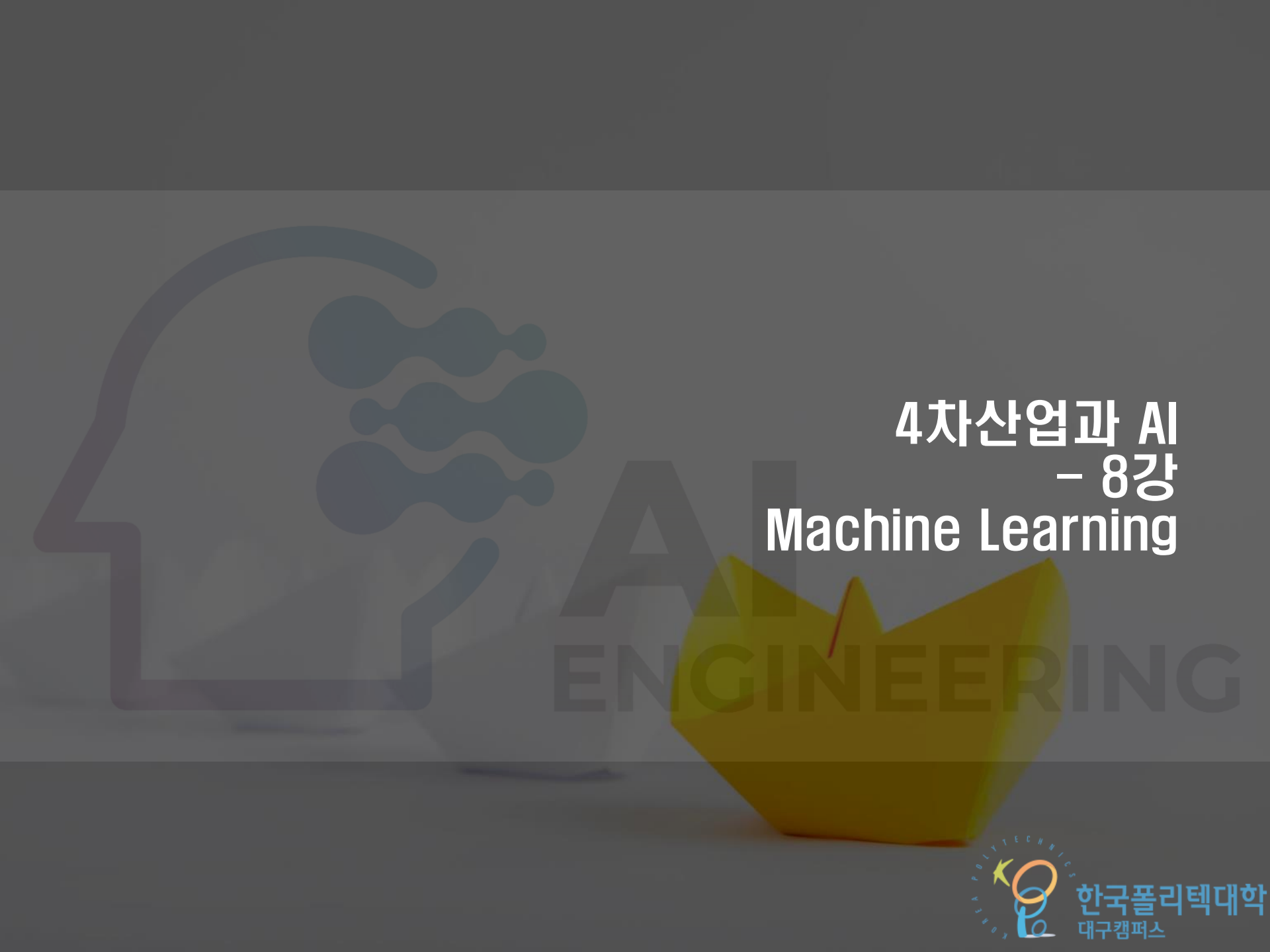


4차산업과 AI

한국폴리텍대학 대구캠퍼스
AI엔지니어링과 강현우



4차산업과 AI - 8강 Machine Learning



기계학습의 등장 배경

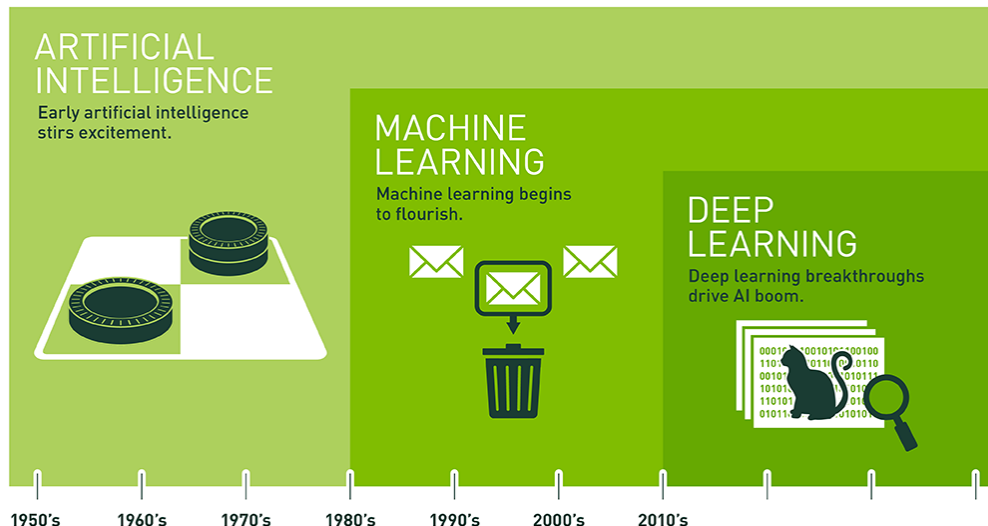
◆ 왜? 기계 학습?

- 1980년대 인공 지능 연구의 대표적인 방법
= 전문가 시스템
- 사람이 직접 많은 수의 규칙을 만드는 것을 전제
- 규칙을 정확하게 규정할 수 없는 분야는 어떻게?
- 사람조차 정확한 원리를 모르는 영역에 대해 요구

기계가 학습한다는 것?

◆ Machine Learning

- 어떤 컴퓨터 프로그램이 T라는 작업을 수행한다.
- 이 프로그램의 성능을 P 라는 척도로 평가했을 때
- 경험 E를 통하여 성능이 개선된다면
- 이 프로그램은 학습을 한다고 말할 수 있다.



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.

