PASONA TECH ウェルカムセミナー

Pythonが熱い！

データ解析から学ぶPython

辻 真吾（@tsjshg）

2016.4.27

自己紹介

- 1975年生まれ
- 東京大学先端科学技術研究センター 特任助教
  - 研究室は癌とゲノム
  - やっていることはPythonでデータ解析
- 昔はJavaでWeb開発とかしていましたが、10年ほど前から全部Python
- http://www.tsjshg.info/

みんなのPython勉強会

- connpassでイベント情報公開
  しています
  - http://
    startpython.connpass.com/
- 誰でも歓迎なPython勉強会
- 12回目の次回は5/10（火）
  19:00〜の予定です

Udemy

2016/5/31までの30%OFFクーポンはこのURLから

https://www.udemy.com/python-jp/?couponCode=pasona1605

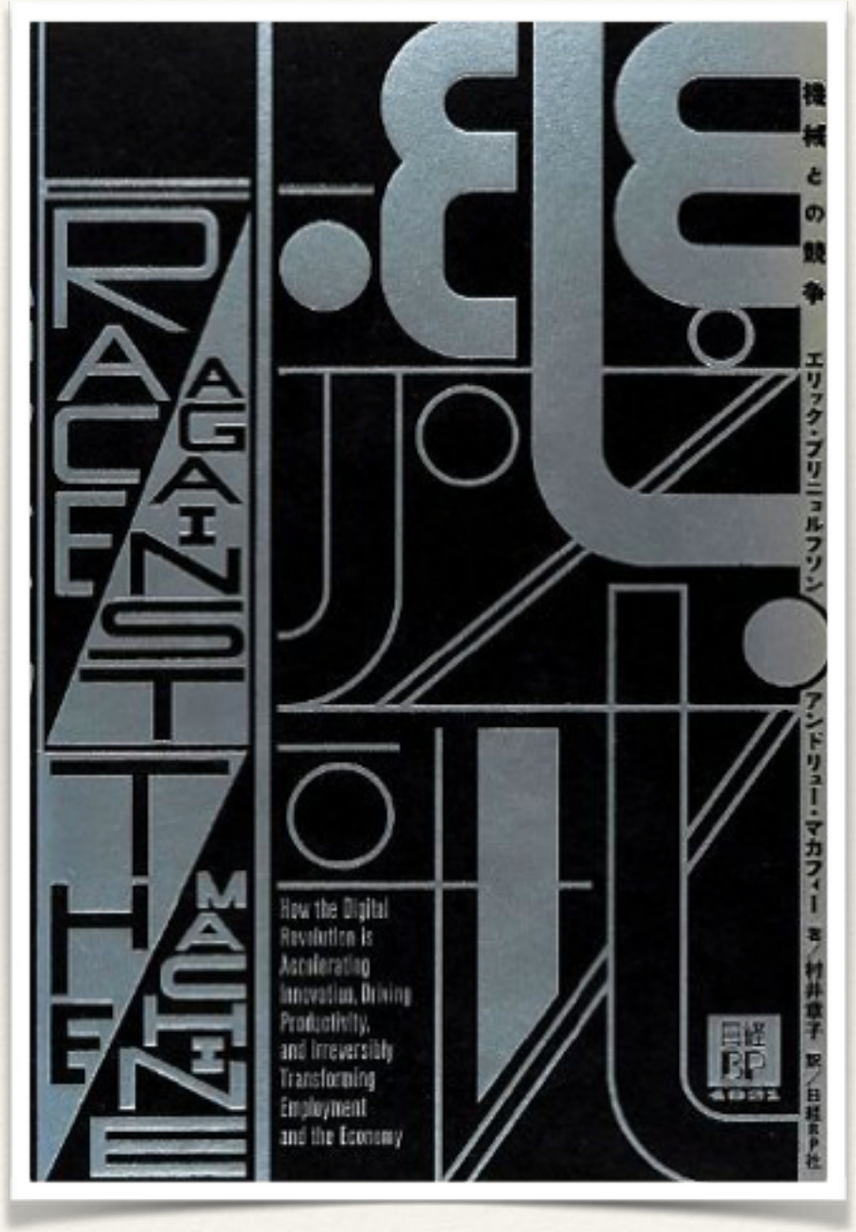
今日の構成

- 導入
- Pythonについて
  - Pythonには勢いがある
  - Pythonの環境構築
- データ解析をとりまく環境
- Pythonデータサイエンスのエコシステム
  - さまざまなライブラリ
  - 機械学習とPython
- 各種ライブラリの使い方を実際のコードを交えて

導入

テクノロジー失業説

ザ・セカンド・マシン・エイジ



機械との競争

この先、誰が生き残れるのか？

- 1998年、当時のチェス世界王者カスパロフがIBMのスーパーコンピュータ「ディープブルー」に敗北
- いま、チェス界はどうなっているのか？
- チェスには今、フリースタイルという分野がある
  - フリースタイルは何でもあり
  - 人だけ、コンピュータだけ、人＋コンピュータ

生き残るのは・・・

- もちろん、人＋コンピュータが最強
- ただ、その組み合わせは
- **チェスの名人**＋コンピュータではなく
- **コンピュータを上手に扱う人**＋コンピュータの組合せ
- コンピュータを使って大量のデータを上手に分析できる奴が生き残る！

ところで、

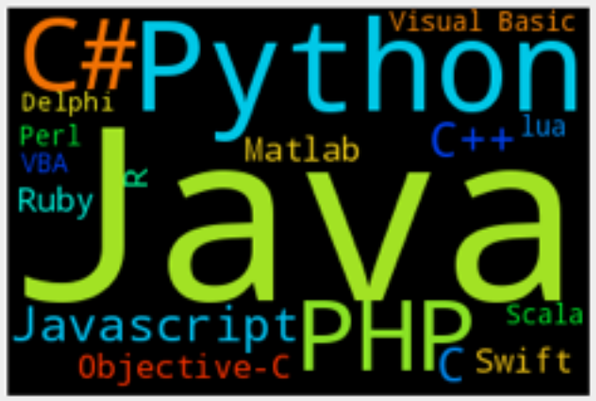
Pythonには勢いがある！

PYPL PopularitY of Programming Language

The PYPL PopularitY of Programming Language Index is created by analyzing how often language tutorials are searched on Google.

