Implicações da Inteligência Artificial e como mitigar vieses com ferramentas open source

Gabriela de Queiroz

@gdequeiroz | linktr.ee/gdq

slides: bit.ly/codabr20

CODA·BR

Gabriela de Queiroz

Sr. Engineering and Data Science Manager, IBM

- Fundadora do R-Ladies (rladies.org)
- Fundadora do AI Inclusive (ai-inclusive.org)

- Graduação em Estatística (UERJ)
- Mestrado em Epidemiologia (ENSP/Fiocruz)
- Mestrado em Estatística (CSUEB)

Data Scientist + Developer Advocate + Open Source Developer + Manager +

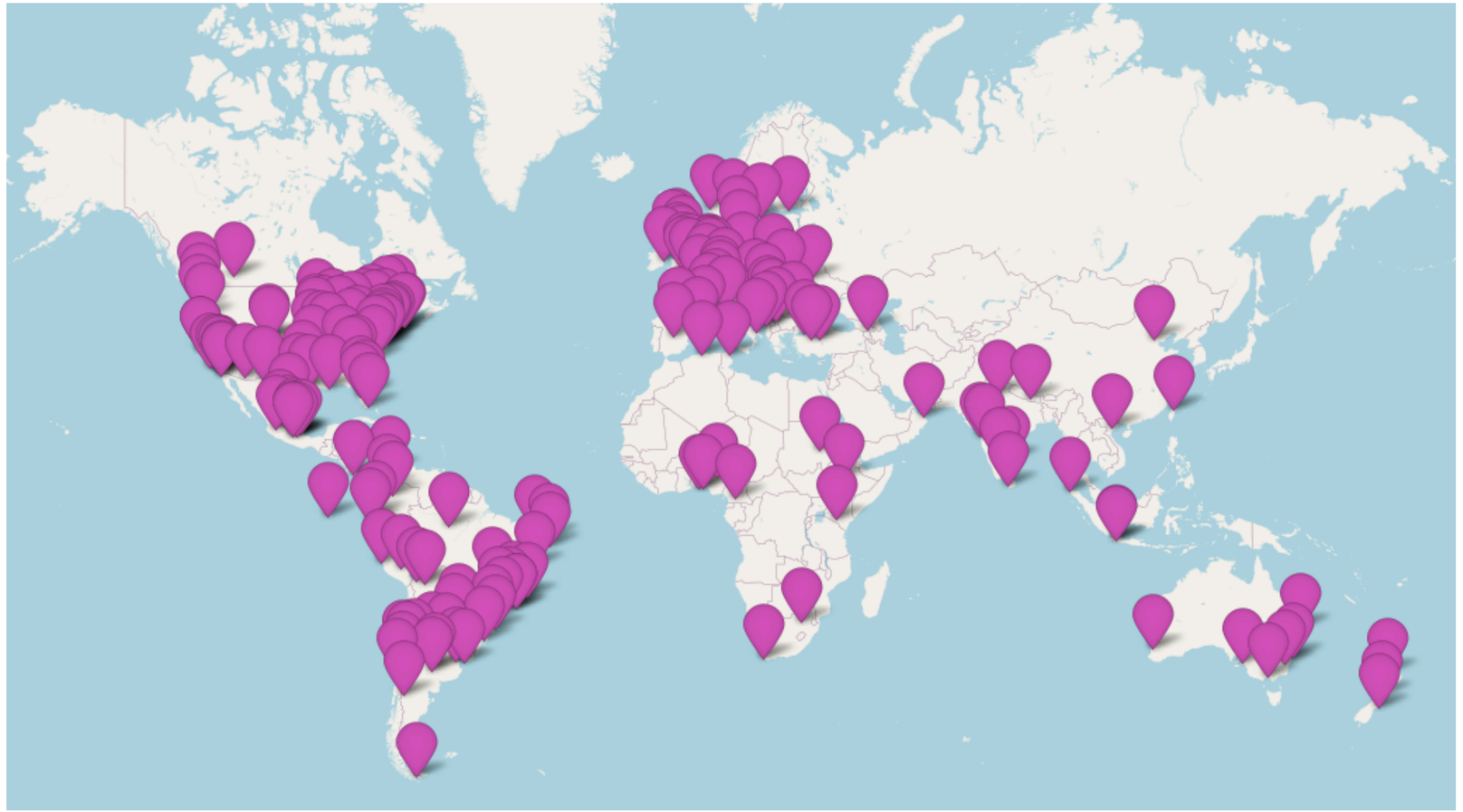
Statistician + Epidemiologist + Community Builder + Mentor + Speaker + Educator

R-Ladies

54
R-Ladies Countries

193
R-Ladies Cities

71618
R-Ladies members on meetup.com



	city	state	country	dt_created	members
41	São Paulo		BR	2018-02-10	1004
38	Belo Horizonte		BR	2018-04-20	800
34	Niterói		BR	2018-06-04	580
51	Rio de Janeiro		BR	2017-02-27	579
28	Florianópolis		BR	2019-04-07	441
43	Porto Alegre		BR	2017-10-30	346
7	Natal		BR	2020-06-07	241
32	Salvador		BR	2018-07-23	192
22	Recife		BR	2019-09-02	147
25	Goiania		BR	2019-05-06	146
18	Vitória		BR	2019-09-29	132
6	Fortaleza		BR	2020-06-09	69
26	Lavras		BR	2019-04-16	41
5	Curitiba		BR	2020-06-12	36
14	Ribeirão Preto		BR	2020-03-06	32
21	Manaus		BR	2019-09-11	29

