빅데이터 분석 프로그래밍

Introduction

2021-Spring 서중원

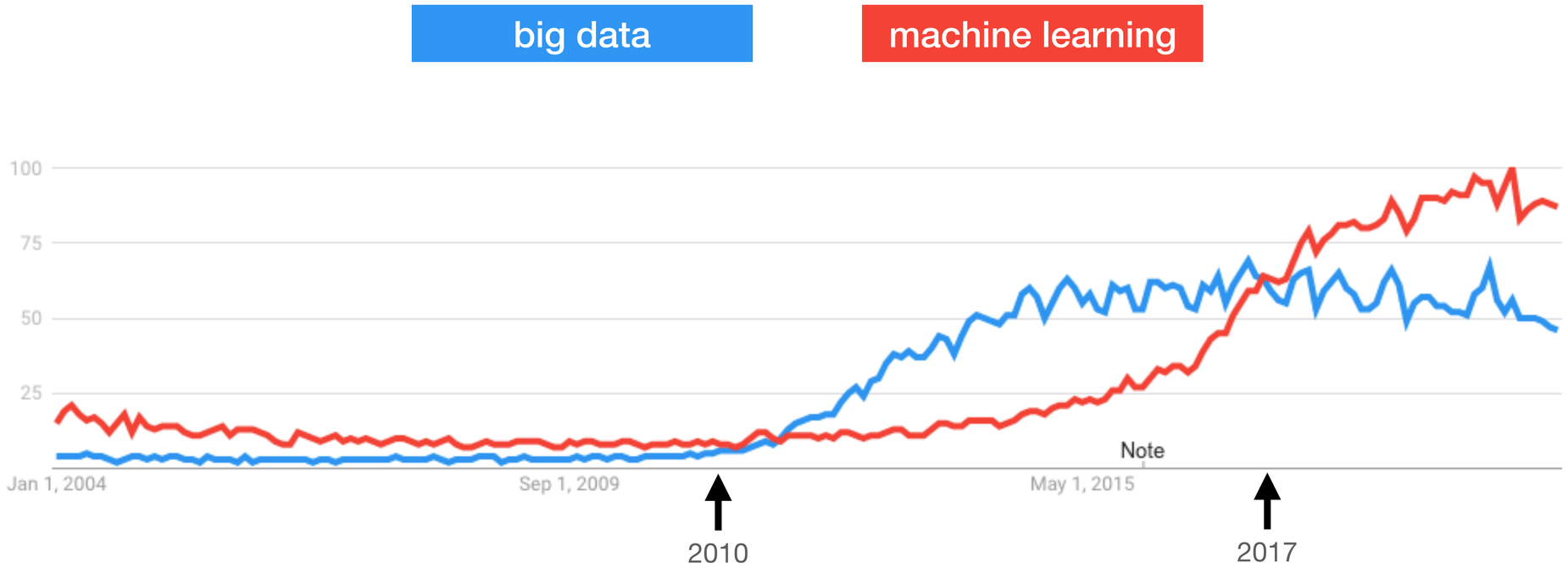
강사소개

서중원

https://github.com/thejungwon

- 경력
  - 연세대 정보대학원 강사 (2019-)
  - IT교육 및 컨설팅 업체 CodeVinci 대표
  - 스타트업 BeBridge CTO
  - 스타트업 코스폴 Lead Developer
- 학력
  - University of Stavanger, 컴퓨터과학 석사
  - 연세대학교 컴퓨터과학 학사
- 기타
  - AWS (Seoul), Equinor (Norway) 인턴
  - NASA-Yonsei 큐브위성 프로젝트 개발자

