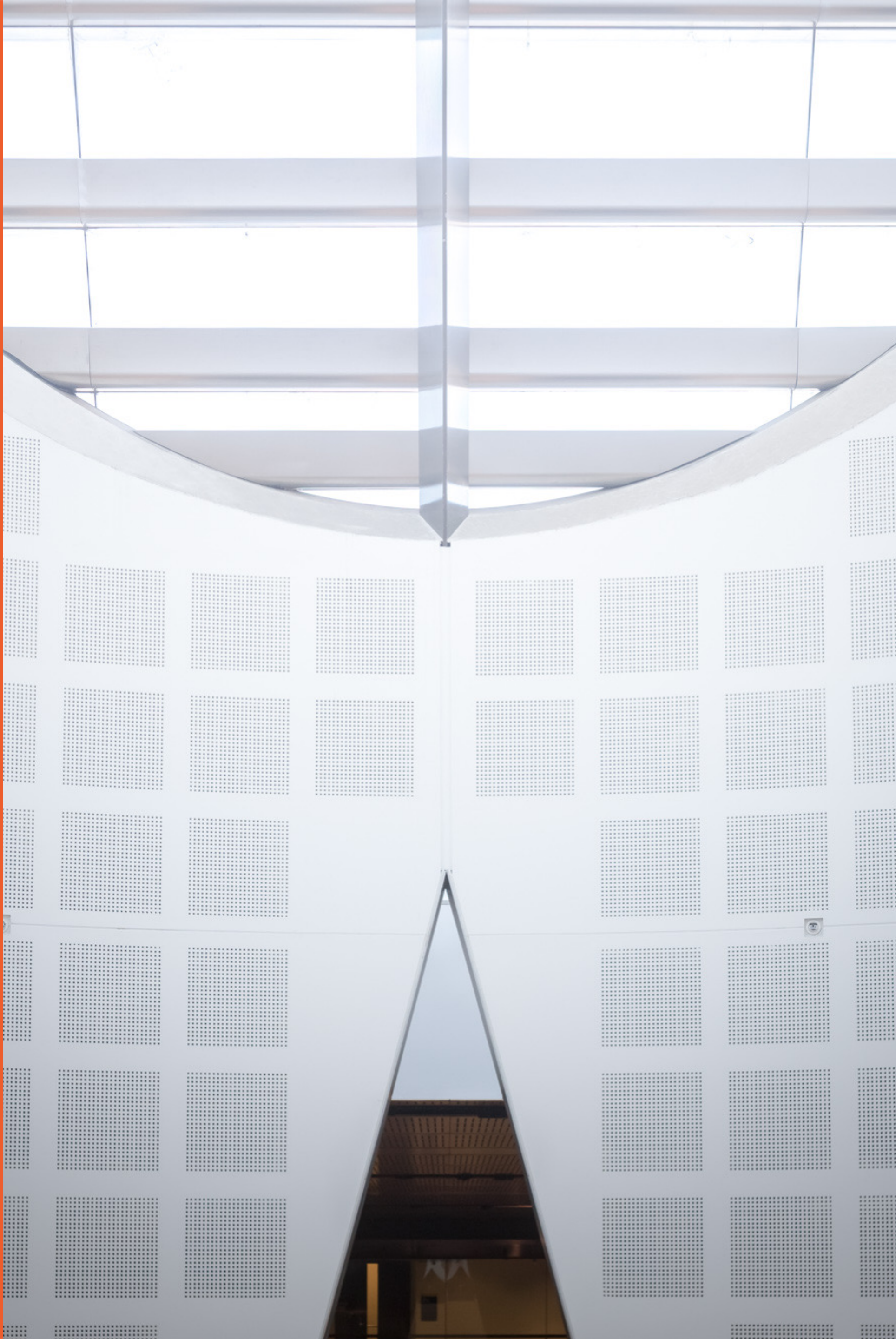
The Machine Learning Landscape

Dr Chang Xu

School of Computer Science

THE UNIVERSITY OF SYDNEY



What is Artificial Intelligence?

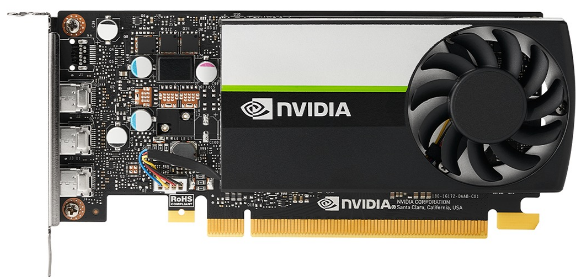
Narrow AI: The theory and development of computer systems that perform tasks that augment for human intelligence such as perceiving, classifying, learning, abstracting, reasoning, and/or acting.

General AI: Full autonomy

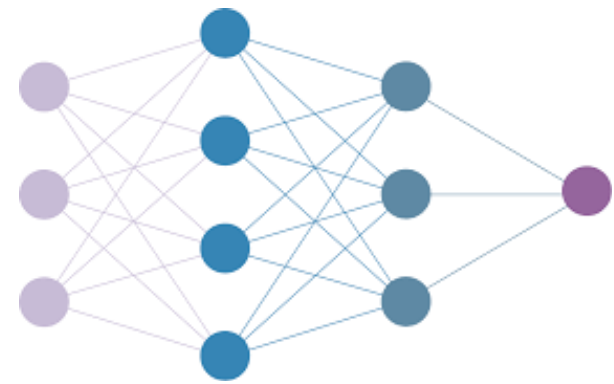
AI. Why Now?

<u>Big Data</u>

<u>Compute Power</u>

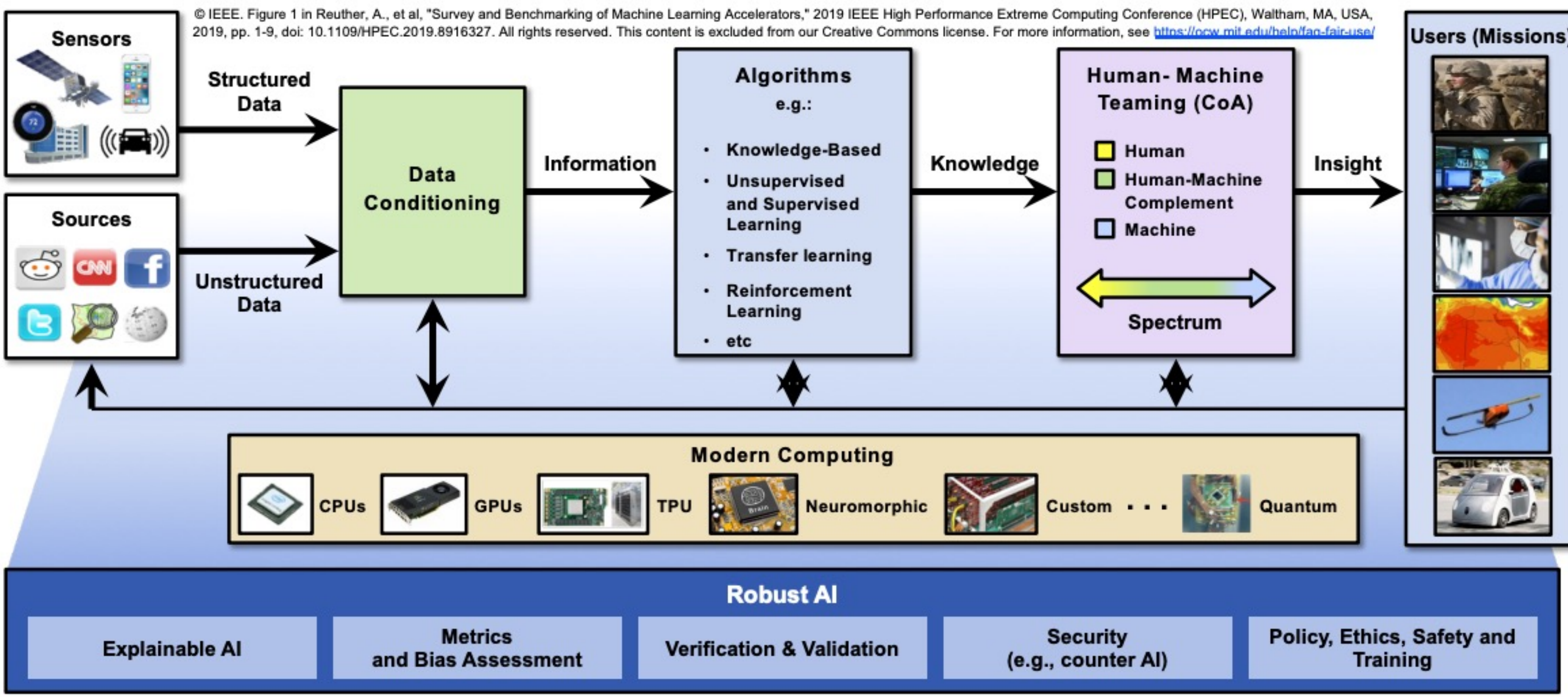


<u>Machine Learning Algorithms</u>

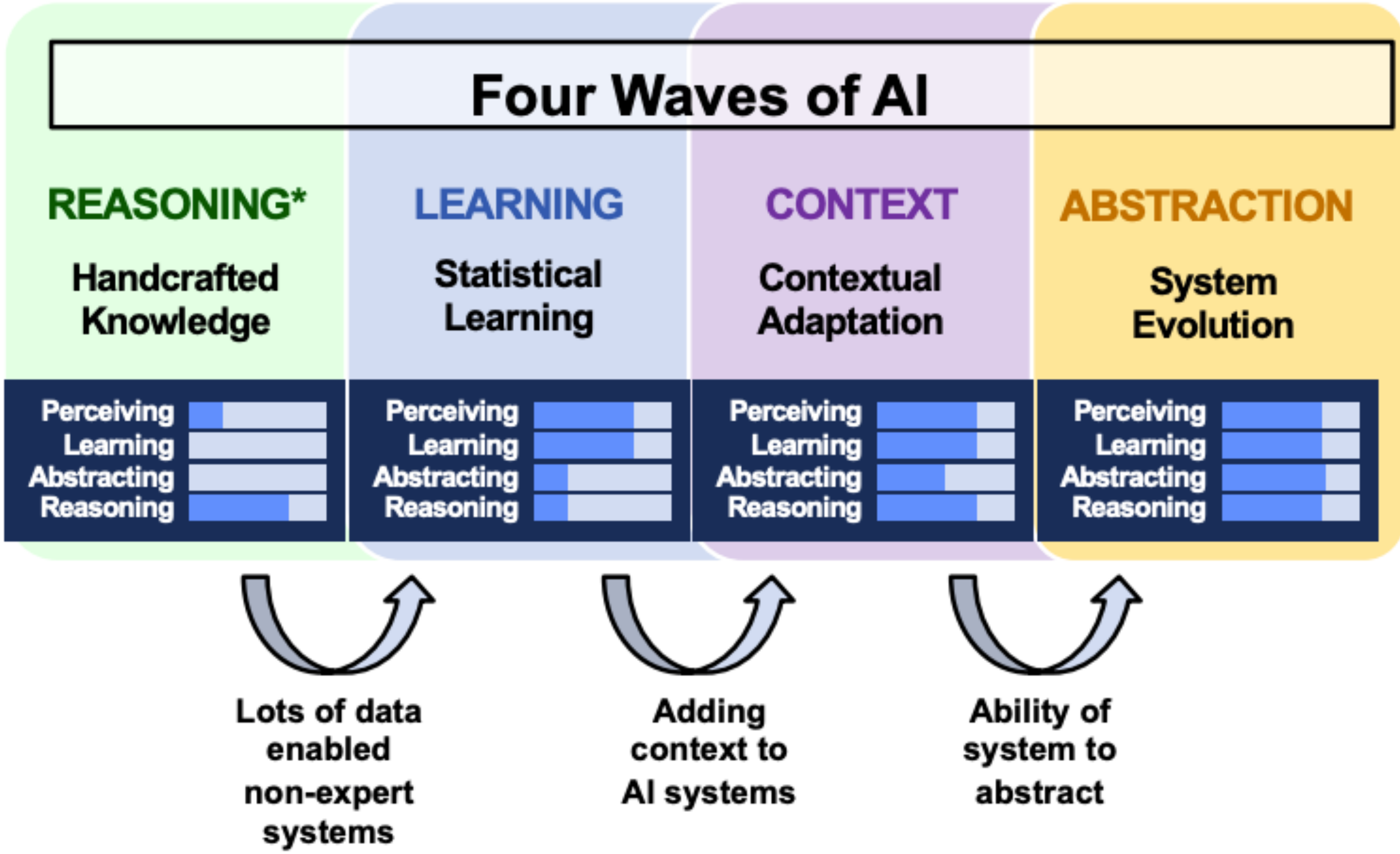


Convergence of High Performance Computing, Big Data and Algorithms that enable widespread AI development

