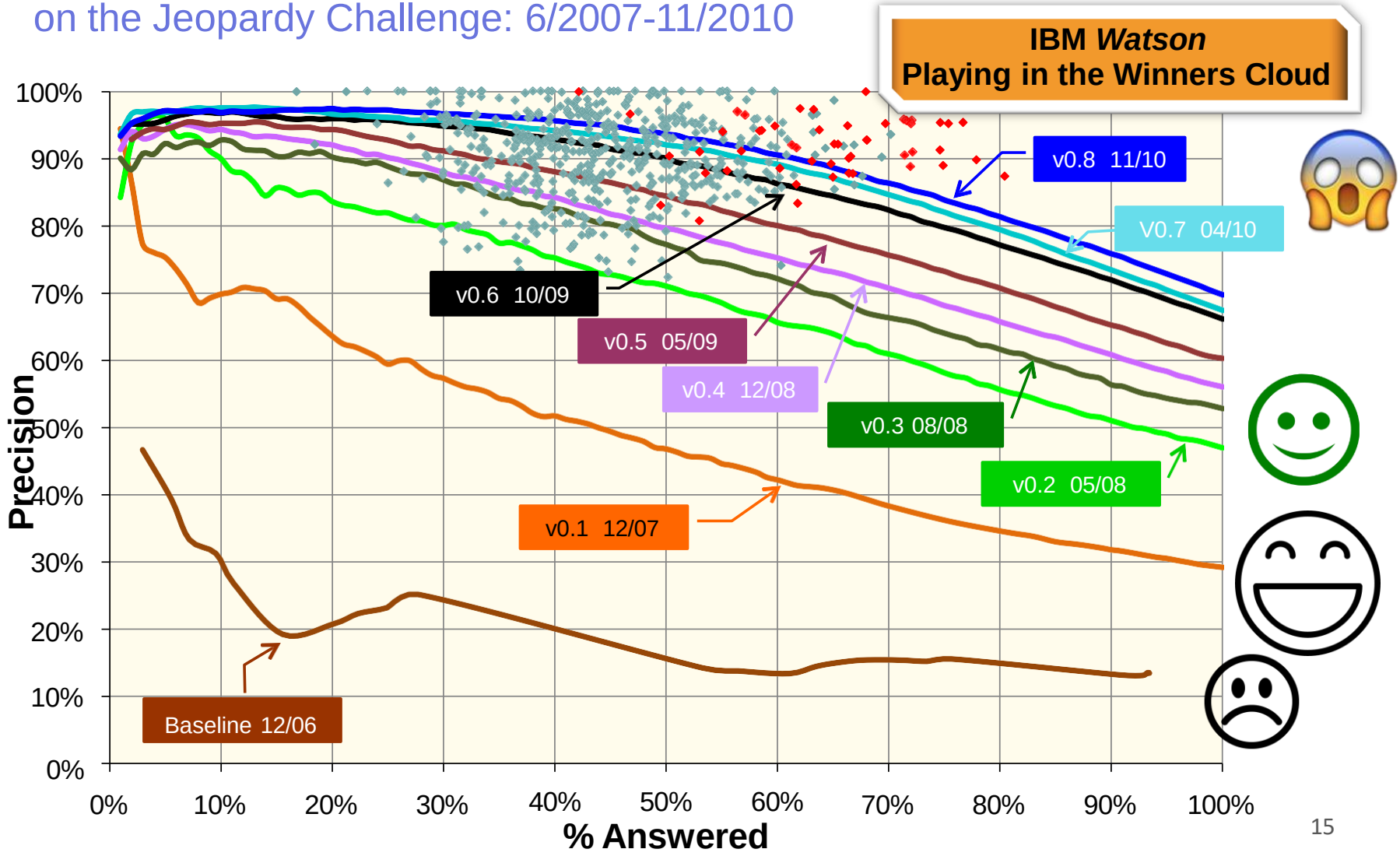
Quy trình thực hiện: hướng sản phẩm



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

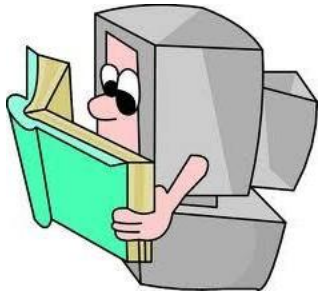
Phát triển sản phẩm: kinh nghiệm từ IBM

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



Machine Learning?

