

Introduction to Data Mining

Joon Young Kim

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

Attendance Check (출석 체크)

스마트캠퍼스 통한 출결 진행 예정

수업 도중 질의/문의사항이 있으면
채팅으로 알려주세요

Q & A 페이지에서는 직접 문의 가능

Data Mining Lecture

Week 1: Course Overview & (Introduction to) Data Mining

Joon Young Kim

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

Week 1 Outline

- Orientation & Syllabus
- Course Overview
- What and Why Data Mining?
- Why Python?
- Conclusion

Week 1 Outline

- Orientation & Syllabus
- Course Overview
- What and Why Data Mining?
- Why Python?
- Conclusion

Brief Orientation

과목명(Course Name): 데이터 마이닝 개론, Introduction to Data Mining

강의 교수(Course Instructor): 김준영

강의 일정 (Week Schedule): W 12:00PM~3:00PM

강의 개요 (Course Description)

- 데이터마이닝 개요 및 개념
 - 주요 데이터마이닝 기법에 대한 방법론 및 실 활용 예제 설명
- 파이썬을 활용한 기본적인 데이터 처리 및 분석을 배울 예정
- 현재 많은 관심을 받고 있는 machine learning의 일부 개념을 배운다.

문의사항: jkim@sungshin.ac.kr

Need-to-Know

- 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

Expectation

- 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 this course

- 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

Course Schedule (1)

- 과제: 2주부터 6주까지 2~3주에 1개씩 (4~5점)
- 중간고사: 1주~7주 내용

일정	내용
1주	수업 개요 & 개발 환경 구축
2주	파이썬빅데이터 프로그래밍
3주	넘파이 (1)
4주	넘파이 (2)
5주	판다스 (1)
6주	판다스 (2)
7주	판다스 고급 (1)
8주	중간고사 (1주 ~ 7주)

Course Schedule (2)

- 과제: 9주부터 13주까지 2~3주에 1개씩 (4~5점)
- 기말고사: 1주~14주 내용

일정	내용
9주	판다스 고급 (2)
10주	matplotlib (1)
11주	matplotlib (2)
12주	시계열
13주	어플리케이션 (1)
14주	어플리케이션 (2)
15주	기말고사 (1~14주)

Textbook

주교재

- 파이썬을 이용한 데이터 분석의 정석 (넴파이, 판다스, 맷플롯립과 실전 예제로 배우는), 채진석 지음, 2021
 - 해당 교재를 기반으로 수업 진도 진행할 예정

넴파이, 판다스, 맷플롯립과 실전 예제로 배우는

채진석 지음



If you are interested

추가문헌

- Building Machine Learning Systems with Python, Richert Coelho, 2013

※ 영문 문헌임을 참고 부탁

Building Machine Learning Systems with Python

Third Edition

Explore machine learning and deep learning techniques for building intelligent systems using scikit-learn and TensorFlow



By Luis Pedro Coelho, Willi Richert
and Matthieu Brucher

Packt
www.packt.com



Community Experience Distilled

Learning scikit-learn: Machine Learning in Python

Experience the benefits of machine learning techniques by
applying them to real-world problems using Python and the
open source scikit-learn library

Raúl Garreta
Guillermo Moncecchi

[PACKT] open source★
PUBLISHING

Resources

실 문제 해결 등의 경우 직접 해결 권장

- 대부분의 문제 해결은 책에 나와 있음
- 필요시 다양한 사이트들을 통해서 해결 필요

모든 문제의 해결 실마리들은 전부 여기로

- <http://www.google.com> (한국어/영어 둘다)

정 안될시 저에게 이메일로 문의 부탁

최소 요구사항 (Minimum Requirement)

- 파이썬 프로그래밍 활용 능력

- 통계 수학에 대한 이해가 있으면 플러스

- 파이썬 프로그래밍 이해 부족 시 본 강좌 수강 지양 권장

- 숙제는 대체로 Python 코드 제출인 점 감안

- 본 강의는 데이터마이닝 과목이지 Python 과목이 아님

- 문제해결 및 필요시 질의응답 꼭

Evaluation for 데이터마케팅개론

대면 상대평가 (A 30% 이하)

- 중간고사: 35%
- 기말고사: 35%
- 과제물: 20%
- 출석: 10%

총합: 100%

A(+/-): 상위 30%

B(+/-)

C(+/-)

D(+/-)

→ 제한없음

출석/과제물 꼭 챙기세요.

Coffee Break

Week 1 Outline

- Orientation & Syllabus
- **Course Overview**
- What and Why Data Mining?
- Why Python?
- Conclusion

데이터 마이닝 개론 과목

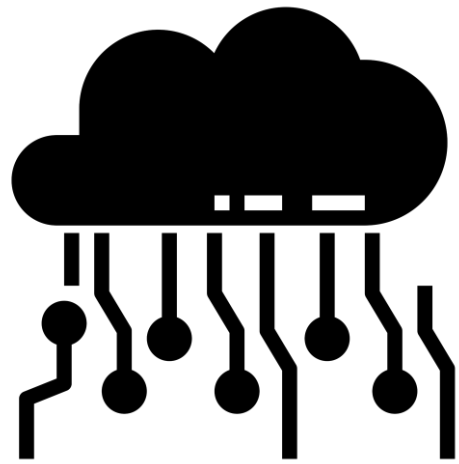
- 데이터마이닝 개요 및 개념
 - 주요 데이터마이닝 기법에 대한 방법론 및 실 활용 예제 설명
- 파이썬을 활용한 기본적인 데이터 처리 및 분석을 배울 예정
- 현재 많은 관심을 받고 있는 machine learning의 일부 개념을 배운다.

정보/지식의 분석 및 추출, 정형화 개념 학습 진행

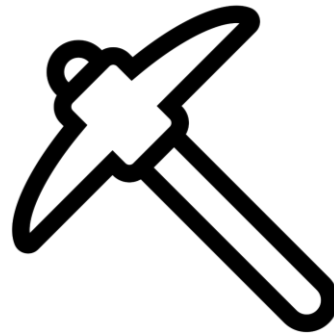
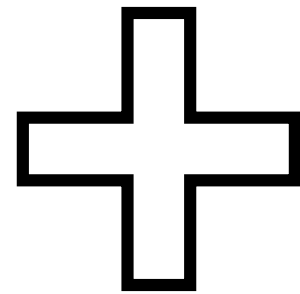
Data Mining

■ Data + Mining

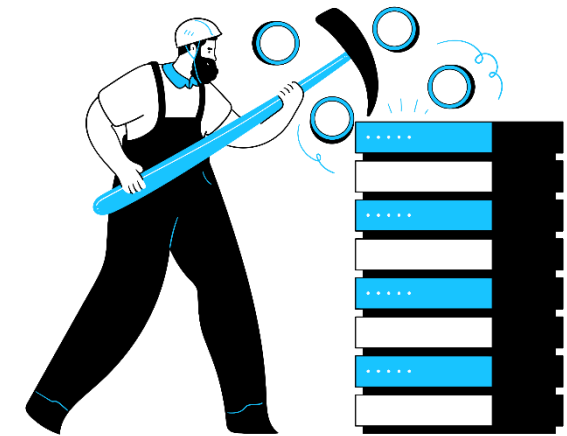
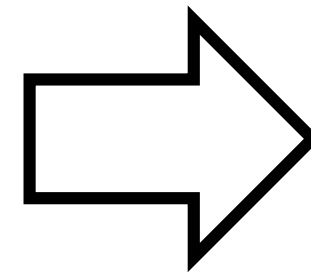
→ 데이터/정보 + 캐내다/발굴하다



<Data>



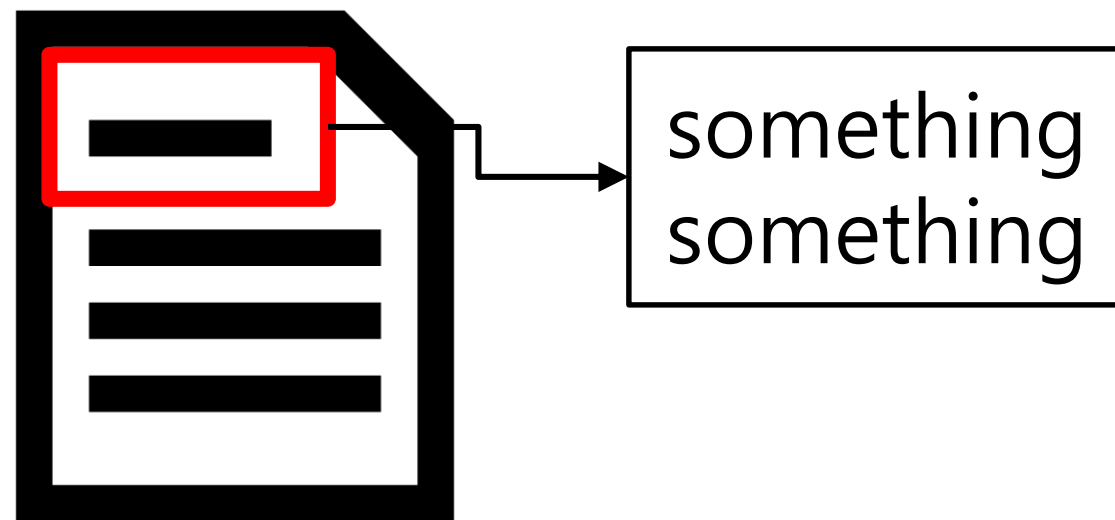
<Mining>



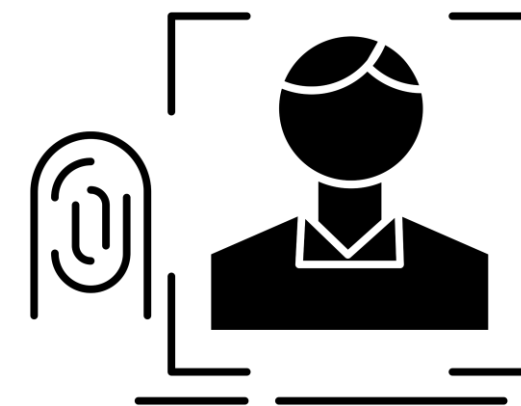
<Data Mining>

Before data mining

- Text Mining(텍스트 마이닝) & Pattern Recognition (패턴 인식)
→ Data Processing/Detection/Classification



<Text Mining>



<Pattern Recognition>

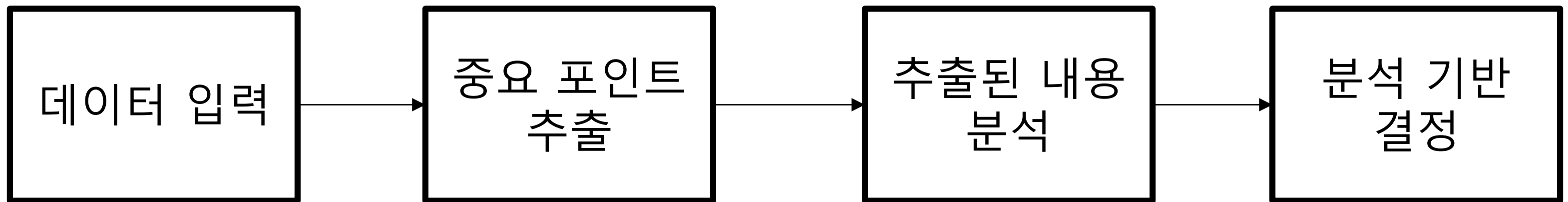
데이터 마이닝 개론 과목

- 데이터마이닝 개요 및 개념
 - 주요 데이터마이닝 기법에 대한 방법론 및 실 활용 예제 설명
- 파이썬을 활용한 기본적인 데이터 처리 및 분석을 배울 예정
- 현재 많은 관심을 받고 있는 machine learning의 일부 개념을 배운다.

