강의 소개

Kyungsik Han

강의 목표

CODING

- 머신러닝 알고리즘에 대한 기초적인 내용을 이해하고, 실습을 통해서 머신러닝 작동 원리에 대해서 학습

- 다음과 같은 학습 결과 기대
  - 머신러닝 학습을 위한 python언어 핵심 요소 이해
  - 필수 지도 및 비지도 머신러닝 알고리즘 이해
  - 실습을 통한 머신러닝 모델 구축 과정 및 프로젝트 진행에 대한 이해
  - 중급 이상의 머신러닝 및 딥러닝 학습을 위한 기반 마련

강의 특징

CODING X

- 머신러닝 학습에 필요한 핵심 내용 파악
- 챕터별 이론과 실습의 유기적인 연결을 통한 기초 필수 머신러닝 개념 이해
- 추가 미니 프로젝트 실습을 통한 머신러닝 활용에 대한 이해

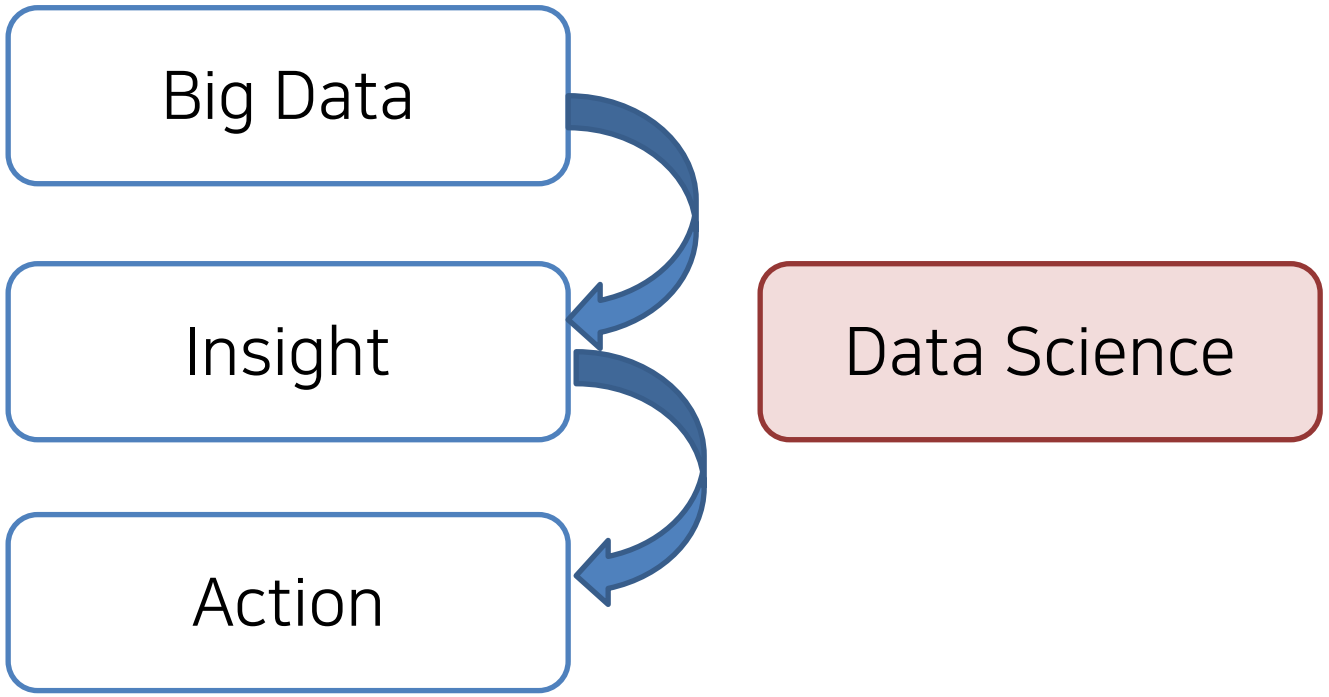
강의자 소개

- 아주대학교 컴퓨터공학 및 인공지능학과 한경식 교수
- 이메일: kyungsikhan@ajou.ac.kr

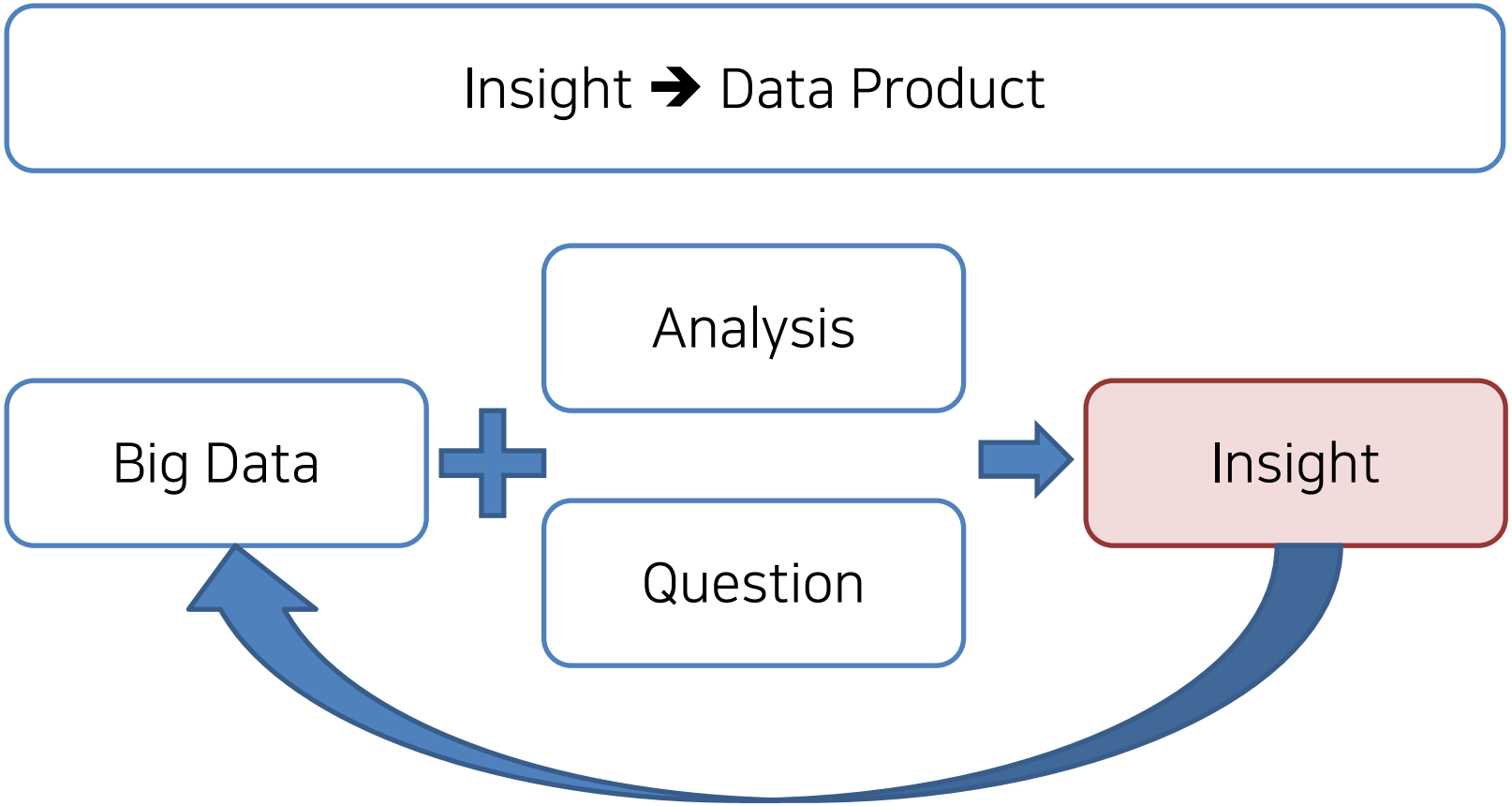
연구실 홈페이지: https://www.ajouhcil.com/

데이터 과학 소개

What is Machine Learning?



