빅데이터 분석 전문가

06차시

파이썬을 활용한 크롤링
pandas 기본
머신러닝 개념

조성진 강사

❑실습해보기

- ❍R에서 크롤링했던 예제를 Python으로 해보기
- ❍JTBC씨의 검색 이용하기
- ❍ 총 검색 개수 얻기

❑판다스(pandas)

- ❍데이터 처리와 분석을 위한 라이브러리
- ❍R의 data.frame을 본떠서 설계한 DataFrame이라는 데이터 구조를 기반으로 만들어짐
- ❍dataframe은 행과 열로 이루어진 형태
- ❍아나콘다 배포판에는 기본으로 포함되어 있음
- ❍pandas의 의미는 panel data(계량경제학)이라는 단어에서 파생.
- ❍panel data란 복수의 시간에 설쳐서 추적하여 얻는 데이터를 뜻함

❑판다스(pandas)의 주요 특징

- ❍자동적/명시적으로 축의 이름에 따라 데이터를 정렬할 수 있는 데이터 구조
- ❍누락 된 데이터의 데이터 정렬 및 통합 처리
- ❍통합된 시계열 기능
- ❍누락된 데이터를 유연하게 처리 가능
- ❍SQL같은 일반 데이터베이스처럼 데이터를 합치고 관계연산을 수행하는 기능

❑기계 학습(머신 러닝Machine learning)

- ❍인공 지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
  - ex) 수신한 이메일의 분류
- ❍기계학습의 핵심은 표현과 일반화
  - 표현: 데이터의 평가
  - 일반화: 아직 알 수 없는 데이터에 대한 처리
- ❍기존 데이터들을 평가하여 향후 데이터를 평가한다.

❑머신 러닝의 기능

❍분류(classification) - 지도학습의 일종

- 주어진 입력 x의 레이블 y를 추정해내는 것
- 소속집단의 정보를 이미 알고 있고, 비슷한 집단으로 묶는 것
- ex) 동물 사진을 입력 받으면 고양이인지 강아지인지를 논리값이나 확률값으로 리턴함

