Data Science Ethics

C. Alex Simpkins Jr., Ph.D

RDPRobotics LLC,

UC San Diego, Department of Cognitive Science

rdprobotics@gmail.com

csimpkinsjr@ucsd.edu

Lectures :http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

Plan for today

- Review of last time
- Announcements
- Lecture 1: Data Science Ethics
- Groups
- Intro to Project proposal and previous project review
- Lecture 2: Data Science Questions, Visualization Part I

Announcements

Due Friday

- D1 (recommended today)
- D2 (recommended Wed)
- A1 (recommended Wed)
- Q1 (recommended Friday)
- Project proposal (to assigned github repo for course)
- Previous project review (google form - 1 per group)

Notes:

- *Projects*: Repo invitation coming (please accept invitation when it appears!)
- *Labs*: D1 answers (Lectures - XX_lab_answers)/scores will be posted (Canvas)
- *Assignments*: D3 today, D4 Wed, A2 available Wed, due next Friday (not this Friday)
- Groups - we have put the groups together, we will discuss the groups form between the first and second lecture

Data Science Ethics

When working on a data science project, data privacy is the primary ethical concern.

○	○	○	○	○
A	B	C	D	E
Strongly Agree	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Strongly Disagree**

“Big data and analytics technology can reap huge benefits to both individuals and organizations – bringing personalized service, detection of fraud and abuse, efficient use of resources and prevention of failure or accident. **So why are there questions being raised about the ethics [of data science]**?”

Source: IBM

A few examples we have compiled...

- OKCupid Data Published [link]
- Equifax Hack [link]
- Google & Pentagon Team Up on Drones [link]
- Amazon and Police Team Up on Facial Recognition & Surveillance [link]
- Amazon scraps secret AI recruiting tool biased against women [link]

- Study of bias in AI [link]
- Pasco County Algorithmic Bias [link]
- Banjo surveillance via fake apps [link]
- ***Google fires AI ethics founder [link] & Timnit Gebru’s firing [link]***

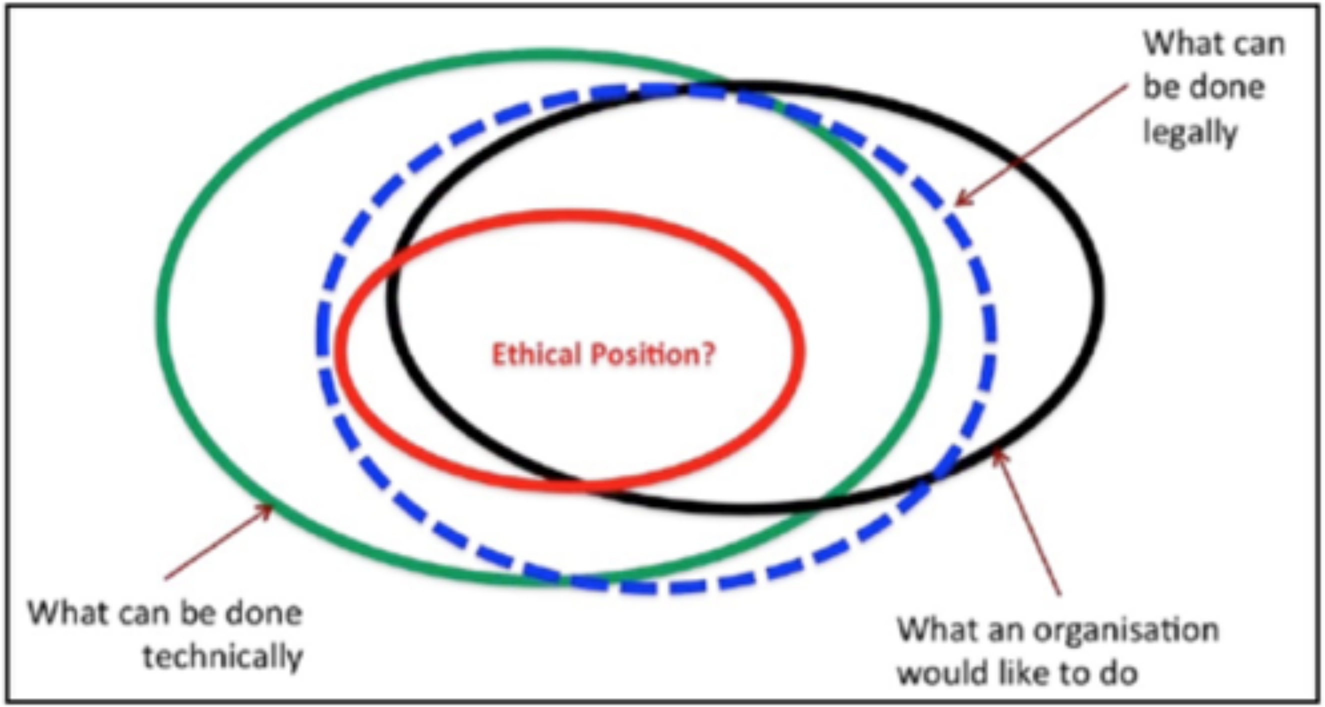
Always consider ethics.

