



# 引言

- 大数据的采集与清洗是进行大数据统计分析、机器学习、可视化的必要前提。
- 数据采集与清洗也是输出高质量数据标注成品的前提。
- 数据产生的主体有哪些？
- 数据获取有哪些方法？
- 数据获取的基本流程是什么？
- 数据清洗有哪些方法？
- 数据清洗的基本流程是什么？
- 常用数据可视化技术有哪些？



# 内容提要

## 第0章 概述

主要的数据来源

数据采集方法及基本流程

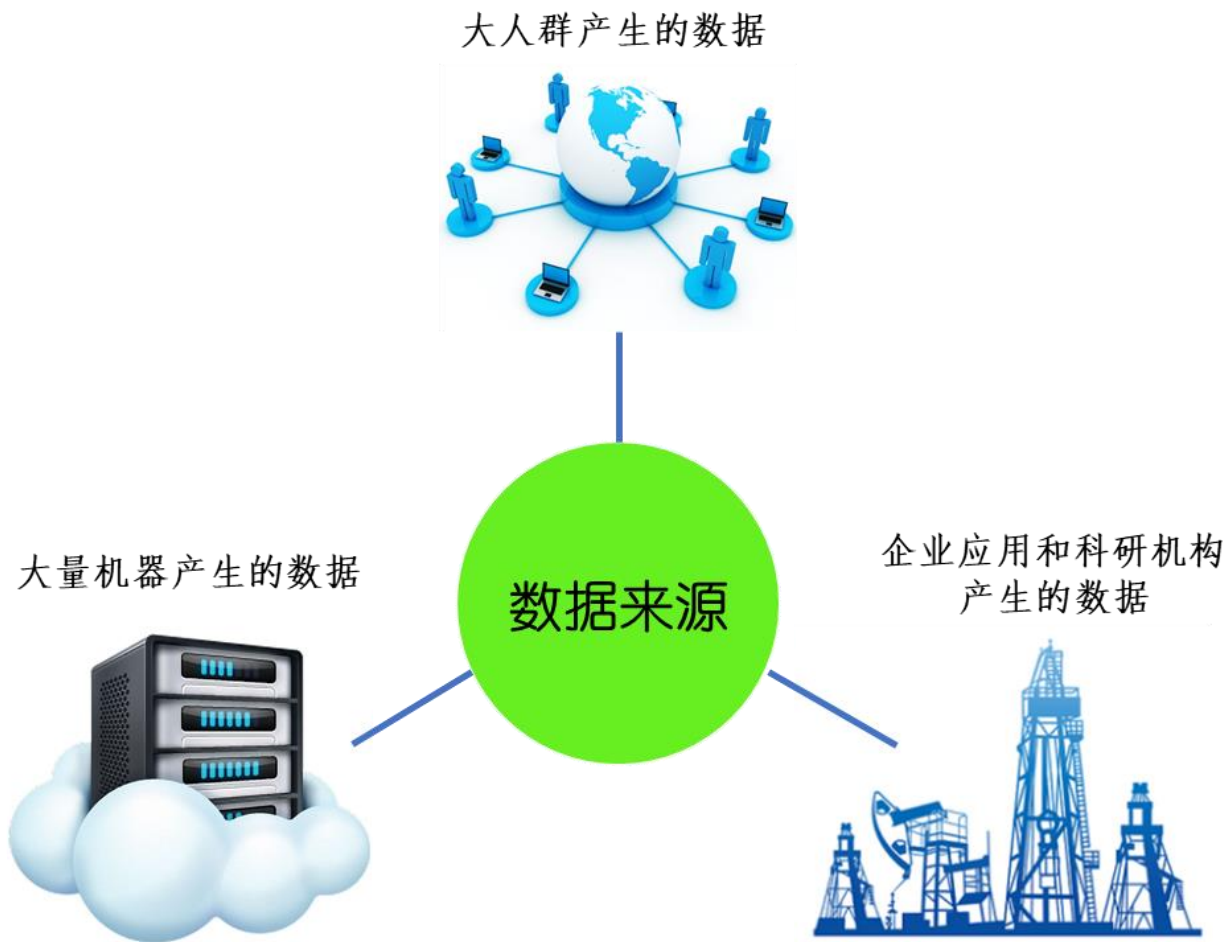
数据清洗方法及基本流程

数据可视化技术

课程主要内容及主要案例

## 0.1 主要的数据来源

- 庞大数据的三大来源



# 0.1 主要的数据来源

## ● 细分（按产生数据的主体）

微博、推特等社交平台数据、在线交易数据、移动通信数据等

大人群产生的数据



大量机器产生的数据



应用服务器日志、各类传感器数据、图像和视频监控数据、二维码和条形码扫描数据等

数据来源

企业应用和科研机构产生的数据



石油、海洋、气象等行业数据和科研机构的再分析数据



## 0.2 数据采集方法

- 系统日志采集
  - 大数据平台下的Kafka、Flume等工具采集实时数据
- 互联网数据采集
  - 编写网络爬虫爬取
  - 通过公开的API，编程来获取，例如欧洲中期天气预报中心 ECMWF 的再分析数据 ERA5数据集。
  - 通过下载工具人工下载
- 通过Web程序，在App移动端或PC端采集



## 0.2 数据采集方法

- 与数据服务机构进行合作
  - 购买数据
  - 共享其全部或部分数据
- 其他：
  - 对人像、车辆、街景等进行现场拍摄
  - 对语音进行人工朗读、转录
  - 直接从书籍、文章中提取特定的文本内容等



## 0.3 数据采集基本流程





## 0.4 数据清洗方法

数据清洗即ETL处理（抽取Extract、转换Transform、加载Load）。

采集端的原始数据需要导入一个专门的数据库中，以便进行有效分析。这些原始数据大体上是不完整、不一致的脏数据，无法直接进行数据分析或挖掘等工作，因此需要进行数据预处理。在导入的同时，应针对缺失信息、不一致信息与冗余信息等完成数据清洗和预处理工作。





## 0.4 数据清洗方法

### ● 缺失值处理

- **删除元组**：直接删除含有缺失属性值的对象。
- **数据补齐**：使用一定的值对缺失属性进行填充补齐，从而使信息表完备化。主要有以下四种方法：
  - ✓ **人工填写**：适用于工作人员非常了解数据相关信息的情况，缺点是效率太低。
  - ✓ **特殊值填充**：例如用“-999”填充，缺点是会导致严重的数  
据偏离。
  - ✓ **平均值填充**：对数值型数据取平均值填充，倾斜分布情况也可以采用中位数填充。非数值型属性采用出现频率最高值填充。
  - ✓ **可能值填充**：通过推断填充缺失值，空值对象周围与其相似的对象值，建立回归模型、贝叶斯模型推理、决策树归纳确定。



## 0.4 数据清洗方法

- 重复数据处理：一般直接合并或者删除
- 删除空行
- 噪声数据处理

噪声数据的出现一般由于收集工具的问题，或数据输入、传输错误，或技术限制等原因。处理方法主要有：

- **回归**：通过函数拟合数据来光滑数据。
- **局部平滑**：通过考察相邻数据来确定最终值。
- **孤立点分析**：通过聚类来检测离群点，落在簇外的数据对象被视为孤立点。

- 数据切割

将包含多个信息的一列文本（即一个列有多个参数）切割成多列的更小的原子数据项

- 数据标准化处理

规范化数据类型、统一格式、统一单位

- 没有列头的字段命名或字段重命名



## 0.4 数据清洗方法

### ● 数据变换

- 数据透视（把垂向显示的两列数据旋转到横向显示，其中一列的不同取值作为新表的多个列的列名，另一列的数据作为这些列对应的值，即把长格式数据变换成宽格式数据）
- 数据融合（也叫数据逆透视，把宽格式数据变换成长格式数据）

Year	Course	Earning
2021	Python	1000.00
2021	Java	2000.00
2021	C++	2500.00
2022	Python	1700.00
2022	Java	2100.00
2022	C++	2400.00

数据透视

Year	Course列的不同取值		
	Python	Java	C++
2021	1000.00	2000.00	2500.00
2022	1700.00	2100.00	2400.00

