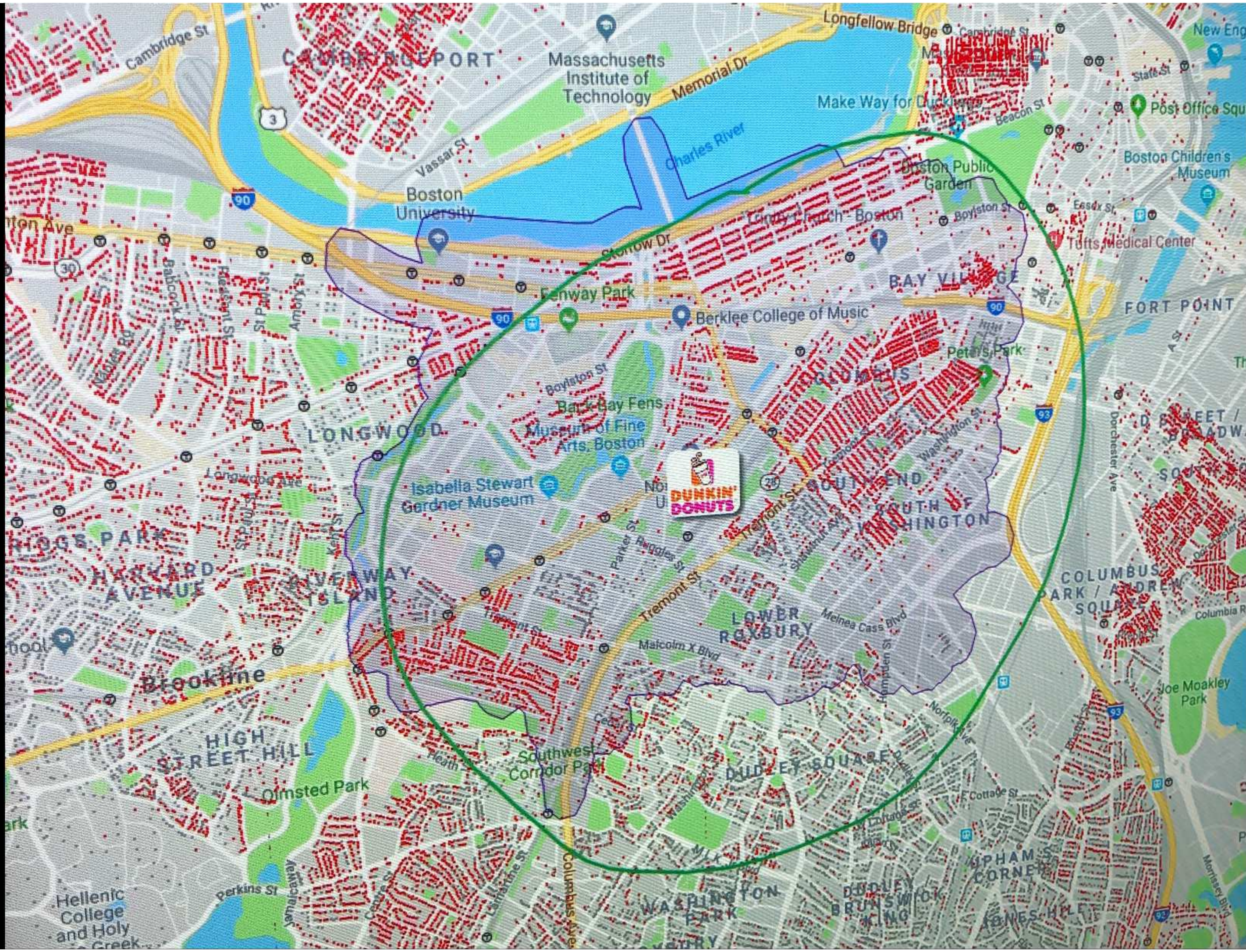
(Un-)trusted AI: Strategien um Diskriminierung durch KI entgegenzuwirken

