Introduction to Data Mining Lecture

Week 2: Python and Data Mining

Joon Young Kim

Assistant Professor, School of AI Convergence
Sungshin Women's University

Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

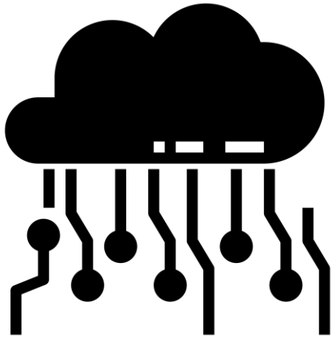
Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

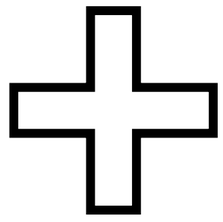
Week 1 Review

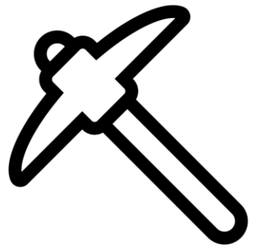
- Orientation and Overview of Data Mining

  → Data Mining 강의에서 배울 내용 및 핵심 요소

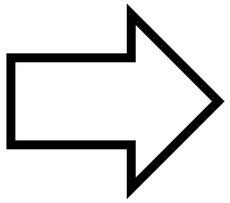


<Data>





<Mining>

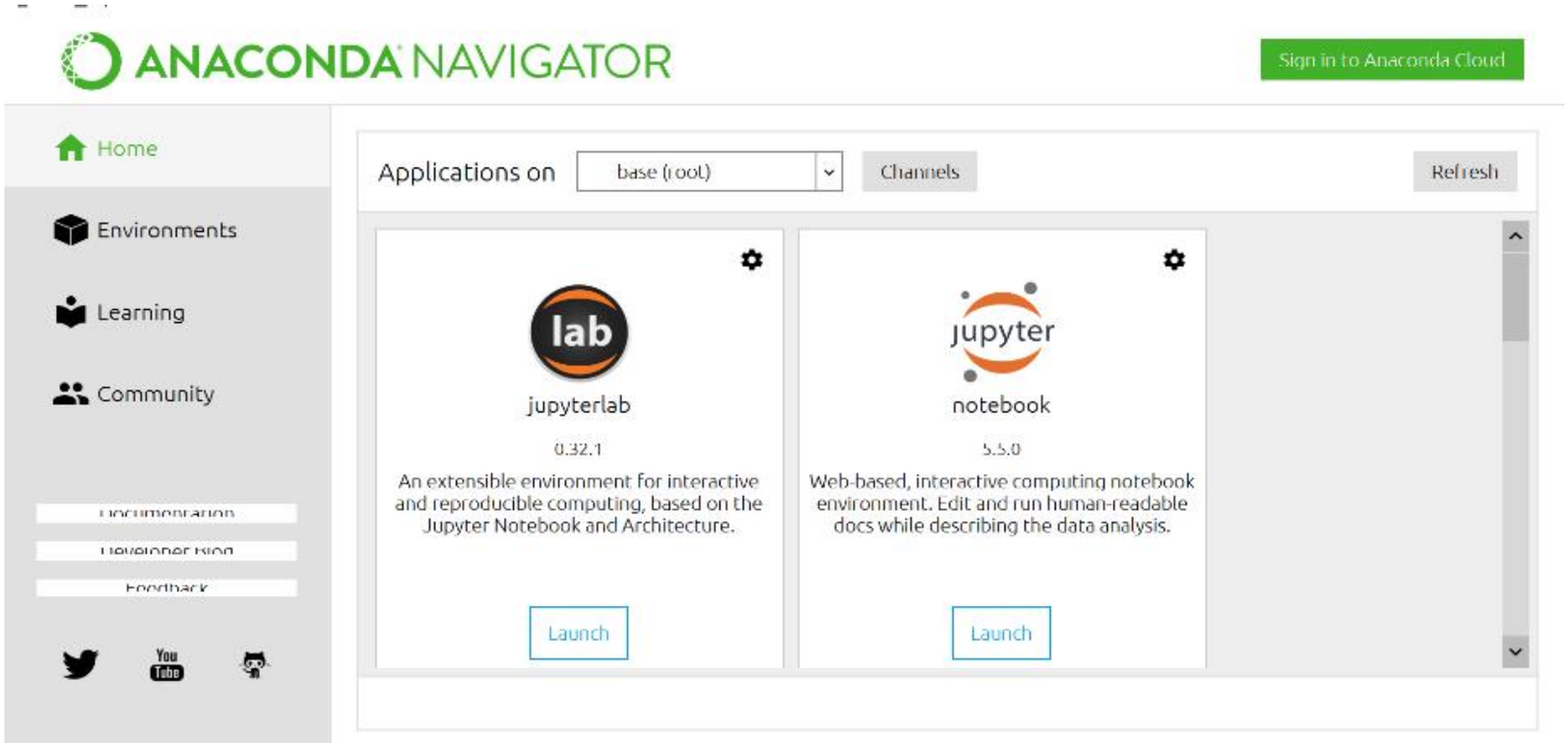


<Data Mining>

Week 1 Review

- Data Mining Concept & Environment Setup

  → 데이터 마이닝 컨셉 및 Anaconda3 설치/Jupyter Notebook 사용법



For those missing Week 1

- LMS 활용 예정
  → 숙제 제출, 질의 응답, 공지 사항, 강의 자료등
- 휴강등 예외 케이스 제외하고 녹화 컨텐츠는 없을 예정 → 반드시 출석 필수
