Responsible Data Science

Final Exam Review

Prof. George Wood

Center for Data Science
New York University

data RESPONSIBLY

NYU

Where are we in the course?

Fairness

Data Protection

Review

1 2 3 4 5 6 7 8 9 10 11 12 13 14

DS Lifecycle

Transparency & Interpretability

r/ai

Key dates

Week 12:

- Draft project report due: Friday, April 22

Week 13:

- Final exam: Monday, April 25

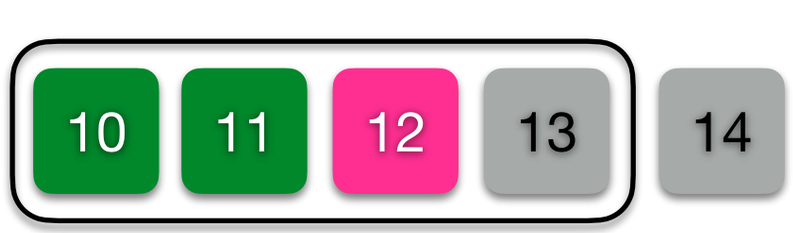
Week 14:

- Homework 3 due: Thursday, May 5

Week 15:

- Project report due: Monday, May 9

Review



Transparency & Interpretability

Final exam logistics

- 30% of the course grade
- Complete online (remotely), 120 minutes
- Exam “window” opens at 8am EDT on Monday, April 25. Submit no later than 10:00pm EDT on Monday, April 25
- I will be available online for clarification questions you may have **as you are working** on the exam, at my office hours Zoom link. Times I will be online to be announced soon

Exam must be completed individually. Do not discuss the problems and the solutions with anyone in the class, even after you submit your solution.

What's covered and how to prepare

- Covers all lectures, labs, and **assigned reading** up to and including Week 11 (transparency in online ad delivery)
- Questions will be similar to those you saw on the homework, go over all homework questions and solutions to prepare
- You will not be asked to write any code during the exam
- You may be asked to answer quantitative or discussion-style questions of the kind we saw on homework assignments

r/ai

fairness

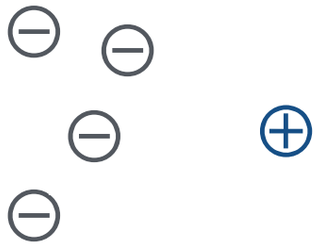
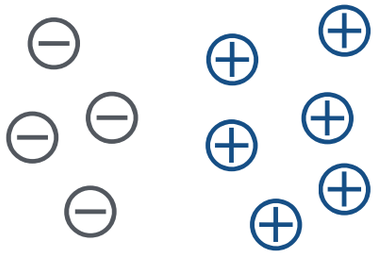
r/ai

Fairness in risk assessment

- A risk assessment tool **gives a probability estimate of a future outcome**
- Used in many domains:
  - insurance, criminal sentencing, medical testing, hiring, banking
  - also in less-obvious set-ups, like online advertising
- Fairness in risk assessment is concerned with how different kinds of error are distributed among sub-populations

Calibration

positive outcomes: do recidivate

	risk score		
	0.2	**0.6**	**0.8**
White			
Black			

given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

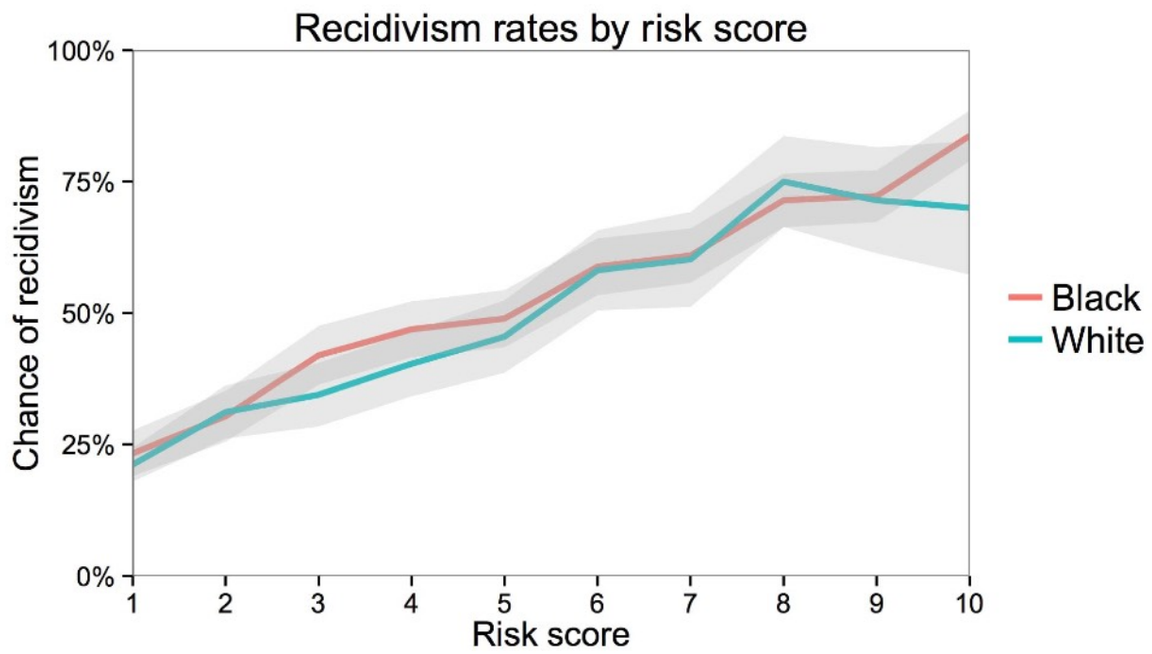
0.6 means 0.6 for any defendant - likelihood of recidivism

why do we want calibration?

COMPAS as a predictive instrument

Predictive parity (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is reasonably well-calibrated



[plot from Corbett-Davies et al.; *WaPo 2016*]

An impossibility result

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups.

Recidivism rates in the ProPublica dataset are higher for the Black group than for the White group

efendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

Achievable only in trivial cases

- Perfect information: the tool knows who recidivates (score 1) and who does not (score 0)
- Equal base rates: the fraction of positive-class people is the same for both groups

a negative result, need tradeoffs

proof sketched out in (starts 12 min in)

https://www.youtube.com/watch?v=UUC8tMNxwV8

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Fairness for whom?

Decision-maker: of those labeled low-risk, how many will recidivate?

Defendant: how likely will I be incorrectly labeled high-risk?

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

based on a slide by Arvind Narayanan

HW1

r/ai

HW 1: 1(a)

Consider again the COMPAS investigation by ProPublica, and additionally consider Northpointe's response to ProPublica (you may wish to consult Northpointe's report). For each criterion A-E below, explain in 1-2 sentences which stakeholders it benefits and why. If you believe that a criterion is not reasonable in this case, state so.

A. **Accuracy** (ACC), defined as (TP + TN) / (P + N): benefits multiple stakeholders, including the software vendor, the decision maker, and the general public. In a sense, high accuracy, like calibration, is a necessary condition that a risk instrument should meet to be considered useful.

B. **Positive predictive value** (PPV), defined as TP / (TP + FP), is higher when the true positives constitute a high proportion of the positive class. PPV benefits society, in the sense that we don't spend resources on incarcerating individuals who do not go on to reoffend (FP). It also clearly benefits the innocent defendants. It also benefits the software vendor (it's an important measure of accuracy).

C. **False positive rate** (FPR), defined as FP / N, does not directly benefit any stakeholder. However, because of the trade-off with other measures, a decision maker may be willing to incur as higher FPR to help lower the false negative rate (FNR).

HW 1: 1(a)

Consider again the COMPAS investigation by ProPublica, and additionally consider Northpointe's response to ProPublica (you may wish to consult Northpointe's report). For each criterion A-E below, explain in 1-2 sentences which stakeholders it benefits and why. If you believe that a criterion is not reasonable in this case, state so.

D. **False negative rate** (FNR), defined as FN / P, benefits society by ensuring that individuals who are likely to recidivate do not have a chance to do so if released. (This is under the assumption that incarcerating individuals actually prevents them from committing crime, either immediately following incarceration or in the long run.) It also benefits a decision maker whose reputation will suffer if an individual whom they release goes on to commit a crime. Finally, it benefits the software vendor that caters to this type of a decision maker.

E. **Statistical parity** (demographic parity among the individuals receiving any prediction), **does not make sense in this application**, because we are not interested in equalizing outcomes between populations per se. Rather, fairness here amounts to balancing the kinds of error that different demographic groups incur - a goal that, as we know from the literature, cannot be met directly and so requires trade-offs.

HW 1: 1(b)

(b) (6 points) Consider a hypothetical scenario in which *TechCorp*, a large technology company, is hiring for data scientist roles. Alex, a recruiter at *TechCorp*, uses a resume screening tool called *Prophecy* to help identify promising candidates. *Prophecy* takes applicant resumes as input and returns them in ranked (sorted) order, with the more promising applicants (according to the tool) appearing closer to the top of the ranked list. Alex takes the output of the *Prophecy* tool under advisement when deciding whom to invite for a job interview.

In their 1996 paper “Bias in computer systems”, Friedman & Nissenbaum discuss three types of bias: **A.** pre-existing, **B.** technical, and **C.** emergent. We also discussed these types of bias in class and in the “All about that Bias” comic.

For each type of bias:

- Give an example of how this type of bias may arise in the scenario described above;
- Name a stakeholder group that may be harmed by this type of bias; and
- Propose an intervention that may help mitigate this type of bias.

