Introduction

Data Management
Fall & Winter 2019
Osaka U

Shuhei Kitamura

About the course

Data Management (B), Data Management & Analysis (M)

- Date & time: Thursdays 1st hour
- Location: Here
- Language: English
- Instructor: Shuhei KITAMURA
- Office hours: Fridays 3-4pm
- TA: Ryo MIKAMI

About the course (cont.)

Objective: Learn together how to conduct empirical research

In particular, your 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 and useful references will be provided

Course materials are downloadable from CLE

Grading

Four assignments (100%)

- Hand-in your code and answers via CLE

Demand for data analysts (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

IT industries have been growing. Tech companies seem eager to recruit those who can handle data

Demand for data analysts (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 have been promoting Evidence-Based Policy Making (EBPM)

Growing demand

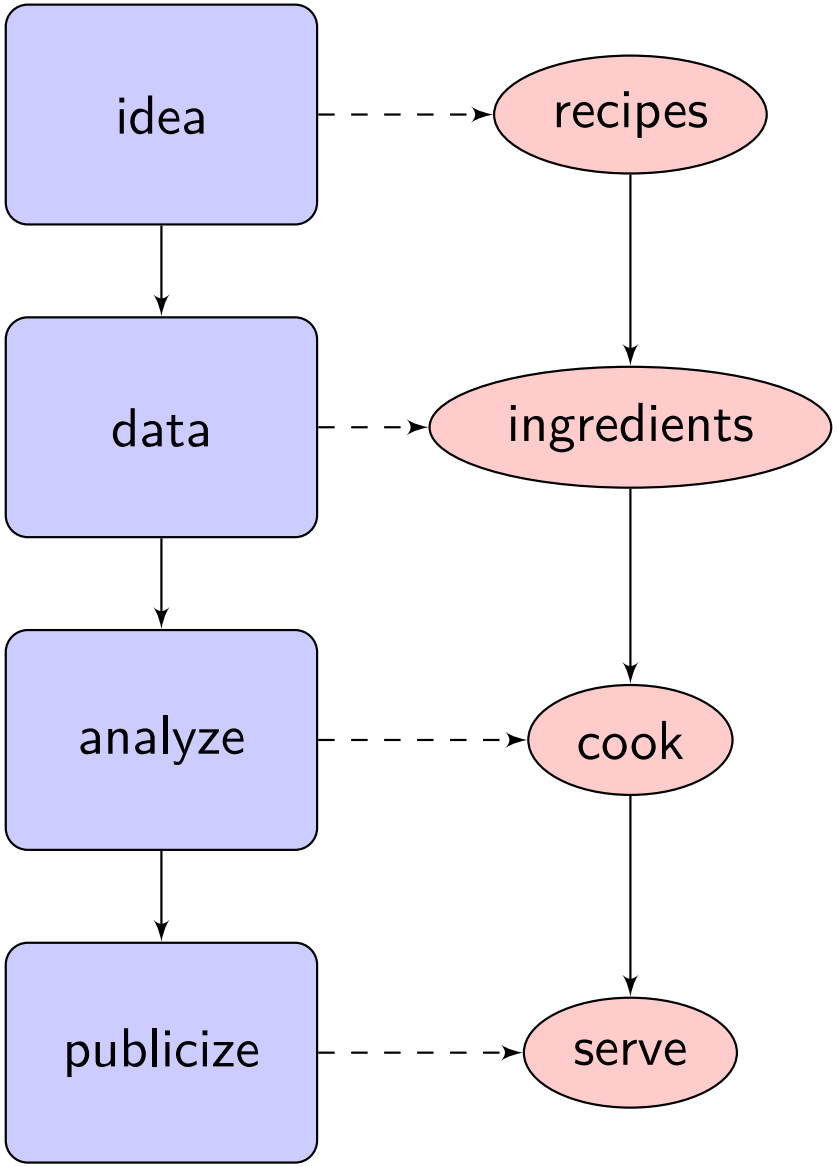
There are (I think) two major types of data specialists (which are not mutually exclusive)

- Programmers
- Data scientists/analysts

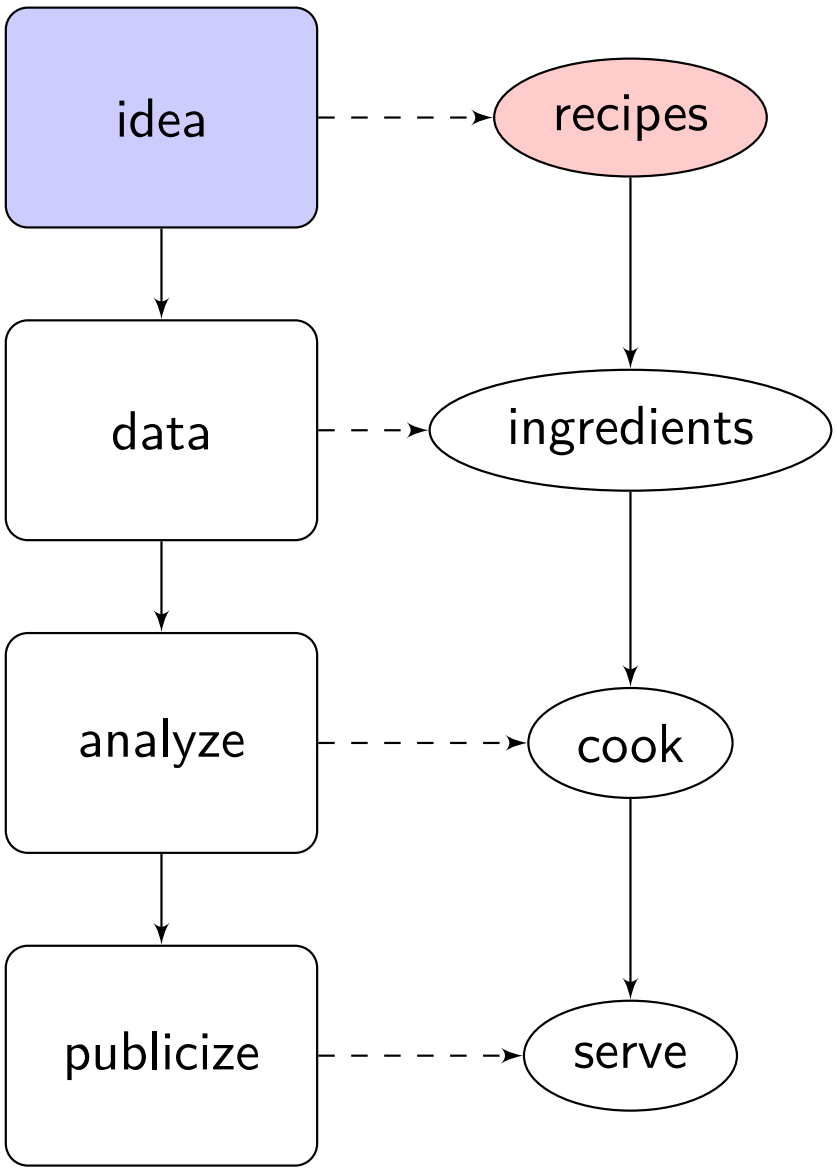
Data science is not only for engineers

- E.g., # of Economics Ph.D. holders finding jobs in tech companies (Airbnb, Amazon, Facebook, etc.) has been increasing

General workflow of empirical work



Step 1



"Good food begins with a good recipe."

