WEEK 1: WEB DATA COLLECTION 1

SECU0050

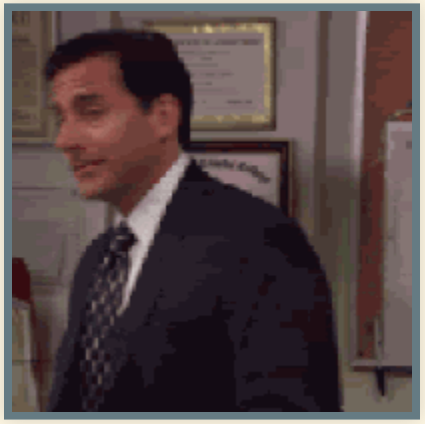
BENNETT KLEINBERG

16 JAN 2020

Data Science for Crime Scientists

What is this? and Why do we need it?

AAAH: SO WE'RE TALKING BIG DATA!



PROBLEMS WITH “BIG DATA”

- what is “big”?
- data = data?
- complexity of data?
- sexiness of small data

THINGS YOU'LL DO

- Accessing news article data and academic literature through APIs
- Building proper web-scrapers to collect data on the FBI's most wanted terrorists
- Analysing patterns of language use in text data
- Utilising Natural Language Processing to understand controversial news coverage
- Identifying the semantic relationships of concepts through automated methods
- Building fake news and sentiment classification systems using machine learning
- Utilising clustering to identify patterns in YouTube videos
- Building neural networks for prediction tasks
- your own project

LEARNING HOURS

Component	Amount	Duration	Total hours
Lectures	10	2h	20h
Tutorials/practicals	10	1h	10h
Assessment: class test	1	1h	1h
Assessment: project	1	47h	47h
Homework/revision/self-study	11	10h	110h
TOTAL	-	-	188h

≈ 17 hours per week

BECOMING A REAL PROBLEM-SOLVER

- Principle 1: There's no magic in Data Science
- Principle 2: Data never come in a spreadsheet
- Principle 3: Data are hidden in front of you
- Principle 4: Programming can only ever be the vehicle

~~It all starts with the data.~~

It all starts with a problem.

WEEK 1: WEB DATA COLLECTION 1

TODAY

- Types of webscraping
- Using APIs
- “Real” webscraping: basics of a webpage

WHAT IS WEBSCRAPING ANYWAY?

THE GAME CHANGER!

- direct broadcasting of ideas
- “unfiltered” and “uncensored” (?)
- location-enabled
- and: *en masse*

TYPES OF WEBSCRAPING

	Data shared	**Data not shared**
Ready-made table	Download	*closed source*
Not ready-made	API	Real webscraping

APPLICATION PROGRAMMING INTERFACES (APIS)

API: BASICS

Goal:

- helping developers interact with the platform
- facilitating interaction in an automatable manner
- analogous to the GUI
- part of it: enabling data access
- contains precise documentation

WHAT AN API DOES NOT DO:

- give you all the data
- be free forever
- give you full control

There’s no free lunch!

CORE ELEMENTS OF AN API:

- GET requests
- POST requests

Implementable in different ways…

CLASSES OF APIS

1. Web APIs
    - send requests through the browser
    - add URL parameters

      ```
      https://data.police.uk/api/crimes-at-locati
      ```
2. Libraries/packages for APIs
    - depending on the API: python, js, php, ruby
    - = frameworks to access the API
    - = methods implemented in different languages
3. API wrappers
    - R packages that use the API

USEFUL WEBSITES THAT HAVE AN API

- Twitter (`twitteR`, `rtweet`)
- YouTube (`tuber`)
- Instagram
- Facebook
- Reddit
- The Guardian

Recommended: try online tutorials

API EXAMPLE 1: NEWSPAPER COVERAGE

Accessing TheGuardian data using `GuardianR`

```
library(GuardianR)

get_guardian("SOME SEARCH TERM",
             from.date="START_DATE",
             to.date="END_DATE",
             api.key="YOUR_ACCESS_KEY")
```

HOW TO OBTAIN THE ACCESS KEY?

https://open-platform.theguardian.com/access/

Get Started

Choose the level of access you need

How it works

The Open Platform is a public web service for accessing all the content the Guardian creates, categorised by tags and section. To get started, You need an key to successfully authenticate against the API.

Choose the level of access required for your application:

Developer

This key is for any non-commercial usage of the content, such as student dissertations, hackathons, nonprofit app developers.

- Up to 12 calls per second
- Up to 5,000 calls per day
- Access to article text
- Access to over 1,900,000 pieces of content
- **Free for non-commercial usage**

Register developer key

Commercial

This key is for any commercial enterprises and developers wanting to monetise our content.

- Custom throttle limit
- Custom quota limit
- Access to article text, images, audio and videos
- Access to entire Guardian content store
- **Price dependent on usage**

Request commercial key

STEP 1: SENDING API REQUESTS

Aim: search for keyword “crime” in Aug 2019

```
request.crime = get_guardian("crime",
                             from.date="2019-08-01",
                             to.date="2019-08-31",
                             api.key="ea765...")
```

Note: save these requests in an R object (variable)

STEP 2: UNDERSTANDING THE DATA

Dimensions (rows, columns) of the `request.crime` object

```
dim(request.crime)
```

```
## [1] 430  27
```

COLUMN NAMES

```
names(request.crime)
```

```
##  [1] "id"                    "sectionId"             "sectionName"
##  [4] "webPublicationDate"    "webTitle"              "webUrl"
##  [7] "apiUrl"                "newspaperPageNumber"   "trailText"
## [10] "headline"              "showInRelatedContent" "lastModified"
## [13] "hasStoryPackage"       "score"                 "standfirst"
## [16] "shortUrl"              "wordcount"             "commentable"
## [19] "allowUgc"              "isPremoderated"        "byline"
## [22] "publication"           "newspaperEditionDate" "shouldHideAdverts"
## [25] "liveBloggingNow"       "commentCloseDate"      "body"
```

DISPLAYING THE DATA

id	sectionName	webTitle	headline
books/2019/aug/07/top-10-true-books-duncan-campbell	Books	Top 10 true crime books	Top 10 true crime books
books/2019/aug/23/best-recent-crime-and-thrilers-review-roundup	Books	The best recent crime and thrillers â€“ review roundup	The best recent crime and thrillers â€“ review roundup
politics/2019/aug/12/boris-johnson-crime-crackdown-policing-prisons	Politics	Johnson's crackdown on offenders will only entrench crime	Johnson's crackdown on offenders will only entrench crime

id	sectionName	webTitle	headline
commentisfree/2019/aug/12/boris-johnson-crime-fighting-crusade-posturing-brexit	Opinion	Boris Johnsonâ€™s â€˜crime fightingâ€™ crusade is mere posturing | Simon Jenkins	Boris Johnsonâ€™s â€˜crime fightingâ€™ crusade is mere posturing
uk-news/2019/aug/05/rural-crime-in-britain-hits-seven-year-high	UK news	Rural crime in Britain hits seven-year high	Rural crime in Britain hits seven-year high

DISPLAYING THE DATA

	headline	wordcount	byline	newspaperEditionDate
11	Surge in violent crime in Barcelona prompts calls for legal reform	451	Stephen Burgen in Barcelona	2019-08-24T00:00:00Z
12	Monica Lewinsky to produce American Crime Story drama about Clinton scandal	505	Gwilym Mumford	2019-08-08T00:00:00Z
13	Eric Cantona speech: humans 'will become eternal' â€“ unless crime or war intervene	397	Guardian sport	NA
14	87 bird crime incidents last year and just one conviction, says RSPB	388	Patrick Barkham	2019-08-29T00:00:00Z
15	Video of trans women forced from LA bar prompts hate crime	994	Sam Levin in Los Angeles	NA

	headline	wordcount	byline	newspaperEditionDate

THE ARTICLE BODY

```
as.character(request.crime$body[33])
```

```
## [1] "<p>The police minister, Kit Malthouse, has insisted that a new wa
```

