

# Data Science

Ethics

# Goals for today

Get you thinking about the ethics of doing Data Science

1. This is not about *morals*, lots of definitions of these things, for us:
  - 1.1 *Morals*: how you, individually, determine what's right or wrong
  - 1.2 *Ethics*: moral principals that govern a person's behavior *when conducting an activity in a professional capacity*
2. I can't tell you what's right or wrong, that's up to your moral worldview
3. I can tell you what the Data Science community is discussing as part of their ethical framework

# Statistical Inference

We've concerned ourselves with building models

1. Models are used to go from world of 'data' back to something we can change about the world
2. If we don't intend on our models being actionable, why have them?
3. Things we may conclude from a model:
  - 3.1 An estimate
  - 3.2 rejection of a hypothesis
  - 3.3 Clustering/classification of data points into groups

# Sampling

We often can't know all the data out there, so we have to *sample*

1. Usually, we want a random sample
2. When might we not want a random sample?
3. When gather data, (e.g. sampling the population), we have to be careful
  - 3.1 How were questions worded?
  - 3.2 How did you poll people?
  - 3.3 How is missing data handled?

# Potential Sources of Bias

## 1. Sample Bias

- 1.1 Selection Bias: some subjects more likely to be selected
- 1.2 Volunteer Bias: people who volunteer are not representative
- 1.3 Nonresponse Bias: people who decline to be interviewed

## 2. Survey/Response Bias

- 2.1 Interviewer Bias
- 2.2 Acquiescence Bias: tendency to agree with all questions
- 2.3 Social Desirability Bias: Reluctance to admit to embarrassing things

## 3. Confirmation Bias

- 4. Anchor Bias: you might say yes to something because a worse alternative was shown first.

# Potential Sources of Bias

How do clinical trials work (often see as the gold standard)

1. Some receive treatment, others in control group
2. Each group is picked completely at random
3. Considerations
  - 3.1 Only ethical of alternatives have a good basis (i.e. we don't ask folks to *start* smoking for a trial)
  - 3.2 Very expensive
  - 3.3 Sometimes impossible!
4. What about how people 'opt in'?
5. Social networks run trials on us all the time, is that okay?

# MISUSE OF STATISTICS

This famous, but old book on statistics goes into detail about How to lie with statistics

**Number of children abused per 1,000 population in 1998 (National average is 12.9)\***

**States with the highest rates**

1. Alaska	37.1
2. Florida	23.2
3. Kentucky	23.1
4. Idaho	22.6
5. Connecticut	21.4

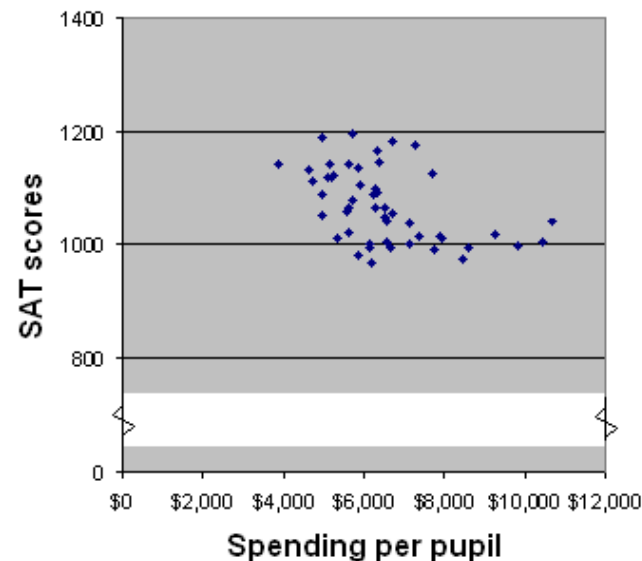
**States with the lowest rates**

45. Wisconsin	6.0
46. Virginia	5.9
47. New Jersey	4.9
48. New Hampshire	3.9
49. Pennsylvania	1.9

\*North Dakota not reporting

Source: U.S. Department of Health and Human Services, Children's Bureau

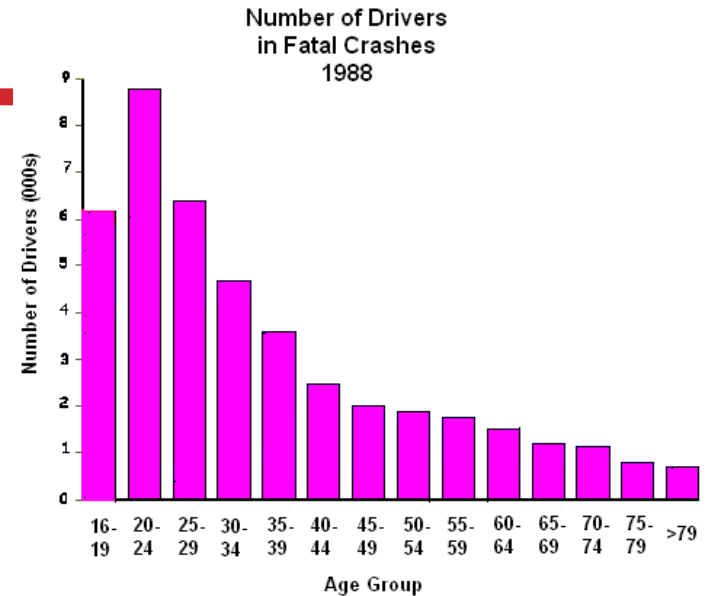
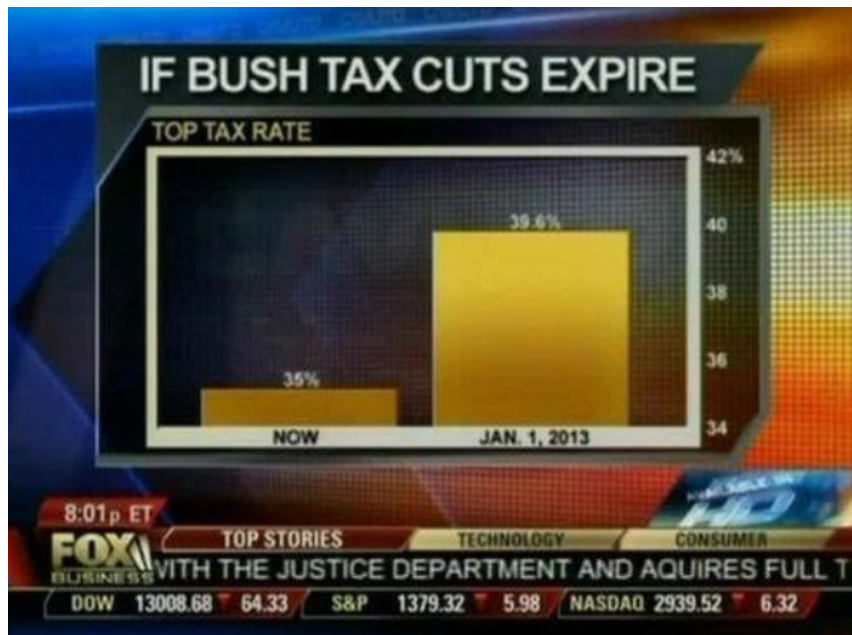
Spending per Pupil and SAT Scores by State, 1998



