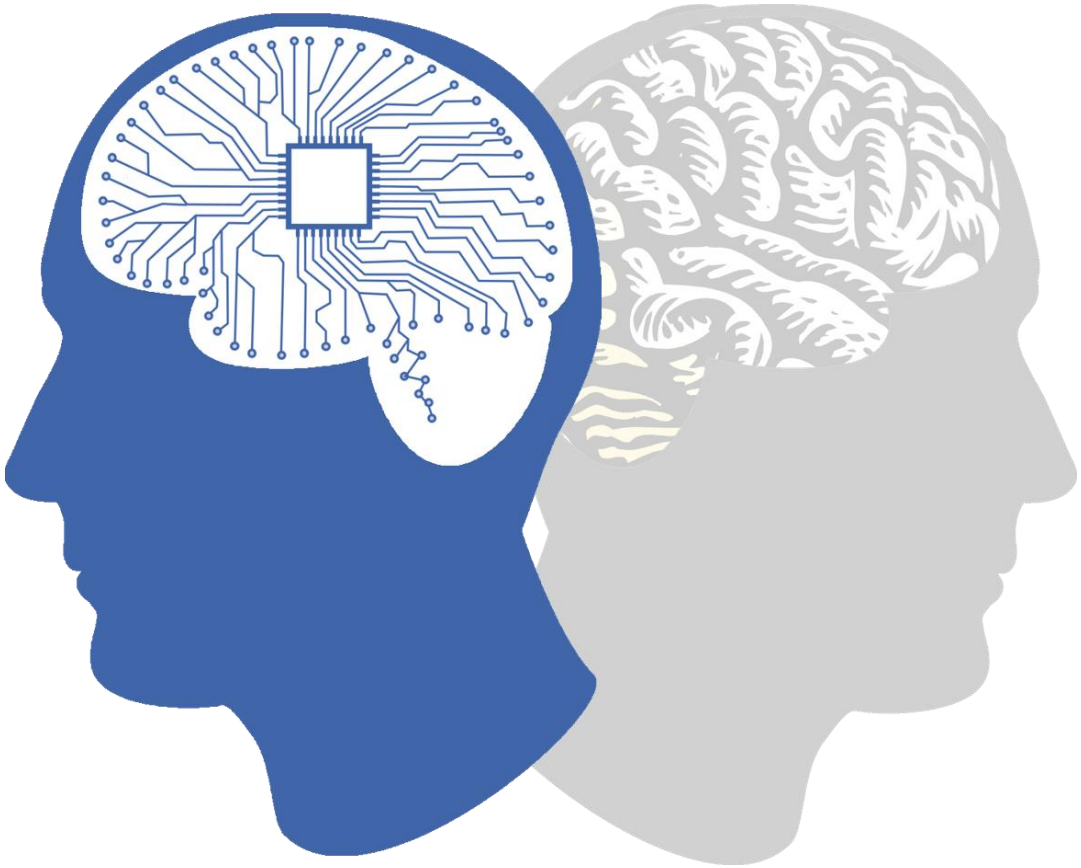
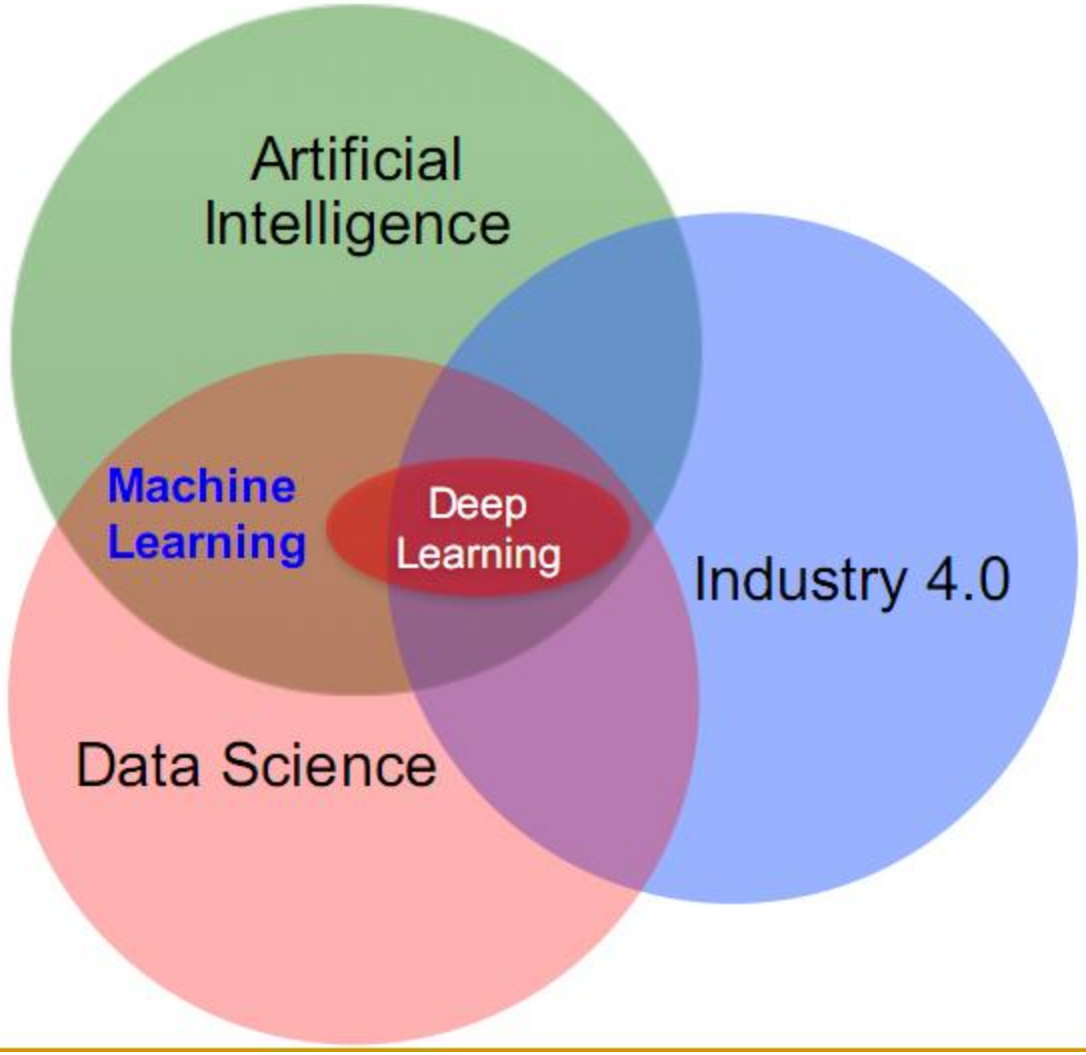
AI

Artificial Intelligence



MACHINE LEARNING



NỘI DUNG

- Tầm quan trọng và một số ứng dụng ML
- Khái niệm ML
- Phân loại ML

MỤC TIÊU

- Tổng quan về ML.
- Một số khái niệm và thuật toán ML.
- Áp dụng một số thuật toán ML vào giải quyết bài toán thực tế.

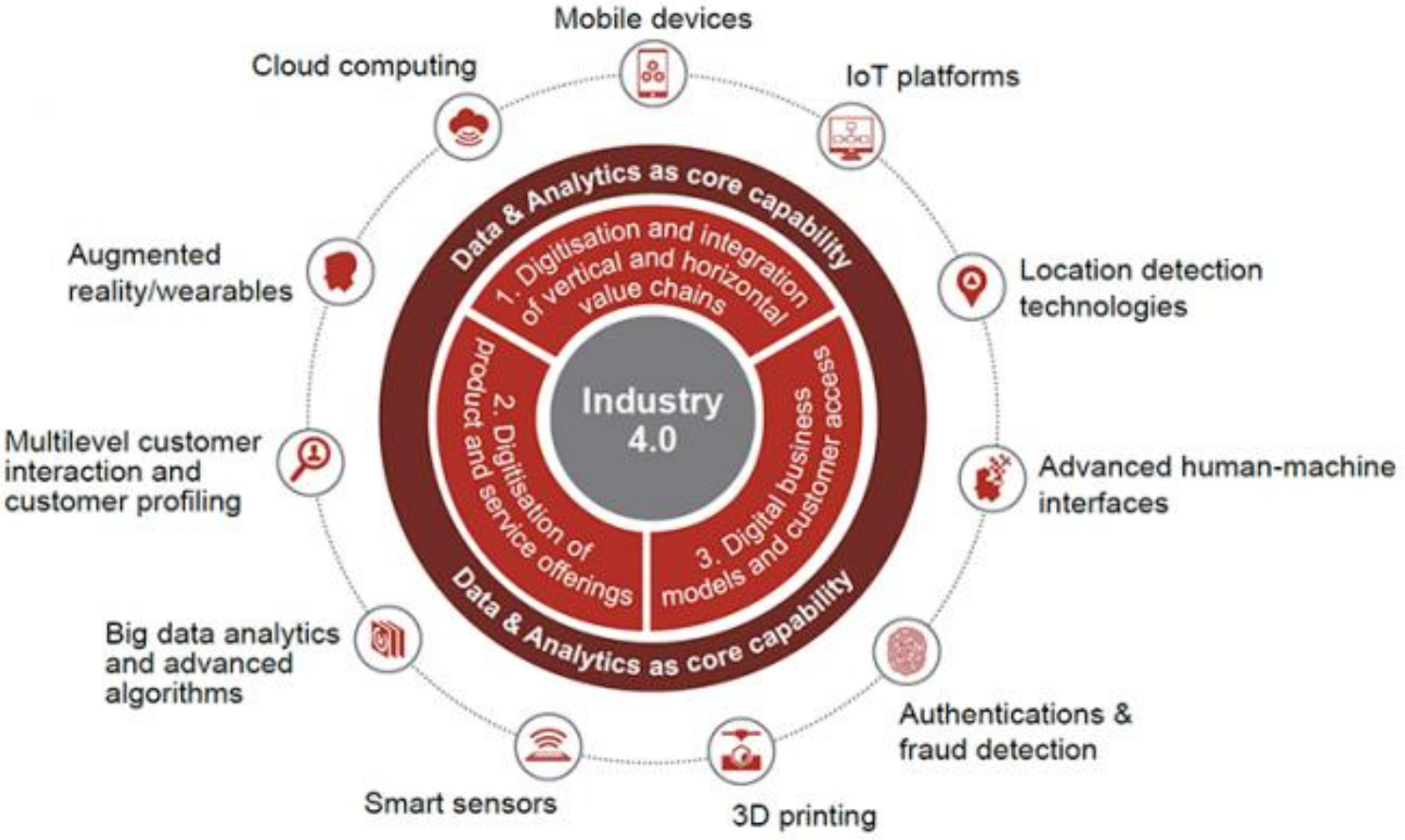
Tại sao nên biết Học Máy?

