



# Data Analytics 개요 및 활용 사례

강필성

고려대학교 산업경영공학부

Bflysoft & WIGO AI LAB

# AGENDA

01 Data Analytics 개요 및 주요 개념

---

02 데이터 과학 프로젝트 절차

---

03 Machine Learning 방법론

---

04 PP 기사 분류 모형

---

# 데이터 기반 의사결정

- 우리는 당신이 무엇을 구매할 지 이미 알고 있다



# 데이터 기반 의사결정

- 우리가 알고 싶은 것



# 데이터 기반 의사결정

- 좀 더 구체적으로 제조업에서 무엇을 할 수 있을까?

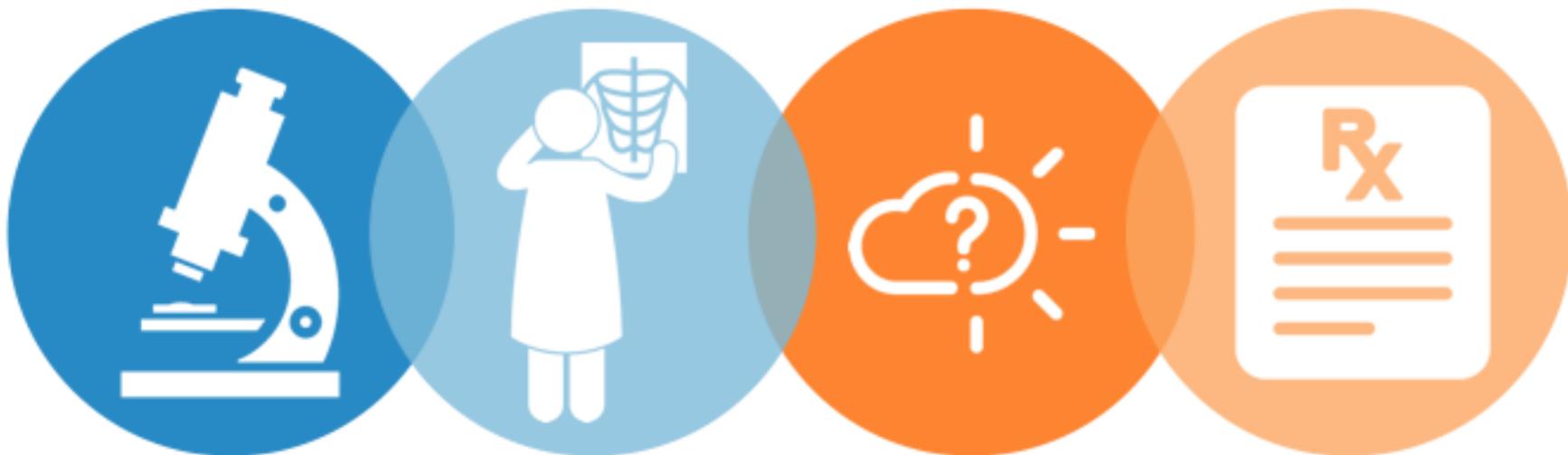
## ✓ Landing.ai

- 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



# 데이터 기반 의사결정

- 네 가지 유형의 Analytics



## Descriptive

Explains what happened.

## Diagnostic

Explains why it happened.

## Predictive

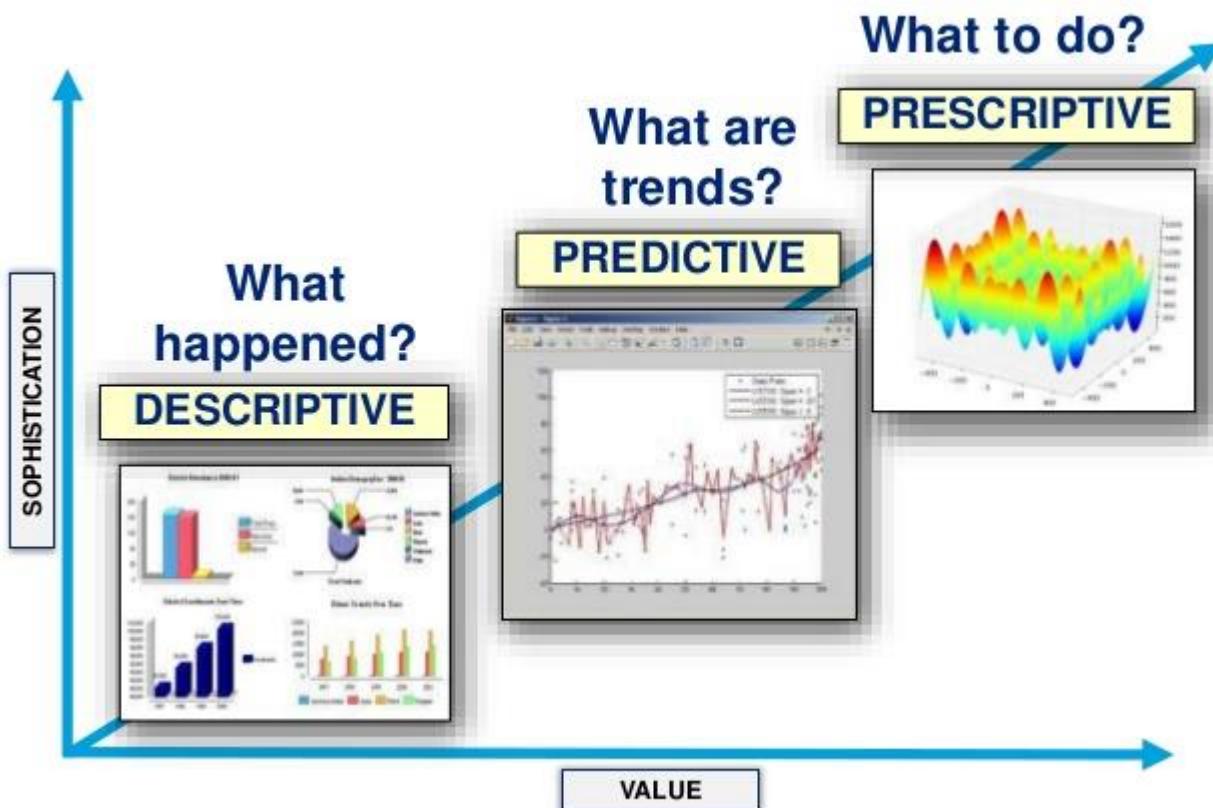
Forecasts what might happen.

## Prescriptive

Recommends an action based on the forecast.

# 데이터 기반 의사결정

- 세 가지 유형의 Analytics



# 데이터 기반 의사결정

## • 세 가지 유형의 Analytics

### Understanding analytics

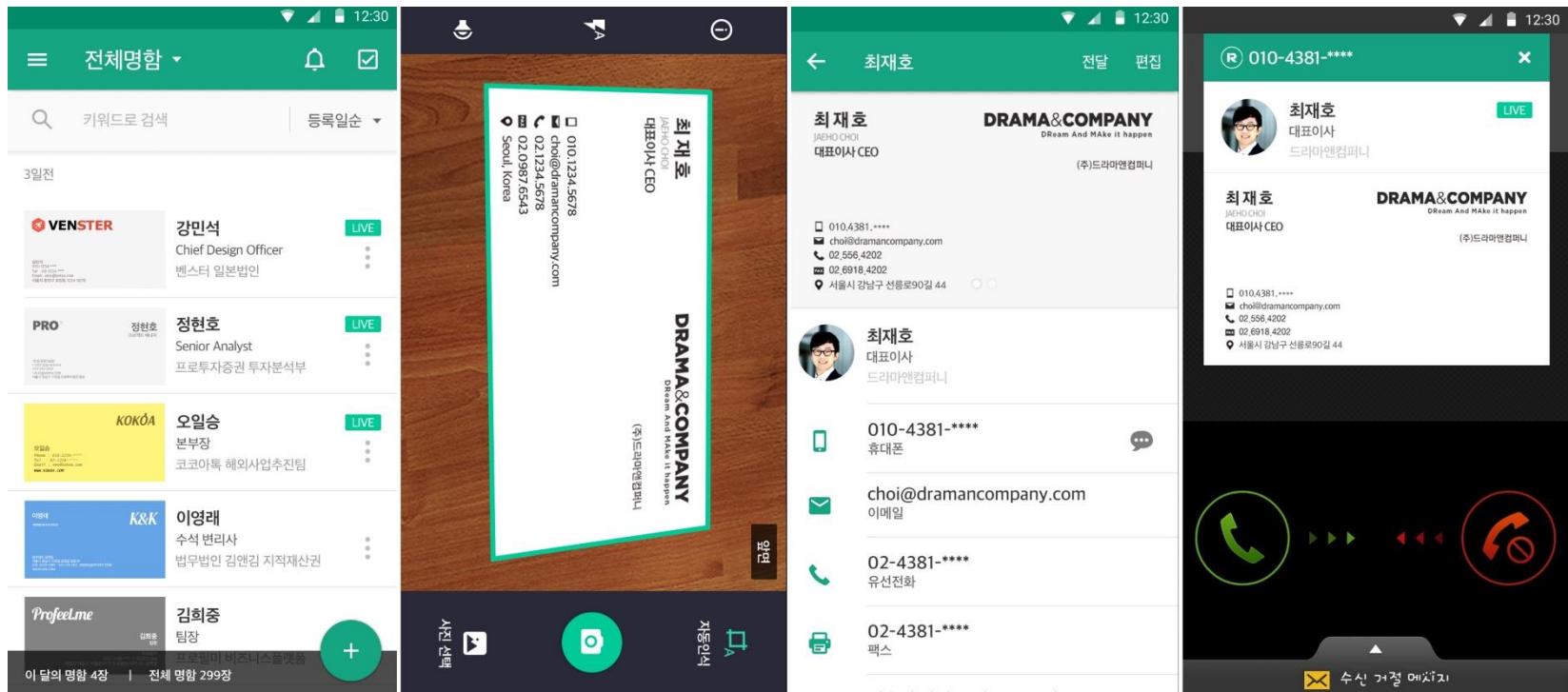
Definitions, sample applications and opportunities, and underlying technologies

	Descriptive	Predictive	Prescriptive
What the user needs to DO	What HAS happened?	What COULD happen?	What SHOULD happen?
What the user needs to KNOW	<ul style="list-style-type: none"><li>Increase asset reliability</li><li>Reduce labor and inventory costs</li></ul>	<ul style="list-style-type: none"><li>Predict infrastructure failures</li><li>Forecast facilities space demands</li></ul>	<ul style="list-style-type: none"><li>Increase asset utilization</li><li>Optimize resource schedules</li></ul>
How analytics gets ANSWERS	<ul style="list-style-type: none"><li>The number and types of asset failures</li><li>Why maintenance costs are high</li><li>The value of the materials inventory</li></ul>	<ul style="list-style-type: none"><li>How to anticipate failures for specific asset types</li><li>When to consolidate underutilized facilities</li><li>How to determine costs to improve service levels</li></ul>	<ul style="list-style-type: none"><li>How to increase asset production</li><li>Where to optimally route service technicians</li><li>Which strategic facilities plan provides the highest long-term utilization</li></ul>
What makes this analysis POSSIBLE	<ul style="list-style-type: none"><li>Standard reporting - What happened?</li><li>Query/drill down - Where exactly is the problem?</li><li>Ad hoc reporting - How many, how often, where?</li></ul>	<ul style="list-style-type: none"><li>Predictive modeling - What will happen next?</li><li>Forecasting - What if these trends continue?</li><li>Simulation - What could happen?</li><li>Alerts - What actions are needed?</li></ul>	<ul style="list-style-type: none"><li>Optimization - What is the best possible outcome?</li><li>Random variable optimization - What is the best outcome given the variability in specified areas?</li></ul>
Business value →			

# 데이터 기반 의사결정: 데이터의 중요성

- 데이터는 수단인가? 아니면 목적인가?

- ✓ 데이터를 얻기 위해 제품/서비스를 판매한다?



- ✓ 결국 2017년 12월 21일 라인플러스와 네이버가 인수함

# 데이터 기반 의사결정: 데이터의 중요성

- 멋있는 알고리즘도 중요하지만 그 전에 먼저 제대로 된 데이터를 수집하자

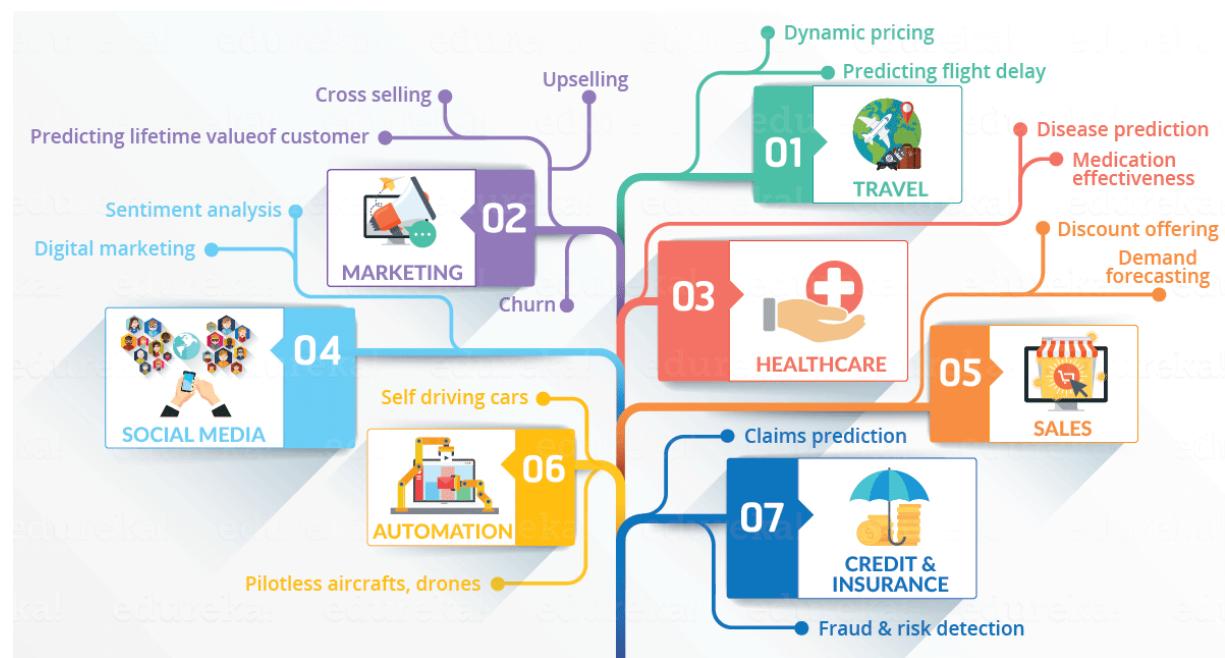
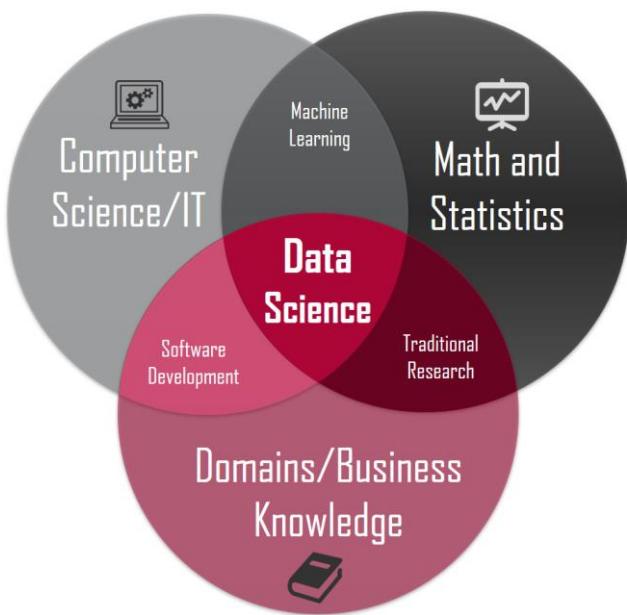


"문제는 한국 사회에서 시스템이 필요하다고 지시를 내릴 사람은 많은데 **전통적으로** '노가다'를 뛸 사람은 없다는 겁니다. 이런 일은 남이 해야 하는 거라 생각하죠. 아니면 남이 했다가 자기한테 해가 되면 안 되니까 오만 가지 이유를 대서 이런 일은 하면 안 된다고 하고, 이런 일이 의료계에서만 있는 줄 알았어요. 사회 전반이 바뀌지 않으면 이 문제는 나아지지 않아요"라고 했다.

# 데이터 과학: 정의

- 데이터 과학이란?

- ✓ 다양한 학제간 학문이 융합되어 데이터 기반 의사결정 및 문제 해결을 목적으로 하는 학문



# 데이터 과학: 연역법 vs. 귀납법

규칙: A속성의 사람들은 인사를 하고 B속성의 사람들은 악수를 한다

A와 B는 무엇일까?



# 데이터 과학: 연역법 vs. 귀납법

아시아계 사람들은 인사를 하고 백인들은 악수를 한다.



# 데이터 과학: 연역법 vs. 귀납법

아시아계 사람들은 인사를 하고 백인들은 악수를 한다.



# 데이터 과학: 연역법 vs. 귀납법

같은 색상의 옷을 입은 사람들은 인사를 하고,  
다른 색상의 옷을 입은 사람들은 악수를 한다.



# 데이터 과학: 연역법 vs. 귀납법

같은 색상의 옷을 입은 사람들은 인사를 하고,  
다른 색상의 옷을 입은 사람들은 악수를 한다.



