Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Os Keyes

Reproducibility & Open Research | Week 3 | October 11, 2018

Overview of the day

- Six Provocations for Big Data: Review & Reflections
- A primer on copyright, licensing, and hosting for code and data
- Introduction to replicability, reproducibility, and open research
- Reproducibility case study: fivethirtyeight.com and the Bechdel Test
- Group activity: assessing reproducibility in data journalism
- Overview of Assignment 1: Data curation

Reading reflections

danah boyd and Kate Crawford, *Six Provocations for Big Data (2011)*

“As the paper mentions, the increase in accessibility and lowering of barriers to working with "big data" has allowed those not traditionally in the sciences/engineering to work with such data. But as the paper also explains, one needs to be careful about the limitations of such "big dataset", and these require mathematical and domain knowledge that these new data consumers may lack. **Therefore, should some basic education on working with data (statistics, for example) be more prevalent or even mandatory?** We see big lack of even the most basic knowledge of statistical concepts from, say journalists today in many mainstream publications.”

- Ryan

“They state that combining datasets creates unique challenges, although they don't really list what those are - so, what are some examples of how not understanding limitations or using inappropriate interpretations could be magnified by using multiple datasets?”

- Kenton

“Analogous to how stocks are traded on the market, do you think having a central source where data can be listed, sold and bought (ethically, of course) would make sense? I feel that data is currently concentrated with a few companies like Facebook, Google etc., and this would help reduce this monopolistic concentration of data with a few partners.”

- Tejas

“Question: How do you get around the fact that money = access and make “data” more accessible? If you have more money, you can buy access to sources others cannot, buy/access better computing equipment, and a better ability to implement/test more accurate and efficient analysis methods?”

- Hannah

Key concepts & themes

danah boyd and Kate Crawford, *Six Provocations for Big Data (2011)*

Bigger data != Better Data

- “Finally, in the era of the computational turn, it is increasingly important to recognize the value of ‘small data’....Take, for example, the work of Tiffany Veinot (2007), who followed one worker-a vault inspector at a hydroelectric utility company-in order to understand the information practices of blue-collar worker. In doing this unusual study, Veinot reframed the definition of ‘information practices’ away from the usual focus on early-adopter, white-collar workers, to spaces outside of the offices and urban context.”

Accessible != Ethical

- OKCupid: 60,000+ accounts’ data stolen
- “but it was online!”

Data copyright & licensing

Why do we care about copyright?

- As a data consumer?
- As a data producer?

What even is copyright, anyway?

- A system of rights afforded to the creators of original works
- Reproduction, modification, money, licensing

“Original works”

- Works must be original to be copyrightable
- Art
- Code
- Not data
  - But data presentation

What is licensing?

- Rights can be waived or sub-licensed.
- Example: right to create derivative works
  - Remixes
- Can come with conditions
- Data releases are licenses (usually)

Licenses for code

- **MIT**
  - Do whatever
- **GPL**
  - Provide attribution in derivatives
  - Release any derivatives under the GPL
  - Do whatever

Licenses for documentation

- Creative Commons suite
  - The build-a-bear of licenses
- Common building blocks:
  - **BY:** must provide attribution
  - **NC:** cannot use commercially
  - **ND:** cannot make derivatives
  - **SA:** must release derivatives under the same/a compatible license
  - **0:** public domain release

Licenses for documentation

- Combinations:
  - CC-BY-SA
  - CC-BY-ND
  - CC-NC
  - CC-0
- Combinations that don’t exist:
  - CC-SA-ND
- All of these work for data!

Using licensed code

- Preference MIT
- GPL problems:
  - Virality
    - Linking issues
  - aGPL

Using licensed data

- What does “attribution” look like?
  - Include any copyright terms
  - “This data was provided by X and can be found at Y URL”
  - Mention if it’s a derivative work

Making your work accessible

Licensing code

License	Reuse	Credit	Pay
MIT	X		
GPL	X	X	
aGPL		X	X

Licensing data

License	Reuse	Credit	Virality
CC-0	X		
CC-BY	X	X	
CC-BY-SA		X	X

Open publishing

- Once you've worked out how to license data, how do you release it?
- Open publishing!
- "Green" OA:
    a. Self-archiving
    b. Can clash with publishing!
- "Gold" OA:
    a. Published archiving
    b. Costs $$$ (with some deferments)

Places to archive

- Figshare (https://figshare.com/)
  a. 100GB free per project - but fees after that
  b. 1TB max
- Dryad (https://datadryad.org/)
  a. Fees for anything - but no size limit!
- OSF (https://osf.io/)
  a. Totally free - 5GB per file
- Zenodo (https://zenodo.org/)
  a. Free - 50GB per dataset - but less reliable

Making your project identifiable

- If your data is free but not findeable, it’s useless
- If it’s free and findeable...until the link breaks...it’s useless
- Digital Object Identifiers (DOIs)
- Unique ID for an artefact
    a. Works even if sites fall over
    b. Single point of reference
- Supported by Dryad, OSF...

