Biases & Ethics

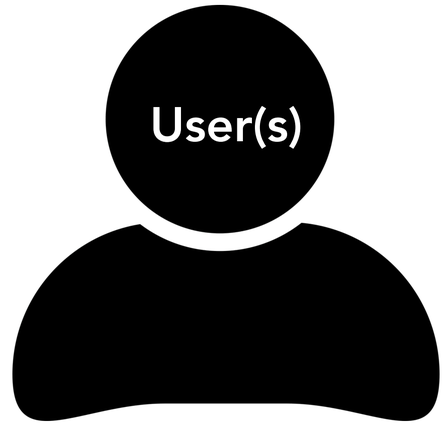
Baptiste Caramiaux

Machine learning has become “real”

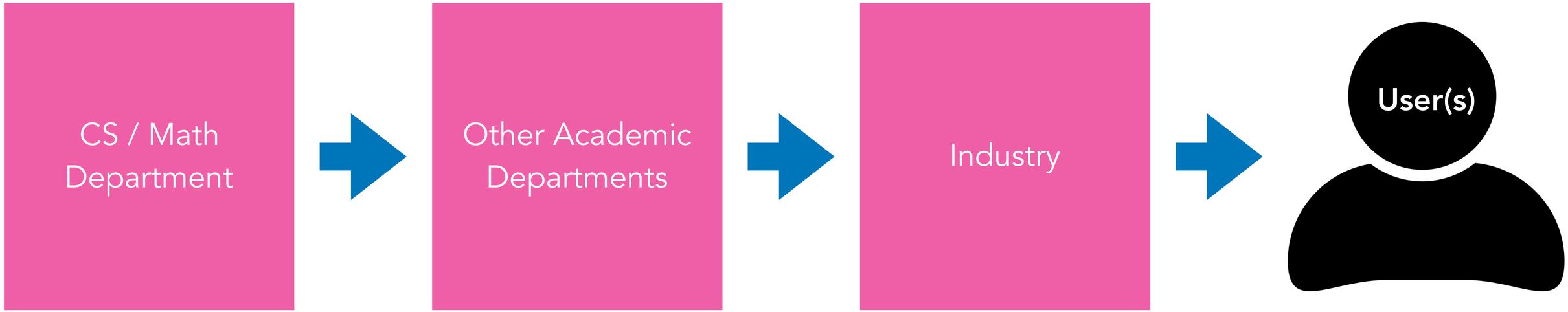
CS / Math Department

Other Academic Departments

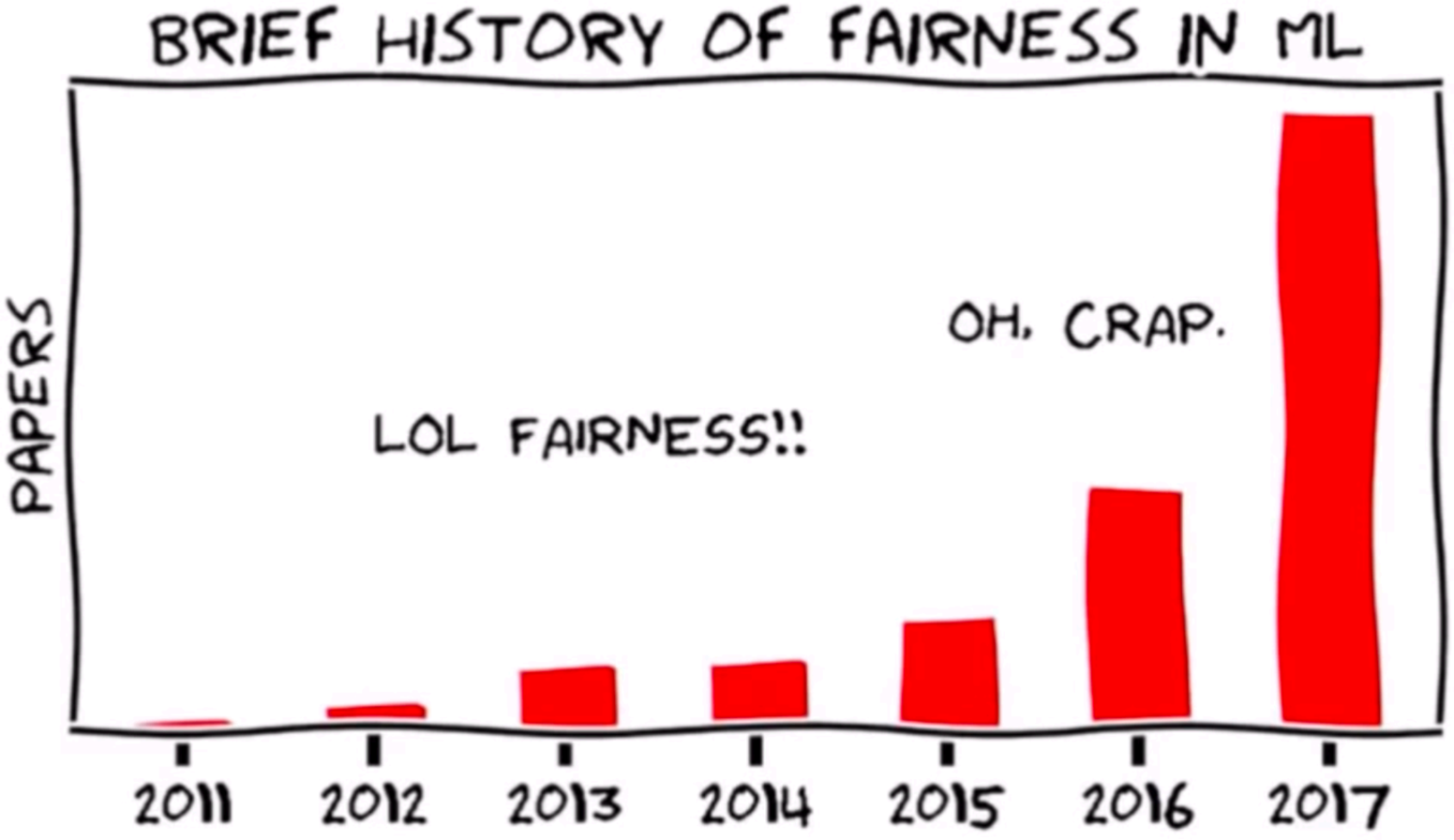
Industry



Machine learning has become “real”



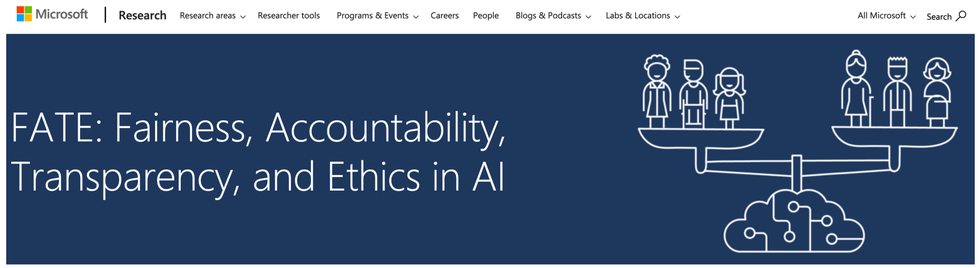
Socio-technical issues linked to ML ?

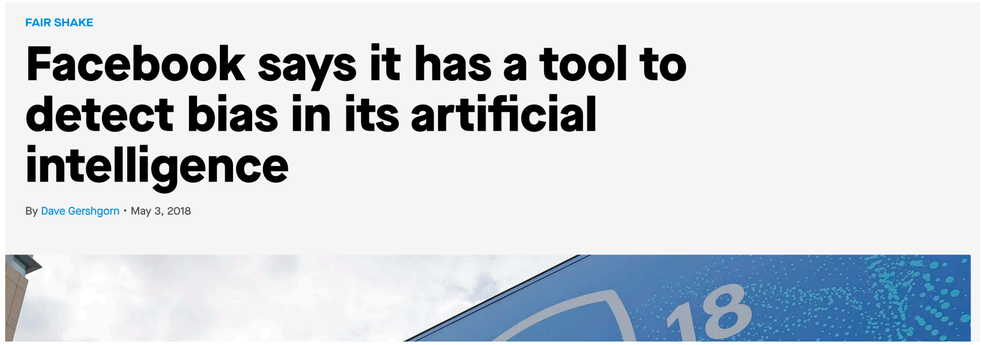


source: Mortiz Hardt

Growing interests in ML/AI/HCI

- Explainability
- Interpretability
- Fairness
- Accountability
- Transparency
- …





Before we begin

Discussion about https://www.excavating.ai/

Excavating AI

The Politics of Images in Machine Learning Training Sets

By Kate Crawford and Trevor Paglen

Terminology

- In the field of ethics, biases, fairness etc, people talk about **AI** systems more than **ML** systems

- **Bias**: hard to define, maybe “a skew that can harm someone”
  Note: A bias can be intentional or unintentional

So what happened?

Examples of “issues” with ML systems

2018

AARIAN MARSHALL TRANSPORTATION 05.29.2018 07:00 AM

False Positives: Self-Driving Cars and the Agony of Knowing What Matters

New details from Uber's self-driving crash highlight the difficulty—and importance—of ignoring what doesn't matter, while recognizing what does.

In self-driving cars, the problem with avoiding both false positives and negatives is that the more you do to get away from one, the closer you get to the other. SPENCER PLATT/GETTY IMAGES

According to a preliminary report released by the National Transportation Safety Board last week, Uber's system detected pedestrian Elaine Herzberg six seconds before striking and killing her. It identified her as an unknown object, then a vehicle, then finally a bicycle. (She was pushing a bike, so close enough.) About a second before the crash, the system determined it needed to slam on the brakes. But Uber hadn't set up its system to act on that decision, the NTSB explained in the report. The engineers prevented their car from making that call on its own "to reduce the potential for erratic vehicle behavior." (The company relied on the car's human operator to avoid crashes, which is a whole separate problem.)

