영등포구

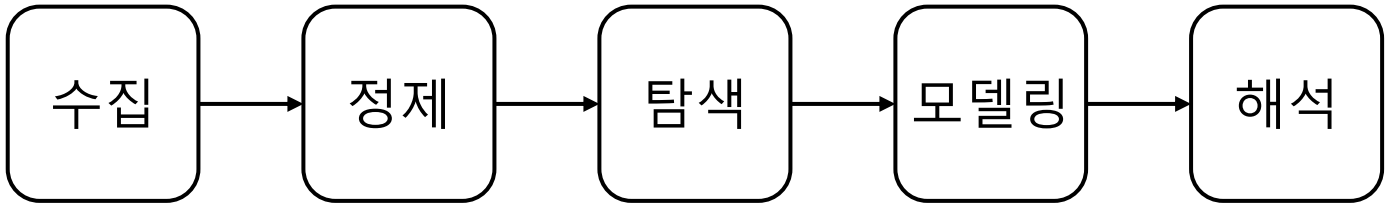
영등포구 청년을 위한

생성형AI 활용 데이터 시각화 교육

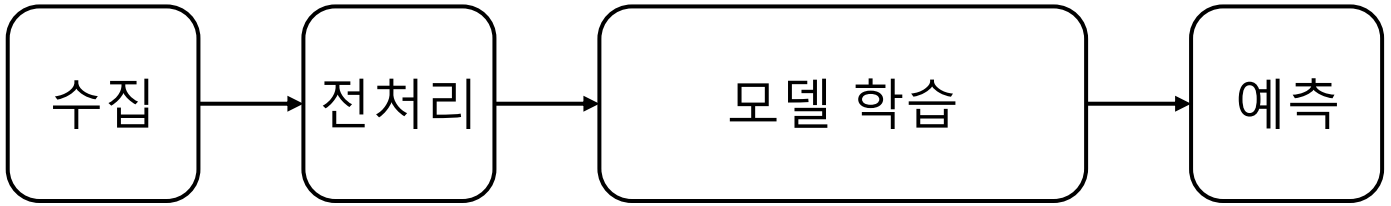
AI·머신러닝 활용 데이터 시각화

데이터 분석 VS. 머신러닝(딥러닝)

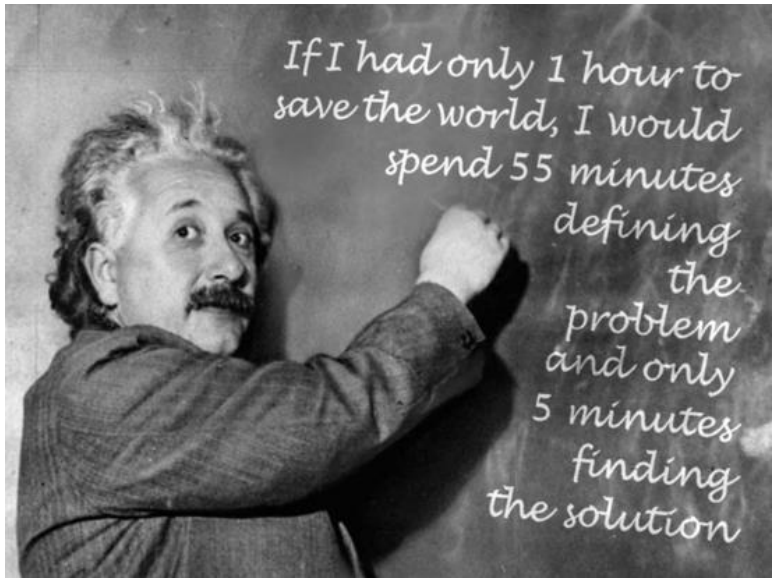
- **데이터 분석(사람)**
  - 데이터 속에서 의미를 찾아냄
  - 어떤 결과를 찾기 위해서 수행



- **머신러닝/딥러닝(기계)**
  - 데이터 속에서 찾아낸 의미를 활용
  - 결과를 내기 위해서 수행



머신러닝 프로세스



나에게 1시간이 주어진다면,

문제가 무엇인지 정의하는데 55분의 시간을 쓰고,

해결책을 찾는데 나머지 5분을 쓸 것이다.