ETHICS

“Moral principles that govern a person's behavior or the conducting of an activity.”

Source: Oxford Dictionary

Big Data Ethics


What can be done legally
Ethical Position?
What can be done technically
What an organisation would like to do

Source: IBM

adapted from Brad Voytek

Ethical Data Science

- Data Science pursued in a manner that
  - Minimizes bias, discrimination and exclusion
  - Respects privacy and consent
  - Minimizes and avoids undue harm now and in the future

On INTENT and OBJECTIVITY

- Intent is not required for harmful practices to occur
- Data, algorithms and analysis are not objective.
  - They are created and executed by people, who have biases
  - They use data, which have biases
- Data Science is powerful
- Bias & discrimination driven by data & algorithms can give new scale to pre-existing inequities, and create new inequalities that never existed

adapted from Thomas Donoghue

NINE THINGS TO CONSIDER TO NOT RUIN PEOPLE’S LIVES WITH DATA SCIENCE

adapted from Thomas Donoghue

NINE THINGS TO CONSIDER TO NOT RUIN PEOPLE’S LIVES WITH DATA SCIENCE

1. THE QUESTION
2. THE IMPLICATIONS
3. THE DATA
4. INFORMED CONSENT
5. PRIVACY
6. EVALUATION
7. ANALYSIS
8. TRANSPARENCY & APPEAL
9. CONTINUOUS MONITORING

adapted from Thomas Donoghue

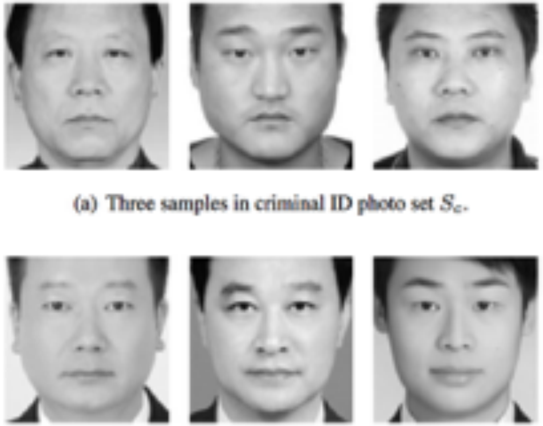
1. THE QUESTION

- What is your question? Is it well-posed?
- Do you know something about the context and background of your question?
- What is the scope your investigation? What correlates might you inadvertently track? Is it possible to answer this question well?

adapted from Thomas Donoghue

Case Study: Labeling Faces

Detecting criminality from faces [link, paper]