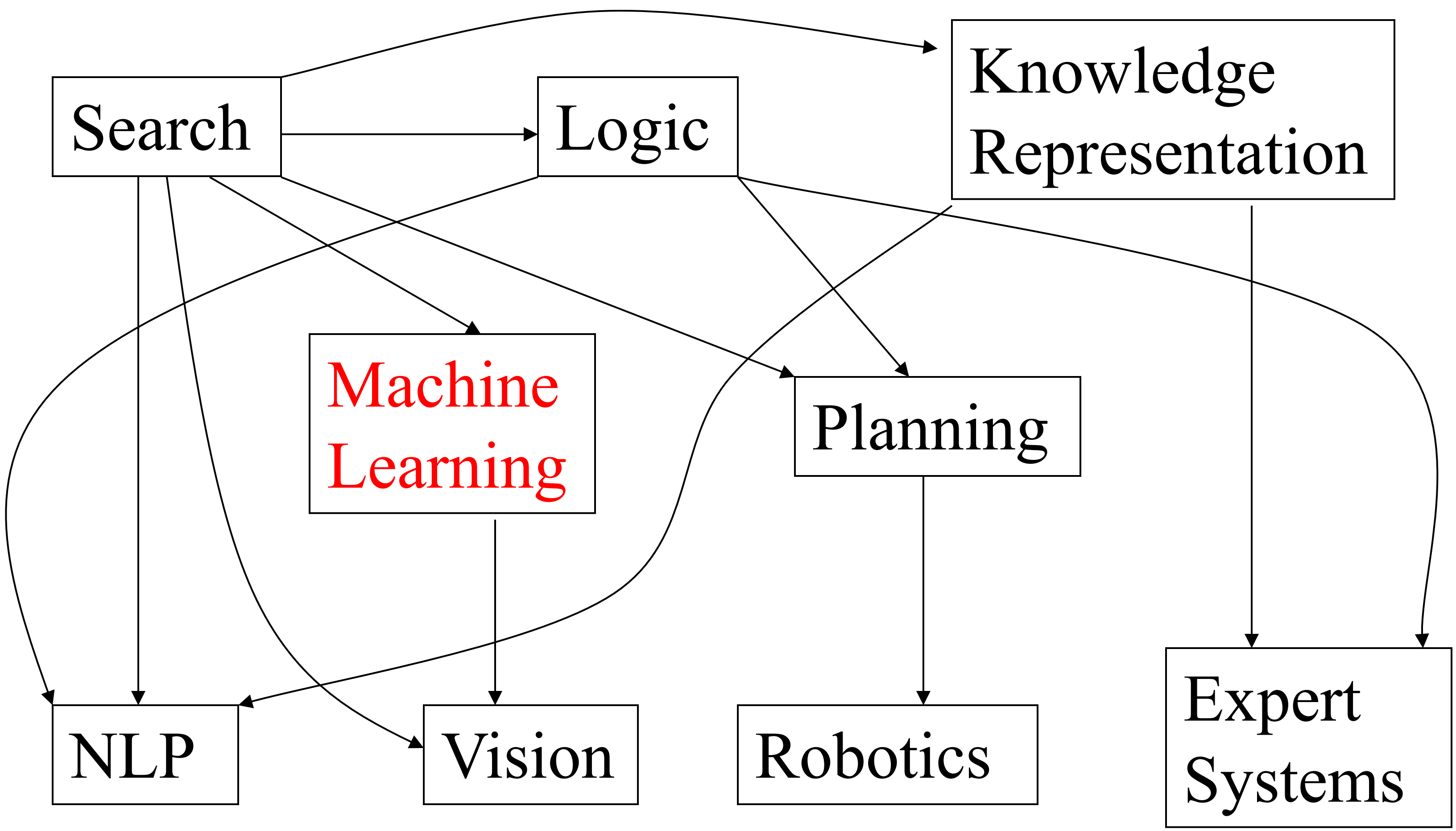
AI Canonical Architecture



Artificial Intelligence Evolution



Areas of AI and Some Dependencies



What is Machine Learning?

Machine Learning

– Study of algorithms that improve their performance at some task with experience (data)

– Optimize based on performance criterion using example data or past experience

• Combination of techniques from statistics, computer science communities

• Getting computers to program themselves

• Common tasks:

– Classification

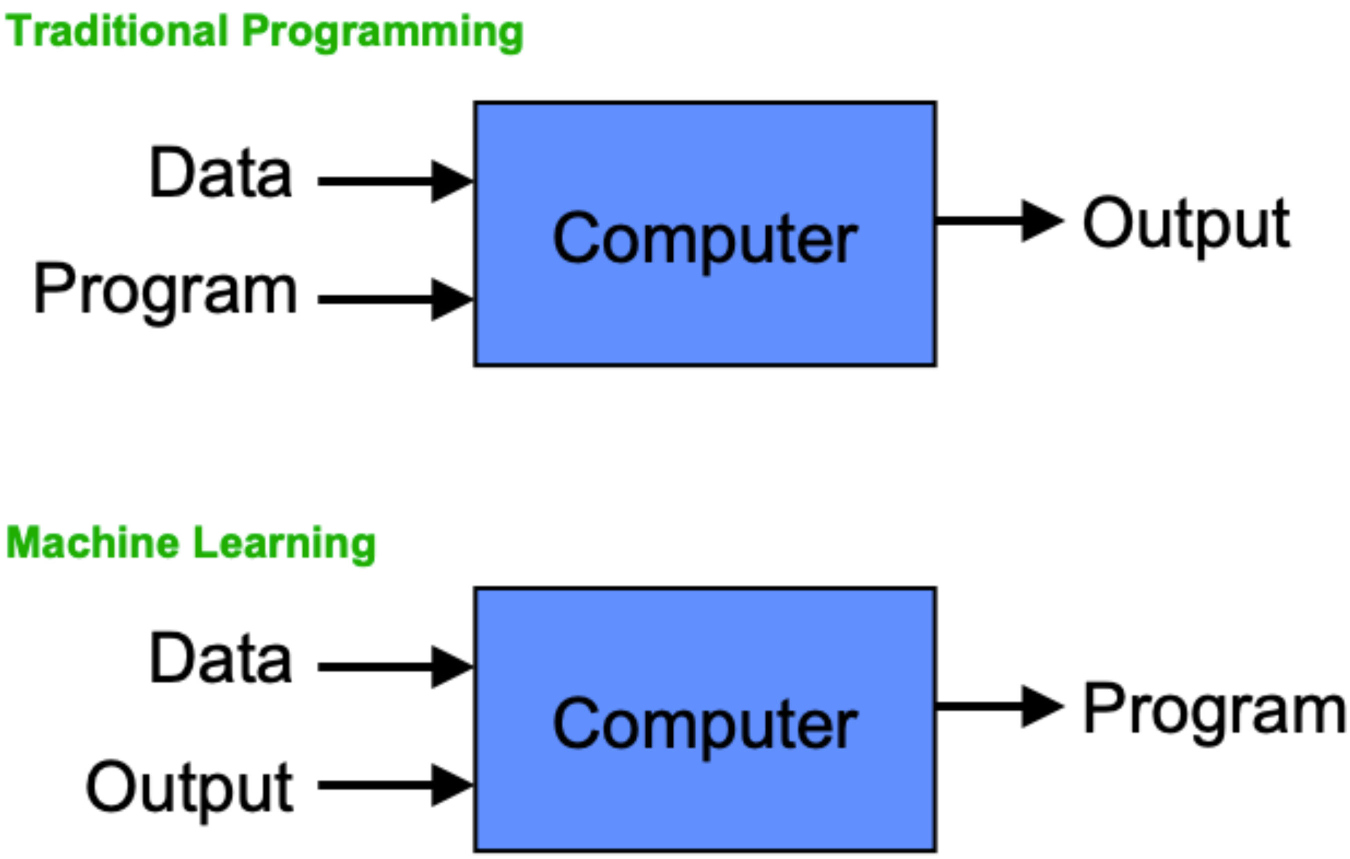
– Regression

– Prediction

– Clustering

– …

Traditional Programming vs. Machine Learning



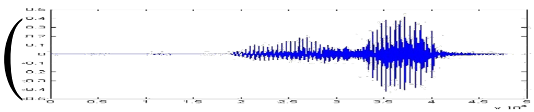
What is Machine Learning?

Informally, machine learning is a function that we do not know the exact formation, but we can learn or find a way to link the input data and corresponding targets.

We want to find a model that can link new input and accurate outputs

Machine Learning
≈ Looking for Function

– Speech Recognition

$f($  $;\theta) =$ "How are you"

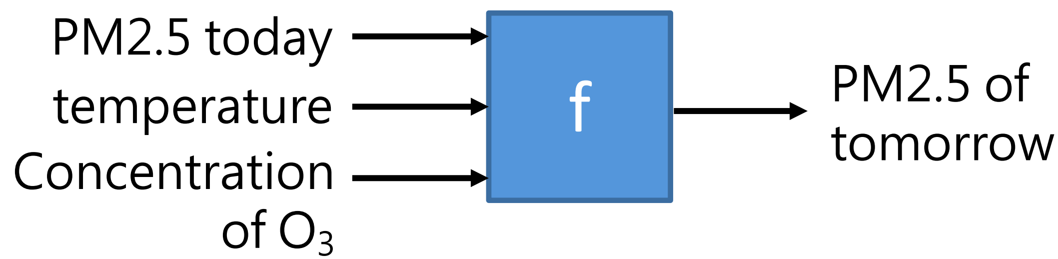
– Image Recognition

$f($  $;\theta) =$ "Cat"

Different types of Functions

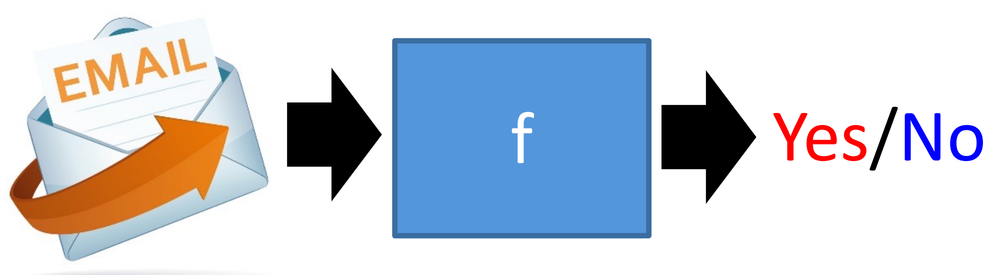
Regression: The function outputs a scalar.