Things to include

- Code
- Data
- Documentation
- *Sampled* data
    a. Lowers the barrier to exploration
- Suggested uses
    a. What did you want to explore but couldn’t?
    b. What else could it be interesting for?

If you can’t publish..

- Sometimes you can’t publish
  a. Private information
  b. Corporate IP
- Release internally
  a. The standards are good for in-org transparency
  b. Helps with project structure (“you in 6 months..”)
- Release samples + instructions
  a. “Here is an example, if you want to use the full dataset..”

An introduction to open research

What is open research?

Open research is research conducted in the spirit of free and open-source software. Much like open-source schemes that are built around a source code that is made public, **the central theme of open research is to make clear accounts of the methodology freely available via the internet, along with any data or results extracted or derived from them.** This permits a massively distributed collaboration, and one in which anyone may participate at any level of the project.

Especially if the research is scientific in nature, it is **frequently referred to as open science**. Open research can also include social sciences, the humanities, mathematics, engineering and medicine.

Source: https://en.wikipedia.org/wiki/Open_research

Scientific impact of OR

Publishing your research openly can increase the **scientific impact** of your work

- It makes it easier for others to **check your work** and **verify your conclusions**
- It helps avoids the “file drawer problem”*
- It allows others to build off what you did more easily

*https://en.wikipedia.org/wiki/Publication_bias

Reproducibility and Replicability

- **Reproducing** a research study involves applying the same methods to the same data and achieving an *identical* result.

- **Replicating** a research study involves applying the same methods to new data and achieving an *identical, commensurate, confirming* result.

The terms are not used consistently, and in reality it's more of a spectrum than buckets, esp in data science. Nevertheless, these are the definitions we'll try to use.

Source:
https://www.practicereproducibleresearch.org/core-chapters/2-assessment.html

Societal impact of OR

Publishing your research openly can increase the **social impact** of your work

- It makes it easier for researchers, journalists, and the public to find your research and use it.
- It helps bolster your reputation as a Serious Scientist (™).

If you're not publishing regularly in peer-reviewed science venues regularly, public documentation of your data, code, and analytical contributions can serve as *alternative metrics* of your impact as a researcher.

- Ex: Downloads, forks, pull requests, citations, and derivative works of your projects, code libraries, and datasets.

OR and alternative impact metrics

Altmetrics expand our view of what impact looks like, but also of what's making the impact. This matters because expressions of scholarship are becoming more diverse. [Peer-reviewed] articles are increasingly joined by:

- The sharing of "raw science" like datasets, code, and experimental designs
- Semantic publishing or "nanopublication," where the citeable unit is an argument or passage rather than entire article.
- Widespread self-publishing via blogging, microblogging, and comments or annotations on existing work.

J. Priem, D. Taraborelli, P. Groth, C. Neylon (2010), Altmetrics: A manifesto, 26 October 2010. http://altmetrics.org/manifesto

Open research in organizations

- Assume *someone* will be reading, re-using, or making decisions based on your research, even if it's corporate IP or sensitive data that you can't share publicly.
- This audience will have diverse expertise and needs: executives, product managers, developers, other scientists and researchers.
- You don't want to field every single request for information about your data, your methods, or your findings.
- You won't always be in your current position. How can you future-proof your work?

How is this a human-centered thing?

Audience: Who are you publishing your research for?

Purpose: How do you want them to use your research?

Context: What factors (under your control) could impact *whether* or *how* they use it? What factors could impact what conclusions they draw from it?

How is this a human-centered thing?

Based on an analysis of human tasks

Built to take account of human skills

Designed to address human needs

Monitored in terms of human benefits

Source: R. Kling and S. L. Star “Human centered systems in the perspective of organizational informatics”. 1997

Literate programming

Let us change our traditional attitude to the construction of programs.

