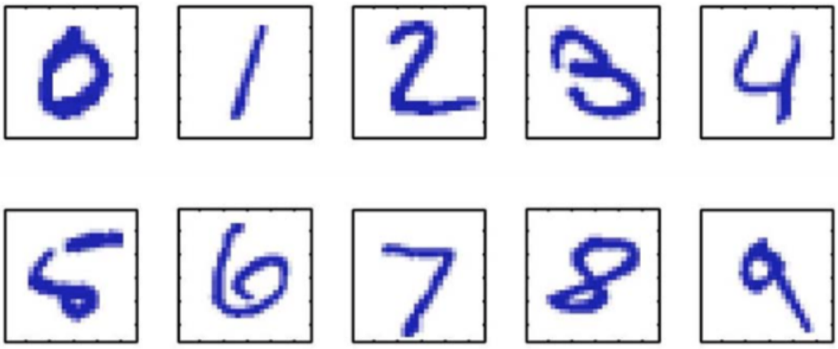
What's ML

- Machine learning (ML) is a type of artificial intelligence (AI)
- are data driven.
- Their primary work is to guess/predict based on past/training data provided to them.
- Output: a program

Example

Example 1: hand-written digit recognition



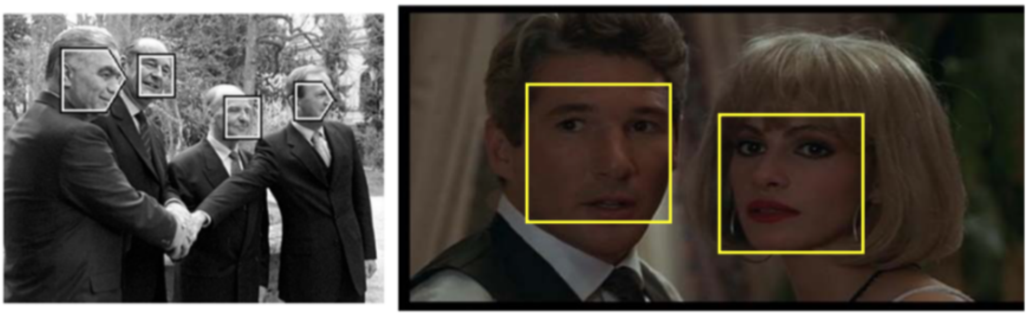
Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

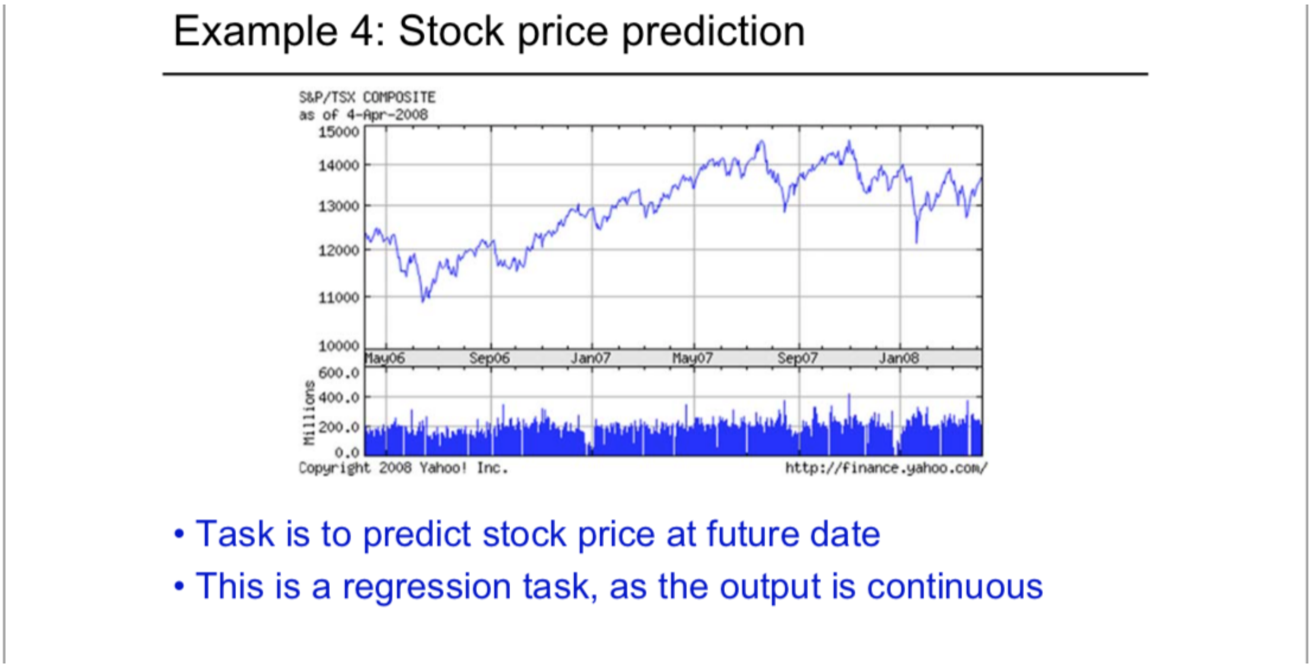
Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Example 2: Face detection



