밑바닥부터 시작하는 데이터 과학

1회차

- 데이터 과학이란 ?
- 파이썬 실습

스터디 소개 - 목표 및 방향

우리는

기초 수학, 통계, 머신러닝 알고리즘까지 직접 코드를 작성하며 이해할 것입니다.

그래서

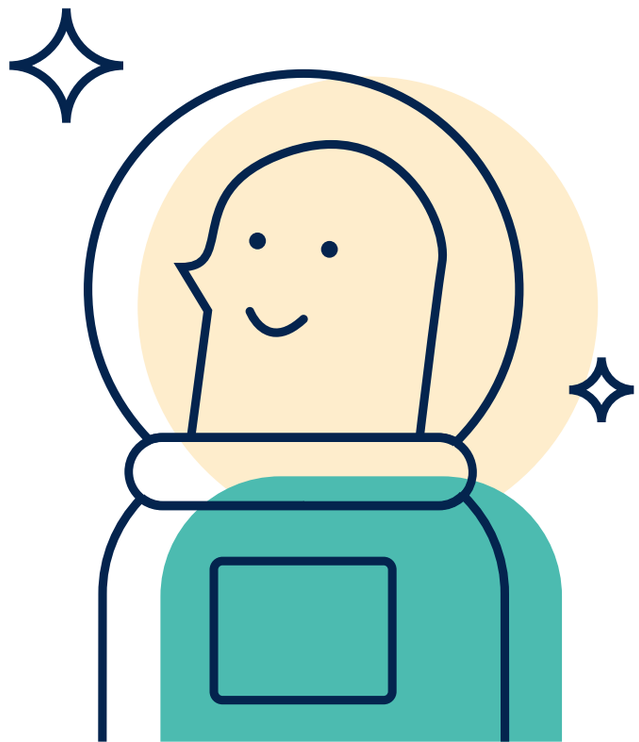
한 분도 빠짐없이 코드를 작성하는 실습 위주로 진행할 것입니다.

물론, 부족한 이론 부분을 채우기 위한 **과제**도 할 것입니다.

스터디가 종료된 후,

데이터 과학을 어떻게 활용할 수 있을지, 쉽게 상상할 수 있습니다 !

그리고, 앞으로 채워나가야할 부분도 좀 더 명확해질 것입니다.



스터디 소개 - 계획

1 회
- 데이터 과학 소개
- Python 속성 강좌

2 회
- 선형대수
- 통계
- 확률
- 가설과 추론

3 회
- 경사하강법
- 데이터 다루기
- 기계학습
- K-NN
- 나이브베이즈

4 회
- 단순 회귀 분석
- 다중 회귀 분석
- 로지스틱 회귀분석

5 회
- 의사결정나무
- 신경망
- 군집화

6 회
- 자연어처리
- 추천 시스템

질문 1)

x = [1, 2, 3, 4, 5, 6] 라는 list 객체가 있습니다.
두 번째 값부터 마지막 값까지 뽑아내고 싶다면,
어떤 명령어를 사용해야 될까요?

질문 2)

회귀 분석과 로지스틱 회귀 분석의 차이를 아시나요?

질문 3)

데이터 과학을 통해 해결하고 싶은 문제가 있나요?

1. 데이터 과학이란?

데이터 과학이란

- 데이터를 수집하고 분석하여 활용하기 위한 모든 기술의 집합을 말하며,
- **컴퓨터 사이언스, 수학, 통계학, 머신 러닝(machine learning)**, 영상 및 신호 처리 등 다양한 학문 분야가 만나는 영역이다.
- 프로그래밍을 포함하는 **컴퓨터 사이언스**는 실제로 데이터를 다루기 위한 필수 기술이며,
- **수학과 통계학**은 데이터 분석 모형의 기반에 깔린 핵심적인 개념을 구체화하는 언어이다.
- **머신 러닝**은 이러한 분석 결과를 활용하여 지금까지 인간이 해오던 각종 분석과 의사 판단을 대신하고자 하는 노력이다.

