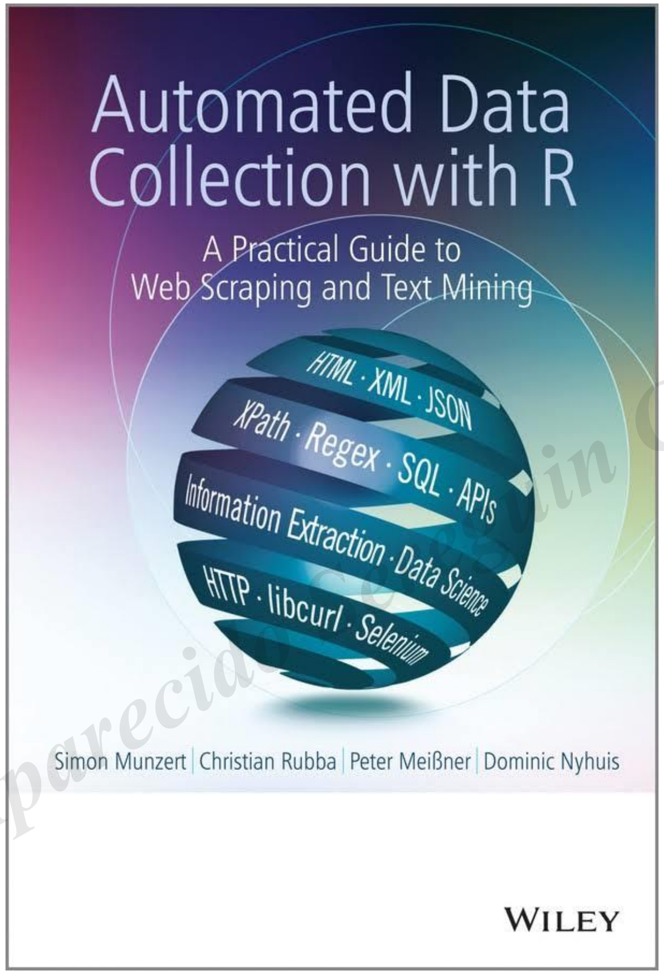
MBA USP ESALQ

DATA COLLECTION: CRAWLERS AND WEB SCRAPING

Prof. Dr. Jeronymo Marcondes

Introduction


Automated Data Collection with R

A Practical Guide to Web Scraping and Text Mining

HTML · XML · JSON

XPath · Regex · SQL · APIs

Information Extraction · Data Science

HTTP · libcurl · Selenium

Simon Munzert | Christian Rubba | Peter Meißner | Dominic Nyhuis

WILEY

Introduction

- Plan of attack:

1. Basic understanding of how the web works.
2. Basic understanding of HTML.
3. Basic understanding of XML and JSON.
4. Web Scraping and the course focus.
5. Using R and packages to perform the operation.