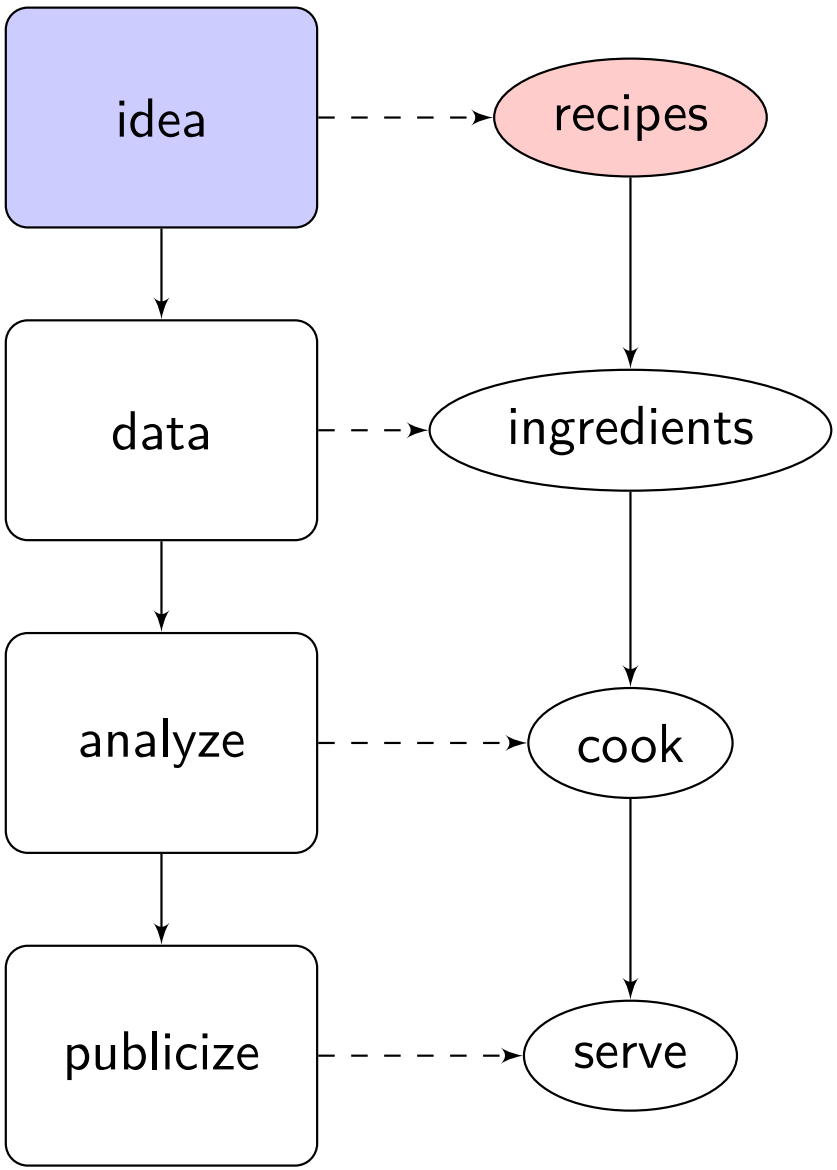
Goal: Get good ideas.

- 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 tool for storing, organizing, and sharing ideas and resources:

- Evernote
- Readcube
- Dropbox

Step 1



"Good food begins with a good recipe."

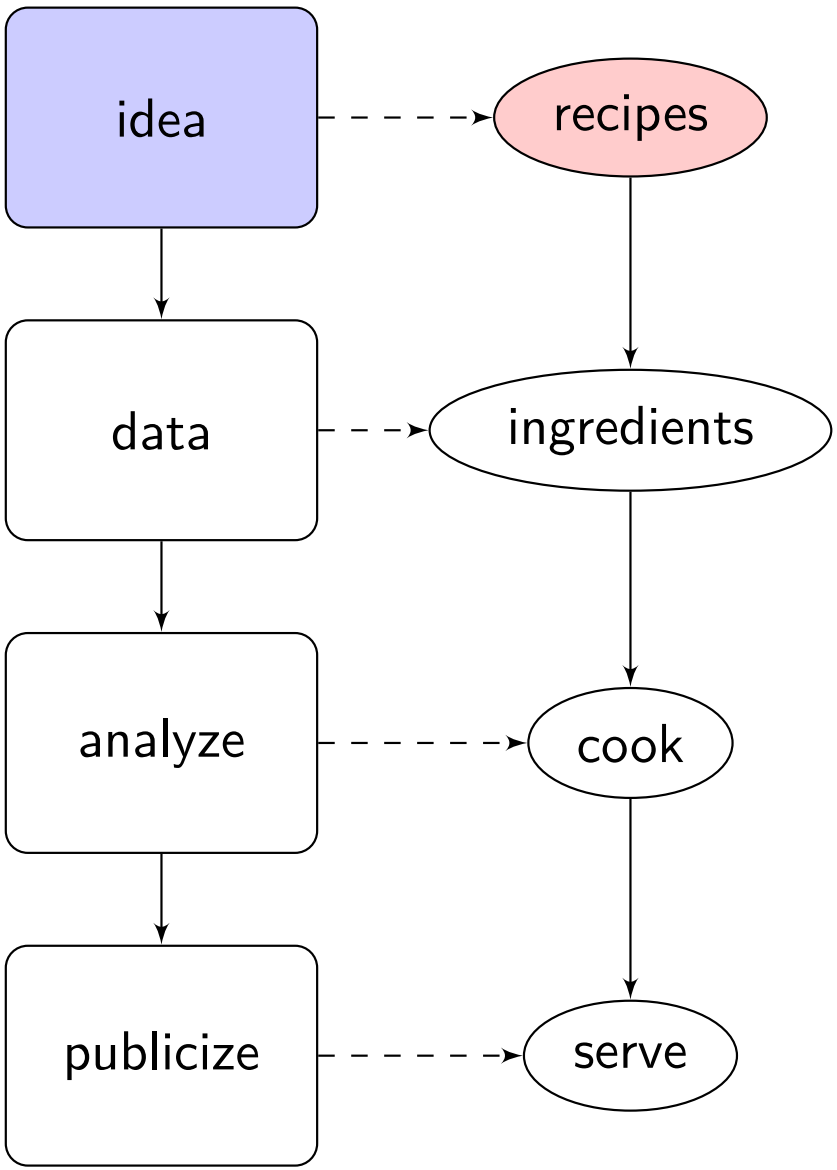
Goal: Get good ideas.

- 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 tool for storing, organizing, and sharing ideas and resources:

- Evernote
- Readcube
- Dropbox

Step 1



"Good food begins with a good recipe."

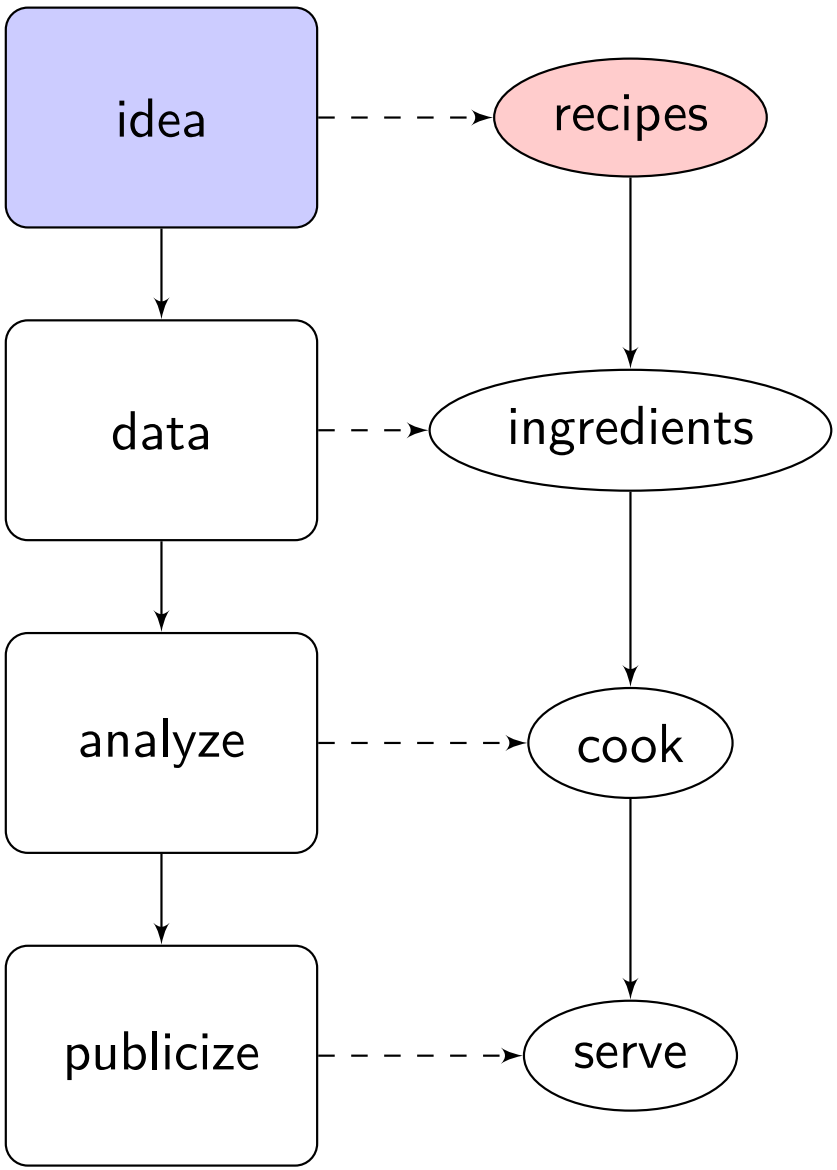
Goal: Get good ideas.