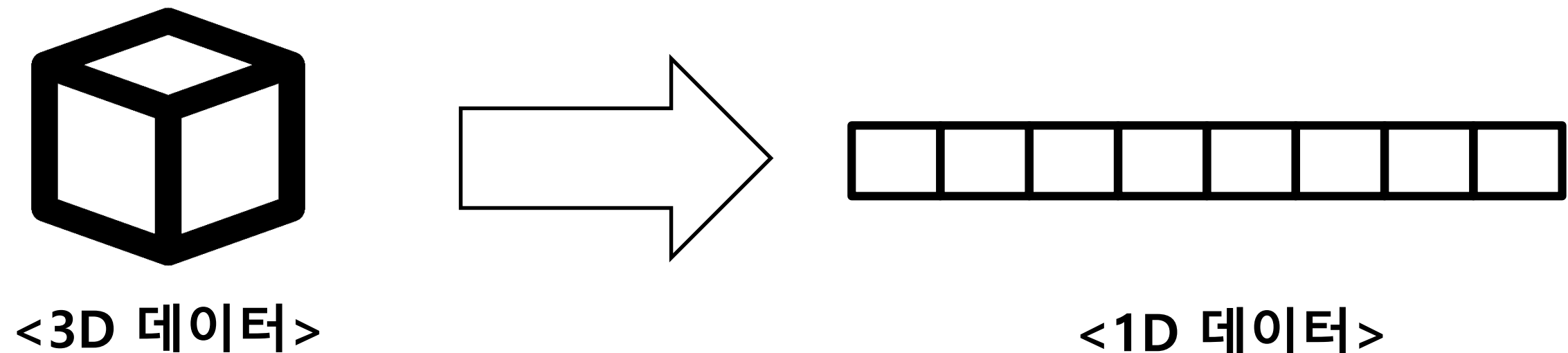
기계/심층 학습 기초를 위한 데이터 마이닝 개념 학습

Main Contents for Data Mining

■ Data Mining Process (In-a-nutshell)

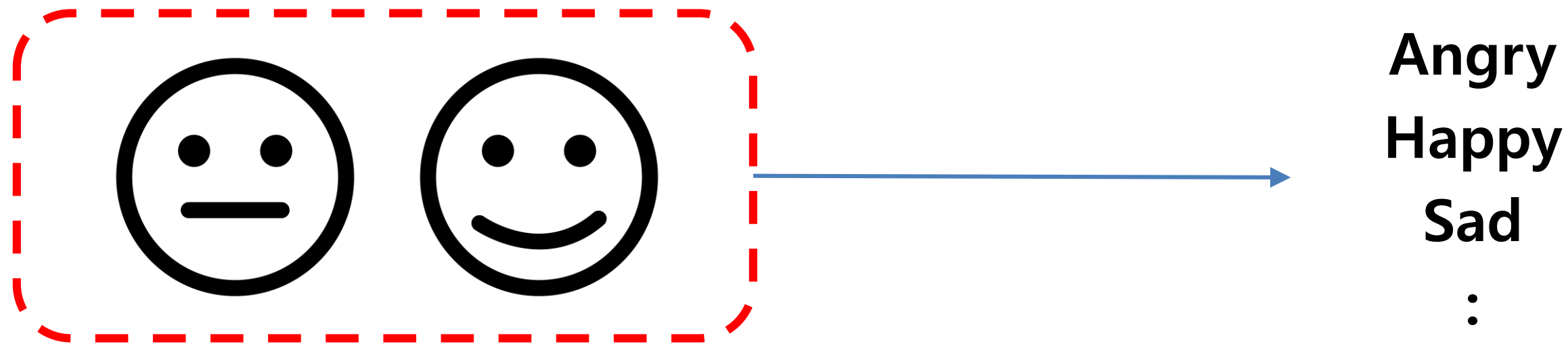


■ Data Compression



Main Contents for Data Mining

■ Detection/Classification



■ Prediction/Estimation



데이터 마이닝 개론 과목

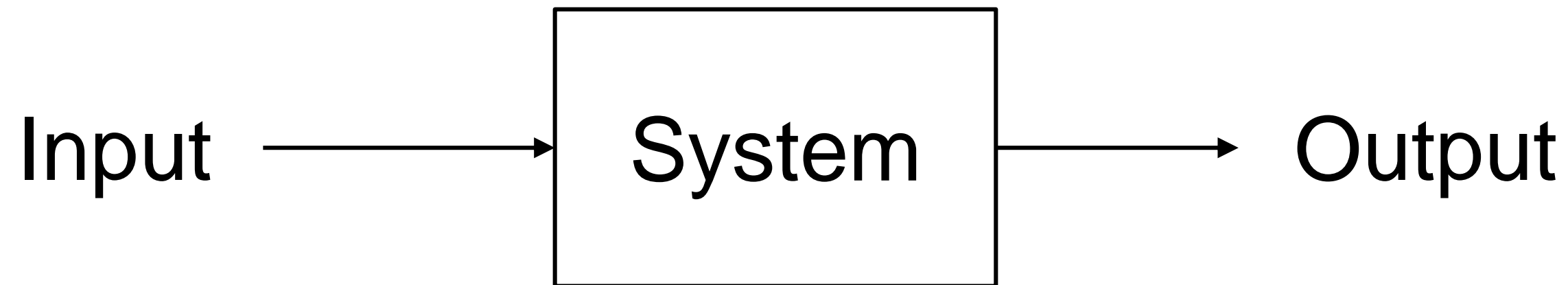
- 데이터마이닝 개요 및 개념
 - 주요 데이터마이닝 기법에 대한 방법론 및 실 활용 예제 설명
- 파이썬을 활용한 기본적인 데이터 처리 및 분석을 배울 예정
- 현재 많은 관심을 받고 있는 machine learning의 일부 개념을 배운다.

기계/심층 학습 기초를 위한 데이터 마이닝 개념 학습

Concept of “System”

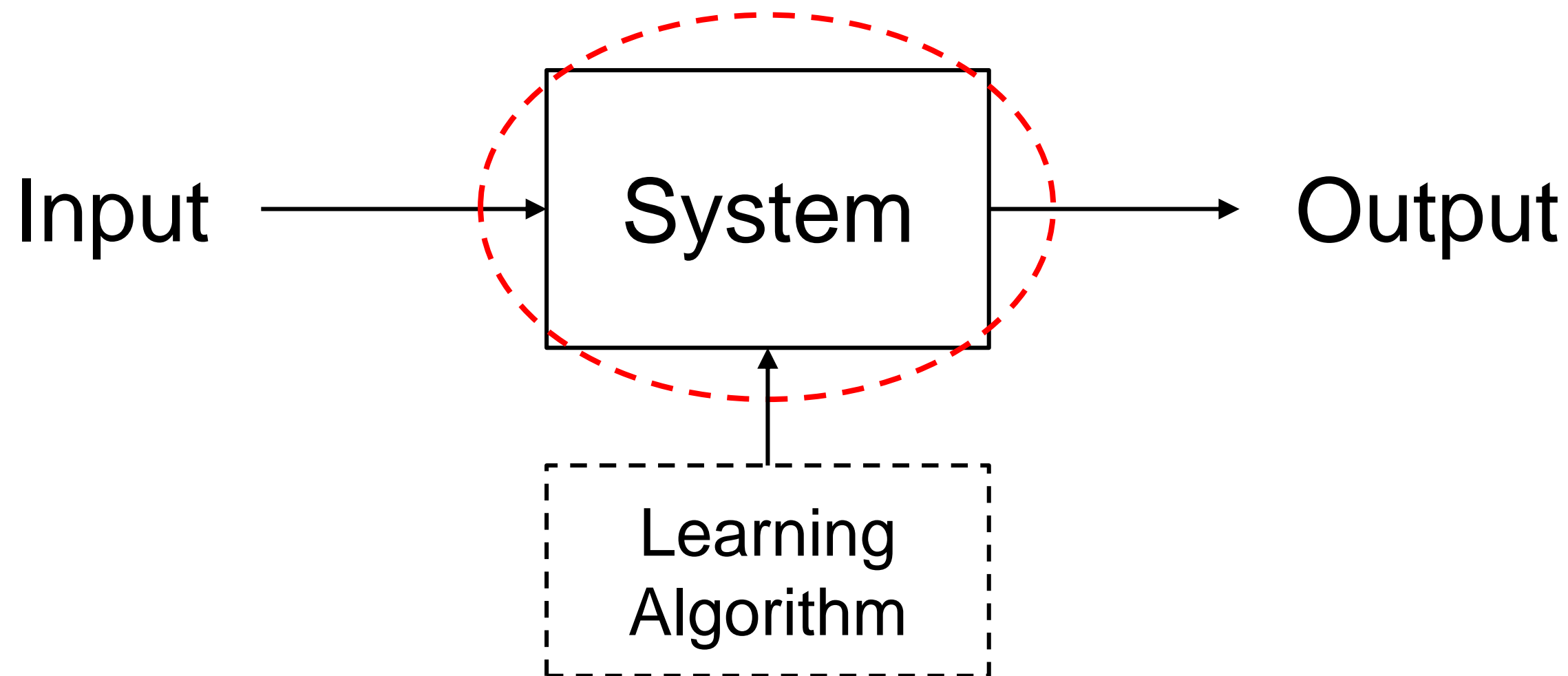
■ 시스템의 기본 구조

→ 입력/시스템/출력으로 구성



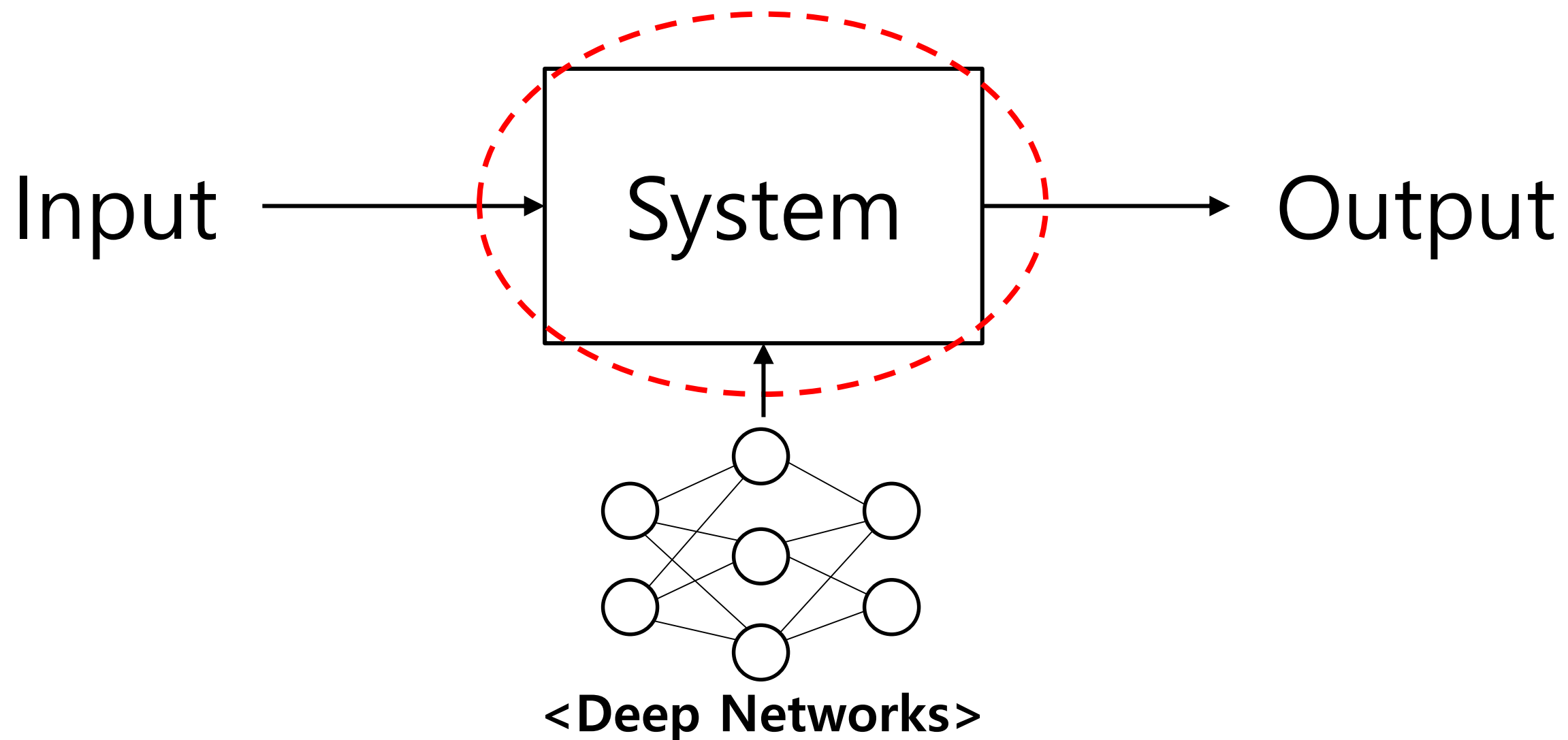
What is Machine Learning?

