KHAI PHÁ DỮ LIỆU

VO DUC QUANG – VINH UNIVERSITY

QUANGVD@VINHUNI.EDU.VN

Thông tin chung

Học phần: INF30034-Khai phá dữ liệu

Số tín chỉ: 03 (35/10/90)

Chuyên cần thái độ	Hồ sơ học phần	Giữa kỳ	Cuối kỳ
10%	20%	20%	50%
- Chuyên cần - Phát biểu	01 Project cá nhân	Trắc nghiệm	- Thực hành trên lớp - Thi tự luận cuối kỳ

Giáo trình:

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, **Introduction to Data Mining**, Pearson, 2013

Tài liệu tham khảo:

[2] Wes McKinney, **Python for Data Analysis**, 2nd Edition, Orailly, 2017

[3] Robert Layton, **Learning Data Mining with Python**, 2nd Edition, Packt Publishing, 2017

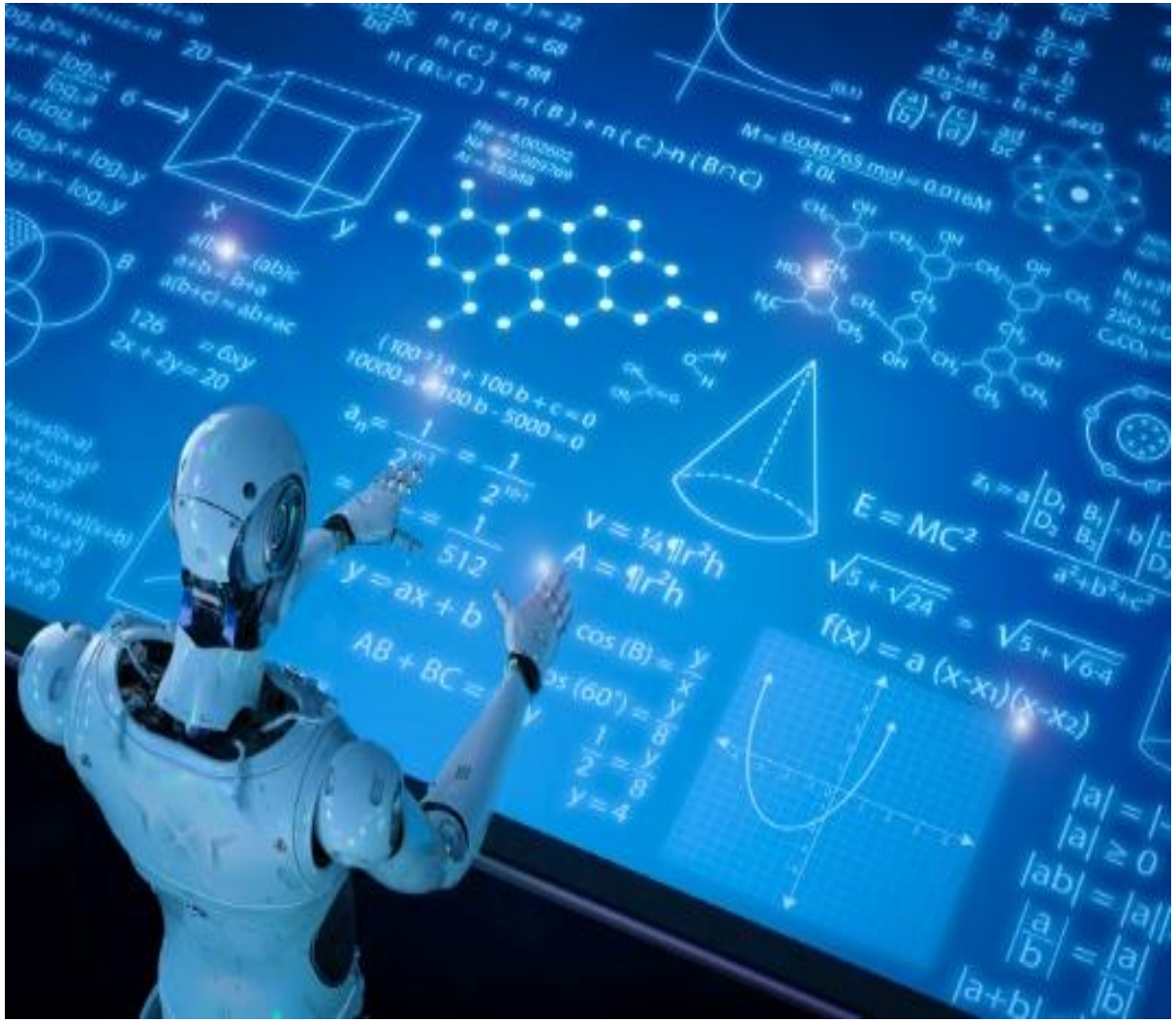
Nội dung

- Chương 1: Tổng quan về Data Mining và Machine Learning
- Chương 2: Dữ liệu và tiền xử lý dữ liệu
- Chương 3: Bài toán phân lớp dữ liệu
- Chương 4: Bài toán phản cụm dữ liệu
- Chương 5: Khai phá luật kết hợp

Chương 1 – Giới thiệu

- Giới thiệu Machine Learning (ML) và Data Mining (DM)
- Phân loại một số bài toàn
  - Học có giám sát (Supervised learning)
  - Học không giám sát (Unsupervised learning)
- Đánh giá hiệu năng (Performance evaluation)
- Một số kinh nghiệm

Trí tuệ nhân tạo - Artificial Intelligence (AI)



Trí tuệ nhân tạo - Artificial Intelligence (AI)

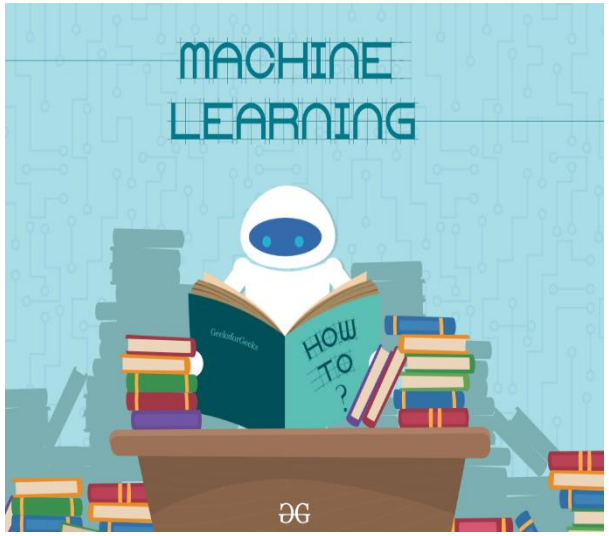
- Trí thông minh được mô tả bằng máy móc (máy tính) có khả năng bắt chước các chức năng "nhận thức" mà con người thường phải liên kết với tâm trí, như "học tập" và "giải quyết vấn đề".
- Phân loại các lĩnh vực AI
  - Nhận thức, Phân tích vấn đề
  - Lấy cảm hứng từ con người: mô phỏng hình thể, nhận thức, cảm xúc, ra quyết định
  - Nhân cách hóa: Tự ý thức, tự nhận thức
- Ví dụ:
  - Deep fake: https://www.youtube.com/watch?v=mUfJOQKdtAk
  - Robot Sophia
  - Phân loại email rác, ô tô tự lái, dự đoán chứng khoán,...

