Human Centered Data Science

DATA 512 — Jonathan T. Morgan & Oliver Keyes & Dr. “The Morten” Warncke-Wang

Machine learning | Week 5 | October 26, 2017

Overview of the day

- Reflecting on week 4 in-class presentation and assigned reading
- A1 review and reflection
- Reading reflections clarification
- A2 check-in and Q&A
- *Dinner break (15 min)*
- Algorithmic black boxes, bias, and ethical AI
- Algorithmic transparency and interpretability
- Evaluating output: Auditing algorithms
- *Break (15 min)*
- In-class activity: ‘toxic’ tweets?
- Overview of A3: Final project plan

Week 4 reflections

Week 4 in-class activity

Are there any questions you wanted to ask Dr. The Morten, but didn’t get to?

Reading reflections - Lam et al. 2011

Andrew Enfield

This article also provides a good template for organizing communication about any DS work. After a typical preamble, the article includes a dedicated section listing and specifying the research questions and hypotheses, including IDs for each. Then the 'Results and Analysis' section reviews each of these questions and hypotheses in detail and in order. Furthermore, the text for each hypothesis is consistently organized into paragraphs first explaining _how_ they tested the hypothesis (with what data, with what statistical techniques, etc.) and then the results of the analysis.

Reading reflections - Lam et al. 2011

Todd Schultz

I think the authors did a great job building their hypotheses on top of each other to have one hypothesis to validate an assumption in another. Validating assumptions will strengthen the trust the readers have and the impact of the conclusions.

How can [women editors] be involved [in follow-up research interviews] while protecting their privacy?

Reading reflections - Lam et al. 2011

Suman Bhagavathula

The authors did ask [what steps can be taken to minimize the gender gap] and attempted a quick suggestion that more research is needed to identify the underlying causes. Have there been any steps taken by wikimedia in that direction?

Reading reflections - Lam et al. 2011

Sha Li

Whether the topic of a Wikipedia article is contentious or not was solely determined by whether it is a candidate for protection or not. I find this argument is a weak assumption.

Reading reflections - Lam et al. 2011

Rex Thompson

I was very surprised to read that the entire study was based on self-reported data for only 2.8% of the population, and of this small fraction, an even smaller fraction (only 25%) was for English Wikipedia. This seemed like an absurdly small fraction of total users, and I couldn't help but wonder if certain populations were more likely to self-report for one reason or another. [several other good points and questions] PS, apologies for the great amount of skepticism. I guess that's just the data scientist in me coming out… :)

Reading reflections - Lam et al. 2011

Aldo Adji

Are the hypotheses, specifically the research questions, itself biased toward proving gender bias exists in Wikipedia community? In other words, do they have more neutral hypotheses disproving gender bias existence?

Reading reflections - Lam et al. 2011

Alyssa Goodrich

I am dubious about the assumption that the population of people who don't reveal their gender has the same gender breakdown as the population of people who do.

I also was not surprised that women had a higher presence in the protected articles. If I were trying to avoid conflict, I would tend to feel more comfortable contributing to articles where I knew that particularly aggressive or malicious people had already been excluded.

I am also not sold on the idea that the lower quality articles are necessarily of more interest to women just because they are more likely to be edited by women.

I would tend to have more confidence in the conclusions of the article if there were more evidence that the subjects of the research were engaged in developing those conclusions.

I am curious about whether there is a difference in the rate of edit reversions among people who later reveal themselves to be female as compared to edits written by someone who is openly female.

Reading reflections - Lam et al. 2011

Erin Orbits

Just because the 48 women who won Nobel Prizes between 1901 and 2017 and the Oscar winners for Best Actress have comparably long articles to their male counterparts, doesn't prove the coverage of the "very important and notable" stuff is all good.

At times, the authors try to move beyond gender stereotypes but they end up with a convoluted linear regression model trying to predict movie article length.

How was this article received at the conference in October 2011?