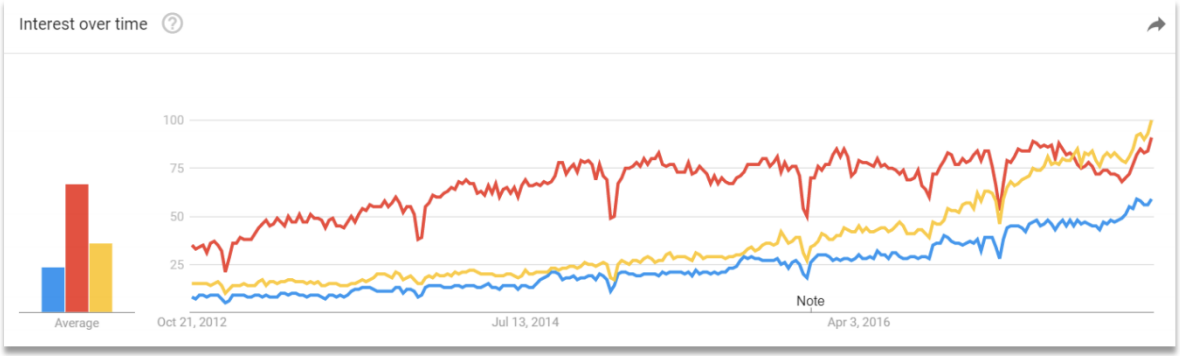
What is Machine Learning?



Trend

CODING

Google Trend: Data Science & Big Data & Machine Learning



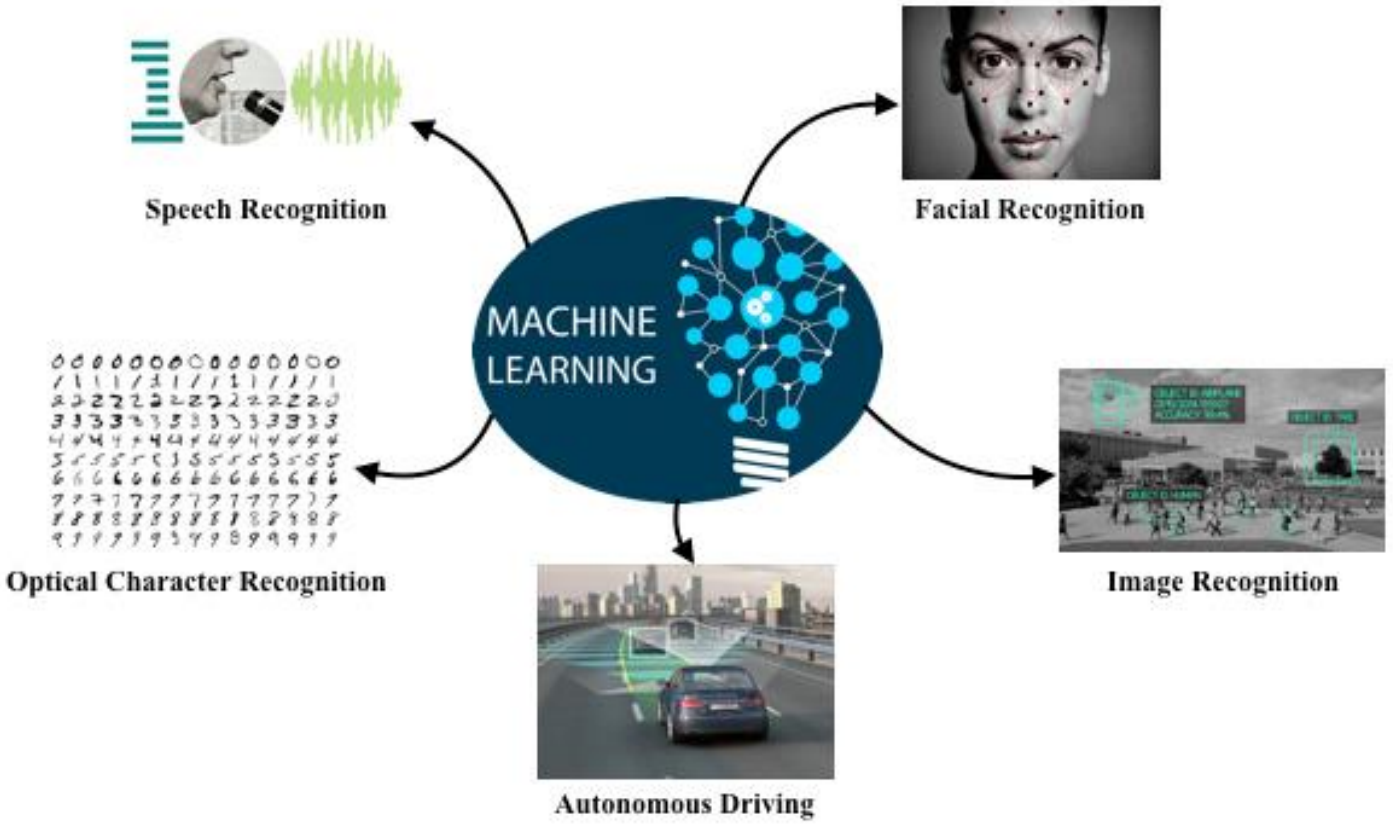
Related queries	
1 computer science	100
2 what is data science	70
3 what is science	70
4 what is data	70
5 data analytics	55

Related queries	
1 big data analytics	100
2 analytics	100
3 data analytics	100
4 hadoop	65
5 big data hadoop	65

Related queries	Top
1 python	100
2 python machine learning	100
3 machine learning pdf	80
4 machine learning google	70
5 machine learning algorithms	60