- 알베트 아인슈타인

머신러닝 프로세스(문제 정의)

- **지도학습 VS 비지도학습**
  - 데이터에 정답이 쓰여져 있는가 여부에 따라 지도와 비지도로 나뉨

	Feature 1	Feature 2	...	Feature 10	스팸메일 여부
Data1	0	0	...	1	N
	0	1	...	1	Y
	1	0	...	0	Y
	1	1	...	0	N

	Feature 1	Feature 2	...	Feature 10
Data2	0	0	...	1
	0	1	...	1
	1	0	...	0
	1	1	...	0

머신러닝 프로세스(문제 정의)

- **분류(Classification) : 데이터를 나누는 것(지도 학습)**
  - 스팸메일 여부(이항 분류)
  - 스팸메일의 종류(다항 분류)

- **회귀(Regression) : 연속적인 숫자 변수들 간의 상관관계 파악(지도 학습)**
  - 음식점 매출 예측
  - 주식가격 예측

- **군집화(Clustering) : 데이터를 유사한 특성 끼리 묶음(비지도 학습)**
  - 관심사나 취미에 따른 사용자 그룹 묶음

머신러닝 프로세스(특징 엔지니어링)

- **모델에 데이터를 입력하기 전, 데이터 특성을 잘 반영하여 성능을 높일 수 있도록 특징(Feature)을 생성하고 가공하는 것**

✓ 특징 선택(Feature Selection)

- 여러 개의 특징 중에서 데이터의 특징을 가장 잘 나타내는 주요 필드 몇개 만을 선택하여 사용

✓ 특징 추출(Feature Extraction)

- 여러 개의 원본 특징 들을 조합하여 새로운 특징을 생성

- 대표적인 방법으로 주성분 분석(PCA)

✓ 특징 생성, 구축(Feature Generation, Construction)

- 데이터에 대한 도메인 지식을 바탕으로 데이터를 합치거나 쪼개는 등의 과정을 통해 새로운 Feature를 생성

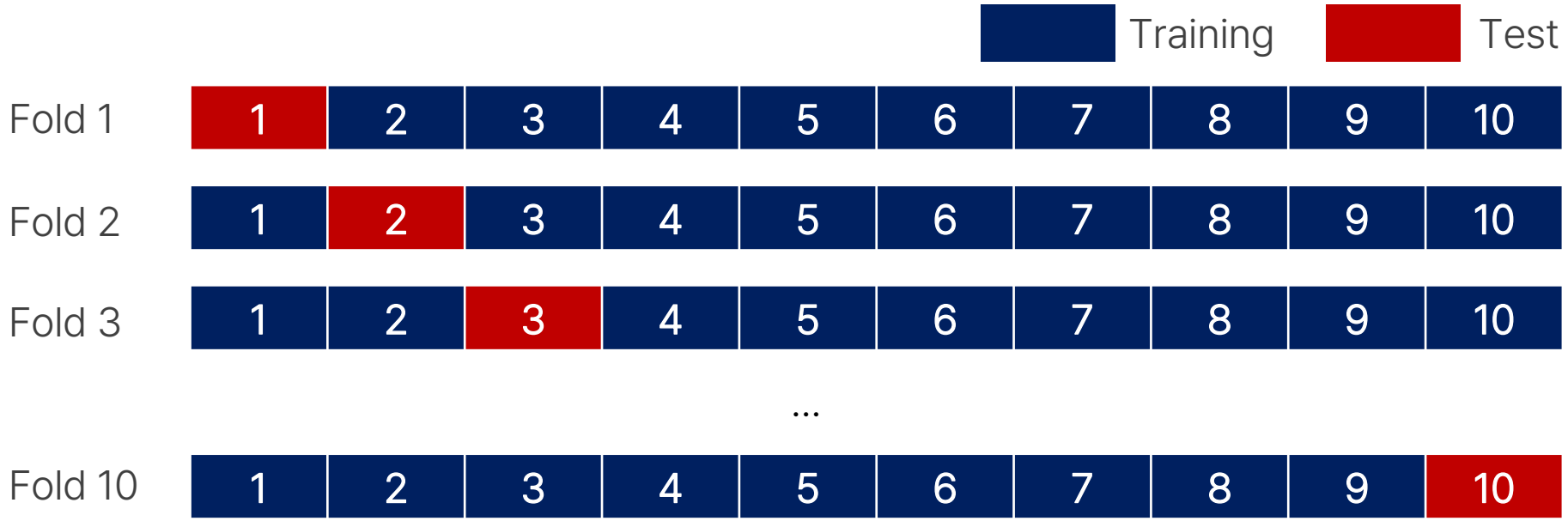
머신러닝 프로세스(학습)

