

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

The screenshot shows the event page for 'みんなのPython勉強会#12' on the connpass platform. The event is scheduled for May 10th (Tuesday) from 19:00 to 21:00. The ticket price is 1000 yen (including venue fee) for the general public, free for students, and free for staff/instructors. The event is organized by Start Python Club. The page also features a 'Start Python Club' group profile with 12 events and 279 members, and a '会場' (venue) section listing 'クリーク・アンド・リバー' in Tokyo.

募集内容	料金	先着順
一般	1000円 (会場払い)	80/100人
学生	無料	7/20人
スタッフ・講師 (関係者限定)	無料	2/10人

# Udemy

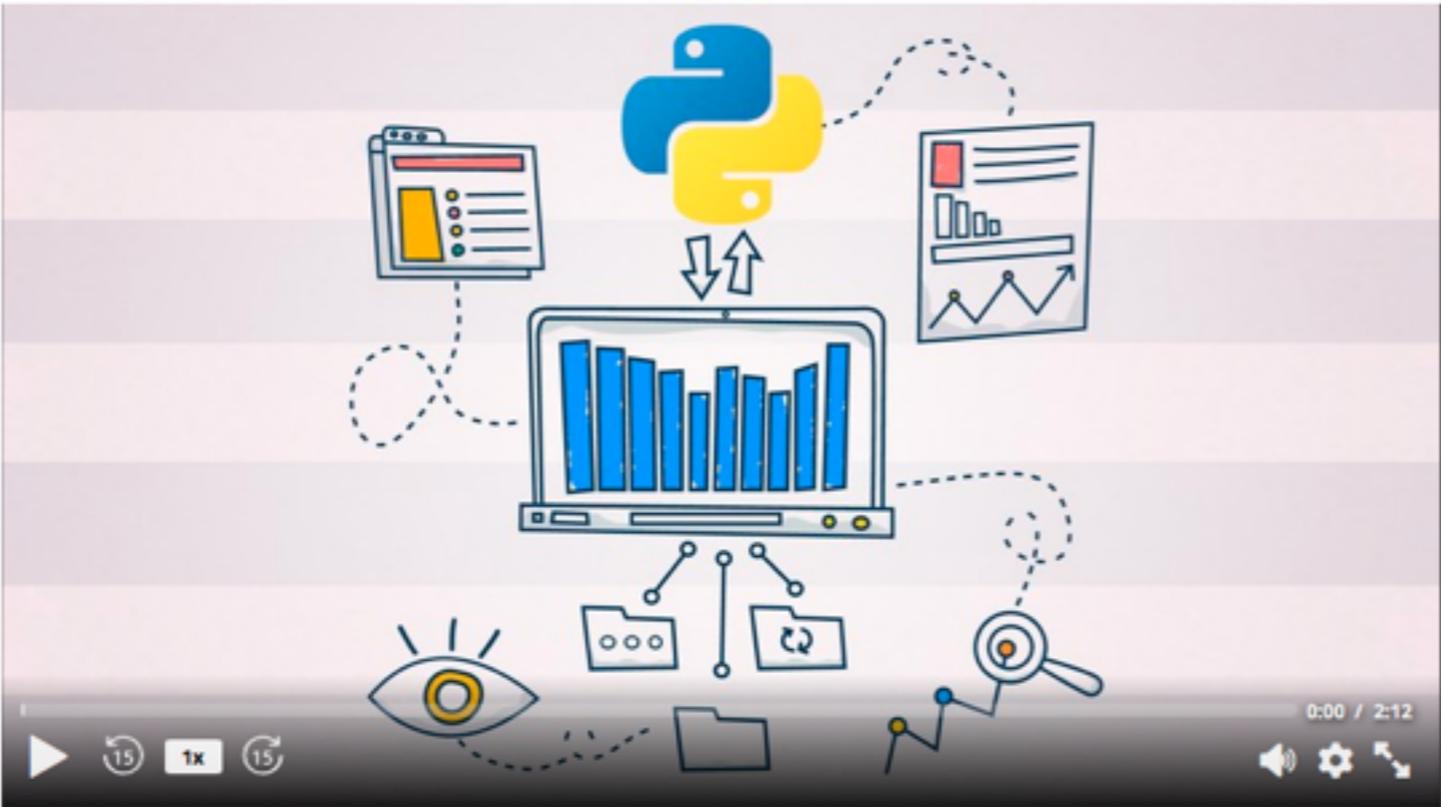
コース検索  **udemy** 講師になりたい方へ

