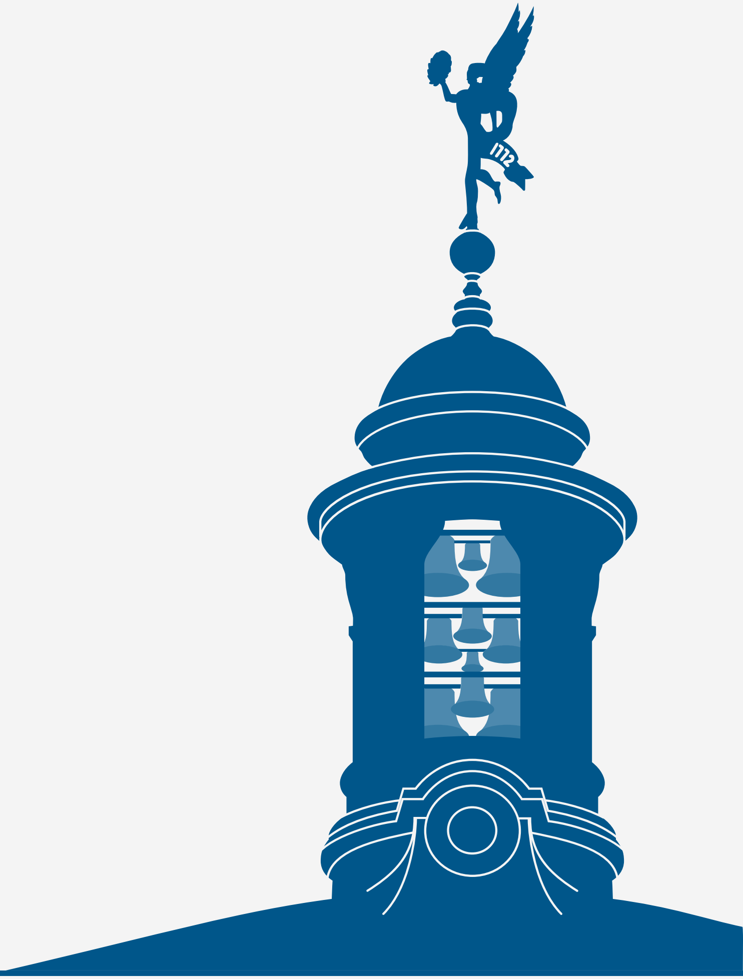
Automated Data Collection

Data extraction with Regexes,

Dr. Jakob Jünger, Chantal Gärtner



Remember: extraction languages

CSS selectors

```
ul#likes li.last
```

XPath

```
//ul/li[@class='last']
```

Regular expressions

```
<li class="last">[^<]*</li>
```

```
<ul id="likes">
      <li class="first">
            <a href="http://example.com/wef>
      </li>
      <li>
          Farmers of the Future
      </li>
      <li class="last">
        Bunny Mcdiarmid
      </li>
</ul>
```

While you listen…

… try to note expressions or techniques that can be used to extract the date from the following URL:

https://www.aljazeera.com/news/2021/12/16/what-you-should-know-about-the-conflict-between-russia-ukraine

Regular Expressions: search patterns with placeholders

- Characters and numbers match themselves	year	matches	year
- A dot . matches an arbitrary character	yea.	matches e.g.	year or yeah or yeas
- A list of characters can bei compiled in square brackets	yea[rh]	matches	year or yeah
- The lists can contain all possible options or ranges	[012345] or [0-5]	matches e.g.	1 or 5 or 0

Regular Expressions: search patterns with placeholders

- Quantifiers determine the number of characters

  * means zero, one or more
  + means at least one
  ? means zero or one
  {3} means exactly three times

Yea[a-z]*	matches e.g.	yea or yearbook
[0-9]+	matches e.g.	1 or 124 or 1782738
years?	matches	year or years
[0-9]{2}	matches e.g.	09 or 15 or 34

What does the following expression match? [a-z]+

Characters

a10Z	Letters or numbers match themselves
\n	Line break (line feed)
\r	Line break (carriage return)
\t	Tabulator
[0-9]	Square brackets define character classes. Ranges can be defined using a hyphen
[^a-z]	The caret within square brackets means **not** to match the character class
\(	Characters with special meanings need to be escaped with a backslash

Quantificators

*	Any number of the previous characters (0 or more)
+	At least one of the previous characters
?	Optional character (0 or 1)
{4}	Exactly four characters

Capture groups

(abc)	Expressions may be grouped with brackets
\1	Groups can be references when doing search & replace by backslash followed by the number of the group

Anchors

^	The caret at the beginning of an expression binds the search to the beginning of the line
$	The dollar sign at the end of an expression binds the pattern to the end of the line

While you listen…

… try to note expressions or techniques
that can be used to extract the date
from the following URL:

https://www.aljazeera.com/news/2021/12/16/what-you-should-know-about-the-conflict-between-russia-ukraine

Practical session with Python