BPStudy#121〜地に足をつけて学ぶ機械学習、データサイエンス

今データサイエンスが必要とされる理由とPythonの役割

辻 真吾 @tsjshg

2017.9.26

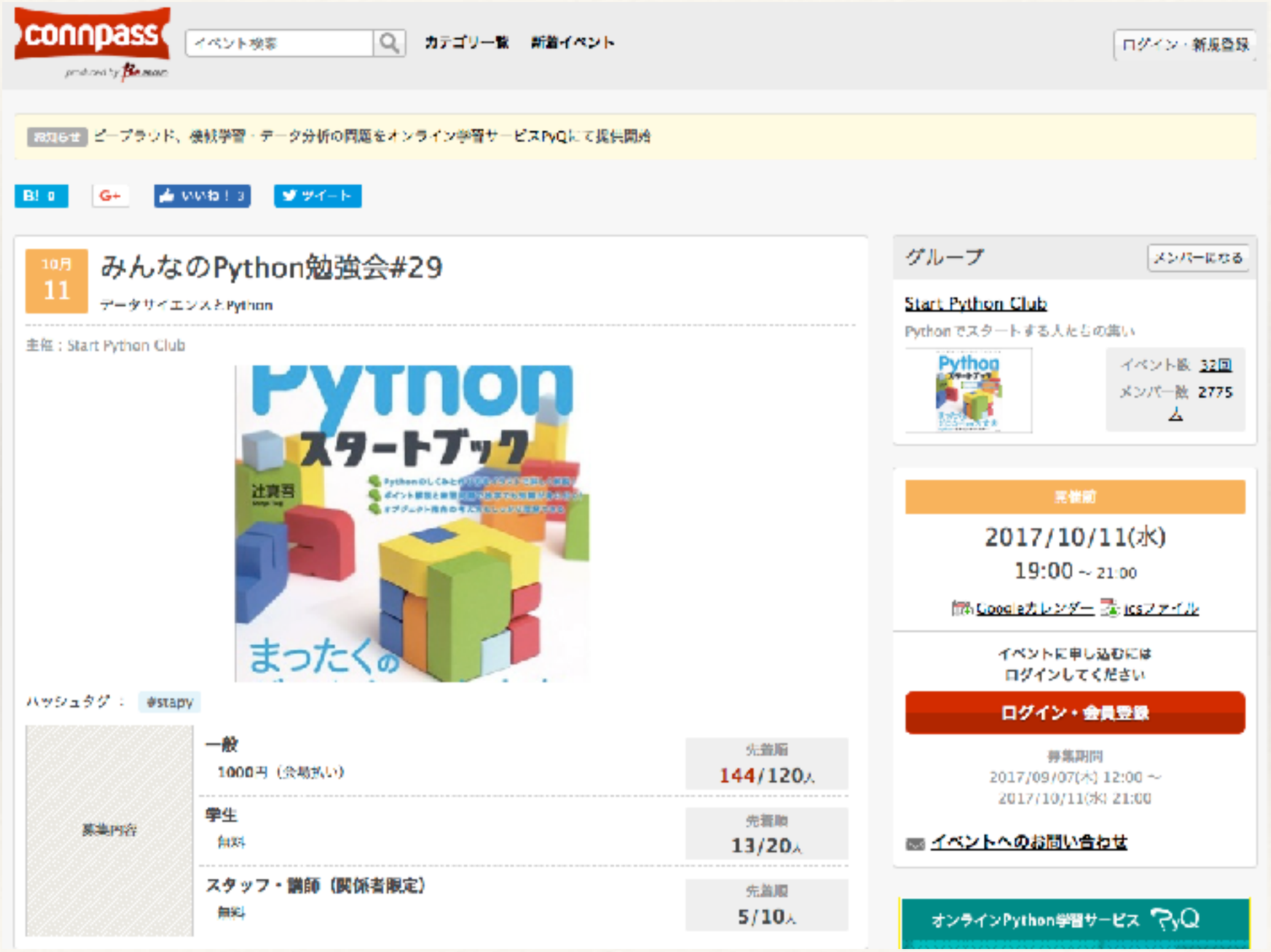
自己紹介

- 1975年生まれ東京都足立区出身
- コンピュータと数学でなんでもやる数理工学を専攻
- 修士終了後、創業間もない（株）いい生活に就職
  - Javaを使ったWebアプリ開発
- 3年弱で会社を辞めて、現在の勤め先（東大先端研）に博士課程の学生として復学
  - バイオインフォマティクス（生命情報科学）
- 42才ですが、人生の次の一手を模索中です

