Introduction to Statistical Learning

various learning tasks & framework

송 준

고려대학교
통계학과 / 융합데이터과학 대학원

Course Information

Meeting

- Time: 09:20 – 12:30, 토요일 (09:20 – 10:50, 15분 휴식, 11:05 – 12:30)
- Location: SK미래관 4119호
- Course Website: 고려대학교 LMS https://canvas.korea.ac.kr/
- Zoom: https://korea-ac-kr.zoom.us/j/83461928135?pwd=wOoXARO3k9EC4GIcmNZ7yWM3YPndAs.1

Instructor: 송준

- Email: junsong@korea.ac.kr
- Office: 정경관 424호

Course Information

학습목표

- 통계적 학습이론 원리의 이해
- 다양한 통계적 학습 방법론들의 이해와 실제 사례 적용
- 모델 구현, 적용 및 분석

Course Information

수업 주요 내용

- *Supervised Learning: Regression & Classification* 방법론 & 예제
- *Unsupervised Learning: Dimension Reduction & Clustering* 방법론 & 예제
- 기계학습 방법론 적용에 있어 고려해야 할 점

참고도서

- *Introduction to Statistical Learning with Python (https://www.statlearning.com/)*
- *Elements of Statistical Learning*
- *Hands-On Machine Learning with Scikit-Learn*

Programming Language

- *Python (VS Code* 사용 권장*:* 안내페이지 참고*)*

평가

중간과제

- 학습한 내용들을 기반으로 한 개념 확인 및 데이터 분석 실습

기말과제

- 통계적 학습의 개념 및 원리
- 데이터 분석 실습

출석

- *LMS* 내에서 간단한 *Quiz* 참여 *(*수업 중 *2*분간 공개*)*

중간과제	시험	출석	*Total*
30%	*40%*	*30%*	*100%*

What’s Machine Learning?

What is Machine Learning?

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

Herbert Simon

What is Machine Learning?

"Machine learning … gives computers the ability to learn without being explicitly programmed."

Arthur Samuel

What is Machine Learning?

Tom Mitchell: *Algorithms that*

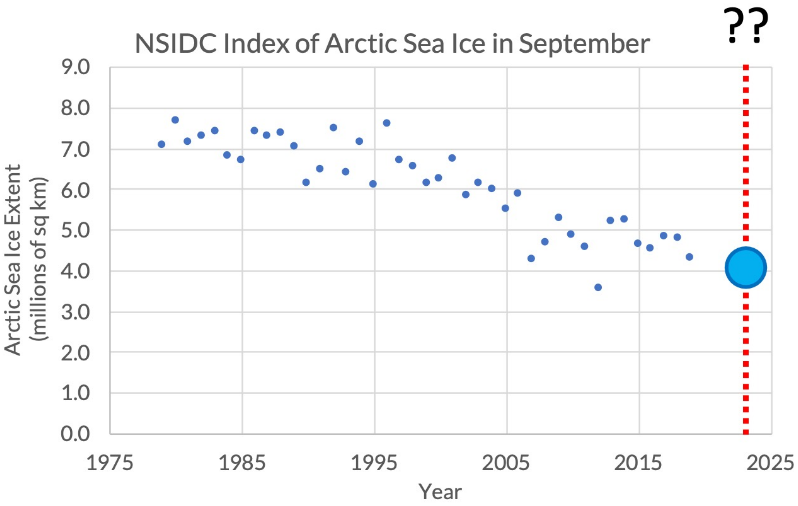
- *improve their* **performance** P
- *at* **task** T
- *with* **experience** E *(데이터)*

