Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Os Keyes

Interrogating datasets | Week 4 | October 18, 2018

Overview of the day

- Final project overview
- Reading reflections review
- Sources of bias in datasets
- Introduction to assignment 2: Bias in data
- Sources of bias in data collection and processing
- In-class exercise: assessing bias in training data

Announcements

Make-up in-class activities

If you are not able to make it to class, and have let Jonathan and Os know ahead of time, we can assign you a ‘bonus’ reading reflection to make up for the in-class activity points you missed.

It’s due the same time that in-class activities are due (11:59pm Friday). The format is the same as for all reading reflections.

Submit it to the Canvas discussion called “Make-up reading reflections”
https://canvas.uw.edu/courses/1244514/discussion_topics/4499233

Heads-up: Overlapping due dates

For awkward scheduling reasons, the next two graded assignments will be due a week apart.

A2: bias in data

- Assigned today
- Due in 2 weeks (Week 6, November 1)

A3: crowdwork ethnography

- Assigned next week
- Due two weeks *after that* (Week 7, November 8)

(We'll try to make the reading reflections somewhat lighter to compensate)

Final project overview

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

Final project plan

- Due Week 9 (November 22)
- 10 points
- Min. 1000 words
- Jupyter Notebook or .md file on GitHub, link submitted to Canvas

Final project plan

- 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 **and link to** 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?
- How do human-centered design considerations inform your decision to pursue this project, and your approach to performing the work?

Final project plan

What type of data and analysis?

- Must use publicly-available and appropriately licensed dataset(s)
- Can be a ‘classic’ statistical analysis, or the design and evaluation of a machine learning model
- Use your own definition of ‘big data’!
- Choose datasets and analyses that are likely to support reproducibility
- Choose datasets and methods that let you answer questions that you find interesting and important
- Visualizations aren’t necessary, but encouraged as an effective way of communicating your findings

Final project presentation

- Due Week 11, December 6
- 5 minute oral presentation
- 10 points
- Submit link to Google Slides (or upload PDF) to Canvas before class starts

Final project presentation

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 9 at 11:59pm)
- No min/max word count--whatever necessary to get the job done
- 15 points
- Should contain, and build from, your Final Project plan
- GitHub repo w/ Jupyter notebook, full datasets and documentation; link to repo submitted via Canvas

Final project report

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

- All your code and data, thoroughly documented and reproducible
- A human-centered argument for why your analysis is important
- 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 7 (November 8):** Project plan assigned.
- **Week 8 (November 16):** We'll set aside time to talk over project ideas individually and answer questions.
- **Week 9 (November 22):** No class session (Thanksgiving holiday). Project plan due by 4:59pm.
- **Week 10 (November 29):** 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 6):** Final project presentations.
- **Week 12 (Sunday, December 9):** Final projects are due by 11:59pm. No late work accepted w/out signed Disability Accommodation agreement.

Reading reflections

Duarte, Lanso, & Loup, *The Limits of Automated Social Media Content Analysis (2018)*

Reading Reflections

“I was a bit surprised and taken aback at the strong (too absolute) phrasing of the "recommendations for policymakers." For example, "Use of automated content analysis tools to detect or remove illegal content should never be mandated in law" and "Any use of automated content analysis tools should be accompanied by human review of the output/conclusions of the tool."... If/when automated tools improve to a point that the foibles/mistakes/biases/coercions/frauds etc. perpetuated by the human-centered solutions offered above are even worse than automation...then we would be eschewing automation due to fears rather than solving problems in reality. What is the threshold of "success" we'd want to reconsider never using a content analysis tool without accompanying human review? Do we really want to put rules down now like that? Should we have put parallel rules that demand we never use an automated flight system without oversight from a human pilot?”

-Patrick

Reading Reflections

“Why are the studies for NLP so limited/non-existent for Non-English texts? Is it because this research/development is only being done in English speaking countries? Are English speaking countries the only ones trying to actively enforce censorship of social media posts?”

-Hannah

Reading Reflections

“How should we consider the natural variations of interpretation among human readers? It is well known that whether a message is offensive is determined by not by what was said, but by how it was received; could this be why tools that focus on what was said are limited in effectiveness?”

-Edmund

Key concepts & themes

Duarte, Lanso, & Loup, *The Limits of Automated Social Media Content Analysis (2018)*

1. Domain specificity

Natural language processing tools perform best when they are trained and applied in specific domains, and cannot necessarily be applied with the same reliability across different contexts

What other areas of data science can you think of where data collected in specific domains may be unreliable in other domains?

2. Disparate impacts

Decisions based on automated social media content analysis risk further marginalizing and disproportionately censoring groups that already face discrimination. NLP tools can amplify social bias reflected in language and are likely to have lower accuracy for minority groups who are underrepresented in training data;

Can you think of examples outside of NLP where training data biases may lead to lower accuracy for marginalized subgroups?

3. Inconsistent definitions

Accurate text classification requires clear, consistent definitions of the type of speech to be identified. Policy debates around content moderation and social media mining tend to lack such precise definitions.

What other data science classification tasks might be affected by a lack of clear and consistent definitions?