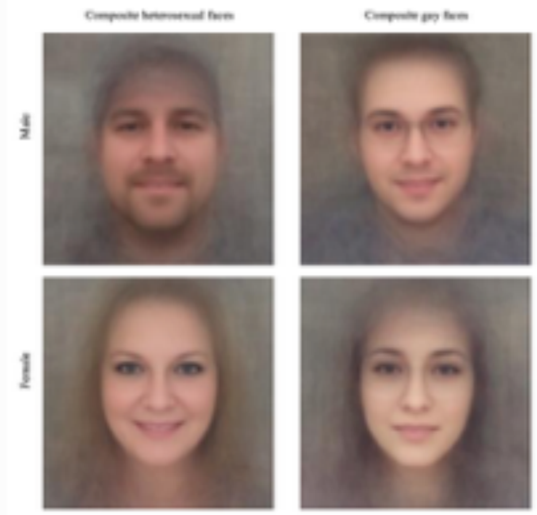


(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

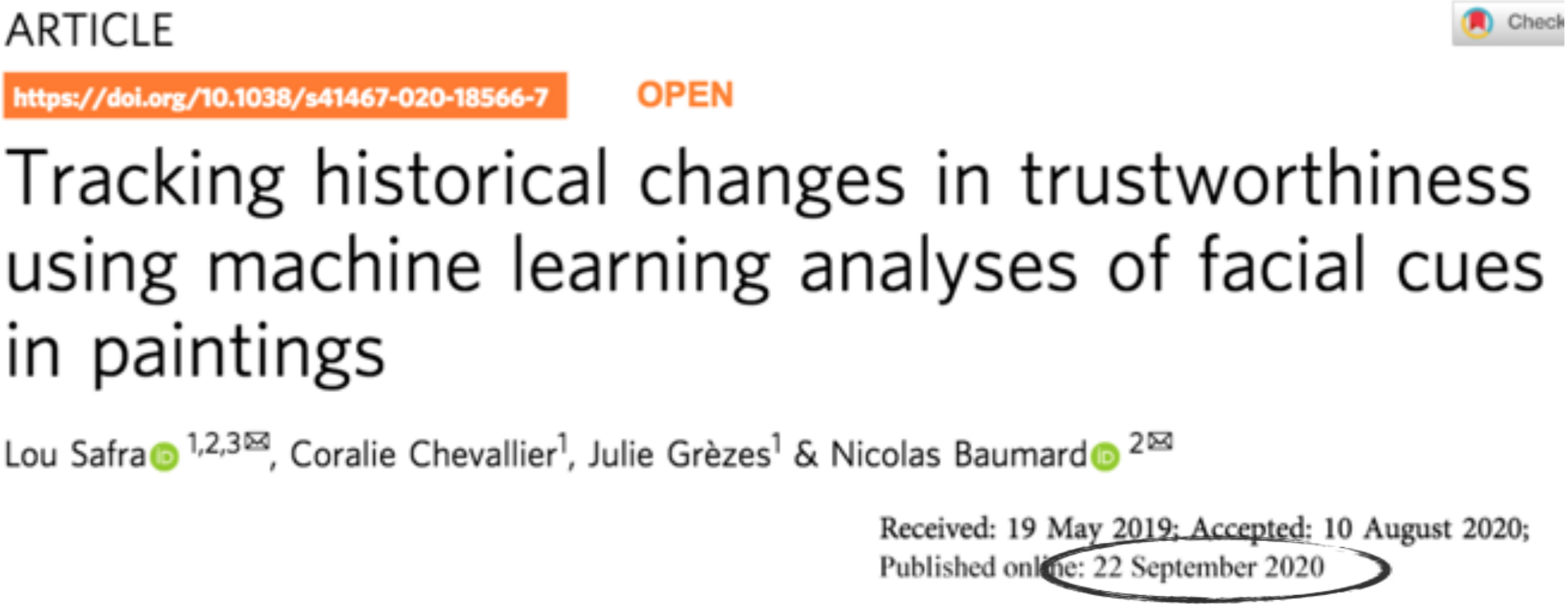
Figure 1. Sample ID photos in our data set.

Detecting Sexual Orientation From Faces with computer vision [link, paper]



adapted from Thomas Donoghue

Published in *Nature*…



ARTICLE

https://doi.org/10.1038/s41467-020-18566-7 OPEN

Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings

Lou Safra[1,2,3]✉, Coralie Chevallier[1], Julie Grèzes[1] & Nicolas Baumard[2]✉

Received: 19 May 2019; Accepted: 10 August 2020;
Published online: 22 September 2020

