

Leonard Traeger M. Sc. Information Systems leonard.traeger@fh-dortmund.de

#### Disclaimer

This lecture part is designed to give you a rough understanding of data in the wild.

If your group aims to gather data via Web-Scraping – this is a good start.

- This lesson is partially based on the Library Carpentry https://librarycarpentry.org/lc-spreadsheets/
- Slides by Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

### Learning Goals Data Formats

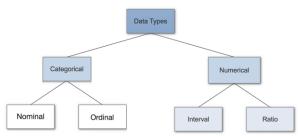
- Give a definition on data.
- **Explain** existing formats to structure data and their core purpose. **Name** advantages and disadvantages.
- Identify data formats given CSV, JSON, and XML samples.
- Describe XPath and how it can be used in Web-Scraping.

### What is data?











#### Data

Even though we can consider a multitude of things as "data": digits, sound, images, text,...

 Common standards exist to store these and have proven themselves in domains.

As data scientists we should be able to

- Transform raw data into a machine readable and optimized representation format.
- Recommend different data format standards.
- Map different data formats: because our analytical application should abstract any standard.

#### data

noun [U, + sing/pl verb]

UK ◀》 /'deɪ.tə/ US ◀》 /'deɪ.tə/ /dæt.ə/

information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer:

https://dictionary.cambridge.org/dictionary/english/data

#### data

¥

#### **Definitions:**

Information in a specific representation, usually as a sequence of symbols that have meaning.

CNSSI 4009-2015 from IETF RFC 4949 Ver 2

A variable-length string of zero or more (eight-bit) bytes.

#### Sources:

NIST SP 800-56B Rev. 2 under Data

Distinct pieces of digital information that have been formatted in a specific way.

Sources:

NIST SP 800-86 under Data

Pieces of information from which "understandable information" is derived.

#### Sources:

NIST SP 800-88 Rev. 1 under Data

Sources:
A subset of information in an electronic format that allows it to be retrieved or transmitted

NIST SP 1800-10B under Data from CNSSI 4009-2015

NIST SP 1800-25B under Data

NIST SP 1800-26B under Data from CNSSI 4009-2015

Representation of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automatic means.

#### Sources:

NIST SP 800-160v1r1

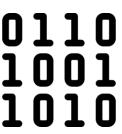
# Data Formats (cont.)

#### **Binary**

- Not human readable
- Memory efficient and fast to parse
- Platform-dependent (negative aspect)
- Difficult format conversion (e.g., open a Word Document in Open Office...)

#### **Text**

- Human readable (mostly)
- Waste more memory and relatively slow to parse
- Platform-independent (positive aspect, but still some encoding problems)
- Easy format conversion.





### **Binary Data**

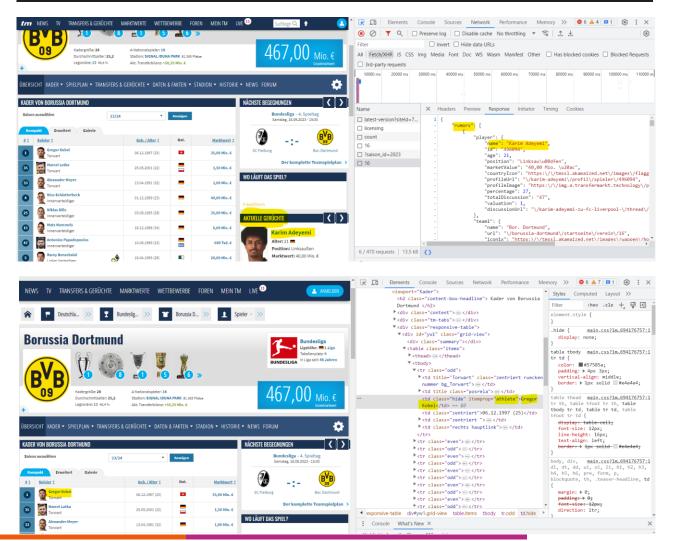
- Binary files are usually thought of as being a sequence of bytes, which means the binary digits (bits) are grouped in eights.
- Binary files typically contain bytes that are intended to be interpreted as something other than text characters.
- Some binary files contain headers to interpret the data in the file. The header often contains a signature or magic number which can identify the format.
- JPEG magic numbers: ff d8 ff e0 or ff d8 ff e1

```
→ dis08 hexdump -n 64 git-meme.jpeg
00000000 ff d8 ff e1 01 08 45 78 69 66 00 00 4d 4d 00 2a
0000010 00 00 08 00 06 01 12 00 03 00 00 01 00 01
0000020 00 00 01 1a 00 05 00 00 00 01 00 00 56 01 1b
0000030 00 05 00 00 00 01 00 00 5e 01 28 00 03 00 00
0000040
→ dis08
```