## 【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](#)

2016/5/31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

---

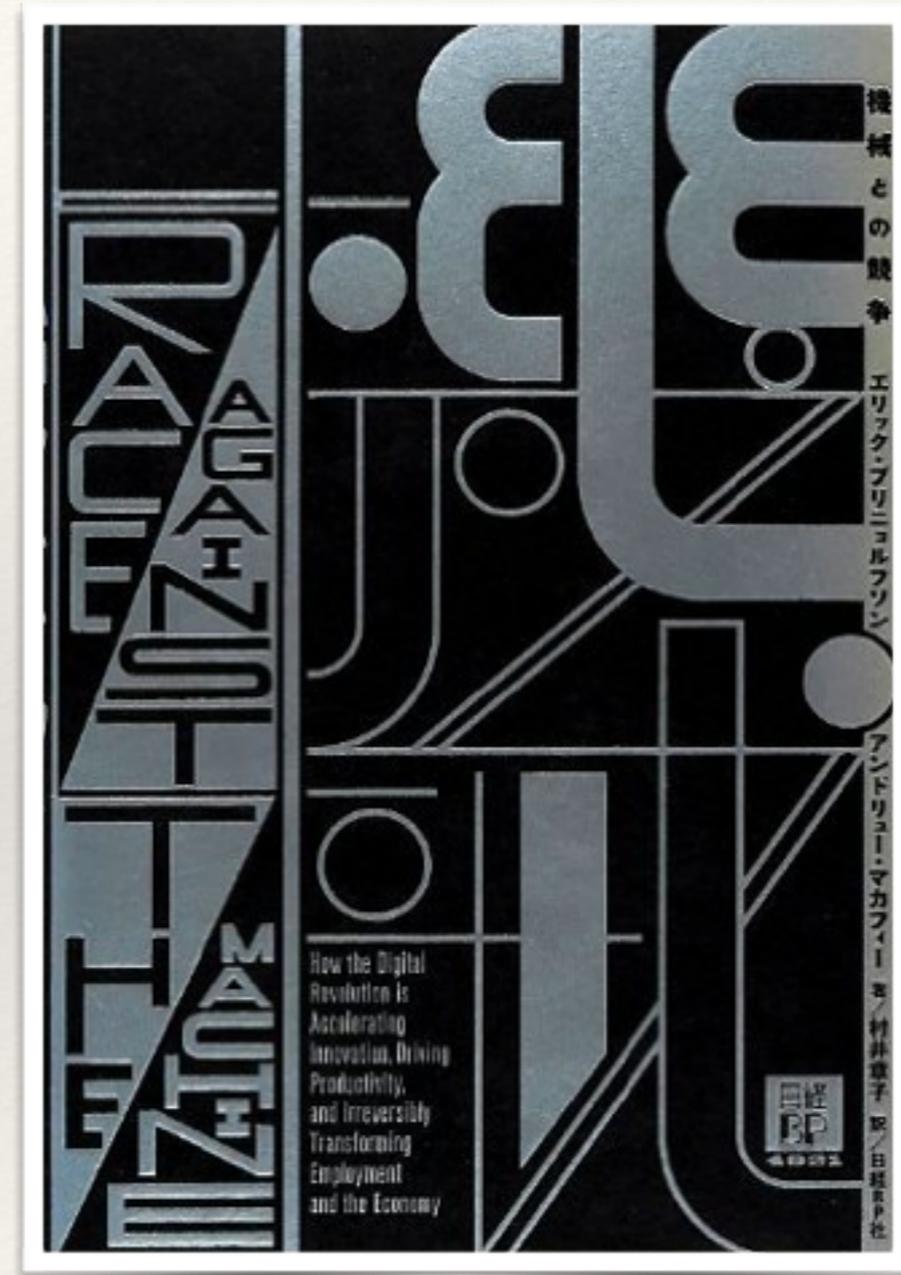
- ❖ 導入
- ❖ Pythonについて
  - ❖ Pythonには勢いがある
  - ❖ Pythonの環境構築
- ❖ データ解析をとりまく環境
- ❖ Pythonデータサイエンスのエコシステム
  - ❖ さまざまなライブラリ
  - ❖ 機械学習とPython
- ❖ 各種ライブラリの使い方を実際のコードを交えて

# 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、

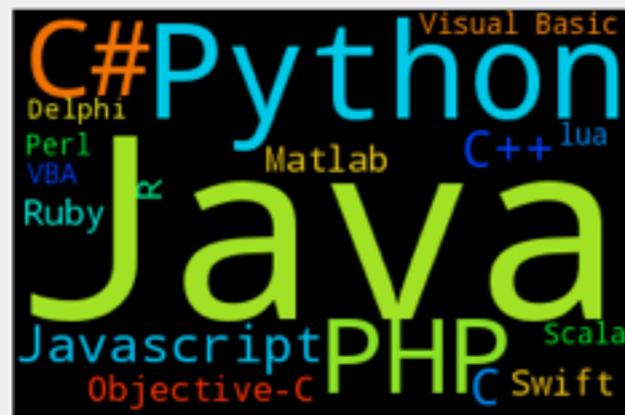
Pythonには勢いがある！

# 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 %

© Pierre Carbonnelle, 2015

# 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	v	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](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勞無く読める
- ❖ 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

