

Plan:

Cover the final 5 ways to *not* ruin peoples
lives with data science

Data Science Ethics II

Shannon E. Ellis, Ph.D
UC San Diego



Department of Cognitive Science
sellis@ucsd.edu

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**

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

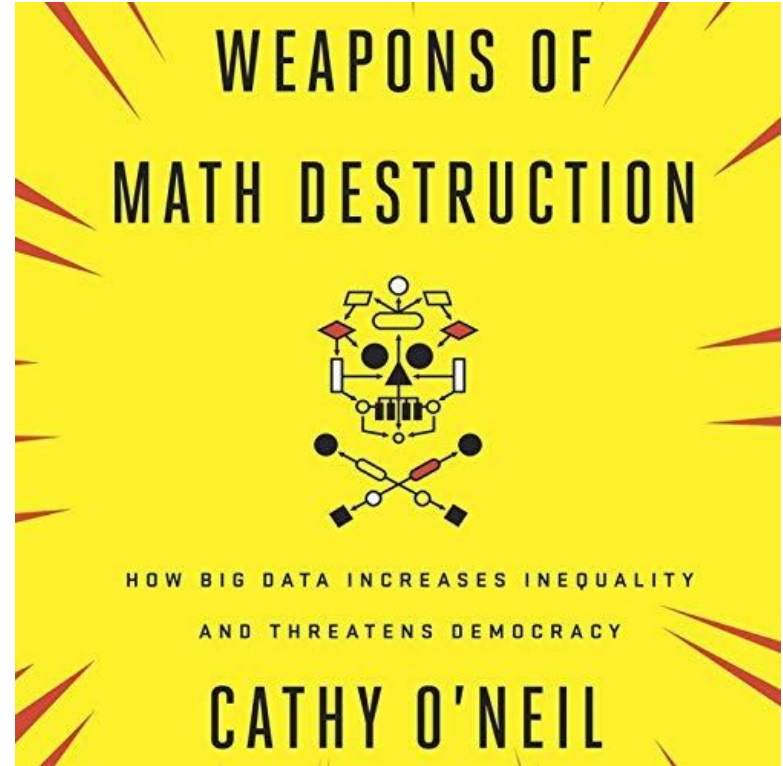
6. EVALUATION

- 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



7. ANALYSIS

- Do your analyses reflect spurious correlations?
 - a. Can you tease apart causation?
- What kind of covariates might you be tracking?
 - a. Are you inferring latent variables from proxies?

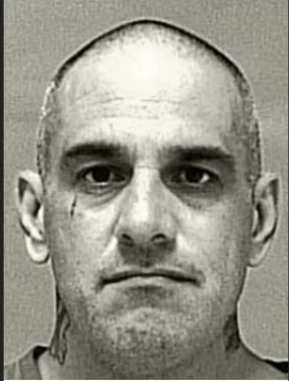

8. TRANSPARENCY & APPEAL

- 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.

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
- Errors and bias typically target the disenfranchised
- 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
- if suggesting causality, have actually established causality
- 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

COGS 9 Examples

- Ashley Madison Hack [[link](#)]
- OKCupid Data Published [[link](#)]
- Equifax Hack [[link](#)]
- Google & Pentagon Team Up on Drones [[link](#)]
- Cambridge Analytica Data Breach To Influence US Elections [[link](#)]
- Amazon and Police Team Up on Facial Recognition & Surveillance [[link](#)]
- Amazon scraps secret AI recruiting tool biased against women [[link](#)]