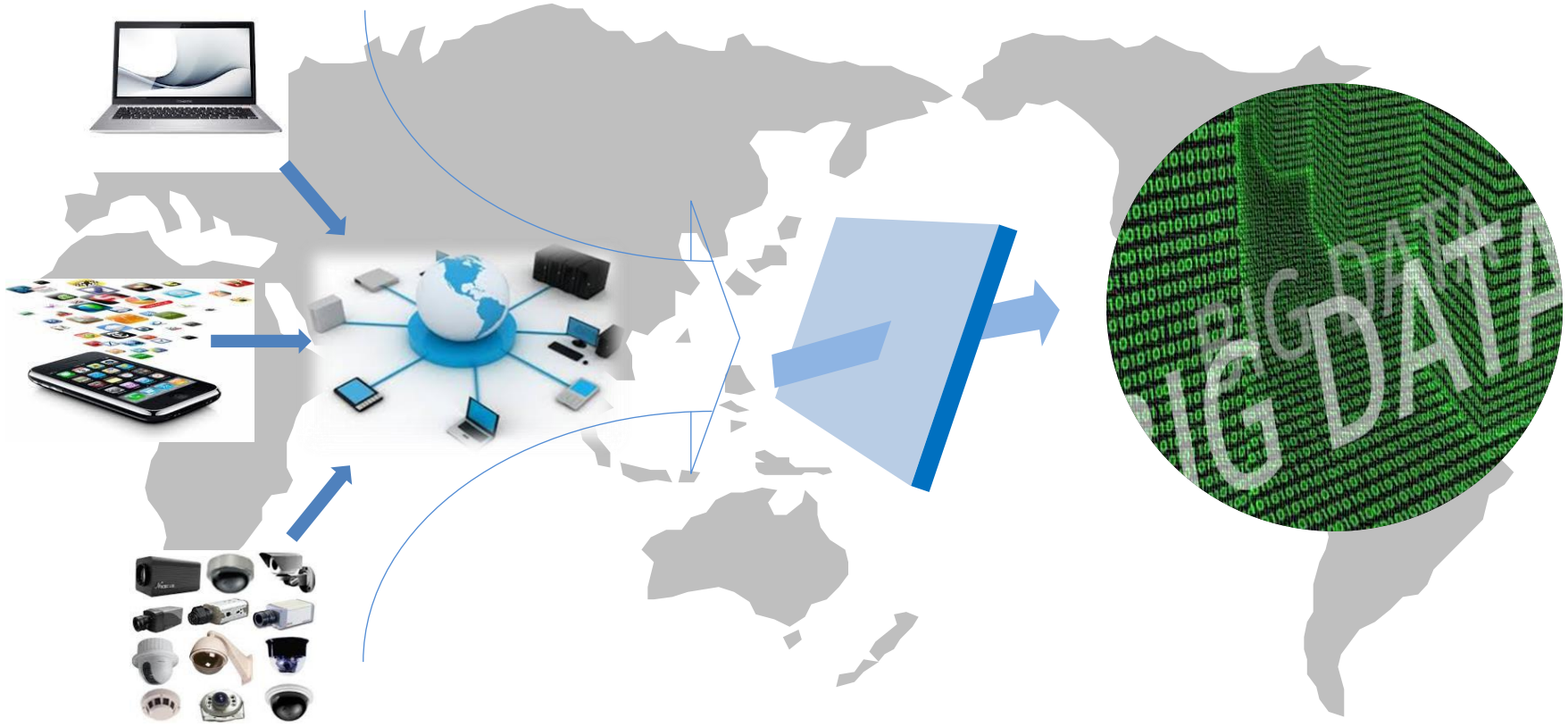
1. 데이터 문제해결

2023

1. 빅데이터와 분석 기획

- 빅데이터 개요: 빅데이터는 기존 DBMS 및 관리도구의 처리 능력을 넘어서는 대량의 정형 및 비정형 데이터를 의미, 3V 특성
- 빅데이터 분석에서 중요한 것은 크기와 종류가 아닌 인사이트의 발견을 통한 문제 해결
- 빅데이터 분석을 위해서는 새로운 관점의 빅데이터 분석과 활용의 기획이 가장 중요함



1. 빅데이터와 분석 기획

Big Data + 새로운 관점

새로운 관점의 빅데이터 분석!

각 분야의 특성을 고려한 기획

새로운 인사이트, 문제 해결과 목적 달성에 기여

1. 빅데이터와 분석 기획

- 데이터 분석 기획: 데이터 분석 과제의 정의 및 기대효과, 목적 달성을 위한 데이터, 분석방안, 관리방안 등을 분석 전에 기획

- 분석방법론과 데이터 분석 기획: 데이터 분석 기획은 실제 분석의 수행 전에 이뤄져야 하며, 분석과 활용에 대한 구체적인 계획이 수립되어야 함

- 분석 기획 시 고려사항: 가용한 데이터 확인, Use Case의 확인, 분석 역량, 기대 효과를 고려해야 함

1. 빅데이터와 분석 기획

빅데이터 분석 기획,

나무를 보지 말고 숲을 보기

데이터 분석 과제의 정의 및 기대효과, 목적 달성을 위한

데이터, 분석방안, 관리방안 등을 분석 전에 기획

1. 빅데이터와 분석 기획

빅데이터 분석 기획

분석 기회 발굴 (Question First!)

