开智学堂

数据科学班 第1讲

工具基础

肖凯

大纲

- Linux基础
- Python数据工具箱
- IPython入门
- 补充阅读材料/练习题

一、Linux 基础

文件操作

- 创建目录：mkdir
- 删除：rm
- 删除非空目录：rm -rf file
- 移动：mv
- 复制：cp (复制目录：cp -r)

文件操作

- 创建目录：mkdir
- 删除：rm
- 删除非空目录：rm -rf file
- 移动：mv
- 复制：cp (复制目录：cp -r)

文件操作

- 找到文件/目录位置：cd
- 显示当前目录下的文件: ls
- 搜寻文件或目录: find
  - find ./ -name "core*"
- 查看文件：cat vi more

开智学堂

数据分析有关操作

- head -n 3 data.csv
- tail -n 3 data.csv
- 使用grep查询文件内容
  - grep 'data' todo.txt

数据分析有关操作

- wc -l file 统计行数
- wc -w file 统计单词数
- wc -c file 统计字符数
- 统计/home/han目录(包含子目录)下的所有js文件：
  - ls -lR /home/han | grep js | wc -l

数据分析有关操作

- sort 排序
  - -n 按数字进行排序 VS -d 按字典序进行排序
  - -r 逆序排序
  - -k N 指定按第N列排序
  - sort -nrk 1 data.txt

数据分析有关操作

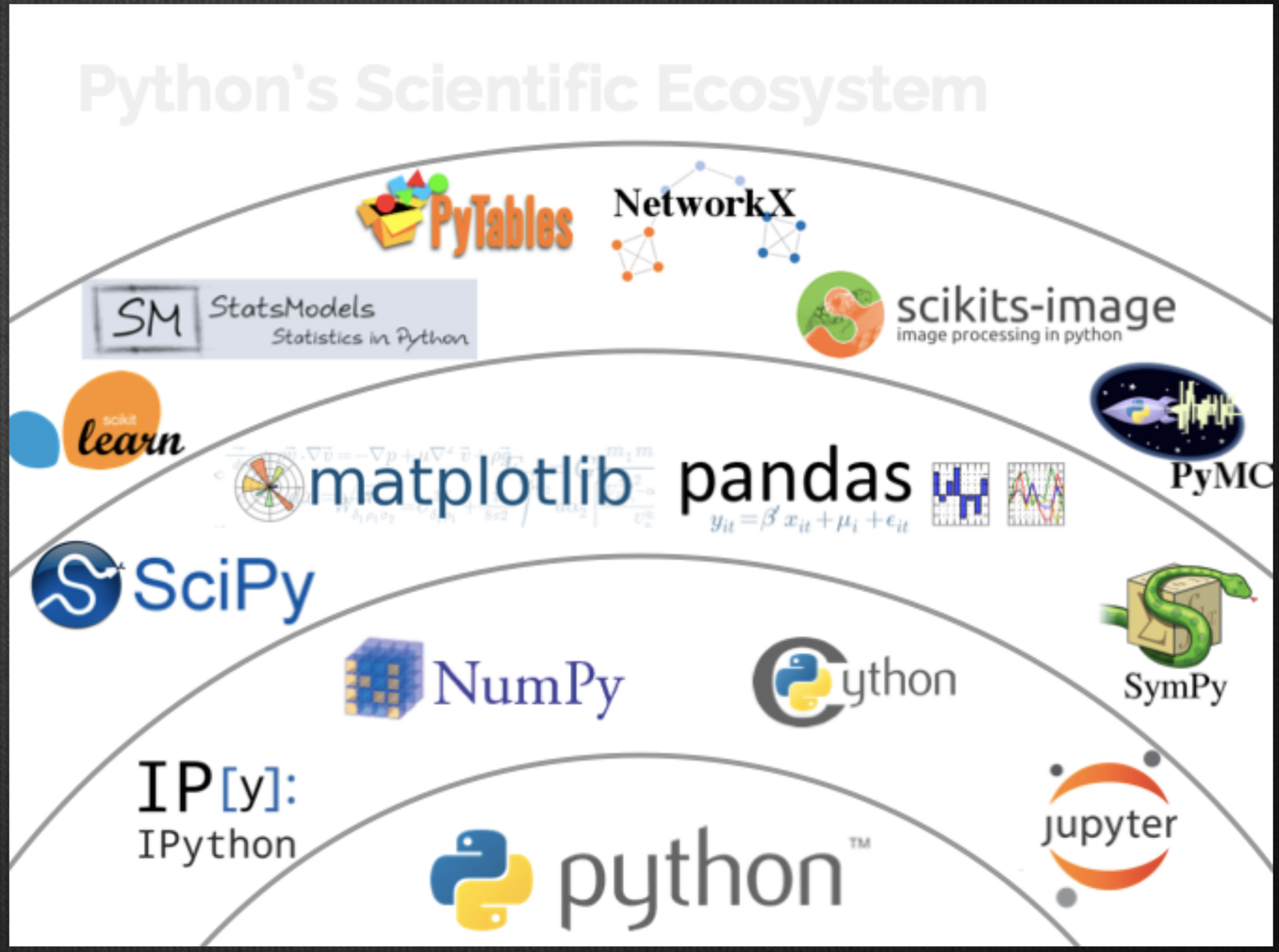
- sort unsort.txt | uniq 消除重复行
- cat text | tr '\t' ' ' 制表符转空格
- cut -f 2,4 filename 截取文件的第2,4列
- paste file1 file2 -d "," 按列拼接两个文件
- 改变文件编码
  - iconv -f GBK -t UTF-8 file1 -o file2

