Welcome to COGS 108!

Data Science in Practice

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Course Info

- Section starts week 2 (no section this week)
- No Wed Office Hours this week (Professor Ellis)
- I don’t have any control over the waitlist.

Scheduling

Lecture: MWF 12-12:50 OR 2-2:50

Office Hours:

Date & Time	Location	Instructional Staff
M 3-5PM & W 3-5PM	CSB 243	Professor Ellis
M 2PM-3PM	TBD	Akshansh Chalal
Tu 11AM-12PM	TBD	Mayank Rajoria

Professor Ellis

What’s your goal?

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

cognitive scientist

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

data scientist

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you're studying to be a:

computer scientist

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

neuroscientist, biologist, or chemist

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

social scientist (linguist?)

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:
statistician or biostatistician

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

CEO/small business owner

adapted from Brad Voytek

Why this course?

You are going to be analyzing lots of data because you’re studying to be a:

political activist

adapted from Brad Voytek

Why this course?

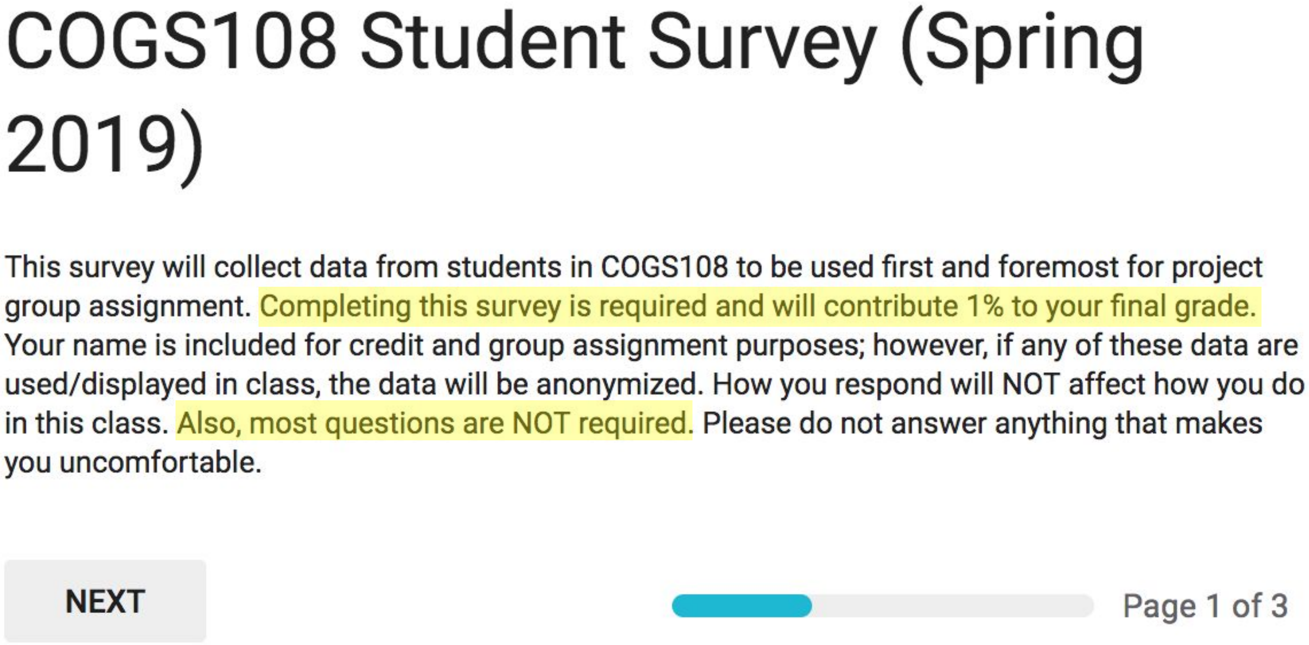
You are going to be analyzing lots of data because you're studying to be:

something else really awesome

adapted from Brad Voytek

Survey : http://bit.ly/cogs108_survey

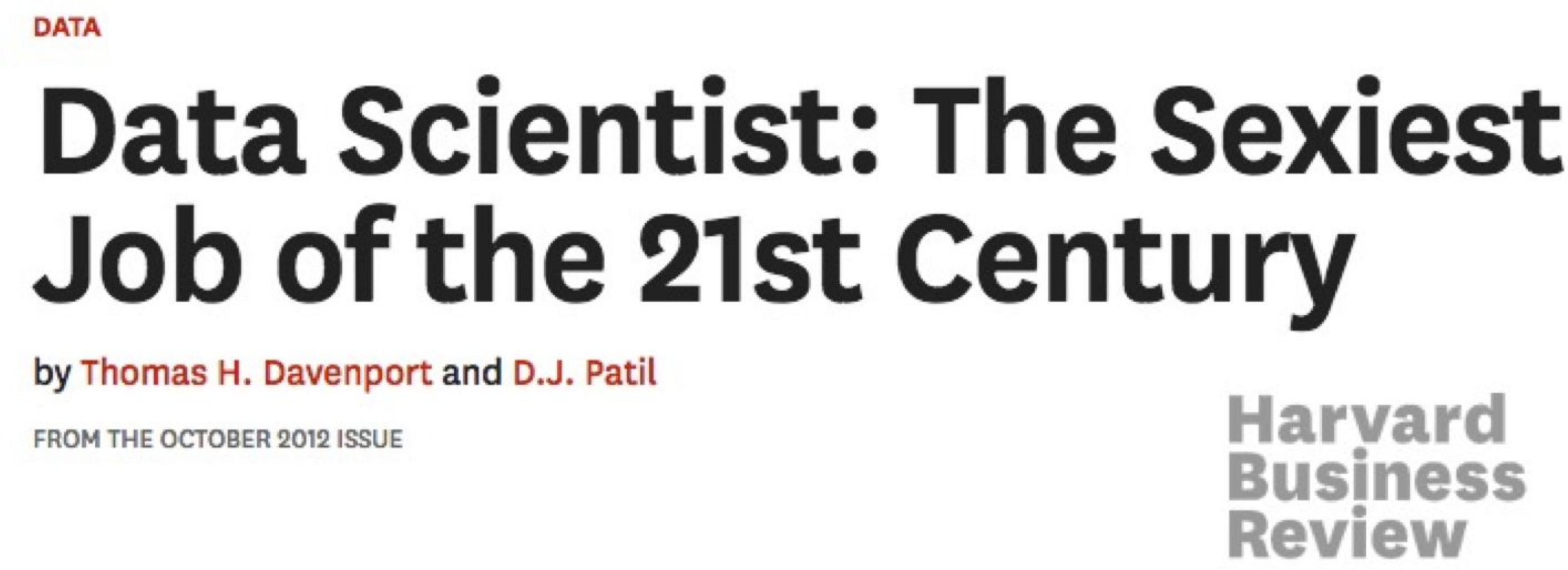
Due 11:59 PM **Friday**



COGS108 Student Survey (Spring 2019)

This survey will collect data from students in COGS108 to be used first and foremost for project group assignment. Completing this survey is required and will contribute 1% to your final grade. Your name is included for credit and group assignment purposes; however, if any of these data are used/displayed in class, the data will be anonymized. How you respond will NOT affect how you do in this class. Also, most questions are NOT required. Please do not answer anything that makes you uncomfortable.

Why this course?



DATA

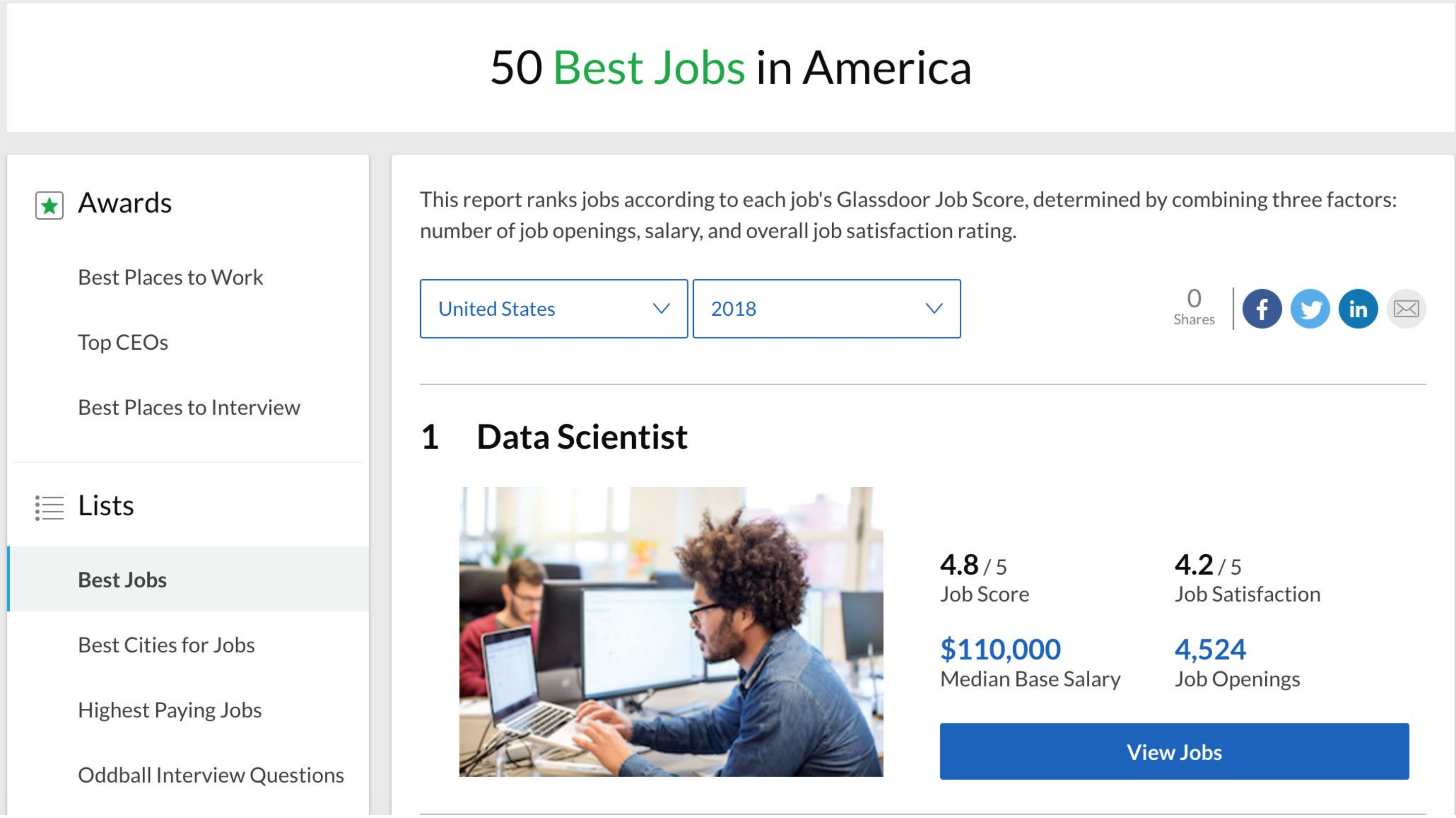
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