- 시스템, 즉 기계를 훈련 → 기계학습
 - 러닝 알고리즘 통한 시스템 내 파라미터/알고리즘 최적화
 - 인간과 비교 시 다수의 한계점 존재(예: 90년대 음성인식)



What is Deep Learning?

- 시스템, 즉 기계를 더 깊은 레벨에서 (**Deep**) 훈련
 - 복잡한/복합적인 러닝 알고리즘 적용
 - 기존 한계점 극복 사례 다수 진행중 (예: 음성/얼굴인식)



So, What now?

주요 사항

Data Mining 기본 개념 및 Python 통한 실 적용 학습

Data Mining을 위한 데이터 처리의 이해 필수

강의+여러분들의 관심으로 전체 커버가 가능

저와 여러분이 같이 배워나가는 과목임을 명심하시길

데이터 마이닝 개론 과목

- 데이터마이닝 개요 및 개념
 - 주요 데이터마이닝 기법에 대한 방법론 및 **실 활용 예제 설명**
- 파이썬을 활용한 기본적인 데이터 처리 및 분석을 배울 예정
- 현재 많은 관심을 받고 있는 machine learning의 일부 개념을 배운다.

실제 어플리케이션 통한 활용 예제 학습

데이터 마이닝 개론 과목

분류	교재명	저자	출판사	출판년도
주교재	(파이썬을 이용한)데이터 분석의 정석	채진석 지음	루비퍼이퍼	2022
주 별 수 업 내 용				
주/회차	수업내용		수업방법	교재진도/과제
1주 1회차	수업 개요 & 개발 환경 구축		대면	
2주 2회차	파이썬빅데이터 프로그래밍		대면	
3주 3회차	넘파이 (1)		대면	
4주 4회차	넘파이 (2)		대면	
5주 5회차	판다스 (1)		대면	
6주 6회차	판다스 (2)		대면	
7주 7회차	판다스 고급 (1)		대면	
8주 8회차	중간고사		대면	
9주 9회차	판다스 고급 (2)		대면	
10주 10회차	matplotlib (1)		대면	
11주 11회차	matplotlib (2)		대면	
12주 12회차	시계열		대면	
13주 13회차	어플리케이션 (1)		대면	
14주 14회차	어플리케이션 (2)		대면	
15주 15회차	기말고사		대면	

Q & A?

Week 1 Outline

- Orientation & Syllabus
- Course Overview
- What and Why Data Mining?
- Why Python?
- Conclusion

What is Data Mining?

Data Mining⁺

The practice of **analyzing large databases** in order to **generate new information**.

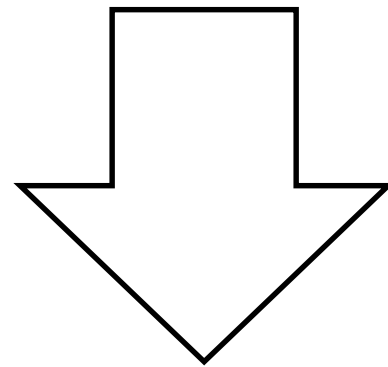


History (?) of Data Mining

- Bayes' Theorem
- Regression Analysis
- Neural Networks
- Evolutionary Computation

History (?) of Data Mining

- Then, here comes “database mining”
 - Recognized as a sub-process or a step within a larger process called Knowledge Discovery in Databases (KDD)

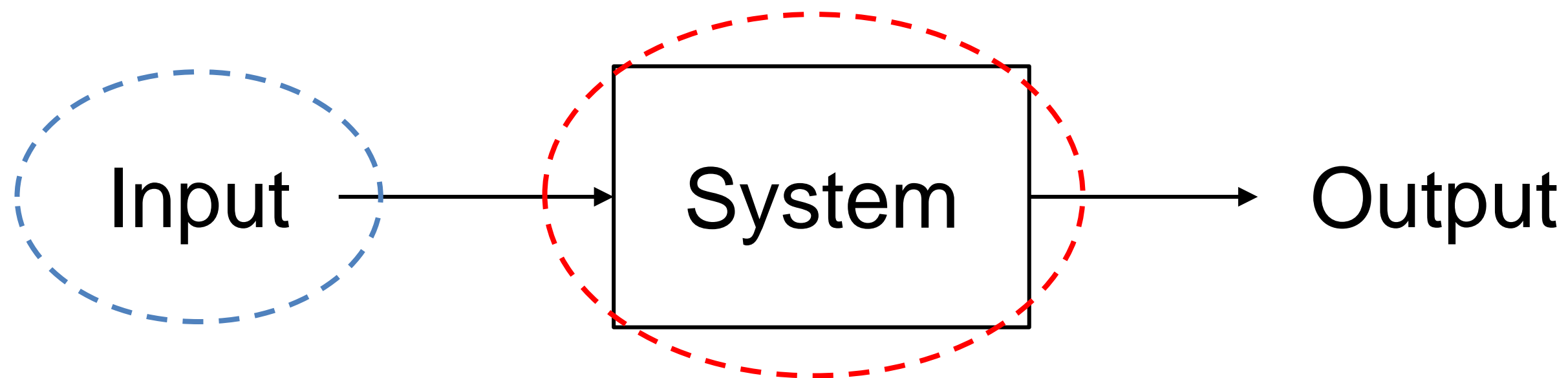


Data Mining

■ Real Definition⁺

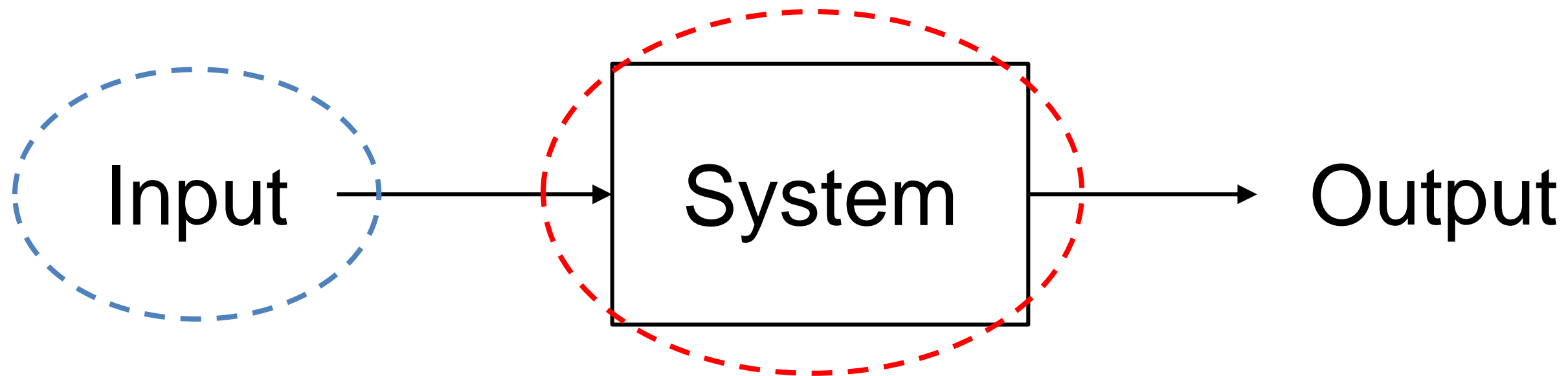
→ The field of **discovering** novel and potentially **useful information** from **large amounts of data**

■ Features



Narrative Example

■ In the Basic System

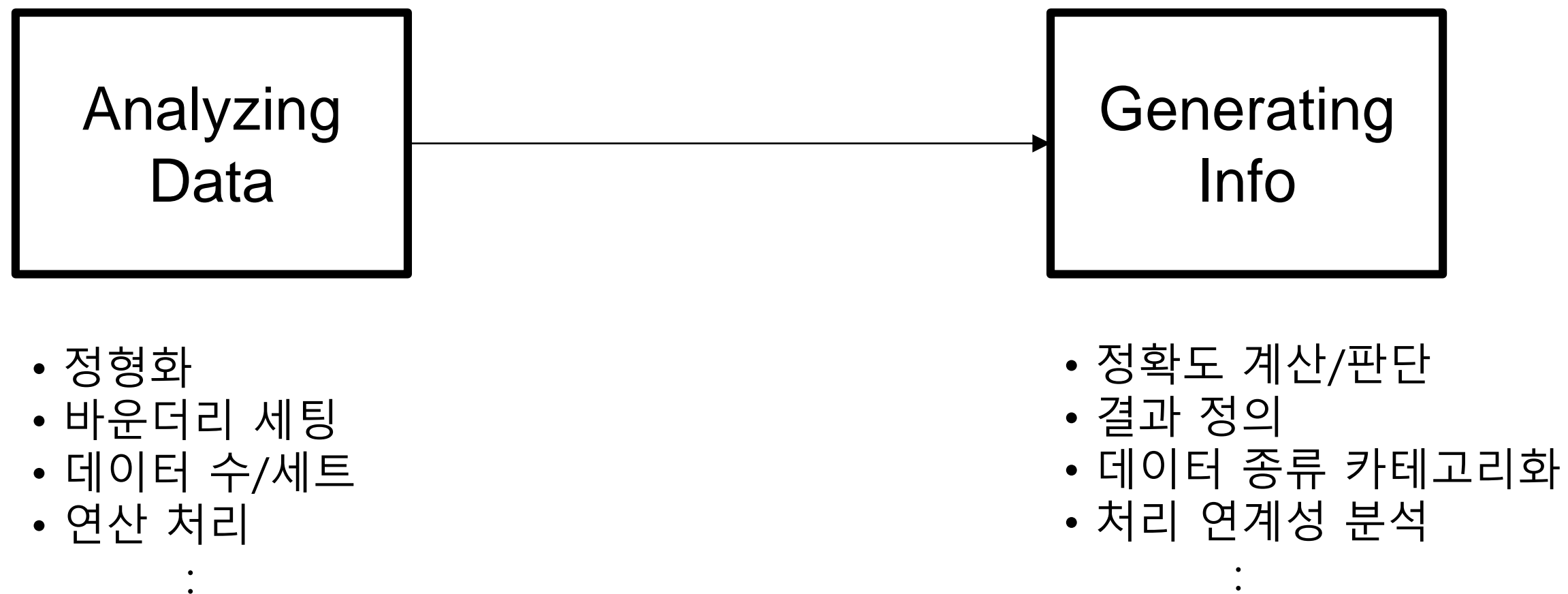


- 텍스트
- 이미지
- 음성 & 소리
- 데이터베이스
- 클라우드 공간
- :

- 데이터 처리
- 알고리즘 학습
- 특징점 추출
- 예측 & 추정
- :

