



深圳技术大学
SHENZHEN TECHNOLOGY UNIVERSITY



大数据与互联网学院
COLLEGE OF BIG DATA AND INTERNET

2023-2024秋季课程：数据科学与大数据导论

Introduction to Data Science and Big data

第二章：数据科学基础

Chapter 2: Data Science Fundamentals

曹劲舟 博士 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2023年9月

大纲 Outline

- 什么是数据科学?
- 数据科学的处理流程
- Python简介

大数据与互联网学院
曹劲舟 版权所有

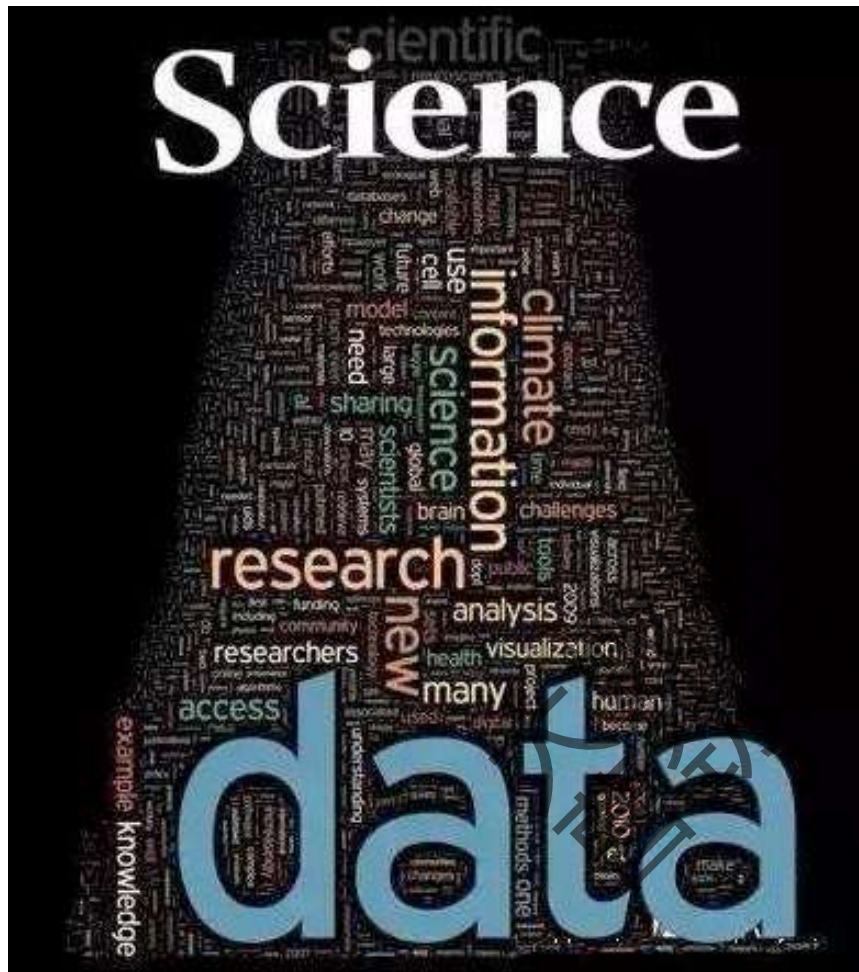
01 Data Science — What IS IT? 什么是数据科学?

大数据与互联网学院
版权所有



What is “Data Science” ?!?

Tools? Big Data?
Machine Learning?



“数据科学”

□2001贝尔实验室科学家威廉.克利夫兰（ William S. Cleveland ）第一次提出：

■ 数据科学应作为由统计学延伸出来的一个独立研究领域，认为统计学中与数据分析有关的技术内容
在下面6个方面扩展后形成了一个新的独立学科——“数据科学”

1. 多学科研究（ Multidisciplinary Investigations ）
2. 数据模型与分析方法（ Models and Methods for Data ）
3. 数据计算（ Computing with Data ）
4. 数据学教程（ Pedagogy ）
5. 工具评估（ Tool Evaluation ）
6. 理论（ Theory ）

