Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Oliver Keyes

Data curation | Week 3 | October 12, 2017

Overview of the day

- Final project overview
- Exercise review
- Reading reflection review
- Intro to open research
- *Dinner break (15 min)*
- Data copyright and licensing
- Supporting reproducibility and replicability
- Making your work accessible
- *Coffee break (15 min)*
- Group activity
- Intro to Assignment 1: Data curation

Final project overview

Final project plan

- Due Week 7 (November 9)
- 10 points
- Min. 1000 words
- Jupyter Notebook or .md file on GitHub, link submitted to Canvas

This proposal should focus on the following questions:

- Why are you planning to do this analysis? Provide background information about the topic, research questions/hypotheses, (imagined) business goals, and anything else that will be required to properly contextualize your study.
- What is your plan? Describe the data sources will you collect, how data will be collected and processed, the analysis you intend to perform, and the outcomes and deliverables you anticipate.
- Are there any unknowns or dependencies that might affect your ability to complete this project?
- **What are some of the human-centered aspects of this project?** How do human-centered design considerations inform your decision to pursue this project, and your approach to performing the work?

Final project presentation

- Due Week 11, December 7
- 5-7 minute oral presentation
- 10 points
- Link to Google Slides to Canvas before class starts

This presentation should demonstrate the following:

- Your ability to present effectively to a professional audience. Imagine that you are pitching your project to directors/execs at a company you work for.
- Your ability to communicate the importance of your research to the specified audience
- Your ability to communicate the nature and implications of your findings accurately and compellingly
- Your ability to do all of the above in a very short time (hint: practice beforehand and time yourself)

Final project report

- Due Week 12 (Sunday, December 10 at 11:59pm)
- No min/max word count--whatever necessary to get the job done
- 15 points
- GitHub repo w/ Jupyter notebook, full datasets and documentation; link to repo submitted via Canvas

A well-written, well-executed research study report that includes:

- All your code and data, thoroughly documented
- The importance of the problem or opportunity you have identified
- Your research question(s)
- The methods, data, and approach that you used to collect and analyze the data
- Findings, implications, and limitations of your study
- A thoughtful reflection that describes the specific ways that human-centered data science principles informed your decision-making in this project—from beginning to end.

Final project timeline

- **Week 5 (October 26):** Project plan assigned. We'll provide examples of datasets and research questions, and talk about project planning.
- **Week 6 (November 2):** We'll set aside time to talk over project ideas individually and answer questions. Come to office hours if you want extra 1:1 time.
- **Week 7 (November 9):** Project plan due before class! We'll set aside time in class to talk over project plans and answer questions.
- **Week 8-9:** Work on your project on your own, come to office hours if you want help.
- **Week 10 (November 30):** Final presentation assigned. Final project workshop—bring your final project progress to class, and be prepared to give and receive feedback with classmates.
- **Week 11 (December 7):** Final project presentations.
- **Week 12 (Sunday, December 10):** Final projects are due by 11:59pm. Absolutely no late work accepted w/out signed Disability Accommodation agreement.

Exercise review

Exercise review

- Lot of awesome submissions!
- More rigorous marking coming in
- Some confusion around informed consent/HCDS
- Further informed consent points:
  - There is a presumption of competence
  - Researchers can’t second guess refusals to participate
  - HCDS means deference to subjects

Reading review

Reading review

Many questions along the lines of:

“Holy heck this is awful, do we need new laws? Do we even have laws to begin with?”

Reading review

“Absent self-interest, how/why should people be motivated to behave ethically?”

Reading review

“As new technologies and research methods are discovered all the time, how do we ensure the our principles are always up to date?”

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

Why OR? Do it for science

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

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

*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 a *commensurate, confirming, or contradictory* result.

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

Why OR? Do it for glory

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 researcher

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.

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) will impact whether/how they use it?

How is this a human-centered thing?

- Publishing your research openly shows that you take **responsibility** for your research. Including the possibility that you might be wrong.
- It provides **transparency** around your values, motivations, and assumptions.
- Having a public audience in mind when designing and publishing your projects encourages you to ***reflect** on* your values, motivations, assumptions, and thought process—and how that might influence your project.

Thinking in terms of HCD can also help you think of trade-offs in open research. For example, it can help you decide when/what **NOT** to publish openly!

Break (15 min)

Data copyright & licensing

Why do we care about copyright?

- As a data consumer:
  - Understanding what you can do
  - Understanding restrictions
- As a data producer:
  - Understanding what rights you have
  - How to release data

What even is copyright, anyway?