Why ML & DM?

- **"The most important general-purpose technology of our era is artificial intelligence, particularly machine learning"** – Harvard Business (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 (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-theworld.html?pagewanted=all&_r=0)
  - https://www.vietnamworks.com/AI-Engineer-kv
  - https://vn.sputniknews.com/vietnam/202101291000741 2-viet-nam-muon-dan-dau-ve-ai-thanh-cuong-quoc-cong-nghe-the-gioi/

Why ML & DM?

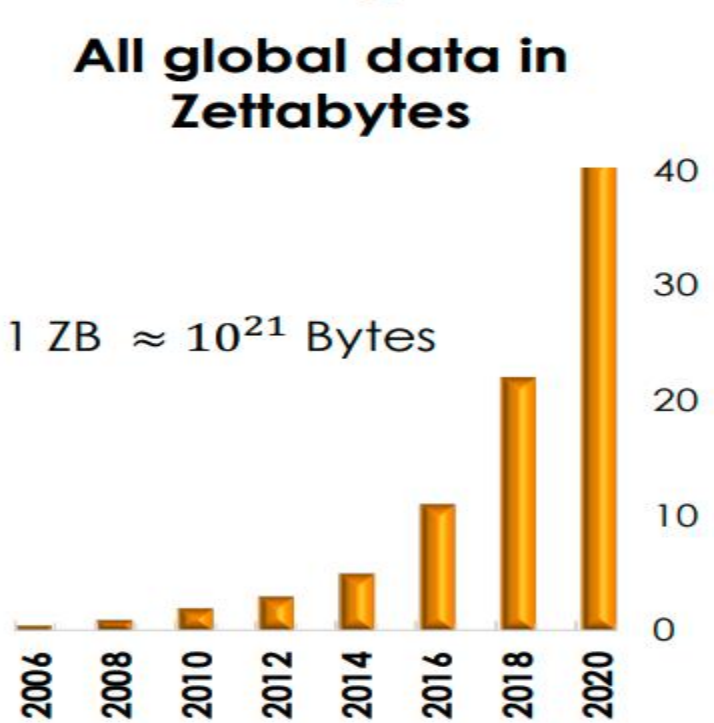
- AI là đích đến
- ML là một cách tiếp cận để huấn luyện xây dựng hệ thống ứng dụng
- DM+BigData là các phương pháp kỹ thuật hỗ trợ (khai phá tri thức mới)



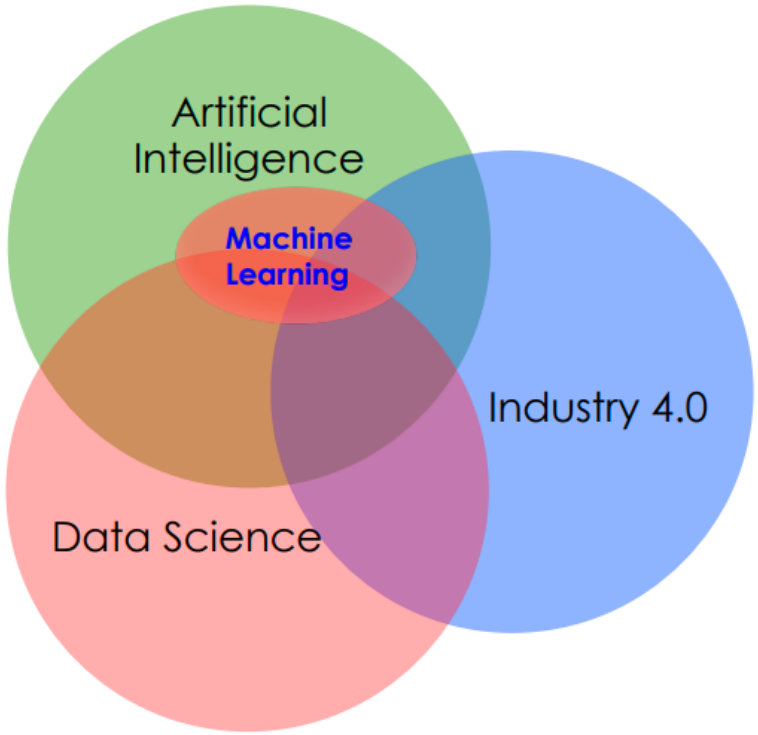
Why ML & DM?