52
R-Ladies groups in Latin America

AI Inclusive

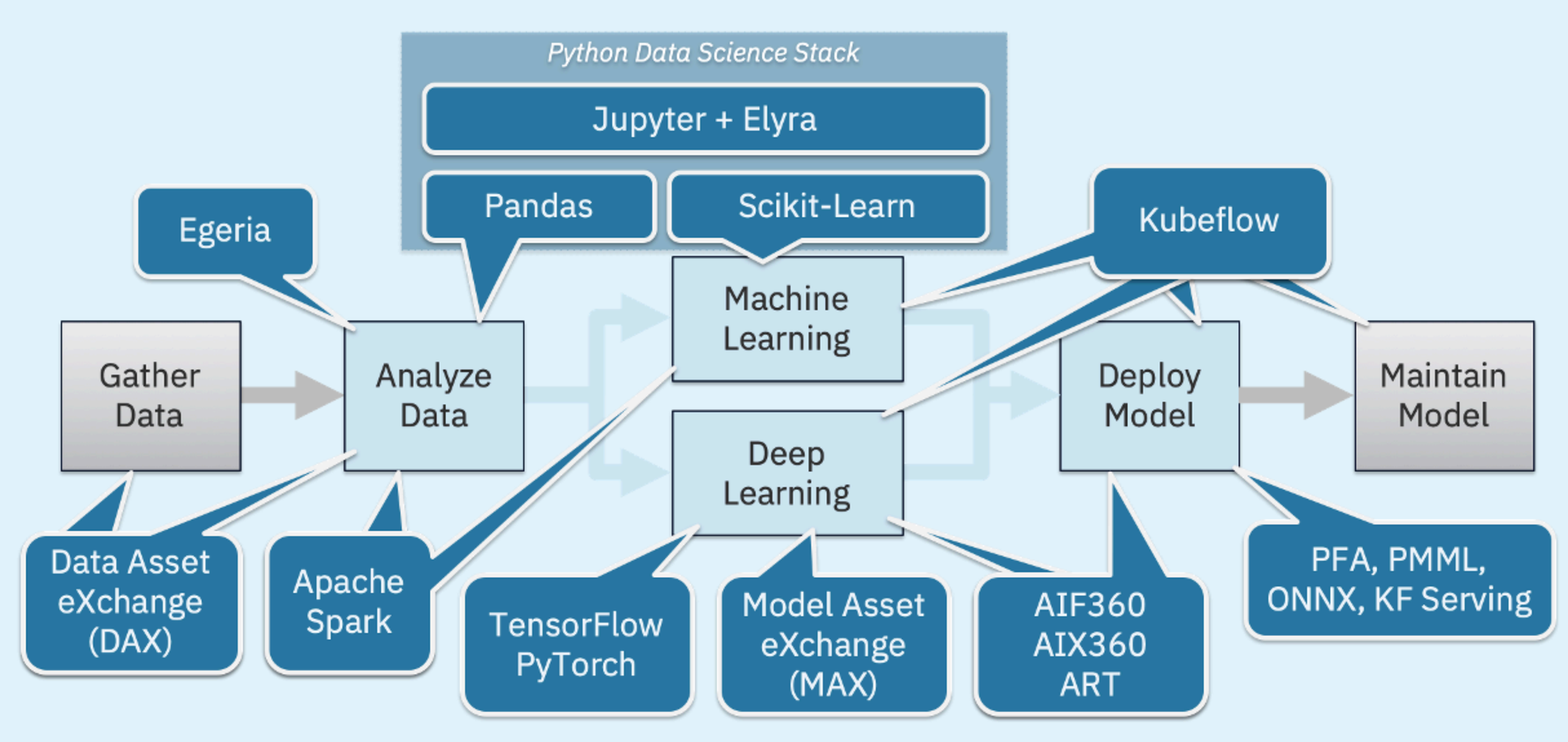
Missão : Aumentar a **representatividade** e **participação** de minorias em Inteligência Artificial

- Website: ai-inclusive.org
- Twitter: bit.ly/ai-inclusive-twitter
- Facebook: bit.ly/ai-inclusive-facebook
- Instagram: bit.ly/ai-inclusive-instagram
- Youtube: bit.ly/ai-inclusive-youtube

Se tiver interesse em criar um capítulo, mande-nos um email: info@ai-inclusive.org

We build tools to make AI accessible and available to everybody

(codait.org)

