

みんなのPython勉強会#29

データサイエンスと

Pythonの役割

辻 真吾 @tsjshg

2017.10.11

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



This is a screenshot of the Udemy website. The browser address bar shows 'udemy.com'. The page features a green header with the Udemy logo and navigation links. The main content area has a dark background and displays a course titled '【世界で5万人が受講】実践 Python データサイエンス' (【50,000 people in the world are enrolled】Practical Python Data Science). Below the title, there is a description of the course, a star rating of 4.3 from 1593 reviews, and the number of students enrolled (10336). The course is created by Shingo Tsuji and Jose Portilla, with the last update on 3/2016. To the right of the course title, there is a video player with a play button and a 'Preview' button. Below the video player, the price is listed as ¥1,200, which is a 90% discount from the original price of ¥12,600. There are buttons for 'Purchase Now' and 'Add to Cart'. At the bottom, there is a list of learning objectives and a link to 'View Curriculum'.

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

本日の目次

- ❖ BPStudy#121～地に足をつけて学ぶ機械学習、データサイエンスの再放送（一部内容を変更）
- ❖ 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

データサイエンスとは？

データ駆動型サイエンス



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](#)

Not logged in

[Article](#) [Talk](#)

[Read](#) [Edit](#) [View history](#)

Data science

From Wikipedia, the free encyclopedia

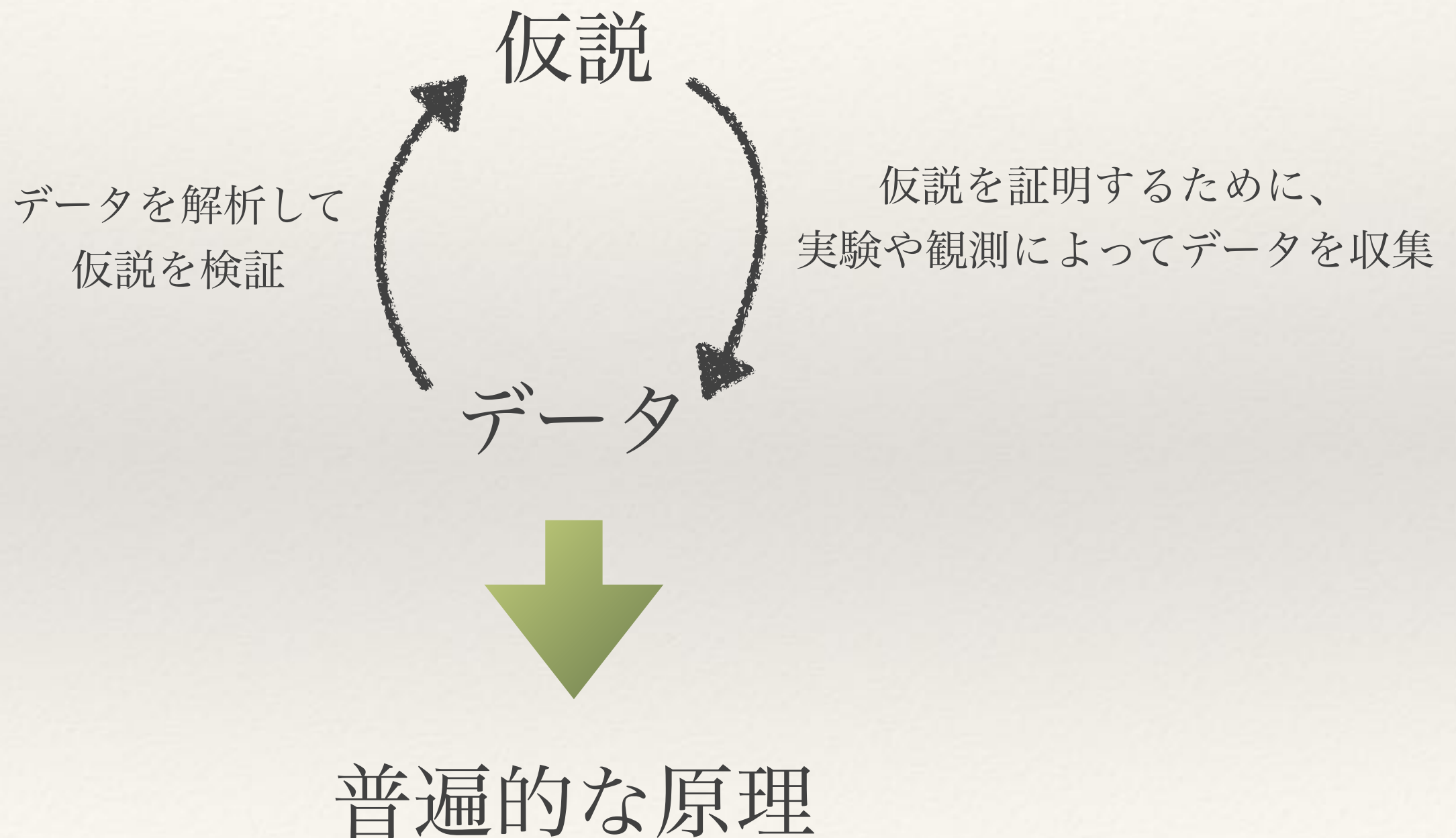
Not to be confused with [information science](#).

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](#).

