KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

MÁY HỌC ỨNG DỤNG

Applied Machine Learning

Giáo viên giảng dạy:
TS. TRẦN NGUYỄN MINH THƯ
tnmthu@cit.ctu.edu.vn

Nội dung

- Ứng dụng của máy học
- Máy học là gì?
- Phân loại máy học
- Link datasets

Ứng dụng của Máy học

Ứng dụng của Máy học

Classification Example: Weather Prediction

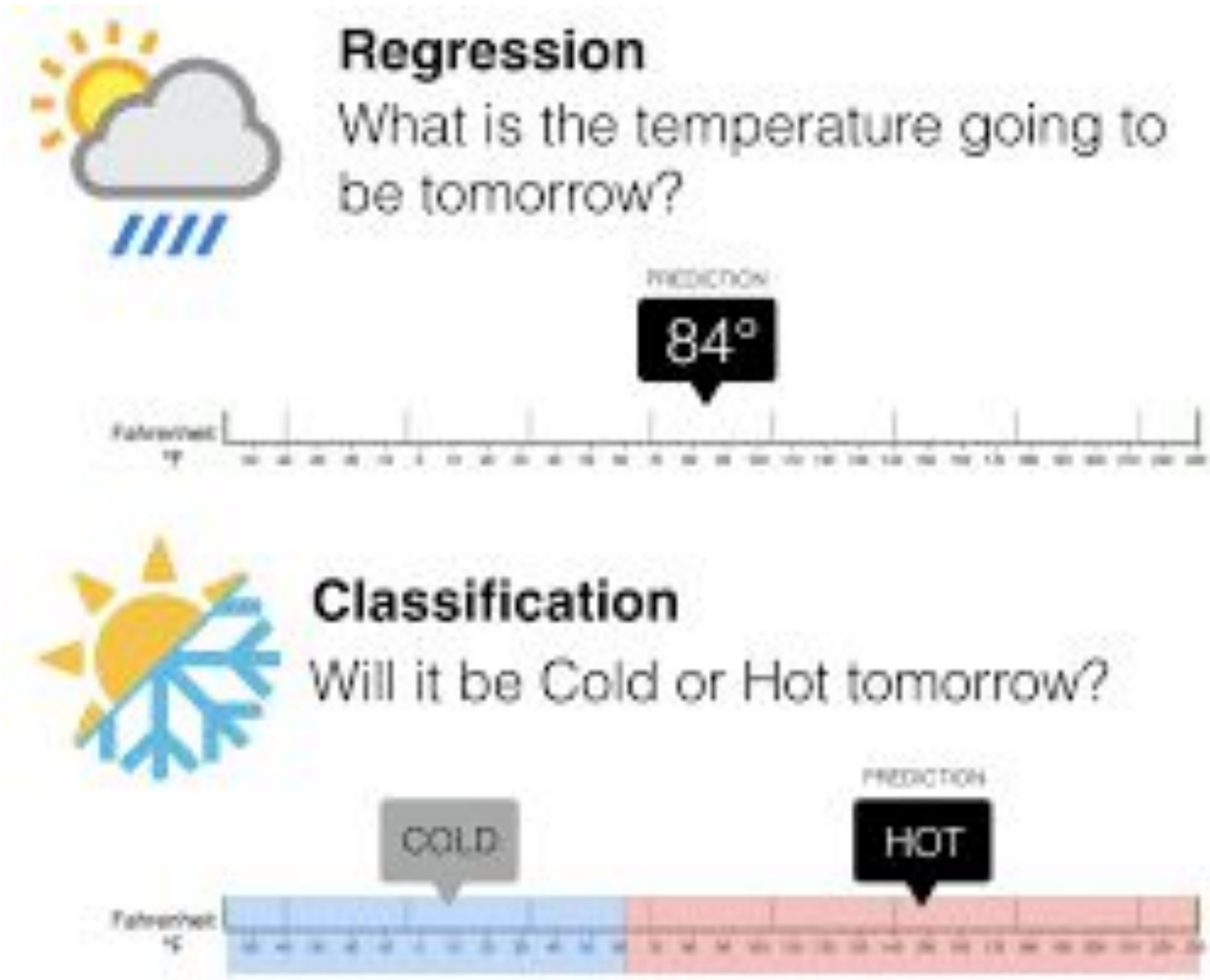
Ứng dụng của Máy học

Regression example: Predicting Gold/Stock prices

Good ML can make you rich (but there is still some risk involved).

