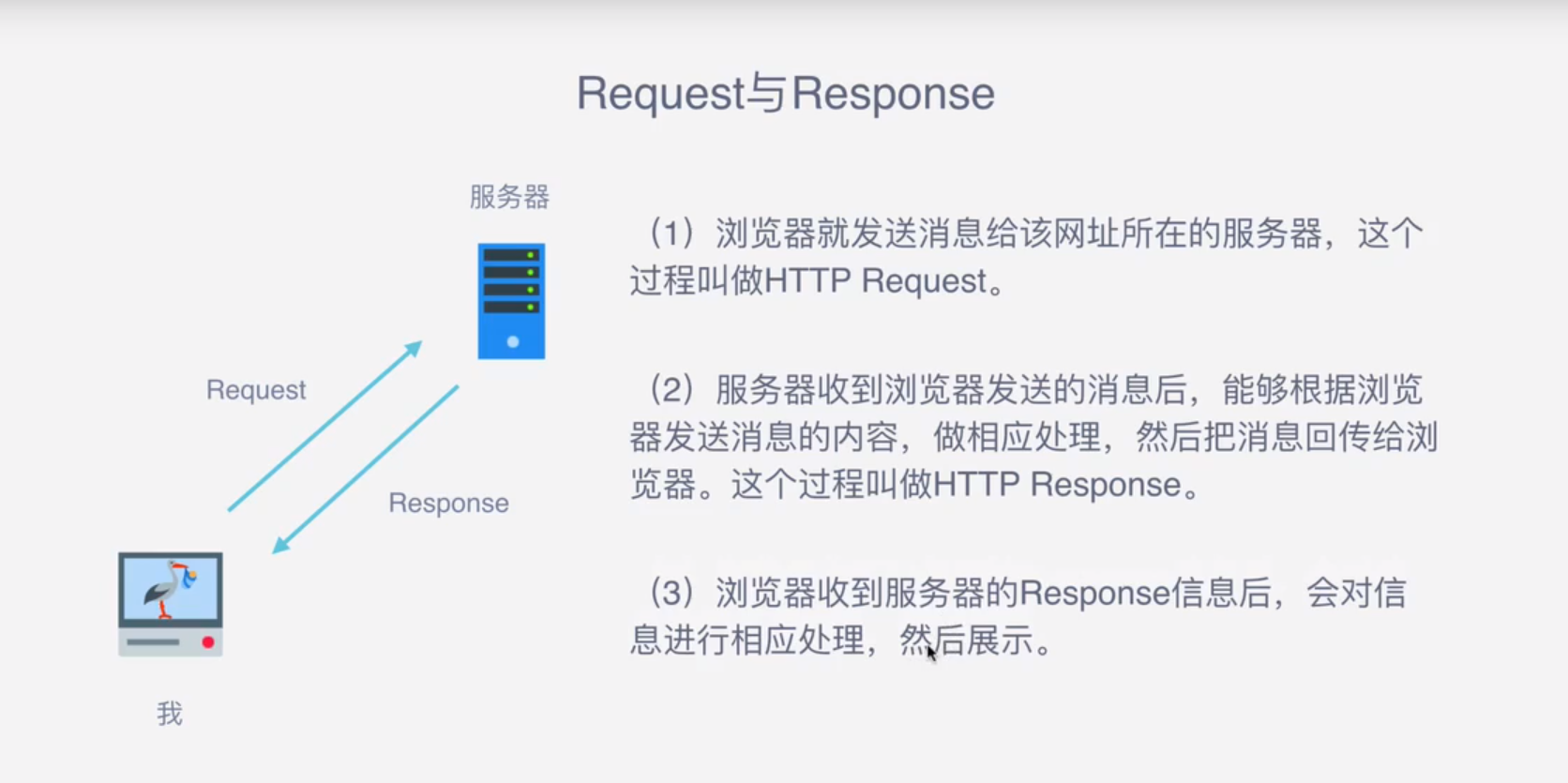
# 爬虫原理：

请求网站并提取数据的自动化程序

爬虫基本流程