- 본 강의에서 강의 슬라이드 자료 업로드는 기본적으로 없음
  → 건별로 요청시 업로드 고려 예정
- 문의사항: jkim@sungshin.ac.kr

For those missing Week 1

- 강의 대상 인원: 3~4학년들 대상 강의

- 기존 데이터마이닝 강의 결과 **파이썬 지식 없는 경우 대부분**

- 이론도 중요하지만 **실질적인 적용 방식을 모르고서는 활용 불가능**
  - 데이터마이닝내 이론 기반이 상당수이며 이론적인 학습만으로는 어려움
  - 적용이 중요하나 **상당한 수의 과제가 엑셀로 제출됨**

- 본 강의에서는 본격적인 파이썬 기반 데이터 활용 및 처리 학습 진행
  - 직접 활용 방법에 대한 실습 다수 진행 예정
  - 빅데이터분석과 비교 시 분류 및 예측 관련된 내용 일부 진행 예정

For those missing Week 1

- 다양한 환경 및 IDE 존재
  - 특정 IDE 강요 하지 않음 (자유롭게 활용 가능)
  - 숙제 제출은 py파일로만 제출 가능. **ipynb등 다른 확장자 제출 절대금지**

- 본 강의는 고학년 학부생들 대상
  - 기본적인 파이썬 프로그래밍 학습 및 IDE 활용 경험 있다고 간주
  - colab 및 외부 클라우드 개발 환경 경우 본 강의 진행시 활용 안할 예정

For those missing Week 1

- 빅데이터 분석 과목 내용 + 데이터마이닝 지식 포함 예정

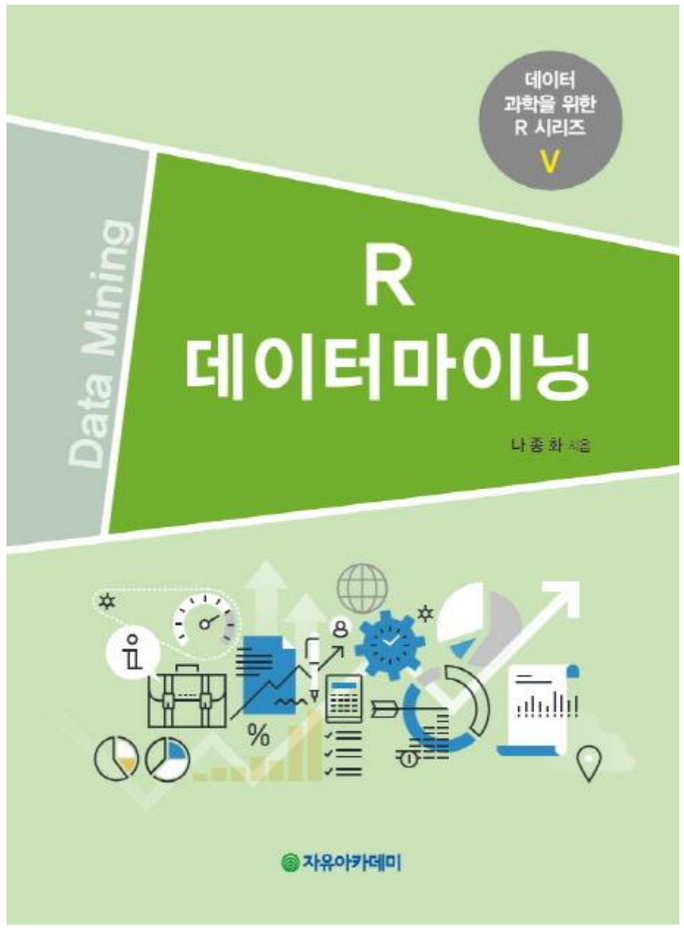
- 커리큘럼상 빅데이터 분석과 동일
  - 중간중간 데이터마이닝 기법 적용 실습 예정
  - 기본적인 패키지 예: scikit-learn

