Are we experimenting on people?

Data Science and AI ethics

NightingaleHQ

CEO

Data Scientist

Microsoft AI & Data Platform MVP

@TheStephLocke

steph@NightingaleHQ.ai

Steph Locke

Nightingale HQ helps businesses develop their strategy, culture and skills for successful AI adoption.

AI Direct

AI is both a risk and an opportunity for your company. Our practical manager's toolkit will help you develop an AI strategy that aligns with your business objectives.

AI Learn

Build core AI competencies inhouse and across-functions with our accelerated training. We deliver a range of online and in-class masterclasses and bootcamps.

AI Connect

Engage with world class data science practitioners to help you deliver AI projects. We take the pain out of contract management and finding the right expertise.

NightingaleHQ

Your business.
Your people.
AI Ready.

NightingaleHQ.ai

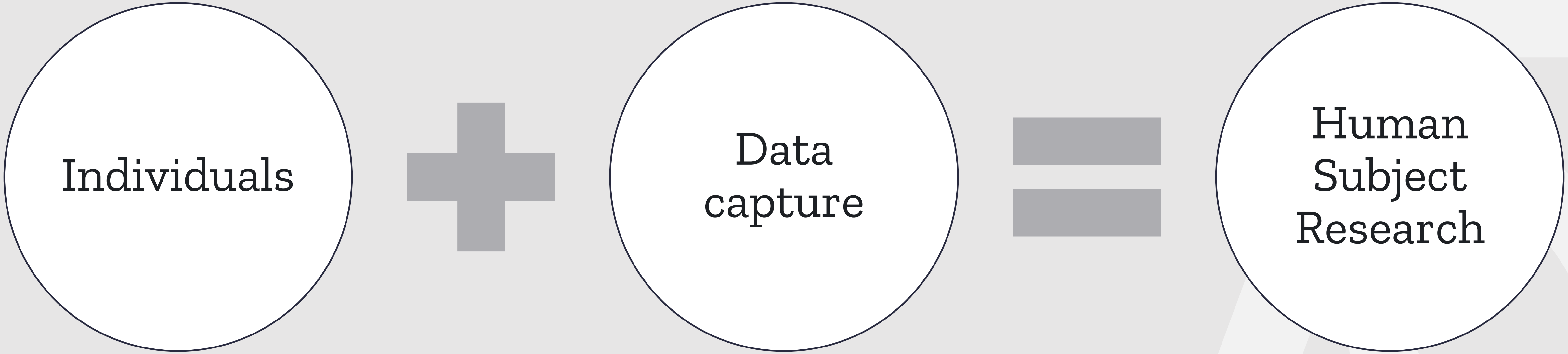
nightingalehqai

nightingale-hq

nightingalehq

- Are we experimenting on people?
- What are our ethical obligations?
- How do we embed ethics into our processes?
- How do we tackle the technical challenges?