- "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 Khoa học dữ liệu (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 nên biết Học Máy?

- Nhu cầu ngày càng tăng tại Việt Nam.

Tại sao? Cách mạng công nghiệp 4.0

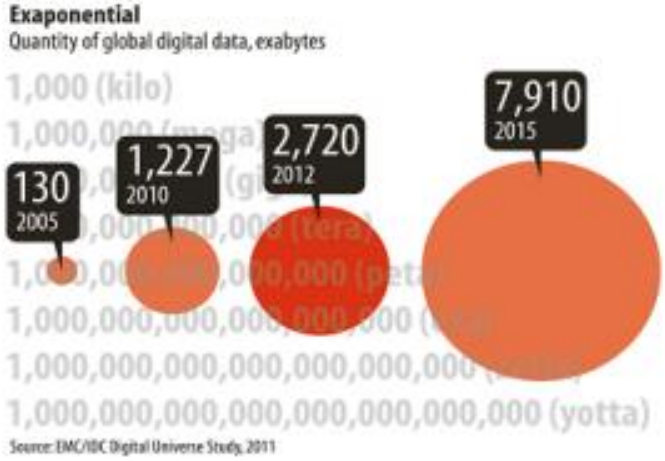


https://www.pwc.com/ca/en/industries/industry-4-0.html

Tại sao nên biết Học Máy?

- Học máy (ML – Machine Learning): data mining, inference, prediction.
- ML là con đường hiệu quả để tạo ra các hệ thống thông minh, dịch vụ thông minh.
- ML cung cấp nền tảng và phương pháp cho Big Data.

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



Vài thành công: GoogleBrain (2012)

Google's Artificial Brain Learns to Find Cat Videos

BY WIRED UK 06.26.12 | 11:15 AM | PERMALINK

By Liat Clark, Wired UK

How Many Computers to Identify a Cat? 16,000

Jim Wilson/The New York Times

An image of a cat that a neural network taught itself to recognize.

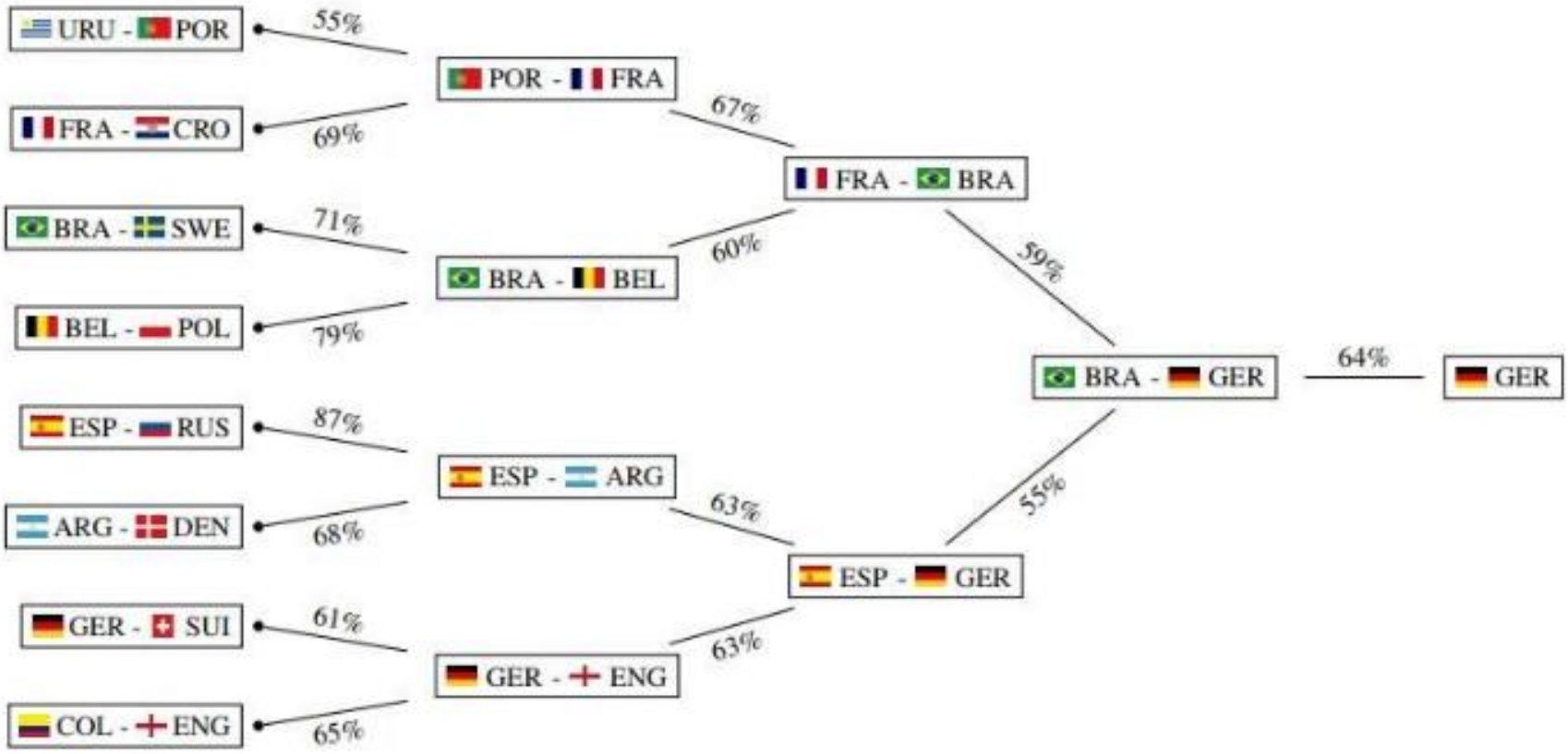
By JOHN MARKOFF

Published: June 25, 2012

Vài thành công: FIFA prediction (2014)

http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/

World Cup 2018



Vài thành công: AlphaGo (2016)

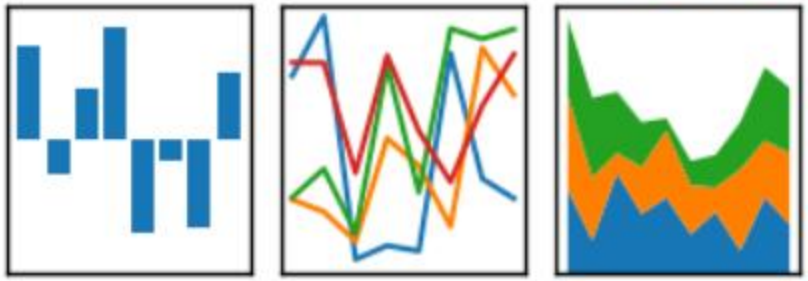
- AlphaGo of Google the world champion at Go (cờ vây), 3/2016
  - Go is a 2500 year-old game.
  - Go is one of the most complex games.
- AlphaGo learns from 30 millions human moves, and plays itself to find new moves.
- It beat Lee Sedol (World champion)
  - http://www.wired.com/2016/03/two-redefined-future/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

python™

ANACONDA
Powered by Continuum Analytics®

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

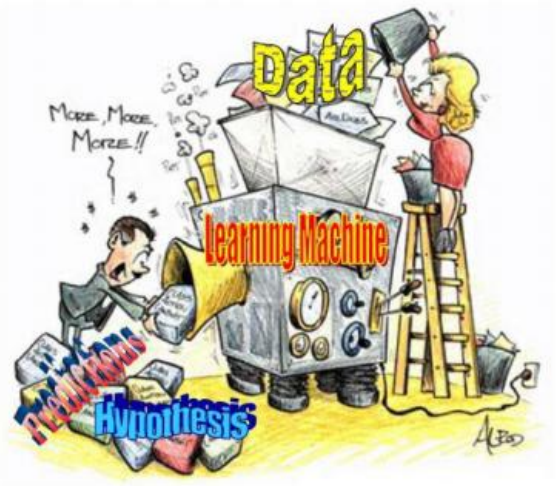


Máy Học Là Gì ?

- Học máy (ML - Machine Learning) là một lĩnh vực nghiên cứu của Trí tuệ nhân tạo (Artificial Intelligence)
- Câu hỏi trung tâm của ML:
  - → *"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"* [Mitchell, 2006]
- Vài quan điểm về học máy:
  - → Một quá trình nhờ đó một hệ thống cải thiện hiệu suất (hiệu quả hoạt động) của nó [Simon, 1983]
  - → Việc lập trình các máy tính để tối ưu hóa một tiêu chí hiệu suất dựa trên các dữ liệu hoặc kinh nghiệm trong quá khứ [Alpaydin, 2010]

Máy Học Là Gì ?