一、Linux 基础

开智学堂

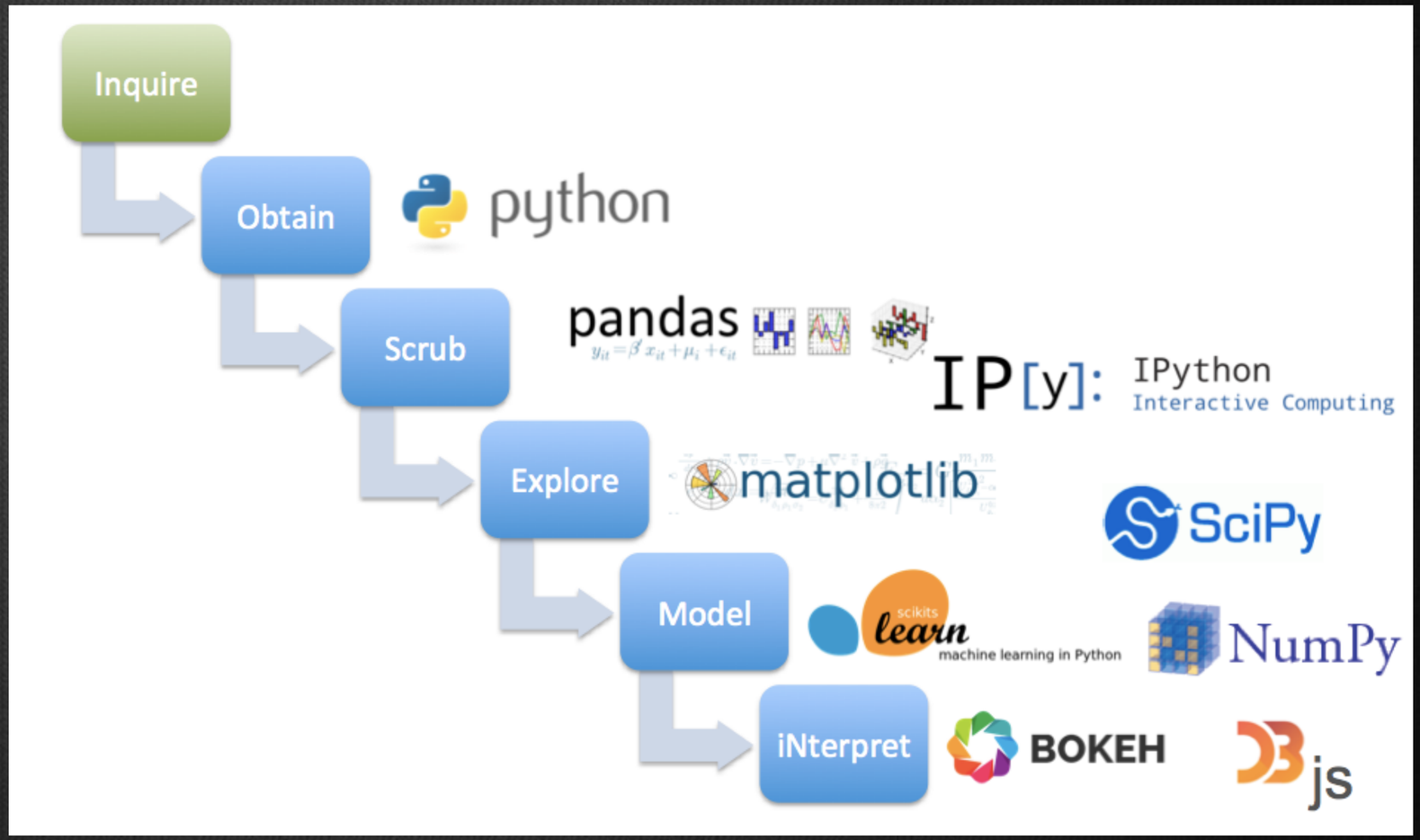
三、Python数据工具箱

Python科学计算库



开智学堂

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
- Beautiful Soup：解析网页数据
- Flask：轻量级的web框架
- sqlite3：轻量级数据库接口

