Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Oliver Keyes

Study design | Week 4 | October 19, 2017

Overview of the day

- How Wikipedia works (and doesn't)
- Guest lecture: Dr. Morten Warncke-Wang, Wikimedia
- In-class activity: Q&A and discussion
- *Dinner break (15 min)*
- Review: Week 3 in-class exercises
- Review: Week 3 reading reflections
- Sources of bias in data science
- *Coffee break (15 min)*
- Sources of bias in Wikipedia
- Week 4 homework overview

How Wikipedia Works (and how it doesn’t)

Wikipedia is big, popular, and useful

- 250+ languages
- 5th most popular site worldwide (18 billion page views per month)
- English is the biggest: 5.5 million articles
- Content is used in a lot of ways
  - Quoted and republished all over the place
  - Used for research
  - Used to power AIs (Deep Blue, Google Knowledge Graph)
- It’s very accurate and up-to-date, at least for ‘major’ topics
- It’s very comprehensive, at least for ‘major’ topics (with some big caveats that we’ll get into later)

Wikipedia is written by volunteers

- No one is paid to write articles*
- Anyone can edit
- Millions and millions of people have edited Wikipedia
- Currently, there are about 100k *active editors* (across all Wikipedias)
  - Active editor: 5 edits per month
  - About 30k on English
- Highly active editor: 100+ edits per month
  - About 3k on English
  - ***Most of the content is created by this much smaller group***

Wikipedia can be *revised* by anyone

- Editors don't just write articles. They also decide (with a few exceptions) what *belongs* on Wikipedia, and what doesn't.
- How do they do it? Simplest example: reverting
  - **P1** edits an article
  - **P2** reverts that edit (restores the article to previous version)
  - Perhaps **P1** comes back and reverts the revert
  - Perhaps **P2** reverts that… this is called an *edit war*

How can this possibly work?

- How did P2 decide that P1's contributions *don't belong* on Wikipedia?
- How did P1 decide what information to add in the first place?
- And considering how easy it is for anyone to change anything, how does Wikipedia maintain its accuracy?

Answer: a bunch of different ways

- **Discussion** (talk pages)
- **Rules** (policies and guidelines)
- **Semi-automated tools** (gadgets, dashboards/feeds)
- **Automated tools** (bots)
- **Moderators** (admins)
- **Committees** (ARBCOM)

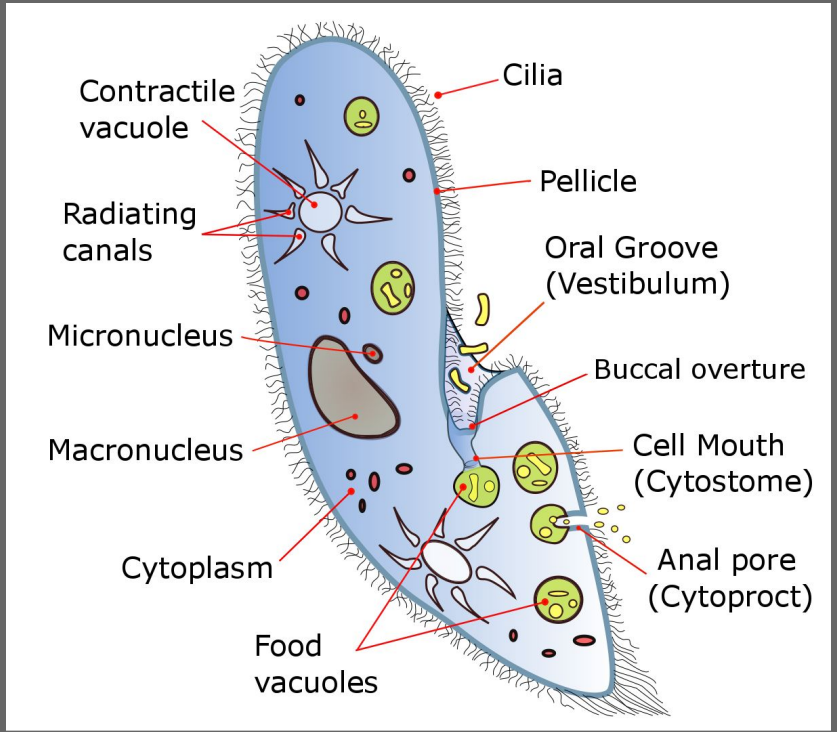
All of this is just as much ‘Wikipedia’ as the articles: you can’t separate the *content* from the *technology*, the *people*, or the *social structures*. It’s all one thing.

