HỌC MÁY
(MACHINE LEARNING)

Nguyễn Thị Hải Bình

Email: nth.binh@hutech.edu.vn

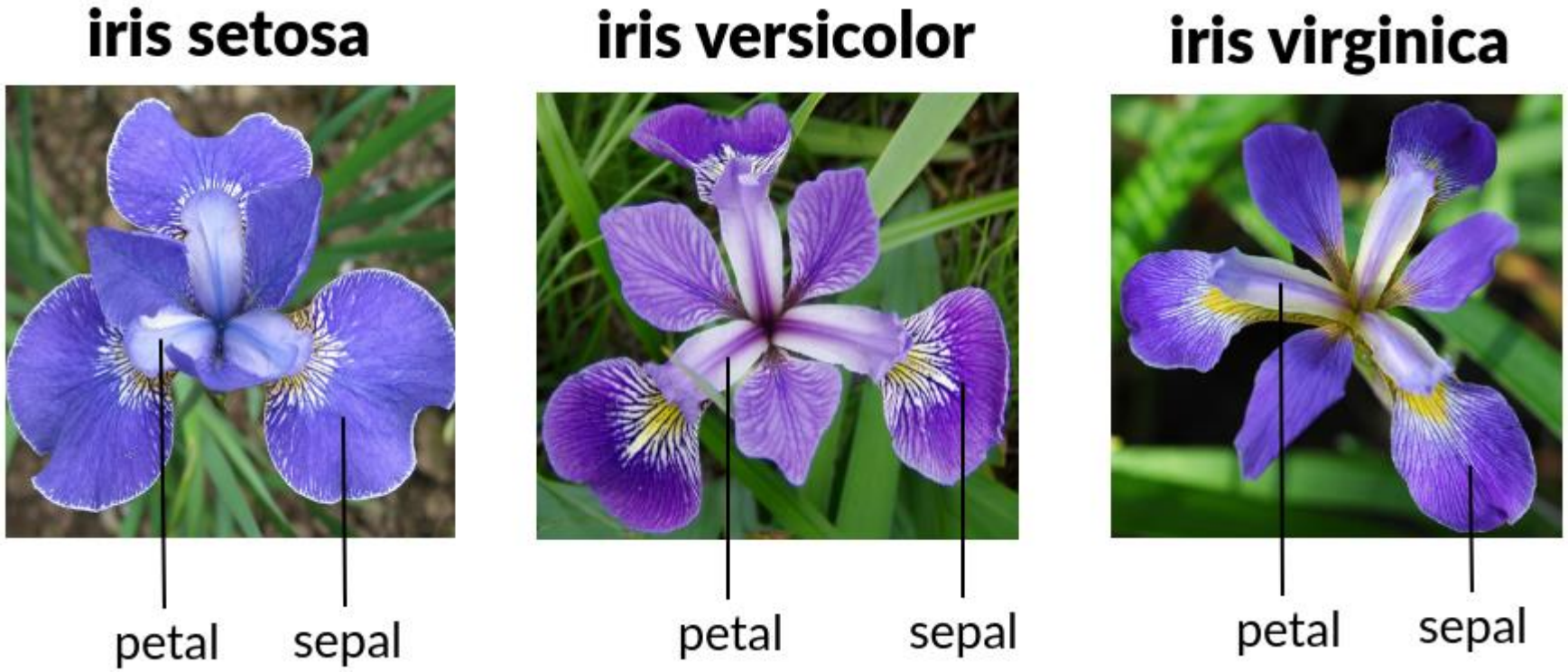
Nội dung chính của bài học

1. Học máy là gì?
2. Quy trình học máy
3. Ứng dụng của học máy
4. Phân loại thuật toán học máy
5. Bài toán phân lớp
6. Bài toán gom cụm dữ liệu

HỌC MÁY LÀ GÌ?

Bài toán phân loại hoa iris

- Phát biểu bài toán:
  - Biết chiều dài và chiều rộng của cánh hoa và đài hoa.
  - Hãy cho biết đó là loài hoa iris nào.

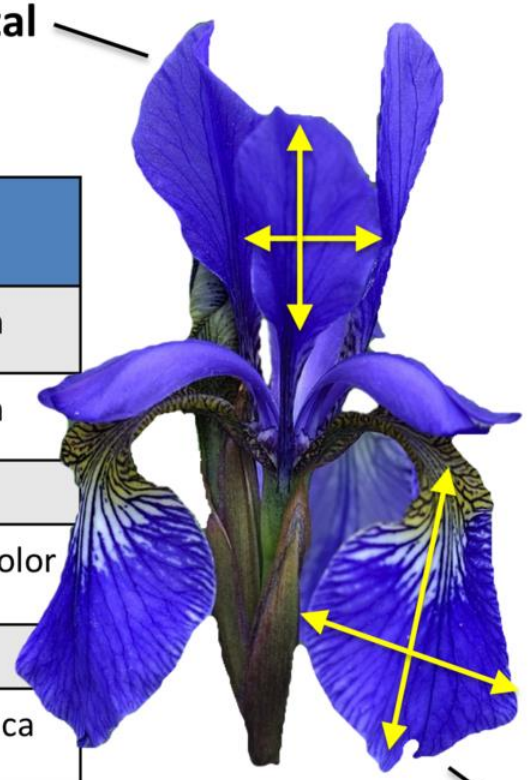


Bài toán phân loại hoa iris

- Để giải bài toán phân loại hoa iris, các nhà nghiên cứu thực hiện đo lường và ghi lại chiều dài và chiều rộng của cánh hoa và đài hoa của 150 bông hoa iris.
- Mỗi bông hoa iris thuộc một trong ba loài setosa, versicolor, hoặc virginica.

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Bài toán phân loại hoa iris

Không sử dụng học máy

- Các nhà nghiên cứu sử dụng bộ dữ liệu gồm 150 quan sát và phân tích.
- Sau khi quan sát giá trị lớn nhất, nhỏ nhất của độ dài cánh hoa và đài hoa của mỗi loại hoa, các nhà nghiên cứu đưa ra quy tắc để phân loại hoa iris như sau:

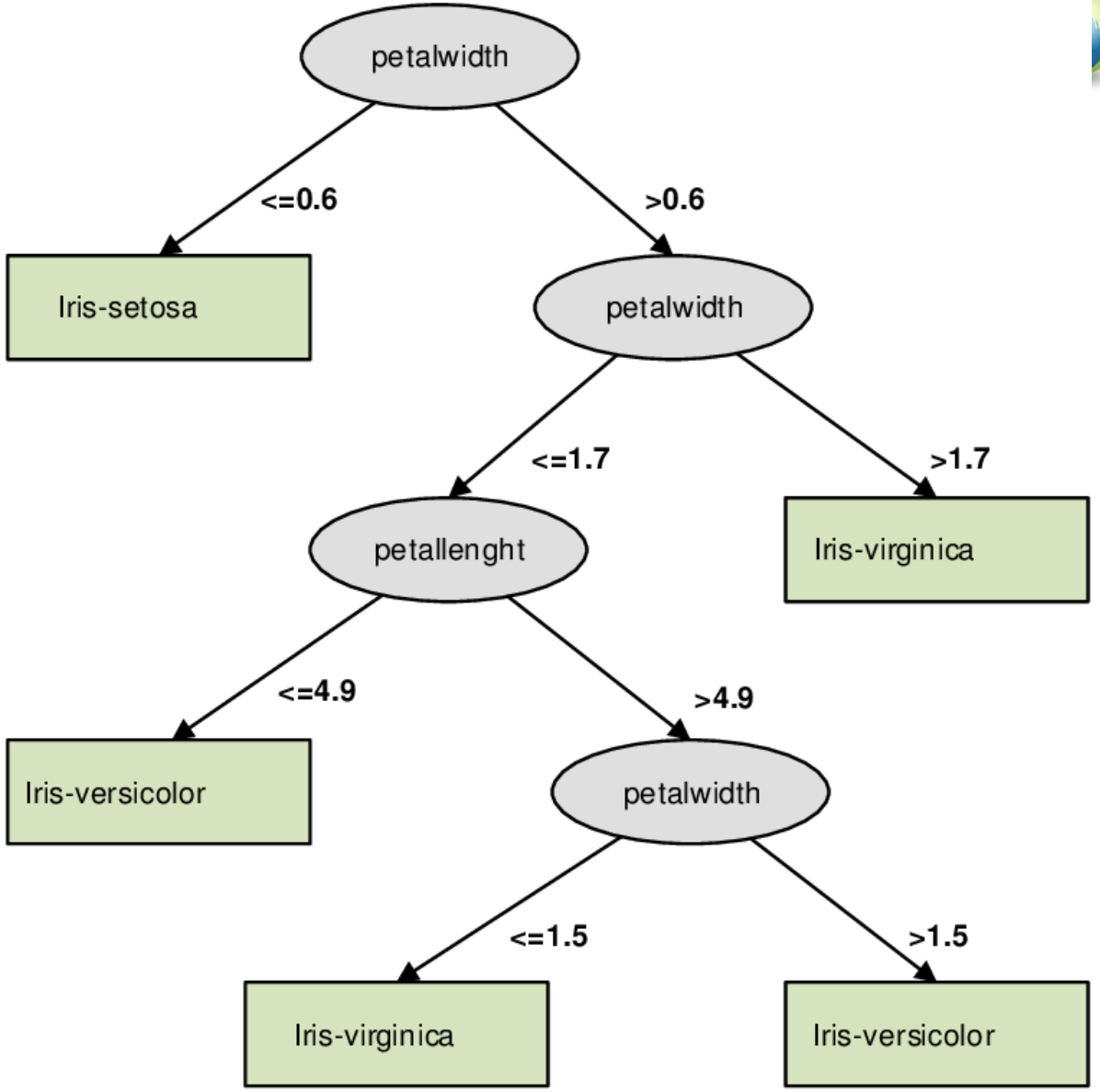
```
R1: If Petal.Length < 2.5cm then species = Setosa

R2: If (Petal.Length > 5.1) OR (2.5 < Petal.Length < 5.1 and Petal.Width > 1.8)
    then species = Virginica

R3: If (2.5 < Petal.Length < 5.1 and Petal.Width < 1.8) then species = Versicolor
```

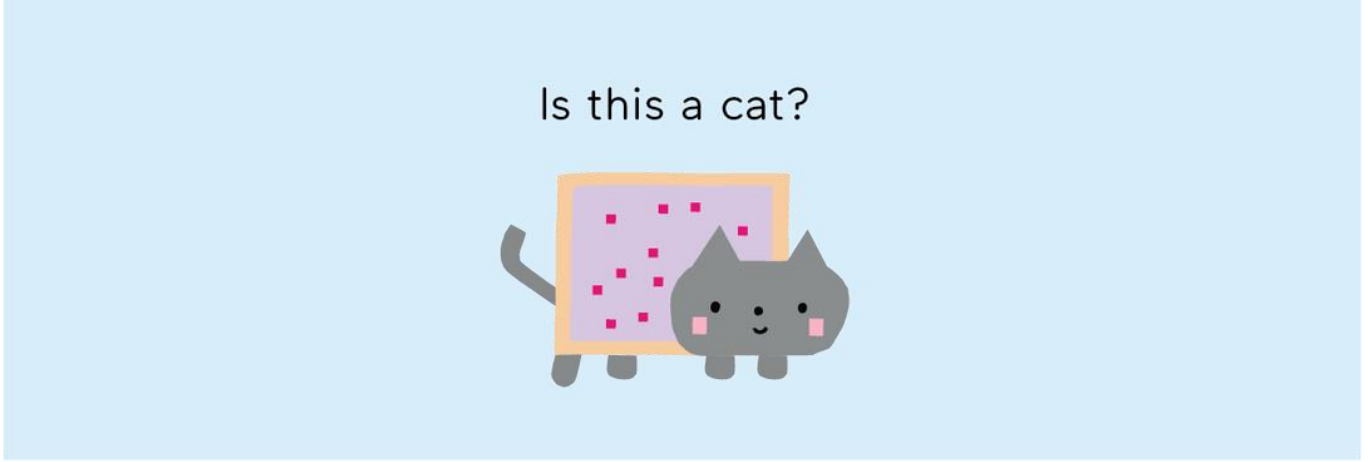
- Ba quy tắc R1, R2, R3 sẽ được sử dụng để máy tính (AI) dự đoán tên loài của hoa iris.

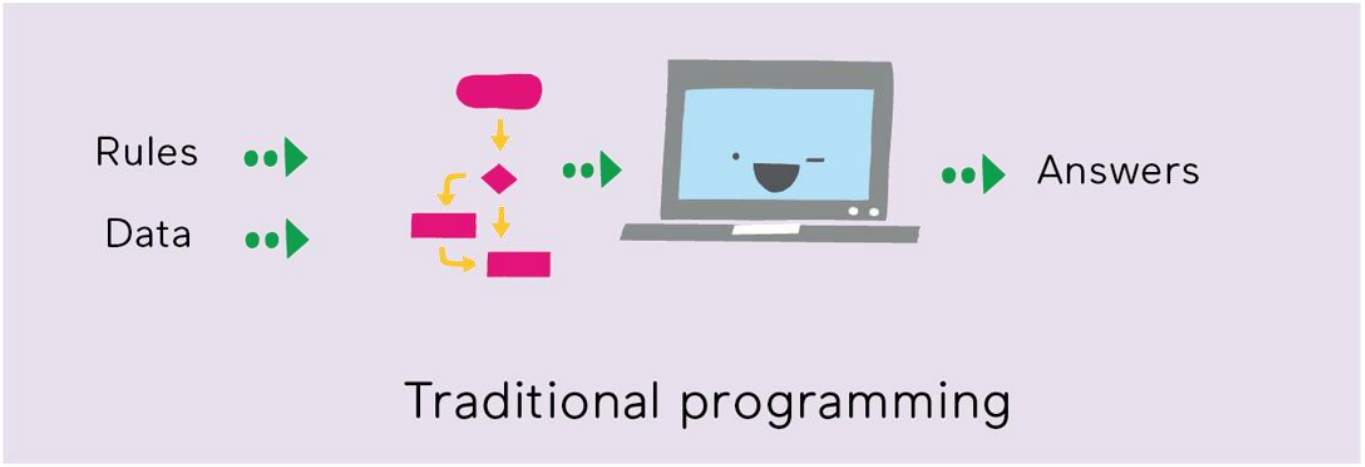
Bài toán phân loại hoa iris

Sử dụng học máy

- 150 quan sát được đưa vào một thuật toán (gọi là thuật toán học máy).
- Thuật toán học máy sẽ sinh ra một tập các quy tắc (hay một mô hình) dùng để dự đoán tên loài của hoa iris.







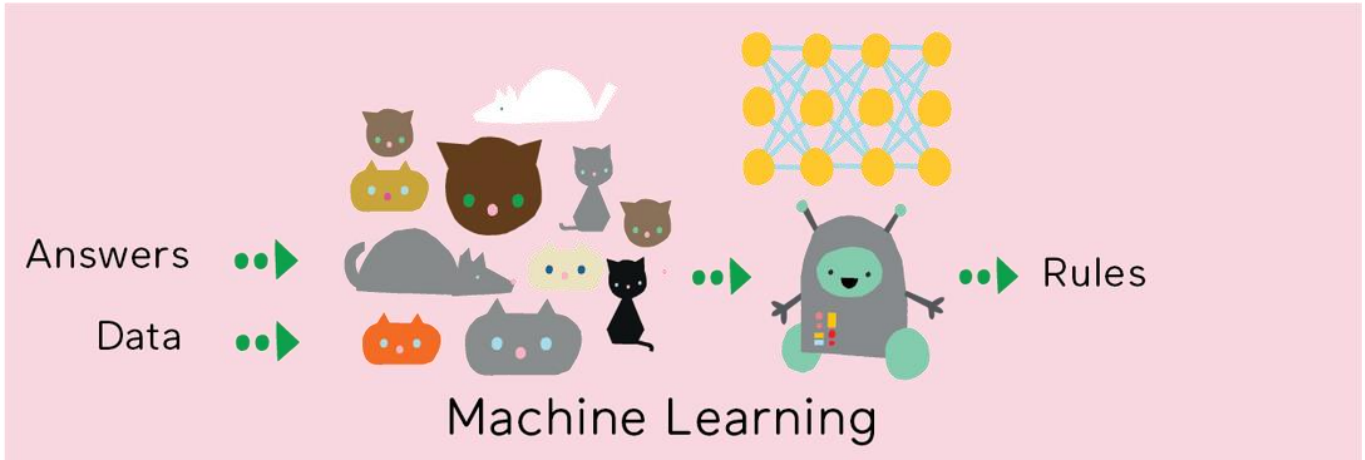


Figure credit: Linda Liukas

Học máy là gì?

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

- Herbert Simon

Definition by Tom Mitchell (1998):

Machine learning is the study of algorithms that:

- improve their performance P
- at some task T
- with experience E

A well-defined learning task is given by $\langle P, T, E \rangle$

Definition of machine learning

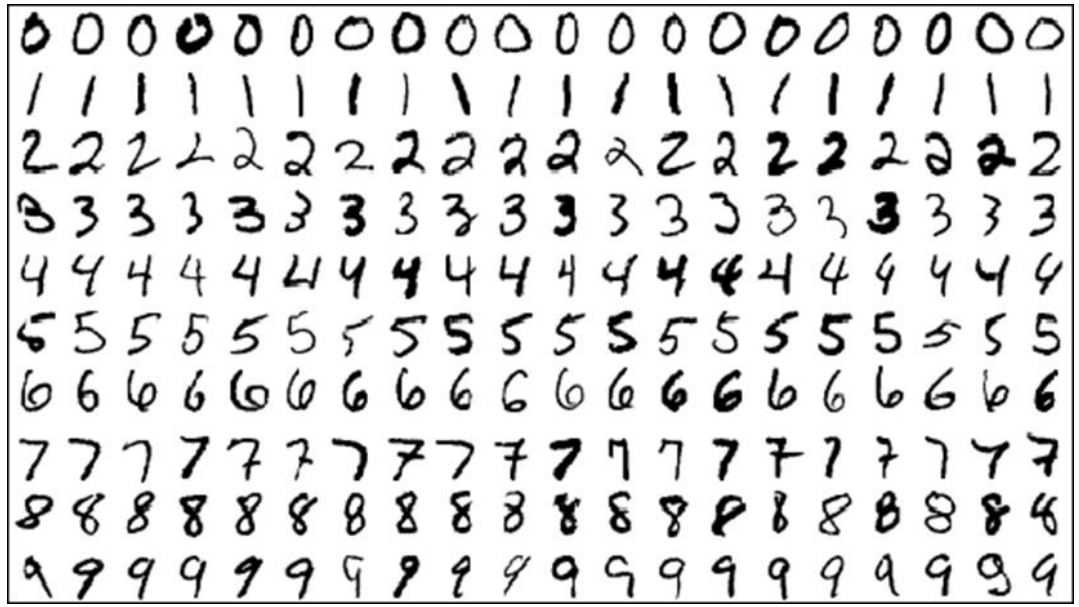
Machine learning is the study of algorithms that improve their performance P at some task T with experience E.