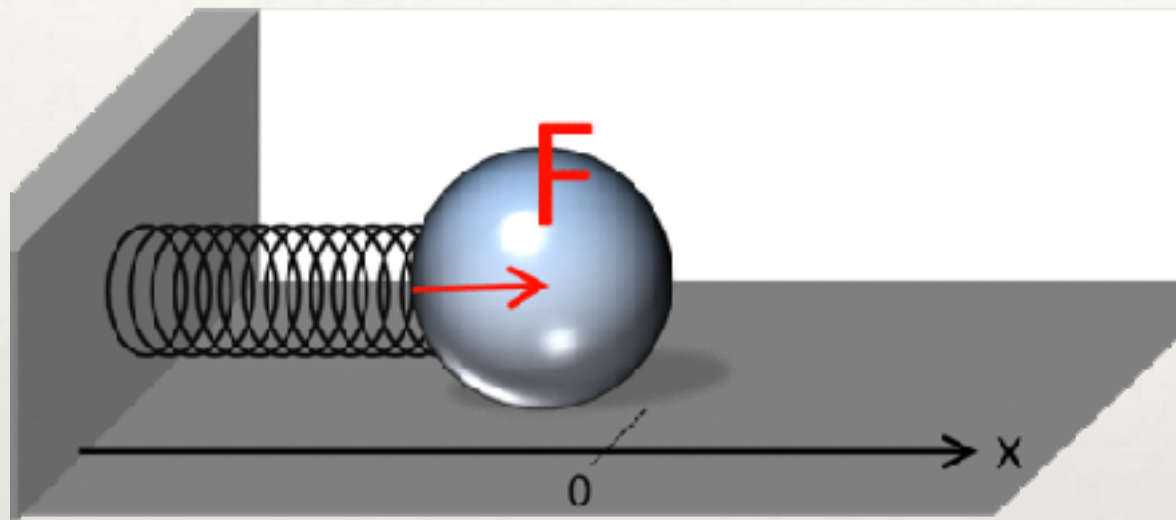
Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.^[3] It employs techniques and theories drawn from many fields within the broad areas of [mathematics](#), [statistics](#), [information science](#), 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