- Học máy (ML - Machine Learning) là một lĩnh vực nghiên cứu của Trí tuệ nhân tạo (Artificial Intelligence)
- Câu hỏi trung tâm của ML: [Mitchell, 2006]
  - *How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?*
- Vài quan điểm về học máy:
  - 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]

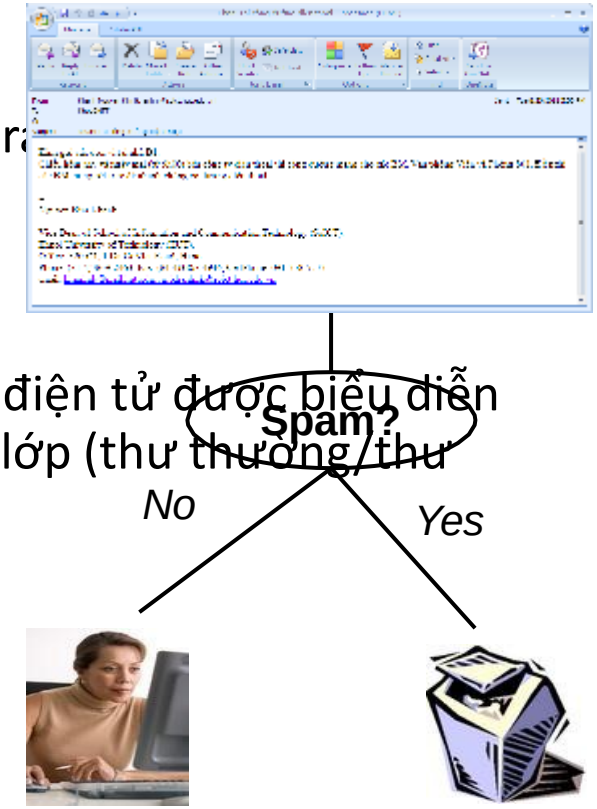


Máy học

- Ta nói một máy tính *có khả năng học* nếu nó tự cải thiện hiệu suất hoạt động ***P*** cho một công việc ***T*** cụ thể, dựa vào kinh nghiệm ***E*** của nó.
- Như vậy *một bài toán học máy* có thể biểu diễn bằng 1 bộ (***T, P, E***)
  - ***T:*** một công việc (nhiệm vụ)
  - ***P:*** tiêu chí đánh giá hiệu năng
  - ***E:*** kinh nghiệm

Ví dụ thực tế (1)

- Lọc thư rác (email spam filtering)
- ***T***: Dự đoán (để lọc) những thư điện tử nào là thư r
- ***P***: số lượng thư điện tử gửi đến được
- phân loại chính xác
- ***E***: Một tập các thư điện tử (emails) mẫu, mỗi thư điện tử được biểu diễn bằng một tập thuộc tính (vd: tập từ khóa) và nhãn lớp (thư thường/thư rác) tương ứng



Ví dụ thực tế (2)

Gán nhãn ảnh

- ***T***: đưa ra một vài mô tả ý nghĩa của 1 bức ảnh
- ***P***: ?
- ***E***: Một tập các bức ảnh, trong đó mỗi ảnh đã được gán một tập các từ mô tả ý nghĩa của chúng

FISH WATER OCEAN TREE CORAL

PEOPLE MARKET PATTERN TEXTILE DISPLAY

BIRDS NEST TREE BRANCH LEAVES

Máy học gì?

- Học một ánh xạ (hàm):

$$f : x \mapsto y$$

  - x: quan sát (dữ liệu), kinh nghiệm
  - y: phán đoán, tri thức mới, kinh nghiệm mới, …

- Hồi quy (regression): nếu y là một số thực
- Phân loại (classification): nếu y thuộc một tập rời rạc (tập nhãn lớp)

Anh ta thích nghe 

+  →Trẻ hay Già?