本とWeb

https://www.udemy.com/python-jp/

お知らせ1



https://startpython.connpass.com/

お知らせ2

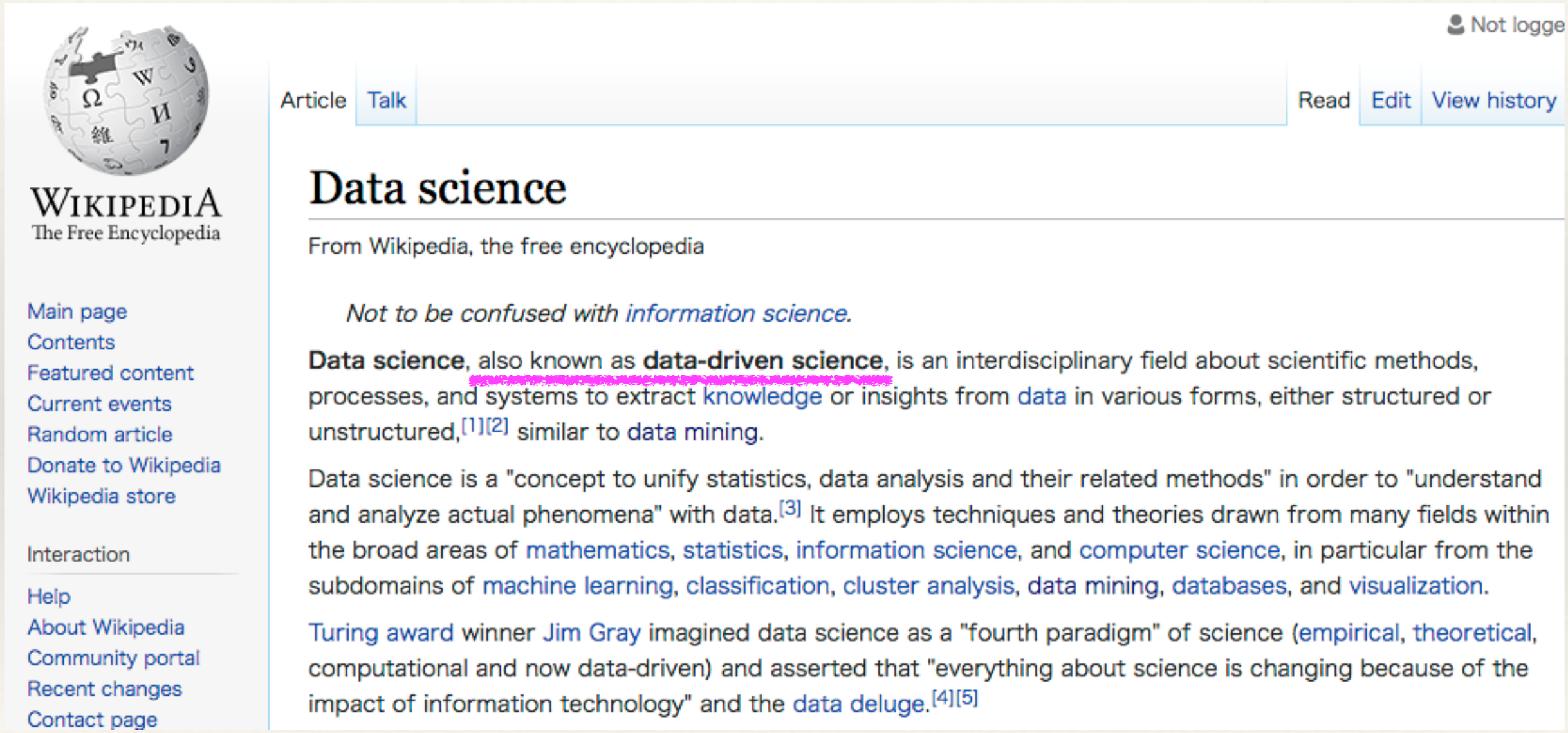
https://edgeai.connpass.com/

本日の目次

- データサイエンスとその必要性
- Pythonの役割
  - 汎用言語でデータ解析
  - 解析の再現性とオープンソース
- データサイエンスのこれから

データサイエンスとは？

データ駆動型サイエンス



Not logge

Article Talk Read Edit View history

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Data science

From Wikipedia, the free encyclopedia

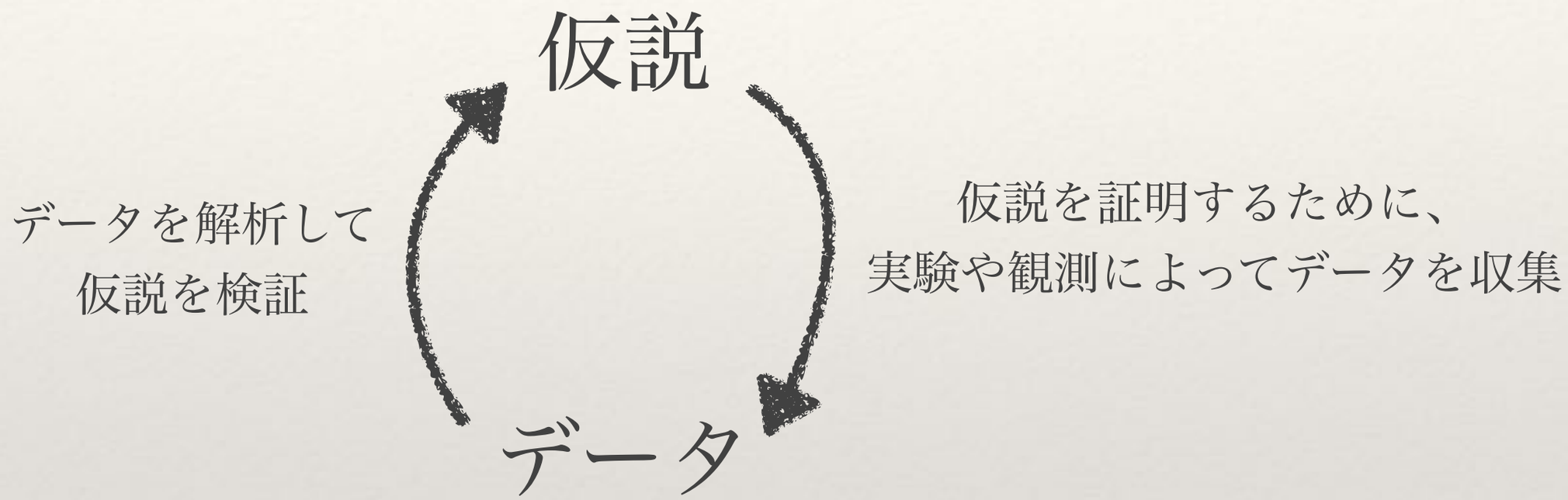
Not to be confused with information science.

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

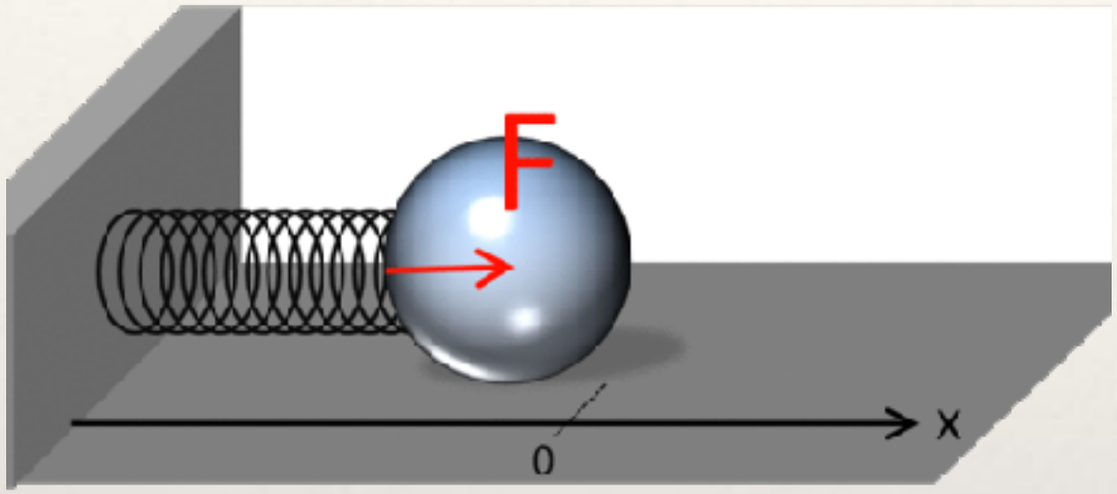
Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

