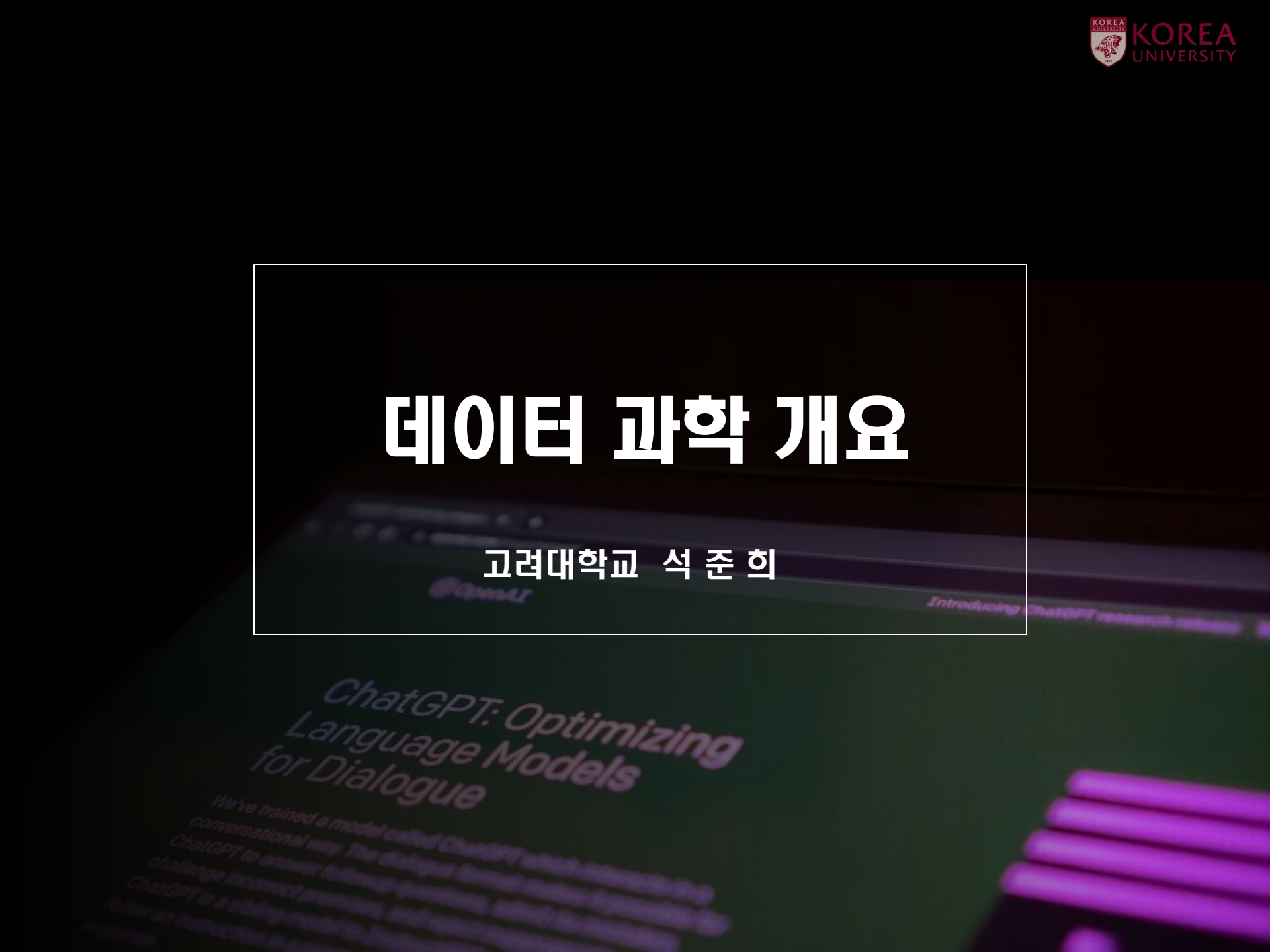
데이터 과학 개요

고려대학교 석 준 희



목차

- 데이터 과학 개요
- 프로그래밍 소개

KOREA UNIVERSITY

데이터 과학 개요

데이터 과학 (Data Science)

- **데이터 과학** 이란?
  - 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는 과정에서 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야 (위키피디아)
  - 누구나 다 해왔고, 하고 있었던 것 아닌가??

- 다양한 분야에서 지속적으로 연구되어 오던 주제
  - 통계학
  - 전자공학: 신호처리, 패턴인식
  - 컴퓨터과학: 기계학습 (인공지능)
  - 산업공학: 데이터마이닝
  - 경영학: MIS (경영정보시스템)

- 그런데 왜 지금?
  - 빅데이터로 촉발
  - 딥러닝으로 가속화
  - 인공지능으로 통합

전통적인 데이터 과학 (1990년대 이전)

- 전통적 데이터 과학 = 통계
  - 전체 데이터를 관측 할 수 없고, 소규모 데이터만 관측할 수 있는 상황
  - 소규모 데이터로부터 어떻게 전체 데이터의 성질을 파악할 수 있는가?

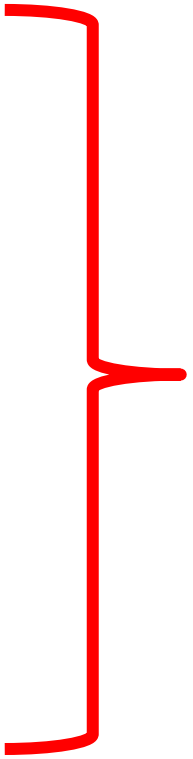
- 예시
  - 남성과 여성의 평균 수명 차이
  - 새로 개발된 신약의 효과 검증

- 상황
  - 낮은 계산 능력
  - 적은 양의 데이터

- 해결책
  - 다양한 가정 (정규분포, 선형성 등등)
  - 이론적 결론

정보화 시대의 도래 (1990년대)

- 컴퓨터의 발전 (무어의 법칙)
  - 계산 및 처리 속도가 급격히 증가
  - 데이터 저장 용량의 증가
  - 가격의 하락
- 인터넷의 발전
  - 전산적으로 처리가 가능한 데이터의 증가
  - 흩어져 있던 데이터의 수집과 공유가 가능
- 데이터 처리 기법의 발전
  - 대용량 데이터 저장, 관리, 분석 기법의 개발
  - 새로운 데이터 예측 기법의 발견



Big Data

빅데이터 (Big Data) – 2000년대

- 복잡한 대규모 데이터(빅데이터)에 새로운 분석 방법론과 관리도구를 적용하여 기존에는 찾지 못했던 새로운 정보를 찾을 수 있음
  - Volume: 대규모 데이터
  - Variety: 언어, 이미지, 비디오 등을 포함하는 다양한 형태
  - Velocity: 빠르게 생성, 유통, 분석, 소비됨

- 예시: 대형마트의 장바구니 분석 – 맥주와 기저귀의 관계

- 데이터 수집
  - 어떻게 대규모 데이터를 효과적으로 수집할 것인가?
- 데이터 관리
  - 대규모 데이터를 어떻게 관리하고, 저장하고, 유통할 것인가?
- 데이터 분석
  - 기존의 데이터에서는 찾지 못했던 새로운 가치를 어떻게 찾을 것인가?