数据相关模块

- 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

运行环境

ipython是一个增强的python shell

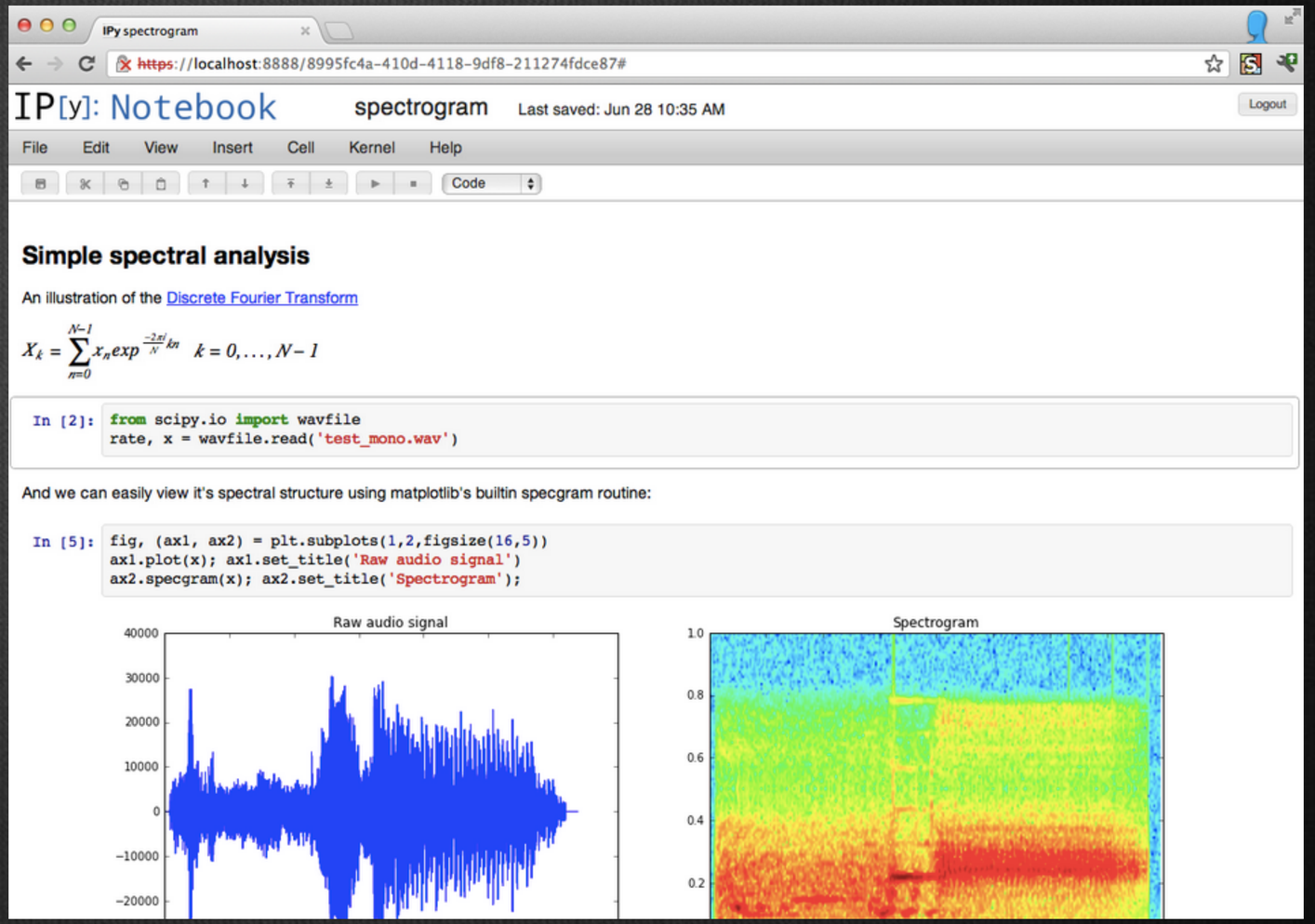
- 提高编写、测试、调度代码的速度
- 提供了IPython Notebook，是一个交互计算平台，也是一个记录计算过程的笔记本