Agenda



Human Subject Research

Churn

PPC optimisation

Credit worthiness

Recommendations

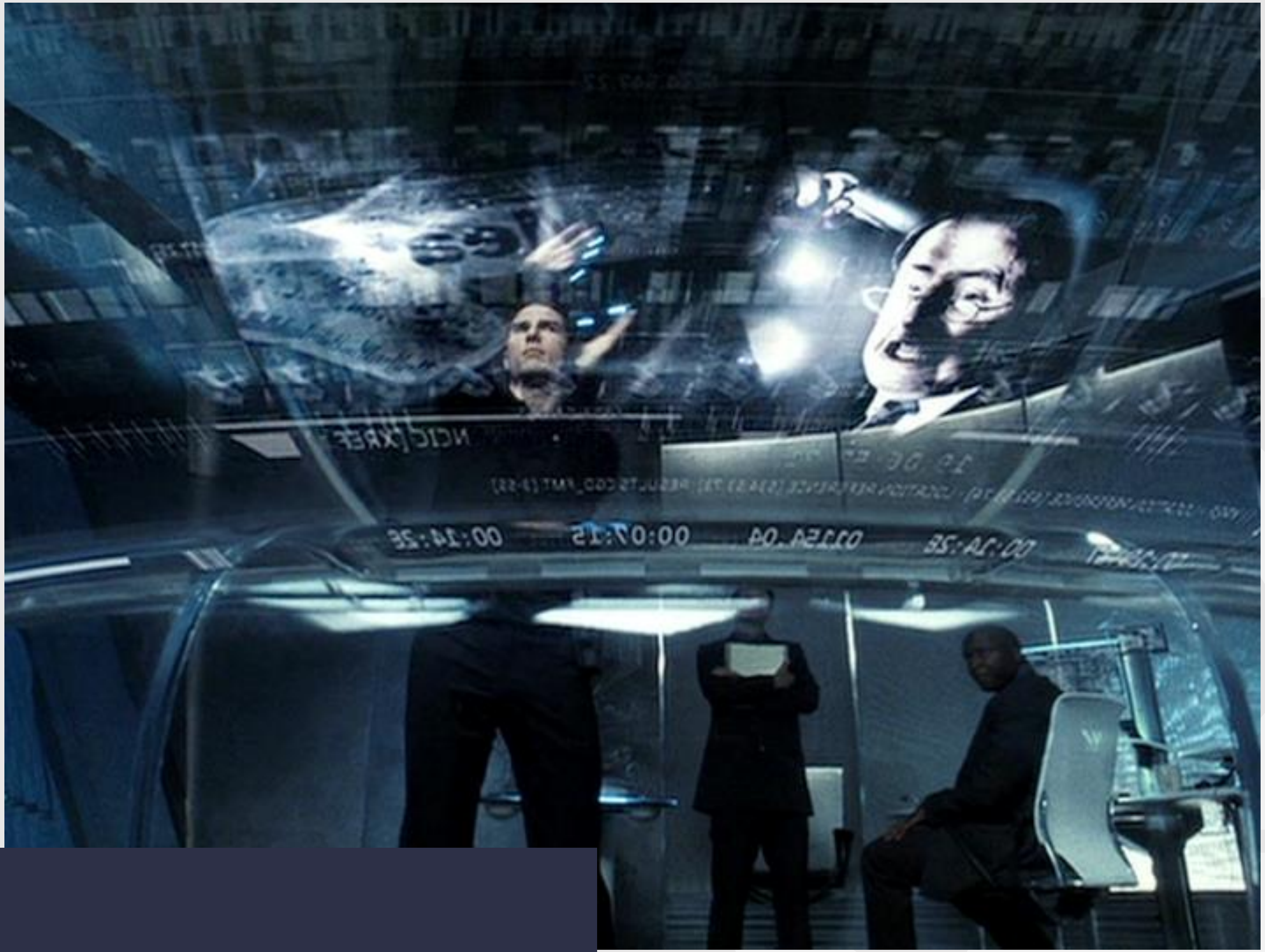
Timeline ordering

Common data science projects

- Informed consent
- Respect
- Right to opt-out
- Benefits > costs
- No harm
- Provide information
- Protect privacy

Researcher obligations

Why do we need to think about this?



To avoid a dystopian future

Chukwuemeka Afigbo
@nke_ise

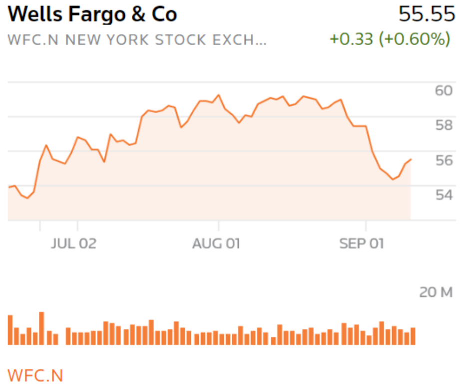
Follow

If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video

To not be (inadvertently) ______ist

Tweet

The $8 million accrual is intended for roughly 625 borrowers who should have qualified for a loan modification under a program the Treasury Department set up in 2009 to help Americans who were struggling to make mortgage payments.



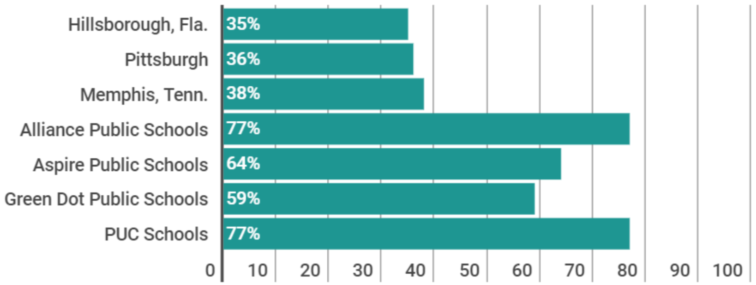
An error in Wells Fargo's underwriting tool improperly excluded those borrowers, 400 of whom eventually had their homes foreclosed upon, the bank said.

