みんなのPython勉強会 in 長野 #3 *3/23, 2019*

データサイエンスチュートリアル

辻 真吾（Start Python Club）
@tsjshg
shingo.tsuji@gmail.com

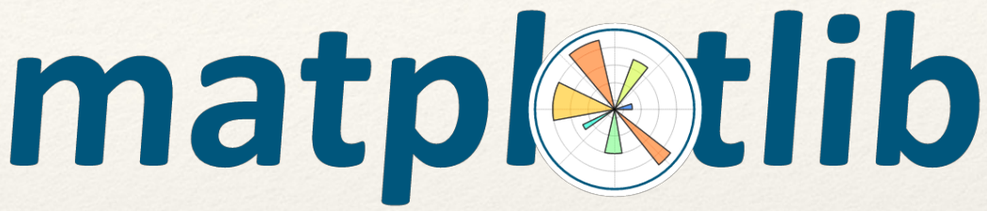
データサイエンスの実際

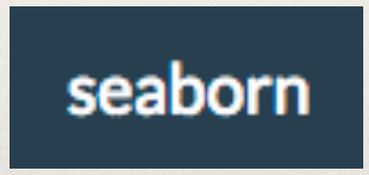
- データを集める（CSV, Excel, DB…）
- 欠損値や異常値の検出と処理
- データが正しく測定されているか？の確認

