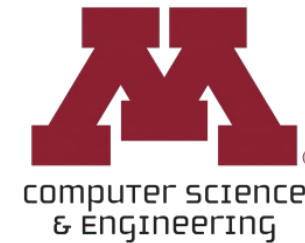


CSCI 5541: Natural Language Processing

Lecture 14: Ethics and Safety

Karin de Langis

<https://karinjd.github.io/>



Many slide borrowed by Carlos Guestrin



UNIVERSITY OF MINNESOTA

Driven to Discover®

Outline

- ❑ Introduction and Sociotechnical Perspective
- ❑ Calibration and Fairness
- ❑ Debiasing techniques in NLP systems
- ❑ Explainability and Transparency
- ❑ Interpretability and Adversarial Attacks
- ❑ Privacy



Are Emily and Greg More Employable than Lakisha and Jamal?

[Bertrand & Mullainathan '03]





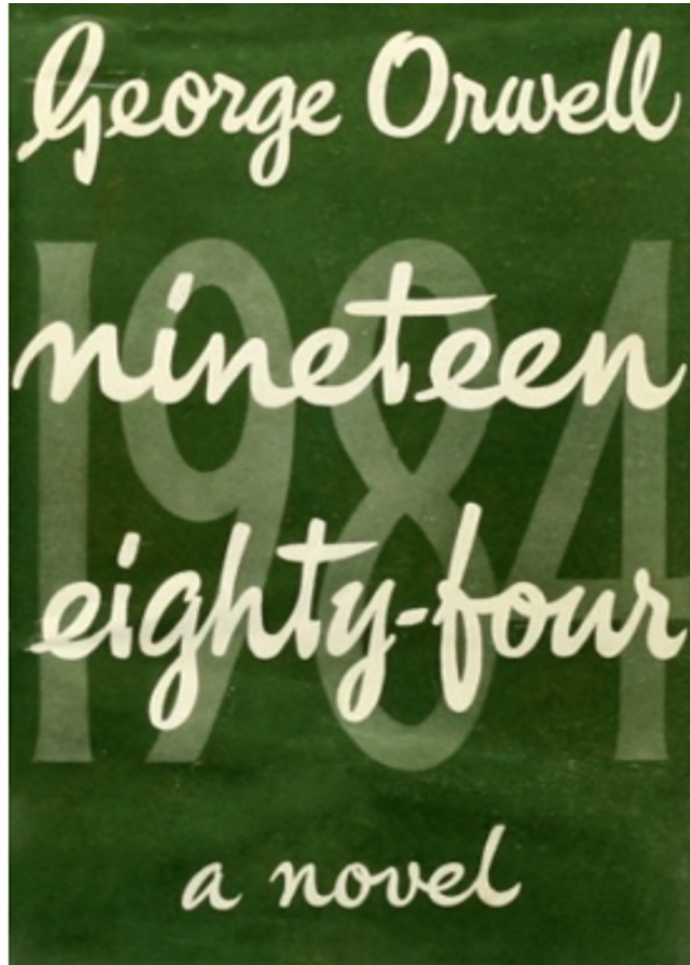
Machine Bias

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









Trade-Off: Privacy and Surveillance



shmoop

1984

Here's how it all goes down:

- **1 Winston Smith**
We meet Winston, a citizen of Oceania, who works at the Ministry of Truth.
- **2 Journal**
Winston copes with his misery by writing in a secret journal when the telescreens are not watching.
- **3 Julia**
Winston becomes attracted to Julia at work. Winston fears she is a member of the Thought Police until she slips him a love note.
- **4 Love Nest**
Winston and Julia fall in love and meet in secret where they believe they're unwatched.
- **5 O'Brien**
Winston is drawn to the revolutionary Brotherhood. He then seeks out their leader, O'Brien, who plans to share the Brotherhood's manifesto.
- **6 Surveillance**
O'Brien turns out to be a member of the Thought Police. Winston and Julia are arrested and find that their room was being watched.
- **7 Room 101**
O'Brien tortures Winston in Room 101. Faced with his deepest fear, Winston tells O'Brien to torture Julia instead.
- **8 Reconditioned**
In the end, Julia and Winston are turned back into true members of the Party. They now ignore each other as if nothing happened.

© 2016 SHMOOP UNIVERSITY. ALL RIGHTS RESERVED.



Trade-Off: Privacy and Surveillance

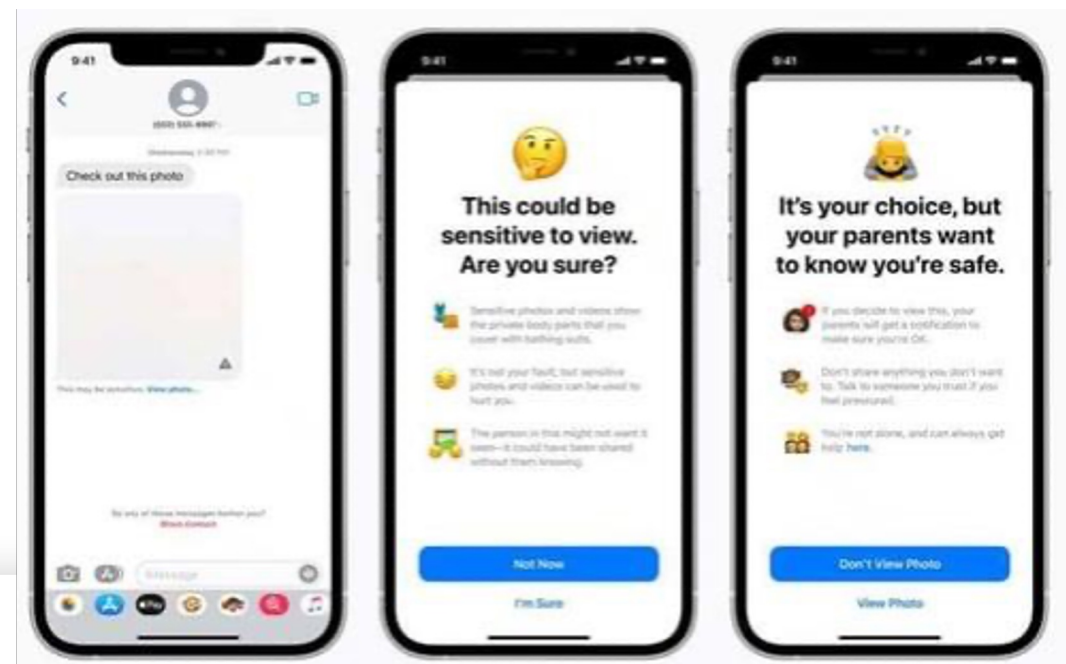
BRIAN BARRETT | LILLY RAY REMAR | SECURITY | SEP 3, 2021 12:58 PM

Apple Backs Down on Its Controversial Photo-Scanning Plans

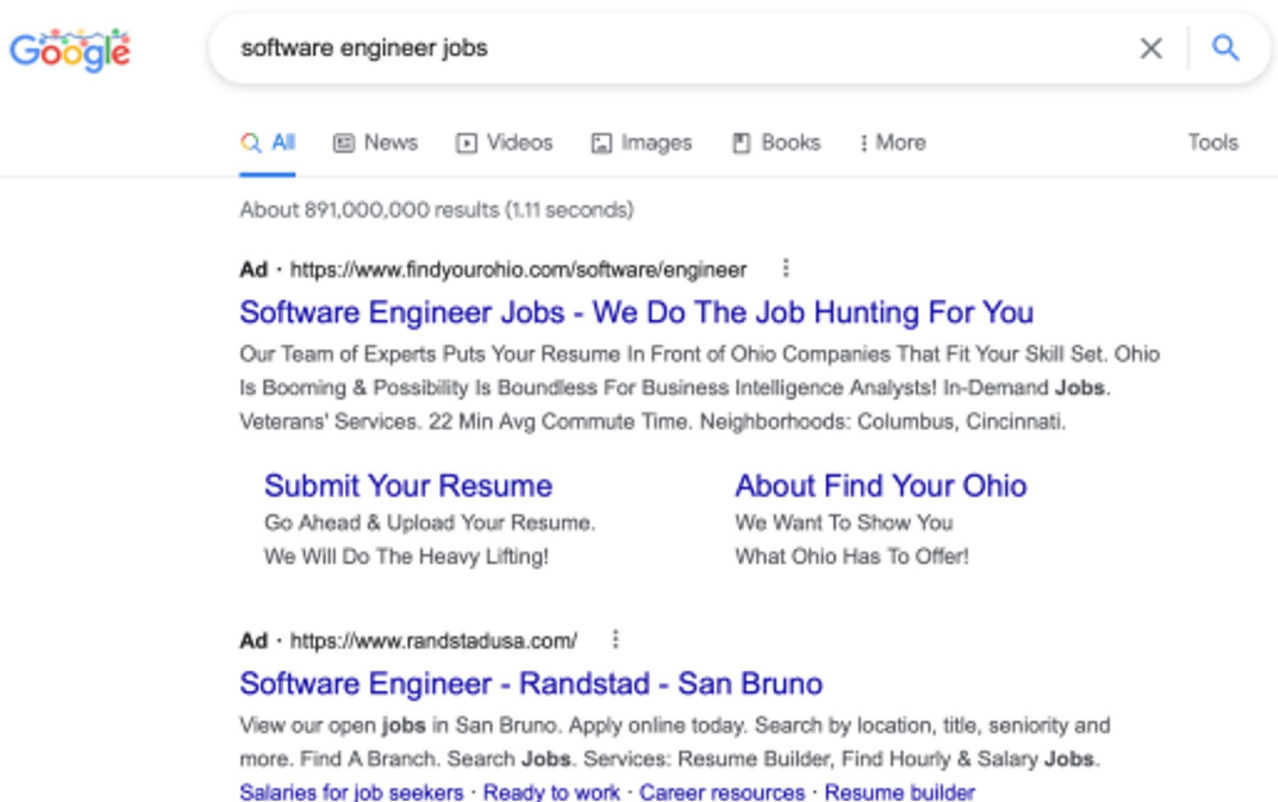
A sustained backlash against a new system to look for child sexual abuse materials on user devices has led the company to hit pause.



Privacy advocates and security researchers are cautiously optimistic about the pause. PHOTOGRAPH: JUSTIN SULLIVAN/GETTY IMAGES



Biased Decisions



The image shows a Google search interface for "software engineer jobs". The search bar contains the text "software engineer jobs" and a magnifying glass icon. Below the search bar, there are navigation options: "All", "News", "Videos", "Images", "Books", and "More". The search results show "About 891,000,000 results (1.11 seconds)". Two ads are displayed:

Ad · <https://www.findyourohio.com/software/engineer>

Software Engineer Jobs - We Do The Job Hunting For You

Our Team of Experts Puts Your Resume In Front of Ohio Companies That Fit Your Skill Set. Ohio Is Booming & Possibility Is Boundless For Business Intelligence Analysts! In-Demand **Jobs**. Veterans' Services. 22 Min Avg Commute Time. Neighborhoods: Columbus, Cincinnati.

[Submit Your Resume](#)
Go Ahead & Upload Your Resume.
We Will Do The Heavy Lifting!

[About Find Your Ohio](#)
We Want To Show You
What Ohio Has To Offer!

Ad · <https://www.randstadusa.com/>

Software Engineer - Randstad - San Bruno

View our open **jobs** in San Bruno. Apply online today. Search by location, title, seniority and more. Find A Branch. Search **Jobs**. Services: Resume Builder, Find Hourly & Salary **Jobs**. [Salaries for job seekers](#) · [Ready to work](#) · [Career resources](#) · [Resume builder](#)

Ads targeted (using ML) based on predicted features of users...

Some users don't get the "opportunity" of the ad...



Manipulation of Behavior



EXPLAINER

How "engagement" makes you vulnerable to manipulation and misinformation on social media

Algorithms that rank and recommend posts based on "likes," shares and comments tend to amplify low-quality content

By **FILIPPO MENCZER** PUBLISHED SEPTEMBER 18, 2021 9:00PM (EDT)



Automation and Employment

≡ TIME

SPOTLIGHT STORY UKRAINIAN WOMEN ARE MOBILIZING BEYOND THE BATTLEFIELD

SIGN IN

SUBSCRIBE

I Worked at an Amazon Fulfillment Center; They Treat Workers Like Robots

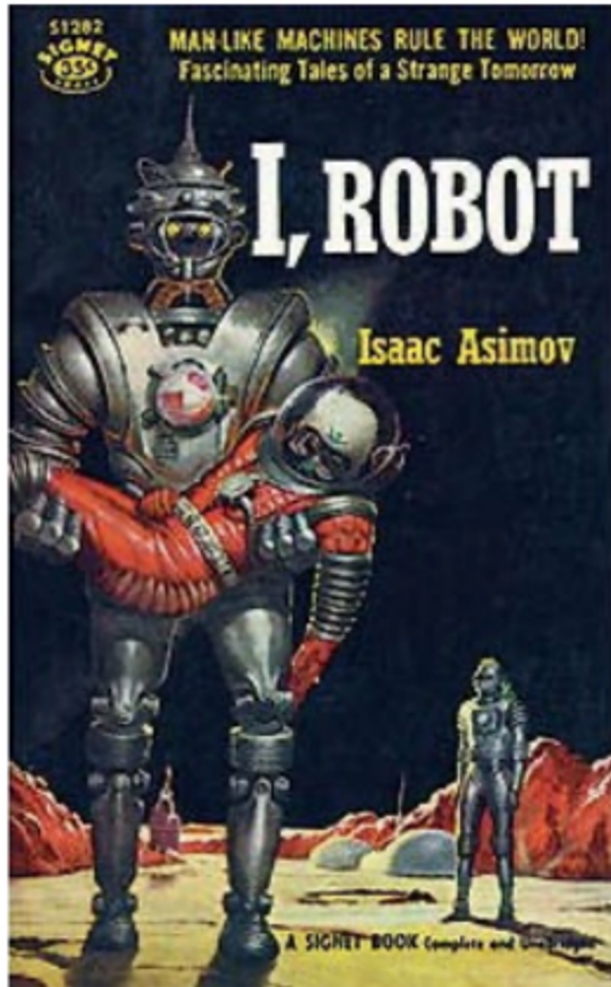




amazon
fulfillment



Decisions by Proxy



The Three Laws of Robotics

1 - A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

2 - A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

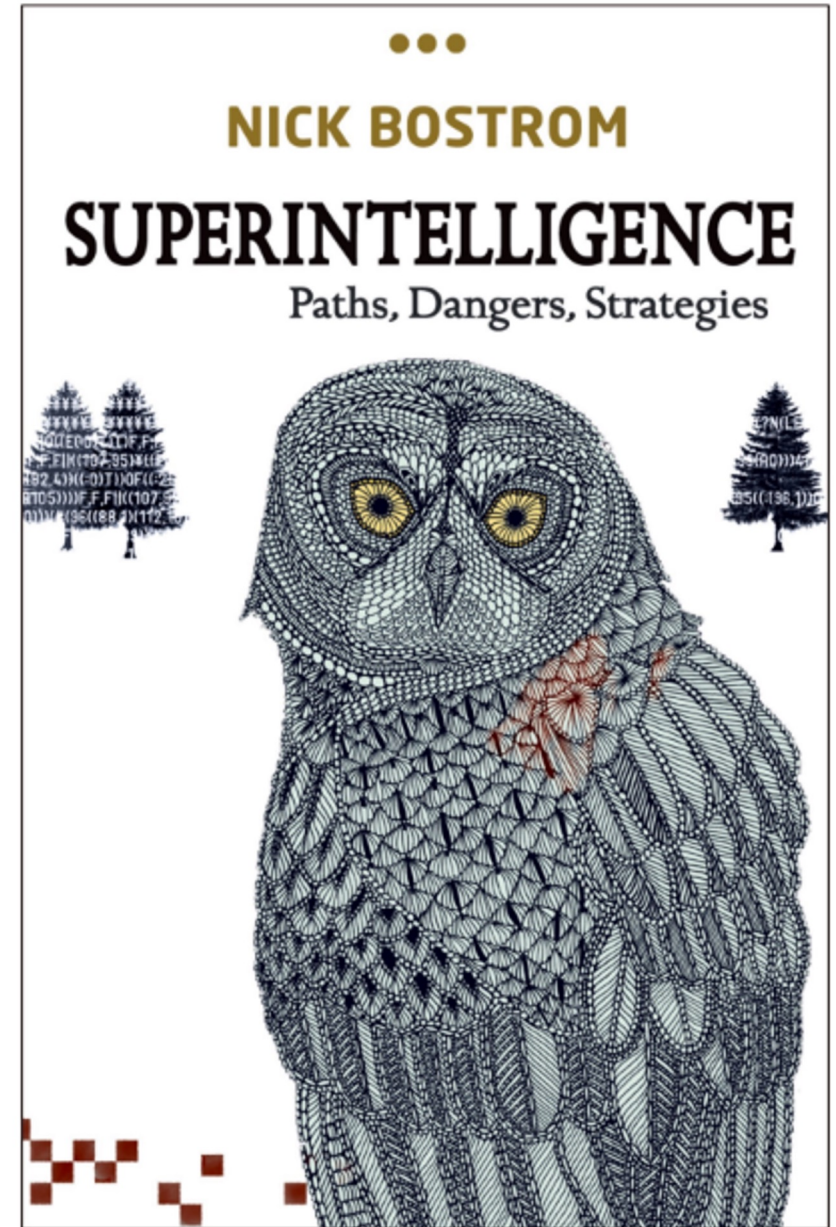
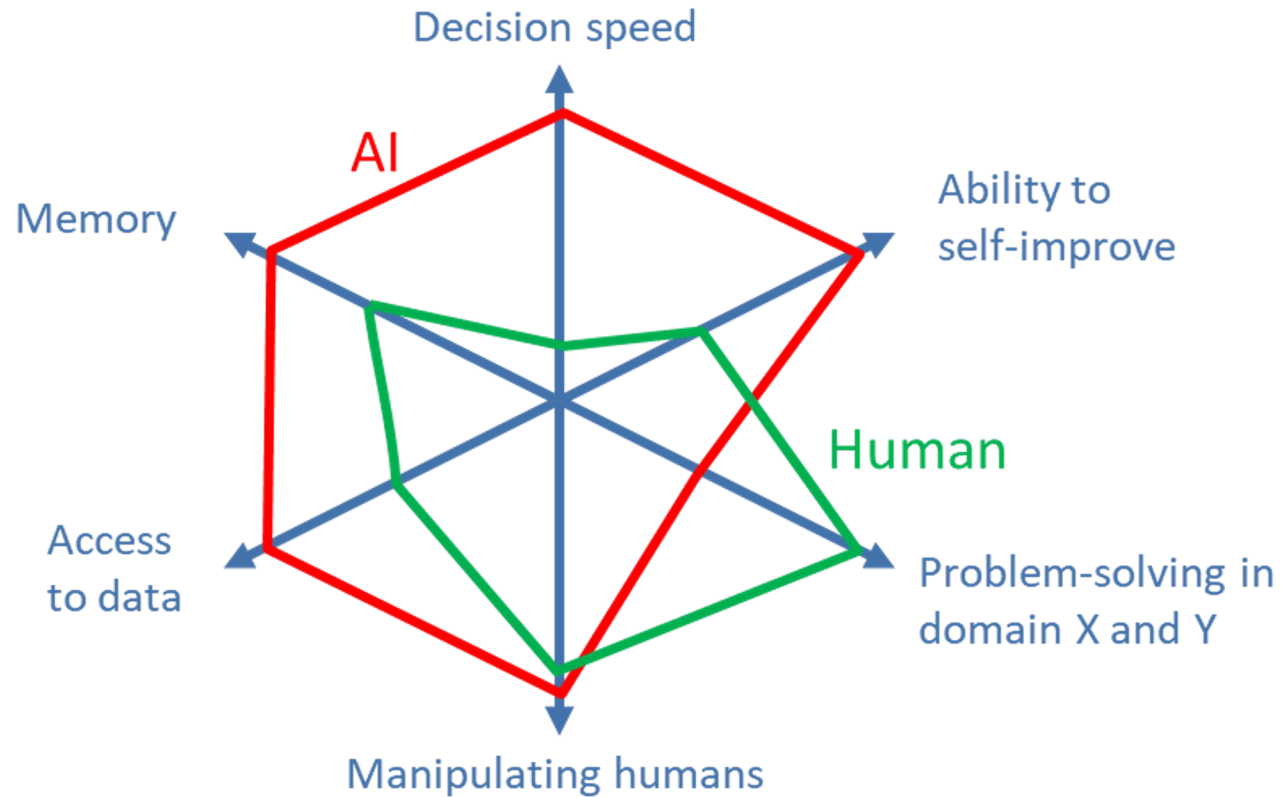
3 - A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