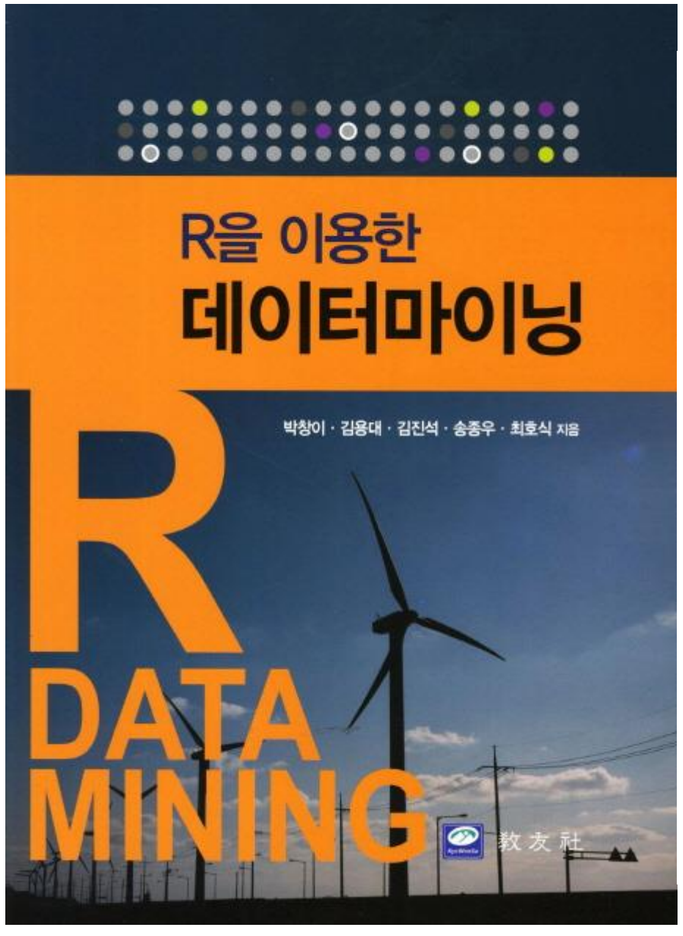
- 본 강의에서는 파이썬이 기본 언어임
  - 파이썬을 모를 경우 수강 지양을 권장함
  - 필요시 "혼자 공부하는 파이썬“ 책을 별도 구매할 것을 권장함
  - 기본적인 파이썬 문법을 알고 있다고 가정하고 수업 예정

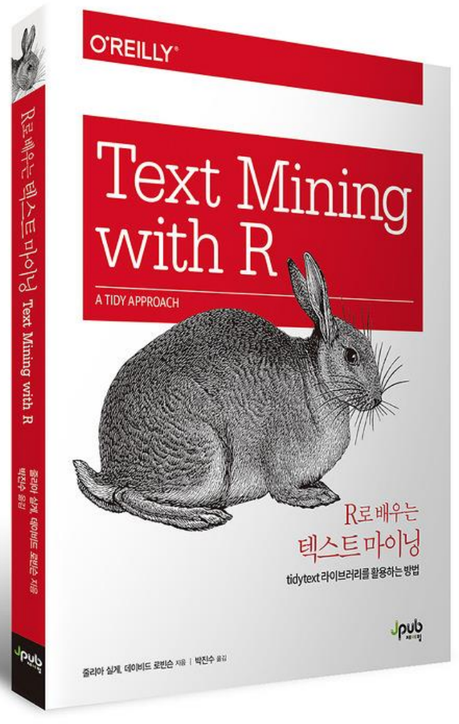
Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

R? Why?


데이터
과학을 위한
R 시리즈
V
Data Mining
R
데이터마이닝
나종화 지음
자유아카데미


R을 이용한
데이터마이닝
박창이 · 김용대 · 김진석 · 송종우 · 최호식 지음
R
DATA
MINING
教友社


O'REILLY
Text Mining
with R
A TIDY APPROACH
R로 배우는
텍스트마이닝
tidytext 라이브러리를 활용하는 방법
R로 배우는 텍스트 마이닝 Text Mining with R

Python v.s. R

- Python
  - 1991년 Guido Van Rossum이 개발
  - 인터프리터 언어
  - 다양한 어플리케이션과 통합운영 용이
  - 코딩과 디버깅이 편리