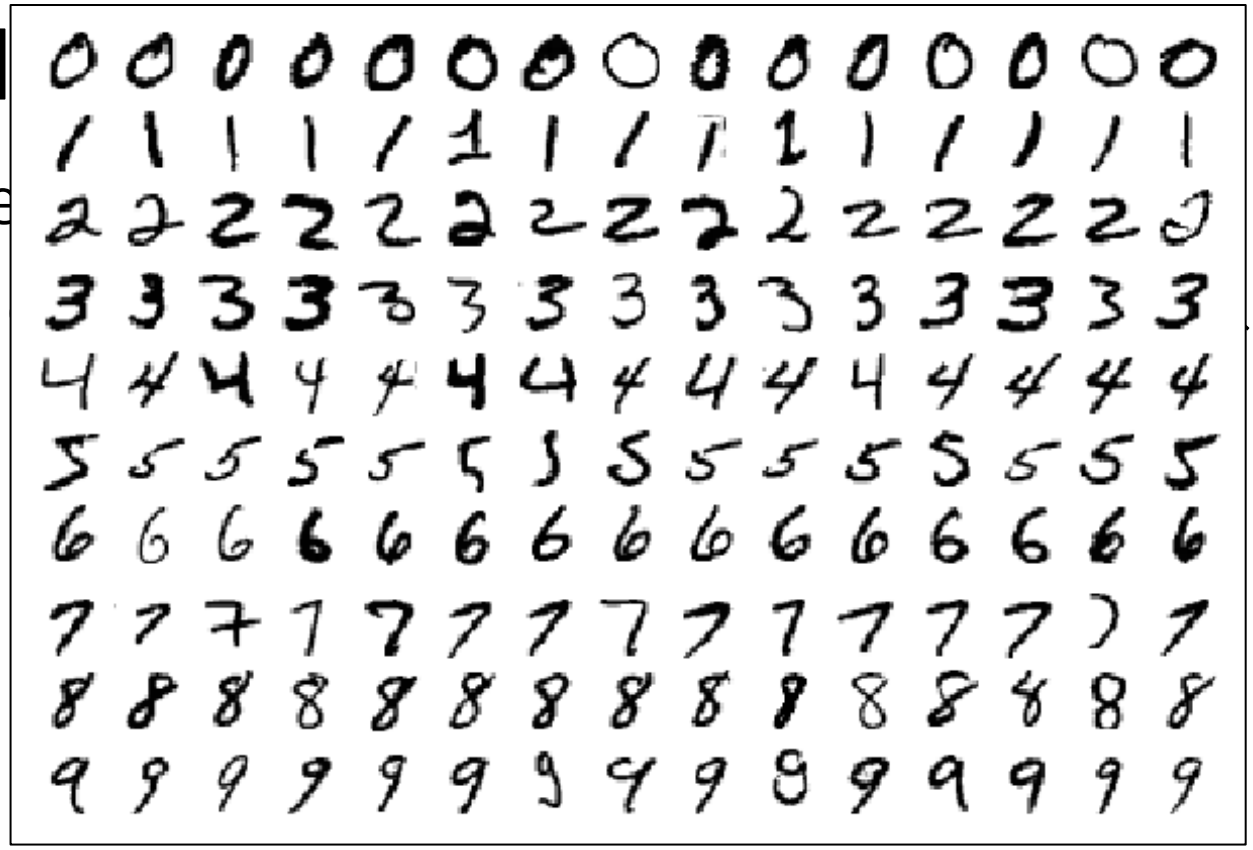
❍군집화(clustering) - 비지도학습의 일종

- 주어진 입력 x와 비슷한 입력들의 군집(cluster)를 추정해내는 것
- 소속집단의 정보가 없고, 모르는 상태에서 비슷한 집단으로 나누는 것
- ex) 내가 키우는 고양이의 사진을 입력했을 때, 동물이라는 집단으로 군집하는 것. 단, DB에 동물에 대한 사진이 있어야함.