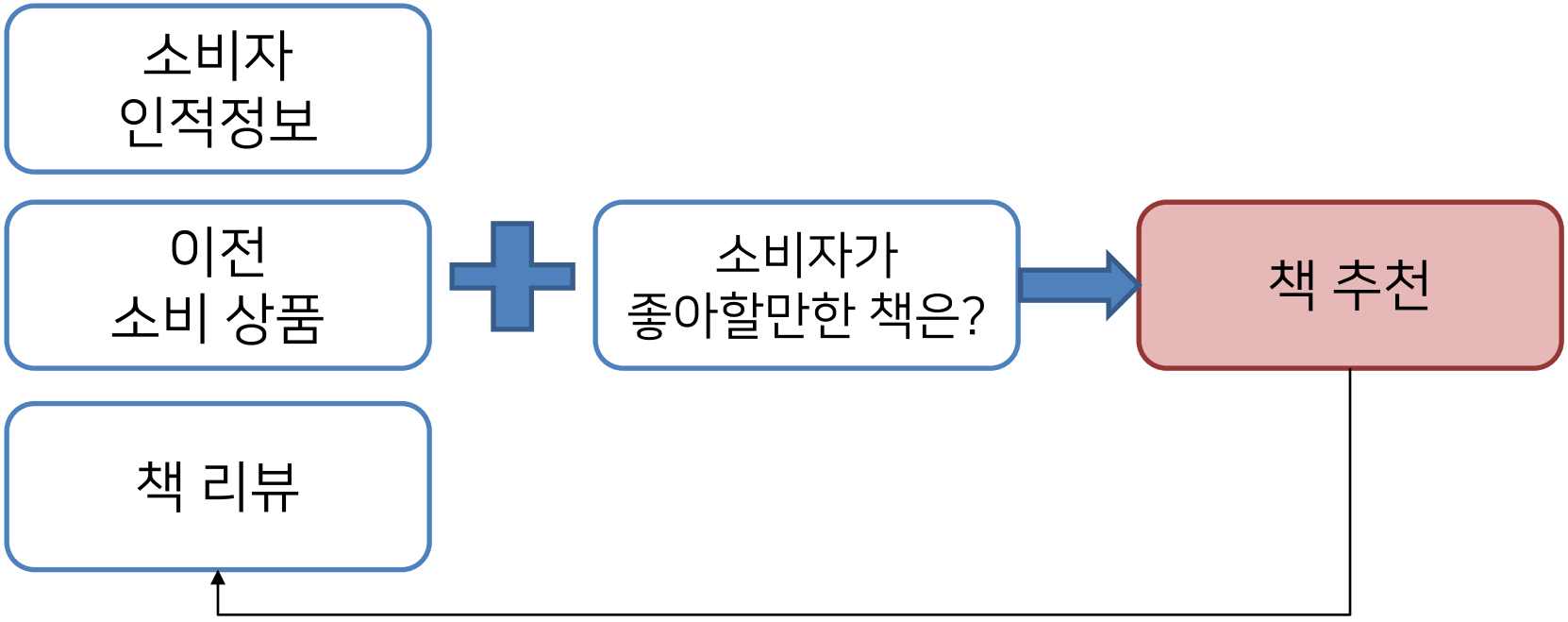
적용 사례

- Web search, Computational biology, Finance, E-commerce, Space exploration, Robotics, Information extraction, Social networks, Debugging
- Your own area!



도메인 예제

Insight ➔ Data Product



amazon.com

도메인 예제

CODING

책 판매를 위한 잠재적 고객 찾기

새로운 책 광고

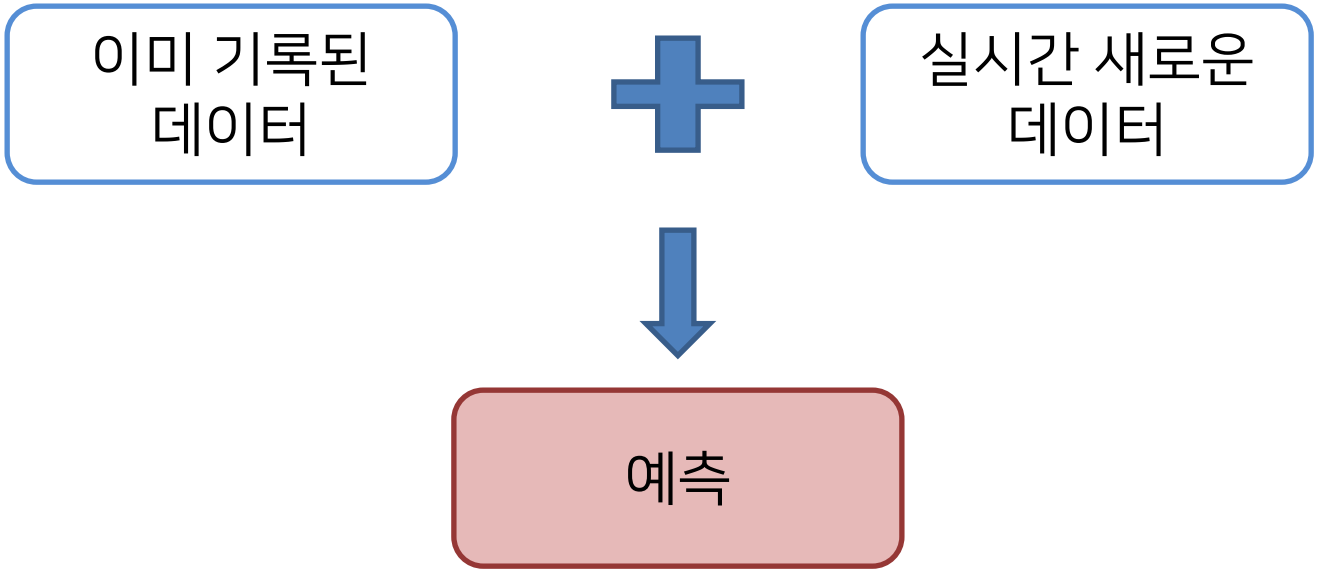
고객 책
선호 모델

새로운 책
정보

이 책을 좋아할
고객은?