https://www.nature.com/articles/s41467-020-18566-7.epdf

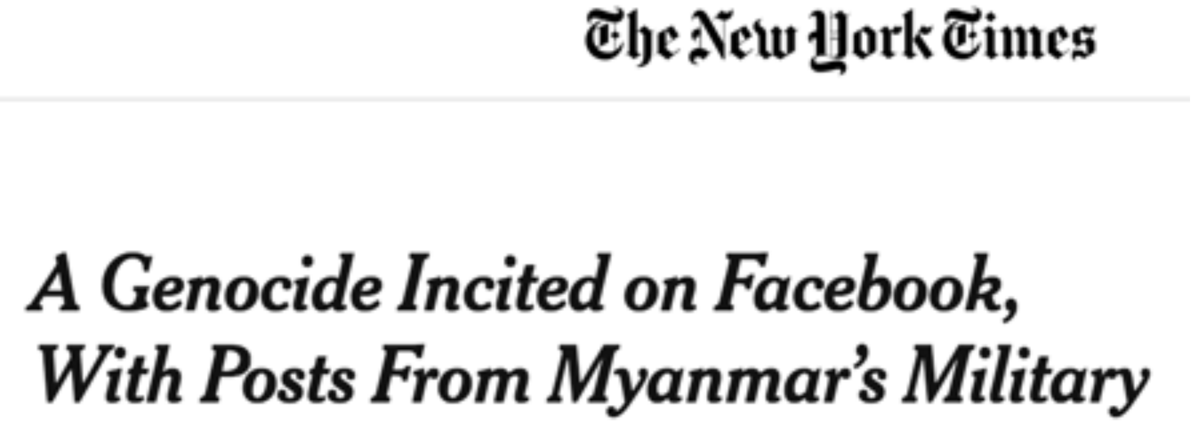
Machine learning does not distinguish between correlations that are causally meaningful and ones that are incidental.

2. THE IMPLICATIONS

- Who are the stakeholders? How does this affect them?
- Could the information you will gain and/or the tool you are building be co-opted for nefarious purposes?
    a. If so, can you protect them from that?
- Have you considered potential unintended consequences?

adapted from Thomas Donoghue

Case Study: Abuse of social networks



The New York Times

A Genocide Incited on Facebook, With Posts From Myanmar's Military

Facebook has been co-opted by military personnel to spread misinformation, hate speech, and promote ethnic cleansing [news link, UN Report]

adapted from Thomas Donoghue

3. THE DATA

- Is there data available? Is this data directly related to your question, or only potentially related through proxies?
- Who do you have data from?
- Do you have enough data to make reliable inferences?
- What biases does your data have?
- If you do not have, and can not get, enough good, appropriate data, you may just have to stop.

adapted from Thomas Donoghue

Case Study: Biomedical Science

Biomedical research has often excluded female subjects

This was based on a (faulty) assumption, among others, that females would be more variable

