

みんなのPython勉強会 in 長野 #3

3/23, 2019

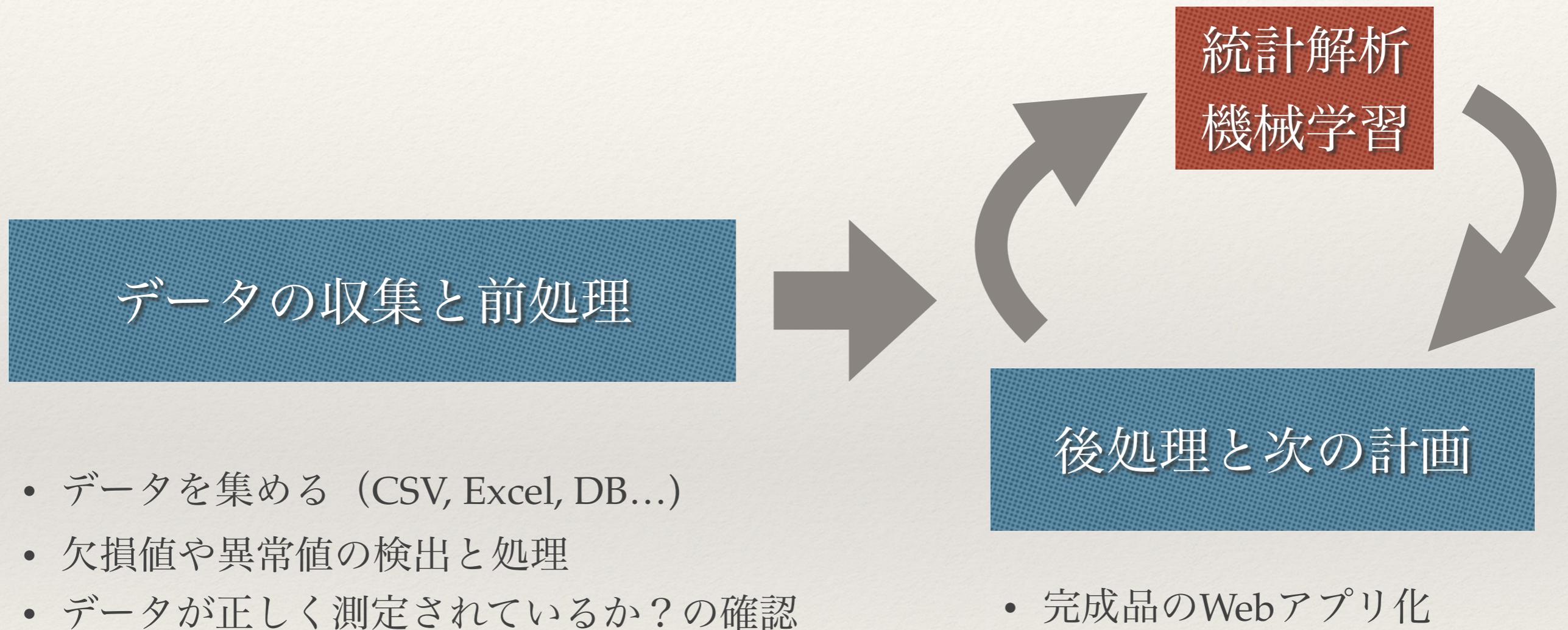
データサイエンス チュートリアル

辻 真吾 (Start Python Club)

@tsjshg

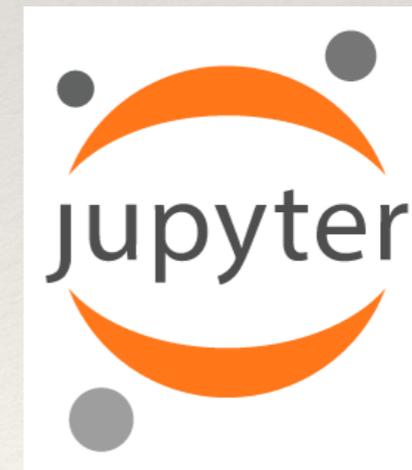
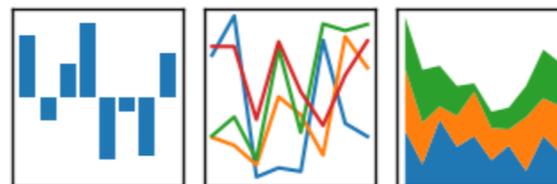
shingo.tsuji@gmail.com

データサイエンスの実際



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

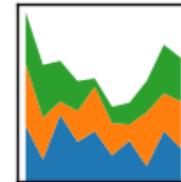
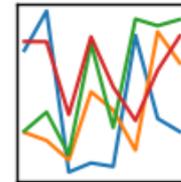


