Lecture 15: Automated data retrieval 1

“Step 0” research programming – reproducible data retrieval

First, the libraries

```
import pandas as pd
import datetime

import pandas_datareader.data as web
from pandas_datareader import wb

import requests
```

Overview

Broad categories of web scraping:

1. Using a data API
2. Accessing a file that is directly part of the URL
3. Parsing data out of the html of a website

Overview

Broad categories of web scraping:

1. Using a data API
2. Accessing a file that is directly part of the URL
3. Parsing data out of the html of a website

Overview

Broad categories of web scraping:

1. Using a data API
2. Accessing a file that is directly part of the URL
3. Parsing data out of the html of a website

Overview

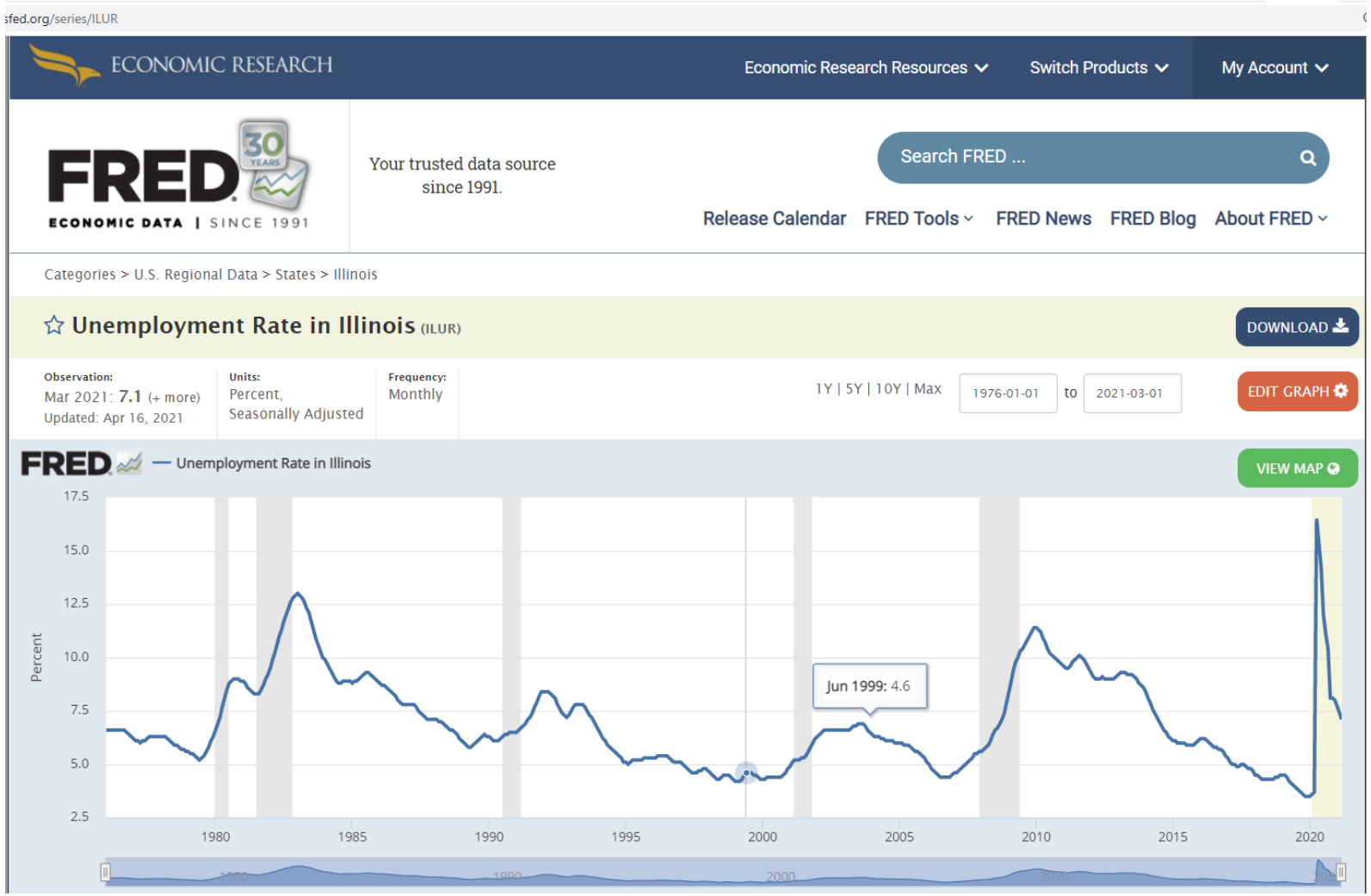
Broad categories of web scraping:

