



Introduction to Business Analytics: Machine Learning Algorithms

강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

AGENDA

01 빅데이터 분석 개요 및 주요 개념

02 데이터 과학 프로젝트 절차

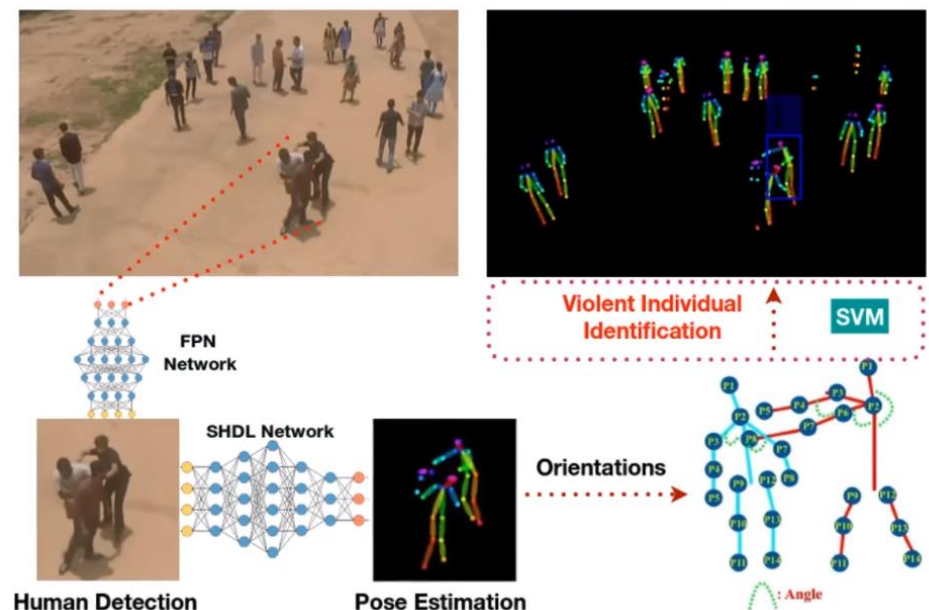
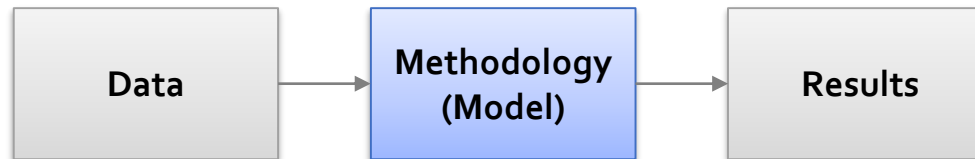
03 기계 학습 방법론

04 제조업 활용 사례 1: 가상 계측 모델 개발

머신 러닝: Machine Learning

- Machine Learning

- ✓ 특정한 과업^{Task}을 달성하기 위해 경험^{Experience}이 축적될수록 과업 수행의 성능^{Performance}이 향상되는 컴퓨터 프로그램 또는 에이전트를 개발하는 것 – Mitchell (1997)

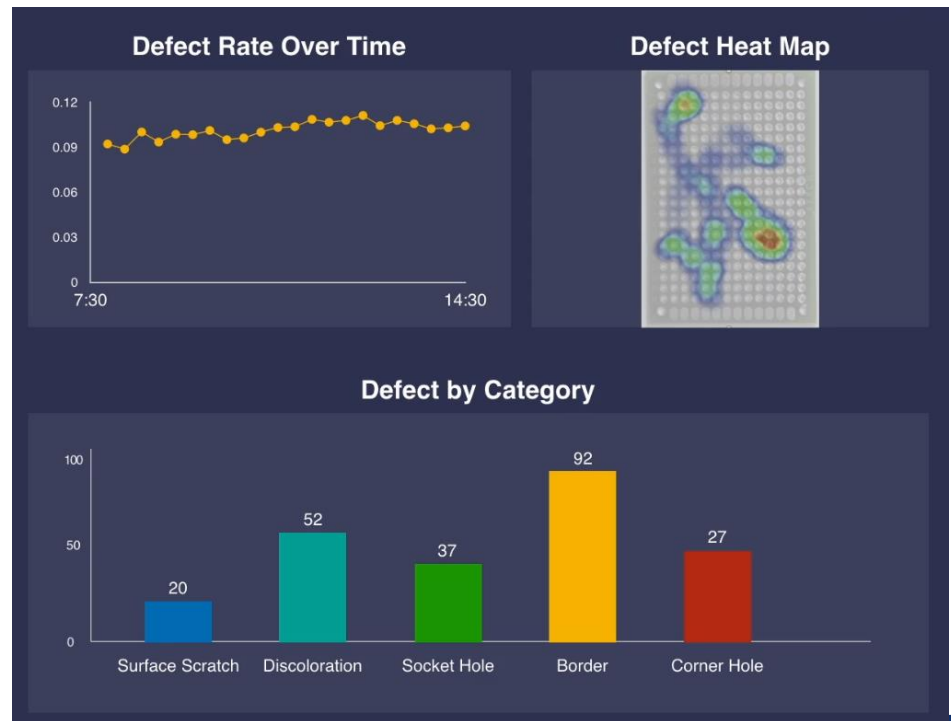


머신 러닝: Machine Learning

• 제조업에서의 머신러닝

✓ Landing.ai

- 인공지능 분야의 세계적 권위자인 Andrew Ng 교수가 인공지능의 제조업 적용을 목표로 세운 스타트업 (대만 폭스콘과 제휴)
- 제품 이미지를 바탕으로 불량 판정 및 불량 의심 영역 판독