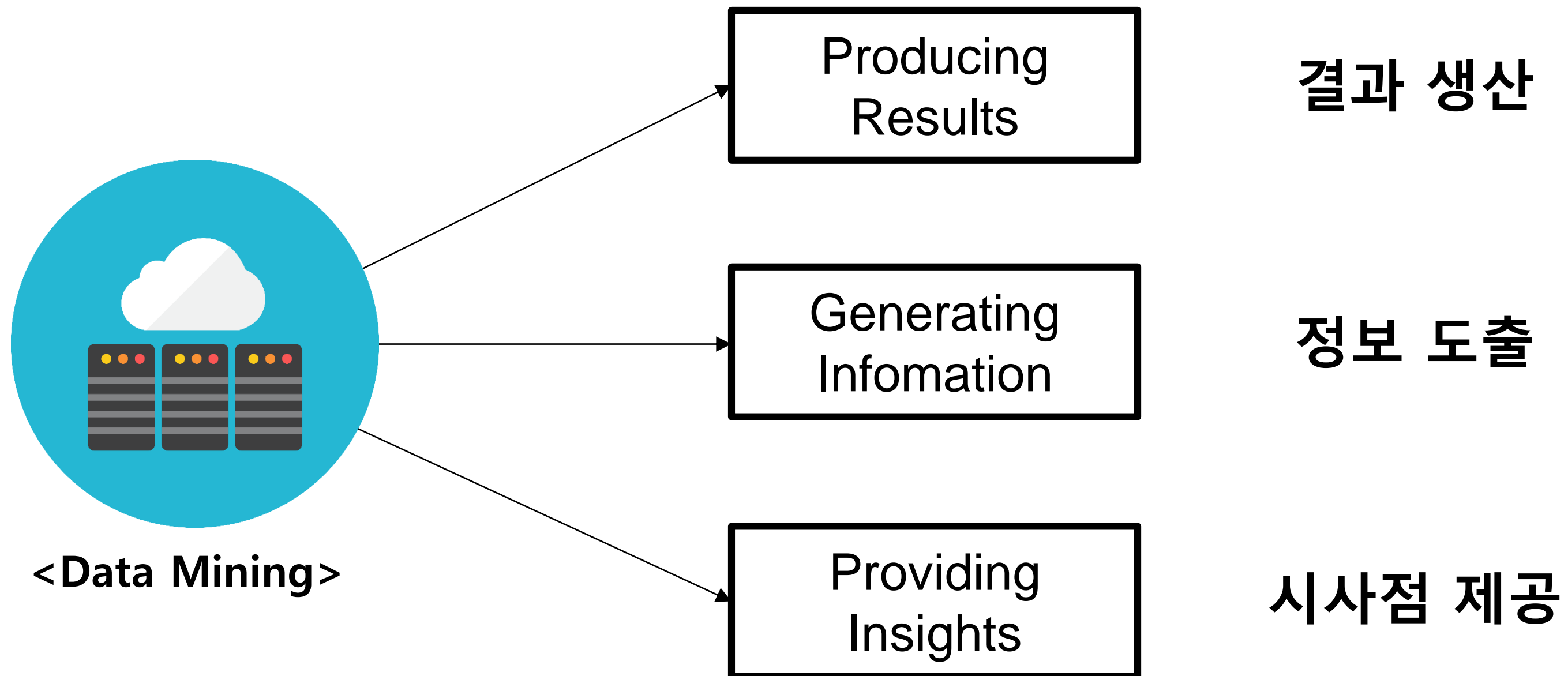
Narrative Example

■ Role of Data Mining



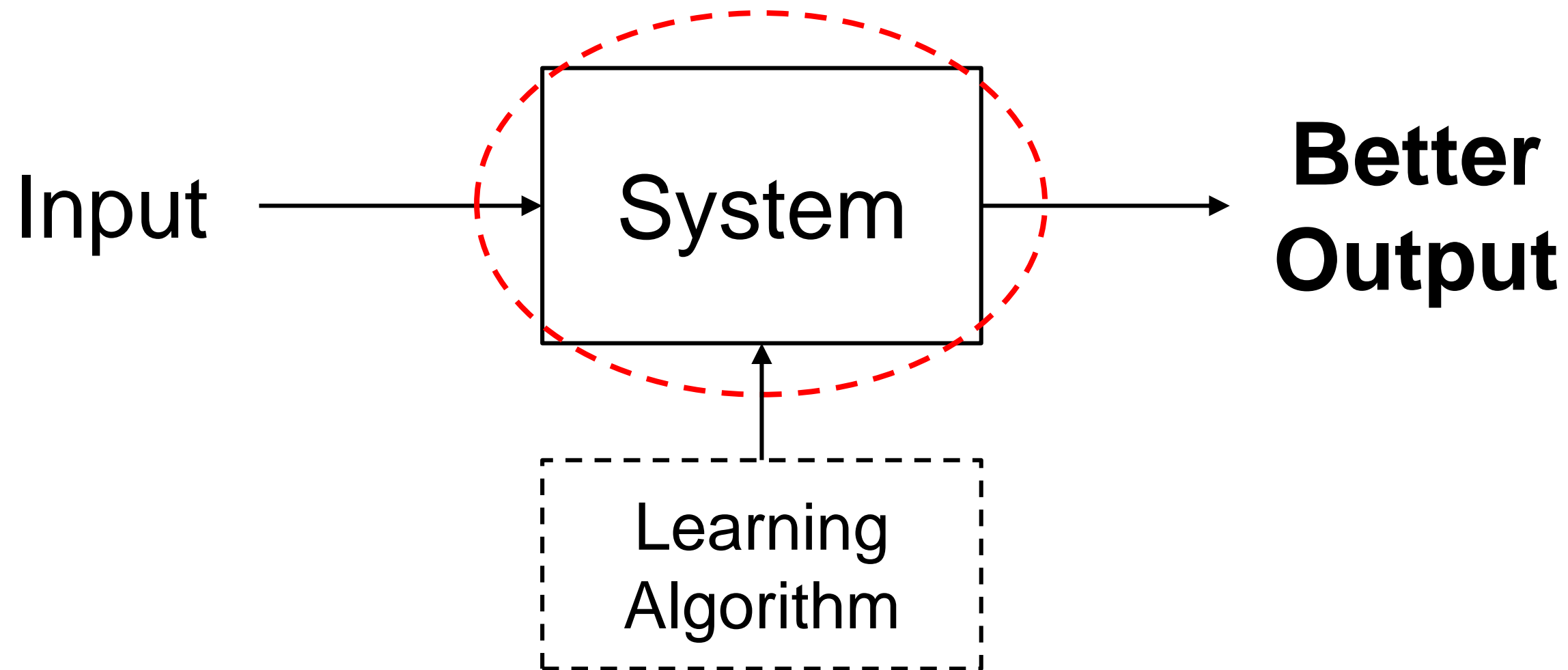
Why Data Mining?

1) We need knowledge and insights between the lines.



Why Data Mining?

2) System needs to be able to learn and produce better.

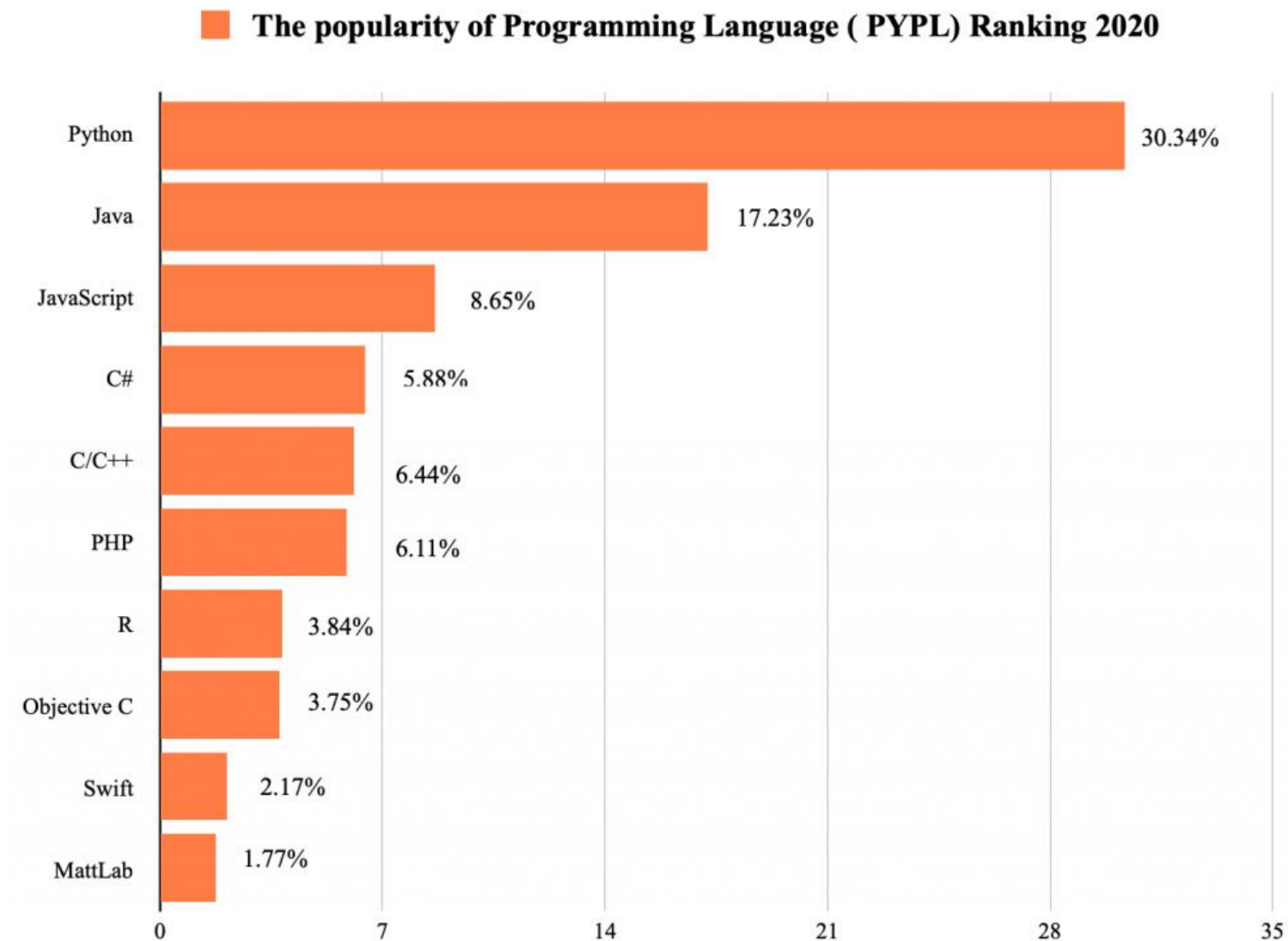


Week 1 Outline

- Orientation & Syllabus
- Course Overview
- What and Why Data Mining?
- **Why Python?**
- Conclusion