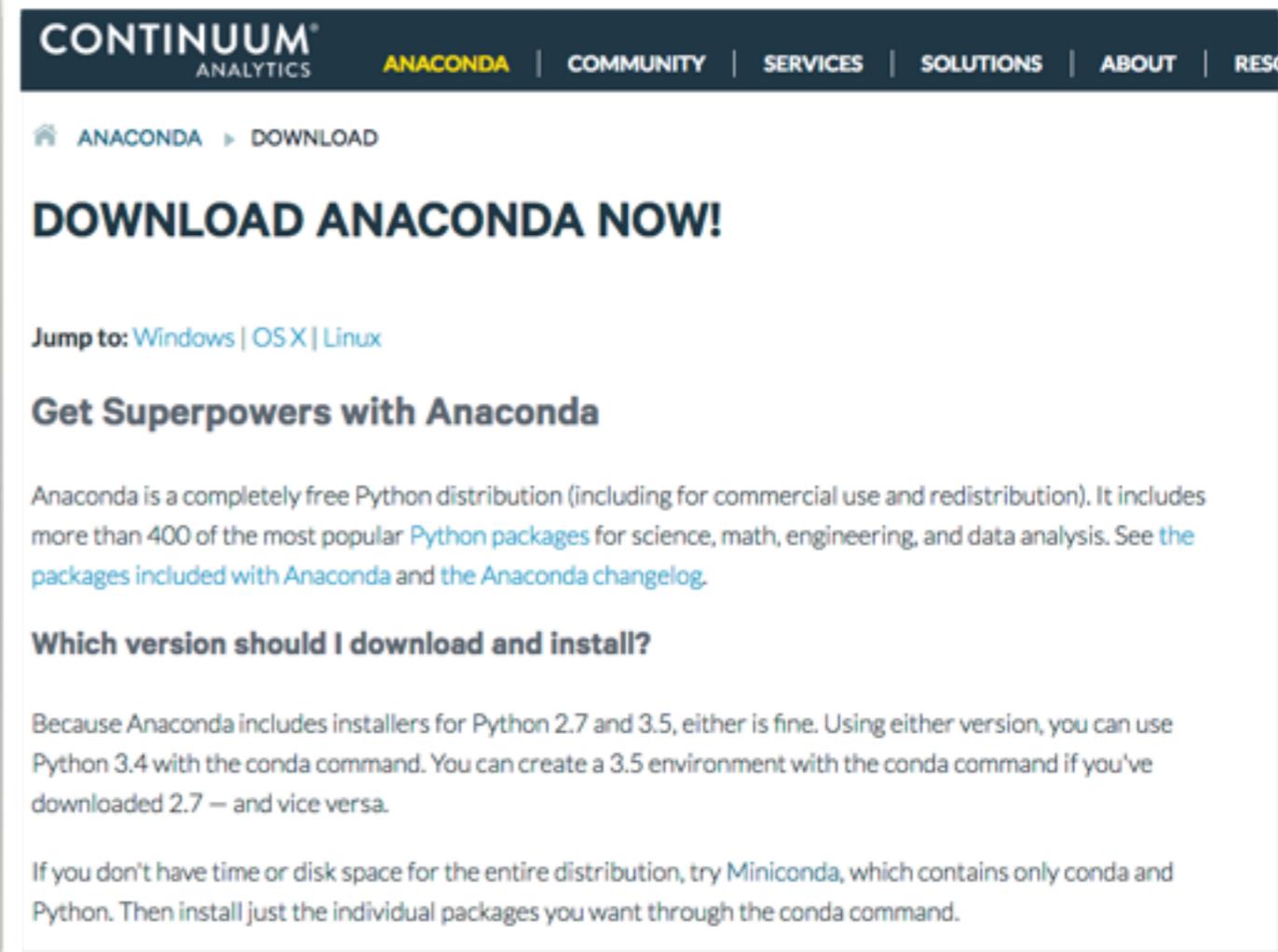
- ❖ 2系と3系がある
  - ❖ 後方互換性がありません
- ❖ 特に理由がなければ3を
- ❖ MacOSXやLinuxのOSにはじめてからインストールされているのが2系（なのが残念）



[www.python.org](http://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（メールアドレスあり）
- ❖ データ解析環境の構築に最適



The screenshot shows the 'ANACONDA' download page. At the top, there is a navigation bar with 'CONTINUUM ANALYTICS' and links for 'ANACONDA', 'COMMUNITY', 'SERVICES', 'SOLUTIONS', 'ABOUT', and 'RESOURCES'. Below the navigation bar, the page title is 'ANACONDA > DOWNLOAD'. The main heading is 'DOWNLOAD ANACONDA NOW!'. There are links to 'Jump to: Windows | OSX | Linux'. The section 'Get Superpowers with Anaconda' describes Anaconda as a free Python distribution with over 400 packages. The section 'Which version should I download and install?' explains that both Python 2.7 and 3.5 installers are available and compatible. At the bottom, it mentions 'Miniconda' as a smaller alternative.

<https://www.continuum.io>

# 豊富な外部ライブラリ

The screenshot shows the PyPI website interface. At the top left is the Python logo and the text "python™". To the right is a search bar with the word "search" next to it. Below the logo is the text "» Package Index".

On the left side, there is a navigation menu with the following items:

- PACKAGE INDEX »
- Browse packages
- Package submission
- List trove classifiers
- List packages
- RSS (latest 40 updates)
- RSS (newest 40 packages)
- Python 3 Packages
- PyPI Tutorial
- PyPI Security
- PyPI Support
- PyPI Bug Reports
- PyPI Discussion
- PyPI Developer Info
- ABOUT »
- NEWS »
- DOCUMENTATION »
- DOWNLOAD »
- COMMUNITY »
- FOUNDATION »
- CORE DEVELOPMENT »

The main content area is titled "PyPI - the Python Package Index". Below the title, it says: "The Python Package Index is a repository of software for the Python programming language. There are currently **78280** packages here. To contact the PyPI admins, please use the [Support](#) or [Bug reports](#) links."

On the right side, there is a "Not Logged In" box with the following links: [Login](#), [Register](#), [Lost Login?](#), [Use OpenID](#), and [Login with Google](#). Below this is a "Status" box with the text "Nothing to report".

Below the main text, there are three boxes:

- Get Packages**: To use a package from this index either "pip install package" (get pip) or download, unpack and "python setup.py install" it.
- Package Authors**: Submit packages with "python setup.py upload". The index [hosts package docs](#). You may also use the [web form](#). You must [register](#). Testing? Use [testpypi](#).
- Infrastructure**: To interoperate with the index use the [JSON](#), [OAuth](#), [XML-RPC](#) or [HTTP](#) interfaces. Use [local mirroring](#) or [caching](#) to make installation more robust.