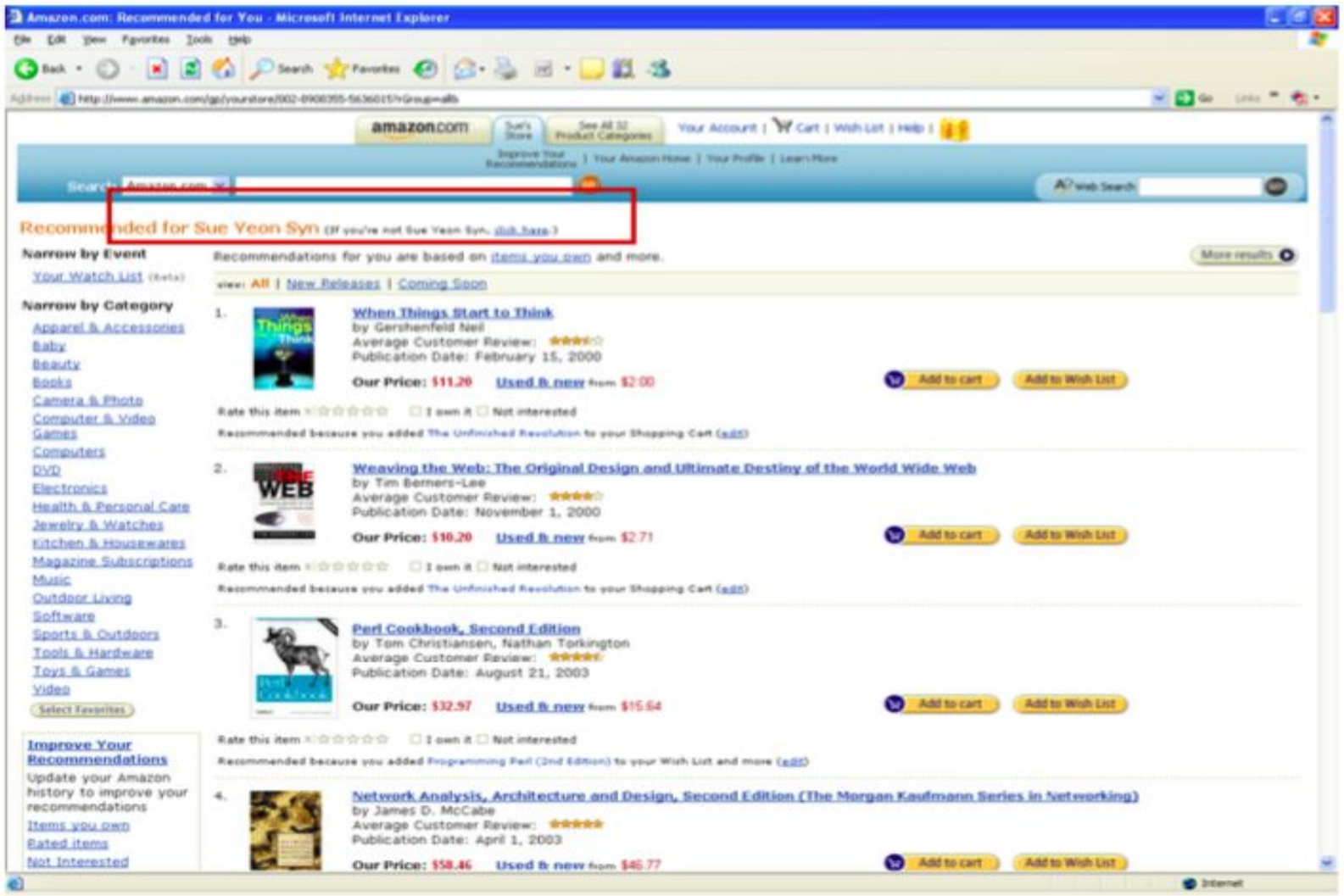
Given historical data on Gold prices, predict tomorrow's price!

