Introduction

Data Management
Fall 2020
Osaka U

Shuhei Kitamura

About the course

Data Management (Undergrads), Data Management & Analysis (Masters)

- Date & time: Tuesday 1st & Thursday 1st
- Location

  (onsite): Toyonaka Sōgō Gakkan 401 (Tue), 302 (Thur)

  (online): Blackboard Collaborate Ultra
- Language: English
- Instructor: Shuhei KITAMURA (kitamura@osipp)
- Office hours: By appointment
- TA: Masateru YAMATANI (u119659b@ecs)

About the course (cont.)

Objective: Learn together how to conduct empirical research

In particular, you are expected to:

- Learn basic knowledge for conducting empirical research
- Obtain skills to write code for making and analyzing data, and writing academic papers and slides

About the course (cont.)

Prerequisite & requirement

- Basic knowledge in statistics and econometrics
- Bring your own laptop to the class
    - Windows/Mac
    - Linux (support not always guaranteed)

No textbook. Lecture slides, code, and useful references will be provided

Course materials are downloadable from CLE

Grading

Three assignments (40%)

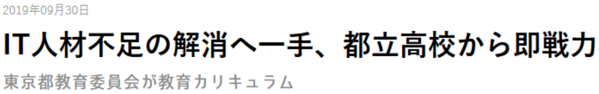
- Write your code to answer the questions
- Work individually
- Hand in your answers via CLE

Final report (60%)

- Find your own research question and data
- Write code to clean data and conduct analyses to find answers to your research question
- Use only Python or R
- Your research idea and data should relate to COVID-19

  - Does a certain policy/event affect the number of confirmed cases and deaths? Is there any association between certain factors and the spread of the virus?, etc.
- Max 3 people can work together
- Graded based on (a) the creativity of your research idea and (b) whether you properly write code to clean data and conduct analyses
- Hand in your ipynb file with code and data via CLE

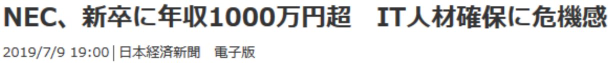
Demand for data scientists (private sector)



2019年09月30日

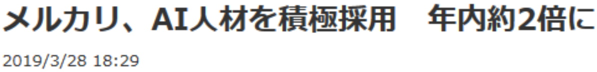
IT人材不足の解消へ一手、都立高校から即戦力

東京都教育委員会が教育カリキュラム



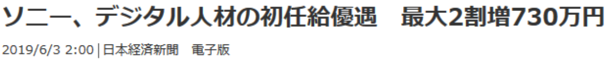
NEC、新卒に年収1000万円超　IT人材確保に危機感

2019/7/9 19:00|日本経済新聞　電子版



メルカリ、AI人材を積極採用　年内約2倍に

2019/3/28 18:29

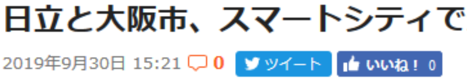


ソニー、デジタル人材の初任給優遇　最大2割増730万円

2019/6/3 2:00|日本経済新聞　電子版

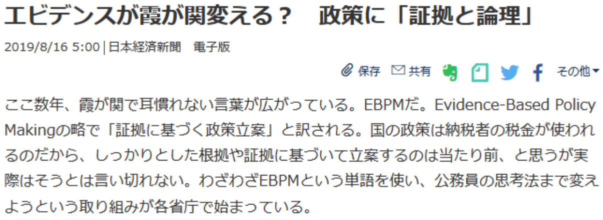
Tech companies are looking for talented individuals who can analyze data

Demand for data scientists (public sector)



日立と大阪市、スマートシティで連携協定

2019年9月30日 15:21 0 ツイート いいね！ 0

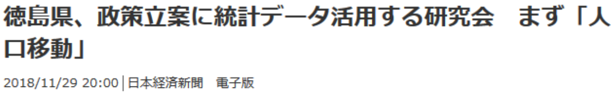


エビデンスが霞が関変える？　政策に「証拠と論理」

2019/8/16 5:00 | 日本経済新聞　電子版

保存　共有　その他

ここ数年、霞が関で耳慣れない言葉が広がっている。EBPMだ。Evidence-Based Policy Makingの略で「証拠に基づく政策立案」と訳される。国の政策は納税者の税金が使われるのだから、しっかりとした根拠や証拠に基づいて立案するのは当たり前、と思うが実際はそうとは言い切れない。わざわざEBPMという単語を使い、公務員の思考法まで変えようという取り組みが各省庁で始まっている。