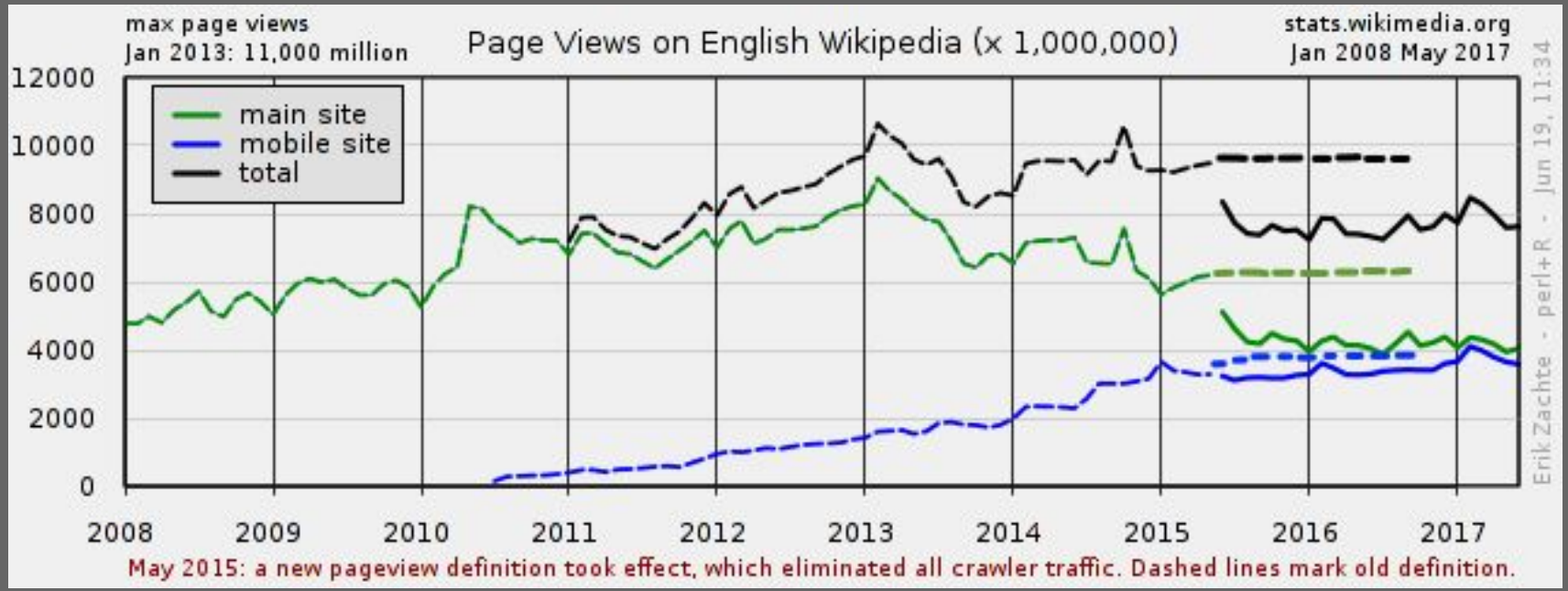
Why wasn't there a single woman among the seven authors?

Why did we read this article?

Break (15 minutes)

Review of Assignment 1

A1: Try to reproduce this graph



...in a reproducible way

Common issues

- Documentation
  - “1 Jupyter notebook named hcds-a1-data-curation that contains all code as well as information necessary to understand each programming step.”
  - “1 README file in .txt or .md format that contains information to reproduce the analysis, including data descriptions, attributions and provenance information, and descriptions of all relevant resources and documentation (inside and outside the repo) and hyperlinks to those resources.”
  - Should be originally written, not from the assignment text
  - Full sentences, in markdown, not inline comments.

Common issues

- Graph
  - “1e10” isn’t clear
  - Not starting at 0
  - No way to distinguish pageviews from pagecounts

Example submissions

- https://github.com/samirpdx/data-512-a1/blob/master/hcds-a1-data-curation.ipynb
  - Great in-notebook documentation
  - Good graph replication
- https://github.com/emontague/DATA_512_Assignment1
  - Great in-notebook documentation (particularly the intro)
- https://github.com/orbitse/data-512-a1
  - Amazing README (above and beyond)
  - Easily readable graph

Reach out if you have problems

- okeyes@uw.edu
- Office hours: 10-12 Monday, 4-7 Tuesday, Sieg 431
- Email, canvas, slack, set up an appointment, anything goes

Reading Reflections

Reflection expectations

- Read the article and:
  - Read all assigned readings.
  - Select a reading to reflect on.
  - In at least 2-3 full sentences, answer the question "How does this reading inform your understanding of human centered data science?"
  - Using full sentences, list at least 1 question that this reading raised in your mind.
- How to answer “how does this reading…”?
- Example: WP:clubhouse

HCD values

- **Ethics**
- **Values**
- **Openness**
- **Communication**
- **Consent**
- **Consequences**
- **Participation**
- **Sustainability**
- **Social impact**
- **Systems thinking**
- **Multidisciplinarity**

This reading informed my understanding by...

