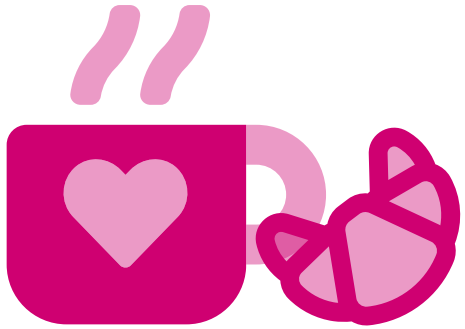
Herzlich Willkommen zum DIVERSITY BREAKFAST



Unconscious Bias und KI

{diversity|think|tank}

der; pluralistisch vielfältig, all gender, Herkunft: total egal; **hochspezialisierte Unternehmensberatung** im Bereich Diversity Management und Inklusion; klar ergebnisorientiert; **innovative Trainings, Tools und Events**; interkulturelle Sensibilisierung; gesamtheitliches Personalmanagement; **Vielfalt als Erfolgsfaktor.**

www.diversitythinktank.at
www.diversitycampus.eu

Diversity Breakfast

- Aktuelle Diversitätsthemen aus mehreren Blickwinkeln
- Good Practices von Unternehmen
- Fragen & Antworten sowie Austauschmöglichkeit
- Vier Mal pro Jahr als Online-Event
- Immer 90 Minuten
- Frühstück an die Tür geliefert und/oder Spende an eine NGO

Die heutige Spende geht an den Verein:

#thenewITgirls

Vielen Dank unseren Partner:innen

Unconscious Bias und KI

Der Apfel fällt nicht weit vom Stamm

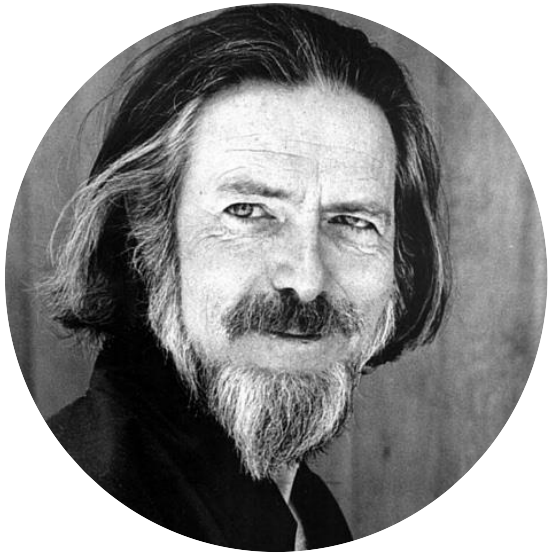
Agenda

1. Unconscious Bias
2. Beispiele für Bias in KI
3. Umgang mit Bias

1. Unconscious Bias

„Das menschliche Bewusstsein ist zwar eine Form des Verstehens, aber zugleich auch immer eine Form der Ignoranz. Unser gewöhnliches, alltägliches Bewusstsein lässt mehr aus, als es aufnimmt.“

