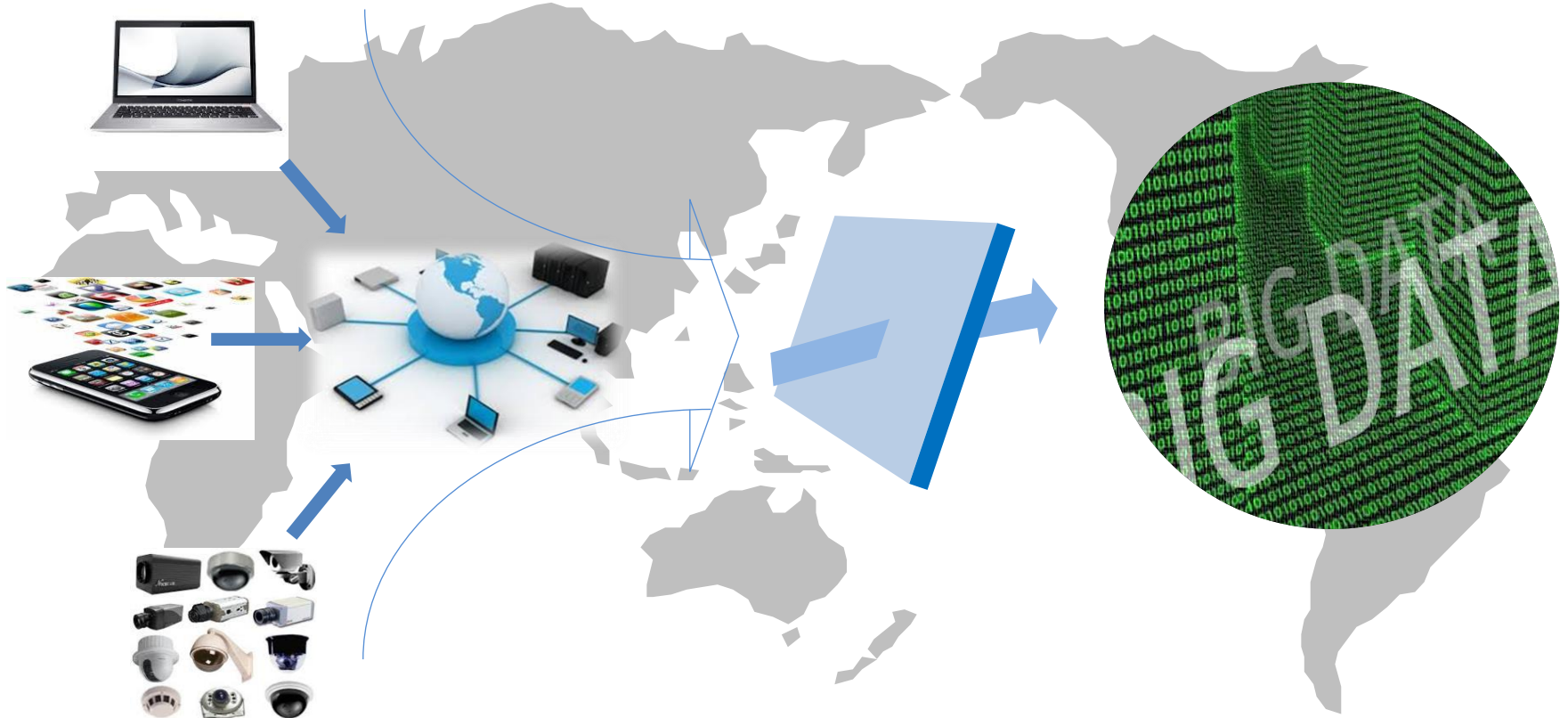


AI 인재육성 프로그램(2차) 심화과정

1. 데이터 분석 기획

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 Data는...

Velocity

Volume

Variety



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 계획 수립
- **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 기획 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

분석 방법론과 빅데이터 분석 기획

– TDSP: Team Data Science Process

- 예측 분석 솔루션 및 지능형 애플리케이션을 효율적으로 제공하는 데이터 과학 방법론
- 팀 협업 및 학습 개선
- By Microsoft

- 
- 데이터 과학 수명 주기 정의
 - 표준화된 프로젝트 구조
 - 데이터 과학 프로젝트에 권장되는 인프라 및 리소스
 - 프로젝트 실행에 권장되는 도구 및 유틸리티

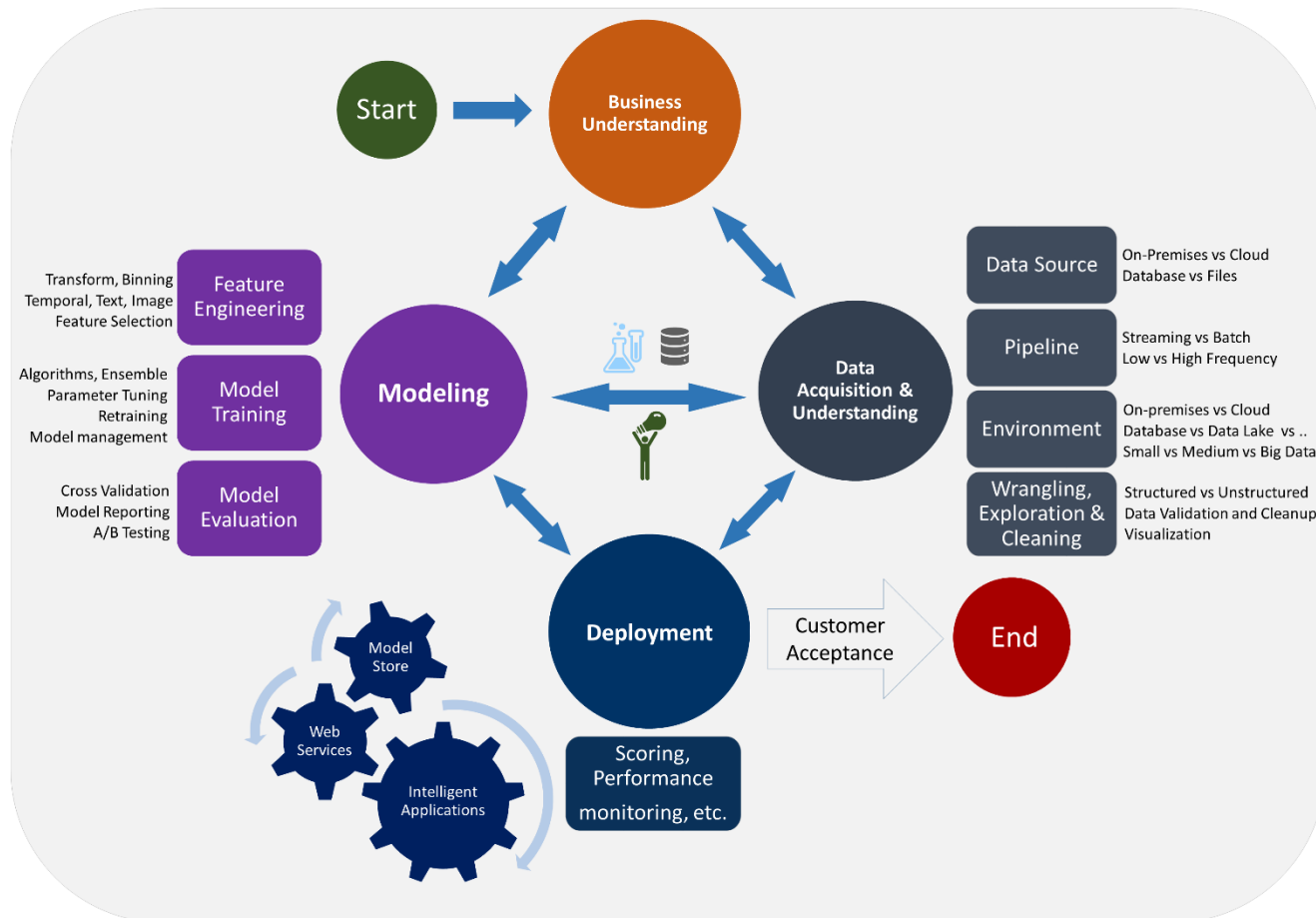
1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획

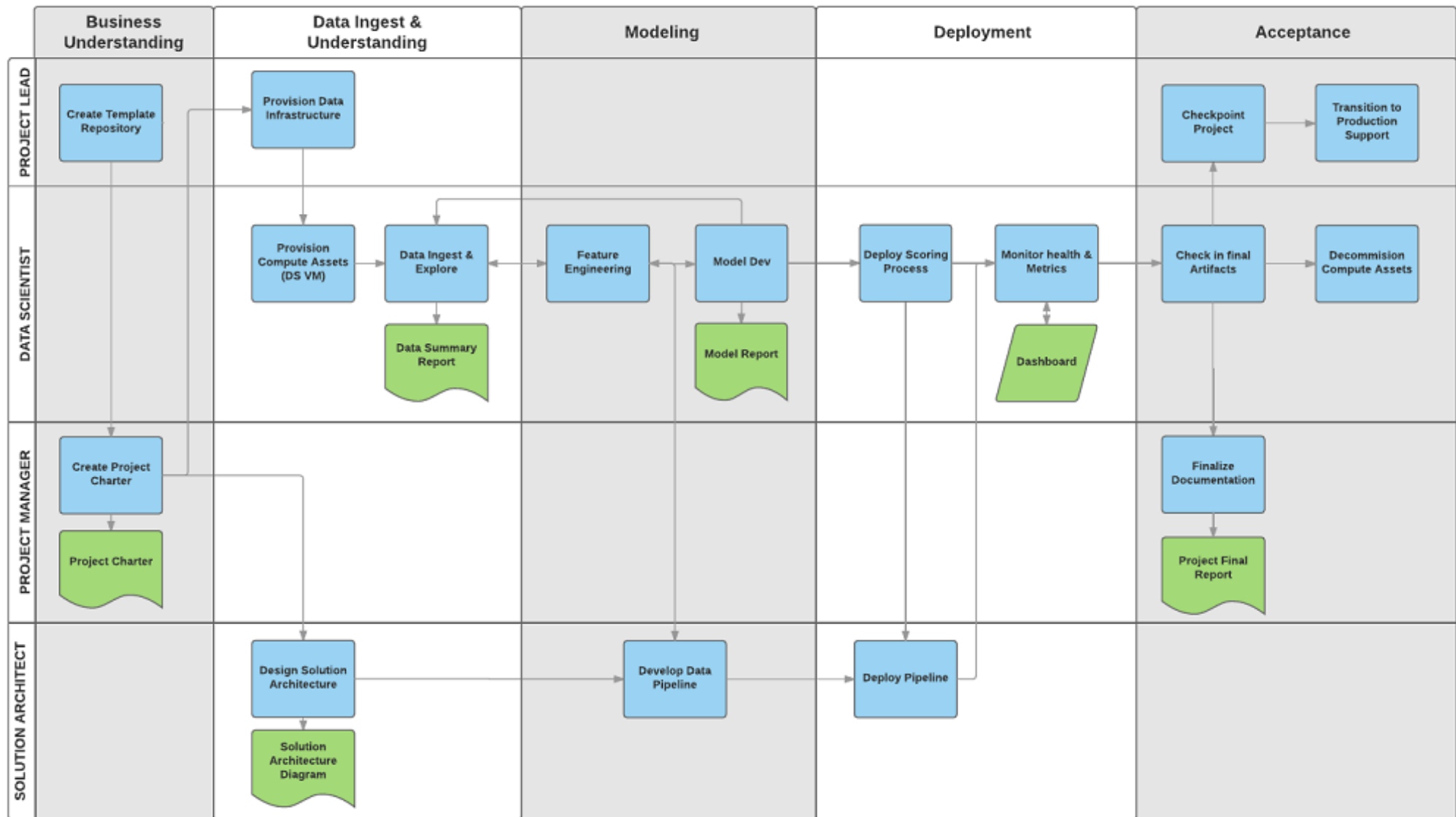
- TDSP

- 데이터 과학 수명 주기: **비즈니스 이해->데이터 취득 및 이해->모델링->배포->Acceptance**

Data Science Lifecycle



1. 빅데이터와 분석 기획

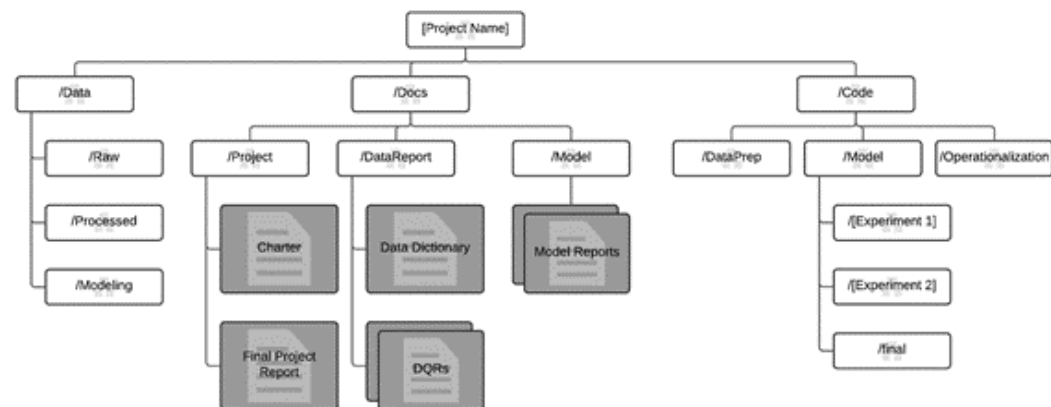


1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획: TDSP

- 표준화된 프로젝트 구조

- 팀 협업 활성화를 위한 Git, TFS 또는 Subversion과 같은 VCS(버전 제어 시스템)
- 버전 관리, 정보 보안 및 협업을 위해 VCS에 각 프로젝트별로 별도 리포지토리
 - » 폴더 구조와 표준 위치에 꼭 있어야 하는 파일들로 된 템플릿을 제공
 - » 템플릿의 예
 - 비즈니스 문제 및 프로젝트의 범위를 문서화하는 프로젝트 헌장
 - 원시 데이터의 구조 및 통계를 문서화하는 데이터 보고서
 - 파생된 기능을 문서화하는 모델 보고서
 - ROC 곡선 또는 MSE와 같은 모델 성능 측정