- 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

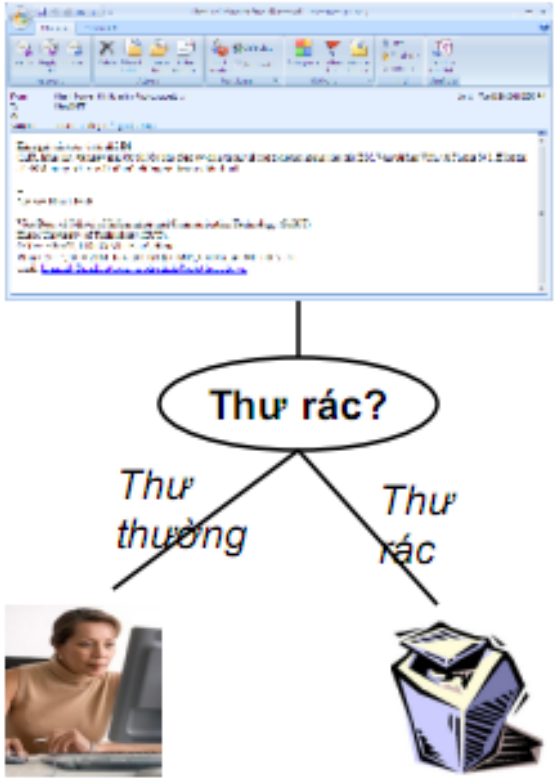


(from Eric Xing lecture notes)

Ví dụ bài toán học máy (1)

Lọc thư rác (email spam filtering)

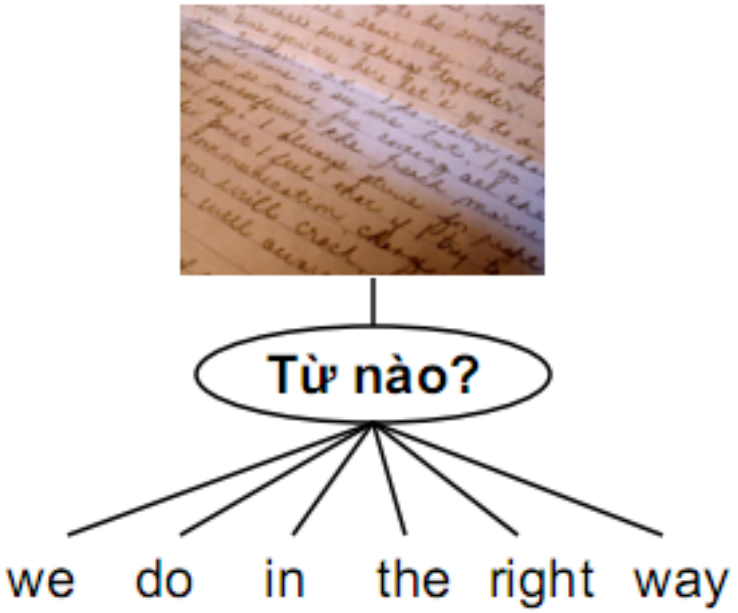
- ***T***: Dự đoán (để lọc) những thư điện tử nào là thư rác (spam email)
- ***P***: số lượng thư điện tử gửi đến được phân loại chính xác
- ***E***: Một tập các thư điện tử (emails) mẫu, mỗi thư điện tử được biểu diễn bằng một tập thuộc tính (vd: tập từ khóa) và nhãn lớp (thư thường/thư rác) tương ứng



Ví dụ bài toán học máy (2)

Nhận dạng chữ viết tay

- ***T***: Nhận dạng và phân loại các từ trong các ảnh chữ viết
- ***P***: Tỷ lệ (%) các từ được nhận dạng và phân loại đúng
- ***E***: Một tập các ảnh chữ viết, trong đó mỗi ảnh được gắn với một định danh của một từ



Ví dụ bài toán học máy (3)

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

Machine learning VS AI

Machine Learning	Artificial Intelligence
use data to compute hypothesis g that approximates target f	compute **something that shows intelligent behavior**

- $g \approx f$ is something that shows intelligent behavior —**ML can realize AI**, among other routes
- e.g. chess playing
  - traditional AI: game tree
  - ML for AI: 'learning from board data'

ML is one possible route to realize AI

Machine learning VS Learning

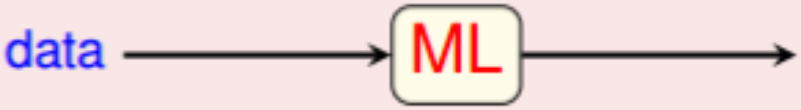
- **Machine learning is concerned with computer programs that automatically improve their performance through experience**

Learning: is any process by which a system improves performance from experience

From Learning to Machine Learning

learning: acquiring skill
with experience accumulated from observations

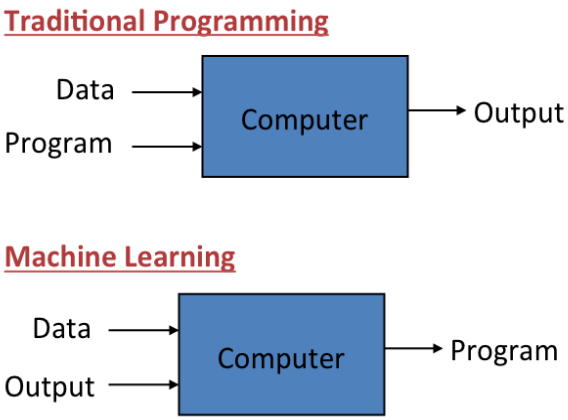
machine learning: acquiring skill
with experience accumulated/**computed** from data



What is skill?

Machine learning VS Programming

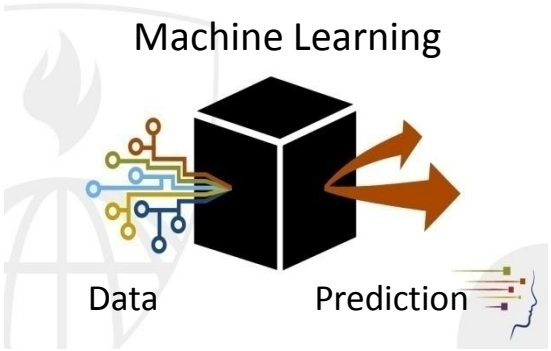
- Programming is telling a computer what to do with the given set of instructions that we call as input.
- On the other hand Machine Learning is making a machine to learn, it involves making a computer to learn from given set of instructions, in this case the computer won't repeat its mistakes, but learn from them as humans



Machine learning VS Data mining

- **Machine learning: Predictions learnt from the training data**

Data mining: Discovery of (previously) unknown properties in the data



Machine Learning and Data Mining

Machine Learning

use data to compute hypothesis *g* that approximates target *f*

Data Mining

use **(huge)** data to **find property** that is interesting

- if 'interesting property' **same as** 'hypothesis that approximate target'
  —**ML = DM** (usually what KDDCup does)
- if 'interesting property' **related to** 'hypothesis that approximate target'
  —**DM can help ML, and vice versa** (often, but not always)
- traditional DM also focuses on **efficient computation in large database**

difficult to distinguish ML and DM in reality

Máy Học Là Gì ?

- 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

Máy Học Là 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 Là Gì ?

- Học từ đâu?
  - Từ các quan sát trong quá khứ (tập học – training 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)*

- 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.
    $y_z = f(z)$

Phân loại dựa vào phương thức học

- Supervised Learning (Học có giám sát)
  - Classification (Phân loại, phân lớp)
  - Regression (Hồi quy)
- Unsupervised Learning (Học không giám sát)
  - Clustering (phân nhóm, cụm)
  - Association (luật)
- Semi-Supervised Learning (Học bán giám sát)
- Reinforcement Learning (Học Củng Cố)

Supervised Learning (Học có giám sát)

- Classification (Phân loại)
- Regression (Hồi quy)

- 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 $y_i \cong f(x_i)$.
  - *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 có giám sát: ví dụ

- Lọc thư rác
- Phân loại trang web
- Dự đoán rủi ro tài chính
- Dự đoán biến động chỉ số chứng khoán
- Phát hiện tấn công mạng

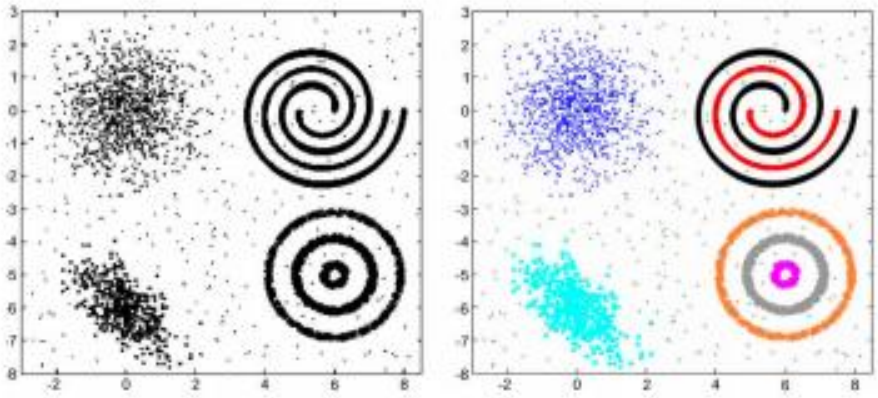
UnSupervised Learning (Học không giám sát)

- Clustering (phân nhóm, phân cụm)
- Association (luật)

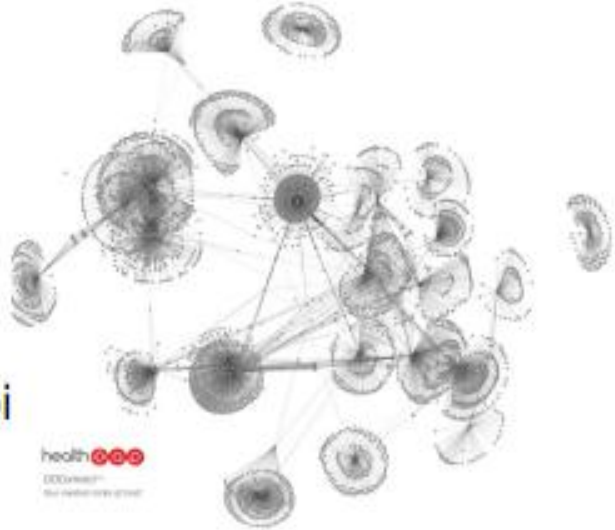
- Học không giám sát (unsupervised learning): cần học một hàm y = f(x) từ tập học cho trước $\{x_1, x_2, \ldots, x_N\}$.
  - Y có thể là các cụm dữ liệu.
  - Y có thể là các cấu trúc ẩn.