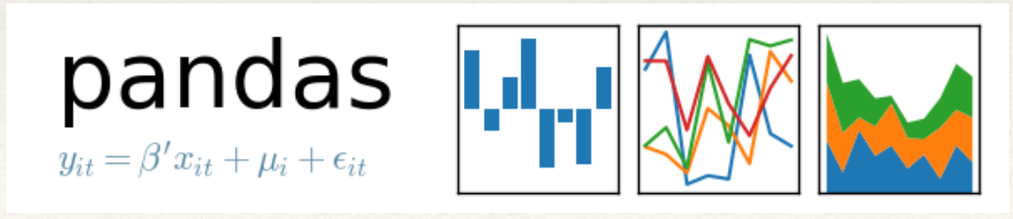
- 完成品のWebアプリ化

Pythonはどの場面でも利用でき、存在感が増している

Pythonはglue（のり）言語







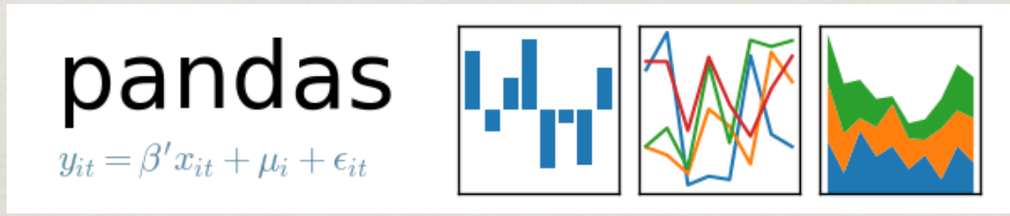
NumPy, SciPy

- Pythonでのデータ解析、科学計算の基礎となるライブラリ

- array（ベクトルや行列）の高速な演算を実現
- 基本的な統計関数や数値積分、最適化なども可能

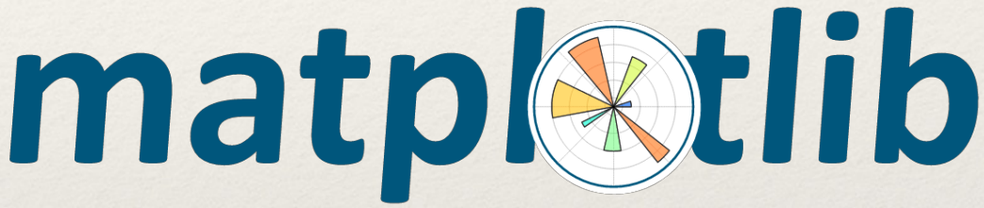
Pandas

- 日常のデータ解析になくてはならない超高性能ライブラリ
- エクセルのシートをイメージすると分かり易いかも
- データの入出力、加工、可視化など幅広く対応



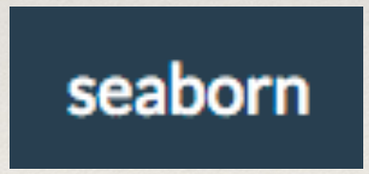
matplotlib, seaborn

- データの可視化に使われる



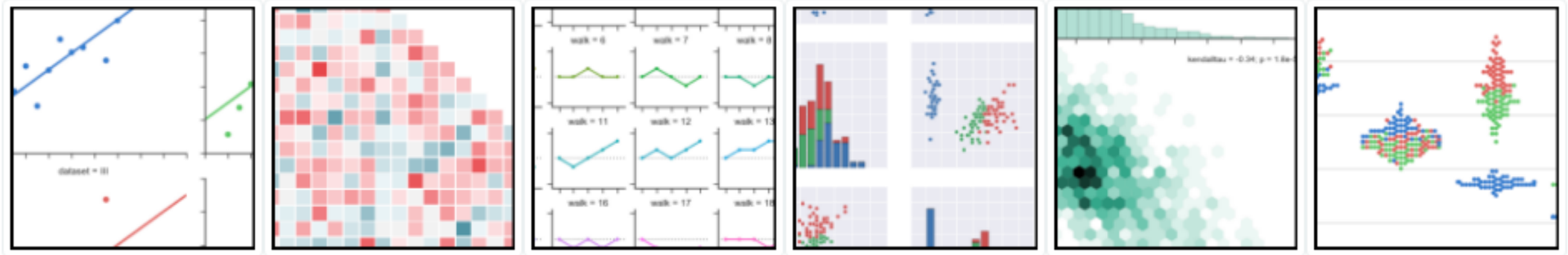
- matplotlibはmatlabの代替を意識

- seabornはmatplotlibを基礎にして、使用しやすく、統計的な機能も取り込む



seaborn

Seaborn: statistical data visualization



Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

For a brief introduction to the ideas behind the package, you can read the introductory notes. More practical information is on the installation page. You may also want to browse the example gallery to get a sense for what you can do with seaborn and then check out the tutorial and API reference to find out how.

To see the code or report a bug, please visit the github repository. General support issues are most at home on stackoverflow, where there is a seaborn tag.

Documentation

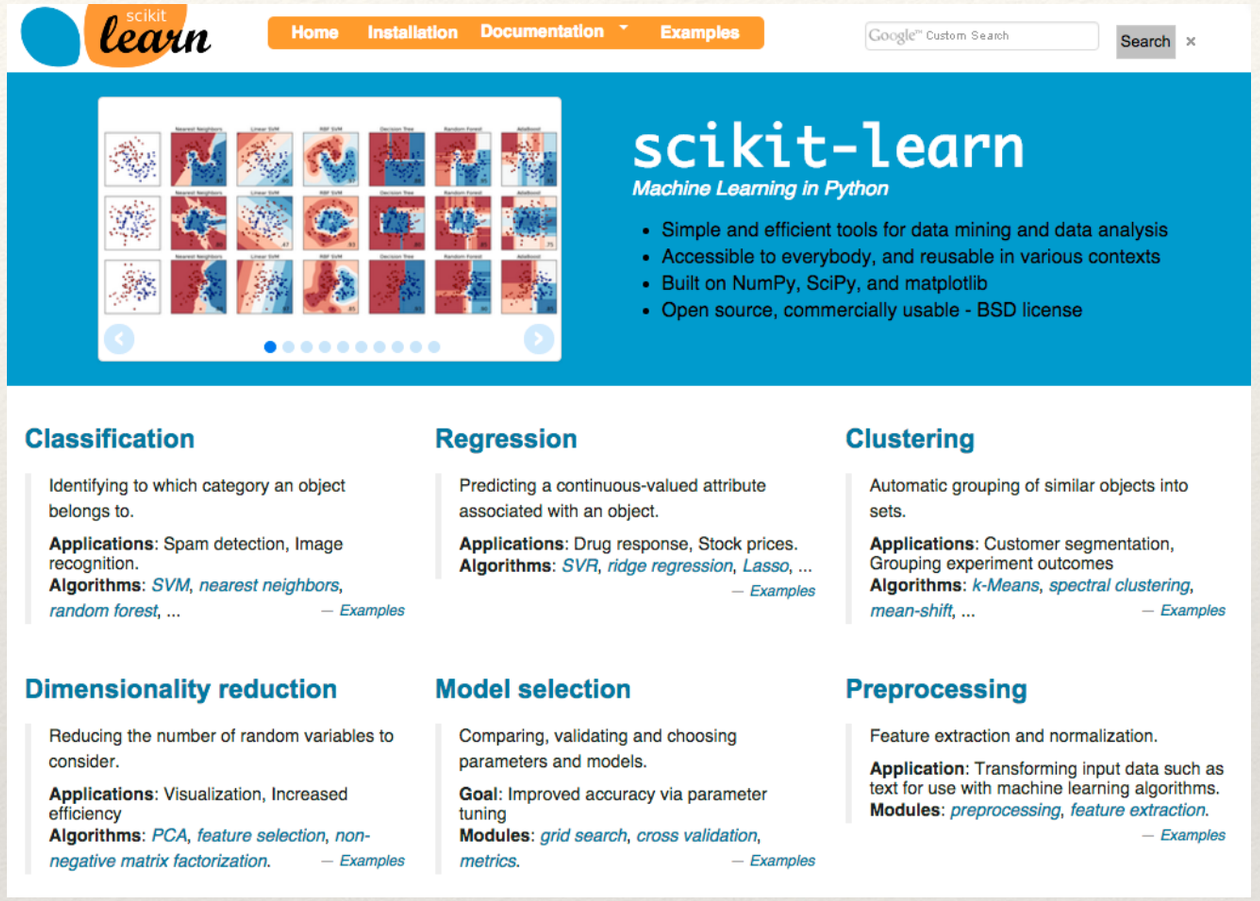
- An introduction to seaborn
- What's new in the package
- Installing and getting started
- Example gallery
- API reference
- Seaborn tutorial