サイエンス



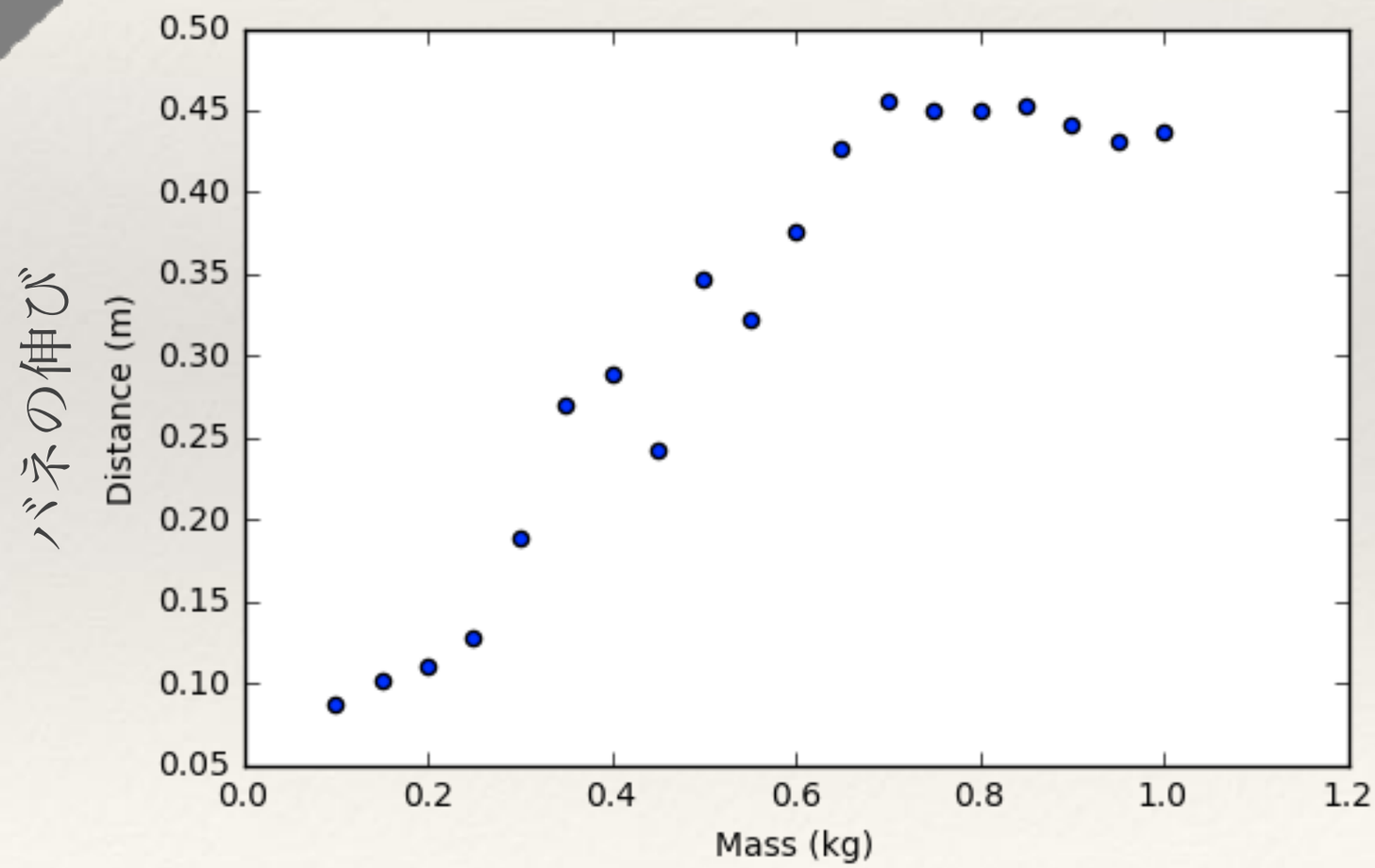
普遍的な原理

たとえば



フックの法則

$F = -k x$

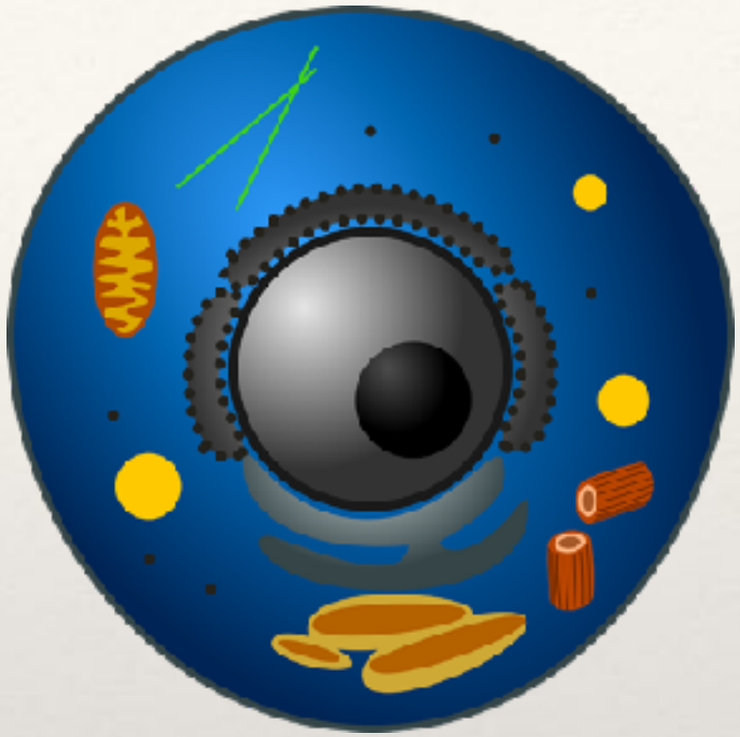


バネの伸び

バネにかかる力

DNA配列だと

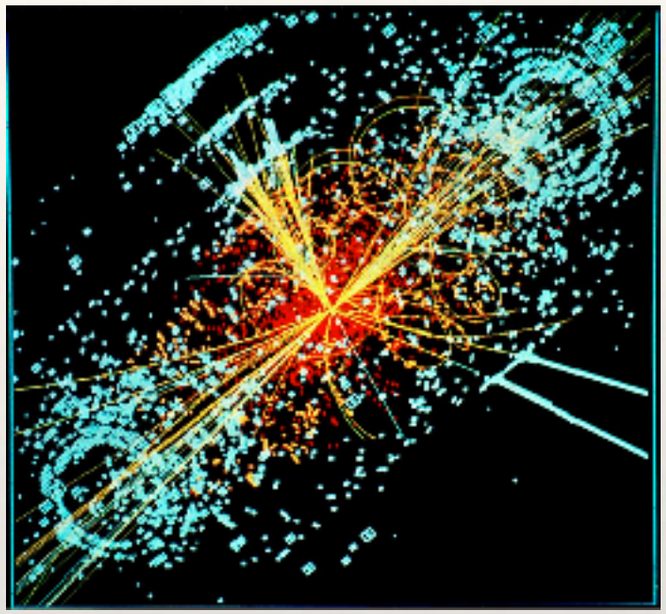
- ヒトの細胞には、約30億文字分のDNA情報が格納されている
- これが最近の装置の進歩によって1週間弱で解読可能
- 100人分やったら、3,000億文字



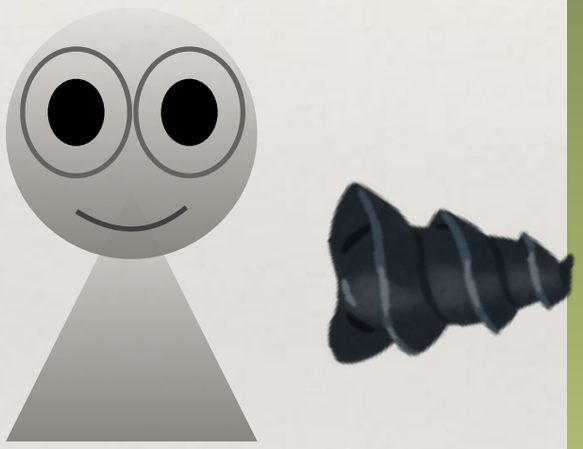
https://jp.illumina.com/systems/sequencing-platforms/hiseq-2500.html

とにかくビッグデータ

- ヒッグス粒子の発見
  - データ解析にはPythonが使われたらしい
- Web上のお客さんの行動履歴
- IoT
  - 地理情報を含む行動履歴
  - エネルギー消費量



データサイエンスの必要性



データ

普遍的な何か

データサイエンスの実際

- データを集める（CSV, Excel, DB…）
- Webスクレイピング（通信とデータの整形）
- 欠損値や異常値の検出と処理

- 完成品のWebアプリ化

汎用言語でデータサイエンスを実践する意味は大きい

Python急成長の真相

IEEE Spectrum

Choose a Ranking (choose a weighting or make your own)

IEEE Spectrum | Trending | Jobs | Open | Custom

Edit Ranking | Add a Comparison |

Language Types (click to hide)

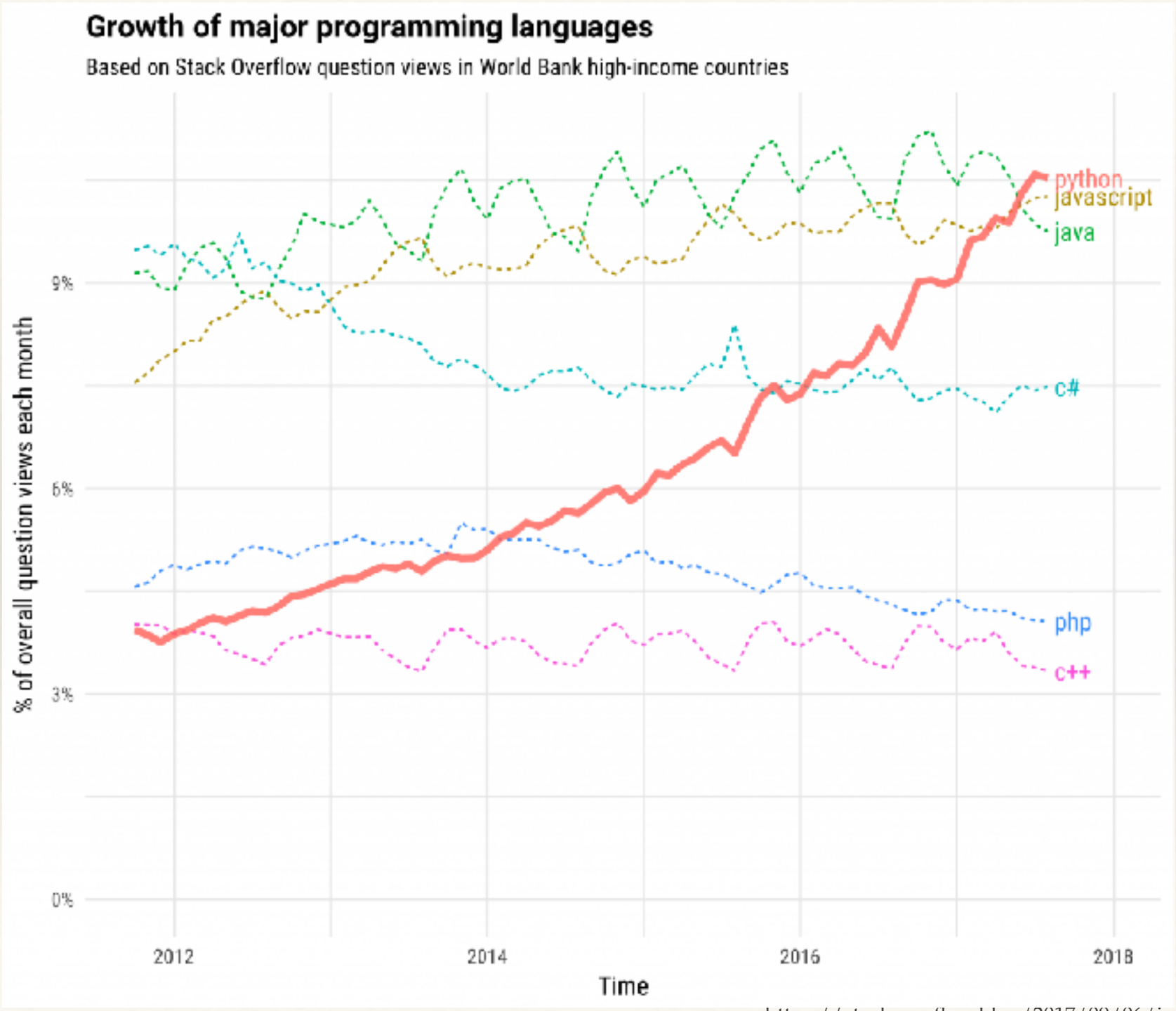
Web | Mobile | Enterprise | Embedded

Language Rank	Types	Spectrum Ranking
1. Python	Web, Enterprise	100.0
2. C	Mobile, Enterprise, Embedded	99.7
3. Java	Web, Mobile, Enterprise	99.4
4. C++	Mobile, Enterprise, Embedded	97.2
5. C#	Web, Mobile, Enterprise	88.6
6. R	Enterprise	88.1
7. JavaScript	Web, Mobile	85.5
8. PHP	Web	81.4
9. Go	Web, Enterprise	76.1
10. Swift	Mobile, Enterprise	75.3
11. Arduino	Embedded	73.0