Học không giám sát: ví dụ (1)

- Phân cụm (clustering)
  - Phát hiện các cụm dữ liệu, cụm tính chất,…

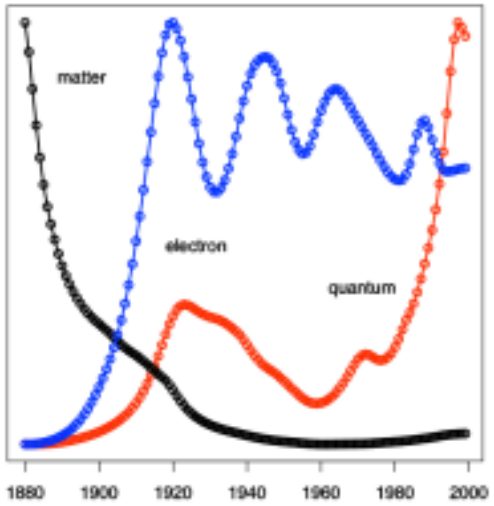


- Community detection
  - Phát hiện các cộng đồng trong mạng xã hội



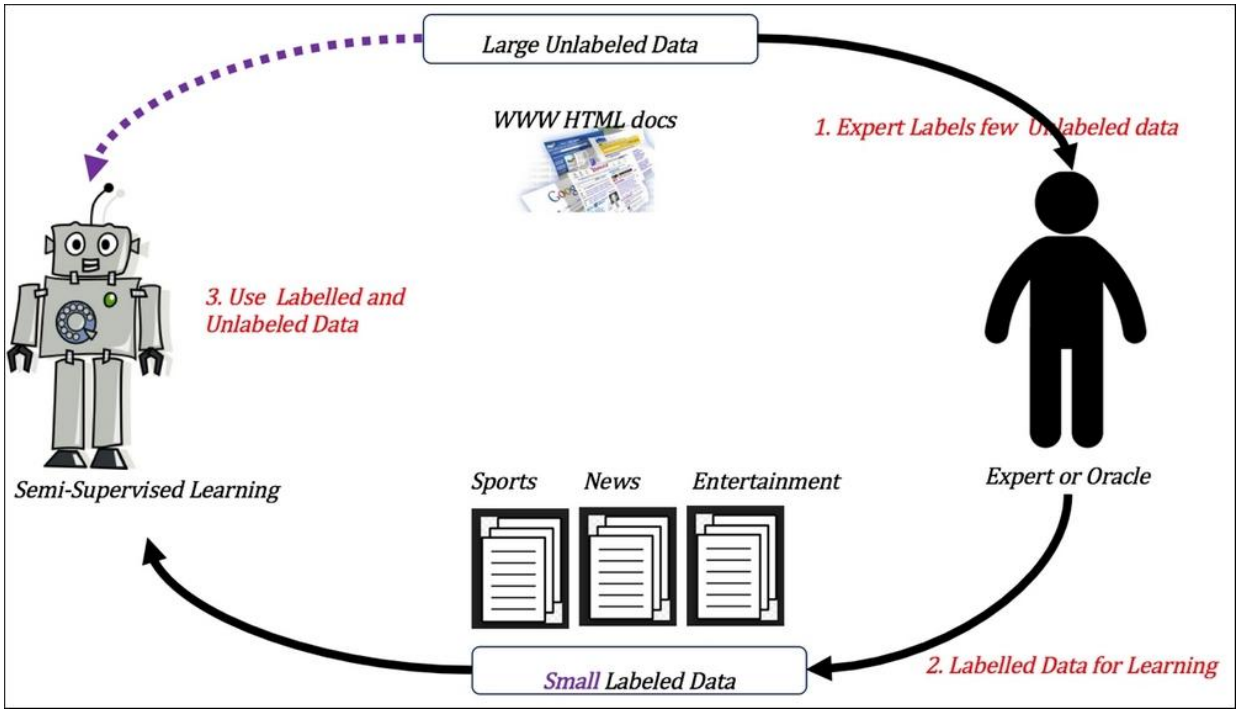
Học không giám sát: ví dụ (2)

- Trends detection
  - Phát hiện xu hướng, thị yếu,…



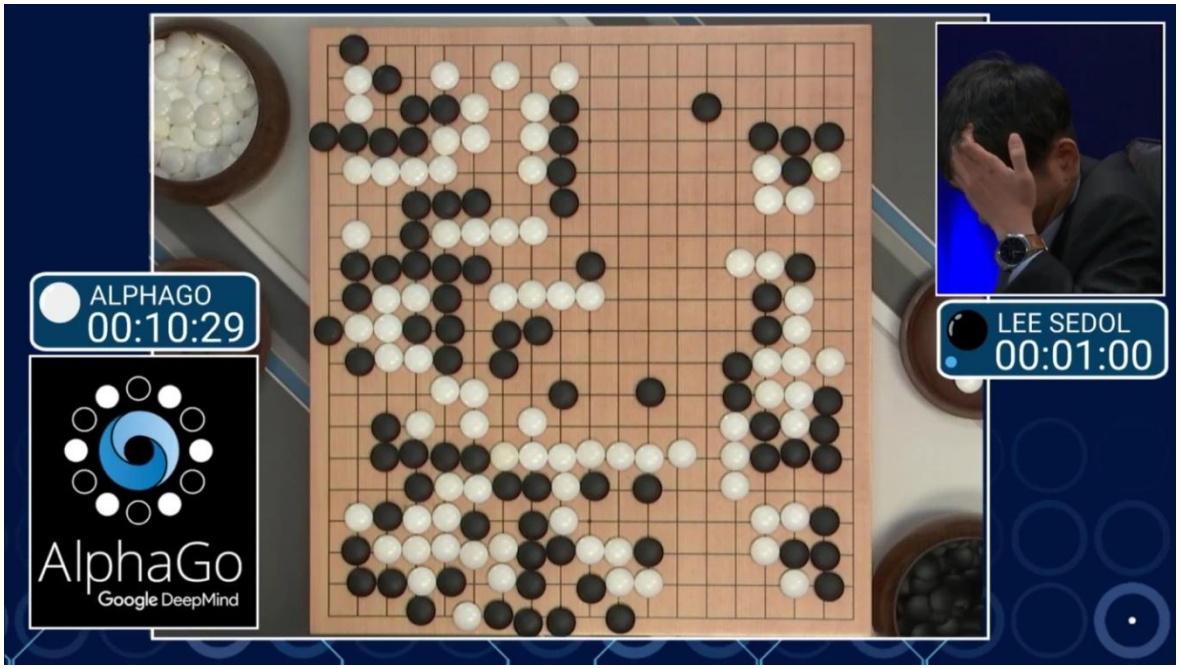
Semi-Supervised Learning (Học bán giám sát)

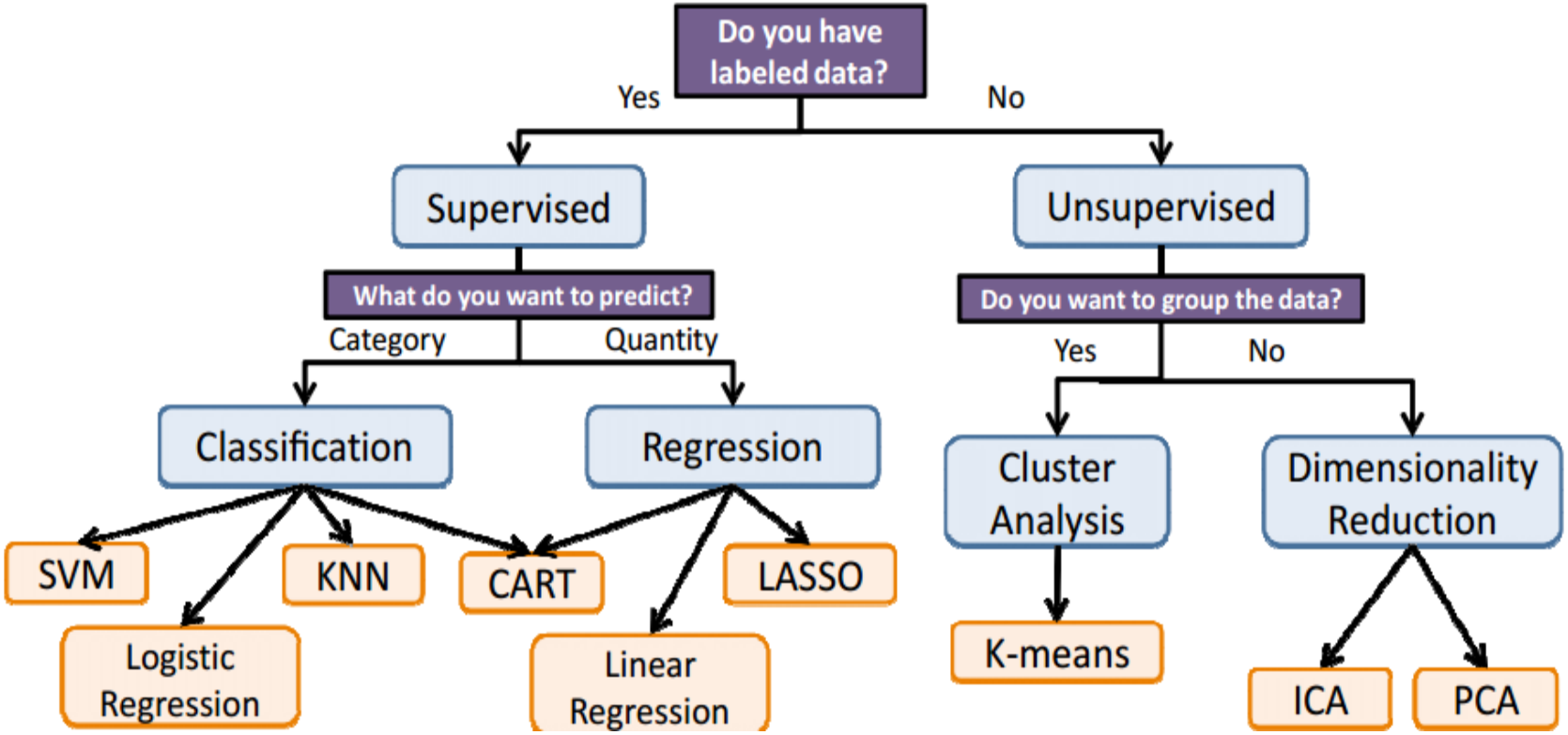
Kết hợp giữa có giám sát và không giám sát trong đó Có một lượng lớn dữ liệu, nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning



Reinforcement Learning (Học Củng Cố)

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance).