HW 1: 1(b)

A. **Pre-existing Bias:**
- Example: If a company has historically had a male-dominant workforce, then the model has few women to learn from and could learn to favor men in its rankings
- Stakeholders: Members of the underrepresented group — in this case, women
- Intervention: Gather a dataset that is more evenly representative or upsample

B. **Technical Bias:**
- Example: The system breaks ties by taking alphabetical order
- Stakeholders: People with names later in the alphabet (can vary by nationality as well)
- Intervention: Break ties using a randomized method or score with more granularity

C. **Emergent Bias:**
- Example: Recruiters use this tool and overtime set up a feedback loop where they pick the top ranked candidates, which in turn leads the system to recommend more people like those already chosen
- Stakeholders: Those underrepresented in the workforce
- Intervention: Don’t only use this tool — also use some separate human judgement

HW 1: 1(c)

(c) (9 points) Consider a hypothetical scenario in which an admissions officer at *Best University* is evaluating applicants based on 3 features: SAT score, high school GPA, and family income bracket (low, medium, high). We discussed several equality of opportunity (EO) doctrines in class and in the "Fairness and Friends" comic: formal, substantive / luck egalitarian, and substantive / Rawlsian.

A. Formal EO would use SAT score or GPA, while income is irrelevant. (Could also say only SAT as it's theoretically more standardized)

B. Substantive/Rawlsian EO and substantive/luck-egalitarian EO are consistent with the goal of correcting these differences.

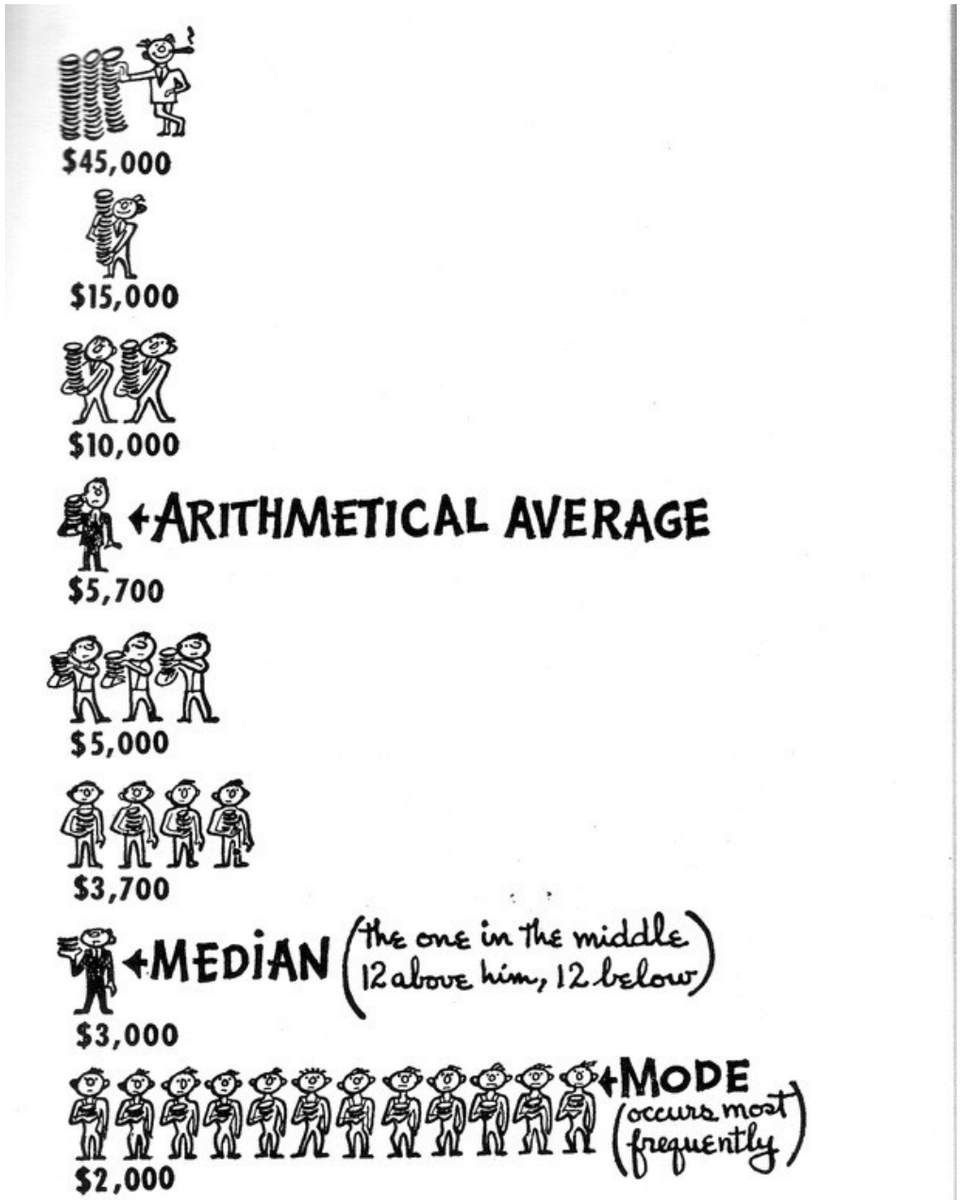
- Rawlsian EO would argue that the advantage of being born into a high-income family has snowballed into an advantage in SAT performance
- Luck-egalitarian EO would argue that the difference in SAT scores between income brackets is a matter of "brute luck."

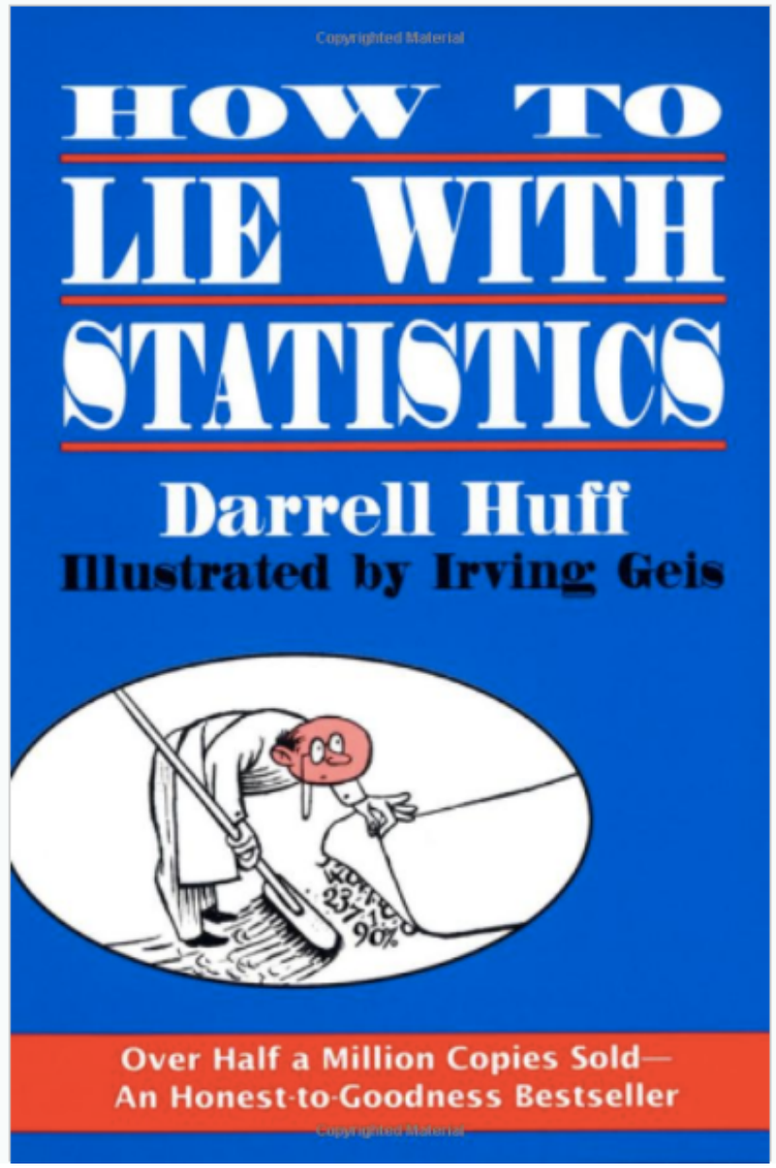
C. Luck-egalitarians would bucket by income and then take the top from each bucket