Máy học từ đâu?

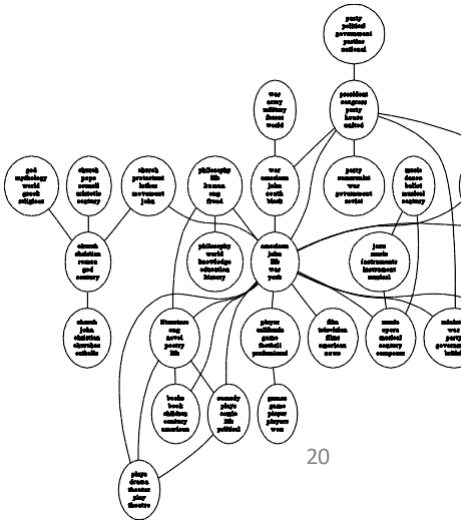
- Học từ đâu?
  - Từ các quan sát trong quá khứ (tập học – training data set).
    $\{\{x_1, x_2, \ldots, x_N\}; \{y_1, y_2, \ldots, y_M\}\}$
  - x_i là các quan sát của x trong quá khứ
  - y_h là *nhãn (label)* hoặc *phản hồi (response)* hoặc *đầu ra (output)* tương ứng với x_h.

- Sau khi đã học:
  - Thu được một mô hình, kinh nghiệm, tri thức mới (f).
  - Dùng nó để suy diễn (infer) hoặc phán đoán (predict) cho quan sát trong tương lai.

$$y_z = f(z)$$

Hai bài toán học cơ bản

- Học có giám sát (supervised learning): cần học một hàm y = f(x) từ tập học $\{\{x_1, x_2, \ldots, x_N\}; \{y_1, y_2, \ldots, y_N\}\}$ sao cho $y_i \; E \; f(x_i)$.
  - *Phân loại* (phân lớp): nếu y chỉ nhận giá trị từ một tập rời rạc, chẳng hạn {cá, cây, quả, mèo}
  - *Hồi quy:* nếu y nhận giá trị số thực
- Học không giám sát (unsupervised learning): cần học một hàm y = f(x) từ tập học cho trước $\{x_1, x_2, \ldots, x_N\}$.
  - Y có thể là các cụm dữ liệu.
  - Y có thể là các cấu trúc ẩn.
- Học bán giám sát (semi-supervised learning)?



Supervised learning: Phân loại

- ***Multi-class*** *classification (phân loại nhiều lớp):* when the output y is one of the pre-defined labels $\{c_1, c_2, \ldots, c_L\}$
  (mỗi đầu ra chỉ thuộc 1 lớp, mỗi quan sát x chỉ có 1 nhãn)
  - Spam filtering: y in {spam, normal}
  - Financial risk estimation: y in {high, normal, no}
  - Discovery of network attacks: ?

- ***Multi-label*** *classification (phân loại đa nhãn):* when the output y is a subset of labels
  (mỗi đầu ra là một tập nhỏ các lớp;
  mỗi quan sát x có thể có nhiều nhãn)
  - Image tagging: y = {birds, nest, tree}
  - sentiment analysis

