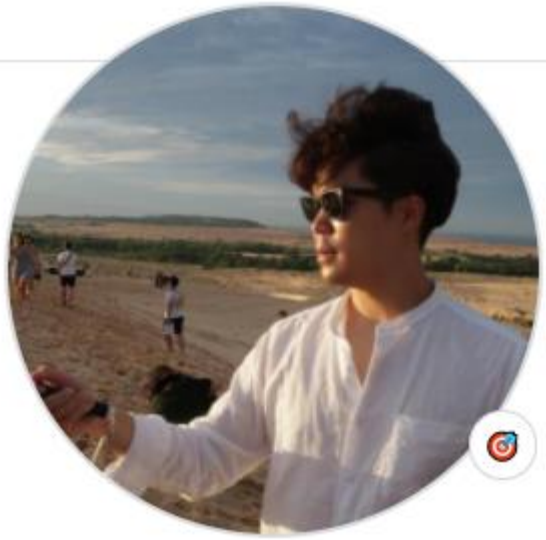
강사소개



GonieAhn

GonieAhn

G I G O 👊

Edit profile

ShinhanCard
Seoul, Korea
gonie32@gmail.com
https://github.com/GonieAhn

Highlights

Developer Program Member

Born To Be Data Scientist 🌱 ***- Garbage In Garbage Out*** 👊

DMQA | gonie32@gmail.com | goniiieee | hits 22 / 215

RESEARCH INTERESTS

- Explainable artificial intelligence for real world problem solving
- Critical feature selection using deep learning for high-dimensional data
- Tree-based interpretable machine learning for structured big data
- Data collection, Loading, Preprocessing, Feature Extraction, Modeling, Validation, Meaning Extraction & Pattern Recognition, Reporting, Decision, Action

SKILLS

Python | Scikit learn | Tensorflow | Keras | PyTorch | R | AWS S3 | Google Cloud Platform | MySQL | Splunk

EMPLOYMENT HISTROY

- **ShinhanCard - 21.10 ~ present**
  - BigData Planning Team, BigData R&D Center, LI
- **LG Electronics - 19.08 ~ 21.10**
  - LG Professional Instructor, Data Scientist
  - BigDataScience TP, AI BigData Team, CDO - 21.07 ~ 21.10
  - Manufacturing Intelligence Task, DXT Center, CTO - 19.08 ~ 21.07

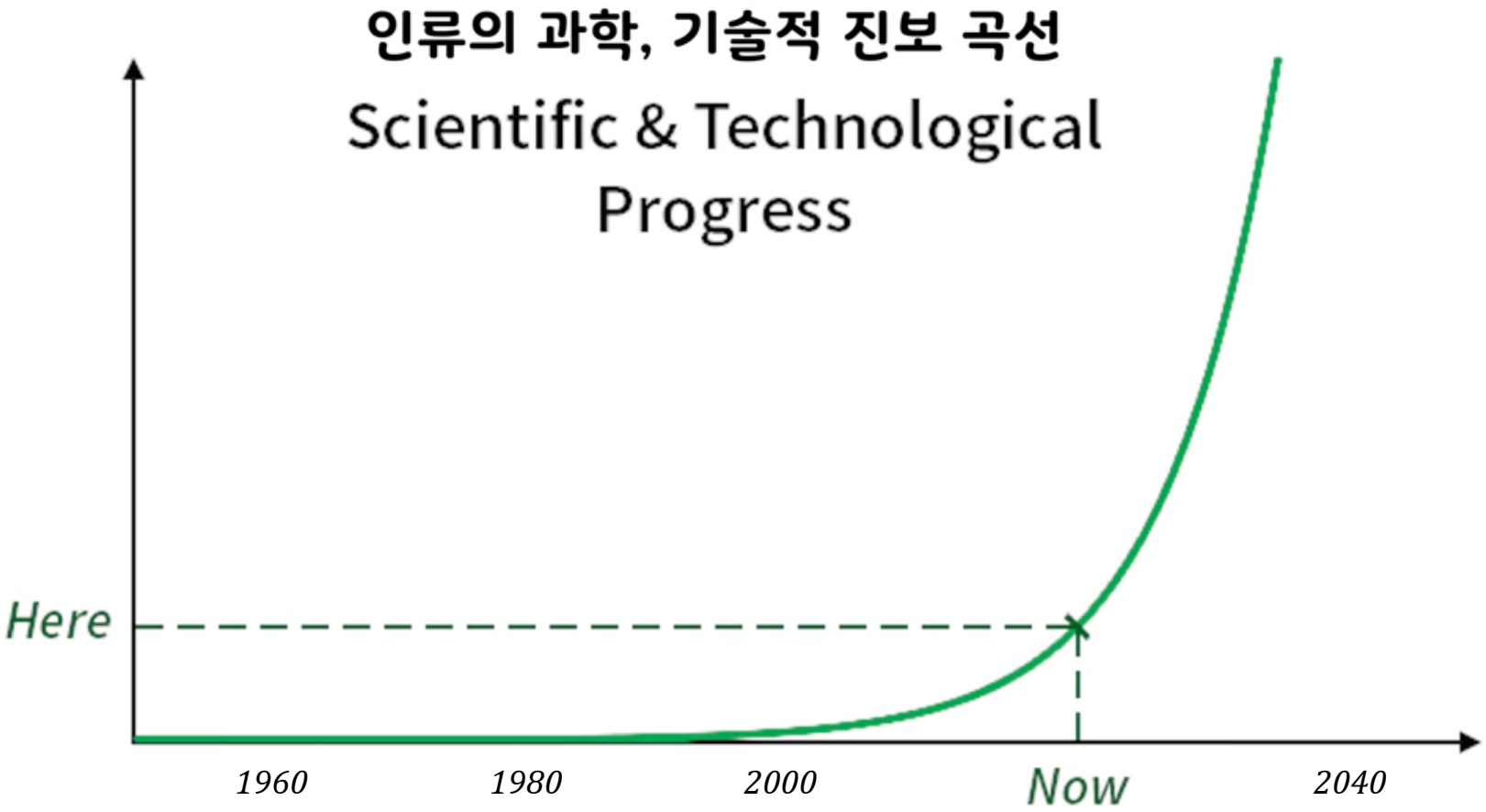
Introduction to Data Analytics

Data Scientist
안건이

목차

AI 시대 흐름

- 갈수록 빨라지는 기술적 진보
  - 15,000년 전의 인류는 농업을 시작함
  - 1,500년 전의 인류는 산업용 기구를 만들기 시작함
  - 150년 전의 인류는 증기기관과 가정용 전기를 만들기 시작함
  - 앞으로 150년 후에는 어떤 세상이 펼쳐 질까?
    - AI기반으로 기존에 발전해온 속도보다 폭발적으로 성장하게 될 것



