Supervised Learning: Basic Concepts

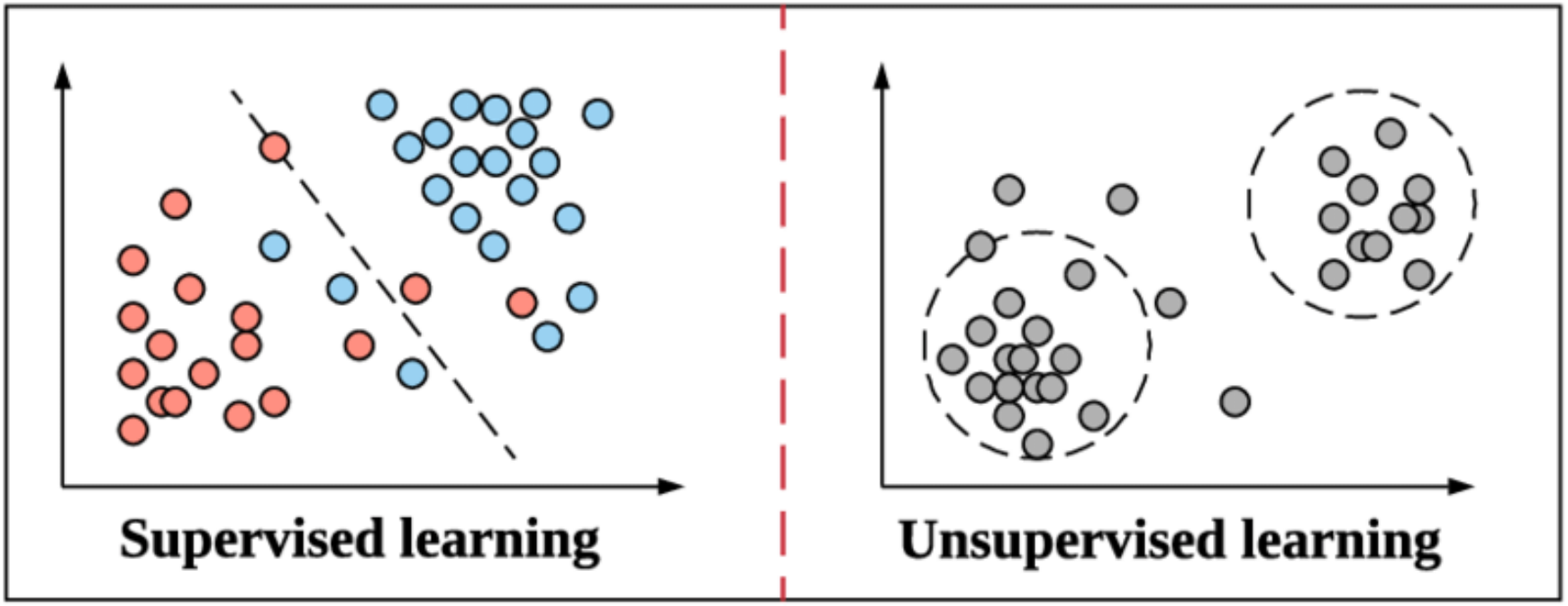
Cheng-Te Li (李政德)

