

みんなのPython勉強会 #13

2016.6.7

scikit-learnで知る 機械学習の全体像

辻 真吾

@tsjshg

自己紹介

- ❖ 1975年生まれ
- ❖ 都内のとある大学で研究やっております
 - ❖ 研究室のテーマは癌とゲノム
 - ❖ 私がやっているのはPythonでデータ解析
- ❖ 昔はJavaやC++も好きでしたが、最近ほぼPython
- ❖ <http://www.tsjshg.info/>

Udemy

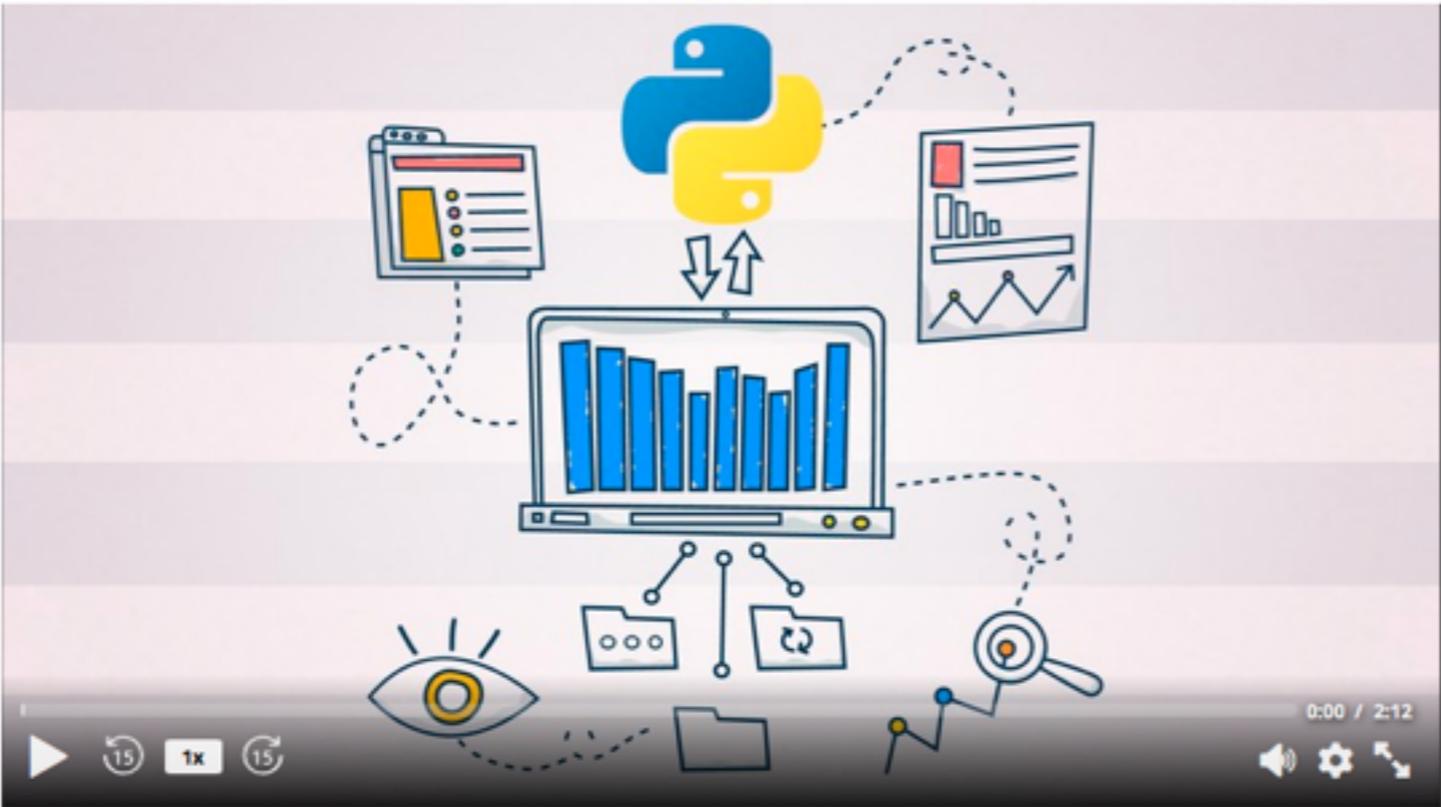
コース検索  **udemy** 講師になりたい方へ

【世界で2万人が受講】 実践 Python データサイエンス

データ解析の基本、可視化、統計、機械学習などデータサイエンスに関するあらゆる実践的なスキルがPythonで身に付く！

★★★★★ 182 評価、1683 生徒が登録済み

講師： Shingo Tsuji, Jose Portilla 開発 / プログラミング言語



¥6,000

[このコースを受講する](#)

[クーポンを利用する](#)
[無料プレビューを開始する](#)
[その他のオプション](#)

レクチャー	104
ビデオファイル	17.5 時間
スキルレベル	すべてのレベル
言語	日本語
その他：	学習期限なし 30日間返金保証 iOS・Androidどちらも受講可能 修了証明書

