Introduction to Business Analytics: Data Science Process

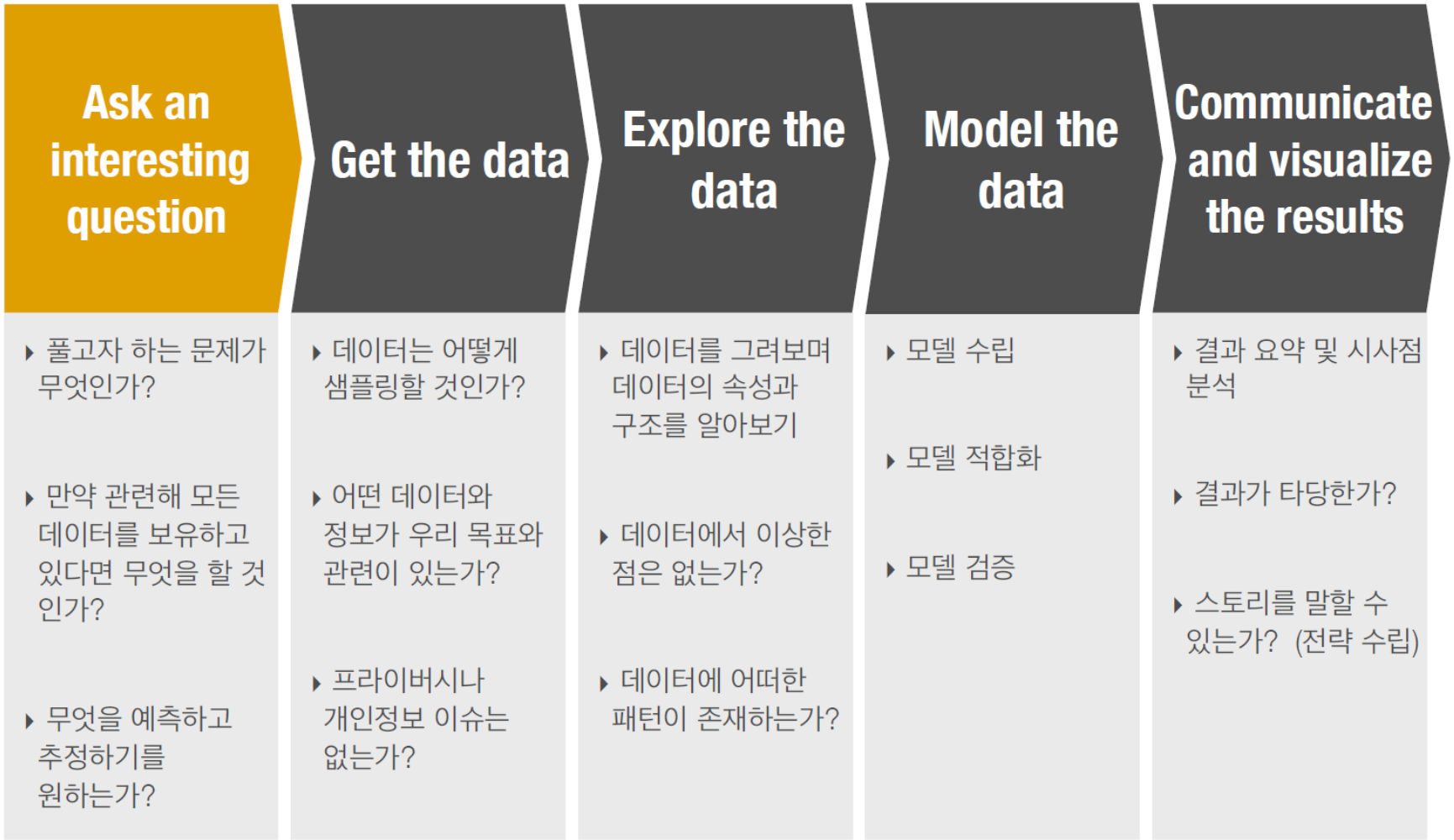
강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