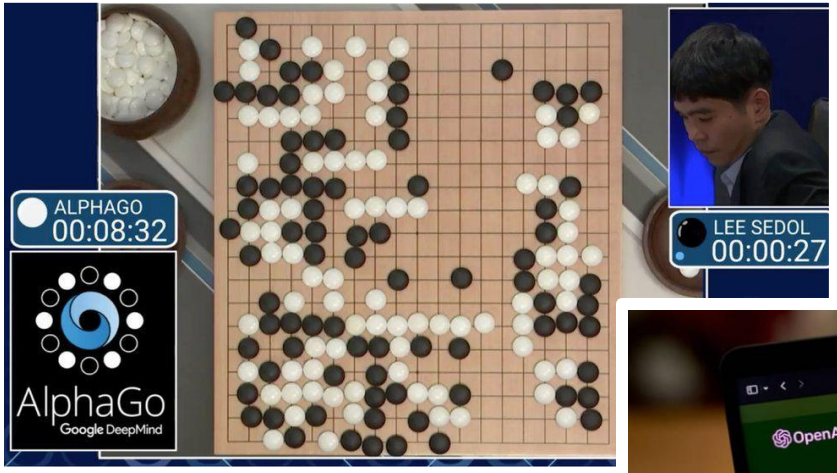
- 2000년대 시작된 빅데이터는 2010년이 지나면서 활용이 점점 중요해짐

인공지능 (AI: Artificial Intelligene) – 2010년대

IBM 왓슨의 퀴즈쇼 (2011)

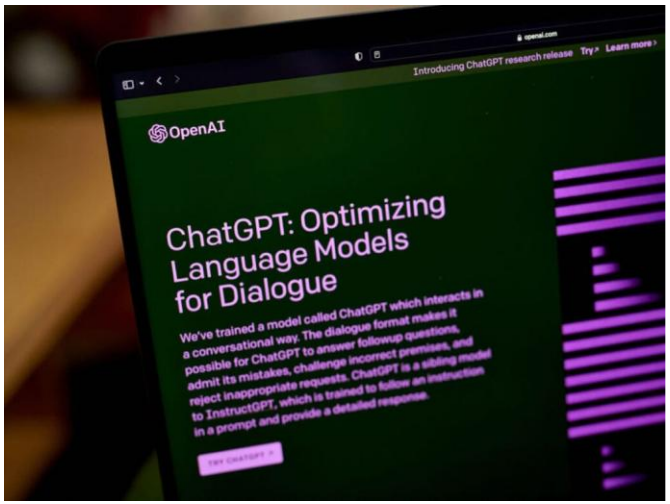
https://www.cbsnews.com/news/ibm-watson-defeats-humans-in-jeopardy/

알파고와의 대국 (2016)



https://www.bbc.com/news/technology-35785875

ChatGPT 열풍 (2023)



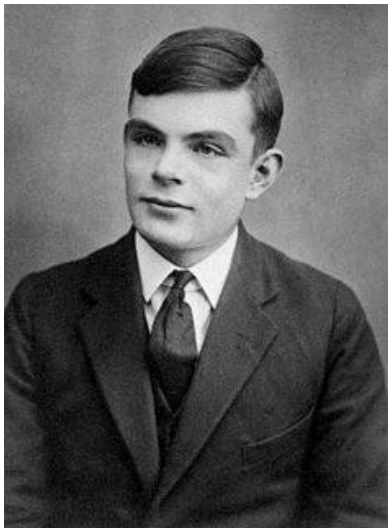
https://chat.openai.com/chat(ChatGPT)

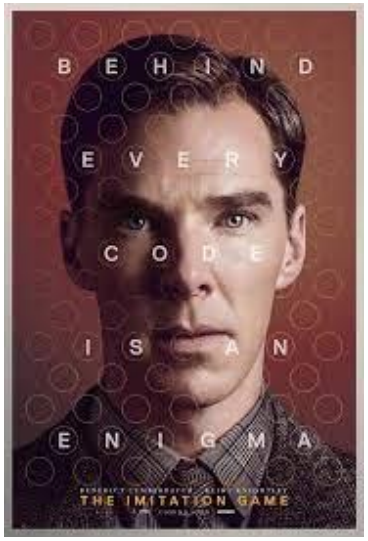
인공지능의 시작

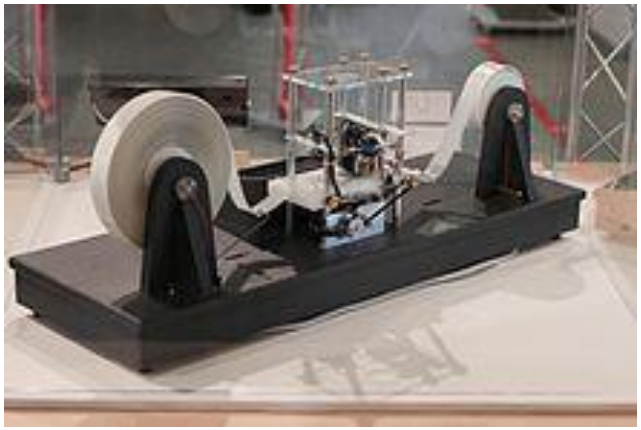
- **'생각하는 기계'** 라는 개념은 앨런 튜링에 의해서 처음으로 제시
  - 앨런 튜링 (Alan Turing, 1912~1954): 컴퓨터 과학의 아버지, 근대적 컴퓨터 모델 (튜링 머신)의 고안

- 계산기 (Calculator): 단순한 수치적 연산, 1640년 파스칼에 의해 발명
- 컴퓨터 (Computer): 논리적 행위의 구현, 튜링에 의해 개념 정립

- **인공지능:** 인간이 하는 논리적 행위(인지, 판단 등)를 모방하는 기계





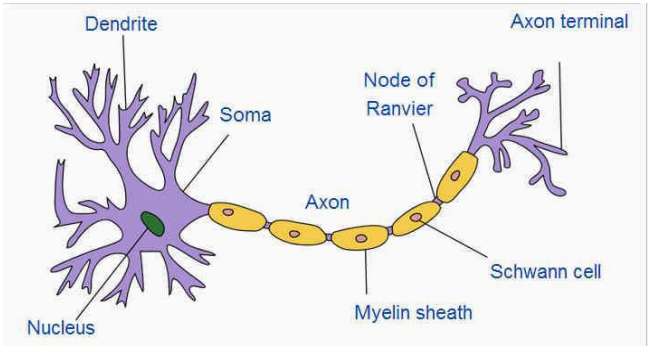


인공신경망 (Artificial Neural Network)

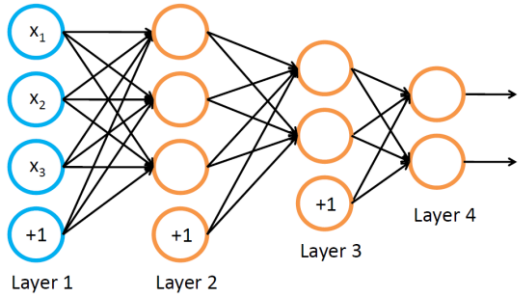
- 인공지능을 어떻게 구현해야 하는지는 명확하지 않음
  - 로켓의 궤적: 어떻게 푸는지는 알지만 계산이 어려움
  - 안면 인식: 5살 어린아이도 잘 하지만 컴퓨터로는 어려움

