Special Topics in Web Scraping with Python

Hi, I'm Isaac Vidas

Software Engineer @ HERE Technologies

🙉 Linkedin: in/isaac-vidas

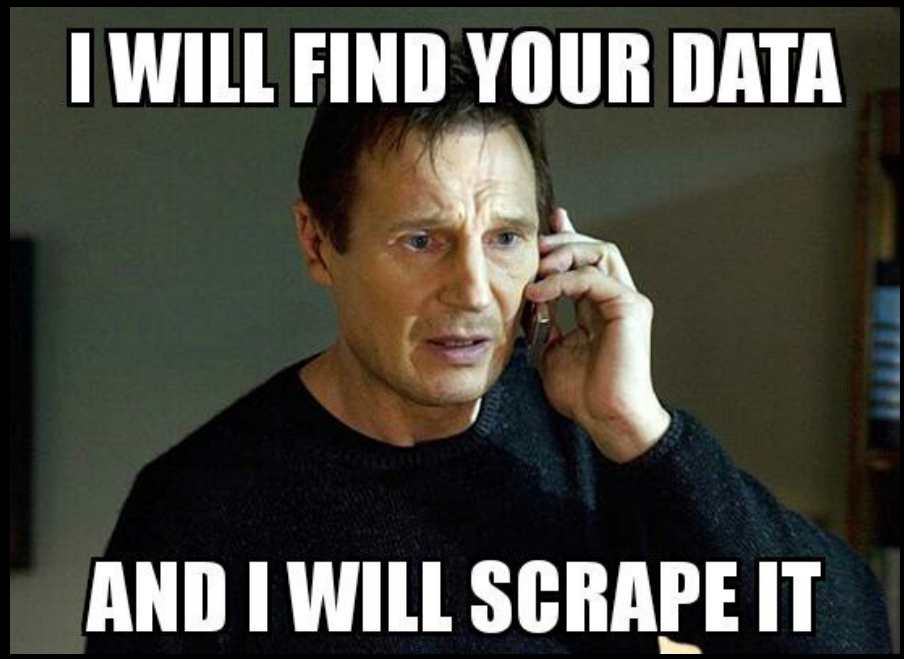
🙈 Twitter: @kazuarous

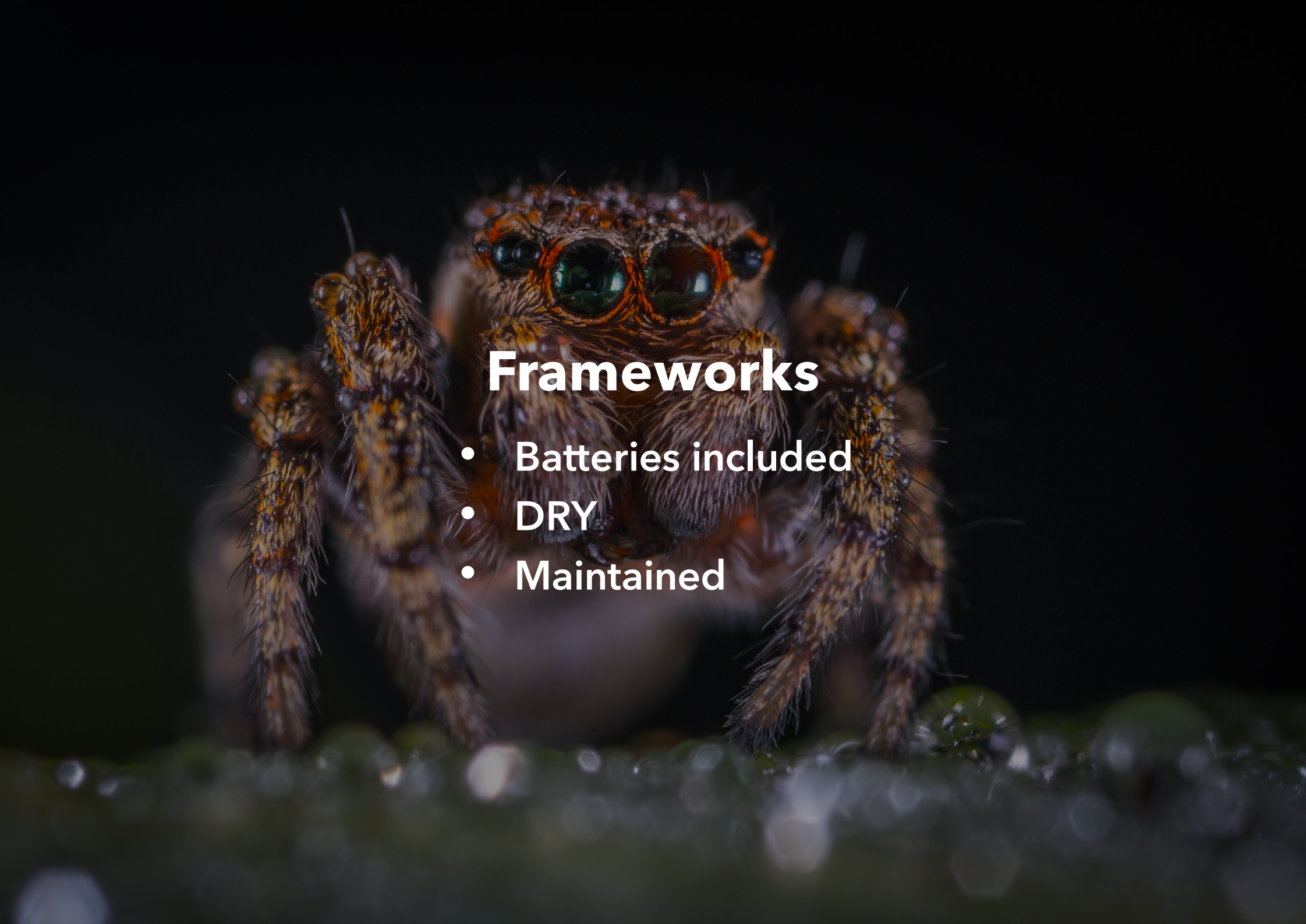
🙈 Blog: http://vid.as

🍌 GitHub: kazuar

Agenda

- Frameworks - why do we need them?
- Static vs. dynamic websites
- Don't be evil
- In production





Frameworks

- Batteries included
- DRY
- Maintained

Scrapy