IEEE SPECTRUM

http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2017

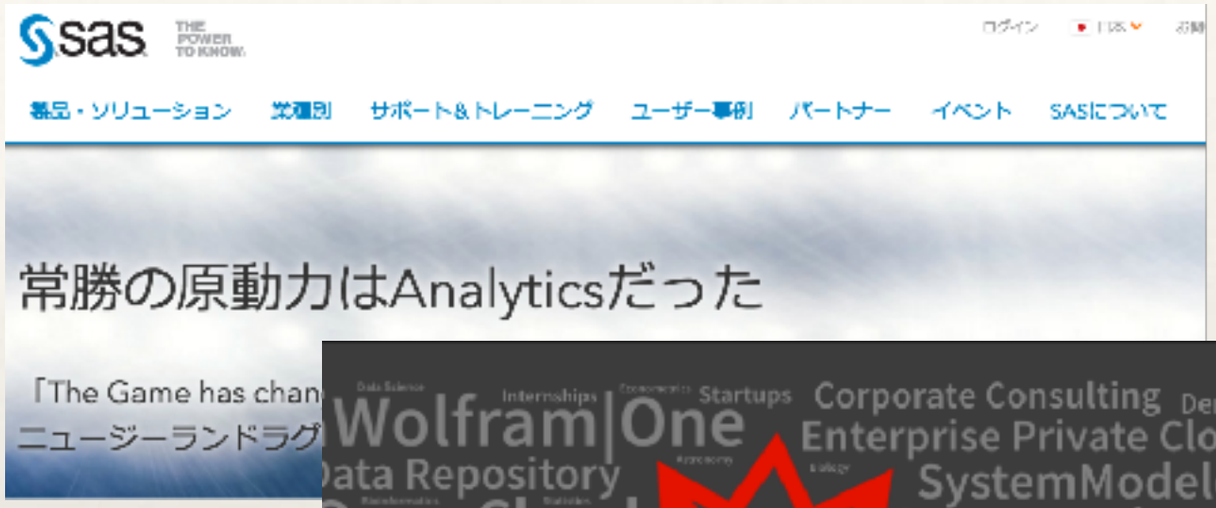
Stack Overflow



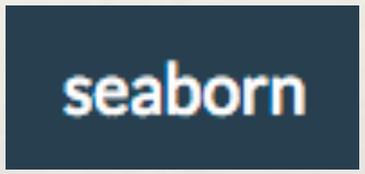
https://stackoverflow.blog/2017/09/06/incredible-growth-python/?cb=1

いくつかの選択肢

- SAS (www.sas.com)
  - BIツールの代表格
- Mathematica（www.wolfram.com）
  - 天才ウルフラムが作った老舗数式処理ソフト
- 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
  - オープンソースのデータ解析専用言語
- Python
  - この中では唯一、オープンかつ汎用



Pythonはglue（のり）言語



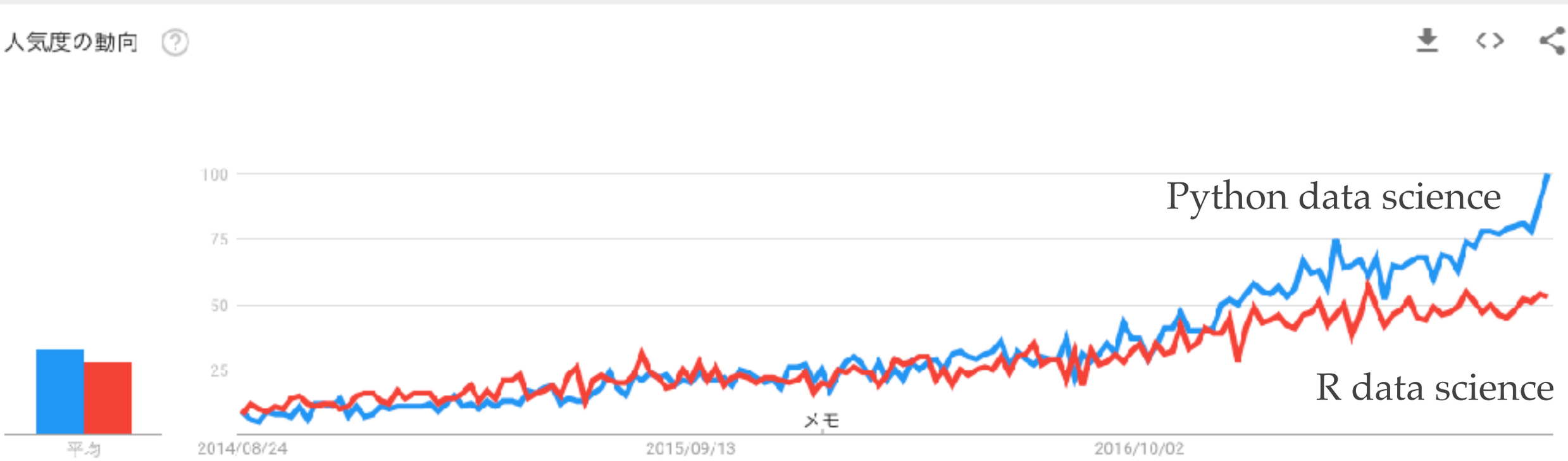
Anacondaがおすすめ！

- 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（データ解析用が中心）を同梱
- 無料
- データ解析環境の構築に最適

https://www.continuum.io

RとPython

過去3年間のGoogle Trendsのデータ



データサイエンス自体が伸びているので、もちろんRも中心的な存在

http://www.datasciencecentral.com/profiles/blogs/python-overtakes-r-for-data-science-and-machine-learning

jupyter Jupyter_notebookのデモ Last Checkpoint: 2 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Code

In [1]:
```
import pandas as pd
```

In [2]:
```
data = pd.read_excel('sample_data.xlsx')
```

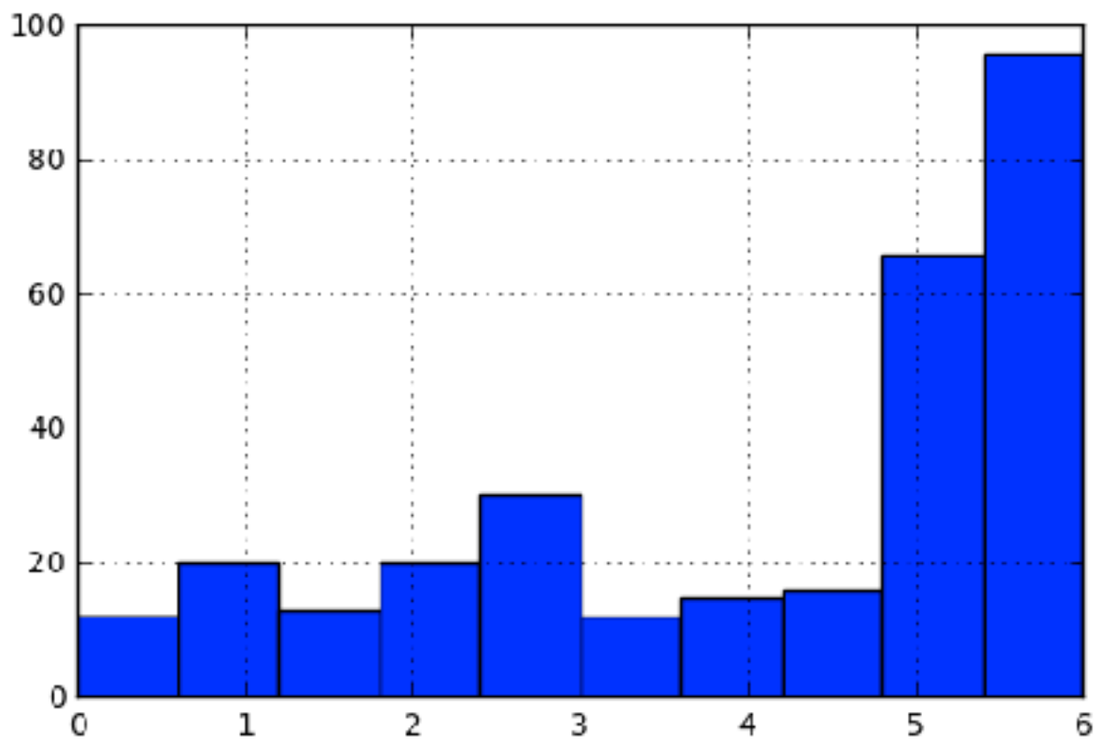
In [3]:
```
data.head()
```

Out[3]:

	sample_0	sample_1	sample_2	sample_3	sample_4	sample_5	sample_6	sample_7	sample_8	sample_9
var_0	1.181546	2.027050	2.307813	1.078164	2.023769	1.083772	0.199090	1.002069	0.715403	1.695127
var_1	0.998445	0.601550	0.054010	0.028059	1.024477	0.737187	2.145073	2.395837	0.594704	0.319473
var_10	5.659790	8.118167	8.333933	8.388697	8.083567	8.349710	8.332364	8.143534	8.277051	8.338494
var_100	2.383074	2.155451	0.948400	2.090212	0.653361	0.897784	1.413244	0.544167	2.484165	0.101348
var_101	5.667063	5.848190	6.075058	5.611951	5.649971	5.803400	5.618007	5.583713	5.629790	5.576541