Dept. of Statistics / Institute of Data Science

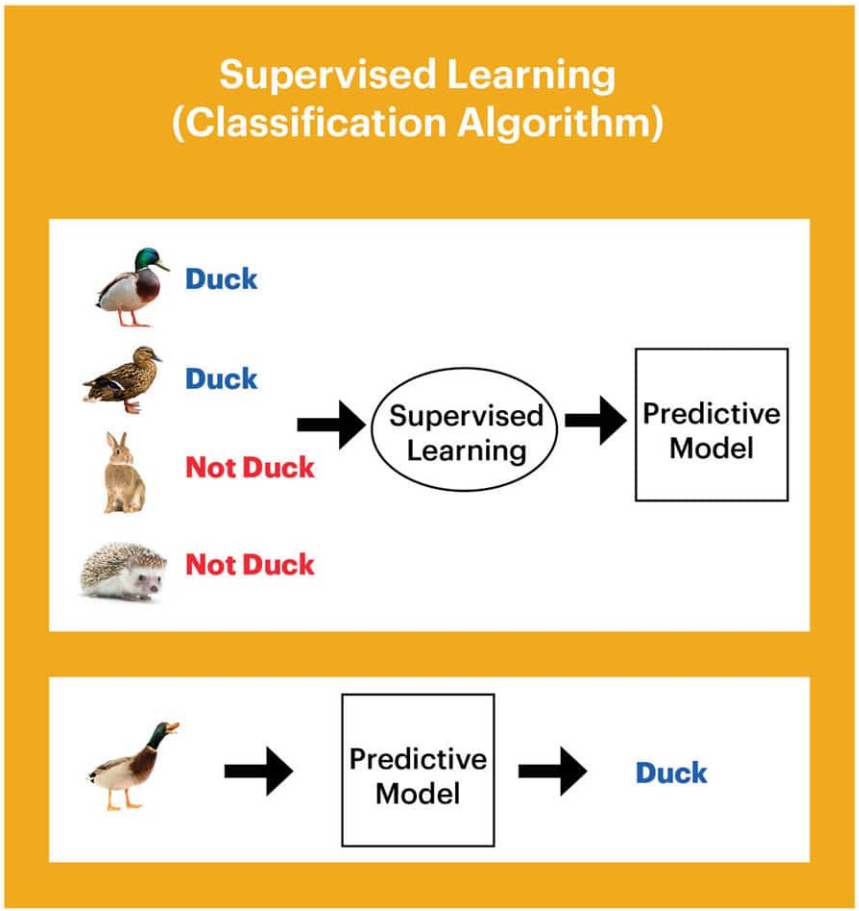
National Cheng Kung University

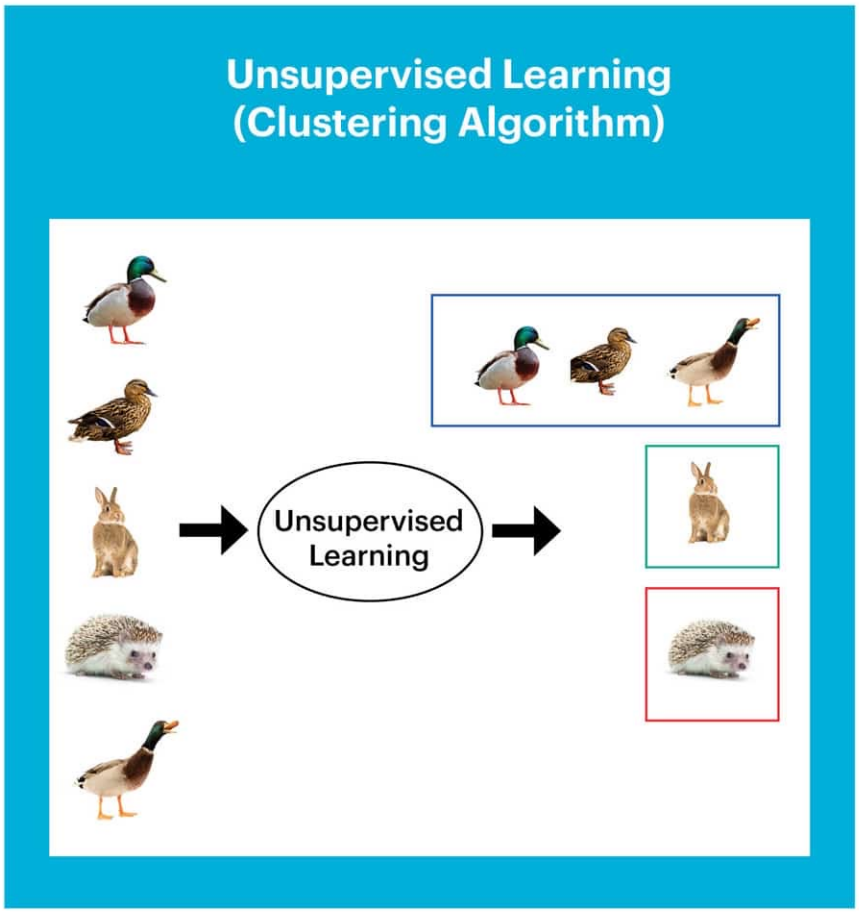
chengte@mail.ncku.edu.tw

Supervised vs. Unsupervised Learning



Supervised vs. Unsupervised Learning

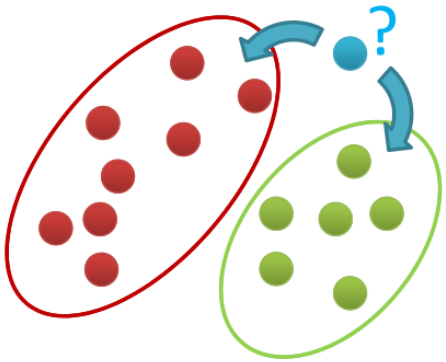




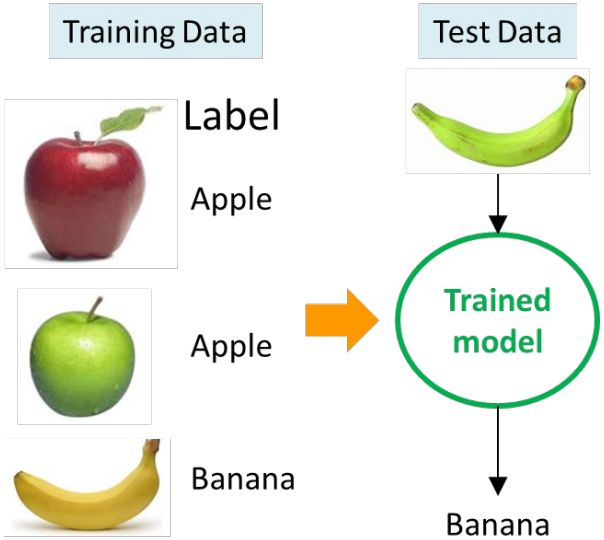
Supervised vs. Unsupervised Learning

Supervised learning:
given data samples with labels

Classification

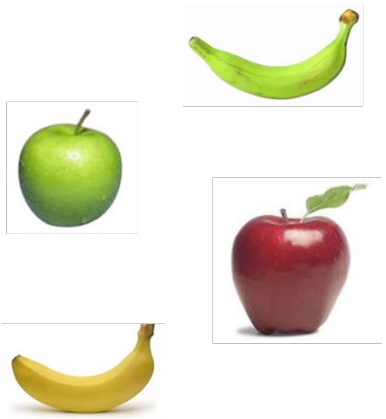


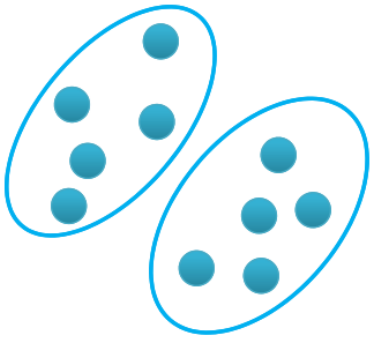
- Groups of data exist
- New data classified to existing groups



