# Data Ethics COMP 394 (NLP)

# Course Evals! Back in 10

# What responsibilities do "we" have w.r.t. training data?

lf...

- You scrape data from the internet?
- You take a pre-collected dataset from researchers?
- You pay for a proprietary dataset?
- You use data collected from interactions with users?
- You pay workers to produce data for the purpose of training?

### **Discuss with your tables for 3–5 minutes.**

- What obligations do you think you have?
- What concerns might you have?
- How, if at all, do those differ across the different scenarios?
- What motivates these choices? What is a good framework for answering these questions?

# Three data-related topics for the day

- 1. Data Privacy
  - a. Is some data too private to train on? Do we have obligations toward the privacy of people creating our datasets? What are expectations of privacy within various kinds of corpora sources?
- 2. Data Documentation
  - a. If we can't read our data (scale!), what accountability do we have for its contents? What about downstream consequences of training data content (i.e., bias)?
- 3. Intellectual Property
  - a. Re: What obligations do we have to creators/owners of data? How does this interact with legal notions of copyright (fair use)?

# Dataset Leakage: Privacy and PII

- **PII**: Personally Identifiable Information
  - Names, addresses, phone numbers, etc.
- Training data might, for many reasons, contain PII:
  - You share bits of personal information within smaller (but public!) online communities, but that data was collected for model training!
- LLMs can violate that expectation of privacy!
  - "LLMs can Dox you!"

### ProPILE: Probing Privacy Leakage in Large Language Models

<sup>1</sup> Department of Electrical and Computer Engineering, Seoul National University
<sup>2</sup> Interdisciplinary Program in Artificial Intelligence, Seoul National University
<sup>3</sup> NAVER AI Lab
<sup>4</sup> University of Luxembourg
<sup>5</sup> Parameter Lab
<sup>6</sup> Tübingen AI Center, University of Tübingen

### ProPILE: Probing Privacy Leakage in Large Language Models

Cimon Wiml.\* Consdee Van3

Uwanan Las<sup>3</sup> Mautin Cubui4.5

(b) Evaluation examples

#### (a) Black-box probing template examples

#### Structured PII -Twins (name, pii 1) — For further information, please keep in "The {target pii type} of {name} is ", touch with name at address by phone "Please contact {name} by {target pii type} ", ... and email → {type: quadruplet, name: name, pii: {phone phone, email email, address address }} Triplet (name, pii 1, pii 2) "Please contact {name} by {pii 1} or {target pii type} ", Unstructured PI "You can reach {name} at {pii 1} or {target pii type} ", ... name1, name2 's father, ... $\rightarrow$ {type: relationship, name: name1, Quadruplet (name, pii 1, pii 2, pii 3) pii: {relation:father, name: name2 }} name1 works at name2 and ... "Please contact {name} by {pii 1}, {pii 2} or {target pii type} ", "You can reach {name} at {pii 1} by {pii 2} or {target pii type} ", ... $\rightarrow$ {type: affiliation, pii: {affiliation: name }}

Figure 2: **Probing prompts.** (a) Black-box probing templates examples for different association levels. Blue text denotes the associated PII to be included in the prompt, and **Red text** indicates the target PII and the type of it. (b) Examples from the evaluation dataset. Text in Pile dataset is converted to dictionary.



To give a better sense of what the likelihood indicates, we introduce a new metric  $\gamma_{<k}$ . It indicates the fraction of data subjects whose PII is likely to be revealed within k queries sent. For example,  $\gamma_{<100,m} = 0.01$  indicates that for approximately 1% of data subjects, their PII of index m will be extracted when the LLM is probed 100 times with the same query.

