

Welcome to COGS 108!

Data Science in Practice

Shannon E. Ellis, Ph.D
UC San Diego



Department of Cognitive Science
sellis@ucsd.edu



Why this course?

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

Why this course?

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

Why this course?

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

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
neuroscientist, biologist, or chemist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
social scientist (linguist?)

Why this course?

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

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
CEO/small business owner


Why this course?

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

Why this course?

You are going to be analyzing lots of data because you're studying to be:
something else really awesome

Survey : http://bit.ly/cogs108_wi20



Due 11:59
PM
Sunday

COGS 108 Student Survey (Winter 2020)

This survey is used to help me get to know you a bit better! Thanks in advance for your participation!

If you complete before Sunday (1/12) at 11:59 PM, there is an opportunity for a bit of extra credit.

If any of these data are used/displayed in class, the data will be anonymized. Please answer as truthfully as possible. **How you respond will NOT affect how you do in this class. Also, many 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

★ Awards

Best Places to Work

Top CEOs

Best Places to Interview

☰ Lists

Best Jobs

Best Cities for Jobs

Highest Paying Jobs

Oddball Interview Questions

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.

United States

2018

0
Shares



1 Data Scientist



4.8 / 5
Job Score

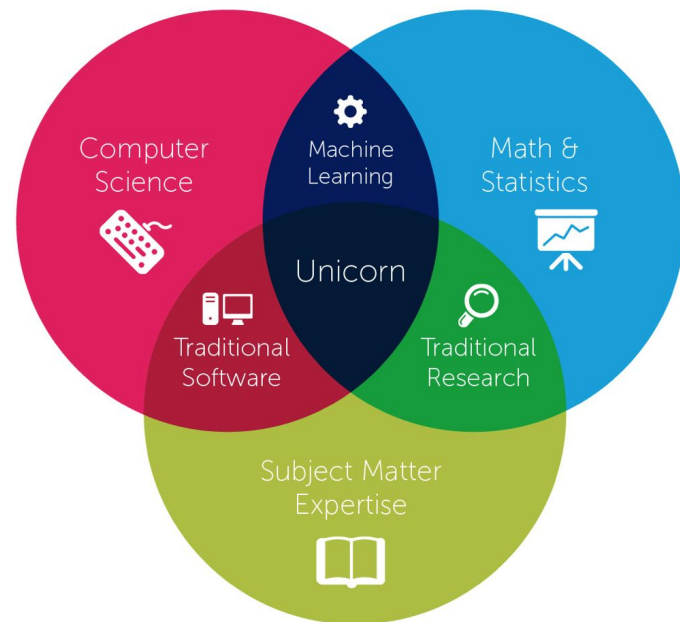
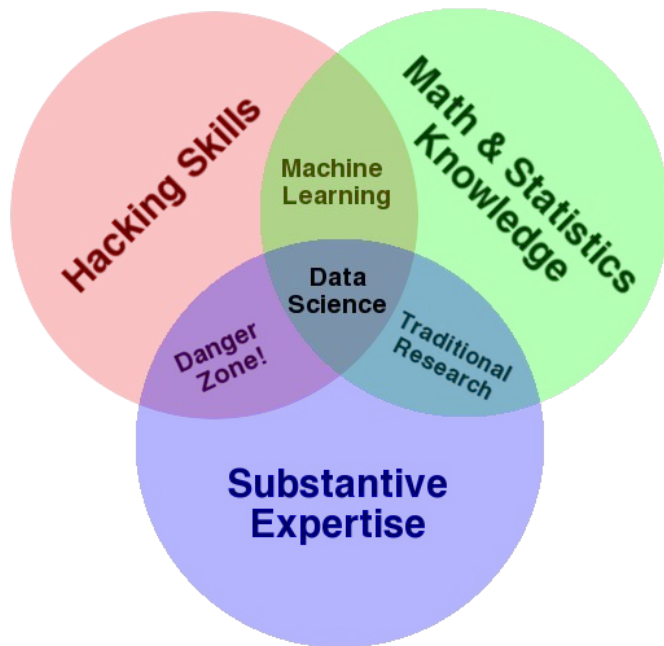
4.2 / 5
Job Satisfaction

\$110,000
Median Base Salary

4,524
Job Openings

[View Jobs](#)

What is data science?