Ứng dụng của Máy học

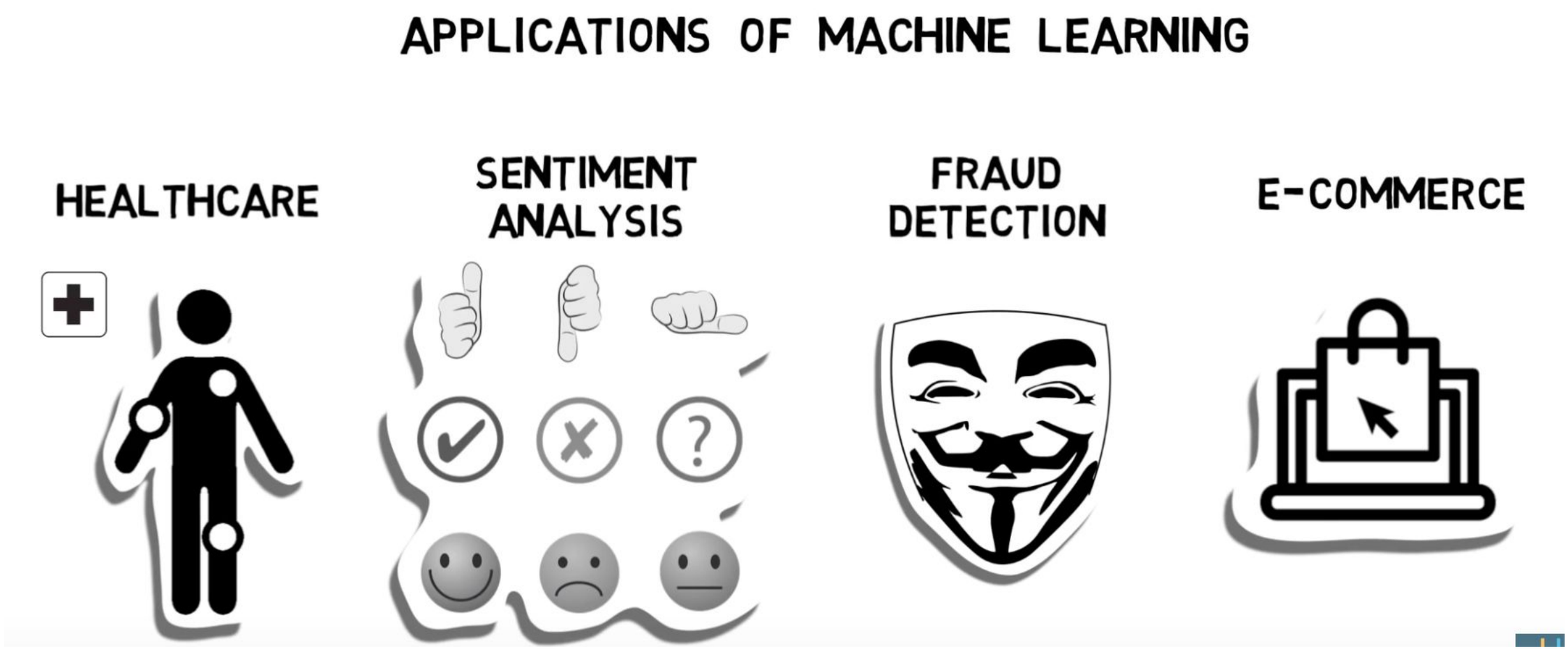


Ứng dụng của Máy học

Collaborative Filtering

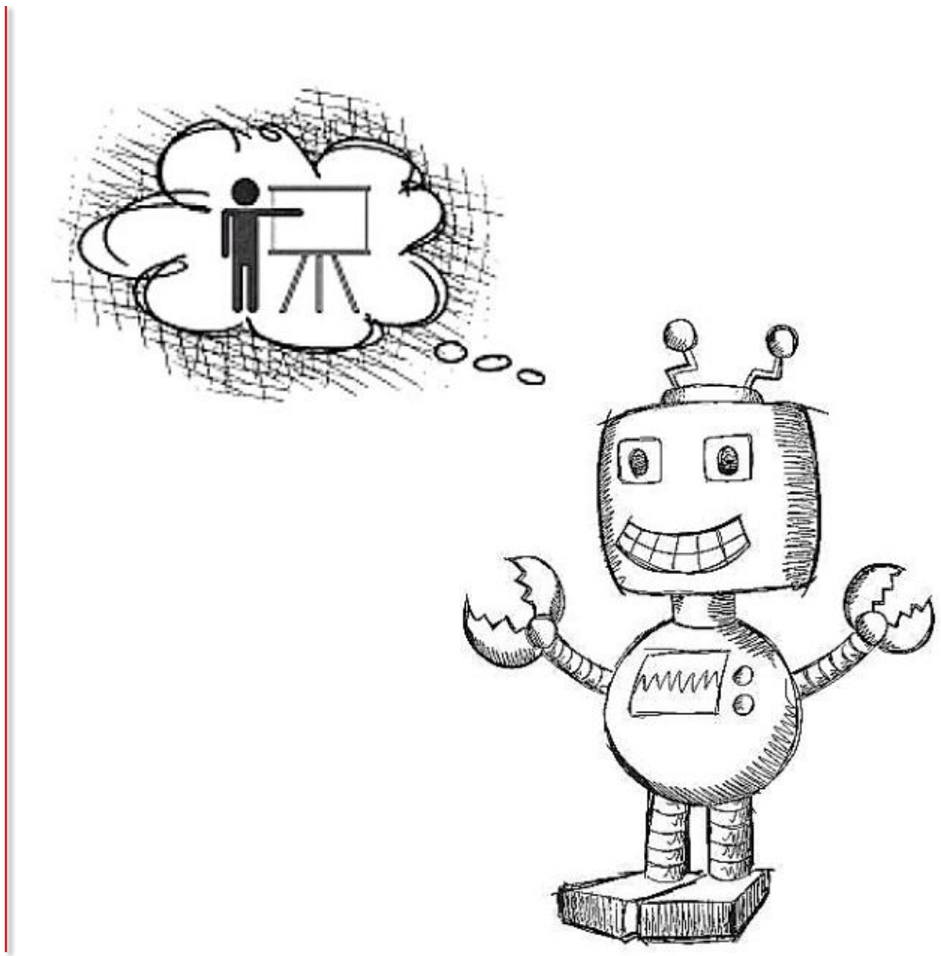


Ứng dụng của Máy học



https://www.kdnuggets.com/2018/11/10-free-must-see-courses-machine-learning-data-science.html

Máy học là gì?