Alan Watts



Wie unsere Denkmuster entstehen

Konstruktivismus

Menschliche Wahrnehmung

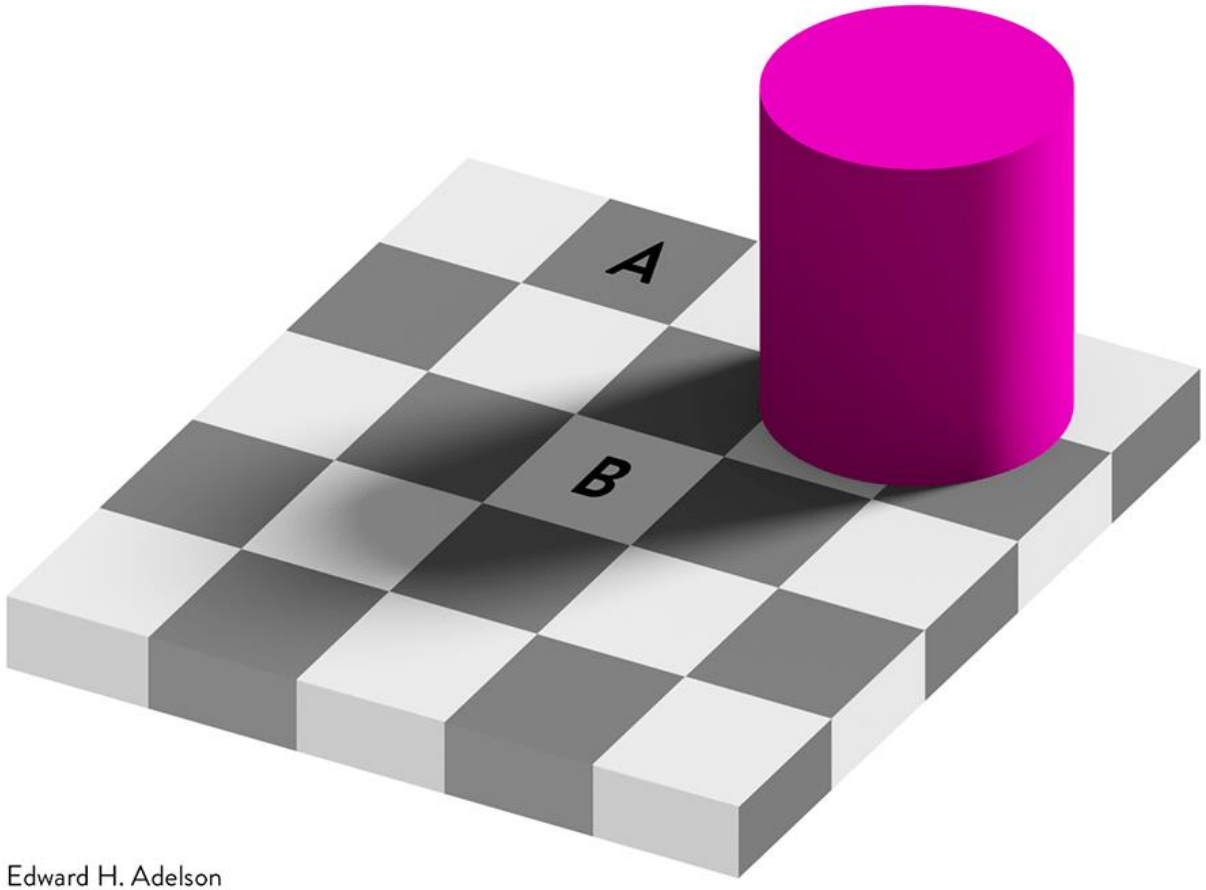
Ludwig Wittgenstein

„Warum sagen die Menschen, es war natürlich zu glauben, die Sonne drehte sich um die Erde?“

„Vermutlich weil es so aussieht, als ob die Sonne um die Erde kreist“

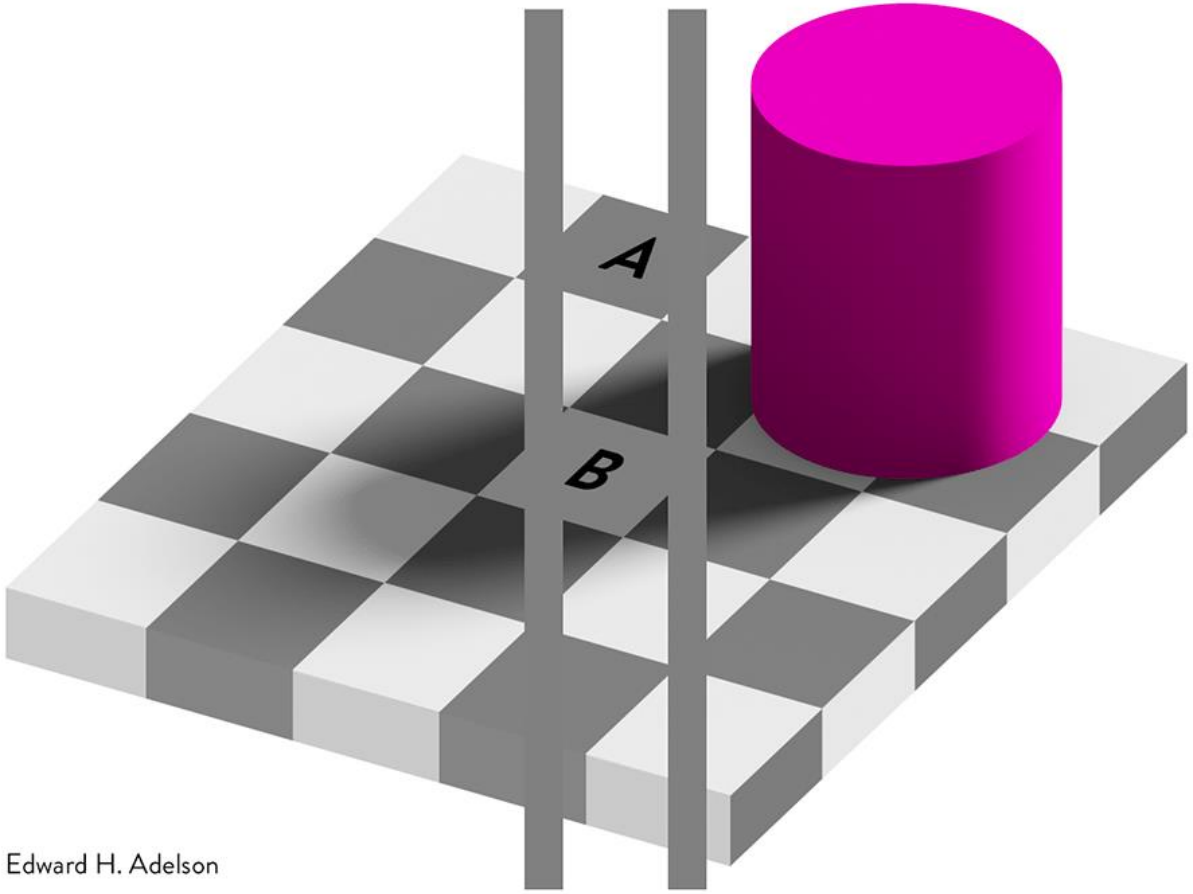
Elisabeth Anscombe

Wahrnehmen



Edward H. Adelson

Und jetzt?



