

# HTML

## 主讲：杨真

## Part 1 基础

- 环境搭建
- **HTML 基础**
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- POSTMAN 工具详解

## Part 2 爬虫

- 网站结构分析
- 抓取方案
- 多线程并行及排重
- 用 MySQL 信息存储

## Part 3 进阶

- 网站服务结构
- Cookie 及 登录处理
- 控制抓取的节奏
- 日志
- 守护进程

## Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

- HTML
- CSS
- Javascript
- DOM 树

# HTML

- HTML 是用来描述网页的一种语言
- HTML 指的是超文本标记语言 (Hyper Text Markup Language)
- HTML 不是一种编程语言，而是一种标记语言 (markup language)
- 标记语言是一套标记标签 (markup tag)
- HTML 使用标记标签来描述网页

- HTML 标签是由尖括号包围的关键词，比如 `<html>`
- HTML 标签通常是成对出现的，比如 `<b>` 和 `</b>`
- 标签对中的第一个标签是开始标签，第二个标签是结束标签
- 开始和结束标签也被称为开放标签和闭合标签

```
<html>
<head>
<meta charset="utf-8" />
</head>
<body>

<h1>我的第一个标题</h1>

<p>我的第一个段落。</p>

</body>
</html>
```

## HTML 链接

```
<a href="http://www.chinahadoop.cn">小象学院</a>
```

## HTML 标题

```
<h1>这是标题</h1>
```

## HTML 段落

```
<p>这是段落</p>
```

## HTML 图片

```

```

## 当前HTML文档的某个位置（成为锚点 anchor）

```
<a href="#title">小象学院</a>
```

这类链接只在当前页面内跳转，例如 #top 就是回到顶部

## 由脚本来处理的链接

```
<a href="javascript:void(0)">登录</a>
```

这类链接表示这是个死的链接，一般有 JavaScript 的脚本注册了对它的监听，所以由监听的响应函数来处理这个链接的点击

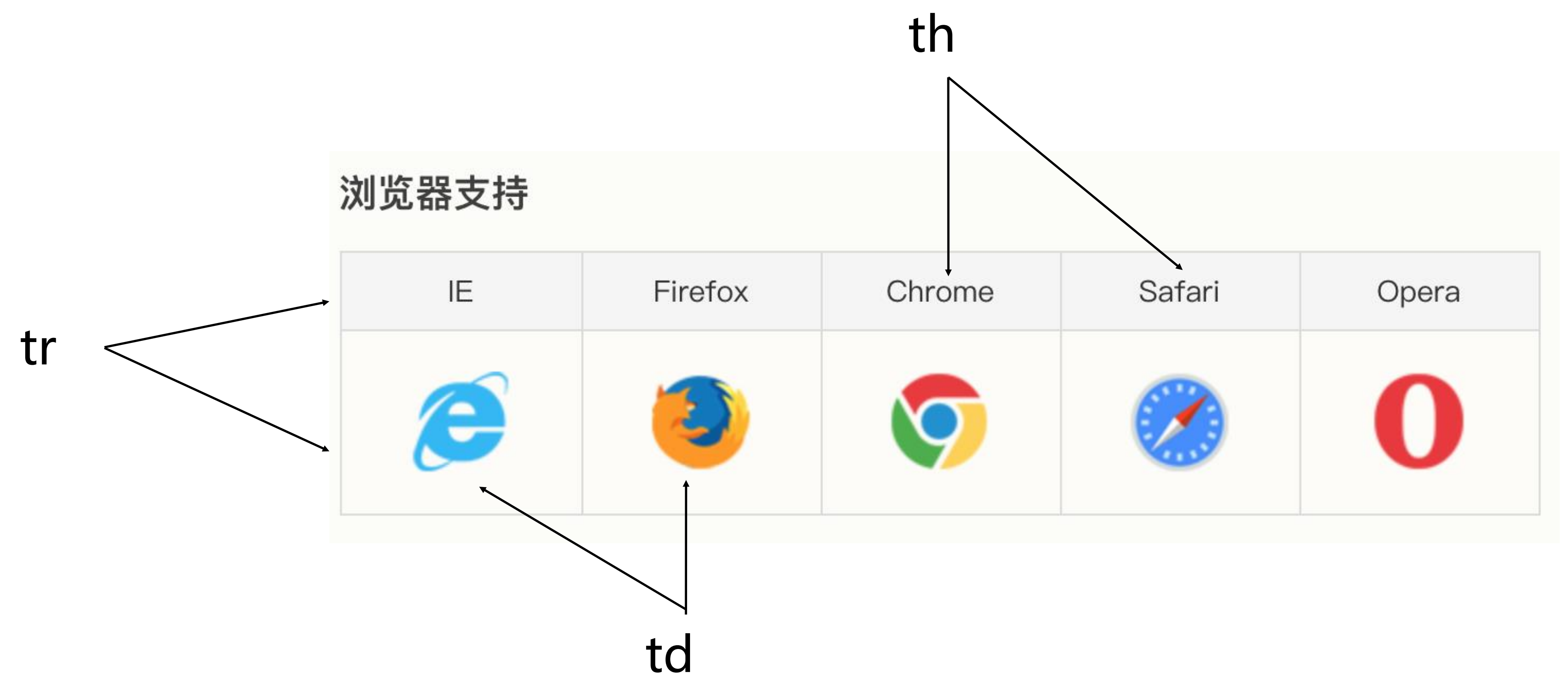


- table 定义表格标签
- tr 行
- th 表头 ( 第一行的单元格 )
- td 表的单元格 ( 表头之外的单元格 )

```
<table border="1">
<caption>Caption</caption>
<tr>
<th>Month</th>
<th>Savings</th>
</tr>
<tr>
<td>January</td>
<td>$100</td>
</tr>
</table>
```

浏览器支持

th				
IE	Firefox	Chrome	Safari	Opera
td	td			



The diagram illustrates the structure of the '浏览器支持' (Browser Support) table. It shows a table with two rows. The first row contains the browser names: IE, Firefox, Chrome, Safari, and Opera. The second row contains the corresponding browser logos. Annotations with arrows point to specific parts of the table: 'th' points to the header row, 'tr' points to the first row (the header row), and 'td' points to the first cell of the second row (containing the IE logo).

```
<a href="http://www.chinahadoop.cn">小象学院</a>
```

```
<h1 align="center">小象学院</h1>
```

属性是在标签里的键值对，对于爬虫来说，最常用的属性是 id，name 及 class

- id：为标签定义的唯一标识，比如用户名的输入框
- name：为标签定义的名字
- class：为标签加上类别属性，在选择某个类型的标签的时候可以利用class的名字一次选中所有这个class的标签，主要用在样式控制上

CSS ( Cascading Style Sheet ) 可译为“层叠样式表”或“级联样式表”，它定义如何显示 HTML 元素，用于控制Web页面的外观，使用方式：

- 可以存在于 HTML 的标签里 `<p style="color:blue;margin-left:20px;">小象学院</p>`
- HTML里用专门的区块来定义 `<style></style>`
- 以独立的 .css 文件存在。在<head> 里定义要引用的 css 文件 `<link rel="stylesheet" type="text/css" href="html_class.css">`

只为遇见明天更优秀的你