Python v.s. R

- R Programming
  - 1995년, Ross Ihaka와 Robert Gentleman이 개발
  - 인터프리터 언어
  - 강력한 통계처리 기능과 운용이 우수
  - 시각화 처리가 용이

Example

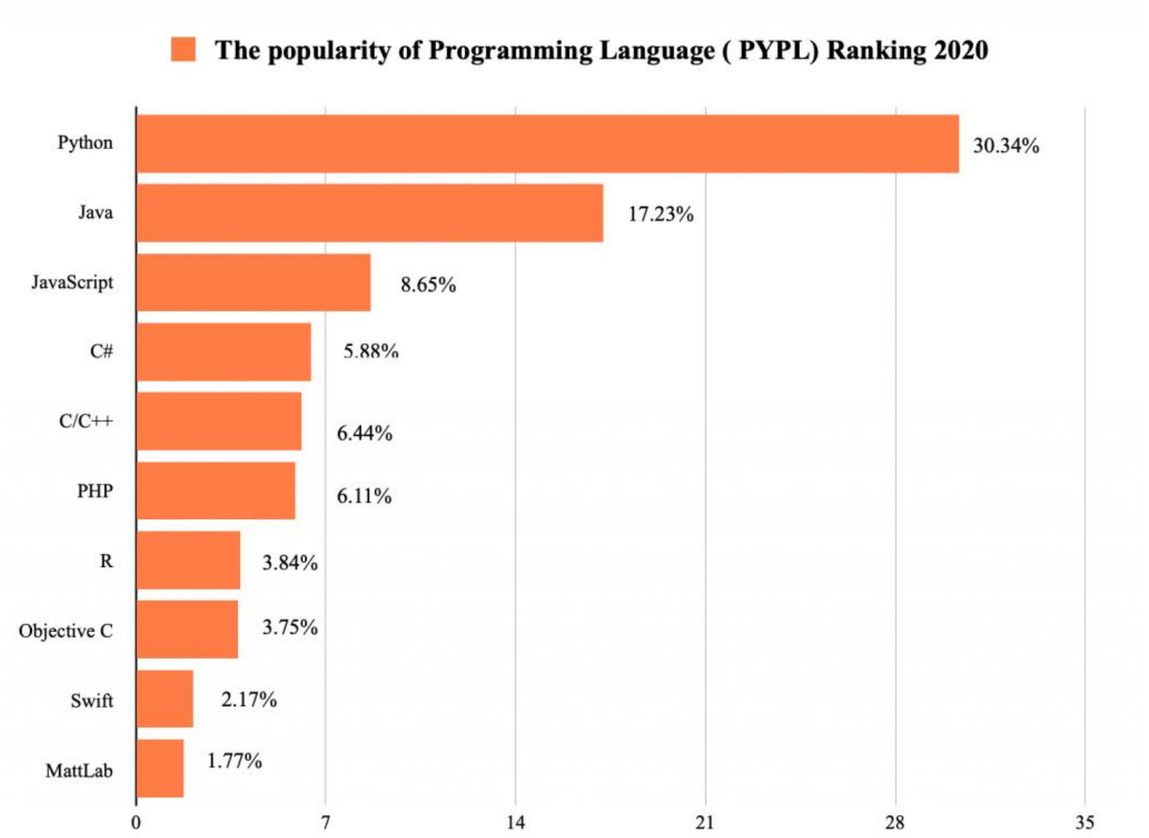
- Python
  - PyCharm
- R
  - R Program

Credentials for Language Choise

- 언어 선택 기준
  - 수행할 업무의 성격
  - 종사분야에서의 사용하는 도구
  - 언어 학습시 소요되는 비용과 시간
  - 종사분야에서 사용되는 어플리케이션과 상관 관계
  - 미래에도 통용되는지 여부

