# Web scraping with `rvest`

Prof. Aarushi Sahejpal
School of Communication, American University

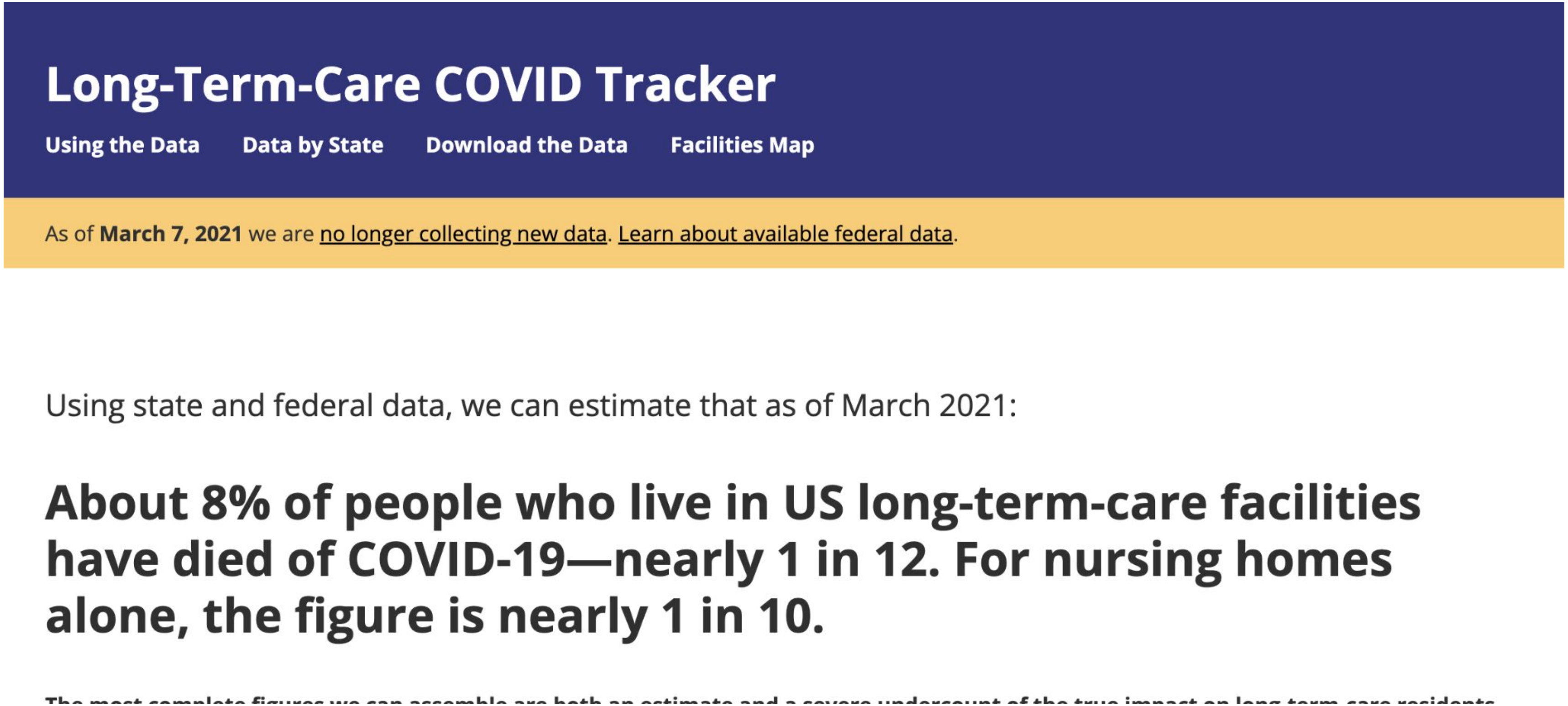
The COVID Tracking Project was cited in **more than 1,000 academic papers**, including major medical journals like *The New England Journal of Medicine*, *Nature*, and *JAMA*.

We received awards for our work from the **Society of Professional Journalists**, the **Sigma Awards**, and the **NYU Journalism Online Awards**.

Our data was used by **two presidential administrations** and an array of federal agencies, including the CDC, HHS, and FDA.

**Federal lawmakers** used our data in at least 11 letters demanding answers on the pandemic response from government leaders and commercial labs.

And our data was cited in **over 7,700 news stories** in publications including *The New York Times*, *The Washington Post*, *CNN*, *Vox*, *ProPublica*, and many more.