The more a language tutorial is searched, the more popular the language is assumed to be. It is a leading indicator. The raw data comes from Google Trends.

If you believe in collective wisdom, the PYPL Popularity of Programming Language index can help you decide which language to study, or which one to use in a new software project.



Worldwide, Apr 2016 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Java	24.0 %	-0.1 %
2	↑	Python	12.3 %	+1.7 %
3	↓	PHP	10.6 %	-1.0 %
4		C#	8.8 %	-0.2 %
5	↑↑	Javascript	7.5 %	+0.4 %
6	↓	C++	7.5 %	-0.3 %
7	↓	C	7.3 %	+0.2 %
8		Objective-C	4.8 %	-0.7 %
9	↑	R	3.1 %	+0.4 %
10	↑	Swift	3.0 %	+0.5 %
11	↓↓	Matlab	2.9 %	-0.1 %
12		Ruby	2.3 %	-0.3 %
13		Visual Basic	1.8 %	-0.4 %
14		VBA	1.5 %	+0.0 %
15		Perl	1.1 %	-0.2 %
16		Scala	0.9 %	+0.2 %
17	↑	lua	0.5 %	+0.1 %
18	↓	Delphi	0.4 %	-0.1 %

http://pypl.github.io/PYPL.html

TIOBE Index for April 2016

Apr 2016	Apr 2015	Change	Programming Language	Ratings	Change
1	1		Java	20.846%	+4.80%
2	2		C	13.905%	-1.84%
3	3		C++	5.918%	-1.04%
4	5	▲	C#	3.796%	-1.15%
5	8	▲	Python	3.330%	+0.64%
6	7	▲	PHP	2.994%	-0.02%
7	6	▼	JavaScript	2.566%	-0.73%
8	12	⇈	Perl	2.524%	+1.18%
9	18	⇈	Ruby	2.345%	+1.28%
10	10		Visual Basic .NET	2.273%	+0.15%

http://www.tiobe.com/tiobe_index

なぜ？

- There should be one-- and preferably only one --obvious way to do it.
  - import this で出てくる The Zen of Python
  - なにかするとき、選択肢が1つ、出来ればそれが唯一の方法である
    こと
- 他人（玄人）が書いたコードがあまり苦労無く読める
- バッテリー同梱（標準モジュールが充実）
- フリーのデータ解析環境として注目
- 機械学習分野では独壇場の立場を築きつつある

