NYU | TANDON SCHOOL OF ENGINEERING

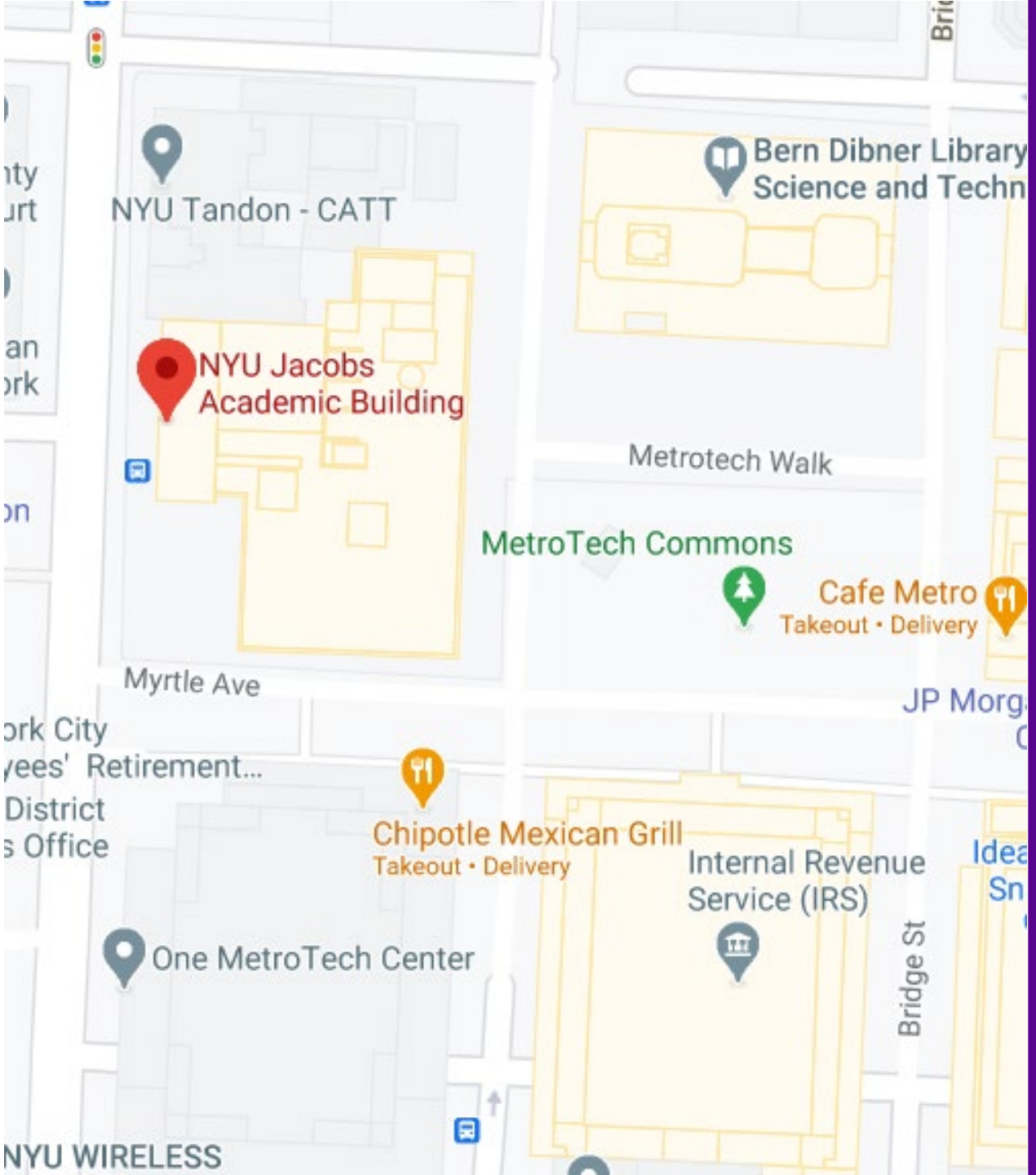
Lecture 7
Websites

Programming for Business Analytics
MG-GY 8401

NYU TANDON SCHOOL OF ENGINEERING

Agenda

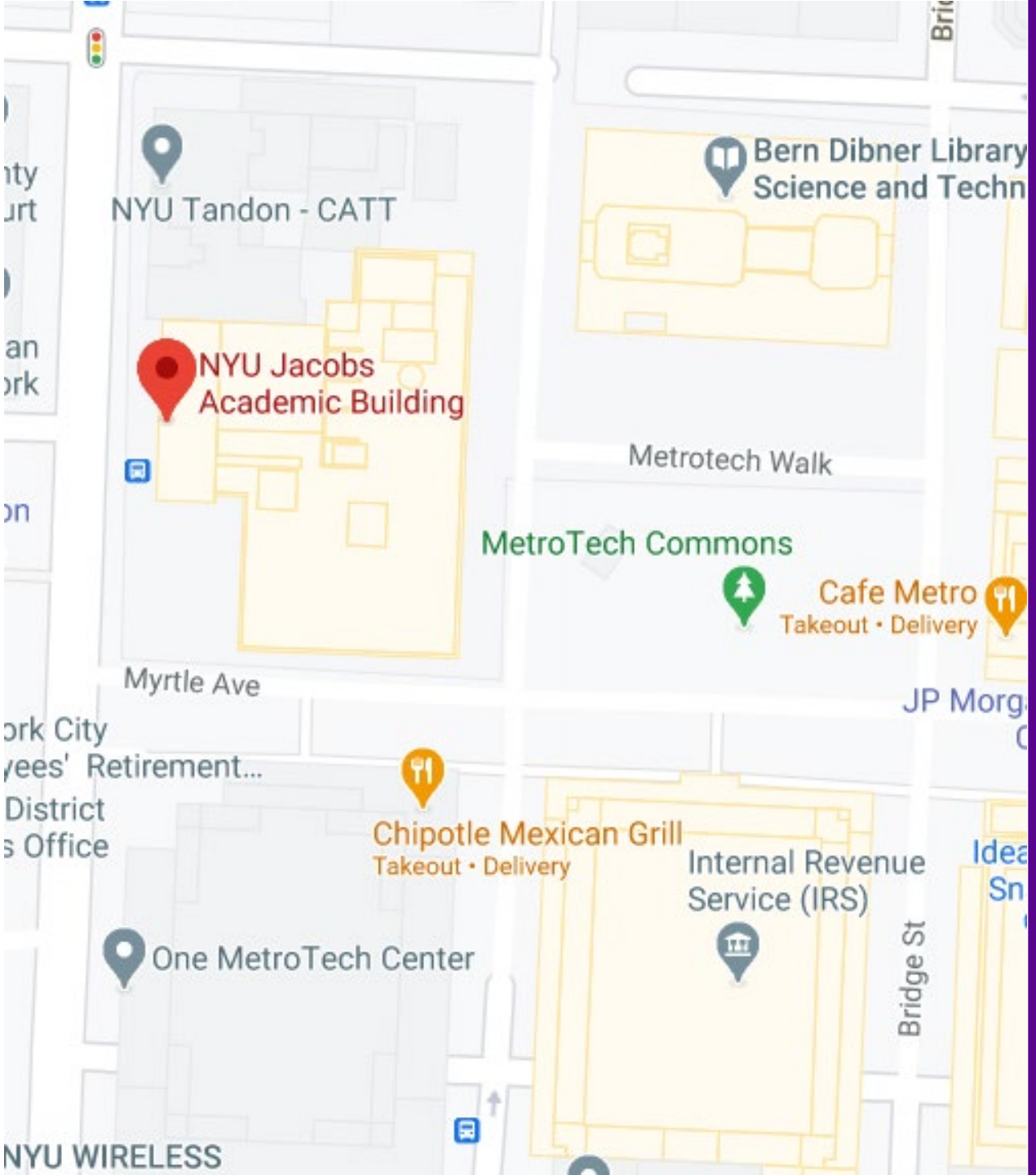
