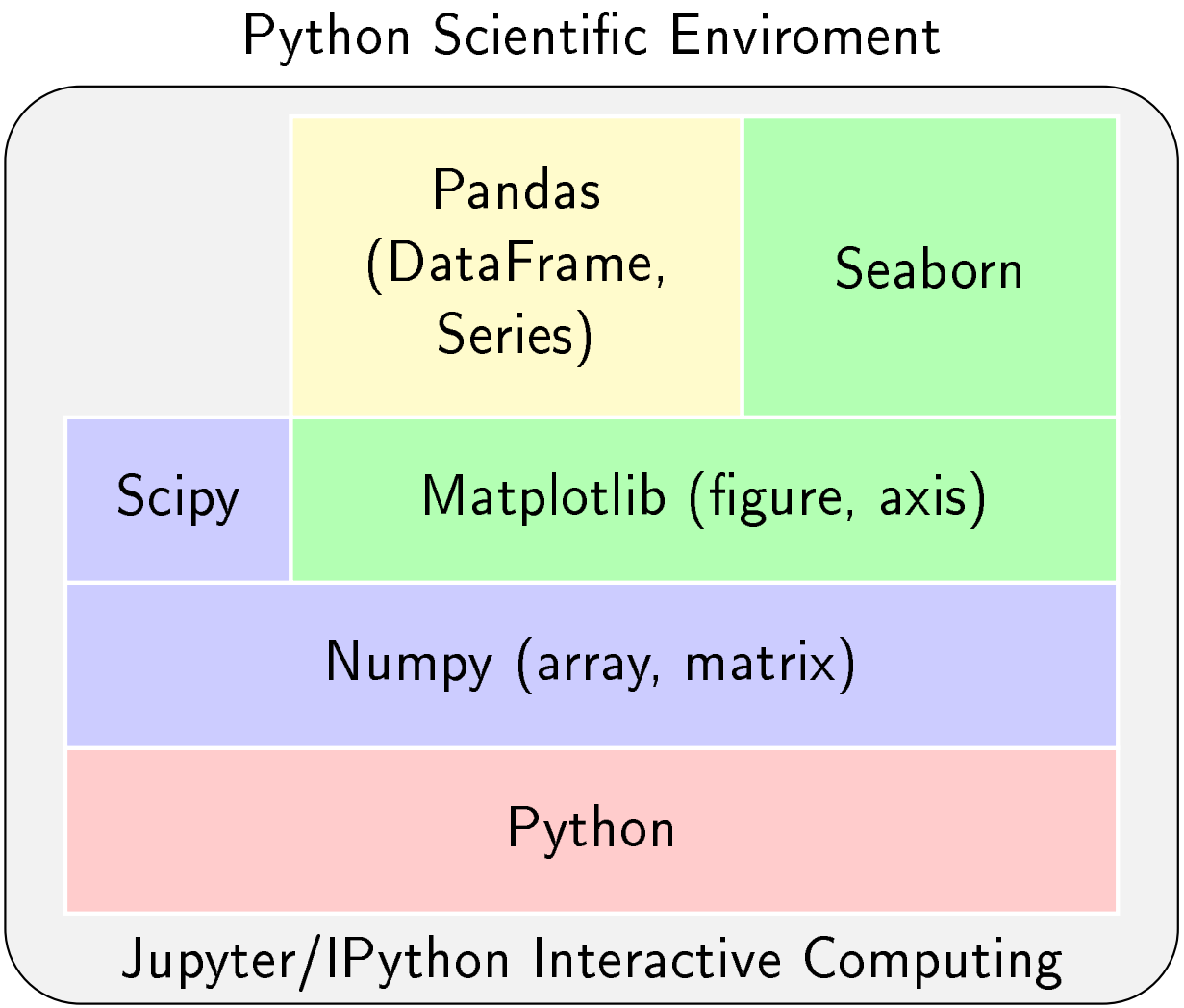
Data Visualization in Python

Árpád Horváth

October 10, 2018

Vázlat

Packages for Data Visualization



- Pylab = Numpy + Matplotlib ≈ MATLAB
- Web-centric plotting libraries: bokeh, plotly

Vázlat

Jupyter notebook vs script

Jupyter notebook

- We need a web browser.
- Set a password (bash commands).
- Easy to run the parts several times with small modifications.
- Titles and texts can be added easily (LaTeX equations as well).
- Easy to share (as is, as HTML, PDF...).
- We can restart the kernel and run all the cells.