STEP 3: USING THE DATA

Note: not all sections are relevant (e.g., book reviews vs politics vs uk news)

Option 1: sending a new API request (R package docs)

```
library(GuardianR)
request.crime.uknews = get_guardian("crime",
                                    from.date="2019-08-01",
                                    to.date="2019-08-31",
                                    section = 'uk-news',
                                    api.key="ea765...")
```

“UK NEWS” CRIME COVERAGE

```
dim(request.crime.uknews)
```

```
## [1] 32 27
```

id	sectionName	webTitle	headline
uk-news/2019/aug/05/rural-crime-in-britain-hits-seven-year-high	UK news	Rural crime in Britain hits seven-year high	Rural crime in Britain hits seven-year high
uk-news/2019/aug/02/east-london-volunteers-set-up-knife-amnesty-bin	UK news	‘There’s not enough police’: East London volunteers take on knife crime	‘There’s not enough police’: East London volunteers take on knife crime
uk-news/2019/aug/08/female-knife-possession-crimes-in-england-rise-by-73	UK news	Female knife possession crimes in England rise by 73%	Female knife possession crimes in England rise by 73%
uk-news/2019/aug/14/nca-freezes-more-than-100m-after-court-order-on-imported-money	UK news	NCA freezes Â£100m suspected to be from corruption overseas	NCA freezes Â£100m suspected to be from corruption overseas

id	sectionName	webTitle	headline
uk-news/2019/aug/07/alesha-macphail-killer-appeals-against-27-year-sentence	UK news	Alesha MacPhail killer appeals against 27-year minimum sentence	Alesha MacPhail killer appeals against 27-year minimum sentence

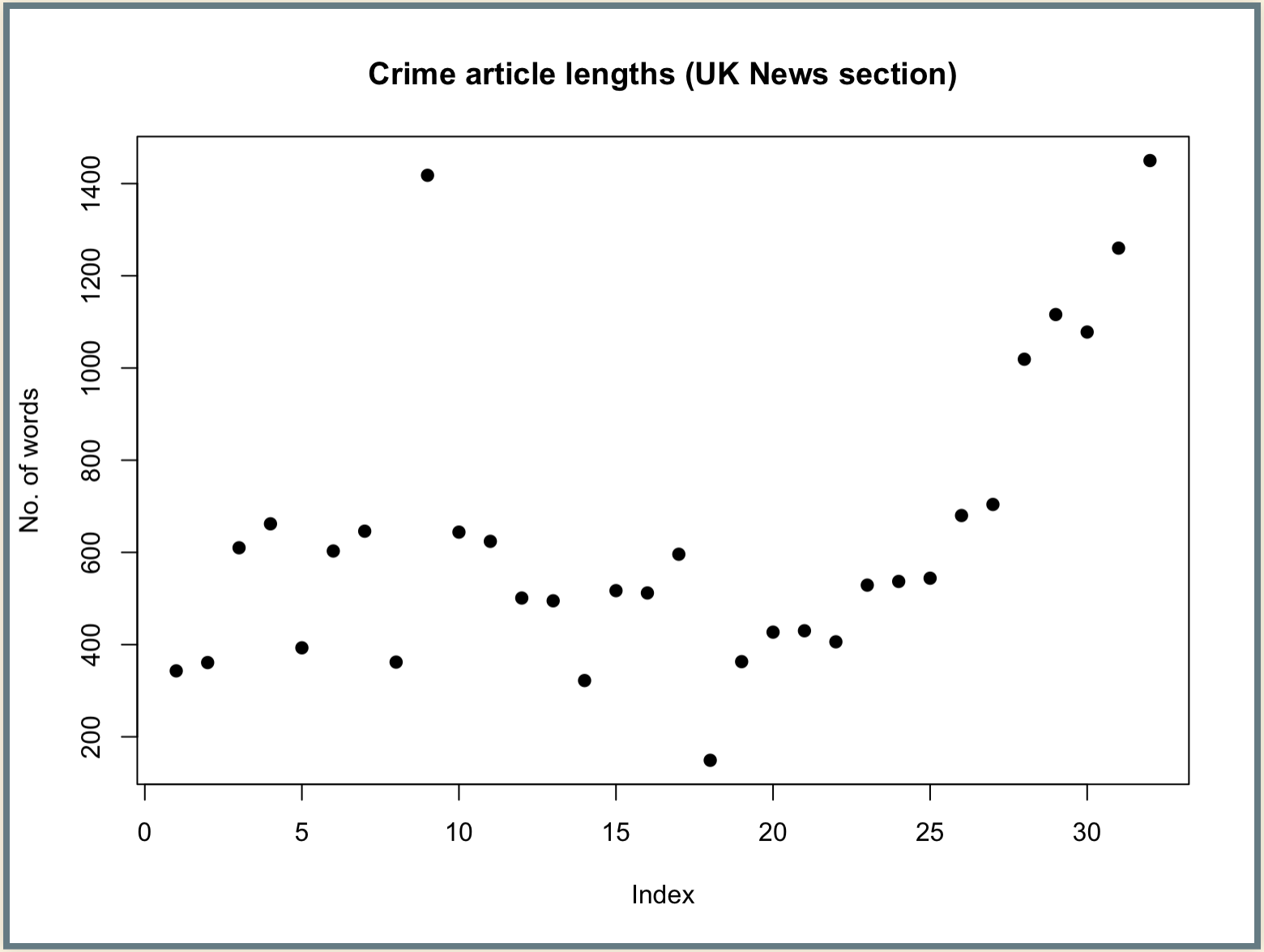
OPTION 2: SUBSETTING THE DATA

```
crime_data_uknews = request.crime[request.crime$sectionId == "uk-news",
dim(crime_data_uknews)
```

