Data Mining & Machine Learning

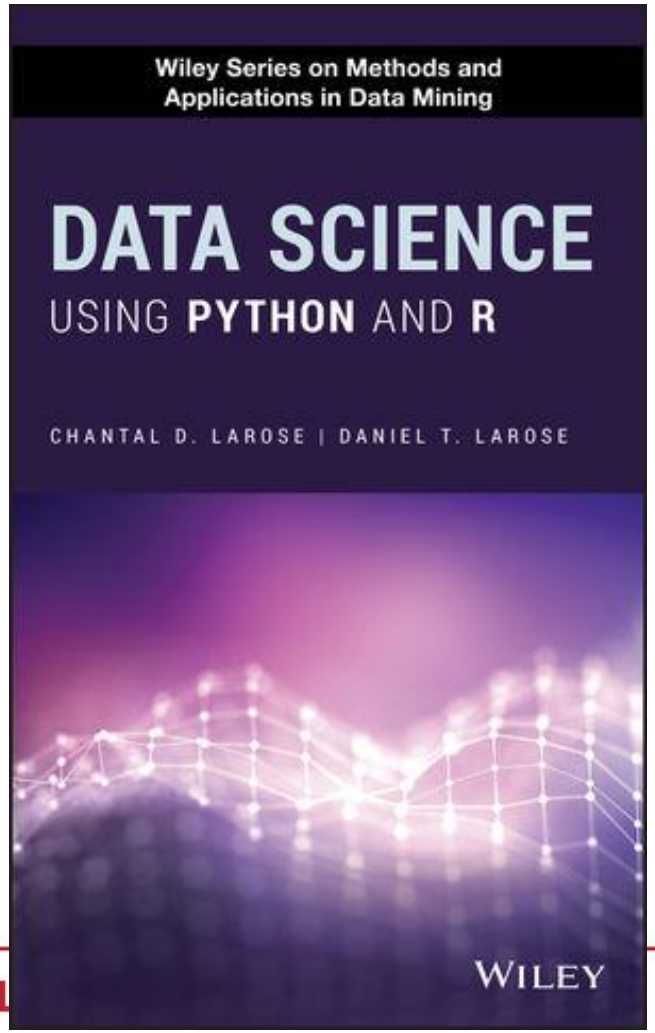
Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH
College of Computing