- A system of rights afforded to the creators of original works
- Right to control:
  - Who reproduces the work
  - Who modifies the work
  - Who can make money from the work
  - Who can license the rights

“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
- BSD
  - GPL with knobs on
- Apache
  - MIT but with preserved attribution and change notes

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

Supporting replicability & reproducibility

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

Source: Justin Kitzes “The Basic Reproducible Workflow” (part of tonight’s reading)

Stage 1: Data acquisition

- Where is your data coming from?
- What are TOU or licenses apply to the source data?
- Who collected your data?
- What client-side tools/environments were used to collect your data?
- If your data is a sample, what were the parameters?
- 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/libraries/environments 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 required for your analysis?
- What tools/libraries/environments were 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, it is helpful to consider multiple scenarios

- **Reproducibility:** To what extent could someone other than you with access to the same data but working on different hardware, reproduce the steps in your process and evaluate their results against yours?
- **Replicability:** To what extent could someone other than you, working on different hardware, 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 my study? Would a fellow data scientist feel confident citing my work, even if they didn't replicate it? Would my 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.

Making your work accessible

Licensing code and data

- What do you want people to be able to do?
- Code:
  a. “Anything”
     - MIT
  b. “Credit me”
     - GPL
  c. “Pay me and credit me”
     - aGPL

Licensing code and data

- What do you want people to be able to do?
- Data:
  a. “Anything”
     - CC-0
  b. “Anything, with credit”
     - CC-BY
  c. “Anything, with credit and licensing”
     - CC-BY-SA
  d. And so on

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

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..”

In-Class Activity

Graded, 30 minutes, groups of 5

https://github.com/fivethirtyeight/data

https://github.com/fivethirtyeight/data

- Read through the article - identify the analysis that was performed
- Read through the readme, the data, and the code
- Read through the previous section slides (available on Canvas) and discuss how the project documentation could be improved to support:
  - **reproducibility?**
  - **Attribution and licensing?**
  - **Understanding of the acquisition, processing, and analysis?**
- Write up as many suggestions as you can think of. Make sure to say *why* you're making the suggestion.

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

Break (15 min)

A1: Data curation

Homework due next week

- Readings (read both, reflect on one)
  - Chapter 2 "Assessing Reproducibility" **AND** Chapter 3 "The Basic Reproducible Workflow Template" from *The Practice of Reproducible Research* University of California Press, 2018.
  - Hickey, Walt. “The Dollars and Cents Case Against Hollywood's Exclusion of Women.” FiveThirtyEight, 2014. **AND** Keegan, Brian. “The Need for Openness in Data Journalism.” 2014.
- Additional (optional) readings that are especially good
  - “Examples of well documented open research projects”
  - “Examples of not-so-well documented open research projects”
- Assignment 1: Data curation (5 points, due next Thursday before class)

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#Week_3:_October_12

A1: data sources

Licensing for Wikipedia data

- All the text of Wikipedia pages (including articles), and all public datasets, are available CC-BY-SA.
- Images and other media licenses vary: need to check on a case-by-case basis.
- 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, 2017. 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:** current and historical traffic data
  - **Page counts:** historical traffic data, less granular
- Article traffic by project, access type, agent type, and time interval
- Top viewed articles by project, access type, and time interval
- Unique device traffic by project, access type, and time interval
- ...and more! See also https://en.wikipedia.org/api/rest_v1/ which has even more data

Source: https://www.mediawiki.org/wiki/REST_API

Wikipedia 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 August 1 2008 - July 31 2016

Wikipedia Page View API

- Historical and current data
  - Developed to replace Page Counts; provides more 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 July 1 2015 - yesterday

Differences: Pageviews vs Pagecounts

- 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’. You will need to combine these for your final dataset and visualization
- A couple small arg key/value differences (e.g. Page Counts uses ‘access-site’ arg where Page Views uses ‘access’)

A1: Starting point

1. Log into your GitHub
2. Log into Canvas
3. Go to https://canvas.uw.edu/courses/1174178/modules/items/7880635
4. Click the link! This should import a the repo as a folder called `data-512-a1` into your Jupyter Hub home folder, and automatically open a notebook called `hcds-a1-data-curation.ipynb`

If something goes wrong, contact Jonathan and Oliver (both) right away so we can help you debug. Ask general questions about the assignment on Slack. Attend office hours for more in-depth support. Email Jonathan and Oliver directly if none of the other options address your needs.

Jupyter notebook intro/demo

I've also put together a simple notebook that shows you how to do some basic stuff with Jupyter (including some special features of our Jupyter Hub)

1. Log into your GitHub
2. Log into Canvas
3. Go to https://canvas.uw.edu/courses/1174178/modules/items/7881421
4. Click the link! This should import a folder called `data-512-jupyter-intro` into your Jupyter Hub home folder, and automatically open a notebook called `Jupyter_intro.ipynb`
5. That notebook contains some interactive demonstrations of Jupyter functionality, and links to external resources you can use to learn more about using Jupyter.