Predict PM2.5



Classification: Given options (**classes**), the function outputs the correct one.

Spam filtering



Data

Data is Critical To Breakthroughs in AI

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level read-speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogleNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolutional neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		**3 years**	**18 years**

AIGC

（AI Generated Content）

a photo of an astronaut riding a horse on mars

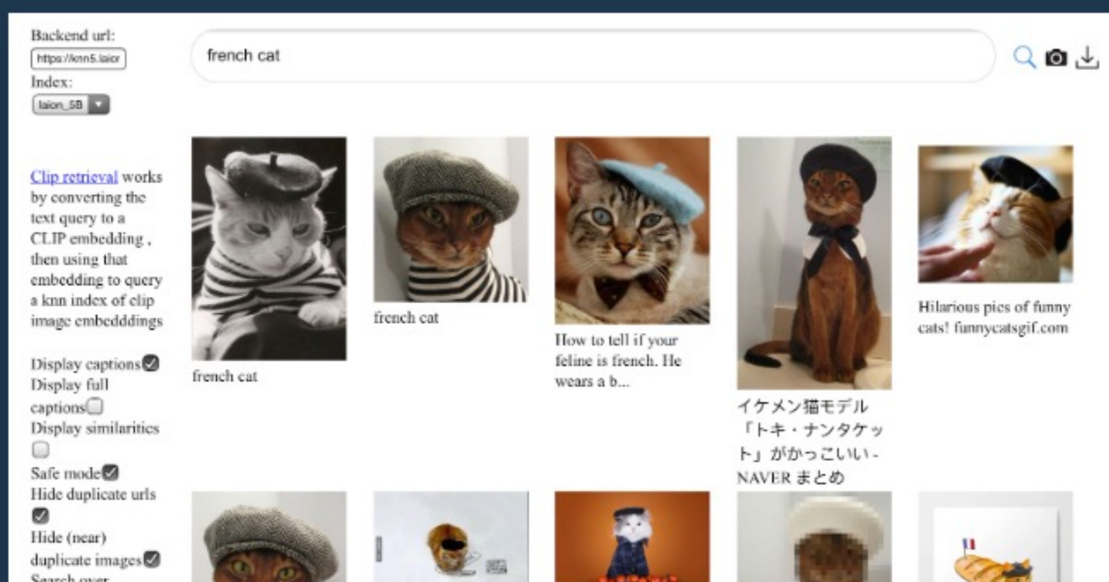
https://replicate.com/cjwbw/stable-diffusion-v2

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 31 Mar, 2022

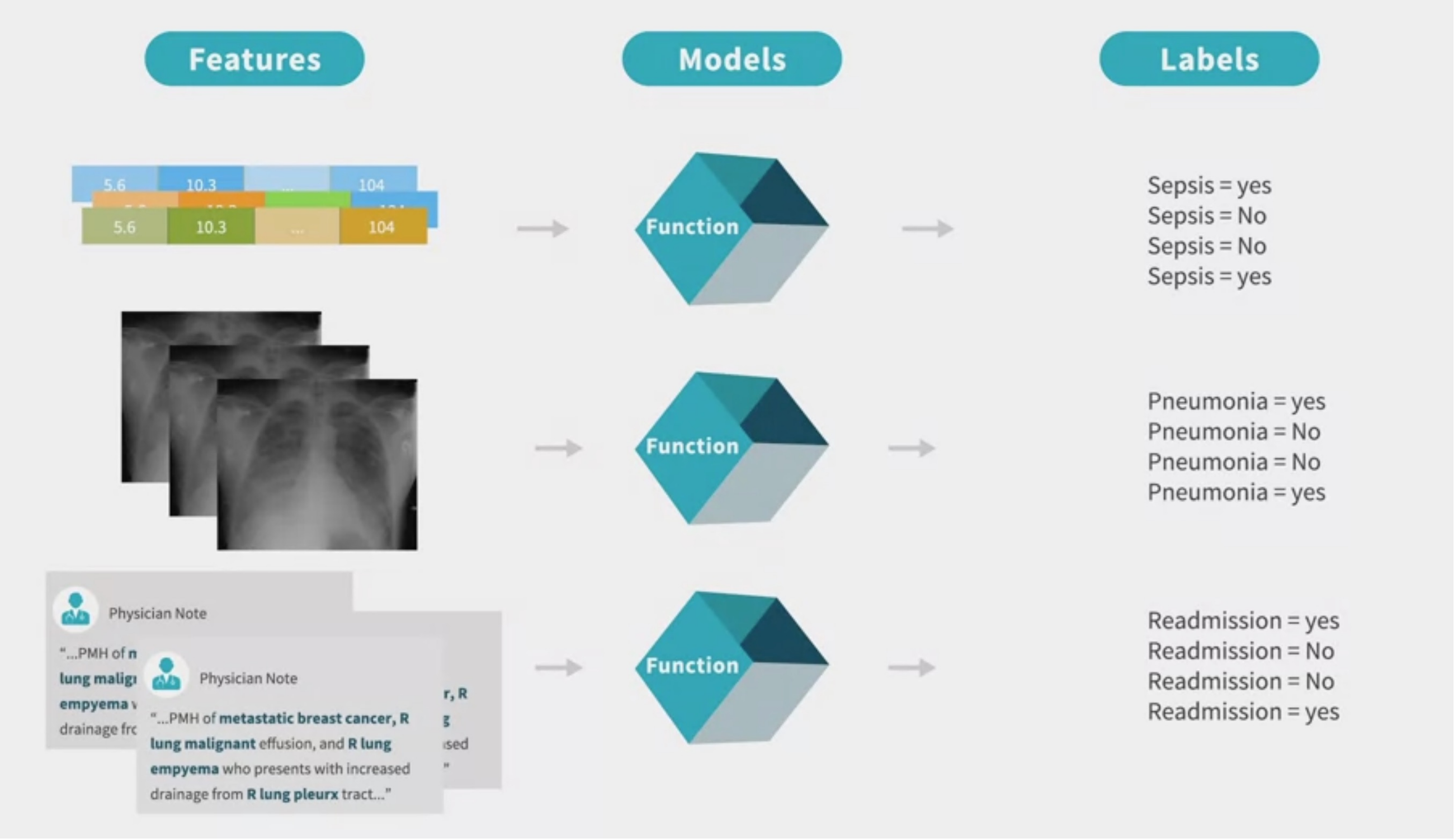
We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world - see also our NeurIPS2022 paper

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev



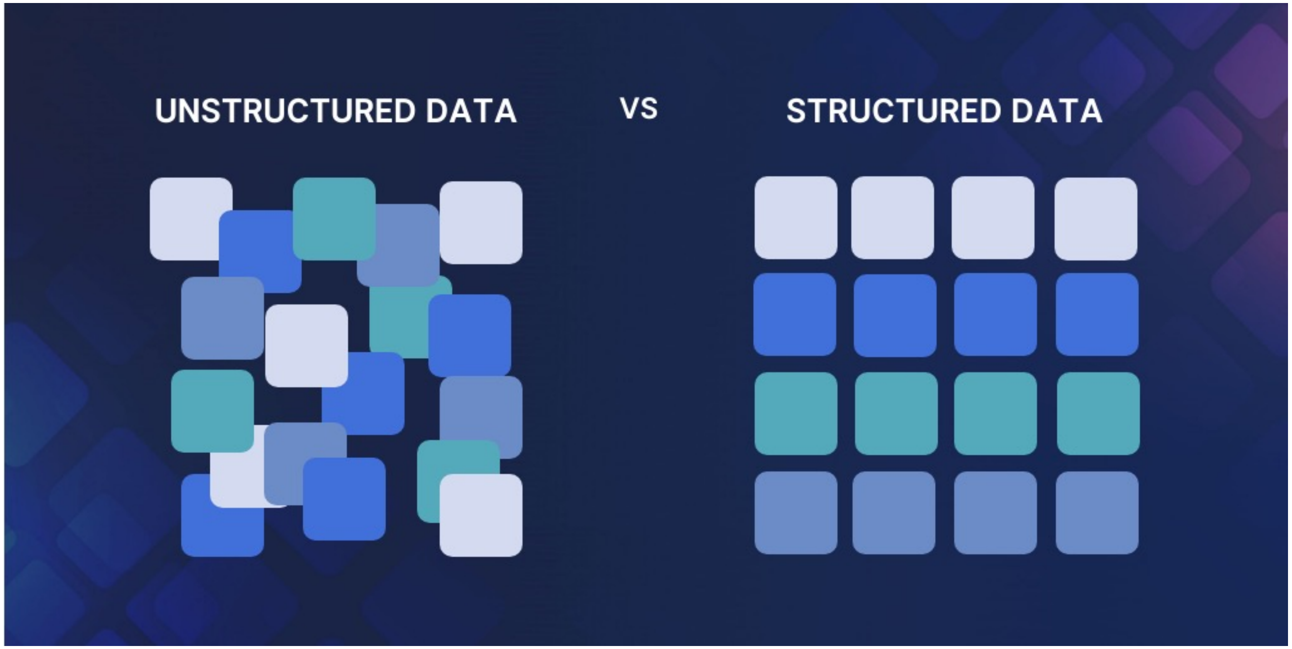
Different Data Types in Healthcare

- There are many different types of data in Healthcare. Most of them can be used to learning a machine learning model.



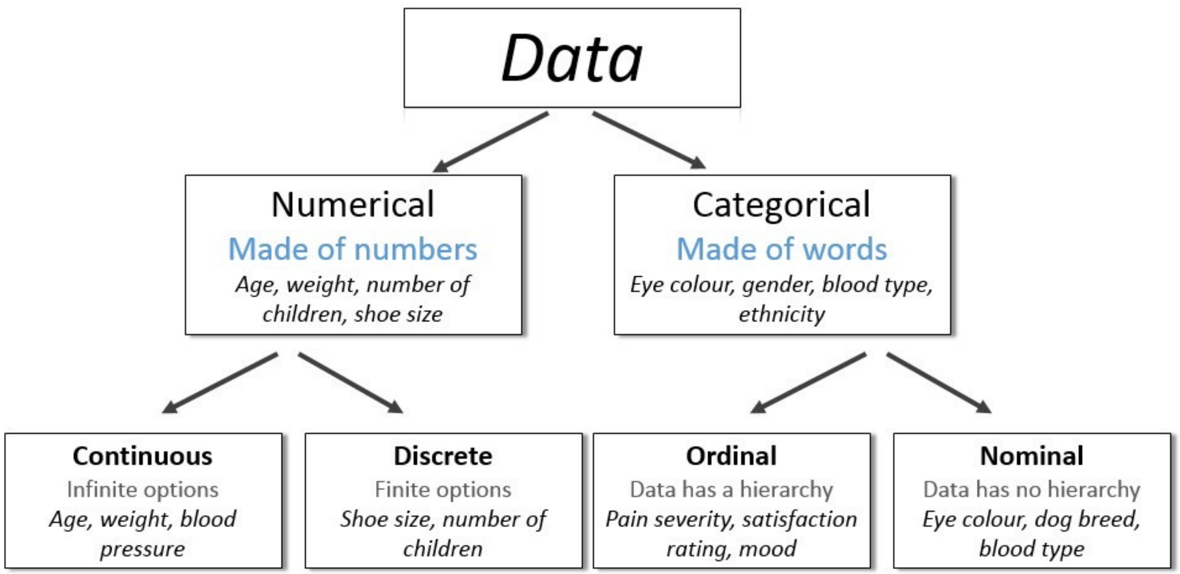
Different Data Types in Healthcare

- Typically, data are divided into 2 groups
  - Structured data: can be easily stored in a data table
  - Unstructured data: what you find in the wild (text, images, signal, video)