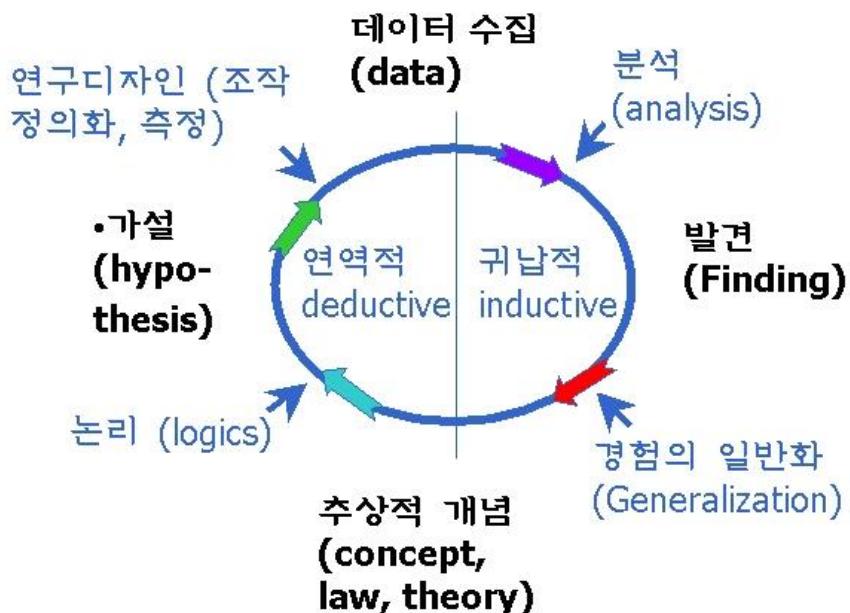
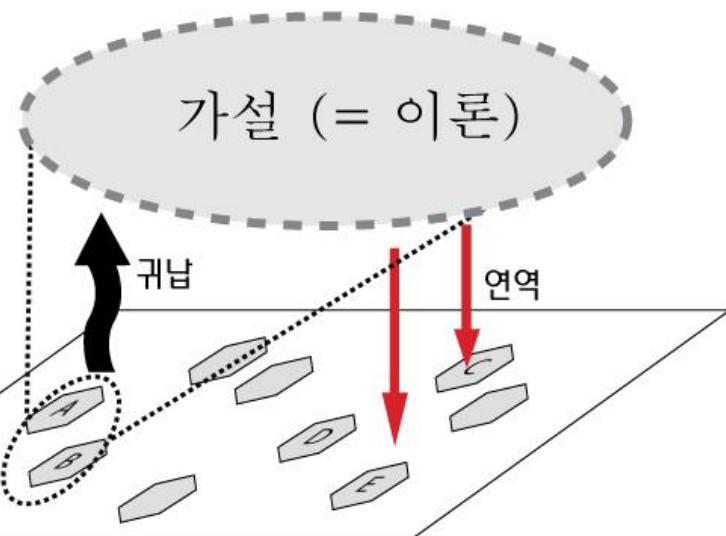
# 데이터 과학: 연역법 vs. 귀납법

- 연역법

- ✓ 일반적 사실이나 원리를 전제로 하여 개별적인 특수한 사실이나 원리를 결론으로 이끌어 내는 추리 방법 (예: 삼단논법)

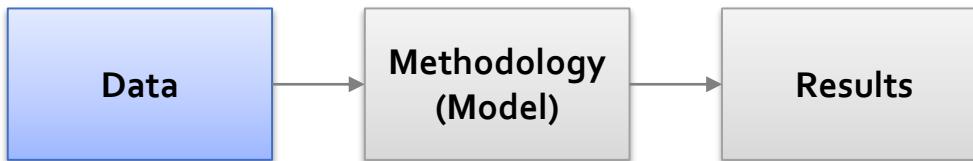
- 귀납법

- ✓ 여러 가지의 관찰된 사실들을 바탕으로 이들의 기저에 깔려 있는 일반적인 원리를 추론해 내는 방법



# 데이터 과학 주요 개념: 빅데이터

- 빅데이터(Big Data)



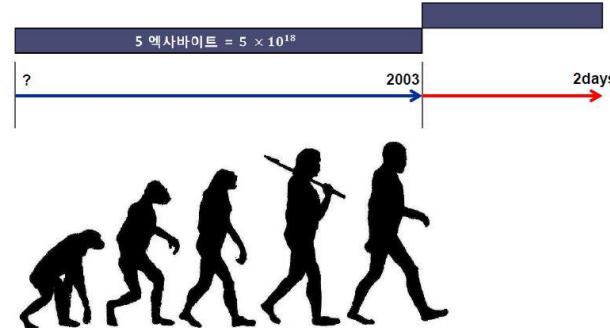
- ✓ 데이터베이스 규모에 초점을 맞춘 정의 (McKinsey, 2011)
  - 일반적인 데이터베이스 SW가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- ✓ 업무 수행에 초점을 맞춘 정의 (IDC, 2011)
  - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처

# 데이터 과학 주요 개념: 빅데이터

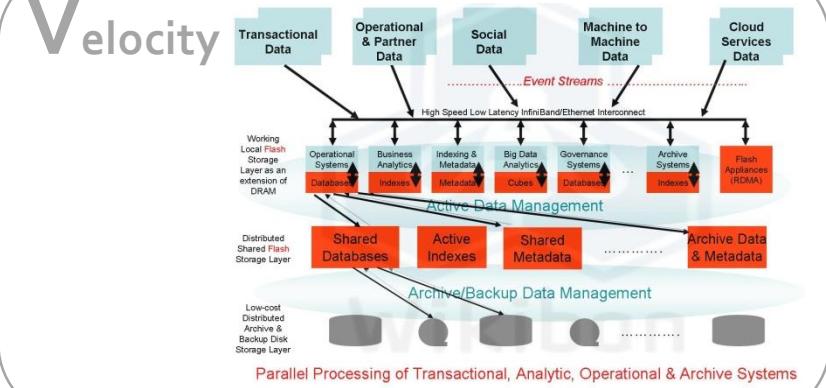
## • 빅데이터의 4V

- ✓ 빅데이터의 특징은 방대한 양(Volume), 빠른 데이터 생성 및 처리 속도(Velocity), 다양한 형태(Variety) 및 데이터에 내재된 잠재 가치(Value)로 정의됨

### Volume



### Velocity



### Variety



### Value



자료: McKinsey (2011.05)

# 데이터 과학 주요 개념: 빅데이터

- 빅데이터의 특징

- ✓ 복잡하고 고도화된 분석 방법론이 아닌 데이터 그 자체로서 가치를 가짐



VS



# 데이터 과학 주요 개념: 빅데이터

## • 빅데이터의 특징

- ✓ 복잡하고 고도화된 분석 방법론이 아닌 데이터 그 자체로서 가치를 가짐



- 데이터에 의한 정량적 유동인구 분포도 작성
- 서울시를 1km 반경의 1,250개 헥사셀 단위로 구분
- KT 휴대전화 이력 데이터로 심야시간(0시~5시) 통화량 분석
- 유동인구 기반 노선 최적화 및 배차간격 조정



# 데이터 과학 주요 개념: 빅데이터

- 빅데이터의 특징

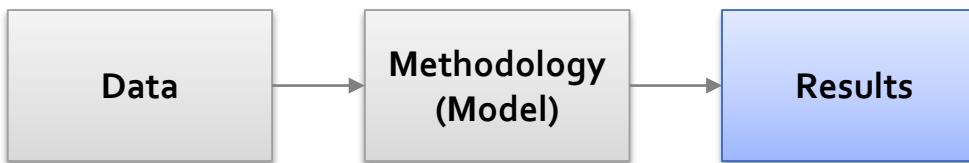
- ✓ 복잡하고 고도화된 분석 방법론이 아닌 데이터 그 자체로서 가치를 가짐
  - 지멘스 암베르크 공장 사례



# 데이터 과학 주요 개념: 데이터 마이닝

- 데이터 마이닝: Data Mining

- ✓ 대량의 데이터로부터 의미있는 규칙이나 패턴을 추출하는 일련의 활동



- ✓ Extracting useful information from large datasets. (Hand et al., 2001)
- ✓ The process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, 2000)
- ✓ The process of discovering meaningful new correlations, patterns and trends by sifting through large amount data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. (Gartner Group, 2004)

# 데이터 과학 주요 개념: 데이터 마이닝

- 데이터 마이닝: Data Mining

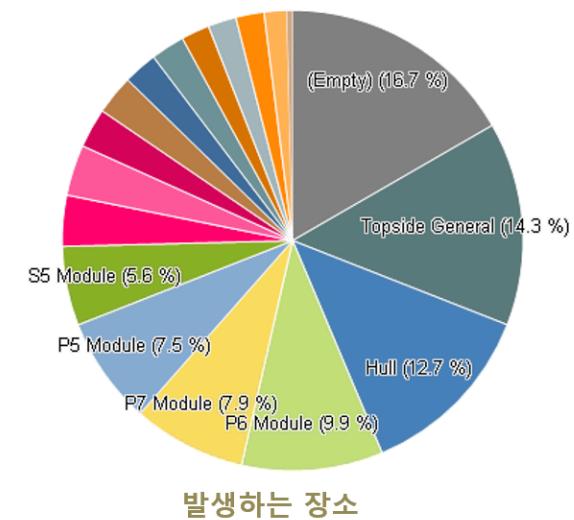
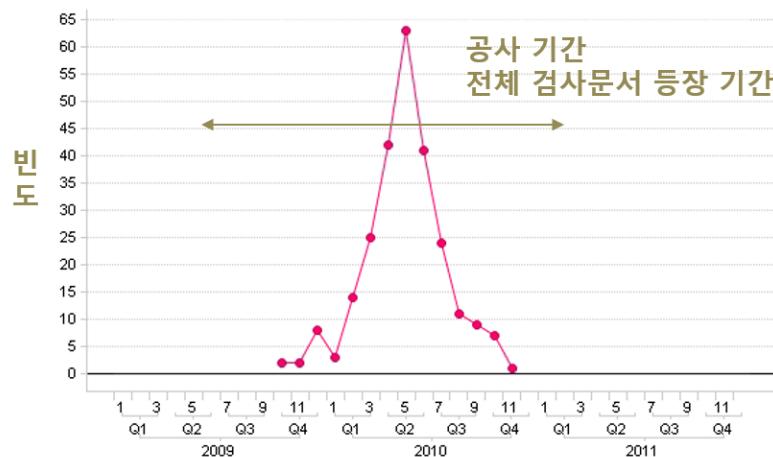
- ✓ 대량의 데이터로부터 의미있는 규칙이나 패턴을 추출하는 일련의 활동



“파이프(pipe)가 흔들리니(shake), 지주(support)를 추가(add)하라”

언제, 어디서?

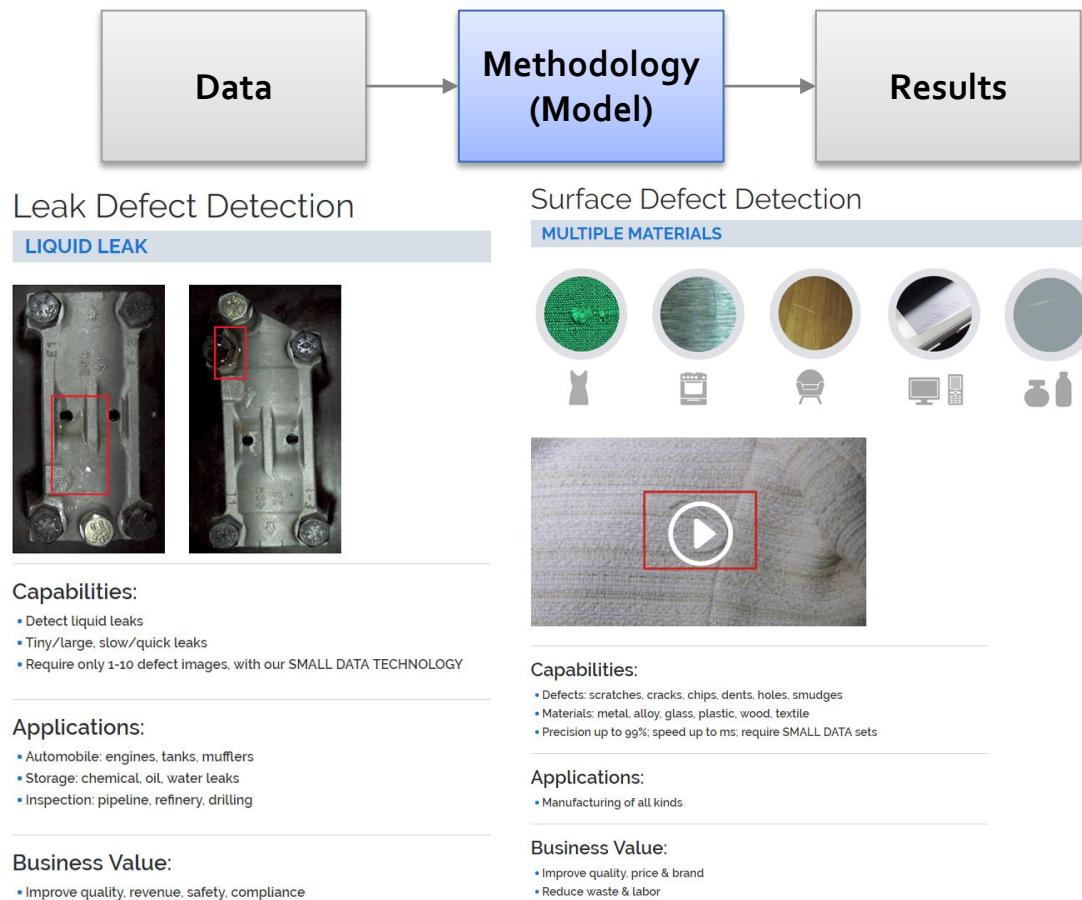
“공사 중반, Topside General, Hull, P5,6,7 Module 등에서 주로 발생한다”



# 데이터 과학 주요 개념: 기계 학습

## • 기계 학습: Machine Learning

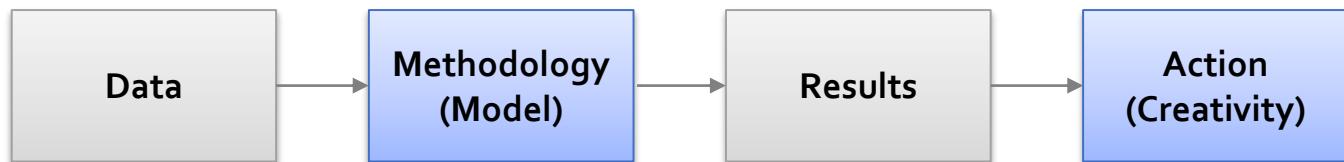
- ✓ 특정한 과업 Task을 달성하기 위해 경험 Experience이 축적될수록 과업 수행의 성능 Performance 이 향상되는 컴퓨터 프로그램 또는 에이전트를 개발하는 것 – Mitchell (1997)



# 데이터 과학 주요 개념: 인공 지능

- 인공 지능: Artificial Intelligence

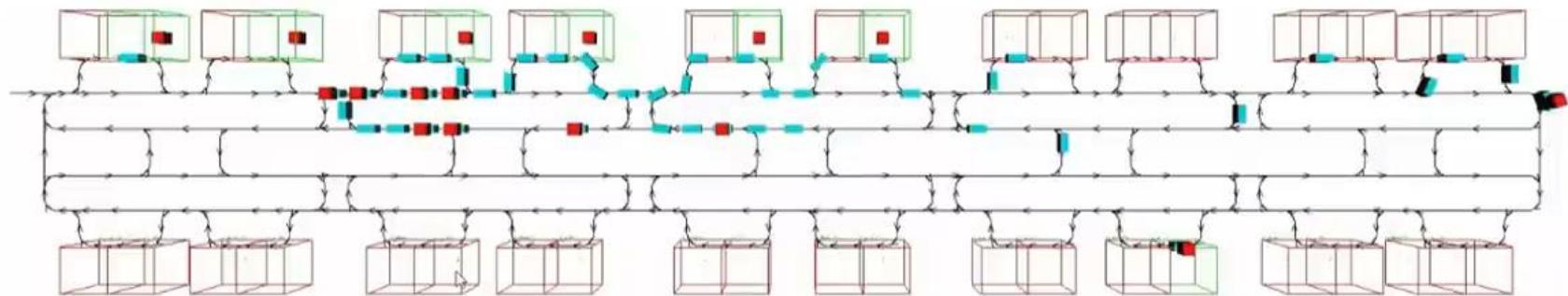
- ✓ 환경을 인지하여 보상이 최대화되는 지능적인 행위를 할 수 있는 컴퓨터 소프트웨어



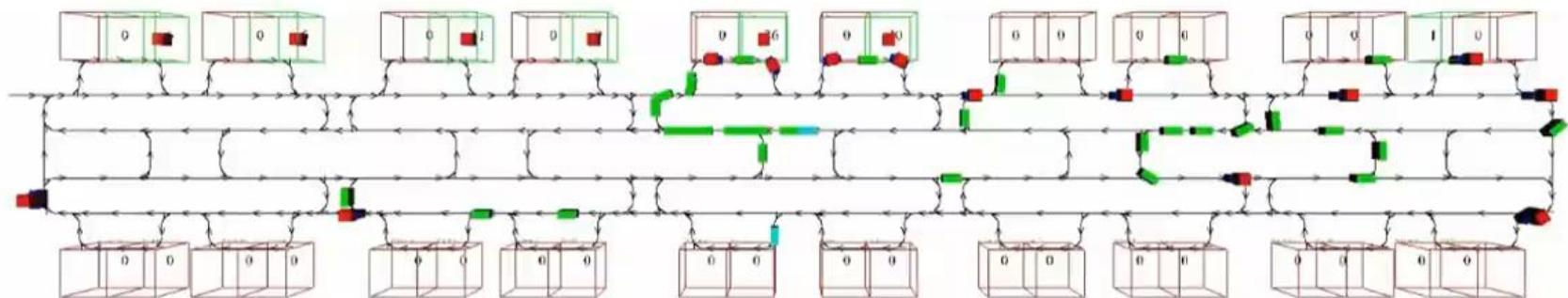
# 데이터 과학 주요 개념: 인공 지능

- 강화학습을 이용한 물류 최적화

Current approach (기존 알고리즘)



Proposed algorithm (카이스트 개발 알고리즘)