AGENDA

DSBA Data Science & Business Analytics

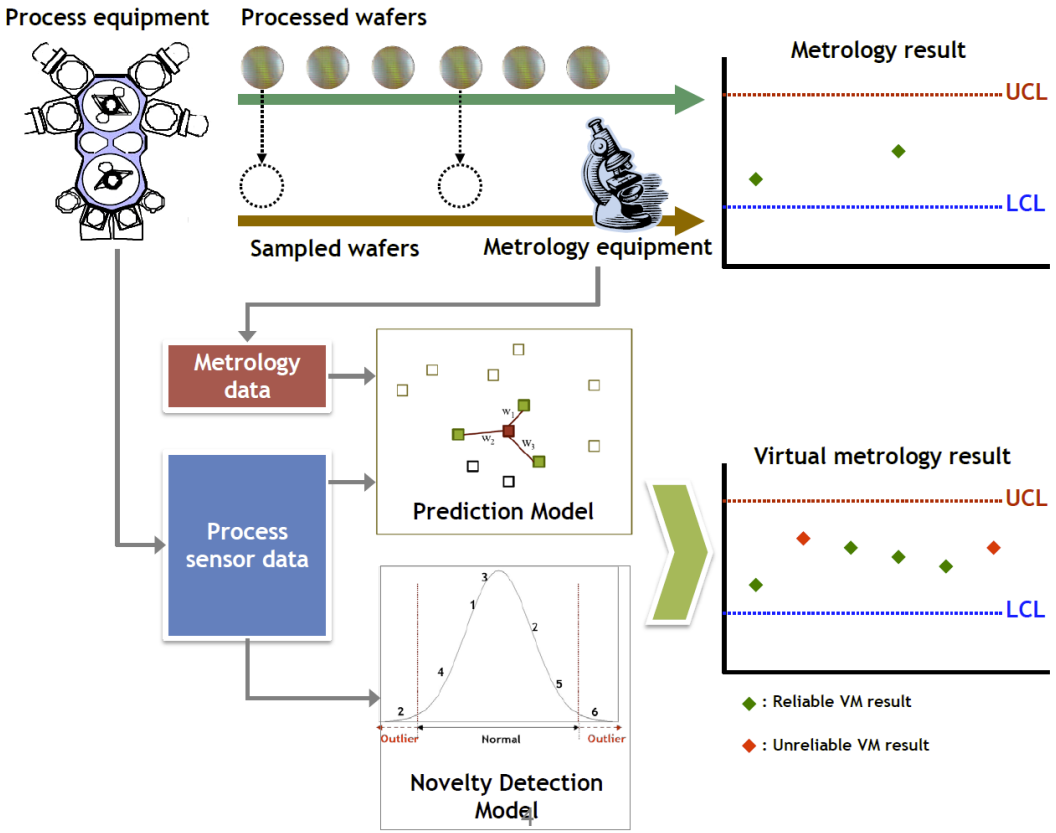
데이터 기반 문제 해결 절차

- 데이터 기반의 문제해결 5단계



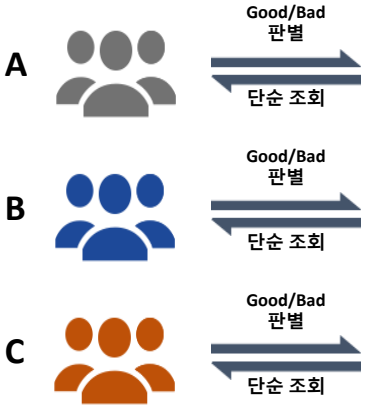
데이터 기반 문제 해결 절차

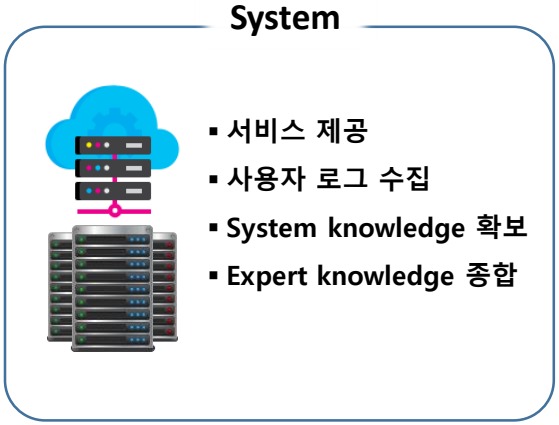
- 1단계: 문제 정의
  - ✓흥미로운 문제(매출 증대, 비용 감소, 공정 단축 등 해결될 경우 조직에 도움이 될 것으로 예상되는 문제)를 발굴할 것
  - ✓예) 공정 중에 발생한 불량 원인은 장비에 기록되는가?

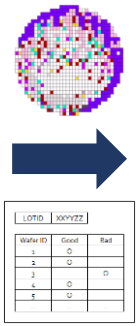


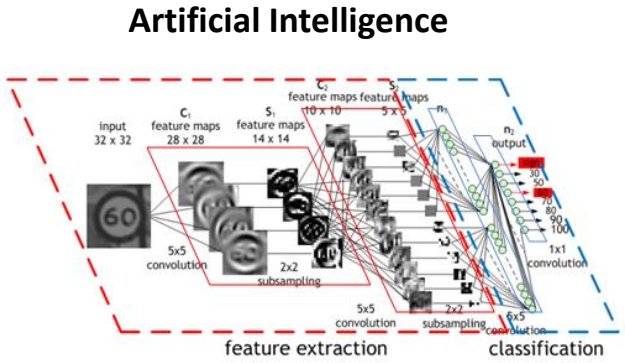
데이터 기반 문제 해결 절차

- 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?

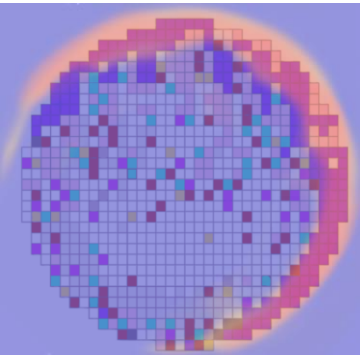
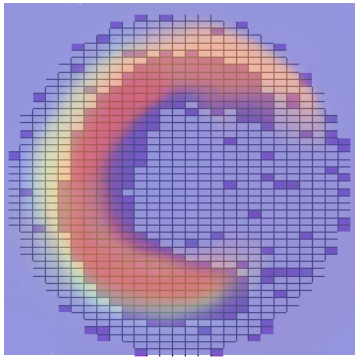
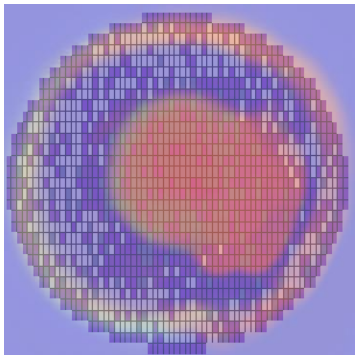
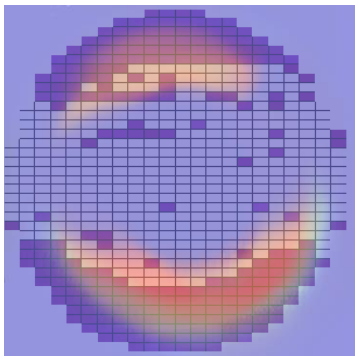
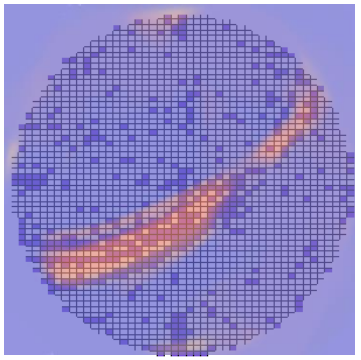






