

# 为什么选择Python

Python在数据科学里的江湖地位

小胖



# 目录

## ONE Python的江湖地位

数据科学家与Python

## TWO Python简介

版本和学习资料

## THREE Python生态

常用的第三方库

# Python的江湖地位

一个认真的玩笑

Python在数据科学里的  
地位，堪比当年的PHP



- Data scientist is statistics on Python and on Mac

数据科学 = Python + Linux + 统计或机器学习

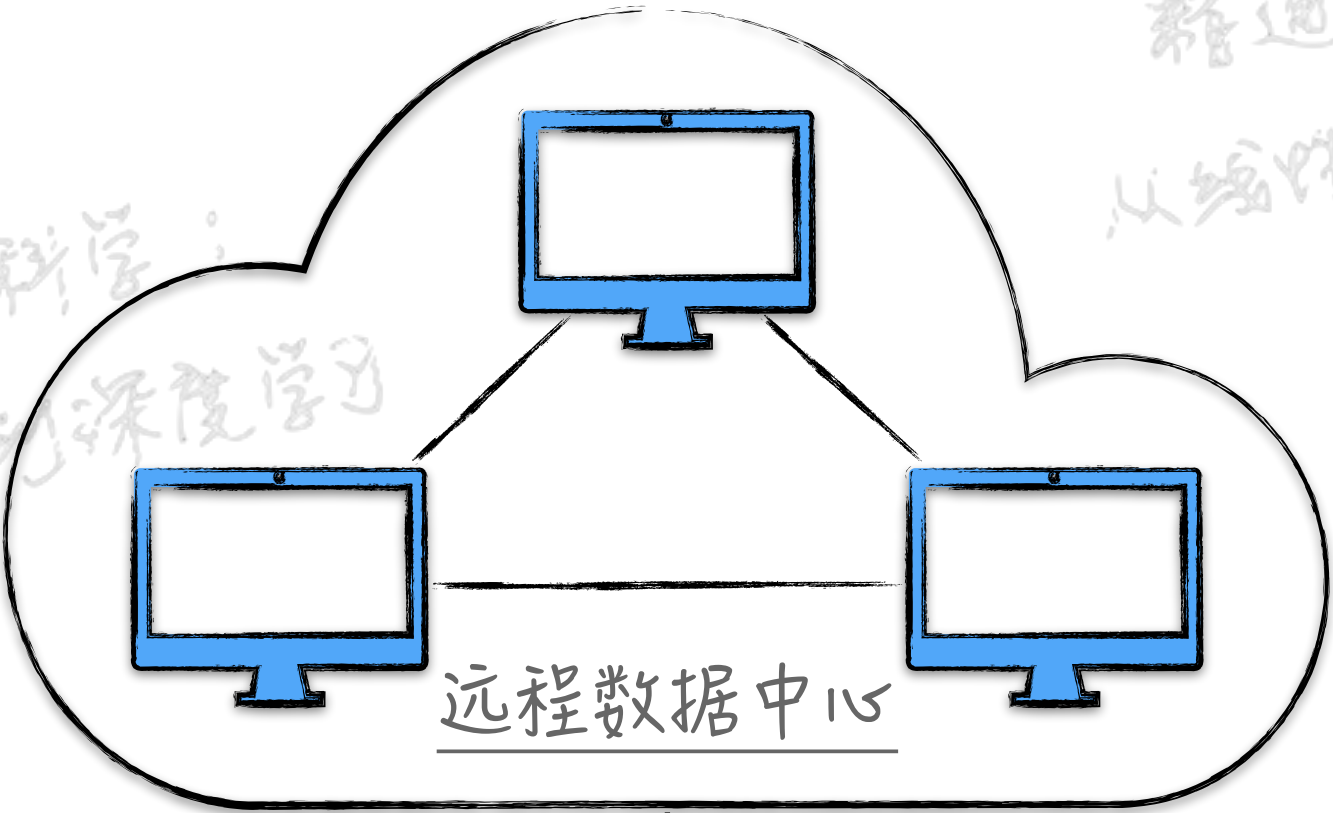
- A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician

数据科学家 = 懂数学的人里代码写得最溜的，写代码的人里数学学得最好的

# Python的江湖地位

数据科学家的工作流程

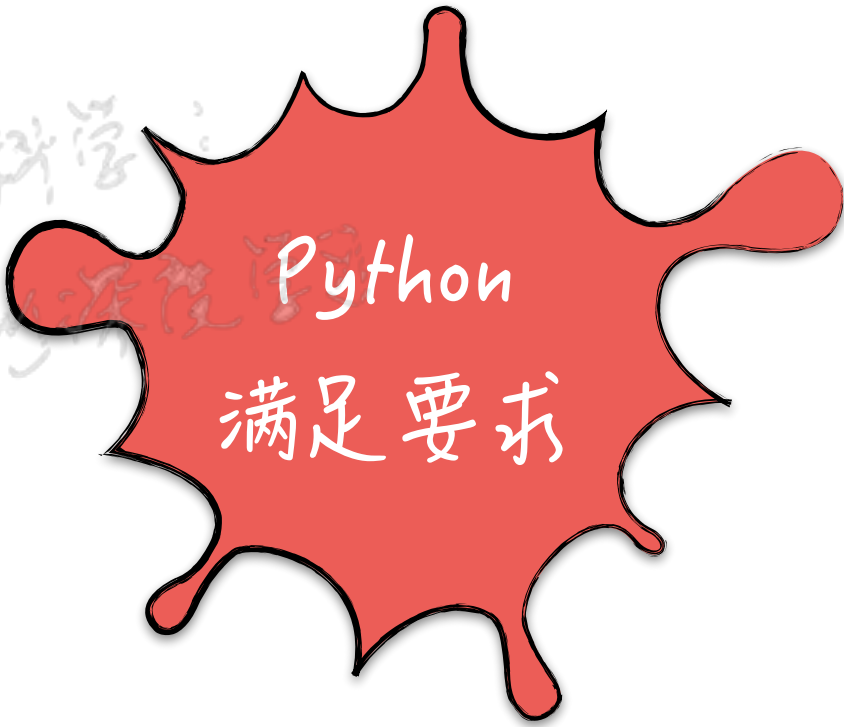
在实际工作中，数据科学家常常面临这样的工作环境



- Linux系统
- 无可视化界面
- 需要编程来操作数据