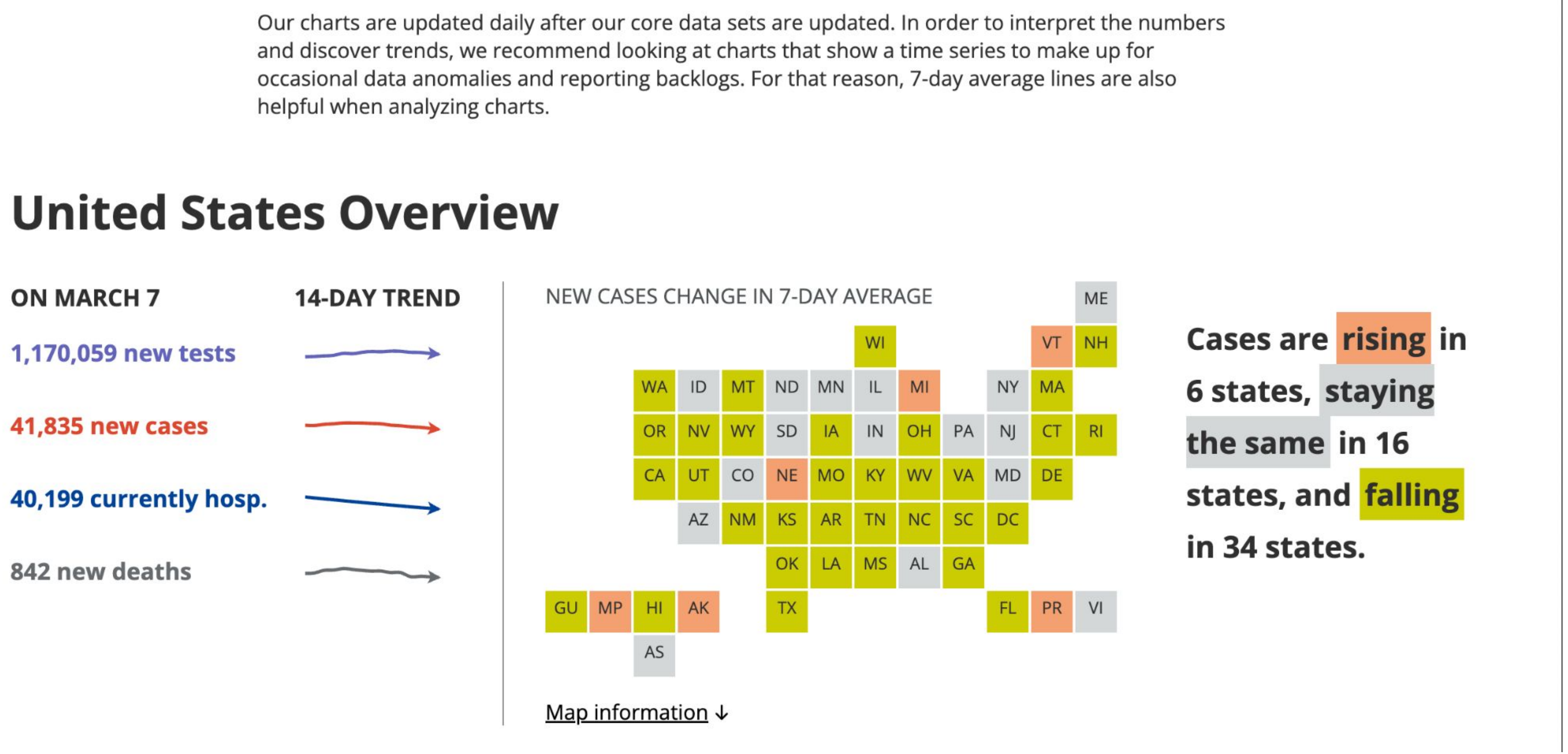
# Long-Term-Care COVID Tracker

As of **March 7, 2021** we are no longer collecting new data. Learn about available federal data.

Using state and federal data, we can estimate that as of March 2021:

**About 8% of people who live in US long-term-care facilities have died of COVID-19—nearly 1 in 12. For nursing homes alone, the figure is nearly 1 in 10.**

**The most complete figures we can assemble are both an estimate and a severe undercount of the true impact on long-term-care residents**



Our charts are updated daily after our core data sets are updated. In order to interpret the numbers and discover trends, we recommend looking at charts that show a time series to make up for occasional data anomalies and reporting backlogs. For that reason, 7-day average lines are also helpful when analyzing charts.