adapted from Brad Voytek



50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

1 Data Scientist

4.8 / 5
Job Score

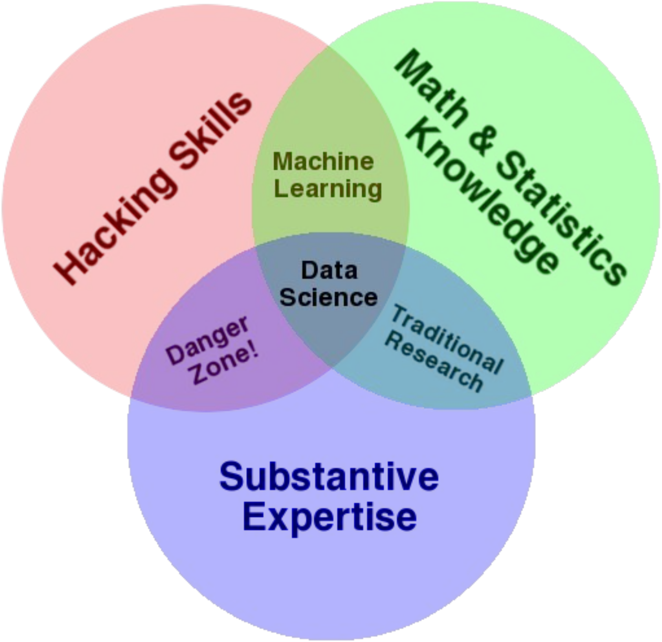
4.2 / 5
Job Satisfaction

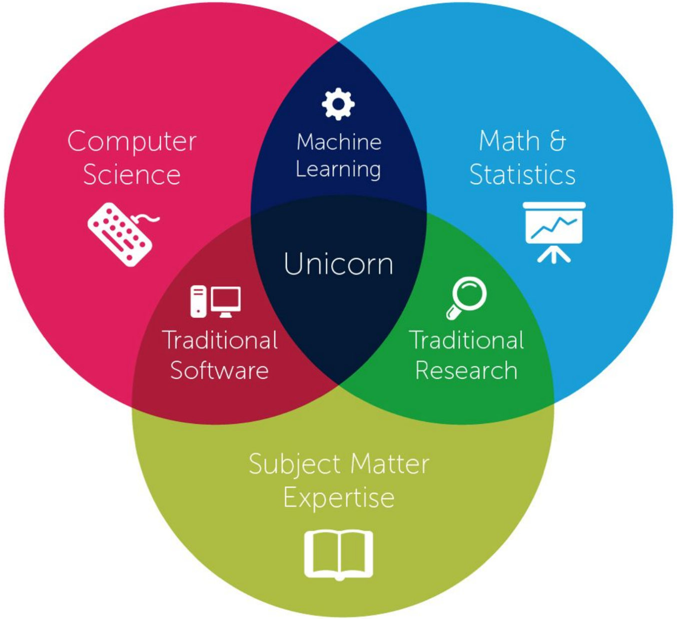
$110,000
Median Base Salary

4,524
Job Openings

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

What is data science?





http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Defining Data Science

a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science. -Wikipedia

"This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary actions." -David Donoho ("*50 years of Data Science*

"an emerging discipline that draws upon knowledge in statistical methodology and computer science to create impactful predictions and insights for a wide range of traditional scholarly fields" - from a panel Rafael Irizarry moderated, shared on SimplyStatistics ("*The role of academia in data science education*")

"an umbrella term used by organizations to describe the processes used to extract value from data" -Rafael Irizarry's personal definition in "*The role of academia in data science education"*

"The study of how the quantification of observable phenomena can lead to human understanding of the processes giving rise to those phenomena—or even the ability to predict future outcomes absent human understanding—and why certain phenomena require more or less data to lead to human understanding and/or prediction accuracy". -Brad Voytek's definition

"The scientific process of extracting value from data"

Data scientists ask interesting questions & answer them with data

How we’ll approach learning about *and doing* data science in COGS 108

The goal in COGS 108 is to *do* data science.

Course Objectives

- Program at a competent level in the Python programming language
- Formulate a plan for & complete a data science project from start (question) to finish (communication)
- Explain and carry out descriptive, exploratory, inferential, and predictive analyses in Python
- Communicate results concisely and effectively in reports and presentations
- Identify and explain how to approach an unfamiliar data science task

Programming Prerequisite

- CSE 7 - MATLAB
- CSE 8A or 11 - Java
- COGS 18 - Python

Bottom line: **we will assume programming knowledge.**

Python will be used for all assignments.

No programming experience?

- *Preferred option*
  - Take a programming course first.
  - COGS 18 : Introduction to Python.
- *Can't wait?*
  - Use online sites like *codecademy.com* or *LearnPython.org*

COGS 108: General Plan

Week	Topic(s)
1	Python, Data Science, & Ethics
2	Data Wrangling & Visualization
3	Introduction to Analysis
4	*Case Study: EDA*
5	Probability & Statistics
6	*Case Study: Inference*
7	Nonparametric Approaches & Text Analysis
8	Machine Learning
9	Geospatial Analysis & Dimensionality Reduction
10	Guest Lectures & Future of Data Science

Course links

GitHub	https://github.com/COGS108	lecture/section materials & final project submission
DataHub	https://datahub.ucsd.edu	assignment submission
Piazza	piazza.com/ucsd/spring2019/cogs108	questions, discussion, and regrade requests
TritonEd	https://tritoned.ucsd.edu	grades
Gradescope	https://www.gradescope.com/courses/46170 Entry Code: M2B7VJ	project proposal & final project submission
Course Podcast	https://podcast.ucsd.edu/	listening to me speak at a rate even faster than normal
Anonymous Feedback	Submit via Google Form	if I ever offend you, use an example you hate, or to provide general feedback

Lecture attendance is *generally* not required

Lecture attendance is *generally* not required

- Attendance *is* required for guest lectures - dates for these will be announced at least a week in advance
    - You must have your iclicker with you the day of the guest lectures to get credit for attendance (2% each)
-

Lecture attendance is *generally* not required

- Attendance *is* required for guest lectures - dates for these will be announced at least a week in advance
  - You must have your iclicker with you the day of the guest lectures to get credit for attendance (2% each)
- Extra credit opportunity: If you answer >=50% of the iclicker questions at more than 75% of the class meetings, you'll have 1% added to your final grade. (Guest lecture days do *not* count toward the 75%.)

Why iclickers?

- There are a whole lot of you!
    - Checks understanding
    - Provides me with feedback
- Aids in critical thinking & allows for application of concepts
- Give you all a break from listening to me (we humans need this!)
- Take attendance without wasting your time

You’ll need to register your iclicker on TritonEd.

If you previously registered an iclicker on TritonEd and are using the same clicker, you are already registered.