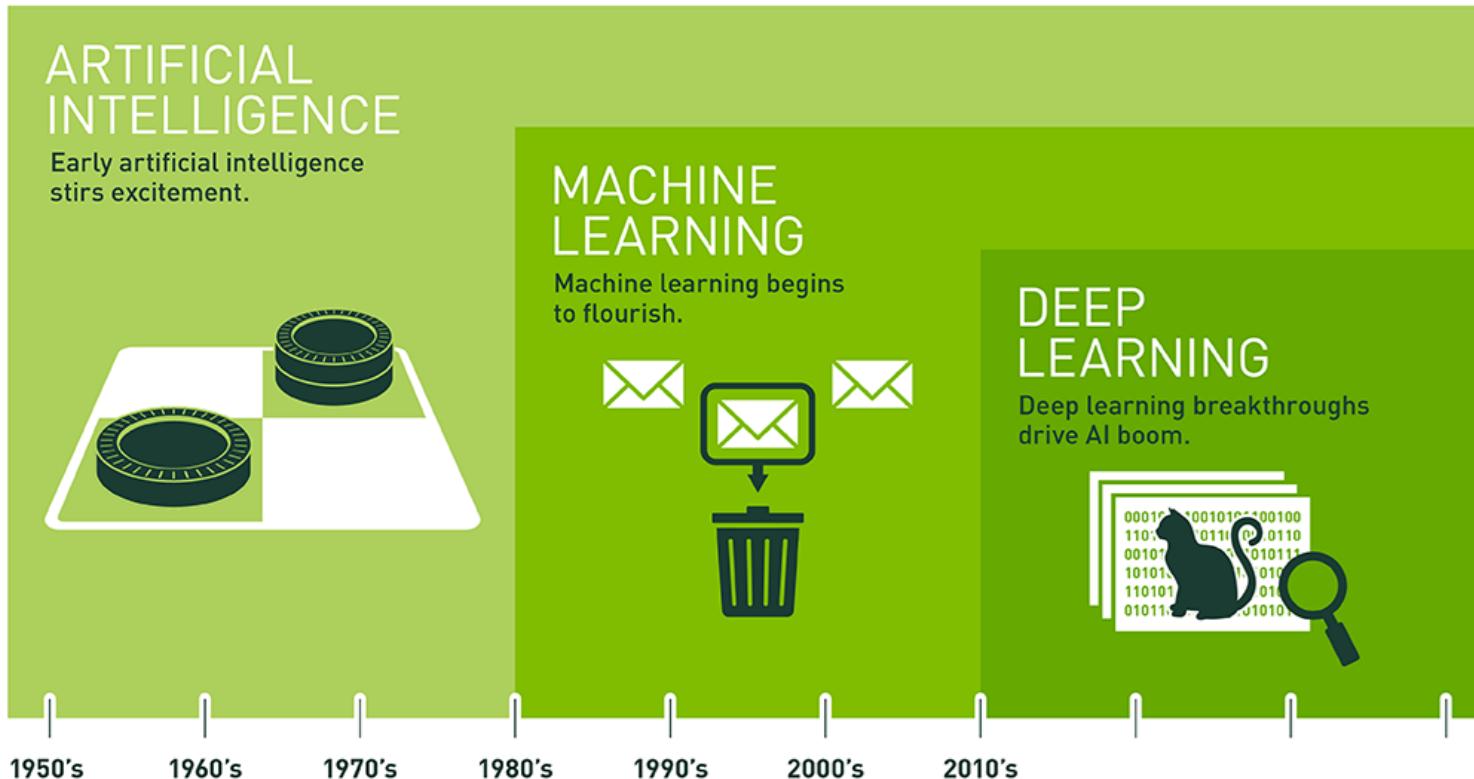
Ilhoe Hwang, Sang Pyo Hong, Young Jae Jang, Sunil Kim and In-Ho Moon, "System Design and Development of the Q-Learning Based Overhead Hoist Transport (OHT) for Semiconductor fabs," International Symposium on Semiconductor Initiatives, 2018

Information: <http://sdm.kaist.ac.kr>

# 데이터 과학 주요 개념: 인공 지능

- 인공지능 vs. 기계학습 vs. 딥러닝

- ✓ 인공지능이 가장 상위 개념이며 딥러닝은 기계학습의 한 종류임

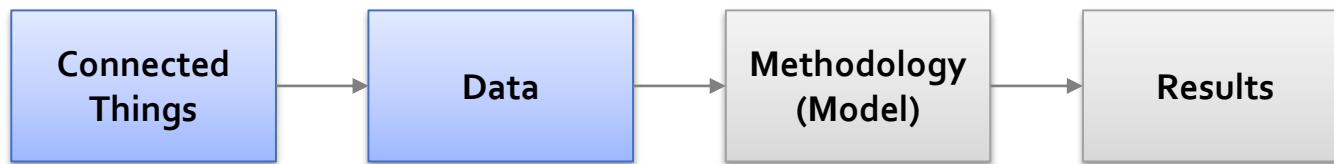


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

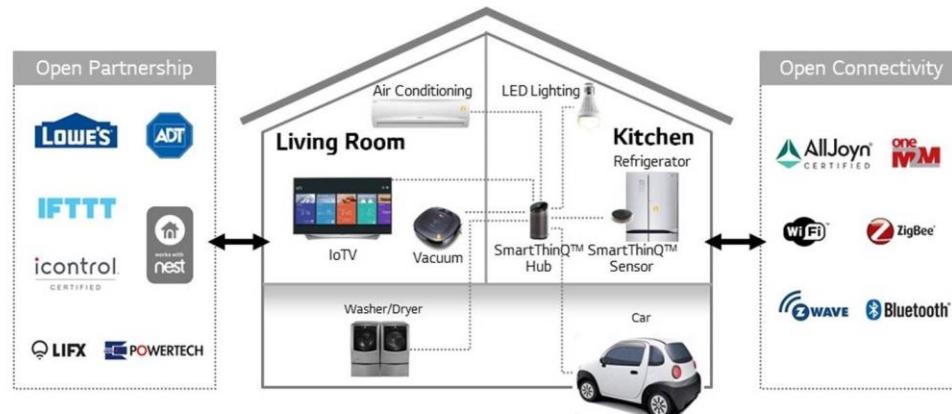
# 데이터 과학 주요 개념: 사물 인터넷

## • 사물 인터넷

- ✓ 센서 및 소프트웨어가 내장된 물리적 개체들이 연결되어 각 개체들간의 통신, 데이터 교환, 컨트롤 등을 지원하는 네트워크 체계



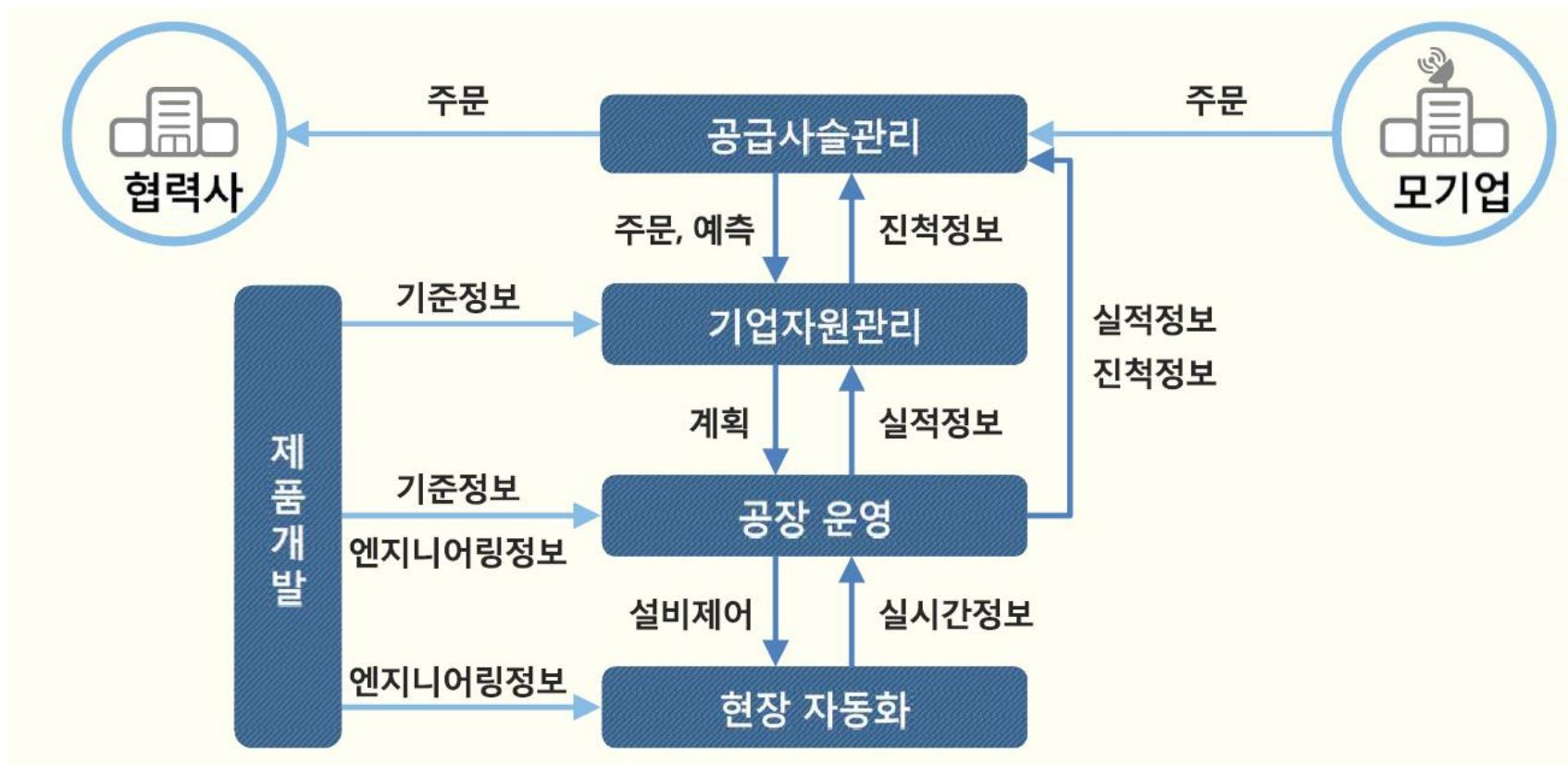
LGE IoT Eco System



# 데이터 과학 주요 개념: 사물 인터넷

- 사물 인터넷: Internet of Things

- ✓ 4차 산업 혁명과 스마트 공장(Smart Factory)의 핵심 구성 요소



# AGENDA

01 Data Analytics 개요 및 주요 개념

---

02 데이터 과학 프로젝트 절차

---

03 Machine Learning 방법론

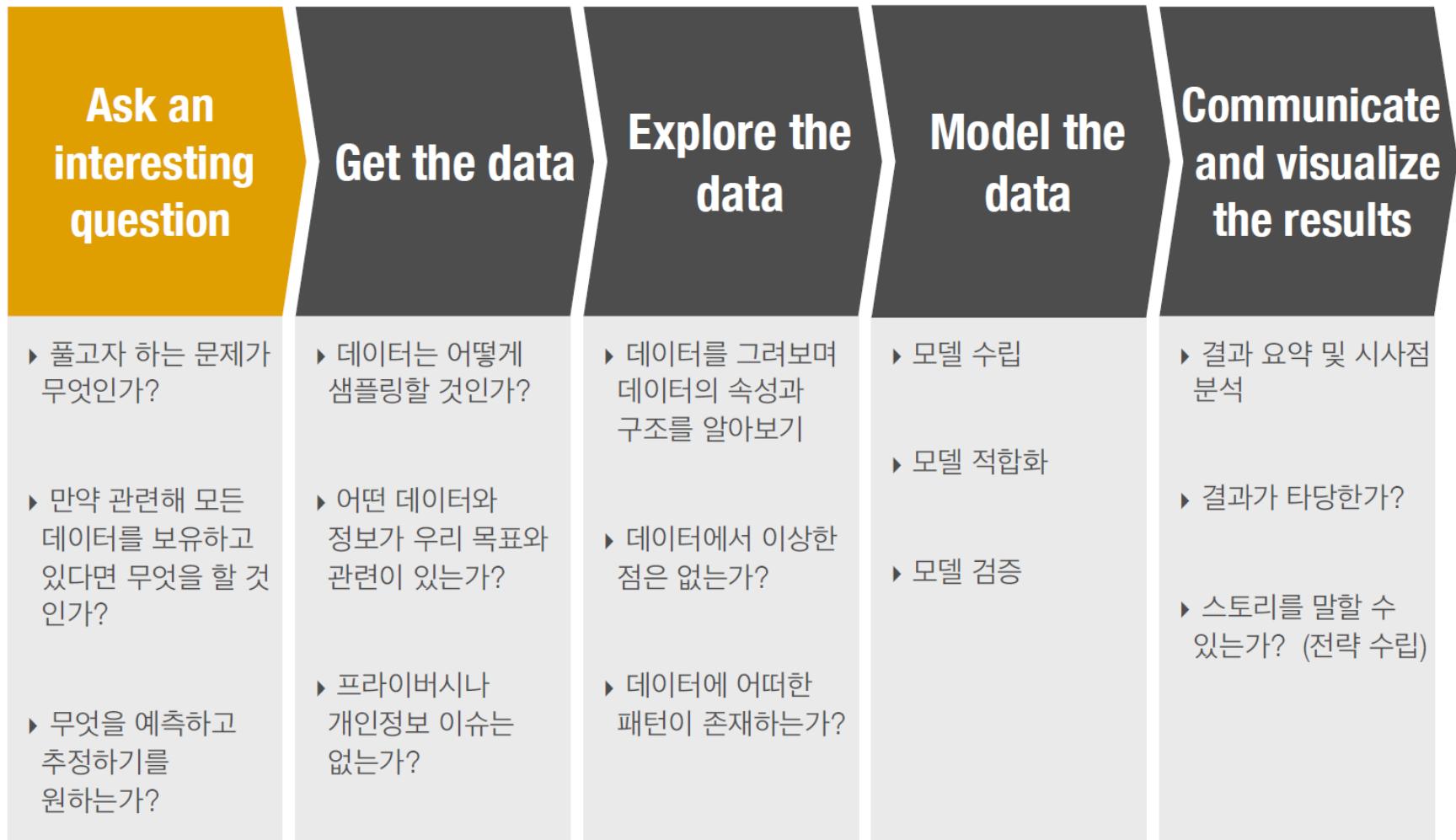
---

04 PP 기사 분류 모형

---

# 데이터 기반 문제 해결 절차

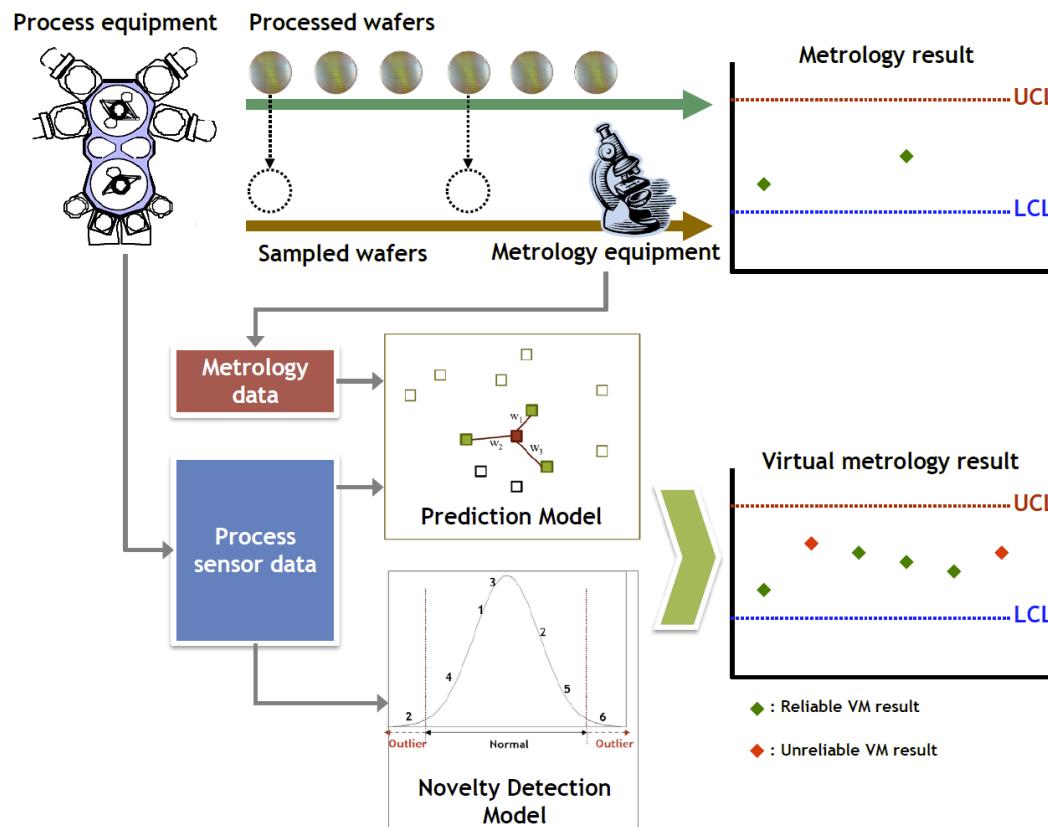
- 데이터 기반의 문제해결 5단계



# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

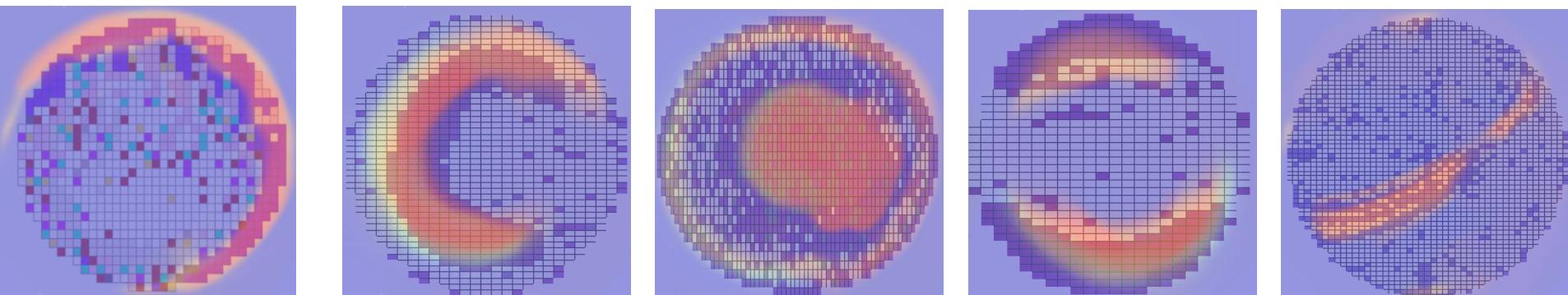
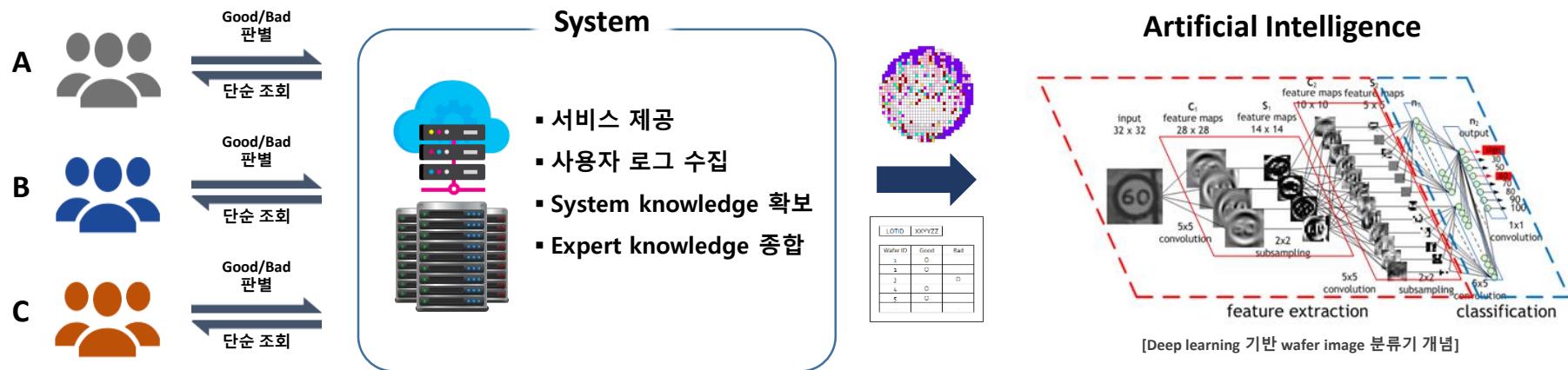
- ✓ 흥미로운 문제(매출 증대, 비용 감소, 공정 단축 등 해결될 경우 조직에 도움이 될 것으로 예상되는 문제)를 발굴할 것
- ✓ 예) 공정 중에 발생한 불량 원인은 장비에 기록되는가?



# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

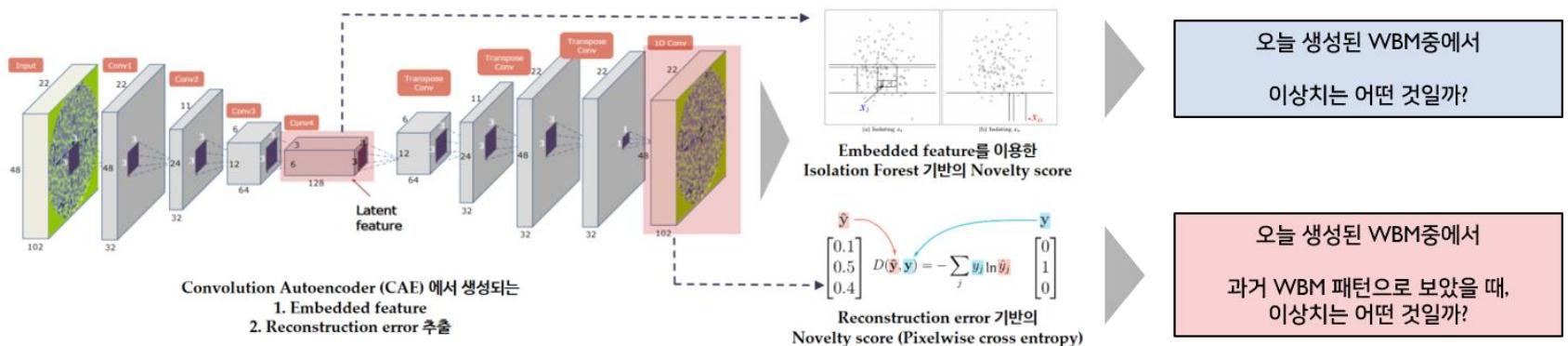
- ✓ 예) Wafer bin map (WBM)의 정상/이상 유무를 자동으로 판별할 수 있는가?
- ✓ 예) 불량일 경우 어느 영역 때문인지를 규명해줄 수 있는가?



# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

- 예) 오늘 생산된 웨이퍼들 중에서 WBM 관점에서 특이한 웨이퍼를 판별할 수 있는가?
- 오늘 생산된 웨이퍼들 중에서 과거 생산된 웨이퍼들과는 다른 WBM을 가지는 웨이퍼들을 파악할 수 있는가?
- 과거와 다른 WBM 패턴으로 판별될 경우, 어떤 칩/다이가 판별에 큰 영향을 미쳤는지 알 수 있는가?



# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

- 예) 중장비 가동 데이터로부터 부품 고장 예측, 유효 알람 파악이 가능한가?

**중장비 고장 예측 모델 구축**

ALARM

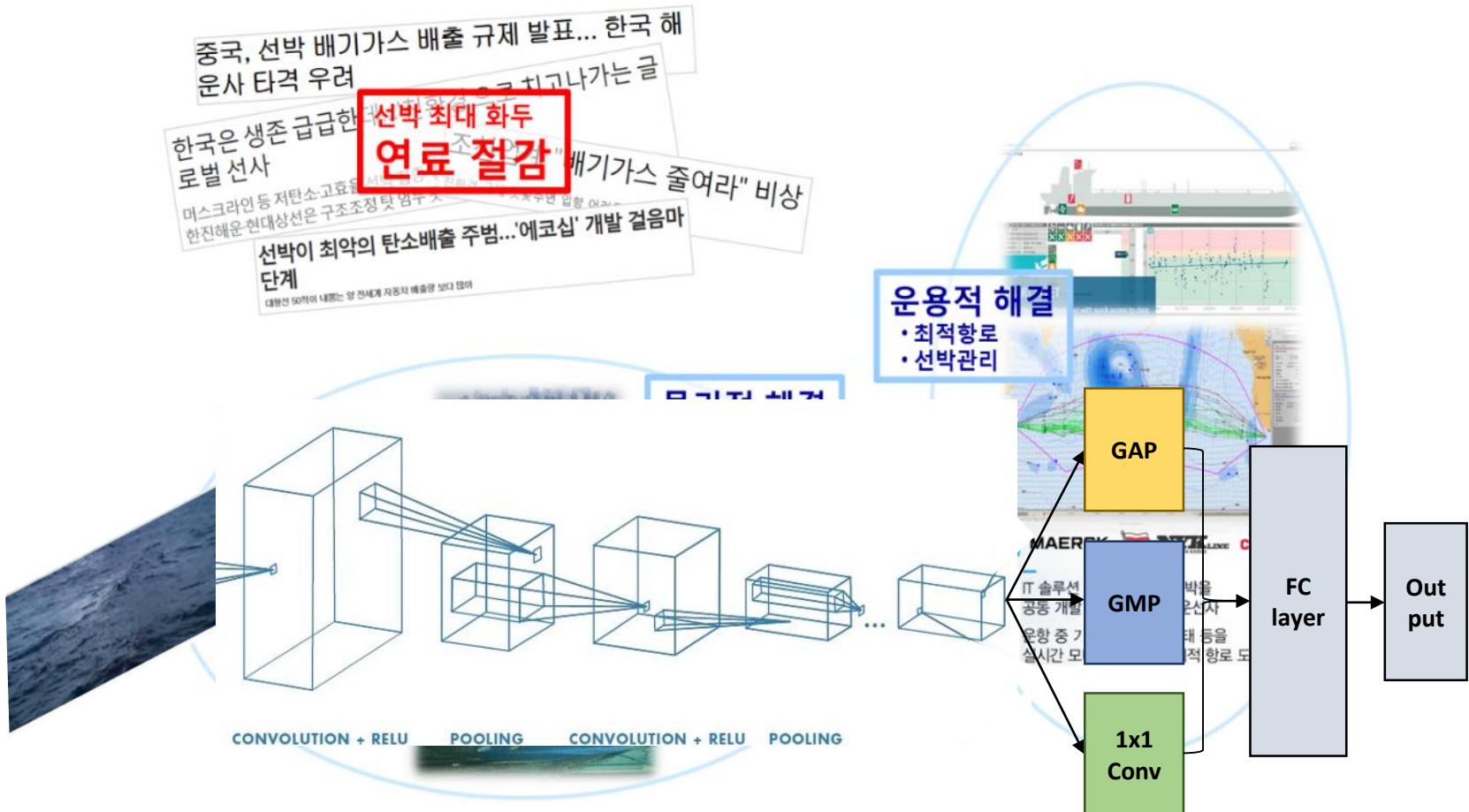
- 하이메이트의 장비 별 데이터와 실제 고장이력을 연동하여 장비상태에 따른 고장 예측 모델 구축
- 장비의 부품 별로 고장을 예측하고 고장에 영향을 주는 변수의 시각화를 통해 고장 패턴에 대한 지식 전달

- 장비의 고장 이력 등을 연동하여 알람의 종류를 정의
- 알람의 종류를 예측하여 불필요한 알람과 중요한 알람에 대한 가이드 제시
- 동시에 발생하는 알람간의 연관성을 분석하여 알람 발생 패턴에 대한 정보 제공

# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

- 예) 해상 이미지 데이터로부터 파고/파향/주기를 예측할 수 있는가?



# 데이터 기반 문제 해결 절차

## • 1단계: 문제 정의

- ✓ 예) 기존 방식보다 더 정확한 제어는 가능한가?

■ 세계 최초 인공지능 제철소로 거듭나는 포스코



▲포스코 기술연구원에서 한 연구원이 철강조직 검사를 실시하고 있다.(왼쪽) 포스코 광양제철소 3CGL(용융아연드금강판공장)의 운전실에서 개발자와 작업자가 '인공지능 기반 도금량 제어 자동화 솔루션'을 모니터링하고 있다.

<http://news.mk.co.kr/newsRead.php?year=2017&no=112444>



Jong-Seok Lee

어제 오전 1:40 ·

도금량 제어 알고리즘 적용에 대한 포스코 내부적인 비용절감액이 산출이 되었다. 언젠가는 공개가 되겠지만 페이스북에 올 수는 없을 것 같고. 스스로 상당히 보람을 느낄 수 있을 정도.

이번 과제 수행을 통해 "체험"한 가장 큰 교훈은  
**큰 일은 혼자 할 수 없다**

는 것이다. 시간이 지나 당시를 뒤돌아 보면 과제를 수행하면서 상황적 도움과 주변 사람들의 도움이 매우 컸다.

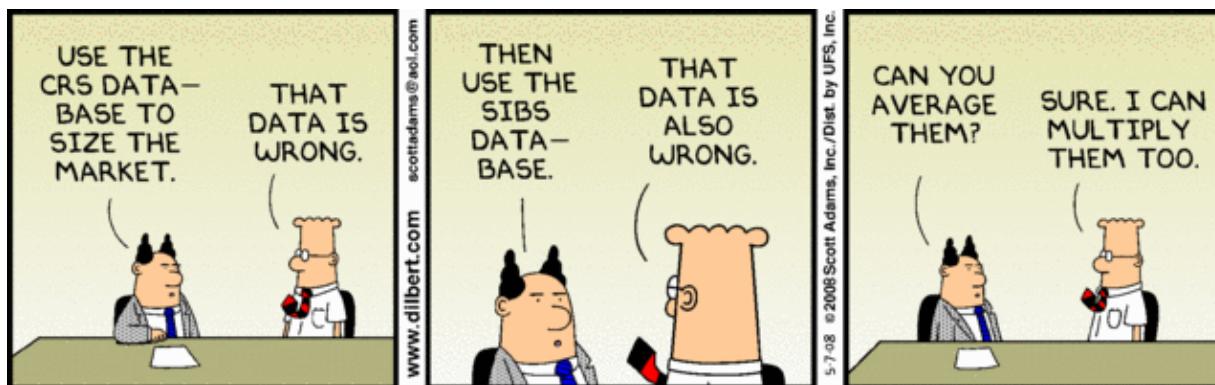
우선 해결해야 하는 문제를 아주 명확히 정의하고 해당 문제를 해결하기 위한 요소들에 대한 자세한 초기 설명을 제공해 준 기술연구소 연구원 분들과 현 광양기술연구소장님께 감사한 마음이 듈다. 이 분들은 꾸준히 문제를 함께 해결하기 위한 관심과 노력을 아끼지 않으신 분들이다. 함께 밤 늦은 시간까지 제철소 내에서 고민했고, 밤을 새기도 했다. 알고리즘을 위한 전용 PLC 설치와 DBMS 설치를 지속적으로 도와준 협력업체 분들께도 감사하다. 데이터 수집항목, 수집주기 등이 과제를 수행하는 동안 몇 번 바뀌었는데 그 요구사항들을 신속하게 처리해 주셨다. 운전실에서 실제 도금업무를 수행하는 조업자 분들도 너무 감사한 분들이다. 이 분들이야 말로 새로운 기술을 받아들이기가 가장 힘든 조업의 최전방에 계신 분들이다. 초기 알고리즘이 불안정할 때도 주의를 기울이며 최대한 사용을 해 주려고 노력하셨고 그 덕분에 알고리즘이 점차 현장 적용에 맞도록 수정되어 갈 수 있었다. 을 여름에 치킨이라도 사서 들고 방문해야지. 그냥 늦은 밤에 감사한 마음이 들어 간단히 그 분들께 감사한 마음을 적어본다.

작년 해당 과제를 수행하면서 만난 "사람"들이 모두 너무나도 좋은 분들이었다는 것은 정말 엄청난 행운이 아닐 수 없다.

큰 일은 혼자 할 수 없다는 것을 단지 글로만 머리속으로만 알고 있던 것을 이렇게 체험할 수 있었다는 것 역시 큰 행운이 아닐 수 없다. 더 겸손하고 더 다가가기 쉬운 사람이 되자.

# 데이터 기반 문제 해결 절차

- 2단계: 분석에 적합한 데이터를 수집하라
  - Garbage in, garbage out



- The larger, the better



**"We don't have better algorithms than anyone else. We just have more data."**

- 데이터가 없다면 수집부터, 수집을 하고 있다면 중앙 집중식 관리를...

# 데이터 기반 문제 해결 절차

- 2단계: 분석에 적합한 데이터를 수집하라
  - ✓ 필요하다면 전문가의 지식을 적극 활용하라 (특히 정답 데이터를 만들 때)

The screenshot shows a research article from JAMA. At the top, there is a dark red horizontal bar with the word "Research" in white. Below it, the journal title "JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY" is displayed. The main title of the article is "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". Below the title, a list of authors is provided: Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD. To the right of the author list, there are two small blue buttons: one labeled "← Editorial" and another labeled "+ Supplemental content". The main text of the article is divided into several sections: "IMPORTANCE", "OBJECTIVE", "DESIGN AND SETTING", etc., each containing a brief description of the methodology or findings.

# 데이터 기반 문제 해결 절차

- 2단계: 분석에 적합한 데이터를 수집하라

Table. Baseline Characteristics<sup>a</sup>

Characteristics	Development Data Set	EyePACS-1 Validation Data Set	Messidor-2 Validation Data Set
No. of images	128 175	9963	1748
No. of ophthalmologists	54	8	7
No. of grades per image	3-7	8	7
Grades per ophthalmologist, median (interquartile range)	2021 (304-8366)	8906 (8744-9360)	1745 (1742-1748)

Patient demographics



<sup>a</sup> Summary of image characteristics and available demographic information in the development and clinical validation data sets (EyePACS-1 and Messidor-2). Abnormal images were oversampled for the development set for algorithm training. The clinical validation sets were not enriched for abnormal images.

<sup>b</sup> Unique patient codes (deidentified) were available for 89.3% of the development set ( $n = 114\,398$  images).

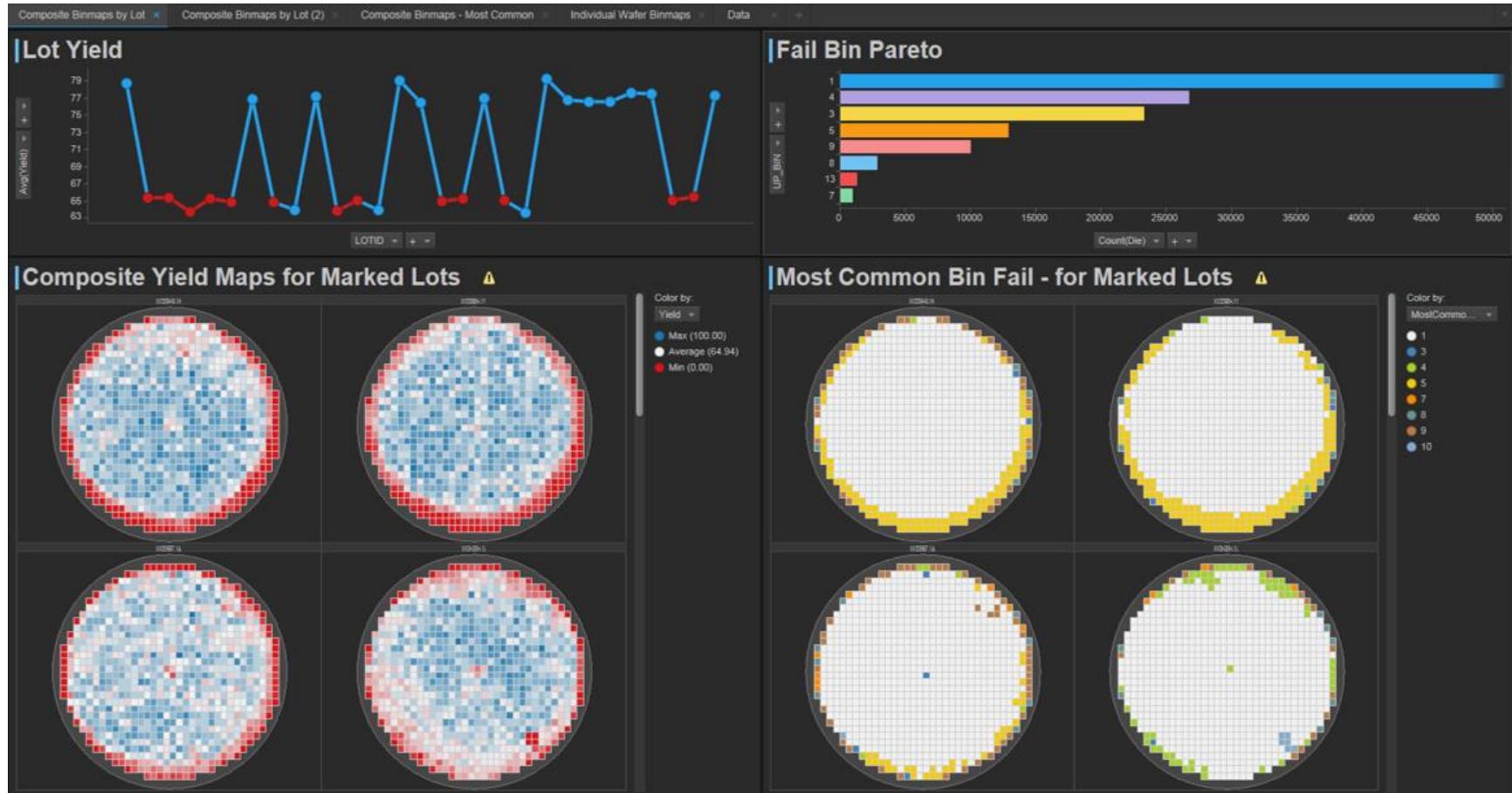
<sup>c</sup> Individual-level data including age and sex were available for 66.1% of the development set ( $n = 84\,734$  images).

