

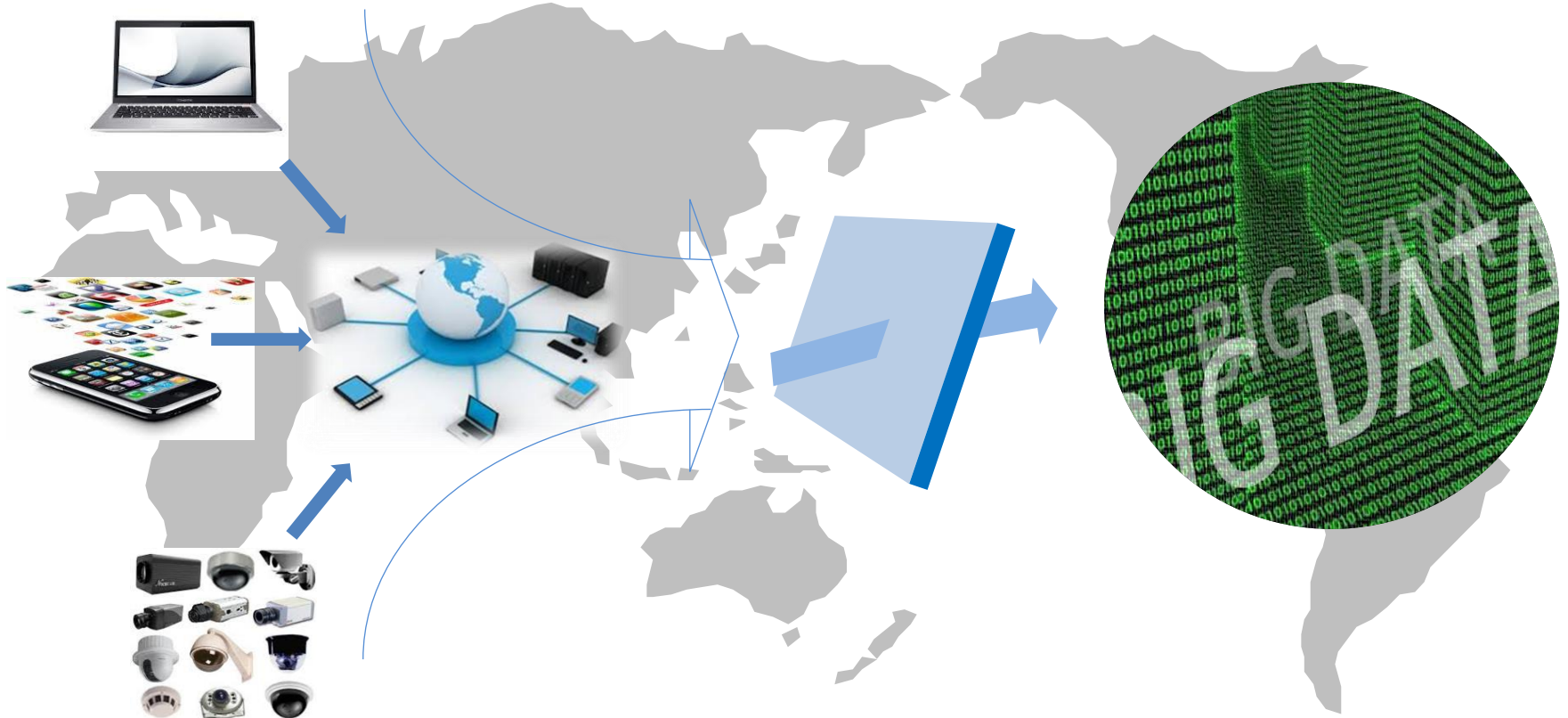


# 1. 데이터 분석 기획

2021

# 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

# 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 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

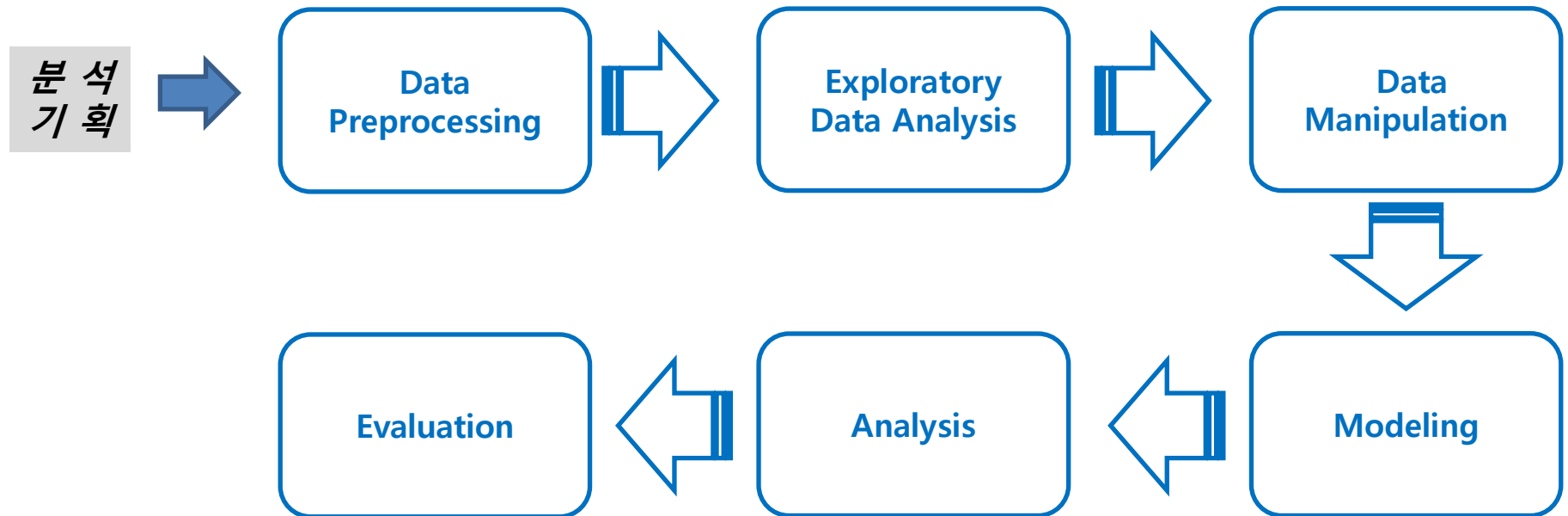
- SEMMA: Sampling Exploration Modification Modeling Assessment