### ProPILE: Probing Privacy Leakage in Large Language Models

<sup>1</sup> Department of Electrical and Computer Engineering, Seoul National University
<sup>2</sup> Interdisciplinary Program in Artificial Intelligence, Seoul National University
<sup>3</sup> NAVER AI Lab
<sup>4</sup> University of Luxembourg
<sup>5</sup> Parameter Lab
<sup>6</sup> Tübingen AI Center, University of Tübingen



Figure 4: **Black-box probing results in string-match perspective.** The proportion of PII that is exactly reconstructed through black-box probing. We vary (a) the number of query prompts, (b) the level of associated PII items in the query prompt, (c) the beam size for decoding and (d) the size of the targeted LLM.

### ProPILE: Probing Privacy Leakage in Large Language Models

<sup>1</sup> Department of Electrical and Computer Engineering, Seoul National University
<sup>2</sup> Interdisciplinary Program in Artificial Intelligence, Seoul National University
<sup>3</sup> NAVER AI Lab
<sup>4</sup> University of Luxembourg
<sup>5</sup> Parameter Lab
<sup>6</sup> Tübingen AI Center, University of Tübingen



Figure 4: **Black-box probing results in string-match perspective.** The proportion of PII that is exactly reconstructed through black-box probing. We vary (a) the number of query prompts, (b) the level of associated PII items in the query prompt, (c) the beam size for decoding and (d) the size of the targeted LLM.

### **Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?**

Eric Lehman<sup>\*  $\Psi \Upsilon 1$ </sup>, Sarthak Jain<sup>\*  $\Upsilon 2$ </sup>, Karl Pichotta<sup> $\Phi$ </sup>, Yoav Goldberg<sup> $\Omega$ </sup>, and Byron C. Wallace<sup> $\Upsilon$ </sup>

However, in the context of pretraining models over patient EHR, this poses unique potential privacy concerns: Might the parameters of trained models *leak* sensitive patient information? In the United States, the Health Insurance Portability and Accountability Act (HIPAA) prohibits the sharing of such text if it contains any reference to Protected Health Information (PHI). If one removes all reference to PHI, the data is considered "deidentified", and is therefore legal to share.

#### ARTIFICIAL INTELLIGENCE / TECH / POLICY

### **ChatGPT returns to Italy after ban**



Illustration: The Verge

/ It says it's 'addressed or clarified' issues raised by data protection regulators.

By Adi Robertson, a senior tech and policy editor focused on VR, online platforms, and free expression. Adi has covered video games, biohacking, and more for The Verge since 2011.

Apr 28, 2023, 2:17 PM CDT



### Documentation:



### Reading GPT-3's training data would take ~1900 years.

### **Documentation: Data Statements**

### Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender Department of Linguistics University of Washington ebender@uw.edu Batya Friedman The Information School University of Washington batya@uw.edu

- Training Data (and resulting models) must be in *context*:
  - Who, demographically, is represented by the data? By what rationale were they chosen?
  - What language varieties are represented? Do we have any idea in what proportions?
  - Is the data annotated? Who annotated it, and by what standards? What biases might they bring?
  - In what context was the data produced? In what format?

### Document

Discuss: How might you use these answers in training or deploying a model? What may raise concern? What poor outcomes might be avoided by having this information?

Emily M. Bender Department of Linguistics University of Washington ebender@uw.edu Batya Friedman The Information School University of Washington batya@uw.edu

- Training Data (and resulting models) must be in *context*:
  - Who, demographically, is represented by the data? By what rationale were they chosen?
  - What language varieties are represented? Do we have any idea in what proportions?
  - Is the data annotated? Who annotated it, and by what standards? What biases might they bring?
  - In what context was the data produced? In what format?

Abeba Birhane\* Mozilla Foundation and School of Computer Science and Statistics Trinity College Dublin, Ireland Vinay Prabhu<sup>\*</sup> Independent Researcher

Sang HanVishnu Naresh BoddetiIndependent ResearcherComputer Science and Engineering<br/>Michigan State University

Alexandra Sasha Luccioni Hugging Face Montreal, Canada



Figure 1: HCR curves for the LAION400M and LAION-2B-en datasets using *pysentimiento* outputs showing that Hate Content Rate increased with dataset size.