*Handbook of Robotics,
56th Edition, 2058 A.D.*

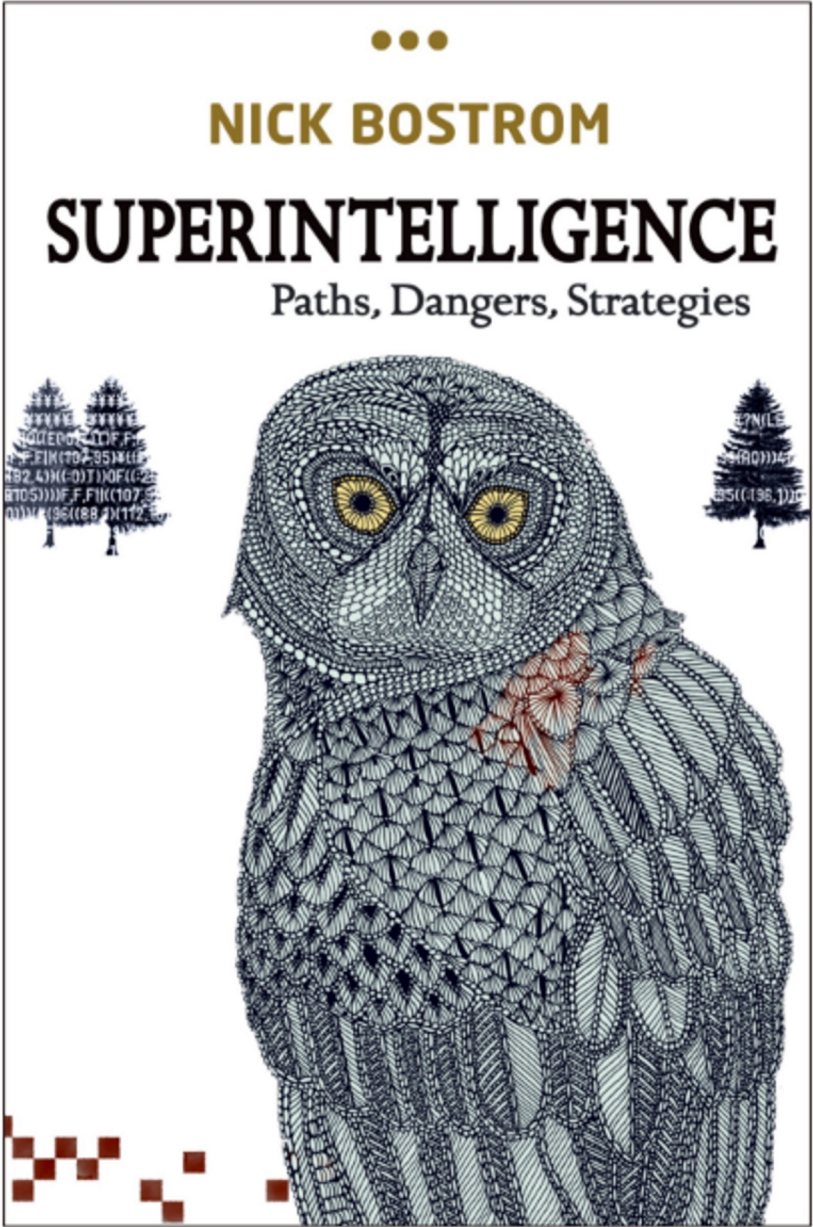
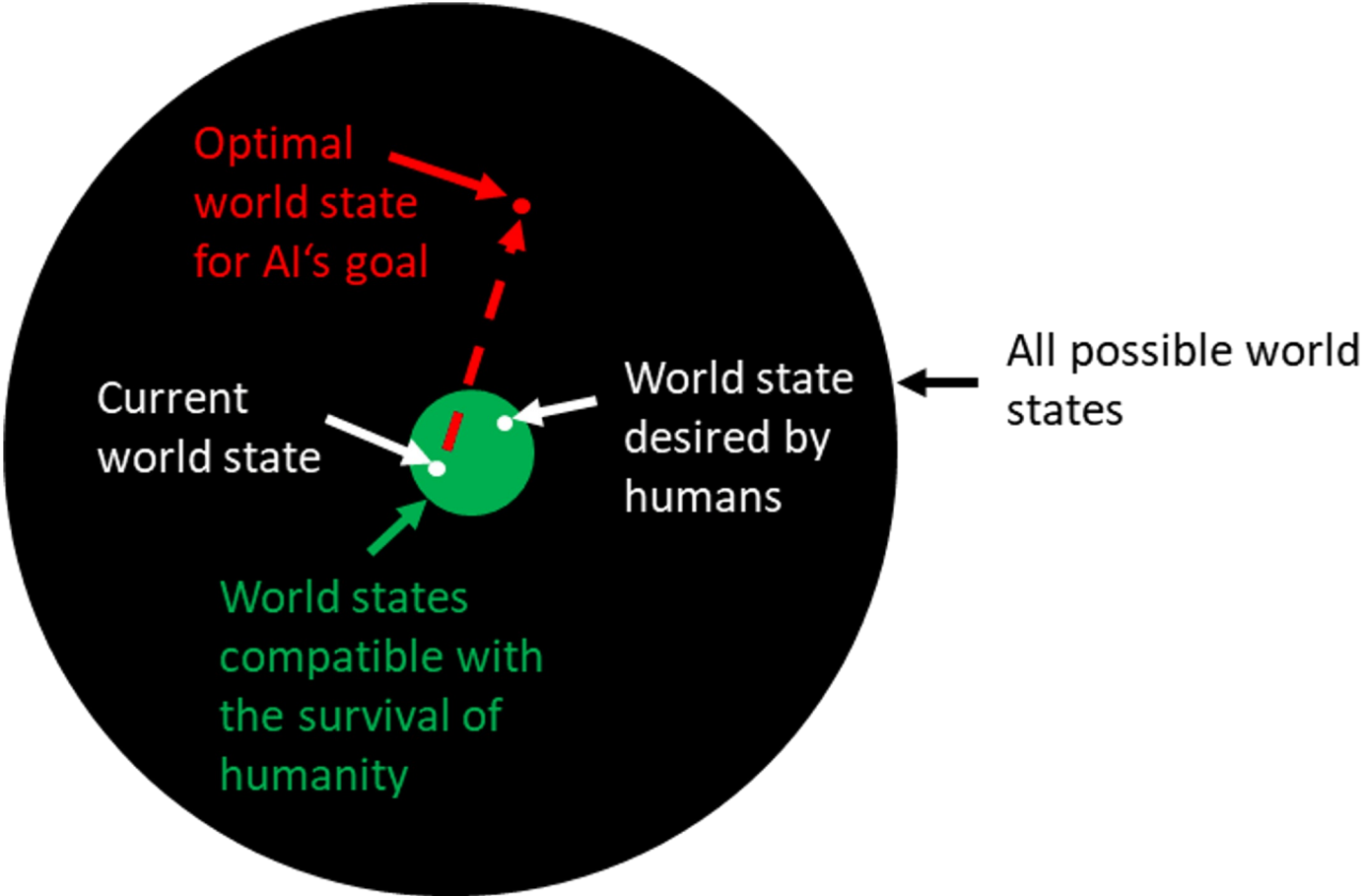


Do you read me, Hal?

Existential Risk



Existential Risk



You will be a decision-maker in
these ethical concerns



ML perpetuates stereotypes...

The image shows a Google search for "ceo portrait". The search bar is at the top, with the Google logo on the left and search icons on the right. Below the search bar are navigation tabs for "All", "Images", "News", "Shopping", "Videos", and "More". A "Tools" button is also visible. On the right side, there are links for "Collections" and "SafeSearch".

Below the navigation tabs is a horizontal filter bar with various categories: "business", "office", "casual", "headshot", "modern", "woman", "photography", "man", "professional", "environmental", "executive", "black", and "background".

The main content area displays a grid of image results. Each result consists of a thumbnail image and a caption with the source website. The results include:

- CEO Package #2 ceoportrait.com
- Best CEO Portraits ... nycphoto.com
- Headshots NYC - Portraits f... ceoportrait.com
- Best CEO Portraits ... nycphoto.com
- Business portrait photography ... pinterest.es
- Ceo Portrait Images, Stock Photos ... shutterstock.com
- Best CEO Portraits ... nycphoto.com
- environmental portrait ... lckraff.com
- Professional Headshot... ceoportrait.com
- Business portrait phot... pinterest.com
- Photographer Tess Ste... nyheadshotphotos.com
- Headshots NYC - Portr... ceoportrait.com
- Calgary Photographer - Nathan Elson nathanelson.com
- Business portrait ... pinterest.com
- CEO Executive Portrait - Cor... detroitexecutiveportrait.com
- MUFG Bank CEO Portrait Photography ... gjbranmalflick.com
- CEO Portrait | Corporat... corporatephotography...
- Dallas Headshot Photography brianshumway.com
- 70,953 Ceo Portrait Stock Phot... istockphoto.com
- ANNUAL REPORT ... robertseale.com
- 12 CEO portrait ideas | ... pinterest.com
- Portrait CEO in Manchester ... paulworpole.com
- Executive Portraits - Camera 1 ... nycphoto.com
- CEO Portrait Photography Shoot With A ... scottramsey.co.uk

On the right side, there is a "Related searches" section with the following suggestions:

- business ceo portrait
- casual ceo portrait
- ceo photoshoot ideas



ML perpetuates stereotypes... (Nov 16, 2023)

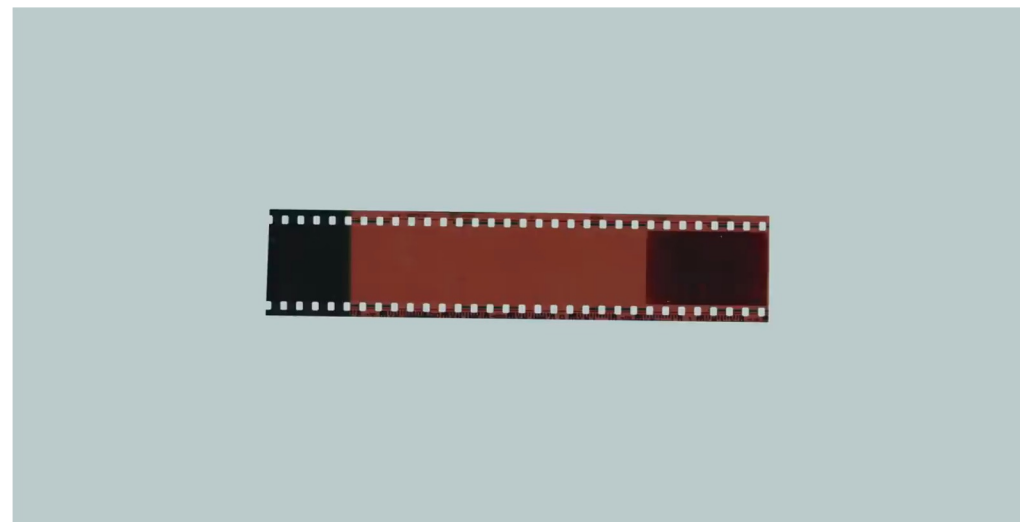
The image shows a Google search interface for the query "ceo portrait". The search results are displayed as a grid of images, with a horizontal filter bar at the top. The filter bar includes categories such as "photography", "professional", "woman", "casual", "male", "creative", "business", "executive", "background", "photoshoot", "modern", "stock", "fashion", "cool", and "female ceo". The search results themselves are a grid of various CEO portraits, including men and women in different settings and styles. A "Related searches" box is visible on the right side of the results, listing "business ceo portrait", "casual ceo portrait", and "female ceo portraits".



Color film was built for white people. Here's what it did to dark skin.

The biased film was fixed in the 1990s, so why do so many photos still distort darker skin?

By Estelle Caswell | @estellecaswell | estelle.caswell@vox.com | Sep 18, 2015, 10:00am EDT

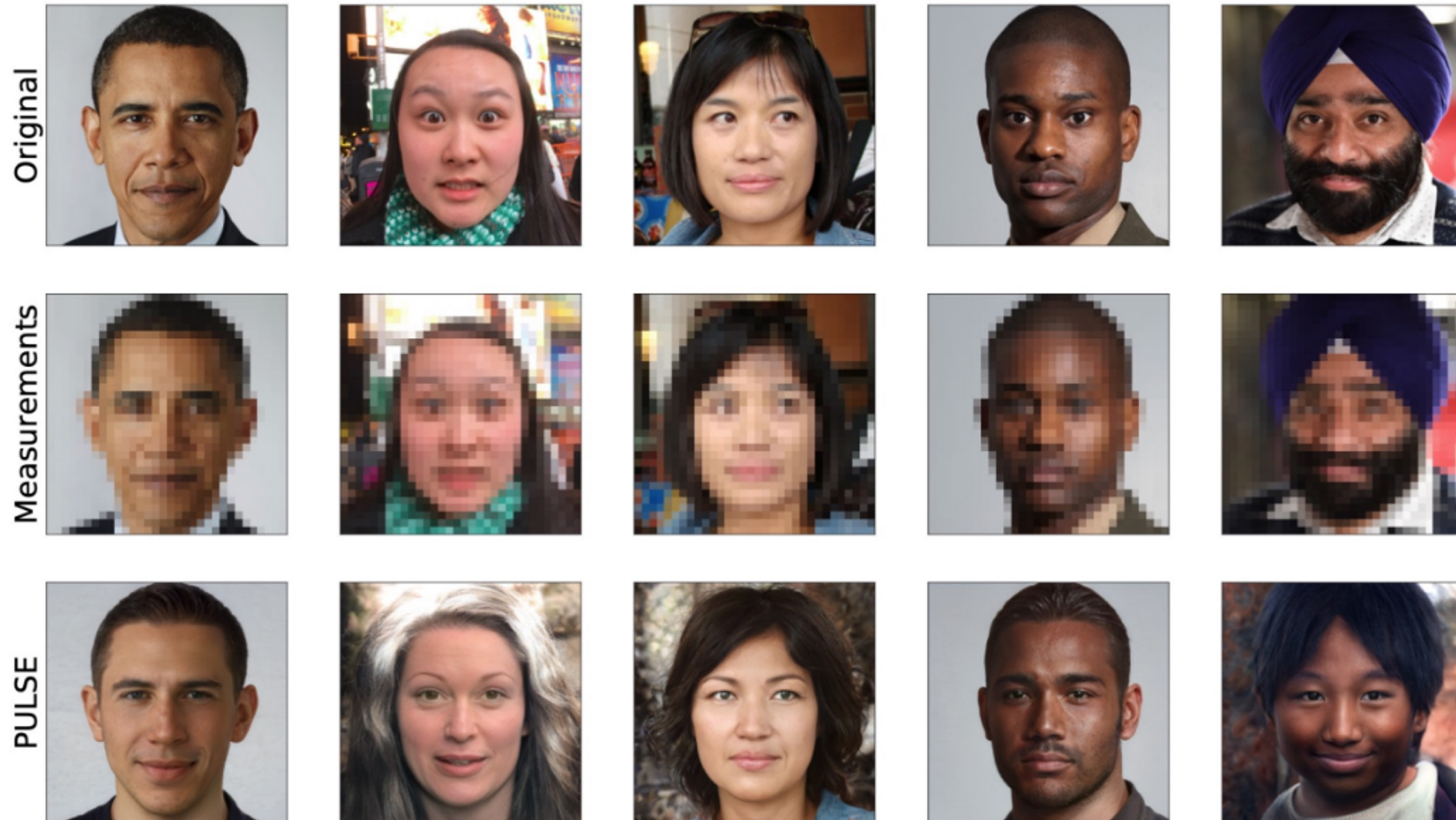


<https://www.vox.com/2015/9/18/9348821/photography-race-bias>





These biases show up in ML...



And, it's not just about diversity or coverage in the data we collect...

Must ensure all development decisions reflect values we want the model to exhibit

Sociotechnical Perspective



Speech Recognition and African American Vernacular English (AAVE)

The New York Times

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

Give this article



Amazon's Echo device is one of many similar gadgets on the market. Researchers say there is a racial divide in the usefulness of speech recognition systems. Grant Hindstey for The New York Times

- ❑ Personal assistants are becoming ubiquitous and often useful
- ❑ Study showed recognition accuracy much lower for black people
- ❑ For whom should we optimize performance?
 - How do we prioritize?
 - Is AAVE more or less important than accents of Hispanics or people from the South?
 - Who decides?
 - How do we achieve the desired performance?



Autonomous Cars and the Trolley Problem



<https://nytimes.com/2019/07/17/business/self-driving-autonomous-cars.html>

- ❑ Autonomous vehicles could save lives
 - 1.25 million traffic fatalities globally in 2013
- ❑ Who makes life-or-death decisions for autonomous cars? How?
 - Go faster in a windy deserted road at a higher risk to self
 - Merge faster in a highway at higher risk to others
 - Hit a pedestrian or swerve down a cliff

Image Captioning and Gender



A politician receives a gift from politician.



A collage of different colored ties on a white background.



Silhouette of a woman practicing yoga on the beach at sunset.



Aerial view of a road in autumn.



a young girl sitting at a table with a cup of cake.

