Combating Bias in AI/ML Applications

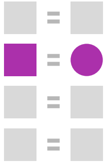
10x Phase 3 Funding Request

xD: U.S. Census Bureau
https://www.xd.gov

10x

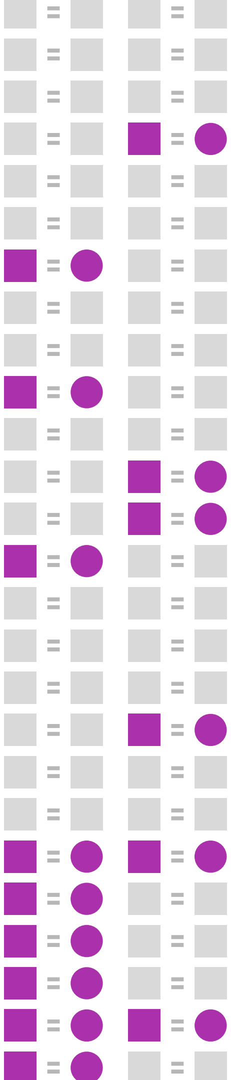
THE CHALLENGE

Data collected by humans in the real world is imperfect.



THE CHALLENGE

When we train algorithms on imperfect data, tiny imperfections can lead to huge systematic inaccuracies.



THE CHALLENGE

We call these systematic inaccuracies **bias**

THE CHALLENGE

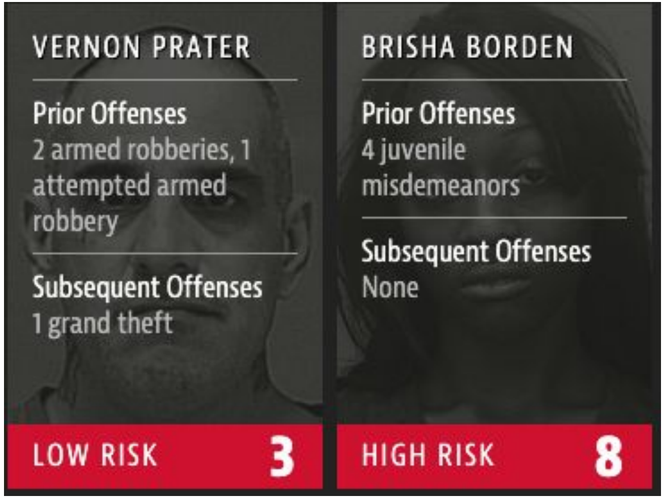
What happens when bias is present in data used by machine learning algorithms to automate decisions and predictions?

THE CHALLENGE: EXAMPLE

COMPAS Recidivism Algorithm

Automated recidivism prediction tool gave harsher sentencing for black defendants who ultimately never committed serious crimes. White defendants were predicted to be less at risk of recidivism than they actually were. Thousands of black defendants were given unwarranted sentences.

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

THE CHALLENGE: CURRENT COSTS

30%

of data scientist time, or about **500 hours/year** per FTE spent on data bias and cleaning issues*

https://hbr.org/2012/09/whos-really-using-big-data, Harvard Business Review, 2012

THE CHALLENGE: FUTURE COSTS

$142.2B in federal research & development in FY 2021

$2B in projected AI spending by agencies 2022

https://insight.ieeeusa.org/articles/fy-2021-rd-budget-proposal/, IEEE USA, 2021

COMBATING BIAS IN AI/ML APPLICATIONS

Key Findings

SUMMARY OF WORK

50+

Subject matter experts and potential users interviewed

7

workshops hosted to identify user needs and solicit feedback on prototypes

5

prototypes developed to solve for common issues of bias in government

30+

resources on machine learning, bias, and AI in government annotated for distribution

KEY AUDIENCES

Technical Audience

Validated the need for technical users to address bias in their data products.

Data scientists, statisticians, and machine learning engineers see bias as a significant problem.

Non-technical Audience

Validated the need for PMs to understand and address bias.

Project/product managers own risk mitigation for technical products, and bias in ML represents a major risk.