徳島県、政策立案に統計データ活用する研究会　まず「人口移動」

2018/11/29 20:00 | 日本経済新聞　電子版

Ministries and municipalities are implementing Evidence-Based Policy Making (EBPM) based on data analyses

Growing demand

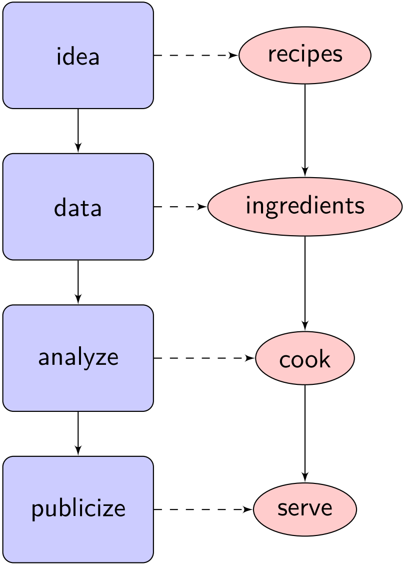
There are two major types of data specialists (which are not necessarily mutually exclusive)

- Programmers
- Data scientists

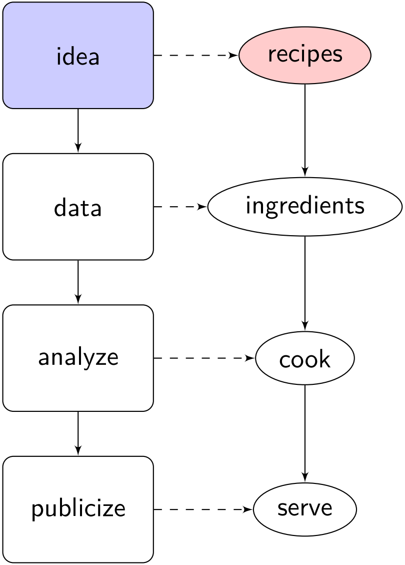
Data science is not only for engineers

- E.g., # of econ PhDs finding jobs in tech companies (Airbnb, Amazon, Facebook, etc.) after graduation is increasing

General workflow of empirical work



Step 1



"Good food begins with a good recipe"

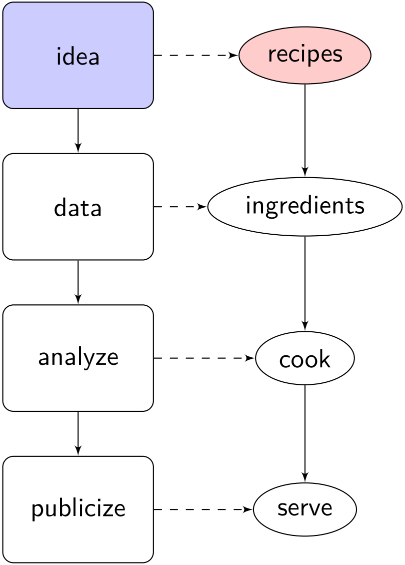
Goal: Get good ideas

- You may get ideas while reading articles, browsing websites, taking shower, etc.
- You can easily forget the idea itself and/or where it is stored
- You can combine a new idea with an old one if both are stored well

Useful tools for storing, organizing, and sharing ideas and resources such as links to web pages:

- Evernote
- Readcube

Step 1



"Good food begins with a good recipe"

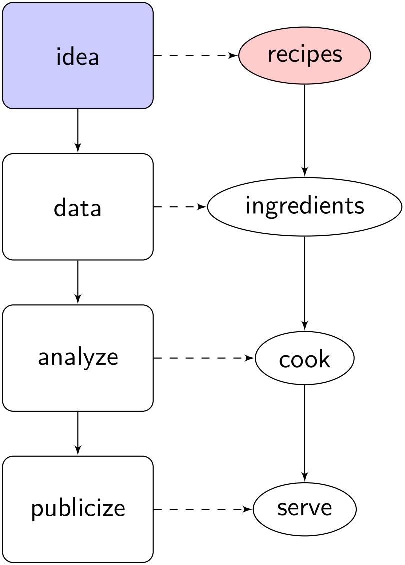
Goal: Get good ideas

- You may get ideas while reading articles, browsing websites, taking shower, etc.
- You can easily forget the idea itself and/or where it is stored
- You can combine a new idea with an old one if both are stored well

Useful tools for storing, organizing, and sharing ideas and resources such as links to web pages:

- Evernote
- Readcube

Step 1



"Good food begins with a good recipe"

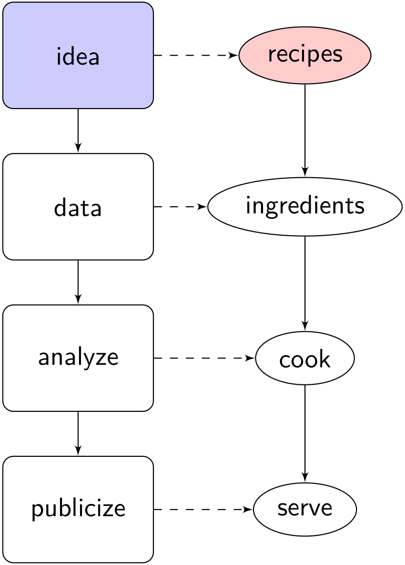
Goal: Get good ideas

- You may get ideas while reading articles, browsing websites, taking shower, etc.
- You can easily forget the idea itself and/or where it is stored
- You can combine a new idea with an old one if both are stored well

Useful tools for storing, organizing, and sharing ideas and resources such as links to web pages:

- Evernote
- Readcube

Step 1



"Good food begins with a good recipe"

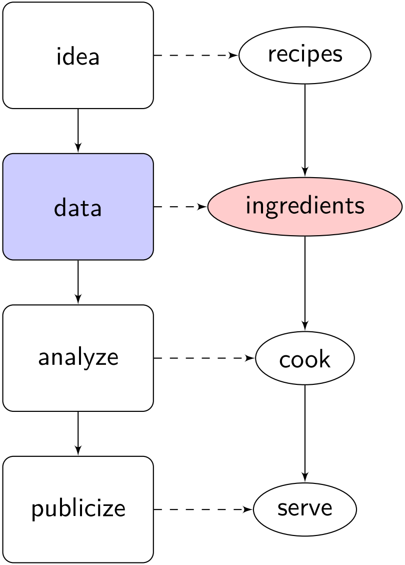
Goal: Get good ideas

- You may get ideas while reading articles, browsing websites, taking shower, etc.
- You can easily forget the idea itself and/or where it is stored
- You can combine a new idea with an old one if both are stored well

Useful tools for storing, organizing, and sharing ideas and resources such as links to web pages:

- Evernote
- Readcube

Step 2



"Good food is made with good ingredients"

Goal: Collect good data and clean them properly

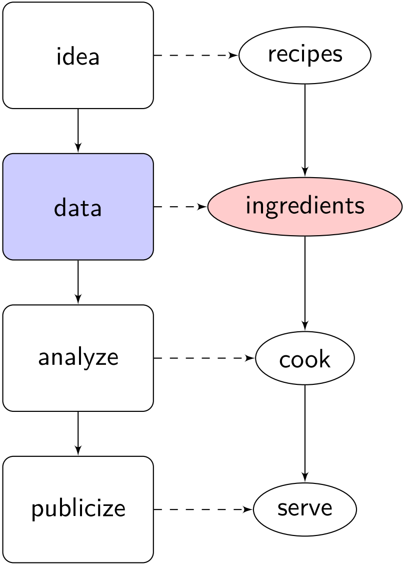
Useful tools for collecting and cleaning data:

- Python & R
- STATA

Useful tools for storing and sharing data:

- Dropbox

Step 2



"Good food is made with good ingredients"

Goal: Collect good data and clean them properly

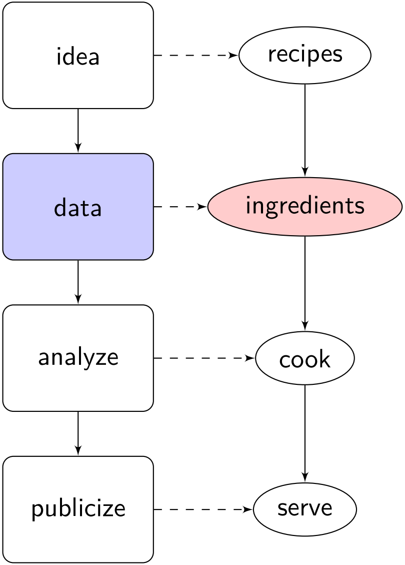
Useful tools for collecting and cleaning data:

- Python & R
- STATA

Useful tools for storing and sharing data:

- Dropbox

Step 2



"Good food is made with good ingredients"