data profiling

r/ai

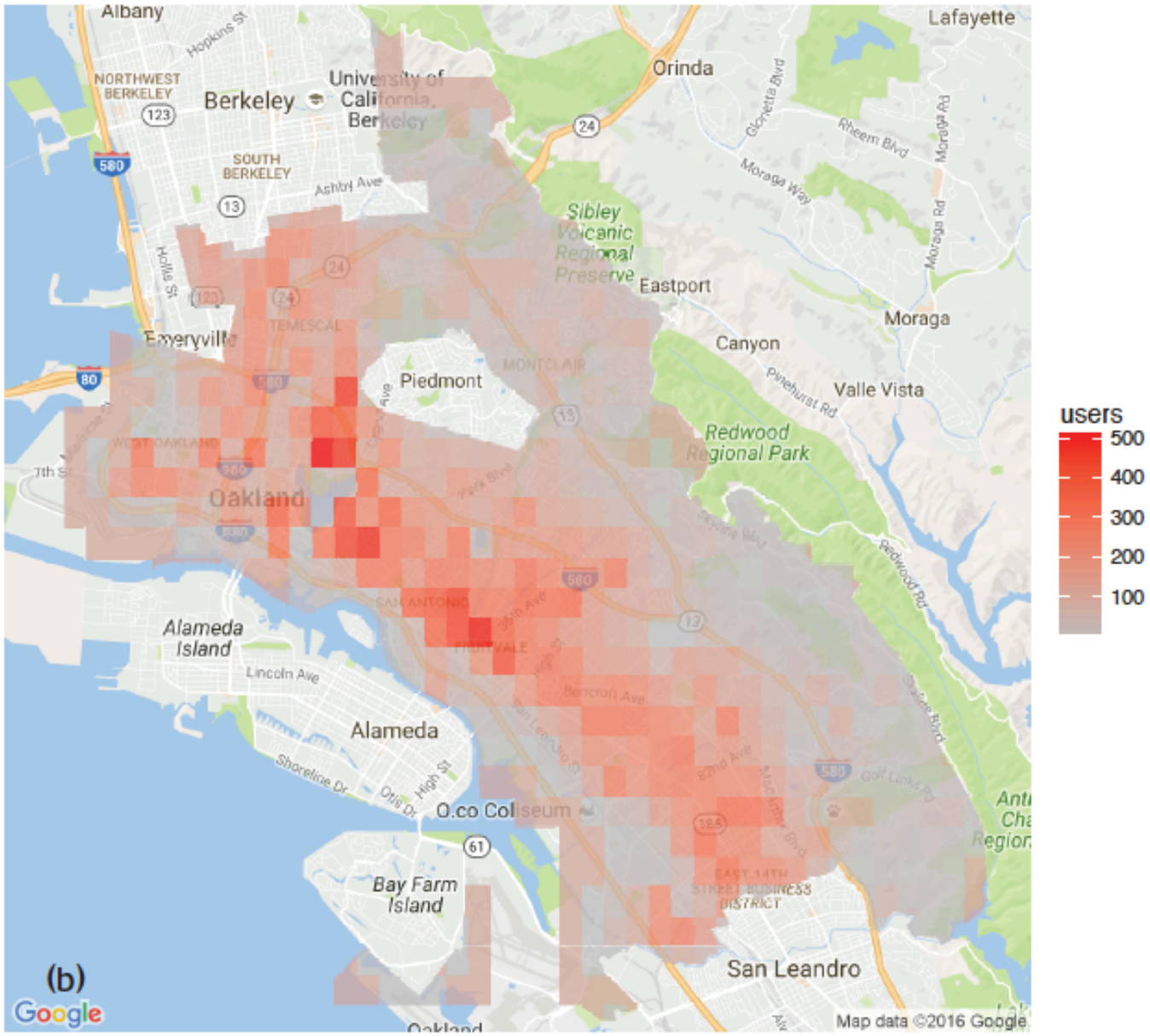
The well-chosen average



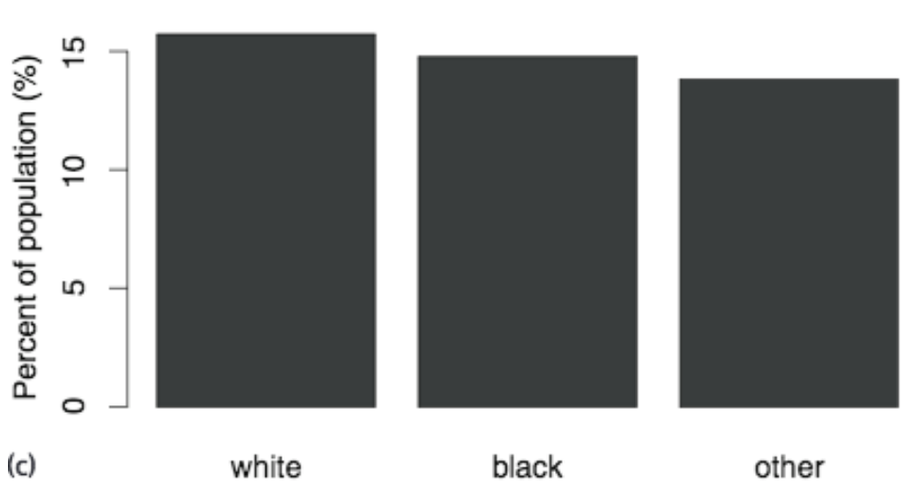


r/ai

Is my data biased? (histograms + geo)



Estimated number of drug users, based on 2011 National Survey on Drug Use and Health, in Oakland, CA

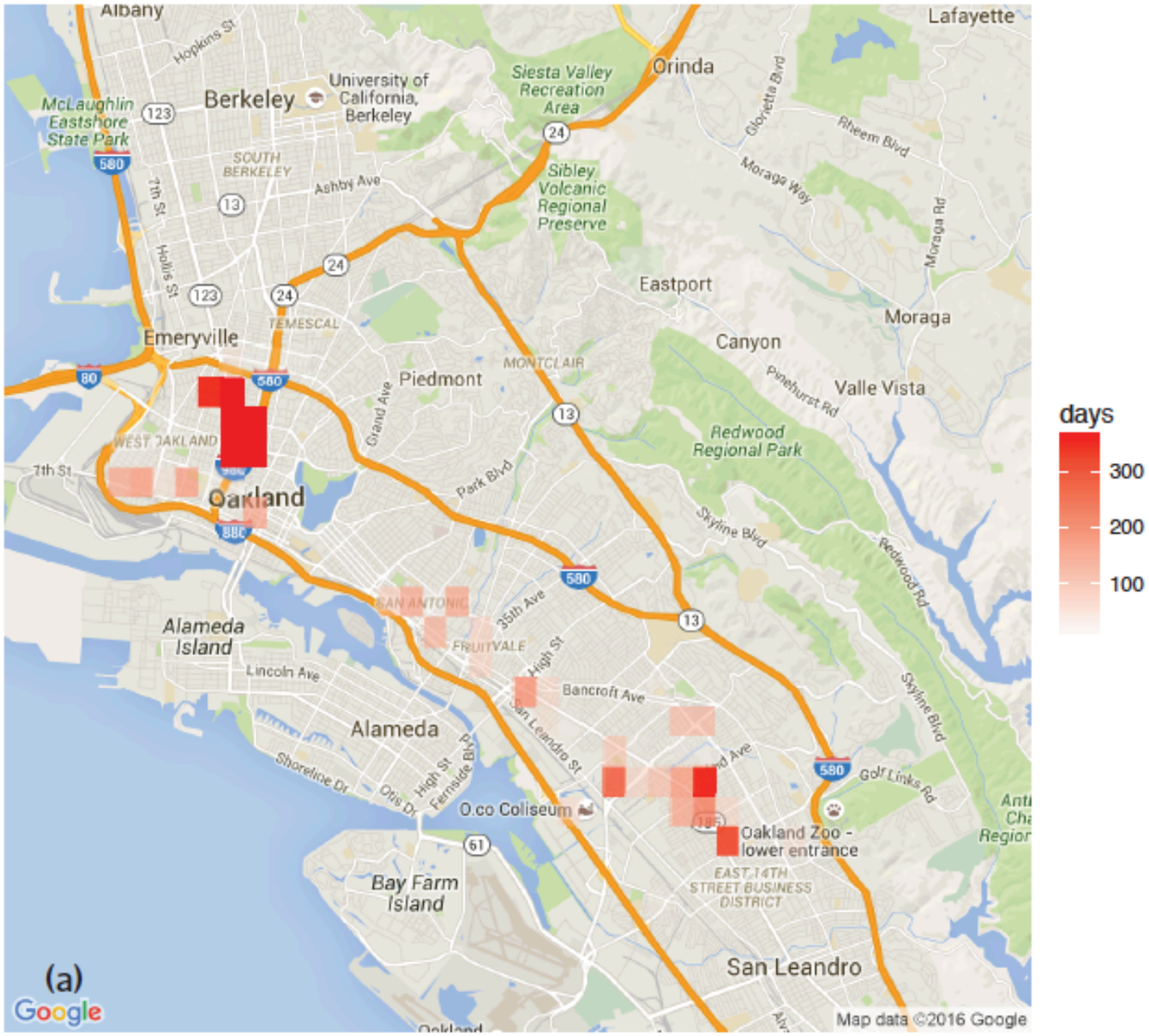


Estimated drug use by race

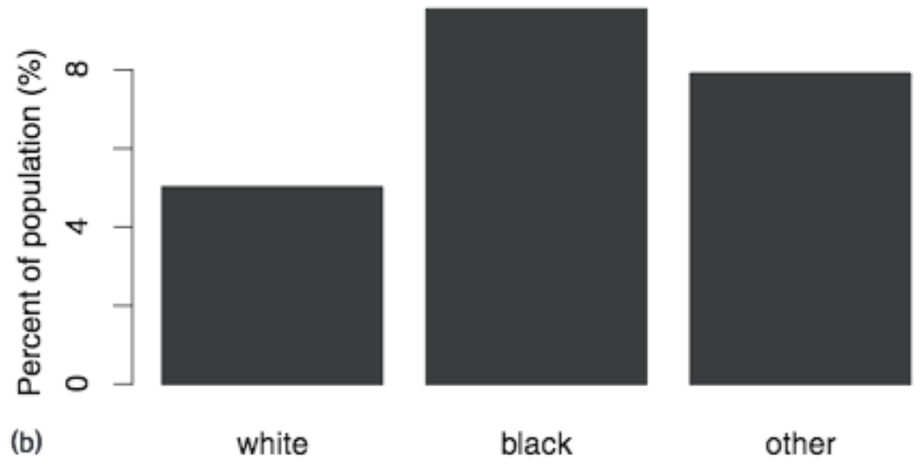
[Lum & Isaac (2016)]

r/ai

Is my data biased? (histograms + geo)



Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland, CA, police data for 2011