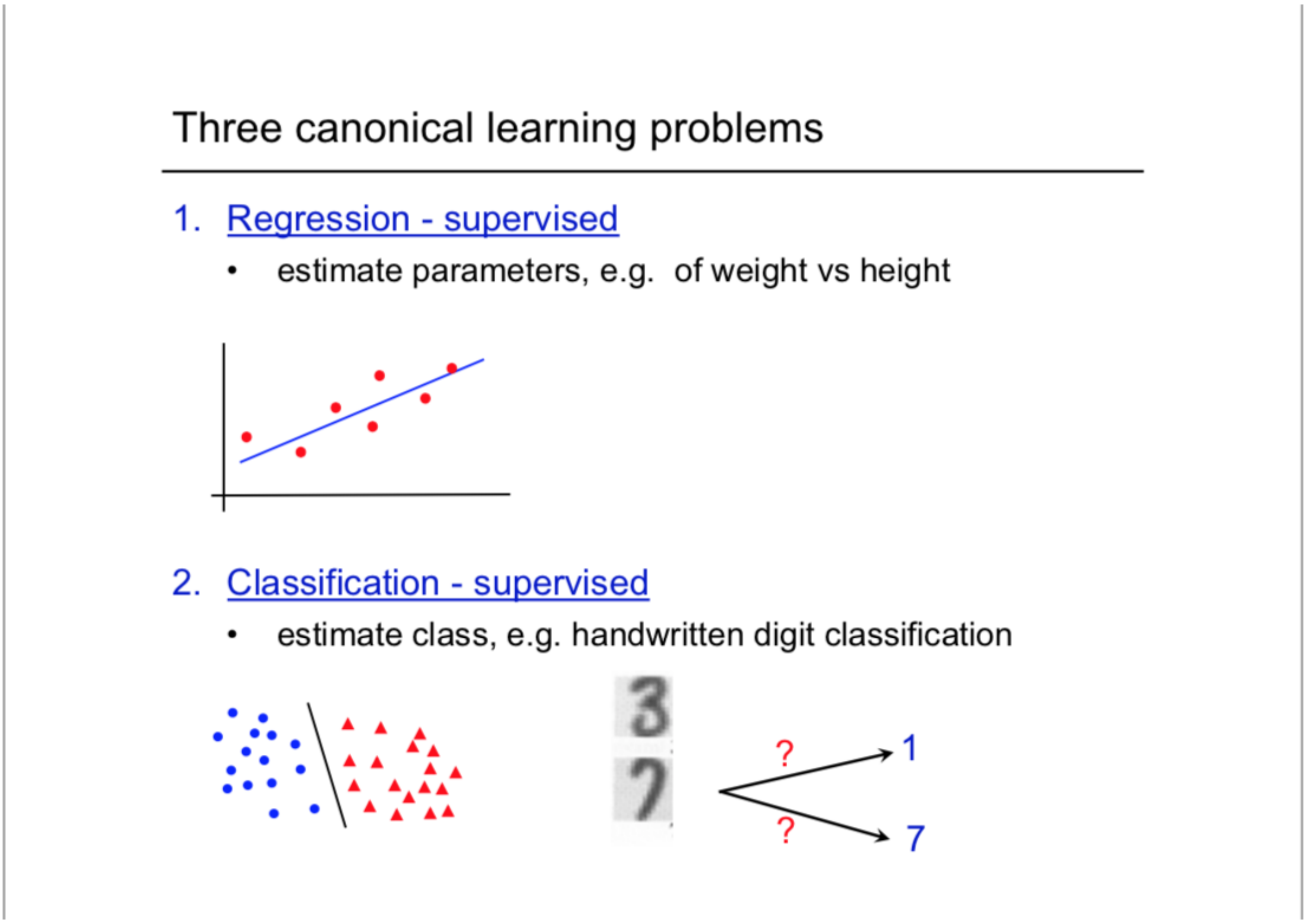
- Again, a supervised classification problem
- Need to classify an image window into three classes:
  - non-face
  - frontal-face
  - profile-face