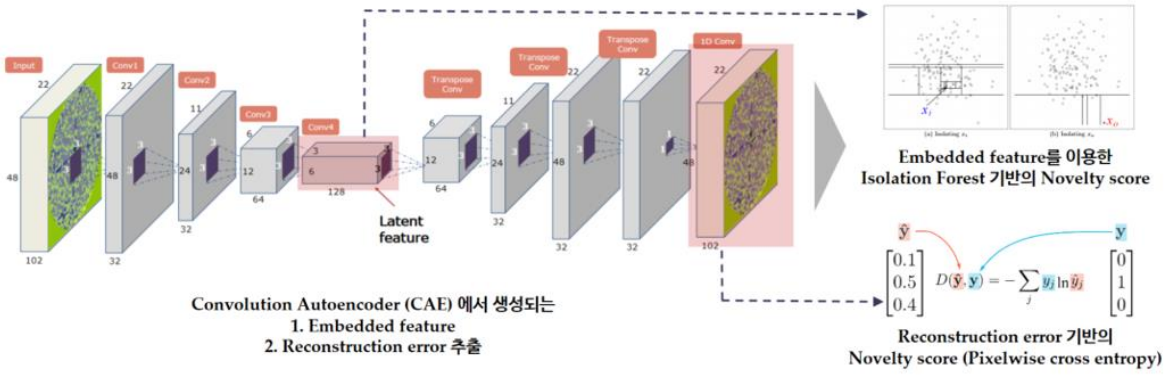


[Deep learning 기반 wafer image 분류기 개념]

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?



오늘 생성된 WBM중에서

이상치는 어떤 것일까?

오늘 생성된 WBM중에서

과거 WBM 패턴으로 보았을 때,
이상치는 어떤 것일까?

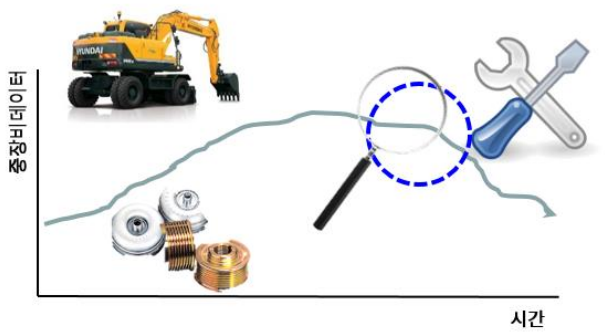
오늘 생성된 WBM중에서

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

데이터 기반 문제 해결 절차

- 1단계: 문제 정의
  - 예) 중장비 가동 데이터로부터 부품 고장 예측, 유효 알람 파악이 가능한가?

중장비 고장 예측 모델 구축



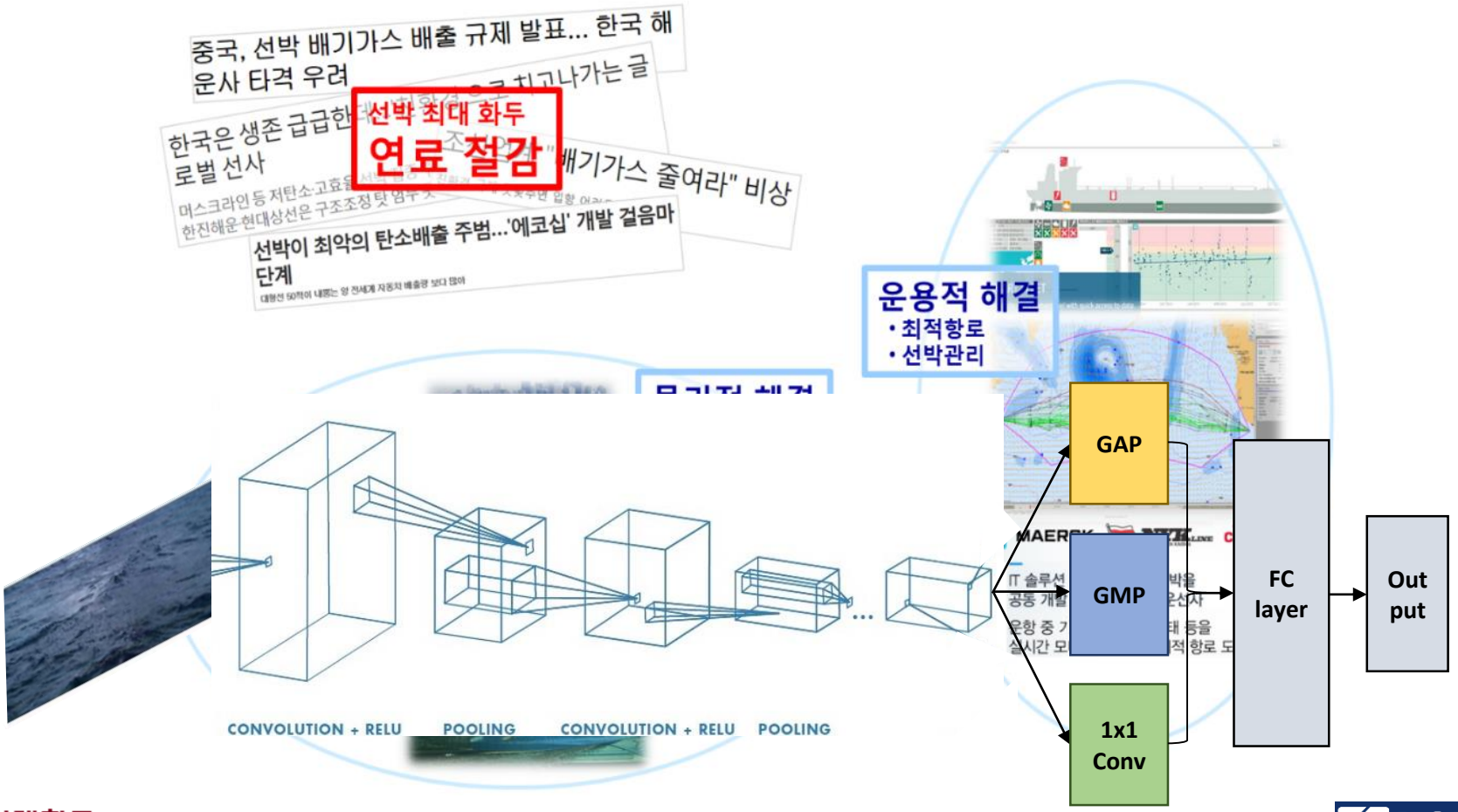
- 하이메이트의 장비 별 데이터와 실제 고장이력을 연동하여 **장비상태에 따른 고장 예측 모델 구축**
- 장비의 부품 별로 고장을 예측하고 고장에 영향을 주는 변수의 시각화를 통해 **고장 패턴에 대한 지식 전달**