#### **Text Formats**

- CSV: strings separated by commas and newlines.
  - → Simple and most common in modelling relational data.
- 2. JSON: uses javascript syntax.
  - → Tree-based and most common in data exchange.
- XML: is a mark-up language (meta) and builds XHTML (language of the web).
  - → Tree-based and most common in (web) structured documents.
- RDF: WWW model to exchange metadata and linkages within graphs
- **OWL**: ontologies in semantic web

club\_name,club\_league,player\_position,player\_number,player\_name,player\_dob,player\_country,player\_value
Borussia Dortmund,Bundesliga,Torwart,1,Gregor Kobel,06.12.1997 (25),Schweiz,"35,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,35,Marcel Lotka,25.05.2001 (22),Deutschland,"1,50 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,33,Alexander Meyer,13.04.1991 (32),Deutschland,"1,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,31,Silas Ostrzinski,19.11.2003 (19),Deutschland,150 Tsd. €
Borussia Dortmund,Bundesliga,Abwehr,4,Nico Schlotterbeck,01.12.1999 (23),Deutschland,"40,00 Mio. €"
Borussia Dortmund,Bundesliga,Abwehr,25,Niklas Süle,03.09.1995 (27),Deutschland,"35,00 Mio. €"



### **CSV**: Comma seperated values

```
club_name,club_league,player_position,player_number,player_name,player_dob,player_country,player_value
Borussia Dortmund,Bundesliga,Torwart,1,Gregor Kobel,06.12.1997 (25),Schweiz,"35,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,35,Marcel Lotka,25.05.2001 (22),Deutschland,"1,50 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,33,Alexander Meyer,13.04.1991 (32),Deutschland,"1,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,31,Silas Ostrzinski,19.11.2003 (19),Deutschland,150 Tsd. €
Borussia Dortmund,Bundesliga,Abwehr,4,Nico Schlotterbeck,01.12.1999 (23),Deutschland,"40,00 Mio. €"
Borussia Dortmund,Bundesliga,Abwehr,25,Niklas Süle,03.09.1995 (27),Deutschland,"35,00 Mio. €"
```

- First row usually stores object / attribute / column description.
- Column separator:
  - Default with comma ("," → CSV)
  - Alternatively: tabulator ("\t" → TSV)
- Row separator: usually newline ("\n")
- Strings can be enclosed in quotation marks to escape special characters.
- Quotation marks are escaped by another quotation marks.

```
club_name,club_league,player_position,player_number,player_name,player_dob,player_country,player_value
Borussia Dortmund,Bundesliga,Torwart,1,Gregor Kobel,06.12.1997 (25),Schweiz,"35,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,35,Marcel Lotka,25.05.2001 (22),Deutschland,"1,50 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,33,Alexander Meyer,13.04.1991 (32),Deutschland,"1,00 Mio. €"
Borussia Dortmund,Bundesliga,Torwart,31,Silas Ostrzinski,19.11.2003 (19),Deutschland,150 Tsd. €
Borussia Dortmund,Bundesliga,Abwehr,4,Nico Schlotterbeck,01.12.1999 (23),Deutschland,"40,00 Mio. €"
Borussia Dortmund,Bundesliga,Abwehr,25,Niklas Süle,03.09.1995 (27),Deutschland,"35,00 Mio. €"
```

## CSV (cont.)

- The advantage of CSV files is simplicity... why?
- CSV files are widely supported by:
  - many types of programs and programming languages.
  - can be viewed and imported in text editors; Excel; databases; and Python ©
  - are a straightforward way to represent data.
- Whenever your data has no nested structure → USE CSV!
- **Comma is default**, but tabs are often better as you rarely have to escape strings in tab separated files.