Pythonの環境構築

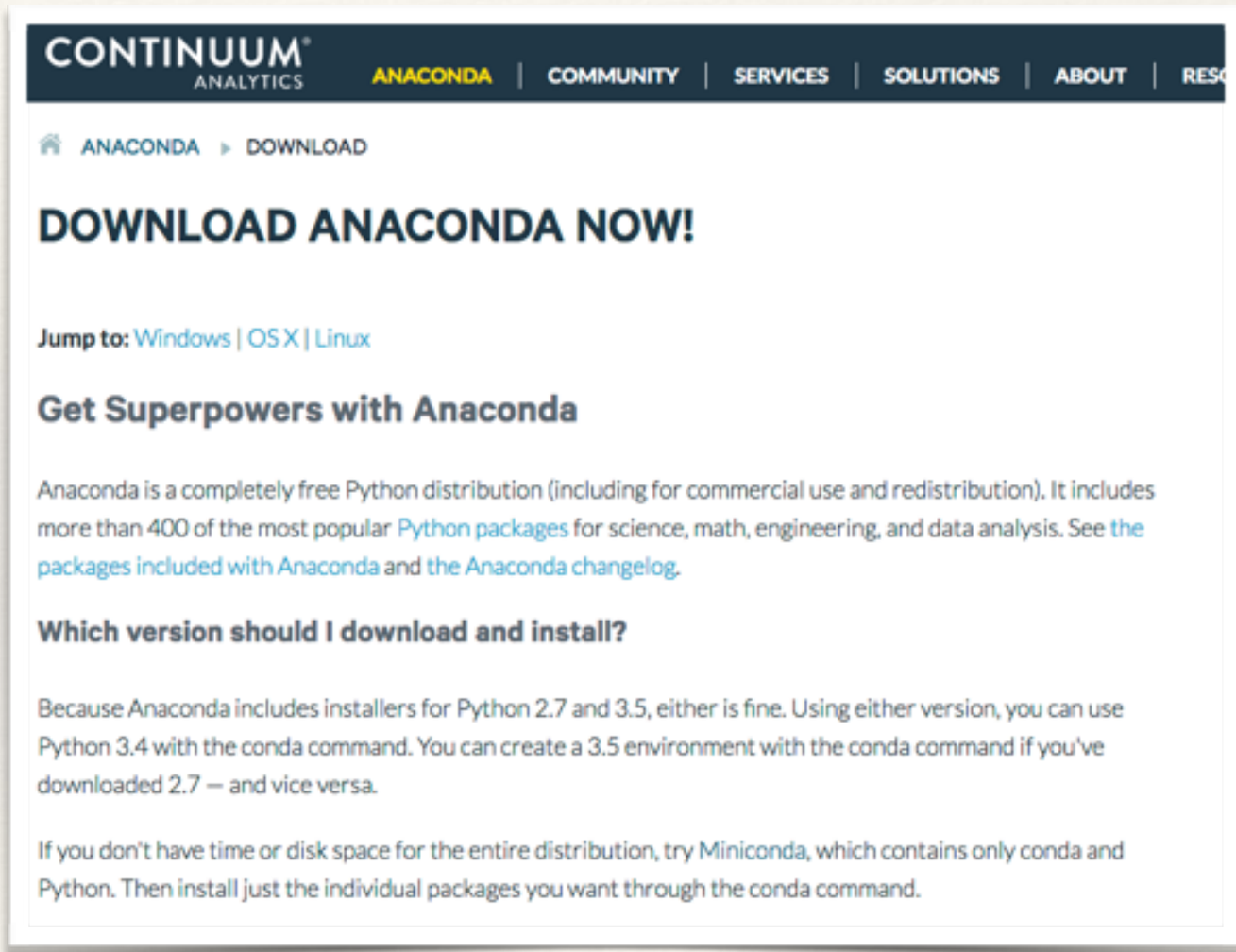
Pythonのインストール

- 2系と3系がある
  - 後方互換性がありません
- 特に理由がなければ3を
- MacOSXやLinuxのOSにはじめからインストールされているのが2系（なのが残念）

www.python.org

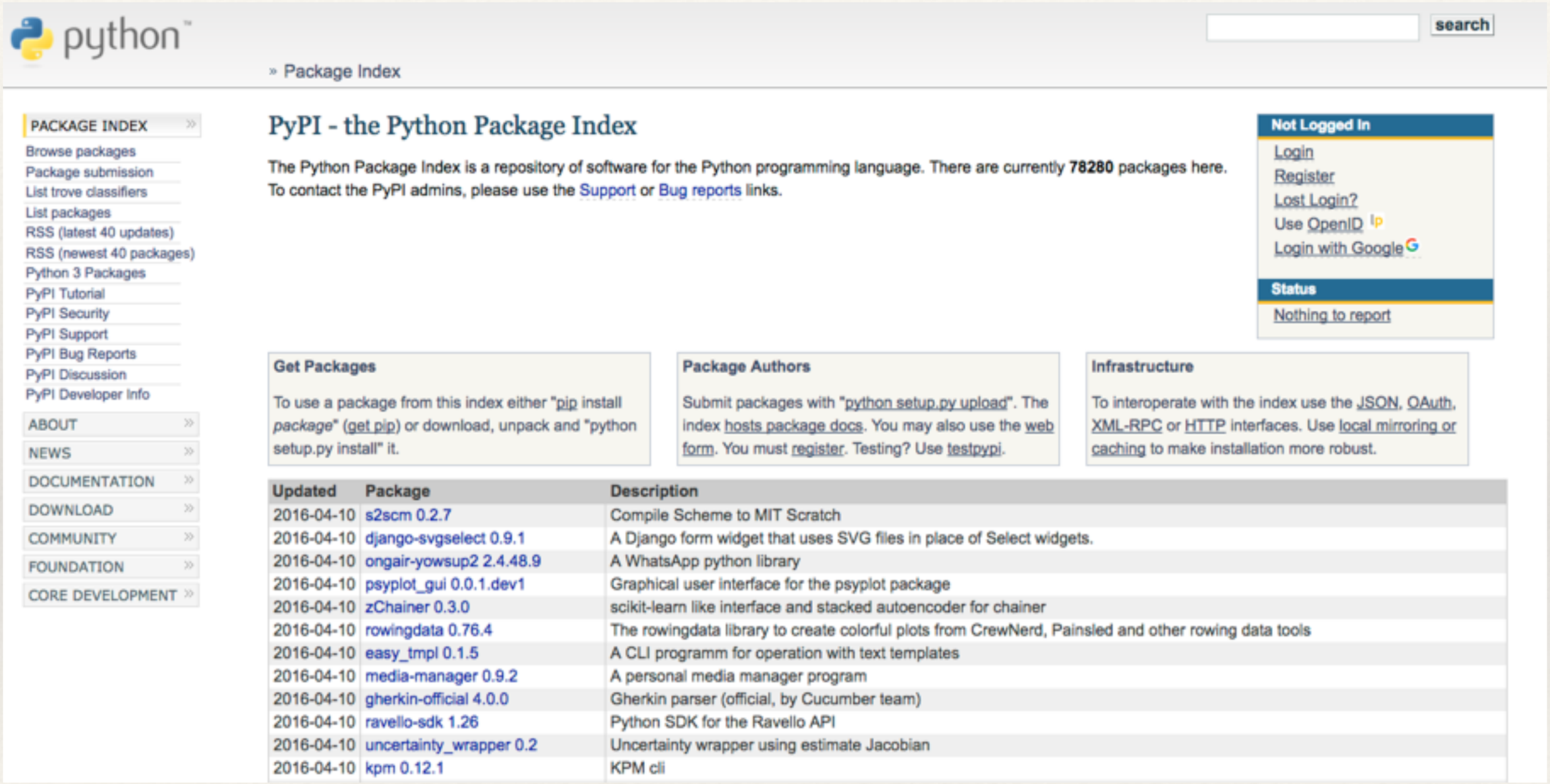
Anacondaがおすすめ！

- 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（メアド登録あり）
- データ解析環境の構築に最適



https://www.continuum.io

豊富な外部ライブラリ



外部モジュールが豊富 *78,280*個（*2016*年*4*月*10*日）

ライブラリの追加方法

ファイルをダウンロードして解凍後

$ python setup.py install

lib/python3.5/site-packages
のようなディレクトリにインストールされる

実際はpipが便利

- Python3.4から標準装備
- PyPIから自動ダウンロード
- 削除にも対応
- 使い方（OSのシェルで）
    - pip install -U django
        - -U or --upgradeで最新版を取得
    - pip freeze
        - いまの状態を表示

しかし・・・

- CやFORTRANで書かれたライブラリがその場でコンパイルされることがある
- Linuxは大丈夫
- MacOSXならXCodeのインストールで対応
- 開発環境を手軽に整えるのが難しいWindowsでこの問題は重い

