Big Data

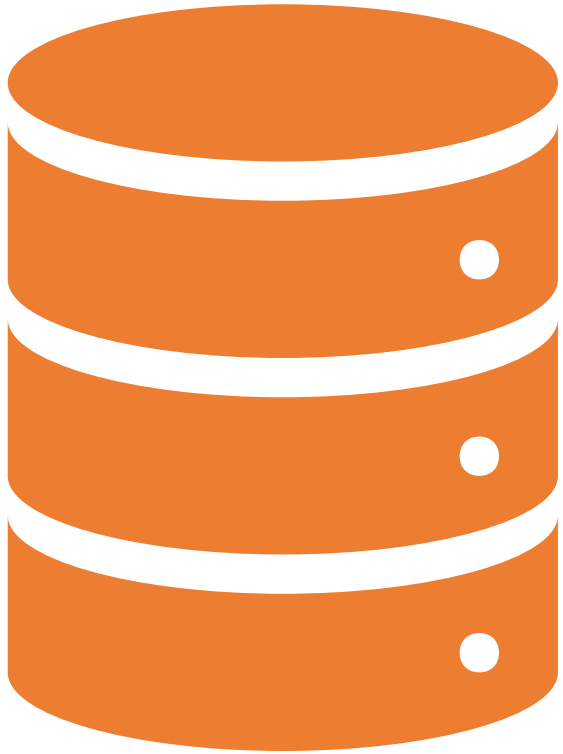
Data Analysis

임보미 선임연구원

Data Analasis

Big Data란?

100 TB



1 MB

Data Analasis

Big Data란?

빅데이터란 기존 데이터베이스 관리도구의 능력을 넘어서는 대량의 정형 또는 데이터베이스 형태가 아닌 비정형의 데이터 집합을 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술이다.

Data Analasis

Big Data란?

- 데이터 저장
- 데이터 관리
- 데이터 추출
- 데이터 처리
- 데이터 시각화
- 데이터 분석

Data Analasis

Big Data란?