Wikipedia: a complex adaptive system

Complex system: composed of distinct, interdependent parts

Adaptive: Adapts to changing environmental conditions

Self organizing: not ‘designed’, build itself

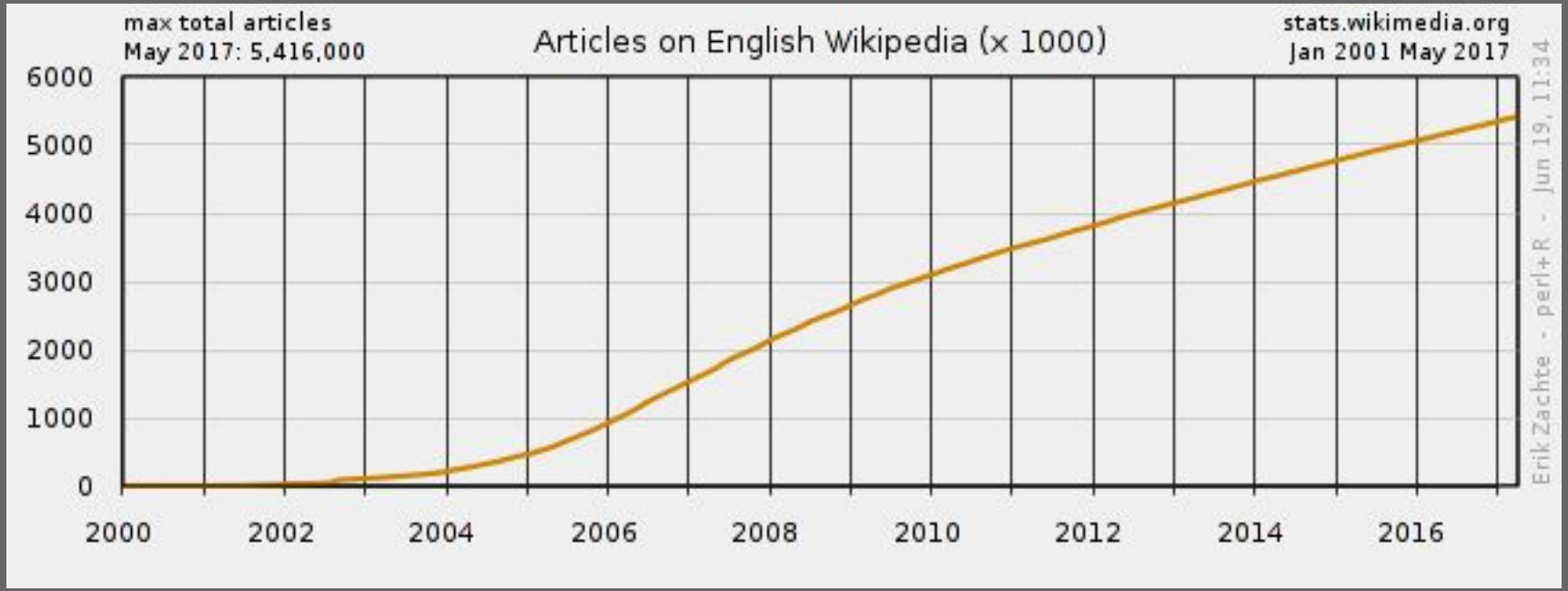


Source: https://commons.wikimedia.org/wiki/File:Paramecium_diagram.png

Wikipedia wants to live

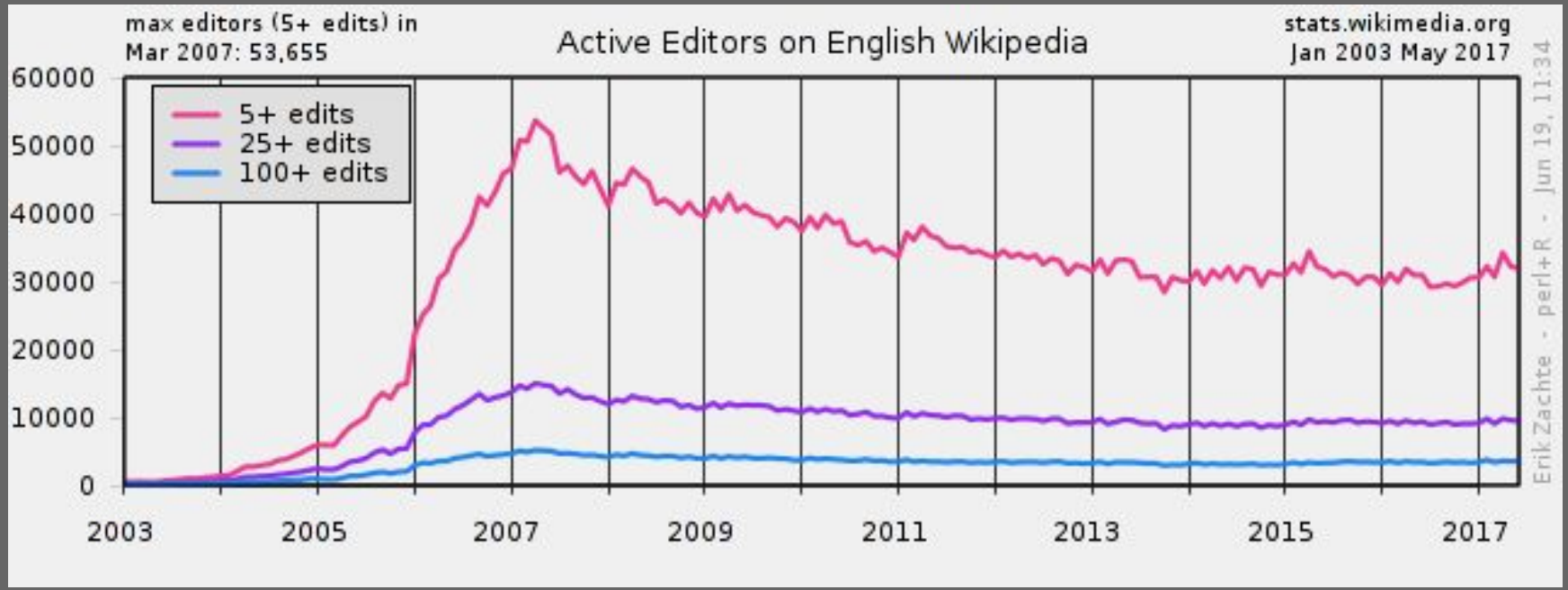
- The first job of any living system: stay alive
- **Homeostasis:** actively keeping internal states within a set of ‘livable’ bounds, even as the organism changes, and the environment changes

Wikipedia has grown



Source: https://stats.wikimedia.org/EN/SummaryEN.htm

Wikipedia has grown



Source: https://stats.wikimedia.org/EN/SummaryEN.htm

The world has noticed

The environment around Wikipedia has changed over 15 years

- More traffic
- More awareness
- More scrutiny
- More people wanting to profit

Wikipedia adapts to its environment

Wikipedia has adapted to these internal and external changes in different ways

- New, specialized quality control mechanisms (adaptive immune system)
- More bureaucracy
  - ARBCOM
  - hundreds of new ‘rules’
  - *“This is how we do things”* becomes *“Do things this way. Or else.”*

Maladaptation

These changes can have unexpected, unpredictable, and often negative consequences

Example: new editor retention

Problem: Wikipedia got popular, and now there are too many newbies! They are writing stuff that doesn't follow all the rules! Some of them are vandalizing articles!

Solution: Assume all newbies are bad, until they prove otherwise.

