Importing flat files from the web

You're already great at importing!

- Flat files such as .txt and .csv

- Pickled files, Excel spreadsheets, and many others!

- Data from relational databases

- You can do all these locally

- What if your data is online?

Can you import web data?

- You can: go to URL and click to download files

- BUT: not reproducible, not scalable

You'll learn how to...

- Import and locally save datasets from the web

- Load datasets into pandas DataFrames

- Make HTTP requests (GET requests)

- Scrape web data such as HTML

- Parse HTML into useful data (BeautifulSoup)

- Use the urllib and requests packages

The urllib package

- Provides interface for fetching data across the web

- `urlopen()` - accepts URLs instead of file names

How to automate file download in Python

```
In [1]: from urllib.request import urlretrieve

In [2]: url = 'http://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv'

In [3]: urlretrieve(url, 'winequality-white.csv')
Out[3]: ('winequality-white.csv', <http.client.HTTPMessage at
0x103cf1128>)
```

IMPORTING DATA IN PYTHON II

Let's practice!

HTTP requests to import files from the web

URL

- Uniform/Universal Resource Locator

- References to web resources

- Focus: web addresses

- Ingredients:

  - Protocol identifier – http:

  - Resource name – datacamp.com

- These specify web addresses uniquely

HTTP

- HyperText Transfer Protocol

- Foundation of data communication for the web

- HTTPS – more secure form of HTTP

- Going to a website = sending HTTP request

  - GET request

- `urlretrieve()` performs a GET request

- HTML – HyperText Markup Language

GET requests using urllib

```
In [1]: from urllib.request import urlopen, Request

In [2]: url = "https://www.wikipedia.org/"

In [3]: request = Request(url)

In [4]: response = urlopen(request)

In [5]: html = response.read()

In [6]: response.close()
```

GET requests using requests

- Used by "her Majesty's Government, Amazon, Google, Twilio, NPR, Obama for America, Twitter, Sony, and Federal U.S. Institutions that prefer to be unnamed"

GET requests using requests

- One of the most downloaded Python packages

```
In [1]: import requests

In [2]: url = "https://www.wikipedia.org/"

In [3]: r = requests.get(url)

In [4]: text = r.text
```

IMPORTING DATA IN PYTHON II

Let's practice!