Thomas Jirku, IBM

where to find discrimination (bias)


people
data
algorithms
models
actions



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


BLICA

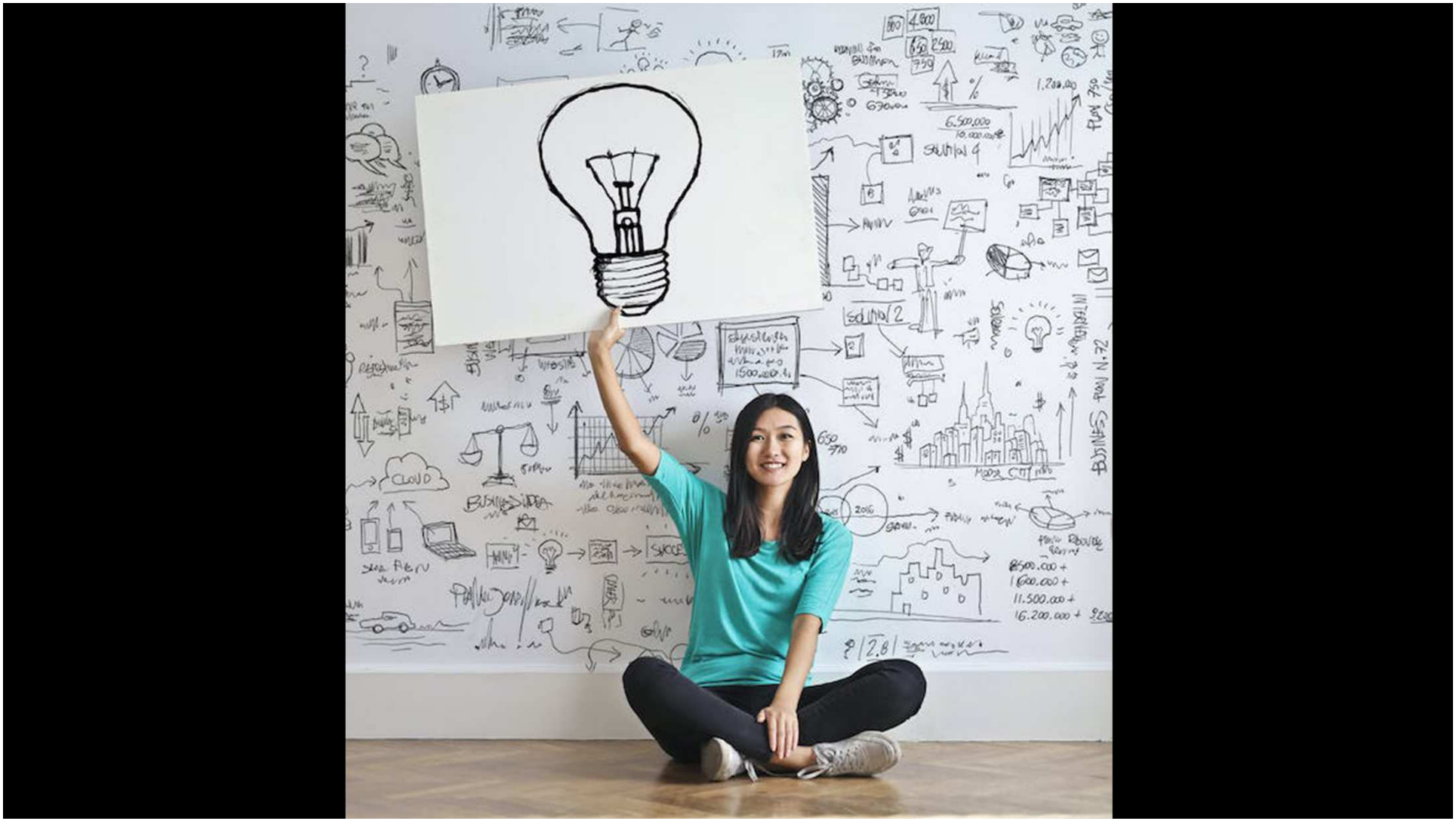
