

정 규 세 셴 1 주 차

ToBig's 11기 유기윤

# 머신러닝을 시작하기 앞서서

데이터 전처리, EDA 를 중심으로

# Contents

---

Unit 01 | Introduction

---

Unit 02 | Data Preprocessing

---

Unit 03 | Exploratory Data Analysis

---

Unit 04 | 과제 설명

---

## Unit 01 | Introduction

빅데이터

딥러닝

기계학습

머신러닝

Where Are We ?

투빅스

## Unit 01 | Introduction

1주차 : 전처리 / EDA / Git      데이터마이닝

2주차 : 회귀분석 / 로지스틱 회귀

3주차 : 나이브베이즈 / SVM / KNN

4주차 : 의사결정나무 / 앙상블 / 클러스터링

5주차 : 클래스 / 차원축소 LDA PCA

## 머신러닝

6주차 : Neural Network 기초 / 프레임워크

7주차 : Neural Network 심화 / 크롤링

8주차 : 이미지(CNN) 기초 / 텍스트(NLP) 기초 / 강화학습 기초

9주차 : 이미지(CNN) 심화 / 알고리즘 원리

10주차 : 텍스트(NLP) 심화(RNN, LSTM, E/D, Attention)

## 딥러닝

## Unit 01 | Introduction

데이터 마이닝 (분석) : 데이터에 대한 인사이트

머신러닝 (기계학습) : 데이터를 통한 예측

인공지능 : 데이터를 통해 Agent 가 행동을 실행 또는 추천

## Unit 01 | Introduction

# 데이터 마이닝 : 데이터에 대한 인사이트

- Statistical inference
- Data visualization
- Experiment design
- Domain knowledge
- Communication

해결하고자 하는 문제 예시:

“A 성질과 B 나이대의 고객들이 상품 구매율이 높다”

구글 딥마인드의 알파고나 스타크래프트 등의 게임 플레이 Agent

## Unit 01 | Introduction

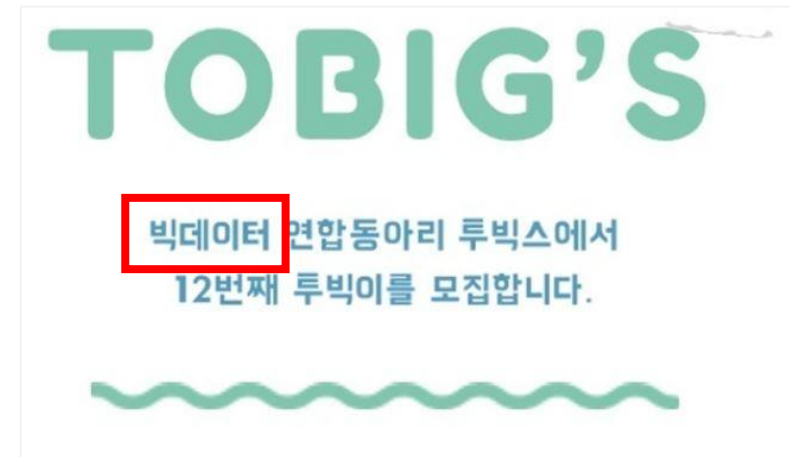
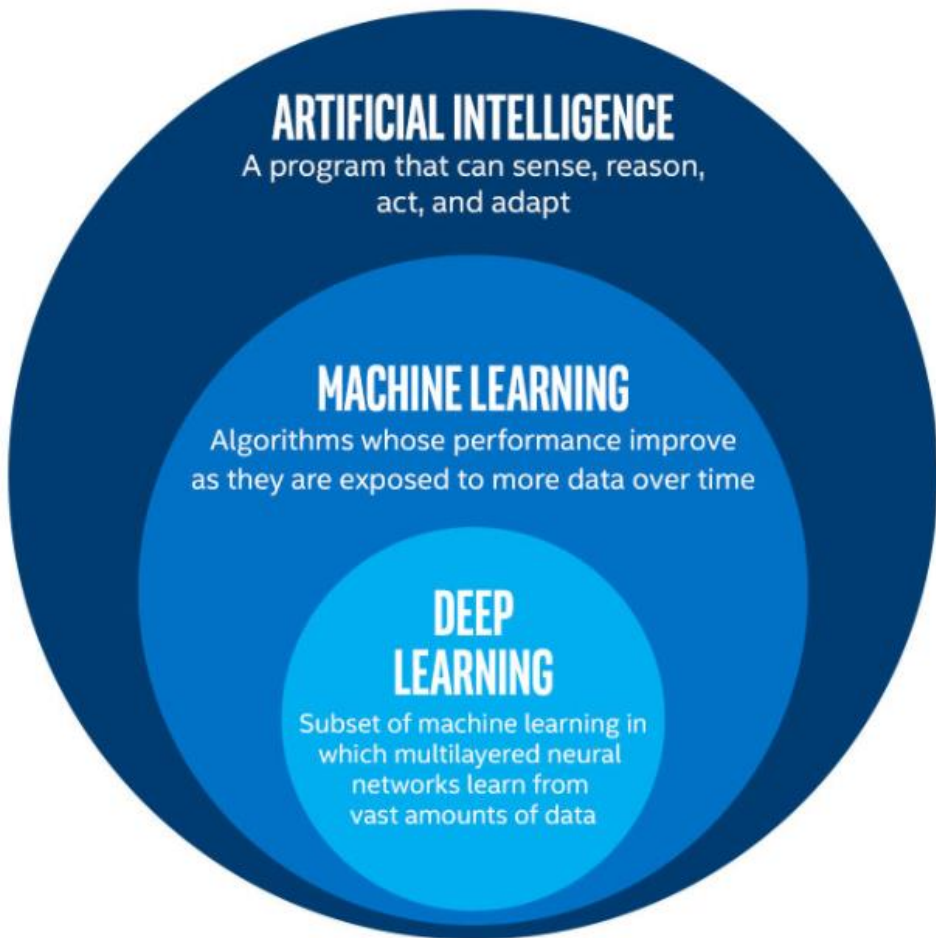
# 머신러닝 (기계학습) : 데이터를 통한 예측

미래에 대한 예측: 특정 환자가 병이 있는지?

컴퓨터에게는 자명하지 않은 사실 예측: 사진 속에 차가 있는가?