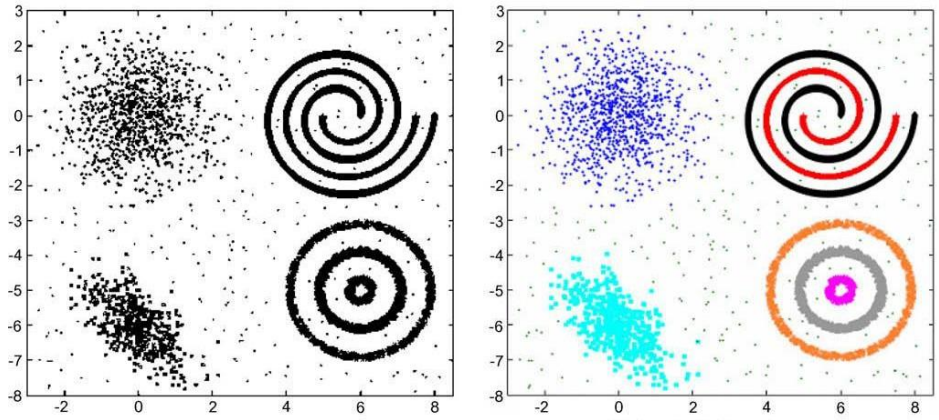
BIRDS NEST TREE

Supervised learning: Hồi quy

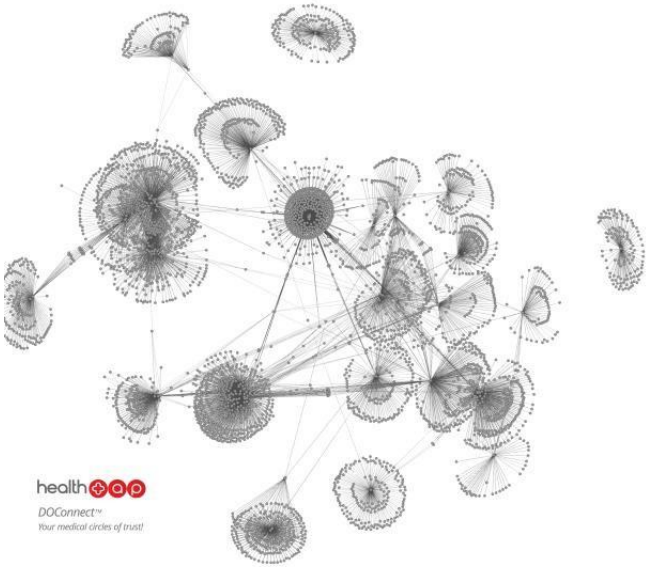
- Phán đoán chỉ số chứng khoán

Unsupervised learning: ví dụ (1)

- Gom nhóm dữ liệu vào các cụm (Clustering)
  - Discover the data groups/clusters

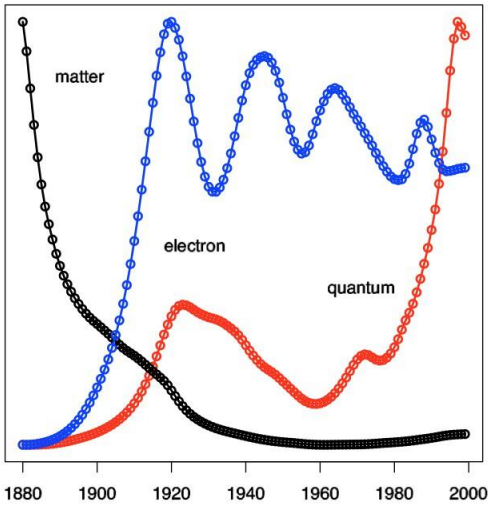


- Phát hiện cộng đồng
  - Detect communities in online social networks



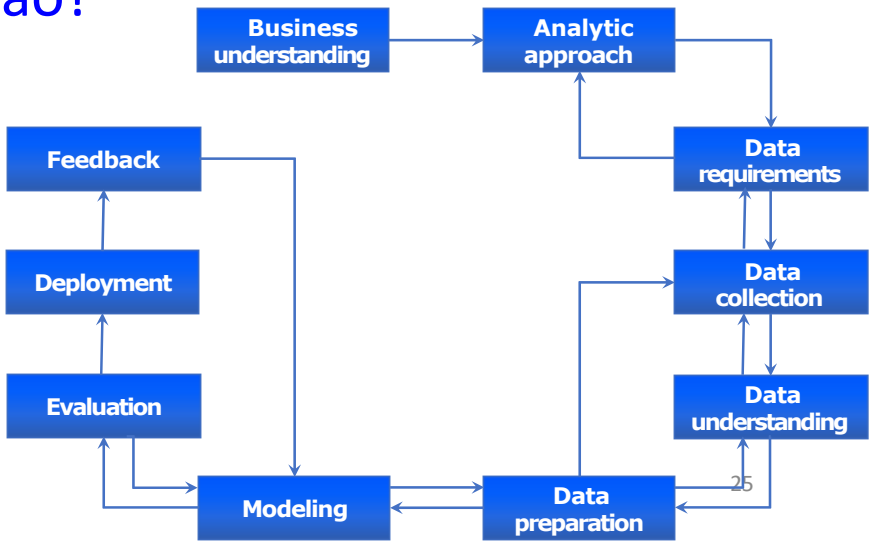
Unsupervised learning: ví dụ (2)

