



大数据，成就未来



Python数据分析概述

目录

1	认识数据分析
2	熟悉Python数据分析的工具
3	安装 Python 的 Anaconda 发行版
4	掌握 Jupyter Notebook 常用功能

认识数据分析

学习目标

- 掌握数据分析的概念与流程
- 了解数据分析的应用场景
- 了解Python在数据分析领域的优势
- 了解Python数据分析常用的类库

认识数据分析

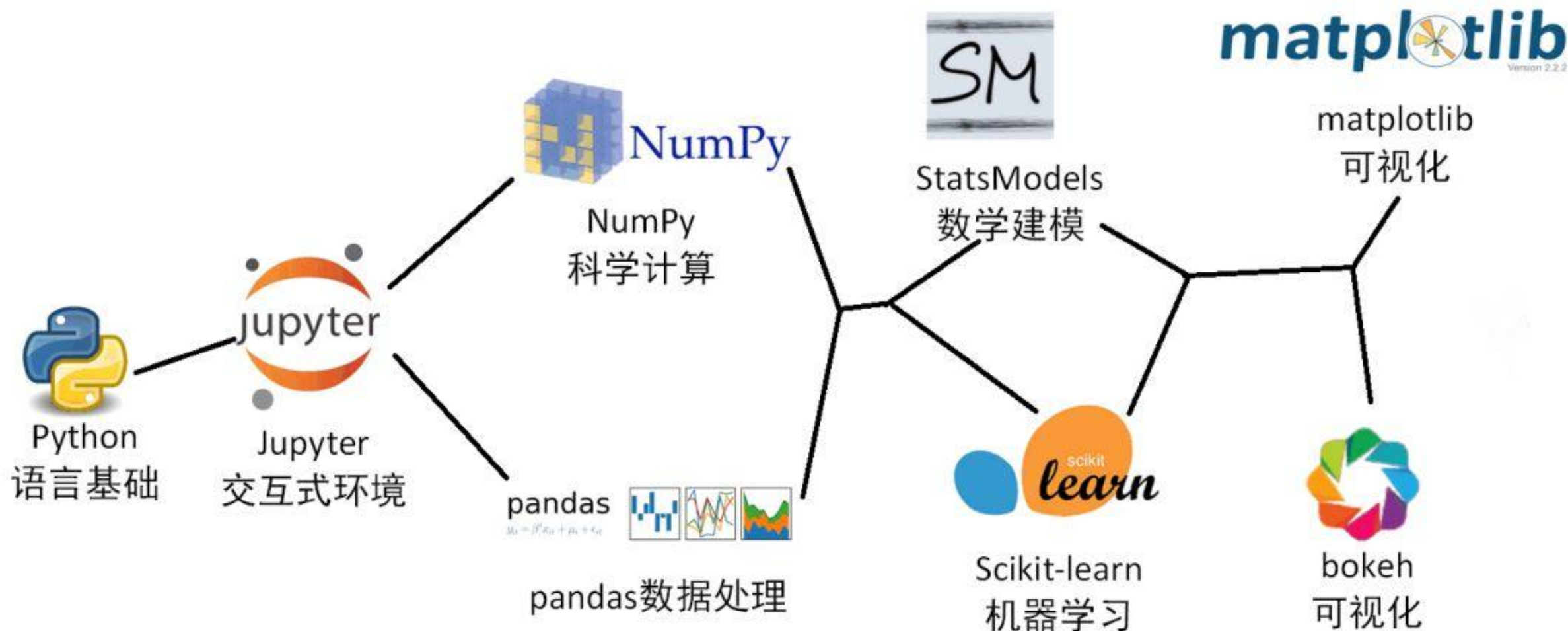
- 网络和信息技术的渗透人类生活的方方面面，数据量呈指数级增长
- 数据的量级超出人力所能处理的范畴
- 如何管理和使用这些数据，成为数据科学领域的一个全新研究课题
- Python有大量的数据科学领域从业者支持并使用，近十年发展迅猛。

