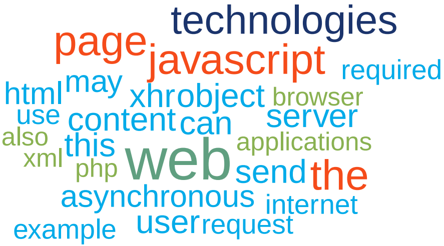
Introduction to Web Scraping with R

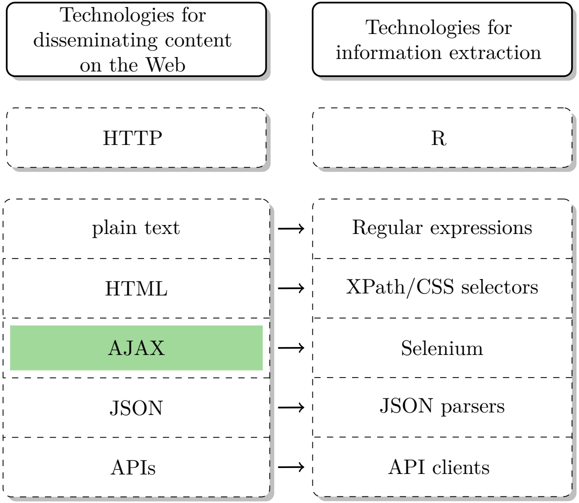
AJAX Technologies



Simon Munzert | IPSDS

AJAX

Technologies of the World Wide Web



Technologies for disseminating content on the Web

Technologies for information extraction

HTTP

R

plain text → Regular expressions

HTML → XPath/CSS selectors

AJAX → Selenium

JSON → JSON parsers

APIs → API clients

What's AJAX?

Purpose

- recall: HTML/HTTP, which are used for static display of content, lack
  1. mechanisms to detect user behavior in the browser (and not only on the server)
  2. a scripting engine that reacts on this behavior
  3. a mechanism for asynchronous queries
- **A**synchronous **J**avaScript **a**nd **X**ML is a set of technologies that serve these purposes

What's AJAX?

Purpose

- recall: HTML/HTTP, which are used for static display of content, lack
  1. mechanisms to detect user behavior in the browser (and not only on the server)
  2. a scripting engine that reacts on this behavior
  3. a mechanism for asynchronous queries
- **A**synchronous **J**avaScript **a**nd **X**ML is a set of technologies that serve these purposes

Components

- HTML/CSS for presentation
- Document Object Model (DOM) for interaction with data ("tree structure")
- JSON/XML for data interchange
- XMLHttpRequest for asynchronous communication
- JavaScript as a scripting language

What's AJAX?

Purpose

- recall: HTML/HTTP, which are used for static display of content, lack
  1. mechanisms to detect user behavior in the browser (and not only on the server)
  2. a scripting engine that reacts on this behavior
  3. a mechanism for asynchronous queries
- **A**synchronous **J**avaScript **a**nd **X**ML is a set of technologies that serve these purposes

Components

- HTML/CSS for presentation
- Document Object Model (DOM) for interaction with data ("tree structure")
- JSON/XML for data interchange
- XMLHttpRequest for asynchronous communication
- **JavaScript** as a scripting language

JavaScript

JavaScript

What's JavaScript?

- Programming language that connects well to web technologies
- W3C web standard
- native browser support, who have a built-in JavaScript engine
- nowadays employed on the majority of websites
- extensible by libraries, e.g. *jQuery* library for DOM manipulation

JavaScript

What's JavaScript?

- Programming language that connects well to web technologies
- W3C web standard
- native browser support, who have a built-in JavaScript engine
- nowadays employed on the majority of websites
- extensible by libraries, e.g. *jQuery* library for DOM manipulation

What's JavaScript?

- core technology to make webpages interactive ("Dynamic HTML")
- allows to...
  - animate page elements
  - offer interactive content (games, video, ...)
  - manipulate page content and communication with the server without reloading the page

JavaScript on the Web

How's JavaScript code embedded in HTML?