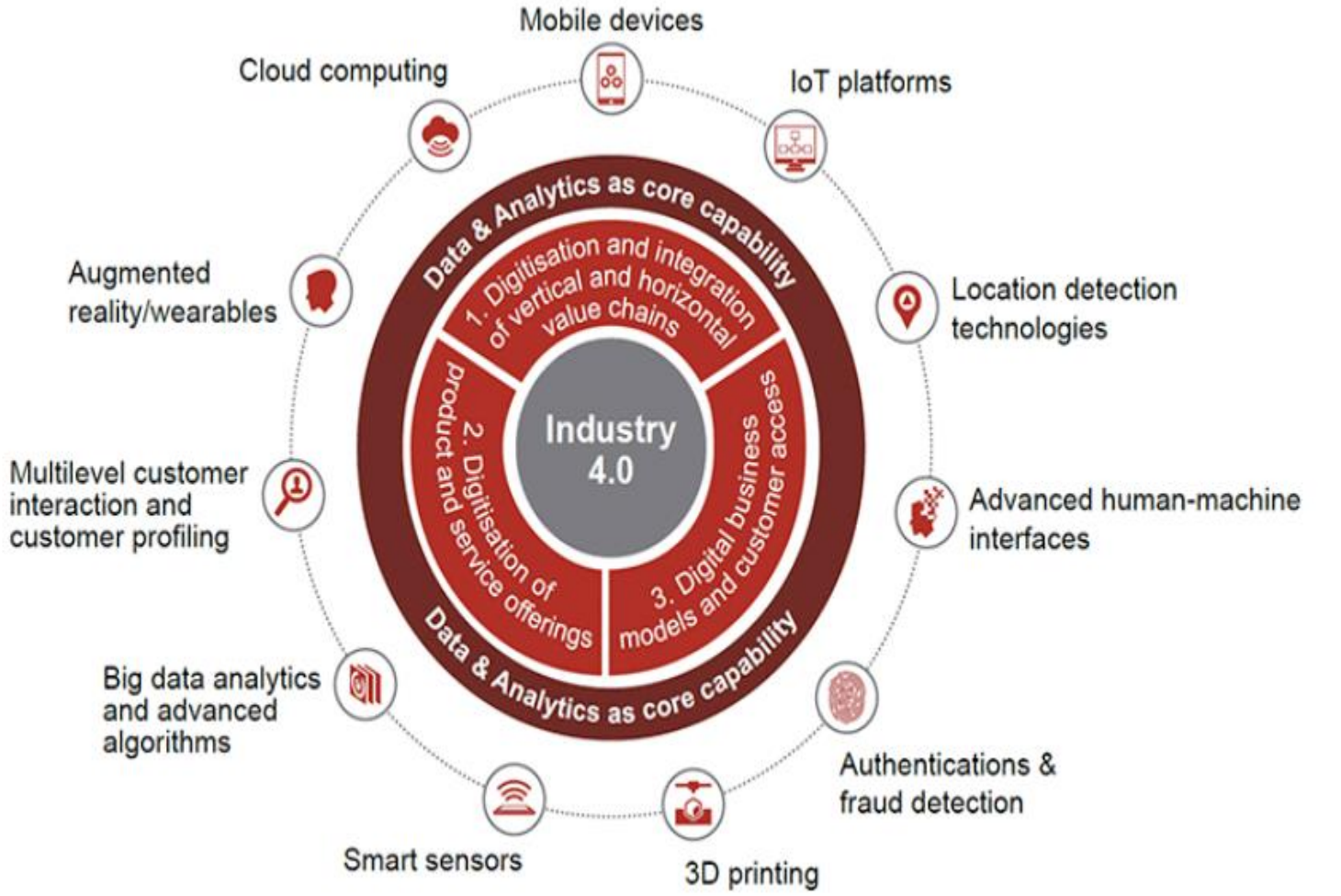
- Data mining, inference, prediction
- ML & DM provides an efficient way to make intelligent systems/services.
- ML provides vital methods and a foundation for **Big Data**.

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



Why Industry 4.0? DS? AI? ML? DM?





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

Một số thành công

- IBM’s Watson (https://vi.wikipedia.org/wiki/Watson_(ph%E1%BA%A7n_m%E1%BB%81m_tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o)
- GAN - Deepfake

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

Một số thành công

- Sophia (Hanson Robotics 2015)
  - Da silicon, 62 sắc thái biểu cảm, chatbot,…
  - Ngày 25 tháng 10 năm 2017, Sophia là Robot đầu tiên được chính phủ Ả Rập Xê Út cấp quyền công dân như con người
- AlphaGO (Google DeepMind 2016)
  - Cờ vây là trò chơi rất phức tạp 2500 năm
  - Đánh bại Lee Sedol (World champion)
- GPT-3
  - Generative Pre-training Transformer OpenAI 2020
  - Tạo khả năng Viết cho máy tính
  - https://quantrimang.com/gpt-3-la-gi-cong-nghe-gp 17

ML & DM?

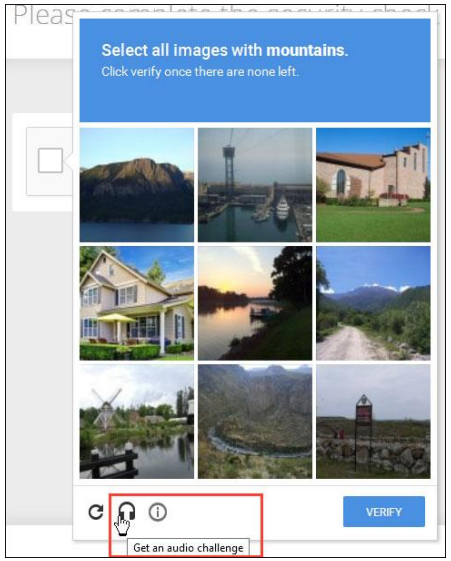
- Machine Learning (ML - Học máy)
  - **To build computer systems that can improve themselves by learning from data.**
  - (Xây dựng những hệ thống mà có khả năng tự cải thiện bản thân bằng cách học từ dữ liệu)
- Data Mining (DM - Khai phá dữ liệu)
  - **To find new and useful knowledge from datasets.**
  - (Tìm ra/Khai phá những tri thức mới và hữu dụng từ các tập dữ liệu lớn)

DATA

- Data?
- Database?
- Dataset?
- DBMS, RDBMS?
- No SQL?

	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


```
{
    "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"
}
```