AI 시대 흐름

- Software 기술적 진보
  - 2015년을 기점으로 Hardware의 뒷받침으로 Software 2.0 시대에 돌입하게 됨
  - Software 1.0시대에서는 Deep Learning이 주목 받지 못했음
    - LeNet-5(1998) – Yann LeCun (OCR)
  - Software 2.0 시대 부터 Deep Learning을 바탕으로 엄청난 양의 데이터를 처리할 수 있게 됨

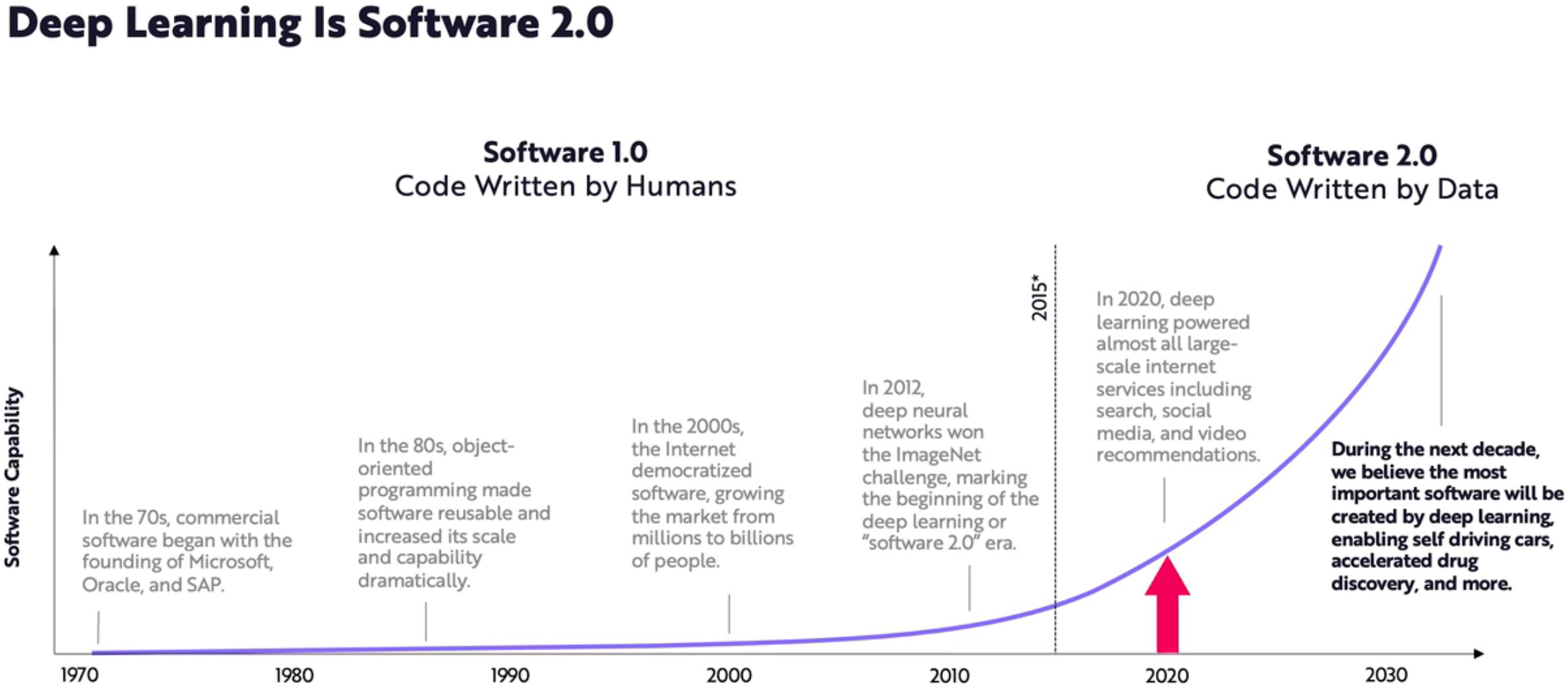
LeNet5 Implementation FROM SCRATCH

This is an implementation of LeNet5 from Yann LeCun's paper in 1998, using Numpy & OOP only (without any auto-differentiate tools or deep learning frameworks).

Yann LeCun's demo in 1993:

AI 시대 흐름

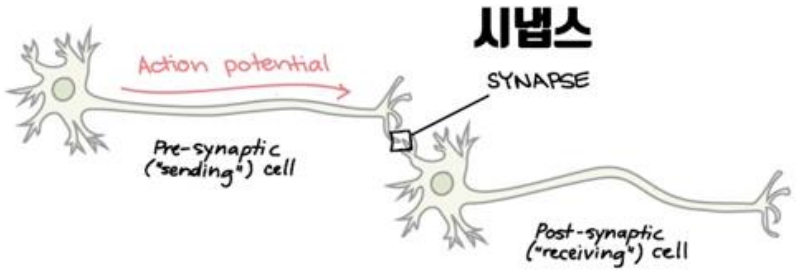
- Software 기술적 진보
  - 2015년을 기점으로 Hardware의 뒷받침으로 Software 2.0 시대에 돌입하게 됨
  - Software 1.0시대에서는 Deep Learning이 주목 받지 못했음
    - LeNet-5(1998) – Yann LeCun (OCR)
  - Software 2.0 시대 부터 Deep Learning을 바탕으로 엄청난 양의 데이터를 처리할 수 있게 됨