- Making me think about the ethics of the methods
- Making me think about the implications of ‘human-centred’ work that doesn’t include any of the humans!
- Data Science to raise awareness of social/systems problems
- Showing how the design of a HCI system can impact people negatively

Assignment 2 Q&A

A2: Measuring bias

Data

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

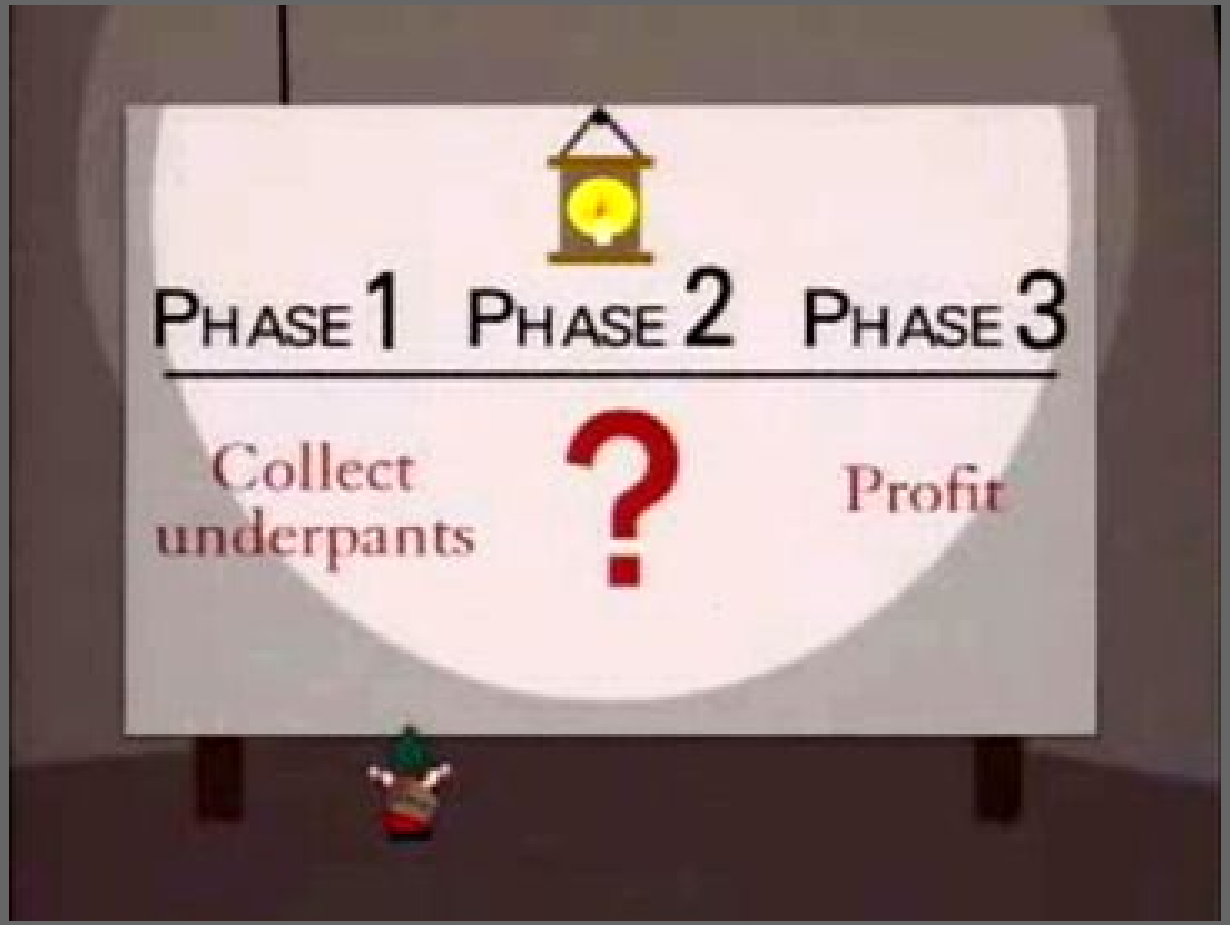
Task

- Merge the two datasets, removing entries that cannot be matched up
- Using the ‘ORES’ system, 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)

Algorithmic black boxes



Source: https://upload.wikimedia.org/wikipedia/en/d/dd/Gnomes_plan.png

Black boxes in science & engineering

Black boxes in everyday life

? $ +/- = $!