Instead of imagining that our main task is to instruct a computer what to do, **let us concentrate rather on explaining to human beings what we want a computer to do.**

- Donald Knuth. "Literate Programming”

https://commons.wikimedia.org/wiki/File:KnuthAtOpenContentAlliance.jpg

Literate programming

“The practitioner of literate programming can be regarded as an essayist, whose main concern is with exposition and excellence of style. Such an author… chooses the names of variables carefully and explains what each variable means.

He or she strives for a program that is comprehensible because its concepts have been introduced in an order that is best for human understanding”

- Donald Knuth. "Literate Programming”

https://commons.wikimedia.org/wiki/File:KnuthAtOpenContentAlliance.jpg

How is this a human-centered thing?

Adopting open research practices supports...

- Research outputs that account for diverse needs, expertise, and use-cases.
- Building in mechanisms for quick iteration and external verification.
- Documentation of goals, values, assumptions, and thought process while you are performing your research, rather than after the fact (or not at all).
- Attribution of the work you're building off of.
- Accountability for your research. Accept that you might be wrong.

Data journalism and data science

“I have found this "new" brand of data journalism disappointing foremost because *it wants to perform science without abiding by* ***scientific norms.”***

Communalism: All science is the product of social collaboration and should contribute to the common enterprise (‘standing on the shoulders of giants’)

Skepticism: Scientific claims must be scrutinized, not accepted uncritically.

Keegan, Brian, 2014. “The need for openness in data journalism”

https://en.wikipedia.org/wiki/Mertonian_norms

Case Study

Open research and data journalism

The Bechdel Test: Origins

Excerpt from “The Rule” (1985). *Dykes to Watch Out For* by Alison Bechdel. Firebrand Books.

“The Dollars and Cents Case Against Hollywood’s Exclusion of Women”

- Used data from BechdelTest.com and TheNumbers.com
- Analyzed 1,615 films released from 1990 to 2013 to examine the relationship between the prominence of women in a film and that film’s budget and gross profits.
- **Claims**
  - the median budget of movies that passed the Bechdel Test was significantly lower than the median budget of all films
  - films that pass the Bechdel Test may in fact have a better return on investment, overall, than those that don’t.

Hickey, Walt. FiveThirtyEight, 2014.
https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

“The Need for Openness in Data Journalism”

- Used data from BechdelTest.com and TheNumbers.com
- Analyzed 1,615 films released from 1990 to 2013 to examine the relationship between the prominence of women in a film and that film’s budget and gross profits.
- **Findings**
  - the median budget of movies that passed the test was substantially lower than the median budget of all films in the sampel e.
  - films that feature meaningful interactions between women may in fact have a better return on investment, overall, than films that don’t.

Keegan, Brian, 2014. http://www.brianckeegan.com/2014/04/the-need-for-openness-in-data-journalism/

Exercise (20 minutes)

Read Keegan's replication and expansion of Hickey, starting with "The Hook: The Bechdel test article in FiveThirtyEight". As you are reading, take notes on...

1. What aspects of Hickey's analysis and write-up…
    a. Made interpreting or reproducing his results difficult?
    b. Made his claims misleading or unverifiable?
2. What aspects of Keegan's re-analysis helps...
    a. Identify the limitations and assumptions of Hickey's work?
    b. Provide a different (better?) account of the phenomenon being studied?
    c. Identify the limitations and assumptions of Keegan's own approach?
    d. Support reproduction and expansion of Keegan's own analysis?

Keegan, Brian, 2014. https://github.com/brianckeegan/Bechdel/blob/master/Bechdel_test.ipynb

Replicability & Reproducibility best practices

Chapter 2 "Assessing Reproducibility" and Chapter 3 "The Basic Reproducible Workflow Template" from *The Practice of Reproducible Research*. Ariel Rokem, Ben Marwick, Valentina Staneva, and Justin Kitzes

Three key practices

- **Clearly separate, label, and document** all data, files, and operations that occur on data and files
- **Document all operations fully**, automating them as much as possible, and avoiding manual intervention in the workflow when feasible
- **Design a workflow as a sequence of small steps** that are glued together, with intermediate outputs from one step feeding into the next step as inputs

Stage 1: Data acquisition

- Where is your data coming from?
- What TOU or licenses apply to the source data?
- Who created your dataset?
- What tools were used to collect your data?
- If your data is a sub-sample, what criteria were used?
- What features are described in your data?
- Is a local copy of your source data available?
- Are there known errors, inconsistencies, or incompletes in your source data?