AI 시대 흐름

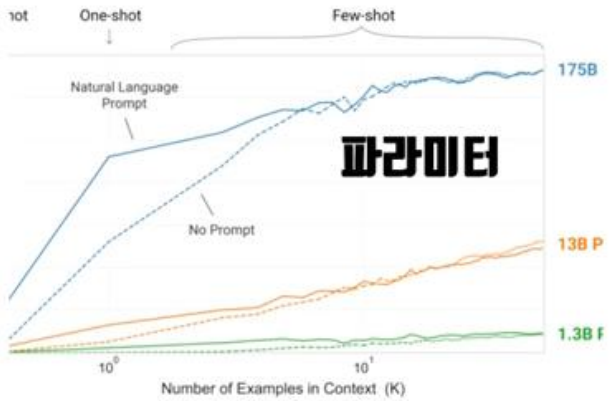
- GPT-4 → For Text Mining, Voice ETC
  - 인간의 스냅스와 GPT-4의 파라미터 수가 매우 가까워짐
  - 파라미터가 인간의 시냅스와 같은 작용을 한다고 말하기 어렵지만 어느 정도 작용을 함
  - 인간보다 더 정확한 AI 스피커, 완벽한 자율 주행으로 진보하고 있음

인간



100 조

GPT-4



1750억*1000 >>> 175조

175조, 인간의 시냅스 수와 비슷해져

데이터 분석이란 무엇인가?

Data

보통 연구나 조사 등의 바탕이 되는 재료

Mining

채굴: 광산에서 광석을 캐내는 것을 의미함

데이터 광산에서 **의미 있는 패턴**을 채굴함

Wal-Mart Legendary Story

BEER AND NAPPIES

맥주 기저귀

Posted on November 25, 2014 by xiaojuntian

THERE ARE MANY BRILLIANT STORIES WHEN BIG DATA COMES. THE MOST FAMOUS one would be the stories of "Beer and Nappies"

Data Mining

"Some of the ways Wal-Mart managers found to exploit their findings are legendary. One such legend is the story, "diapers and beer". Wal-Mart discovered through data mining that the sales of diapers and beer were correlated on Friday nights. It determined that the correlation was based on working men who had been asked to pick up diapers on their way home from work. On Fridays the men figured they deserved a six-pack of beer for their trouble; hence the connection between beer and diapers. By moving these two items closer together, Wal-Mart reportedly saw the sales of both items increase geometrically."

Decision & Explanation

A version with a slightly different view of the roles involved suggests that the men are sent to the supermarket for the diapers and, because there's no time left to go to a bar, take beer home with them.

Data Mining을 통해

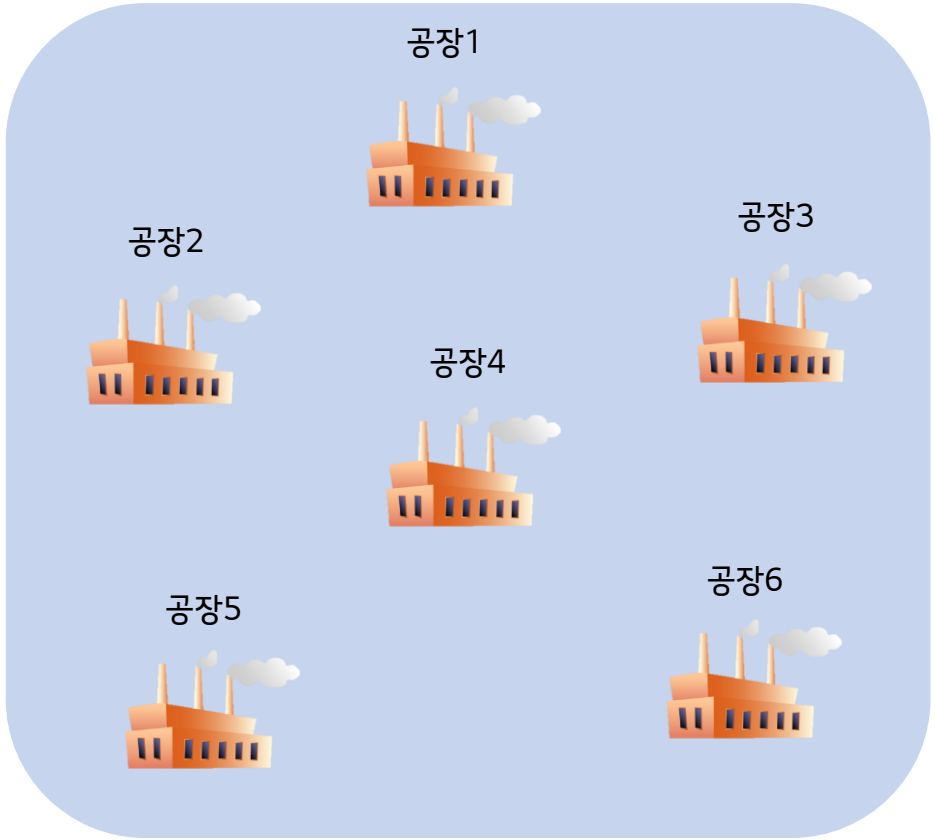
새로운 Insight를 도출

이렇게 간단하고 쉬운데 왜 안되는 것 일까?

