bit.ly/pyladies_pandas

Exploring Wikipedia with Pandas

Zareen Farooqui

What is Pandas?

Open-source Python library for
data analysis and modeling

panel **da**ta

Goals

- High-level overview of Pandas library
- Write code on your own and complete challenge questions
- Lots of ideas, code & data to play around with after tutorial

Introducing the Data

Pageviews	Clickstream
~150 MB	1.18 GB
~ 5 million rows	25,615,007 rows

Let’s explore what’s happening in the world right now….

pageviews-20160921-010000.gz	21-Sep-2016 02:48	38565958
pageviews-20160921-020000.gz	21-Sep-2016 03:31	38032592
pageviews-20160921-030000.gz	21-Sep-2016 04:27	38119794
pageviews-20160921-040000.gz	21-Sep-2016 05:26	37877998
pageviews-20160921-050000.gz	21-Sep-2016 06:22	38632298
pageviews-20160921-060000.gz	21-Sep-2016 07:58	41147282
pageviews-20160921-070000.gz	21-Sep-2016 08:54	45359742
pageviews-20160921-080000.gz	21-Sep-2016 09:17	49001483
pageviews-20160921-090000.gz	21-Sep-2016 10:21	51429348
pageviews-20160921-100000.gz	21-Sep-2016 11:27	52671794
pageviews-20160921-110000.gz	21-Sep-2016 12:20	53004512
projectviews-20160901-000000	01-Sep-2016 02:24	21101
projectviews-20160901-010000	01-Sep-2016 03:37	21316
projectviews-20160901-020000	01-Sep-2016 06:47	20516
projectviews-20160901-030000	01-Sep-2016 07:38	21307
projectviews-20160901-040000	01-Sep-2016 08:33	21520

Download last pageviews file
(not projectviews)