- 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 tool for storing, organizing, and sharing ideas and resources:

- Evernote
- Readcube
- Dropbox

Step 1



"Good food begins with a good recipe."

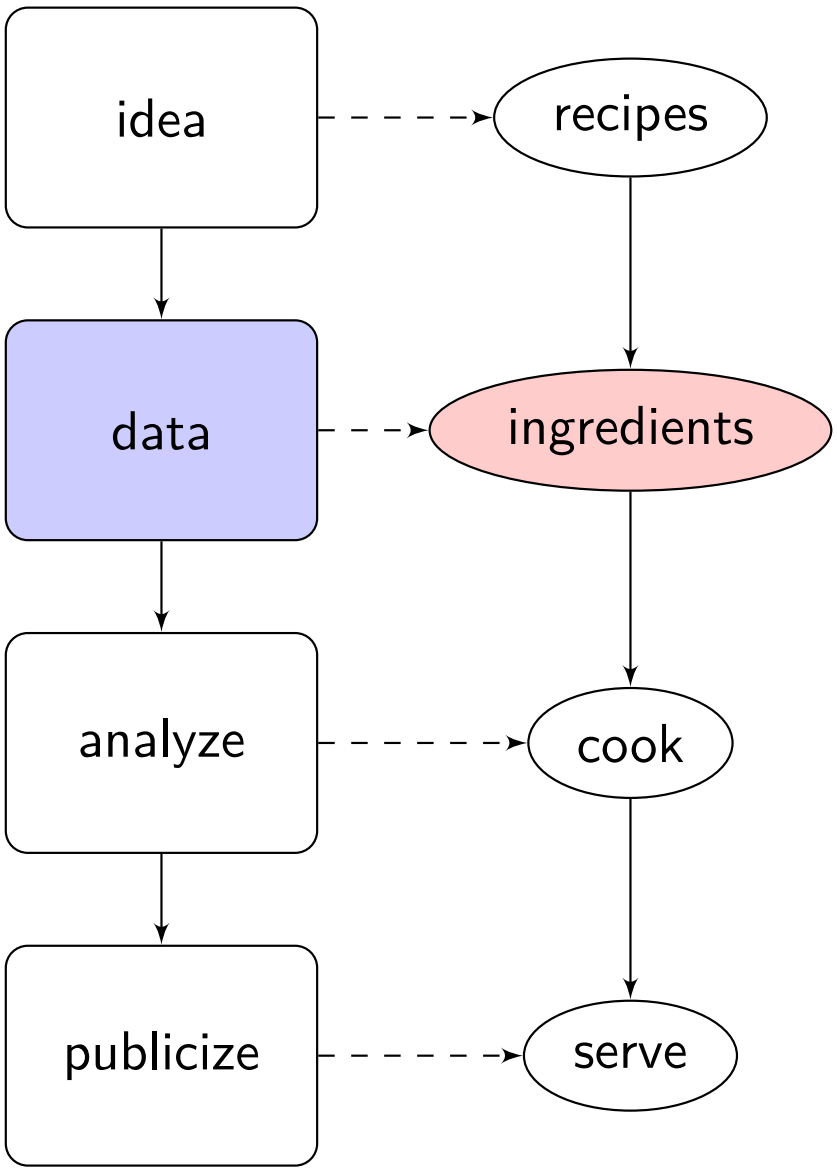
Goal: Get good ideas.

- 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 tool for storing, organizing, and sharing ideas and resources:

- Evernote
- Readcube
- Dropbox

Step 2



"Good food is made with good ingredients."

Goal: Collect good data and clean them properly.

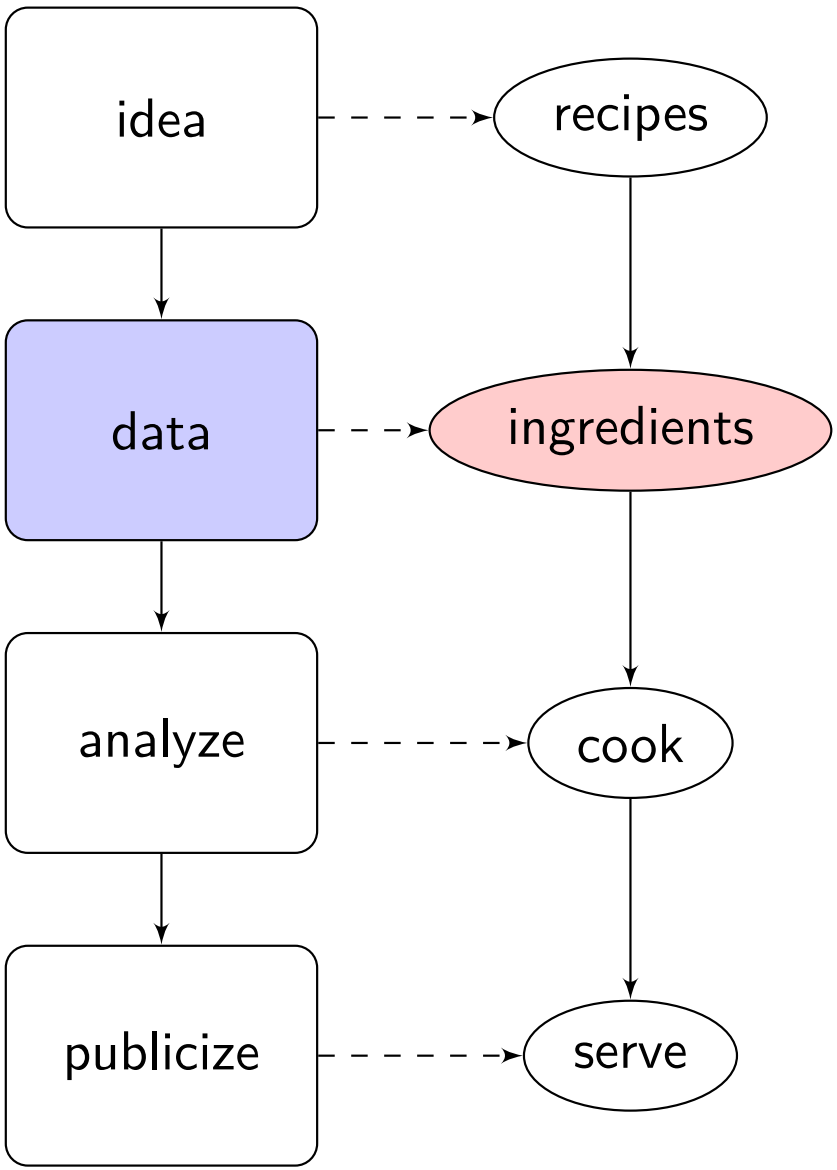
Useful tool for collecting and cleaning data:

- Python & R

Useful tool for storing and sharing data:

- Dropbox

Step 2



"Good food is made with good ingredients."

Goal: Collect good data and clean them properly.