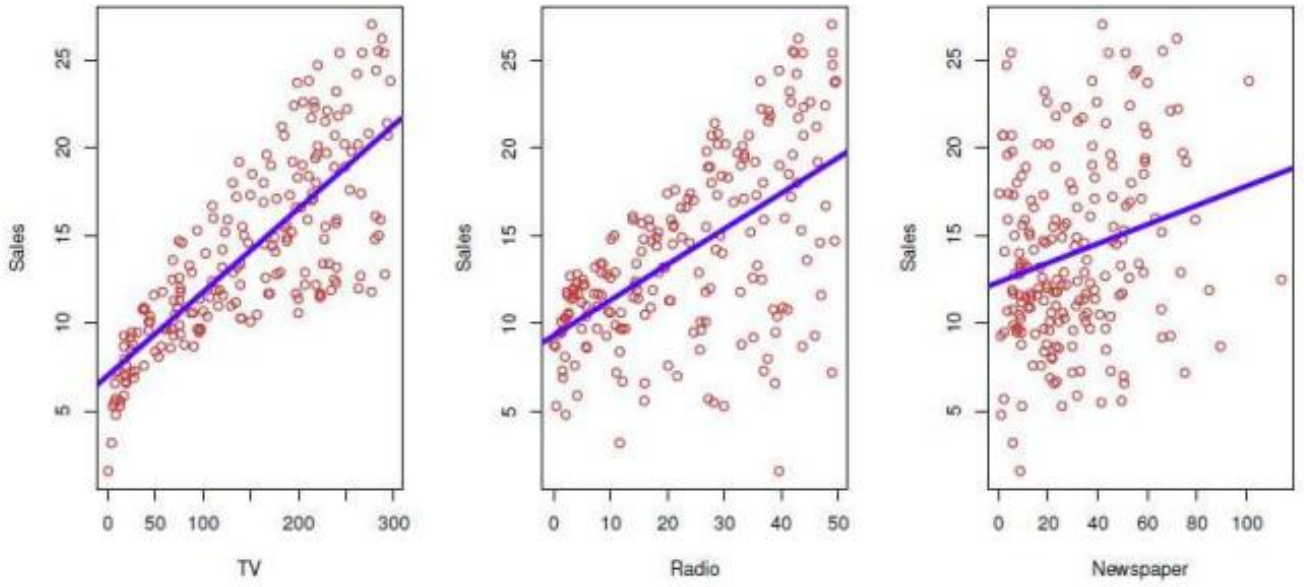

Do you have labeled data?
Yes
No
Supervised
Unsupervised
What do you want to predict?
Category
Quantity
Do you want to group the data?
Yes
No
Classification
Regression
Cluster Analysis
Dimensionality Reduction
SVM
KNN
CART
LASSO
Logistic Regression
Linear Regression
K-means
ICA
PCA

PHÂN L0ẠI DỰA TRÊN CHỨC NĂNG

- 1. Phân nhóm dựa trên phương thức học
  - Supervised Learning (Học có giám sát)
    - Classification (Phân loại)
    - Regression (Hồi quy)
  - Unsupervised Learning (Học không giám sát)
    - Clustering (phân nhóm)
    - Association
  - Semi-Supervised Learning (Học bán giám sát)
  - Reinforcement Learning (Học Củng Cố)
- 2. Phân nhóm dựa trên chức năng
  - Regression Algorithms
  - Classification Algorithms
  - Instance-based Algorithms
  - Regularization Algorithms
  - Bayesian Algorithms
  - Clustering Algorithms
  - Artificial Neural Network Algorithms
  - Dimensionality Reduction Algorithms
  - Ensemble Algorithms

VÍ DỤ BÀI TOÁN THỰC TẾ

- Doanh nghiệp có thể điều chỉnh chiến lược quảng cáo sản phẩm (advertising) để tăng doanh số bán hàng (sales).
- Dữ liệu: Doanh số bán hàng và ngân sách quảng cáo cho 3 phương tiện truyền thông (TV, radio, newspaper).



VÍ DỤ BÀI TOÁN THỰC TẾ

- Trong ví dụ trên đâu là biến đầu vào, đầu ra ?
- Hãy lấy ví dụ về yêu cầu dự đoán, và suy diễn mà ta có đươc dữ liệu lời giải từ dữ liệu trên ?

VÍ DỤ BÀI TOÁN THỰC TẾ

- Trong ví dụ về quảng cáo, đâu là biến đầu vào/đầu ra?
  - *Biến đầu ra* : doanh số bán hàng
  - *Biến đầu vào*: ngân sách quảng cáo trên TV, ngân sách quảng cáo trên Radio, ngân sách quảng cáo trên báo chí
- Hãy lấy ví dụ về yêu cầu dự đoán và suy diễn mà ta có được lời giải từ dữ liệu này.
  - Dự đoán:
    - Số liệu về doanh số bán hàng ở thị trường A dự kiến thế nào khi biết ngân sách đầu tư quảng cáo trên TV, radio và báo chí?
  - Suy diễn:
    - Doanh số bán hàng tăng bao nhiêu nếu tăng ngân sách 10% cho quảng cáo trên TV?
    - Phương tiện truyền thông nào (TV, radio, báo) tạo ra sự thúc đẩy lớn nhất trong bán hàng?

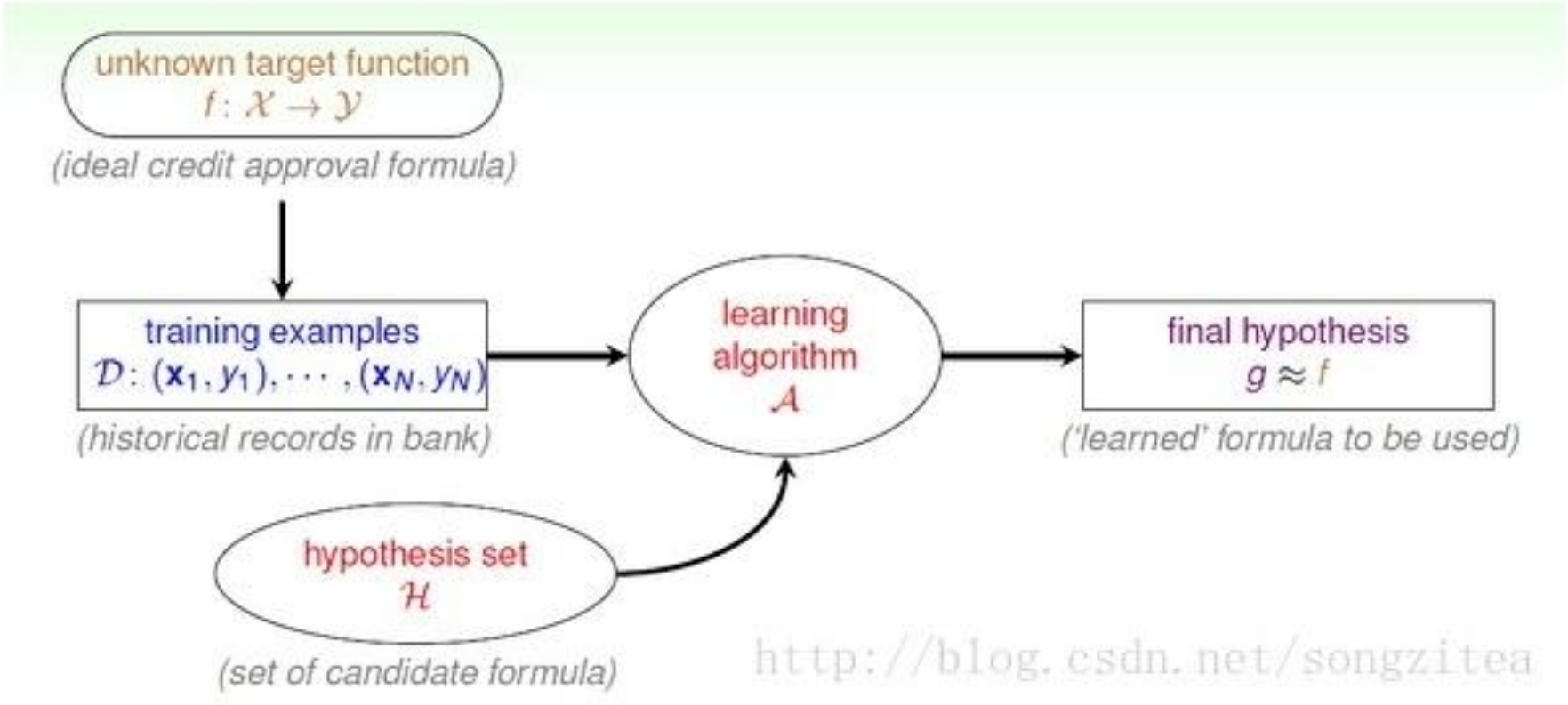
QUÁ TRÌNH HOC MÁY CƠ BẢN

Basic Notations

- input: $\mathbf{x} \in \mathcal{X}$ (customer application)
- output: $y \in \mathcal{Y}$ (good/bad after approving credit card)
- unknown pattern to be learned $\Leftrightarrow$ target function:
  $f: \mathcal{X} \to \mathcal{Y}$ (ideal credit approval formula)
- data $\Leftrightarrow$ training examples: $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$
  (historical records in bank)
- hypothesis $\Leftrightarrow$ skill with hopefully good performance:
  $g: \mathcal{X} \to \mathcal{Y}$ ('learned' formula to be used)