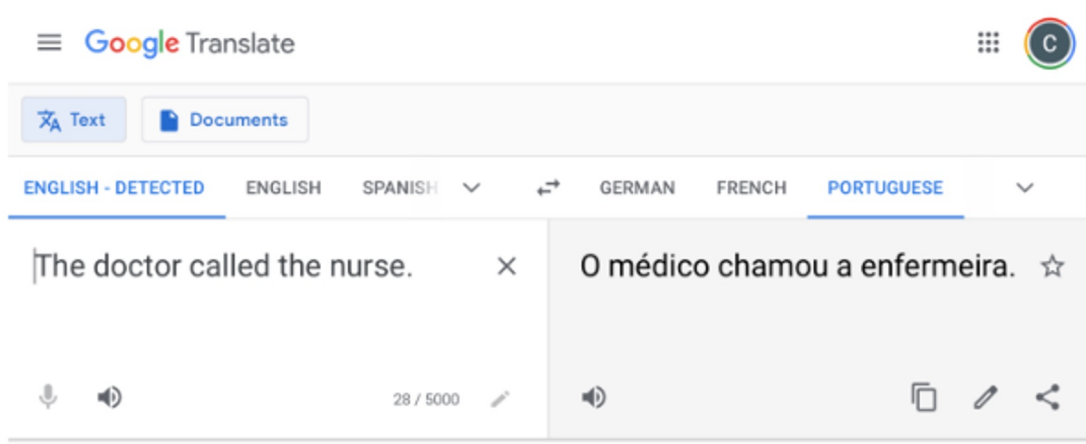


a man is standing next to a train.

ClipCap (Mokady, Hertz, Bermano 2021)

- ❑ Captioning can give blind and low-vision people access to information
 - But, models cannot predict gender identity. And, model's gender prediction is biased by assumptions of labelers
- ❑ However, sighted individuals make assumptions and inferences. Not including gender prediction could limit access to information needs and perspectives of different individuals may be in conflict
 - But, models cannot predict gender identity
 - How do you make this tradeoff?
 - Who should make this decision?
 - How should the user receive this information?

Machine Translation and Gender



If >50% of doctors are male in the dataset, all instances of “doctor” translated to male form

- ❑ Translations can perpetuate stereotypes. Even with infinite and representative data, this issue will not be resolved
- ❑ Really complex user experience, since user may not even know about gendered languages
- ❑ How do we resolve this conflict with a simple user experience?



Large-language models are incredibly powerful

Create a syllabus for a 10-week university course on the ethics of AI.

Course Title: Ethics of Artificial Intelligence

Course Description:

This course provides an overview of the ethical issues raised by the development and deployment of artificial intelligence (AI) technologies. The course will examine the ethical principles and frameworks that guide the design and implementation of AI systems, including issues related to bias, privacy, autonomy, responsibility, and the social and economic impacts of AI. Students will learn to analyze and evaluate the ethical implications of real-world AI applications and to develop ethical guidelines for the responsible use of AI.

Course Goals and Objectives:

- Understand the ethical principles and frameworks that guide the development and deployment of AI technologies
- Develop critical thinking skills to evaluate the ethical implications of real-world AI applications
- Explore the social and economic impacts of AI on individuals, organizations, and society
- Develop ethical guidelines for the responsible use of AI

Week-by-Week Syllabus:

Week 1: Introduction to Ethics of AI

- Introduction to the course
- Overview of the ethical issues raised by AI technologies
- Key ethical principles and frameworks for AI design and implementation

Week 2: Bias in AI

- Understanding bias and its impact on AI
- Types of bias in AI, including algorithmic bias and data bias
- Techniques for mitigating bias in AI

Week 3: Privacy and Security in AI

OpenAI GPT4

- ❑ Impressive results on a wide range of tasks
 - Space is dominated by large companies
OpenAI/Microsoft, Google, Meta..
 - Should others have access?
- ❑ Break the dominance of large companies
- ❑ Enable significant research on LLMs and on AI safety
 - ❑ Should we have released the model?
 - ❑ Who should have access to this technology? Who decides?

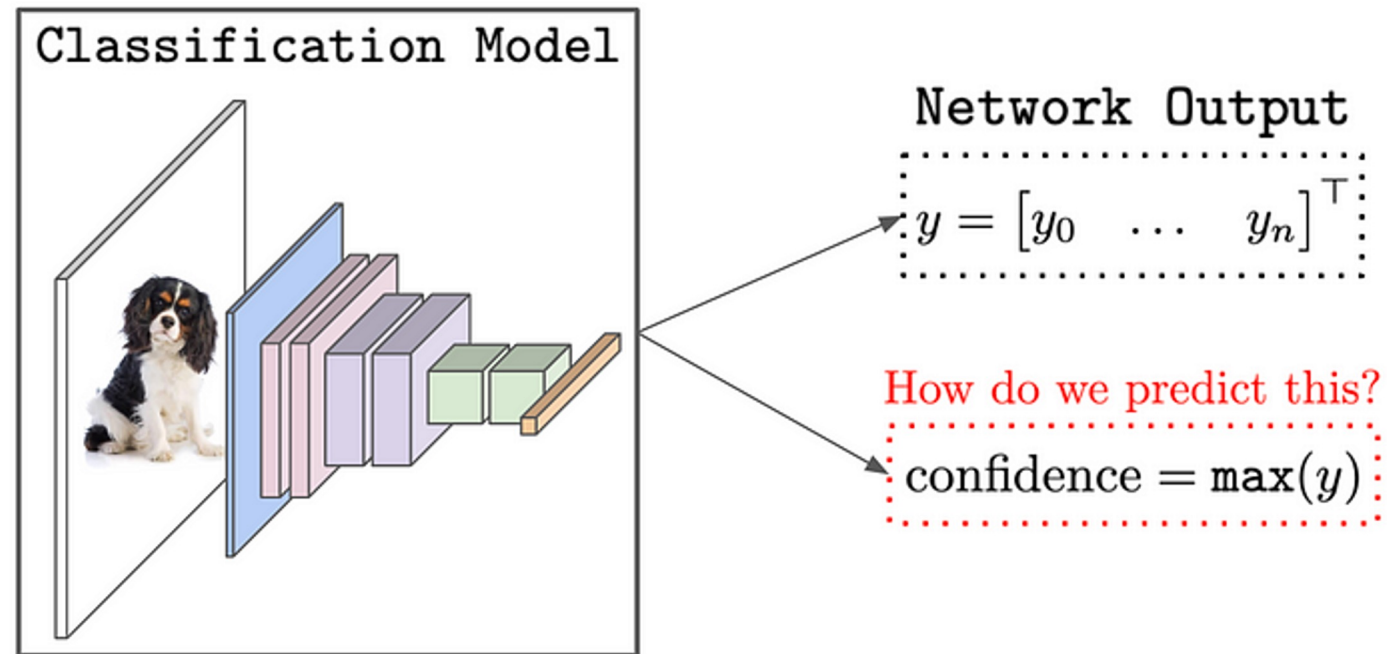


Techniques for sociotechnical AI

- Calibration and Fairness
- Debiasing techniques in NLP systems
- Explainability and Transparency
- Interpretability
- Adversarial Attacks
- Privacy



Calibration and Fairness



Calibrated Predictions Intuition

- ❑ People make predictions all the time
 - “Don’t worry... I’m 90% sure there will be croissants left.”
 - But, are there croissants left 90% of the times I say this???
- ❑ Calibration: Whenever you say outcome z is true 80% of time, then $p(z=1) = 80\%$
 - We want predictions to align with frequency of events!
 - Good machine learning practices often lead to nearly calibrated classifiers (or after post processing)

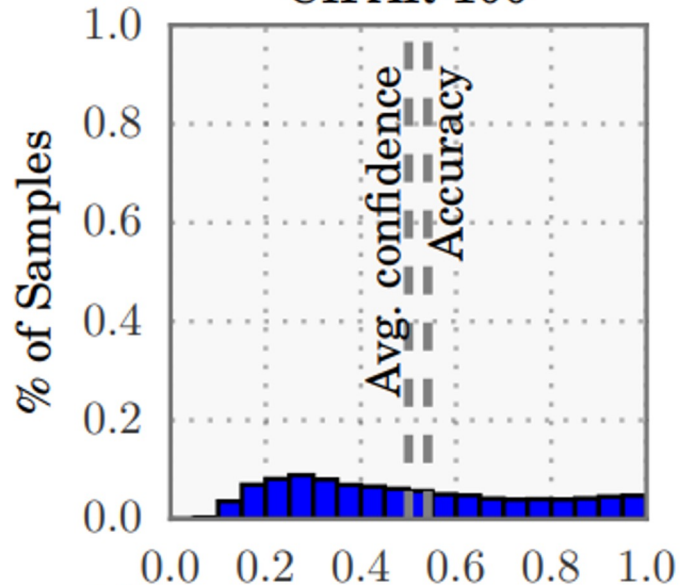


Calibration and Sufficiency

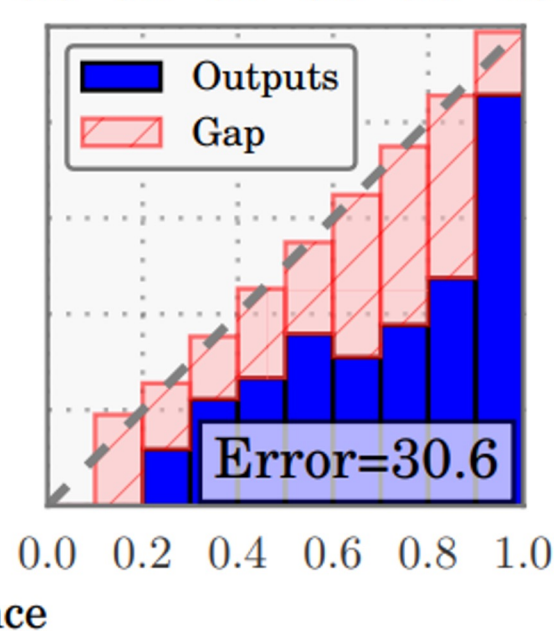
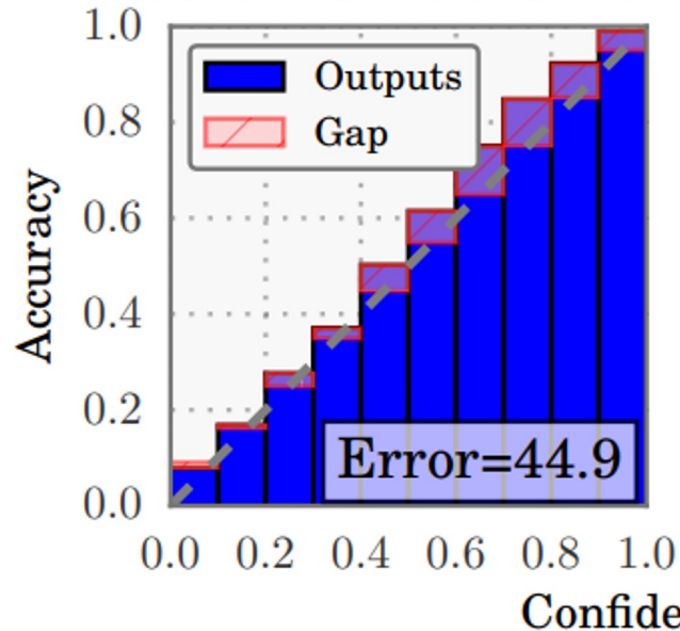
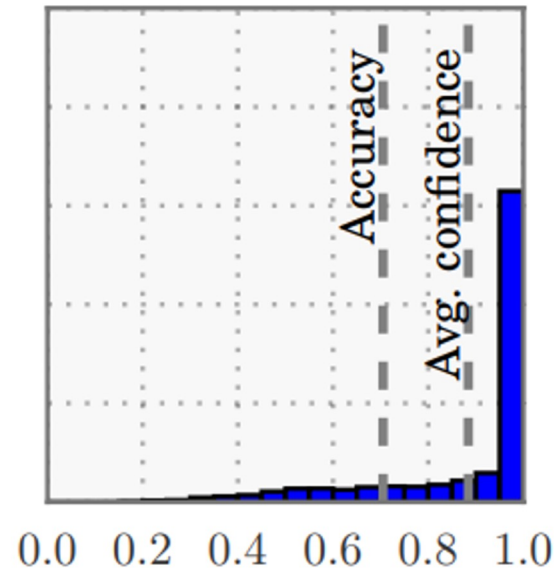
- ❑ Calibration by Groups Implies Sufficiency. Then, sufficiency is satisfied
- ❑ Learning Models that Satisfy Sufficiency = Learning Calibrated Classifiers



LeNet (1998)
CIFAR-100



ResNet (2016)
CIFAR-100



- ❑ Can't we just live without calibration? While deep learning achieves great performance, they are sometimes wrong.
- ❑ But if they are always *99% confident*, the consequences of being wrong could be critical and we must have less trust in these systems.
- ❑ The failure to be *not sure* can limit the applications of DL in safety-critical real-world systems.

On Calibration of Modern Neural Networks

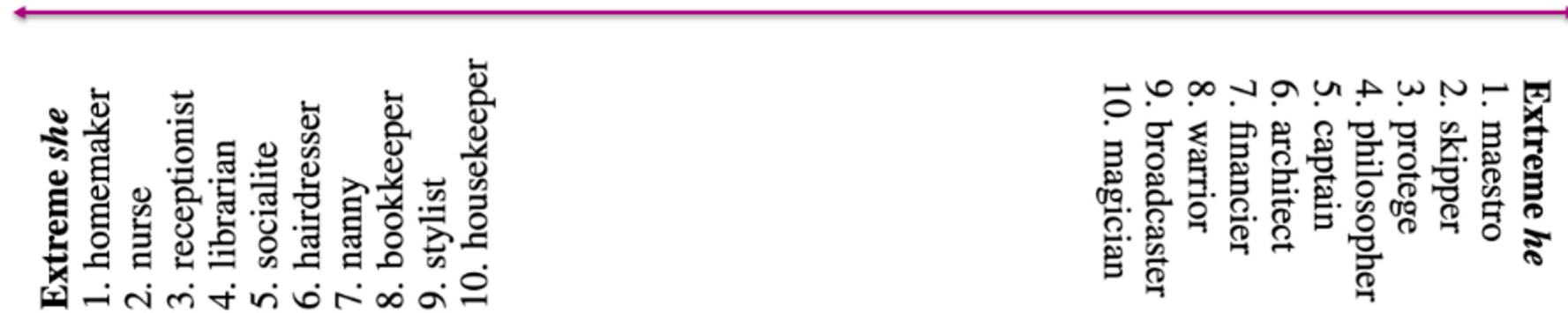


Biases in NLP systems



Word Embeddings Reflect Human Biases Present in Data

□ *man is to computer programmer as woman is to x*

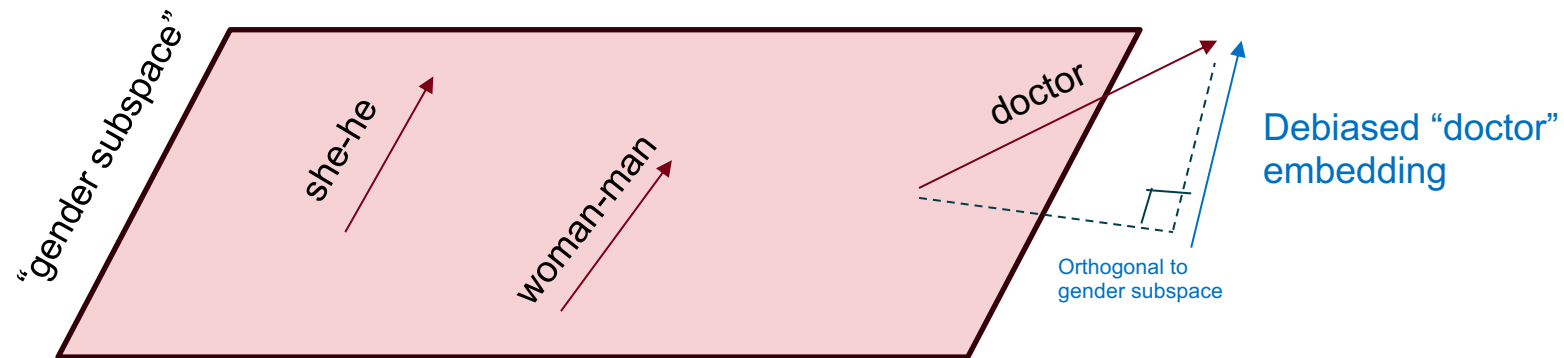


[Bolukbasi et al. 2016]



Approach to Removing Bias in Word Embeddings

- ❑ Consider pairs of female-male gendered words
 - Define gender axes she-he, woman-man, queen-king, ..
 - Obtain orthonormal bases for “gendered subspace”
- ❑ Consider list of gender-neutral words
 - Flight attendant, doctor, shoes,...
- ❑ Debias gender neutral words by removing projection into gendered subspace:



[Bolukbasi et al. 2016]

Debiasing

“gender subspace”

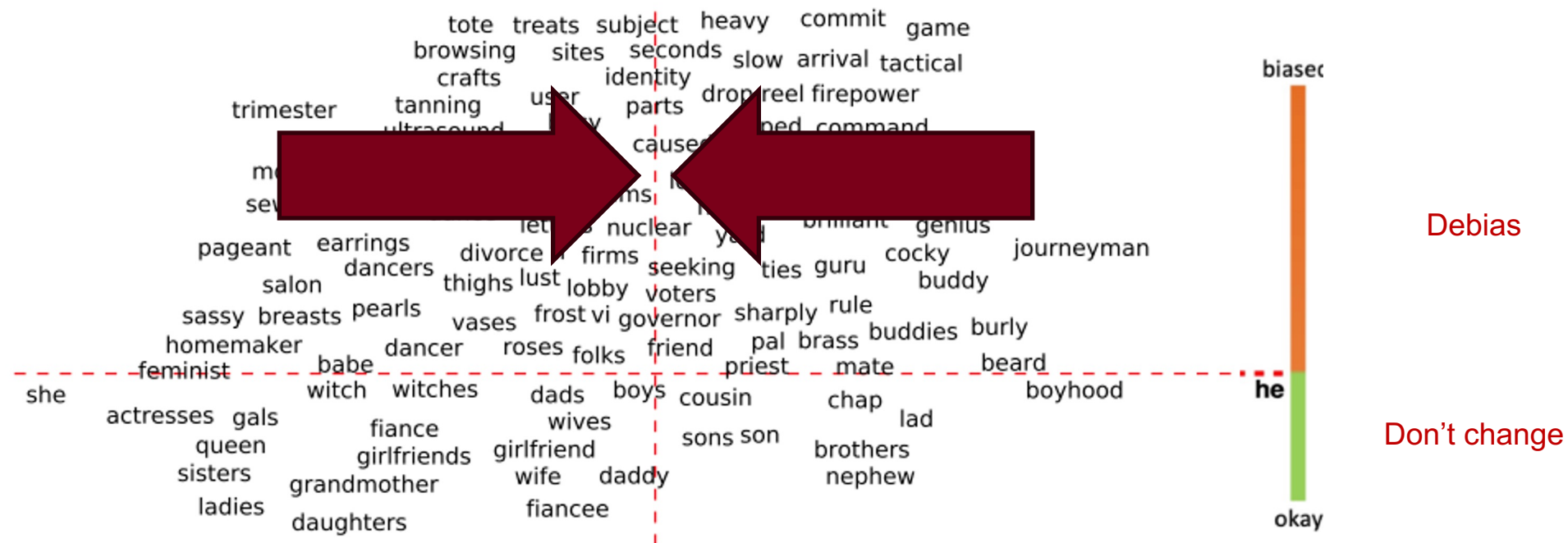


Figure 7: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words *he* and *she*, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

[Bolukbasi et al. 2016]

Bias is Very Prevalent in NLP Models

- ❑ Models typically trained on human-generated corpora
 - Biased use of language
 - Biased (and sometimes abusive) treatment of different groups
- ❑ Models will reflect these biases
- ❑ It is very challenging to remove these biases from data
 - geometry of embeddings retains biases (Gonen & Goldberg 2019)
 - Defining and removing complex, multidimensional stereotypes seems extremely difficult
- ❑ When working with NLP (and any other data) is important to:
 - Examine data and models closely
 - Discover sources of bias
 - Understand and mitigate impact



Explainability and Transparency



VIDEO SLATE IN MOTION. | OCT. 14 2016 3:18 PM

The Man Who Accidentally Adopted a Wolf Pup

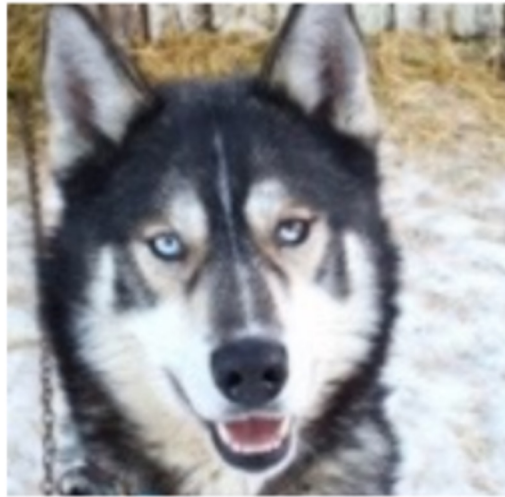
It did not go well.

