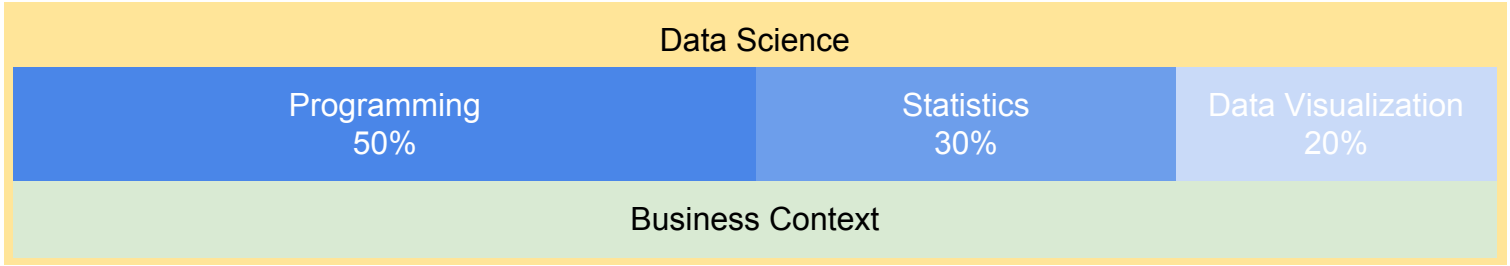
국민대 빅데이터경영MBA 파이썬 프로그래밍을 통한 데이터 분석

2016년 1학기 강의

정광윤 initialkommit@gmail.com

강의 계획

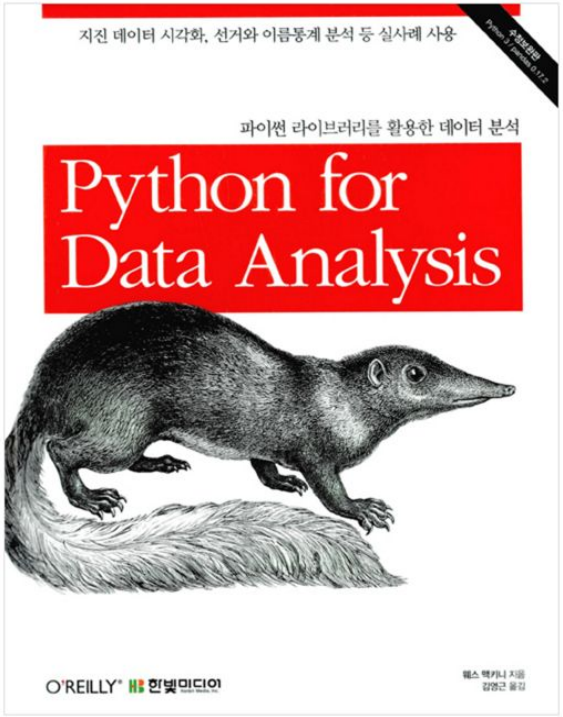
We Focus on

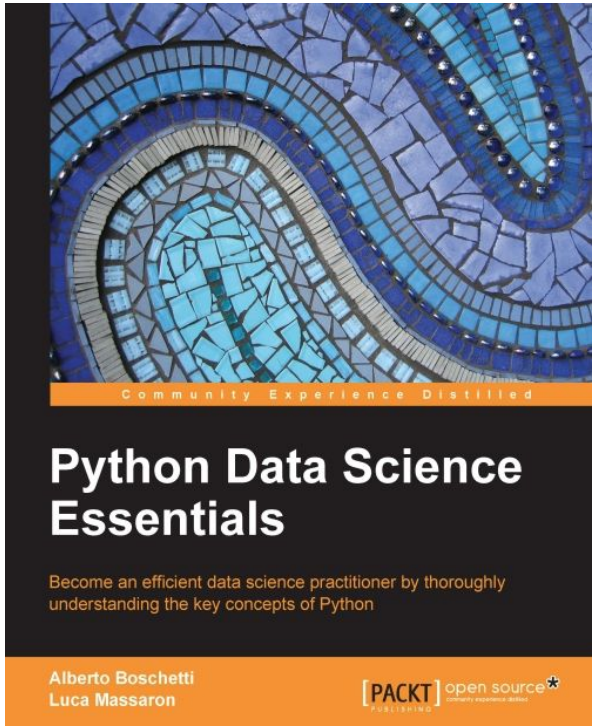


	First Half	First PT	Second Half	Final PT
Date	7개 주차(3/5~4/16)	8주차(4/23)	7개 주차(4/30~6/11)	16주차(6/18)
Goals	1. 개발환경에 친숙해지기 2. Python 익숙해지기 3. 개발에 필요한 도구 학습 (iPython, Github 등) 4. NumPy, Pandas 등 시작하기	지금까지 배운 내용을 토대로 자유 주제 발표	1. NumPy, Pandas 학습 2. 시각화 - Matplotlib, Bokeh 등 학습 3. Tutorial 학습	자유 주제로 종합 발표

국민대
파이썬
데이터

레퍼런스





국민대
파이썬
데이터

철학

We Learn . . .
10% of what we read
20% of what we hear
30% of what we see
50% of what we see and hear
70% of what we discuss
80% of what we experience
95% of what we teach others.