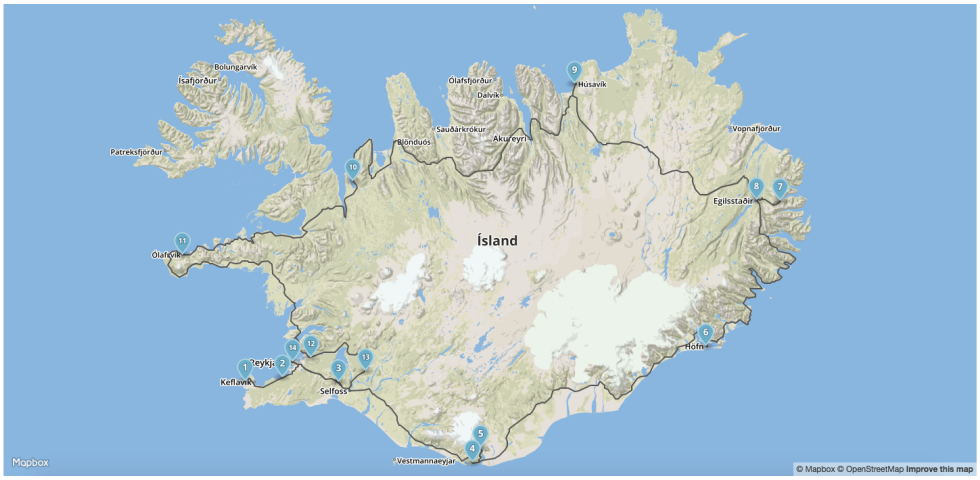
```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
  name = 'blogspider'
  start_urls = ['https://blog.scrapinghub.com']

  def parse(self, response):
    for title in response.css('.post-header>h2'):
      yield {'title': title.css('a ::text').get()}

    for next_page in response.css('a.next-posts-link'):
      yield response.follow(next_page, self.parse)
EOF
$ scrapy runspider myspider.py
```