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

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**

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

Course Objectives

- Formulate a plan for and 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

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

Scheduling

Lecture: Tu/Th 8-9:20

Discussion Sections: M, W, or F

Office Hours:

Date & Time	Location	Instructional Staff
W 2 - 4 PM	CSB 243	Professor Ellis
TBD	TBD	Matthew Feigelis
Th 11 AM - 12 PM	SSRB 100	Sam Lau
TBD	TBD	Shreenivas Venkataramanan

COGS 108: General Plan

Week	Topic(s)
1	Data Science, Ethics, & Version Control
2	Programming & Python
3	Data, Parameterization & Data Wrangling
4	Data Intuition & Data Visualization
5	Descriptive & Exploratory Data Analysis
6	Inference
7	Machine Learning
8	Text Analysis
9	Geospatial Analysis
10	Future of Data Science

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 (or you forget it all)?

- *Preferred option*
 - Take a programming course first
 - COGS 18 : Introduction to Python
- *Can't wait?*
 - Use online sites like [codecademy.com](https://www.codecademy.com) or [LearnPython.org](https://www.learnpython.org)
 - [Python Data Science Handbook](#)

Course links

GitHub	https://github.com/COGS108	lecture/section materials & final projects
DataHub	https://datahub.ucsd.edu	assignment submission
Piazza	http://piazza.com/ucsd/winter2020/cogs108	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/8920	grades
Course Podcast	https://podcast.ucsd.edu/?q=COGS108	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

General grading:

	% of Total Grade
Class Participation (9) Weekly Quizzes OR Lecture Attendance	10
(6) Assignments	45
(1) Project Proposal	10
(1) Project Check-In	2
(1) Project Survey	3
(1) Final Project	25
(2) Guest Lectures	5

Lecture attendance is *generally* not required

- Attendance *is* required for guest lectures - dates for these will be announced at least a week in advance
- Participation Credit:
 - If you answer $\geq 50\%$ of the iclicker questions at more than 75% of the class meetings, you'll get full credit for attendance.
 - Otherwise, the average of your weekly quizzes will be used (lowest weekly quiz is dropped)
- Lectures: most topics will have a conceptual and a technical lecture

Weekly Quizzes: If coming to class at 8 AM is not your thing...

- (9) weekly quizzes (first one due Friday of Week 2)
- Goal: to help you keep on top of the material covered in lecture
- Why?: experience + student feedback
- How:
 - Taken on Canvas
 - Single Attempt
 - ~10 Questions
 - Timed : 15 minutes
 - Lowest quiz dropped
 - Posted by Friday @ 5 PM (after each week of lecture); due the following Friday
 - Meant to test concepts from previous week's lecture

Note: If you complete weekly quizzes *and* come to class, your quiz average can replace your lowest assignment score

Why iclickers in COGS 108?

- 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 (starting Thursday)

You'll need to [register your iclicker on Canvas](#).

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

If you registered it on TritonEd but not Canvas, you must register it on Canvas

(6) Assignments

Assignments are completed individually and graded programmatically.

- These are meant to get you practice programming around the topics covered in class.
- The first two are much simpler than the following four and should take less time.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed to receive credit.
- You'll have the opportunity to practice in discussion section.