2. 데이터 과학 학습

- **기초 수학 이론**
  - 선형대수
  - 미분과 적분
  - 최적화
  - 확률론

- **데이터 분석 이론**
  - 확률 모형
  - 검정 및 추정
  - 회귀 분석과 분류, 클러스터링

- **컴퓨터 관리 및 프로그래밍 기술**
  - 리눅스 운영체제 사용법
  - 프로그래밍 언어
  - 데이터베이스 시스템
  - 병렬처리, 가상화, 클라우드 사용법

- **해당 분야에 대한 전문 지식**
  - 해당 분야의 정보를 이해하고 분석 결과가 올바른지 판단할 수 있는 능력
  - 이미지 처리, 음성/음향 처리, 텍스트 처리 등의 자료 전처리 기술

2. 데이터 과학 학습

이번 스터디에서는,

<table>
<tr><th colspan="3">수학 및 통계 분석</th><th colspan="2">머신 러닝</th></tr>
<tr><td colspan="3">선형대수</td><td colspan="2">클러스터링</td></tr>
<tr><td colspan="3">미적분</td><td rowspan="3">예측</td><td rowspan="3">분류</td></tr>
<tr><td colspan="3">최적화</td></tr>
<tr><td>확률론</td><td>확률 모형</td><td>검정과 추정</td></tr>
<tr><td colspan="5">프로그래밍 (Python)</td></tr>
</table>

3. 데이터 과학 활용

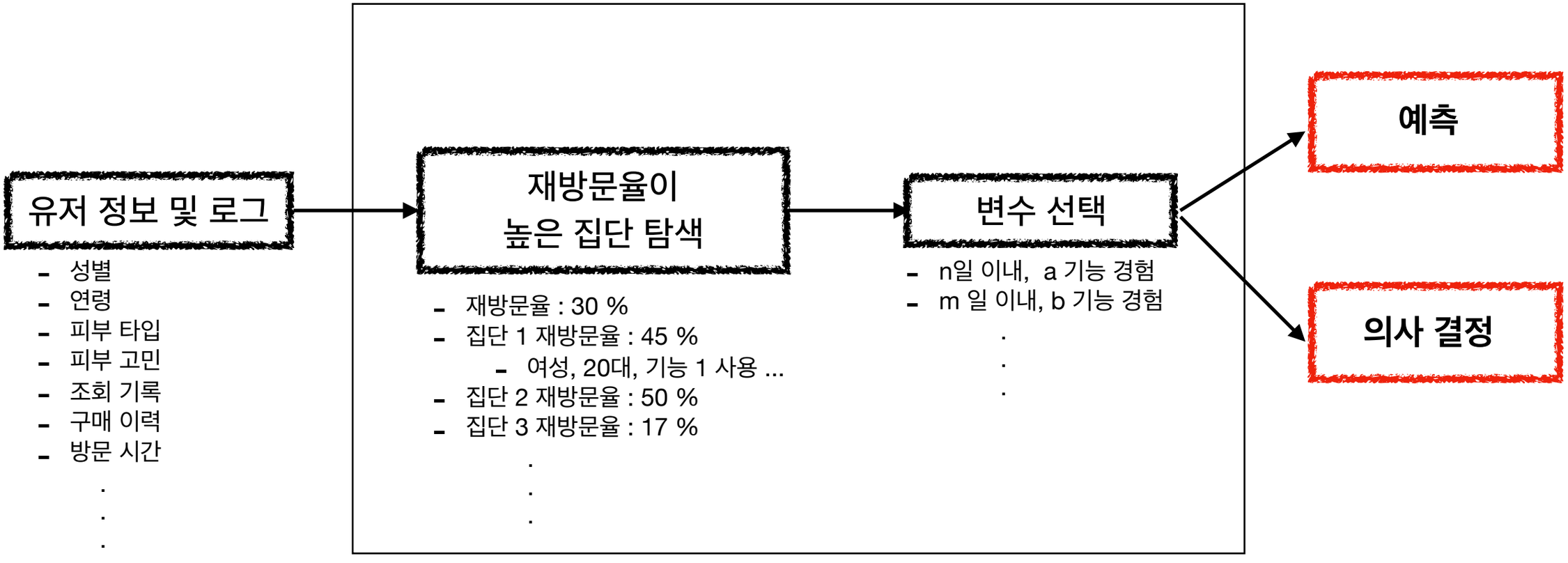
이 많은 방법론들을 공부하면 우리는 무엇을 할 수 있을 까요?