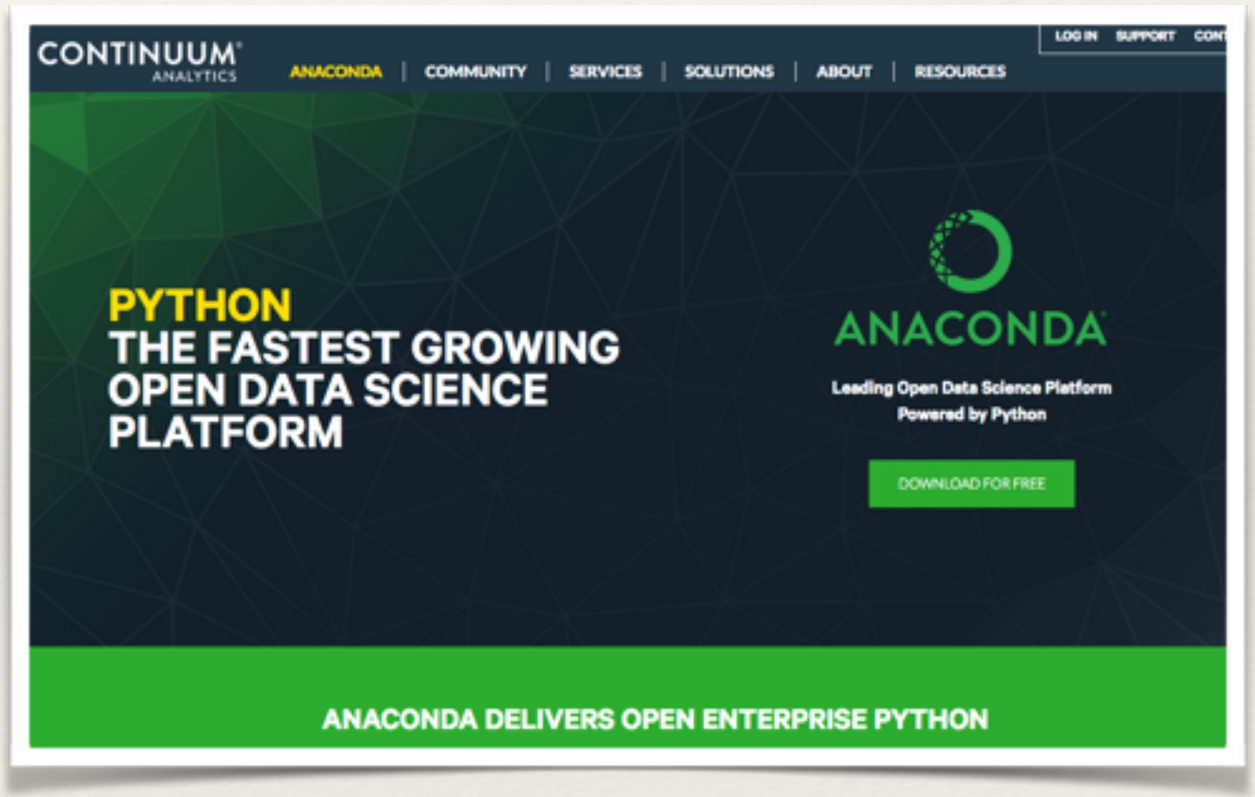
condaが便利

- $ conda install django
  - pipと同じような方法で、外部モジュールを追加
  - PyPIではなく、Continuum Analytics社のレポジトリに接続し、Windowsでもコンパイル済みのバイナリをとってきてくれる

まとめ

- Python + pip + conda+外部ライブラリ
- データサイエンスに必要な外部ライブラリはほとんど

  Anacondaに入っている！
- Continuum Analytics社のページからダウンロード可能です



https://www.continuum.io/

IPython (Jupyter) notebookが便利

[Demo] IPython (Jupyter) notebookが便利

- 高性能なPythonインタラクティブシェルIPython
  - コマンドラインでipython
- Webブラウザで利用できるIPython notebook
  - Jupyterという名前で、Python以外の言語も利用可能
  - コードの入力、実行、結果の表示、保存などができる
  - Web技術的にも先端的