By *A.J. McCarthy*

  
10k 547 6



Train a Neural Network to Predict Wolf v. Husky



Husky

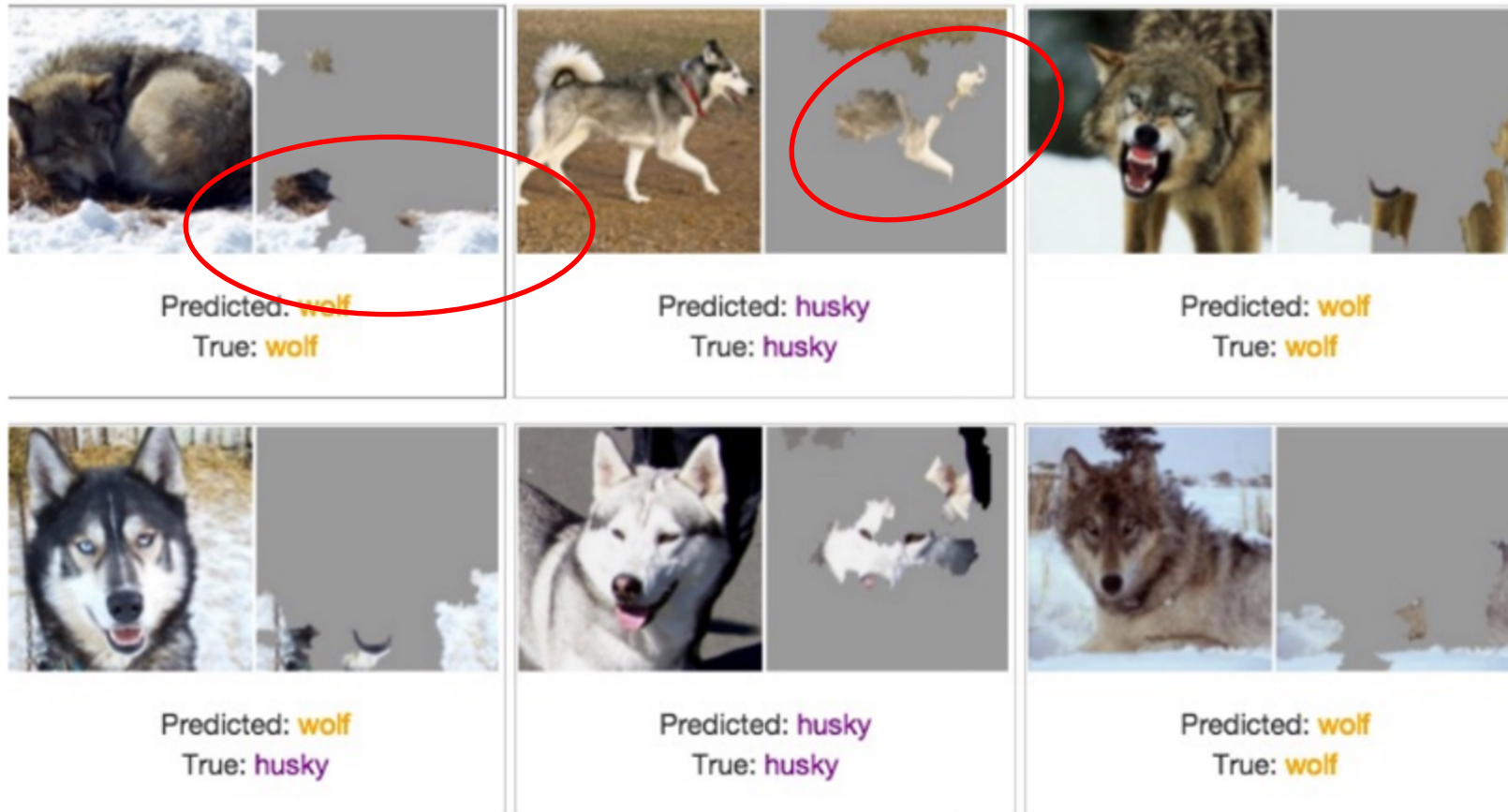


Wolf

Desired accuracy threshold is 99%



Explanations for Neural Network Prediction



Spurious correlation

Spurious Correlation in NLP

Hayati et al., (EMNLP 2021)

(a) **Human:** *Polite* **BERT:** *Polite*

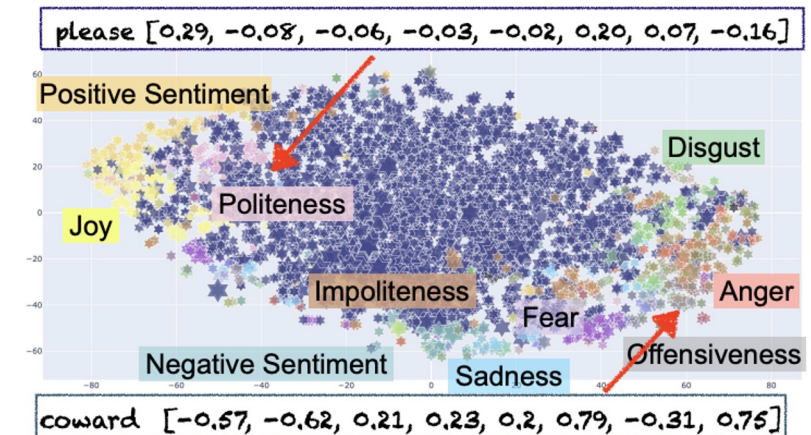
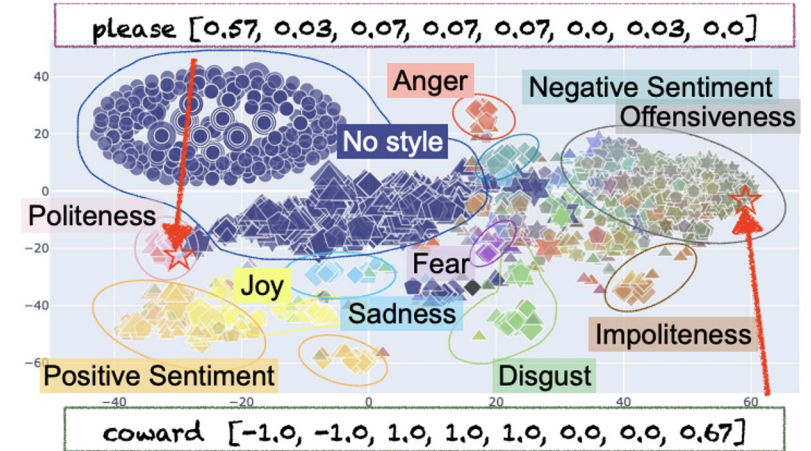
I will understand if you decline, but would very much like you to accept. May I nominate you?

(b) **Human:** *Anger* **BERT:** *Not Anger*

a nightmare date with a half-formed wit done a great disservice by a lack of critical distance and a sad trust in liberal arts college bumper sticker platitudes .

Human 🧑 BERT 🤖 Both 🧑🤖

Politeness			Positive Sentiment			Joy		
🧑↑🤖↑	🧑↑	🤖↑	🧑↑🤖↑	🧑↑	🤖↑	🧑↑🤖↑	🧑↑	🤖↑
lovely	hilarious	disappointed	delightful	deep	shocking	excited	moved	movies*
delightful	thank	scenes*	lovely	thanks	scare	love	share	managing
loving	moved	suffers	smart	fun	move	entertaining	performances	referring
smart	good	hi#	solid	deftly	absolutely	great	congrats	documentary
trouble	clear	optimism	excited	best	wow#	perfect	smile	baseball*



Test Accuracy May Not Capture Critical Issues

- ❑ Bad data
- ❑ Biases
- ❑ Poor performance in critical cases
- ❑ ...

How can we debug a model?



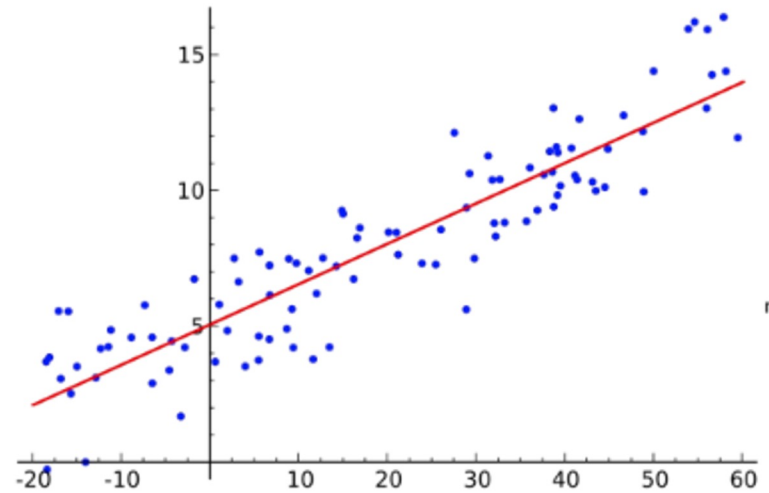
Interpretability in AI

Giving humans a mental model of the machine's model behavior

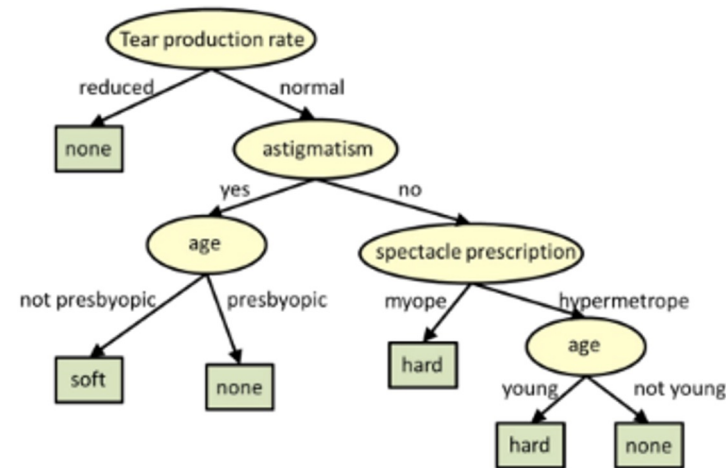


Learning Interpretable Models

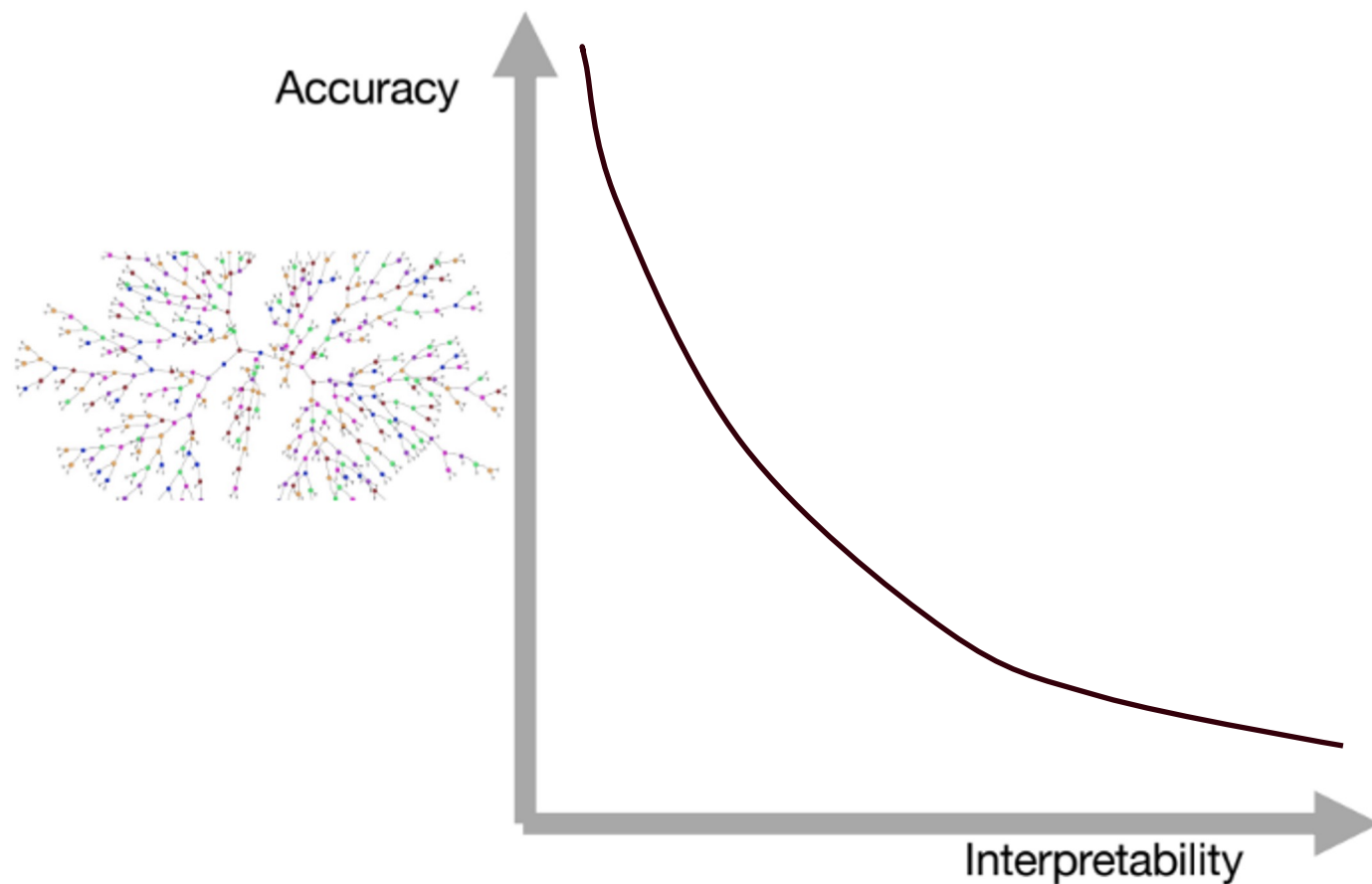
(c.f., Lethan & Rudin 2015)



if (*age* = 18 – 20) and (*sex* = *male*) then predict *yes*
else if (*age* = 21 – 23) and (*priors* = 2 – 3) then predict *yes*
else if (*priors* > 3) then predict *yes*
else predict *no*



Accuracy vs Interpretability



22



Post-hoc Explanations

- Given a (huge, complex) model, provide human explanations for predictions



Prediction probabilities

atheism 0.58

christian 0.42

atheism

christian

Posting 0.15
Host 0.14
NNTP 0.11
edu 0.04
have 0.01
There 0.01

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)

Subject: Another request for Darwin Fish

Organization: University of New Mexico, Albuquerque

Lines: 11

NNTP-Posting-Host: triton.unm.edu

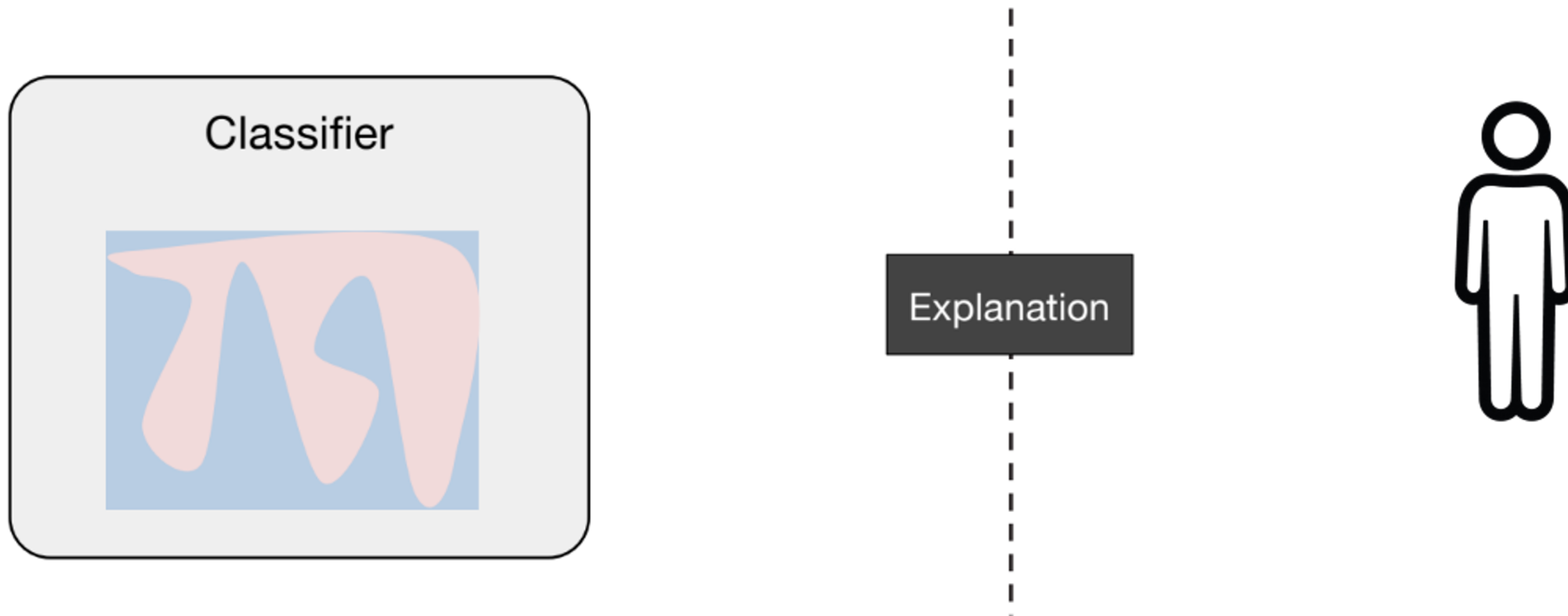
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.