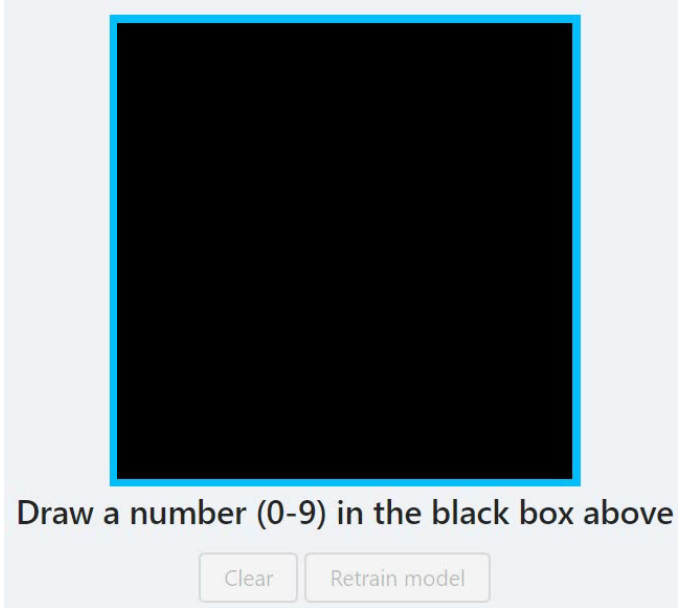
Handwritten digit recognition

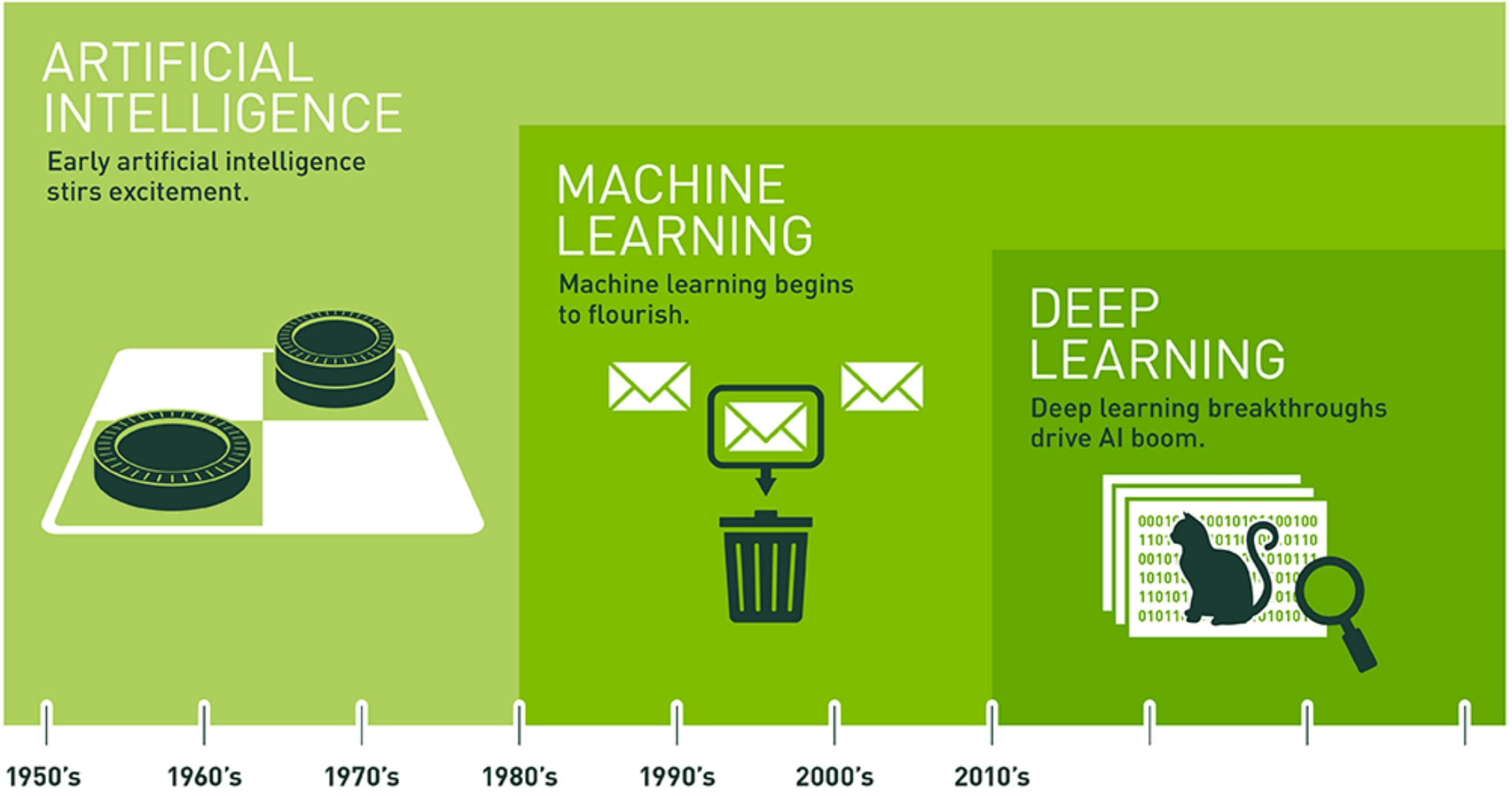
T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of handwritten digits







Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

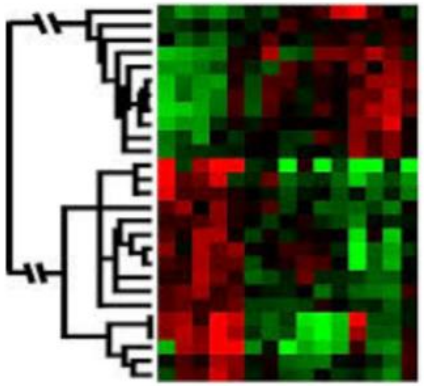
Figure from https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

When Do We Use Machine Learning?

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)



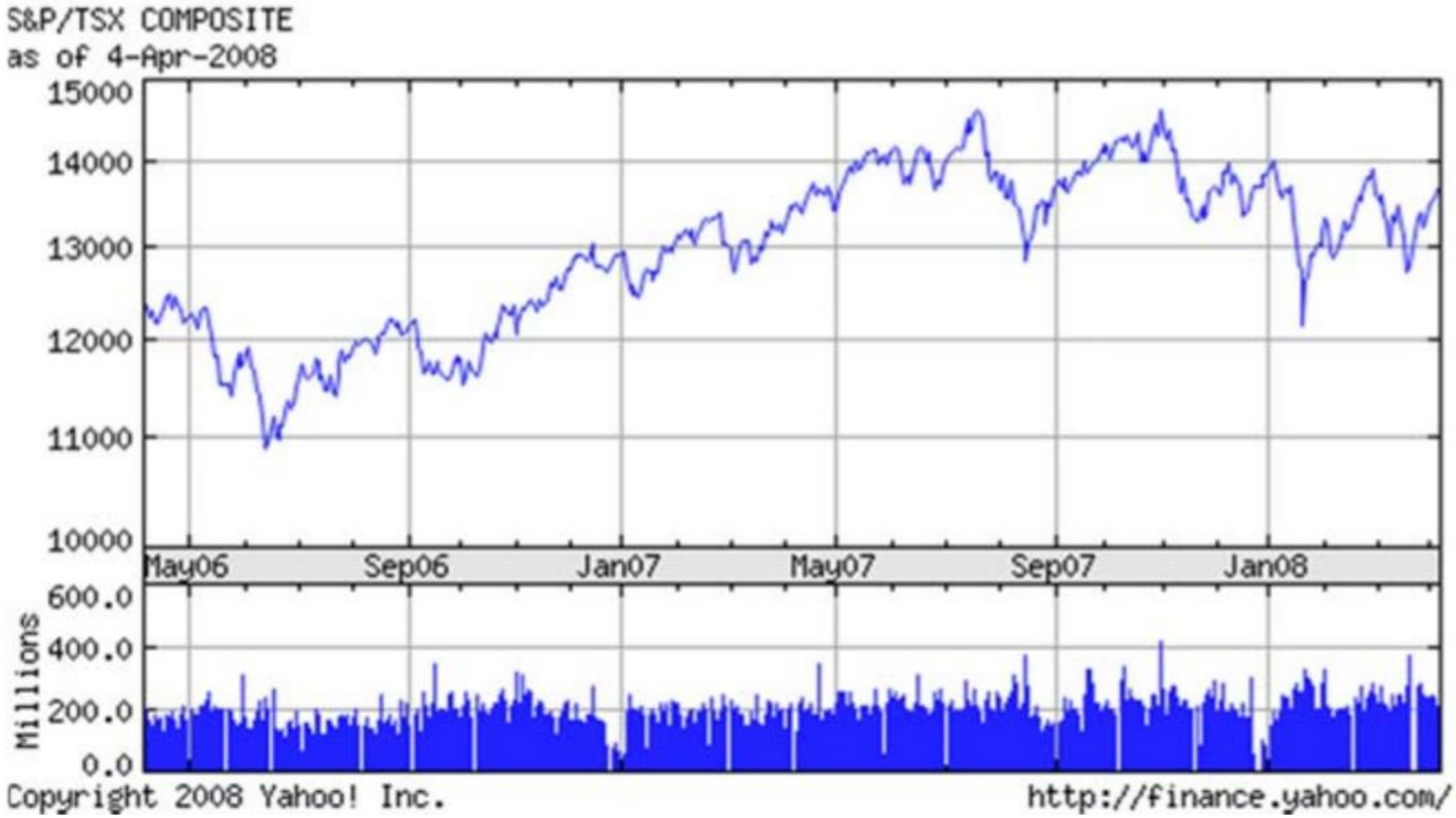
Learning isn't always useful:

- There is no need to "learn" to calculate payroll

Examples of tasks that are best solved by using a learning algorithm

- Recognizing patterns:
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns:
  - Generating images or motion sequences
- Recognizing anomalies:
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction:
  - Future stock prices or currency exchange rates

Stock price prediction



Machine translation

Web Images Maps News Shopping Mail more ▾ Help

Google Translate BETA

Home | Text and Web | Translated Search | Dictionary | Tools

Translate text or webpage

Enter text or a webpage URL.

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

French > English swap Translate

Translation: French » English

Under the new proposals, what is the cost of collection of fees?

Suggest a better translation

Google Home - About Google Translate

Recommender systems

People who bought Hastie …

Frequently Bought Together

Customers buy this book with Pattern Recognition and Machine Learning (Information Science and Statistics) (Information Science and Statistics) by Christopher M. Bishop

Price For Both: £104.95

Add both to Basket

Customers Who Bought This Item Also Bought

Page 1

Pattern Recognition and Machine Learning (Infor...
by Christopher M. Bishop
★★★★☆ (4) £48.96

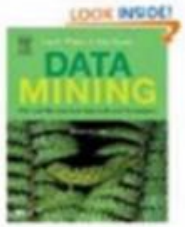
Show related items

MACHINE LEARNING (Mcgraw-Hill International Edit) by Thom M. Mitchell
★★★★★ (3) £42.74

Show related items

Pattern Classification, Second Edition: 1 (A Wi...
by Richard O. Duda
★★★★★ (1) £78.38

Show related items

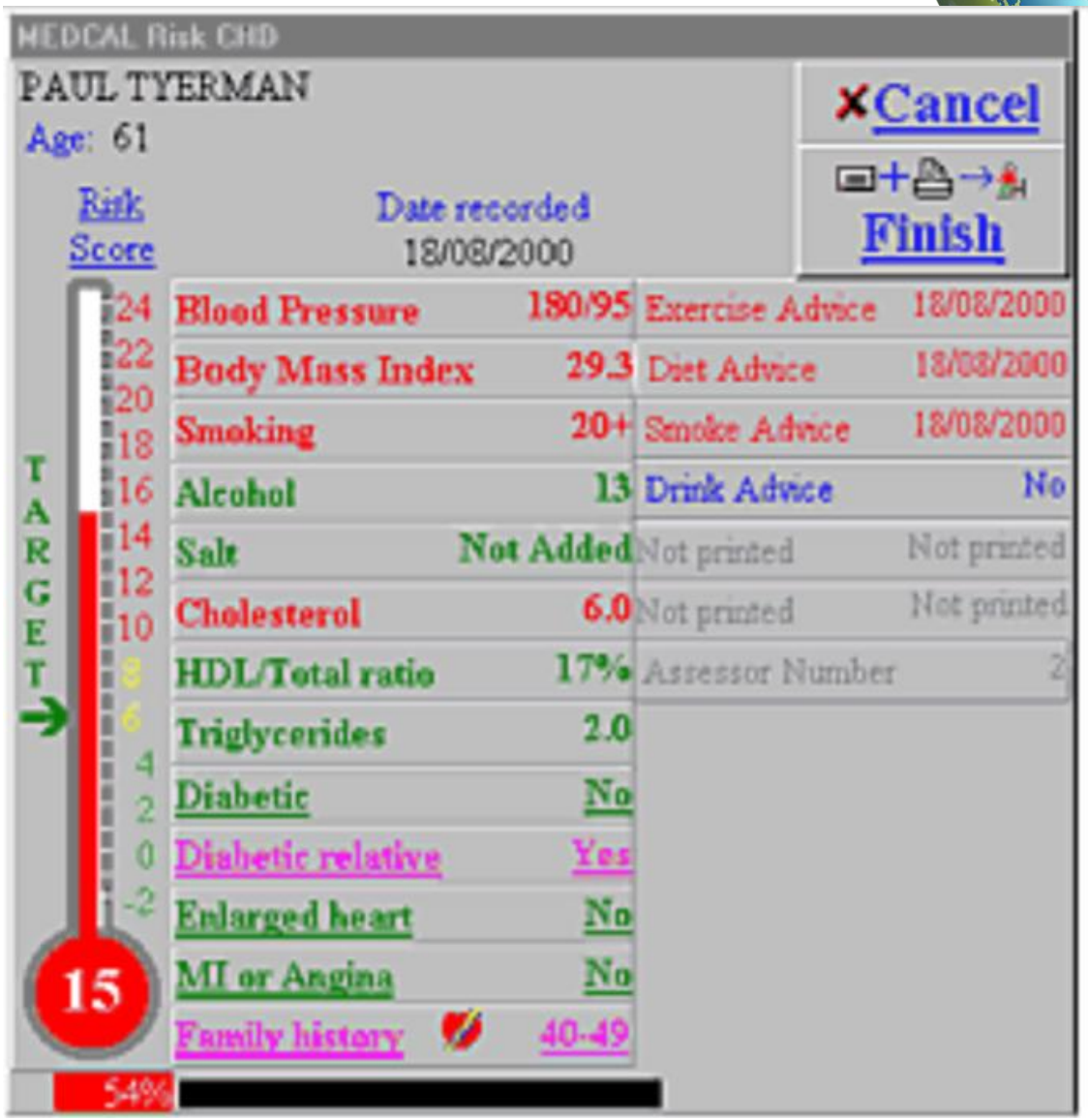


Data Mining: Practical Machine Learning Tools a...
by Ian H. Witten
★★★★★ (1) £37.04

Show related items

Medical Diagnosis

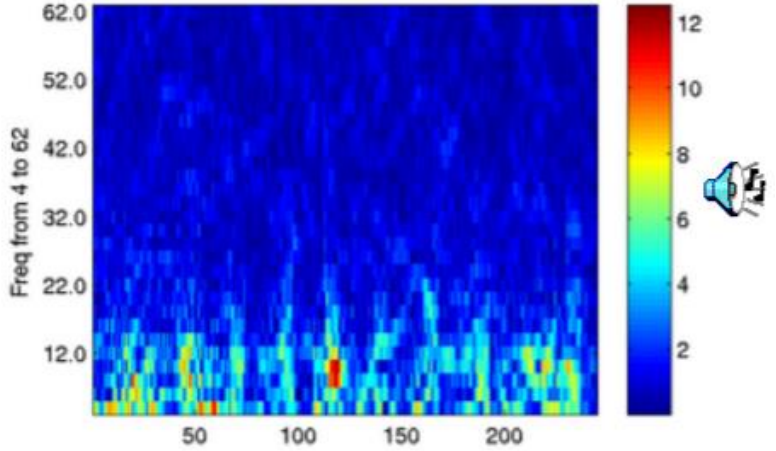
- Inputs: relevant info about patient, symptoms, test results, etc.
- Output: Expected illness or risk factors



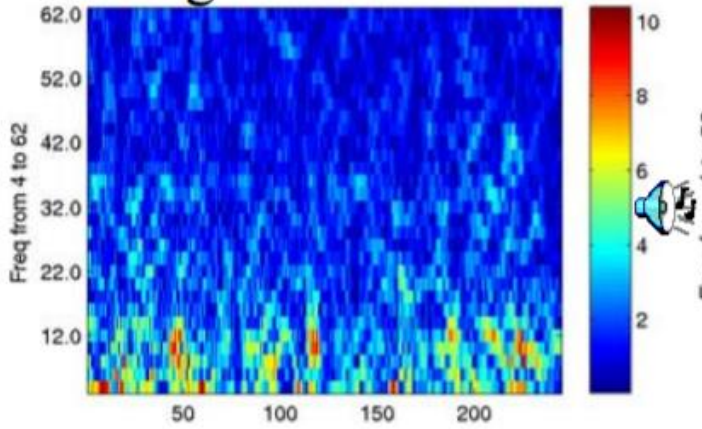
Interpreting Brainwaves

EEG electrodes reading brain waves:

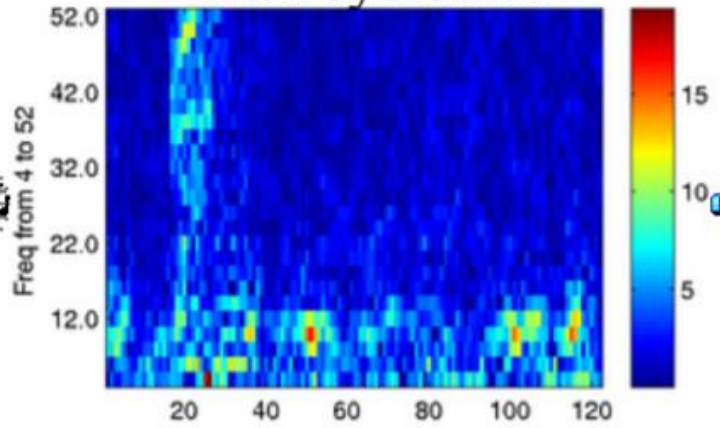
- Rotation task, left brain



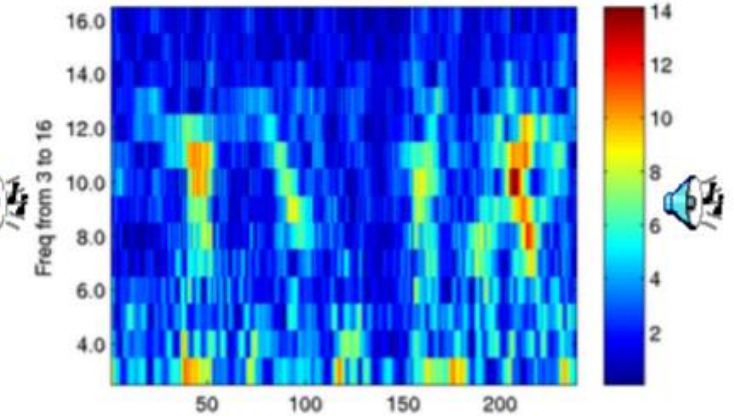
- Rotation task, right brain



- Resting task, with eye blink



- Counting task

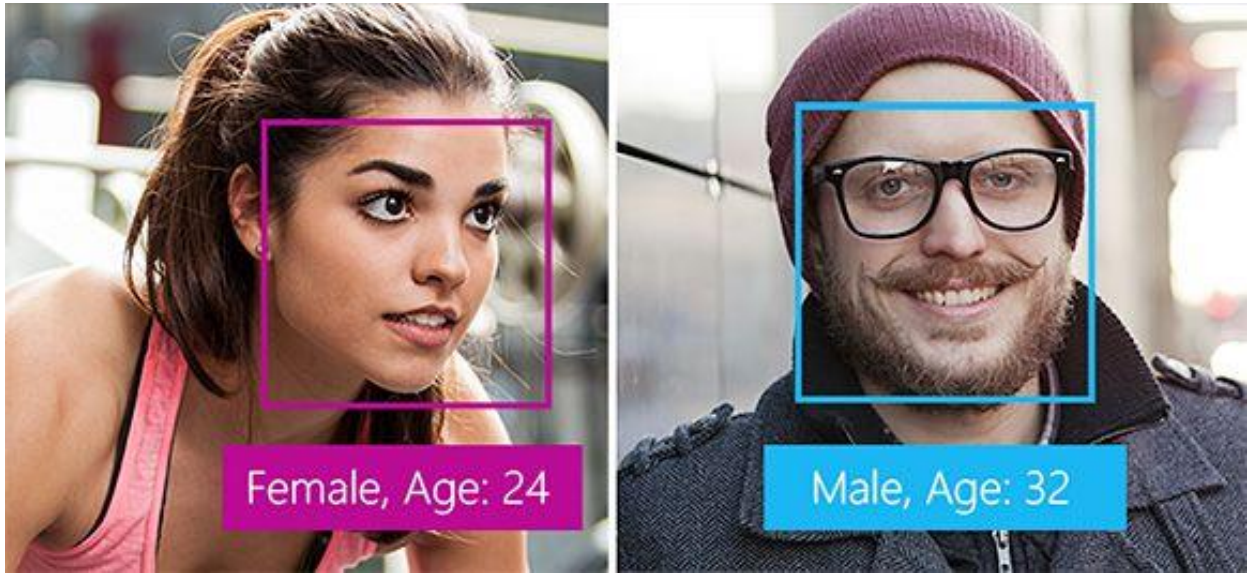


Speech Recognition

Slide credit: Li Deng, MS Research

Azure Face API from Microsoft

- The Face API can detect human faces in an image and return the rectangle coordinates of their locations.