运行环境

- 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

开智学堂

IPython



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

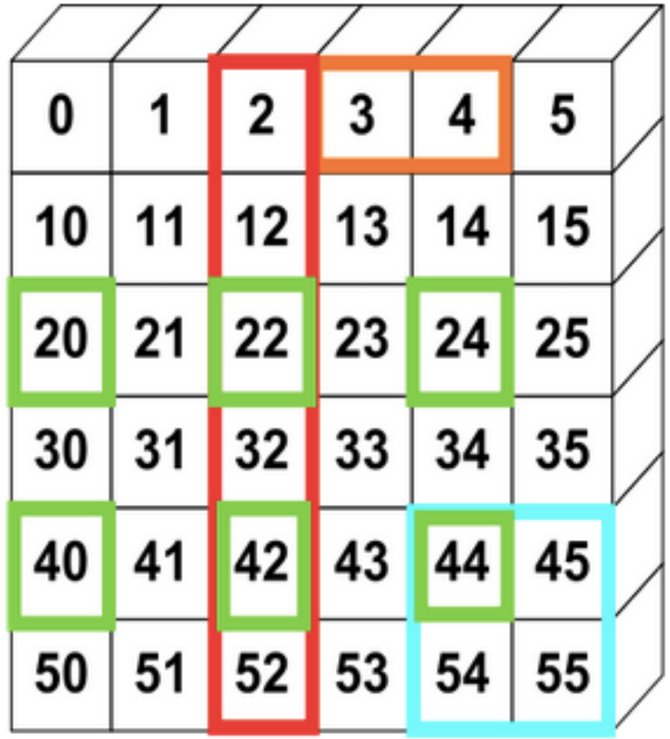
NumPy

```
>>> 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]])
```

