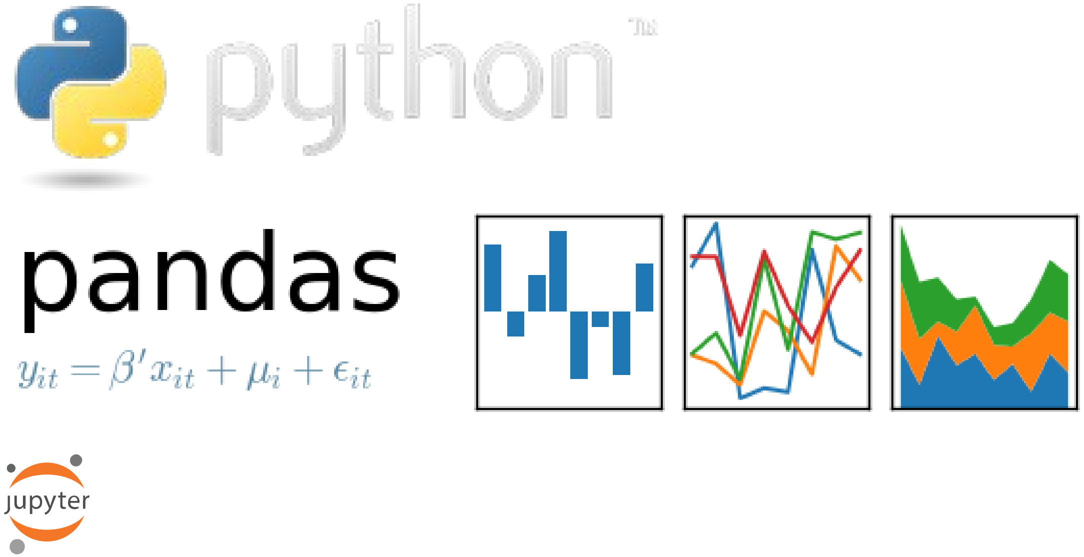
Analysis with Jupyter and Pandas



24 Oct. 2018

Paul Rodgers @prodg
Cloud RI

Who am I?

- Sr. Data Analyst at Virgin Pulse
- Early adopter
- Booted a PDP-8 and a Raspberry Pi (a few months apart)
- Generalist
    - Database architect for early web app (MySQL/mod_perl)
    - ETL veteran (without SSIS or Informatica etc.)
    - Operations automation
- Evangelist

What is Jupyter?

- Interactive code execution environment
- Tells a story
  - Allows the use of data, code and rich content
  - Enables the author to create a narrative
  - Engages the audience
  - Increases comprehension
  - Memorialize all aspects of the project

Who's using Jupyter?

- Academics
  - Paul Romer, Nobel Economist 2018
- Journalists
  - Los Angeles Times Data Desk (github.com/datadesk)
- Data Scientists
- Netflix!!
  - Papermill, nteract, Commuter

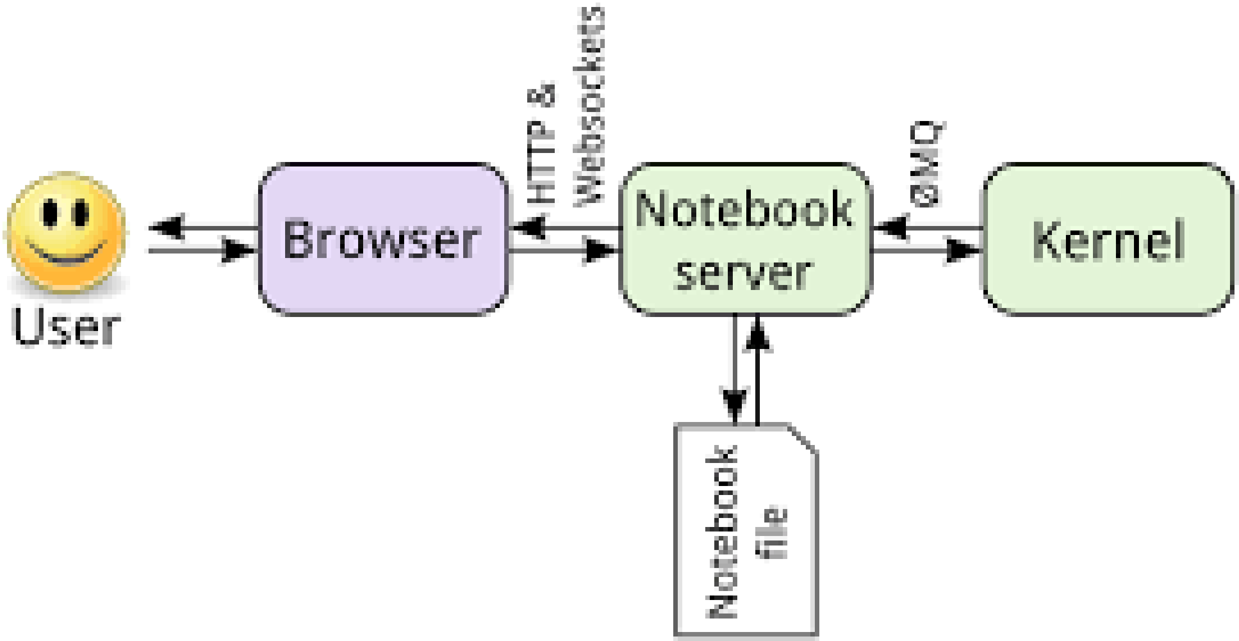
What is Pandas?

- Programmable two dimensional tabular data management tool (Excel optimized beyond description)
- Similar to R dataframes
- Leverages Numpy (fast array math)
- Top notch CSV importer
- RAM based
- Rich data manipulation, categorization and period tools
- Database style joins

Architecture

- Kernel (Ipython for our demonstration)
- Messaging with ZeroMQ
- Webserver – Tornado

Architecture



Who's using Pandas?

- Financial Analysts (Pandas was born here)
  - Time series and period savvy
- Data Scientists
- Me and (hopefully) you!

This environment

- Python 3.7.0
- Pandas 0.23.4
- Jupyter 4.4.0
- Ipython 6.5.0
- Cookiecutter
  - https://github.com/drivendata/cookiecutter-data-science
- Simple-salesforce 0.74.2

Basic Data Importing

- IPython Magics
  - %history
- Shell interaction
  - ! cmd execution
  - %% shell command ‘stack’
- Shell output assignment
-

Passing data to/from the shell

- Assignment to python variable
- Passing python output to the shell
  - {variable}
  - {command result}