Discussion Section: Attendance

- Attend the one you registered for if possible
- Attendance is not technically mandatory
- No section during week one

Discussion Section: Course Tutorials and Workbooks

- **Tutorial 00** - start here to make sure you're ready to go in lecture and for week 2 in section
  - Python 3.6 Anaconda distribution
  - Jupyter Notebooks
  - GitHub

Each week there will be a workbook provided to you in section to work through. Often code in these workbooks will help you complete your assignments. Other times, they'll get you working with data in a way different than we're able to accomplish in the assignments.

Week	Topic(s)	Covered in Section
1	Python, Data Science, & Ethics	--
2	Data Wrangling & Visualization	Python Basics
3	Introduction to Analysis	Data Wrangling
4	*Case Study: EDA*	EDA & Data Visualization
5	Probability & Statistics	*Case Study: EDA*
6	*Case Study: Inference*	Inference
7	Nonparametric Approaches & Text Analysis	*Case Study: Inference*
8	Machine Learning	Text Analysis
9	Geospatial Analysis & Dimensionality Reduction	Machine Learning
10	Guest Lectures & Future of Data Science	Work on Projects

General grading:

	% of Total Grade
(5) Assignments	50
(1) Project Proposal	10
(1) Final Project	35
(2) Survey & Guest Lectures	5

(5) Assignments

Assignments are completed individually and graded programmatically.

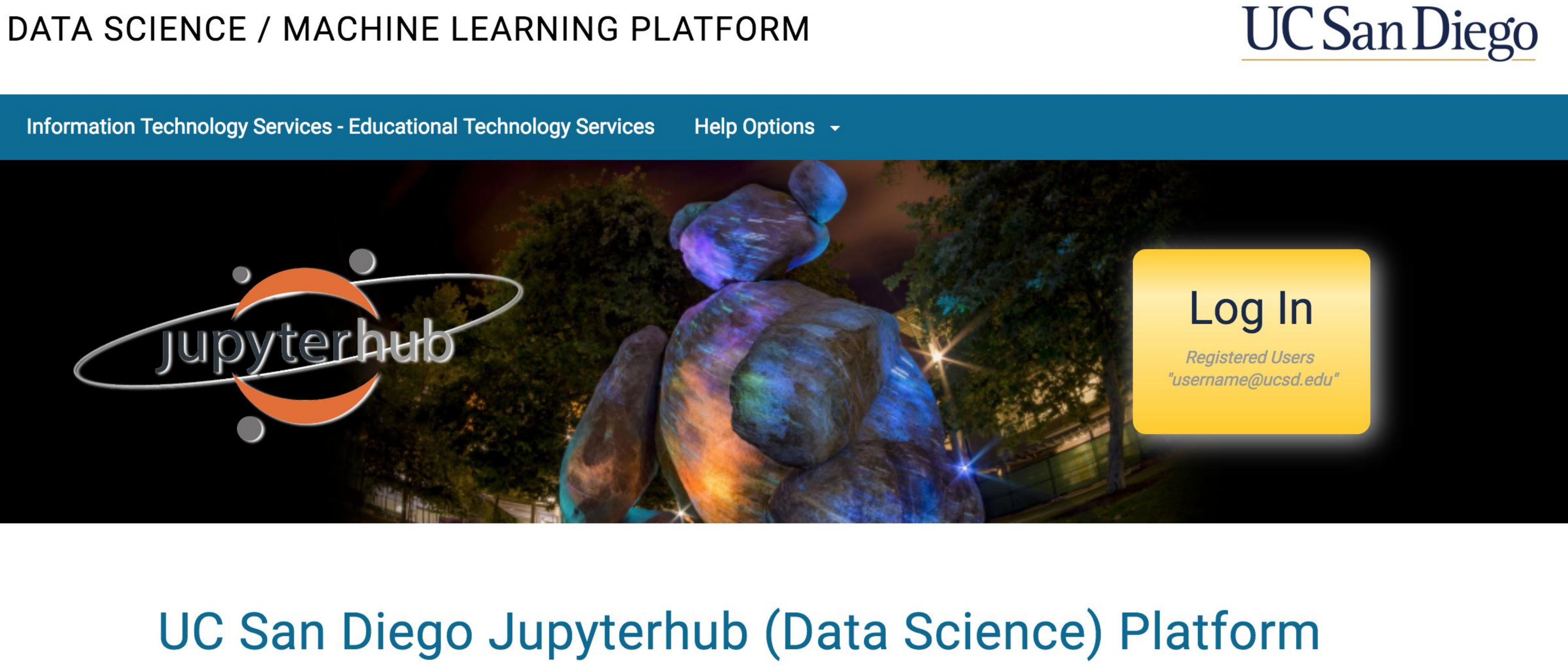
Instructions must be followed to receive credit.

These are meant to get you practice programming around the topics covered in class.

The first one is much simpler than the following four and will take less time.

Assignments will be due on Sundays by 11:59 PM

Assignment Submission @ Datahub: https://datahub.ucsd.edu



DATA SCIENCE / MACHINE LEARNING PLATFORM

UC San Diego

Information Technology Services - Educational Technology Services
Help Options
jupyterhub
Log In
Registered Users
"username@ucsd.edu"

UC San Diego Jupyterhub (Data Science) Platform

(1) Project Proposal (10%)

Project Proposals are Completed in *Assigned Groups* of 4-5.

These are **due Sunday (11:59 PM) after Week 3**.

These include a proposal of what question you are setting out to answer for your final project and what data you plan to use to complete this project..

(1) Final Project (35%)

The final project is a Jupyter notebook completed with your assigned group carrying out a *complicated data science project* from start to finish.