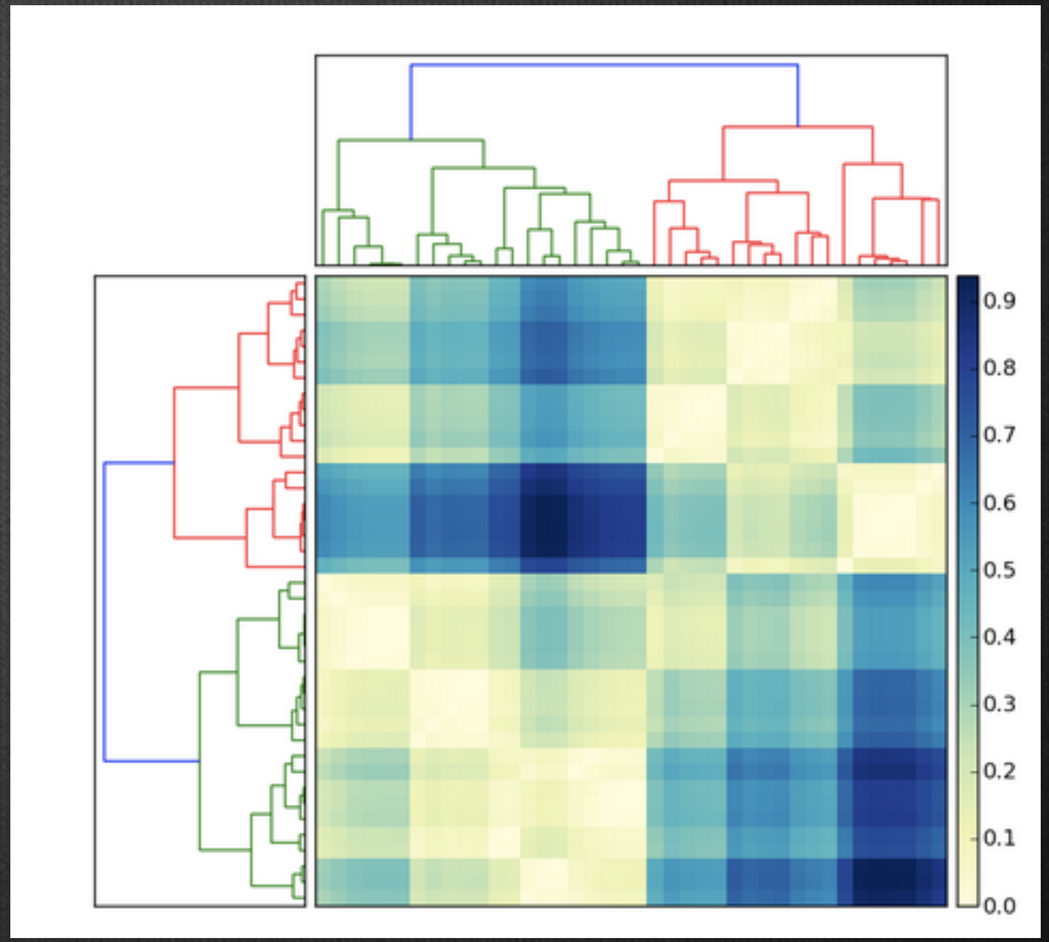


SciPy

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



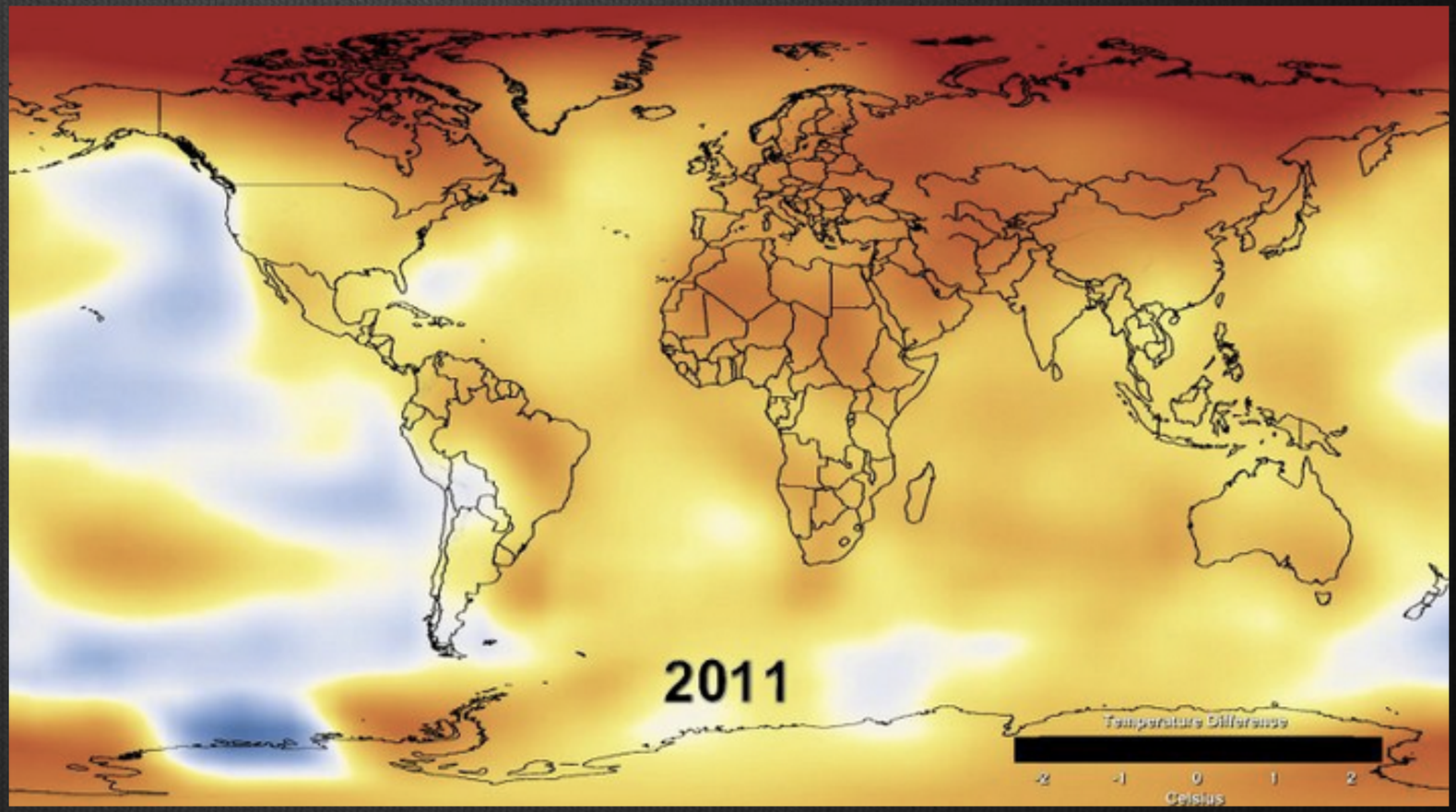