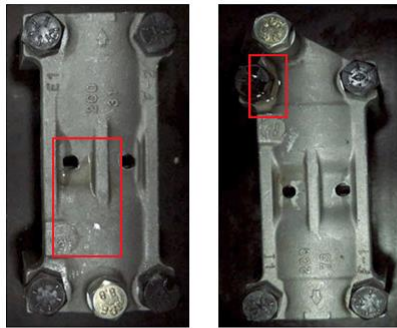
머신 러닝: Machine Learning

• 제조업에서의 머신러닝

✓ Landing.ai

Leak Defect Detection

LIQUID LEAK



Capabilities:

- Detect liquid leaks
- Tiny/large, slow/quick leaks
- Require only 1-10 defect images, with our SMALL DATA TECHNOLOGY

Applications:

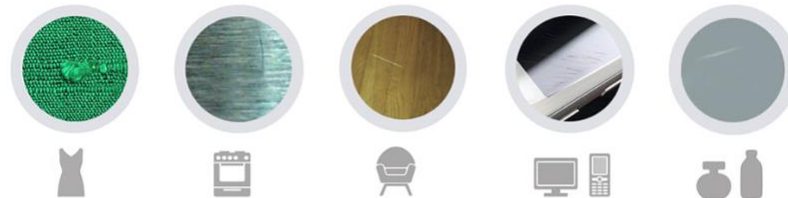
- Automobile: engines, tanks, mufflers
- Storage: chemical, oil, water leaks
- Inspection: pipeline, refinery, drilling

Business Value:

- Improve quality, revenue, safety, compliance

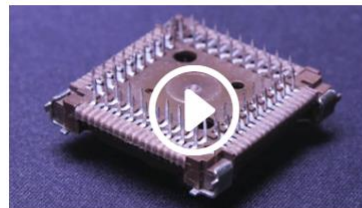
Surface Defect Detection

MULTIPLE MATERIALS

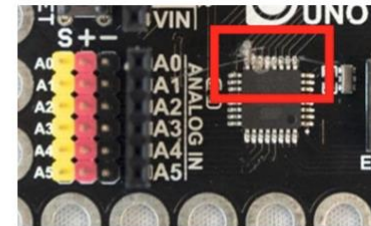


ELECTRONIC COMPONENTS

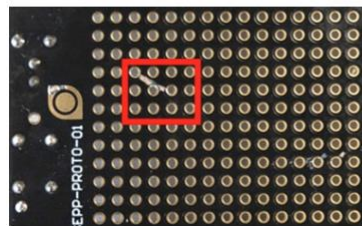
SCRATCH



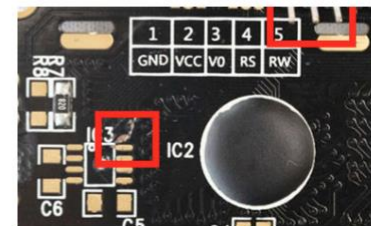
SOLDERING



SOCKET



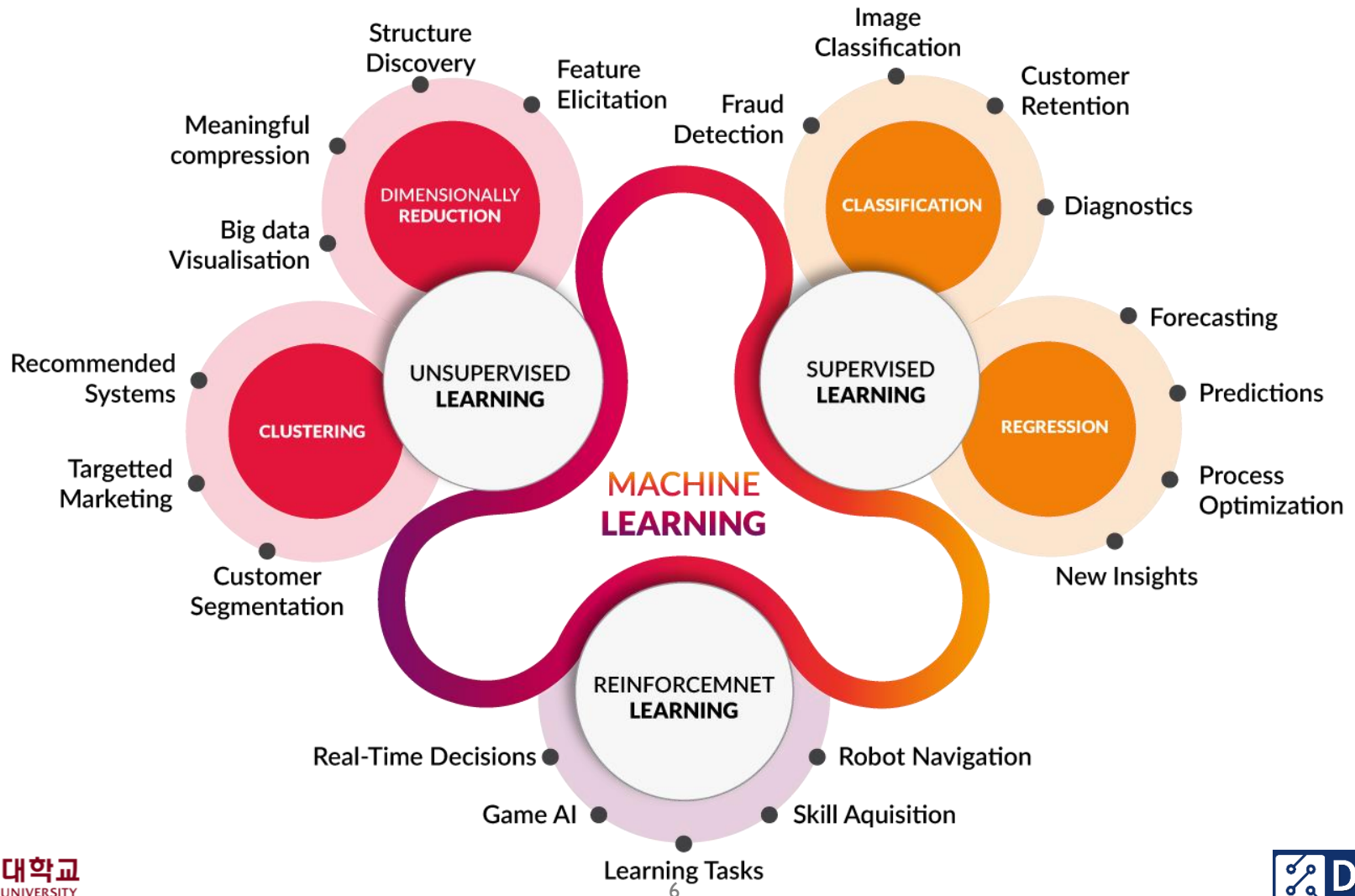
BENDING



- Reduce waste & labor

머신 러닝의 종류

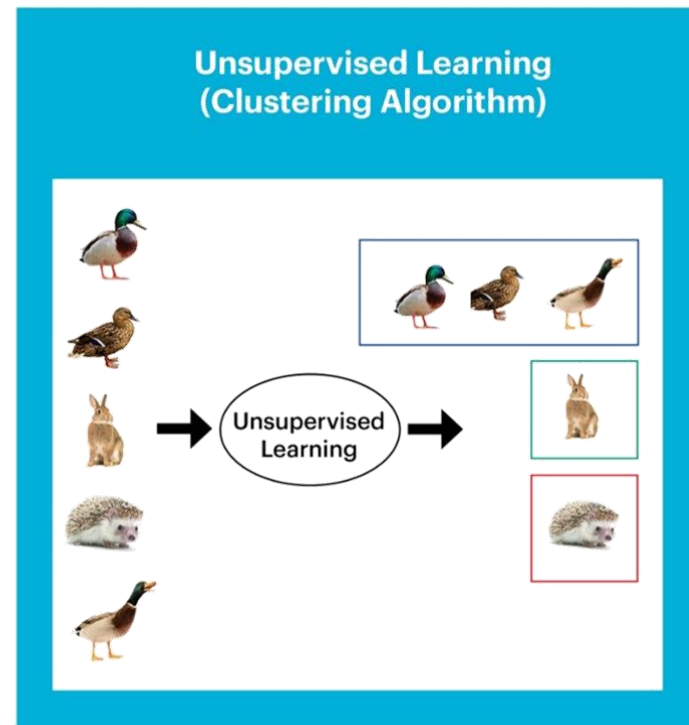
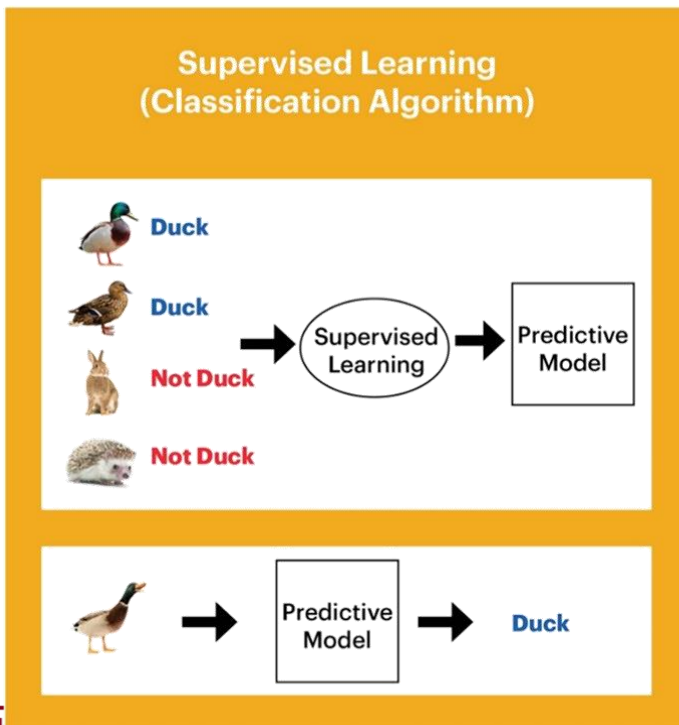
- Machine Learning의 종류