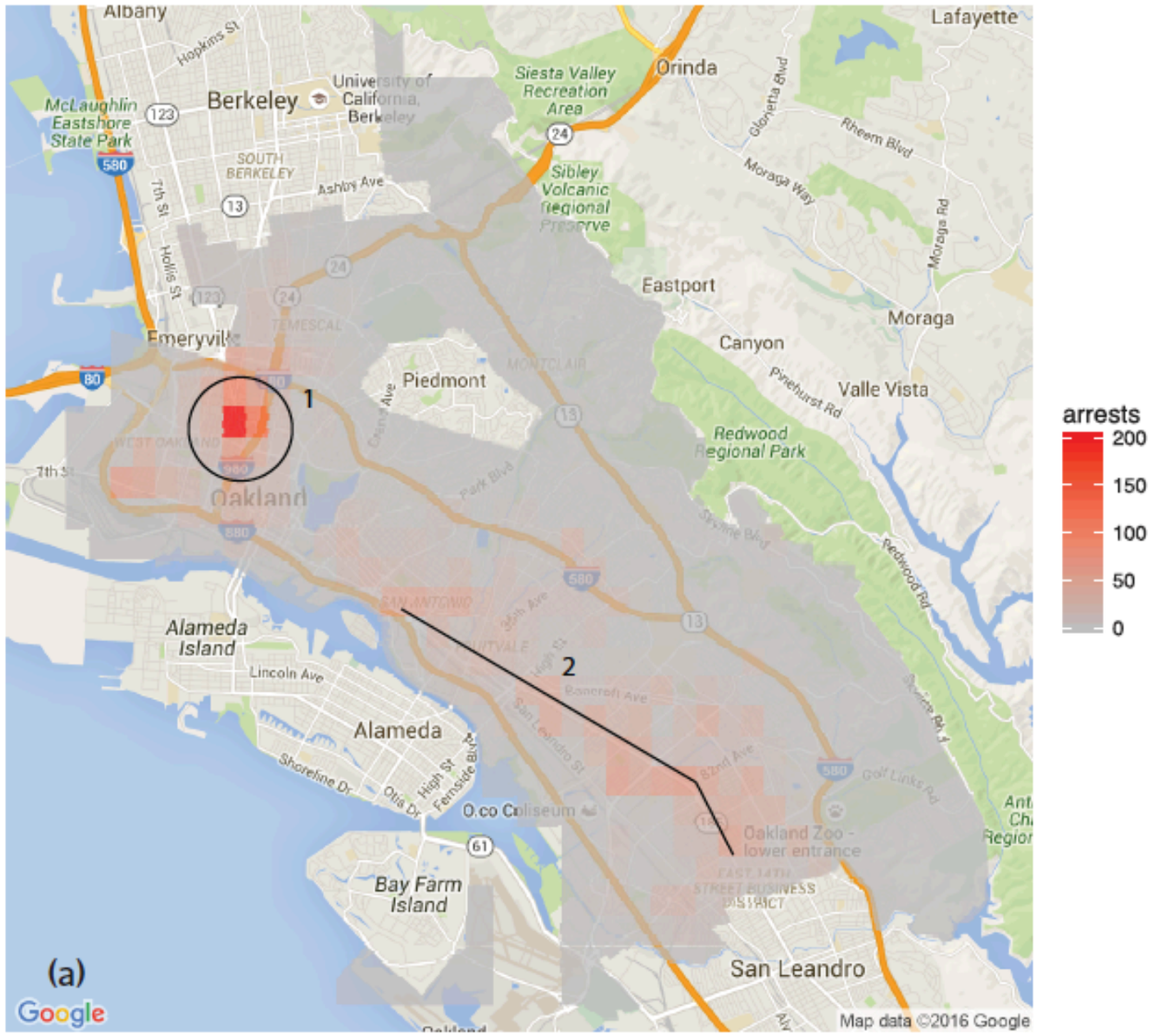


Targeted policing for drug crimes by race

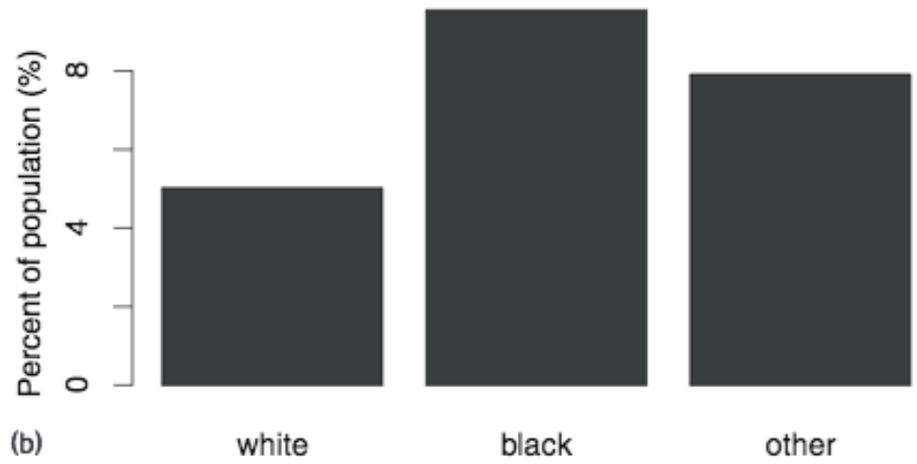
[Lum & Isaac (2016)]

r/ai

Is my data biased? (histograms + geo)



Number of drug arrests made by the Oakland, CA, police department in 2010



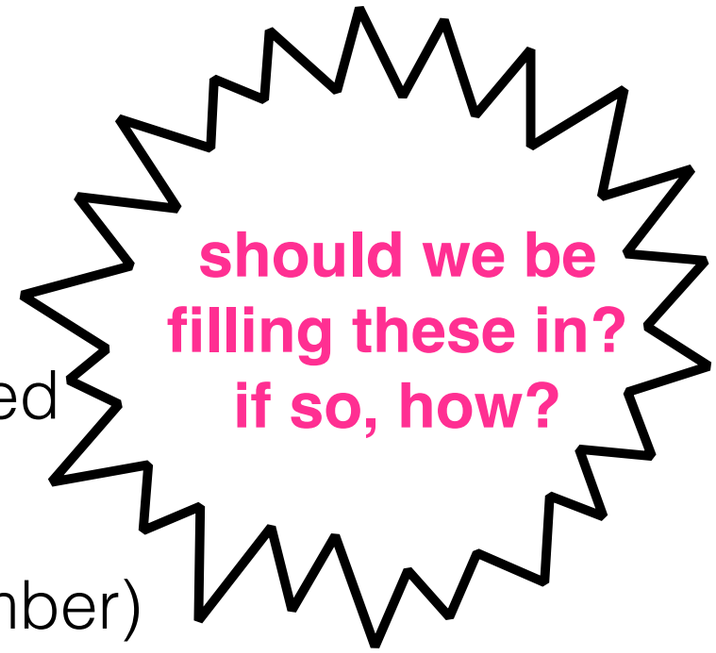
Targeted policing for drug crimes by race

[Lum & Isaac (2016)]

r/ai

50 shades of *null*

- **Unknown** - some value definitely belongs here, but I don't know what it is (e.g., unknown birthdate)
- **Inapplicable** - no value makes sense here (e.g., if marital status = single then spouse name should not have a value)
- **Unintentionally omitted** - values is left unspecified unintentionally, by mistake
- **Optional** - a value may legitimately be left unspecified (e.g., middle name)
- **Intentionally withheld** (e.g., an unlisted phone number)
- …..



https://www.vertabelo.com/blog/technical-articles/50-shades-of-null-or-how-a-billion-dollar-mistake-has-been-stalking-a-whole-industry-for-decades

Missing value imputation

are values **missing at random** (e.g., gender, age, disability on job applications)?

are we ever interpolating **rare categories** (e.g., Native American)

are **all categories** represented (e.g., non-binary gender)?

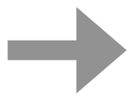
Data filtering

“filtering” operations (like selection and join), **can arbitrarily change demographic group proportions**

select by zip code, country, years of C++ experience, others?

age_group	county
60	CountyA
60	CountyA
20	CountyA
60	CountyB
20	CountyB
20	CountyB

50% vs 50%



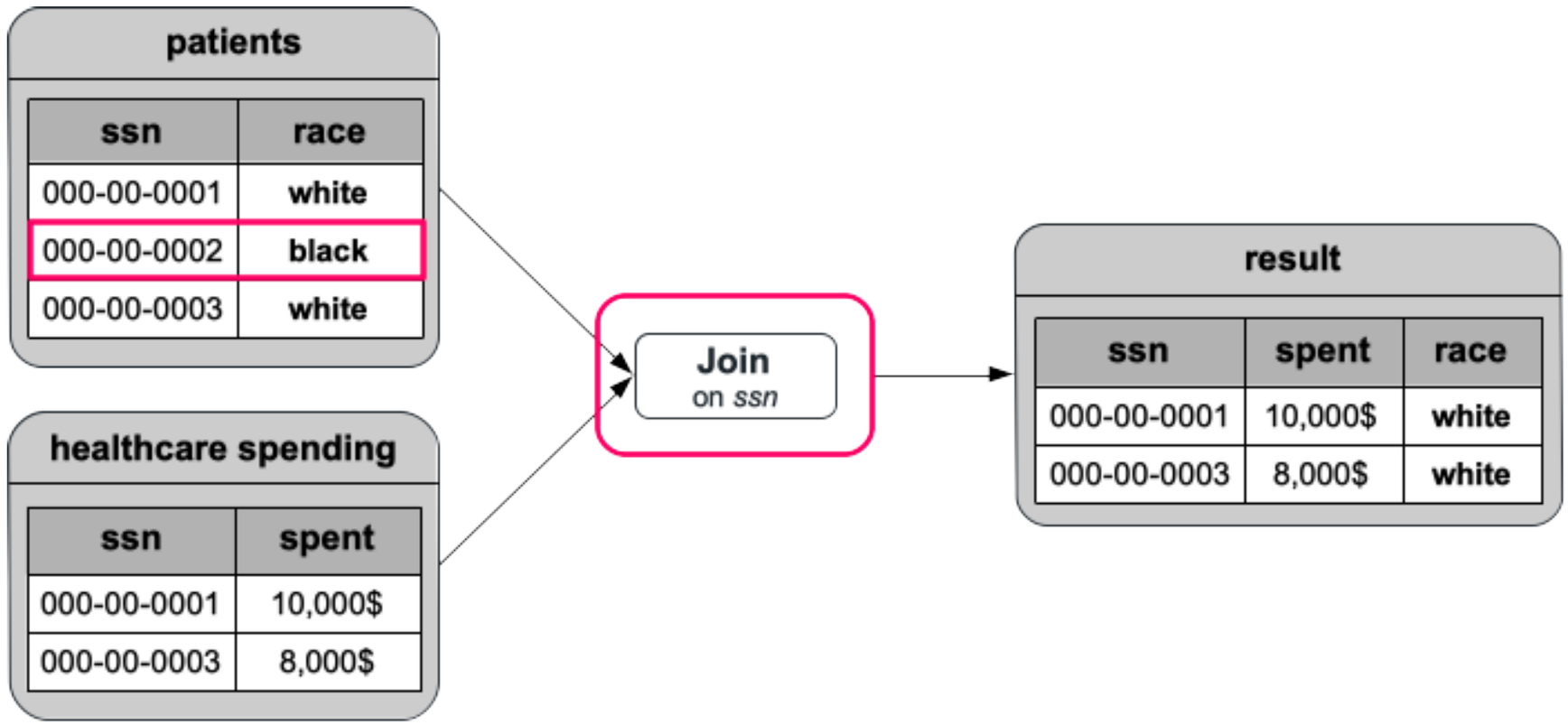
age_group	county
60	CountyA
60	CountyA
20	CountyA