数据可视化

Matplotlib是python下最著名的绘图库

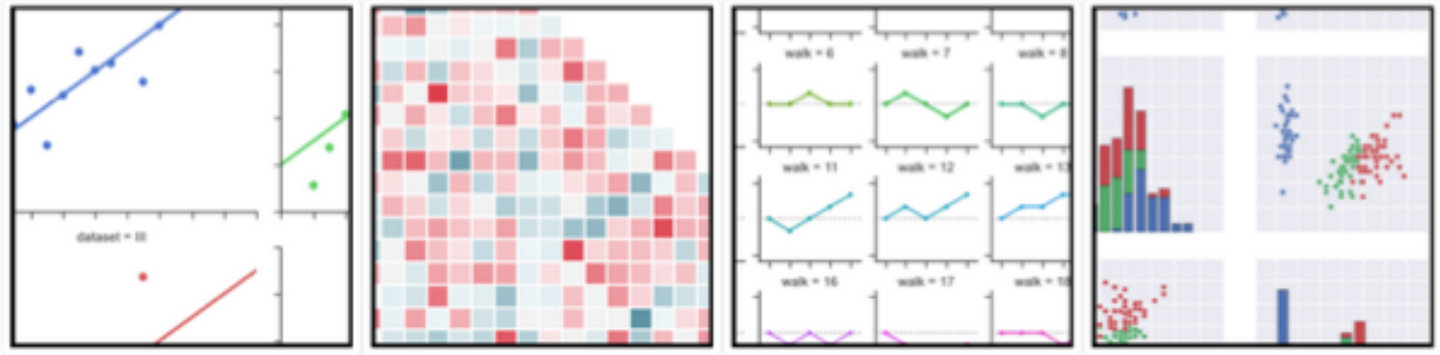
- 提供了一整套和matlab相似的命令API
- 十分适合交互式绘图
- 也可将它作为绘图控件，嵌入GUI应用程序中