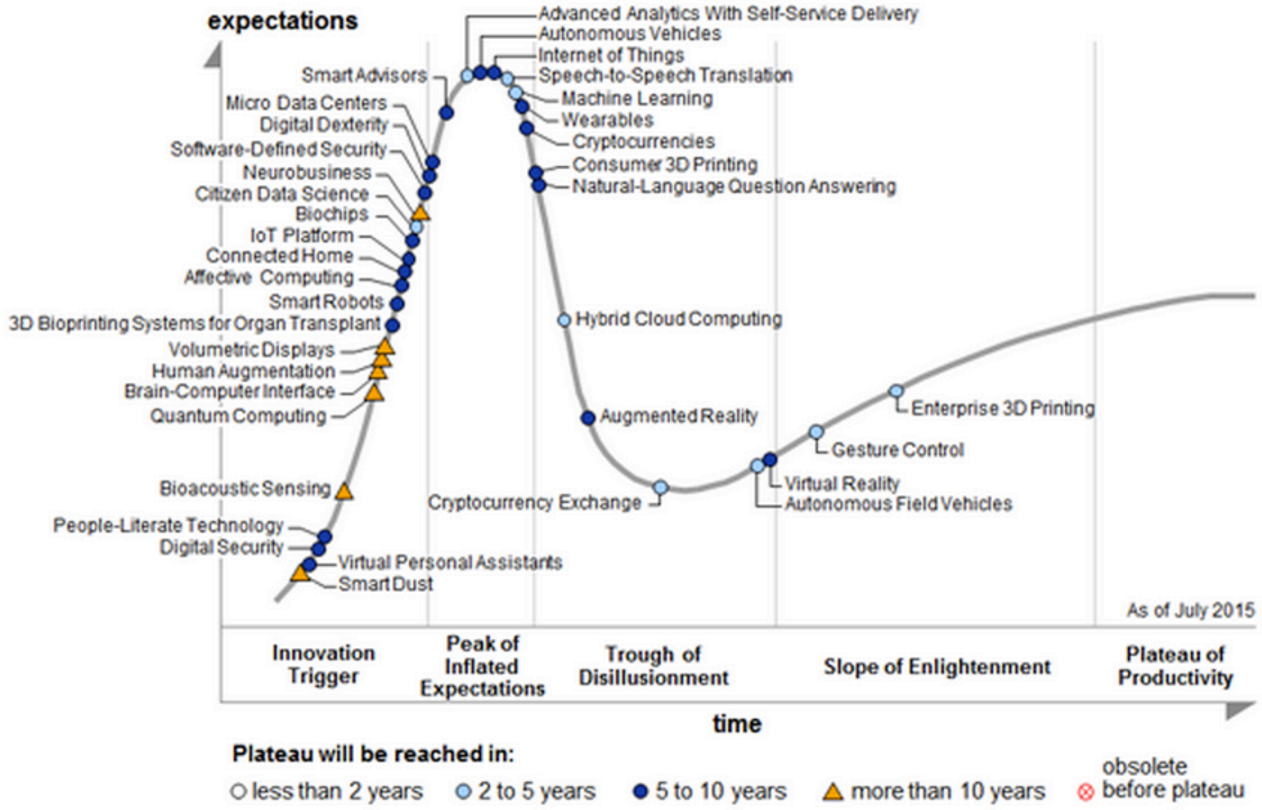
Big Data and Machine Learning



Source: Google Trend, 2020-08

빅데이터의 시대는 끝난건가?