Earning列的值

对应新表中Course列的不同取值

Year	Python	Java	C++
2021	1000.00	2000.00	2500.00
2022	1700.00	2100.00	2400.00

对应新表中Earning列的值

数据融合

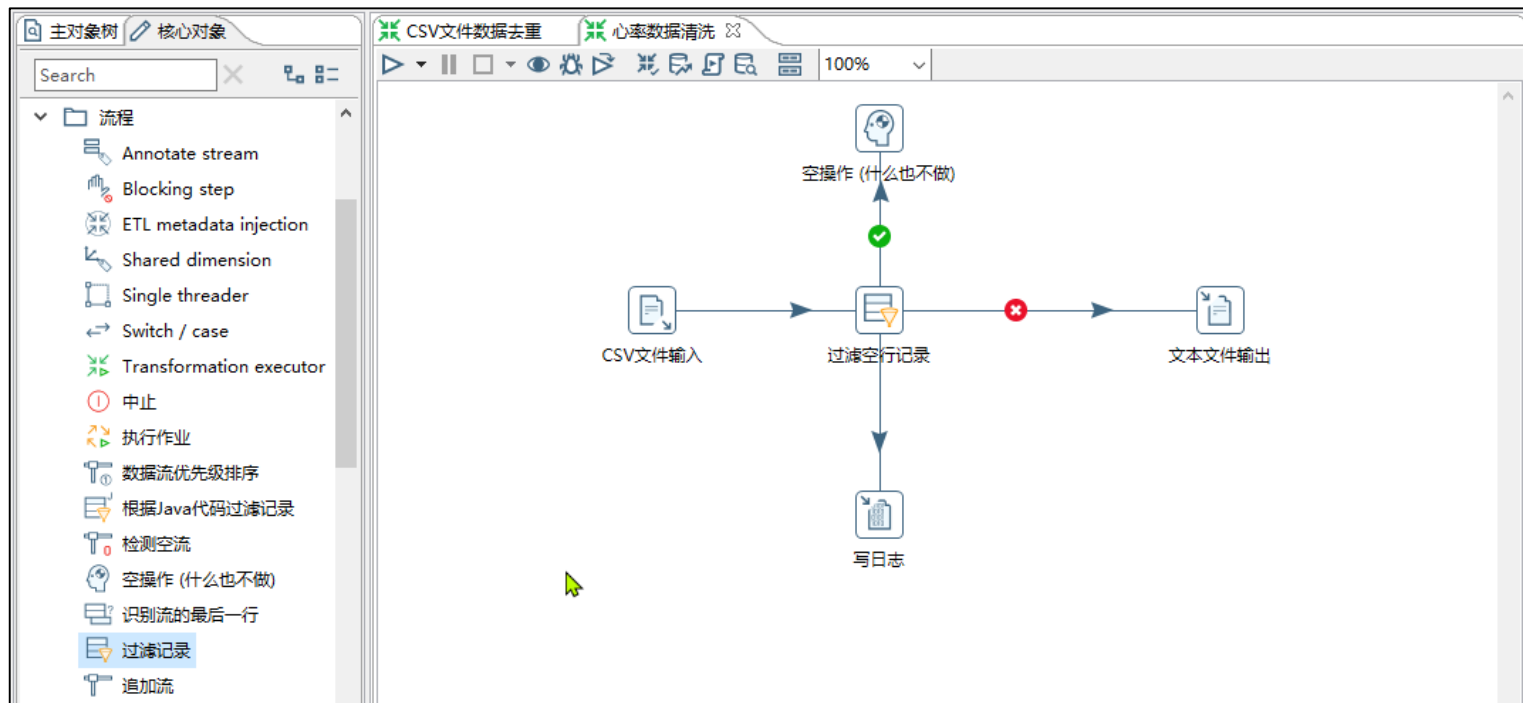
Year	Course	Earning
2021	Python	1000.00
2022	Python	1700.00
2021	Java	2000.00
2022	Java	2100.00
2021	C++	2500.00
2022	C++	2400.00



## 0.4 数据清洗方法

### ● 清洗工具

- Excel
- Pandas工具包
- 可视化的ETL工具Kettle ( 改名为PDI, Pentaho Data Integration )



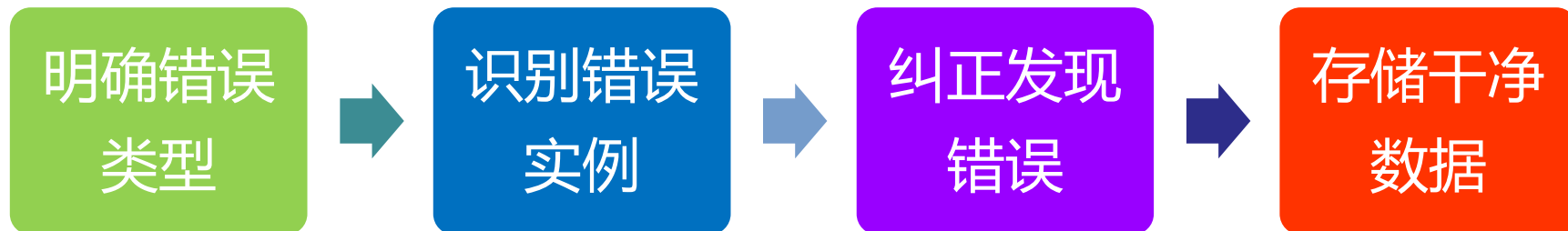


## 0.5 数据清洗的基本流程

### 第0章 概述

- 手动检查或数据样本等数据分析方式
- 定义清洗转换规则与 workflow，根据情况决定数据转换和清洗步骤

- 按预定义的清洗规则和工作流有序进行



- 人工，但耗时耗力准确率低
- 通过统计、聚类或关联规则，自动检测

- 存储备用，避免重复清洗



## 0.6 数据可视化技术

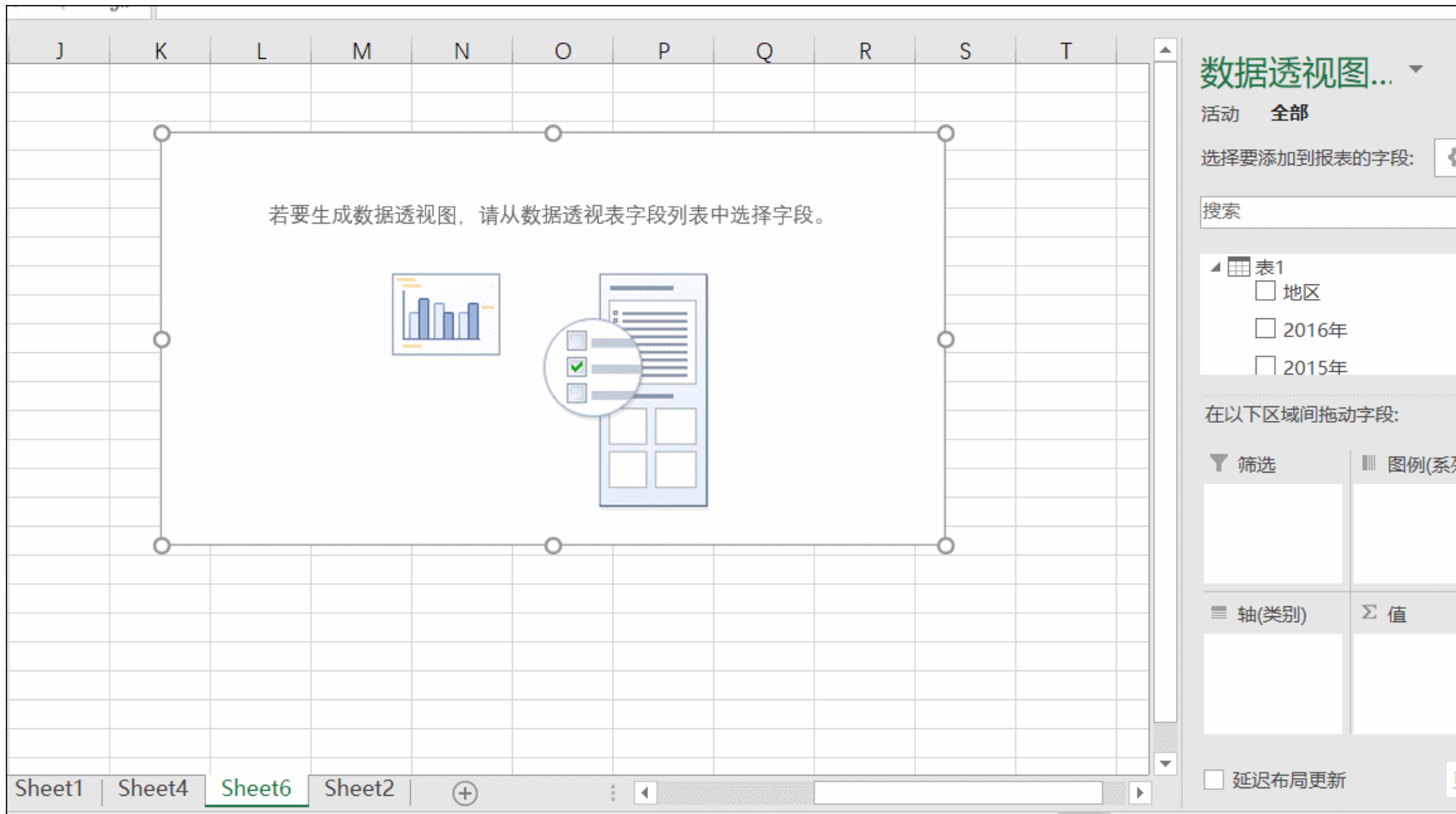
### 第0章 概述

- 使用软件工具可视化数据
  - Excel可视化数据
  - Panoply可视化NetCDF文件中的数据
- 编写程序代码
  - 高级语言C++.NET、Java等使用底层GDI+绘图
  - 高级语言C++.NET、Java等使用插件绘图，例如TeeChart、OxyPlot
  - 脚本语言Matlab、Python、R等使用工具包可视化数据，例如Matplotlib、Pycharts、Seaborn、ggplot