A well-defined machine learning task is given by P, T, E

Example

북극 해빙 면적 변화

Photo by NASA Goddard



Example

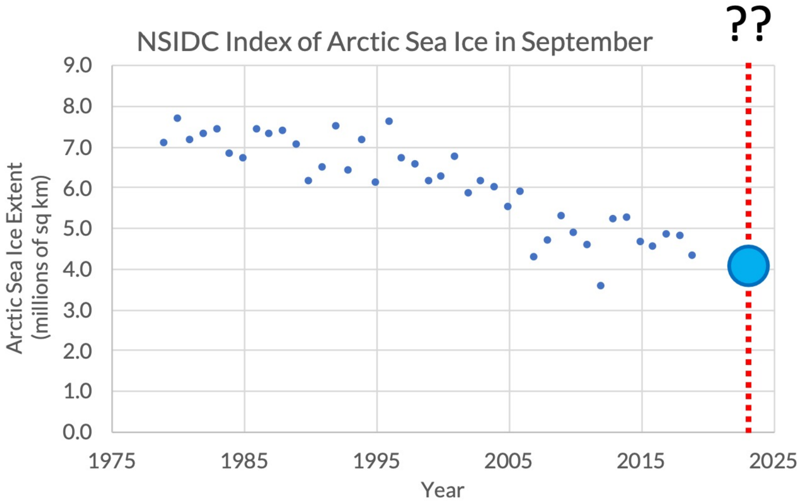
Tom Mitchell: Algorithms that

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

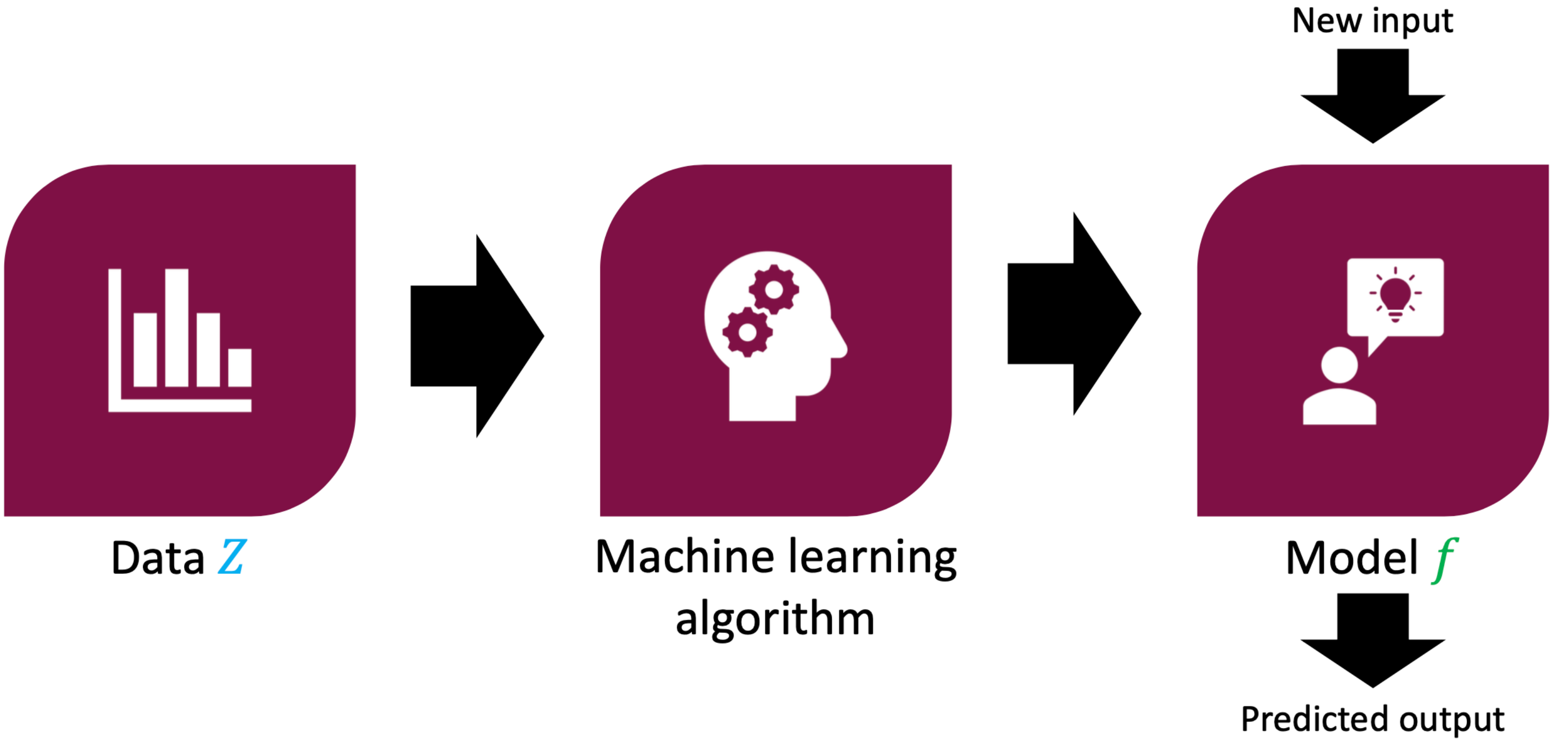
A well-defined machine learning task is given by P,T,E

➡ **북극 해빙 면적 변화**

- T : 2025년 북극 해빙 면적 예측
- P : 오차
- E : 과거 데이터



What is Machine Learning?



New input

Data Z

Machine learning algorithm

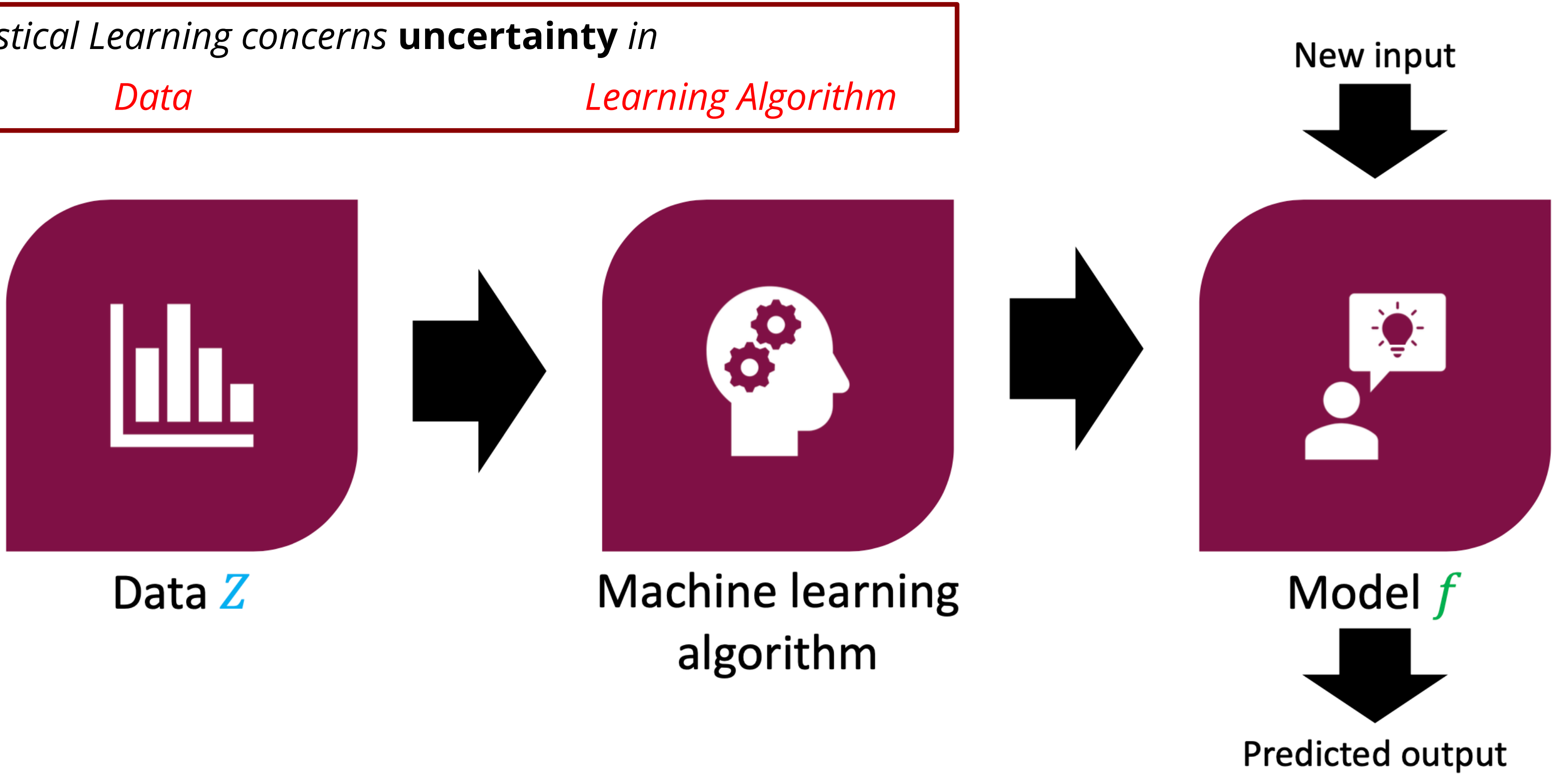
Model f

Predicted output

What is Machine Learning?