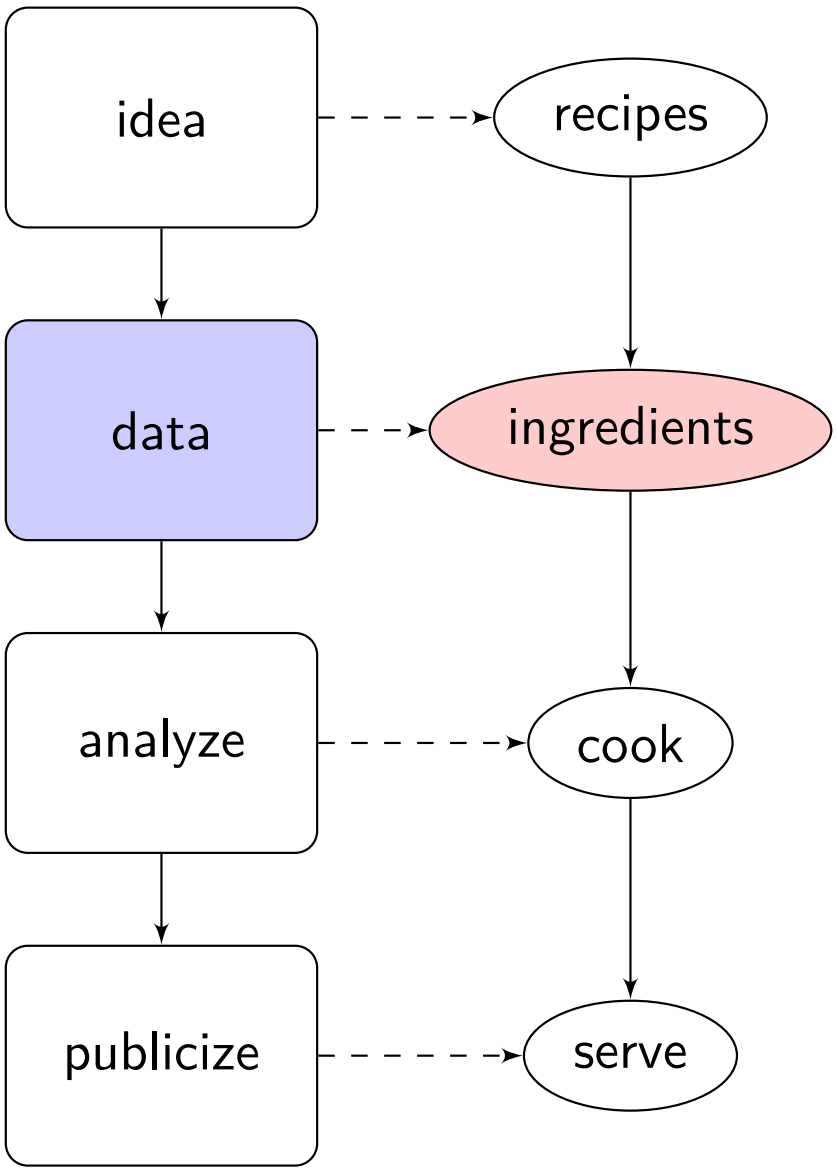
Useful tool for collecting and cleaning data:

- Python & R

Useful tool for storing and sharing data:

- Dropbox

Step 2



"Good food is made with good ingredients."

Goal: Collect good data and clean them properly.

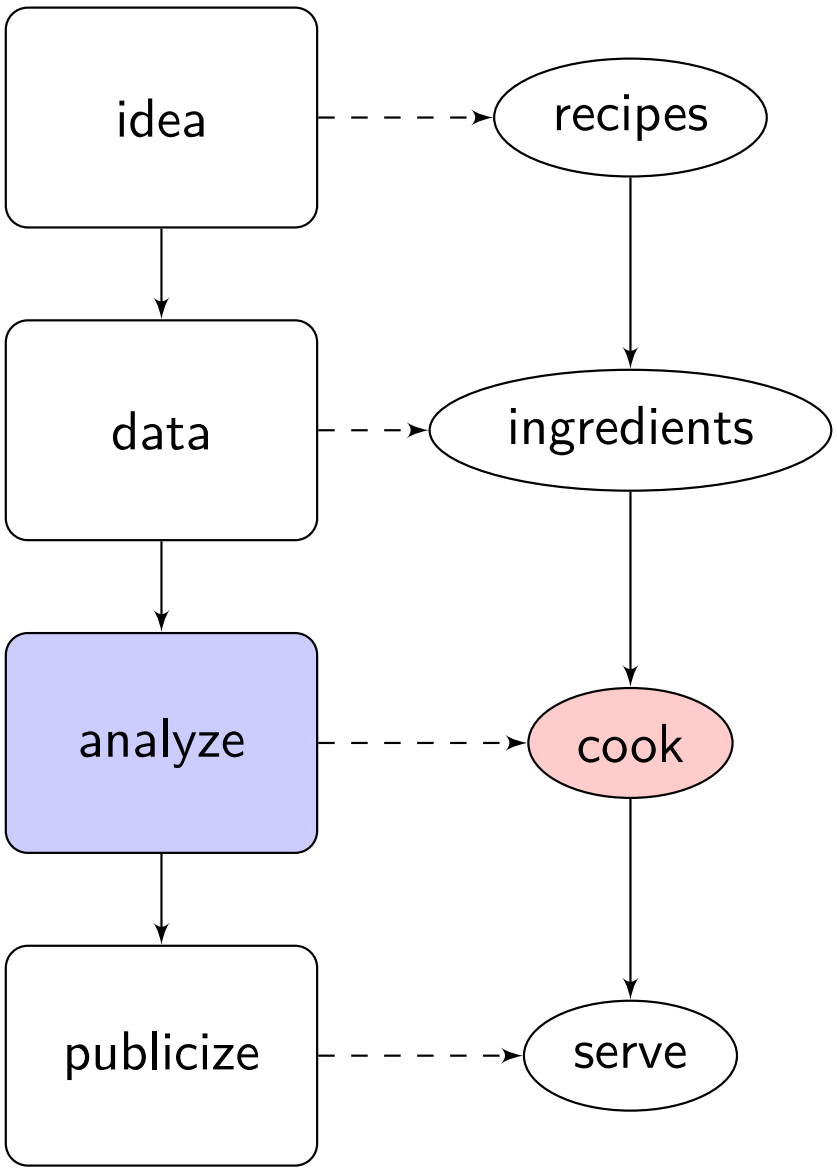
Useful tool for collecting and cleaning data:

- Python & R

Useful tool for storing and sharing data:

- Dropbox

Step 3



"Good food is cooked well."

Goal: Analyze data properly.

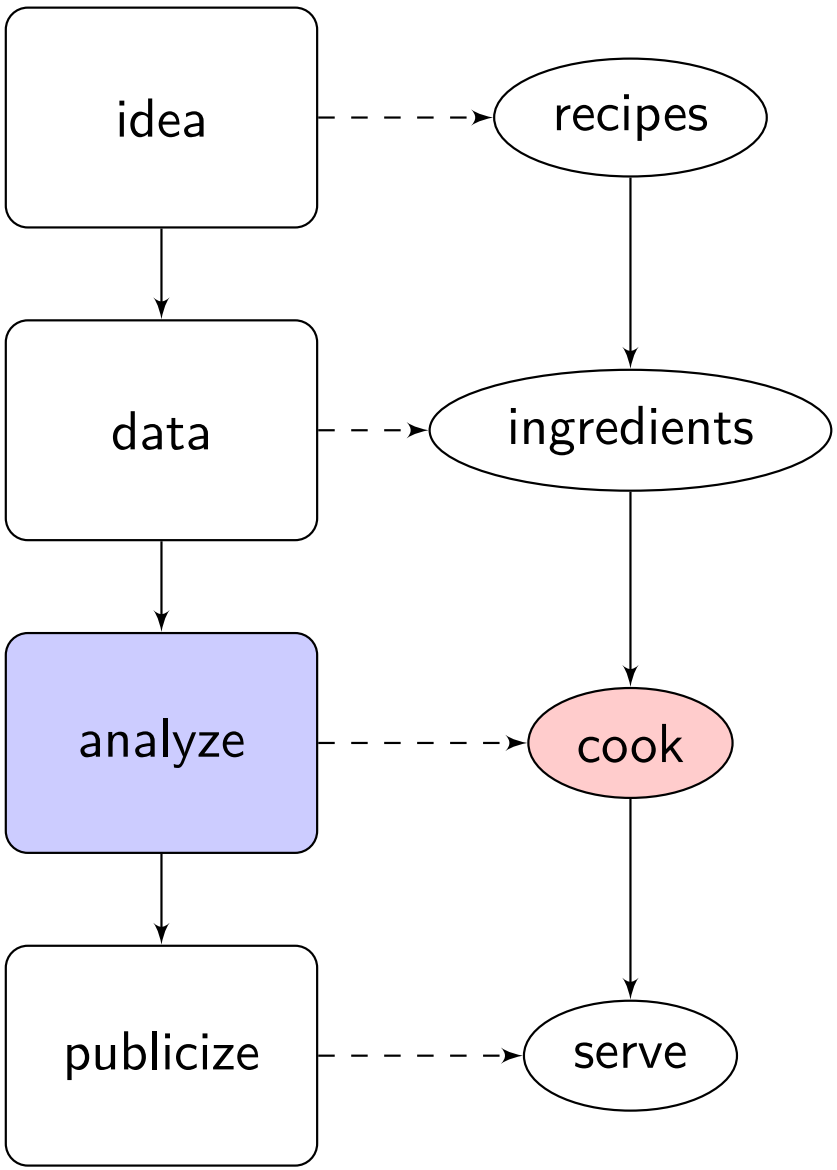
Useful tool for analyzing data:

- Python & R
- Stata

+ Knowledge in statistics and econometrics

(+ Knowledge in AI/machine learning/deep learning)

Step 3



"Good food is cooked well."

Goal: Analyze data properly.

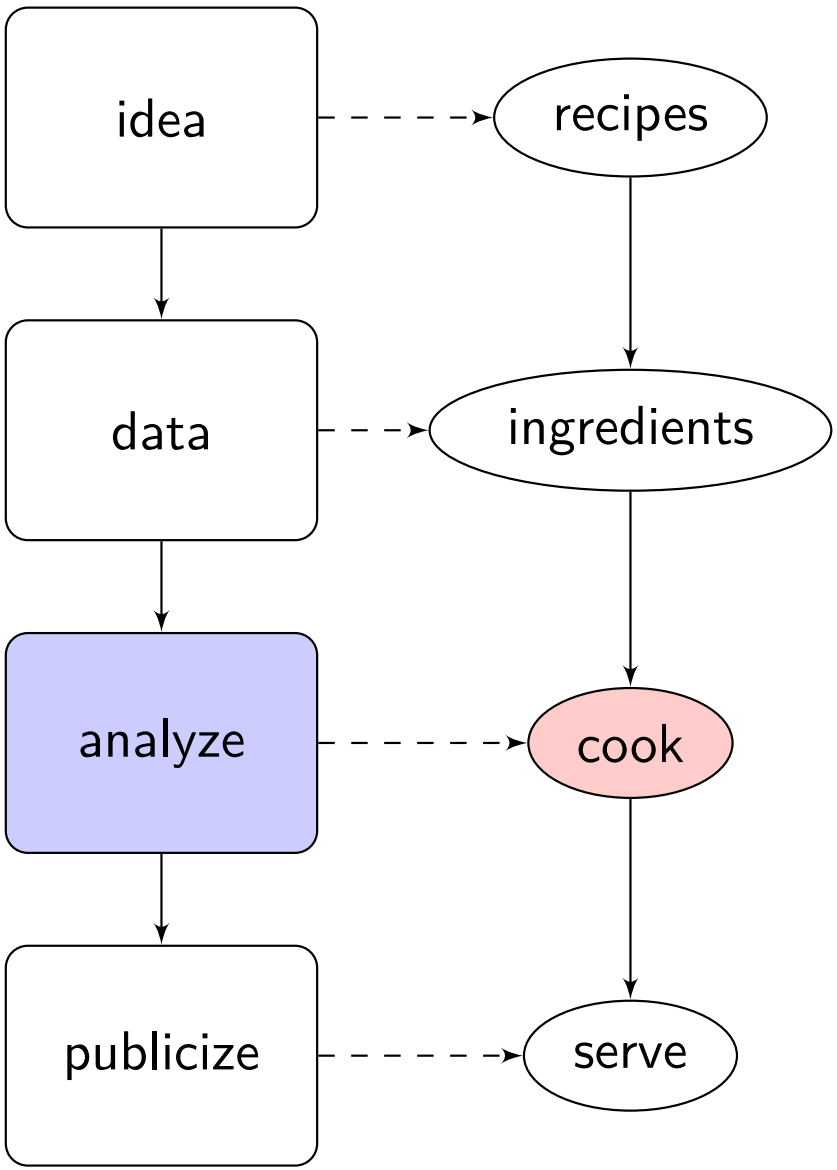
Useful tool for analyzing data:

- Python & R
- Stata

+ Knowledge in statistics and econometrics

(+ Knowledge in AI/machine learning/deep learning)

Step 3



"Good food is cooked well."

Goal: Analyze data properly.

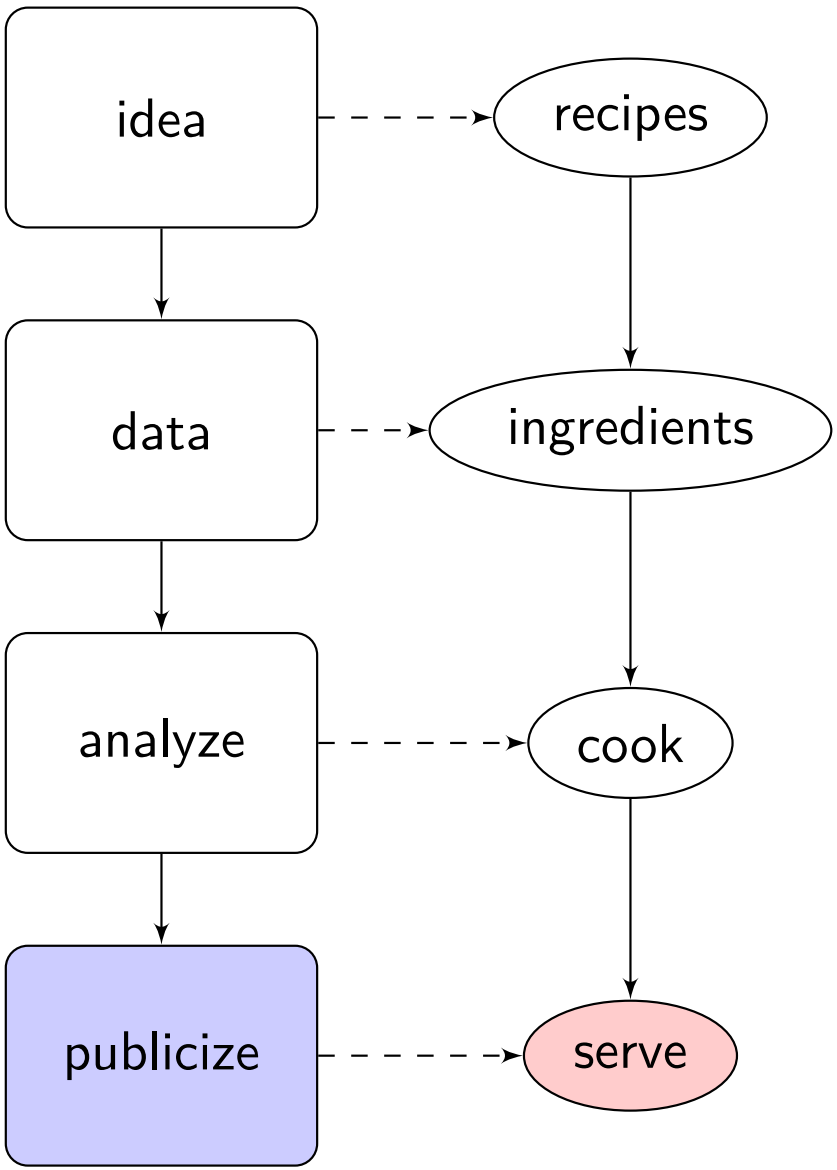
Useful tool for analyzing data:

- Python & R
- Stata

+ Knowledge in statistics and econometrics

(+ Knowledge in AI/machine learning/deep learning)