Sample-Exploration-Modification-모델링-평가



# 1. 빅데이터와 분석 기획

---



# 1. 빅데이터와 분석 기획

---

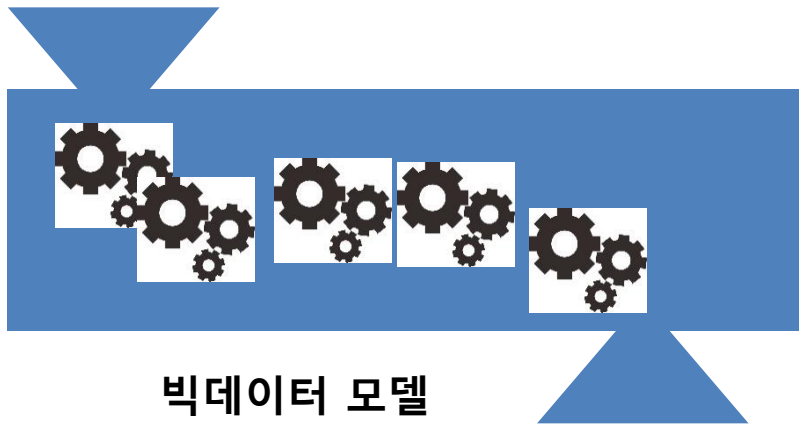
## 데이터 분석 기획

- 분석 과제
- 데이터
- 어떤 기법
- 기대효과
- 업데이트는 어떻게!

# 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 데이터로부터 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

## 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. 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, 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”

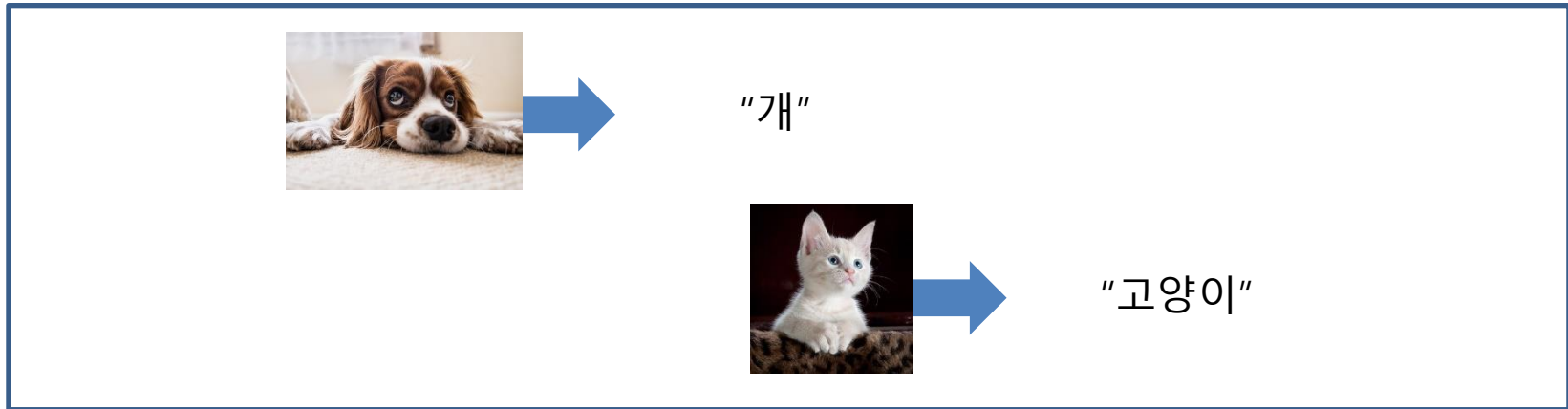


“고양이”

### 3. 분석 기획을 위한 데이터 이해

---

예



"한 데이터로 여러 분석을 할 수 있음!"

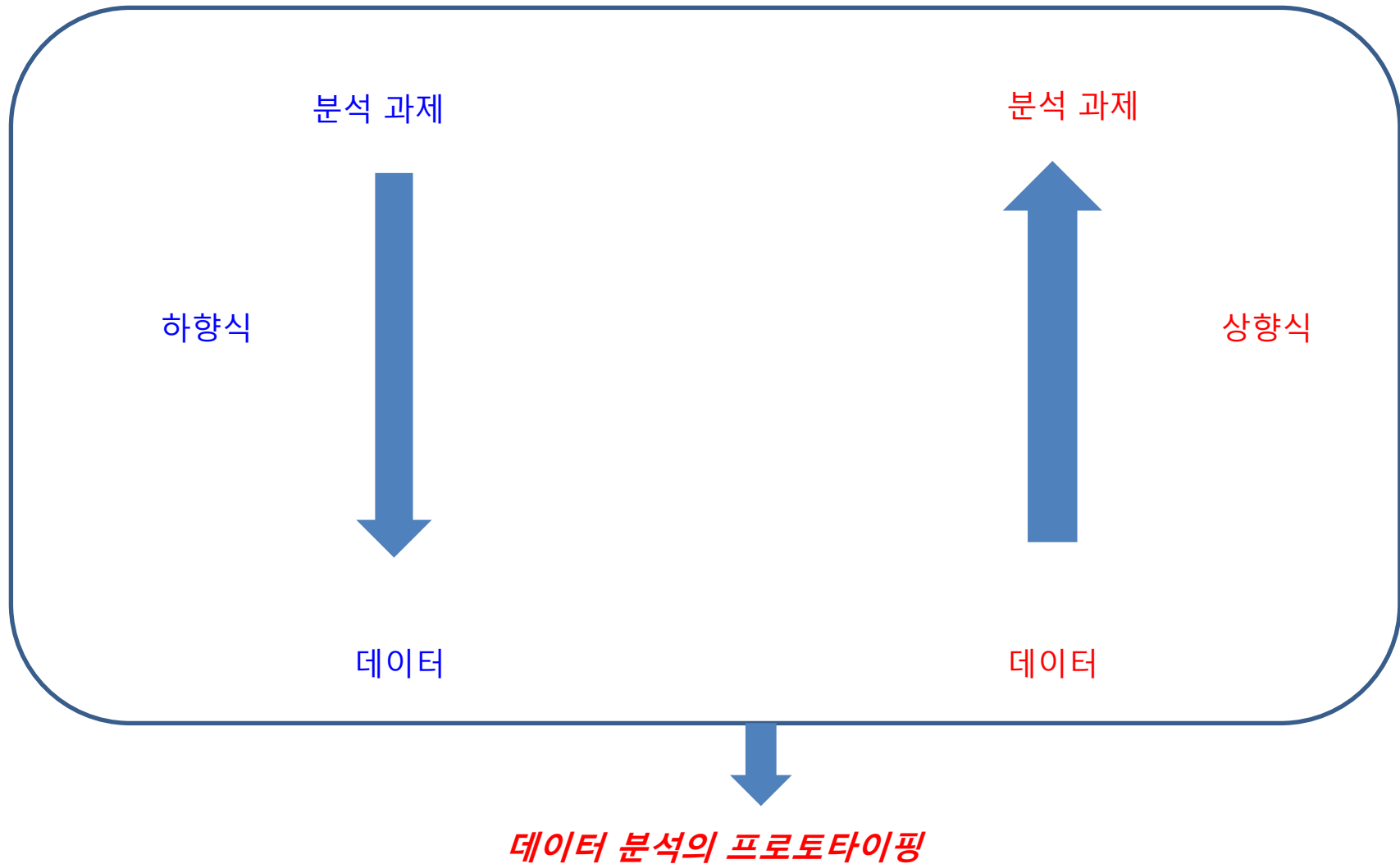
동물 분류 딥러닝!