Reminder

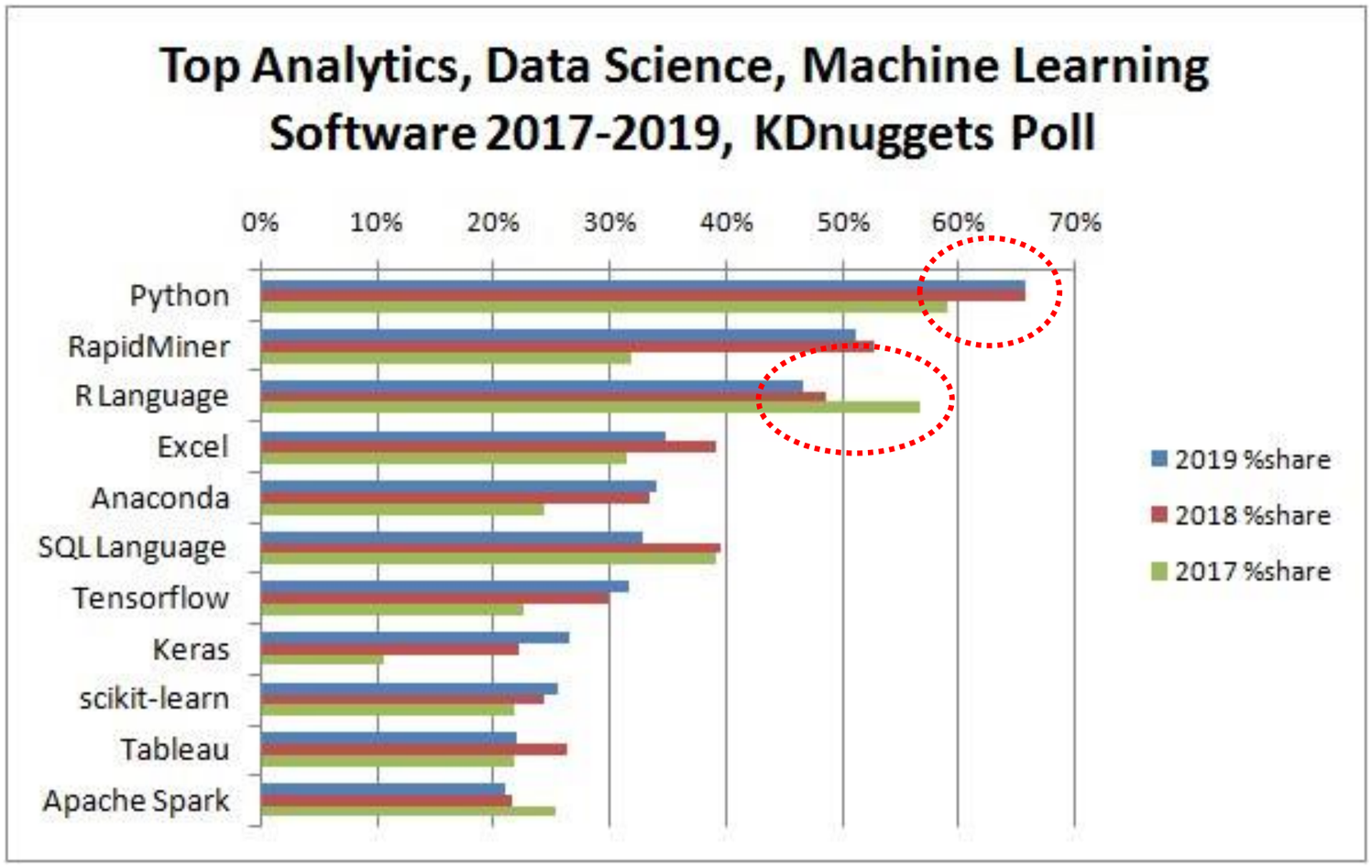
- 파이썬의 활용도 높음
  - 2020년 파이썬 랭킹 1위



https://darly.solutions/the-most-popular-programming-languages-in-2021/

Reminder

- 파이썬의 점유율
  - 2019년 KDnuggets Pool



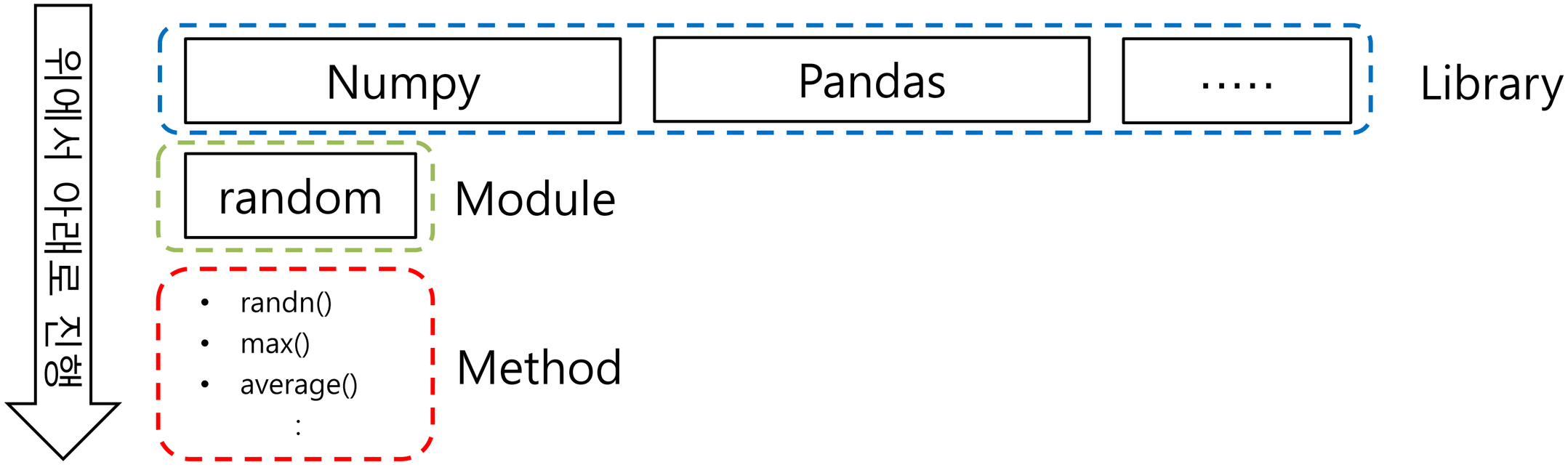
Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know about Python
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