Unsupervised learning:
given data, i.e. samples,
but no labels

Clustering

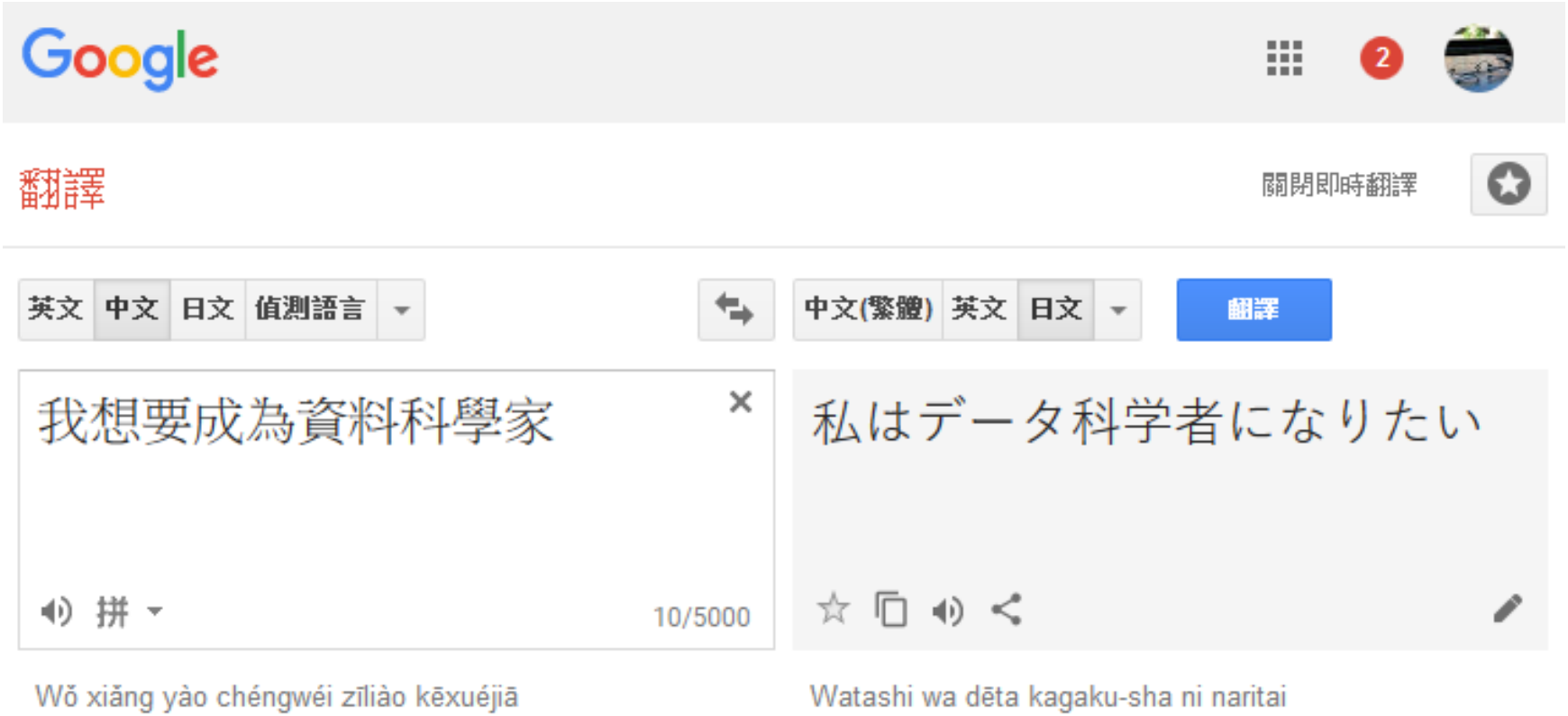




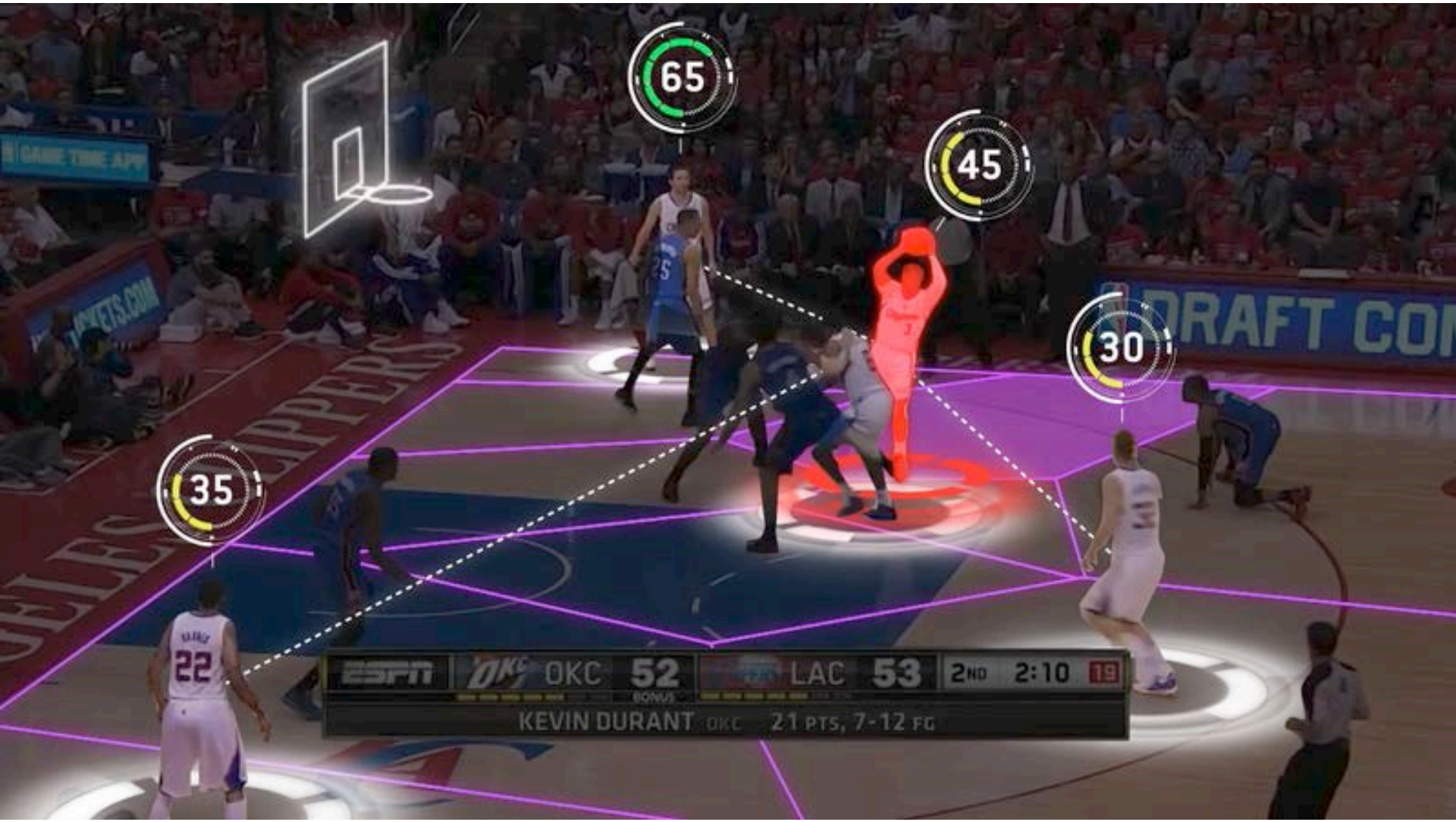
No groups of data exist
Create groups from
data close to each other