기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음

현업관리자



데이터분석가

기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음
  - 대부분의 Data Scientist는 Reporting이 마지막 임무라고 생각함

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → **Action → Feedback → Action → Feedback ...**

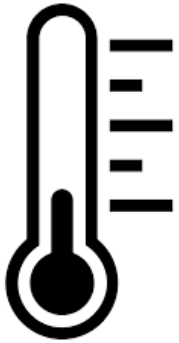
기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음

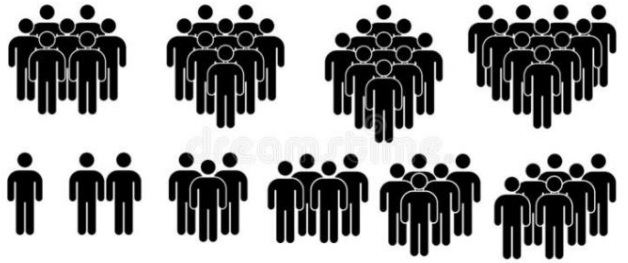
해파리 인명 피해 원인분석

해파리 피해 원인 인자

아이스크림 판매량



기온



해변에 몰리는 인파

기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

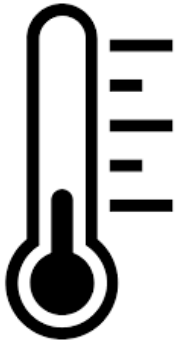
- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음

해파리 인명 피해 원인분석

해파리 피해 원인 인자

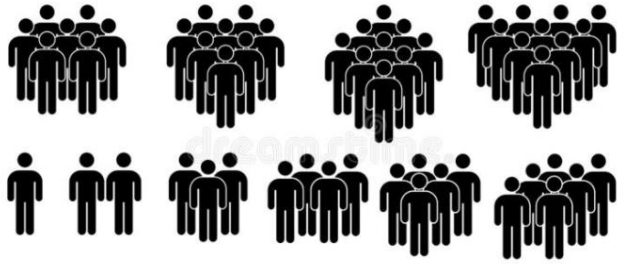
아이스크림 판매량

상관관계는 크지만
인과관계가 아님



기온

상관관계는 크지만
제어 불가능함



해변에 몰리는 인파

상관관계는 크고 제
어가 가능함

기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음

해파리 피해 원인분석

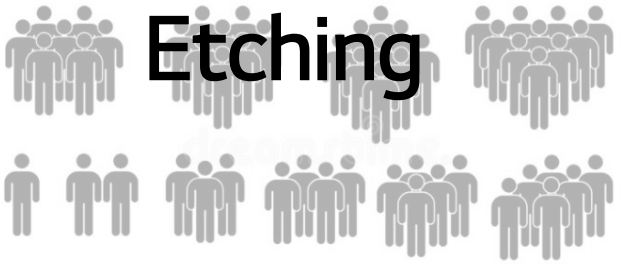
해파리 피해 원인 인자

아이스크림 판매량

상관관계는 크지만
인과관계가 아님

기온

상관관계는 크지만
제어 불가능함



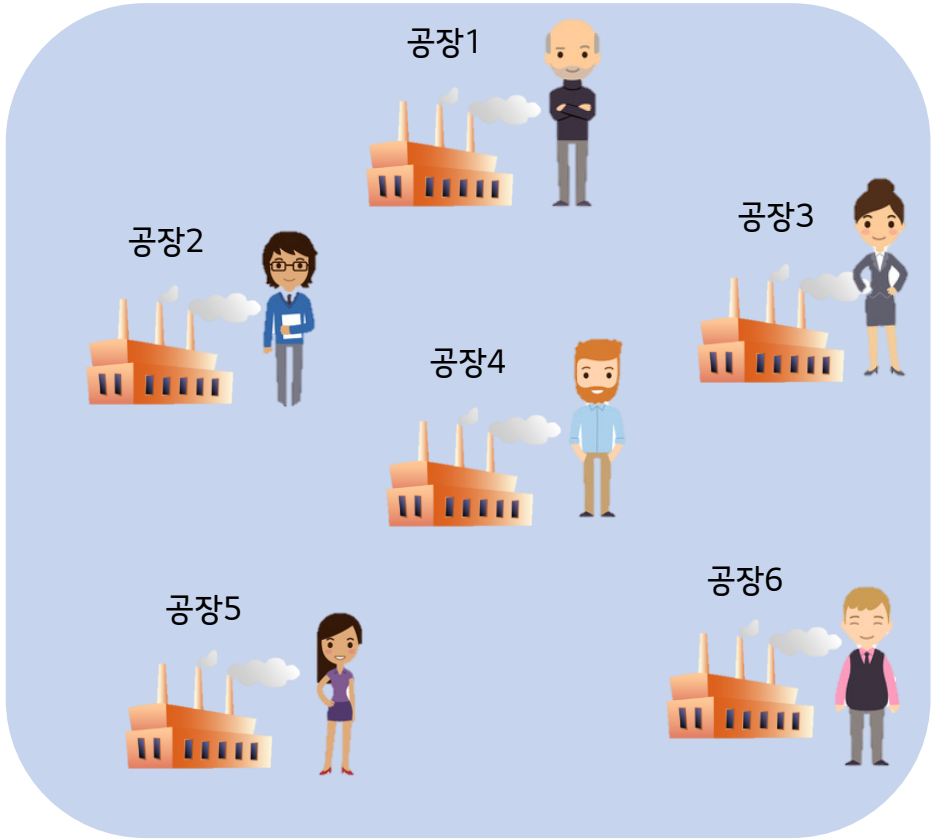
해변에 몰리는 인파

상관관계는 크고 제
어가 가능함

기업에서 데이터 분석이 실패하는 이유 – 현업관리자 vs 데이터분석가

- 비즈니스를 모르는 데이터분석가
  - 상관관계와 인과관계를 이해하지 못하는 분석가
  - 시시각각 바뀌는 현장을 꿰뚫어 보아야 Action을 할 수 있음
- **"데이터분석가는 현업전문가가 되어야 한다" or "현업전문가 중 데이터분석가를 육성해야 한다"**
  - **현업전문가를 대상으로 데이터 분석 교육 진행 → 놀라운 결과가 도출 됨**

현업관리자



데이터분석가

기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- 2015년 이전
  - 있는 데이터 모두 수집해 !!
- 2015년 이후
  - 우리가 필요한 데이터만 수집해 !! → 잘 수집해 !!
  - 비즈니스를 이해하는 분석가 + 전략가

Garbage In Garbage Out

데이터 분석의 가장 중요한 Point

양질의 데이터(재료) 만들기
데이터에 대한 이해
데이터 전 처리

기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- AI system Error
  - 과거에는 code 혹은 Algorithm을 개선하려는 노력을 많이 했음
  - 현재는 Data를 개선하는 쪽으로 사고방식을 바꿔야함