Statistical Learning concerns **uncertainty** *in*

Data *Learning Algorithm*



New input

Data Z

Machine learning algorithm

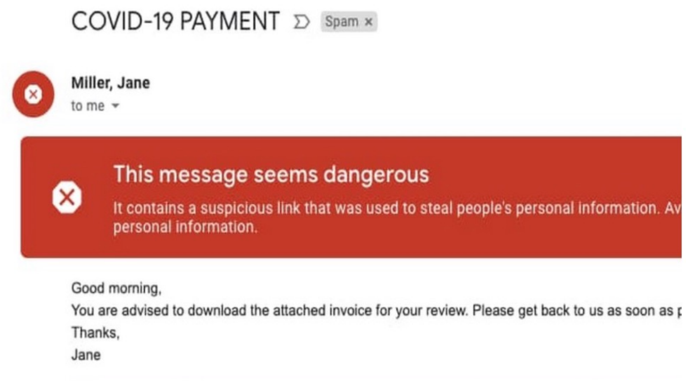
Model f

Predicted output

- 많은 ML/AI 방법론들은 확률 기반의 통계학적 방법론을 근간으로 개발됨

Machine Learning in Action

Daily Life



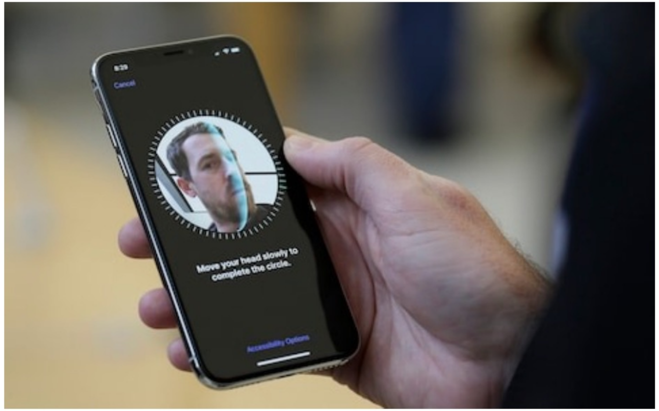
COVID-19 PAYMENT

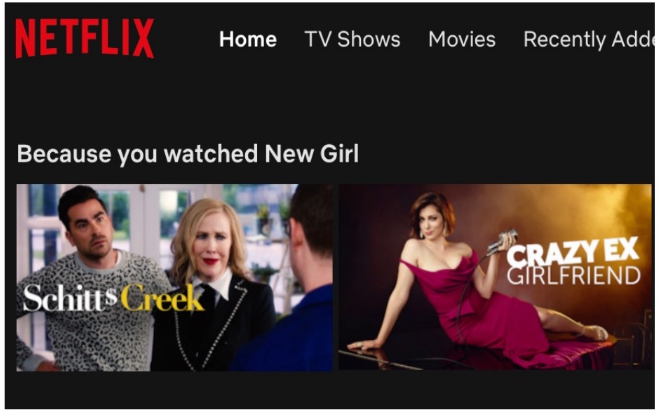
Miller, Jane
to me

This message seems dangerous

It contains a suspicious link that was used to steal people's personal information. Av
personal information.

Good morning,
You are advised to download the attached invoice for your review. Please get back to us as soon as p
Thanks,
Jane





Because you watched New Girl

Document Classification

Document classification

⇨

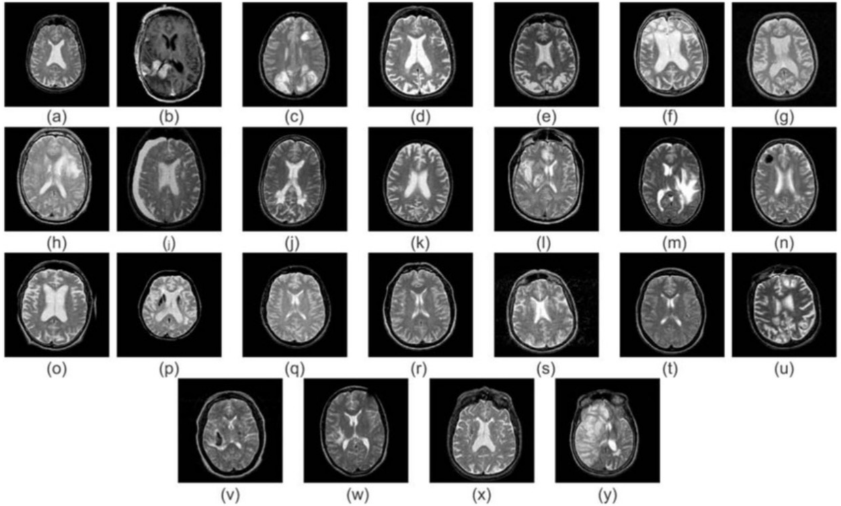
Sports

Science

News

Radiology and Medicine

Input: Brain scans



Output: Neurological disease labels

Machine learning studies on major brain diseases: 5-year trends of 2014–2018

https://www.nature.com/articles/s41573-019-0024-5

Application of machine learning in drug discovery and development

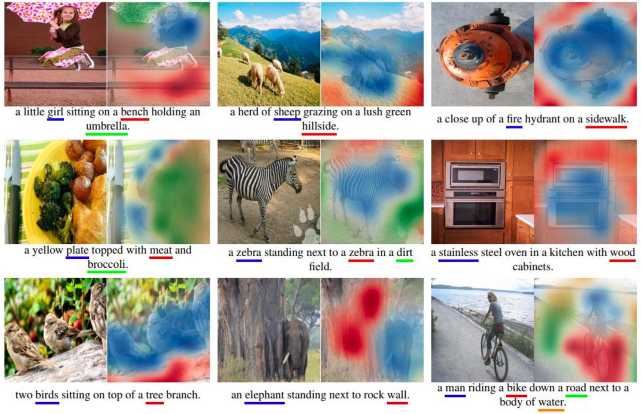
Machine Learning in Action

ML on Images

- □ Face detection
- □ Handwritten text detection
- □ Image super-resolution

Hello Alex,
Hello
Alex
how is your
how
is
your
projekt going?
Projekt
going
?

Image caption generation



a little girl sitting on a bench holding an umbrella.

a herd of sheep grazing on a lush green hillside.

a close up of a fire hydrant on a sidewalk.

a yellow plate topped with meat and broccoli.

a zebra standing next to a zebra in a dirt field.