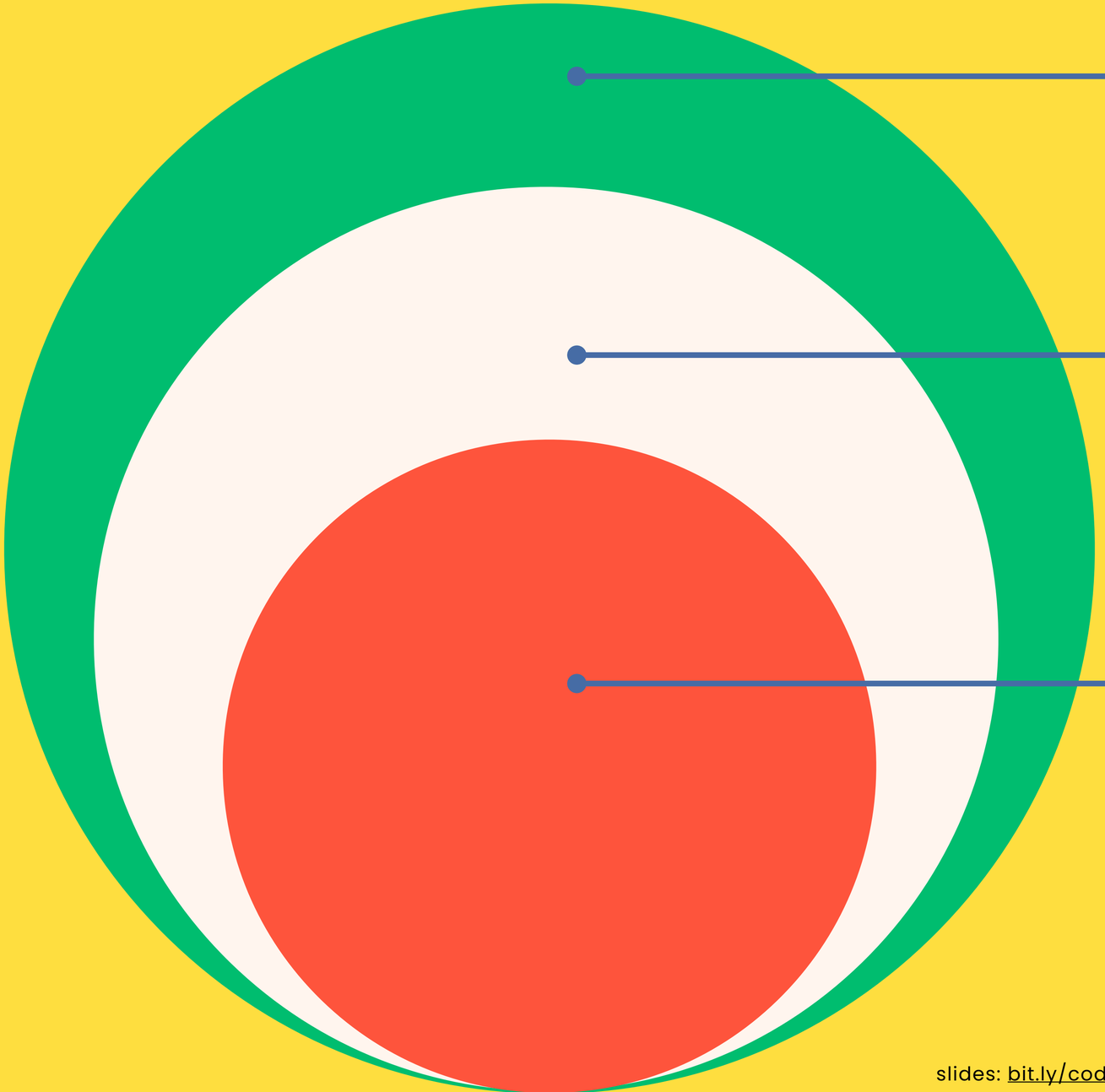


We contribute to and advocate for the open-source technologies

codait.org

O que é Inteligência Artificial?

slides: bit.ly/codabr20



Inteligência Artificial

qualquer coisa que permita que os computadores se comportem como humanos

Aprendizado de Máquina

subconjunto da Inteligência Artificial que lida com a extração de padrões de conjunto de dados

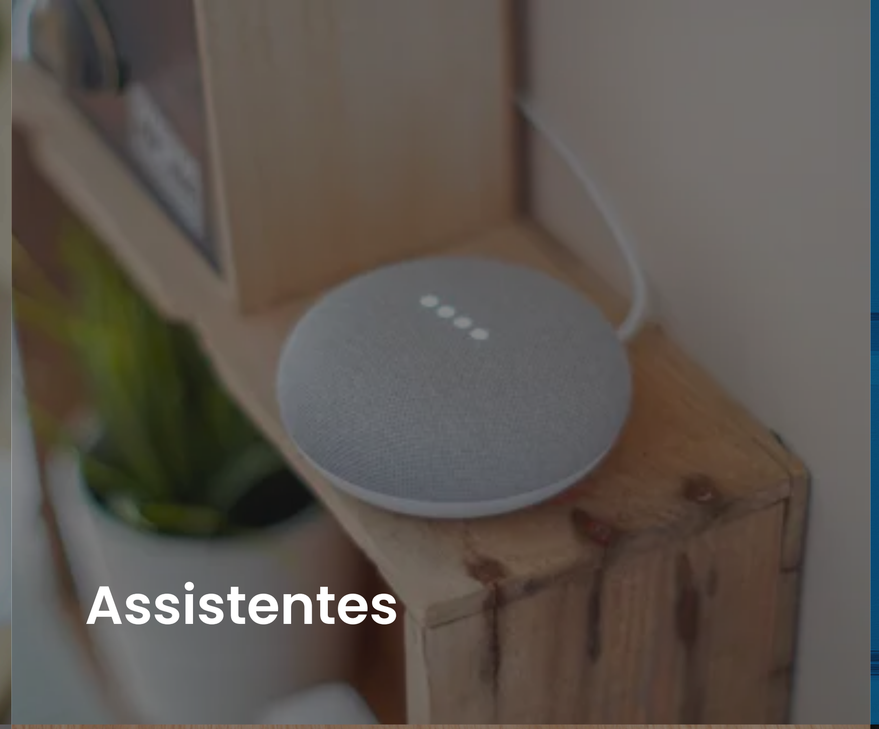
Aprendizagem Profunda

subconjunto do aprendizado de máquina que usa redes neurais complexas

slides: bit.ly/codabr20

Onde podemos encontrar inteligência artificial?