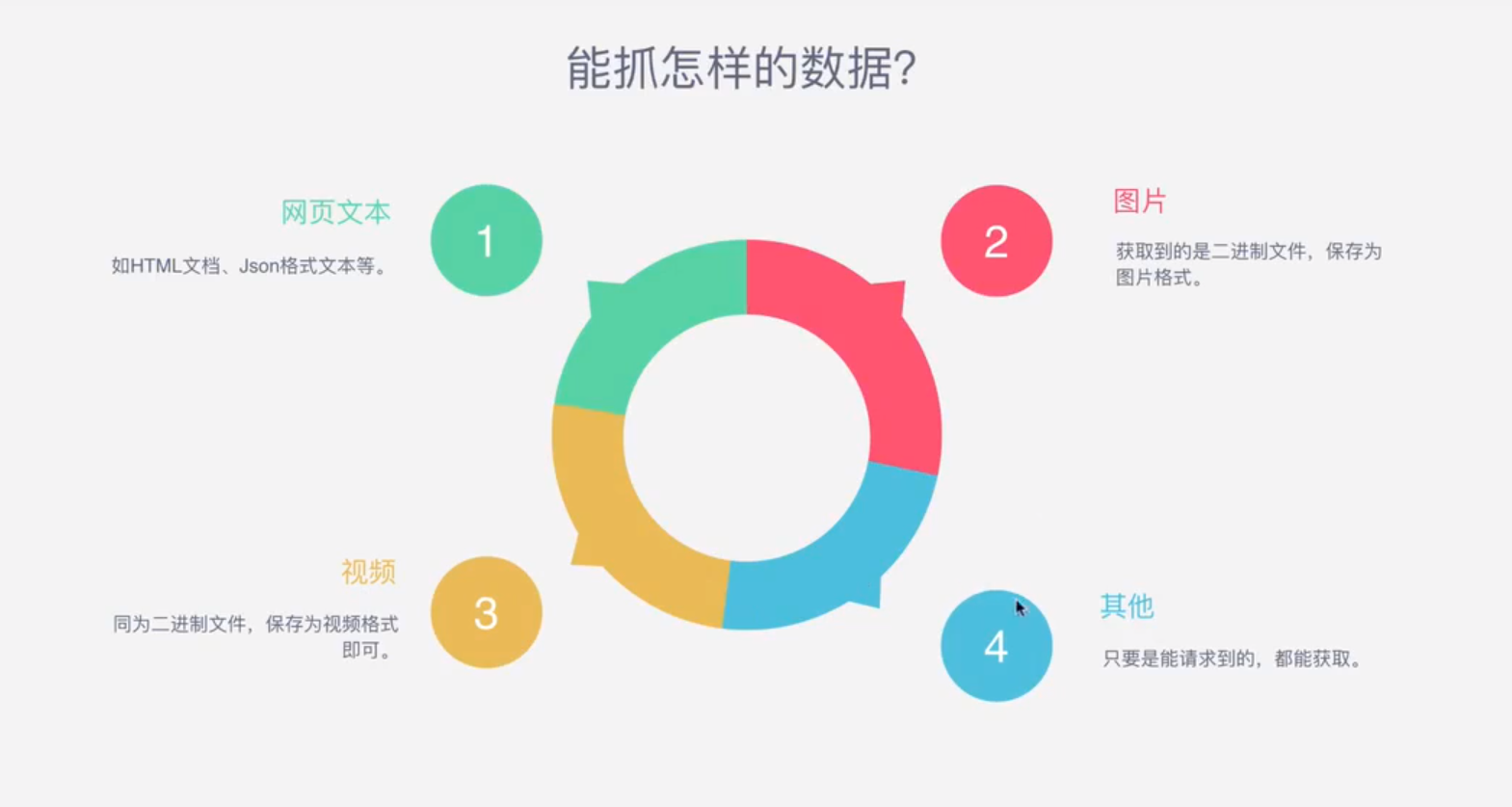
**Request与Response**