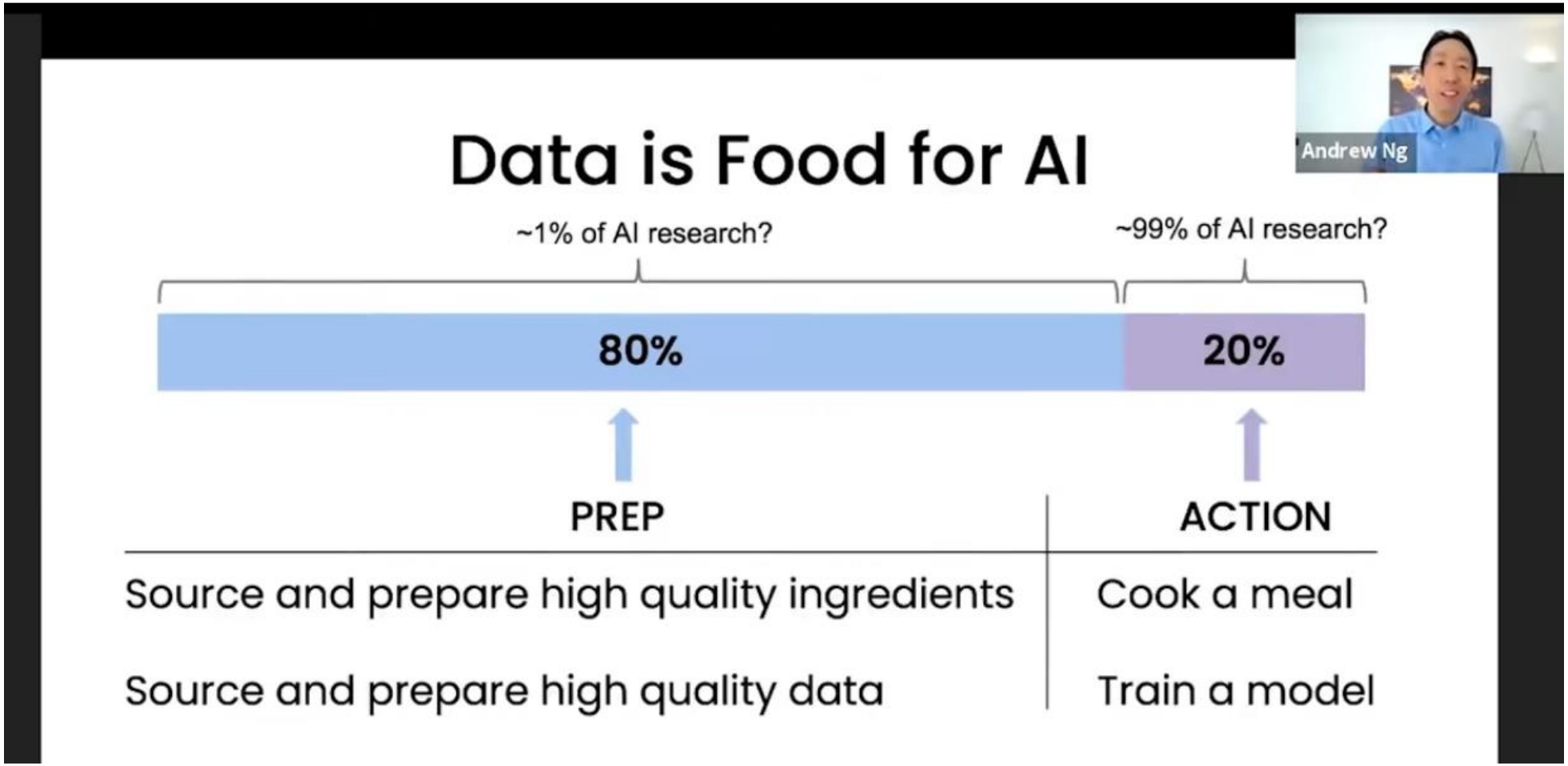
기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- AI system Error
  - 과거에는 code 혹은 Algorithm을 개선하려는 노력을 많이 했음
  - 현재는 Data를 개선하는 쪽으로 사고방식을 바꿔야함
    - 실제 대다수의 프로젝트에서 데이터 개선에 집중하는 것이 효과적이었음을 증명하고 있음

Source. Deeplearning.Ai	강철 결함감지	태양광 패널	표면검사
기본 baseline	76.2%	75.68%	85.05%
코드 – 개선 Model-Centric	+0%	+0.04%	+0.00%
데이터 – 개선 Data-Centric	+16.9%	+3.06%	+0.4%

기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- Andrew Ng
  - 실제 AI 프로젝트의 성패는 데이터 정리에서 80퍼센트가 결정 된다고 말함
  - 20%는 Google의 BERT, Open AI의 GPT-3와 같은 선진적인 모델임
  - 보통 우리는 모델이 99%를 좌우 한다고 생각함
  - 따라서, AI 프로젝트에서 성공하려면 데이터 준비를 잘해야 함



기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- Andrew Ng
  - 실제 AI 프로젝트의 성패는 데이터 정리에서 80퍼센트가 결정 된다고 말함
  - 20%는 Google의 BERT, Open AI의 GPT-3와 같은 선진적인 모델임
  - 보통 우리는 모델이 99%를 좌우 한다고 생각함
  - 따라서, AI 프로젝트에서 성공하려면 데이터 준비를 잘해야 함

머리가 100배 더 뛰어난 사람이 저품질 정보로 학습하는 것보다 머리가 나쁘더라도 고품질데이터로 학습하는 것이 더 좋다

기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- Andrew Ng 교수가 제시한 ML 데이터 개선책 6가지