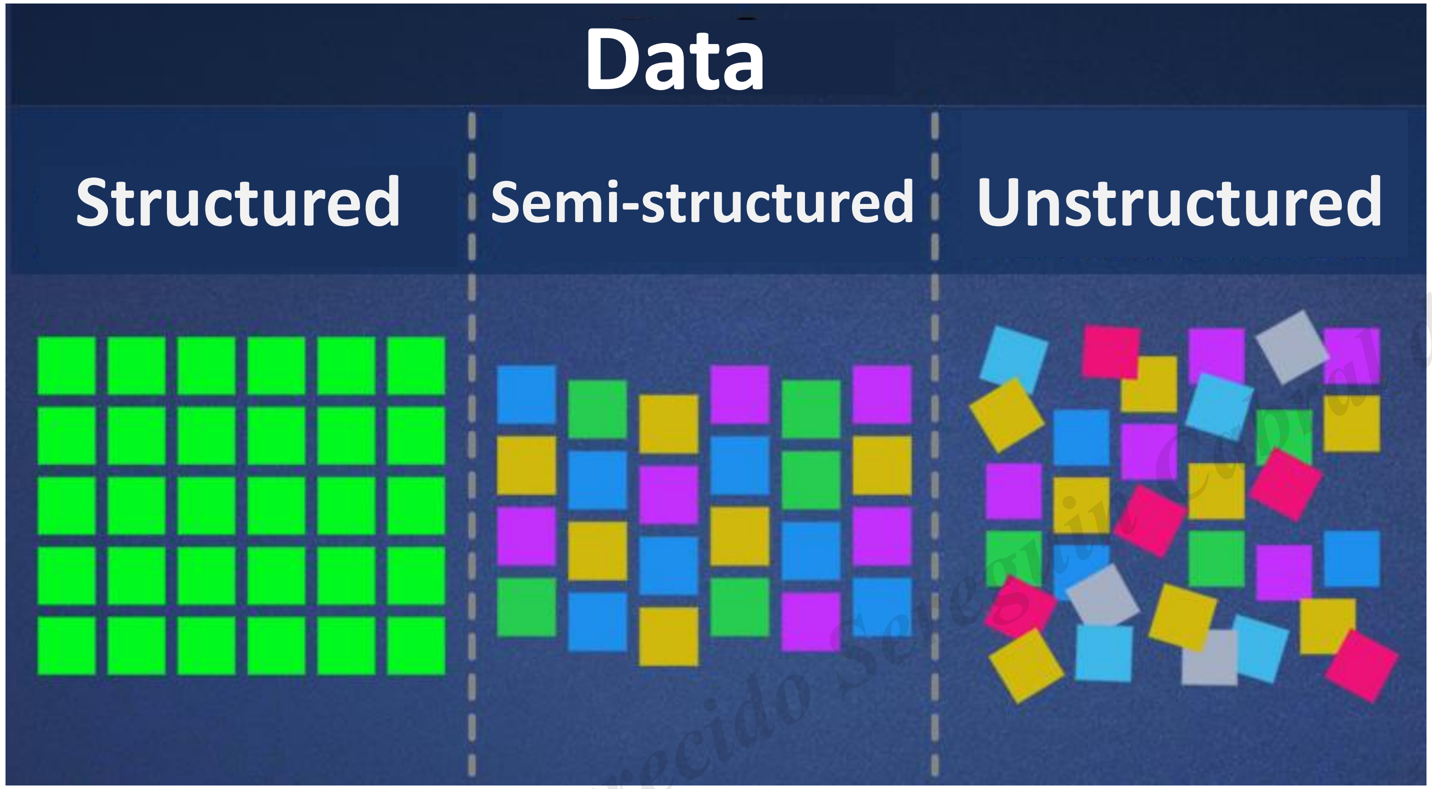
Need for Data

- Several important data = spread on the web
- Most common places: webpages and APIs.
- Other possibilities: FTP Servers, for example.
- How to recover web data and treat the information in order to use it in models and analysis?

Data Structures

The data we can use are divided into:

- Structured Data.
- Semi-structured Data.
- Unstructured data.



Source: https://universidadedatecnologia.com.br/

- Structured - data that have well-defined formats, such as those extracted from spreadsheets or relate databases in SQL.
- Semi-structured – Similar to structured data, but not obedient in the totality regarding the format. In this line are the records of languages based on HTML and XML.
- Unstructured or NoSQL - do not have a specific format, these are data collected in their original form, such as a text, a video, an email fragment, a system log or a simple photo.

Data Structures

- How to use semi-structured or unstructured information that is on the web, collect it and store it?