Defining Data Science

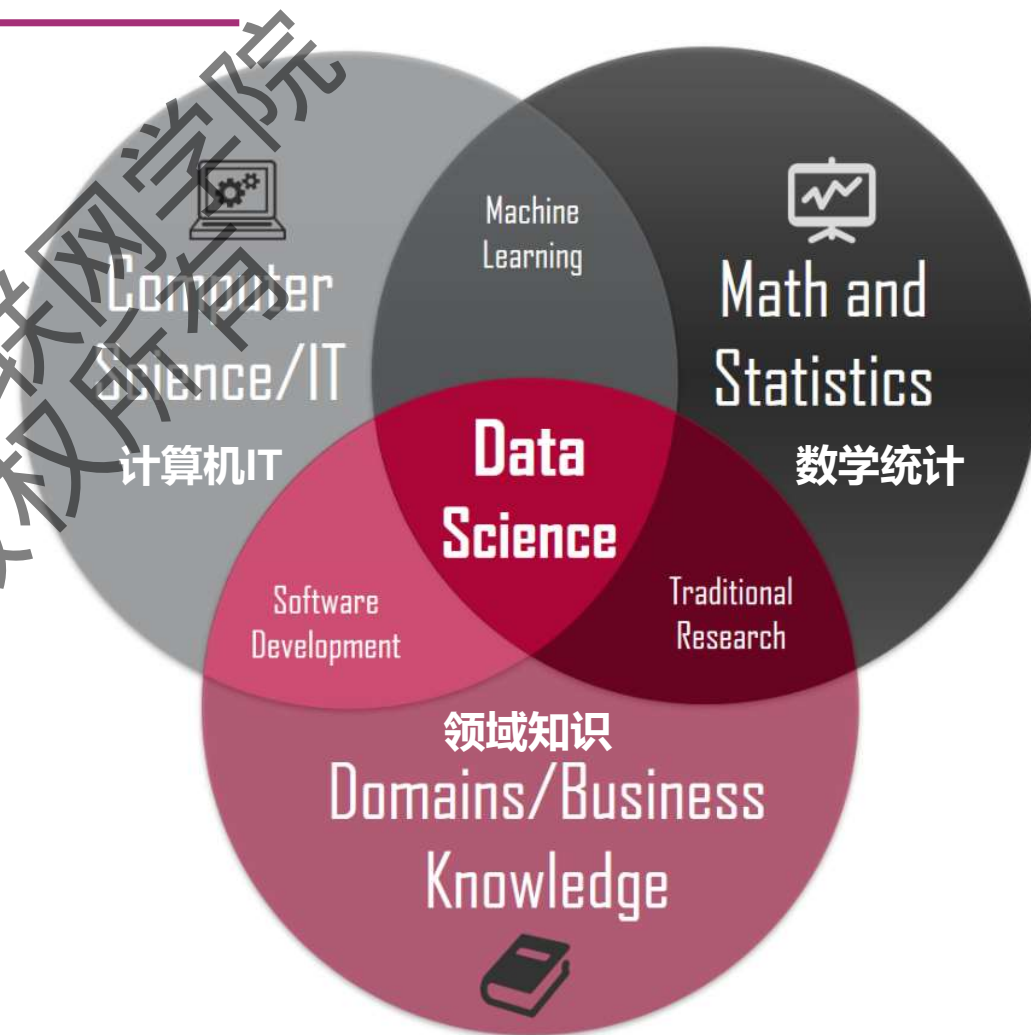
□ 数据科学研究的就是从数据形成知识的过程

- 通过假定设想、分析建模等处理方法，从数据中发现可使用的知识、改进关键决策过程

□ 数据科学的最终产物是数据产品

- 表现为一种发现、预测、服务、推荐、决策、工具或者系统。

Data science is an **inter-disciplinary** field that uses scientific methods, processes, algorithms and systems to extract **knowledge** and **insights** from many structural and unstructured data.



新兴跨领域综合性学科

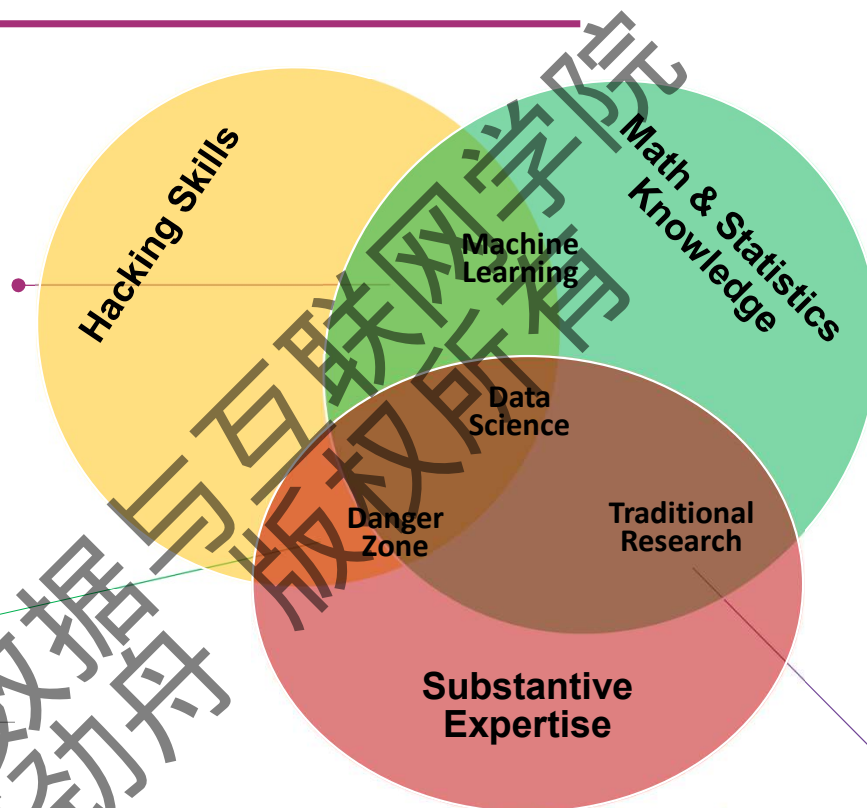
机器学习

侧重计算机编程和数学知识，忽略领域知识的掌握，容易使得数据科学走向机器学习。

危险区

通过计算机解决一些专门领域的问题，但缺乏对问题的数学解释能力，将是非常危险的。

2010年, Drew Conway, 韦恩图



计算机编程(精神)
数学/统计学(理论)
领域知识(实践)