1990~2000

컴퓨터보급

엑셀,
데이터베이스

2000~2010

인터넷보급
웹기반데이터

야후,아마존 등

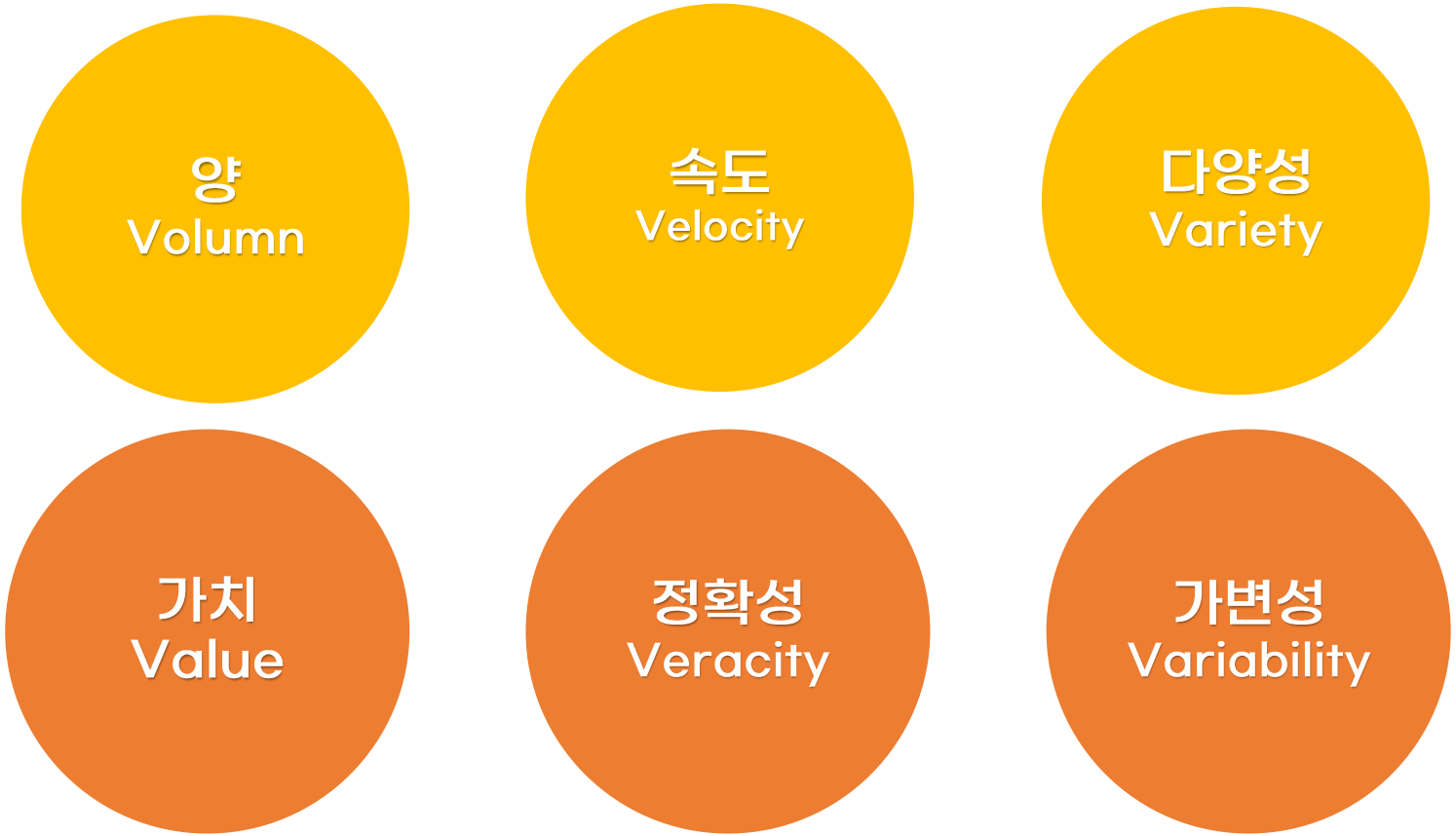
2010~

모바일
IoT기기발달

구글,페이스북,
인스타그램

Data Analasis

Big Data 특징 (3V+3V)



Data Analasis

Big Data 특징 (3V+3V)

대량의 데이터

빠른데이터 유입
실시간 처리속도

다양한 데이터

Data Analasis

Big Data 특징 (3V+3V)