66% vs 33%

Data filtering

"filtering" operations (like selection and join), **can arbitrarily change demographic group proportions**

select by zip code, country, years of C++ experience, others?



privacy

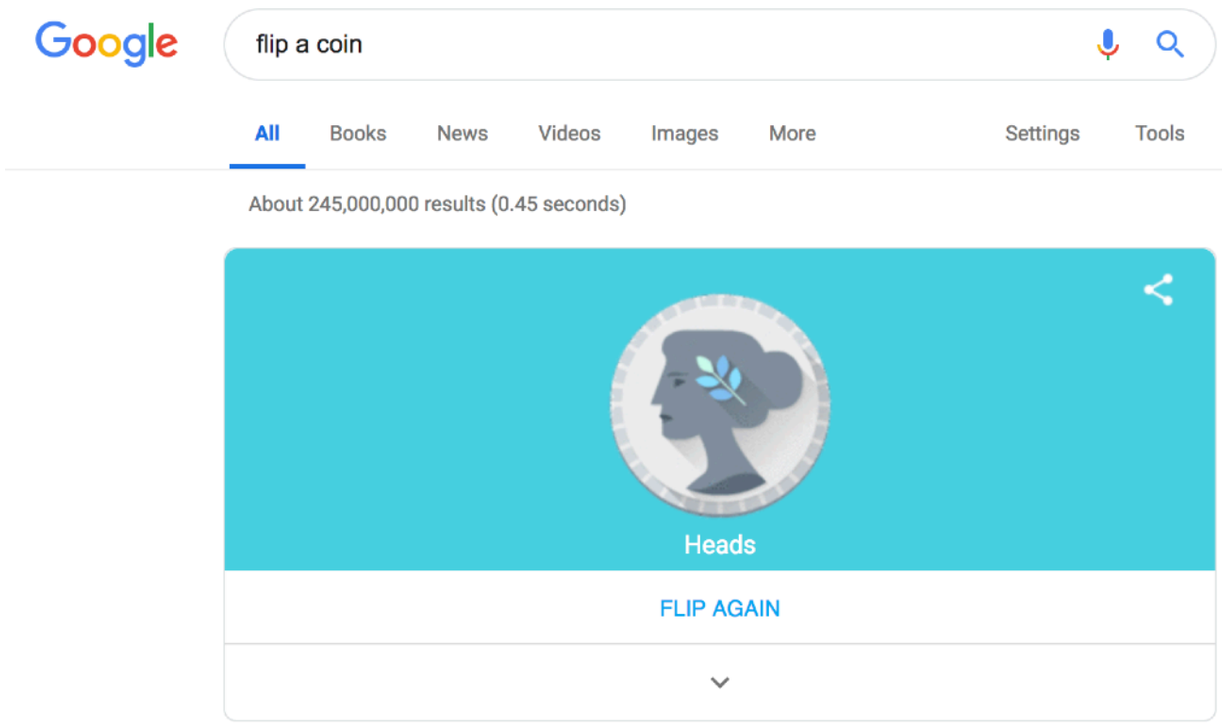
r/ai

Truth or dare?

Did you go out drinking over the weekend?

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the class for whom **P** holds

1. flip a coin **C1**
    1. if **C1** is tails, then **respond truthfully**
    2. if **C1** is heads, then flip another coin **C2**
        1. if **C2** is heads then **Yes**
        2. else **C2** is tails then respond **No**



the expected number of **Yes** answers is:

$$A = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{p}{2}$$

thus, we estimate **p** as:

$$\tilde{p} = 2A - \frac{1}{2}$$

r/ai

Randomized response

Did you go out drinking over the weekend?

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the class for whom **P** holds

1. flip a coin **C1**
   1. if **C1** is tails, then **respond truthfully**
   2. if **C1** is heads, then flip another coin **C2**
      1. if **C2** is heads then **Yes**
      2. else **C2** is tails then respond **No**

randomization - adding noise - is what gives plausible deniability a process privacy method

the expected number of **Yes** answers is:

$$A = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{p}{2}$$

privacy comes from plausible deniability

r/ai

Do we really need randomization?

- Data release approaches that fail to protect privacy (these are prominent classes of methods, there are others):
  - **sampling** (“**just a few**”) - release a small subset of the database
  - **aggregation** (e.g., **k-anonymity** - each record in the release is indistinguishable from at least k-1 other records)
  - **de-identification** - mask or drop personal identifiers
  - **query auditing** - stop answering queries when they become unsafe

Aggregation without randomization

- Alice and Bob are professors at State University.
- In March, Alice publishes an article: "…. the current freshman class at State U is **3,005** students, **202** of whom are from families earning over $1M per year."
- In April, Bob publishes an article: "… **201** families in State U's freshman class of **3,004** have household incomes exceeding $1M per year."
- Neither statement discloses the income of the family of any one student. But, taken together, they state that **John, a student who dropped out at the end of March**, comes from a family that earns $1M. Anyone who has this **auxiliary information** — that John dropped out at the end of March — will be able to learn about the income of John's family.

this is known as a problem of **composition**, and can be seen as a kind of a **differencing attack**

A linkage attack: Governor Weld

In 1997, Massachusetts Group Insurance Commission released "anonymized" data on state employees that showed every single hospital visit!

Latanya Sweeney, a grad student, sought to show the ineffectiveness of this “anonymization.”

She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes.

For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter.

Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code.

Follow up: ZIP code, birthdate, and sex sufficient to identify 87% of Americans!