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



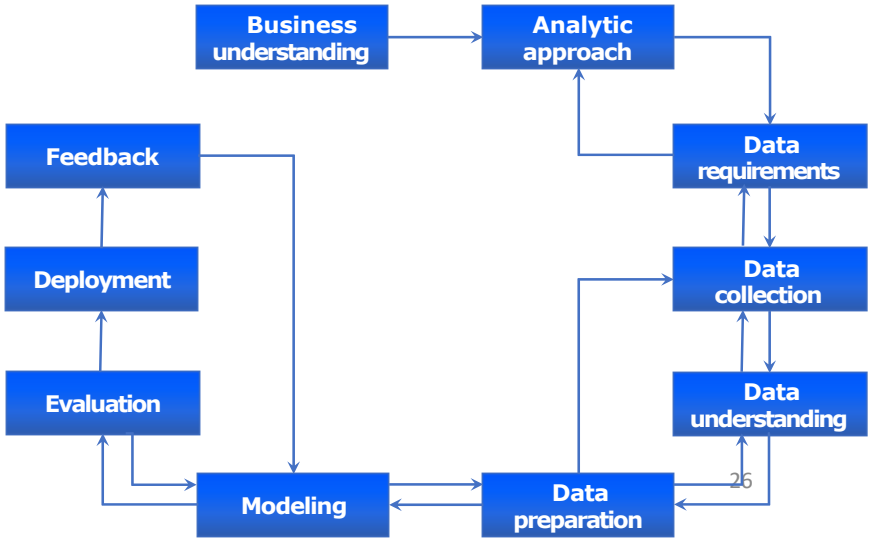
Thiết kế một hệ thống học (1)

- Một số vấn đề quan trọng cần được xem xét kỹ
- Lựa chọn tập học (training examples/data):
  - Tập học có ảnh hưởng lớn đến hiệu quả của hệ thống học.
  - Liệu ta có thu thập được nhãn cho dữ liệu huấn luyện?
  - Các ví dụ học nên tương thích với (đại diện cho) các ví dụ sẽ được làm việc bởi hệ thống trong tương lai (future test examples)
- Xác định được bài toán học máy nào?
  - Phân loại? F: X → {0,1}
  - Hồi quy? F: X → R
  - Phân cụm?



Thiết kế một hệ thống học (2)

- Lựa chọn cách biểu diễn cho hàm mục tiêu cần học
  - Hàm đa thức (a polynomial function)
  - Một tập các luật (a set of rules)
  - Một cây quyết định (a decision tree)
  - Một mạng nơ-ron nhân tạo (an artificial neural network)
  - …
- Lựa chọn một giải thuật học máy có thể học (xấp xỉ) được hàm mục tiêu
  - Hồi quy Ridge?
  - Back-propagation?
  - SGD?



Vài vấn đề trong Học máy (1)

- Giải thuật học máy (Learning algorithm)
  - Những giải thuật học máy nào có thể học (xấp xỉ) một hàm mục tiêu cần học?
  - Với những điều kiện nào, một giải thuật học máy đã chọn sẽ hội tụ (tiệm cận) đến hàm mục tiêu cần học?
  - Đối với một lĩnh vực cụ thể và đối với một cách biểu diễn các ví dụ (đối tượng) cụ thể, giải thuật học máy nào thực hiện 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)

Vài vấn đề trong Học máy (2)

- Các ví dụ học (Training examples)
  - Bao nhiêu ví dụ học là đủ?
  - Kích thước của tập học (tập huấn luyện) ảnh hưởng thế nào đối với độ chính xác của hàm mục tiêu học được?
  - Các ví dụ lỗi (nhiễu) và/hoặc các ví dụ thiếu giá trị thuộc tính (missing-value) ảnh hưởng thế nào đối với độ chính xác?