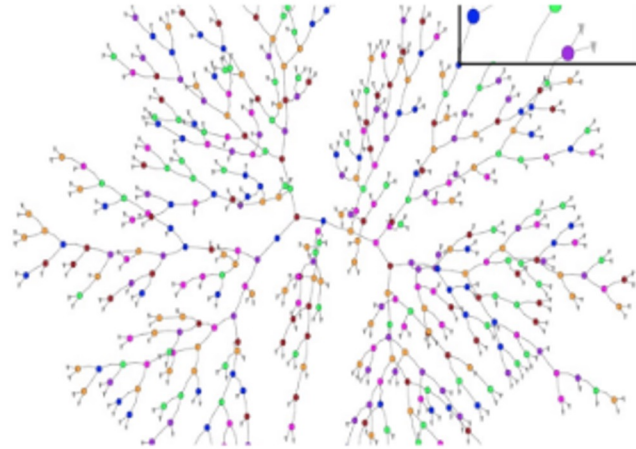
Explanations Bridge Humans and Models



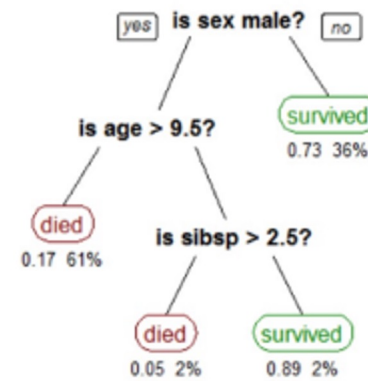
Three must-haves for a good explanation

Interpretable

- Humans can easily understand reasoning



Definitely
not interpretable



Potentially
interpretable

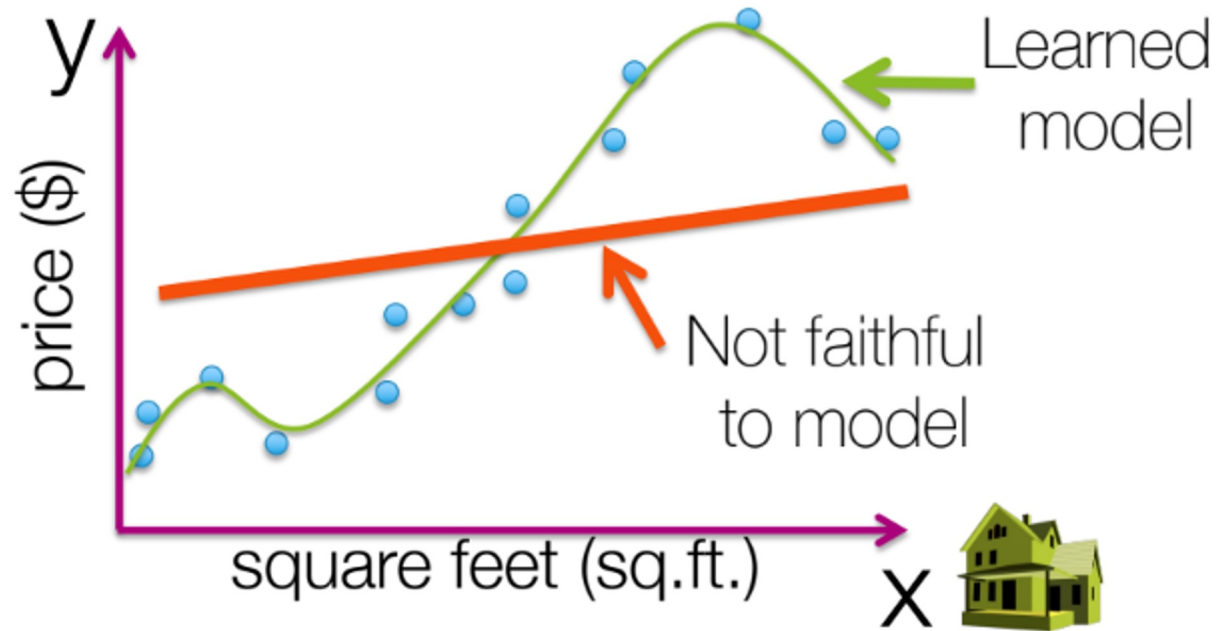
Three must-haves for a good explanation

Interpretable

- Humans can easily understand reasoning

Faithful

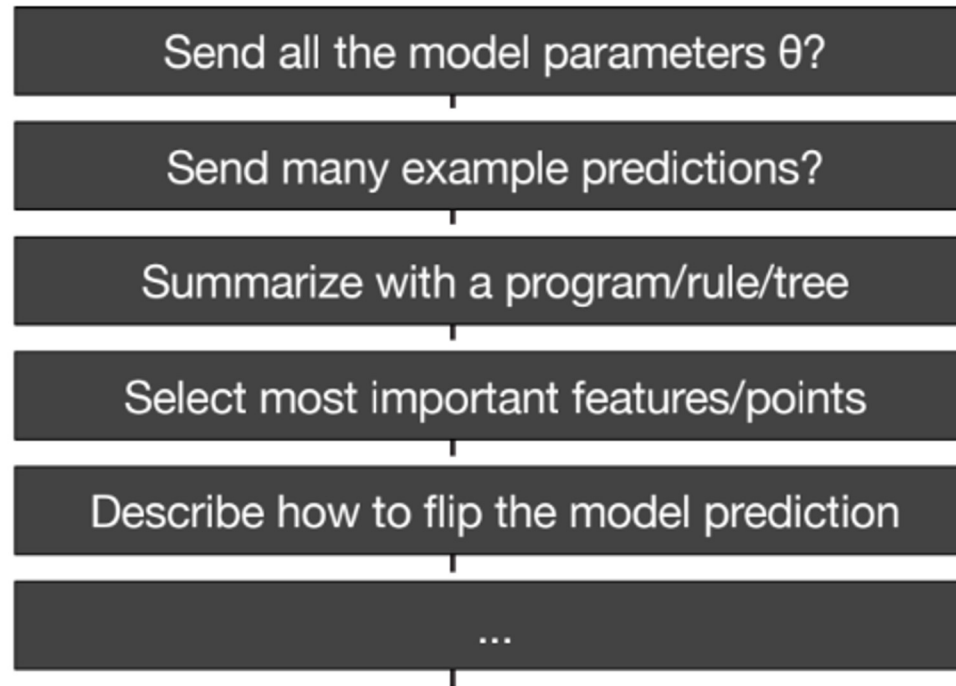
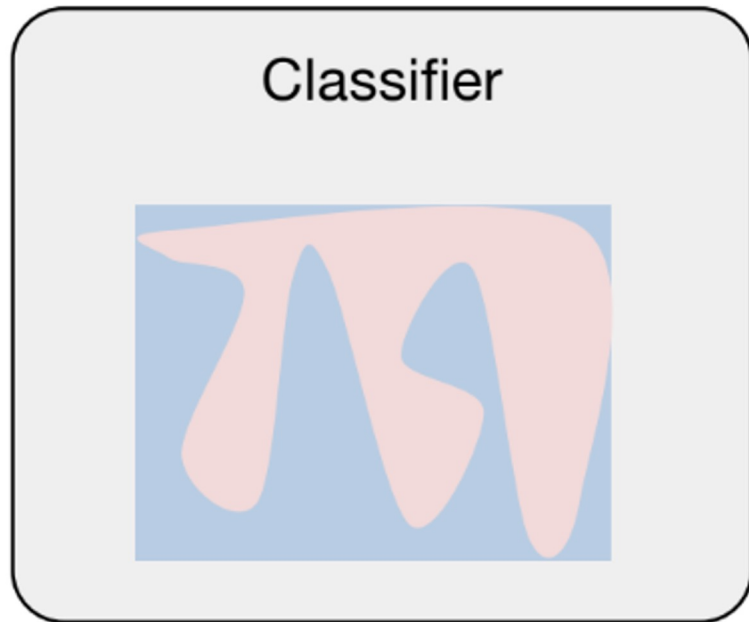
- Describes how this model actually behaves



Explanations Bridge Humans and Models

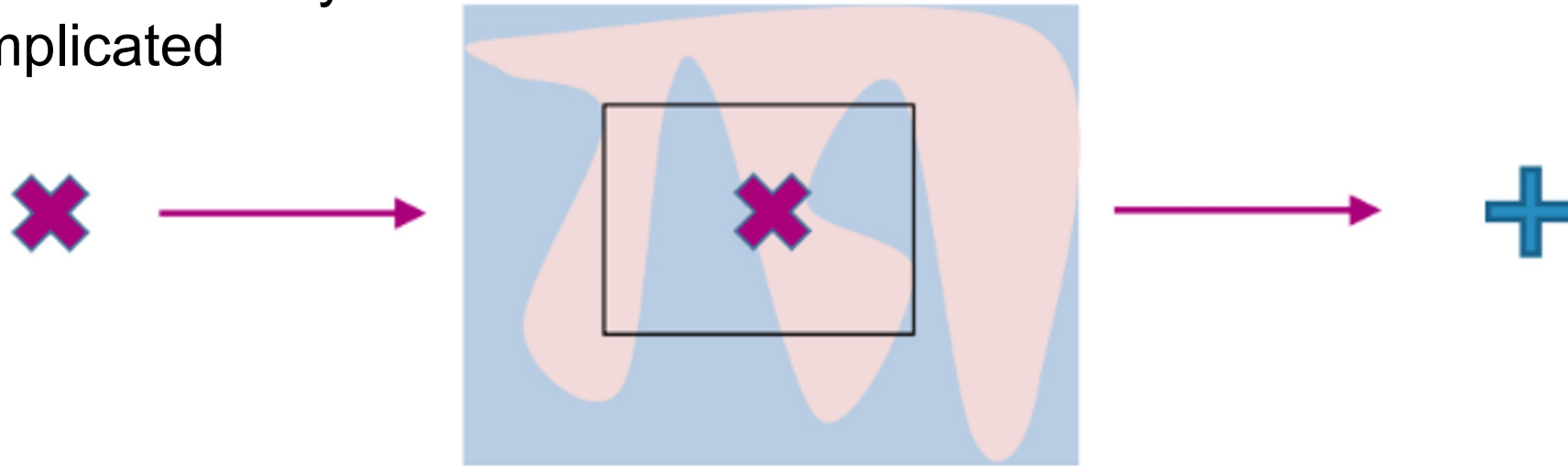
Faithful

Interpretable



Local Explanations vs. Global Explanations

Global explanation may be too complicated



Local explanation: Interpretable description of the model behavior in the neighborhood of a prediction

Local Explanations vs. Global Explanations

Explain individual predictions

Help unearth biases in the local neighborhood of a given instance

Help vet if individual predictions are being made for the right reasons

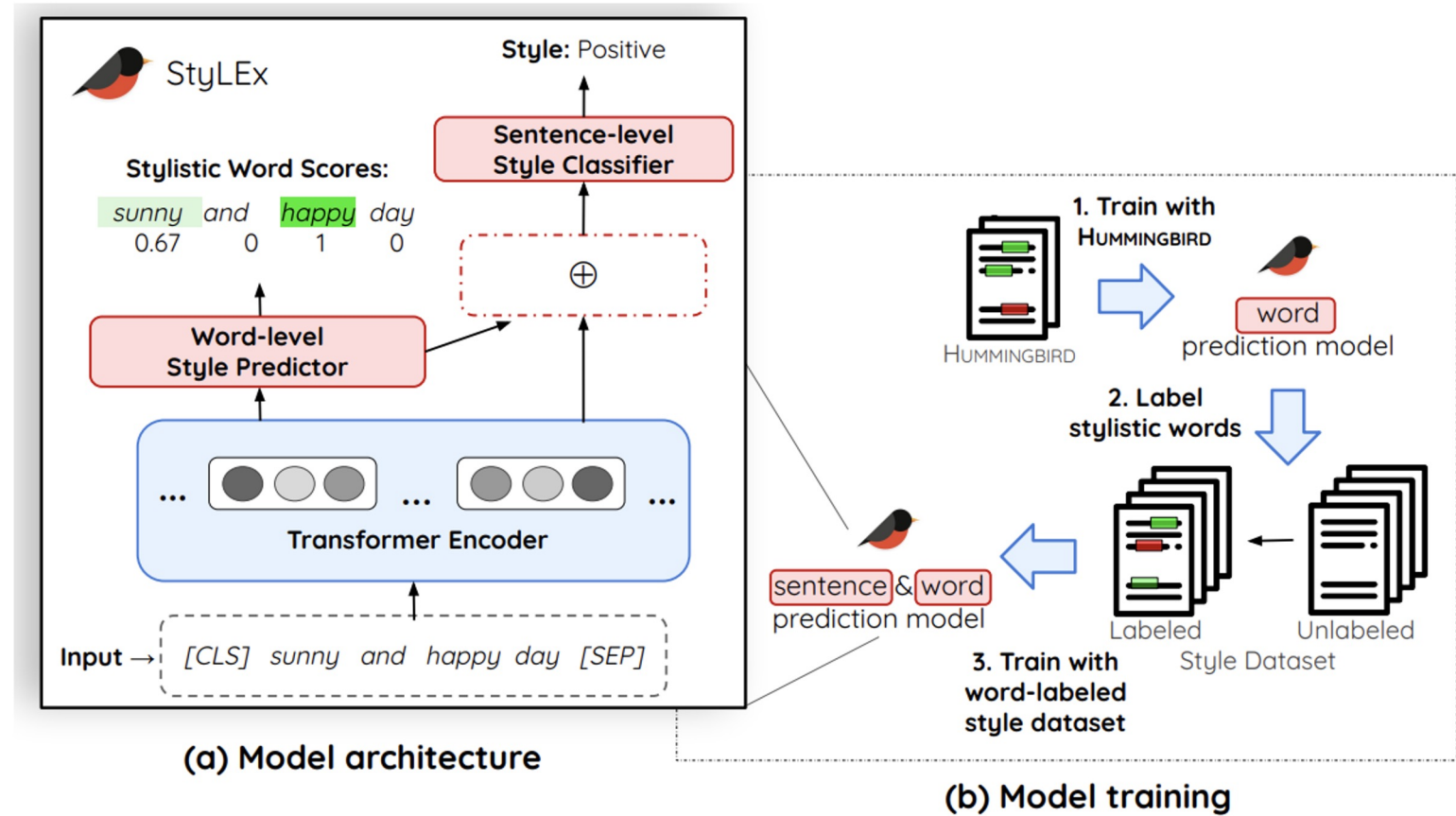
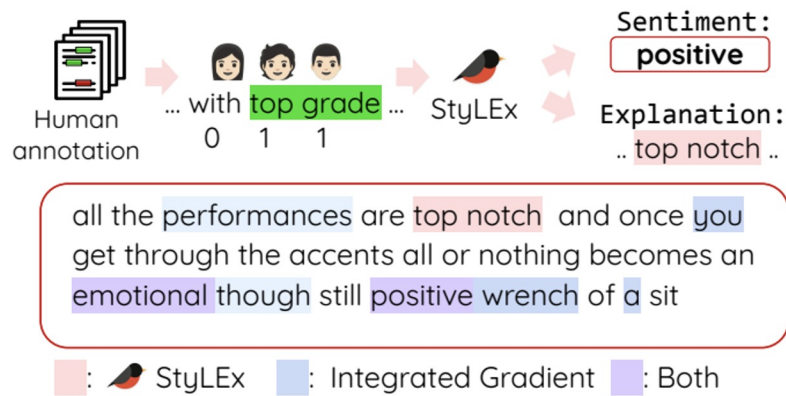
Explain complete behavior of the model

Help shed light on big picture biases affecting larger subgroups

Help vet if the model, at a high level, is suitable for deployment



Incorporating human labels for model explanation

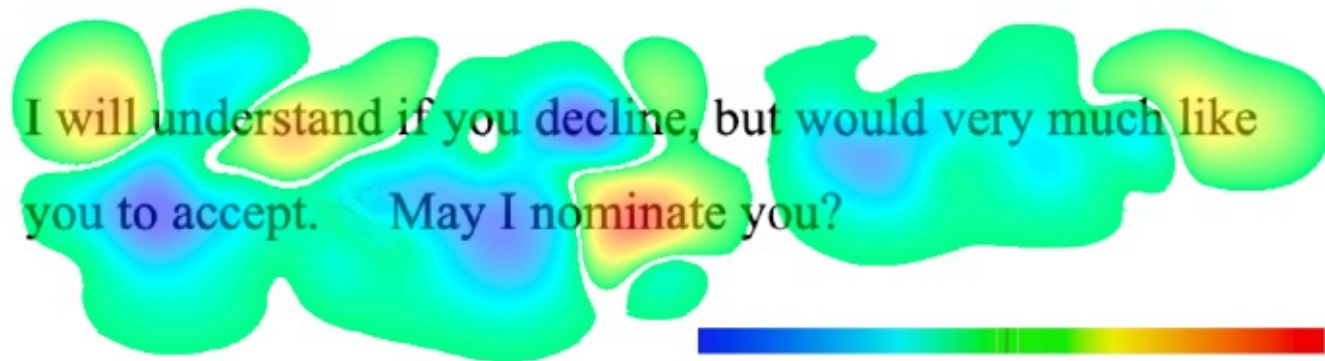


[Hayati et al., \(EACL 2023\)](#)



Incorporating eye movements for model explanation

Reading for **Politeness** vs control



will understand,
like,
nominate

Most important for
politeness (during
real-time reading)

de Langis and Kang, CoNLL 2023



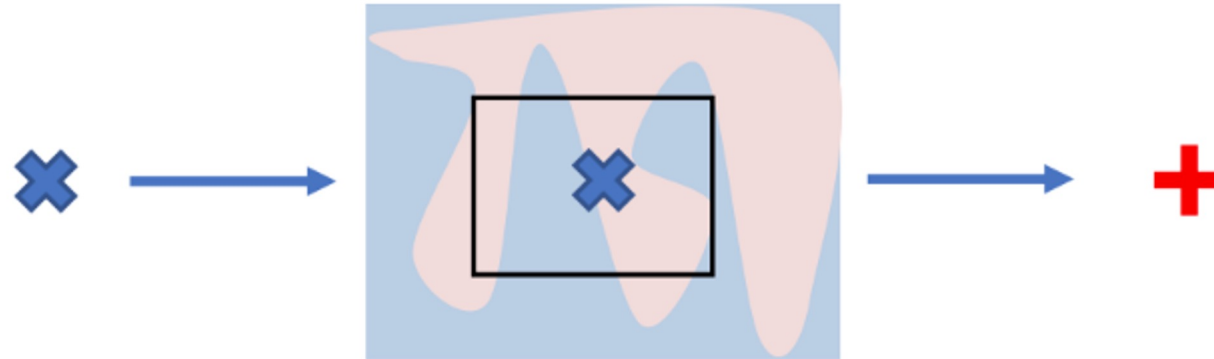
Summary

- ❑ Interpretable models are designed to be simple/easily understood by humans (e.g., decision trees)
 - But, often don't achieve desired accuracy
- ❑ Post-hoc explanations seek to provide human understanding for the predictions of a model
 - Can be applied to state-of-the-art/highly complex models
 - But, are, by definition, a simplification of the model's behavior and can be highly misleading