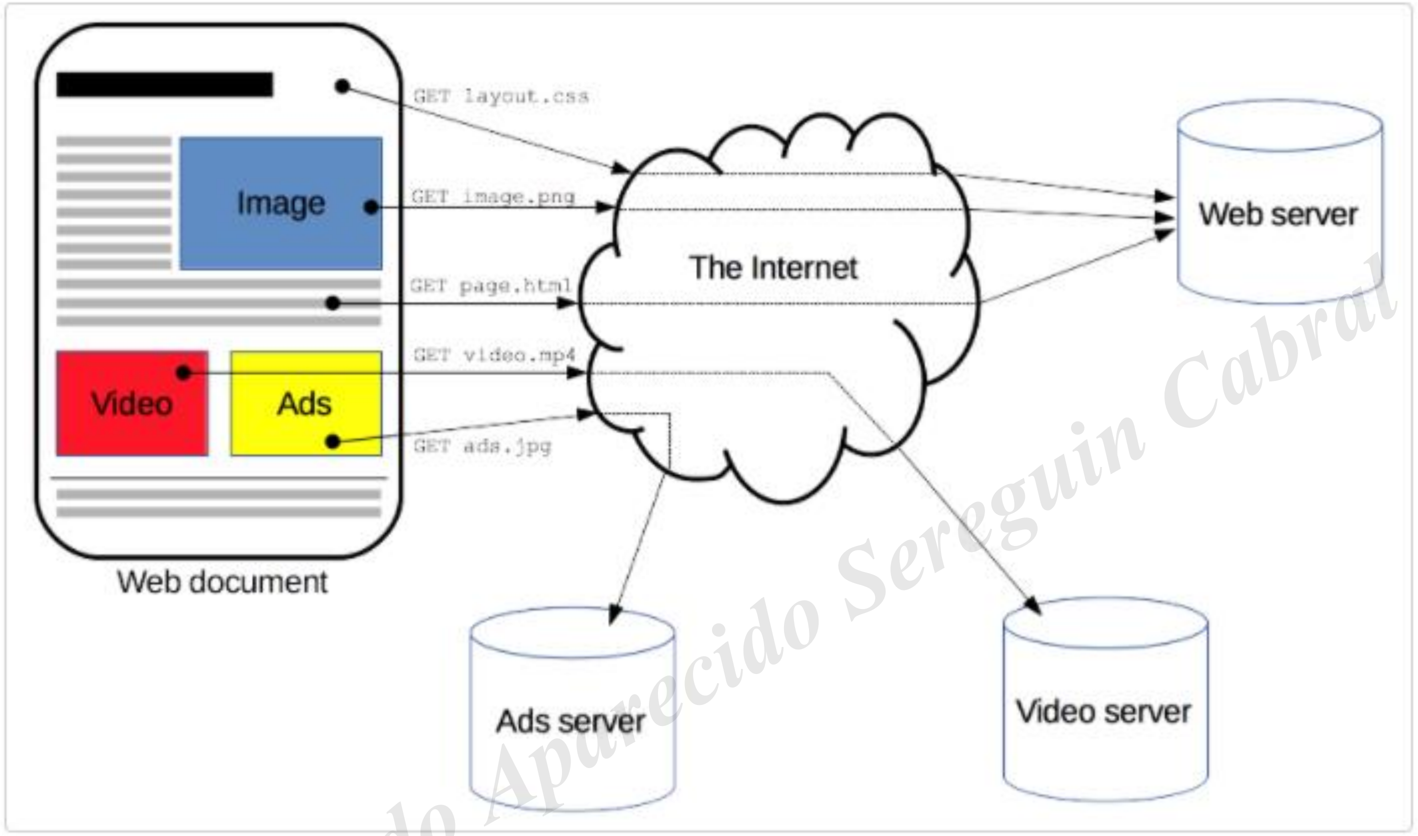
- WebScraping: The process of extraction and combination of Web interest content, in a systematic way. In such process, a software agent, also known as robot, imitates the interaction of navigation between Web servers and humans. Step by step, the robot accesses the websites as necessary, analyzes its content to find and extract data of interest and structure these contents, as desired (LOURENÇO, 2013).

Introduction to the web

- HTTP

HTTP is a protocol which allows to obtain resources, such as HTML documents. It is the basis of any data exchange on the Web, and a customer-server protocol, which means that requests are started by the recipient, usually a Web browser.

Introduction to the web



Source: https://developer.mozilla.org/pt-BR/docs/Web/HTTP/Overview

Introduction to the web

Customers and servers communicate exchanging individual messages. Messages sent by the customer, usually a Web-browser**,** are called **requests**, and the messages sent by the server as a answer are called **responses**.

Request methods

- Set of requests types that indicate what to do
- HTTP verbs
- We are not going to study all of them

Request methods

GET

The GET method asks the representation of a specific resource. Requests using the GET method must only return data.

POST

The POST method is used to submit an entity to a specific resource, often causing a change in the resource state, or side effects on the server.

Request methods

DELETE

The DELETE method removes a specific resource.

CONNECT

The CONNECT method establishes a tunnel for the server identified by the destination resource.

HTML

- HTML
- HyperText Markup Language
- Positional language

HTML

http://www.r-datacollection.com/materials/html/OurFirstHTML.html

HTML

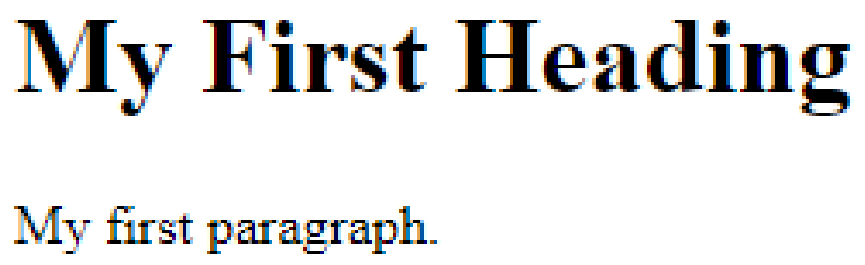
- More information than what is shown
- Labels that show where the content goes to
- Explain the document structure – without a layout

https://www.w3schools.com/html/html_basic.asp

Basic HTML

Simple example:

```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>
</body>
</html>
```



My First Heading

My first paragraph.