1. 고품질의 데이터

2. 레이블링(Y, 라벨링)의 일관성

3. 모델 최신 여부를 따지기 보다 데이터 품질을 개선 시켜야함
   데이터 품질을 개선 시키기 위해서는 Domain 적인 지식이 필수적임

4. 오류 발생시 코드를 개선하기보다 데이터 품질을 개선하는 것이 급선무

5. 노이즈가 많은 소규모 데이터셋은 집중 관리해야함

6. 데이터 품질을 높이기 위한 도구와 서비스를 갖춰야함
   품질 불량이 나면 Line을 Stop하듯 데이터 품질에 불량이 나면 거기에 상응하는 대응이 꼭 필요함

기업에서 데이터 분석이 실패하는 이유 – 크기만 Big Data

- Machine Learning System의 부채는 대부분 코드 레벨 보다 시스템 레벨에 존재함
- 사실 ML 코드는 큰 System 중 정말 일부분에 지나지 않음
- 대부분의 회사는 이러한 기술의 부채를 정확히 파악하지 못하고 있음 → 실패하는 이유
- 따라서, 어디에 힘을 실어야 하는지 판단하지 못함

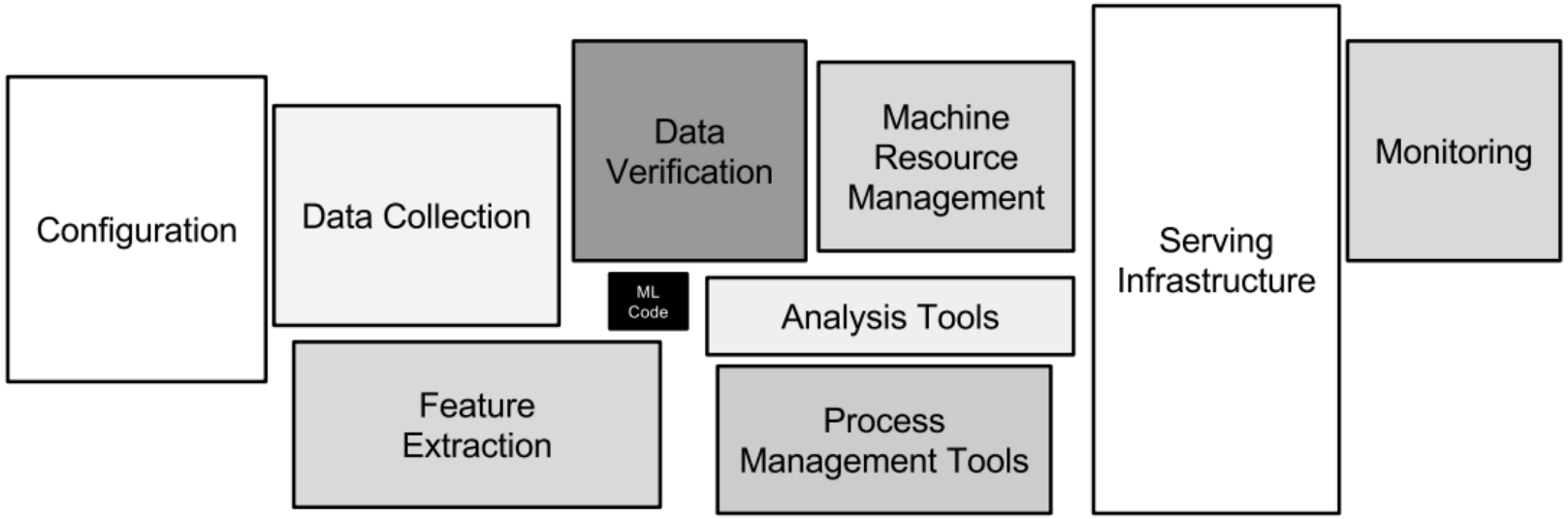


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

기업에서 데이터 분석이 실패하는 이유 – Key Thinking

현업관리자 vs 데이터분석가

크기만 Big Data

All-in-One(Full-Stack) Data Scientist

Computer Programming Skill +
MATH & STATISTICS +
DOMAIN / BUSINESS KNOWLEDGE +
DATA ENGINEERING

기업에서 데이터 분석이 실패하는 이유 – 경험의 자산화

- 실패한 프로젝트의 자산화
  - 과거의 실패를 반복적으로 발생시킴
    - 쓸모 없는 데이터를 수집하는데 투자함 → 결국 Garbage 분석 결과를 불러옴
    - 투자를 했기 때문에 사람을 녹여서 무엇이라도 해보려고 함 → 직원 이탈
  - 과거의 실패를 기반으로 문제 해결에 성공해 제대로 된 예측 모형을 만들어 낼 수 있다면 대규모 투자 집행

- Fast Fail 전략이 필수
  - 안되는 것은 안되는 것임
  - “하면 된다” → 1980 ~ 2000년 대 마인드

- 한방을 노리는 윗 분들
  - 기획 → 비즈니스 수립 → 인프라 준비 → 데이터 수집 → 데이터 분석 → Action → 성과
  - 작은 성공을 발판 삼아 점진적으로 확장
  - 테슬라, 구글, 아마존 모두 아주 오랫동안 점진적으로 데이터 분석을 발전시켜 왔음
  - 한방은 없다