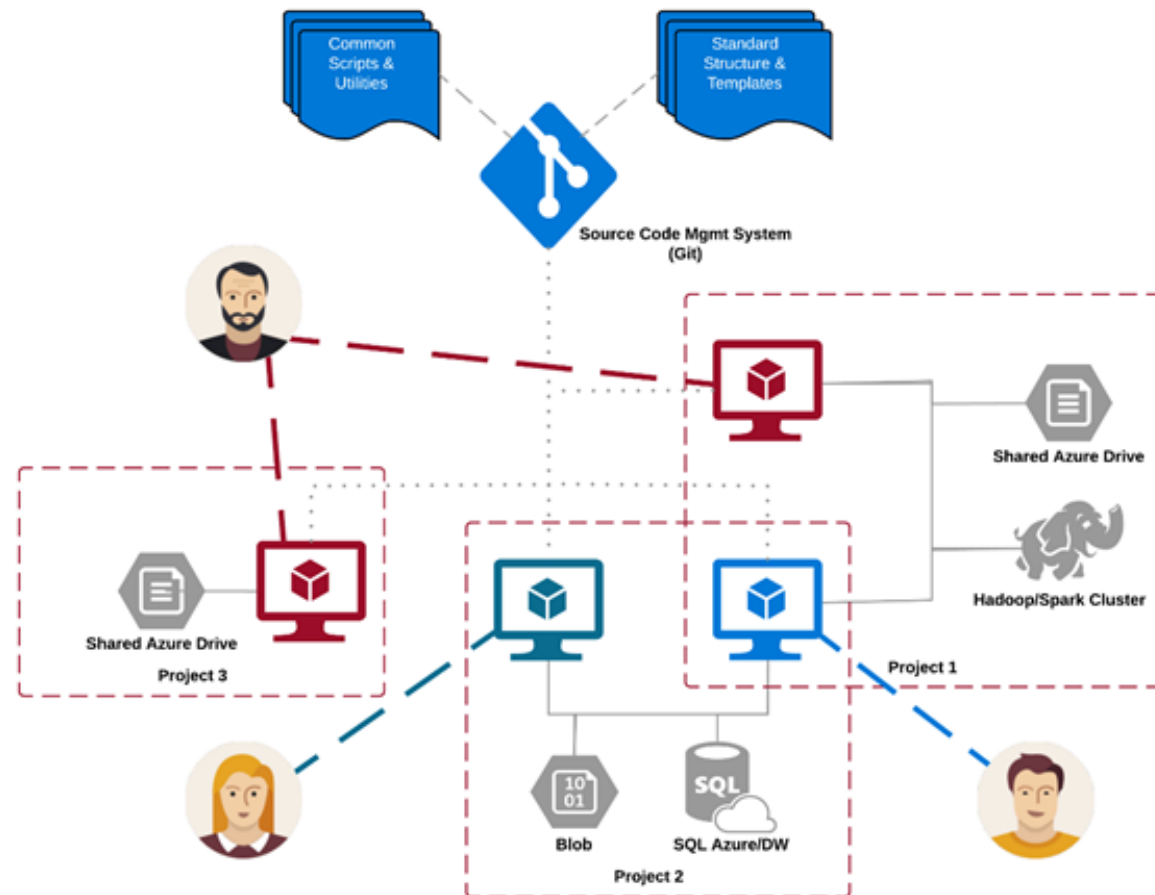


1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획

– TDSP

- 데이터 과학 프로젝트에 필요한 인프라 및 리소스
 - 데이터 세트를 저장하기 위한 클라우드 파일 시스템
 - 데이터베이스
 - 빅 데이터(SQL 또는 Spark) 클러스터
 - 기계 학습 서비스



1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획

– TDSP

- 프로젝트 실행에 권장되는 도구 및 유틸리티
 - TDSP의 도입을 신속하게 시작하도록 도구와 스크립트 기본 리소스 제공
 - 데이터 탐색이나 기준 모델링 같은 일반적인 작업 중 일부를 자동화에 기여
 - 공유 도구와 유틸리티를 팀의 공유 코드 리포지토리로 배포 및 업데이트

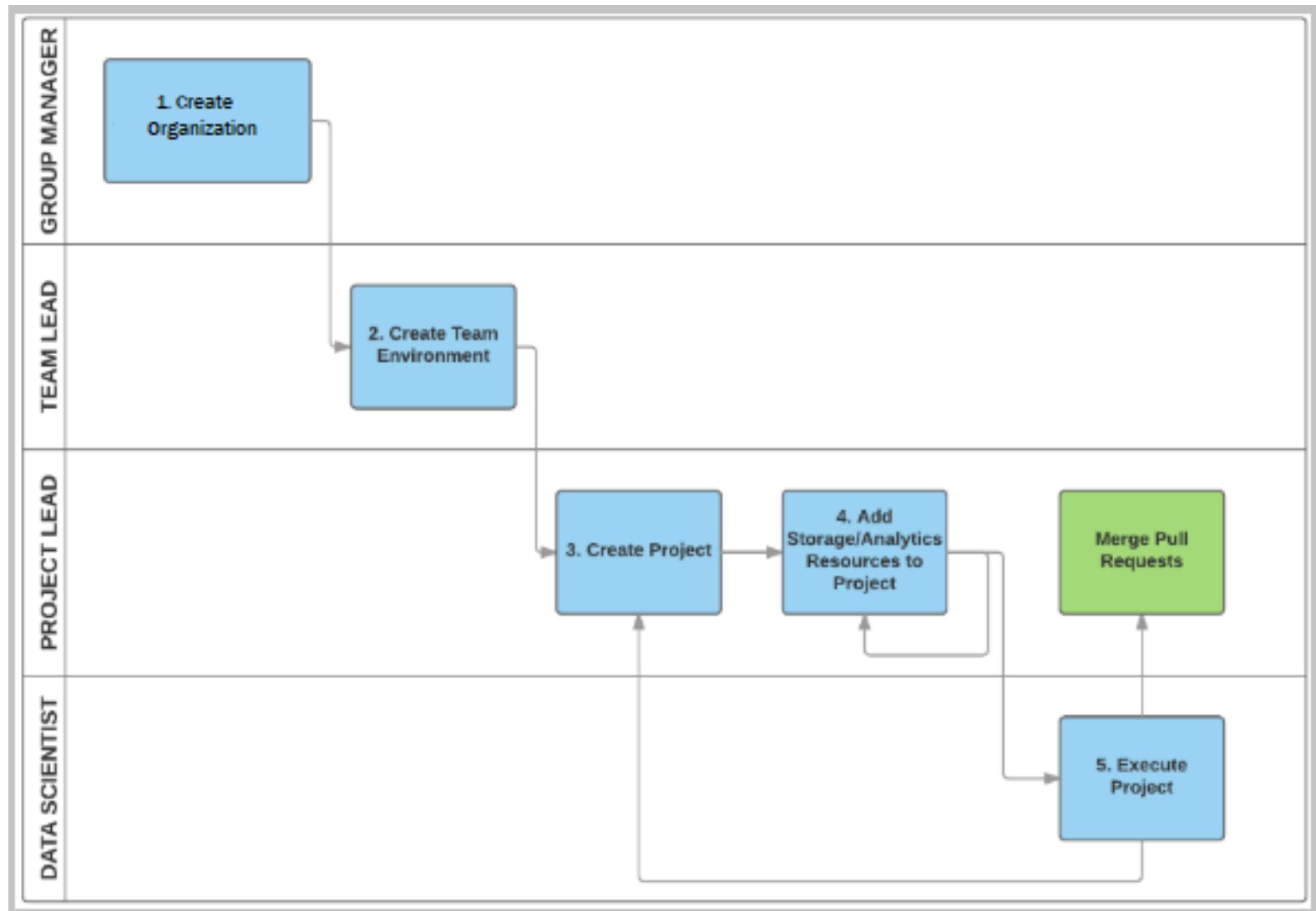
1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획: TDSP 역할 구분

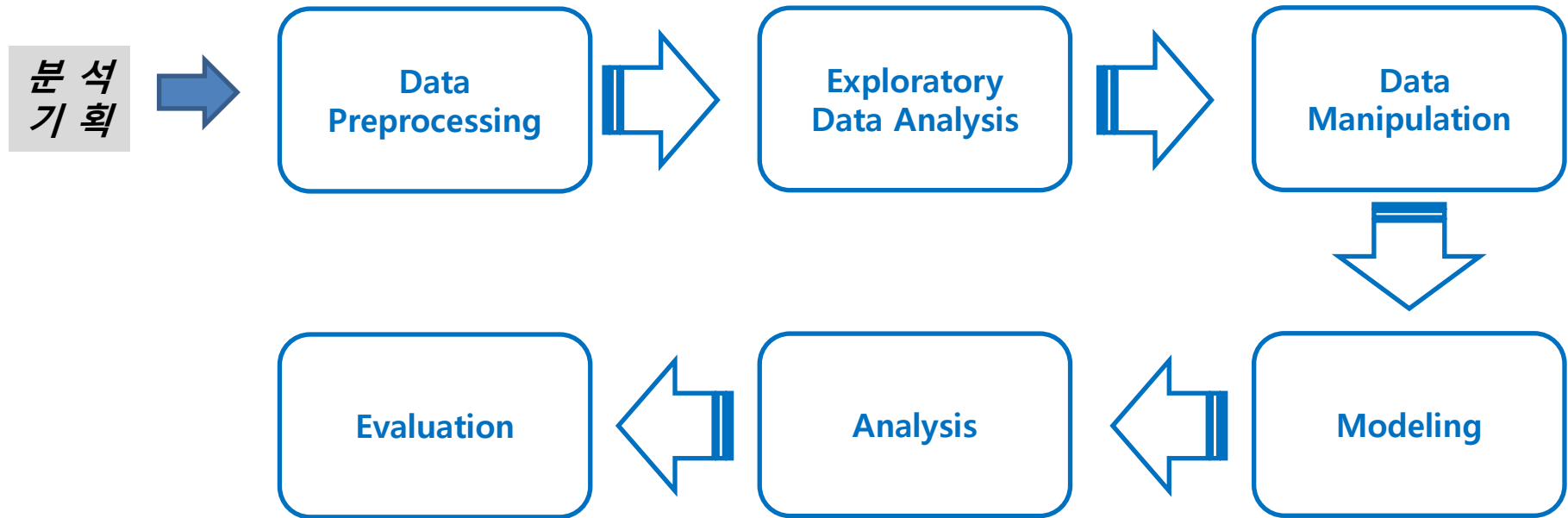
- 그룹 관리자: 기업 내의 전체 데이터 과학 단위를 관리. 데이터 과학 단위에는 여러 팀이 있고 각 팀은 고유 비즈니스 도메인에서 데이터 과학 프로젝트 수행
- 팀 리더: 기업의 데이터 과학 단위 팀 관리. 팀은 여러 명의 데이터 과학자로 구성. 소규모 데이터 과학 단위의 경우 그룹 관리자와 팀 리더가 동일할 수 있음
- 프로젝트 리더: 특정 데이터 과학 프로젝트에서 개별 데이터과학자의 작업 관리.
- 프로젝트 개별 기여자: 데이터 과학 프로젝트를 수행하는 데이터 과학자, 비즈니스 분석가, 데이터 엔지니어, 설계자 등

1. 빅데이터와 분석 기획

분석 방법론과 빅데이터 분석 기획: TDSP 역할 구분



1. 빅데이터와 분석 기획



1. 빅데이터와 분석 기획

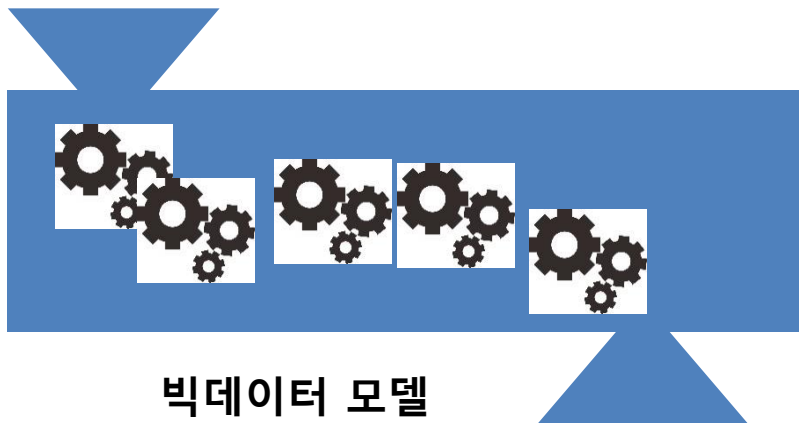
데이터 분석 기획

- 분석 과제
- 데이터
- 어떤 기법
- 기대효과
- 업데이트 계획 수립

1. 빅데이터와 분석 기획

분석 기획

고객	요금제	시청시간/일	커멘트	탈퇴
1	10	1	너무 좋아요	X
2	5	2	불만해요	X
3	1	1	좋아요	X
4	5	0.2	그럭저럭	O



주기적 갱신

좋은 성능!
목적의 달성!

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 것
- 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

2. 분석 기획 시 고려 사항

Inference VS Prediction

2. 분석 기획 시 고려 사항

“회사의 0000분야의 000에 대한 데이터를 분석해서, 기존 프로세스의 문제점을 발견하고, 향후 개선점을 제시하여 0000분야의 수익을 증대하고자 합니다 ”

VS

“회사의 0000분야의 000에 대한 데이터를 분석해서 중요한 변수들을 스크리닝 하고 검토하여 모델링에 활용하고자 합니다. 그 결과로 높은 정확도의 모형을 수립하고자 합니다”

2. 분석 기획 시 고려 사항

데이터 사이언티스트?!

1. 분석할 사람이 없으면, 분석가를 뽑을까?
2. 아니면, 컨설팅 맡길까?
3. 채용된 데이터 사이언티스트가 다 할 거야?!

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 아키텍처
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, 비정형 데이터 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, 프로
토타이핑을 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

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



Machine Learning Repository
Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)



☒ Repository ☐ Web 

[View ALL Data Sets](#)

Check out the [beta version](#) of the new UCI Machine Learning Repository we are currently testing! [Contact us](#) if you have any issues, questions, or concerns. [Click here to try out the new site.](#)

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 622 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With: 

Latest News:

09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!

04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!

03-01-2010: [Note](#) from donor regarding Netflix data


10-16-2009: Two new data sets have been added.

09-14-2009: Several data sets have been added.

03-24-2008: New data sets have been added!

