



# Python的数据工具箱

肖凯

# 关于我

- 喜欢折腾数据
- 1号店-商务智能部
- 《数据科学中的R语言》

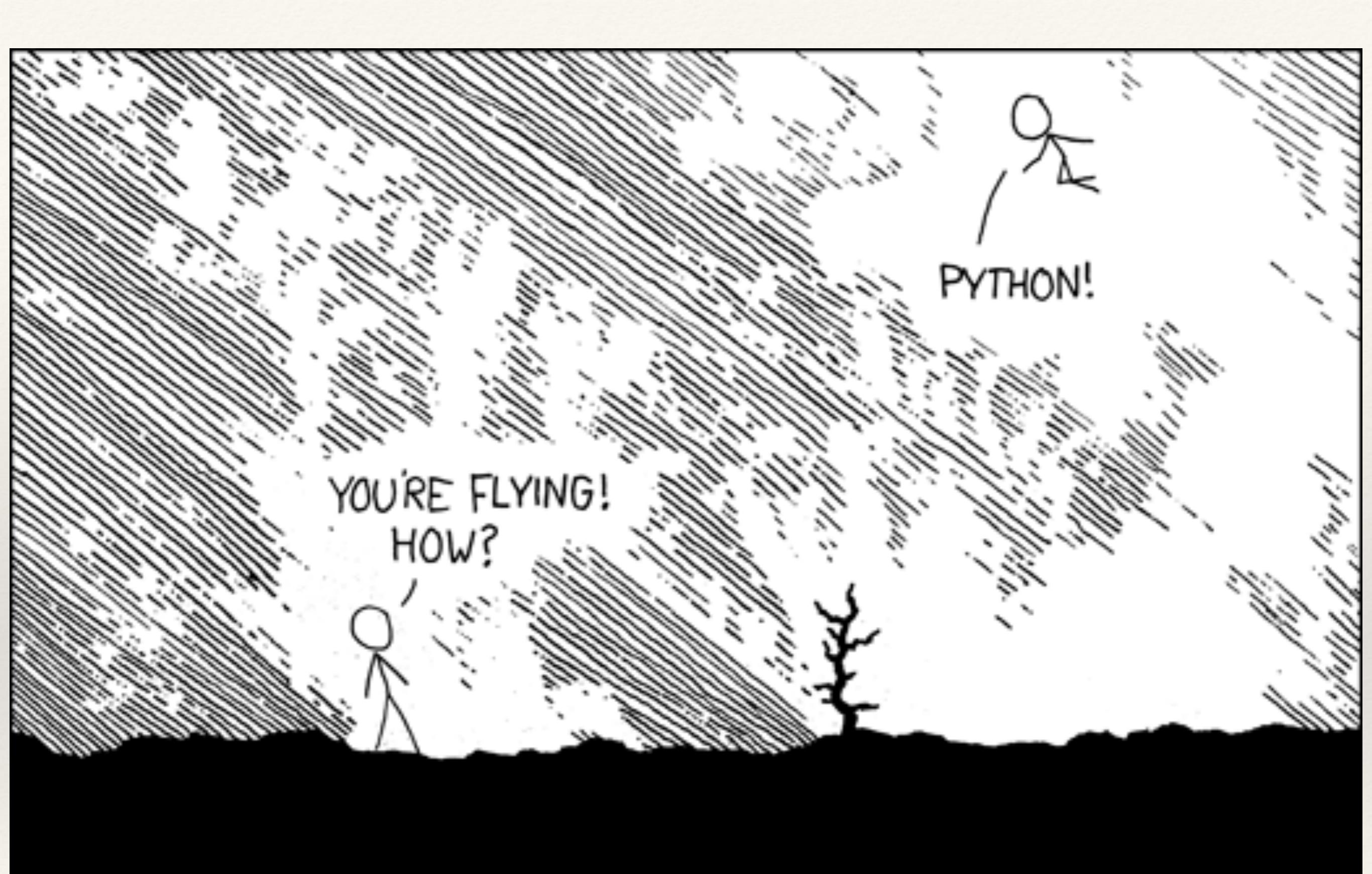


---

# 哪一项最重要

---

- 问题
- 数据
- 方法
- 工具



YOU'RE FLYING!  
HOW?

PYTHON!

# 提纲

---

- 为何使用Python
- Python VS R
- 数据相关模块
- 几个小例子
- 学习资源

# 为何使用Python

- 开源的通用性语言
- 整合能力强
- 扩展模块丰富
- 入门比较简单，语法比较优美
- 解释性语言，不需编译，比较灵活
- 能快速产生结果，从原型到产品

Why don't you use C instead  
of Python? It's so much faster!



Why don't you commute by  
airplane instead of by car? It's  
so much faster!



# 所谓的“弱点”

- 解释型语言，数据全部读入内存，计算速度较慢
- **人的时间比计算机的时间宝贵。**设想下面的场景：
  - 例如某零售公司要研究客户流失情况，想知道哪些因素比较重要，并形成一套可用的预测模型
  - C/C++：开发用了20小时，运行0.1秒，维护困难
  - Python：开发用了2小时，运行60秒，许多库包含现成的统计工具和可视化工具

# 为何使用Python

- 填补数据研究和产品开发之间的鸿沟
- 配合数据科学家完成多领域任务 ( Bigdata/Deeplearning/NLP )

## Analyst

- Uses graphical tools
- Can call functions, cut & paste code
- Can change some variables

Gets paid for:  
**Insight**

**Excel, VB, Tableau,**  
**Python**

## Analyst / Data Developer

- Builds simple apps & workflows
- Used to be "just an analyst"
- Likes coding to solve problems
- Doesn't want to be a "full-time programmer"

Gets paid (like a rock star) for:  
**Code that produces insight**

**SAS, R, Matlab,**  
**Python**

## Programmer

- Creates frameworks & compilers
- Uses IDEs
- Degree in CompSci
- Knows multiple languages