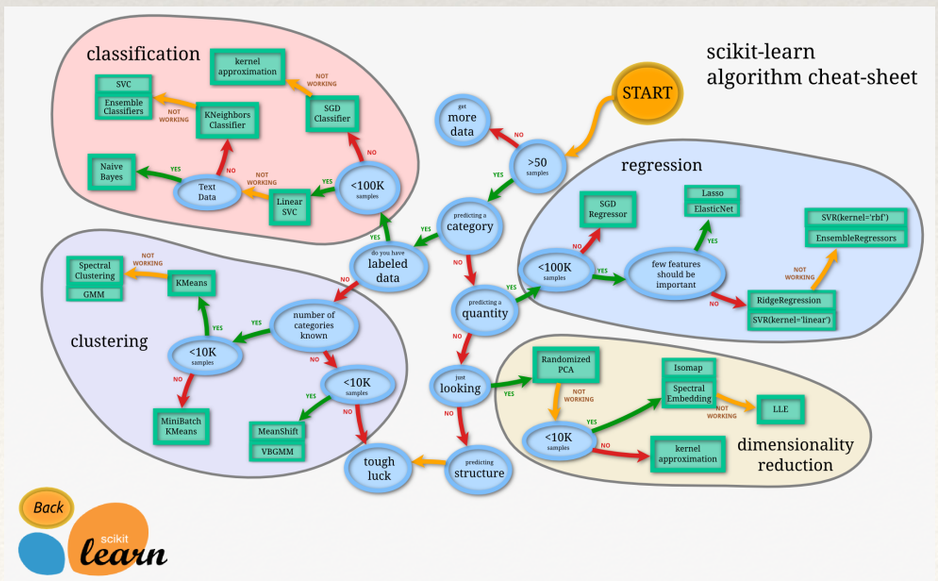
Features

- Style functions: API | Tutorial
- Color palettes: API | Tutorial
- Distribution plots: API | Tutorial
- Regression plots: API | Tutorial
- Categorical plots: API | Tutorial
- Axis grid objects: API | Tutorial

scikit-learn

- 進化し続けるPythonの機械学習ライブラリ
- ドキュメントとコード例が豊富にある
- cheat sheetなどアルゴリズムを学べる資料も多数





The Iris Dataset

- よく使われるサンプルデータ
- アヤメの花に関するデータ
- 3種類x50サンプル
  - 説明変数は4つ
- wikipedia（英語）の記事が詳しいです
- http://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html#example-datasets-plot-iris-dataset-py

I. setosa

versicolor

virginica

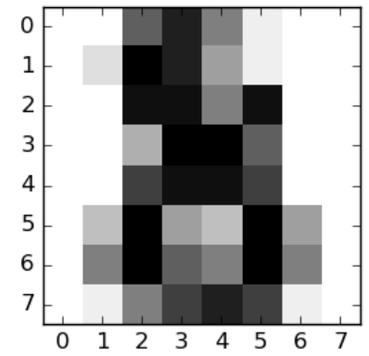
The Digit Dataset

- よく使われるサンプルデータ
- 手書きの数字データ
- 扱いやすいように8x8のグレースケールになっている
- http://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html

The Digit Dataset

This dataset is made up of 1797 8x8 images. Each image, like the one shown below, is of a hand-written digit. In order to utilize an 8x8 figure like this, we'd have to first transform it into a feature vector with length 64.

See here for more information about this dataset.