This should be completed throughout the quarter. Being a contributing member of a team is a requirement of this project. You should *start on this as early as possible and work on it a little each week.*

Week	Topic(s)	Due
1	Python, Data Science, & Ethics	Course Survey Fri 4/5 11:59 PM
2	Data Wrangling & Visualization	A1 Sun 4/14 11:59 PM
3	Introduction to Analysis	Project Proposal* Sun 4/21 11:59 PM
4	*Case Study: EDA*	A2 Sun 4/28 11:59 PM
5	Probability & Statistics	--
6	*Case Study: Inference*	A3 Sun 5/12 11:59 PM
7	Nonparametric Approaches & Text Analysis	--
8	Machine Learning	A4 Sun 5/26 11:59 PM
9	Geospatial Analysis & Dimensionality Reduction	--
10	Guest Lectures & Future of Data Science	A5 Sun 6/9 11:59 PM

Final Project*: due Wednesday of finals week by 11:59 PM

*indicates group submission. All other assignments/surveys are completed & submitted individually.

(1) Survey (1%) & (2) Guest Lectures (4%)

Survey must be filled out by this **Friday (11:59 PM)**. This will provide me with information about you all and help in group assignments. (1%)

There will be two **Guest Lectures** (2% each) through the quarter. Dates will be announced at least 1 week in advance in lecture and on Piazza. Attendance will be taken by iclicker on these days. If you forget your iclicker on the day of the lecture, you will not receive credit for attendance.

Course Confusion

- If something in lecture, a section workbook, or an assignment is unclear:
    - *ask in class*
    - *ask during section*
    - *post on Piazza*
    - *ask a classmate*
    - *come to office hours*

CLASS CONDUCT

In all interactions in this class, you are expected to be respectful. This includes following the UC San Diego principles of community.

This class will be a welcoming, inclusive, and harassment-free experience for everyone, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, ethnicity, religion (or lack thereof), political beliefs/leanings, or technology choices

At all times, you should be considered and respectful. Always refrain from demeaning, discriminatory, or harassing behavior and speech. Last of all, **take care of each other**.

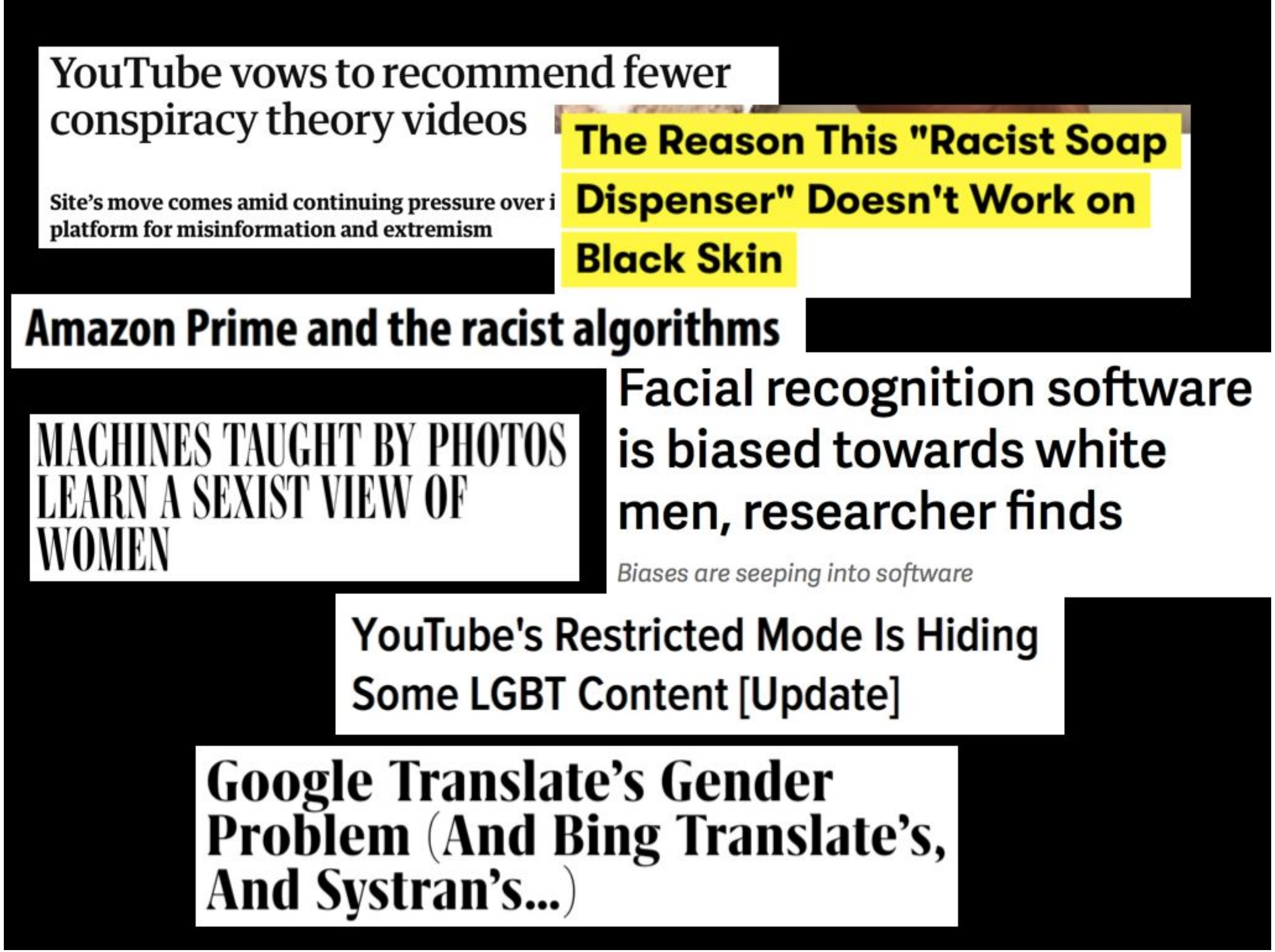
If you have a concern, please speak with Dr. Ellis, your TAs, or IAs. If you are uncomfortable doing so, the OPHD and/or CARE are wonderful resources on campus.

Are there any COGS 108 logistics questions ?