KEY LEARNINGS

Many of our research subjects wanted to better understand bias but didn’t know where to start. We believe their needs can be best met through:

Modular + Reusable Solutions

Publish Jupyter notebooks with customizable code

Reduce Burden of Choice

Guidance what approaches one should take, what algorithmic choices to make

Customize tools for government with features not seen in industrial solutions

Curated Library of Resources

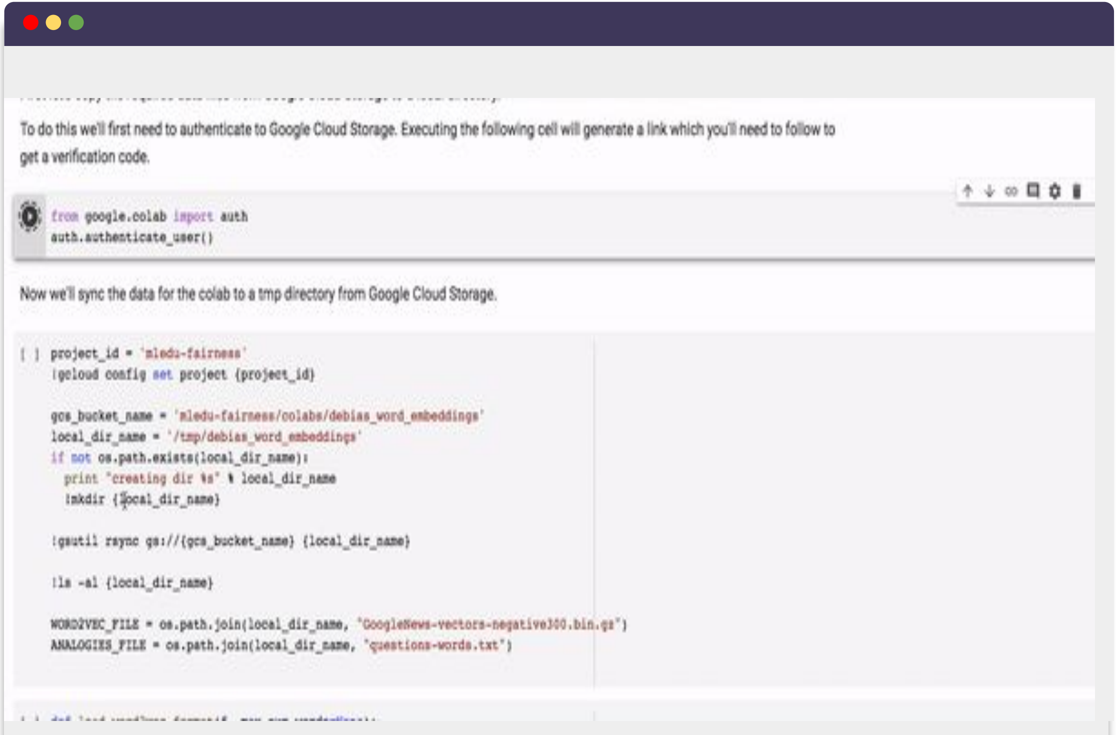
Papers, government papers, etc. with annotations that make it easy for technical and non-technical audiences to learn

Upskilling resources for established government training programs (e.g., Census DS)

COMBATING BIAS IN AI/ML APPLICATIONS

Solutions + Future State

SOLUTION: PROTOTYPE REUSABLE + MODULAR CODE



To do this we'll first need to authenticate to Google Cloud Storage. Executing the following cell will generate a link which you'll need to follow to get a verification code.

```
from google.colab import auth
auth.authenticate_user()
```

Now we'll sync the data for the colab to a tmp directory from Google Cloud Storage.

```
[ ] project_id = 'mledu-fairness'
    !gcloud config set project {project_id}

    gcs_bucket_name = 'mledu-fairness/colabs/debias_word_embeddings'
    local_dir_name = '/tmp/debias_word_embeddings'
    if not os.path.exists(local_dir_name):
      print "creating dir %s" % local_dir_name
      !mkdir {local_dir_name}

    !gsutil rsync gs://{gcs_bucket_name} {local_dir_name}

    !ls -al {local_dir_name}

    WORD2VEC_FILE = os.path.join(local_dir_name, "GoogleNews-vectors-negative300.bin.gz")
    ANALOGIES_FILE = os.path.join(local_dir_name, "questions-words.txt")
```

FEATURED NOTEBOOKS

- NLP Task: Bias in applications tutorial
- Coding Task: Supervised classification notebook
- Adversarial methods notebook
- Separating fraud from bias
- Synthetic data generation approach to bias mitigation