Assignments will be due on Fridays 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 ▾



Log In

Registered Users
"username@ucsd.edu"

UC San Diego Jupyterhub (Data Science) Platform

Before Thurs: log onto datahub & have a working [installation of Jupyter](#) on your computer

Week	Topic(s)	Due (Fri @ 11:59 PM)
1	Data Science, Ethics, & Version Control	--
2	Programming & Python	--
3	Data, Parameterization & Data Wrangling	A1 - git + python (1/24)
4	Data Intuition & Data Visualization	Project Proposal* (1/31)
5	Descriptive & Exploratory Data Analysis	A2 - pandas (2/7)
6	Inference	A3 - Data Exploration (2/14)
7	Machine Learning	A4 - Data Privacy Project Check-In* (2/21)
8	Text Analysis	A5 - Data Analysis (2/28)
9	Geospatial Analysis	A6 - NLP (3/6)
10	Future of Data Science	Group Work Survey (3/13)

Final Project*: due Thursday March 19th of finals week by 11:59 PM

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

Discussion Section

- Goal:
 - help with technical aspects of the course
 - assignment help
- No section during week one
- Attend the one you registered for, if possible
- Attendance is not *technically* mandatory
 - *But*, if you attend and participate in 8 weeks, you can replace your lowest assignment score with a full mark
 - Note: there are two weeks without Mon sections. You can attend Wed or Fri on those weeks.
 - If you're interested in the course material, enjoy learning about data science, and/or are struggling on the assignments, attend discussion section.

Week	Topic(s)	Discussion Section	Due (Fri @ 11:59 PM)
1	Data Science, Ethics, & Version Control		--
2	Programming & Python	intro + datahub + git (A1)	--
3	Data, Parameterization & Data Wrangling	project groups/proposals	A1 - git + python (1/24)
4	Data Intuition & Data Visualization	pandas I (A2)	Project Proposal* (1/31)
5	Descriptive & Exploratory Data Analysis	pandas II (A2)	A2 - pandas (2/7)
6	Inference	dataviz/seaborn (A3)	A3 - Data Exploration (2/14)
7	Machine Learning	web pages + python (A4)	A4 - Data Privacy Project Check-In* (2/21)
8	Text Analysis	inference (A5)	A5 - Data Analysis (2/28)
9	Geospatial Analysis	text data (A6)	A6 - NLP (3/6)
10	Future of Data Science	advice	Group Work Survey (3/13)

Final Project*: due Thursday March 19th of finals week by 11:59 PM

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

(1) Project Proposal (10%)

Project Proposals are completed in groups of 4-5.

These are **due Friday (11:59 PM) of Week 4.**

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) Project Check-In (2%)

Project Proposals are completed in groups of 4-5.

These are **due Friday (11:59 PM) of Week 7.**

These include a brief update about what your group has accomplished up to that point and what your plan is for successful completion at the end of the quarter

(1) Final Project (35%) + (1) Survey (3%)

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*. This is due the day of the scheduled final.

The survey is completed individually. It reports information about working with your group throughout the quarter.

Individuals who do not contribute equally are at risk of receiving a lower grade than other members of the group

Week	Topic(s)	Discussion Section	Due (Fri @ 11:59 PM)
1	Data Science, Ethics, & Version Control		--
2	Programming & Python	intro + datahub + git (A1)	--
3	Data, Parameterization & Data Wrangling	project groups/proposals	A1 - git + python (1/24)
4	Data Intuition & Data Visualization	pandas I (A2)	Project Proposal* (1/31)
5	Descriptive & Exploratory Data Analysis	pandas II (A2)	A2 - pandas (2/7)
6	Inference	dataviz/seaborn (A3)	A3 - Data Exploration (2/14)
7	Machine Learning	web pages + python (A4)	A4 - Data Privacy Project Check-In* (2/21)
8	Text Analysis	inference (A5)	A5 - Data Analysis (2/28)
9	Geospatial Analysis	text data (A6)	A6 - NLP (3/6)
10	Future of Data Science	advice	Group Work Survey (3/13)

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

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

(2) Guest Lectures (5%)

There will be (2) **Guest Lectures** (2.5% each) through the quarter.