Gets paid for:  
**Code**

**C, C++, Java, JS,**  
**Python**

---

# Python VS R

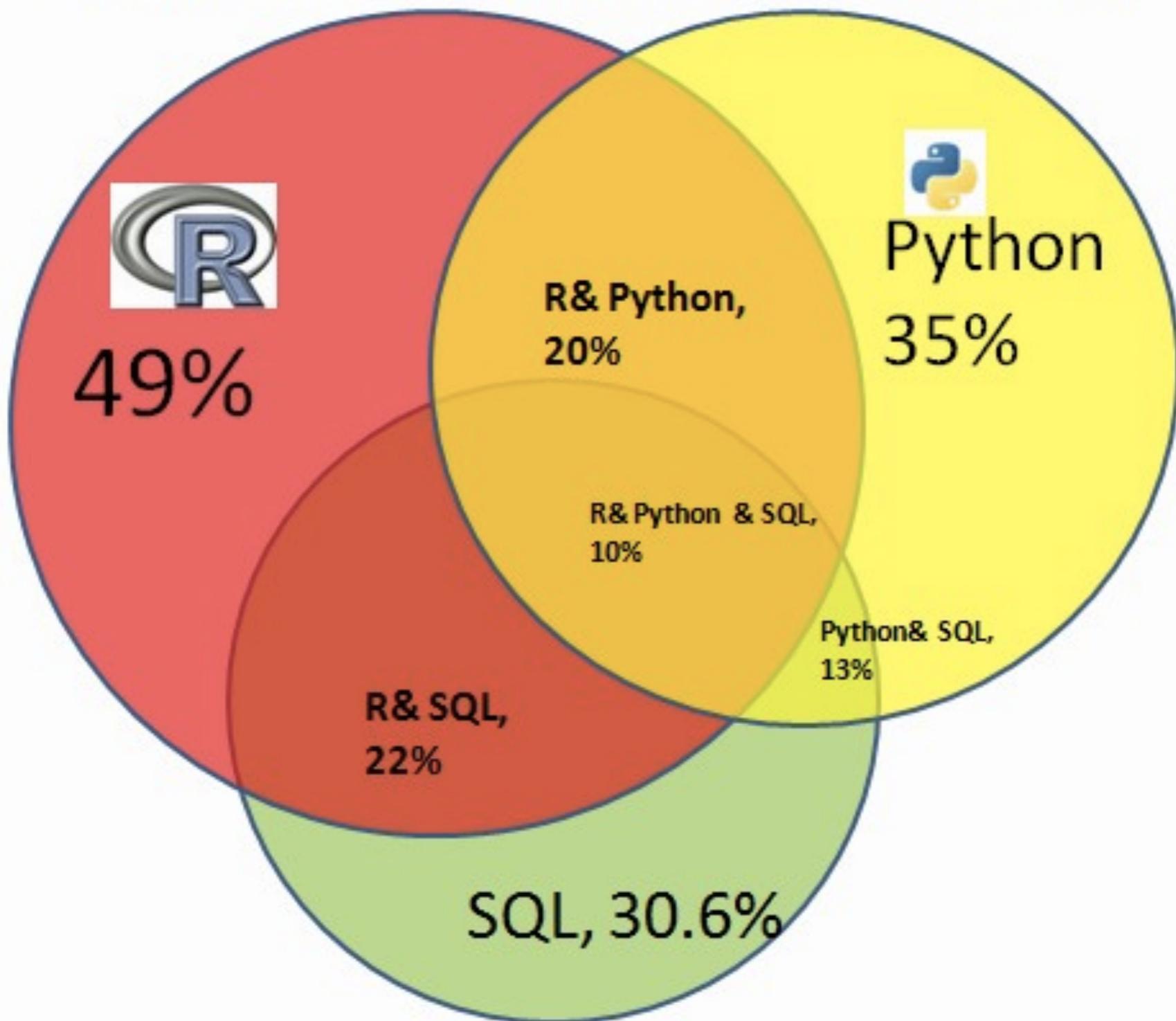
---

两种工具的相同点：

- 均为开源免费
- 均可在三种操作系统中运行
- 均有大量的用户群和社区支持
- 均有大量的扩展包和教程资源
- 调查显示它们是业界人士最为喜爱的两种工具

# Python VS R

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining



---

# Python VS R

---

两种工具的差异点：

- Python是一种通用编程工具，R偏向于统计专业
- R有更为丰富的统计分析函数，Python长于机器学习
- R有更好的可视化包，Python正在进步
- Python和R的核心语法非常简洁，R包的语法兼收并蓄，错综复杂

---

# Python VS R

---

R:

```
results <- lm(y ~ x1 + x2 + x3, data=dataframe)
```

Python:

```
results = sm.OLS(y, X).fit()
```

---

# Python VS R

---

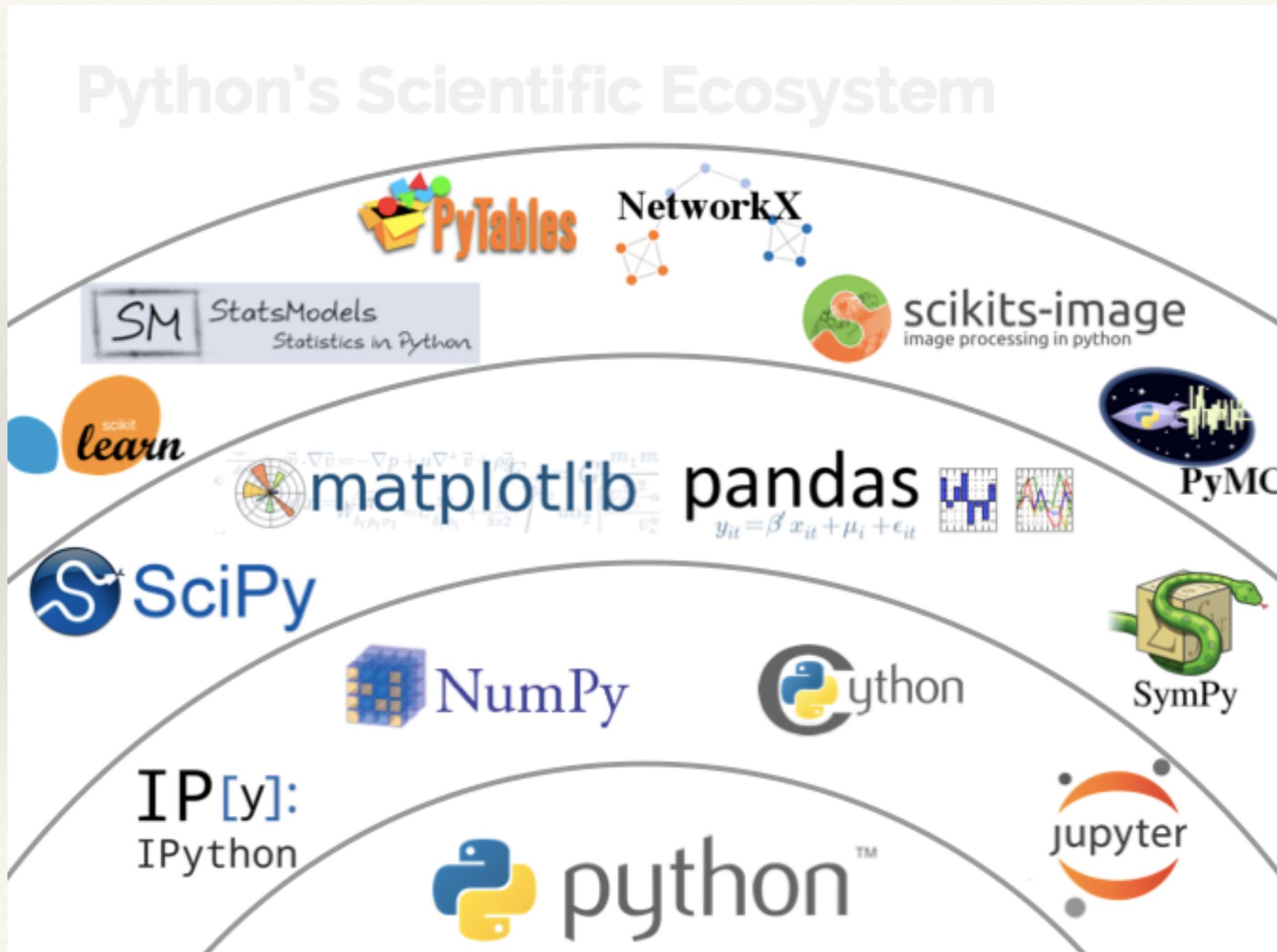
R:

- 如果你不是计算机背景
- 未来有很强的学术化需求

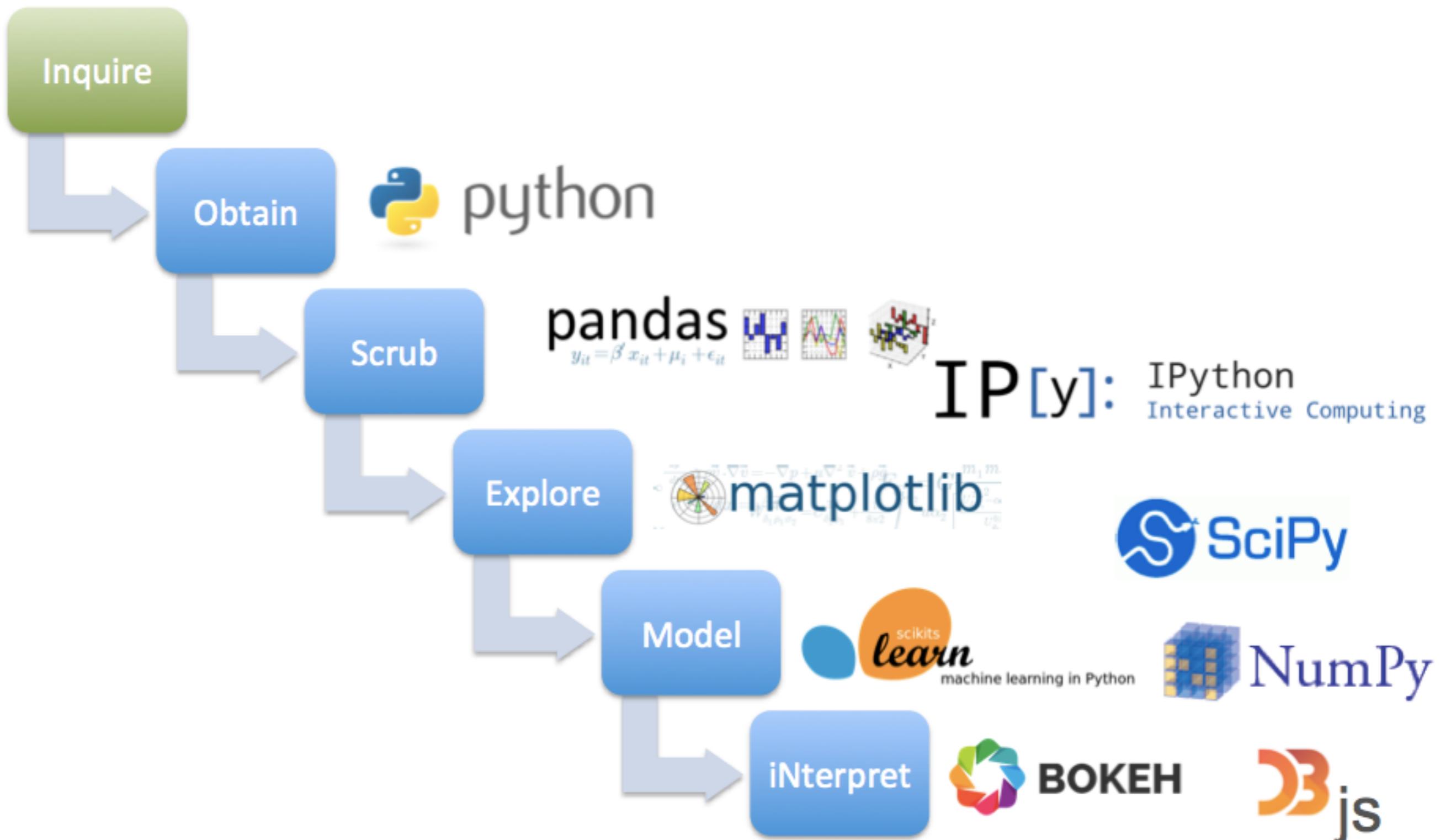
Python:

- 如果你是计算机背景
- 未来有很强的工业化需求

# Python科学计算库



# 分析流程中的Python



# 数据相关模块

- IPython: 增强的交互式运行环境
- NumPy : 数组数据结构和矩阵计算
- SciPy : 科学计算
- Matplotlib : 数据绘图
- Pandas : 提供data frames数据结构
- Statsmodels: 统计模型
- Scikit-learn: 机器学习

# 数据相关模块

- Requests: 网页数据抓取
- BeautifulSoup: 解析网页数据
- Flask: 轻量级的web框架
- sqlite3: 轻量级数据库接口
- Pyspark: Spark的Python接口
- nltk: 自然语言处理
- networkx: 社交网络分析
- theano: 深度学习