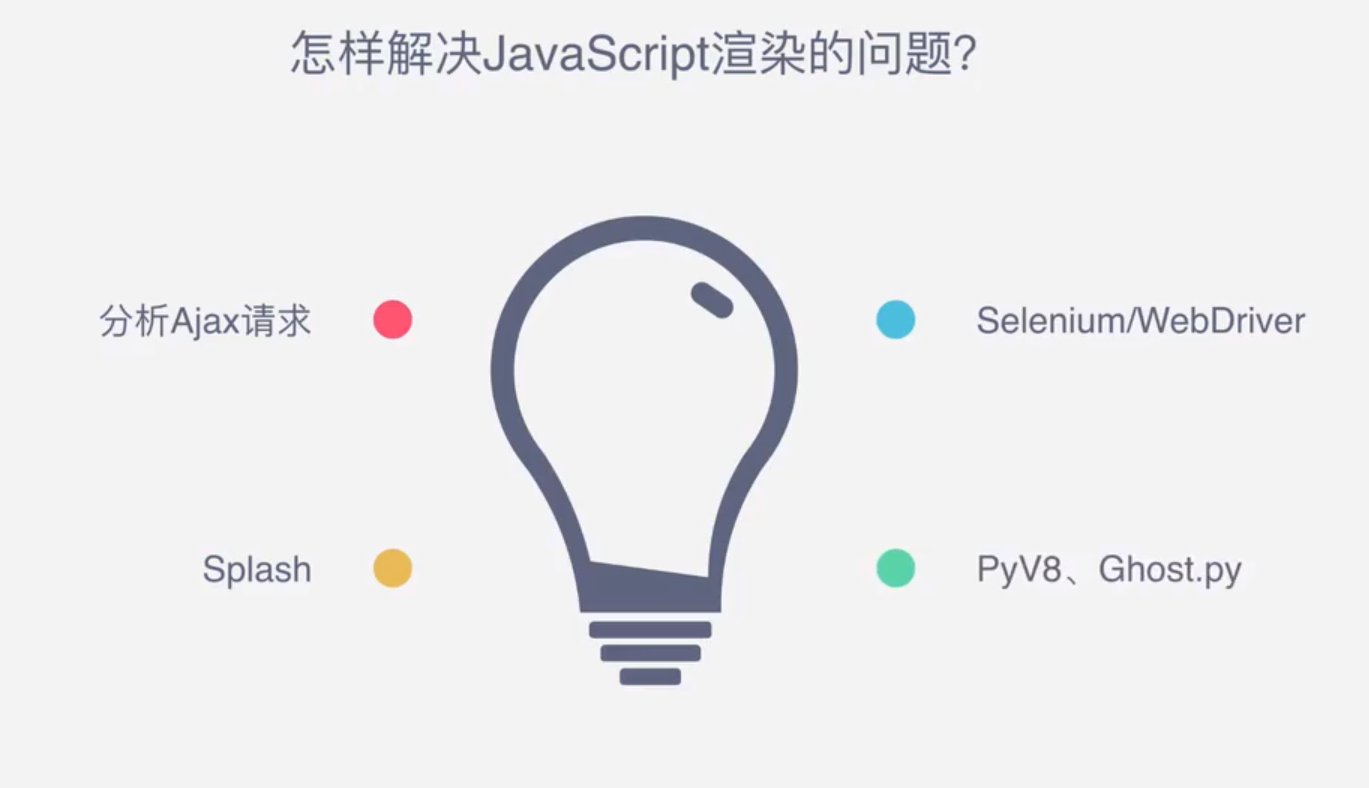


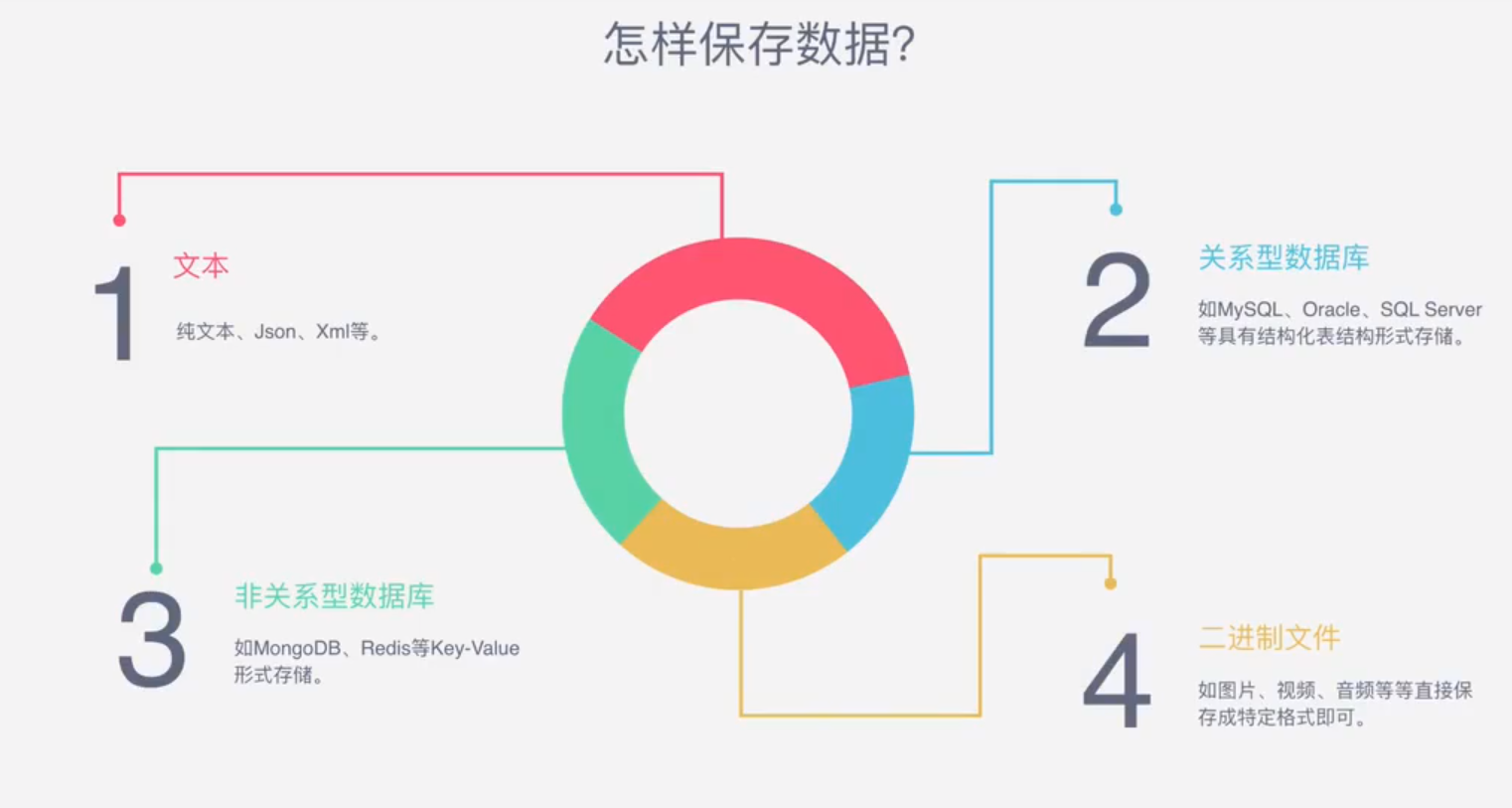