홈 CCTV 비즈니스에 응용!

### 3. 분석 기획을 위한 데이터 이해

---

#### 하향식 VS 상향식



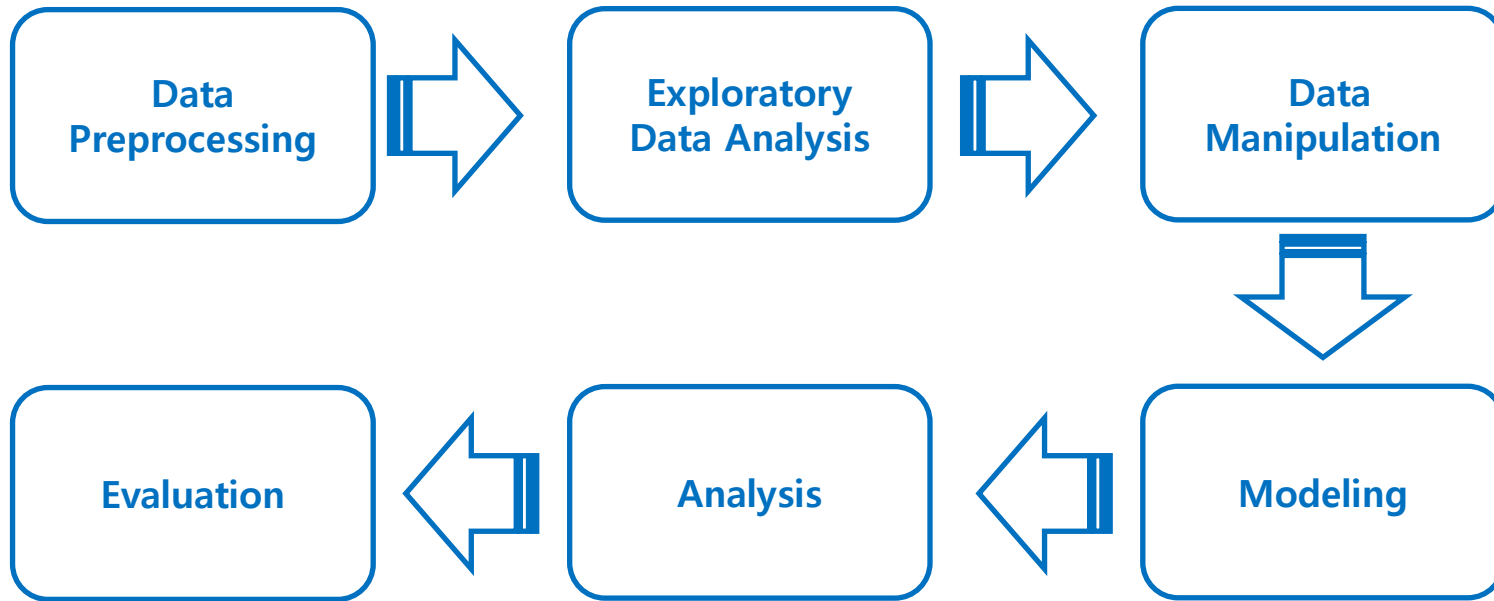
## 4. 분석 절차 1

---

- 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

---

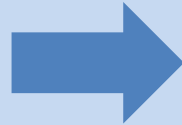


## 4. 분석 절차 1

---

### 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 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

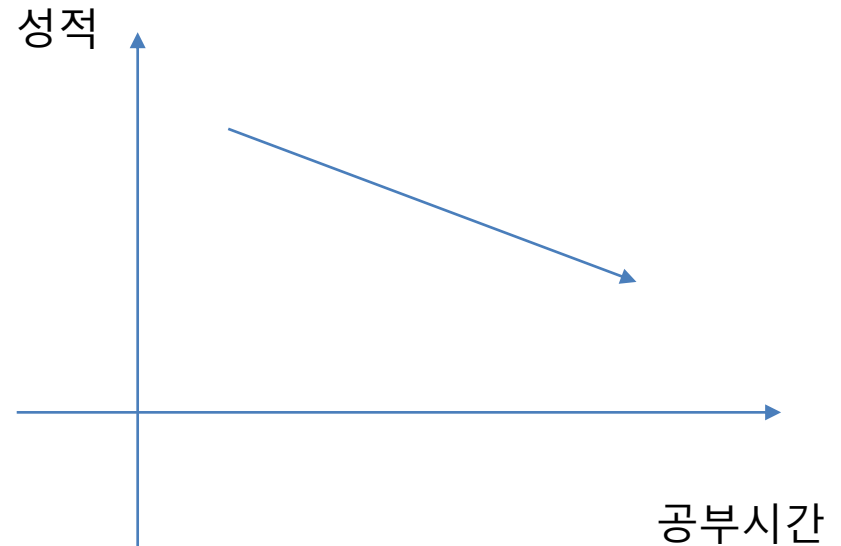
### Exploratory Data Analysis(EDA)

