SPIA 586a: Machine Learning for Policy Analysis
Week 11: Fairness

Brandon Stewart

Princeton

April 7–9, 2025

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - privacy

Where We've Been and Where We're Going...

- Last Week
  - privacy
- This Week
  - fairness, racial justice, rule lists

Where We've Been and Where We're Going...

- Last Week
  - privacy
- This Week
  - fairness, racial justice, rule lists
- Next Week
  - interpretability

Where We've Been and Where We're Going...

- Last Week
  - privacy
- This Week
  - fairness, racial justice, rule lists
- Next Week
  - interpretability
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Three Broad Categories of Values

- Privacy

  what information is exposed?
- Fairness

  are gains and costs distributed equitably?
- Interpretability

  do we know what the machine algorithm is doing?

Decision-making is Hard

Decision-making is Hard

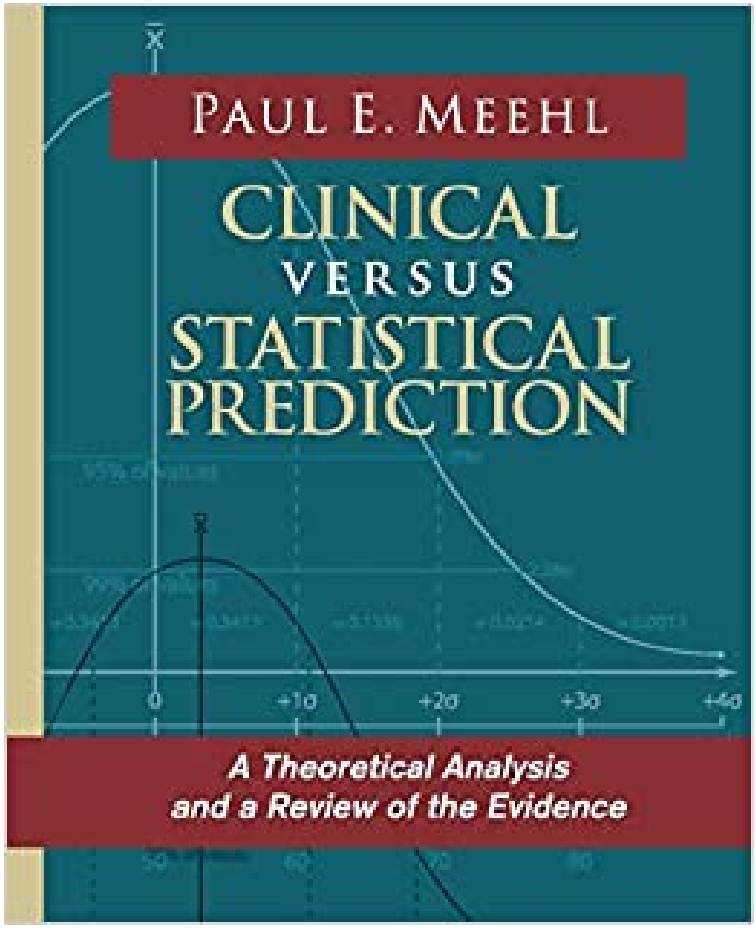
- People are inconsistent in making decisions and have well-studied biases.

Decision-making is Hard

- People are inconsistent in making decisions and have well-studied biases.
- There is an intuitive appeal to a decision-maker who could be more: consistent and possibly more accurate.

Decision-making is Hard

- People are inconsistent in making decisions and have well-studied biases.
- There is an intuitive appeal to a decision-maker who could be more: consistent and possibly more accurate.
- Work going back to the 1950's on the possibility of algorithms as a way of making decisions.



What is an Algorithm?

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.
- In this sense, delegating a task to a person is an algorithm.

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.
- In this sense, delegating a task to a person is an algorithm.
- In practice, we tend to think of an algorithm as a case of taking a set of inputs and rendering an automated (provisional) decision.

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.
- In this sense, delegating a task to a person is an algorithm.
- In practice, we tend to think of an algorithm as a case of taking a set of inputs and rendering an automated (provisional) decision.
- This could be in the form of a simple checklist.

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.
- In this sense, delegating a task to a person is an algorithm.
- In practice, we tend to think of an algorithm as a case of taking a set of inputs and rendering an automated (provisional) decision.
- This could be in the form of a simple checklist.
- Part of the contribution of Meehl (and validated many times since!) is that even very simple scoring algorithms can beat average judgment of experts.

What is an Algorithm?

- An algorithm is just a set of instructions for taking a set of inputs and producing an output.
- In this sense, delegating a task to a person is an algorithm.
- In practice, we tend to think of an algorithm as a case of taking a set of inputs and rendering an automated (provisional) decision.
- This could be in the form of a simple checklist.
- Part of the contribution of Meehl (and validated many times since!) is that even very simple scoring algorithms can beat average judgment of experts.
- Some evidence that experts consider (approximately) the right factors, but are bad at making systematic judgments.

Pretrial Risk Assessments

DANE COUNTY CLERK OF COURTS
Public Safety Assessment – Report

215 S Hamilton St #1000
Madison, WI 53703
Phone: (608) 266-4311

Name: [redacted] **Spillman Name Number:** [redacted]
DOB: [redacted] **Gender:** Male
Arrest Date: 03/25/2017 **PSA Completion Date:** 03/27/2017

New Violent Criminal Activity Flag

No

New Criminal Activity Scale

1	2	3	4	5	6

Failure to Appear Scale

1	2	3	4	5	6

Charge(s):