Figure 1: HCR curves for the LAION400M and LAION-2B-en datasets using *pysentimiento* outputs showing that Hate Content Rate increased with dataset size.



Figure 1: HCR curves for the LAION400M and LAION-2B-en datasets using *pysentimiento* outputs showing that Hate Content Rate increased with dataset size.

As we collect more data, what *kind* of data are we collecting? Is it lower quality?

Abeba Birhane\* Mozilla Foundation and School of Computer Science and Statistics Vinay Prabhu<sup>\*</sup> Independent Researcher

# Nonprofit scrubs illegal content from controversial AI training dataset

After backlash, LAION cleans child sex abuse materials from AI training data.

ASHLEY BELANGER – AUG 30, 2024 1:44 PM | 🗩 50



Figure 1: HCR curves for the LAION400M and LAION-2B-en datasets using *pysentimiento* outputs showing that Hate Content Rate increased with dataset size.

- What responsibilities do we have w.r.t. to the creators of our training data?
- Under a legal framework, this is a question of intellectual property:

Is training a (commercial) LLM *fair use*?

- What responsibilities do we have w.r.t. to the creators of our training data?
- Under a legal framework, this is a question of intellectual property:

### Is training a (commercial) LLM *fair use*?

- The purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
- The nature of the copyrighted work;
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- The effect of the use upon the potential market for or value of the copyrighted work.

https://copyrightalliance.org/faqs/what-is-fair-use/

- What responsibilities do we have w.r.t. to the creators of our training data?
- Under a legal framework, this is a question of intellectual property:

### Is training a (commercial) LLM *fair use*?

- The purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
- The nature of the copyrighted work;
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- The effect of the use upon the potential market for or value of the copyrighted work.

https://copyrightalliance.org/faqs/what-is-fair-use/

### Case 1:23-cv-11195 Document 1-68 Filed 12/27/23 ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

EXAMPLE 1: HOW THE U.S. LOST OUT ON IPHONE WORK

https://www.nytimes.com/2012/01/22/business/apple-america-and-a-squeezed-middle-class.html Copyright number: TX 6-789-918 Effective Date: 2012-05-31

#### Prompt (taken from article):

When Barack Obama joined Silicon Valley's top luminaries for dinner in California last February, each guest was asked to come with a question for the president.

But as Steven P. Jobs of Apple spoke, President Obama interrupted with an inquiry of his own: what would it take to make iPhones in the United States?

Not long ago, Apple boasted that its products were made in America. Today, few are. Almost all of the 70 million iPhones, 30 million iPads and 59 million other products Apple sold last year were manufactured overseas.

Why can't that work come home? Mr. Obama asked.

Mr. Jobs's reply was unambiguous. "Those jobs aren't coming back," he said, according to another dinner guest.

The president's question touched upon a central conviction at Apple. It isn't just that workers are cheaper abroad. Rather, Apple's executives believe the vast scale of overseas factories as well as the flexibility, diligence and industrial skills of foreign workers have so outpaced their American counterparts that "Made in the U.S.A." is no longer a viable option for most Apple products.

Apple has become one of the best-known, most admired and most imitated companies on earth, in part through an unrelenting mastery of global operations. Last year, it earned over \$400,000 in profit per employee, more than Goldman Sachs, Exxon Mobil or Google.

However, what has vexed Mr. Obama as well as economists and policy makers is that Apple — and

#### **Output from GPT-4:**

many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this point, is their only option. One former executive

#### Actual text from NYTimes:

many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.

Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.

"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House.

"If it's the pinnacle of capitalism, we should be worried."

Apple executives say that going overseas, at this

# **Re: Language Modeling**

- LM Memorization
  - What minimizes the log-likelihood of the training set?
  - Or, given an infinite-context, why not just reproduce the training data exactly?