- Connecting to Websites
- Web-Scraping
- Parsing Text as Data



NYU TANDON SCHOOL OF ENGINEERING

Logistics

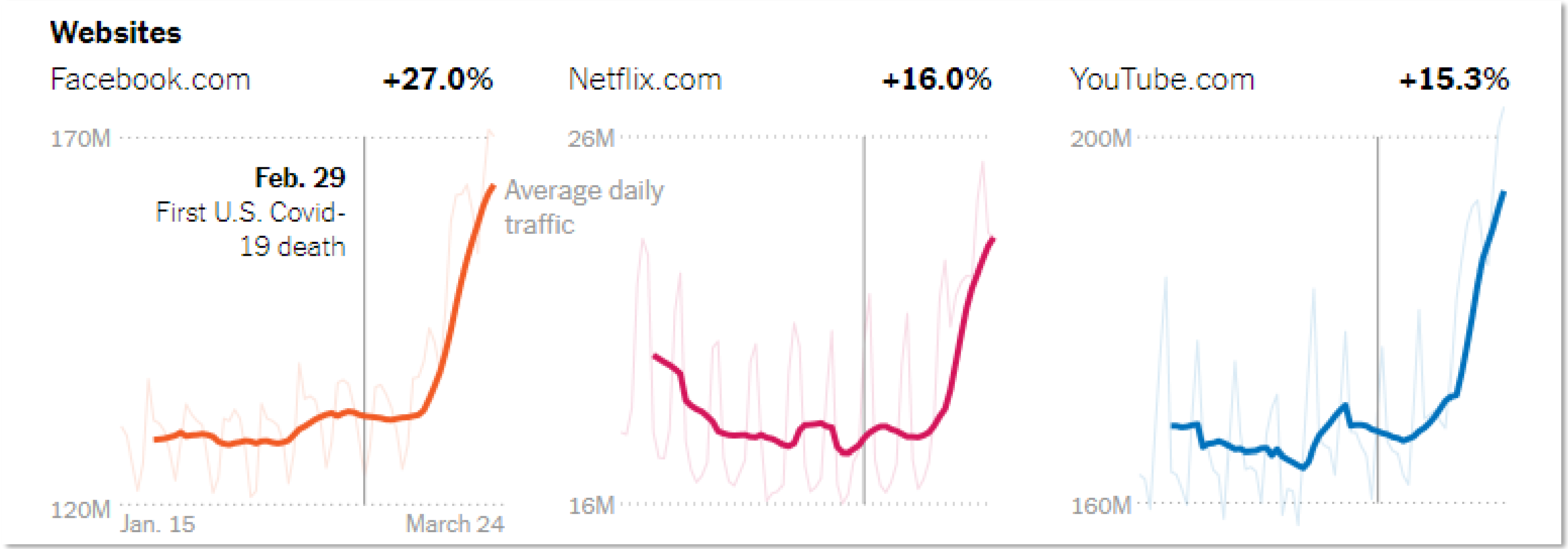
- Homework
  - Homework 6
  - Homework 5