철학

1. see & experience (눈 & 손)
2. teach others (과제)
3. 숲과 나무
4. 정광윤 보다 파이썬

과제 및 발표

- 과제 40% 발표 40% 출석 10% 참여 10%
- 과제 40%: 주제 선정의 뚜렷한 목적 등 과제의 절차와 방법에 집중
- 발표 40%: 타 수강생에게 본인의 지식을 잘 전달하는지에 집중
- 주제
  - 데이터 분석
  - 수강생 개인이 관심이 있거나 유익이 되는 것으로 자유롭게 과제 선정
- 모든 자료는 수강생과 자료 공유 (Github 이용)
- 예시
  - Yes24의 추천 알고리즘 (http://hyunje.com/data%20analysis/2015/12/21/yes24-recommendation-1)
    - 추천 알고리즘을 보자는 것이 아님
  - Python Korea 2015, 유재명, Pay-thon (http://www.pycon.kr/2015/program/79)

강사 프로필

정광윤

현) SCM 솔루션, (주)드림아이즈 컨설팅팀 책임연구원 (www.dreamize.co.kr)
전) 시간 분석 서비스 스타트업 대표
전) SCM 시스템 개발 2년
전) 반도체 회사 생산관리 2년

산업정보시스템공학 전공
주중 9시 이후 답장 가능
initialkommit@gmail.com

프로젝트

현) W 반도체 회사 SCM BI/DW 개발중
제일모직 SCM
삼성SDI SCM
앰코코리아 SCM(SAP APO)
까사미아 SCM
대성전기 수주관리시스템

강의

2015.08~12 한국데이터베이스진흥원 ‘Python 데이터 분석’ 전문가 컬럼 연재
2015.10 Django Girls Seoul 코치
2015.08-09 초보자 대상 Django 웹 프로그래밍
2015,04 문과생 대상 Python을 이용한 데이터 분석

01
Please Introduce Yourself

공부하기 전에 먼저 서로 알아가겠습니다.

Kookmin University
Graduate School of Business Administration
MBA in Big Data Analytics

자기 소개

1. 이름
2. 현재 직업(학생, 직장인(직장인 이라면 어떤 분야) 등)
3. 왜 빅데이터경영MBA를 하게 되었는지
4. 개인 컴퓨터(노트북) 운영체제(윈도우, 맥, 리눅스 등)
5. 언어 경험(파이썬, 루비, SQL, 자바스크립트 등)
6. 데이터베이스 경험(오라클, MSSQL, PostgreSQL, MongoDB 등)
7. 관심사(관심있는 전공, 취미, 사회 현상 등)
8. 어떤 데이터를 분석해보고 싶은지

02
Introduction to Data Science

최근 유행하는 용어인 데이터 과학 혹은 데이터 사이언스는 무엇인지 살펴보면서 우리가 해야할 역할을 파악해보도록 하겠습니다.

Kookmin University
Graduate School of Business Administration
MBA in Big Data Analytics

Why Now Data Science?

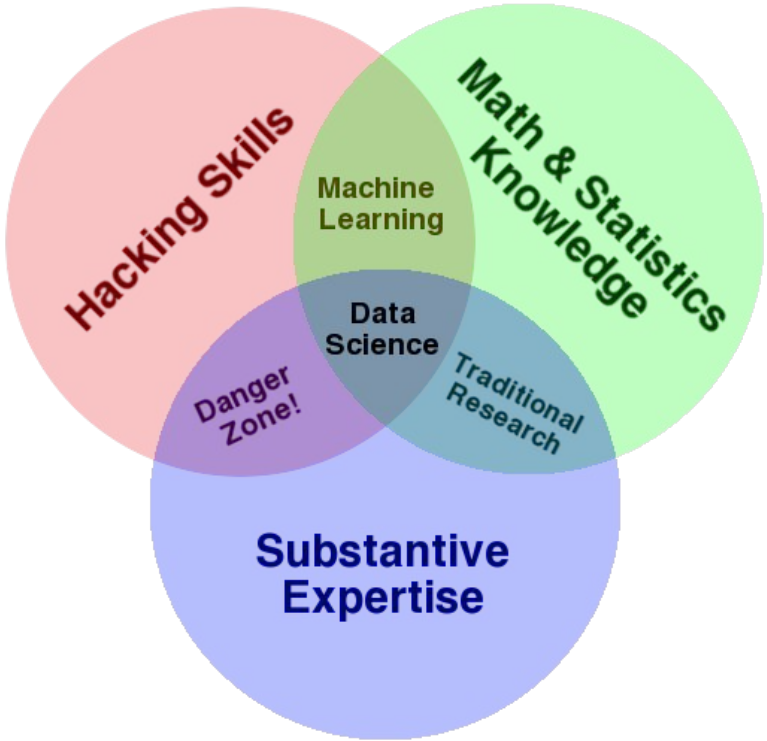
- 삶의 많은 측면에 관한 대량의 데이터를 갖고 있다.
- 저비용의 컴퓨터 처리 능력을 갖고 있다.
- 모든 온라인 활동이 기록되고 있다.
- 모든 온, 오프라인에서 데이터 수집 혁명과 비슷하게 '**데이터화**'되기 시작했다.
- 거의 모든 부문과 산업에서 데이터의 영향이 커지고 있다.
- 데이터 자체가 실시간으로 데이터 상품의 소재가 되고 있다.