Data Science Ethics

“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



YouTube vows to recommend fewer conspiracy theory videos

Site's move comes amid continuing pressure over i 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...)

adapted from Thomas Donoghue

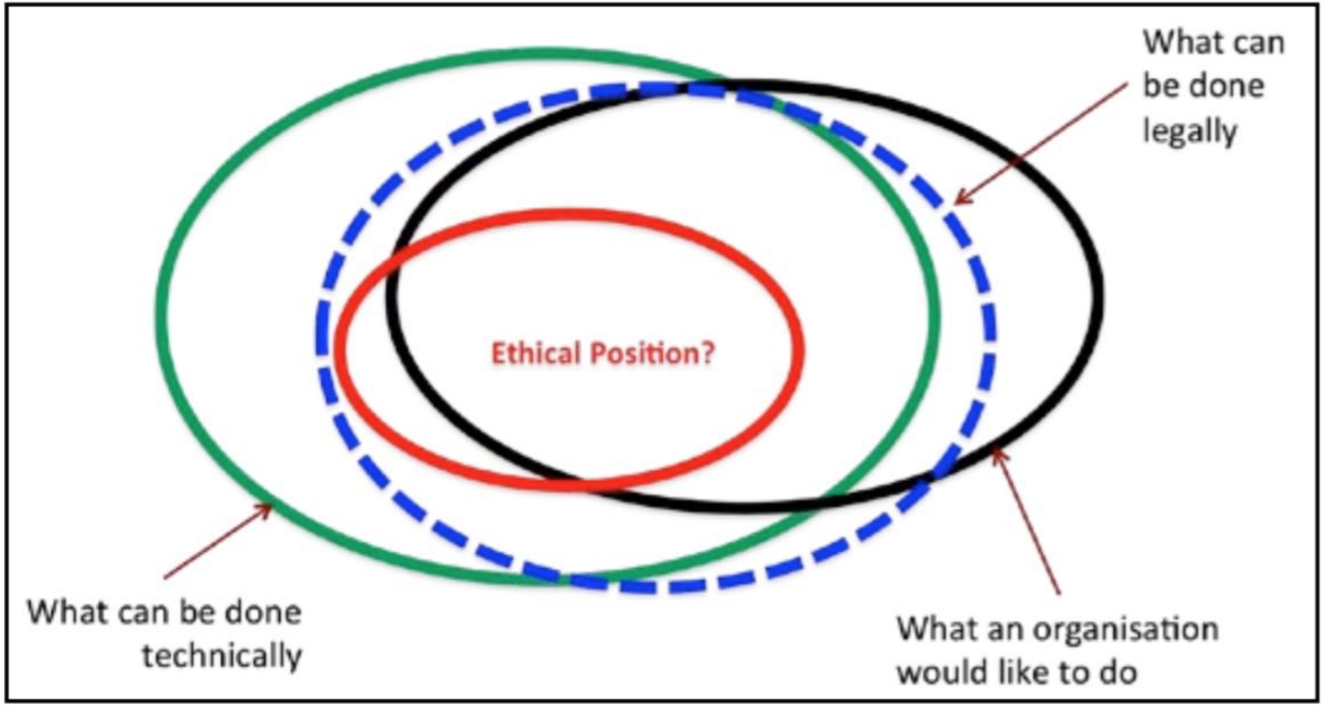
Always consider ethics.

ETHICS

“Moral principles that govern a person's behaviour 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 so that is equitable, with respect for privacy and consent, so as to ensure that it does not cause undue harm.

adapted from Thomas Donoghue

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

adapted from Thomas Donoghue

On the LAW

- There are laws that cover research, privacy and other aspects of data use, data science, & algorithms
- We’re not really talking about the law today
- Regulations are needed

adapted from Thomas Donoghue

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

adapted from Thomas Donoghue

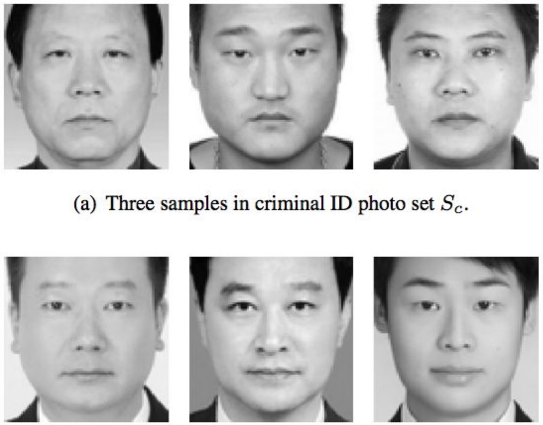
1. THE QUESTION

- 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: Labelling Faces

Detecting criminality from faces [link, paper]

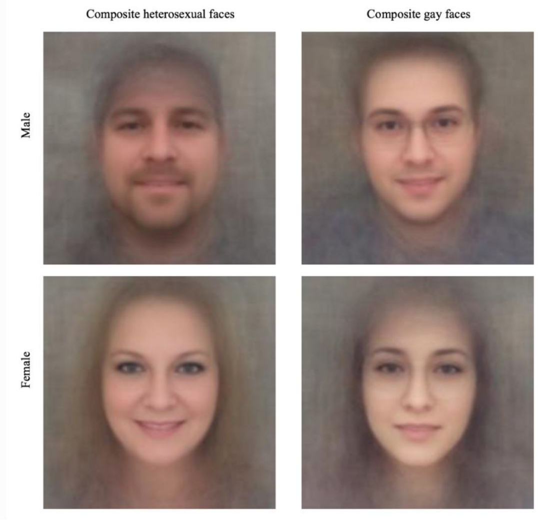


(a) Three samples in criminal ID photo set S_c.