Stage 1: Data acquisition

What mechanism was used to gather the data?

Scraping: when was it scraped? Is a static archive of the original web page available?

Queries: what is the schema of the database/API? What query was used? When was the query run?

Dumps: what is the schema of the dump? File format? Version number?

Streams: during what time interval was the data collected from the stream? How was the stream accessed?

Stage 2: Data processing

- What tools were used in the processing of your data?
- What sub-sampling, filtering, aggregation, or transformation steps were performed?
- What order were processing steps performed in?
- Why were these processing steps performed?
- How were errors, inconsistencies, or incompletes discovered, and how were they addressed?
- Did your data processing involve any manual (i.e. non-programmatic) steps?
- Are you making incremental datasets available?
- Are you making a final processed dataset available?

Stage 3: Data analysis

- What are the goals of your analysis?
- What is the nature of your analysis?
- What assumptions about your data are assumed in your analytical approach?
- What tools used in the analysis of your data?
- What order were analysis steps performed in?
- Why and how was each analytical step performed?
- Are you making samples, demos, or test sets available?
- How are the results of your analysis presented?
- Are you making a final analyzed dataset available?

Overall goal

When designing and documenting your acquisition/processing/analysis workflow, consider multiple scenarios

- **Reproducibility:** To what extent could someone else with access to the same data reproduce the steps in your process and evaluate their results against yours?
- **Replicability:** To what extent could someone else with different hardware/software and with *similar but not identical data* reproduce the steps in your process and evaluate their results against yours?
- **Other forms of reuse:** Would a data-savvy journalist be able to write an accurate description of your study? Would a fellow data scientist feel confident citing your work, even if they didn't replicate it? Would your mom/dad understand?

A few more best practices

- Version your code and data
- Explain each step that allow others understand your thought process
- Describe complex steps or concepts at multiple levels with
    a. a grammatical prose description of what you are doing
    b. clear function-level I/O descriptions (e.g. docstrings)
    c. liberal use of inline comments
- Use descriptive names for files, functions, and variables
- Provide real data examples in context
- Describe/demonstrate the output of each step
- **Document the unexpected**: anything counterintuitive or potentially surprising about your code, methods, or data.

In-Class Activity

Graded, 45 minutes, groups of 4-5

Assessing reproducibility in data journalism

https://github.com/fivethirtyeight/data

In-class activity instructions

Full instructions and links in the *Week 3 in-class activity* discussion thread

1. Download and unpack data-master.zip from Files/Datasets on Canvas
2. Select a repo from the spreadsheet
3. Read the instructions doc

Deliverables (post in the Week 3 activity thread):

1. At least 3 specific ways in which the artifacts in this repo support reproducibility of the analysis presented in the article. For each, say *how* it supports reproducibility.
2. At least 5 concrete suggestions for making the analysis in this repo *more reproducible*. For each suggestion, say *how* this suggestion will improve reproducibility.

Choose one person from your team to submit your deliverables to Canvas. Include the link to the dataset AND all group members' names in the post.