분석의 전제조건! 데이터, 필요기법 등으로 확장

분석 목적, 데이터, 처리 및 분석 절차 등의 빅데이터 분석 라이프사이클에 걸친 구체적인 방안 수립

의사결정과 목표 달성 실행 과정에 필요한 인사이트를 과학적인 분석으로 제공하는 체계

1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획

- KDD : Knowledge Discovery in Database

  선택-전처리-변환-데이터마이닝-해석/평가

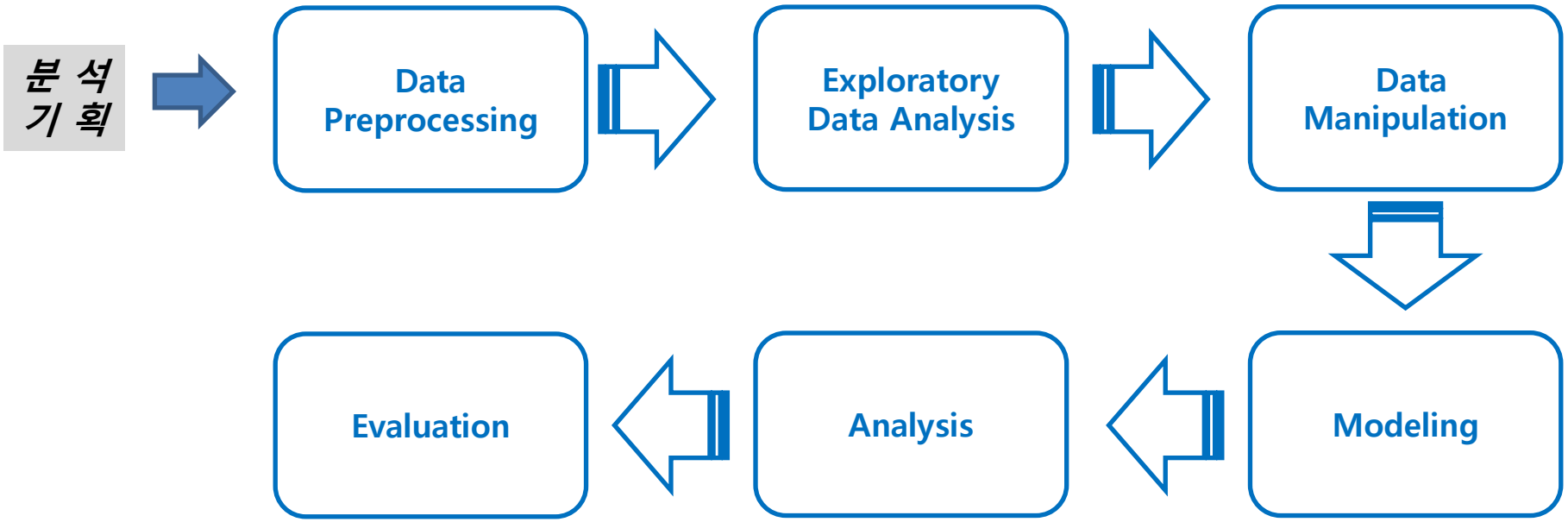
- CRISP-DM: Cross-Industry Standard Process for Data Mining

  비즈니스 이해-데이터 이해-데이터 준비-모델링-평가-전개

- SEMMA: Sampling Exploration Modification Modeling Assessment

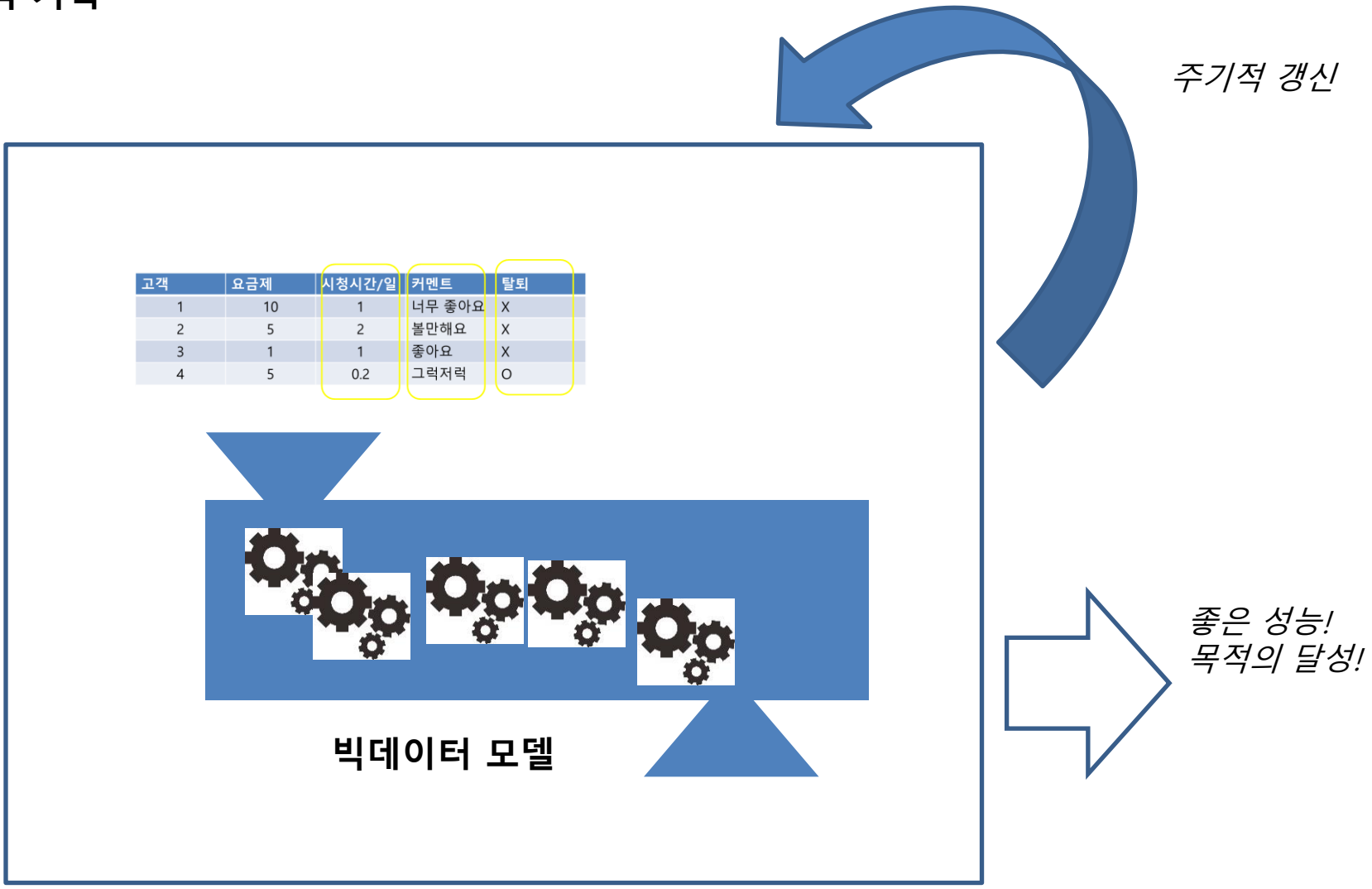
  Sample-Exploration-Modification-모델링-평가

1. 빅데이터와 분석 기획


분 석
기 획
Data Preprocessing
Exploratory Data Analysis
Data Manipulation
Modeling
Analysis
Evaluation

1. 빅데이터와 분석 기획

분석 기획



1. 빅데이터와 분석 기획

주의 사항?!

- **분석 과제에 가용한 데이터!**
- **기존 Use Case의 연구**
- **실질적인 분석 절차에 대한 계획 수립**
- **분석 역량의 고려**

2. 분석 기획 시 고려 사항

- 분석의 목적: 데이터 분석의 목적은 데이터로 부터 Inference를 하거나, 혹은 Prediction을 하는 것임.