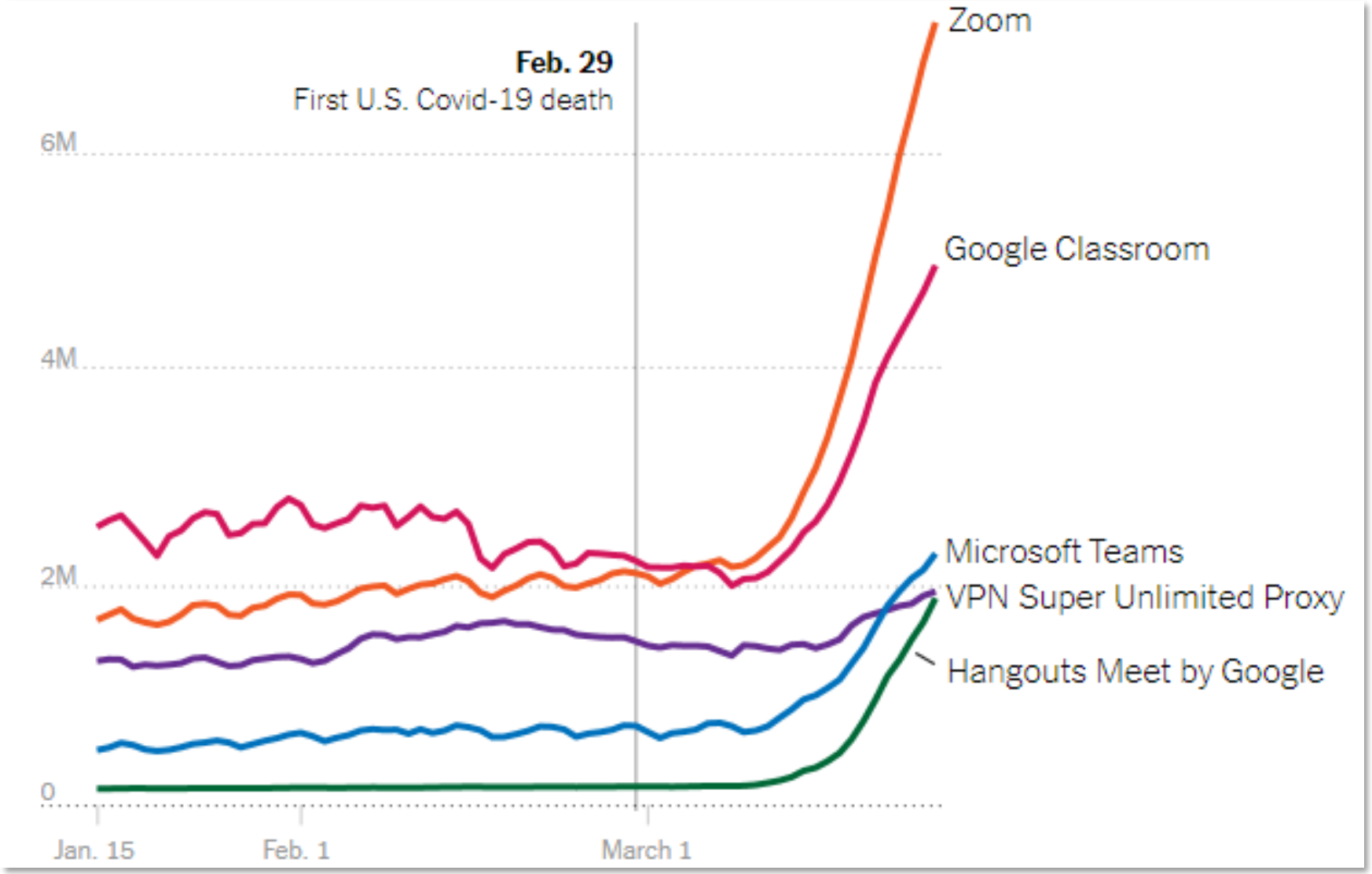


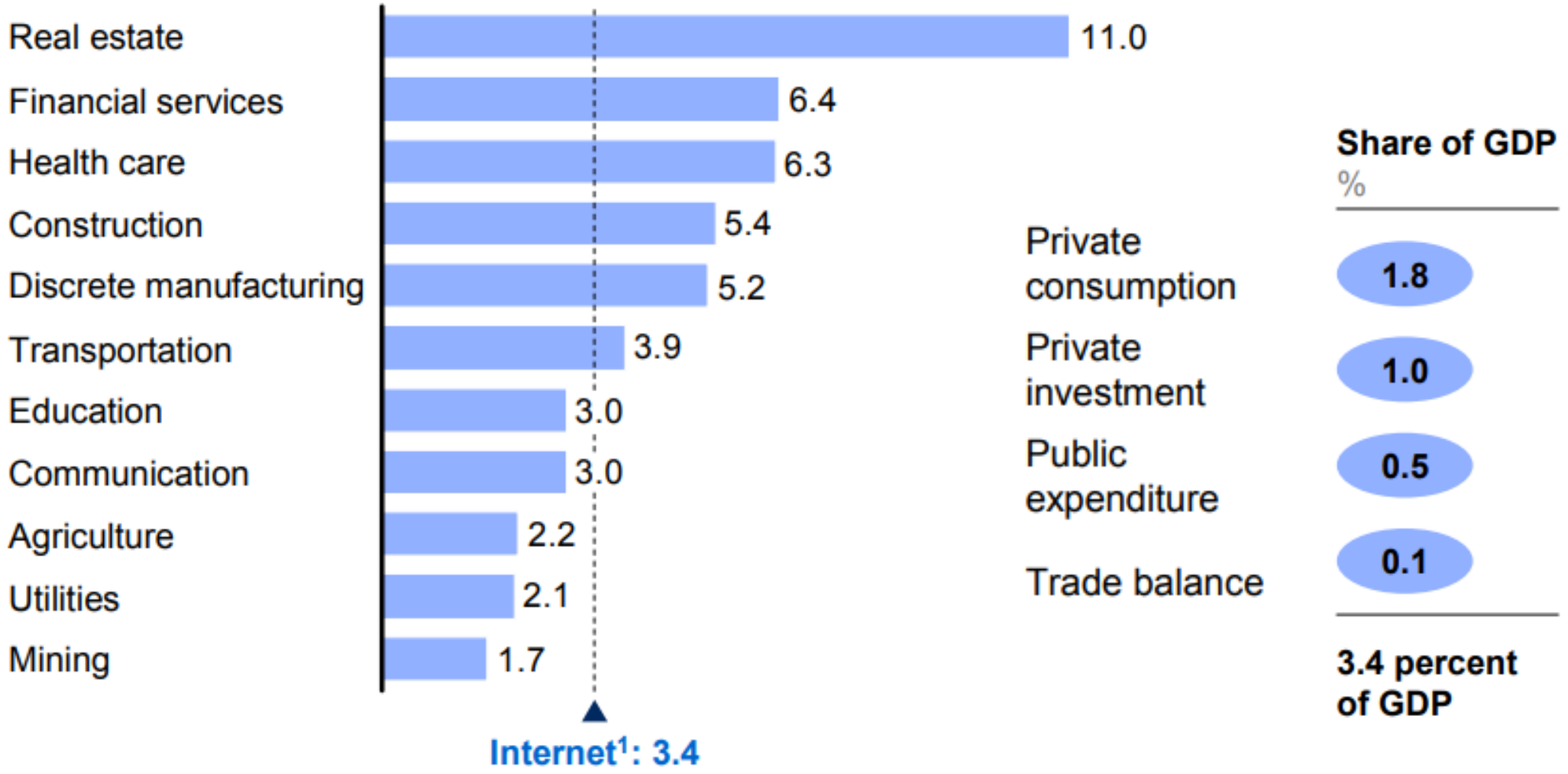
NYU TANDON SCHOOL OF ENGINEERING

Logistics