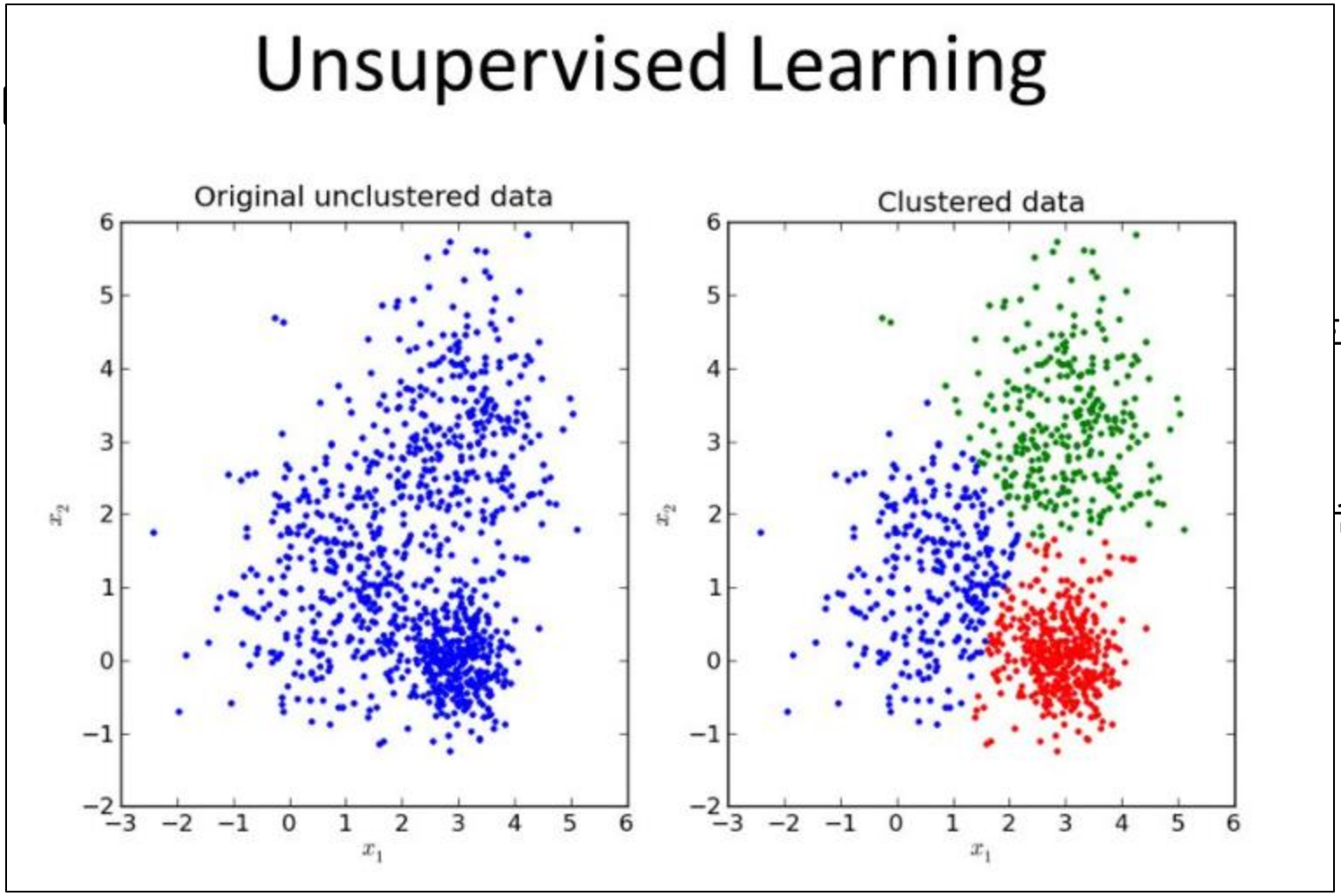
❑머신 러

❍Supe

– 레 시키는 방법



4를 입력했을 때 4를 예측하도록 학습시키는 것



❑scikit-learn 사이킷런

- ❍2007년 구글 썸머 코드에서 처음 구현
- ❍파이썬용 기계 학습 오픈 소스 라이브러리
- ❍다른 라이브러리와 호환성이 뛰어남
- ❍탄탄한 학습 알고리즘이 장점