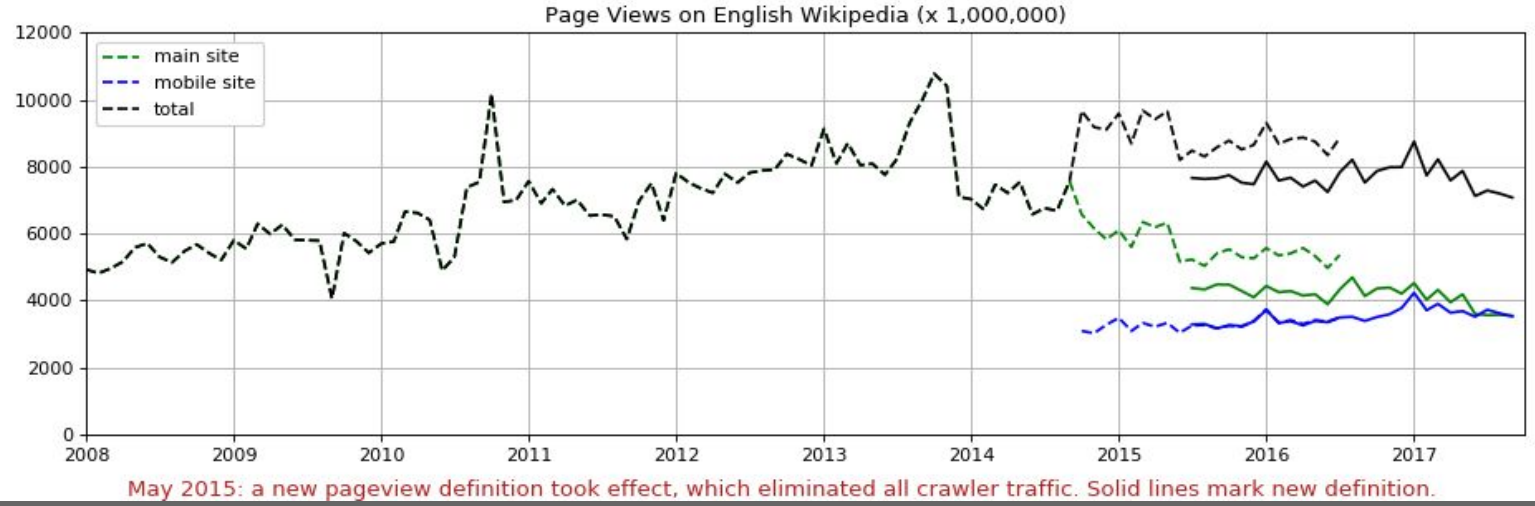
Homework due next week

- Read Duarte, N., Llanso, E., & Loup, A. *Mixed Messages? The Limits of Automated Social Media Content Analysis.* FAT '18
  - Submit reflection via Canvas

- Assignment 1: Data Curation
  - Submit link to a GitHub repository named data-512-a1 via Canvas: https://canvas.uw.edu/courses/1244514/assignments/4376106

A1: Data curation

https://wiki.communitydata.cc/Human_Centered_Data_Science_(Fall_2018)/Assignments#A1:_Data_curation



Goal: make a graph like this one, using the Wikimedia REST API as a data source, and document your process and outcomes according to best practices for open, reproducible research.

Deliverables: all code, data, and documentation in your Github repo. With a graph that looks (something) like this one.

Graph: Rex Thompson, 2017. MIT License. Used with permission.

A1: Goal

The **goal** for this assignment is to construct, analyze, and publish a dataset of **monthly** English Wikipedia mobile and desktop page traffic from the earliest month where data is available through the most recent month where data is available.

The **purpose** of the assignment is to demonstrate that you can follow best practices for open scientific research in designing and implementing your project, so that anyone can understand your process and reproduce your results.

A1: licensing

Licensing for Wikipedia data

- All the text of Wikipedia pages (including articles), and all public datasets, are available CC-BY-SA.
- See the Wikimedia Terms of Use for more details: https://wikimediafoundation.org/wiki/Terms_of_Use/en

Example (for this assignment): *"Data was gathered from the Wikimedia REST API, Wikimedia Foundation, 2018. CC-BY-SA 3.0"*

A1: data sources

Wikipedia public data: REST API: https://wikimedia.org/api/rest_v1/

Page traffic by project, access type, agent type, and time interval

- **Page views data:** current and historical traffic data
- **Legacy data:** (a.k.a “page counts”) historical traffic data, less granular

...and more! See also https://en.wikipedia.org/api/rest_v1/ which has even more data

Additional documentation: https://www.mediawiki.org/wiki/REST_API

Page View API

- Historical and current data
  - Developed to replace Legacy Page Counts; provides granular traffic data
  - Views per project or article(s)
    - Aggregated by hourly/daily/monthly
    - Filterable by
      - Agent: Spider vs user
      - Access: Desktop/mobile-app/mobile-web
  - Data available from mid 2015 - last month