认识数据分析

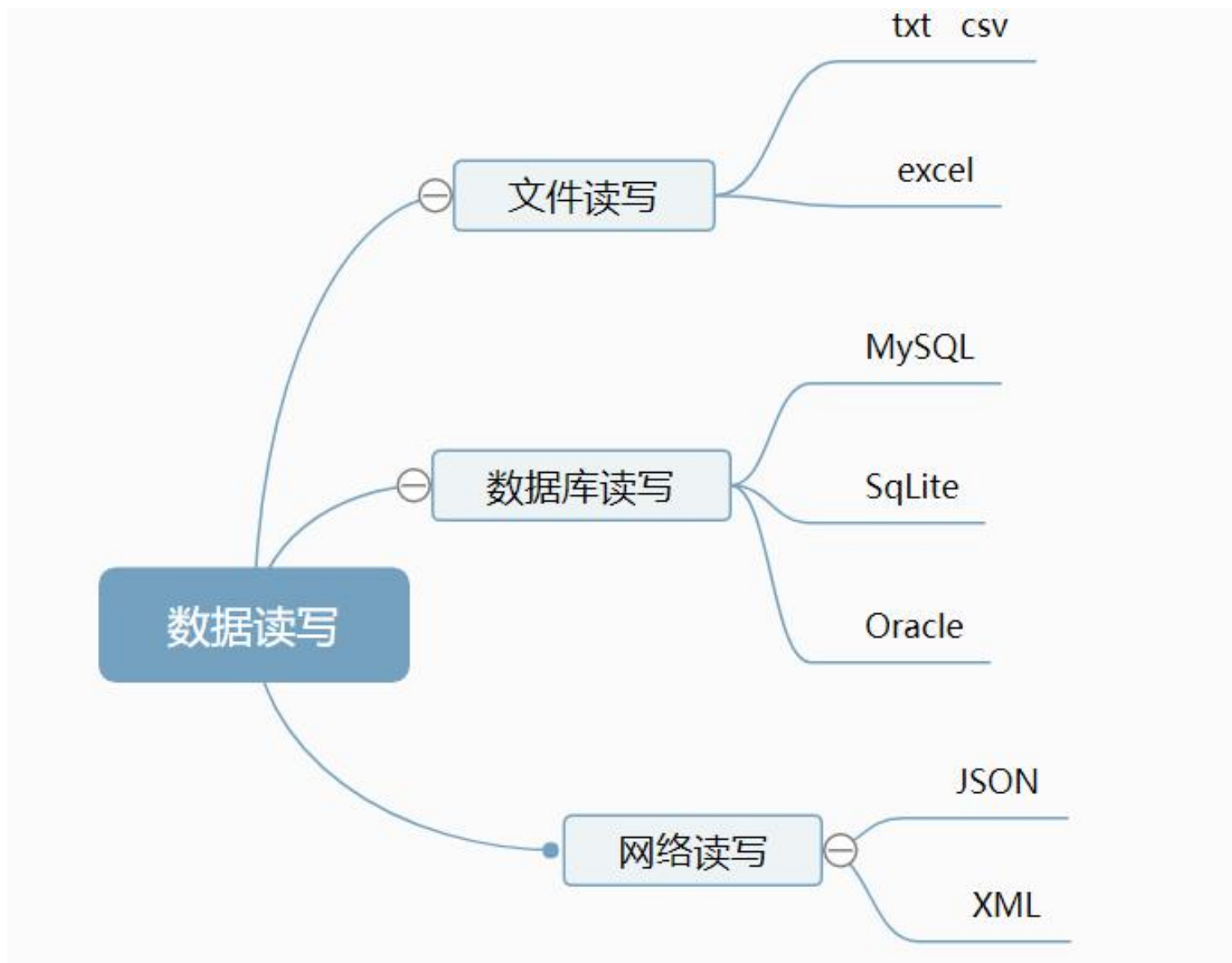
- **数据分析是大数据技术的重要组成部分**
- **数据分析技能被认为是数据科学领域中数据从业人员的必备项**
- **数据分析师应运而生，逐渐热门**
- **掌握数据分析技能，要明确数据分析概念、分析流程和分析方法**