生活中的ML/DS應用: 商品推薦

生活中的ML/DS應用: 機器翻譯

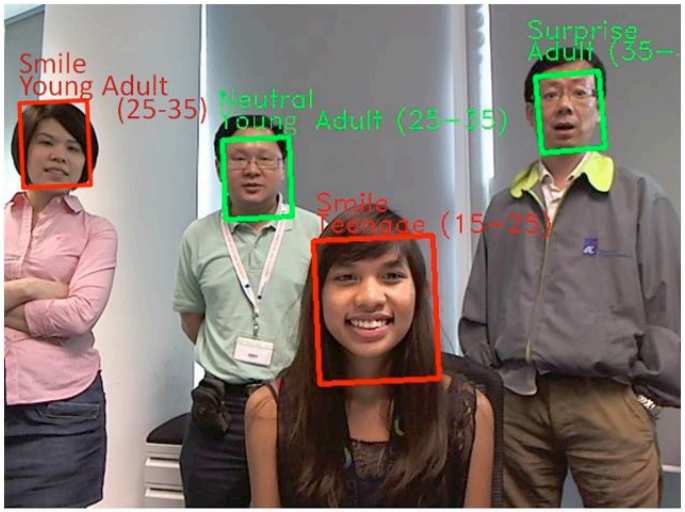


生活中的ML/DS應用: 籃球戰術分析

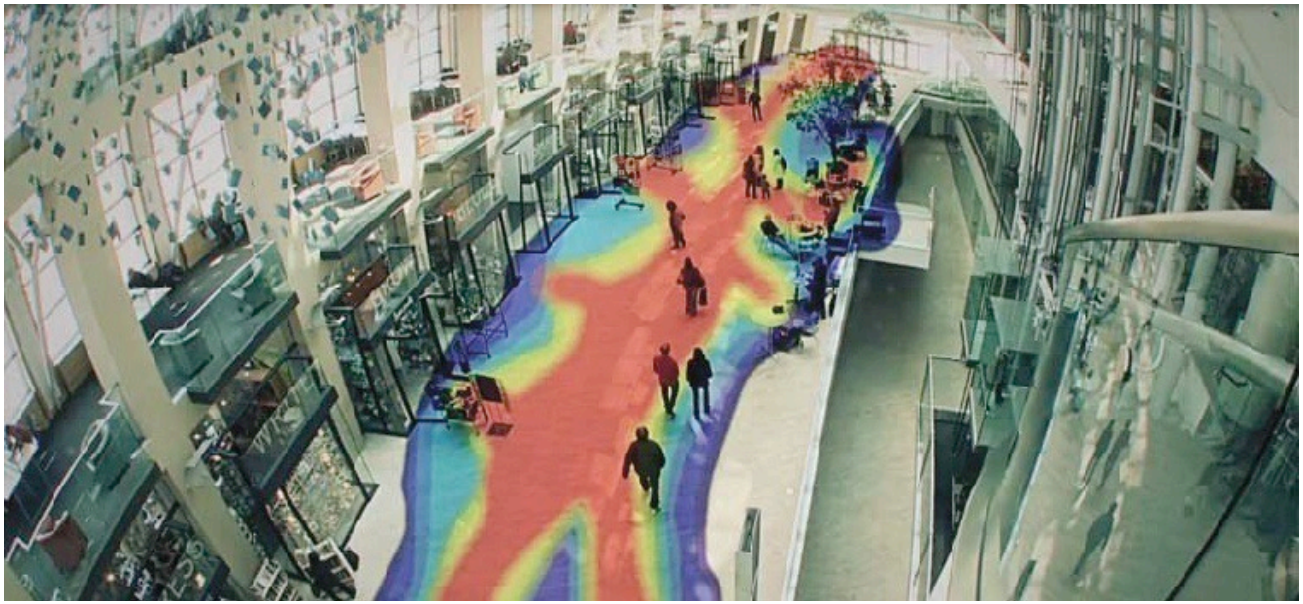


http://www.secondspectrum.com/

生活中的ML/DS應用: 人臉辨識/人流分析



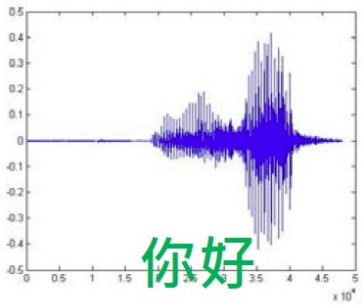
Female Male

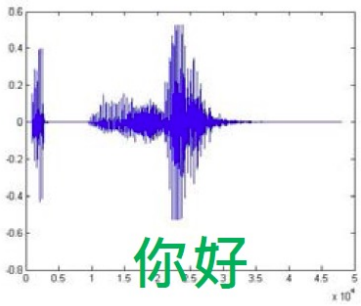


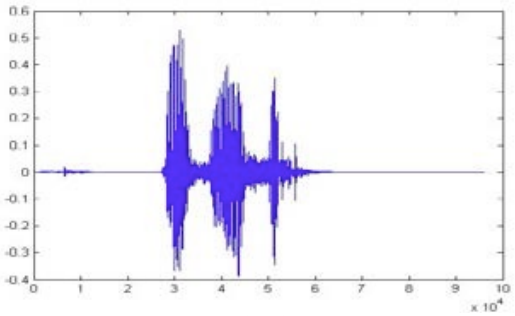
生活中的ML/DS應用: 健康管理/測謊

https://www.youtube.com/watch?v=QbXgEbeceJI

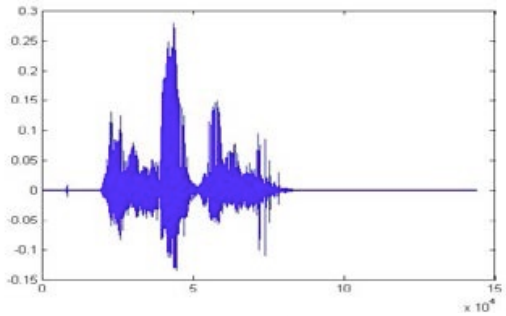
生活中的ML/DS應用: 語音辨識/聲紋辨識



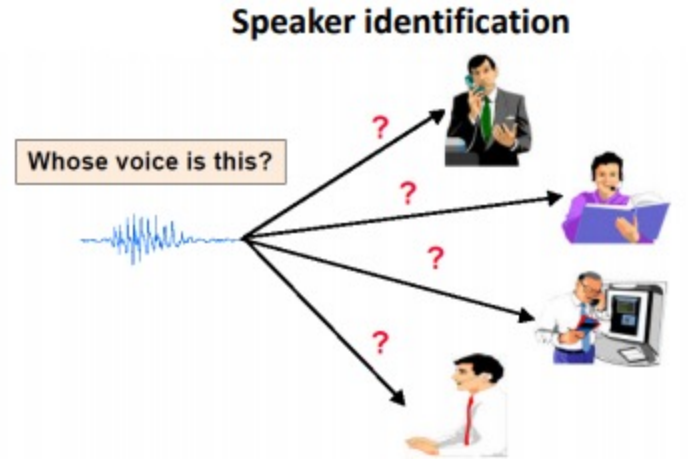




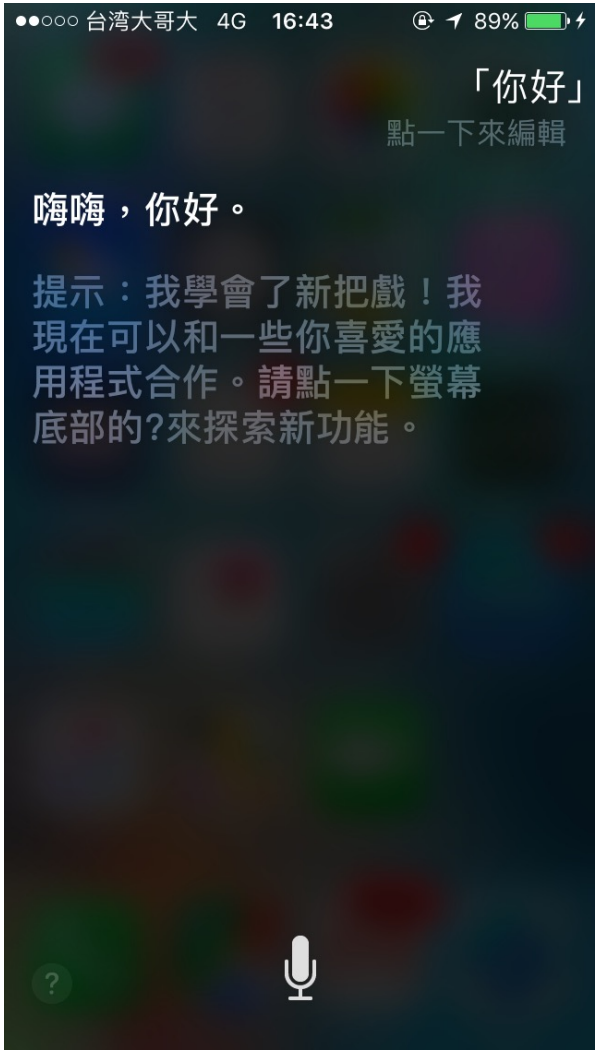
大家好



人帥真好

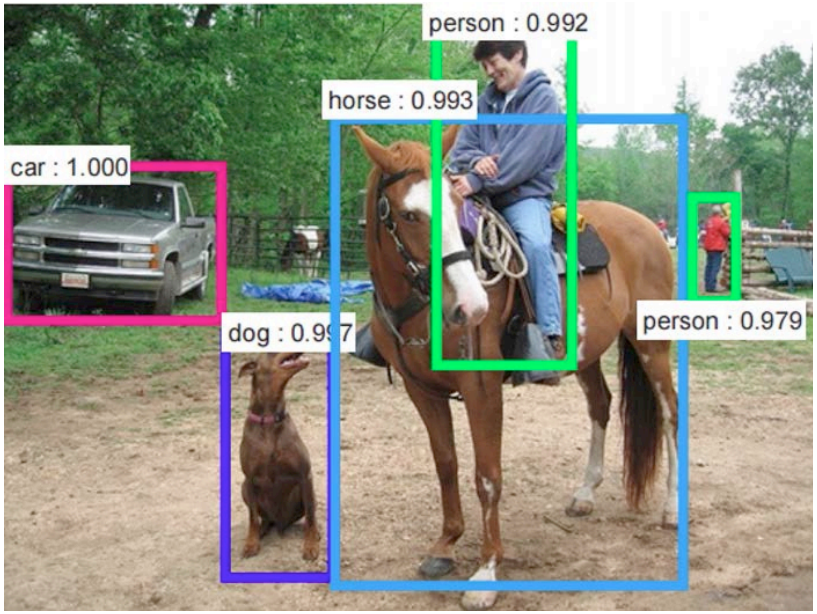


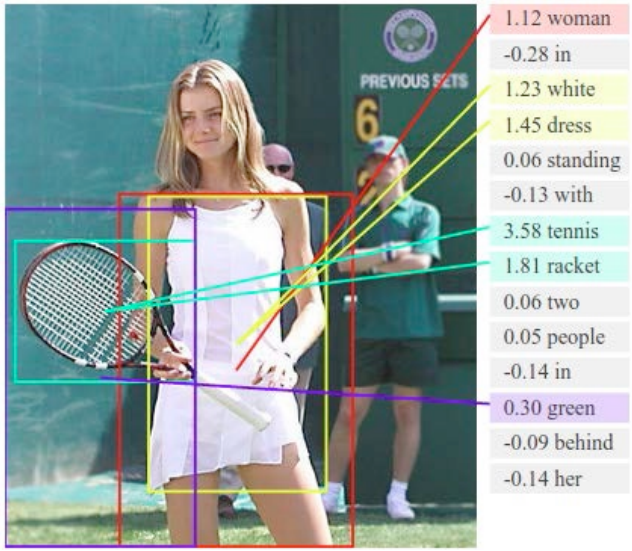
聲紋辨識



(Credit: Hung-Yi Lee)

生活中的ML/DS應用: 物件辨識/車牌辨識



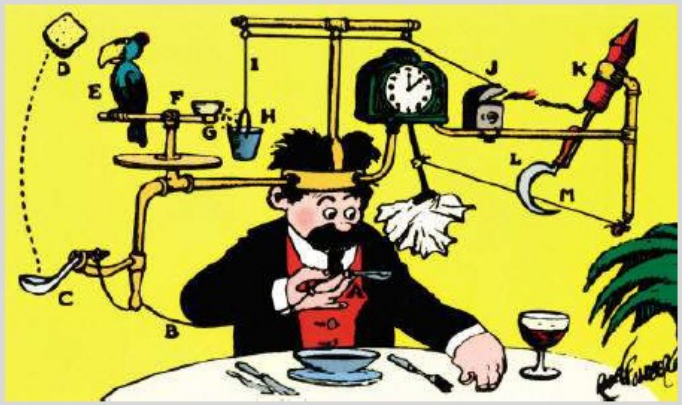


https://imagga.com/auto-tagging-demo

當代AI: [監督式]機器學習 (Machine Learning)