Legacy Page Count API

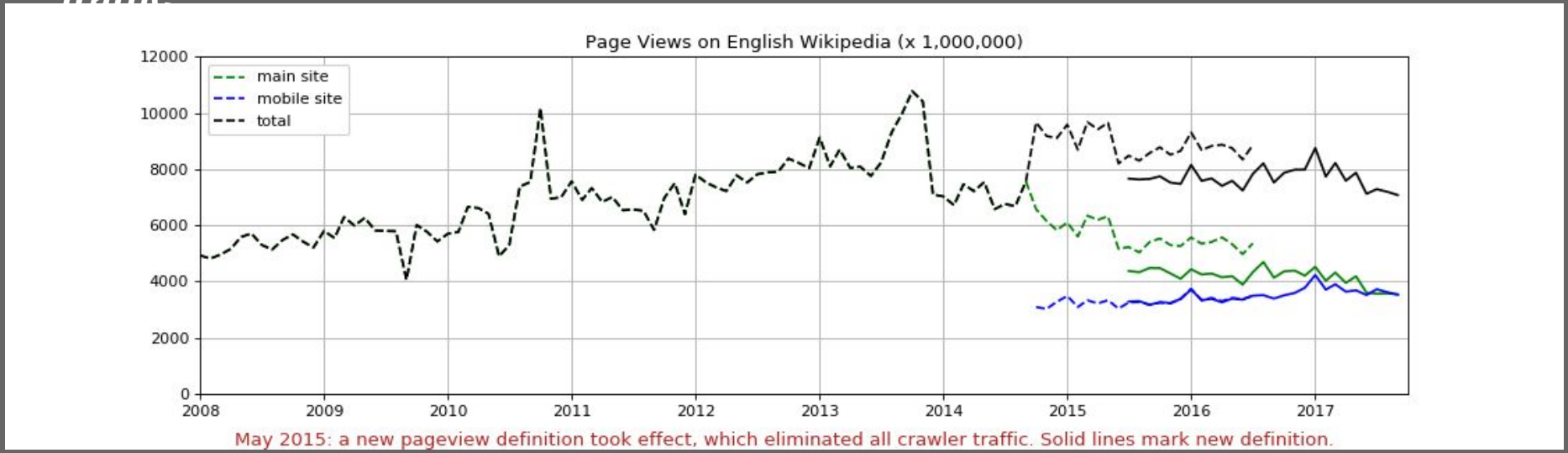
- Legacy traffic data - no longer updated
- Views per project (e.g. en.wikipedia.org)
  - Aggregated by hourly/daily/monthly
  - Filterable by
    - Access-site: Desktop-site/mobile-site
- Data available from late 2007 - mid 2016 (desktop) and late 2014 - mid 2016 (mobile)

Differences: Pageviews vs Pagecounts

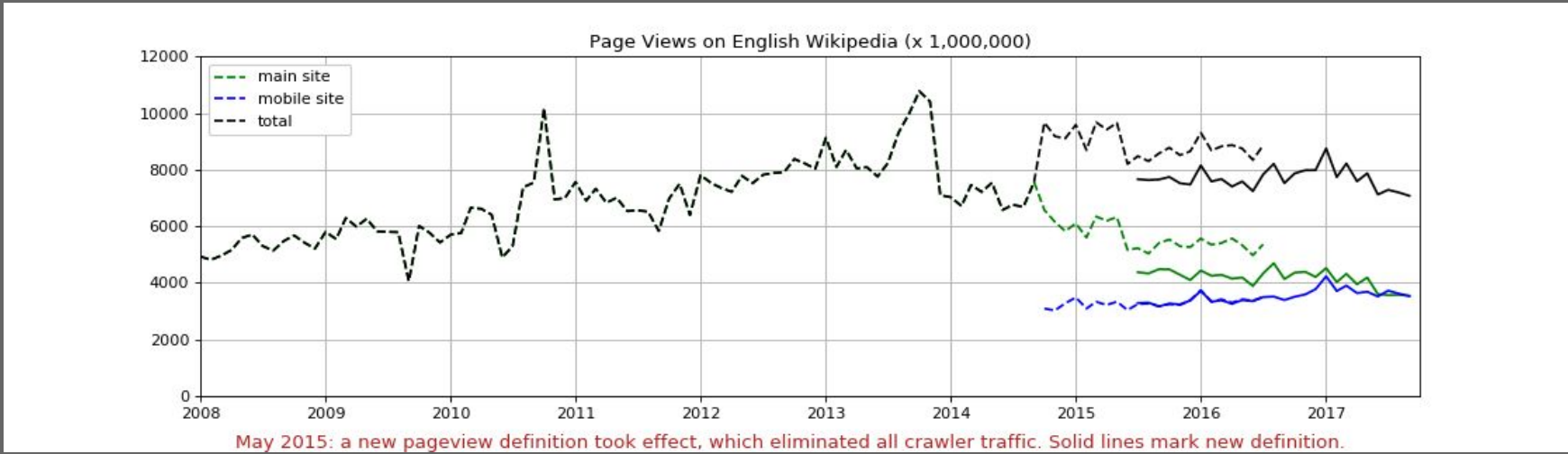
- Legacy Page Counts does not let you filter out web spiders, so it overcounts ‘organic’ traffic.
- Page Views provides options to filter by ‘spider’ and ‘user’ traffic.
- Page Views divides mobile by ‘mobile-app’ and ‘mobile-web’.
- There are a couple small arg key/value differences (e.g. Legacy Page Counts uses ‘access-site’ arg where Page Views uses ‘access’)