출처: Nvidia.com



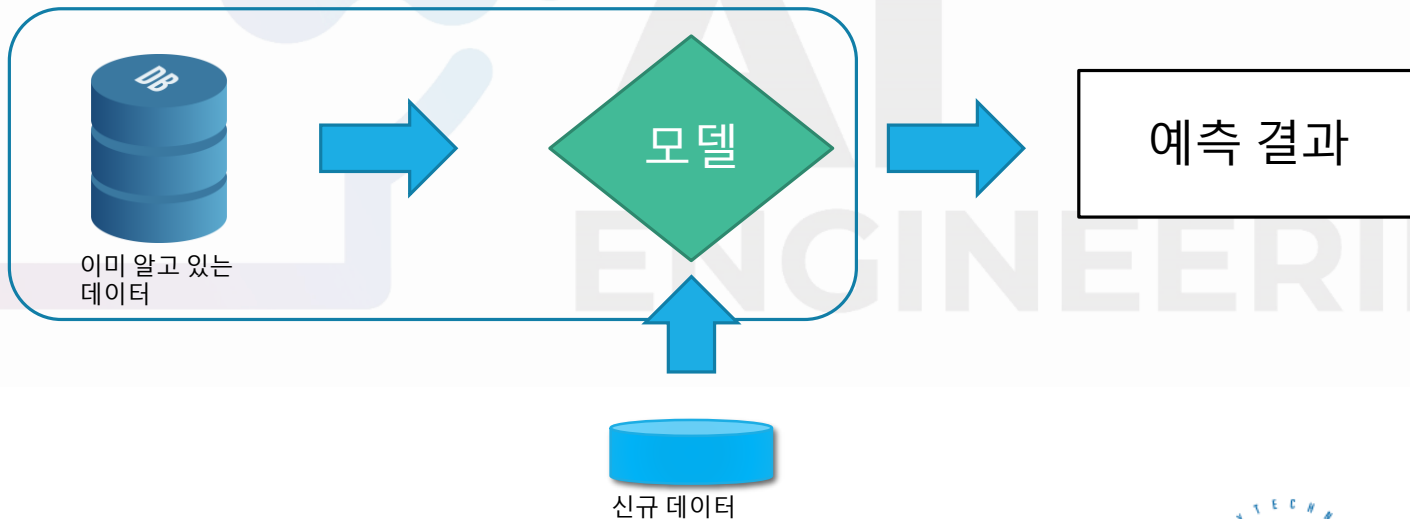
한국폴리텍대학
대구캠퍼스

기계 학습

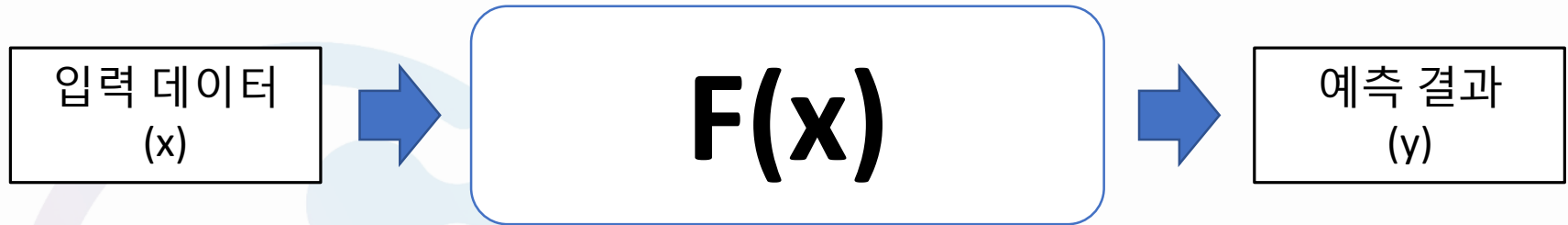
◆ Machine Learning

➤ 이미 알고 있는 데이터 (학습 데이터)로
모델을 생성해내는 과정

✓ 데이터에서 패턴을 추출하여 스스로 지식을 획득



기계 학습



- ◆ 과거에는 $F(x)$ 를 만드는데 집중
- ◆ 머신 러닝은 알고있는 데이터 x 와 결과 y 로 $F(x)$ 를 만들어 내는 것

So What?

◆ 그래서, 무슨 문제를 풀고 싶은 건데?

➤ 분류 (Classification)

➤ 군집화 (Clustering)

➤ 회귀 (Regression)

➤ ...

➤ 세상은 넓고 문제는 많다.

세상은 넓고 문제는 많다

◆ Kaggle

- <https://www.kaggle.com/>
- 예측 모델 및 분석 대회 플랫폼

◆ 현재 진행 중 Competitions

- NFL 건강 및 안전 – \$100,000
- NFL 빅 데이터 볼 2022 – \$100,000
- ...