Máy học là gì?


Learn from experience
Learn from data
Follow instructions

Máy học là gì?

Máy học là chương trình máy tính cho phép **học tự động** từ **dữ liệu** để nhận dạng các mẫu phức tạp, tạo ra hành vi ứng xử thông minh với trường hợp mới đến [T. Mitchell, 1997]

T. Mitchell: machine Learning: improving performance via **experience**

Học= Cải thiện tác vụ (task) nào đó bằng kinh nghiệm

Máy học là gì?

T. Mitchell:
– machine learning: improving performance via experience
– Formally, A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, it its performance at tasks in T as measured by P, improves with experience.

(Mitchell, 1997): một chương trình máy tính được gọi là học từ kinh nghiệm E với một vài lớp của vấn đề T và độ đo hiệu quả P, nếu hiệu năng của vấn đề trong T, đánh giá theo tiêu chí P, được cải thiện từ kinh nghiệm E

– Cải thiện tác vụ T,
– Với độ đo hiệu quả P
– Dựa trên kinh nghiệm E

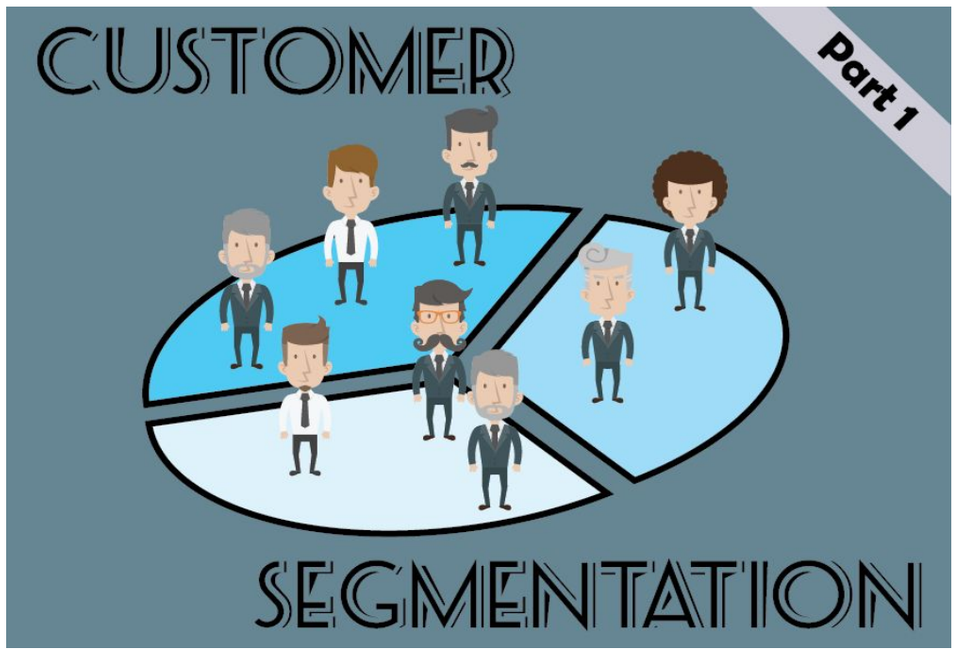
Các kiến thức khác có liên quan

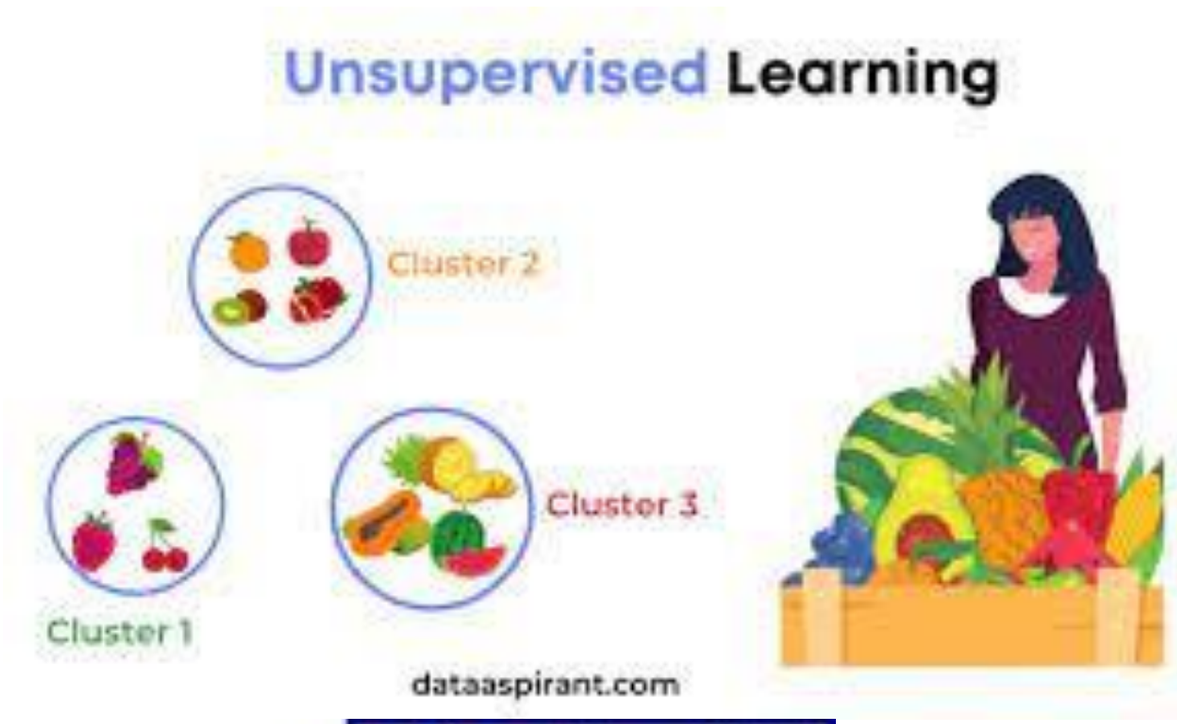
- Trí tuệ nhân tạo
- Phương pháp Bayes
- Lý thuyết độ phức tạp tính toán
- Lý thuyết điều khiển
- Lý thuyết thông tin
- Triết học
- Tâm lý học và thần kinh học (neurobiology)
- Thống kê
- …

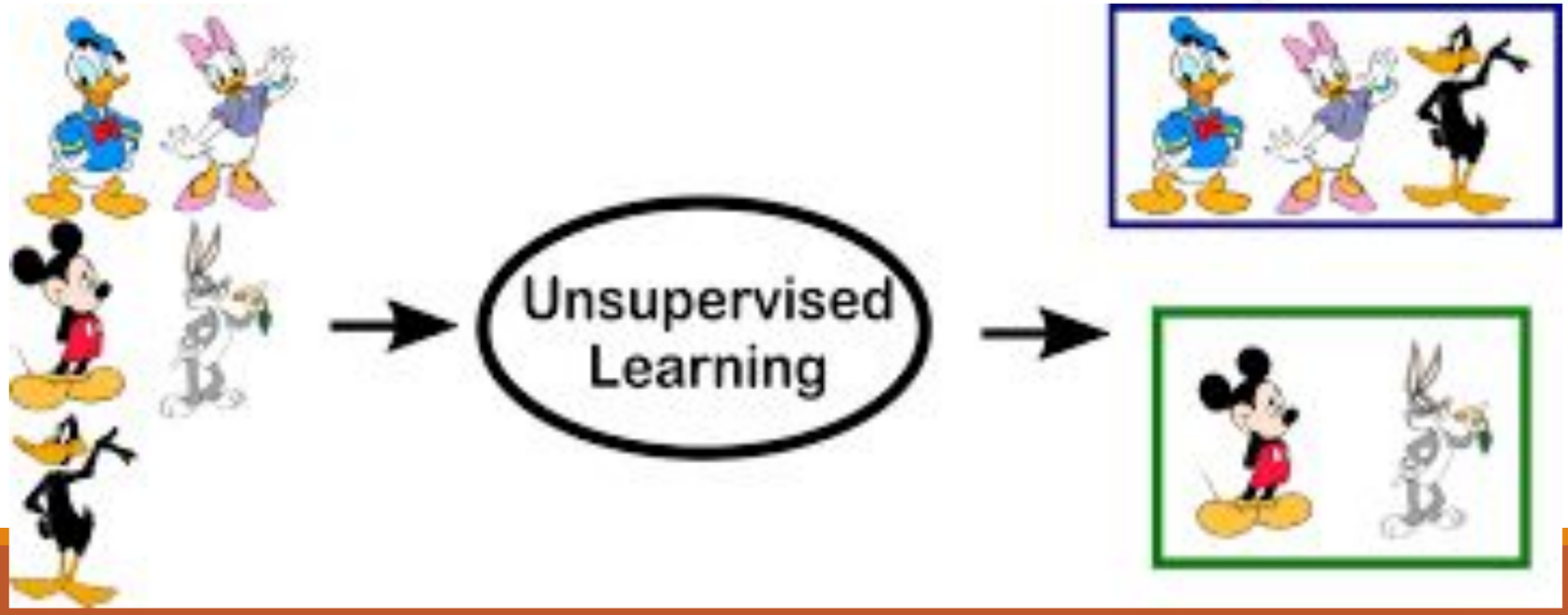
Phân loại học máy

1. **Học không có giám sát – unsupervised learning**
2. Học có giám sát – supervised learning
3. Học bán giám sát- semi- supervised learning
4. Học củng cố / học tăng cường: reinforcement learning