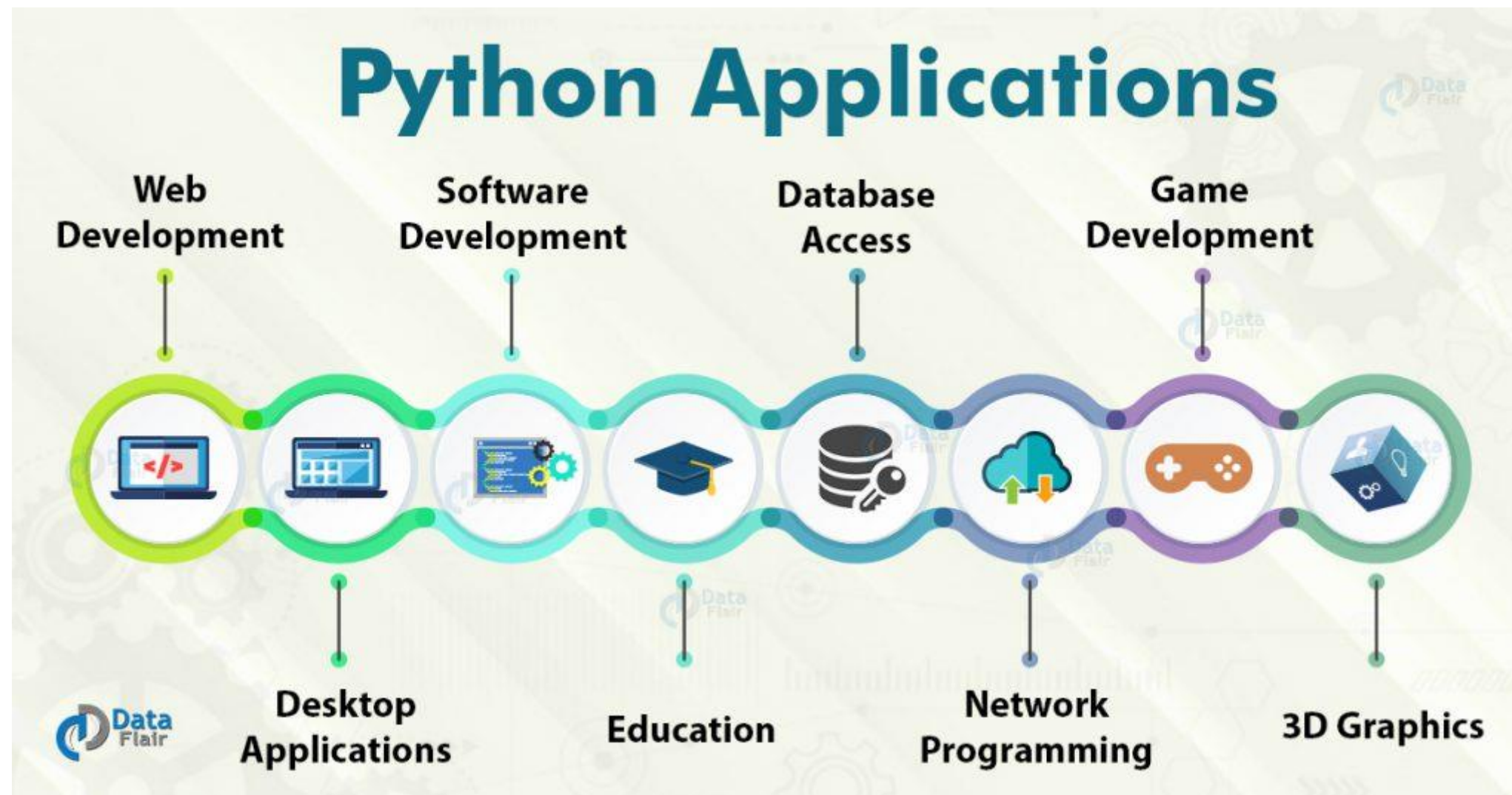
Why Python?

- 파이썬의 활용도 높음
 - 2020년 파이썬 랭킹 1위



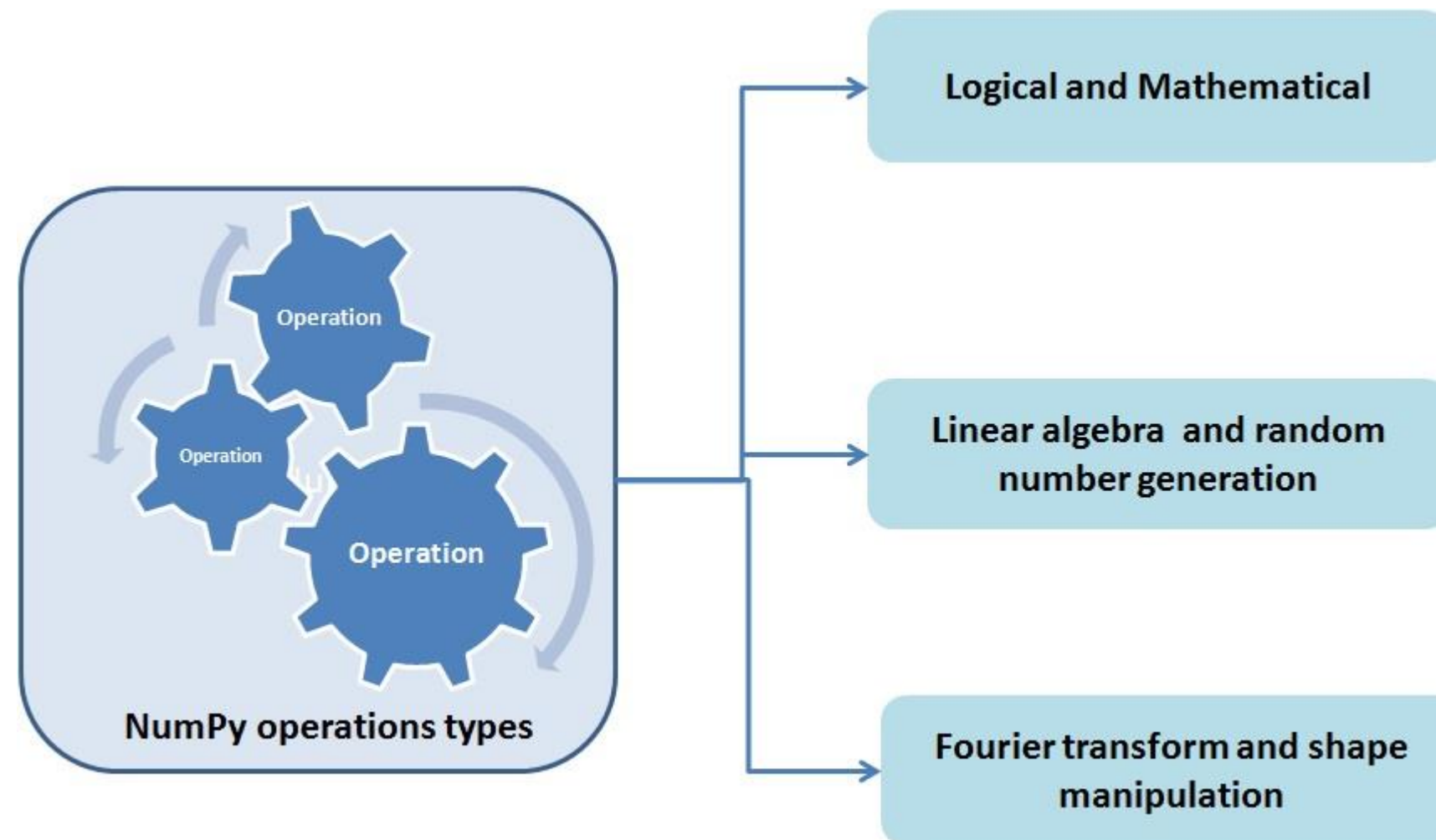
Why Python?

- 파이썬의 넓은 활용 범용성
 - 하기 어플리케이션 이외에도 다양한 어플리케이션에 적용중



Why Python?

- 파이썬의 특징들과 빅데이터와 매칭
 - 오픈소스, 라이브러리 Support, 진행속도, 범위 및 데이터 처리 Support



Why Python?

- 딥러닝에서 파이썬 위치는 독보적
 - 타 언어 대비 큰 장점으로 인해 활용됨

1. Python offers a rich library ecosystem

2. Python has a low entry barrier

3. It's amazingly flexible

4. It doesn't depend on any particular platform

5. Python is easy to read

6. It offers a variety of visualization options for data scientists

7. It's represented by a large community

8. Python's popularity keeps growing among scientists, professors, and large corporations

Limitation on Python?

- 그럼에도 파이썬의 한계점은 1) 속도 , 2) 무거움
 - 실제 서비스/제품 출시시 처리 속도 및 경량화에 문제 가능



Week 1 Outline

- Orientation & Syllabus
- Course Overview
- What and Why Data Mining?
- Why Python?
- Conclusion