# United States Overview

| ON MARCH 7 | 14-DAY TREND |
|---|---|
| 1,170,059 new tests | → |
| 41,835 new cases | → |
| 40,199 currently hosp. | → |
| 842 new deaths | → |

NEW CASES CHANGE IN 7-DAY AVERAGE

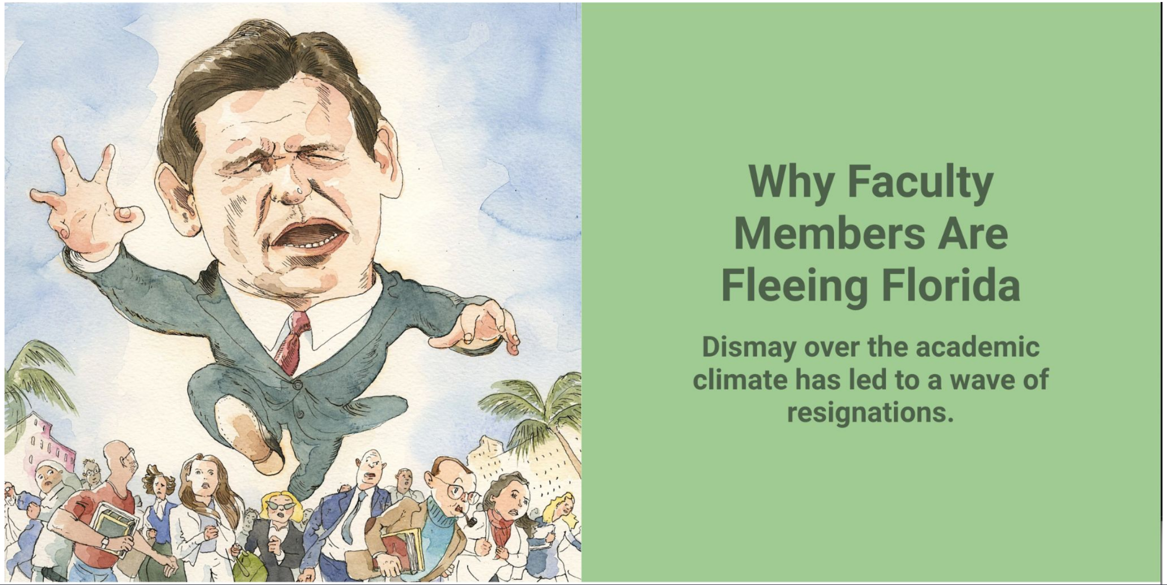
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | ⬜ ME |
| | | | | | | | 🟨 WI | | | | 🟧 VT | 🟨 NH |
| | | 🟨 WA | ⬜ ID | 🟨 MT | ⬜ ND | ⬜ MN | ⬜ IL | 🟧 MI | | ⬜ NY | 🟨 MA | |
| | | 🟨 OR | 🟨 NV | 🟨 WY | ⬜ SD | 🟨 IA | ⬜ IN | 🟨 OH | ⬜ PA | ⬜ NJ | 🟨 CT | 🟨 RI |
| | | 🟨 CA | 🟨 UT | ⬜ CO | 🟧 NE | 🟨 MO | 🟨 KY | 🟨 WV | 🟨 VA | ⬜ MD | 🟨 DE | |
| | | | ⬜ AZ | 🟨 NM | 🟨 KS | 🟨 AR | 🟨 TN | 🟨 NC | 🟨 SC | 🟨 DC | | |
| | | | | | 🟨 OK | 🟨 LA | 🟨 MS | ⬜ AL | 🟨 GA | | | |
| 🟨 GU | 🟧 MP | 🟨 HI | 🟧 AK | | 🟨 TX | | | | | 🟨 FL | 🟧 PR | ⬜ VI |
| | | ⬜ AS | | | | | | | | | | |