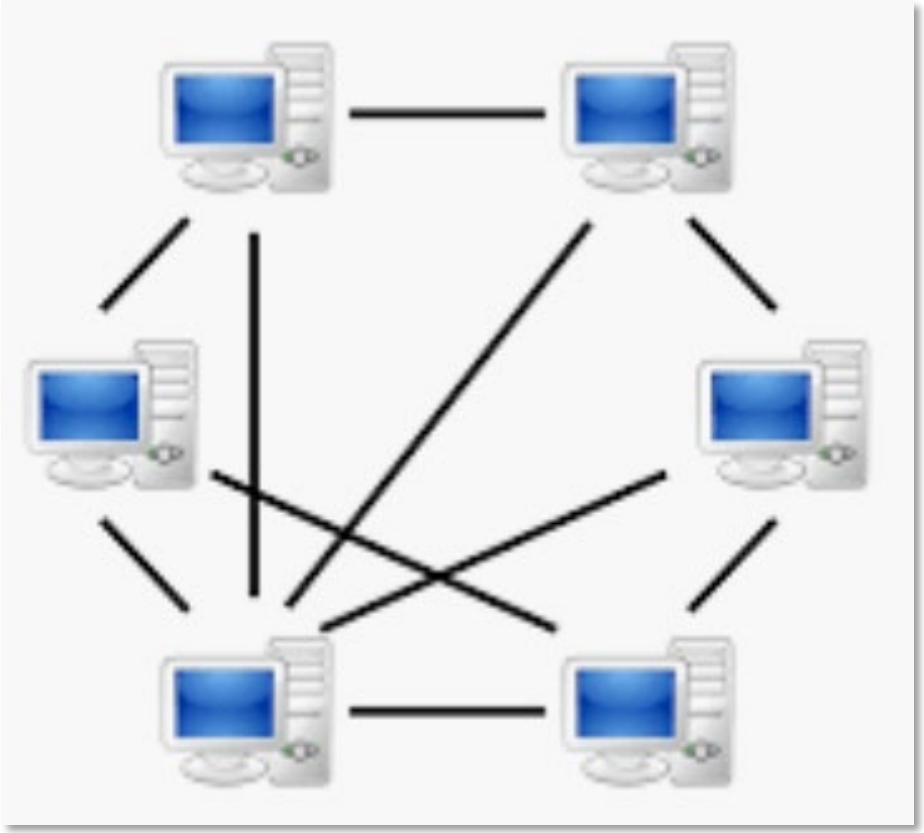
- Homework
  - Homework 7
  - Project

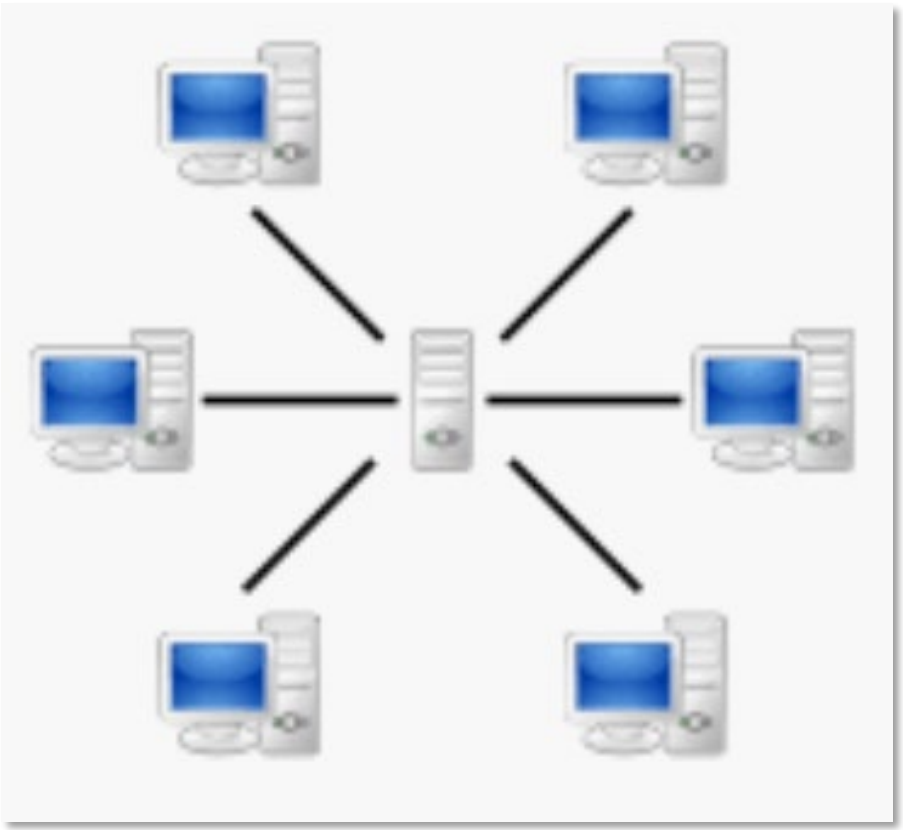

Websites
Facebook.com
+27.0%
Netflix.com
+16.0%
YouTube.com
+15.3%
170M
120M
Jan. 15
March 24
Feb. 29
First U.S. Covid-19 death
Average daily traffic
26M
16M
200M
160M