4. Accuracy and reliability

The accuracy and intercoder reliability challenges documented in NLP studies warn against widespread application of the tools for consequential decision-making.

What are some other kinds of ‘high-stakes’ labeling situations where it might be difficult to achieve high intercoder agreement?

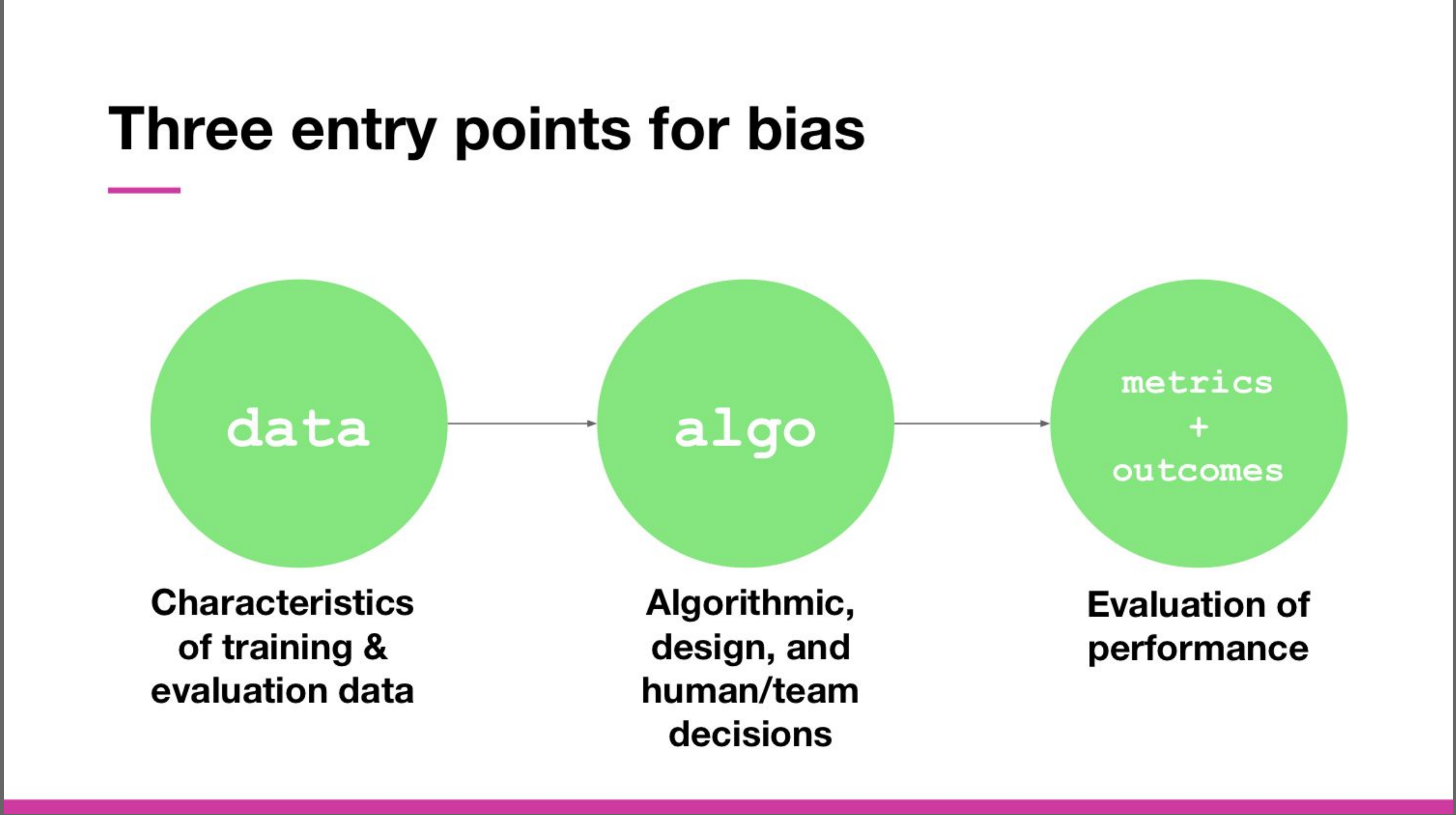
Gamability and comprehension

Text filters remain easy to evade and fall far short of humans’ ability to parse meaning from text.

What other situations can you think of where the context, meaning, intention behind an utterance (or any kind of document) might be important for algorithmic or human decision-making?