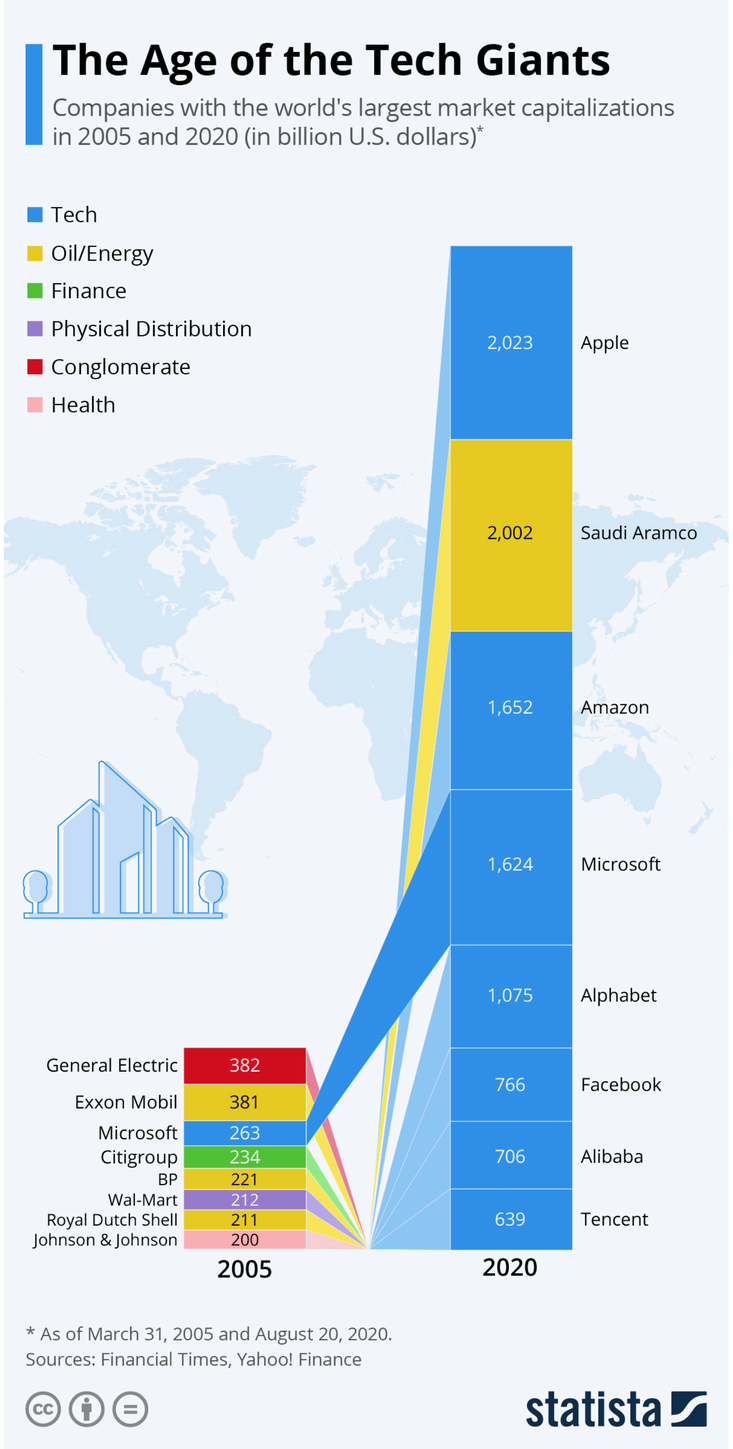
Figure 1. Hype Cycle for Emerging Technologies, 2015



> **"Big data…has become prevalent in our lives"**
> **-- Betsy Burton, Gartner Analyst**

https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/

주 산업이
급격하게 변하고 있다.



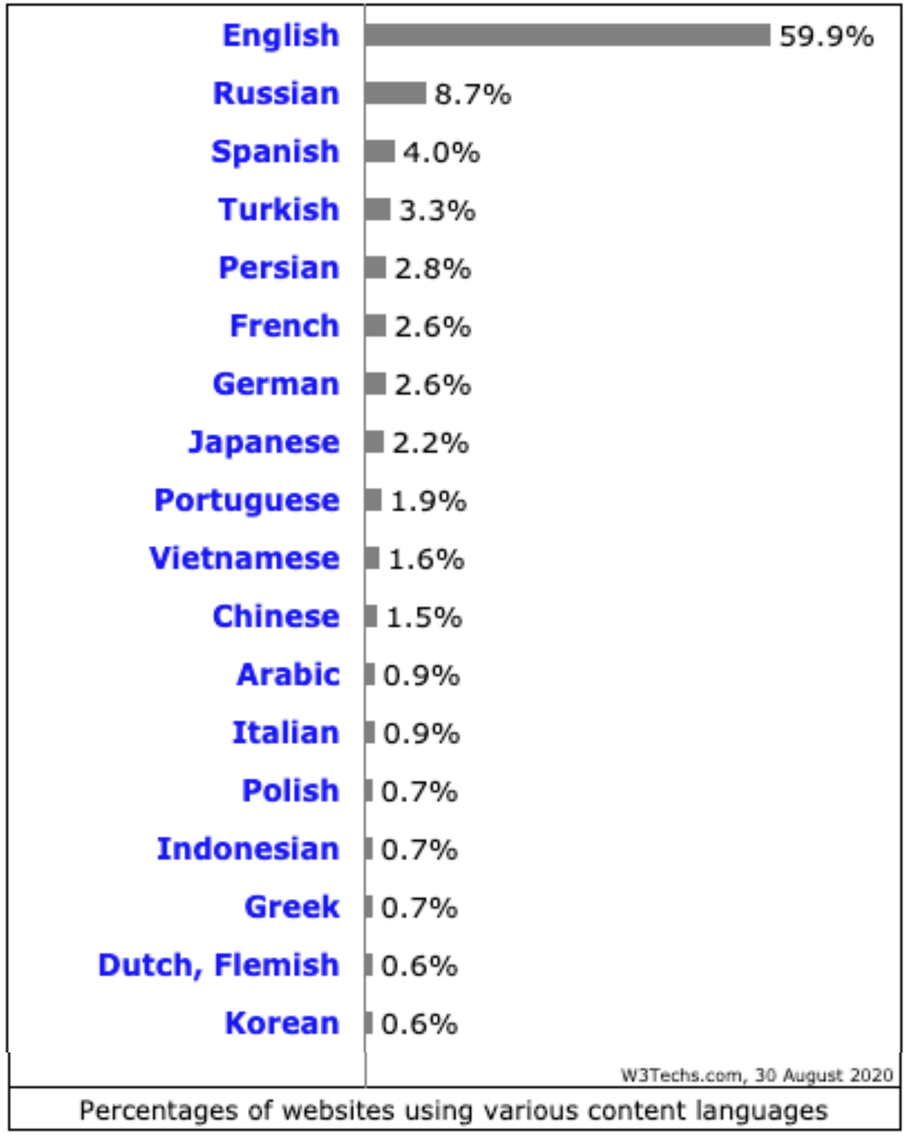
* Martin Armstrong, “The Age of the Tech Giants”, Statista, 2020-08-25