Need-to-know about Python

- Object Oriented Programming Langauge

  → 주요 구성: Function, Method, Attribute, Property



Need-to-know about Python

- Object Oriented Programming Langauge

  → 주요 구성: Function, Method, Attribute, Property

```
from numpy import random

x = random.randn()

print(x)
```

Need-to-know about Python

■ 주요 라이브러리

→ Numpy: 과학기술 및 수학 계산

→ Pandas: 데이터 처리 및 분석

→ Matplotlib: 데이터 시각화

→ Seaborn: 데이터 시각화

→ Scipy: 신호처리, 최적화, 과학계산 및 통계 처리

→ scikit-learn: 최적화된 기계학습 제공

그냥 scikit-learn 쓰면 끗?

- Scikit-learn v.s. Tensorflow

  → ……**Scikit-learn** is used in practice with **a broader range of models**, whereas **TensorFlow**'s implied use is for **neural networks**…….

https://www.simplilearn.com/scikit-learn-vs-tensorflow-article#:~:text=Scikit%2Dlearn%20and%20TensorFlow%20were,use%20is%20for%20neural%20networks.

Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know about Python
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

Class

- Object and Class
  - 객체

    → 모든 것이 객체이고 클래스도 객체
  - 클래스

    → 과업을 수행하기 위해 코드를 블록으로 구성한 것

    → 메서드, 함수 및 속성을 가지는 코드 템플릿 또는 원형

Class

- Code Example

```
class MyPerson:
    i = 5
def __init__(self, name, age):
    self.name = name
    self.age = age

def bluechip(self):
    return “what is your name ?”
```

Instance

- Instance
  - 클래스에 있는 메서드나 속성을 사용하기 위한 절차
  - 변수 선언처럼 다른 변수 이름으로 반복적으로 선언 가능

```
p1 = MyPerson()
p2 = MyPerson()
      :
pn = MyPerson()
```

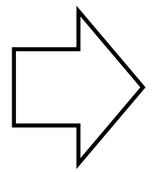
Instance

■ Code Example

```
class MyPerson:
    i = 5
def __init__(self, name, age):
    self.name = name
    self.age = age

def bluechip(self):
    return “what is your name ?”
```

```
p1 = MyPerson(“Jin”, 37)
```



```
p1.name
p1.age
p1.i
p1.bluechip()
```

Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

Module

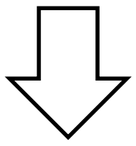
- Code Example
  - 클래스, 변수, 메서드, 함수 및 실행 코드를 포함하는 파일/패키지
  - 사용 방법: import 모듈명
  - Import를 발견하면 search path에 존재하는 모듈을 호출

→ 예) sys.path

Module

- Code Example

```
import sys
sys.path
```



```
['',
 'C:₩₩Anaconda3₩₩python36.zip',
 'C:₩₩Anaconda3₩₩DLLs',
 'C:₩₩Anaconda3₩₩lib',
 'C:₩₩Anaconda3',
 'C:₩₩Anaconda3₩₩lib₩₩site-packages']
```

Function and Method

- Function and Method
  - Java → Method
  - Python → Function, Method

- Python Function
  - User Defined Function
  - Nested (Inner) Function

Function and Method

- Python Function Example
  - User Defined Function

```
def add(a, b):
return a+b

add(3, 4)
```

  - Nested (Inner) Function

```
sm = sum([5, 15, 2])
print(sm)

mx = max(15, 6)
print(mx)
```

So….what's the difference?