<sup>d</sup> Image quality was assessed for a subset of the development set.

<sup>e</sup> Referable diabetic retinopathy, defined as the presence of moderate and worse diabetic retinopathy and/or referable diabetic macular edema according to the International Clinical Diabetic Retinopathy Scale,<sup>14</sup> was calculated for each ophthalmologist before combining them using a majority decision. The 5-point grades represent the grade that received the highest number of votes for diabetic retinopathy alone. Hence, the sum of moderate, severe, and proliferative diabetic retinopathy for the 5-point grade differs slightly from the count of referable diabetic retinopathy images.

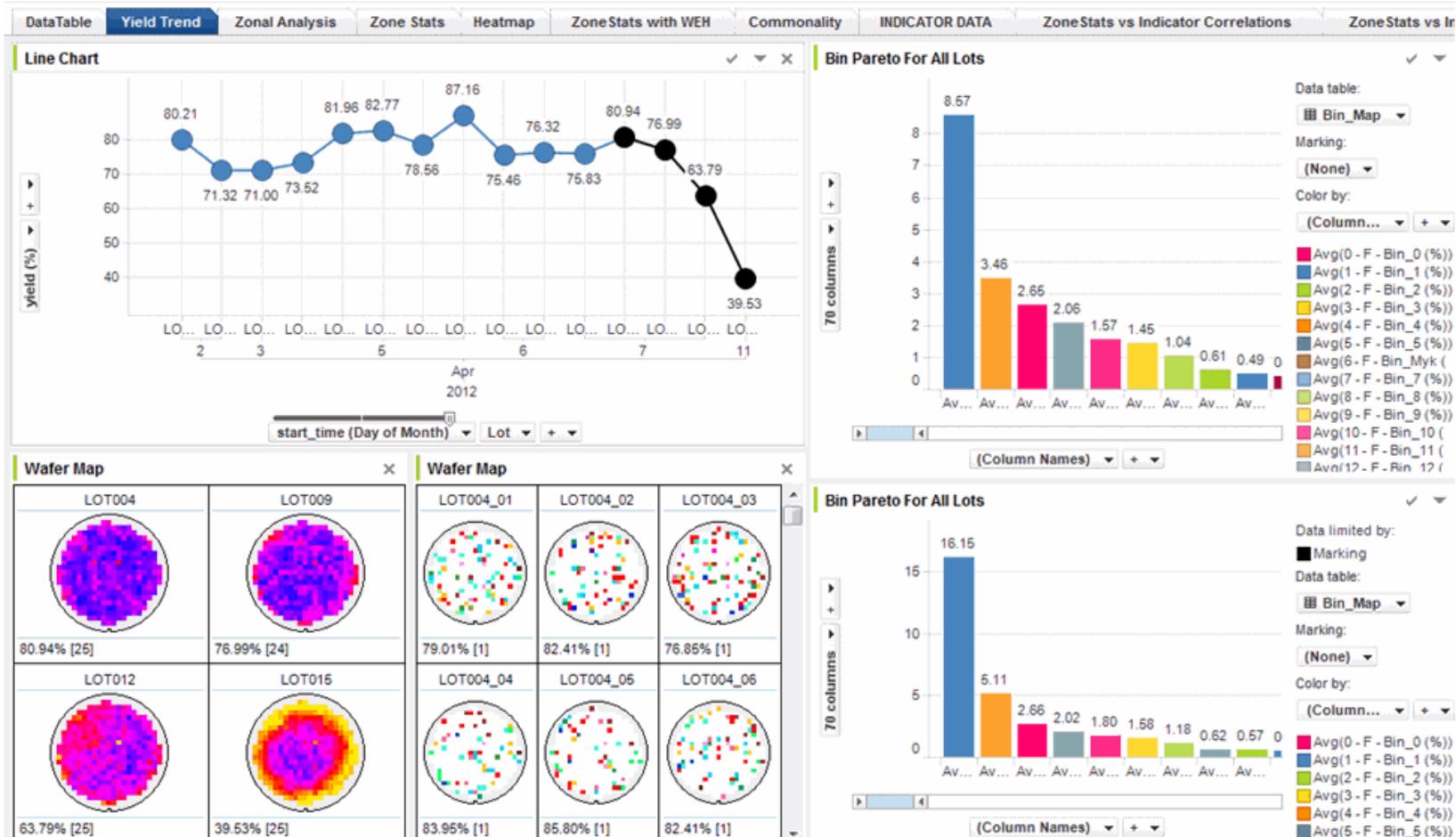
# 데이터 기반 문제 해결 절차

- 3단계: 성급한 모델링 이전에 충분히 데이터를 탐색하라
  - ✓ 데이터 시각화 툴 사용 권장



# 데이터 기반 문제 해결 절차

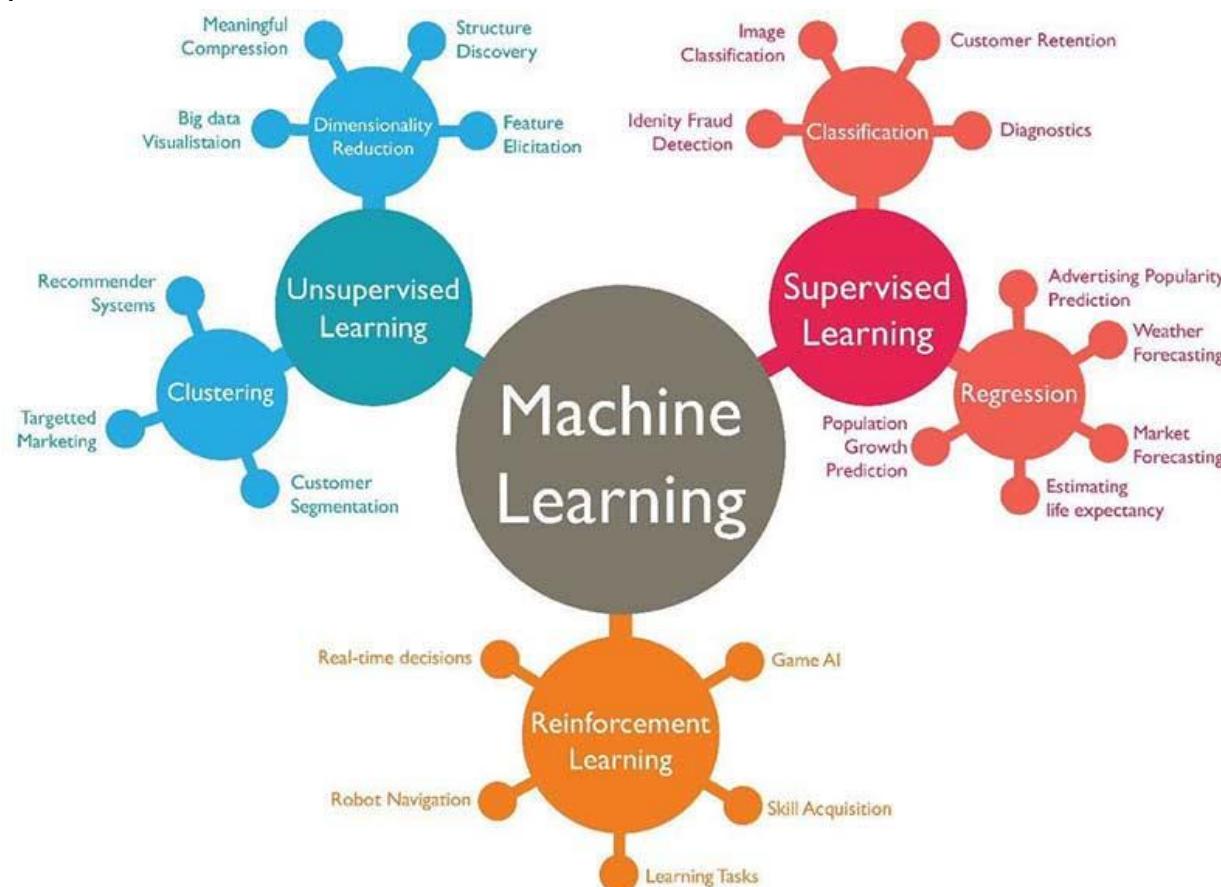
- 3단계: 성급한 모델링 이전에 충분히 데이터를 탐색하라
  - ✓ 데이터 시각화 툴 사용 권장



# 데이터 기반 문제 해결 절차

- 4단계: 모델 구축

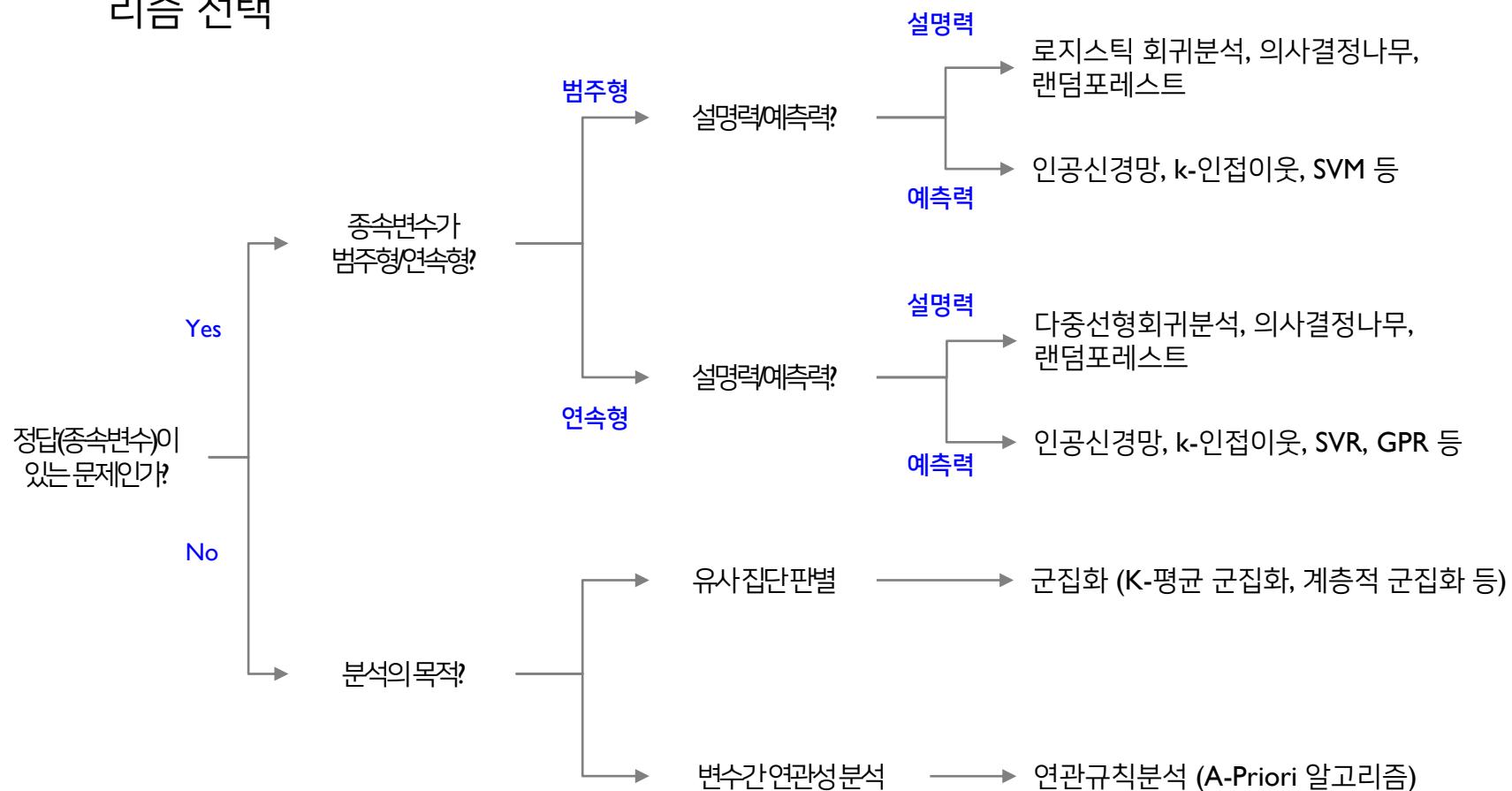
- ✓ 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



# 데이터 기반 문제 해결 절차

## • 4단계: 모델 구축

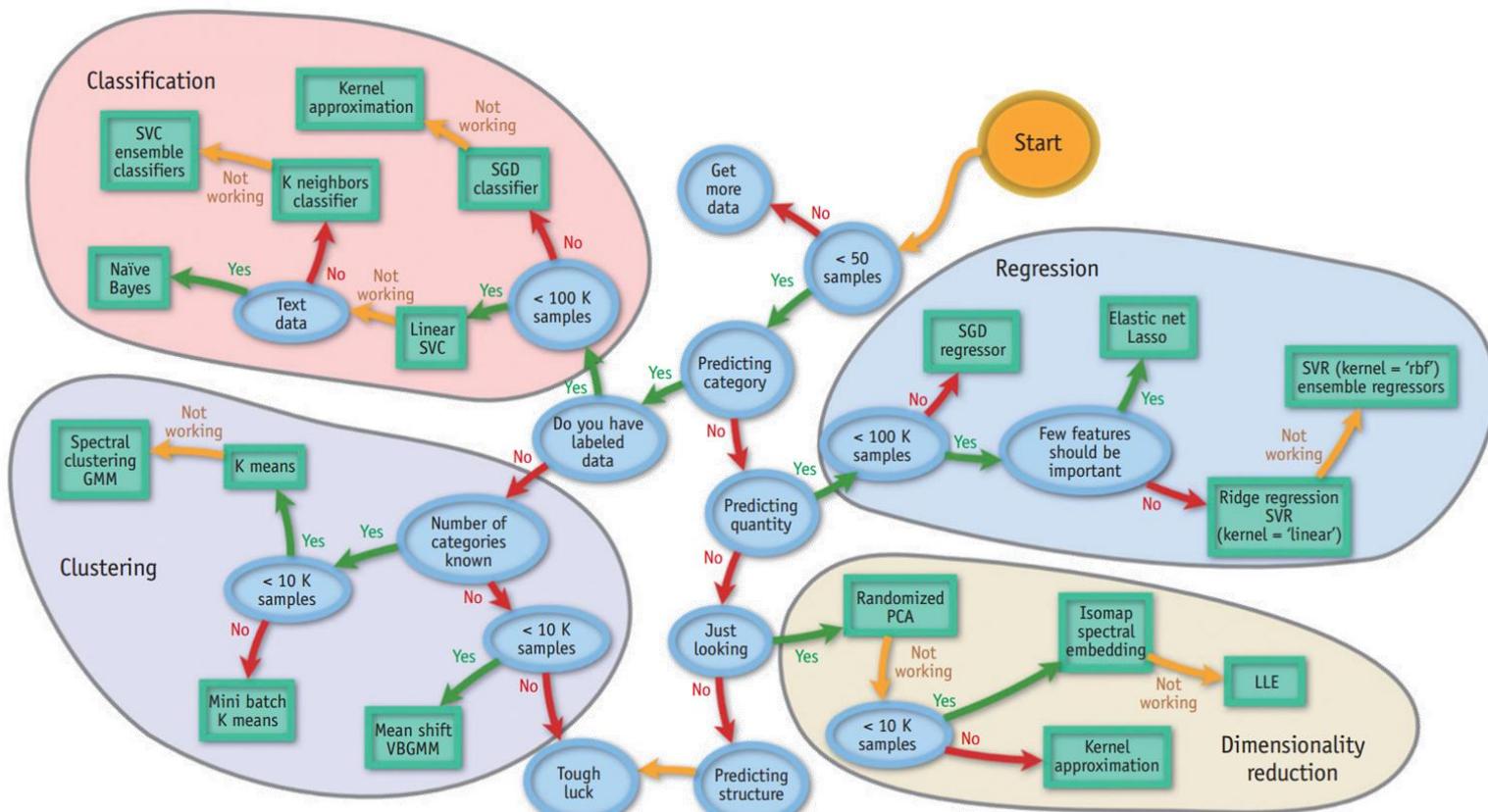
- ✓ 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