알람 패턴 분석

- 장비의 고장 이력 등을 연동하여 **알람의 종류를 정의**
- **알람의 종류를 예측**하여 불필요한 알람과 중요한 알람에 대한 가이드 제시
- 동시에 발생하는 **알람간의 연관성을 분석**하여 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(용융아연도금강판공장)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 쓸 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.

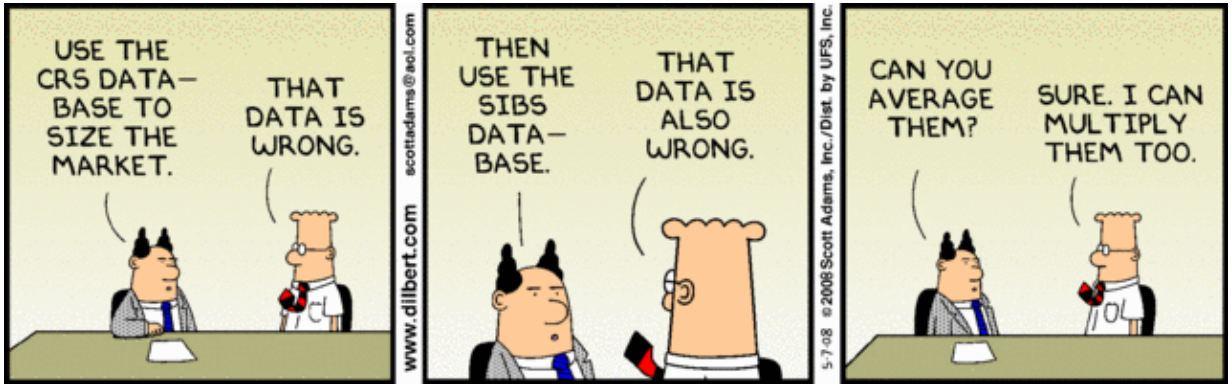
우선 해결해야 하는 문제를 아주 명확히 정의하고 해당 문제를 해결하기 위한 요소들에 대한 자세한 초기 설명을 제공해 준 기술연구소 연구원 분들과 현 광양기술연구소장님께 감사한 마음이 든다. 이 분들은 꾸준히 문제를 함께 해결하기 위한 관심과 노력을 아끼지 않으신 분들이다. 함께 밤 늦은 시간까지 제철소 내에서 고민했고, 밤을 새기도 했다. 알고리즘을 위한 전용 PLC 설치와 DBMS 설치를 지속적으로 도와준 협력업체 분들께도 감사하다. 데이터 수집항목, 수집주기 등이 과제를 수행하는 동안 몇 번 바뀌었는데 그 요구사항들을 신속하게 처리해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