$\{(\mathbf{x}_n, y_n)\}$ from f → ML → g

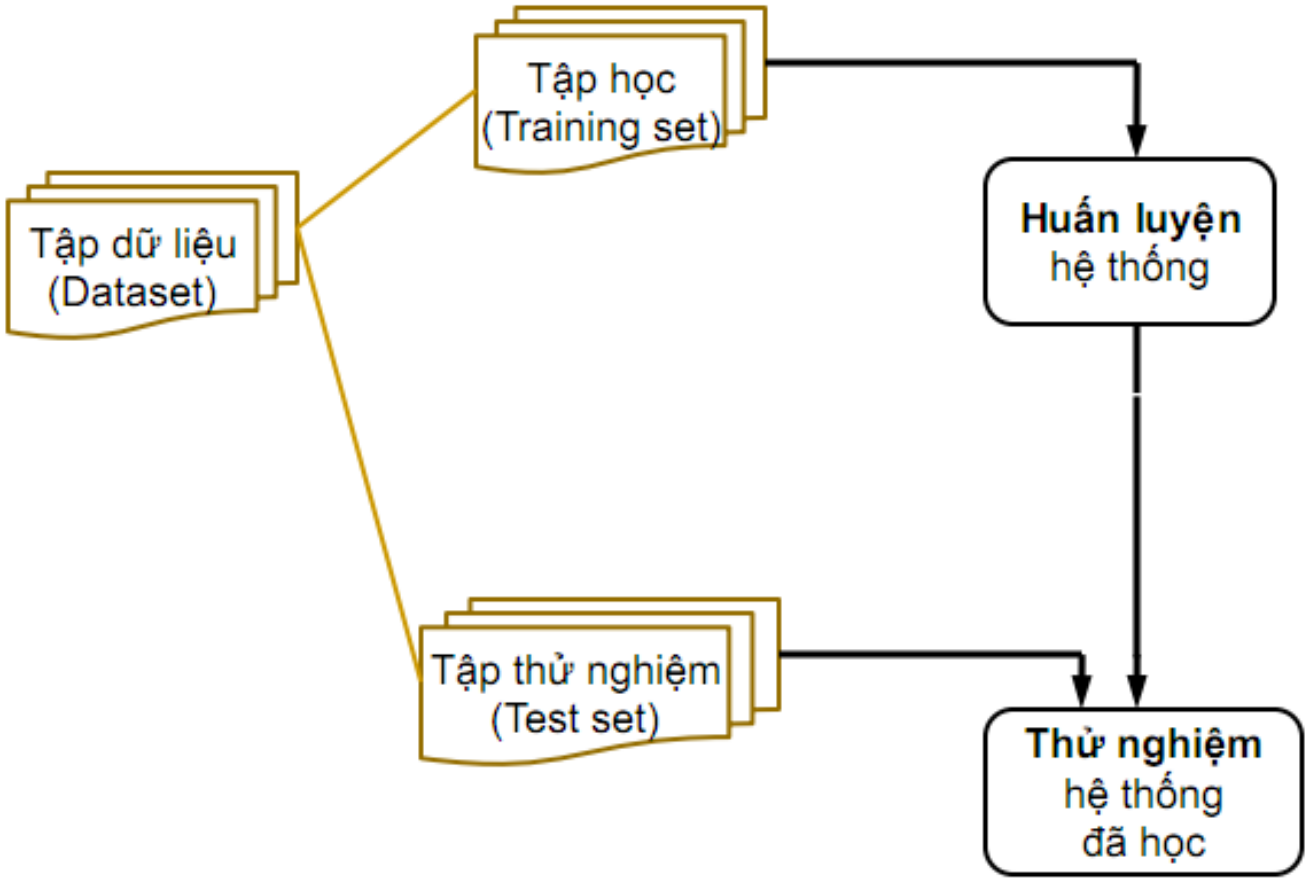
QUÁ TRÌNH HOC MÁY CƠ BẢN



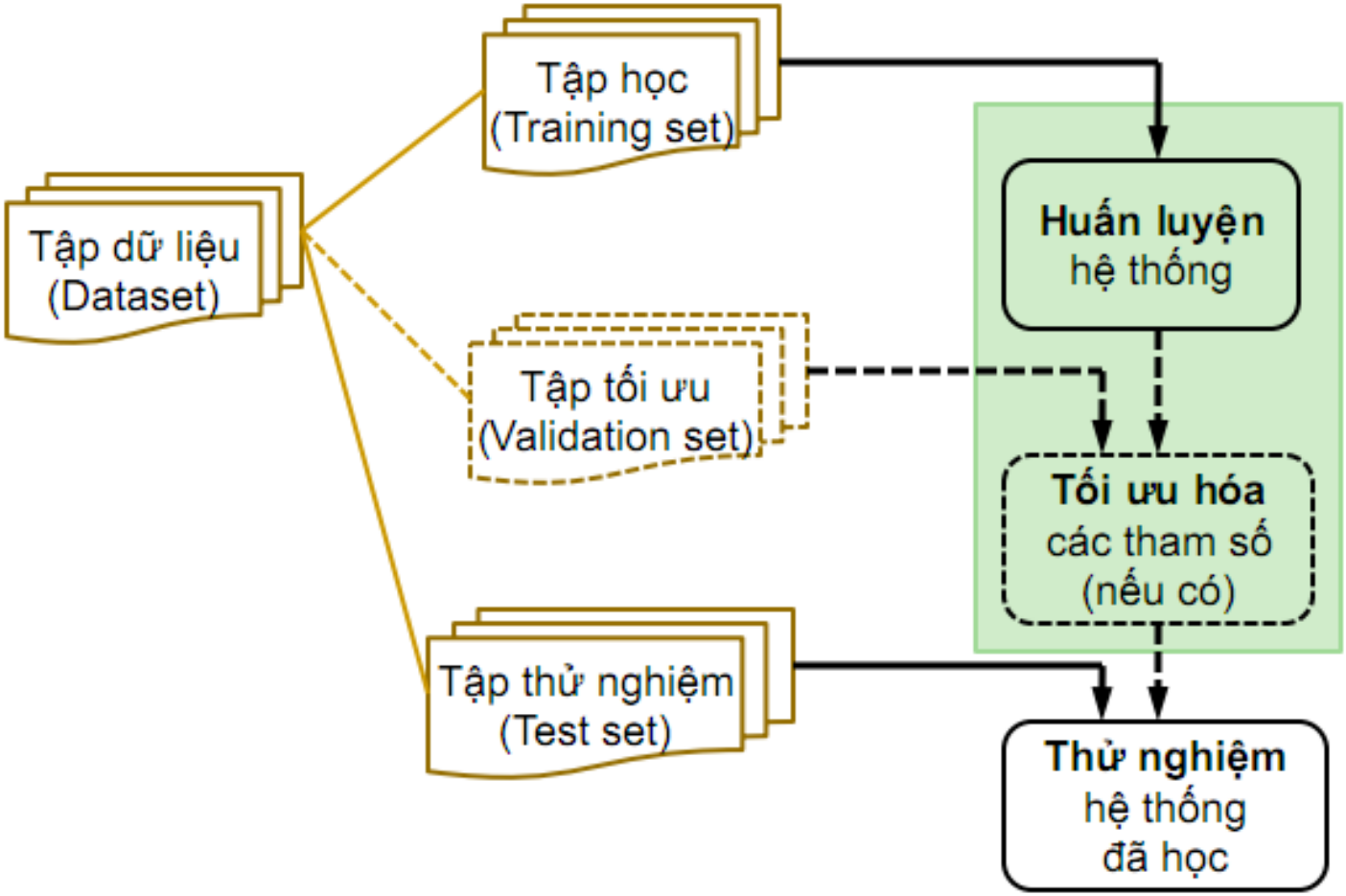
DỮ LIỆU HUẤN LUYỆN

- *Dữ liệu huấn luyện (Training data)*
  - Tập các quan sát (bản ghi) được sử dụng để xây dựng (học) mô hình.
- *Dữ liệu kiểm chứng (Validation data)*
  - Tập các quan sát dùng để ước lượng lỗi nhằm tìm tham số hoặc lựa chọn mô hình.
- *Dữ liệu kiểm thử (Test data)*
  - Tập các quan sát dùng để đánh giá hiệu năng trên dữ liệu chưa biết (unseen) trong tương lai.
  - Dữ liệu này không sử dụng cho giải thuật học máy trong quá trình xây dựng mô hình.