Different Data Types in Healthcare

- Typically, data are divided into 2 groups
  - Structured data: can be easily stored in a data table
    - Numerical data such as blood pressure, weight, height
    - Categorical data: gender, blood type, pain severity
  - Unstructured data: what you find in the wild (text, images, signal, video)



Different Data Types in Healthcare

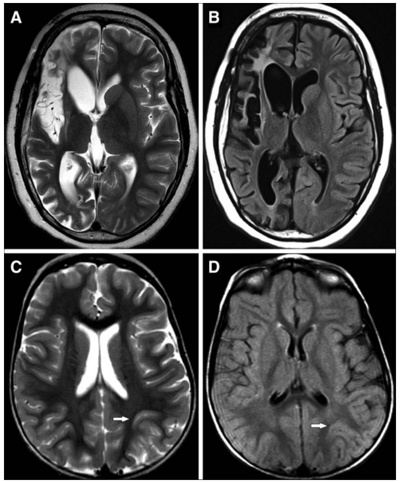
- A typical structured dataset in healthcare
  - ***Diagnosis*** column is called **targets/labels**
  - ***Other*** columns are called **features**
  - Each row is call one **data example**

id	diagnosis	# radius_mean	# texture_m...	# perimeter_...	# area_mean
842302	M	17.99	10.38	122.8	1001
842517	M	20.57	17.77	132.9	1326
84300903	M	19.69	21.25	130	1203
84348301	M	11.42	20.38	77.58	386.1
84358402	M	20.29	14.34	135.1	1297
843786	M	12.45	15.7	82.57	477.1
844359	M	18.25	19.98	119.6	1040
84458202	M	13.71	20.83	90.2	577.9
844981	M	13	21.82	87.5	519.8
84501001	M	12.46	24.04	83.97	475.9

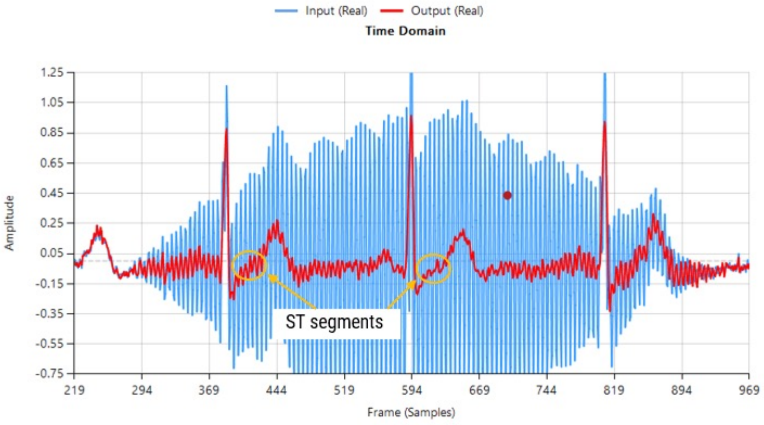
Breast Cancer Wisconsin (Diagnostic) Data Set

Different Data Types in Healthcare

- Typically, data are divided into 2 groups
  - Structured data: can be easily stored in a data table
  - Unstructured data: what you find in the wild
    - biomedical images
    - biomedical signal
    - clinical text
    - …..



MCI Scan



ECG Signal

CHIEF COMPLAINT: Ms. NAME is a AGE-year-old woman who presents with newly diagnosed stage IV metastatic non-small cell lung carcinoma here for further treatment options.

HISTORY OF PRESENT ILLNESS: Here today for evaluation. She developed dyspnea and was found to have a right sided pleural effusion on chest x-ray. Thoracentesis cytology was indicative of malignant cells consistent with adenocarcinoma.

She underwent a CT scan of the chest that demonstrated a left lower lung nodule measuring 1.2cm. A CT scan of the abdomen and pelvis was negative in detail.

PAST MEDICAL HISTORY: Hypertension, kidney stones. Breast lump removed DATE. Hysterectomy DATE, cesarean section DATE. Right leg surgery after an accident. Hyperlipidemia.

SOCIAL HISTORY: No history of alcohol or tobacco use. Patient lives alone in Manhattan. She has two adult children who live nearby. She works at a law firm.

FAMILY HISTORY: No family history in first-degree relatives. History of esophageal cancer in aunt, melanoma in uncle. Father died of heart attack at AGE.

REVIEW OF SYSTEMS: She denies weight loss, headaches, dizziness. Has dyspnea on exertion but no SOB at rest. No changes in urination pattern.

PHYSICAL EXAMINATION:

GENERAL: In no acute distress, well appearing.
LUNGS: Decreased breath sounds on the left.
ABDOMEN: Soft, nontender. Bowel sounds present.

IMPRESSION & PLAN: The patient is a AGE-year-old woman without history of smoking now with recently diagnosed stage IV non-small cell lung adenocarcinoma with malignant pleural effusion. During this visit we discussed the extent of disease and role of palliative systemic chemotherapy. Her cancer is incurable. I suggest that she undergo a biopsy of the left lower lung mass to obtain more tissue for EGFR mutational testing. She may be a candidate for targeted oral therapy. If not found to have EGFR mutation then a platinum doublet may be used. She will return after the biopsy to discuss final treatment options. In general, we also discussed the prognosis related to stage IV non-small cell lung carcinoma.

Clinical Text

Machine learning Pipeline

Let’s create a dataset!

https://forms.gle/XnpMNYrv1eFisEePA

Framework of ML

Training data: $\left\{\left(\boldsymbol{x}^{\mathbf{1}}, y^{1}\right), \left(\boldsymbol{x}^{\mathbf{2}}, y^{2}\right), \ldots, \left(\boldsymbol{x}^{\boldsymbol{N}}, y^{N}\right)\right\}$

Testing data: $\left\{\boldsymbol{x}^{\boldsymbol{N+1}}, \boldsymbol{x}^{\boldsymbol{N+2}}, \ldots, \boldsymbol{x}^{\boldsymbol{N+M}}\right\}$

BMI (body mass index) prediction

$\boldsymbol{x}$: $\rightarrow$ y

[Height, Weight, Gender] BMI

Framework of ML

Training data: $\{(\boldsymbol{x}^1, y^1), (\boldsymbol{x}^2, y^2), \ldots, (\boldsymbol{x}^N, y^N)\}$

Training:

- Step 1: function with unknown — $y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$
- Step 2: define loss from training data — $L(\boldsymbol{\theta})$
- Step 3: optimization — $\boldsymbol{\theta}^* = arg \min_{\boldsymbol{\theta}} L$

Testing data: $\{\boldsymbol{x}^{N+1}, \boldsymbol{x}^{N+2}, \ldots, \boldsymbol{x}^{N+M}\}$

Use $y = f_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ to label the testing data

$\{y^{N+1}, y^{N+2}, \ldots, y^{N+M}\} \Rightarrow$?

1. Function with Unknown Parameters

$y = f($  $;\theta)$

$(\boldsymbol{w}; b)$

Model $\hat{y} = b + \boldsymbol{w}^T \boldsymbol{x}$

$\hat{y}$: predicted BMI value
$\boldsymbol{x}$: individual feature vector representation

$\boldsymbol{w}$ (weight vector) and b (bias) are unknown parameters (learned from data)

2. Define Loss from Training Data

- Loss: how good a set of values is. Loss: $L = \frac{1}{N}\sum_n e_n$

$e = |y - \hat{y}|$ L is mean absolute error (MAE)

$e = (y - \hat{y})^2$ L is mean square error (MSE)

$\widehat{y_b} = f($  $;\theta)$ 18.2 19.4 y_b

$\widehat{y_r} = f($  $;\theta)$ 20.3 24.3 y_r

$\widehat{y_g} = f($  $;\theta)$ 30.6 28.2 y_g

Prediction Ground-truth label

Source of image: http://chico386.pixnet.net/album/photo/171572850

3. Optimization

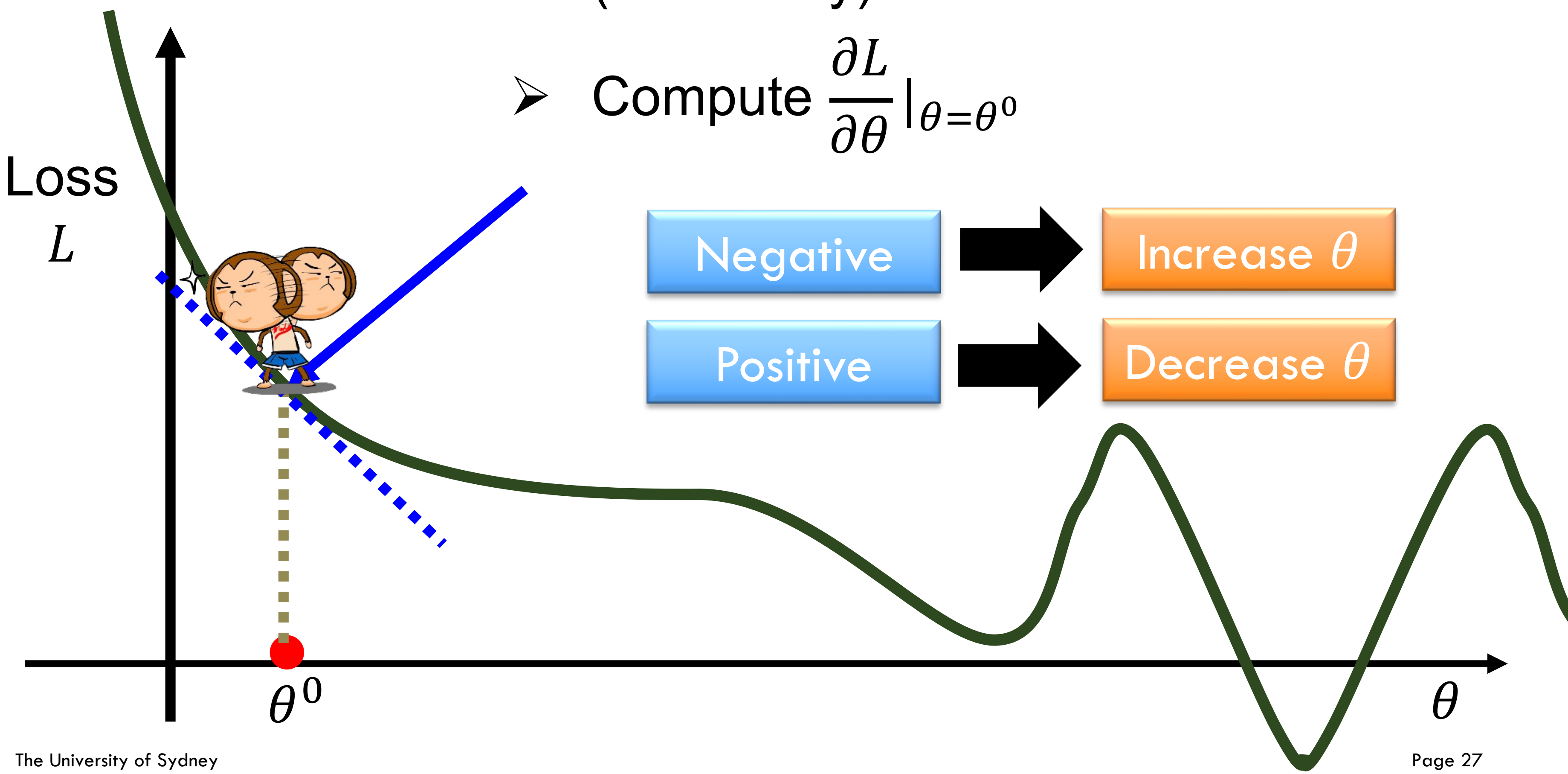
$$\theta^* = arg\min_{\theta} L$$

Gradient Descent

- (Randomly) Pick an initial value θ^0
- Compute $\frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$