Goal: Collect good data and clean them properly

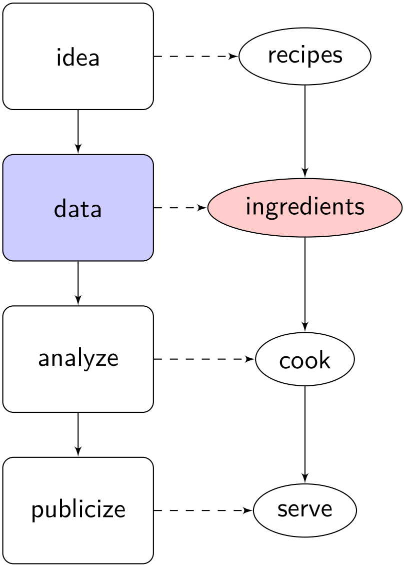
Useful tools for collecting and cleaning data:

- Python & R
- STATA

Useful tools for storing and sharing data:

- Dropbox

Step 2



"Good food is made with good ingredients"

Goal: Collect good data and clean them properly

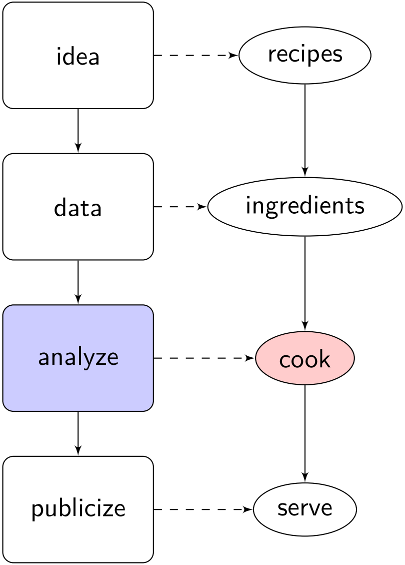
Useful tools for collecting and cleaning data:

- Python & R
- STATA

Useful tools for storing and sharing data:

- Dropbox

Step 3



"Good food is cooked well"

Goal: Analyze data properly

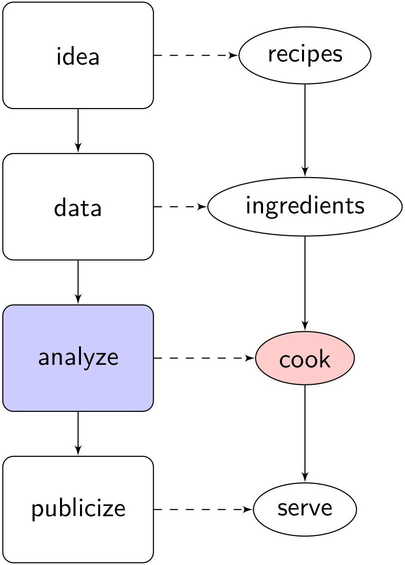
Useful tool for analyzing data:

- Python & R
- STATA

Useful knowledge for analyzing data:

- Statistics
- Econometrics
- AI/machine learning/deep learning, etc.

Step 3



“Good food is cooked well”

Goal: Analyze data properly

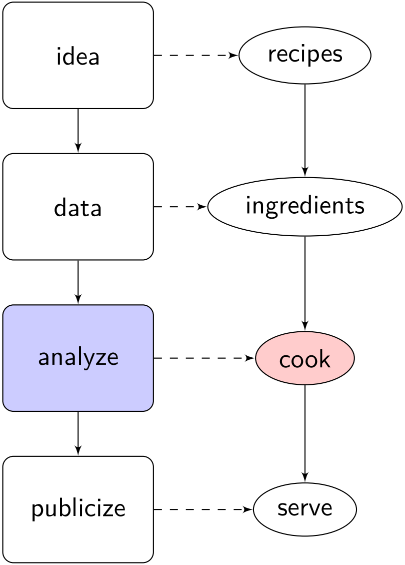
Useful tool for analyzing data:

- Python & R
- STATA

Useful knowledge for analyzing data:

- Statistics
- Econometrics
- AI/machine learning/deep learning, etc.