머신 러닝의 종류

- 구분기준 1: Target (정답)의 유무에 따른 구분

- ✓ **Supervised learning (지도 학습)**: 입력과 출력 변수가 정해져 있고 둘 사이의 관계를 규명하는 것을 주 목적으로 하는 학습
- ✓ **Unsupervised learning (비지도 학습)**: 출력 변수가 없는 데이터의 특질이나 특성을 파악하는 것을 주 목적으로 하는 학습



머신 러닝의 종류

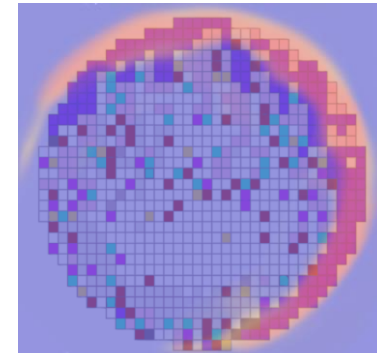
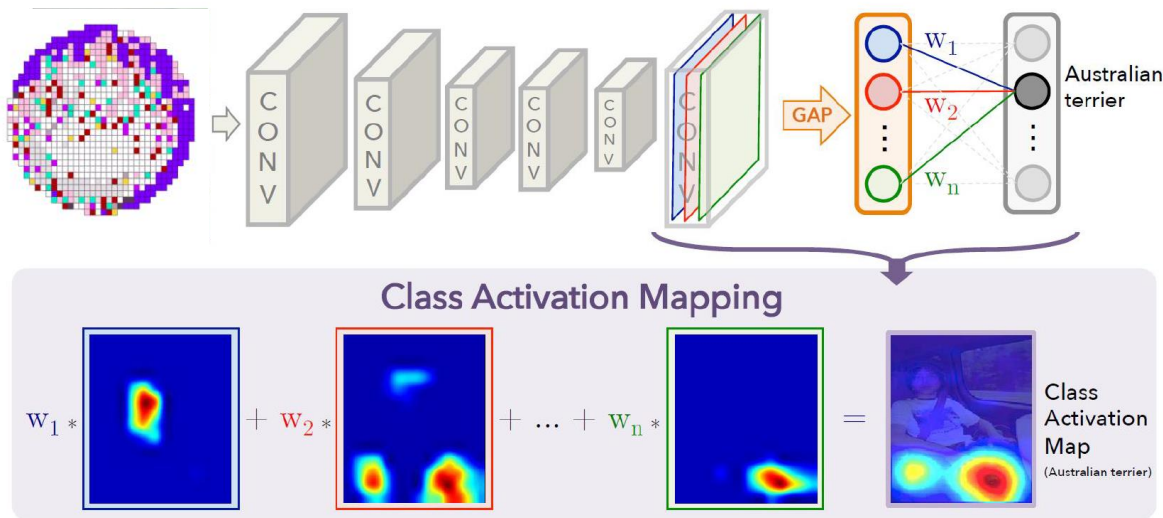
- 구분기준 1: Target (정답)의 유무에 따른 구분