データ解析をとりまく環境

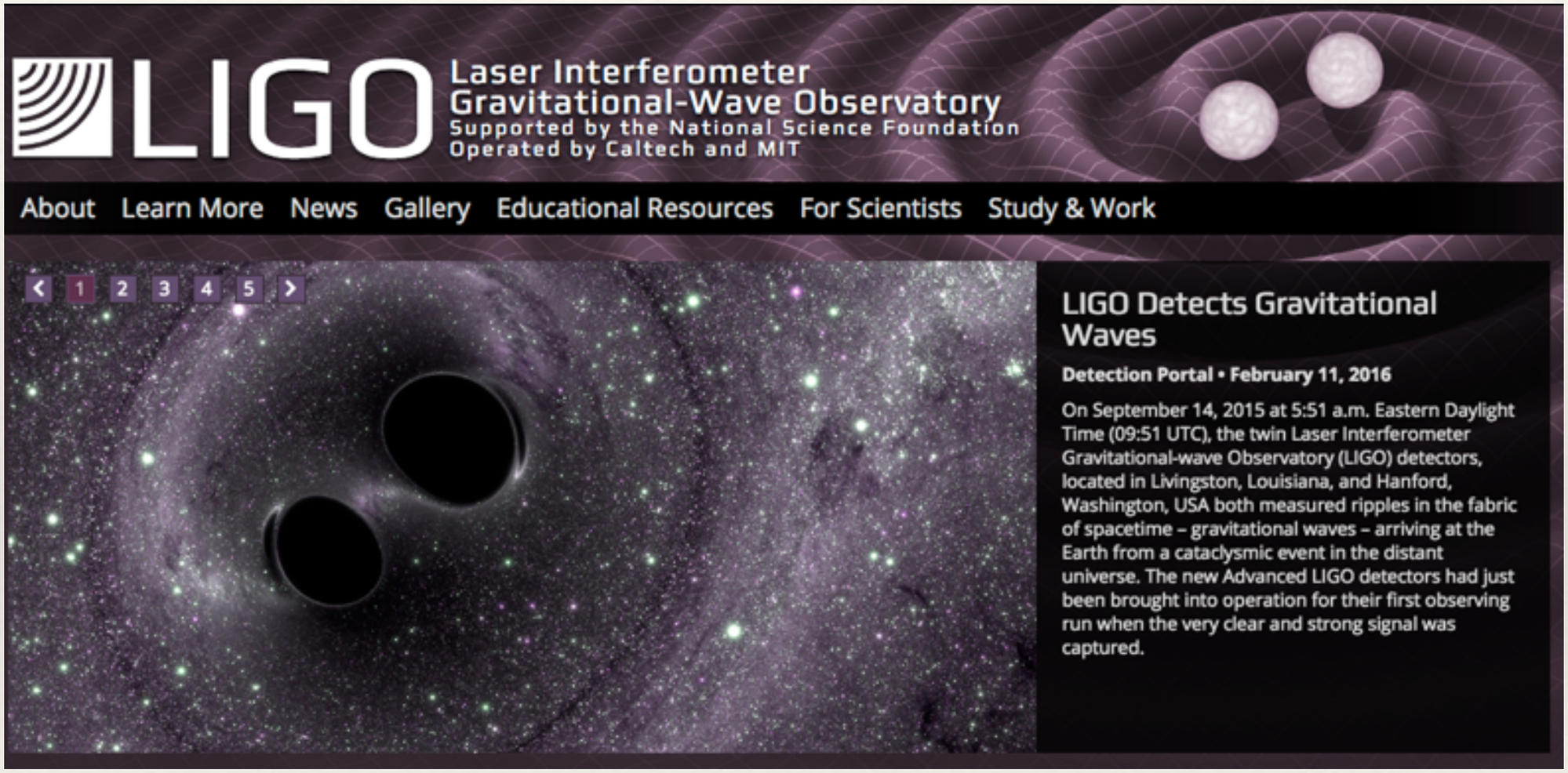
いくつかの選択肢

- SAS (www.sas.com)
  - BIツールの代表格
- Matlab (jp.mathworks.com)
  - 数値シミュレーションなどに利用されている
- R
  - RStudio（IDE）は便利
- Python

我田引水

	オープン？	汎用言語？
SAS	×	×
Matlab	×	×
R	○	×
Python	○	○

重力波を初めて観測



質量を持った物体が動くことで、時空のゆがみが波のように光速で伝わる現象を
LIGO（ライゴ：レーザー干渉計重力波天文台）で初観測

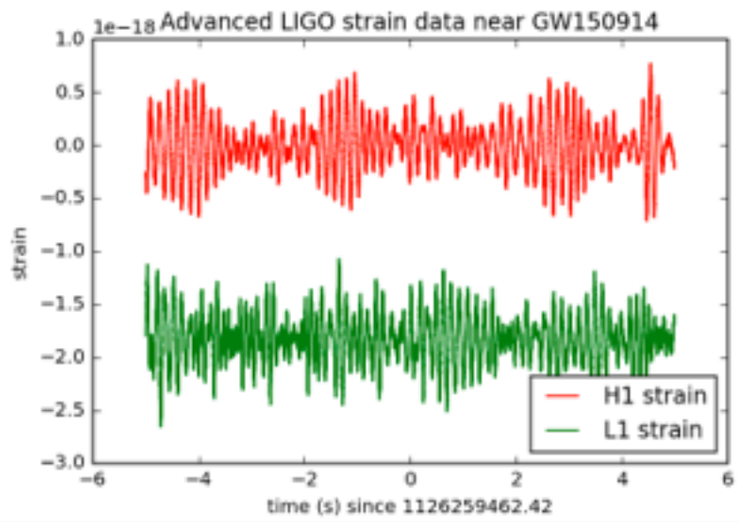
オープン＆汎用である事の利点

- LIGOではいたるところで Pythonが活躍
  - 機器の制御からデータ解析まで
- データはIPython notebookとともに公開されている
- 同じ解析を世界中の人が再現できる

```
H1 CBC_CAT2: len, min, mean, max =  32 1 1.0 1
L1     DATA: len, min, mean, max =  32 1 1.0 1
L1 CBC_CAT1: len, min, mean, max =  32 1 1.0 1
L1 CBC_CAT2: len, min, mean, max =  32 1 1.0 1
In both H1 and L1, all 32 seconds of data are present (DATA=1),
and all pass data quality (CBC_CAT1=1 and CBC_CAT2=1).
```

```
# plot +- 5 seconds around the event:
tevent = 1126259462.422       # Mon Sep 14 09:50:45 GMT 2015
deltat = 5.                   # seconds around the event
# index into the strain time series for this time interval:
indxt = np.where((time_H1 >= tevent-deltat) & (time_H1 < tevent+deltat))
```

```
plt.figure()
plt.plot(time_H1[indxt]-tevent,strain_H1[indxt],'r',label='H1 strain')
plt.plot(time_L1[indxt]-tevent,strain_L1[indxt],'g',label='L1 strain')
plt.xlabel('time (s) since '+str(tevent))
plt.ylabel('strain')
plt.legend(loc='lower right')
plt.title('Advanced LIGO strain data near GW150914')
plt.savefig('GW150914_strain.png')
```



https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html

Rの人気

KDnuggets

KDnuggets™ **Data Mining, Analytics, Big Data, and Data Science**

Subscribe to **KDnuggets News** | Follow | **Contact**

search KDnuggets Search

Data Mining Software | News | Top stories | Opinions | Tutorials | Jobs | Academic | Companies | Courses | Datasets | Education | Meetings | Polls | Webinars

