

數據科學方法

Python工具箱

Python 3.5+

- [Official python](#)
- [Anaconda](#) or [miniconda](#)
- [docker datascience-notebook](#)

Modules

- numpy and scipy
- matplotlib, Seaborn, and Bokeh (perhaps)
- Pandas and StatsModels
- scikit-learn
- jupyter notebook

Miniconda

1. Download and install miniconda
2. Create a virtual environment:
`conda create -n datascience python=3.5`
3. Activate the above virtual environment:
`(source) activate datascience`
4. Install required packages:
`pip3 install xxx`

Miniconda

- Pros
 - Optimized version of python for numerical computations
 - Easy to use
- Cons
 - Maintained by commercial company

Docker

- Light-weight and fast virtualization technique
- Focus on applications instead of operation system
- Deployable on different platforms

Basic ideas for docker

- Image file: a file used to define and run a container
- Container: a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

Docker datascience-notebook

1. Download and install docker
2. Get the datascience-notebook image file:
`docker pull jupyter/datascience-notebook`
3. Create and run a container:
`docker run -d -p 8888:8888 jupyter/
datascience-notebook`

Docker datascience–notebook

- [Further commands](#)
- Pros
 - sandbox
- Cons
 - fat
 - may require more knowledge for customization