데이터 가치

데이터 품질

데이터 의미

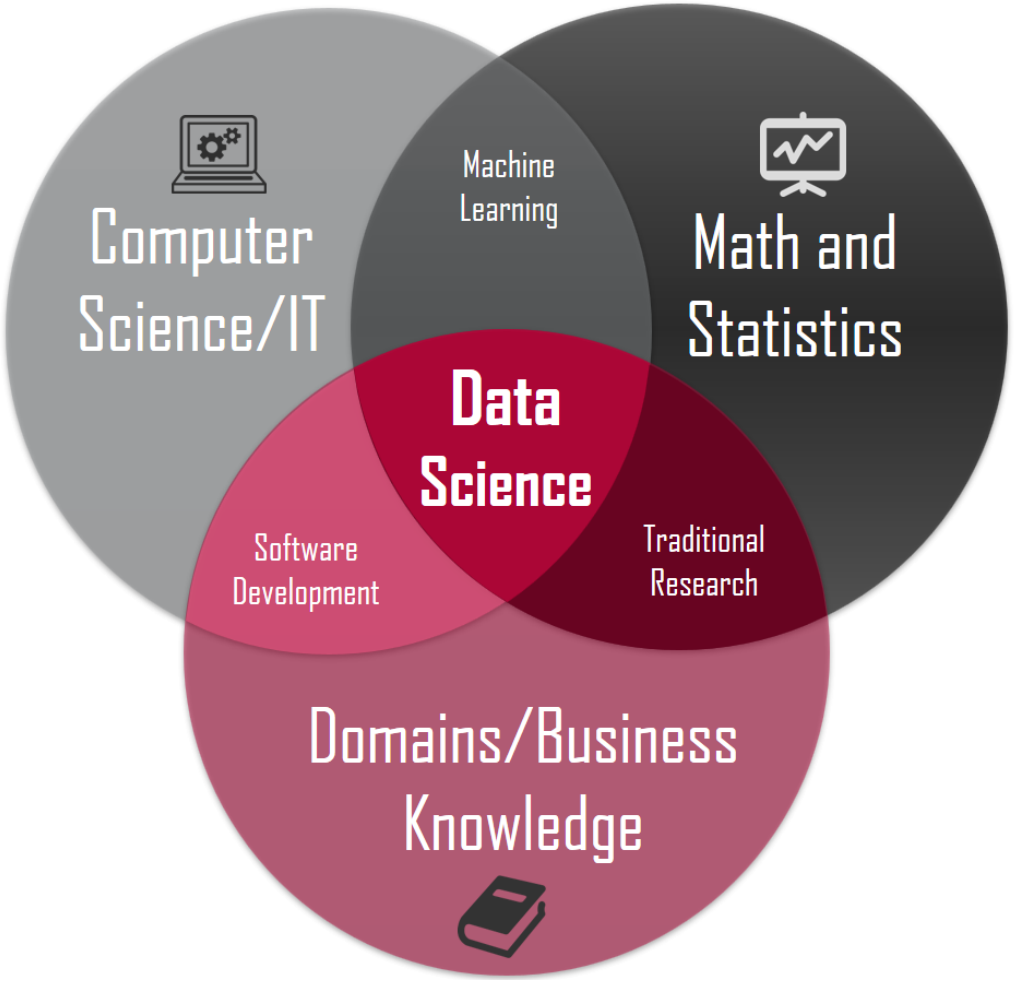
Data Analasis

Data Science

- 데이터를 단순히 분류하거나 분석하는 것 말고 데이터 속에 담긴 패턴이나 미래 예측에 도움이 되는 신호를 찾는 것
- 통계학 : 표본 조사를 통해 전체를 추론하고 예측
- 데이터과학 : 통계학의 분석방법론, 머신러닝, 딥러닝, 텍스트마이닝, 추천시스템 등

Data Analasis

Skillset



Data Analasis

Data Science Process

Data Analasis

Collecting data sets

- Database
- File(CSV,XML,JSON)
- Web crawling
- IoT sensor data
- Survey

Data Analasis

Preprocessing

- 결측치 처리 : 데이터 삭제, 다른 값으로 대체(최대값,최소값,중앙값 등), 예측 모델을 활용한 값 삽입
- 이상치 처리 : 입력오류(데이터 삭제, 다른 값으로 대체), 자연발생(feature 추가)
- Feature Engineering : scaling(feature의 단위를 변경), binning(수치형 -> 범주형),
  transform(feature를 분리하거나 연산 - 날짜,주중/주말),
  encoding(범주형 -> 수치형)

Data Analasis

EDA (Exploratory Data Analysis)

- 각 변수가 어떤 의미인지 파악 : meta data, data description
- 범주형(categorical data) : 성별,혈액형,성공여부(nominal data), 학점,지역
- 수치형(numerical data) : 발생횟수,학급인원(discrete data), 키,몸무게,혈압
- 기술통계 : 최대값, 최소값, 최빈값, 평균값, 중앙값, 분산, 표준편차, 사분위수
- 변수간 상관관계, 독립여부 확인

Data Analasis

Machine Learning

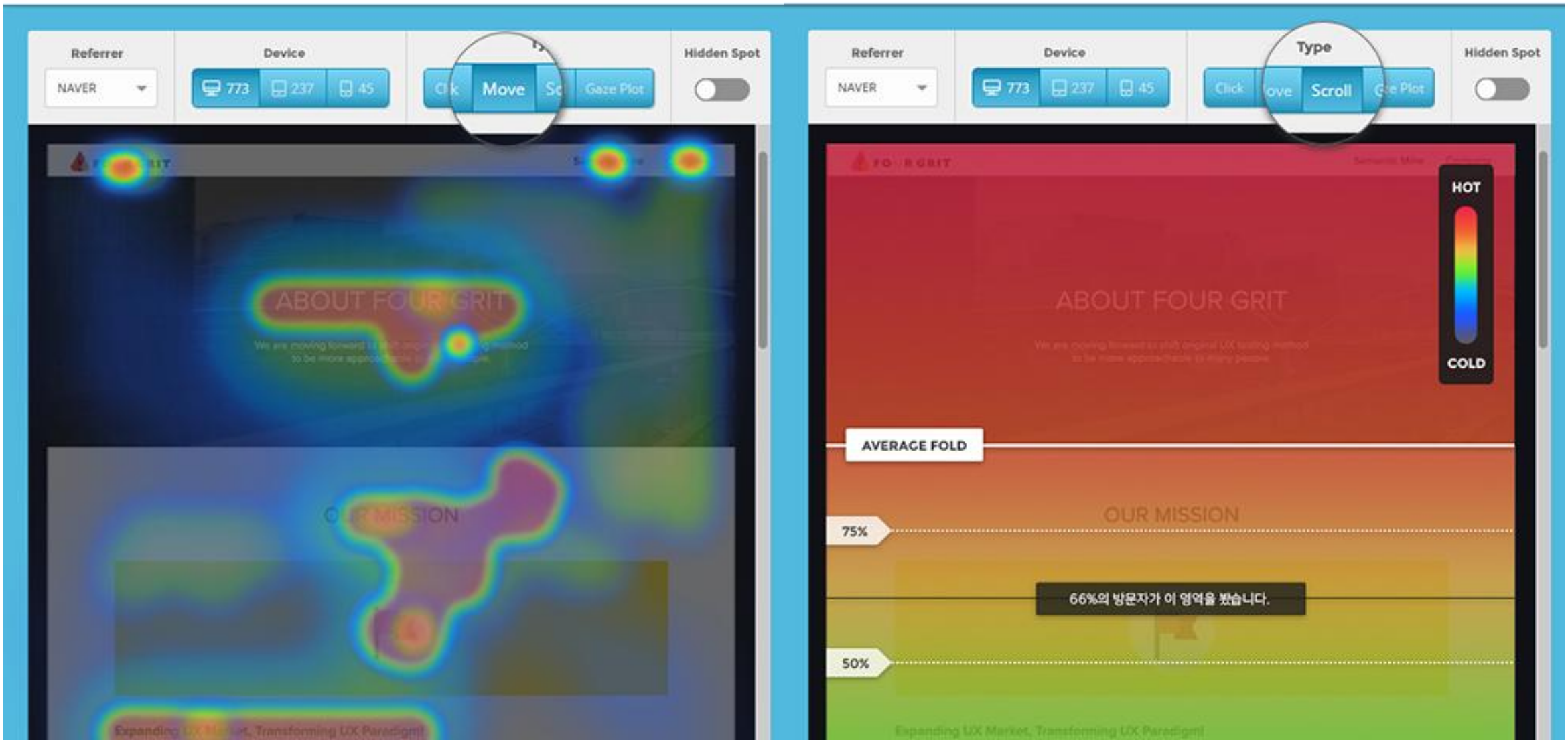
- Supervised learning(지도학습) : Linear regression, SVM, Decision tree, KNN, Ensemble model etc.
- Unsupervised learning(비지도학습) : Clustering, Dimensionality reduction etc.

