

Data Visualization with Python – Day 1 파이썬 기본

Dec 2019

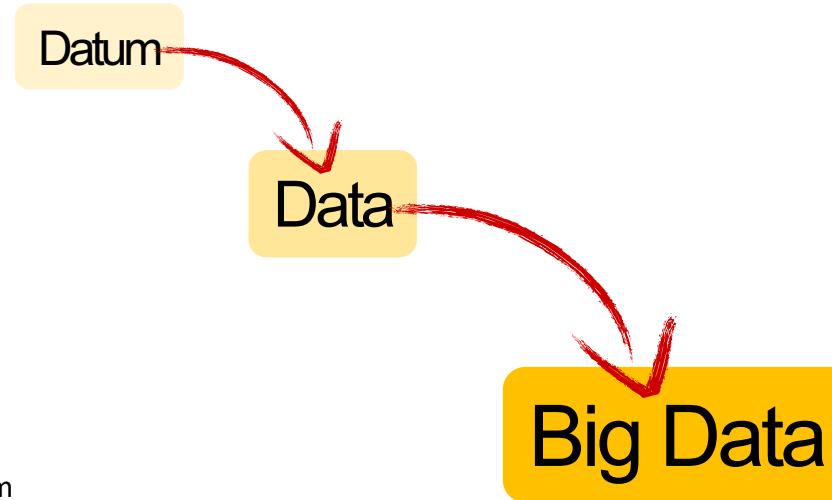
Agenda

1. 데이터란 무엇인가
2. 데이터 사이언스와 수학
3. 데이터의 수집, 가공 및 활용
4. 데이터 기반 업무의 이해
5. 데이터 분석의 이론과 실전
6. 실무 데이터 분석의 프로세스

1. 데이터란 무엇인가 | Datum, Data and Big Data

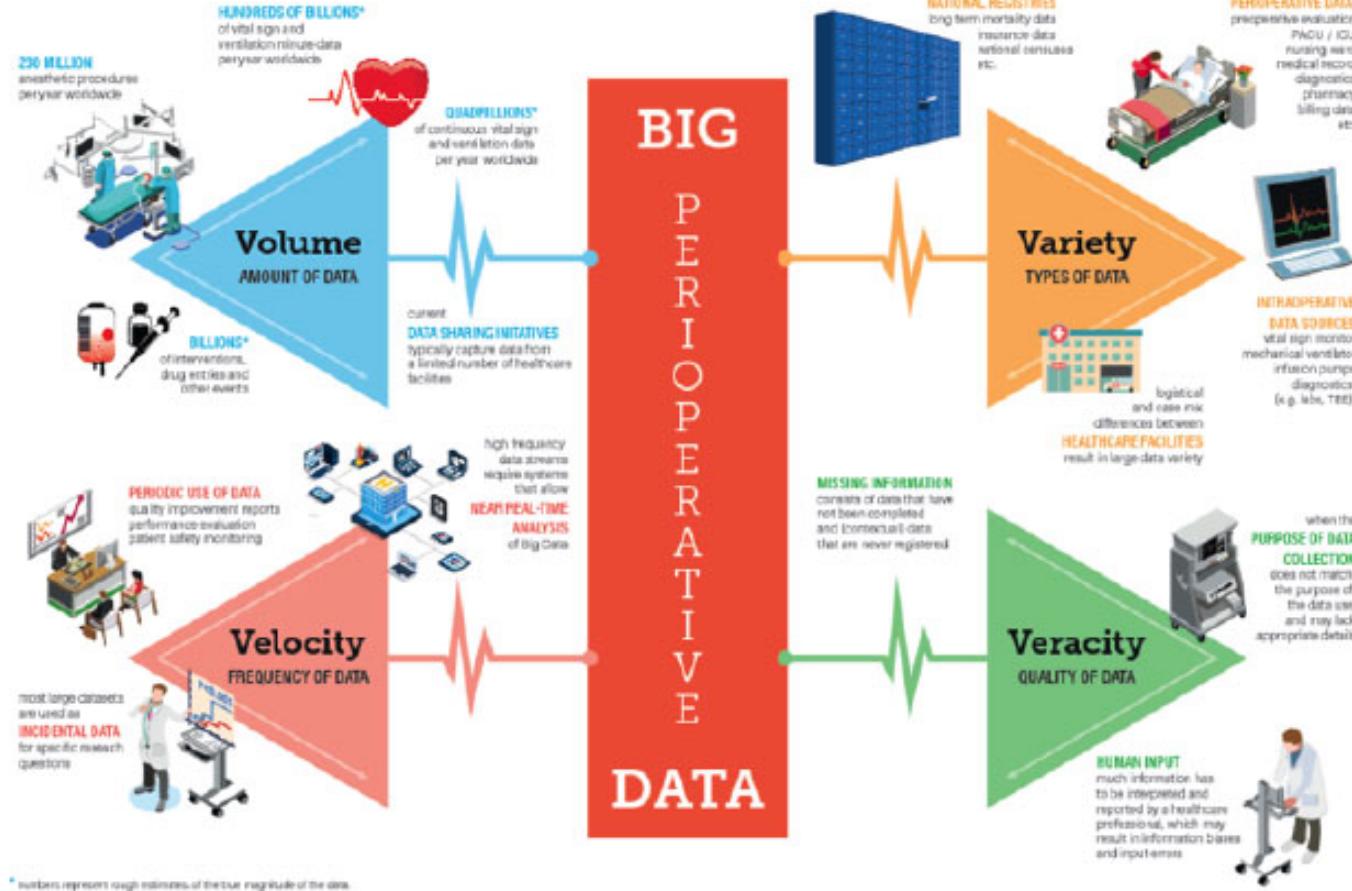
작은 데이텀, 데이터, 큰 데이터

Datum
Datum Datum
Datum Datum Datum
Datum Datum Datum Datum
Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum Datum Datum Datum Datum
Datum Datum Datum Datum Datum Datum Datum Datum Datum Datum Datum