# 데이터 기반 문제 해결 절차

## • 4단계: 모델 구축

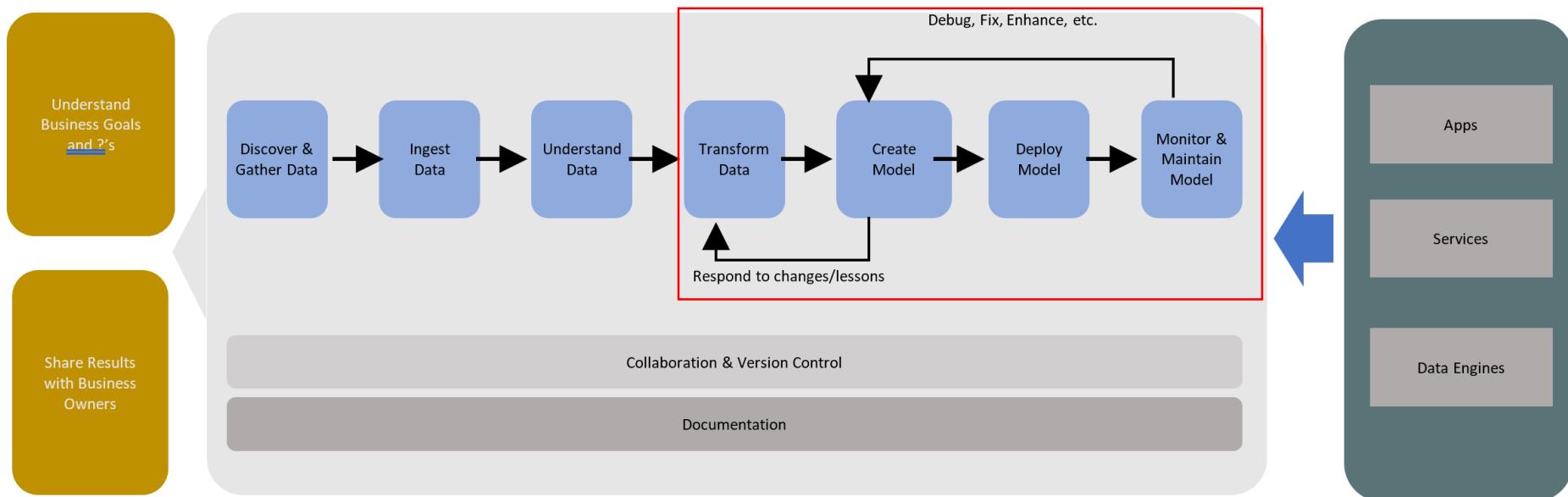
- ✓ 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



# 데이터 기반 문제 해결 절차

- 5단계: 결과 적용

- ✓ 구축된 모델의 시스템 탑재, 시간에 따른 성능 모니터링, 업데이트 주기 결정 등



# 데이터 기반 문제 해결 절차

## • 각 단계별 주요 과업 및 산출물



# AGENDA

01 Data Analytics 개요 및 주요 개념

---

02 데이터 과학 프로젝트 절차

---

03 Machine Learning 방법론

---

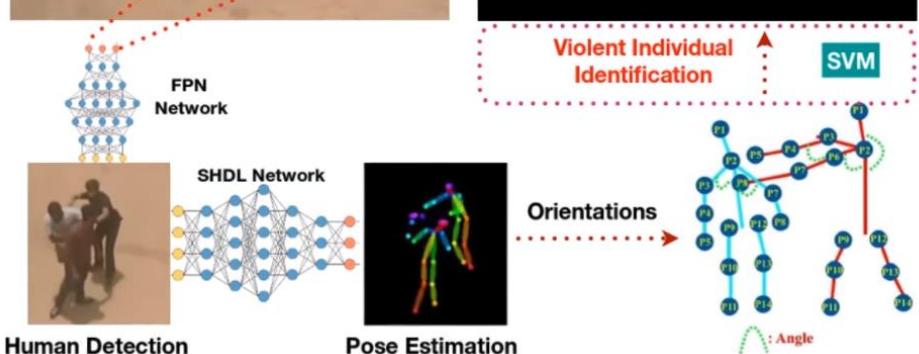
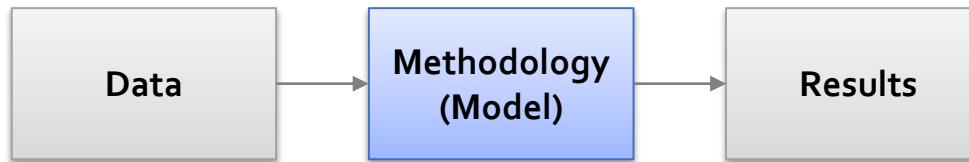
04 PP 기사 분류 모형

---

# 머신 러닝: Machine Learning

- Machine Learning

- ✓ 특정한 과업 Task을 달성하기 위해 경험 Experience이 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



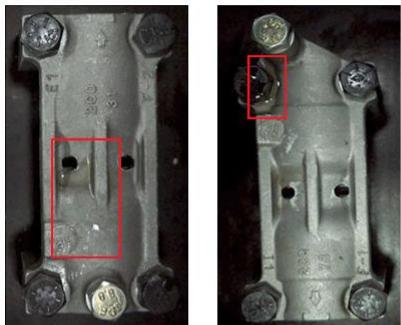
# 머신 러닝: Machine Learning

- 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

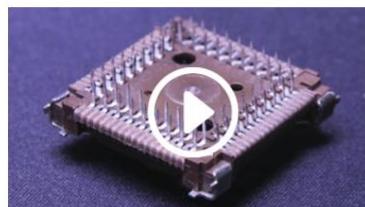
### Surface Defect Detection

#### MULTIPLE MATERIALS

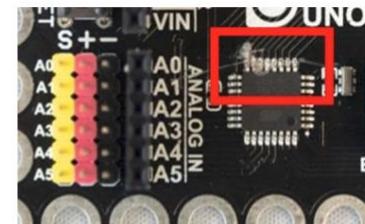


#### ELECTRONIC COMPONENTS

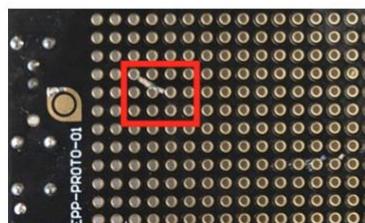
#### SCRATCH



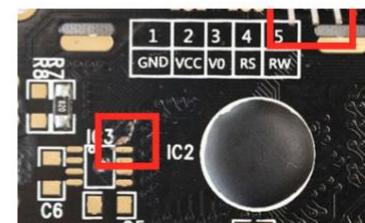
#### SOLDERING



#### SOCKET

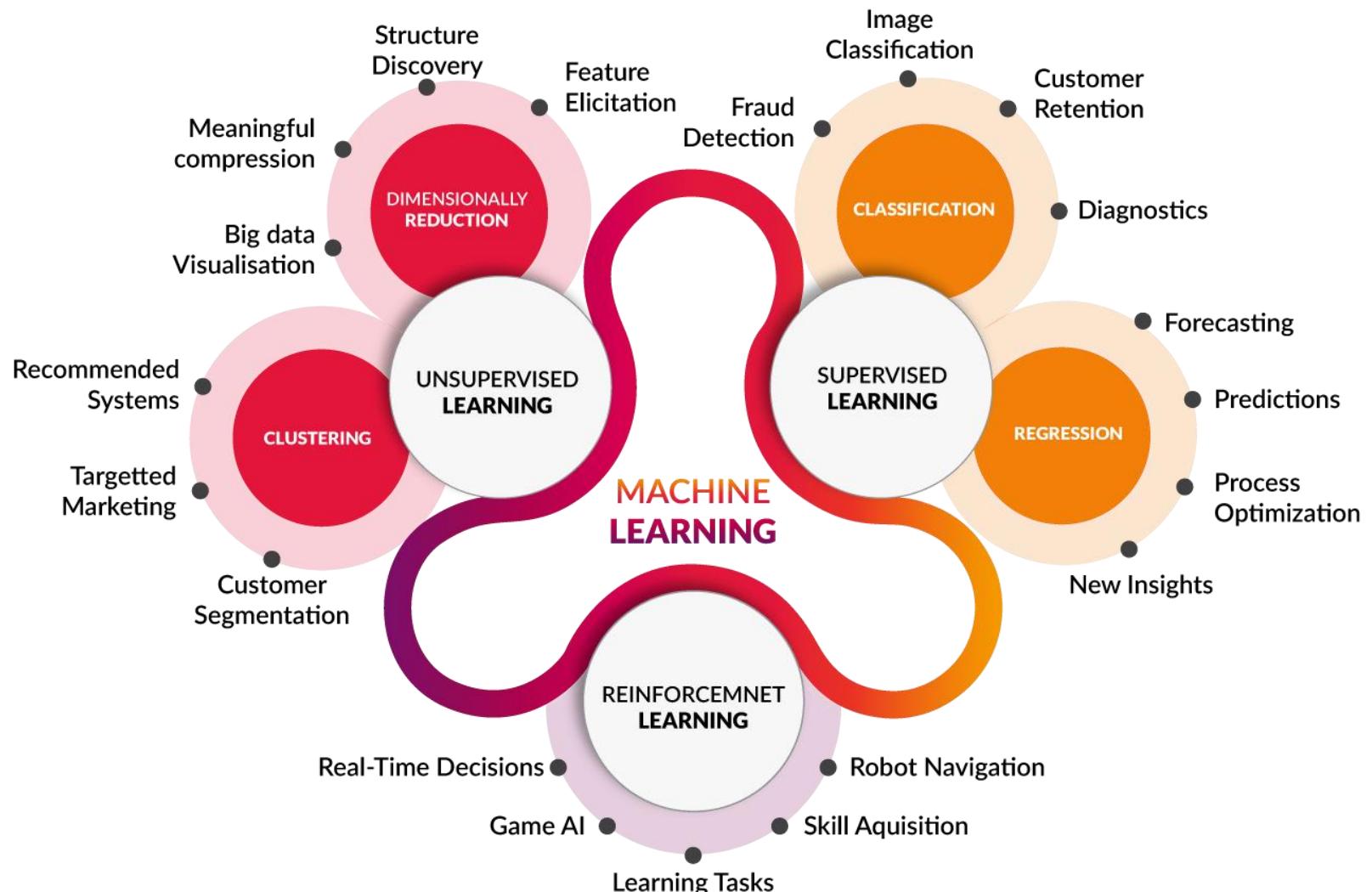


#### BENDING



# 머신 러닝의 종류

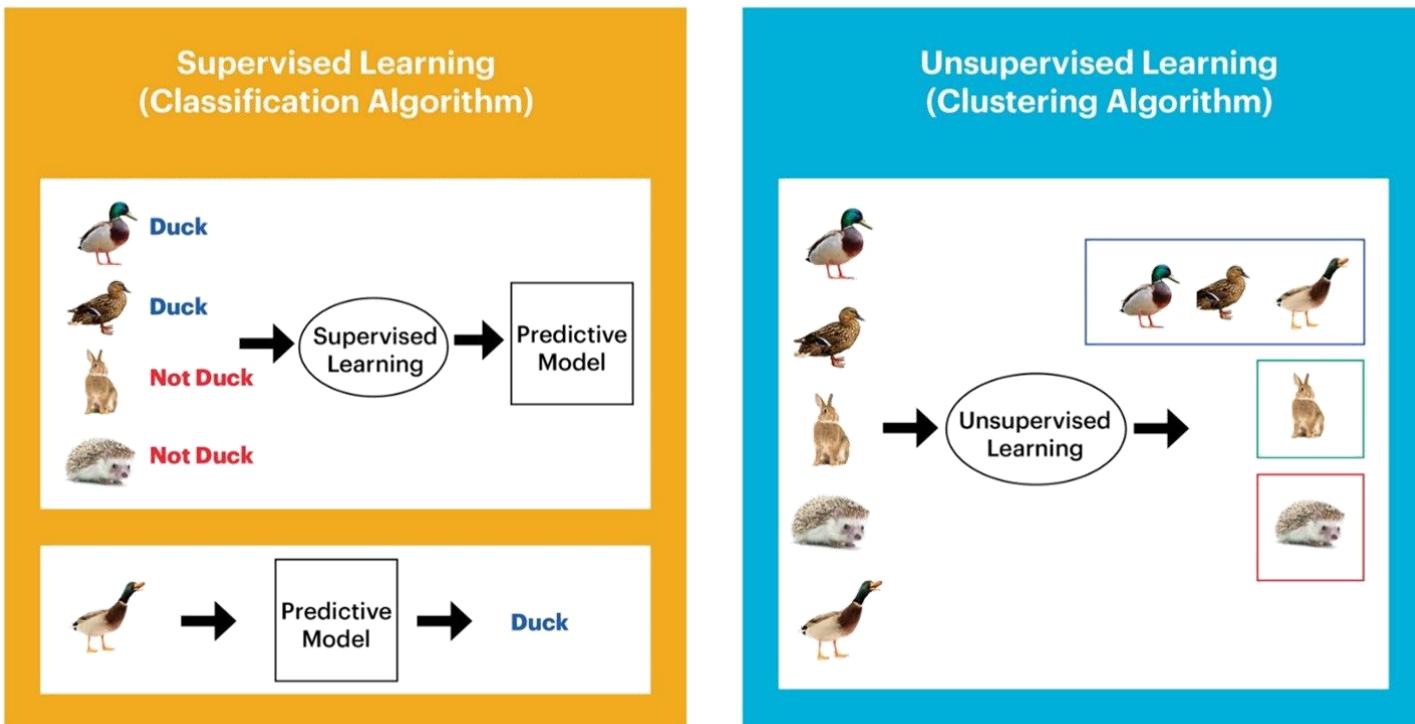
- Machine Learning의 종류



# 머신 러닝의 종류

- 구분기준 I: 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

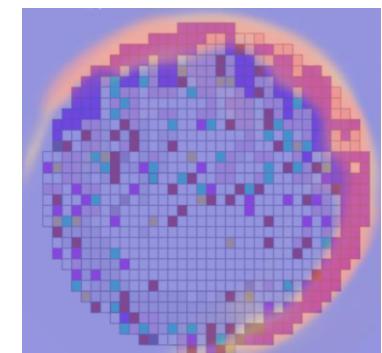
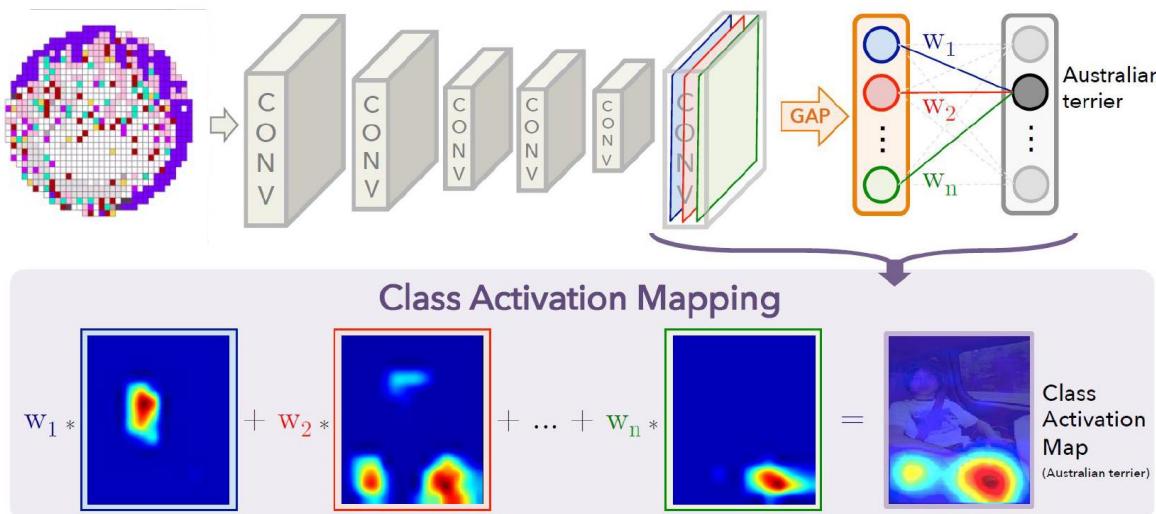
- 구분기준 I: Target (정답)의 유무에 따른 구분

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

입력: WBM

머신러닝 알고리즘: 합성곱 신경망

출력: 불량 유무 및 영역

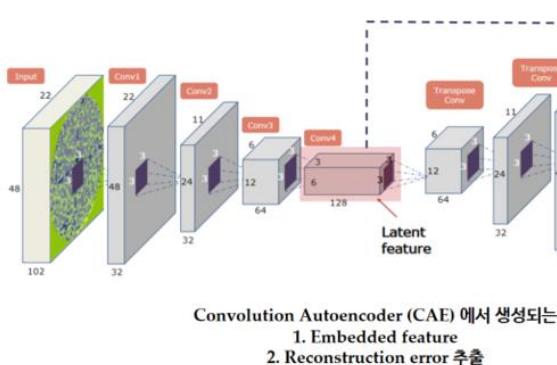


# 머신 러닝의 종류

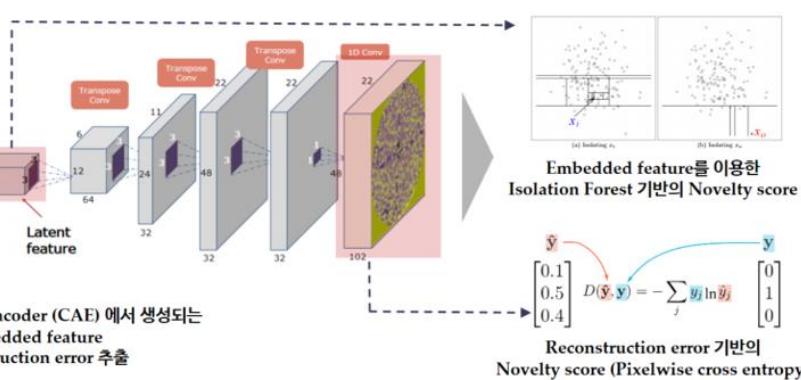
- 구분기준 I: Target (정답)의 유무에 따른 구분

✓ 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

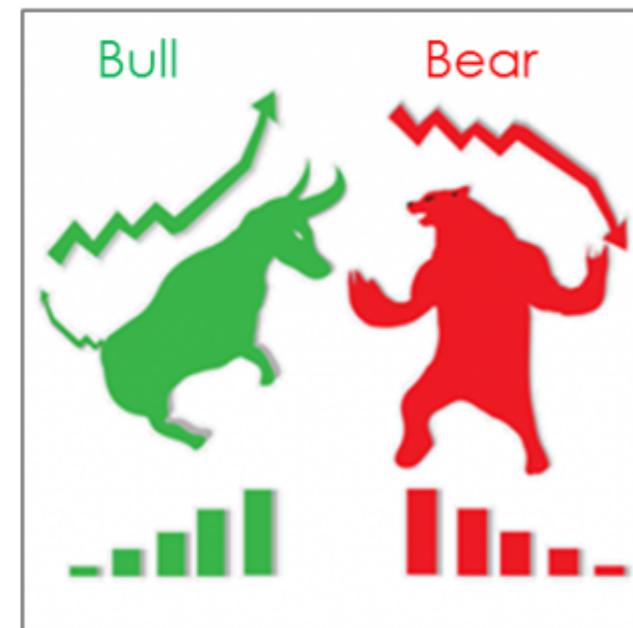
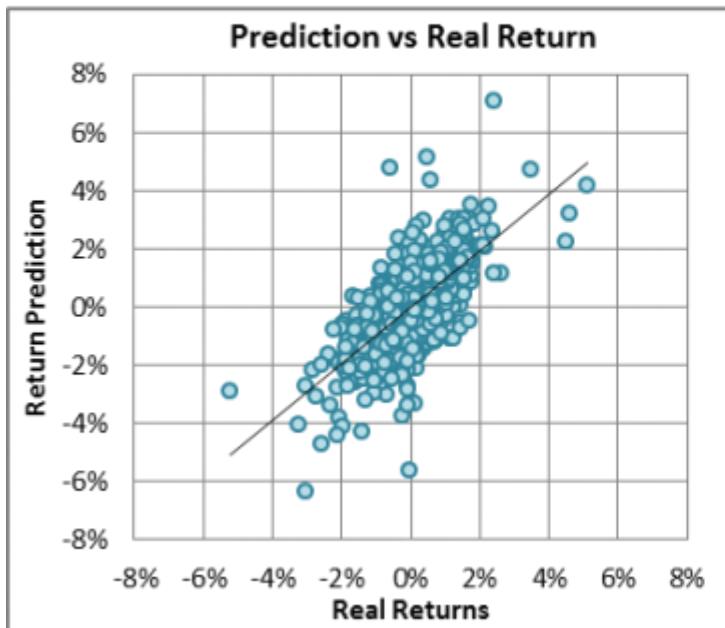
- ✓ Classification (분류) vs. Regression (회귀)