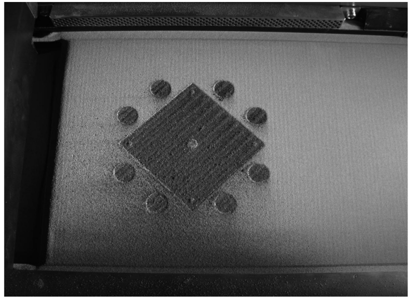
a stainless steel oven in a kitchen with wood cabinets.

two birds sitting on top of a tree branch.

an elephant standing next to rock wall.

a man riding a bike down a road next to a body of water.

Anomaly Detection for Quality Control in Additive Manufacturing



	Spatter	Clean Powder
Spatter	99.0%	2.3%
Clean Powder	1.0%	97.7%

Types of Learning

Types of Learning

- Supervised learning
  - **Input:** Examples of inputs (x) and outputs (y)
  - **Output:** Model that predicts unknown output given a new input

- Unsupervised learning
  - **Input:** Examples of some data (x) (output is not specified)
  - **Output:** Representation of structure in the data and further

Types of Learning

- **Supervised learning (with responses or labels (y))**
  - *Regression, classification*
- **Unsupervised learning (without responses or labels (y))**
  - *Density estimation, clustering, dimension reduction*

$\Bigg\}$ **Foundational problem**

Types of Learning

- **Supervised learning (with responses or labels (y))**
  - *Regression, classification*
- **Unsupervised learning (without responses or labels (y))**
  - *Density estimation, clustering, dimension reduction*

Foundational problem

As SL/ML have become highly developed and more sophisticated, more problems have arisen in a variety of scenarios

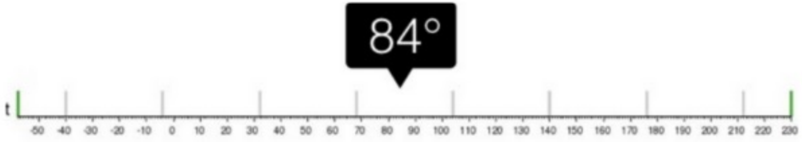
- *Reinforcement learning (interactive, maximizing reward)*
- *Semi-supervised learning (y's are partially observed)*
- *Self-supervised learning (no y, but give y manually)*
- *Active learning (interactive, machine-human)*
- *Online learning (incremental, update pre-fitted model(large) with a new data(small))*
- *Transfer learning (using pre-trained model in a new problem)*
- *Multitask learning (multi-task from one model)*
- *Federated learning (multi-source, privacy consideration)*

Supervised Learning

Supervised Learning : Regression vs. Classification

Regression

What will be the temperature tomorrow?

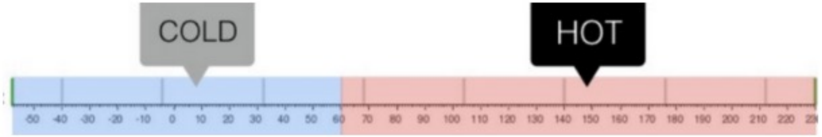


Fahrenheit

Predict variable : continuous

Classification

Will it be hot or cold tomorrow?



Fahrenheit

Predict variable : categorical

Supervised Learning : Regression vs. Classification

- *Where does Y reside?*
  - **Regression** *: Real vector space*
  - **Classification***: A finite set. {c1,c2, …, ck}*

- *Real Number: Math operations!(+, -, *, /)*
- *finite set don't have the math operations. cat+dog? cat-dog?*

- *Differently treated in*
  - *modeling*
  - *(E)* **data** *coding*
  - *(T) developing a* **method** *to* **do the task**
  - *(P)* **measuring the performance** *of the method*
  - *etc*

Supervised Learning : Regression

Regression

- **Population-level** Regression Setting

$$Y = f(\mathbb{X}) + \varepsilon$$

where X is a random element in $\mathbb{R}^p$

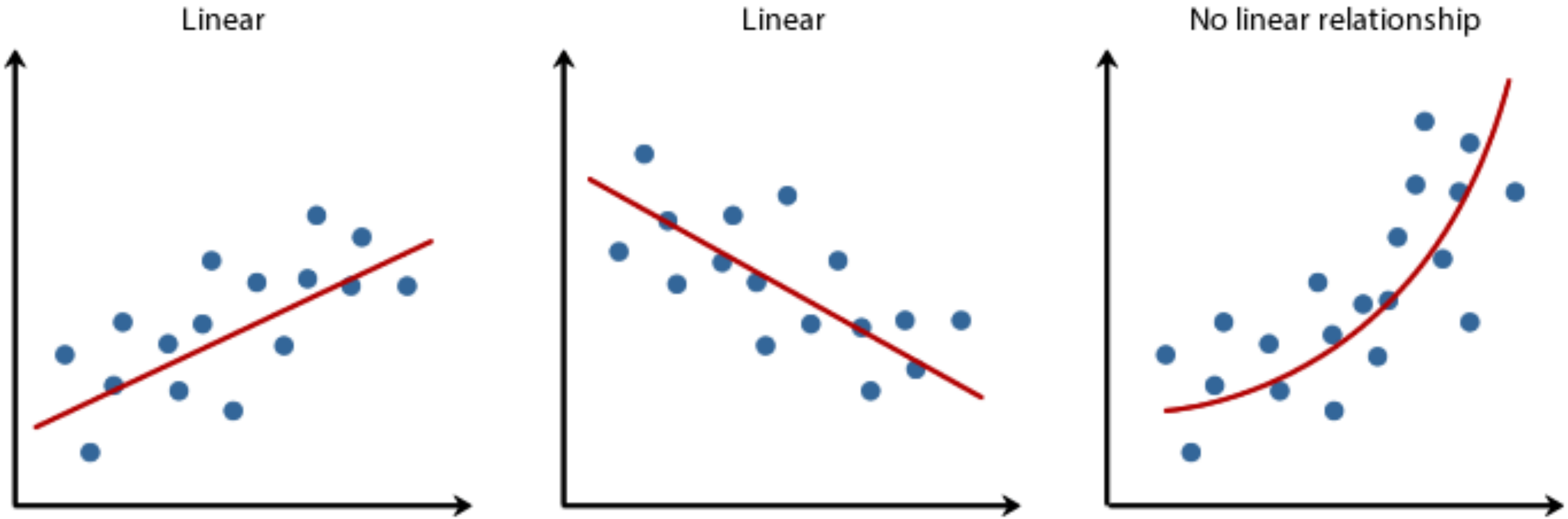
- **Sample-level**: Based on $n -$ paired observation, $(Y_1, \mathbb{X}_1), (Y_2, \mathbb{X}_2), , , (Y_n, \mathbb{X}_n)$, say D_n.