传统研究

重视领域知识、数学理论的学习和掌握，忽视计算机技术，数据难以计算、处理和呈现。

对学科发展的影响

□对数学、统计学

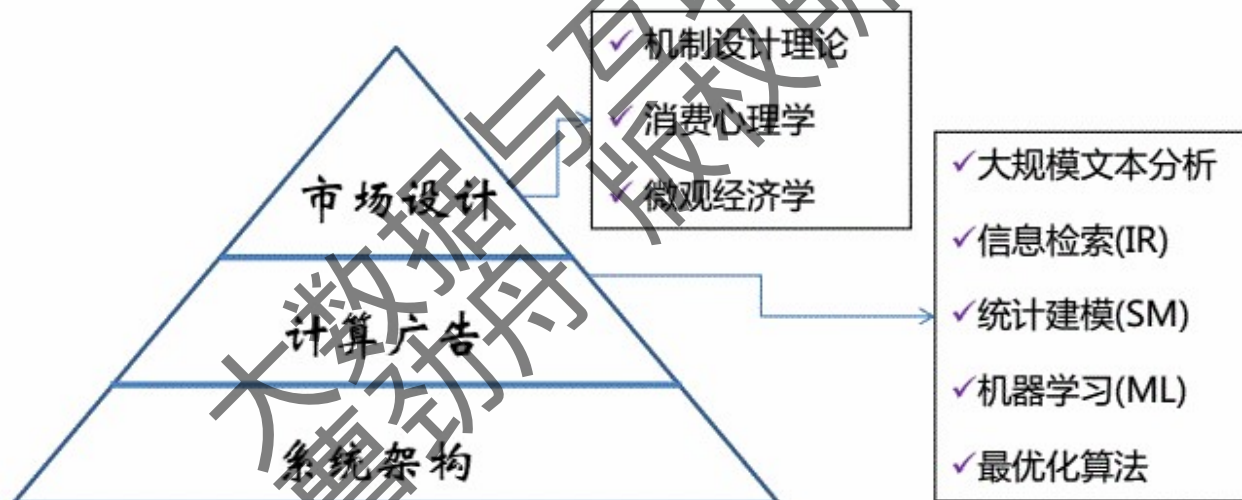
- 数据科学在数学和实际应用间建立一个直接的桥梁
- 拓展统计学研究范围

□对计算机科学

- 复杂性理论拓展
- 大数据数据管理与实时处理，对探索高效的大规模分布式数据存储、查询与处理带来了需求和挑战
- 以机器学习特别是深度学习为代表的数据分析技术在应对规模化数据集的训练和检验，已经催生了软硬件一体的跨层设计和面向数据处理的高性能、低功耗定制芯片、定制服务器设计
- 此外，数据科学与其他基础和应用学科的交叉，将使数据在计算机科学中的地位进一步加强

对学科发展的影响

□促进传统学科的“数据化”，催生新的交叉学科



数据科学基本内涵：两个方面

用数据的方法研究科学

对海量数据挖掘、计算，发现新模式、新知识和新规律，研究不同的学科领域，包括生物信息学、天体信息学、数字地球等领域

科学研究第四范式

从实验、理论、计算到数据科学
(Data-intensive Scientific Discovery)

数据驱动发现

1618年，开普勒第三定律的发现

行星	周期(年)	平均距离	周期 ² /距离 ³
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165.0	30.06	1.00