Input
Input
Input
Input

X



Y

Outcome
Outcome
Outcome
Outcome

Supervised Learning

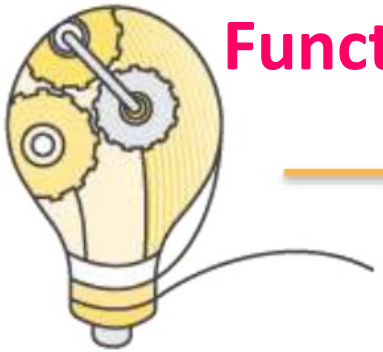
Known Historical Data

Training Data (訓練資料)

Test Data (測試資料)

Unseen Input

X'



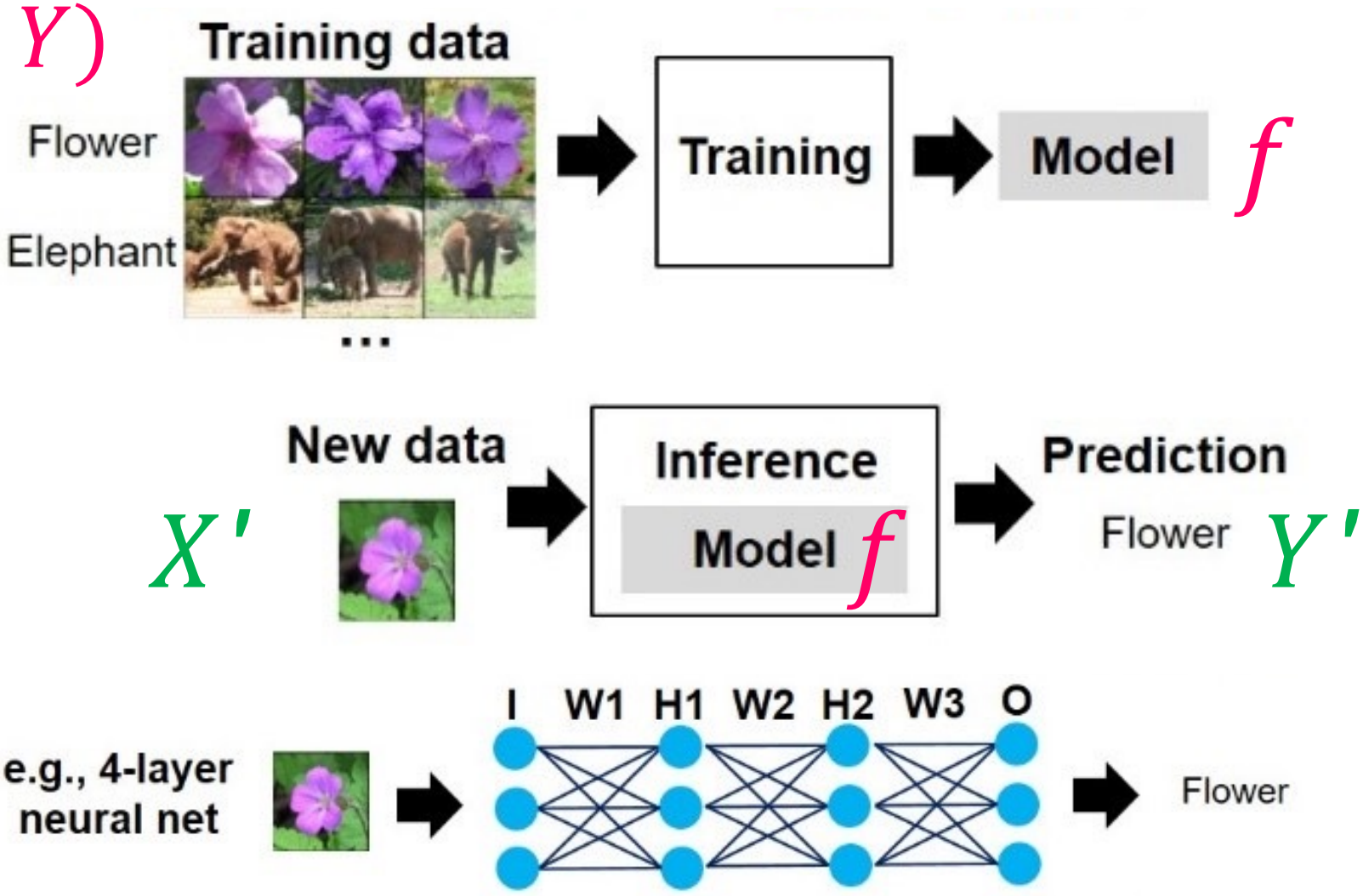
Function f: $f(X) \rightarrow Y$

Same Outcome

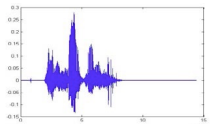
$f(X')$

AI = 從資料中[監督式]學習出函數: $f(X) \rightarrow Y$

(X, Y)

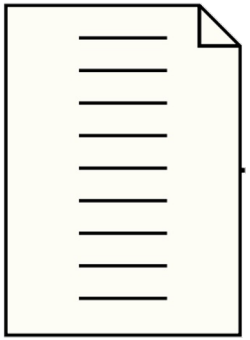


AI = 從資料中[監督式]學習出函數: $f(X) \rightarrow Y$

	學習$f(X)$函數	預測目標Y
商品推薦	f([滑鼠, 鍵盤, 硬碟])	[攝影機, 手寫板]
機器翻譯	f("你好嗎?")	"$How\ are\ you?$"
戰術分析	f(CPaul[…])	DBooker[…]
人臉辨識	f()	Female, Smile, 15-25
心率評估	f()	128 BPM
語音辨識	f()	"人帥真好"
車牌辨識	f()	9A-0265

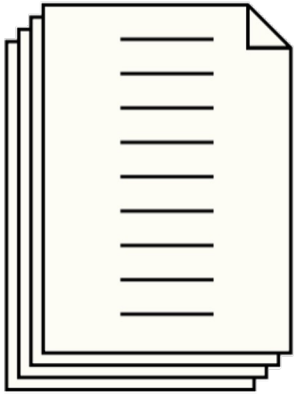
Instances and Labels

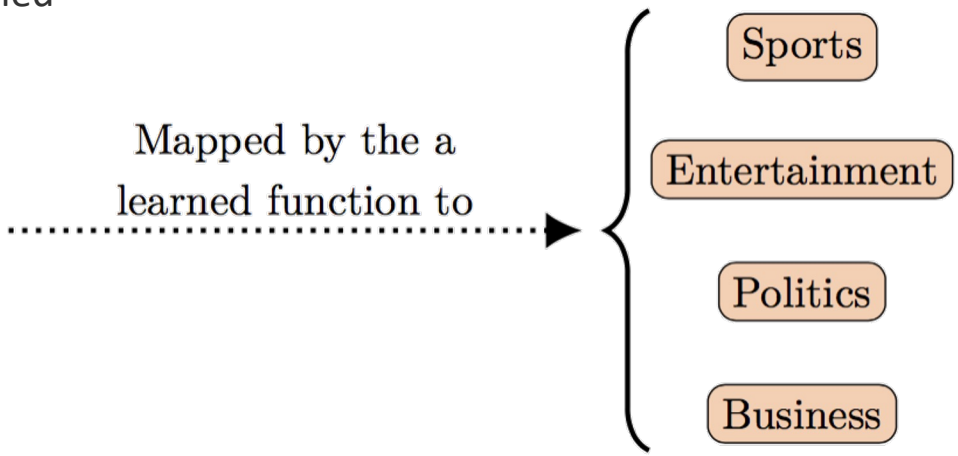
Running example: Automatically tag news articles



An instance of a news article that needs to be classified

A label





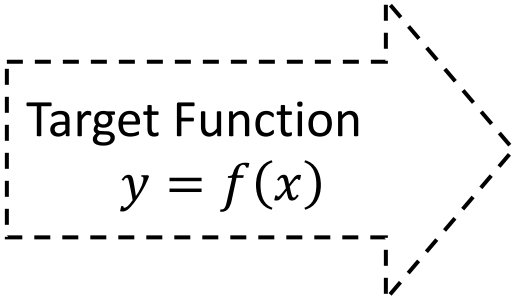
Instance Space:
All possible news articles

Label Space:
All possible labels

Instances and Labels

Feature Space
$\mathcal{X}$: **Instance Space**

The set of samples that need to be classified



$\mathcal{Y}$: **Label Space**

The set of all possible labels

The goal of learning:
Find this target function

E.g.: The set of all possible names, documents, sentences, images, emails, etc.