Easy – 1. Using a data API

Easy – 2. Accessing a file that is directly part of the URL

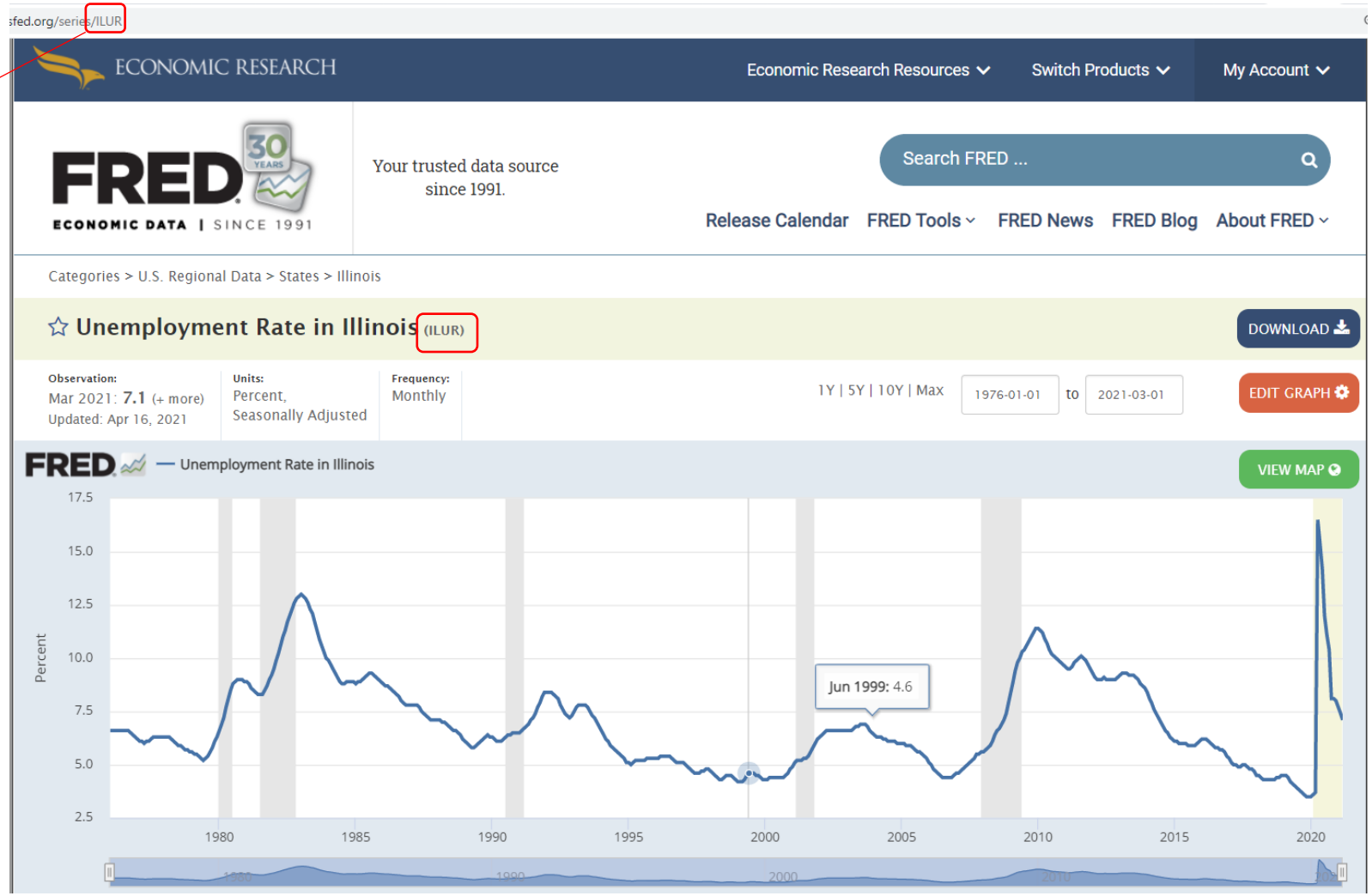
Can be crazy hard – 3. Parsing data out of the html of a website

Pandas DataReader: FRED



Pandas DataReader: FRED

Series name is the end of the URL



Pandas DataReader: FRED

```
start = datetime.date(year=2000, month=1,  day=1)
end   = datetime.date(year=2010, month=12, day=31)
series = 'ILUR'
source = 'fred'
```

Pandas DataReader: FRED

```
start = datetime.date(year=2000, month=1,  day=1)
end   = datetime.date(year=2010, month=12, day=31)
series = 'ILUR'
source = 'fred'
```

```
df = web.DataReader(series, source, start, end)
df.head()
```

Pandas DataReader: FRED

```
start = datetime.date(year=2000, month=1,  day=1)
end   = datetime.date(year=2010, month=12, day=31)
series = 'ILUR'
source = 'fred'
```

```
df = web.DataReader(series, source, start, end)
df.head()
```