- between `<script>` tags
- as an external reference in the `src` attribute of a `<script>` element
- directly in certain HTML attributes ('event handler')

JavaScript on the Web

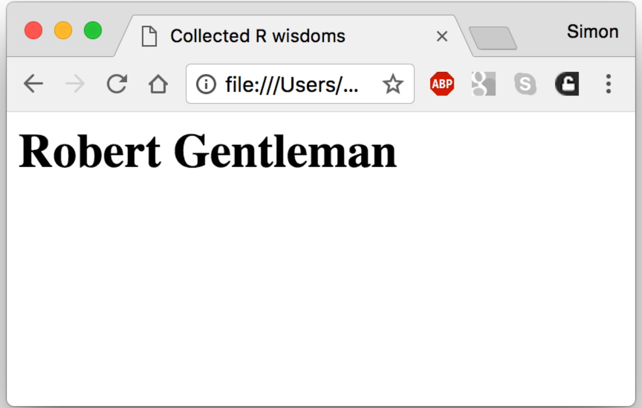
DOM manipulation with JavaScript

- adding/removing HTML elements
- changing attributes
- modification of CSS styles
- . . .

Example:

```
<script type="text/javascript" src="jquery-1.8.0.min.js"></script>
<script type="text/javascript" src="script1.js"></script>
```

JavaScript on the Web



Robert Gentleman

JavaScript on the Web

```html
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>

<script type="text/javascript" src="jquery-1.8.0.min.js"></script>
<script type="text/javascript" src="script1.js"></script>

<head>
<title>Collected R wisdoms</title>
</head>

<body>
<div id="R Inventor" lang="english" date="June/2003">
  <h1>Robert Gentleman</h1>
  <p><i>'What we have is nice, but we need something very different'</i></p>
  <p><b>Source: </b>Statistical Computing 2003, Reisensburg</p>
</div>
</body>
</html>
```

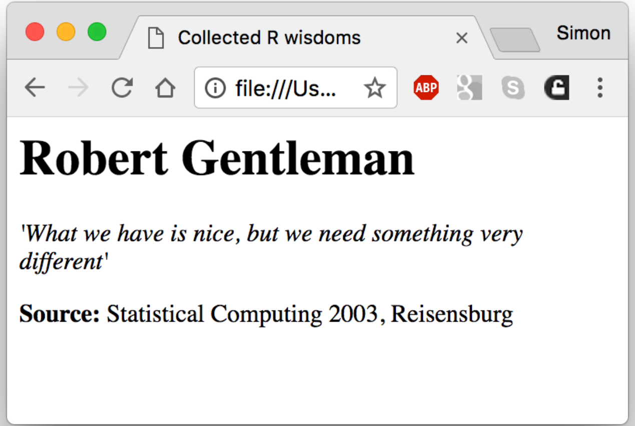
JavaScript on the Web

A JavaScript code snippet

```
$(document).ready(function() {
    $("p").hide();
    $("h1").click(function(){
        $(this).nextAll().slideToggle(300);
    });
});
```

- `$()` operator: addresses DOM elements
- `ready()`: JavaScript execution starts when the complete DOM is ready, i.e. fetched from the server
- `hide()`: element is hidden at first place
- `click` event: identifies mouse click and executes a certain action
- `nextAll()`: all subsequent elements in the DOM are addressed
- `slideToggle()`: Toggle effect, 300 milli-seconds

JavaScript on the Web



Robert Gentleman

'What we have is nice, but we need something very different'

Source: Statistical Computing 2003, Reisensburg

Summary

Summary

- AJAX technologies allow for a dynamic manipulation of the HTML tree
- what the user sees in the browser is not necessarily what's in the static HTML source code
- elements can be added, removed, or changed
- the "live" DOM tree that you saw in the Web Developer Tools takes track of such changes
- but you cannot simply access this live HTML with R and `rvest`
- we need another technology that helps us mimic a session in the browser to render all dynamic content