- 분석 목적의 구체화: 분석 목적은 모호하거나 '분석'에만 초점을 맞춘 것이 아닌, 비즈니스 프로세스의 관점에서 성과 개선에 도움을 주어야 함

- 조직 정비: 데이터 분석과 관련된 다양한 영역의 인력이 유기적으로 협업해야 하며 Cross Functional Team을 구성

Inference VS Prediction

2. 분석 기획 시 고려 사항

분석을 위한 Teaming **Cross Functional Team!**

1. 도메인 경험 및 현장의 이슈
2. 데이터 엔지니어링 및 매니지먼트
3. 알고리즘에 대한 이해와 분석 역량
4. 시스템 및 아키텍쳐
5. 비즈니스 컨설팅

3. 분석 기획을 위한 데이터 이해

- 데이터: 값의 기록인 데이터에는, 수치형, 범주형, 텍스트 등의 값이 들어갈 수 있음
- 데이터의 형태 이해: 데이터를 이루는 값들은 다양한 형태로 구성될 수 있으며, 크게는 정형데이터, 반정형데이터, 비정형데이터 등.
- 고려사항: 데이터 가용 여부, 데이터 사용에 대한 허용과 관련 법 등에 대한 검토가 필요

3. 분석 기획을 위한 데이터 이해

데이터의 값

- 수치형: 1,2,3,4,5,.... 1.1,2.4,3.1,...
- 논리형: True or False
- 범주형: “합격” 또는 “불합격“ 등
- 텍스트: “오늘의 뉴스는...“

정형 / 반정형 / 비정형 데이터

Structured / Semi Structured / Unstructured

3. 분석 기획을 위한 데이터 이해

확인 사항!

“ 이 데이터 써도 되나?”

동의

GDPR(유럽연합 일반 데이터 보호 규칙)

이용 허가

개망신법

비식별화

3. 분석 기획을 위한 데이터 이해

정형화 시 고려사항!

1. 같은 분석 대상은 같은 줄(행)에 표현하기
2. 같은 종류의 값들은 같은 열에 표시하기, 열의 이름은 변수라고 부르기
3. 변수 명칭은 일관성있게 만들기
4. 범주는 그대로 표시하되 분석 시에는 숫자로 변환하여 처리하기(One hot encoding)
5. 텍스트는 나누고 정리하여 컬럼처럼 사용하기

3. 분석 기획을 위한 데이터 이해

- 데이터 큐레이션: 데이터의 가치를 제고해주는 데이터 관련 활동

- 데이터 활용: 주어진 데이터로 모델링 하고 비즈니스에 활용할 수 있는 시나리오를 통해 보다 구체성있는 분석을 기획함

- 하향식 VS 상향식: 분석 과제에 맞는 데이터를 찾아 분석해나가거나, 혹은 데이터로 부터 이슈를 찾는 방식으로, 프로토타이핑을 통해 갭을 줄여나감

3. 분석 기획을 위한 데이터 이해

데이터 큐레이션

- 데이터를 수집하고 처리하여 정제하며, 분석 알고리즘의 적용을 위한 활용, 그리고 모형의 성능을 평가하기 위한 활용 등 데이터의 가치를 제고해주는 데이터 관련 활동

- 비즈니스와 데이터, 알고리즘과 시스템을 연결

3. 분석 기획을 위한 데이터 이해

데이터 큐레이션의 예

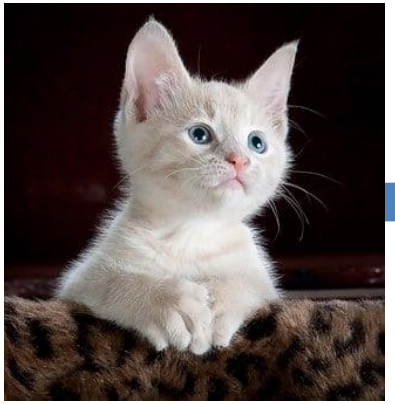
- 분석 목적에 사용할 내부 데이터를 위한 RDBMS 접근
- 외부 데이터를 위한 API와 웹 수집
- 수집된 데이터를 정형화
- .....

3. 분석 기획을 위한 데이터 이해

데이터 큐레이션의 또 다른 예, “Data Annotation”

다량의 이미지를 바탕으로 사물인식 모델링을 위해 각 이미지에 라벨링을 해주기

“개”



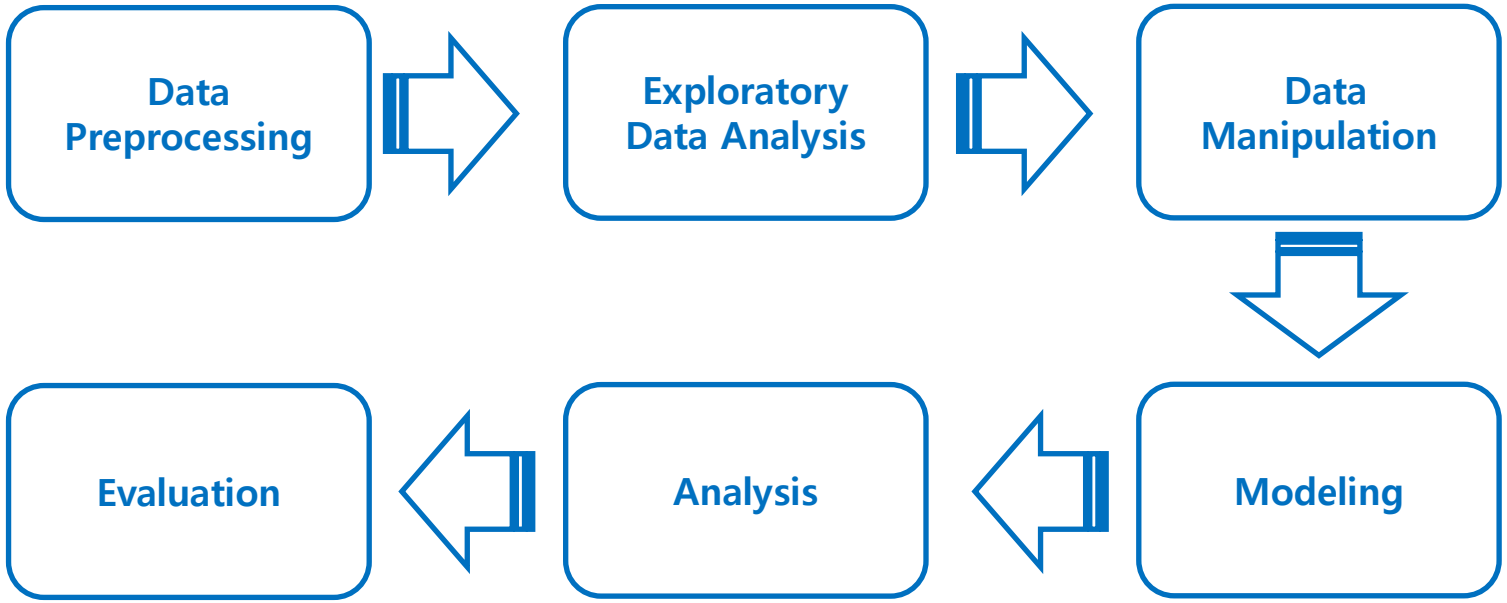
“고양이”

4. 분석 절차 1

- Data Preprocessing: 분석에 필요한 데이터를 핸들링이 가능하도록 처리하는 과정을 의미

- Exploratory Data Analysis: 데이터를 요약하거나 시각화하여 분석에 필요한 인사이트를 발견
- Data Manipulation: 데이터에서 필요한 변수를 선정하거나 변수를 가공하여 분석에 활용할 수 있도록 함

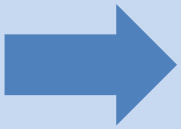
4. 분석 절차 1


Data Preprocessing
Exploratory Data Analysis
Data Manipulation
Modeling
Analysis
Evaluation

4. 분석 절차 1

Data Preprocessing

"Data는 깔끔하지 않습니다!"



