

Course Announcements

Due Tomorrow

- D1
- A1
- Resubmission

Notes:

- *Projects*: Repos (please accept invitation!), Grader Assigned, Project to Review Assigned
- *Labs*: D1 answers (Lectures - XX_lab_answers)/scores will be posted (Canvas)
- *Assignments*: A2 available next week, you'll have a longer time

Data Science Ethics

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

• • •

Department of Cognitive Science
rdprobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures : <https://github.com/COGS108/Lectures-Wi23>

A word of caution first...

- WARNING - if you have a tendency toward epilepsy, do not watch these videos as they contain patterns of color and light
- Otherwise enjoy!

What does this have to do with data science?

<https://www.youtube.com/watch?v=fnuSrhGWqu4>

<https://www.youtube.com/watch?v=LhOSM6uCWxk>

<https://www.youtube.com/watch?v=rGwwydEWLlI>

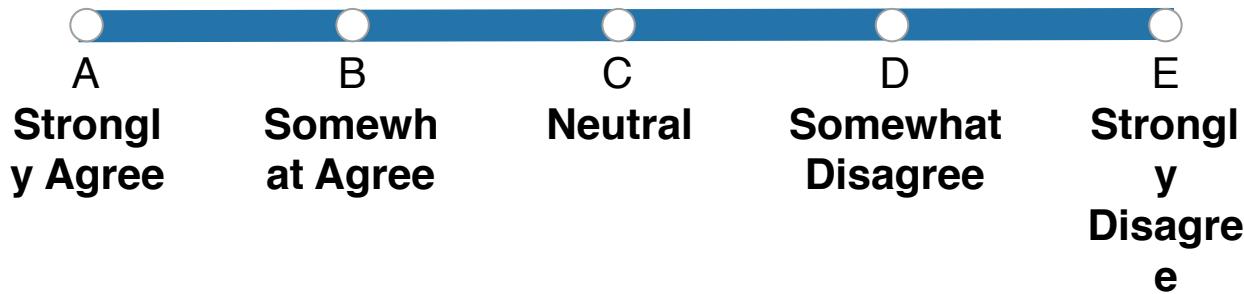
<https://www.youtube.com/watch?v=8YZEp4lhRk>

$$z_{n+1} = z_n^2 + c$$



Data Science Ethics

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



“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]?**”

YouTube vows to recommend fewer conspiracy theory videos

Site's move comes amid continuing pressure over platform for misinformation and extremism

The Reason This "Racist Soap Dispenser" Doesn't Work on Black Skin

Amazon Prime and the racist algorithms

MACHINES TAUGHT BY PHOTOS
LEARN A SEXIST VIEW OF WOMEN

Facial recognition software is biased towards white men, researcher finds

Biases are seeping into software

YouTube's Restricted Mode Is Hiding Some LGBT Content [Update]

Google Translate's Gender Problem (And Bing Translate's, And Systran's...)

COGS 9 Examples

- 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](#)]

A few additional examples we have compiled...