- 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

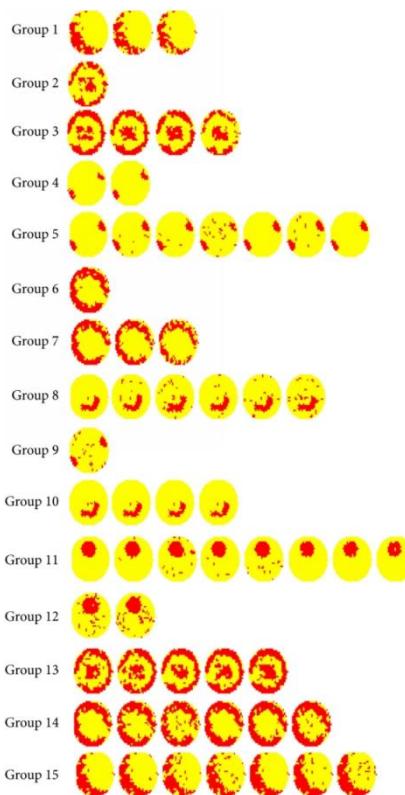


# 머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 군집화(Clustering)

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



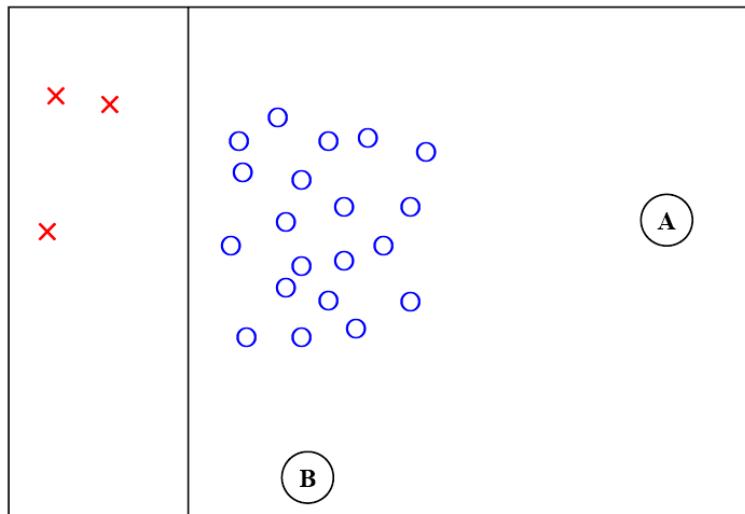
Pattern	Original Map	ESRN (p1=4, p2=6)	ESRN (p1=4, p2=5)	ESRN (p1=5, p2=6)	ESRN (p1=5, p2=5)
Checkerboard					
Ring					
Right-Down Edge					
Composite Pattern					

# 머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

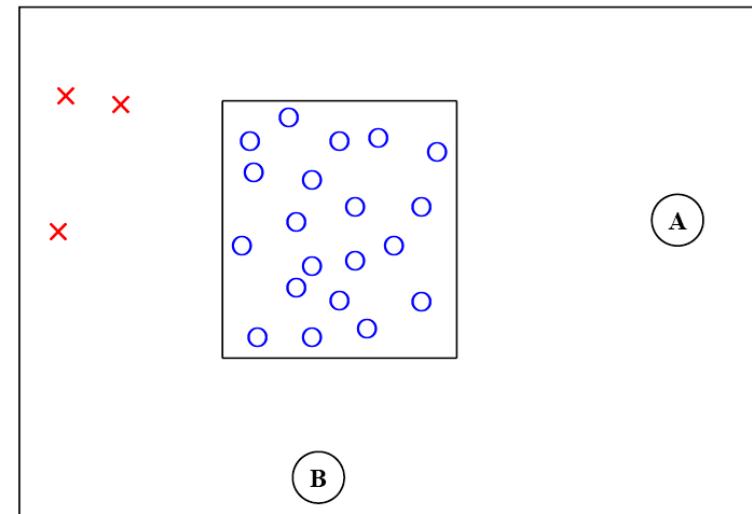
- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection)

- 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

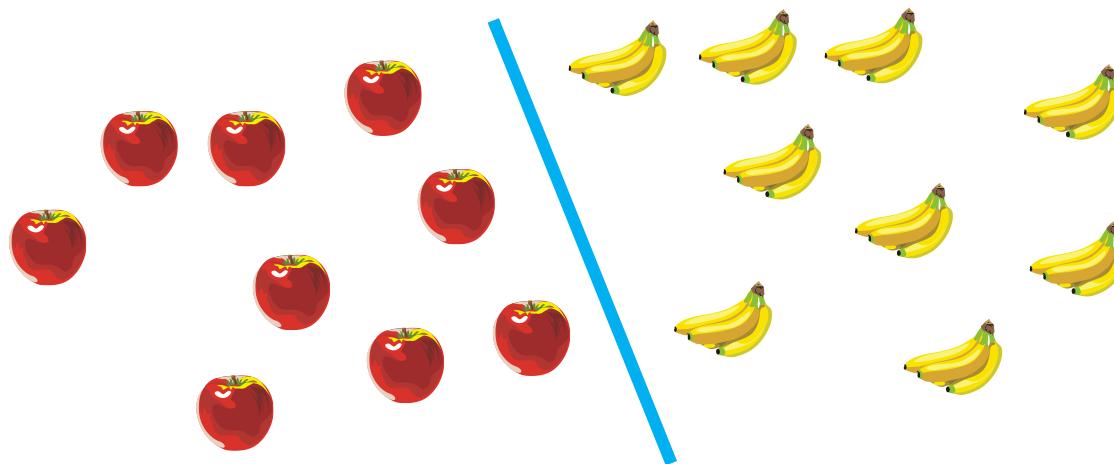
정상 범주에 속하는지 아닌지를 판단

# 머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

- ✓ 이상치 탐지(Novelty Detection, Anomaly Detection)

- 분류 알고리즘이 학습하는 방식

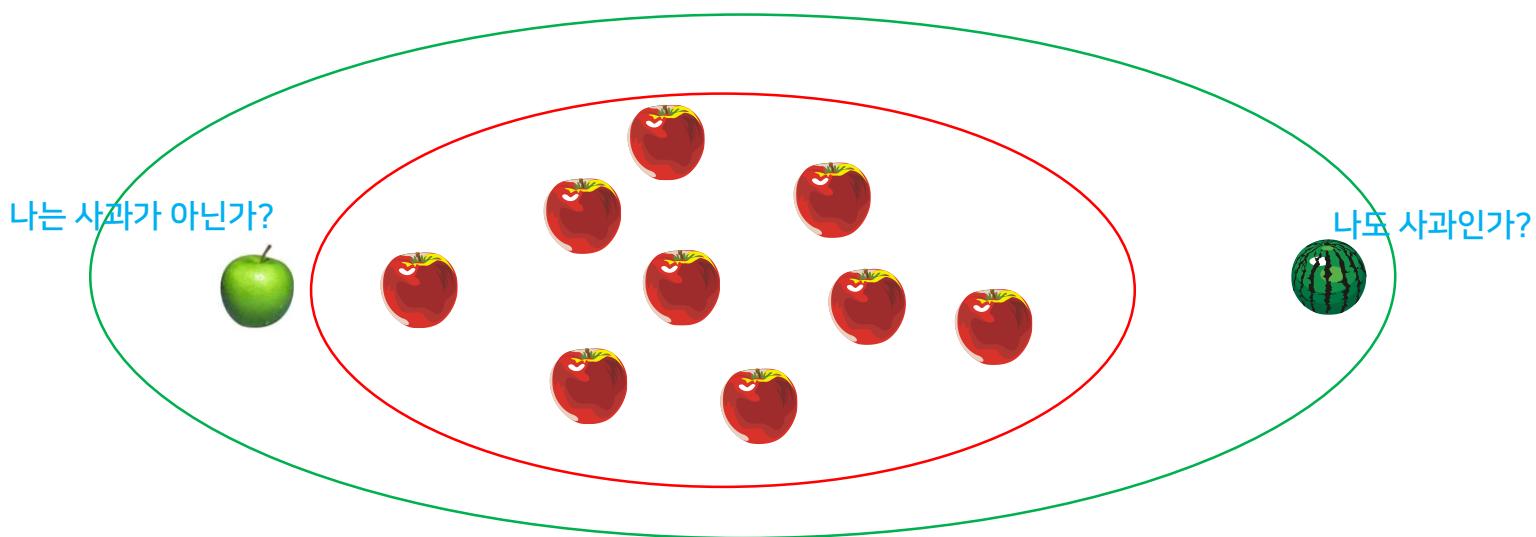


# 머신 러닝의 종류

- 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

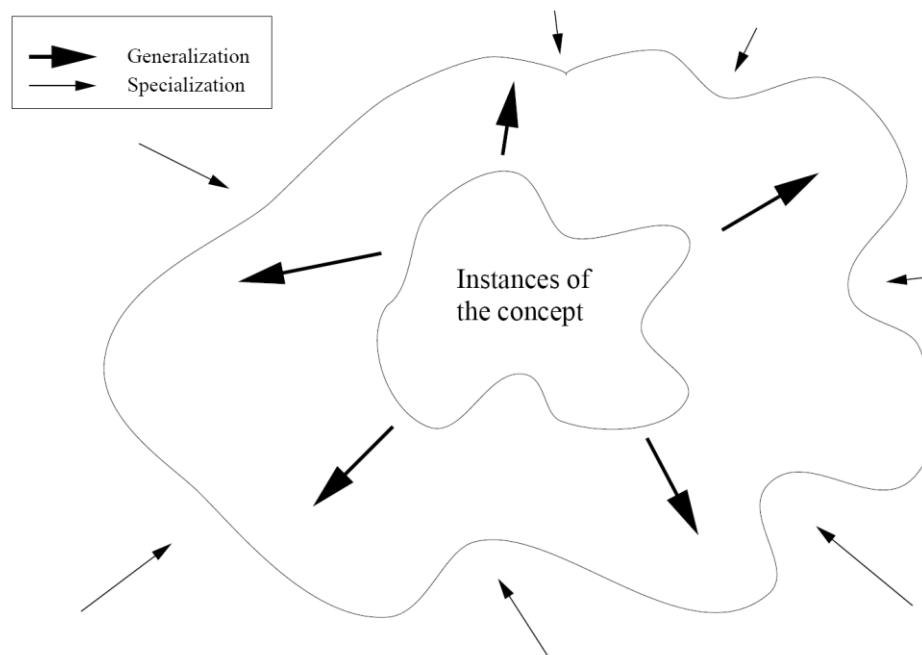


# 머신 러닝의 종류

- 구분기준 2: 학습 목적에 따른 구분

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

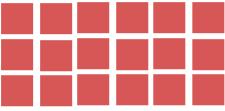
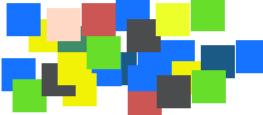
- 일반화: 주어진 데이터로부터 정상 범주의 개념을 확장해 가는 것
    - 특수화: 주어진 데이터로부터 정상 범주의 개념을 좁혀 가는 것
    - 일반화에 치중할 경우 이상치 데이터 판별이 어렵게 되며, 특수화에 치중할 경우 과적합의 위험(빈번한 false alarm)에 빠질 수 있음



# 머신 러닝의 종류

## • 구분기준 3: 사용 데이터에 따른 구분

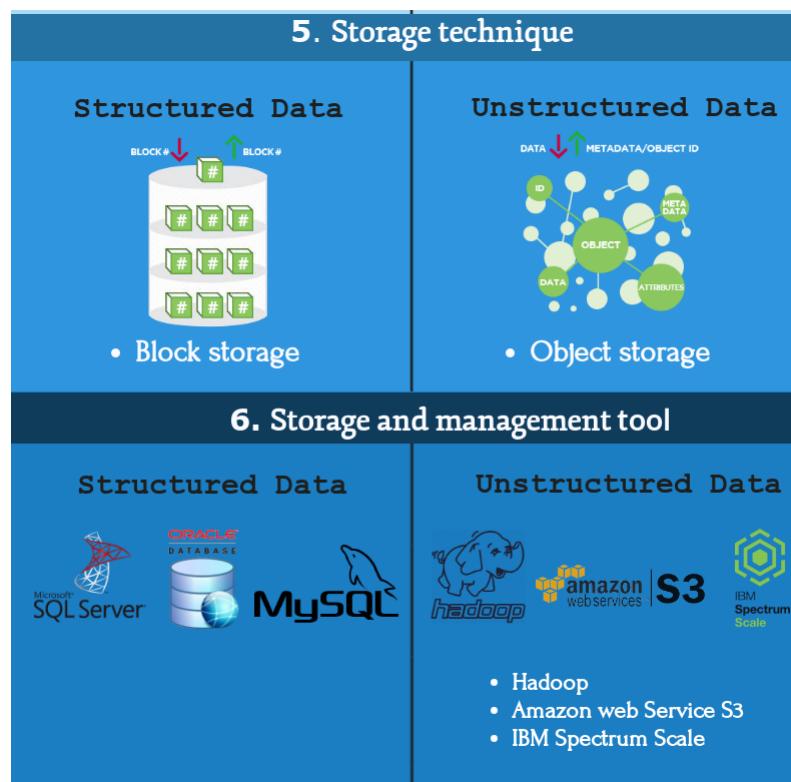
- ✓ 정형 데이터(Structured Data): 기존 방식으로 테이블에 적재된 수치 데이터
- ✓ 비정형 데이터(Unstructured Data): 이미지, 음성, 텍스트 등 숫자가 아닌 형태의 정보가 구조화되지 않은 형태로 존재하는 데이터

1. Definition	3. Growth
<p><b>Structured Data</b></p>  <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> <p><b>Unstructured Data</b></p>  <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>	<p><b>Structured Data</b></p>  <p>Structure data accounts for about 20% of the total existing data.</p> <p><b>Unstructured Data</b></p>  <p>Experts estimate that 80% of the data in any organization is unstructured.</p>
2. Example	4. Characteristic
<p><b>Structured Data</b></p>  <ul style="list-style-type: none"><li>Databases (structuring fields)</li><li>Meta-data (Time and date of creation, File size, Author etc.)</li><li>Census records (birth, income, employment, place etc.)</li></ul> <p><b>Unstructured Data</b></p>  <ul style="list-style-type: none"><li>Website Data which are present in the form of HTML Pages.</li><li>Media ( MP3, digital photos, audio and video files )</li><li>Text files (Word processing, spreadsheets, presentations etc. )</li></ul>	<p><b>Structured Data</b></p> <ul style="list-style-type: none"><li>Schema dependent.</li><li>Scaling DB schema is difficult.</li><li>Robust.</li><li>Structured query allows complex joins.</li><li>Easy to access.</li><li>Organized.</li><li>Efficient to analysis.</li></ul> <p><b>Unstructured Data</b></p> <ul style="list-style-type: none"><li>Absence of schema.</li><li>Very flexible.</li><li>Highly scalable.</li><li>Only textual query possible.</li><li>Hard to access.</li><li>Scattered and dispersed.</li><li>Additional preprocessing is needed.</li></ul>

# 머신 러닝의 종류

- 구분기준 3: 사용 데이터에 따른 구분

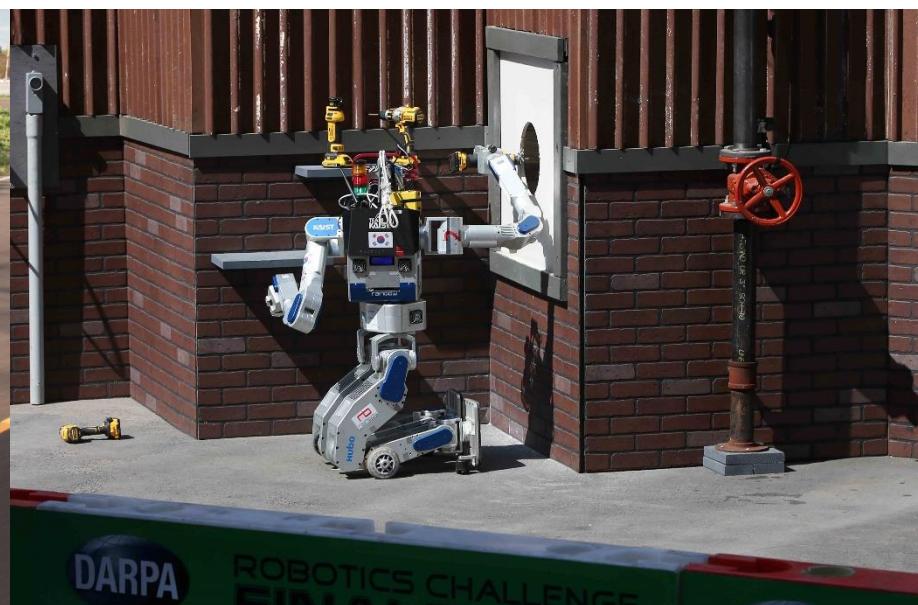
- ✓ 정형 데이터(Structured Data): 기존 방식으로 테이블에 적재된 수치 데이터
- ✓ 비정형 데이터(Unstructured Data): 이미지, 음성, 텍스트 등 숫자가 아닌 형태의 정보가 구조화되지 않은 형태로 존재하는 데이터



# 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



# AGENDA

01 Data Analytics 개요 및 주요 개념

---

02 데이터 과학 프로젝트 절차

---

03 Machine Learning 방법론

---

04 PP 기사 분류 모형

---

# PP 기사 분류 모델

- 문제 인식

- ✓ 전체 기사 중에서 PP기사로 선택되는 비중은 2017년 1월 1일부터 2020년 1월 30일까지 총 1,090,698건 중 52,899건(4.85%)으로 매우 낮은 비중을 가짐

- 가정

- ✓ 스크랩 업무를 하는 전문 인력들은 각자가 PP기사로 선택 기준이 있을 것이다.
  - ✓ 그 기준으로부터 정량화할 수 있는 변수를 생성하여 Machine Learning 모형을 학습하게 되면 높은 정확도로 PP기사를 자동으로 선별할 수 있는 방법론을 개발할 수 있을 것이다.