- Python Methods v.s. Functions

Difference between Python Methods vs Functions

METHODS	FUNCTIONS
Methods definitions are always present inside a class.	We don't need a class to define a function.
Methods are associated with the objects of the class they belong to.	**Functions are not associated with any object.**
A method is called 'on' an object. We cannot invoke it just by its name	**We can invoke a function just by its name.**
Methods can operate on the data of the object they associate with	**Functions operate on the data you pass to them as arguments.**
Methods are dependent on the class they belong to.	Functions are independent entities in a program.
A method requires to have 'self' as its first argument.	Functions do not require any 'self' argument. They can have zero or more arguments.

Source: https://techvidvan.com/tutorials/python-methods-vs-functions/

Function and Method

- Python Example
  - Function

```
def add(a, b):
return a+b

add(3, 4)
```

  - Method

```
class mymath:
    def add(a, b):
        return a+b

p1 = mymath()
p1.add(3, 4)
```

Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

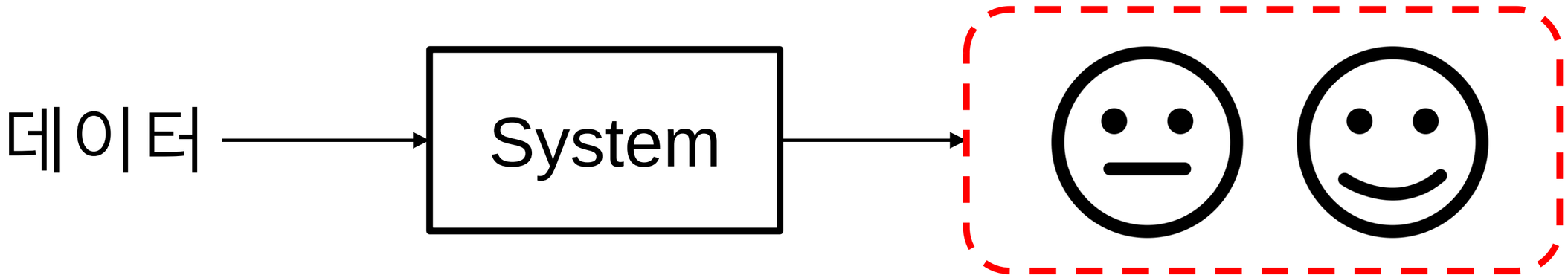
Key Ideas of Data Mining

- 주요 컨셉
  - 감지, 분류, 예측
  - 연관 규칙
  - 데이터 축소 및 탐색
  - 데이터 시각화
  - 기타등등

→ 이외에도 다양한 아이디어들이 존재한다. 여기서 끝이 아니다.

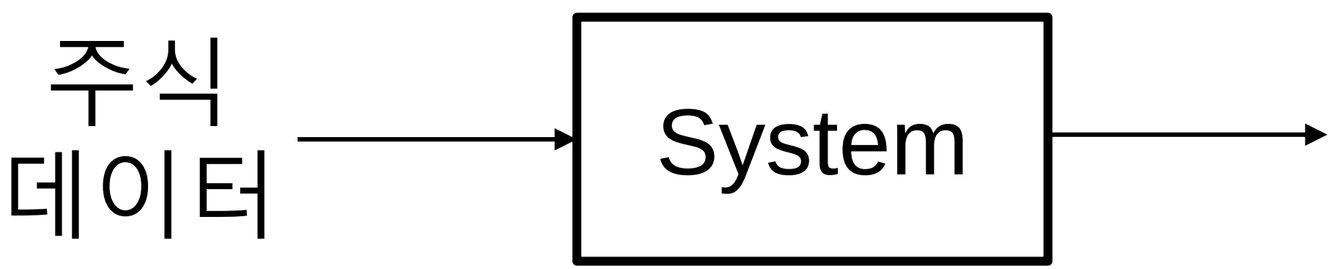
Core Ideas of Data Mining

- Detection/Classification



Happy, Sad, …

- Prediction



상승장, 하락장, 보합세, …

Core Ideas of Data Mining

- Association Rules

  → 데이터 사이의 연계 룰 및 세팅

  (예: 넷플릭스 → 보던 영화/드라마 연계 프로그램 전시)

Core Ideas of Data Mining

- Data Reduction

  → 데이터 셋의 양/사이즈 축소화

  1) 복잡성으로 인한 적절한 데이터 처리 불가능

  2) 데이터 사이즈로 인한 처리 시간 증대 문제

  (예: 짜장면, 짬뽕, 유산슬, 탕수육 ….. → 짜장면, 짬뽕 in 광역시 only)