Python source code: plot_digits_last_image.py

```
print(__doc__)


# Code source: Gaël Varoquaux
# Modified for documentation by Jaques Grobler
# License: BSD 3 clause

from sklearn import datasets

import matplotlib.pyplot as plt

#Load the digits dataset
digits = datasets.load_digits()

#Display the first digit
plt.figure(1, figsize=(3, 3))
plt.imshow(digits.images[-1], cmap=plt.cm.gray_r, interpolation='nearest')
plt.show()
```

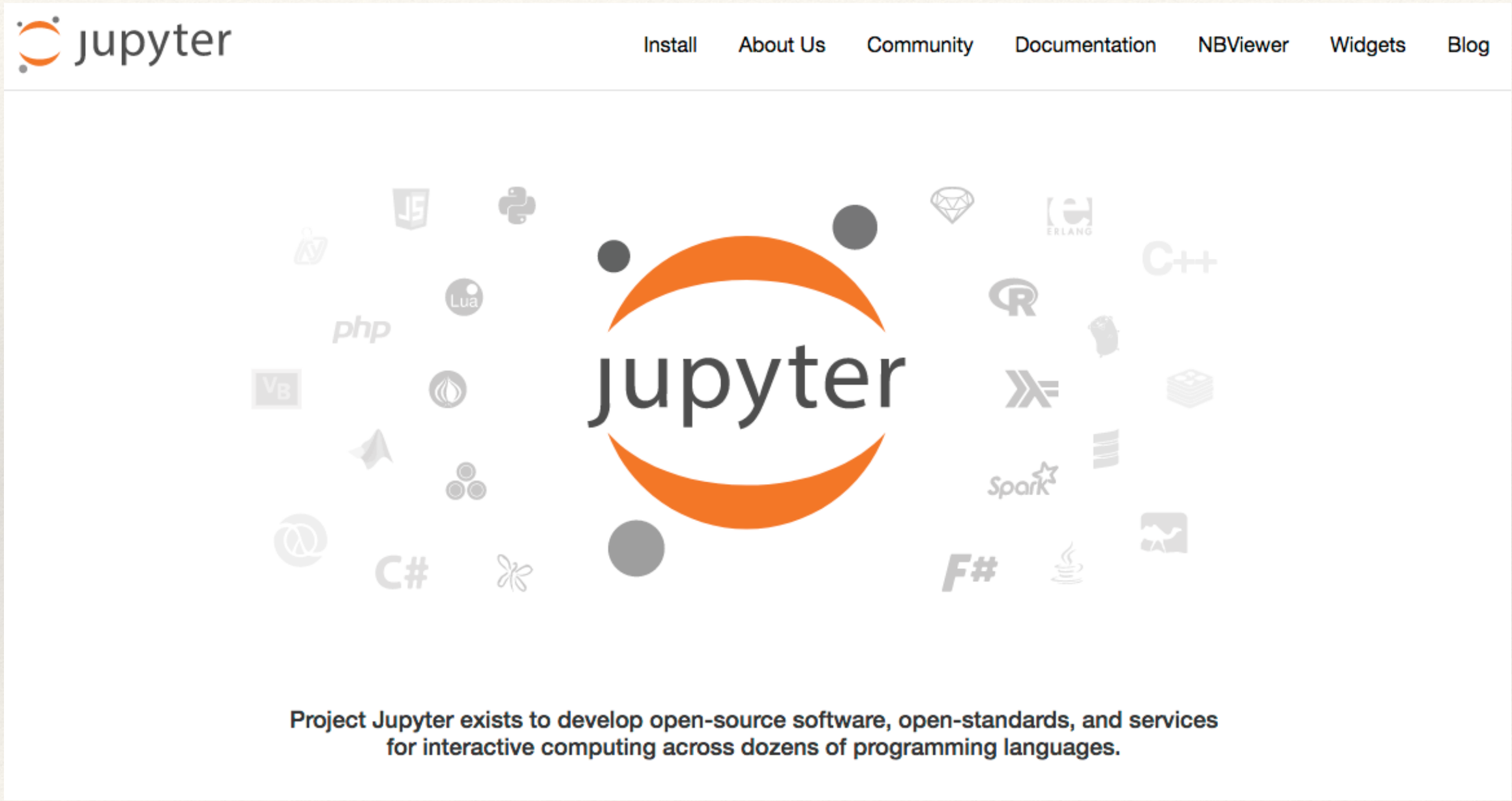
Total running time of the example: 0.32 seconds (0 minutes 0.32 seconds)

© 2010 - 2014, scikit-learn developers (BSD License). Show this page source

Deep Learningの主なライブラリ

- Caffe　最も古いライブラリの1つ
  - http://caffe.berkeleyvision.org/
- theano　Deep Learningの計算の基本を実装。ちょっと使いにくい
  - http://deeplearning.net/software/theano/
- Pylearn2　theanoを使いやすくした感じ
  - http://deeplearning.net/software/pylearn2/
- Keras　非常に使いやすいライブラリ。作者が最近、Googleに転職
  - https://keras.io/ja/
- Chainer　日本のPreferred Networks社が開発
  - http://chainer.org/
- TensorFlow　Googleの機械学習アルゴリズムライブラリ
  - https://www.tensorflow.org/

jupyterがすごい



昔は、IPython notebookだったが、Python以外の言語の実行にも対応している。

たとえば、Rはすぐ設定できる。

とりあえず、やってみよう