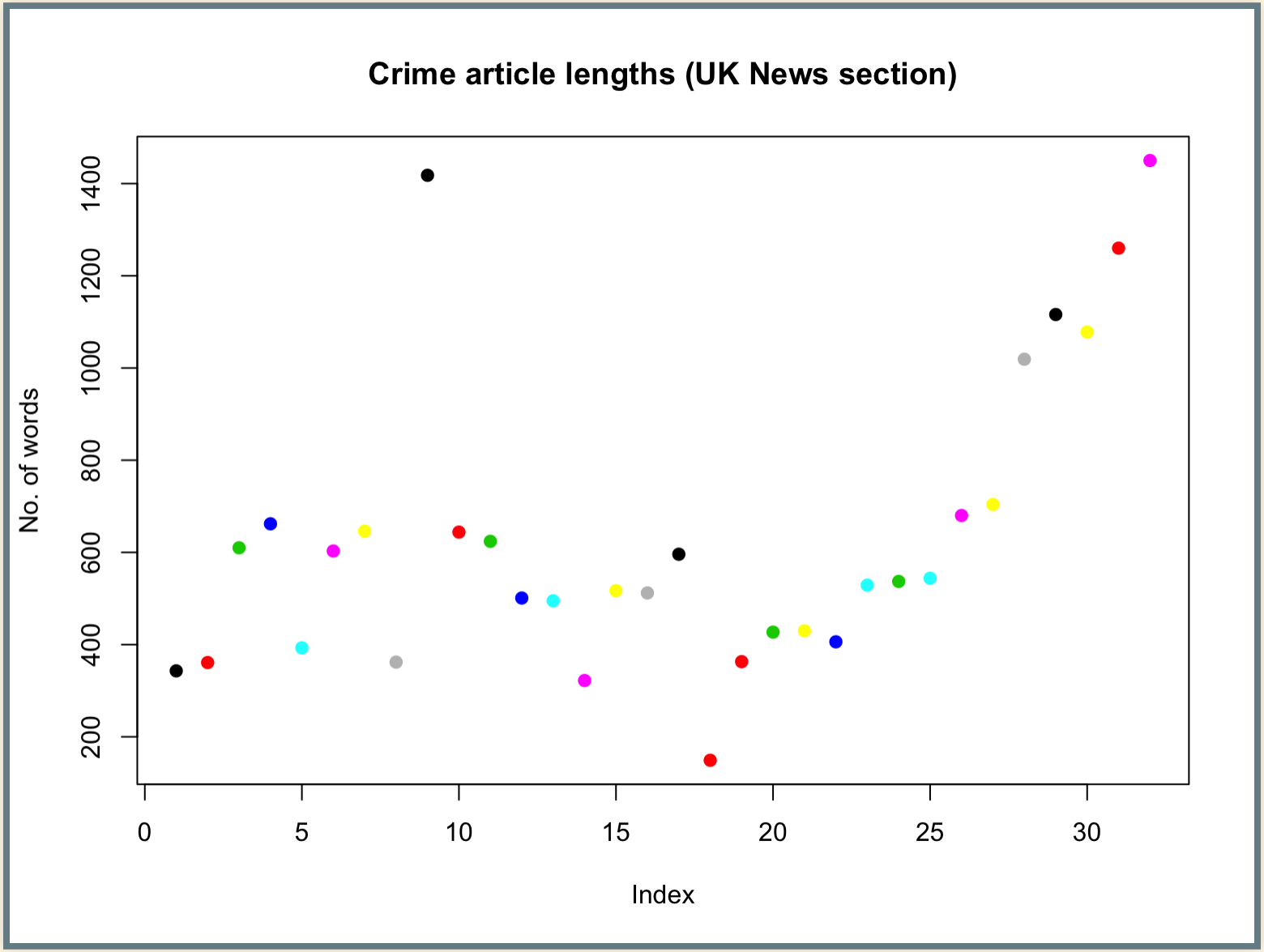
```
## [1] 32 27
```

STEP 3: USING THE DATA

Aim: investigate article length


Crime article lengths (UK News section)
No. of words
200
400
600
800
1000
1200
1400
0
5
10
15
20
25
30
Index

... BY AUTHOR


Crime article lengths (UK News section)
No. of words
200
400
600
800
1000
1200
1400
0
5
10
15
20
25
30
Index

NUMERICALLY

```
tapply(X = as.numeric(as.character(request.crime.uknews$wordcount))
       , INDEX = request.crime.uknews$byline
       , FUN = mean)
```

```
##                                    Frances Perraudin
##                                                343.0
##                                          Amy Walker
##                                                362.0
##           Jamie Grierson Home affairs correspondent
##                                                573.5
##                  Rupert Neate Wealth correspondent
##                                                662.0
##                Libby Brooks Scotland correspondent
##                                                393.0
##                 Helen Pidd North of England editor
##                                               1026.5
##         Vikram Dodd Police and crime correspondent
##                                                862.0
##                                             Letters
##                                                362.0
##                                       Duncan Campbell
##                                                1418.0
```

API EXAMPLE 2: ACADEMIC LITERATURE

arxiv.org/list/cs.CL/recent

Cornell University

arXiv.org > cs > cs.CL

Computation and Language

Authors and titles for recent submissions

- Wed, 1 Jan 2020
- Mon, 30 Dec 2019
- Wed, 25 Dec 2019
- Tue, 24 Dec 2019
- Mon, 23 Dec 2019

[total of 108 entries: **1-25** | 26-50 | 51-75 | 76-100 | 101-108]
[showing 25 entries per page: fewer | more | all]

Wed, 1 Jan 2020 (showing first 25 of 34 entries)

[1] **arXiv:1912.13415** [**pdf**, **other**]

End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models

John Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D. Bader, Bo Wang

Comments: 12 pages, 2 figures

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)

[2] **arXiv:1912.13413** [**pdf**, **other**]

Semantics- and Syntax-related Subvectors in the Skip-gram Embeddings

Maxat Tezekbayev, Zhenisbek Assylbekov, Rustem Takhanov

Comments: 2 pages, 1 figure, Student Abstract

Subjects: **Computation and Language (cs.CL)**

[3] **arXiv:1912.13362** [**pdf**]

Text Classification for Azerbaijani Language Using Machine Learning and Embedding

Umid Suleymanov, Behnam Kiani Kalejahi, Elkhan Amrahov, Rashid Badirkhanli

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG); Machine Learning (stat.ML)

UTILISING THE `arxiv` PACKAGE

- Docs: https://ropensci.org/tutorials/arxiv_tutorial/
- Repo: https://github.com/ropensci/aRxiv

SENDING QUERIES

1. Author search with full data

```
library(aRxiv)
arxiv_author_search = arxiv_search(query = 'au:"Hinrich Schuetze"'
                                   , limit=50)
```

id	submitted	updated	title
cmp-lg/9503009v1	1995-03-08 18:36:40	1995-03-08 18:36:40	Distributional Part-of-Speech Tagging
cmp-lg/9707002v1	1997-07-08 21:08:34	1997-07-08 21:08:34	Automatic Detection of Text Genre
1301.2811v3	2013-01-13 19:33:31	2013-04-26 12:33:50	Cutting Recursive Autoencoder Trees
1301.3627v2	2013-01-16 08:37:39	2013-05-11 12:17:44	Two SVDs produce more focal deep learning representations
1610.00479v3	2016-10-03 10:30:13	2017-05-01 14:30:00	Nonsymbolic Text Representation

SENDING QUERIES

2. Author search with count data

```
arxiv_count(query = 'au: Hinrich Schuetze')
```

```
## [1] 229
```

SENDING QUERIES

3. Articles submitted by category and date

All articles in the “Computation & Language” subcategory submitted in the 1st of May 2019.

```
arxiv_count(query = 'cat:cs.CL AND submittedDate:20190501*')
```

```
## [1] 10
```

SENDING QUERIES

4. Articles submitted by category and date

Do NLP researchers work over the Christmas holidays?