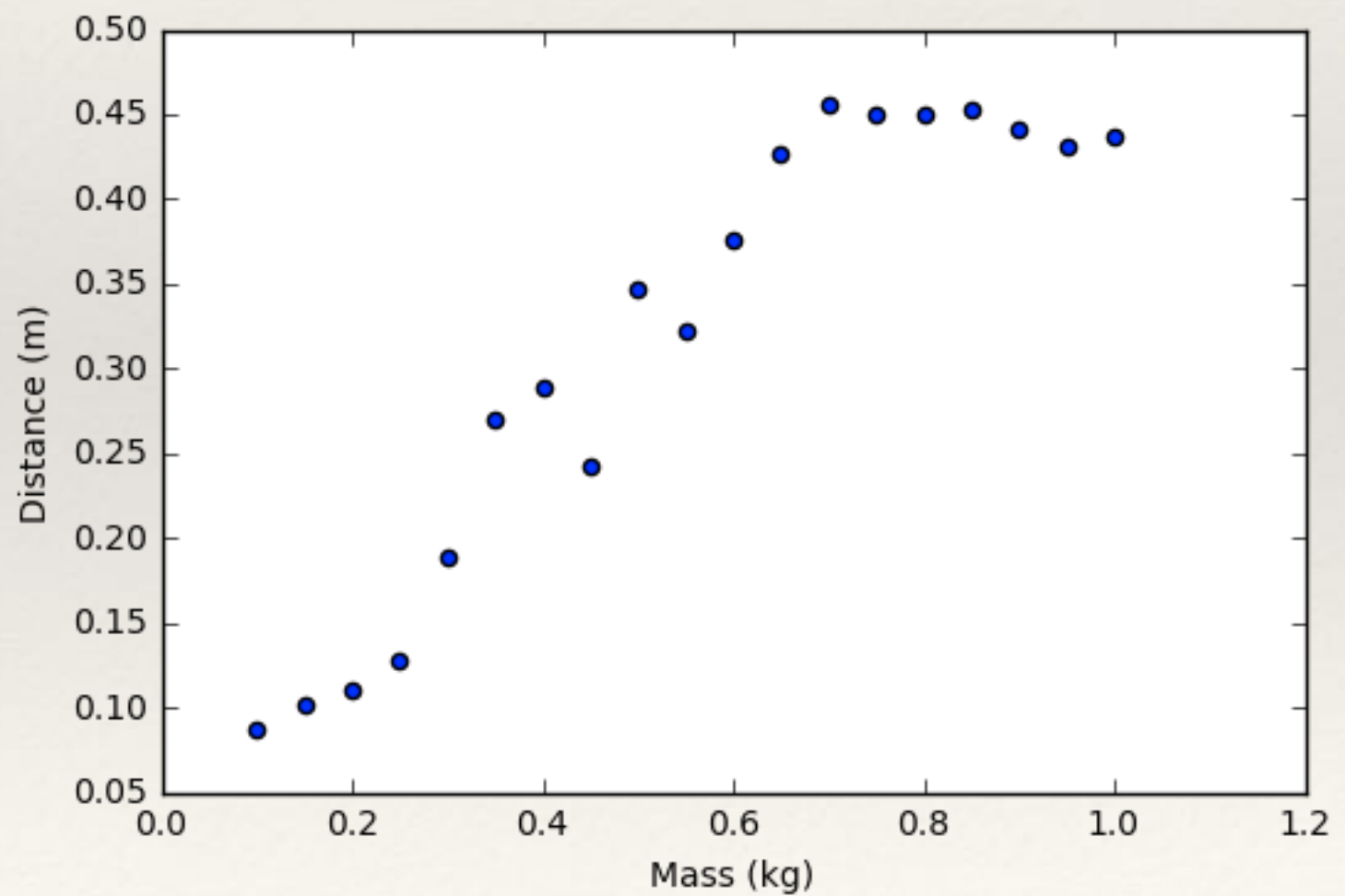
たとえば



フックの法則

$$F = -k x$$

バネの伸び



バネにかかる力

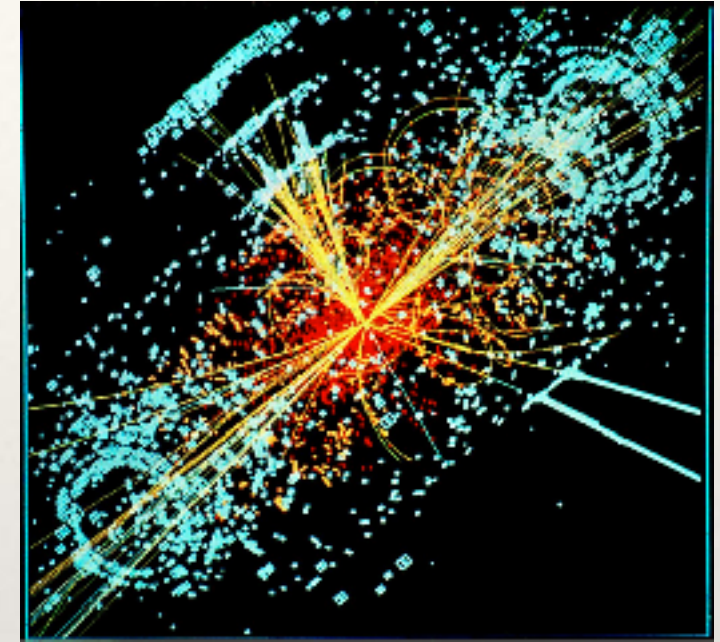
DNA配列だと

- ❖ 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



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