Python数据分析的学习路径

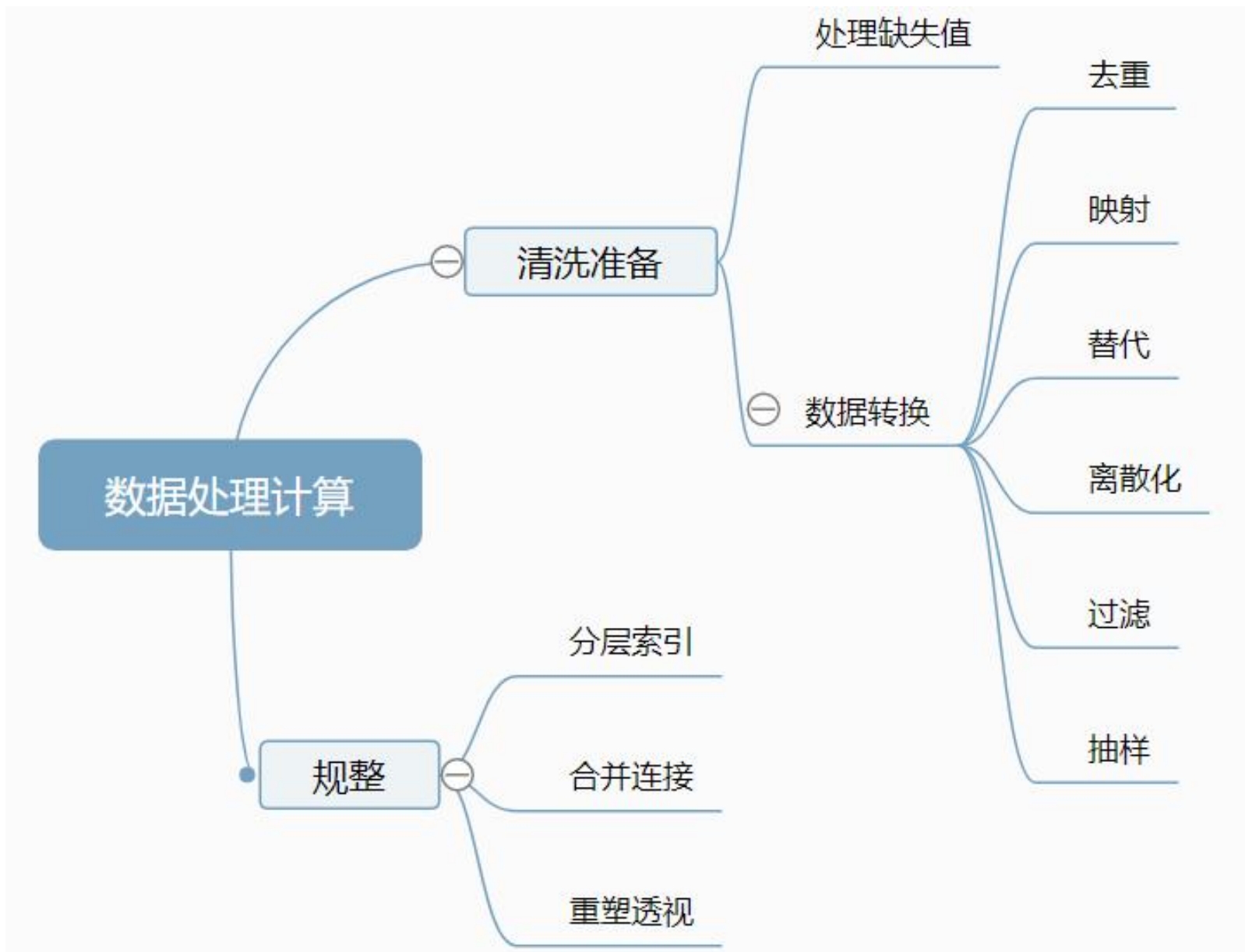
根据每个部分需要用到的工具，Python数据分析的学习路径如下：



利用Python读写数据

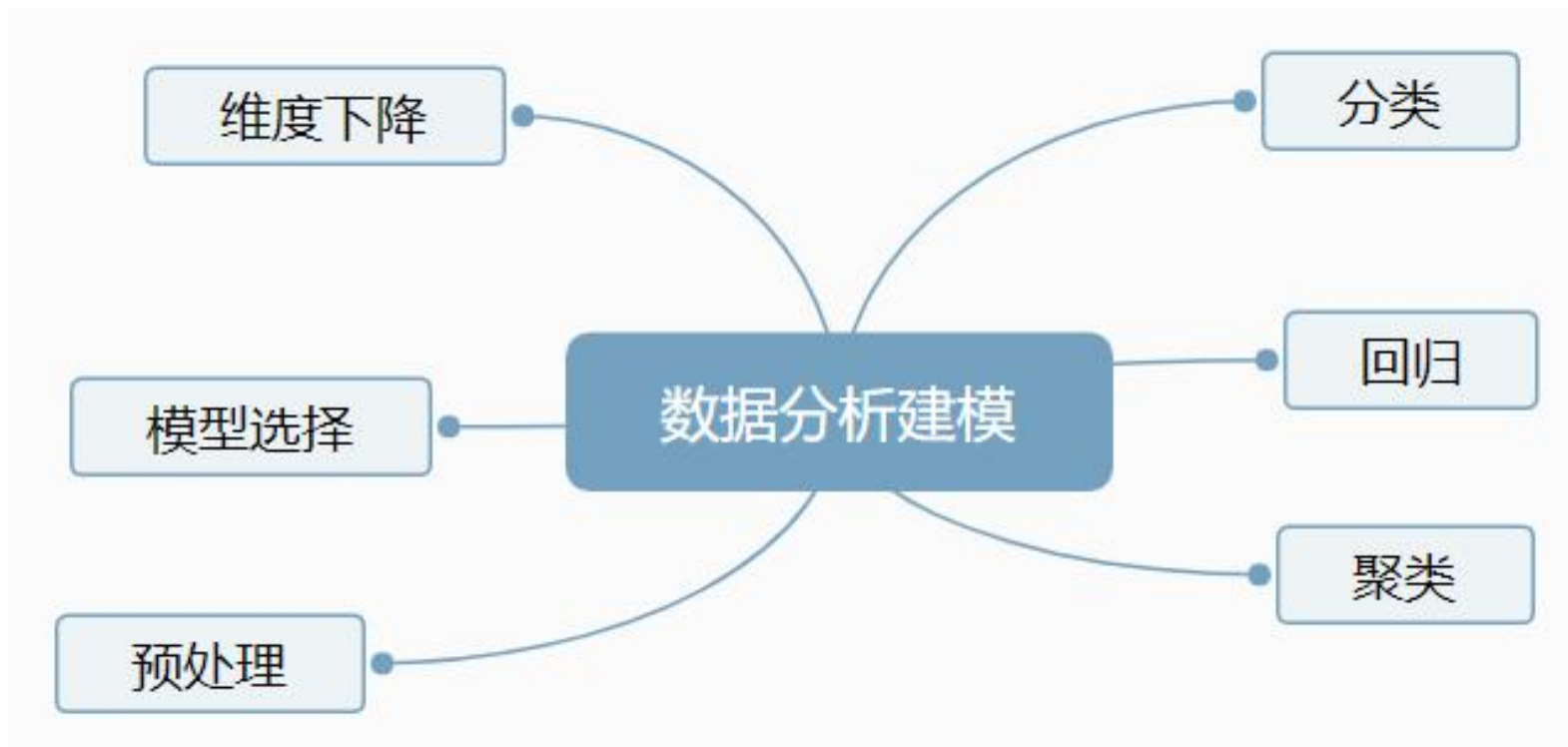


利用Python处理和计算数据



利用Python分析建模

在分析和建模方面，主要用Statsmdels和Scikit-learn两个库

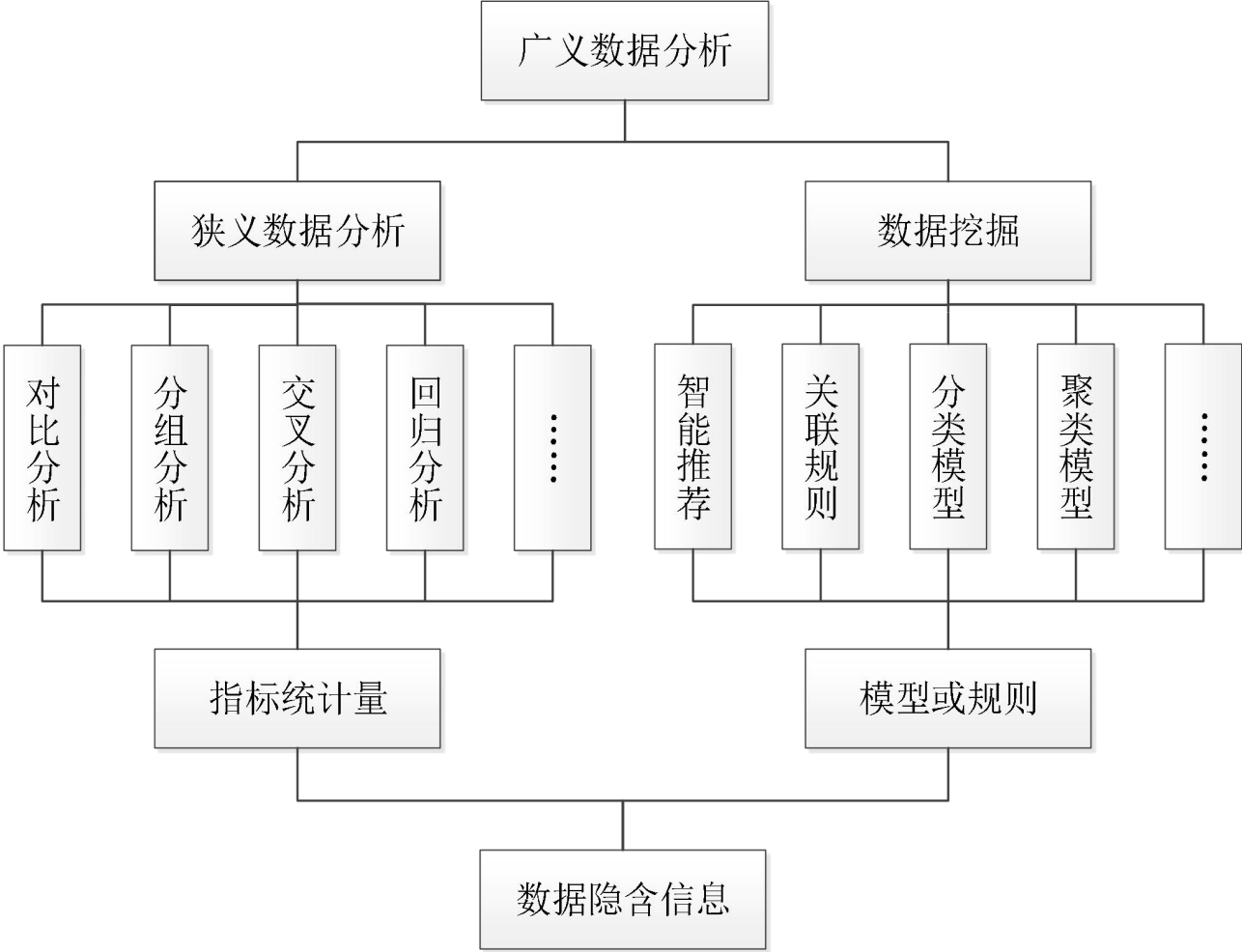


数据分析的概念

广义的数据分析包括狭义数据分析和数据挖掘。