### JSON: JavaScript Object Notation

- JavaScript notation. Human readable but not easy to parse.
- Popular use: Client and web server often use JSON to communicate and exchange data.
- JavaScript lists and arrays are naturally represented with JSON.
- Almost compatible to Python's syntax of datatypes (booleans, numbers, strings) and containers (arrays, lists, tuples, dictionaries), and object-oriented programming.

# JSON (cont.)

#### JSON is built on two structures:

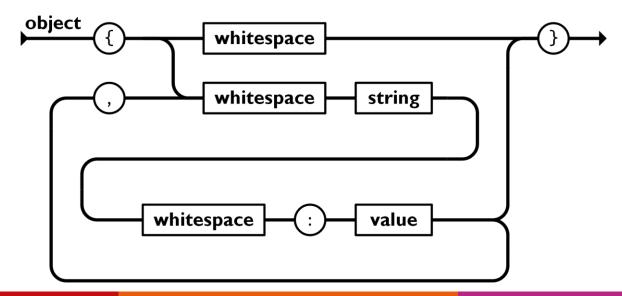
- 1. A collection of name/value pairs. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array.
- 2. An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.

All details can be found online: https://www.json.org

### JSON Object #1

An object is an unordered set of name/value pairs.

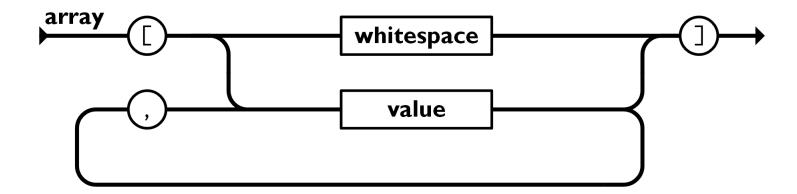
- An object begins with { (left curly parenthesis) and ends with } (right curly parenthesis).
- Each name is followed by: (colon) and its value
- Pairs are separated by , (comma).



### JSON Arrays #2

An array is an ordered collection of values.

- Begins with [ (left bracket) and ends with ] (right bracket).
- Values are separated by , (comma).