> Uber's engineers decided not to let the car auto-brake because they were worried the system would overreact to things that were unimportant or not there at all. They were, in other words, very worried about false positives.

https://www.wired.com/story/self-driving-cars-uber-crash-false-positive-negative/

› Technology Intelligence

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours

share

Save

Microsoft's new teenage chat-bot CREDIT: TWITTER

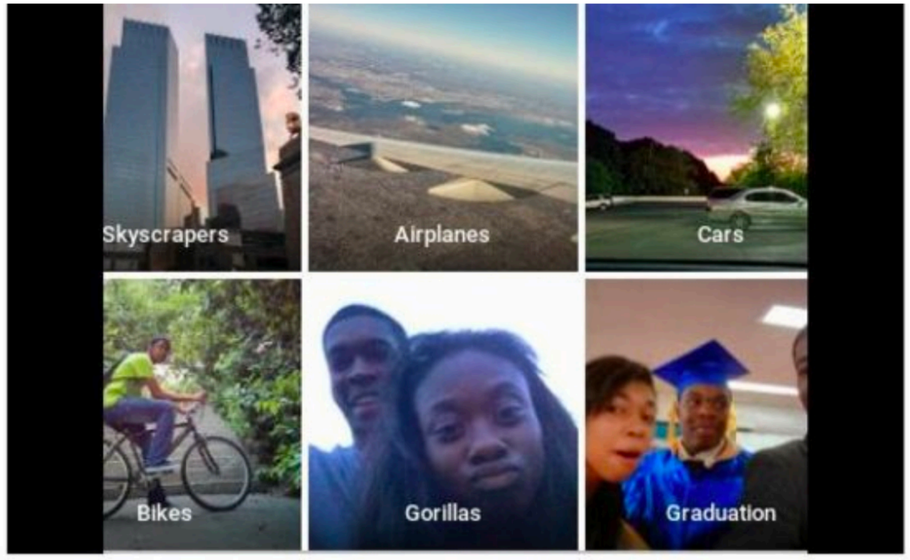
https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