✓ **Supervised learning**: Wafer별 불량 유무에 대한 Label 정보를 알고 있음

입력: WBM

머신러닝 알고리즘: 합성곱 신경망

출력: 불량 유무 및 영역



머신 러닝의 종류

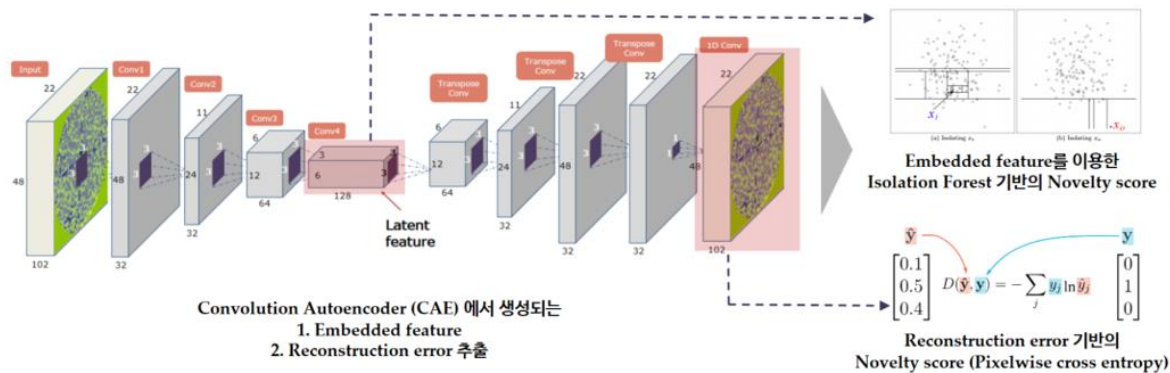
- 구분기준 1: Target (정답)의 유무에 따른 구분

✓ **Unsupervised learning**: Wafer별 특이 웨이퍼 유무 Label이 없음

입력: WBM

머신러닝 알고리즘: 합성곱 신경망

분석 결과: 특이 웨이퍼



오늘 생성된 WBM중에서
이상치는 어떤 것일까?

오늘 생성된 WBM중에서
과거 WBM 패턴으로 보았을 때,
이상치는 어떤 것일까?

오늘 생성된 WBM중에서
과거 WBM 패턴으로 보았을 때,
WBM중 어느 칩에서 이상치가
크게 발생 하였을까?

머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

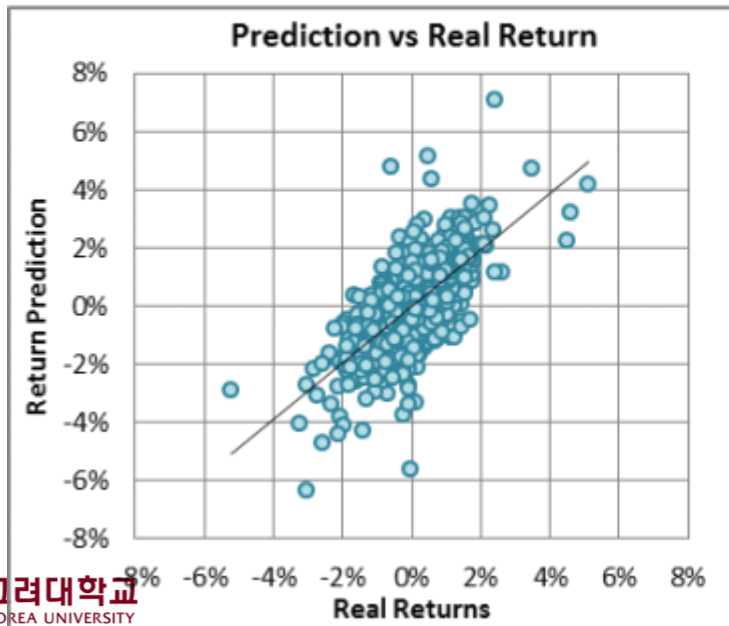
- ✓ Classification (분류) vs. Regression (회귀)

- **Classification**: 명목형(categorical) 변수를 예측하는 방법론 (예: 웨이퍼 단위 불량/정상 유무 (good/bad))
- **Regression**: 연속형(continuous) 변수를 예측하는 방법론 (예: 웨이퍼별 수율 (0~100%))

Regression

vs

Classification

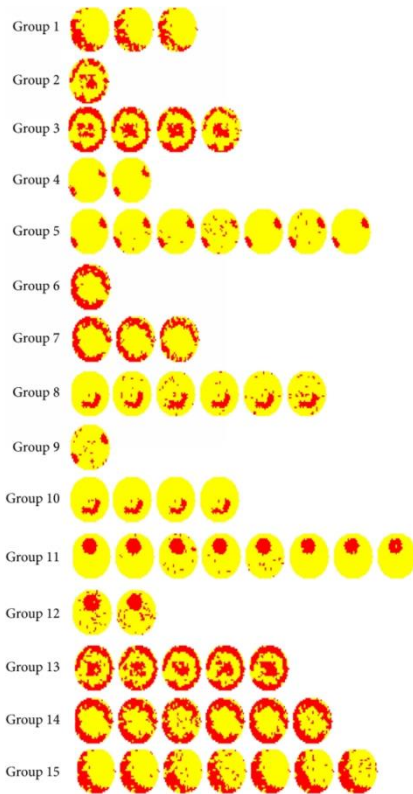


머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 군집화(Clustering)