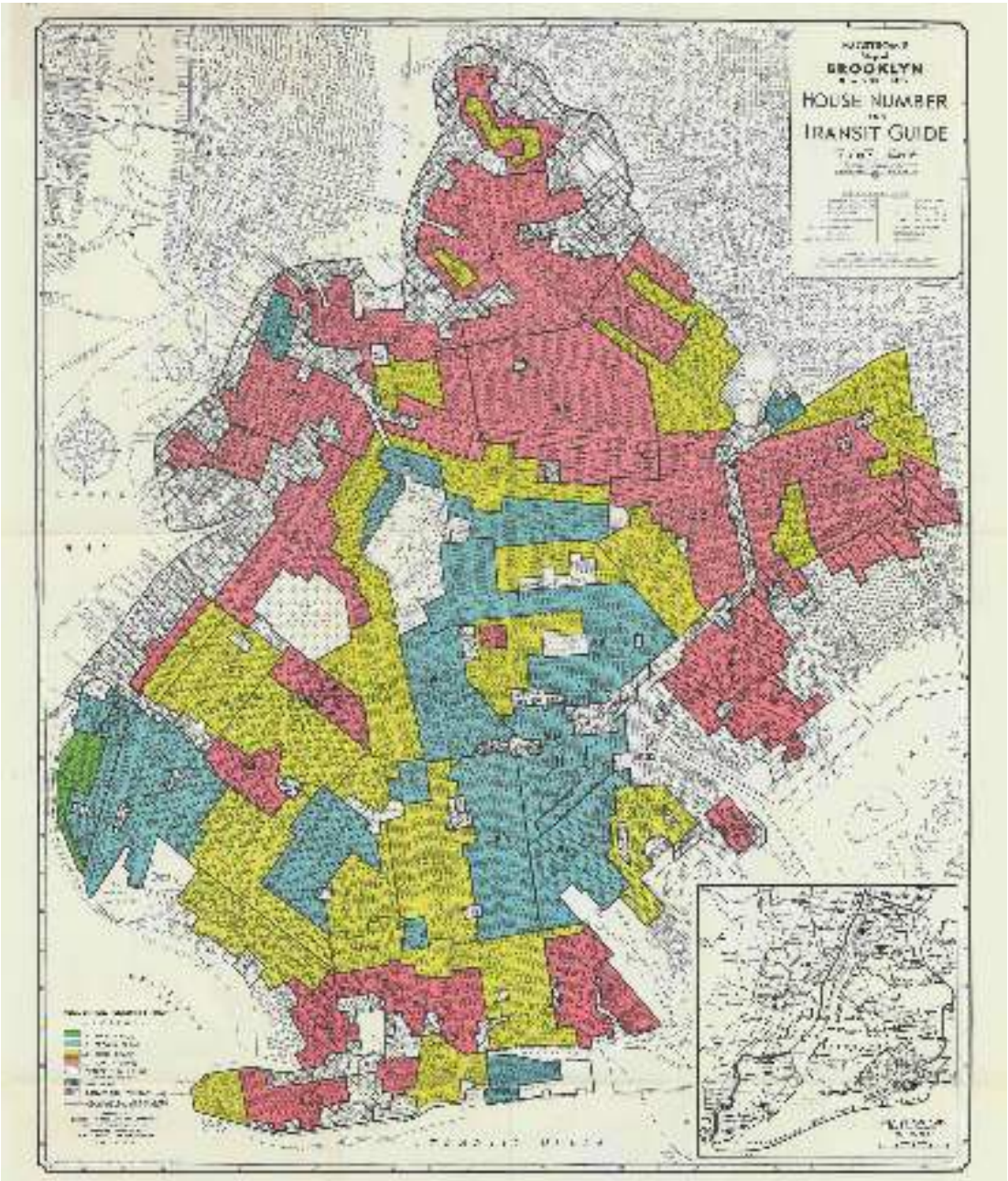
961.41(1)(D)(1) MFC DELIVER HEROIN <3 GMS F 3

Risk Factors:	Responses:
1. Age at Current Arrest	23 or Older
2. Current Violent Offense	No
a. Current Violent Offense & 20 Years Old or Younger	No
3. Pending Charge at the Time of the Offense	No
4. Prior Misdemeanor Conviction	Yes
5. Prior Felony Conviction	Yes
a. Prior Conviction	Yes
6. Prior Violent Conviction	2
7. Prior Failure to Appear Pretrial in Past 2 Years	0
8. Prior Failure to Appear Pretrial Older than 2 Years	Yes
9. Prior Sentence to Incarceration	Yes

Recommendations:

Release Recommendation - Signature bond
Conditions - Report to and comply with pretrial supervision

Redlining



More Complex Algorithms

More Complex Algorithms

- Back to the first day: major transition from rule-based learning to example-based learning.

More Complex Algorithms

- Back to the first day: major transition from rule-based learning to example-based learning.
- The difficulty of learning from examples is that algorithms can learn unintended things from examples and the examples we choose matter a lot!

More Complex Algorithms

- Back to the first day: major transition from rule-based learning to example-based learning.
- The difficulty of learning from examples is that algorithms can learn unintended things from examples and the examples we choose matter a lot!
- The level of abstraction (humans aren't writing the rules!) has caused some people to believe that these approaches should be 'neutral' or 'unbiased'

Concerns


Google
Translate
Spanish
English
French
Turkish - detected
English
Spanish
Turkish
O bir hemşire
O bir doktor
26/5000
She is a nurse
He is a doctor

An Explanation

Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

Image Credit: Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186.