Sources of bias in social media data

Olteanu, Castillo, Diaz, Kiciman 2016. *Social data: Biases, methodological pitfalls, and ethical boundaries*



Three entry points for bias

Characteristics of training & evaluation data

Algorithmic, design, and human/team decisions

Evaluation of performance

Source: https://piret.gitlab.io/fatrec2018/program/fatrec2018-garciagathright-slide.pdf

Biases threaten validity

Internal validity

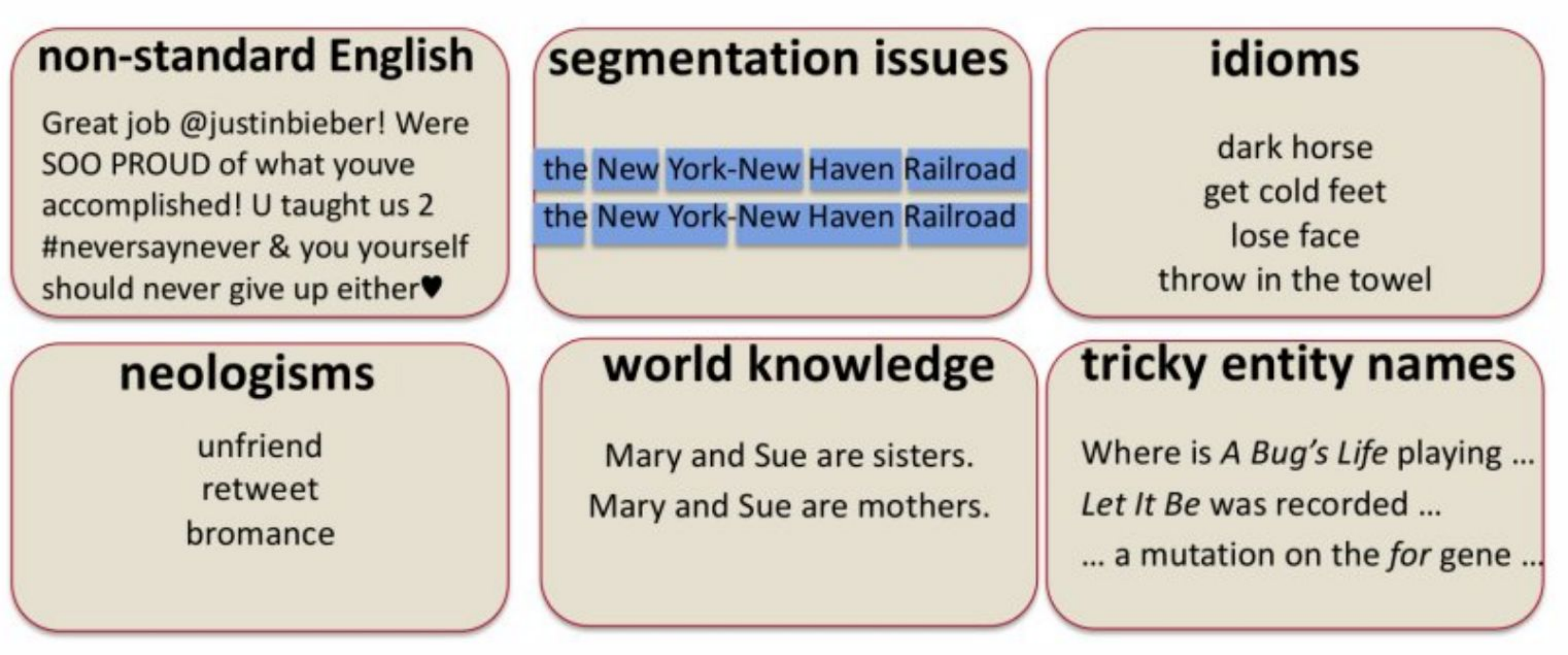
- Do the conclusions accurately reflect real relationships in the sample data? → *reproducibility*

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*



non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

Source: Dan Jurafsky, Stanford CS124 *From Languages to Information.*

Population and platform bias

Population bias: Differences in demographics or other user characteristics between a population of users represented in a dataset or platform and a target population.

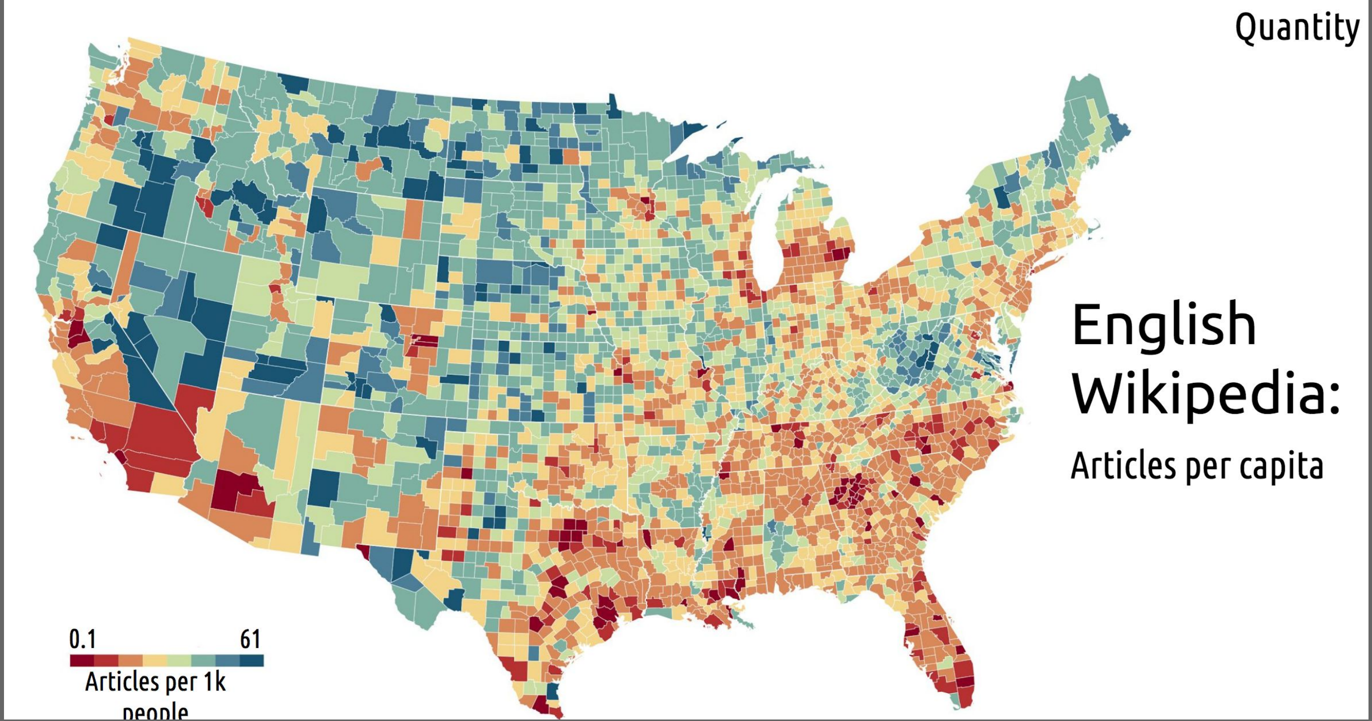
- Using Twitter sentiment to infer national attitudes re: political candidates or public policy
- Using data on public tweets to infer characteristics of all (public and private) tweets
- Using data from Twitter’s public API to infer characteristics of all public tweets