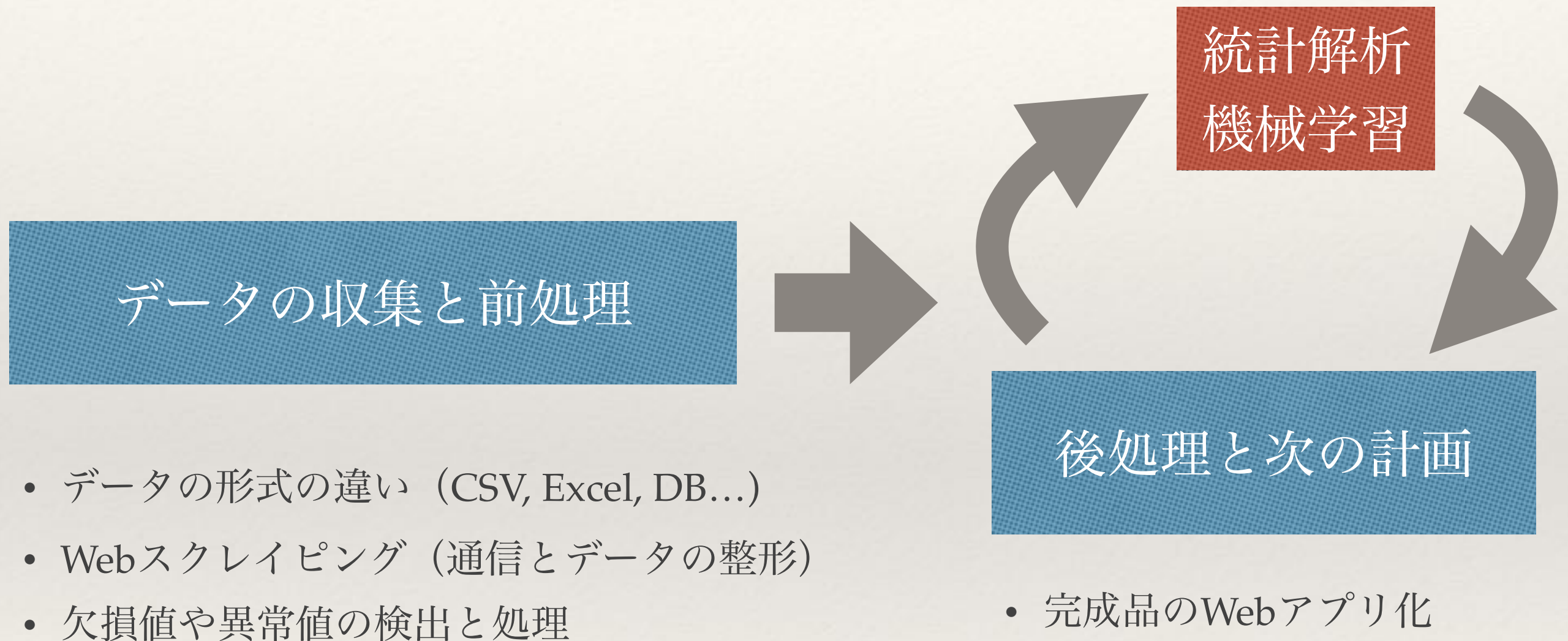
データサイエンスの必要性



データ

普遍的な何か

データサイエンスの実際



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

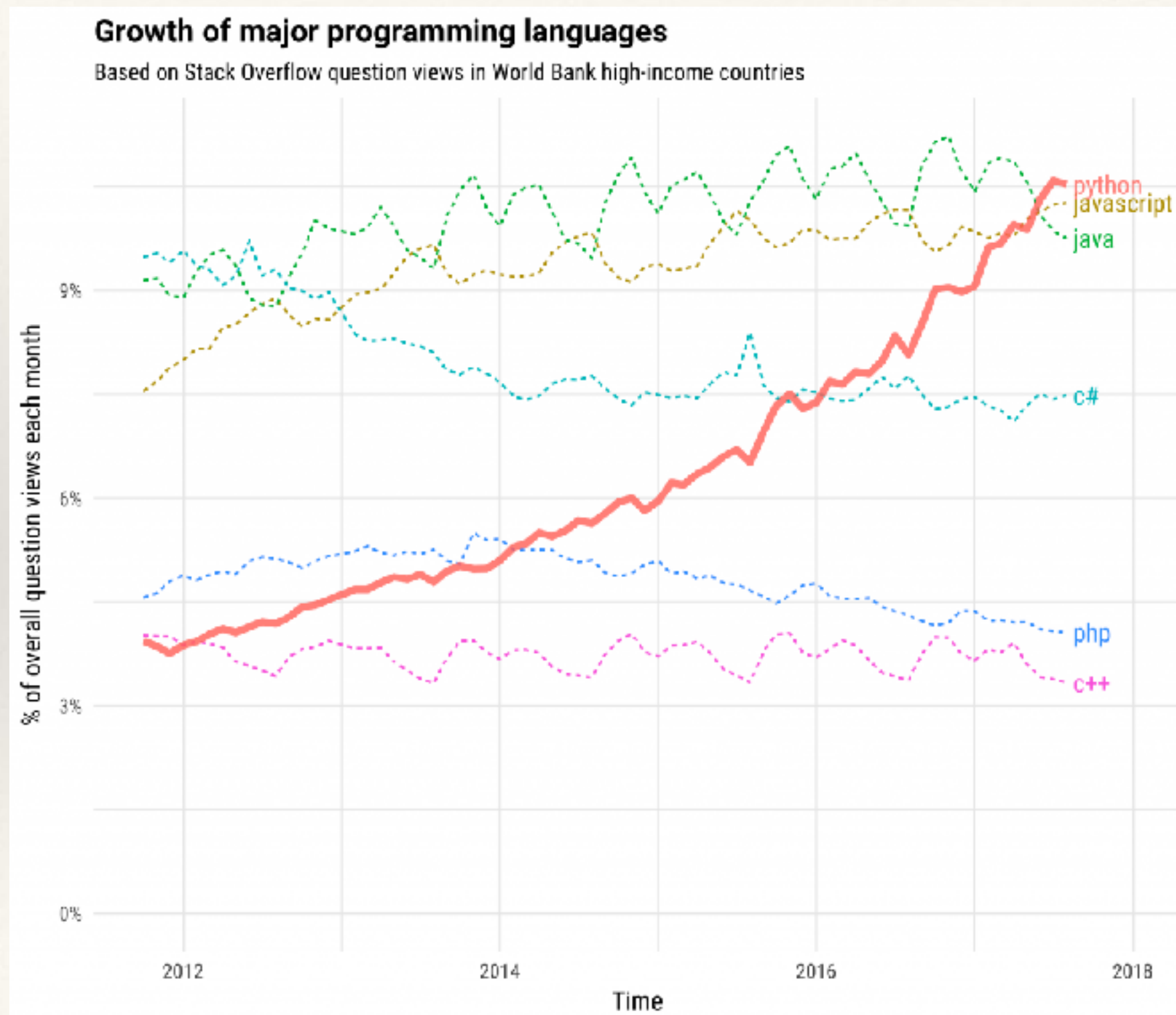
[Edit Ranking](#) | [Add a Comparison](#) | [Twitter](#) | [Facebook](#)

Language Types (click to hide)