---

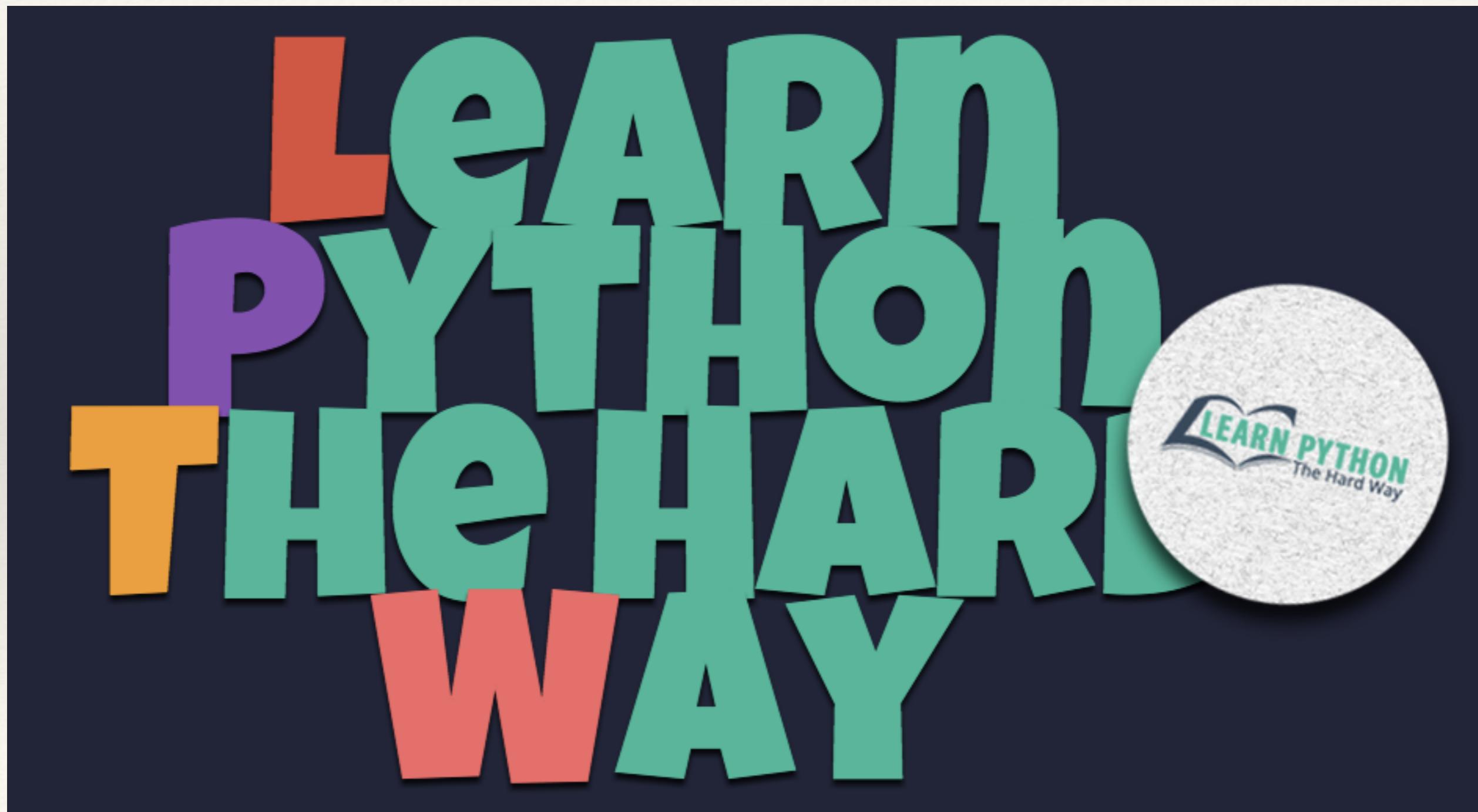
# 科学计算套件

---



**Anaconda**

# 先修知识



# 运行环境

ipython是一个增强的python shell

- 提高编写、测试、调度代码的速度
- 提供了IPython Notebook，是一个交互计算平台，也是一个记录计算过程的笔记本
- 满足交互计算和批处理计算，同时能保存脚本文件以记录计算过程
- 能兼容markdown等语法，满足可重复数据分析的需求，以及课程教学、博客写作
- 能在本地的计算机上对远程服务器中的数据进行分析

# IPython

IPy spectrogram

https://localhost:8888/8995fc4a-410d-4118-9df8-211274fdce87#

IP[y]: Notebook spectrogram Last saved: Jun 28 10:35 AM Logout

File Edit View Insert Cell Kernel Help

Code

## Simple spectral analysis

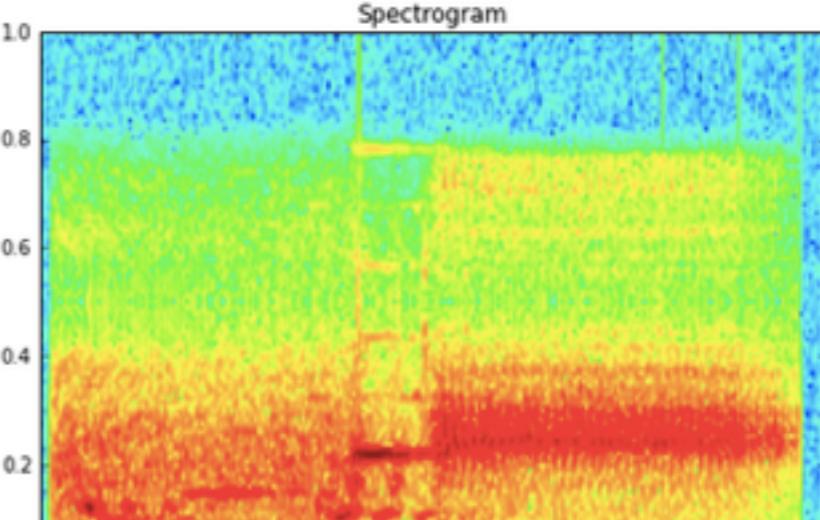
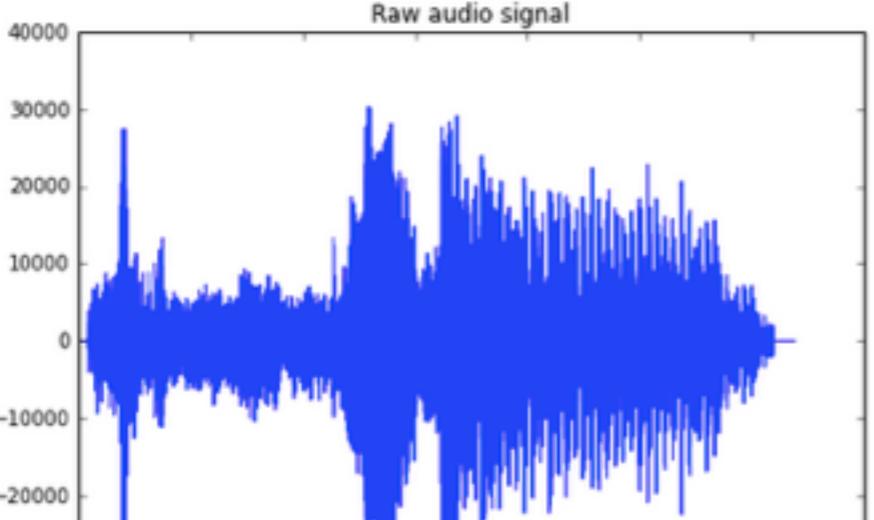
An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n \exp \frac{-2\pi i}{N} kn \quad k = 0, \dots, N-1$$

```
In [2]: from scipy.io import wavfile  
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))  
ax1.plot(x); ax1.set_title('Raw audio signal')  
ax2.specgram(x); ax2.set_title('Spectrogram');
```



# 数值计算

numpy：科学计算的基础包

- 快速高效的多维数组对象
- 可执行向量化计算
- 提供线性代数等矩阵运算
- 可集成C的代码

# NumPy

```
>>> a[0,3:5]  
array([3,4])
```

```
>>> a[4:,:4:]  
array([[44, 45],  
       [54, 55]])
```

```
>>> a[:,2]  
array([2,22,52])
```

```
>>> a[2::2,:,:2]  
array([[20,22,24],  
       [40,42,44]])
```

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

# Numba

**Numba:** with a simple decorator, Python JIT compiles to LLVM and executes at near C/Fortran speed!

```
@numba.jit
def fib(n):
    a, b = 0, 1
    for i in range(n):
        a, b = b, a + b
    return a

@timeit(fib(50))
```

1 loops, best of 3: 468 ns per loop

20x speedup!