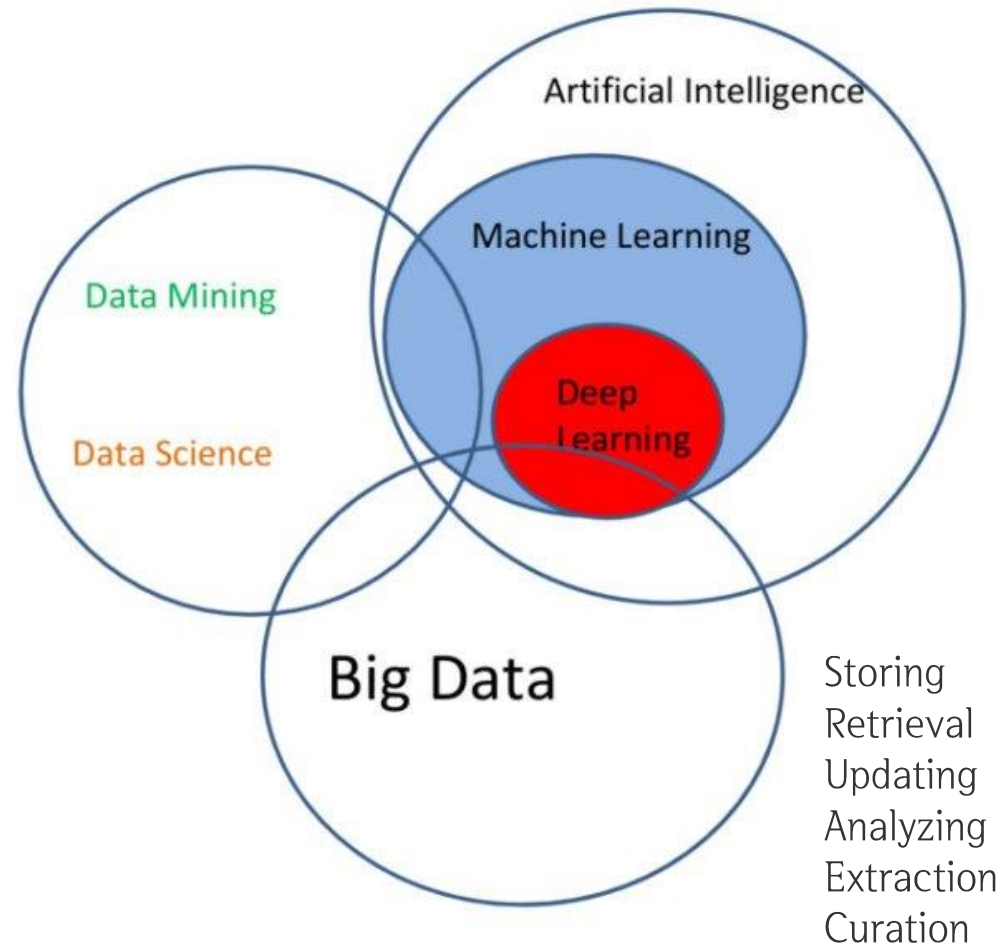
## 딥러닝 : 머신러닝 기법 중 Neural Net 을 통해 모델링

## Unit 01 | Introduction





## Unit 01 | Introduction



## Unit 01 | Introduction



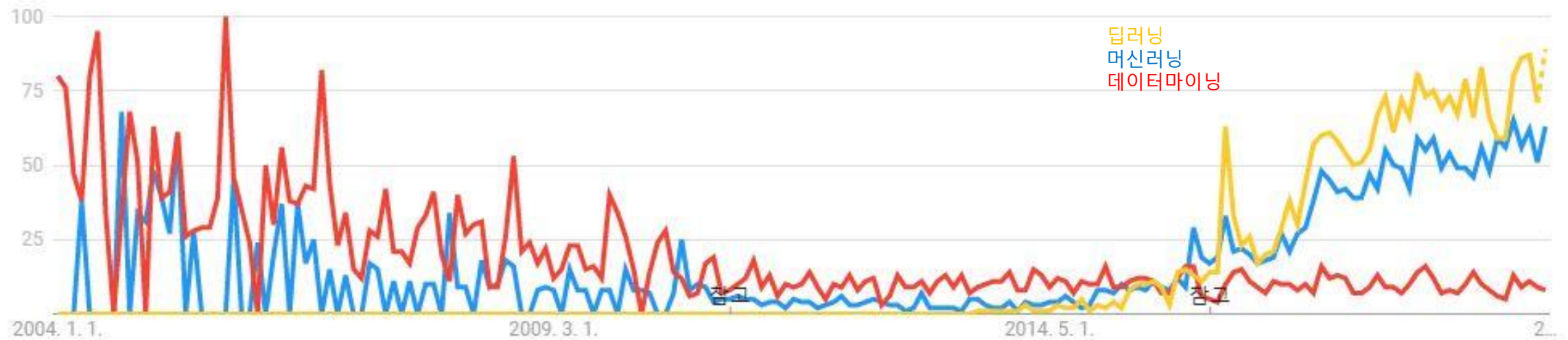
그렇다면 왜 '빅데이터' ?

Big Data

storing,  
retrieval,  
updating,  
analyzing,  
extraction ,  
curation

## Unit 01 | Introduction

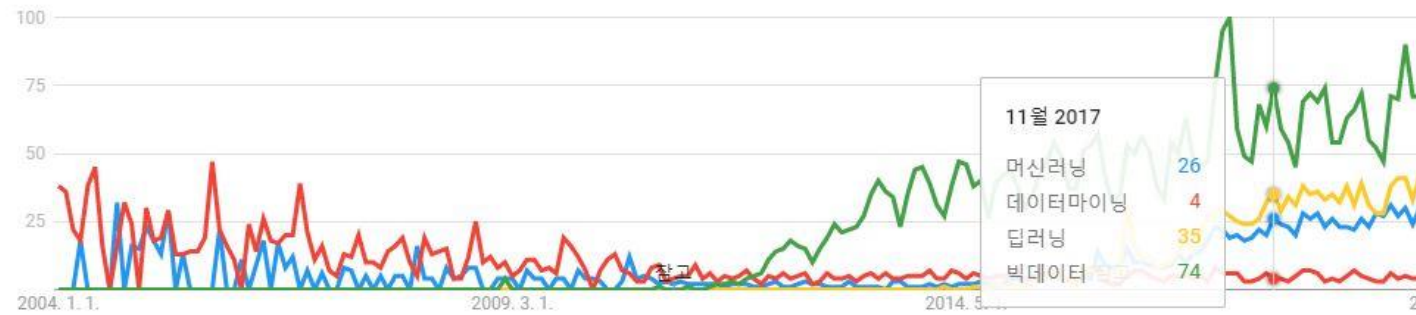
## 구글 아날리틱스 (2004년-2019년):