Telefone

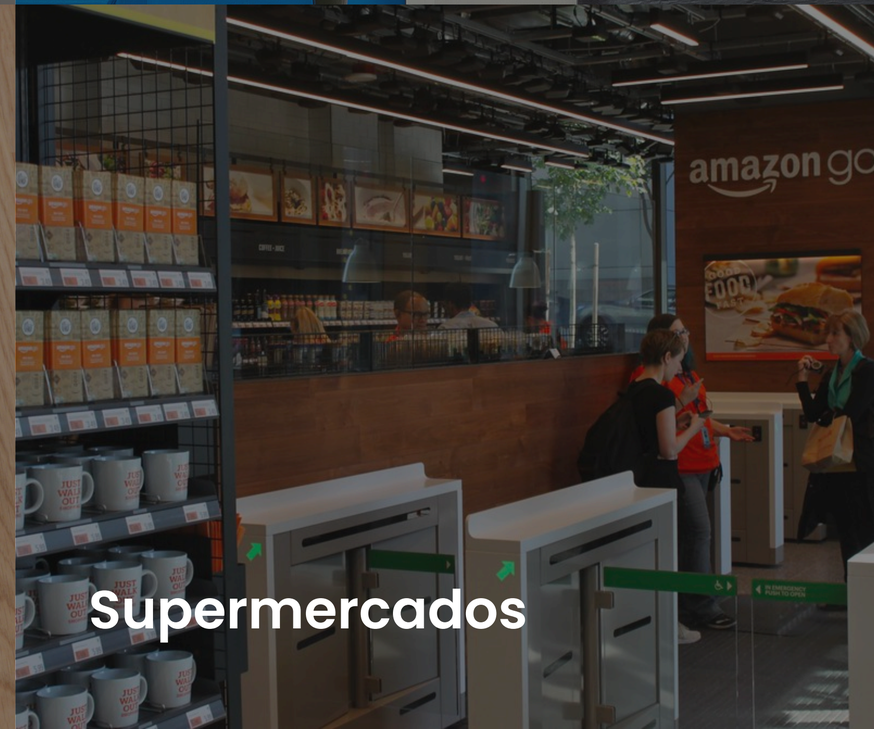


Assistentes

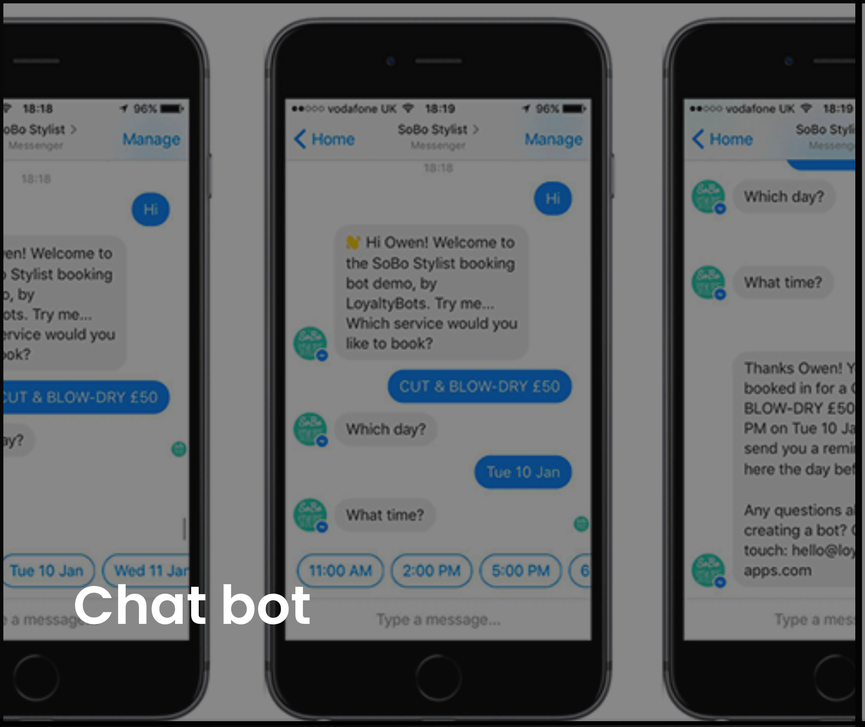
Campainha

Smartwatch

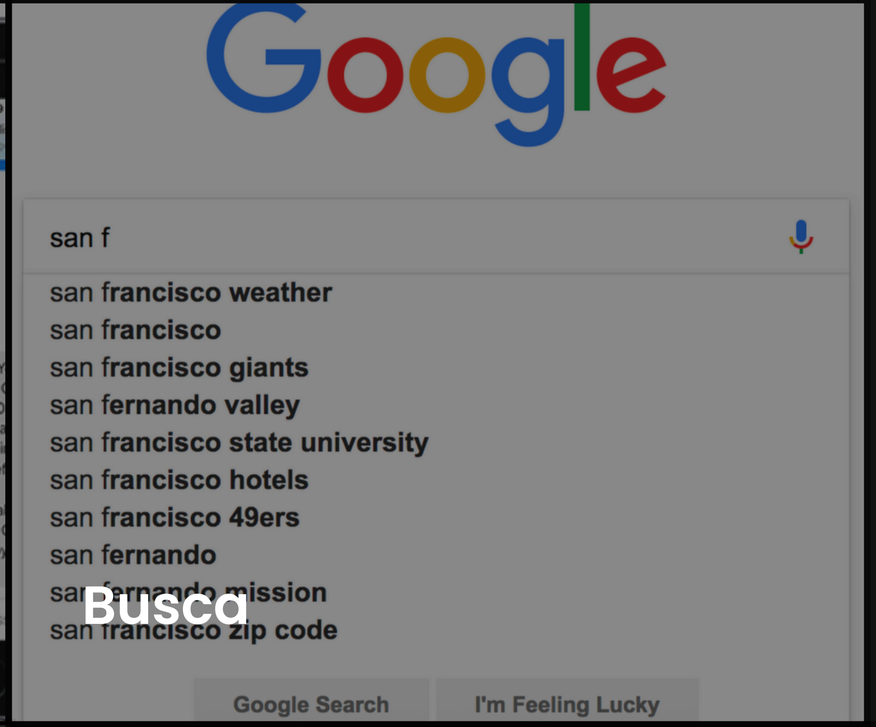
Aspirador inteligente



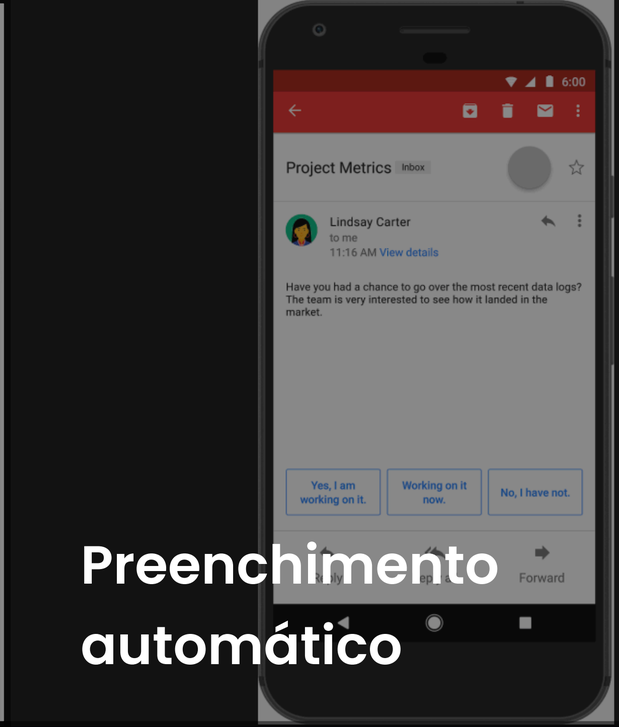
Supermercados



Chat bot



Busca



Preenchimento automático

Recomendações

Ridesharing

Compras online

Não existe tecnologia neutra

não existe ferramenta de tecnologia que seja apenas uma ferramenta.
Os aplicativos que usamos diariamente não são neutros.
Eles têm preconceitos embutidos, incentivos que nos fazem tomar certas ações.