In Computer Vision

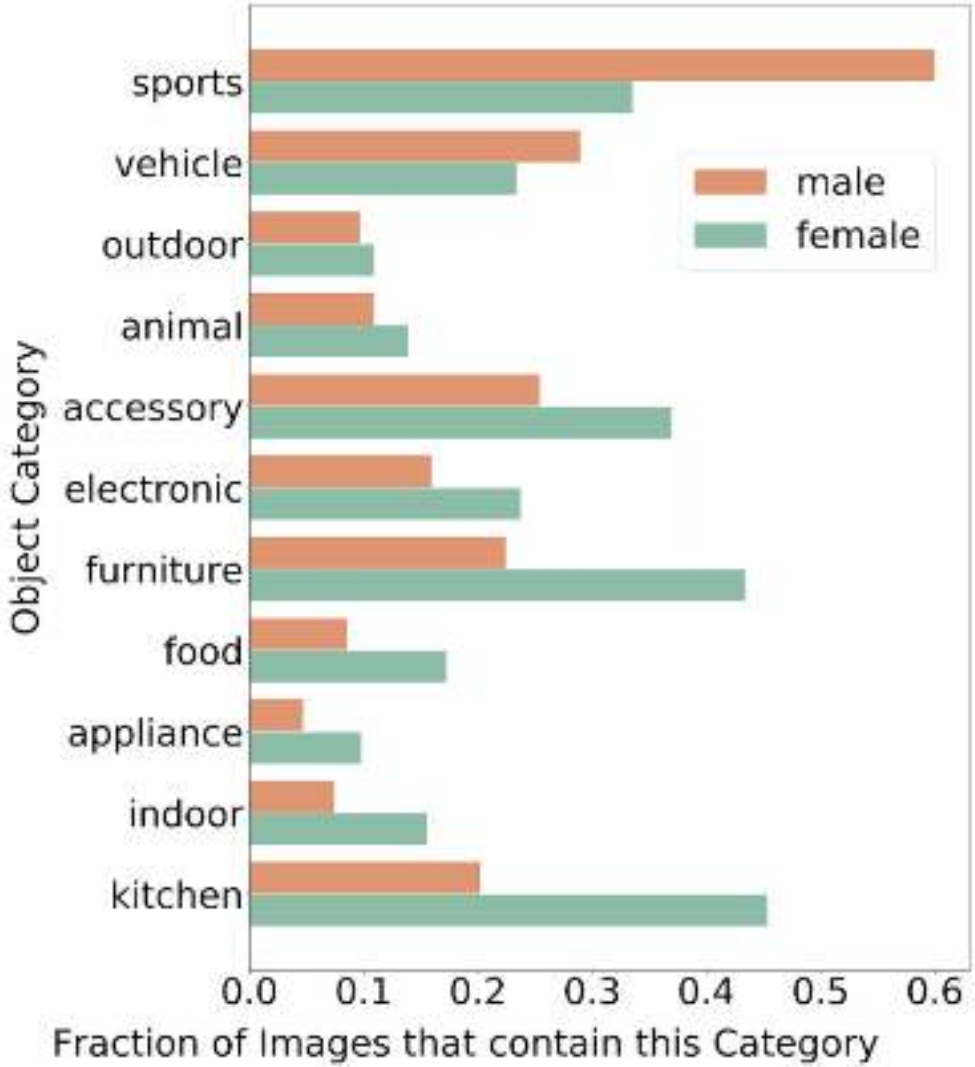


Image Credit: Wang, Angelina, Arvind Narayanan, and Olga Russakovsky. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets." European Conference on Computer Vision. Springer, Cham, 2020.

1. Fairness: Concerns
2. Sources of Concern
3. On the Dangers of Stochastic Parrots
4. Case Study: Censorship of Online Encyclopedias
5. Race After Technology
6. Rule Lists

Sources of Concern

- Clarity about key stages:
  - Task

    how is it measured? what kind of prediction?

Sources of Concern

- Clarity about key stages:
  - Task

    how is it measured? what kind of prediction?
  - Data Source

    does the data reflect what we want?

Sources of Concern

- Clarity about key stages:
  - Task
    how is it measured? what kind of prediction?
  - Data Source
    does the data reflect what we want?
  - Method
    is the method amplifying the bias?

Sources of Concern

- Clarity about key stages:
  - Task
    how is it measured? what kind of prediction?
  - Data Source
    does the data reflect what we want?
  - Method
    is the method amplifying the bias?
- Challenge: fixing imperfect data source with a method is very hard, fixing a bad task is almost impossible.

Sources of Concern

- Clarity about key stages:
  - Task
    how is it measured? what kind of prediction?
  - Data Source
    does the data reflect what we want?
  - Method
    is the method amplifying the bias?
- Challenge: fixing imperfect data source with a method is very hard, fixing a bad task is almost impossible.
- Key component of all of these: who is in the room

Example

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The pre-

We Must Actively Guard Against These Concerns

There is every reason to believe that if we don't actively guard against it, we will reproduce the status quo.

Who Is in the Room Where It Happens

Who Is in the Room Where It Happens

- It matters a lot who is in the room for intrinsic and extrinsic reasons.

Who Is in the Room Where It Happens

- It matters a lot who is in the room for intrinsic and extrinsic reasons.
- Many organizations looking to change the system: Black in AI, Girls Who Code, Queer in AI, LatinX in AI

Who Is in the Room Where It Happens

- It matters a lot who is in the room for intrinsic and extrinsic reasons.
- Many organizations looking to change the system: Black in AI, Girls Who Code, Queer in AI, LatinX in AI
- Organization with special Princeton connections: AI4All founded by Fei-Fei Li, Olga Russakovsky and Rick Sommer.

Who Is in the Room Where It Happens

- It matters a lot who is in the room for intrinsic and extrinsic reasons.
- Many organizations looking to change the system: Black in AI, Girls Who Code, Queer in AI, LatinX in AI
- Organization with special Princeton connections: AI4All founded by Fei-Fei Li, Olga Russakovsky and Rick Sommer.
- Goal: train the next generation of leaders in AI.

"Until this program, I never thought that people who look like me could succeed in computer science and AI."

- AI4ALL 2016 student

Image Credit: Olga Russakovsky and AI4All

1. Fairness: Concerns
2. Sources of Concern
3. On the Dangers of Stochastic Parrots
4. Case Study: Censorship of Online Encyclopedias
5. Race After Technology
6. Rule Lists