Map information ↓

**Cases are 🟧 rising in 6 states, ⬜ staying the same in 16 states, and 🟨 falling in 34 states.**

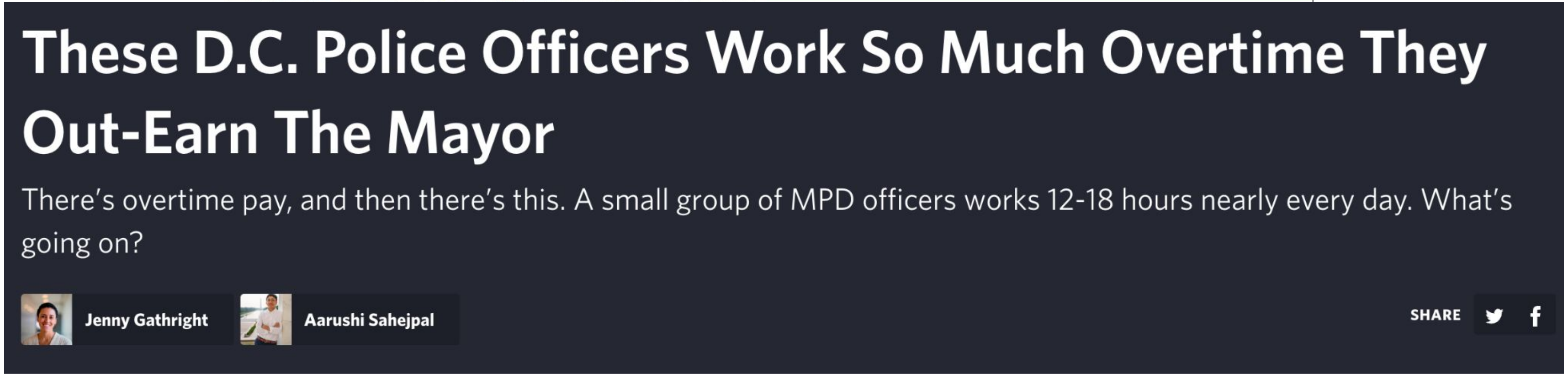
# Blueprint for an AI Bill of Rights

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE



## Why Faculty Members Are Fleeing Florida

Dismay over the academic climate has led to a wave of resignations.

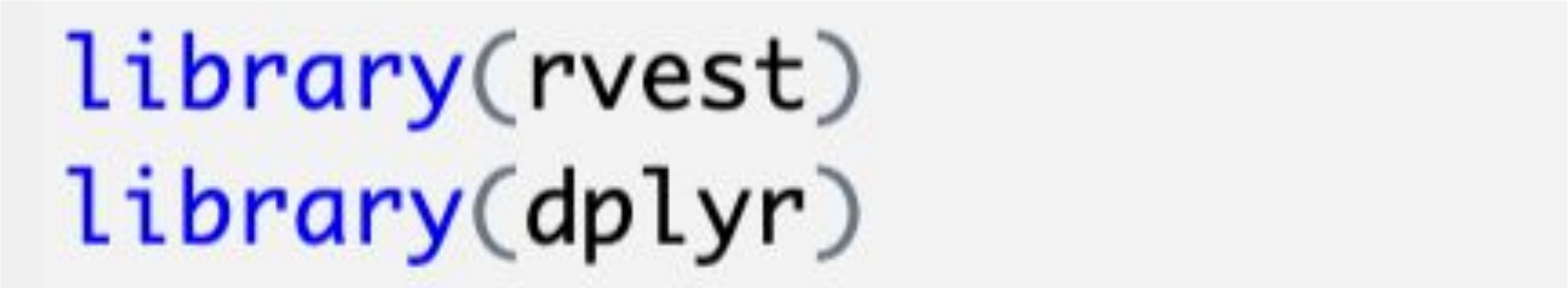


## These D.C. Police Officers Work So Much Overtime They Out-Earn The Mayor

There’s overtime pay, and then there’s this. A small group of MPD officers works 12-18 hours nearly every day. What’s going on?