기계 학습 준비

◆ 어떤 문제를 머신 러닝으로 풀고 싶다면

➤ 어떤 부류의 문제인지 파악

➤ **데이터 세트**

- ✓ 학습 데이터
- ✓ 테스트 데이터
- ✓ Optional – 검증 데이터

➤ **모델을 설계**

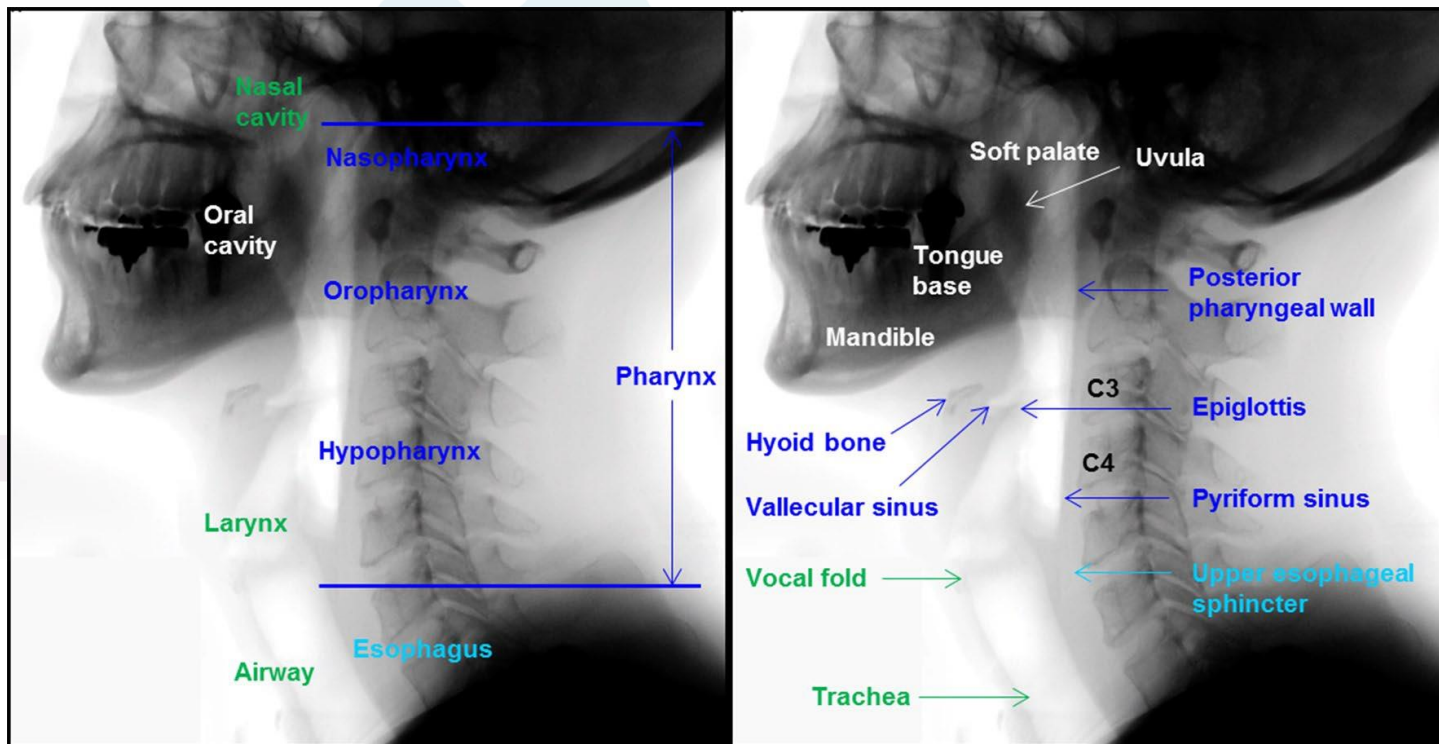
- ✓ 알고리즘



Domain Knowledge

◆ Engineer 는 엔지니어...?