06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: [UJI Pen Characters \(Version 2\)](#)




Task: Classification

Data Type: Multivariate, Sequential

Instances: 11640


Newest Data Sets:

06-05-2021:




[Average Localization Error \(ALE\) in sensor node localization process in WSNs](#)

05-25-2021:




[9mers from culpdb](#)

05-18-2021:




[TamilSentiMix](#)

05-02-2021:




[Accelerometer](#)

04-21-2021:



[Synchronous Machine Data Set](#)


04-21-2021:



[Synchronous Machine Data Set](#)


Most Popular Data Sets (hits since 2007):

4476669:




[Iris](#)

2382387:




[Adult](#)

1842408:




[Wine](#)

1772912:




[Wine Quality](#)

1760410:



[Heart Disease](#)

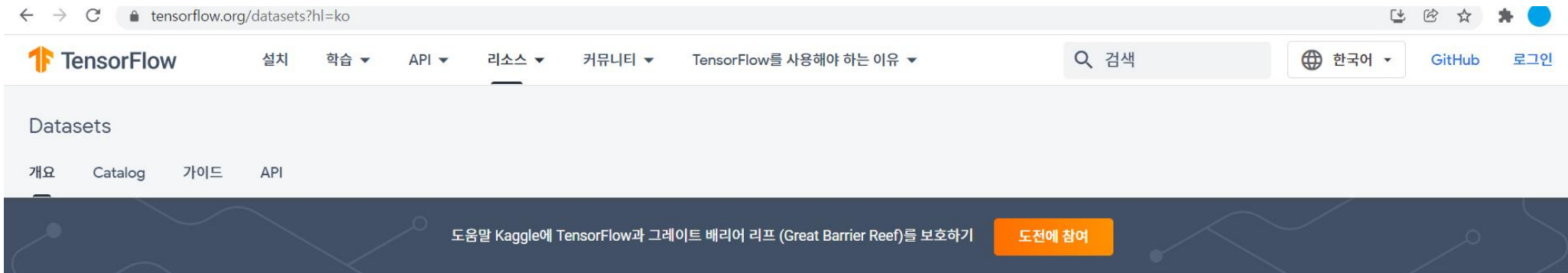
1665486:



[Bank Marketing](#)

<https://archive.ics.uci.edu/ml/index.php>

3. 분석 기획을 위한 데이터 이해



TensorFlow 데이터 세트: 바로 사용할 수 있는 데이터 세트 컬렉션

TensorFlow Datasets는 TensorFlow 또는 Jax와 같은 다른 Python ML 프레임워크와 함께 사용할 준비가 된 데이터 세트 컬렉션입니다. 모든 데이터세트는 `tf.data.Datasets` 로 노출되므로 사용이 간편한 고성능 입력 파이프라인이 가능합니다. 시작하려면 [가이드](#) 및 [데이터세트 목록](#)을 참조하세요.

```
import tensorflow.compat.v2 as tf
import tensorflow_datasets as tfds

# Construct a tf.data.Dataset
ds = tfds.load('mnist', split='train', shuffle_files=True)

# Build your input pipeline
ds = ds.shuffle(1024).batch(32).prefetch(tf.data.experimental.AUTOTUNE)
for example in ds.take(1):
    image, label = example["image"], example["label"]
```

CO 노트북에서 실행



<https://www.tensorflow.org/datasets?hl=ko>

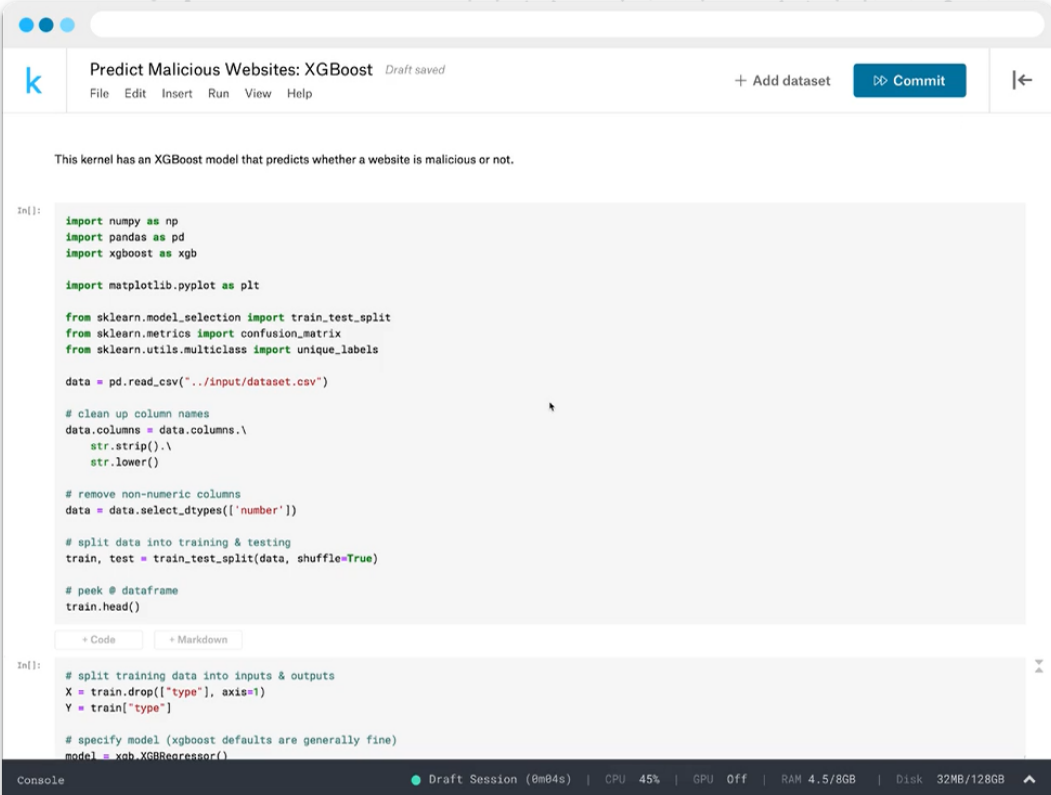
3. 분석 기획을 위한 데이터 이해

Start with more than
a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email



The screenshot shows a Kaggle Jupyter Notebook interface. The title bar reads "Predict Malicious Websites: XGBoost" with a "Draft saved" status. Below the title bar are tabs for "File", "Edit", "Insert", "Run", "View", and "Help". The main area contains two code cells. The first cell starts with imports for numpy, pandas, xgboost, and matplotlib. It then loads a dataset from a local file, cleans up column names, removes non-numeric columns, and splits the data into training and testing sets. The second cell continues by splitting the training data into inputs and outputs, and specifying an XGBRegressor model. The bottom status bar shows "Draft Session (9m84s)" and resource usage: CPU 45%, GPU Off, RAM 4.5/8GB, and Disk 32MB/128GB.

```
In[]: import numpy as np
import pandas as pd
import xgboost as xgb

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.utils.multiclass import unique_labels

data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()

In[]: # split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

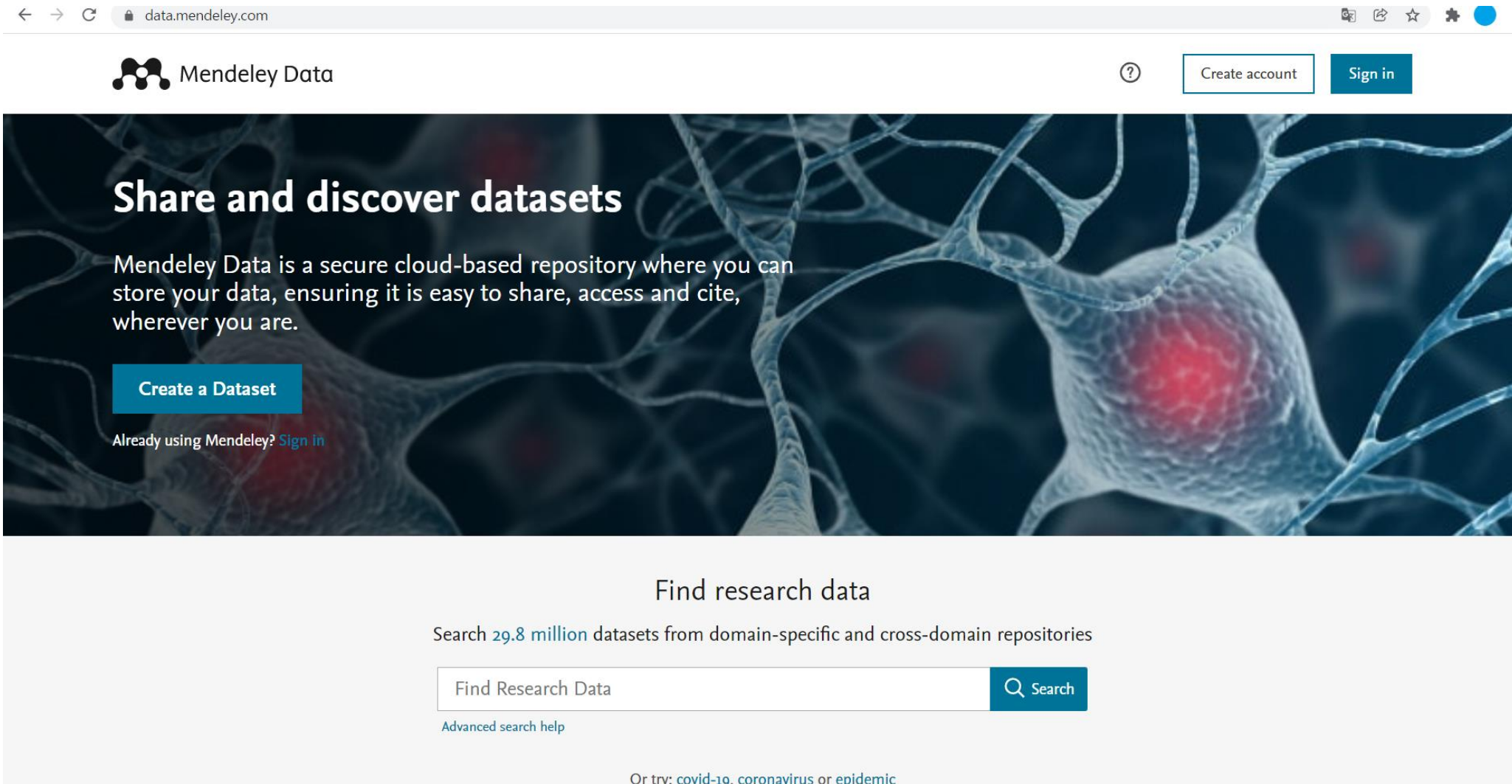
# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor()
```

We use cookies on Kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies.


Got it Learn more

<https://www.kaggle.com/>

3. 분석 기획을 위한 데이터 이해



data.mendeley.com

 Mendeley Data

?

Create account

Sign in

Share and discover datasets

Mendeley Data is a secure cloud-based repository where you can store your data, ensuring it is easy to share, access and cite, wherever you are.


[Create a Dataset](#)

Already using Mendeley? [Sign in](#)

Find research data

Search **29.8 million** datasets from domain-specific and cross-domain repositories

Find Research Data

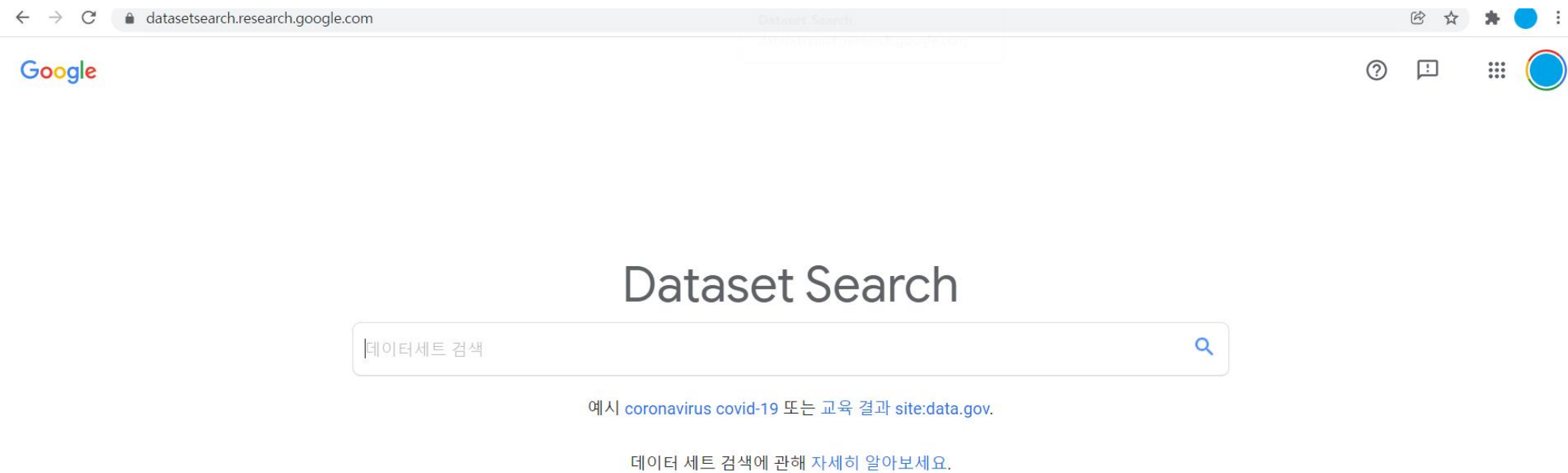
 Search

[Advanced search help](#)

Or try: [covid-19](#), [coronavirus](#) or [epidemic](#)

<https://data.mendeley.com/>

3. 분석 기획을 위한 데이터 이해



<https://datasetsearch.research.google.com/>

3. 분석 기획을 위한 데이터 이해

데이터 큐레이션의 또 다른 예, “Data Annotation”

다량의 이미지를 바탕으로 사물인식 모델링을 위해 각 이미지에 라벨링을 해주기



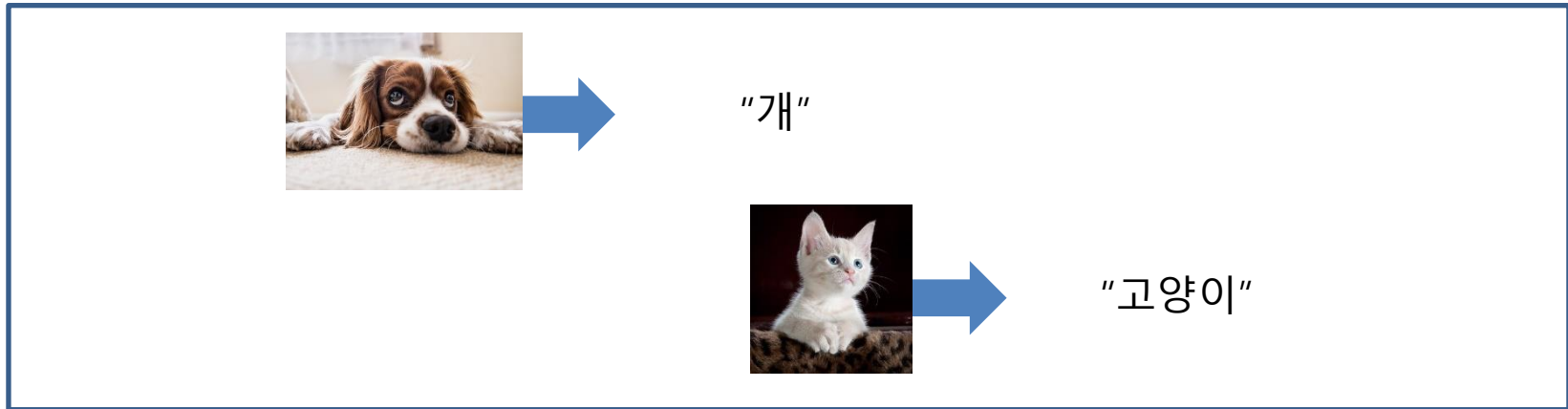
“개”



“고양이”