What would the query need to look like?

```
arxiv_count(query = 'cat:cs.CL AND submittedDate:[20191224* TO 20191231*
```

```
arxiv_count(query = 'cat:cs.CL AND submittedDate:[20191224* TO 20191231*
```

```
## [1] 55
```

APIS: PROS & CONS

Pro

- easy to access
- nicely documentation
- **works even if website changes**

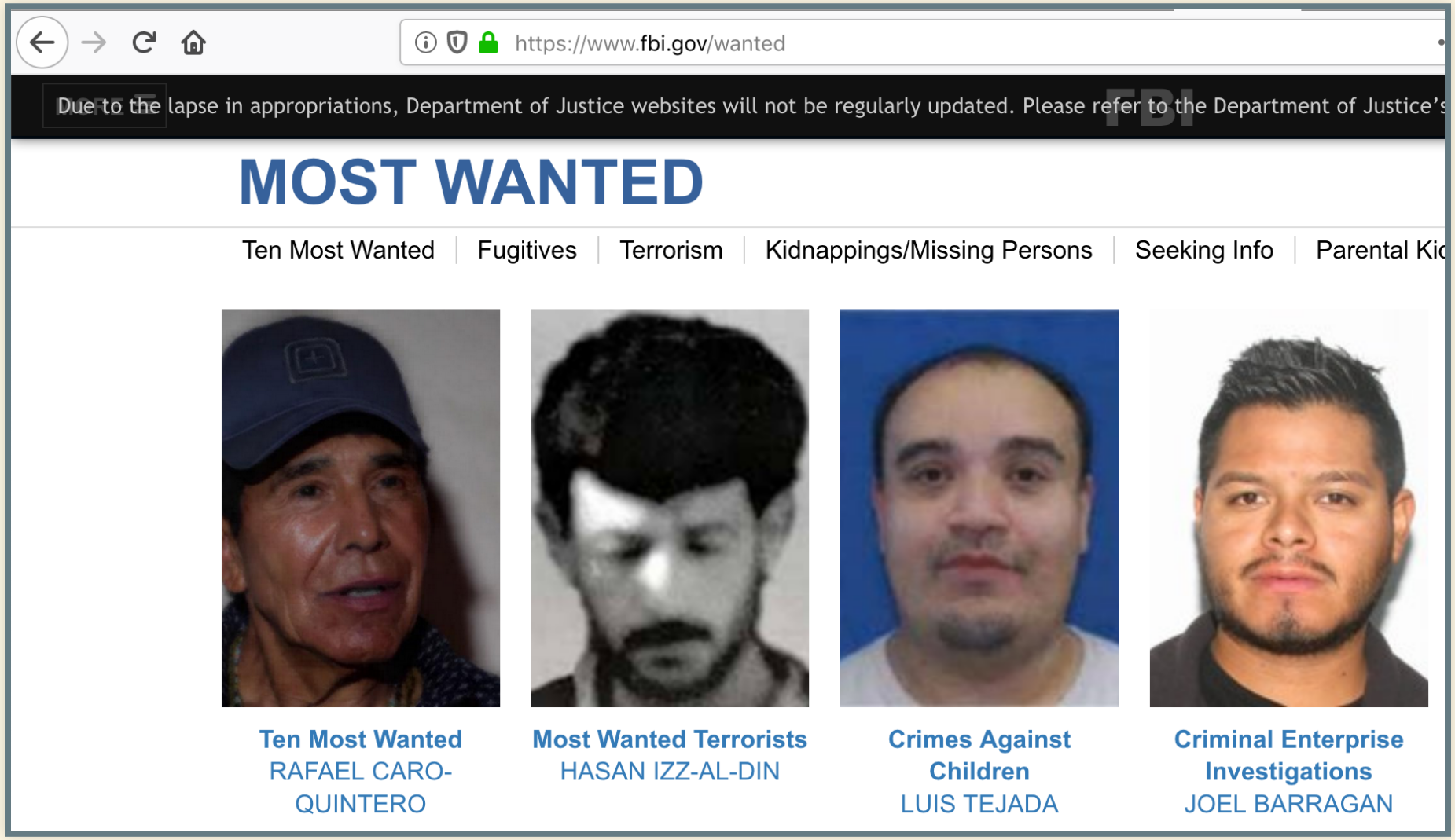
Cons

- quota limits ($ $ $)
- under the platforms' control
- only for few platforms

DON’T LET THE DATA DETERMINE YOUR RESEARCH!

BEYOND APIS

What about:



UK MISSING PERSONS UNIT

National Crime Agency (GB) | https://www.missingpersons.police.uk/en-gb/case-

Bureau Reference:	18-009282
Location	London
Gender	Male
Date found	21 September 2018
Age	30 - 40
Ethnicity	White European

View case details

View case details

Bureau Reference:	18-006195
Location	Dunsden
Gender	Male
Date found	15 May 2018
Age	16 - 100
Ethnicity	White European

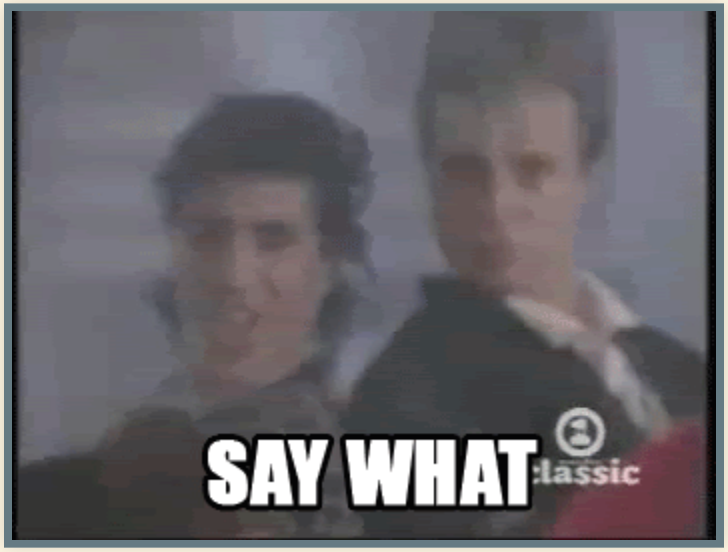
View case details

NO APIS

- incels.me
- Stormfront
- 4chan
- **APIs are restrictive!**

... WHAT ABOUT:

YOUR RESEARCH PROJECT -> NO API?



Main problem:

Really ‘juicy’ data of the Internet vs **APIs**

“REAL” WEBSCRAPING: BASICS OF A WEBPAGE

THREE ELEMENTS OF A WEBPAGE

1. Structure
2. Behaviour
3. Style

THREE ELEMENTS OF A WEBPAGE

1. Structure
2. Behaviour
    - JavaScript (**!= Java**)
    - user interaction
    - examples: alerts, popups, server-interaction
3. Style

THREE ELEMENTS OF A WEBPAGE

1. Structure
2. Behaviour
3. Style
    - CSS (Cascading Style Sheets)
    - formatting, design, responsiveness
    - examples: submit buttons, app interfaces

THREE ELEMENTS OF A WEBPAGE

1. Structure
    - HTML (hypertext markup language)
    - structured with `<tags>`
    - contains the pure content of the webpage
2. Behaviour
3. Style

FOR NOW: HTML

The very basics of HTML:

Raw architecture of a webpage

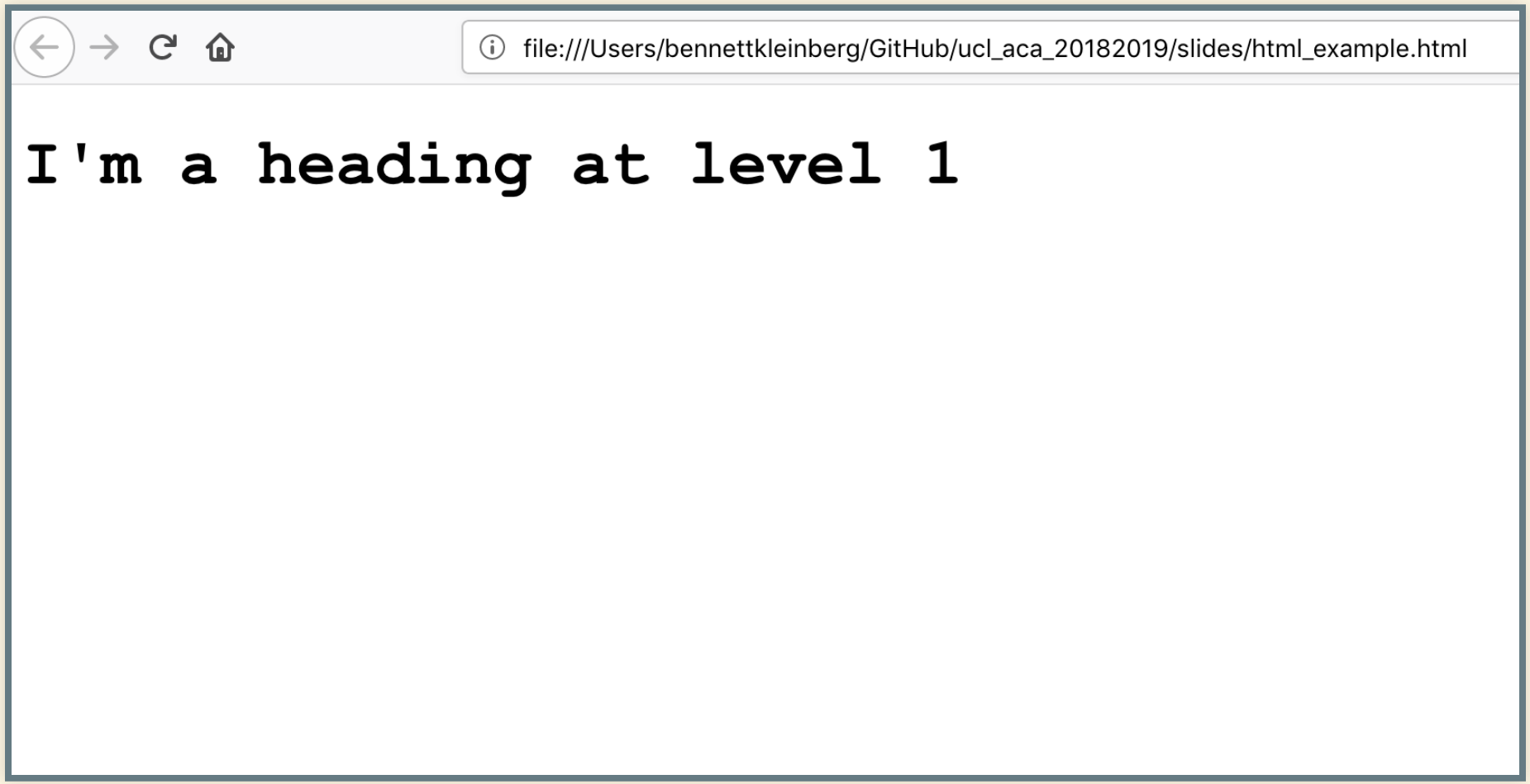
```
<!DOCTYPE html>
<html>
<body>

HERE COMES THE VISIBLE PART!!

</body>
</html>
```

Note: Every tags `< >` is closed `< />`. Content is contained within the tag.

HTML BASICS

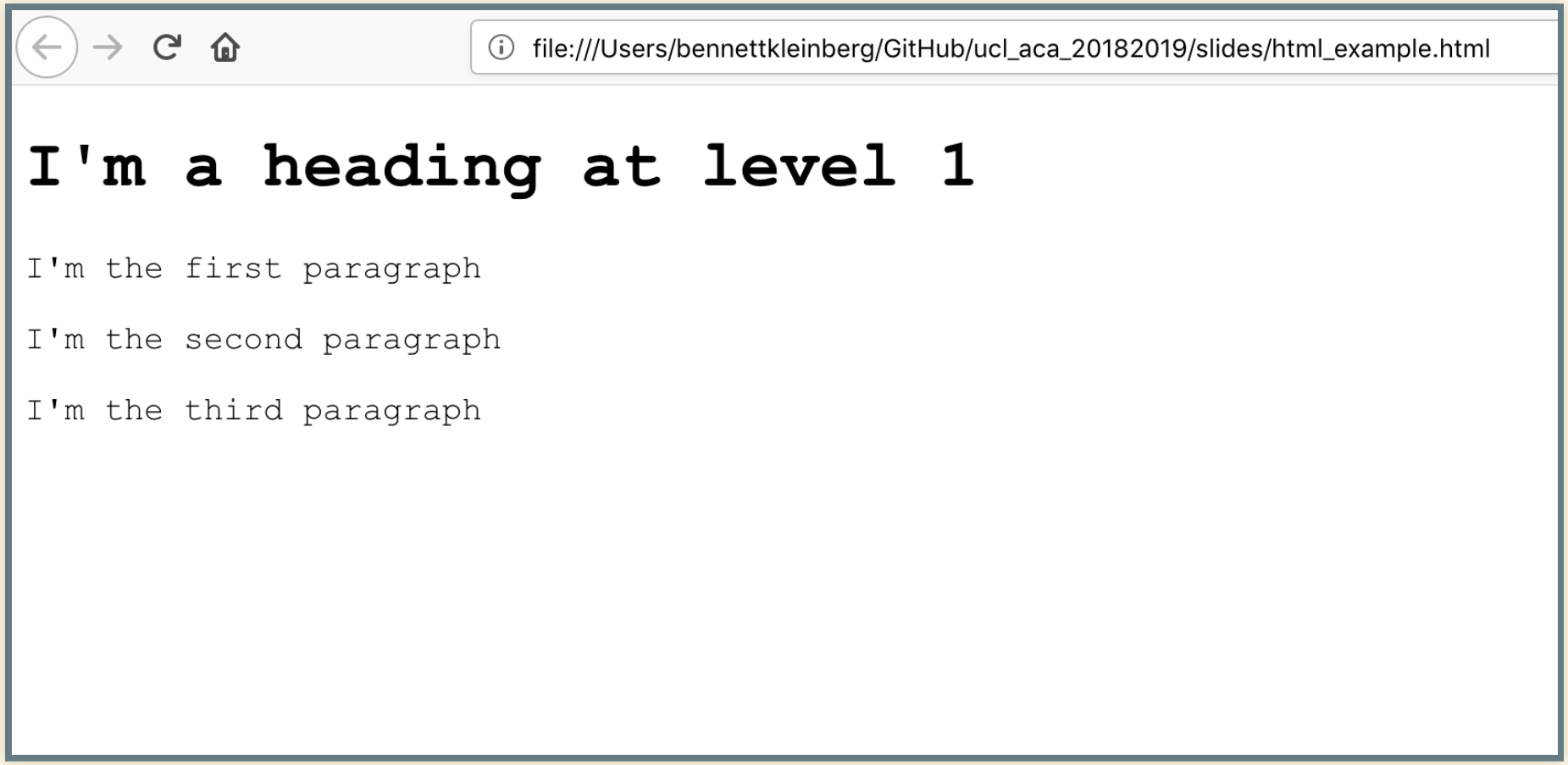
Ways to put content in the `<body> ... </body>` tag:

- headings: `<h1>I'm a heading at level 1</>`



CONTENT IN THE BODY TAG

- paragraphs: `<p>This is a paragraph</p>`



CONTENT IN THE BODY TAG

- images: `<img src="../img/ucl.jpg">`

file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html

I'm a heading at level 1

I'm the first paragraph

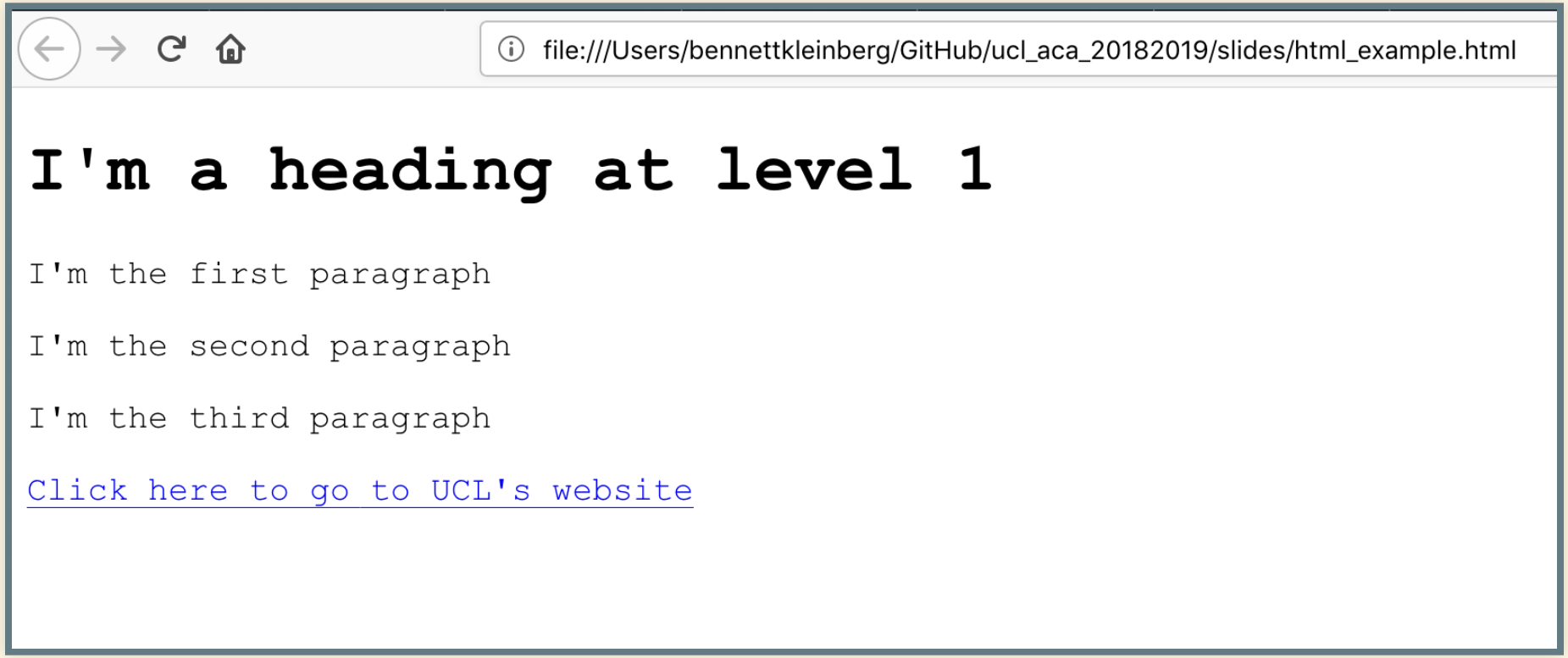
I'm the second paragraph

I'm the third paragraph

CONTENT IN THE BODY TAG

- links:

```
<a href="https://www.ucl.ac.uk/">Click here t
```



WEB SCRAPING LOGIC

If all webpages are built in this structure…

… then we could access this structure programmatically.

BUT WHERE DO I FIND THAT STRUCTURE?

Is it just “there”?

YES!!

HOW TO SEE THE HTML STRUCTURE?

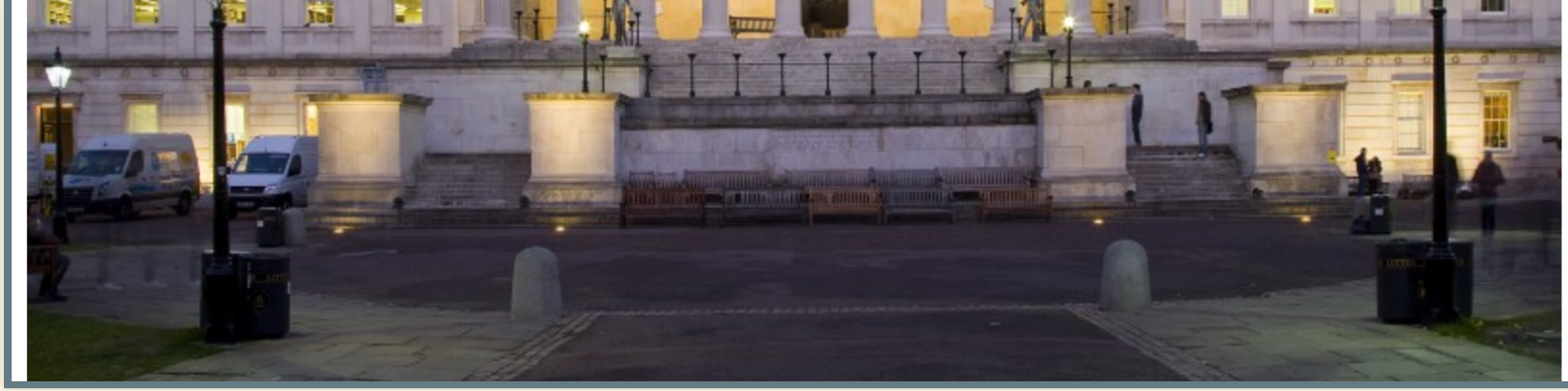
file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html

I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

I'm the third paragraph



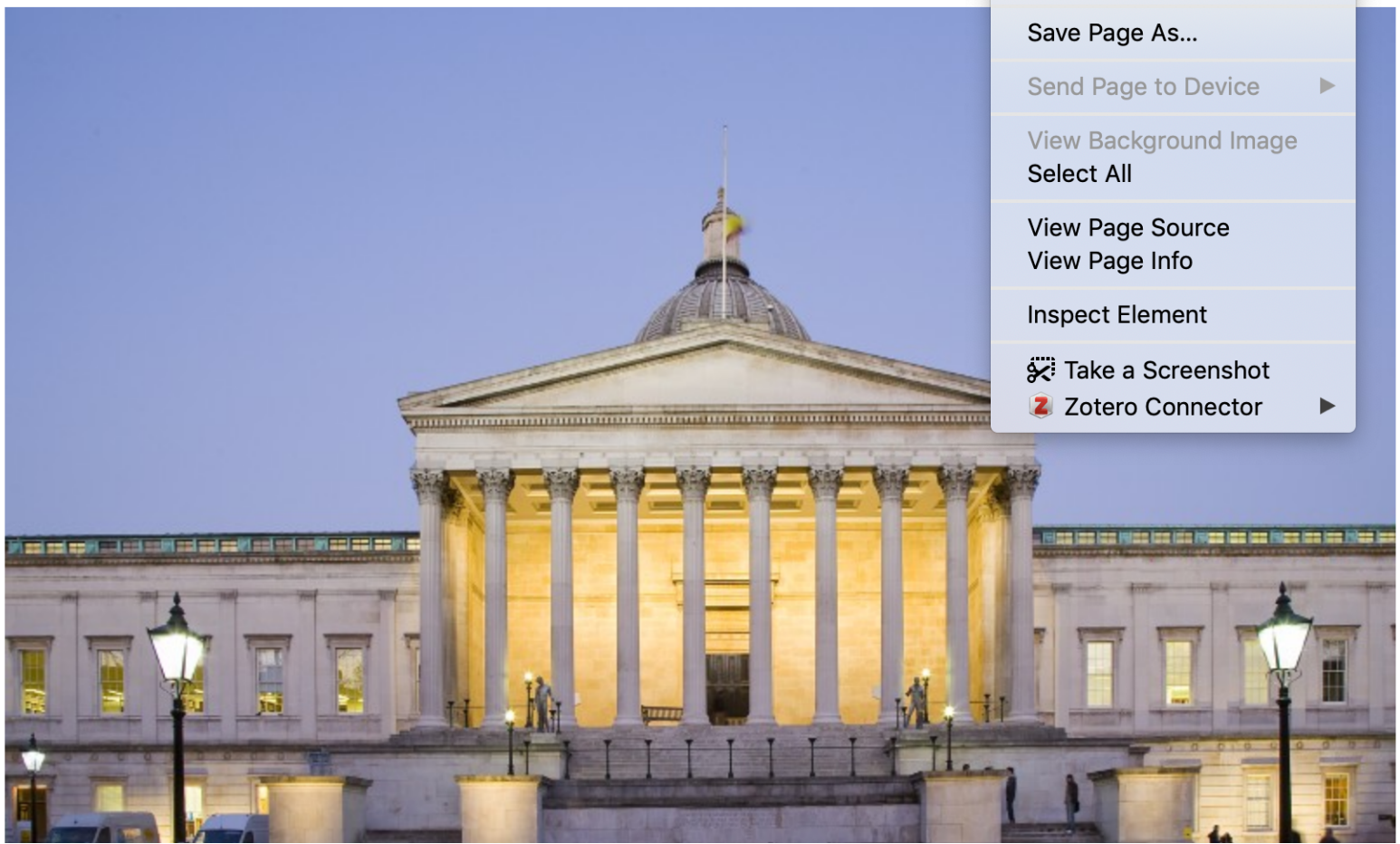
file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html