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost

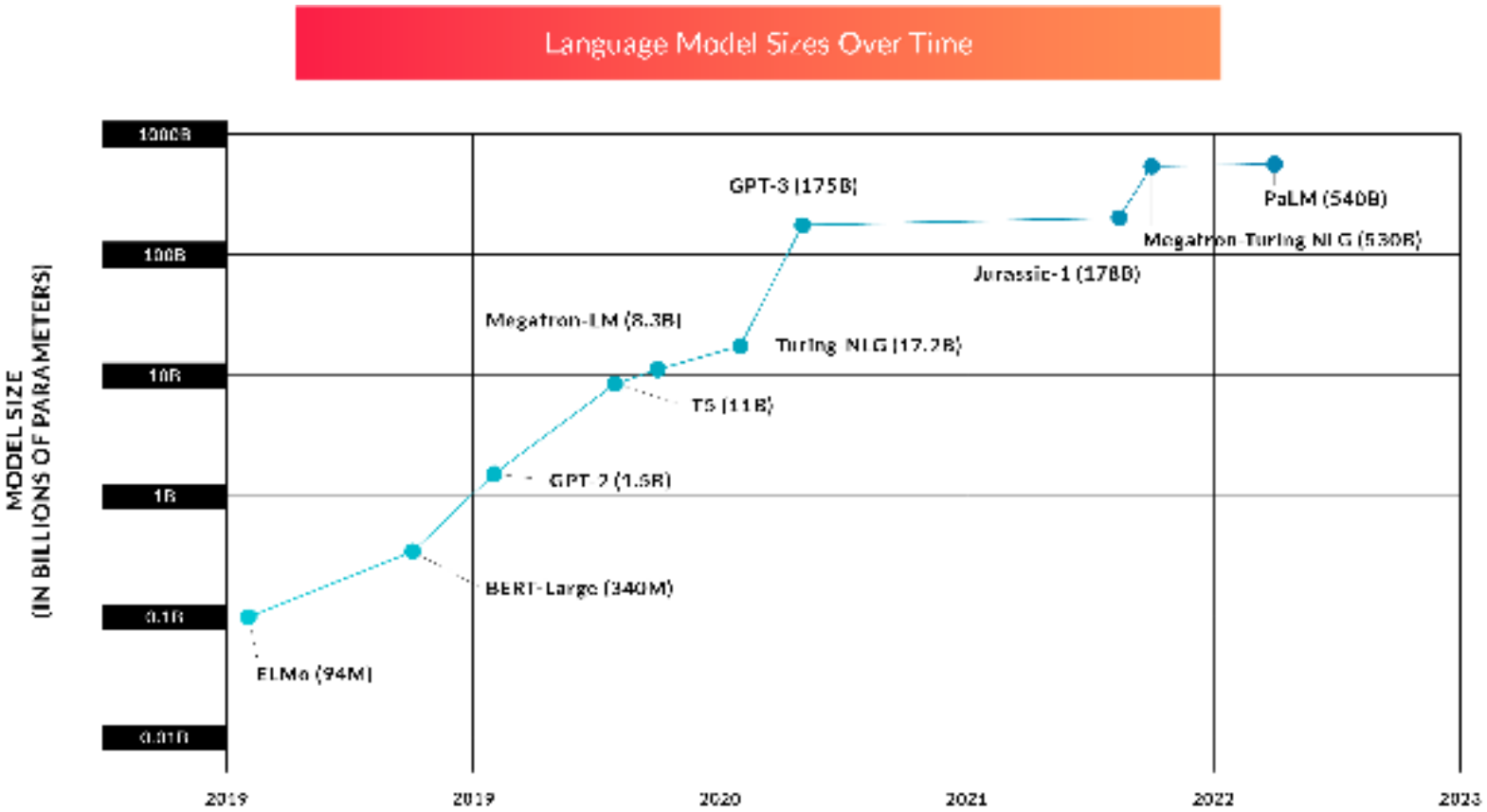
A Complicated Backstory

THE SEATTLE TIMES BUSINESS DEC 3, 2020 4:20 PM

Behind the Paper That Led to a Google Researcher's Firing

Timnit Gebru was one of seven authors on a study that examined prior research on training artificial intelligence models to understand language.

Growing Scale



Bender, Gebru et al. 2021

Bender, Gebru et al. 2021

- Core Ideas:

Bender, Gebru et al. 2021

- Core Ideas:
  - is the growth in scale inevitable or necessary?

Bender, Gebru et al. 2021

- Core Ideas:
  - is the growth in scale inevitable or necessary?
  - what are the risks and who bears them?

Bender, Gebru et al. 2021

- Core Ideas:
  - is the growth in scale inevitable or necessary?
  - what are the risks and who bears them?
- Core risks

Bender, Gebru et al. 2021

- Core Ideas:
  - is the growth in scale inevitable or necessary?
  - what are the risks and who bears them?
- Core risks
  - environmental
  - unmanageable training data
  - harms of use

Environmental Concerns

Environmental Concerns

- Literature kicked off by Emma Strubell et al. 2019: showed large financial and environmental costs associated with compute from these models.

Environmental Concerns

- Literature kicked off by Emma Strubell et al. 2019: showed large financial and environmental costs associated with compute from these models.
- Some degree to which we can quibble with specific numbers, but unambiguous that access is limited and compute energy consumption is non-trivial.

Environmental Concerns

- Literature kicked off by Emma Strubell et al. 2019: showed large financial and environmental costs associated with compute from these models.
- Some degree to which we can quibble with specific numbers, but unambiguous that access is limited and compute energy consumption is non-trivial.
- As in many contexts, energy use is an underpriced externality.

Training Data Concerns

Training Data Concerns

- Data reflects speech on the internet, but who produces that speech?

Training Data Concerns

- Data reflects speech on the internet, but who produces that speech?
- Big $\neq$ Representative and even defining what we want it to be representative of is complicated

Training Data Concerns

- Data reflects speech on the internet, but who produces that speech?
- Big $\neq$ Representative and even defining what we want it to be representative of is complicated
- Conceptually, we could see 'value lock' where we reinforce existing (including problematic) hegemonic views.

Training Data Concerns

- Data reflects speech on the internet, but who produces that speech?
- Big $\neq$ Representative and even defining what we want it to be representative of is complicated
- Conceptually, we could see 'value lock' where we reinforce existing (including problematic) hegemonic views.
- Evaluating bias in language models requires knowing what to look for $\rightsquigarrow$ we need social science.

Training Data Concerns

- Data reflects speech on the internet, but who produces that speech?
- Big $\neq$ Representative and even defining what we want it to be representative of is complicated
- Conceptually, we could see 'value lock' where we reinforce existing (including problematic) hegemonic views.
- Evaluating bias in language models requires knowing what to look for $\rightsquigarrow$ we need social science.
- Documentation debt $\rightsquigarrow$ data and models are 'moving too fast' and 'too big' to be documented, but then the work is never done!

Harms of Use Concerns

Harms of Use Concerns

- Generation of harmful/hateful speech

Harms of Use Concerns

- Generation of harmful/hateful speech
- Release of private information

Harms of Use Concerns

- Generation of harmful/hateful speech
- Release of private information
- Flood of mis/disinformation

Harms of Use Concerns

- Generation of harmful/hateful speech
- Release of private information
- Flood of mis/disinformation
- ...