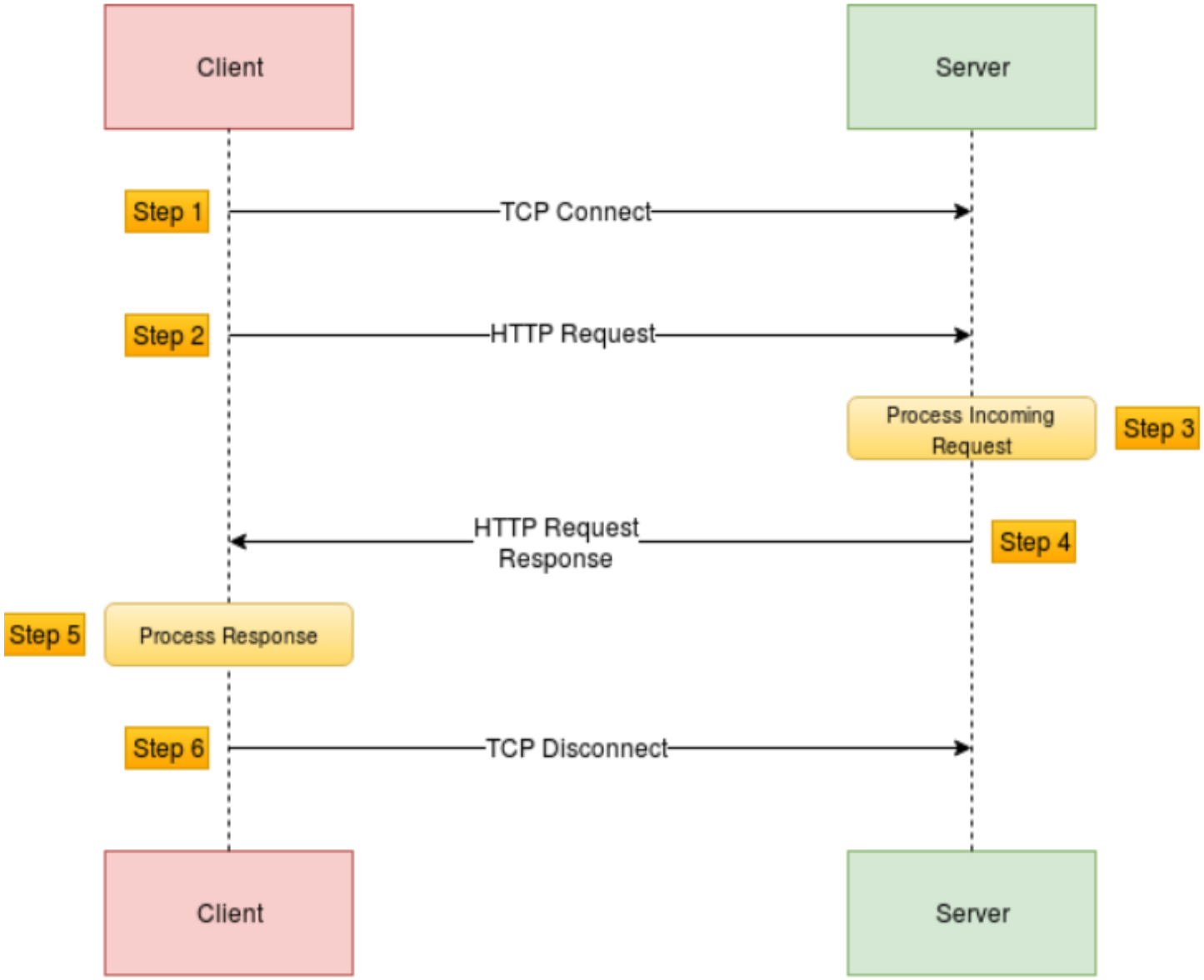

Feb. 29
First U.S. Covid-19 death
6M
4M
2M
0
Jan. 15
Feb. 1
March 1
Zoom
Google Classroom
Microsoft Teams
VPN Super Unlimited Proxy
Hangouts Meet by Google


Real estate
11.0
Financial services
6.4
Health care
6.3
Construction
5.4
Discrete manufacturing
5.2
Transportation
3.9
Education
3.0
Communication
3.0
Agriculture
2.2
Utilities
2.1
Mining
1.7
Internet[1]: 3.4
Share of GDP
%
Private consumption
1.8
Private investment
1.0
Public expenditure
0.5
Trade balance
0.1
3.4 percent of GDP