用科学的方法研究数据

方法多，如统计学、机器学习、数据挖掘、数据库、可视化等。

数据全生命周期

数据采集、清理、存储和分析、可视化、有效治理等。

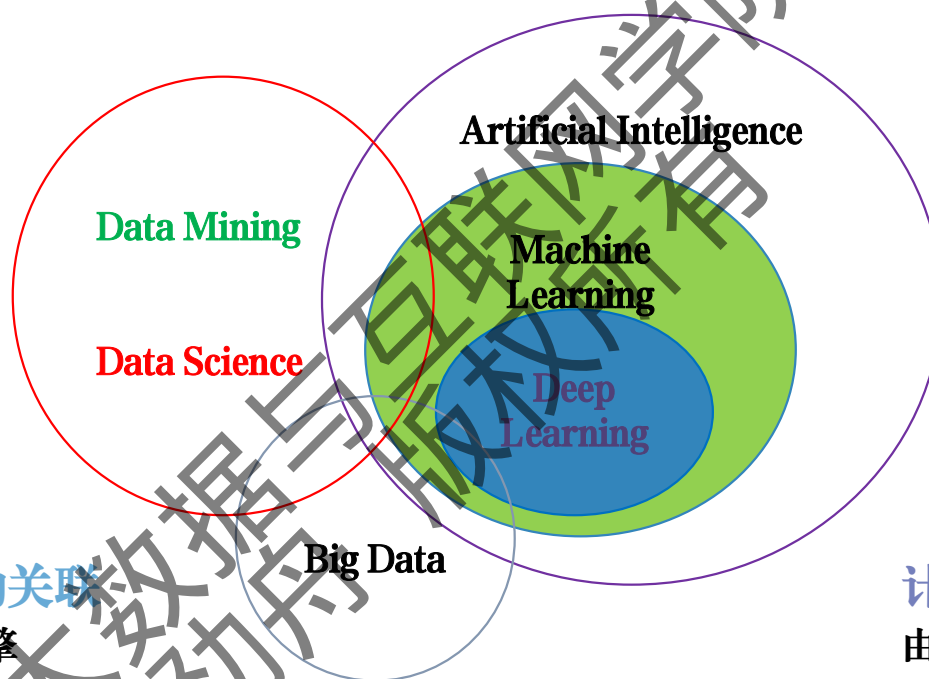
数据离不开科学

2020年，新冠疫情的统计分析、溯源、预测

与其他概念的关联

目标与核心不变

与Drew Conway韦恩图相比



与大数据的关联

在平台技术、模型方法上既有联系又有区别

与机器学习、人工智能的关联

机器学习：数据挖掘算法的引擎

人工智能：基于数据预测的动力来源

计算vs 数据为中心

由学科领域的属性、目标决定
不存在互斥关系，可良性迁移

新诠释：与先进理论和技术相结合

---KDnuggets公司总裁、数据科学家 Gregory Piatetsky-Shapiro

数据分析与数据挖掘

分析

描述和探索性分析，评估现状和修正不足

实际的业务知识

统计学、数据库、Excel、可视化等

需结合业务知识解读统计结果

定义

侧重点

技能

结果

挖掘

技术性的“采矿”过程，发现未知的模式和规律

挖掘技术的落地，完成“采矿”过程

过硬的数学功底和编程技术

模型或规则

02 Doing Data SciENCE

数据科学处理流程

大数据与互联网学院
数据科学处理流程



数据科学的一般过程

数据工程

数据加工、计算、存储和管理、初步分析



原始数据

商业数据、互联网数据、传感器数据等



机器学习

根据历史数据或模拟数据，发现模式并对数据判别和预测



数据操控

模型部署、高级数据分析、可视化、决策



数据科学的一般过程

1

数据收集

系统日志、网页、电子文档、社交媒体

3

数据计算

计算模式趋向云计算、智能计算
google、MapReduce、Spark、YARN

5

初步分析

对象类型：定性分析、定量分析
统计分析：描述性、推论性

7

模型部署

机器学习模型的合理选择和优化

9

数据可视化

数据以图形方式呈现，发现潜在价值和模式，更好理解数据

2

数据加工

数据识别、清洗、变换、集成、脱敏、规约、标注

4

数据存储和管理

关系数据库、格式文件
NoSQL、NewSQL、关系云

6

机器学习

针对复杂问题、海量数据
模型训练，判别和预测

8

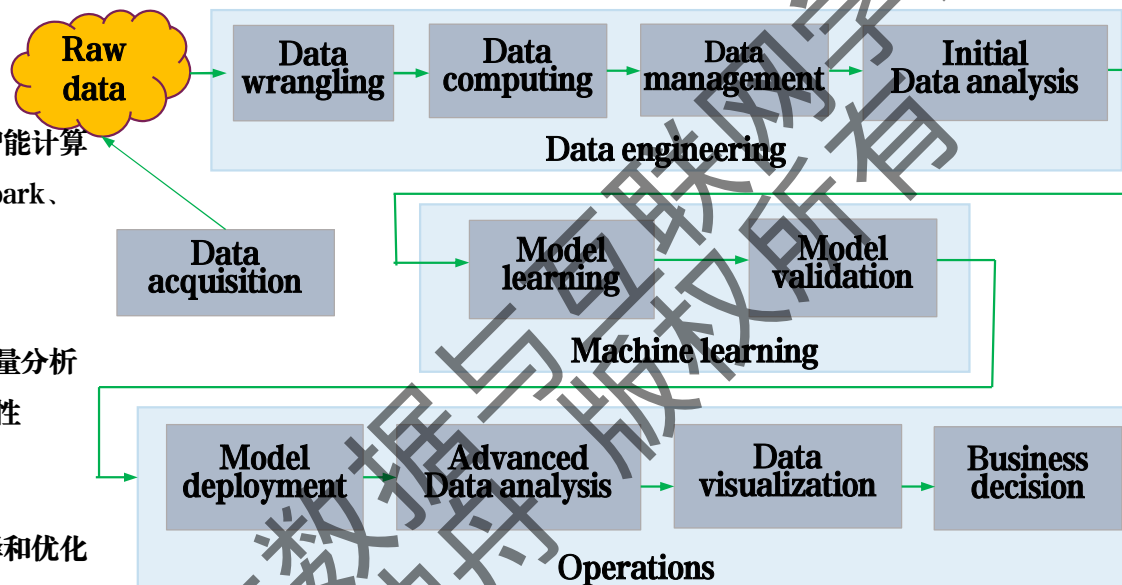
高级分析

分析智能化、自动化
模式识别、理解过程不断迭代

10

决策支持

趋向数据为中心、扁平化范式
数据驱动的决策



数据科学关键技术

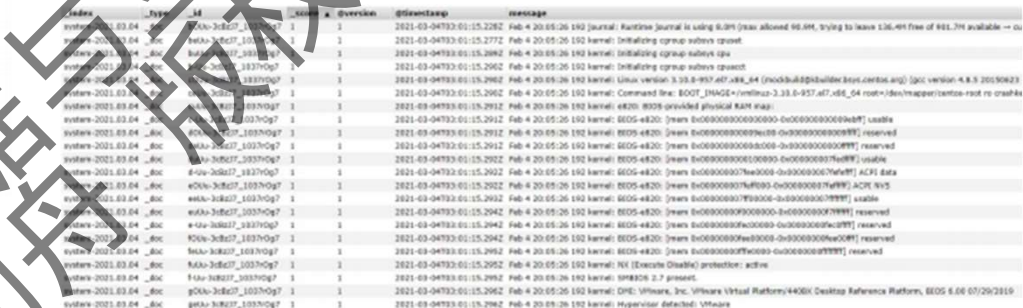
- 数据采集
- 数据预处理
- 数据存储与管理
- 数据分析

大数据与互联网学院
曹劲舟 版权所有

• 人工采集

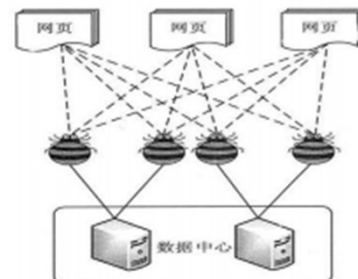


• 传感器采集



- 系统日志

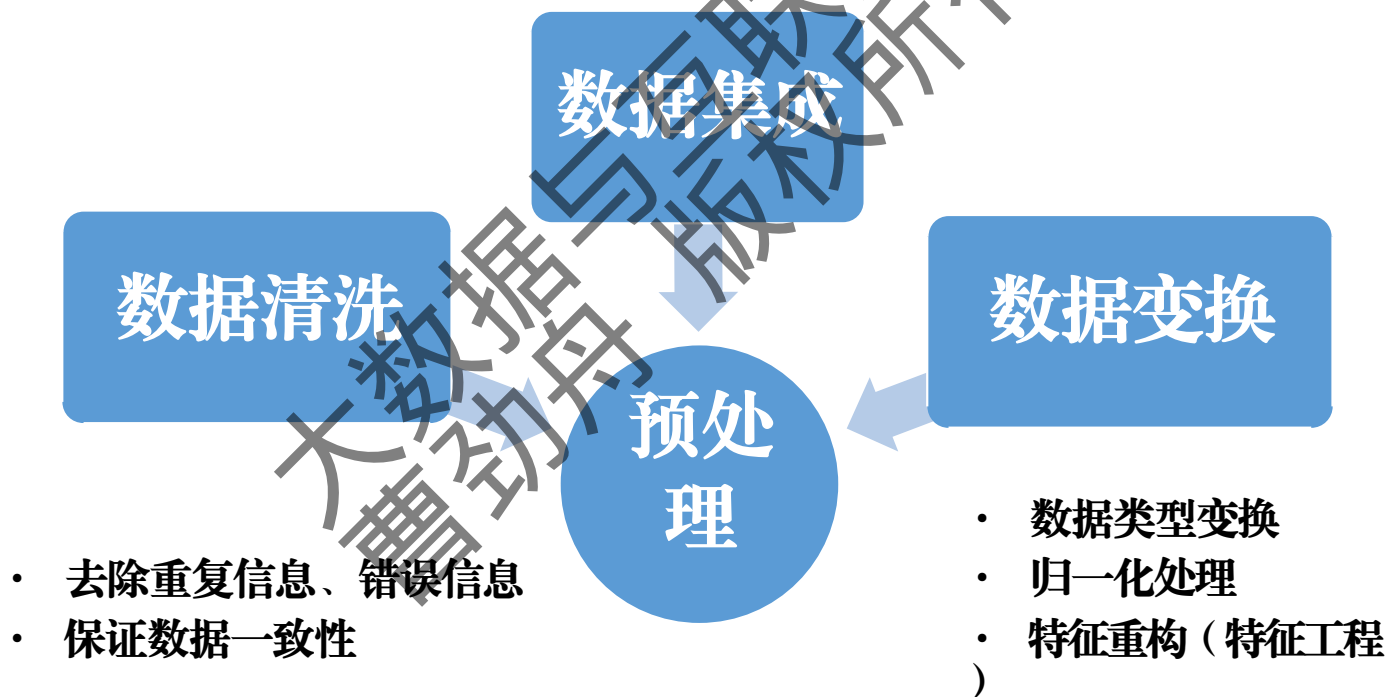
