Cogs 9 Discussion Section

FA22 Week 3
Will McCarthy

Upcoming due dates

This Thursday October 13th

Reading Quiz 2

This Friday Oct 14th:

Group Assignment 1

Assignment 2 due Oct 28th

This week's content

Reading 2(b): Narayanan and Shmatikov, Privacy & Security Myths & Fallacies of "PII"

- Data merging demo

Collecting Data

Data collection demo

Reading 2(b): Narayanan and Shmatikov, Privacy & Security Myths & Fallacies of “PII”

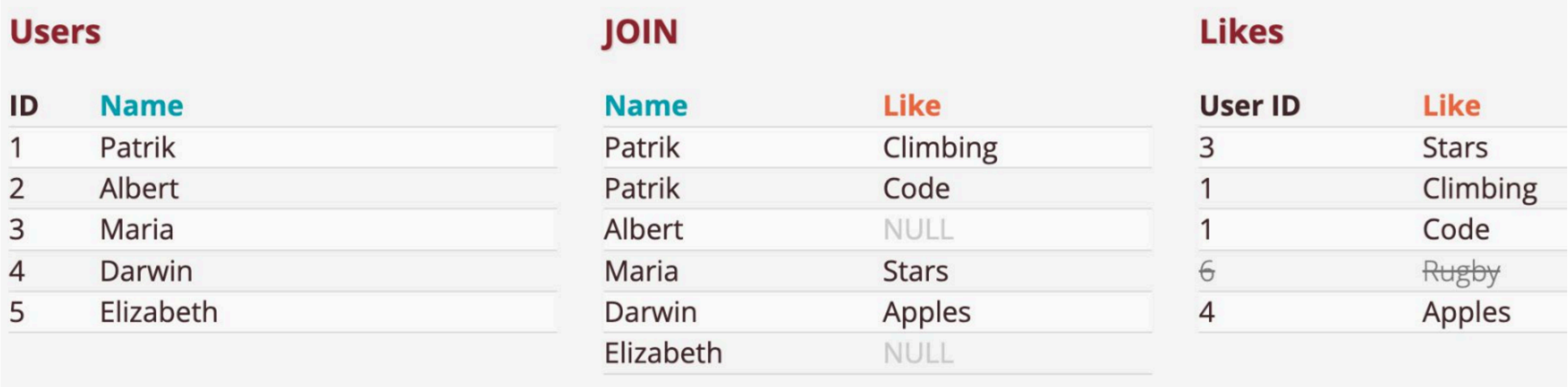
What is Personally Identifiable Information (PII)?

A couple of definitions exist: *breach notification* law and *privacy* law

Privacy sense is broader: “Identified, **directly** or **indirectly**” (p 25)

“account should be taken of all the means likely reasonably to be used either by the controller a or by any other person to identify the said person.” (p 25)

Identification by combining data from multiple sources (joins)