SOLUTION: COLLECTION OF OPEN-SOURCE MATERIALS

Government White Papers

Essentials of Bias in ML

Overview of bias in ML/AI

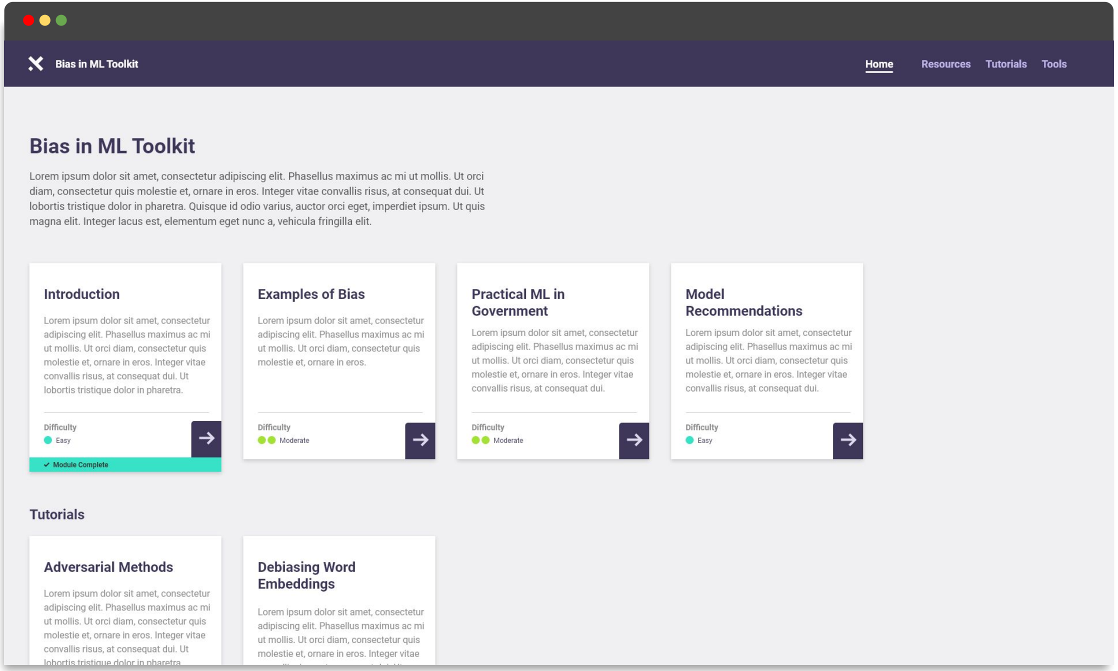
Practical ML and Bias in government

Tool Overview: Bias in AI/ML

Datasets and further Reading

Note: Above Icons courtesy 18F 10x presentation

FUTURE STATE: BRINGING IT ALL TOGETHER



Bias in ML Toolkit

Bias in ML Toolkit

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae convallis risus, at consequat dui. Ut lobortis tristique dolor in pharetra. Quisque id odio varius, auctor orci eget, imperdiet ipsum. Ut quis magna elit. Integer lacus est, elementum eget nunc a, vehicula fringilla elit.

Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae convallis risus, at consequat dui. Ut lobortis tristique dolor in pharetra.

Difficulty
● Easy

✓ Module Complete

Examples of Bias

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros.

Difficulty
●● Moderate

Practical ML in Government

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae convallis risus, at consequat dui.

Difficulty
●● Moderate

Model Recommendations

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae convallis risus, at consequat dui.

Difficulty
● Easy

Tutorials

Adversarial Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae convallis risus, at consequat dui. Ut lobortis tristique dolor in pharetra

Debiasing Word Embeddings

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus maximus ac mi ut mollis. Ut orci diam, consectetur quis molestie et, ornare in eros. Integer vitae

EMPHASIS ON UX

- A simple, easy-to-use interface for organizing content
- Enable bias tools to be executed within the website
- Simplify decision-making through step-by-step model recommendation
- Focus on implementation and tooling for government context

FUTURE STATE: DETAILS

- **Build a user-friendly library** of tools and resources that addresses needs of technical and non-technical users.
- **Expand range of notebooks** to address additional common tasks in government
- **Build automated no-code tools** to detect bias in datasets and models