- **인공신경망**: 인간의 두뇌를 모방하여 만든 수학적 모델

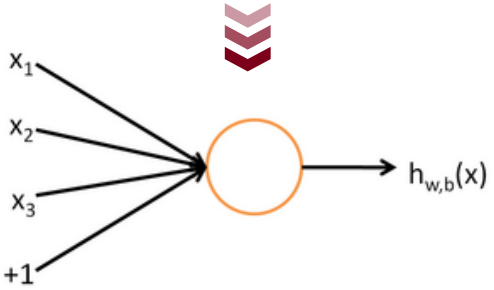
인간의 두뇌



신경세포

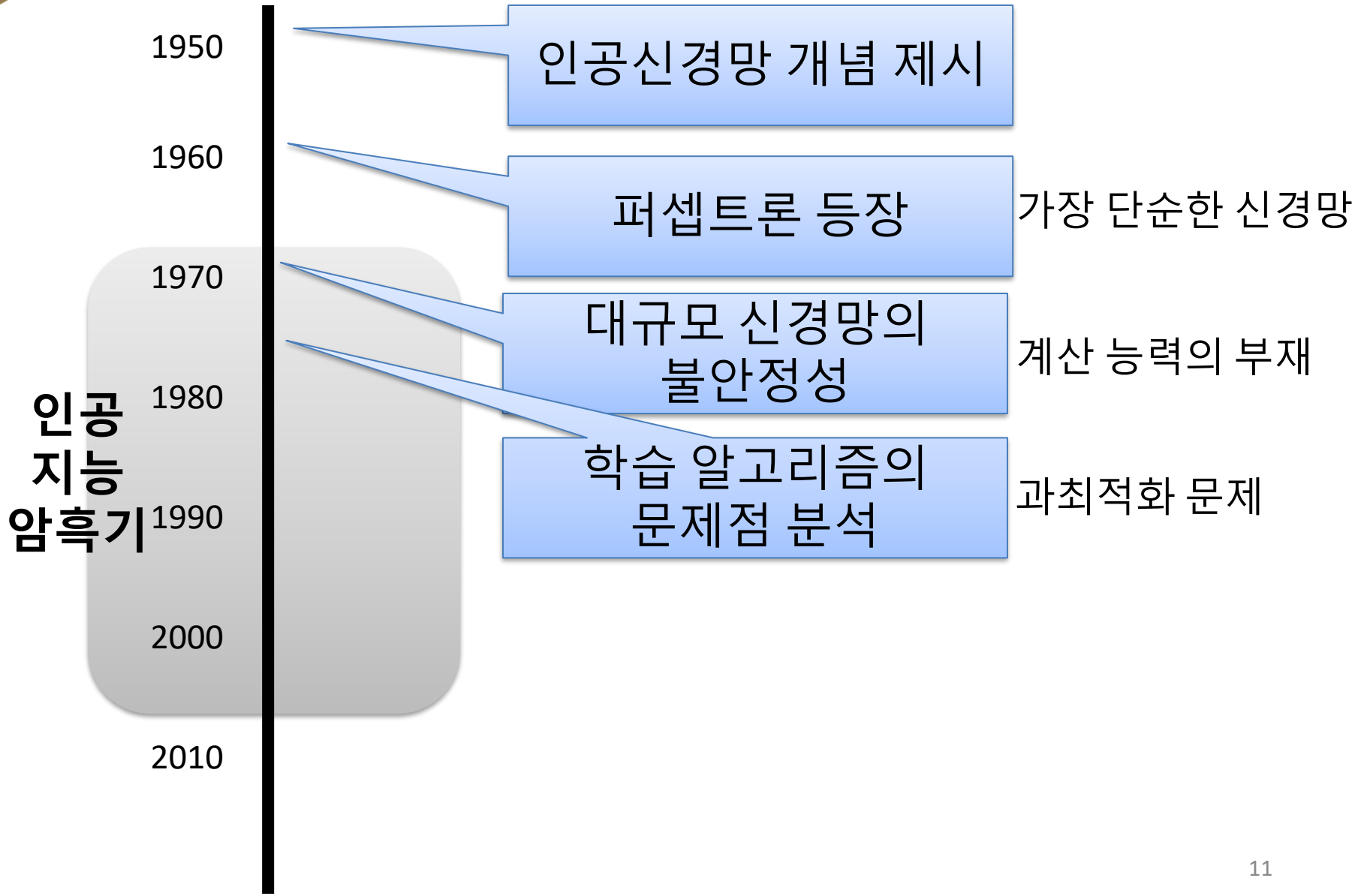


인공신경망



퍼셉트론

인공지능의 암흑기


1950
1960
1970
1980
1990
2000
2010
인공
지능
암흑기
인공신경망 개념 제시
퍼셉트론 등장
가장 단순한 신경망
대규모 신경망의
불안정성
계산 능력의 부재
학습 알고리즘의
문제점 분석
과최적화 문제