Data Analasis

Reporting & Business service

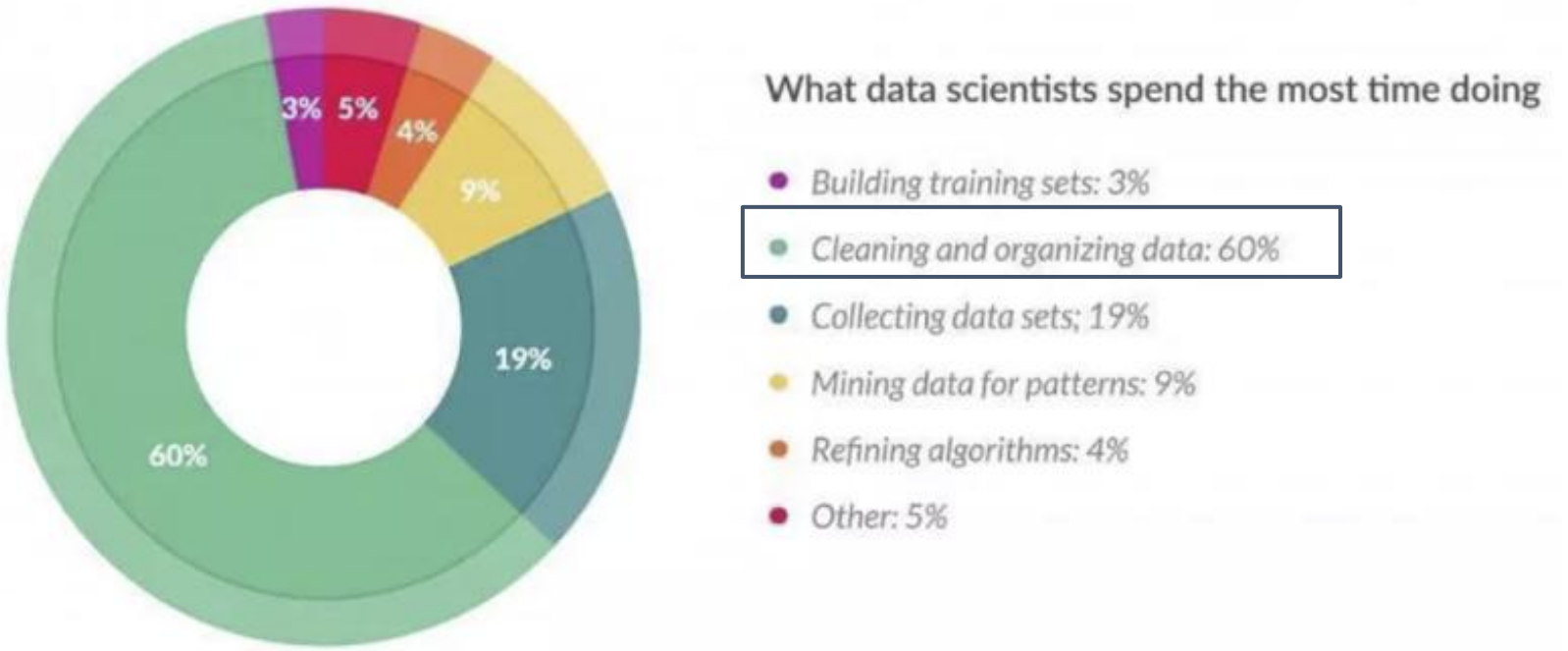
- 시각화 : scatter plot, bar chart, pie chart, box plot, histogram, heatmap
- 서비스화 : Web Service, Mobile Service, Desktop Application