Functional bias: Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.

- Twitter’s character limit causing people to use more *telegraphic speech*
- The many different ways people use “like”, “love”, and “favorite” buttons in social media

Behavioral biases

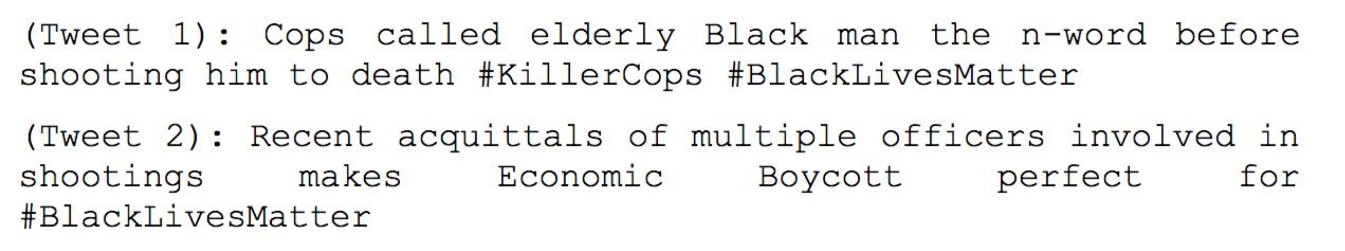
Differences in user behavior across platforms or contexts, or across users represented in different datasets.



Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. DOI: https://doi.org/10.1145/2858036.2858123

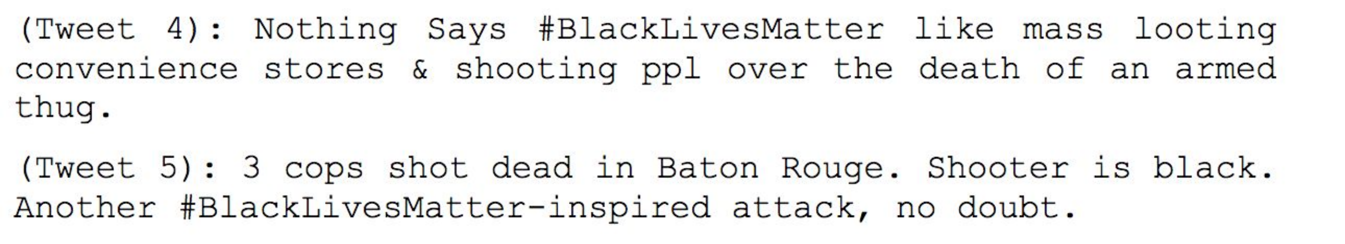
External bias

Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, social pressure, privacy concerns, topical interests, language, personality, and culture.



(Tweet 1): Cops called elderly Black man the n-word before shooting him to death #KillerCops #BlackLivesMatter

(Tweet 2): Recent acquittals of multiple officers involved in shootings makes Economic Boycott perfect for #BlackLivesMatter