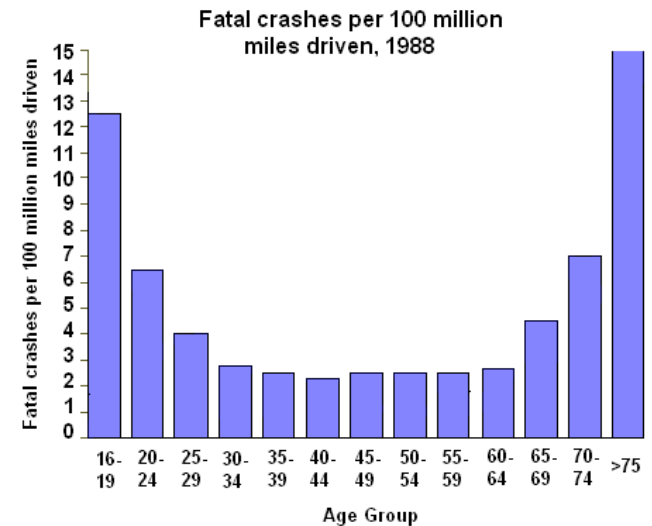
SAT Scores, 1998

State	Verbal	Math	Participation Rate
North Dakota	590	599	5%
New Jersey	497	508	79%

# BEWARE OF CHART



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.