Step 3



"Good food is cooked well"

Goal: Analyze data properly

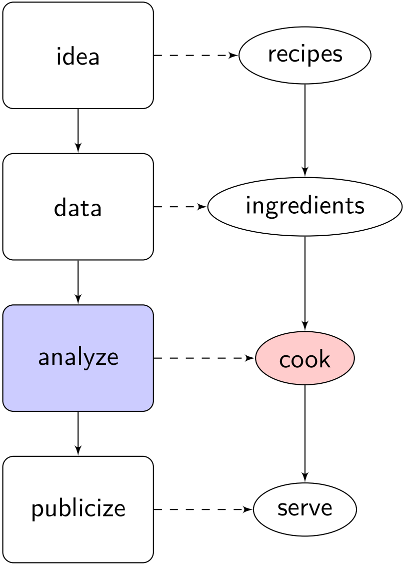
Useful tool for analyzing data:

- Python & R
- STATA

Useful knowledge for analyzing data:

- Statistics
- Econometrics
- AI/machine learning/deep learning, etc.

Step 3



"Good food is cooked well"

Goal: Analyze data properly

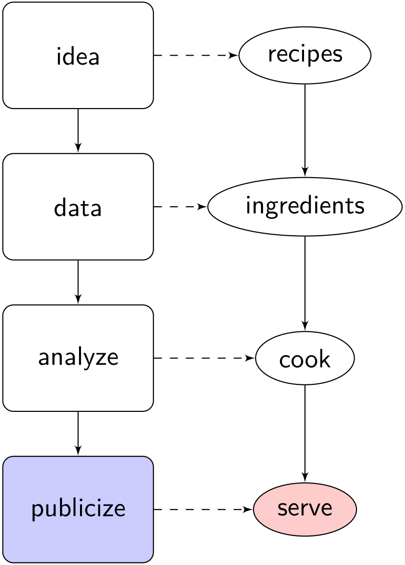
Useful tool for analyzing data:

- Python & R
- STATA

Useful knowledge for analyzing data:

- Statistics
- Econometrics
- AI/machine learning/deep learning, etc.

Step 4



"Good food is served well"

Goal: Show results nicely and intuitively

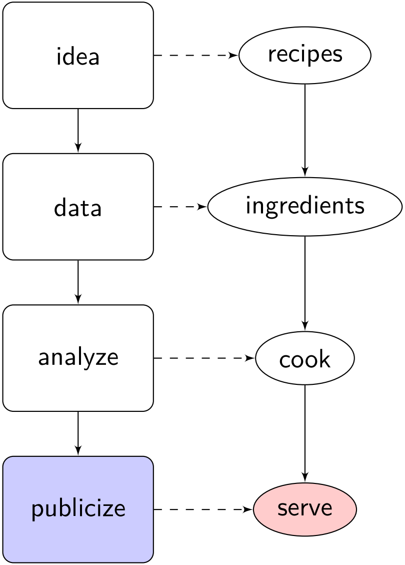
Useful tool for making figures and tables:

- Python & R
- STATA

Useful tool for writing academic papers and slides:

- TEX

Step 4



"Good food is served well"

Goal: Show results nicely and intuitively

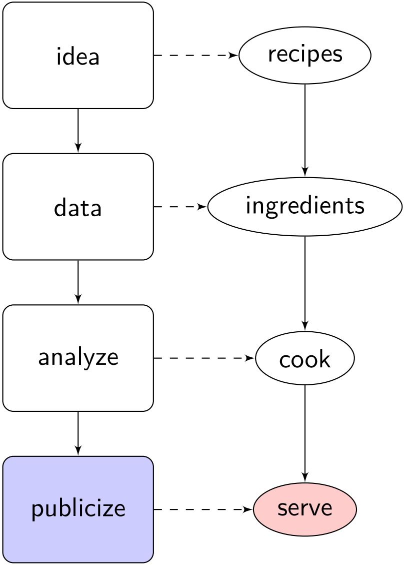
Useful tool for making figures and tables:

- Python & R
- STATA

Useful tool for writing academic papers and slides:

- TEX

Step 4



"Good food is served well"

Goal: Show results nicely and intuitively

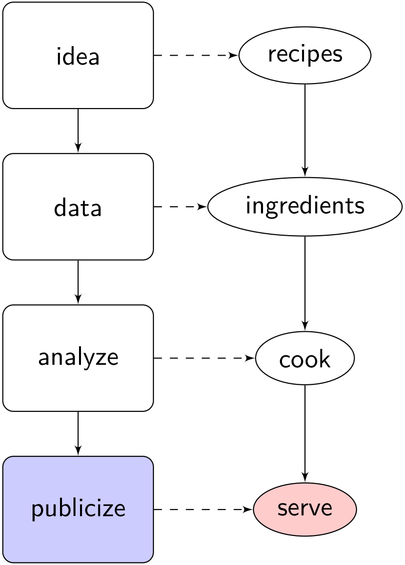
Useful tool for making figures and tables:

- Python & R
- STATA

Useful tool for writing academic papers and slides:

- $\TeX$

Step 4



"Good food is served well"

Goal: Show results nicely and intuitively

Useful tool for making figures and tables:

- Python & R
- STATA

Useful tool for writing academic papers and slides:

- TEX

Course plan

1. Idea [Recipes] (this lecture) (5%)
2. Making data [Ingredients] Python & R (60%)
3. Analyzing data [Cooking] Python & R (30%)
   - Making figures and tables, running regressions, etc.
4. Publicizing results [Serving] $\LaTeX$ (5%)

Course plan (cont.)

Each class has coding exercises

We will use Jupyter Notebook (for Python) and RStudio (for R)

1. Idea

Idea

Idea → often downplayed, but ideas are very important for empirical work

A good idea per 100 mediocre ideas (say)

A good idea is *important* and *feasible*

- Important = Contribution to the literature, important for business strategies or policy-making
- Feasible = It is possible to test the idea using data

How to come up with a good idea?

- Frequently ask empirical questions to yourself (while reading news articles, etc.)
- Store ideas nicely and revisit them occasionally
- Read academic articles just to know ongoing discussions and knows and unknowns in the literature

How to store resources

Recommend: Store resources (ideas, articles, data, etc.) in cloud

Why cloud?

- Handy (easy to store, access, organize, and share)
- Relatively secure
- Saving local spaces

A nice tool for storing ideas, web resources, etc.

Evernote

- Free
- Unlimited storage
- Web Clipper available for Firefox and Chrome
- Basic account allows sync only between two devices

A nice tool for storing journal articles

Readcube

- Free, for local use. Online version, free for 30 days, then \$3-5/month
- Unlimited storage
- Web Importer available for Chrome
- Easy to make reference lists

Other options: Mendeley, Endnote, etc.

A nice tool for storing data

Dropbox

- Free (Basic, 2GB), $12/month (Plus, 2TB)

Other options (free plan): Google One (15GB), Amazons (5GB), Box (10GB), etc.

2. Making data

Python and R

In this course, we will use Python and R

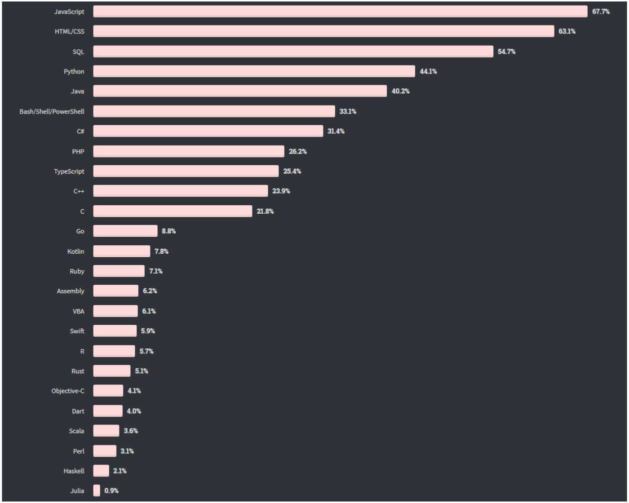
- Both are free and suitable for handling data
- Python is a popular language for programmers
- R is getting more popular in academia

We start from Python, then move on to R

Focus more on Python (70%). Three reasons:

- Python is a general language
- There are many overlaps
- Many of you may work in non-academic sectors after graduation

Popularity of languages



Source: Stack Overflow Developer Survey 2020

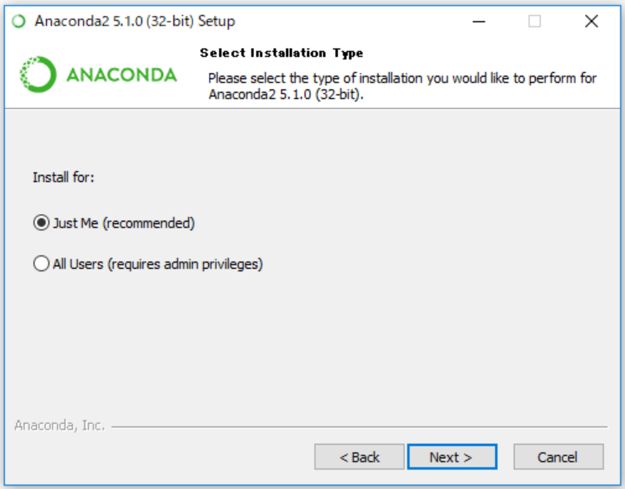
Install Jupyter Notebook (Python)

Install Jupyter Notebook using the Anaconda Distribution

Which Python version should I use?