HTML - elements

- An element begins with a start tag <“tag name”> and finishes with an end tag </”tag name>

```
<tagname>Content goes here...</tagname>
```

- Nested tags - previous example

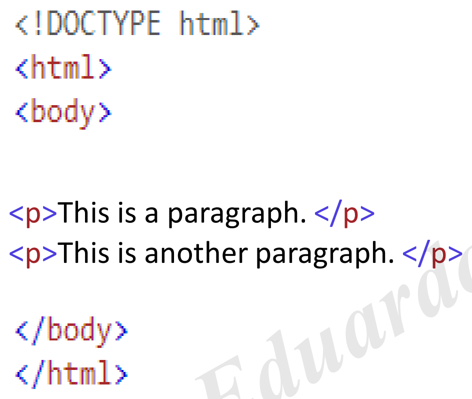
HTML

- Doctype
- <html>and</html>
- <body>and</body> - all elements in a page, such as headings, paragraphs, images, hyperlinks, etc

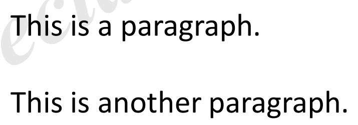
HTML

- Heading (h1, h2, h3, h4, h5 and h6)

- Paragraphs <p>and</p>

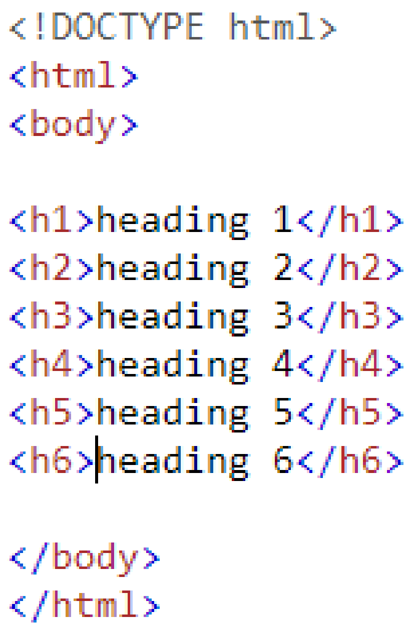


```
<!DOCTYPE html>
<html>
<body>

<p>This is a paragraph. </p>
<p>This is another paragraph. </p>

</body>
</html>
```

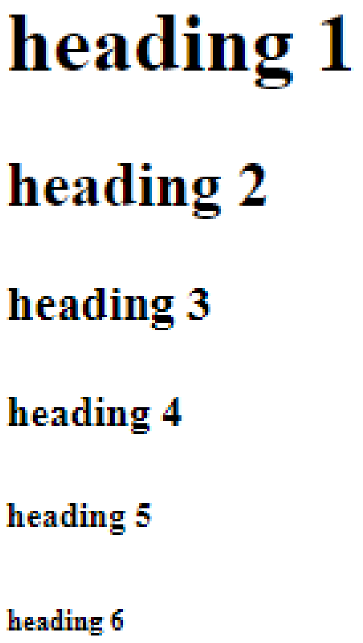


This is a paragraph.

This is another paragraph.



```
<!DOCTYPE html>
<html>
<body>

<h1>heading 1</h1>
<h2>heading 2</h2>
<h3>heading 3</h3>
<h4>heading 4</h4>
<h5>heading 5</h5>
<h6>heading 6</h6>

</body>
</html>
```



heading 1

heading 2

heading 3

heading 4

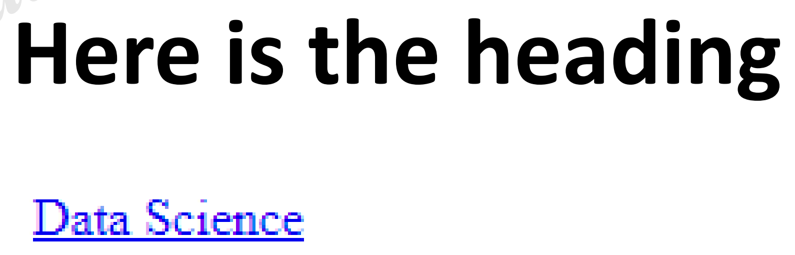
heading 5

heading 6

HTML

- Hiperlink: <a>and</a> :

```
<!DOCTYPE html>
<html>
<body>

<h1>Here is the heading</h1>

<a href="https://mbauspesalq.com/">Data Science</a>

</body>
</html>
```