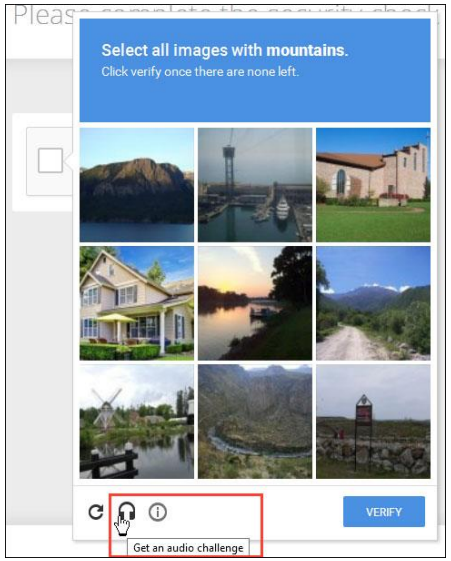



DATA

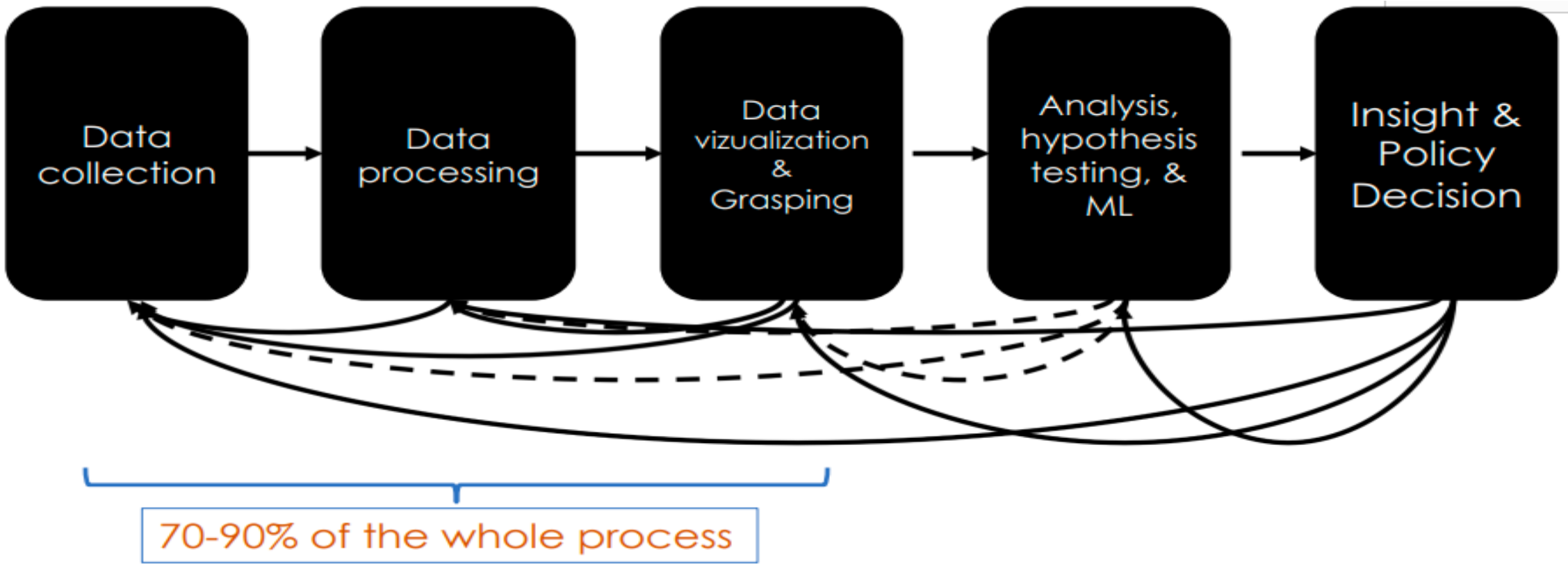
- Structured
- Un-Strucred
- Text
- Images
- Voice
- …

	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


```
{
    "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"
}
```

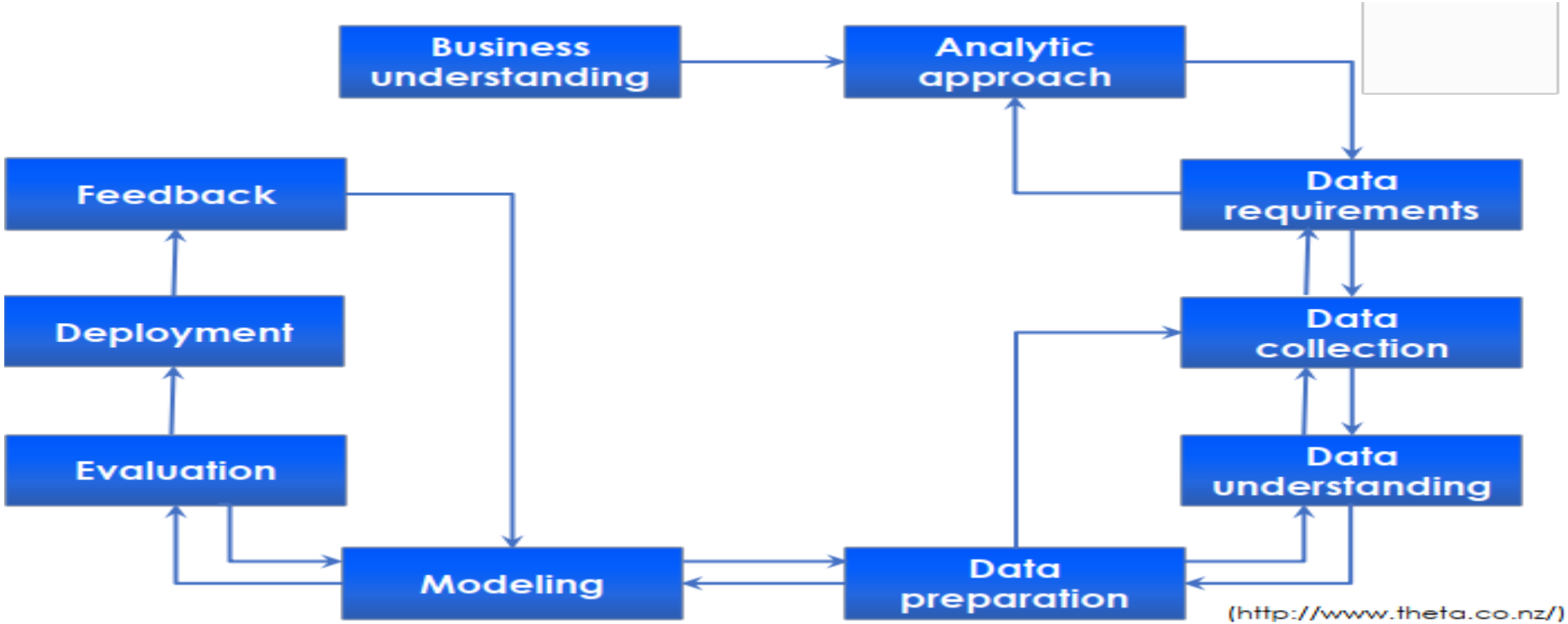
Tiến trình khai phá dữ liệu



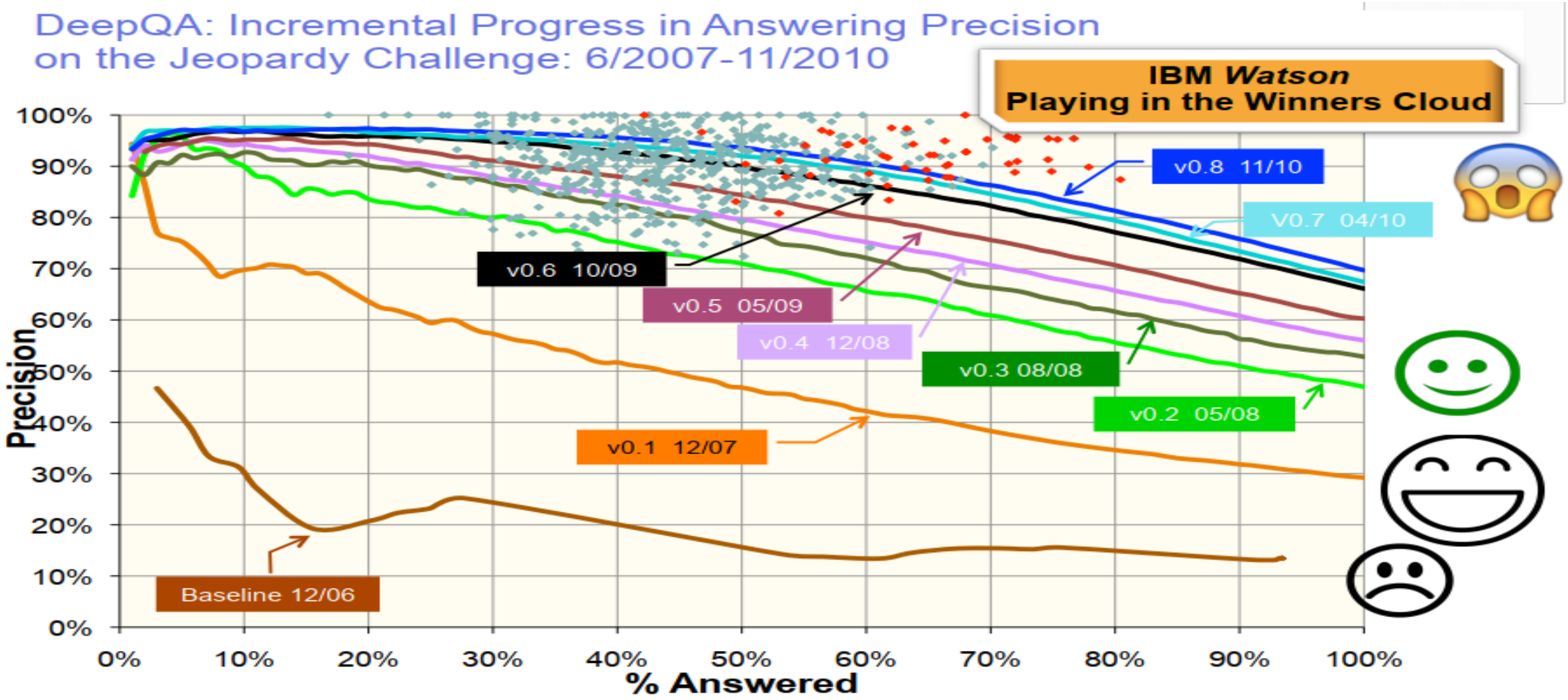
(John Dickerson, University of Maryland)