## 0.6 数据可视化技术

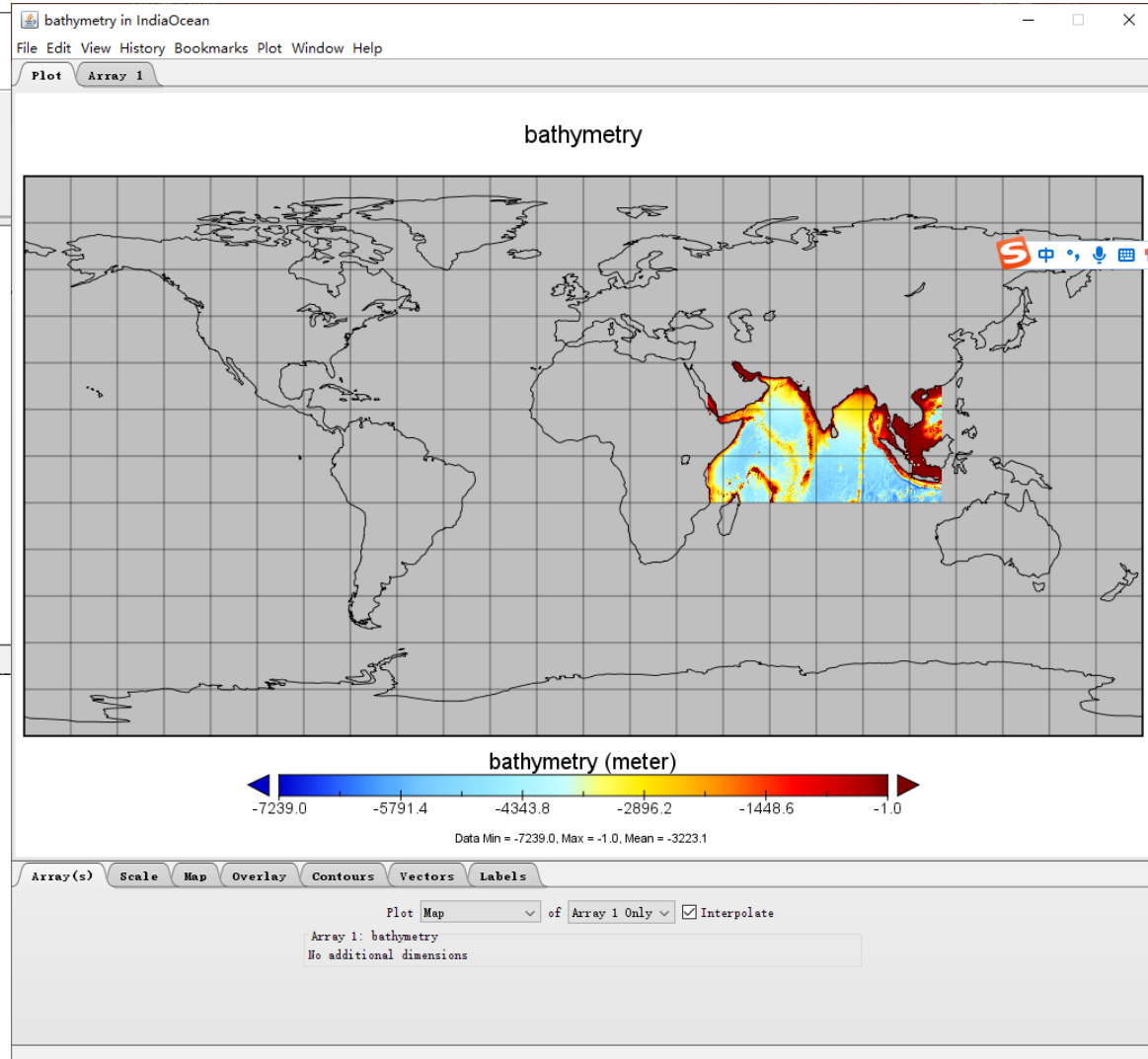
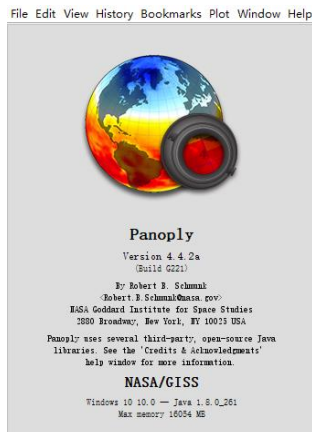
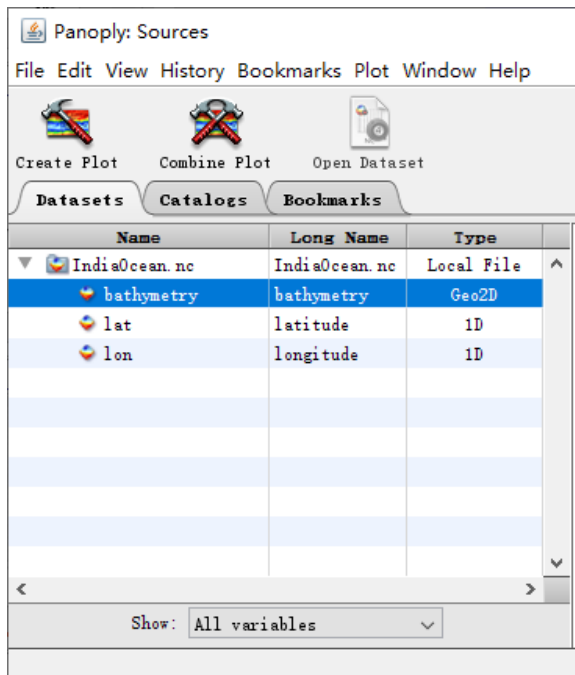
### 第0章 概述