과거의 인공지능

- 단순한 규칙기반 혹은 탐색 기반의 인공지능이 선호됨.
  - 복잡한 지능을 묘사하는 것은 불가능

럭키證券 人工지능「브레인스」개발

기술지표·개별재료등 綜合분석 有望종목추천

情報입력「인덱스펀드」보다 한발 앞선 기법

화제

1990년 2월 9일 매일경제

딥러닝(Deep Learning)의 출현

1950
1960
1970
1980
1990
2000
2010

인공지능 암흑기

인공신경망 개념 제시

퍼셉트론 등장

가장 단순한 신경망

대규모 신경망의 불안정성

계산 능력의 부재

학습 알고리즘의 문제점 분석

과최적화 문제

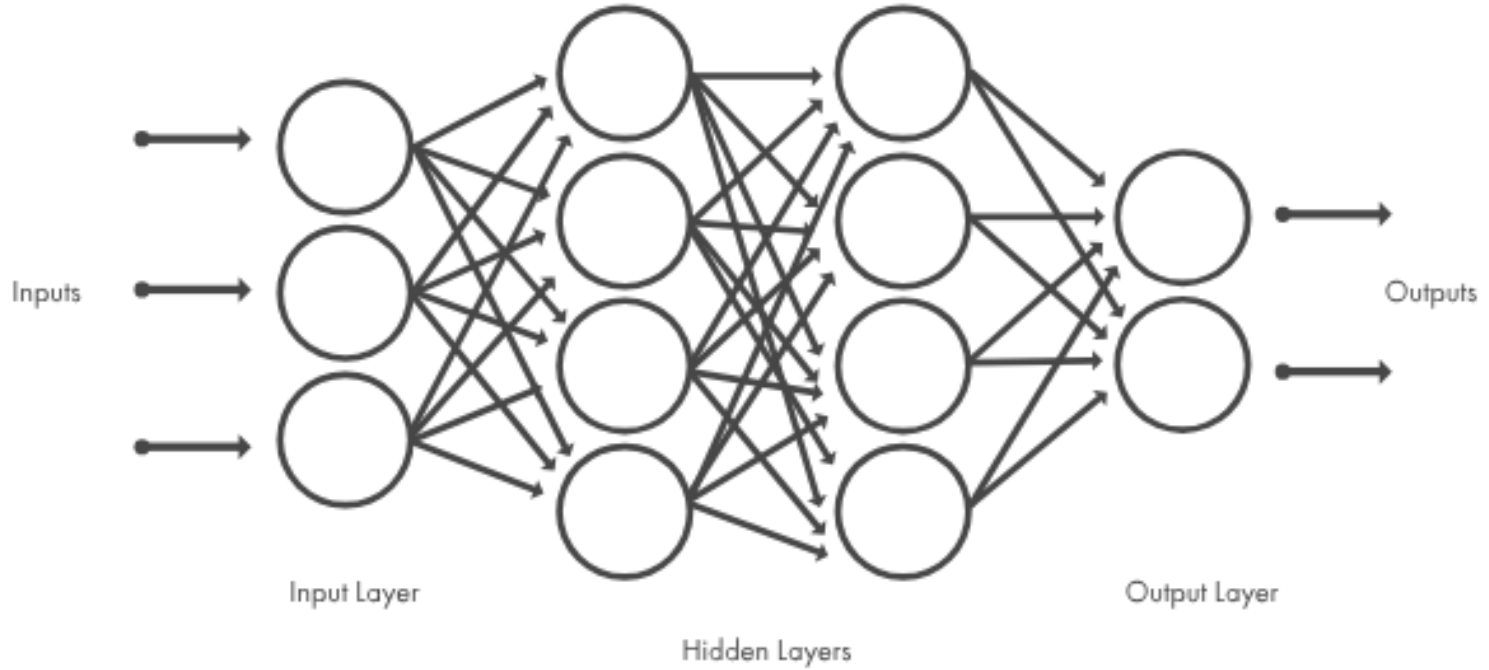
Deep Belief Network

Geoffrey Hinton

KOREA UNIVERSITY

딥러닝 모델

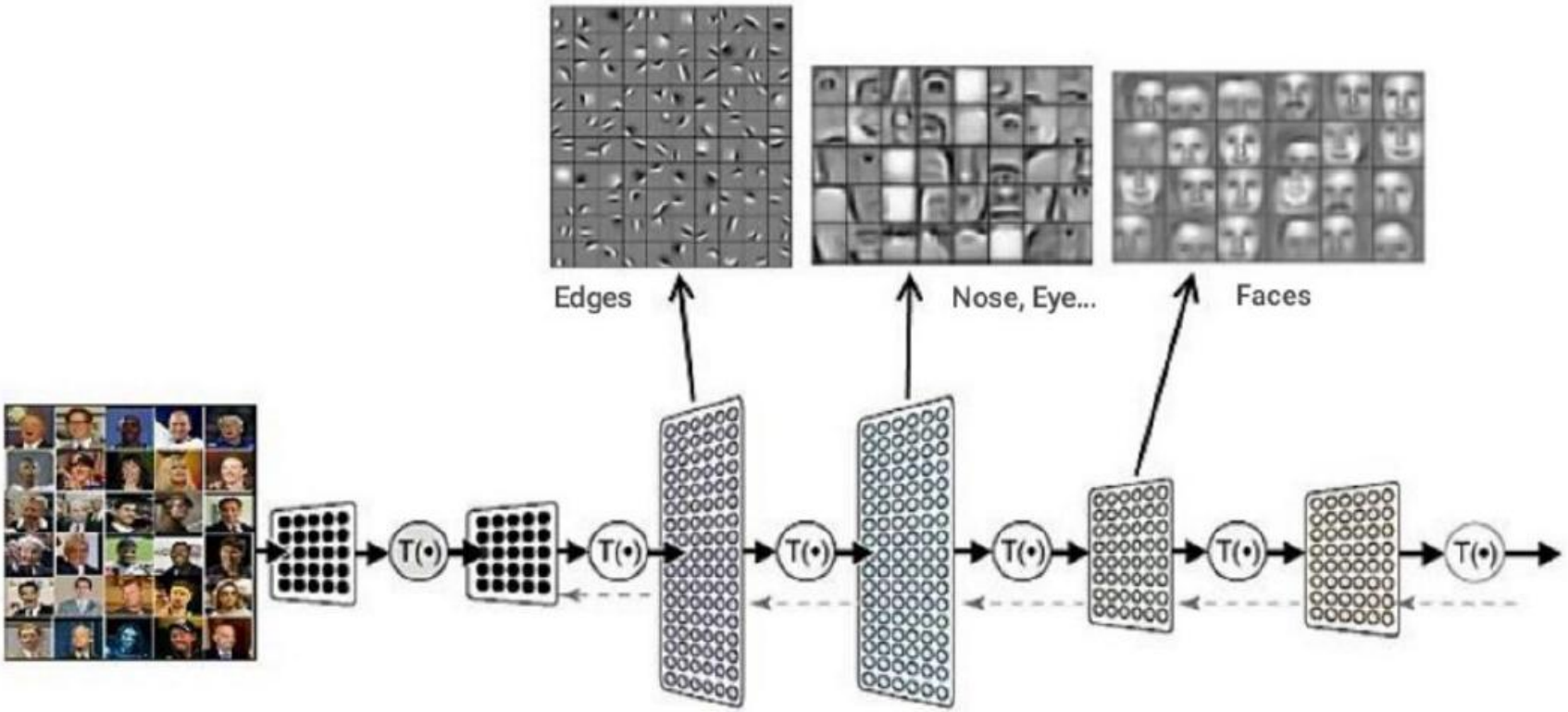
- 딥러닝 모델
  - 외부로 나타나지 않고 숨겨진 계층(hidden layer)가 다수 존재하는 신경망 모델
  - 숨겨진 변수가 학습을 통해 필요한 정보(feature)를 추출하는 모델
- 전통적 접근 vs. 딥러닝 접근
  - 전통적 접근: 인간 전문가가 유용한 정보를 선정하고 이를 데이터에서 계산
  - 딥러닝 접근: 모델이 자동으로 유용한 정보를 추출하여 사용