  Then we estimate $f(\cdot)$ using the observed data D_n.

$$\hat{f} = g(D_n) : \mathbb{R}^p \rightarrow \mathbb{R}$$

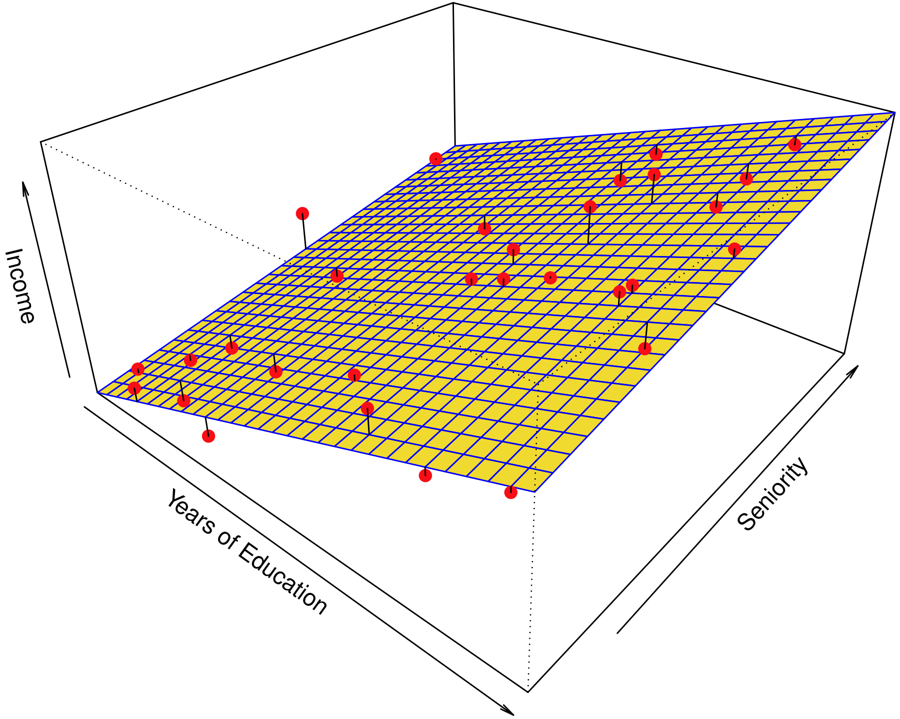
- Prediction of a new data point $\mathbb{X}^*$ is $\hat{Y}^* = \hat{f}(\mathbb{X}^*) = g(D_n)(\mathbb{X}^*)$

Supervised Learning : Linear Regression



Supervised Learning : Multiple Linear Regression

Regression **linear** *regression example*



- *Y: Income*
- *X=(X1,X2)=(Years of Ed., Seniority)*

- *red dots: observed data*

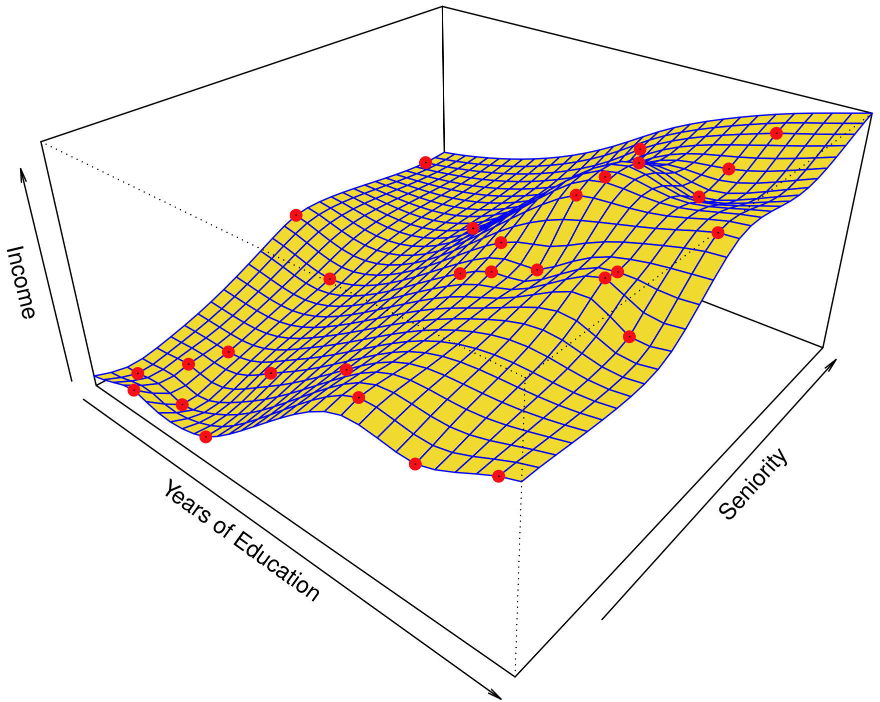
$$(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n),$$