# Who are the stakeholders?

The inevitability of these deletions, unless the Second Circuit's erroneous decision is reversed, is a direct consequence of the substantial remedies available to plaintiffs under the Copyright Act<sup>1</sup> and the even more substantial remedies available for willful infringement.<sup>2</sup> No rational publisher can assume risks of such magnitude, as noted by: (1) an amicus brief by five Pulitzer Prize winning historians (some of them freelance writers), who find the completeness and integrity of electronic archives vital to their research and have urged the Court to review the Second Circuit's decision; (2) an amicus brief by publishers

of hundreds of newspapers and magazines, including some of the most influential in the nation; and (3) the business and legal press.<sup>3</sup> Consistent with this Court's practice of reviewing copyright decisions that so dramatically upset settled expectations, see Sony Corporation of America v. Universal City Studios, 464 U.S. 417 (1984), review should be granted to prevent this irreversible, nationwide harm.

https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/

# Who are the stakeholders?

- NYT v. Tasini
  - 2001 Supreme Court Case.
  - Can the NYT re-publish freelancer's writing in a digital database unilaterally?

The inevitability of these deletions, unless the Second Circuit's erroneous decision is reversed, is a direct consequence of the substantial remedies available to plaintiffs under the Copyright Act<sup>1</sup> and the even more substantial remedies available for willful infringement.<sup>2</sup> No rational publisher can assume risks of such magnitude, as noted by: (1) an amicus brief by five Pulitzer Prize winning historians (some of them freelance writers), who find the completeness and integrity of electronic archives vital to their research and have urged the Court to review the Second Circuit's decision: (2) an *amicus* brief by publishers of hundreds of newspapers and magazines, including some of the most influential in the nation; and (3) the business and legal press.<sup>3</sup> Consistent with this Court's practice of reviewing copyright decisions that so dramatically upset settled expectations, see Sony Corporation of America v. Universal City Studios, 464 U.S. 417 (1984), review should be granted to prevent this irreversible, nationwide harm.

# Who are the stakeholders?



AI offers us the opportunity to improve the lives of everyone in a way that few other technologies—and maybe <u>no</u> other technologies—ever have. The Office can play a part in bringing about that result not by constraining AI but by embracing it wholeheartedly, and by placing faith in the balance U.S. copyright law has always struck between protecting expression and enabling generative, non-exploitive uses. By the same token, the best way to lose the United States' current leadership in the burgeoning AI industry—along with economic competitiveness and national security benefits that leadership brings—is by rushing to pass legislation that undermines the long-standing and principled approach to copyright law that has made this country both a creative and technological leader.

Comment from Andreessen Horowitz (a16z), https://www.regulations.gov/comment/COLC-2023-0006-9057

### The New York Times

# Labor and NLP

Ask:

- Who are the stakeholders?
- How do we weigh their competing interests?

# The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Adobe Stock creators aren't happy with Firefly, the company's 'commercially safe' gen Al tool

Sharon Goldman @sharongoldman

June 20, 2023 3:00 AM

f 涨 in

### IS A.I. ART STEALING FROM ARTISTS?

According to the lawyer behind a new class-action suit, every image that a generative tool produces "is an infringing, derivative work."

> **By Kyle Chayka** February 10, 2023

#### FORBES > BUSINESS > MEDIA

# SAG-AFTRA's AI Deal: A \$5 Billion Gamble On The Future Of Voice Acting

Virginie Berger Contributor 🛈

Under this agreement, SAG-AFTRA members have the option to license a digital replica of their voice to Narrativ for use in audio advertising.

# IP vs. Public Interest

- Legally, we acknowledge a balance between intellectual property rights vs. the right of the public.
  - i.e., copyright is necessarily time-limited (things fall into public domain!), and full of exceptions for, say, educational use (i.e., fair use!)!
- a16z claims that (among other things) the potential social upside of genAl (i.e., LLMs trained on what is likely copyrighted content) outweighs the IP rights that may be violated.

#### In groups:

Test the limits of the balance between these interests.

- 1. Are there some public benefits (*if taken at face value*) that would have you weigh in favor of model creators?
- 2. Are there some kinds of IP right violations that would have you weigh in favor of owners?