 Web |  Mobile |  Enterprise |  Embedded

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

Stack Overflow

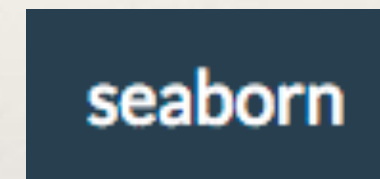


いくつかの選択肢

- ❖ 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

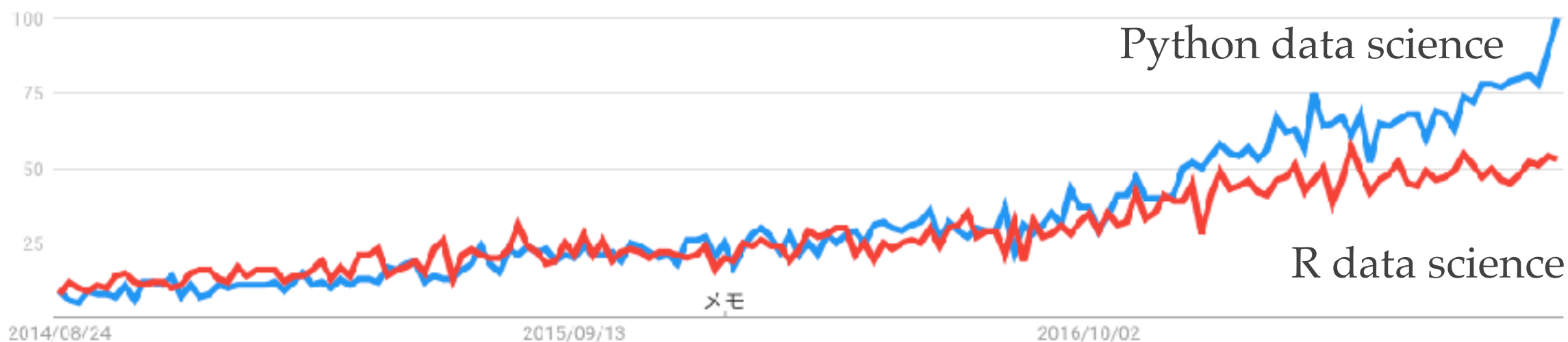


<https://www.continuum.io>

R と Python

過去3年間のGoogle Trendsのデータ

人気度の動向 ?



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

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

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

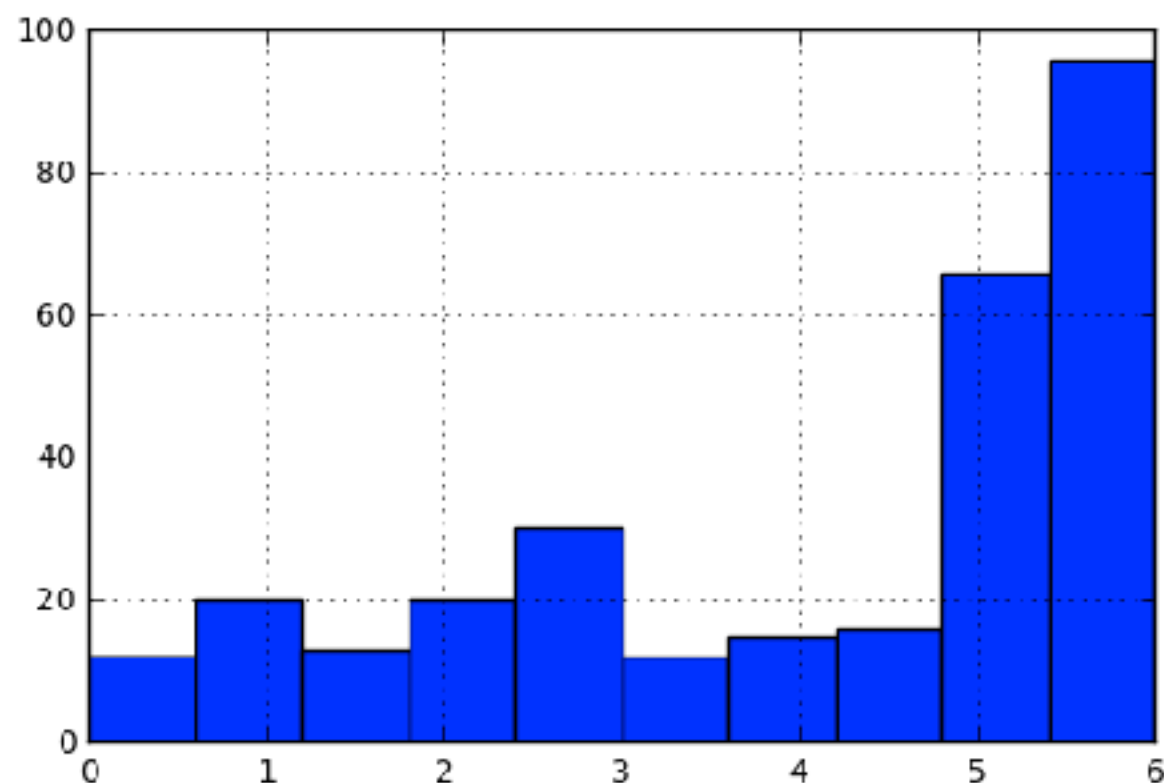
```
In [3]: data.head()
```

```
Out[3]:
```