Internet Protocol

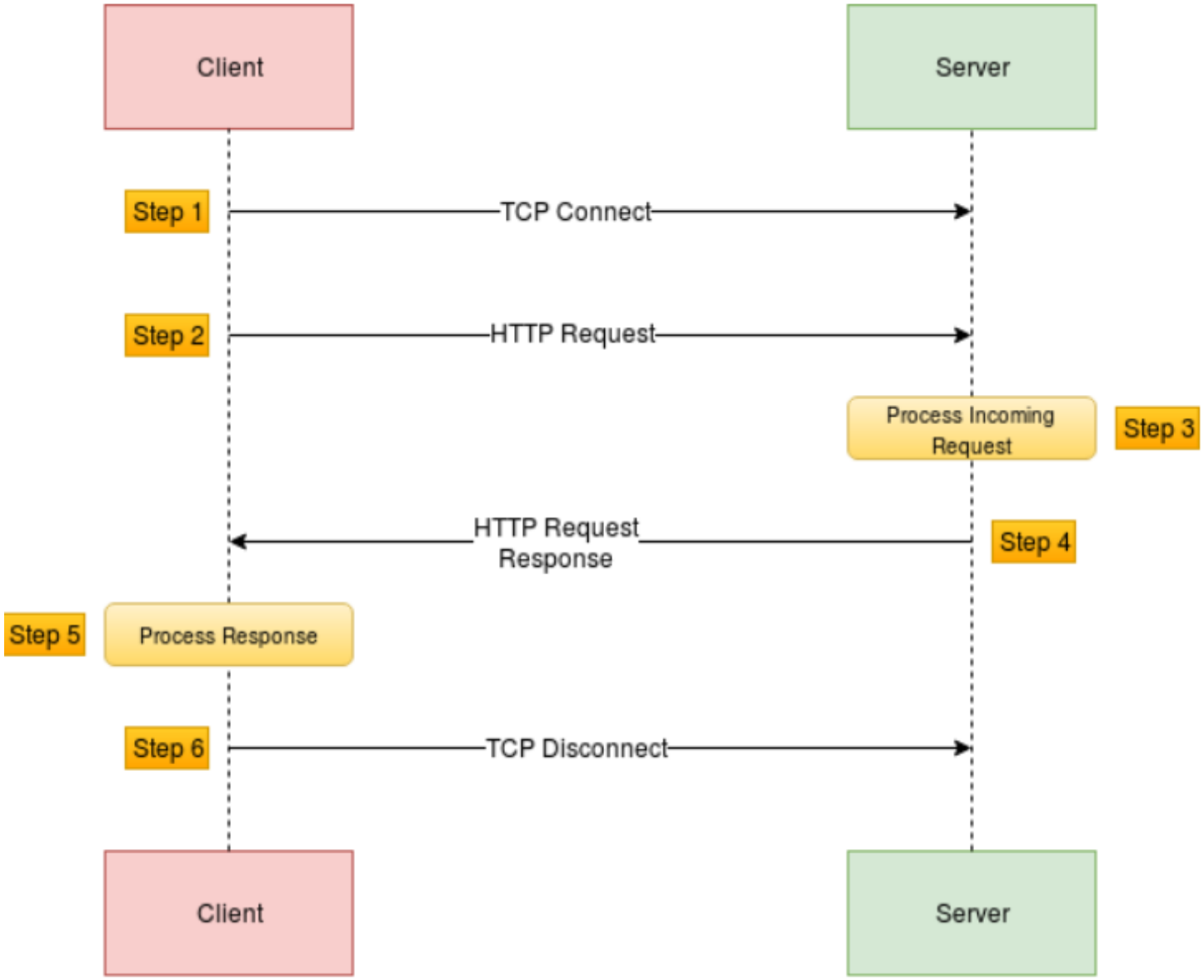






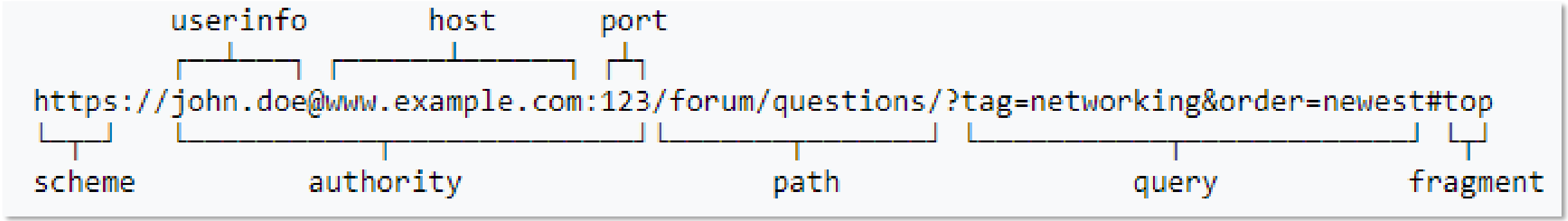
Transmission Control Protocol

- ordering of requests and responses
- splitting large packets into small packets
- removal of duplicate packets
- retransmission of lost packets



HyperText Transfer Protocol

- client sends request to server
- server runs an application to process the request
- server sends a response to client



Uniform Resource Locator

- credentials for authentication
  - username + password
- location on the website
- query to filter content on website

Which of the following would match the regular expression jo*hn?

1. joohn
2. jon
3. jhn
4. john
5. jooooooohnn

Which of the following would match the regular expression jo*hn?

Try it on regex101.com where you can for experimenting with regular expressions

1. joohn
2. jon
3. jhn
4. john
5. jooooooohnn

HTTP Commands

DELETE	`/delete`	"The request's DELETE parameters.
GET	`/get`	The request's query parameters.
PATCH	`/patch`	The request's PATCH parameters.
POST	`/post`	The request's POST parameters.
PUT	`/put`	The request's PUT parameters.

HTTP Status Codes

Which of the following would match revolution, revolutionary, revolutionaries?

A. revolution[a-z]?
B. revolution[a-z]*
C. revolution[a-z]+

Which of the following would match revolution, revolutionary, revolutionaries?