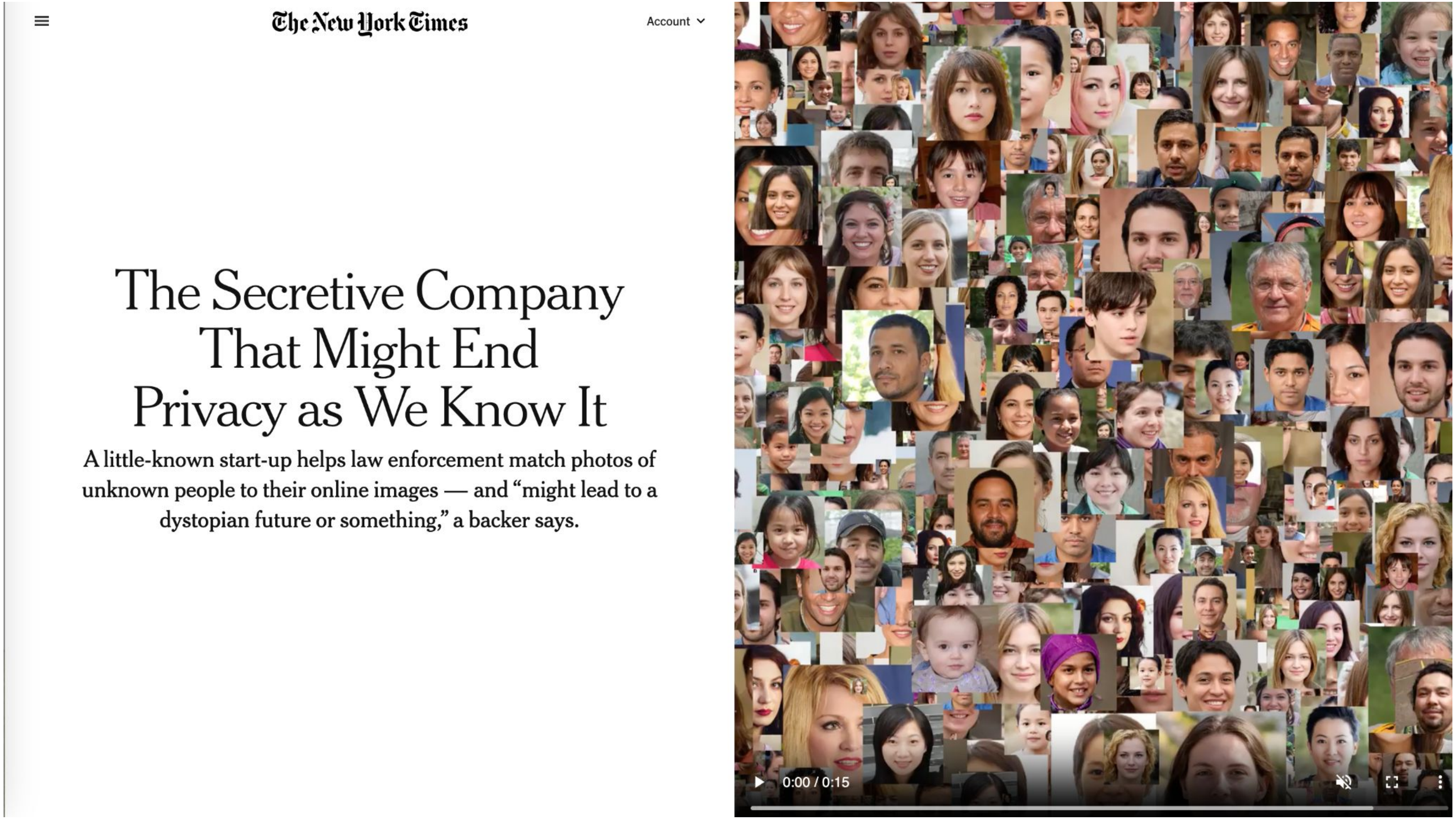
Jenny Gathright Aarushi Sahejpal

Web scraping is a very useful tool to extract data from webpages, that don’t allow easy access.