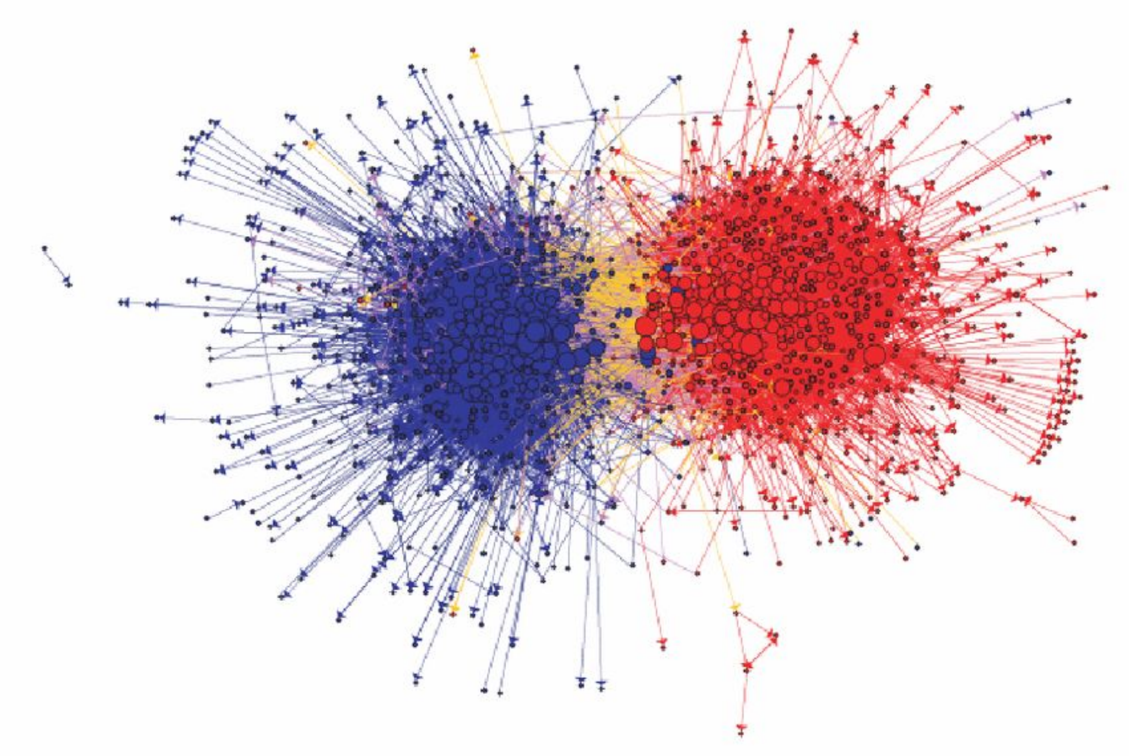
(Tweet 4): Nothing Says #BlackLivesMatter like mass looting convenience stores & shooting ppl over the death of an armed thug.

(Tweet 5): 3 cops shot dead in Baton Rouge. Shooter is black. Another #BlackLivesMatter-inspired attack, no doubt.

Leo Graiden Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. Drawing the Lines of Contention: Networked Frame Contests Within #BlackLivesMatter Discourse. *C*SCW 2017. DOI: https://doi.org/10.1145/3134920

Content production and link bias

Behavioral biases that are expressed as lexical, syntactic, semantic, and structural differences in the contents generated by users.



Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. (LinkKDD '05). DOI=http://dx.doi.org/10.1145/1134271.1134277

Normative bias

Biases that are a result of written norms or unwritten norms describing acceptable patterns of behavior on a given platform.

“Weasel words are words and phrases aimed at creating an impression that something specific and meaningful has been said, when in fact only a vague or ambiguous claim has been communicated… Articles including weasel words should ideally be rewritten such that they are supported by reliable sources.”

https://en.wikipedia.org/wiki/WP:WEASEL

Temporal variation

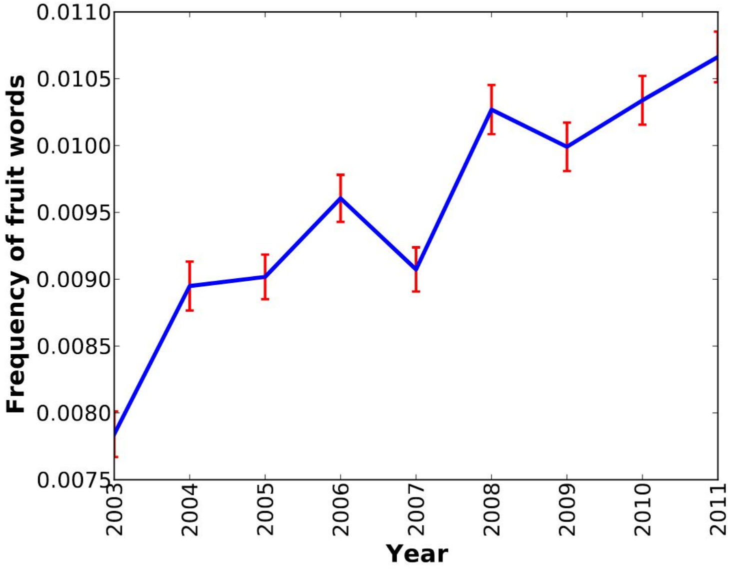


Figure 4: Example of community-level change: The usage of fruit words (e.g., *peach, pineapple, berry*) increases on BeerAdvocate. (Same trend holds for RateBeer.)

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. (WWW '13). DOI: https://doi.org/10.1145/2488388.2488416

Data collection bias

- **How was sampling done?**
    - Ex. political polls that only contact people with land-lines, or that only contact people during dinner time.
- **How were metadata categorized?**
    - Ex. demographic surveys that force people to pick a single answer from a fixed list of identity categories, like gender identity or race.