狭义的数据分析是指根据分析目的，采用对比分析、分组分析、交叉分析和回归分析等分析方法，对收集来的数据进行处理与分析，提取有价值的信息，发挥数据的作用，得到一个特征统计量结果的过程。

数据挖掘则是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，通过应用聚类、分类、回归和关联规则等技术，挖掘潜在价值的过程。



数据分析的概念

对比分析法：也叫比较分析法，是把客观事物加以比较，以达到认识事物的本质和规律并做出正确评价的方法。一般通过实际数与基数（标准数据）的对比来提示实际数与基数之间的差异，借以了解经济活动的成绩和问题的一种分析方法。

分组分析：是指通过统计分组的计算和分析，来认识所要分析对象的不同特征，不同性质及相互关系的方法。

交叉分析法：又称立体分析法，是在纵向分析法和横向分析法的基础上，从交叉、立体的角度出发，由浅入深、由低级到高级的一种分析方法。这种方法虽然复杂，但它弥补了“各自为政”分析方法所带来的偏差。

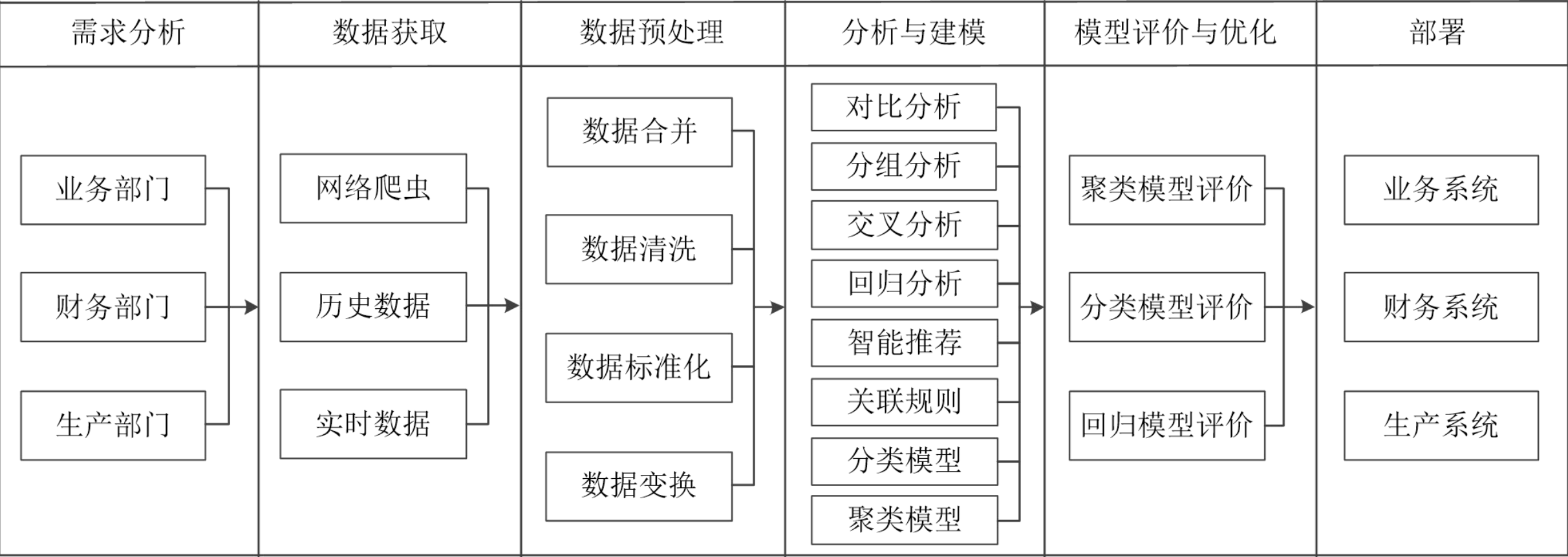
数据分析的概念

回归分析 (regression analysis): 研究一个随机变量Y对另一个(X)或一组(X_1, X_2, \dots, X_k)变量的相依关系的统计分析方法。它是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法。运用十分广泛, 回归分析按照涉及的自变量的多少, 可分为一元回归分析和多元回归分析; 按照自变量和因变量之间的关系类型, 可分为线性回归分析和非线性回归分析。