A1: Analysis

Your analysis will consist of developing a time series visualization of Wikipedia article traffic by **month**, divided by: ***desktop traffic***, ***mobile traffic***, and ***all traffic***.



Graph: Rex Thompson, 2017. MIT License. Used with permission.



- **Where possible, you must filter out web spiders**, in order to represent ‘organic’ readership traffic to Wikipedia.
- **Where necessary, you must combine individual sources of mobile traffic** (app and web) to display total counts for all mobile traffic in a given month.
- **You must collect data for all months for which data is available.** Some months have traffic data from both PageViews and PageCounts.

http://paws-public.wmflabs.org/paws-public/User:Jtmorgan/data512_a1_example.ipynb

English Wikipedia page views 2007 - 2018

Sample API code


```
import json
import requests
```

```
endpoint_legacy = 'https://wikimedia.org/api/rest_v1/metrics/legacy/pagecounts/aggregate/{project}/
{access-site}/{granularity}/{start}/{end}'

endpoint_pageviews = 'https://wikimedia.org/api/rest_v1/metrics/pageviews/aggregate/{project}/{acces
s}/{agent}/{granularity}/{start}/{end}'
```

```
# SAMPLE parameters for getting aggregated legacy view data
# see: https://wikimedia.org/api/rest_v1/#!/Legacy_data/get_metrics_legacy_pagecounts_aggregate_pro
ject_access_site_granularity_start_end
example_params_legacy = {"project" : "en.wikipedia.org",
                 "access-site" : "desktop-site",
                 "granularity" : "monthly",
                 "start" : "2001010100",
                # for end use 1st day of month following final month of data
                 "end" : "2018100100"
                    }

# SAMPLE parameters for getting aggregated current standard pageview data
# see: https://wikimedia.org/api/rest_v1/#!/Pageviews_data/get_metrics_pageviews_aggregate_project_
access_agent_granularity_start_end
example_params_pageviews = {"project" : "en.wikipedia.org",
                   "access" : "desktop",
                   "agent" : "user",
                   "granularity" : "monthly",
                   "start" : "2001010100",
                   # for end use 1st day of month following final month of data
                   "end" : '2018101000'
                       }

# Customize these with your own information
headers = {
    'User-Agent': 'https://github.com/yourusername',
    'From': 'youremail@uw.edu'
```

A1: Required deliverables

Your GitHub repo should contain...

1. **Source *and* final data files** that follow the specified conventions for file type, file names, column headers, and column values, and contain the correct number of rows.
2. **A Jupyter notebook** in which all data processing and analysis steps are clearly presented and documented and the sequence of steps is clearly communicated.
3. **A README.md** file that contains all data and code descriptions, attributions and provenance information, and hyperlinks to all relevant resources and documentation (inside and outside the repo).
4. **A LICENSE file** that specifies the license under which you are releasing your code.
5. **A .png image of your visualization** that follows the specified naming convention

https://wiki.communitydata.cc/Human_Centered_Data_Science_(Fall_2018)/Assignments#A1:_Data_curation

A1: Tips and hints

- The first full month for which mobile data is available is October 2014
- Some months may return 0s or error messages from the API. Read the docs carefully so you know what to watch out for.
- Your chart should be the right scale to view the data, all units, axes, and values should be clearly labeled, and it should possess a key and a title.
- Use a generic API library like `requests`, rather than something you found on GitHub--external libraries may not work as expected.
- Re-check the requirements before you submit. Ask questions on Slack if you're unsure about something.
- When in doubt, document it.

Homework due next week

- Read Duarte, N., Llanso, E., & Loup, A. *Mixed Messages? The Limits of Automated Social Media Content Analysis.* FAT '18
  - Submit reflection via Canvas

- Assignment 1: Data Curation
  - Submit link to a GitHub repository named data-512-a1 via Canvas: https://canvas.uw.edu/courses/1244514/assignments/4376106

Unused slides