Azure Face API from Microsoft

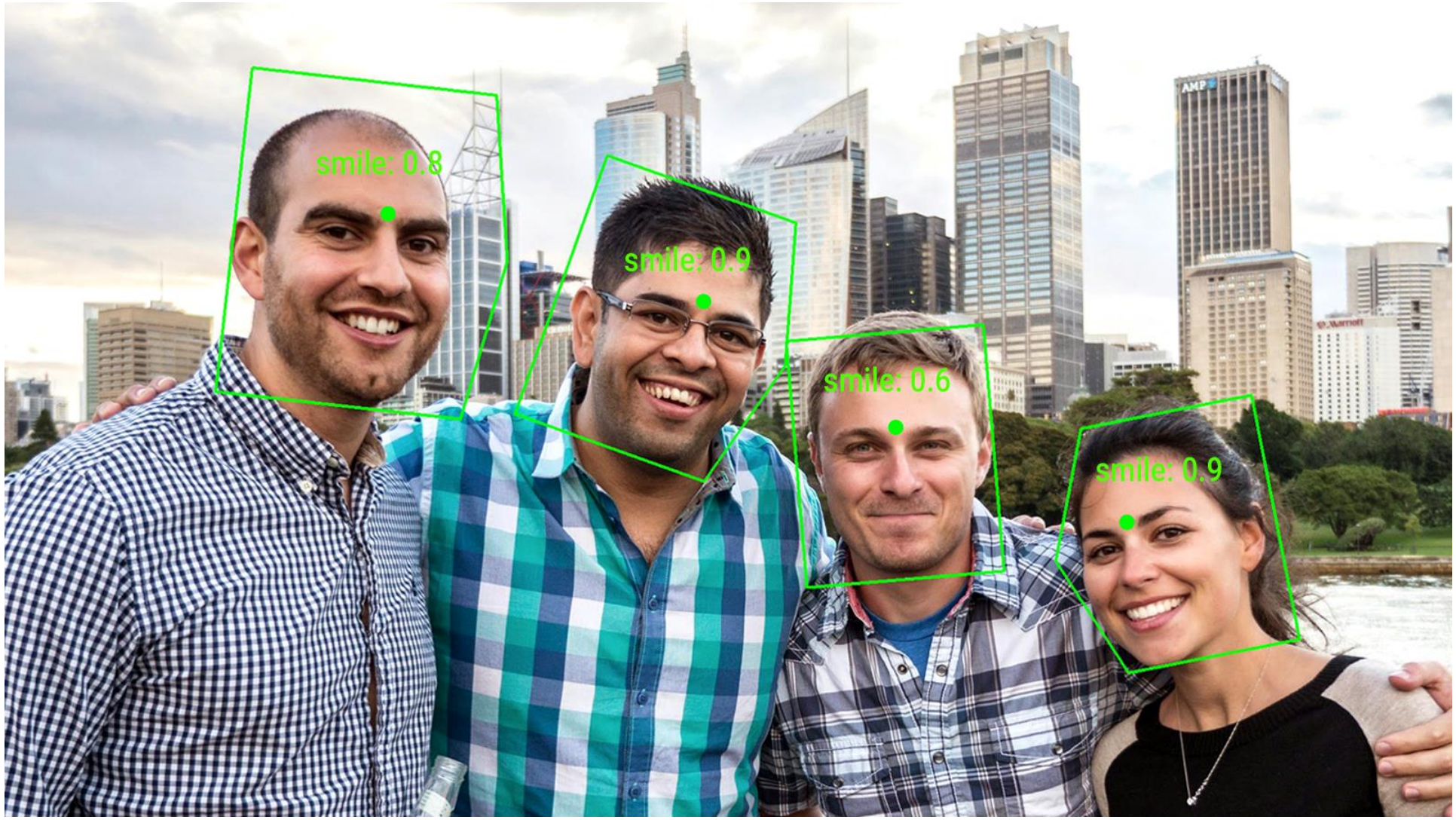
Target face

Find similar faces

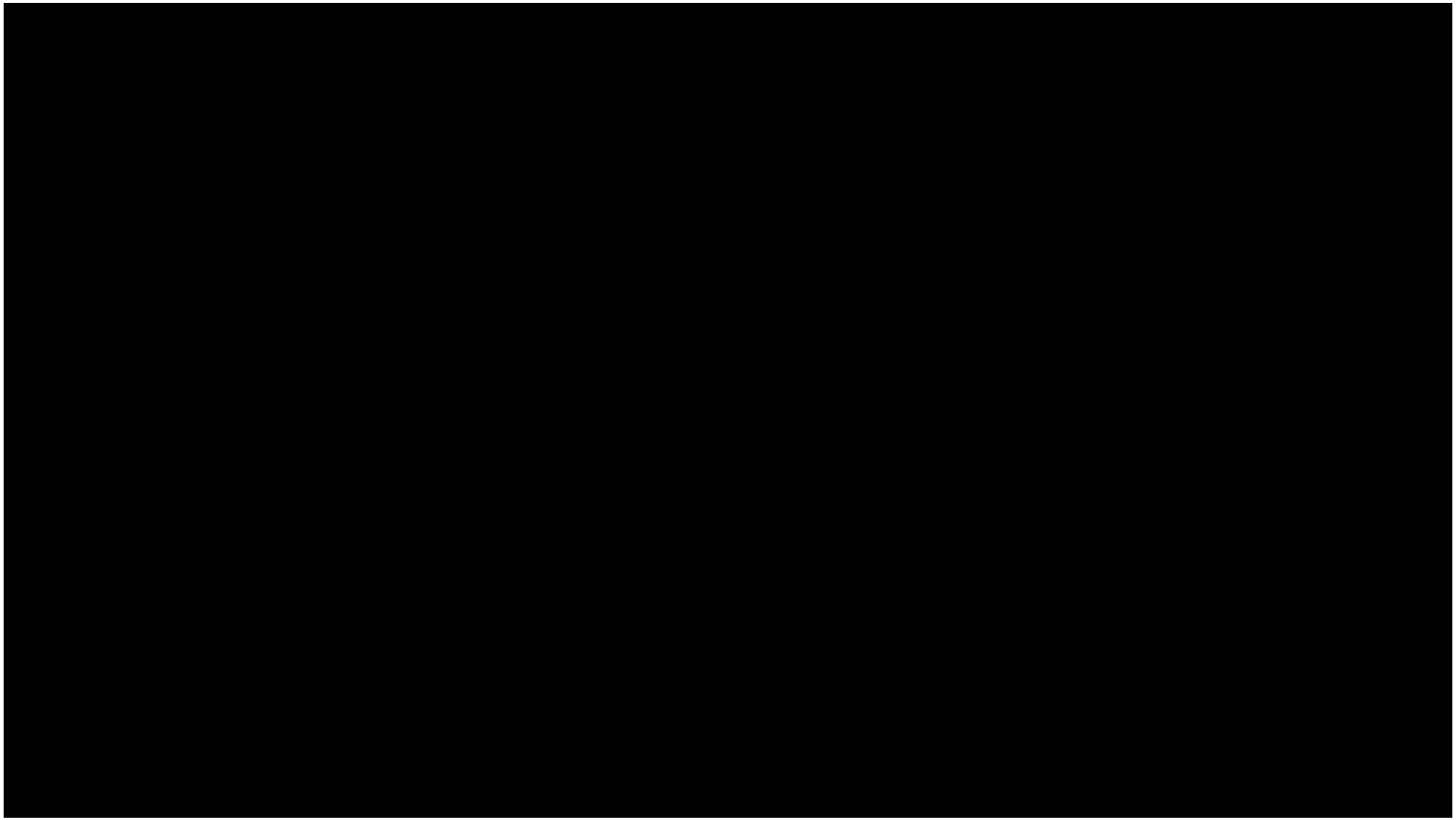
The Find Similar API takes a target face, and a set of candidate faces and finds a smaller set of faces that look most similar to the target face.

Candidate faces

Face API from Google



Autonomous Cars



Deep Learning in the Headlines

BUSINESS NEWS

MIT Technology Review

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to acquire DeepMind Technologies, a startup based in

BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. "DeepMind is bona fide in terms of its research capabilities and depth," says Peter Lee, who heads Microsoft Research.

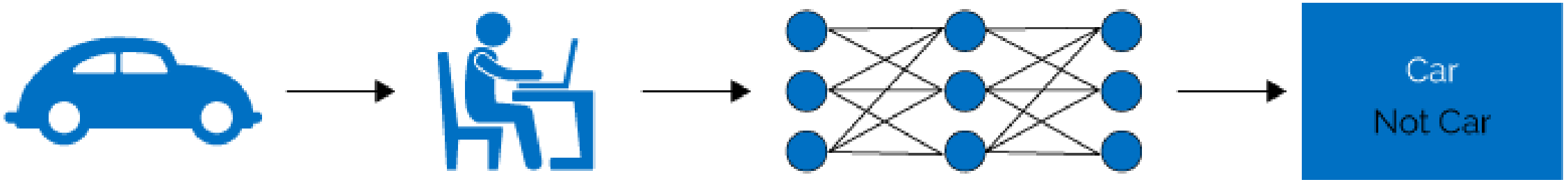
According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. "We would have more if the talent was there to

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIG

INNOVATION INSIGHTS | community content | featured

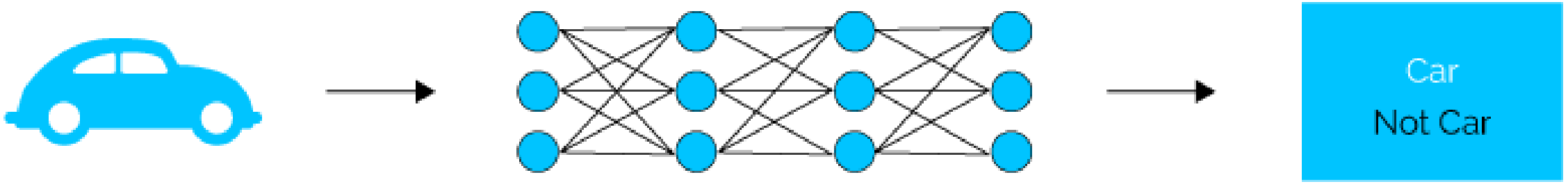
Deep Learning's Role in the Age of Robots

Machine Learning

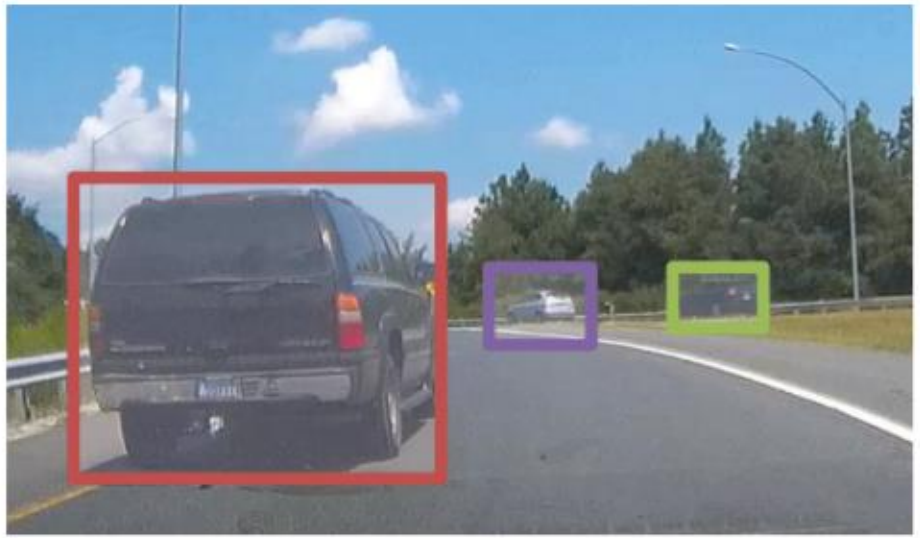


Input → Feature extraction → Classification → Output

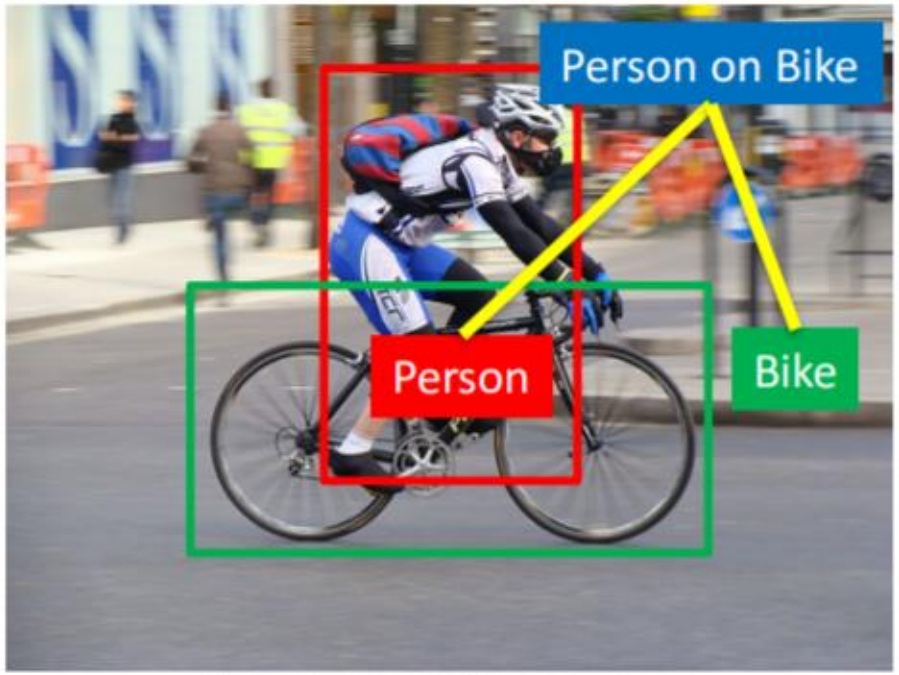
Deep Learning



Input → Feature extraction + Classification → Output



- Object detection
- Action classification
- Image captioning
- …



Types of Learning

Supervised (inductive) learning

- Given: training data + desired outputs (labels)

Unsupervised learning

- Given: training data (without desired outputs)

Semi-supervised learning

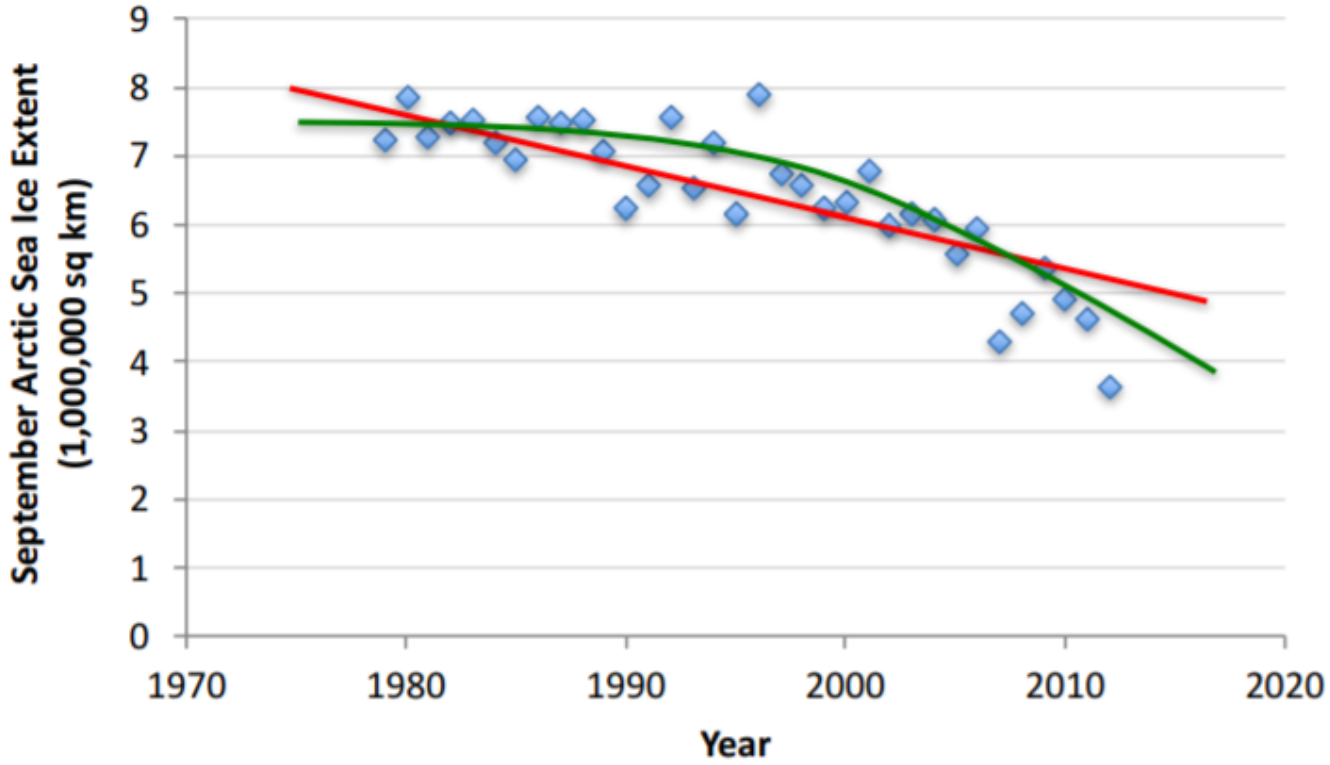
- Given: training data + few desired outputs

Reinforcement learning

- Rewards from sequence of actions

Supervised Learning: Regression

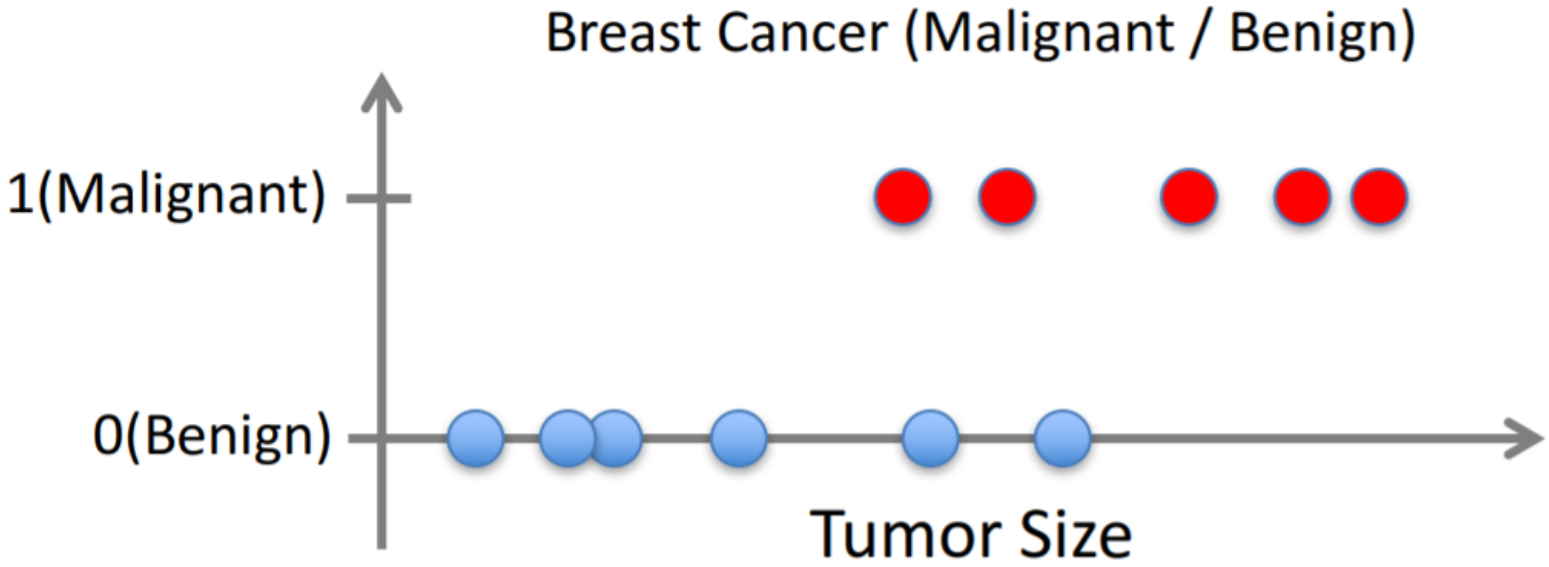
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Slide credit: Eric Eaton