Phân loại học máy – **học không giám sát**







Phân loại học máy – **học không giám sát**

- Học không giám sát là thuật toán học thực hiện mô hình hoá một tập dữ liệu đầu vào**, không được gán nhãn** (lớp, giá trị cần dự báo)
  - **gom cụm, nhóm** (clustering, unsupervised classification): xây dựng mô hình gom cụm dữ liệu tập học (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau
- Xây dựng **mô hình H** từ ***tập dữ liệu*** $(X^1, X^2, .., X^m)$
  - Hierachical Clustering
  - Kmeans, …

Phân loại học máy

1. **Học không có giám sát – unsupervised learning**
2. **Học có giám sát – supervised learning**
3. Học bán giám sát- semi- supervised learning
4. Học củng cố / học tăng cường: reinforcement learning

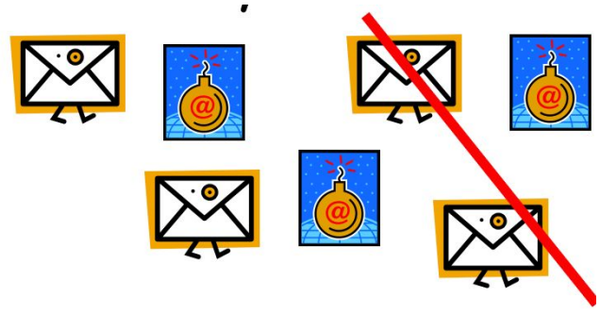
Phân loại học máy – **học có giám sát**

- Học có giám sát là thuật toán học tạo ra một hàm ánh xạ dữ liệu đầu vào tới kết quả đích mong muốn (nhãn, lớp, giá trị cần dự báo). Trong học có giám sát, tập dữ liệu dùng để huấn luyện phải **được gán nhãn, lớp hay giá trị cần dự báo**
- Xây dựng mô hình H được huấn luyện từ tập dữ liệu $\{(\mathbf{X}^1, \mathbf{y}^1), (\mathbf{X}^2, y^2), .., (\mathbf{X}^n, y^n)$
  - **Bài toán hồi quy** (regression): y là giá trị liên tục
  - **Bài toán phân lớp (**classification): y là giá trị **không** liên tục

Phân loại học máy – **học có giám sát**

- **phân lớp** (classification, supervised learning) : xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có **nhãn** (lớp) **là kiểu liệt kê**

  VD: có sẵn tập dữ liệu thư điện tử, mỗi thư có 1 nhãn là thư rác hay thư bình thường, mục tiêu là xây dựng mô hình phân lớp tập dữ liệu thư điện tử thành thư rác hay thư bình thường để khi có một thư điện tử mới đến thì mô hình dự báo được thư này có phải là thư rác hay không



- **hồi quy (regression)** : xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp) là **giá trị liên tục**.

VD. Xd mô hình dự báo mực nước sông Mekong từ các yếu tố thời tiết, mùa,…

Từ tập dữ liệu học/huấn luyện $\{ (x^1, y^1), (x^2,y^2),..,(x^m,y^m) \}$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