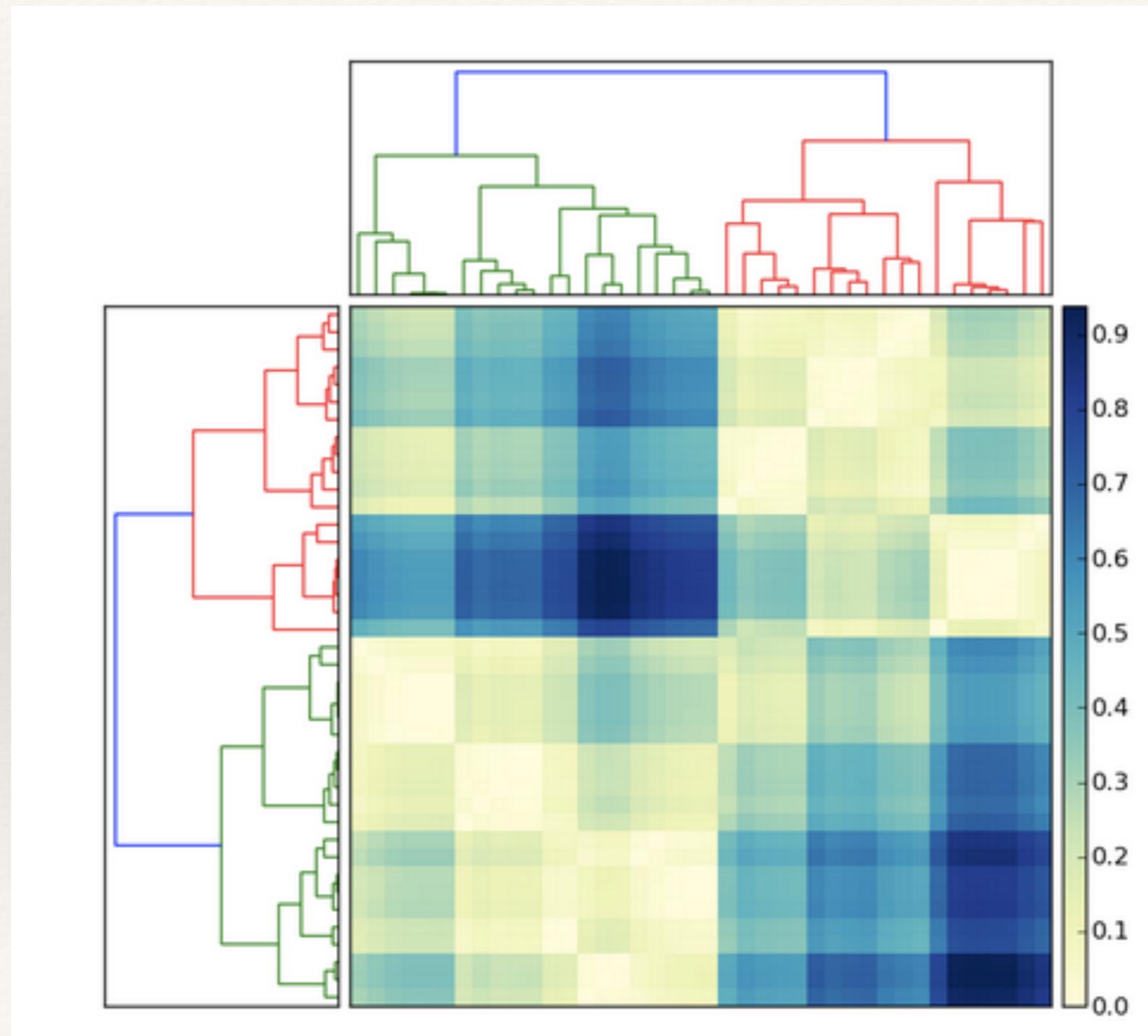
# SciPy

---

science python简称，用于解决科学计算中标准问题

- 数值积分和微分方程求解
- 扩展的矩阵计算功能
- 最优化工具
- 概率分布计算和统计函数
- 信号处理函数

# SciPy

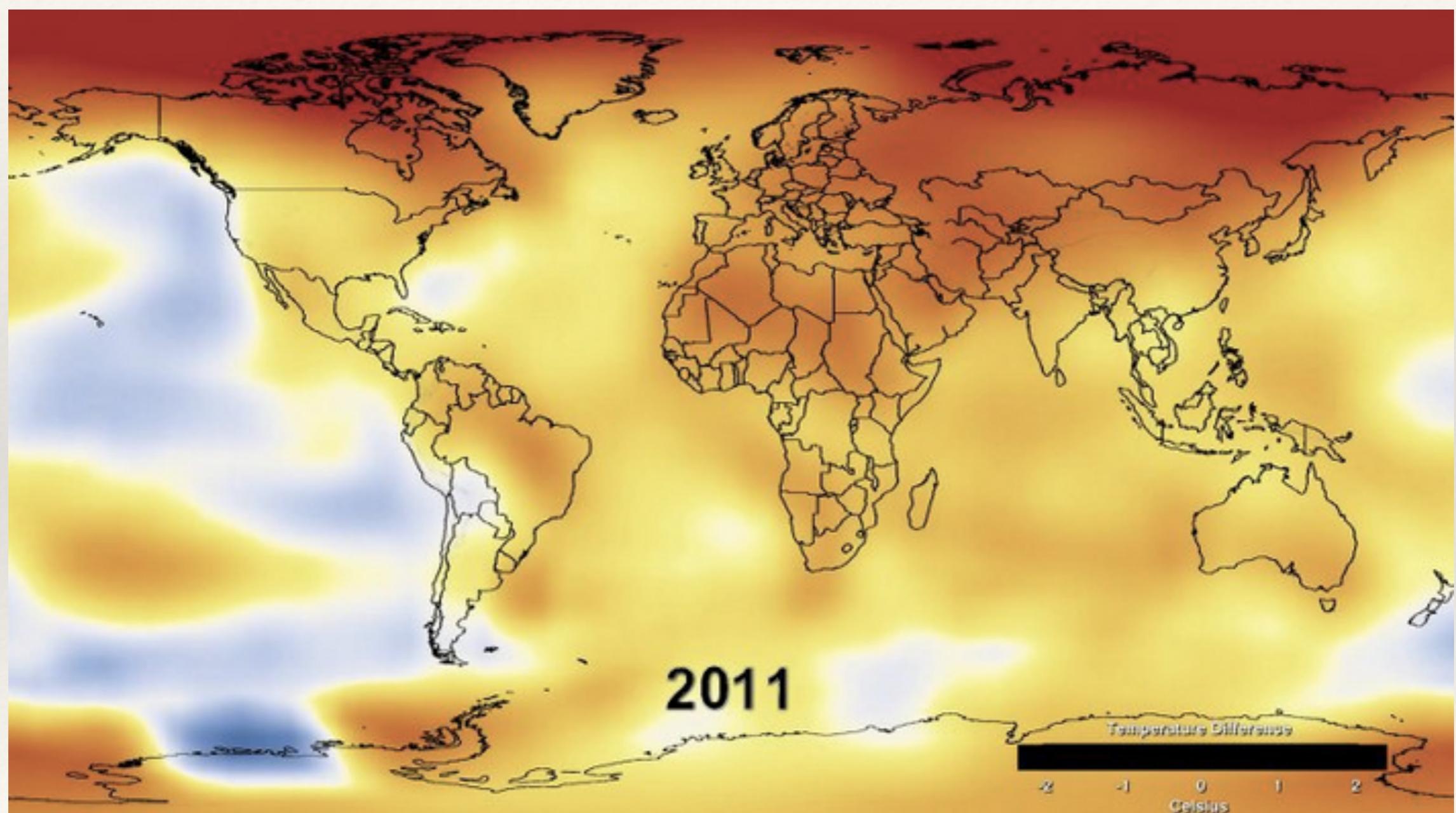


# 数据可视化

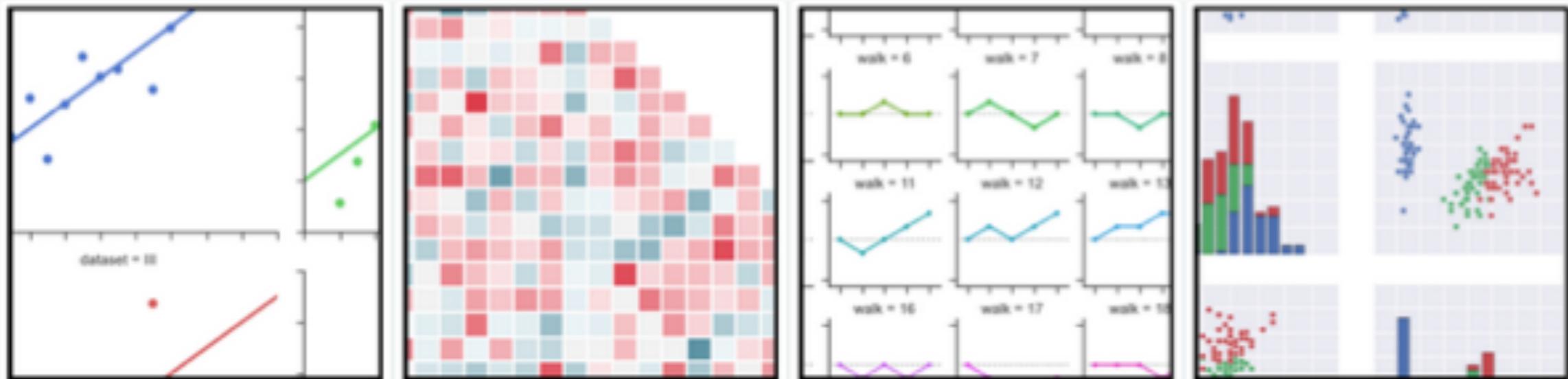
Matplotlib是python下最著名的绘图库

- 提供了一整套和matlab相似的命令API
- 十分适合交互式绘图
- 也可将它作为绘图控件，嵌入GUI应用程序中

# Matplotlib



# Seaborn



- built on top of **matplotlib**: able to use any of its backends & output formats
- **pandas**-aware: quick plotting of labeled data
- provides beautiful, well-thought-out default plot styles