Supervised Learning: Classification

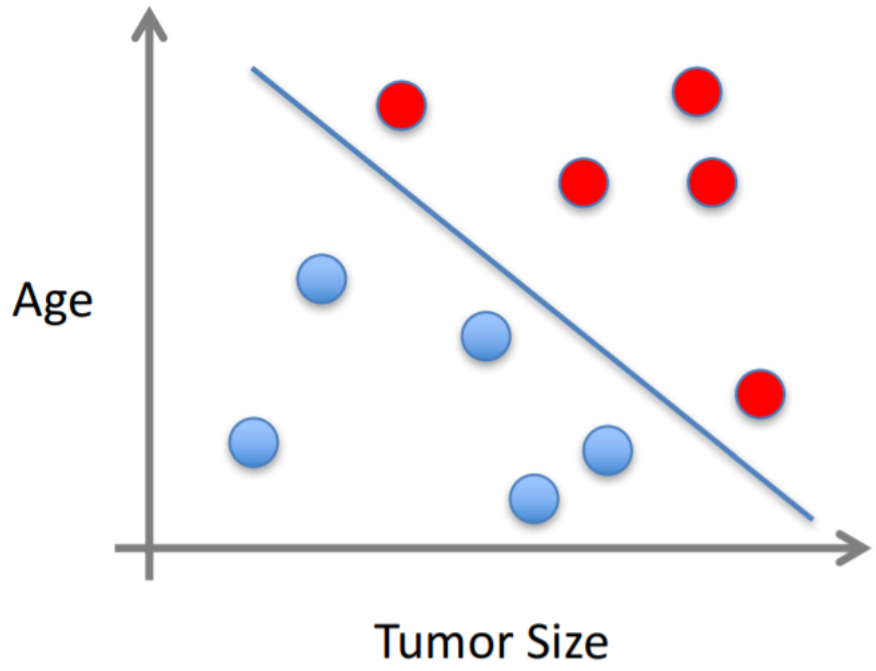
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Slide credit: Eric Eaton

Supervised Learning

- x can be multi-dimensional
  - Each dimension corresponds to an attribute



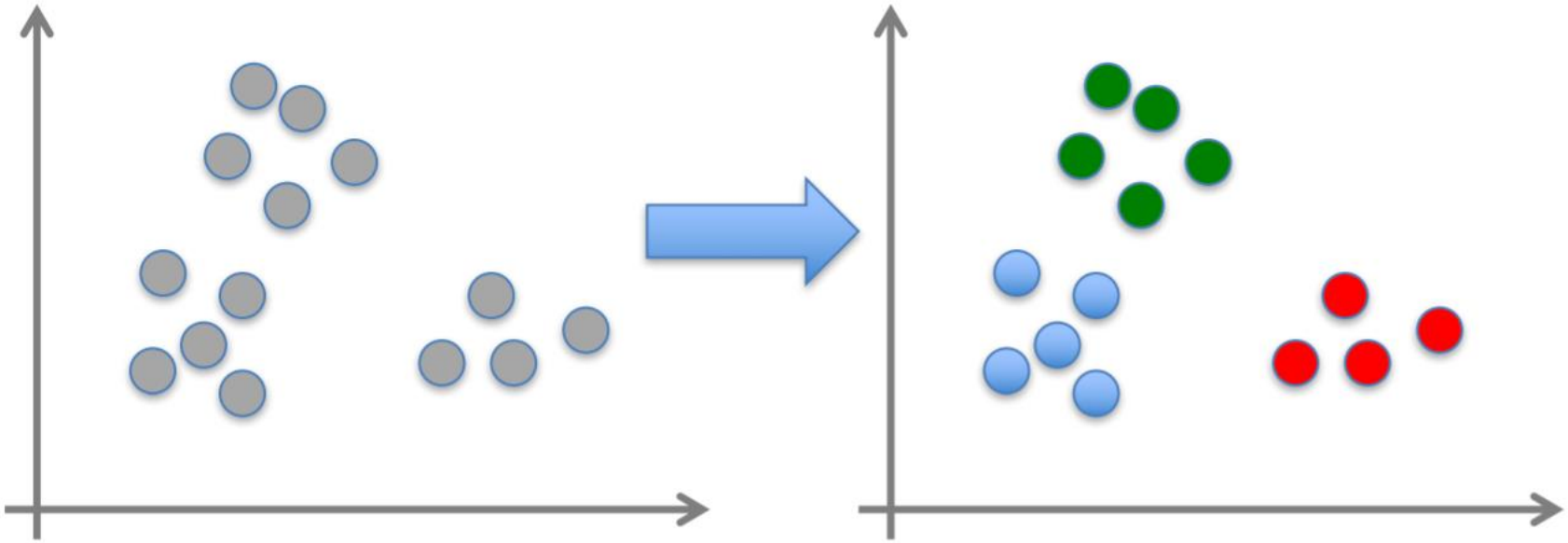
- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

...

Slide credit: Eric Eaton

Unsupervised Learning

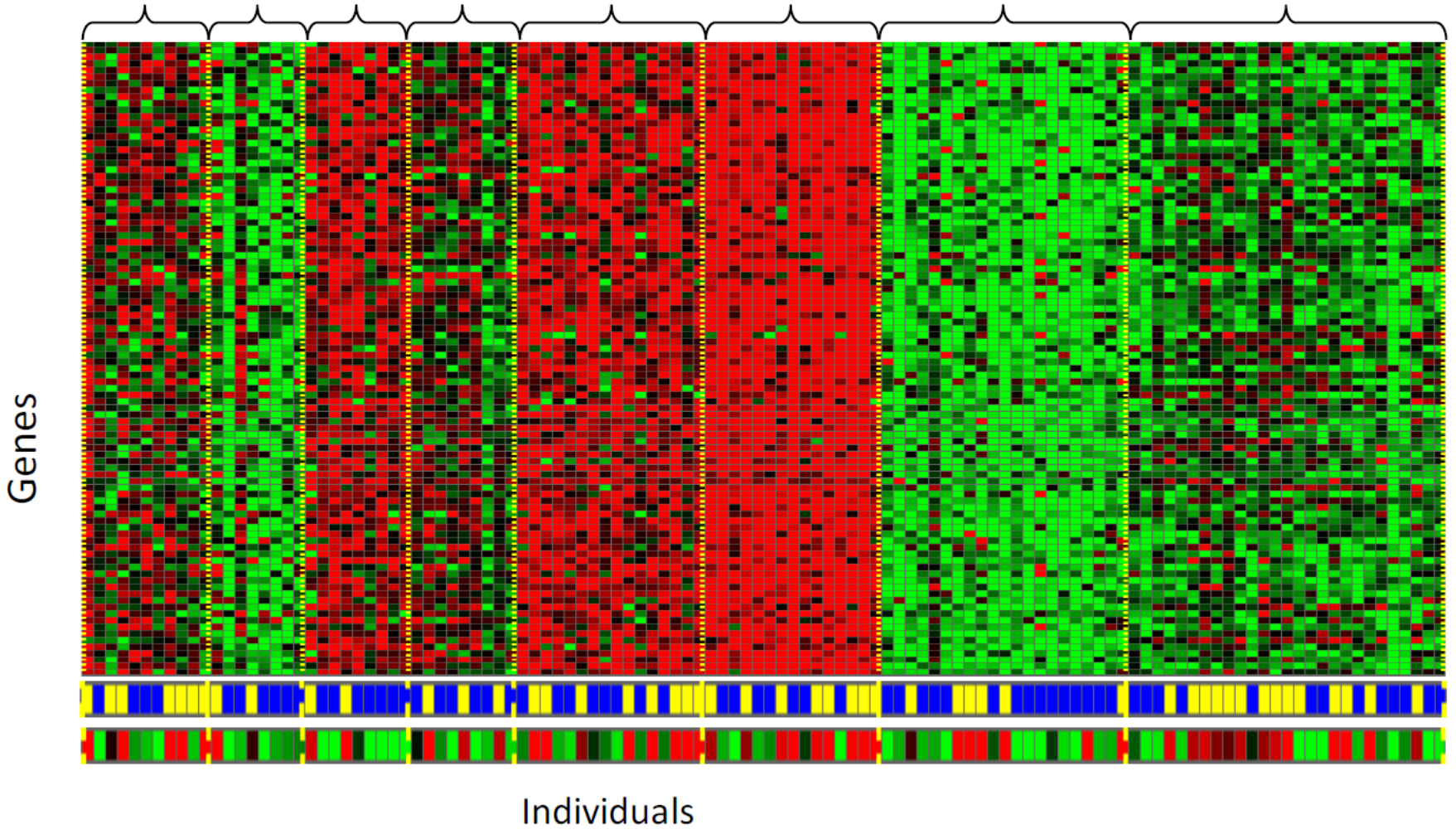
- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering



Slide credit: Eric Eaton

Unsupervised Learning

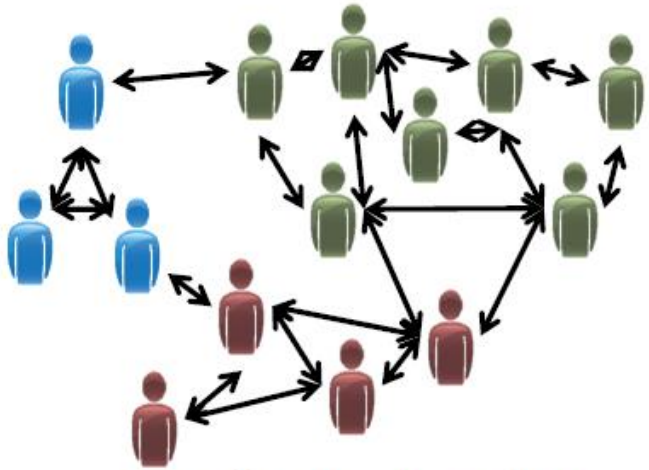
Genomics application: group individuals by genetic similarity



Slide credit: Eric Eaton

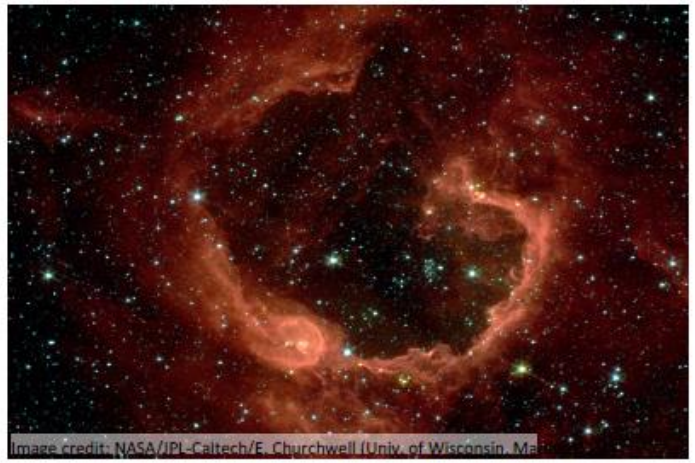
Unsupervised Learning

Organize computing clusters



Social network analysis

Market segmentation

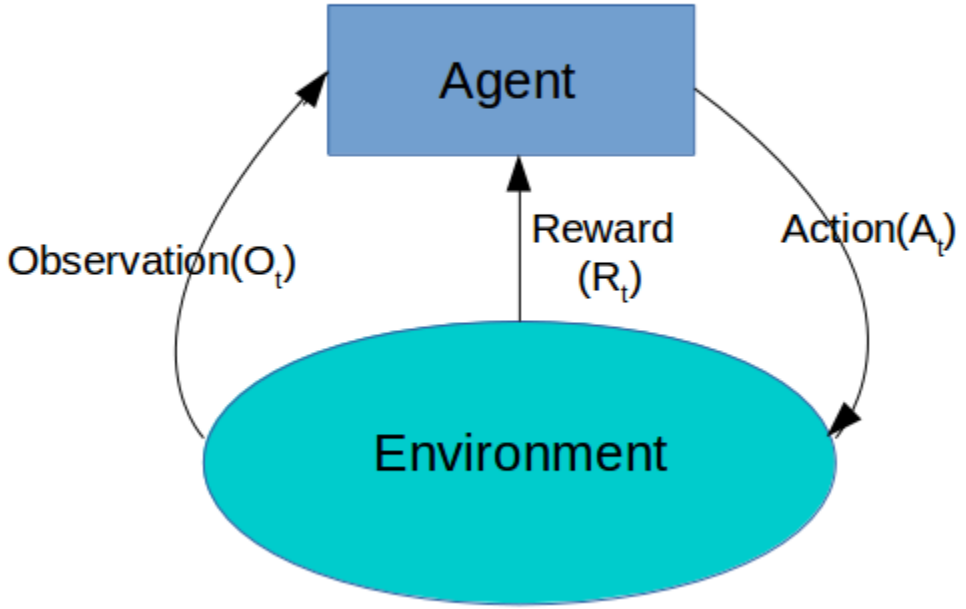


Astronomical data analysis

Slide credit: Eric Eaton

Reinforcement Learning

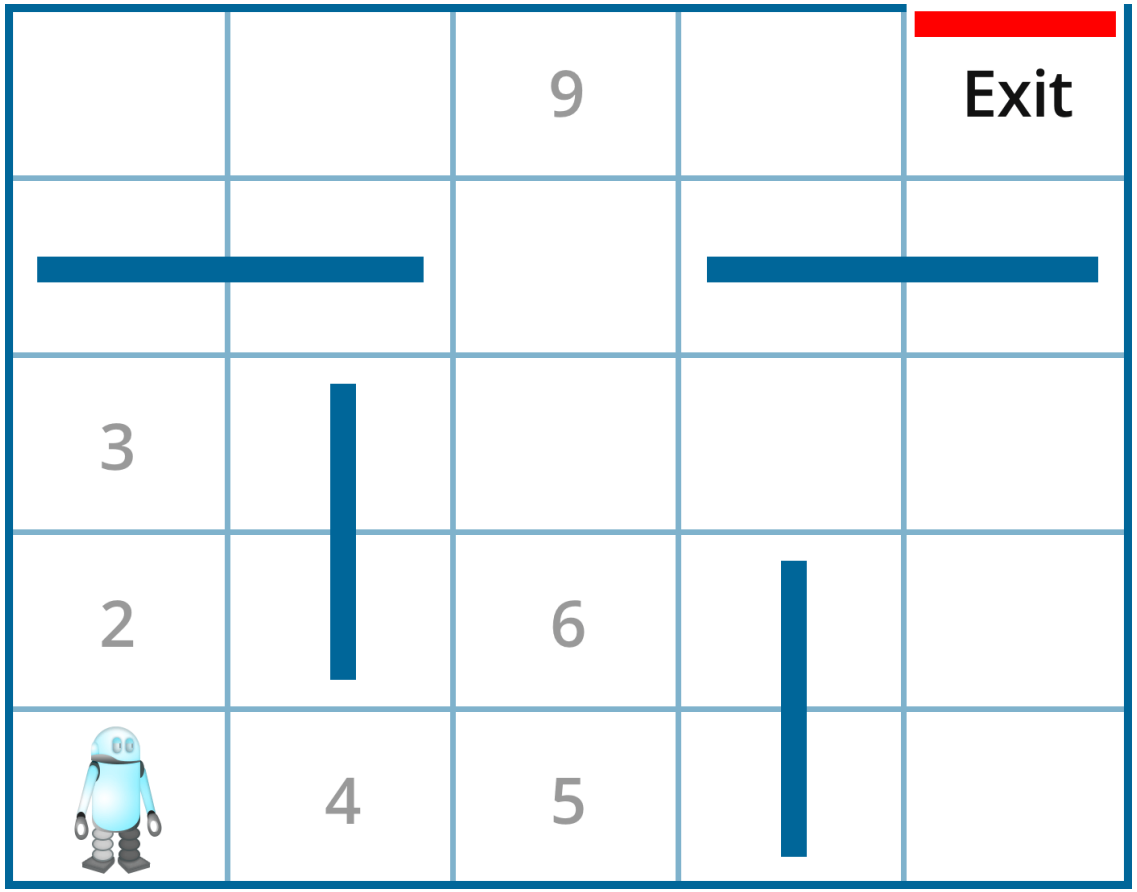
- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from state → actions that tells you what to do in a given state



Reinforcement Learning

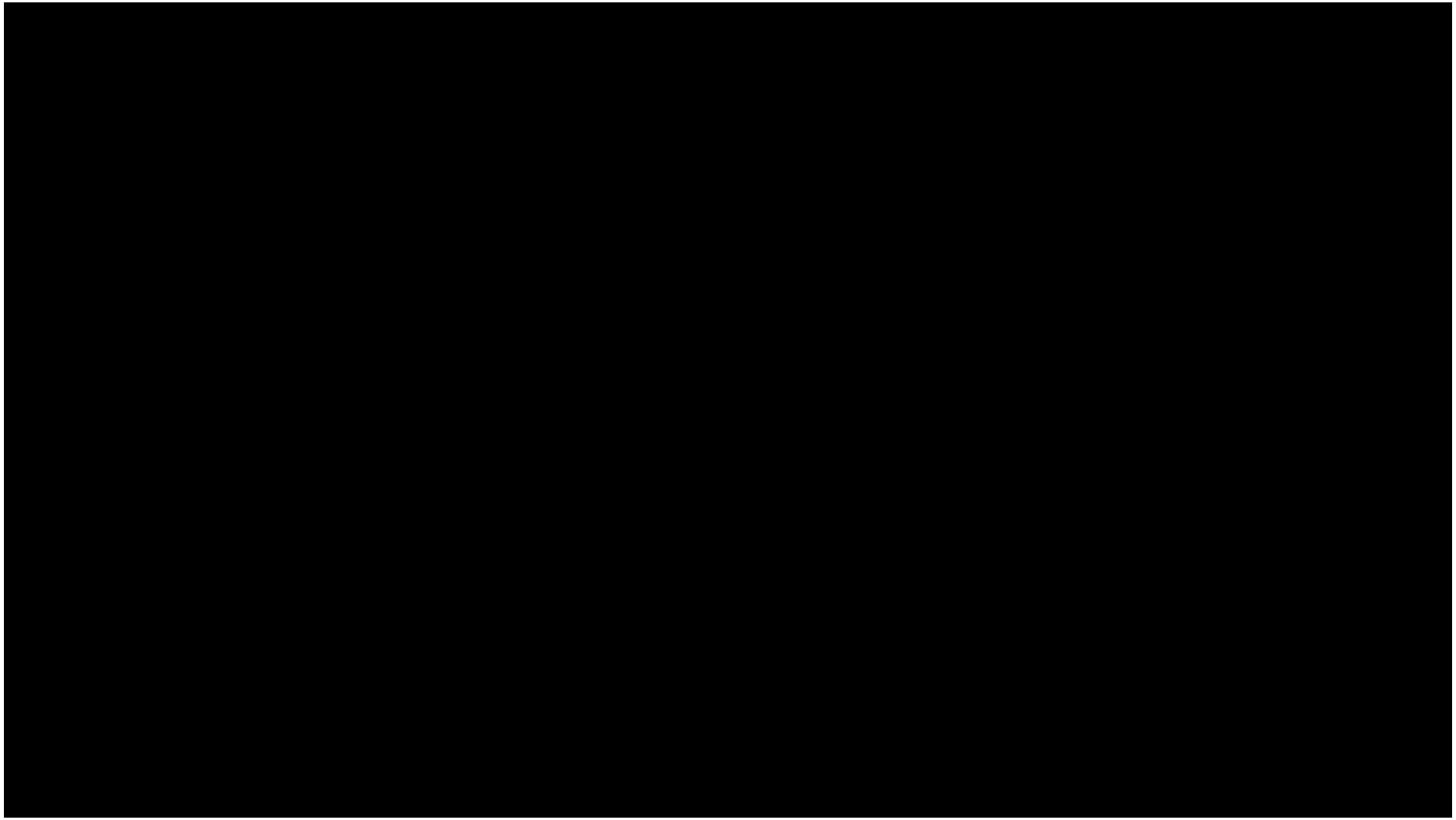
A robot learns about the paths through a maze.

The robot starts from the lower left corner of the maze.
Each location (state) is indicated by a number. There are four action choices (left, right, up, down).
When the robot hits a wall, it receives reward -1.
When it reaches an open location, it receives reward 0.
When it reaches the exit, it receives reward 100.