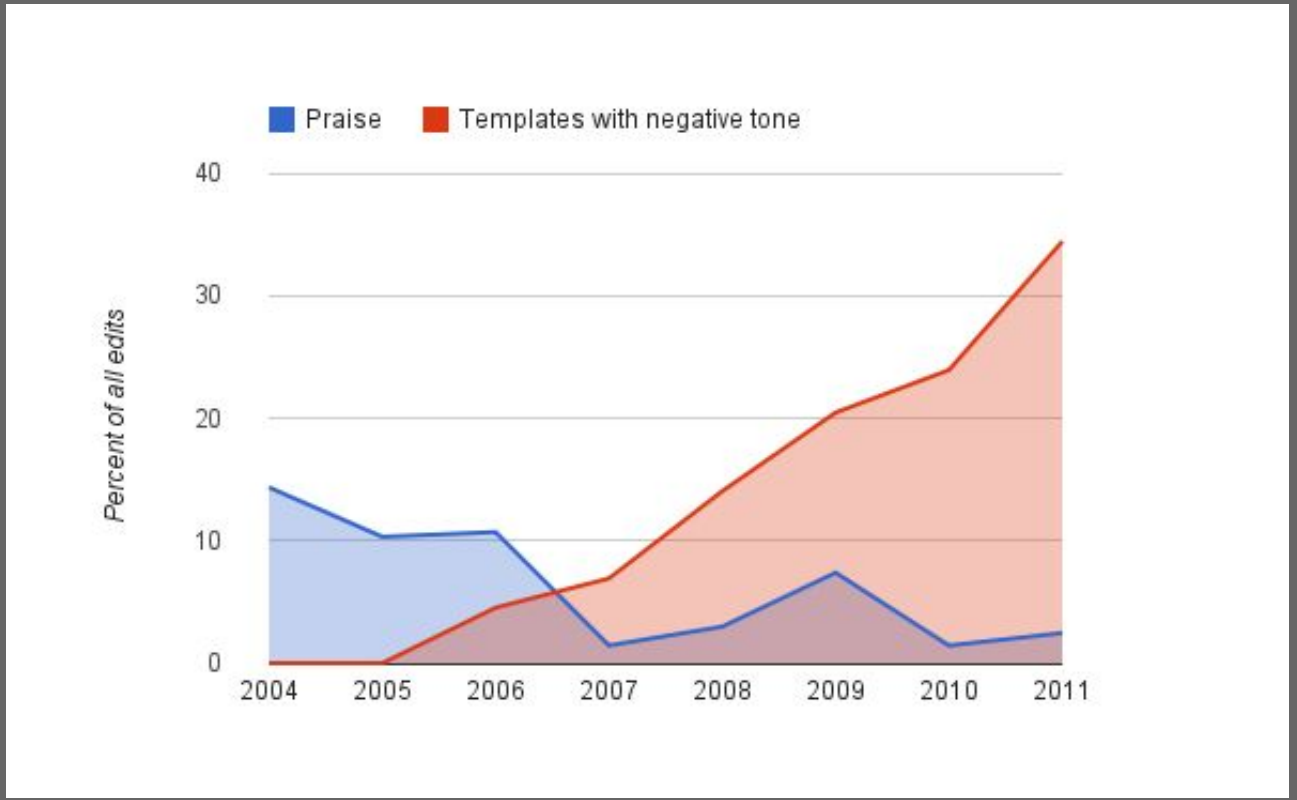
Messages to new editors

October 2017 [edit source]

Please stop adding unsourced content, as you did to Big Hero 6 (film). This contravenes Wikipedia's policy on verifiability. If you continue to do so, you may be blocked from editing Wikipedia. Geraldo Perez (talk)
5:46 pm, 14 October 2017, last Saturday (2 days ago) (UTC−7)