1. 데이터란 무엇인가 | 빅데이터의 특성

빅데이터 해부하기



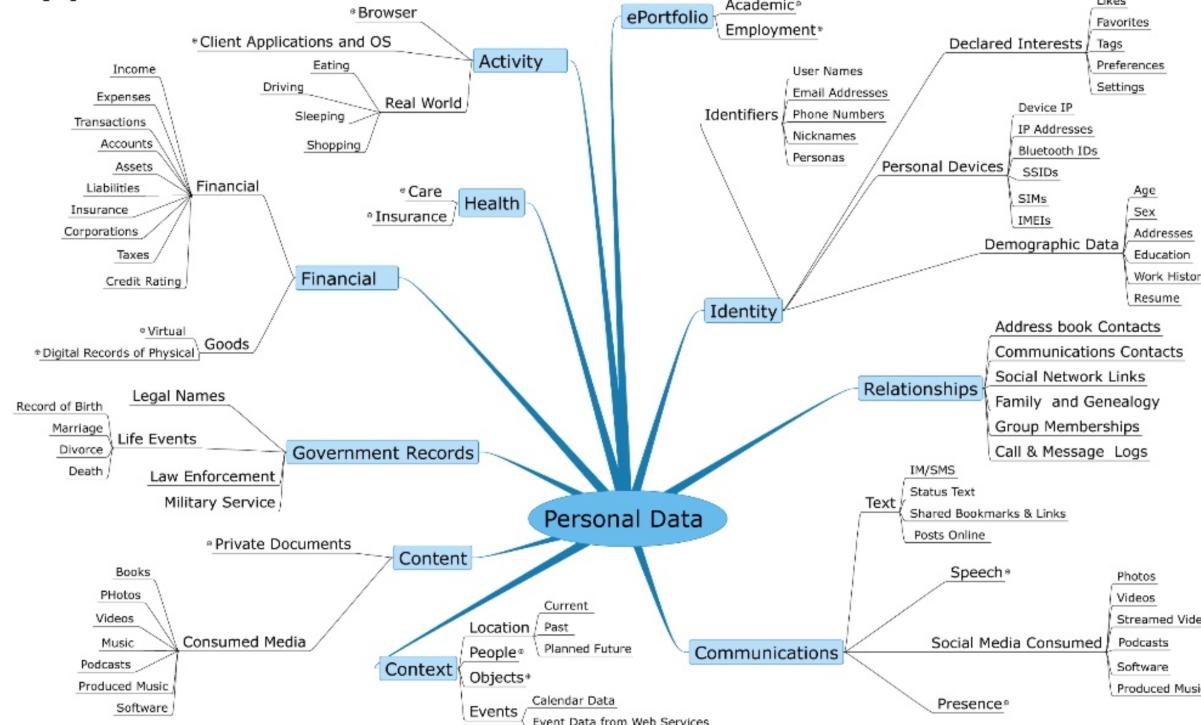
빅데이터의 특성 : 3V + 2V

- Volume - 양
- Variety - 다양성
- Velocity - 속도
- Variability - 변동성
- Veracity - 정확성

1. 데이터란 무엇인가 | 데이터의 속성

내가 그의 이름을 불러주기 전에는 그는 다만 하나의 사실에 지나지 않았다

Types of Personal Data



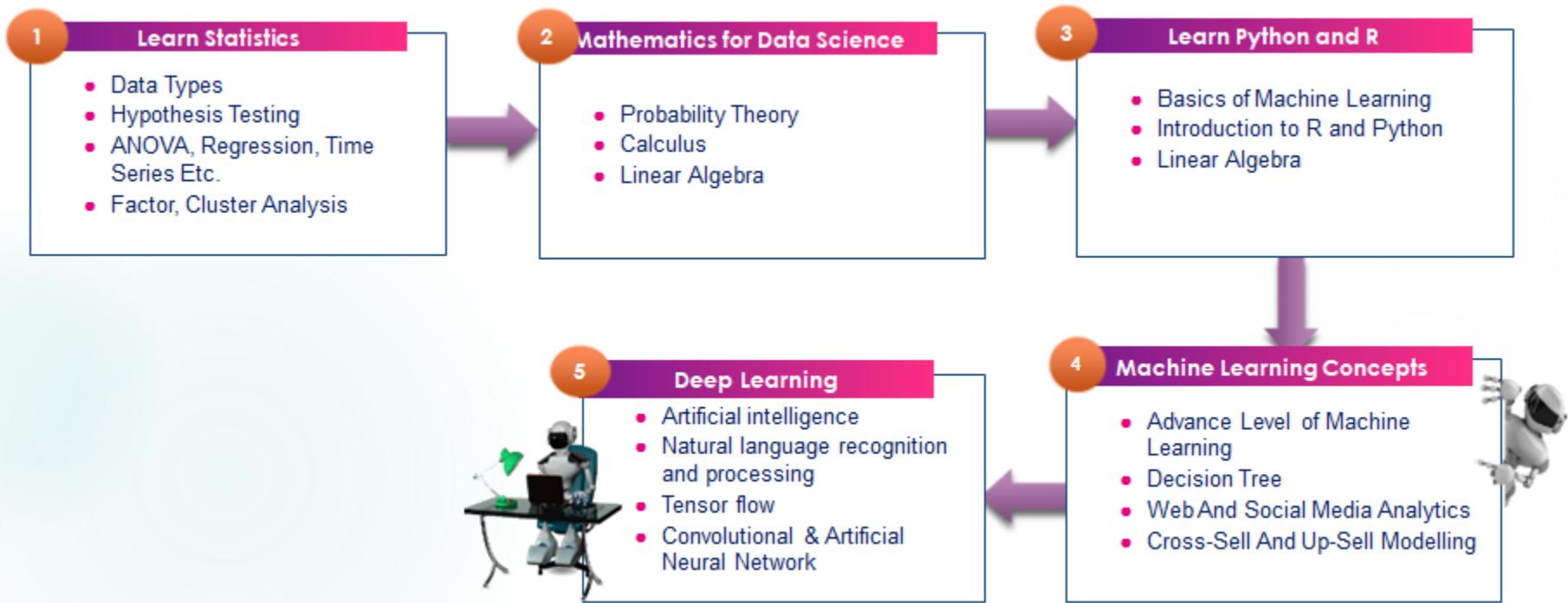
NSTIC Privacy Workshop, June 28, 2011 PERSONAL DATA ECOSYSTEM CONSORTIUM