- 网络爬虫



数据预处理

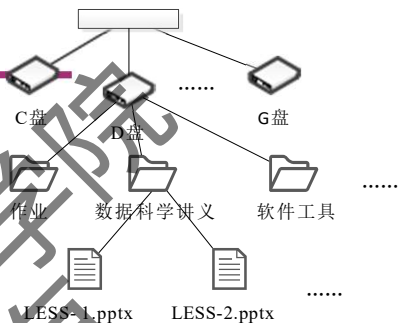
□综合多种采集数据，提高数据质量

- 多种数据源数据
- 集成工具，如ETL等



数据存储与管理

- 文件系统 – 树



- 数据库 – 表

学号	350101	410501	450206	310103	360104	450301	350301
2012011	70	85	77	90	82	84	89
2012012	60	64	80	75	80	92	90
2012013	90	93	88	87	86	90	91
2012014	80	82	91	88	83	86	80
2012015	88	72	78	90	91	73	80

学号	姓名	性别	籍贯
2012011	王微	男	山东青岛
2012012	肖良英	女	上海
2012013	方绮雯	女	湖北荆州
2012014	刘旭阳	男	山西太原
2012015	钱易铭	男	广东茂名

编号	课程名称	学时	学分	学院
350101	Math	64	4	理学院
410501	English	32	2	外语学院
310103	Chinese	32	2	文学院
350301	Physics	48	3	理学院
360104	Art	32	2	文学院
450206	Python	64	3	计算机学院
450301	Database	48	3	计算机学院

数据分析技术

□统计分析

- 描述统计、假设检验、置信度分析、相关分析、因子分析、方差分析、回归分析、判别分析、主成分分析等、结构方程分析

□基于计算机求解

- 数据挖掘
- 关联规则、回归分析、分类、聚类
- 分类：决策树、朴素贝叶斯、最近邻、支持向量机（SVM）

□可视化技术

- 直观反应数据统计特性、关联关系

□机器学习方法

- 神经网络、深度学习



Python

Python是由荷兰人Guido van Rossum于1989年发明的，并在1991年首次公开发行。它是一款简单易学的编程类工具，同时，其编写的代码具有简洁性、易读性和易维护性等优点，也受到广大用户的青睐。

借助于pandas、statsmodels、scipy等模块用于数据处理和统计分析；matplotlib、seaborn、bokeh等模块实现数据的可视化功能；sklearn、PyML、keras、tensorflow等模块实现数据挖掘、深度学习等操作。



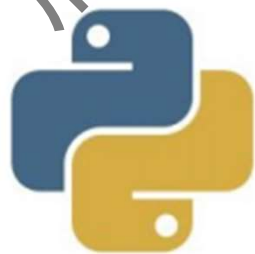
Python数据分析工具

□强大的数据分析工具

- Numpy 、 SciPy 、 Pandas 、 SciKit 、 sklearn、 matplotlib，可用于 数值计算、机器学习和图表绘制
- 专注数据分析方法和模式，代码量很少
- 适用于初学者，同样也适用专家

□与Matlab 、 R语言比较

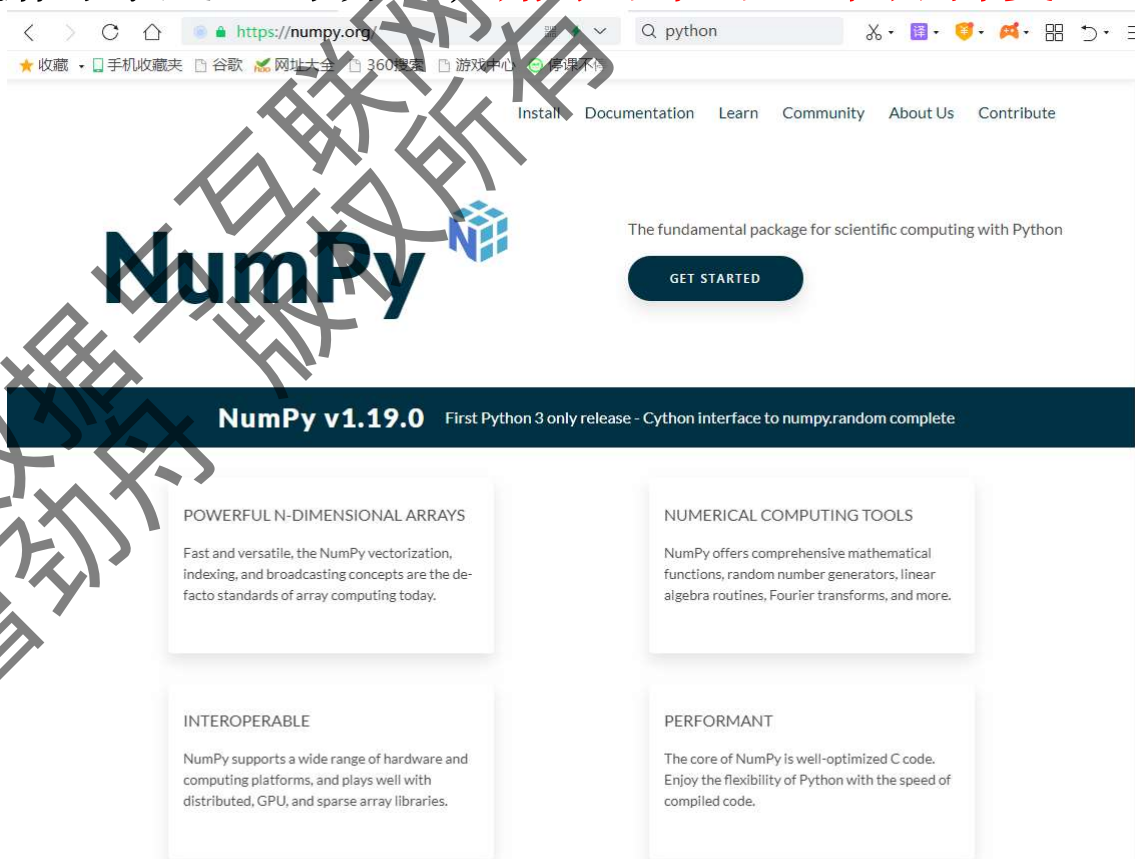
- 良好的可扩展性
- 丰富的第三程序库，紧跟最新技术发展
- 具有速度优势，能够处理大数据