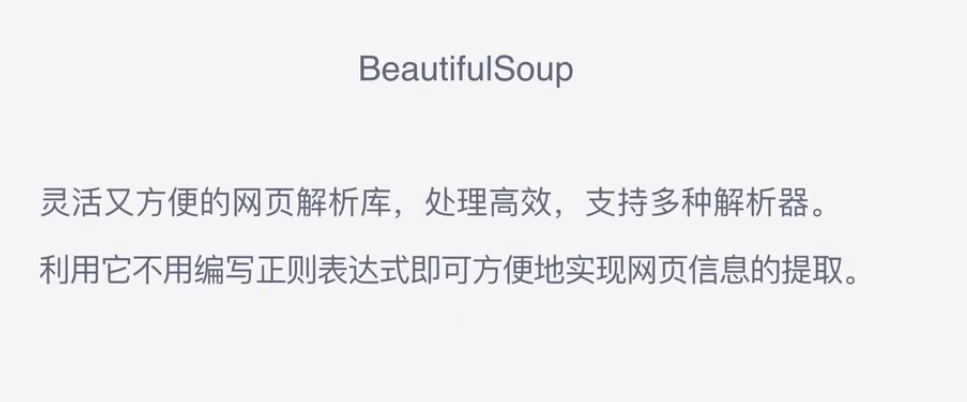
**HTTP库：urllib，httplib2,requests,treq等**