데이터의 속성

- 가치적 측면**
 - ✓ 사실로서의 데이터
 - ✓ 정보로서의 데이터
- 형식적 측면**
 - ✓ 질적자료/비계량형자료
 - ✓ 양적자료/계량형자료
- 형태적 측면**
 - ✓ 정형 데이터
 - ✓ 비정형 데이터
- 관리(저장 및 처리)적 측면**
 - ✓ 관계형 DB(Relational DB)
 - ✓ 비관계형 DB(Non-Relational DB)

2. 데이터 사이언스와 수학 | 데이터 사이언스의 필수 요소들

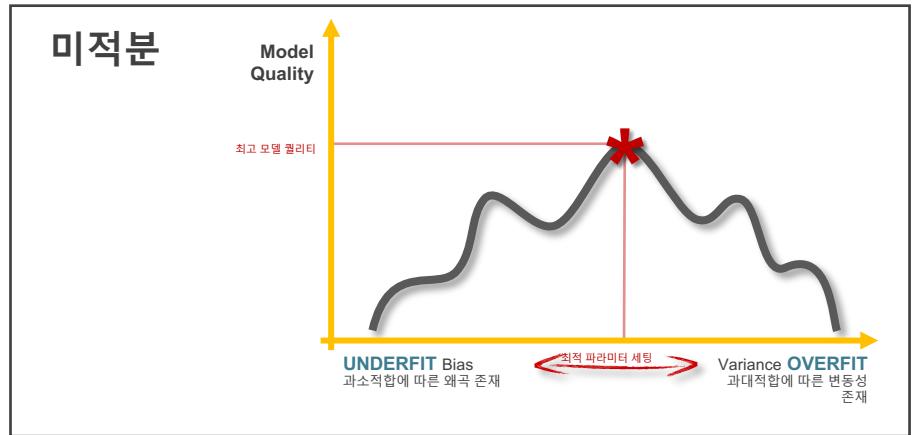
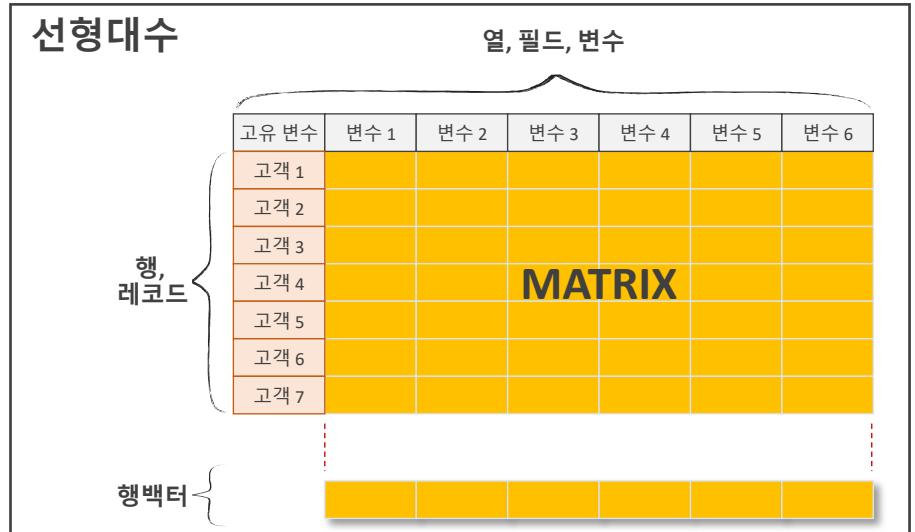
날개를 달자



출처: 5 Step Learning Path to Become Data Scientist in 2019, <https://sixsigmastats.com/5-step-learning-path-become-data-scientist-2019/>