Latest News, Stories

- 5,000 KDnuggets Posts – Examining Our Most Popul...
- Microsoft is Becoming M(ai)crosoft
- Top stories for Apr 18-24: Top 15 Machine Learning Fra...
- Schwab: Director, Institutional Analytics
- How machine learning is making bots more human

R, Python users show surprising stability, but strong regional differences

◀ Previous post Next post ▶

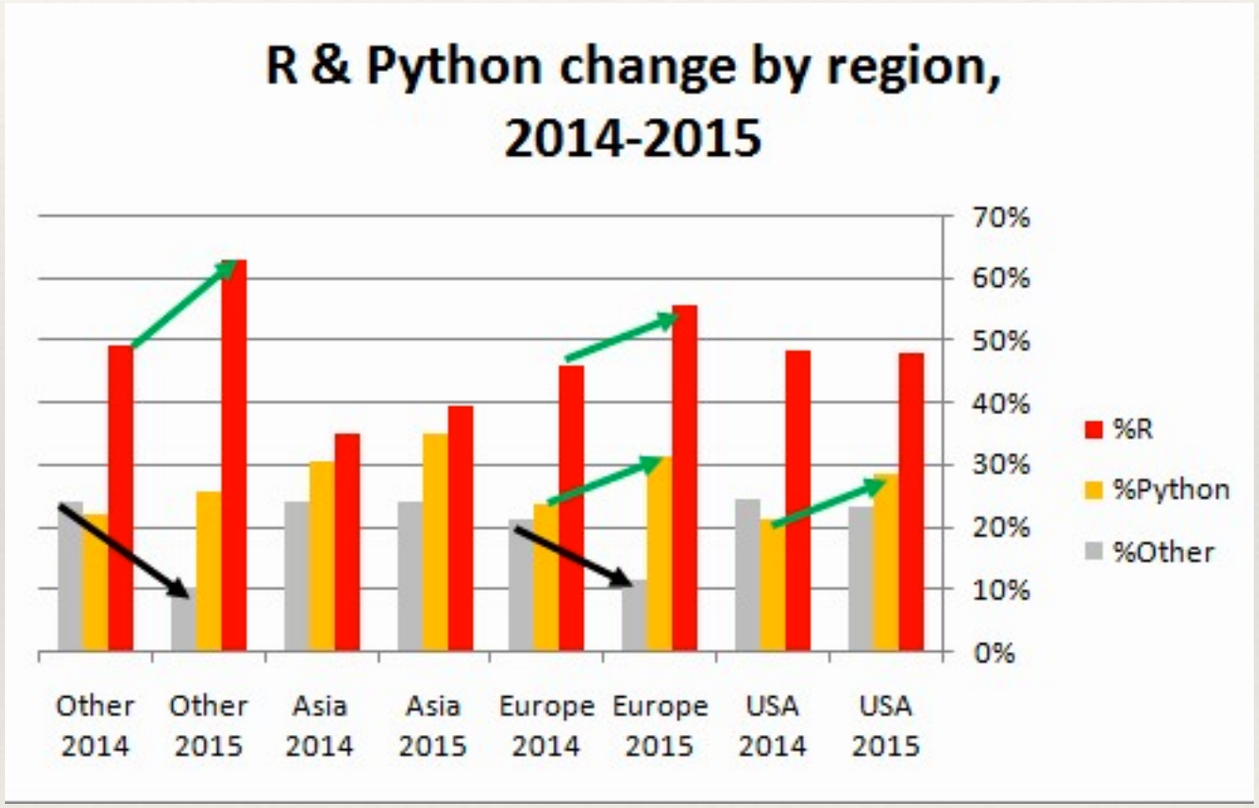
Tags: Poll, Python, Python vs R, R

R remains the dominant language, with Python slowly catching up, and other languages shrinking. We also found surprising stability, with about 90% of R and Python users staying with that language, and strong regional differences.

By Gregory Piatetsky, KDnuggets

RとPythonの比較

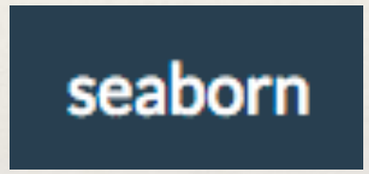
- RとPythonは全世界的に伸びている
  - データサイエンス自体の拡大が要因か？
- USAでのRの伸びに注目
  - 鈍化している
- この流れが、次に来る（かも）



http://www.kdnuggets.com/2015/07/poll-primary-analytics-language-r-python.html

Pythonデータサイエンスの
エコシステム

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

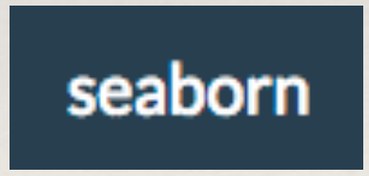
- データ解析なくてはならない超高性能ライブラリ
- エクセルのシートをイメージすると分かり易いかも
- データの入出力、加工、可視化など幅広く対応