- *yellow surface: graph of fitted $\hat{f}$*

$$\hat{f} = g(D_n) : \mathbb{R}^p \to \mathbb{R}$$

Supervised Learning : Nonlinear Regression

Regression **nonlinear** *regression example*



- *Y: Income*
- *X=(X1,X2)=(Years of Ed., Seniority)*

- *red dots: observed data*

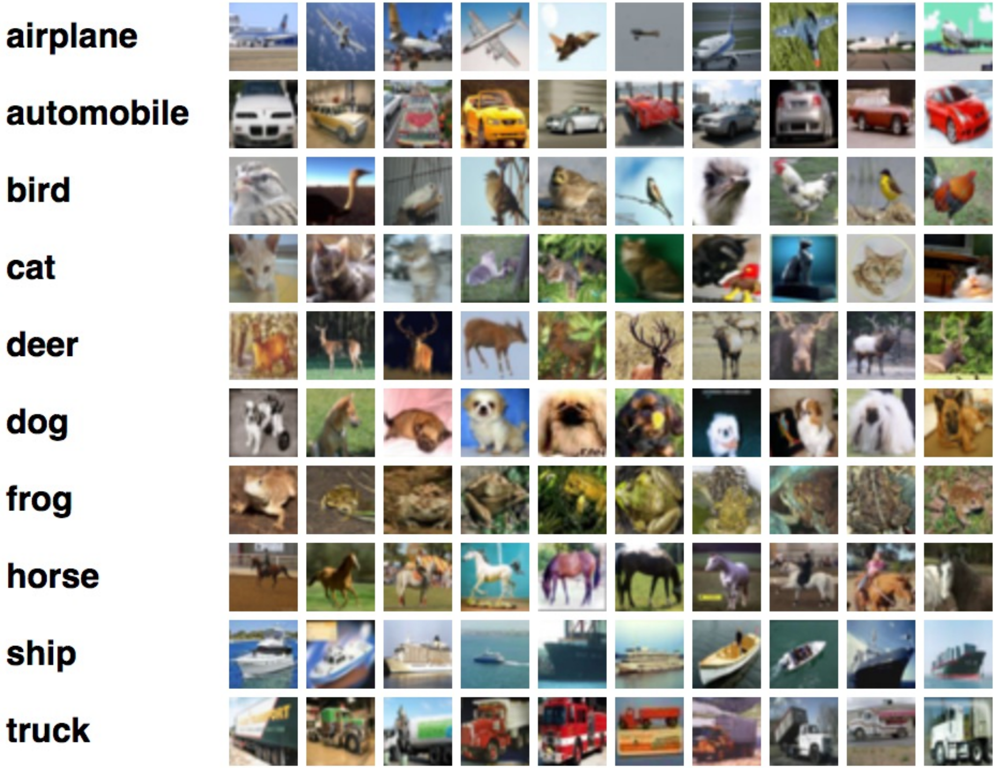
$$(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n),$$

- *yellow surface: graph of fitted $\hat{f}$*

$$\hat{f} = g(D_n) : \mathbb{R}^p \to \mathbb{R}$$

Supervised Learning : Classification

Classification y x



Data

(x,y)

(image1, 'airplane'),

(image2, 'airplane'),

.

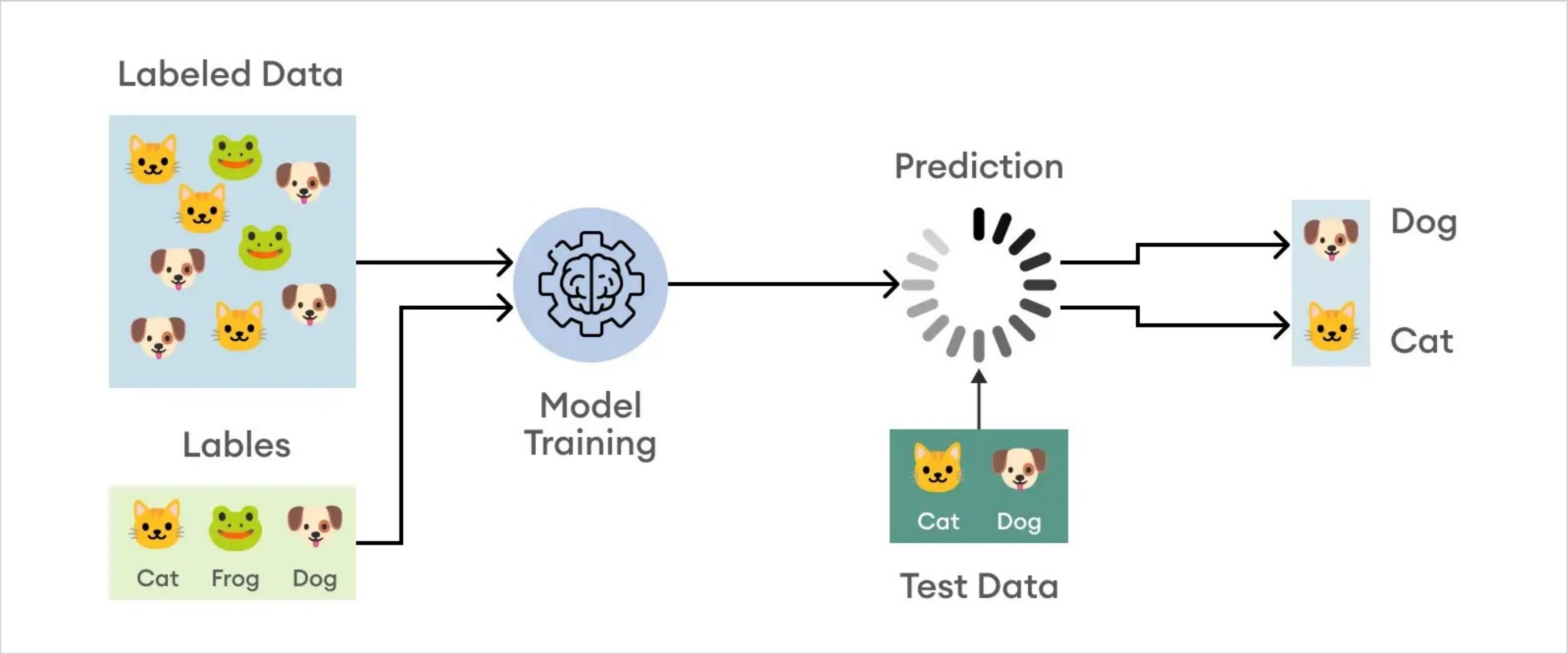
.

.

(image100, 'truck')