Example 4: Stock price prediction

- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

Supervised Learning (Labelled Data)



Three canonical learning problems

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height
2. Classification - supervised
   - estimate class, e.g. handwritten digit classification

$$\text{Function to learn}: f: X \rightarrow \mathcal{Y}$$

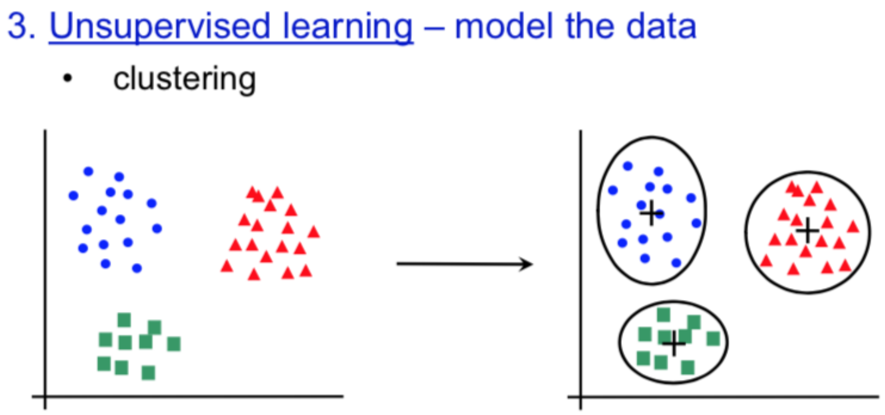
$$\text{Training Data } Z = \{(X^1, Y^1)\ldots(X^n, Y^n)\}$$

$$\text{Objective}: \text{minimize some loss function} \min \mathcal{L}(Y, \hat{Y})$$

EPE: Expected (squared) prediction error

$$\mathrm{EPE}(f) = \mathbb{E}[Y - f(X)]^2 = \int p(x) \int [y - f(x)]^2 p(y|x) dy dx = \mathbb{E}[\ \mathbb{E}[(Y - f(X))^2|X]\ ]$$
$$\Rightarrow \mathbb{E}[(Y - f(x))^2|x] = \mathbb{E}[Y^2|x] + f(x)^2 - 2f(x)\mathbb{E}[Y|x]$$
$$\Rightarrow f(x) = \mathbb{E}[Y|x]$$

Unsupervised



> 3. Unsupervised learning – model the data
>
> - clustering

Kmeans: Distance Based Clustering

Given a sample data, , *k*-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets, so that the

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2$$

Reinforcement Learning (AlphaGo)

The idea behind Reinforcement Learning is that an agent will learn from the **environment** by interacting with it and receiving rewards for performing actions.