1. Fairness: Concerns
2. Sources of Concern
3. On the Dangers of Stochastic Parrots
4. Case Study: Censorship of Online Encyclopedias
5. Race After Technology
6. Rule Lists

Censorship of Online Encyclopedias: Implications for NLP Models

Eddie Yang*
z5yang@ucsd.edu
University of California, San Diego
La Jolla, California

Margaret E. Roberts*
meroberts@ucsd.edu
University of California, San Diego
La Jolla, California

ABSTRACT

While artificial intelligence provides the backbone for many tools people use around the world, recent work has brought to attention that the algorithms powering AI are not free of politics, stereotypes, and bias. While most work in this area has focused on the ways in which AI can exacerbate existing inequalities and discrimination, very little work has studied how governments actively shape training data. We describe how censorship has affected the development of Wikipedia corpuses, text data which are regularly used for pre-trained inputs into NLP algorithms. We show that word embeddings trained on Baidu Baike, an online Chinese encyclopedia, have very different associations between adjectives and a range of concepts about democracy, freedom, collective action, equality, and people and historical events in China than its regularly blocked but uncensored counterpart – Chinese language Wikipedia. We examine the implications of these discrepancies by studying their use in downstream AI applications. Our paper shows how government repression, censorship, and self-censorship may impact training data and the applications that draw from them.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Information systems** → **Content analysis and feature selection**; • **Social and professional topics** → **Political speech**.

use daily. NLP impacts how firms provide products to users, content individuals receive through search and social media, and how individuals interact with news and emails. Despite the growing importance of NLP algorithms in shaping our lives, recently scholars, policymakers, and the business community have raised the alarm of how gender and racial biases may be baked into these algorithms. Because they are trained on human data, the algorithms themselves can replicate implicit and explicit human biases and aggravate discrimination [6, 8, 39]. Additionally, training data that over-represents a subset of the population may do a worse job at predicting outcomes for other groups in the population [13]. When these algorithms are used in real world applications, they can perpetuate inequalities and cause real harm.

While most of the work in this area has focused on bias and discrimination, we bring to light another way in which NLP may be affected by the institutions that impact the data that they feed off of. We describe how censorship has affected the development of online encyclopedia corpuses that are often used as training data for NLP algorithms. The Chinese government has regularly blocked Chinese language Wikipedia from operating in China, and mainland Chinese Internet users are more likely to use an alternative Wikipedia-like website, Baidu Baike. The institution of censorship has weakened Chinese language Wikipedia, which is now several times smaller than Baidu Baike, and made Baidu Baike - which is subject to pre-censorship - an attractive source of train-

Thanks for Eddie and Molly for slides that follow!

AI is Political

- AI is not free of politics and bias

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)
  2. gender bias:

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)
  2. gender bias:
     - word embeddings (Caliskan et al, 2017)

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)
  2. gender bias:
     - word embeddings (Caliskan et al, 2017)
     - labeling of images (Zhao et al, 2017)

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP
  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)
  2. gender bias:
     - word embeddings (Caliskan et al, 2017)
     - labeling of images (Zhao et al, 2017)

- They explore another institution that impacts AI:

AI is Political

- AI is not free of politics and bias

  Human bias $\rightarrow$ training data $\rightarrow$ AI $\rightarrow$ bias

- Existing work focuses on racial and gender bias $\rightarrow$ AI/NLP

  1. racial discrimination:
     - ad delivery (Sweeney 2013)
     - speech recognition (Koenecke et al 2020)
     - hate speech detection (Davidson et al 2019)
  2. gender bias:
     - word embeddings (Caliskan et al, 2017)
     - labeling of images (Zhao et al, 2017)

- They explore another institution that impacts AI:

  censorship $\rightarrow$ training data $\rightarrow$ NLP models/applications

Today: Censorship's and AI

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications
  - Entertainment applications

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications
  - Entertainment applications
  - Productivity applications

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications
  - Entertainment applications
  - Productivity applications
  - Algorithmic governance

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications
  - Entertainment applications
  - Productivity applications
  - Algorithmic governance
  - Social media/media curation

Today: Censorship's and AI

- Large user-generated datasets are the building blocks for AI
  - Wikipedia corpuses
  - social media datasets
  - government curated data
- Governments around the world influence these datasets (Roberts 2018):
  1. Fear – threats/laws that create self-censorship
  2. Friction – mandated deletion and blocking
  3. Flooding – coordinated addition of information
- This data is then used in other user-facing applications
  - Entertainment applications
  - Productivity applications
  - Algorithmic governance
  - Social media/media curation
- What implications does censorship have for downstream applications?

Today: Censorship's Implications for AI

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike
- They measure: political word associations

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike
- They measure: political word associations
- They find:

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike
- They measure: political word associations
- They find:
  1. Baidu Baike word embeddings:

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike
- They measure: political word associations
- They find:
  1. Baidu Baike word embeddings:

Today: Censorship's Implications for AI

- They study: censorship of Chinese online encyclopedias and its implications for Chinese language NLP
- They use: word embeddings trained on two Chinese online encyclopedia corpuses (Li et al 2018)
  1. Chinese language Wikipedia
  2. Baidu Baike
- They measure: political word associations
- They find:
  1. Baidu Baike word embeddings:
     democracy $\rightarrow$ bad, CCP and social control $\rightarrow$ good
  2. tangible effect on downstream NLP applications

Censorship of Training Data

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese
  2. Commonly used as training data for NLP

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese
  2. Commonly used as training data for NLP
- Chinese Wikipedia: uncensored, blocked

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese
  2. Commonly used as training data for NLP
- Chinese Wikipedia: uncensored, blocked
- Baidu Baike: censored, unblocked

Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese
  2. Commonly used as training data for NLP
- Chinese Wikipedia: uncensored, blocked
- Baidu Baike: censored, unblocked