4차 산업혁명에서의 인재상은?

Required Skills: Programming

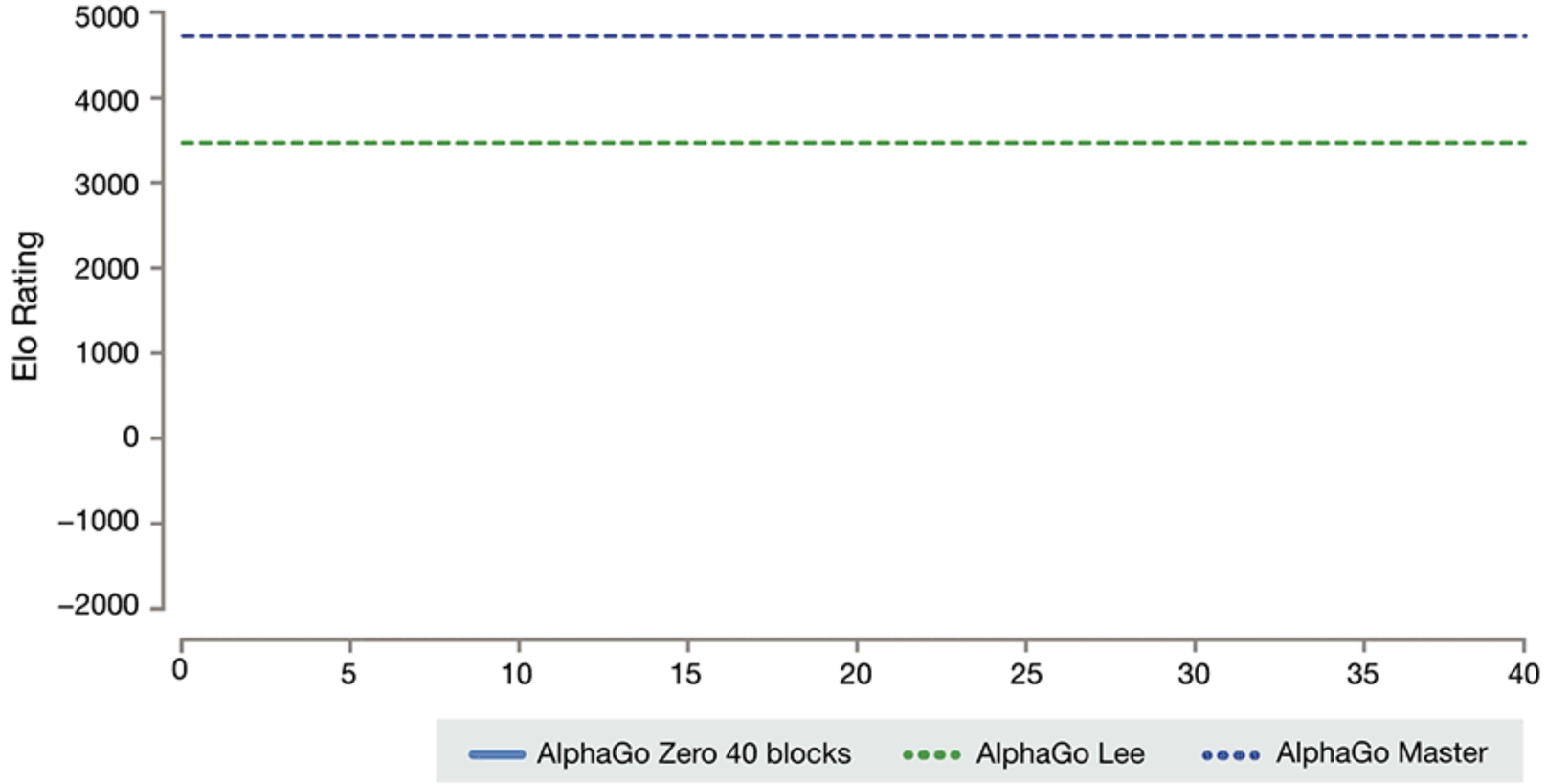
Preferred Skills: ?