https://sarvagyavaish.github.io/FlappyBirdRL/

Actions

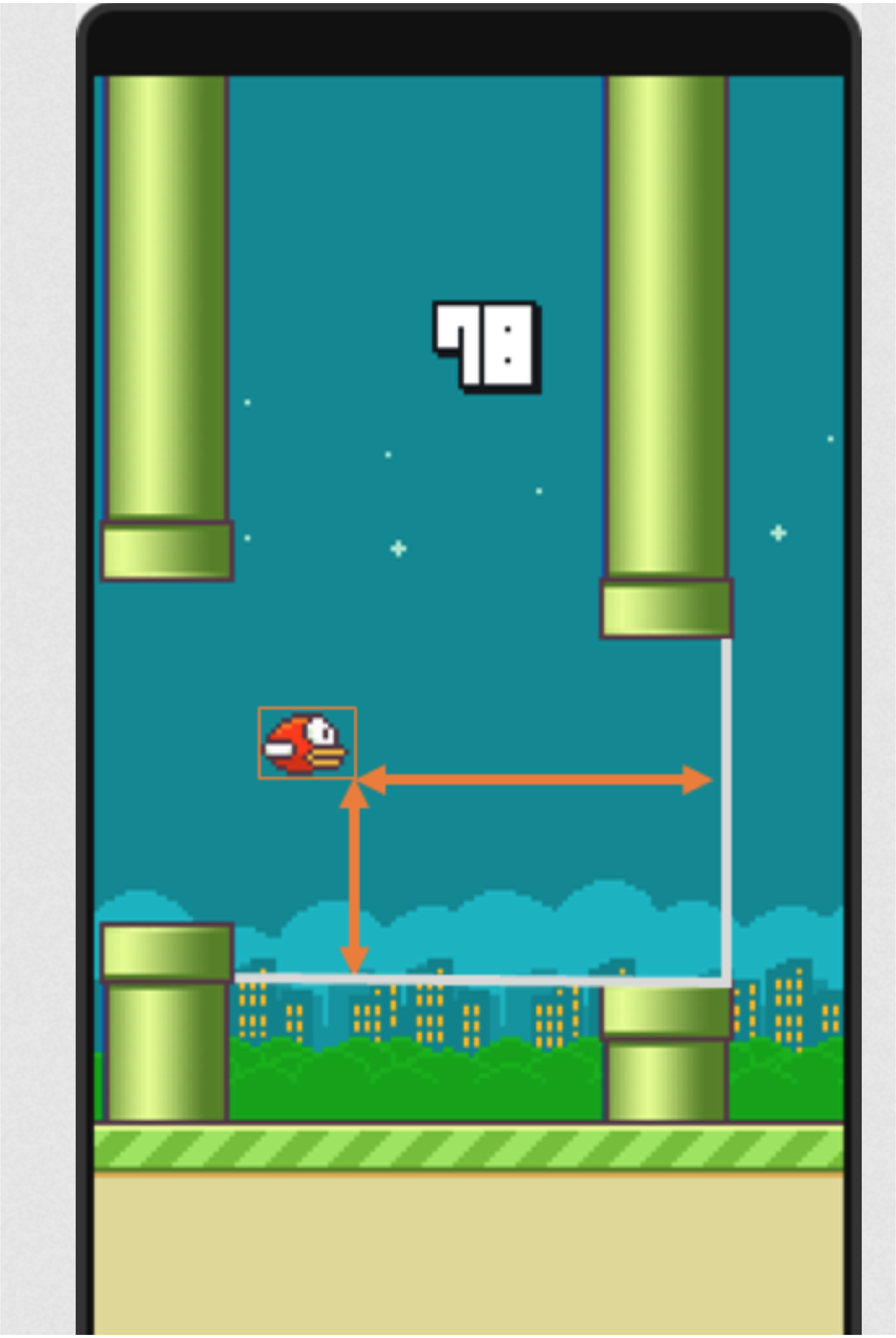
For each state, I have two possible actions

- Click
- Do Nothing

Rewards

The reward structure is purely based on the "Life" parameter.