In [5]:
```
%matplotlib inline
data['sample_0'].hist()
```

Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x1146362b0>



jupyter notebook

エクセルファイルをpandasのDataFrameとして読み込んで、ヒストグラムを描く

In [6]:
```
import seaborn as sns
sns.pairplot(data)
```

Out[6]:
```
<seaborn.axisgrid.PairGrid at 0x1146b7c88>
```

ペアプロットを描く。なんか、データが1つおかしそう・・・

R

```r
library(GGally)

ggpairs(nba[,c("ast", "fg", "trb")])
```

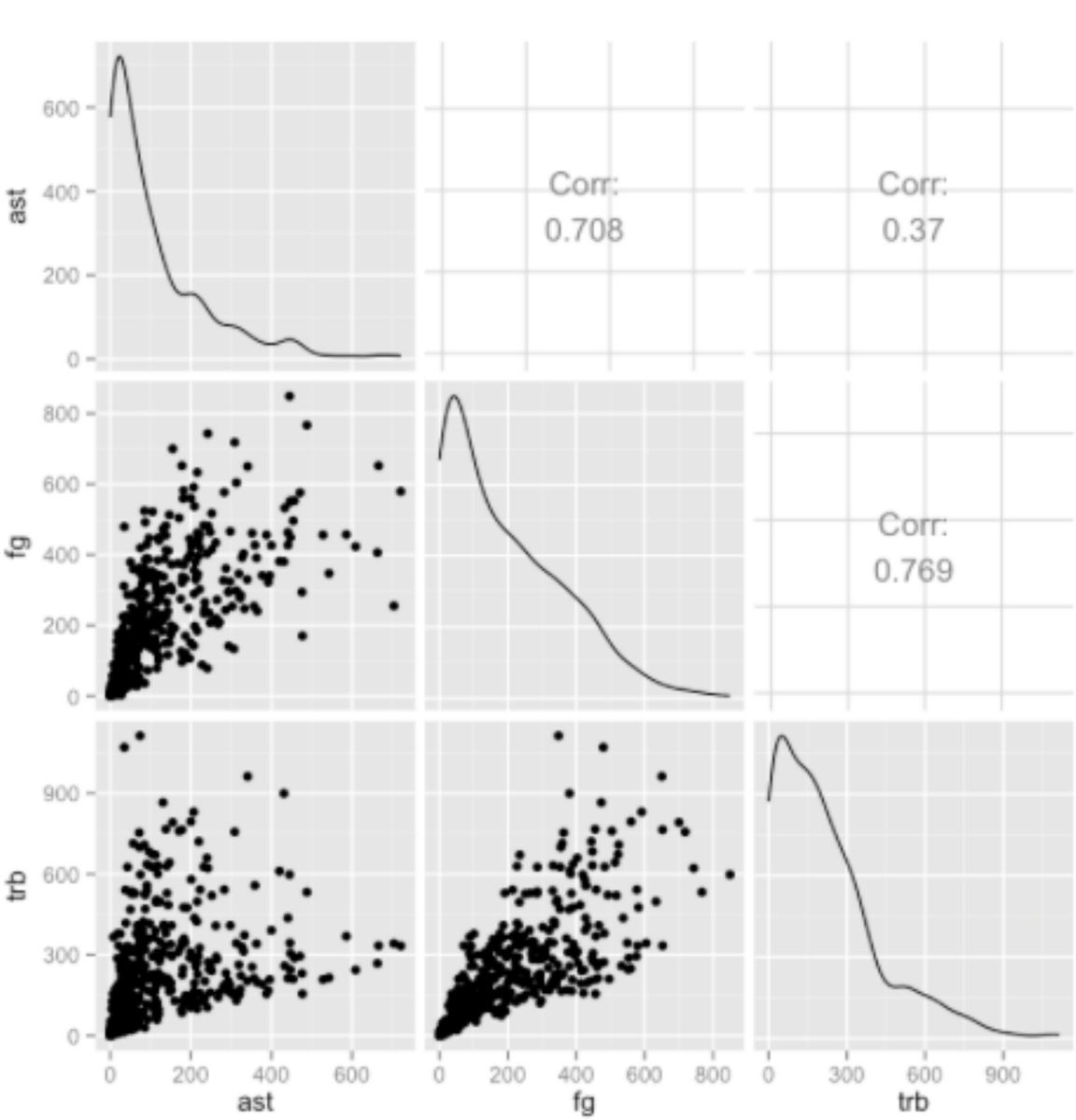
Python

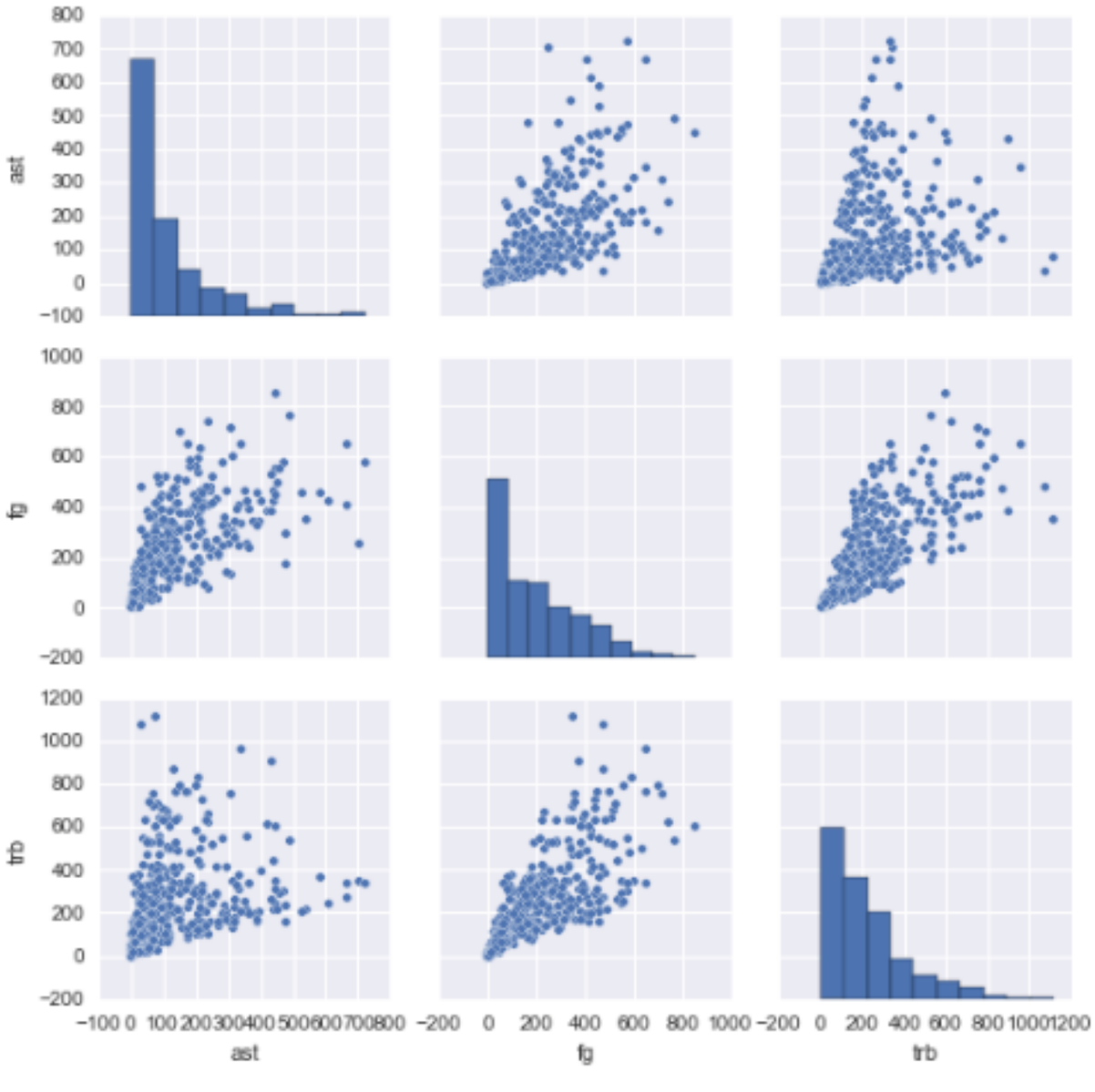
```python
import seaborn as sns

import matplotlib.pyplot as plt

sns.pairplot(nba[["ast", "fg", "trb"]])

plt.show()
```