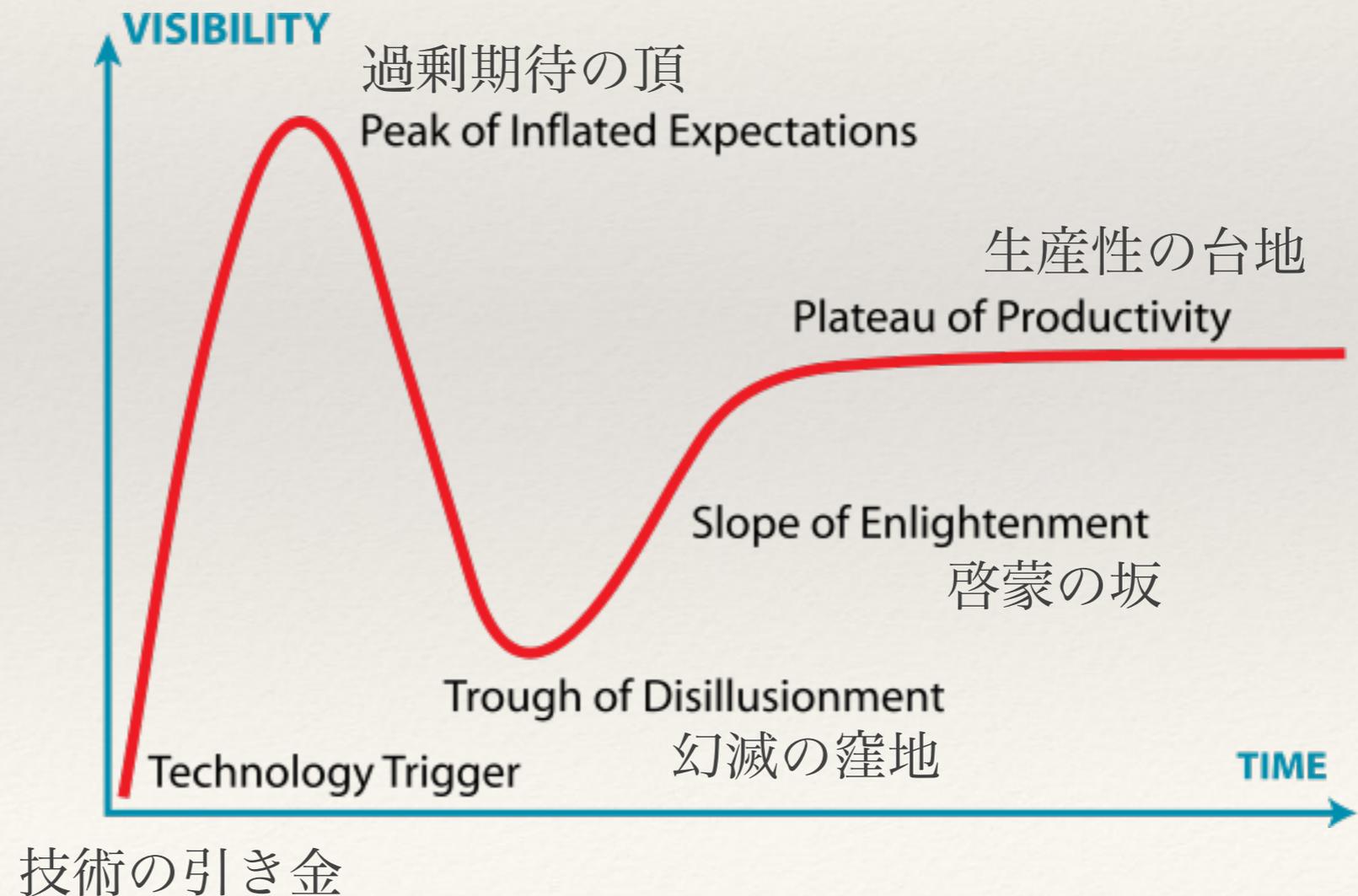
[❤️ ほしい物リスト](#)

2016/6/30までの30% OFF クーポンです。

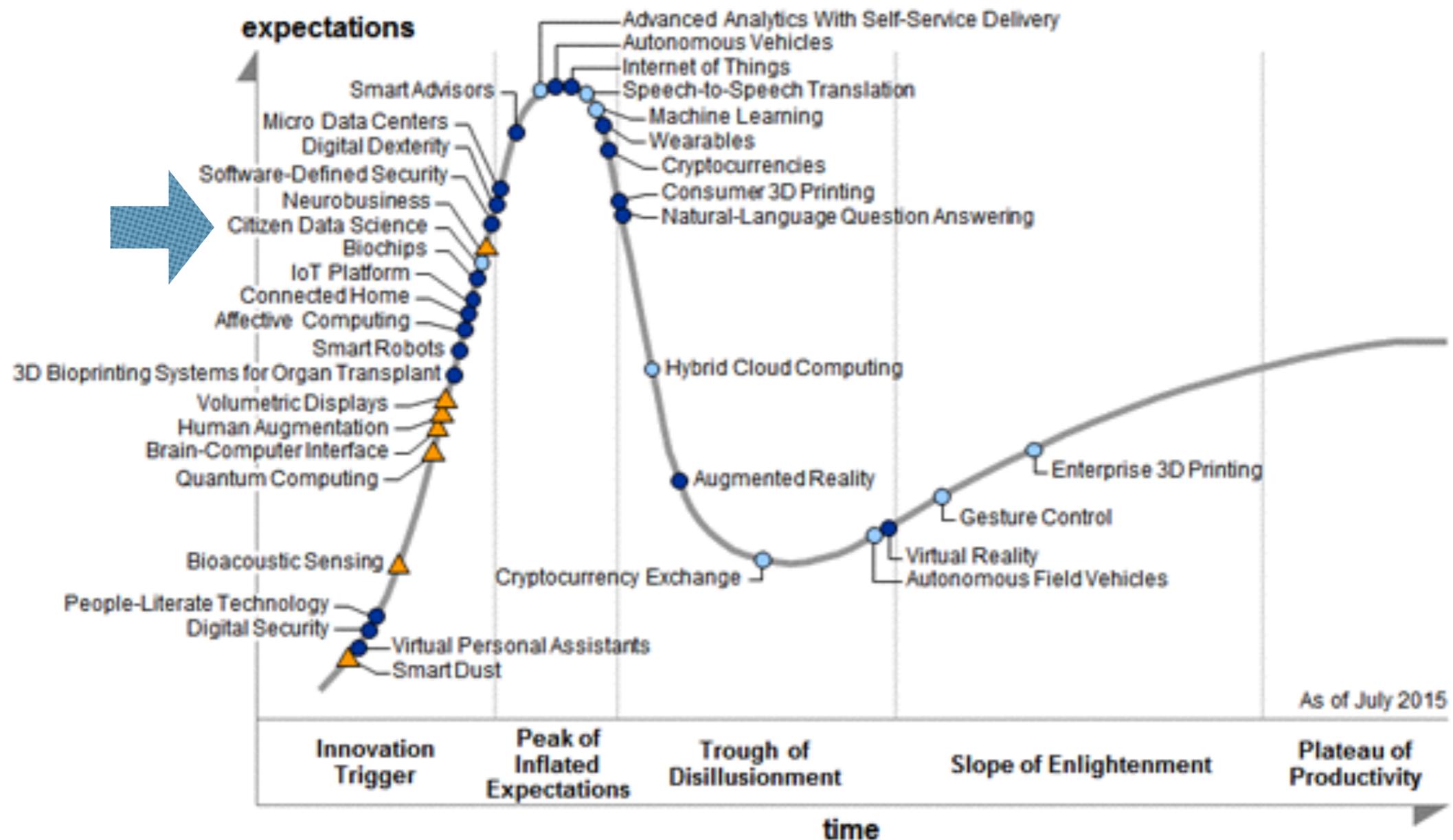
<https://www.udemy.com/python-jp/?couponCode=stapy13-35d>