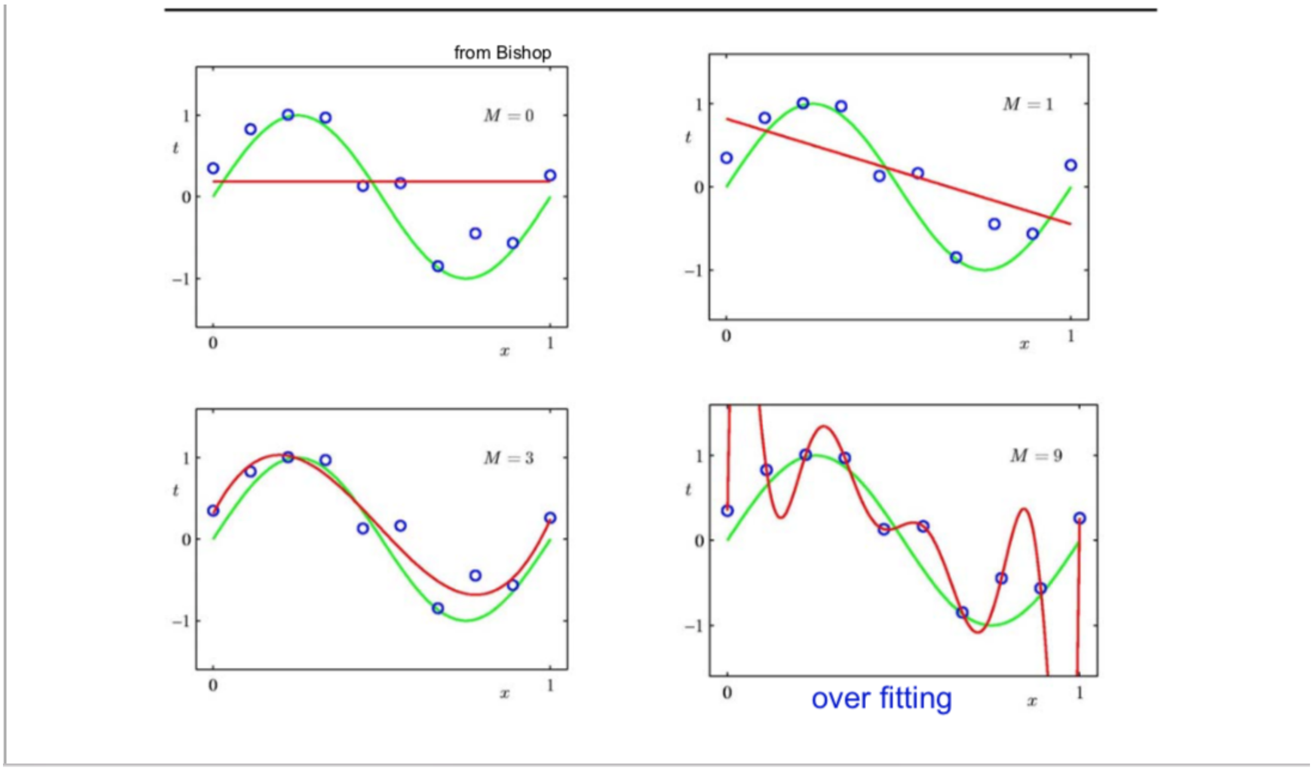
- **+1** if Flappy Bird is still alive
- **-1000** if Flappy Bird is dead


78

Model Capability: the ability to learn a function

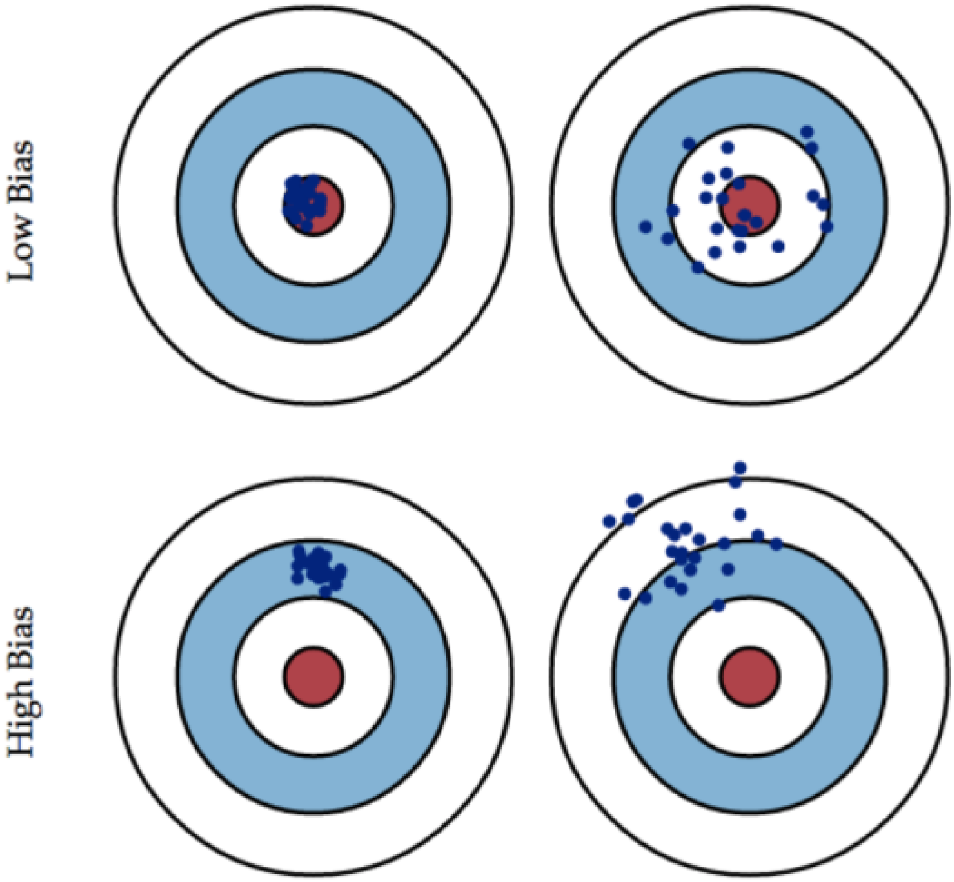
$$M = 1 : f(x) = wx$$
$$M = 3 : f(x) = w_1 x + w_2 x^2 + w_3 x^3$$