추천할만
고객에게 추천

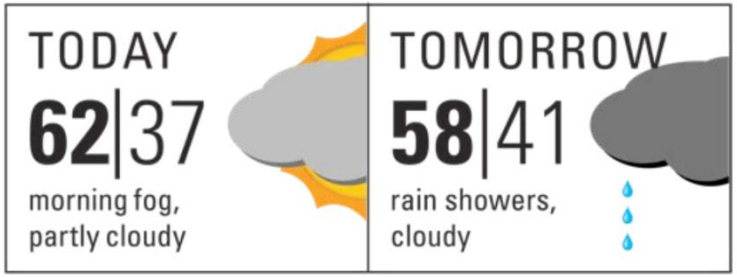
실행할 수 있는 정보



amazon.com

실행할 수 있는 정보

예측



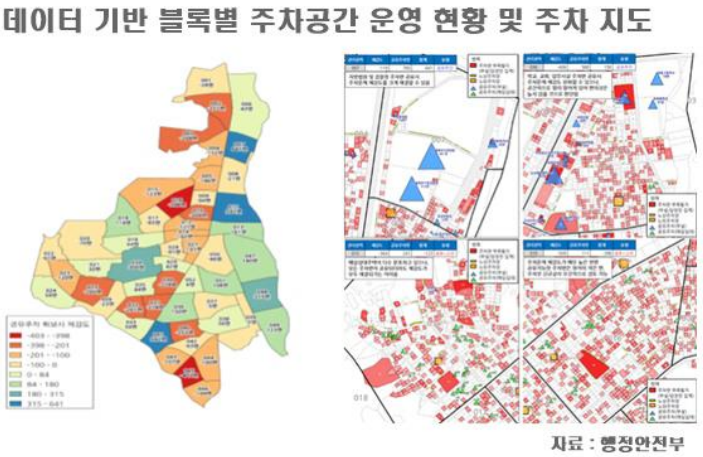
실행

Data Everywhere

CODING

Smartphone, Tablets, PC, SNS, Game, …

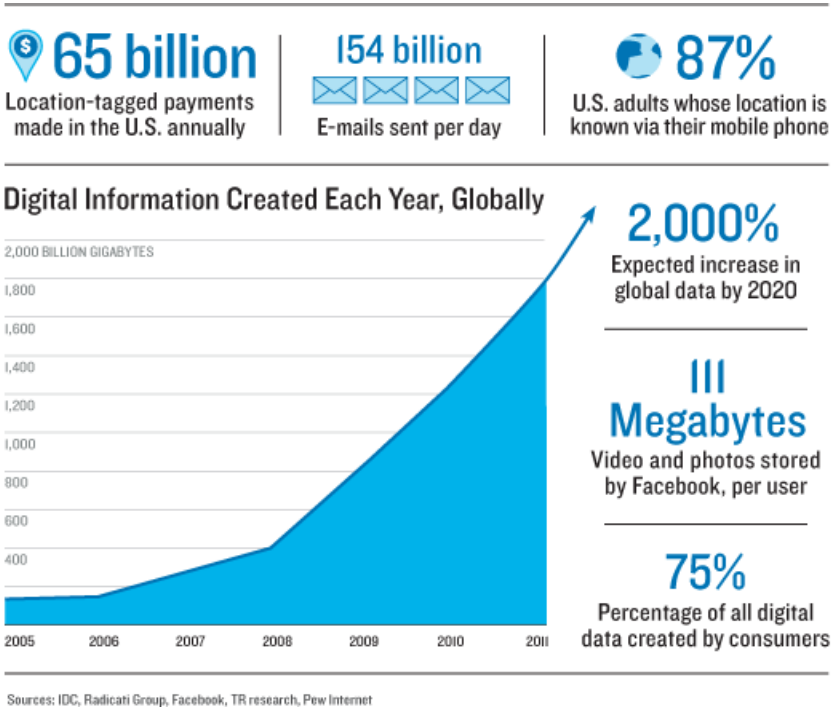
공공 빅데이터



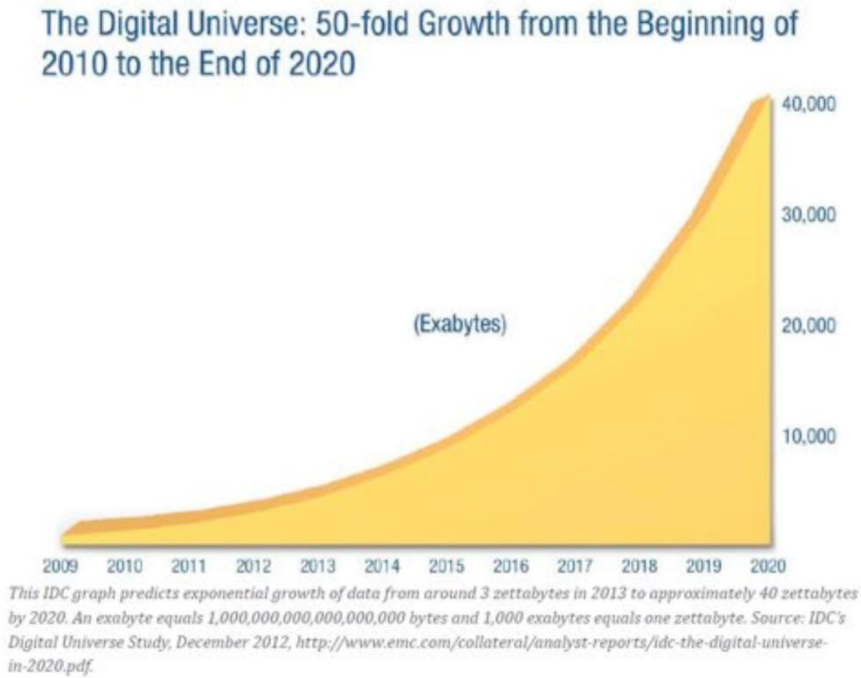
의료 데이터

금융 데이터

기하급수적으로 증가하는 데이터량

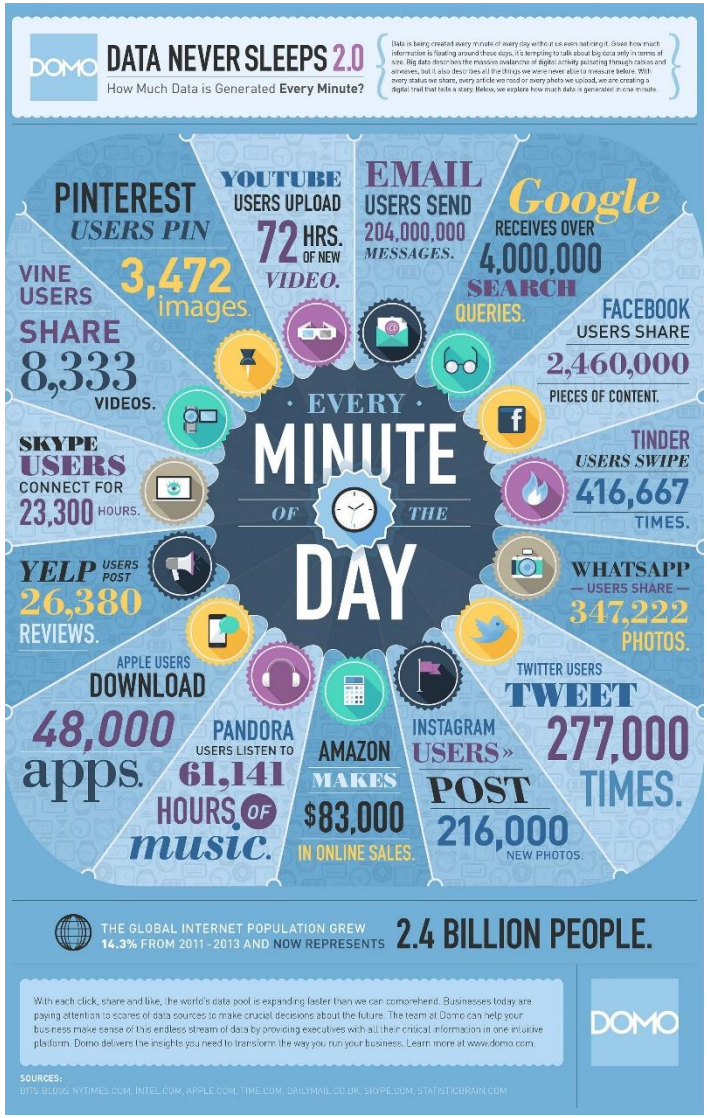


Source: IDC's Digital Universe Study

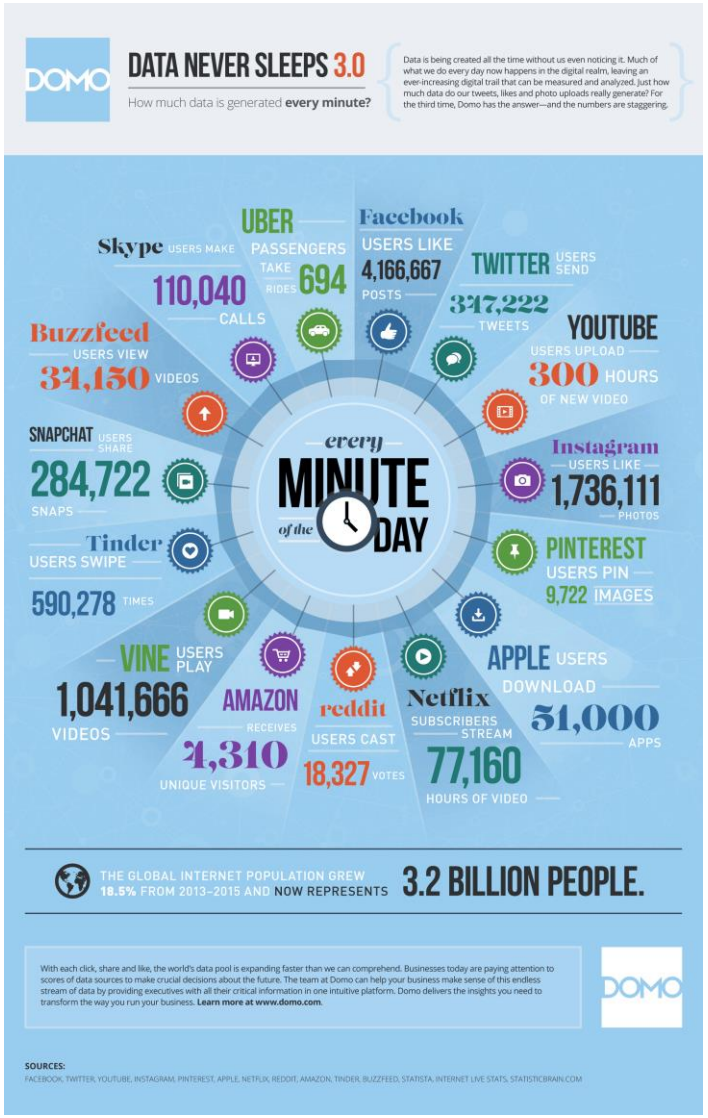


Source: IDC's Digital Universe Study

생산되는 데이터 량 (분당)