MONEYBOX

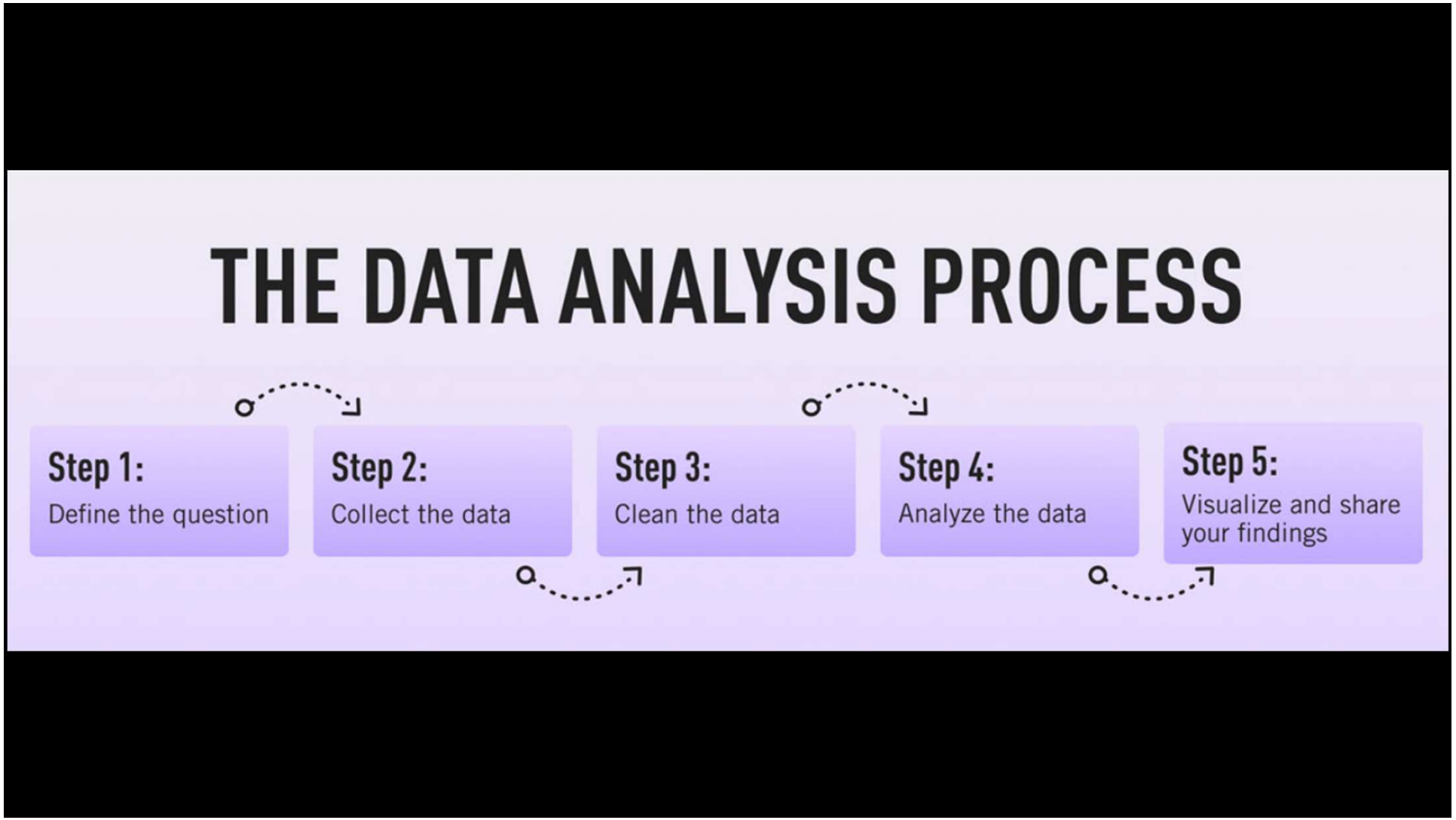
Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

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)