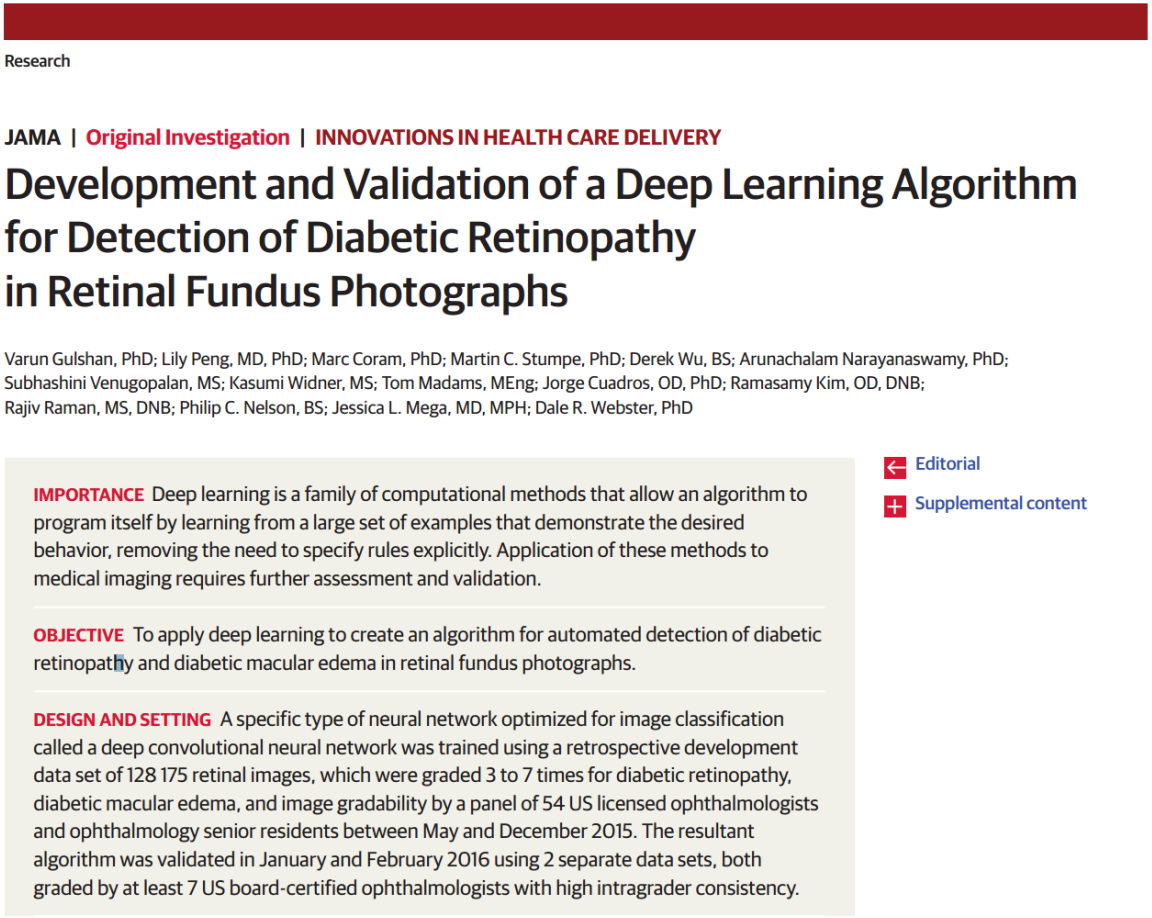
  - The larger, the better

"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)



Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

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

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

Editorial

Supplemental content

데이터 기반 문제 해결 절차

- 2단계: 분석에 적합한 데이터를 수집하라

Table. Baseline Characteristics[a]

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			

[a] 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.

[b] Unique patient codes (deidentified) were available for 89.3% of the development set (n = 114 398 images).

[c] Individual-level data including age and sex were available for 66.1% of the development set (n = 84 734 images).

[d] Image quality was assessed for a subset of the development set.

[e] 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,[14] 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.