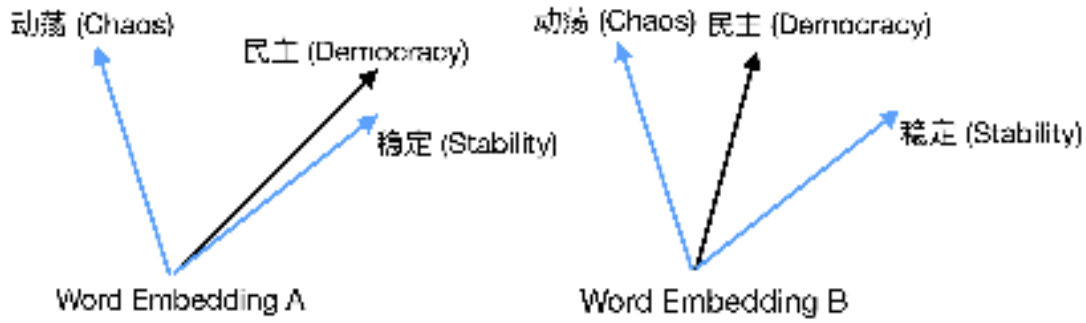
Censorship of Training Data

- Chinese Wikipedia and Baidu Baike:
  1. Both are online encyclopedias in Chinese
  2. Commonly used as training data for NLP
- Chinese Wikipedia: uncensored, blocked
- Baidu Baike: censored, unblocked

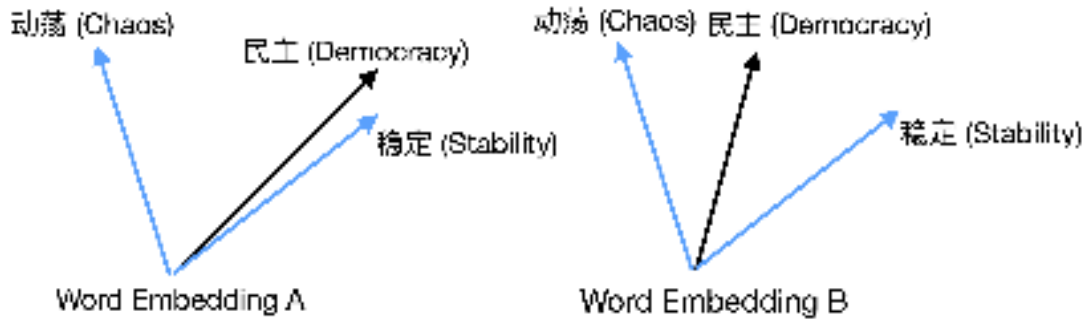
- Baidu Baike many times larger than Chinese Wikipedia $\rightarrow$ increasingly attractive source of training data

1. Comparing Word Embeddings

1. Comparing Word Embeddings



1. Comparing Word Embeddings



- Target words (black):

1. Comparing Word Embeddings

- Target words (black):
  - democratic concepts and ideas:
    - democracy, freedom, election

1. Comparing Word Embeddings

- Target words (black):
  - democratic concepts and ideas:
    - democracy, freedom, election
  - known targets of propaganda:
    - social control, surveillance, collective action, political figures, CCP, historical events

1. Comparing Word Embeddings

- Target words (black):
  - democratic concepts and ideas:
    - democracy, freedom, election
  - known targets of propaganda:
    - social control, surveillance, collective action, political figures, CCP, historical events
- Attribute words (blue):

1. Comparing Word Embeddings

- Target words (black):
  - democratic concepts and ideas:
    - democracy, freedom, election
  - known targets of propaganda:
    - social control, surveillance, collective action, political figures, CCP, historical events
- Attribute words (blue):
  1. "propaganda attributes words"

1. Comparing Word Embeddings

- Target words (black):
  - democratic concepts and ideas:
    - democracy, freedom, election
  - known targets of propaganda:
    - social control, surveillance, collective action, political figures, CCP, historical events
- Attribute words (blue):
  1. "propaganda attributes words"
  2. general evaluative words from ANTUSD

1. Comparing Word Embeddings

Table: Theoretical Expections

Category	Sign
Freedom	−
Democracy	−
Election	−
Collective Action	−
Negative Figures	−
Social Control	+
Surveillance	+
CCP	+
Historical Events	+
Positive Figures	+

1. Comparing Word Embeddings

	Propaganda Words		Evaluative Words	
	effect size	p-value	effect size	p-value
Freedom	-0.62	0.01	0.06	0.60
Democracy	-0.50	0.05	-0.56	0.03
Election	-0.27	0.13	-0.33	0.05
Collective Action	-0.66	0.00	-0.09	0.34
Negative Figures	-0.91	0.00	0.50	0.99
Social Control	0.70	0.04	0.68	0.01
Surveillance	0.09	0.32	0.73	0.00
CCP	1.05	0.02	1.39	0.00
Historical Events	0.14	0.19	0.27	0.01
Positive Figures	0.59	0.00	1.17	0.00

2. Downstream Effect

2. Downstream Effect

- Task: sentiment classification of news headlines

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset
- Test data: over-sample of news headlines containing target words

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset
- Test data: over-sample of news headlines containing target words
- Models: Naive Bayes, SVM, TextCNN

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset
- Test data: over-sample of news headlines containing target words
- Models: Naive Bayes, SVM, TextCNN
- Evaluation metric: classification error

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset
- Test data: over-sample of news headlines containing target words
- Models: Naive Bayes, SVM, TextCNN
- Evaluation metric: classification error
  - Do models using Baidu Baike word embeddings tend to classify headlines with democracy target words more negatively?

2. Downstream Effect

- Task: sentiment classification of news headlines
- Training data: general Chinese news headlines dataset
- Test data: over-sample of news headlines containing target words
- Models: Naive Bayes, SVM, TextCNN
- Evaluation metric: classification error
  - Do models using Baidu Baike word embeddings tend to classify headlines with democracy target words more negatively?
  - Do models using Chinese Wikipedia word embeddings tend to classify headlines with social control target words more negatively?