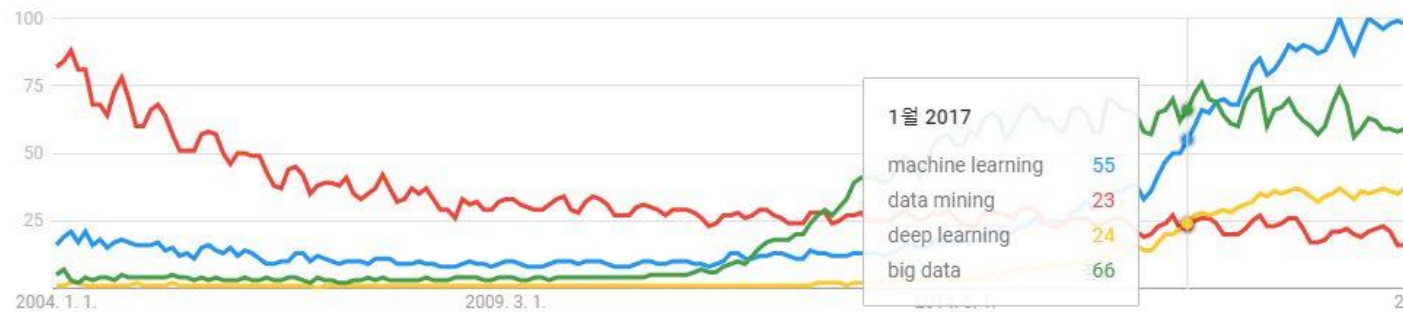
## Unit 01 | Introduction

## 구글 아날리틱스 (2004년-2019년):

한국



전세계



## Unit 01 | Introduction

NAVER

빅데이터분석



파워링크 '빅데이터분석' 관련 광고입니다. ⓘ

**빅데이터분석, 퓨처젠** [www.futuregen.co.kr](http://www.futuregen.co.kr)

빅데이터분석, 제품판매, IBM파트너, 견적상담, 유지보수, IT솔루션제공

**국가인증 코리아 IT아카데미** [www.koreaitacademy.com](http://www.koreaitacademy.com)

수강료조회 · 수강평 보기 · TV광고보기 · 학원시설

빅데이터분석, 브랜드대상, 여름방학특강 사전등록이벤트, 국비무료, 수강료&amp;

**그린컴퓨터아트학원공식홈페이지** [www.greenart.co.kr](http://www.greenart.co.kr) 

국비지원안내 · 추천교육과정 · 수강료조회 · 지점조회

취준생&amp;직장인 국비100% 무료지원, 실무양성교육, 1:1맞춤취업지원, 빅데이터

**NO.1 더조은컴퓨터아카데미** [www.tjoen.co.kr](http://www.tjoen.co.kr)

국비교육과정 · 취업스토리 · 포트폴리오 · 온라인상담

취업이잘되는 빅데이터분석, 실무형교육, 압도적인 수강생, 포트폴리오 역시 다

**SAP 코리아 공식 홈페이지** [www.sap.com/korea/](http://www.sap.com/korea/)

머신 러닝 · 데이터 관리 · 비즈니스분석

SAP Leonardo 기술을 통한 혁신, 통합된 비즈니스, 성장하는 기업!

국비교육과정 · 취업스토리 · 포트폴리오 · 온라인상담  
취업이잘되는 빅데이터분석, 실무형교육, 압도적인 수강생, 포트폴리오 역시 더조은!**SAP 코리아 공식 홈페이지** [www.sap.com/korea/](http://www.sap.com/korea/)

머신 러닝 · 데이터 관리 · 비즈니스분석

SAP Leonardo 기술을 통한 혁신, 통합된 비즈니스, 성장하는 기업!

**데이터바우처 데이터메이커** [www.datavoucher.or.kr/](http://www.datavoucher.or.kr/)

데이터 바우처 지원사업, AI가공 부터 코딩, 크롤링, 시각화 등 일반가공까지!

**메가IT아카데미 강남역** [megaitacademy.com/](http://megaitacademy.com/)

친구 동행 시 7월 전과목 10만원! 메가스터디에서 설립한 IT전문 교육기관!

**RSN 알에스엔** [www.realsn.com](http://www.realsn.com)

실시간 수집, 분석 소셜 빅데이터 를 연구개발 12개 특허 200개 고객사 보유

**에이콘 아카데미** [acornacademy.co.kr](http://acornacademy.co.kr)

국비지원 무료교육, 취업지원, 빅데이터 분석, 머신러닝, 딥러닝, 실무프로젝트

**이든티앤에스 E2D2** [www.e2d2.co.kr](http://www.e2d2.co.kr)

E2D2, 지능형 독립 데이터 센터 솔루션, 설치 구성 최적화, 통합구축 비용절감



## Unit 01 | Introduction

A.I.  
인공지능 네이밍  
노하우와 데이터 베이스로  
맘에 쏘옥 드는 브랜드와 만남

A.I.

K1455070025  
A.I. 네이밍 서비스, 인공지능처럼  
노하우와 데이터베이스로 쏘옥 ...

90,000 원 ~

★★★★★ (2)

인공지능 AI  
별자리 운세

퓨처닥터  
AI [인공지능] 별자리 프로그램을  
활용해 보다 정확한 진로-연애-...

20,000 원 ~

★★★★★ (5)



Amy Hoy  
@amyhoy



by today's definition,  $y=mx+b$  is an artificial intelligence bot that can tell you where a line is going