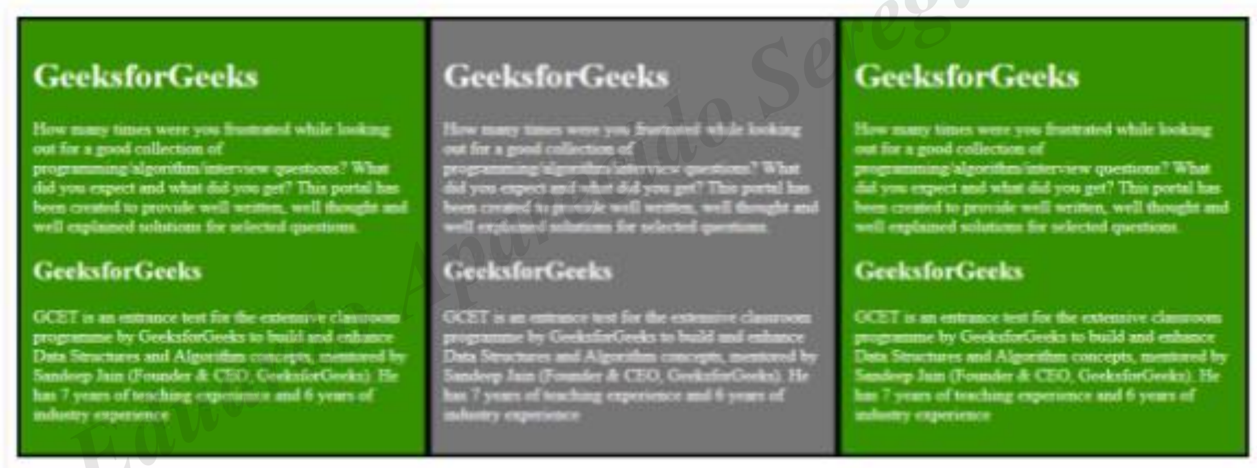


Here is the heading

[Data Science](https://mbauspesalq.com/)

HTML

- div: <div>and</div>: it defines a division or section of a HTML page
- Container for a set of HTML tags – specific characteristics of a specific part of the page.



GeeksforGeeks

How many times were you frustrated while looking out for a good collection of programming/algorithm/interview questions? What did you expect and what did you get? This portal has been created to provide well written, well thought and well explained solutions for selected questions.

GeeksforGeeks

GCET is an entrance test for the extensive classroom programme by GeeksforGeeks to build and enhance Data Structures and Algorithm concepts, mentored by Sandeep Jain (Founder & CEO, GeeksforGeeks). He has 7 years of teaching experience and 6 years of industry experience

GeeksforGeeks

How many times were you frustrated while looking out for a good collection of programming/algorithm/interview questions? What did you expect and what did you get? This portal has been created to provide well written, well thought and well explained solutions for selected questions.

GeeksforGeeks

GCET is an entrance test for the extensive classroom programme by GeeksforGeeks to build and enhance Data Structures and Algorithm concepts, mentored by Sandeep Jain (Founder & CEO, GeeksforGeeks). He has 7 years of teaching experience and 6 years of industry experience

GeeksforGeeks

How many times were you frustrated while looking out for a good collection of programming/algorithm/interview questions? What did you expect and what did you get? This portal has been created to provide well written, well thought and well explained solutions for selected questions.