A. revolution[a-z]?

B. revolution[a-z]*

C. revolution[a-z]+

JavaScript Object Notation

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 27,
  "address": {
    "city": "New York",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": []
}
```

- JSON format resembles dictionaries in Python with key-value pairs
- Keys are strings
- Values are
  - Numbers
  - Strings
  - Boolean
  - List
  - Another Dictionary

How could we match social security numbers in text?

Note that social security numbers have the form

###-##-####

1. [0 to 9]{3}-[0 to 9]{2}-[0 to 9]{4}
2. {0-9}[3]-{0-9}[2]-{0-9}[4]
3. [0-9]{3}-[0-9]{2}-[0-9]{4}
4. [0-9]{3}-[0-9]{3}-[0-9]{3}

How could we match social security numbers in text?

Note that social security numbers have the form

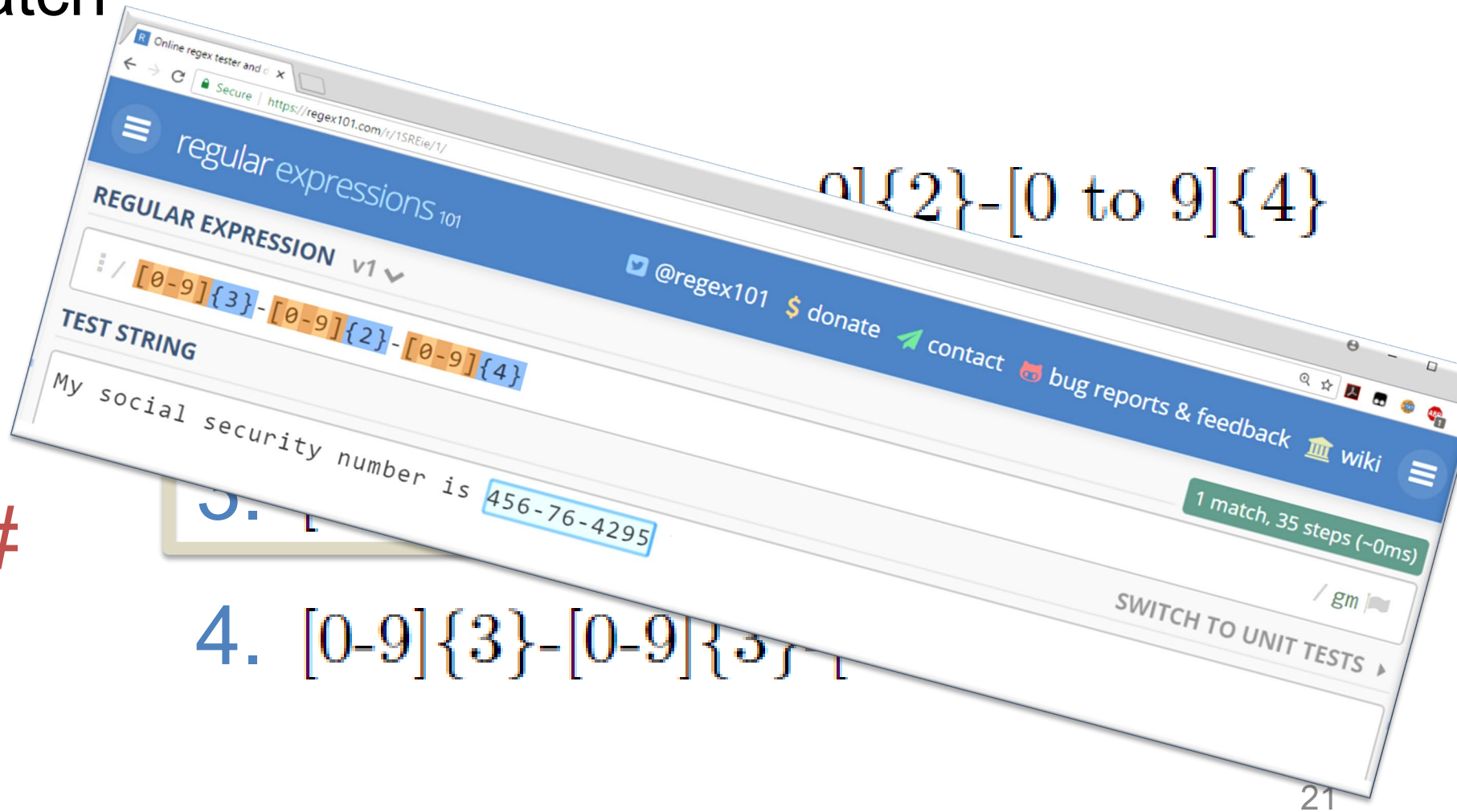
###-##-####

1. [0 to 9]{3}-[0 to 9]{2}-[0 to 9]{4}
2. {0-9}[3]-{0-9}[2]-{0-9}[4]
3. [0-9]{3}-[0-9]{2}-[0-9]{4}
4. [0-9]{3}-[0-9]{3}-[0-9]{3}

How could we match social security numbers in text?

Note that social security numbers have the form

###-##-####



```
<catalog>
    <plant>
        <common>Bloodroot</common>
        <botanical>Sanguinaria canadensis</botanical>
        <zone>4</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$2.44</price>
        <availability>031599</availability>
    </plant>
    <plant type='a'>
        <common>Columbine</common>
        <botanical>Aquilegia canadensis</botanical>
        <zone>3</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$9.37</price>
        <availability>030699</availability>
    </plant>
</catalog>
```

```
<catalog>
    <plant>
        <common>Bloodroot</common>
        <botanical>Sanguinaria canadensis</botanical>
        <zone>4</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$2.44</price>
        <availability>031599</availability>
    </plant>
    <plant type="a">
        <common>Columbine</common>
        <botanical>Aquilegia canadensis</botanical>
        <zone>3</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$9.37</price>
        <availability>030699</availability>
    </plant>
</catalog>
```

Start tag

End tag

Content

```
<catalog>
    <plant>
        <common>Bloodroot</common>
        <botanical>Sanguinaria canadensis</botanical>
        <zone>4</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$2.44</price>
        <availability>031599</availability>
    </plant>
    <plant type='a'>
        <common>Columbine</common>
        <botanical>Aquilegia canadensis</botanical>
        <zone>3</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$9.37</price>
        <availability>030699</availability>
    </plant>
</catalog>
```

Start tag

Content consists of multiple elements

End tag

```
<catalog>
    <plant>
        <common>Bloodroot</common>
        <botanical>Sanguinaria canadensis</botanical>
        <zone>4</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$2.44</price>
        <availability>031599</availability>
    </plant>
    <plant type='a'>
        <common>Columbine</common>
        <botanical>Aquilegia canadensis</botanical>
        <zone>3</zone>
        <light>Mostly Shady</light>
        <price currency="USD">$9.37</price>
        <availability>030699</availability>
    </plant>
</catalog>
```

Attribute

NYU TANDON SCHOOL OF ENGINEERING

Review

- HyperText Transfer Protocol
- Javascript Object Notation
- HyperText Markup Language

NYU TANDON SCHOOL OF ENGINEERING

References

- Mitchell, *Web-Scraping with Python*

(Chapter 1 + 2)

Questions

- Describe the learning objectives.
- Summarize the relevant take-aways.
- Ask about unclear information.