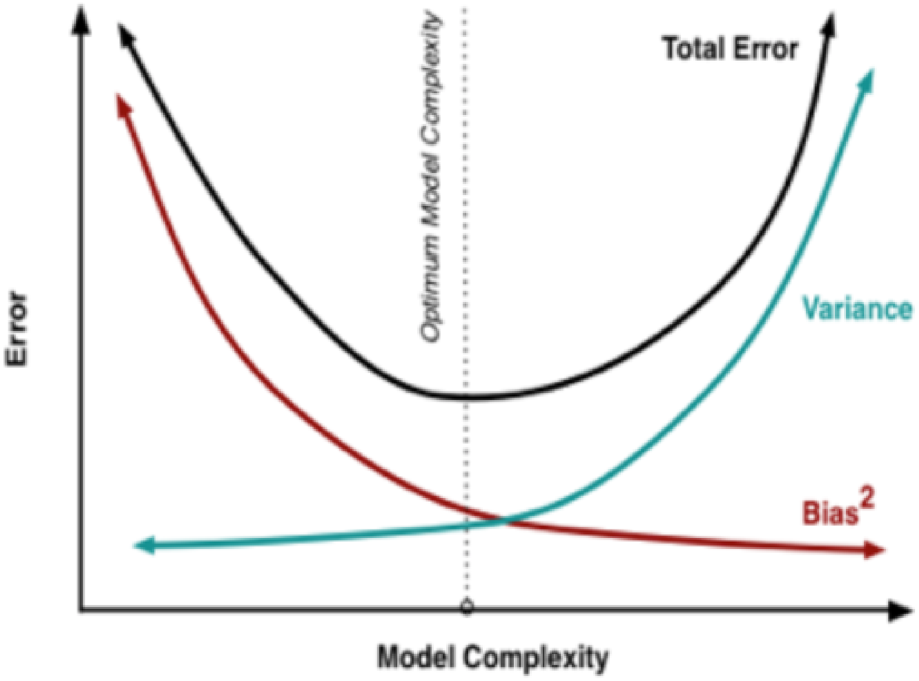
Overfit/underfit (Supervised Learning)