- 要求编程语言：
- 易学易用
  - 跨操作系统运行

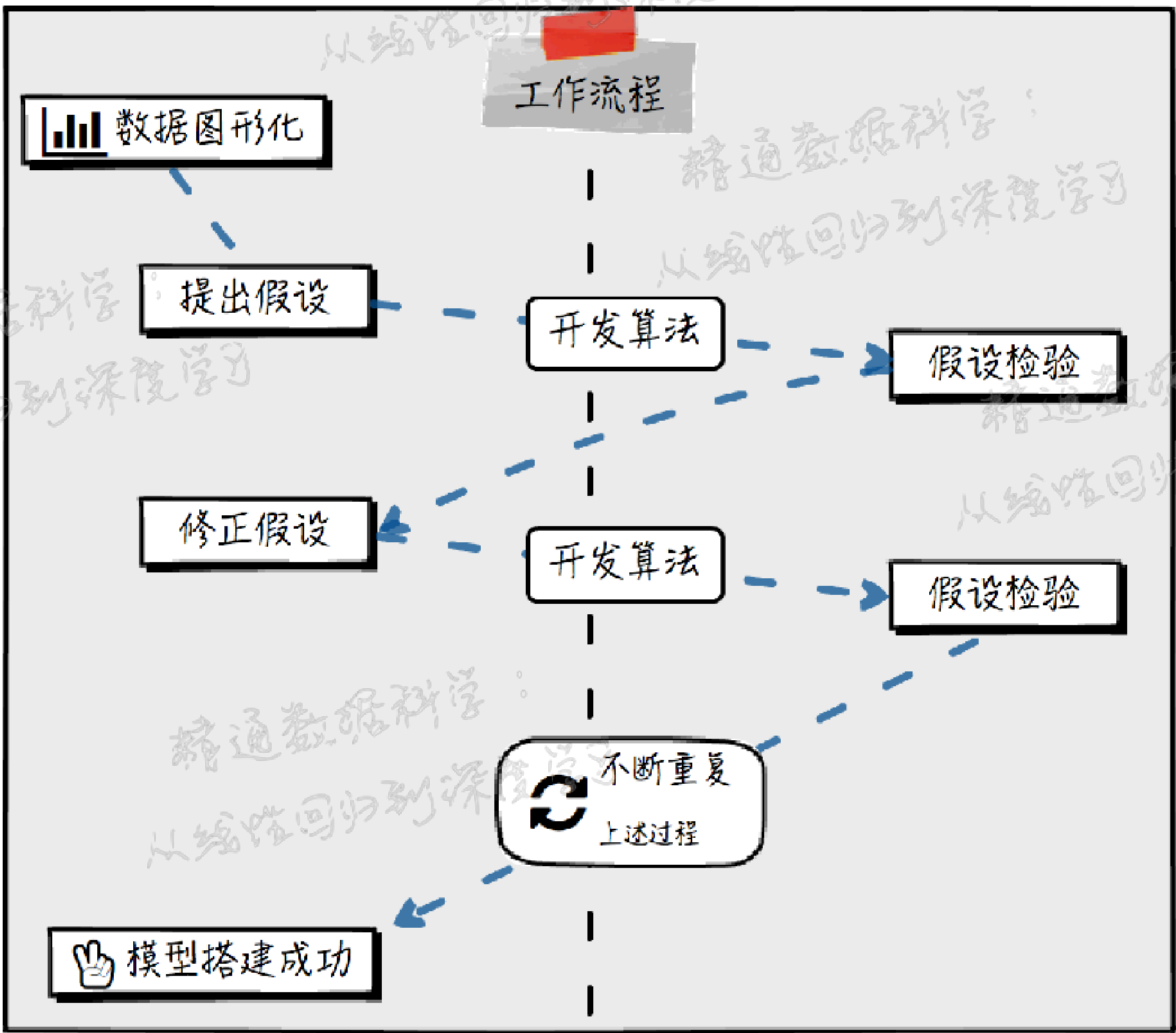




# Python的江湖地位

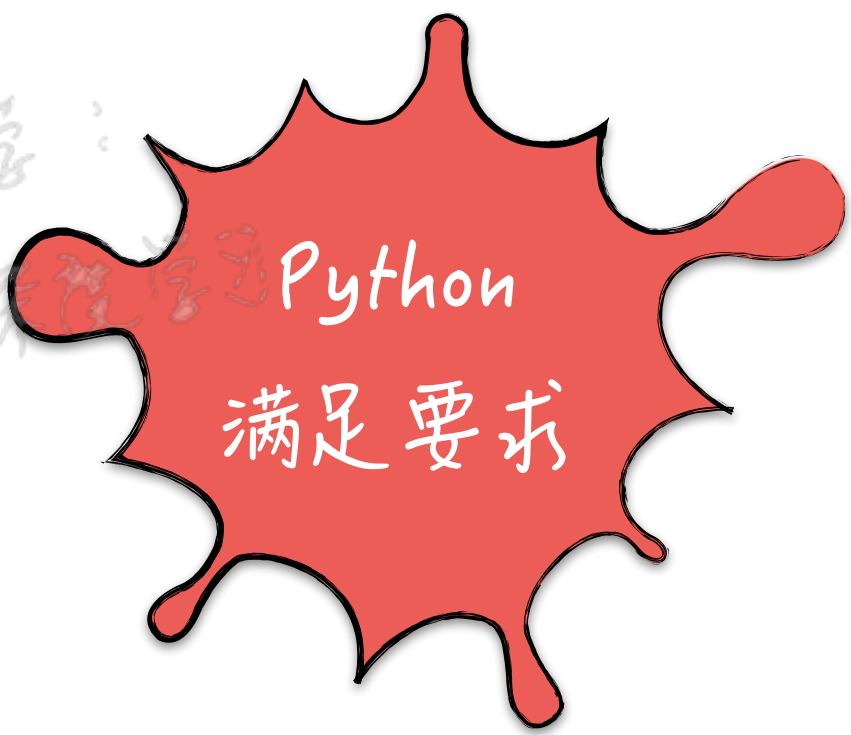
数据科学家的工作流程

类似于乒乓球



要求编程语言:

- 即改即用
- 成熟的算法库



# 目录

## ONE Python的江湖地位

数据科学家与Python

## TWO Python简介

版本和学习资料

## THREE Python生态

常用的第三方库

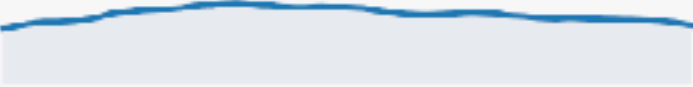
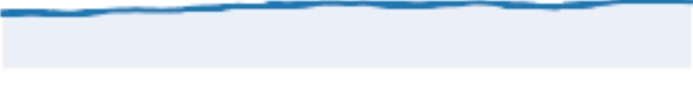

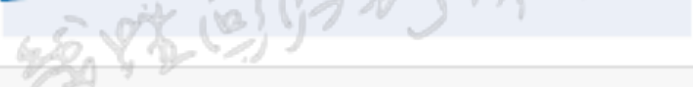
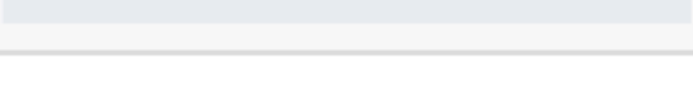
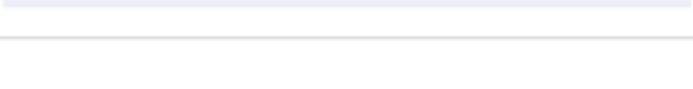
# Python简介

## Python 2 vs. Python 3

Python是一种计算机编程语言



- 应用范围十分广泛，常年位于流行语言TOP10榜单
- 语法简单、学习简单、专为非专业程序员设计
- 强大的生态系统，事半功倍

Rank	Language	Monthly Active Users	Trend
1	JavaScript	22.63%	
2	Python	14.75%	
3	Java	14.01%	
4	C++	8.45%	
5	C	6.03%	
6	PHP	5.85%	



目前Python有两个主要的版本，Python2和Python3

- 两个版本并不兼容，Python3是较新的版本
- 虽然Python3发布已近10年，但在生产环境中，Python2仍大量使用
- 为了面向未来，本课程将使用Python3，但提供配套代码兼容Python2



# Python简介

学习资源

入门教程	廖雪峰的Python教程 <a href="https://www.liaoxuefeng.com/">https://www.liaoxuefeng.com/</a>
	一个非常好的入门教程网站 <a href="https://www.tutorialspoint.com/python/">https://www.tutorialspoint.com/python/</a>
	电子书《Learn Python the hard way》 <a href="https://learnpythonthehardway.org/book/">https://learnpythonthehardway.org/book/</a>
进阶教程	《Dive into python》免费中文版 <a href="http://www.tlsla.com/docs/dive-into-python3/">http://www.tlsla.com/docs/dive-into-python3/</a>
	Python官网 <a href="https://www.python.org/">https://www.python.org/</a>