3. 데이터 과학 활용

수학 및 통계 분석을 공부하면,

수 많은 데이터 속에서 꼭 필요한 데이터만 선택할 수 있게 됩니다.

비즈니스 활용 사례 (재방문율 분석)



3. 데이터 과학 활용

예측 모델을 공부하면,

수 많은 (과거) 데이터를 바탕으로 비어있는 값 (미래)를 예측할 수 있게 됩니다.

예측 문제의 수학적 표현

예측 모형은 입력 데이터 X를 받아서 출력 데이터Y 를 만든다는 점에서 **수학의 함수(function)**와 유사하다.
예측 문제의 최종 목표는 **X와 Y의 관계 함수 f(X)** 를 구하는 것이다.

$$Y=f(X) \qquad Y=a_1x_1+a_2x_2+a_3x_3+a_4x_4+....$$

하지만 현실적으로는 데이터의 개수가 너무 많고, x 변수의 개수가 한정적인 경우가 많으므로 정확한 f를 구할 수 없다.
(**연립 방정식**을 생각하면 쉽다 !)
따라서, 우리는 정확한 f 보다는 f와 가장 유사하고, 재현 가능한 $\hat{f}$ 를 구하는 것에 집중한다.
이러한 $\hat{f}$ 을 가지고 있다면, 그리고 f와 아주 유사하다면, 새로운 데이터가 들어왔을 때,
실제 y와 아주 유사한 $\hat{y}$ 을 구할 수 있다.

$$Y\approx\hat{f}(X)$$

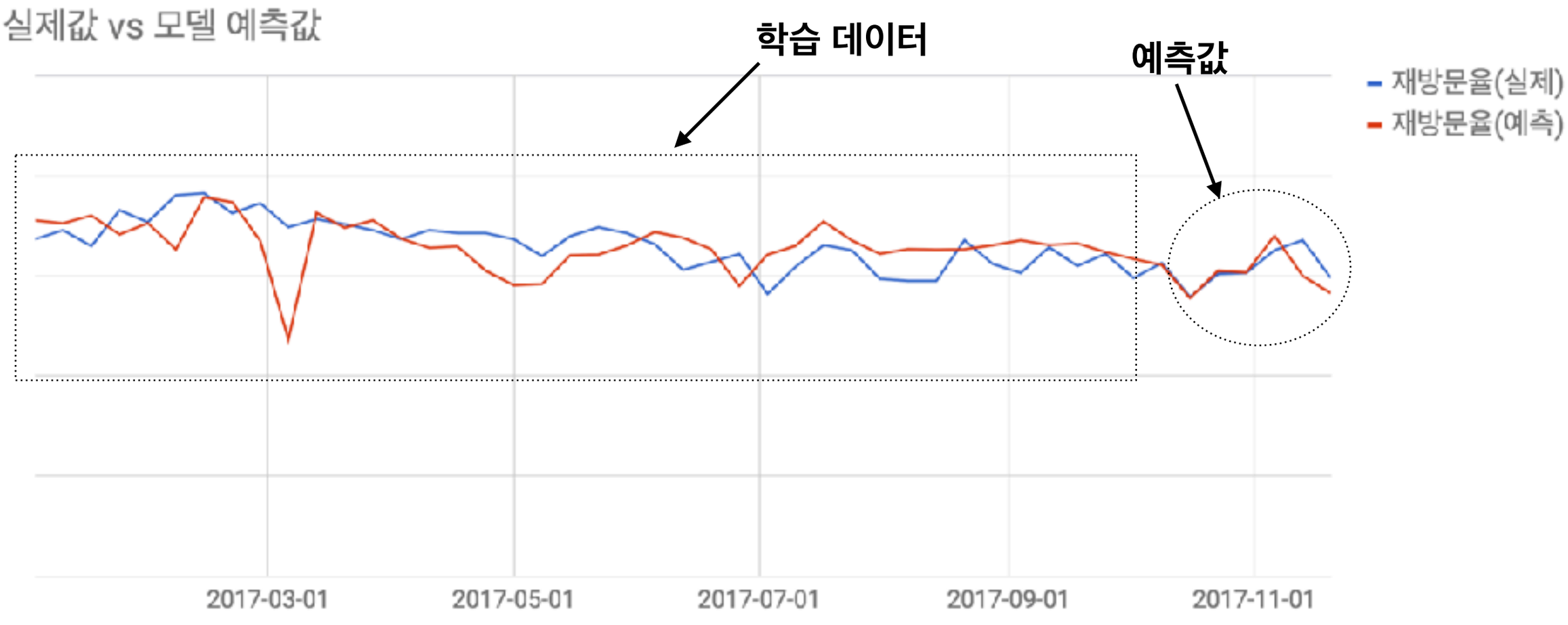
여기서 우리는, f도 중요하지만 X도 매우 중요하다는 사실을 꼭 기억하자.