2015

Technology

Google apologises for Photos app's racist blunder

1 July 2015 Share



Mr Alcine tweeted Google about the fact its app had misclassified his photo

https://www.bbc.com/news/technology-33347866

Facial Recognition Is Accurate, if You're a White Guy

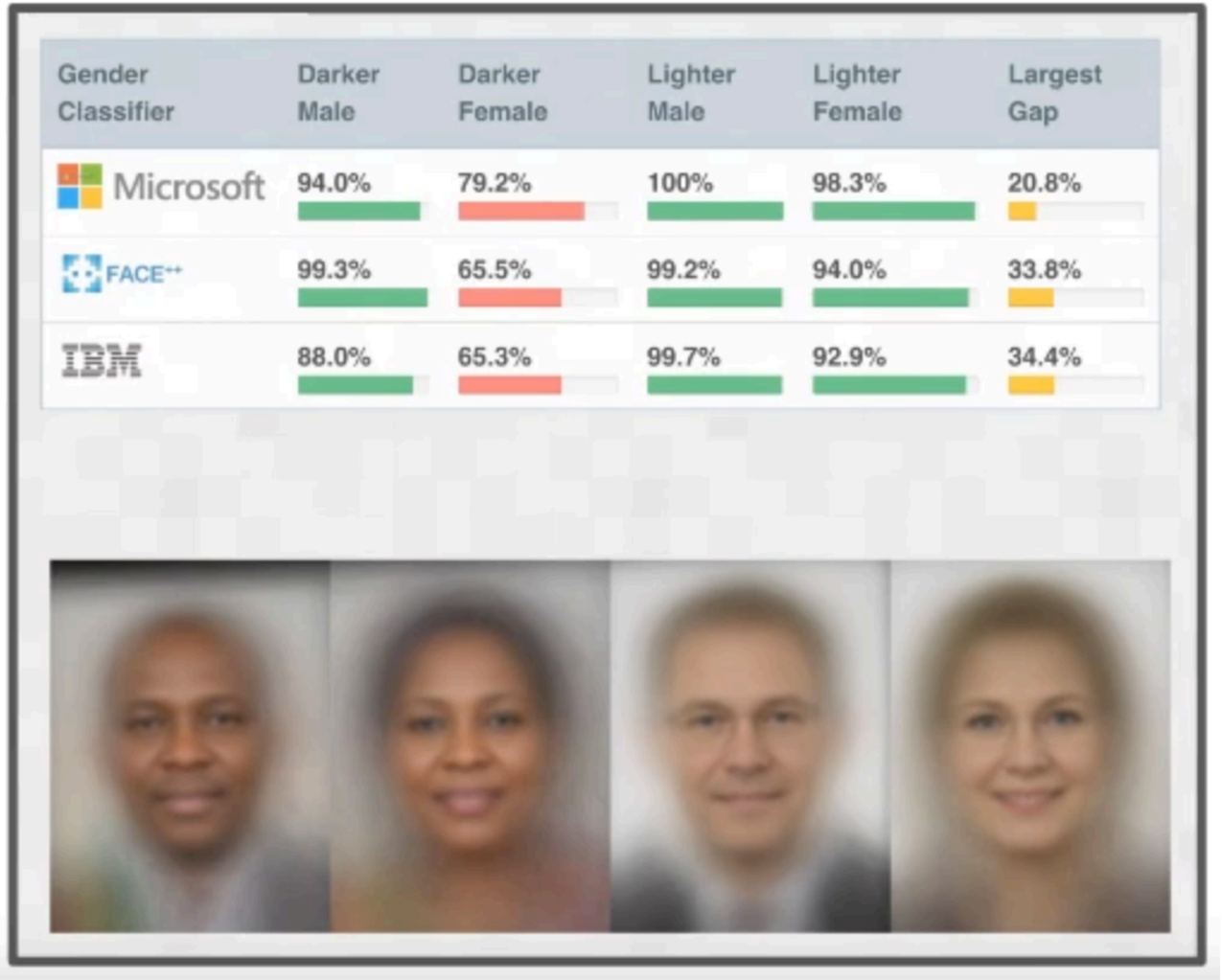
By Steve Lohr

Feb. 9, 2018

Facial recognition technology is improving by leaps and bounds. Some commercial software can now tell the gender of a person in a photograph.

When the person in the photo is a white man, the software is right 99 percent of the time.

But the darker the skin, the more errors arise — up to nearly 35 percent for images of darker skinned women, according to a new study that breaks fresh ground by measuring how the technology works on people of different races and gender.



https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

TECHNOLOGY NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

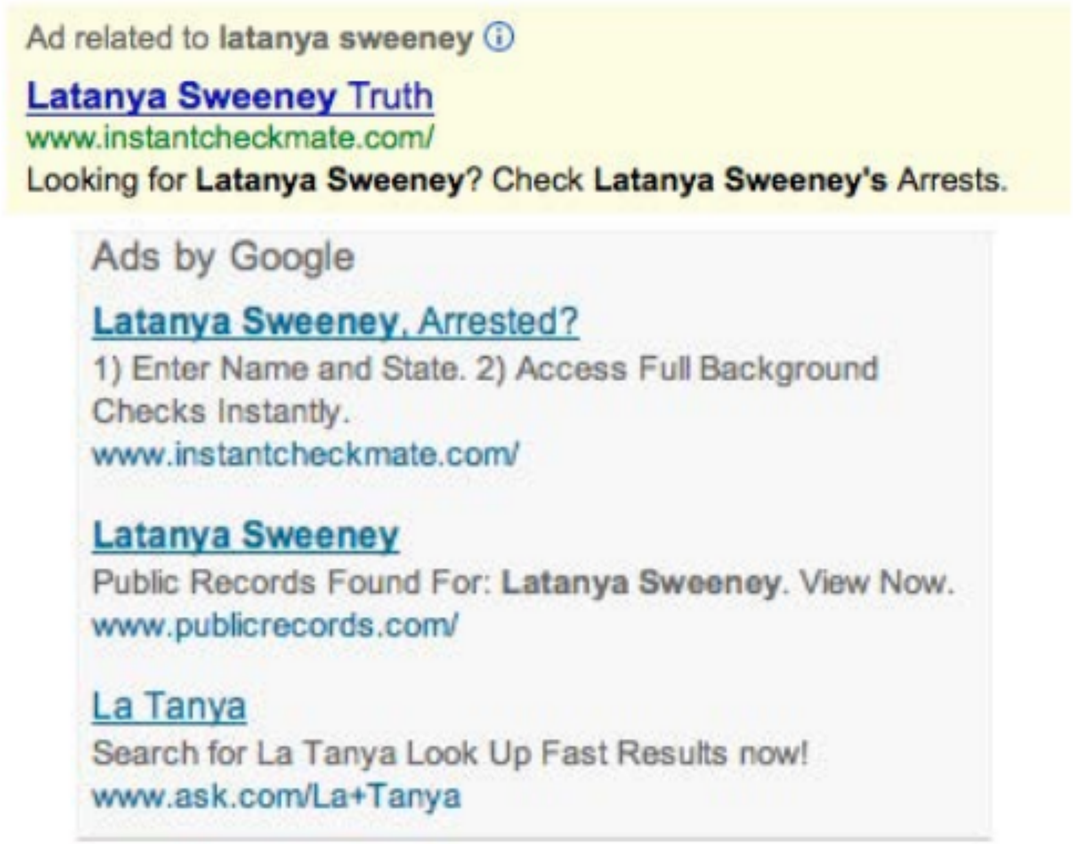
That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Percentage of women in top 100 Google image search results for CEO: 11%
Percentage of U.S. CEOs who are women: 27%

https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/

Kay et al. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations”. In *CHI’15*