- "Python 2.x is legacy, Python 3.x is the present and future of the language."
- E.g., $3/2 = 1$ in Python 2.x but $3/2 = 1.5$ in Python 3.x
- Unless Python 2.x is required (e.g., ArcGIS Desktop), choose Python 3.x

Install Jupyter Notebook (Python) (cont.)



Anaconda2 5.1.0 (32-bit) Setup

Select Installation Type

Please select the type of installation you would like to perform for Anaconda2 5.1.0 (32-bit).

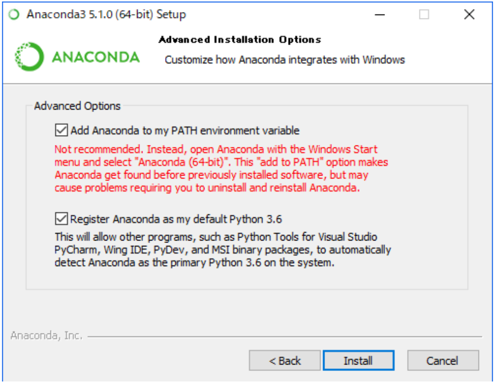
Install for:

● Just Me (recommended)

○ All Users (requires admin privileges)

Choose “Just Me (recommended)”

Install Jupyter Notebook (Python) (cont.)



Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

☑ Add Anaconda to my PATH environment variable

Not recommended. Instead, open Anaconda with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

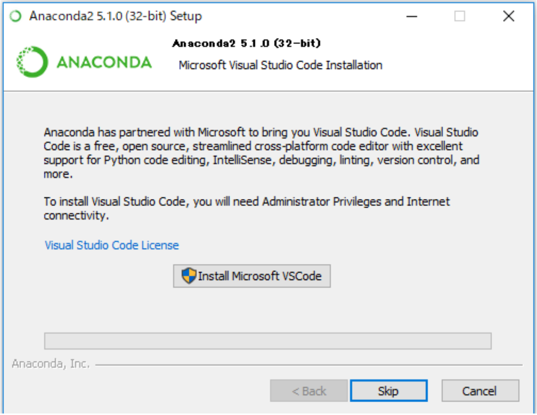
☑ Register Anaconda as my default Python 3.6

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.6 on the system.

Anaconda, Inc.

- Check **BOTH**
- **CAUTION**: Double-check whether any previously installed software uses Python (e.g. ArcGIS) before starting installation
- For ArcGIS, a path file should be added to Anaconda3/Lib/site-packages after installation. Otherwise, Python won't recognize ArcPy

Install Jupyter Notebook (Python) (cont.)



Anaconda2 5.1.0 (32-bit)

Microsoft Visual Studio Code Installation

Anaconda has partnered with Microsoft to bring you Visual Studio Code. Visual Studio Code is a free, open source, streamlined cross-platform code editor with excellent support for Python code editing, IntelliSense, debugging, linting, version control, and more.

To install Visual Studio Code, you will need Administrator Privileges and Internet connectivity.

Visual Studio Code License

If you like, install Visual Studio Code (a text editor for coding)

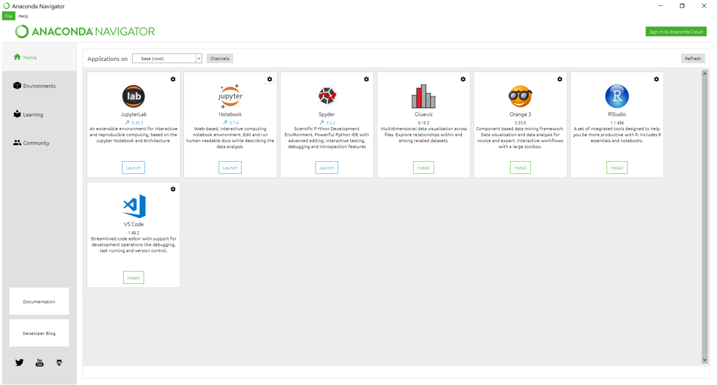
- Other options: Sublime Text, Vim

Launch Jupyter Notebook

There are two ways to launch Jupyter Notebook

- 1. Using Anaconda Navigator
- 2. Using Command Prompt/Terminal

1. Using Anaconda Navigator



JupyterLab
0.35.3
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook
5.7.4
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Spyder
3.3.2
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz
0.15.2
Multidimensional data visualization across files. Explore relationships within and among related datasets.