영어를
모른다는 것은?



https://w3techs.com/technologies/overview/content_language

코딩을 할 줄 모른다?

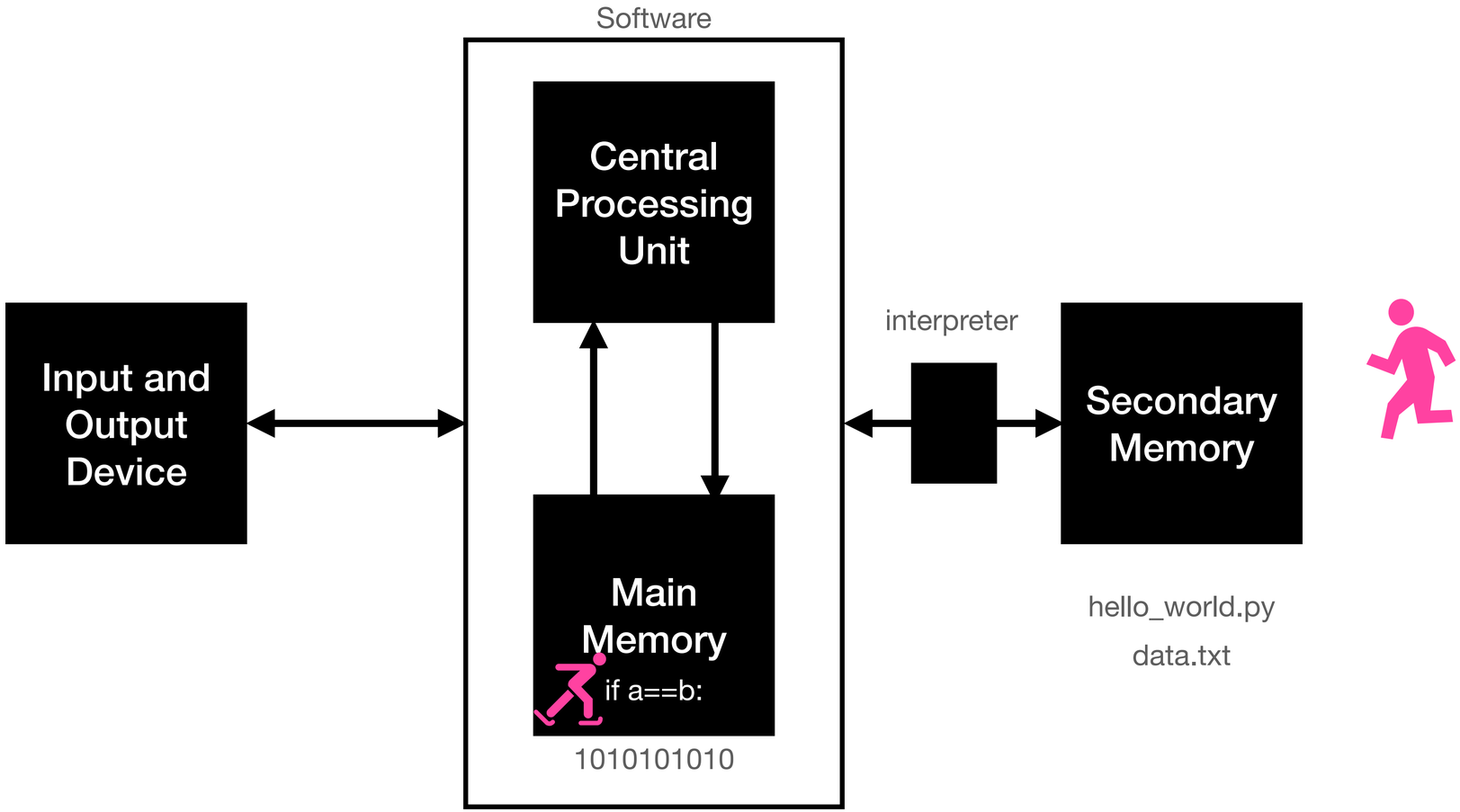


https://deepmind.com/blog/article/alphago-zero-starting-scratch

강의계획

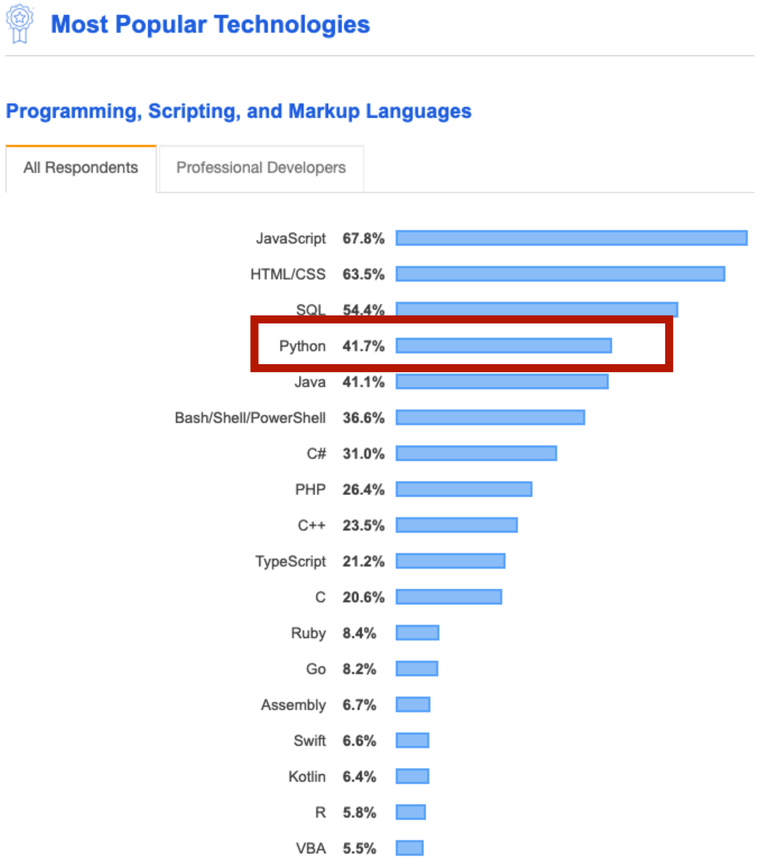
- 1주차
  - 환경설정, Github, Slack 사용법
- 2주차
  - Python 변수, 기본 연산, 출력
- 3주차
  - Python 문자열
- 4주차
  - 조건문, 반복문
- 5주차
  - 함수
- 6주차
  - 클래스 I
- 7주차
  - 클래스 II
- 8주차
  - 데이터 전처리
- 9주차
  - 데이터 시각화
- 10주차
  - 빅데이터 분석

프로그램 실행과정


Software
Central Processing Unit
Input and Output Device
interpreter
Secondary Memory
Main Memory
if a==b:
1010101010
hello_world.py
data.txt

Python

- Guido van Rossum이 1991년에 최초 배포
- 코드의 가독성을 높이는 것이 파이썬의 철학
- 언어 특징
  - Interpreted
  - High Level
  - General Purpose



https://insights.stackoverflow.com/survey/2019#most-popular-technologies

Python을 이용해서

- Logic을 만들 수 있어야 하며
  - 간단한 if-else
  - 자료구조/알고리즘
  - 수학적 연산
  - 모델 정의

- 데이터/파일을 Read/Write 할 수 있어야 한다.
  - 로컬 파일 읽기
  - 외부 파일 (e.g., 클라우드, 웹사이트) 읽기
  - 데이터 변형
  - 모델 저장

머신러닝 (Machine Learning)

- 머신러닝 : 데이터 + 알고리즘
  - 기존에 단순 rule-base 또는 통계 기반으로 해결하지 못한 문제를 데이터의 힘을 빌려 해결한다.
- Python 기반의 라이브러리들을 활용
  - Numpy, Pandas, Matplotlib, Scikit Learn
  - Tensorflow, Pytorch

빅데이터

얼마나 커야 “빅” 데이터인가?