Edward H. Adelson

DIVERSITY CAMPUS

DIVERSITY THINK TANK AUSTRIA

Welche Farbe hat das Kleid?

"Wir sehen die Welt nicht so, wie sie ist.
Wir sehen die Welt, wie wir sind."

Anaïs Nin

„Die meisten Menschen finden nicht nur Trost in ihrer Unwissenheit. Sondern sie zeigen auch Feindseligkeit gegenüber denjenigen, die ihre vorgefassten Meinungen in Frage stellen.“

Platon

2. Beispiele für Bias in KI

AI – Racial Bias

Rona Wang tried to use Playground AI to created a professional LinkedIn photo. Courtesy of Rona Wang

DIVERSITY CAMPUS

DIVERSITY THINK TANK AUSTRIA

Wie nehmen wir KI war?

3. Umgang mit Bias

Strategien zum Umgang

1. **Bewusstsein schaffen:** Schulen wir uns in unserer Wahrnehmung
2. **Präsenz:** mehr im Hier und Jetzt sein
3. **Dialog:** lasst uns mehr in den Dialog kommen

1. Bewusstsein schaffen

Beobachten

Interpretieren

Bewerten

DIVERSITY CAMPUS

DIVERSITY THINK TANK AUSTRIA

Konstruktivistische Bescheidenheit

2. Im Hier und Jetzt sein

3. In den Dialog kommen

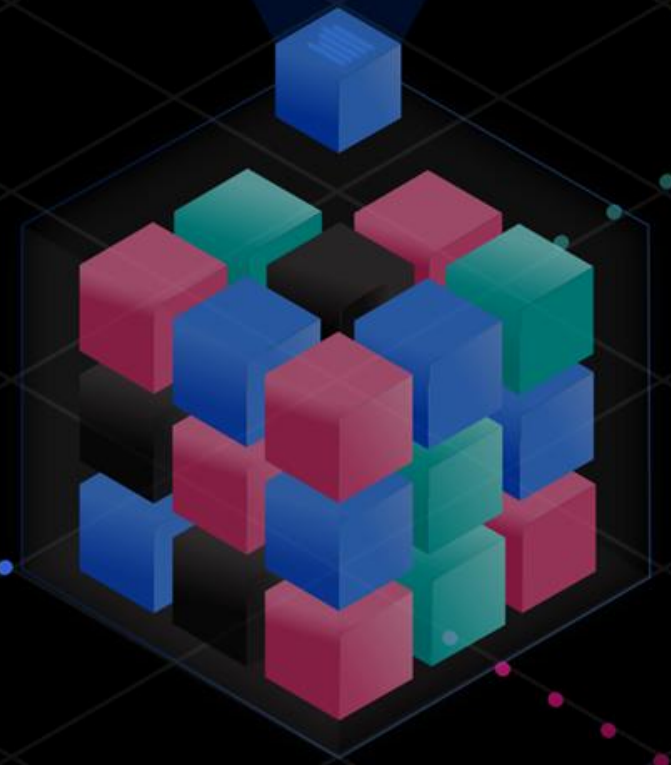
Wenn du einen Menschen siehst, frage dich, ob du ihn/sie wirklich siehst, oder ob du nur deine Gedanken über ihn/sie siehst.

Kann KI uns helfen, weniger gebiased durchs Leben zu gehen?