Users

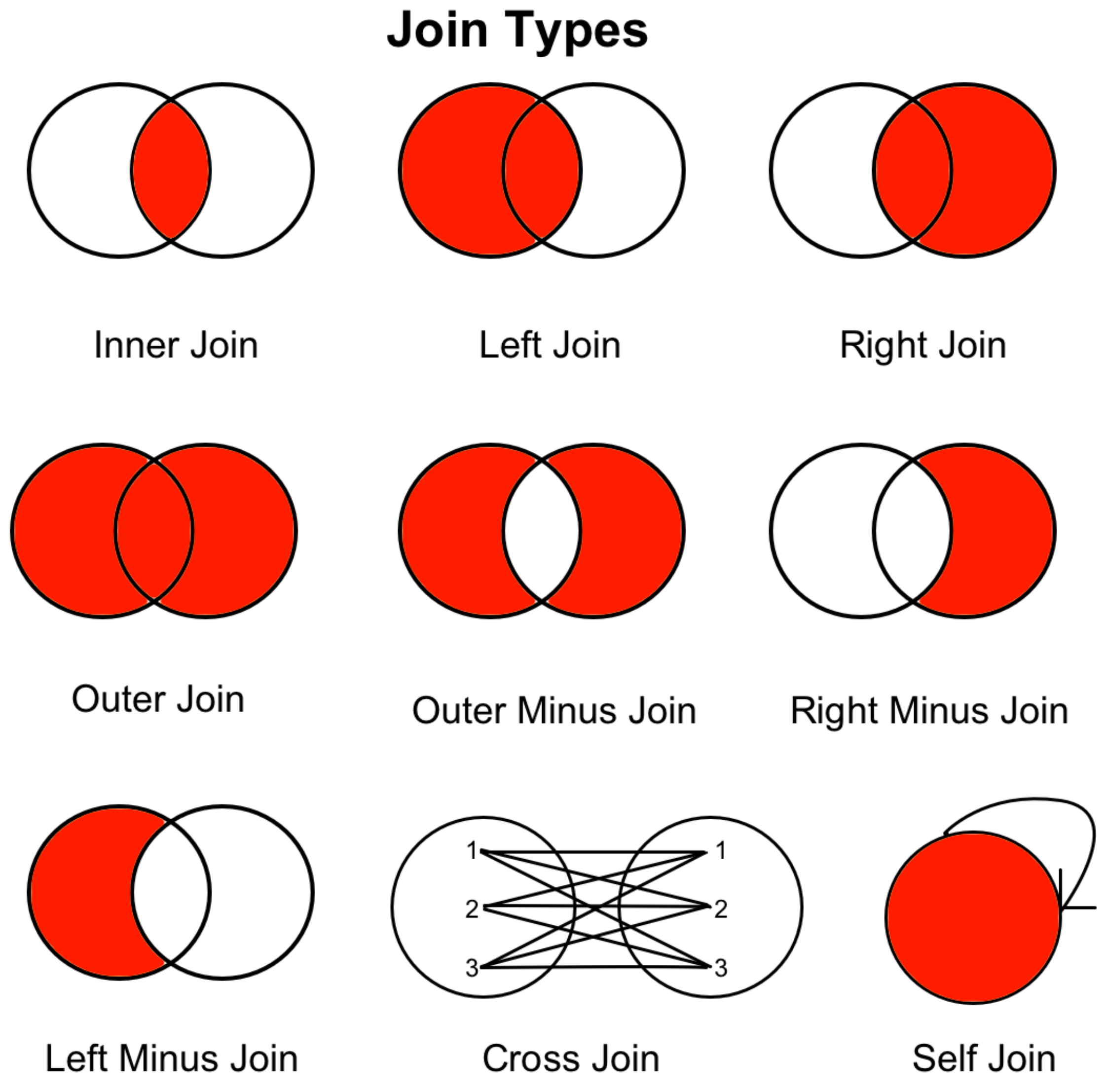
ID	Name
1	Patrik
2	Albert
3	Maria
4	Darwin
5	Elizabeth

JOIN

Name	Like
Patrik	Climbing
Patrik	Code
Albert	NULL
Maria	Stars
Darwin	Apples
Elizabeth	NULL

Likes

User ID	Like
3	Stars
1	Climbing
1	Code
~~6~~	~~Rugby~~
4	Apples



datascienceexamples.com

Data Merging Demo

Takeaway: PII not just about the data you’re collecting, it’s how that data can be used in combination with other data

Reading 2(b): Narayanan and Shmatikov, Privacy & Security Myths & Fallacies of "PII"

Strategies for dealing with this

k-anonymity: "every combination of quasi-identifier values occurring in the dataset must occur at least k times."

> "It turns out there is a wide spectrum of human characteristics that enable **re-identification**: consumption preferences, commercial transactions, Web browsing, search histories, and so forth. Their two key properties are that
>
> (1) they are reasonably stable across time and contexts, and
>
> (2) the corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability." (P26)

Reading 2(b): Narayanan and Shmatikov, Privacy & Security Myths & Fallacies of "PII"

'"personally identifiable" and "quasi-identifier" simply have no technical meaning.' (p26)