COMBATING BIAS IN AI/ML APPLICATIONS

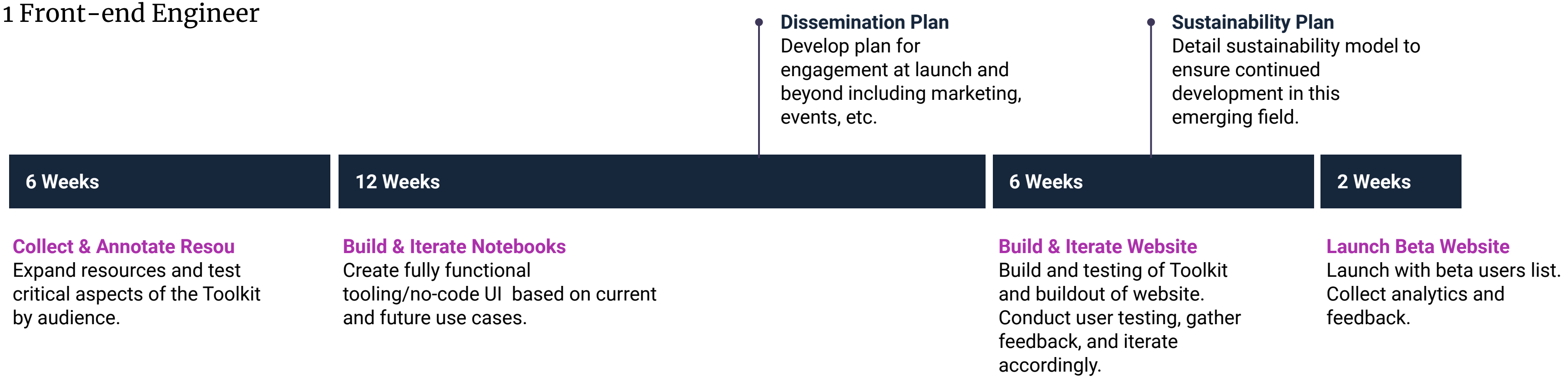
Phase 3 Plan

PHASE 3 ROADMAP

1 Project Manager

2 ML Engineers/ Data Scientists

1 Front-end Engineer



Dissemination Plan
Develop plan for engagement at launch and beyond including marketing, events, etc.

Sustainability Plan
Detail sustainability model to ensure continued development in this emerging field.

6 Weeks

Collect & Annotate Resou
Expand resources and test critical aspects of the Toolkit by audience.

12 Weeks

Build & Iterate Notebooks
Create fully functional tooling/no-code UI based on current and future use cases.

6 Weeks

Build & Iterate Website
Build and testing of Toolkit and buildout of website. Conduct user testing, gather feedback, and iterate accordingly.

2 Weeks

Launch Beta Website
Launch with beta users list. Collect analytics and feedback.

KEY ASSUMPTIONS/RISKS

✘ RISK	✓ RESPONSE
Toolkits exist in the private sector.	Outside resources do not meet the full needs of government users or use cases or effectively accommodate their significant technical limitations.
Users lack incentive to expose bias in their data or models.	Explore institutionalizing bias mitigation and auditing work at federal agencies, and engaging with scientific agencies as early adopters. Incorporation into training/upskilling programs/employee education initiatives.
Bias is meant to be reduced, not completely eliminated.	Use examples to show that even improvements in an imperfect dataset has significant financial and reputational effects; our project will help create a cultural norm around addressing bias in ML applications in government
This field is evolving quickly	Leverage deep connections with academic and research partners to keep abreast of developments in the rapidly-evolving field

IT TAKES A VILLAGE

Technical Partners

MIT

Carnegie Mellon

Dissemination Partners

TTS AI COE/COP

Georgetown

VA (Pilot partner)

GAO

Training Partners

Data Science Advisory Council (OPM/Census)

Data Science Users Group (Bureau of Labor Statistics)

COLLABORATION (TTS AI COE/COP)

40+

Early Adopters: Partner on expressed need from Ethics Working Group to deliver toolkit prototype and engage the group as key early adopters in build out.

1

Roadmap: Use this moment to pilot/develop roadmap of larger engagement strategy for Community of Practice/Working Groups

3

Case Studies: Produce a set of Case Studies on how agencies are leveraging TTS AI COE/COP and the Bias Toolkit to combat bias in their work.