# 0.6 数据可视化技术

## 第0章 概述







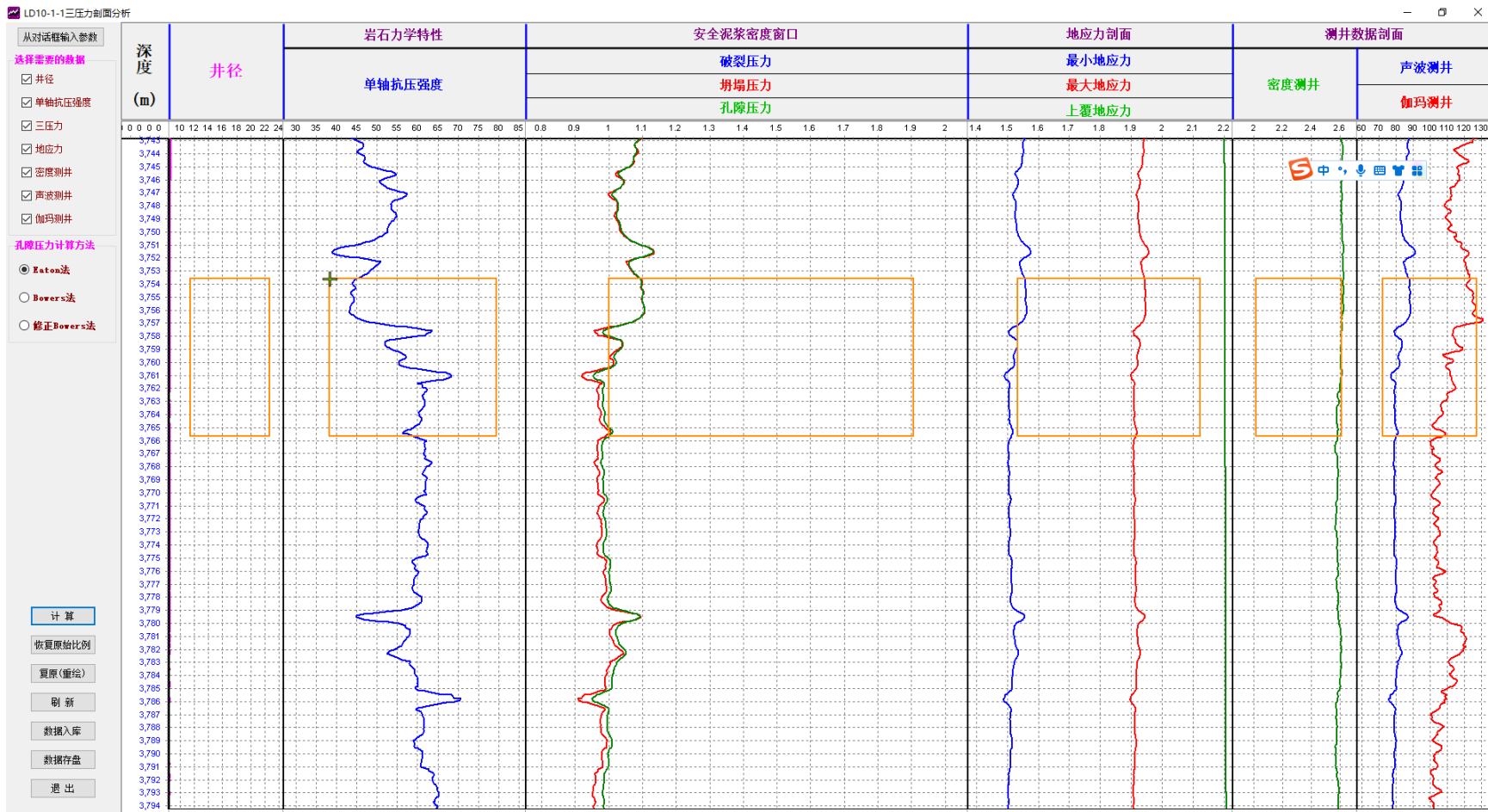
## 0.6 数据可视化技术





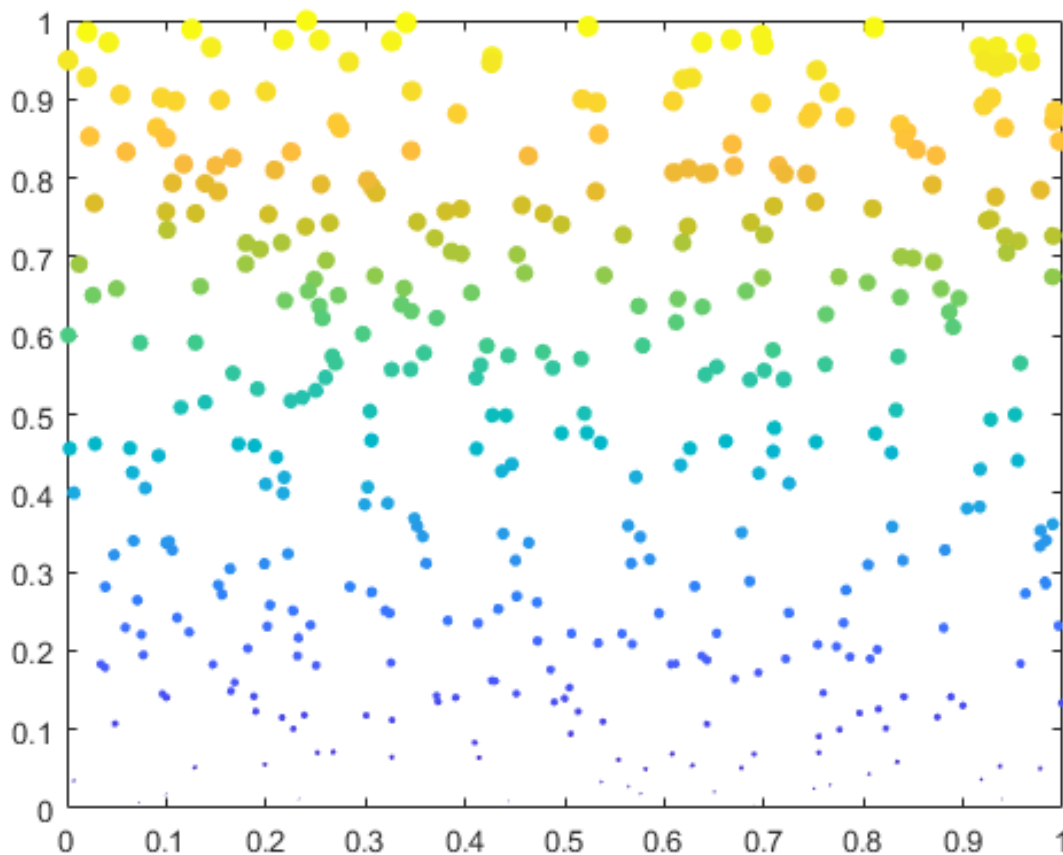
# 0.6 数据可视化技术

## 第0章 概述





## 0.6 数据可视化技术



Matlab绘制散点图



# 0.6 数据可视化技术

## 第0章 概述

Gallery — Matplotlib 3.3.0

<https://matplotlib.org/gallery/index.html>

# matplotlib

Version 3.3.0

Installation Documentation Examples Tutorials Contributing

home | contents » modules | index

## Gallery

This gallery contains examples of the many things you can do with Matplotlib. Click on any image to see the full image and source code.

For longer tutorials, see our [tutorials page](#). You can also find [external resources](#) and a [FAQ](#) in our [user guide](#).

### Lines, bars and markers

- Lines, bars and markers
- Images, contours and fields
- Subplots, axes and figures
- Statistics
- Pie and polar charts
- Text, labels and annotations
- Pyplot
- Color
- Shapes and collections
- Style sheets
- Axes Grid