准与确

Low Variance High Variance





Bias

$$\mathbb{E}[\hat{f}(X)] - f(X)$$

Variance

$$\mathbb{E}[(\hat{f}(X) - \mathbb{E}[\hat{f}(X)])^2]$$

models are trained with different training data

Error

$$
\begin{aligned}
\mathbb{E}_\tau[(Y-\hat{f}\,(X))^2] &= \mathbb{E}_\tau[(Y-\hat{Y})^2] \\
&= \mathbb{E}[\hat{Y}^2 + Y^2 - 2Y\hat{Y} + \mu_{\hat{Y}}^2 - \mu_{\hat{Y}}^2 - 2\hat{Y}\mu_{\hat{Y}} + 2\hat{Y}\mu_{\hat{Y}}] \\
&= \mathbb{E}[(\hat{Y}-\mu_{\hat{Y}})^2] + \mathbb{E}[Y^2 - 2Y\hat{Y} - \mu_{\hat{Y}}^2 + 2\hat{Y}\mu_{\hat{Y}}] \\
&= \mathbb{E}[(\hat{Y}-\mu_{\hat{Y}})^2] + Y^2 - 2Y\mu_{\hat{Y}} - \mu_{\hat{Y}}^2 + 2\mu_{\hat{Y}}^2 \\
&= \underbrace{\mathbb{E}[(\hat{Y}-\mu_{\hat{Y}})^2]}_{variance} + \underbrace{|Y-\mathbb{E}[\hat{Y}]|^2}_{bias^2}
\end{aligned}
$$

On the Safety of Machine Learning

the minimization of both risk and uncertainty of harms,

Risk Function

$$R(h) = \mathbb{E}[L(h(X),Y)] = \int_{\mathcal{X}}\int_{\mathcal{Y}} L(h(x),y)f(x,y)dxdy$$

Empirical Risk Function

$$R^{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i),y_i)$$

Limitation of risk minimization approach -cannot capture the issues related to the uncertainty

1. assume data samples are **independent** drawn from the **identical** true underlying probability distribution, but may not always be the case.
    1. **not identical** : Training samples may not be representative of the testing samples
    . **not independent** : the model have high variance
2. training samples are absent from large parts of the X × Y space
3. The statistical learning theory analysis utilizes laws of large numbers to study the effect of

    finite training data and the convergence of risk, but in practice, the data is usually insufficient

STRATEGIES FOR ACHIEVING SAFETY

1 Inherently Safe Design

- Black models such as deep neural networks are so complex that it is very difficult to understand how they will react to **unknown unknowns**

use models that can be interpreted by people $y = wx + b$

exclude features that are not causally related to the outcome

incorporate interpretability and causality into the model formulation

for example, for decision tree, the number nodes is incorporated into the minimization.

2 Safety Reserves

The uncertainty in the matching of training and test data distributions or in the instantiation of the test set can be parameterized with the symbol θ

$R^*(\theta)$: the risk of the risk-optimal model if the θ were known.

$R(h, \theta)$: the risk of of model h

Objective

$$\min(\max_{\theta} \frac{R(h,\theta)}{R^*(\theta)}) \ or \ \min(\max_{\theta} R(h,\theta) - R^*(\theta))$$

Partition input space as protected $\mathcal{X}_p$ and unprotected groups $\mathcal{X}_u$,

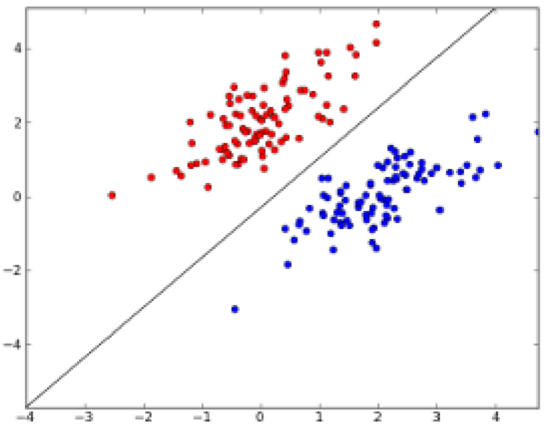
constraint the relative risk of harm for the protected versus unprotected group to a maximum value

$$\frac{\mathbb{E}_{\mathcal{X}_p}[R(h)]}{\mathbb{E}_{\mathcal{X}_u[R(h)]}} \leq z$$

3 Safe Fail

reject option : When the model selects the reject option, typically a human operator intervenes, examines the test sample, and provides a manual prediction.

classification problem



parts of X with low density may not contain any training samples at all and the decision boundary may be completely based on an inductive bias, thereby containing much epistemic uncertainty.