Negative → Increase θ

Positive → Decrease θ



Source of image: http://chico386.pixnet.net/album/photo/171572850

3. Optimization

$$\theta^* = arg \min_{\theta} L$$

Gradient Descent

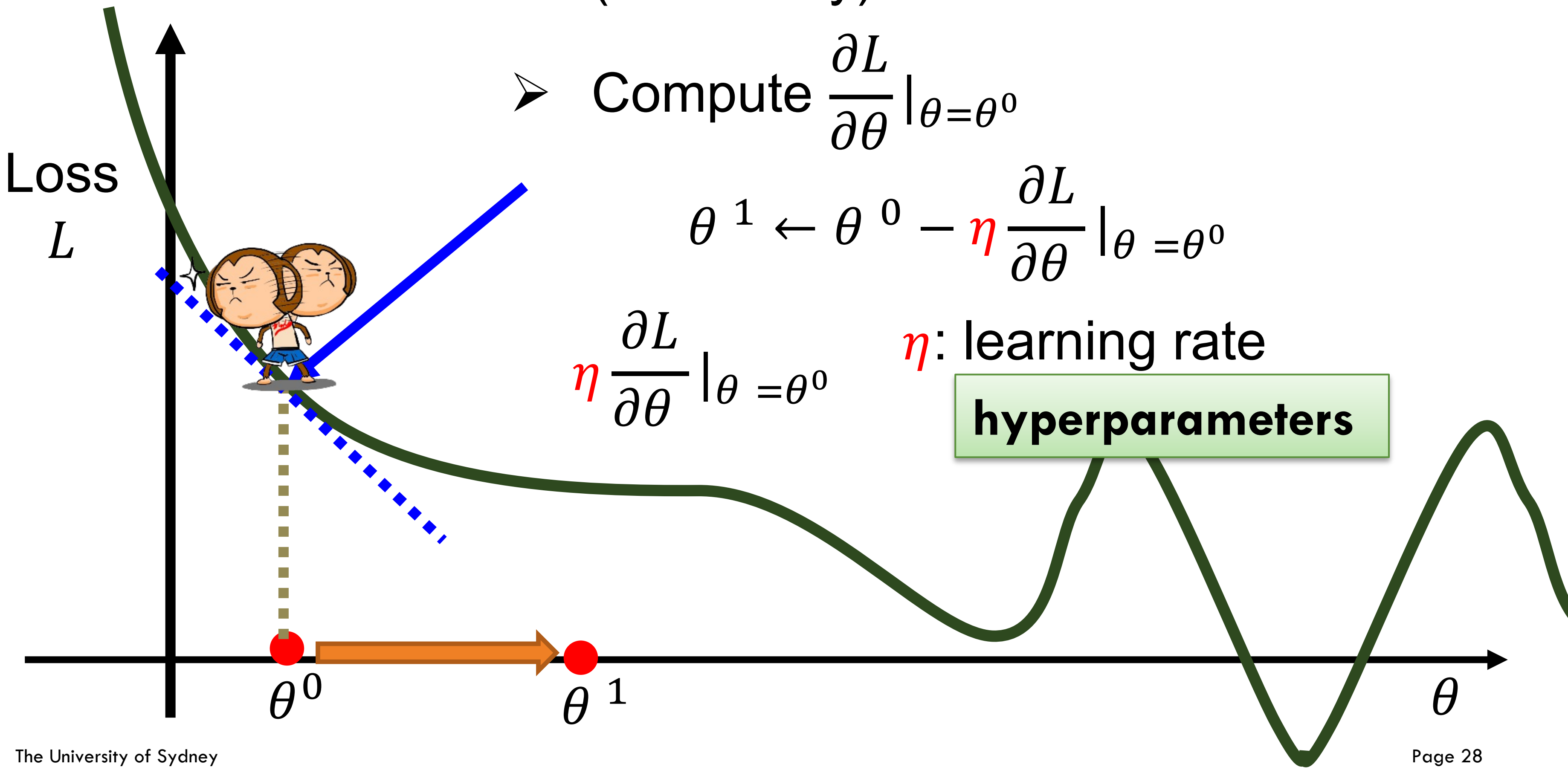
- (Randomly) Pick an initial value θ^0
- Compute $\frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$

$$\theta^1 \leftarrow \theta^0 - \eta \frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$$

$$\eta \frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$$

η: learning rate

hyperparameters



Source of image: http://chico386.pixnet.net/album/photo/171572850

3. Optimization

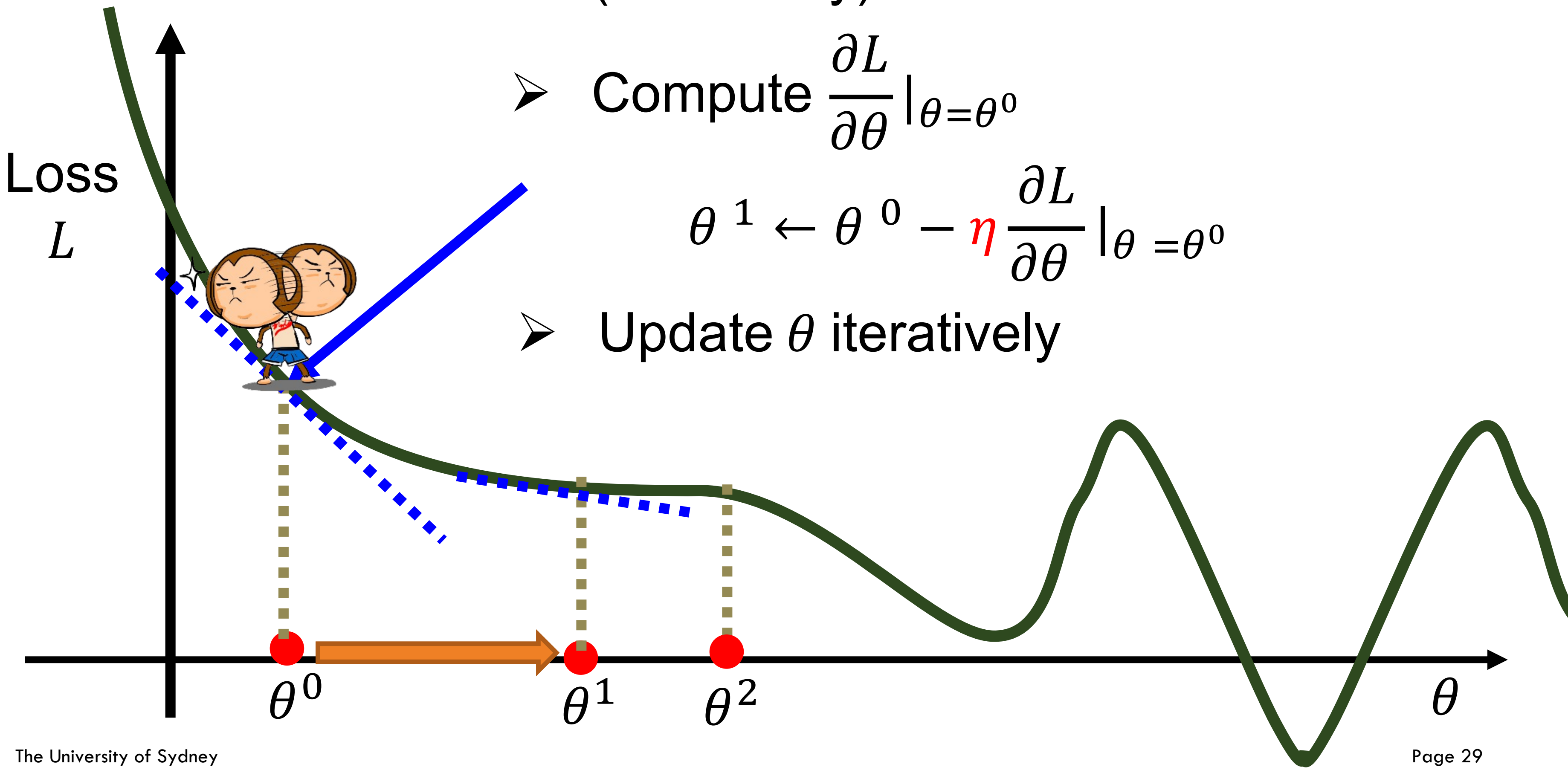
$$\theta^* = arg\min_{\theta} L$$

Gradient Descent

- (Randomly) Pick an initial value θ^0
- Compute $\frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$

$$\theta^1 \leftarrow \theta^0 - \eta \frac{\partial L}{\partial \theta}|_{\theta=\theta^0}$$

- Update θ iteratively



3. Optimization

$$\theta^* = arg \min_{\theta} L$$

- (Randomly) Pick initial values θ^0
- Compute

$$\boxed{\frac{\partial L}{\partial \theta}|_{\theta = \theta^0}}$$

$$\theta^1 \leftarrow \theta^0 - \eta \frac{\partial L}{\partial \theta}|_{\theta = \theta^0}$$

$\downarrow$

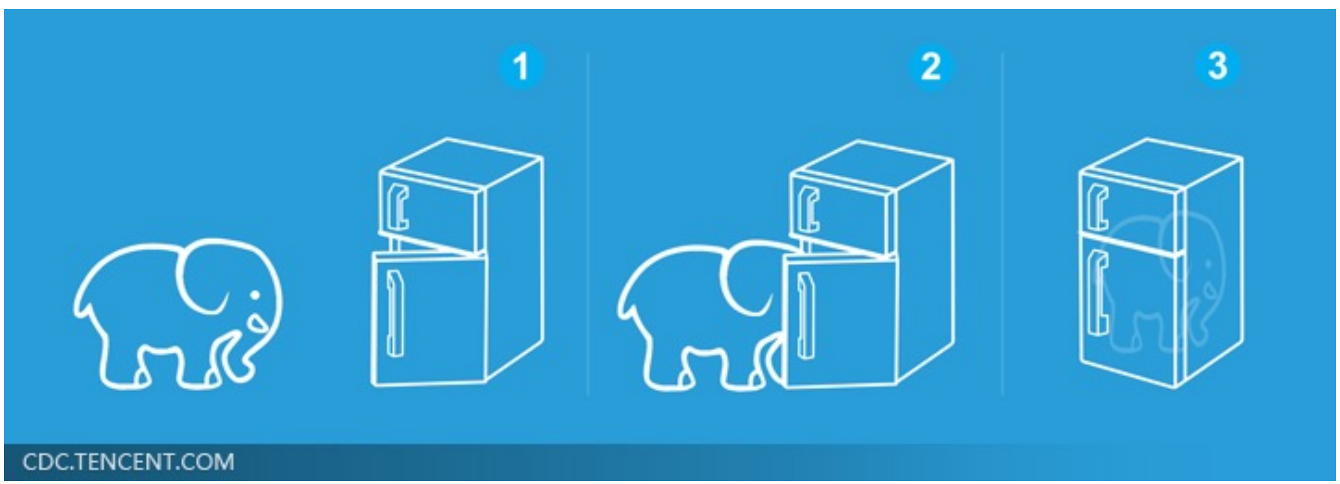
Can be done in one line in most deep learning frameworks

- Update θ interatively

Machine Learning is so simple ......

$$\hat{y} = f(x; \theta)$$

Step 1: function with unknown → Step 2: define loss from training data → Step 3: optimization



Introduce yourself: your name, background, any ML/data science/healthcare experience/story to share, etc.