♡ 6,002 11:44 PM - Mar 29, 2017

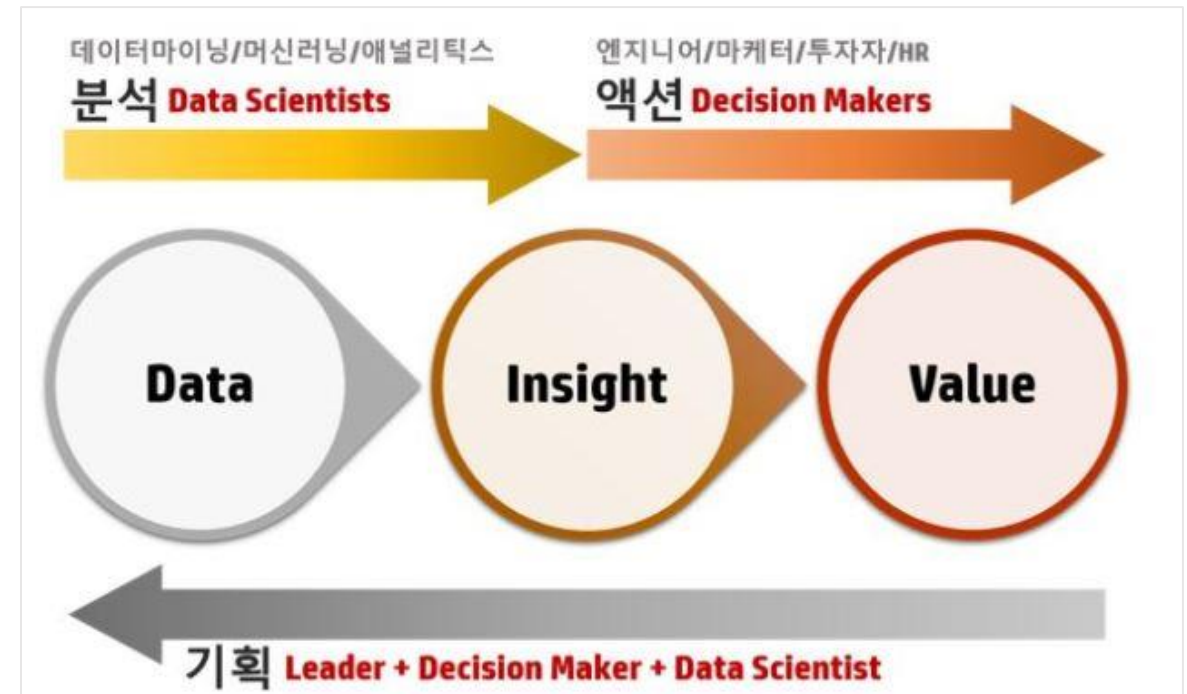


💬 3,629 people are talking about this



## Unit 01 | Introduction

빅데이터 :  
비즈니스 밸류 창출의  
근거가 되는 인사이트의 재료

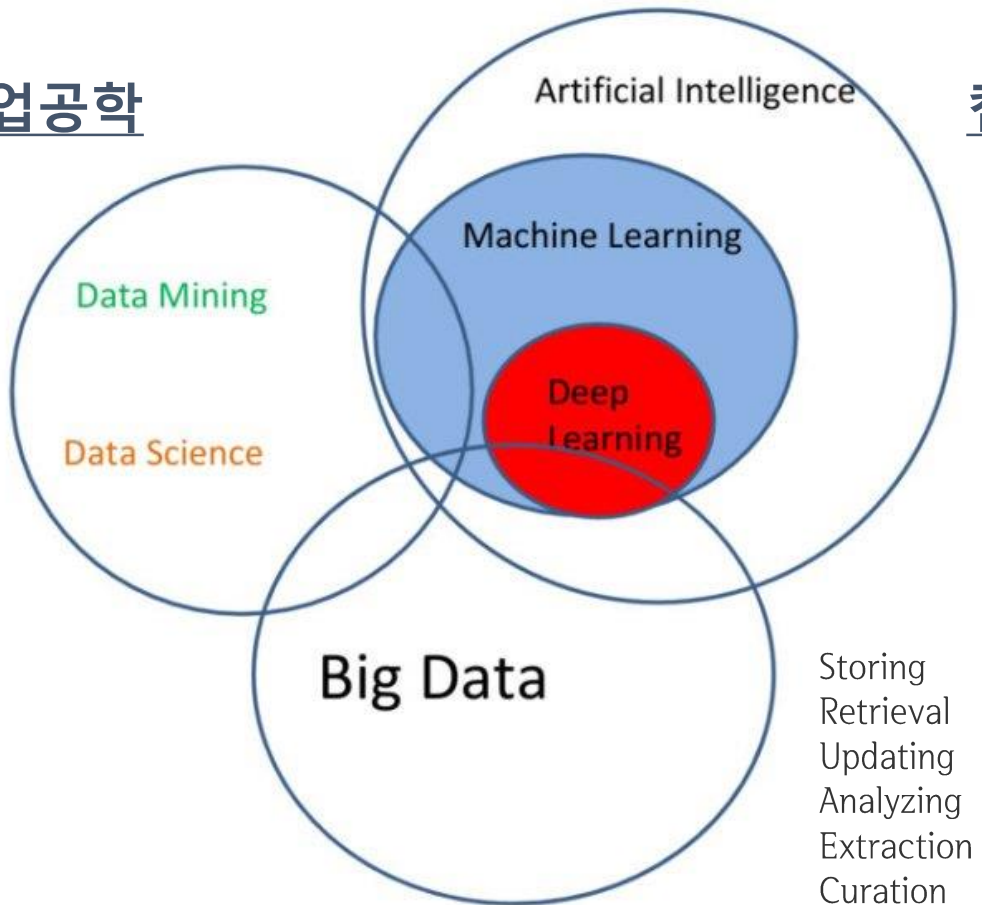


### 데이터마이닝과 빅데이터와의 관계?

최근 분산 데이터베이스 (distributed database) 기술 및 병렬처리 (parallel processing) 기술 발전에 힘입어, 대량의 (volume), 빨리 생성되는 (velocity), 숫자 뿐만 아니라 텍스트와 이미지 동영상 같은 다양한 (variety) 데이터로 대표되는 빅데이터 (big data) 도 분석 가능해졌습니다. 바로 이 빅데이터 분석(analytics)이 데이터마이닝입니다.

# Unit 01 | Introduction

# 산업공학



## 컴퓨터공학, 전기전자

## Deep Blue, AlphaGo (Reinforcement Learning)

Siri, Bixby (Natural language processing & Natural language understanding)

자율주행 (Object Detection, RL)