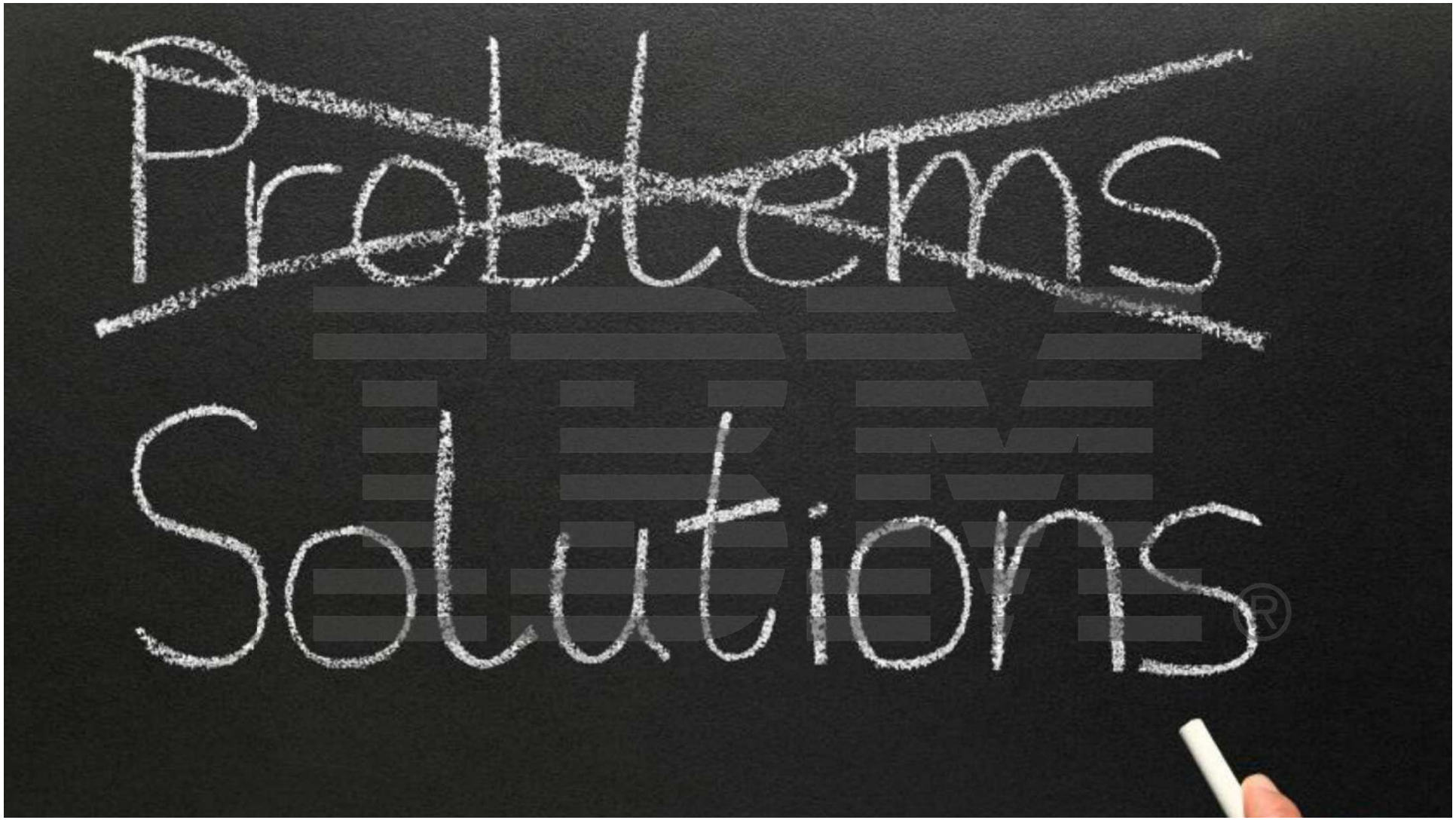

CLOUD
SUCCESS
BUSINESS PLAN N.2


THE DATA ANALYSIS PROCESS
Step 1:
Define the question
Step 2:
Collect the data
Step 3:
Clean the data
Step 4:
Analyze the data
Step 5:
Visualize and share your findings

It's OK!