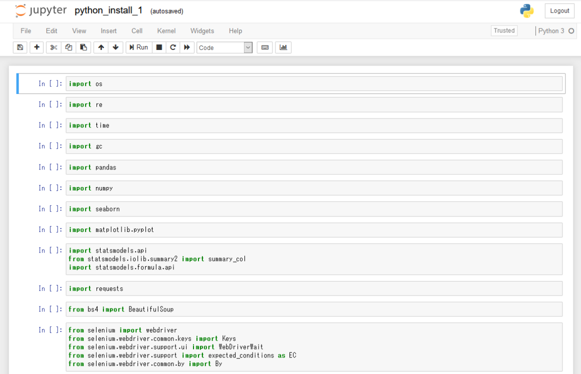
Orange 3
3.23.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

VS Code
1.48.2
Streamlined code editor with support for development operations like debugging, task running and version control.

- Search and click Anaconda Navigator (AN)
- In AN, press the “Launch” button under Jupyter Notebook (JN)
- In JN, navigate to the downloaded folder and click python_install_1.ipynb

2. Using Command Prompt/Terminal



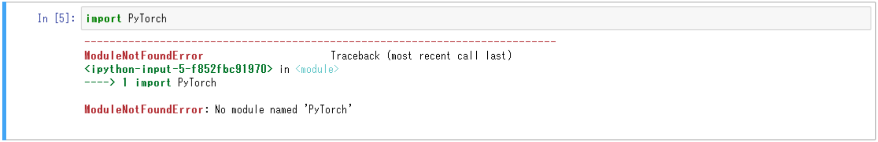
```
In [ ]: import os

In [ ]: import re

In [ ]: import time

In [ ]: import gc

In [ ]: import pandas

In [ ]: import numpy

In [ ]: import seaborn

In [ ]: import matplotlib.pyplot

In [ ]: import statsmodels.api
        from statsmodels.iolib.summary2 import summary_col
        import statsmodels.formula.api

In [ ]: import requests

In [ ]: from bs4 import BeautifulSoup

In [ ]: from selenium import webdriver
        from selenium.webdriver.common.keys import Keys
        from selenium.webdriver.support.ui import WebDriverWait
        from selenium.webdriver.support import expected_conditions as EC
        from selenium.webdriver.common.by import By
```

- Launch Command Prompt/Terminal. In the command line, type

```
jupyter notebook
```

- In JN, go to the downloaded folder and click python_intall_1.ipynb
- A screen like the above picture shows up

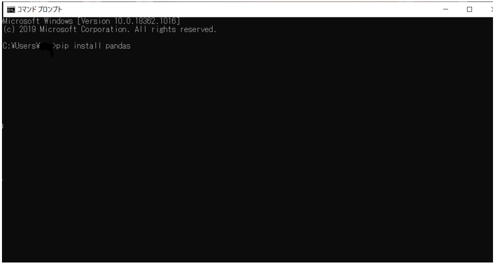
Error



```
In [5]: import PyTorch

        ---------------------------------------------------------------------------
        ModuleNotFoundError                       Traceback (most recent call last)
        <ipython-input-5-f852fbc91970> in <module>
        ----> 1 import PyTorch

        ModuleNotFoundError: No module named 'PyTorch'
```

- Run all the cells in python_intall_1.ipynb (Cell → Run All)
- If you get an error message like this, you need to install the package/module using Command Prompt/Terminal

Import packages/modules



```
Microsoft Windows [Version 10.0.18362.1016]
(c) 2019 Microsoft Corporation. All rights reserved.

C:¥Users¥[redacted]>pip install pandas
```

- Launch Command Prompt/Terminal ($\neq$ the one you used to launch JN. You should launch new one) and type

```
conda install xxx
```

or

```
pip install xxx
```

where xxx is the name of the package/module for which you get an error

Summary

- General workflow
- Course plan
- 1. Idea
- 2. Making data

References: Python

Online sources:

- QuantEcon (English)
- Matsuo Lab at U of Tokyo (Japanese)
- Online tutorial services (most of them are not free)

  - E.g., DataCamp, PyQ, progate, TechAcademy

Books:

- *Python Data Science Handbook: Essential Tools for Working with Data* (web)
- *Fundamentals of Data Visualization* (web)