# BEWARE OF CHARTS !

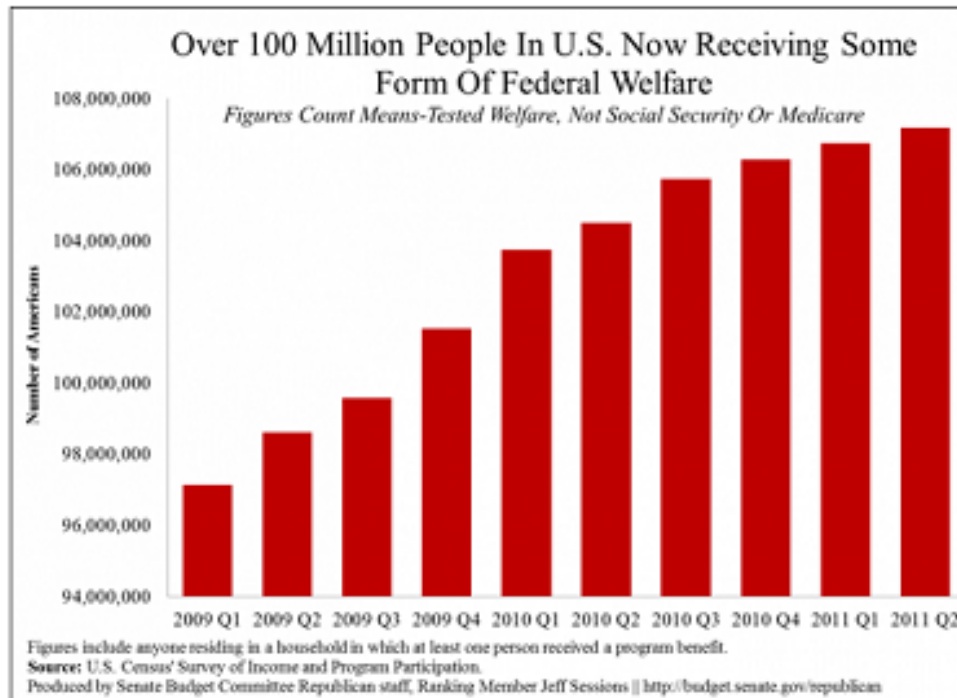
## THE BLOG

### Over 100 Million Now Receiving Federal Welfare

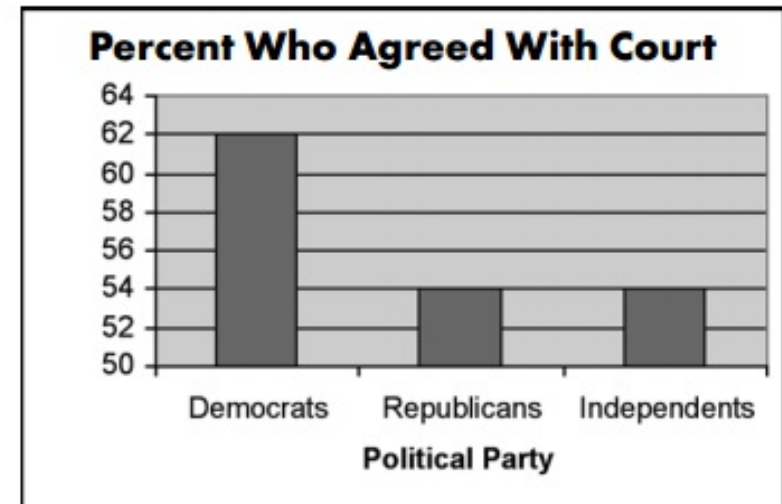
2:40 PM, AUG 8, 2012 • BY DANIEL HALPER

SHARE PAGE PRINT LARGER TEXT SMALLER TEXT PAGE 12

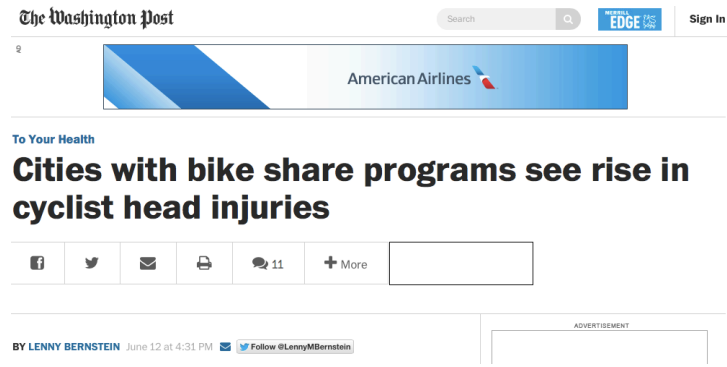
A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."



### Terry Schiavo Case



# NEWSPAPERS EVEN MORE



## Source

**A Washington Post article says: In the first study of its kind, researchers from Washington State University and elsewhere found a 14 percent greater risk of head injuries to cyclists associated with cities that have bike share programs. In fact, when they compared raw head injury data for cyclists in five cities before and after they added bike share programs, the researchers found a 7.8 percent increase in the number of head injuries to cyclists.**

**Actually: head injuries declined from 319 to 273, and overall injuries declined from 757 to 545**

- So the proportion of head injuries went up !!

# **CASE STUDY: FACEBOOK EMOTIONAL EXPERIMENT**

**Facebook routinely does A/B testing to test out new features (e.g., layouts, features, fonts, etc)**

**In 2014: intentionally manipulated news feeds of 700k users**

- Changed the number of positive and negative stories the users saw
- Measured how the users themselves posted after that

**Hypothesis: Emotions spread over the social media**

**Huge outcry**

**Facebook claims it gets the “consent” from the user agreement**

# OKCUPID EXPERIMENTS

## Experiment 1: Love is Blind

- Turned off photos for a day
- Activity went way down, but deeper conversations, better responses
- Deeper analysis at the link below

## Experiment 2:

- Turned off text or not – kept picture
- Strong support for the hypothesis that the words don't matter

## Experiment 3: Power of Suggestion

- Told people opposite of what the algorithm suggested

**<https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5>**

# GDPR AND CONSENT

**General Data Protection Regulation – new law in EU that recently went into play**

**Requires unambiguous consent**

- data subjects are provided with a clear explanation of the processing to which they are consenting
- the consent mechanism is genuinely of a voluntary and "opt-in" nature
- data subjects are permitted to withdraw their consent easily
- the organisation does not rely on silence or inactivity to collect consent (e.g., pre-ticked boxes do not constitute valid consent);

# DATA OWNERSHIP

## Consider your “biography”

- About you, but is it yours?
- No, the authors owns the copyright – not much you can do

## If someone takes your photo, they own it

- Limits on taking photos in private areas
- Can't use the photo in certain ways, e.g., as implied endorsement or implied libel

## Intellectual Property Basics:

- Copyright vs Patent vs Trade Secret
- Derivative works

# DATA OWNERSHIP

**Data Collection and Curation takes a lot of effort, and whoever does this usually owns the data “asset”**

**Crowdsourced data typically belongs to the facilitator**

- Rotten tomatoes, yelp, etc.

**What about personal data though?**

- e.g., videos of you walking around a store, etc?
- Written contracts in some cases, but not always

**New regulations likely to come up allowing customers to have more control over what happens with their data (e.g., GDPR)**

# PRIVACY

## First concern that comes to mind

- How to avoid the harms that can occur due to data being collected, linked, analyzed, and propagated?
- Reasonable rules ?
- Tradeoffs?

## No option to exit

- In the past, could get a fresh start by moving to a new place, waiting till the past fades
- big data is universal and never forgets
- Data science results in major asymmetries in knowledge



# WAYBACK MACHINES

**Archives pages on the web (<https://archive.org/web/> - 300 billion pages saved over time)**

- almost everything that is accessible
- should be retained forever

**If you have an unflattering page written about you, it will survive for ever in the archive (even if the original is removed)**

# RIGHT TO BE FORGOTTEN

Laws are often written to clear a person's record Law in EU and Argentina since 2006 after some years.

impacts search engines (not removed completely, but hard to find)

## Collection vs Use

- Privacy usually harmed upon use of data
- Sometimes collection without use may be okay
- Survenillance:
  - By the time you know what you need, it is too late to go back and get it

# WHY PRIVACY?

Data subjects have inherent right and expectation of privacy

“**Privacy**” is a complex concept

- **What** exactly does “privacy” mean? **When** does it apply?
- Could there exist societies without a concept of privacy?