- 유사한 개체들의 집단을 판별하는 방법론
 - K-평균 군집화, 계층적 군집화 등



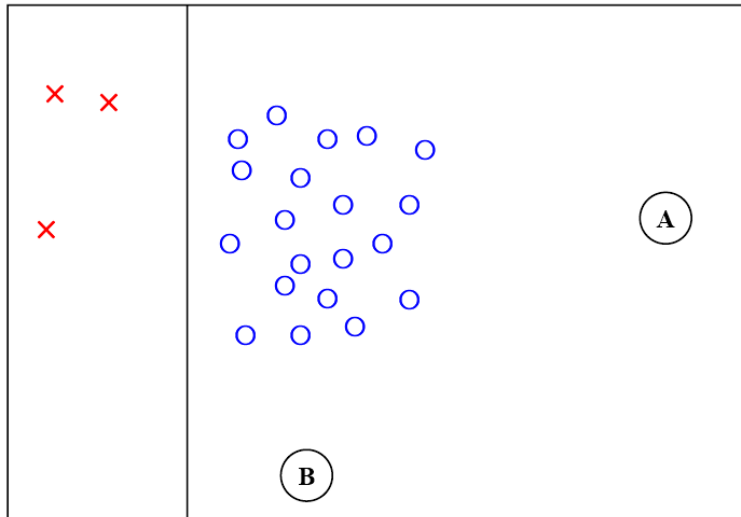
Pattern	Original Map	ESRN ($\rho_1=4, \rho_2=6$)	ESRN ($\rho_1=4, \rho_2=5$)	ESRN ($\rho_1=5, \rho_2=6$)	ESRN ($\rho_1=5, \rho_2=5$)
Checkerboard					
Ring					
Right-Down Edge					
Composite Pattern					

머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

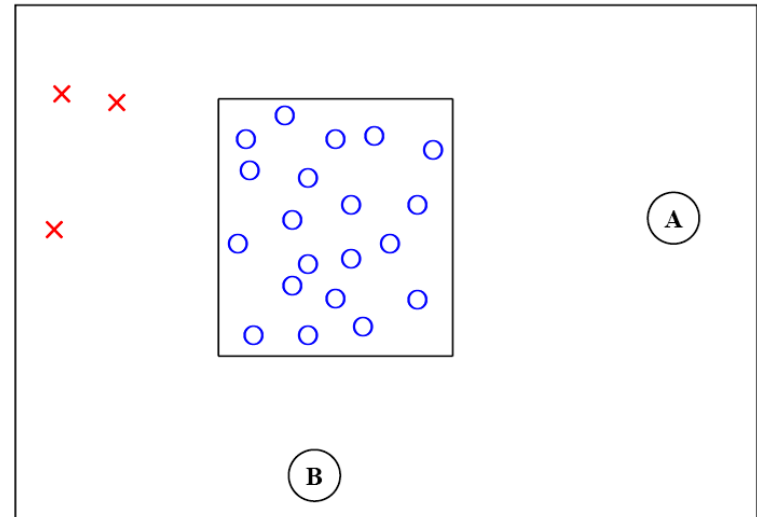
- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection)

- 대부분이 정상 데이터인 상황에서 매우 낮은 확률로 발생하는 이상치 데이터를 탐지하는 방법론 (예: 반도체 공정의 불량 웨이퍼 탐지)



Binary classification

두 범주 중 하나의 범주로 할당



Novelty detection

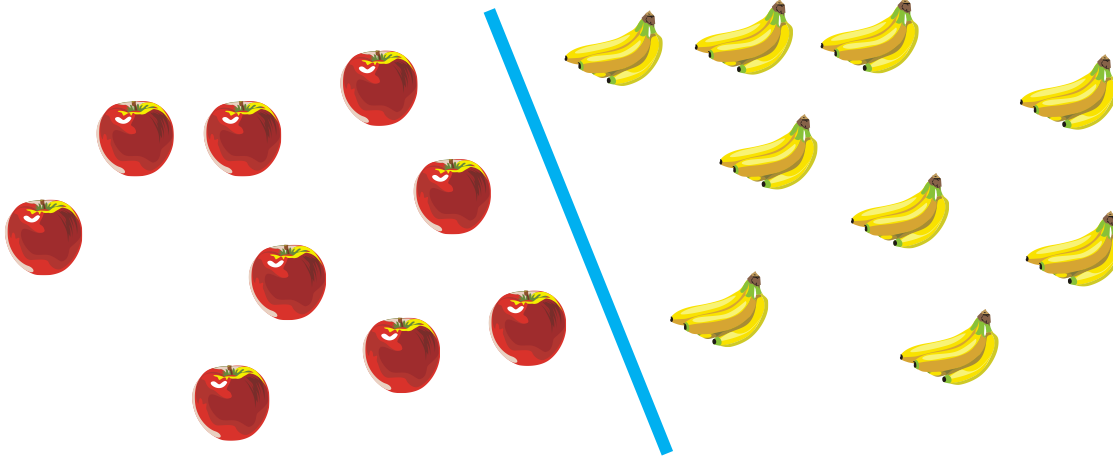
정상 범주에 속하는지 아닌지를 판단

머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection)