Preprocessing으로
Data와 분석을 연결

4. 분석 절차 1

Data Preprocessing의 역할

1	1	1	3	9999
2	2	%@$%#	2	1
3		3	1	2
4	3	4	3	^--^
5	4		4	3

Preprocessing 방안:

1) 빈 값에 대한 처리: 해당 행 삭제, 치환, 등
2) 이상한 값: 해당 행 삭제, 치환, 등
3) 범위 외의 값: 해당 데이터 생성 환경 검토

4. 분석 절차 1

Data Preprocessing 中

Data Partitioning

-모형을 구축하고 모형의 성능을 평가하기 위해 주어진 데이터를

train 데이터와 test데이터로 나누는 것

-train 데이터와 test 데이터는 랜덤하게 선택되며, 서로 중복되지 않음

4. 분석 절차 1

Exploratory Data Analysis(EDA)

Exploration

데이터에서 변수 발견

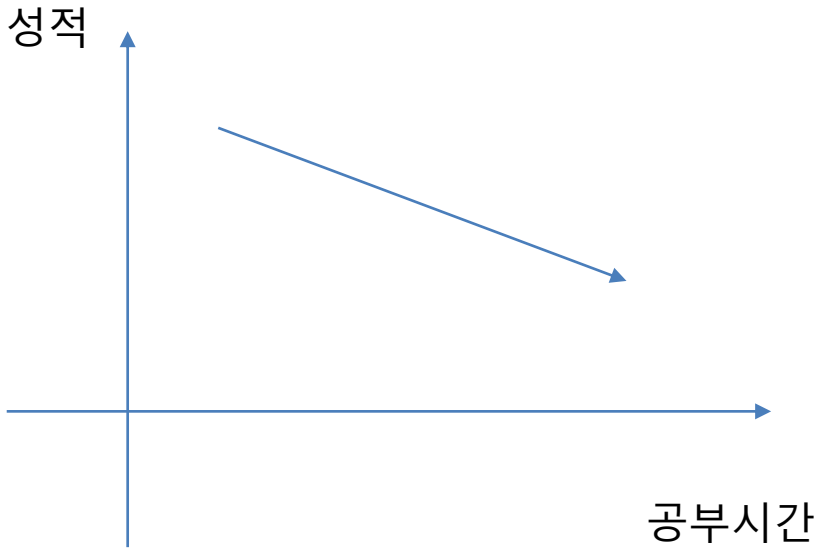
- 변수 단위의 요약값 확인(평균, 최대, 최소, 표준편차 등)
- 변수 단위의 그래프 그리기
- 두 변수에 대한 요약값 확인
- 두 변수에 대한 그래프 그리기

4. 분석 절차 1

Exploratory Data Analysis(EDA)

공부시간	성적
10	70
9	80
8	90
7	100

공부시간 평균: 8.5 시간
성적 평균: 85점



다양한 통계량(한 변수, 두 변수)
다양한 그래프

4. 분석 절차 1

Data Manipulation

변수

Target 또는 Y = Output = Dependent

X =Input = Independent = Exploratory

4. 분석 절차 1

Data Manipulation

변수에 대한 선택:

모델링 전 가장 중요한 단계!

기획된 분석 목적의 이해가 중요! (지도 VS 비지도)

<지도 학습>

Target(=Y변수)은?

X 변수 중 어떤 것을 선택할까?

5. 분석 절차 2

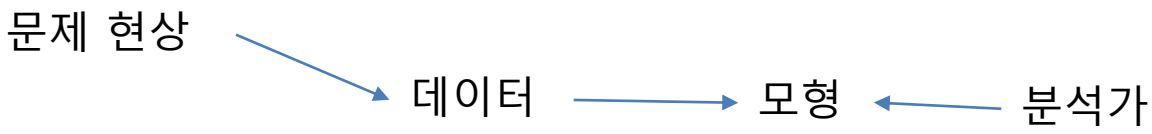
- **Modeling: 주어진 데이터로 기획된 분석 목적에 부합한 기법을 선택하는 단계**
- **Analysis: 선택된 기법을 바탕으로 실제 분석을 수행하여 모형을 수립하며, 주로 훈련 데이터를 사용하여 분석**
- **Evaluation:평가 데이터를 바탕으로 모형의 성능을 파악함**

5. 분석 절차 2

모형

- **분석 목적에 맞는 적절한 모형 선택이 중요**
- **추론과 예측 중 하나에 특화된 모형들**
- **선택된 변수를 고려!**

모형 / 모델 : 데이터를 바라보는 우리의 관점



5. 분석 절차 2

➢ **Data Analytics 모형 구분**

지도학습 (Supervised Learning)

종속 및 독립변수를 이용하여 주어진 독립(설명)변수를 바탕으로 종속(반응)변수 예측 모형 제시

예: 회귀/분류 모형

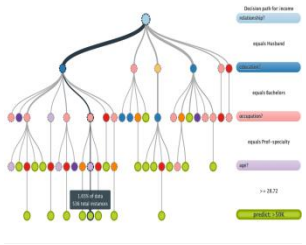
+

비지도학습 (Unsupervised Learning)

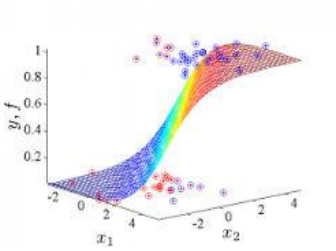
Target(종속변수/반응변수)이 없으며, 독립(설명)변수 간의 관계나 이를 바탕으로 개체들을 구분하여 의미 있는 결과를 제시

예:군집 분석, 연관성 분석, 주성분 / 요인분석

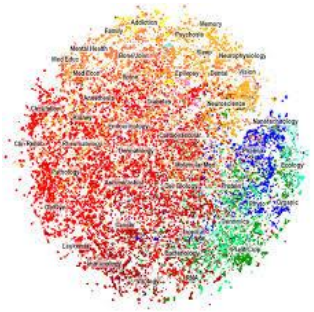
[decision tree]



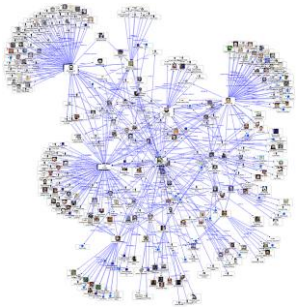
[logistic regression]



[clustering analysis]



[link analysis]



5. 분석 절차 2

모형 선택의 예

분석 상황-지도학습

Y변수는 어떤 성격인가? 수치 VS 범주

X변수로 Y변수를 잘 설명해야할까?
예측해야할까?