These findings do not generalize as well

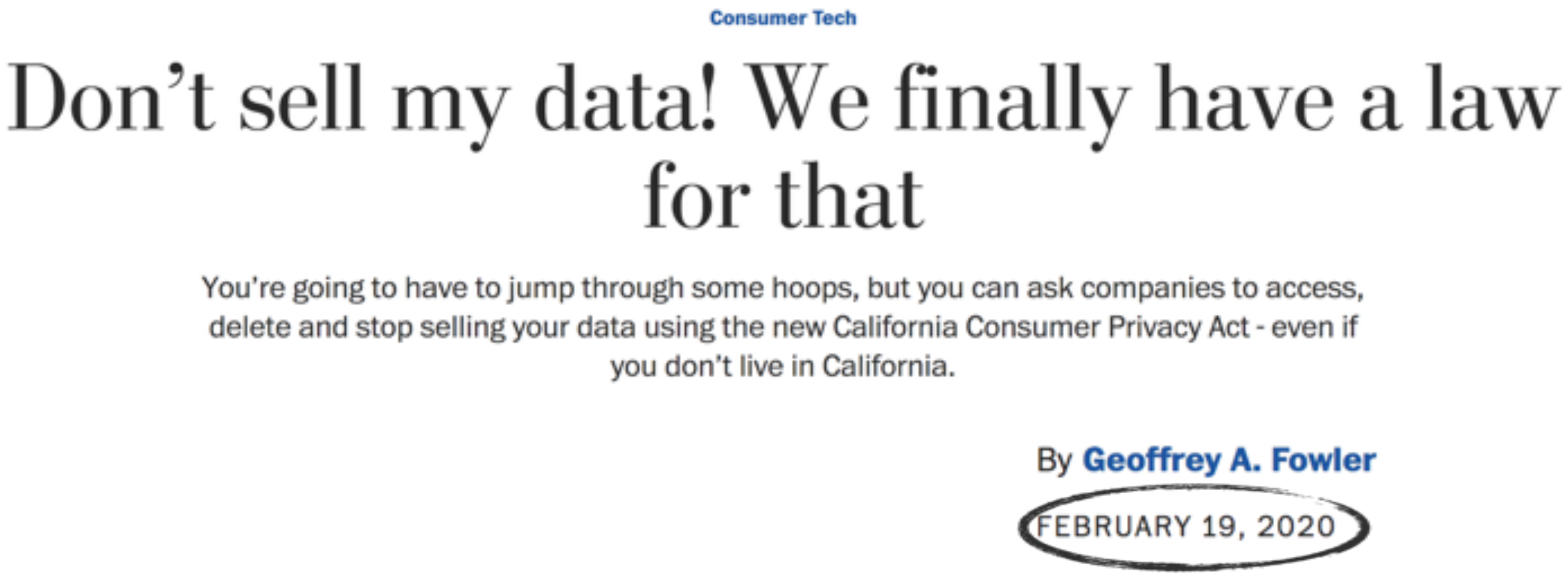
Sources: link, link, link

adapted from Thomas Donoghue

Complete guide to GDPR compliance

GDPR.eu is a resource for organizations and individuals researching the General Data Protection Regulation. Here you'll find a library of straightforward and up-to-date information to help organizations achieve GDPR compliance.

https://gdpr.eu/



Consumer Tech

Don’t sell my data! We finally have a law for that

You’re going to have to jump through some hoops, but you can ask companies to access, delete and stop selling your data using the new California Consumer Privacy Act - even if you don’t live in California.

By **Geoffrey A. Fowler**

FEBRUARY 19, 2020

https://www.washingtonpost.com/technology/2020/02/06/ccpa-

4. INFORMED CONSENT

INFORMED CONSENT: the voluntary agreement to participate in research, in which the subject has an understanding of the research and its risks

Informed consent can be withdrawn at any point in time

adapted from Thomas Donoghue

Case Study: Emotional Contagion

Facebook conducted an experiment investigating whether they could manipulate people’s emotions by manipulating the content displayed on one’s newsfeed. [link, paper]

adapted from Thomas Donoghue

5. PRIVACY

- Can you guarantee privacy?
- What is the level of risk of your data, and how will you mitigate the risks? Are all subjects equally vulnerable?
- Anonymization: the process of removing personally identifiable information from datasets (PII)
- Use “secure” data storage, with appropriate access rights