HTTPS://FOURWEEKMBA.COM/WHY-TECHNOLOGY-IS-NOT-NEUTRAL-AND-WHY-YOU-SHOULD-CARE/

Exemplos de preconceito digital e de violação de privacidade

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

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

Março, 2020

Estados Unidos

audio clip of a 40-year-old black man speaking

• WHAT WAS SAID • WHAT MACHINE HEARD

"I~~me~~ mean, I knew~~know~~ I~~'m~~ was kinda tall for high school~~asking~~. I didn't wanna play center. I didn't because center don't~~send it on~~ have the ball that much. You get the ball occasionally when you in the post, I mean, but I didn't want to play it."

https://www.nytimes.com/2020/03/23/technology/speech-recognition-bias-apple-amazon-google.html

Wrongfully Accused by an Algorithm

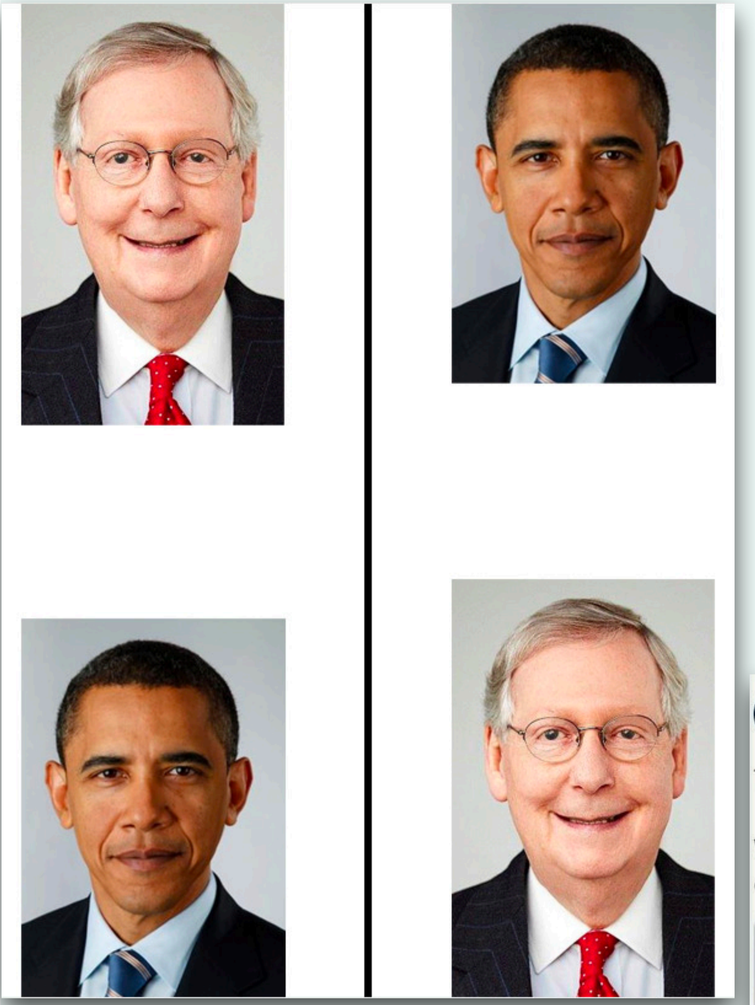
In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Junho, 2020

Estados Unidos

"This is not me," Robert Julian-Borchak Williams told investigators.
"You think all Black men look alike?"

https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html



Twitter investigates racial bias in image previews

Setembro, 2020

Estados Unidos

https://www.bbc.com/news/technology-54234822

Live facial recognition is tracking kids suspected of being criminals

In Buenos Aires, the first known system of its kind is hunting down minors who appear in a national database of alleged offenders.

Rights group criticizes Buenos Aires for using face recognition tech on kids

Outubro, 2020

Buenos Aires

HTTPS://WWW.TECHNOLOGYREVIEW.COM/2020/10/09/1009992/LIVE-FACIAL-RECOGNITION-IS-TRACKING-KIDS-SUSPECTED-OF-CRIME/

Algumas perguntas que devemos fazer

SOBRE IA

- Será que devemos fazer isso? E qual será a implicação?
- Que tipo de viés existe nesses dados?
- Quais são as taxas de erro para diferente subgrupos?
- Podemos tirar o software em produção se ele estiver se "comportando mal"?
- Temos um mecanismo de reparação caso as pessoas sejam prejudicadas pelo resultados?
- Quão diversa é a equipe que o construiu?