Sampling bias - common causes

- **Selection bias**
  - Convenience sampling
  - Human/instrument errors
  - Non-response bias
- **Performance bias**
  - Hawthorne effect
- **Survivorship bias**
  - Attrition bias
  - Excluding outliers
- *Failure to capture sufficient data*
- *Failure to capture data with sufficient granularity*
- *Failure to capture all relevant variables*

You can never eliminate bias

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

Gathering data, 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 and 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.*

Reducing bias through documentation

Bender and Friedman, 2018. Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science.

NLP Data Statements

Anatomy of a data statement

1. *Curation rationale*
2. *Language variety*
3. *Speaker demographics*
4. *Annotator demographics*
5. *Speech situation*
6. *Text characteristics*
7. *Recording quality*

“Drawing on value sensitive design, this paper contributes one new professional practice— called data statements—which we argue will bring about improvements in engineering and scientific outcomes while also enabling more ethically responsive NLP technology.

A data statement is a characterization of a dataset which provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.”

Bender and Friedman, 2018. *Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science.*

NLP Data Statements

1. **Curation rationale:** Which texts were collected, and for what purpose? What were the sampling criteria?
2. **Language variety:** What kind of language? What regional or social dialects are represented?
3. **Speaker demographics:** Age, gender, race/ethnicity, native language, SES, etc.
4. **Annotator demographics:** *Technical or subject matter expertise* + age, gender, race/ethnicity, native language, SES, etc.
5. **Speech situation:** Intended audience, time, place, spoken/written, synch/asynchronous, scripted/spontaneous
6. **Text characteristics:** Topic and genre, what is are people talking about, and what forms are they using?
7. **Recording quality:** Fidelity, noise, transcription errors, accidental duplication, gaps

Case study:
Sources of bias in Wikipedia data

Wikipedia as a dataset: Content

- 30+ million articles
    - Full text/media
    - version histories
    - Internal link structure
- 200+ language versions
    - Semantically linking across languages
- Descriptive metadata
    - Categories, named entities, geolocations

Wikipedia as a dataset: Content

- 30+ million articles
  - Full text/media… but not for deleted articles
  - version histories… but not for deleted revisions
  - Internal link structure… constantly in flux
- 200+ language versions
  - Semantically linking across languages… for some content
- Descriptive metadata
  - Categories, named entities, geolocations… for some content only, AND constantly in flux