가용한 모형들!

분석 상황-메시지 내용으로 스팸메일 발견!

Y변수? 스팸메일 VS 정상메일

X변수: 메시지 내용
스팸메일을 잘 예측하는 것이 중요

가용한 모형들!
-분류모형 / SVM, DNN, NB 등의 모형들!

5. 분석 절차 2

모형 선택의 예2

분석 상황-지도학습

Y변수는 어떤 성격인가? 수치 VS 범주

X변수로 Y변수를 잘 설명해야할까?
예측해야할까?

가용한 모형들!

분석 상황-금리에 따른 기업 부도 여부

Y변수? 기업 부도 여부

X변수: 금리
금리에 따른 부도 발생을 설명하는 것이 중요

가용한 모형들!
-분류모형 / Logistic Regression!

5. 분석 절차 2

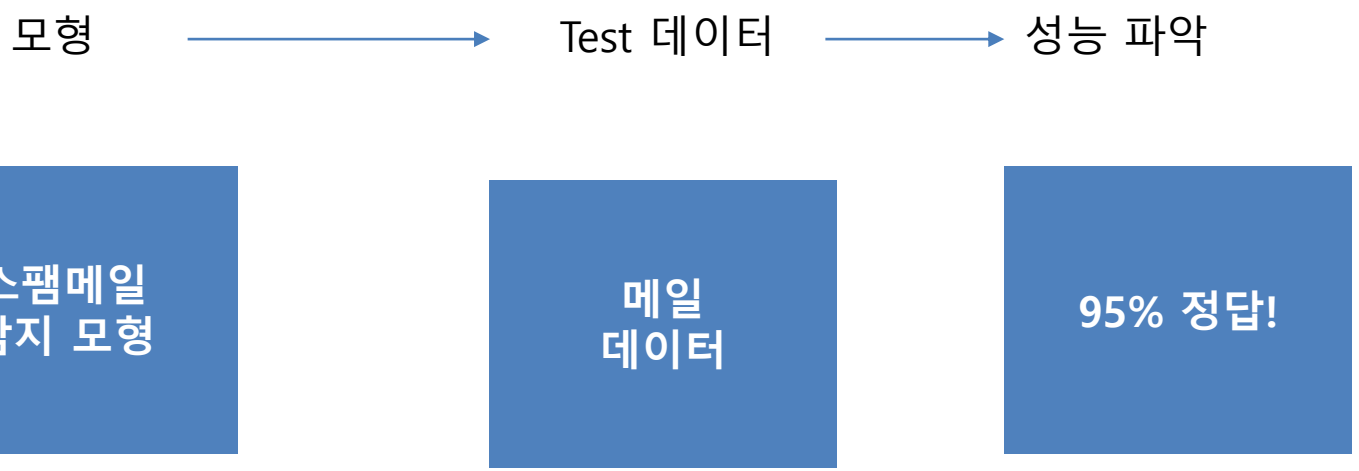
Evaluation

“ 이 모형 써도 되나?”