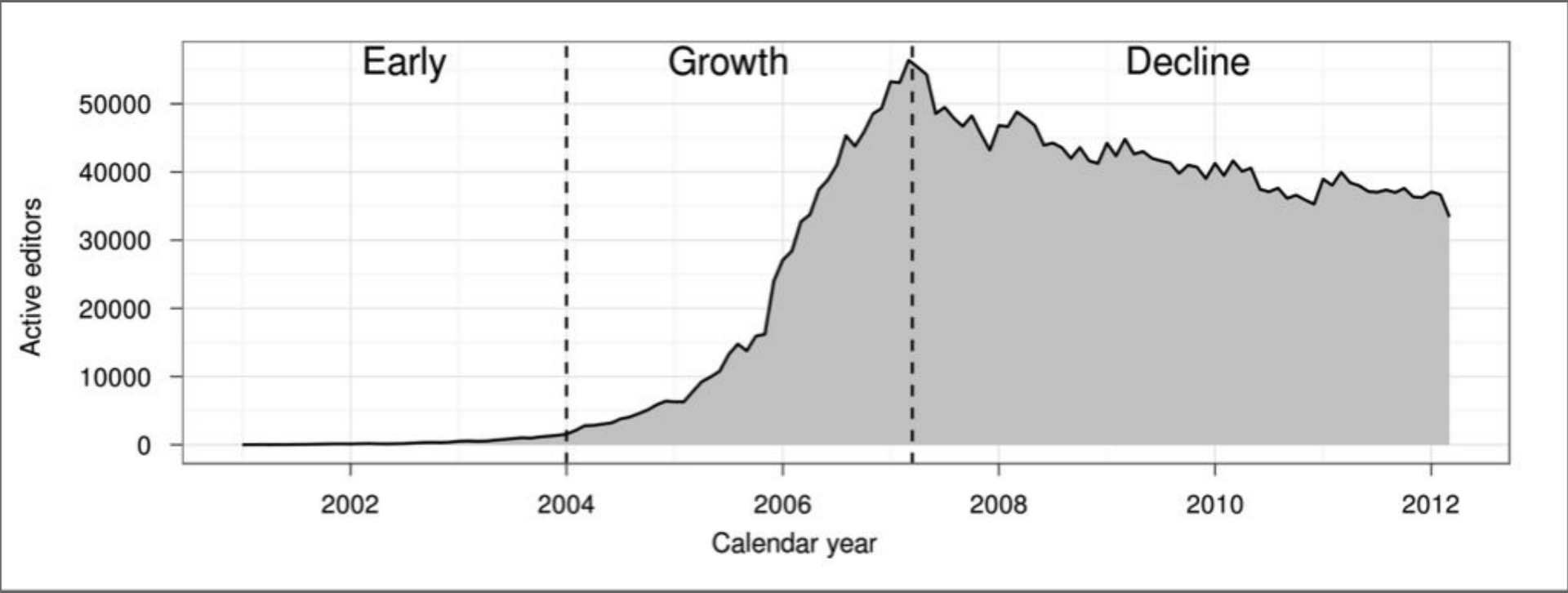
You may be **blocked from editing without further warning** the next time you vandalize Wikipedia, as you did with this edit to List of most expensive animated films. –**Miles Edgeworth** Talk 4:50 pm, Today (UTC−7)

Messages to new editors 2004-2011



Source: https://blog.wikimedia.org/2011/05/02/neweditorwarnings/

of active editors 2001-2012



Source: Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, *57*(5), 664-688.

Variables, correlation, causation

Challenge: how do you perform *data science research* on a system with hundreds of potentially important variables that may or may not relate to one another?

Example: How do we know how much hostility towards newbies contributed to the editor decline? What other factors might have been involved?

- Maybe there was just less and less stuff to write about
- Maybe only ~30k people in the world are interested in writing encyclopedias
- Maybe only ~30k people in the world are capable of writing encyclopedias
- Maybe Facebook came along and distracted everyone
- *[your pet theory here]*

Observer effects

Challenge: how do you perform *data science research* on a system that adapts to changes in external conditions? Publishing your research could change Wikipedia!

Example 1: publishing methods for using public Wikipedia data to identify the geographic locations of ‘anonymous’ IP editors could discourage editing from:

- people who live under authoritarian regimes
- people who just don’t like being doxxed

Example 2: publishing findings that misrepresents Wikipedia (as either better or worse than it is) can affect public perception of Wikipedia, affecting peoples’ decisions to read or edit.

Design interventions

Challenge: how do you perform *data-driven design interventions* on a system that adapts to changes in internal conditions in unpredictable ways?

Even a seemingly trivial intervention can have serious, and often unpredictable, consequences down the line.