3. 분석 기획을 위한 데이터 이해

예



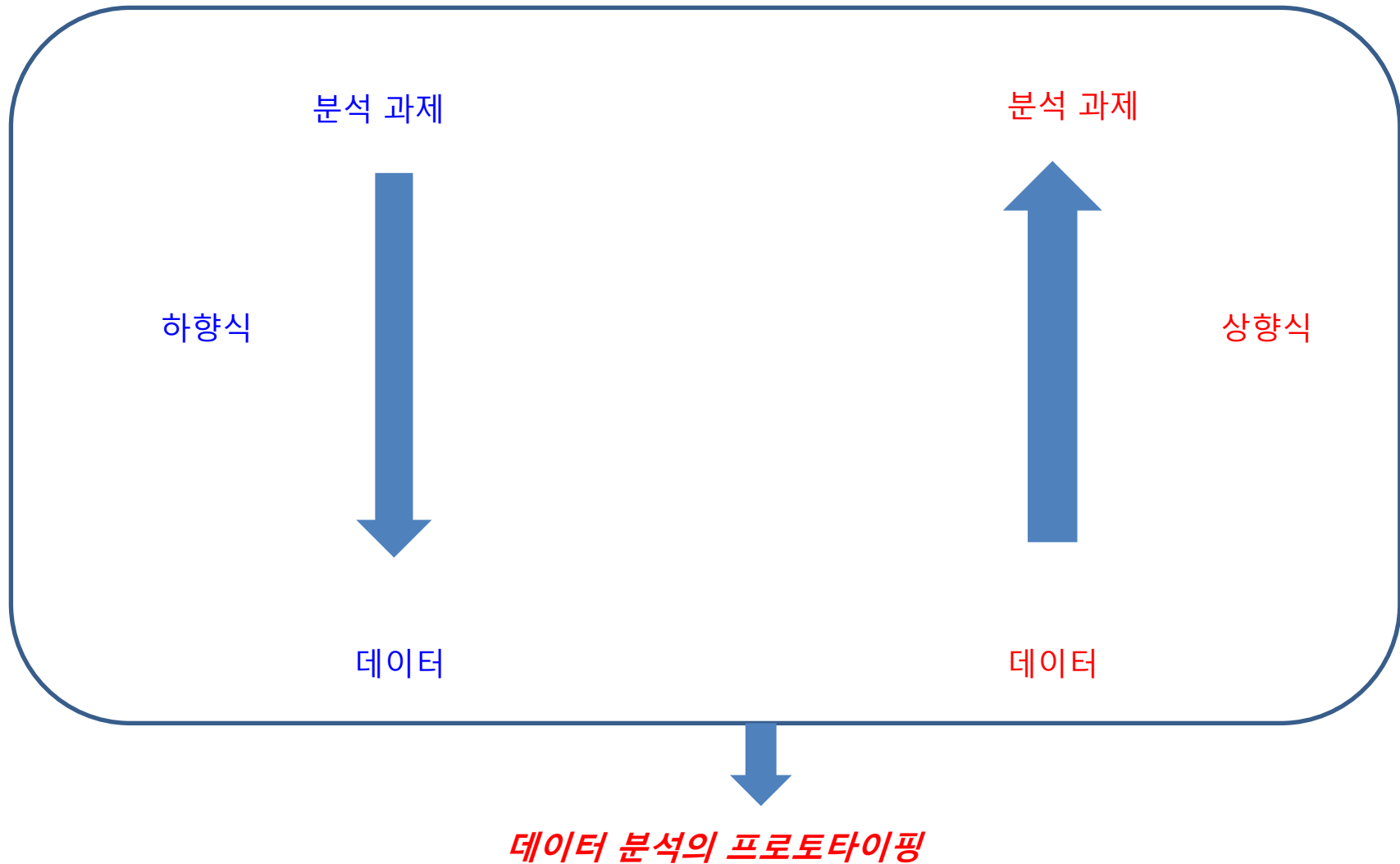
"한 데이터로 여러 분석을 할 수 있음!"

동물 분류 딥러닝!

홈 CCTV 비즈니스에 응용!

3. 분석 기획을 위한 데이터 이해

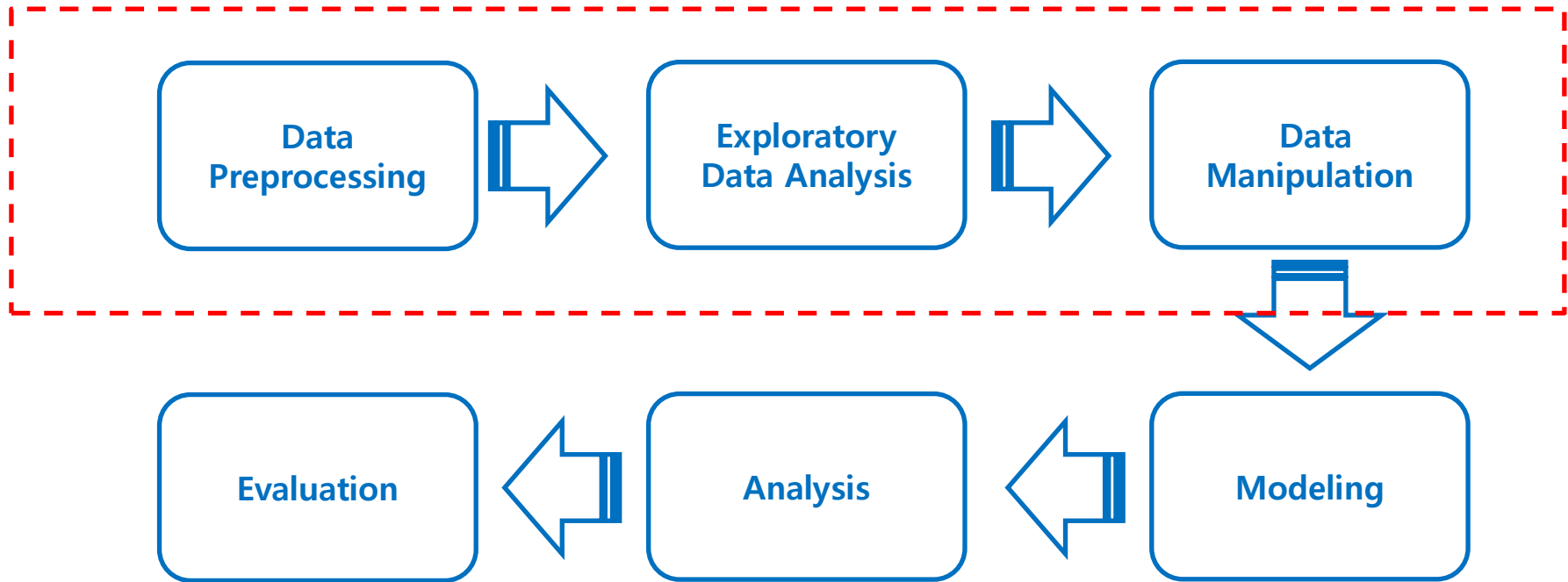
하향식 VS 상향식



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

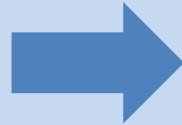
4. 분석 절차 1



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, validation, test로 구분하기도 함
- 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

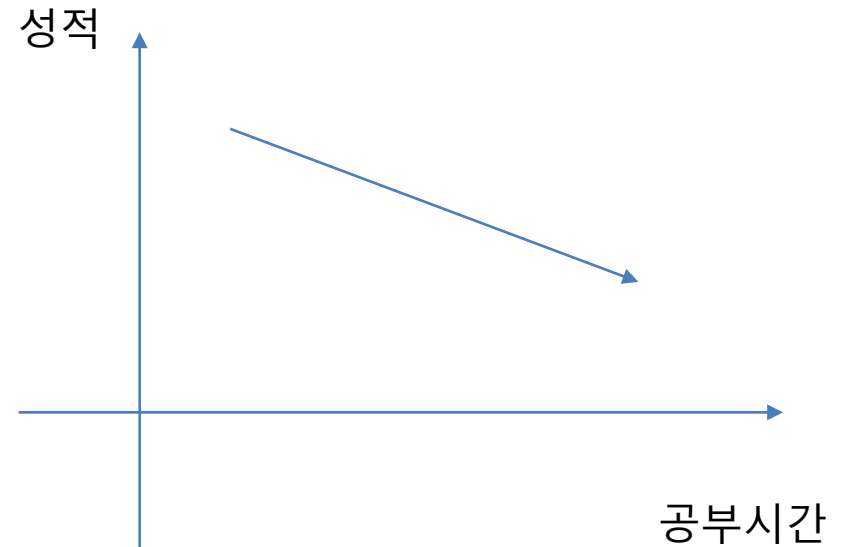
Exploratory Data Analysis(EDA)

공부시간	성적
10	70
9	80
8	90
7	100



다양한 통계량(한 변수, 두 변수)
다양한 그래프

공부시간 평균: 8.5 시간
성적 평균: 85점



4. 분석 절차 1

Data Manipulation

변수

Target 또는 Y = Output = Dependent

X = Input = Independent = Exploratory

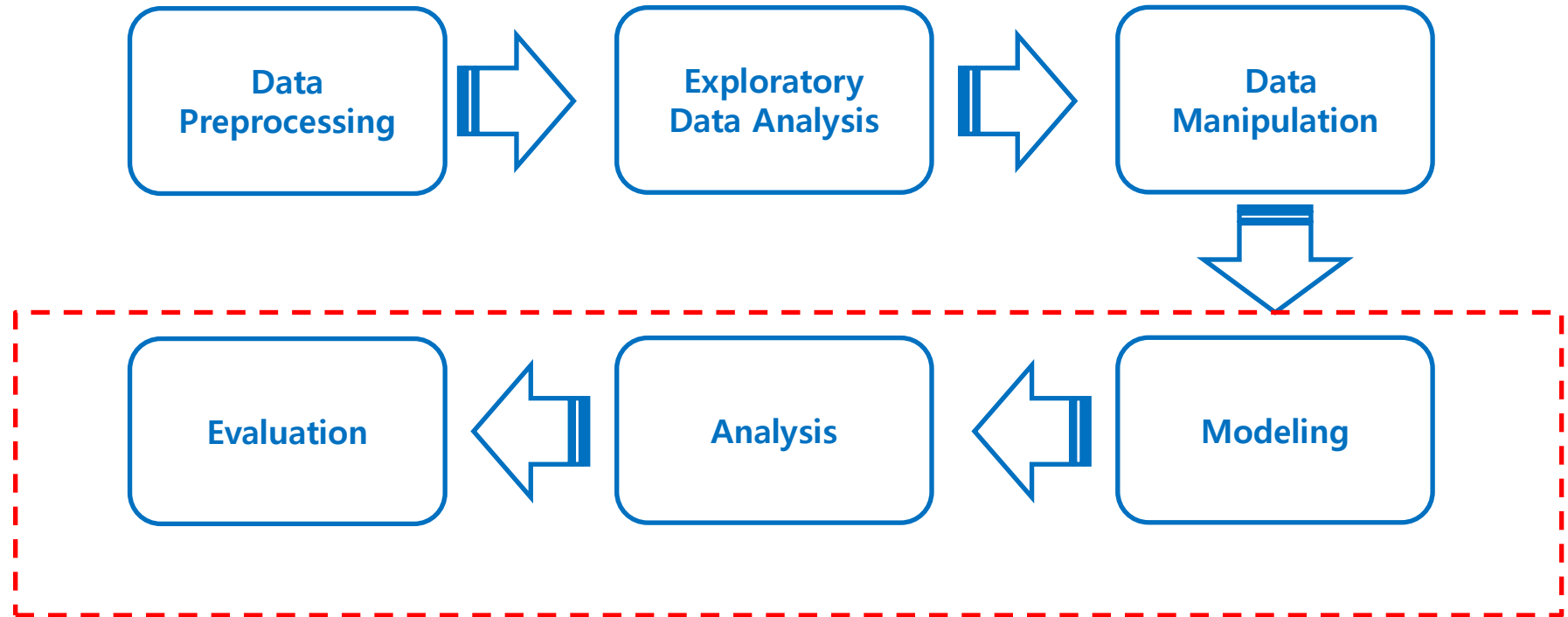
4. 분석 절차 1

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



5. 분석 절차 2

- **Modeling**

- 주어진 데이터로 기획된 분석 목적에 부합한 기법을 선택하는 단계

- **Analysis**

- 선택된 기법을 바탕으로 실제 분석을 수행하여 모델을 수립하며, 주로 훈련 데이터를 사용하여 분석

- **Evaluation**

- 평가 데이터를 바탕으로 모형의 성능을 파악함

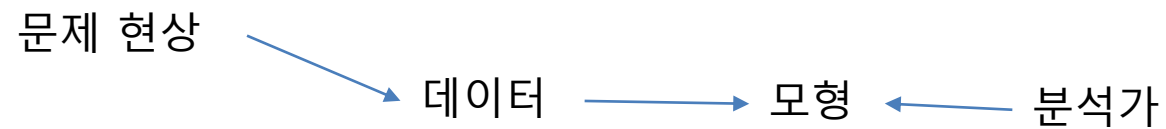
5. 분석 절차 2



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

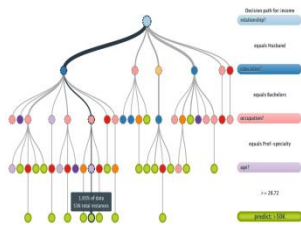


비지도학습 (Unsupervised Learning)

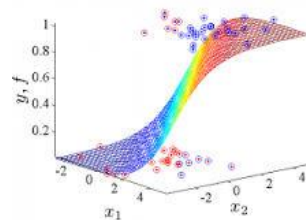
Target(종속변수/반응변수)이 없으며, 독립(설명)변수 간의 관계나 이를 바탕으로 개체들을 구분하여 의미 있는 결과를 제시

예: 군집 분석, 연관성 분석, 주성분 / 요인분석

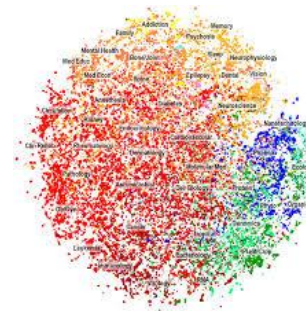
[decision tree]



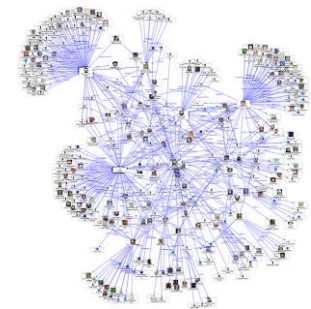
[logistic regression]



[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

Analysis단계를 위한 기획 포인트!

- *전처리된 훈련 데이터를 사용*
- *좋은 분석 도구를 활용하는 것이 중요!*
- *빅데이터의 경우 계산 이슈를 고려*

5. 분석 절차 2

Evaluation

“ 이 모형 써도 되나?”

평가를 위한 대표적인 지표

Accuracy

Mean Squared Error

모형



Test 데이터



성능 파악

스팸메일
탐지 모형

메일
데이터

95% 정답!