Step 4



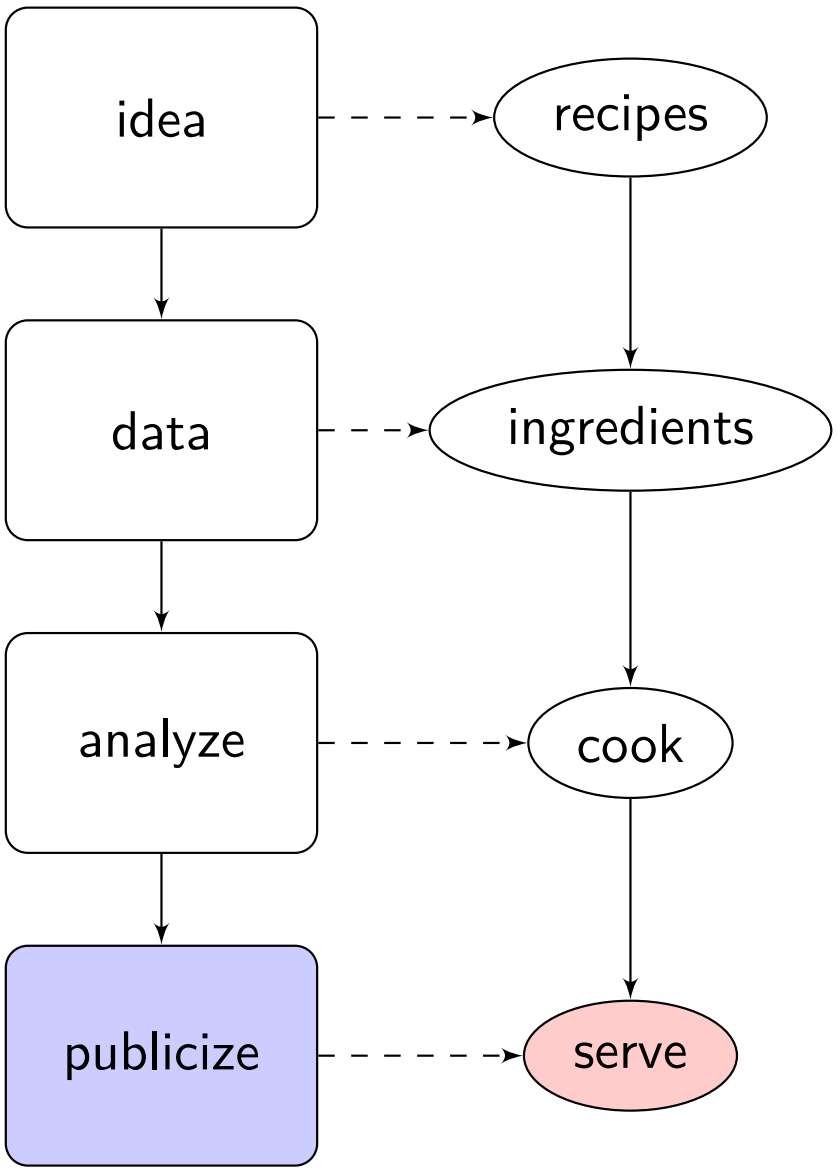
"Good food is served well."

Goal: Summarize results intuitively.

Useful tool for writing a paper and slides:

- TEX

Step 4



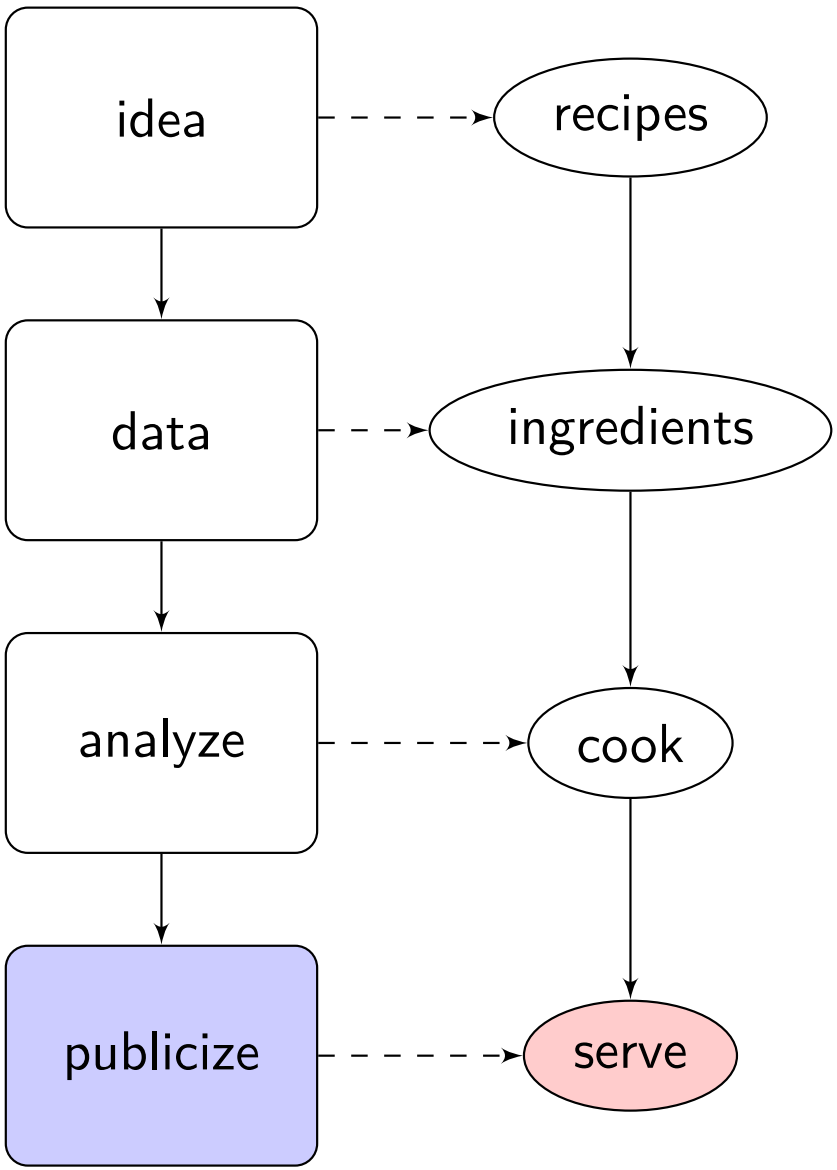
"Good food is served well."

Goal: Summarize results intuitively.

Useful tool for writing a paper and slides:

- TEX

Step 4



"Good food is served well."

Goal: Summarize results intuitively.

Useful tool for writing a paper and slides:

- TEX

Course plan

1. Idea [Recipes] (this lecture) (10%)
2. Making data [Ingredients] Python & R (50%)
3. Analyzing data [Cooking] Python & R (30%)
   - Making tables and figures, running regressions, etc.
4. Publicizing results [Serving] LaTeX (10%)

Course plan (cont.)

Each class consists of a lecture and coding exercises

Feel free to study together

BUT, you need to hand in **your own code** for assignments

1. Idea

Idea

Idea $\rightarrow$ often downplayed, but is 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 well and revisit them sometimes
- Read journal articles just to know what is going on and knows and unknowns

Read articles, but should not be overwhelmed by them

A bad example: The pile of papers I have read during my Ph.D.

How to store resources and data

Recommend: Store resources and data in cloud

Why cloud?