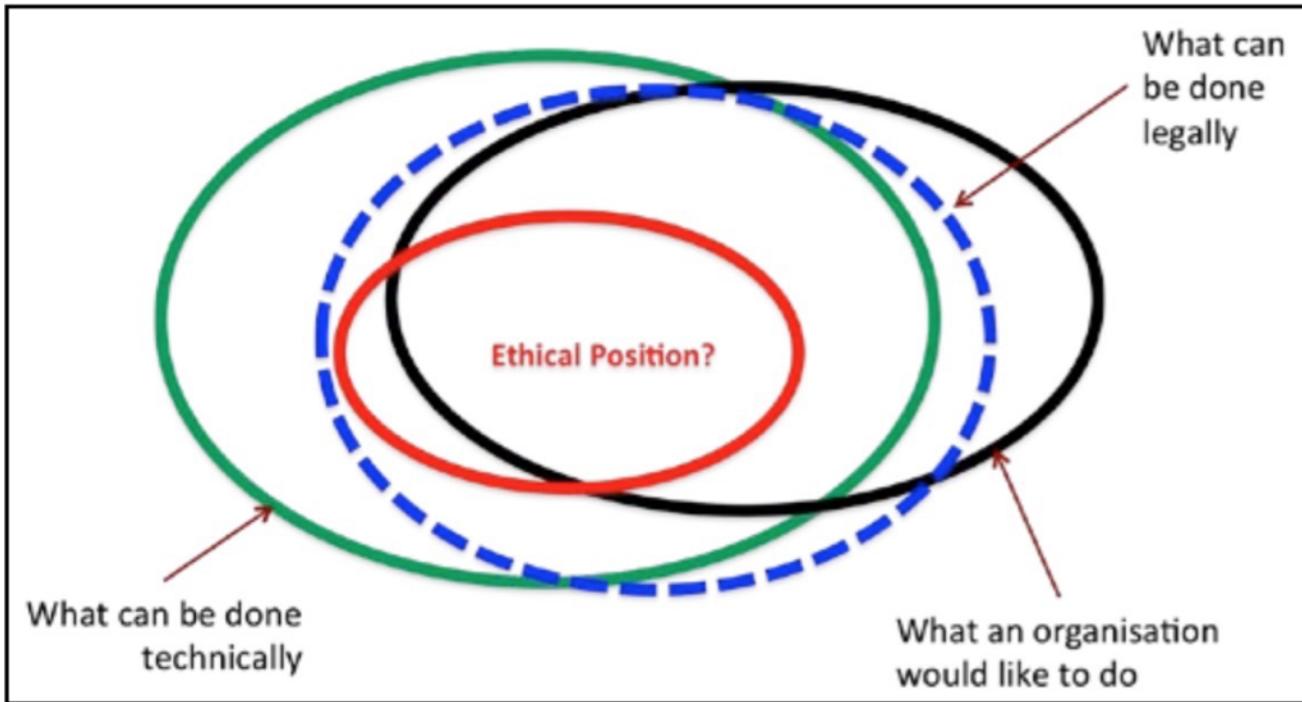
- Study of bias in AI [[link](#)]
- Pasco County Algorithmic Bias [[link](#)]
- Banjo surveillance via fake apps [[link](#)]
- Google fires AI ethics founder [[link](#)] & Timnit Gebru's firing [[link](#)]

Always consider ethics.

ETHICS

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

Big Data Ethics



Ethical Data Science

Data science pursued in a manner so that is equitable, with respect for privacy and consent, so as to ensure that it does not cause undue harm.

On INTENT and OBJECTIVITY

- Intent is not required for harmful practices to occur
- Data, algorithms and analysis are not objective.
 - It is done by people, who have biases
 - It uses 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

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

adapted from Thomas
Donohue

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

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

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?

Case Study: Labelling Faces

Detecting criminality from faces [[link](#), [paper](#)]



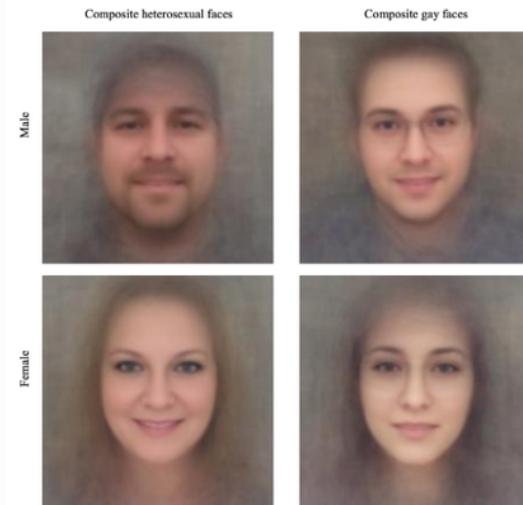
(a) Three samples in criminal ID photo set S_c .



(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

This stuff just doesn't go away...

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 Chevallier¹, Julie Grèzes¹ & Nicolas Baumard²

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

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?