At the bottom, there is a table of recently updated packages:

Updated	Package	Description
2016-04-10	<a href="#">s2scm 0.2.7</a>	Compile Scheme to MIT Scratch
2016-04-10	<a href="#">django-svgselect 0.9.1</a>	A Django form widget that uses SVG files in place of Select widgets.
2016-04-10	<a href="#">ongair-yowsup2 2.4.48.9</a>	A WhatsApp python library
2016-04-10	<a href="#">psyplot_gui 0.0.1.dev1</a>	Graphical user interface for the psyplot package
2016-04-10	<a href="#">zChainer 0.3.0</a>	scikit-learn like interface and stacked autoencoder for chainer
2016-04-10	<a href="#">rowingdata 0.76.4</a>	The rowingdata library to create colorful plots from CrewNerd, Painsled and other rowing data tools
2016-04-10	<a href="#">easy_tmpl 0.1.5</a>	A CLI program for operation with text templates
2016-04-10	<a href="#">media-manager 0.9.2</a>	A personal media manager program
2016-04-10	<a href="#">gherkin-official 4.0.0</a>	Gherkin parser (official, by Cucumber team)
2016-04-10	<a href="#">ravello-sdk 1.26</a>	Python SDK for the Ravello API
2016-04-10	<a href="#">uncertainty_wrapper 0.2</a>	Uncertainty wrapper using estimate Jacobian
2016-04-10	<a href="#">kpm 0.12.1</a>	KPM cli

外部モジュールが豊富 78,280個 (2016年4月10日)

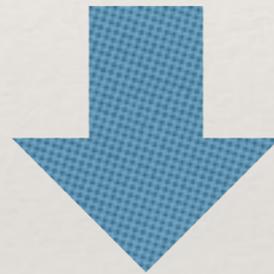
---

# 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

---

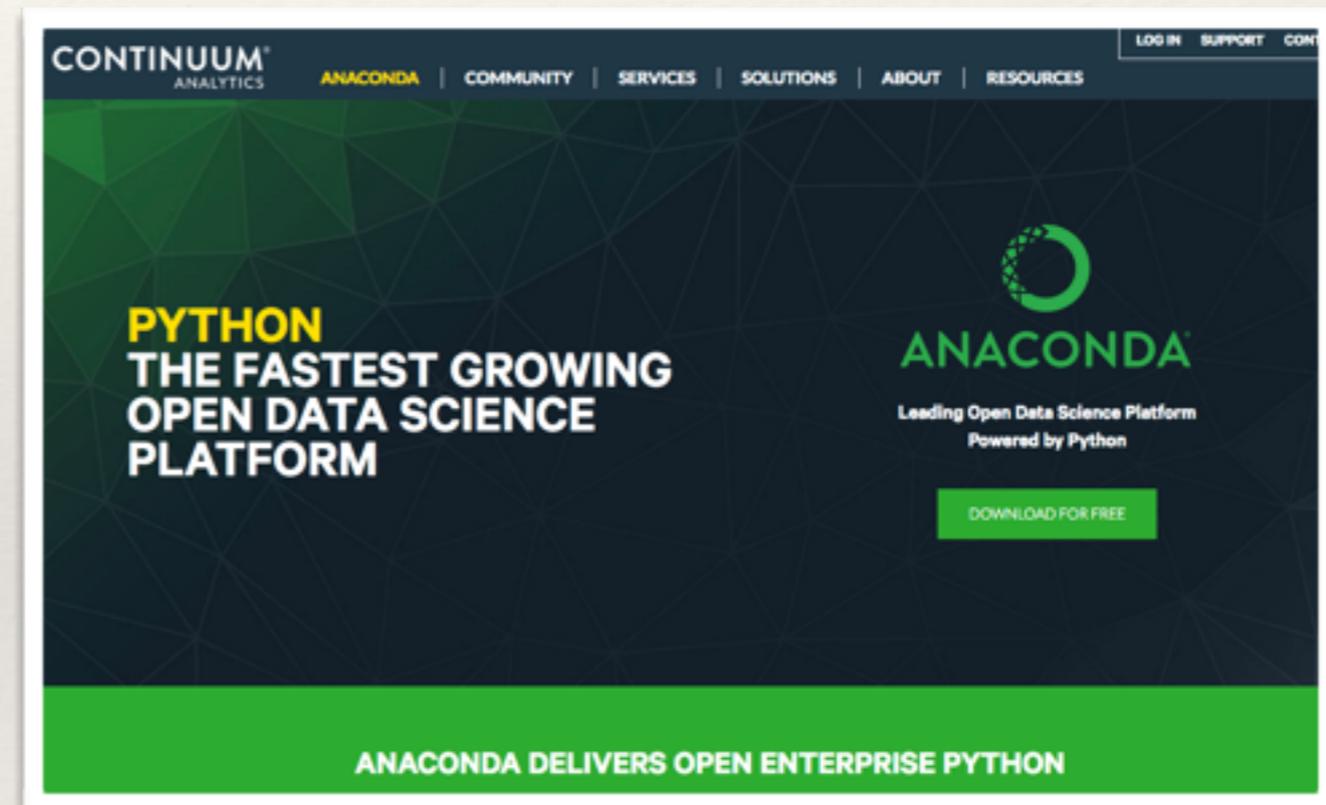
# condaが便利

---

- ❖ `$ 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](http://www.sas.com))
  - ❖ BIツールの代表格
- ❖ Matlab ([jp.mathworks.com](http://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