Example: *preventing new editors from creating articles*

Lecture:
Dr. Morten Warncke-Wang

In-Class Activity

Graded, ***Individual***

Reflect on the lecture

- Listen to Morten's lecture and take notes
- In at least 2-3 full sentences, answer the question "*How does this lecture inform your understanding of human centered data science?*"
- Using full sentences, list at least 1 question that you have for the Morten.

Submit your reflection and questions to the 'week 4 in-class activity' discussion on Canvas.

Remember: *this is an individual assignment.*

Q&A

Break (15 minutes)

Review: Week 3 reading reflections

Hickey & Keegan

Andrew Enfield: “What can we do to shift people toward releasing more? How can we reward and make it ok to do the '80% solution' without requiring something closer to perfection?”

Gary Greg: “Will standards for data research, and results presentation evolve over time so that expositions such as Walter Hickey's be easily, and widely identified to have a missing component of openness?”

Samir Patel: “since consumers of media may not be interested or technically up-to-speed on the methods used to create analysis pieces, what can be used to help them assess validity?”

Todd Schultz: “I’m left wondering if there still a better presentation for the general audience that has the precision and openness of Keegan’s study with accessibility and appeal of Hickey’s article?”

Rokem et al. & Kitzes

Sha Li: “There are many questions listed in class and in the article to be addressed during each stage of the research workflow. How do we decide which questions would be priorities to be addressed or most relevant when there are ambiguities?”

Rokem et al. & Kitzes

Libby Montague: “Automation is a powerful tool that can allow the user to reproduce the researchers work with one command. However, this can obscure the underlying processes and instead makes the process more akin to calling an API or running executing a piece of proprietary software. How does the research trade off ease of delivering easily followed steps and creating reproducible work?”

Rokem et al. & Kitzes

Alyssa Goodrich: “What is the best alternative to the word “learnings” without using the first person (“Things I learned”). All the words that come to mind (“findings”, “take-aways”) I think have the same grammatical issues as “learnings” (in that they are supposed to be a verb but we turn them into a noun).”

Review: Week 3 in-class activity

US weather history

https://github.com/fivethirtyeight/data/tree/master/us-weather-history
https://fivethirtyeight.com/features/what-12-months-of-record-setting-temperatures-looks-like-across-the-u-s/

Pew religions

https://github.com/fivethirtyeight/data/tree/master/pew-religions
https://fivethirtyeight.com/features/evangelical-protestants-are-the-biggest-winners-when-people-change-faiths/

Study drugs

https://github.com/fivethirtyeight/data/tree/master/study-drugs
https://fivethirtyeight.com/features/college-students-arent-the-only-ones-abusing-adderall/

A statistical analysis of the work of Bob Ross

https://github.com/fivethirtyeight/data/tree/master/bob-ross
https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/

Repeated phrases GOP

https://github.com/fivethirtyeight/data/tree/master/repeated-phrases-gop
https://fivethirtyeight.com/features/these-are-the-phrases-each-gop-candidate-repeats-most/

Classic rock

https://github.com/fivethirtyeight/data/tree/master/classic-rock
https://fivethirtyeight.com/features/why-classic-rock-isnt-what-it-used-to-be/

Obama commutations

https://github.com/fivethirtyeight/data/tree/master/obama-commutations
https://fivethirtyeight.com/features/obama-granted-clemency-unlike-any-other-president-in-history/

???

Group: Niharika Sharma, Khyati Parekh, Sha Li, Elham Rezvani

Sources of bias in data science

Sampling bias

Problem: The sample does not reflect the population it's drawn from

Causes

- Non-random sampling by the researcher (Selection bias)
  - Convenience sampling
  - Human/instrument errors
- Non-random sampling by the participants
  - Non-response bias
- Failure to capture sufficient data
- Failure to capture data with sufficient granularity
- Failure to capture all relevant variables

Biases in study design

Problem: The way the study is designed or conducted biases the findings

Causes

- Data dredging
  - Multiple comparisons w/out hypotheses or statistical correction
- Performance bias
  - Hawthorne effect
- Survivorship bias
  - Attrition bias
  - Excluding outliers