The bank also updated disclosures on issues it has discovered in auto lending, wealth management, fiduciary and custody accounts, foreign exchange trading, mortgage rate-lock extensions, "add-on" products like identity theft protection, and frozen or closed bank accounts.

Teachers' Perceptions of Evaluation Systems

In spring 2016, researchers asked teachers whether they thought the consequences tied to evaluation results are reasonable, fair, and appropriate. Consequences can include possible termination or increased compensation.

Teachers in charter schools were much more likely to believe the high-stakes consequences were fair.

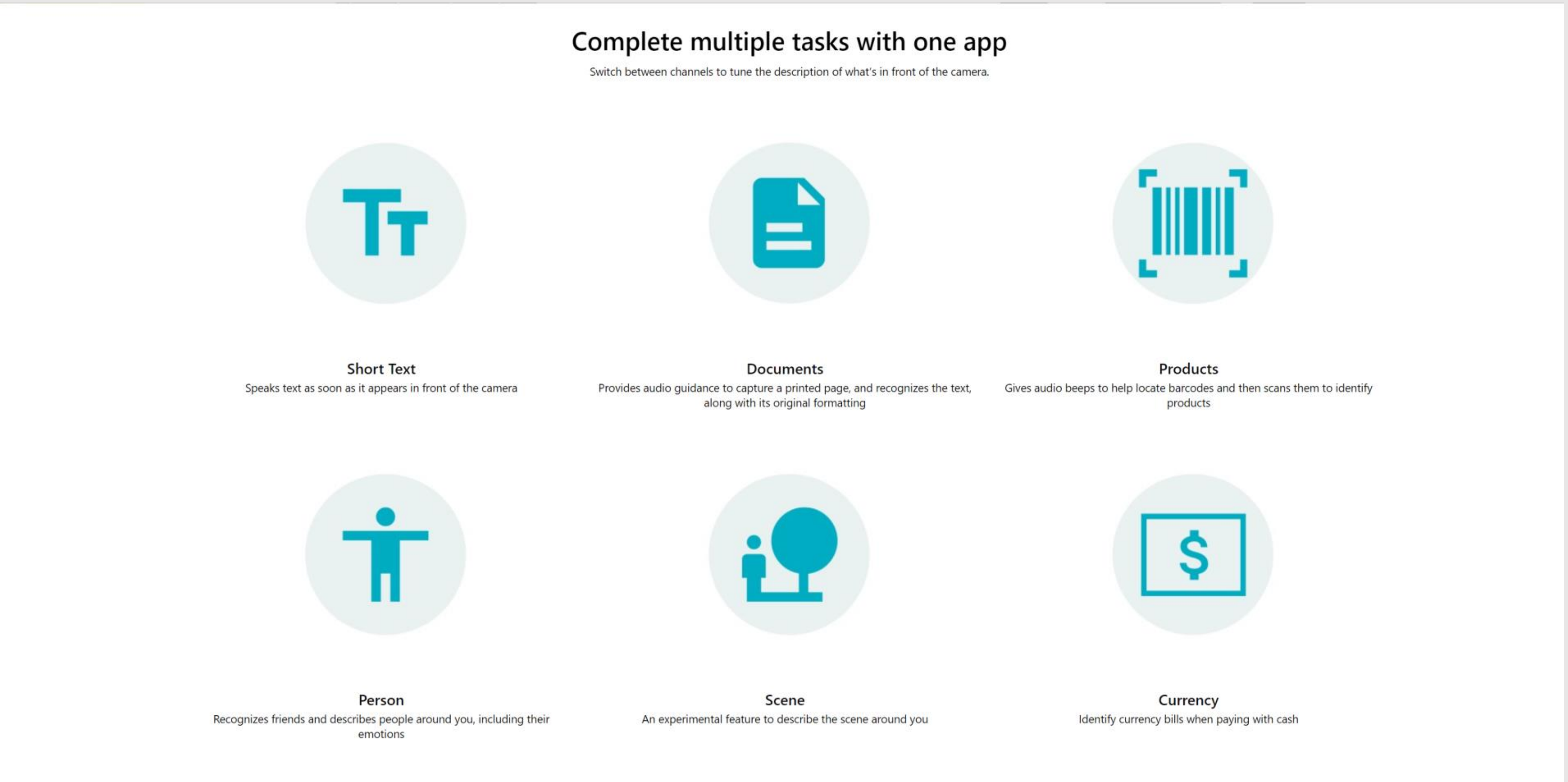


SOURCE: RAND Corporation and the American Institutes for Research

EDUCATION WEEK

To avoid destroying lives

https://money.cnn.com/2018/08/04/news/companies/wells-fargo-mortgage-modification/index.html
https://www.edweek.org/ew/articles/2018/06/21/an-expensive-experiment-gates-teacher-effectiveness-program-show.html



To feel good about what we do

https://www.microsoft.com/en-us/seeing-ai

How does this translate to a business context?