- Data Exploration

  → 데이터 셋의 의미 파악 및 차이 분석

  → 이를 통한 데이터 자체의 스케일 혹은 차원 축소 가능

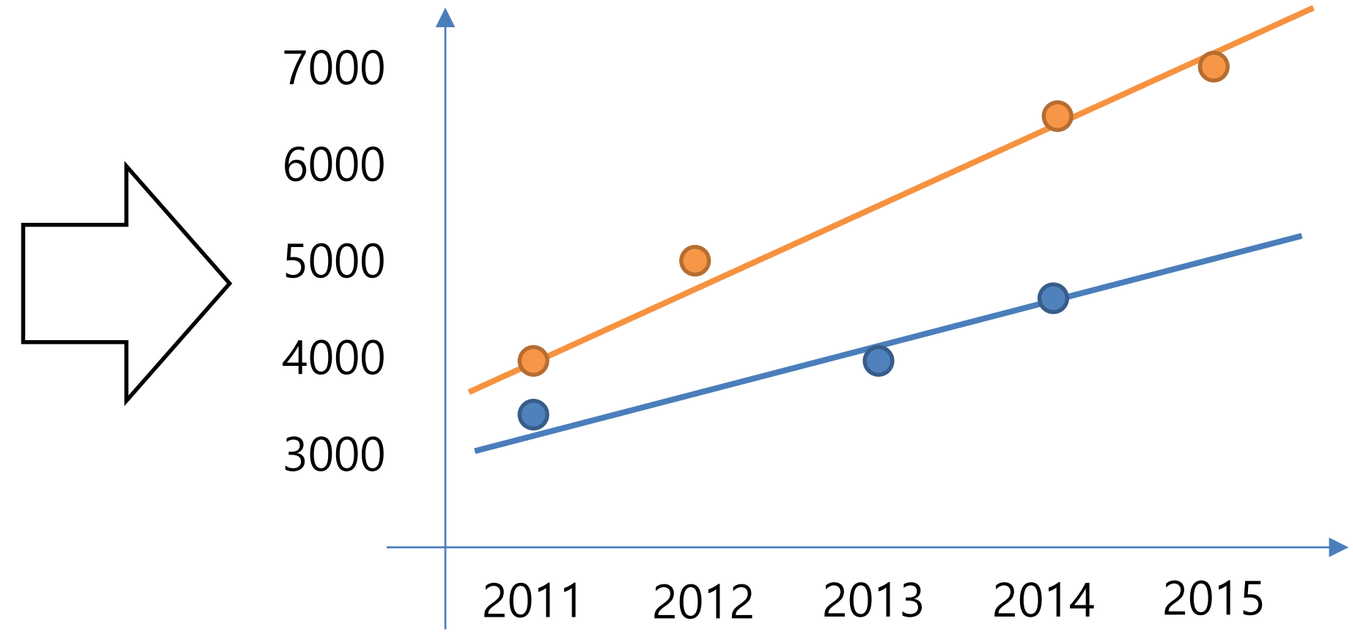
  (예: 짜장면/짬뽕의 소비자지수, 이미지내 얼굴 인식 기법)

Core Ideas of Data Mining

- Data Visualization

  → 데이터 셋 기반의 시각화를 통한 직관적 데이터 연결성/인사이트 도출

종류	년도	가격(원)
짬뽕	2011	4000
짜장면	2011	3500
짬뽕	2012	5000
짜장면	2013	4000
짬뽕	2014	6500
짜장면	2014	4500
짬뽕	2015	7000



Core Ideas of Data Mining

- Data Visualization

  → 데이터 셋 기반의 시각화를 통한 직관적 데이터 연결성/인사이트 도출

시도별	품목별	2000. 01	2000. 02	2000. 03	2000. 04
전국	자장면	58.003	58.295	58.413	58.472
서울특별시	자장면	60.939	60.939	60.939	60.939
부산광역시	자장면	59.409	59.409	59.409	59.409
대구광역시	자장면	59.869	59.869	59.869	59.869
인천광역시	자장면	56.795	56.795	56.795	56.795
광주광역시	자장면	47.204	47.204	47.204	47.204
대전광역시	자장면	48.783	48.783	48.783	48.783
울산광역시	자장면	49.864	53.666	53.666	53.666

Source: KOSIS, https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1J15112&conn_path=I2

Week 2 Outline

- Week 1 Review
- Python v.s. R
- Need-to-know for Python in Big Data
- Class and Instance
- Module, Function and Method
- Key Ideas of Data Mining
- Conclusion

Week 2 Key Takeaway

- 데이터마이닝를 위한 파이썬 개요

  → 파이썬/R 비교 및 함수, 메소드등 객체지향 언어 내용

- 함수/메소드 실제 코드 실습

  → 함수와 메소드의 차이점 및 라이브러리부터 메소트까지 구조 학습

다음 장에서는

- numpy, numpy, and numpy