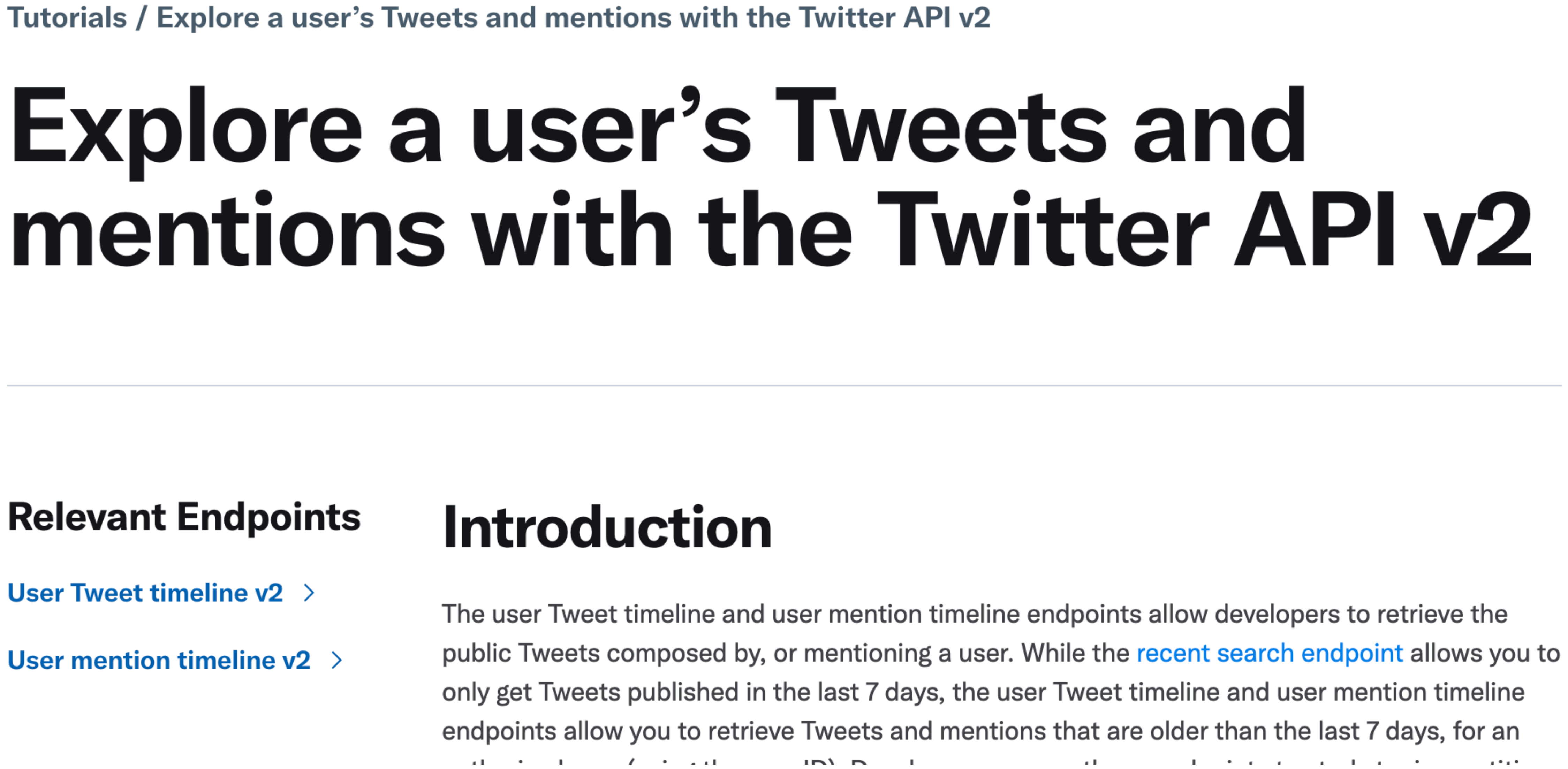
"*any attribute can be identifying in combination with others.*" (p26)

Activity: discuss how people could be identified with the data you use for your project (directly or indirectly). What other datasets could be *joined with* yours to identify people?

Three common ways of collecting data from the web

1. Download a **dataset** from a **website** (easiest)
2. Use an **API** to request data from a server (reqs coding knowledge)
3. Use **web scraping** techniques to download data from a website (reqs coding + html)