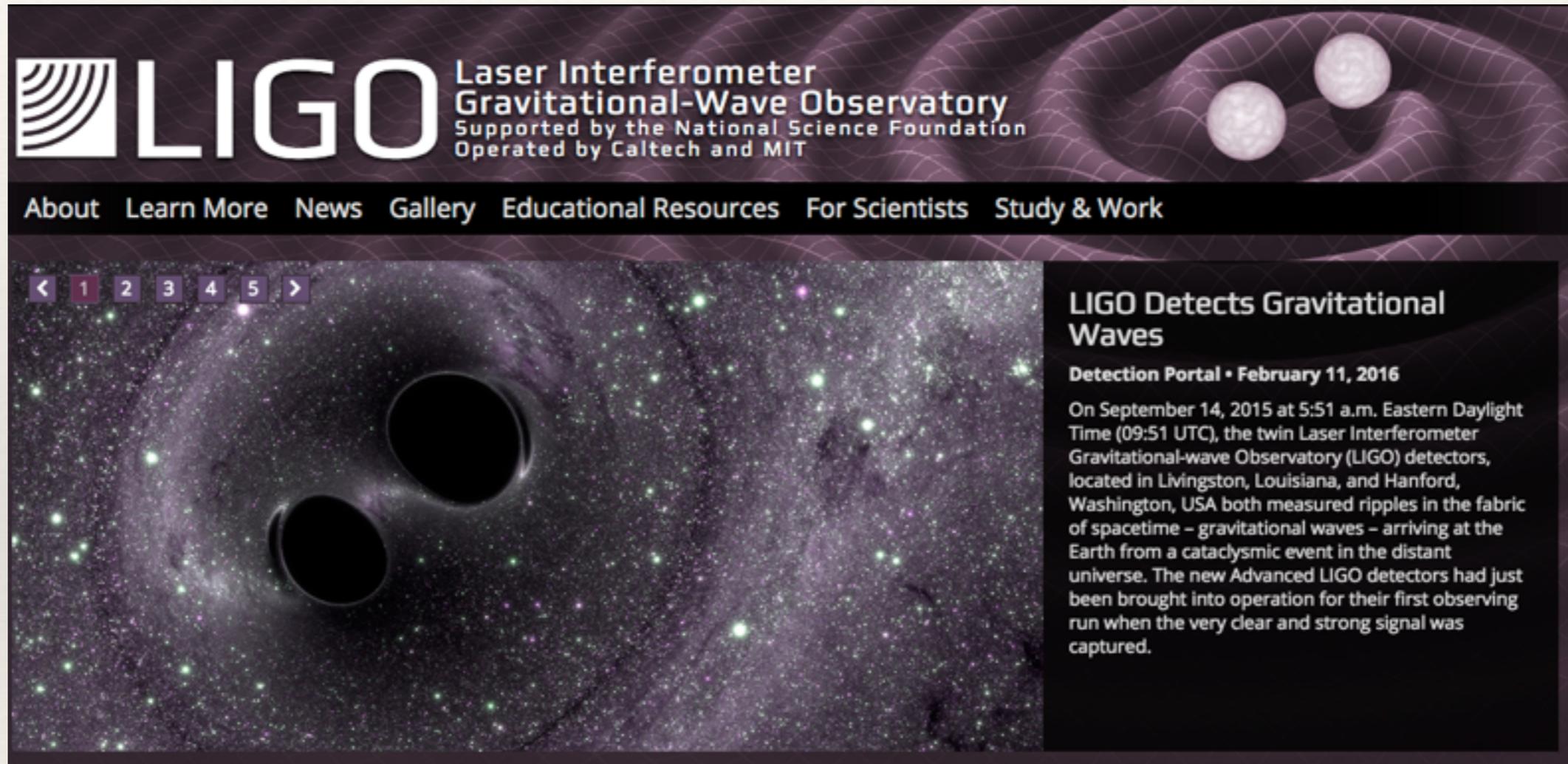
---

オープン？

汎用言語？

SAS	×	×
Matlab	×	×
R	○	×
Python	○	○

# 重力波を初めて観測



**LIGO** Laser Interferometer Gravitational-Wave Observatory  
Supported by the National Science Foundation  
Operated by Caltech and MIT

About Learn More News Gallery Educational Resources For Scientists Study & Work

< 1 2 3 4 5 >

## LIGO Detects Gravitational Waves

Detection Portal • February 11, 2016

On September 14, 2015 at 5:51 a.m. Eastern Daylight Time (09:51 UTC), the twin Laser Interferometer Gravitational-wave Observatory (LIGO) detectors, located in Livingston, Louisiana, and Hanford, Washington, USA both measured ripples in the fabric of spacetime – gravitational waves – arriving at the Earth from a cataclysmic event in the distant universe. The new Advanced LIGO detectors had just been brought into operation for their first observing run when the very clear and strong signal was captured.