slides: bit.ly/codabr20

12% of leading machine learning researchers are female

- Research by WIRED and Element AI 2018

22% of jobs in AI are held by women,

with even fewer holding senior roles

- Global Gender Gap Report 2018

15% of AI research staff at Facebook are women

- AI Now Institute 2019

10% of AI research staff at Google are women

- AI Now Institute 2019

slides: bit.ly/codabr20

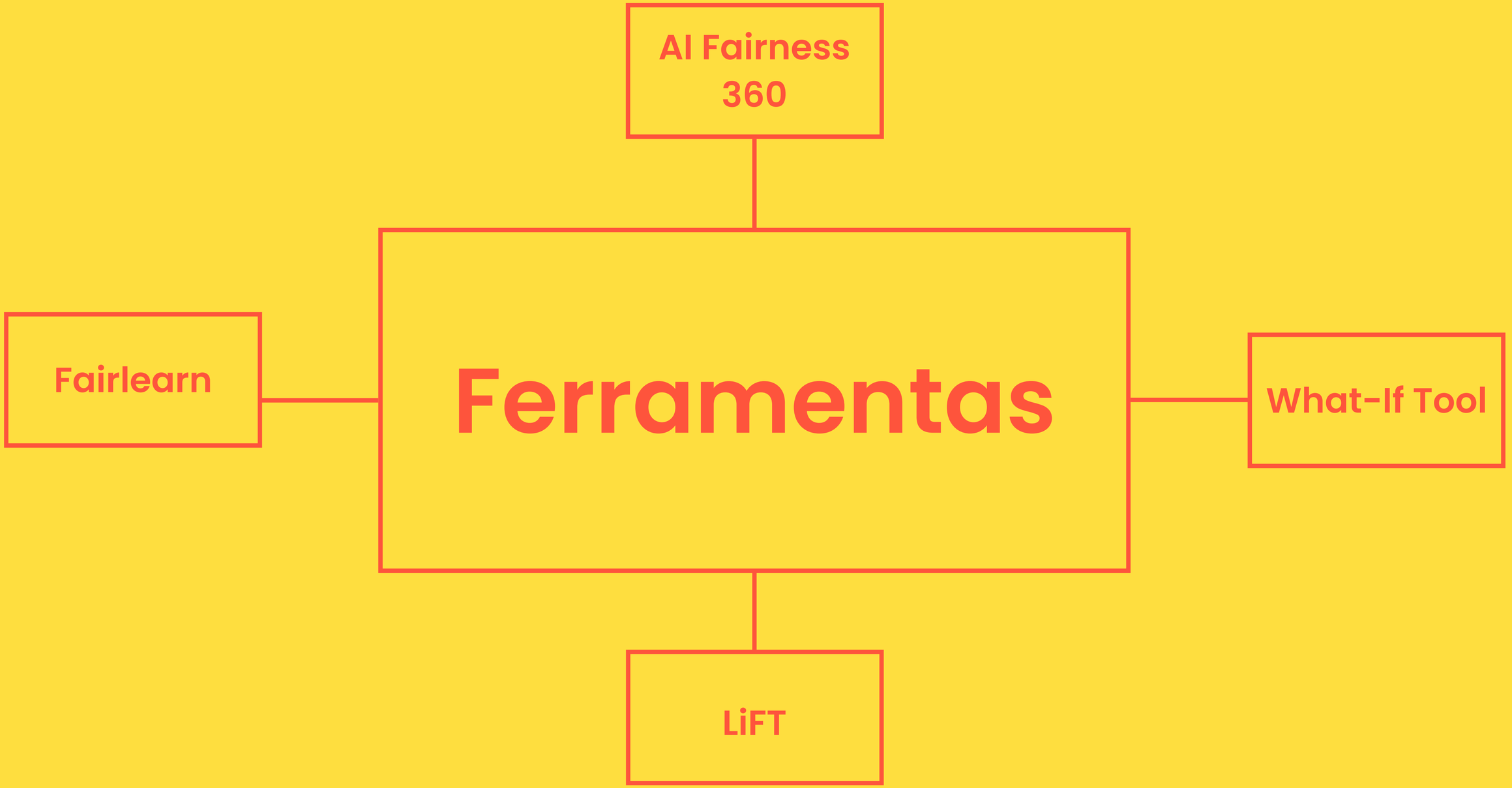
Ainda temos um caminho longo para melhorar a diversidade nessas áreas

O que podemos fazer enquanto isso?

FERRAMETAS PARA MITIGAR VIESES

Como podemos calcular e mitigar vieses em nossas análises?

slides: bit.ly/codabr20



E ALGUMAS OUTRAS ...

OPEN SOURCE

AI Fairness 360

aif360.mybluemix.net

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗ | Get Python Code ↗ | Get R Code ↗

Read More
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

Watch Videos
Watch videos to learn more about AI Fairness 360.

Read a paper
Read a paper describing how we designed AI Fairness 360.

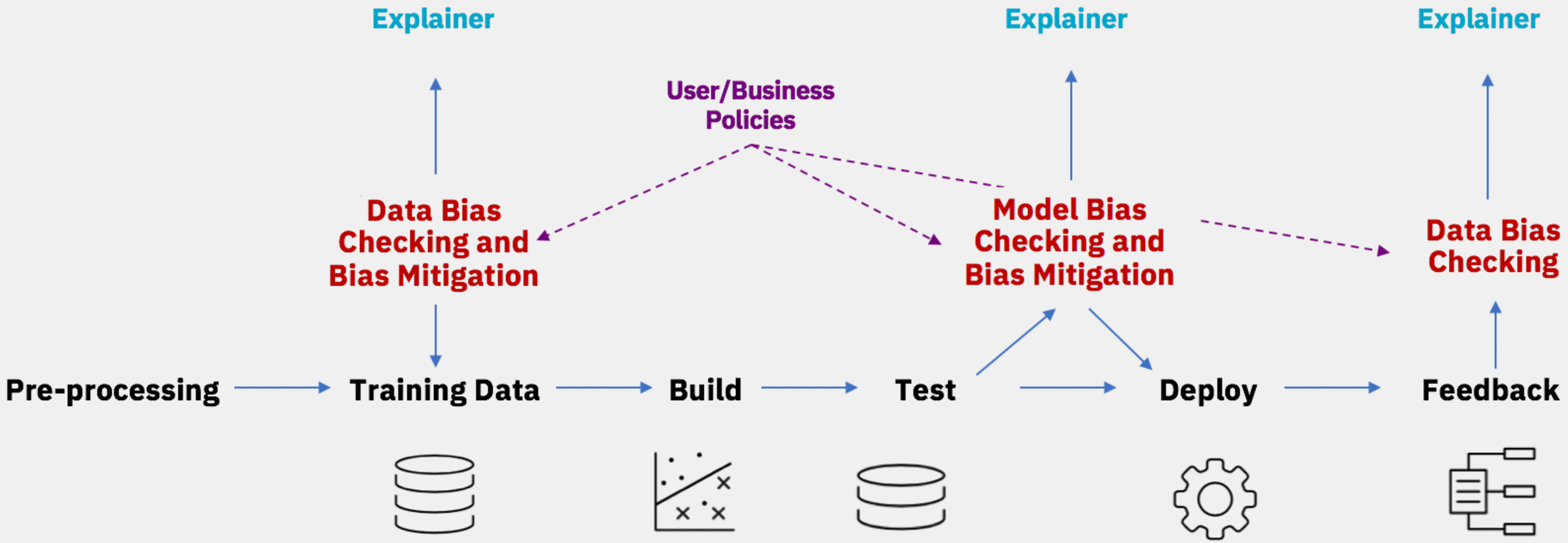
Use Tutorials
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

Ask a Question
Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.