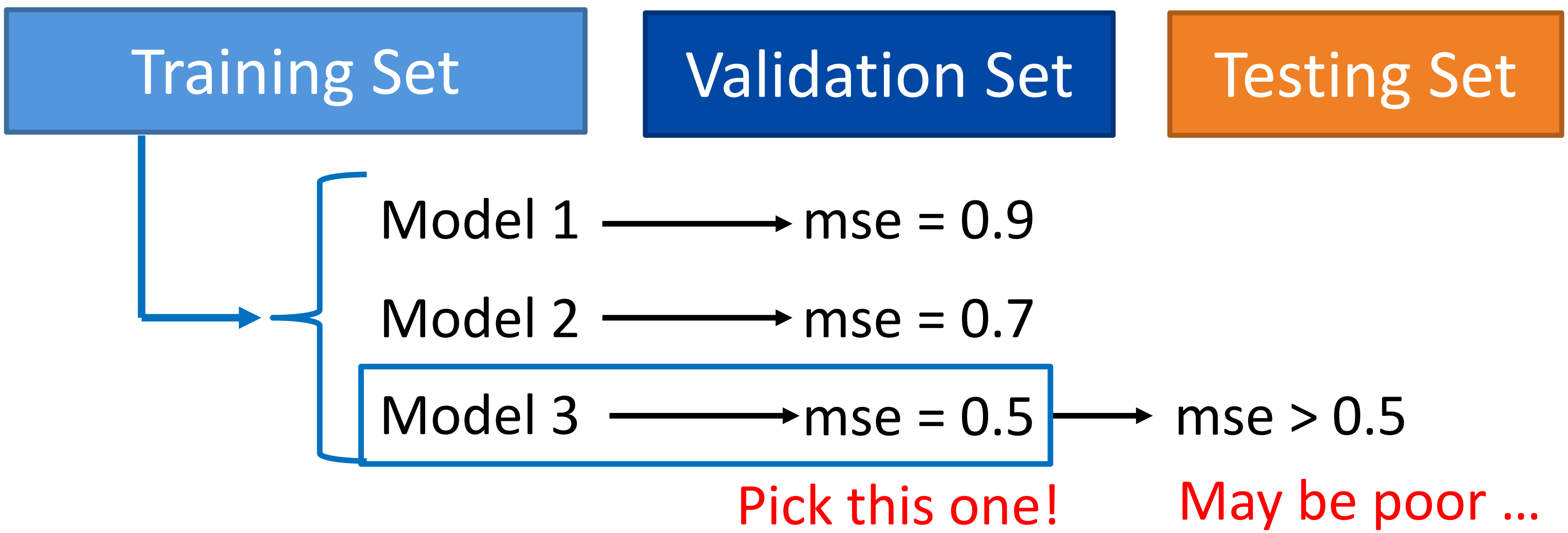
Make new friends.

Need a team member for the group assignment? The paper you like to pick?

Cross Validation

Dataset are first divided into three parts

- Training dataset: data examples used learn function
- Validation dataset: data examples used to periodically assess generalization performance and choose hyperparameters
- Test dataset: data examples evaluated only at the very end of model development (complete unseen during the development process)



$$\theta^{1} \leftarrow \theta^{0} - \eta \frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{0}}$$

η: learning rate

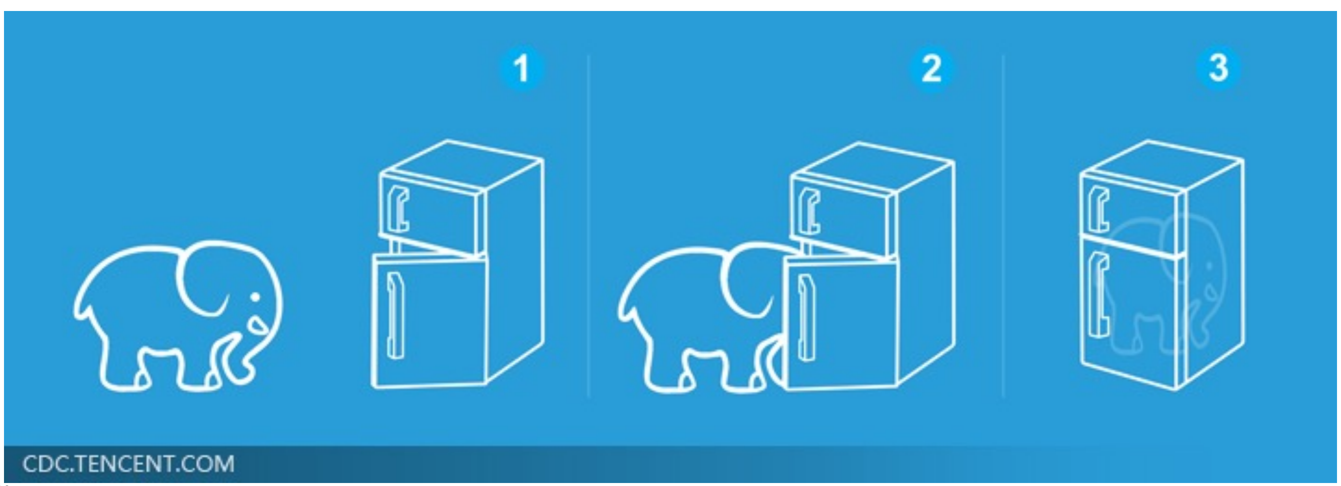
hyperparameters

Machine Learning is so simple ......

$$\hat{y} = f(x; \theta)$$

Step 1: function with unknown → Step 2: define loss from training data → Step 3: optimization

Training



Machine Learning is so simple ……

$$\hat{y} = f(x; \theta)$$

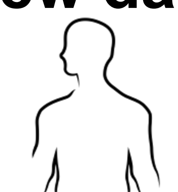
Step 1: function with unknown → Step 2: define loss from training data → Step 3: optimization

$\underbrace{\qquad\qquad\qquad\qquad}_{\textit{Training}}$

Training

θ^* achieves the smallest loss on **training data**:

How about **unseen new data**?



$; y$

Overfitting

- Small loss on training data, large loss on testing data. Why?

An extreme example

Training data:

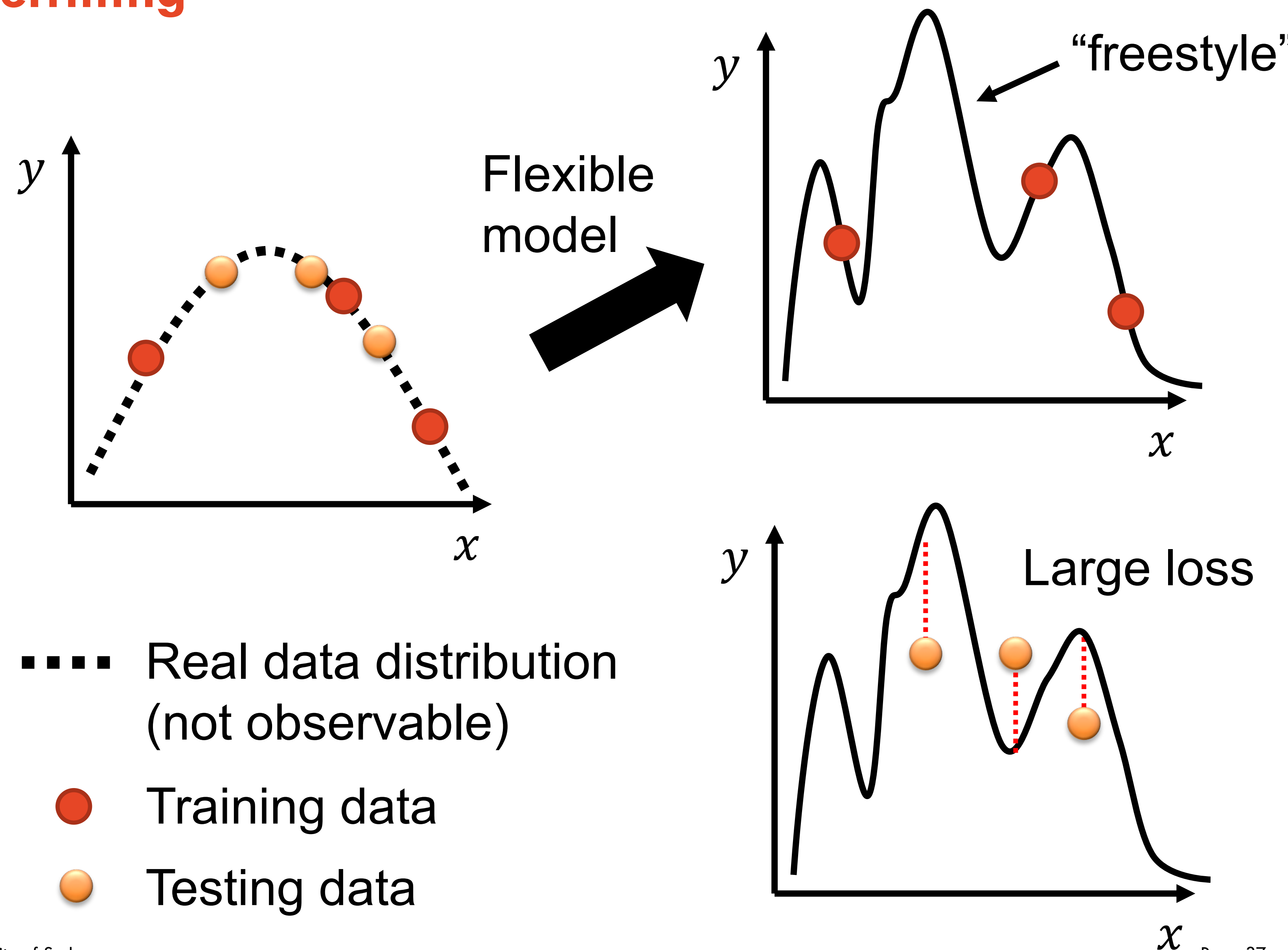
$$S = \{(\boldsymbol{x^1}, y^1), (\boldsymbol{x^2}, y^2), \dots, (\boldsymbol{x^N}, y^N)\}$$

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \begin{cases} y^i & \forall \boldsymbol{x^i} \in \boldsymbol{S} \\ random & otherwise \end{cases}$$