This is the step that goes online

Pandas DataReader: FRED

```
start = datetime.date(year=2000, month=1,  day=1)
end   = datetime.date(year=2010, month=12, day=31)
series = 'ILUR'
source = 'fred'
```

```
df = web.DataReader(series, source, start, end)
df.head()
```

```
            ILUR
DATE
2000-01-01   4.3
2000-02-01   4.3
2000-03-01   4.4
2000-04-01   4.4
2000-05-01   4.4
```

Pandas DataReader: FRED

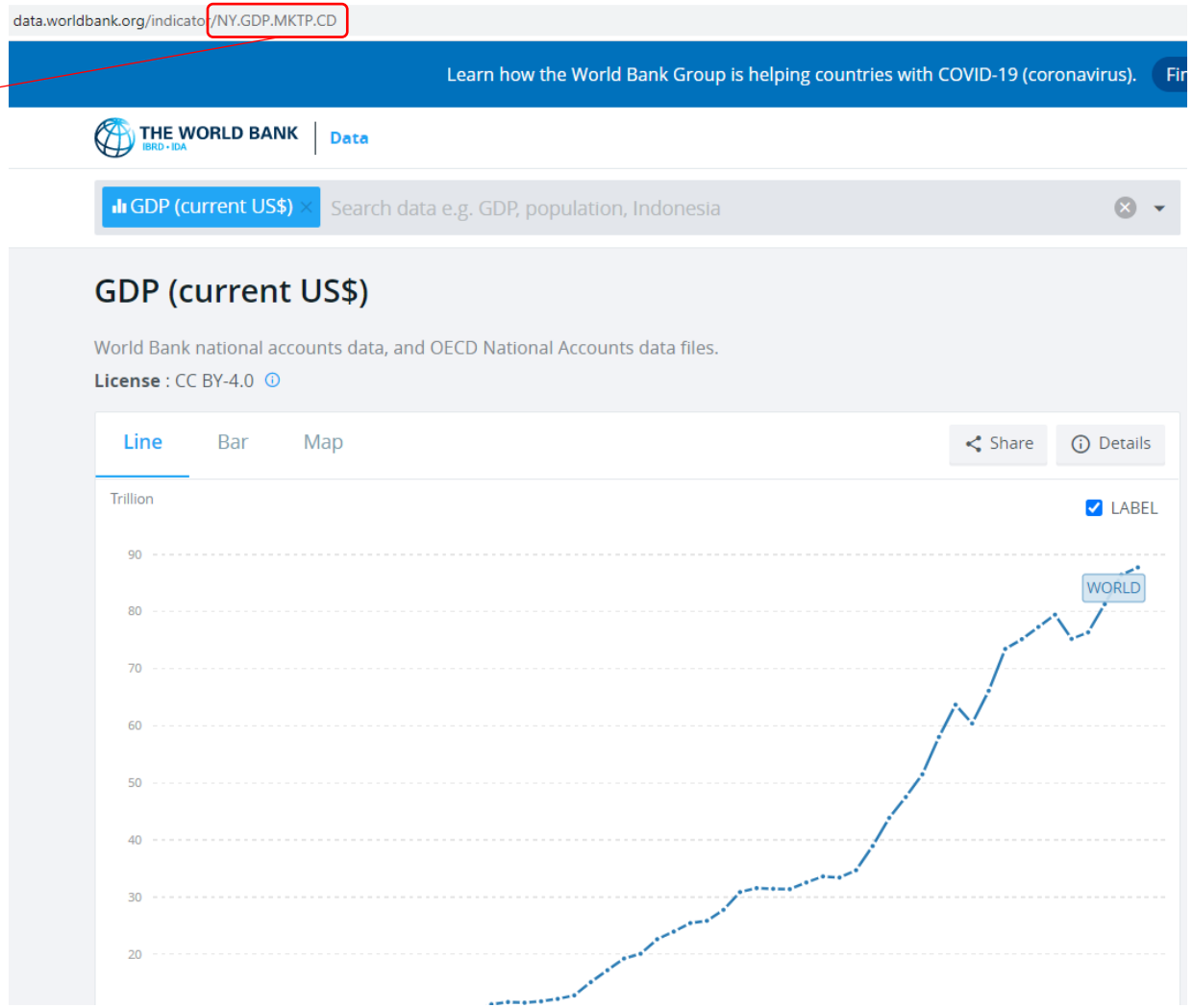
```
series = ['ILUR', 'WIUR', 'MIUR']
df = web.DataReader(series, source, start, end)
df.head()
```

Change “series” from a string, to a list of strings

	ILUR	WIUR	MIUR
DATE			
2000-01-01	4.3	3.1	3.3
2000-02-01	4.3	3.1	3.2
2000-03-01	4.4	3.1	3.3
2000-04-01	4.4	3.2	3.4
2000-05-01	4.4	3.3	3.5