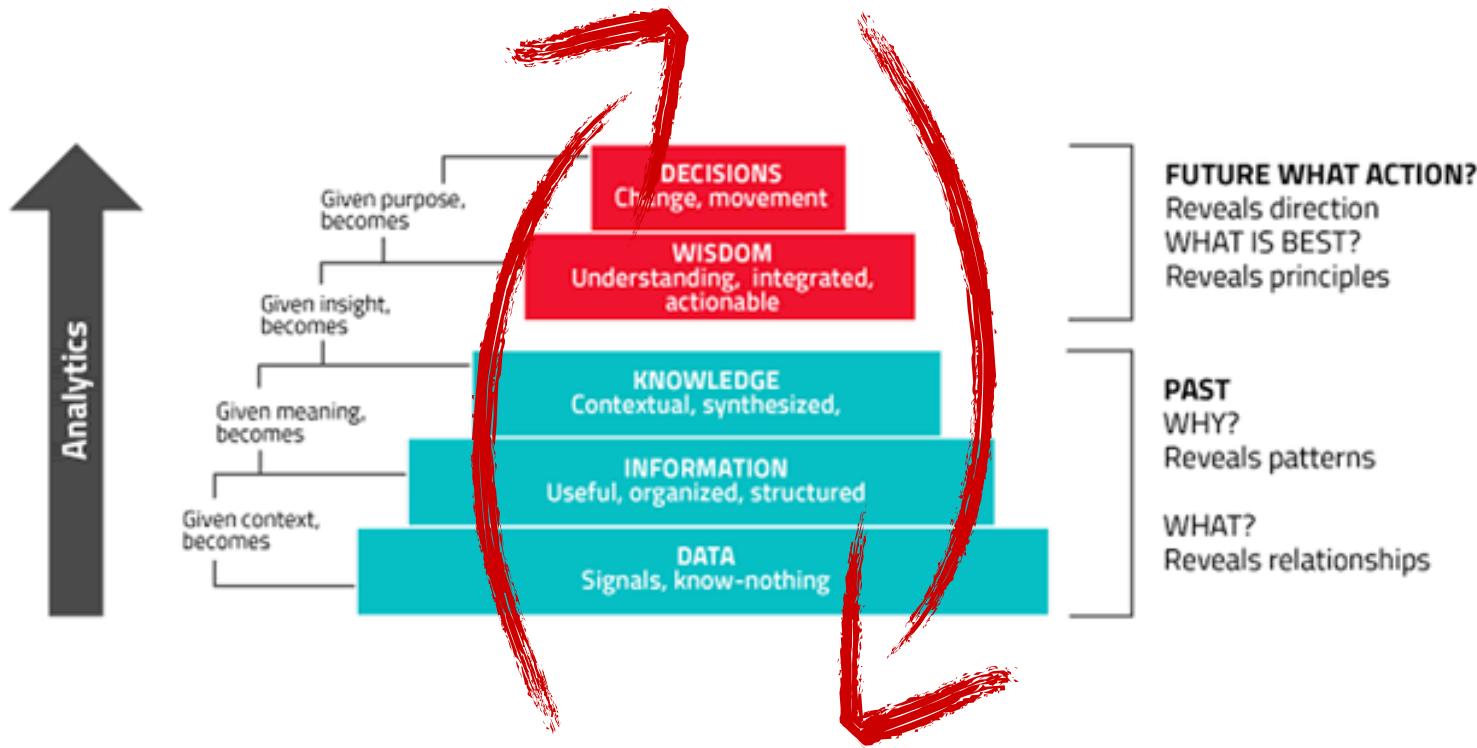
2. 데이터 사이언스와 수학 | Linear Algebra & Calculus

수학 없는 데이터 분석은, 장님이 코끼리 만지기



3. 데이터의 수집, 가공 및 활용 | Data Lifecycle

새벽부터 황혼까지 1



3. 데이터의 수집, 가공 및 활용 | Analytics Lifecycle

새벽부터 황혼까지 2

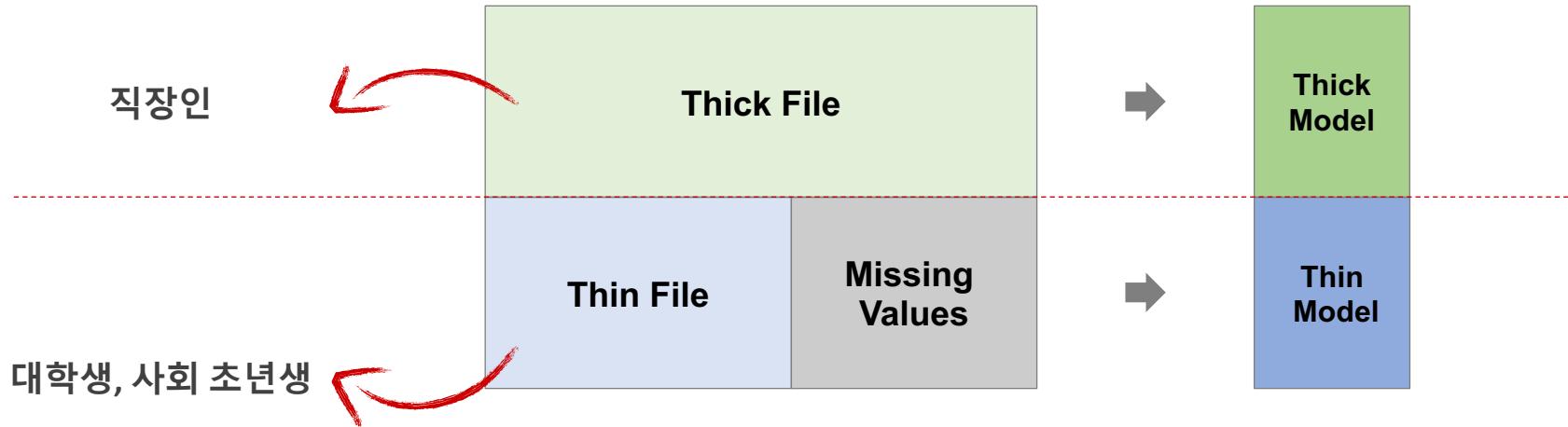


출처: 미상

3. 데이터의 수집, 가공 및 활용 | Thick File & Thin File의 이해

풍요속의 빈곤

Ex.