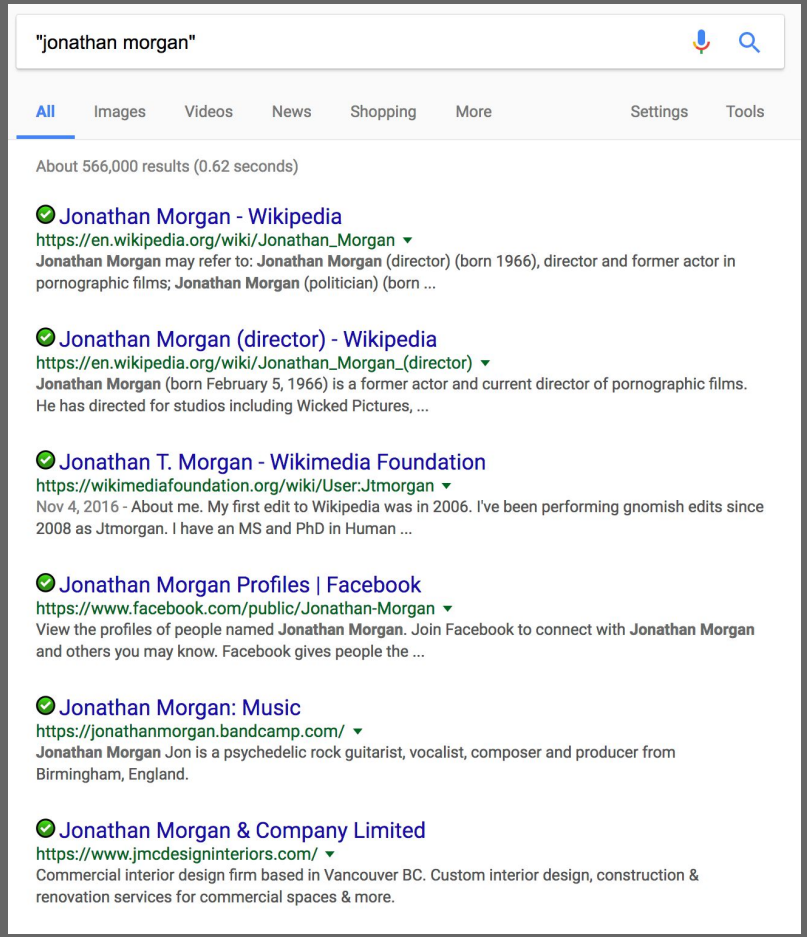
Source: Pixabay.com CC0

Machine learning blackens the box

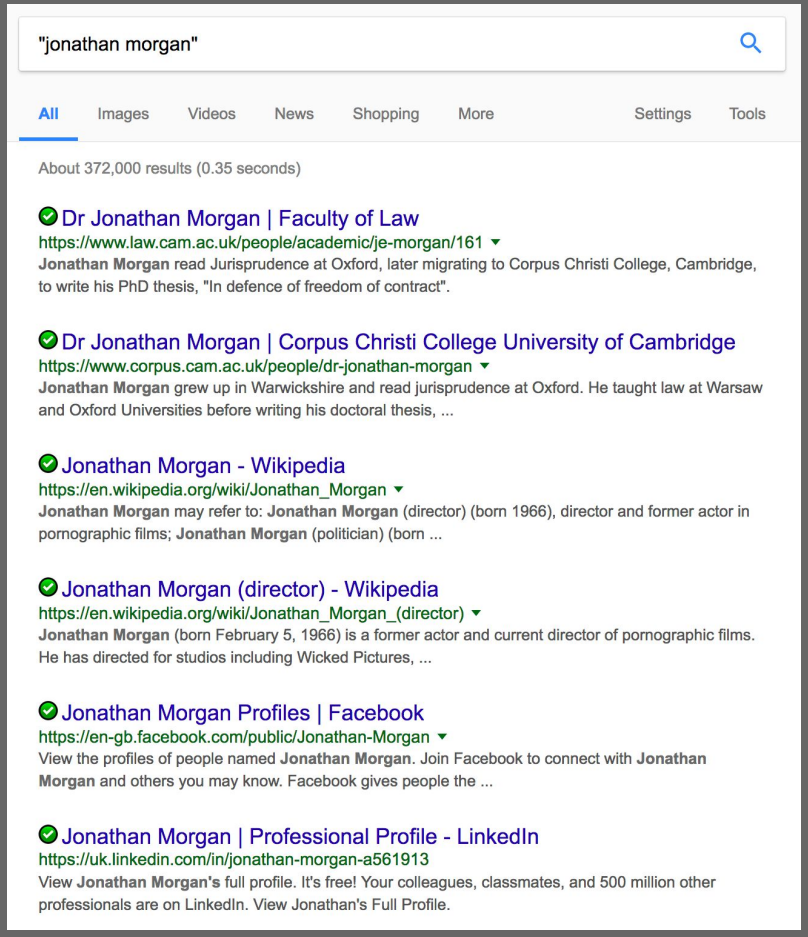
- Input is a dataset, with human-selected features that describe some stuff
- Training data may be labeled by a human (supervised), by another algorithm (semi-supervised), or not labeled at all (unsupervised)
- Output is generally probabilistic, not deterministic
- Depending on the approach used, even the algorithm's designers may not be able to reverse-engineer how a particular prediction/classification was made
- Algorithms are constantly 'tweaked'
- Algorithms are often kept secret