https://kr.mathworks.com/discovery/deep-learning.html

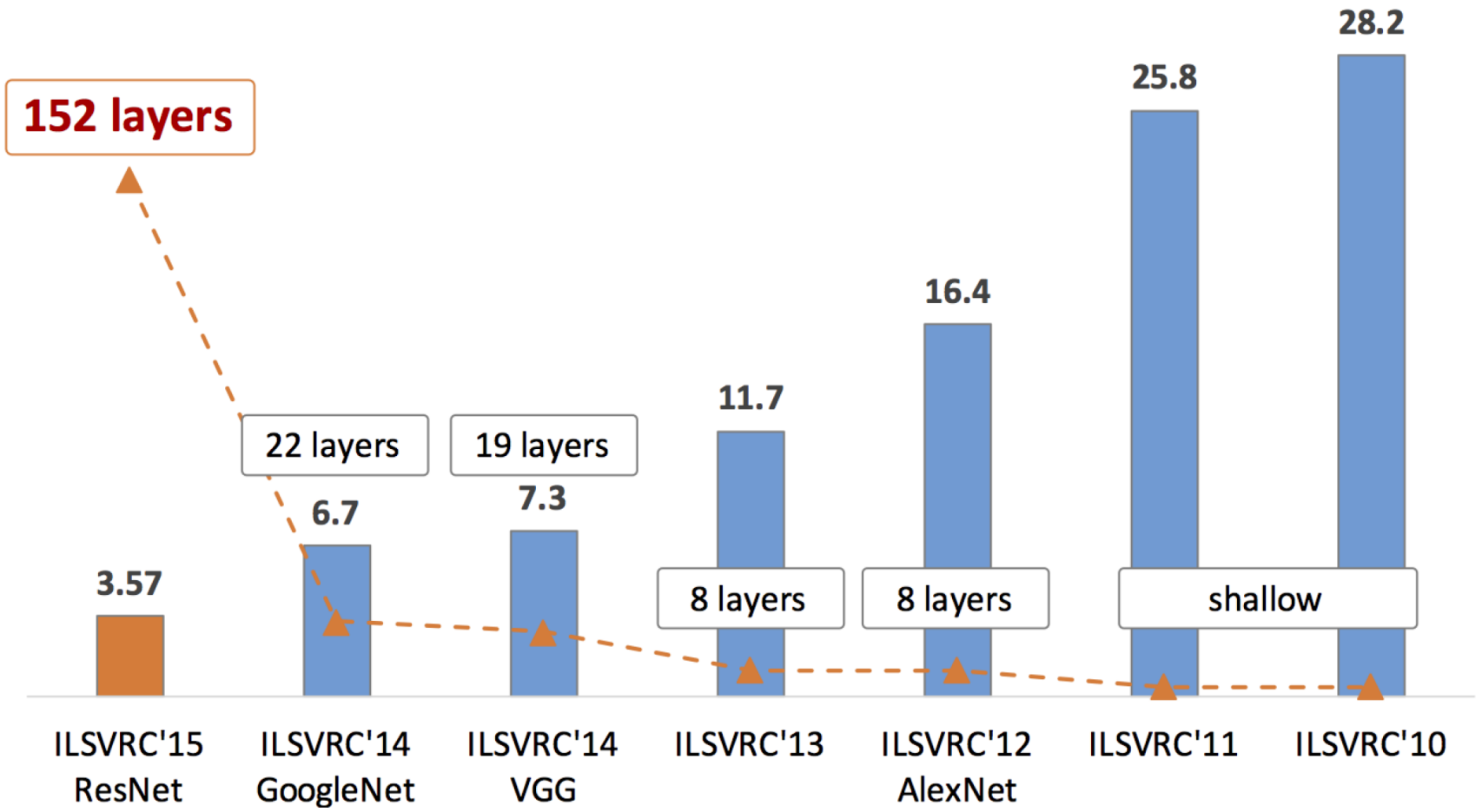
딥러닝 모델

- 딥러닝을 이용한 영상 인식의 예



딥러닝 모델

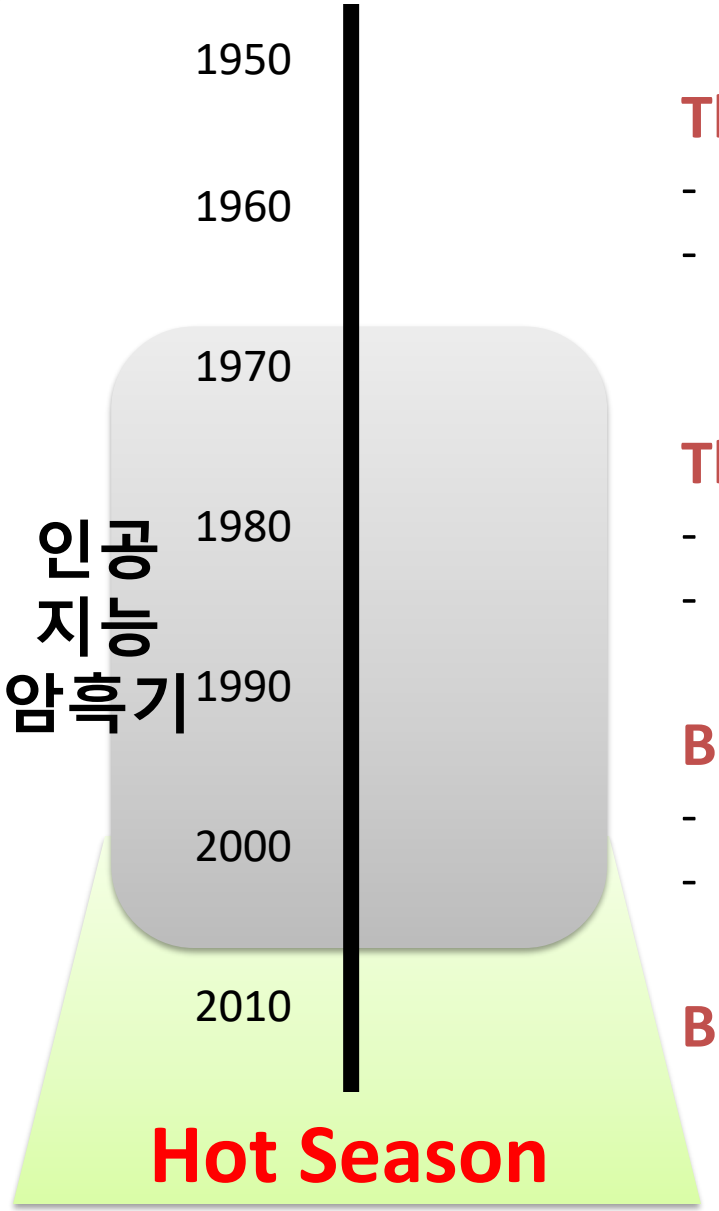
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



Top-5 error rate, (Human : 5.1%)

Kaiming He, ICML 2016 Tutorial

인공지능의 봄-여름



The hottest topic in speech recognition

- Keep breaking the previous records
- MS and Google deployed DL-based speech recognition in their products

The hottest topic in computer vision

- Top recorder holder in competition
- Image search of Google, Baidu, and Facebook.

Becoming hot in natural language

- Semantic search & deep Q&A in IBM Watson
- Large scale language model

Becoming hot in applied mathematics

인공지능의 구현

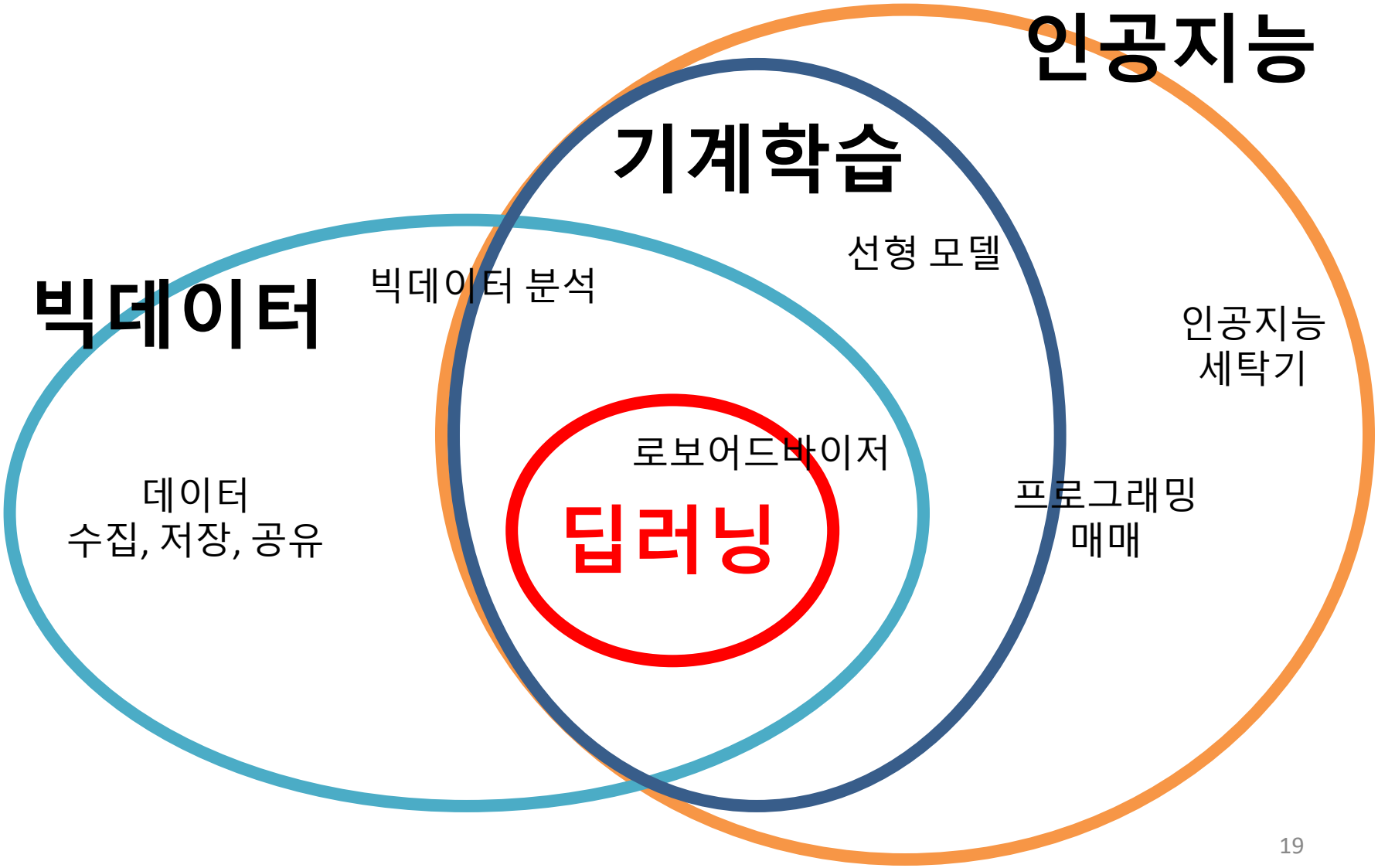
- 인공지능을 어떻게 구현해야 하는지는 명확하지 않음.
  - 우리는 지능이 판단을 한다는 것은 알지만 어떻게 판단하는지는 모름