Hype Cycle

- ❖ 先端技術がたどる典型的な道筋
- ❖ 米国の調査会社Gartnerが提唱



Gartner Hype Cycle 2015



Citizen Data Science

- ❖ (またも) Gartner社の提唱した概念
- ❖ 統計やプログラミングを専門にしないが、自分の分野のデータを解析する人
- ❖ データサイエンティストの不足を補う

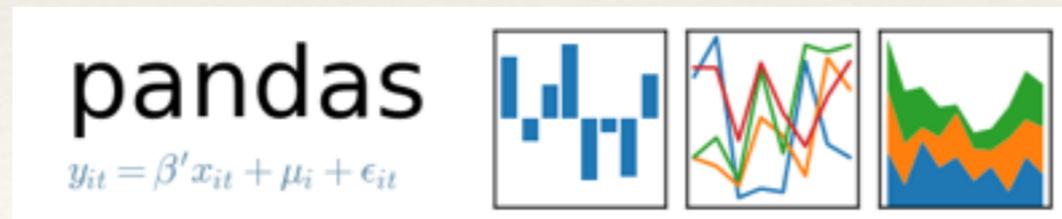
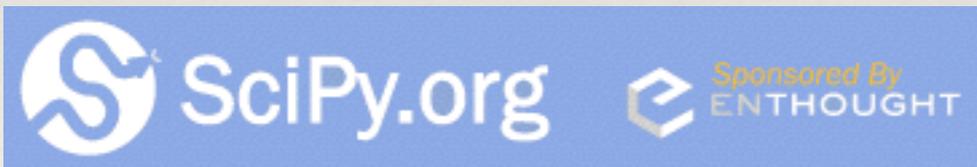
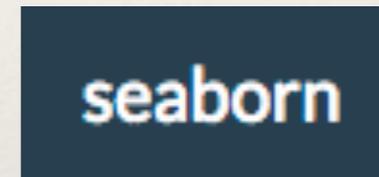
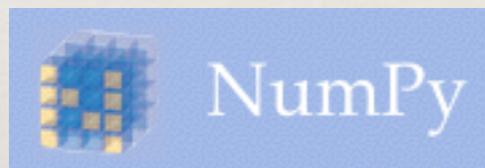
データサイエンスやりたい？