Sweeney et al. (2013) Discrimination in Online Ad Delivery
https://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf

Bigger picture

Consequences of biases

- Harm of allocation
  withhold opportunities or ressources

- Harm of representation
  over/under representation, stereotypes, denigration

The Trouble with Bias - Kate Crawford - N(eur)IPS 2017 Keynote
https://www.youtube.com/watch?v=fMym_BKWQzk&t=2047s

Ethics and innovation

In development, innovation most often means:

- Working fast and flexibly like a Silicon Valley startup (process)
- Incorporating the latest technologies to transform the way we engage populations and measure everything (technology)
- Taking user needs into account, make sure that solutions reflect real user needs (design)
- New models of shared value partnership, thinking more like an ‘incubator’ than a serial-process driven system (investing)

From https://ssir.org/articles/entry/the_ethics_of_innovation

Some principles

Operational principles for design in innovation:

- Design with the user.
- Understand the existing ecosystem.
- Design for scale.
- Build for sustainability.
- Be data driven.
- Use open standards, open data, open source, and open innovation.
- Reuse and improve.
- Do no harm.
- Be collaborative.

From https://ssir.org/articles/entry/the_ethics_of_innovation

Decision making

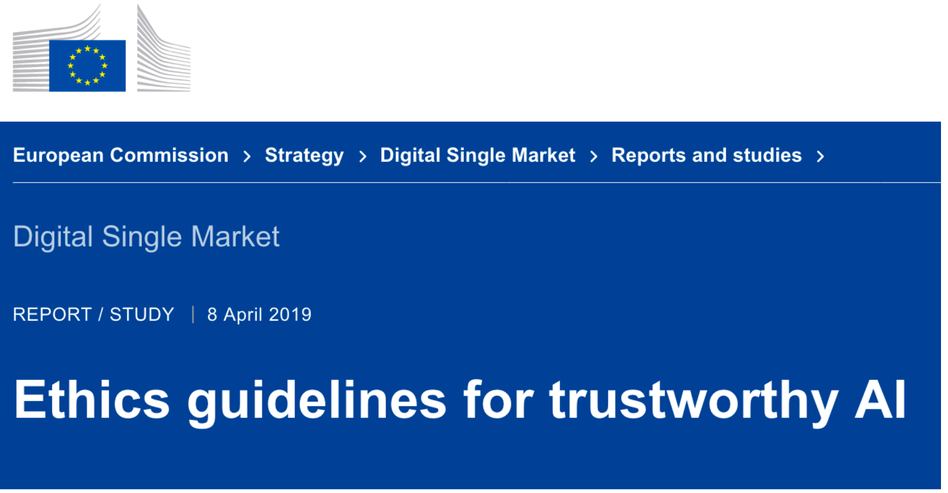
- Existing ethical frameworks require a **person to make a decision**
- Why?
  - Decisions without a person's judgement have no responsibility.
  - Nearly any ethical frameworks disallows "I followed the rules" as a defense.

From http://www.humanaiclass.org/schedule/

Policy making

European Commission

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai



European Ethics guidelines for trustworthy AI

Realisation of trustworthy AI

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

What interaction design with ML can bring?

IML for FAT AI

- Methodologies
  - Qualitative research
  - Field study
  - User experience
- Interactive tools
  - Make dataset, model and its parameters interactive !
- Feedback and guidance