# Bokeh



- HTML5 output, both server and client-side
- Flexible in-browser interactivity
- Fundamentally a **Javascript library** with Python bindings

# 数据分析

Pandas：用于数据处理和分析

- 易用、高效的数据操作函数库
- 执行join以及其他SQL类似的功能来重塑数据
- 提供包括dataframe在内的数据结构
- 支持各种格式（包括数据库）输入输出数据
- 支持时间序列
- 拥有基本绘图功能和统计功能

# Pandas

```
In [17]: df_concat = pd.concat([df_1, df_2, df_3])
```

```
In [18]: df_concat.head()
```

Out[18]:

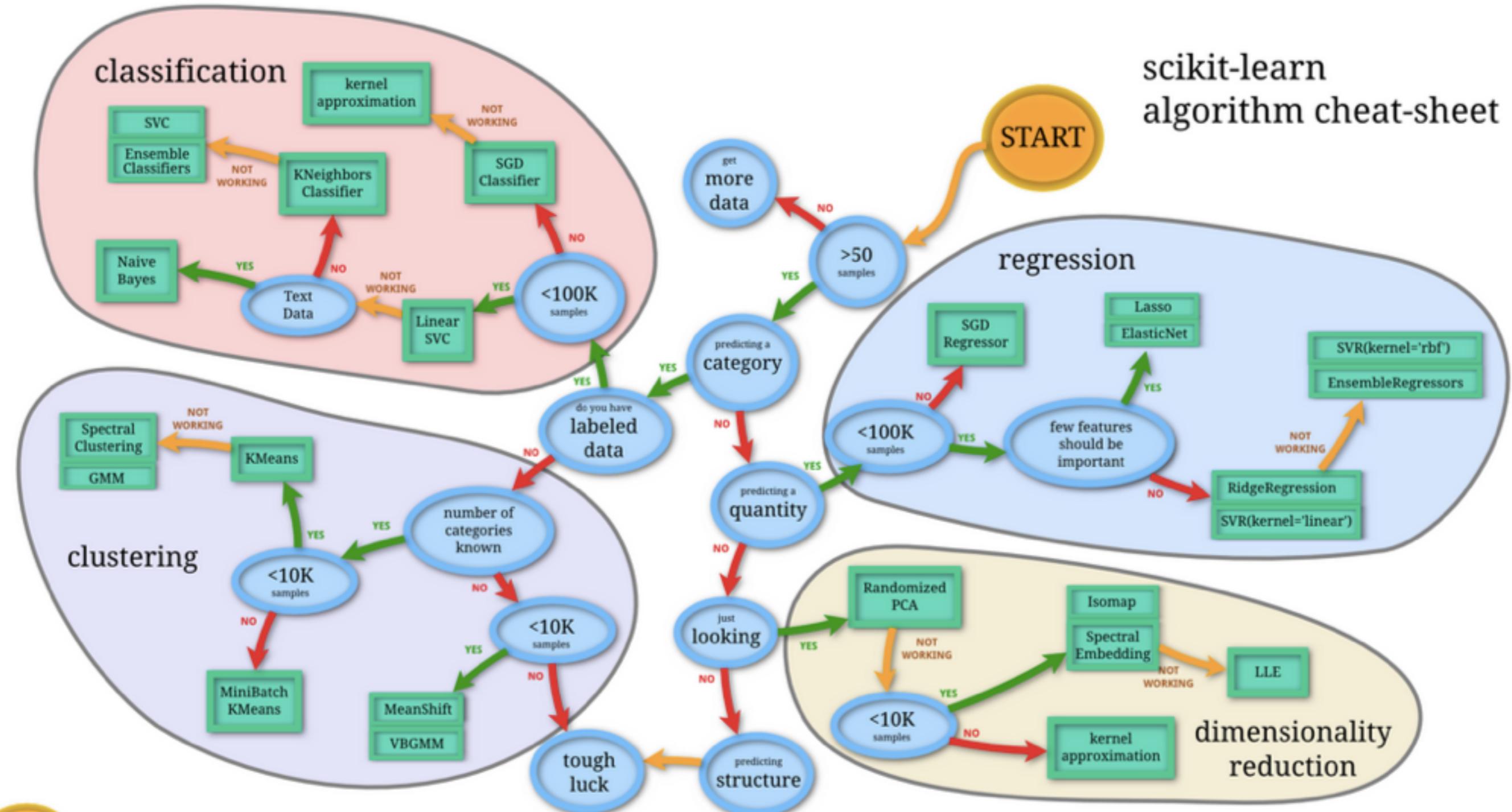
	Words	Counts
0	also	9
1	cells	8
2	one	8
3	two	7
4	expression	7