But... we still kinda think we know what's going on, most of the time. Right?

Chrome; Seattle IP; Logged in @wikimedia.org



Firefox; London proxy IP; Private browsing



We make assumptions about black boxes

- We assume a **shared context**
  - When my mom asks “why didn’t you like my post on Facebook?”
- We assume **legitimacy**
  - Even though people constantly try to game the system to deceive or confuse us
- We assume **good faith**
  - Even when Facebook messes with our emotions
- We assume **competence**
  - Even when we’re shown evidence that algorithm designers frequently mess up

Biases are embedded in algorithms

Any model, by definition, is a simplification or abstraction of the real world, made by humans. Bias can be introduced all over the place.

- Sample (training data) does not represent population
- Human-labeled data embeds subjective judgements
- Model may memorize instead of learn (over-fitting)
- Feature selection, weighting embeds assumptions of importance
- Models can degrade without maintenance/monitoring (see: Google Flu Trends)

Social consequences of bias

Biased algorithms can *actively* discriminate against people, the discrimination may be intentional or unintentional (on the part of the designers).

- Ex 1: using race as a features in an actuarial model
- Ex 2: building a face-recognition algorithm that doesn’t recognize darker skin

Perpetuating systemic bias

Even algorithms that don't *explicitly* discriminate based on things like race, gender, etc. can reinforce systemic biases.

- Ex 1: US News College Rankings incentivize colleges to admit wealthier students
- Ex 2: Amazon 1-day delivery less available in non-white neighborhoods

Creating filter bubbles

Algorithms that determine what we see (and don’t) skew our view of the world.

- They decide what information we want and need to see for us
- They hide opposing viewpoints, opinions, and perspectives
- They deprive us of information necessary to make informed decisions
- They create a picture of reality that reinforces our existing beliefs, or change our beliefs without our knowledge

Ethical AI

There is no formal definition of what it means to do AI in an ethical manner, but based on these examples and what you already know about bias, we can derive some ‘best practices’:

- “First, do no harm”
- Get informed consent, when applicable
- Consider sources of bias
- Consider social consequences (discrimination, disenfranchisement, automation)
- Take responsibility for negative outcomes (even unintended outcomes) of what you build, and attempt to address them

What is ‘Human centered design of AI’?

Beyond ethical considerations, there are other aspects of human centered data science that pertain to the development and deployment of algorithms

- Understand audience, purpose, and context of your algorithms
- Design to address human needs, take human skills into account
- Monitor for performance in terms of human benefits
- Make your algorithms as transparent and interpretable as possible

Opening up the black box: Algorithmic transparency and interpretability

Algorithmic transparency

Algorithmic transparency

- The the model (code) and training/test data are publicly inspectable
- Individual decisions are reproducible
- Changes are logged and version controlled

Algorithmic interpretability

Your intended audience can...