A1: Goal

The **goal** for this assignment is to construct, analyze, and publish a dataset of English Wikipedia page traffic from July 1 2008 through September 30 2017

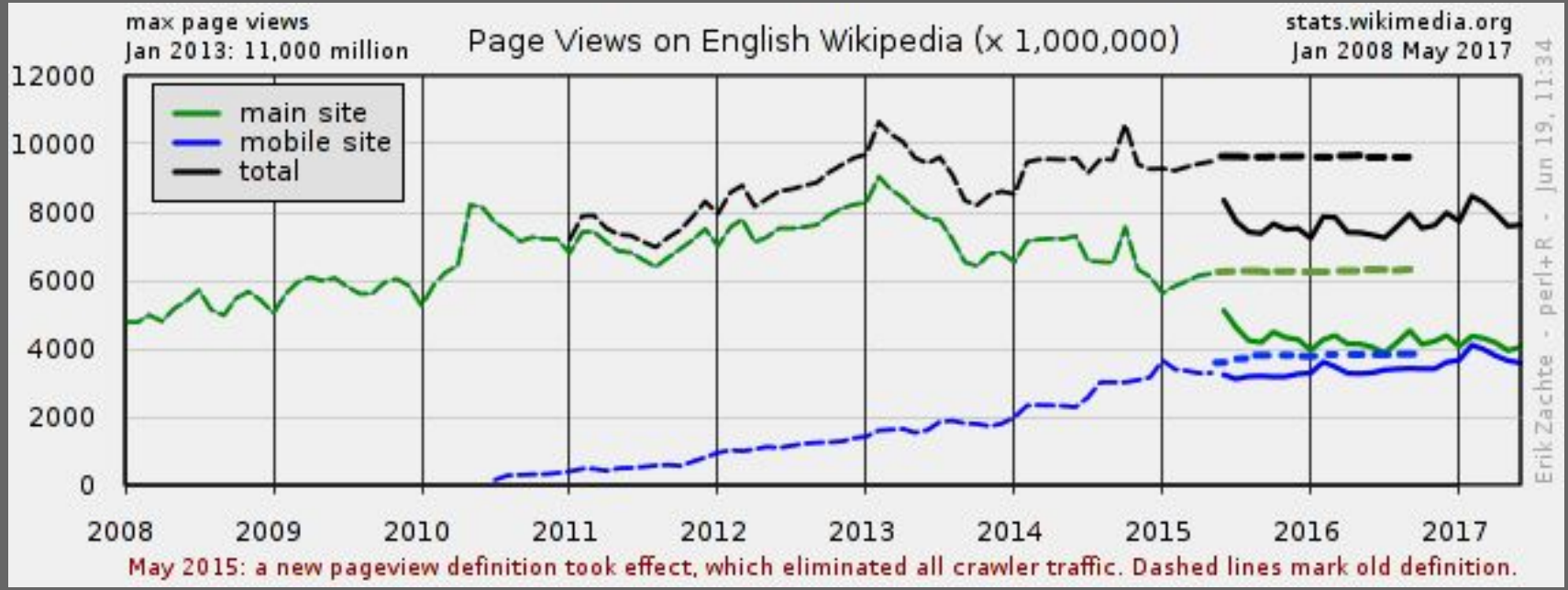
The **purpose** of the assignment is to demonstrate that you can follow best practices for open scientific research in designing and implementing your project, and make your project fully reproducible.

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***.

- **When possible, you must filter out web spiders**, in order to represent ‘organic’ readership traffic to Wikipedia.
- **When necessary, you must combine individual sources of mobile traffic** (e.g. 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.

A1: Analysis



Adapted from: https://stats.wikimedia.org/EN/PlotPageviewsEN.png Erik Zachte CC-BY-SA 3.0

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

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)/Assignments#A1:_Data_curation

A1: Required deliverables

Note: If you use Google Sheet or other open, public data visualization platform to build your visualization, make sure to link to it in the README, and make sure sharing settings allow anyone to view and download the data!

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)/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.
- Use the first day of the following month for `end=` to get the correct monthly count (e.g. `end=20171001` for September 2017 traffic).
- 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.

A1: Submission instructions

1. Complete you Notebook and datasets in Jupyter Hub
2. Download the data-512-a1 directory to your computer
3. Create the data-512-a1 repository on GitHub w/ your code and data
4. Complete and add your README.md and LICENSE file
5. Submit the link to your GitHub repo to: https://canvas.uw.edu/courses/1174178/assignments/3876066

See:
https://wiki.communitydata.cc/HCDS_(Fall_2017)/Assignments#Submission_instructions

A1: Grading rubric

Late work won't be accepted!

Points	Criteria
5	All deliverables completed per requirements.
4	Most deliverables completed per requirements.
3	Some deliverables completed per requirements.
2	A few deliverables completed per requirements.
1	Something resembling the assignment was submitted.

https://wiki.communitydata.cc/HCDS_(Fall_2017)/Assignments#A1:_Data_curation

Questions?

Wikipedia public data: types

- The full content of all (non-deleted) edits to all (non-deleted) pages on every wiki (Wikipedia, Wiktionary, Wikimedia Commons, etc)
  - Articles
  - Talkpage discussions
  - User pages
- Rich metadata about each edit to each page
- Hosted media files (images, audio, video, PDF, etc)
- Some basic metadata about individual editors
- Structured knowledge base (see WikiData.org)
- Aggregated page views