30+

Connected Resources: Link our Bias Toolkit and curated resources from TTS AI's planned central AI/ML repository

PHASE 3 SUCCESS

We're excited to get critical tools and resources in the hands of those that need them to address bias in AI.

MVP & Feedback

A fully functional MVP of a Bias Toolkit that helps users mitigate bias in government data and algorithms

Delivery & Adoption

An engaged group of early adopters that will produce case studies for how they are mitigating bias in their work. A clear indication of potential for increased adoption.

Sustained Engagement

How will this toolkit evolve in the future as AI matures? Who should be involved? What new tools and resources might be most valuable to our users?

POTENTIAL IMPACT

Significant Cost Savings

- Huge savings in the time of highly-skilled FTEs
- Significant savings in not having to change datasets and applications retroactively

Increased Build Integrity

- As outlined in the OMB's Guidance on AI, the government must instill confidence in the public about its use of AI
- Auditing capabilities can also increase the trust of federal employees in externally-built acquired products

Building Public Trust

- Ensure fairness in data-driven policy outcomes
- Vastly increase public confidence in the government, which has been downtrending

Increased Confidence in Datasets

- Auditing mechanisms would secure the integrity of datasets released to academic partners and research institutions
- Federal government datasets form the basis of millions of downloads and millions of dollars in research funding each year, but many datasets have been found to have significant biases and quality issues

COMBATING BIAS IN AI/ML APPLICATIONS

Thank You!

10x

Appendix

EXAMPLES

Amazon

- Hiring algorithms discriminated significantly against applicants whose resumes mentioned “Women’s” activities
- ML trained on data that largely excluded women, and thus learned to exclude female-identified applicants

Department of Education

- School district matching algorithm used across the country found to be systematically biased in placing minority students

Kentucky PD

- Gaps in open 311 and police datasets in neighborhoods that a predominantly black versus predominantly white influence models built for civic applications

Customs and Border Protection

- Facial recognition and risk assessment done using externally-acquired software experienced systematic failures that could not be explained because of black-box nature of software

Citation, Citation,

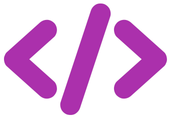
ICON ATTRIBUTION

Alice Design - Noun Project

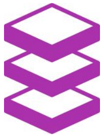
Srinivas Agra - Noun Project

Les vieux garçons - Noun Project

Tomas Knopp - Noun Project



Shashank Singh - Noun Project



Noun Project

Citation, Citation,

SOLUTION: BIAS MITIGATION TECHNIQUES

Dataset debiasing:

Generating synthetic data to mimic “good” data your data loses after you discard bad data

Speeded-up testing of all possible subpopulations in a dataset for imbalances

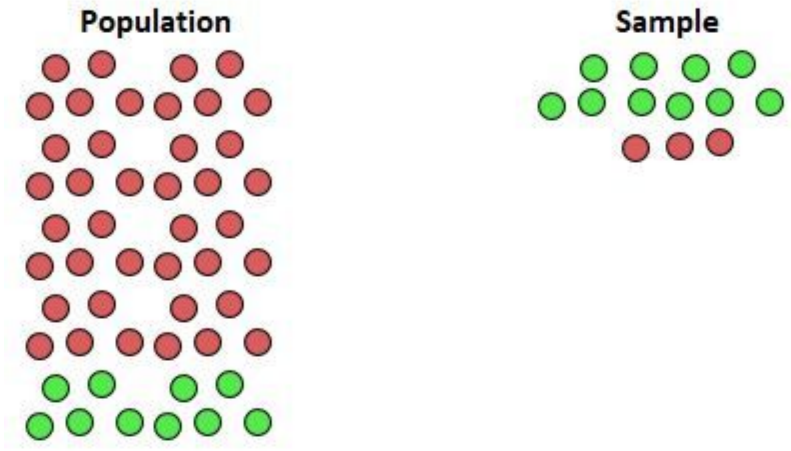
In-process model debiasing:

Protecting data components from discovery by an adversary ensures consistency

Post-process model debiasing:

Deleting malignant outliers post-hoc

The “shape” of bias (sampling bias in training data)



Source: Zach Bobbitt, *Statology*

Combating Bias in AI/ML Applications

10x Phase 3 Funding Request

xD: U.S. Census Bureau
https://www.xd.gov

10x