Source: DOMO

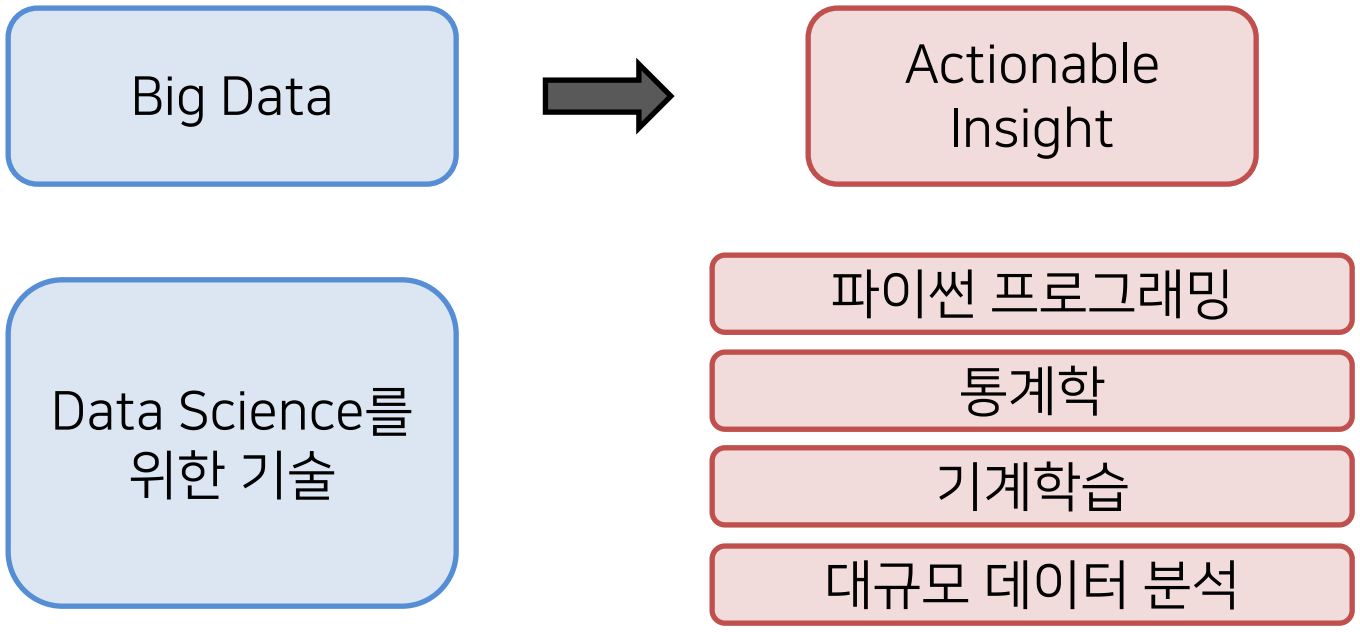


Source: DOMO

데이터 과학을 위한 기술

CODING

- Python (파이썬) 프로그래밍
- 통계학
- 기계학습
- 대규모 데이터 분석

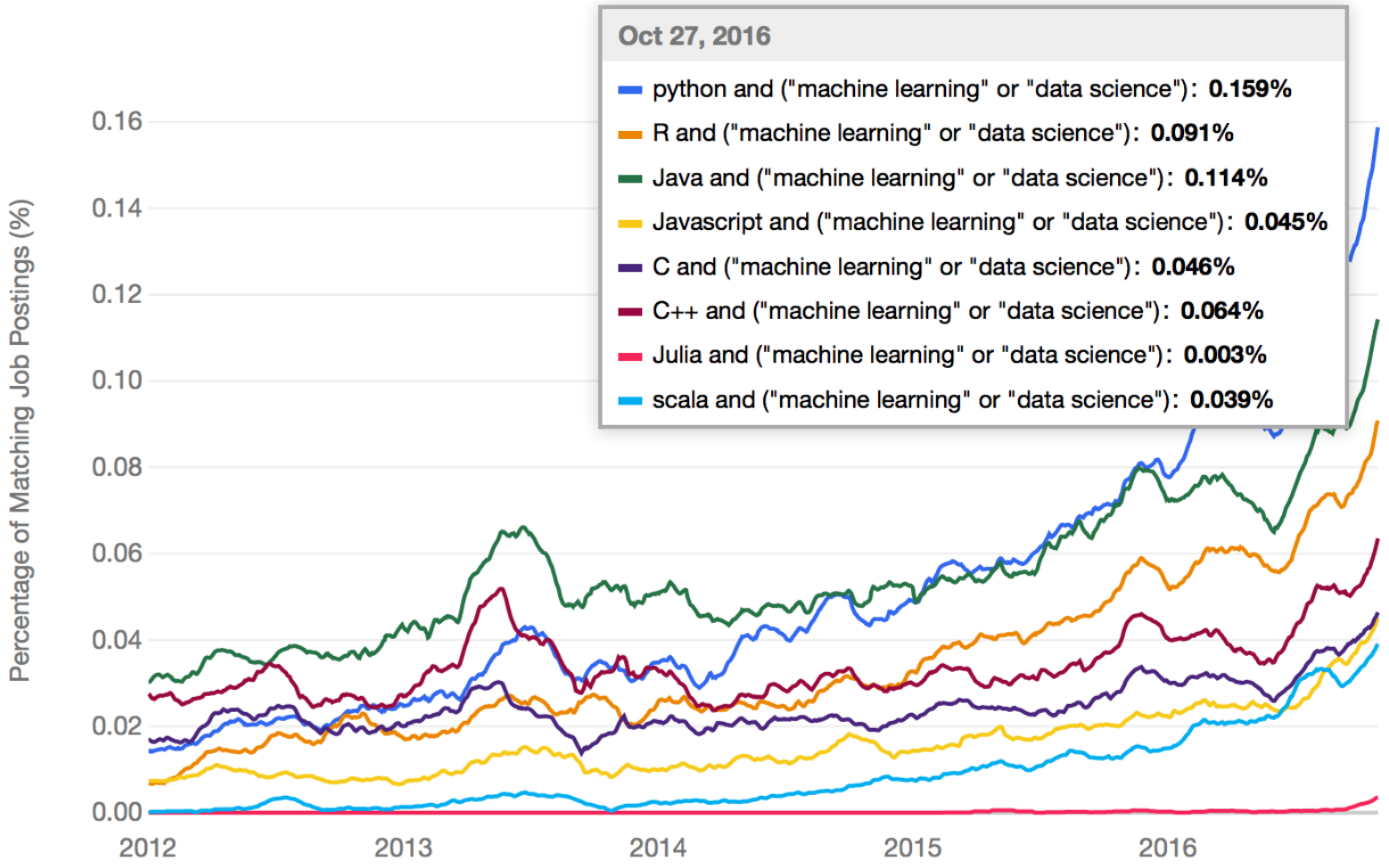


데이터 과학

CODING ✖

- 데이터에 대한 열정
- 문제를 분석으로 연결
- 공학적인 솔루션
- 호기심을 보여줌
- 팀원들과의 커뮤니케이션

데이터 과학: 언어



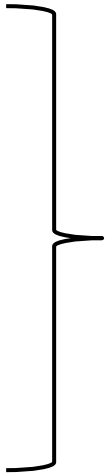
https://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html