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](#)]

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.

Case Study: Biomedical Science



Biomedical research has often excluded female subjects

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

These findings do not generalize as well

Sources: [link](#), [link](#), [link](#)

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.

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

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

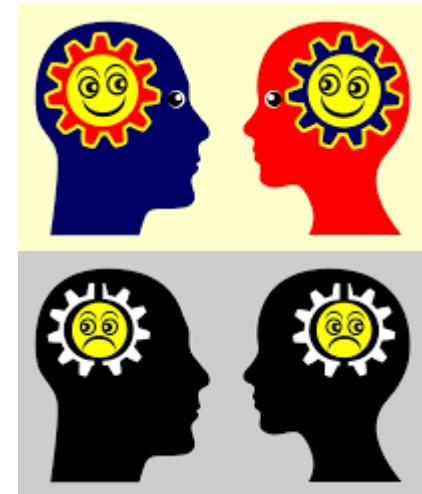
Informed consent can be withdrawn at any point in time



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

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

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

yahoo, alibaba,
linkedin, facebook,
marriott, yahoo,
myspace, adobe, 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 users 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.

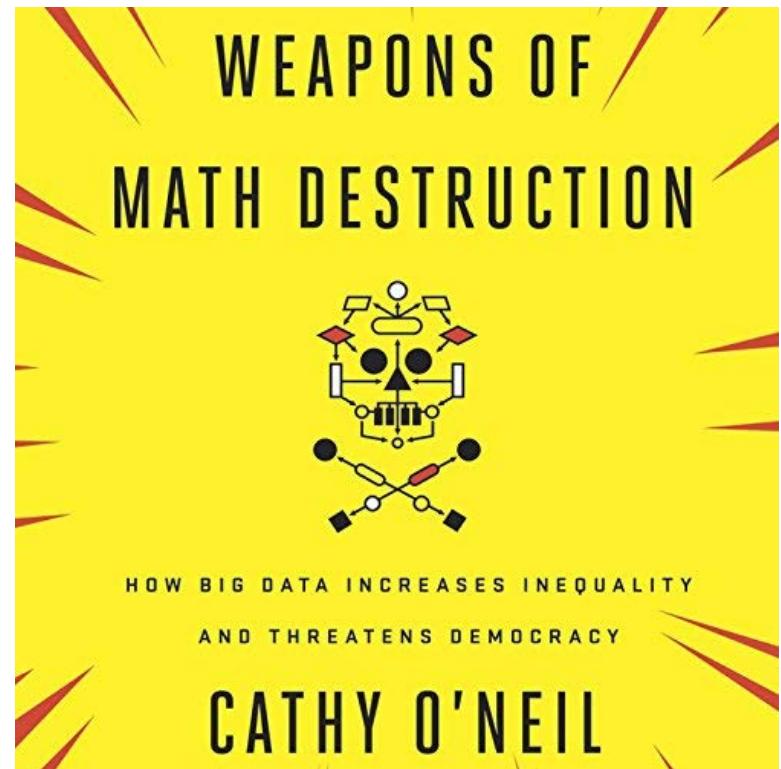
6. EVALUATION

- 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.

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

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?

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?
 - a. Is it interpretable as to how it came to any particular decision?
- Is there a way to appeal a model decision?
 - a. What kind of evidence would you need to refute a decision?

Case Study: Predictive Policing

- Predictive policing uses algorithms to predict crime, and recidivism
- Input data can be highly correlated [[link](#)] with race & SES, reflecting spurious correlations and leading to discriminatory decisions.
- These algorithms and decisions are often opaque and un-appealable.

Two Petty Theft Arrests



VERNON PRATER BRISHA BORDEN

RISK: 3 RISK: 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

adapted from Thomas
Donoghue

9. CONTINUOUS MONITORING

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

Case Study: NEWS SHARING

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



ON SYSTEMS & INCENTIVE STRUCTURES

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

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

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

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

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



Data Science Ethics

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