View Notebooks
Open a directory of Jupyter Notebooks in GitHub that provide working examples of bias detection and mitigation in sample datasets. Then share your own notebooks!

Contribute
You can add new metrics and algorithms in GitHub. Share Jupyter notebooks showcasing how you have examined and mitigated bias in your machine learning application.

aif360.mybluemix.net



Mitigating bias throughout the AI lifecycle

Exemplo

Credit Scoring

German credit scoring

Predict an individual's credit risk.

Protected Attributes:

- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

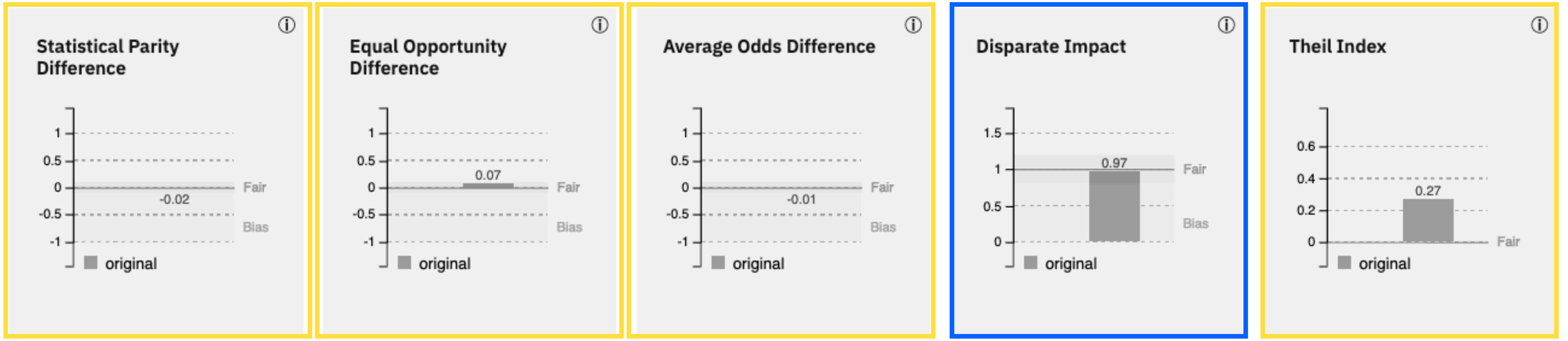
Demo: aif360.mybluemix.net

2. Check bias metrics

Dataset: German credit scoring
Mitigation: none

Protected Attribute: Sex

Privileged Group: ***Male***, Unprivileged Group: ***Female***
Accuracy with no mitigation applied is 75%
With default thresholds, bias against unprivileged group detected in 0 out of 5 metrics



O número ideal é zero

O número ideal é um

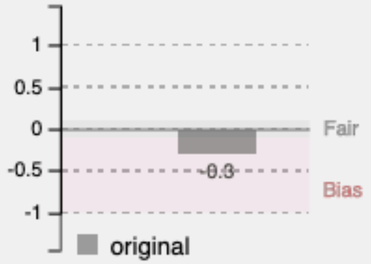
Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

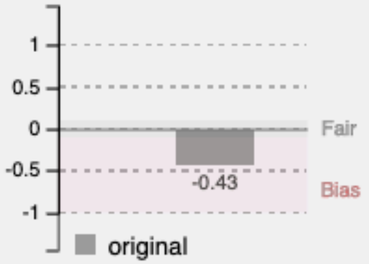
Accuracy with no mitigation applied is 75%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

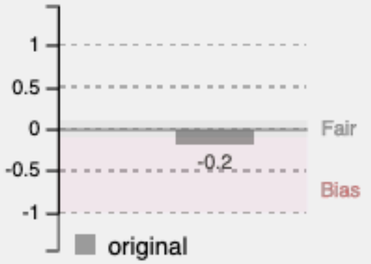
Statistical Parity Difference



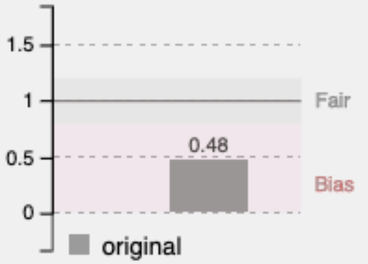
Equal Opportunity Difference



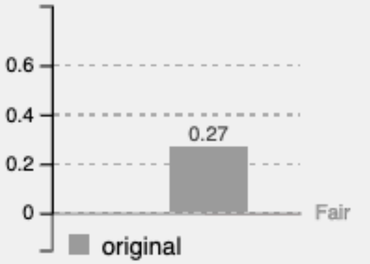
Average Odds Difference



Disparate Impact



Theil Index



3. Choose bias mitigation algorithm

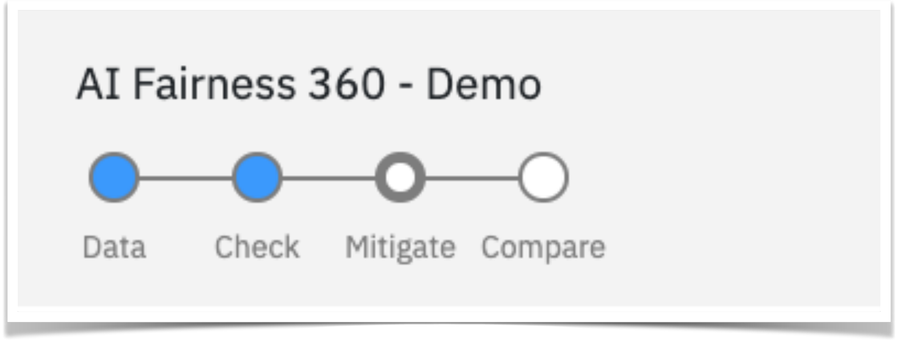
A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process). Learn more about how to choose.