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

---

- **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

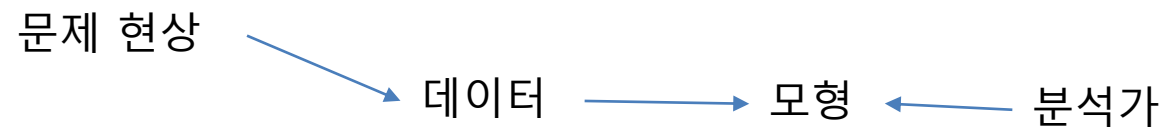
## 5. 분석 절차 2



### 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

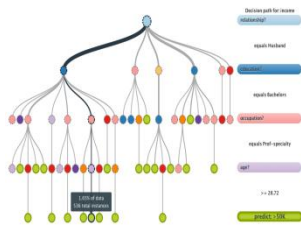


#### 비지도학습 (Unsupervised Learning)

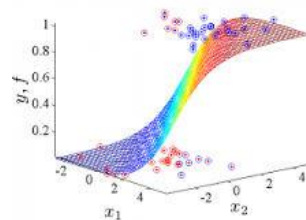
Target(종속변수/반응변수)이 없으며, 독립(설명)변수 간의 관계나 이를 바탕으로 개체들을 구분하여 의미 있는 결과를 제시

예: 군집 분석, 연관성 분석, 주성분 / 요인 분석

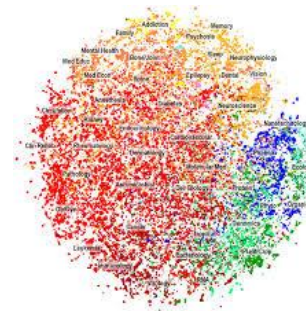
[decision tree]



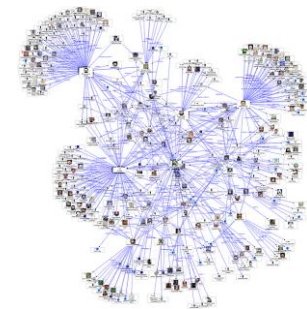
[logistic regression]



[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

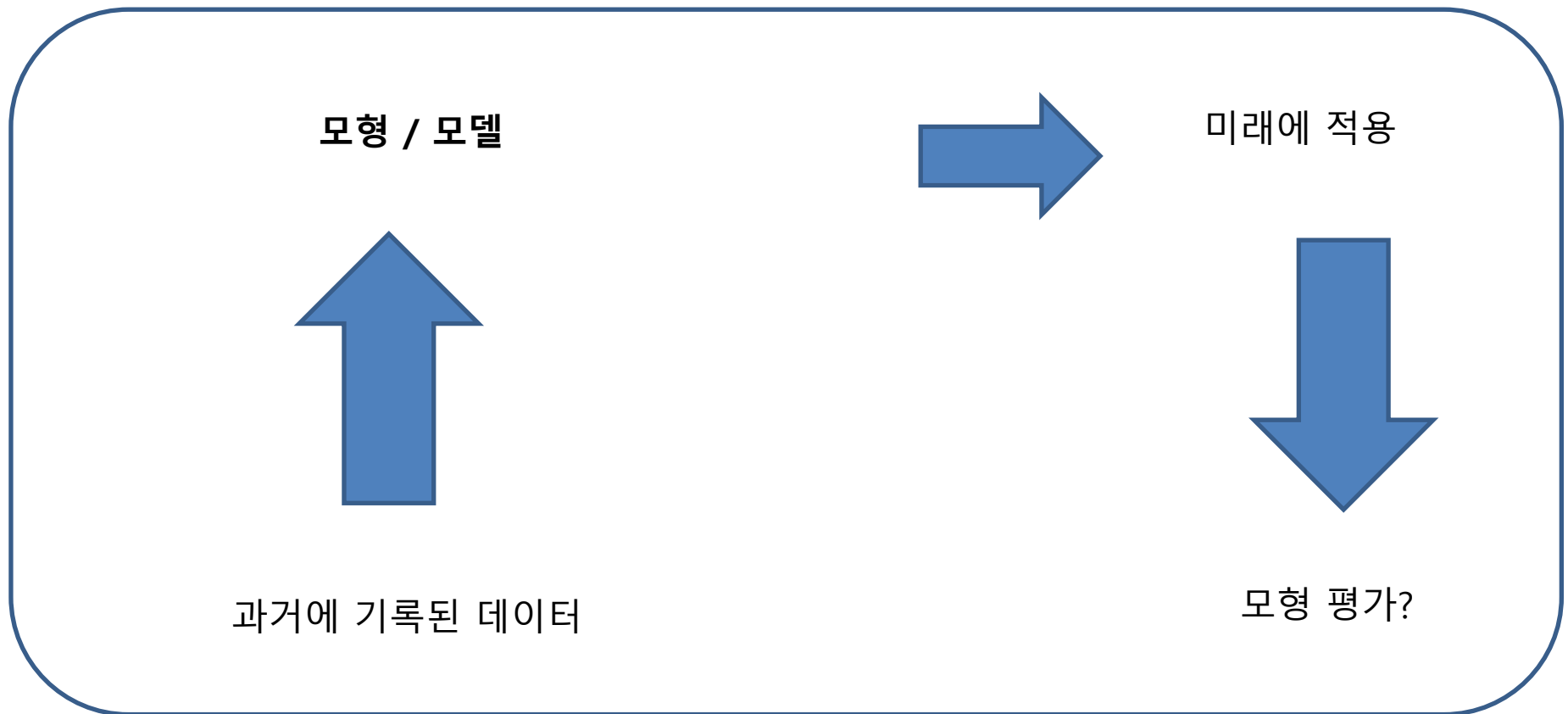
---

- 분석 모형 평가를 위한 데이터 파티셔닝: 주어진 데이터를 Train 데이터와 Test 데이터로 나눠, 모델링 결과에 Test 데이터를 적용해 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