데이터화Datafication

- Foreign Affairs 2013년 5/6월에서 '빅데이터의 등장'이라는 논문 발표(Kenneth Neil Cukier, Victor Mayer-Schoenberger)
- 데이터화를 삶의 모든 측면을 포착해서 그것을 데이터로 바꾸는 과정이라고 정의한다. 예를 들어 '구글의 증강현실 안경은 시선을 데이터화하고, 트위터는 생각의 조각들을 데이터화하며, 링크드인은 전문가 네트워크를 데이터화한다'고 말한다. 우리 자신이 데이터화되고 있거나, 아니면 우리 행동이 데이터화되고 있다. 예컨데 온라인상의 누군가 혹은 무엇을 '좋아할' 때 우리는 데이터화되는 것이다.
- 일단 대상을 데이터화하면, 우리는 그것의 사용 목적을 바꾸고 그 정보를 새로운 형식의 가치로 전환할 수 있다.

Doing Data Science, O'Reilly, 2014, Cathy O'Neil and Rachel Schutt

Data Science



American Data Scientist, Drew Conway, 2013
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The Origin of “Data Scientist” Term

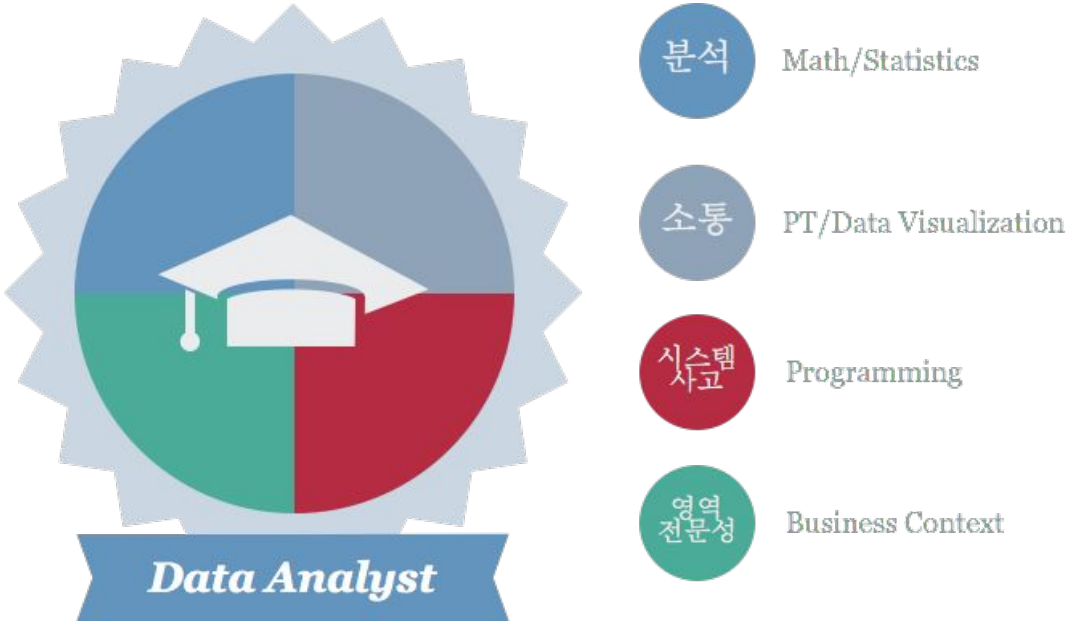
“In 2008, DJ Patil and Jeff Hammerbacher used the term ‘data scientist’ to define their jobs at LinkedIn and Facebook, respectively.”

- en.wikipedia.org/wiki/Data_science -

- DJ Patil은 2015년 2월 백악관 최초의 Chief Data Scientist로 임명되었다.
(www.wired.com/2015/02/white-house-names-dj-patil-first-us-chief-data-scientist)

- Email from DJ Patil: "How I Became Chief Data Scientist"
(www.whitehouse.gov/blog/2015/05/06/email-dj-patil-how-i-became-chief-data-scientist)

Data Analyst



Udacity, Data Analyst
https://www.udacity.com/career-paths/data-analyst

Sexy Skills of Data Geeks

STATISTICS

traditional analysis you're used to thinking about

DATA MUNGING

parsing, scraping, and formatting data

VISUALIZATION

graphs, tools, etc.

Nathan Yau, 2009, 'Rise of the Data Scientist'
https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/

Data Scientist

2015/02/07, 넘버웍스 대표 하용호, 어떻게 하면 데이터 사이언티스트가될 수 있나요?

I am a data scientist!

항목	나
통계학	
커뮤니케이션 능력	
프로그래밍 스킬	
전문 분야	

위 항목에 맞춰 스스로 본인을 솔직하게 평가해봅시다.

I am a data scientist!

