Python tools for data science

Jeffrey Salmond

July 3, 2017

Overview

Transforming Data

- numpy

Organising Data

- pandas
- pytables

Analysing Data

- scipy
- scikit-learn
- opencv

Plotting Data

- matplotlib
- bokeh
- seabourn

And more

- sympy
- cython

numpy extends Python with arrays

```
import numpy as np
x = np.array([[ 1, 0, 0], [0, 1, 2]])
```

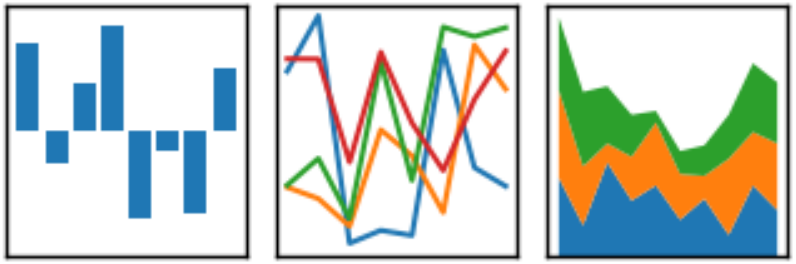
- operations on arrays
- indexing, slicing, reshaping
- views and fancy indexing

builds on numpy to provide tools for data Series and Tables

- easy import/export to csv and other formats
- functions to select and aggregate data using joins and pivots
- functions to deal with time-series data
- easy interface to rapidly plot data

pandas

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



adds functionality to save large datasets to compressed binary files

- built on the HDF5 data format
- good for *hierarchical* data
- can efficiently deal with very large amounts of data by not loading it all in memory

an ecosystem of packages for science, mathematics and engineering

wraps c, c++ and fortran code to provide

- integration
- optimisation
- interpolation
- Fourier transforms
- linear algebra
- and many more...

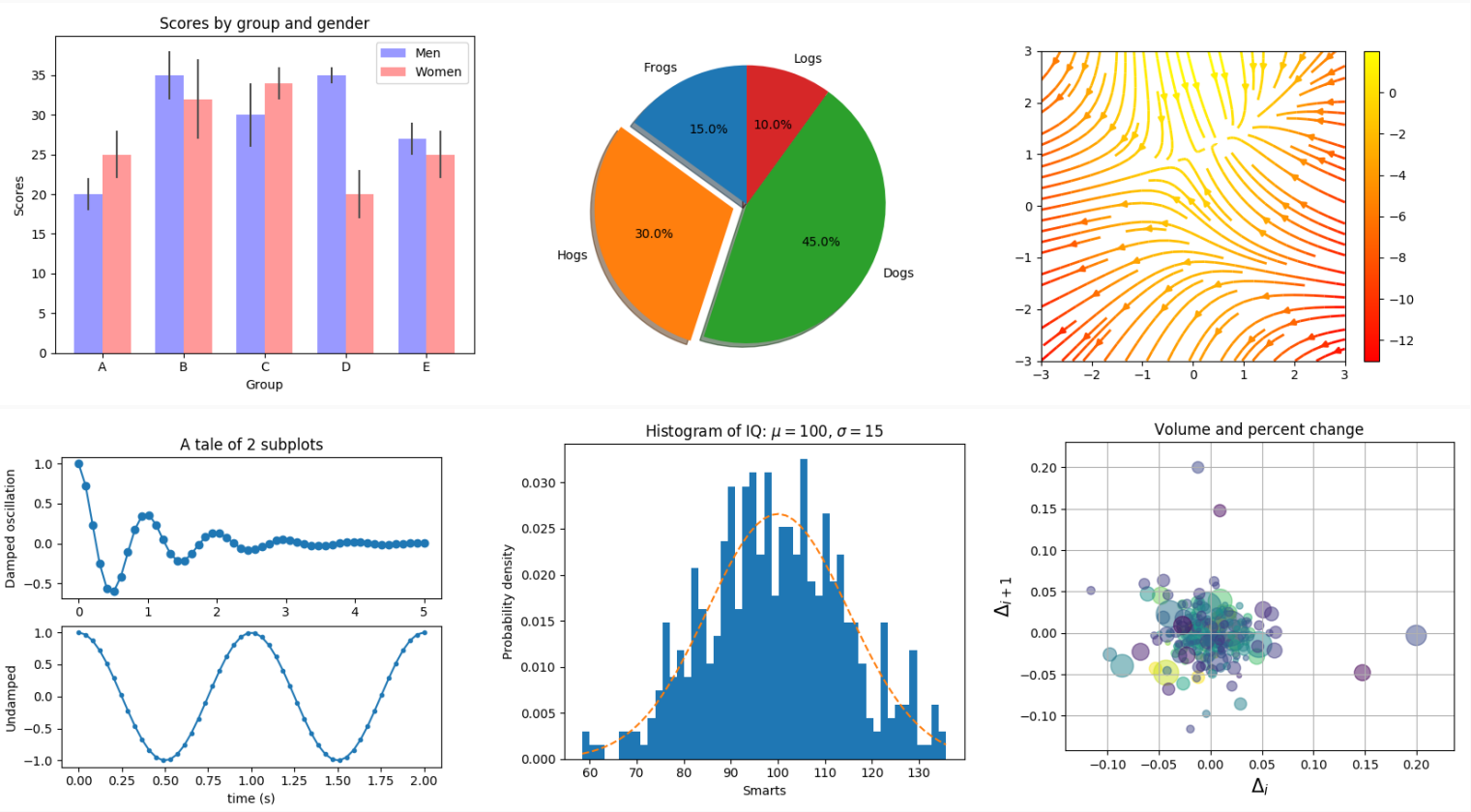
provides tools for data mining and data analysis, including

- classification
- regression
- clustering
- dimensionality reduction
- model selection

provides tools centred around computer vision, including

- image/video input and output
- 3D reconstruction
- feature extraction
- object detection
- GPU acceleration

the most popular 2D plotting library for python



matplotlib

An alternative to matplotlib

- built to let you construct interactive visualisations
- targets browsers for viewing and interacting with the data

builds on top of matplotlib to make plots of common statistical analysis easier, including

- plotting distributions of datasets
- displaying categorical data

a computer algebra system, or library for symbolic mathematics

- can automatically rearrange and solve algebraic equations
- can automatically differentiate and integrate expressions
- can take these generated expressions and generate python, C or fortran code

cython

cython is a set of tools to speed up python programs

- can call from C to python or python to C
- can annotate python code and compile it for more performance
- underpins many of the libraries we have talked about

Deep learning frameworks

tensorflow `tensorflow.org`

a new set of tools from Google

caffe `caffe.berkeleyvision.org`

a deep/machine learning framework

theano `deeplearning.net/software/theano`

another deep/machine learning framework