Coming soon: Project Bias Explorer

KI = bias

Thomas Jirku, IBM



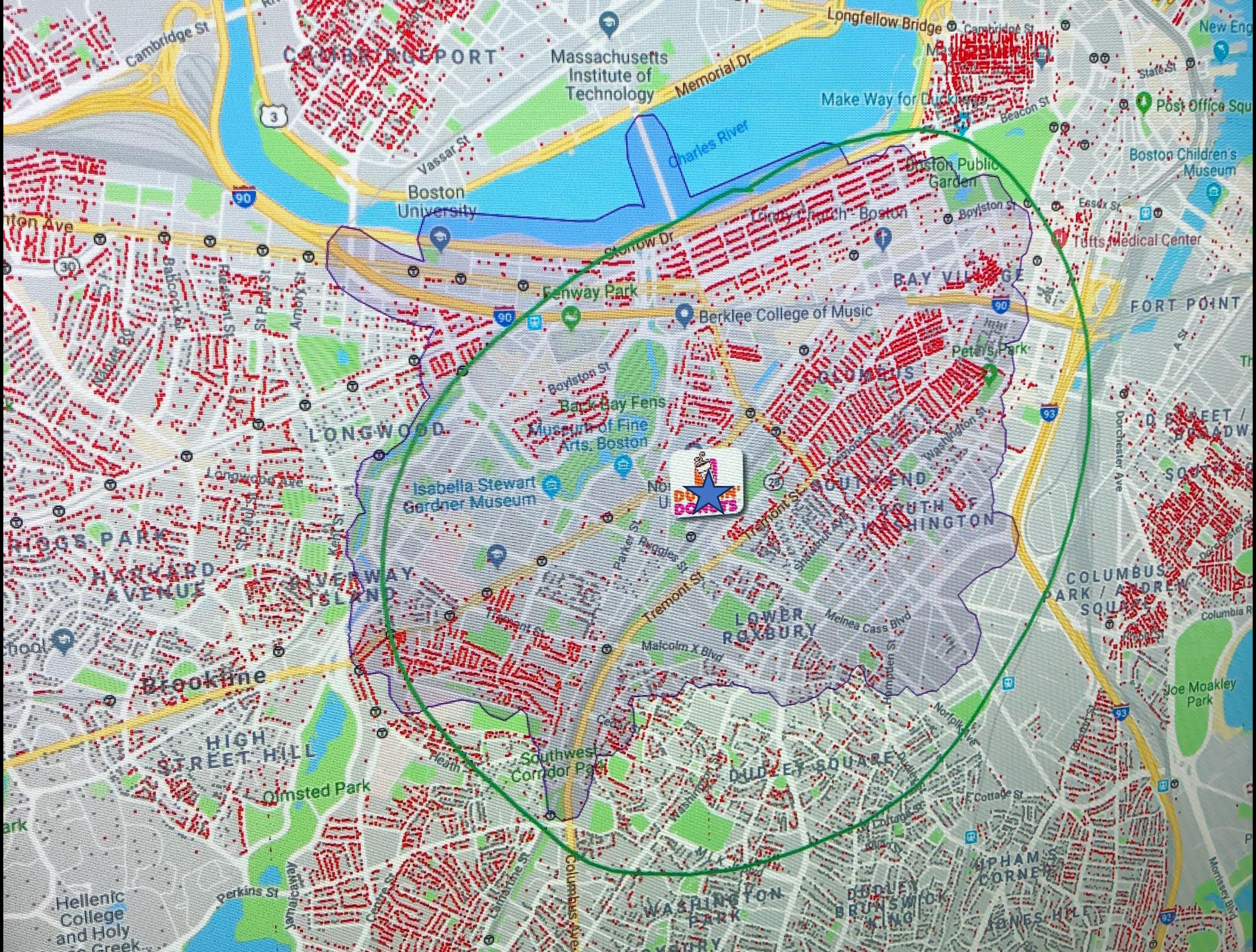
people

data

algorithms

models

actions


Cambridge St
CAMBRIDGEPORT
Massachusetts Institute of Technology
Memorial Dr
Longfellow Bridge
Make Way for Ducklings
Post Office Square
Charles River
Vassar St
Boston University
Boston Public Garden
Boston Children's Museum
Trinity Church Boston
Tufts Medical Center
Storrow Dr
Boylston St
BAY VILLAGE
Fenway Park
Berklee College of Music
FORT POINT
Back Bay Fens
Museum of Fine Arts, Boston
LONGWOOD
Longwood Ave
Isabella Stewart Gardner Museum
SOUTH END
Tremont St
Ruggles St
Dorchester Ave
LOWER ROXBURY
Malcolm X Blvd
Melnea Cass Blvd
COLUMBUS PARK / ANDREW SQUARE
Columbia Rd
Joe Moakley Park
Brookline
HIGH STREET HILL
Olmsted Park
Southwest Corridor Park
DUDLEY SQUARE
UPHAMS CORNER
Perkins St
Columbus Ave

MONEYBOX

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

By JORDAN WEISSMANN

OCT 10, 2018 • 4:52 PM

BE_INT
= f (0,10
– 0,14 x GESCHLECHT_WEIBLICH
– 0,13 x ALTERSGRUPPE_30_49
– 0,70 x ALTERSGRUPPE_50_PLUS
+ 0,16 x STAATENGRUPPE_EU
– 0,05 x STAATENGRUPPE_DRITT
+ 0,28 x AUSBILDUNG_LEHRE
+ 0,01 x AUSBILDUNG_MATURA_PLUS
– 0,15 x BETREUUNGSPFLICHTIG
– 0,34 x RGS_TYP_2
– 0,18 x RGS_TYP_3
– 0,83 x RGS_TYP_4
– 0,82 x RGS_TYP_5
– 0,67 x BEEINTRÄCHTIGT
+ 0,17 x BERUFSGRUPPE_PRODUKTION
– 0,74 x BESCHÄFTIGUNGSTAGE_WENIG
+ 0,65 x FREQUENZ_GESCHÄFTSFALL_1
+ 1,19 x FREQUENZ_GESCHÄFTSFALL_2
+ 1,98 x FREQUENZ_GESCHÄFTSFALL_3_PLUS
– 0,80 x GESCHÄFTSFALL_LANG
– 0,57 x MN_TEILNAHME_1
– 0,21 x MN_TEILNAHME_2
– 0,43 x MN_TEILNAHME_3)

3. Geschlecht?


CLOUD
SOLUTION 4
SOLUTION 2


SCIENCE RULES!

THE DATA ANALYSIS PROCESS

It's OK!