2. Downstream Effect

Example 1:
"Tsai Ing-wen: Hope Hong Kong Can Enjoy Democracy as Taiwan Does"
Baidu Baike Label: - Wikipedia Label: + Human Label: +

Example 2:
"Cancel Culture Spreading through the Western World, Is It the Fault of Freedom?"
Baidu Baike Label: - Wikipedia Label: + Human Label: -

Examples of Headlines Labeled Differently By Baidu Baike and Wikipedia

2. Downstream Effect

Table: Model Accuracy in Test Set

	Model	Accuracy
Naive Bayes		
	Baidu Baike	76.90
	Wikipedia	76.22
SVM		
	Baidu Baike	77.59
	Wikipedia	77.02
TextCNN		
	Baidu Baike	82.34
	Wikipedia	81.37

2. Downstream Effect

	Naive Bayes		SVM		TextCNN	
	estimate	p-value	estimate	p-value	estimate	p-value
Freedom	-0.13	0.00	-0.06	0.00	-0.04	0.04
Democracy	-0.08	0.00	-0.05	0.04	-0.04	0.06
Election	-0.11	0.00	-0.06	0.03	-0.02	0.48
Collective Action	-0.13	0.00	-0.07	0.00	-0.05	0.01
Negative Figures	-0.04	0.03	0.00	0.96	-0.01	0.54
Social Control	0.03	0.12	0.00	0.93	0.03	0.13
Surveillance	-0.01	0.68	-0.01	0.80	0.00	0.91
CCP	0.03	0.21	0.01	0.65	0.03	0.05
Historical Events	-0.04	0.04	0.01	0.75	-0.02	0.26
Positive Figures	0.06	0.00	0.06	0.00	0.06	0.00

Summary

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:
  1. public opinion monitoring

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:
  1. public opinion monitoring
  2. conversational agents

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:
  1. public opinion monitoring
  2. conversational agents
  3. social media curation

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:
  1. public opinion monitoring
  2. conversational agents
  3. social media curation
- AI $\rightarrow$ automation of politics

Summary

- data reflects the institutional and political context in which it is created and impacts the tools that it powers
  - word embeddings
  - downstream NLP applications
- likely to have effect in a wide array of areas:
  1. public opinion monitoring
  2. conversational agents
  3. social media curation
- AI $\rightarrow$ automation of politics
- hard to de-bias

1. Fairness: Concerns
2. Sources of Concern
3. On the Dangers of Stochastic Parrots
4. Case Study: Censorship of Online Encyclopedias
5. Race After Technology
6. Rule Lists

The New Jim Code

the employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era

- Benjamin 2019 pg 5-6

The New Jim Code

- 'New Jim Crow' implicit legal racism designed to maintain power of explicit legal racism (Jim Crow)

The New Jim Code

- 'New Jim Crow' implicit legal racism designed to maintain power of explicit legal racism (Jim Crow)
- 'New Jim Code' technological racism which reproduces and maintains existing inequalities.

The New Jim Code

- 'New Jim Crow' implicit legal racism designed to maintain power of explicit legal racism (Jim Crow)
- 'New Jim Code' technological racism which reproduces and maintains existing inequalities.
- "The animating force of the New Jim Code is that tech designers encode judgments into technical systems but claim that the racist results of their designs are entirely exterior to the encoding process. Racism thus becomes doubled – magnified and buried under layers of digital denial."

Coded Inequity

Coded Inequity

- Perceptions of tech as:
  - Impartial
  - Personalized
  - Meritocratic
  - Predictive

Coded Inequity

- Perceptions of tech as:
  - Impartial
  - Personalized
  - Meritocratic
  - Predictive
- Contrast between these progressive ideals and coded inequity.

Coded Inequity

- Perceptions of tech as:
  - Impartial: subjective based on access, who is in data
  - Personalized
  - Meritocratic
  - Predictive
- Contrast between these progressive ideals and coded inequity.

Coded Inequity

- Perceptions of tech as:
  - Impartial: subjective based on access, who is in data
  - Personalized: group-based stereotypes
  - Meritocratic
  - Predictive
- Contrast between these progressive ideals and coded inequity.

Coded Inequity

- Perceptions of tech as:
  - Impartial: subjective based on access, who is in data
  - Personalized: group-based stereotypes
  - Meritocratic: 'meritocratic' notations shaped by who has power
  - Predictive
- Contrast between these progressive ideals and coded inequity.

Coded Inequity

- Perceptions of tech as:
  - Impartial: subjective based on access, who is in data
  - Personalized: group-based stereotypes
  - Meritocratic: 'meritocratic' notations shaped by who has power
  - Predictive: can reproduce historical inequalities
- Contrast between these progressive ideals and coded inequity.

Coded Inequity

- Perceptions of tech as:
  - Impartial: subjective based on access, who is in data
  - Personalized: group-based stereotypes
  - Meritocratic: 'meritocratic' notations shaped by who has power
  - Predictive: can reproduce historical inequalities
- Contrast between these progressive ideals and coded inequity.
- "Coded inequity makes discrimination easier, faster, and even harder to challenge, because there is not just a racist boss, banker, or shopkeeper to report."

Private Choices Public Consequences

Private Choices Public Consequences

- Tech industry values: "They are animated by political values influenced strongly by libertarianism, which extols individual autonomy and corporate freedom from government regulation."

Private Choices Public Consequences

- Tech industry values: "They are animated by political values influenced strongly by libertarianism, which extols individual autonomy and corporate freedom from government regulation."
- Democratization of voices, but also provides hate-speech etc. to connect and grow.

Private Choices Public Consequences

- Tech industry values: "They are animated by political values influenced strongly by libertarianism, which extols individual autonomy and corporate freedom from government regulation."
- Democratization of voices, but also provides hate-speech etc. to connect and grow.
- Ubiquity of systems also makes it essentially impossible to opt out of the system, but private industry means that the public doesn't have a voice in governance.

Abolitionist Toolkit

Abolitionist Toolkit

- Using an abolitionist lens to everyday production, deployment, and interpretation of data.

Abolitionist Toolkit

- Using an abolitionist lens to everyday production, deployment, and interpretation of data.
- "Justice, in this sense, is not a static value but an ongoing methodology that can and should be incorporated into tech design."

Abolitionist Toolkit

- Using an abolitionist lens to everyday production, deployment, and interpretation of data.
- "Justice, in this sense, is not a static value but an ongoing methodology that can and should be incorporated into tech design."
- Fairness (as with all our values) is something that we have actively strive for in addition to customary focuses like accuracy. Must be willing to prioritize equity over efficiency.