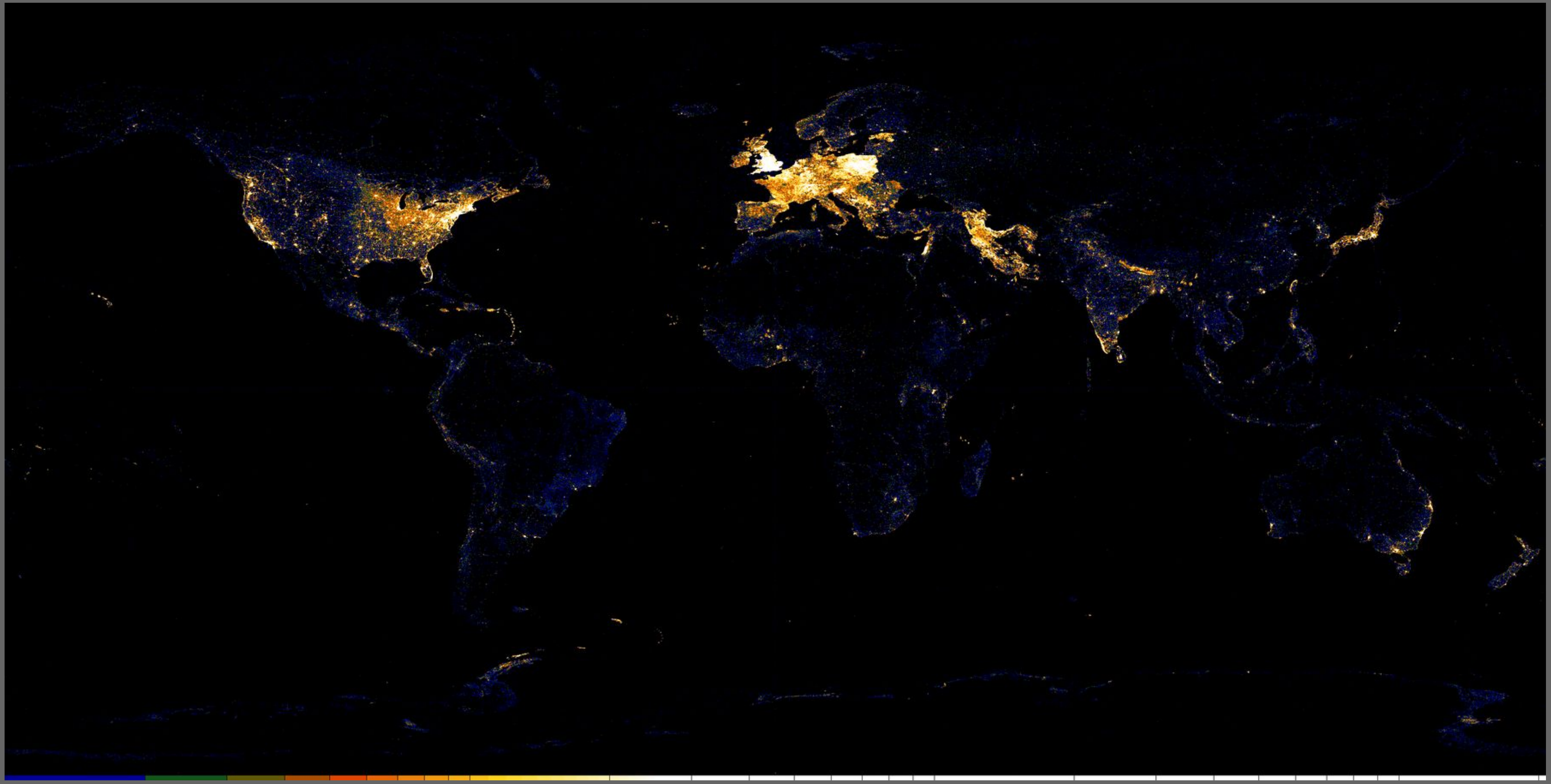
Wikipedia as a dataset: Community

- User metadata
  - Unique identifiers
  - Edit history
  - Join date
  - Volunteered demographic information
- Communication
  - Discussions about articles
  - Meta-discussions
  - Rules, standards, and process documentation

Wikipedia as a dataset: Community

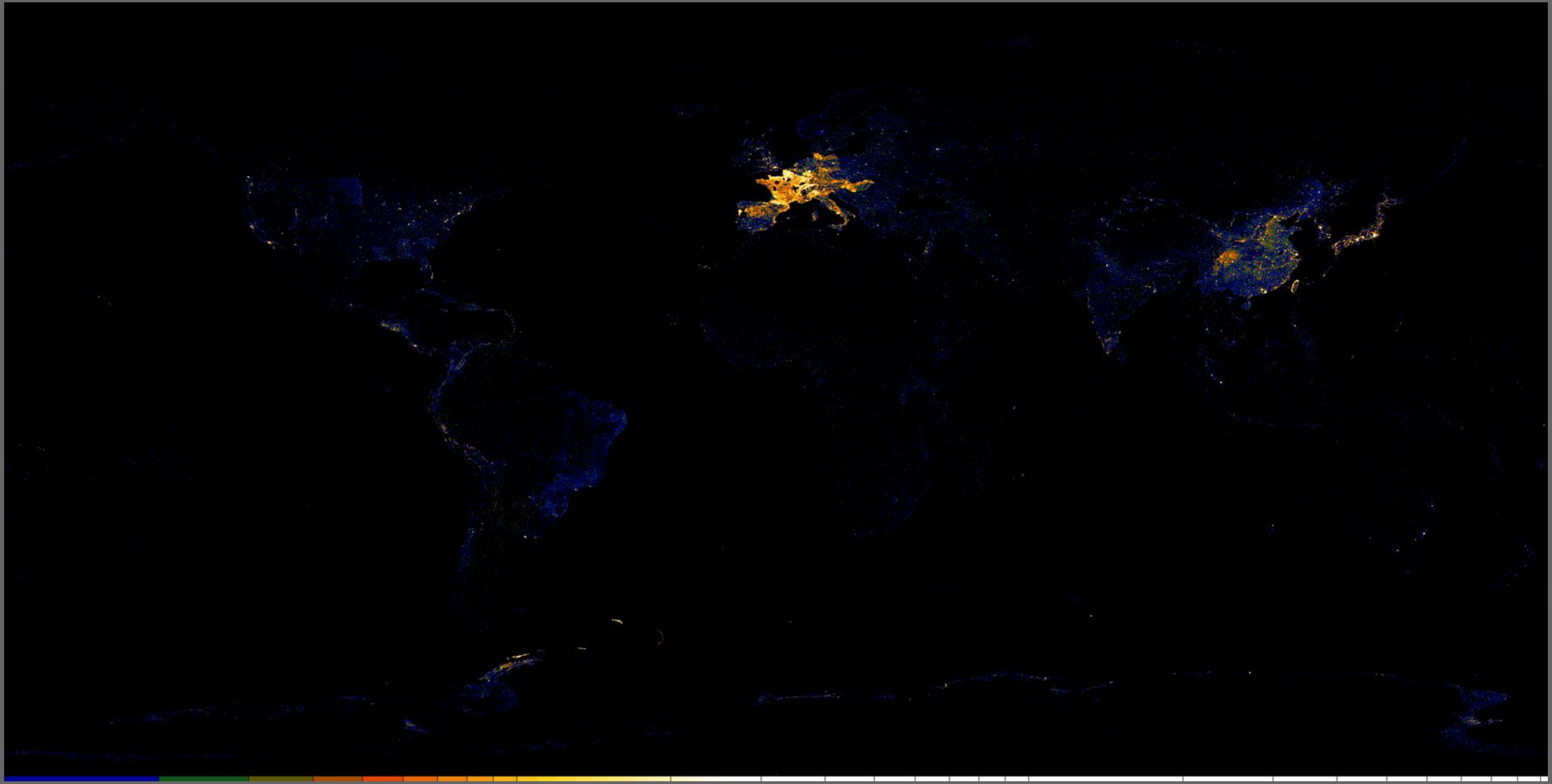
- User metadata
  - Unique identifiers… trivially easy to switch identities
  - Edit history… except deleted edits
  - Join date… trivially easy to switch identities
  - Volunteered demographic information… totally unverified
- Communication
  - Discussions about articles… poorly structured, hard to parse
  - Meta-discussions… poorly structured, hard to parse
  - Rule and process documentation… deeply idiosyncratic