- Type I and II errors
- Change in behaviours
- Financial, medical, or political impacts

(Unintended) Consequences

- Consider and plan
- Guide implementation
- Refuse to do harmful work and/or whistleblow?
- Determine monitoring to identify potential harm

The responsibilities of the (data) scientist

- Legal
- Shareholders
- Be customer-centric?

Responsibilities of the company

What do we need to do?

- Communicate about ethics internally
- Run workshops to help understand implications *scu.edu/ethics-in-technology-practice*
- Share stories about what data scientists fail to be ethical

Education

- Initial impact assessments
- Plans and checklists
- Test strategies and harnesses
- ainowinstitute.org/aiareport2018.pdf
- www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

Frameworks

- **group unaware** - same cutoff points / decision boundary
- **group thresholds** - different cutoff points to allow different volumes in
- **demographic parity** - different cutoff points to end up with distribution like overall demographics
- **equal opportunity** - the same true positive rate holds across groups
- **equal accuracy** - the same overall accuracy rate holds across groups

5 concepts of fairness

- Legal and contractual requirements
- Anonymisation and maintaining privacy
- Consultation processes

Data

Show me the tech!

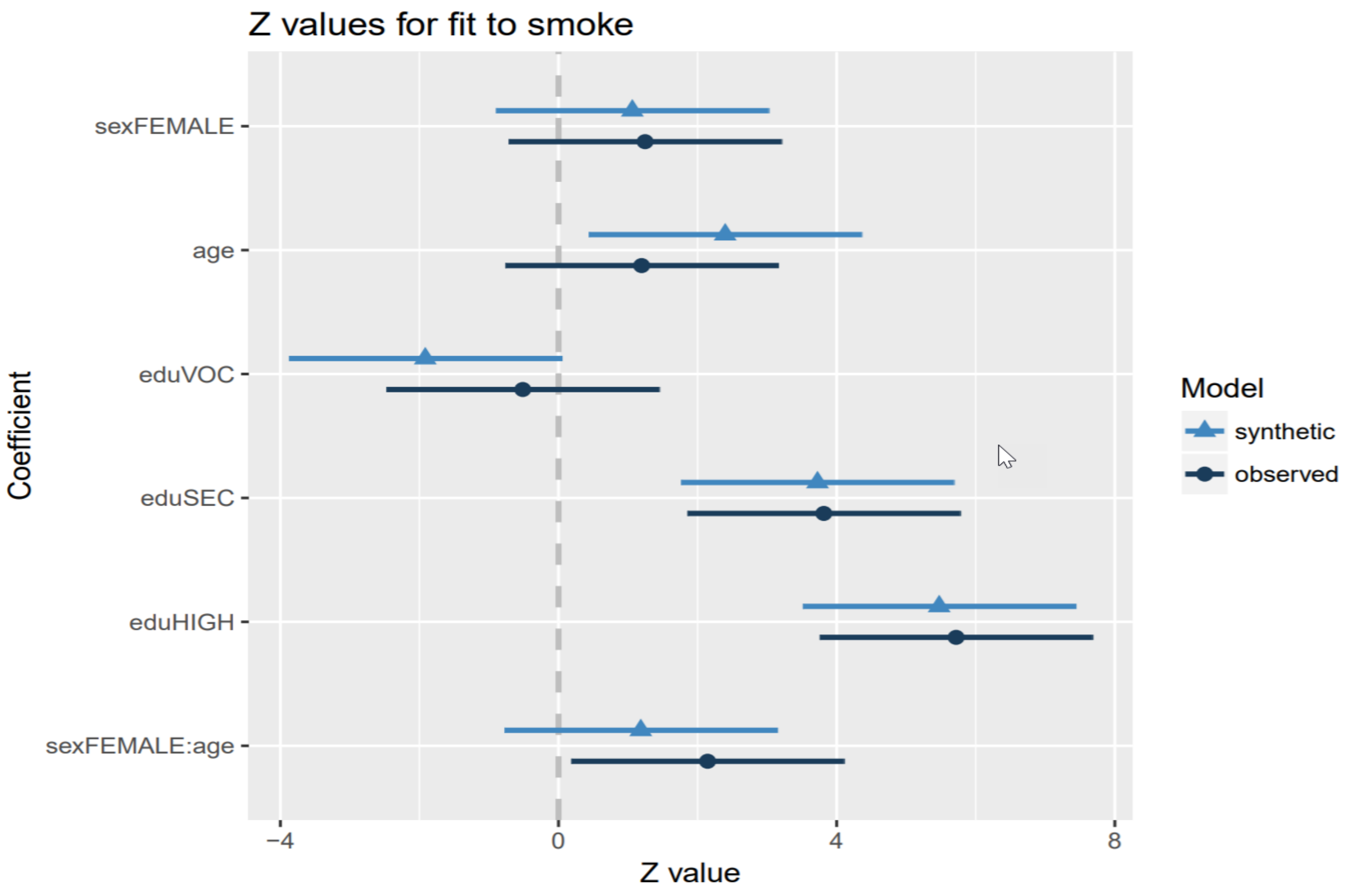


Figure 1: Comparison of intervals for fit f8 and original data.

Synthetic data

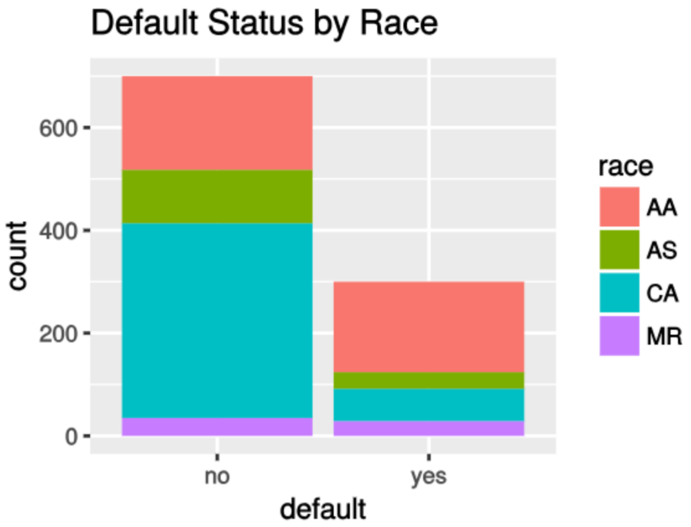
https://cran.r-project.org/package=synthpop

As you can see in the histogram below, the majority of our data set is Caucasian (**CA**) at 44.2%, closely followed by African American (**AA**) at 35.8%, and then Asian American (**AS**) at 13.6% and Hispanic or Mixed Race (**MR**) at 6.4%.

In [8]:
```
table(credit$race, credit$default) / 10

ggplot(credit, aes(default)) + geom_bar(aes(fill = race)) + ggtitle("Default Status by Race")
```

```
     no  yes
AA 18.2 17.6
AS 10.4  3.2
CA 37.9  6.3
MR  3.5  2.9
```



EDA

https://github.com/ropenscilabs/proxy-bias-vignette/blob/master/EthicalMachineLearning.ipynb

- Functions for fairness concepts
- Analysis by groups
- Pen portrait / “Typical” entries

Testing

FairML: Auditing Black-Box Predictive Models

FairML is a python toolbox auditing the machine learning models for bias.