(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]



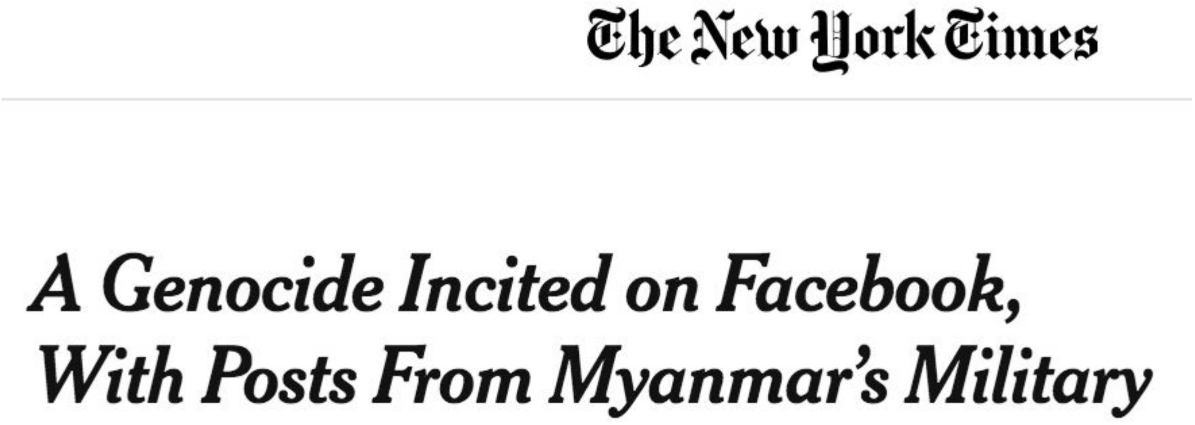
adapted from Thomas Donoghue

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 that females would be more variable

These findings do not generalize as well

Sources: link, link, link

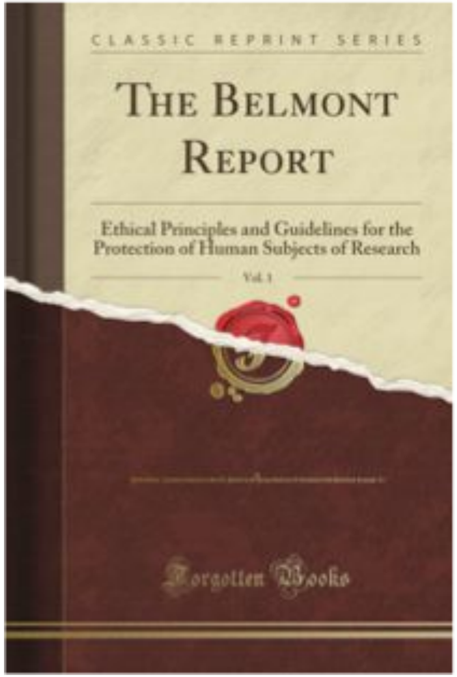
adapted from Thomas Donoghue

ASIDE: RESEARCH ETHICS

RESEARCH: A systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.

Investigations that qualify as research have their own regulatory structure that must be followed

The Belmont Report



CLASSIC REPRINT SERIES

THE BELMONT REPORT

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

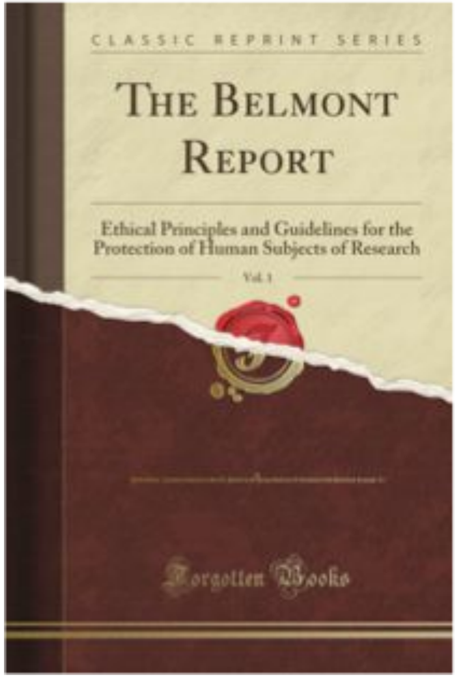
Vol. 1

Forgotten Books

The Belmont Report, published in 1974 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, is a statement of basic ethical principles and guidelines that should assist in resolving the ethical problems that surround the conduct of research with human subjects.

adapted from Brad Voytek

The Belmont Report


CLASSIC REPRINT SERIES
The Belmont Report
Ethical Principles and Guidelines for the Protection of Human Subjects of Research
Vol. 1
Forgotten Books

The Belmont Report presents three principles ; serve as basic concepts in the conduct of research with human subjects:

- Beneficence
- Respect for Persons
- Justice

The principles are intended to "*assist scientists, subjects, reviewers and interested citizens to understand the ethical issues inherent in research involving human subjects.*" It is generally accepted that the principles share equal weight, and the evaluation and conduct of human subjects research are done in the context of all three.

adapted from Brad Voytek

The Common Rule

The Federal Policy for the Protection of Human Subjects

- Requirements for assuring compliance by **research institutions**
- Requirements for researchers' obtaining and documenting **informed consent**
- Requirements for Institutional Review Board (IRB) membership, function, operations, review of research, and record keeping.
- Protections for certain vulnerable research subjects
    - pregnant women
    - *in vitro* fertilization and fetuses
    - prisoners
    - children