Tiến trình khai phá dữ liệu

- Tiếp cận hướng sản phẩm

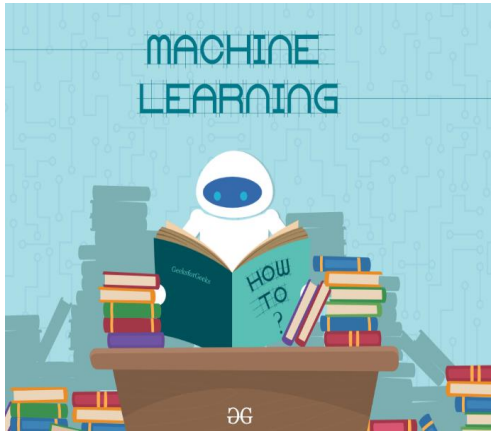


Ví dụ dự án sản phẩm AI



Bản chất của Machine Learning

- **Learns** ???!!!
- We say that a machine learns if the system reliably improves its performance **P** at task **T**, following experience **E**
- **P**: hiệu năng, hiệu suất
- **T**: Công việc, nhiệm vụ
- **E**: Kinh nghiệm



Ví dụ: Phân loại email

- **(P, T, E) ???**
- T : Phân loại lọc các email spam (rác)
- P : Tỷ lệ phân loại chính xác (accuracy) các email vào đúng nhóm normal/spam
- E : Tập các email cũ đã được nhận biết (gán nhãn) chính xác normal/spam

Ví dụ gán nhãn ảnh (tagging)

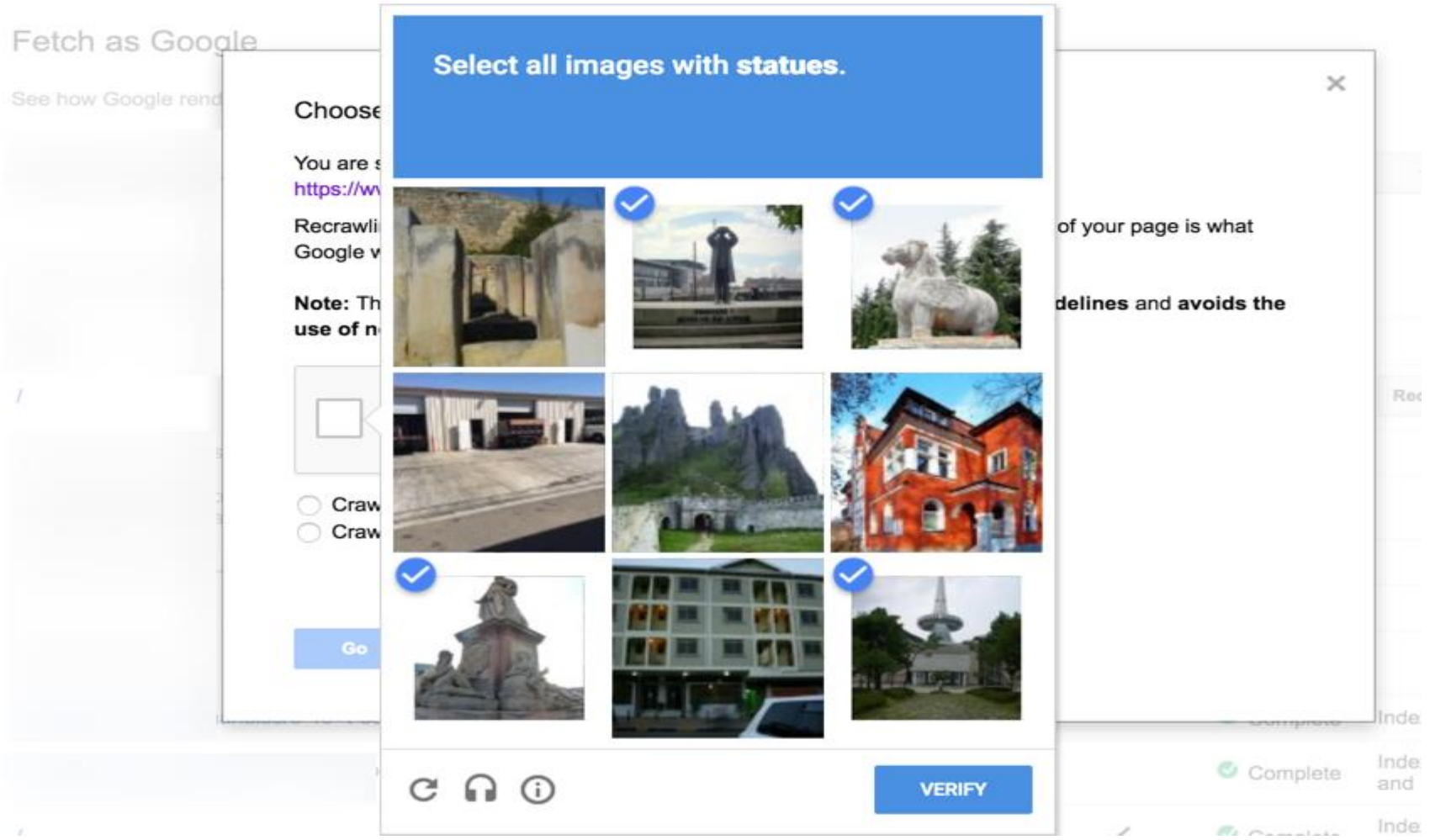
- Image tagging
  - **T**: give some words that describe the meaning of a picture.
  - **P**: ?
  - **E**: set of pictures, each has been labelled with a set of words.