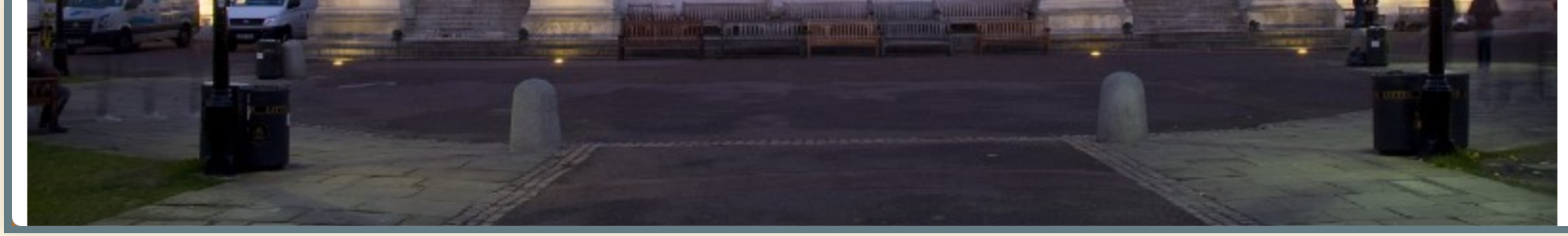
I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

I'm the third paragraph





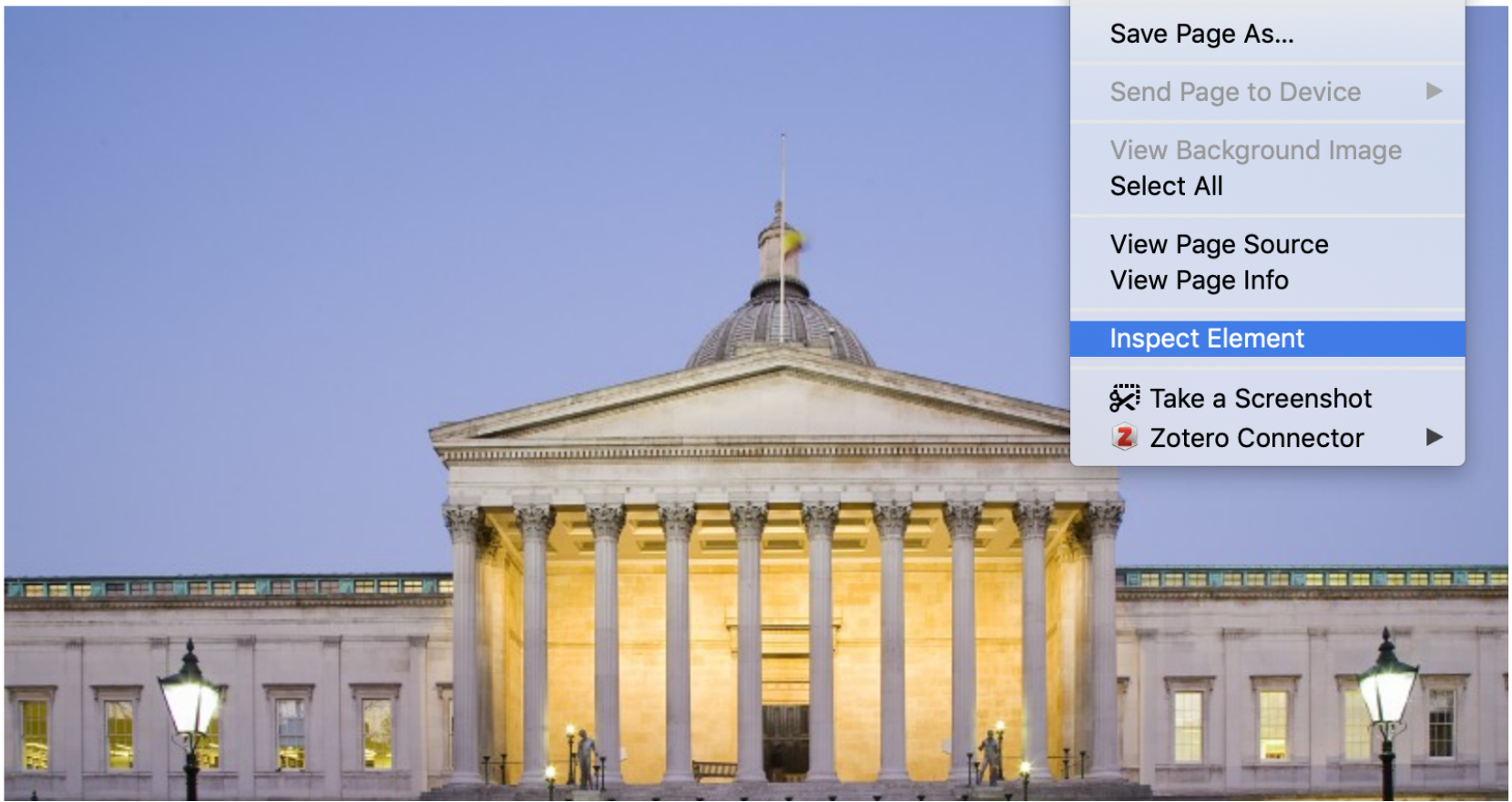
file:///Users/bennettkleinberg/GitHub/ucl_aca_20182019/slides/html_example.html