Learning is **search** over functions

E.g.: {Spam, Not-Spam}, {+, -}, {Sports, Political, Business, Health}

Supervised Learning

Feature Space
$\mathcal{X}$: **Instance Space**

The set of samples that need to be classified

Target Function
$y = f(x)$

$\mathcal{Y}$: **Label Space**

The set of all possible labels

Learning algorithm only sees samples generated by the function f in action

$x_1, f(x_1)$
$x_2, f(x_2)$
$x_3, f(x_3)$
$\vdots$
$x_n, f(x_n)$

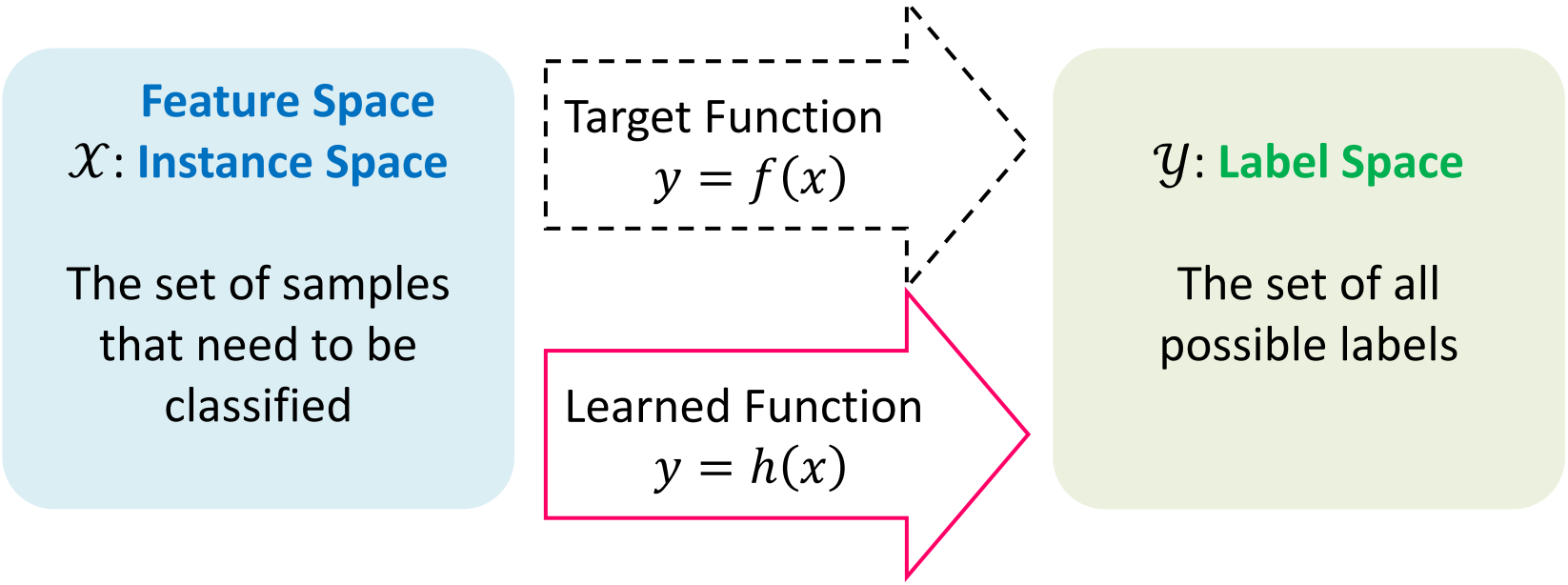
Labeled training data

→ Learning algorithm →

A learned function $h: \mathcal{X} \to \mathcal{Y}$

This is the training phase.

Supervised Learning: Evaluation



Draw *test* sample $x \in \mathcal{X}$ → $f(x)$, $h(x)$ → Are they different? How different?

Apply the model to many test samples and compare to the target's prediction
Aggregate these results to get a quality measure, e.g., **accuracy**

Can we use these test samples during training phase? → *Semi-Supervised Learning*

Supervised Learning: General Setting

Given: Training samples that are pairs of the form $(x, f(x))$

The function f is unknown

Typically the input x is represented as *feature vectors*

- Example: $x \in \{0,1\}^d$ or $x \in \mathbb{R}^d$ (d-dimensional vectors)
- A deterministic mapping from instances in your problem (e.g., news articles) to features

For a training sample $(x, f(x))$, the value of $f(x)$ is called its *label*

The goal of learning: Use the training samples to find a good approximation h for f

The label determines the kind of problem, we can have:

- **Binary classification**: label space = {0,1}
- **Multiclass classification**: label space = {1, 2, 3, ⋯, K}
- **Regression**: label space = $\mathbb{R}$

Example of Binary Classification

- Spam filtering
  - Is an email spam or not?
- Recommendation systems
  - Given user's movie preferences, will she like a new movie?
- Anomaly detection
  - Is a smartphone app malicious?
  - Is a Twitter user a bot?
- Authorship identification
  - Were these two documents written by the same person?
- Time series prediction
  - Will the future value of a stock increase or decrease with respect to its current value?