python

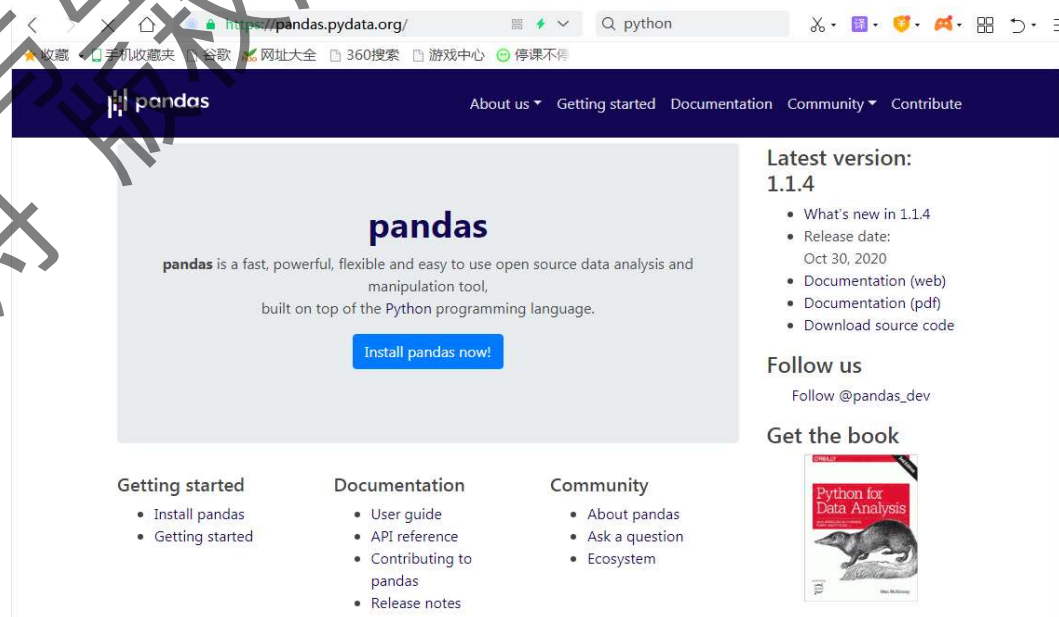
Numpy

NumPy是Python 的**数据分析**的基本库，是在Python的Numeric数据类型的基础上，引入Scipy模块中针对数据对象处理的功能，**用于数值数组和矩阵类型的运算、矢量处理等**。



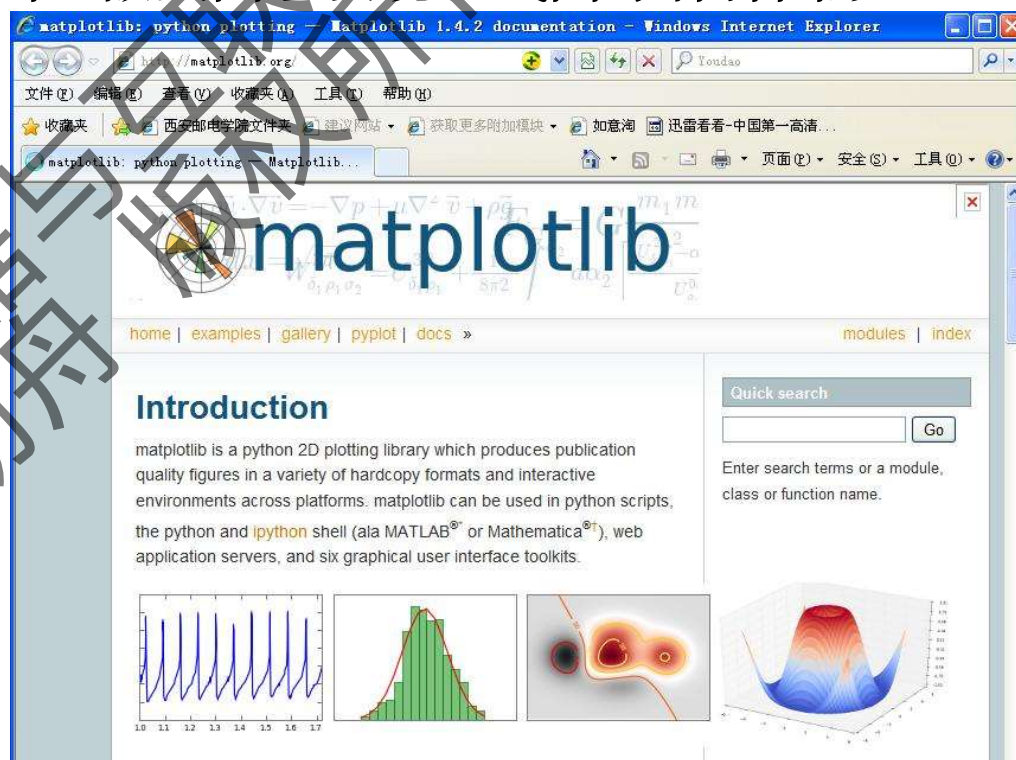
Pandas

Pandas的名称来源于面板数据（Panel Data）和Python数据分析（Data Analysis），作为Python进行数据分析和挖掘时的数据基础平台和事实上的工业标准，支持关系型数据的增、删、改、查，具有丰富的数据处理函数，支持时间序列分析功能，灵活处理缺失数据等。