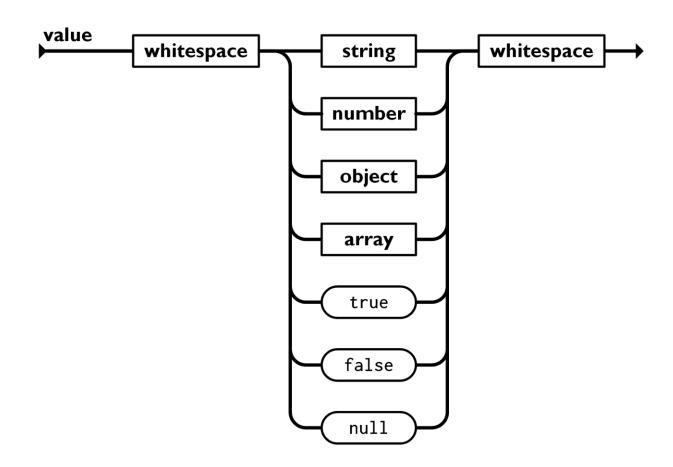
#### **JSON Values**

#### A value can be a

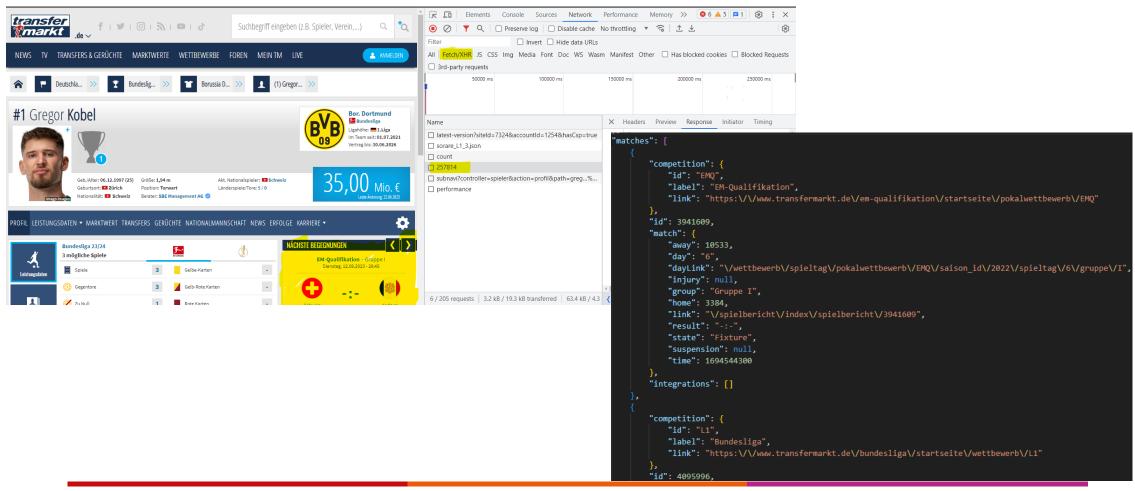
- String in double quotes
- Number
- True or false or null
- Object
- Array.

These structures can be nested.

And so on...



# JSON: Example







"...defines a **syntax** to provide **structured data** sets of any kind with simple, **understandable markups**, which can be **evaluated by applications** of various kinds."

What is only indirectly stated here:

- XML is the eXtensible Markup Language.
- XML is an international W3C standard.
- XML documents are human and machine readable!
- XML is a document description language.
- XML separates structure from presentation, or content from presentation.
- XML documents are developed according to a document model.

## XML (cont.)

#### An XML document may contain:

- **Elements**, possibly with attributes
- Processing instructions
- Comments
- Entity references

An XML document must be well-formed and can be validated.

- XML attribute values must be in " (double quotes).
- XML documents are encoded as linear strings.
- XML documents begin with a special processing instruction, the prologue / header.

#### XML (cont.) root lectures <lectures> element <lecture title="xml"> lecture <day>thursday</day> attribute <slides>many</slides> slides title: xml day </lecture> value </lectures> closing tag thursday many

- Very common data format for web content files.
- Solves a lot of problems (namespaces, encoding, embedded data).
- Not very readable, very verbose.

### XML: Example

```
<?xml Version="1.0" Encoding="UTF-8"?>
<results>
    <result>
       <episode>
           <title>Star Trek - Deep Space Nine</title>
           <title eng>Star Trek - Deep Space Nine</title eng>
           <eptitle>1.1/1.2 Der Abgesandte</eptitle>
           <eptitle eng>Emissary</eptitle eng>
           <description>Der Pilotfilm wurde bei der Erstausstrahlung in Deutschland an einem Stück, später in
           <rating_text>Alles in allem ist "Der Abgesandte" eine gelungene Einführung in die neue Serie. DS9
       </episode>
       <episode>
           <title>Star Trek - Deep Space Nine</title>
           <title_eng>Star Trek - Deep Space Nine</title_eng>
           <eptitle>1.3 Die Khon-Ma</eptitle>
           <eptitle_eng>Past Prologue</eptitle_eng>
           <description>Ein bajoranischer Aufklärer taucht in unmittelbarer Nähe von DS9 auf, verfolgt von e.
           <rating text>Im Wesentlichen bot diese Episode nur Star-Trek-Hausmannskost. Sie bot weder besonde
       </episode>
       <episode>
           <title>Star Trek - Deep Space Nine</title>
           <title eng>Star Trek - Deep Space Nine</title eng>
           <eptitle>1.4 Unter Verdacht</eptitle>
           <eptitle_eng>A Man Alone/eptitle_eng>
           <description>Lt. Jadzia Dax benutzt ihre Konzentrationskräfte in der Holosuite auf Deep Space 9,
           <rating_text>Diese Folge macht deutlich, dass Odo für DS9 das darstellt, was Spock für TOS und Da
       </episode>
       <episode>
           <title>Star Trek - Deep Space Nine</title>
           <title_eng>Star Trek - Deep Space Nine</title_eng>
           <eptitle>1.5 Babel</eptitle>
           <eptitle_eng>Babel</eptitle_eng>
           <description>Es ist ein schlechter Tag für Miles O'Brien auf DS9. Er wusste, es würde technologis-
           <rating_text>Man merkt, dass diese Folge zu den ersten der Serie gehört. Aber die Ecken und Kanter
       </episode>
       <episode>
           <title>Star Trek - Deep Space Nine</title>
           <title_eng>Star Trek - Deep Space Nine</title_eng>
           <eptitle>1.6 Tosk, der Gejagte</eptitle>
```