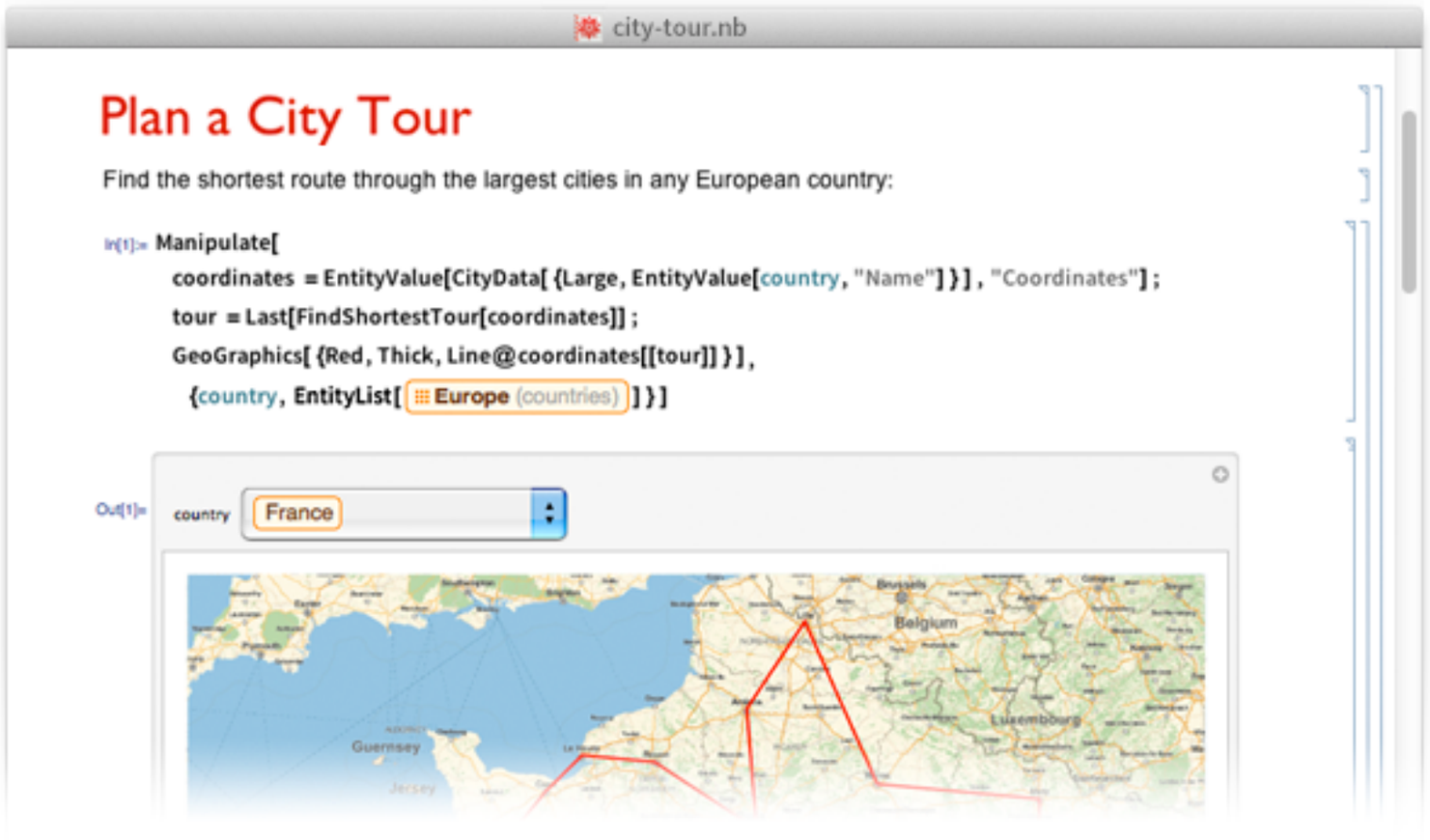




Mathematicaで使われるWolfram言語のドキュメントより

ノートブックドキュメント

Wolframノートブックは，テキスト，グラフィックス，インターフェース等とコードを組み合せることができ，デスクトップでもWebでも使えます：



PYTHONのプログラマー向けの注意事項

Wolframノートブックは25年以上に渡り常に開発され続けており，Wolfram言語のデスクトップ版とクラウド版のどちらにおいても完全に統合されています．Wolframノートブックはjupyterのようなノートブックライブラリのきっかけとなっています．

http://www.wolfram.com/language/fast-introduction-for-programmers/ja/notebook-documents/

Pythonが台風の目に

- Pythonとそれを支えるデータサイエンスのエコシステムは、良いものをどんどん取り入れ進化が加速
- もちろん、それぞれの特技はある
  - たとえば、Rは統計に強い

マニアックな統計関数

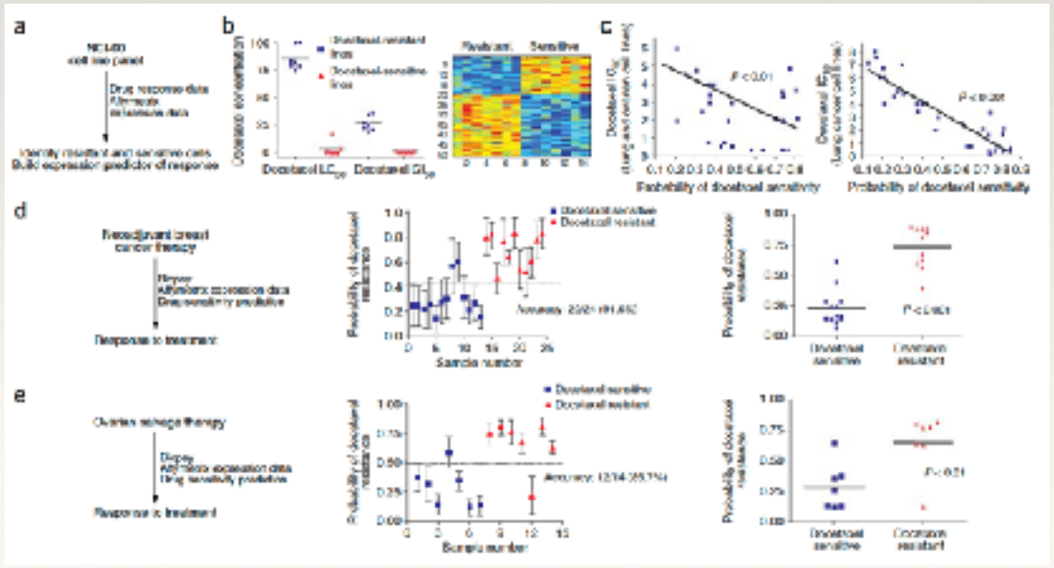
	R	Python
Kolmogorov–Smirnov検定	○	scipy.stats
Shapiro-Wilk検定	○	scipy.stats
Hosmer-Lemeshow検定	ResourceSelection	?

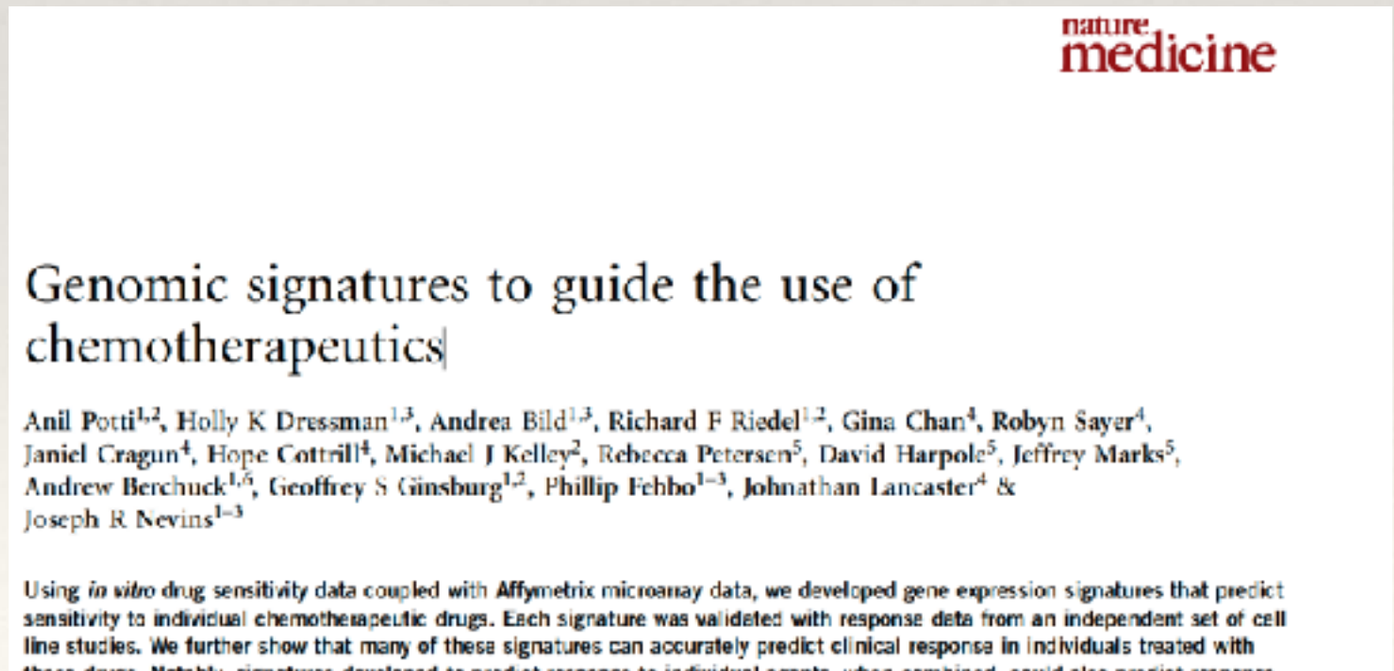
適切なツールを選べる知識と柔軟性が重要（だと思う）

サイエンスと再現性

本当にあった怖い話

- 2006年、名門Nature Medicine誌に、1本の論文が掲載される
- 抗がん剤の効きを、細胞株（試験管で培養できる細胞）のデータを使って予測できるとする当時としては画期的な内容
- 教授にこの論文の再現を命じられる私