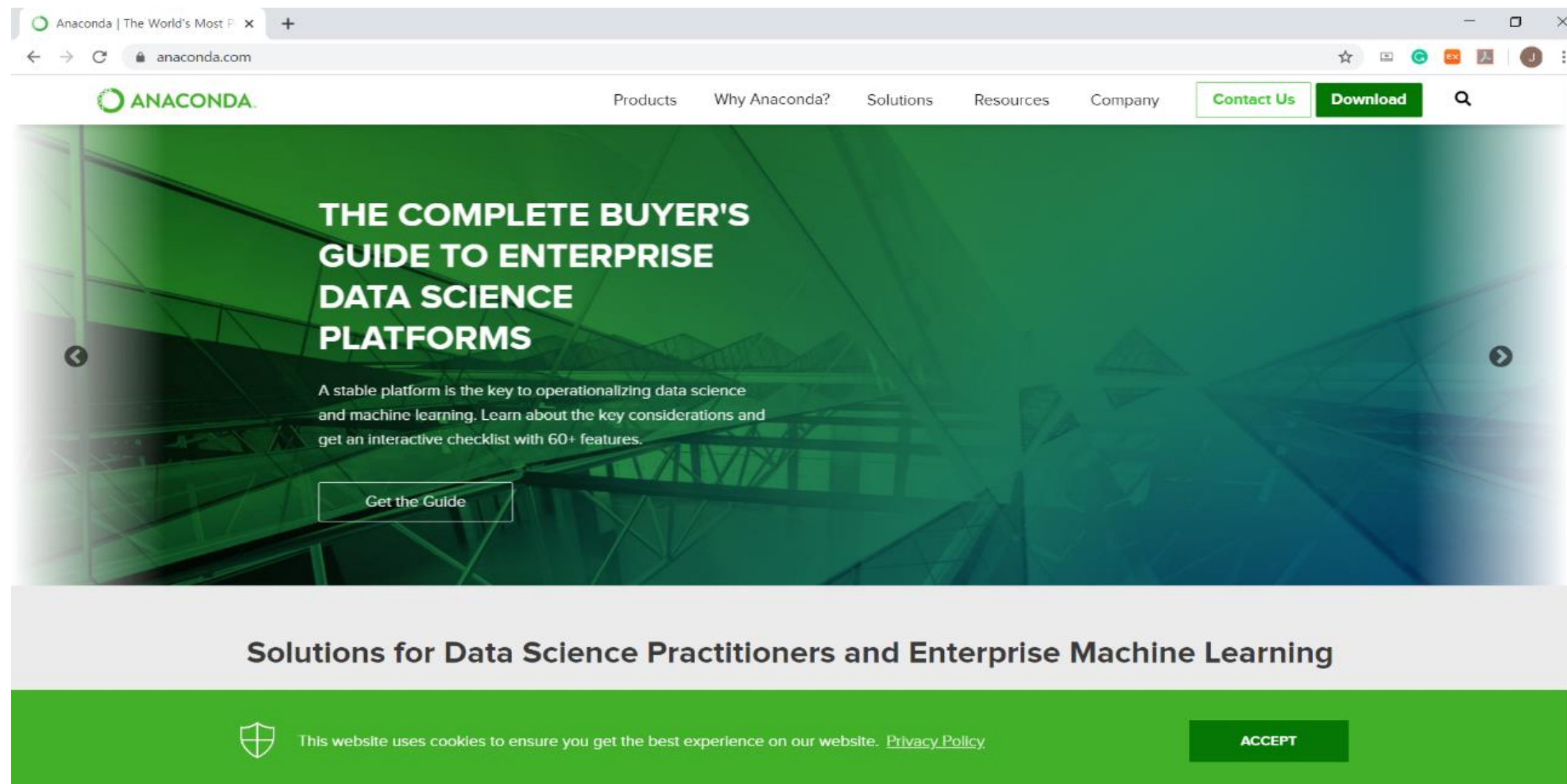
환경 세팅

전체 기본 사항

- Anaconda3 설치
- Python 버전: Python 3.9

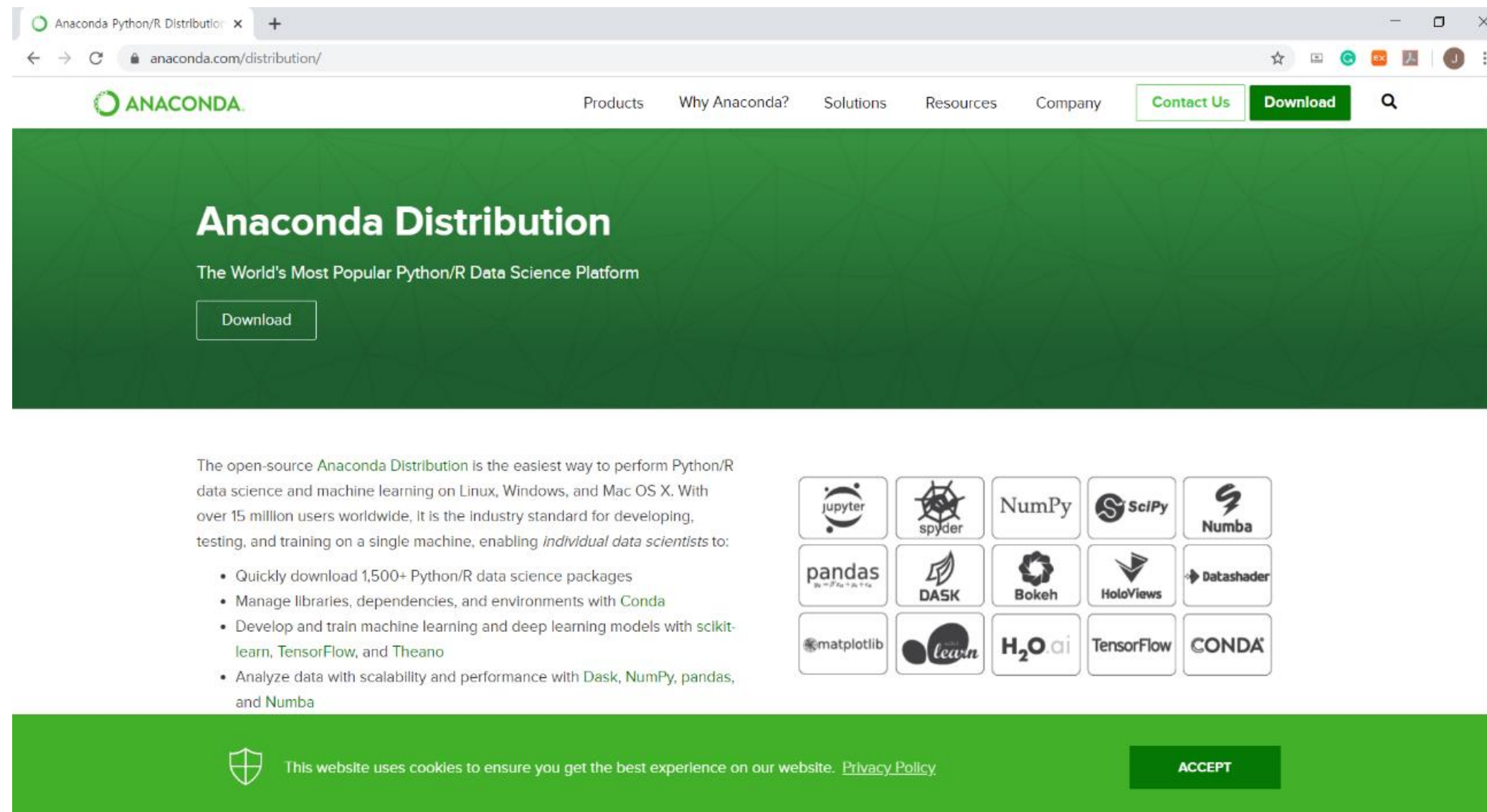
Anaconda 설치

- 주소창에 `https://www.anaconda.com`
- download 메뉴 선택



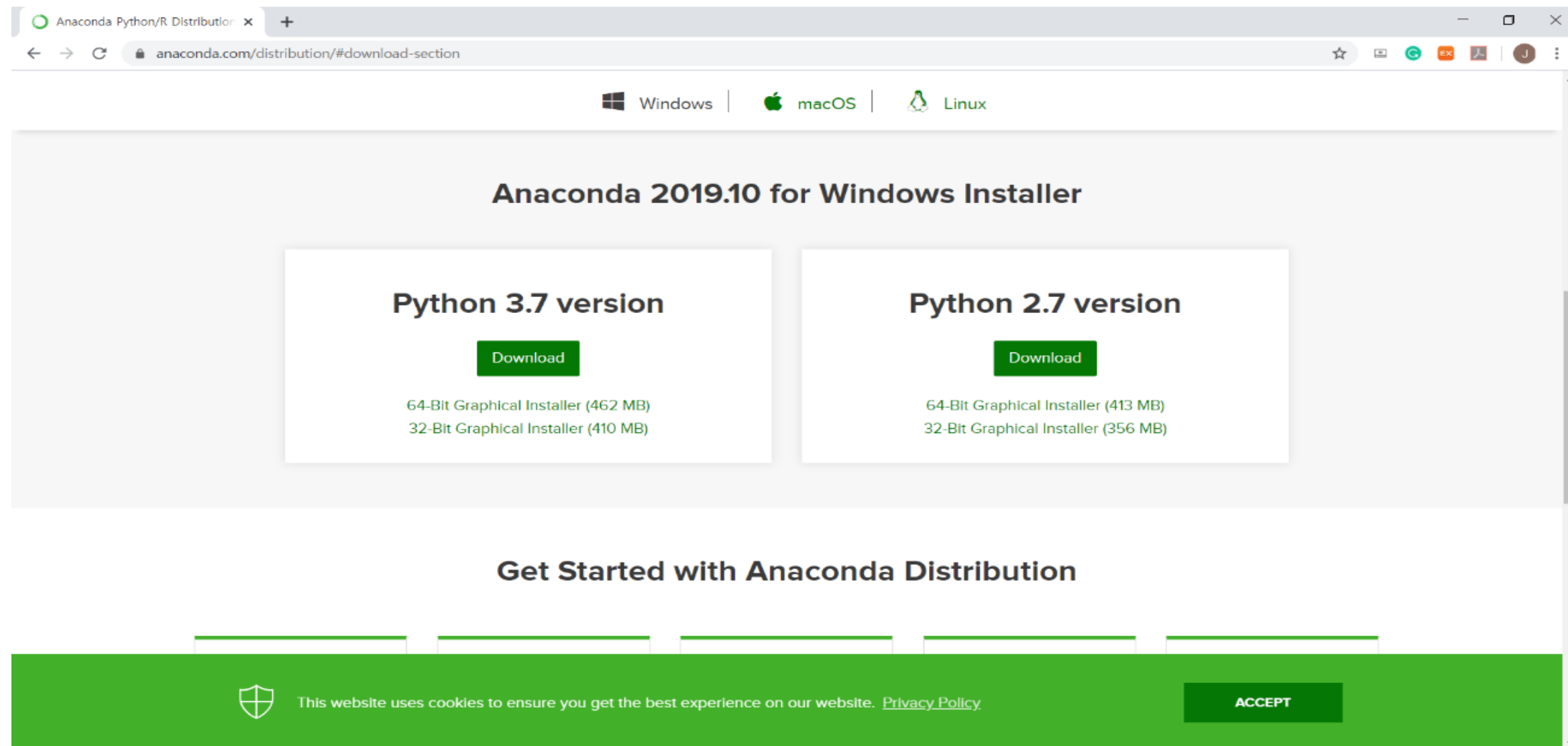
Anaconda 설치

■ 본문의 Download 버튼 선택



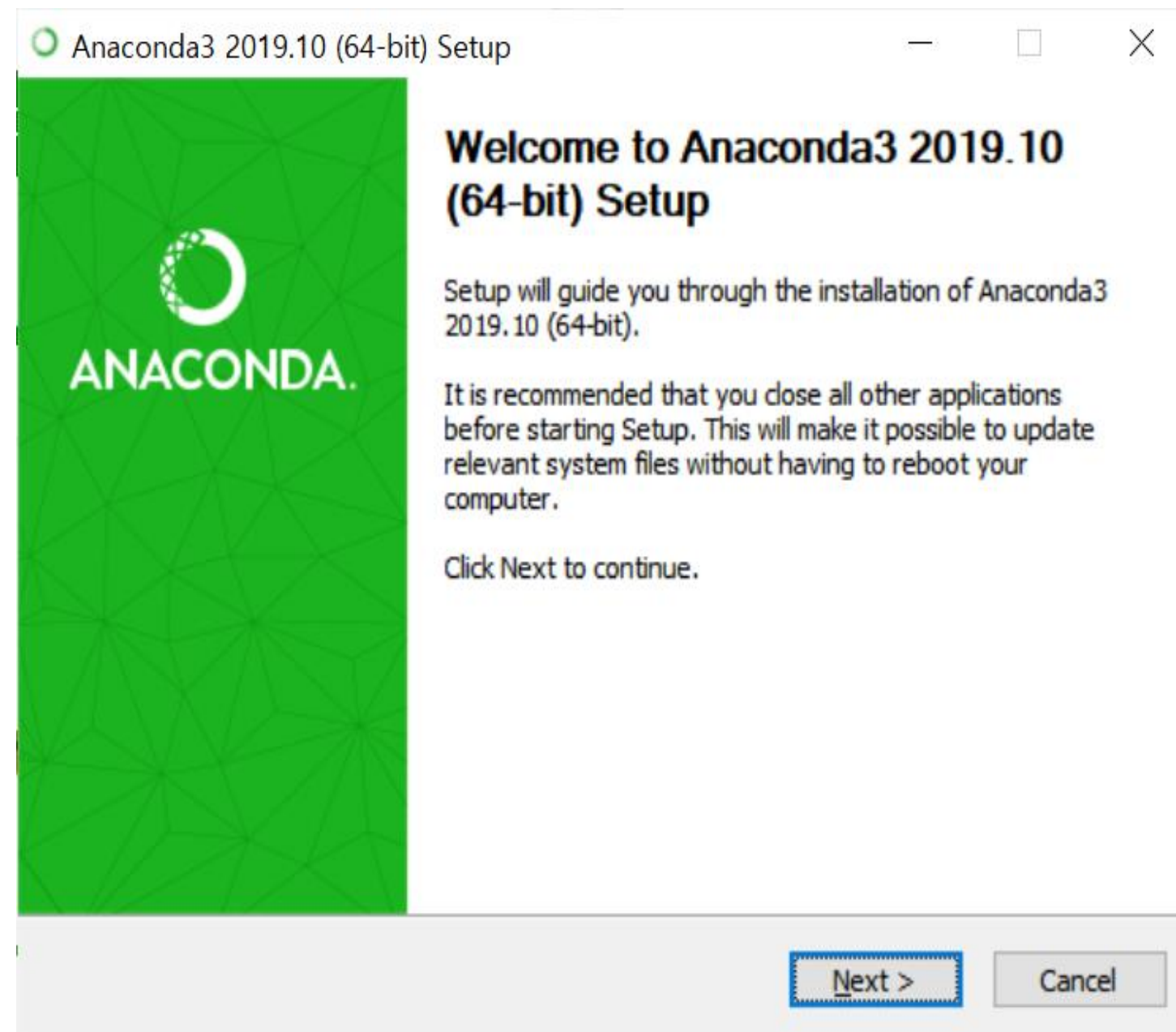
Anaconda 설치

- Python 3.7 Version을 선택하고 32비트, 64비트 중 선택



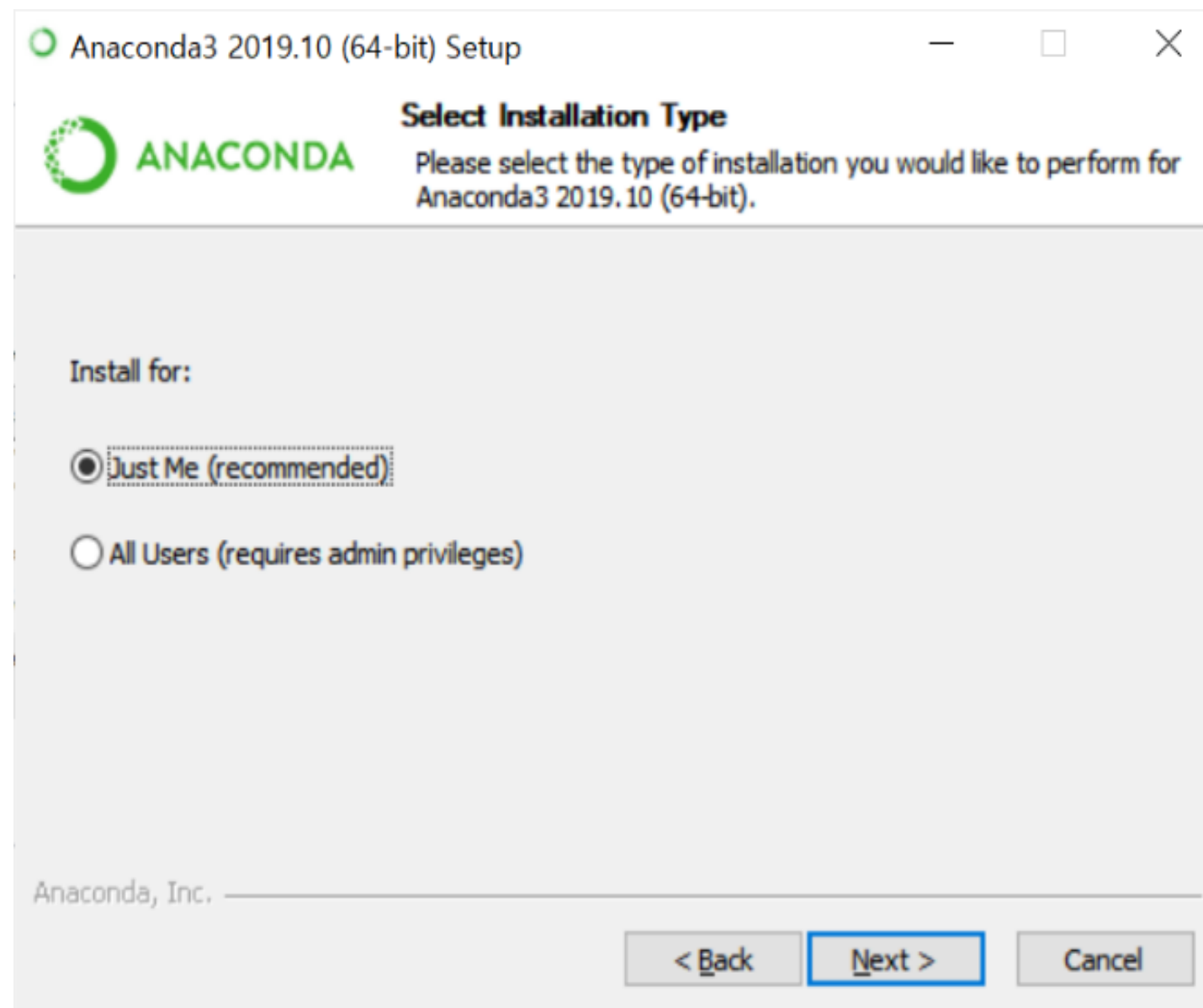
Anaconda 설치

- 다운로드 완료 후 PC좌측 하단 열어 실행한 다음 Next 버튼 선택



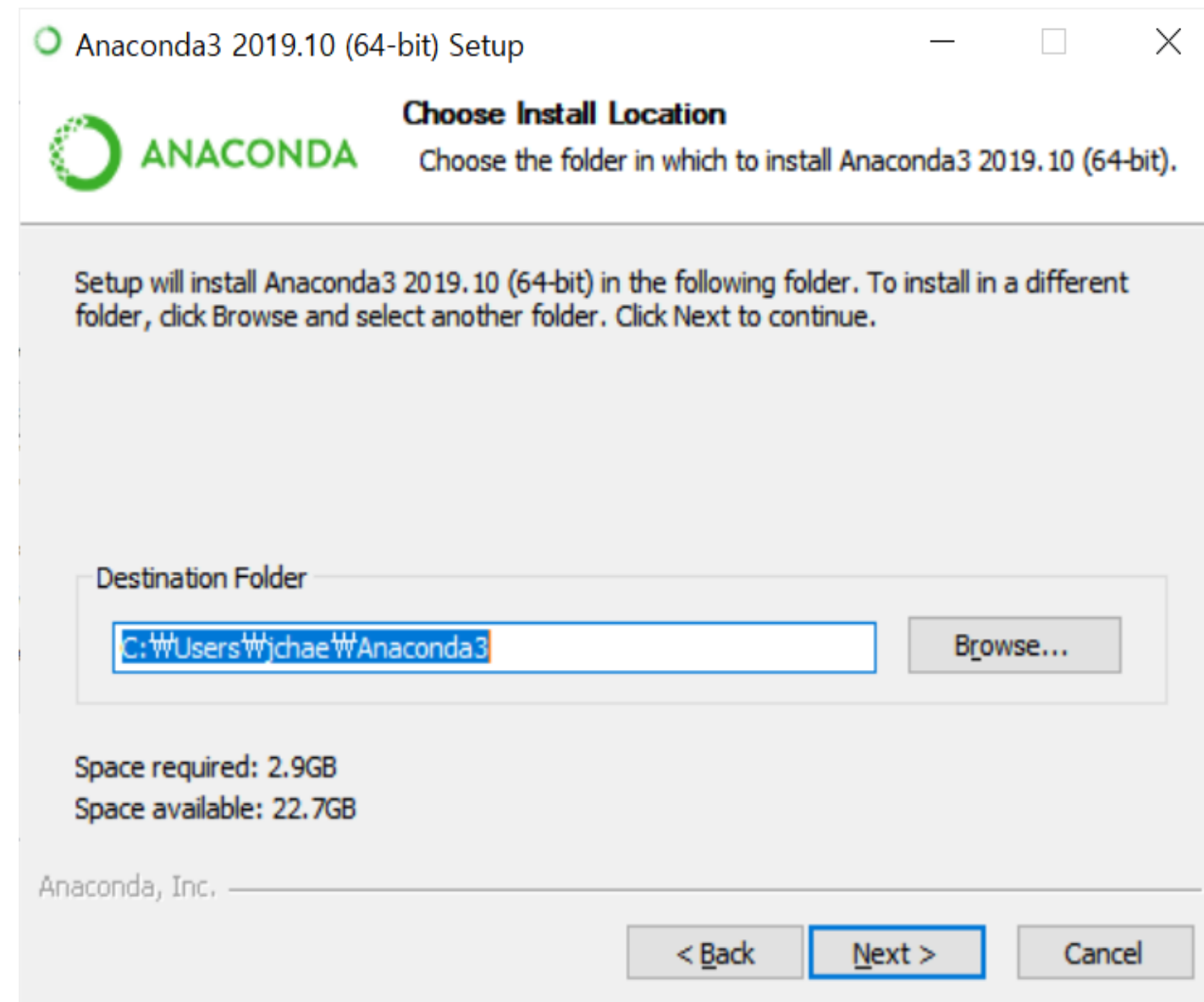
Anaconda 설치

- Select Installation Type창에서 Just Me 라디오버튼 선택