- **규칙기반 인공지능**
  - 명확한 규칙이 존재하고 그 규칙에 따라서 판단
  - 다양한 소스로부터 규칙을 수집하고 정리하는 것이 주된 요구사항
  - E.g. clinical decision support system

- **데이터기반 인공지능** (**머신러닝, 기계학습**)
  - 대략적인 규칙은 존재하지만 명확하지는 않음
  - 많은 사례로부터 대략적인 규칙을 배우는 것이 주된 요구사항
  - E.g. face recognition

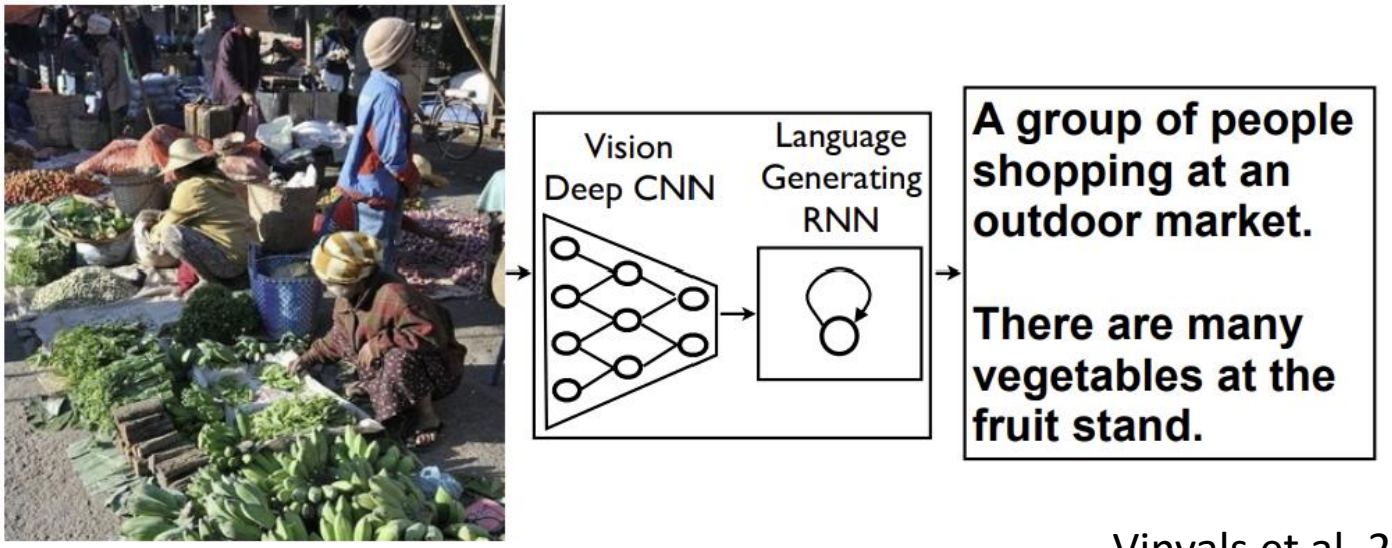
- 데이터기반의 인공지능을 위한 필수 요소
  - 양질의 대규모 데이터 → 빅데이터
  - 높은 연산 능력 → GPU 및 AI 반도체
  - 새로운 학습 방식과 모델 → 딥러닝 모델

인공지능, 기계학습, 빅데이터, 딥러닝??



인공지능의 발전 방향

- 범용인공지능 (AGI: Artificial General Intelligence)
  - 현재의 인공지능: 하나의 프로그램이 하나의 문제를 해결
  - 인간의 지능: 한 사람이 다양한 문제를 해결하는 것이 가능
  - AGI: 인간과 같이 한 프로그램이 다양한 문제를 해결
- 멀티모달 학습 (Multimodal Learning)
  - 다양한 종류의 데이터를 학습하여 다양한 문제를 해결
  - E.g. Image captioning, Video Q&A, ChatGPT v4