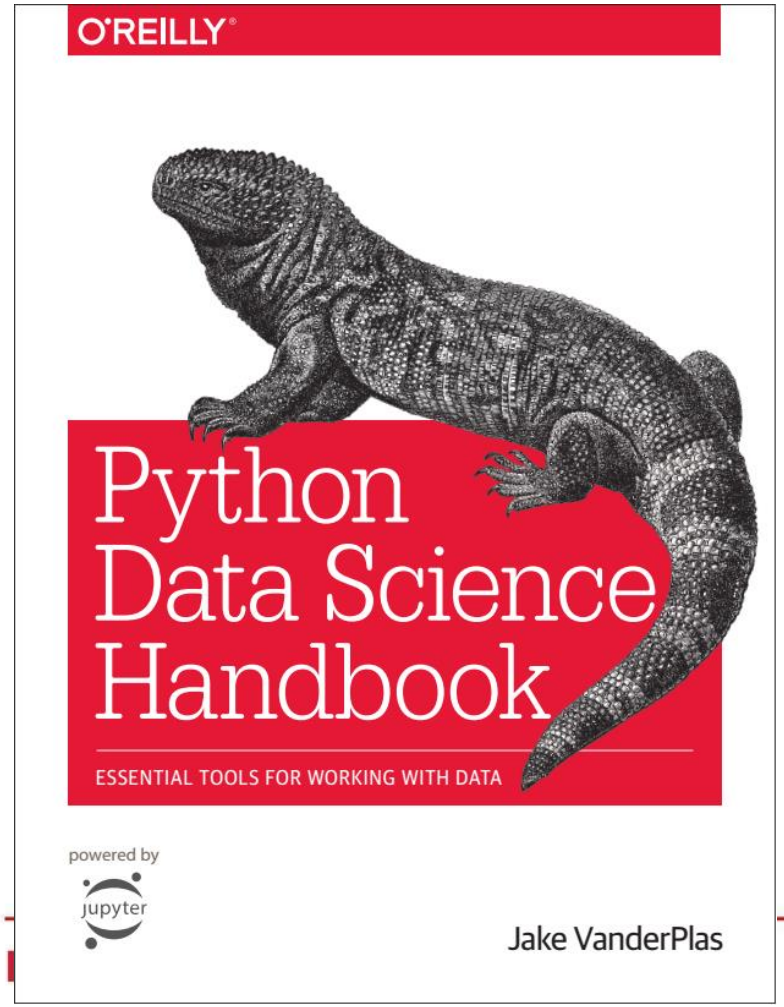
Python for Data Science: Books

- General Books: Coding by R and Python



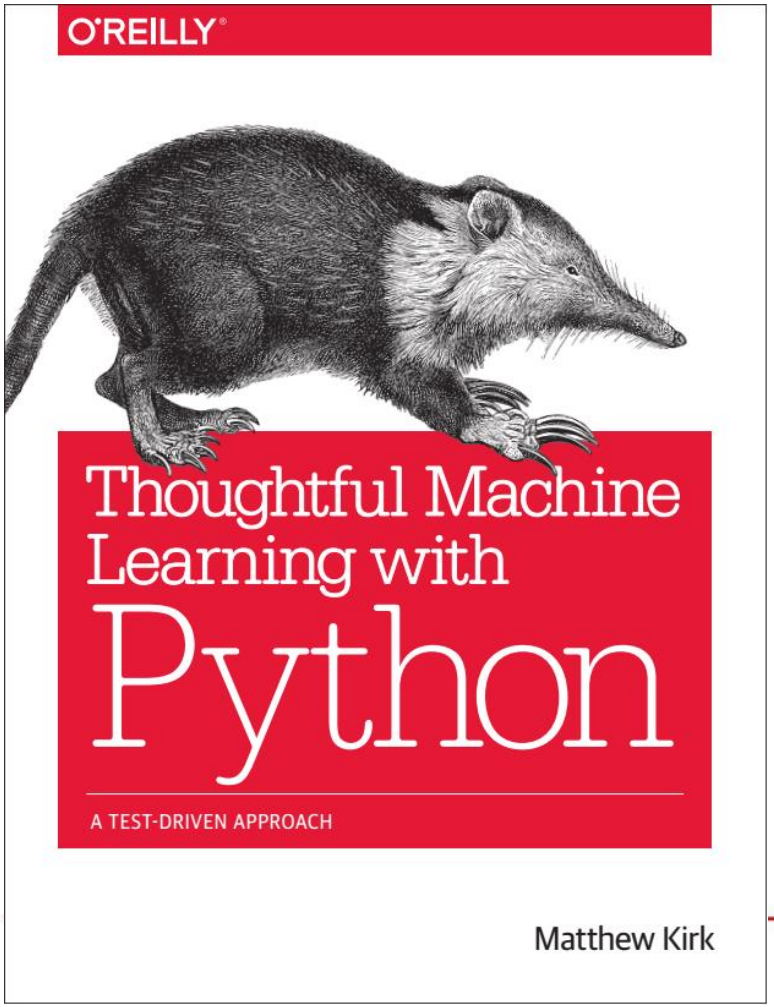
Python for Data Science: Books

- General Books: Knowledge + Python Practice



Python for Data Science: Books

- General Books: Knowledge + Python Practice



Python for Data Science: Videos

- Youtube Videos https://www.youtube.com/watch?v=OGxgnH8y2NM&list=PLQVvvaa0QuDfKTOs3Keq_kaG2P55YRn5v&index=1

Python

- Anaconda and Jupyter Notebook
- Data Science Libraries
- Python for Data Science

Python

- Anaconda and Jupyter Notebook
- Data Science Libraries
- Python for Data Science