Dates will be announced at least 1 week in advance in lecture and on Piazza.

Multiple Paths to Success (ok...a good grade) in COGS 108

	Quizzes	Lecture Participation	Assignments	Guest Lectures	Discussion Section	Final Project	Final Grade
<u>Hind</u> (enjoys lecture; quizzes are stressful; busy in the afternoons)	0%	100%	100%	100%	0%	100%	100%
<u>Tung</u> (super busy and/or not a morning person)	100%	0%	100%	100%	0%	100%	100%
<u>Sarah</u> (struggling on the assignments)	100%	100%	<< 100%	100%	100%	100%	100%

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 Prof. Ellis, your TAs, or IAs. If you are uncomfortable doing so, the OPHD and/or CARE are wonderful resources on campus.

Prof Pet Peeve: Please do not pack up before lecture is over.
I have never kept a class past class time.

The (dreaded) waitlist

1. I know this matters to you and is a source of stress (and I hate that).
2. I have no control over the waitlist
 - a. I know in other departments profs have control of this
 - b. I quite literally do not have access to the system
3. A few people in each section typically get off the waitlist, but that number varies each quarter.
 - a. Number of seats in lecture hall
 - b. Number of seats in discussion section room
 - c. Number of people who drop the class
4. The waitlist settles after week 2.
5. Our staff (cogsadvising@ucsd.edu) take care of this. They'll be able to help you via email or in COGS drop-in office hours more than I can.

**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]?**”

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]**

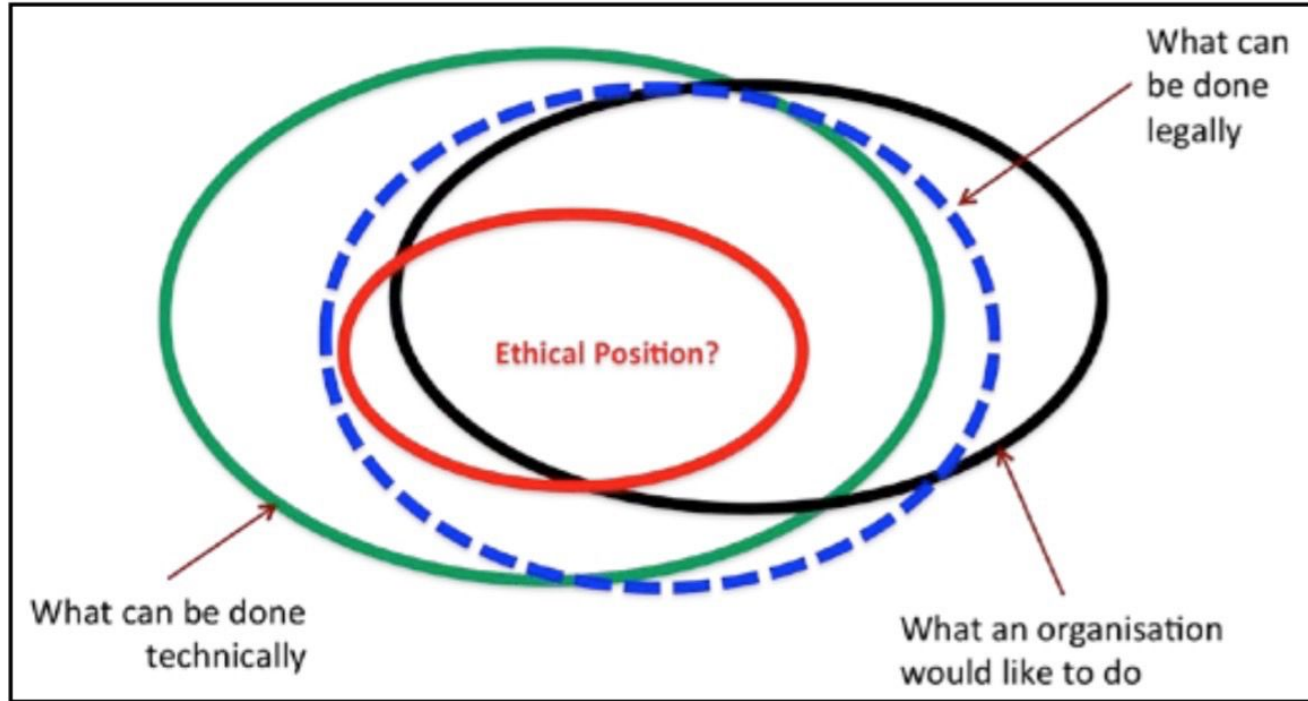
**Google Translate's Gender
Problem (And Bing Translate's,
And Systran's...)**