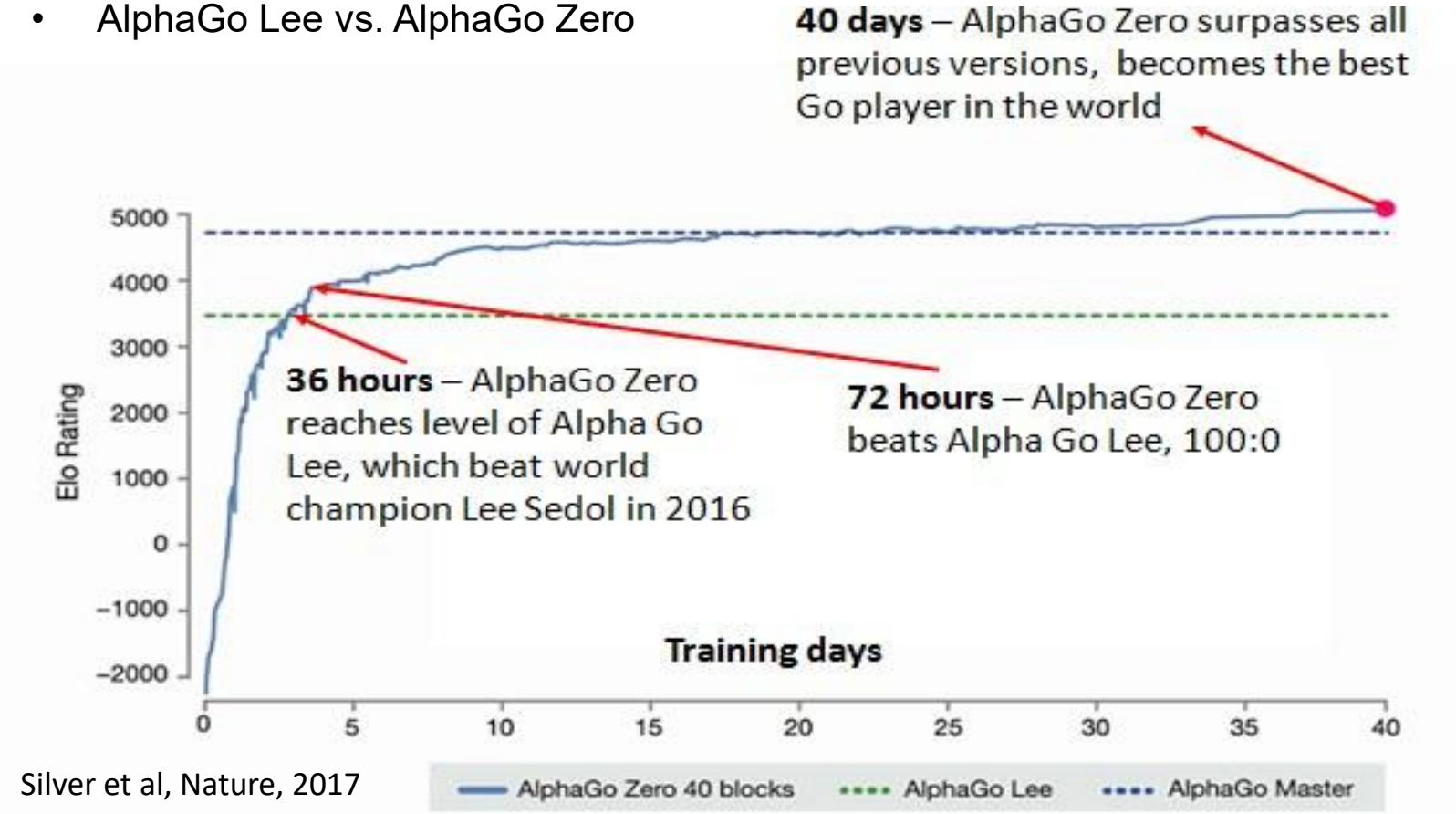


Vinyals et al, 2014

인공지능의 발전 방향

- 인공초지능 (ASI: Artificial Super Intelligence)
  - 기존의 인공지능: 인간의 해결법을 학습 (지도학습)
  - 미래의 인공지능: 새로운 해결법을 발견 (**강화학습**)

- AlphaGo Lee vs. AlphaGo Zero



Silver et al, Nature, 2017

인공지능, 빅데이터, 기계학습, 딥러닝??

- 일반적으로 산업계에서 생각하는 현실적 구분

- 빅데이터 = 데이터 엔지니어링
  - 데이터의 수집, 가공, 저장, 공유
  - 데이터베이스 및 자동화 플랫폼관련 기술

- **데이터 과학 = 데이터 분석 (우리 수업의 내용!!)**
  - 주로 정형 데이터 위주 + 약간의 비정형 데이터
  - 인간이 인지하기 어려운 데이터(많은 숫자들)에 대한 분석
  - 공장에서 불량률 예측, 스타벅스의 입점 위치 분석 등
  - 통계 및 전통적 기계학습 + 약간의 딥러닝에 대한 기술

- 인공지능 = 기계학습 = 딥러닝
  - 비정형 데이터 위주, 인간이 쉽게 인지하는 데이터를 다룸
  - 이미지 분류, 문장에 대한 답변 등
  - 딥러닝 및 인공지능에 대한 기술

데이터 과학 개요

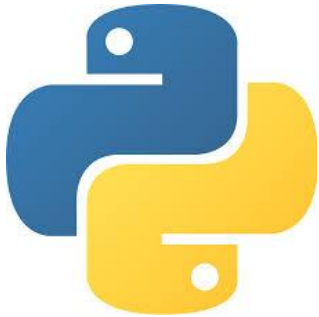
프로그래밍 소개

데이터 과학 프로그래밍

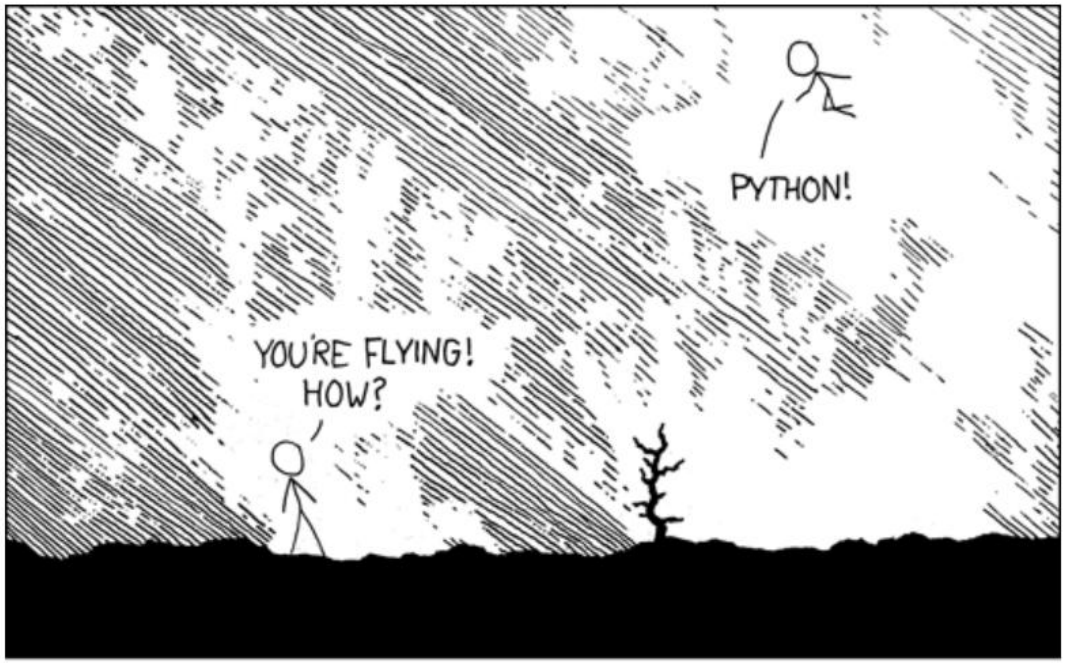
- 일반적인 코딩/프로그래밍
  - Java, C, C++, Javascript 등등

- R
  - 통계학자들이 개발한 통계분석 전문 언어
  - 통계 전문 언어 중에서는 어렵지만, 일반적인 프로그래밍 언어로서는 쉬움
  - 시각화와 통계분석이 장점
  - 빅데이터 분석에 많이 사용

- **파이썬 (Python)**
  - 컴퓨터 학자들이 개발한 일반적인 언어
  - R보다는 어렵지만 프로그래밍 언어 중에서는 쉬움
  - 빠른 처리와 범용성이 장점
  - 인공지능에 많이 사용