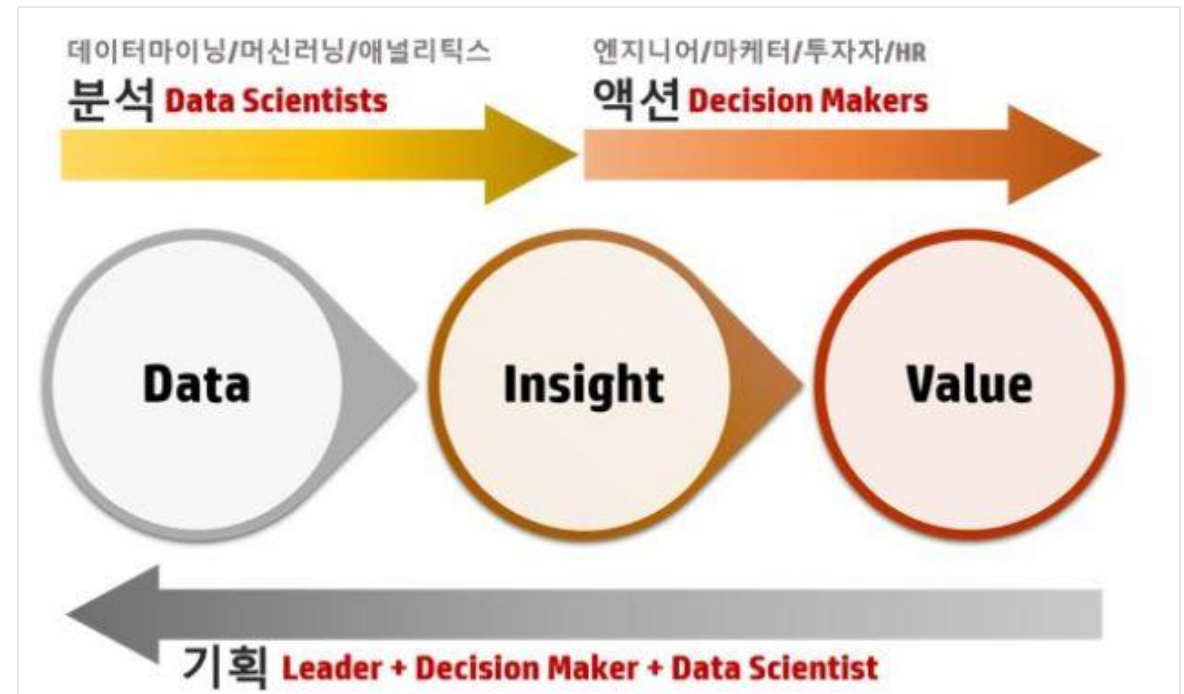
## Unit 01 | Introduction

데이터 전처리 :

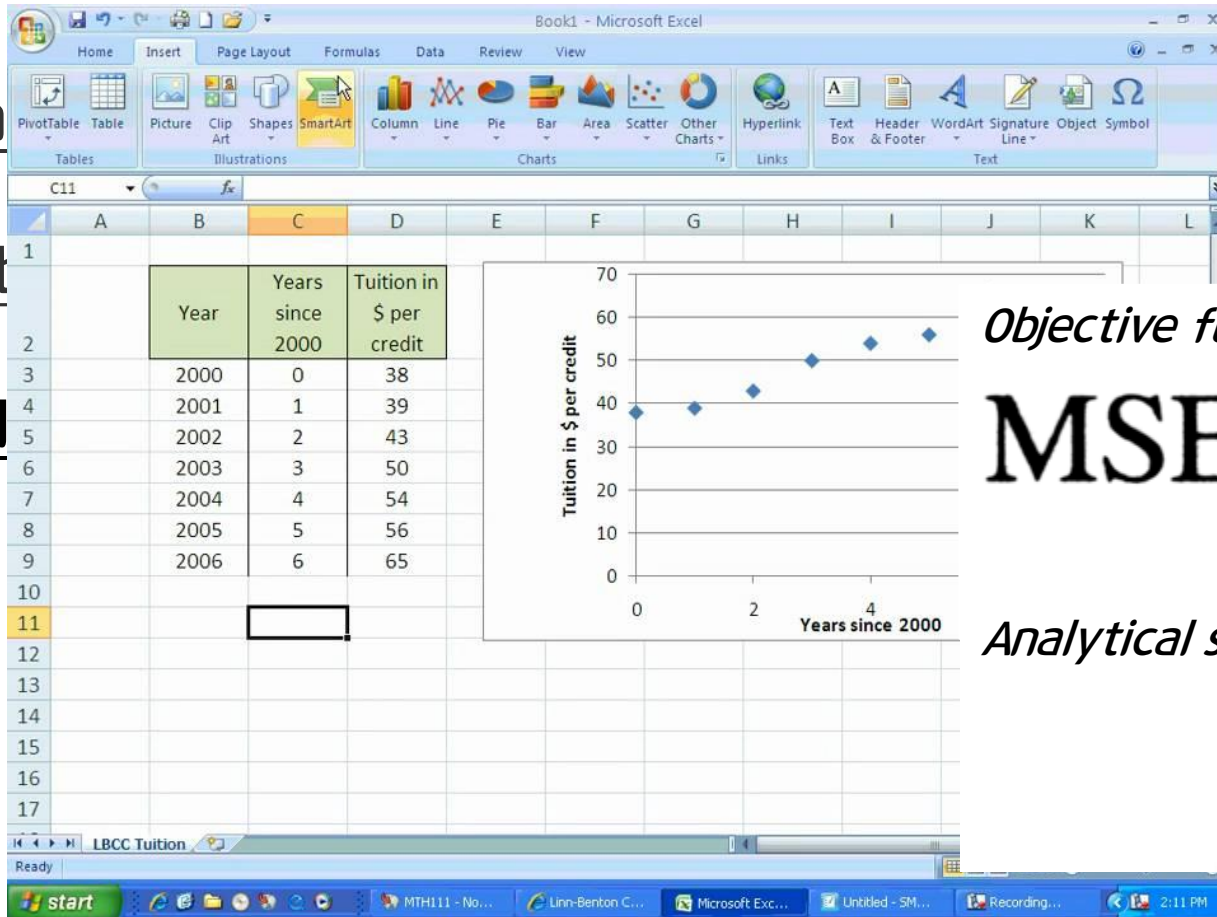
모델에 Input 값으로 데이터를 넣기 위해  
정제하는 과정

Exploratory Data Analysis:

탐색적 데이터 분석: 여러 방면에서 데이터를  
관찰하고 이해하는 과정



## Unit 01 | Introduction

In  
Ot  
Al*Objective func.*

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

*Analytical solution*

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

## Unit 01 | Introduction

1주차 : 전처리 / EDA / Git

2주차 : 회귀분석 / 로지스틱 회귀

3주차 : 나이브베이즈 / SVM / KNN

4주차 : 의사결정나무 / 앙상블 / 클러스터링

5주차 : 클래스 / 차원축소 LDA PCA

6주차 : Neural Network 기초 / 프레임워크

7주차 : Neural Network 심화 / 크롤링

8주차 : 이미지(CNN) 기초 / 텍스트(NLP) 기초 / 강화학습 기초

9주차 : 이미지(CNN) 심화 / 알고리즘 원리

10주차 : 텍스트(NLP) 심화(RNN, LSTM, E/D, Attention)