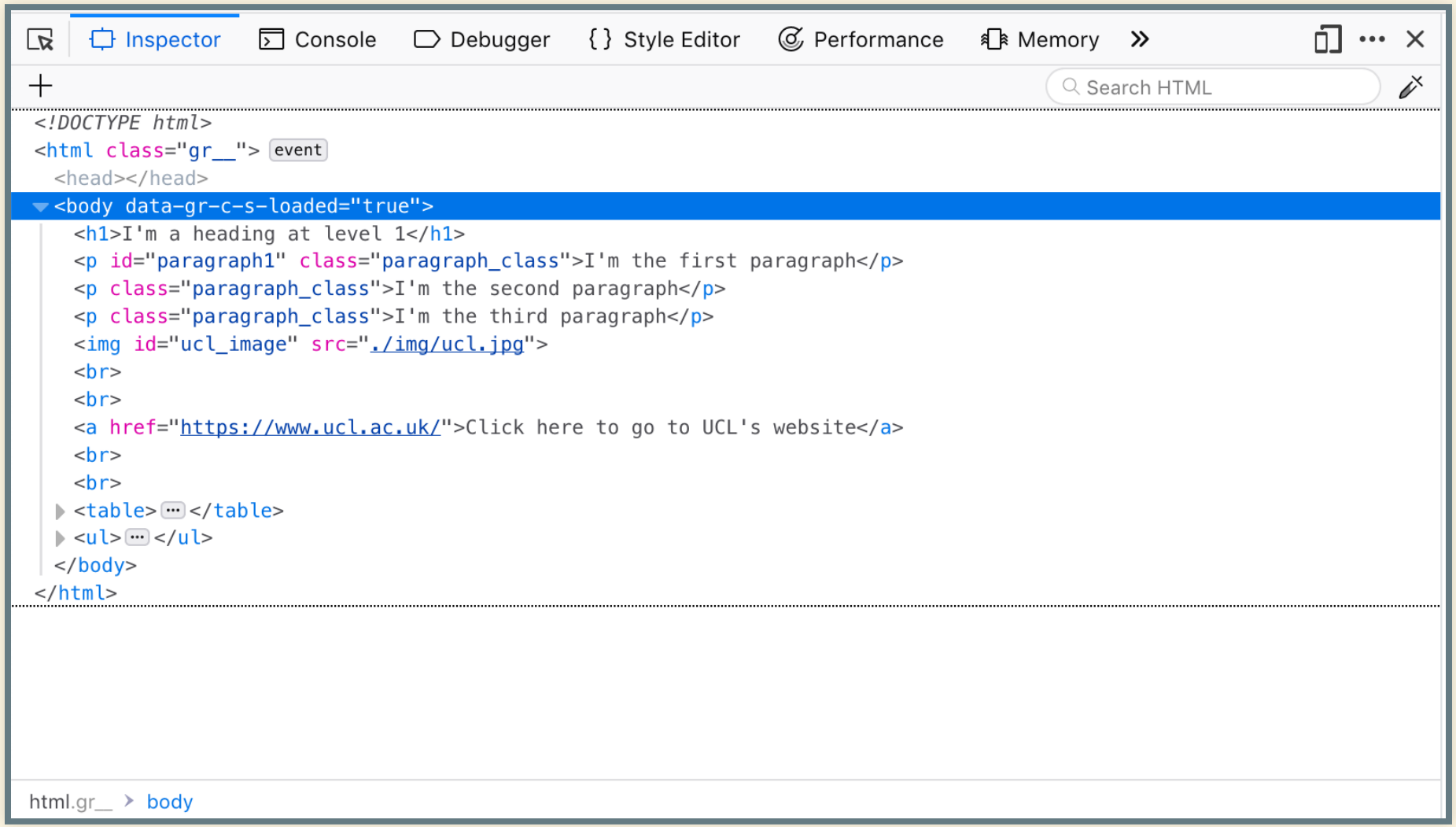
I'm a heading at level 1

I'm the first paragraph

I'm the second paragraph

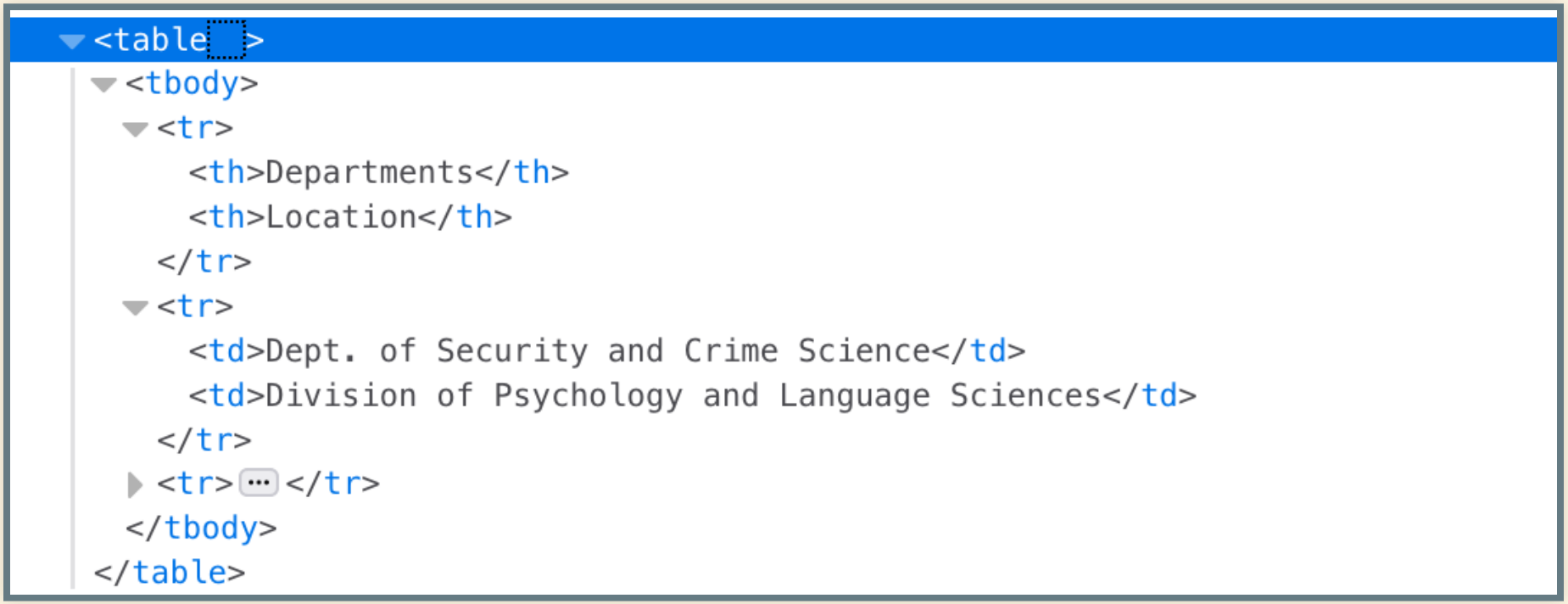
I'm the third paragraph




Inspector
Console
Debugger
Style Editor
Performance
Memory
Search HTML
<!DOCTYPE html>
<html class="gr__"> event
<head></head>
<body data-gr-c-s-loaded="true">
<h1>I'm a heading at level 1</h1>
I'm the first paragraph
I'm the second paragraph
I'm the third paragraph
<img id="ucl_image" src="./img/ucl.jpg">

Click here to go to UCL's website

<table> ··· </table>
 ···
</body>
</html>
html.gr__ > body


<table>
<tbody>
<tr>
<th>Departments</th>
<th>Location</th>
</tr>
<tr>
<td>Dept. of Security and Crime Science</td>
<td>Division of Psychology and Language Sciences</td>
</tr>
<tr>…</tr>
</tbody>
</table>

WEBSCRAPING IN A NUTSHELL

1. understand the structure of a webpage
2. exploit that structure for web-scraping

WHAT'S NEXT?

- Today's tutorial: building your own API scraper, using the data, some HTML
- Homework: arxiv scraping, Twitter access, project start

Next week: Web scraping 2 (more HTML, browser simulation)