- Understand how the model works (input data, features, basic mechanics)
- Understand how specific determinations/predictions/classifications were made
- Glean insights from the model and communicate those insights effectively to *other* non-subject-matter experts

Algorithmic interpretability

There are no hard rules for making an model more interpretable, because it's a context thing (and an audience thing, and a purpose thing…). But some strategies include:

- Use simpler models
- Use more transparent models
- Use concrete features of the input data, rather than composite features
- Use fewer features
- Provide supporting documentation written for non-data scientists
- Make it easy for people to explore various inputs/outputs of your model

Example: ORES

The ORES (Objective Revision Evaluation Service) platform you are using for assessing article quality in Assignment 2 is an example of a very transparent algorithm:

- 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

Example: ORES

ORES is also a fairly *interpretable* model

- The classification task the model performs is relatively straightforward
- The features and scores used to determine the quality of a specific article are presented in a fairly human-readable way
- You can test it out yourself, on real data, to understand its strengths and weaknesses

Example: ORES

But interpretability is contextual. So we still need to ask

- Who needs to be able to interpret this ORES's output (audience)?
- What task do they need to interpret it for (purpose)? and
- How/where/when is it important for them to be able to interpret it (context)?

I don't have a ready answer for this. Fortunately, the designer of ORES will be here next week. Hint: you should ask him questions like these :)

Why be transparent and interpretable?

- It may soon be required by law, in some cases
  - Example: EU General Data Protection Regulation gives *data subjects* of machine learning systems a right to explanation
- Sometimes it's the ethical thing to do
  - Example: Facebook suggests you 'friend' a long-lost relative, but won't tell you what information they used to make the recommendation
- If other people understand your model, they can give you useful feedback

Why be transparent and interpretable?

If your audience believes they understand your model, they are more likely to trust your model, and use your model.

“A model’s total performance is the product of the model predictive performance times the probability that the model will be used. One needs to optimize for both.”

- Carl Anderson, *The role of model interpretability in data science*

Legitimate trade-offs and limitations

There are legitimate reasons scientists and companies do not make their models fully transparent and interpretable.

- You need to use a more complex/opaque model because simpler/more interpretable models don’t perform well enough
- You are concerned about people gaming or undermining the system if they know exactly how the model works
- You are concerned other people could use your model for nefarious purposes
- Your model is your intellectual property and you need to make a living

Illegitimate trade-offs and limitations

There are also *less than legitimate* reasons scientists and companies do not make their models fully transparent and interpretable.

- If people knew how your model worked they would not use your product; you would be publicly shamed and/or arrested.
- You think you can get away with some ‘token’ transparency, which may or may not provide people with accurate or useful information about how your model works.

Your information

Close

About you | **Your categories**

The categories in this section help advertisers reach people who are most likely to be interested in their products, services, and causes. We've added you to these categories based on information you've provided on Facebook and other activity.

- Away from family
- Away from hometown
- Birthday in November
- US politics (very liberal)
- Facebook access (mobile): smartphones and tablets
- Frequent Travelers
- African American (US)
- Facebook access (mobile): smartphones
- Facebook access (OS): Mac OS X
- Gmail users
- Facebook access (network type): WiFi
- Facebook access (mobile): all mobile devices

See More

Evaluating the black box from the outside: Auditing algorithms

History of audits

Audit: “an official examination by an independent body”

Audit study (social science): “a field experiment where researchers participate in a *social process* that they suspect to be corrupt in order to diagnose harmful discrimination” - Sandvig et al. 2014 (tonight’s reading)

History of audits

Audit studies leave the black box intact

- analyzes output based on controlled input
- evaluates output against pre-defined normative or empirical criteria

‘Scholarly consensus’ is that these studies do not require informed consent as long as the benefits to society (revealing and discouraging discriminating behavior) outweigh the potential harms to the people or organizations that were audited:

- Embarrassment
- Wasted time
- Lost reputation or revenue

Types of algorithm audits

From Sandvig et al. *Auditing Algorithms,* 2014

- **Code audit:** code & data published, or shared with independent investigators
- **Non-invasive user audit:** individual users share their interactions with a platform, or allow their behaviors (model inputs) and the results of those behaviors (model outputs) to be tracked or recorded by researchers
- **Scraping audit:** researchers programmatically provide the algorithm with a large variety of different inputs and record the outputs
- **Sockpuppet audit:** researchers programmatically emulate the behaviors of real users and record the result
- **Crowdsourced audit:** researchers recruit a large number of confederates to interact with the platform and track/survey the results

Uses of auditing

- Detecting discriminatory bias (duh)
- Sanity checking results
- Identifying limitations, edge cases
- Evaluating use cases
- Providing feedback to algorithm designers

Limits of auditing

- Scale
- Non-random sampling
- Less control over experimental conditions
- May be in violation of TOU/federal law

Break (15 minutes)

Auditing case study: Google's Perspective API

https://www.perspectiveapi.com

Overview

From perspectiveapi.com:

> *“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 public online comment datasets, including Wikipedia talk page discussions and New York Times comment section

Overview

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.