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

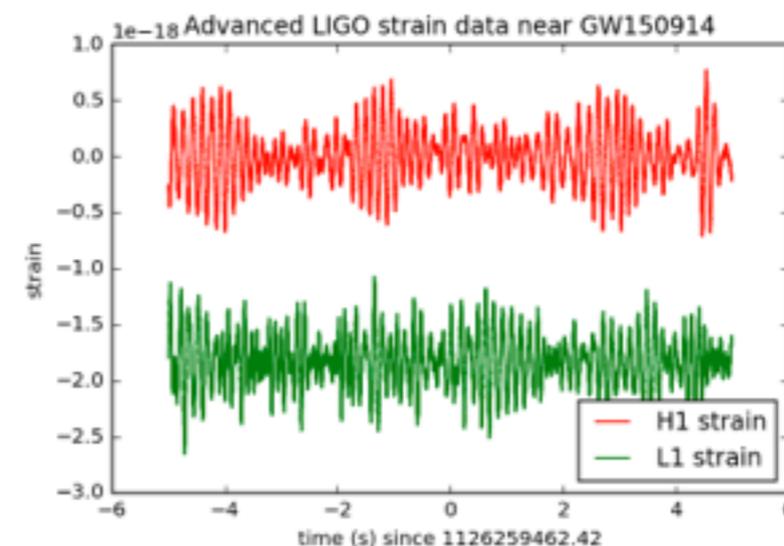
# オープン & 汎用である事の利点

- ❖ 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')
```



# Rの人気

# KDnuggets



Data Mining, Analytics, Big Data, and Data Science

Subscribe to [KDnuggets News](#) | Follow [Twitter](#) [Facebook](#) [LinkedIn](#) | [Contact](#)

[Data Mining Software](#) | [News](#) | [Top stories](#) | [Opinions](#) | [Tutorials](#) | [Jobs](#) | [Academic](#) | [Companies](#) | [Courses](#) | [Datasets](#) | [Education](#) | [Meetings](#) | [Polls](#) | [Webinars](#)



a podcast on everything from trending topics in data to interviews with some of the most influential people in the industry



[Afternoon Analytics](#) - a podcast on trending topics in data, interviews with influential people, and more

[KDnuggets Home](#) » [News](#) » [2015](#) » [Jul](#) » [News, Features](#) » [R, Python users show surprising stability, but strong regional differences \( 15:n22 \)](#)