- **모델을 학습하기 전 데이터를 학습(Train) 데이터와 테스트(Test) 데이터로 분할**
  - 모델 검증을 위해서 데이터를 나눠서 사용함
  - 성능을 높이기 위해 K-Fold Validation 등 다양한 기법 이용

머신러닝 프로세스(학습)

▪ K-Fold Cross Validation

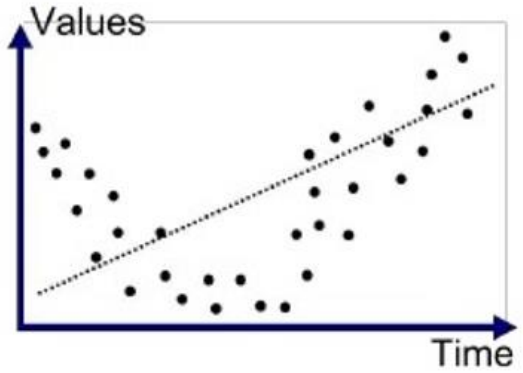
- 많은 데이터가 없을 경우, 한정된 데이터를 최대한 효율적으로 사용할 수 있는 방법
- 모든 데이터를 모델 검증에 활용 할 수 있음
- 단, 시간과 비용이 많이 든다는 단점



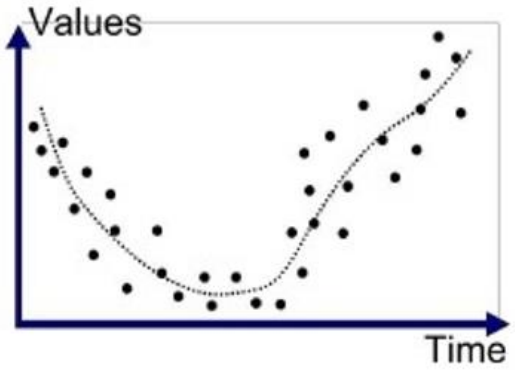
머신러닝 프로세스(학습)

▪ 일반화(Generalization)

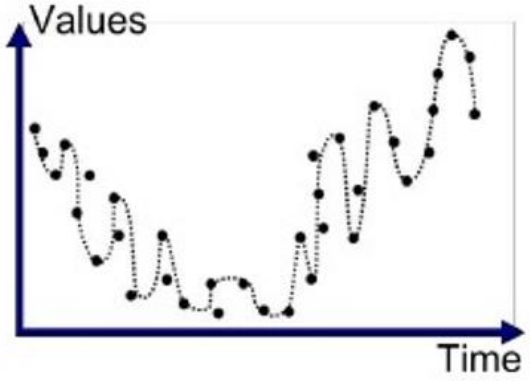
- 모델은 가능한 한 일반화 시키는게 좋음
- 오버피팅(Overfitting) : 학습 데이터에 딱 맞춰서 모델이 학습 된 경우
- 언더피팅(Underfitting) : 모델 학습이 덜 된 경우



Underfitted



Good Fit/Robust



Overfitted

머신러닝 프로세스(검증)

▪ Confusion Matrix(분류 문제 한정)

- Accuracy(탐지율) : P,N을 맞게 예측한 비율 - (TP + TN) / (TP + TN + FP + FN)
- Precision(정확도) : P로 예측한 것 중 실제 P의 비율 - TP / (TP + FP)
- Recall(재현율) : 실제 P를 P로 예측한 비율 - TP / (TP + FN)

Confusion Matrix		ACTUAL	
		Positive	**Negative**
PREDICT	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

머신러닝 프로세스(검증)