Concretely: at collection “small print” outlines privacy rules

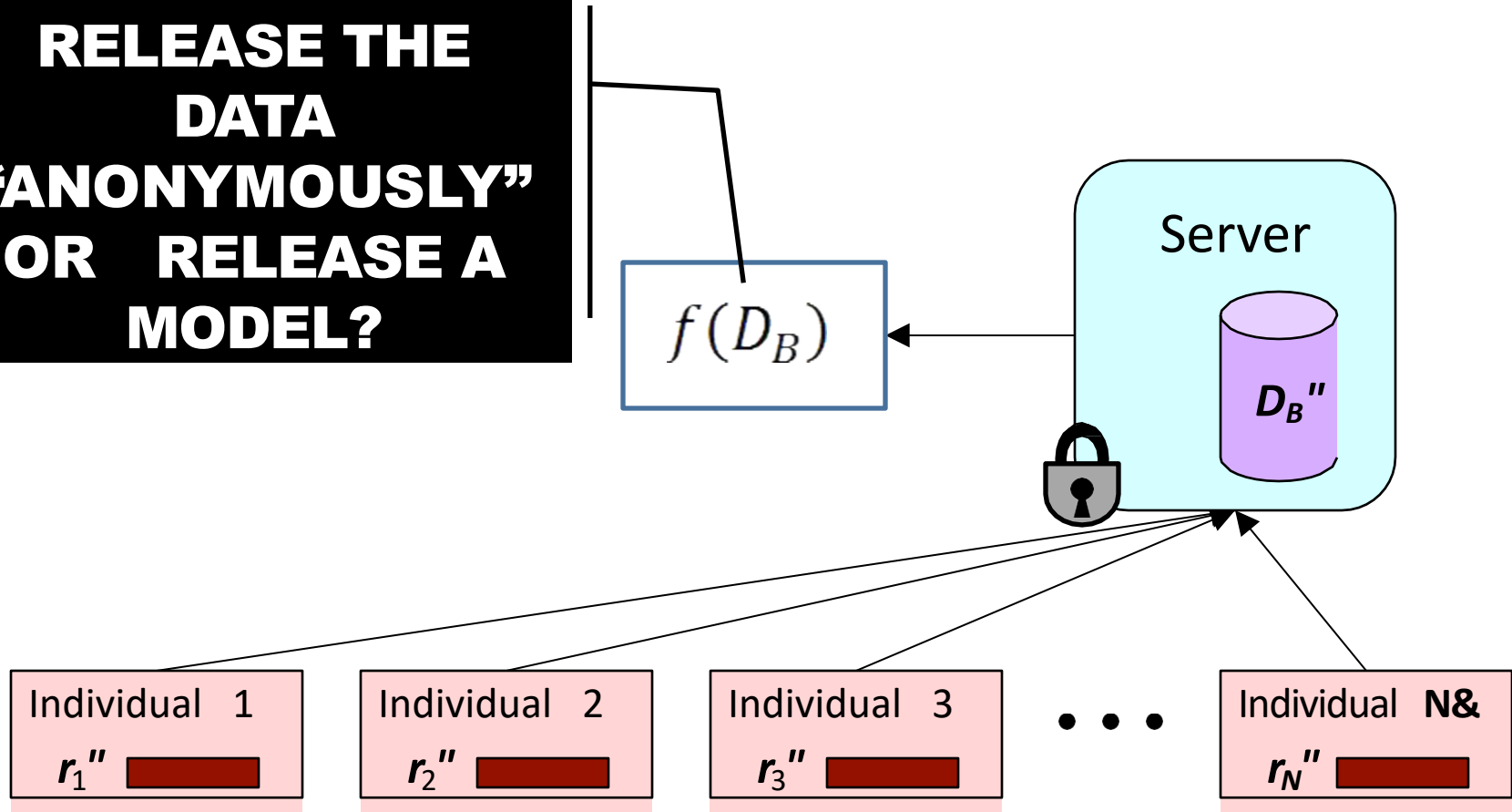
- Most companies have adopted a **privacy policy**
- E.g. AT&T privacy policy [att.com/gen/privacy-policy?pid=2506](http://att.com/gen/privacy-policy?pid=2506)

Significant legal framework relating to privacy

- UN Declaration of Human Rights, US Constitution
- HIPAA, Video Privacy Protection, Data Protection Acts



**RELEASE THE  
DATA  
“ANONYMOUSLY”  
OR RELEASE A  
MODEL?**



# WHY ANONYMIZE?

## For Data Sharing

- Give real(istic) data to others to study without compromising privacy of individuals in the data
- Allows third-parties to try new analysis and mining techniques not thought of by the data owner

## For Data Retention and Usage

- Various requirements prevent companies from retaining customer information indefinitely
- E.g. Google progressively anonymizes IP addresses in search logs
- Internal sharing across departments (e.g. billing → marketing)

# WHY ANONYMIZE?

## *2.1. Definitions in the EU Legal Context*

Directive 95/46/EC refers to anonymisation in Recital 26 to exclude anonymised data from the scope of data protection legislation:

*“Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;”.<sup>1</sup>*

# Releasing data is bad?



*What if we ensure our names and other  
identifiers are never released?*

# CASE STUDY: US CENSUS



**Raw data:** information about every US household

- Who, where; age, gender, racial, income and educational data

**Why released:** determine representation, planning

**How anonymized:** aggregated to geographic areas (Zip code)

- Broken down by various combinations of dimensions
- Released in full after 72 years

**Attacks:** no reports of successful deanonymization

- Recent attempts by FBI to access raw data rebuffed

**Consequences:** greater understanding of US population

- Affects representation, funding of civil projects
- Rich source of data for future historians and genealogists



# CASE STUDY: NETFLIX PRIZE

The Netflix logo, consisting of the word "NETFLIX" in white, bold, sans-serif capital letters with a black outline, set against a red rectangular background.