2000년
HDD: 20GB
RAM: 64MB

2020년
SSD: 512GB
RAM: 16GB

https://www.apple.com/shop/buy-mac/macbook-pro

빅데이터

- 저장소의 용량은 기하급수적으로 증가하는 반면 처리량은 선형적으로 증가
  - 하드 드라이브에 있는 데이터를 모두 처리하는데 걸리는 시간은 매년 증가
  - SSD의 개발로 더 복잡한 데이터 처리에 더 복잡한 알고리즘과 프레임워크를 사용할 수 있게되었지만, 아래와 같은 동향은 변하지 않음 (Wiktorski, 2019)

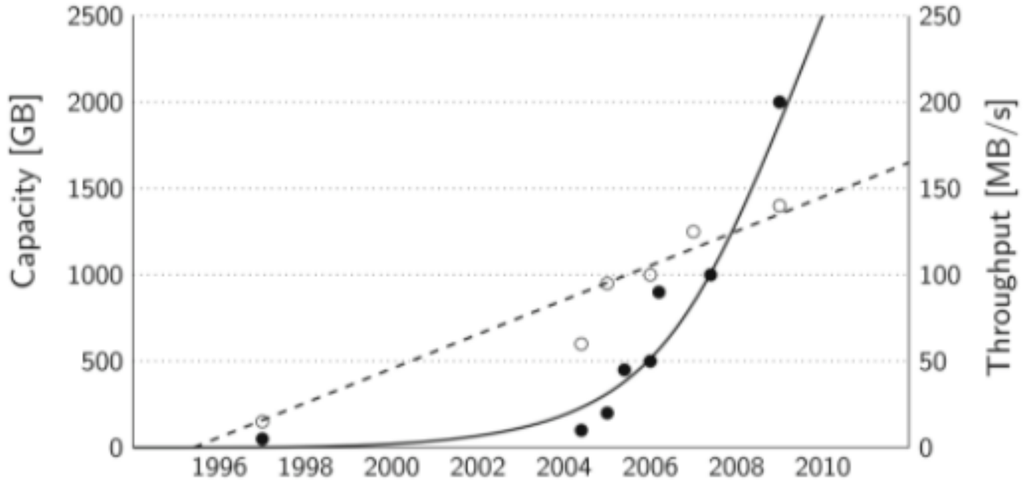


Fig. 2.2 Historical capacity versus throughput for HDDs. *Source* Leventhal (2009)

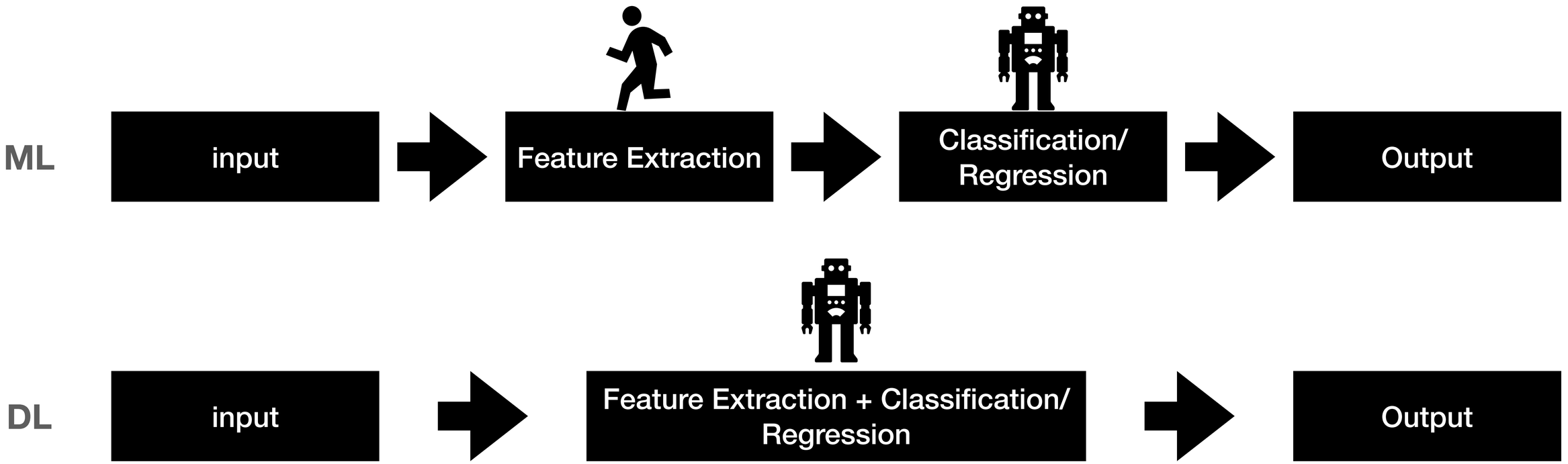
* Wiktorski, Tomasz. Data-intensive Systems: Principles and Fundamentals Using Hadoop and Spark. Springer, 2019.