Pythonが便利

- ❖ 最近、怒濤の勢い
- ❖ これから始めるな3系を
- ❖ 汎用言語なのでなんでもできますが、データ解析分野で大きな存在感



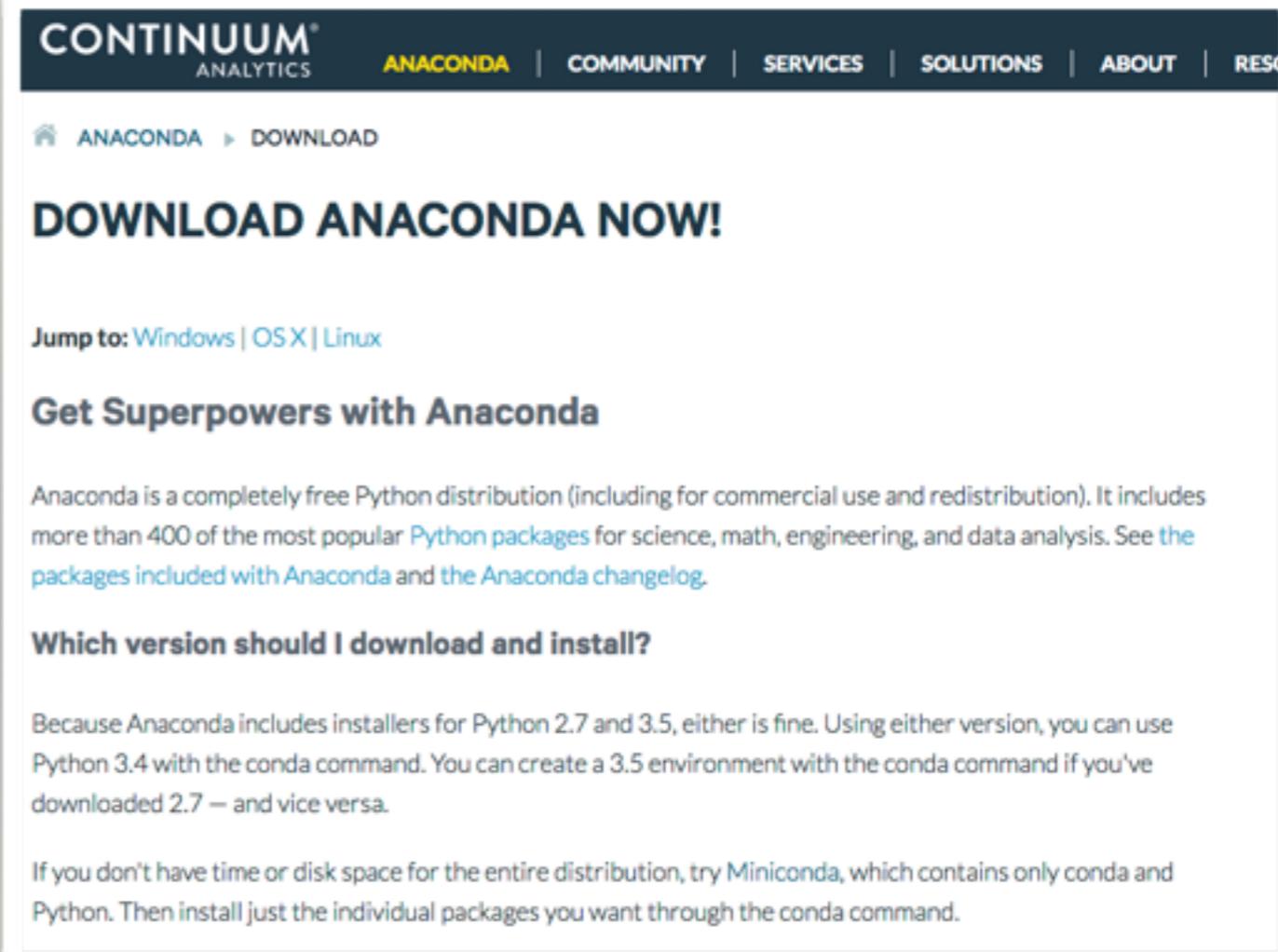
Pythonはglue (のり) 言語



IP[y]:
IPython

Anacondaがおすすめ！

- ❖ Continuum Analytics社が配布するPython
- ❖ 標準のPythonにcondaをはじめとして多くの外部ライブラリを同梱
- ❖ 無料です
- ❖ データ解析環境があっという間に整います



The screenshot shows the 'Download Anaconda Now!' page from the Continuum Analytics website. The page features a dark blue header with the Continuum Analytics logo and navigation links for ANACONDA, COMMUNITY, SERVICES, SOLUTIONS, ABOUT, and RESOURCES. Below the header, there is a breadcrumb trail 'ANACONDA > DOWNLOAD' and a prominent 'DOWNLOAD ANACONDA NOW!' button. The page also includes links to download for Windows, OSX, and Linux, and a section titled 'Get Superpowers with Anaconda' which describes Anaconda as a free Python distribution with over 400 packages. A section titled 'Which version should I download and install?' explains that both Python 2.7 and 3.5 installers are available and compatible. Finally, it mentions Miniconda as a lighter alternative for users with limited disk space.