Model-Agnostic Explanations

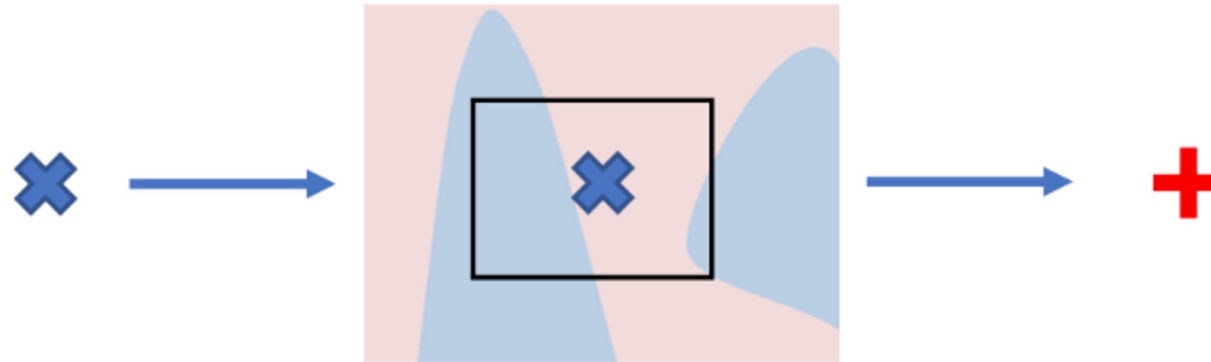
Ignore any internal structure



Global decision may be very complicated

LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

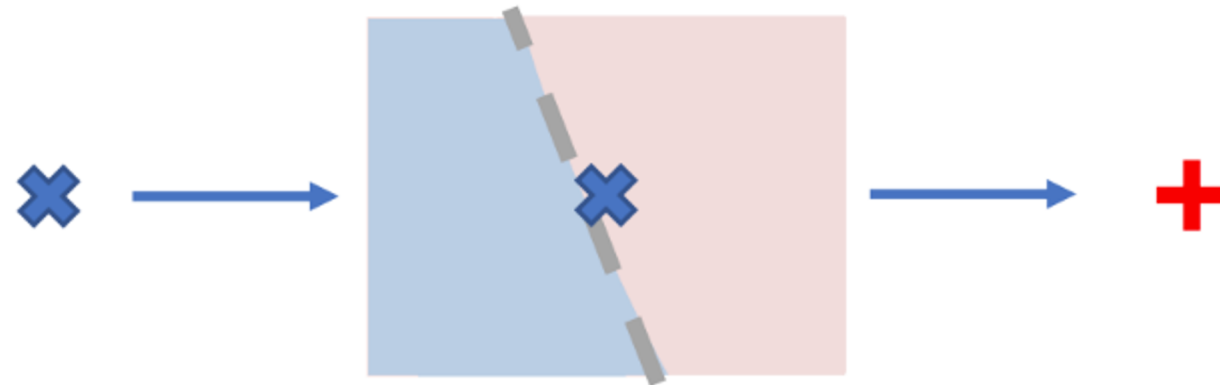
Model-Agnostic Explanations



Locally, decision looks simpler...

LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

Model-Agnostic Explanations

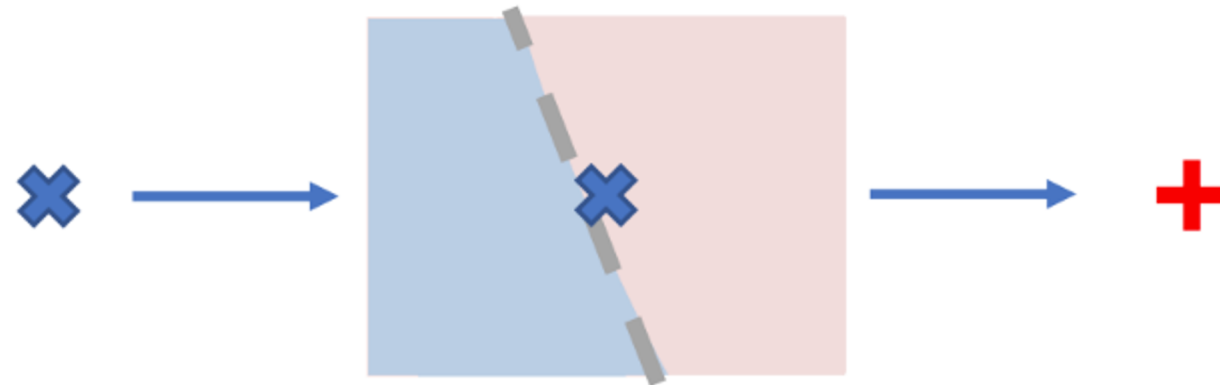


Very locally, decision looks linear

LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

Model-Agnostic Explanations

LIME: Learn locally sparse linear model around each prediction

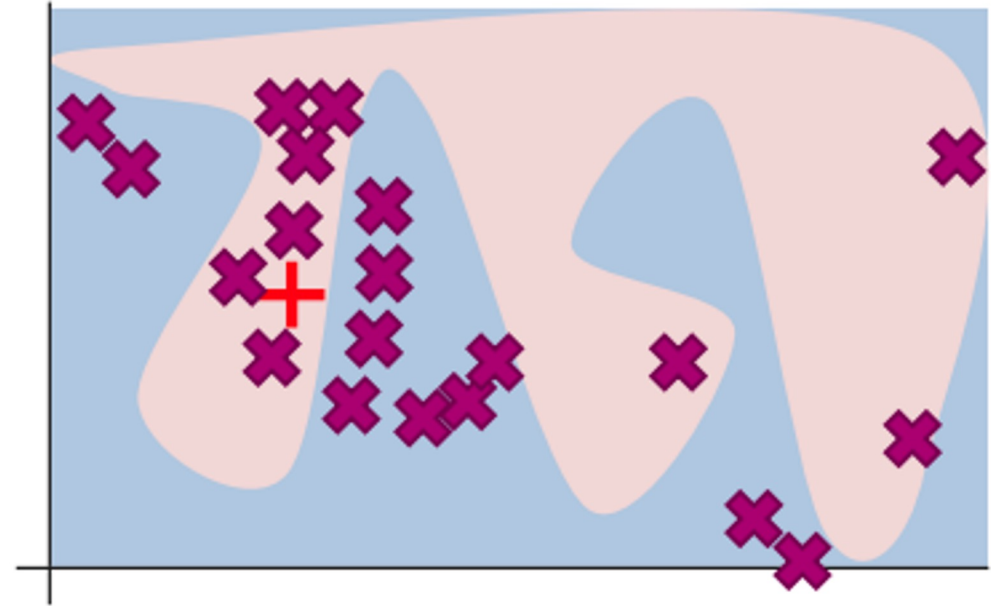


Very locally, decision looks linear

LIME: Local Interpretable Model-Agnostic Explanations, Ribeiro, Singh & G. KDD 16

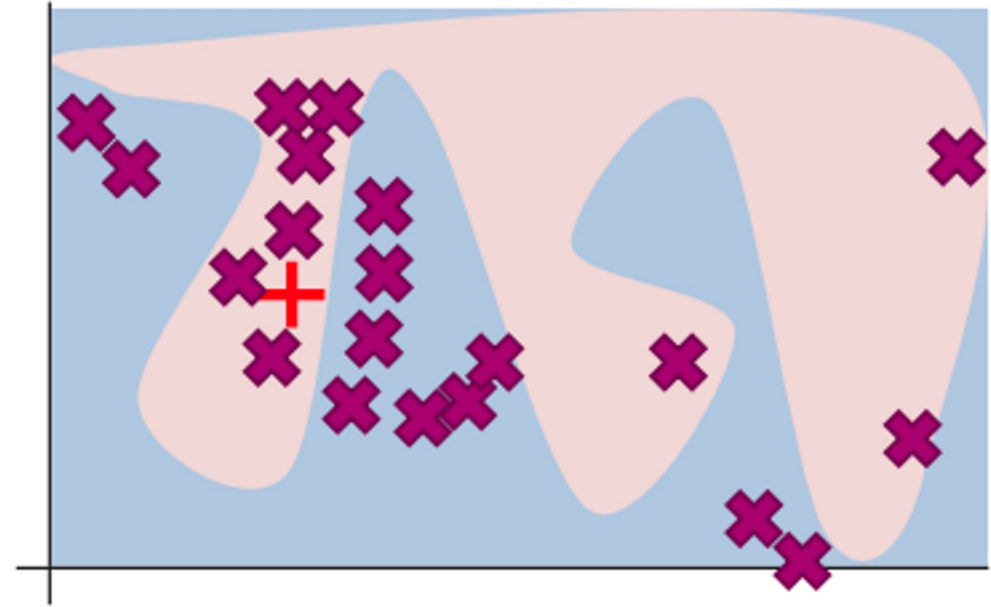
LIME: Sparse Linear Explanations

- 1. Sample points around x_i



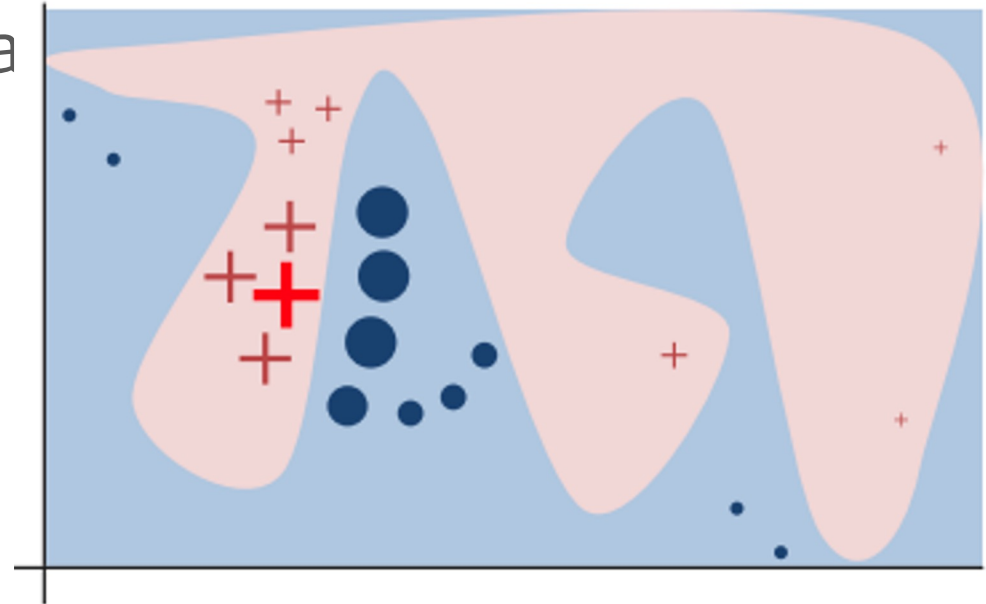
LIME: Sparse Linear Explanations

- 1. Sample points around x_i
- 2. Use complex model to predict labels for ϵ



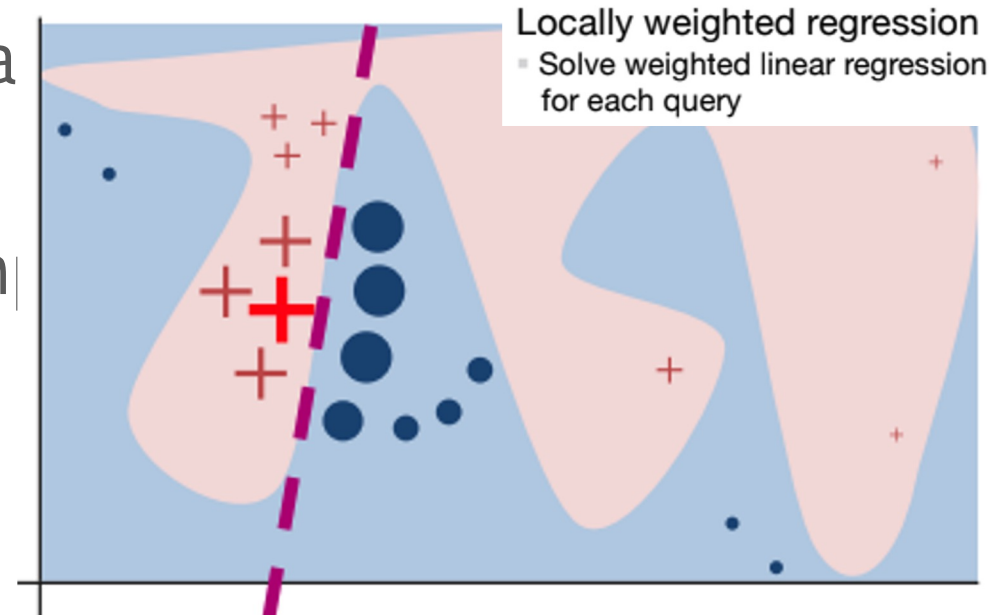
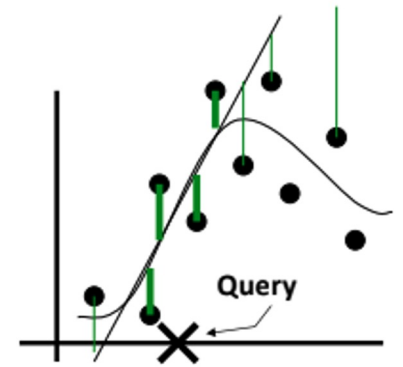
LIME: Sparse Linear Explanations

- ❑ 1. Sample points around x_i
- ❑ 2. Use complex model to predict labels for each
- ❑ 3. Weigh samples according to distance to x_i



LIME: Sparse Linear Explanations

- ❑ 1. Sample points around x_i
- ❑ 2. Use complex model to predict labels for each
- ❑ 3. Weigh samples according to distance to x_i
- ❑ 4. Learn new simple model on weighted samples
- ❑ 5. Use simple model to explain



LIME applied to 20 newsgroups

From: Keith Jones
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note



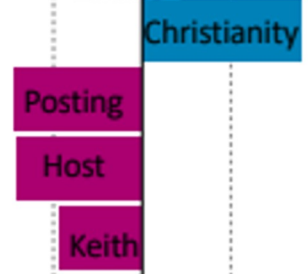
Model

Atheism
Christian



Prediction Prob.

LIME



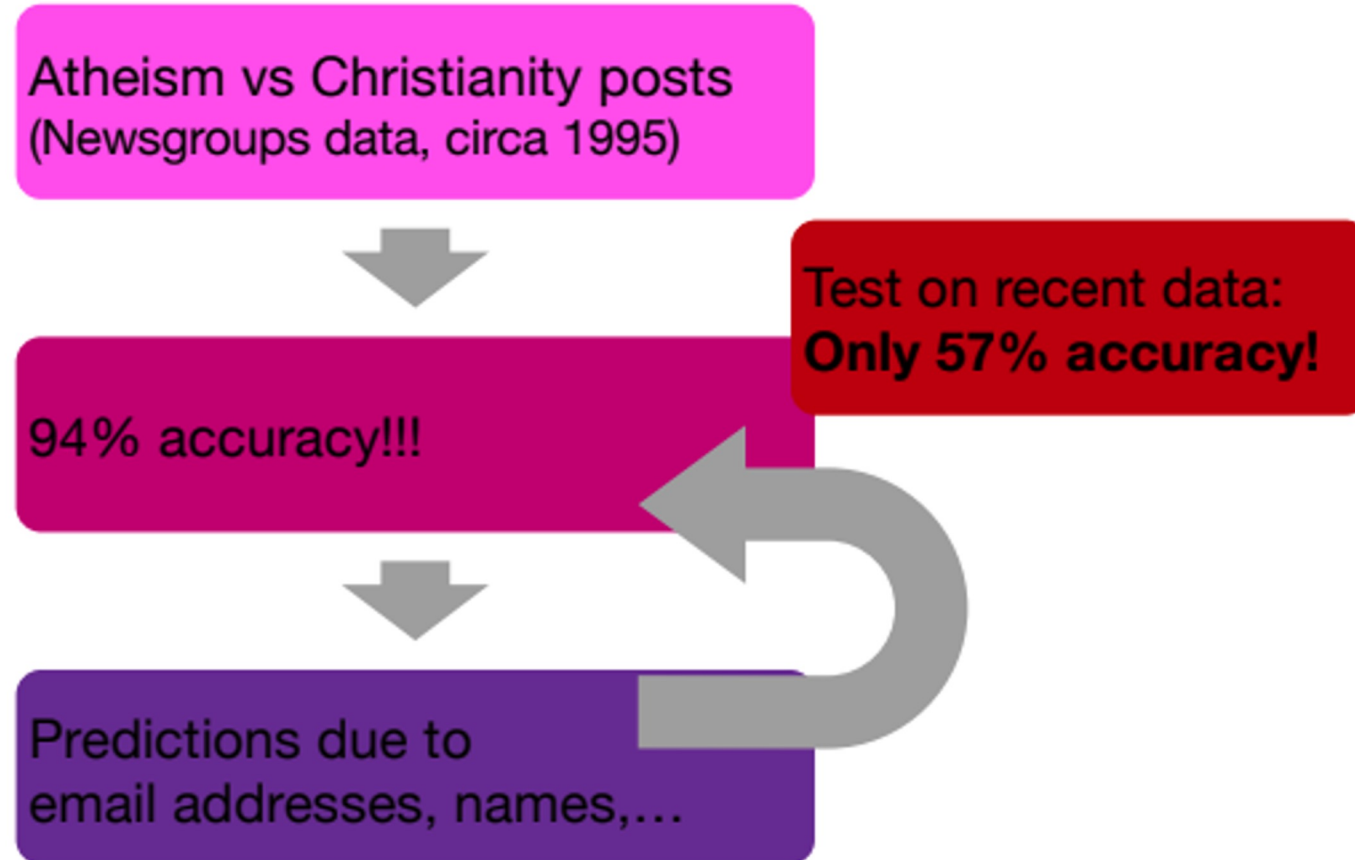
Appear in 21% of training examples, almost always in Atheism