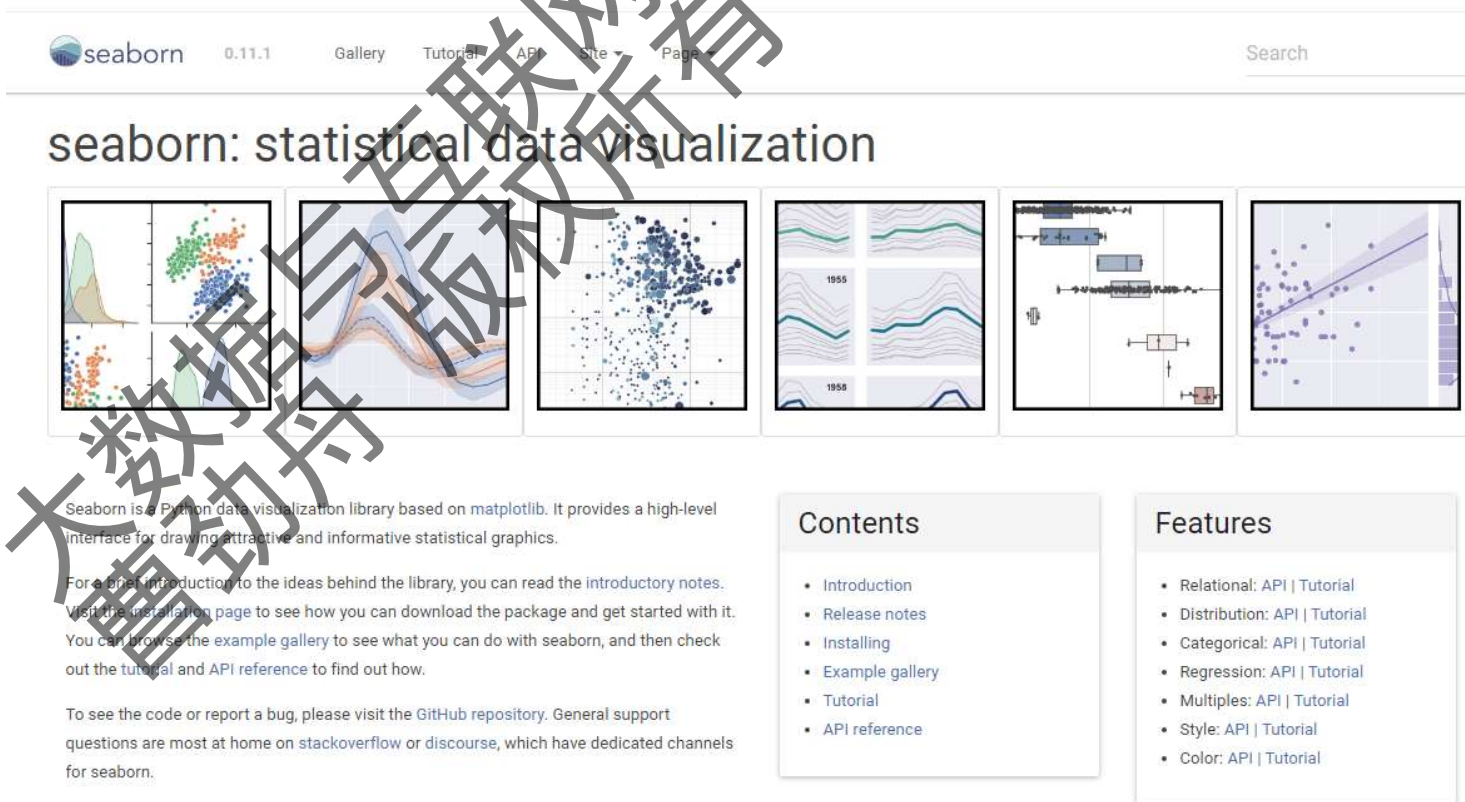
Matplotlib

Matplotlib具有两个重要的模块——pylab和pyplot。PyLab实现了MATLAB的绘图功能，就是MATLAB的Python版本。pyplot主要用于将NumPy统计结果可视化，可以绘制线图、直方图、饼图、散点图以及误差线图等各种图形。



seaborn

Seaborn是图形可视化python包，作为matplotlib的补充，在其基础上进行了更高级的API封装，高度兼容numpy与pandas数据结构以及scipy等统计模式，能做出具有吸引力的图。



The screenshot shows the Seaborn website homepage. At the top, there's a navigation bar with links for 'Gallery', 'Tutorial', 'API', 'Site', and 'Page'. The main heading is 'seaborn: statistical data visualization'. Below this, there's a row of six example plots: a joint plot with marginal histograms, a density plot, a scatter plot, a faceted line plot, a box plot, and a regression plot. Below the plots, there's a paragraph describing Seaborn as a Python data visualization library based on matplotlib. To the right, there are two columns: 'Contents' and 'Features'. The 'Contents' column lists links for Introduction, Release notes, Installing, Example gallery, Tutorial, and API reference. The 'Features' column lists links for Relational, Distribution, Categorical, Regression, Multiples, Style, and Color, each followed by 'API' and 'Tutorial' links.

seaborn 0.11.1 Gallery Tutorial API Site Page Search

seaborn: statistical data visualization



Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the [introductory notes](#). Visit the [installation page](#) to see how you can download the package and get started with it. You can browse the [example gallery](#) to see what you can do with seaborn, and then check out the [tutorial](#) and [API reference](#) to find out how.

To see the code or report a bug, please visit the [GitHub repository](#). General support questions are most at home on [stackoverflow](#) or [discourse](#), which have dedicated channels for seaborn.