matplotlib, seaborn

- データの可視化に使われる

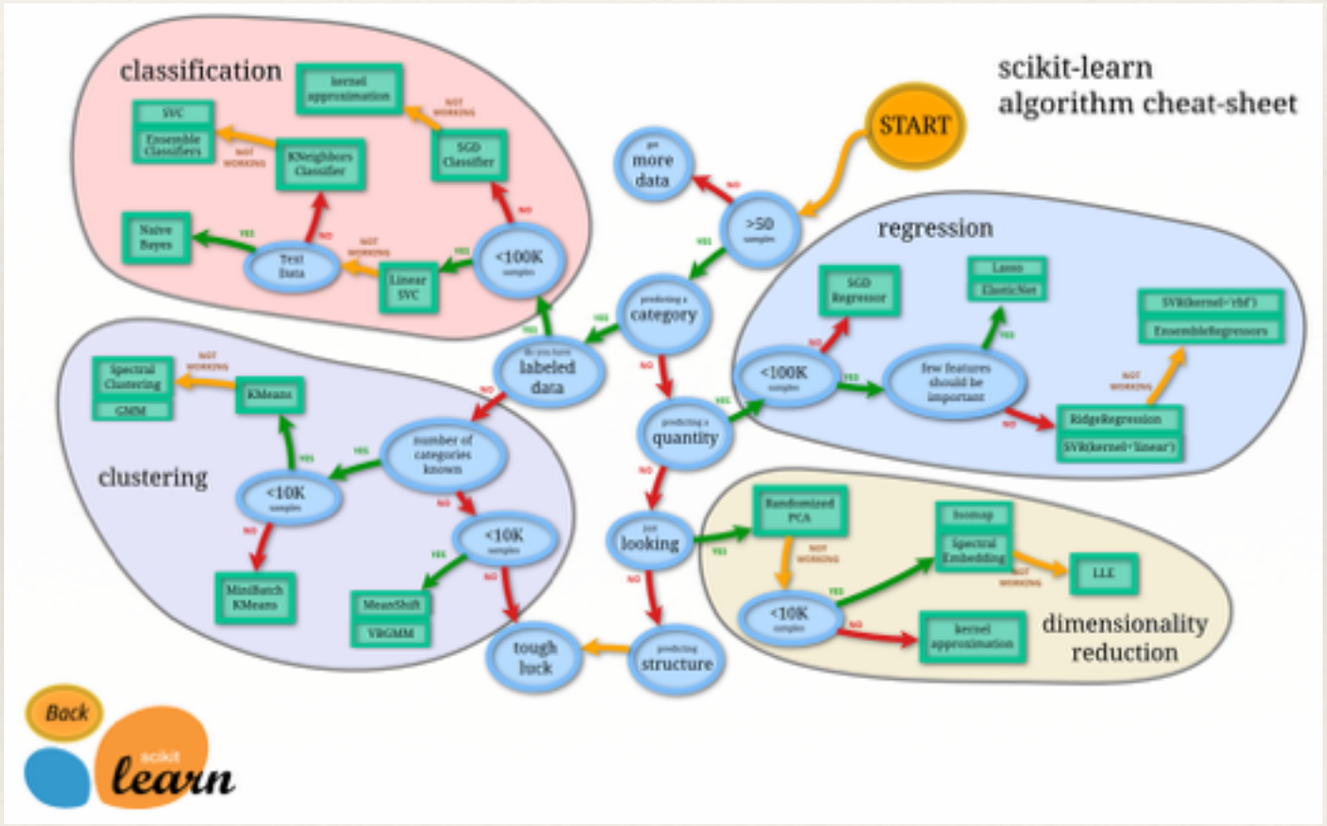
- matplotlibはmatlabの代替を意識

- 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



scikit-learn

- 進化し続けるPythonの機械学習ライブラリ
- 豊富なドキュメントとコード例で、利用するには便利
- cheat sheetなどアルゴリズムを学べる資料も多数



chainer

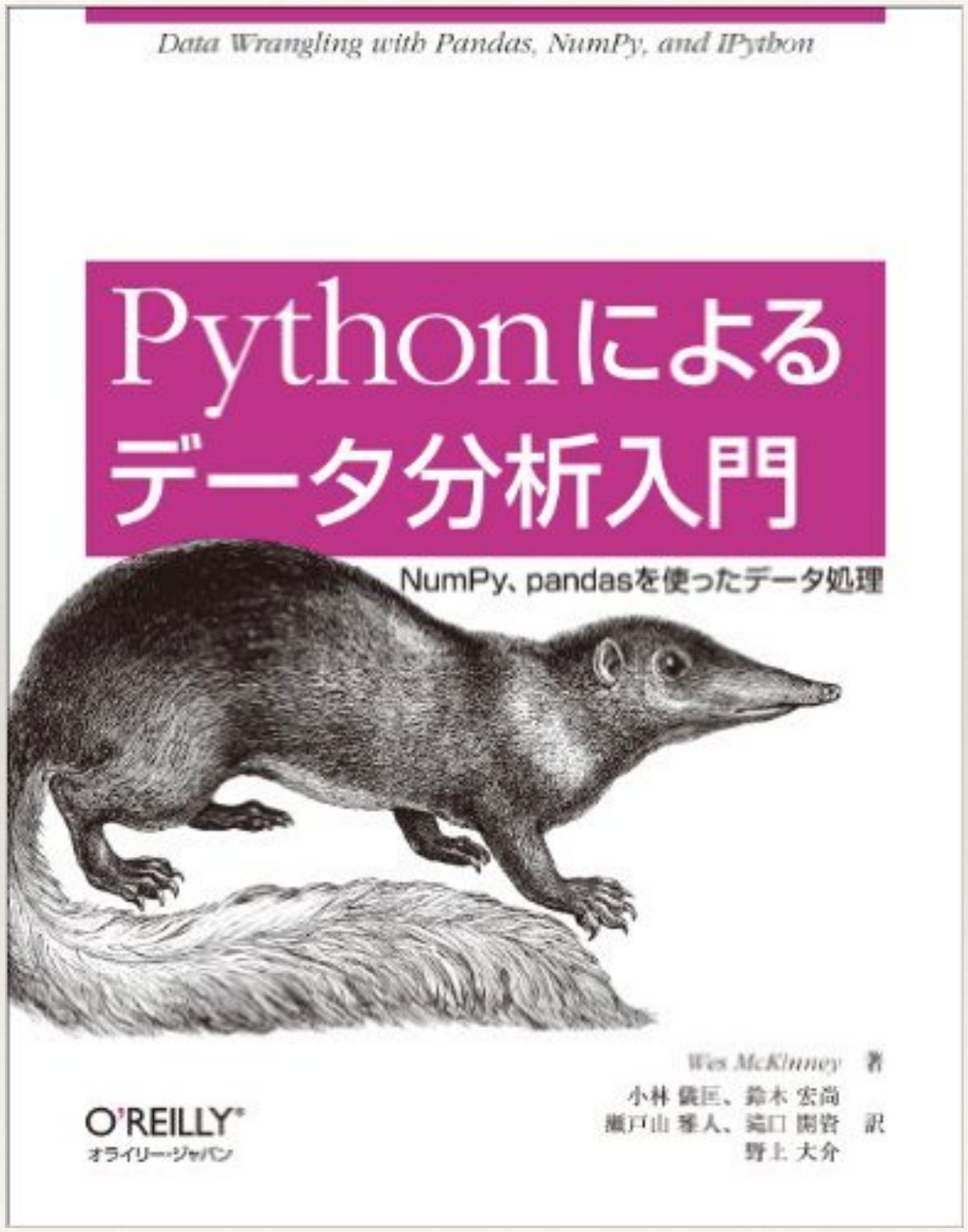
- Preferred Networksが開発する
  Deep Learning用のライブラリ
- 利用しやすい
- GPUの計算、マルチGPUにも
  対応