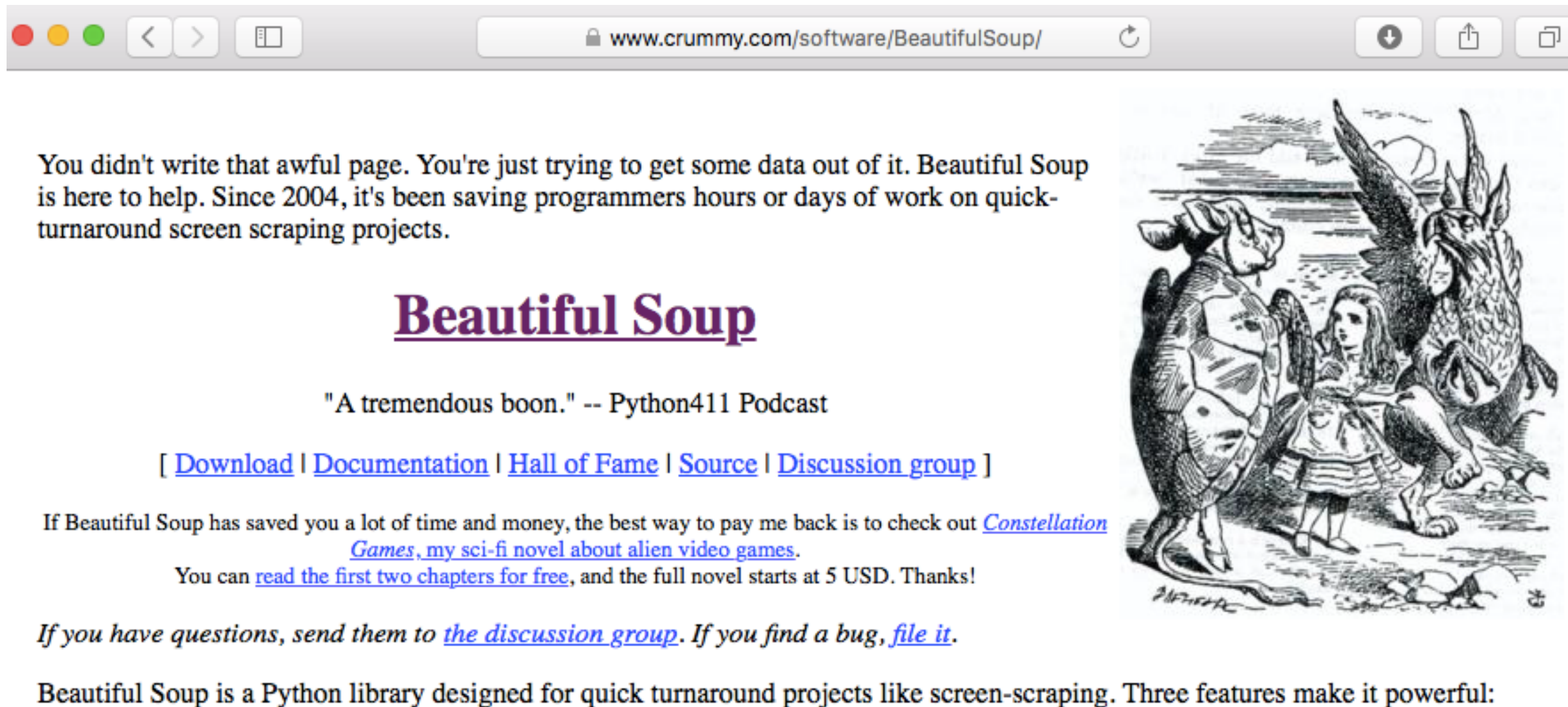# Scraping the web in Python

HTML

- Mix of unstructured and structured data

- Structured data:

    - Has pre-defined data model, or

    - Organized in a defined manner

- Unstructured data: neither of these properties

```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <!-- SEO -->
5      <meta charset="utf-8" />
6  <title>DataCamp: The Easy Way To Learn R &amp; Data Science Online</title>
```

BeautifulSoup

- Parse and extract structured data from HTML

- Make tag soup beautiful and extract information

BeautifulSoup

```
In [1]: from bs4 import BeautifulSoup

In [2]: import requests

In [3]: url = 'https://www.crummy.com/software/BeautifulSoup/'

In [4]: r = requests.get(url)

In [5]: html_doc = r.text

In [6]: soup = BeautifulSoup(html_doc)
```

Prettified Soup

```
In [7]: print(soup.prettify())
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40
<html>
 <head>
  <meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>
  <title>
   Beautiful Soup: We called him Tortoise because he taught us.
  </title>
  <link href="mailto:leonardr@segfault.org" rev="made"/>
  <link href="/nb/themes/Default/nb.css" rel="stylesheet" type="text/css"/>
  <meta content="Beautiful Soup: a library designed for screen-scraping HTML and XML." name="D
  <meta content="Markov Approximation 1.4 (module: leonardr)" name="generator"/>
  <meta content="Leonard Richardson" name="author"/>
 </head>
 <body alink="red" bgcolor="white" link="blue" text="black" vlink="660066">
  <img align="right" src="10.1.jpg" width="250"/>
  <br/>
  <p>
```

Exploring BeautifulSoup

- Many methods such as:

```
In [9]: print(soup.title)
<title>Beautiful Soup: We called him Tortoise because he taught
us.</title>
```

```
In [8]: print(soup.get_text())

Beautiful Soup: We called him Tortoise because he taught us.

You didn't write that awful page. You're just trying to get some
data out of it. Beautiful Soup is here to help. Since 2004, it's
been saving programmers hours or days of work on quick-turnaround
screen scraping projects.
```

Exploring BeautifulSoup

- find_all()

```
In [10]: for link in soup.find_all('a'):
    ....:      print(link.get('href'))
    ....:
bs4/download/
#Download
bs4/doc/
#HallOfFame
https://code.launchpad.net/beautifulsoup
https://groups.google.com/forum/?fromgroups#!forum/beautifulsoup
http://www.candlemarkandgleam.com/shop/constellation-games/
http://constellation.crummy.com/Constellation%20Games
%20excerpt.html
https://groups.google.com/forum/?fromgroups#!forum/beautifulsoup
https://bugs.launchpad.net/beautifulsoup/
http://lxml.de/
http://code.google.com/p/html5lib/
```

IMPORTING DATA IN PYTHON II

Let's practice!