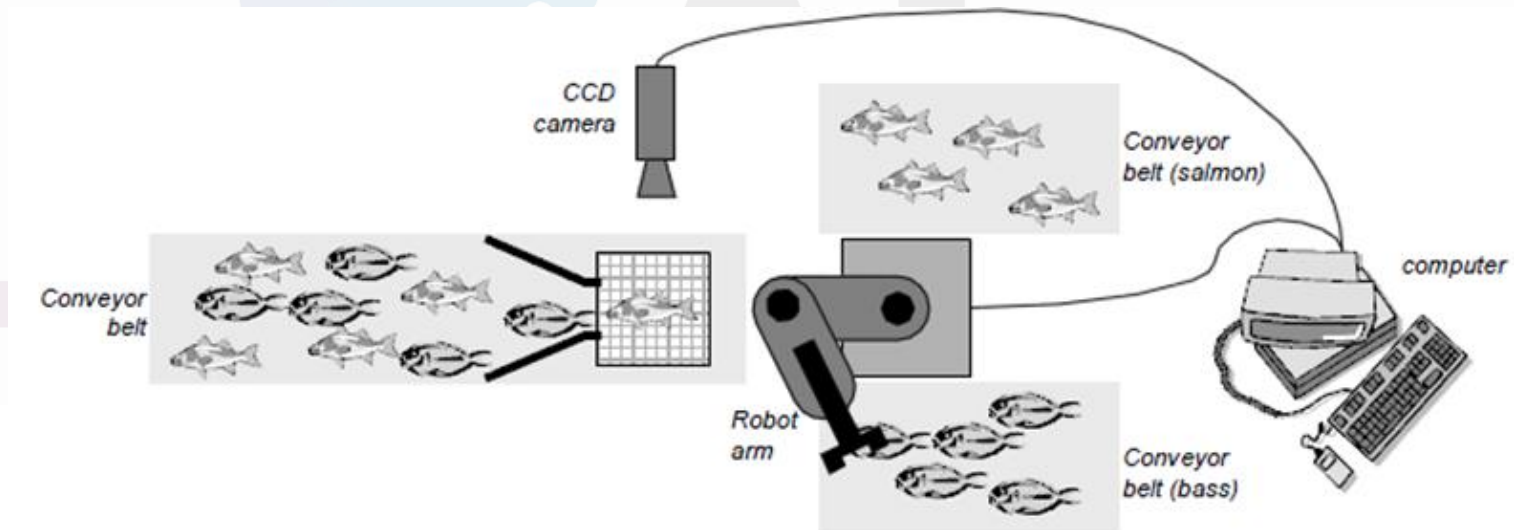
- 이 사진은 뭐죠? → 논문
- Kaggle에서 제공되는 것이 뭐였죠?



기계 학습 예제

◆ 시나리오

- 생선처리 공장에서 연어(salmon), 농어(sea bass)를 분류
- CCD 카메라를 갖춘 비전 시스템
- 영상을 분석하여 로봇 암을 제어하여 생선을 이동



기계 학습 설계

◆ 데이터 수집



◆ 전처리



◆ 특징?

➤ 길이, 밝기 ... Domain 지식이 없으니까...