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

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


Problems
Solutions

explainability

M **Watson OpenScale** | *beta*

Explain a transaction

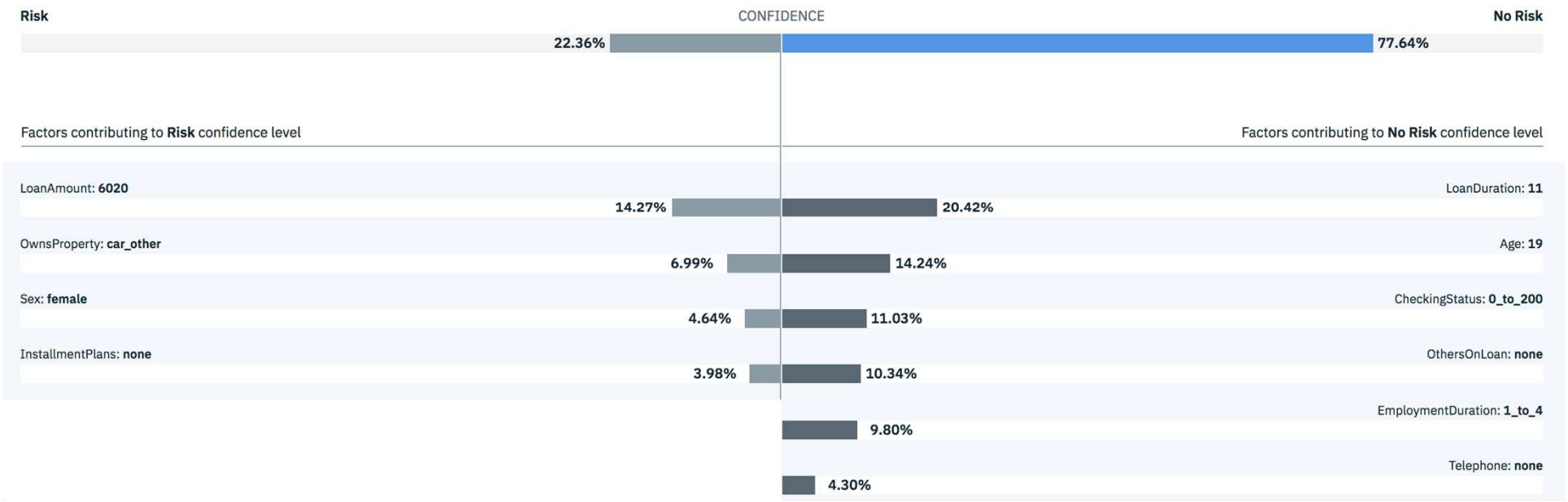
0b9b0e1d-7022-4efc-... × 688bffd7-1fe1-4d9c-a... × 01a5ea94-e77a-4bda-... × 01a5ea94-e77a-4bda-... ×