4. 데이터 기반 업무의 이해 | Data Engineering, Analytics, and Science 1

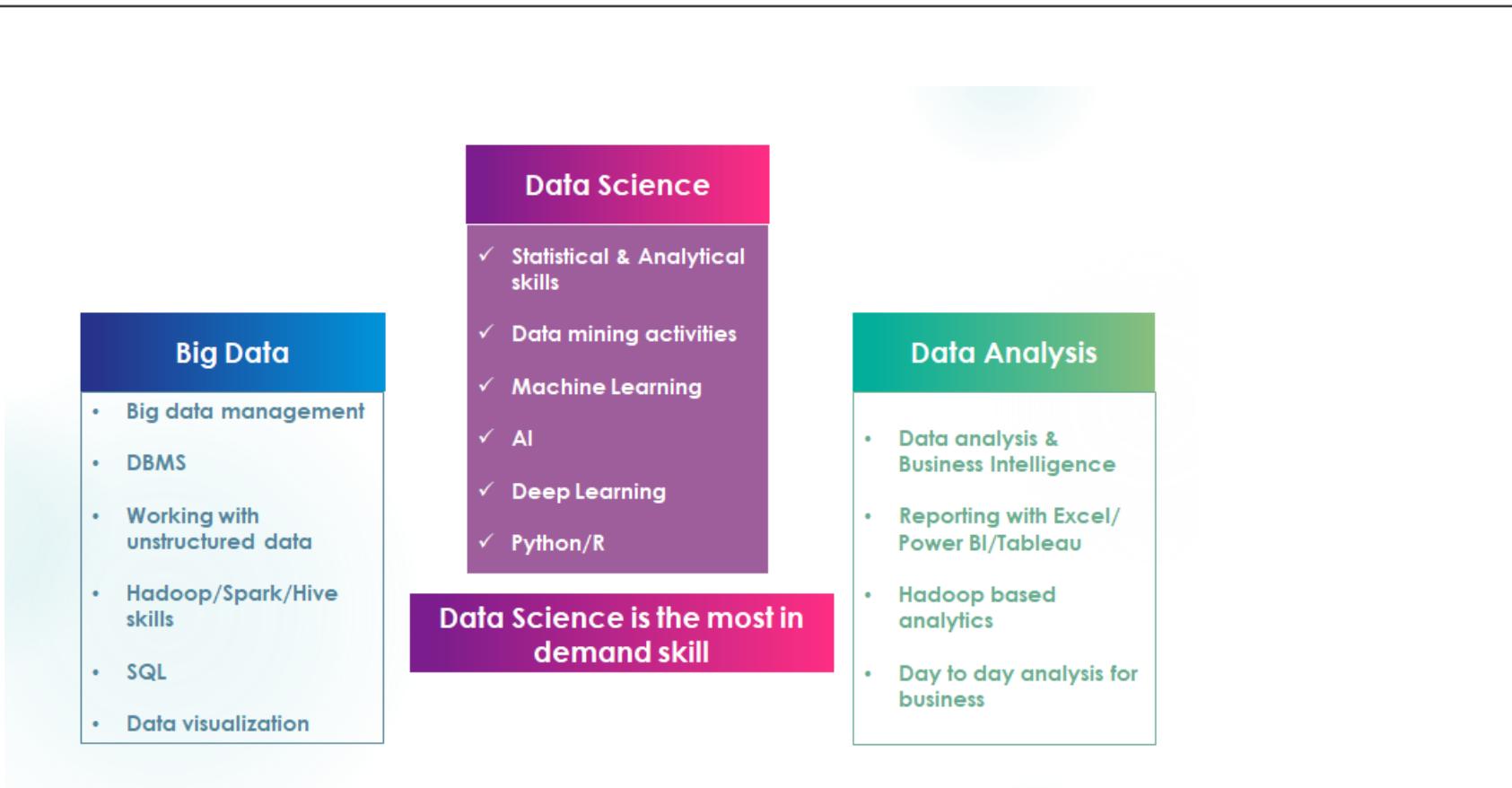
아다르고 어다르다 1

데이터 엔지니어 - 데이터 애널리스트 - 데이터 사이언티스트

Engineering Friendly Biz Friendly Science Friendly

4. 데이터 기반 업무의 이해 | Data Engineering, Analytics, and Science 2

아다르고 어다르다 2



출처: 5 Step Learning Path to Become Data Scientist in 2019, <https://sixsigmastats.com/5-step-learning-path-become-data-scientist-2019/>

4. 데이터 기반 업무의 이해 | 데이터 분석 직무

a.k.a. 핵인싸

데이터 엔지니어

데이터 애널리스트

비즈니스 애널리스트

비즈니스 인텔리전트/인사이트 매니저

데이터 비즈니스 기획자

그로스 해커

데이터 사이언티스트

AI 개발자

AI 기획자

AI 리서쳐

...

4. 데이터 기반 업무의 이해 | 데이터 분석 직무의 분류

초록은 동색

Engineering Friendly
데이터 엔지니어 - 데이터 애널리스트 - 데이터 사이언티스트
Biz Friendly

데이터 엔지니어	데이터 애널리스트	데이터 사이언티스트
AI개발자	비즈니스 애널리스트	AI 리서쳐
	비즈니스 인텔리전트 매니저	
	비즈니스 인사이트 매니저	
	데이터 비즈니스 기획자	
	그로스 해커	
	AI 기획자	

4. 데이터 기반 업무의 이해 | Data Driven

데이터 분석은 거의 모든 업무와 연관



5. 데이터 과학의 이론과 실전 | 데이터 기반의 커뮤니케이션

우리는 데이터로 이야기 한다.

Empiricism, Heuristics and Cognitive Bias



정량적 vs. 정성적

5. 데이터 과학의 이론과 실전 | 재료로서의 데이터

우리는 데이터로 이야기 한다.