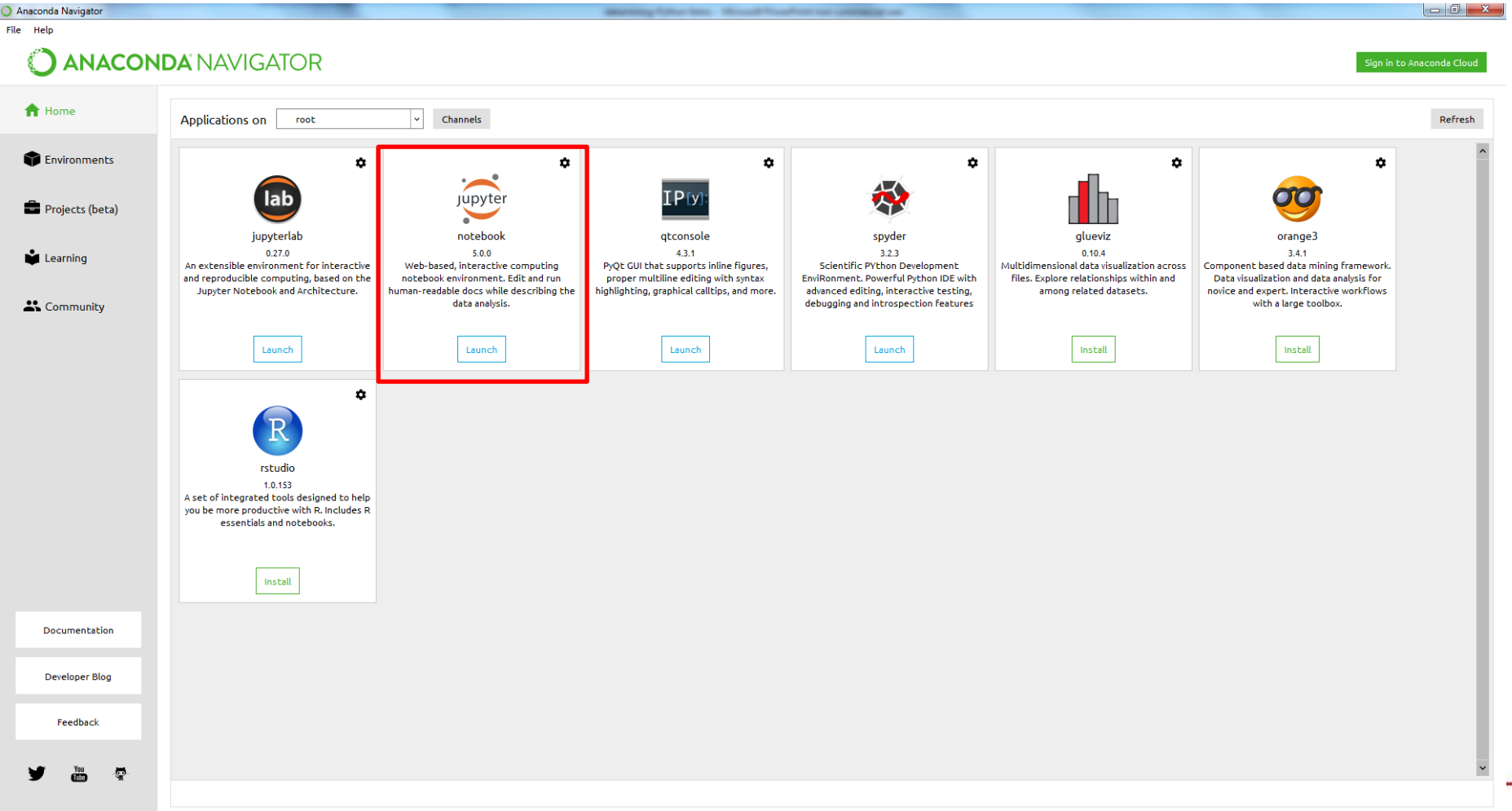
Python

- In the last few years there is an increasing community that creates Data Mining tools in Python
- The major reasons why
  - It is much more convenient in coding
  - There are multiple libraries
  - They do support big data and multiple data preprocessing

Installing Python

- Installing libraries in Python is complex, so you should download the Anaconda Scientific Python distribution which will install most of the libraries that we will use.
  - There are two versions, Python 2.7 and Python 3.0 and they are not compatible. We will use Python 3.0

The Anaconda Navigator



Starting iPython notebook

Anaconda - ipython notebook

```
Added C:\Anaconda3 and C:\Anaconda3\Scripts to PATH.

C:\Anaconda3>ipython notebook
[I 20:22:58.688 NotebookApp] Using MathJax from CDN: https://cdn.mathjax.org/mat
hjax/latest/MathJax.js
[I 20:22:58.824 NotebookApp] Serving notebooks from local directory: C:\Anaconda
3
[I 20:22:58.824 NotebookApp] 0 active kernels
[I 20:22:58.824 NotebookApp] The IPython Notebook is running at: http://localhos
t:8888/
[I 20:22:58.824 NotebookApp] Use Control-C to stop this server and shut down all
 kernels (twice to skip confirmation).
```