- 분류 알고리즘이 학습하는 방식

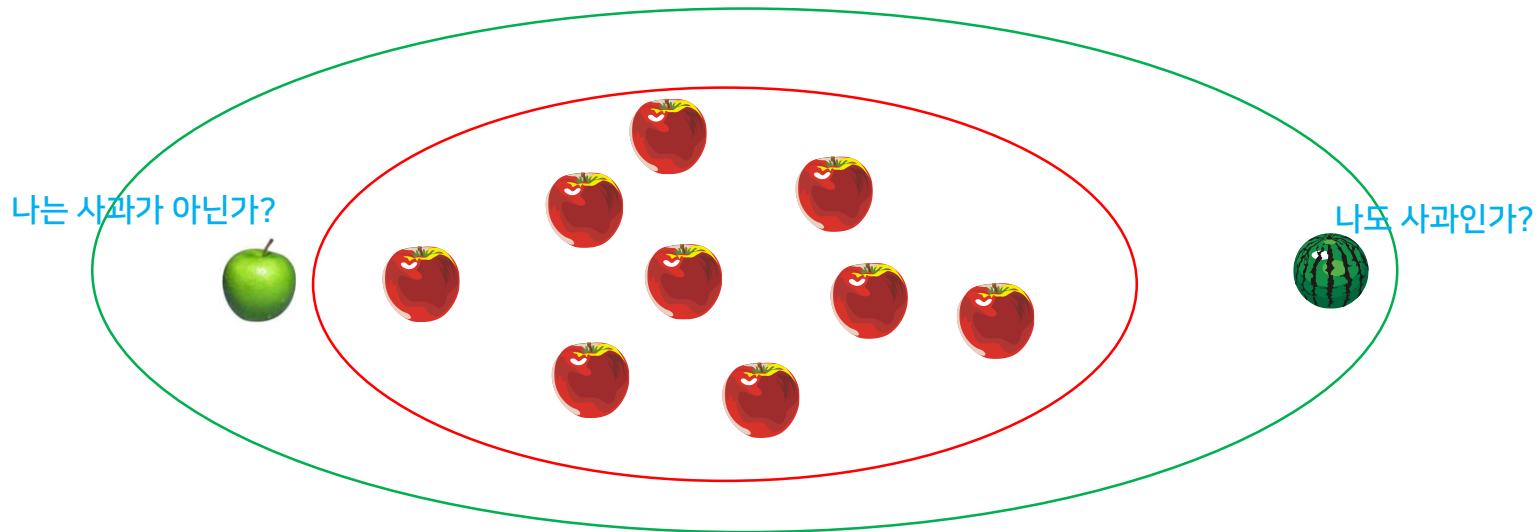


머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection)

- 이상치 탐지 알고리즘이 학습하는 방식
- 사과(normal)와 사과가 아닌 것(abnormal)을 구분하라
- 기준 1: 동그란 과일은 사과
- 기준 2: 동그란 과일이면서 색깔이 빨간 과일이 사과

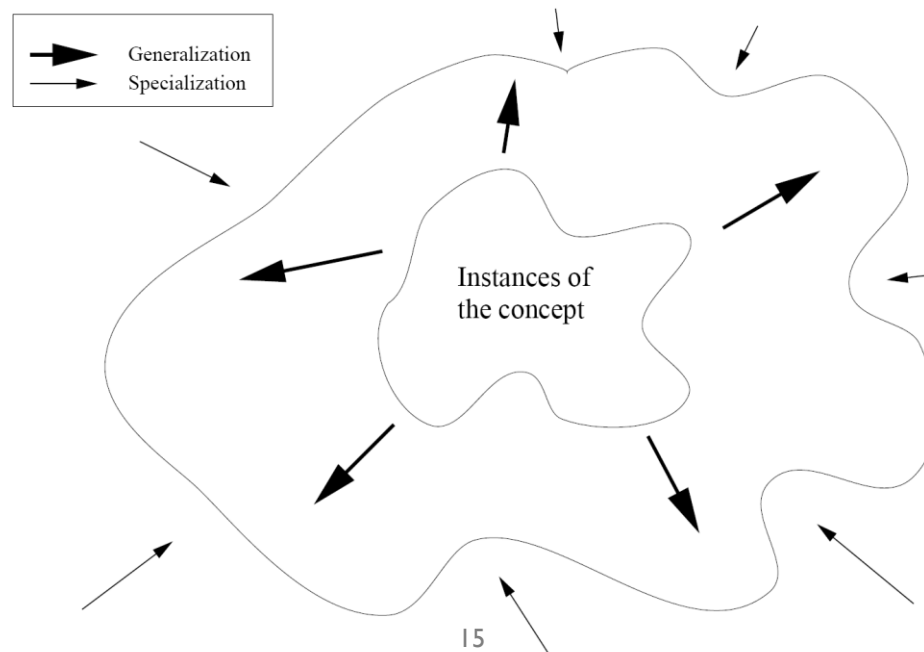


머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection): 일반화 vs. 특수화

- 일반화: 주어진 데이터로부터 정상 범주의 개념을 확장해 가는 것
- 특수화: 주어진 데이터로부터 정상 범주의 개념을 좁혀 가는 것
- 일반화에 치중할 경우 이상치 데이터 판별이 어렵게 되며, 특수화에 치중할 경우 과적합의 위험(빈번한 false alarm)에 빠질 수 있음