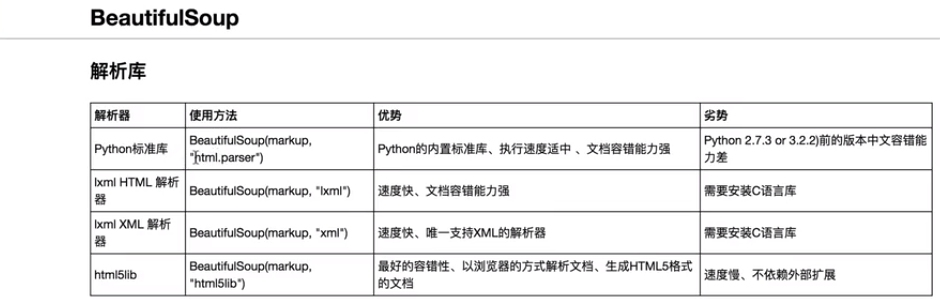
## Urllib:

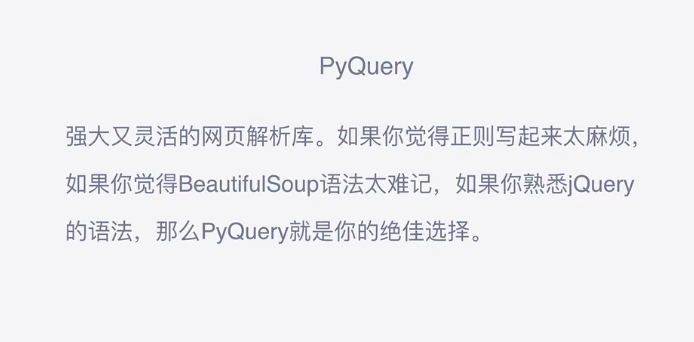
包含了四个模块：

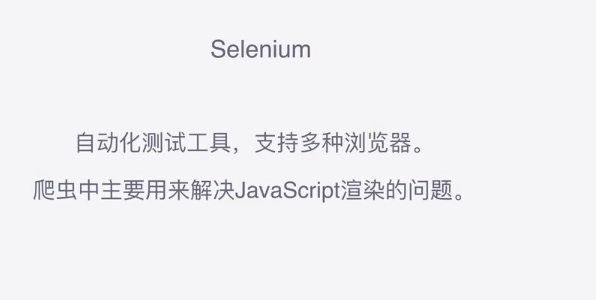
* request :最基本HTTP请求模块
* error：异常处理模块
* parse:一个工具模块，提供了URL处理方法
* robotparser:主要识别网站的robots.txt,看哪些网站可以爬











Selenium可以驱动浏览器执行特定的动作，同时可以获取浏览器当前呈现的页面的源代码，做到可爬可见。

from selenium import webdriver

from selenium.webdriver.common.by import By

from selenium.webdriver.common.keys import Keys

from selenium.webdriver.support import expected\_conditions as EC

from selenium.webdriver.support.wait import WebDriverWait

Splash

利用Splash，我们可以实现如下功能：

* 异步方式处理多个网页渲染过程；
* 获取渲染后的页面的源代码或截图；
* 通过关闭图片渲染或者使用Adblock规则来加快页面渲染速度；
* 可执行特定的JavaScript脚本；
* 可通过Lua脚本来控制页面渲染过程；
* 获取渲染的详细过程并通过HAR（HTTP Archive）格式呈现。