6. 분석 시 고려 사항

- **분석 모형 평가를 위한 데이터 파티셔닝**

- 주어진 데이터를 Train 데이터와 Test 데이터로 나눠, 모델링 결과에 Test 데이터를 적용해 성능 가늠

- **지도 학습 VS 비지도 학습의 평가 차이**

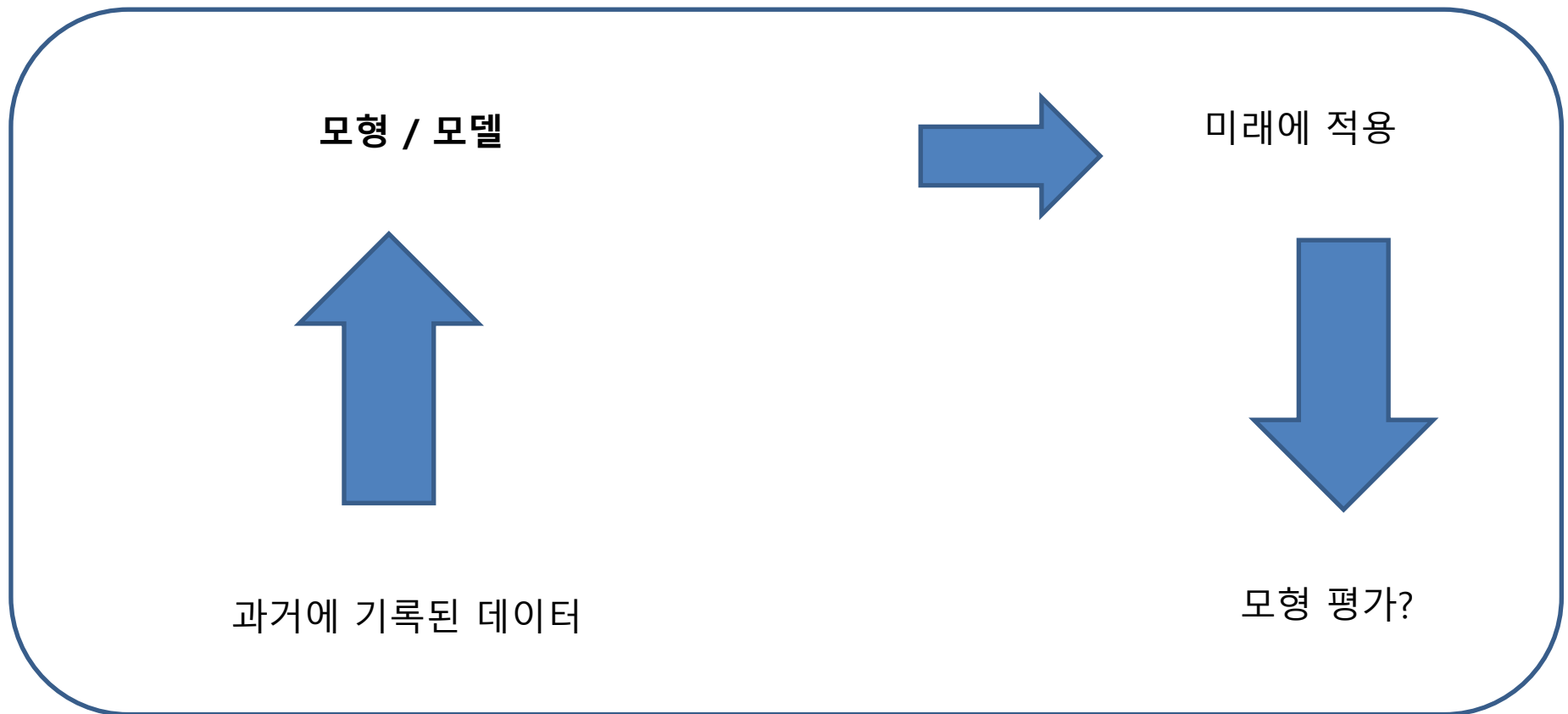
- 지도학습은 성능이 수치화되어 파악되지만, 비지도 학습은 분석 환경, 분석 목적 등을 고려해서 평가될 수 있음

- **모형 평가 시 주의점**

- 현재 평가는 앞으로의 성능에 대한 추정이므로 맹신하기 보다는, 향후 지속적인 모니터링과 업데이트를 기획해야 함

6. 분석 시 고려 사항

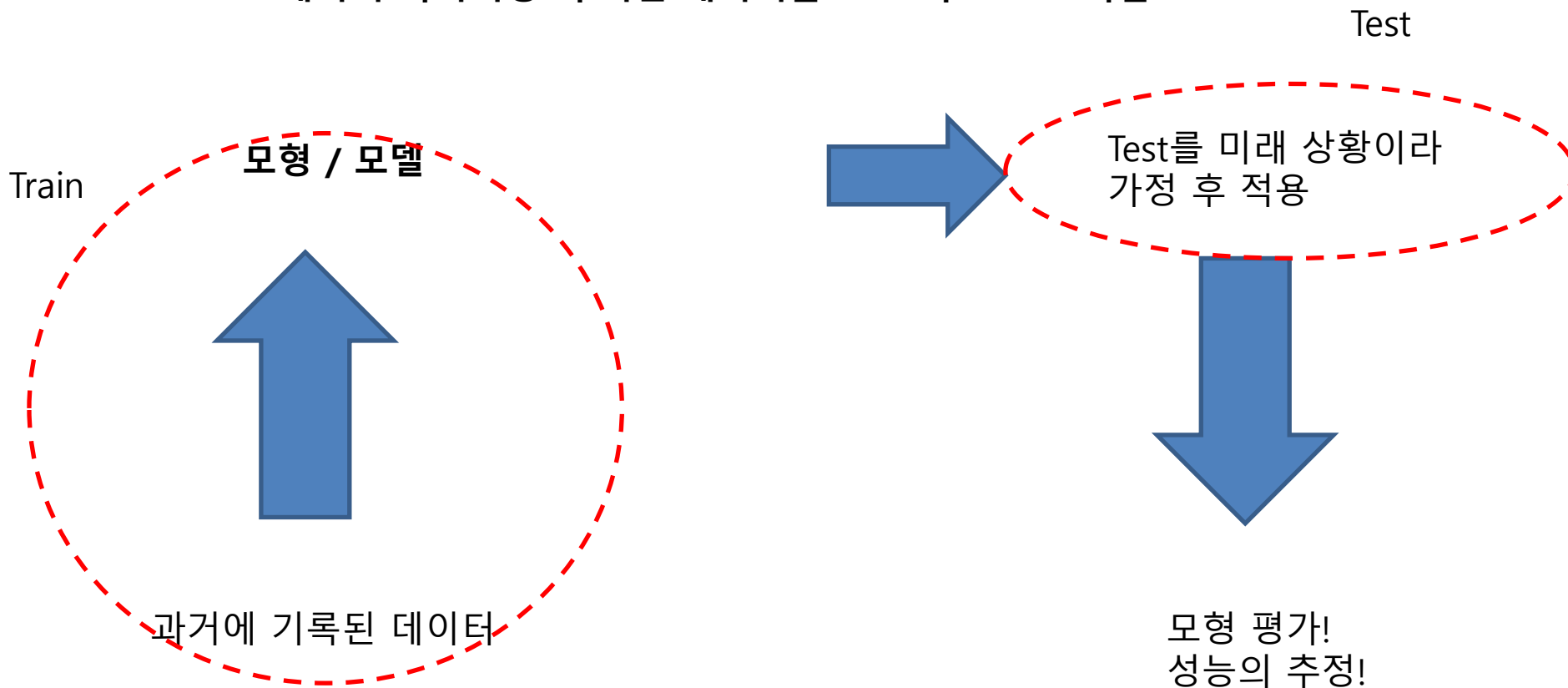
모형 평가



6. 분석 시 고려 사항

데이터 파티셔닝과 모형 평가

데이터 파티셔닝: 주어진 데이터를 Train과 Test로 나눔



6. 분석 시 고려 사항

지도학습 VS 비지도학습

주가 등락 여부에 대한 분류 모형의 성능 97%

비지도 학습 중 군집 분석 결과, 15000명의 고객에서 찾은 군집은 32개

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 분석 기획의 목적

- 단순히 데이터 분석을 수행하는 것을 넘어서서 비즈니스 상황을 이해하고 유기적인 기획과 분석의 수행을 통해 좋은 성과를 가져다 주는 것

- 비즈니스 아이디어의 도출

- 분석을 통해 얻은 인사이트는 비즈니스 모델의 고도화나 개선, 또는 새로운 비즈니스 모델의 제안으로 이어질 수 있음

7. 분석 기획과 비즈니스 아이디어

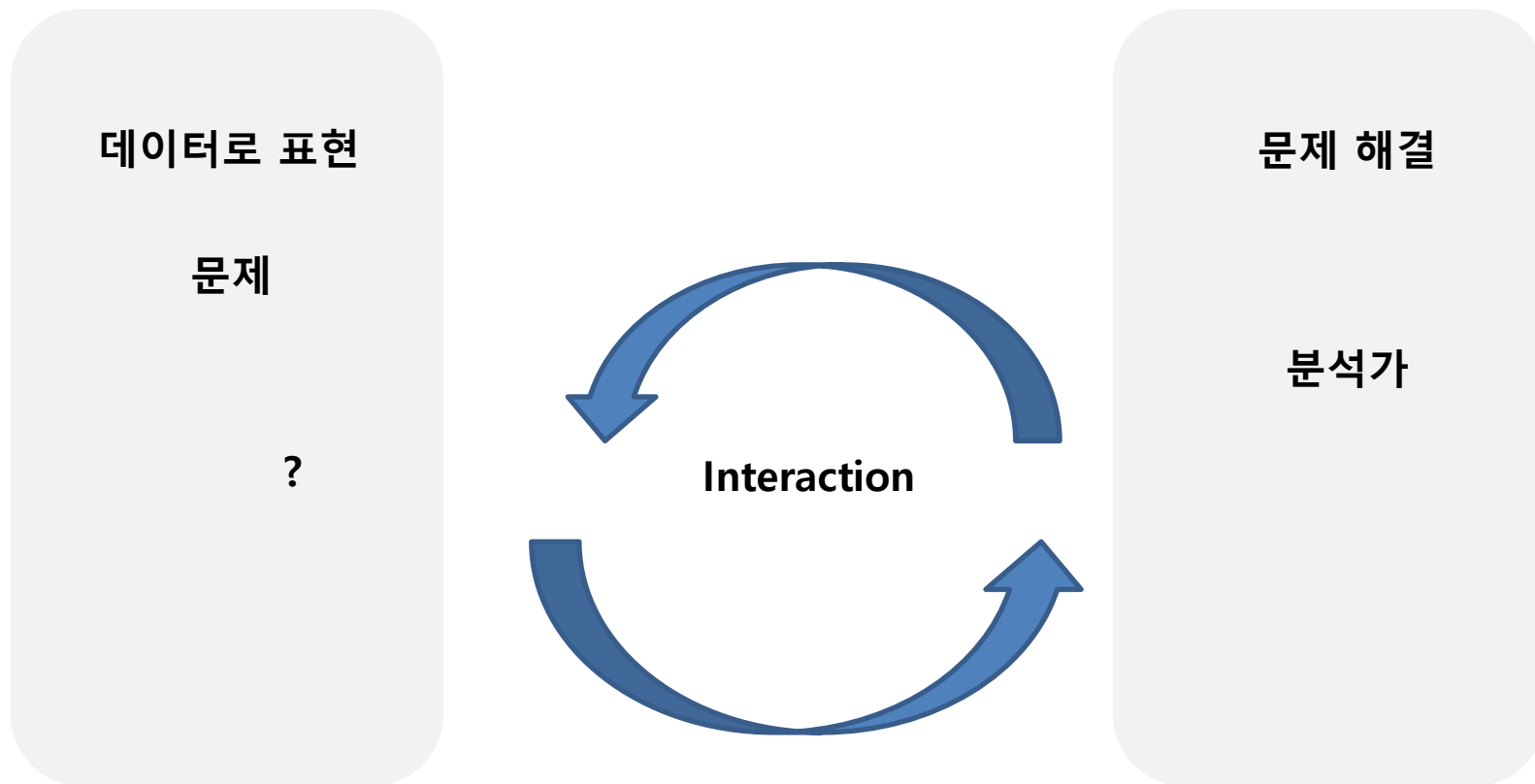
데이터 분석은 왜 하는 것일까?



데이터 분석 = 데이터로 표현된 문제 해결

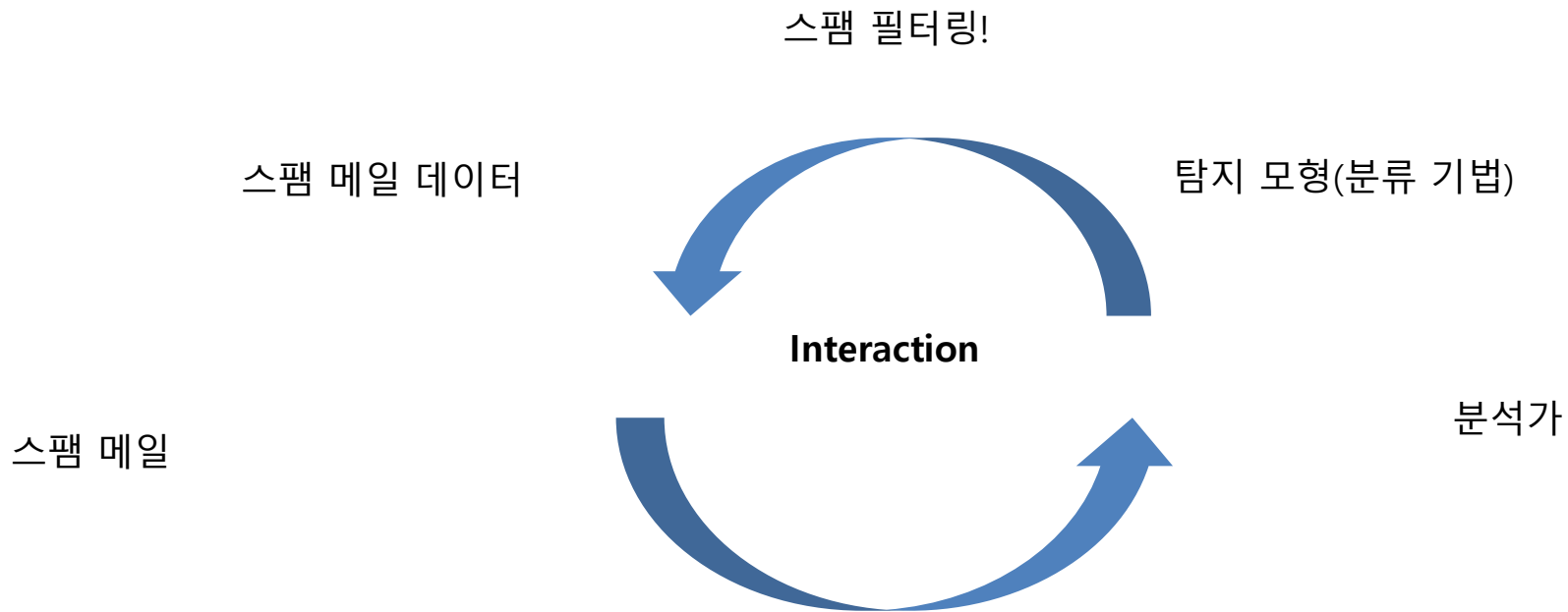
7. 분석 기획과 비즈니스 아이디어

데이터 분석 기획: 문제 상황과의 지속적인 Interaction



7. 분석 기획과 비즈니스 아이디어

분석 결과로 부터 비즈니스 아이디어 도출



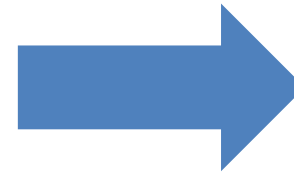
스팸 탐지 기능을 활용하는 비즈니스: 메일 서비스, 스팸 트렌드, 융합 보안 등

7. 분석 기획과 비즈니스 아이디어

- 현업 적용을 위한 Ideation
 - 현업 적용을 위해서는, 기술보다 현업에서의 Ideation이 가장 중요하며, 이를 위한 방법론이 필요
- Design Thinking
 - 기존과는 다른 창의적인 문제 해결을 위한 접근법으로 문제해결에 많이 활용
- Business Idea 발굴 시 주의점
 - Data로 표현될 수 있는 문제를 대상으로 AI 서비스를 기획하는 것이 중요

7. 분석 기획과 비즈니스 아이디어

AI 현업 적용 = Ideation이 핵심!