Slide credit: Eric Eaton

Reinforcement Learning



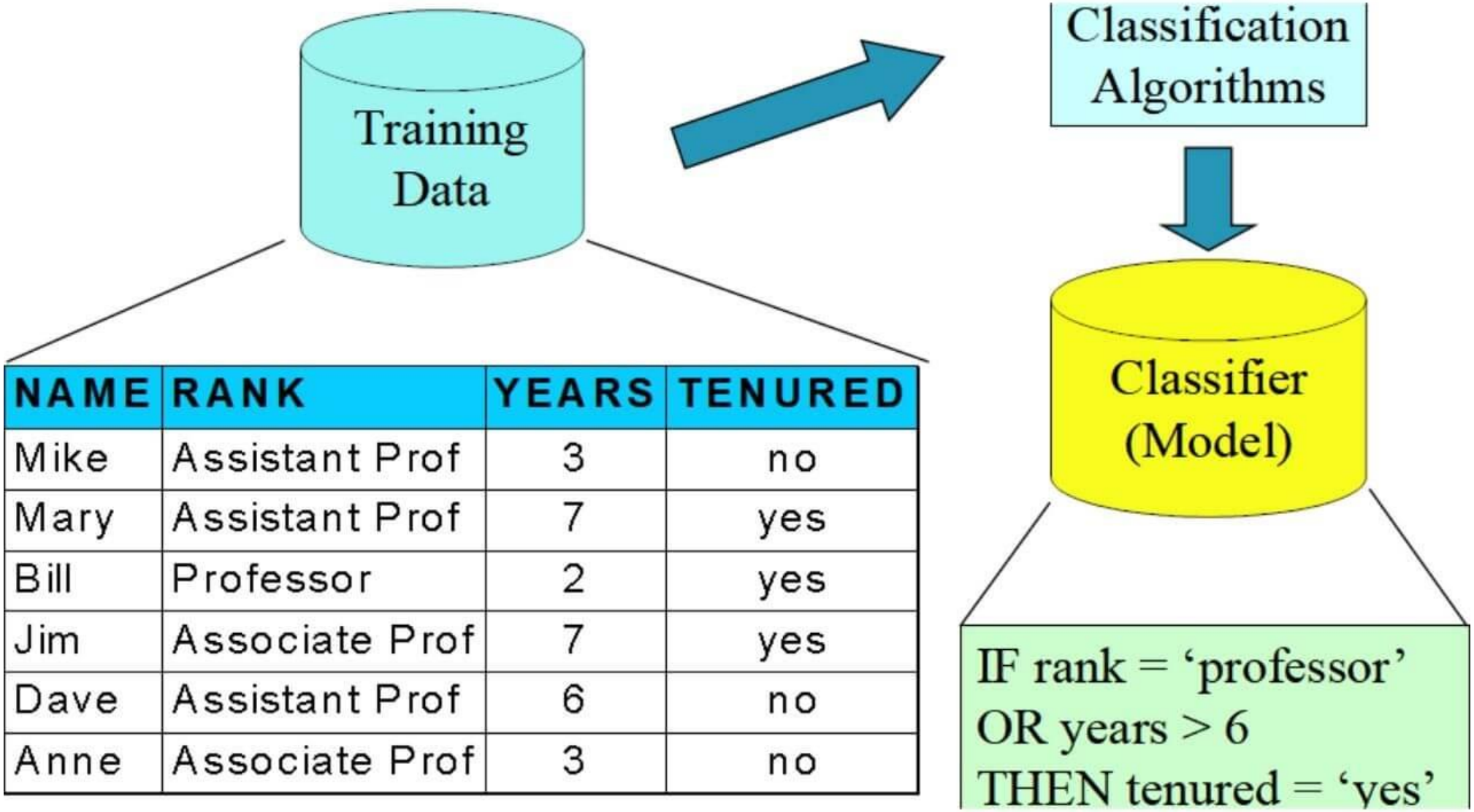
Slide credit: Eric Eaton

Quy trình học máy

- Quy trình học máy chia làm 2 giai đoạn
- Giai đoạn huấn luyện (training/learning)
  - Sử dụng một bộ dữ liệu huấn luyện để xây dựng tập quy tắc (hoặc mô hình) để giải bài toán.
- Giai đoạn thử nghiệm (testing)
  - Kiểm nghiệm tập quy tắc (hoặc mô hình) xây dựng được ở giai đoạn huấn luyện trên một bộ dữ liệu (gọi là bộ dữ liệu kiểm chứng)
  - Một số phương pháp đánh giá mô hình phổ biến: độ chính xác (accuracy), confusion metric, precision, recall, F1-Score, MSE.

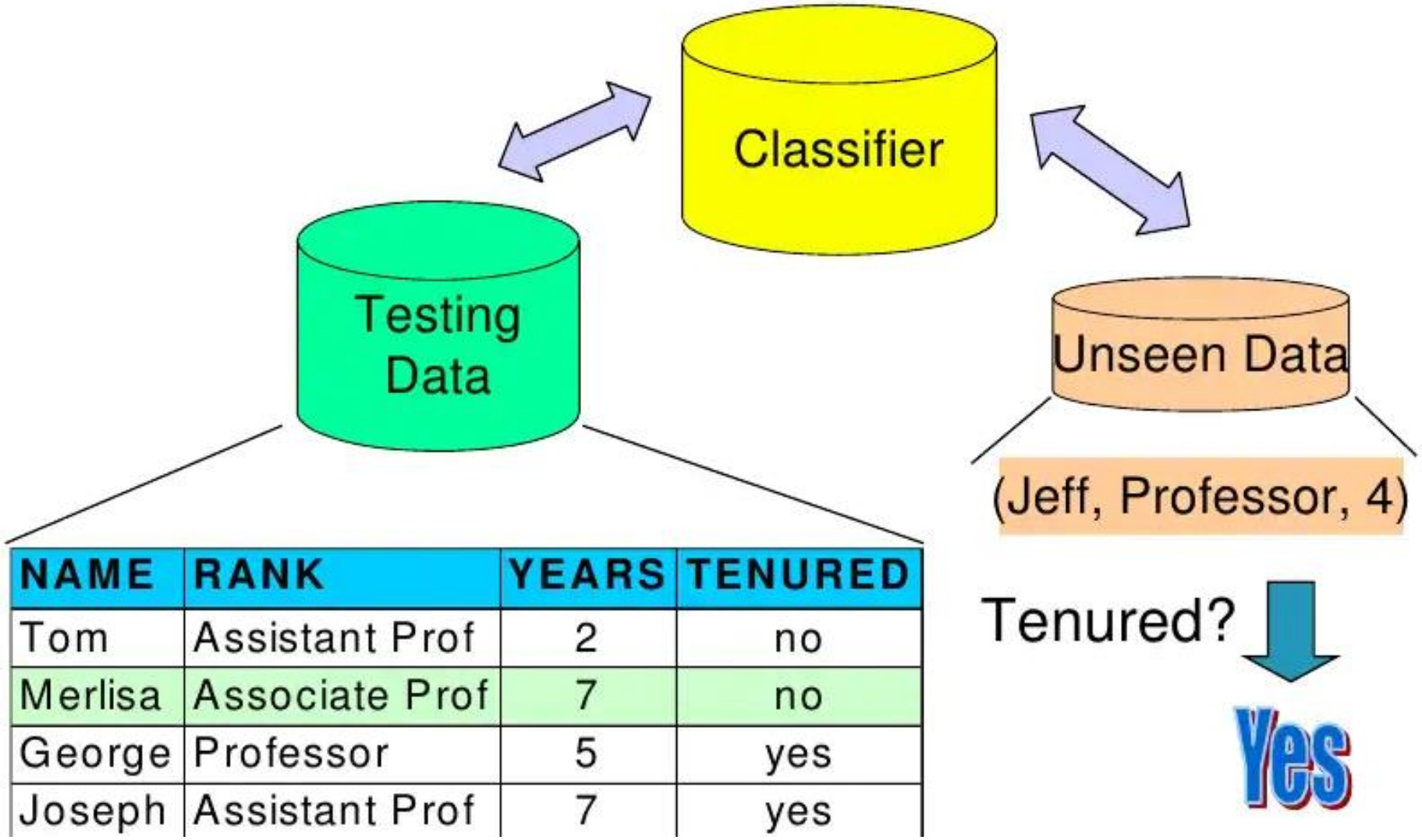
BÀI TOÁN PHÂN LỚP (CLASSIFICATION)

Giai đoạn huấn luyện (training phase)



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Giai đoạn kiểm nghiệm (testing phase)



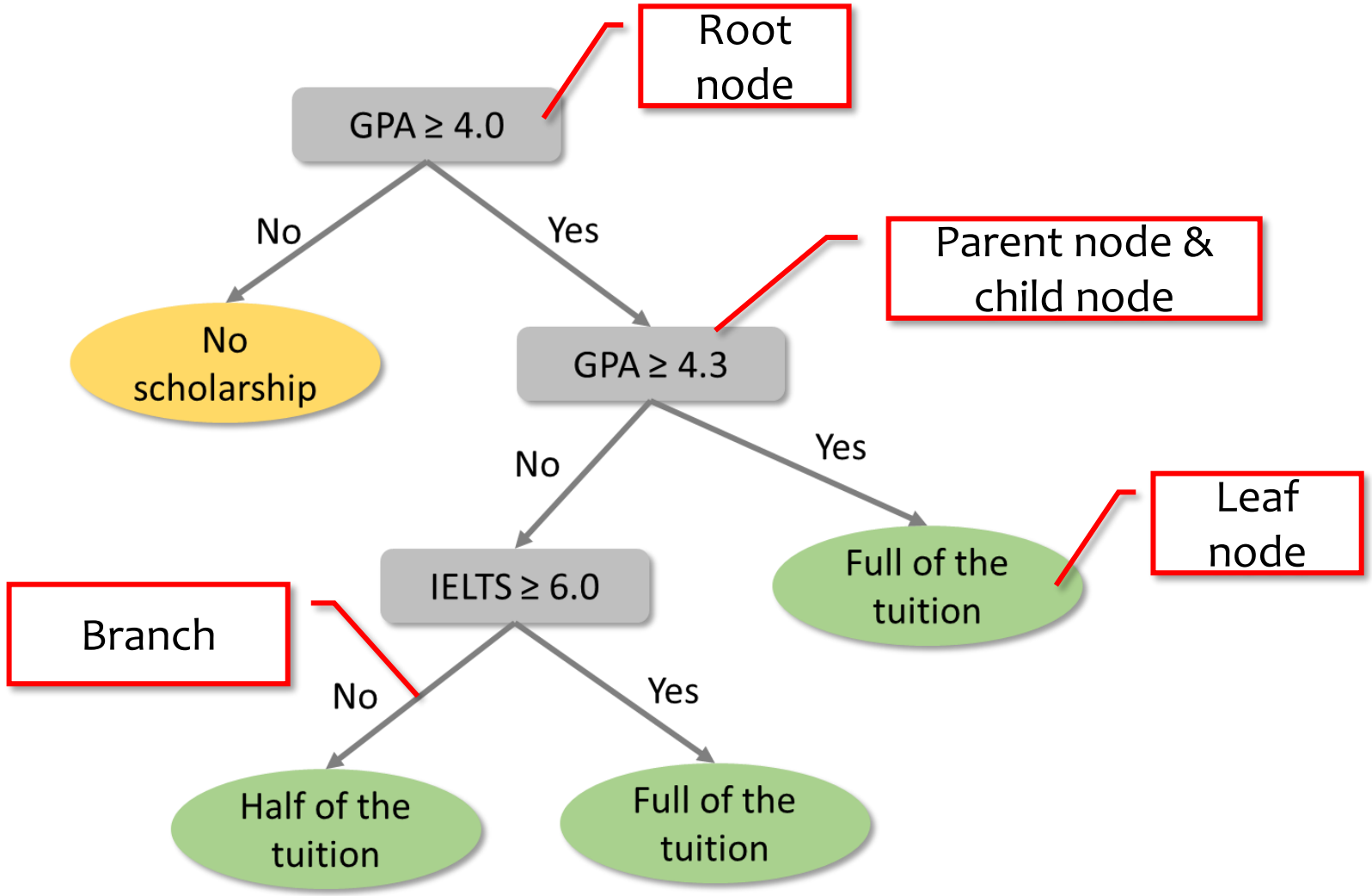
NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Thuật toán phân lớp

- Cây quyết định (Decision tree)
- Bộ phân lớp Bayes (Naive Bayes)
- Hồi quy logistic (Logistic regression)
- Random forests
- Mạng neuron
- …

CÂY QUYẾT ĐỊNH

Biểu diễn cây quyết định



Ví dụ: Cây quyết định cho bài toán *PlayTennis*

- Bộ dữ liệu huấn luyện

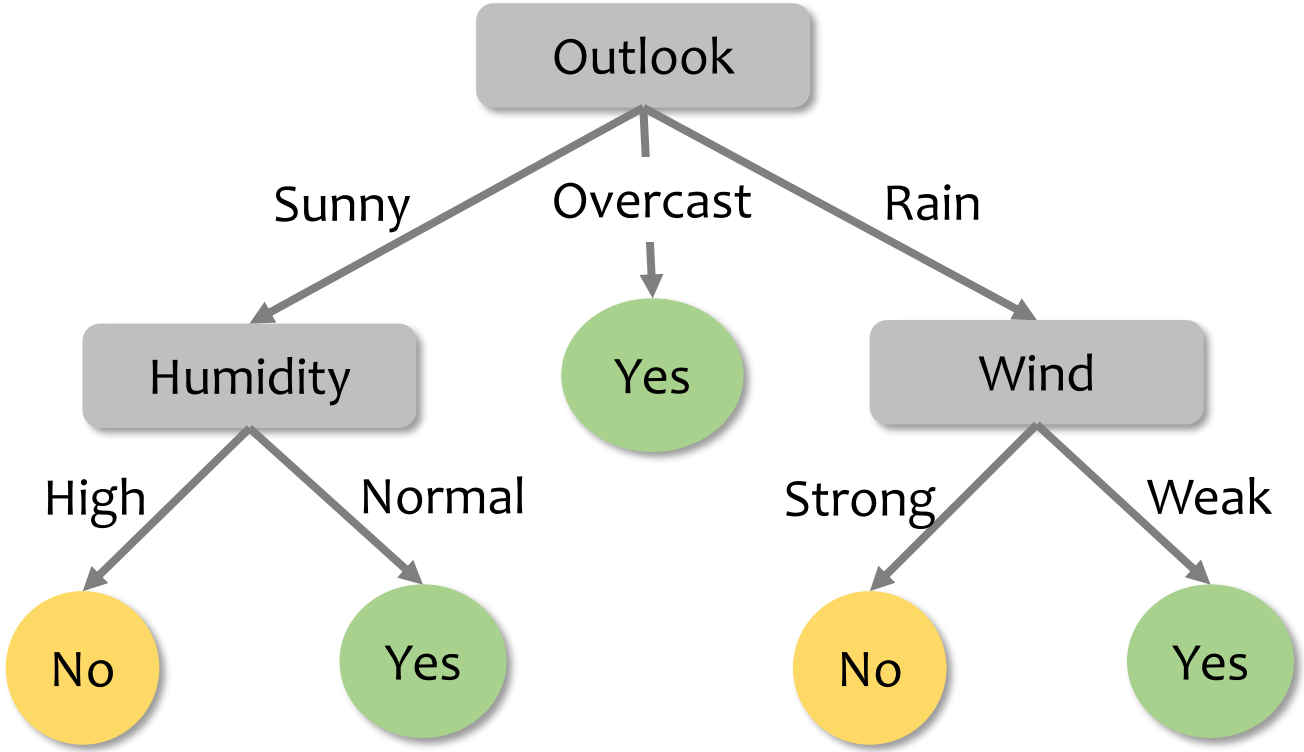
Day	Outlook	Temperature	Humidity	Wind	Playtennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision Tree for *PlayTennis*

Day	Outlook	Temperature	Humidity	Wind	Playtennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

All observations in the data set are perfectly described by the tree.

Question: How do we build such trees?



Xây dựng cây quyết định

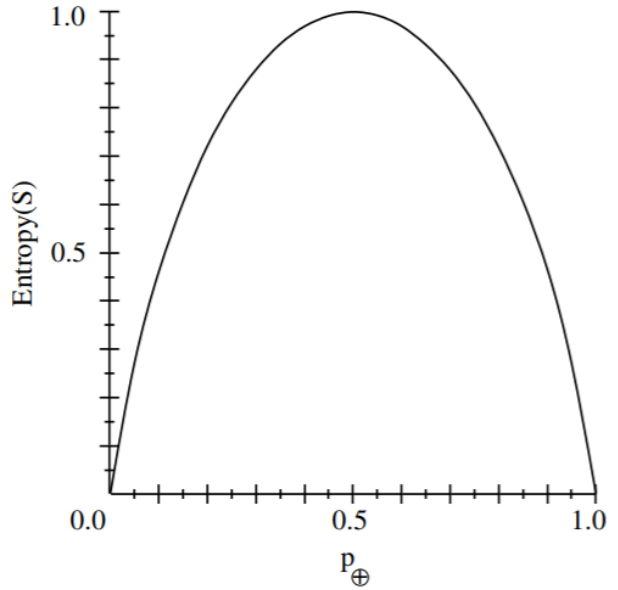
Thuật toán xây dựng cây quyết định nhỏ nhất là bài toán thuộc nhóm NP-problem.

- Áp dụng giải thuật tham lam để xây dựng cây quyết định:
  - Khởi tạo cây quyết định rỗng
  - Lần lượt chọn các thuộc tính tốt nhất để đưa vào cây quyết định.
  - Lặp tới khi cây quyết định được xây dựng xong.

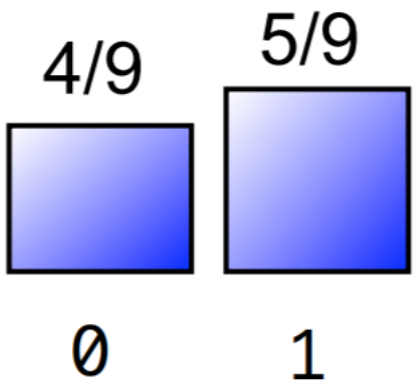
- Thế nào là thuộc tính tốt nhất?
  - Sử dụng lý thuyết Entropy để xác định.

Entropy

- S = bộ dữ liệu huấn luyện.
  - S gồm các quan sát thuộc một trong hai lớp + (positive) hoặc – (negative).
- $p_{\oplus}$ = số lượng quan sát thuộc lớp positive / tổng số quan sát.
- $p_{\ominus}$ = số lượng quan sát thuộc lớp negative / tổng số quan sát.
- $Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$
- Entropy đo lường độ đồng nhất của S.