GeeksforGeeks

GCET is an entrance test for the extensive classroom programme by GeeksforGeeks to build and enhance Data Structures and Algorithm concepts, mentored by Sandeep Jain (Founder & CEO, GeeksforGeeks). He has 7 years of teaching experience and 6 years of industry experience

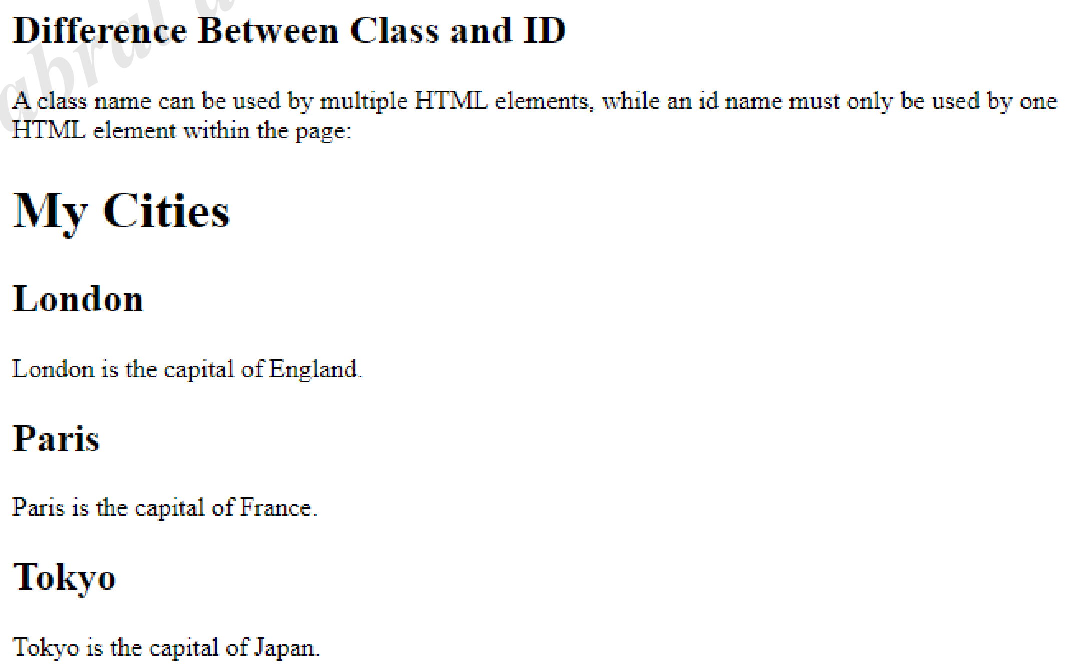
Source: https://www.geeksforgeeks.org/

HTML

- Attributes: additional information on elements
- Id and Class

HTML

```html
<!DOCTYPE html>
<html>
<body>

<h2>Difference Between Class and ID</h2>
<p>A class name can be used by multiple HTML elements, while an id name must
only be used by one HTML element within the page:</p>

<!-- An element with a unique id -->
<h1 id="myHeader">My Cities</h1>

<!-- Multiple elements with same class -->
<h2 class="city">London</h2>
<p>London is the capital of England.</p>

<h2 class="city">Paris</h2>
<p>Paris is the capital of France.</p>

<h2 class="city">Tokyo</h2>
<p>Tokyo is the capital of Japan.</p>

</body>
</html>
```



Difference Between Class and ID

A class name can be used by multiple HTML elements, while an id name must only be used by one HTML element within the page:

My Cities

London

London is the capital of England.

Paris

Paris is the capital of France.

Tokyo

Tokyo is the capital of Japan.

API

- New ways to share data on the web
- Increasing use of APIs: Application Programming Interface
- Set of definitions, protocols, and rules to create integrations between systems.

API

- How it works:

1. Costumer application send request for API
2. API verifies if the request is valid
3. If it is valid, it processes what is necessary for response
4. The API transfers the data to customer application

- Example: IBGE (The Brazilian Institute of Geography and Statistics)

API

- REST Architecture

- Same verbs of HTTP – uses their characteristics, for example:

1. Customer-server protocol
2. HTTP Messages

API

- Advantages for companies – microservices
- Data return: most of the time JSON
- However, return type can be varied

XML

- One of the most popular ways of crossing data on the web
- Same markup basis of HTML, but instead of showing information, the goal is to store data.
- Data stored in a pure text – any browser or operational system is able to understand it
- Self-descriptive Tags