Static vs. Dynamic

Isaac Vidas
This is not for you

Blog About

Visualize your trip with Flask and Mapbox



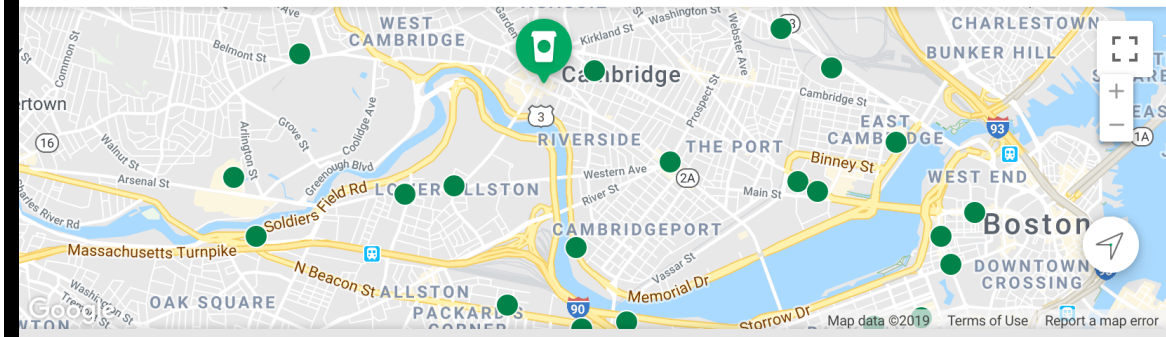
Following my recent trip to Iceland (which I highly recommend), I wanted to visualize the trip using a map, complete with the route and stop locations along the way.

Source code

Sign in

COFFEE TEA MENU COFFEEHOUSE SOCIAL IMPACT BLOG GIFT CARDS

Harvard Square, Cambridge, MA, USA Filter



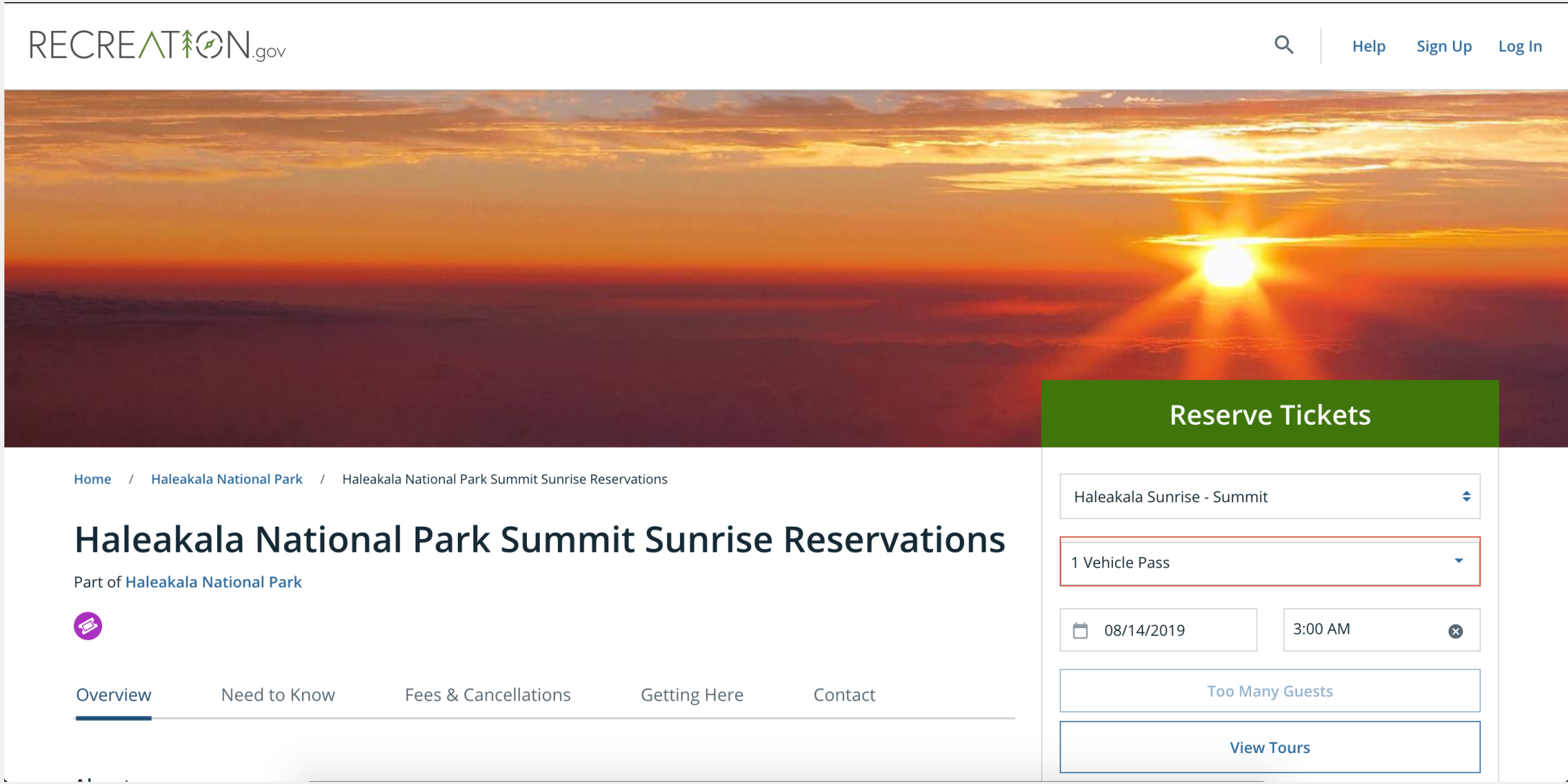
Harvard Yard
Open until 11:00 PM
0.5 miles away