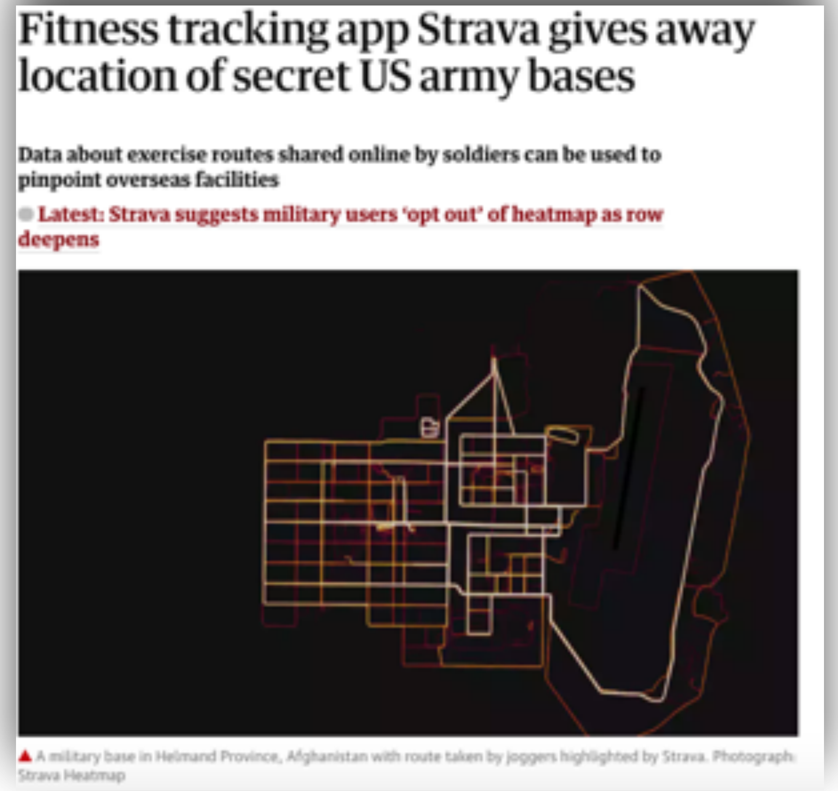
adapted from Thomas Donoghue

Privacy

- Keep in mind
  - **No system is completely secure**
    - *As long as a system exists, it can be accessed, if it can be accessed, data can be compromised [Simpkins]*
  - Cloud-based systems and networked systems have the potential to be compromised ***massively*** and ***rapidly*** if a security breach occurs
  - Be extremely cautious storing sensitive data on the cloud
    - How many of you (do not raise your hands, just consider this) store sensitive information on your mac/pc *and* have iCloud/OneDrive sync on? How many have a smartphone? How much sensitive info is on there?
    - Apple breach, Facebook breach, LinkedIn breach, Equifax breach, Adobe breach,

Case Study: Running Data

Strava, a company who made an app that released running data, geotagged from around the world [link]



Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

● Latest: Strava suggests military users 'opt out' of heatmap as row deepens

▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

adapted from Thomas Donoghue

Case Study: 15 worst data breaches of the 21st century

yahoo, alibaba, linkedin, facebook, marriott, myspace, adobe, apple, equifax, and more [link]

- **YAHOO** - took place in 2013 – announced in December 2016. Account information of more than a billion of its customers had been accessed by a hacking group. Actual figure of user accounts exposed was 3 billion.
- **Linkedin** 2021: 700 million of its user account credentials posted on a dark web forum in June 2021, impacting more than 90% of its user base.
- **FACEBOOK** : In April 2019, 2 datasets from Facebook exposed to the public internet. More than 530 million Facebook users and included phone numbers, account names, and Facebook IDs. Later posted online
- **ADOBE** : 2013 “appears to include more than 150 million username and hashed password pairs taken from Adobe.” Hack had also exposed customer names, password, and debit and credit card information.

adapted from Thomas Donoghue

6. EVALUATION

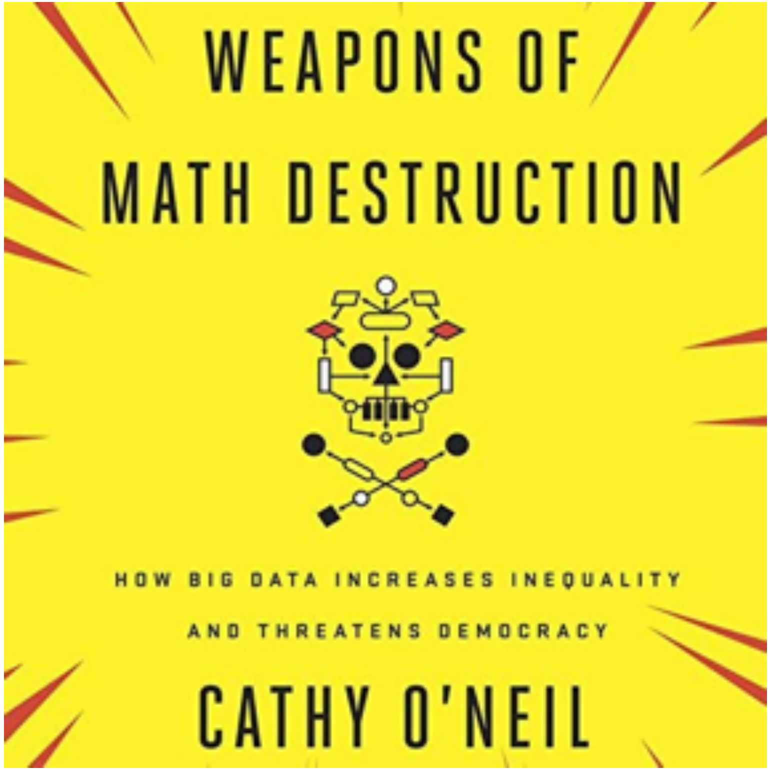
- How will you evaluate the project?
  a. Do you have a verifiable metric of success?
- Goodhart's Law: when a measure becomes a target, it ceases to be a good measure.
  - Named after British economist Charles Goodhart
- In general ***a measurement should not change the system***, or it will tend to produce invalid measurements

adapted from Thomas Donoghue

Case Study: Teacher Rating

Washington, DC school district used an algorithm to rate teachers, based on test scores. Scores from this algorithm were used to fire ‘low performers’