◯ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

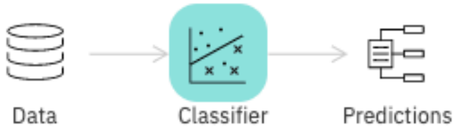
◯ **Optimized Pre-Processing**

Learns a probabilistic transformation that can modify the features and the labels in the training data.



◯ **Adversarial Debiasing**

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.



◯ **Reject Option Based Classification**

Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

4. Compare original vs. mitigated results

Dataset: German credit scoring

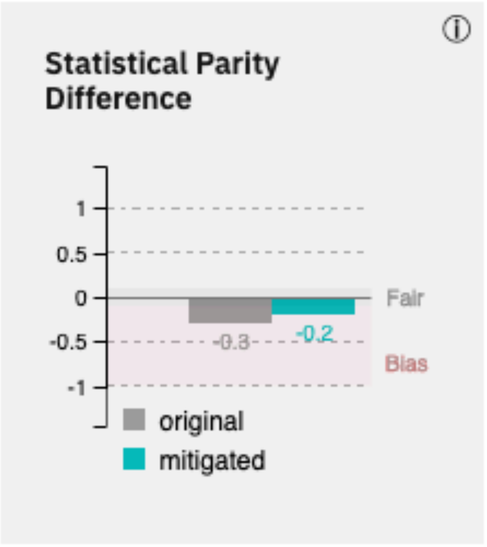
Mitigation: **Reweighing algorithm applied**

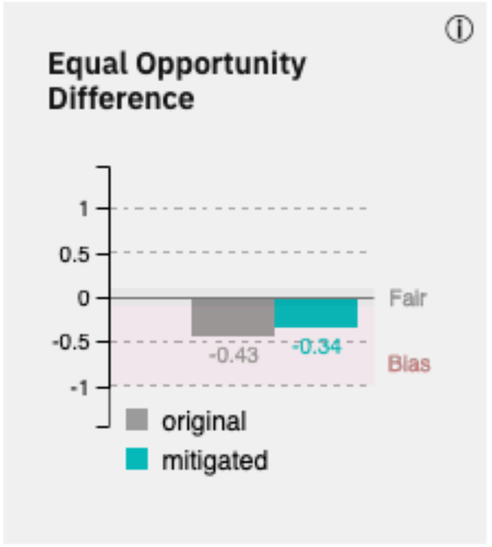
Protected Attribute: Age

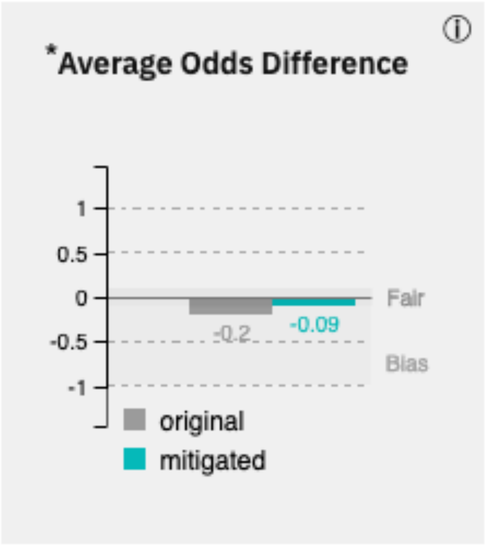
Privileged Group: ***Old***, Unprivileged Group: ***Young***

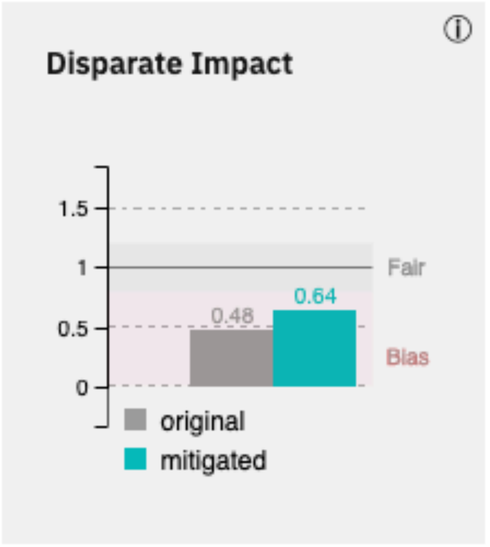
Accuracy after mitigation changed from 75% to 74%

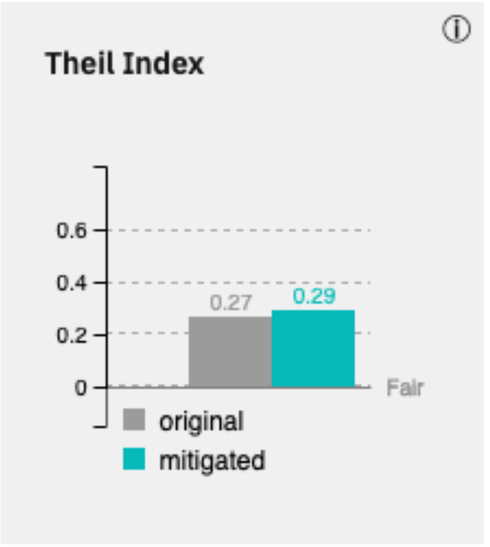
Bias against unprivileged group was reduced to acceptable levels* for 1 of 4 previously biased metrics (3 of 5 metrics still indicate bias for unprivileged group)











É nossa responsabilidade construir sistemas que sejam justos.

FACES OF OPEN SOURCE

Photo by Peter Adams - http://www.facesofopensource.com/

OBRIGADA!

K-ROZ.COM

@GDEQUEIROZ

CONTACTKROZ@GMAIL.COM

AI Inclusive
ai-inclusive.org

ai-inclusive.org

AI_INCLUSIVE

INFO@AI-INCLUSIVE.ORG