❑머신러닝 3단계

- 데이터 정리 및 이해

시작하기 전에 꼭 기억하세요

❍모델의 학습

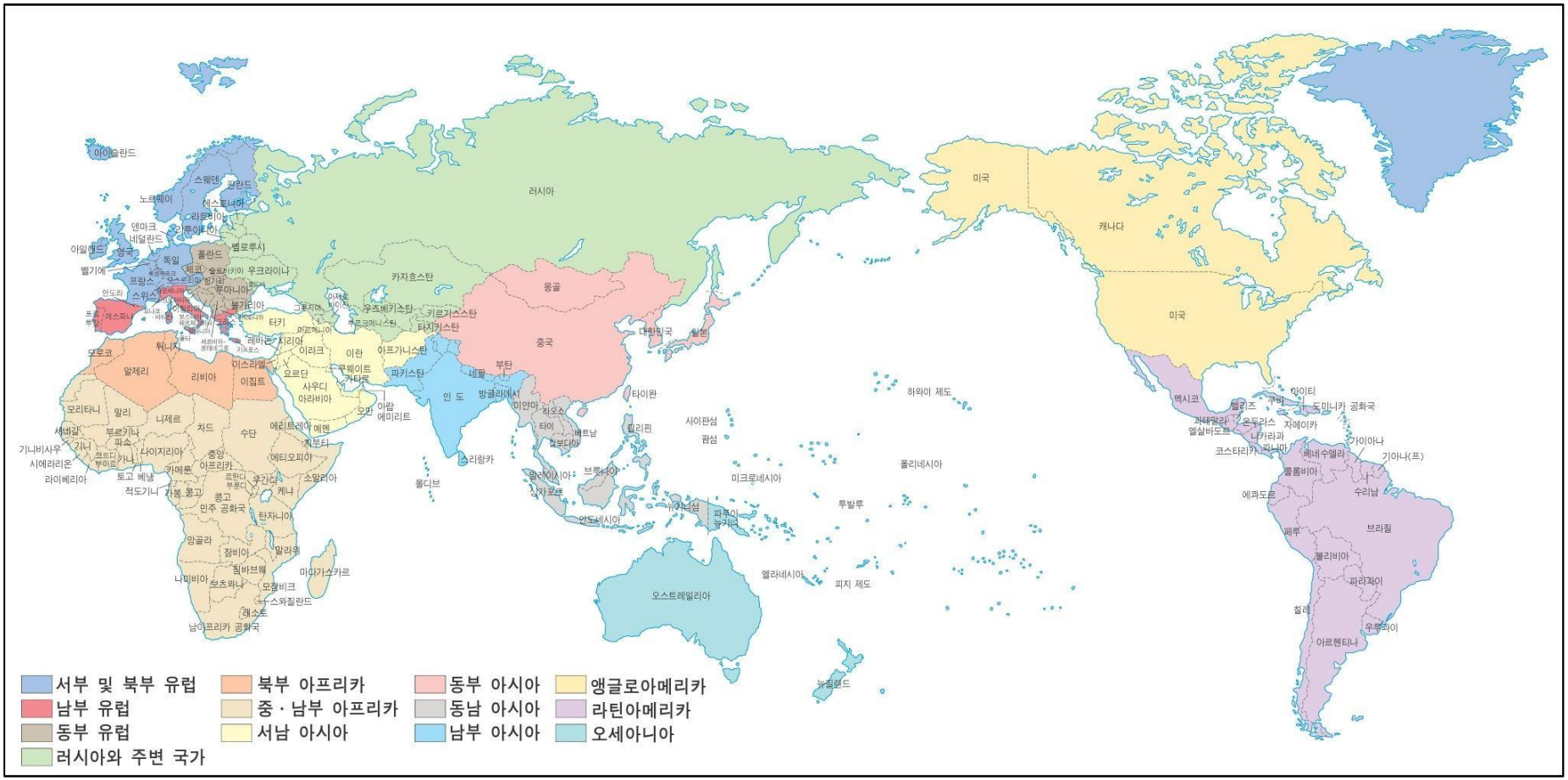
– 머신러닝

❍모델의 평가

– 예측 결과를 보고 평가

- 데이터를 충분히 보세요.
- 데이터로만 이해하세요.
- 데이터를 객관적으로 보세요.
- 당신은 생각보다 잘못 알고 있는게 많습니다.

❑러시아가 큰가요? 아프리카 대륙이 큰 가요?



❑17,100,000 VS 30,370,000 뭐가 더 큰가요?