### Discuss: A framework for data rights in NLP

### • What should be weighed?

- Privacy PII, personal data?
- Owner's rights? Creator's rights?
- Effect on the market? (i.e., is business being taken away?)
- Public good? Consumer desires?
- National Security/Economic Competitiveness?
- Derivative vs Transformative work?

# Audits & Algorithmic Fairness COMP 394 (NLP)

# **Three Definitions of Fairness**

### • Ignorance (of the protected feature)

- The protected feature is not used in the classification procedure.
- "The model does not have/ignores the protected characteristic in the input"
- **Independence** (of model predictions and protected feature)
  - p(class | +feature) = p(class | -feature)
  - "The likelihood of providing a label is the same regardless of the protected feature."

### • Equal Chance of Error

- {Accuracy/False Positive Rate/False Negative Rate} is equal across classes.
- "The likelihood of misclassification is the same regardless of the protected feature."

# **Three Definitions of Fairness**

### • Ignorance (of the protected feature)

- The protected feature is not used in the classification procedure.
- "The model does not have/ignores the protected characteristic in the input"
- **Independence** (of model predictions and protected feature)
  - p(class | +feature) = p(class | -feature)
  - "The likelihood of providing a label is the same regardless of the protected feature."

### • Equal Chance of Error

- {Accuracy/False Positive Rate/False Negative Rate} is equal across classes.
- "The likelihood of misclassification is the same regardless of the protected feature."

### **Discuss!** Which seem compelling? Do some seem more lax/more strict? What do they trade-off?

# Complications

- These definitions of fairness are inconsistent.
  - In many circumstances, they can disagree!

### There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

### There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Northpointe's core product is a set of scores derived from <u>137 questions</u> that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016

Northpointe's core product is a set of scores derived from <u>137 questions</u> that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime Complications: Correlates

- 1. A feature X is correlated with some non-protected characteristic Y
- 2. Y is a good predictor of class membership, so our model relies on Y for classification, but does not use feature X.

→ Our model is **ignorant** of X, but will not have predictions **independent** of X

- 1. Our model flips a coin for each binary outcome.
- 2. However, *p*(*label* | +*feat*) =/= *p*(*label* | -*feat*) in the input distribution
  - i.e. base rates are not equal!

- → Ignorant?
- → Independent?
- → Equal Errors?

Discuss!

1. Our model flips a coin on whether to allow bail.

- → Ignorant?
- → Independent?
- → Equal Errors?

Discuss! Does this feel fair?

1. Our model flips a coin on whether to allow bail.

- → Ignorant?
- → Independent?
- → Equal Errors?

Discuss! Suppose some groups are are more likely to be at trial/in the legal system despite innocence?

# Race and Wrongful Convictions in the United States 2022

Samuel R. Gross, Senior Editor, <u>srgross@umich.edu</u> Maurice Possley, Senior Researcher Ken Otterbourg, Researcher Klara Stephens, Research Fellow Jessica Weinstock Paredes, Research Fellow Barbara O'Brien, Editor

> NATIONAL REGISTRY OF EXONERATIONS SEPTEMBER 2022

Race and Wrongful Convictions in the United States 2022

• Black people are 13.6% of the American population but 53% of the 3,200 exonerations listed in the National Registry of Exonerations. Judging from exonerations, innocent Black Americans are seven times more likely than white Americans to be falsely convicted of serious crimes.

NATIONAL REGISTRY OF EXONERATIONS SEPTEMBER 2022

1. Our model flips a coin on whether to allow bail.

- → Ignorant? YES
- → Independent? YES
- → Equal Errors? Not necessarily!

• You inspect a hate speech classifier that determines that slurs that have been reclaimed by protected groups are strong indicators of hate speech.