Script

- Easy to structure (modules, packages).
- Doesn't need Jupyter to install.
- IDE-s with more capabilities (version control, renaming).

ms.version

Imports used here

```
from matplotlib import pyplot as plt
import numpy as np
import pandas as pd
from pandas import DataFrame, Series
import seaborn as sns
```

In interactive sessions, instead of the first two lines:

```
from pylab import *
```

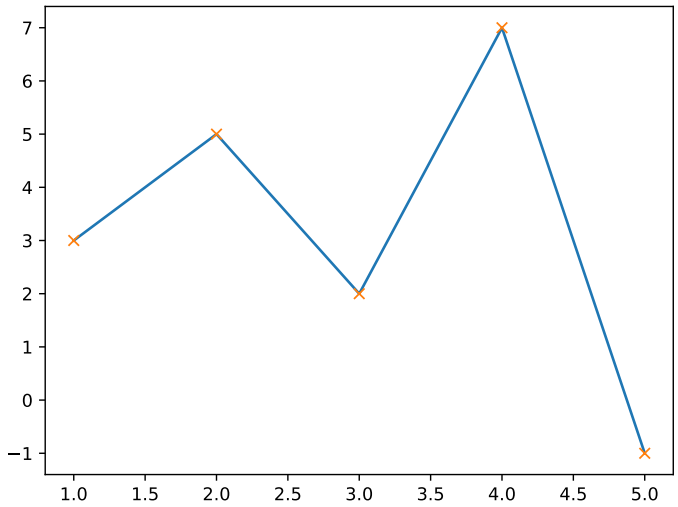
Do not use in scripts!

Vázlat

Vázlat

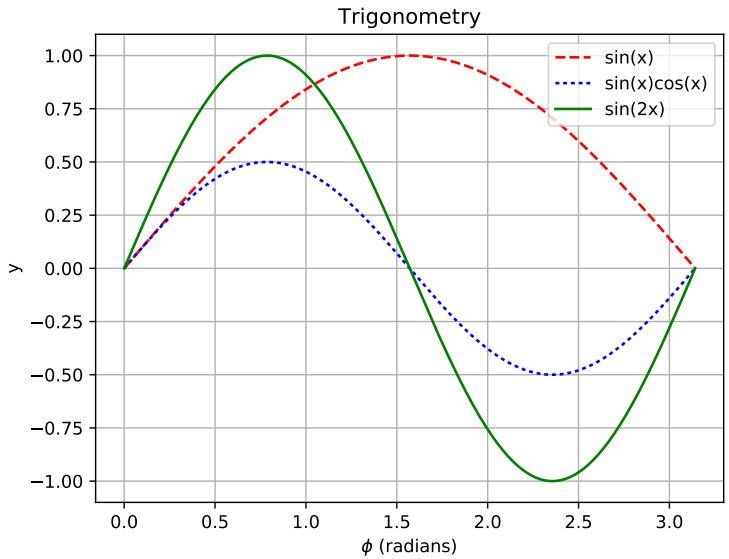
First plot with matplotlib

```
x = [1, 2, 3, 4, 5]
y = [3, 5, 2, 7, -1]
plt.plot(x, y)
plt.plot(x, y, 'x')
plt.savefig('first.pdf')
plt.close()
```



2nd plot

```
x = np.linspace(0, np.pi, 777)
y = np.sin(x)
plt.plot(x, np.sin(x), 'r--', label='sin(x)')
plt.plot(x, np.sin(x)*np.cos(x), 'b:', label='sin(x)cos(x)')
plt.plot(x, np.sin(2*x), 'g-', label='sin(2x)')
plt.legend(loc='upper right')
plt.title('Trigonometry')
plt.xlabel('$\phi$ (radians)')
plt.ylabel('y')
plt.grid(True)
```



pylab

No needs of `np.` and `plt.` with `from pylab import *` or

```
from numpy import linspace, pi
from matplotlib.pyplot import (plot, legend, savefig,
           close, xlabel, ylabel, title, grid)
```