데이터 분석 모델

CODING

- 통계 분석

- 군집 분석
- 연관 분석
- 분류 분석
- 회귀 분석



데이터마이닝, 머신러닝, 딥러닝

- 시각화

파이썬을 쓰는 이유

- 읽고 배우기가 쉬움
- 활발한 온라인 커뮤니티
- 지속적인 라이브러리 업데이트
- 새로운 라이브러리 제공
  - 데이터 관리, 분석
  - 시각화
- 데이터사이언스를 위한 모든 단계에 최적화
- Notebooks

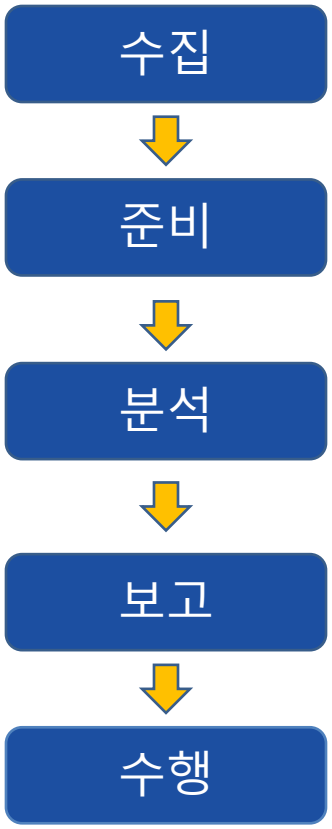
Example
(축구 데이터 분석)

케이스 스터디 1: 축구 데이터

- 의미있는 축구 그룹 생성
- 내가 선호하는 선수와 비슷한 성향을 가진 선수 찾기
- 분석을 통한 팀 구성

데이터사이언스 프로젝트에서의 중요 단계

CODING



정제되지 않은 데이터(raw data)를 수집

데이터 확인, 시각화, 전처리 작업 진행

특성 선택, 모델 선택, 결과 분석

결과 보고 (제 3자가 이해할 수 있도록)

결과 내용을 실제로 사용할 수 있도록 함

다양한 데이터 수집

CODING

- 데이터베이스
  - 관계형 데이터베이스
  - 비관계형 데이터베이스 (NoSQL)
- 텍스트 파일
  - CSV, Text
- 실시간 피드
  - 센서 데이터
  - 온라인 플랫폼
    - 트위터
    - 실시간 날씨 정보

데이터 준비: 통계를 활용

CODING

```
In [8]: df.describe().transpose()
```

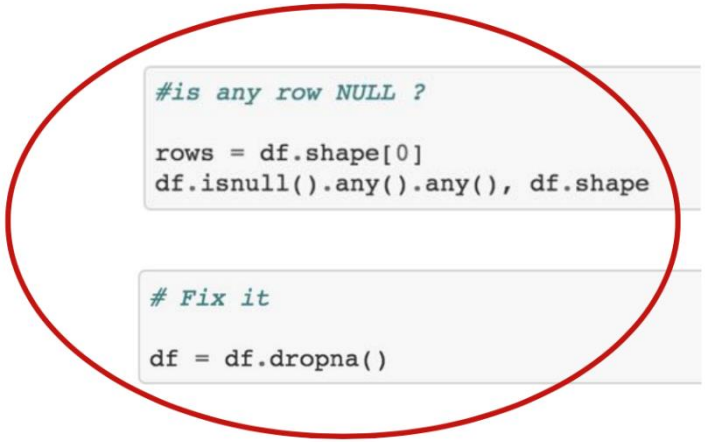
Out[8]:

	count	mean	std	min	25%	50%	75%	max
id	183978.0	91989.500000	53110.018250	1.0	45995.25	91989.5	137983.75	183978.0
player_fifa_api_id	183978.0	165671.524291	53851.094769	2.0	155798.00	183488.0	199848.00	234141.0
player_api_id	183978.0	135900.617324	136927.840510	2625.0	34763.00	77741.0	191080.00	750584.0
overall_rating	183142.0	68.600015	7.041139	33.0	64.00	69.0	73.00	94.0
potential	183142.0	73.460353	6.592271	39.0	69.00	74.0	78.00	97.0
crossing	183142.0	55.086883	17.242135	1.0	45.00	59.0	68.00	95.0
finishing	183142.0	49.921078	19.038705	1.0	34.00	53.0	65.00	97.0
heading_accuracy	183142.0	57.266023	16.488905	1.0	49.00	60.0	68.00	98.0
short_passing	183142.0	62.429672	14.194068	3.0	57.00	65.0	72.00	97.0
volleys	181265.0	49.468436	18.256618	1.0	35.00	52.0	64.00	93.0
dribbling	183142.0	59.175154	17.744688	1.0	52.00	64.0	72.00	97.0

데이터 정리

CODING

- 데이터 정리가 필요한 이유?
  - 누락된 데이터
  - 필요없는 데이터
  - NULL 데이터
- 데이터 정리 방법
  - 정리하고자하는 데이터 삭제
  - Impute these entries with a counterpart
    - 평균값
    - 0, -1 과 같은 값

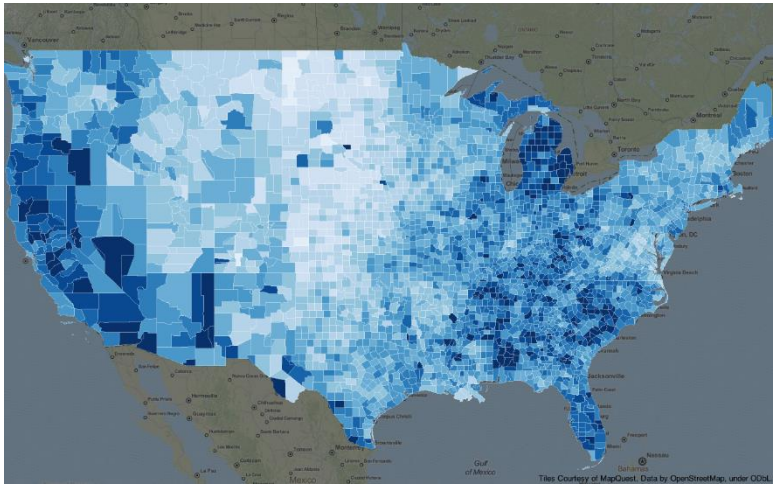


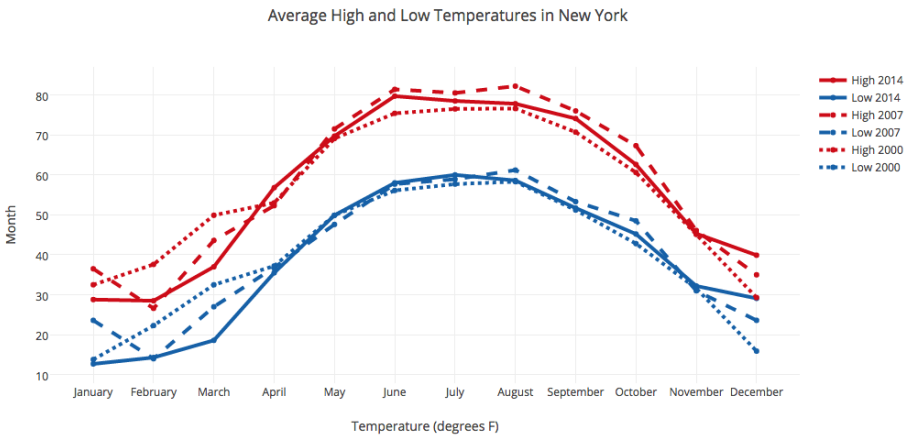
```
#is any row NULL ?

rows = df.shape[0]
df.isnull().any().any(), df.shape
```