Data Analasis



Data Analasis

데이터의 중요성



- Forbes에서 인용한 CrowdFlower의 설문 결과에 따르면 데이터분석가는 80% 이상의 시간을 데이터 수집/전처리 과정에 사용
- Garbage in garbage out : 좋은 자료를 모으고, 적정하게 정리하여 넣지 않으면 가치를 발견하기 어려움

Data Analasis

단계별 python 패키지

Collecting:

- BeautifulSoup, Selenium, PyMySQL, PyMongo etc.

Preprocessing:

- Numpy, Pandas etc.

Explorary Data Analysis:

- Numpy, Pandas, Matplotlib, Seaborn etc.

Machine Learning:

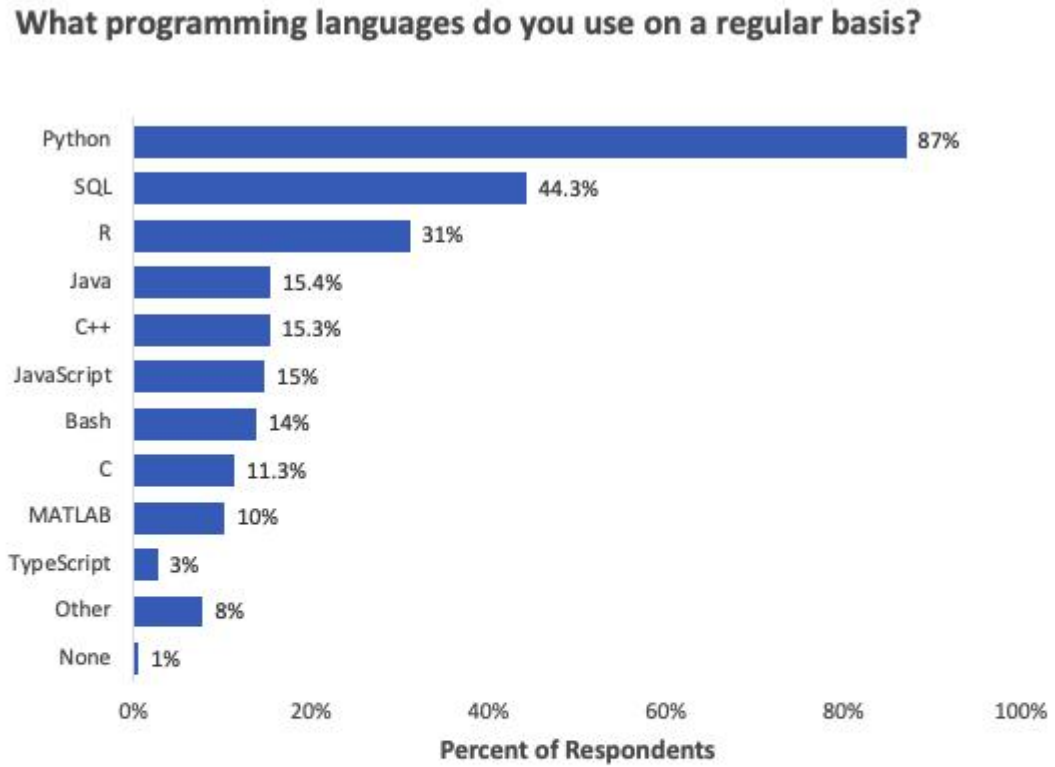
- Scikit Learn, SparkML, Tensorflow, Keras, Pytorch etc.

Communication/Reporting/Building Data Product:

- Matplotlib, Seaborn, Django, Flask etc.

Data Analasis

Kaggle 사이트에서 data scientists에게 주로 사용하는 tool에 대해 설문조사 진행 (2020년)



Note: Data are from the 2019 Kaggle ML and Data Science Survey. You can learn more about the study here: https://www.kaggle.com/c/kaggle-survey-2019.
A total of 19717 respondents completed the survey; the percentages in the graph are based on a total of 14762 respondents who provided an answer to this question.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Big Data

Numpy

임보미 선임연구원