Biases in analysis & interpretation

Problem: The way the data are analyzed or interpreted biases the conclusions

Causes

- Post-hoc analysis
- Confirmation bias
  - cherry-picking
- Publication bias, a.k.a “file-drawer problem”
- ‘Time will tell’ bias
  - Rescue bias
- Over-generalization

Biases threaten validity

Internal validity

- Do the conclusions accurately reflect real relationships in the sample data?
  - Reproducibility

External validity

- Do the conclusions accurately reflect real relationships in the population? Or in other similar populations?
  - Replicability, generalizability

Ecological validity

- Do the conclusions accurately reflect the phenomena being studied outside of a controlled experimental context?
  - Real-world applicability

You can never eliminate bias

- Data are abstractions of phenomena in the real world, made by humans
- Instruments have limited sensitivity, are error-prone, and are made by humans
- *Humans are biased*

Formulating research questions, designing studies, and interpreting results are *inherently subjective* processes.

You can never eliminate bias

"Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."

- Geoffrey Bowker, *Memory Practices in the Sciences (2006)*

You can *sometimes* correct for bias, if you

- understand your data, methods, and *instruments* (including your stats)
- understand your own cognitive biases (including values, beliefs & attitudes)
- solicit input from peers (including subject matter experts)
- follow scientific best practices
- follow open research best practices

You should *always* report known biases

Report any potential limitations of your study design, your source data, or your methods that could bias your conclusions—*even if you think they didn't.*

Exploratory analysis is great!

You can, and should, *explore* the phenomena you're interested in analyzing before you run a formal—*confirmatory*—study. This is what scientists do.

- Exploratory analysis helps you learn...
  - Features of the phenomena that you might want to analyze (possible variables)
  - Faulty assumptions you might have had about the phenomena or the dataset
  - The relative frequency of various measurable quantities in your population
  - Errors or gaps in the data you have available
  - **What's missing in your dataset:** potentially important phenomena that you aren't capturing at all, or not capturing with enough granularity to answer your research questions

Break (15 minutes)

Bias in Wikipedia data

Sources of bias: demographics

English Wikipedia editors are mostly…

- White
- highly educated
- economically privileged
- Liberal (mostly in the ‘free speech’ sense)
- Live in North American and Western Europe
- **Identify as men (80-90%!)**

Geographic coverage (English)

Source: https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en

Geographic coverage (Chinese)

Source: https://iccl.inf.tu-dresden.de/w/images/5/51/Wikidata-20150622-map-items-zhwiki-2880x1440.png

Addressing bias

Problem: Wikipedia editors write about what they know, and what they are interested in.

Solution: Recruit more new editors who...

- know about different things
- are interested in different things

BUT… As a complex, adaptive system Wikipedia resists change. And it may ‘perceive’ external attempts to increase editor diversity as threats to *homeostasis.*

Homework

Homework due next week

Reading for reflection

- Shyong (Tony) K. Lam, et al. 2011. *WP:clubhouse?: an exploration of Wikipedia's gender imbalance.*

Additional (optional) readings that are especially good

- Aschwanden, Christie. “Science Isn't Broken.” FiveThirtyEight, 2015.
- Halfaker, Aaron et al. *The Rise and Decline of an Open Collaboration Community: How Wikipedia's reaction to sudden popularity is causing its decline.* American Behavioral Scientist, 2012.
- Warnke-Wang, Morten. *Autoconfirmed article creation trial.* Wikimedia, 2017.

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#Week_4:_October_19

A2: Measuring bias

Measuring bias

Data

- https://figshare.com/articles/Untitled_Item/5513449
- http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=14

Task

- Merge the two datasets, removing entries that cannot be matched up
- Using the ‘ORES’ system (example coming later), identify the quality of each article
- Visualise how average quality varies depending on country, and how coverage aligns with the population of each country.
- Report back in an iPython notebook

Data limitations

- Some data missing from each category
- Some countries missing (geopolitics sucks)

Measuring bias

- 10 points, due Thursday 2 November before class
- Office hours: Monday, 10am-1pm, Tuesday, 4pm-7pm, Sieg 431
- Use slack/email with wanton abandon
- Reach out if you need help!
- iPython examples (with R support) to come this weekend.

Questions?