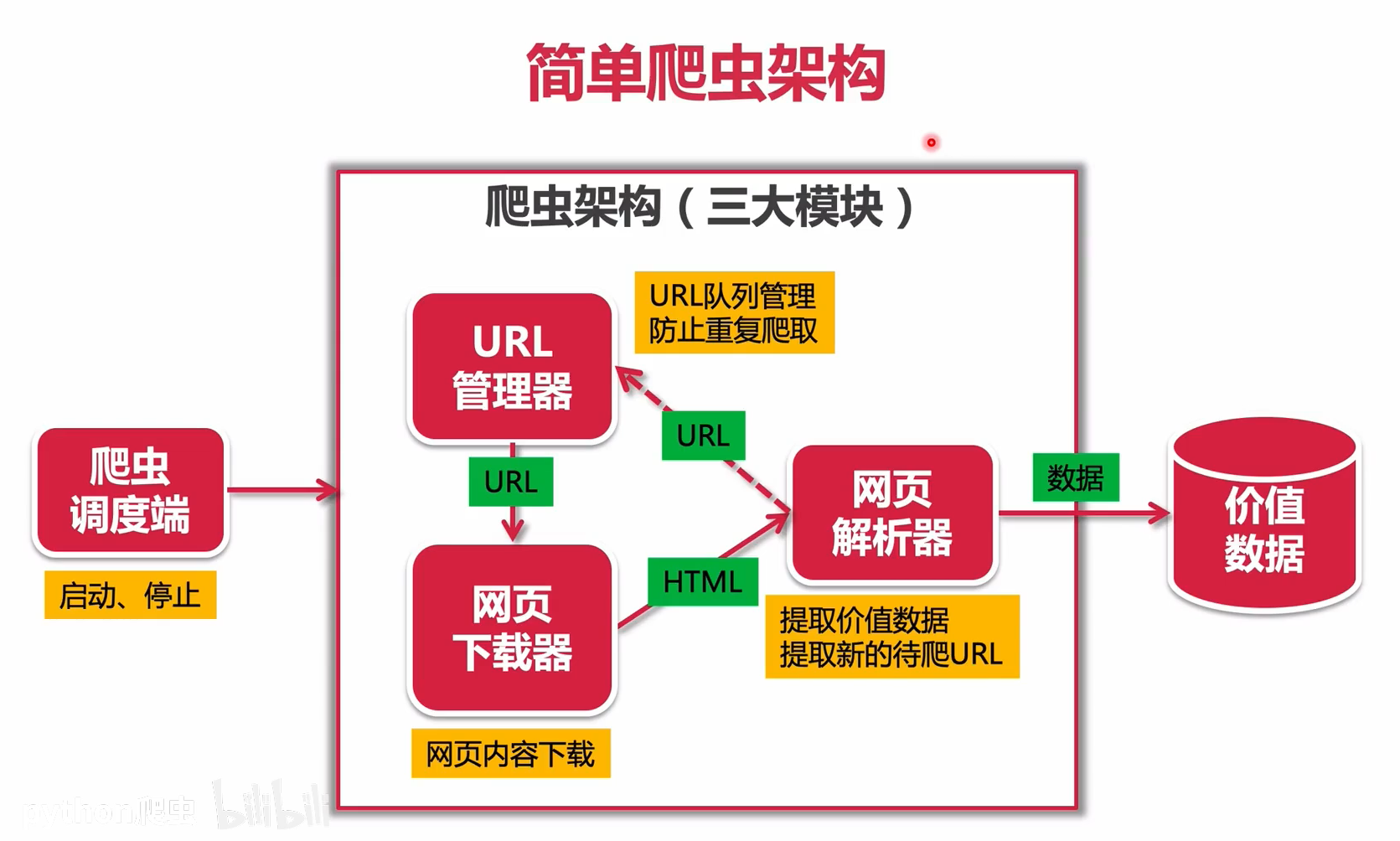
## 1. 爬虫架构



## 2. 下载器模块





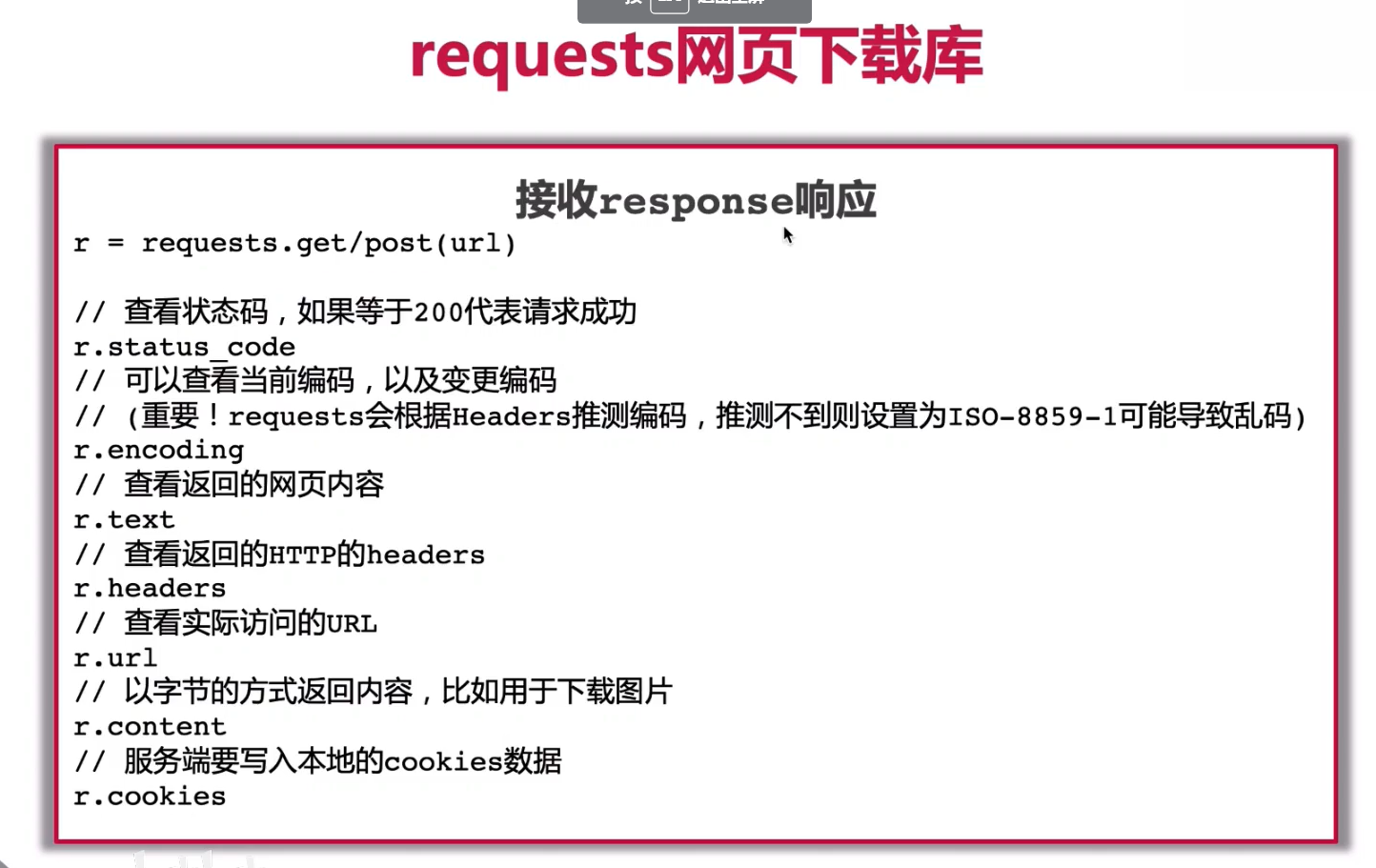