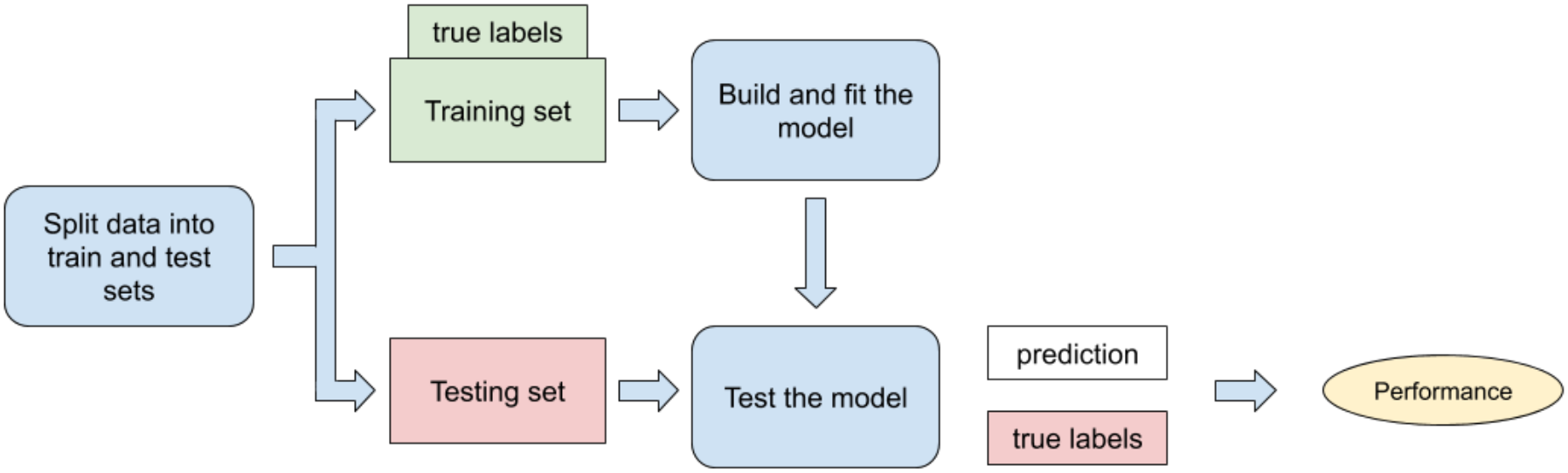
Supervised Learning

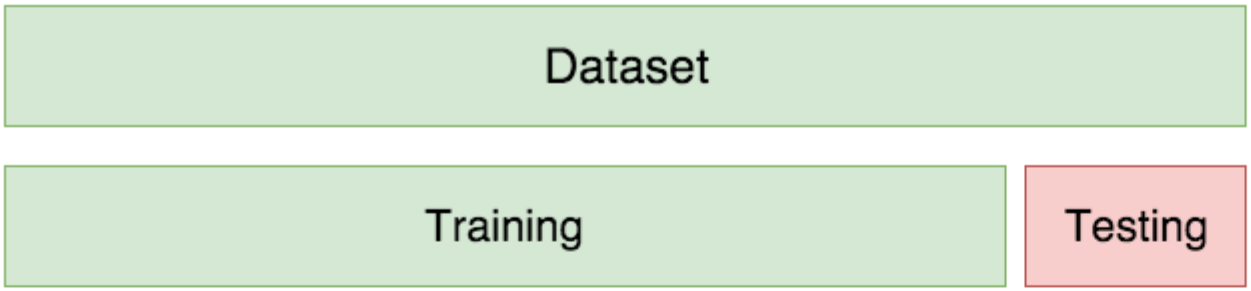
- Given: training samples $(\mathbf{x}, f(\mathbf{x}))$ generated by some unknown function f
- Find: a good approximation h to f

Example Applications

- Credit risk assessment
  - $\mathbf{x}$: attributes of customer and proposed purchase
  - $f(\mathbf{x})$: approve purchase or not
- Disease diagnosis
  - $\mathbf{x}$: attributes of patient (symptoms, lab tests)
  - $f(\mathbf{x})$: disease (or maybe, recommended therapy)
- Face recognition
  - $\mathbf{x}$: bitmap picture of person's face
  - $f(\mathbf{x})$: name of the person
- Automatic Steering
  - $\mathbf{x}$: bitmap picture of road surface in front of car
  - $f(\mathbf{x})$: degrees to turn the steering wheel

Basic Flow of Supervised Learning





On Supervised Learning

1) What is our instance/feature space?
   - What are the inputs to the problem?
   - What are the features?
2) What is our label space?
   - What is the prediction task?
3) What is our hypothesis space?
   - What functions should the learning algorithm search over?
4) What is our learning algorithm?
   - How do we learn from the labeled data?
5) What is our loss function or evaluation metric?
   - What is the goodness of a learning algorithm?
   - What is success?

Basic Flow of Supervised Learning

x: [65, 87]

Input: Features (特徵)

y: - (negative)

Output: Prediction Target

ML Training Process

各種參數組合而成的函數:
$\{h_\theta^1, h_\theta^2, h_\theta^3, \dots\}$

訓練目標: 尋找一個最能夠縮小$h_\theta(x) \rightarrow y$預測誤差的函數 h_θ^*

Training Data
$\{(x_1, y_1), (x_2, y_2), \dots\}$

找到**最佳**函數 h_θ^*

Test Data
$\{(\hat{x}, ?), \dots\}$

Predict: $h_\theta^*(\hat{x}) = \hat{y}$ → + (Positive)

訓練(training): 挑選最能夠表示從訓練資料之特徵到預測目標的最佳函數

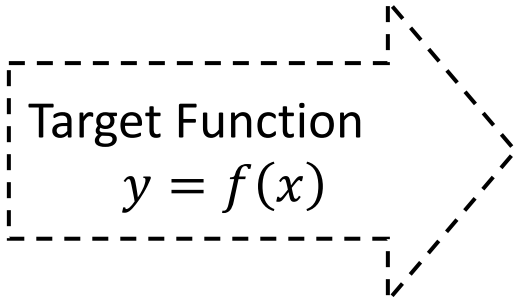
測試(test): 根據測試資料之特徵以及學到的函數，來產生預測結果

1) Instance/Feature Space $\mathcal{X}$

$\mathcal{X}$: **Instance Space**

The set of samples that need to be classified

E.g.: The set of all possible names, documents, sentences, images, emails, etc.



The goal of learning:
Find this target function

$\mathcal{Y}$: **Label Space**

The set of all possible labels

E.g.: {Spam, Not-Spam}, {+, -}, {Sports, Political, Business, Health}

- Designing an appropriate feature representation of the instance space is very crucial
- Instances $x \in \mathcal{X}$ are defined by features/attributes
- Features could be Boolean
  - Example: Does the email contain the word “free”?
- Features could be real valued
  - Example: What is the height of the person?
  - Example: What was the stock price yesterday?

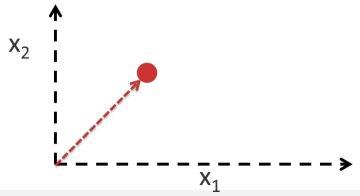
Features could be **hand-crafted** or **themselves learned**

Instances as Feature Vectors

An input to the problem (E.g., emails, names, images) → Feature function → A feature vector

Feature functions, also known as feature extractors

- Convert data samples to vectors (or a collection of attributes)
- Each $x \in \mathcal{X}$ & $x \in \mathbb{R}^d$ is a feature vector (d-dimensional space)
  - Each $x = [x_1, x_2, \cdots, x_d]$ is a point in the vector space with d dimensions
- Often deterministic, but could also be learned
- Typically thought of as **high-dimensional vectors**



* Features are supposed to capture all the information needed for a learned system to make its prediction
→ Think of them as the sensory inputs for the learned system

* Not all information about the instances is necessary or relevant
→ Bad features could even confuse a learner

Tabular Data

Somebody has defined features for you. But you can still extend it by defining your own!

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Sentiment Text Classification

Can you define features to depict texts?

Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter...I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mofette briliant! May the fourth be with you #starwarsday #starwars	Positive
Good morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

2) Label Space $\mathcal{Y}$

$\mathcal{X}$: **Instance Space**

The set of samples that need to be classified

E.g.: The set of all possible names, documents, sentences, images, emails, etc.

Target Function
$y = f(x)$

The goal of learning:
Find this target function

$\mathcal{Y}$: **Label Space**

The set of all possible labels

E.g.: {Spam, Not-Spam}, {+, -}, {Sports, Political, Business, Health}

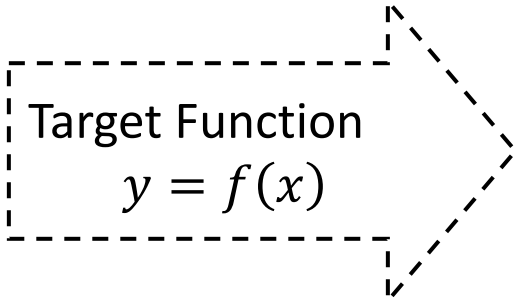
The label space depends on the nature of the problem

- Classification: labels are categorical
  - Binary classification: Two possible labels
  - Multiclass classification: K possible labels
  - Structured classification: Graph valued outputs → e.g., Machine Translation, Speech Recognition, Image Style Transfer
- Regression (numerical): label space $\mathcal{Y}$ is the set (or subset) of real numbers
- Ranking (ordinal)
  - Labels are ordinal (an ordering over the labels)
  - E.g. A Yelp 5-star review is only slightly different from a 4-star review, but very different from a 1-star review

3) Hypothesis Space

$\mathcal{X}$: **Instance Space**

The set of samples that need to be classified



$\mathcal{Y}$: **Label Space**

The set of all possible labels

The goal of learning:
Find this target function

E.g.: The set of all possible names, documents, sentences, images, emails, etc.