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Ví dụ gán nhãn ảnh (tagging)



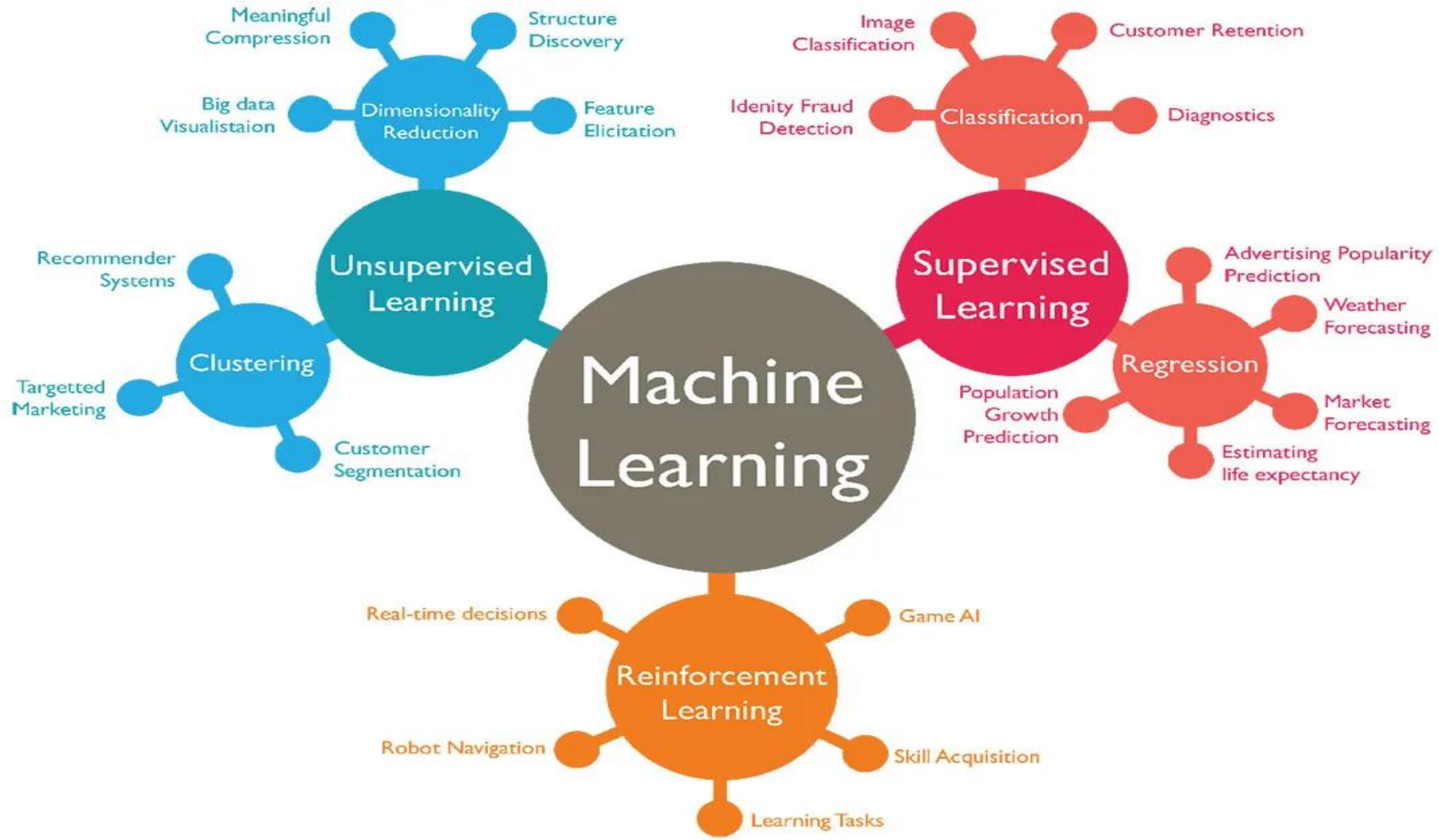
Bản chất của Machine Learning

- Đi tìm một mô hình (model)

$$f : x \longmapsto y$$

- X: những quan sát (observations) data, những kinh nghiệm trong quá khứ
- y: đầu ra: dự đoán (prediction), những kiến thức mới, những kinh nghiệm mới…
- F ???
- Ta đôi khi giả sử rằng dữ liệu thường tuân theo một mô hình nào đó
- Học (trong học máy) là học ra những mô hình, hay chính xác hơn là học ra các tham số xác định mô hình đó

Các bài toán trong Machine Learning



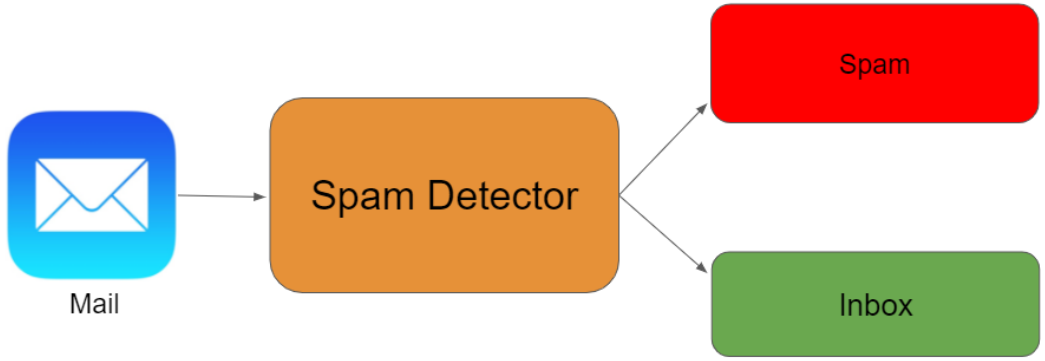
Phân loại các phương pháp học máy

- **Supervised learning (học có giảm sát):** đưa vào tập các ví dụ mẫu có đáp án để máy học và đưa ra dự đoán cho những ví dụ khác chưa có câu trả lời
  - Bài toán phân lớp (Classification)
  - Bài toán hồi quy (Regression)
- **UnSupervised learning (học không giảm sát):** giải thuật cố gắng khai thác cấu trúc ẩn của một tập dữ liệu mà không cần câu trả lời mẫu
  - Bài toán phân cụm (Clustering)
- **Semi-Supervised learning (học bán giám sát)**
- **Reinforcement learning (học tăng cường):** giải thuật ghi nhận được phản hồi cho mỗi hành động để điều chỉnh hoạt động cho phù hợp