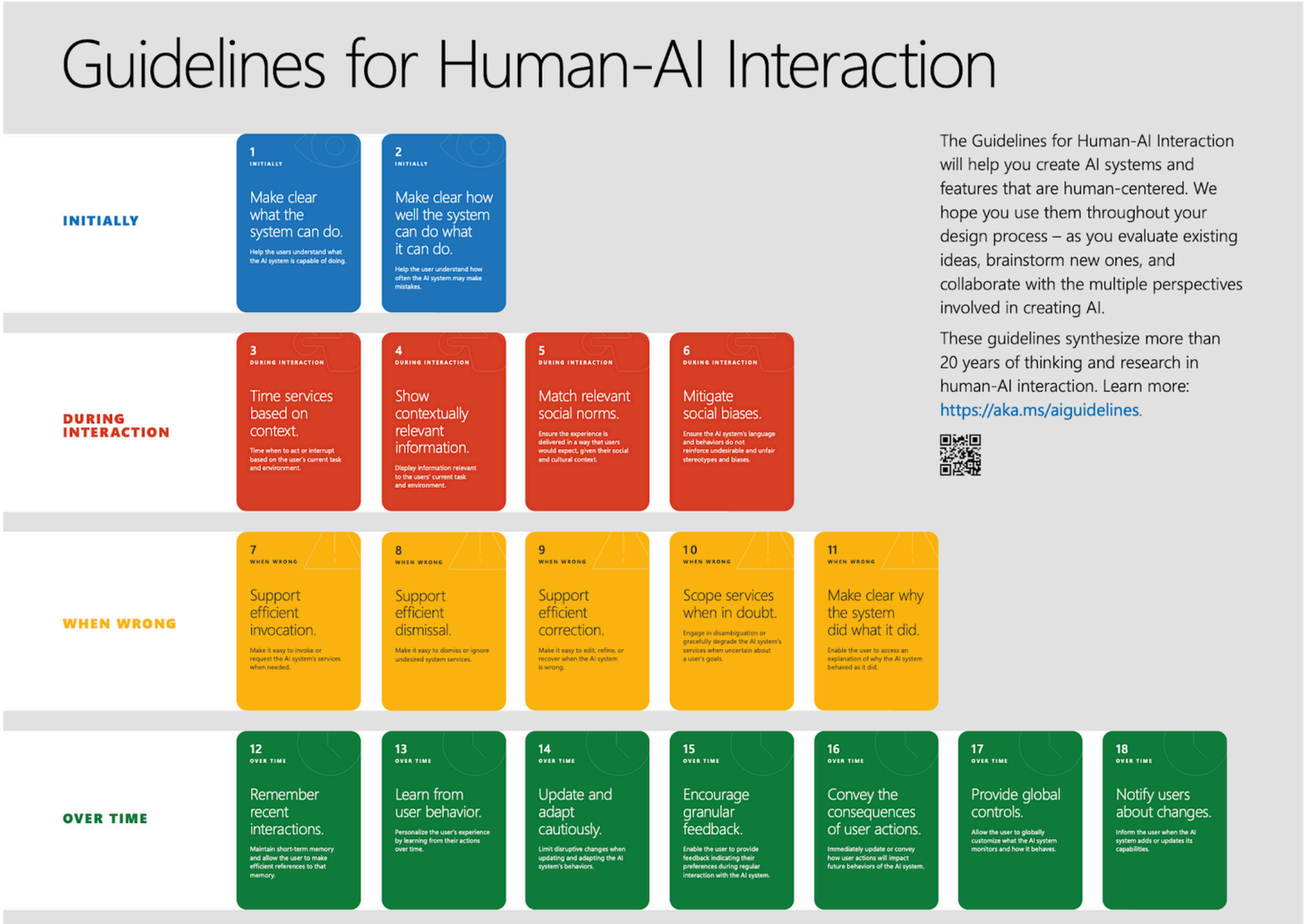
Methodology: an example

- Example in medical application
  Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-24. https://dl.acm.org/doi/pdf/10.1145/3359206

- Method: "*interview of 21 pathologists before, during, and after being presented deep neural network (DNN) predictions for prostate cancer diagnosis*"

- Results: "*far beyond understanding the local, case-specific reasoning behind any model decision, clinicians desired upfront information about basic, global properties of the model, such as its known strengths and limitations, its subjective point-of-view, and its overall design objective—what it's designed to be optimized for.*"

Guidelines for Human-AI Interaction



The Guidelines for Human-AI Interaction will help you create AI systems and features that are human-centered. We hope you use them throughout your design process – as you evaluate existing ideas, brainstorm new ones, and collaborate with the multiple perspectives involved in creating AI.

These guidelines synthesize more than 20 years of thinking and research in human-AI interaction. Learn more: https://aka.ms/aiguidelines.

A case study: COMPAS

Correctional Offender Management Profiling for Alternative Sanctions

FREE

COMPAS Recidivism Risk Score Data and Analysis

Source	Broward County Clerk's Office, Broward County Sherrif's Office, Florida Department of Corrections, ProPublica
Date Released	January 2020
Related Content	Machine Bias
Methodology	How We Analyzed the COMPAS Recidivism Algorithm

DOWNLOAD ON GITHUB.COM

Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood to re-offend. There are dozens of these risk assessment algorithms in use, including two leading nationwide tools offered by commercial vendors. Our story, "Machine Bias," set out to assess one of the commercial tools, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), made by Northpointe, Inc. to discover the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

The linked data includes: a database containing the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County from 2013 and 2014; code in R and Python; a Jupyter notebook; and other files needed for the analysis.