# 机器学习

Scikit-learn：机器学习库

- 建立在NumPy, SciPy基础上的机器学习库
- 通过一个统一的接口来使用，有助于迅速地在数据集上实现流行的算法。
- 含了许多用于标准机器学习任务的工具，如：聚类、分类和回归等。

# Scikit-learn



# 案例1：数据抓取和绘图

- 用python抓取豆瓣电影数据
- 用R绘图来分析探索评分

豆瓣电影  

影讯&购票 选电影 电视剧 排行榜 分类 影评 预告片 问答

## 豆瓣电影TOP250

排名	电影名称	导演	主演	年份	国家	类型	评分	评价数
1	肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]	导演: 弗兰克·德拉邦特 Frank Darabont	主演: 蒂姆·罗宾斯 Tim Robbins / ...	1994	美国	犯罪 剧情	★★★★★ 9.6	622659人评价

□ 我没看过的

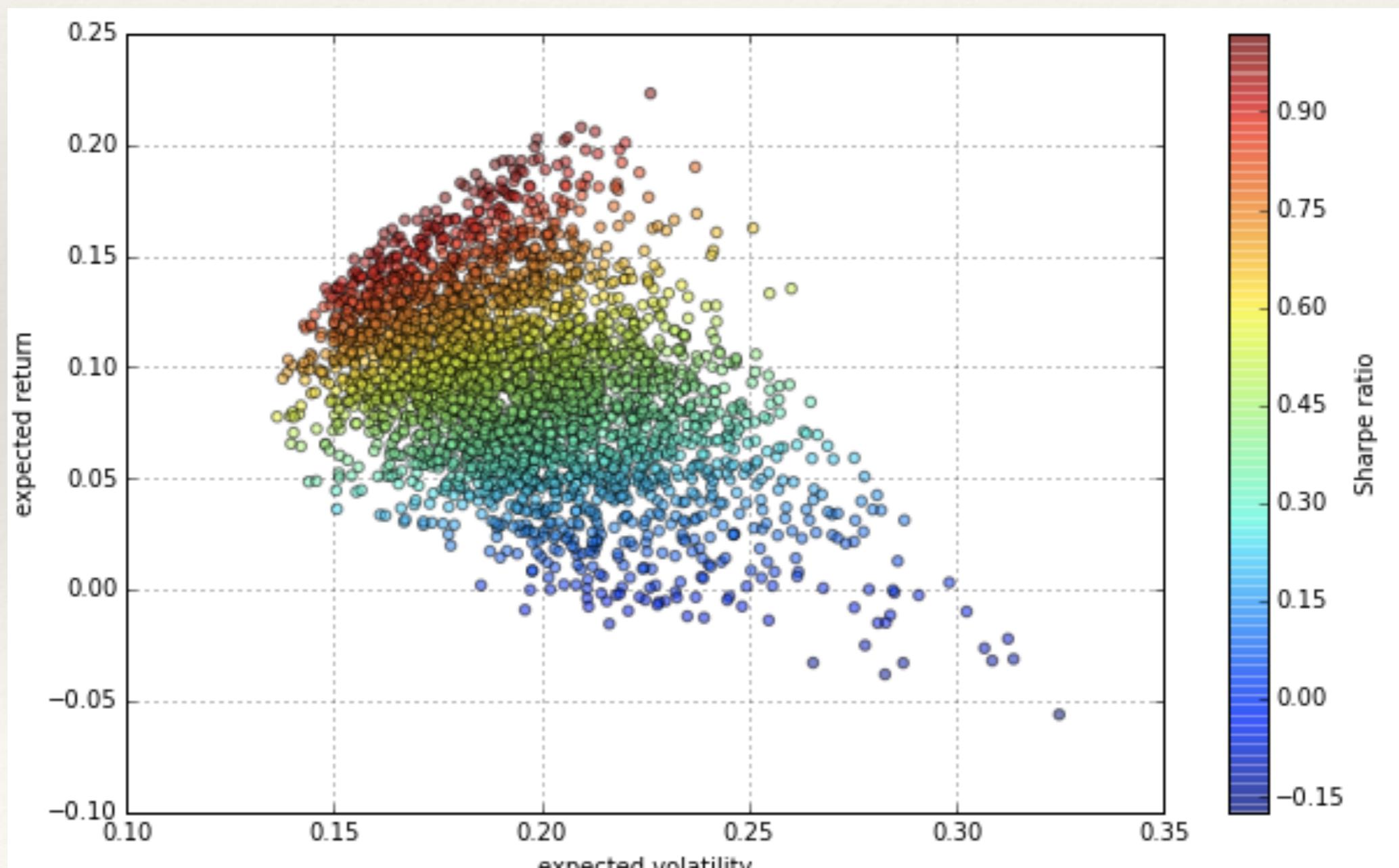
豆瓣用户每天都在对“看过”的电影进行“很差”到“力荐”的评价，豆瓣根据每部影片看过的人数以及该影片所得的评价等综合数据，通过算法分析产生豆瓣电影250。

 **豆瓣电影客户端**  
让买票看电影更简单 >

“希望让人自由。”

# 案例2：选择最优资产组合

- 获取四种股票的历史数据
- 使用最优化函数来计算最优资产组合



# 案例3：新闻分类

- 用文本函数处理搜狗新闻语料
- 建立新闻自动分类模型
- 用Flask快速构建数据产品原型

	precision	recall	f1-score	support
0	0.91	0.88	0.89	576
1	0.86	0.83	0.84	604
2	0.88	0.83	0.86	616
3	0.99	0.97	0.98	580
4	0.87	0.88	0.88	597
5	0.88	0.80	0.83	607
6	0.78	0.89	0.83	599
7	0.74	0.79	0.76	613
8	0.92	0.93	0.92	579

avg / total    0.87    0.86    0.87    5371

## 新闻自动分类

请输入新闻

据解放军空军政治部出版的《空军报》7月23日报道，7月21日，空军将官军衔晋升仪式在京举行，中央军委委员、空军司令员马晓天宣读中央军委命令，空军政委于忠福主持仪式。