Vài vấn đề trong Học máy (3)

- Quá trình học (Learning process)
  - Chiến lược tối ưu cho việc lựa chọn thứ tự sử dụng (khai thác) các ví dụ học?
  - Các chiến lược lựa chọn này làm thay đổi mức độ phức tạp của bài toán học máy như thế nào?
  - Các tri thức cụ thể của bài toán (ngoài các ví dụ học) có thể đóng góp thế nào đối với quá trình học?

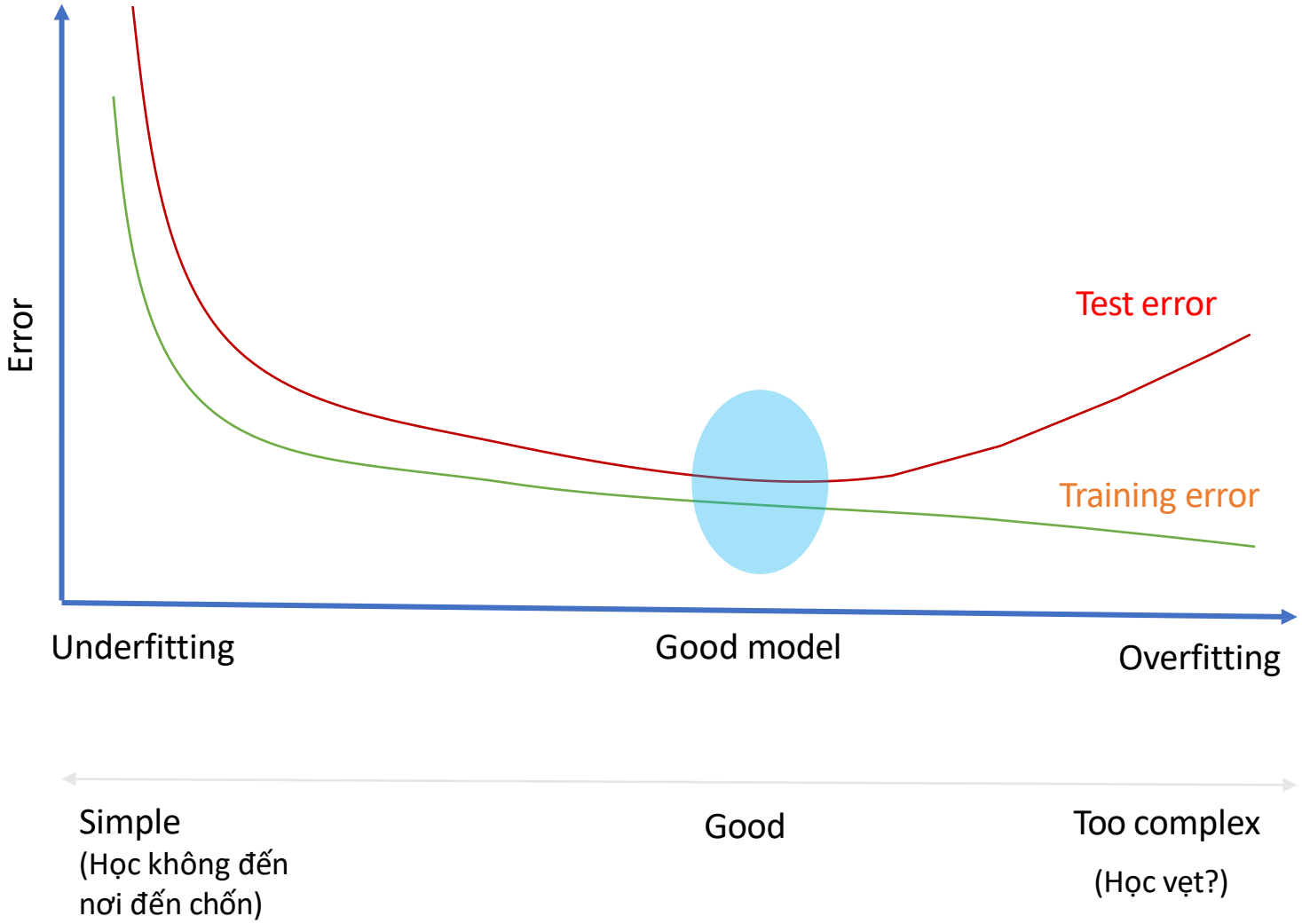
Vài vấn đề trong Học máy (4)