<https://www.continuum.io>

今日はscikit-learnの話

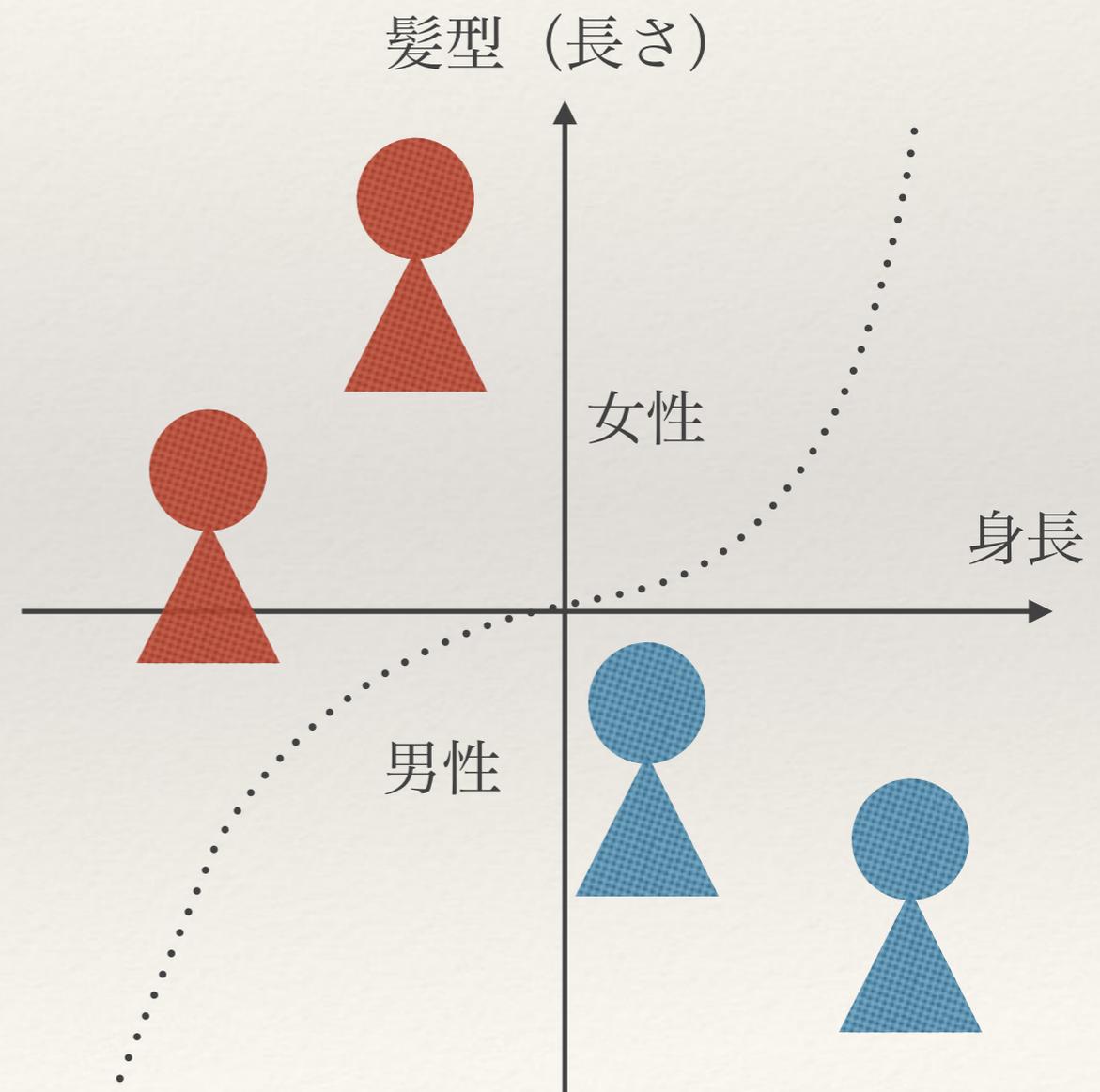
(その前に) 機械学習とは？

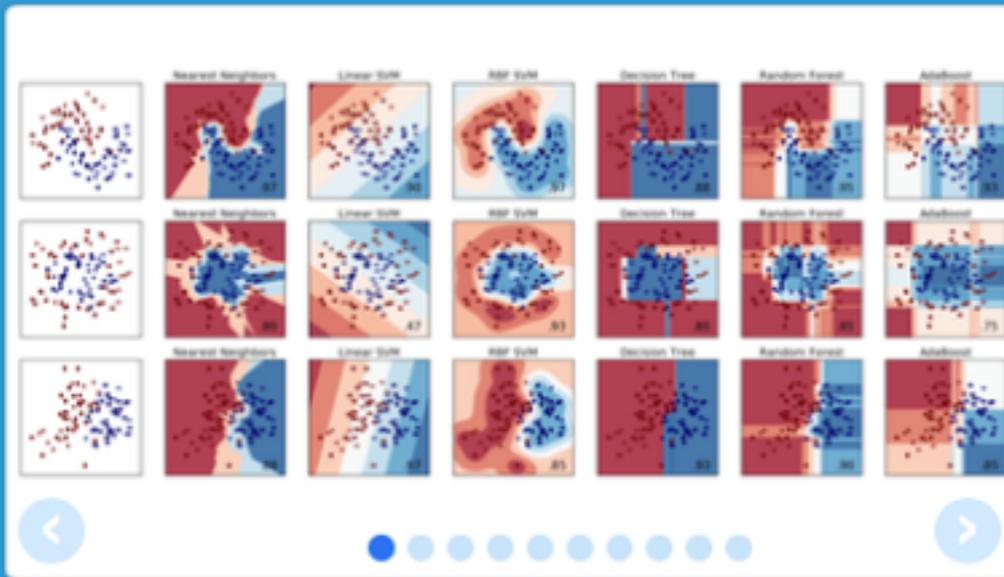
身長182cmで角刈りの人は男性か？女性か？

目的変数

説明変数

性別	身長(cm)	髪型	眼鏡
男	182	角刈り	なし
男	170	七三分け	あり
女	160	ロング	あり
女	153	ショートヘア	なし





scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

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

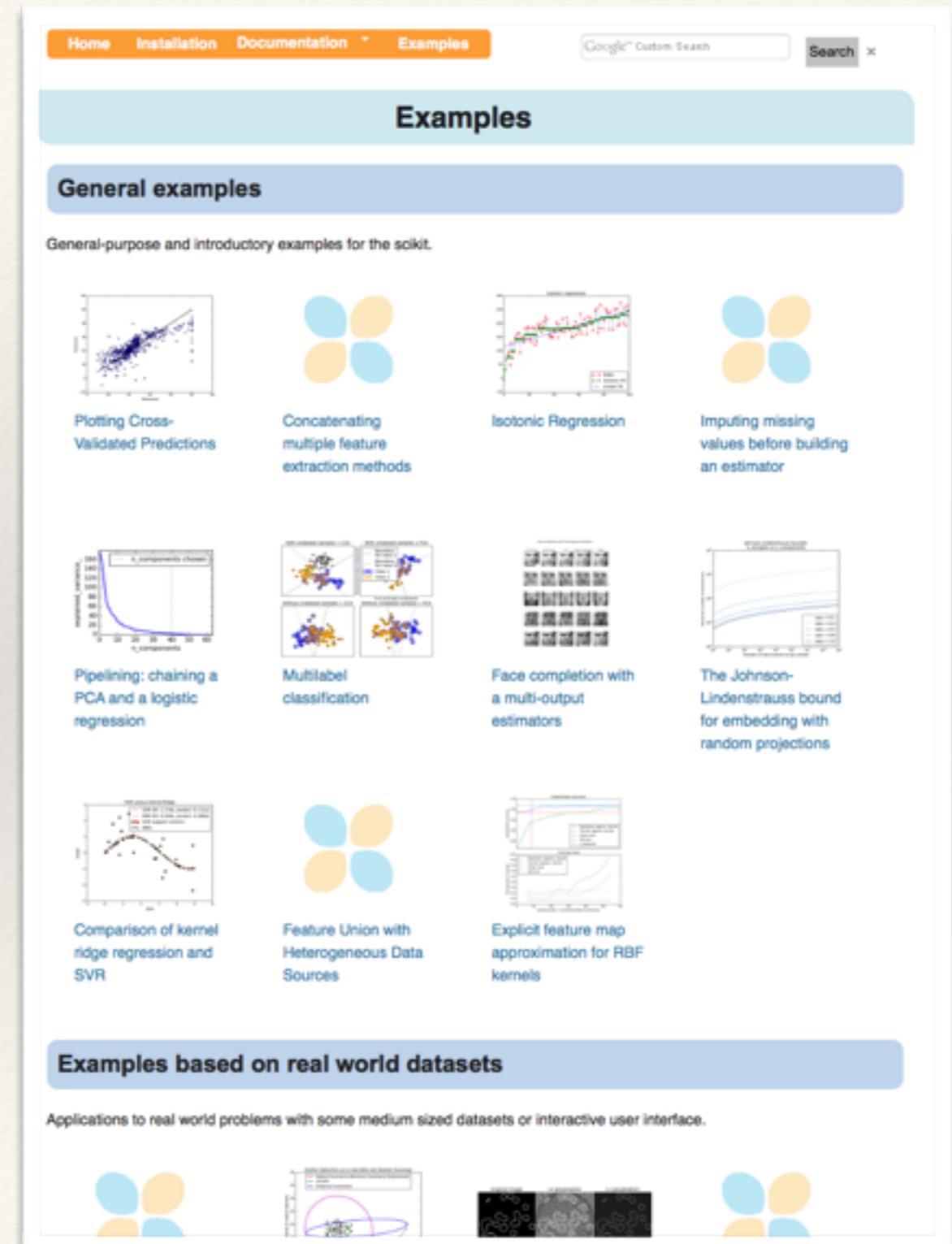
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

膨大すぎる . . .