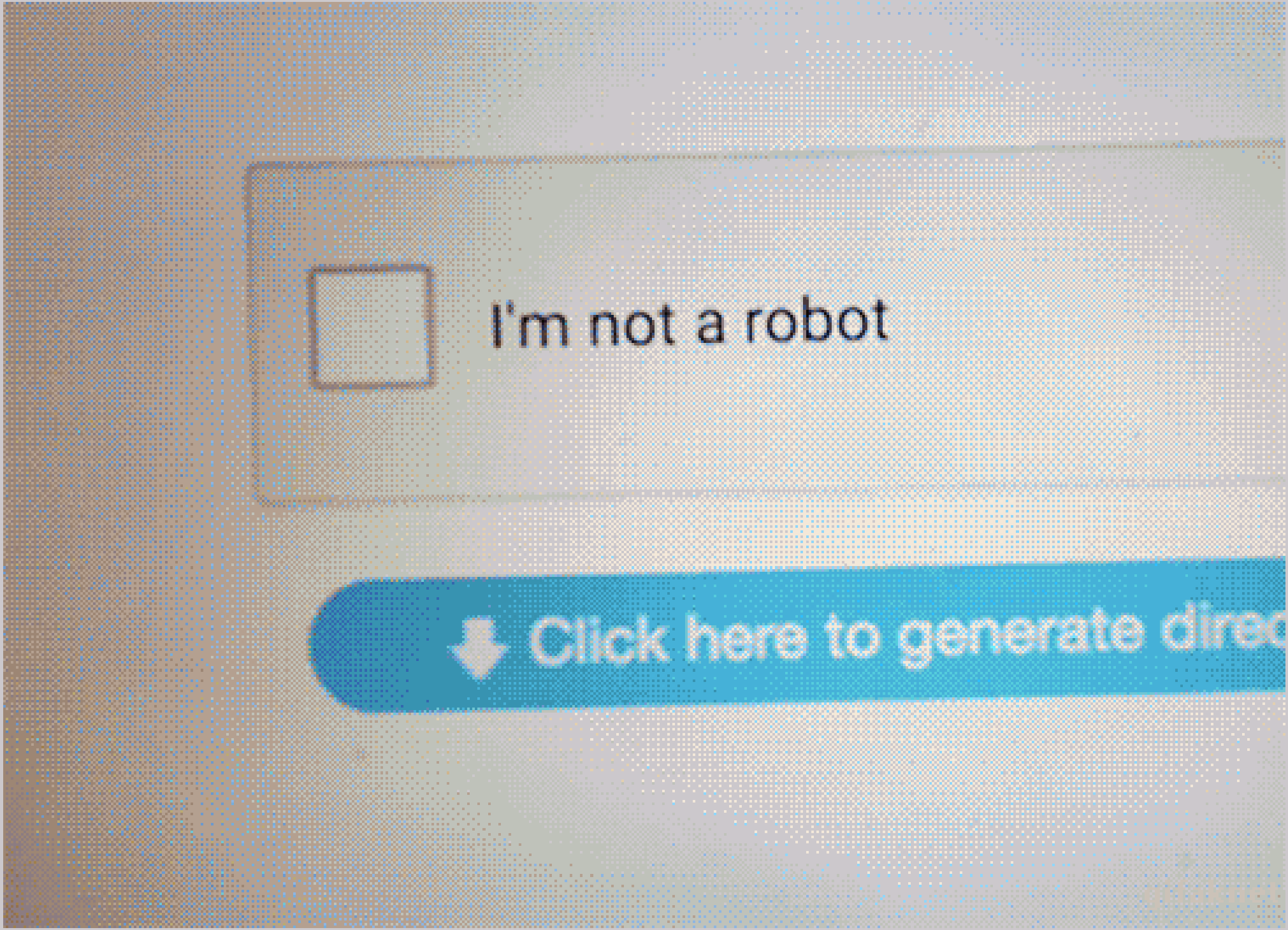
Broadway Marketplace
Open until 9:00 PM
0.6 miles away