nature medicine

Genomic signatures to guide the use of chemotherapeutics

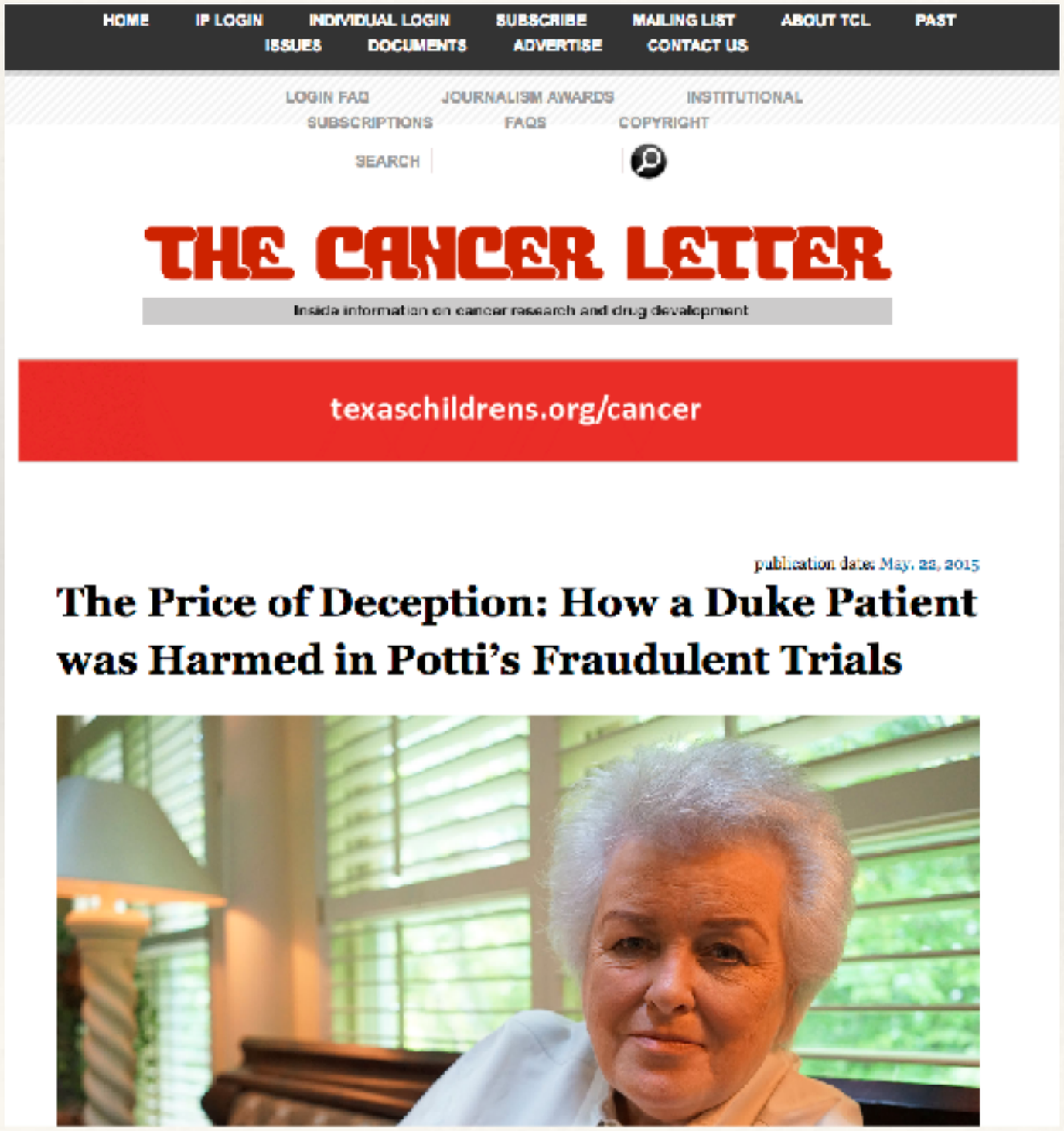
Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,5], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression

なにかおかしい・・・

- データ解析だけの論文だが、複雑な計算方法の詳細が書かれていない
- 業界の常識的な方法論でいろいろ試してみる
- まったく再現しない・・・
- 世界中の研究者が疑いはじめる

全部うそ


HOME
IP LOGIN
INDIVIDUAL LOGIN
SUBSCRIBE
MAILING LIST
ABOUT TCL
PAST ISSUES
DOCUMENTS
ADVERTISE
CONTACT US
LOGIN FAQ
JOURNALISM AWARDS
INSTITUTIONAL SUBSCRIPTIONS
FAQS
COPYRIGHT
SEARCH
THE CANCER LETTER
Inside information on cancer research and drug development
texaschildrens.org/cancer
publication date: May. 22, 2015
The Price of Deception: How a Duke Patient was Harmed in Potti's Fraudulent Trials


WIKIPEDIA
The Free Encyclopedia
Article
Talk
Anil Potti
From Wikipedia, the free encyclopedia
Anil Potti is a physician and former Duke
oncogenomics. He, along with Joseph Nevi
scandal at Duke University[1][2][3] and many
retracted. On November 9, 2015, the Office
engaged in research misconduct.[4] Accord
with ORI, Potti can continue to perform res
year 2020, while he "neither admits nor de
Contents [hide]
1 Biography
2 Scientific misconduct
3 Consequences
4 Research questions and Institute of Medi
5 Continuing controversy
6 Retracted papers
About Wikipedia
Community portal
Recent changes
Contact page
Tools

再現性とオープンなこと

- データサイエンスと言うからには、再現性が必要
  - もちろん、ビジネスの現場でも
- 解析を完全に再現するには、オープンな基盤は重要
  - Pythonが果たす役割
- オープンサイエンスという試みが世界的に動き出している
  - STAP細胞ってあるんですかね？

なにがどこまで出来るのか？

自然言語処理

- 人が使う言語を計算機に理解させるための技術
- 人工知能研究の黎明期（1950年代）からある
- そんな簡単じゃない

1.職場の上司とW不倫中です。

2.職場の上司と付き合っています。私は既婚者です。

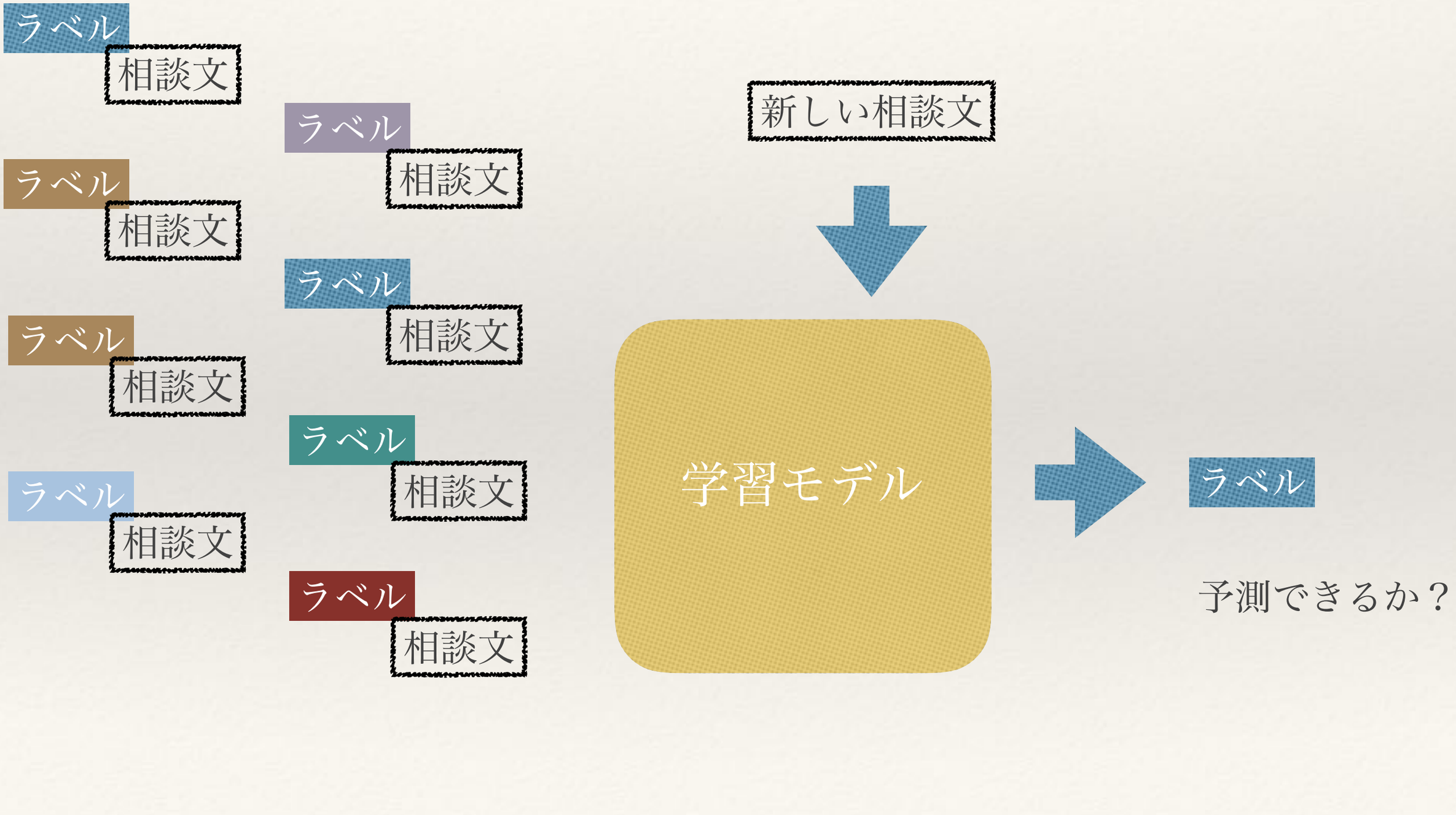
＊占いコンテンツに強い会社さんとの協業事例なので、例文が若干人間味溢れておりますが、ご了承下さい。