#### XML: well formed

An element always has a start tag and an end tag.

- This is a <StartTag>.
- This is the end of the </StartTag>.
- The tag names are case-sensitive.
- <tag></tag> and not <tag></Tag>.

Empty elements, called milestone elements, can be abbreviated

- <emptyTag />.
- Meaning: attributes may be included, but no content may be placed between the tags.

## XML: well formed (cont.)

All elements must be **nested correctly!** 

```
<up><up><up></up></up></up></up></down>Text</down></up></down>
```

#### **Element names**

- Must begin with a letter, underscore, or colon.
- Can contain letters, numbers, hyphens, periods, or underscores, as well as umlauts and accents.

An XML document has only exactly one root node!

#### XML: Attributes

Elements can be defined in more detail by attributes.

- Attributes are always in the start tag!
- <name AttrName="Attributwert"> </name>

#### Properties of attributes

- Choice of designer on what information is element-worthy and which is attribute-worthy.
- Multiple attributes are allowed per element.
- Same named attributes in different elements

## XML: Prolog

An XML document always starts with a prolog that contains

Declaration

```
<?xml version="1.0"?>
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
```

Processing information (optional)

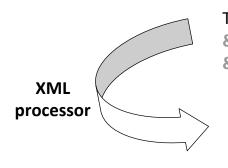
```
<?xsl-stylesheet type="text/xsl" href="myown.xsl"?>
```

Embedding of the document model (optional)

DTD, XML Schema, RelaxNG, Schematron

```
<!DOCTYPE tei SYSTEM "tei.dtd">
```





This element is encoded as <code&gt;&lt;Element&gt;... &lt;/Element&gt;&lt;/code&gt;

This element is encoded as <code><Element>...</Element></code>

#### XML: Entities

Entities stand for something else.

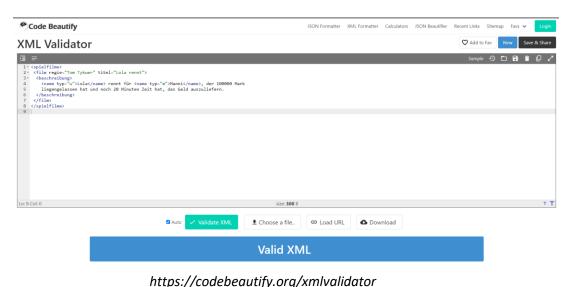
- To avoid conflicts during XML processing!
- XML allows to set up a reference that points to an entity.
- The **XML processor replaces** the references.

Some predefined entity references e.g., special characters:

#### XML: Verification

The verification is done by an XML parser. An XML document

- Must be well-formed and can be valid.
- Is well-formed if it complies with the rules of the XML standard.
- Is valid if it is **well-formed and conforms to the grammar** of the XML schema.



# JSON versus XML

It is based on JavaScript language.	It is derived from SGML.
It is a way of representing objects.	It is a markup language and uses tag structure to represent data items.
It does not provide any support for namespaces.	It supports namespaces.
It supports arrays.	It doesn't support arrays.
Its files are very easy to read as compared to XML.	Its documents are <b>comparatively difficult to read</b> and interpret.
It doesn't use end tag.	It has <b>start</b> and <b>end tags</b> .
It is less <b>secured</b> .	It is more secured than JSON.
It doesn't support <b>comments</b> .	It supports comments.
It supports only UTF-8 encoding.	It supports various encoding.