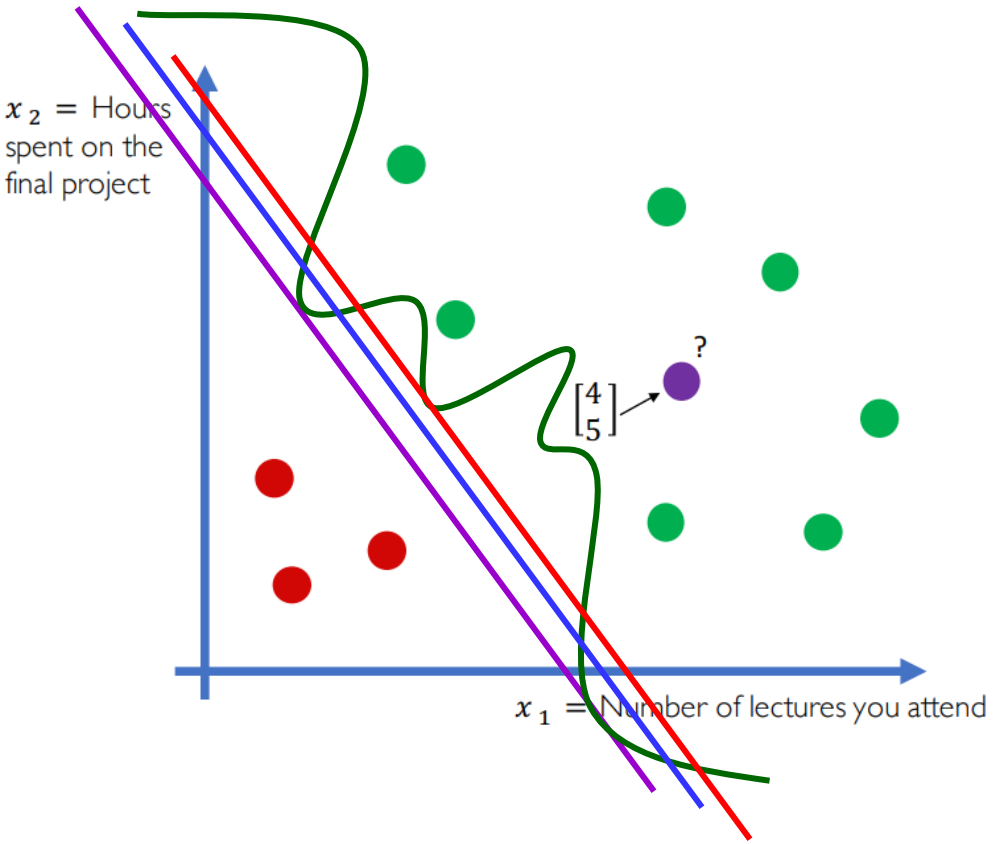
Learning is **search** over functions

E.g.: {Spam, Not-Spam}, {+, -}, {Sports, Political, Business, Health}

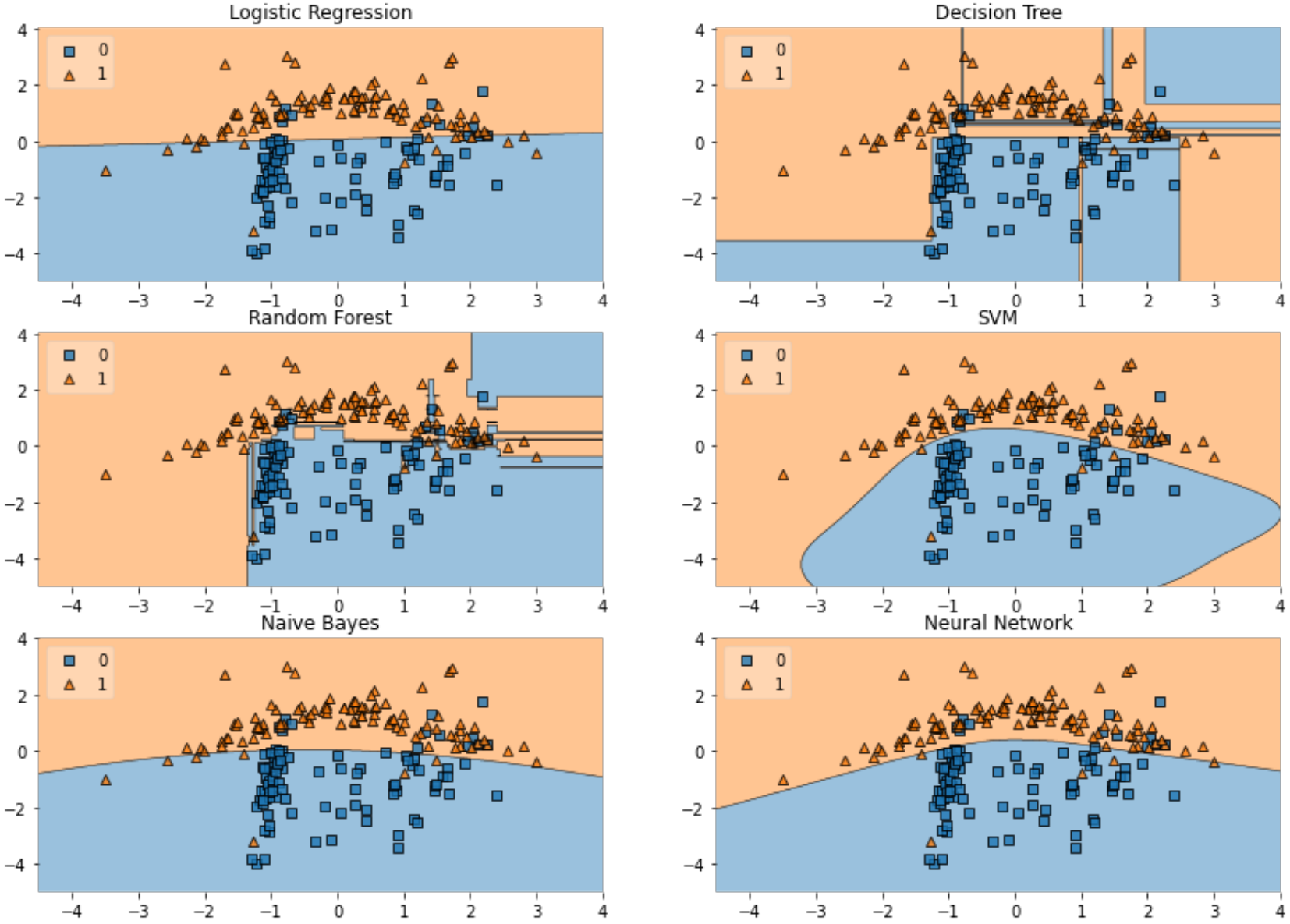
Hypothesis Space

- Hypothesis: a proposed function h **believed to be similar/approximate to target function f**
- Hypothesis space: the space of all hypotheses that can, in principle, be **output by a learning method**
- Learning is a search through hypothesis space

Learning problem: Find a function that best separates the data



Different Hypothesis Spaces



On Supervised Learning

1) What is our instance/feature space?
   - What are the inputs to the problem?
   - What are the features?
2) What is our label space?
   - What is the prediction task?
3) What is our hypothesis space?
   - What functions should the learning algorithm search over?
4) What is our learning algorithm?
   - How do we learn from the labeled data?
5) What is our loss function or evaluation metric?
   - What is the goodness of a learning algorithm?
   - What is success?

Will be introduced in this semester

List of Supervised Learning Methods

- KNN
- Naïve Bayes
- Decision Tree
- Random Forest
- Gradient Boosting ML
  - GBDT, XGBoost, CatBoost, LightGBM
- Logistic Regression
- Support Vector Machine
- Overfitting and Regularization
- MLP (Shallow Neural Network)
- Deep Neural Network