休憩のあと

各ライブラリをデモします

参考文献など

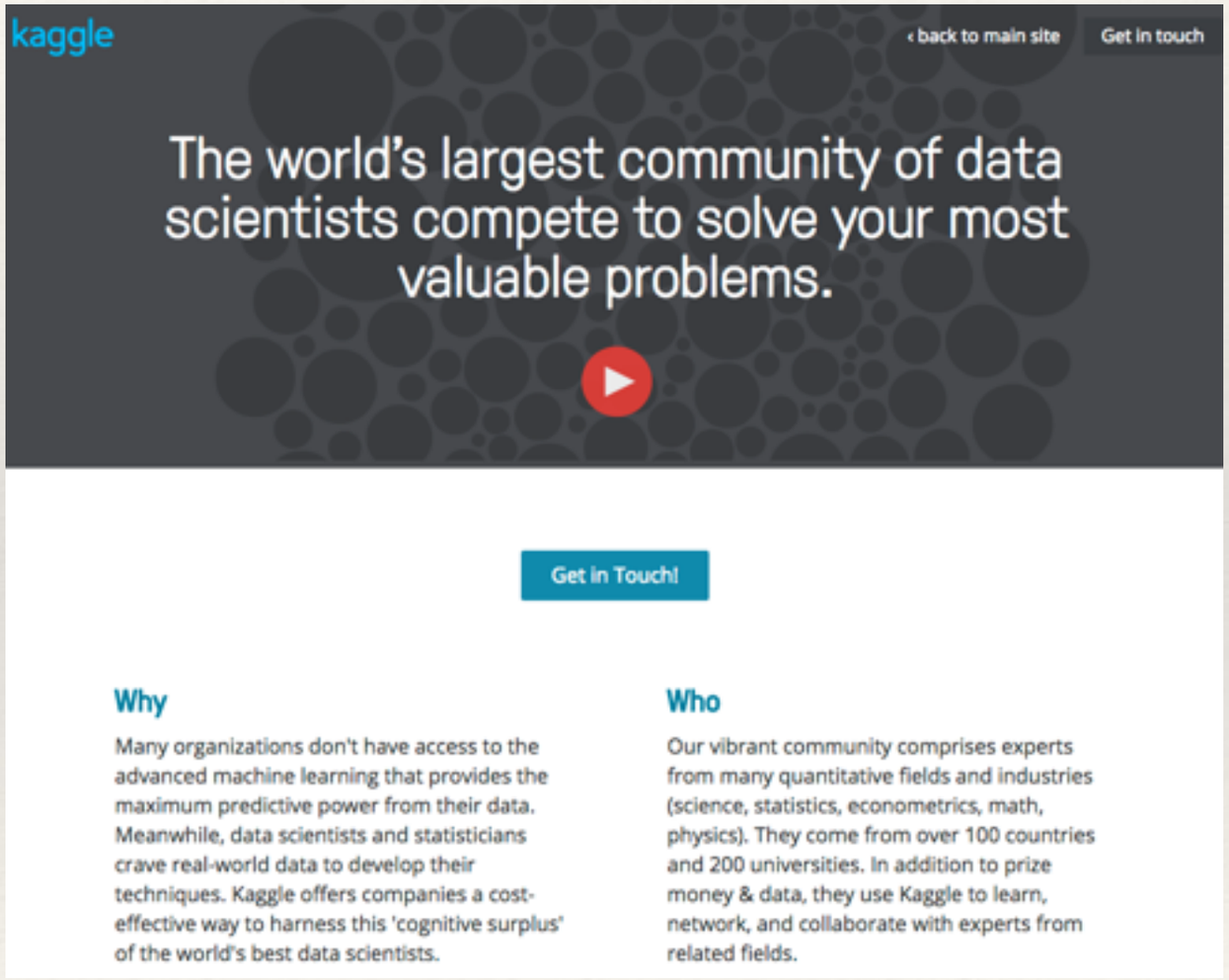
- pandasの作者であるWes McKinneyの著書
- 分量多いですが、網羅的に分かります



- もちろんPython
- 少し高度ですが、実際のデータと豊富なコード例が参考になります

kaggle

- 豊富なサンプルデータ
- 機械学習の性能を競い合える



まとめ

- これからは、データを上手に解析できる人が有利
- データ解析のソフトウェアにはいくつかの選択肢
- Pythonとそれを取り巻くエコシステムは、オープンなデータサイエンスプラットフォームのなかで非常に強力
- 是非、Pythonでデータサイエンスを！

ありがとうございました