Ideation!



그런데, 현업 적용은?



7. 분석 기획과 비즈니스 아이디어

Design Thinking!

- 전통적인 문제해결이 아닌, 창의적인 접근 방법
- 미국 디자인업체 IDEO 창업자에 의한 구체화
- 디자인에 먼저 적용되었으며, 문제해결의 프레임워크로 발전

Design Thinking + 현업

Emphasize (공감하기)

Define(문제정의)

Ideate (아이디어 창출)

Prototype(프로토타이핑)

Test(테스트)

현장에서의 다양한 업무

도메인의 특수성이 고려된 기획

7. 분석 기획과 비즈니스 아이디어

- **Business Process**

- 특정 업무를 위한 활동이나 태스크의 구성, 그 관계에 대한 집합이며, 논리적으로 관련성 있는 작업 및 조직이 비즈니스 결과를 생성하기 위해 개발됨

- **Process Mining**

- 업무 시스템의 이벤트 로그로 부터 지식을 찾아내어 업무 절차를 발견하고 모니터링하고 개선하기 위한 분석

- **Process Data 기반 Ideation**

- Process mining의 결과에서 얻은 인사이트를 통해 업무 개선의 아이디어를 제안

7. 분석 기획과 비즈니스 아이디어

Process Mining

Process Innovation

Business Process Re-engineering

Business Process Management

데이터 기반 프로세스 마이닝!
-이벤트 로그, 시스템의 데이터들

Process Mining=

기업 데이터로 부터 프로세스에 대한 인사이트

프로세스에 대한 모니터링과 개선

- ERP 등 비즈니스 시스템의 데이터와 로그
- 업무 활용 채널의 이벤트 로그
- 구성원 간 인터랙션에 대한 데이터와 로그

7. 분석 기획과 비즈니스 아이디어

Process Mining!

문제 해결을 위한 Event Log를 처리

-Traffic

-Status

-Prediction

-AI + X

7. 분석 기획과 비즈니스 아이디어

Process Data의 관리

다양한 곳에 산재된 Process Data의 Sourcing 과 Management



7. 분석 기획과 비즈니스 아이디어

- AI 서비스 기획 단계
 - 아이디어 도출과 프로세스 분석을 통한 서비스 주제 구체화, 이후 관련 AI 기술의 매칭과 프로토타이핑
- 기술 주도 VS 기획 주도
 - AI 서비스를 바라보는 기술 위주 관점과 기획 위주 관점을 비교
- 작지만 큰 AI 서비스 기획
 - 현업의 작은 아이디어부터 AI 현업 적용 기획을 시작하여 점차 확대하는 접근 방식

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. 데이터 매니지먼트

- Data Management?



THE UNIVERSITY
of EDINBURGH

8. 데이터 매니지먼트

- 데이터 매니지먼트 정책

- 데이터 매니지먼트를 위한 기관의 프로세스 수립
- Johnston 외(2017)
 - 미네소타대학교, 미시간 대학교, 워싱턴 대학교, 일리노이 대학교, 코넬 대학교, 펜실베이니아 주립대학교에서의 연구데이터 관리 프로세스와 현황을 비교
- Grguric 외(2016)
 - 연구 과정에서 발생하는 데이터에 대한 관리 프로세스 중요성

- 의료 분야에서 사용되는 MRI 이미지들로 구성된 폴더일 수도 있고, 문서로 된 신약의 효능에 대한 내용, 종양 크기를 표현하는 엑셀 형태의 데이터이거나, 논문에 최종적으로 사용된 데이터일 수 있음
- 예를 들어 데이터 내 변수에 대한 이름과 설명 등과 실제 파일과 폴더 명이 일관되게 표현, 또한 레코드 마다 고유한 식별자를 통해 같은 개체의 다른 형식의 데이터도 연결될 수 있어야 하고, 이러한 일련의 과정은 연구 절차와 함께 잘 설명되어야 함

8. 데이터 매니지먼트

- 데이터 매니지먼트 계획(Data Management Plan)
 - 미네소타 대학교 연구데이터 가이드라인 화면

The screenshot shows a website titled "DATA MANAGEMENT PLAN EXAMPLES". On the left is a sidebar menu under the heading "MANAGING YOUR DATA" with links: Home, 1. Before Your Research, 2. During Your Research, 3. After Your Research Ends, Training and Workshops, and About Us. The main content area has a dark header with the title. Below the header, there is introductory text about writing a data management plan. To the right of this text is a "Need help?" section with a "Contact Us" link. The main content area is divided into three sections: "Education and Human Resources", "Health and Medical Science (Human Studies)", and "Physical Samples and Non-Digital Objects". Each section contains introductory text and a list of questions to address in a DMP. The "Physical Samples and Non-Digital Objects" section includes a list of example links: "Bone/wood lander studies example", "Cataloging insects example", and "Colorado School of Mines example". The bottom section is "Physical Sciences and Engineering", which includes a paragraph about links to data management plans for physical sciences and engineering research projects.

MANAGING YOUR DATA

- Home
- 1. Before Your Research
- 2. During Your Research
- 3. After Your Research Ends
- Training and Workshops
- About Us

DATA MANAGEMENT PLAN EXAMPLES

Not sure where to start writing your data management plan? Managing data in different disciplines can sometimes require very different strategies, standards, and considerations. Here are several examples of plans written across different disciplines to guide your own thinking.

Education and Human Resources

The NSF directorate lists [several context-specific questions](#) to consider when writing DMPs.

Health and Medical Science (Human Studies)

[NIH provides several examples of DMPs for studies involving human subjects.](#)

Physical Samples and Non-Digital Objects

Here are some questions to address in your DMP:

1. Are the samples already being stored by someone else? (e.g., Many DNA centers keep DNA samples indefinitely. Some samples may be in museum collections.)
2. Unambiguous identifiers for physical samples is important. Bar coding is a great option if available.
3. Photos can be a surrogate or to enhance the physical sample (e.g., colors fade in preserved fish).
4. Describe how the samples can be reused. (e.g., is destructive sampling allowed (DNA for instance always uses at least a little? Can the items be shipped or must the researcher be shipped (travel) to the sample?)
5. Do the samples degrade over time? If so, what's the lifespan of the objects.
6. If preserving/sharing samples is not possible, how will the researcher help others to replicate the sample?

- [Bone/wood lander studies example](#)
- [Cataloging insects example](#)
- [Colorado School of Mines example](#)

Physical Sciences and Engineering

The following links provide data management plans written for a variety of physical sciences and engineering research projects:

8. 데이터 매니지먼트

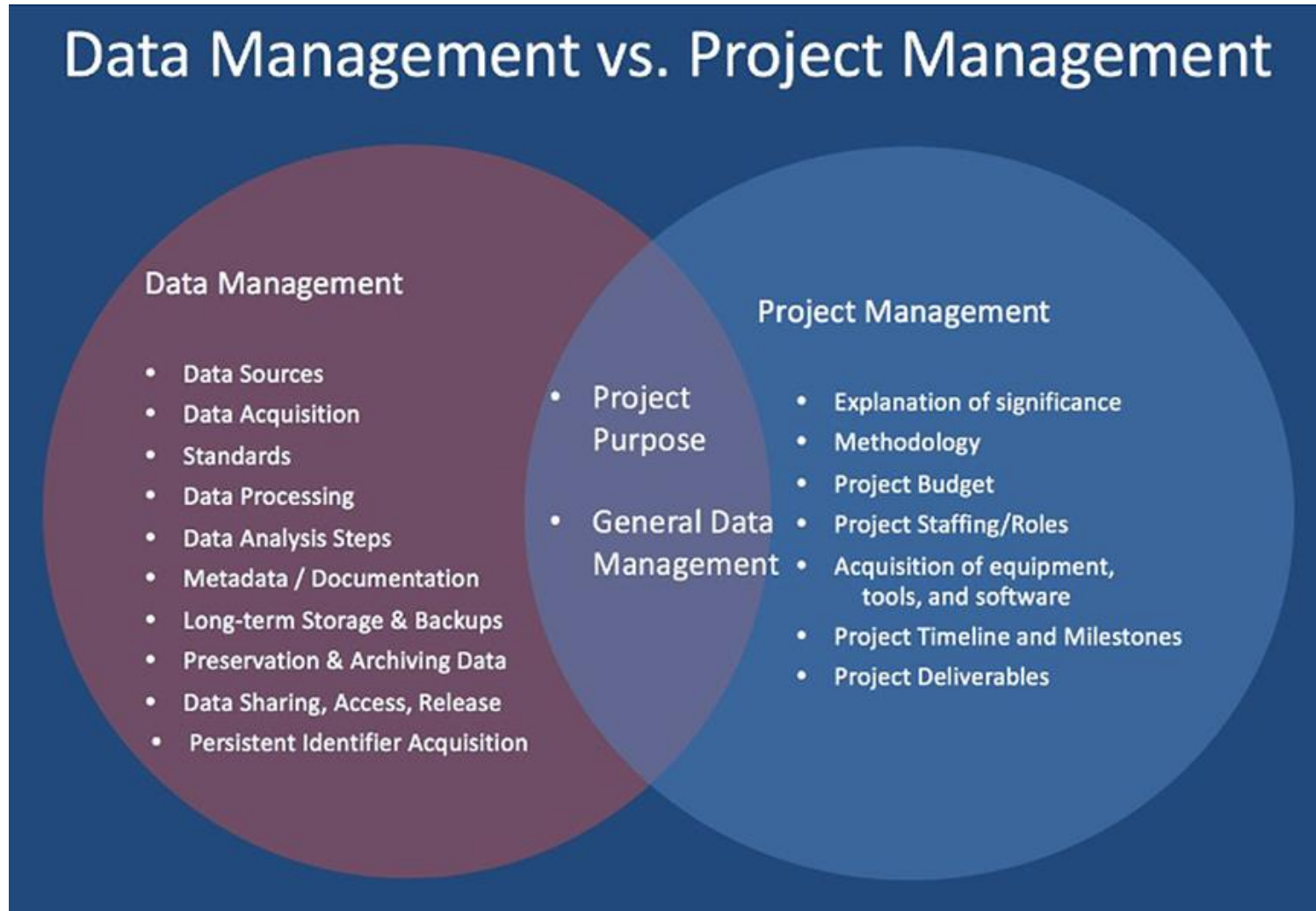
- 데이터 매니지먼트 계획(Data Management Plan)

프로젝트 중, 프로젝트 후의 데이터를 다루는 방안에 대한 계획, 체계적이고 안전한 데이터의 활용을 촉진

- ① Data Collection and Documentation
 - How are the data generated and which data are re-used?
 - How will the data be documented?
 - What metadata are needed to sufficiently describe and thus understand the data?
- ② Ethics, legal and security issues?
 - Are the data subject to personal rights or copyrights?
 - Are there other legal contracts that have to be respected?
 - Do the data have to be modified in a way (e.g. anonymization) that they can be shared?
- ③ Data Storage and Preservation
 - How and where are data stored?
 - How often are back ups performed and by whom?
- ④ Data sharing and re-use
 - How and where are data shared?
 - Who is allowed to access the data?
 - How are sensitive data protected?

8. 데이터 매니지먼트

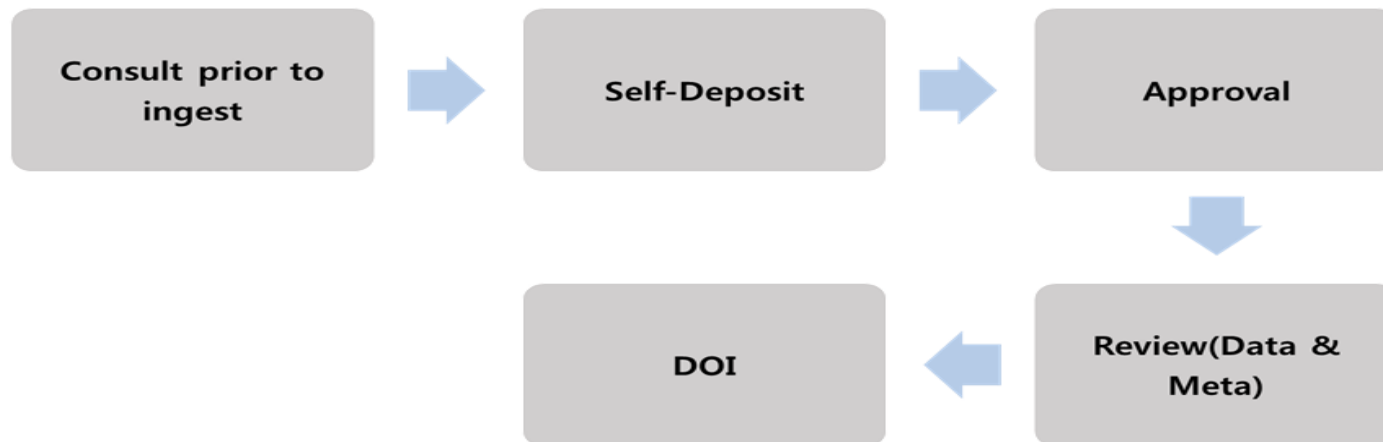
- 데이터 매니지먼트 계획(Data Management Plan)



8. 데이터 매니지먼트

- 프로세스

- ① 데이터 반입 전 자문
- ② Self Deposit
- ③ 자동 승인
- ④ 서비스
- ⑤ 데이터 및 메타데이터에 대한 리뷰
- ⑥ DOI 등 접근점 추가
- ⑦ 승인 후 데이터와 메타데이터가 적절한지 확인



The background features a glowing blue sphere composed of binary code (0s and 1s) against a dark blue space-like background with light rays. A dark horizontal band is positioned across the middle of the image.