3. 데이터 과학 활용

예측 모델을 공부하면,

수 많은 (과거) 데이터를 바탕으로 비어있는 값 (미래)를 예측할 수 있게 됩니다.

비즈니스 활용 사례 (재방문율 예측)

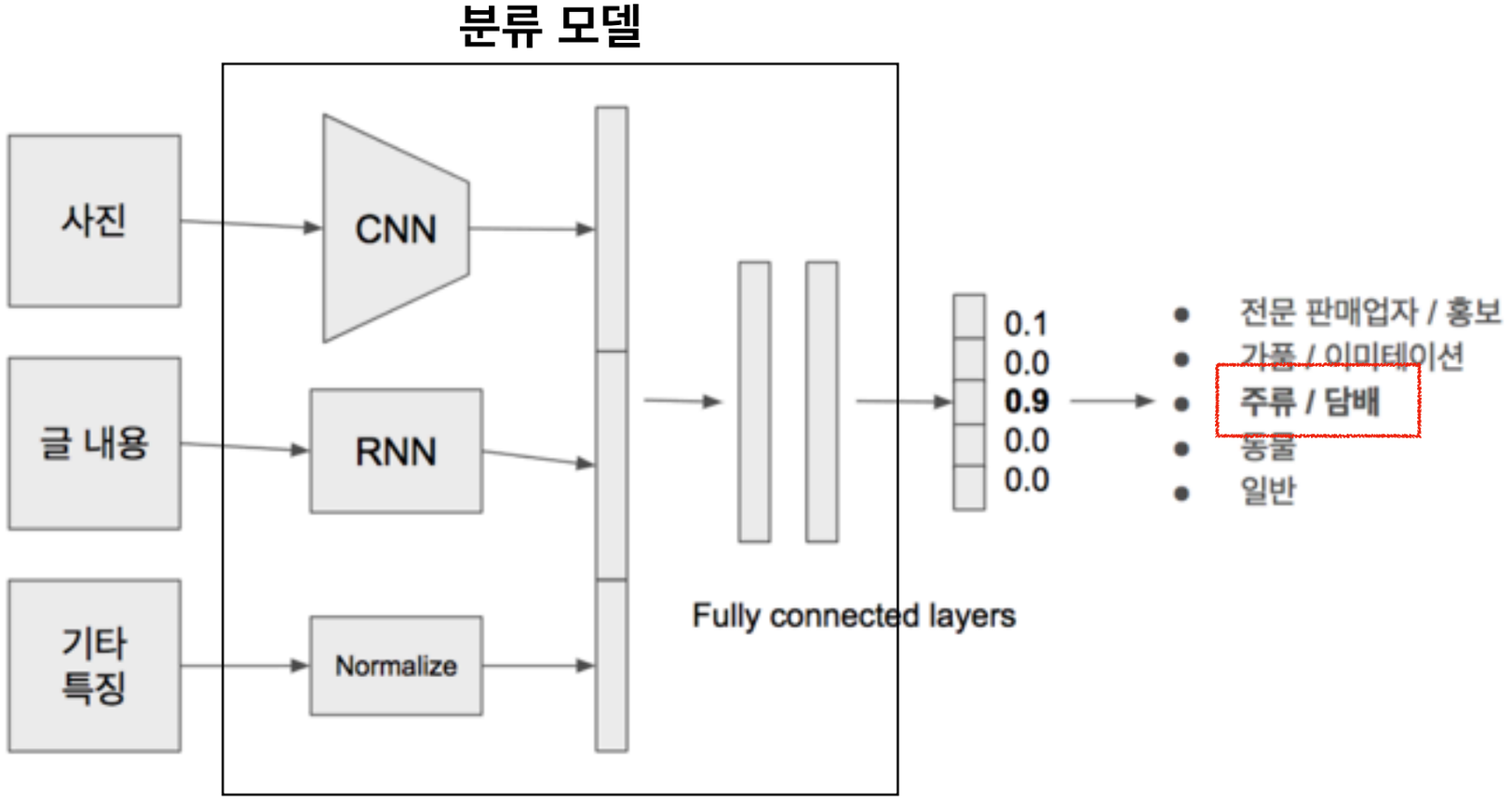


3. 데이터 과학 활용

분류 모델을 공부하면,

수 많은 데이터를 바탕으로 **패턴**을 찾아서, 어떠한 값으로 분류할 수 있게 됩니다.

비즈니스 활용 사례 (게시글 분류)



게시글 분류 모델 네트워트

출처 : https://medium.com/n42-corp/당근마켓에서-딥러닝-활용하기-3b48844eba62

3. 데이터 과학 활용

분류 모델을 공부하면,

수 많은 데이터를 바탕으로 **패턴**을 찾아서, 어떠한 값으로 분류할 수 있게 됩니다.

비즈니스 활용 사례 (게시글 분류)

대부분의 머신 러닝 / 딥러닝은 바로 이 분류 문제를 풀기 위한 방법이다.

그렇다면 A.I 는 무엇 일까요 ????

3. 데이터 과학 활용

분류 모델을 공부하면,

수 많은 데이터를 바탕으로 **패턴**을 찾아서, 어떠한 값으로 분류할 수 있게 됩니다.

A . I (Artificial Intelligence)

Raymond "Ray" Kurzweil (레이 커즈와일)

- “특이점이 온다” 저자
- “ 뇌는 패턴 인식기이다” 라고 주장함.
  - 대뇌 신피질은 같은 패턴인식기 3억개가 펼쳐져 있는 구조.
  - 이 패턴 인식기의 구조와 작동방식만 이해하면 뇌의 정보처리 과정을 분석할 수 있다.
  - 즉, 인간의 뇌를 모사할 수 있게되고, 인간과 같이 판단할 수 있는 **A.I**가 만들어진다.