항목	정광윤
통계학	산업공학 전공했는데도 여전히 어려움
커뮤니케이션 능력	좋아함, 데이터 시각화는 전문성을 더 높여야
프로그래밍 스킬	Python, SQL, PL/SQL, Javascript etc.
전문 분야	반도체, SCM

위 항목에 맞춰 스스로 본인을 솔직하게 평가해봅시다.

Data Scientist - 대학에서

사회과학에서 생물학까지 어떤 분야에서건 훈련된 과학자고, 대량의 데이터를 분석하며, 현실 세계의 문제를 해결함과 동시에 데이터의 구조, 크기, 무정형성, 복잡성과 같은 성격 때문에 발생하는 전산화의 문제들을 다뤄야만 한다.

Doing Data Science, O’Reilly, 2014, Cathy O'Neil and Rachel Schutt

Data Scientist - 산업에서

최고위직의 데이터 과학자는 회사의 데이터 전략을 세워야 한다.

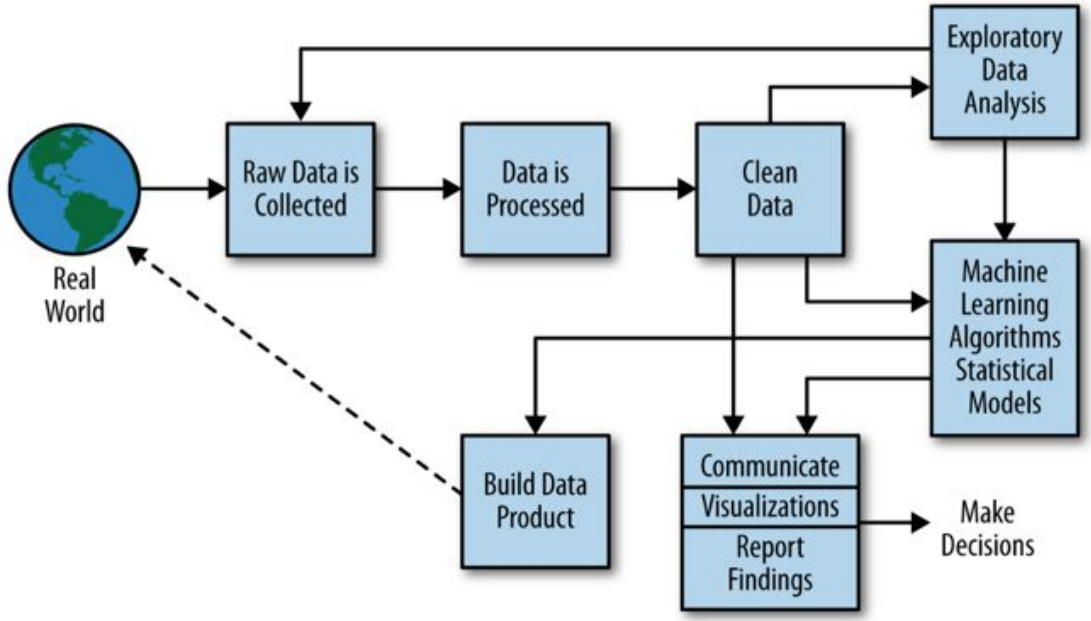
거기에는 다음과 같은 다양한 역할이 포함된다. 데이터를 수집하고 기록하기 위한 공학과 인프라부터 프라이버시 문제까지 어떤 데이터가 사용자 대면이고, 의사결정을 위해 데이터가 어떻게 사용될 것인지, 그것이 어떻게 제품에 반영될 것인지에 이르는 모든 사항을 기획해야 한다. 그는 공학자, 과학자, 분석가로 구성된 팀을 관리해야 하고, 최고경영자, 최고기술책임자, 제품 책임자 등 기업 내 다양한 리더들과 커뮤니케이션해야 한다.

일반적으로, 데이터 과학자는 데이터에서 의미를 뽑아내고 해석하는 방법을 아는 사람이다.

그러기 위해서는 인본주의적이어야 하고, 통계학과 기계학습의 도구와 방법 모두를 알아야 한다. 그는 데이터를 수집하고 정제하고 변환하는 과정에 많은 시간을 보낸다. 데이터는 항상 깔끔하지 않기 때문이다. 이 과정은 집념, 통계학, 소프트웨어공학 스킬을 요구한다. 이 스킬들은 또한 데이터에 내재된 편의(bias)를 이해하고 코드에서 생성된 결과를 디버깅하는 데 필수적이다. 일단 데이터를 다 매만지고 나면, 핵심 작업은 시각화와 데이터 감각을 포함한 탐색적 데이터 분석이다. 패턴을 발견하고, 모형을 만들며, 알고리즘을 고안한다. 이것은 프로토타입으로 활용되며 제품 용례, 제품의 전반적인 건강성 등을 이해하는 데 도움을 준다. 또한 그는 실험을 설계하고 데이터 주도 의사결정을 하는 데 있어 중요한 역할을 수행한다. 그는 명료한 해석과 시각화를 통해서 팀원, 공학자, 혹은 리더와 커뮤니케이션해야 한다. 그렇게 함으로써 데이터 자체에 관여하지 않는 동료들도 데이터에 담긴 뜻을 이해할 수 있을 것이다.