### XML: as a tree structure

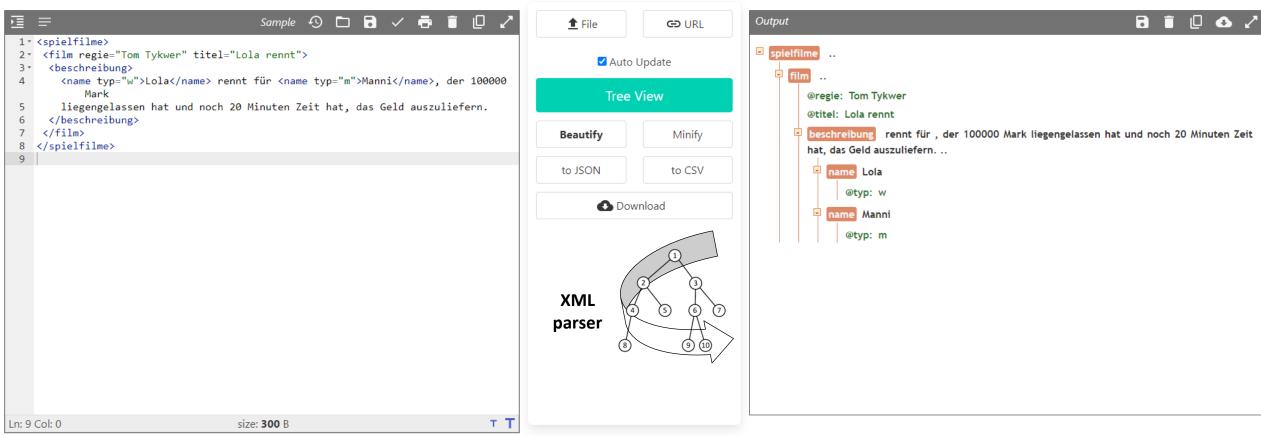
**Code Beautify** 

XML Formatter Calculators JSON Beautifier Recent Links Sitemap

Add to Fav

Save & Share

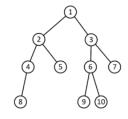
#### **XML Viewer**



https://codebeautify.org/xmlviewer

## XML: Xpath Example

Xpath uses path expressions to select nodes or node-sets in an XML document.

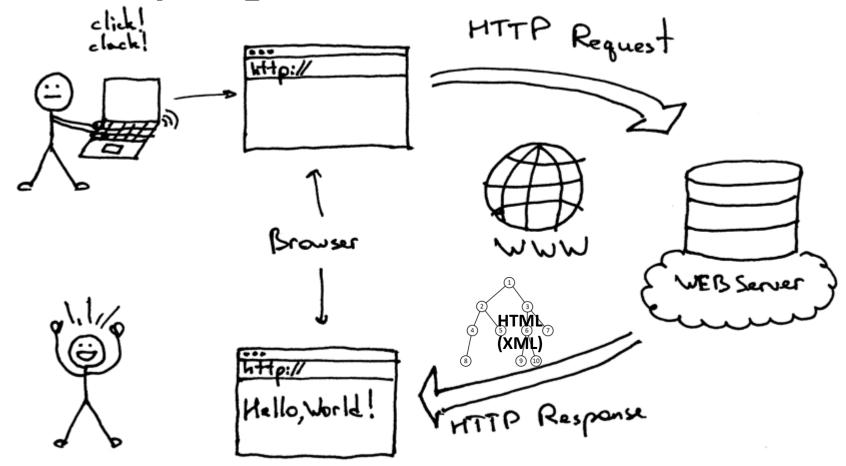


/film[@regie='Tom Tykwer']/@titel

```
/film/beschreibung
```

...

# Web Scraping



https://monashdatafluency.github.io/python-web-scraping/section-2-HTML-based-scraping/

## Training #1



Your web browser interprets HTML files (a similar type as XML).