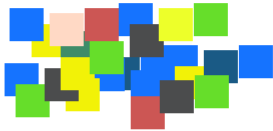






머신 러닝의 종류

구분기준 3: 사용 데이터에 따른 구분

✓ 정형 데이터(Structured Data): 기존 방식으로 테이블에 적재된 수치 데이터

✓ 비정형 데이터(Unstructured Data): 이미지, 음성, 텍스트 등 숫자가 아닌 형태의 정보가 구조화되지 않은 형태로 존재하는 데이터

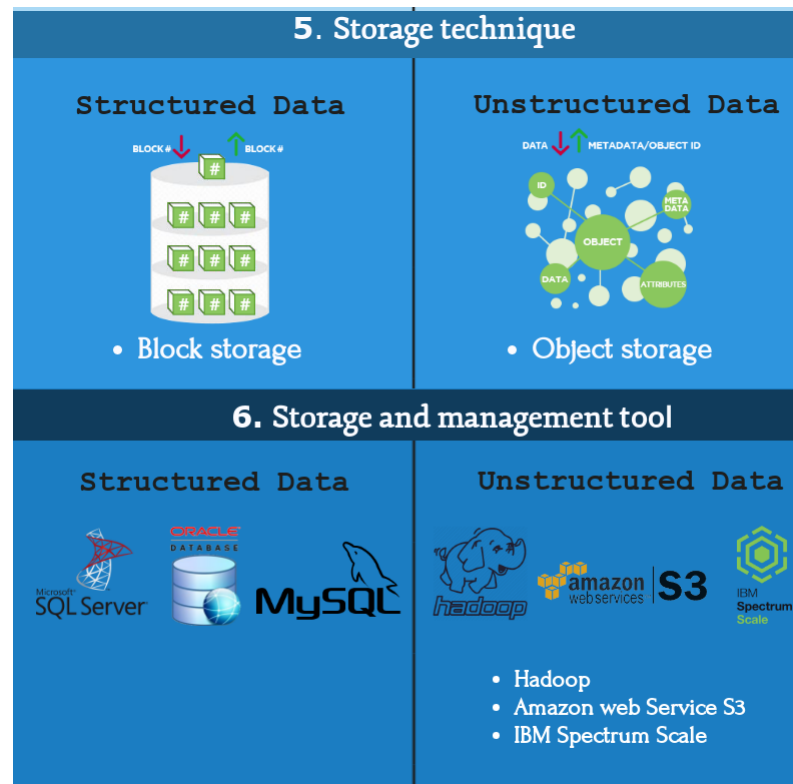
1. Definition		3. Growth	
Structured Data  <p>Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.</p>	Unstructured Data  <p>Unstructured data (or unstructured information) is information that either does not have a predefined data model or is not organized in a pre-defined manner.</p>	Structured Data  <ul style="list-style-type: none"> • Structure data accounts for about 20% of the total existing data. 	Unstructured Data  <ul style="list-style-type: none"> • Experts estimate that 80% of the data in any organization is unstructured.
2. Example		4. Characteristic	
Structured Data  <ul style="list-style-type: none"> • Databases (structuring fields) • Meta-data (Time and date of creation, File size, Author etc.) • Census records (birth, income, employment, place etc.) 	Unstructured Data  <ul style="list-style-type: none"> • Website Data which are present in the form of HTML Pages. • Media (MP3, digital photos, audio and video files) • Text files (Word processing, spreadsheets, presentations etc.) 	Structured Data <ul style="list-style-type: none"> • Schema dependent. • Scaling DB schema is difficult. • Robust. • Structured query allows complex joins. • Easy to access. • Organized. • Efficient to analysis. 	Unstructured Data <ul style="list-style-type: none"> • Absence of schema. • Very flexible. • Highly scalable. • Only textual query possible. • Hard to access. • Scattered and dispersed. • Additional preprocessing is needed.

머신 러닝의 종류

- 구분기준 3: 사용 데이터에 따른 구분

- ✓ 정형 데이터(Structured Data): 기존 방식으로 테이블에 적재된 수치 데이터

- ✓ 비정형 데이터(Unstructured Data): 이미지, 음성, 텍스트 등 숫자가 아닌 형태의 정보가 구조화되지 않은 형태로 존재하는 데이터



머신 러닝의 종류

- 구분기준 4: 학습 목적과 모델 업데이트 방식에 따른 구분

	Static Learning	Incremental (online) Learning	Reinforcement Learning
Objective Function	Short-term (snapshot)	Short-term (snapshot)	Long-term
Model update	Fully Updated	Partially Updated	Adaptively updated