Always consider ethics.

ETHICS

“Moral principles that govern a person's behaviour 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

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

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 .

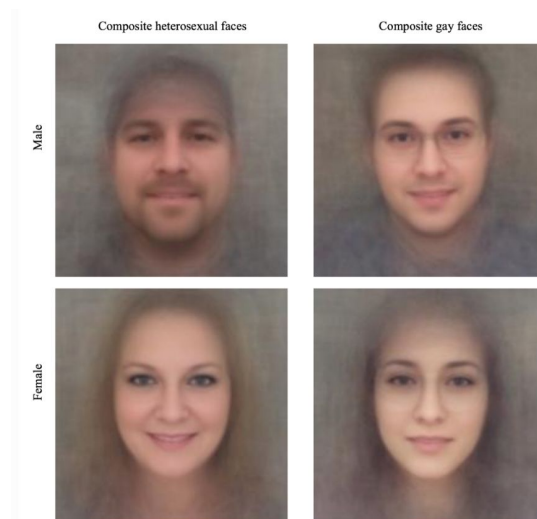


(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Detecting Sexual Orientation From

Faces with computer vision [[link](#), [paper](#)]



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](#)

ASIDE: RESEARCH ETHICS

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

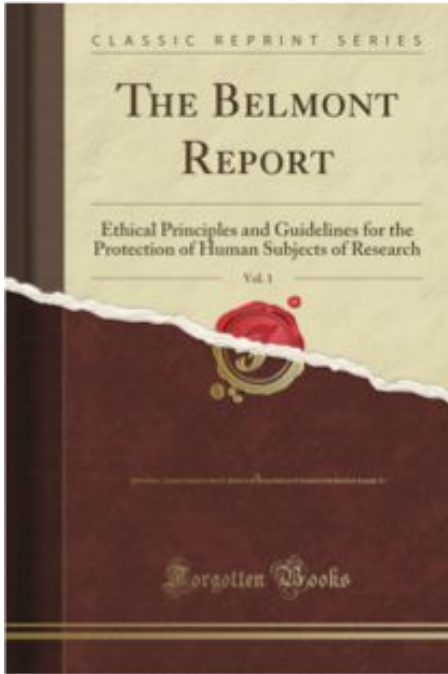
Investigations that qualify as research have their own regulatory structure that must be followed

The Belmont Report



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.

The Belmont Report



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.

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

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

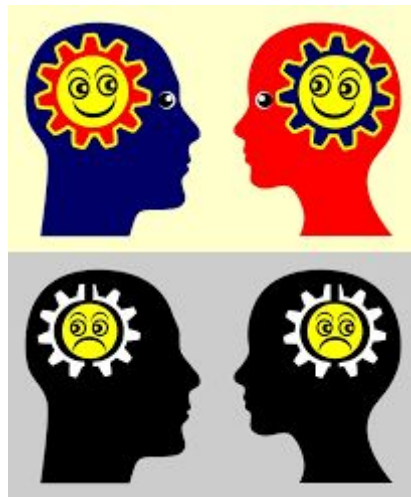
Informed consent can be withdrawn at any point in time



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](#)]