XML

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<bond_movies>
  <movie id="1">
    <name>Dr. No</name>
    <year>1962</year>
    <actors bond="Sean Connery" villain="Joseph Wiseman"/>
    <budget>1.1M</budget>
    <boxoffice>59.5M</boxoffice>
  </movie>
  <movie id="2">
    <name>Live and Let Die</name>
    <year>1973</year>
    <actors bond="Roger Moore" villain="Yaphet Kotto"/>
    <budget>7M</budget>
    <boxoffice>126.4M</boxoffice>
  </movie>
  <movie id="3">
    <name>Skyfall</name>
    <year>2012</year>
    <actors bond="Daniel Craig" villain="Javier Bardem"/>
    <budget>175M</budget>
    <boxoffice>1108.6M</boxoffice>
  </movie>
</bond_movies>
```

XML

- Characteristics:

1. XML Version
2. Start and end Tags
3. Elements
4. Attributes
5. Data
6. Tree structure - us parents and children

XML

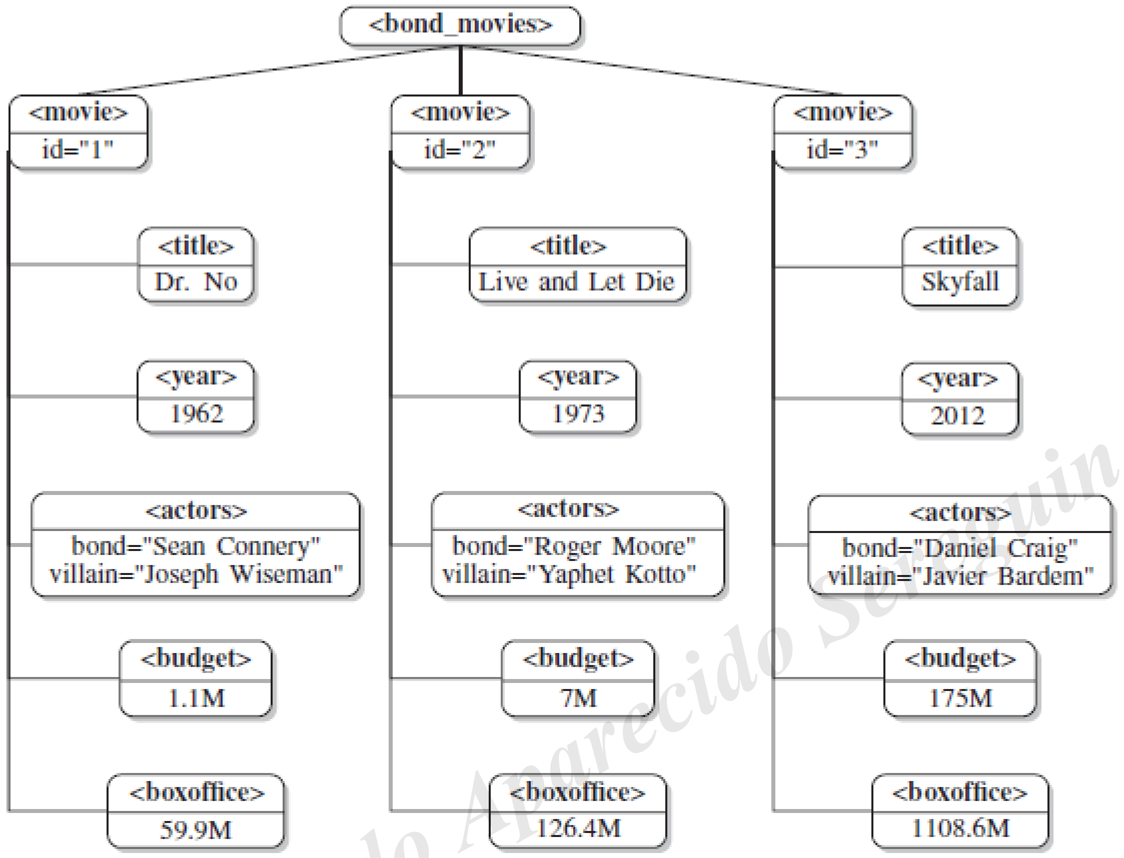
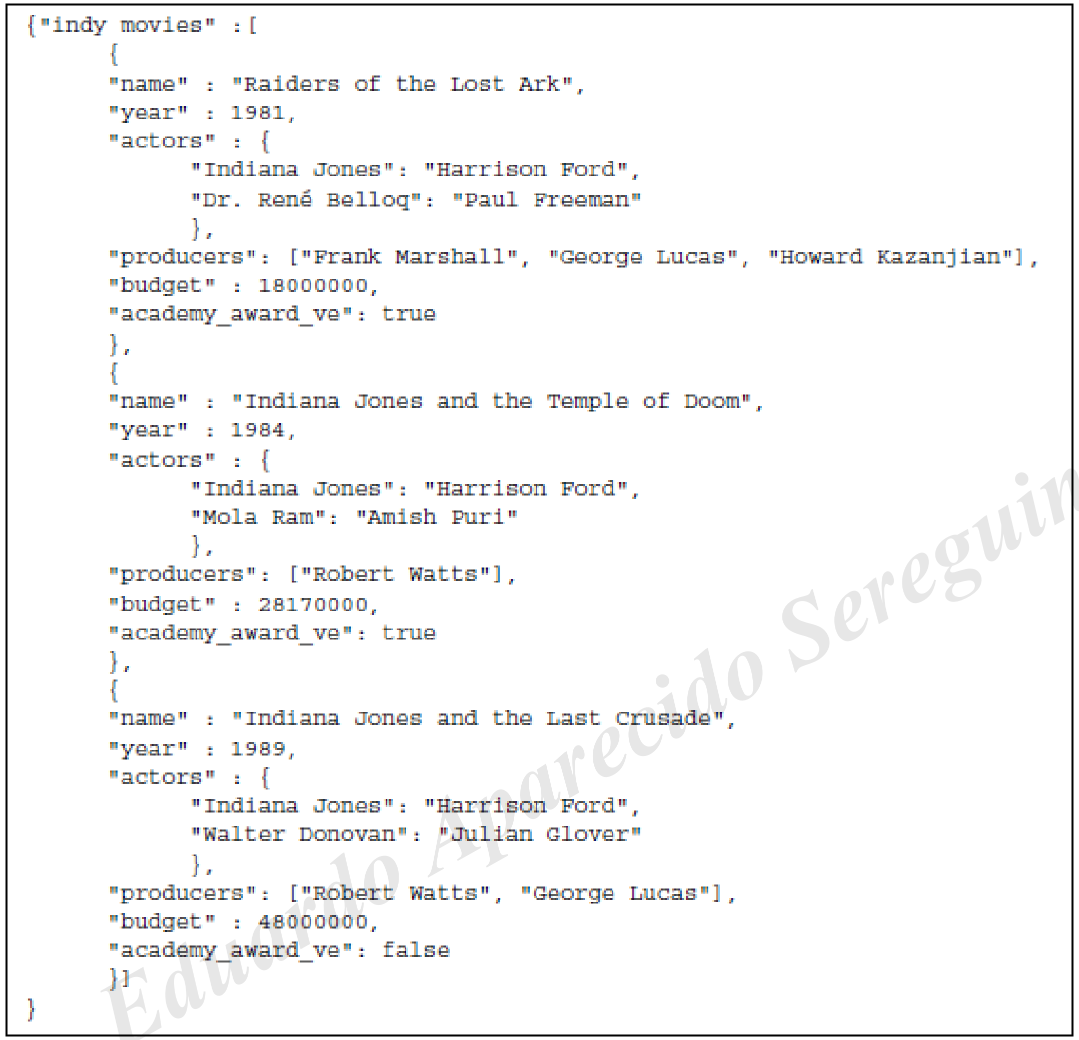


Figure 3.2 Tree perspective on an XML document

JSON