- Open a web site of your interest, e.g.,
  - Buecher.de
  - Transfermarkt and your favourite club
  - and click on Option+Command+U (Mac) or F12 (Windows) on your keyboard to open the developer mode (DevTools).
- Click on the "Select element" button: and click on something on the web-page.







Right click on "Copy > XPath" of the element in the HTML document and note it down.

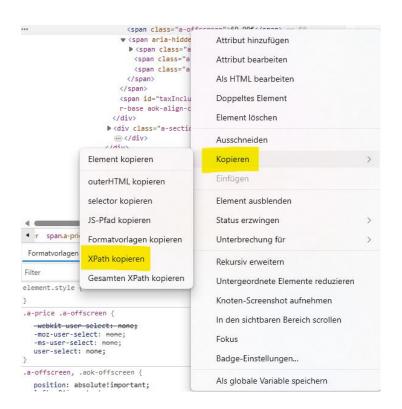
What element(s) interest you?

→ Imagine you can extract this/these element/s from various categories/products/players.

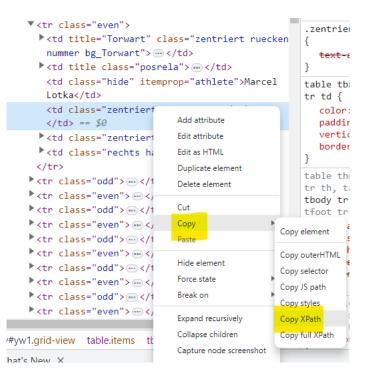
How could these elements benefit someone from an analytical perspective?

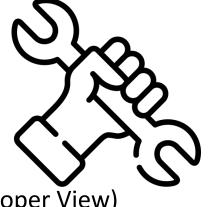
# Training #1 (cont.)

#### Microsoft Edge

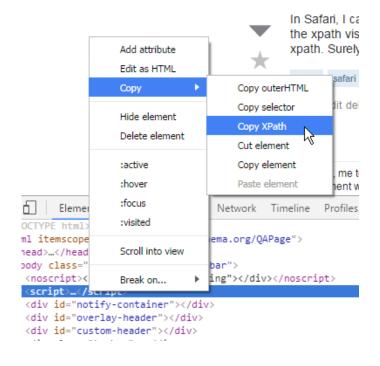


#### Chrome

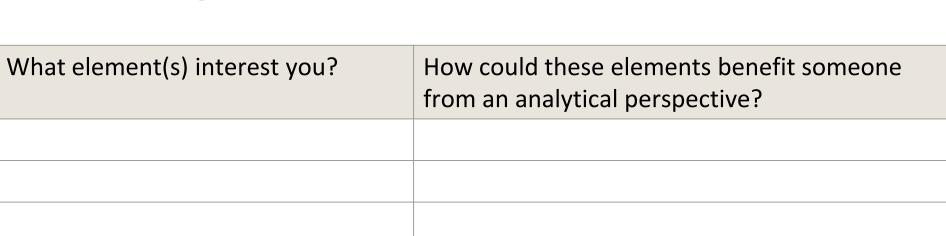




Safari (Turn on Developer View)