2.1 表单→字典；json打包为字符串

2.2 避免反扒识别

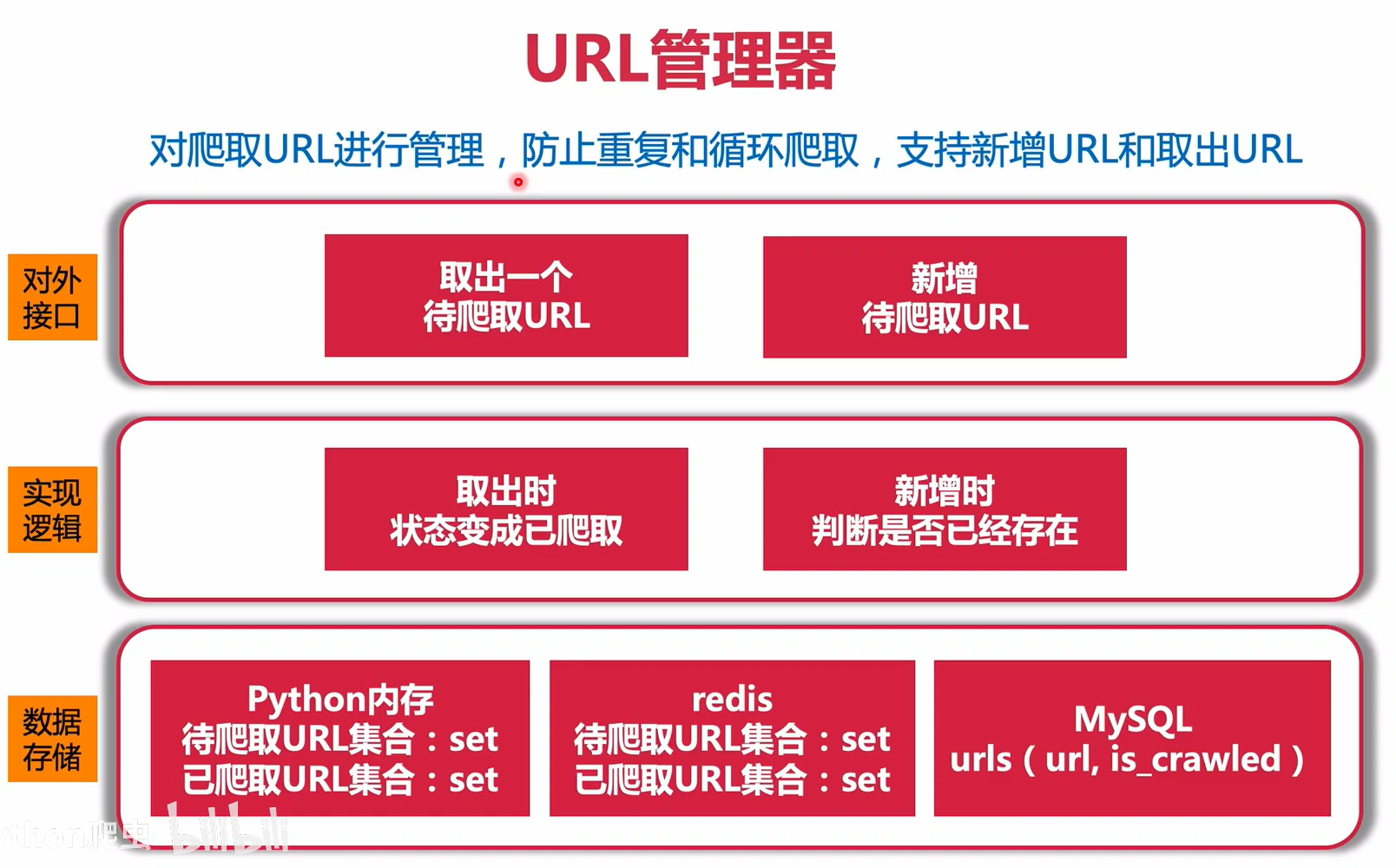
2.3 避免时间卡死

2.4 verify:true本地保存证书；设置false绕开验证