- Handy (easy to store, access, organize, and share)
- Hard to lose
- Save space

A nice tool for storing ideas & web resources

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

Other options (free plan): Google One (15GB), Amazon (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 has been getting more popular in academia

We start by Python, then move on to R

Focus more on Python (60%). Two reasons:

- Many of you may work in non-academic sectors after graduation
- There are some overlaps

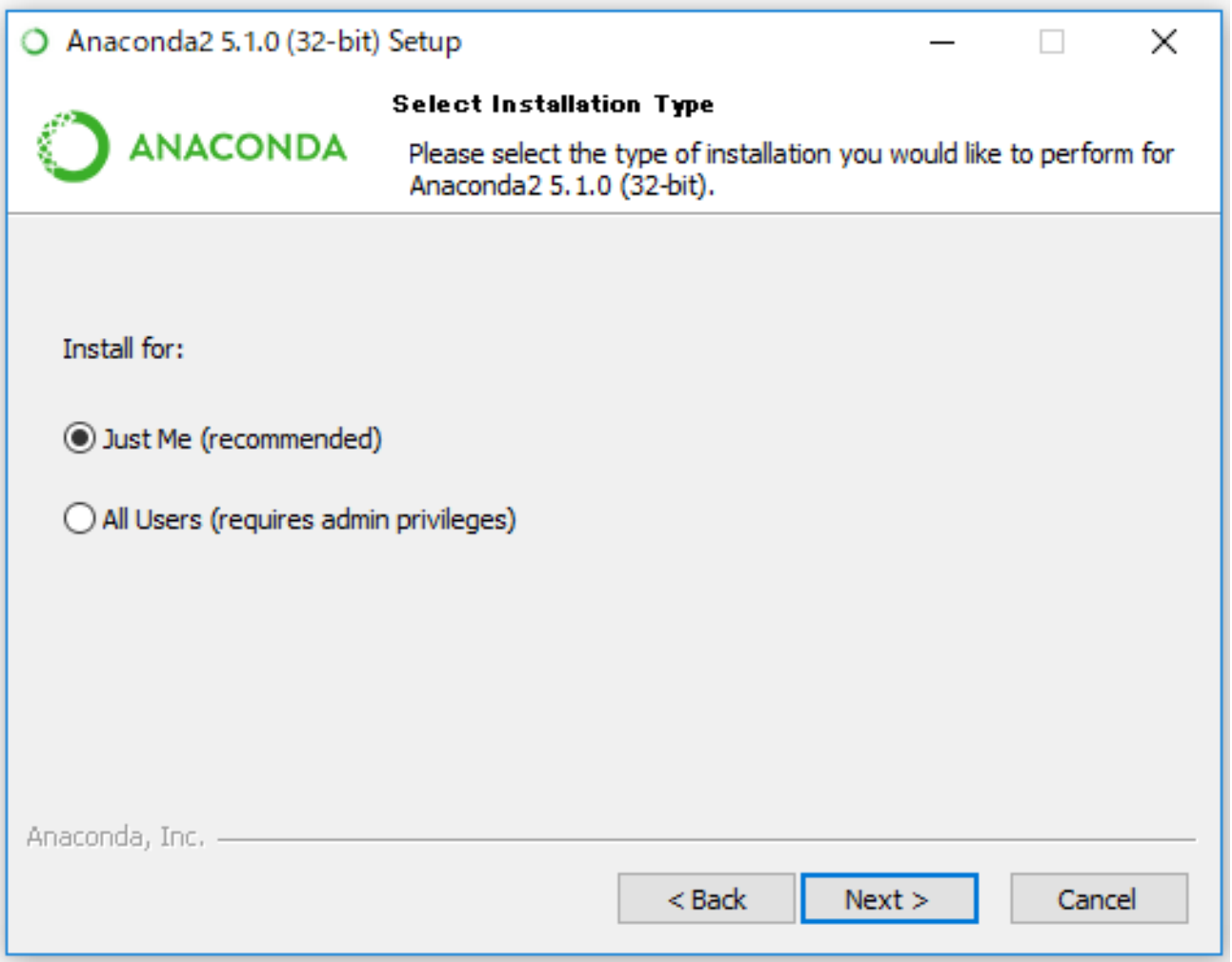
Install Jupyter Notebook (Python)

Install Jupyter Notebook using the Anaconda Distribution

Which Python version should I use?

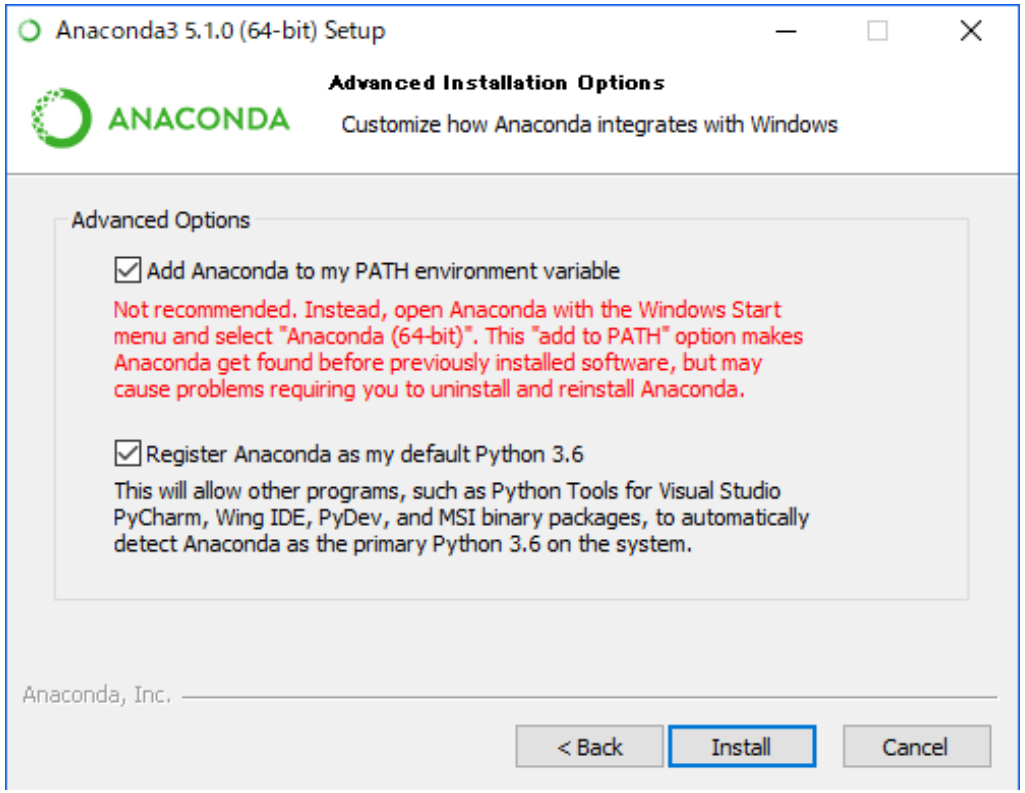
- "Python 2.x is legacy, Python 3.x is the present and future of the language."
- E.g., $3/2 = 1$ in Python 2.x but $3/2 = 1.5$ in Python 3.x
- Unless Python 2.x is required (e.g. ArcGIS Desktop), choose Python 3.x

Install Jupyter Notebook (Python) (cont.)



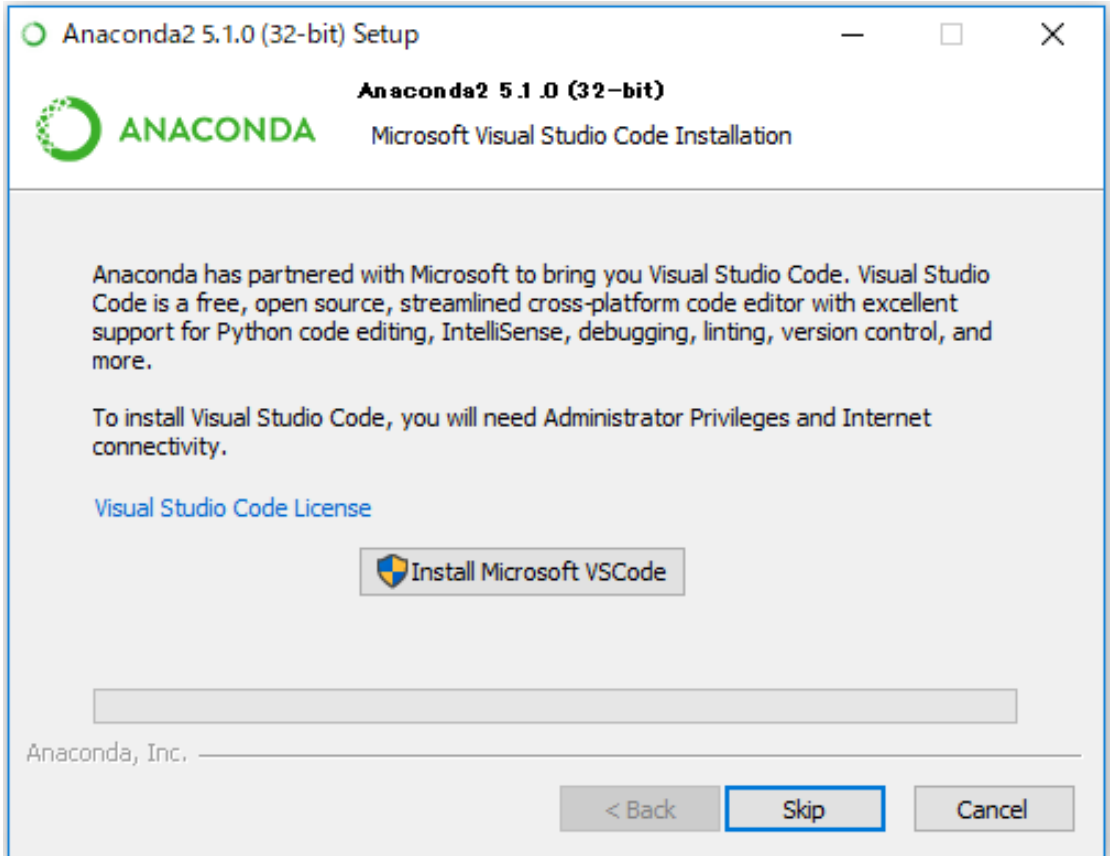
Choose "Just Me (recommended)"