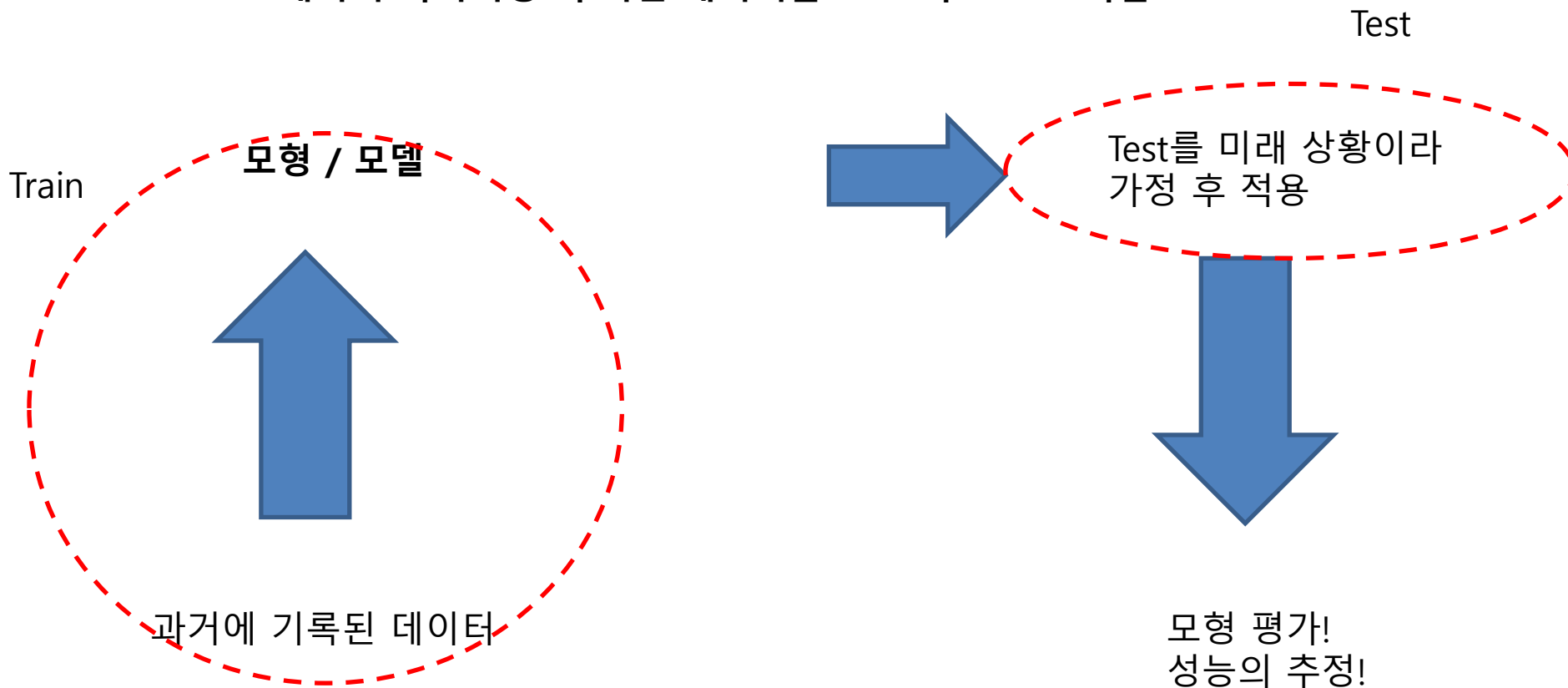


## 6. 분석 시 고려 사항

---

### 데이터 파티셔닝과 모형 평가

데이터 파티셔닝: 주어진 데이터를 Train과 Test로 나눔



## 6. 분석 시 고려 사항

---

### 지도학습 VS 비지도학습

*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?