NEW 惨! 纽约时尚女总裁被大白鲨活活咬死, 水中正嬉笑瞬间被拖入海底

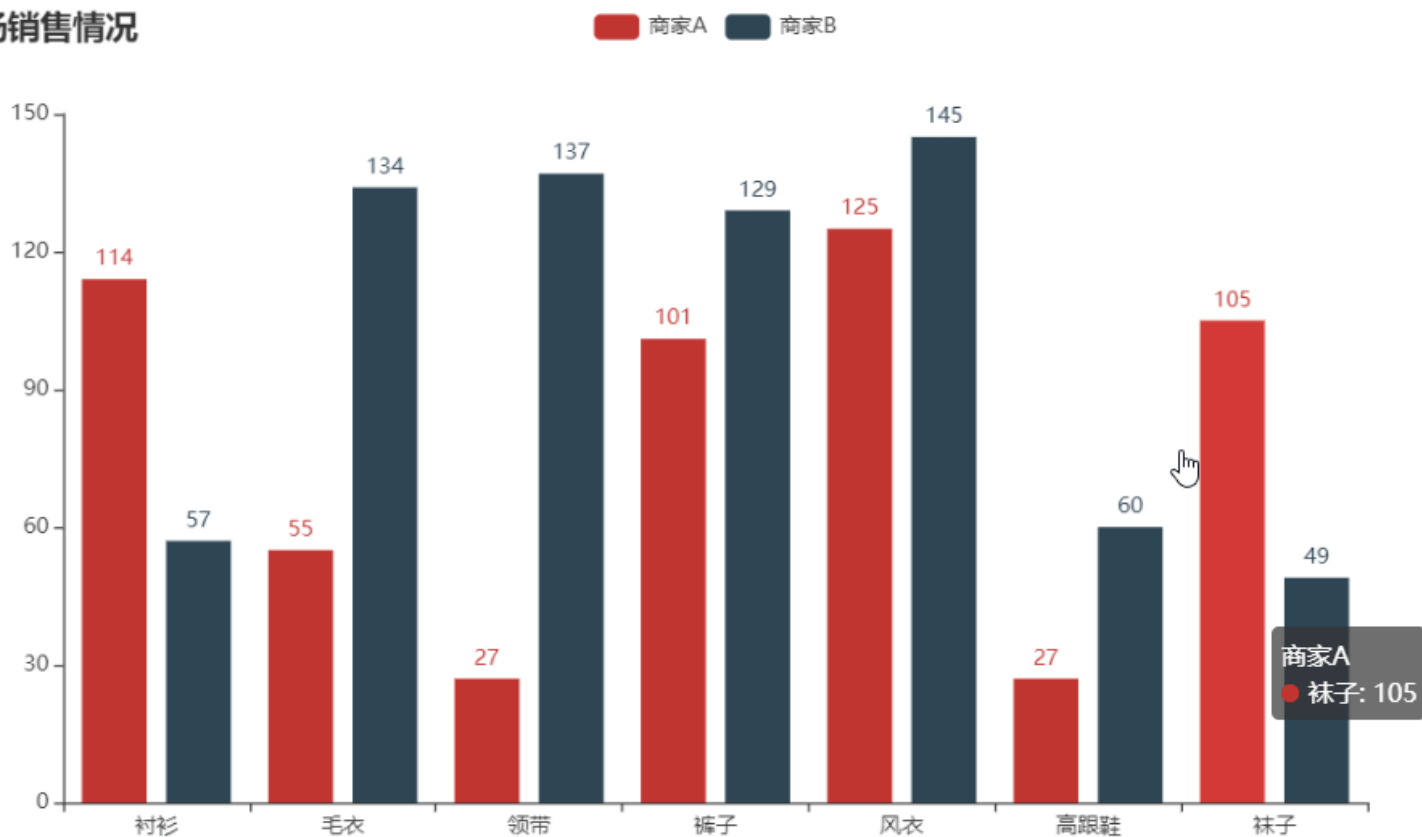
我的视频

Python使用Matplotlib工具包可视化数据



## 0.6 数据可视化技术

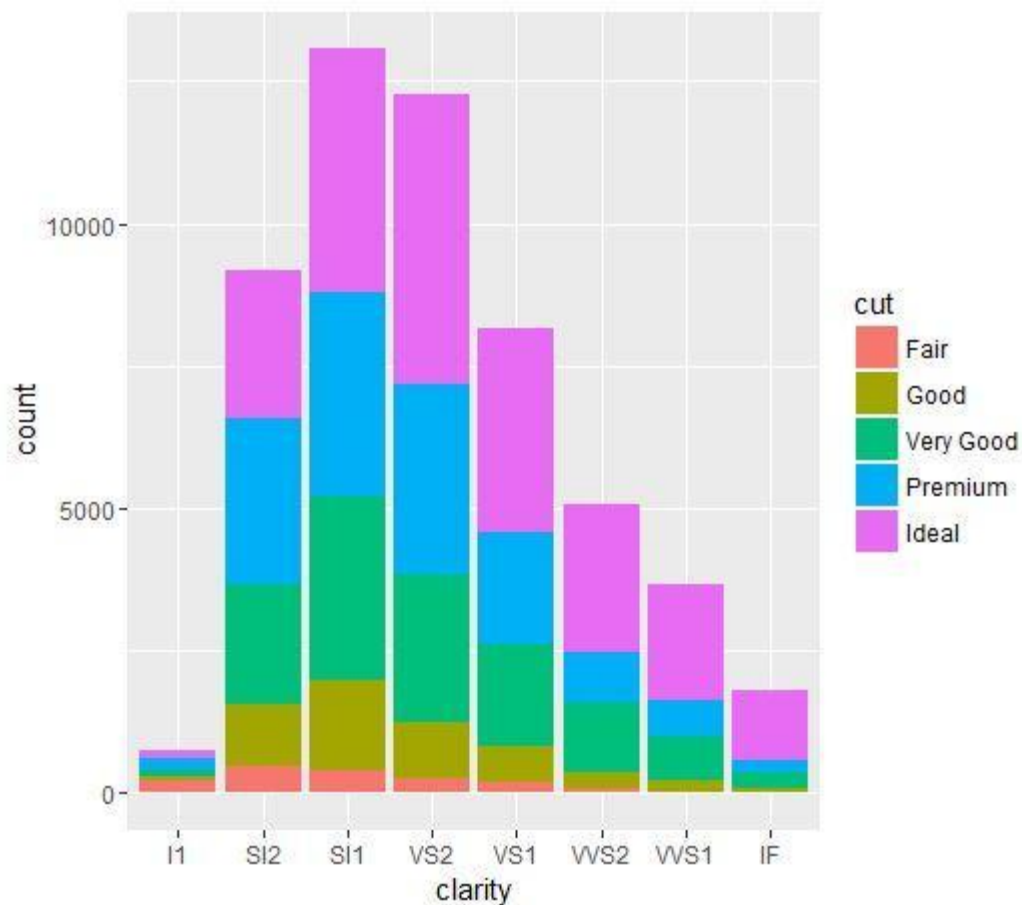
某商场销售情况



Python使用Pyecharts工具包可视化数据



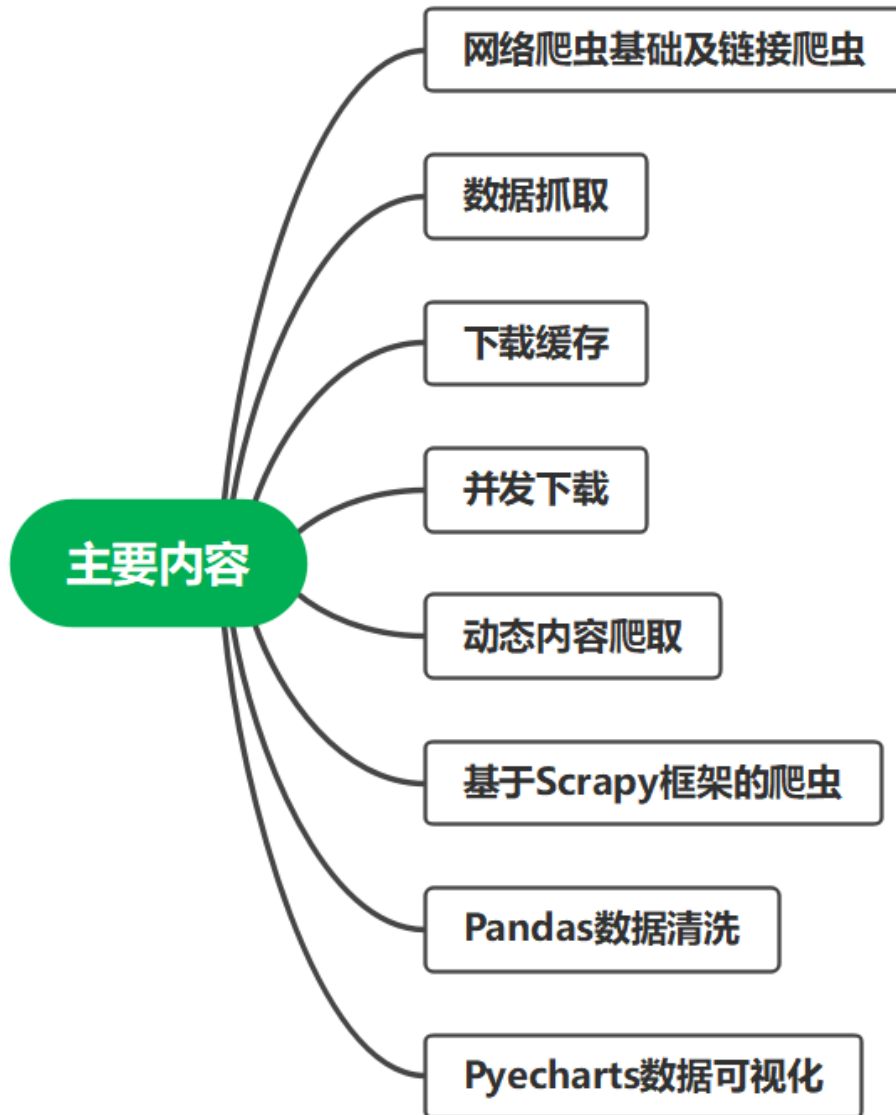
## 0.6 数据可视化技术



Python使用ggplot工具包可视化数据

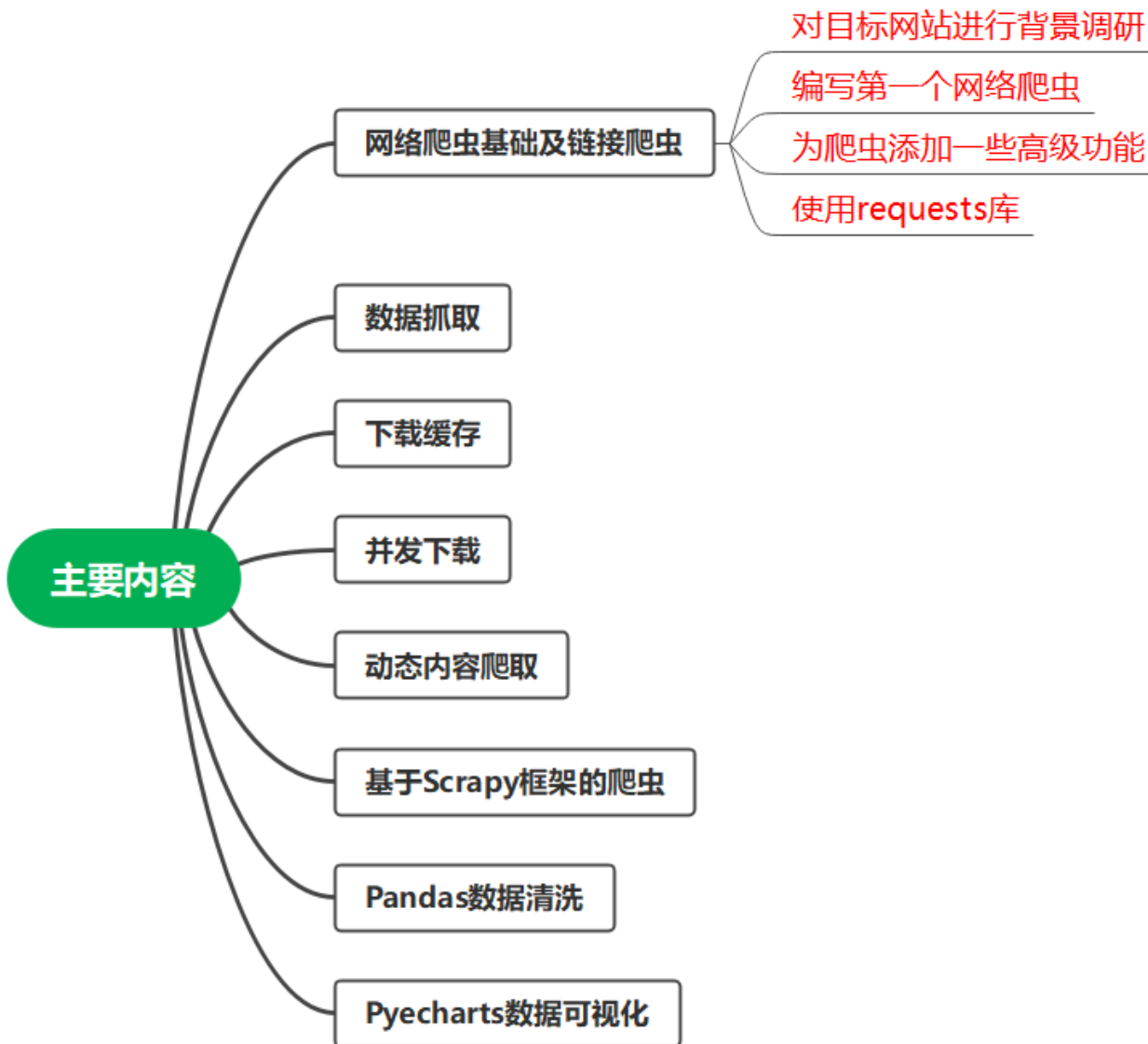


## 0.7 本课程主要内容





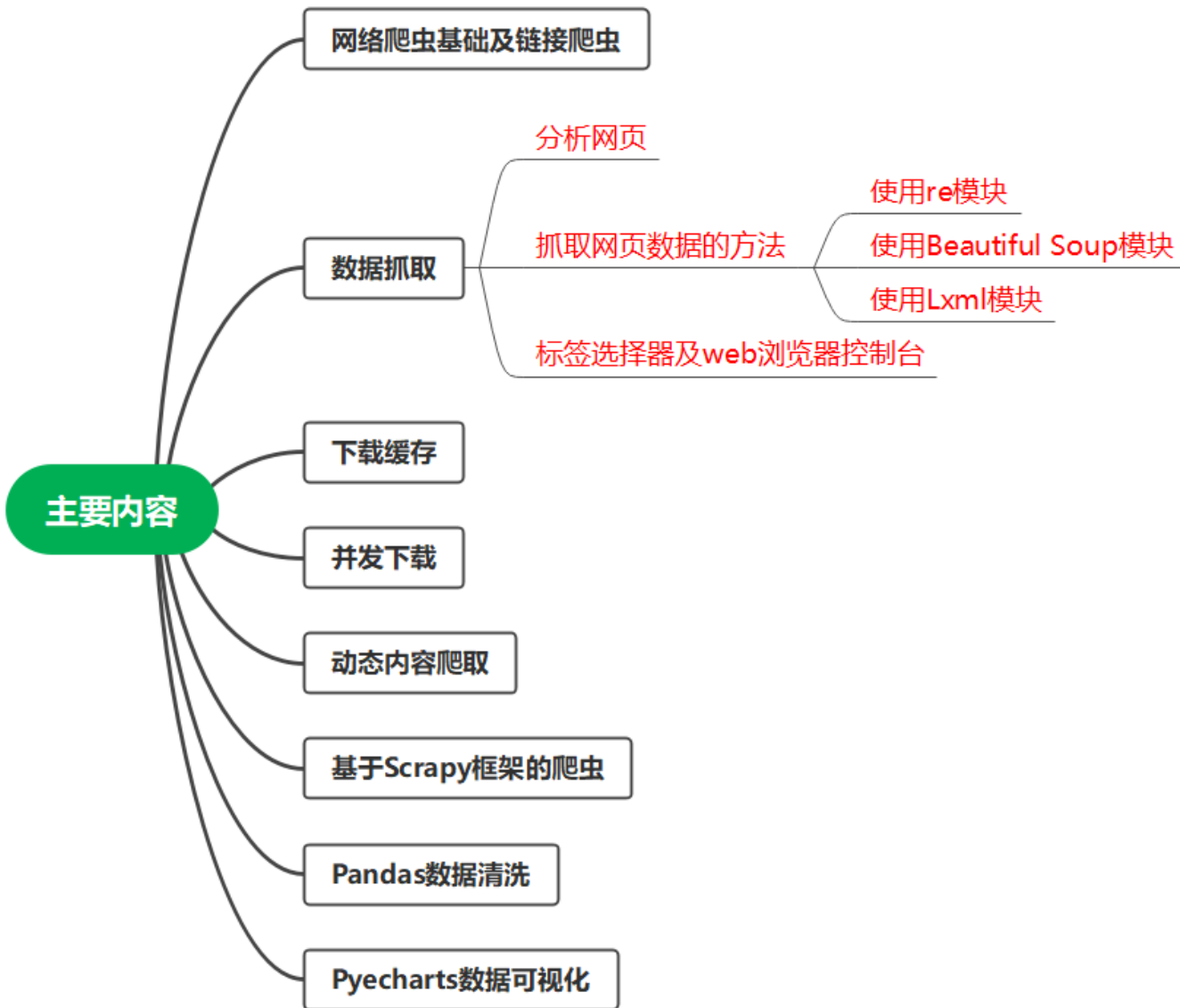