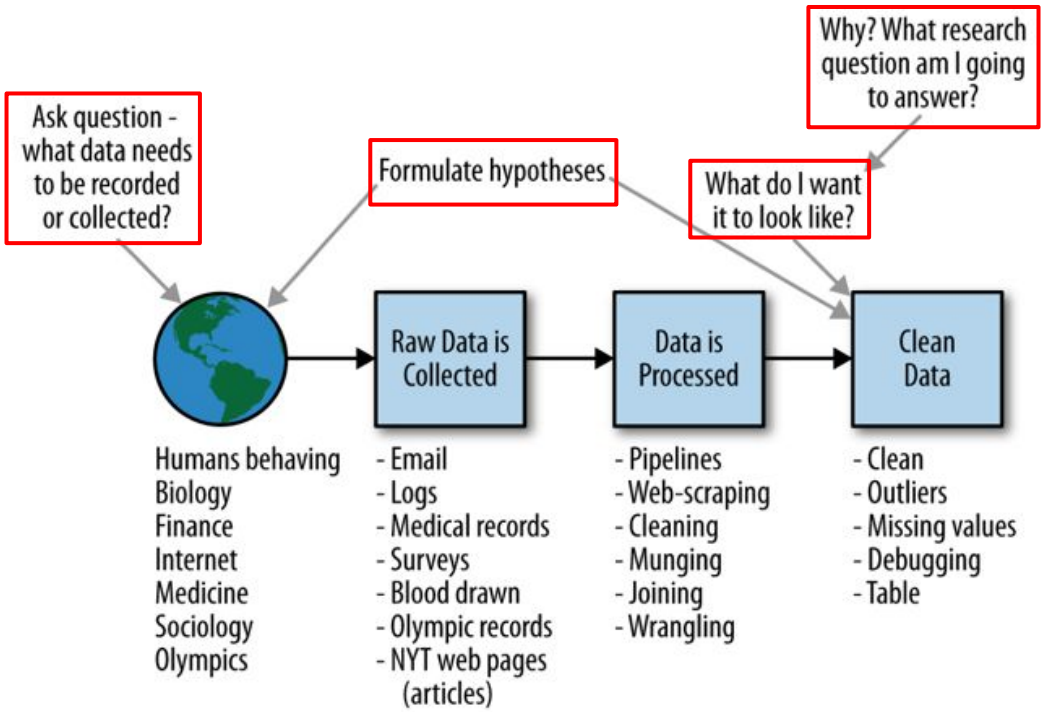
Doing Data Science, O'Reilly, 2014, Cathy O'Neil and Rachel Schutt

Data Science Process



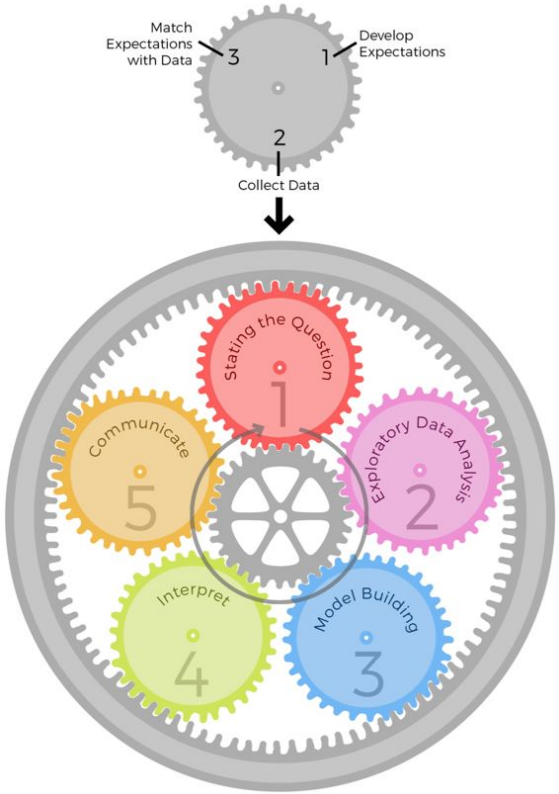
Doing Data Science, O'Reilly, 2014, Cathy O'Neil and Rachel Schutt