Home

localhost:8888/tree

jupyter

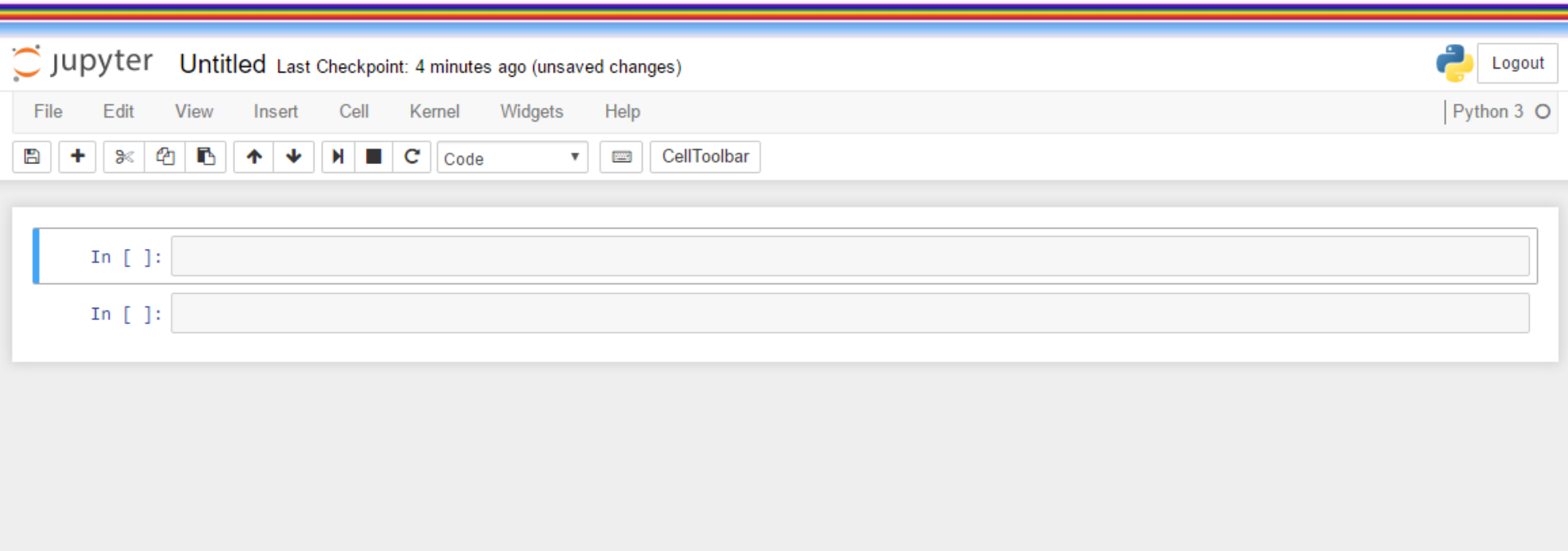
Files | Running | Clusters

Select items to perform actions on them.

Upload | New

- DLLs
- Doc
- Examples
- Lib
- Library
- Menu
- Scripts
- Tools
- conda-meta
- envs
- include
- info
- libs
- licenses
- node-webkit
- pkgs
- share
- tcl


jupyter
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
/ Desktop / Untitled Folder
..
Untitled.ipynb
Text File
Folder
Terminals Unavailable
Notebooks
Python 3



Click + to add a block
You can run codes block by block
Or, you can run the codes for all the blocks

The variables in the previous block can also be used in the following blocks

Python

- Anaconda and Jupyter Notebook
- Data Science Libraries
- Python for Data Science

Data Science Libraries in Python

Many popular Python toolboxes/libraries:

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Visualization libraries

- matplotlib
- Seaborn

Data Science Libraries in Python

NumPy:

- introduces objects for multidimensional arrays, vectors and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects
- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance
- many other python libraries are built on NumPy

Data Science Libraries in Python

Pandas:

- adds data structures (data frame) and tools designed to work with table-like data
- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.
- allows handling missing data

Data Science Libraries in Python

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more
- part of SciPy Stack
- built on NumPy
- SciPy and NumPy are usually used for matrix-based operations, such as matrix factorization

Data Science Libraries in Python

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.

- built on NumPy, SciPy and matplotlib

Data Science Libraries in Python

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization

Data Science Libraries in Python

Seaborn:

- based on matplotlib
- provides high level interface for drawing attractive statistical graphics
- Similar (in style) to the popular ggplot2 library in R

Python

- Anaconda and Jupyter Notebook
- Data Science Libraries
- Python for Data Science

Python for Data Science

- Data Manipulation by Python
  01. Data Manipulation with Pandas.ipynb
- Python for Data Preprocessing
  02. Data Preprocessing 1.ipynb