- Khả năng/giới hạn học (Learnability)
  - Hàm mục tiêu nào mà hệ thống cần học?
    - Biểu diễn hàm mục tiêu: Khả năng biểu diễn (vd: hàm tuyến tính / hàm phi tuyến) vs. Độ phức tạp của giải thuật và quá trình học
  - Các giới hạn đối với khả năng học của các giải thuật học máy?
  - Khả năng **Tổng quát hóa (generalization)** của hệ thống?
    - Để tránh vấn đề "overfitting" (đạt độ chính xác cao trên tập học, nhưng đạt độ chính xác thấp trên tập thử nghiệm)
  - Khả năng hệ thống tự động thay đổi (thích nghi) biểu diễn (cấu trúc) bên trong của nó?
    - Để cải thiện khả năng (của hệ thống đối với việc) biểu diễn và học hàm mục tiêu

Overfitting (quá khớp, quá khít)

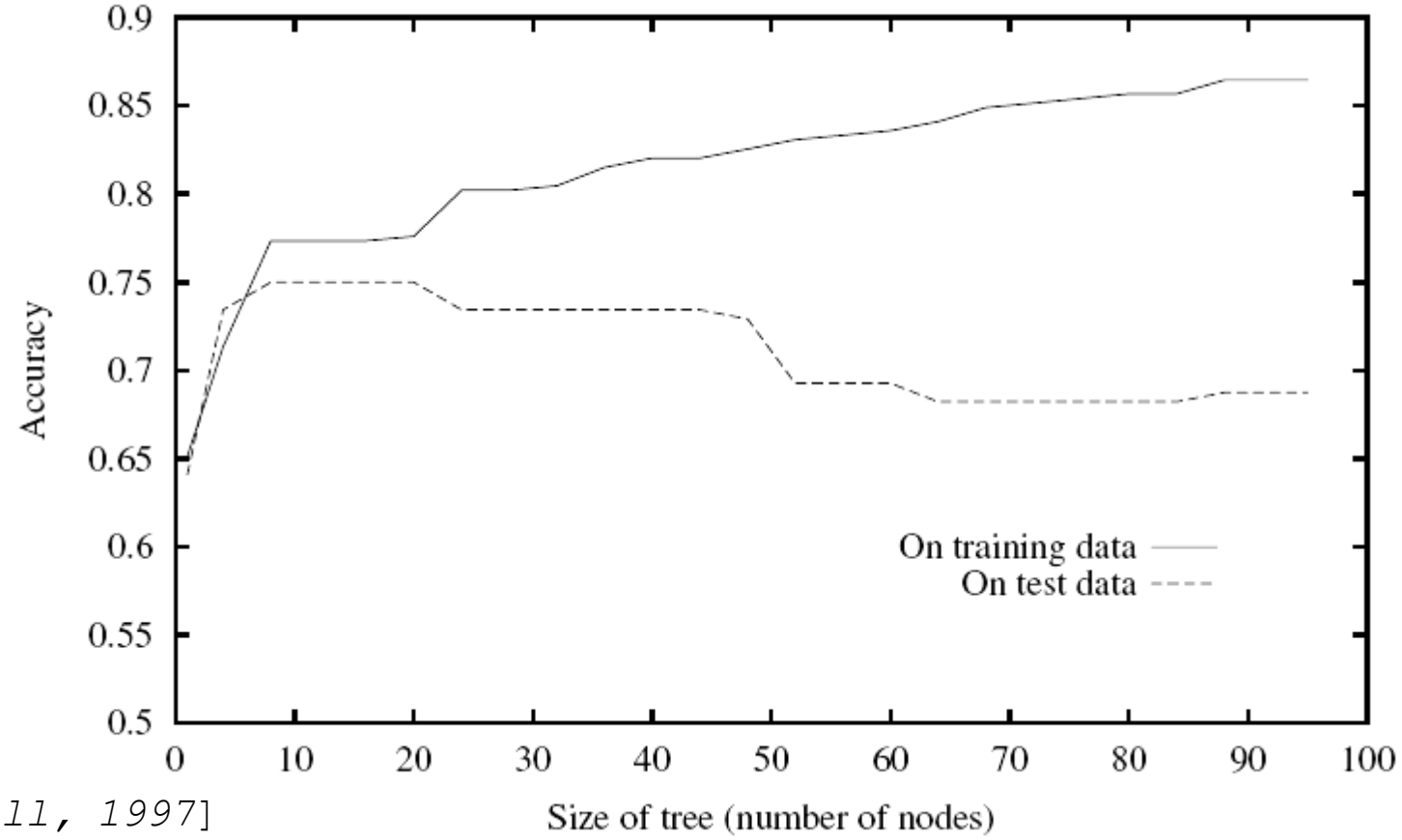
- Hàm h được gọi là *overfitting* nếu tồn tại hàm g mà:
  - g có thể tồi hơn h đối với tập huấn luyện,
  - nhưng g tốt hơn h đối với dữ liệu tương lai.
- A learning algorithm is said to overfit relative to another one if it is *more accurate in fitting* known data, but *less accurate in predicting* unseen data.
- Vài nguyên nhân gây ra Overfitting:
  - Hàm h quá phức tạp
  - Lỗi (nhiễu) trong tập huấn luyện (do quá trình thu thập/xây dựng tập dữ liệu)
  - Số lượng các ví dụ học quá nhỏ, không đại diện cho toàn bộ tập (phân bố) của các ví dụ của bài toán học