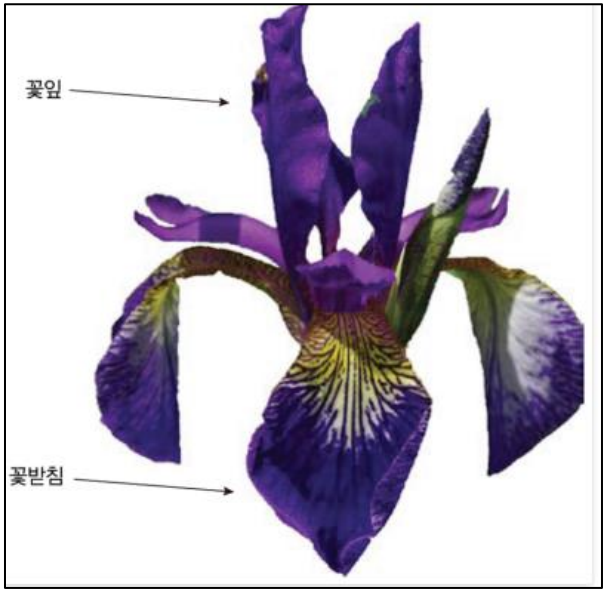
❑ 러시아 면적: 17,100,000

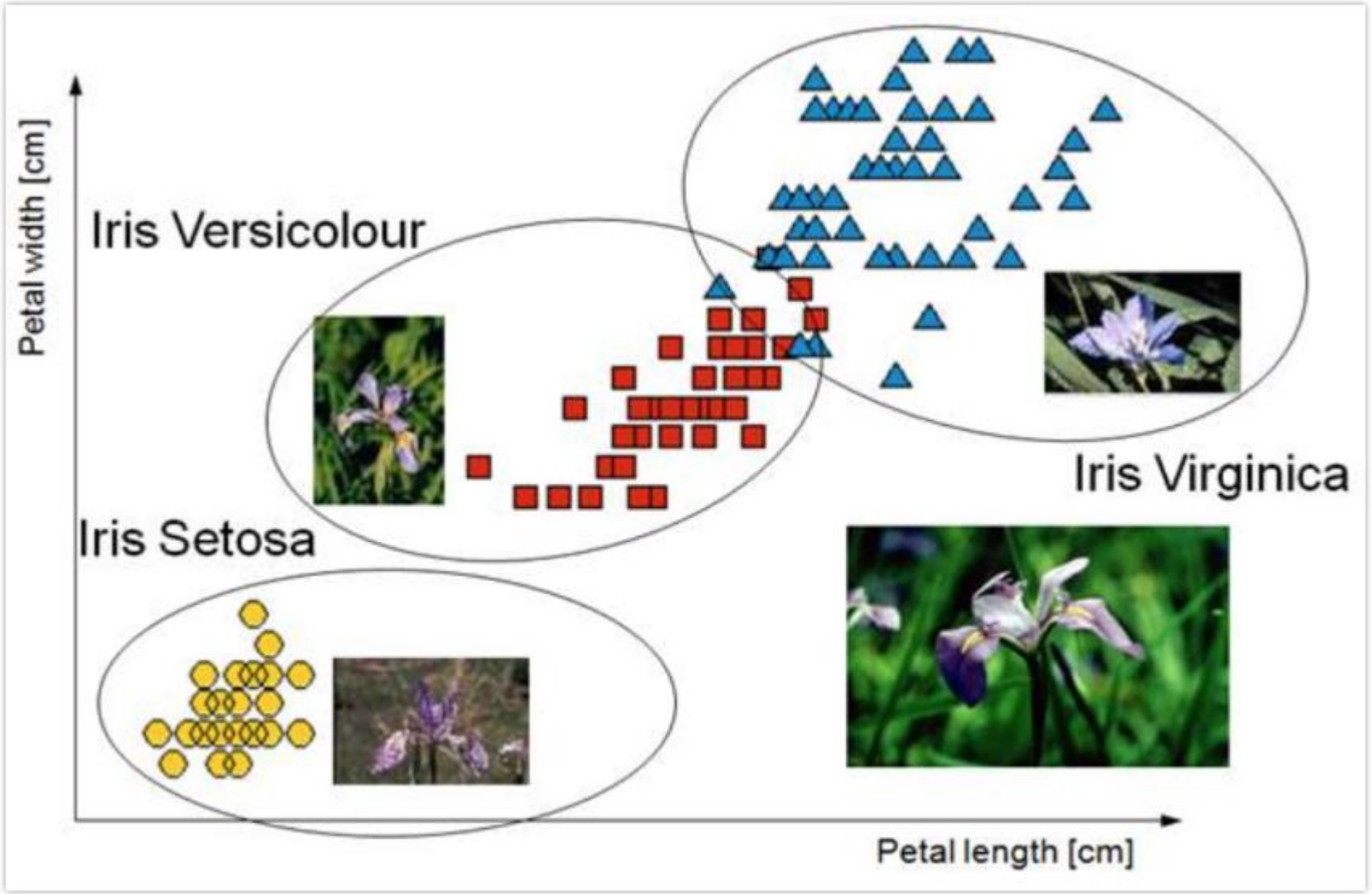
❑아프리카 면적: 30,370,000

❑메르카토르 도법 찾아보세요

❑데이터에 주관을 개입시키면 안 됩니다.

- 머신러닝 모델 만들기
- scikit learn에서 기본 제공하는 iris를 이용
- 꽃잎(petal)과 꽃받침(Sepal)의 길이와 폭을 가지고 세 개의 종을 분류하는 모델 만들기
  - Setosa, Vericolor, Virginica





출처 : http://articles.concreteinteractive.com/