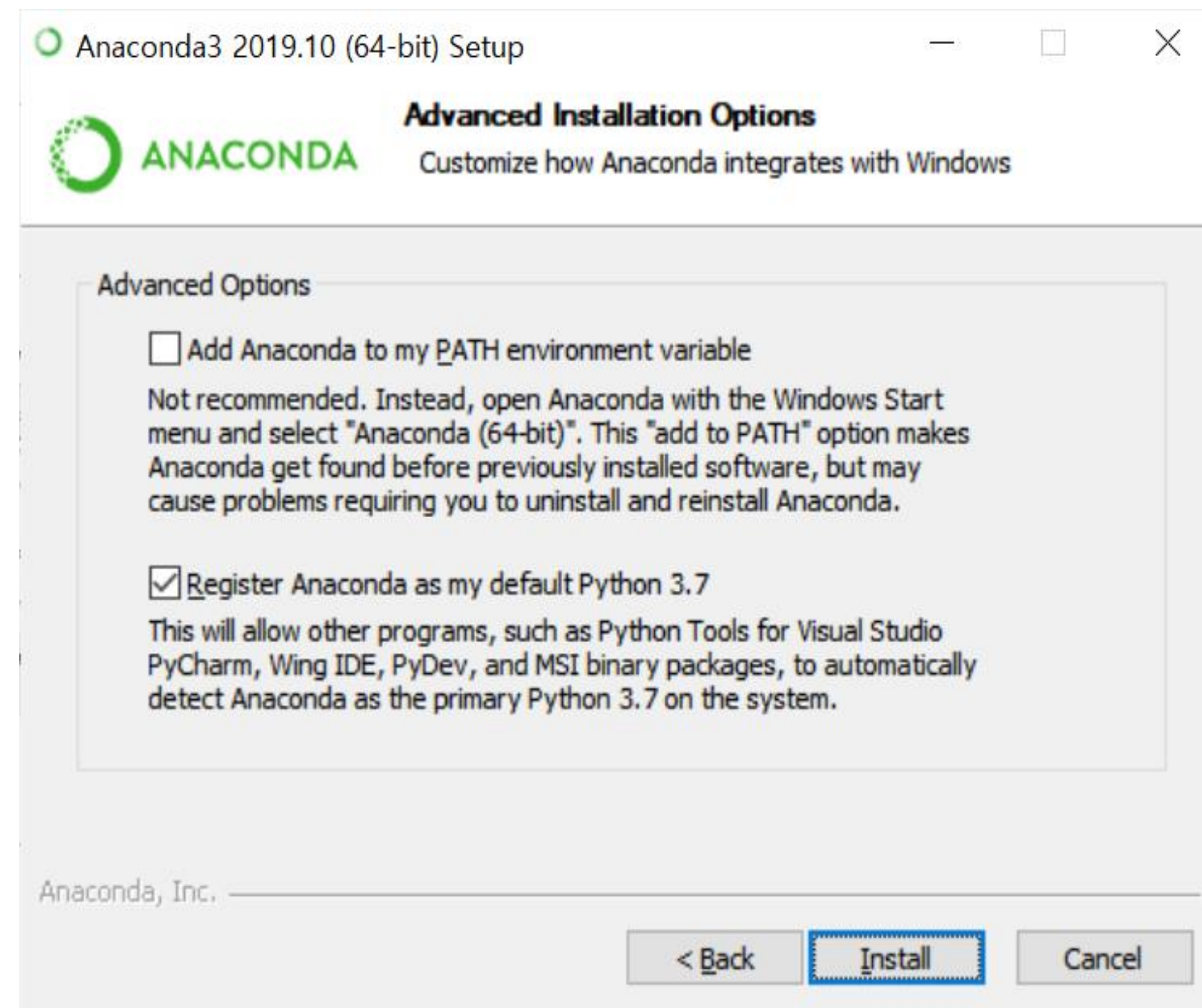
Anaconda 설치

- Choose Install Location 창에서 Next 버튼 선택



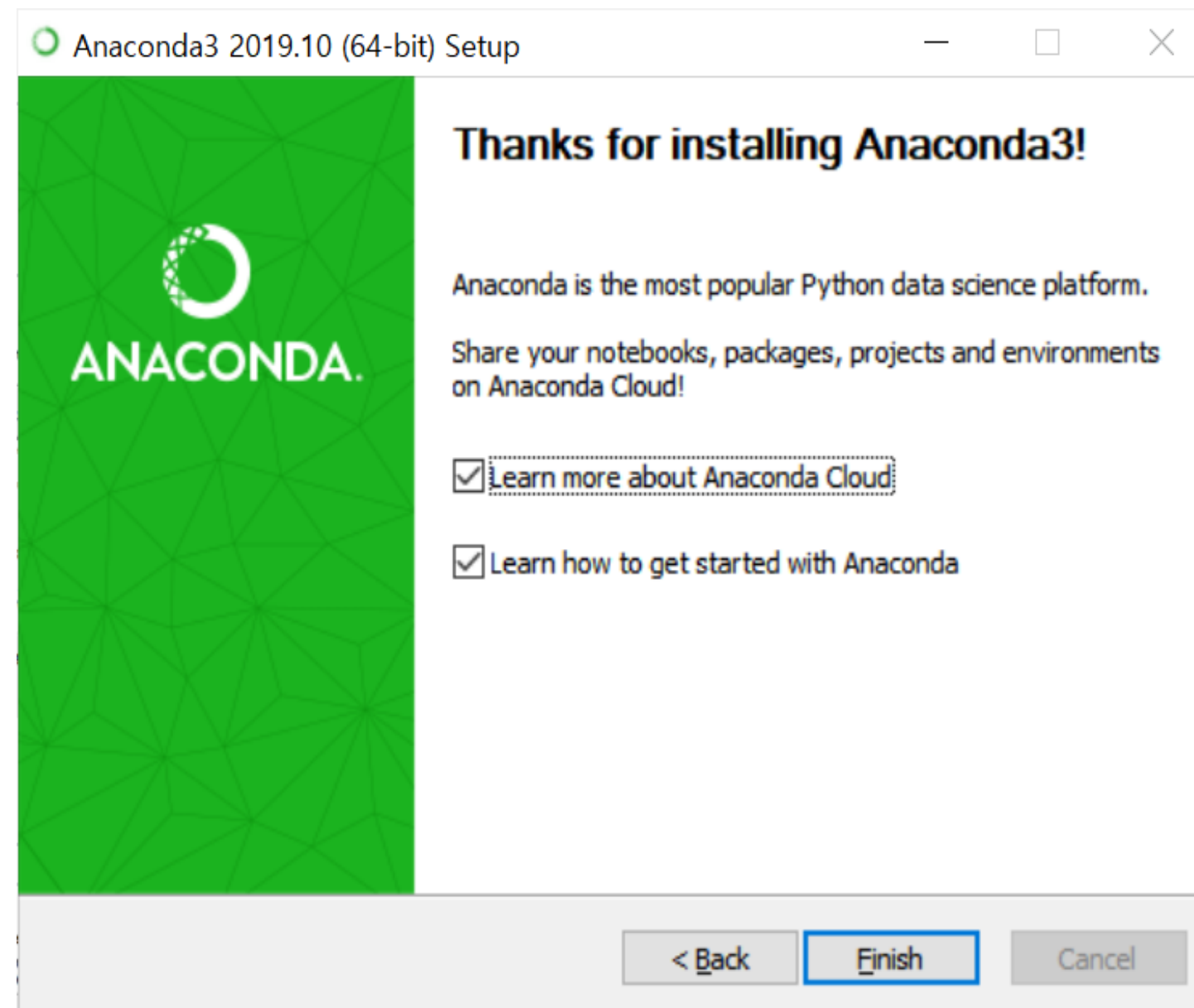
Anaconda 설치

■ Install 버튼 선택



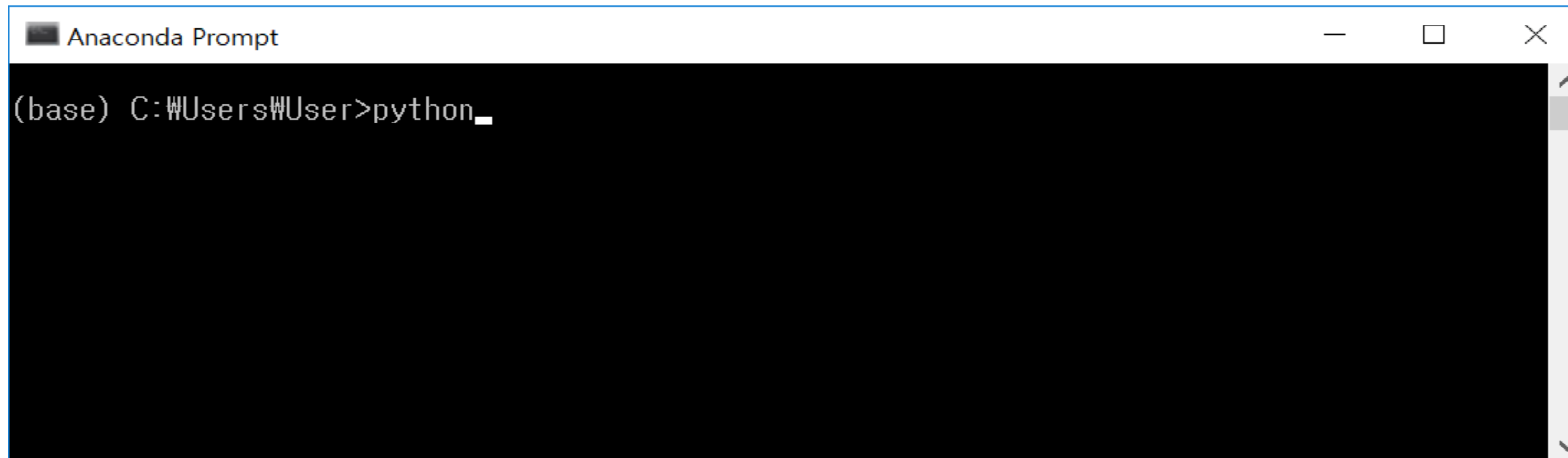
Anaconda 설치

■ 설치완료 – Finish 버튼 선택



Anaconda 설치

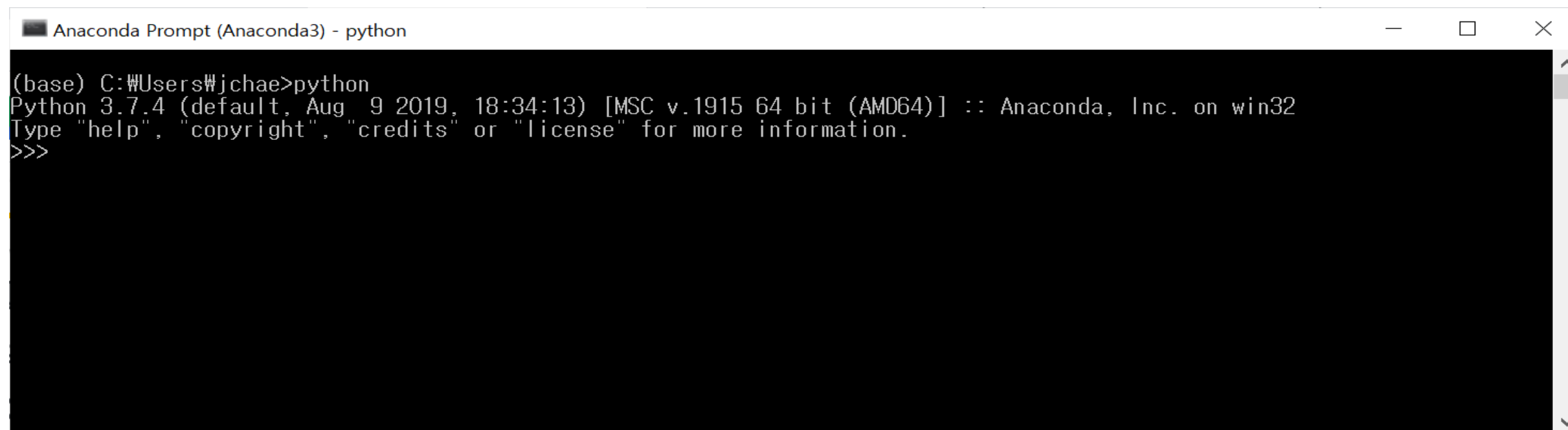
- PC에서 Anaconda폴더 확장 후 Anaconda Prompt 선택하고 입력



```
Anaconda Prompt
(base) C:\Users\User>python_
```

Anaconda 설치

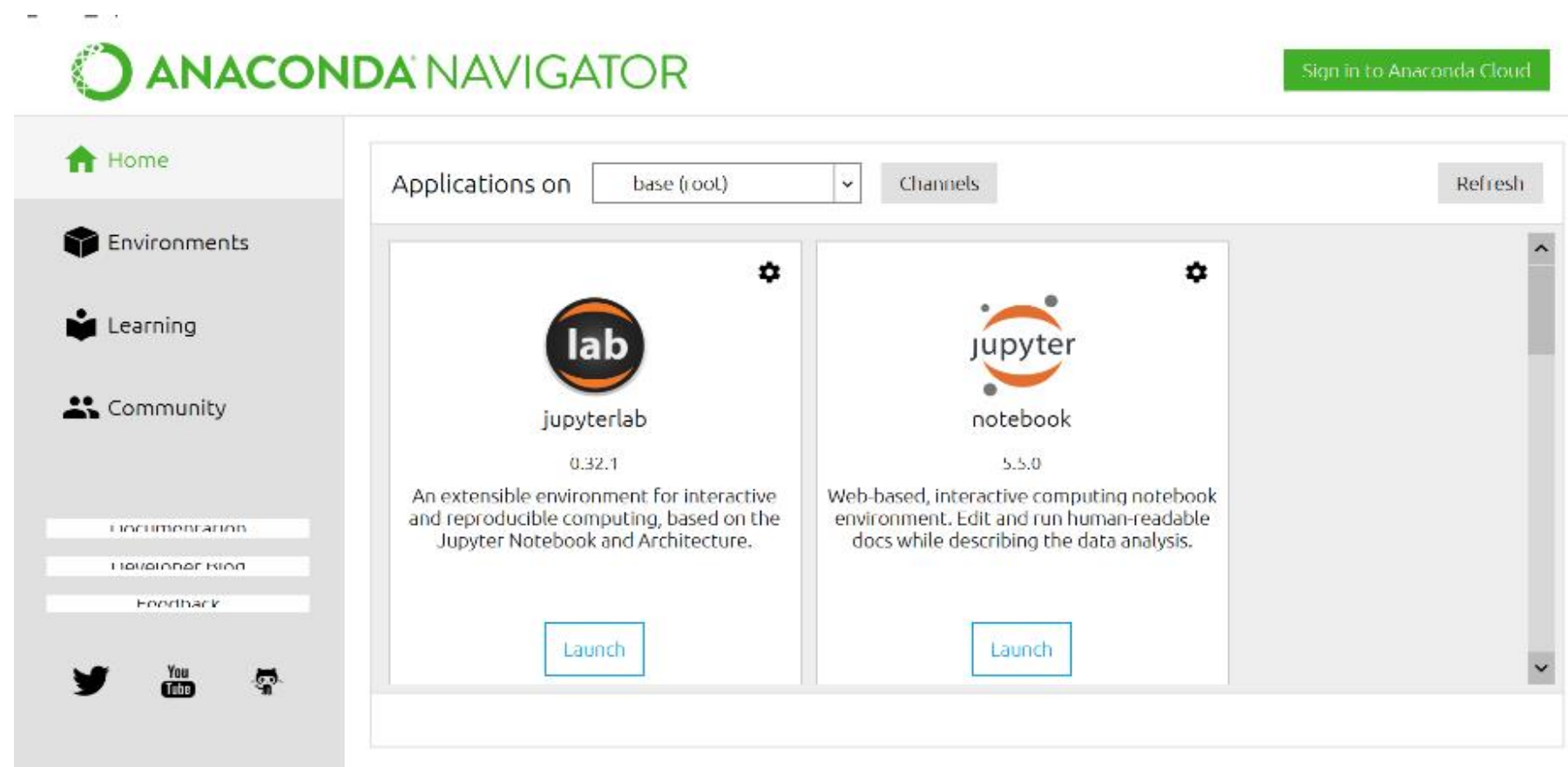
- PC에서 Anaconda폴더 확장 후 Anaconda Prompt 선택하고 입력



```
Anaconda Prompt (Anaconda3) - python
(base) C:\Users\jchae>python
Python 3.7.4 (default, Aug 9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Jupyter Notebook 사용하기

- Anaconda Prompt를 선택 후 Jupyter Notebook 입력
- PC의 시작 메뉴에서 Anaconda폴더를 확장 후 Jupyter Notebook 선택
- PC의 시작 메뉴에서 Anaconda 폴더를 확장 후 Anaconda Navigator를 선택하여 Jupyter Notebook의 Launch 버튼을 선택



Week 1 Outline

- Orientation & Syllabus
- Course Overview
- Big Data Analysis
- Why Python?
- Environment Setup
- Conclusion

Week 1 Key Takeaway

■ Orientation and Overview of Data Mining

→ Data Mining 강의에서 배울 내용 및 핵심 요소

■ Data Mining Concept & Environment Setup

→ 데이터 마이닝 컨셉 및 Anaconda3 설치/Jupyter Notebook 사용법

다음 장에서는

- 파이썬 빅데이터 프로그래밍