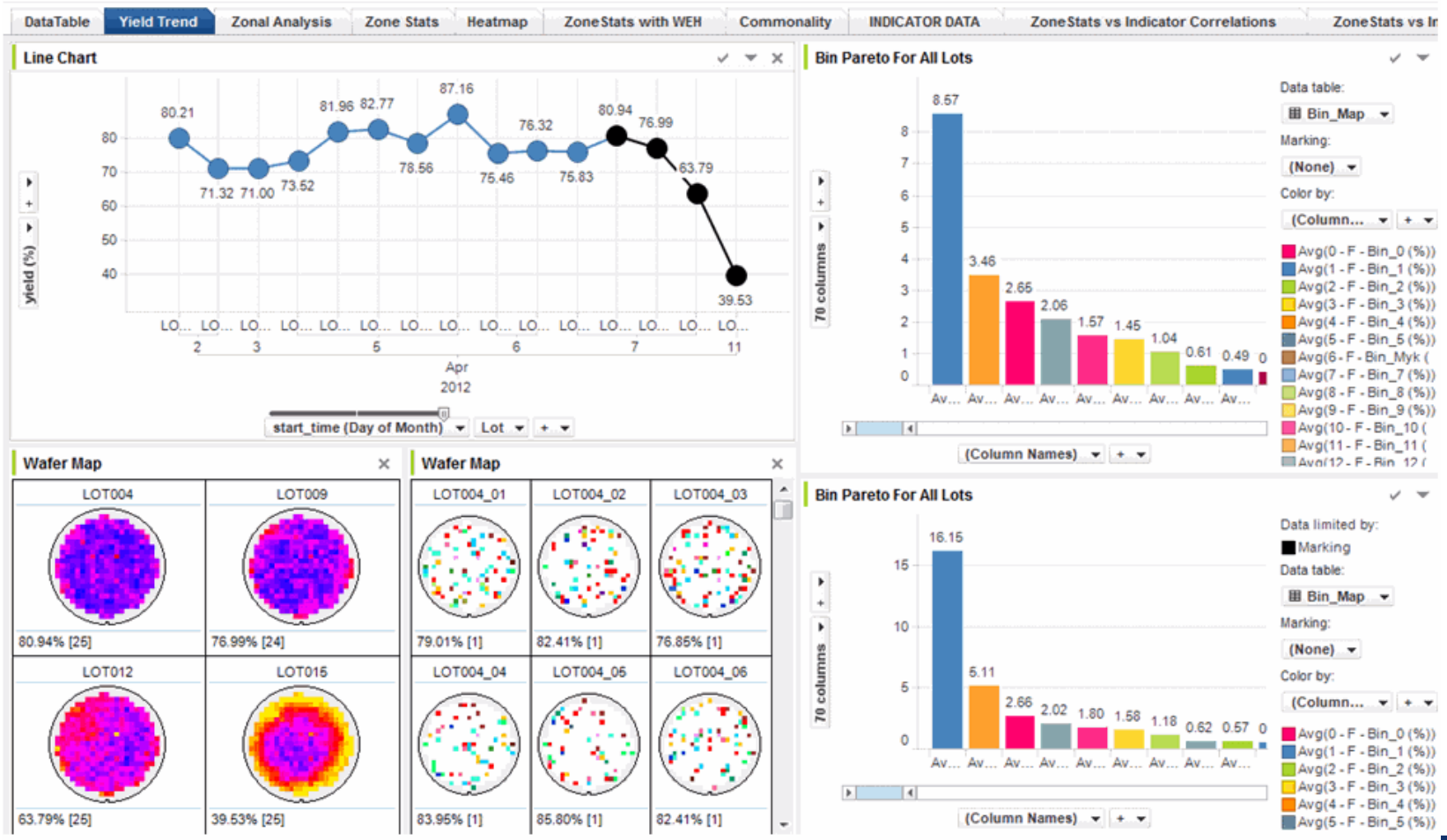
데이터 기반 문제 해결 절차

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

데이터 기반 문제 해결 절차

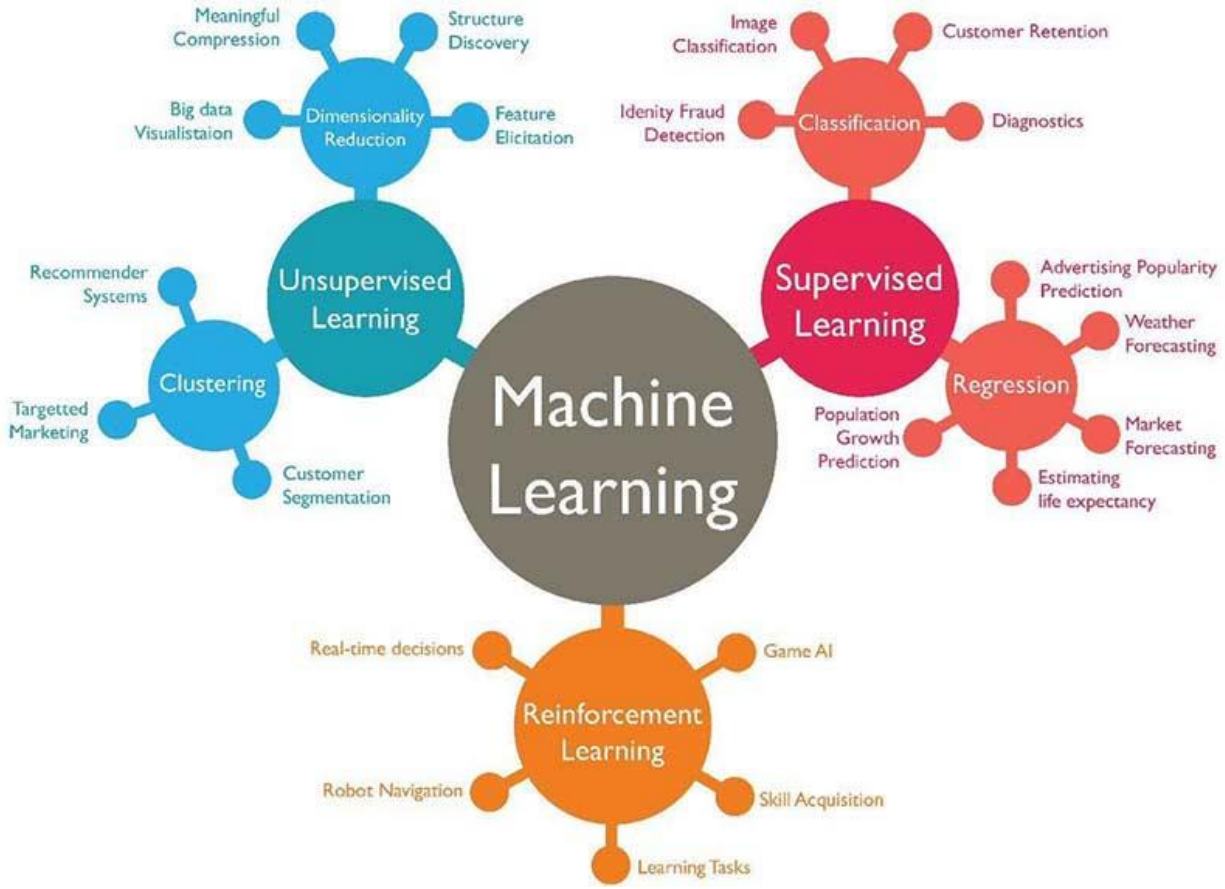