```
{"indy movies" :[
      {
      "name" : "Raiders of the Lost Ark",
      "year" : 1981,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Dr. René Belloq": "Paul Freeman"
            },
      "producers": ["Frank Marshall", "George Lucas", "Howard Kazanjian"],
      "budget" : 18000000,
      "academy_award_ve": true
      },
      {
      "name" : "Indiana Jones and the Temple of Doom",
      "year" : 1984,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Mola Ram": "Amish Puri"
            },
      "producers": ["Robert Watts"],
      "budget" : 28170000,
      "academy_award_ve": true
      },
      {
      "name" : "Indiana Jones and the Last Crusade",
      "year" : 1989,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Walter Donovan": "Julian Glover"
            },
      "producers": ["Robert Watts", "George Lucas"],
      "budget" : 48000000,
      "academy_award_ve": false
      }]
}
```

JSON

- Characteristics:

1. Based on objects (similar to XML elements)
2. Based on the division of key\value pairs
3. Keys are under double quotes
4. Values are preceded by two points
5. Values will be under double quotes if they are text fields
6. Array are defined by brackets

Web Scraping

- How to search for information?
- Packages that execute the verbs on HTTP, such as GET
- Httr and rvest packages

Web Scraping

- Return as text
- PARSE
- Searches the information in a given point

Web Scraping

Discussion: is web scraping illegal?