## 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](#)

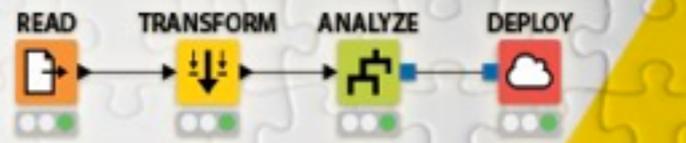
[Next post ▶](#)

[f](#) [in](#) [G+1](#) 6 [Share](#) 25 [Tweet](#)

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](#)



**KNIME Analytics Platform** helps solve your most complex data puzzles

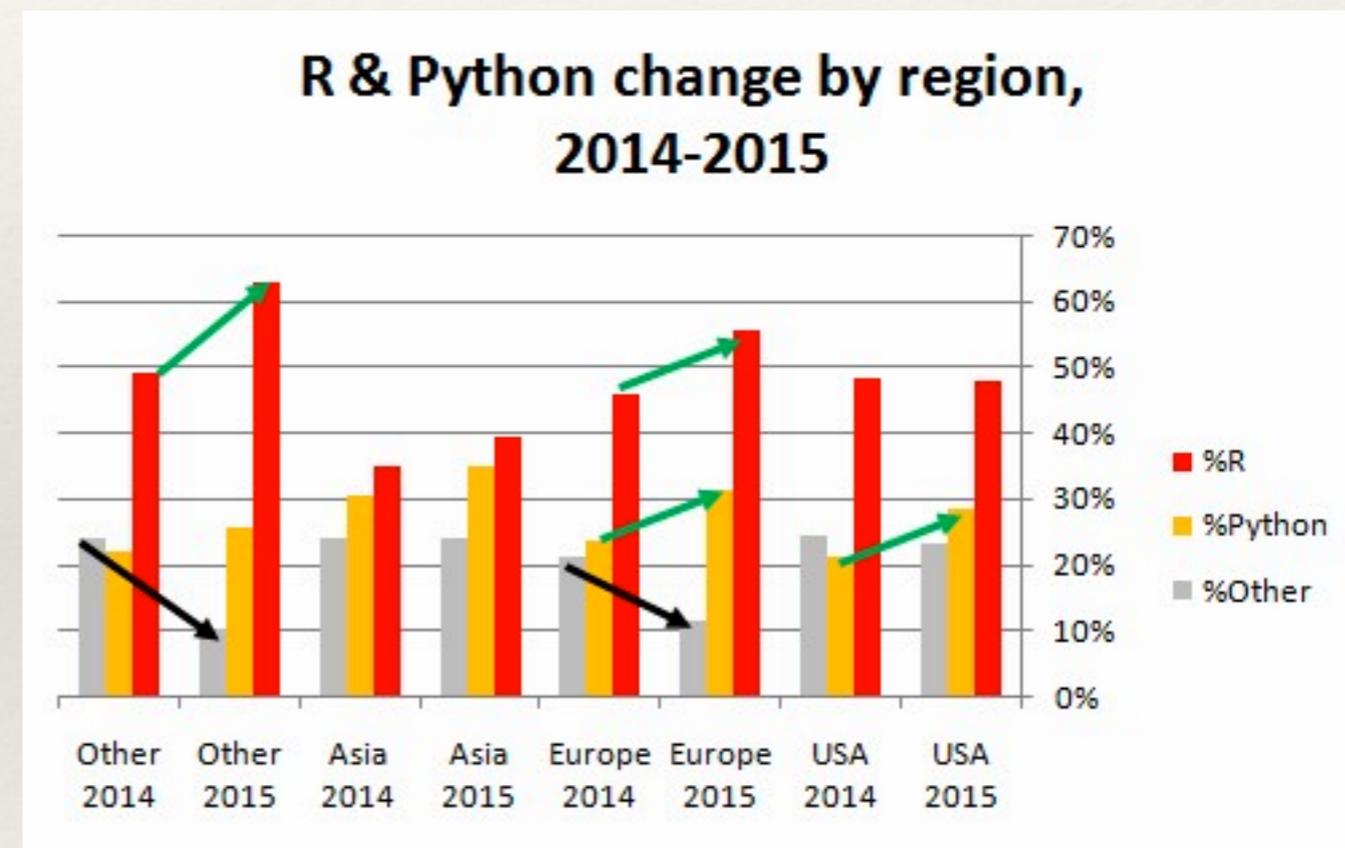
*Integrating R, Python, Spark, MLlib & more*



Open for Innovation  [LEARN MORE](#)

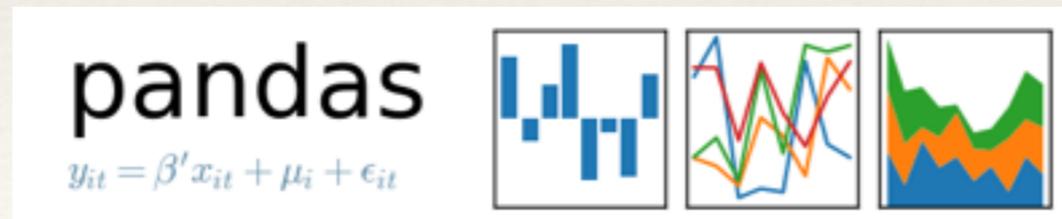
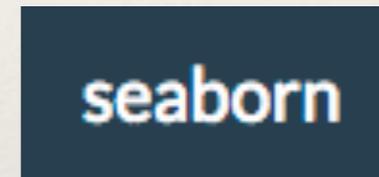
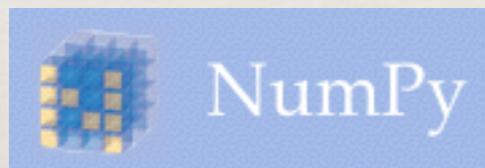
# RとPythonの比較

- ❖ 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）



# Pythonデータサイエンスの エコシステム

# Pythonはglue (のり) 言語



IP[y]:  
IPython

---

# 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

---

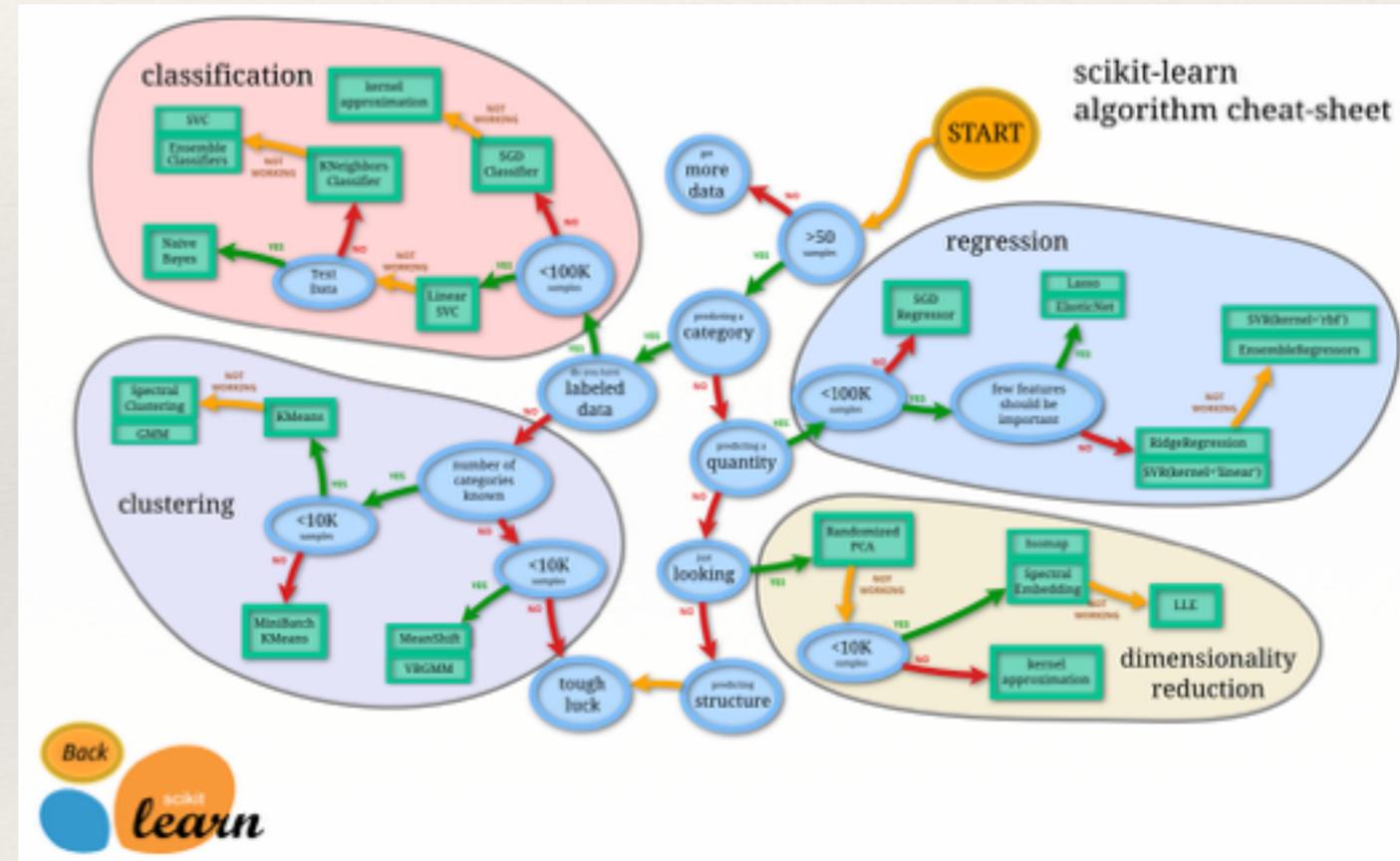
- ❖ データの可視化に使われる
- ❖ matplotlibはmatlabの代替を意識
- ❖ seabornはmatplotlibを基礎にして、使用しやすく、統計的な機能も取り込む

**matplotlib**

**seaborn**

# 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

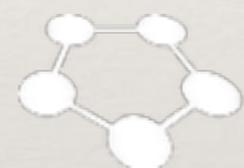


---