Contents

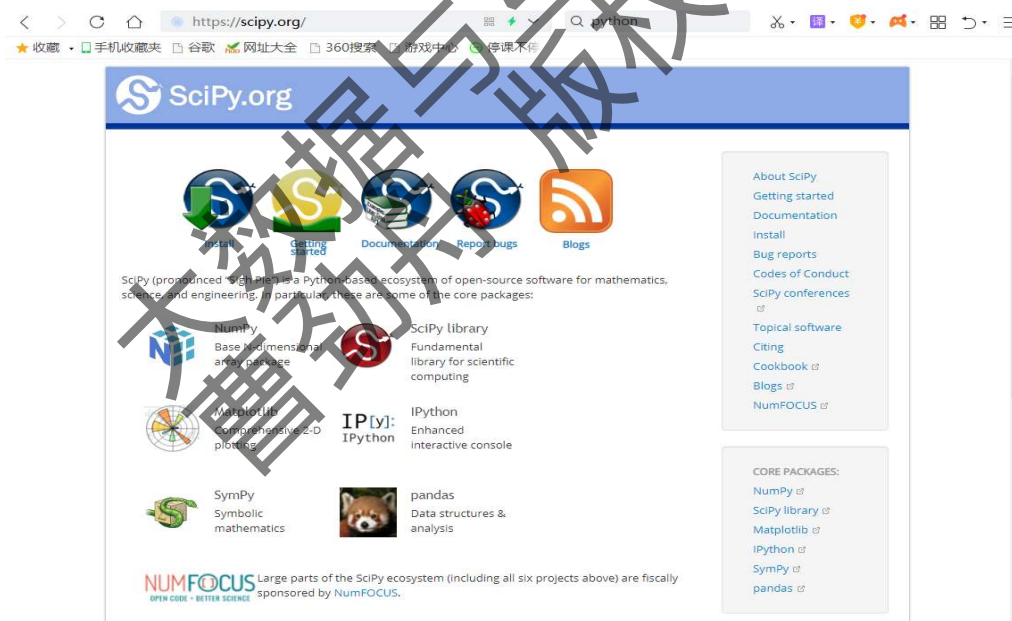
- [Introduction](#)
- [Release notes](#)
- [Installing](#)
- [Example gallery](#)
- [Tutorial](#)
- [API reference](#)

Features

- [Relational: API | Tutorial](#)
- [Distribution: API | Tutorial](#)
- [Categorical: API | Tutorial](#)
- [Regression: API | Tutorial](#)
- [Multiples: API | Tutorial](#)
- [Style: API | Tutorial](#)
- [Color: API | Tutorial](#)

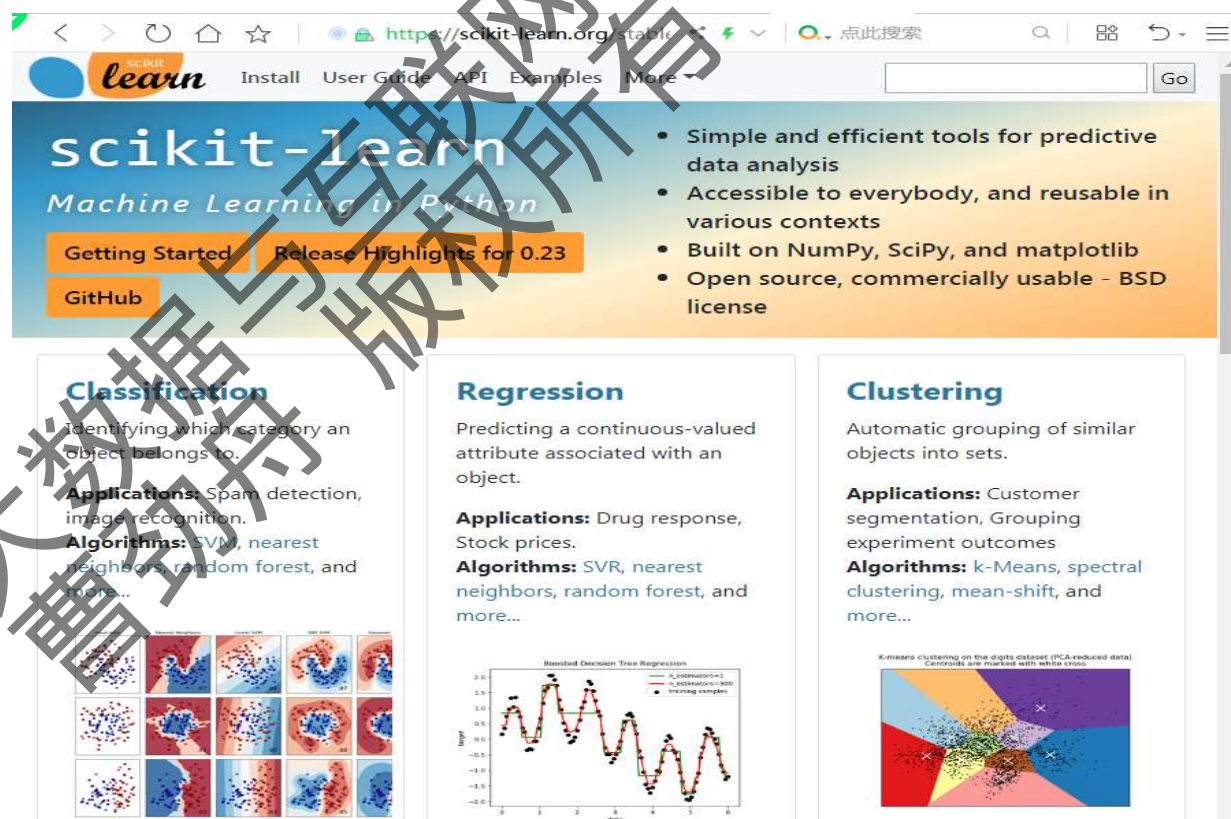
Scipy

Scipy是2001年发行的类似于Matlab和Mathematica等数学计算软件的Python 库，用于统计、优化、整合、线性代数模块、傅里叶变换、信号和图像处理等数值计算。scipy具有stats（统计学工具包）、scipy.interpolate（插值，线性的，三次方）、cluster（聚类）、signal（信号处理）等模块。



Sklearn

Sklearn（又称为scikit-learn）是简单高效的数据挖掘和数据分析工具，基于python语言的NumPy、SciPy 和 matplotlib库之上，是当前较为流行的机器学习框架。



Python编辑器

□ 自带IDE编辑器

□ VScode

□ Pycharm

□ Eclipse+pydev

□ Ulipad

□ Anaconda

大数据与互联网学院
曹劲舟 版权所有