The data scientist is involved in every part



Doing Data Science, O'Reilly, 2014, Cathy O'Neil and Rachel Schutt

Epicycles of Analysis



1. Develop Expectations
2. Collect Data
3. Match Expectations with Data

1. Stating the Question
2. Exploratory Data Analysis
3. Model Building
4. Interpret
5. Communicate

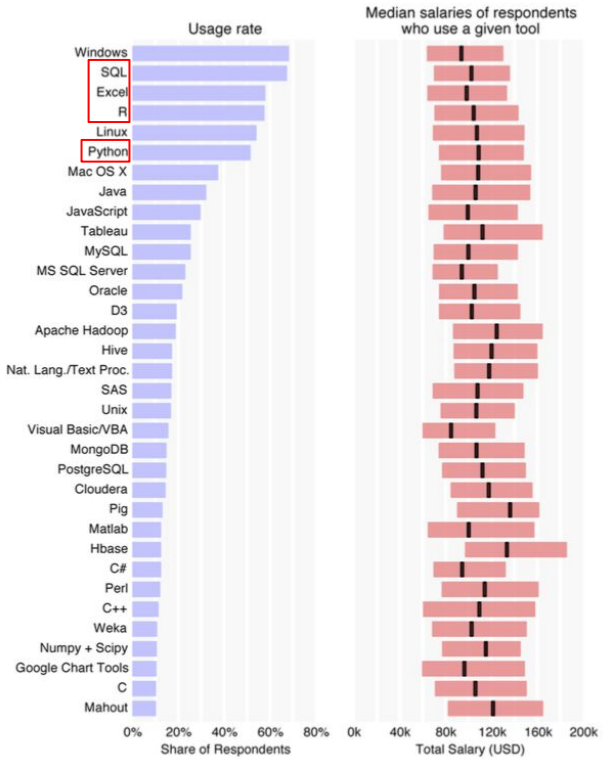
The Art of Data Science, 2015.12, Roger D. Peng, Data Science in Coursera

Epicycles of Analysis

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

The Art of Data Science, 2015, Roger D. Peng, Data Science in Coursera

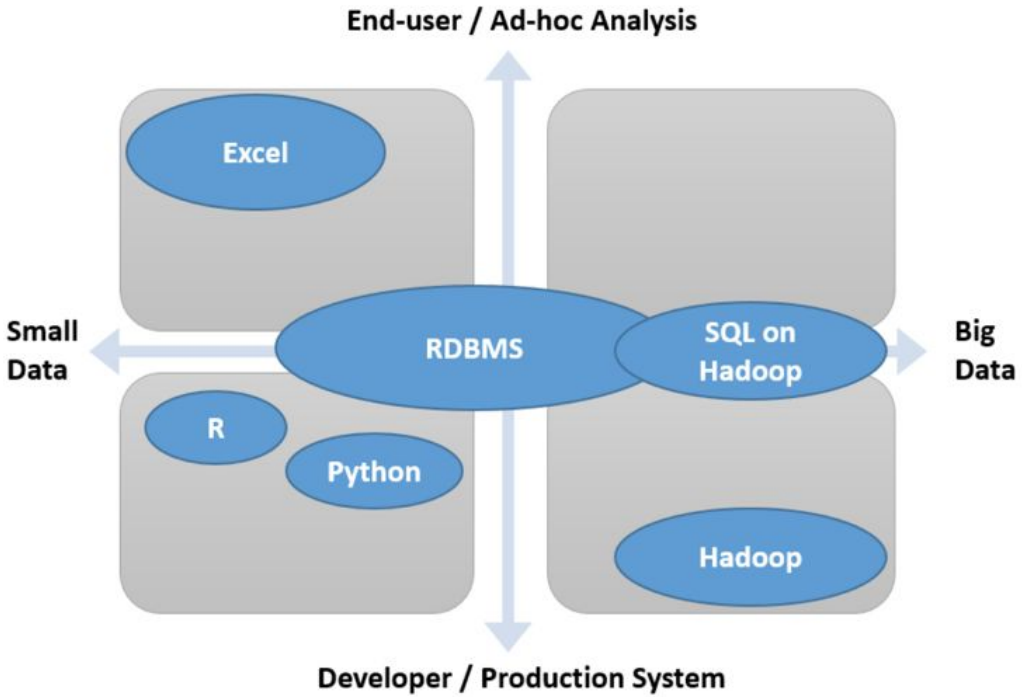
Most commonly used tools



2014 Data Science Salary Survey, O'Reilly
http://www.oreilly.com/data/free/2014-data-science-salary-survey.csp

국민대
파이썬
데이터

Tool Positioning



헬로데이터과학, 김진영
http://www.hellodatascience.com/?p=136

국민대
파이썬
데이터

03
Setting Development Environment

본격적으로 Python으로 개발 혹은 데이터 분석을 하기 전에 필수적으로 갖춰야할 개발 환경을 구축해보도록 하겠습니다. 이번 페이지에서는 간단히 소개하고 계속 실습하면서 조금씩 구체적으로 알아보겠습니다.

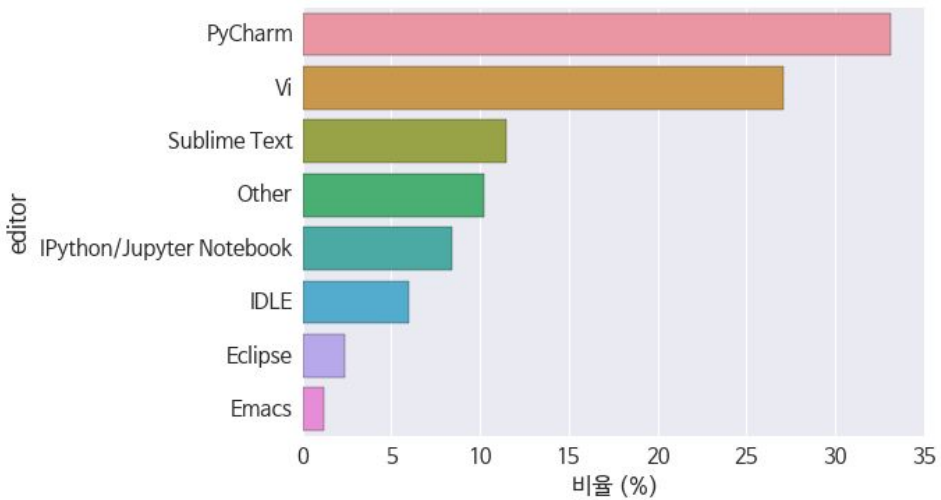
Kookmin University
Graduate School of Business Administration
MBA in Big Data Analytics