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Labels: Likelihood (points to  $P(x|c)$ ), Class Prior Probability (points to  $P(c)$ ), Posterior Probability (points to  $P(c|x)$ ), Predictor Prior Probability (points to  $P(x)$ )

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Maximize (in  $\alpha_i$ )

$$\tilde{L}(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to (for any  $i = 1, \dots, n$ )

$$\alpha_i \geq 0,$$

and to the constraint from the minimization in  $b$ 

$$\sum_{i=1}^n \alpha_i y_i = 0.$$

- The loss is

$$\ell_{\log\text{-loss}}(\mathbf{w}) = - \sum_{i=1}^N t^{(i)} \log p(C=1|\mathbf{x}^{(i)}, \mathbf{w}) - \sum_{i=1}^N (1-t^{(i)}) \log p(C=0|\mathbf{x}^{(i)}, \mathbf{w})$$

where the probabilities are

$$p(C=0|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-z)} \quad p(C=1|\mathbf{x}, \mathbf{w}) = \frac{\exp(-z)}{1 + \exp(-z)}$$

and  $z = \mathbf{w}^T \mathbf{x} + w_0$ 

- We can simplify

$$\begin{aligned} \ell(\mathbf{w})_{\log\text{-loss}} &= \sum_i t^{(i)} \log(1 + \exp(-z^{(i)})) + \sum_i t^{(i)} z^{(i)} + \sum_i (1 - t^{(i)}) \log(1 + \exp(-z^{(i)})) \\ &= \sum_i \log(1 + \exp(-z^{(i)})) + \sum_i t^{(i)} z^{(i)} \end{aligned}$$

## Unit 01 | Introduction

Book1 - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

PivotTable Table Picture Clip Art Shapes SmartArt Column Line Pie Bar Area Scatter Other Charts Hyperlink Text Box Header & Footer WordArt Signature Object Symbol

C11 fx

Data columns (total 41 columns):

Auction_key	1933	non-null	int64	Final_auction_date	1933	non-null	object
Auction_class	1933	non-null	object	Final_result	1933	non-null	object
Bid_class	1933	non-null	object	Creditor	1933	non-null	object
Claim_price	1933	non-null	int64	addr_do	1933	non-null	object
Appraisal_company	1933	non-null	object	addr_si	1933	non-null	object
Appraisal_date	1933	non-null	object	addr_dong	1933	non-null	object
Auction_count	1933	non-null	int64	addr_li	23	non-null	object
Auction_misconriage_count	1933	non-null	int64	addr_san	1933	non-null	object
Total_land_gross_area	1933	non-null	float64	addr_bunji1	1929	non-null	float64
Total_land_real_area	1933	non-null	float64	addr_bunji2	889	non-null	float64
Total_land_auction_area	1933	non-null	float64	addr_etc	1933	non-null	object
Total_building_area	1933	non-null	float64	Apartment_usage	1933	non-null	object
Total_building_auction_area	1933	non-null	float64	Preserve_regist_date	1933	non-null	object
Total_appraisal_price	1933	non-null	int64	Total_floor	1933	non-null	int64
Minimum_sales_price	1933	non-null	int64	Current_floor	1933	non-null	int64
First_auction_date	1933	non-null	object	Specific	64	non-null	object
				Share_auction_YorN	1933	non-null	object
				road_name	1933	non-null	object

그런데 만약 주어진 데이터

라면?

면?

≒ 범주형?

가 너무 많아  
들다면?

## Unit 01 | Introduction

1주차 : 전처리 / EDA / Git

2주차 : 회귀분석 / 로지스틱 회귀

3주차 : 나이브베이즈 / SVM / KNN

4주차 : 의사결정나무 / 앙상블 / 클러스터링

5주차 : 클래스 / 차원축소 LDA PCA



## 머신러닝

6주차 : Neural Network 기초 / 프레임워크

7주차 : Neural Network 심화 / 크롤링

8주차 : 이미지(CNN) 기초 / 텍스트(NLP) 기초 / 강화학습 기초

9주차 : 이미지(CNN) 심화 / 알고리즘 원리

10주차 : 텍스트(NLP) 심화(RNN, LSTM, E/D, Attention)



## 딥러닝

# Contents

---

Unit 01 | Introduction

---

Unit 02 | Data Preprocessing

---

Unit 03 | Exploratory Data Analysis

---

Unit 04 | 과제 설명

---

## Unit 02 | Data Preprocessing

- 필요한 라이브러리:

- Pandas : 데이터 분석을 위한 툴

- A fast and efficient DataFrame object for data manipulation with integrated indexing;
    - Tools for reading and writing data between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format;
    - Intelligent data alignment and integrated handling of missing data: gain automatic label-based alignment in computations and easily manipulate messy data into an orderly form