# 目录

## ONE Python的江湖地位

数据科学家与Python

## TWO Python简介

版本和学习资料

## THREE Python生态

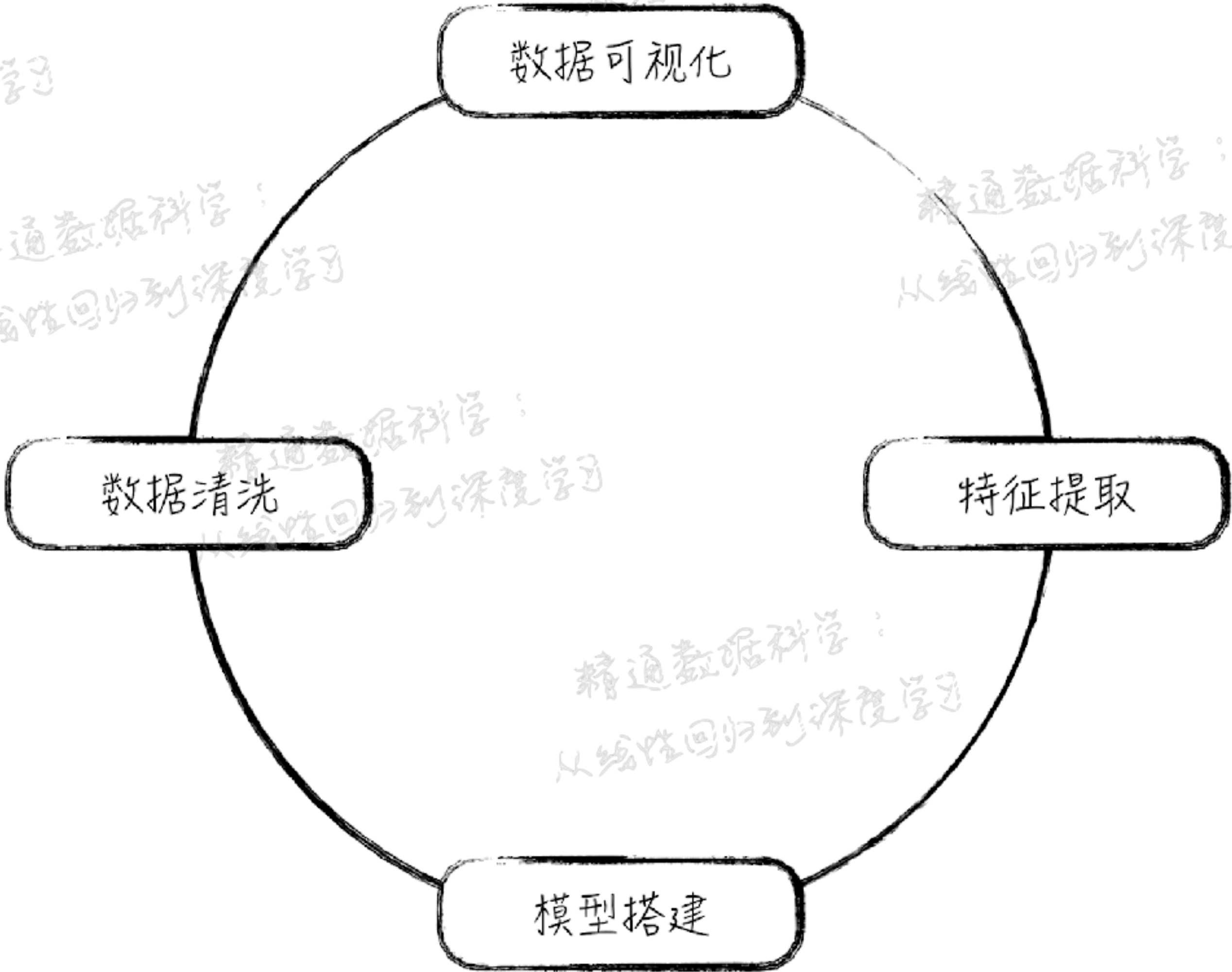
常用的第三方库

# Python生态

数据科学的四大任务

数据科学家们几乎天天都和  
如下的四项任务打交道：

- 数据可视化
- 数据清洗
- 特征提取
- 搭建模型





# Python生态

常用的第三方库

数据预处理	NumPy	科学计算基础库。它提供高效的 $N$ 维数组和向量运算
	SciPy	科学计算库。它依赖于 NumPy，提供高效的数值计算，以及用于函数最优化、数值积分等任务的模块
	pandas	数据结构和数据分析库。包含高级数据结构和类 SQL 语句，让数据处理变得快速、简单
数据可视化	Matplotlib	数据可视化库。它提供大量专业数据图形制作工具
标准模型库	scikit-learn	标准机器学习库。它主要用于分类、回归和聚合等，依赖于 NumPy、SciPy、Matplotlib
	Statsmodels	标准统计模型库。它主要用于假设检验和参数置信区间分析
	Spark ML	分布式机器学习算法库。它可在分布式集群上，如 Hadoop，对大量数据建模。Spark ML 由 Scala 开发，但提供 Python API
	TensorFlow	成熟的深度学习算法库。它提供 GPU 运算模块

精通数据科学；  
从线性回归到深度学习

# THANK YOU