https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

Privacy: two sides of the coin

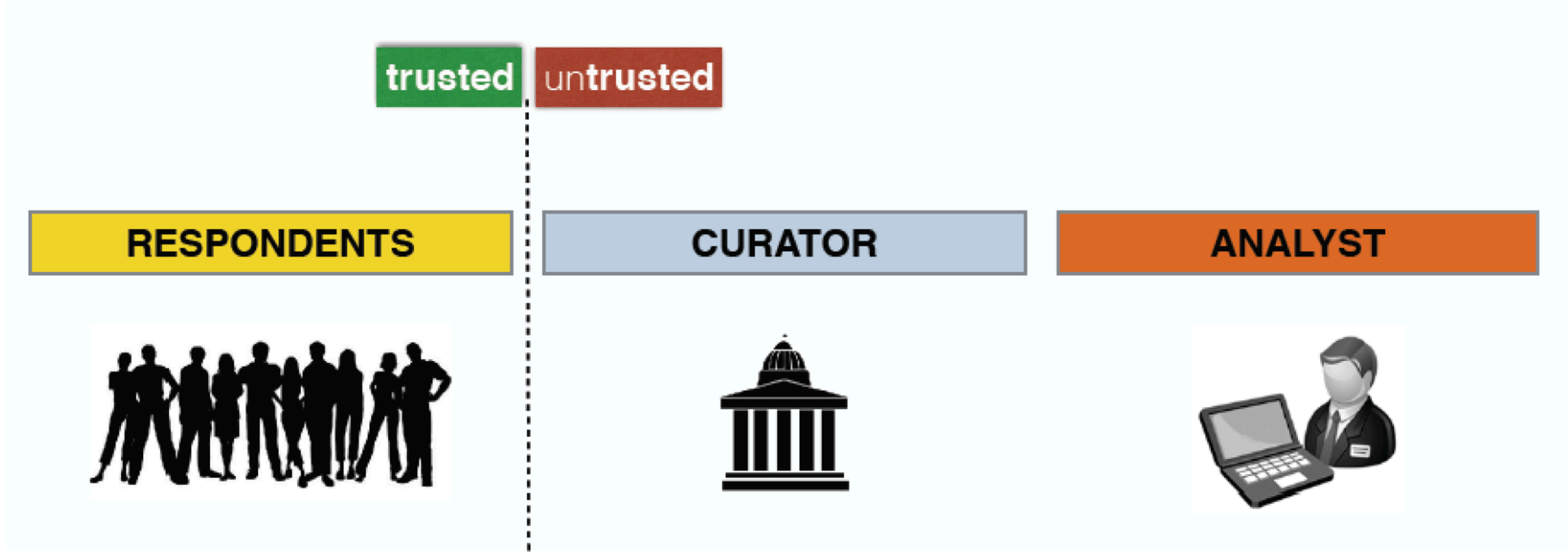
protecting an individual

plausible deniability

learning about the population

noisy estimates

Privacy-preserving data analysis

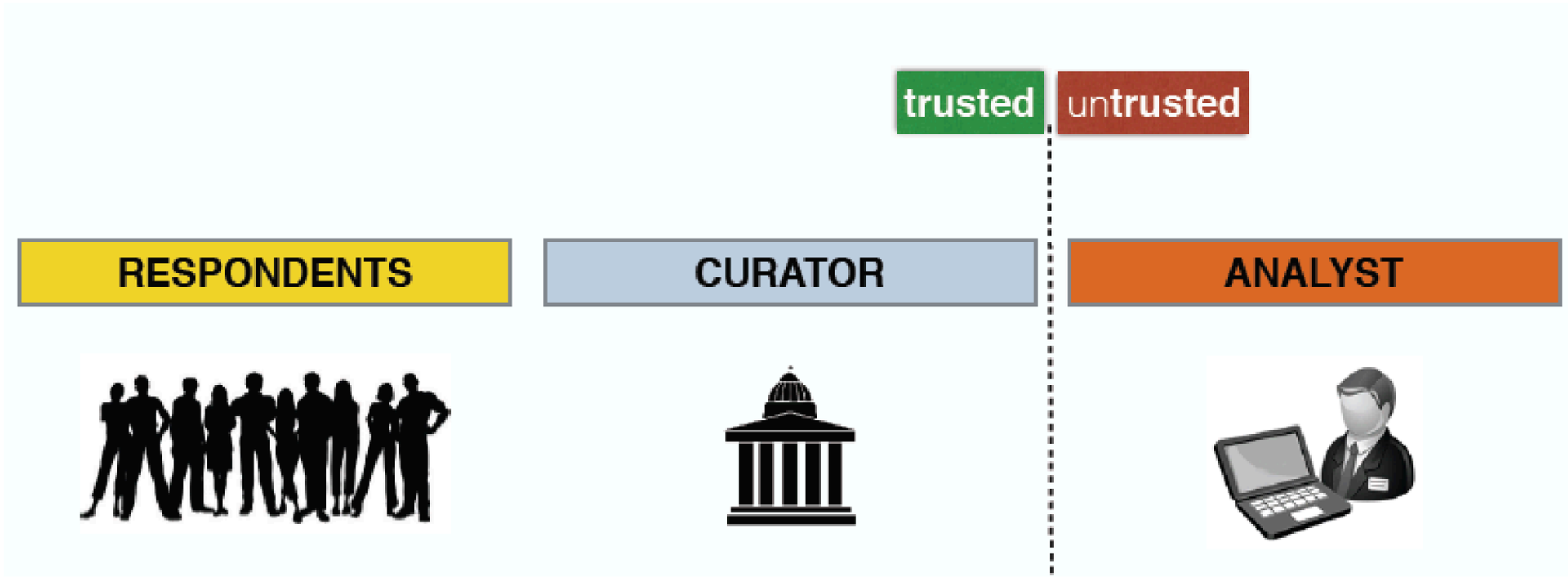


respondents contribute their personal data

the **curator** is **untrusted**, collects data, releases it to analysts

the **analyst** is **untrusted**, extracts value from data

Privacy-preserving data analysis



respondents in the population seek protection of their personal data

the **curator** is **trusted** to collect data and is responsible for safely releasing it

the **analyst** is **untrusted** and wants to gain the most accurate insights into the population

Differential privacy: the formalism

The information released about the sensitive dataset is virtually indistinguishable **whether or not a respondent's data is in the dataset**. This is an informal statement of **differential privacy**. That is, no information **specific to an individual** is revealed.

A randomized algorithm ***M*** provides **ε-differential privacy** if, for all neighboring databases D_1 and D_2, and for any set of outputs ***S***:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

ε (epsilon) is a privacy parameter

lower ε = stronger privacy

The notion of **neighboring databases** is integral to plausible deniability: D_1 can represent a database with a particular respondent's data, D_2 can represent a neighboring database but without that respondent's data

r/ai

Back to randomized response

Did you go out drinking over the weekend?

1. flip a coin **C1**
    1. if **C1** is tails, then **respond truthfully**
    2. if **C1** is heads, then flip another coin **C2**
        1. if **C2** is heads then **Yes**
        2. else **C2** is tails then respond **No**

Denote:

- Truth=Yes by **P**
- Response=Yes by **A**
- **C1**=tails by **T**
- **C1**=heads and **C2**=tails by **HT**
- **C1**=heads and **C2**=heads by **HH**

A randomized algorithm ***M*** provides **ε-differential privacy** if, for all neighboring databases D_1 and D_2, and for any set of outputs ***S***:

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \Pr[M(D_2) \in S]$$

$$\Pr[A \mid P] = \Pr[T] + \Pr[HH] = \frac{3}{4}$$

$$\Pr[A \mid \neg P] = \Pr[HH] = \frac{1}{4}$$

$$\Pr[A \mid P] = 3 \Pr[A \mid \neg P]$$

$$\Rightarrow \varepsilon = \ln 3$$

our version of randomized response is $(\ln 3)$-differentially private

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} \left| q(D) - q(D') \right|$$

lower ε = stronger privacy

- Example 1: counting queries
  - "How many elements in ***D*** satisfy property ***P***?" **What's *Δq*?**
  - "What fraction of the elements in ***D*** satisfy property ***P***?"
- Example 2: max / min
  - "What is the maximum employee salary in ***D***?" **What's *Δq*?**

Intuition: for a given ε, the higher the sensitivity, the more noise we need to add to meet the privacy guarantee

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} \left| q(D) - q(D') \right|$$

query *q*	query sensitivity *Δq*
select count(*) from D	1
select count(*) from D where sex = Male and age > 30	1
select MAX(salary) from D	*MAX(salary)-MIN(salary)*
select gender, count(*) from D group by gender	1 (disjoint groups, presence or absence of one tuple impacts only one of the counts)

r/ai

Query sensitivity

The ℓ_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} \left| q(D) - q(D') \right|$$

query *q*	query sensitivity *Δq*
select gender, count(*) from D group by gender	**1** (**disjoint groups**, presence or absence of one tuple impacts only one of the counts)
an arbitrary list of *m* counting queries	***m*** (no assumptions about the queries, and so a single individual may change the answer of **every query** by 1)

r/ai

HW2

r/ai

HW2 solutions posted soon

Solutions will for HW2 will be posted on Saturday

ethical

frameworks

r/ai

Two ethical frameworks

Consequentialism (Jeremy Bentham, John Stuart Mill): Take actions that lead to better states in the world

Deontology (Immanuel Kant): Focus on ethical duties, independent of their consequences

Deontologists focus on ***means***, **consequentialists** focus on ***ends***

"Arguments between consequentialists and deontologists are like two ships passing in the night."

https://www.bitbybitbook.com/en/1st-ed/ethics/

r/ai

Two ethical frameworks

Deontologists focus on ***means*, consequentialists** focus on ***ends***

Individuals, to the degree that they are capable, should be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.

Both consequentialism and deontology support **informed consent**, but for different reasons.

Two ethical frameworks

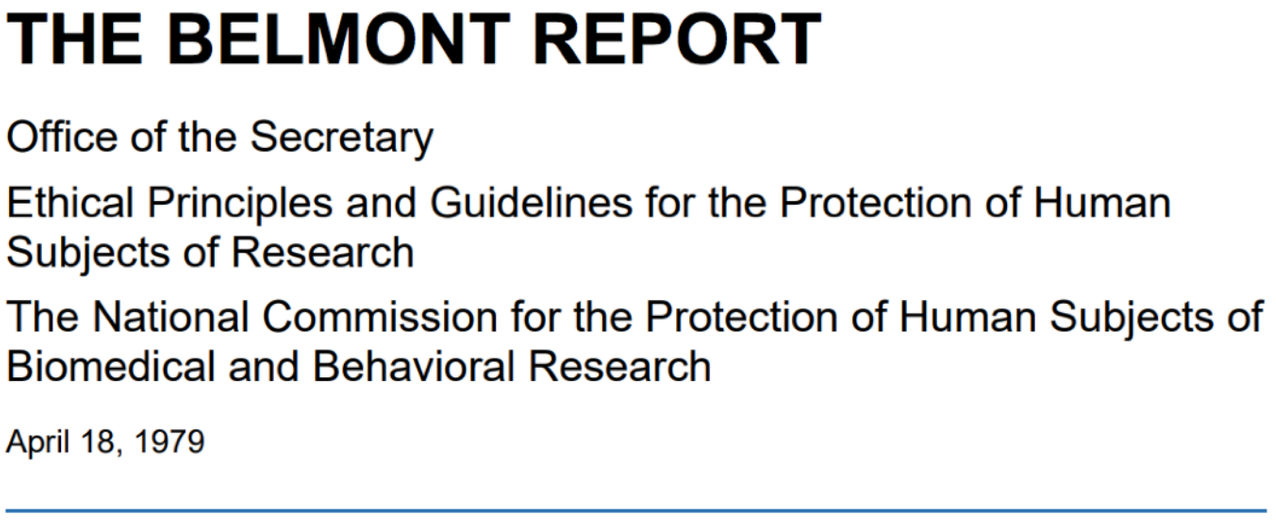
Deontologists focus on ***means*, consequentialists** focus on ***ends***

A **consequentialist** argument: Informed consent helps prevent harm to participants by prohibiting research that does not properly balance risk and anticipated benefit. In other words, consequentialist thinking would support informed consent because it helps **prevent bad outcomes** for participants.

A **deontological** argument for informed consent focuses on a researcher's duty to respect the **autonomy** of participants.

Given these arguments, a pure consequentialist might be willing to waive the requirement for informed consent in a setting where there was no risk, whereas a pure deontologist would not.