**Raw data:** 100M dated ratings from 480K users to 18K movies

**Why released:** improve predicting ratings of unlabeled examples

**How anonymized:** exact details not described by Netflix

- All direct customer information removed
- Only subset of full data; dates modified; some ratings deleted,
- Movie title and year published in full

**Attacks:** dataset is claimed vulnerable [Narayanan Shmatikov 08]

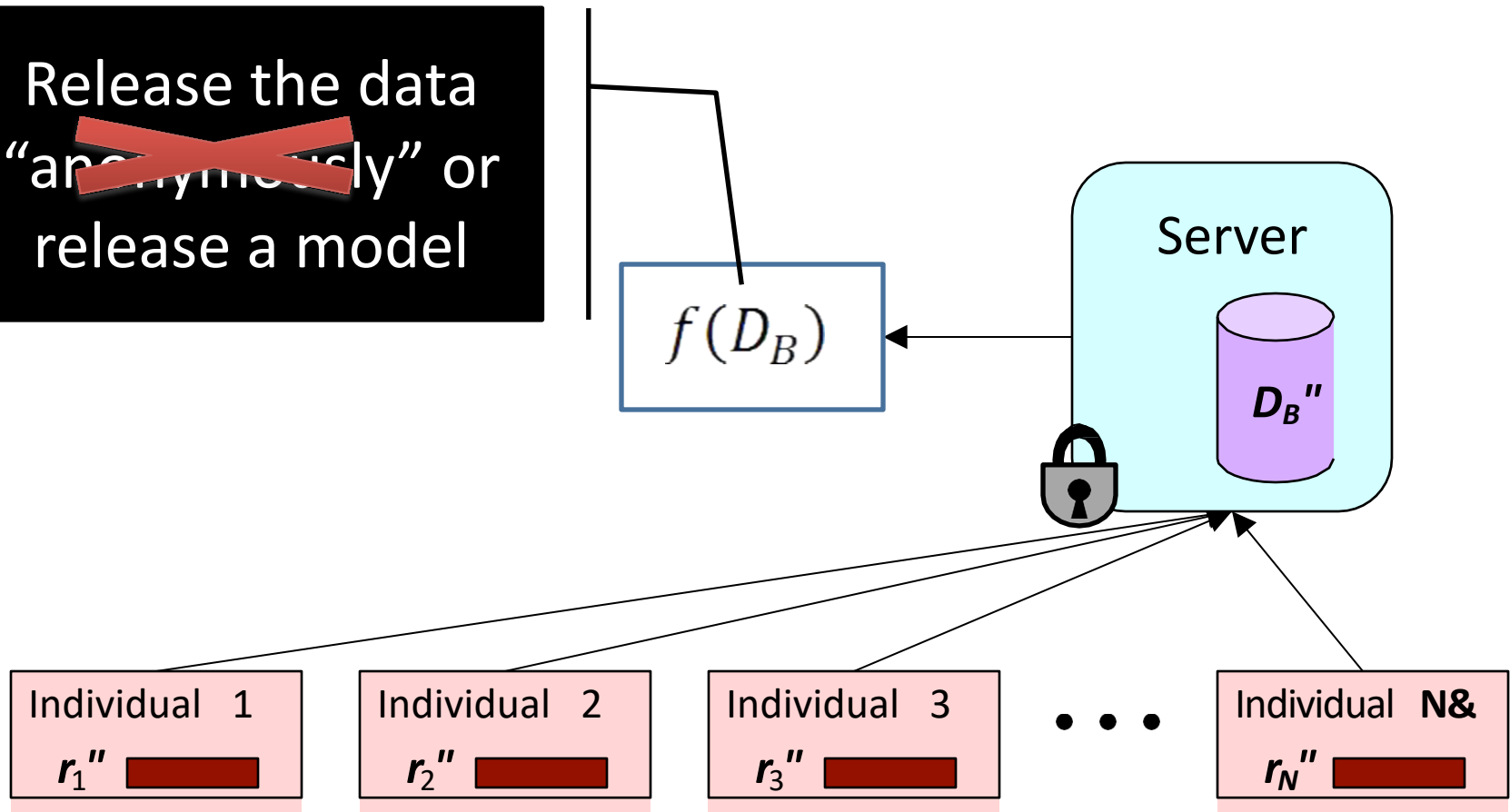
- Attack links data to IMDB where same users also rated movies
- Find matches based on similar ratings or dates in both

**Consequences:** rich source of user data for researchers

- unclear if attacks are a threat—no lawsuits or apologies yet

# CAN WE RELEASE A MODEL ALONE?

Release the data  
~~“anonymously”~~ or  
release a model



# RELEASING A MODEL CAN ALSO BE BAD

[Korolova JPC 2011]

Facebook profile



+

Online Data



- who live in the **United States**
- who live within 50 miles of **Staten Island, NY**
- between the ages of **23 and 27** inclusive
- who are **female**
- who are connected to **DogAnd PonyShow**
- in one of the categories: **Pop Culture, Science Fiction/Fantasy, Alternative, Rock, Classic Rock or iPhone**



+ Who are  
interested in  
**Men**

Number of  
Impressions

25

+ Who are  
interested in  
**Women**

0

Facebook's learning algorithm uses private information to predict match to ad

# Model Inversion

[Frederickson et al., USENIX Security 2014]

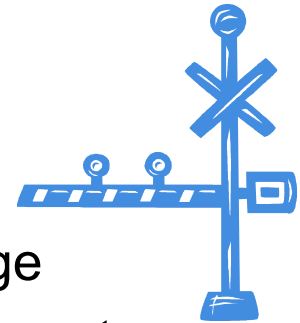
- An attacker, given the model and some demographic information about a patient, can predict the patient's genetic markers.

We show, however, that warfarin models do pose a privacy risk (Section 3). To do so, we provide a general model inversion algorithm that is optimal in the sense that it minimizes the attacker's *expected misprediction rate* given the available information. We find that when one knows a target patient's background and stable dosage, their genetic markers are predicted with significantly better accuracy (up to 22% better) than guessing based on marginal distributions. In fact, *it does almost as well as regression models specifically trained to predict these markers (only ~5% worse)*, suggesting that model inversion can be nearly as effective as learning in an “ideal” setting. Lastly, the inverted model performs measurably better for members of the training cohort than others (yielding an increased 4% accuracy) indicating a leak of information specifically about those patients.