평가를 위한 대표적인 지표

Accuracy

Mean Squared Error

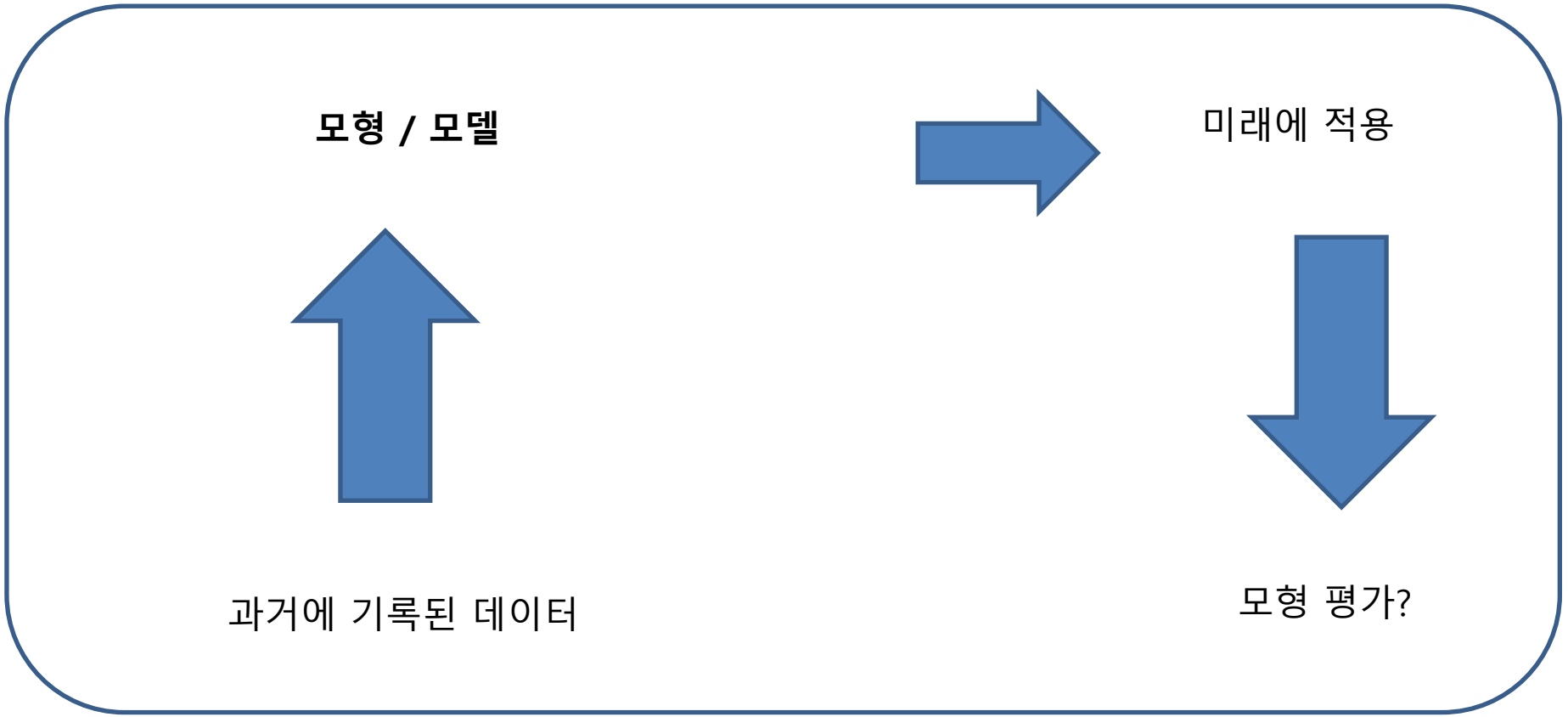


6. 분석 시 고려 사항

- 분석 모형 평가를 위한 데이터 파티셔닝: 주어진 데이터를 Train 데이터와 Test 데이터로 나눠, 모델링 결과에 Test 데이터를 적용해 성능을 가늠
- 지도 학습 VS 비지도 학습의 평가 차이: 지도학습은 성능이 수치화되어 파악되지만, 비지도 학습은 분석 환경, 분석 목적 등을 고려해서 평가될 수 있음
- 모형 평가 시 주의점: 현재 평가는 앞으로의 성능에 대한 추정이므로 맹신하기 보다는, 향후 지속적인 모니터링과 업데이트를 기획해야 함

6. 분석 시 고려 사항

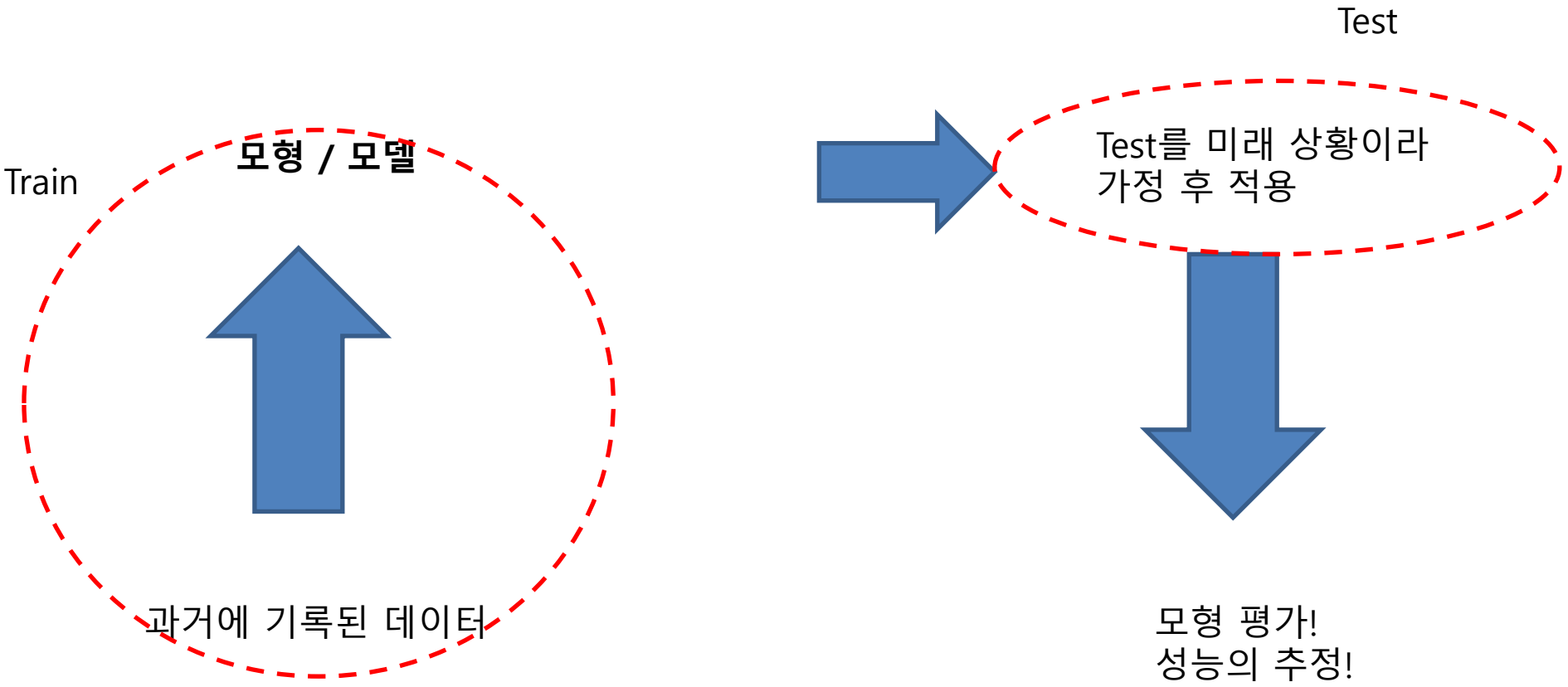
모형 평가



6. 분석 시 고려 사항

데이터 파티셔닝과 모형 평가

데이터 파티셔닝: 주어진 데이터를 Train과 Test로 나눔



6. 분석 시 고려 사항

지도학습 VS 비지도학습

지도학습의 모형 평가

- Target이 있는 분석
- 구체적인 평가 기준
- 수치화된 성능-정분류율, RMSE 등

"얼마나 잘 맞추는가?"

6. 분석 시 고려 사항

지도학습 VS 비지도학습

비지도학습의 모형 평가

- **Target이 없는 분석**
- **구체적인 평가 기준 없음**
- **상대적이고 주관적인 평가**

“얼마나 분석 목적과 기획 의도에 부합하는 결과인가?”

6. 분석 시 고려 사항

모형 평가 시 주의사항

- 분석의 목적을 고려해야 함!
- 성능이 너무 좋아도, 성능이 너무 나빠도 주의!
- Test 데이터를 통해 추정된 모형의 성능을 맹신하지 말 것!
- 결국은 분석가와 분석팀에 의한 정성적인 해석 필요!

7. 분석 기획과 비즈니스 아이디어

데이터 분석은 왜 하는 것일까?

데이터 분석 = 데이터로 표현된 문제 해결

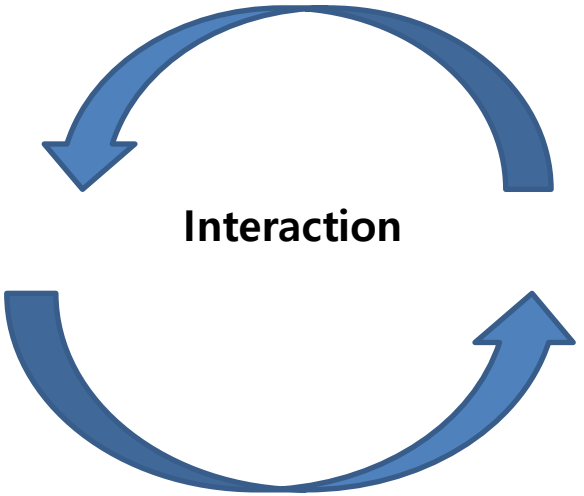
7. 분석 기획과 비즈니스 아이디어

데이터 분석 기획: 문제 상황과의 지속적인 Interaction

데이터로 표현

문제

?