# 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



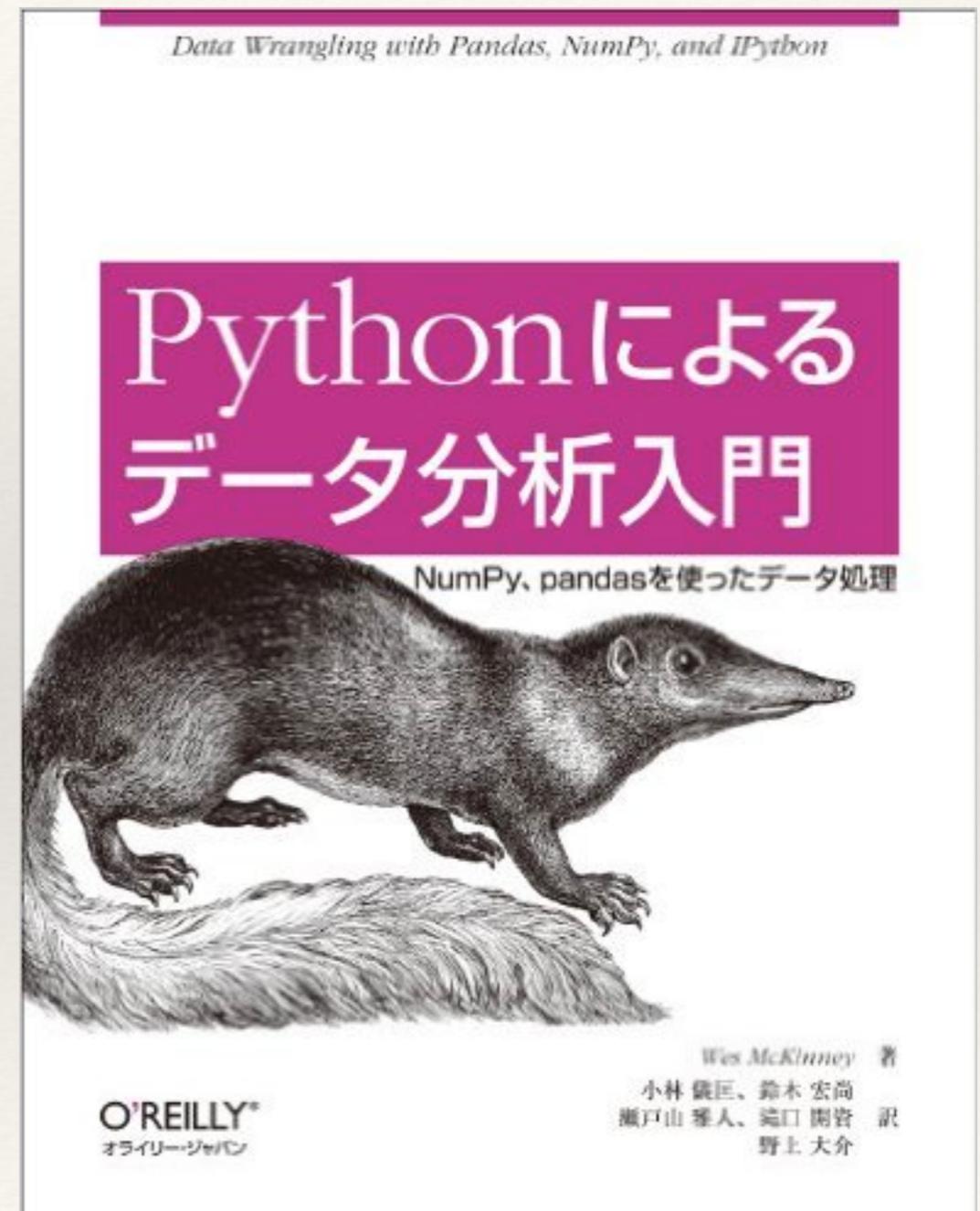
Chainer

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

kaggle [back to main site](#) [Get in touch](#)

The world's largest community of data scientists compete to solve your most valuable problems.

[Get in Touch!](#)

**Why**  
Many organizations don't have access to the advanced machine learning that provides the maximum predictive power from their data. Meanwhile, data scientists and statisticians crave real-world data to develop their techniques. Kaggle offers companies a cost-effective way to harness this 'cognitive surplus' of the world's best data scientists.

**Who**  
Our vibrant community comprises experts from many quantitative fields and industries (science, statistics, econometrics, math, physics). They come from over 100 countries and 200 universities. In addition to prize money & data, they use Kaggle to learn, network, and collaborate with experts from related fields.

---

# 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