Shepard Post
Open until 10:00 PM
0.1 miles away

Automate your browser with Selenium

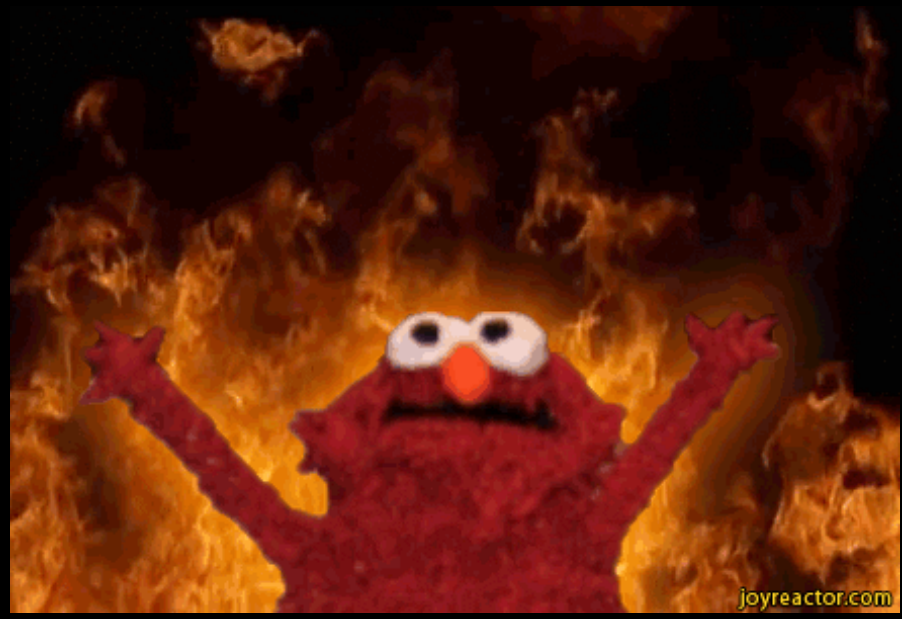
Too Many Guests




I'm not a robot
Click here to generate

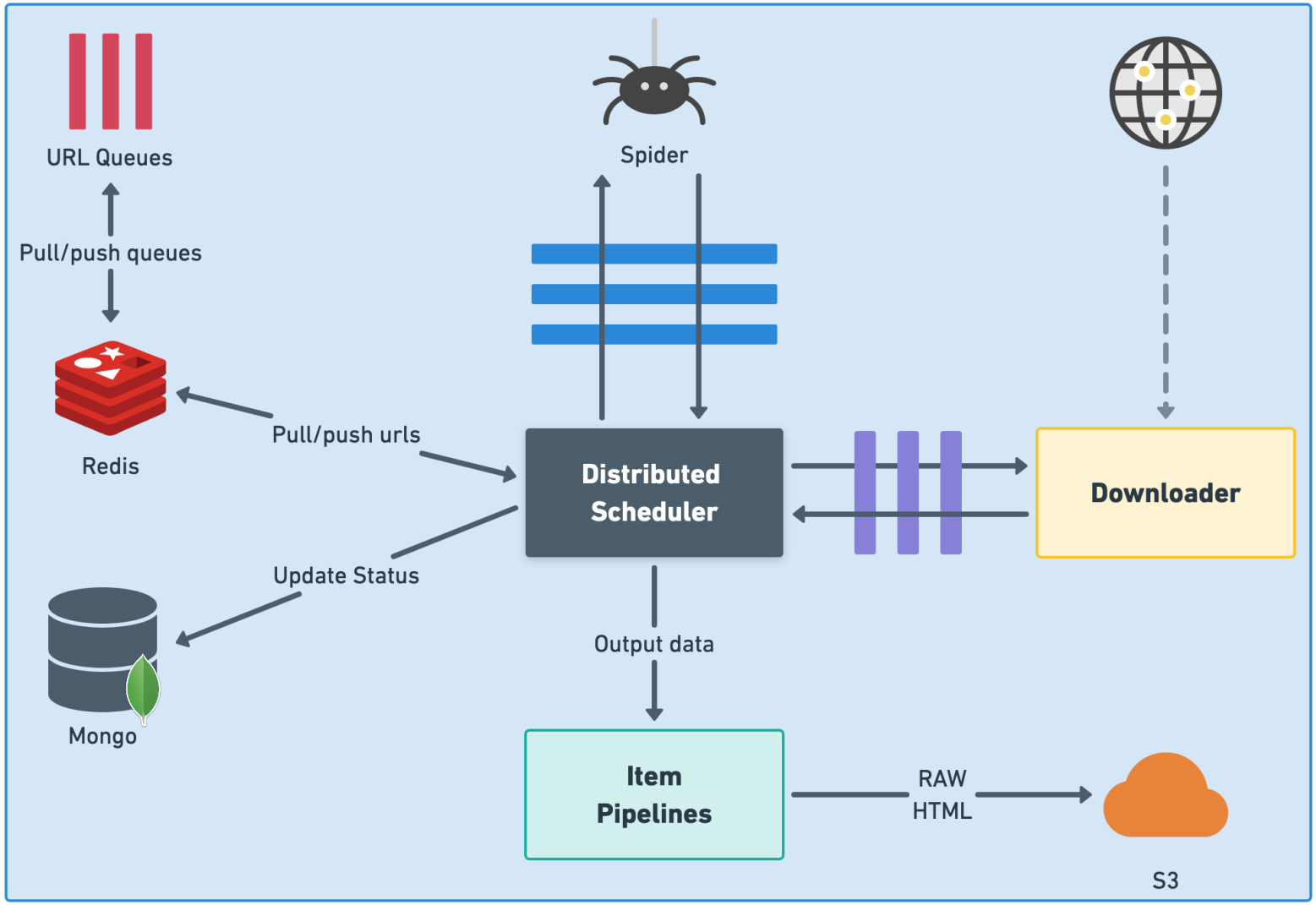
Don't Be Evil

- Politeness
- Robots.txt
- Hits per second
- Duplicate page filter
- User agent
- Alternatives to crawling



Web Scraping in Production

- Scaling
- Fault tolerance
- Monitoring
- Testing strategy
- Refresh policy



Thanks!

Resources

- Scrapy - https://scrapy.org/
- Selenium - https://www.seleniumhq.org/
- Login with Python - http://kazuar.github.io/scraping-tutorial/
- Crawl politely - https://blog.scrapinghub.com/2016/08/25/how-to-crawl-the-web-politely-with-scrapy