晋衔仪式在庄严的《中华人民共和国国歌》声中开始。空军司令员马晓天上将宣读了中央军委主席习近平签署的命令，马晓天、于忠福分别为晋升军衔的军官颁发了命令状。

提交

新闻类别属于：军事

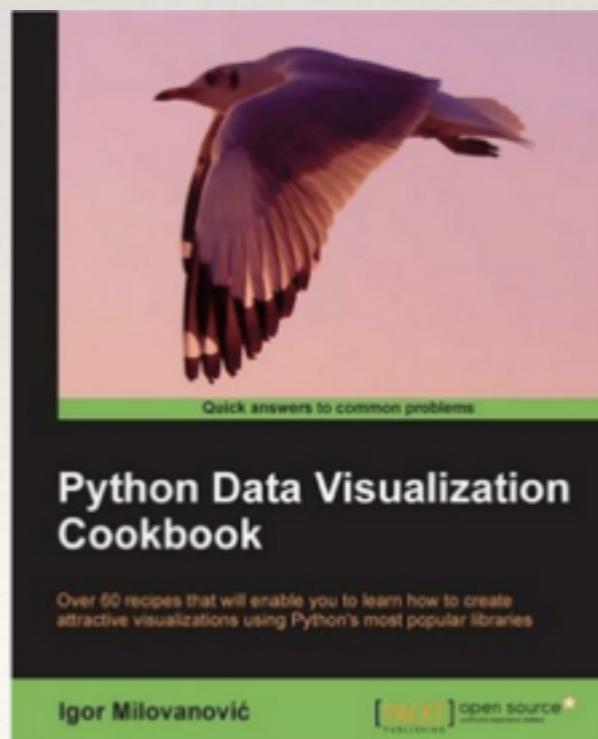
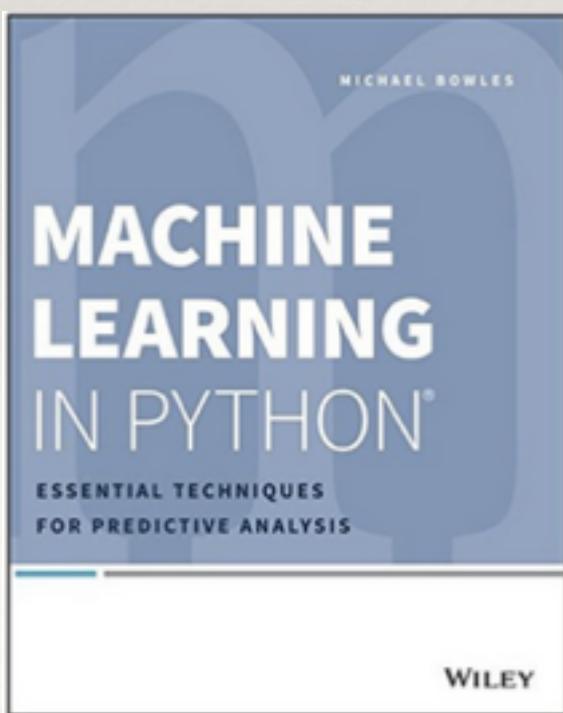
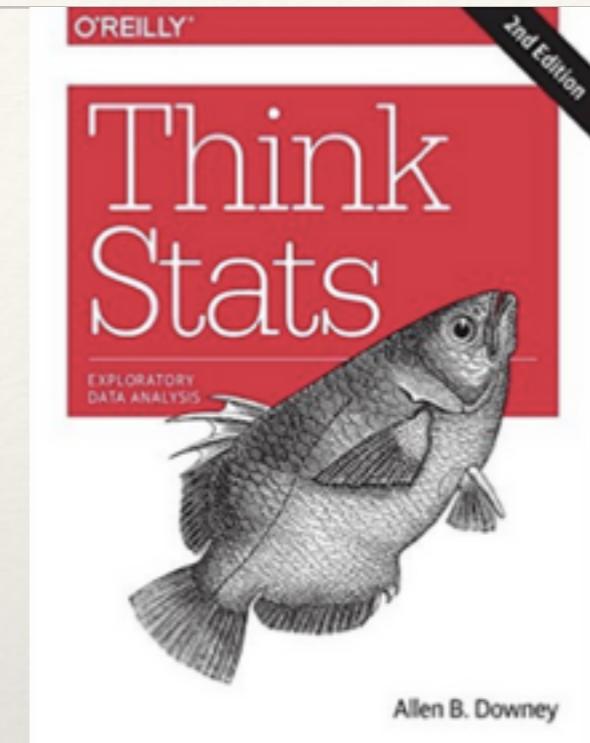
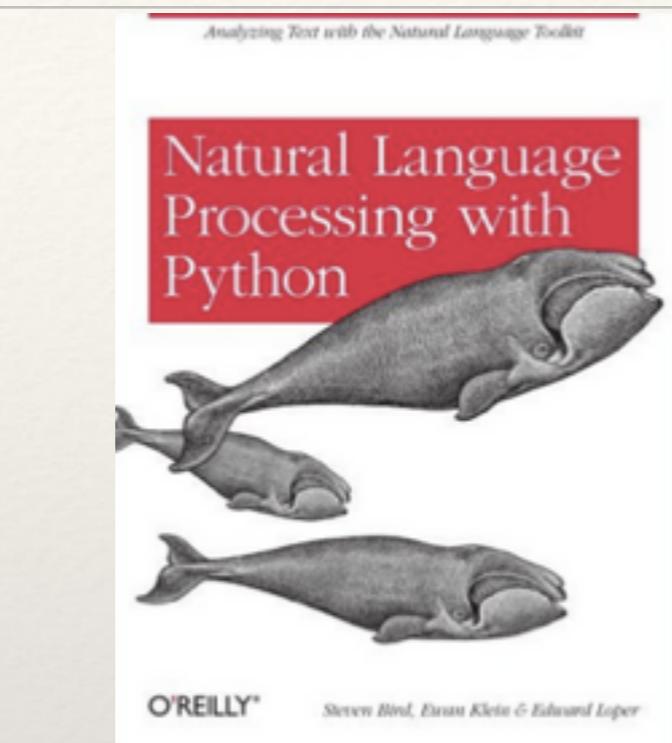
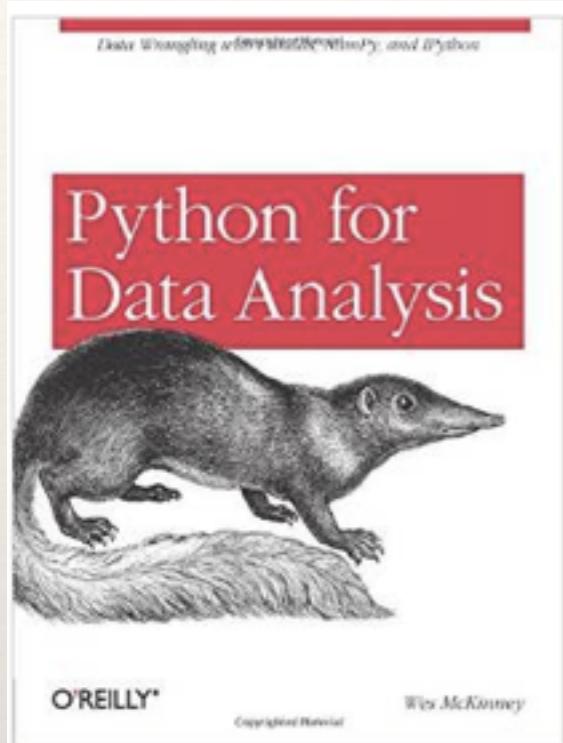
# 如何学习



# 学习的方法

- 通过**阅读**来学习
  - 包括了阅读经典的教材、代码、论文、学习公开课。
- 通过**牛人**来学习
  - 包括同行的聚会、讨论、大牛的博客、微博、twitter、RSS。
- 通过**练习**来学习
  - 包括代码练习题、参加kaggle比赛、解决实际工作中的难题。
- 通过**分享**来学习
  - 包括自己写笔记、写博客、写书、翻译书，和同伴分享交流、培训新人。

# 推荐阅读



---

# The End

---