Abolitionist Toolkit

- Using an abolitionist lens to everyday production, deployment, and interpretation of data.
- "Justice, in this sense, is not a static value but an ongoing methodology that can and should be incorporated into tech design."
- Fairness (as with all our values) is something that we have actively strive for in addition to customary focuses like accuracy. Must be willing to prioritize equity over efficiency.
- Include all stake-holders, not just technologists into the design process.

Fairness Literature

Fairness Literature

- Much broader fairness literature which runs gamut from mathematical theory results to applied work documenting discriminatory system to activism pushing for change.

Fairness Literature

- Much broader fairness literature which runs gamut from mathematical theory results to applied work documenting discriminatory system to activism pushing for change.
- Growing in importance but also many different approaches to how to think about fairness.

Fairness Literature

- Much broader fairness literature which runs gamut from mathematical theory results to applied work documenting discriminatory system to activism pushing for change.
- Growing in importance but also many different approaches to how to think about fairness.
- Formalisms can be helpful, but be sure to keep focus on real people!

1. Fairness: Concerns
2. Sources of Concern
3. On the Dangers of Stochastic Parrots
4. Case Study: Censorship of Online Encyclopedias
5. Race After Technology
6. Rule Lists

Core Idea

Core Idea

Classifier as a list of rules

Core Idea

Classifier as a list of rules

$\rightsquigarrow$ Optimal decision trees.

Interpretability

Interpretability

- Interpretable machine learning models are explicitly constrained to ensure that they:
  a) satisfy domain knowledge
  b) have an easily understandable form

Interpretability

- Interpretable machine learning models are explicitly constrained to ensure that they:
  a) satisfy domain knowledge
  b) have an easily understandable form
- 'Easy to understand' is in the eye of the beholder.

Interpretability

- Interpretable machine learning models are explicitly constrained to ensure that they:
  a) satisfy domain knowledge
  b) have an easily understandable form
- 'Easy to understand' is in the eye of the beholder.
- Simple form helps ensure that everyone understands what is happening and can build trust in high-stakes decisions.

Interpretability

- Interpretable machine learning models are explicitly constrained to ensure that they:
  a) satisfy domain knowledge
  b) have an easily understandable form
- 'Easy to understand' is in the eye of the beholder.
- Simple form helps ensure that everyone understands what is happening and can build trust in high-stakes decisions.
- Interpretable models give users the choice to trust them.

Interpretability

- Interpretable machine learning models are explicitly constrained to ensure that they:
  a) satisfy domain knowledge
  b) have an easily understandable form
- 'Easy to understand' is in the eye of the beholder.
- Simple form helps ensure that everyone understands what is happening and can build trust in high-stakes decisions.
- Interpretable models give users the choice to trust them.
- Key is creating models which don't sacrifice accuracy but are easier to understand.

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref. [33] is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

Slide Credit: 0

Rule Lists

- Rule lists use sequences of ‘or’, ‘and’, ‘if’ and ‘else’ conditions to output predictions.

Rule Lists

- Rule lists use sequences of ‘or’, ‘and’, ‘if’ and ‘else’ conditions to output predictions.
- These are essentially decision trees just reexpressed in a different way.

Rule Lists

- Rule lists use sequences of ‘or’, ‘and’, ‘if’ and ‘else’ conditions to output predictions.
- These are essentially decision trees just reexpressed in a different way.
- Algorithm: Certifiably Optimal Rule Lists(CORELS) which looks for an optimal set of if-then patterns in the data (Angelino et al 2018)

Rule Lists

- Rule lists use sequences of 'or', 'and', 'if' and 'else' conditions to output predictions.
- These are essentially decision trees just reexpressed in a different way.
- Algorithm: Certifiably Optimal Rule Lists(CORELS) which looks for an optimal set of if-then patterns in the data (Angelino et al 2018)
- Avoids problems of previous decision trees being greedy which results in a very difficult optimization/search problem (it could take until end of the universe to enumerate all possibilities!).

Rule Lists

- Rule lists use sequences of 'or', 'and', 'if' and 'else' conditions to output predictions.
- These are essentially decision trees just reexpressed in a different way.
- Algorithm: Certifiably Optimal Rule Lists(CORELS) which looks for an optimal set of if-then patterns in the data (Angelino et al 2018)
- Avoids problems of previous decision trees being greedy which results in a very difficult optimization/search problem (it could take until end of the universe to enumerate all possibilities!).
- "We would like models that look like they are created by hand, but they need to be accurate, full-blown ML models" (Rudin 2019)

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

This model from ref. [33] is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

Slide Credit: 0

Objective Follows a Similar Structure

Objective Follows a Similar Structure

$$\arg\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{\text{obs i is misclassified}} + \lambda \text{size}(f) \right)$$

- size(f) is number of rules

Objective Follows a Similar Structure

$$\arg\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{\text{obs i is misclassified}} + \lambda \text{size}(f) \right)$$

- size(f) is number of rules
- λ is percent training accuracy sacrificed to reduce size of model by one.

Other Styles of Systems (Scoring)

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	⋯
2.	Prior arrests ≥ 5	1 point	+⋯
3.	Prior arrests for local ordinance	1 point	+⋯
4.	Age at release between 18 to 24	1 point	+⋯
5.	Age at release ≥ 40	−1 point	+⋯
		Score	= ⋯

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [2], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

Slide Credit: 0

Other Styles of Systems (Falling Rule Lists)

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density $\geq$ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age $\geq$ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Table 1: Falling rule list for mammographic mass dataset.

Image Credit: Wang and Rudin 2015

Summary

Summary

- Rule lists use discrete optimization to provide more optimal decision trees.

Summary

- Rule lists use discrete optimization to provide more optimal decision trees.
- CORELS as a nice algorithm with great software implementations in many languages (including R!)

Summary

- Rule lists use discrete optimization to provide more optimal decision trees.
- CORELS as a nice algorithm with great software implementations in many languages (including R!)

Going Deeper:

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*.

This Week in Review

This Week in Review

- Fairness and Racial Justice

This Week in Review

- Fairness and Racial Justice

Next:

- Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*, 1(5), 206-215.