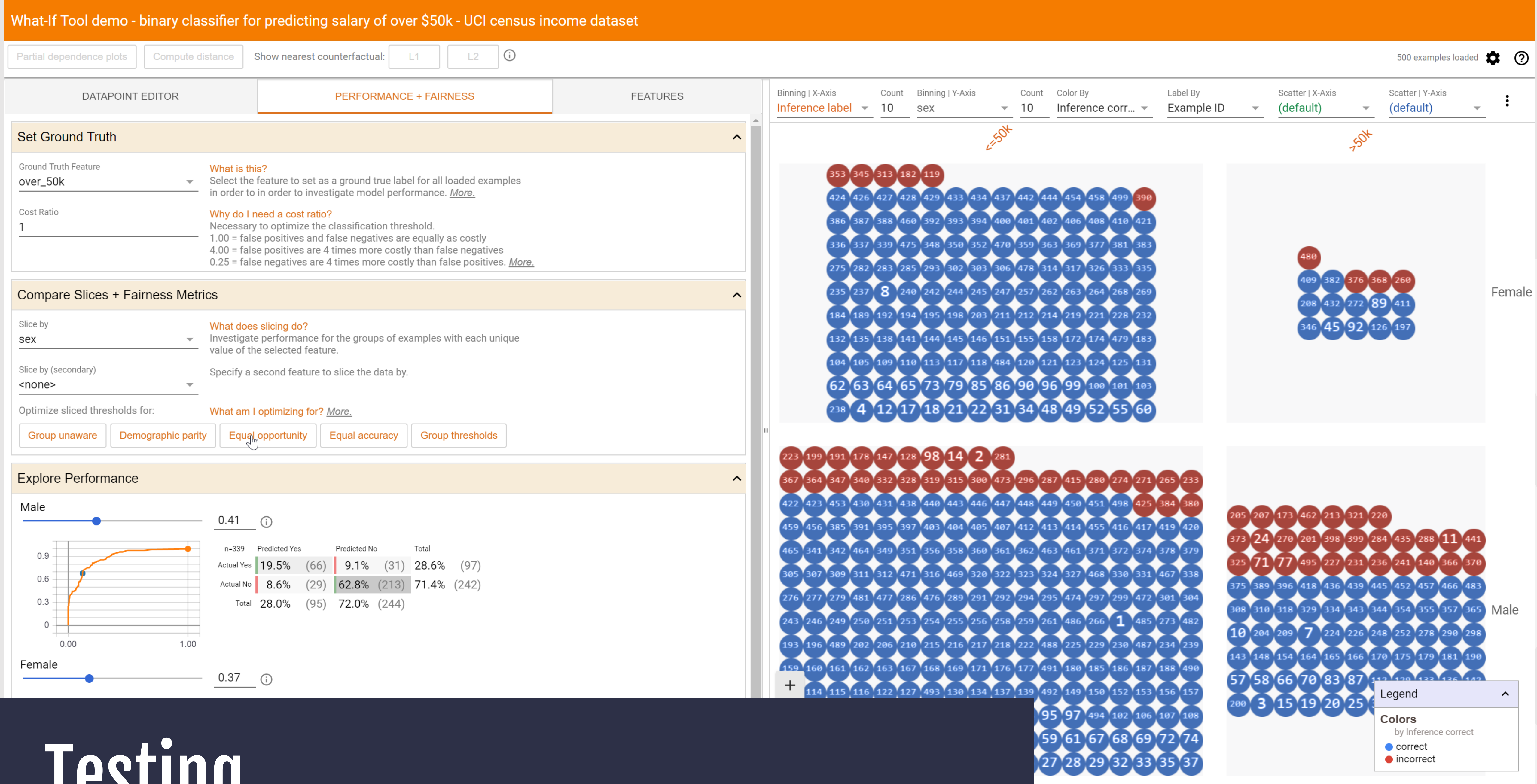
Description

Predictive models are increasingly been deployed for the purpose of determining access to services such as credit, insurance, and employment. Despite societal gains in efficiency and productivity through deployment of these models, potential systemic flaws have not been fully addressed, particularly the potential for unintentional discrimination. This discrimination could be on the basis of race, gender, religion, sexual orientation, or other characteristics. This project addresses the question: how can an analyst determine the relative significance of the inputs to a black-box predictive model in order to assess the model's fairness (or discriminatory extent)?

We present FairML, an end-to-end toolbox for auditing predictive models by quantifying the relative significance of the model's inputs. FairML leverages model compression and four input ranking algorithms to quantify a model's relative predictive dependence on its inputs. The relative significance of the inputs to a predictive model can then be used to assess the fairness (or discriminatory extent) of such a model. With FairML, analysts can more easily audit cumbersome predictive models that are difficult to interpret.s of black-box algorithms and corresponding input data.

Testing

https://github.com/adebayoj/fairml



Testing

https://pair-code.github.io/what-if-tool/uci.html

- Protected characteristics
- Spot checks and counter factuals

Monitoring

- Adversarial behaviours
- Robustness
- Retro-engineering
- Lessons from security – Red and Blue Teams?

Security

@theStephLocke

bit.ly/ethicaldatasciencelinks

Thanks!