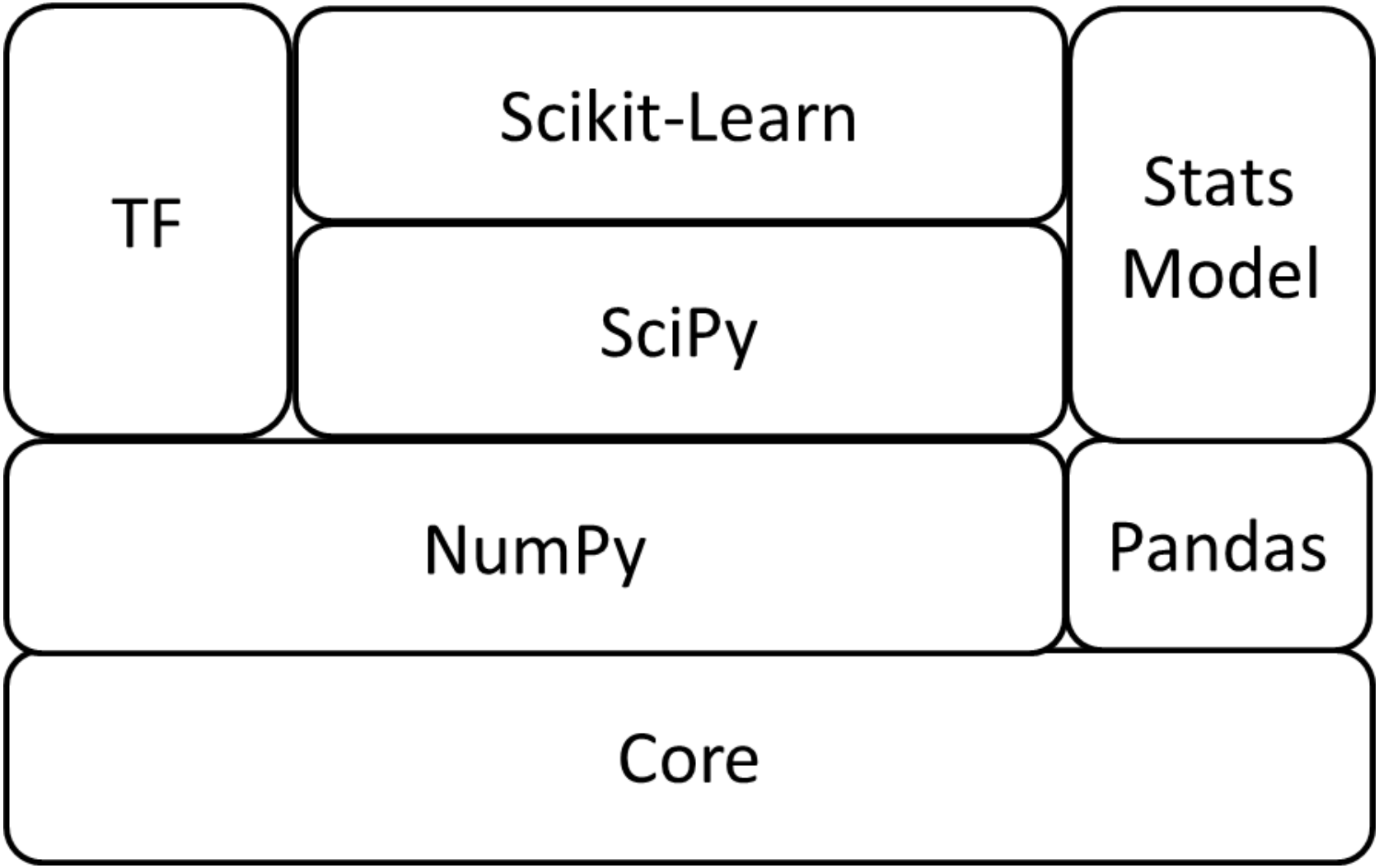
왜 파이썬인가?



https://xkcd.com/353/

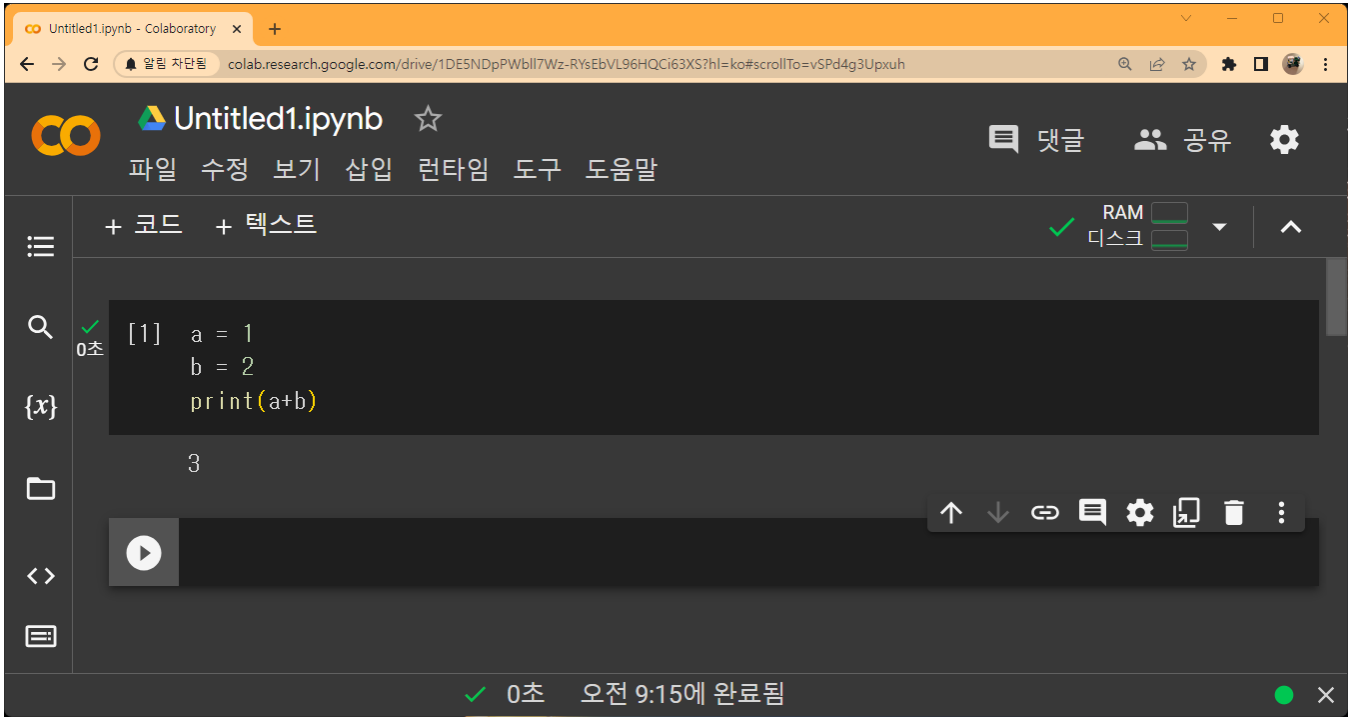
- 쉽고 빠르게 학습 및 이용이 가능
- 범용적 프로그래밍 언어
- 다양한 최신 프로그래밍 기법이 도입 (객체지향, 동적타이핑, ...)
- 잘 조성된 생태계

인공지능을 위한 파이썬 라이브러리



파이썬의 사용

- 컴퓨터에 직접 설치 후 사용
  - 아나콘다 패키지 설치 (https://www.anaconda.com/download)
  - Jupyter notebook을 사용하여 파이썬 실행
- 온라인 클라우드 사용
  - 구글 코랩을 이용 (https://colab.research.google.com/?hl=ko)
  - Jupyter notebook과 유사한 인터페이스
  - 구글 계정이 필요



KOREA UNIVERSITY

감사합니다