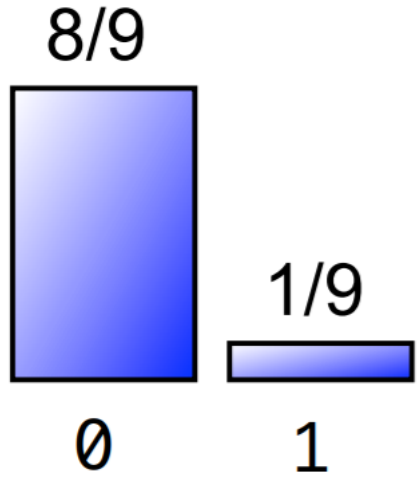


Entropy



Bộ dữ liệu có tính đồng nhất thấp

$$-\frac{4}{9}\log_2\frac{4}{9}-\frac{5}{9}\log_2\frac{5}{9}\approx 0.99$$

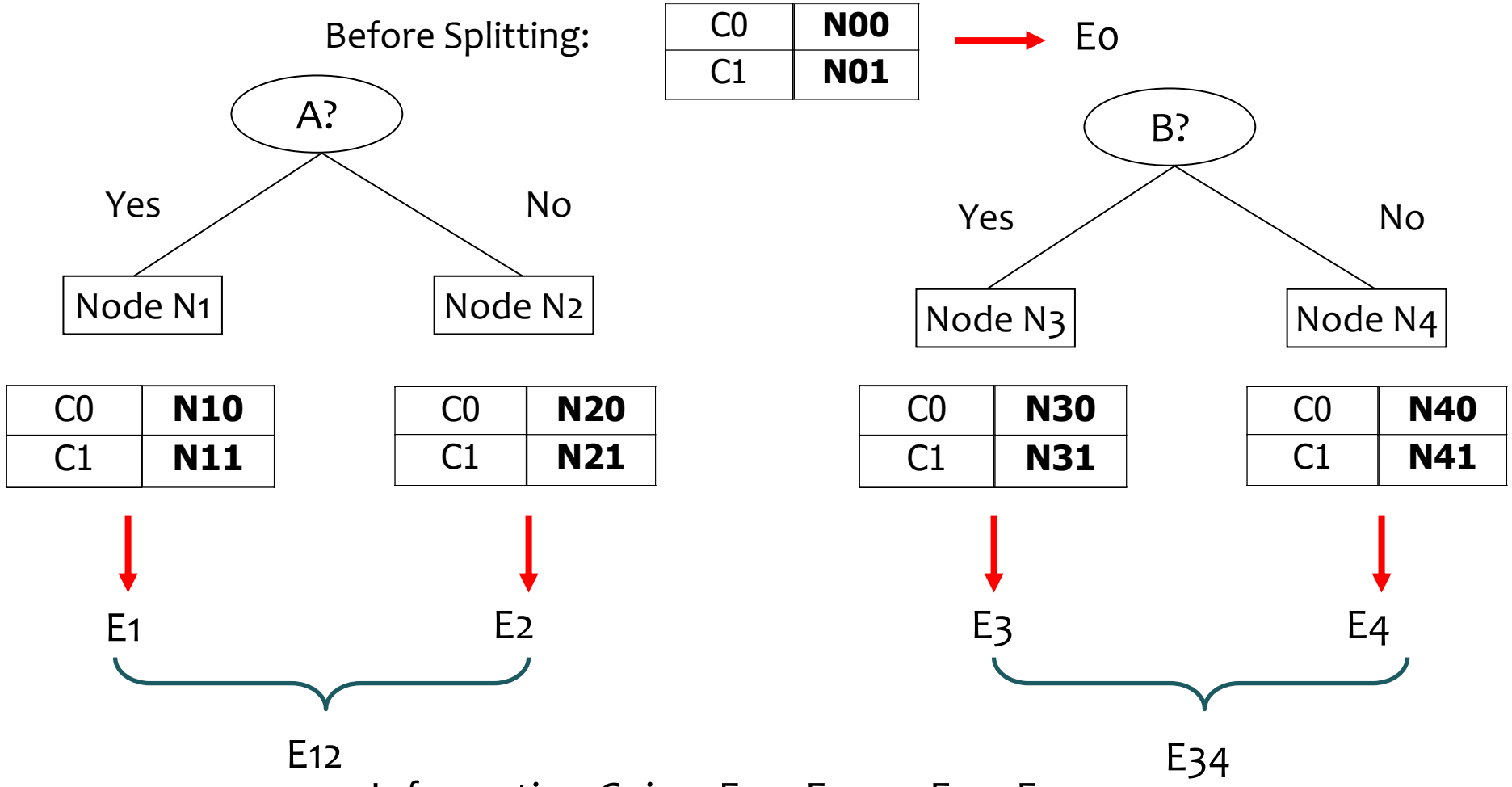


Bộ dữ liệu có tính đồng nhất cao

$$-\frac{8}{9}\log_2\frac{8}{9}-\frac{1}{9}\log_2\frac{1}{9}\approx\frac{1}{2}$$

Information Gain

Cách tìm thuộc tính tốt nhất?



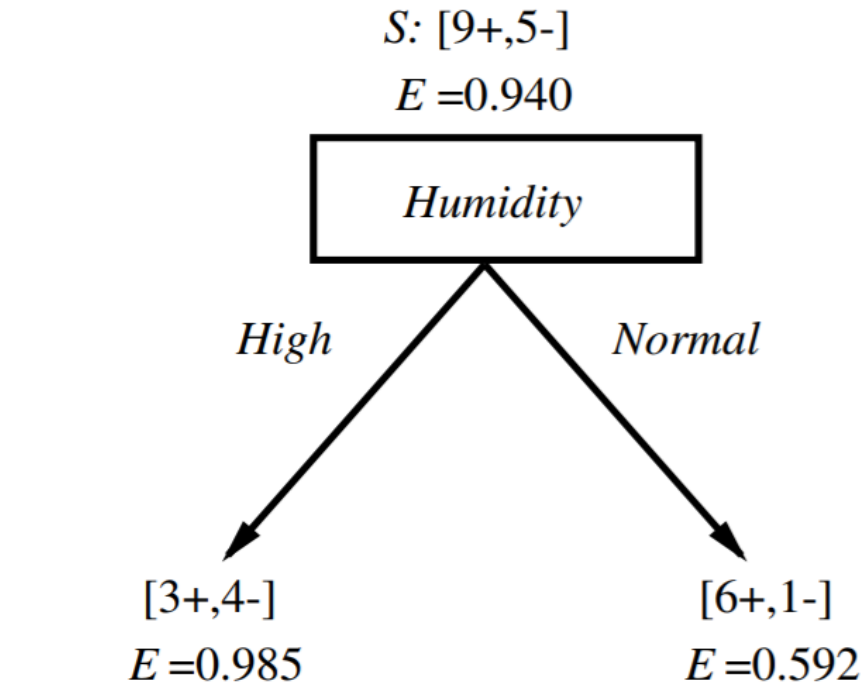
Information Gain = E0 – E12 vs E0 – E34

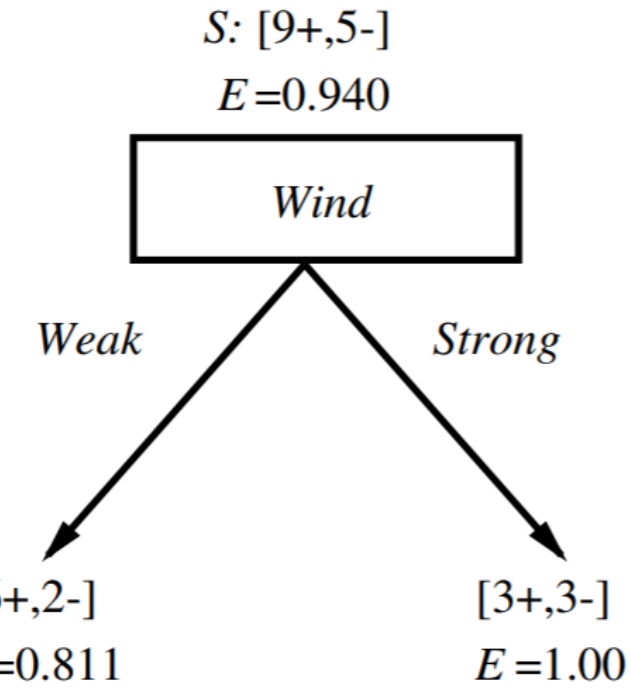
Information Gain

- Let p is a parent node, p is split into k partitions, n_i is number of records in partition i, $n = \sum_{i=1}^{k} n_i$.

$$GAIN = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

Information Gain





Gain (S, Humidity)

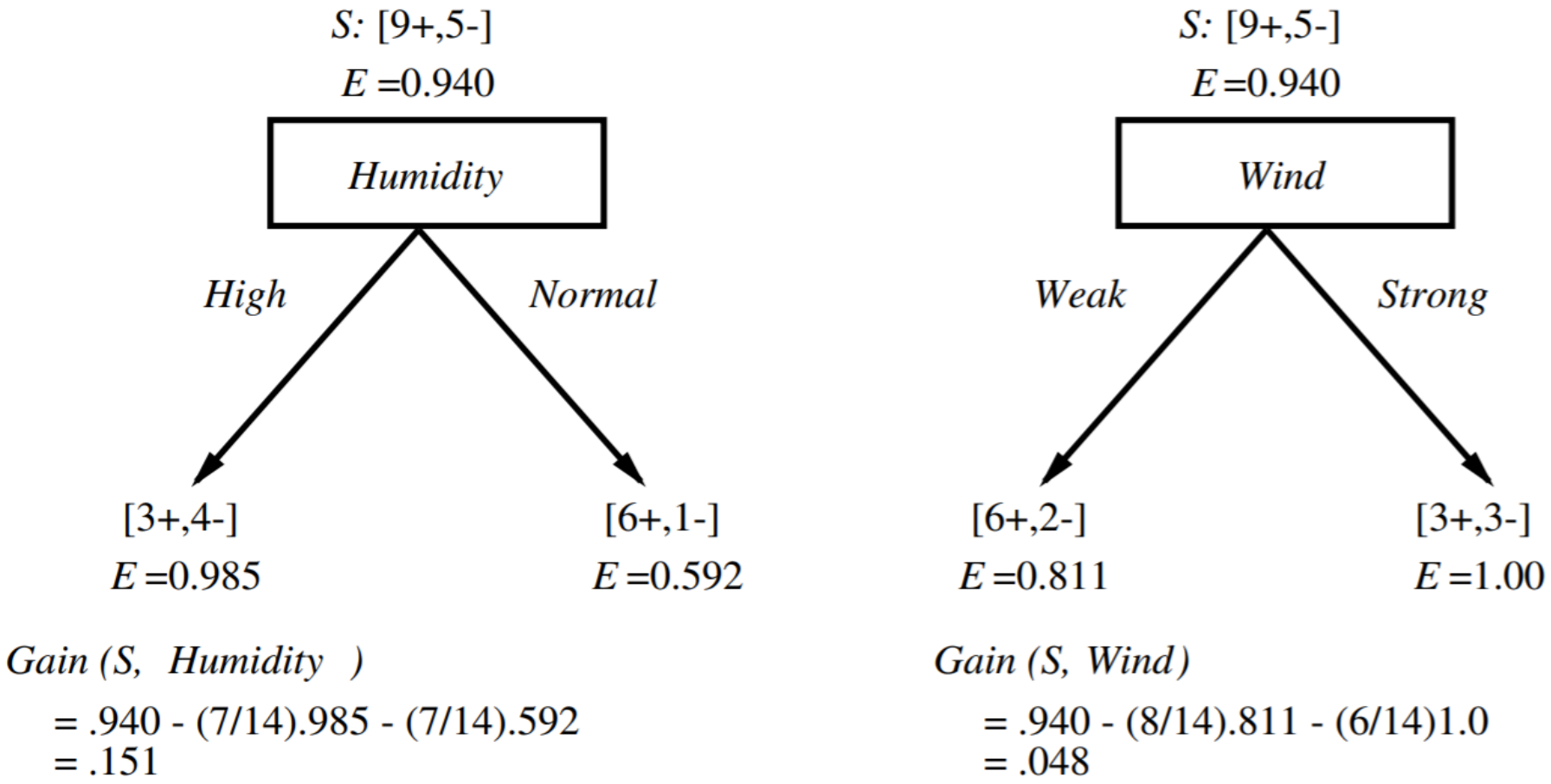
= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

$$GAIN = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

Information Gain



Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

Which attribute is the best classifier?

ID3 Algorithm

$ID3(Examples, Target\_attribute, Attributes)$

- Create a $Root$ for the tree
- If all examples are **positive**, Return single-node tree $Root$, with $label = +$
- If all examples are **negative**, Return single-node tree $Root$, with $label = -$
- If $Attributes$ is empty, Return single-node tree $Root$, with $label =$ most common value of $Target\_attribute$ in $Examples$
- **otherwise**, Begin
  - $A \leftarrow$ attribute in $Attributes$ that best classifies $Examples$
  - decision attribute for $Root \leftarrow A$
  - **For each possible value** v_i **of** A
    - Add new branch below $Root$ with $A = v_i$
    - Let $Examples_{v_i}$ be the subset of $Examples$ with v_i for A
    - If $Examples$ is empty
      - Then add a leaf node with label = most common value of $Target\_attribute$ in $Examples$
      - Else add ID3($Examples_{v_i}, Target\_Attribute, Attributes - \{A\}$)
- Return $Root$

Ví dụ

Day	Outlook	Temperature	Humidity	Wind	Playtennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Bài tập

Age	Income	Student	Credit_rating	Buys_computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31…40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31…40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31…40	Medium	No	Excellent	Yes
31…40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

Bài tập

- Cho tập dữ liệu huấn luyện được thể hiện trong bảng sau:
- Xây dựng cây quyết định hai mức dự đoán một chuyến bay có bị chậm trễ không
- Tính error rate trên tập dữ liệu huấn luyện.

Thuộc tính	Giá trị của thuộc tính	Số lượng chuyến bay bị trễ	Số lượng chuyến bay không bị trễ
Mưa	YES	30	10
	NO	10	30
Gió	YES	25	15
	NO	15	25
Mùa hè	YES	5	35
	NO	35	5
Mùa đông	YES	20	10
	NO	20	30
	NO	25	30

BỘ PHÂN LỚP BAYES

Định lý Bayes

Likelihood

Prior

$$P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)}$$

Normalization Constant

$$P(X_1,\ldots,X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

(Giả sử X1, …, Xn là độc lập theo điều kiện Y)

Day	Outlook	Temperature	Humidity	Wind	Playtennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

A new day				
outlook	temperature	humidity	windy	play
sunny	cool	high	true	?

Cần tính:

P(play = Yes | outlook = sunny, temperature = cool, humidity = high, windy = true)

P(play = No | outlook = sunny, temperature = cool, humidity = high, windy = true)

Huấn luyện bộ phân lớp Bayes

- Tính P(Y=v):

$$P(Y = v) = \frac{Count(Y = v)}{\#\ records}$$

- Tính $P(X_i=u|Y=v)$:

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v)}{Count(Y = v)}$$

The weather data, with counts and probabilities													
outlook			**temperature**			**humidity**			**windy**			**play**	
	yes	**no**		**yes**	**no**		**yes**	**no**		**yes**	**no**	**yes**	**no**
sunny	2	3	**hot**	2	2	**high**	3	4	**false**	6	2	9	5
overcast	4	0	**mild**	4	2	**normal**	6	1	**true**	3	3		
rainy	3	2	**cool**	3	1								
sunny	2/9	3/5	**hot**	2/9	2/5	**high**	3/9	4/5	**false**	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**mild**	4/9	2/5	**normal**	6/9	1/5	**true**	3/9	3/5		
rainy	3/9	2/5	**cool**	3/9	1/5								

P(play = Yes) = 9/14
P(play = No) = 5/14
P(outlook = sunny | play = yes) = 2/9
P(outlook = sunny | play = no) = 3/5

P(outlook = overcast | play = yes) = 4/9
P(outlook = overcast | play = no) = 0/5
P(outlook = rainy | play = yes) = 3/9
P(outlook = rainy | play = no) = 2/5

Cần tính:

P(play = Yes | outlook = sunny, temperature = cool, humidity = high, windy = true)

$$= \frac{P(\text{outlook = sunny, temperature = cool, humidity = high, windy = true}|\text{play = Yes}) \times P(\text{play = Yes})}{M}$$

$$= \frac{P(\text{outlook = sunny}|\text{play = Yes}) \times P(\text{temperature = cool}|\text{play = Yes}) \times P(\text{humidity = high}|\text{play = Yes}) \times P(\text{windy = true}|\text{play = Yes}) \times P(\text{play = Yes})}{M}$$

P(play = No | outlook = sunny, temperature = cool, humidity = high, windy = true)

$$= \frac{P(\text{outlook = sunny, temperature = cool, humidity = high, windy = true}|\text{play = No}) \times P(\text{play = No})}{M}$$

$$= \frac{P(\text{outlook = sunny}|\text{play = No}) \times P(\text{temperature = cool}|\text{play = No}) \times P(\text{humidity = high}|\text{play = No}) \times P(\text{windy = true}|\text{play = No}) \times P(\text{play = No})}{M}$$

M = P(outlook = sunny, temperature = cool, humidity = high, windy = true|play = Yes) x P(play = Yes) +
P(outlook = sunny, temperature = cool, humidity = high, windy = true|play = No) x P(play = No)

P(outlook = sunny, temperature = cool, humidity = high, windy = true|play = Yes) x P(play = Yes)
$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$
= 0.0053

P(outlook = sunny, temperature = cool, humidity = high, windy = true|play = No) x P(play = No)
$= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}$
= 0.0206

P(play = Yes | outlook = sunny, temperature = cool, humidity = high, windy = true)
= 0.0053 / (0.0053 + 0.0206)
= 0.205

P(play = No | outlook = sunny, temperature = cool, humidity = high, windy = true)
= 0.0206 / (0.0053 + 0.0206)
= 0.795

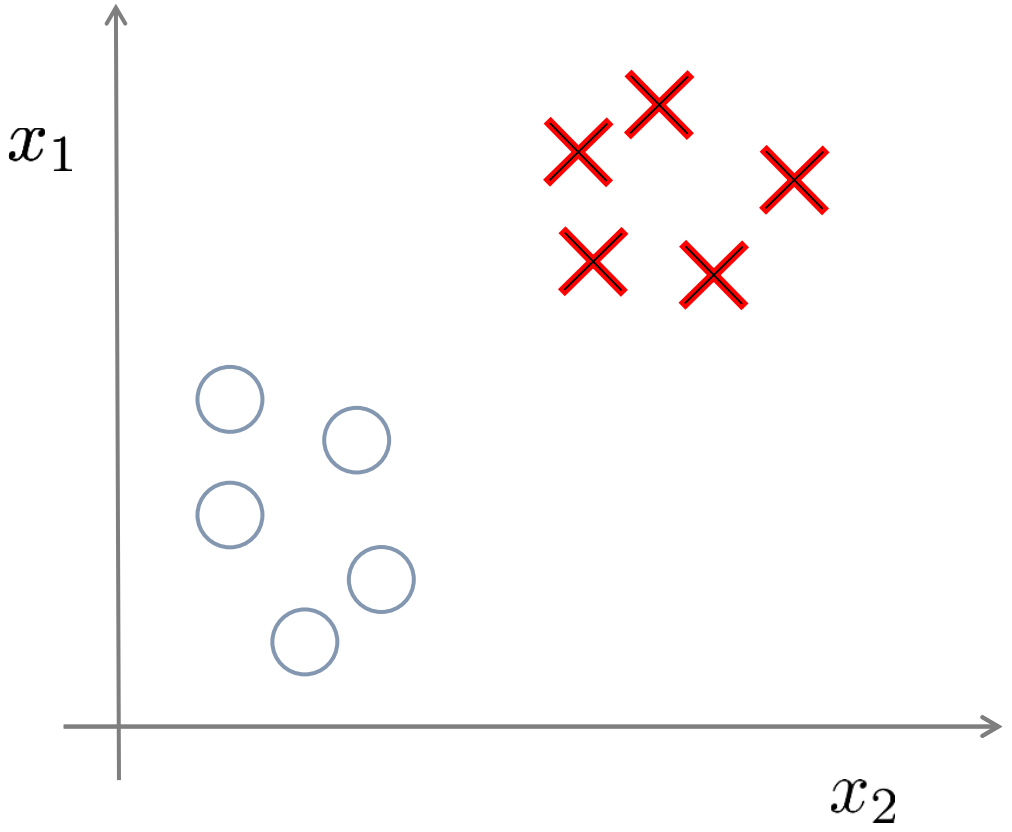
→ Dự đoán: play = No

Bài tập

Thuộc tính	Giá trị của thuộc tính	Số lượng chuyến bay bị trễ	Số lượng chuyến bay không bị trễ
Mưa	YES	30	10
	NO	10	30
Gió	YES	25	15
	NO	15	25
Mùa hè	YES	5	35
	NO	35	5
Mùa đông	YES	20	10
	NO	20	30
	NO	25	30