- 3단계: 성급한 모델링 이전에 충분히 데이터를 탐색하라

  ✓데이터 시각화 툴 사용 권장



데이터 기반 문제 해결 절차

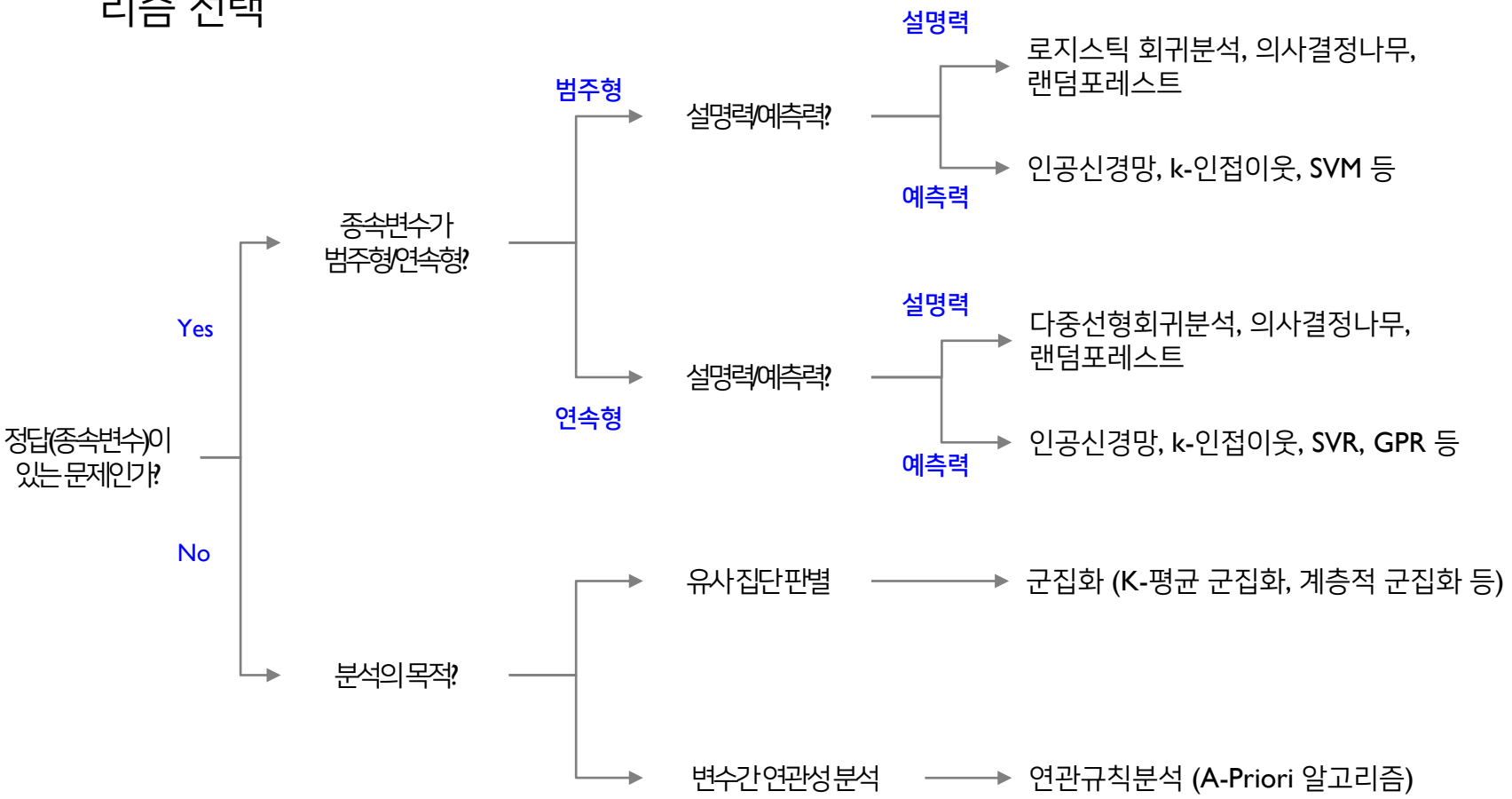
- 4단계: 모델 구축
  - ✓ 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