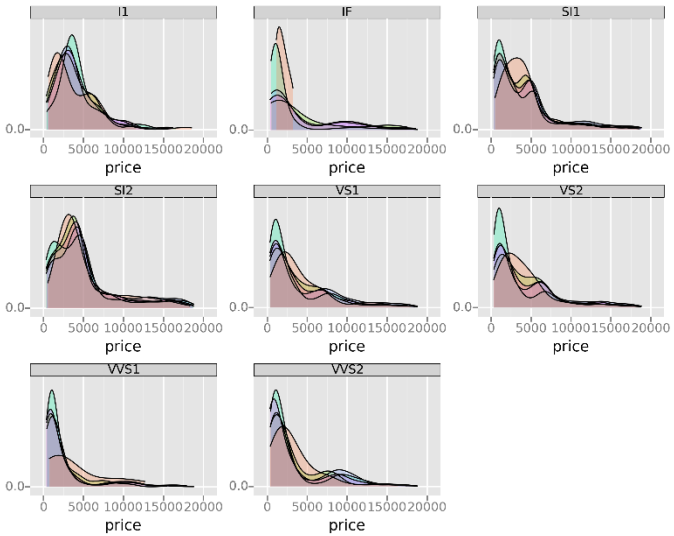
```
# Fix it

df = df.dropna()
```

데이터 시각화

CODING



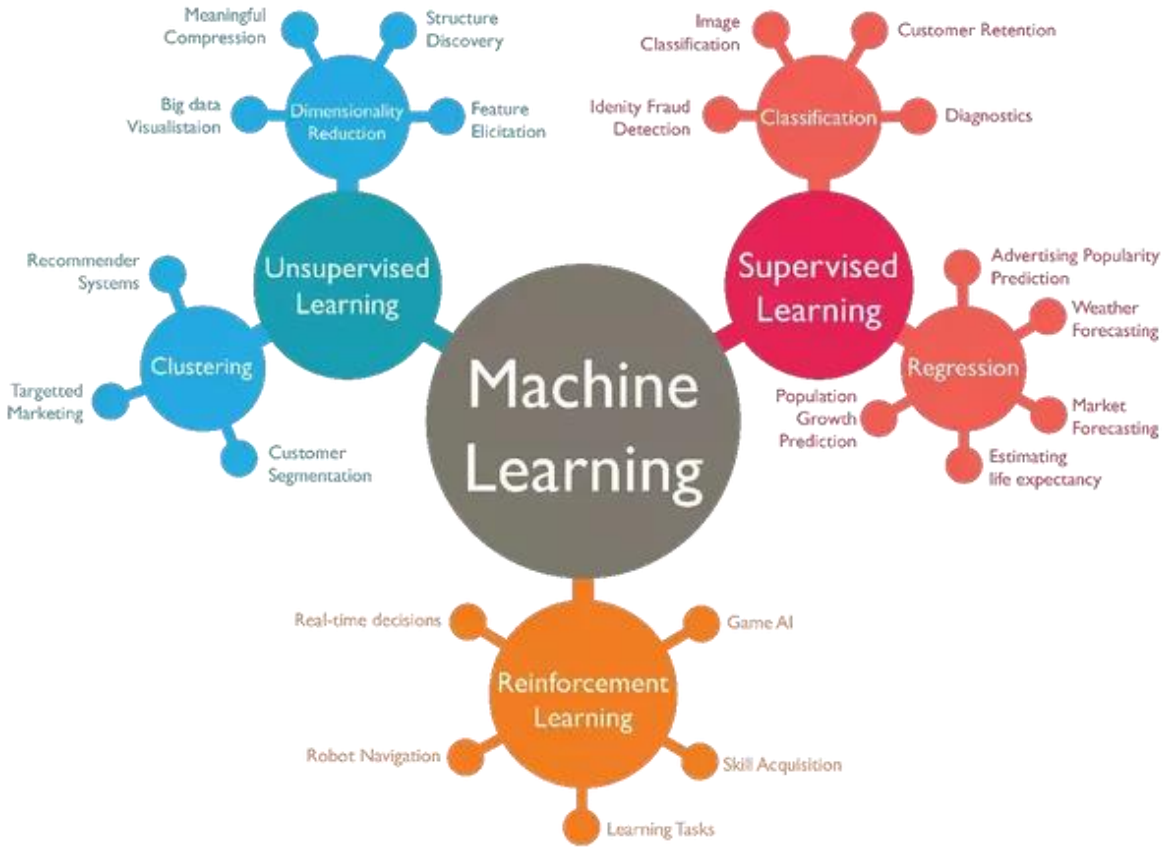




시각적인 분석 가능

모델링을 위한 알고리즘 선택

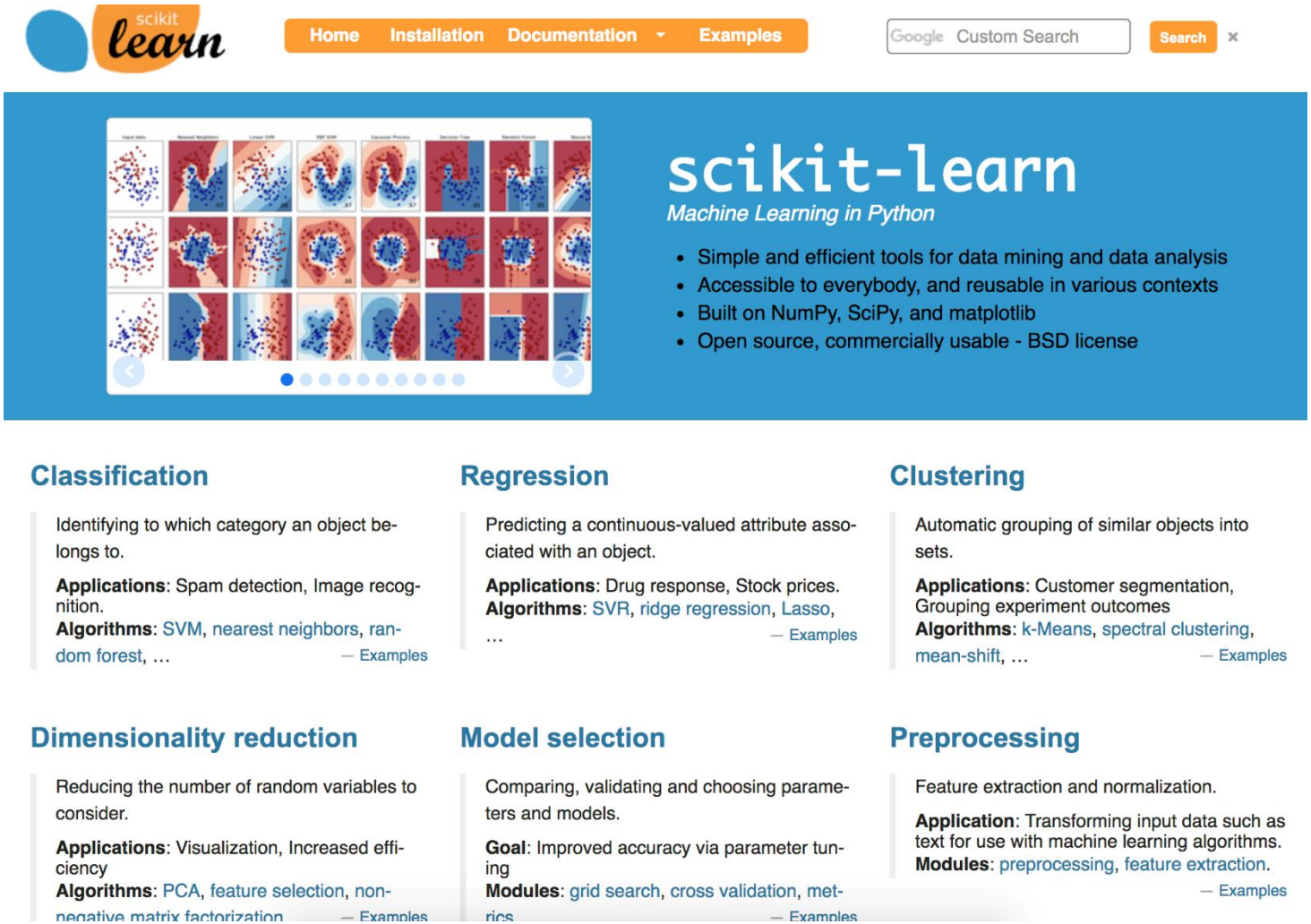
- 지도 학습(Supervised Learning)
- 비지도 학습(Unsupervised Learning)



라이브러리 활용

CODING

- scikit-learn (http://scikit-learn.org/stable/)



축구 데이터 분석

CODING

- 어떤 내제적인 요소를 바탕으로 축구 선수들을 그룹핑할 것 인가?
  - Agility
  - Reaction Time
  - Shot Power
  - Sprint Speed
  - Hair Style?
  - Movies the player likes?
- 기존 요소를 바탕으로 새로운 요소를 만들 수 있음
  - f(shot power, reaction time)

축구 데이터 분석: 군집화 (Clustering)

CODING