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback …

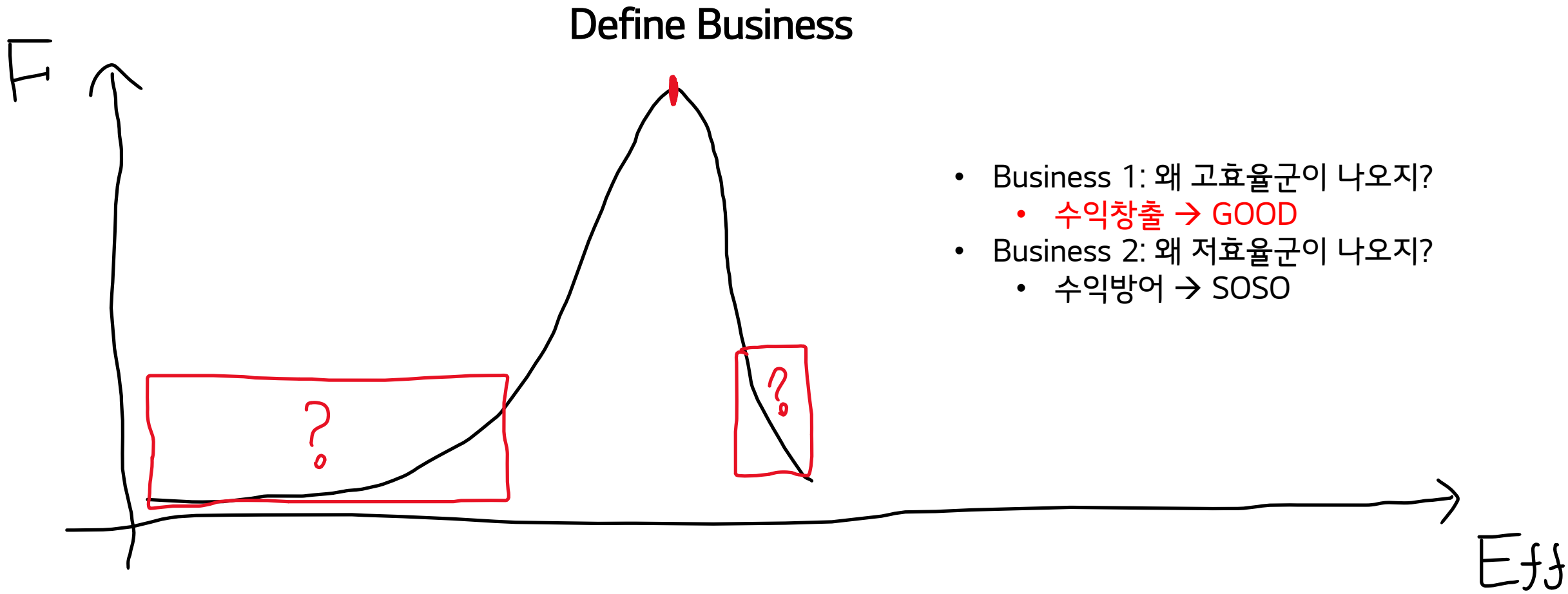
- Solar Project
  - 태양빛을 전기로 얼마만큼 전환 시킬 수 있을까 ?
    - “효율”이라고 정의함

- 첫 단추를 잘 끼우자
  - Define Business: 약 1개월 – 현업전문가 vs 분석전문가
    - 현업전문가
      - 요구사항 : 스마트 팩토리, 효율 극대화 등 매우 광범위한 Business 이야기를 함
      - 데이터 분석이 모든 걸 해결해줄 수 있다는 착각을 대부분 하고 있음 (특히 AI…)
    - 분석전문가
      - 냉정하고 실현 가능한 분석 Business 유도 → 도메인이 필수적으로 필요함
      - 초창기에는 공장에서 살아야 함 (Daily Meeting, Weekly Meeting 참석하며 Pain Point 발굴)

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback ...

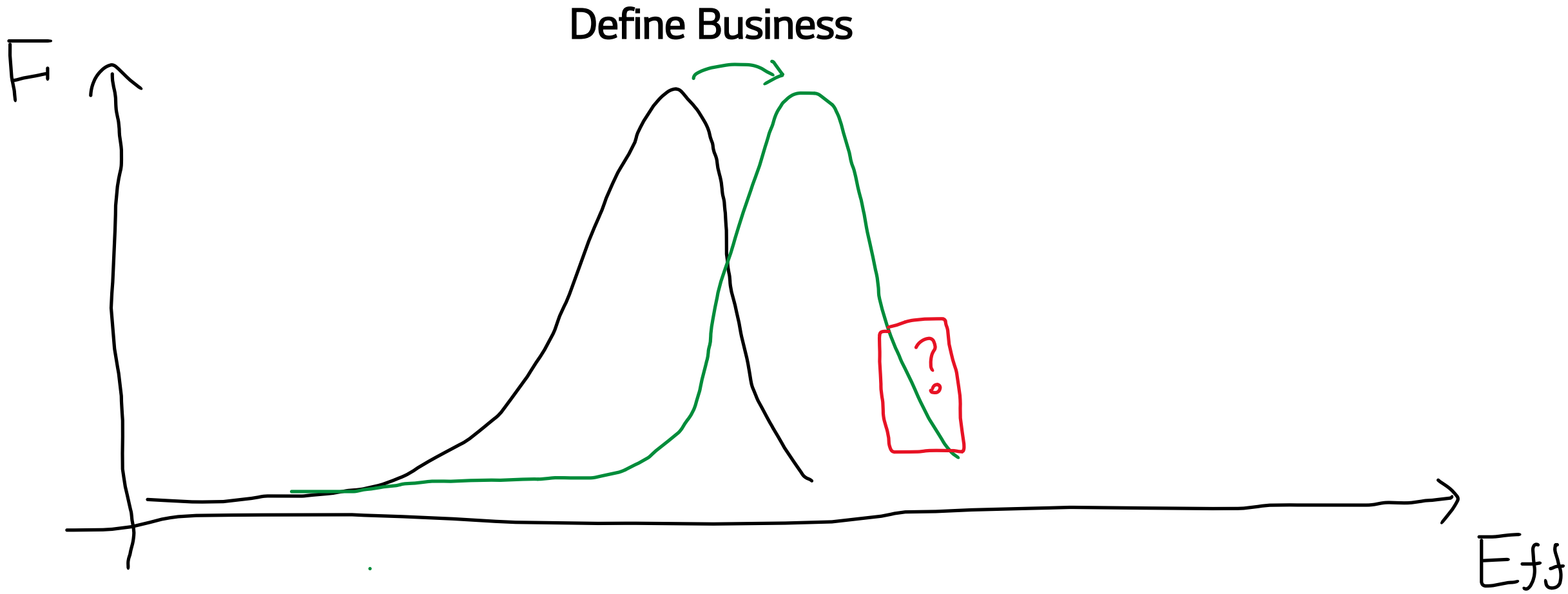


- Business 1: 왜 고효율군이 나오지?
  - 수익창출 → GOOD
- Business 2: 왜 저효율군이 나오지?
  - 수익방어 → SOSO