빅데이터

- 어떻게 많은 양의 데이터를 효과적으로 저장/처리/분석 할 수 있을까?
- 분산 컴퓨팅
  - Hadoop, Spark

딥러닝

딥러닝과 머신러닝의 차이는?



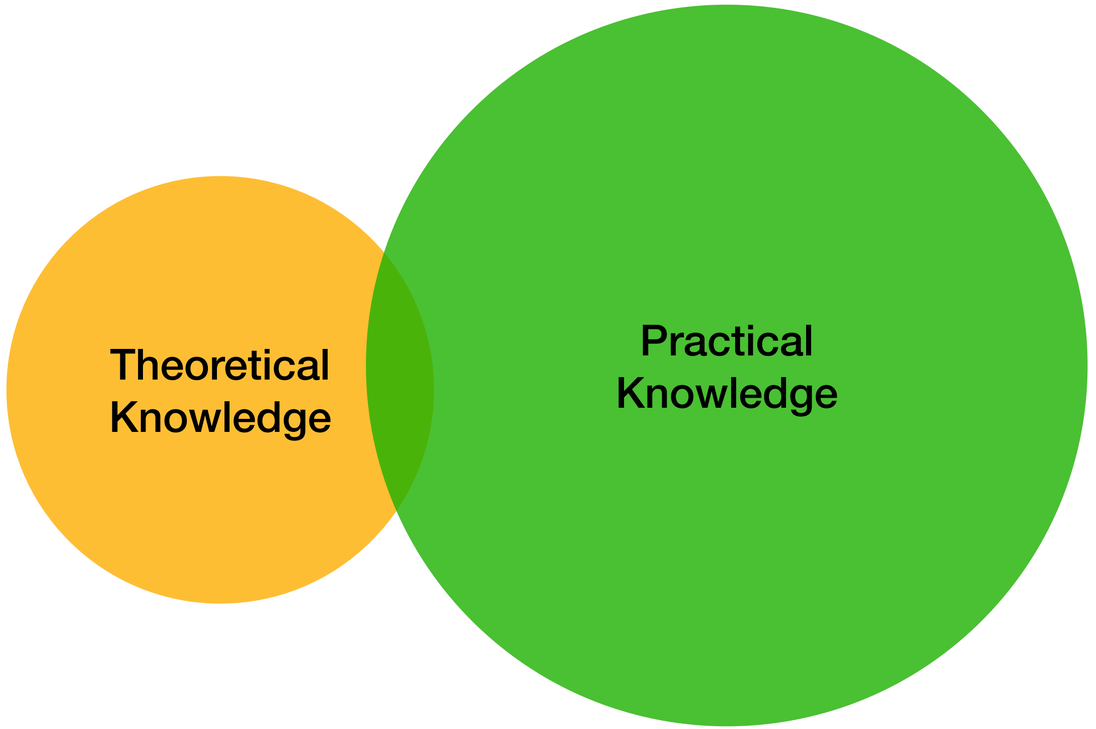
배울 내용들

- 지식적인 측면
  - Python 프로그래밍에 익숙해지기
  - 빅데이터 기술에 대한 지식
- 실무적인 측면
  - Github 활용법
  - Shell 활용법

성적평가

- 1시간 이론강의, 2시간 실습강의
- 매 수업 퀴즈
  - 성적평가 미반영 (점검용)
- 참여도 (40%)
  - 출석 + Q&A 답변
- 과제 (60%)
    - 매주 작은 과제

수업 목표



E.O.D