I removed the **[GENDER]** *column!*

```
ALTER TABLE "client_data_for_model" DROP "gender";
```

Proxy features

Model features that are highly correlated
or
associated with another feature


Problems
Solutions

explainability

Explain a transaction

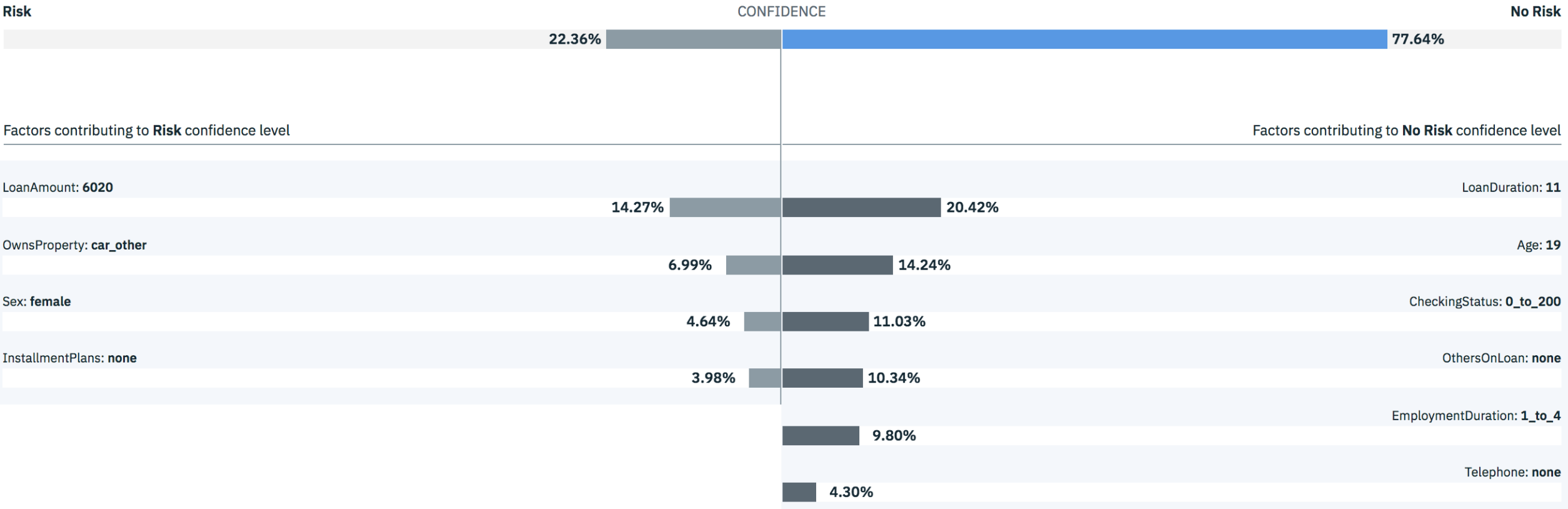
0b9b0e1d-7022-4efc-... × 688bffd7-1fe1-4d9c-a... × 01a5ea94-e77a-4bda-... × 01a5ea94-e77a-4bda-... ×

Details

Transaction	0b9b0e1d-7022-4efc-a74c-338bbcd41647-3
Deployment	GermanCreditRiskModel
Model Name	GermanCreditRiskModel

Maximum changes allowed for the same outcome

Age	19
LoanDuration	11
CheckingStatus	no_checking



fairness

Data Set

- ○ Payload + Perturbed
- ○ Payload
- ○ Training
- ◉ Debiased

Monitored Feature

Sex

Date and Time

5/31/2019

7:00 PM

After de-bias ▲

82% of the group **female** received favorable outcomes.