- ❖ Examplesのページが秀逸
- ❖ 目的別にサンプルコードとその出力が並んでいる
- ❖ 入門するならここからを選びました



The screenshot shows the 'Examples' page from the scikit-learn documentation. At the top, there are navigation links for 'Home', 'Installation', 'Documentation', and 'Examples', along with a search bar. The main heading is 'Examples'. Below it, there is a section titled 'General examples' with the subtitle 'General-purpose and introductory examples for the scikit-learn.' This section contains a grid of 12 example cards, each with a small thumbnail image and a title: 'Plotting Cross-Validated Predictions', 'Concatenating multiple feature extraction methods', 'Isotonic Regression', 'Imputing missing values before building an estimator', 'Pipelining: chaining a PCA and a logistic regression', 'Multilabel classification', 'Face completion with a multi-output estimators', 'The Johnson-Lindenstrauss bound for embedding with random projections', 'Comparison of kernel ridge regression and SVR', 'Feature Union with Heterogeneous Data Sources', and 'Explicit feature map approximation for RBF kernels'. Below this grid is another section titled 'Examples based on real world datasets' with the subtitle 'Applications to real world problems with some medium sized datasets or interactive user interface.' This section also contains a grid of example cards, with the first one visible showing a scatter plot.

よく使われるデータセット

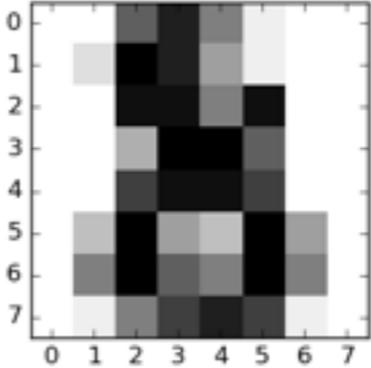
The Digit Dataset

- ❖ 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](#)