聚类分析: 指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。聚类是将数据分类到不同的类或者簇这样的一个过程, 所以同一个簇中的对象有很大的相似性, 而不同簇间的对象有很大的相异性。

数据分析的流程

典型的数据分析的流程



数据分析的流程

典型的数据分析的流程

- 需求分析：数据分析中的需求分析也是数据分析环节的第一步和最重要的步骤之一，决定了后续的分析的方向、方法。
- 数据获取：数据是数据分析工作的基础，是指根据需求分析的结果提取，收集数据。
- 数据预处理：数据预处理是指对数据进行数据合并，数据清洗，数据变换和数据标准化，数据变换后使得整体数据变为干净整齐，可以直接用于分析建模这一过程的总称。
- 分析与建模：分析与建模是指通过对比分析、分组分析、交叉分析、回归分析等分析方法和聚类、分类、关联规则、智能推荐等模型与算法发现数据中的有价值信息，并得出结论的过程。
- 模型评价与优化：模型评价是指对已经建立的一个或多个模型，根据其模型的类别，使用不同的指标评价其性能优劣的过程。
- 部署：部署是指将通过了[正式应用数据分析结果与结论应用至实际生产系统的过程](#)。¹⁴

了解数据分析应用场景

1. 客户分析：主要是客户的基本数据信息进行商业行为分析

- 首先界定目标客户，根据客户的需求，目标客户的性质，所处行业的特征以及客户的经济状况等基本信息使用统计分析方法和预测验证法，分析目标客户，提高销售效率。
- 其次了解客户的采购过程，根据客户采购类型、采购性质进行分类分析制定不同的营销策略。
- 最后还可以根据已有的客户特征，进行客户特征分析、客户忠诚分析、客户注意力分析、客户营销分析和客户收益分析。



了解数据分析应用场景

2. 营销分析：囊括产品分析，价格分析，渠道分析，广告与促销分析这四类分析

- **产品分析**主要是竞争产品分析，通过对竞争产品的分析制定自身产品策略。
- **价格分析**又可以分为成本分析和售价分析，成本分析的目的是降低不必要成本，售价分析的目的是制定符合市场的价格。
- **渠道分析**目的是指对产品的销售渠道进行分析，确定最优的渠道配比。
- **广告与促销分析**则能够结合客户分析，实现销量的提升，利润的增加。



了解数据分析应用场景

3. 社交媒体分析

以不同社交媒体渠道生成的内容为基础，实现不同社交媒体的用户分析，访问分析，互动分析等。同时，还能情感和舆情监督提供丰富的资料。

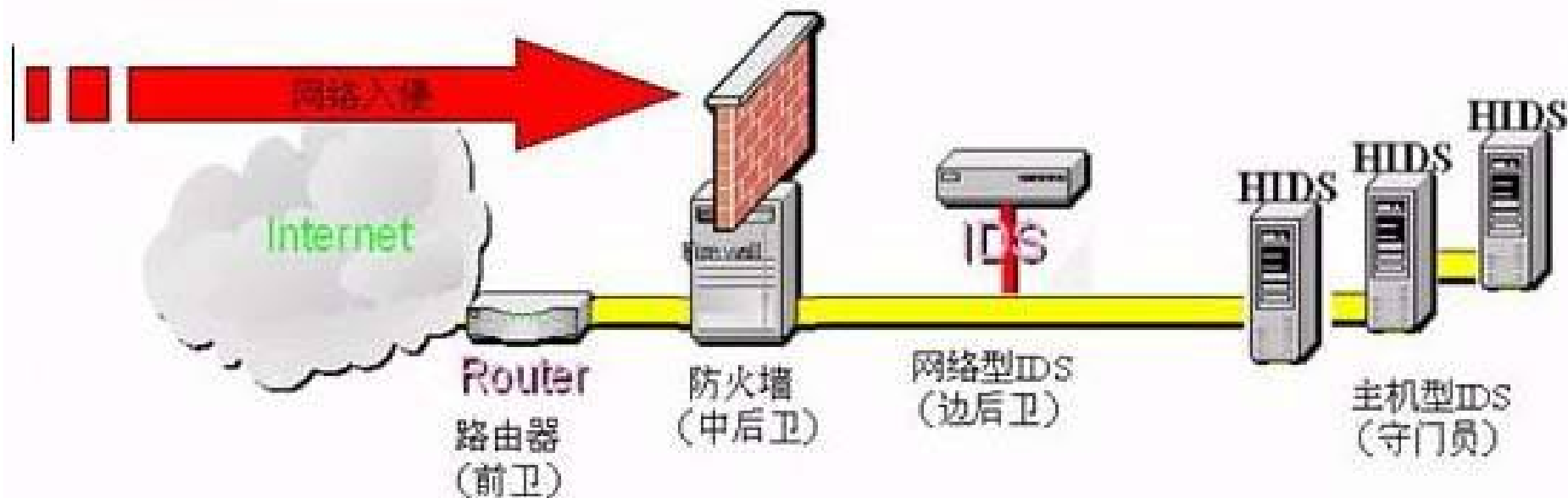
- **用户分析** 主要根据用户注册信息，登录平台的时间点和平时发表的内容等用户数据，分析用户个人画像和行为特征。
- **访问分析** 则是通过用户平时访问的内容，分析用户的兴趣爱好，进而分析潜在的商业价值。
- **互动分析** 根据互相关注对象的行为预测该对象未来的某些行为特征。