Age Fairness ⚠

100% before

97% after

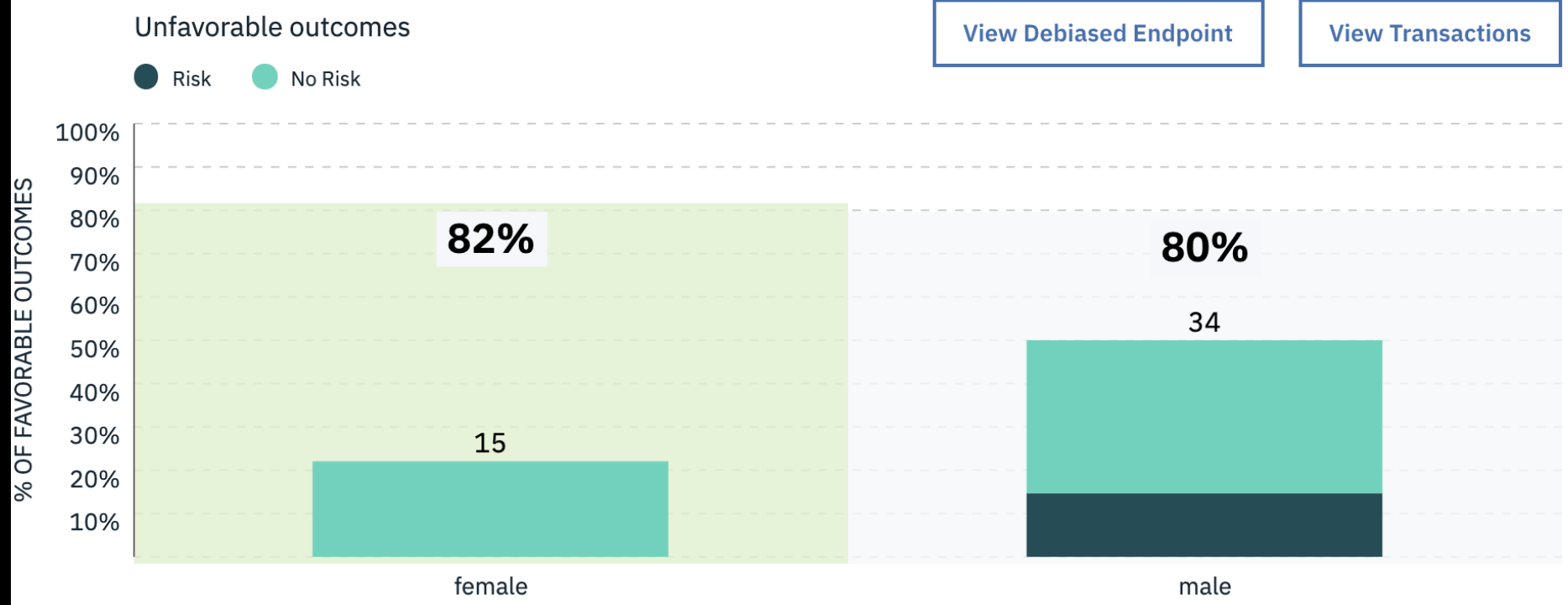
Sex Fairness

85% before

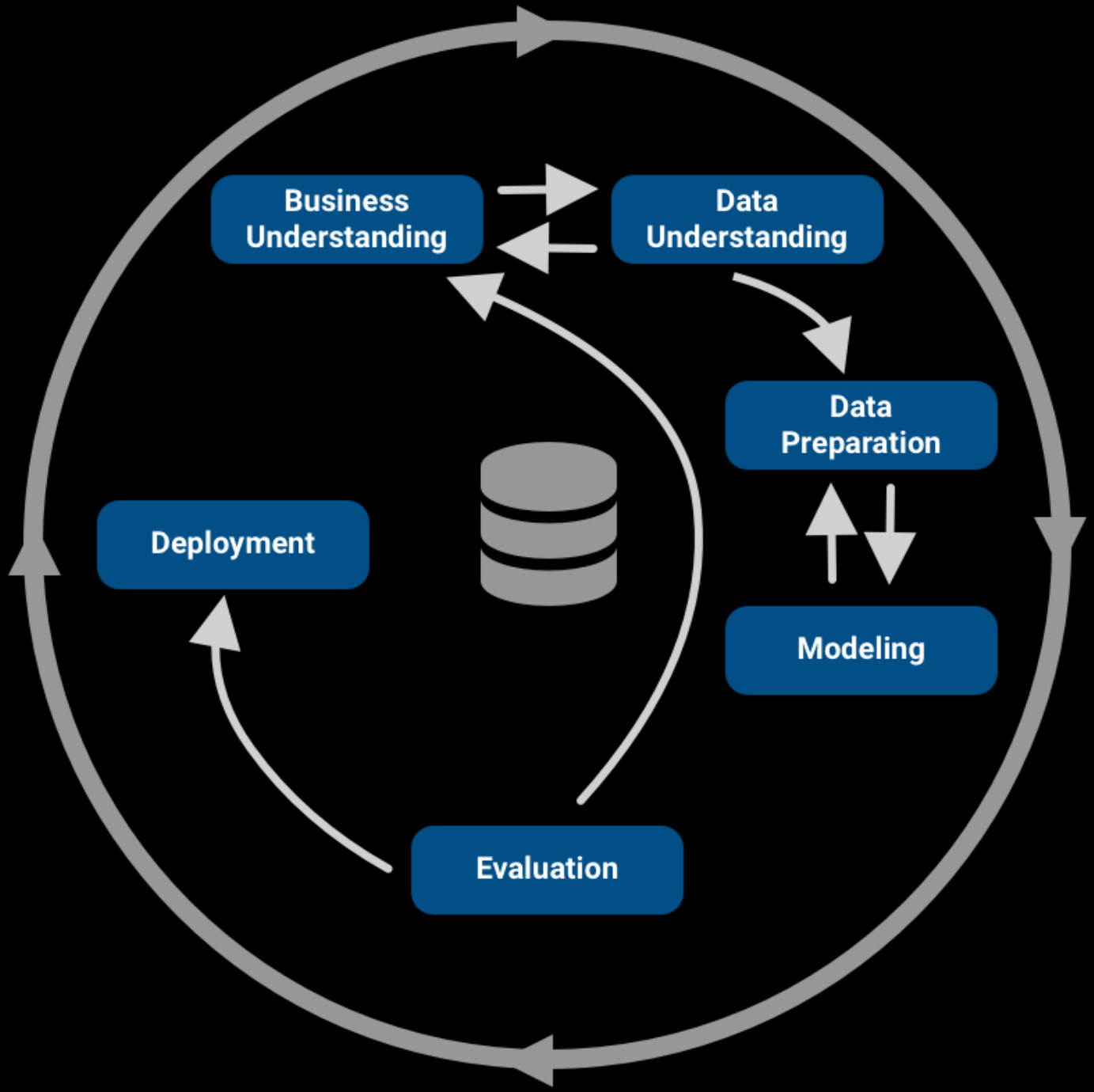
103% after

View Debiased Endpoint

View Transactions



robustness


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

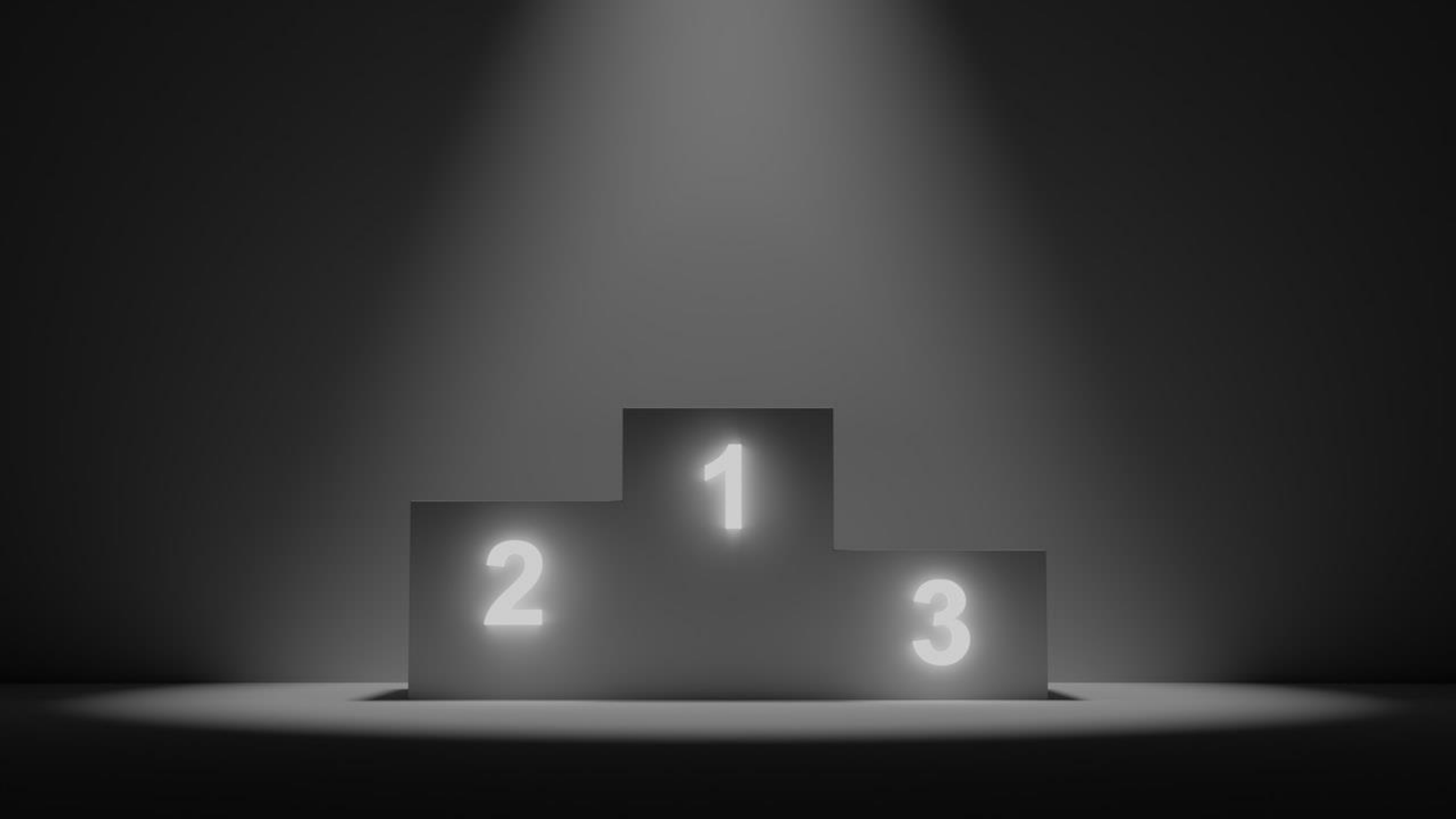
transparency

AI FACTSHEET

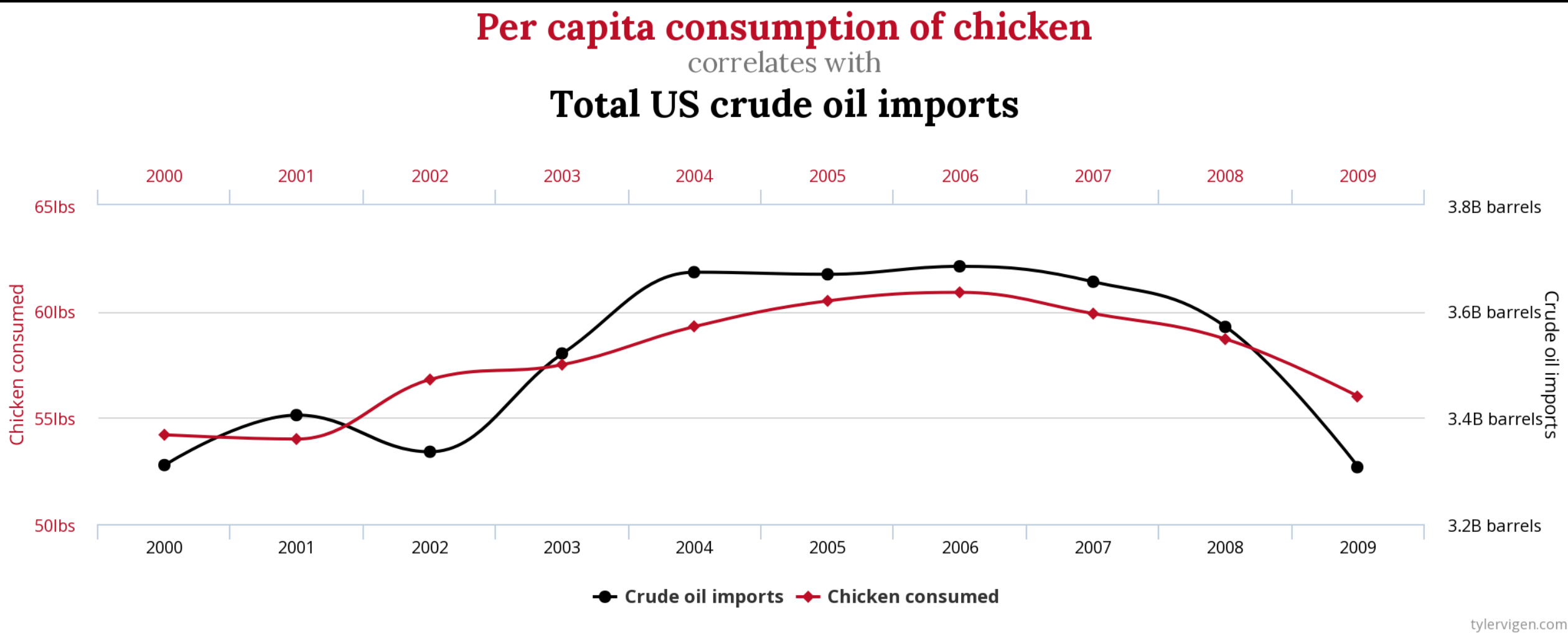
Model Name	Mortgage Evaluator Privacy
Target Audience	Model Risk Officer
Overview	This evaluation tests the vulnerability of the model to select inference attacks and measures certain risk factors that are known to be strongly correlated with privacy risk.