2



Tutorials / Explore a user's Tweets and mentions with the Twitter API v2

Explore a user's Tweets and mentions with the Twitter API v2

Relevant Endpoints

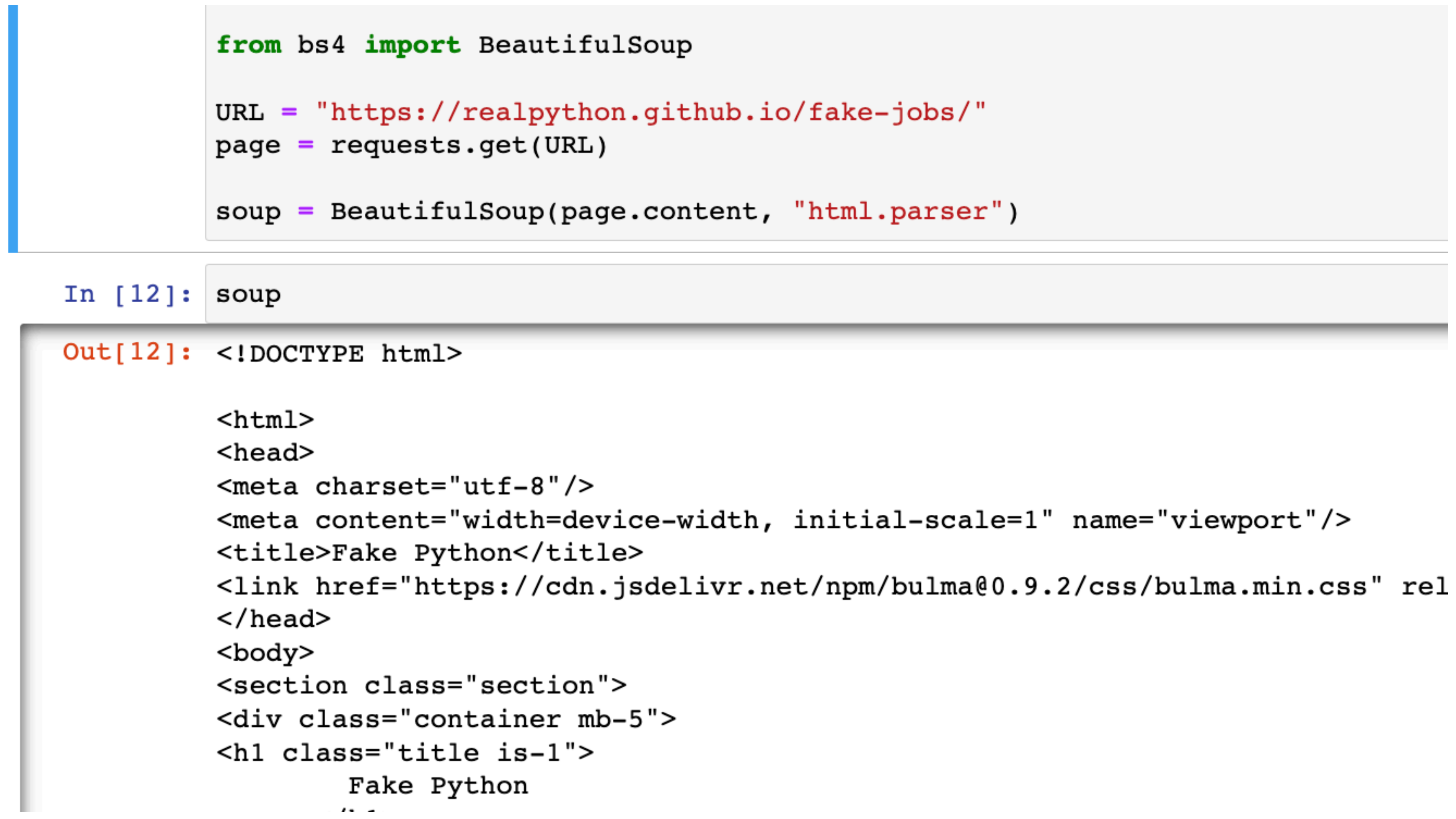
User Tweet timeline v2 ›

User mention timeline v2 ›

Introduction

The user Tweet timeline and user mention timeline endpoints allow developers to retrieve the public Tweets composed by, or mentioning a user. While the recent search endpoint allows you to only get Tweets published in the last 7 days, the user Tweet timeline and user mention timeline endpoints allow you to retrieve Tweets and mentions that are older than the last 7 days, for an

3



```
from bs4 import BeautifulSoup

URL = "https://realpython.github.io/fake-jobs/"
page = requests.get(URL)

soup = BeautifulSoup(page.content, "html.parser")
```

```
In [12]: soup

Out[12]: <!DOCTYPE html>

         <html>
         <head>
         <meta charset="utf-8"/>
         <meta content="width=device-width, initial-scale=1" name="viewport"/>
         <title>Fake Python</title>
         <link href="https://cdn.jsdelivr.net/npm/bulma@0.9.2/css/bulma.min.css" rel
         </head>
         <body>
         <section class="section">
         <div class="container mb-5">
         <h1 class="title is-1">
                 Fake Python
```

Web Scraping Demo

1. Find a website you might be able to scrape to answer your data science question.
2. Can you identify the html element of the website you would scrape to collect that data?

Next week: Data Wrangling

Readings on Tidy Data and Spreadsheets.

Tidy Data demo.

Group work / questions

Future Readings

3(a): Hadley Wickham, 2014, Tidy Data

3(b): Broman & Woo, 2018, Data organization in spreadsheets

4(a): Evan M. Peck, et al., 2019, Attitudes and Perceptions of Data Visualization

4(b): Hadley Wickham, et al., 2010, Graphical Inference for Infovis

5(a): Nicholas Diakopoulos, 2016, Accountability in Algorithmic Decision Making

5(b): Julia Angwin, et al., 2016, Machine Bias