https://dumps.wikimedia.org/other/pageviews/2016/2016-09/

Pageviews Data

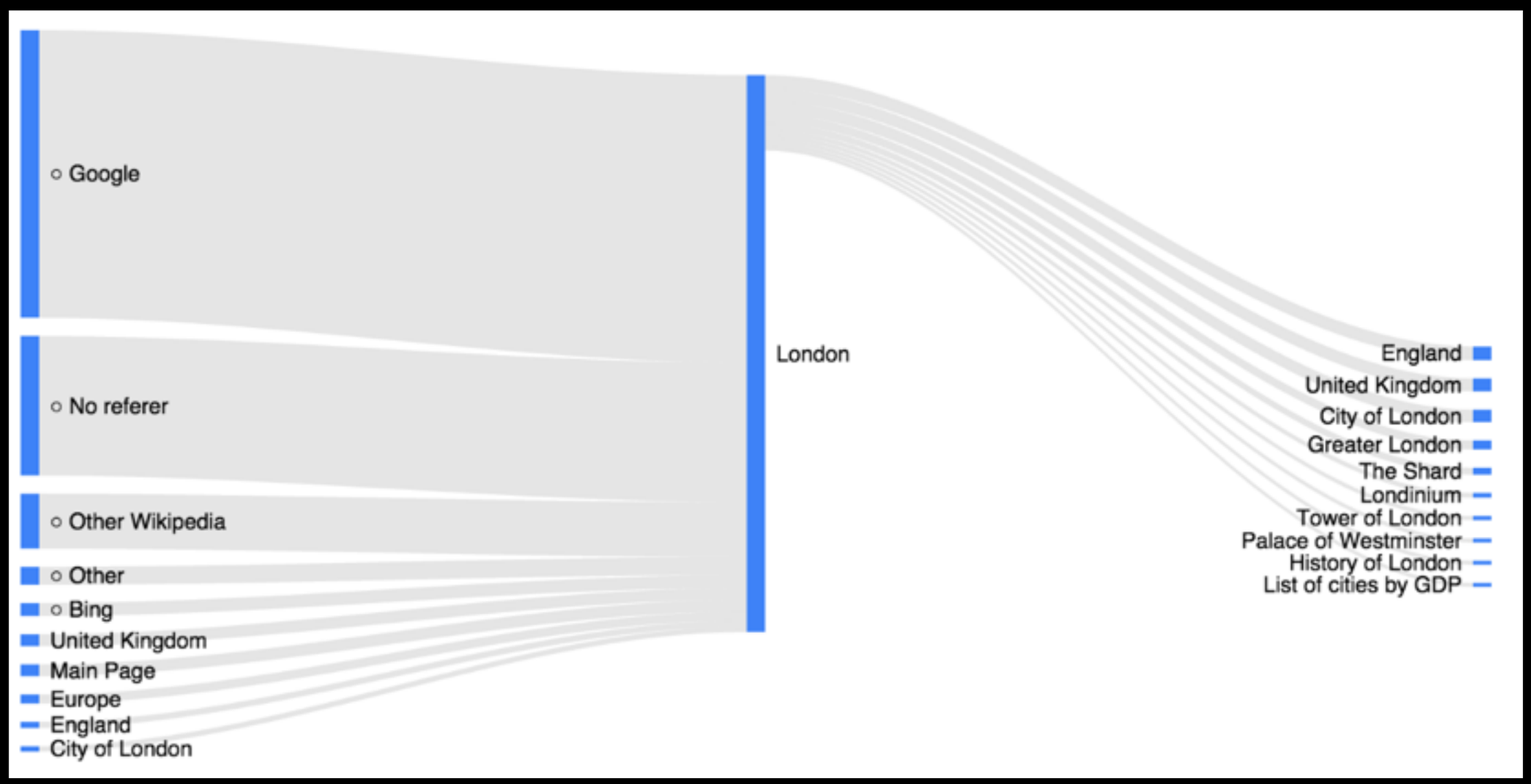
	project	article	requests	bytes_served
1228192	en	Main_Page	306095	0
2315811	en.m	Main_Page	185612	0
465978	de.m	Wikipedia:Hauptseite	66556	0
1555004	en	Special:Search	56479	0
3221294	es.m	Wikipedia:Portada	45600	0

- **project**: Wikimedia project (290+ languages, many different project like wikibooks, wikinews, wikiquotes)
- **article**: title of current article
- **requests**: number of times that link was visited
- **bytes_served**: total size of content returned

Let’s explore what was happening in the world March 2016….