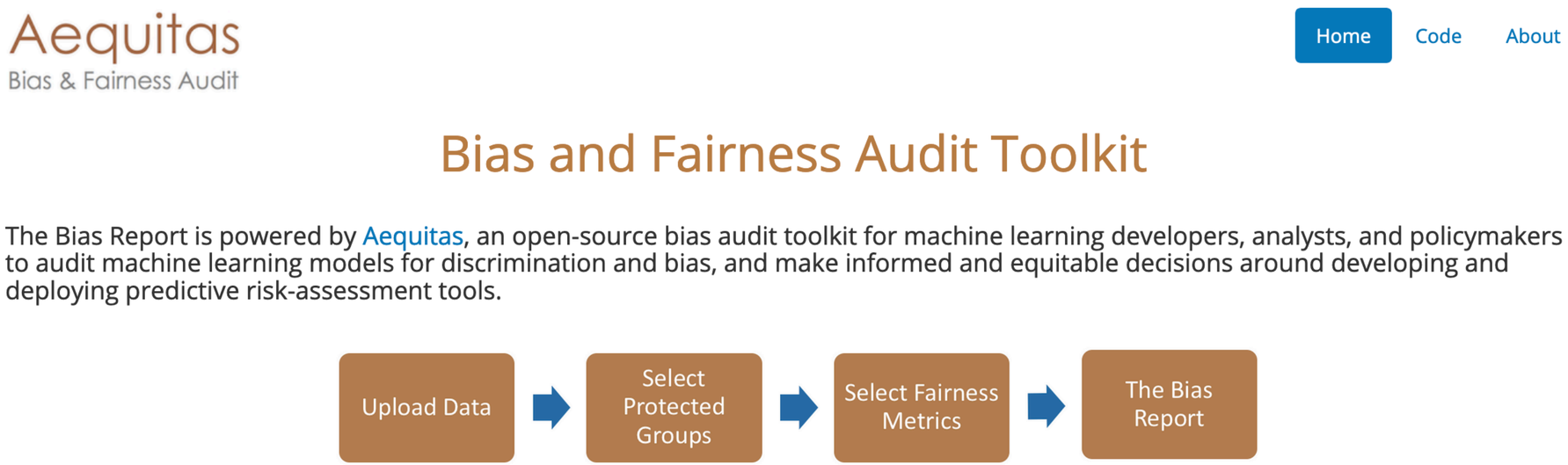
https://www.kaggle.com/danofer/compass
https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

	id	name	first	last	sex	dob	age	age_cat	race	juv_fel_count	decile_score	juv_misd_count	juv_other_count	priors_count	days_b_screening_arrest	c_jail_in
0	1.0	miguel hernandez	miguel	hernandez	Male	18/04/1947	69	Greater than 45	Other	0	1	0	0	0	-1.0	13/08/2013 6:03
1	2.0	miguel hernandez	miguel	hernandez	Male	18/04/1947	69	Greater than 45	Other	0	1	0	0	0	-1.0	13/08/2013 6:03
2	3.0	michael ryan	michael	ryan	Male	06/02/1985	31	25 - 45	Caucasian	0	5	0	0	0	NaN	NaN
3	4.0	kevon dixon	kevon	dixon	Male	22/01/1982	34	25 - 45	African-American	0	3	0	0	0	-1.0	26/01/2013 3:45
4	5.0	ed philo	ed	philo	Male	14/05/1991	24	Less than 25	African-American	0	4	0	1	4	-1.0	13/04/2013 4:58
5	6.0	ed philo	ed	philo	Male	14/05/1991	24	Less than 25	African-American	0	4	0	1	4	-1.0	13/04/2013 4:58
6	7.0	ed philo	ed	philo	Male	14/05/1991	24	Less than 25	African-American	0	4	0	1	4	-1.0	13/04/2013 4:58
7	8.0	ed philo	ed	philo	Male	14/05/1991	24	Less than 25	African-American	0	4	0	1	4	-1.0	13/04/2013 4:58
8	9.0	ed philo	ed	philo	Male	14/05/1991	24	Less than 25	African-American	0	4	0	1	4	-1.0	13/04/2013 4:58
9	10.0	marcu brown	marcu	brown	Male	21/01/1993	23	Less than 25	African-American	0	8	1	0	1	NaN	NaN
10	11.0	bouthy pierrelouis	bouthy	pierrelouis	Male	22/01/1973	43	25 - 45	Other	0	1	0	0	2	NaN	NaN
11	12.0	marsha miles	marsha	miles	Male	22/08/1971	44	25 - 45	Other	0	1	0	0	0	0.0	30/11/2013 4:50
12	13.0	edward riddle	edward	riddle	Male	23/07/1974	41	25 - 45	Caucasian	0	6	0	0	14	-1.0	18/02/2014 5:08
13	14.0	edward riddle	edward	riddle	Male	23/07/1974	41	25 - 45	Caucasian	0	6	0	0	14	-1.0	18/02/2014 5:08
14	15.0	steven stewart	steven	stewart	Male	25/02/1973	43	25 - 45	Other	0	4	0	0	3	-1.0	29/08/2013 8:55

...