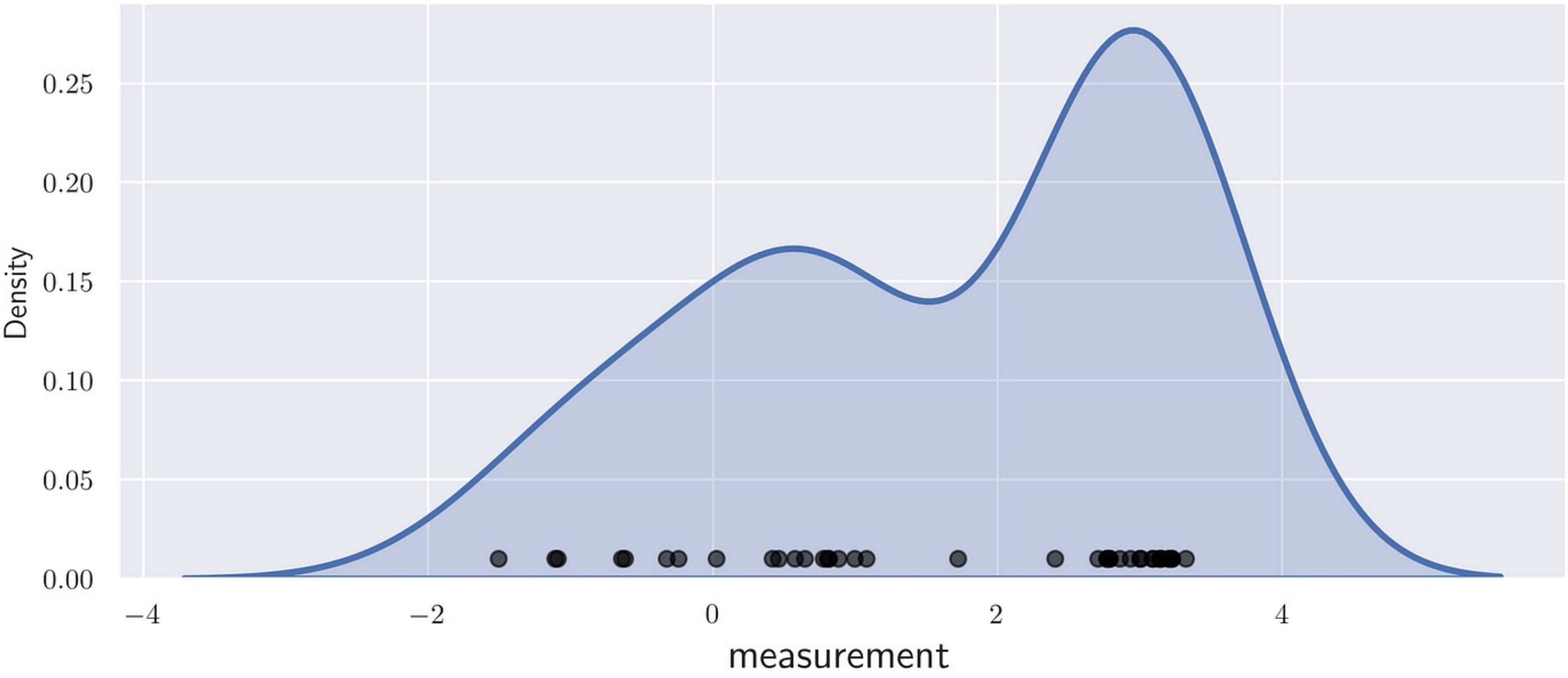
Supervised Learning : Classification

Classification



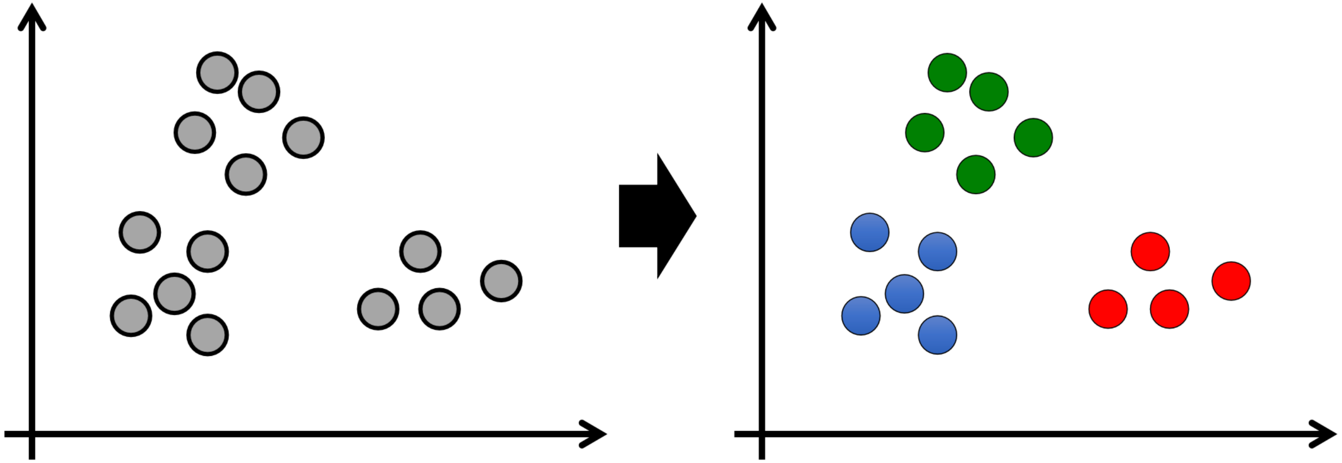
Unsupervised Learning

Unsupervised Learning : Probability Density Estimation



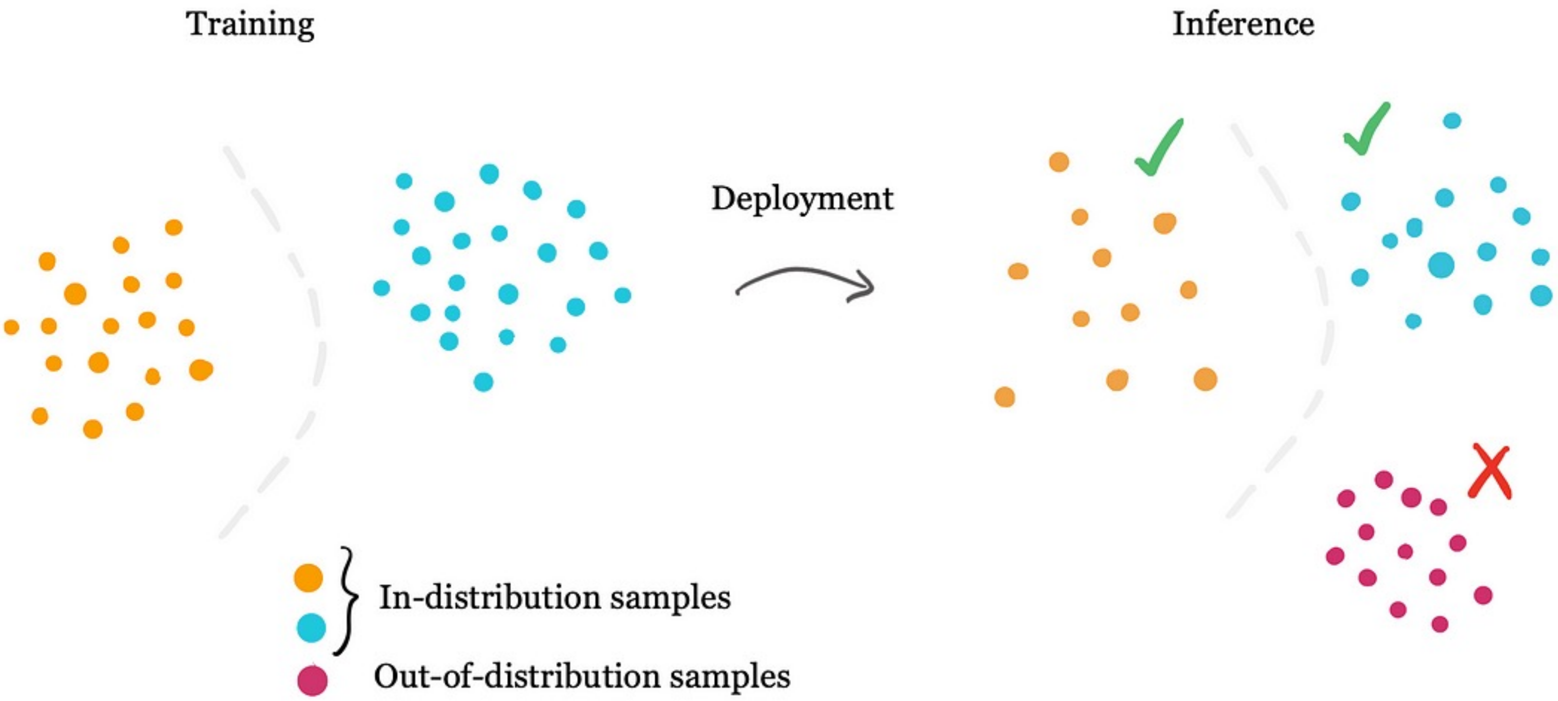
Unsupervised Learning : Clustering

- Given $x_1, \ldots, x_n$ (no labels), output hidden structure in x's
  - E.g., clustering



Things to Consider

Danger of Out-of-Domain Application



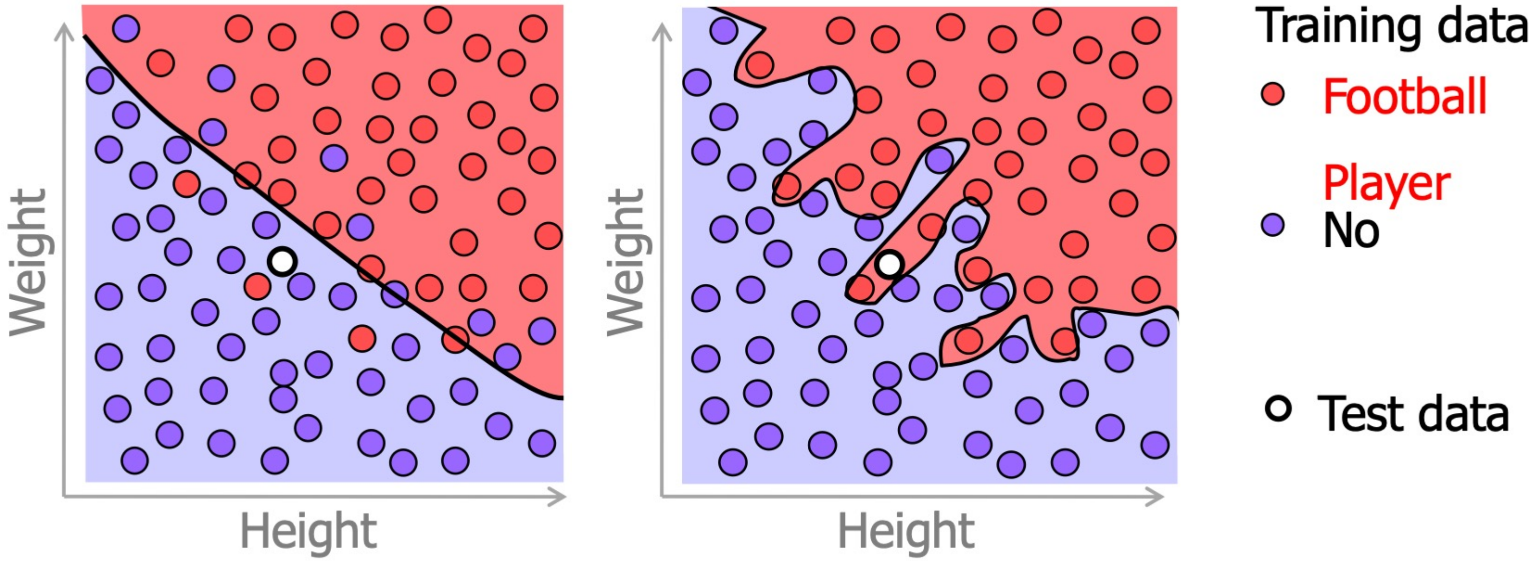
This can easily happen with high-dimensional data in ML algorithms.

Overfitting Problem

A good machine learning algorithm

- *Does not overfit training data*

- *Generalizes well to test data*



Ethical Consideration

- **편향과 차별**: What if we have biased data? Is it okay to learn the algorithm with this?
  - 입학/채용 서류절차에 ML 적용시: 과거의 인종/국적/성별에 대한 편향이 포함된 데이터로 훈련되었을 경우 특정 인종/국적/성별에 불이익을 줄 수 있음.