What are some other potential use cases for an online toxicity detector?

Reception

Wired magazine, as usual, is naively optimistic

> *“The numbers reveal everything from the trolliest time of day to the nastiest state in the union.”*

...which, apparently, is Vermont with neighboring New Hampshire taking the ‘least toxic’ title.

Reception

Violet Blue, writing for Engadget, takes a more critical perspective:

> *“My experience typing ‘I am a black trans woman with HIV’ got a toxicity rank of 77 percent. ‘I am a black sex worker’ was 89 percent toxic, while ‘I am a porn performer’ was scored 80. When I typed ‘People will die if they kill Obamacare’ the sentence got a 95 percent toxicity score.”*

In-Class Activity: Auditing Trump

Graded, Groups of 5, 30 min

Source: http://www.trumptwitterarchive.com/

Toxic tweets

- Read through the tweets in your group's spreadsheet
- Identify at least 10 tweets that you think are mis-classified (more or less toxic than predicted). In a sentence or two, *provide a reason for your decision in the 'notes' column of the spreadsheet.*
- Based on your audit of Perspective, describe in your Canvas post:
    - Some features/characteristics of the text the model *may* be using to predict high- and low-toxicity tweets
    - At least 1 scenario where you think this model could be beneficial, and why
    - At least 1 scenario where you think this model could be harmful, and why

Choose 1 member of your group to submit your results to Canvas before tomorrow night. Include the link to your spreadsheet and the names of all group members.

Homework

Homework due next week

Readings (read and reflect)

- Christian Sandvig, Kevin Hamilton, Karrie Karahalios, Cedric Langbort. *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. ICA 2014*

Assignment 2: Measuring bias in data

- 10 points, due next Thursday before class
- Post a link to GitHub repo to designated Canvas submission form
- Refer to the wiki for full assignment description. Make sure to take advantage of Slack and Oliver's office hours if you need help!

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#Week_5:_October_26

Also assigned this week

A3: Final project plan

- 10 points, due Thursday, November 9th, before class
- Post a GitHub repo link to designated Canvas submission form

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#A3:_Final_project_plan (JM will update this)

Final project plan

- **Length:** Min. 1000 words
  - because it's really about quality, not quantity
- **Deliverable:** Jupyter Notebook or .md file
  - because you'll probably want to re-use a lot of this text in your final report, which will be in a Jupyter notebook

This plan 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 plan

Some important points to hit in your project plan:

- **Motivating your study**: why should we care about this? What do we know about it already? What don't we know?
- **Defining your research questions:** what, specifically, do you want to learn from this study?
- **Framing hypotheses:** based on existing evidence, what do you expect to find about this research question?
- **Describing data sources and preparation:** what data are you using? Where did you came from? What are you doing to it?
- **Describing analytical methods:** what methods are you planning to use, and why are they appropriate?
- **Noting limitations:** anything that threatens the external validity of your study

Final project plan

Choosing a dataset

- Make sure you can legally use the data (don't use datasets of unknown provenance)
- Make sure you have access to the dataset
- Make sure you understand what is contained in the dataset
- Make sure you can process the dataset using tools you're familiar with
- Make sure you can ask (and hopefully, answer) interesting questions with the data

We will provide a list with some potential open online data sources in it tomorrow. If you find a dataset and have questions about whether you can/should use it, ask us—ideally, on the Slack channel so that others can also benefit from our *sage advice.*

Final project plan

We'll have at least 30 minutes next Thursday to talk with you about data sources, research questions, etc. This is the best opportunity to get timely, useful feedback from us before the project plan is due!

If you end up changing your project a little bit after you submit your project plan, you don't need to get approval or submit a new project plan.

- **Example:** you add some new researcher questions, or your change your a bit analysis

If you end up changing your project a lot after you submit your project plan, you *still* don't need to get approval or submit a new plan. **But you should definitely check in with us, so we can give advice.**

- **Example:** you choose a new dataset, or completely change your analytical approach

If you change your plan, make sure your final report describes *why* you did so.

Questions?