Evaluation Details	• **Membership inference** is a type of attack where, given a trained model and a data sample, one can deduce whether or not that sample was part of the model's training. • **Attribute inference** is an attack where certain sensitive features may be inferred about individuals whose data was included in training a model.

Results

Evaluation	Result
Training set size	650,877
Overfitting	0.0011
Black-box membership inference (BBMI)	0.55
Black-box attribute inference (BBAI) - Age	0.046
Black-box attribute inference - Debt-to-income ratio	0.069
Feature influence - Age	0.003
Feature influence - Debt-to-income ratio	0.678


2
1
3

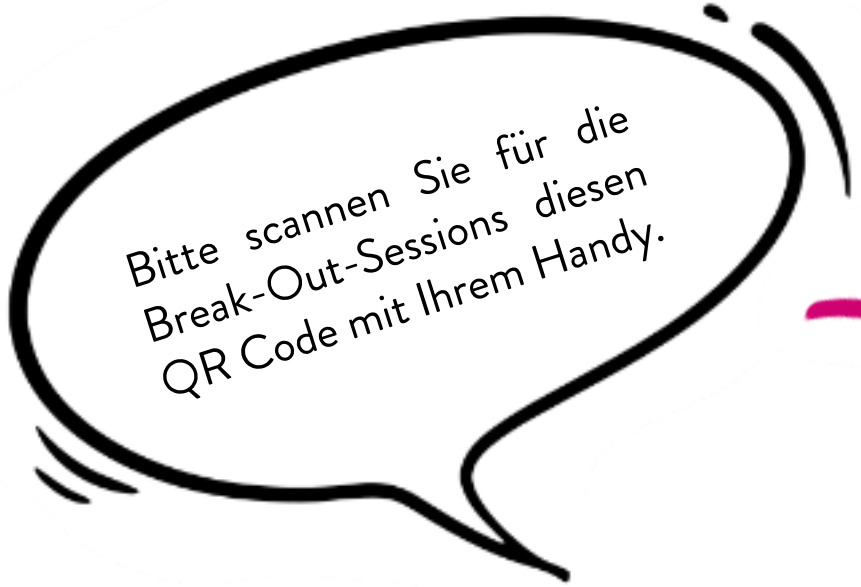

MIND


Per capita consumption of chicken
correlates with
Total US crude oil imports
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
65lbs
60lbs
55lbs
50lbs
Chicken consumed
3.8B barrels
3.6B barrels
3.4B barrels
3.2B barrels
Crude oil imports
Crude oil imports
Chicken consumed
tylervigen.com

Thomas Jirku, IBM

LinkedIn

Thank you!



Frage für den Austausch:

Was können wir tun um Bias in der KI zu verringern?

Ausblick Online-Termine 2024

14. März 2024 Diversity Lunch Break: ESG: S-Faktor im Fokus

25. April 2024 Diversity Breakfast

20. Juni 2024 Diversity Breakfast

Anmeldung auf www.diversitythinktank.at/events

Vielen Dank unseren Partner:innen