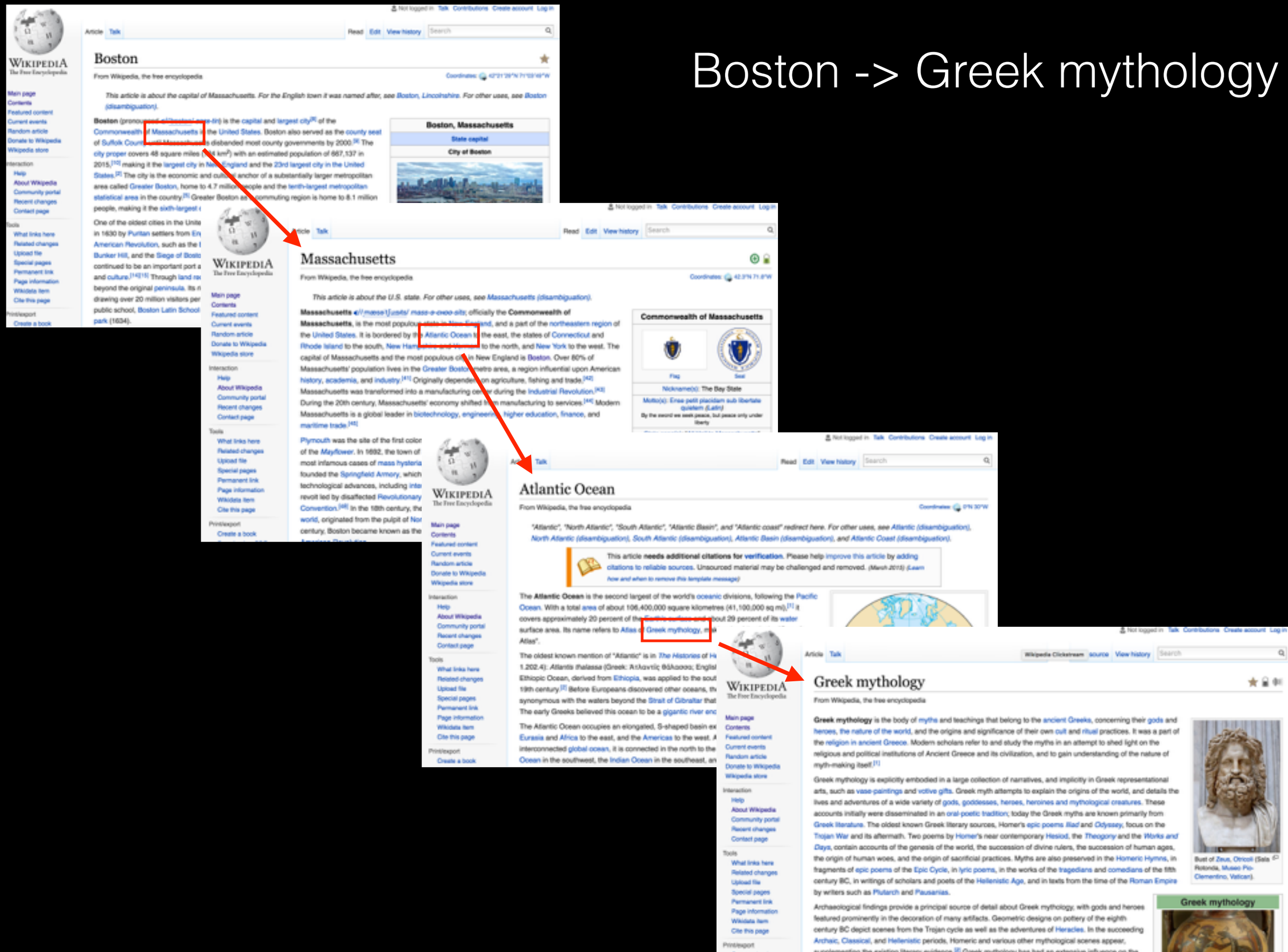
Clickstream

how people get to a Wikipedia article and what links they click on



Source: https://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/

Boston -> Greek mythology



Clickstream Data

	prev	curr	type	n
0	Wild_Bill_Hickok	Deadwood,_South_Dakota	link	18
1	(Ghost)_Riders_in_the_Sky:_A_Cowboy_Legend	Stan_Jones_(songwriter)	link	410
2	(Hey_You)_The_Rock_Steady_Crew	Rock_Steady_Crew	link	75
3	(Miss)understood	Heaven_(Ayumi_Hamasaki_song)	link	13
4	(Theme_from)_Valley_of_the_Dolls	Do_You_Know_the_Way_to_San_Jose	link	12

- **prev**: where the hit came from
  - another Wikipedia articles
  - external sources (Google, Facebook, Twitter)
- **curr**: title of current article
- **type**:
  - link: prev and curr are both Wikipedia articles,
  - external: prev is a non-Wikipedia source (Google, Facebook, etc.),
  - other: prev and curr are both Wikipedia articles but the curr article was searched for or the prev article was spoofed
- **n**: number of times that link was visited that month

Data Structures

- **series:** one-dimensional object similar to a labeled array, list or column in a table

```
s = pd.Series([22, 'Boston', 3.14, -87264034, 'Hello there!'])
s
```

```
0              22
1          Boston
2            3.14
3        -87264034
4    Hello there!
dtype: object
```

- **dataframe:** 2 dimensional labeled data structure with rows and columns (spreadsheet, SQL table, dict of Series)

Thanks!

@iamzareenf

@zareenf

zareenf@gmail.com