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

http://scikit-learn.org/stable/modules/clustering.html

파이썬에서 k-Means clustering

CODING

```
from sklearn.cluster import Kmeans

…

Y = KMeans(n_clusters=3,random_state=random_state).fit_predict(X)

…
```

군집화 결과 해석

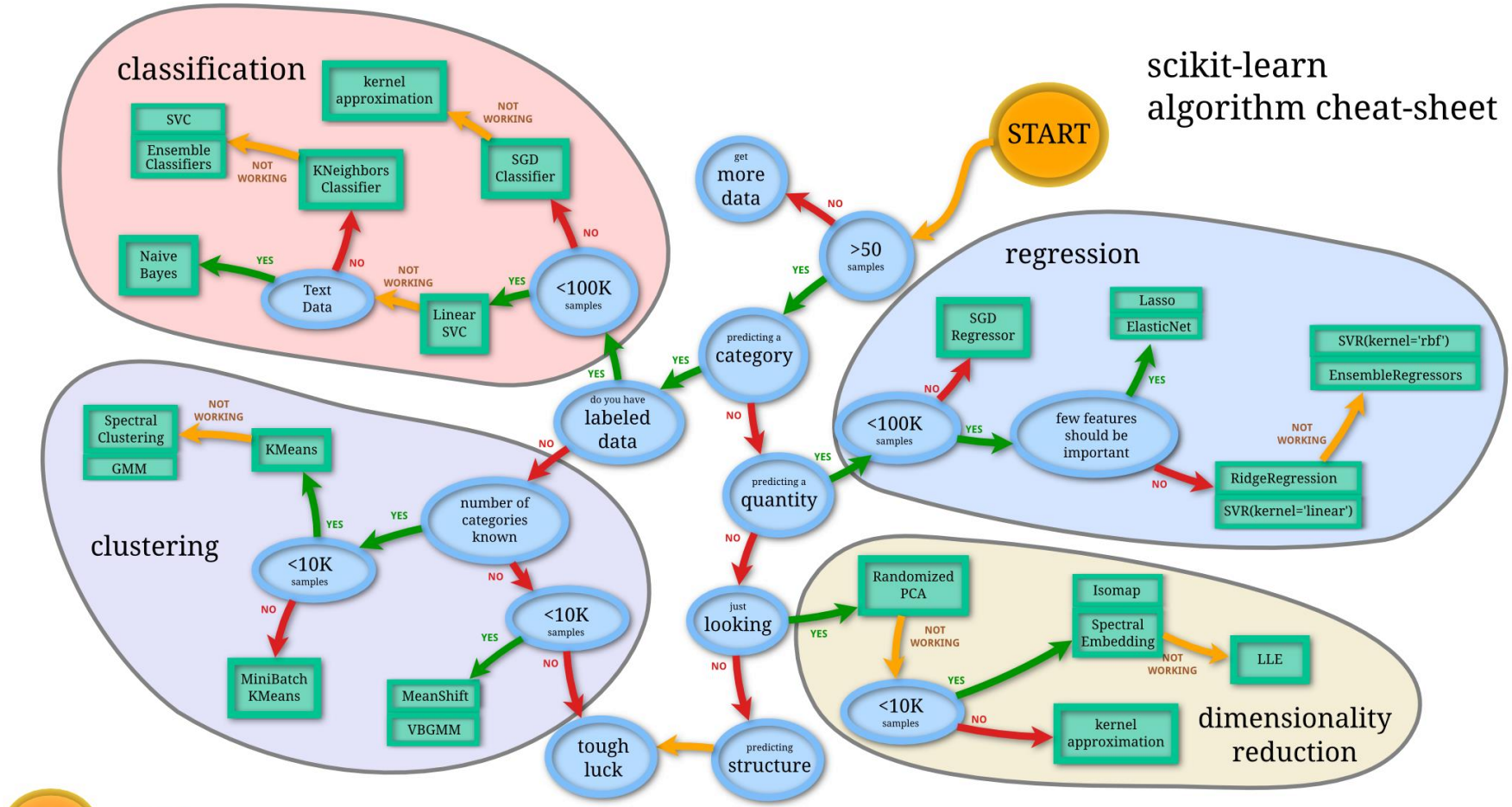
CODING

- 한 그룹당 선수는 몇 명인가?
  - 그룹이 너무 적은가? 너무 많은가?
  - 그룹을 나누는 기준은 무엇인가?

- 군집화에 사용했던 요소를 변경해주어 새로운 요소를 도출
  - 새로운 요소는 군집화 결과에 어떤 영향을 미치는가?
  - 추가적인 데이터 수집이 필요한가?

좋은 알고리즘을 찾는 방법

CODING



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

다음 영상에서 배울 내용

CODING

- 데이터 관련 내용 학습
  - 수집
  - 종류
  - 형식
  - 타입
  - 행렬
  - 품질

수고하셨습니다.