Vấn đề overfitting: minh hoạ



Overfitting: ví dụ

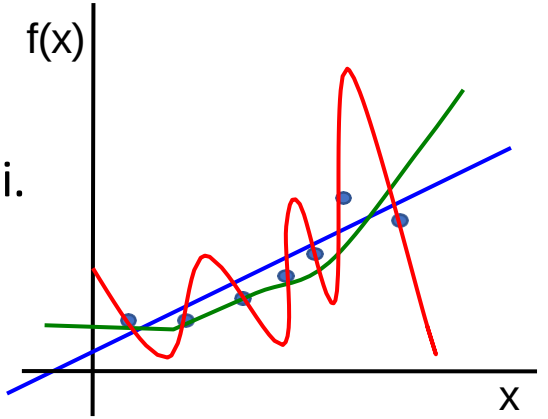
- Khi tăng cỡ lớn của một Cây quyết định thì chất lượng phán đoán của nó có thể giảm dần, mặc dù độ chính xác trên tập huấn luyện tăng dần

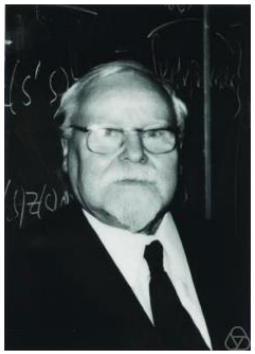


[*Mitchell, 1997*]

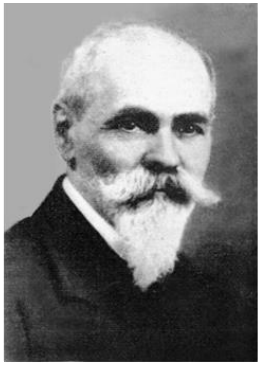
Overfitting: Regularization

- Trong số rất nhiều hàm thì hàm nào có khả năng tổng quát cao nhất khi học từ tập dữ lieu cho trước?
  - *Tổng quát hoá là mục tiêu chính của học máy.*
  - Tức là, khả năng phán đoán tốt với dữ liệu tương lai.
- **Regularization:** cách dùng phổ biến
  - Là cách hạn chế không gian chứa hàm f.





Tikhonov, smoothing an ill-posed problem



Zaremba, model complexity minimization

Bayes: priors over parameters

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

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

Tài liệu tham khảo

- 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.
- Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", *IEEE Transactions on Evolutionary Computation* **1**, 67.