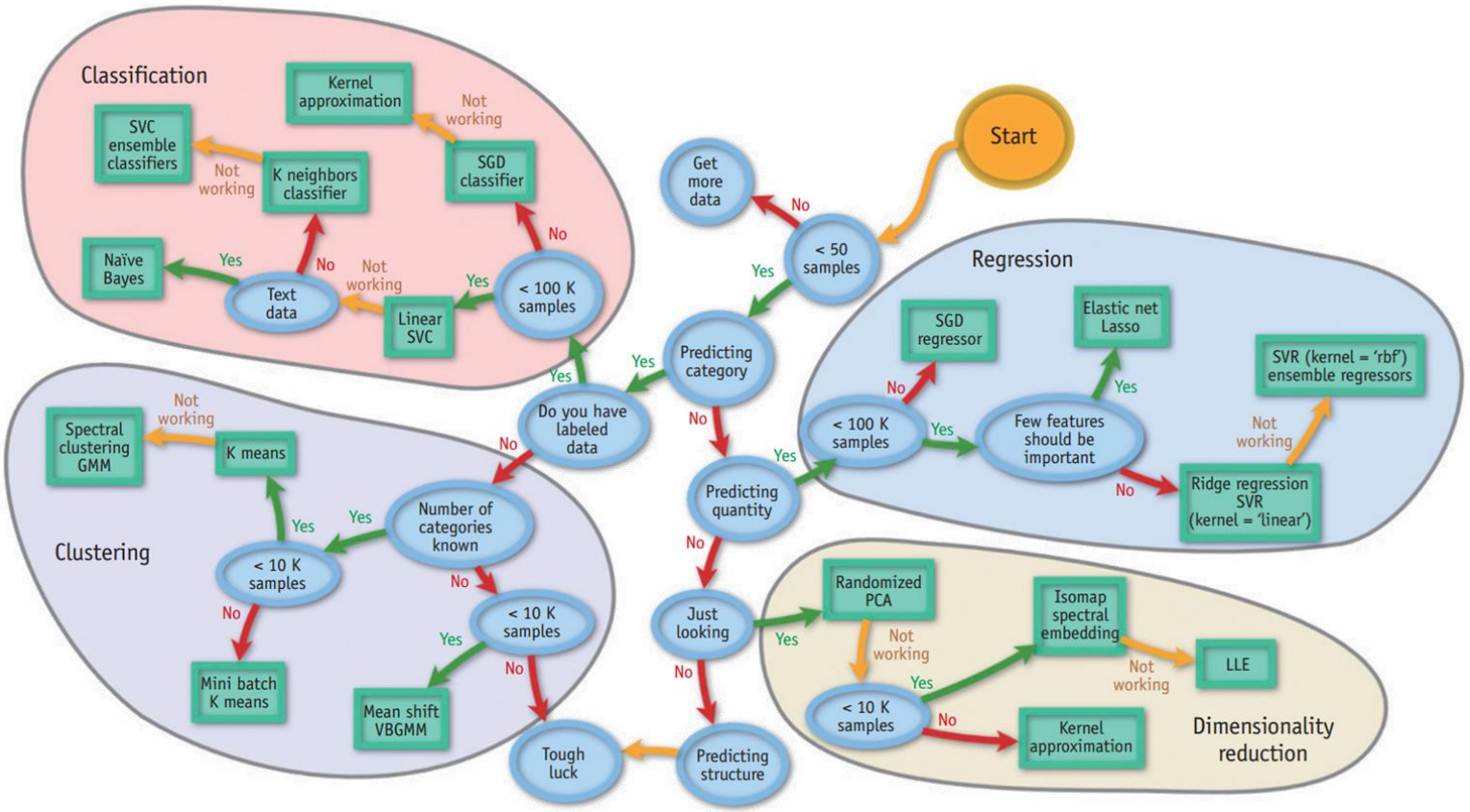
데이터 기반 문제 해결 절차

- **4단계: 모델 구축**
  - ✓ 질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



데이터 기반 문제 해결 절차

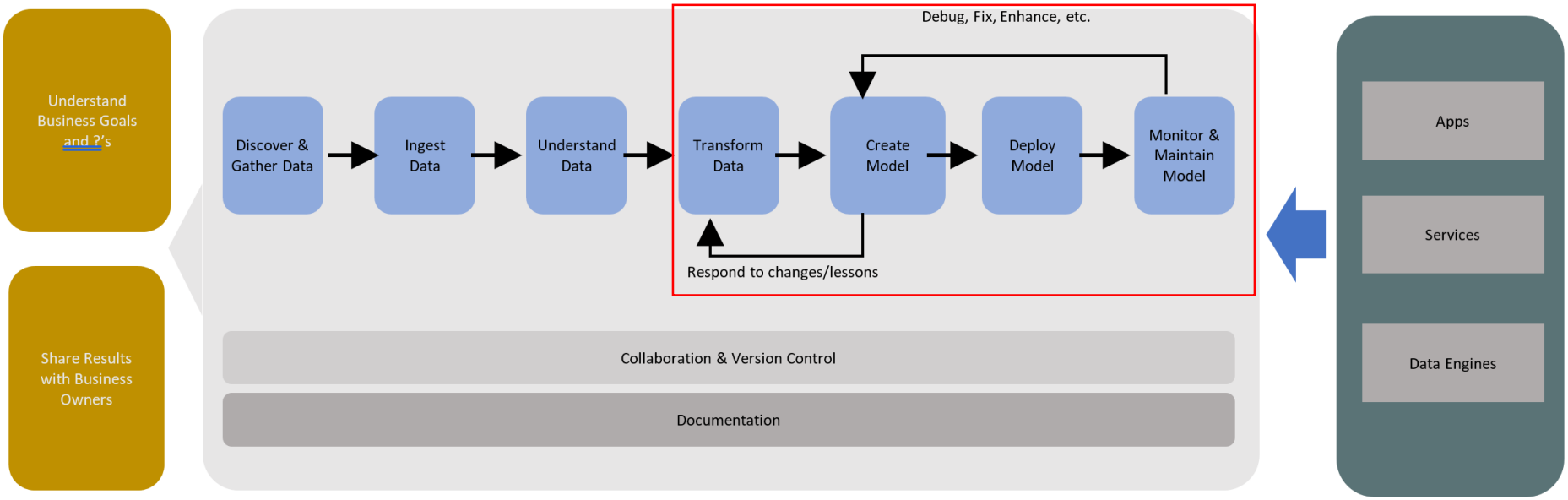
- 4단계: 모델 구축
  - ✓질문의 속성, 데이터의 특징, 결과의 설명력 포함 유무 등을 고려하여 적합한 분석 알고리즘 선택



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

	목적 및 문제 정의	데이터 수집/검증/수정	데이터 전처리	모형 구축	평가 및 해석
주요 활동	• 데이터분석을통해 달성하고자하는목표 구체화	• 데이터원천확인 • 독립변수/종속변수정의 • 변수별이상치/결측치 탐지및제거	• 불필요한변수삭제 • 변수변환 • 비지도방식의변수선택 및추출 • 데이터분할	• 모델학습 • 최적파라미터선택	• 모델링결과평가 • 개선안수립
주 사용 기법			• 기초통계분석을포함한 EDA • 주성분분석	• 분류알고리즘 • 회귀알고리즘 • 군집화알고리즘 • 이상치탐지알고리즘	
산출물	• 문제기술서 • 모형의유형(분류/회귀등)	• 행렬형태의모델링기초 데이터(행:레코드, 열: 변수)	• 정제된모델링용데이터	• 구축된모형 • 성능평가결과	• 모델결과평가표 • 개선아이디어리스트
고려 사항	• 현재보유데이터로달성 가능한목적인가?	• 최대한많은레코드와 변수를이단계에서수집	• 사용모형에따른데이터 분할비율 • 문제에따른적절한변수 수	• 다양한알고리즘시도 • 최적파라미터선택시 충분한영역탐색	• 모델의결과가현장에서 수용가능한수준인가?


ANY
questions?