了解数据分析应用场景

4. 网络安全

新型的病毒防御系统可使用数据分析技术，建立潜在攻击识别分析模型，监测大量网络活动数据和相应的访问行为，识别可能进行入侵的可疑模式，做到未雨绸缪。



了解数据分析应用场景

5. 设备管理

通过物联网技术能够收集和分析设备上的数据流，包括连续用电、零部件温度、环境湿度和污染物颗粒等无数潜在特征，建立设备管理模型，从而预测设备故障，合理安排预防性的维护，以确保设备正常作业，降低因设备故障带来的安全风险。



了解数据分析应用场景

6. 交通物流分析

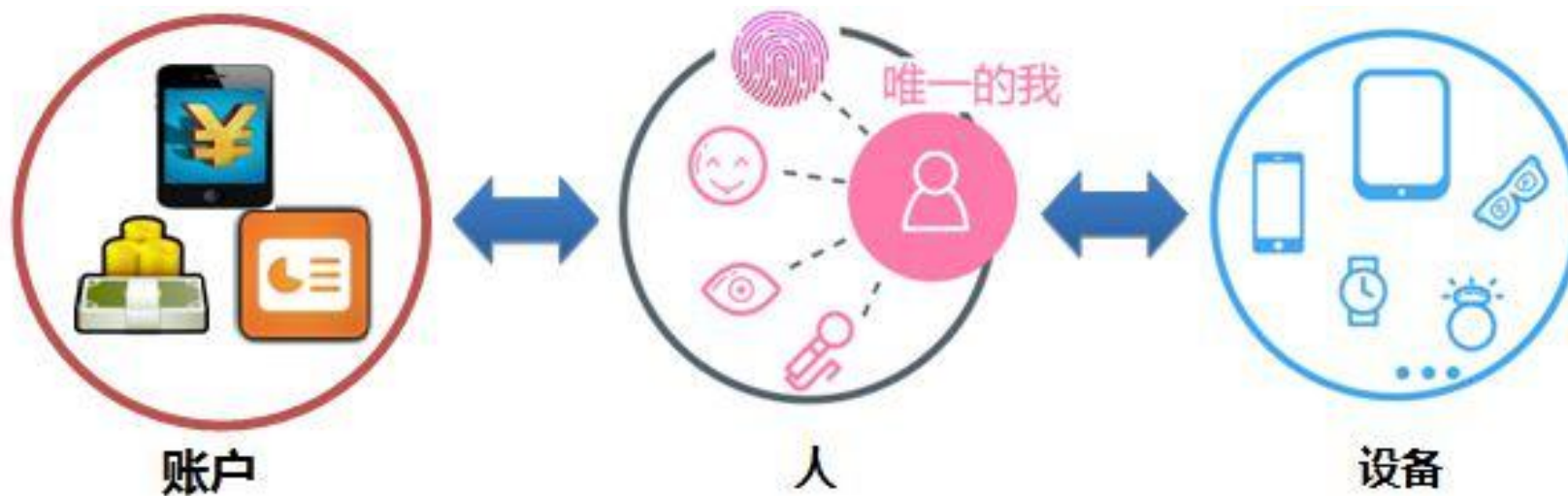
物流是物品从供应地向接收地的实体流动。通过业务系统和GPS定位系统获得数据，对于客户使用数据构建交通状况预测分析模型，有效预测实时路况、物流状况、车流量、客流量和货物吞吐量，进而提前补货，制定库存管理策略。



了解数据分析应用场景

7. 欺诈行为检测

身份信息泄露盗用事件逐年增长，随之而来的是欺诈行为和交易的增多。公安机关，各大金融机构，电信部门可利用用户基本信息，用户交易信息，用户通话短信信息等数据，识别可能发生的潜在欺诈交易，做到提前预防未雨绸缪。



目录

1	认识数据分析
2	熟悉Python数据分析的工具
3	安装 Python 的 Anaconda 发行版
4	掌握 Jupyter Notebook 常用功能

了解数据分析常用工具

目前主流的数据分析语言有R，Python，MATLAB三种程序语言。

	R	Python	MATLAB
语言学习难易程度	入门难度低	入门难度一般	入门难度一般
使用场景	数据分析，数据挖掘，机器学习，数据可视化等。	数据分析，机器学习，矩阵运算，科学数据可视化，数字图像处理，web应用，网络爬虫，系统运维等。	矩阵计算，数值分析，科学数据可视化，机器学习，符号计算，数字图像处理，数字信号处理，仿真模拟等。
第三方支持	拥有大量的Packages，能够调用C，C++，Fortran，Java等其他程序语言。	拥有大量的第三方库，能够简便地调用C，C++，Fortran，Java等其他程序语言。	拥有大量专业的工具箱，在新版本中加入了对C，C++，Java的支持。
流行领域	工业界~学术界	工业界 > 学术界	工业界 ≤ 学术界
软件成本	开源免费	开源免费	商业收费