Python

- Windows
  - http://www.python.org
  - 설치 시 pip이 포함되어 설치할 것
- OSX
  - El Capitan에서는 Python 2.7이 기본적으로 설치되어 있음
  - brew or brew cask 등을 이용해 설치
  - Source를 다운받아 설치할 수도 있음
- 그러나 우리는 Python 3.5를 설치해 사용함

Editor

IDE: Integrated Development Environment 통합 개발 환경



Python Korea 2015 유재명 대표
http://nbviewer.jupyter.org/gist/euphoris/5b451790dd1dd7b5f7f1

PyCharm

Vi IMproved

Sublime Text 2 or 3

IP[y]: IPython
Interactive Computing

ATOM

Emacs

국민대
파이썬
데이터

Editor - Sublime Text 3

- 쉽게 사용할 수 있음
- 무료와 유료에 기능적 차이가 없음
- 다양한 패키지를 갖고 있음(기능을 추가 설치할 수 있고 커스터마이징 가능)
- 많은 사용자들이 사용하고 있음

설치방법

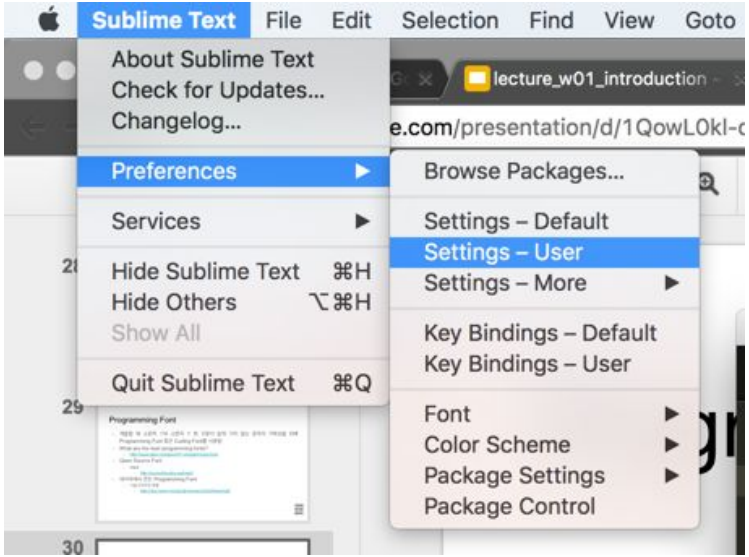
- www.sublimetext.com/3
- 본인의 OS에 맞게 다운로드

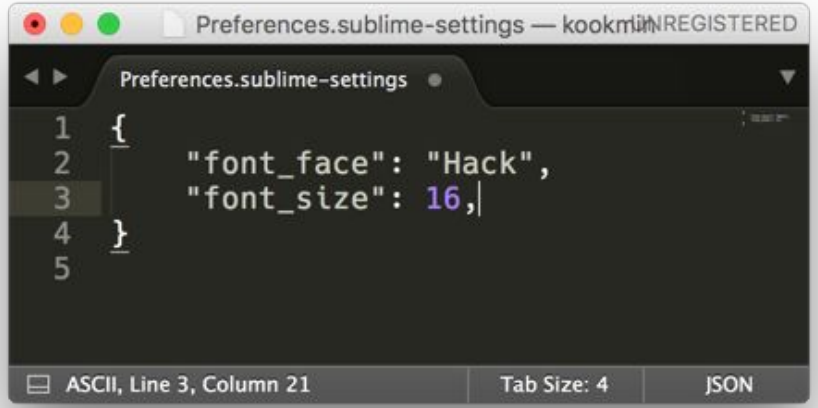
1. 패키지 컨트롤 설치(packagecontrol.io)
2. 패키지 설치
   a. Python Auto-Complete
   b. Jedi - Python Autocompletion
   c. SublimeREPL
   d. SublimeLinter
      i. 이 기능을 사용하기 위해서 pip으로 pylint를 설치해야함
   e. SublimeLinter - pylint
   f. pep8
   g. ConvertToUTF-8
   h. 이 외에 본인이 원하는 패키지를 검색해 설치

Programming Font

- 개발할 때 소문자 'i'와 소문자 'l' 등 구분이 쉽게 가지 않는 문자의 가독성을 위해 Programming Font 혹은 Coding Font를 사용함
- What are the best programming fonts?
  - http://www.slant.co/topics/67/~programming-fonts
- Open Source Font
  - Hack
    - http://sourcefoundry.org/hack/
- 네이버에서 만든 Programming Font
  - 나눔고딕코딩 글꼴
    - http://dev.naver.com/projects/nanumfont/download