# MODELS OF ANONYMIZATION

## Interactive Model (akin to statistical databases)

- Data owner acts as “gatekeeper” to data
- Researchers pose queries in some agreed language
- Gatekeeper gives an (anonymized) answer, or refuses to answer



## “Send me your code” model

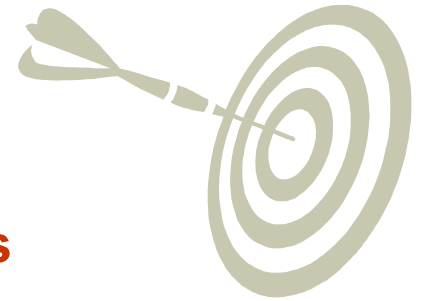
- Data owner executes code on their system and reports result
- Cannot be sure that the code is not malicious

## Offline, aka “publish and be damned” model

- Data owner somehow anonymizes data set
- Publishes the results to the world, and retires
- Our focus in this tutorial – seems to model most real releases



# OBJECTIVES FOR ANONYMIZATION



## Prevent (high confidence) inference of **associations**

- Prevent inference of salary for an individual in “census”
- Prevent inference of individual’s viewing history in “video”
- Prevent inference of individual’s search history in “search”
- All aim to prevent **linking** sensitive information to an individual

## Prevent inference of **presence** of an individual in the data set

- Satisfying “presence” also satisfies “association” (not vice-versa)
- Presence in a data set can violate privacy (eg STD clinic patients)

## Have to model what knowledge might be known to attacker

- **Background knowledge**: facts about the data set (X has salary Y)
- **Domain knowledge**: broad properties of data (illness Z rare in men)

# UTILITY

**Anonymization is meaningless if **utility** of data not considered**

- The empty data set has perfect privacy, but no utility
- The original data has full utility, but no privacy

**What is “**utility**”? Depends what the application is...**

- For fixed query set, can look at max, average distortion
- Problem for publishing: want to support unknown applications!
- Need some way to **quantify** utility of alternate anonymizations

# PRIVACY IS NOT ANONYMITY

- Bob's record is indistinguishable from records of other Cancer patients
  - We can infer Bob has Cancer !
- “New Information” principle
  - Privacy is breached if releasing  $D$  (or  $f(D)$ ) allows an adversary to learn sufficient new information.
  - *New Information = distance(adversary's prior belief, adversary's posterior belief after seeing  $D$ )*
  - *New Information* can't be 0 if the output  $D$  or  $f(D)$  should be useful.



# PRIVACY DEFINITIONS

- Many privacy definitions
  - L-diversity, T-closeness, M-invariance,  $\epsilon$ - **Differential privacy**, E- Privacy, ...
- Definitions differs in
  - What information is considered sensitive
    - Specific attribute (disease) vs all possible properties of an individual
  - What is the adversary's prior
    - All values are equally likely vs Adversary knows everything about all but one individuals
  - How is new information measured
    - Information theoretic measures
    - Pointwise absolute distance
    - Pointwise relative distance

# NO FREE LUNCH

- Why can't we have a single definition for privacy?
  - For every adversarial prior and every property about an individual, new information is bounded by some constant.
- No Free Lunch Theorem: For every algorithm that outputs a  $D$  with even a sliver of utility, there is some adversary with a prior such that privacy is not guaranteed.

# RANDOMIZED RESPONSE MODEL

- $N$  respondents asked a sensitive “yes/no” question.
- Surveyor wants to compute fraction  $\pi$  who answer “yes”.
- Respondents don't trust the surveyor.
- What should the respondents do?

# RANDOMIZED RESPONSE MODEL

- Flip a coin
  - heads with probability  $p$ , and
  - tails with probability  $1-p$  ( $p > \frac{1}{2}$ )
- Answer question according to the following table:

	True Answer = Yes	True Answer = No
Heads	Yes	No
Tails	No	Yes

# DIFFERENTIAL PRIVACY

- Typically achieved by adding controlled noise (e.g., Laplace Mechanism)
- Some adoption in the wild:
  - US Census Bureau
  - Google, Apple, and some others have used this for collecting data
- Issues:
  - Effectiveness in general still unclear

# THE DREAM

**You run your ML algorithm(s) and it works well (?!)**

**Still: be skeptical ...**

**Very easy to accidentally let your ML algorithm cheat:**

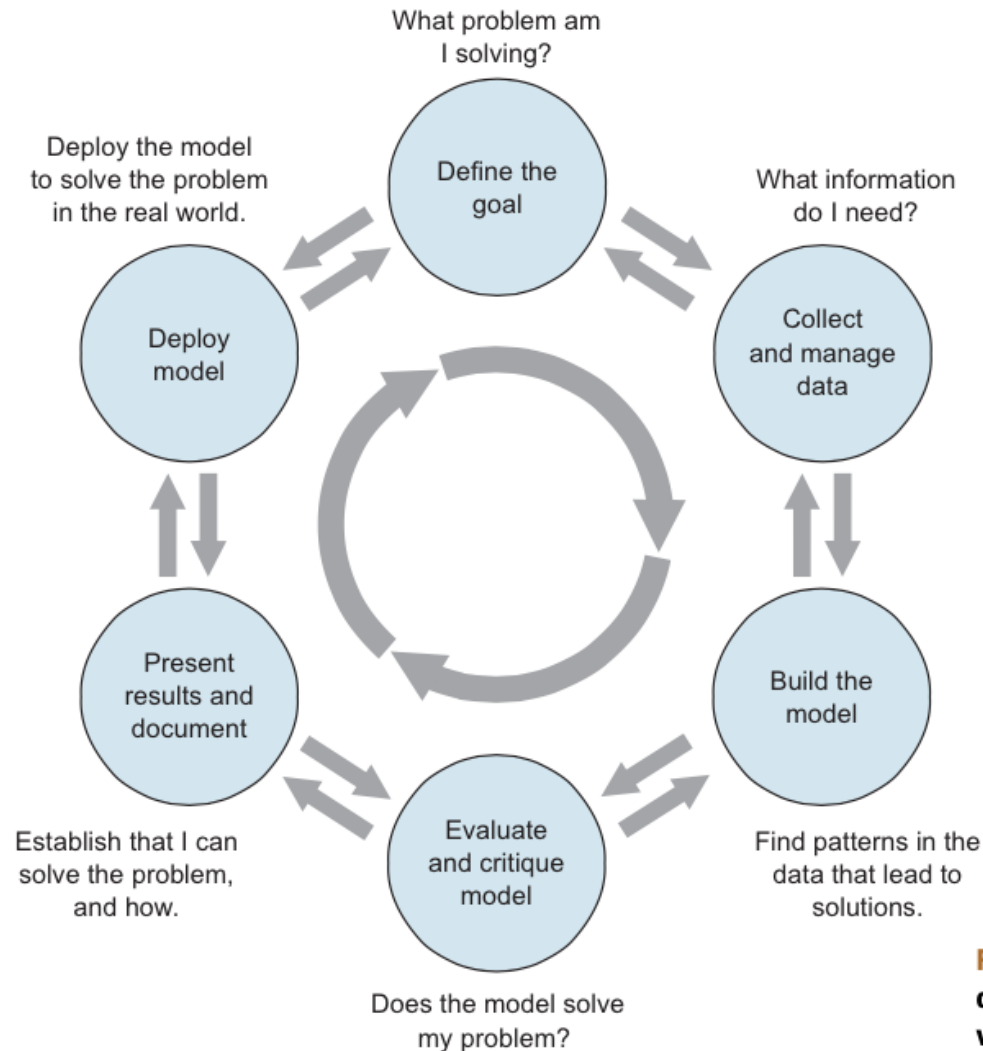
- Peaking (train/test bleedover)
- Including output as an input feature explicitly
- Including output as an input feature implicitly