◆ 분류기 설계

➤ 모델 선정, 분류기 학습

◆ 성능 테스트

➤ 학습에 사용하지 않은 데이터 사용

모델 설계

◆ Domain Knowledge

➤ 농어(sea bass)는 연어(salmon)보다 일반적으로 길다

◆ 선정 특징: Length

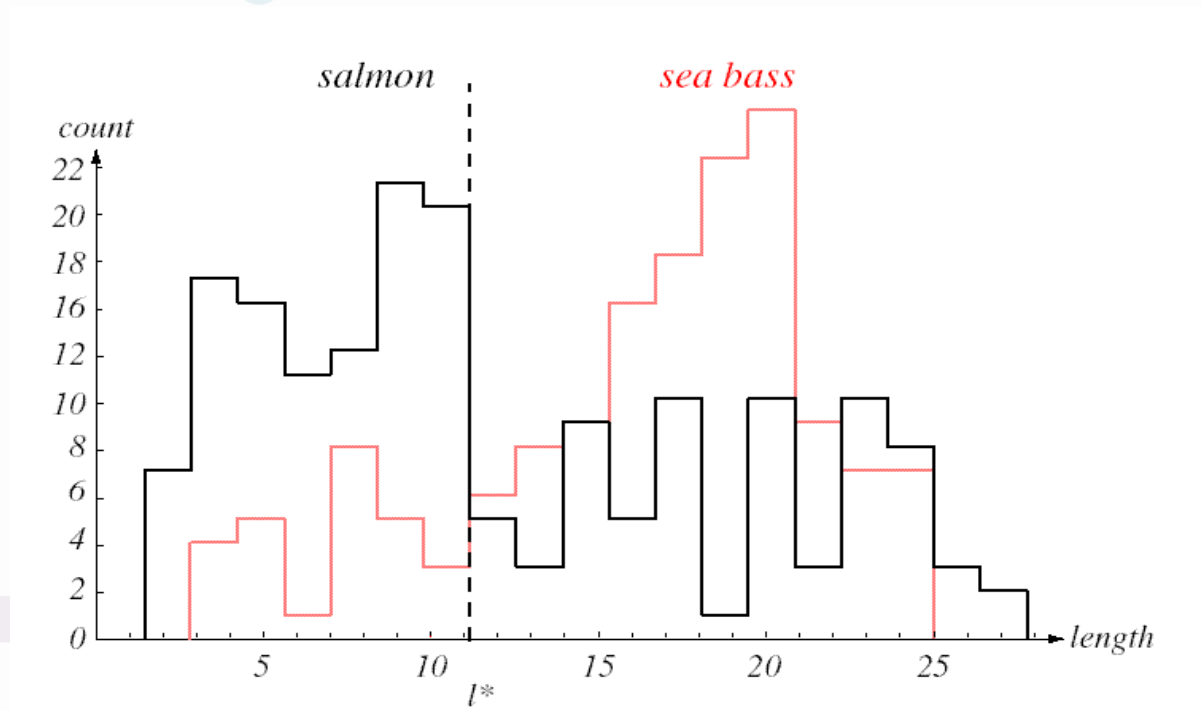
◆ 분류 규칙

If *Length* $\geq l^*$ then *sea bass*
otherwise *salmon*

◆ l^* 를 고르는 방법?

모델 학습

◆ 두 생선에 대한 Length 히스토그램



오분류가 제일 적은 지점

Training error: $90 / 316 = 28\%$



학습 결과

◆ 실험 결과

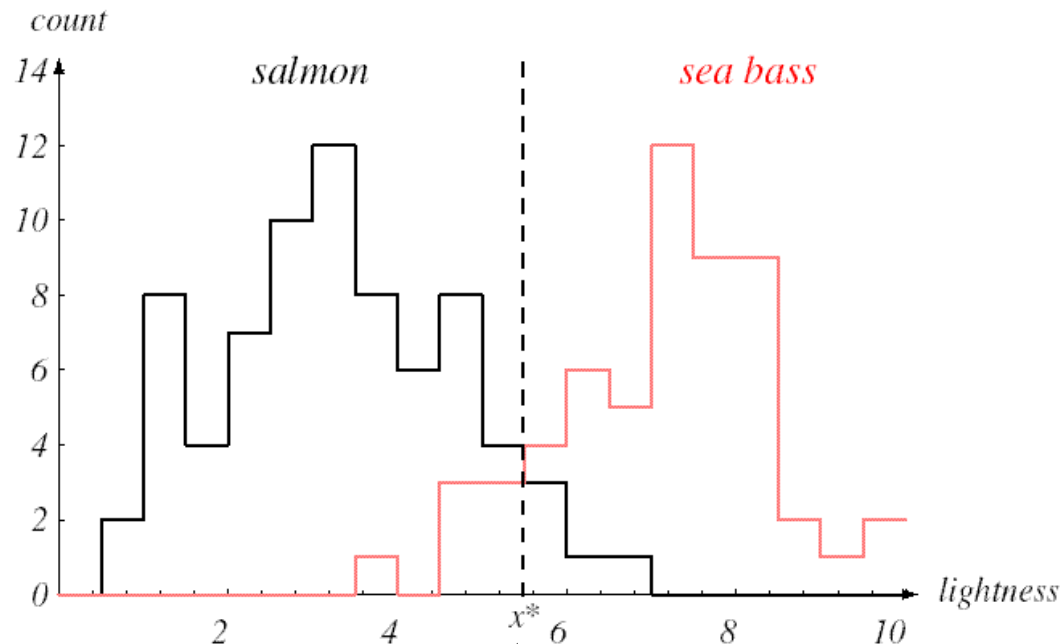
- 학습 데이터에 대한 분류율: 28%
- 너무 낮다!!!
- 다른 특징을 시도

◆ 밝기?

