

Python for Data Analysis and Visualization

Professor Widom's Instructional Odyssey

www.professorwidom.org



Python

- Very popular general-purpose programming language
- Used from introductory programming courses to production systems

Python Features

- Dynamically typed
(rather than statically typed like Java or C/C++)
- Interpreted
(rather than compiled like Java or C/C++)

Python programs are comparatively...

- + Quicker to write
- + Shorter
- More error-prone
- Slower to run

Python for Data

- Fairly easy to read/write/process data using standard features
- Plus special packages for...
 - Numerical and statistical manipulations
NumPy
 - Visualization (“plotting”)
matplotlib
 - Relational database like capabilities
pandas (dataframes)
 - Machine learning
scikit-learn

Data Sets

Europe Temperatures

Cities: city, country, latitude, longitude, temperature

Countries: country, population, EU, coastline

2010 World Cup

Teams: team, ranking, games, wins, draws, losses, goalsFor, goalsAgainst, yellowCards, redCards

Players: surname, team, position, minutes, shots, passes, tackles, saves

Jupyter Notebooks

(formerly iPython notebooks)

- Modeled after “laboratory notebooks”
- In one notebook can combine text boxes (“markdown”) with boxes containing executable code in a wide variety of languages
- Can run/re-run boxes (cells) individually, or run/re-run entire notebook

Rapid adoption in many sectors

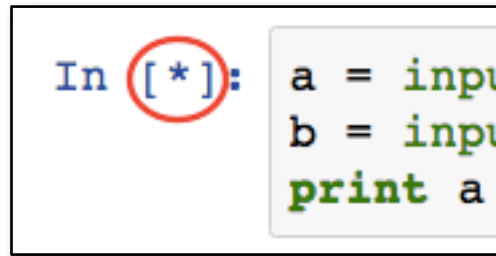
Jupyter Notebooks

- Can download to your computer (recommend *Anaconda*) but no one-button download yet
- We will use notebooks in the cloud, courtesy *Instabase*, *Google Cloud*, and *Amazon Web Services*
- Either way, notebooks run in a web browser

Jupyter Notebook Hints

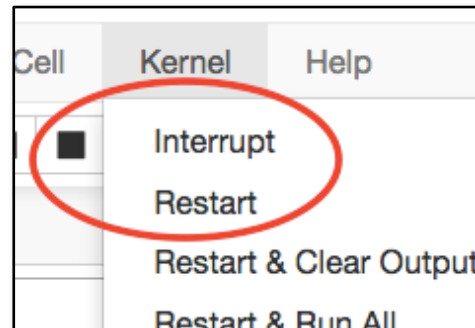
To execute a cell, click inside the box then type **shift-enter** (or **shift-return**)

If nothing happens, some cell is probably still executing



A screenshot of a Jupyter Notebook cell. On the left, the prompt 'In [*]:' is shown, with the asterisk and brackets '[*]' circled in red. To the right of the prompt, the code is displayed: `a = input`, `b = input`, and `print a`.

Try Kernel > Interrupt or Kernel > Restart



Agenda

1. Python basics
2. Data manipulation
3. Plotting
4. Pandas

(more in Machine Learning, Data Mining,
Network Analysis modules)

Plenty of your turn!

For help while working with Python:

Tutorials and help pages

(website Course Materials)

➤ **Web search**