**Try to solve the problem by hand;**

**Try to interpret the ML algorithm / output**

**Continue being skeptical. Always be skeptical.**

# DATA SCIENCE LIFECYCLE: AN ALTERNATE VIEW



**Figure 1.1** The lifecycle of a data science project: loops within loops

# COMBATING BIAS

## Fairness through blindness:

- Don't let an algorithm look at **protected attributes**

## Examples currently in use ???????????

- Race
- Gender
- Sexuality
- Disability
- Religion

## Problems with this approach ???????????



# COMBATING BIAS

*“After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. ‘This program had absolutely nothing to do with race... but multi-variable equations,’ argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound.”*

# COMBATING BIAS

**If there is bias in the training data, the algorithm/ML technique will pick it up**

- Especially social biases against minorities
- Even if the the protected attributes are not used

**Sample sizes tend to vary drastically across groups**

- Models for the groups with less representation are less accurate
- Hard to correct this, and so fundamentally unfair
- e.g., a classifier that performs no better than coin toss on a minority group, but does very well on a majority group

# COMBATING BIAS

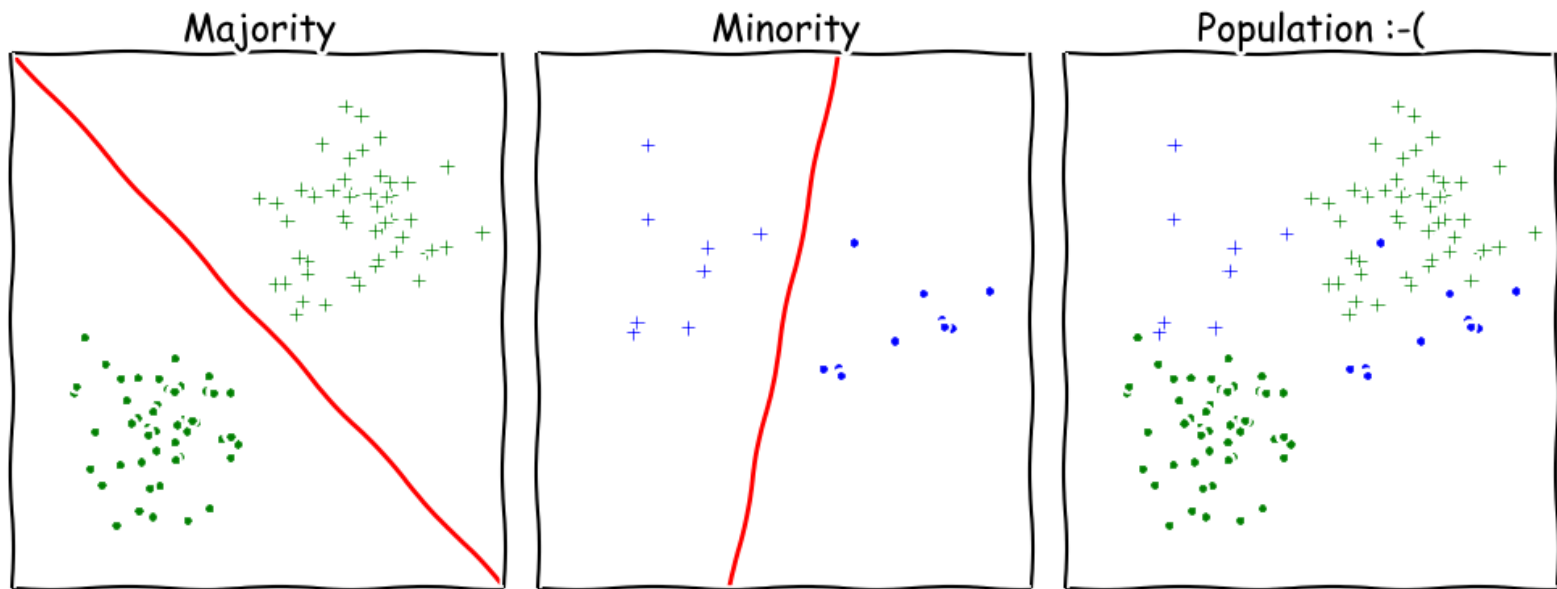
## Cultural Differences

- Consider a social network that tried to classify user names into real and fake
- Diversity in names differs a lot – in some cases, short common names are 'real', in others long unique names are 'real'

# COMBATING BIAS

## Undesired complexity

- Learning combinations of linear classifiers much harder than learning linear classifiers



# FATML

**This stuff is really tricky (and really important).**

- It's also not solved, even remotely, yet!