Belmont Report: Summary



THE BELMONT REPORT

Office of the Secretary

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

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

April 18, 1979

- Boundaries between research and practice
- Ethical principles
    - Respect for Persons
    - Beneficence
    - Justice
- Applications

r/ai

The Menlo Report: Summary

Principle	Application
Respect for Persons	Participation as a research subject is voluntary, and follows from informed consent; Treat individuals as autonomous agents and respect their right to determine their own best interests; Respect individuals who are not targets of research yet are impacted; Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection.
Beneficence	Do not harm; Maximize probable benefits and minimize probable harms; Systematically assess both risk of harm and benefit.
Justice	Each person deserves equal consideration in how to be treated, and the benefits of research should be fairly distributed according to individual need, effort, societal contribution, and merit; Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects.
Respect for Law and Public Interest	*Engage in legal due diligence; Be transparent in methods and results; Be accountable for actions.*

http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

r/ai

Recall: Racial bias in resume screening

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

September 2004

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. **White names receive 50 percent more callbacks for interviews.** Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U. S. labor market.

Back to Informed Consent

Research question: Does an employer unlawfully discriminate against applicants based on membership in protected groups?

Employers don't provide consent, in fact, they are actively deceived!

Field experiments to study discrimination are legally permissible **if**:

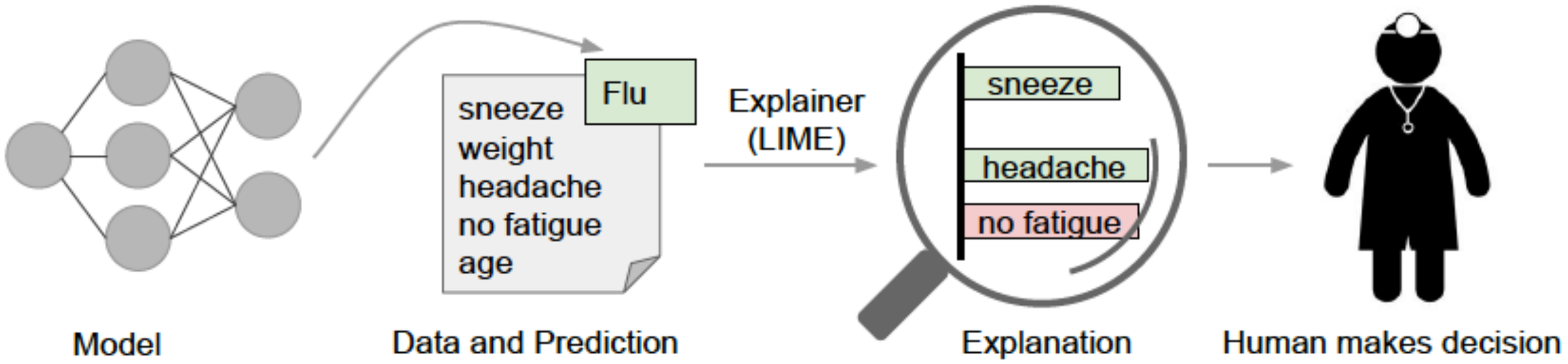
1. the harm to employers is limited, **and**
2. there is great social benefit to having a reliable measure of discrimination, **and**
3. other methods of measuring discrimination are weak; **and**
4. deception does not strongly violate the norms of that setting.

transparency

r/ai

Explanations based on features

- **LIME** (Local Interpretable Model-Agnostic Explanations): to help users trust a prediction, explain individual predictions
- **SP-LIME**: to help users trust a model, select a set of representative instances for which to generate explanations



features in green ("sneeze", "headache") support the prediction ("Flu"), while features in red ("no fatigue") are evidence against the prediction

what if patient id appears in green in the list? - an example of "data leakage"

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Key idea: Interpretable representation

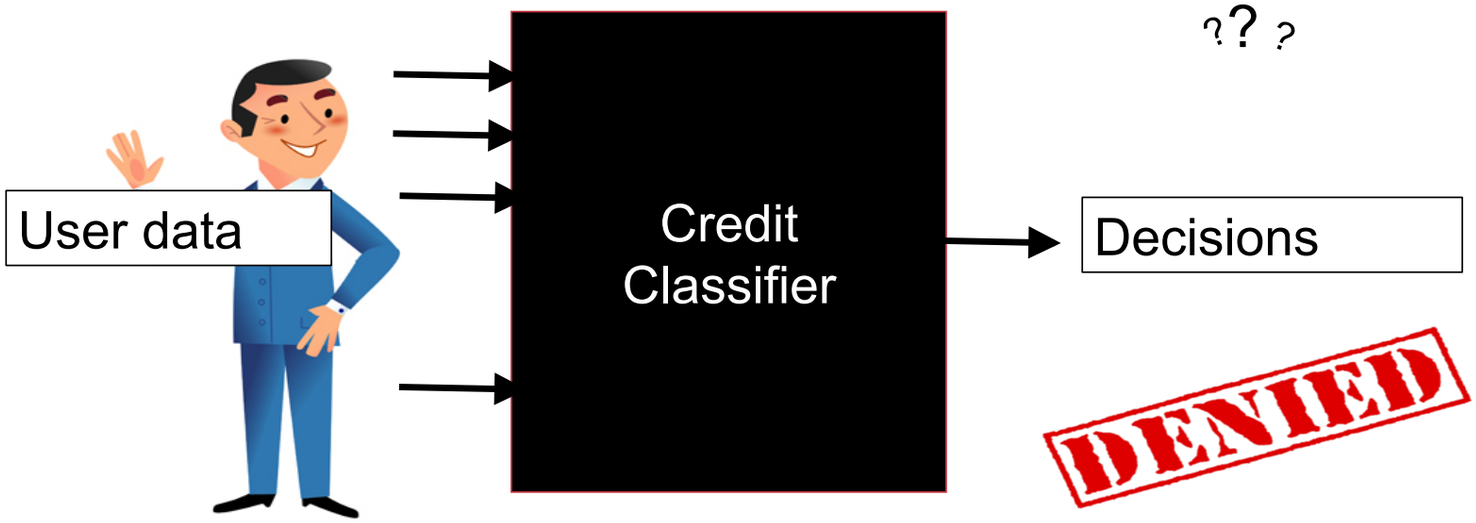
"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classier."

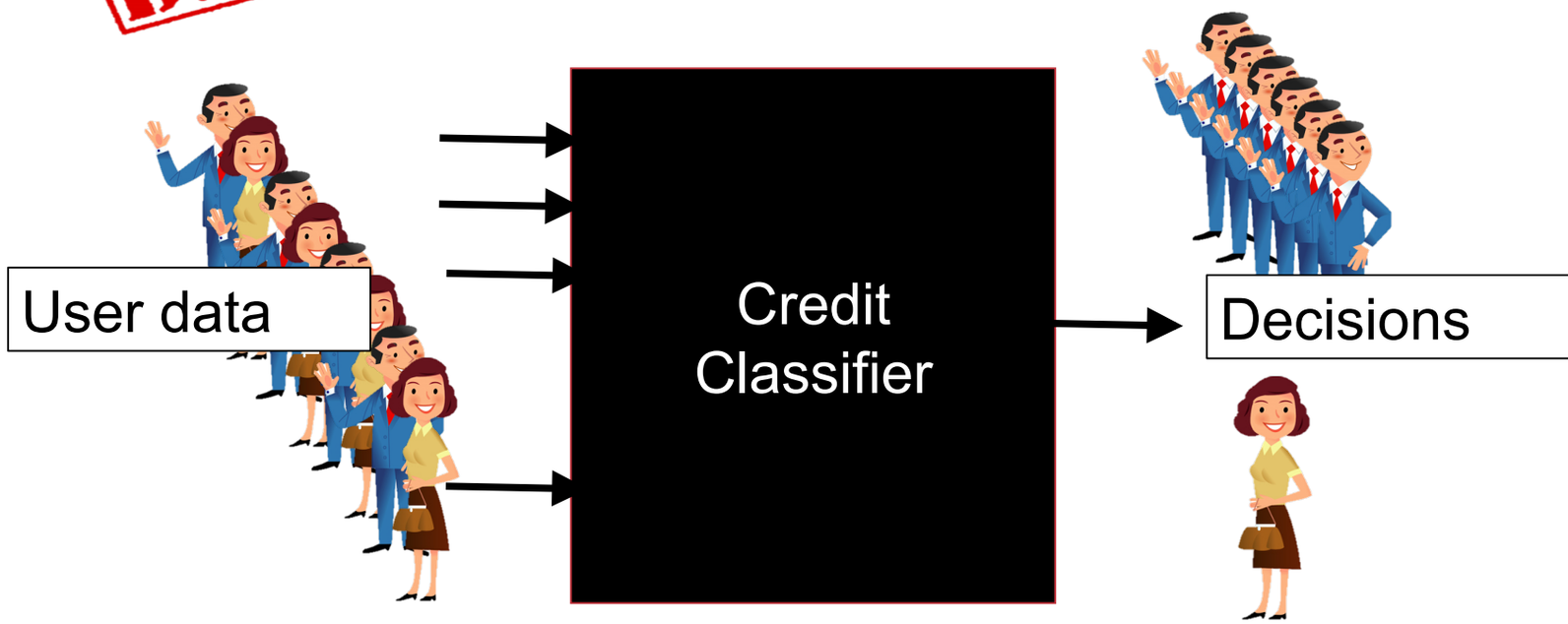
- LIME relies on a distinction between **features** and **interpretable data representations**; examples:
  - In text classification features are word embeddings; an interpretable representation is a vector indicating the presence of absence of a word
  - In image classification features encoded in a tensor with three color channels per pixel; an interpretable representation is a binary vector indicating the presence or absence of a contiguous patch of similar pixels
- **To summarize**: we may have some ***d*** features and ***d'*** interpretable components; interpretable models will act over domain $\{0, 1\}^{d'}$ - denoting the presence of absence of each of d' interpretable components

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Auditing black-box models





images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

QII: Quantitative Input Influence

Goal: determine how much influence an input, or a set of inputs, has on a **classification outcome** for an individual or a group

Transparency queries / quantities of interest

Individual: Which inputs have the most influence in my credit denial?