2.5 allow\_redirects:false避免非需求跳转



## 3. 管理器

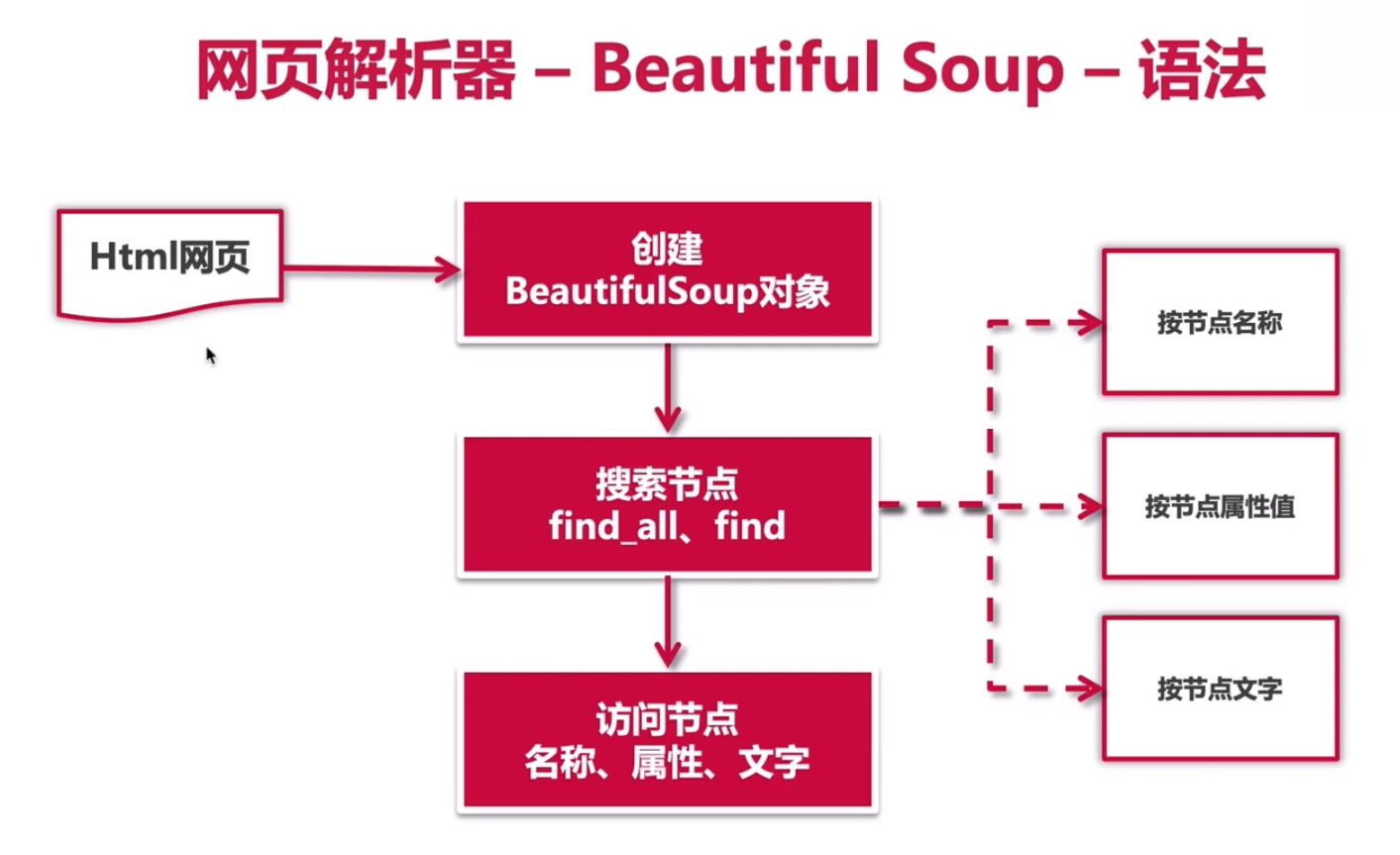