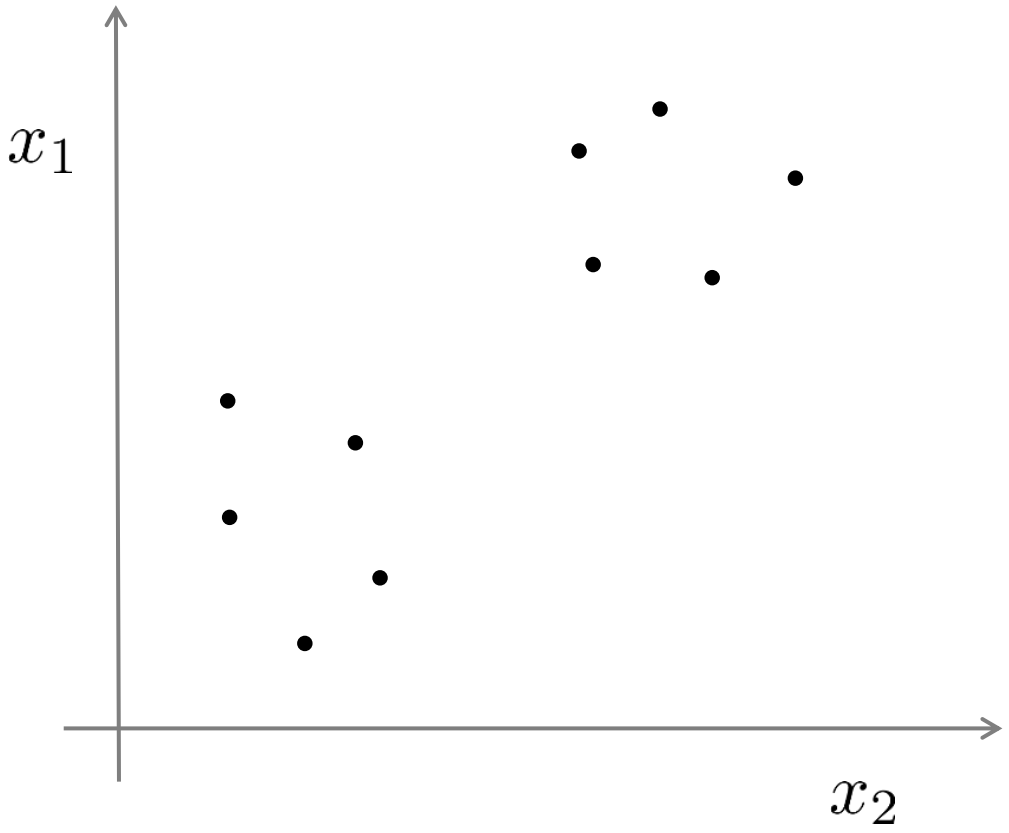
GOM CỤM DỮ LIỆU

Học có giám sát (supervised learning)

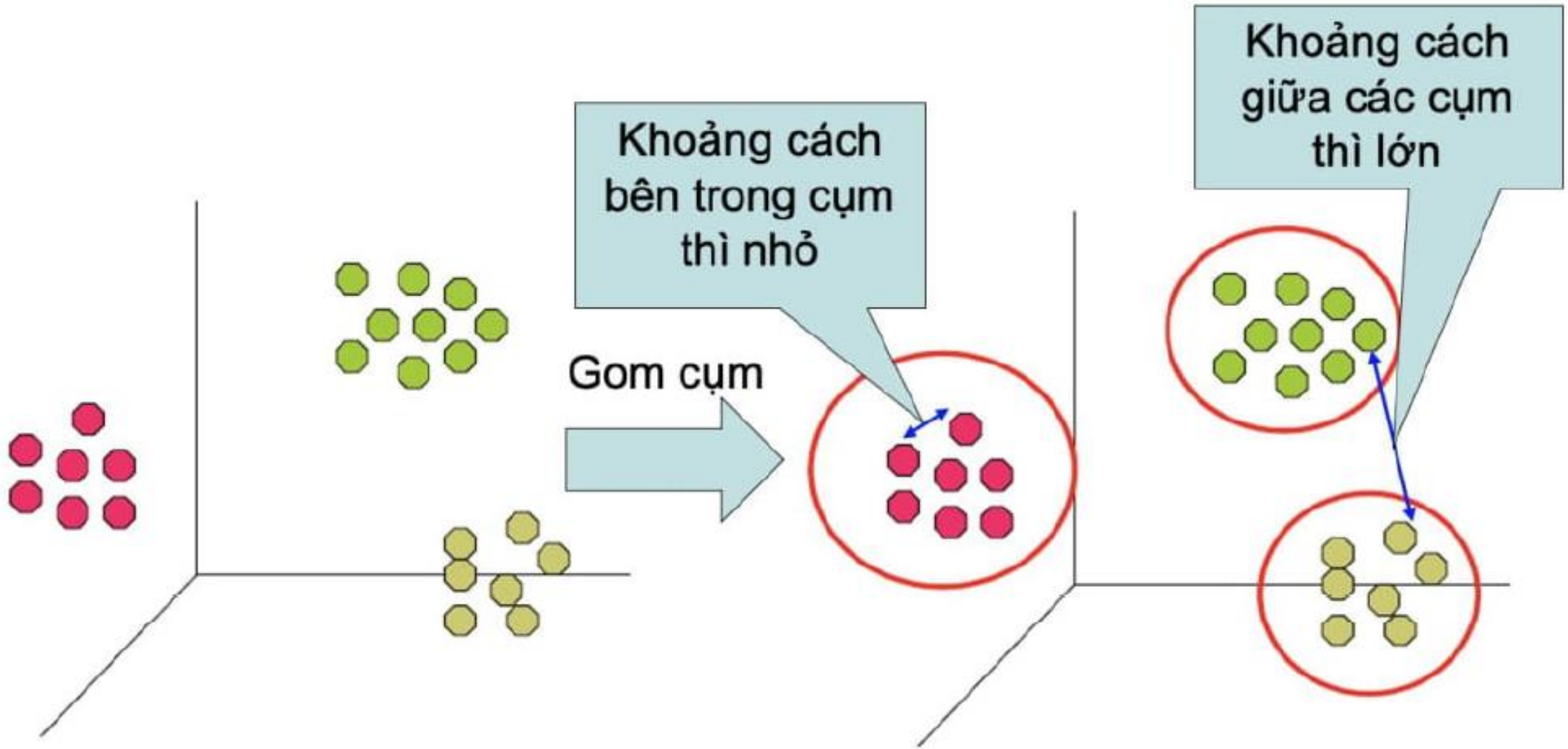


Dữ liệu huấn luyện: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

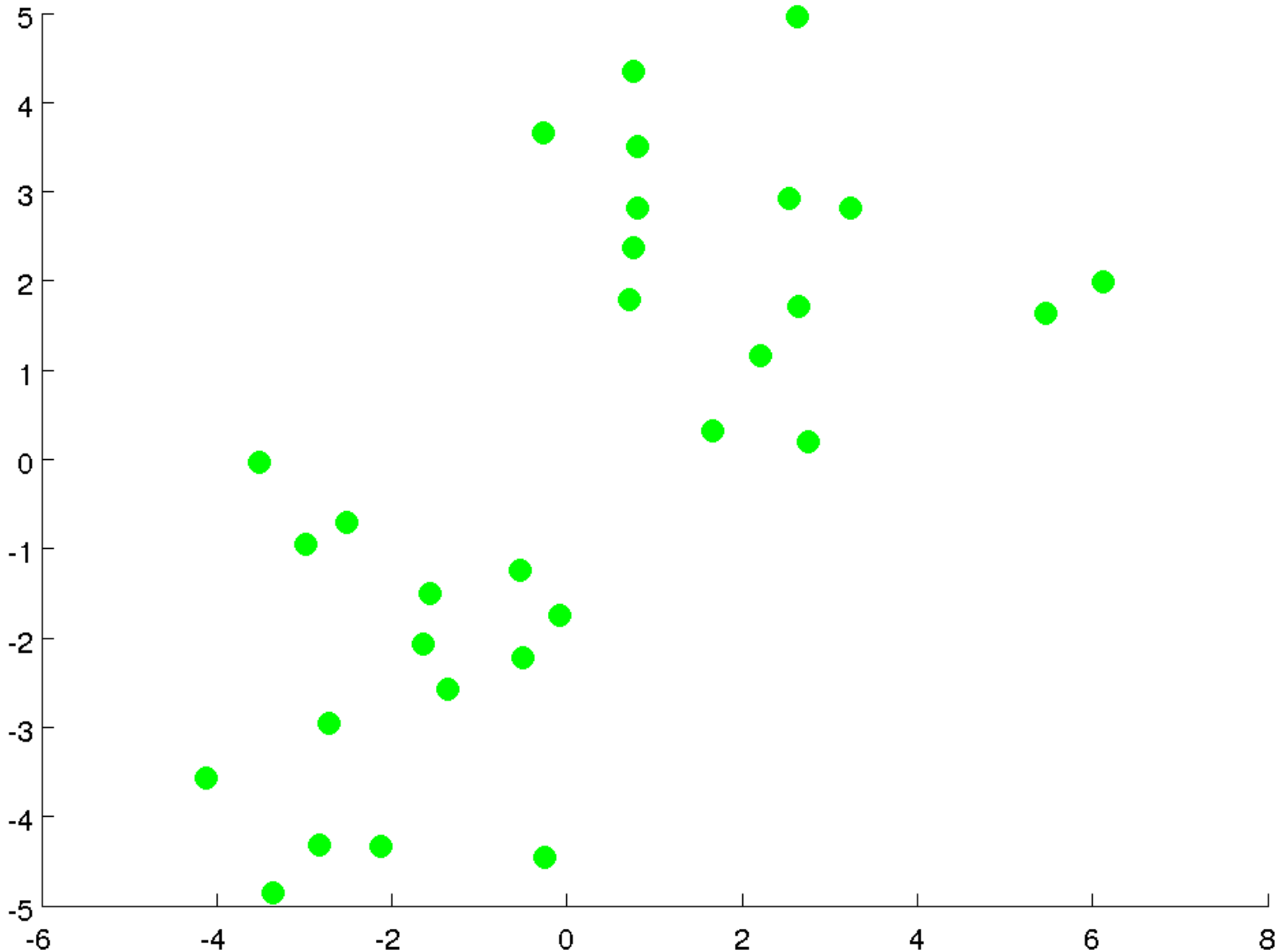
Học không giám sát (unsupervised learning)

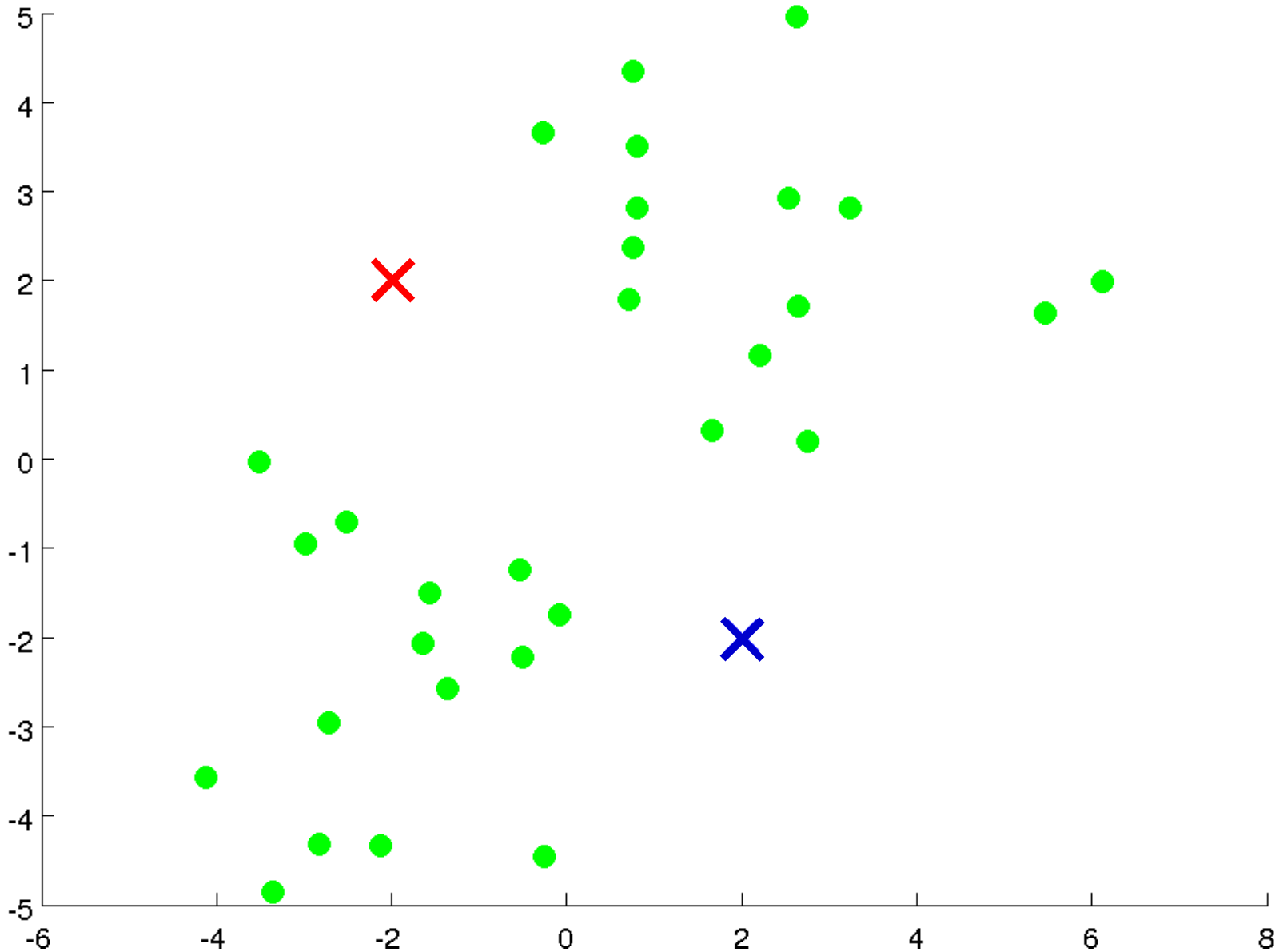


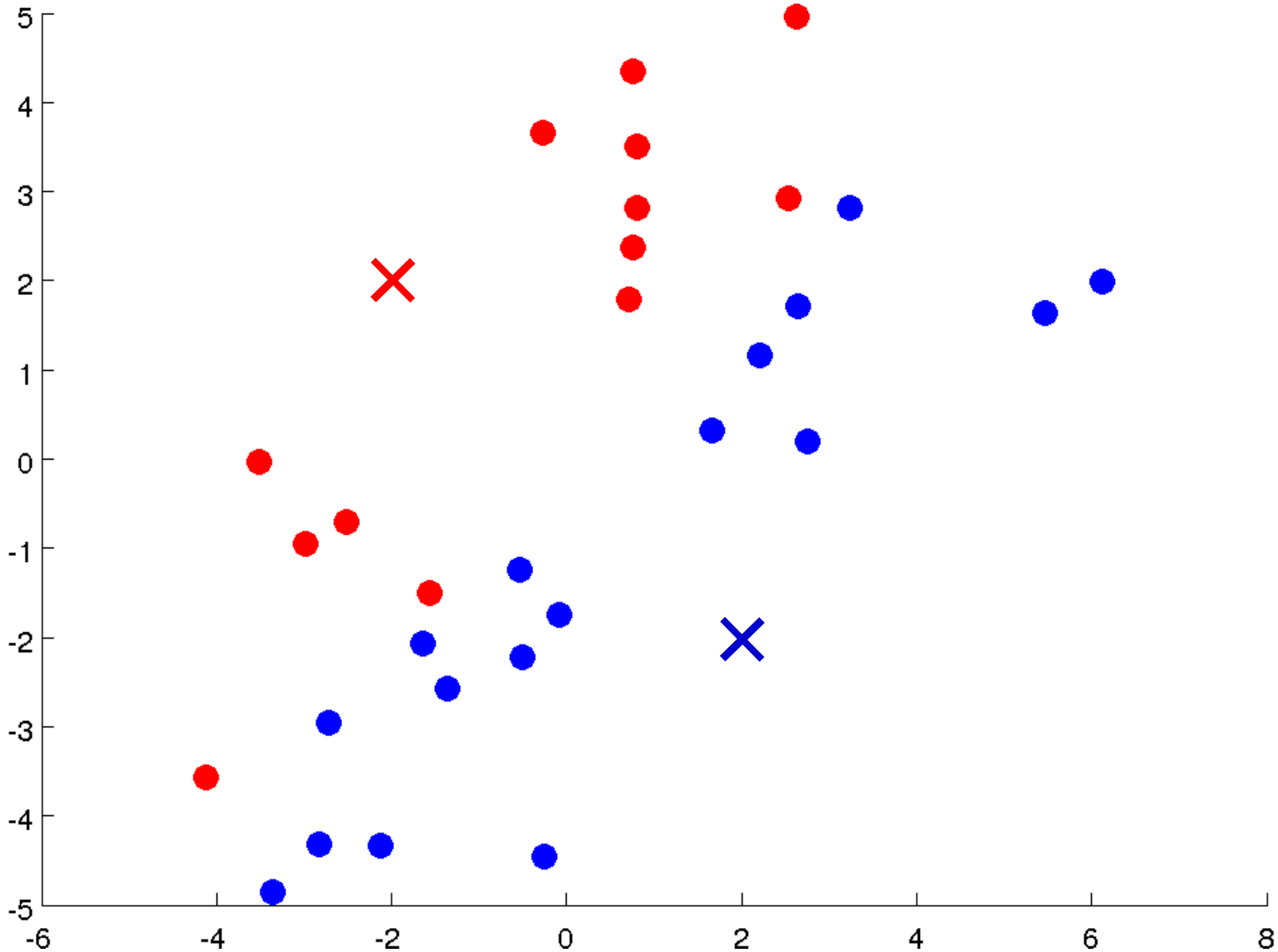
Dữ liệu huấn luyện: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