Programming Font





Windows에서는 상단 메뉴에 Preferences를 선택

Terminal

- 개발할 때 가장 먼저 친숙해져야할 대상
- 서버(iPython) 구동 등 개발 환경을 컨트롤할 때 사용
- Windows
  - cmd
  - Windows에서 Unix 명령어 가능하게 해주는 유틸리티
    - GOW (GNU On Windows)
- OSX
  - Terminal
  - iTerm 2

Package Manager

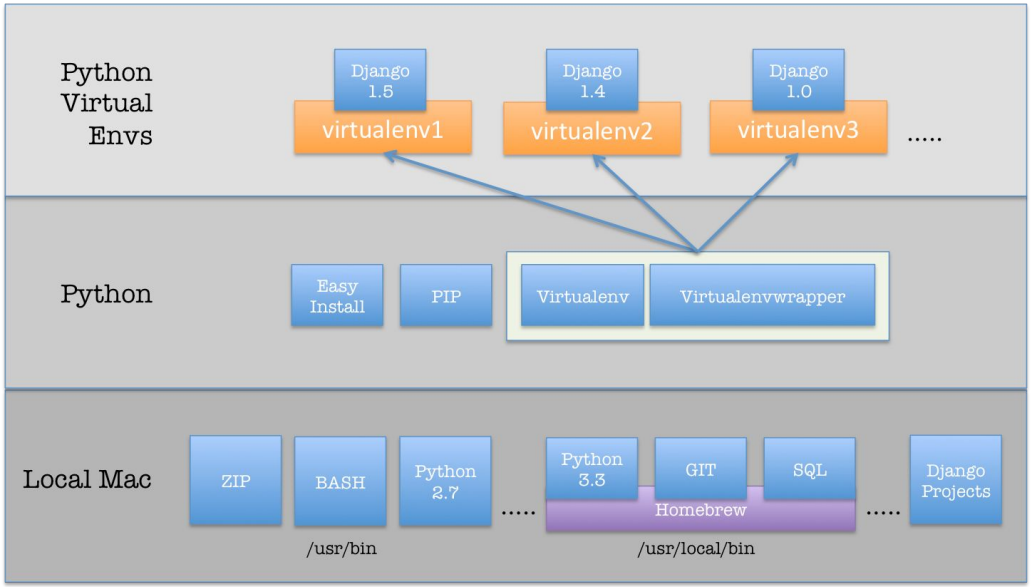
- OS
  - Ubuntu: apt-get
  - CentOS: yum
  - OSX: brew
- Language
  - Front-End for Web: npm, bower
  - Python: pip
  - Ruby: gem
- 각 운영체제 혹은 언어에 따라 오픈 소스 패키지를 간편하게 다운받아 설치 및 관리를 용이하게 해주는 것을 패키지 매니저라고 한다.

Package Manager - PIP

- Pip Installs Package or Python
- Python 설치와 함께 설치됨
- Python은 수없이 많고 다양한 Package가 존재하며 이를 PIP이라는 패키지 매니저로 쉽고 간편하게 설치 및 관리를 할 수 있음
- 사용법
  - pip search 패키지명
  - pip freeze
    - 설치한 Package 목록보기
  - pip install 패키지명
  - pip install 패키지명==특정 버전
  - pip uninstall 패키지명

가상 환경 - 개념

- 프로젝트별로 패키지를 관리
- 다양한 프로젝트를 하면서 각 프로젝트에 설치한 패키지를 프로젝트별/버전별로 관리
- 물리적 공간(하드디스크)를 나눠 사용하는 것이 아님



http://www.jackalventure.com/blog/tag/python-programming-language/

가상 환경 - 설치

- **virtualenv + virtualenvwrapper**
  - 참고: http://docs.python-guide.org/en/latest/dev/virtualenvs/
  - virtualenvwrapper는 virtualenv 명령어를 쉽게 사용할 수 있도록 wrap.
- autoenv
- pyenv

설치방법

1. virtualenv
   a. pip install virtualenv
2. virtualenvwrapper
   a. OSX - pip install virtualenvwrapper
   b. Windows - pip install virtualenvwrapper-win
3. 가상 환경 파일이 들어갈 디렉토리 생성
   a. C:\.virtualenvs

가상 환경 - 사용 방법

- 가상 환경 생성
  - mkvirtualenv 가상환경이름
- 가상 환경 삭제
  - rmvirtualenv 가상환경이름
- 가상 환경 바꾸기
  - workon 가상환경이름

iPython(Jupyter)

- http://jupyter.org/
- iPython + Julia 기능이 포함되면서 Jupyter라는 이름으로 바뀌게 됨
- 터미널창에서 ipython이라 입력 가능
- ipython notebook은 deprecated, 대신 jupyter notebook 사용
- iPython
  - Interactive computing interpreter
- jupyter notebook
  - Server-client application that allows editing and running notebook documents via a web browser
- 앞으로 모든 분석은 jupyter notebook을 사용할 예정

03
Next Lecture

Kookmin University
Graduate School of Business Administration
MBA in Big Data Analytics

2주차 강의

1. 가상 환경+PIP 실습
2. Git, Github 실습
3. Python
   a. 객체 지향 프로그래밍 개념 학습 및 실습
   b. 기본 문법 학습 및 실습

국민대
파이썬
데이터