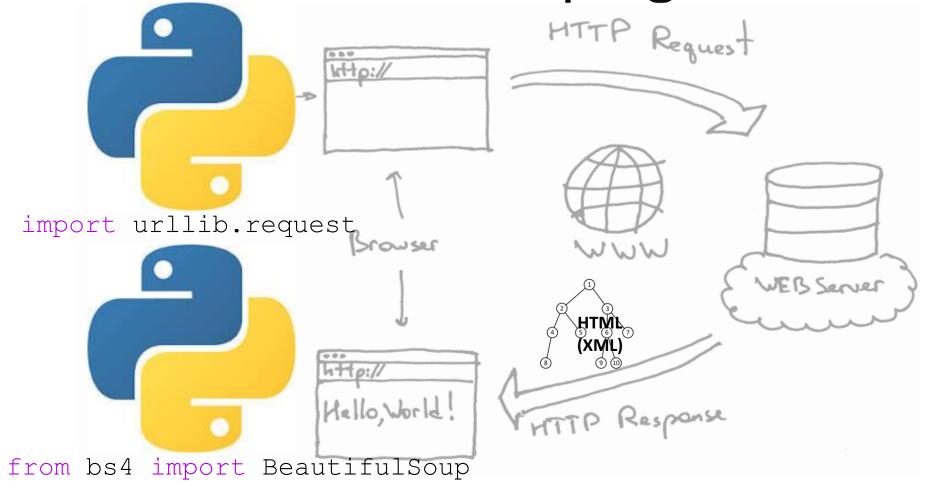


# Training#1



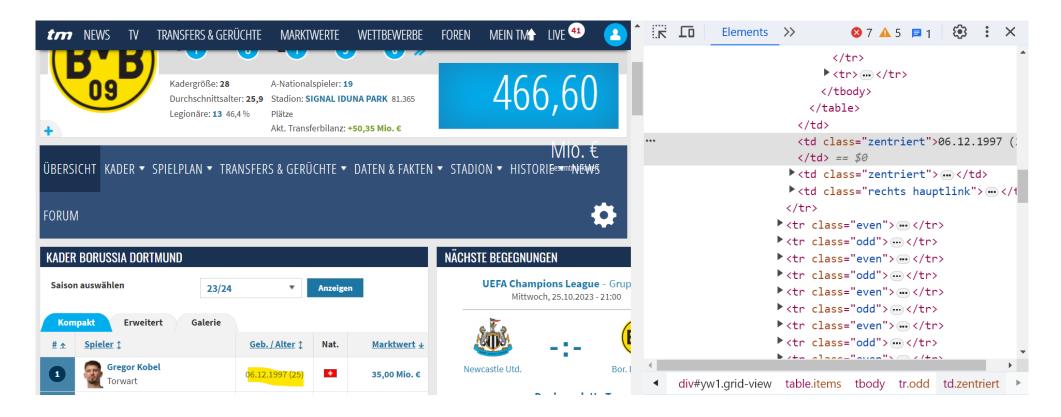


### Demo: First Web Scraping



https://monashdatafluency.github.io/python-web-scraping/section-2-HTML-based-scraping/

### Demo: First Web Scraping



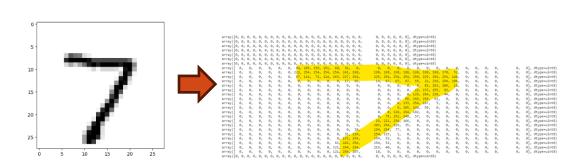
Xpath:[@id="yw1"]/table/tbody/tr[1]/td[3]

## Takeaways

- CSV, JSON, and XML are typical text-based file formats that you will encounter in the wild.
- All have their use cases and applications, and it always depends on what you would like to accomplish!
- It's good practice to know all three one of these is usually available and applicable.

### Outlook

- In the next week #4, we will dive deep into Python NumPy.
- This is the core library for numerically typed arrays. You might use it in your project for Data Wrangling purposes.



```
usp_height = np.array([189, 170, 189, 163, 183, 171, 185, 168, 173, 183,
                       173, 173, 175, 178, 183, 193, 178, 173, 174, 183,
                       183, 168, 170, 178, 182, 180, 183, 178, 182, 188,
                       175, 179, 183, 193, 182, 183, 177, 185, 188, 188,
                       182, 185, 191, 182])
print("Mean height: ", usp_height.mean())
                                               #Prints "180.04.."
print("Standard deviation:", usp_height.std()) #Prints "6.983.."
print("Minimum height: ", usp_height.min())
                                               #Prints "163"
print("Maximum height: ", usp height.max())
                                               #Prints "193"
print("25th: ", np.percentile(usp_height, 25)) #Prints "174.75"
print("Median: ", np.median(usp_height))
                                               #Prints "182.0"
print("75th: ", np.percentile(usp_height, 75)) #Prints "183.5"
```

 In weeks 6+ we will dive deep into Python Pandas' data frames to arrange tabular data to analytical goals.

# See you again next week online!

Questions?