엄청난 성능을 보이는 패턴 인식기 = Deep Learning -> A.I

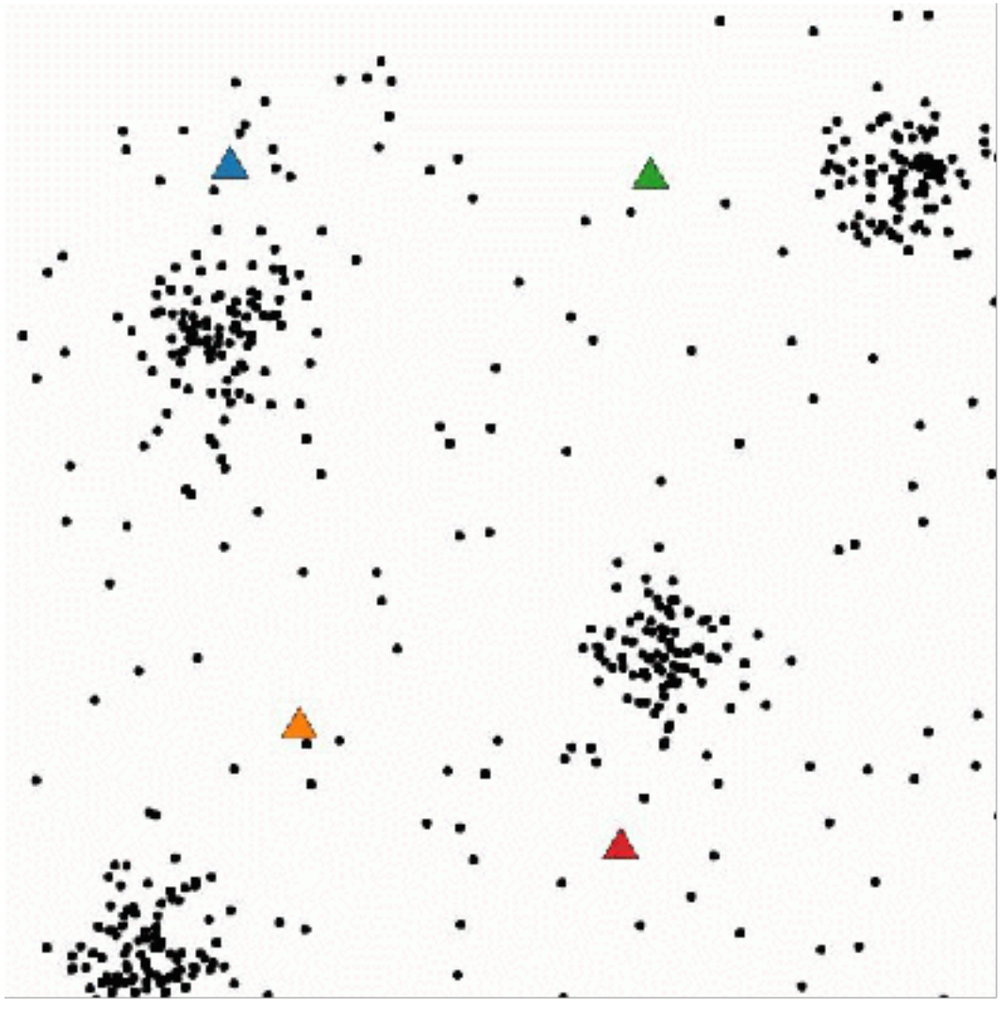
3. 데이터 과학 활용

클러스터링을 공부하면,

수 많은 데이터들을 몇 개의 그룹으로 묶을 수 있습니다.

게임 유저 클러스터링 분석

k-means 클러스터링 알고리즘이 클러스터를 나눠가는 과정



(출처 : 위키피디아 Incheol, CC BY-SA 4.0)