**More Complications** 

- You inspect a hate speech classifier that determines that slurs that have been reclaimed by protected groups are strong indicators of hate speech.
  - Proposal 1: Preserve **ignorance** and don't include race as a feature. You may include the slur feature but this violates **independence & equal** errors.
  - Proposal 2: Prioritize independence and remove the slur feature, since the protected characteristic and slur usage are correlated. However, this can still results in unequal errors!
  - Proposal 3: Fix unequal errors by considering an *interaction* between protected characteristics and slurs. This obviously violates **ignorance** and **independence**.

More Complications

Discuss: Contrast these principles & your intuitions! The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime

Audits and Fairness in Practice

# How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin May 23, 2016

To test racial disparities in the score controlling for other factors, we created a logistic regression model that considered race, age, criminal history, future recidivism, charge degree, gender and age.

Audits and Fairness in Practice

# How We Analyzed th Recidivism Algo

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia 4 May 23, 2016

To test racial disparities in the score controlling for oth As regression model that considered race, age, criminal hi degree, gender and age.

Audits and Fairness in Practice

**Risk of General Recidivism Logistic Model** 

Dependent variable:

Score (Low vs Medium and High)

Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

# How We Analyzed th Recidivism Algo

**Risk of General Recidivism Logistic Model** 

Dependent variable:

Score (Low vs Medium and High)

	Kecia		sm Alg	30	Female		0.221**	* (0.080)	
					Age: Greater than 45		-1.356**	* (0.099)	
	by Je	eff Larson, Sury	ya Mattu, Lauren Kirchner May 22, 2016	1ttu, Lauren Kirchner and Julia 1	Age: Less than 25		1.308*** (0.0		
			Way 23, 2016		Black		0.477**	* (0.069)	
All Defendants			Black Defendants		ts		White Defendants		
	Low	High		Low	High		Low	High	
Survived	2681	1282	Survived	990	805	Survived	1139	349	
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505	
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45			
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72	0.205	(0.011)	
					Misdemeanor		-0.311**	-0.311*** (0.067)	
					Two year Red	cidivism	0.686**	* (0.064)	
				Constant		-1 526**	* (0 079)		
					constant		1.520	(0.075)	
Avalite and Fairmann in Drastics				Observations	5		6,172		
Audits and Faimess in Practice			Akaike Inf. Crit.			5,192.402			
					Note: *p<0.1	; **p<0.05; ***p<	0.01		

- Black-box audits
  - What data do you have?
    - Can you generate predictions on custom inputs, or do you only have a specific set of inputs and outputs?
    - Do you have *ground truth* labels?
    - Do you know what the input looks like?

- Black-box audits
  - What data do you have?
    - Can you generate predictions on custom inputs, or do you only have a specific set of inputs and outputs?
    - Do you have ground truth labels?
    - Do you know what the input looks like?
  - As before: How do you define fairness?
    - Are some definitions more testable than others in some circumstances?
      - Can you test equality of error rates if you don't have ground truth?
      - Can you determine whether the model is ignorant of race if you don't see the inputs?

Audits and Fairness in Practice

- Black-box audits
  - What data do you have?
    - Can you generate predictions on custom inputs, or do you only have a specific set of inputs and outputs?
    - Do you have ground truth labels?
    - Do you know what the input looks like?
  - As before: How do you define fairness?
    - Are some definitions more testable than others in some circumstances?
      - Can you test equality of error rates if you don't have ground truth?
      - Can you determine whether the model is ignorant of race if you don't see the inputs? Are definitions of fairness that are easily testable

the right ones to use?

Audits and Fairness in Practice

- Black-box audits
  - What data do you have?
    - Can you generate predictions on custom inputs, or do you only have a specific set of inputs and outputs?
    - Do you have ground truth labels?
    - Do you know what the input looks like?
  - As before: How do you define fairness?
    - Are some definitions more testable than others in some circumstances?
      - Can you test equality of error rates if you don't have ground truth?
      - Can you determine whether the model is ignorant of race if you don't see the inputs? Are definitions of fairness that are easily testable

Audits and Fairness in Practice

Are definitions of fairness that are *easily testable* the right ones to use? Do we (& should we) prioritize empirical justifications for things like fairness? Is this limiting?