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

A2: Bias in data

Measuring bias

Data

- https://figshare.com/articles/Untitled_Item/5513449
- https://www.dropbox.com/s/5u7sy1xt7g0oi2c/WPDS_2018_data.csv?dl=0

Task

- Merge the two datasets, removing entries that cannot be matched up
- Using the ‘ORES’ system (see examples on the wiki!), identify the quality of each article
    - Model info: https://ores.wmflabs.org/v2/scores/enwiki/wp10/?model_info
    - Model code: https://github.com/wiki-ai
    - Documentation: https://www.mediawiki.org/wiki/ORES
    - Sandbox: https://ores.wikimedia.org/v3/#!/scoring/get_v3_scores_context_revid_model

Measuring bias

Task

- 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, ugh)

Measuring bias

- 10 points, **due Thursday November 1 before class (2 weeks)**
- Office hours: Monday, 6pm-7pm, Wednesday, 5pm-7pm, Sieg 431; additional time available upon request!
- Use slack/email with wanton abandon
- Reach out if you need help!
- iPython examples (with R & Python support) linked on the wiki: https://github.com/Ironholds/data-512-a2

Homework

Homework due next week

- Read both, reflect on one
    - Wang, Tricia. *Why Big Data Needs Thick Data*. Ethnography Matters, 2016.
    - Sen, Giesel, Gold, Hillmann, Lesicko, Naden, Russell, Wang, and Hecht. 2015. *Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards*. CSCW 2015.

Homework due in two weeks

Assignment 2: Bias in Data

- Submit GitHub repository link to: https://canvas.uw.edu/courses/1244514/assignments/4376107

Assessing bias case study

The Wikipedia Talk Corpus and Perspective API

Wikipedia Talk corpus

What it is: An annotated dataset of 1m crowd-sourced annotations that cover 100k talk page diffs (with 10 judgements per diff) for personal attacks, aggression, and toxicity.

Why it is: Discussions on Wikipedia are a crucial mechanism for editors to coordinate their work of curating the world's knowledge. Unfortunately discussions are not only the locus of coordination and cooperation; they are also a major avenue by which editors experience toxicity and harassment… in collaboration with Jigsaw, Wikimedia Research is developing tools for automated detection of toxic comments using machine learning models.

Source: https://meta.wikimedia.org/wiki/Research:Detox

Perspective API

"Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Perspective is an API that makes it easier to host better conversations. The API uses machine learning models to score the perceived impact a comment might have on a conversation."

The Perspective API was trained on the Wikipedia talk page corpus and other public online comment datasets like the New York Times comment section.

Source: https://www.perspectiveapi.com/#/

Perspective API

The Perspective API predicts the toxicity of online comments.

> *“This model was trained by asking people to rate internet comments on a scale from "Very toxic" to "Very healthy" contribution.”*
>
> *“Toxic is defined as... "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."*

Intended use cases include automated and semi-automated comment moderation and filtering.

… Given the training data, how could this backfire?

… What are some other potential use cases?

In-Class Activity

Graded, 1.5 hours, groups of 4-5

Assessing sources of bias in the Wikipedia Talk corpus

In-class activity instructions

You will be randomly assigned one of the 3 datasets listed in the candidates spreadsheet: “toxicity”, “aggression”, and “personal attacks”.

1. Download the corresponding dataset from the Datasets folder on Canvas.
2. Read the instructions Google doc.

Deliverables (post **links** to these in the “Week 4 in-class activity” Canvas thread):

1. A Google Doc (shared with the instructors) that documents the work you performed for each of the three parts of the activity--"Data statement", "Analyze labels", "Document usage considerations"
2. If possible, post any notebooks, code, new datasets, or figures you generated during your analysis to a public GitHub repository and include the link to that repository in your Google Doc.

Choose one person from your team to submit your group deliverables to Canvas.

Homework

Homework due next week

- Read both, reflect on one
  - Wang, Tricia. *Why Big Data Needs Thick Data*. Ethnography Matters, 2016.
  - Sen, Giesel, Gold, Hillmann, Lesicko, Naden, Russell, Wang, and Hecht. 2015. *Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards*. CSCW 2015.

Homework due in two weeks

Assignment 2: Bias in Data

- Submit GitHub repository link to: https://canvas.uw.edu/courses/1244514/assignments/4376107

Unused slides