开智学堂

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

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

数据分析

Pandas：用于数据处理和分析

- 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

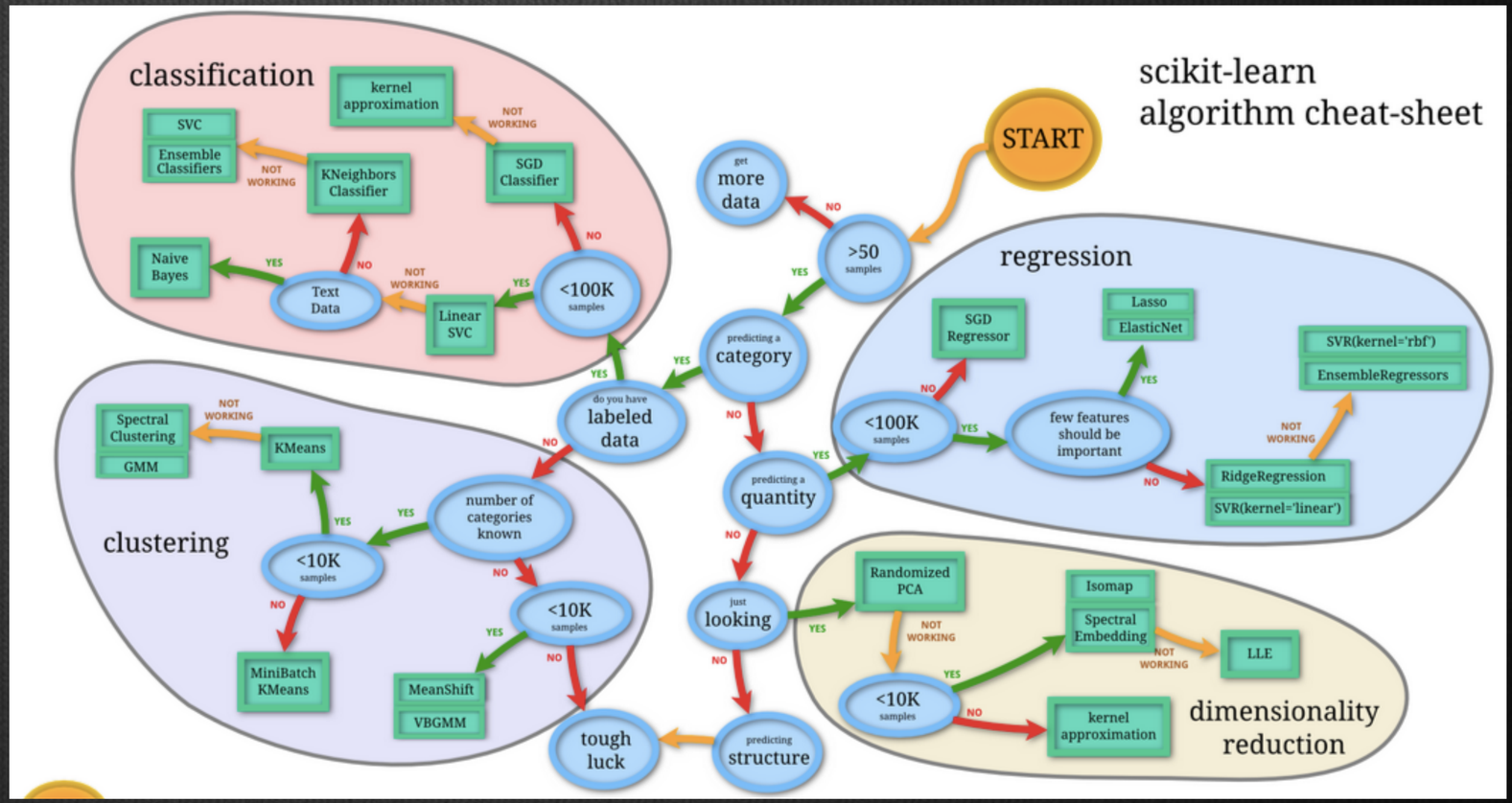
机器学习

Scikit-learn：机器学习库

- 建立在NumPy, SciPy基础上的机器学习库
- 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。

开智学堂

Scikit-learn



三、Python数据工具箱

三、IPython入门

In [9]:
```
display(i)
```

IP[y]: IPython
Interactive Computing

In [3]:
```
from IPython.display import SVG
SVG(filename='python-logo.svg')
```

Out[3]:

开智学堂

第1课录像

开智学堂

四、补充阅读/练习作业

开智学堂

补充阅读材料

- http://ipython.org
- https://damontallen.github.io/IPython-quick-ref-sheets/
- 《利用python进行数据分析》第3章
- 《Data Science at the Command Line》
- 《Numerical Python》第1章
- 《python for scientists》第2章

练习

- 在自己的本机上安装好ipython环境，打开本课程附带的notebook文件自己运行一遍；
- 尝试自己在notebook中录入公式、代码等内容，并将文件上传到github上去，看看是什么效果。

总结

- Linux基础
- Python数据工具箱
- IPython入门
- 补充阅读材料/练习题