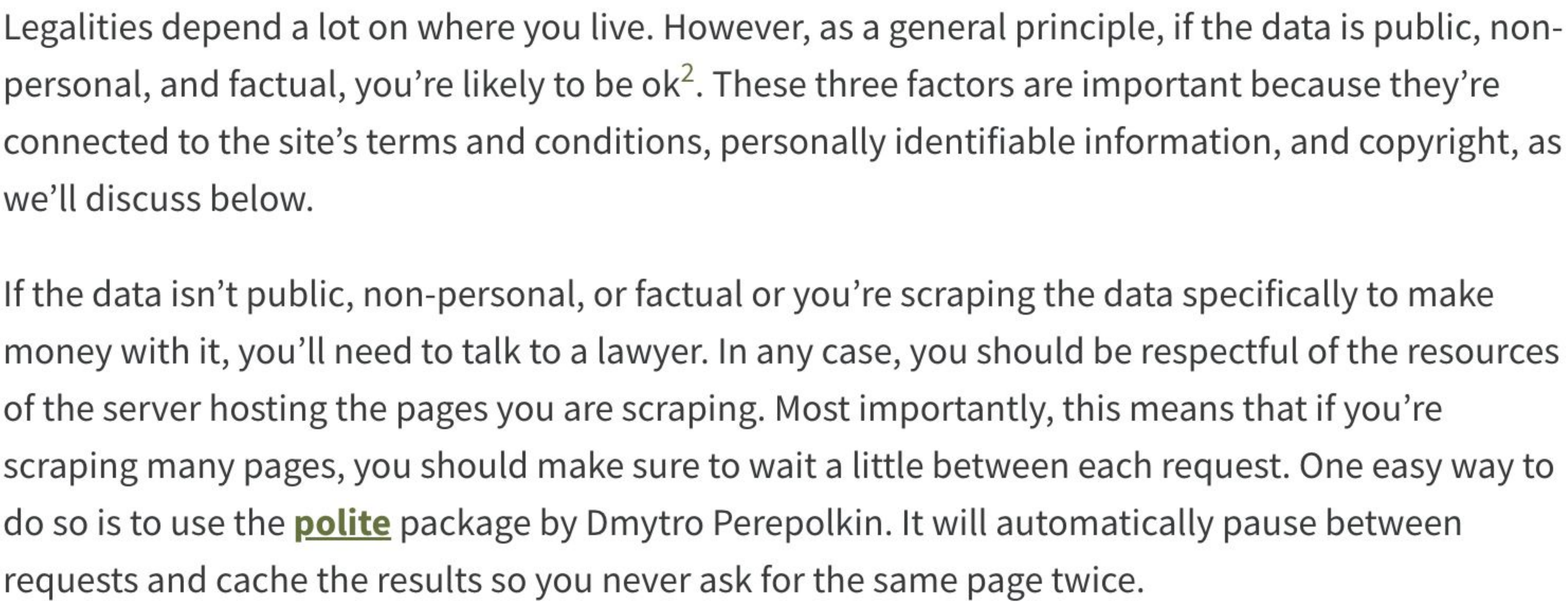
```r
library(rvest)
library(dplyr)
```

What is web scraping and why do some sites hate it..



# The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.



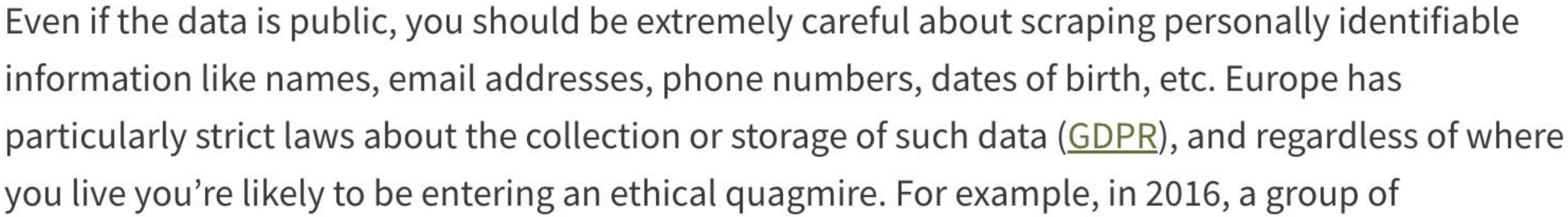
Legalities depend a lot on where you live. However, as a general principle, if the data is public, non-personal, and factual, you're likely to be ok[2]. These three factors are important because they're connected to the site's terms and conditions, personally identifiable information, and copyright, as we'll discuss below.

If the data isn't public, non-personal, or factual or you're scraping the data specifically to make money with it, you'll need to talk to a lawyer. In any case, you should be respectful of the resources of the server hosting the pages you are scraping. Most importantly, this means that if you're scraping many pages, you should make sure to wait a little between each request. One easy way to do so is to use the **polite** package by Dmytro Perepolkin. It will automatically pause between requests and cache the results so you never ask for the same page twice.

R for Data Science (2e) - Wickham et al.

# Terms and services

# Copyright



Even if the data is public, you should be extremely careful about scraping personally identifiable information like names, email addresses, phone numbers, dates of birth, etc. Europe has particularly strict laws about the collection or storage of such data (GDPR), and regardless of where you live you're likely to be entering an ethical quagmire. For example, in 2016, a group of

R for Data Science (2e) - Wickham et al.

# Why do journalists do it?

```
<html>
<head>
  <title>Page title</title>
</head>
<body>
  <h1 id='first'>A heading</h1>
  <p>Some text &amp; <b>some bold text.</b></p>
  <img src='myimg.png' width='100' height='100'>
</body>
```

R for Data Science (2e) -  Wickham et al.

# Rvest

# Let’s try it