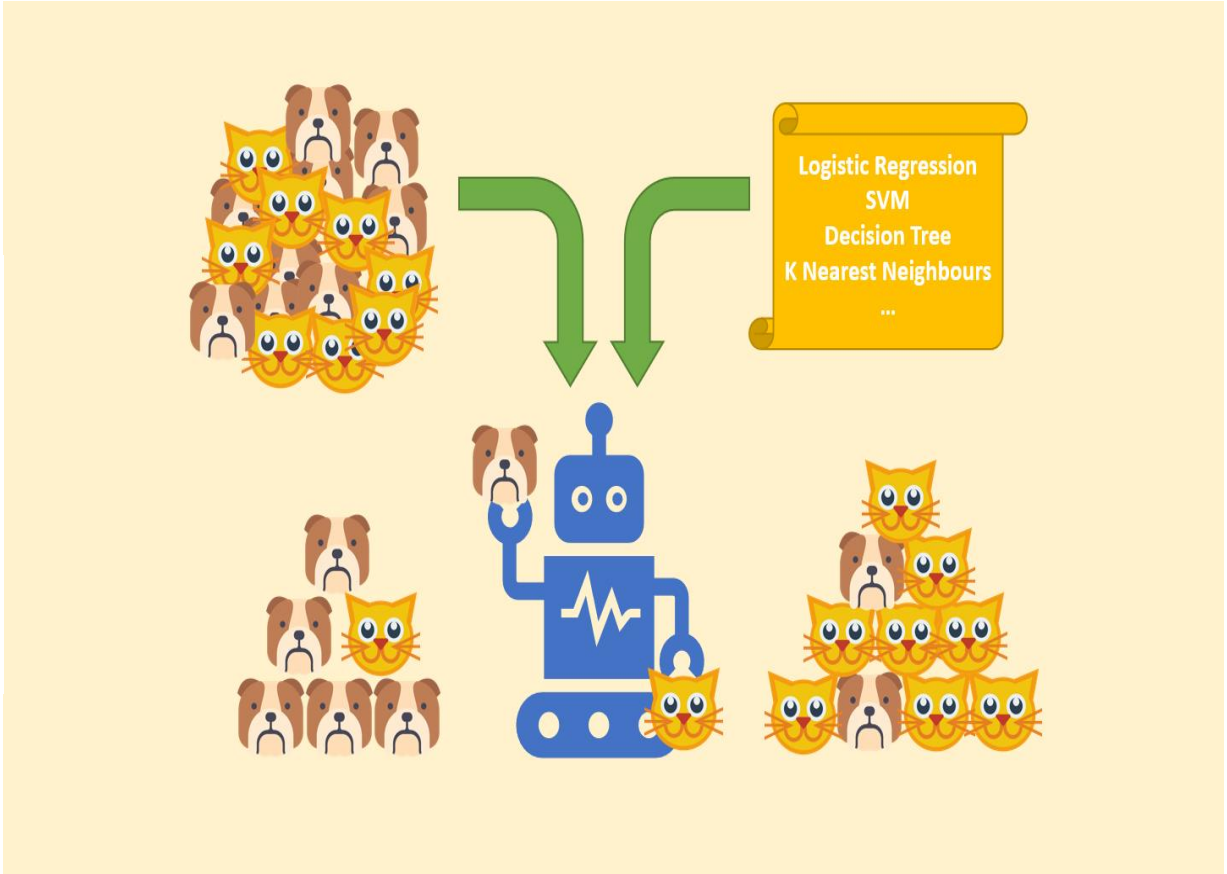
Ví dụ: Phân lớp - Classification

$$y = f(x)$$

- Y: Nhãn lớp

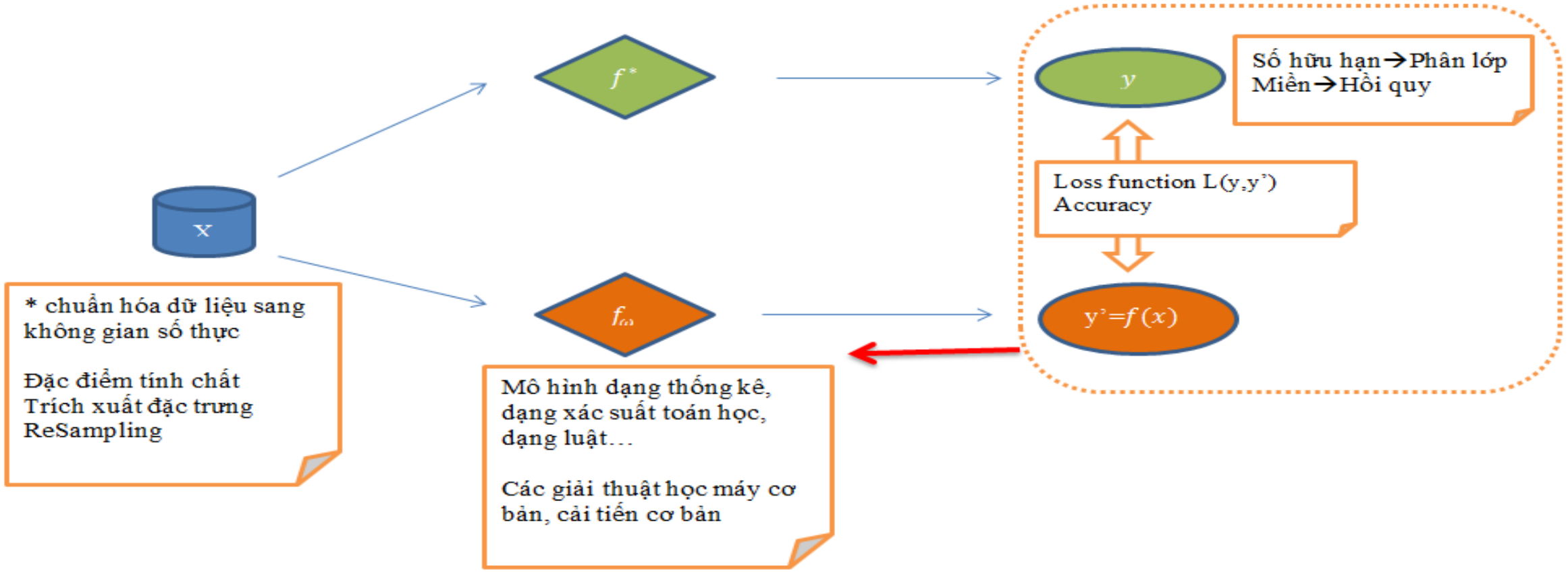


- Nhận dạng ảnh động vật



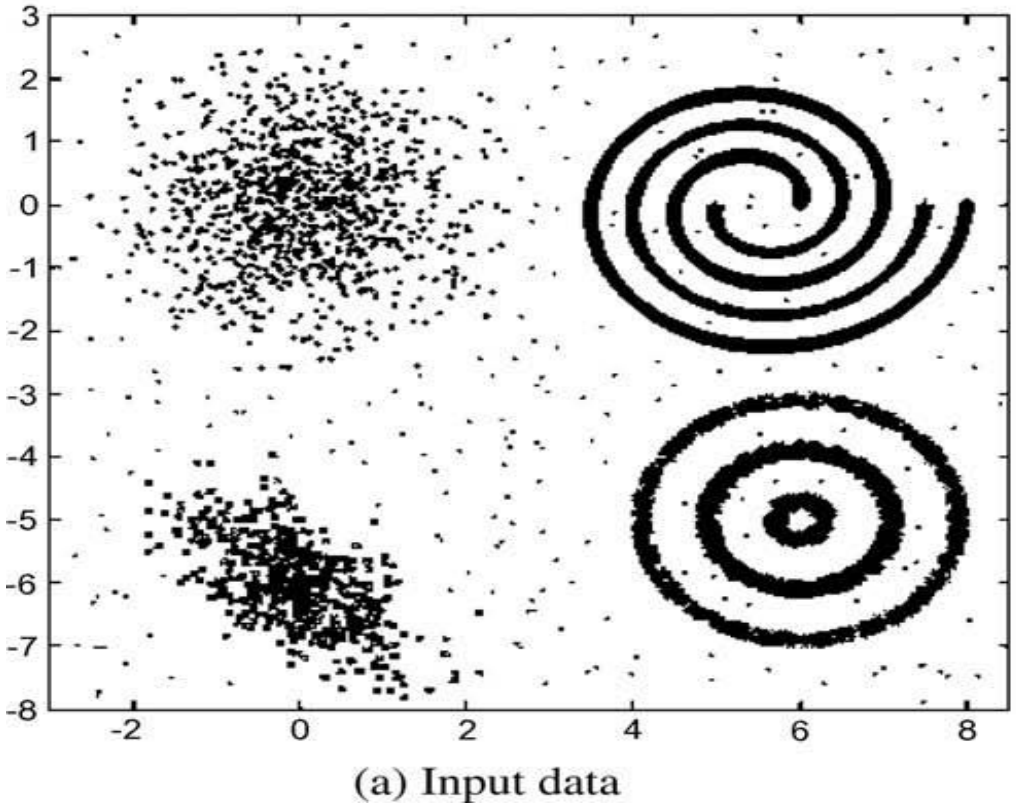
Mô hình Phân lớp

- **Training set** (Tập mẫu huấn luyện): {(x1,y1); (x2,y2), …, (xN,yN)}

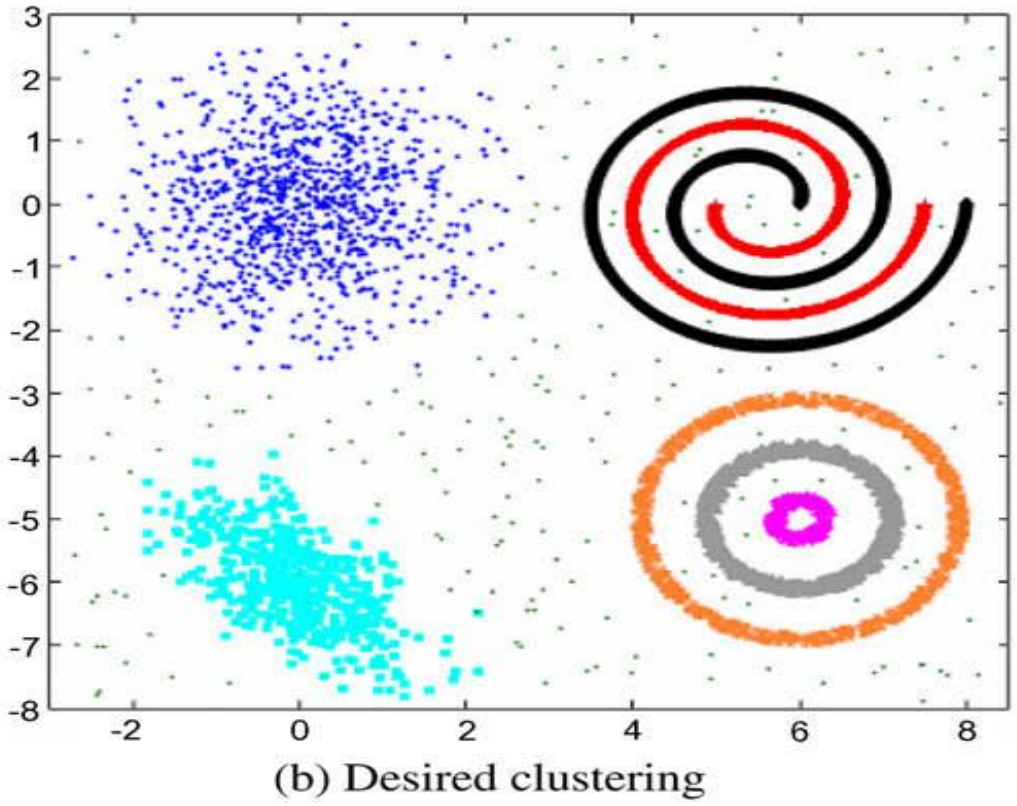


Why ML & DM?

- Training set (Tập mẫu huấn luyện): {(x1,); (x2,), …, (xN,)}



(a) Input data



(b) Desired clustering

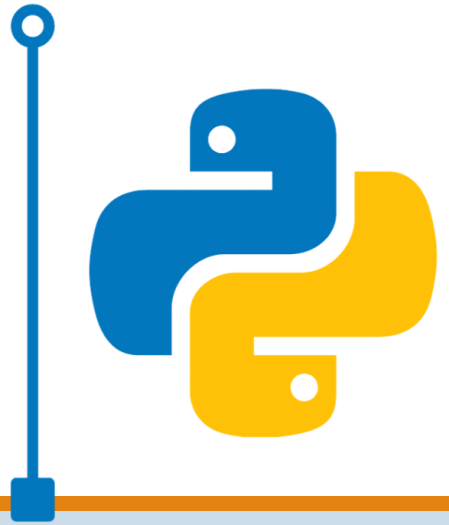
Ví dụ: Hồi quy - Regression

$$y = f(x)$$

- y: Là một số thực

Chuẩn bị về công cụ và công nghệ

- Software:
  - WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
  - Anaconda, Python (https://www.anaconda.com/)
  - Scikit-Learn (http://scikit-learn.org/)
  - Visual Studio Code
- Data for experiments:
  - UCI repository: http://archive.ics.uci.edu/ml/



Yêu cầu:

- Cài đặt Anaconda và các môi trường thư viện
  - Python
  - Numpy
  - Pandas
  - Scikit-learn
  - Tensorflow