2. IMRaD

1. IMRaD

- IMRaD? Scientific Communication!
 - Introduction
 - Method
 - Results
 - (and)
 - Discussions
- 그리고
- References

1. IMRaD

- *Introduction*
 - 무엇에 대한 연구/분석인지
 - 연구/분석의 배경이 있는지
 - 연구/분석이 왜 중요한지
 - 기대효과는 무엇인지

- (Optional) *Literature Review 또는 선행연구*
 - 이론적인 배경이 있다면 무엇인지
 - 이 주제와 관련되어 먼저 진행된 연구/분석들은 무엇이 있는지
 - 연구/분석의 이슈 도출과 구체화된 내용을 제시하고 있는지

1. IMRaD

- *Method : Data & Methodology*
 - 연구/분석에서 사용된 데이터에 대한 기술
 - 데이터의 출처 및 해당 데이터를 사용한 기존 분석
 - 데이터의 소개: 크기, 변수, 변수에 대한 설명 등
 - 데이터에 대한 요약 (Descriptive Statistics)
 - 연구/분석에서 사용된 방법론(모형, 기법 등)에 대한 소개
 - 사용한 방법론에 대한 소개 및 Reference
 - 왜 이 방법론을 사용해야 했는지 (연구/분석 주제와 연관되도록 제시)

1. IMRaD

- *Results*

- 연구/분석의 결과에 대한 기술
 - 모형/기법에 사용된 하이퍼 파라미터 / 컴퓨팅 환경 / 소요기간 등
 - 분석의 결과만 기술
 - 가능한 경우 도식화

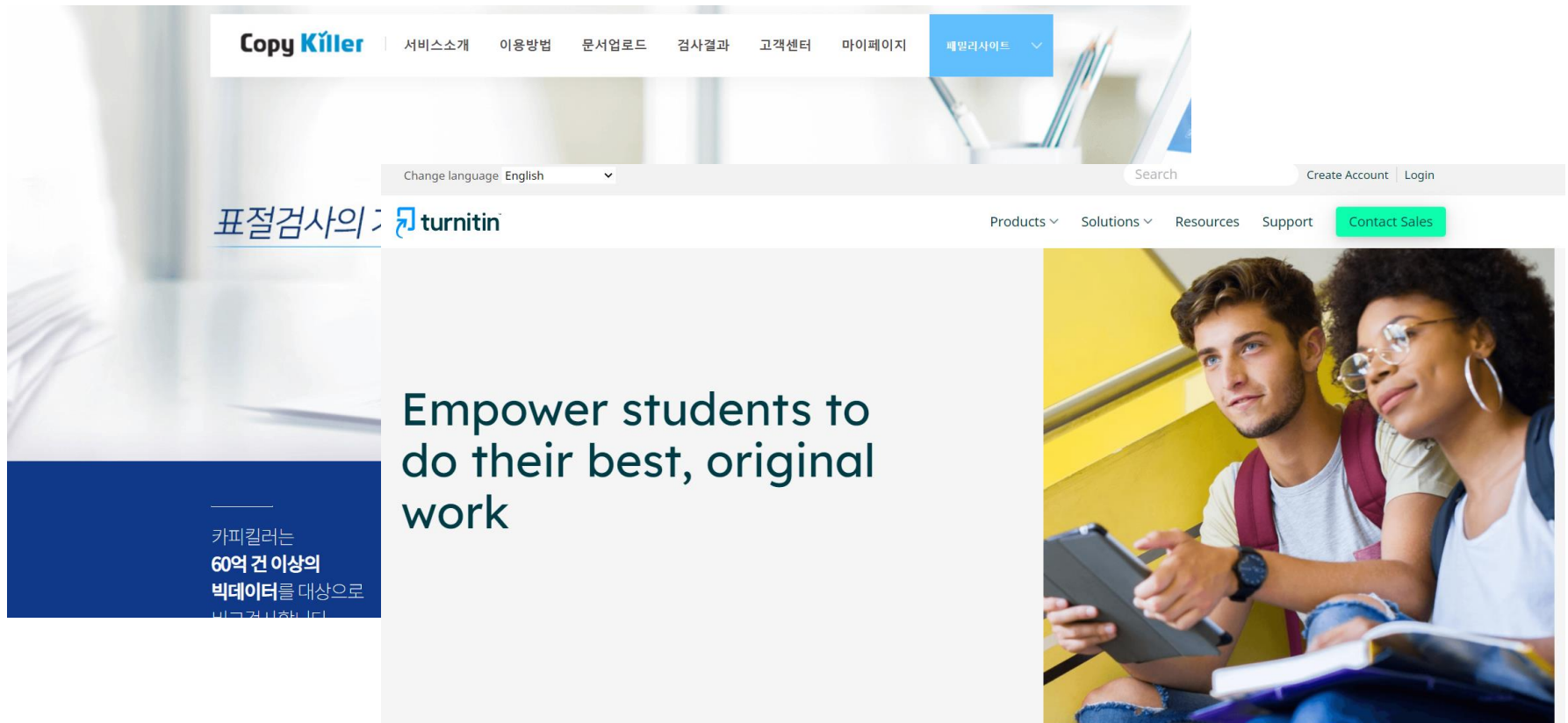
1. IMRaD

- *Discussions*
 - 연구/분석의 결과에 대한 논의
 - 앞서 제기한 연구/분석의 이슈 관점에서 결과 논의
 - 모형 성능에 대한 언급
 - 연구/분석의 기여점과 한계 제시
 - Business/Policy Implications의 제시

1. IMRaD

- References

- 연구/분석에 사용된 기존의 연구/분석들을 언급: 표절(Plagiarism) 방지!



1. IMRaD

- **References**

- 연구/분석에 사용된 컨퍼런스/논문/Technical Report 등의 문헌들

- APA 스타일: APA(American Psychological Association, 미국 심리학회) Style은 사회과학 분야에서 일반적으로는 많이 사용

[책] Deep learning

J.D Kelleher - 2019 - books.google.com

An accessible introduction to the artificial intelligence technology that enables computer vision, speech recognition, machine translation, and driverless cars. Deep learning is an artificial intelligence technology that enables computer vision, speech recognition in mobile ...

☆ 저장 인용 117회 인용 관련 학술자료 전체 6개의 버전

Deep learning

N Rusk - Nature Methods, 2016 - nature.com

and high computational costs are being tackled as in startup companies such as Deep Genomics authors of DeepBind, will increasingly apply deep learning to other domains.

☆ 저장 인용 107회 인용 관련 학술자료 전체 1개의 버전

X 인용

MLA	Rusk, Nicole. "Deep learning." <i>Nature Methods</i> 13.1 (2016): 35-35.
APA	Rusk, N. (2016). Deep learning. <i>Nature Methods</i> , 13(1), 35-35.
ISO 690	RUSK, Nicole. Deep learning. <i>Nature Methods</i> , 2016, 13.1: 35-35.

BibTeX EndNote RefMan RefWorks

[HTML] **Deep learning in agriculture**

A Kamilaris, FX Prenafeta-Boldú - Computers & Electronics Agriculture, 2016 - Elsevier

Deep learning constitutes a recent, modern paradigm in machine learning analysis, with promising results and large potential. In the past few years it has been applied in various domains, it has recently emerged as one of the most powerful tools available.

☆ 저장 인용 1412회 인용 관련 학술자료 전체 2개의 버전

Deep learning: yesterday, today, and tomorrow

K Yu, L Jia, Y Chen, W Xu - Journal of computer research and applications, 2013 - crad.ict.ac.cn

Machine **learning** is an important area of artificial intelligence. Since 1980s, huge success has been achieved in terms of algorithms, theory, and applications. From 2006, a new machine **learning** paradigm, named **deep learning**, has been popular in the research ...

☆ 저장 인용 208회 인용 관련 학술자료 전체 2개의 버전

1. IMRaD

- References

- 본문 슬라이드 중 해당하는 내용에 직접 레퍼런스를 표시
- 표시 예:
 - 홍길동(2021)
 - 홍길동과 김철수(2021)
 - 홍길동 외 (2021)
 - 등에 사용되었음 (홍길동, 2021)
 - ... 등의 연구가 있음 (홍길동, 2021; 김철수 외, 2020)

2. Reproducibility & Implications

- “Reproducibility” : 재현가능성

the original researcher's data and computer codes are used to regenerate the results

2. Reproducibility & Implications

- “Reproducibility” *for Open Science*
 - 연구/분석에 대한 신뢰도
 - 연구/분석에 대한 동료들의 이해 및 개선
 - 소속 기관 및 참여 커뮤니티에 기여

2. Reproducibility & Implications

- “Reproducibility”
 - IMRaD의 M과 관련!
 - 사용된 코드와 데이터를 공유
 - 코드에 주석을 통해 작성자/소속/일시 등에 대한 메모
 - 데이터의 경우 출처 및 사용 가능성에 대한 확인 필요

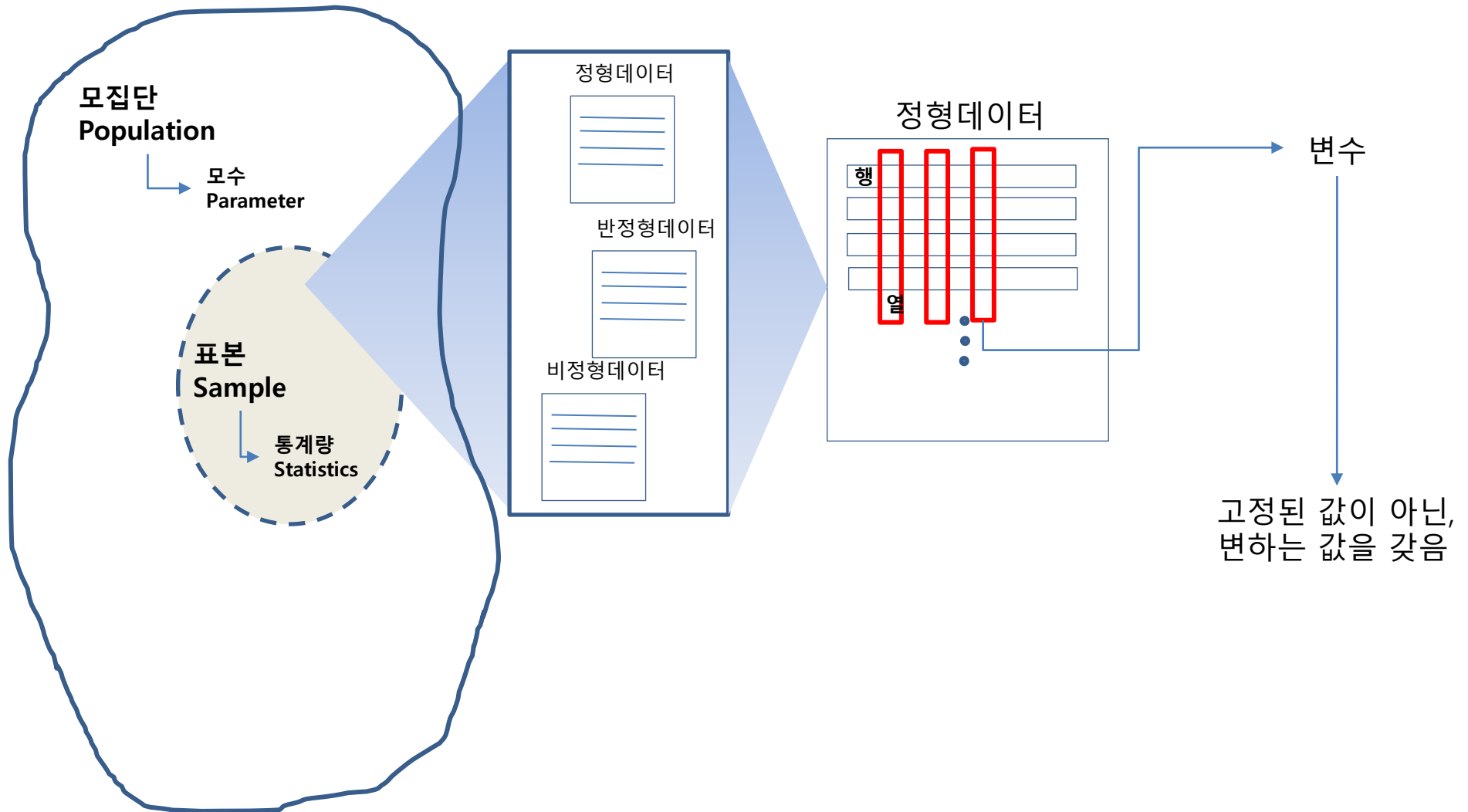
2. Reproducibility & Implications

- 연구/분석의 주제와 결과에 대한 해석과 활용
 - 연구/분석의 결과를 활용
 - 연구/분석의 결과의 필요성에 대한 리뷰

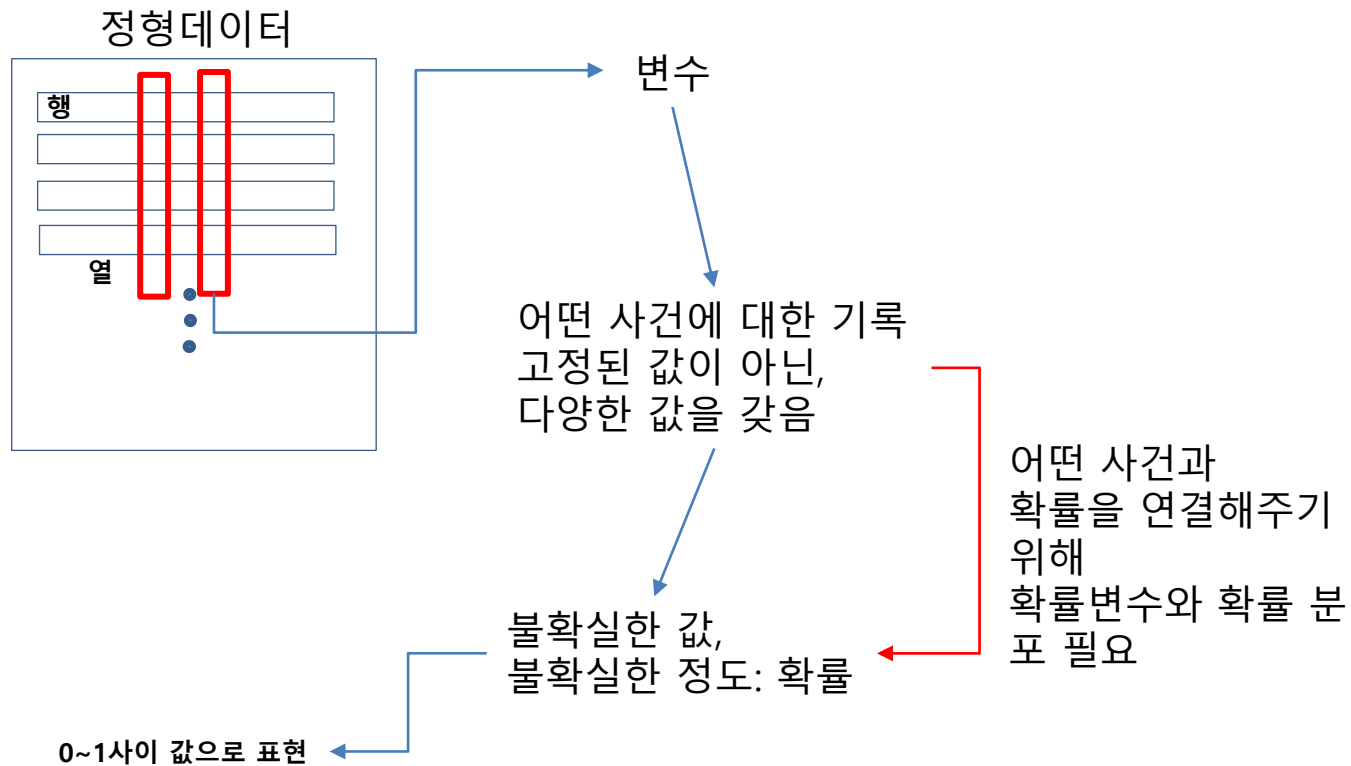
*“연구/분석의 결과가 회사의 0000 문제나 프로젝트에 대해
기여하고, 해결책을 제시할 수 있는지!”*