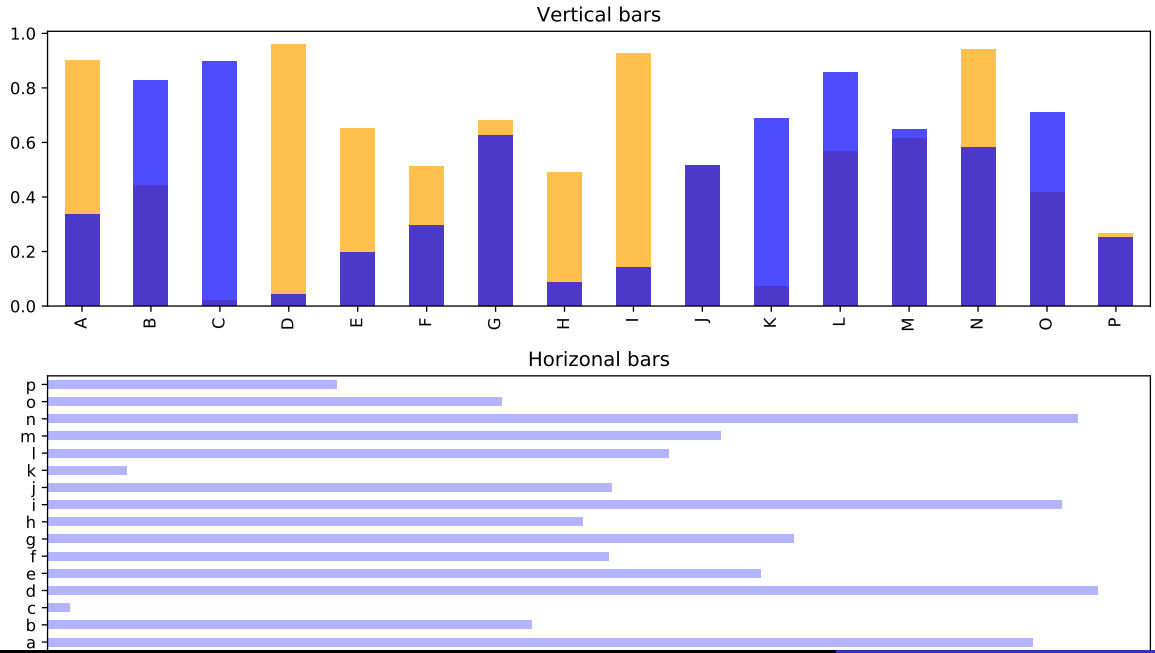
For Jupyter notebook

```
%matplotlib notebook
```

is enough.

Bar plots and colors, tight layout

```
fig, axes = plt.subplots(2, 1)
data = pd.Series(np.random.rand(16), index=list('ABCDEFGHIJKLMNOP'))
data2 = pd.Series(np.random.rand(16), index=list('abcdefghijklmnop'))
data2.plot.bar(ax=axes[0], color='orange', alpha=0.7)
data.plot.bar(ax=axes[0], color='b', alpha=0.7)
data2.plot.barh(ax=axes[1], color='#0000ff', alpha=0.3)
axes[0].set_title('Vertical bars')
axes[1].set_title('Horizonal bars')
fig.tight_layout()
```



sharex, sharey

```
fig, axes = plt.subplots(2, 2, sharex=True, sharey=True)
for i in range(2):
    for j in range(2):
        axes[i, j].hist(np.random.randn(500), bins=50, color='k', alpha=0.5)
plt.subplots_adjust(wspace=0, hspace=0)
```

sharex, sharey : bool or {'none', 'all', 'row', 'col'}

Vázlat

Vázlat

Wes McKinney: Python for Data Analysis, 2nd edition

- `pyedu.hu/pandas2toc` Chapter 9: Visualization
- `github.com/wesm/pydata-book` ,
  - notebooks
    Notebooks for 1st edition works well in MS, but not Seaborn in it, some problems with the notebooks of 2nd edition
    I can share a fixed one
  - data sets