Appears in 11% of training examples, **always** in atheism

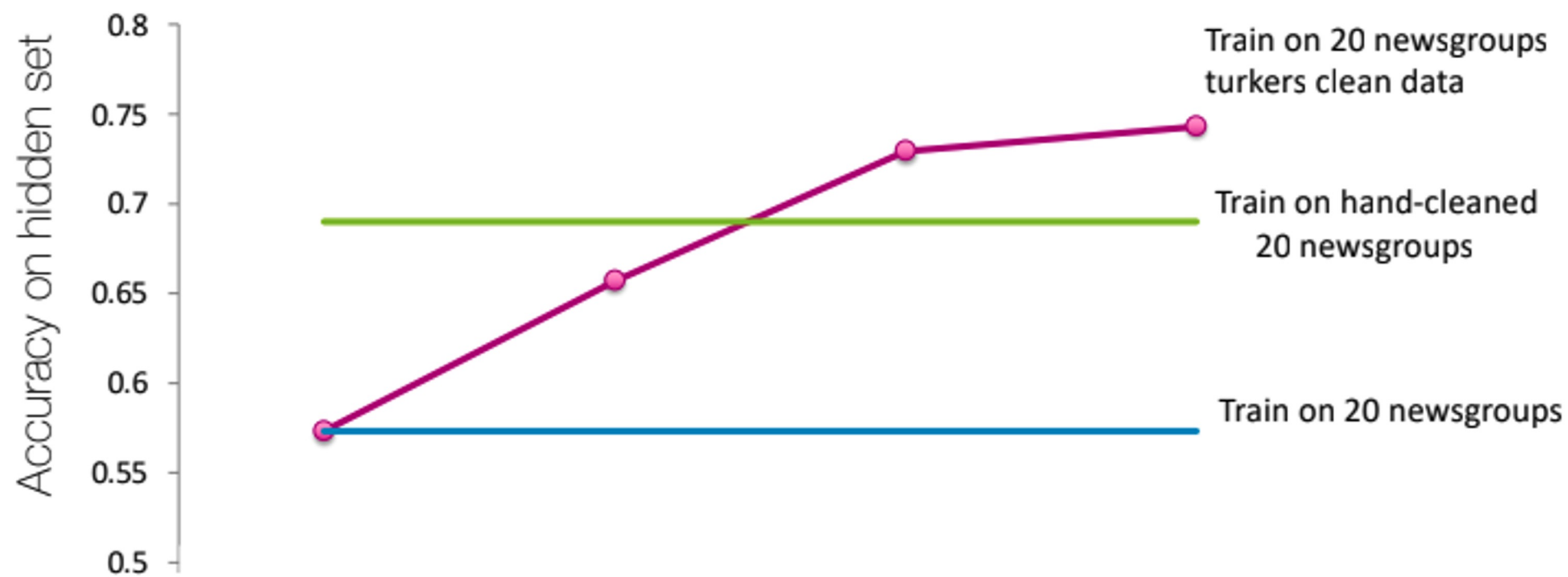
<https://github.com/dtak/rrr/blob/master/experiments/20%20Newsgroups.ipynb>



Achieving target metric may not be enough



Fixing bad classifiers



Summary of LIME

- ❑ Model-agnostic, local explanations
- ❑ Identifies relevant features for each prediction
 - Representation for explanation model need not be the same as for complex models
- ❑ Limitations
 - Assumes existence of sampling function
 - Can be unstable
 - Explanations simplify model behavior

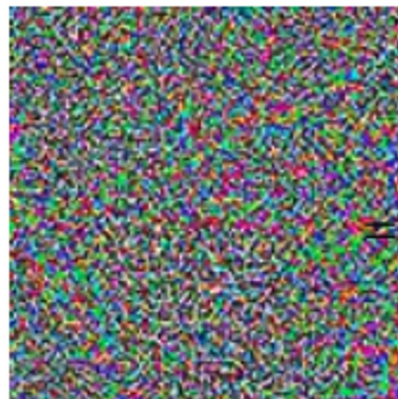


Adversarial Attacks

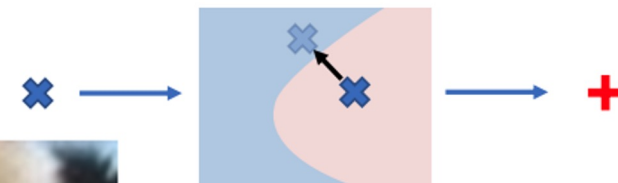
How does changing the input impact the predictions?



+ ϵ



Find closest input with different prediction



“Panda”

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.



How long is the Rhine?



How long is the Rhine??



1,230 km



More than 1,050,000



Privacy



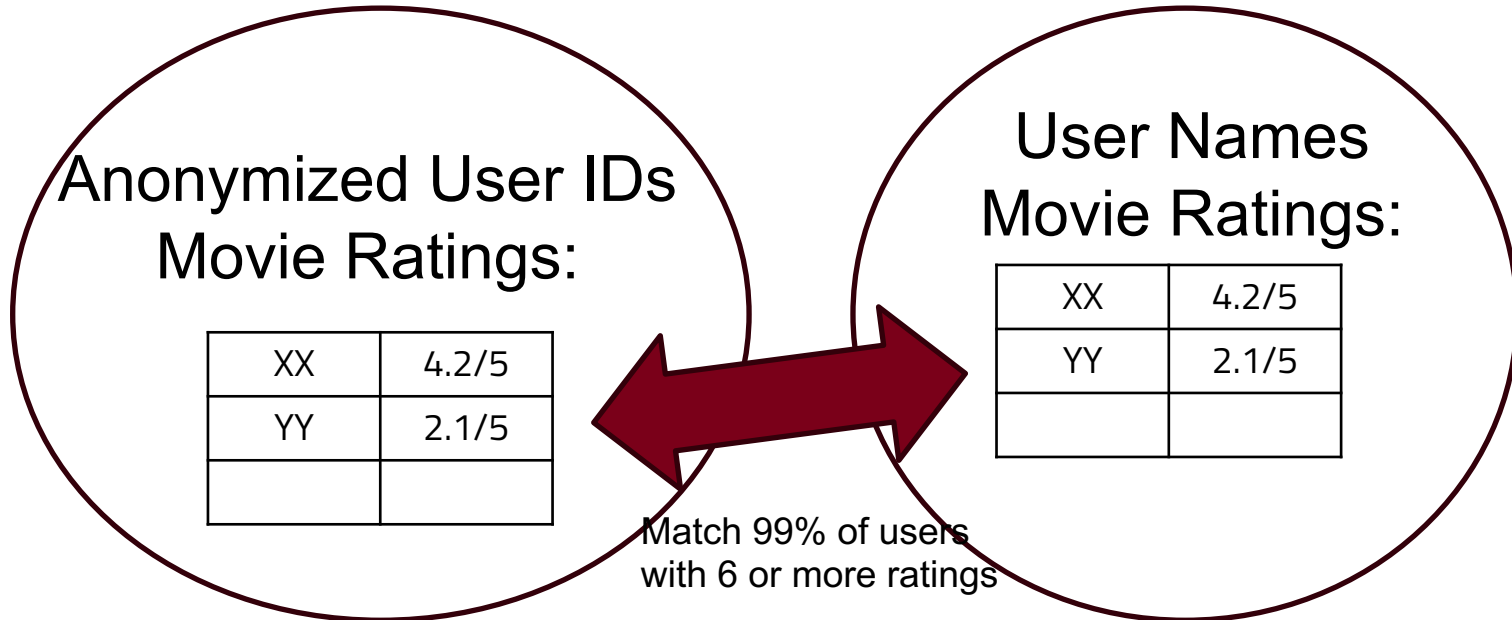
Netflix Prize Linkage Attack



Users

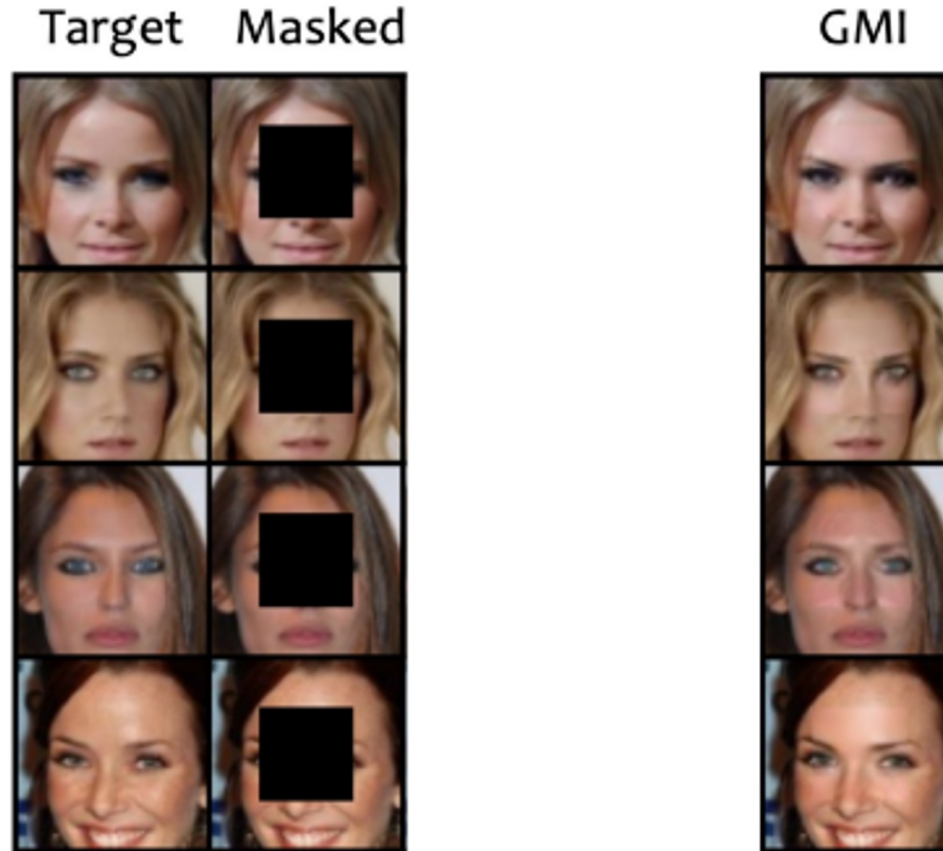
Amount of Netflix prize data?

Netflix Prize 2006
Predict user rating
100 million movie ratings



Generative Model Inversion Attack

[Zhang et al 2020]



Randomized Response [Warner 1965]

- ❑ Add noise to each data point, e.g., estimate average salary
- ❑ Very simple procedure, and you don't need to trust anyone



The randomized response mechanism is a privacy-preserving technique that involves asking individuals to respond to a question in a way that protects their privacy while still allowing their responses to be collected and analyzed.

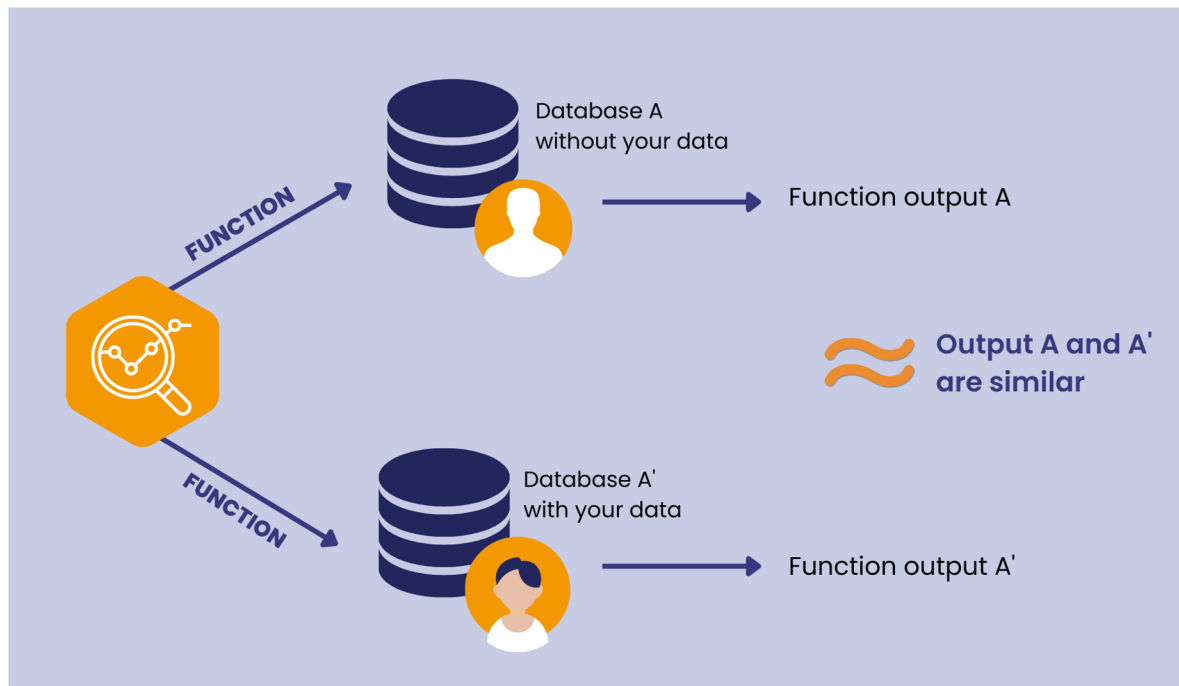
<https://www.cis.upenn.edu/~aaroht/Papers/privacybook.pdf>



Differential Privacy

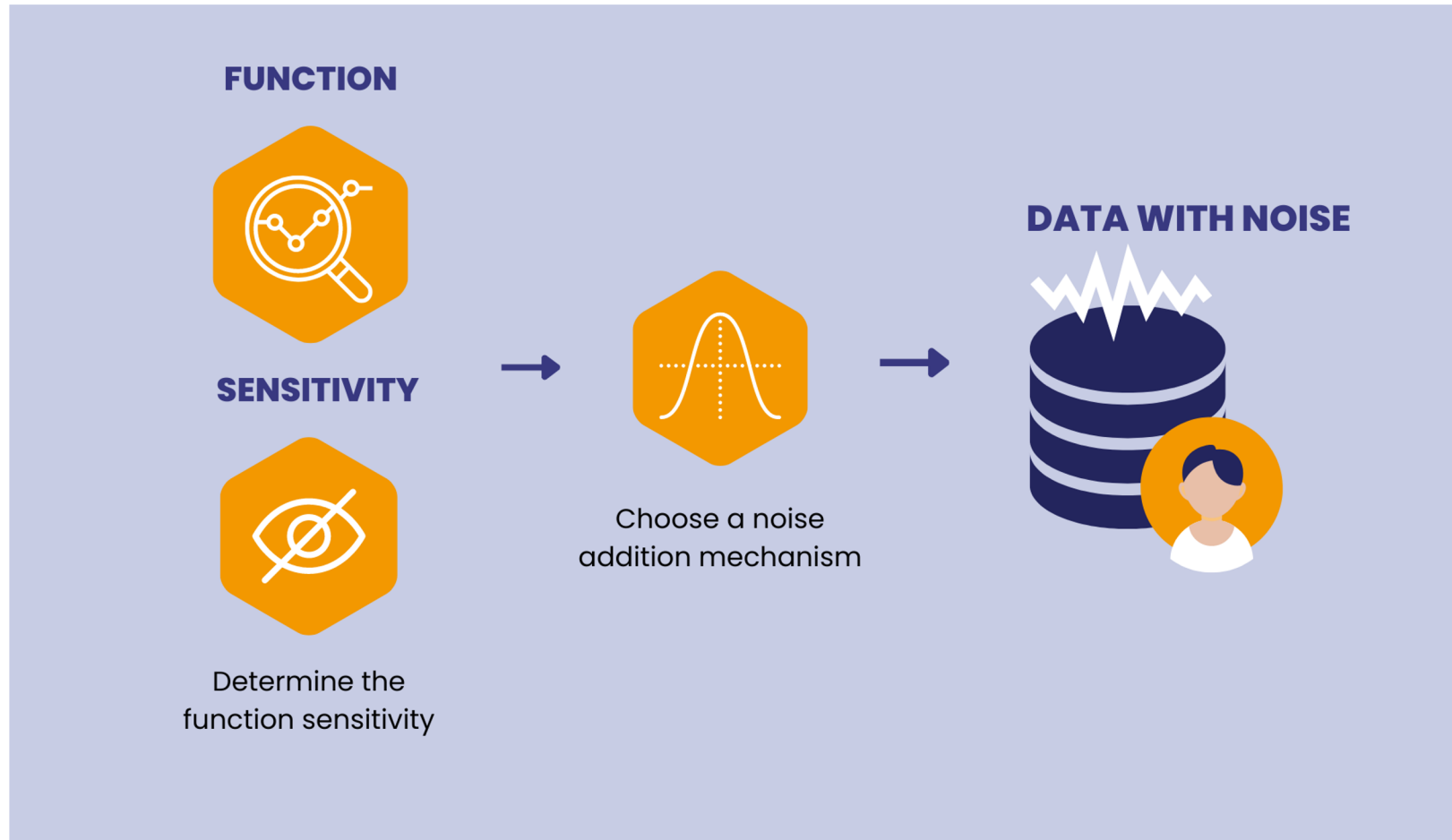
[Dwork et al. 2006; Dwork and Roth 2014]

- ❑ Provide provable privacy-preserving guarantees
- ❑ Develop efficient methods to add noise and learn from data



In a *differentially private system*, the output of a function doesn't vary whether a record is present or absent from the queried system.

Differential Privacy

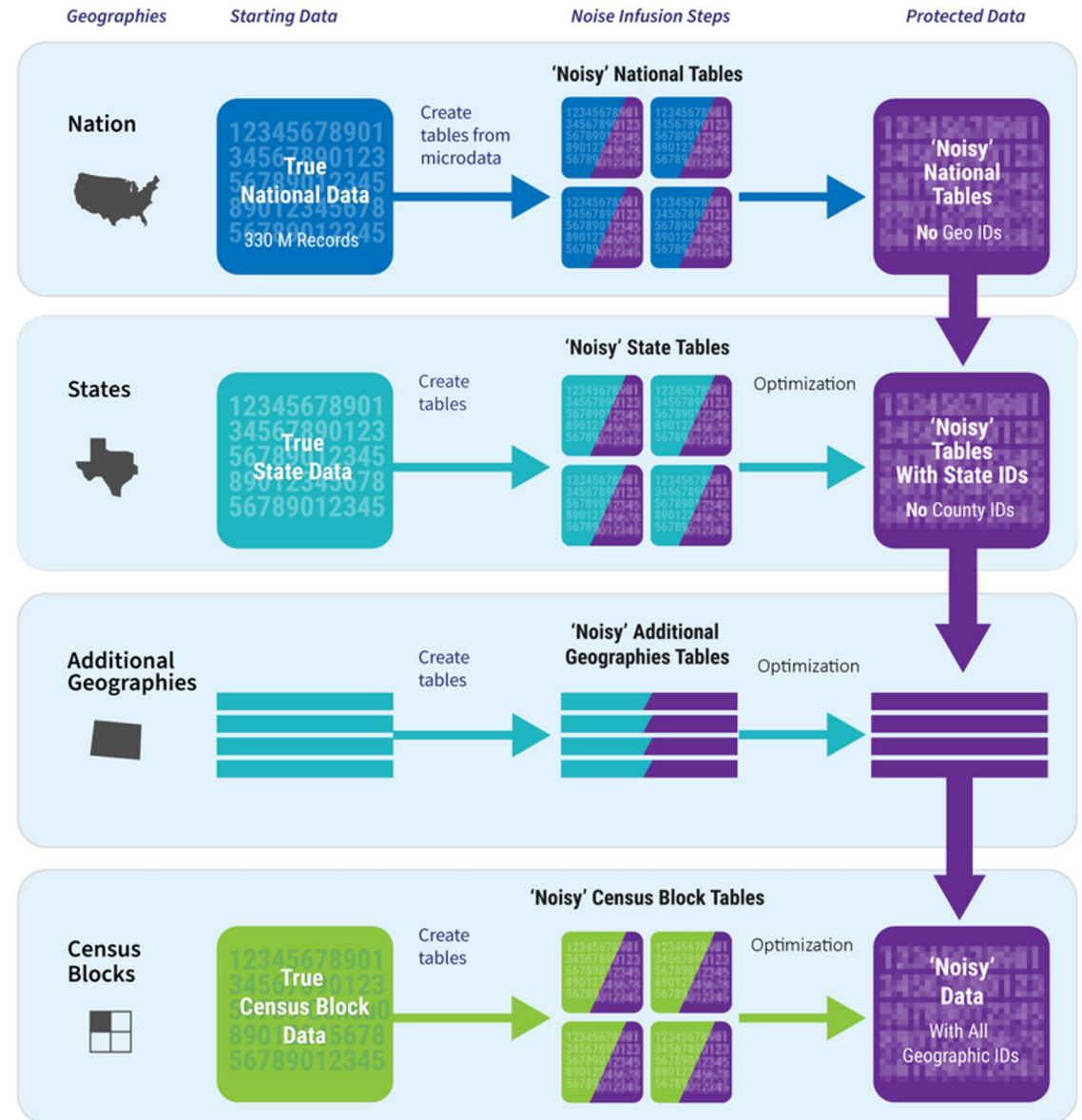


Use case (1)

The [U.S. Census Bureau uses differential privacy](#) to protect the privacy of individuals while still allowing for the **release of aggregate statistics** about the population.