## 0.7 本课程主要内容







## 0.7 本课程主要内容



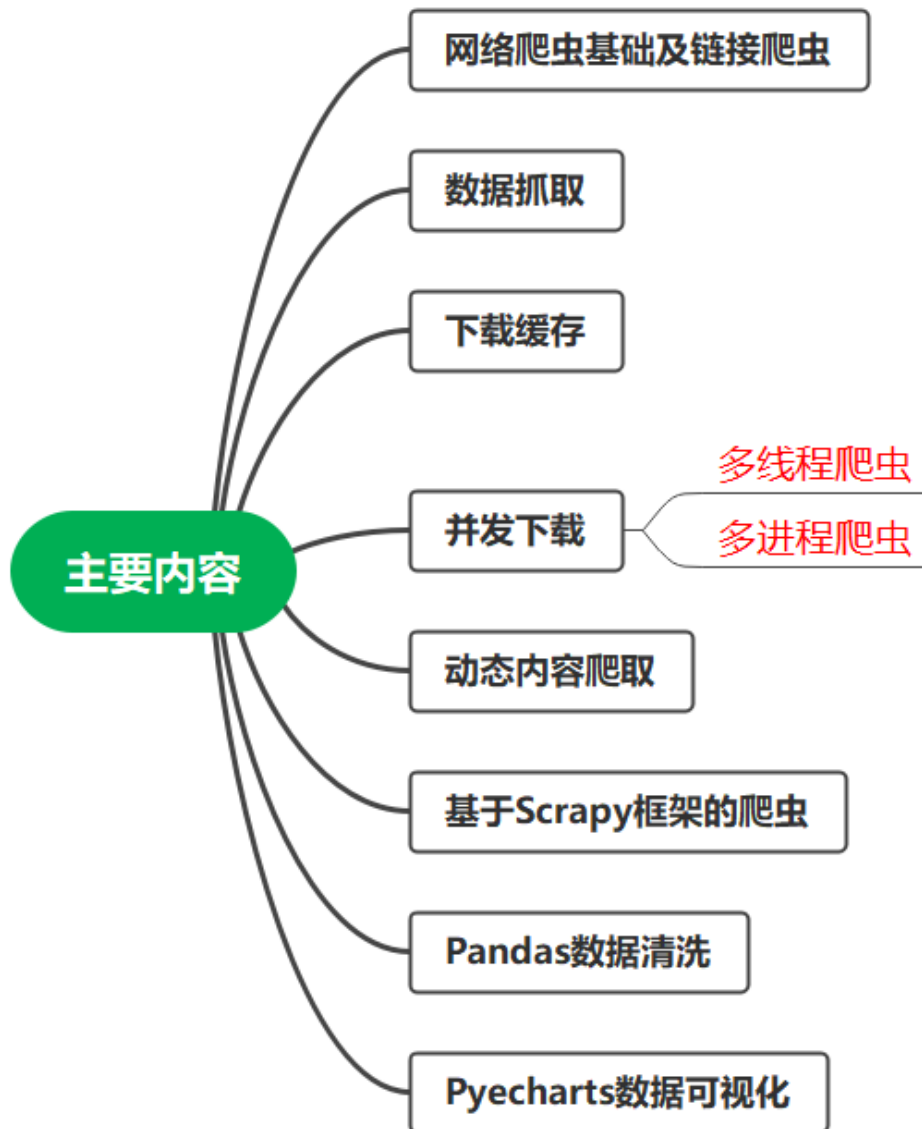


## 0.7 本课程主要内容



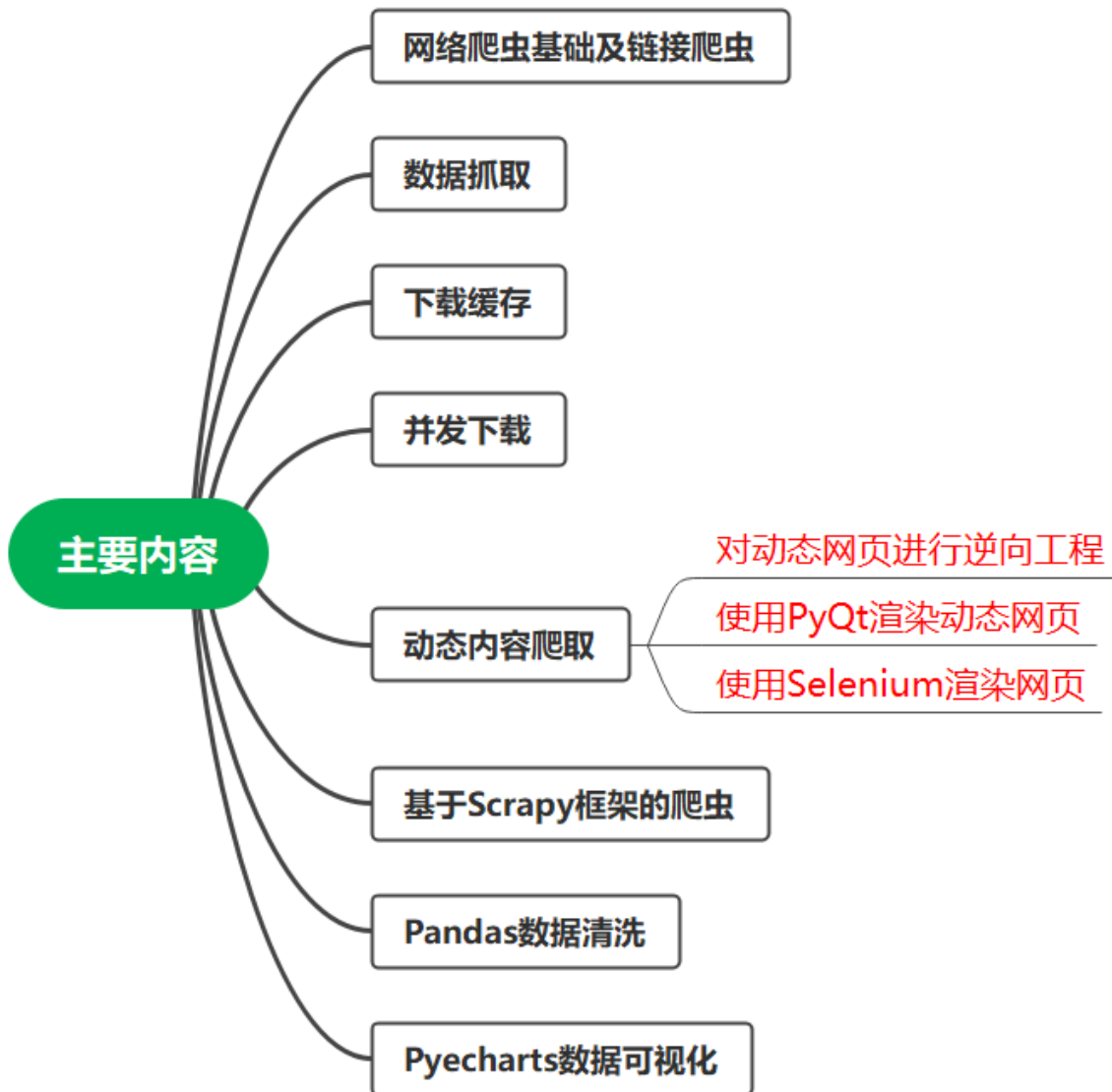


## 0.7 本课程主要内容





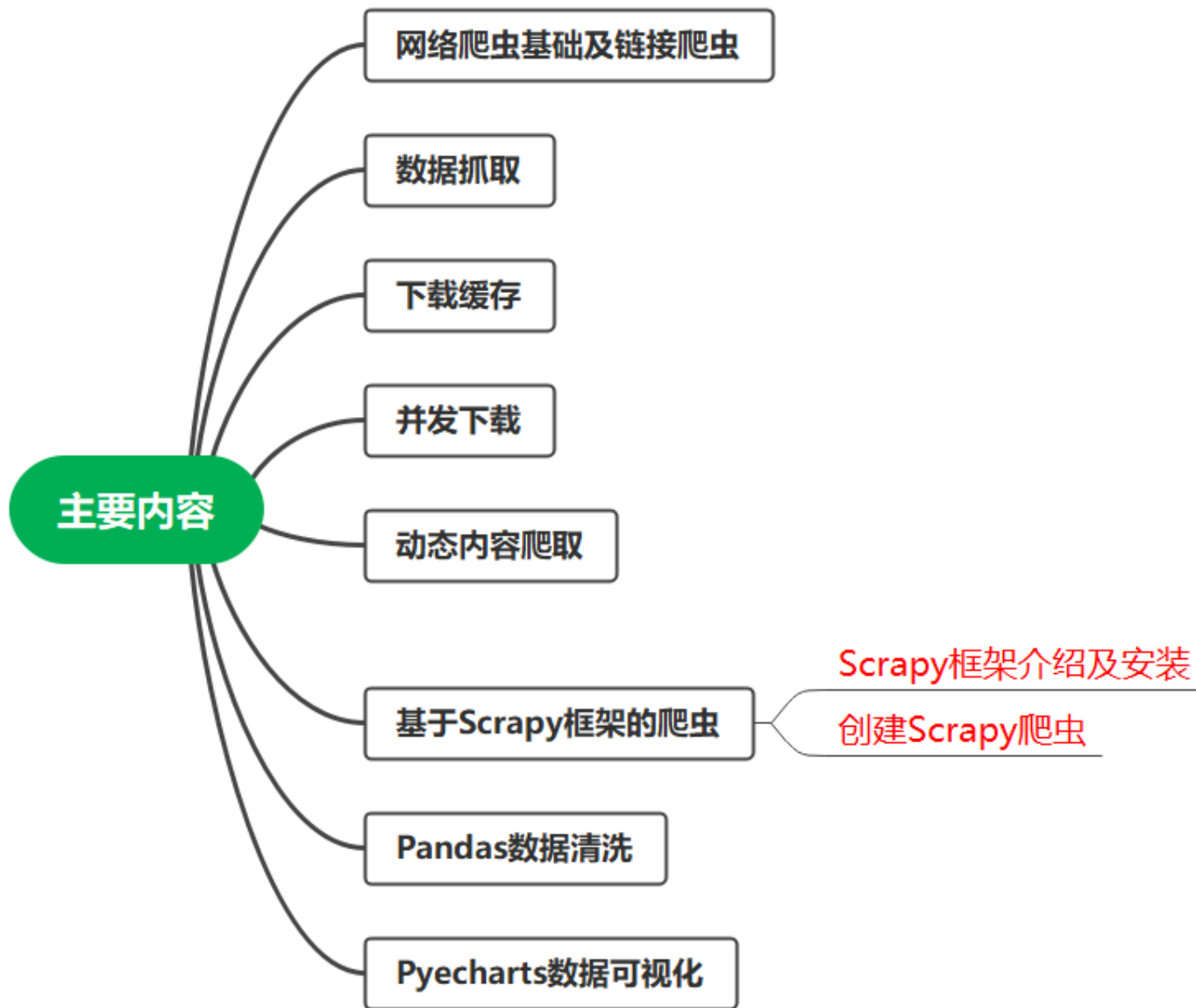
## 0.7 本课程主要内容





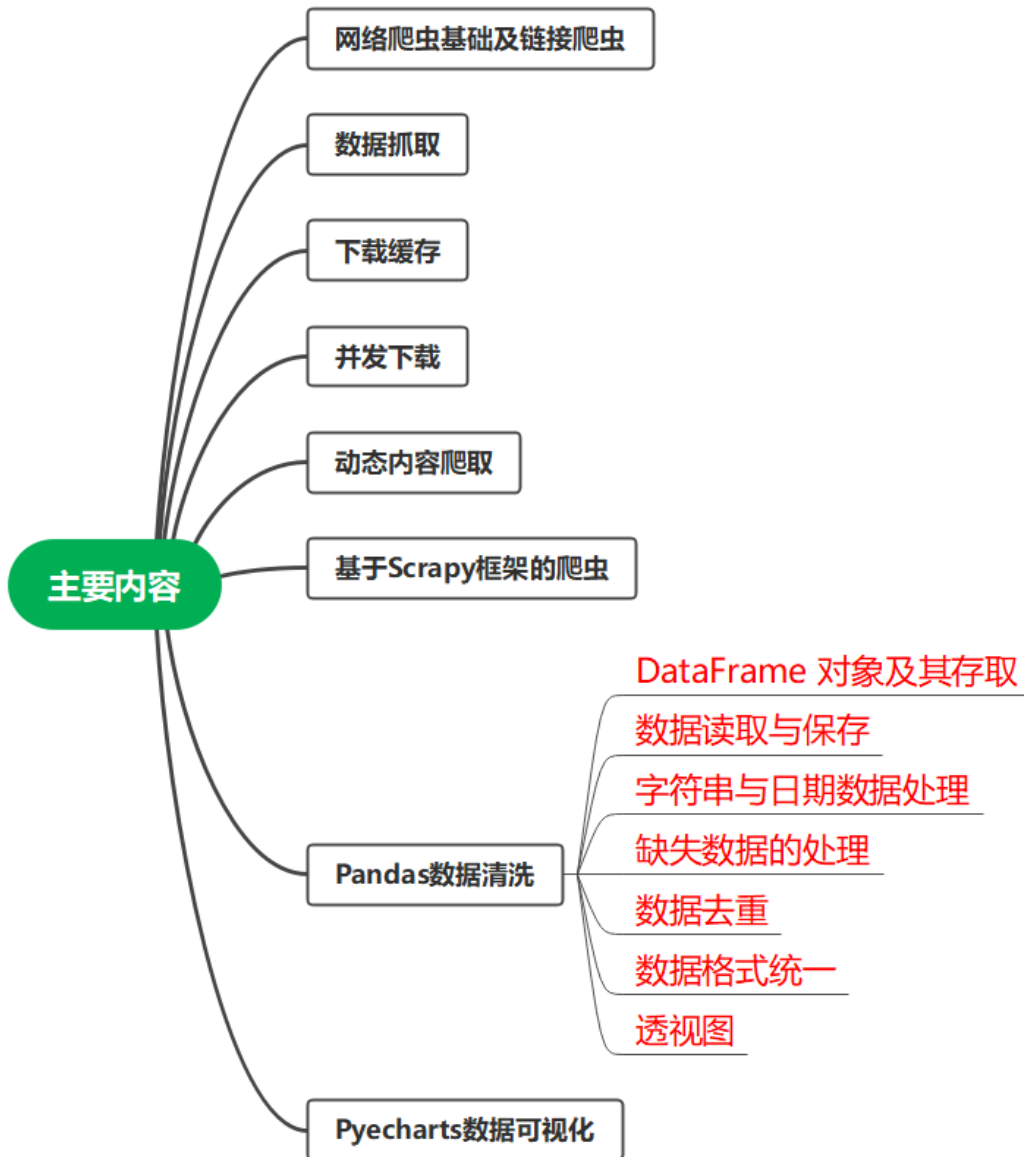
## 0.7 本课程主要内容

### 第0章 概述





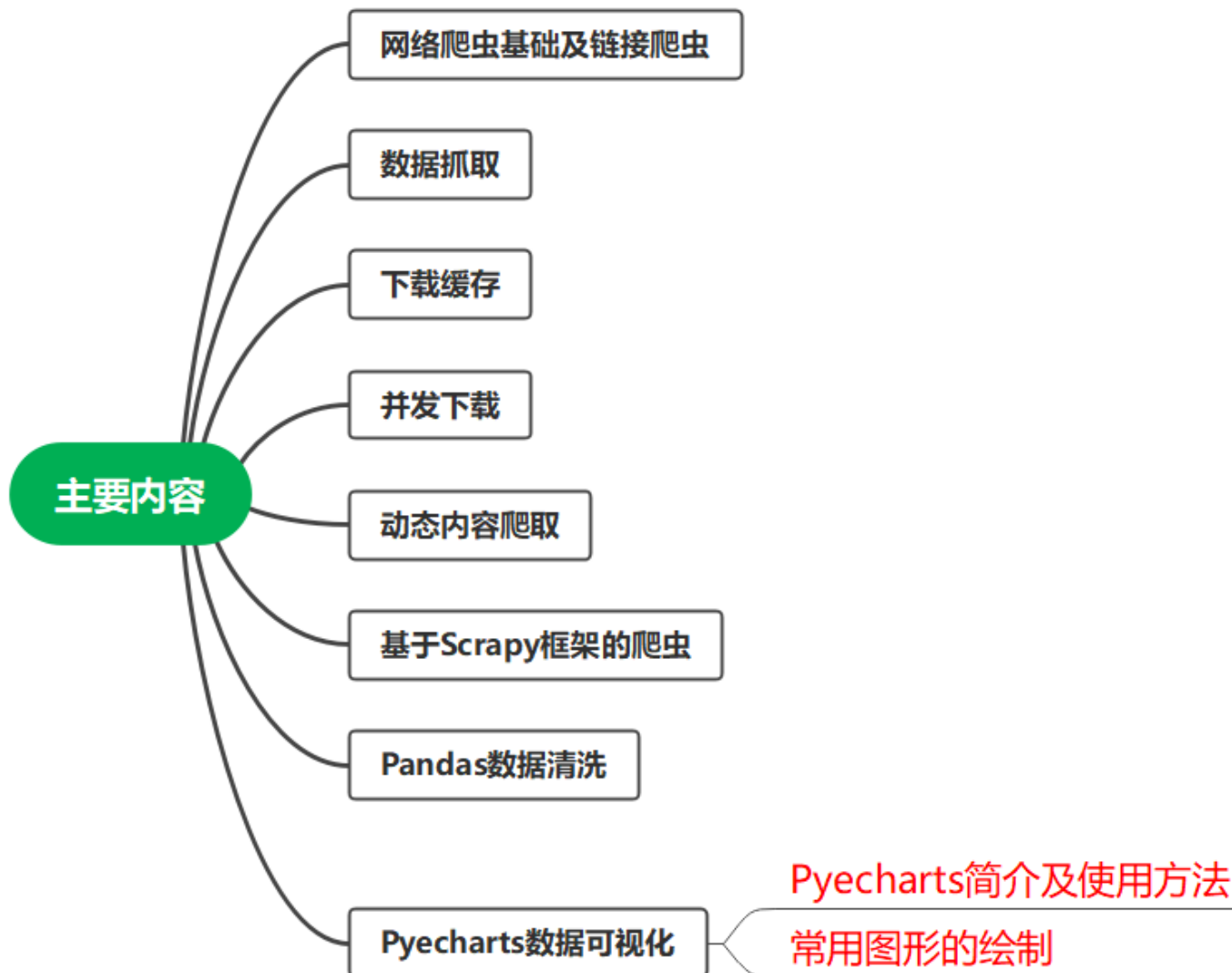
## 0.7 本课程主要内容

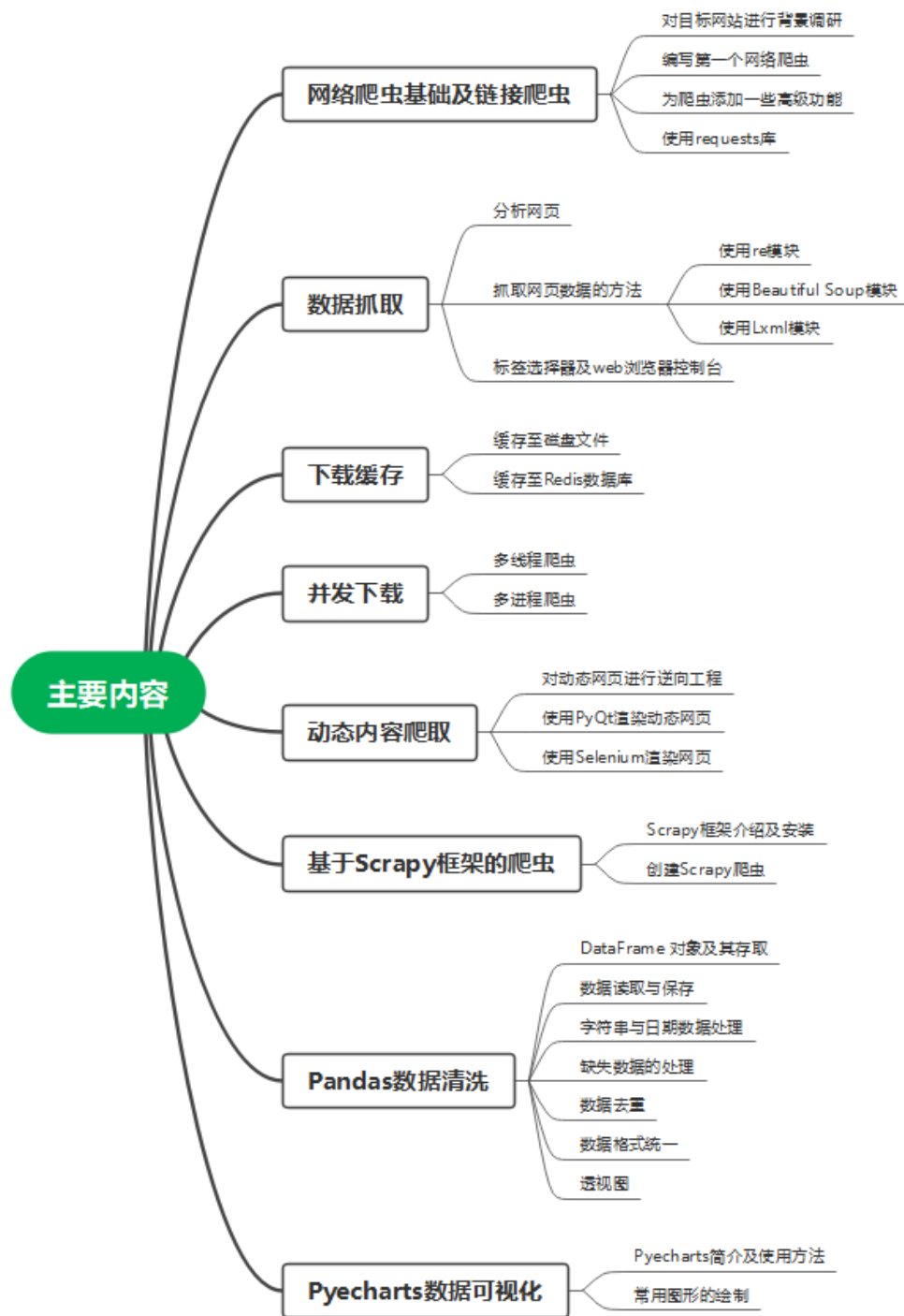




## 0.7 本课程主要内容

### 第0章 概述



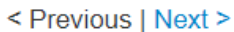