일반적으로 생각하는 데이터

VS.

분석용 데이터

다양하고, 지저분하고, 부족하고, 빠져있고, 잘못된



5. 데이터 과학의 이론과 실전 | 도구로서의 데이터

우리는 데이터로 이야기 한다.



일반적으로 생각하는 데이터 분석

VS.

실무 데이터 분석

빠르고 유효하고 정확하고 효율적이며 설득력 있지만 제한적이고 어려운?

5. 데이터 과학의 이론과 실전 | 업무로서의 데이터

우리는 데이터로 이야기 한다.

일반적으로 생각하는 데이터 관련 직무

VS.

실제 데이터 관련 직무



제한적이고, 전문적이며 폐쇄적이고 특이한?

6. 실무 데이터 과학 프로세스 | 6단계 프로세스

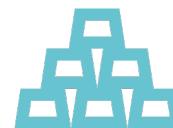
그림으로 배우는 데이터 분석 프로세스

1



이슈 도출 > 가설 설정 > 실험 설계

2



데이터 수집 > 데이터 가공

3



EDA > 데이터 요약, 집계

4



모델링 > 사후 분석

5



시각화 > 리포트 작성

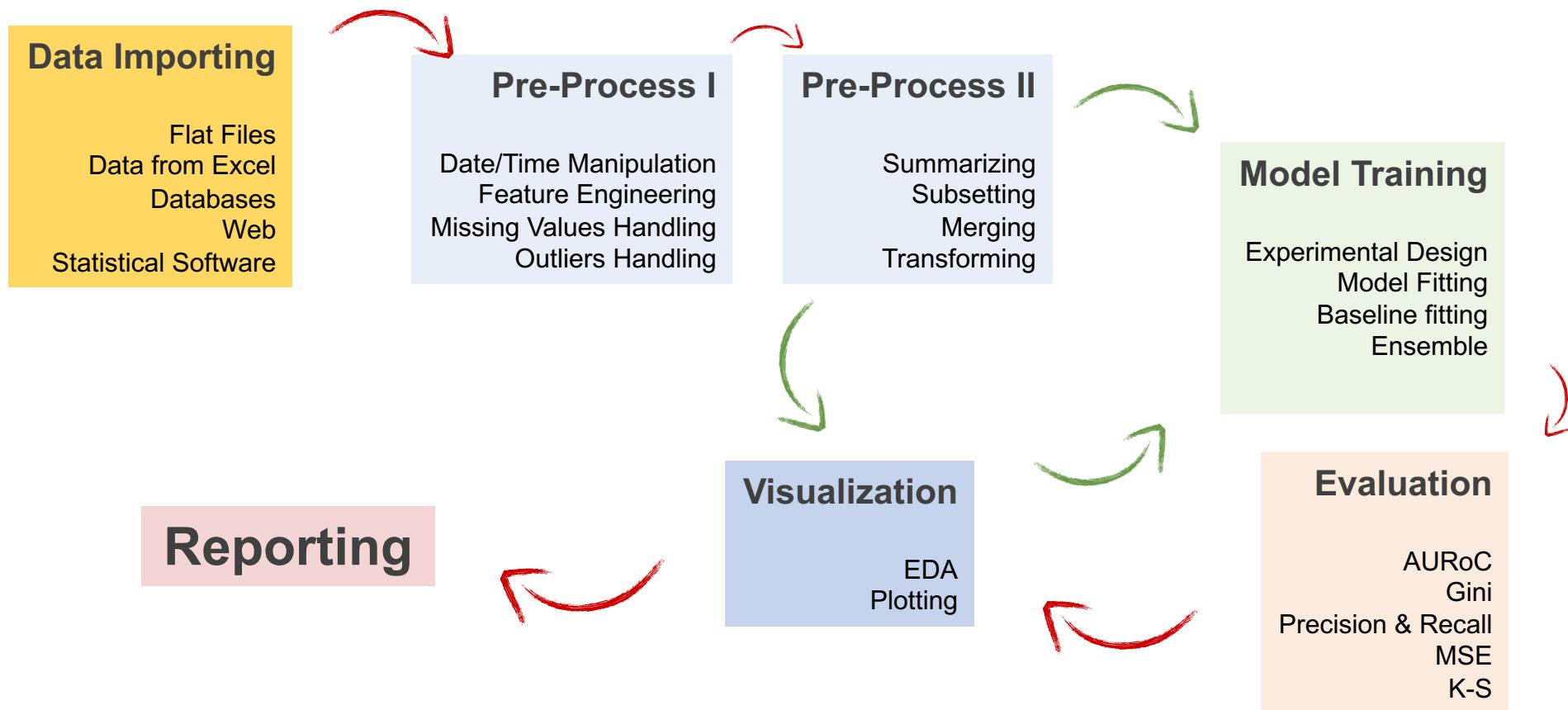
6



회고 및 공유

6. 실무 데이터 과학 프로세스 | 프로세스 상세

수집부터 리포팅까지



6. 실무 데이터 과학 프로세스 | 협업 형태에 따른 분류

어떻게 일을 하느냐에 따라 프로세스도 달라진다



독립형

- ✓ 상이한 업무 담당자로 팀이 구성된 경우 혹은 소규모 스타트업에 해당
- ✓ 주로 서비스 기준 조직
- ✓ **수평적** 업무 커뮤니케이션이 특징이며, PM은 직책보다는 직무로서의 역할에 초점
- ✓ 각 담당자가 one person one team 개념으로 각 업무를 전담
- ✓ 독립적으로 업무를 처리
- ✓ 재량권이 많으나 동시에 책임 역시 증가