- **개인정보 보호**

- **안정성과 보안**
  - 시스템에 결함이나 취약점에 이해 예기지 않은 행동으로 인해 사고 발생 가능 (특히 Blackbox-type learning algorithm 을 사용할 때)

- **의사결정의 투명성과 설명가능성**
  - ML/AI 모델은 종종 '블랙 박스'로 작동하여, 그 결정 과정이 불투명
  - 예를 들어, 은행이 ML을 사용하여 대출 승인을 결정할 경우, 모델이 어떻게 그 결정에 도달했는지 설명하기 어려울 수 있고 이는 고객의 불만을 초래

Tentative list of topics and schedule

주차	날짜	주제	내용
1	7월 5일 (토)	Introduction	Introduction to various learning tasks
2	7월 12일 (토)	Regression 1	Linear methods for regression
3	7월 19일 (토)	Regression 2	Regularization & Feature selection
4	7월 26일 (토)	Model Assessment	Model selection & tuning methods with real examples
5	8월 2일 (토)	Classification 1	Introduction to Classification and Loss Functions
6	8월 9일 (토)	Classification 2	Various classification methods and application
7	8월 16일 (토)	기말고사	Take Home 과제 온라인 제출 (강의없음)

세부적인 강의 진행은 바뀔 수 있으며 변경 시 공지가 올라올 예정입니다.