The Iris Dataset

- ❖ 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



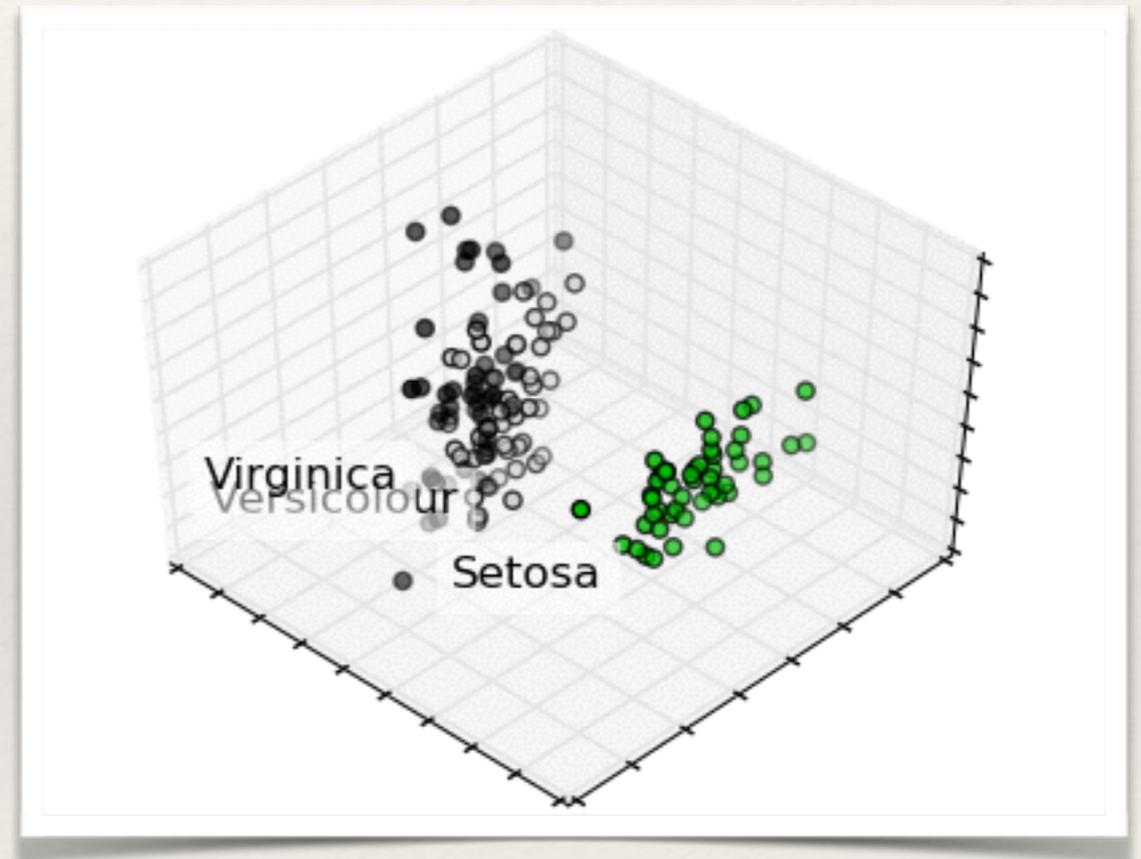
virginica



サンプルの全体像を把握する

PCA example with Iris Data-set

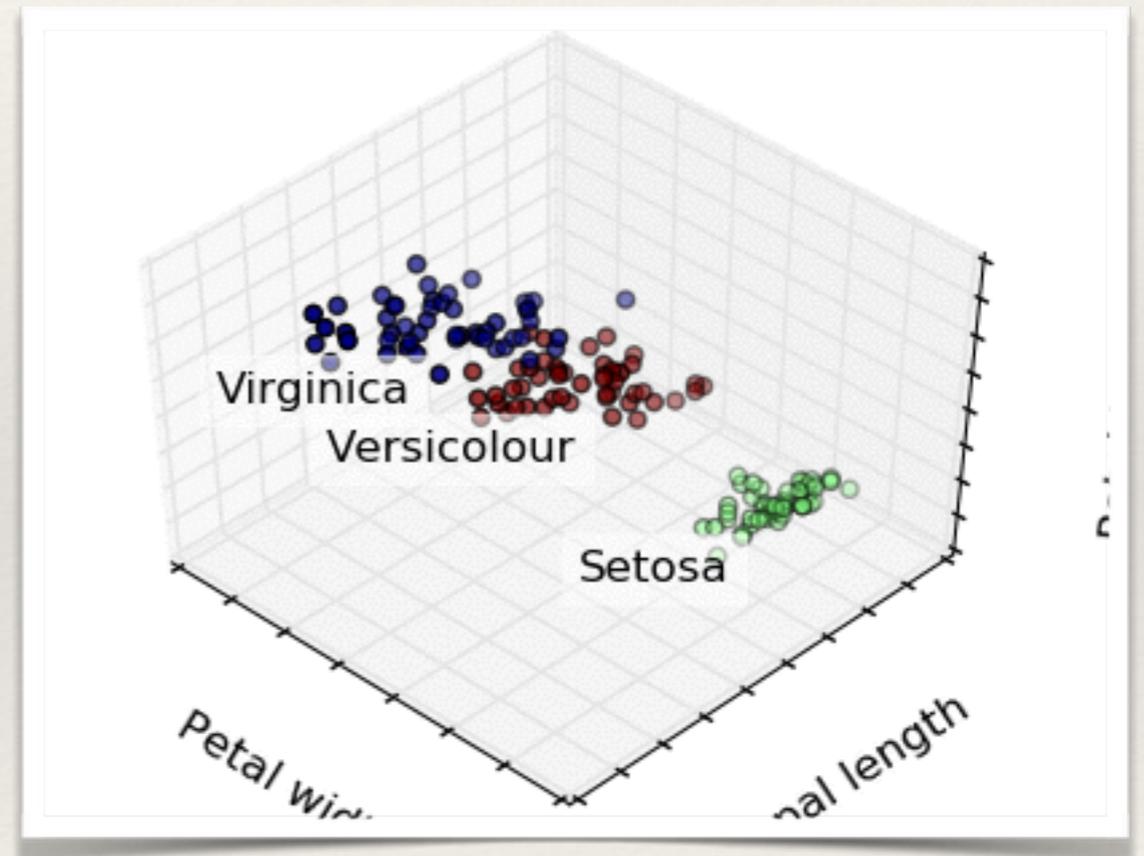
- ❖ Principal Component Analysis(PCA、主成分分析)
- ❖ 多数の説明変数を縮約し、データセットの全体像を把握するのに便利
- ❖ http://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html



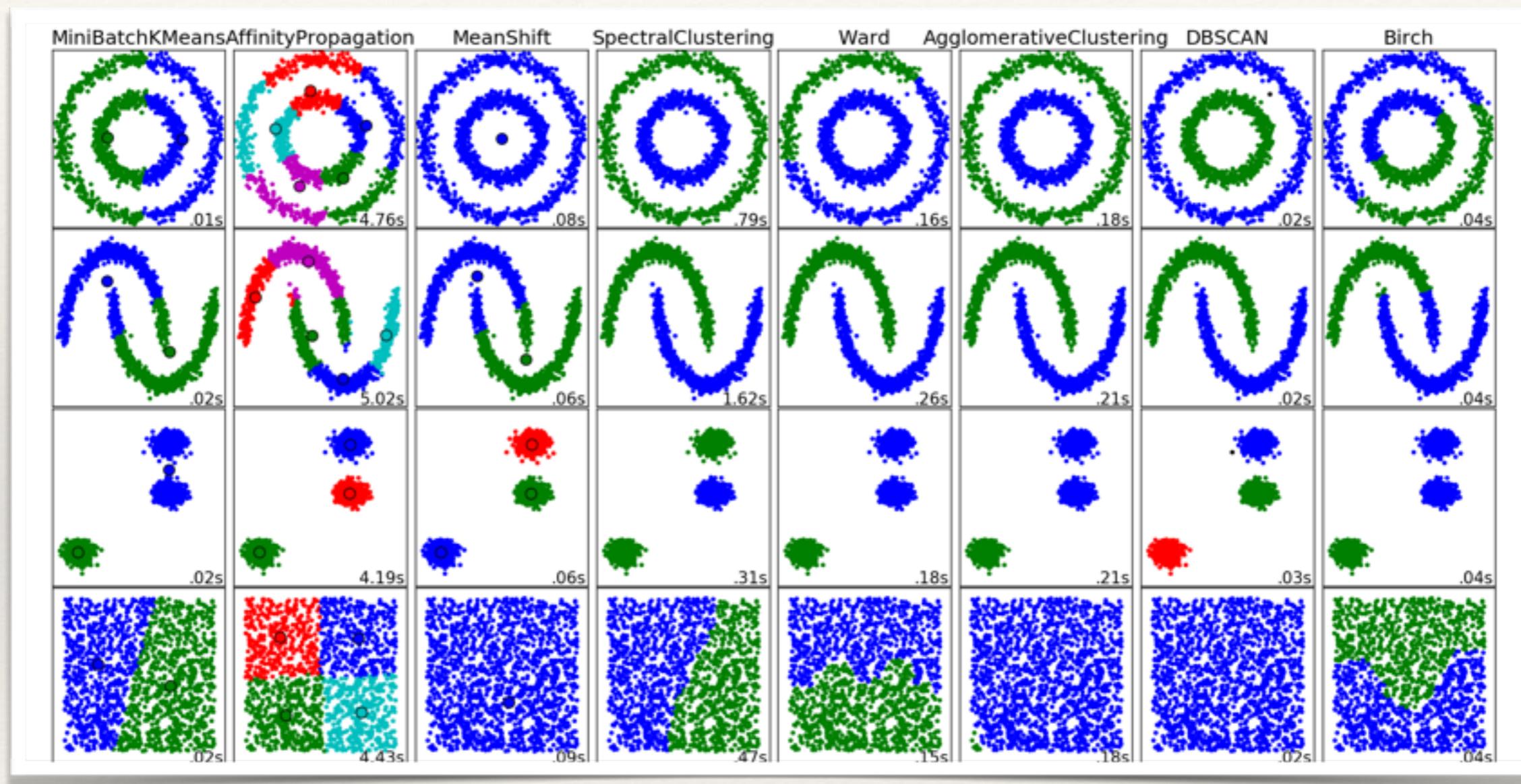
教師無しでクラスタリング

K-means clustering

- ❖ 単純なアルゴリズムながら、よく使われている方法の1つ
- ❖ アルゴリズムのパラメータについて学習できる
- ❖ http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html#example-cluster-plot-cluster-iris-py



Comparing different clustering algorithms on toy datasets



クラスタリングのアルゴリズムって沢山あるなーというのが分かる

教師有り学習

予測モデルを作りたい

Digits Classification Exercise

- ❖ k-近傍法とロジスティック回帰を使ったモデルの作成と予測
- ❖ 珍しくシンプルなコードなので、理解しやすい
- ❖ http://scikit-learn.org/stable/auto_examples/exercises/digits_classification_exercise.html#example-exercises-digits-classification-exercise-py

Digits Classification Exercise

A tutorial exercise regarding the use of classification techniques on the Digits dataset.