문제 해결

분석가

7. 분석 기획과 비즈니스 아이디어

AI 현업 적용 = Ideation이 핵심!

Ideation!

그런데, 현업 적용은?

7. 분석 기획과 비즈니스 아이디어

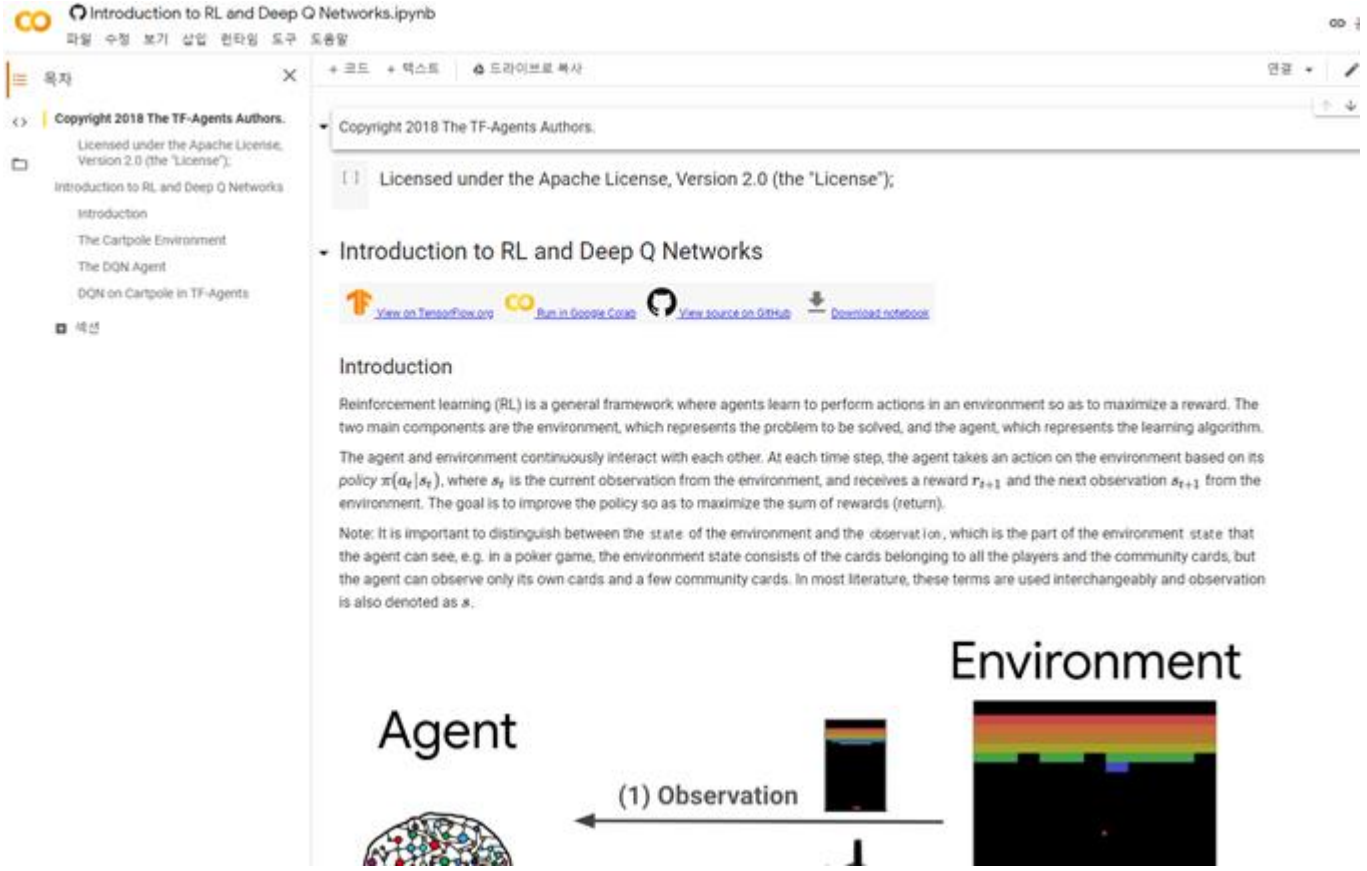
작지만 큰 AI 서비스 기획

1. 거창한 서비스는 나중에
2. 기술을 바로 응용할 수 있는 기획
3. 모니터링을 통한 지속적 개선
4. 역량 내재화를 통한 서비스 확대

8. 실습 도구

➢ 구글 Colab?

- 클라우드 上 파이썬 환경
- 쥬피터노트북과 동일!



8. 실습 도구

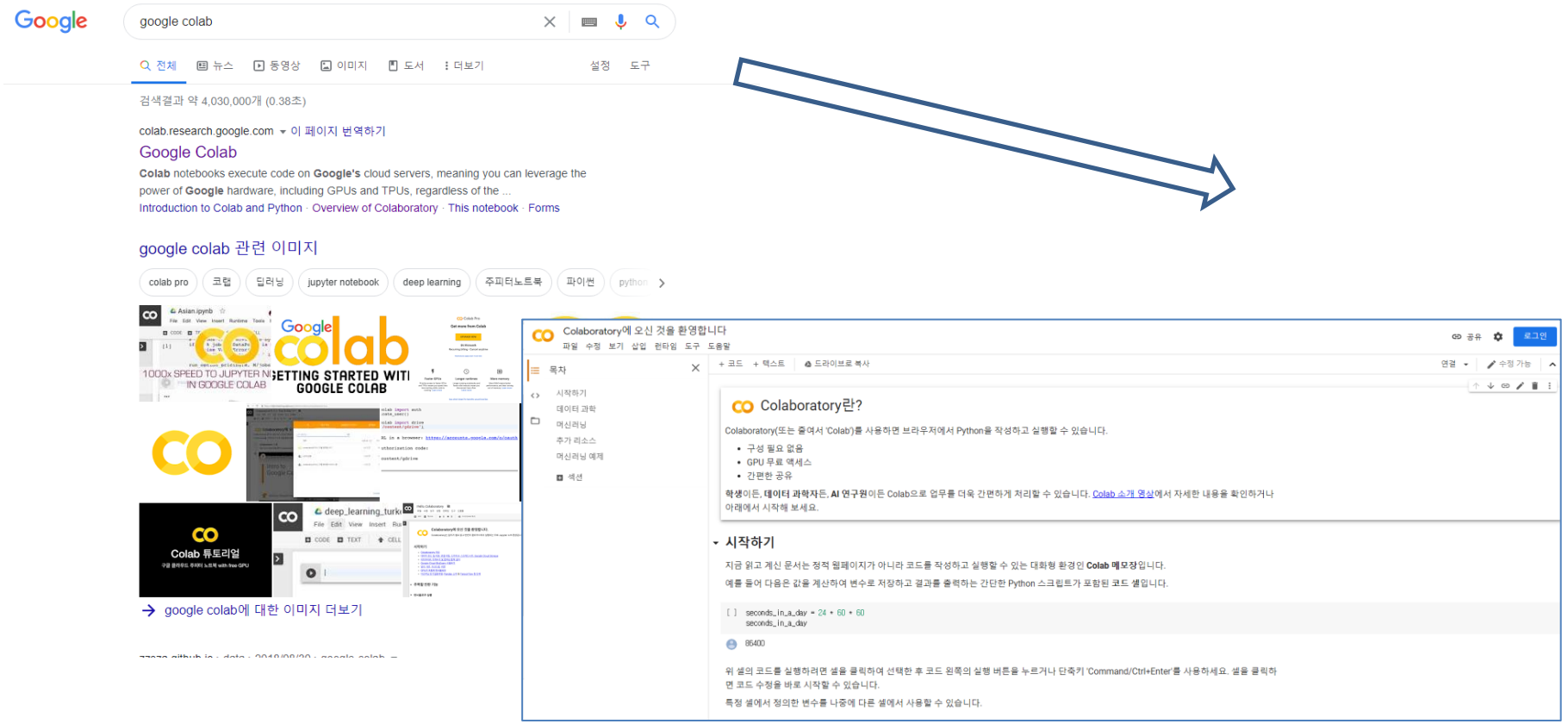
- 구글 Colab?