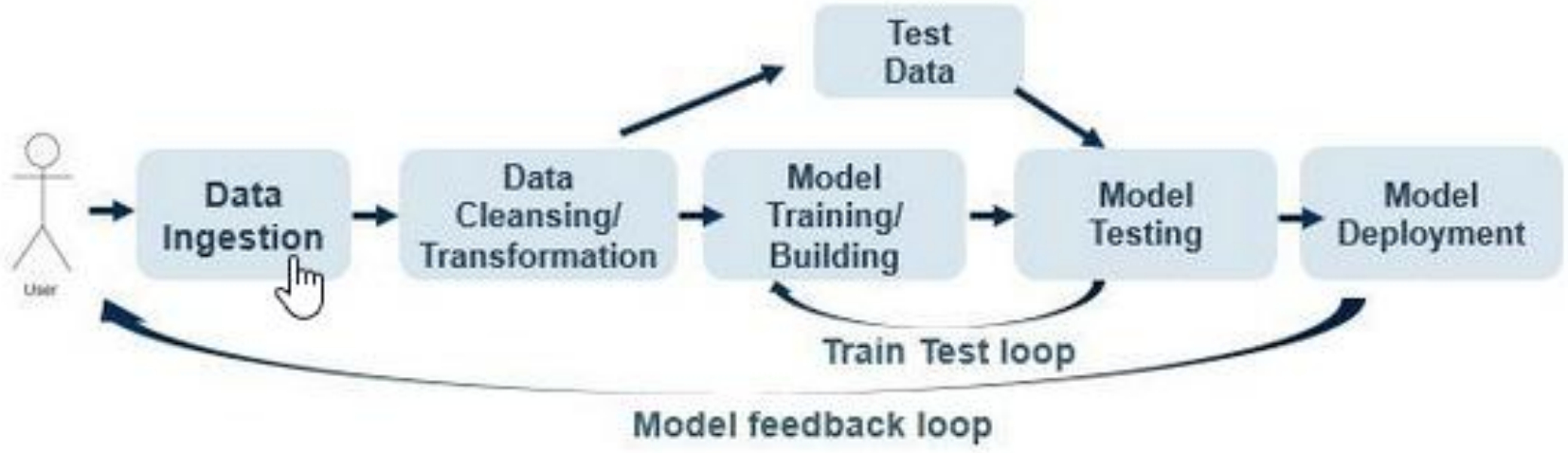
QUÁ TRÌNH HOC MÁY CƠ BẢN



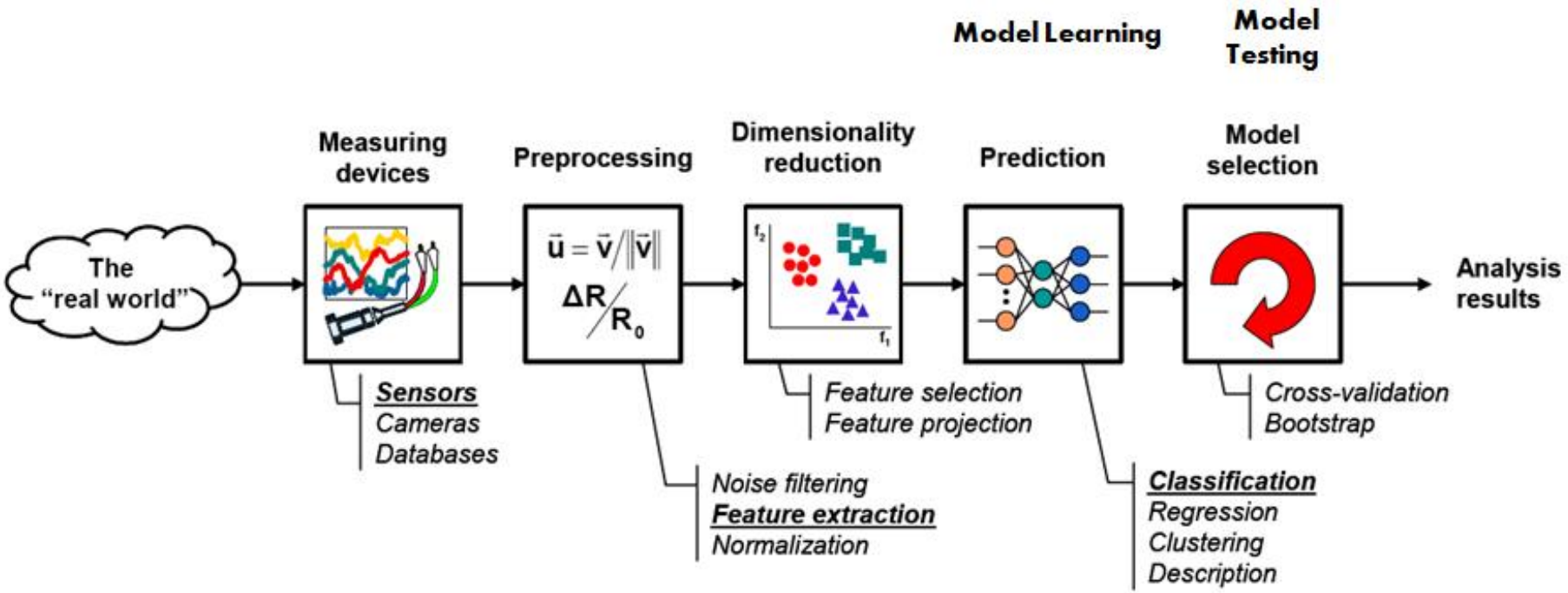
QUÁ TRÌNH HOC MÁY CƠ BẢN



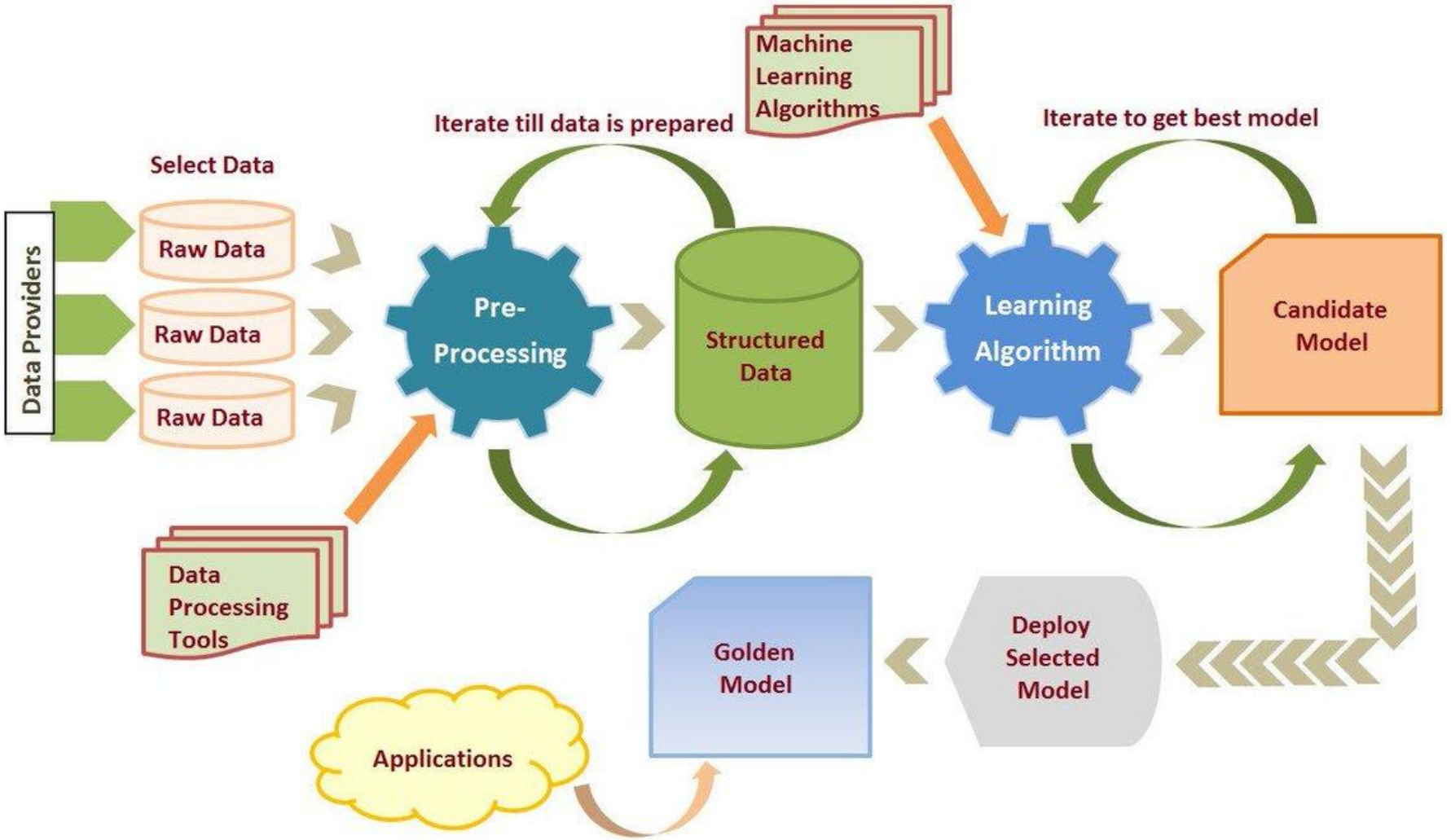
QUÁ TRÌNH HOC MÁY CƠ BẢN



QUÁ TRÌNH HOC MÁY CƠ BẢN

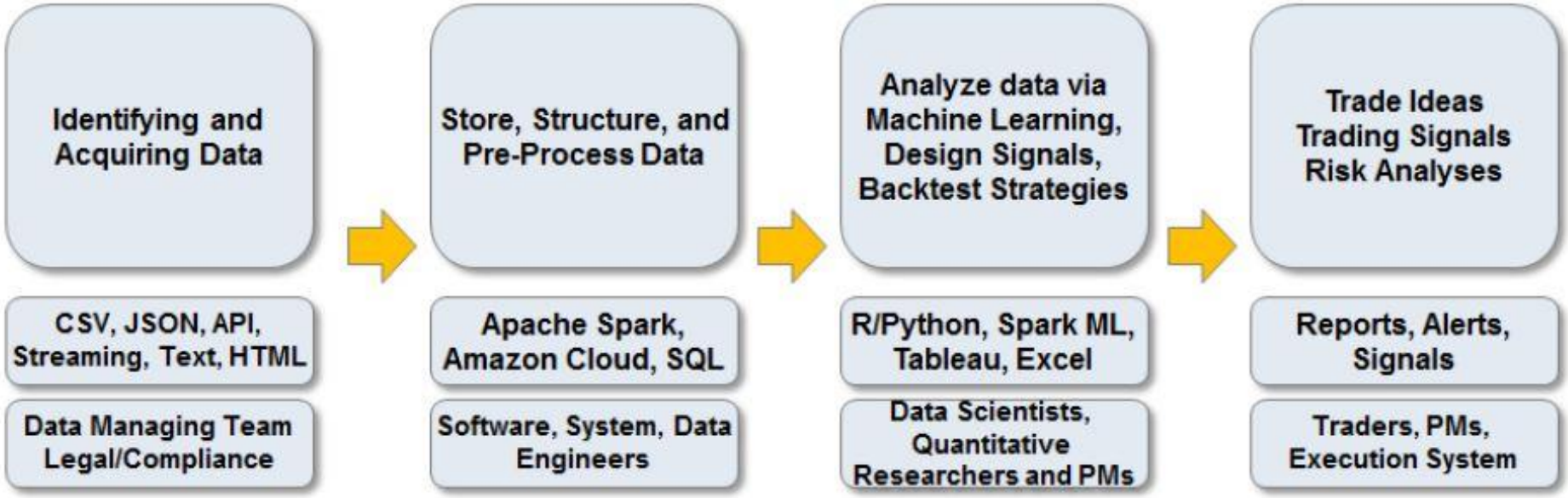


QUÁ TRÌNH HOC MÁY CƠ BẢN



QUÁ TRÌNH HOC MÁY CƠ BẢN

Figure 9: Big Data workflow for investment managers



Source: J.P.Morgan Macro QDS

CÁC BƯỚC THIẾT KẾ MỘT HỆ THỐNG HỌC

Chuẩn bị dữ liệu

- Các giải thuật học máy cần phải có dữ liệu!
- Tiền xử lý dữ liệu để chuyển đổi dữ liệu trước khi áp dụng vào giải thuật học máy
  - Lấy mẫu: chọn tập con các quan sát/mẫu
  - *Trích chọn thuộc tính*: Chọn các biến đầu vào
  - *Chuẩn hóa dữ liệu (Normalization)* (standardization, scaling, binarization)
  - Xử lý dữ liệu thiếu và phần tử ngoại lai (missing data and outliers)
- Ngoài ra, còn phụ thuộc vào giải thuật học máy
  - Cây quyết định có thể xử lý dữ liệu thiếu/phần tử ngoại lai
  - PCA yêu cầu dữ liệu đã được chuẩn hóa

CÁC VẤN ĐỀ LƯU Ý TRONG HỌC MÁY

- 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?