	sample_0	sample_1	sample_2	sample_3	sample_4	sample_5	sample_6	sample_7	sample_8	sample_9
var_0	1.161546	2.027050	2.307813	1.078164	2.023768	1.093772	0.199090	1.002068	0.715403	1.895127
var_1	0.998445	0.601550	0.054010	0.028059	1.024477	0.737187	2.145073	2.395837	0.594704	0.319478
var_10	5.659790	8.118167	8.338933	8.388697	8.083567	8.349710	8.332364	8.143534	8.277051	8.338494
var_100	2.383074	2.155451	0.943400	2.090212	0.653381	0.897784	1.413244	0.544167	2.494165	0.101343
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)
```

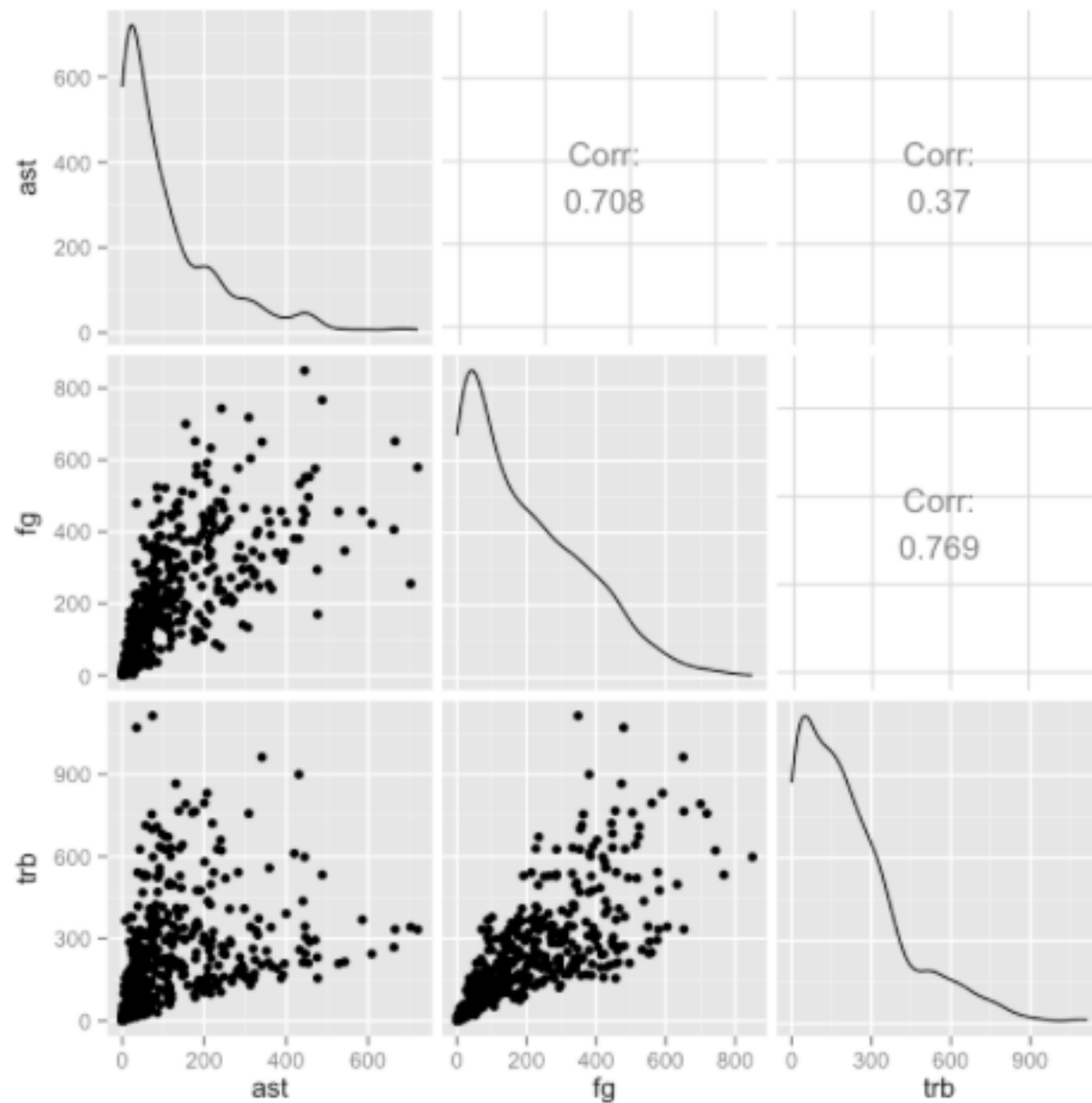
```
Out[6]: <seaborn.axisgrid.PairGrid at 0x1146b7c88>
```



ペアプロットを描く。なんか、データが1つおかしそう・・・

R

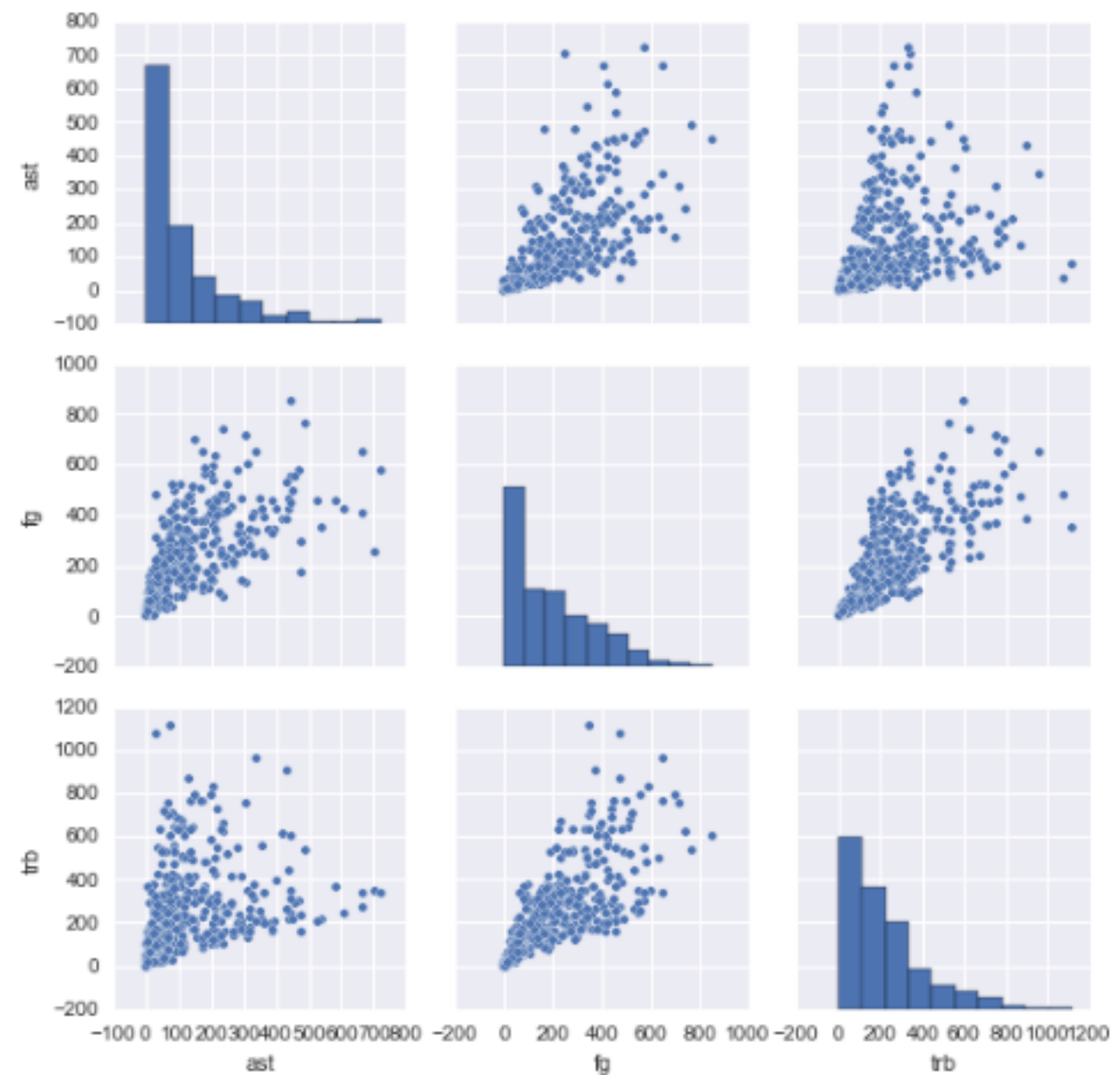
```
library(GGally)
ggpairs(nba[,c("ast", "fg", "trb")])
```



Python

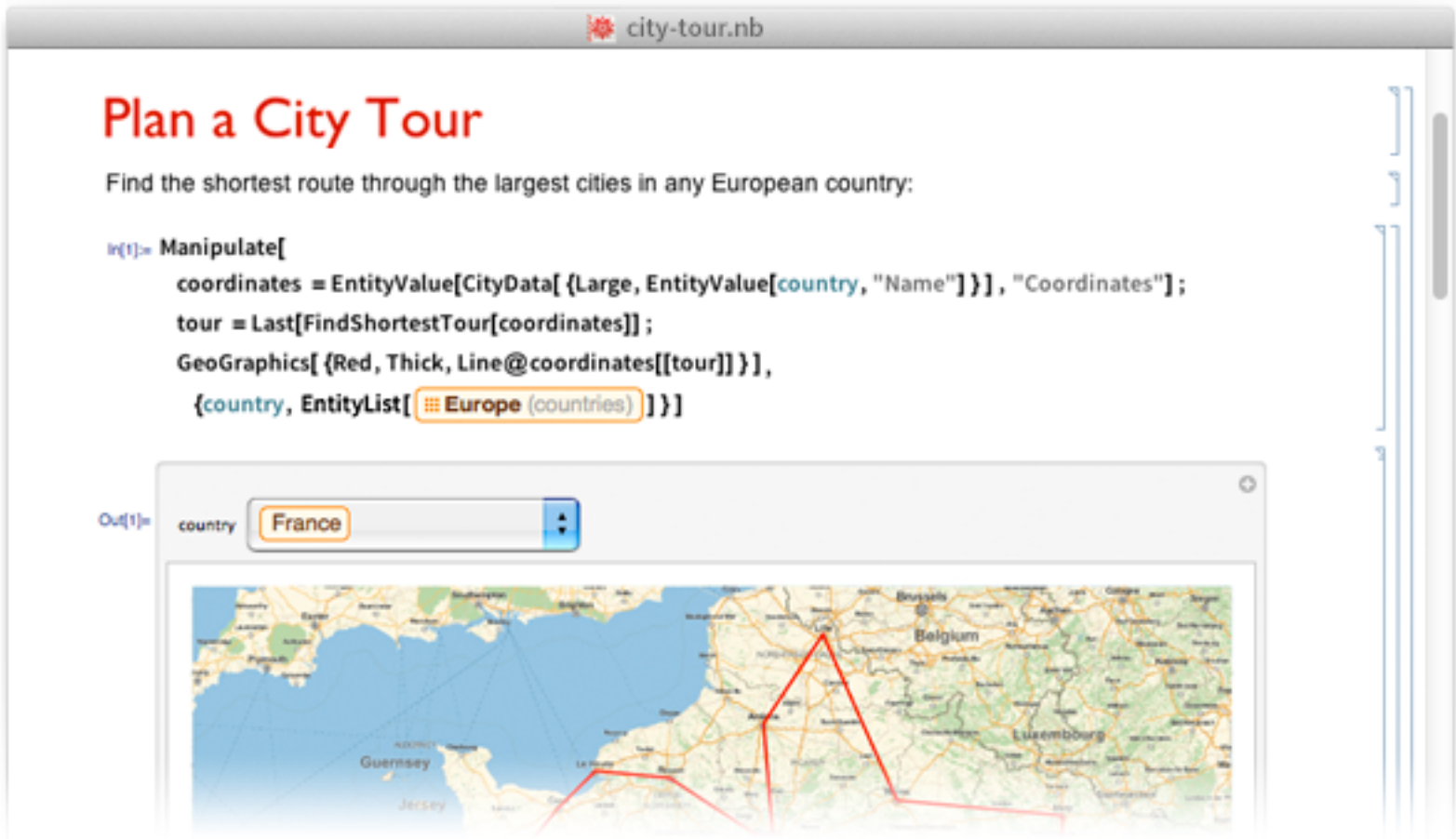
```
import seaborn as sns
import matplotlib.pyplot as plt