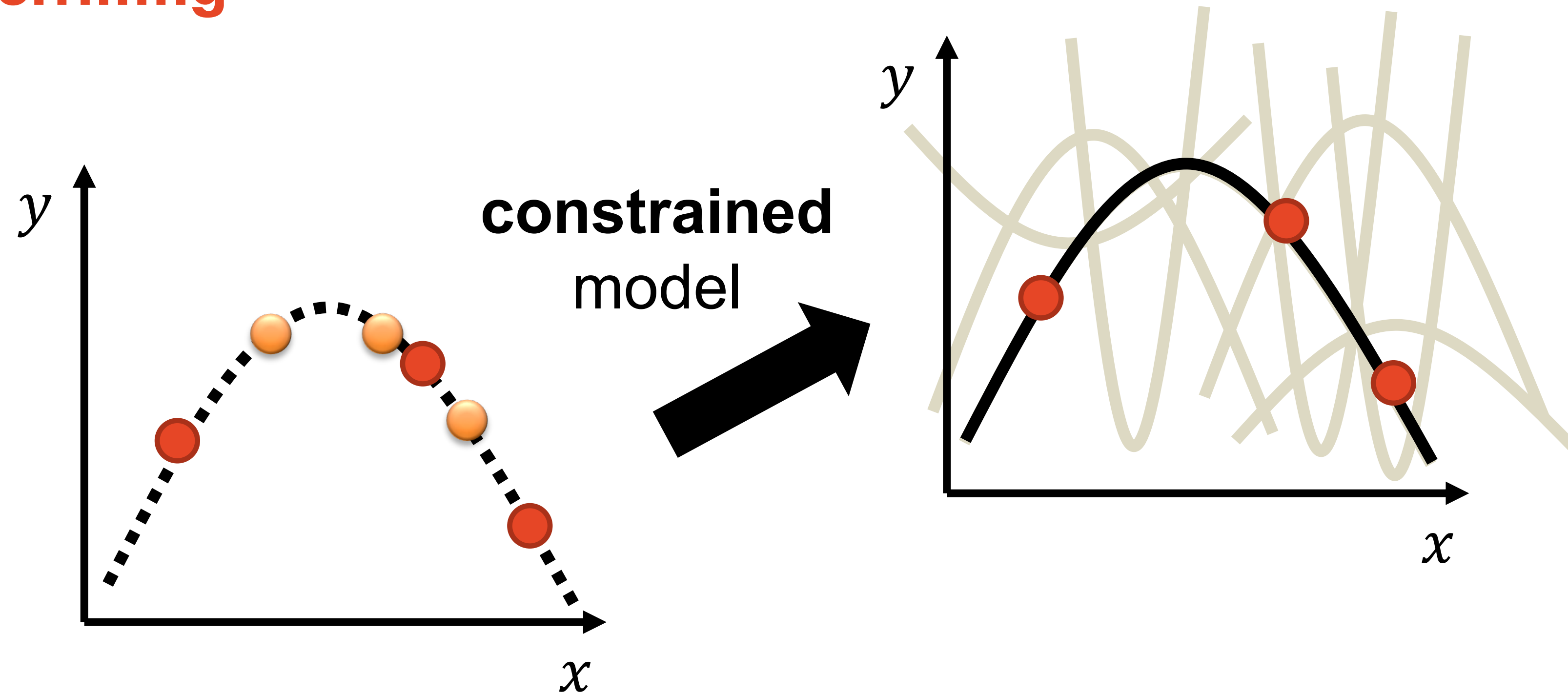
Less than useless …

This function obtains **zero training loss**, but **large testing loss**.

Overfitting



Overfitting



- Less parameters
- Less features
- Early stopping
- Regularization
- Dropout

Regularization

Different Types of Learning

Different Types of Learning

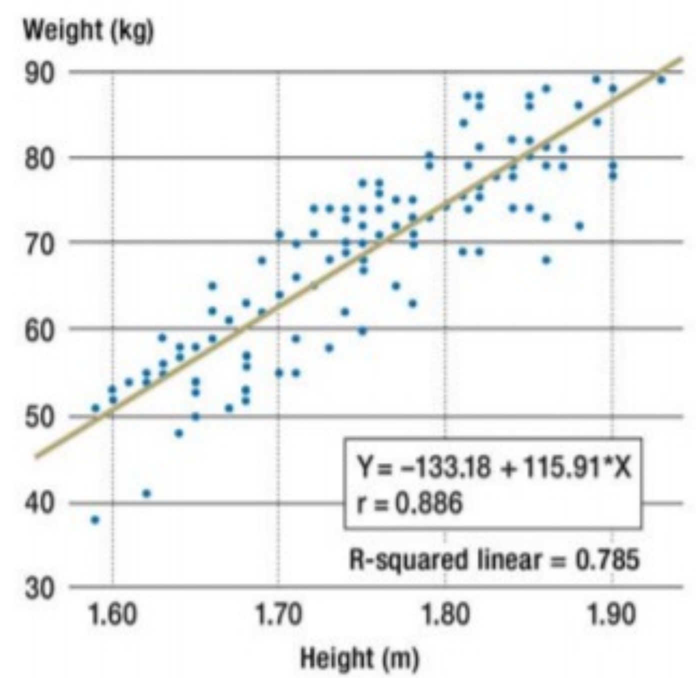
- Supervised Learning (This course foucus on)
  - Given: training data + desired outputs (labels)
- Unsupervised Learning
  - Given: training data (without desired outputs)
- Reinforcement Learning
  - Rewards from sequence of actions

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

Regression: y is real-valued

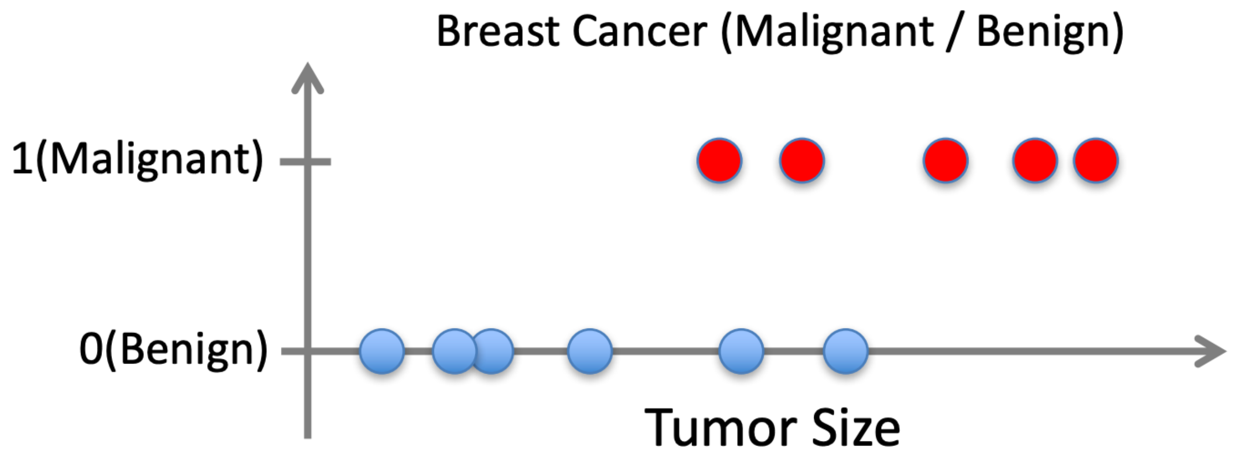


Supervised Learning: Classification

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Learn a function $f(x)$ to predict y given x

Classification: y is categorical

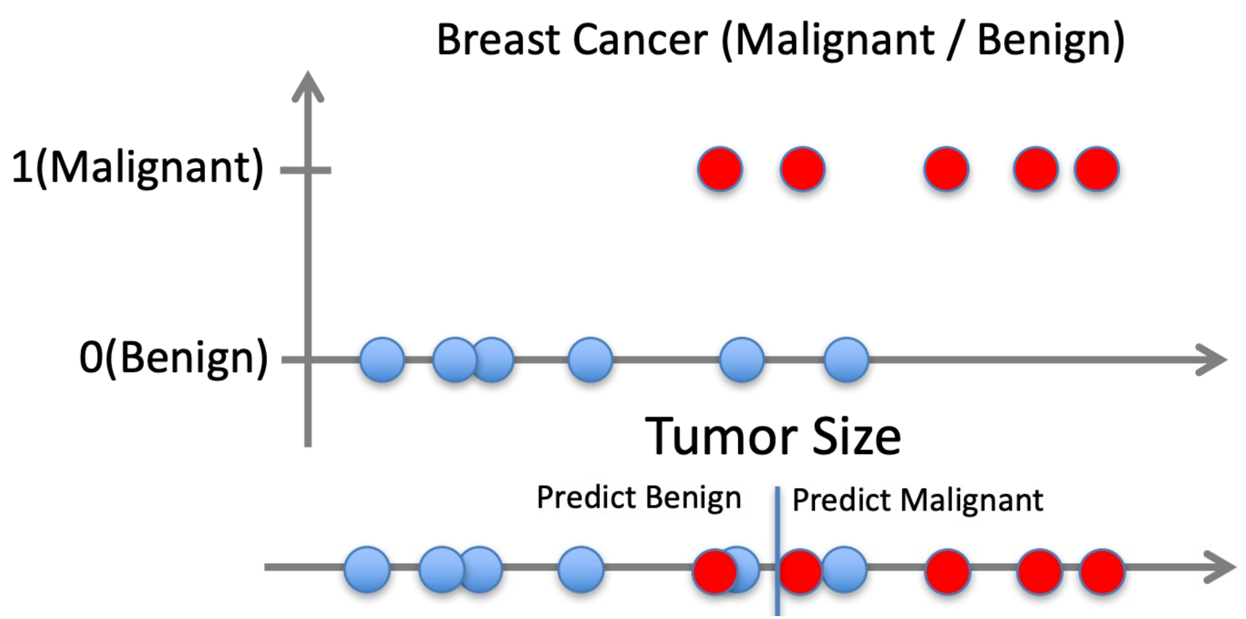


Supervised Learning: Classification

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Learn a function $f(x)$ to predict y given x

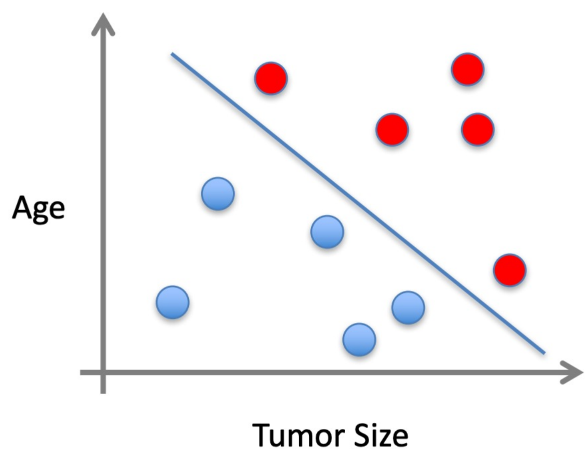
Classification: y is categorical



Supervised Learning

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

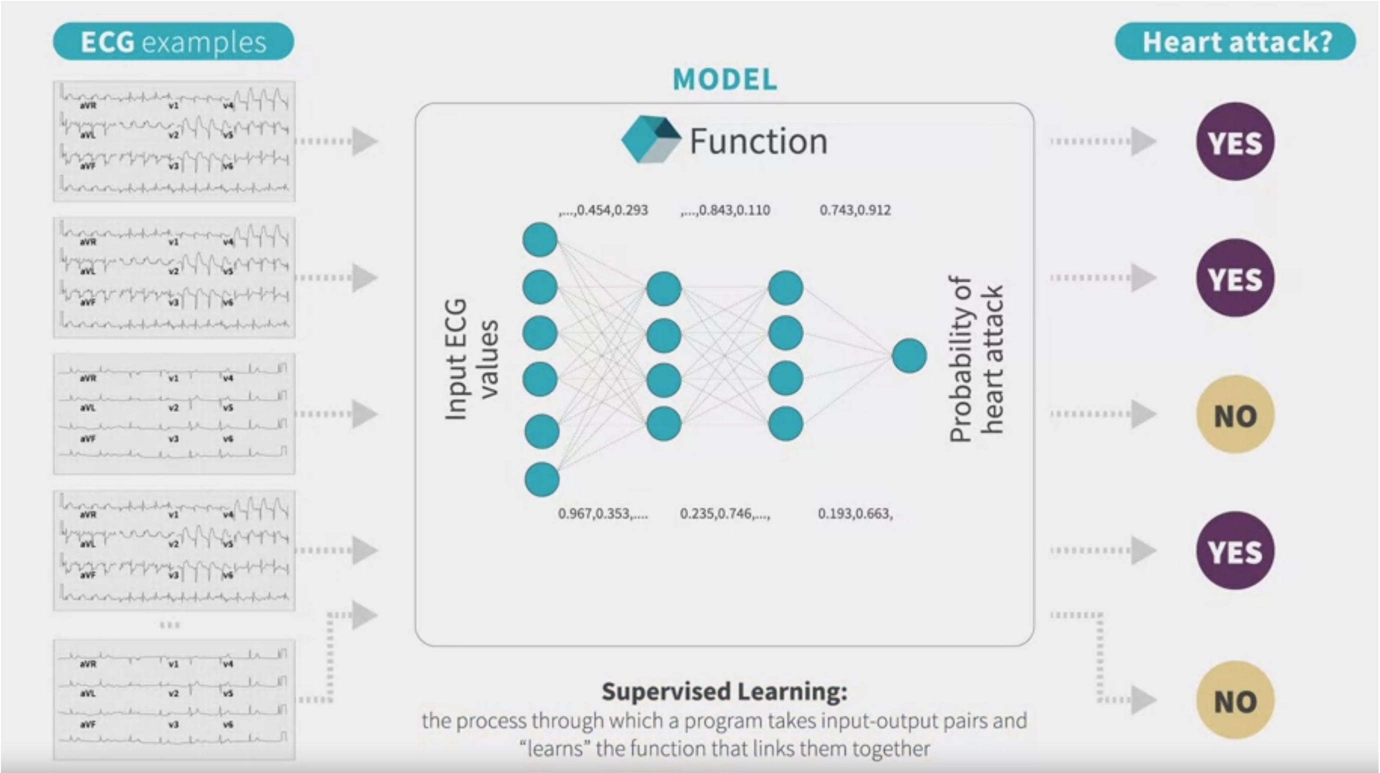
x can be multi-dimensional
– Each dimension corresponds to an attribute