- Numpy : 매트릭스, 벡터 연산 적합한 툴

- Matplotlib, Seaborn : 시각화 패키지

- Jupyter Notebook : 대화형 인터페이스

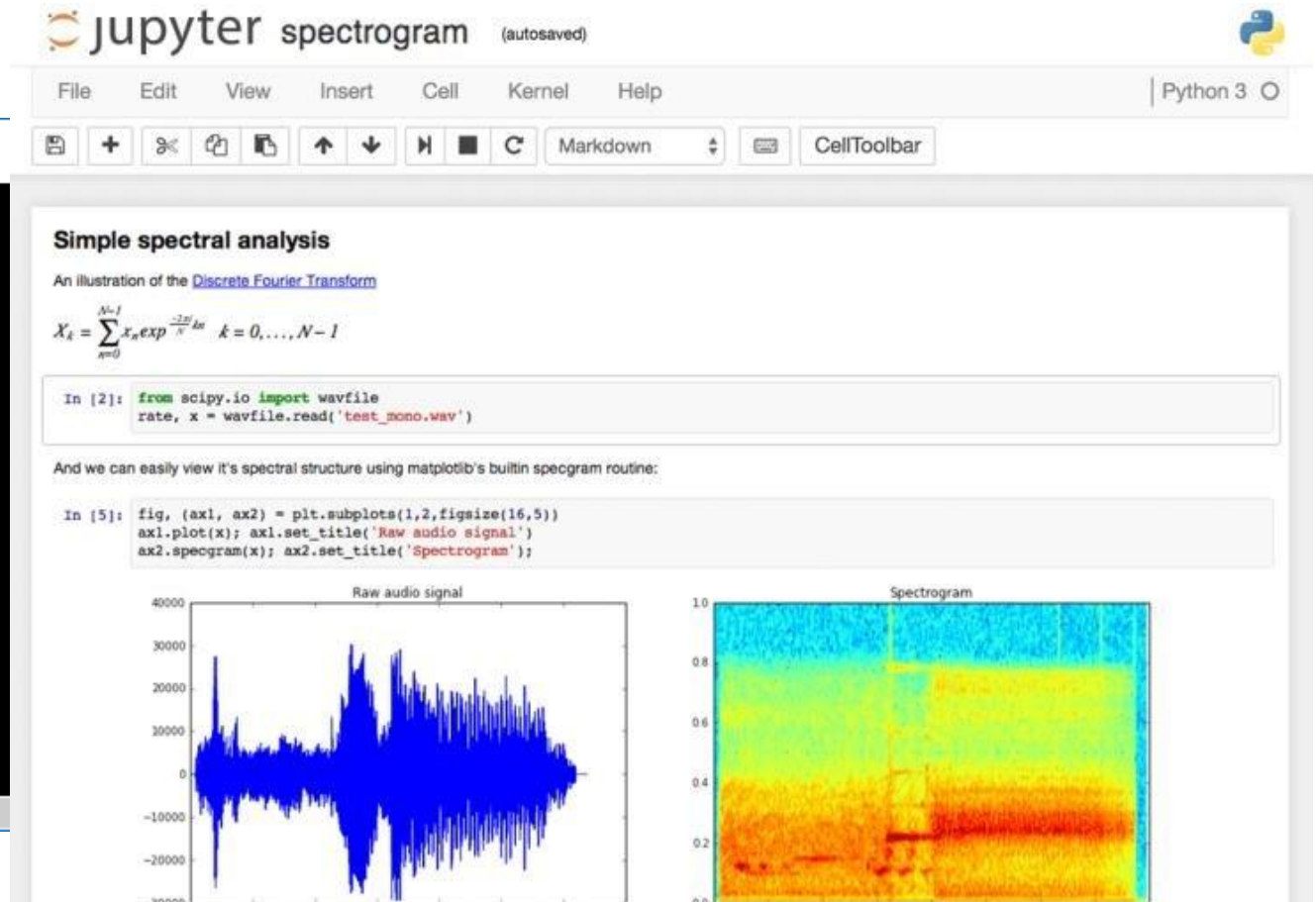


## Unit 02 | Data Preprocessing

## • 필요한 라이브러리:

```
Command Prompt
Microsoft Windows [Version 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\Brennan>
```





## Unit 02 | Data Preprocessing

The Kaggle logo is displayed in a large, light blue, lowercase sans-serif font.


2010년 설립된 예측모델 및 분석 대회 플랫폼이다. 기업 및 단체에서 데이터와 해결과제를 등록하면, 데이터 과학자들이 이를 해결하는 모델을 개발하고 경쟁한다. (위키백과)


Competitions Datasets Kernels Discussion Courses


대회, 데이터셋, 커널, 강의까지 제공


## Unit 02 | Data Preprocessing


15 Active Competitions

  
TWO SIGMA



  
BIG DATA  
COMPETITION





Two Sigma  
Use news  
Featured

Air pollutants measured in Seoul

Trending YouTube Video Statistics

Shinzo Abe (Japanese Prime Minister) Twitter NLP  
Team AI  
2 years 22 KB 7.1 1 File (CSV)

SIIM-AI  
Identify Pr  
Featured

Open Inria  
Detect obj  
Research

20  
sample submission  
4h ago with multiple data

19  
Adversarial IEEE  
7h ago with multiple data

IEEE Fraud De

Electric Motor Temperature  
Kirsan  
a month 47 MB 10.0 1 File (CSV)

Chennai Water Management  
SRK  
a month 75 KB 8.8 2 Files (CSV)

50

104

27

1109

120

12

63

9  
7h ago with multiple data

109  
Finding Pneumonia  
12h ago with multiple data

9  
Instagram to APT  
10h ago with multiple data

4  
Reducing Memory S  
2h ago in ieee-fraud-dete

City of New York  
a day 3 MB 8.5 3 Files (other, JSON, CSV)

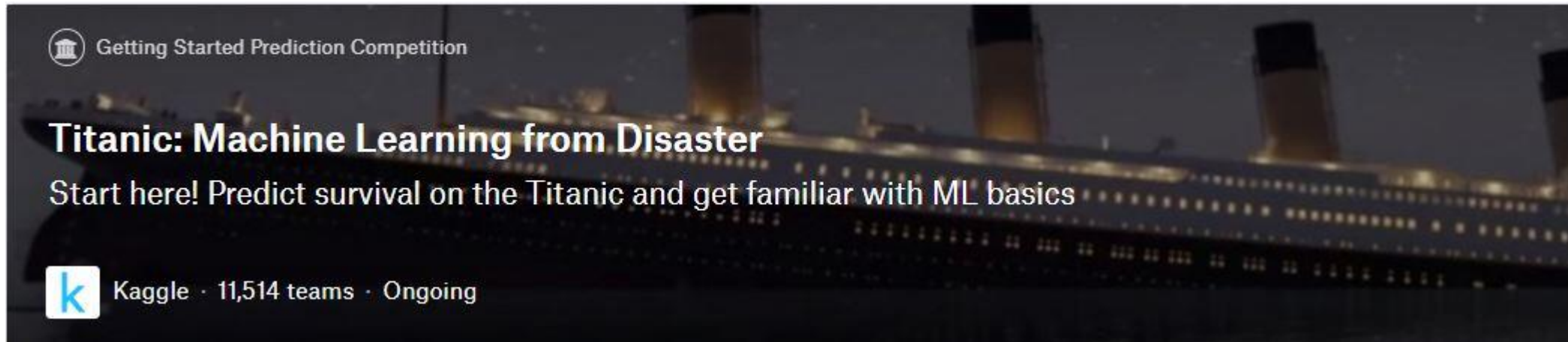
Search Engine Results - Flights & Tickets Keywords  
Elias Dabbas  
20 hours 21 MB 9.1 16 Files (CSV)

Crimes in Boston  
Analyze Boston  
10 months 10 MB 8.2 2 Files (CSV)

20

19

## Unit 02 | Data Preprocessing



Feature : 이름 / 나이 / 성별 / 좌석 / 동행 가족 수 / 탑승지역 등등

Target : 생존 유무

## Unit 02 | Data Preprocessing

- 전처리 :
  - 결측치 처리
  - 범주형 변수 처리
    - One-hot encoding
    - Ordinal encoding
    - Target encoding