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



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わせることができ、デスクトップでもWebでも使えます：



The screenshot shows a Wolfram Notebook interface with the title bar "city-tour.nb". The notebook content includes a title "Plan a City Tour" in red, followed by the instruction "Find the shortest route through the largest cities in any European country:". Below this is a code cell with the following Wolfram code:

```
In[1]:= Manipulate[
  coordinates = EntityValue[CityData[ {Large, EntityValue[country, "Name"]}], "Coordinates"];
  tour = Last[FindShortestTour[coordinates]];
  GeoGraphics[ {Red, Thick, Line@coordinates[[{tour}]]},
    {country, EntityList[ Europe (countries) ]}]
```

The output cell shows a dropdown menu for "country" with "France" selected. Below the dropdown is a map of France with a red line indicating the shortest route through several cities. The map also shows neighboring countries like Belgium and Luxembourg, and islands like Guernsey and Jersey.

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。

Pythonが台風のように

- ❖ 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

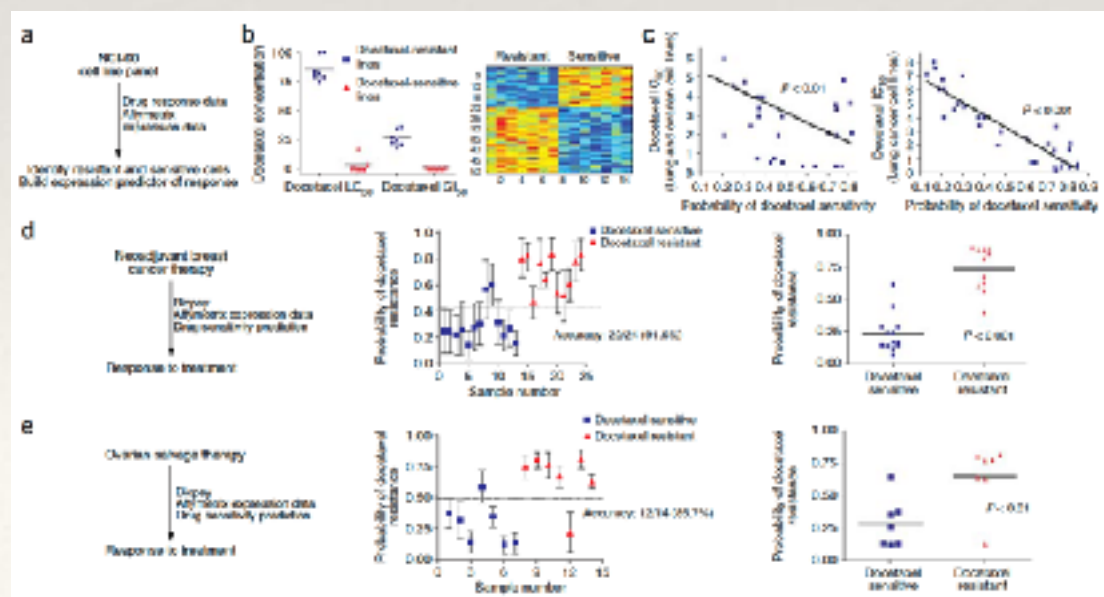
	R	Python
Kolmogorov–Smirnov検定	○	scipy.stats
Shapiro-Wilk検定	○	scipy.stats
Hosmer-Lemeshow検定	ResourceSelection	?

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 Chan⁴, Robyn Sayer⁴, Daniel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo¹⁻³, Johnathan Lancaster⁴ & Joseph R Nevins¹⁻³

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](#) [IF 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



[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)

[What links here](#)

[Article](#)

[Talk](#)

Anil Potti

From Wikipedia, the free encyclopedia

Anil Potti is a physician and former [Duke University School of Medicine](#) [oncogenomics](#). He, along with Joseph Nevins, was involved in a [scandal at Duke University](#)^{[1][2][3]} and many [papers](#) were [retracted](#). On November 9, 2015, the [Office of Research Integrity](#) [engaged in research misconduct](#).^[4] According to the [ORI](#), Potti can continue to perform research until the year 2020, while he "neither admits nor denies

Contents [\[hide\]](#)

- [1 Biography](#)
- [2 Scientific misconduct](#)
- [3 Consequences](#)
- [4 Research questions and Institute of Medicine](#)
- [5 Continuing controversy](#)
- [6 Retracted papers](#)

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？

まとめ

- ❖ データサイエンスはデータ駆動型サイエンス
 - ❖ 大量データを巧みに処理して真理を探究
- ❖ 前処理負担が重いので、汎用言語が役に立つ
 - ❖ データの整形やスクレイピング、Webアプリ化など
- ❖ 再現性は重要
 - ❖ オープンな環境がそれを支える

ご静聴ありがとうございました