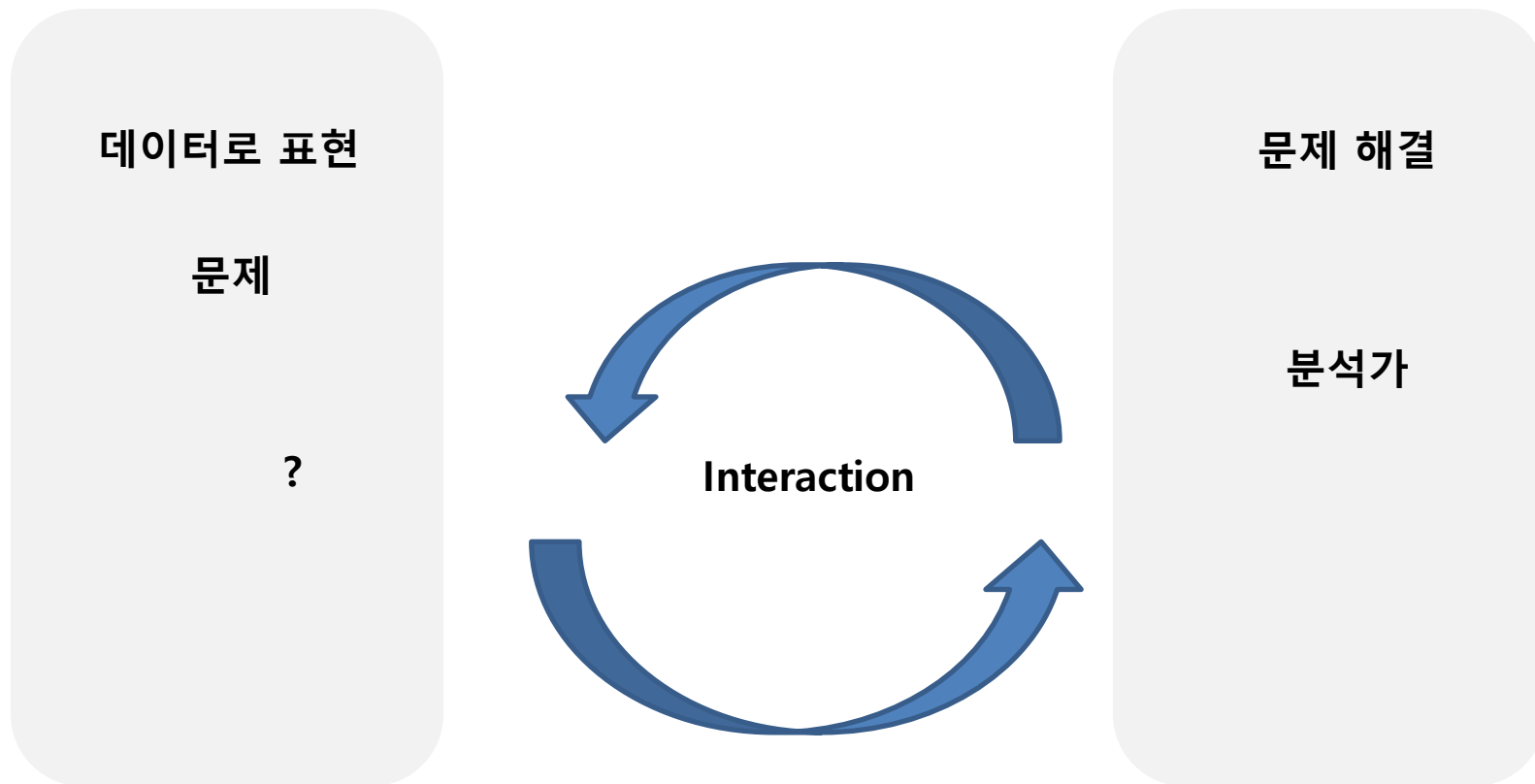


데이터 분석 = 데이터로 표현된 문제 해결

## 7. 분석 기획과 비즈니스 아이디어

---

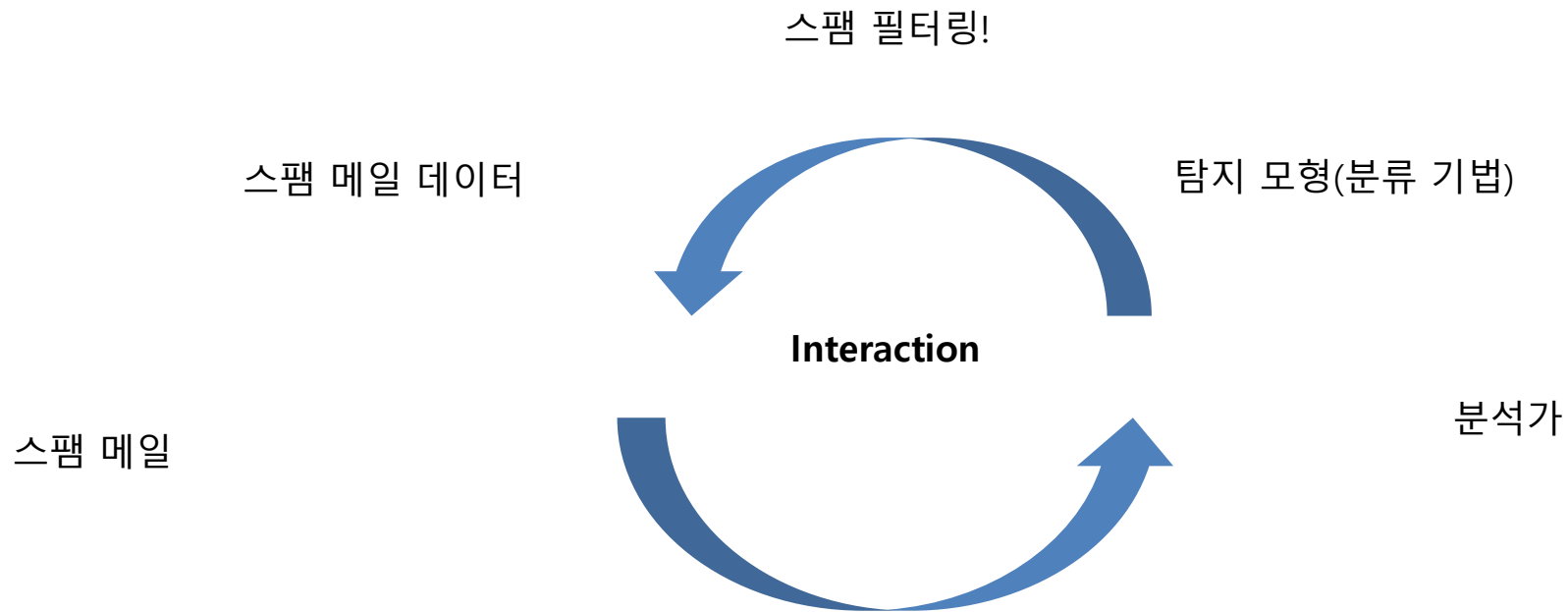
데이터 분석 기획: 문제 상황과의 지속적인 Interaction



## 7. 분석 기획과 비즈니스 아이디어

---

분석 결과로 부터 비즈니스 아이디어 도출



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

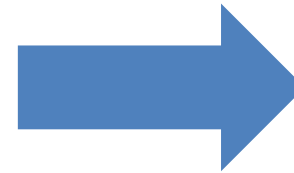
## 7. 분석 기획과 비즈니스 아이디어

---

- 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 레코드마다 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 is organized into three sections: "Education and Human Resources", "Health and Medical Science (Human Studies)", and "Physical Samples and Non-Digital Objects". Each section provides specific guidance and links to example DMPs. The "Physical Samples and Non-Digital Objects" section includes a numbered list of six questions to address in a DMP, followed by three example links: "Bone/wood lander studies example", "Cataloging insects example", and "Colorado School of Mines example". The final section is "Physical Sciences and Engineering", which provides links to DMPs 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:

~ π ~>

## 8. 데이터 매니지먼트

---

- 데이터 매니지먼트 계획(Data Management Plan)