了解Python数据分析的优势

Python 数据分析主要包含以下 5 个方面优势

- 语法简单精练。对于初学者来说，比起其他编程语言，Python更容易上手。
- 有很强大的库。可以只使用Python这一种语言去构建以数据为中心的应用程序。
- 功能强大。Python是一个混合体，丰富的工具集使它介于传统的脚本语言和系统语言之间。
Python不仅具备所有脚本语言简单和易用的特点，还提供了编译语言所具有的高级软件工程工具。
- 不仅适用于研究和原型构建，同时也适用于构建生产系统。研究人员和工程技术人员使用同一种编程工具，会给企业带来非常显著的组织效益，并降低企业的运营成本。
- Python是一门胶水语言。Python程序能够以多种方式轻易地与其他语言的组件“粘接”在一起。

了解Python数据分析的优势

Python中有关数据分析&挖掘相关的库

- numpy: 数组支撑, 可用来做高维度矩阵计算
- Scipy: 提供矩阵支持, 以及矩阵相关的数值计算和科学计算
- pandas: 强大灵活的数据分析和探索工具
- matplotlib: 强大的数据可视化工具、作图
- scikit-learn: 支持回归、分类、聚类等强大的机器学习与数据挖掘库

了解Python数据分析常用类库

1. NumPy(Numerical Python)—— Python 科学计算的基础包

- 快速高效的多维数组对象 ndarray。
- 对数组执行元素级的计算以及直接对数组执行数学运算的函数。
- 读写硬盘上基于数组的数据集的工具。
- 线性代数运算、傅里叶变换，以及随机数生成的功能。
- 将 C、C++、Fortran 代码集成到 Python 的工具。

了解Python数据分析常用类库

2. SciPy——专门解决科学计算中各种标准问题域的模块的集合

SciPy 主要包含了 8 个模块，不同的子模块有不同的应用，如插值、积分、优化、图像处理和特殊函数等。

- `scipy.integrate` 数值积分例程和微分方程求解器
- `scipy.linalg` 扩展了由 `numpy.linalg` 提供的线性代数例程和矩阵分解功能
- `scipy.optimize` 函数优化器（最小化器）以及根查找算法
- `scipy.signal` 信号处理工具
- `scipy.sparse` 稀疏矩阵和稀疏线性系统求解器
- `scipy.special` SPECFUN（这是一个实现了许多常用数学函数的 Fortran 库）的包装器
- `scipy.stats` 检验连续和离散概率分布、各种统计检验方法，以及更好的描述统计法
- `scipy.weave` 利用内联 C++ 代码加速数组计算的工具

了解Python数据分析常用类库

3. Pandas——数据分析核心库

pandas是Python的数据分析核心库，最初被作为金融数据分析工具而开发出来。

- 为时间序列分析提供了很好的支持。
- 提供了一系列能够快速、便捷地处理结构化数据的数据结构和函数。
- 兼具 Numpy高性能的数组计算功能以及电子表格和关系型数据库（如 SQL）灵活的数据处理功能。
- 复杂精细的索引功能，以便便捷地完成重塑、切片和切块、聚合及选取数据子集等操作。

了解Python数据分析常用类库

4. Matplotlib——绘制数据图表的 Python 库

Matplotlib是最流行的用于绘制数据图表的 Python库，是 Python的2D绘图库

- 非常适合创建出版物上用的图表。
- 操作比较容易，只需几行代码即可生成柱状图、功率谱图、条形图、错误图和散点图等图形。
- 提供了pylab的模块，其中包括了NumPy和pyplot中许多常用的函数，方便用户快速进行计算和绘图。
- Matplotlib与 IPython结合得很好，交互式的数据绘图环境，绘制的图表也是交互式的。

了解Python数据分析常用类库

5. scikit-learn——简单有效的数据挖掘和数据分析工具

- 简单有效，可以供用户在各种环境下重复使用。
- scikit - learn建立在 NumPy、Scipy和 Matplotlib基础之上，封装了一些常用的算法方法。
- 基本模块主要有数据预处理、模型选择、分类、聚类、数据降维和回归 6 个。
- 在数据量不大的情况下，scikit-learn可以解决大部分问题。
- 对算法不精通的用户在执行建模任务时，并不需要自行编写所有的算法，只需要简单的调用scikit - learn库里的模块就可以。