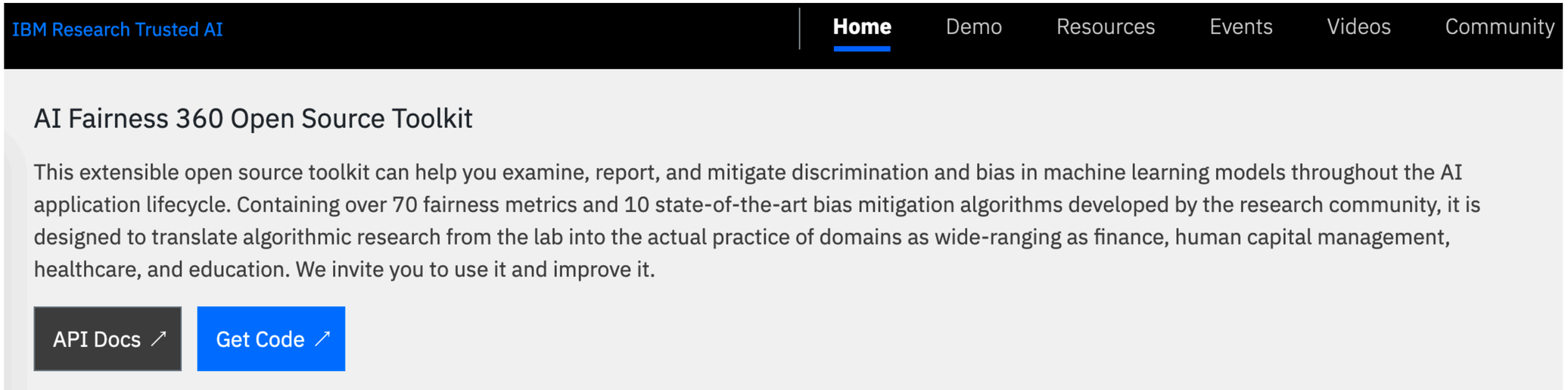
https://www.kaggle.com/danofer/compass
https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

Example #1



http://aequitas.dssg.io/

Example #2



https://aif360.mybluemix.net

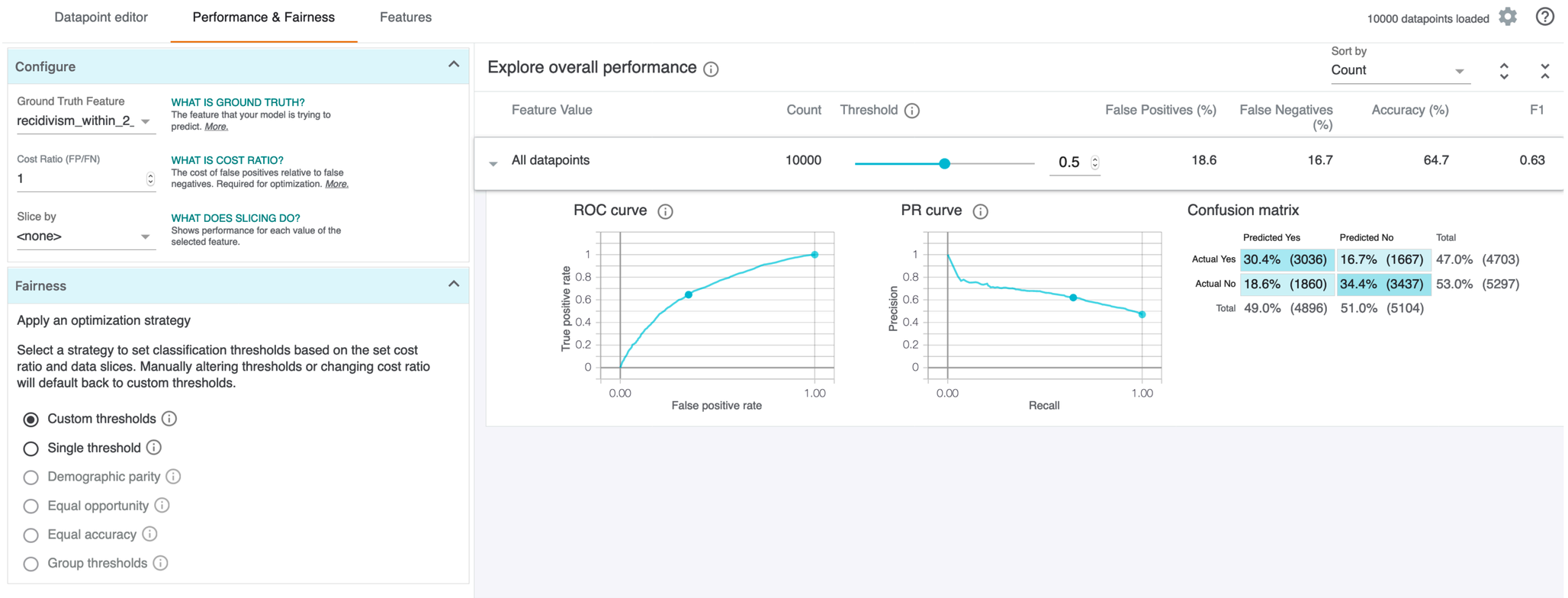
Example #3

What If...

you could inspect a machine learning model,
with minimal coding required?

https://pair-code.github.io/what-if-tool/index.html#demos

Exercise



How the users think this tool work?
How does the user know when it finds a better (fairer) model?
How can we improve the interaction?

Ressources

Project specifications

Project preparation

https://www.notion.so/marcellejs/Public-572f3f36344d400e943061a4ee0ed022