CÁC BƯỚC THIẾT KẾ MỘT HỆ THỐNG HỌC

- Lựa chọn các ví dụ học (training/learning examples)
  - Các thông tin hướng dẫn quá trình học (training feedback) được chứa ngay trong các ví dụ học, hay là được cung cấp gián tiếp (vd: từ môi trường hoạt động)
  - Các ví dụ học theo kiểu có giám sát (supervised) hay không có giám sát (unsupervised)
  - 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 hàm mục tiêu (giả thiết, khái niệm) cần học
  - F: X → {0,1}
  - F: X → Một tập các nhãn lớp
  - F: X → R^+ (miền các giá trị số thực dương)
  - …

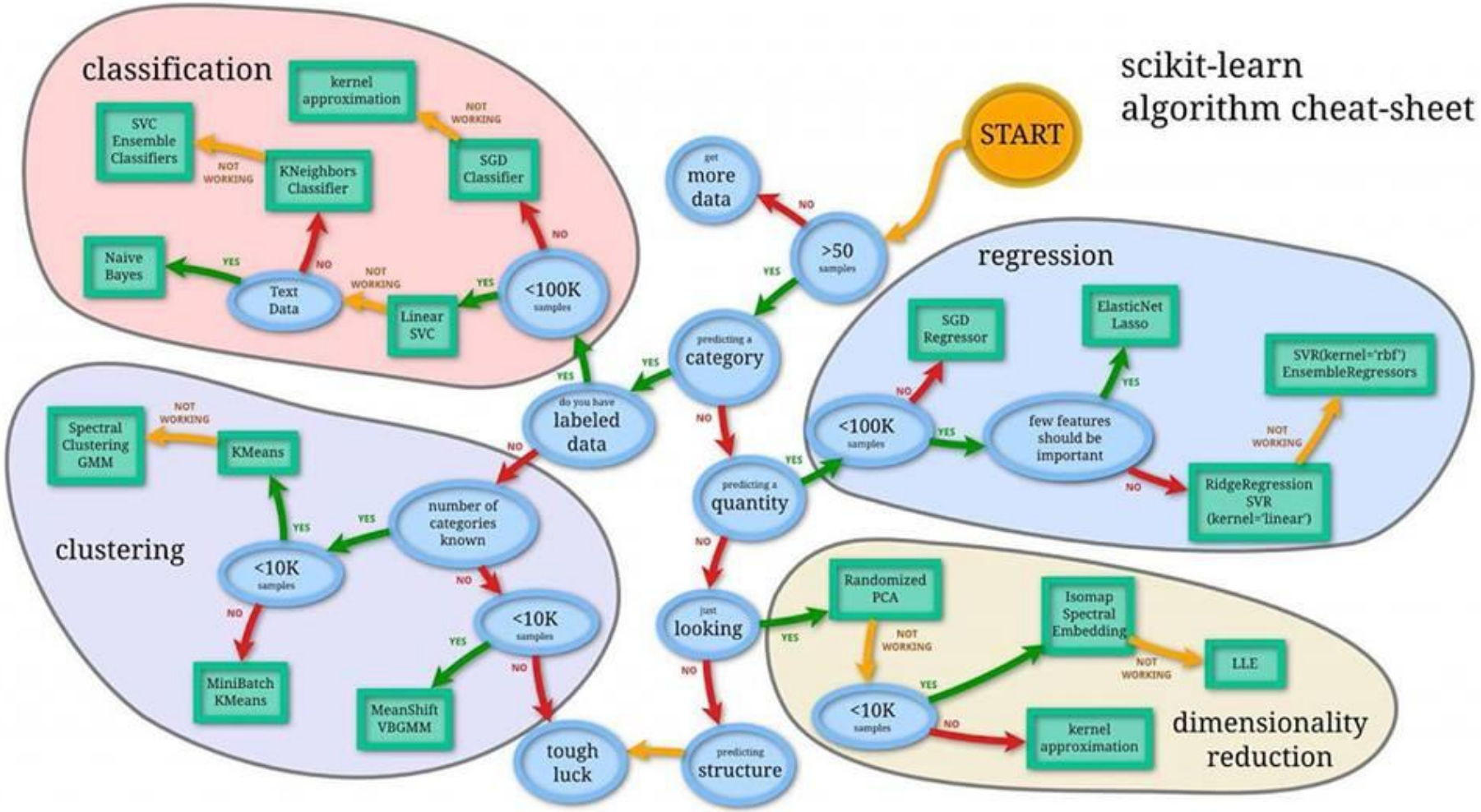
CÁC BƯỚC THIẾT KẾ MỘT HỆ THỐNG HỌC

- 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
  - Phương pháp học hồi quy (Regression-based)
  - Phương pháp học quy nạp luật (Rule induction)
  - Phương pháp học cây quyết định (ID3 hoặc C4.5)
  - Phương pháp học lan truyền ngược (Back-propagation)
  - …

CÁC VẤN ĐỀ LƯU Ý TRONG HỌC MÁY

- 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) 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, 2005]:
  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)

QUÁ TRÌNH HOC MÁY CƠ BẢN



QUÁ TRÌNH HOC MÁY CƠ BẢN

CÁC VẤN ĐỀ LƯU Ý TRONG HỌC MÁY

- 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?

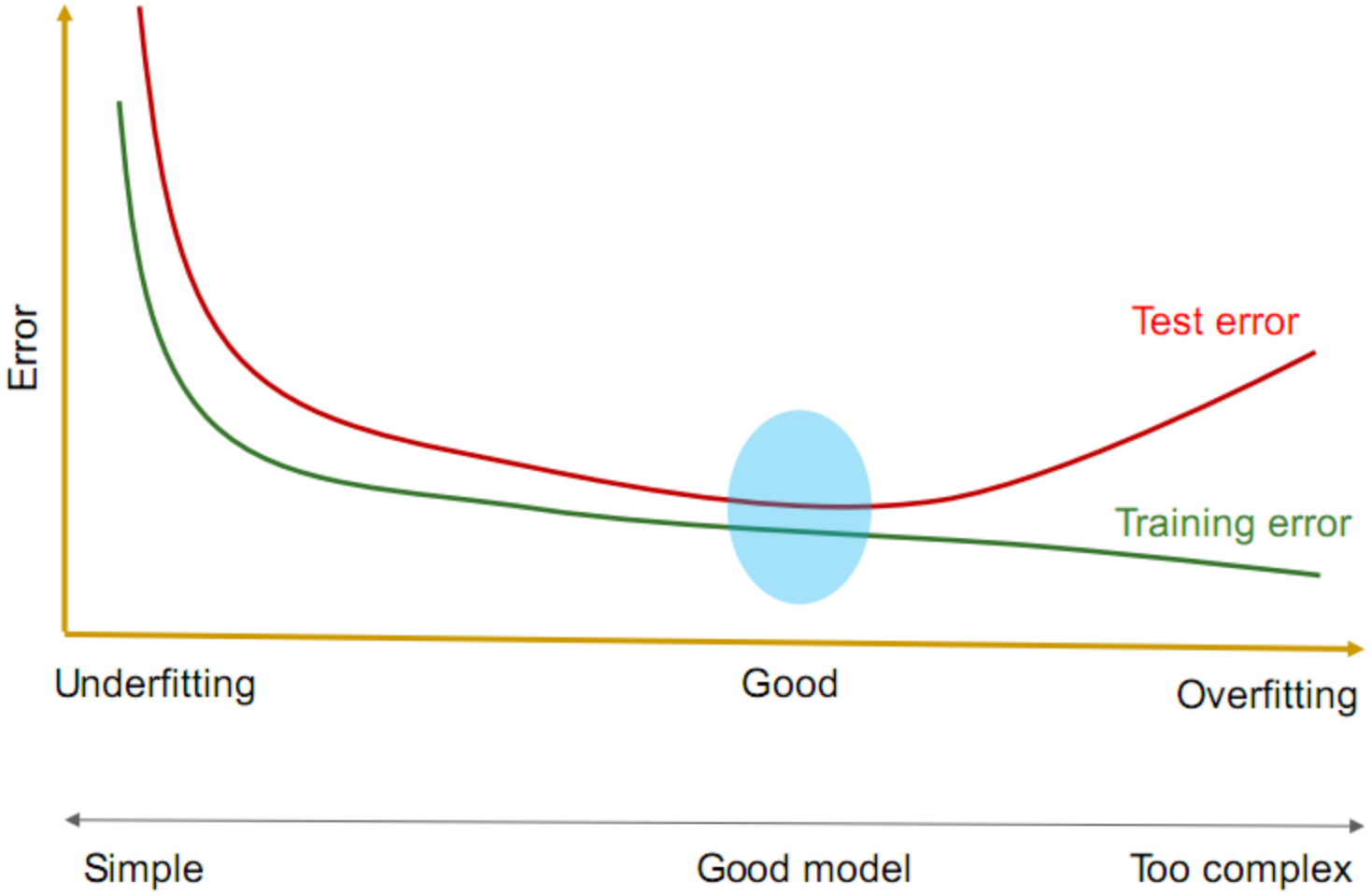
CÁC VẤN ĐỀ LƯU Ý TRONG HỌC MÁY

- 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 (trên lý thuyết) đối với khả năng học của các giải thuật học máy?
  - Khả năng **khái quát hóa (generalization)** của hệ thống?
    - Để tránh vấn đề "over-fitting" (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

OVERFIT

- Một hàm mục tiêu (một giả thiết) học được *h* sẽ được gọi là **quá khớp/quá phù hợp (overfit)** với một tập học nếu tồn tại một hàm mục tiêu khác *h'* sao cho:
  - *h'* kém phù hợp hơn (đạt độ chính xác kém hơn) *h* đối với tập học, nhưng
  - *h'* đạt độ chính xác cao hơn *h* đối với toàn bộ tập dữ liệu (bao gồm cả những ví dụ được sử dụng sau quá trình huấn luyện)
- Vài nguyên nhân:
  - 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

OVERFIT


Error
Test error
Training error
Underfitting
Good
Overfitting
Simple
Good model
Too complex