관계형

- ✓ 동일하거나 유사한 업무 담당자로 팀이 구성된 경우
- ✓ 주로 기능 기준 조직
- ✓ **수직적** 업무 커뮤니케이션이 특징이며, Lead 혹은 Chief가 팀을 대표
- ✓ 각 구성원이 유사 업무를 분할하여 수행
- ✓ 팀의 시니어는 곧 팀원들 업무의 시니어와 일치, 업무 방향성 제시나 기타 조언 제공 가능

6. 실무 데이터 과학 프로세스 | 데이터 사이언티스트의 협업

어떻게 일을 하느냐에 따라 프로세스도 달라진다

- ✓ 협업이 필요한 이유:

협업이 필요한 것은 모든 직군, 직무에 다 해당하지만 데이터 사이언티스트는 데이터 자체를 위한 업무보다 데이터를 활용하여 의사결정을 돋거나, 결정을 고도화 하는 등 '의사 결정의 기준'으로 작용하는 경우가 많으므로, 타 부서 혹은 직군과 협업이 빈번

- ✓ R&R(Role and Responsibility)과 DRI(Directly Responsible Individual):

- ✓ Waterfall과 Agile:

6. 실무 데이터 과학 프로세스 | 분석 준비하기

시작 전 마지막 작업

업무 기본 설정

- ✓ 업무 목표 및 범위 산정
- ✓ 업무 필요요소 파악
- ✓ 유관부서 혹은 협력요소 파악
- ✓ 산출물 수준 결정
- ✓ MD 산정 및 일정 파악
- ✓ 업무 우선순위 산정

Agenda

1. Echo Systems of Python
2. Python Basic(by J. Notebook)

1. Echo Systems of Python | Introduction

R이냐 Python



- **R은** 1992년 오클랜드대학 통계학과에서 Ross Ihaka와 Robert Gentleman에 의해 개발
- 벨 연구소의 S가 R의 전신
- S는 벨 연구소의 사내 통계분석환경 구축을 위해 1976년 개발, 기원은 포트란 라이브러리



- **Python은** 1991년 네덜란드의 귀도 반 로섬에 의해 개발된 오픈소스 언어

1. Echo Systems of Python | Key Features of R

R?



- 특징 1: 명령어를 만날 때마다 즉시 실행하는 인터프리터 언어
- 특징 2: 프로그래밍 언어들 중에서 코드가 굉장히 직관적
- 특징 3: 프로그래밍 언어들과 비교하여 성능은 떨어지나 높은 생산성 보유
고성능의 계산이 필요한 부분은 C로 만들어서 작성하여 성능과 생산성을 모두 만족
- 특징 4: 오픈소스기반, 사용자가 패키지 개발 및 추가를 통해 기능 확장 가능
- 특징 5: 분석 환경의 물리 메모리 사이즈에 종속적
- 특징 6: 사용자 커뮤니티의 집단지성이 힘의 원천
- 특징 7: 뛰어난 데이터 시각화
- 특징 8: 간편한 환경 설정과 관리
- 활용분야 : 데이터 사이언스, AI, BA

1. Echo Systems of Python | Key Features of Python

Python?



- 특징 1: 명령어를 만날 때마다 즉시 실행하는 인터프리터 언어
- 특징 2: 프로그래밍 언어들 중에서 코드가 굉장히 직관적
- 특징 3: 프로그래밍 언어들과 비교하여 성능은 떨어지나 높은 생산성 보유
고성능의 계산이 필요한 부분은 C로 만들어서 작성하여 성능과 생산성을 모두 만족
- 특징 4: 오픈소스기반, 사용자가 패키지 개발 및 추가를 통해 기능 확장 가능
- 특징 5: 분석 환경의 물리 메모리 사이즈에 종속적
- 특징 6: 사용자 커뮤니티의 집단지성이 힘의 원천
- 특징 7: 애초에 데이터 사이언스를 위해 개발되지 않아 상대적으로 R에 비해 다양한 개발 가능
- 특징 8: 다소 난해한 환경 설정과 관리
- 활용분야 : 데이터 사이언스, AI, BA, 웹 프로그래밍...

1. Echo Systems of Python | Comparison

R 혹은 P

Difference between R and Python

Parameter	R 	Python 
Objective	Data analysis and statistics	Deployment and production
Primary Users	Scholar and R&D	Programmers and developers
Flexibility	Easy to use available library	Easy to construct new models from scratch. I.e., matrix computation and optimization
Learning curve	Difficult at the beginning	Linear and smooth
Popularity of Programming Language. Percentage change	4.23% in 2018	21.69% in 2018
Average Salary	\$99.000	\$100.000
Integration	Run locally	Well-integrated with app
Task	Easy to get primary results	Good to deploy algorithm
Database size	Handle huge size	Handle huge size
IDE	Rstudio	Spyder, IPython Notebook
Important Packages and library	tidyverse, ggplot2, caret, zoo	pandas, scipy, scikit-learn, TensorFlow, caret
Disadvantages	Slow High Learning curve Dependencies between library	Not as many libraries as R
Advantages	<ul style="list-style-type: none">•Graphs are made to talk. R makes it beautiful•Large catalog for data analysis•GitHub interface•RMarkdown•Shiny	<ul style="list-style-type: none">•Jupyter notebook: Notebooks help to share data with colleagues•Mathematical computation•Deployment•Code Readability•Speed•Function in Python

출처: R Vs Python: What's the Difference?, <https://www.guru99.com/r-vs-python.html>