adapted from Brad Voytek

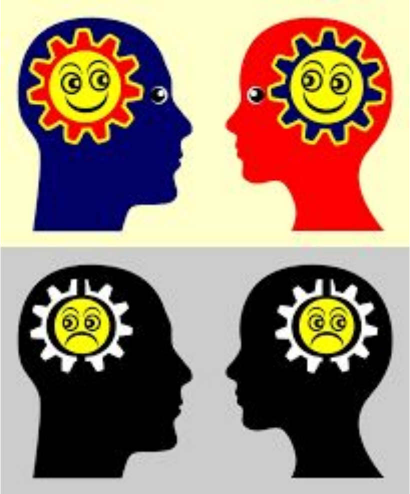
4. INFORMED CONSENT

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

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

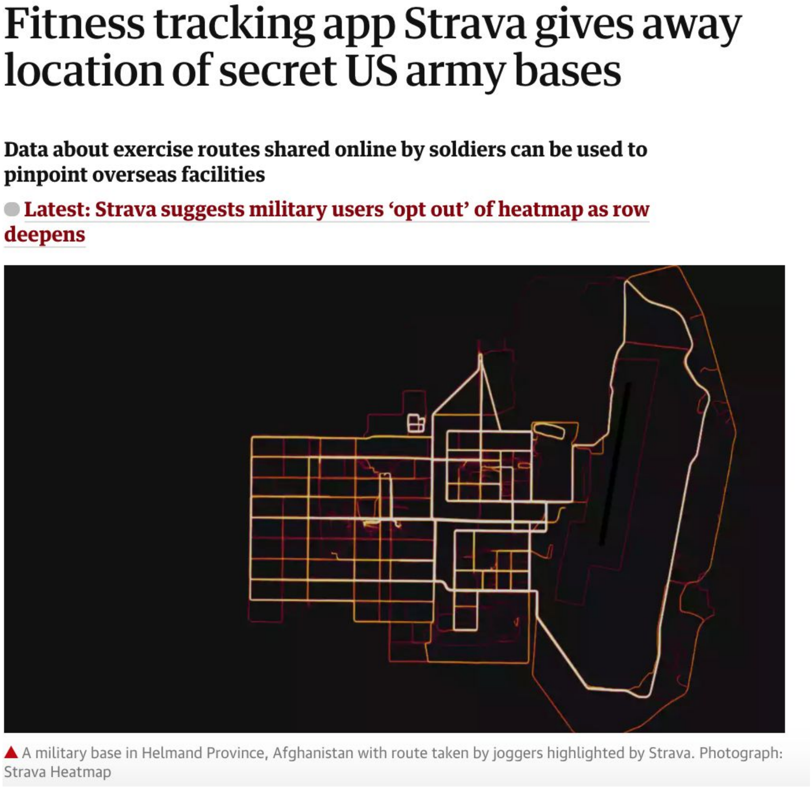
5. PRIVACY

- 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

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

6. EVALUATION

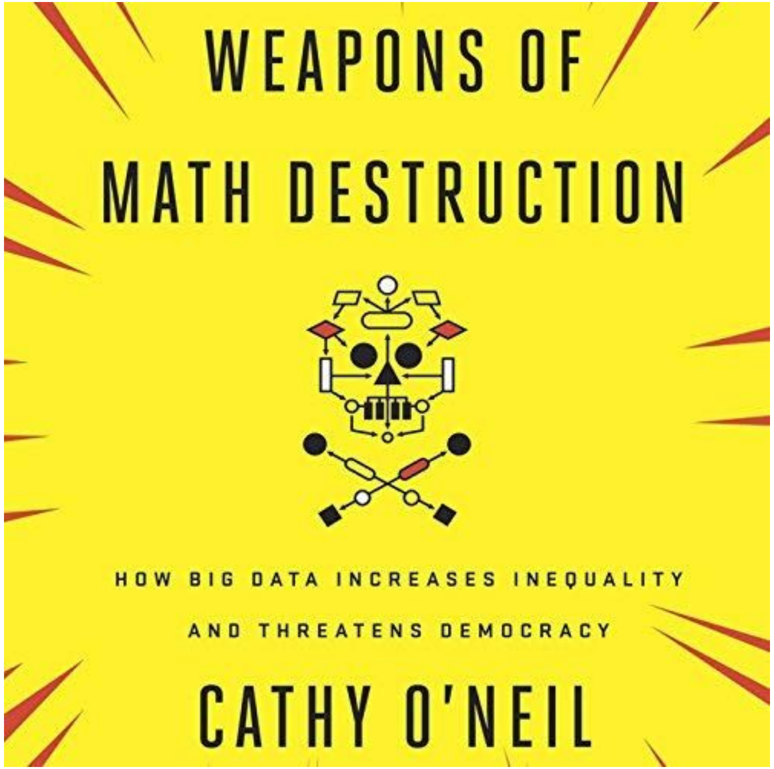
- How will you evaluate the project?
  a. Do you have a verifiable metric of success?
- Goodhart’s Law: when a measure becomes a target, it ceases to be a good measure.

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

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?

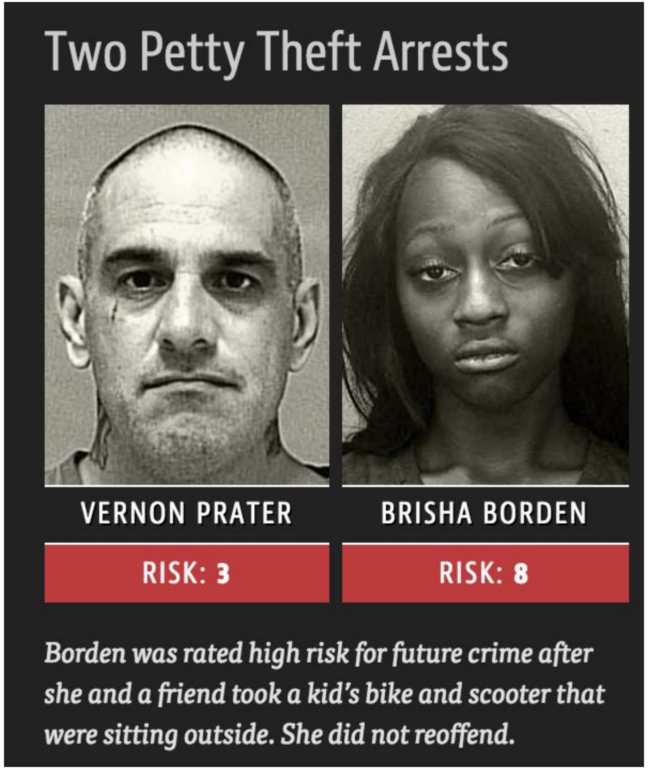
adapted from Thomas Donoghue

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

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?

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 behaviour 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 working on these systems may have conflicts of interest with respect to the incentive structures imposed by the system and/or the business

adapted from Thomas Donoghue

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

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
- if suggesting causality, have actually established causality
- 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

COGS9 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]

I’m excited to have you all in COGS 108!