They had no independent measure of whether this measure improved teaching



adapted from Thomas Donoghue

UCSD CAPES

- No independent or objective measure of learning outcome
  - Self report has been determined over decades of research to not be an objective measure
- Not connected directly to any sort of learning, only opinion at the time, no measure of sample quality, sample representation, not sampled in an objective way, students are not made aware of their significance
- No identification of confounding variables or any control
- Yet used as if it is a direct measure as an administrative tool to excuse arbitrary decisions regarding non-tenured instruction
- >40% of part time instructors are not rehired in UC after 1 year, regardless of evaluation or educational outcome
- *Never designed to be an objective measure*, but as feedback for Tenured faculty years ago by students, as there was no such opportunity
  - They are important, but should not be used as a decision criterion

7. ANALYSIS

- Do your analyses reflect spurious correlations?
  a. Remember the rule: “Correlation does not equal causation”
- What kind of covariates might you be tracking?
  a. Are you inferring latent variables from proxies?

adapted from Thomas Donoghue

Example

- Four types of logical errors here: (link)
    - Third unobserved variable
    - x might not cause y, y might cause x
    - Sample size/selection
        - Height bias - might think the average male human height is 182.9cm (~6ft), but that is if you sample from the Netherlands
    - Measurement error
        - People at the front of the room example

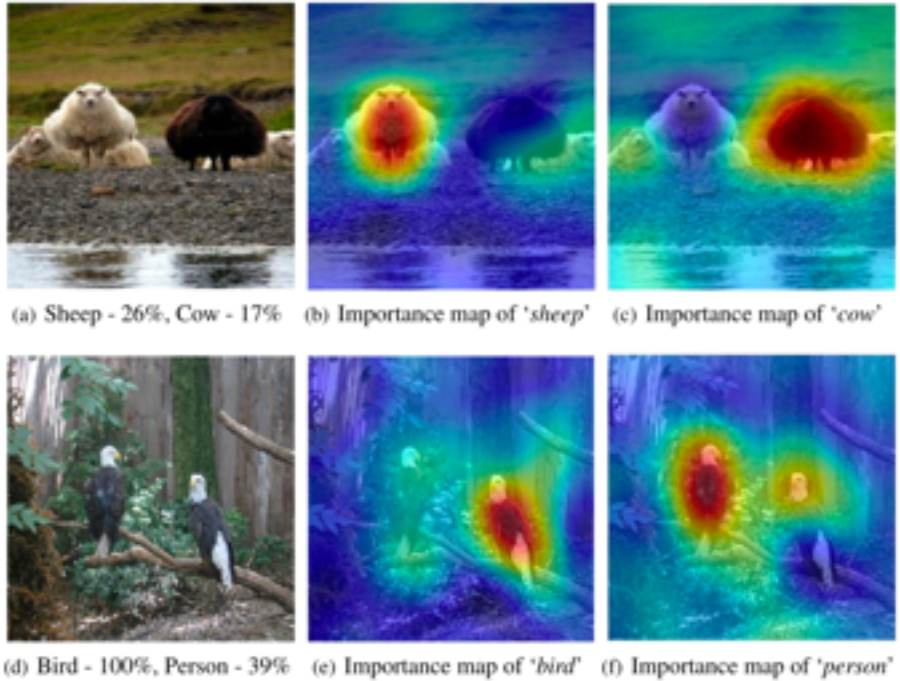
8. TRANSPARENCY & APPEAL

- Is your model a black box?
    - Is it interpretable as to how it came to any particular decision?
    - Do you understand how information and decision process is represented?
- Is there a way to appeal a model decision?
    - What kind of evidence would you need to refute a decision?