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

Supervised Learning: Classification

We can even use signal data like electrocardiogram (ECG) to train a heart attack classification model.

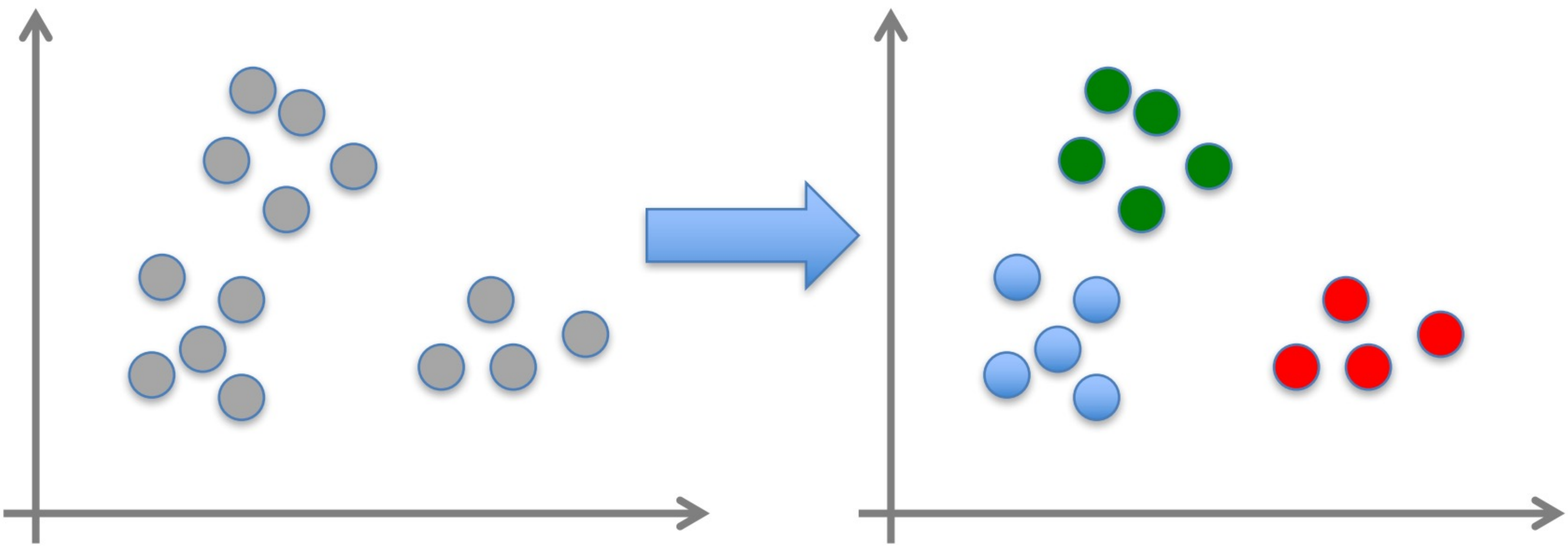


Unsupervised Learning

Given $x_1, x_2, \ldots, x_n$ without labels

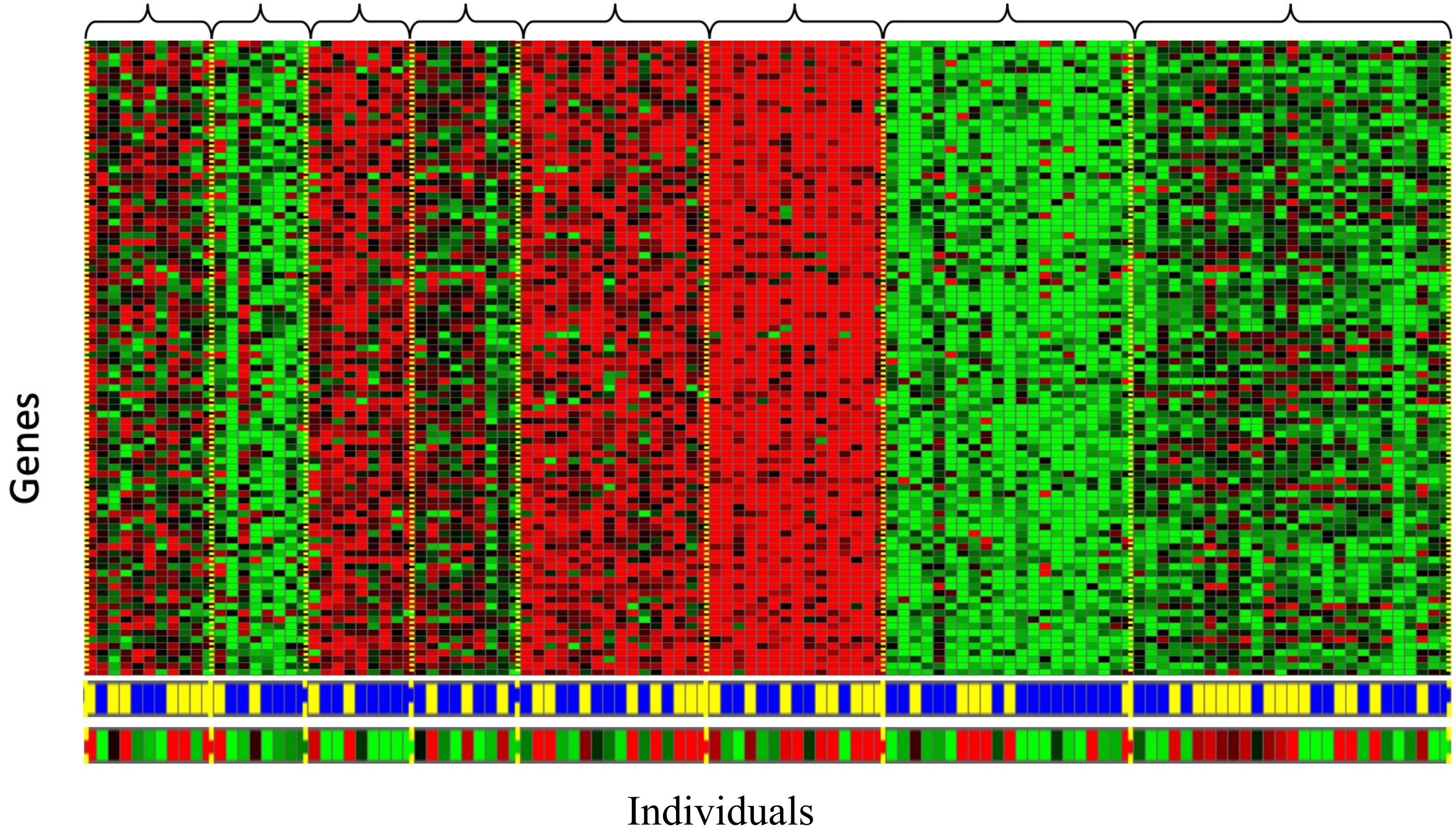
Output hidden structure behind the x's

– E.g., clustering



Unsupervised Learning

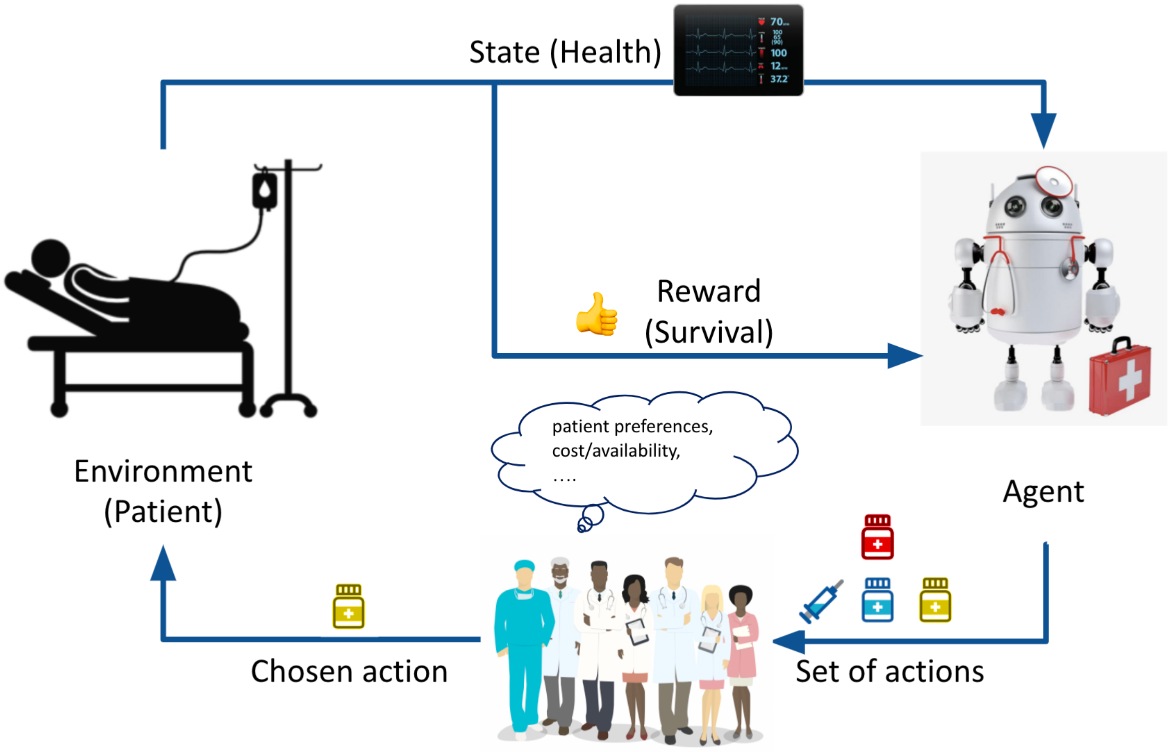
Genomics application: group individuals by genetic similarity



Reinforcement Learning

Given a sequence of states and actions with (delayed) rewards, output a policy

– Policy is a mapping from states → actions that tells you what to do in a given state



Reinforcement Learning

https://www.youtube.com/watch?v=4cgWya-wjgY