3. 데이터사이언스 기본 리뷰

1. 데이터부터 회귀분석까지



1. 데이터부터 회귀분석까지

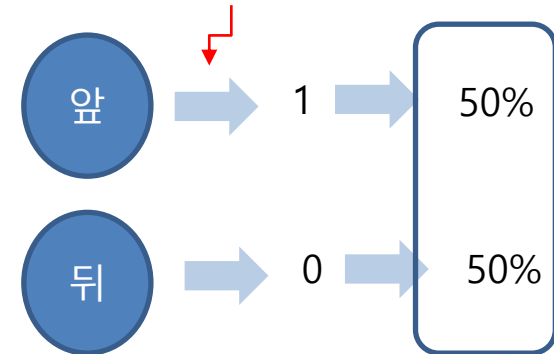


예를 들어, 동전을 던지면,



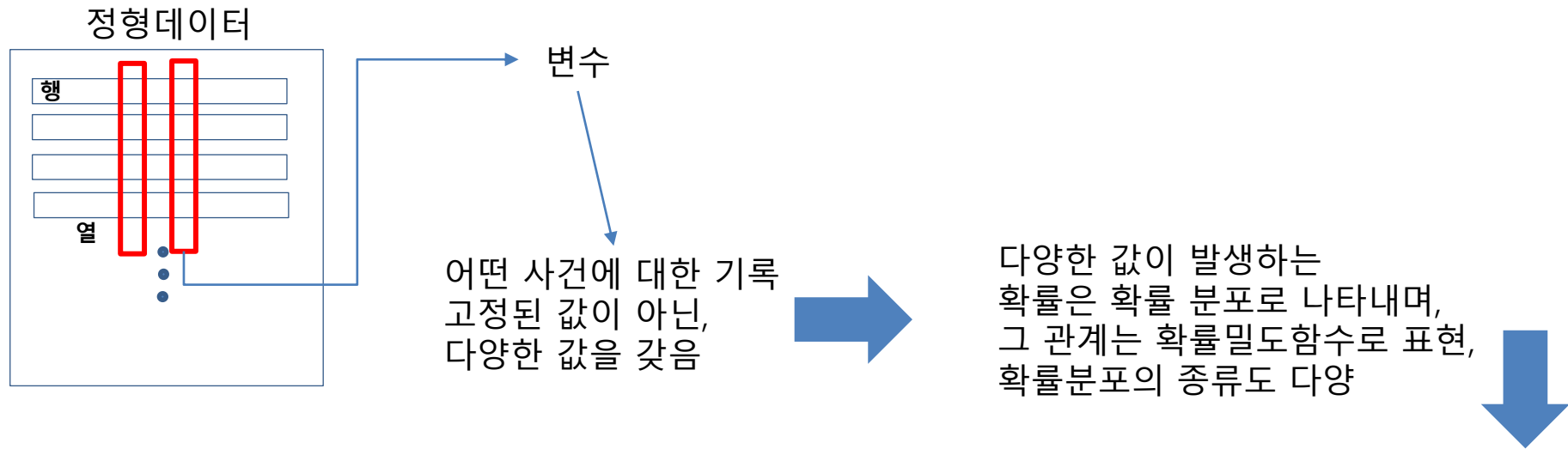
앞과 뒤를 숫자로 표현하면
다루기 편해지는...

앞면이 나온 횟수를 세어
숫자로 표현하는 대응 규칙을
고려할 수 있음: 확률변수



이때 각 경우에 대한 확률을
표시해줌: 확률 분포

1. 데이터부터 회귀분석까지



이산형 확률분포: 어떤 사건이 갖는 값이 셀 수 있는 경우, 이항분포, 포아송 분포

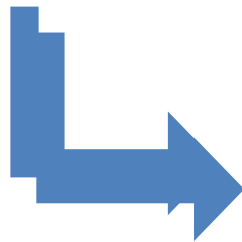
연속형 확률분포: 어떤 사건이 갖는 값이 셀 수 없는 경우, 정규분포, 표준 정규분포

1. 데이터부터 회귀분석까지

다양한 값이 발생하는
확률은 확률 분포로 나타내며,
그 관계는 확률밀도함수로 표현,
확률분포의 종류도 다양



이 자료를 효과적으로
이해하려면: 요약이 필요



통계량
통계량

집중화경향
어디에 값이
주로 몰려있는지...
:mean, median

산포도
평균을 중심으로
값이 얼마나 퍼져
있는지...
:분산, 표준편차

표본분포

-모집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때 그 표본의 특성치(통계량)의 확률 분포

중심극한정리

- 표본을 뽑았을 때, n 이 충분히 크다면 모집단의 분포모양에 관계없이 표본평균 \bar{X} 는 근사적으로 정규분포

t - 분포

-서로 다른 두 집단의 평균의 통계 검정

χ^2 - 분포

-서로 다른 2개 이상 집단의 비율의 통계 검정

F - 분포

-서로 다른 2개 이상 집단의 분산의 균질성 검증

1. 데이터부터 회귀분석까지



과연 모집단의
특성인 모수를 잘 나타낼까?
확인하는 방법은?...

통계량으로 모수를 추정하는 것을 통계적 추론이라 하며,
가설검정을 통해 할 수 있음

정규성 검정

귀무가설(H_0) : 정규분포를 따른다.

대립가설(H_1) : 정규분포를 따르지 않는다.

t 검정

귀무가설(H_0) : $\mu_1 = \mu_2$ (두 모집단의 평균은 같다.)

대립가설(H_1) : $\mu_1 \neq \mu_2$ (두 모집단의 평균은 다르다.)

Paired t 검정

귀무가설(H_0) : $\delta = 0$ (두 모집단의 평균은 같다.)

대립가설(H_1) : $\delta \neq 0$ (두 모집단의 평균은 다르다.)

F 검정

귀무가설(H_0) : (두 모집단의 산포는 같다.)

대립가설(H_1) : (두 모집단의 산포는 다르다.)

카이제곱 검정

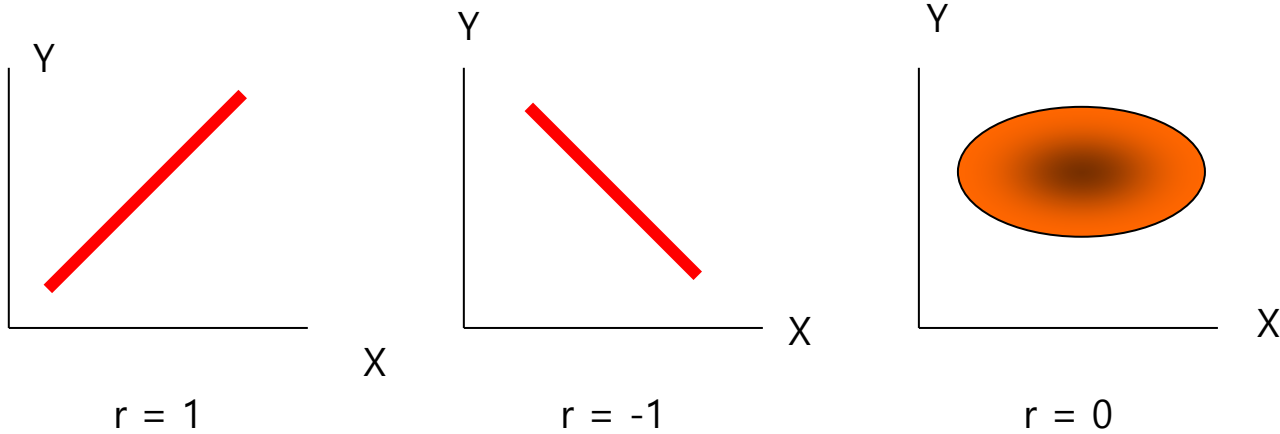
귀무가설(H_0) : 두 모집단은 독립적이다.

대립가설(H_1) : 두 모집단은 종속적이다.

- | | |
|----------|---|
| •Step #1 | 가설 설정
귀무가설(H_0)과 대립가설(H_1)을 세운다. |
| •Step #2 | 유의 수준(α) 결정 |
| •Step #3 | P-Value 산출 |
| •Step #4 | 귀무가설(H_0)의 기각 여부 결정
If P-Value < 유의수준(α)이면, 귀무가설 (H_0) 기각 |

1. 데이터부터 회귀분석까지

Review:상관분석(Correlation Analysis)



Y와 X의 관계가 있음을 알았는데, 과연 둘의 인과관계는?

=> X로 인해 Y는 어떤 영향을 받을까?

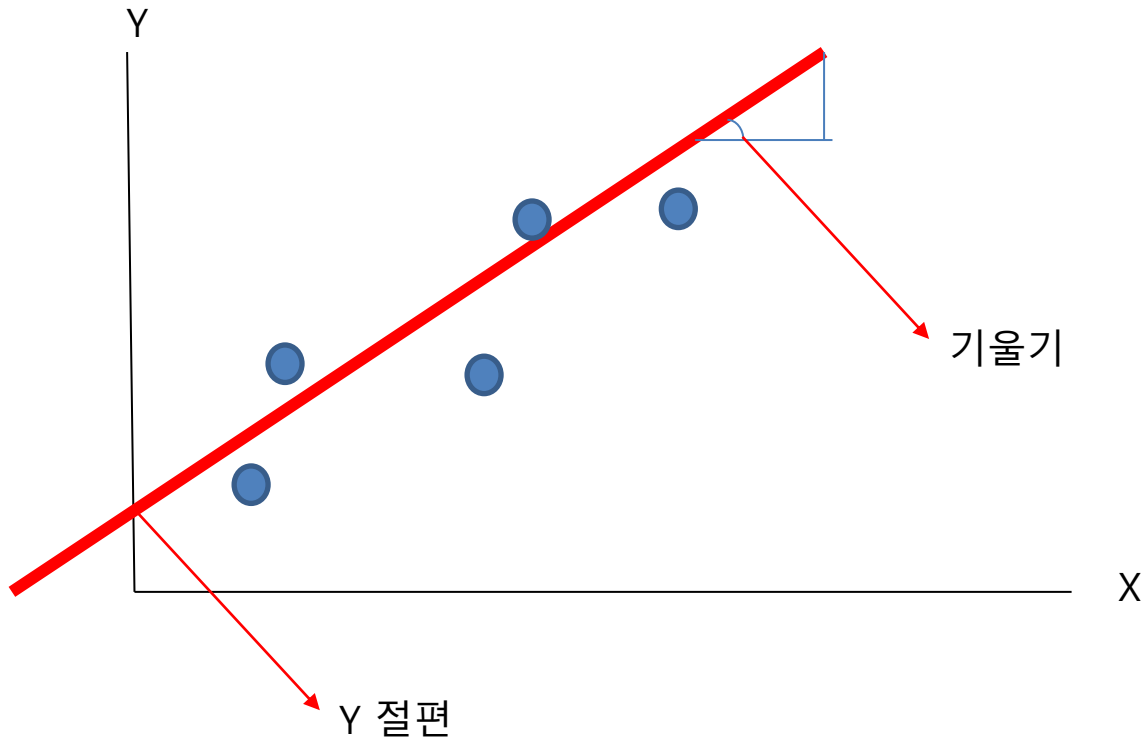
?

Y: 종속변수
X: 독립변수



1. 데이터부터 회귀분석까지

주어진 X와 Y의 값을 그려본다면, 아래처럼 표시됨



X와 Y 각각의 값은 좌표가 되어 점으로 표시

여러 개의 점을 한 번에 설명할 수 있는
방법이 필요

-효율적이지만, 아주 정확하지는 않음

점들을 가장 잘 나타내는
하나의 직선으로 표시해 보기!

직선은 기울기와 Y절편만 있으면 그릴 수 있음

더 나아가 X와 Y의 관계도 표시

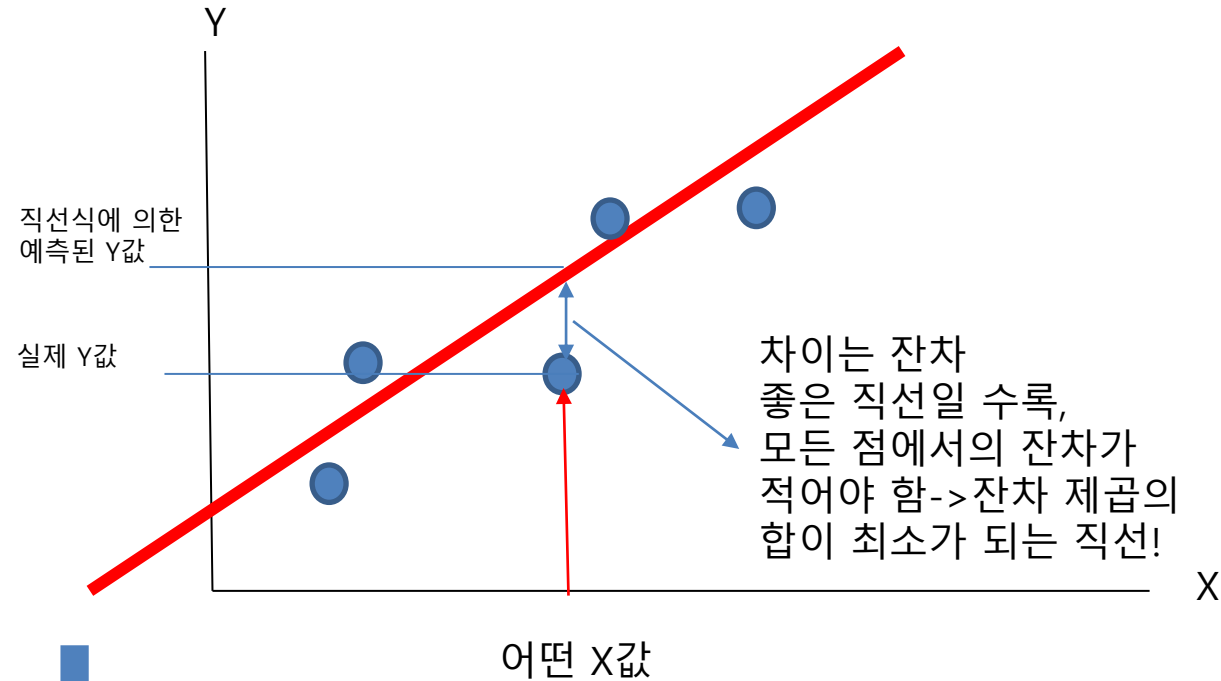
$Y = \text{기울기} \times X + Y \text{ 절편}$

선형회귀분석

참고로 비선형회귀분석도 있고, 다른 방식의
회귀분석도 많이 있음.

1. 데이터부터 회귀분석까지

선형회귀분석



이 모형이 자료 전체를 잘 설명하는지는
결정계수로 파악
결정계수: 0~1사이 값, 선택한 X변수가 Y 변수의
잔차제곱의 합을 잘 설명해주는 정도를 알려줌

선형회귀분석

기울기의 해석: X가 1단위 증가 시
Y의 변화

모집단에도 우리가 사용한 X, Y에 상응하는
값들이 있고 모집단에서의 기울기가 있음

그렇지만, 우리는 주어진 자료(표본)로만
기울기를 알아내야 함....

기울기를 추정해야 하는 문제가 되며
통계량이 모수를 잘 나타내는지를 알기 위해
가설검정을 함

예:

추정된 기울기가 0.5
 H_0 : 기울기=0
 H_1 : 기울기!=0

이때 이 기울기의 p-value는 0.01이라면,
귀무가설이 기각되어, 추정된 기울기는
통계적으로 유의!

Q&A