- New Feature → Lightness

모델 학습

◆ 두 생선에 대한 Lightness 히스토그램



오분류가 제일 적은 지점 Training error: $16 / 316 = 5\%$

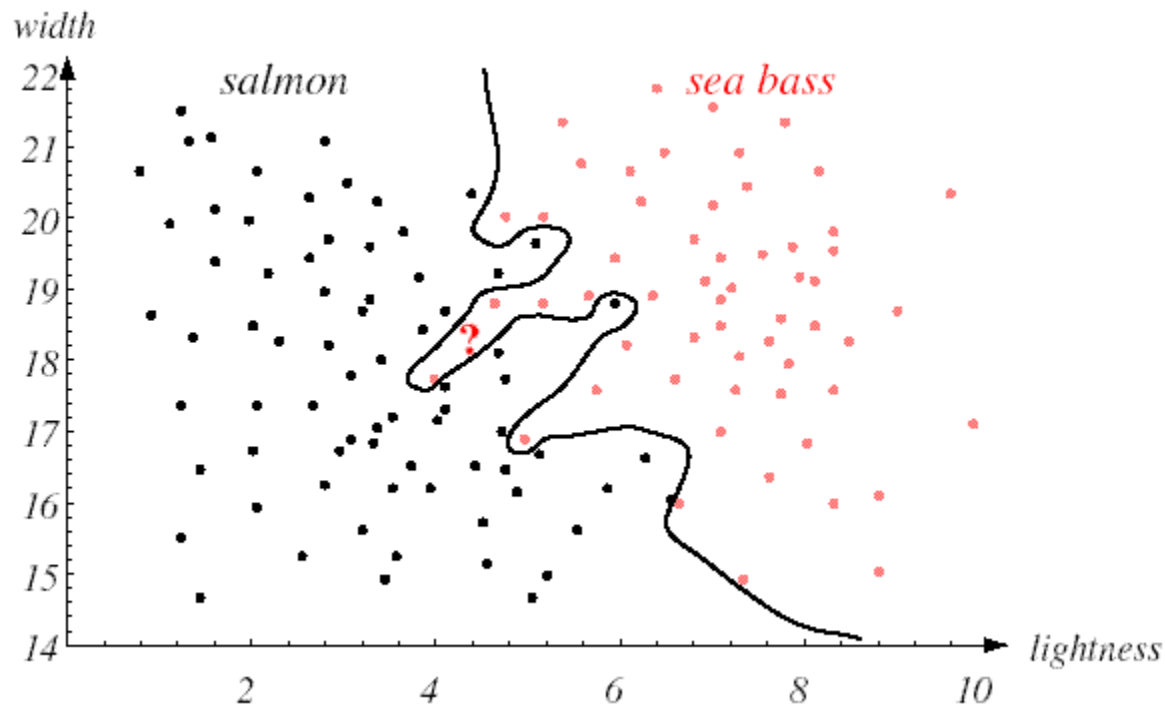
분류가 아까 보다 잘 되었음!

최선입니까?

- ◆ 단일 특징으로는 만족스럽지 못하다
- ◆ 복합 특징을 사용
 - Sea bass 가 salmon 보다 보통 폭이 넓다
- ◆ 일반적으로 특징 공간의 차원이 높을 수록
 - 분리에 유리
- ◆ 보다 복잡한 분류 모델을 사용할 필요

분류 함수를 좀 더 복잡하게?