1. Echo Systems of Python | Python 설치 및 환경설정

Python 설치 및 환경설정



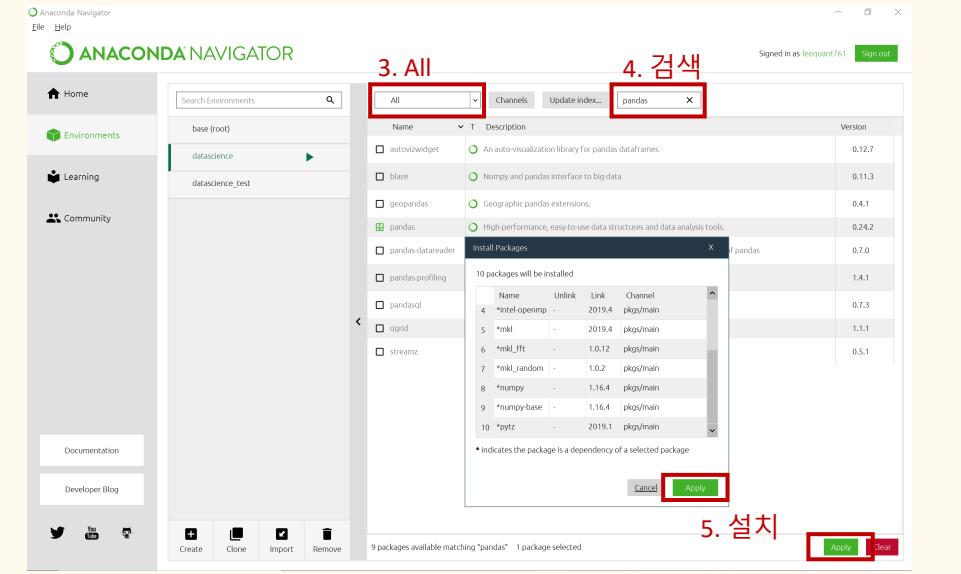
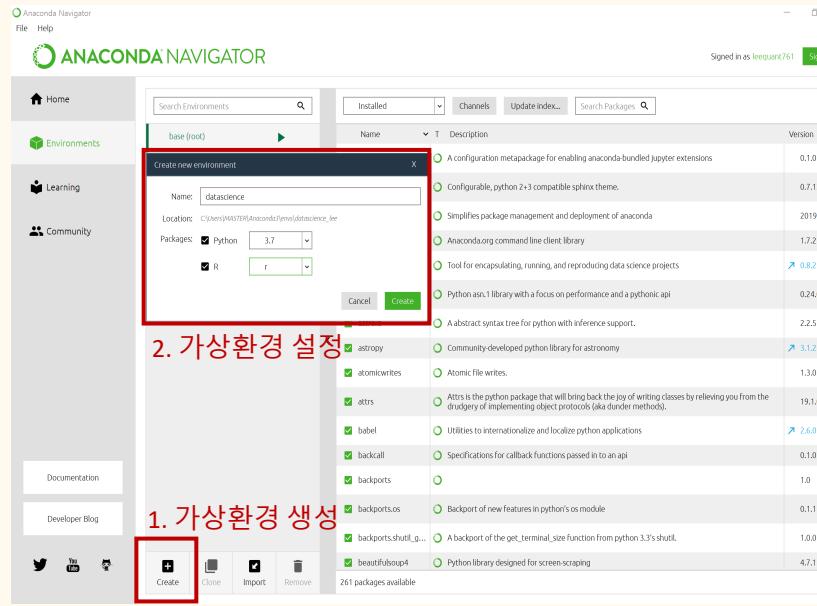
- 절차1 : anaconda(데이터과학 패키지 설치 및 관리 파이썬 배포판) 접속 : <https://www.anaconda.com/distribution/>
패키지 간의 종속성 문제 해결에 매우 유용하다.
- 절차2 : 운영체제 선택 : Windows, Linux, and Mac OS X
- 참고1 : Install for Just Me를 권장 관리자 권한 문제에서 자유롭다.
- 참고2 : Advanced Options 모두 선택 권장 환경 변수를 추가하면 cmd 창에서 anaconda 명령어 인식
- 참고3 : 기존에 아나콘다가 아닌 파이썬을 설치한 경우 제거하고 설치 권장

1. Echo Systems of R Python | Python 설치 및 환경설정

Python 가상 환경설정



- 가상환경은 진행하는 프로젝트에 맞는 파이썬 개발환경을 만들기 위해 사용한다.
- 절차1 : Anaconda Navigator에 들어간다.
- 절차2 : 가상환경 생성 및 설정
- 절차3 : 가상 환경에 pandas & scikit-learn 설치



1. Echo Systems of Python | Python 실행 모드

Python 실행 모드



대화식/스크립트

모드	대화식(Interactive)	스크립트(Script)
방식	입력 즉시 응답	.py에 코드 작성 후 파일 실행
용도	데이터 분석에 용이	프로젝트 개발에 용이
추천 환경	Jupyter Notebook	Pycharm, VS code, Spyder

1. Echo Systems of Python | Interface of Jupyter Notebook

기본 화면 구성



jupyter Untitled Last Checkpoint: 1분 전 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Logout

Trusted Python 3

In []:

셀 실행 : Shift + Enter
셀 추가 : a, b

셀을 마크다운으로 : m

타이틀 : #, ##, ###, #####

점 : *

들여쓰기 : >

Jupyter notebook 실행 방법

1. cmd 실행
2. conda activate datascience
3. jupyter notebook

2. Python Basic | Object I

Definition



Class / Instance / Object

Class	Instance	Object
<ul style="list-style-type: none">뭔가를 찍어내는 틀어떤 것을 만들기 위한 설계도톰	<ul style="list-style-type: none">틀로 만들어낸 실체설계도가 구현된 것톰's son	<ul style="list-style-type: none">붕어빵제품잭

2. R & Python Basic | Object I

Definition

R