Pandas DataReader: World Bank

Series name is again part of the URL



Pandas DataReader: World Bank

```
indicator = 'NY.GDP.MKTP.CD'
country = 'CL'
```

Pandas DataReader: World Bank

```
indicator = 'NY.GDP.MKTP.CD'
country = 'CL'
```

```
df = wb.download(indicator=indicator,
                 country=country,
                 start=2000, end=2010)
df.head()
```

Pandas DataReader: World Bank

```
indicator = 'NY.GDP.MKTP.CD'
country = 'CL'
```

```
df = wb.download(indicator=indicator,
                 country=country,
                 start=2000, end=2010)
df.head()
```

```
                NY.GDP.MKTP.CD
country year
Chile   2010    2.185376e+11
        2009    1.723895e+11
        2008    1.796385e+11
        2007    1.736060e+11
        2006    1.547880e+11
```

Pandas DataReader: World Bank

```
country = ['CL', 'AR', 'BR']
df = wb.download(indicator=indicator,
                 country=country,
                 start=2000, end=2010)
df.head()
```

Change “country” from a string to a list of strings

(“indicator” can also be a list of strings)

Pandas DataReader: World Bank

```
country = ['CL', 'AR', 'BR']
df = wb.download(indicator=indicator,
                 country=country,
                 start=2000, end=2010)
df.head()
```

```
                 NY.GDP.MKTP.CD
country   year
Argentina 2010    4.236274e+11
          2009    3.329765e+11
          2008    3.615580e+11
          2007    2.875305e+11
          2006    2.325573e+11
```

Pandas DataReader: World Bank

```
country = ['CL', 'AR', 'BR']
df = wb.download(indicator=indicator,
                 country=country,
                 start=2000, end=2010)
df.head()
```

		NY.GDP.MKTP.CD
country	year	
Argentina	2010	4.236274e+11
	2009	3.329765e+11
	2008	3.615580e+11
	2007	2.875305e+11
	2006	2.325573e+11

```
In [5]: df.reset_index()['country'].unique()
Out[5]: array(['Argentina', 'Brazil', 'Chile'], dtype=object)
```

Introducing requests: PDF document

```
42  url = 'http://standupeconomist.com/pdf/misc/interstellar.pdf'
43  response = requests.get(url)
```

Introducing requests: PDF document

```
42  url = 'http://standupeconomist.com/pdf/misc/interstellar.pdf'
43  response = requests.get(url)
```

- Contains information about the search (e.g. response code)
- Contains the contents of the headers sent back and forth
- Contains the actual data that makes up the web page

Introducing requests: PDF document

```
42  url = 'http://standupeconomist.com/pdf/misc/interstellar.pdf'
43  response = requests.get(url)
```

```
44  data = response.content
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

Introducing requests: PDF document

```
42  url = 'http://standupeconomist.com/pdf/misc/interstellar.pdf'
43  response = requests.get(url)
```

```
44  data = response.content
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

This content is binary, and not readable text

Introducing requests: PDF document

```
42  url = 'http://standupeconomist.com/pdf/misc/interstellar.pdf'
43  response = requests.get(url)
```

```
44  data = response.content
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

“Readers may, however, wish to use general relativity to extend the analysis to trade between planets with large relative motion. This extension is left as an exercise for interested readers because the author does not understand the theory of general relativity, and therefore cannot do it himself.”

Aside: Context management

```
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

```
1  ofile = open(r'c:/users/jeff levy/desktop/interstellar.pdf', 'wb')
2  ofile.write(data)
3  ofile.close()
```

Aside: Context management

```
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

```
1  ofile = open(r'c:/users/jeff levy/desktop/interstellar.pdf', 'wb')
2  ofile.write(data)
3  ofile.close()
```

Aside: Context management

```
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

```
1  ofile = open(r'c:/users/jeff levy/desktop/interstellar.pdf', 'wb')
2  ofile.write(data)
3  ofile.close()
```

Aside: Context management

```
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

```
1  ofile = open(r'c:/users/jeff levy/desktop/interstellar.pdf', 'wb')
2  ofile.write(data)
3  ofile.close()
```

Aside: Context management

```
45  with open(r'c:\users\jeff levy\desktop\interstellar.pdf', 'wb') as ofile:
46      ofile.write(data)
```

```
1  ofile = open(r'c:/users/jeff levy/desktop/interstellar.pdf', 'wb')
2  ofile.write(data)
3  ofile.close()
```

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?

Responsible automatic data retrieval

- Do the terms of service forbid it?
- Does the robots.txt forbid it?
- How many times does your code query a website?
- How fast is your code running its queries?
- Is it a big website, that can handle a large load, or a small website that you might create problems for?
- How many times will your code be run?