**New community:** **F**airness, **A**ccountability, and **T**ransparency in **M**achine **L**earning (aka **FATML**)

“... policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.”



**Fairness, Accountability,  
and Transparency  
in Machine Learning**

# **F IS FOR FAIRNESS**

**In large data sets, there is always proportionally less data available about minorities.**

**Statistical patterns that hold for the majority may be invalid for a given minority group.**

**Fairness can be viewed as a measure of diversity in the combinatorial space of sensitive attributes, as opposed to the geometric space of features.**

# **A IS FOR ACCOUNTABILITY**

**Accountability of a mechanism implies an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.**

- Current accountability tools were developed to oversee human decision makers
- They often fail when applied to algorithms and mechanisms instead

**Example, no established methods exist to judge the intent of a piece of software. Because automated decision systems can return potentially incorrect, unjustified or unfair results, additional approaches are needed to make such systems accountable and governable.**

# T IS FOR TRANSPARENCY

**Automated ML-based algorithms make many important decisions in life.**

- Decision-making process is opaque, hard to audit

**A transparent mechanism should be:**

- understandable;
- more meaningful;
- more accessible; and
- more measurable.



# DATA COLLECTION

**What data should (not) be collected**

**Who owns the data**

**Whose data can (not) be shared**

**What technology for collecting, storing, managing data**

**Whose data can (not) be traded**

**What data can (not) be merged**

**What to do with prejudicial data**

# DATA MODELING

## Data is biased (known/unknown)

- Invalid assumptions
- Confirmation bias

## Publication bias

- WSDM 2017: <https://arxiv.org/abs/1702.00502>

## Badly handling missing values

# DEPLOYMENT

**Spurious correlation / over-generalization**

**Using “black-box” methods that cannot be explained**

**Using heuristics that are not well understood**

**Releasing untested code**

**Extrapolating**

**Not measuring lifecycle performance (concept drift in ML)**

**We will go over ways to counter  
this in the ML/stats/hypothesis  
testing portion of the course**

# **GUIDING PRINCIPLES**

**Start with clear user need and public benefit**

**Use data and tools which have minimum intrusion necessary**

**Create robust data science models**

**Be alert to public perceptions**

**Be as open and accountable as possible**

**Keep data secure**



**GOV.UK**

Thanks to: UK cabinet office

# SOME REFERENCES

Presentation on ethics and data analysis, Kaiser Fung @ Columbia Univ. [http://andrewgelman.com/wp-content/uploads/2016/04/fung\\_ethics\\_v3.pdf](http://andrewgelman.com/wp-content/uploads/2016/04/fung_ethics_v3.pdf)

O'Neil, Weapons of math destruction.  
<https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815>

UK Cabinet Office, Data Science Ethical Framework.  
<https://www.gov.uk/government/publications/data-science-ethical-framework>

Derman, Modelers' Hippocratic Oath.  
<http://www.ijournals.com/doi/pdfplus/10.3905/jod.2012.20.1.035>

Nick D's MIT Tech Review Article.  
<https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/>

# FINDING A JOB

## **Make a personal website.**

- Free hosting options: GitHub Pages, Google Sites
- Pay for your own URL (but not the hosting).
- Make a clean website, and make sure it renders on mobile:
  - Bootstrap: <https://getbootstrap.com/>
  - Foundation: <http://foundation.zurb.com/>

**Highlight relevant coursework, open source projects, tangible work experience, etc**

**Highlight tools that you know (not just programming languages, but also frameworks like TensorFlow and general tech skills)**

# “REQUIREMENTS”

**Data science job postings – and, honestly, CS postings in general – often have completely nonsense requirements**

1. The group is filtering out some noise from the applicant pool
2. Somebody wrote the posting and went buzzword crazy

**In most cases (unless the position is a team lead, pure R&D, or a very senior role) you can work around requirements:**

- A good, simple website with good, clean projects can work wonders here ...
- Reach out and speak directly with team members
- Alumni network, internship network, online forums

# INTERVIEWING

**We saw that there is no standard for being a “data scientist” – and there is also no standard interview style ...**

**... but, generally, you’ll be asked about the five “chunks” we covered in this class, plus core CS stuff:**

- Software engineering questions
- Data collection and management questions (SQL, APIs, scraping, newer DB stuff like NoSQL, Graph DBs, etc)
- General “how would you approach ...” EDA questions
- Machine learning questions (“general” best practices, but you should be able to describe DTs, RFs, SVM, basic neural nets, KNN, OLS, **boosting**, PCA, **feature selection**, clustering)
- Basic “best practices” for statistics, e.g., hypothesis testing

**Take-home data analysis project (YMMV)**



# GRADUATE SCHOOL, ACADEMIA, R&D, ...

Data science isn't really an academic discipline by itself, but it comes up **everywhere** within and without CS

- Modern science is built on a “CS and Statistics stack” ...

Academic work in the area:

- Outside of CS, using techniques from this class to help fundamental research in that field
- Within CS, fundamental research in:
  - Machine learning
  - Statistics (non-pure theory)
  - Databases and data management
  - Incentives, game theory, mechanism design
- Within CS, trying to automate data science (e.g., Google Cloud's Predictive Analytics, “Automatic Statistician,” ...)

# Final Thoughts

1. No easy answers
2. Play, explore, think
3. Use off-the-shelf technologies wherever possible
4. Think about possible introduction of biases and be skeptical of 'clear' results

# Course Evals

1. If you're able, do the course evals
2. I try to take feedback seriously (though feedback doesn't always agree!)



Thanks for your time!

Have a good summer!