Install Jupyter Notebook (Python) (cont.)



- 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

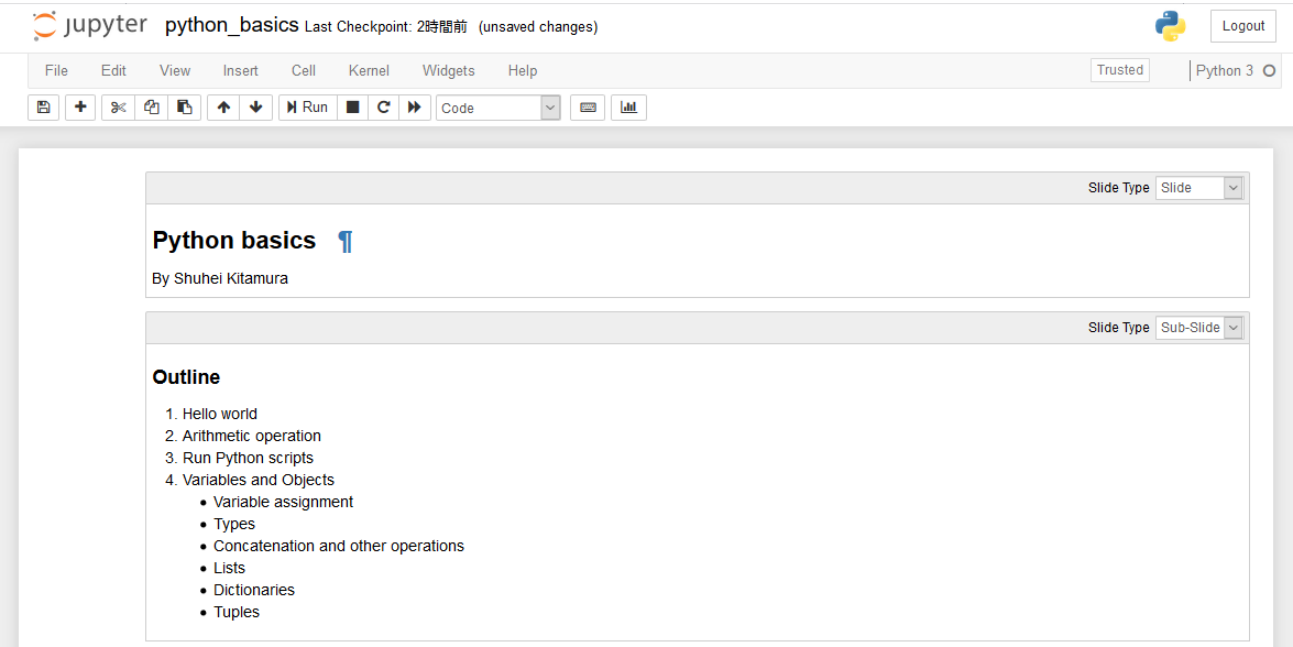
Install Jupyter Notebook (Python) (cont.)



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

- Other options: Sublime Text, Vim

Launch Jupyter Notebook



- In the command line, type

```
jupyter notebook
```

- Go to the local folder where you saved downloaded files
- Click python_basics_1.ipynb. A screen like the above picture shows up

Install JupyterLab (not required)

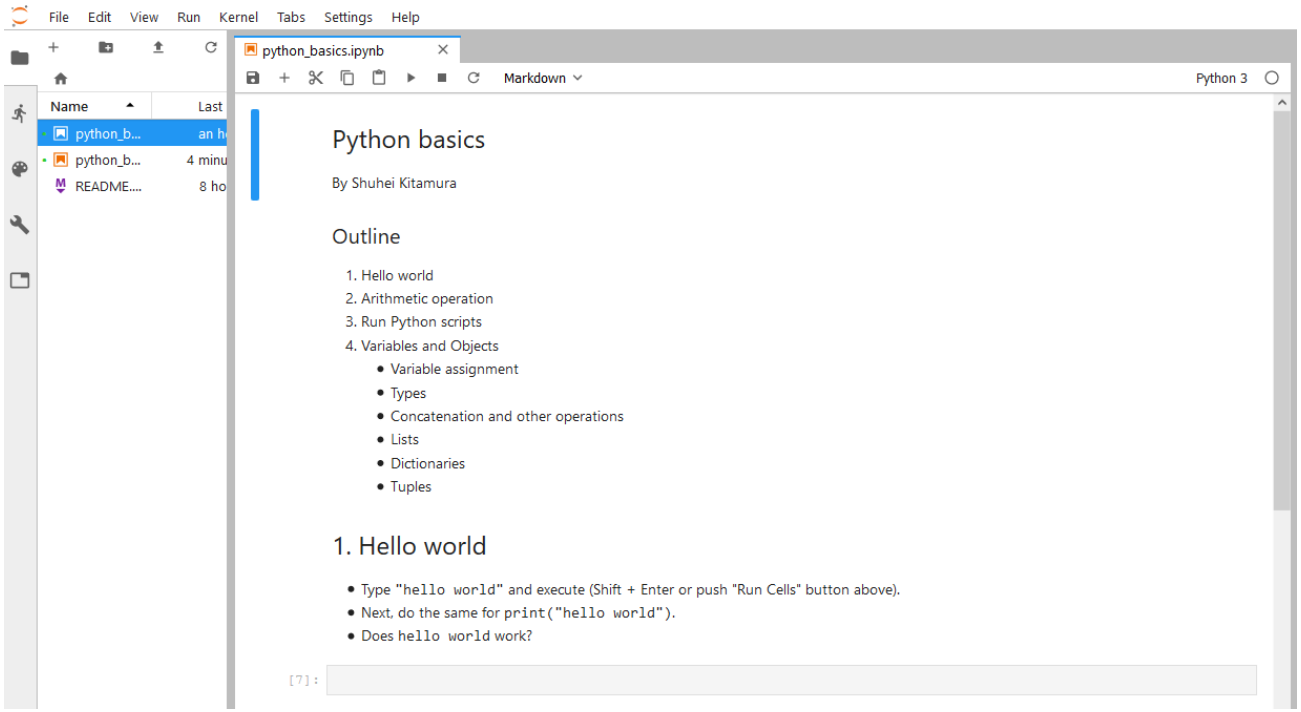
- Type "cmd" in the search box at the bottom of your screen. Command Prompt pops up
- Then, in the command line, type

```
conda install -c conda-forge jupyterlab
```

- For other OSs, see this guide
- To update all packages and modules, type

```
conda update --all
```

Launch JupyterLab (not required)



- In the command line, type

```
jupyter lab
```

- Go to the local folder where you saved downloaded files
- Click python_basics_1.ipynb. A screen like the above picture shows up

Summary

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

References: Python

Online sources:

- Lectures in Quantitative Economics: (English)
- DataCamp (some courses are free): (English)
- An Introduction to Python for Economists: (English)
- Matsuo Lab at U of Tokyo: (Japanese)

Books:

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