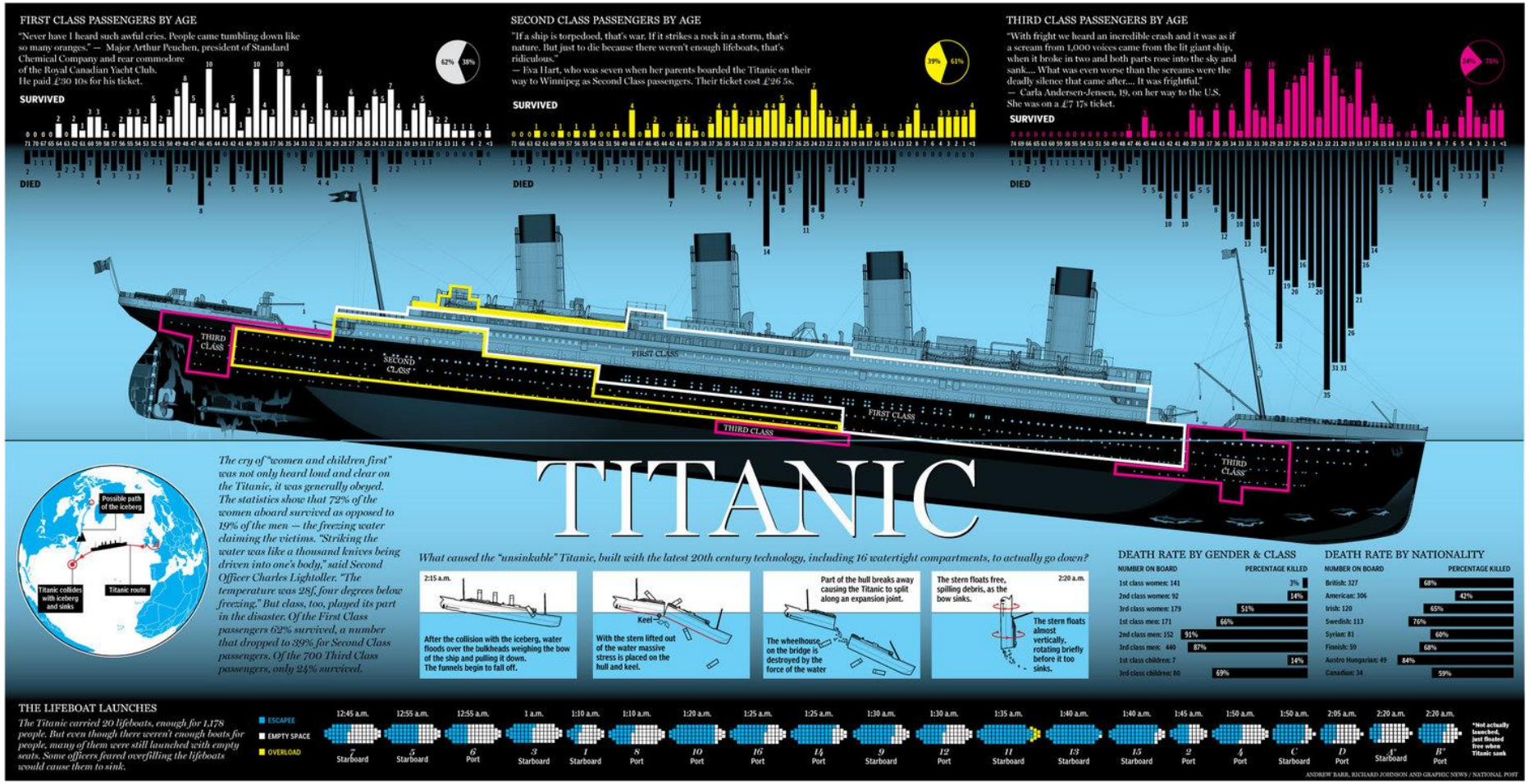
▪ Confusion Matrix

- FPR(False Positive Rate) : 실제 N인데 P로 예측한 비율 - FP / (FP+TN)
- FNR(False Negative Rate) : 실제 P인데 N으로 예측한 비율 - FN / (TP+FN)

Confusion Matrix		ACTUAL	
		Positive	**Negative**
PREDICT	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

타이타닉 생존자 예측하기

- 타이타닉 호에 탑승했던 사람들의 정보를 바탕으로 생존자를 예측



감사합니다.