*DMP: 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

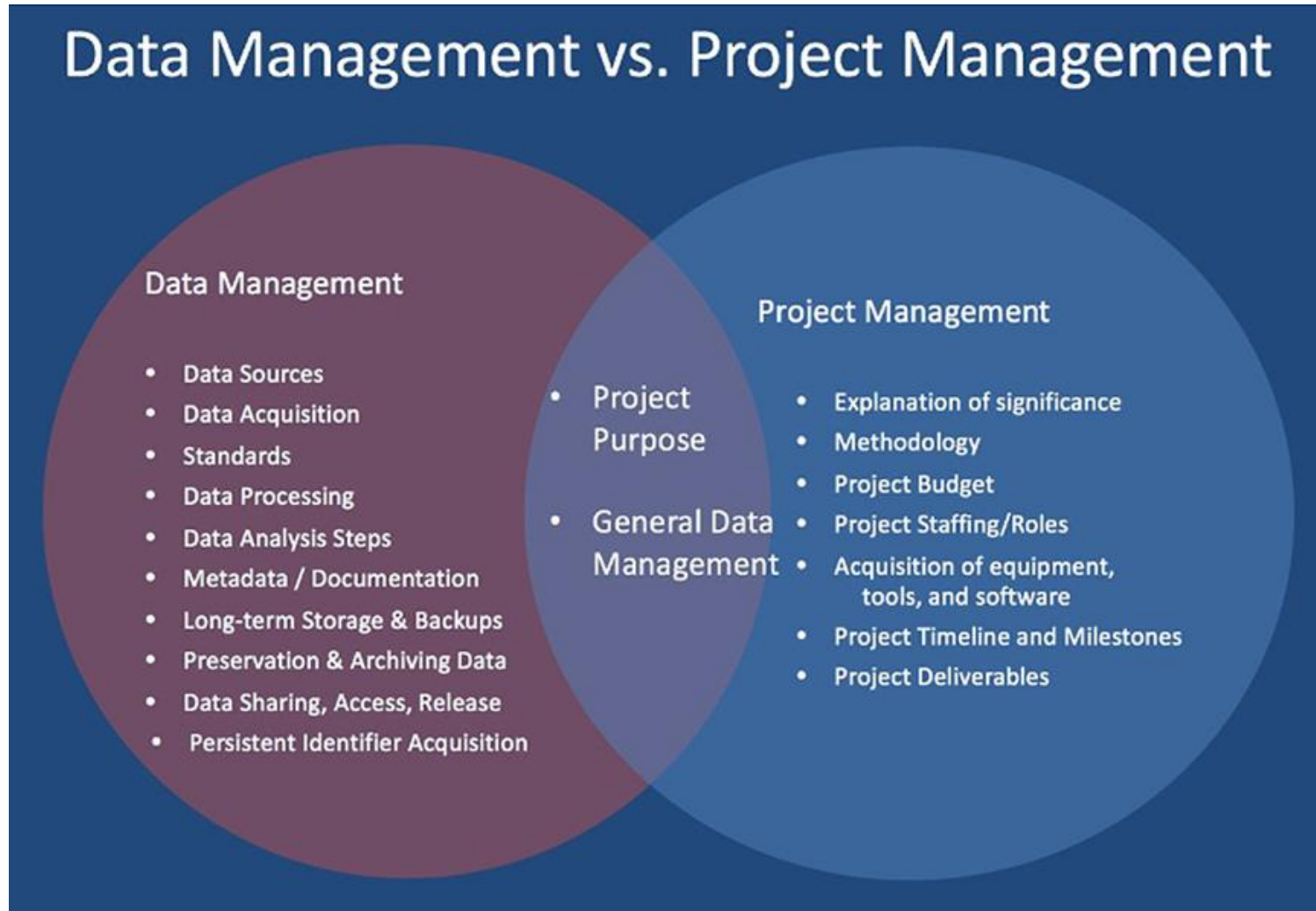
- How and where are data stored?
- How often are back ups performed and by whom?

- 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



## 9. 실습 도구

### ➤ 구글 Colab?

- 클라우드 상 파이썬 환경
- 쥬피터노트북과 동일!

Introduction to RL and Deep Q Networks.ipynb

파일 수정 보기 삽입 런타임 도구 도움말

목차

- Copyright 2018 The TF-Agents Authors.  
Licensed under the Apache License, Version 2.0 (the "License");
- Introduction to RL and Deep Q Networks
- Introduction
- The Cartpole Environment
- The DQN Agent
- DQN on Cartpole in TF-Agents

섹션

Copyright 2018 The TF-Agents Authors.  
Licensed under the Apache License, Version 2.0 (the "License");

Introduction to RL and Deep Q Networks

[View on TensorFlow.org](#) [Run in Google Colab](#) [View source on GitHub](#) [Download notebook](#)

### Introduction

Reinforcement learning (RL) is a general framework where agents learn to perform actions in an environment so as to maximize a reward. The two main components are the environment, which represents the problem to be solved, and the agent, which represents the learning algorithm. The agent and environment continuously interact with each other. At each time step, the agent takes an action on the environment based on its policy  $\pi(a_t | s_t)$ , where  $s_t$  is the current observation from the environment, and receives a reward  $r_{t+1}$  and the next observation  $s_{t+1}$  from the environment. The goal is to improve the policy so as to maximize the sum of rewards (return).

Note: It is important to distinguish between the state of the environment and the observation, which is the part of the environment state that the agent can see, e.g. in a poker game, the environment state consists of the cards belonging to all the players and the community cards, but the agent can observe only its own cards and a few community cards. In most literature, these terms are used interchangeably and observation is also denoted as  $s$ .

Agent



(1) Observation



## 9. 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

## 9. 실습 도구

### ➤ 구글 Colab?

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



검색결과 약 4,030,000개 (0.38초)

colab.research.google.com ▾ 이 페이지 번역하기

Google Colab

Colab notebooks execute code on Google's cloud servers, meaning you can leverage the power of Google hardware, including GPUs and TPUs, regardless of the ...

[Introduction to Colab and Python](#) · [Overview of Colaboratory](#) · [This notebook](#) · [Forms](#)