adapted from Thomas Donoghue

Interpretability vs. explainability: Widespread AI use concerns

- AI algorithms can be either too complicated for humans to understand how they make decisions or can be proprietary
- **Interpretable** AI - clear explanation of decision making process
  - Decision trees (explained in a later lecture)
  - Linear regression
- **Explainable** AI - don’t give clear explanation of decisions
  - Saliency maps - highlight pixels that contribute to AI’s output
- ***White box, grey box, black box modeling***



(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39% (e) Importance map of '*bird*' (f) Importance map of '*person*'

Source: https://bdtechtalks.com/2020/07/27/black-box-ai-models/

9. CONTINUOUS MONITORING

- Healthy models maintain feedback with the thing(s) in the world they are trying to understand.
- Are you monitoring changes related to your data, assumptions, and evaluation metrics?
- Are you proactively looking for potential unintended side effects of your model or harmful outputs?
- Do you have a mechanism to fix and update your algorithm?

adapted from Thomas Donoghue

Case Study: NEWS SHARING

- Facebook is continuously making predictions about what you are going to do, which it uses to try to influence behavior and then update its models based on the results
- Models optimize for engagement and sharing - can promote the spreading of misinformation

adapted from Thomas Donoghue

ON SYSTEMS & INCENTIVE STRUCTURES

- Novel systems are not, *de facto*, equalizers. They will tend toward propagating existing inequalities.
- Companies/academic institutions/researchers working on these systems may have conflicts of interest with respect to the incentive structures imposed by the system

adapted from Thomas Donoghue

ON PERPETUATING INEQUALITY

- Data & Algorithms can & will entrench social disparities
    - Your care and efforts can offset this
- Errors and bias typically target the disenfranchised
    - Use rigorous methods to minimize this- also intuition
- The combination of damage, scale, and opacity can be incredibly destructive
- They can introduce feedback in such a way as to enact self-fulfilling prophecies

adapted from Thomas Donoghue

PUTTING IT ALL TOGETHER (GOOD)

- Well-posed question that you know something about
- Have considered implications of work
- Adequate data, covering population of interest, with known and manageable biases
- Allowed to use the data
- Have de-identified data, stored securely
- Defined metrics for success, objectively measured
- Cannot establish causality with certainty
- Model is understandable, has procedure for appeal
- Will monitor system for changes, have way & plan to update

adapted from Thomas Donoghue

HOW TO BE BAD WITH DATA SCIENCE

- Ill-posed question you know nothing about
- Don’t consider implications
- Haphazardly collected biased data
- Didn’t check or are not allowed to use data for this purpose
- Un-anonymized, identifiable data, stored insecurely
- No clear metric for success (meh, it ‘seems to work’)
- Present spurious correlations as meaningful
- Model is a black box, no method for appeal in place
- No monitoring, no way to identify biases or update model

adapted from Thomas Donoghue

Integrity

- The quality of being honest and having strong moral principles, moral uprightness.
- The state of being whole and undivided

Integrity

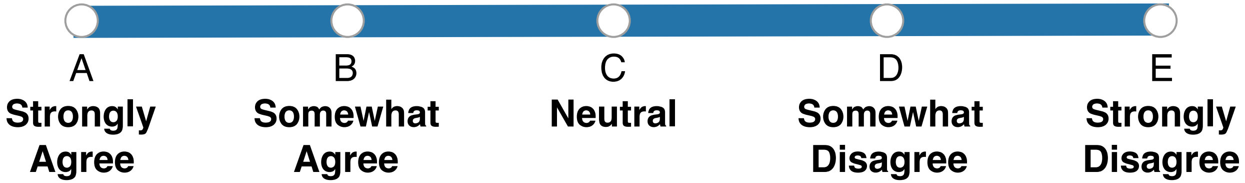
- Integrity is very important
- It can help you make decisions when life gets murky
- Maintain your integrity
- It is difficult to get back once lost (but possible)
- One particular position is less important than your integrity

Integrity

- Doing the right thing leads you in directions you cannot always see, but can trust
- When you are at one with yourself and your moral center, you can make grounded decisions

Data Science Ethics

When working on a data science project, data privacy is the primary ethical concern.



A
Strongly Agree

B
Somewhat Agree

C
Neutral

D
Somewhat Disagree

E
Strongly Disagree

This is traditionally the end of this lecture

- We will take a break, then have a lecture on questions and P1 of visualization
- I have another way to turn this around and get you thinking, we will discuss next time!