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

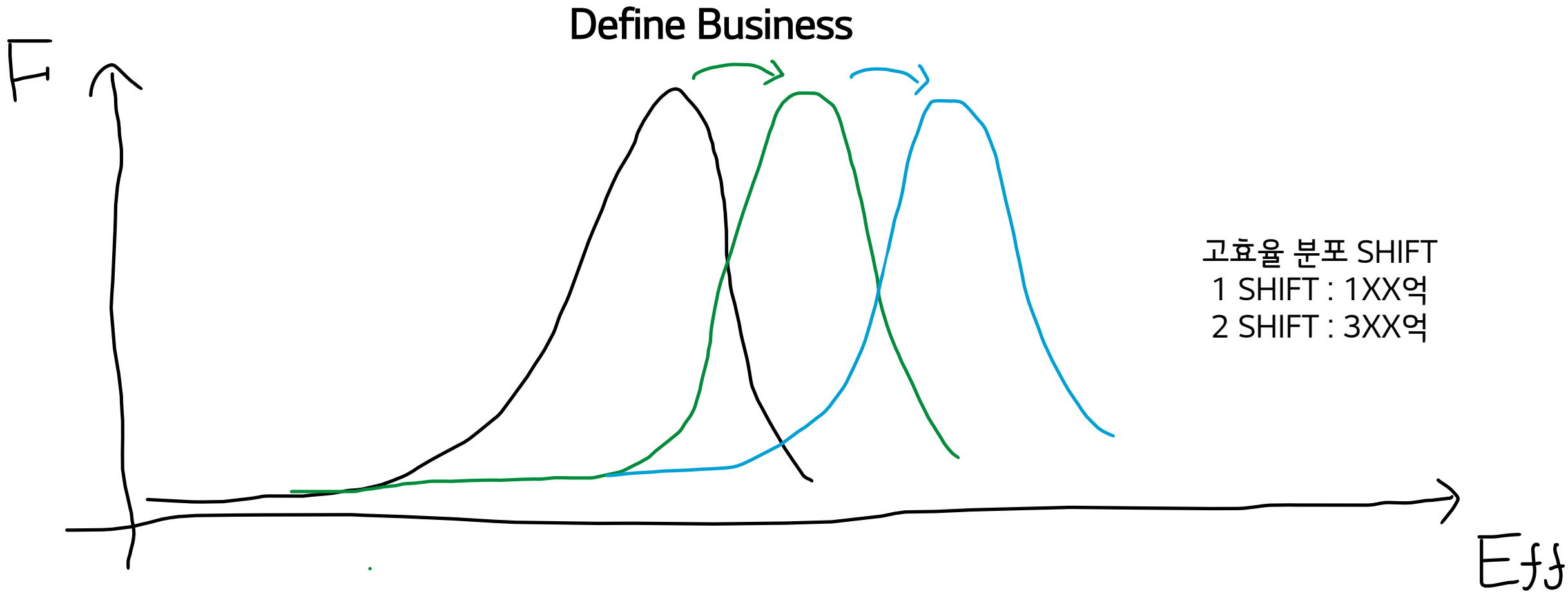
→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback ...



데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback ...



데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback ...

- Data Preprocessing – GIGO 약 3개월 & 분석과정 내내 진행
  - MES → Server → AWS(S3) → 분석환경
    - Data Moving Point가 많을 수록 Data 손실이 많이 발생함
    - Missing Value 구하기
  - Feature Extraction Based on Domain Knowledge
    - 고효율에 영향을 미치는 Hidden Factor를 찾고 싶어함
    - Data에 Hidden Factor가 없는데 어떻게 찾아요?
      - 현업전문가 vs 분석전문가 - 피 터지는 눈치 작전 시작

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback …

- Data Analytic – 약 2주
  - 상황1: 하루에 매우 큰 데이터가 쌓임 → 1주일 데이터만 모아도 약 XXXG
    - 방대한 데이터를 빠르게 처리할 Algorithm Research
      - Complexity 무한대 + 빠르게 처리 ➔ LightGBM 적용 결정
  - 상황2 : 특정 데이터(Locally, 고효율군)에 대한 해석 필요
    - 상위 5% 효율에 대한 원인분석
      - SHAP를 적용하여 상위 5%에 대한 중요인자 추출 진행함
  - 상황3 : 상위 5%에서도 다양한 원인이 존재할 것이라고 가정
    - 상위 5%는 하루 약 25000개 발생
      - 상위 5% SHAP Value를 HDBSCAN 군집화 알고리즘을 활용하여 군집화 진행
        - 어느 군집에도 속하지 않은 노이즈 고효율군은 제거
        - 뚜렷한 군집만 추출
      - 각 군집에 대한 고효율 원인 도출

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback …

- Reporting – 약 1주
  - 상위 5% 고효율군에 대한 원인 분석
    - 총 3개의 뚜렷한 군집 도출
      - 고효율 군집1
        - 원인 1 xxxxxxxxxxxxxxx
        - 원인 2 xxxxxxxxxxxxxxx
        - …
      - 고효율 군집2
        - 원인 1 xxxxxxxxxxxxxxx
        - 원인 2 xxxxxxxxxxxxxxx
        - …
      - 고효율 군집3
        - 원인 1 xxxxxxxxxxxxxxx
        - 원인 2 xxxxxxxxxxxxxxx
        - …

데이터 분석 성공 사례

Define Business → Data Collection → Data Cleaning → Feature Extraction → Algorithm Research

→ Modeling → Validation → Decision → Reporting → Action → Feedback → Action → Feedback …

- Action → Feedback – 약 2개월 & Define Business 부터 다시 설정하는 경우도 발생함
  - 군집 3개 중 1개만 실험 가능
    - 실험 불가능 이유
      - 이유1 xxxxxxxxxxxxxx
      - 이유2 xxxxxxxxxxxxxx
      - 이유3 xxxxxxxxxxxxxx
      - ….
    - 실험 결과
      - 효율상승 or 효율유지 or 효율하락

설비 투자 없이 이익 약 XXX억 증가

Q & A