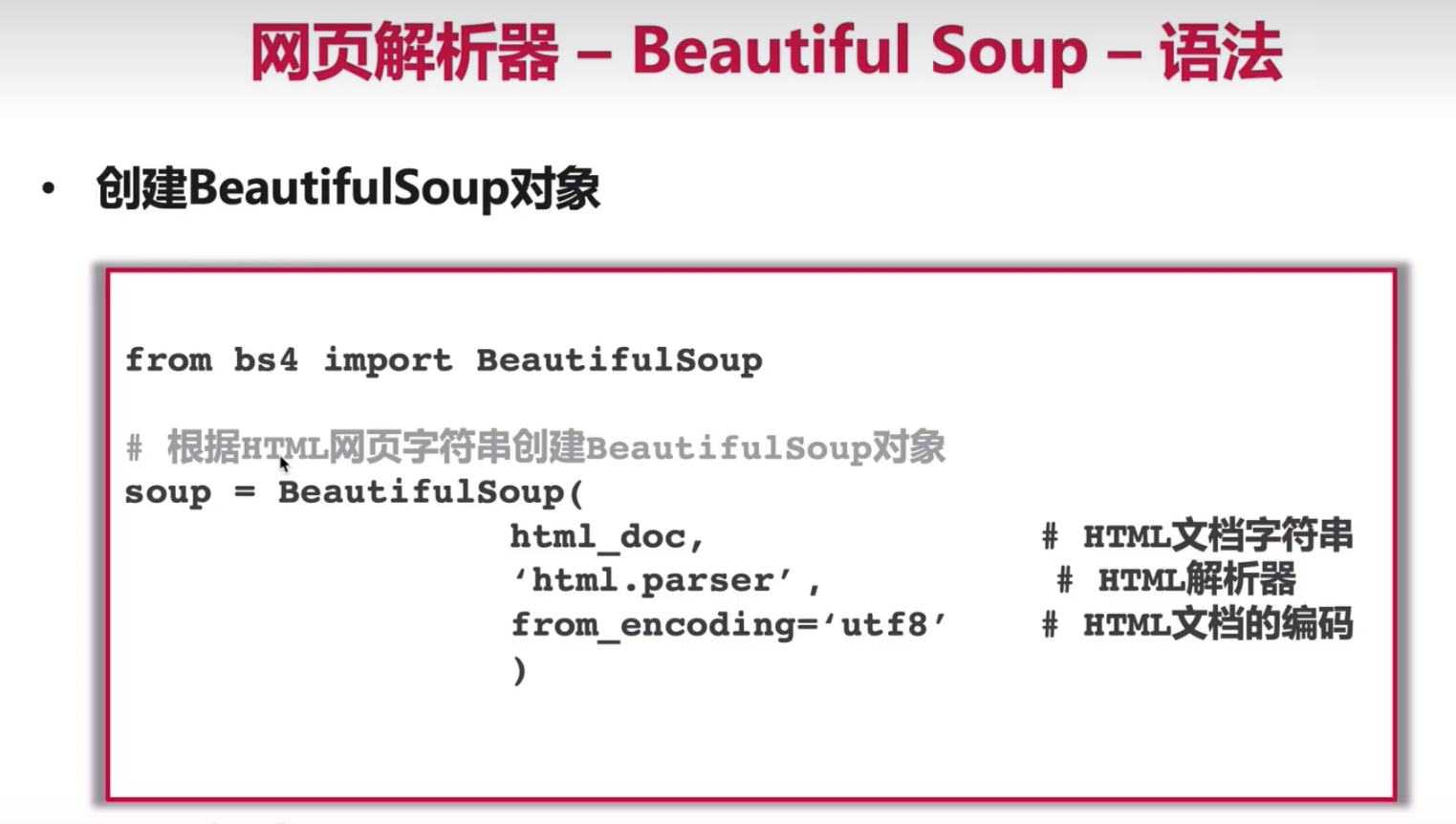
3.1 不同set存储不同状态的URL

## 4. HTML



## 5. 网页解析器——Beautiful Soup











## 6. 网站分析

6.1 content-type

6.2 Cookie

6.3 User-Agent

6.4 右键-展示源代码/鼠标移到-检查→未经过JS处理过的HTML文件

## 7. 实例