Pandas

- ❖ 日常のデータ解析になくっ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

pandas

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



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

matplotlib

seaborn

seaborn

seaborn

0.8.dev

API

Tutorial

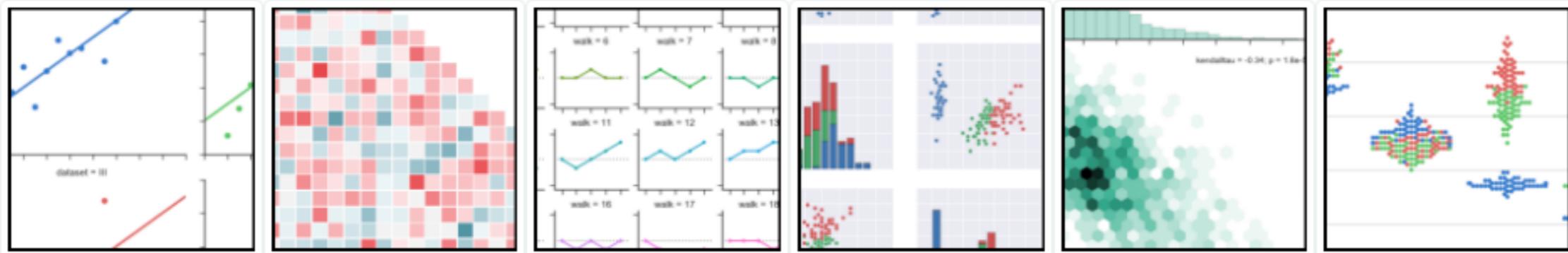
Gallery

Site ▾

Page ▾

Search

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 Features

- [An introduction to seaborn](#)
- [What's new in the package](#)
- [Installing and getting started](#)
- [Example gallery](#)
- [API reference](#)
- [Seaborn tutorial](#)
- [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 screenshot shows the scikit-learn website homepage. At the top, there is a navigation bar with links for Home, Installation, Documentation, and Examples. A search bar is also present. Below the navigation bar is a large blue banner with the scikit-learn logo and the tagline "Machine Learning in Python". To the right of the banner, there are several bullet points highlighting the library's features: "Simple and efficient tools for data mining and data analysis", "Accessible to everybody, and reusable in various contexts", "Built on NumPy, SciPy, and matplotlib", and "Open source, commercially usable - BSD license". Below the banner, the website is organized into six main sections: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing. Each section provides a brief description, applications, and algorithms.

Classification
Identifying to which category an object belongs to.
Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

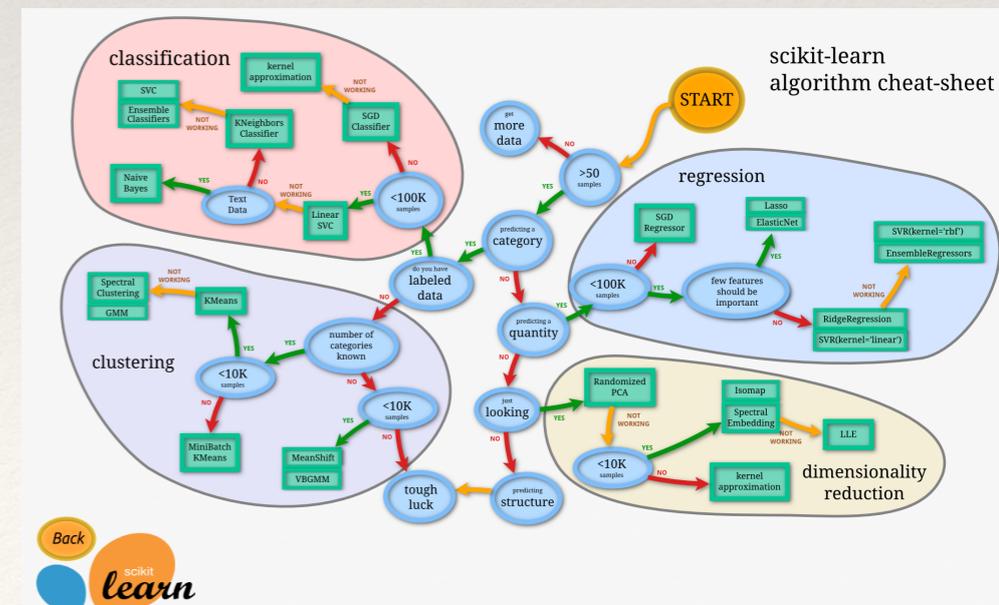
Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction
Reducing the number of random variables to consider.
Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection
Comparing, validating and choosing parameters and models.
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing
Feature extraction and normalization.
Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples



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



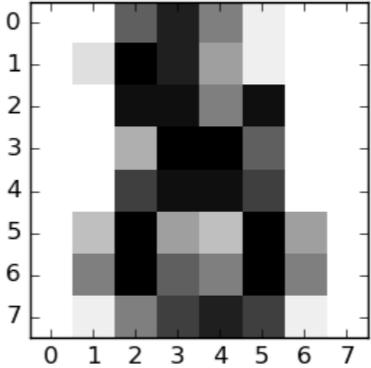
The Digit Dataset

- ❖ 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

[Install](#)[About Us](#)[Community](#)[Documentation](#)[NBViewer](#)[Widgets](#)[Blog](#)

Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

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