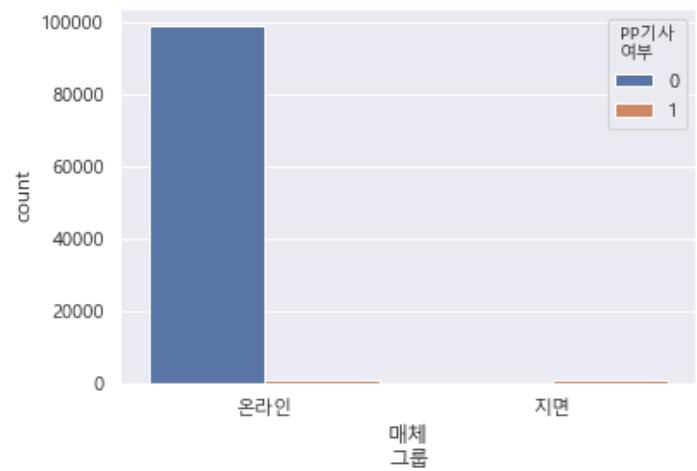
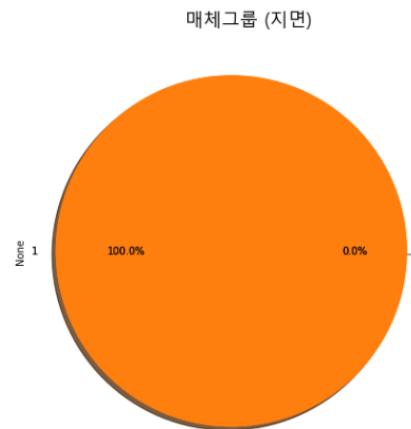
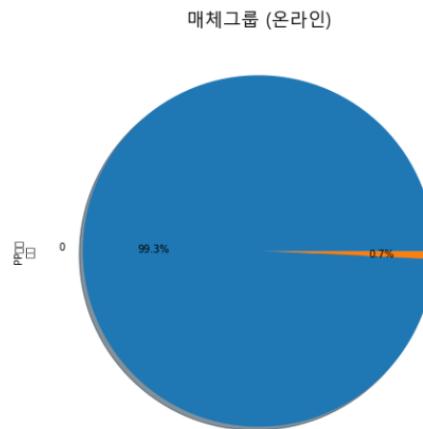
## PP 기사 분류 모델

#### • 데이터 예시

# PP 기사 분류 모델

- 데이터 탐색 (by 임지수 프로)

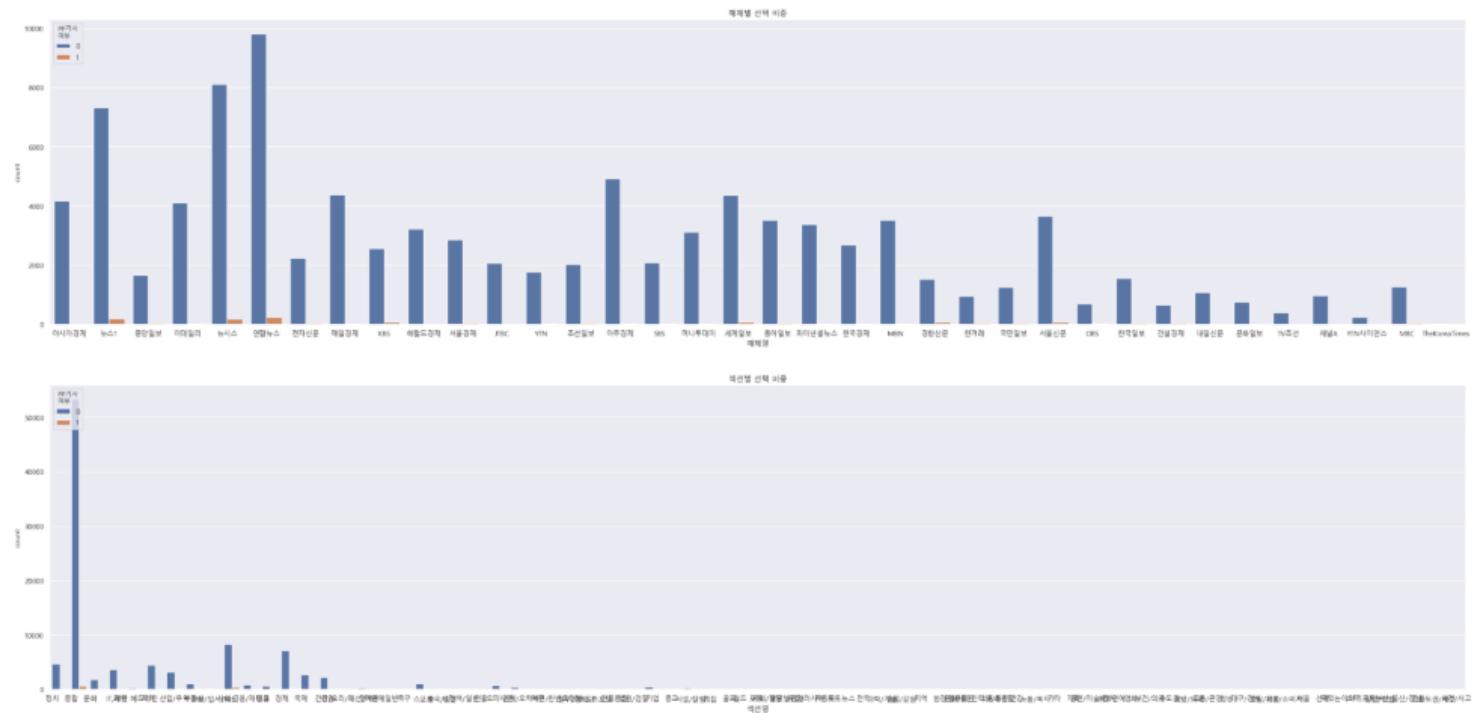
✓ Finding I: 지면 기사는 전부 PP 기사임



## PP 기사 분류 모델

- 데이터 탐색 (by 임지수 프로)

✓ Finding 2: 매체별로 PP기사로 선택될 확률이 다름



# PP 기사 분류 모델

## • 설문조사를 통한 변수 생성

✓ 총 51명의 대상자 중 27명이 응답

### ■ 설문지 예시

스크랩 할 때 기사가 고객과 관련이 있는지를 판단할 때 어느 부분을 살펴보는지 다음 항목 중에서 모두 골라주세요.\*

(중복 체크 가능)

- 1. 기사 제목
- 2. 기사 부제목 (리드문)
- 3. 기사 본문 첫 문단
- 4. 기사 본문 마지막 문단
- 5. 기사 본문 전체
- 6. 기사 내 사진 및 그래프
- 7. 기사가 들어간 섹션(정치, 경제, 사회....)
- 기타: \_\_\_\_\_

앞선 질문의 보기 중에서 여러분들이 스크랩을 할 때 살펴 보는 가장 많이 보는 것 3 개를 순서대로 적어주세요.\*

예시) 1 - 3 - 2

내 답변 \_\_\_\_\_

기사가 고객과 관련된 기사라고 판단해서 1차로 워크보드로 선별할 때 어떤 기준을 활용해서 판단하시는지 다음 항목 중에서 모두 골라주세요.\*

(중복 체크 가능)

- 1. 고객 관련 핵심단어의 일치
- 2. 고객 관련 핵심단어의 반복
- 3. 고객 관련 기사 섹션
- 4. 기사의 주제, 핵심 문장의 고객사 관련 여부

# PP 기사 분류 모델

- 설문조사를 통한 변수 생성

- ✓ 총 51명의 대상자 중 27명이 응답: 응답 결과 예시

1번 문항	
답안	빈도
0. 기타	2
1. 기사 제목	26
2. 기사 부제목(리드문)	21
3. 기사 본문 첫 문단	14
4. 기사 본문 마지막 문단	7
5. 기사 본문 전체	17
6. 기사 내 사진 및 그래프	9
7. 기사가 들어간 섹션(정치, 경제, 사회....)	3

2번 문항					
답안	1순위	2순위	3순위	4순위	총점
0. 기타	1	1	0	0	7
1. 기사 제목	21	4	1	0	98
2. 기사 부제목(리드문)	1	13	5	0	53
3. 기사 본문 첫 문단	0	6	6	0	30
4. 기사 본문 마지막 문단	0	2	0	0	6
5. 기사 본문 전체	4	2	7	1	37
6. 기사 내 사진 및 그래프	0	0	4	0	8
7. 기사가 들어간 섹션(정치, 경제, 사회....)	0	1	0	0	3

\* 1,2번 문항 기타  
1순위 : 고객사명이 들어간 모든 기사(키워드)  
2순위 : 검색 키워드가 들어간 문단

# PP 기사 분류 모델

- Machine Learning 학습 데이터셋 구조

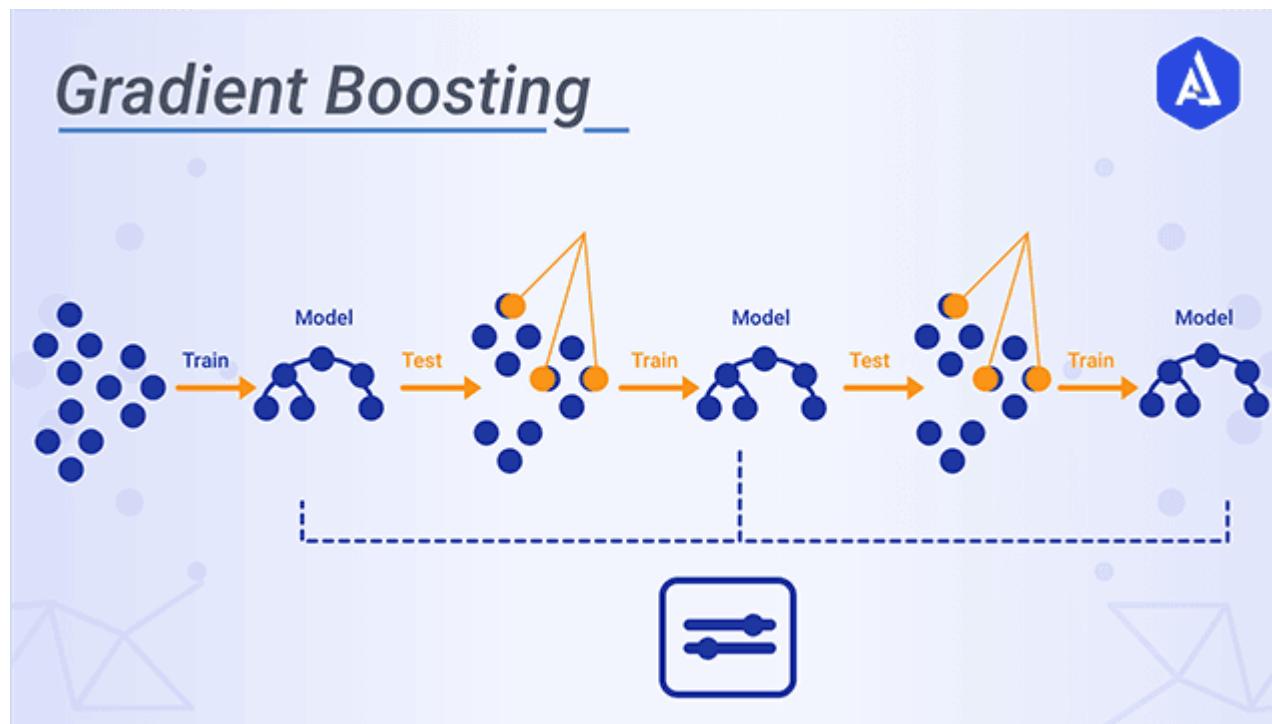
변수명	Type	설명
Ggroup	Category	기사 구분 (지면, 온라인, 사용자요청매체 등)
매체 선호도	Continuous	온라인 기사들 중에서 SCT 코드를 기준으로 PP 데이터 선정률을 변수로 사용 예시1) AB006 (매일경제) 온라인 기사 1,000건, PP 포함 10건, 변수값 = $100*10/1000 = 1.00 (\%)$ 예시2) AB031 (이투뉴스) 온라인 기사 1,000건, PP 포함 1건, 변수값 = $100*1/1000 = 0.10 (\%)$
카테고리 선호도	Continuous	온라인 기사들 중에서 NCT 코드를 기준으로 PP 데이터 선정률을 변수로 사용 예시1) AB01 (정부/청와대) 온라인 기사 1,000건, PP 포함 50건, 변수값 = $100*50/1000 = 5.00 (\%)$ 예시2) AY01 (스타포토) 온라인 기사 1,000건, PP 포함 1건, 변수값 = $100*1/1000 = 0.10 (\%)$
제목 키워드 포함 여부	Binary	제목에 키워드가 1번 이상 포함되어 있으면 1, 아니면 0
제목 키워드 포함 수	Integer	제목에 키워드가 포함된 횟수 (중복 고려)
본문 키워드 포함 여부	Binary	본문에 키워드가 1번 이상 포함되어 있으면 1, 아니면 0
본문 키워드 포함 수	Integer	본문에 키워드가 포함된 횟수 (중복 고려)
기사 본문 첫 문단 키워드 포함 여부	Binary	기사 본문 첫 문단에 키워드가 1번 이상 포함되어 있으면 1, 아니면 0
기사 본문 첫 문단 키워드 포함 수	Integer	기사 본문 첫 문단에 키워드가 포함된 횟수 (중복 고려)
사진 유무	Binary	1: 기사에 사진 포함, 0: 기사에 사진 없음
제목에 "보도자료" 단어 포함 유무	Binary	1: 있으면 1, 없으면 0
제목에 "특집" 단어 포함 유무	Binary	1: 있으면 1, 없으면 0
제목에 "기획" 단어 포함 유무	Binary	1: 있으면 1, 없으면 0
본문 길이 (글자수)	Integer	본문의 총 글자 수 (단어 기준 아님, 글자 기준, 공백 포함)
본문 길이 (단어 수)	Integer	본문의 총 단어 수 (공백 제외)
본문 명사 비중 (단어 기준)	Integer	(본문에 사용된 총 명사 수)/(본문에 사용된 단어 수)
본문 Unique 명사 비중 (단어 기준)	Integer	(본문에 사용된 Unique 명사 수)/(본문에 사용된 단어 수)

# PP 기사 분류 모델

- Machine Learning 모델

- ✓ Logistic Regression, Decision Tree, Random Forest, XGBoost 등의 분류 알고리즘 중 최종적으로 XGBoost 알고리즘 사용

- 학습 데이터: 2017년 1월 1일~2018년 12월 31일 발생 기사 (24개월치)
    - 테스트 데이터: 2019년 1월 1일~2020년 1월 31일 발생 기사 (13개월치)



# PP 기사 분류 모델

- 모델 보정

- ✓ Ggroup 변수값이 F가 아니고 XGBoost 알고리즘 예측이 1이 아닐 경우, Logistic Regression, Decision Tree, Random Forests 세 가지 알고리즘 중 두 개 이상이 1로 예측하면 최종 예측을 1로 수정

- ✓ Rule 추가 전

		모델 예측	
		PP기사	Non-PP 기사
정답	PP 기사	23,381	2,418
	Non-PP	29,833	282,311

- ✓ Rule 추가 후

		모델 예측	
		PP기사	Non-PP 기사
정답	PP 기사	23,496	2,303
	Non-PP	30,327	281,817

# PP 기사 분류 모델

## • 예측 모델의 효과

✓ 비지면 기사에 대한 Random model과의 비교

- 비지면 기사의 경우 테스트기간 동안 총 325,293건의 기사가 존재하며, 이 중 13,149건만 PP기사로 선택됨 (선택율 4.21%)

		모델 예측		Random Model	
		PP기사	Non-PP 기사	PP기사	Non-PP 기사
정답	PP 기사	23,496	2,303	577	13,118
	Non-PP	30,327	281,817	13,118	298,480

- 예측 모델 적용 시 PP기사 검출율은 4.21% → 91.07%로 86.86%p 상승
- 실제 pp기사 중 8.94%가 검출되지 않음 (실질적인 작업 수행 시 고려사항)
- Non-PP 기사에 대해서는 9.72% 정도가 PP 기사로 잘못 예측됨
- 모델이 예측한 PP 기사 중 43.65%가 실제 PP 기사임



ANY  
questions?