## 0.8 主要案例

- ## ● 爬取Python爬虫示例网站的国家的国家（或地区）基本信息

# Example web scraping website



Flag:	
Area:	244,820 square kilometres
Population:	62,348,447
Iso:	GB
Country (District):	United Kingdom
Capital:	London
Continent:	<a href="#">EU</a>
Tld:	.uk
Currency Code:	GBP
Currency Name:	Pound
Phone:	44
Postal Code Format:	@# #@ @ @## #@ @ @# #@ @ @#
Postal Code Regex:	^((([A-Z]\d{2}[A-Z]{2})) ([A-Z]\d{3}[A-Z]{2}) ([A-Z]{2})(GIR0AA))\$
Languages:	en-GB,cy-GB,gd
Neighbours:	<a href="#">IE</a>
<a href="#">Edit</a>	

包含200多个国家或地区数据



## 0.8 主要案例

### ● 爬取中国铁路12306网站火车票信息





## 0.8 主要案例

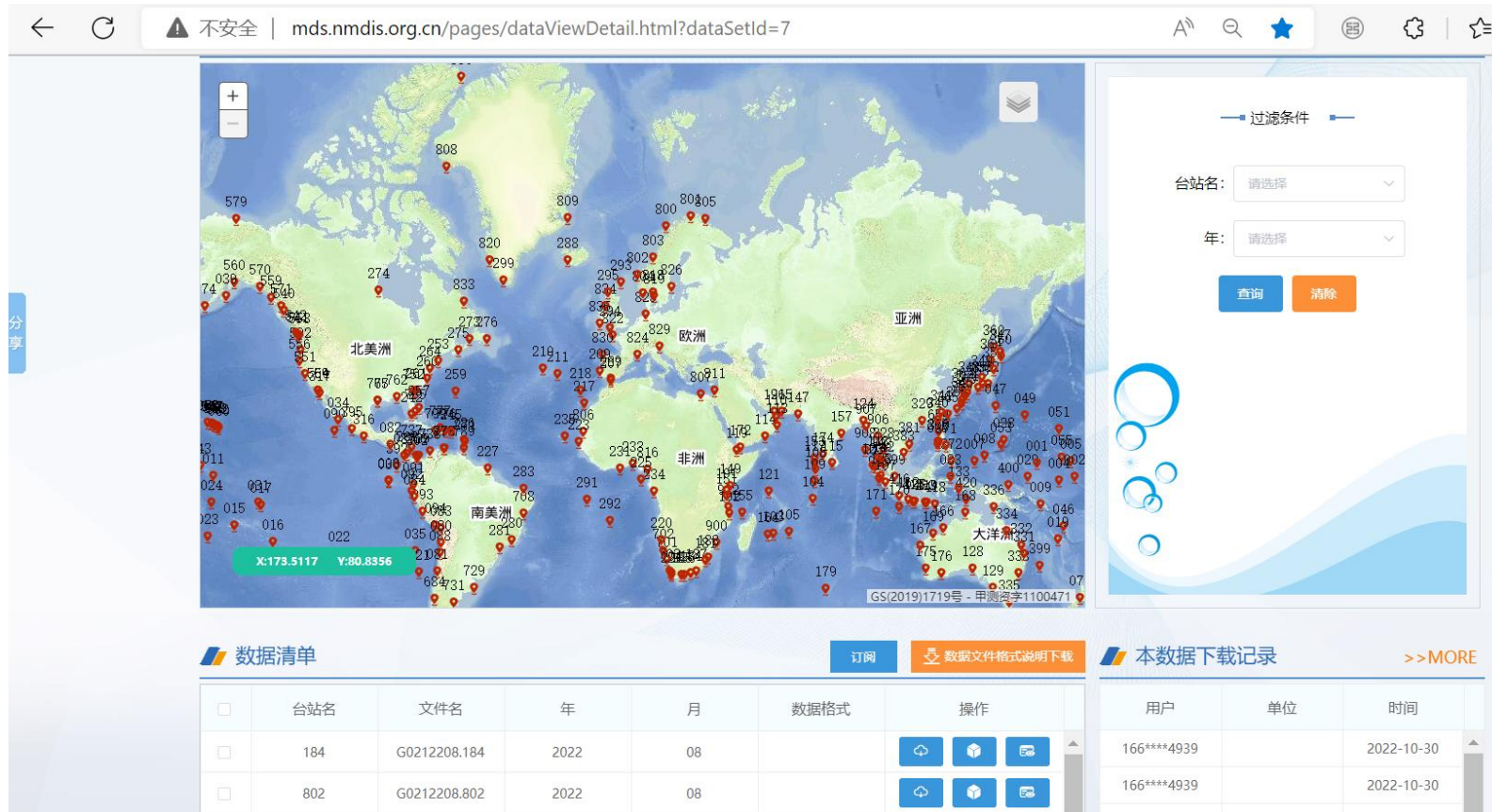
### ● 爬取京东商品信息





## 0.8 主要案例

### ● 下载国家海洋科学信息全球水位观测数据



目前文件数量: 16799个





## 0.8 主要案例

- 爬取石油大学新闻网新闻，清洗并可视化



目前新闻数量： 6000多条



## 0.9 本章小结

- 本章介绍了主要数据来源、数据采集方法与基本流程、数据清洗的方法与基本流程、可视化技术、课程主要内容及案例。
- 下一章将介绍网络爬虫的基本知识并创建一个链接爬虫。



谢谢大家!