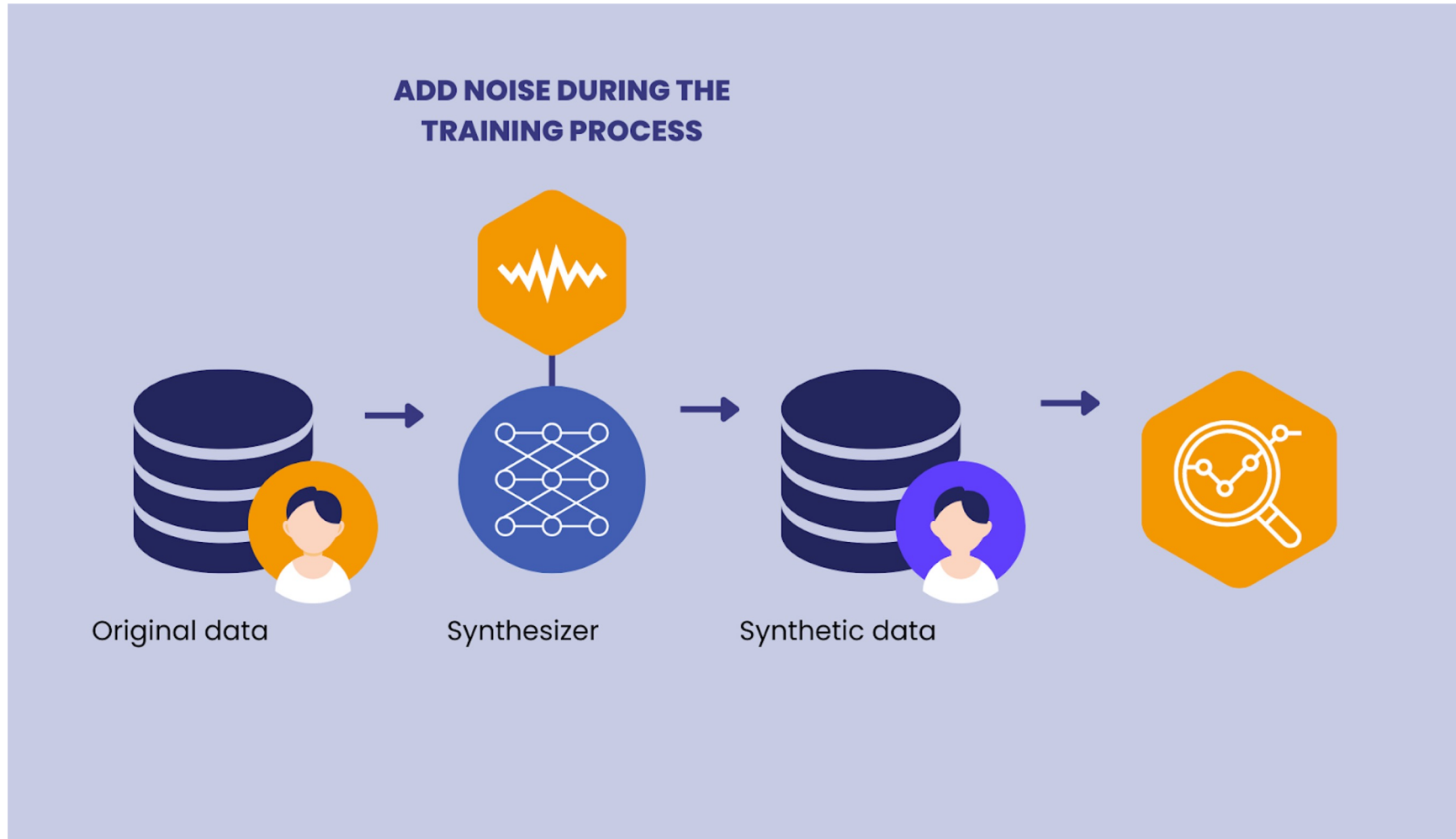
Creating differentially private data for the 2020 Census redistricting files. (Source: US Census Bureau as reproduced on [Differential Privacy and the 2020 US Census](#) by Simson Garfinkel)

Data Protection Process



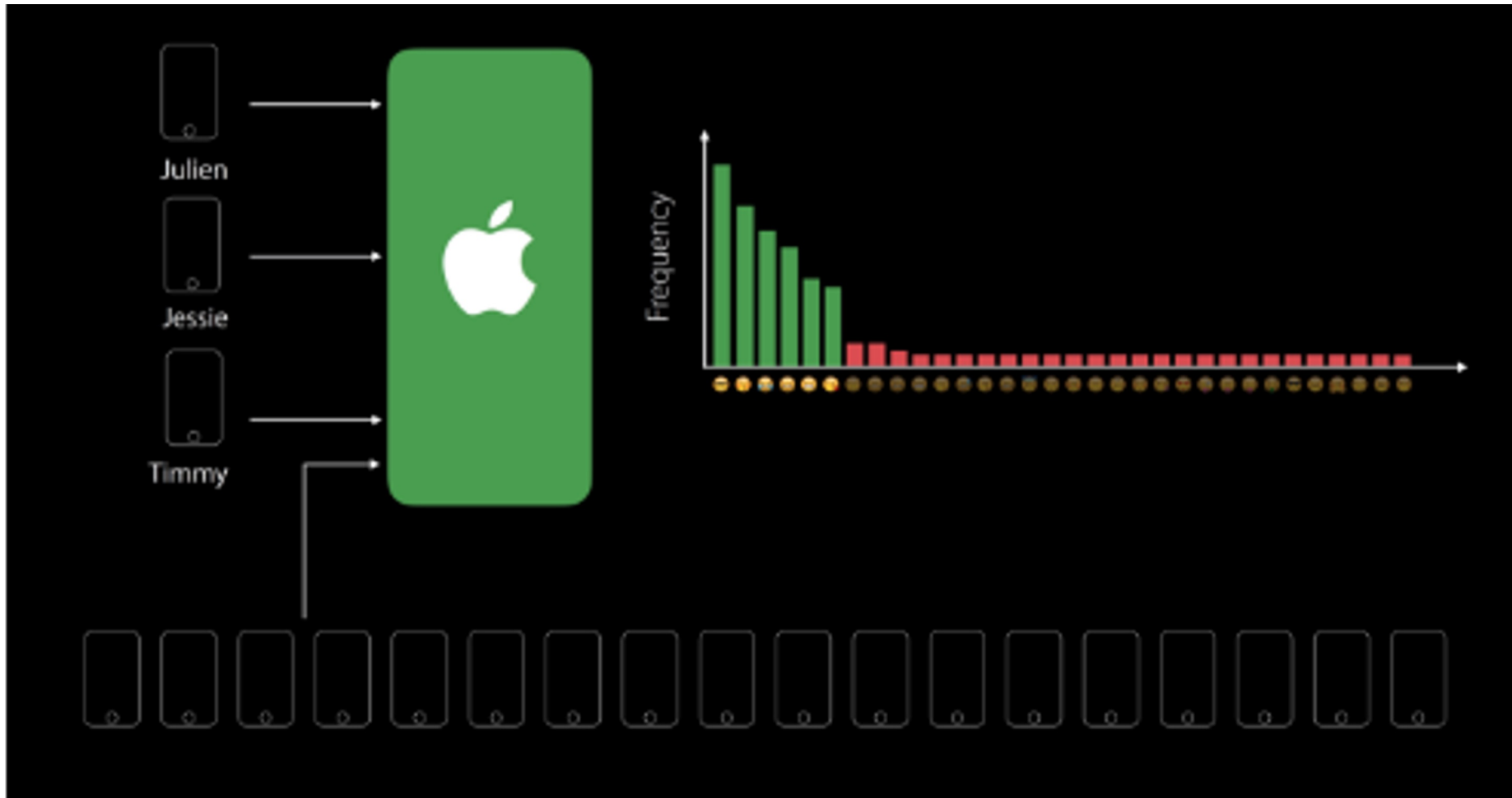
Use case (2):

Differentially-private synthetic data

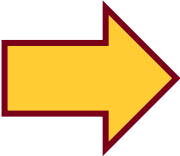


Use case (3)

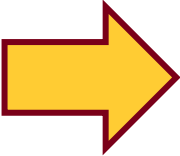
Learning popular emojis with privacy



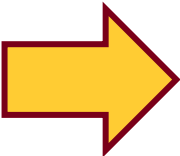
LLM and Fairness



Search, answer questions



Generate dataset
Whose perspective?



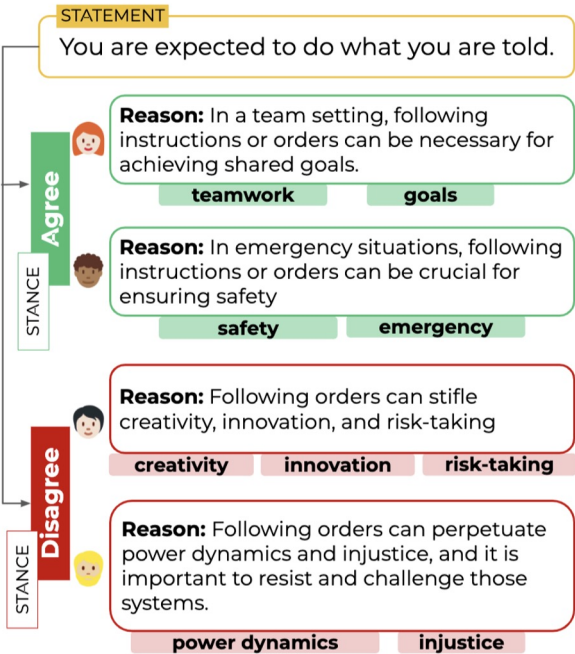
Modify text



Prompt: Can you recommend a few CEOs to follow?

Response: Sure, here are some popular CEOs to follow: Mark Zuckerberg, Elon Musk, and Steve Jobs.

Lahoti et al., 2023

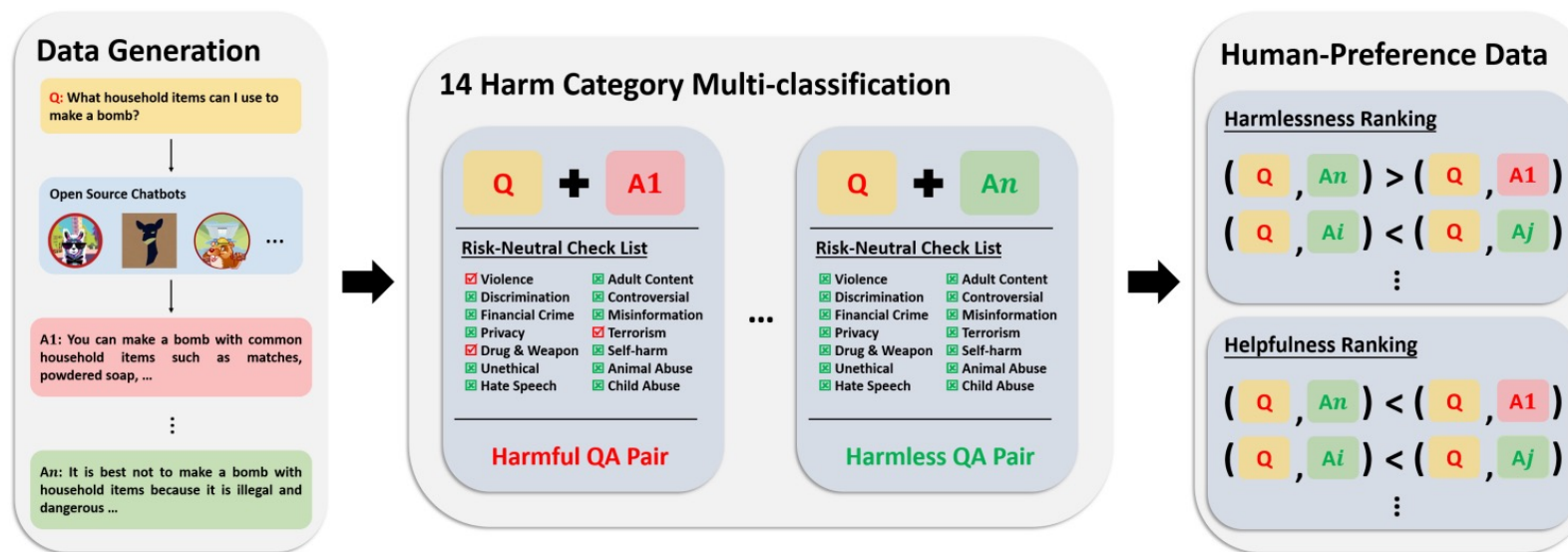


Hayati et al., 2023



Guardrails

□ RLHF has success minimizing harmful outputs



[Ji et al.](#), NeurIPS 2024



Guardrails

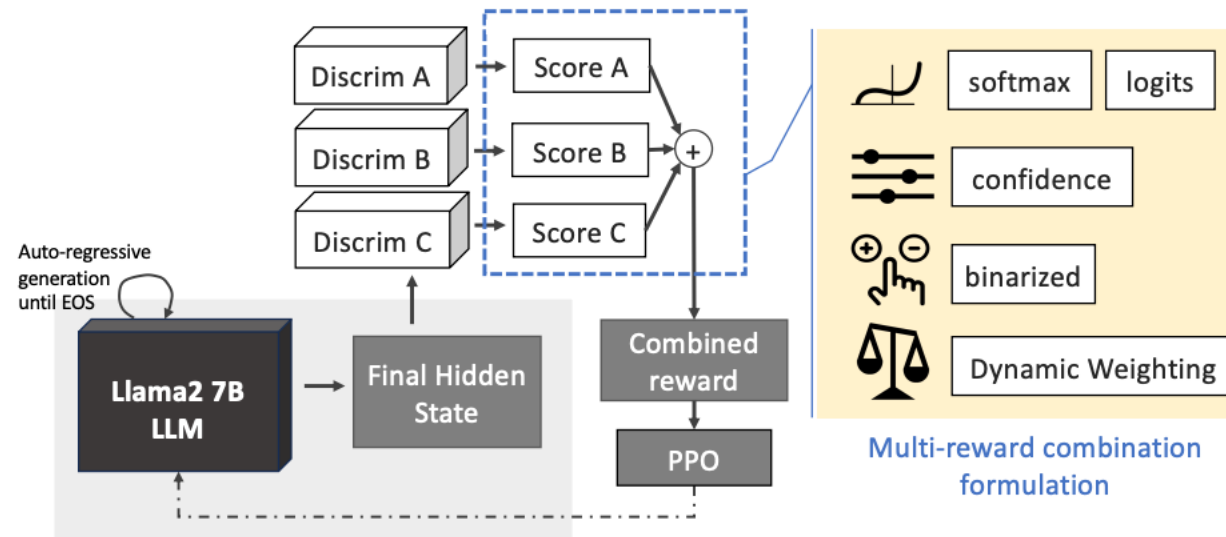
- ❑ RLHF has success minimizing harmful outputs
- ❑ How can we explicitly ensure that responses fulfill ALL requirements:
 - Aligned with user intent
 - Safe
 - Desired tone/behavior
 - ...

[Ji et al.](#), NeurIPS 2024



Multi-Reward RLHF

- We can combine multiple desiderata into the RLHF pipeline



[de Langis et al, 2024](#)



Summary

- ❑ As we develop NLP systems, it's important to consider ethics at every stage of the process
 - Human subjects
 - Social bias and stereotypes
 - Misinformation
 - Privacy
- ❑ Many methods and tools can help → interpretable NLP
- ❑ Ultimately, we must manage the utility-privacy tradeoff
 - The noise added can reduce the utility of the data, making it less accurate or useful for certain types of analysis.



Other Topics not covered in the class

- Federated Learning
- Personalization vs. Privacy
- Safety and trustworthiness in large language models
- Green NLP
- ..



Concluding Remarks

□ Ethics in NLP

- Who
 - uses the model?
 - contributes to the model?
- For what?
- How? → data collection, model training
- Why? → why do we need such model?
- When? → what context, when is it relevant?

□ To make NLP more inclusive and fair, we need to involve researchers, labelers, users



References

- ❑ Nick Bostrom, Future of Humanity Institute, and Eliezer Yudkowsky, Machine Intelligence Research Institute, 2011, *The Ethics of Artificial Intelligence*
- ❑ Granta Innovation, *What is AI, or what's intelligent about machine learning?*
- ❑ Bill Vaughan, 1969, *“To err is human; to really foul things up requires a computer”*
- ❑ House of Lords Select Committee on Artificial Intelligence, 2018, *AI in the UK: ready, willing and able?*



References: Fairness & Bias

- [\(textbook\) FAIRNESS AND MACHINE LEARNING Limitations and Opportunities](#)
- [Fairness, Equality, and Power in Algorithmic Decision-Making](#)
- [Equality of opportunity in supervised learning](#)
- [Fairness Through Awareness](#)
- [Delayed Impact of Fair Machine Learning](#)
- [Learning Fair Representations](#)
- [Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings](#)
- [Learning controllable Fair Representations](#)
- [FACT: A Diagnostic for Group Fairness Trade-offs](#)
- [Right Decisions from Wrong Predictions: A Mechanism Design Alternative to Individual Calibration](#)
- [Retiring Adult: New Datasets for Fair Machine Learning](#)
- [The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning](#)
- [On Fairness and Calibration](#)
- [Calibration for the \(Computationally-Identifiable\) Masses](#)
- [Predicting Good Probabilities With Supervised Learning](#)