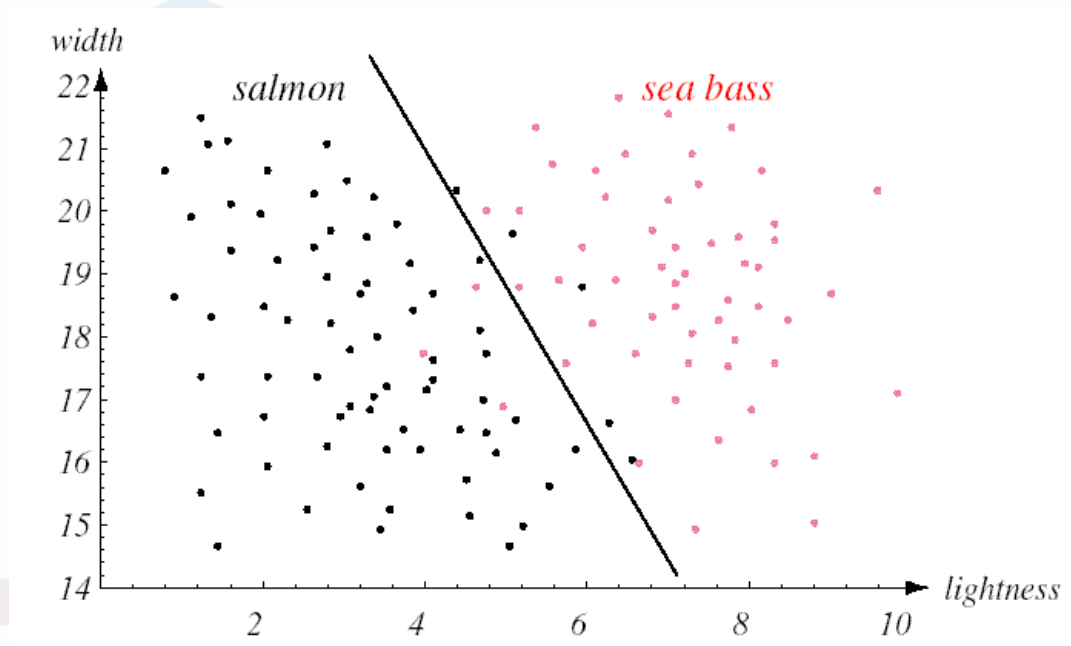
◆ 학습 데이터를 완벽하게 분류하는 모델



Complex decision function
Training error: $0 / 316 = 0\%$

일반화

◆ 앞 슬라이드의 분류 모델은 Overfitting



Linear decision function

Training error: $8 / 316 = 2.5\%$



Overfitting

◆ Overfitting 은 왜 생기나?

- 우리가 가지고 있는 데이터는 전체 데이터 중 얼마나 될까?
- 얼마나 일반화가 잘 되었는지는 어떻게 알지?
- 우리가 가지고 있는 데이터가 무진장 많다면?

모델 – ML Algorithm

◆ 분류

- 지도 학습 (Supervised Learning)
- 비지도 학습 (Unsupervised Learning)
- 강화 학습 (Reinforcement Learning)

◆ Algorithm

- Support Vector Machine (SVM), Bayesian Network, Decision Tree, Random Forest, Neural Network...

◆ Deep Neural Network (심층 신경망)



ML Application

◆ 컴퓨터 비전

- 컴퓨터에서 카메라 등의 시각 매체를 통해 입력 받은 영상을 분석하여 유용한 정보를 생성하는 기술
[ex. 보행자 검출, 얼굴 인식, 번호판 인식 등등]

◆ 데이터 마이닝

- 데이터 베이스 내에서 유용한 정보를 발견하는 기술
[ex. 상품 추천, 마케팅]

◆ 자연어 처리

- 컴퓨터를 이용해 사람의 자연어를 분석하고 처리하는 기술
- 대량의 말뭉치 데이터를 활용하는 기계 학습 기반의 자연어 처리 기법이 주류

Summary

◆ 기계 학습이란?

- 데이터를 이용하여 일반화된 모델을 만드는 것
- 데이터, 알고리즘이 필요

◆ Domain Knowledge

- 해결하고자 하는 문제가 속한 분야의 지식

◆ Overfitting

- 학습데이터에 과하게 적합된 모델