This exercise is used in the [Classification](#) part of the [Supervised learning: predicting an output variable from high-dimensional observations](#) section of the [A tutorial on statistical-learning for scientific data processing](#).

Python source code: [digits_classification_exercise.py](#)

```
print(__doc__)

from sklearn import datasets, neighbors, linear_model

digits = datasets.load_digits()
X_digits = digits.data
y_digits = digits.target

n_samples = len(X_digits)

X_train = X_digits[:.9 * n_samples]
y_train = y_digits[:.9 * n_samples]
X_test = X_digits[.9 * n_samples:]
y_test = y_digits[.9 * n_samples:]

knn = neighbors.KNeighborsClassifier()
logistic = linear_model.LogisticRegression()

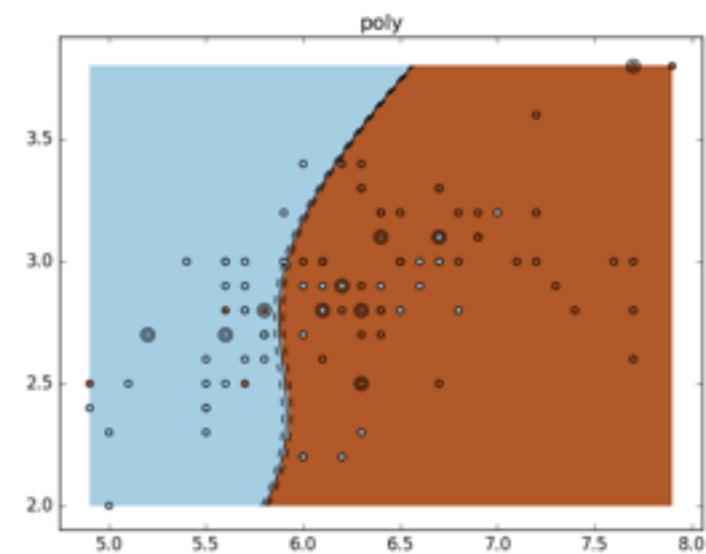
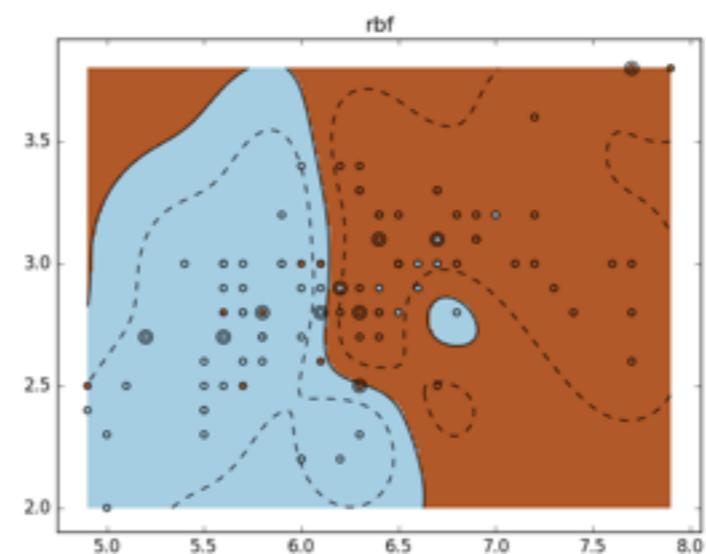
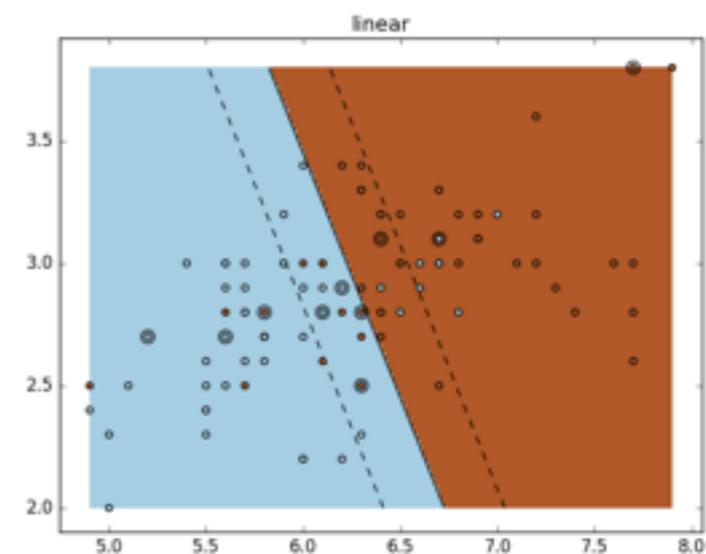
print('KNN score: %f' % knn.fit(X_train, y_train).score(X_test, y_test))
print('LogisticRegression score: %f' % logistic.fit(X_train, y_train).score(X_test, y_test))
```

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

SVM exercise

- ❖ Support Vector Machine(SVM)
- ❖ カーネルによる分離面の違いがわかるプロット付き

- ❖ http://scikit-learn.org/stable/auto_examples/exercises/plot_iris_exercise.html



まとめ

- ❖ scikit-learnのExampleから入門によさそうなものを紹介
- ❖ その他、さまざまなアルゴリズムが揃っている
- ❖ ドキュメントも豊富
 - ❖ 加藤先生が読み方伝授してくれるはず
- ❖ アルゴリズムの数学的な背景は別
- ❖ 使うだけなら、慣れれば結構簡単