## Unit 02 | Data Preprocessing : 결측치 처리

- Real-world 데이터는 결측치가 존재하는 데이터셋이 다수
- 값이 아예 비어있거나, Nan, None 등 다양한 형식으로 숨어있음
  - Nan : Not a number ; float
  - None : Pythonic object
- 따라서 각 변수가 의미하는 것을 확인하는 것이 중요

isnull(): Generate a boolean mask indicating missing values

notnull(): Opposite of isnull()

dropna(): Return a filtered version of the data

fillna(): Return a copy of the data with missing values filled or imputed

## Unit 02 | Data Preprocessing : 결측치 처리

결측치 확인을 위한  
가장 간단한 방법 :

```
# Null 값 개수 복습!  
data.isnull().sum()
```

```
PassengerId    0  
Survived        0  
Pclass          0  
Name            0  
Sex             0  
Age            177  
SibSp           0  
Parch           0  
Ticket         0  
Fare            0  
Cabin          687  
Embarked        2  
dtype: int64
```

발생할 수 있는 문제점 ?

‘Null’, ‘없음’ 처럼 숨어져 있는 결  
측치는 파악하기 어려움

## Unit 02 | Data Preprocessing : 결측치 처리

결측치 확인을 위한  
가장 간단한 방법 :

```
# Null 값 개수 복습!  
data.isnull().sum()
```

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

dtype: int64

옵션 1 : 데이터 제외

옵션 2 : Imputing

## Unit 02 | Data Preprocessing : 결측치 처리

## 옵션 1 : 데이터 제외 (행/열)

```
# Null 값 개수 복습!  
data.isnull().sum()
```

```
PassengerId    0  
Survived        0  
Pclass          0  
Name            0  
Sex             0  
Age            177  
SibSp           0  
Parch           0  
Ticket          0  
Fare            0  
Cabin          687  
Embarked        2  
dtype: int64
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NA	
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	2	1	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NA	
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NA	S



## Unit 02 | Data Preprocessing : 결측치 처리

## 옵션 1 : 데이터 제외 (행/열)

```
# Null 값 개수 복습!  
data.isnull().sum()
```

```
PassengerId    0  
Survived        0  
Pclass          0  
Name            0  
Sex             0  
Age            177  
SibSp           0  
Parch           0  
Ticket          0  
Fare            0  
Cabin          687  
Embarked        2  
dtype: int64
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500		S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833		C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250		S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500		S

## Unit 02 | Data Preprocessing : 결측치 처리

### 옵션 2 : Imputation

통계치, 외부 정보 활용

실습으로 ~

# Contents

---

Unit 01 | Introduction

---

Unit 02 | Data Preprocessing

---

Unit 03 | Exploratory Data Analysis

---

Unit 04 | 과제 설명

---

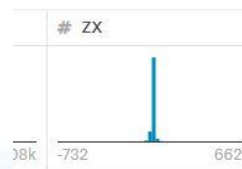


## Unit 03 |EDA

실습으로 ~

## Unit 03 | EDA의 필요성

## Data Sources

dipole_moments.csv	85.0k x 4
magnetic_shielding...	1.53m x 11
mulliken_charges.csv	1.53m x 3
potential_energy.csv	85.0k x 2
sample_submission...	2.51m x 2
scalar_coupling_co...	4.66m x 8
structures.csv	2.36m x 6

# ZX	molecule_name	# atom_index	# mulliken_charge
	85003 unique values		
0	-0.0001		
317	4.0544		
173	-4.1474		
386	1.2723		
342	-1.1793		
303	0.0133	dsgdb9nsd_000001	0
126	-2.1025	dsgdb9nsd_000001	1
327	-2.0299	dsgdb9nsd_000001	2
.13	4.1318	dsgdb9nsd_000001	3
0	0	dsgdb9nsd_000001	4
0	0	dsgdb9nsd_000002	0
0	0	dsgdb9nsd_000002	1
0	0	dsgdb9nsd_000002	2
304	0.0008	dsgdb9nsd_000002	3
305	0.0009	dsgdb9nsd_000003	0
394	3.3643		

Feature : ...

Target : scalar coupling constant (?)

## Unit 03 | 결론

1. Feature 종류 파악 : 쓸모없는 변수 제거 (e.g. ID), 연속형, 범주형 변수
2. 결측치 처리
3. 범주형 변수 처리
4. EDA : 모든 feature 를 target variable 와 시각화 / feature 간 상관관계 파악
5. Feature Engineering : 파생변수 생성
6. 새로 만들어진 변수에 대해서 EDA
7. 모델 선정 및 학습

# Contents

---

Unit 01 | Introduction

---

Unit 02 | Data Preprocessing

---

Unit 03 | Exploratory Data Analysis

---

Unit 04 | 과제 설명

---

## Unit 04 | 과제 1 (~30분)

- 실습파일에서 Target Encoding 부분을 구현하세요.
- Target Encoding

$x_0$	$x_1$	$y$
a	c	1
a	c	1
a	c	1
a	c	1
a	c	0
b	c	1
b	c	0
b	c	0
b	c	0
b	d	0

```
df.groupby('x_0')['y'].mean()  
{a: 0.8 , b: 0.2}
```

```
df.groupby('x_1')['y'].mean()  
{c: 0.44 , d: 0}
```



## Unit 04 | 과제 1 (~30분)

- 실습파일에서 Target Encoding 부분을 구현하세요. (추가적인 라이브러리 사용금지)
- Target Encoding

$x_0$	$x_1$	$y$
0.8	0.444	1
0.8	0.444	1
0.8	0.444	1
0.8	0.444	1
0.8	0.444	0
0.2	0.444	1
0.2	0.444	0
0.2	0.444	0
0.2	0.444	0
0.2	0	0

```
df.groupby('x_0')['y'].mean()  
{a: 0.8 , b: 0.2}
```

```
df.groupby('x_1')['y'].mean()  
{c: 0.44 , d: 0}
```

## Unit 04 | 과제 2 (2시간~)

- 아파트가 데이터셋을 이용해서 다음을 해보세요:
  - 결측치 처리 (변수 지정)
  - 범주형 변수 인코딩 (변수 지정)
  - 2차원 feature 시각화 (5개)
  - 3차원 feature 시각화 (3개)
  - 파생변수 2개 생성 후, 설명
- 모든 시각화 결과를 짚막하게 설명해주세요.

Q & A

들어주셔서 감사합니다.