google colab 관련 이미지

colab pro 크렙 딥러닝 jupyter notebook deep learning 주피터노트북 파이썬 python >



→ google colab에 대한 이미지 더보기

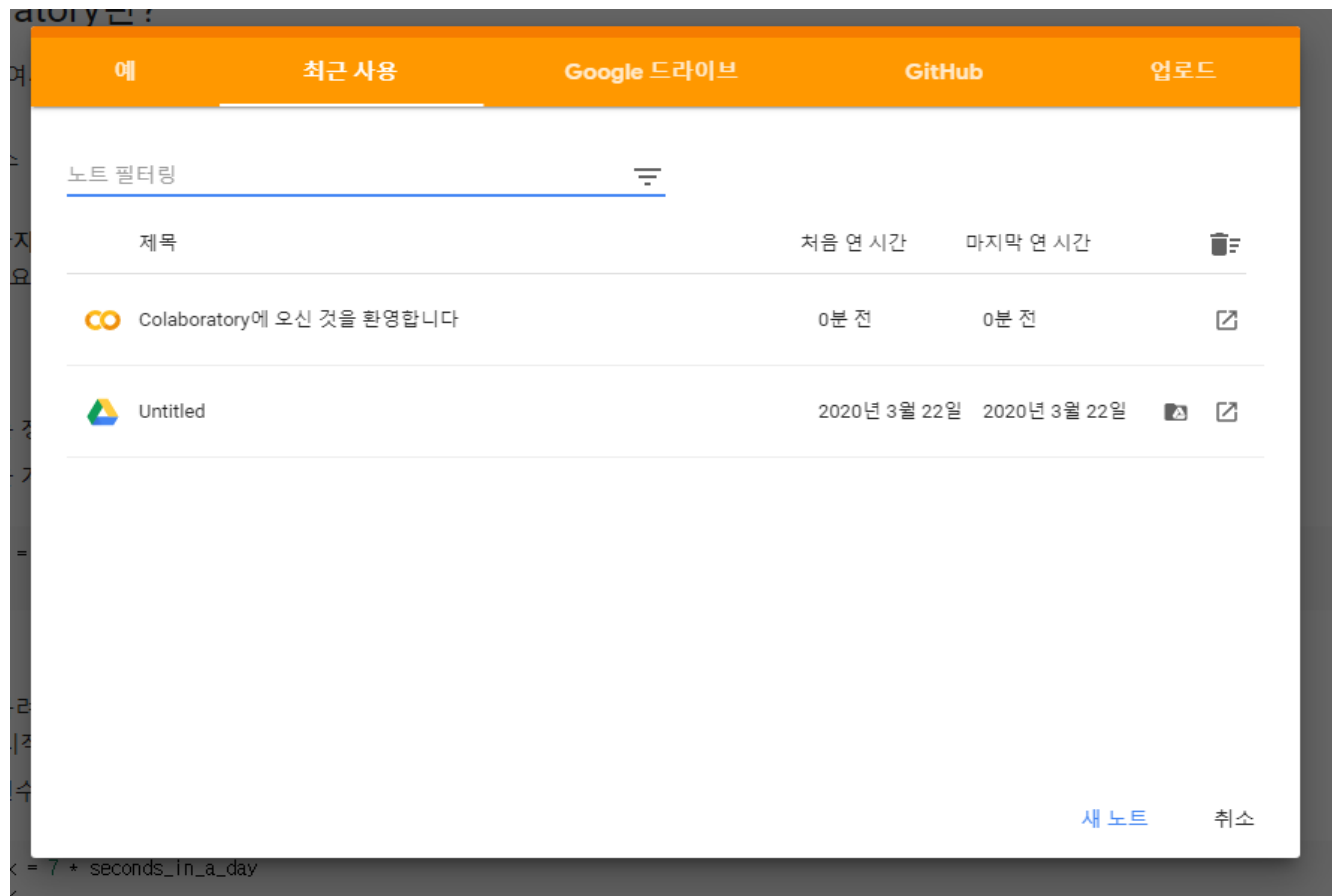




## 9. 실습 도구

### ➤ Colab

- 데이터+분석/활용 플랫폼



# 10. 향후 일정

구분	일정			과정	데이터 셋	이론&실습	내용
데이터 전처리	1주차	10/13	13:00~17:00 (4Hr)	분석 기획 개요	-	이론	분석 방법론 개요 분석 기획 실습 데이터 매니지먼트 플랜 작성 이해
		10/14	08:00~17:30 (8Hr)	데이터 전처리I	로그 파일	실습	파일 입출력을 통한 데이터 전처리 폴더 내 파일 읽기
				데이터 전처리II	제품 이미지	이론+실습	이미지 파일 입력 이미지 파일 전처리
	Review	10/20(수)	13:00~17:00 (4Hr)	과제 점검 및 리뷰			
	2주차	10/26	08:00~17:30 (8Hr)	텍스트 데이터 전처리	제품 리뷰	이론	텍스트 데이터 전처리 방법 이해 텍스트 데이터 전처리(영문) 텍스트 데이터 전처리(국문)
				텍스트 데이터 분석		이론+실습	문헌용어행렬 생성 고빈도 단어 분석 문헌 클러스터링
		10/27	08:00~17:30 (8Hr)	웹 수집I	국내 뉴스	이론	웹 데이터 수집 방법 리뷰 웹 데이터 수집 실습 수집 데이터 파싱
				웹 수집II	해외 뉴스	실습	반복문 활용 웹 데이터 수집 실습 수집 데이터 파싱 셀레니움 활용 수집 방법 실습
	Review	11/3 (수)	13:00~17:00 (4Hr)	과제 점검 및 리뷰			

## 10. 향후 일정

구분	일정			과정	데이터 셋	이론&실습	내용
모델링	3주차	11/10	08:00~17:30 (8Hr)	Feature Engineering & EDA		이론+실습	변수 처리 및 변환 개요, 주성분 분석 텍스트, 이미지에서의 변수 처리
						이론+실습	EDA 개요 및 데이터 탐색
		11/11	08:00~17:30 (8Hr)	머신러닝 적용	판매/제조/운영 관련 데이터	이론+실습	정형/반정형/비정형 데이터에 대한 머신러닝 적용 계획 머신러닝 기법 리뷰 지도학습 적용 및 기법별 성능 평가 비교 비지도학습 적용과 결과 리뷰
	Review	11/17 (수)	13:00~17:00 (4Hr)	과제 점검 및 멘토 학습 / 1~3주차 내용 리뷰			
모델링	4주차	11/24	08:00~17:30 (8Hr)	딥러닝 적용		이론	정형/반정형/비정형 데이터에 대한 딥러닝 적용 계획 딥러닝 기법 리뷰
텍스트/이미지 관련 데 이터					이론+실습	주요 딥러닝 모형의 적용 및 성능 개선을 위한 하이퍼파라 미터 튜닝	
결과 해석 및 시각화		11/25	08:00~17:30 (8Hr)	시각화		이론	시각화 방안 리뷰 분석 결과 시각화를 위한 라이브러리 학습 전 단계 분석 결과 시각화 실습
						실습	웹 시각화를 위한 dash 등의 라이브러리 소개 기본 사용법 이해 시각화 실습
모델링 결과 해석 및 시각화	Review	12/1 (수)	13:00~17:00 (4Hr)	과제 점검 및 멘토 학습 / 1~4주차 내용 리뷰			

## 10. 향후 일정

구분	일정			과정	데이터 셋	이론&실습	내용
Workshop	5주차	12/08 ~ 12/09	08:00~17:30 (16Hr)	1차: 전처리+EDA	기존 데이터 확장	실습	데이터 분석 문제 발굴 및 분석 기획 필요 데이터에 대한 데이터 매니지먼트 플랜 작성 및 리뷰 데이터 수집 및 전처리 탐색적 데이터분석 수행 중간 결과 발표 및 피드백 Feature Engineering 적용 모형 선택 및 적용, 모형 비교 모형 튜닝을 통한 성능개선 결과 보고서 작성 및 비즈니스 시사점 도출, 발표 및 피드백
				2차: 모델링 및 시각화			

### 워크숍 주제 예시:

- 00 분야에 대한 국내외 뉴스 웹수집 및 텍스트 마이닝을 통한 이슈 발견과 예측
- 제조 데이터 또는 관련 이미지를 활용한 불량 탐지 기반 품질 관리
- 실외/실내 공간 이미지 전처리 및 분류 모델링

---

Q&A