Group: Which inputs have the most influence on credit decisions for women?

Disparity: Which inputs influence men getting more positive outcomes than women?

[Datta, Sen & Zick, 2016]

r/ai

QII: Quantitative Input Influence

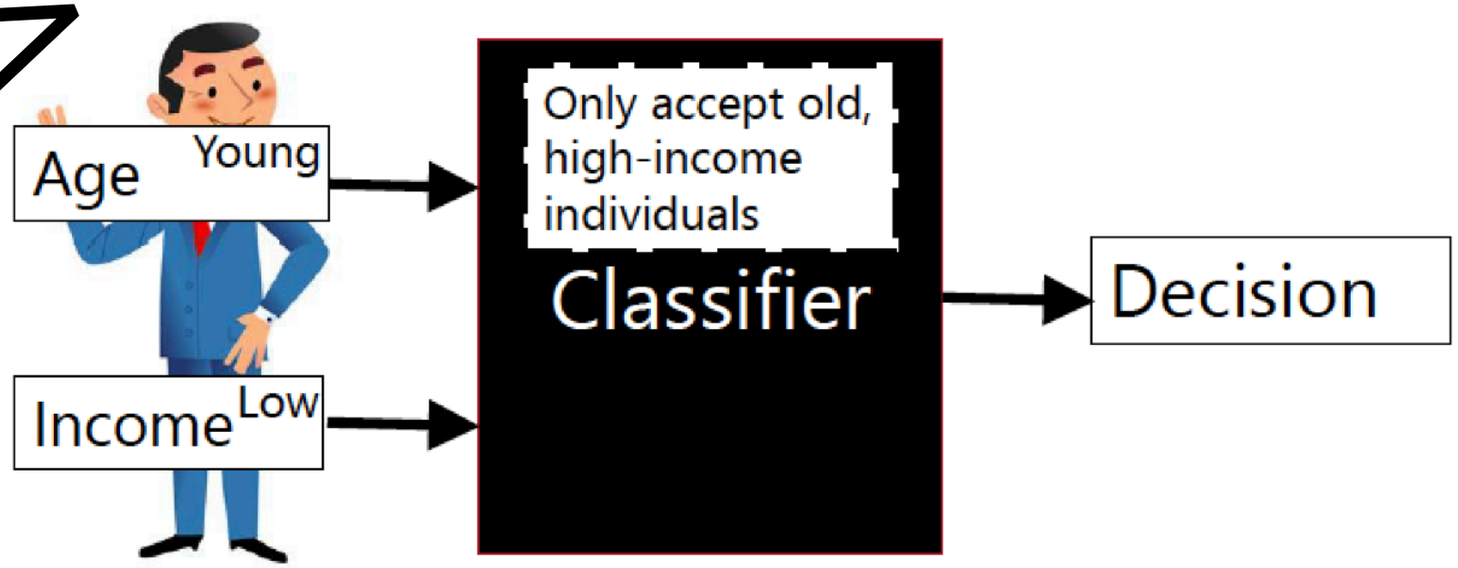
images by Anupam Datta

For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.

Key ideas

intervene on an input feature, measure its **importance**

aggregate feature importance using its **Shapley value**



in this case, intervening on one feature at a time will have no effect

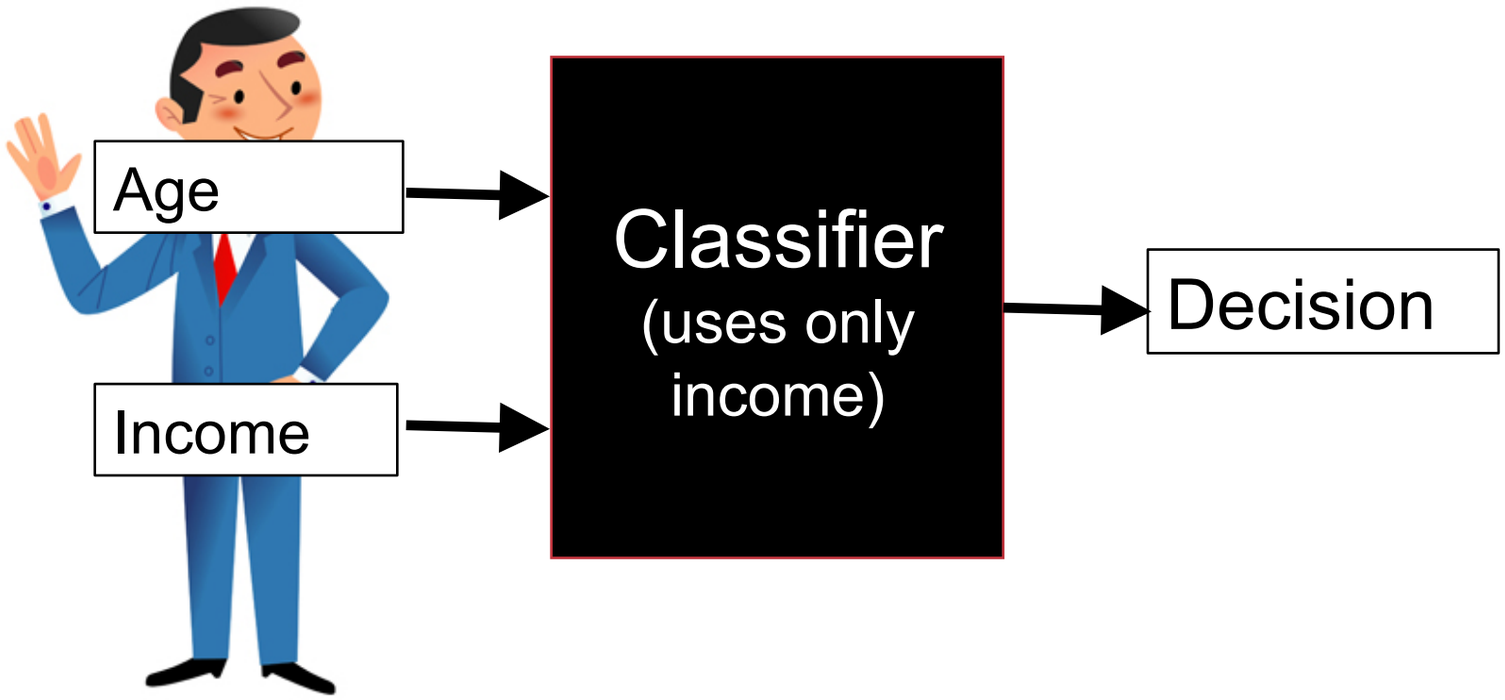
[Datta, Sen & Zick, 2016]

r/ai

Unary QII

images by Anupam Datta

For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.



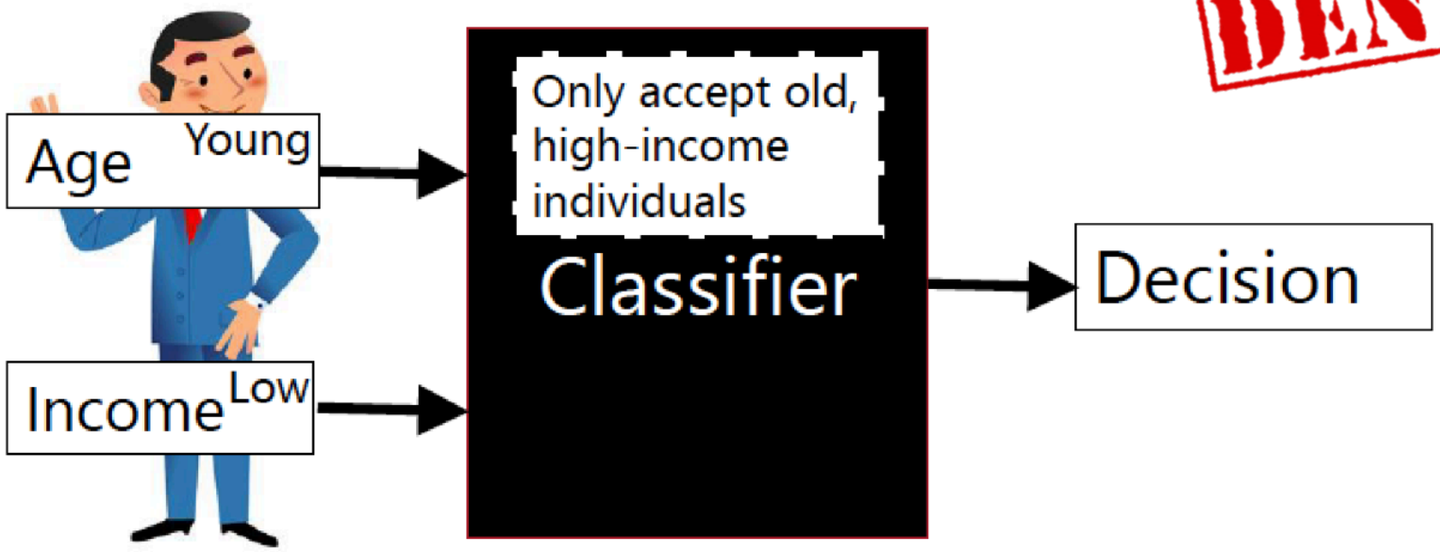
replace features with random values from the population, examine the distribution over outcomes

[Datta, Sen & Zick, 2016]

r/ai

Limitations of unary QII

For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.



intervening on one feature at a time will not have any effect

images by Anupam Datta

[Datta, Sen & Zick, 2016]