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

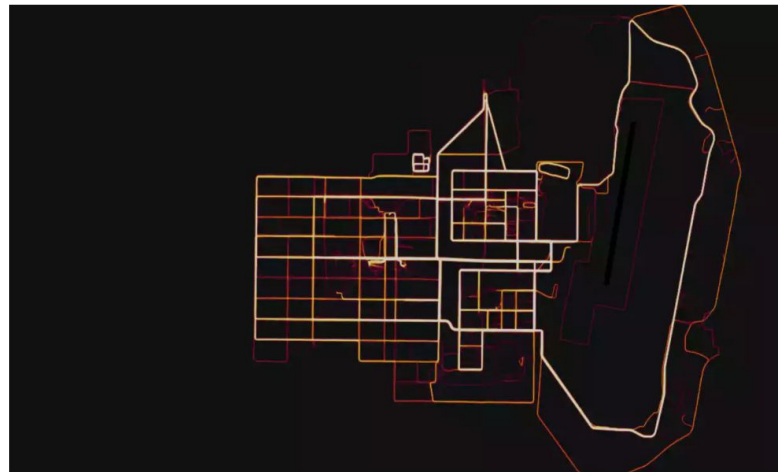
Case Study: Running Data

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

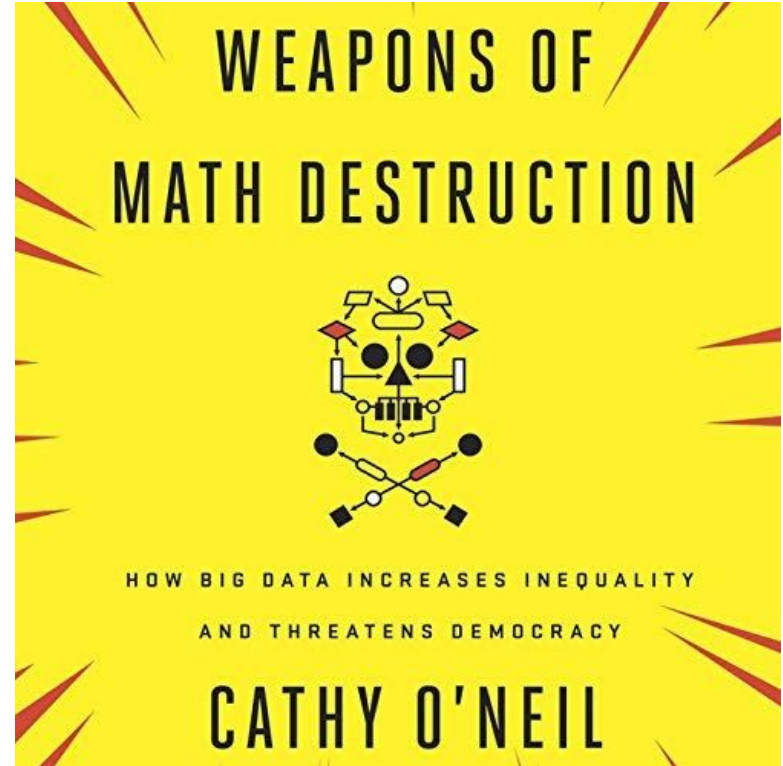
6. EVALUATION

- 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?

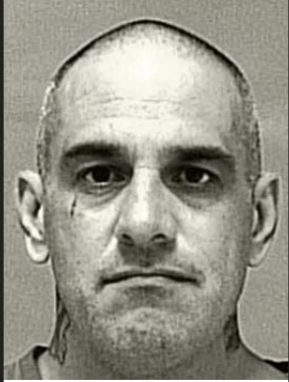

8. TRANSPARENCY & APPEAL

- 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](#)]

**I'm excited to have you
all in COGS 108!**

Reminder: I don't have any control over the waitlist.
Please contact cogsadvising@ucsd.edu or go to
their drop-in advising hours with questions.