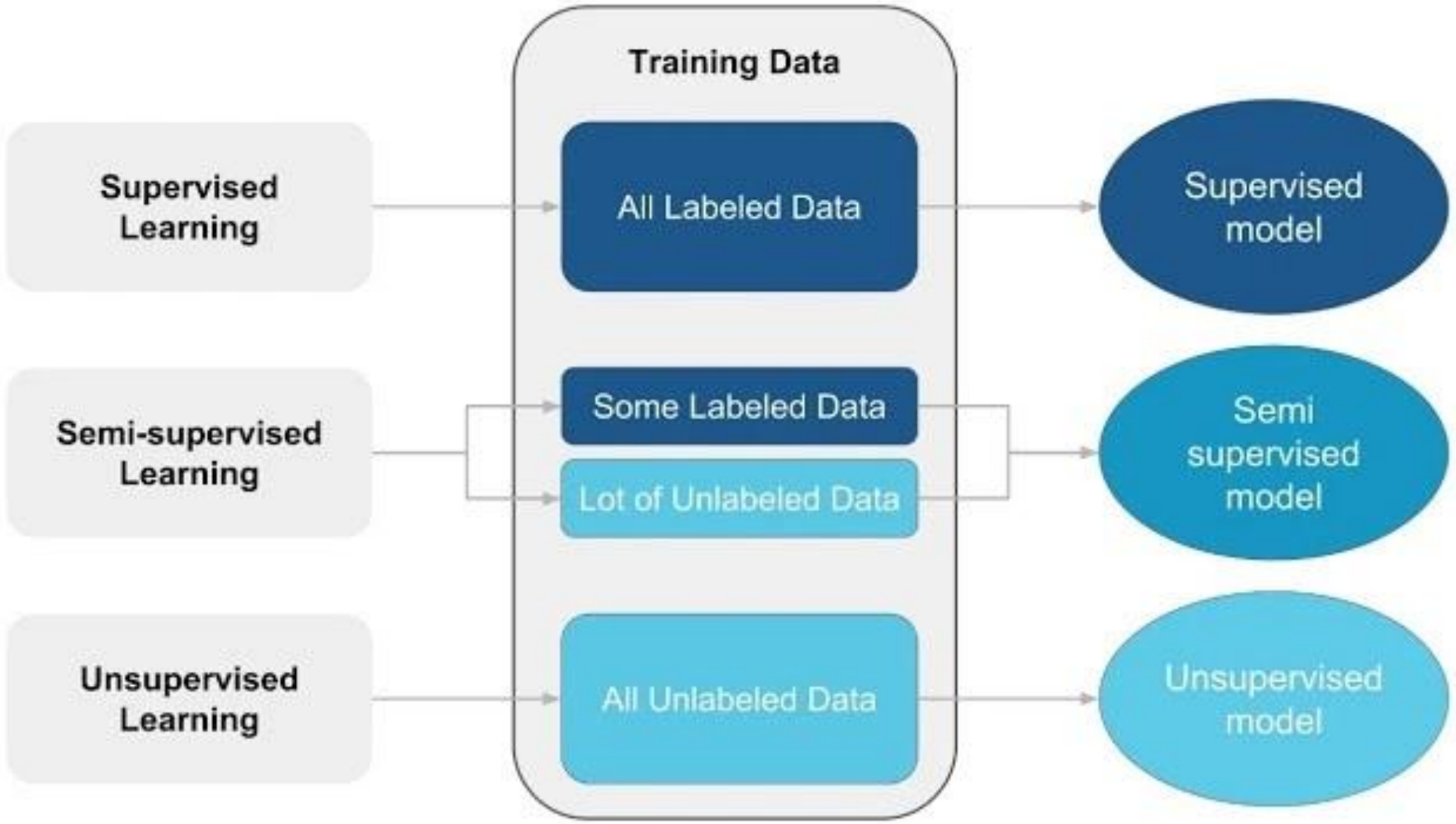
[See: Tom M. Mitchell, *Machine Learning,* McGraw-Hill, 1997]

Chỉ ra thuộc tính? Nhãn/lớp của tập dữ liệu thời tiết trong bảng trên

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

Phân loại học máy – **học bán giám sát**

Học bán giám sát: Học bán giám sát đối với trường hợp dữ liệu thu thập được có một phần nhỏ đã được gán nhãn, phần lớn còn lại chưa được gán nhãn trong quá trình học.