たとえば・・・

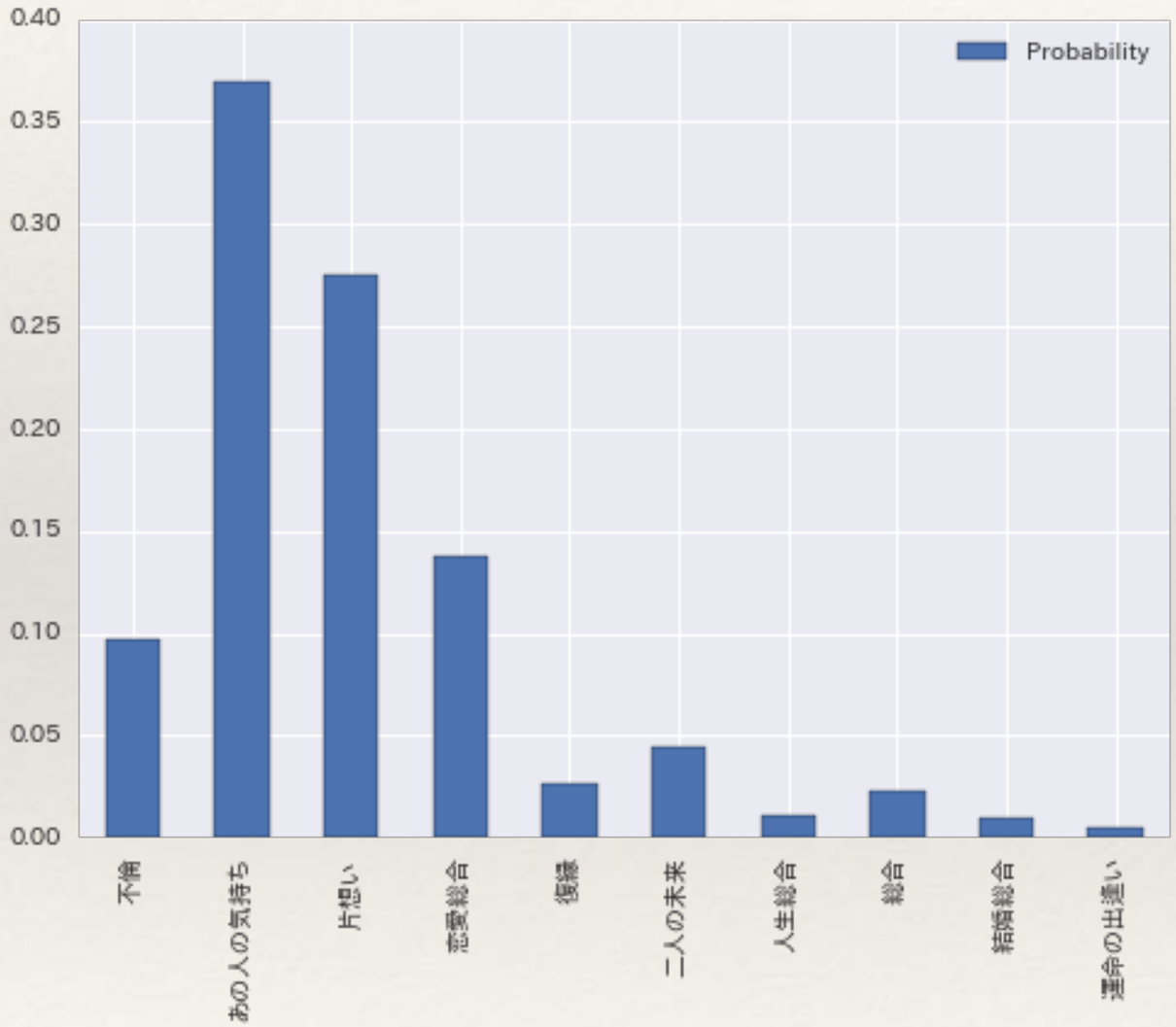
実際の相談内容です。公開出来ないのですみませんが、割愛

これは「片想い」の相談（ユーザーの入力）

大量のデータを使った学習



意味を漠然と捉えることはできる

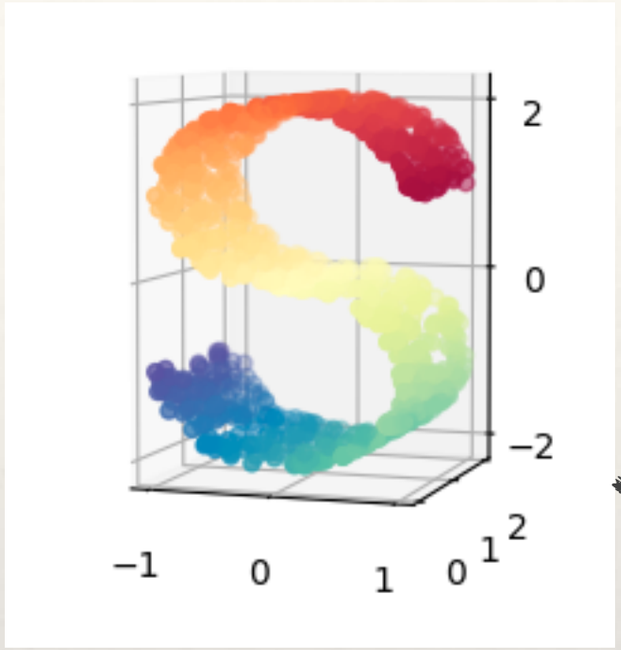


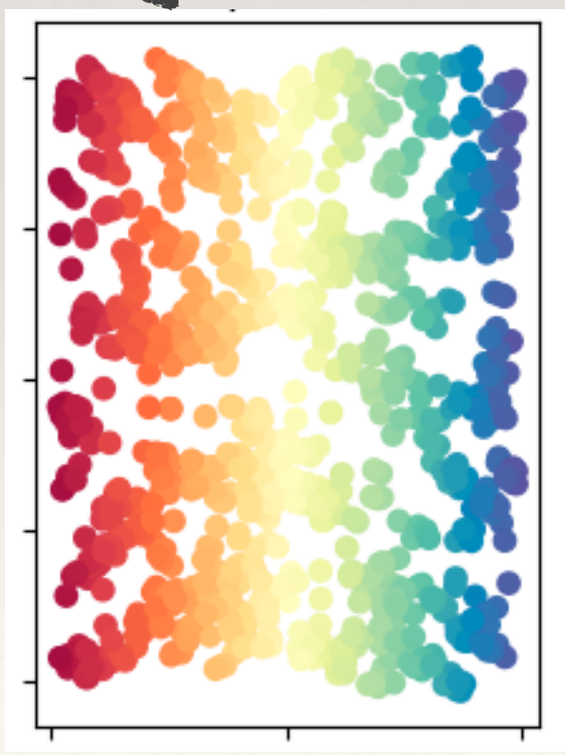
10カテゴリ、3万件の教師データを使ったRandom Forestsモデルの予測

あの人の気持ちが気になる片想い。
ちょっと不倫の要素も。

多様体仮説（学習）

- 単語の数は沢山あるけど、トピックの数は、はるかに少ない
- 高次元データでも、それを生成する本質は低次元で表現できる（はず）
- 計算してデータを加工するのは計算機、推測するのは人間
  - PCA, Autoencoder





http://scikit-learn.org/stable/auto_examples/manifold/plot_compare_methods.html

まとめ

- データサイエンスはデータ駆動型サイエンス
- 前処理負担が重いので、汎用言語が役に立つ
- 再現性は重要。オープンな環境がそれを支える
- 大量データの背後に何があるのか意識する

ご静聴ありがとうございました