SciPy2016

ANALYZING AND MANIPULATING DATA WITH PANDAS

July 12th, 2016

Jonathan Rocher (@jonrocher)
Principal Software Architect, KBI-Biopharma

KBI BIOPHARMA

Thanks!

Pandas

- Started by Wes McKinney in 2009.
- Emerged from the finance industry. Motivated by the toolbox in R for manipulating data easily. But Python is a better language…
- First public release is 0.3 in Feb 2011.
- Grown and maintained by a huge community now, headed by Jeff Reback.
- Last release is 0.18.1 in May 2016.
- Open source (BSD).
- Has become a **corner stone of the SciPy ecosystem** for all things data!

Pandas' mission

"To provide high-performance, easy-to-use data structures and data analysis tools [in Python]."

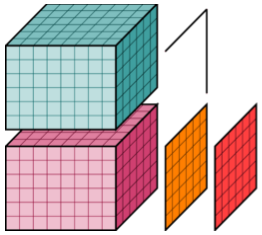
Easy-to-use and performant:

- **Self-describing** data structures to understand, explore and clean the data : 1D, 2D, 3D.
- Data loaders to/from common file formats (CSV, Excel json, SQL, SAS, Stata, …).
- Plotting functions to visualize the data. See `Seaborn`, `Bokeh`… for more.
- Basic statistical analysis tools. See `statsmodels` and `sklearn`+`sklean-pandas` for more.

KBI BIOPHARMA

Pandas' ecosystem growing quickly

GeoPandas



xarray

Plotly

Vincent

Bokeh

Dask

code. ENTHOUGHT .com

sklearn-pandas

Seaborn

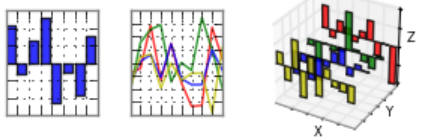
Altair

statsmodels

IP[y]: IPython
Interactive Computing

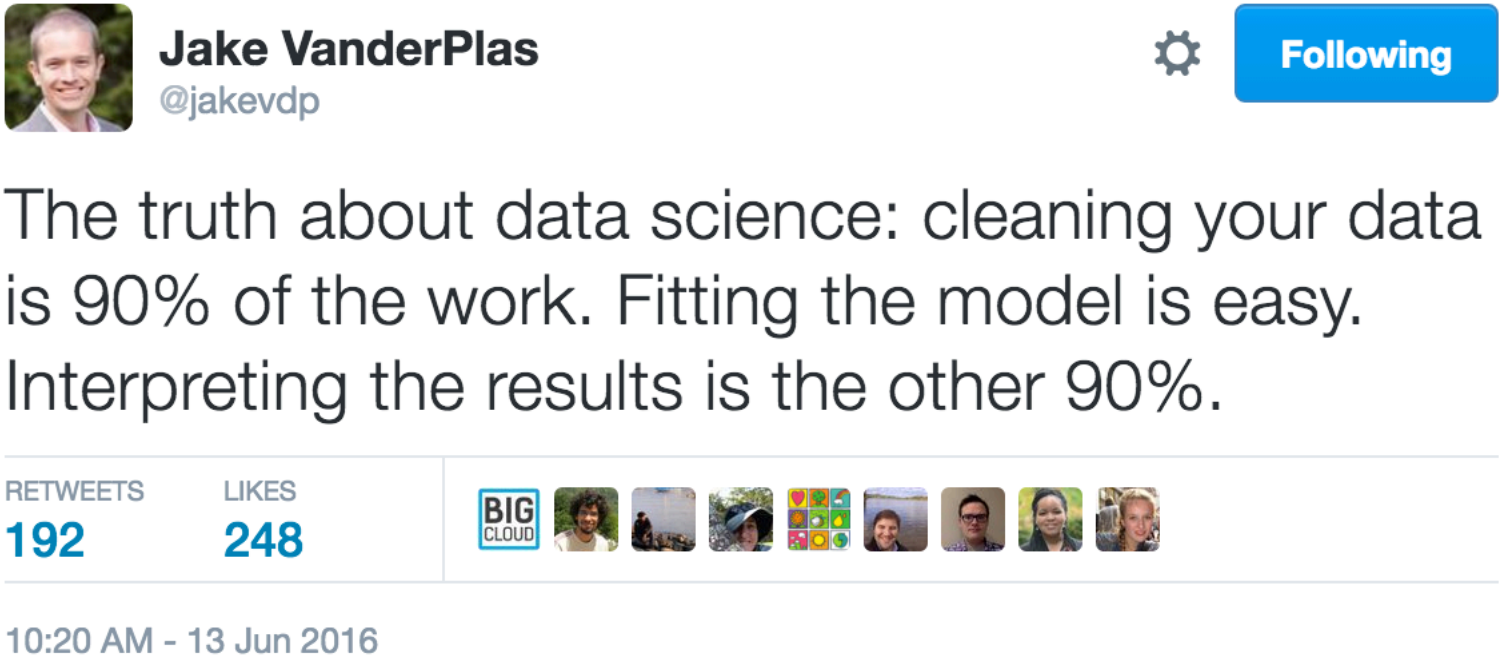
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



KBI BIOPHARMA

This tutorial's mission



KBI BIOPHARMA

The story we will follow

To learn about Pandas, we will explore some climate data, mostly timeseries.
Goal: become better-informed citizens explore data on global temperatures, greenhouse gas and sea-level: load, clean, plot, correlate, search, resample, and model.

Full disclosure: I am not a climate scientist ! (if you are, come talk to me…)

Off to https://github.com/jonathanrocher/pandas_tutorial.git

KBI BIOPHARMA