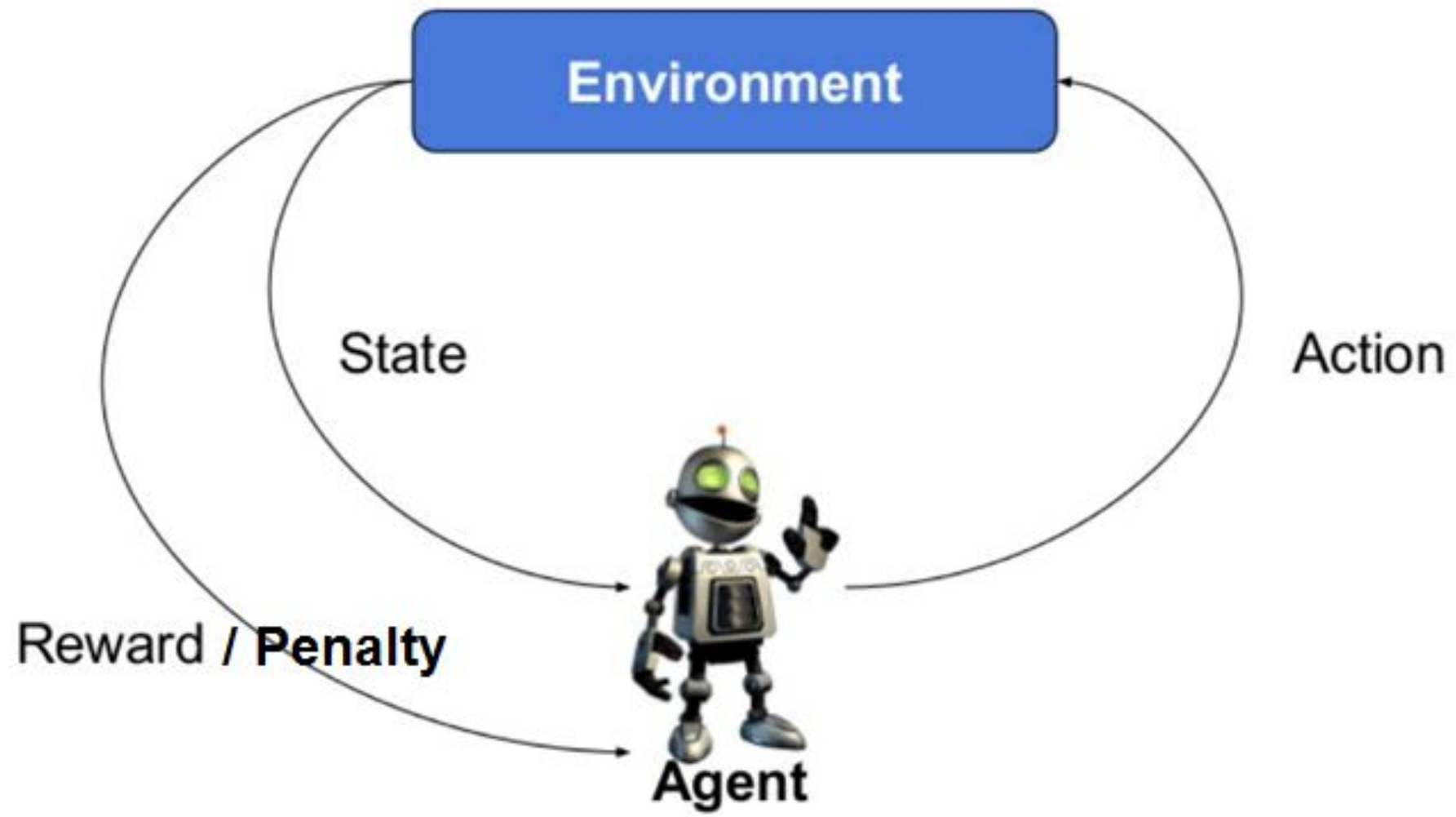

Training Data
Supervised Learning
All Labeled Data
Supervised model
Semi-supervised Learning
Some Labeled Data
Lot of Unlabeled Data
Semi supervised model
Unsupervised Learning
All Unlabeled Data
Unsupervised model

Phân loại học máy – **học tăng cường**

Học củng cố / học tăng cường: reinforcement learning:

- là một cách tiếp cận tập trung vào việc học để hoàn thành được mục tiêu bằng việc tương tác trực tiếp với môi trường.
- Đây là các bài toán giúp cho một hệ thống tự động xác định hành động dựa vào môi trường cụ thể để đạt được hiệu quả cao nhất.
- Bản chất của học tăng cường là trial-and-error, nghĩa là thử đi thử lại và rút ra kinh nghiệm sau mỗi lần thử như vậy. Đây là một nhánh học khá hấp dẫn trong máy học.

Typical RL scenario



Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/
- Kaggle: https://www.kaggle.com/datasets

b. Dựa vào thông tin: số ca nhiễm, số người tử vong, số lượt di chuyển của người dân trong thành phố, dân số của thành phố, người ta cần xác định mức độ lây nhiễm của dịch Covid-19 theo 3 mức: nguy cơ thấp, nguy cơ, nguy cơ cao. Theo anh/chị, chúng ta nên sử dụng giải thuật gì (clustering/classification/regression) để xác định mức độ lây nhiễm của dịch Covid-19? Anh/chị hãy giải thích cho lựa chọn của mình?

Bài tập

Cho tập dữ liệu hình bên, cần dự đoán giá trị y cho phần tử có x1 =2, x2=4, x3=0.7. Bài toán trên thuộc dạng nào?

x1	x2	x3	y
3	4	0	A
4	6	0.5	C
3.5	5	1.5	A
2	7	2.5	B
6	3	3	A

Bài tập

Một tập dữ liệu được cho trong bảng sau (từ A đến H) dùng để phân loại nấm ăn được hay không (giá trị Edible lần lượt là 1 hoặc 0) dựa vào các thuộc tính sau: NotHeavy, Smelly, Spotted, Smooth

Cần xây dựng mô hình dự đoán xem một loại nấm có ăn được hay ko?
Bài toán trên thuộc dạng nào?

Example	*NotHeavy*	*Smelly*	*Spotted*	*Smooth*	*Edible*
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

Bài tập

Cho tập dữ liệu hình bên, xác định Y và giải thuật cần sử dụng

STT	Rented Bike Count	Hour	Temp	Rain
1	**1241**	23	28.5	No
2	**1003**	13	31.1	No
3	**943**	23	15.8	No
4	**353**	16	3.5	Yes
5	**78**	4	-6	Yes


The End