4 Procedural Safeguards:

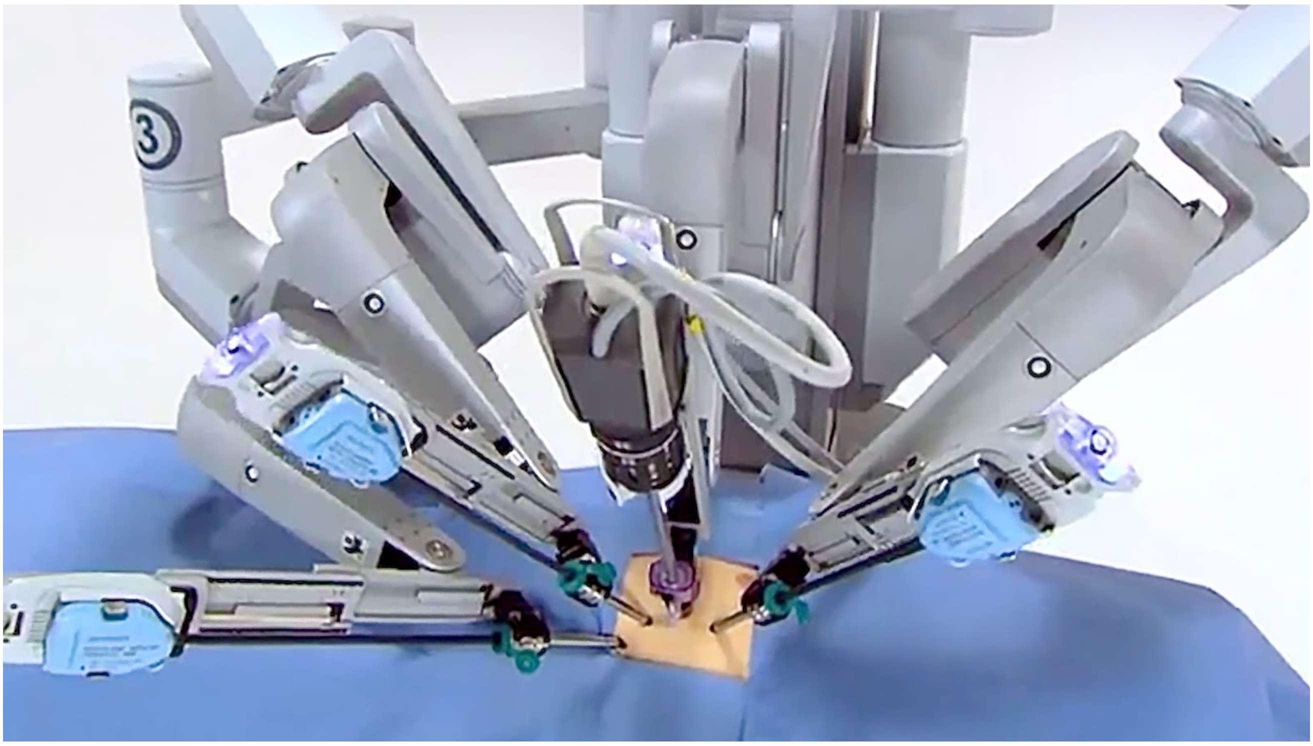
decision science applications： User experience design can be used 。

open source, Open data

EXAMPLE APPLICATIONS

Cyber-Physical Systems

Surgical Robots:



The robot control system receives the surgeon’s commands issued using the teleoperation console and translates the surgeon’s hand, wrist, and finger movements into precisely engineered movements of miniaturized surgical instruments inside patient’s body.

A machine learning enabled surgical robot continuously estimates the state of the environment (e.g., length or thickness of soft tissues under surgery) based on the measurements from sensors (e.g., image data or force signals) and generates a plan for executing actions (e.g., moving the robotic instruments along a trajectory).

Safety

uncertainty and large variability in the **operator actions and behavior**, **organ/tissue movements and dynamics**

Large Outcome Space: difficult to elicit all different outcomes and characterize which tasks or actions are costly enough to represent safety issues

The training data often consists of samples collected from a select set of surgical tasks performed by well-trained surgeons, which might not represent the variety of actions and tasks performed during a real procedure.

Possible Soulutions

1. One solution for dealing with these uncertainties is to assess the robustness of the system in the presence of unwanted and rare hazardous events by **simulating such events** in virtual environments
2. Another solution currently adopted in practice is through **supervisory control** of automated surgical tasks instead of **fully autonomous surgery** .

For example, if the robot generates a gemetrically optimized suture plan based on sensor data or surgeon input, it should still be tracked and updated in real time because of possible tissue motion and deformation during surgery

Decision Sciences: (Loan Approval)