	Google Colaboratory	Kaggle Notebooks	Amazon SageMaker
BackGround Platform	Google Cloud	Google Cloud	AWS
Notebooks Types	Jupyter Notebook	Jupyter Notebook, Scripts	Jupyter Notebook, Scripts
Support Language	Python	Python, R	Python,R
GPU	Nvidia K80, T4, P4 and P100 GPU quota: vary overtime Free Edition: No choise for GPU Type	NVIDIA Tesla P100 GPU quota: 30 hours per week	NVIDA K80, Tesla V100, Tesla M60, T4 and Tesla A100 GPU quota: On-demand
TPU	Support TPU TPU quota: vary overtime	Support TPU v3-8 TPU quota:30 hours per week	No
Memory	vary overtime	Max 16 GB	On-demand
Disk Space	Follow your Google Driver	Max20 GB	On-demand
LifeTime & Idle Time	Lifetime, Max 12 hours	Lifetime, Max 9 hours Idletime, Max 1 hour	On-demand
Deep learning Framework	TensorFlow, PyTorch···	TensorFlow, PyTorch···	MXNet, TensorFlow, PyTorch···
Documents & Course	Basic	Rich	Basic
Free Edition	Free Edition and Colab Pro	Free	No Free Edition

8. 실습 도구

➢ 구글 Colab?

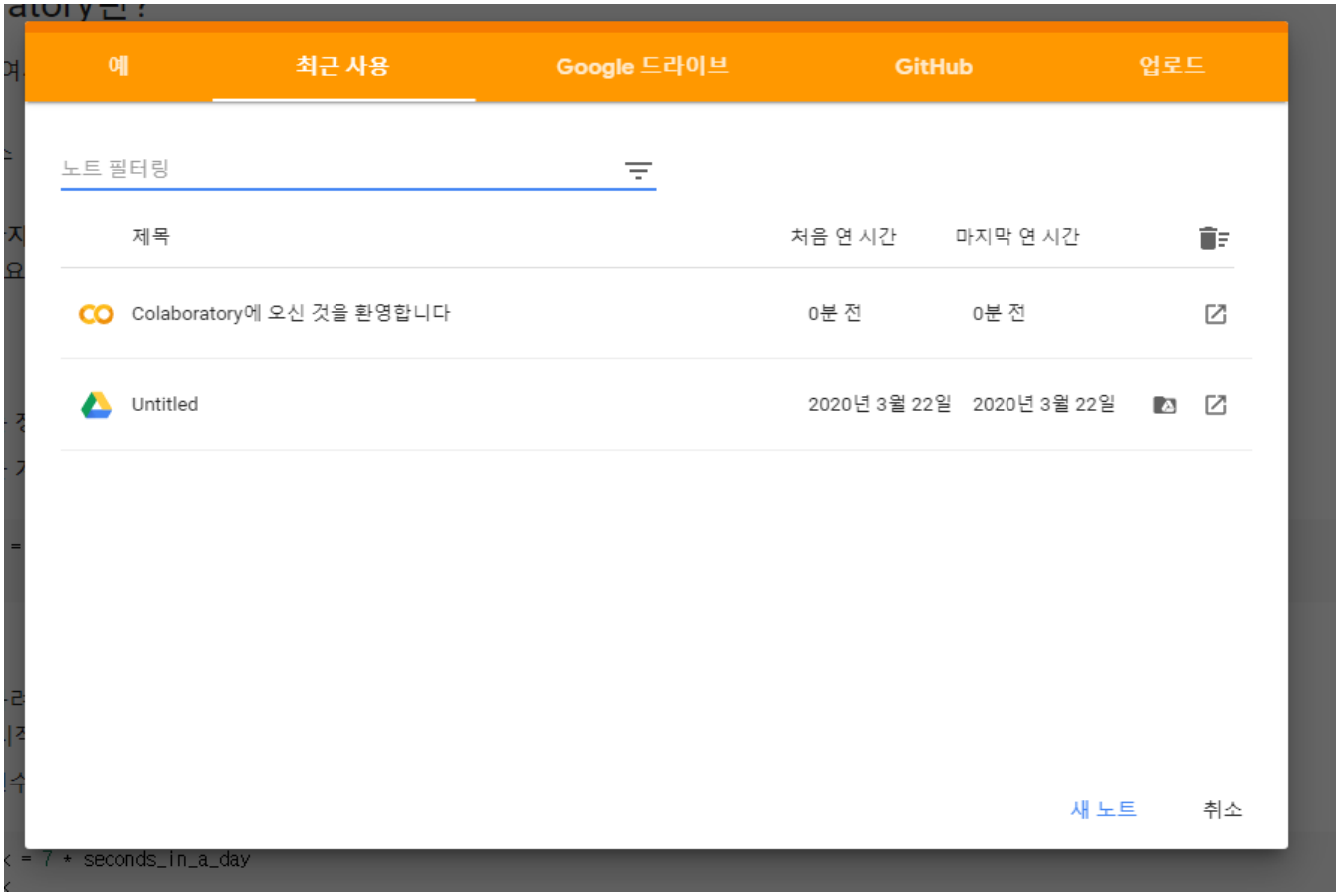
- 클라우드 상에서 GPU를 지원
- 12시간 세션시간의 제한
- 구글 드라이브



8. 실습 도구

➢ Colab

- **데이터+분석/활용 플랫폼**



Q&A