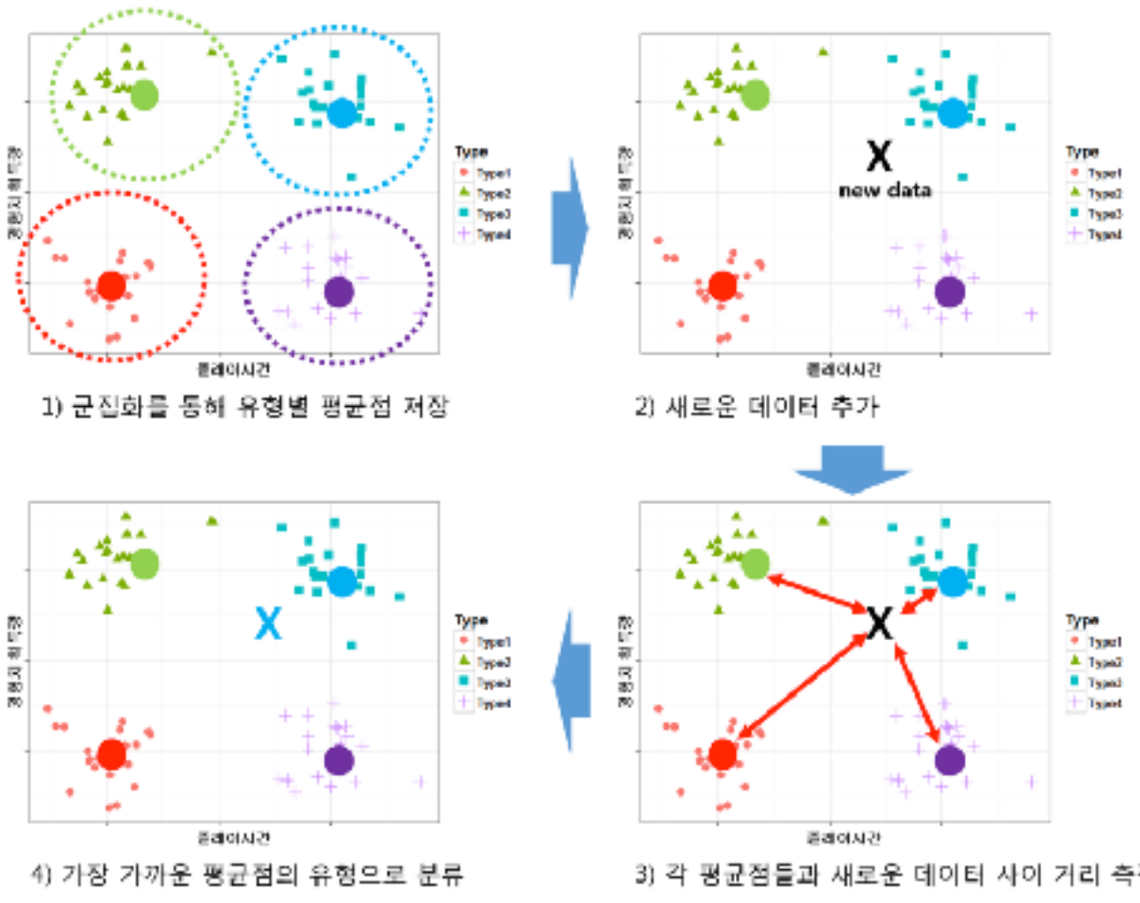
3. 데이터 과학 활용

클러스터링을 공부하면,

수 많은 데이터들을 몇 개의 그룹으로 묶을 수 있습니다.

게임 유저 클러스터링 분석

클러스터별 차이를 분석해,
의사결정에 도움을 준다.



k 평균 군집화로 유형을 분류하는 과정

(출처 : http://blog.ncsoft.com/?p=25333

3. 데이터 과학 활용

마지막 시간에는,

풀고 싶은 문제를 정의하고,

그 문제를 풀기 위한 데이터 과학 방법론들을 충분히 상상하고,

데이터 과학을 활용해 꼭 풀수있게 되길 바랍니다.

3개월간 잘 부탁드립니다 !

잠깐
쉬어요! 우리 :)

learnit Flip your study

다음에
또!
같이!
만나요!

learnit Flip your study

https://github.com/surprisoh/datascience_scratch_1/