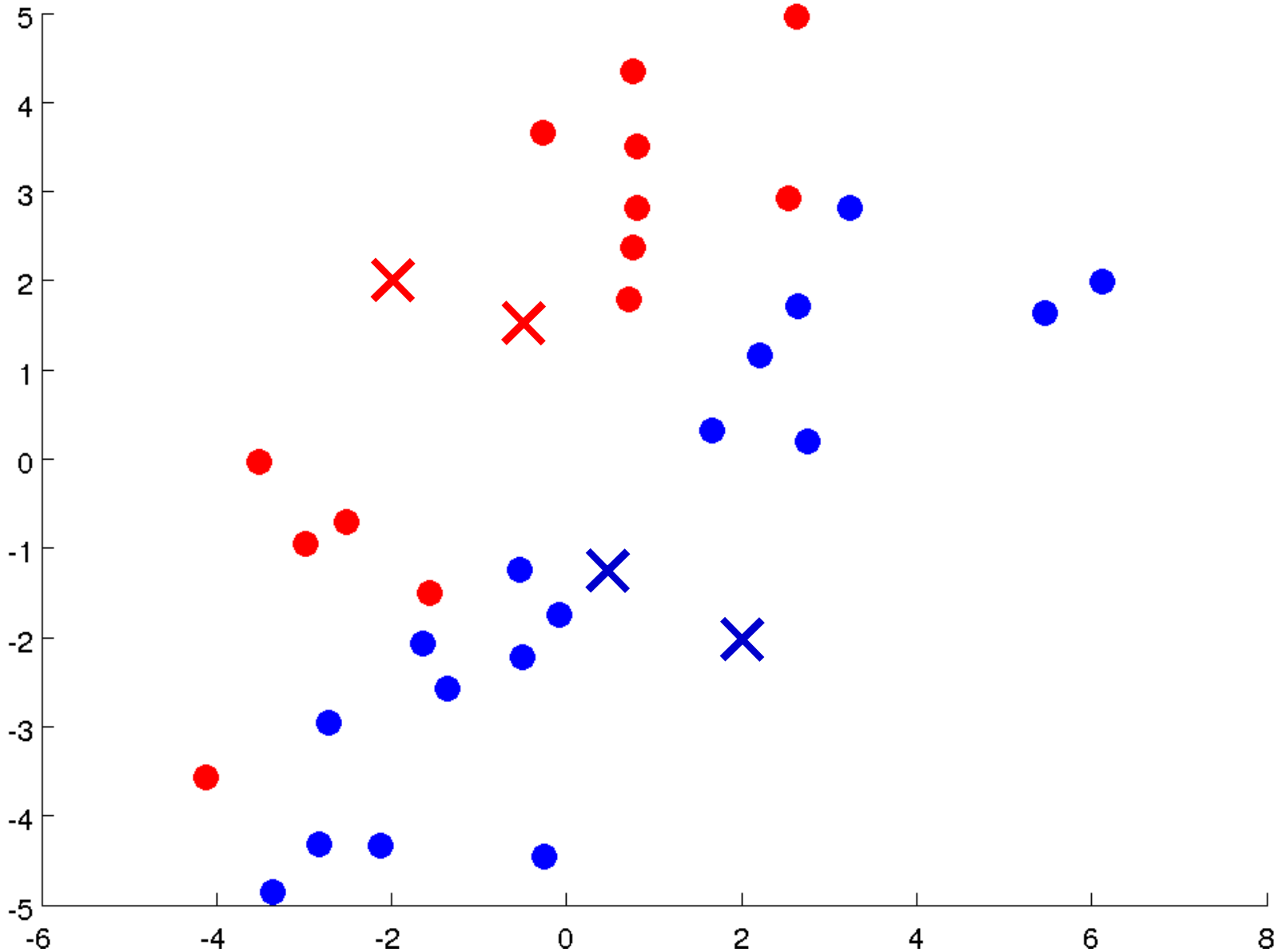

Khoảng cách giữa các cụm thì lớn
Khoảng cách bên trong cụm thì nhỏ
Gom cụm

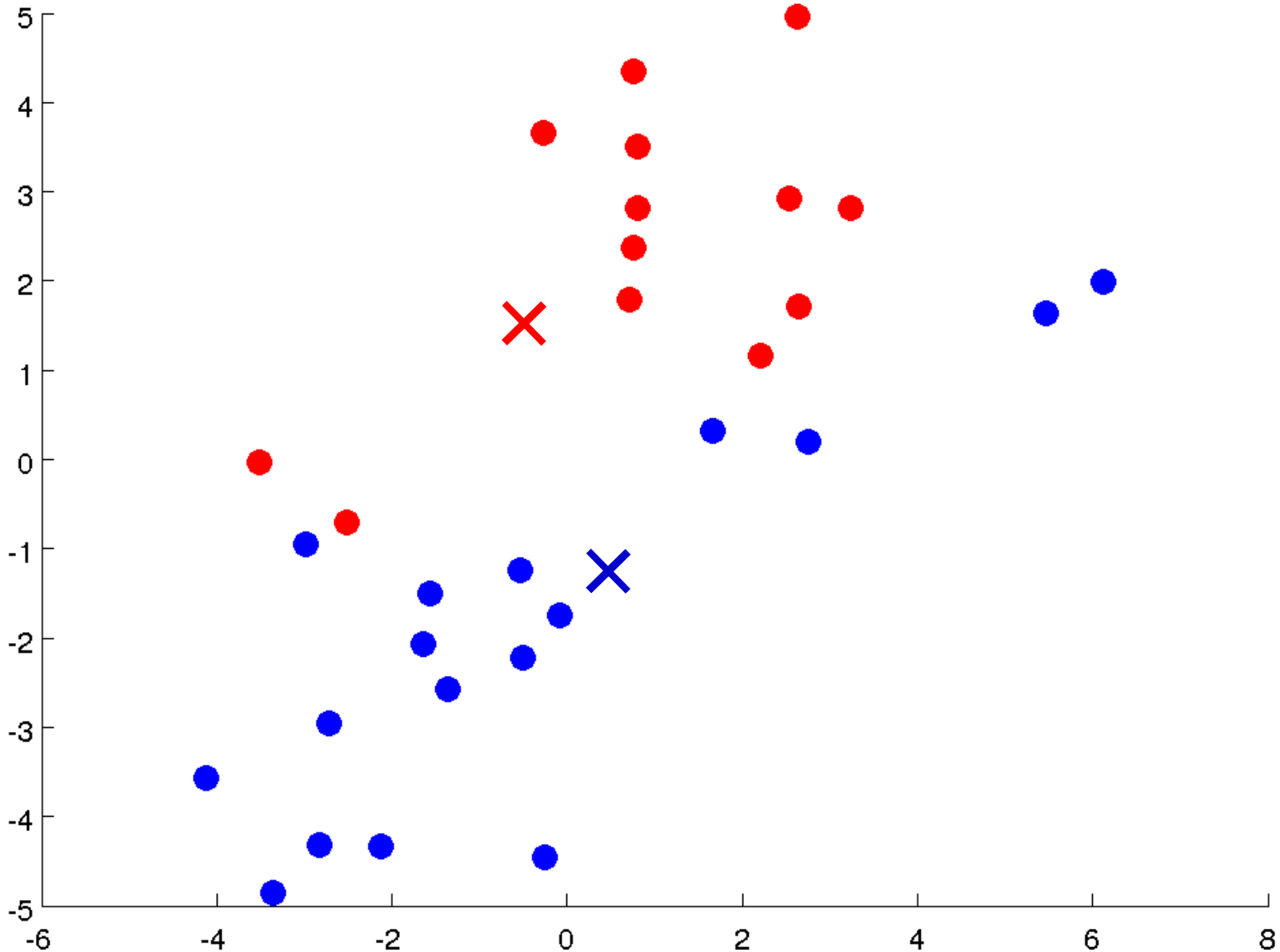
THUẬT TOÁN K-MEANS

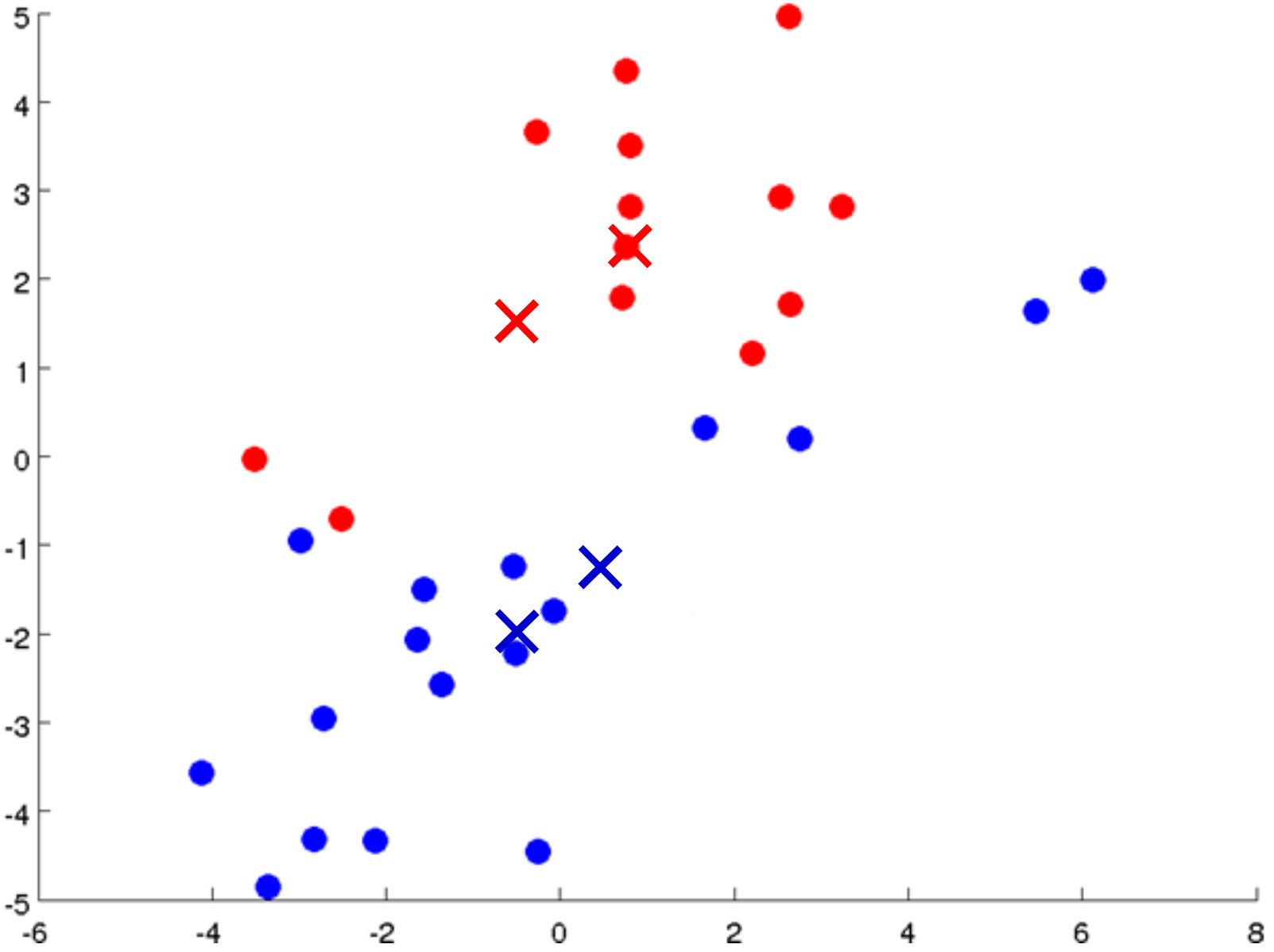

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

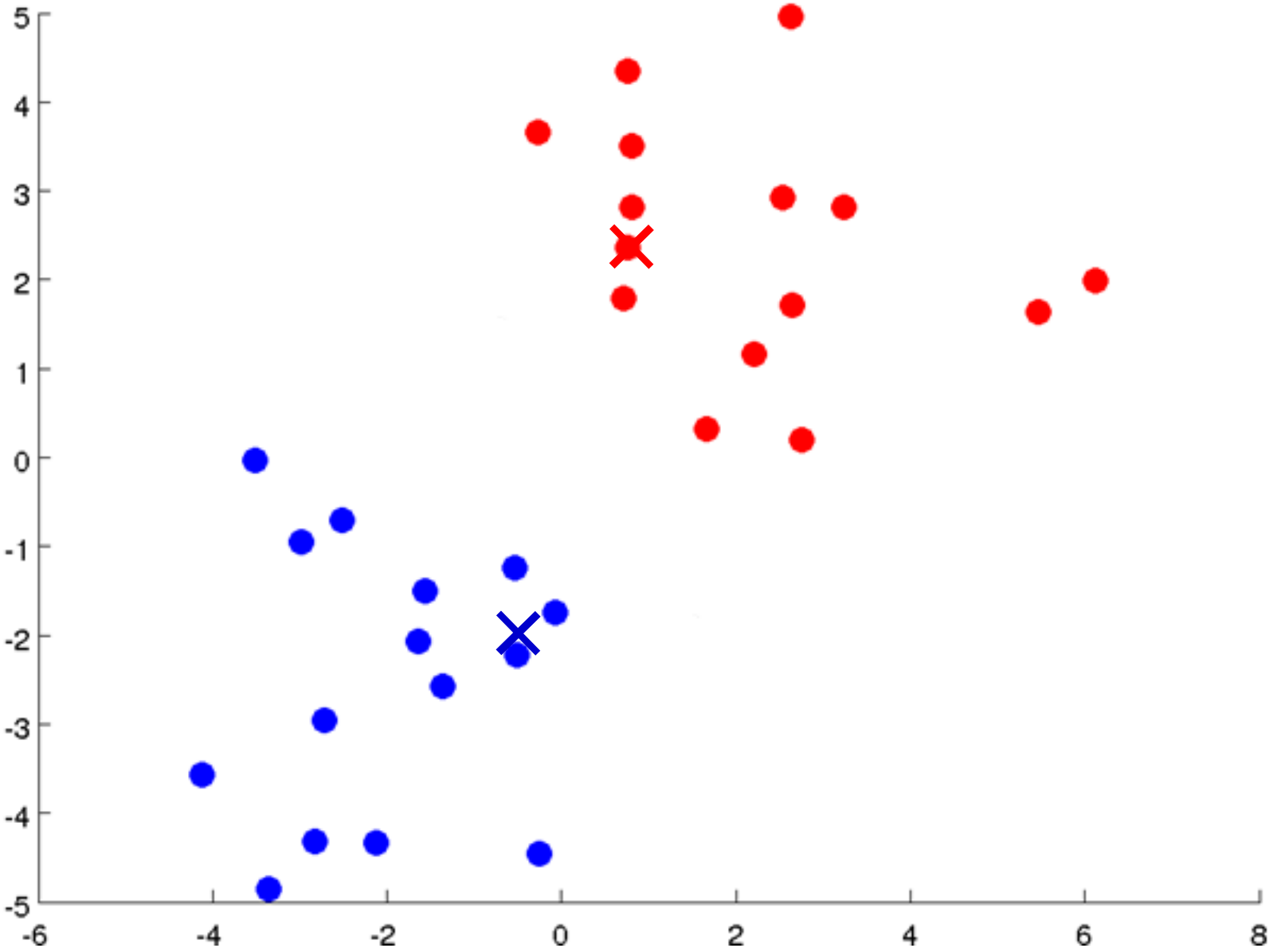

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

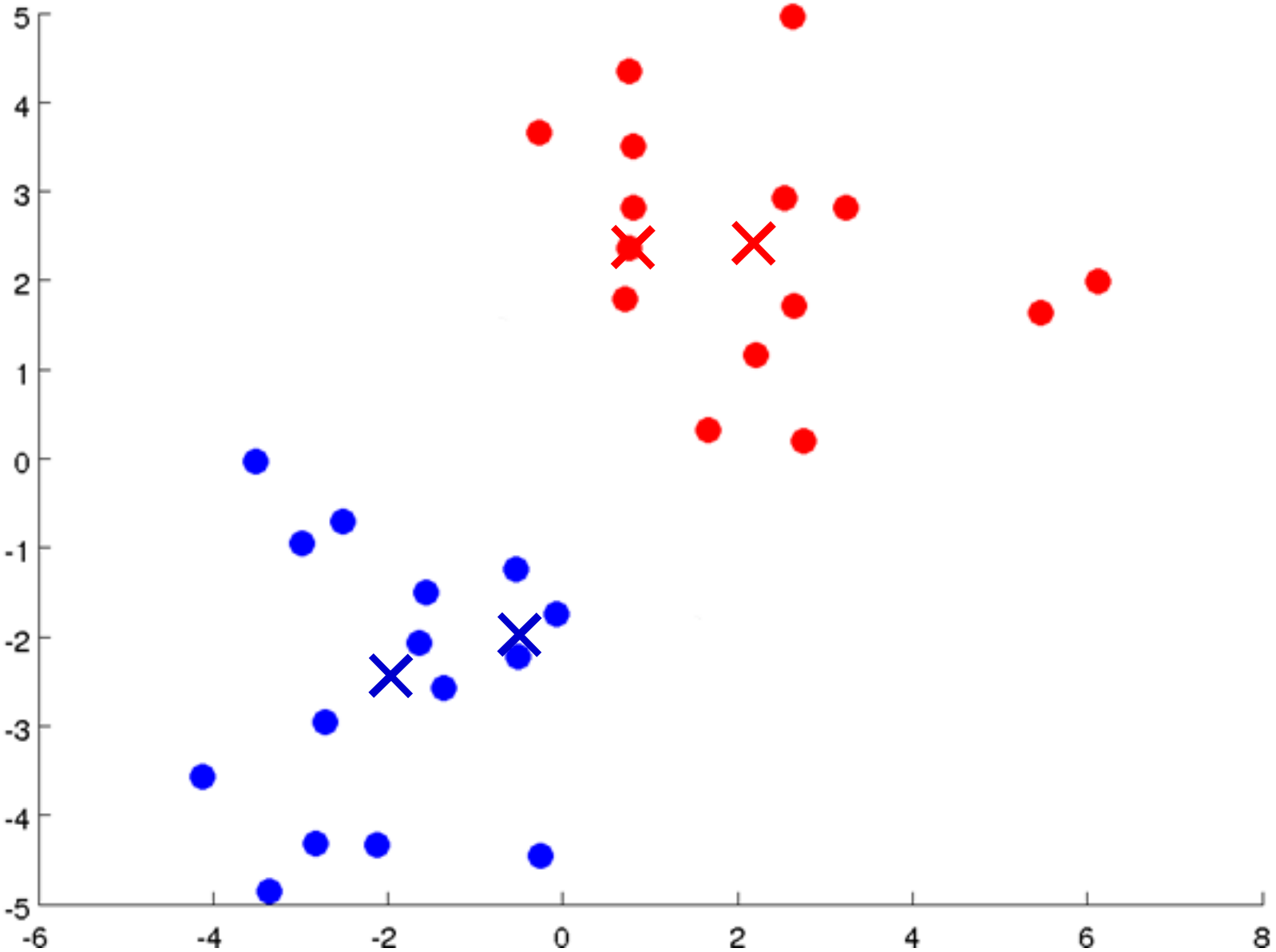

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

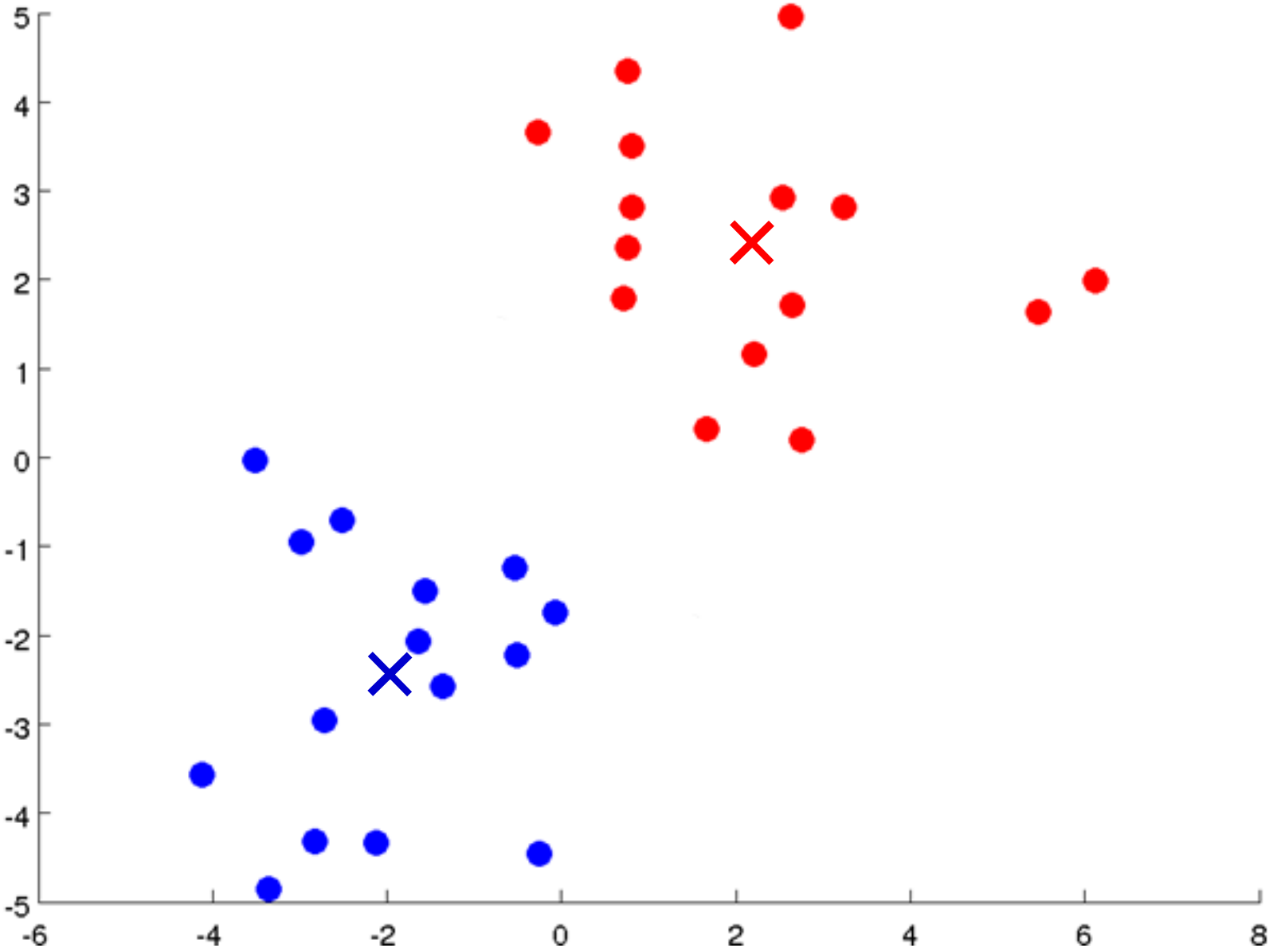

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

Thuật toán K-means

Input:

- K (số lượng cụm)
- Dữ liệu huấn luyện: $\{x^1, x^2, \dots, x^m\}$

Các bước của thuật toán:

- Khởi tạo trọng tâm của K cụm: $m_1, m_2, \dots, m_k$
- Lặp tới khi trọng tâm cụm không thay đổi:
  - Với mọi $i \in [1, m]$, gán x^i vào phân cụm có trọng tâm gần x^i nhất.
  - Tính lại trọng tâm của mỗi cụm.

Ví dụ

- Gom dữ liệu sau thành 2 cụm dựa trên weight index và pH.

Object	Weight index	pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4