Details

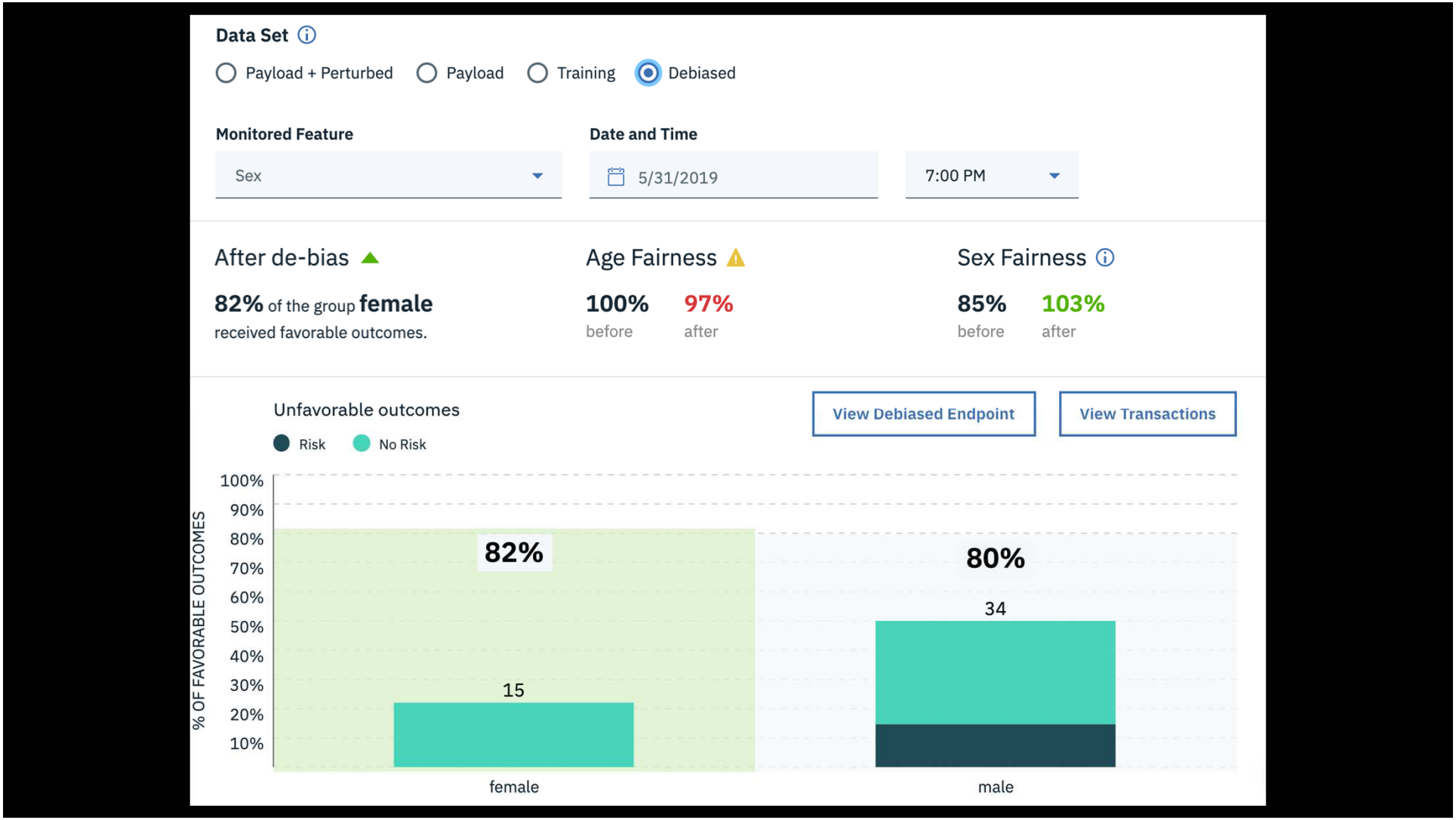
Transaction	0b9b0e1d-7022-4efc-a74c-338bbcd41647-3
Deployment	GermanCreditRiskModel
Model Name	GermanCreditRiskModel

Maximum changes allowed for the same outcome

Age	19
LoanDuration	11
CheckingStatus	no_checking

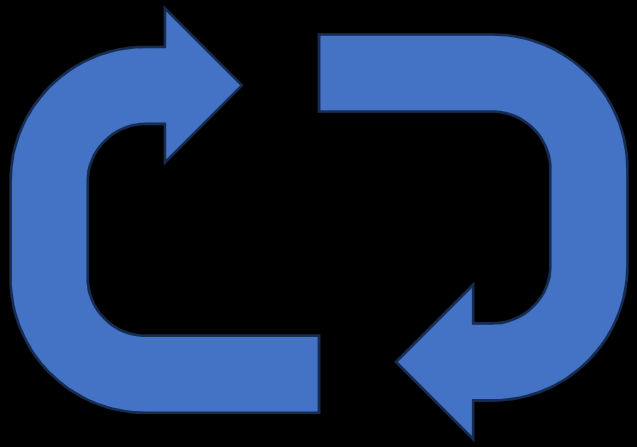


fairness


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
Unfavorable outcomes
View Debiased Endpoint
View Transactions
Risk
No Risk
% OF FAVORABLE OUTCOMES
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
82%
15
80%
34
female
male

robustness

continuous



evaluation

transparency

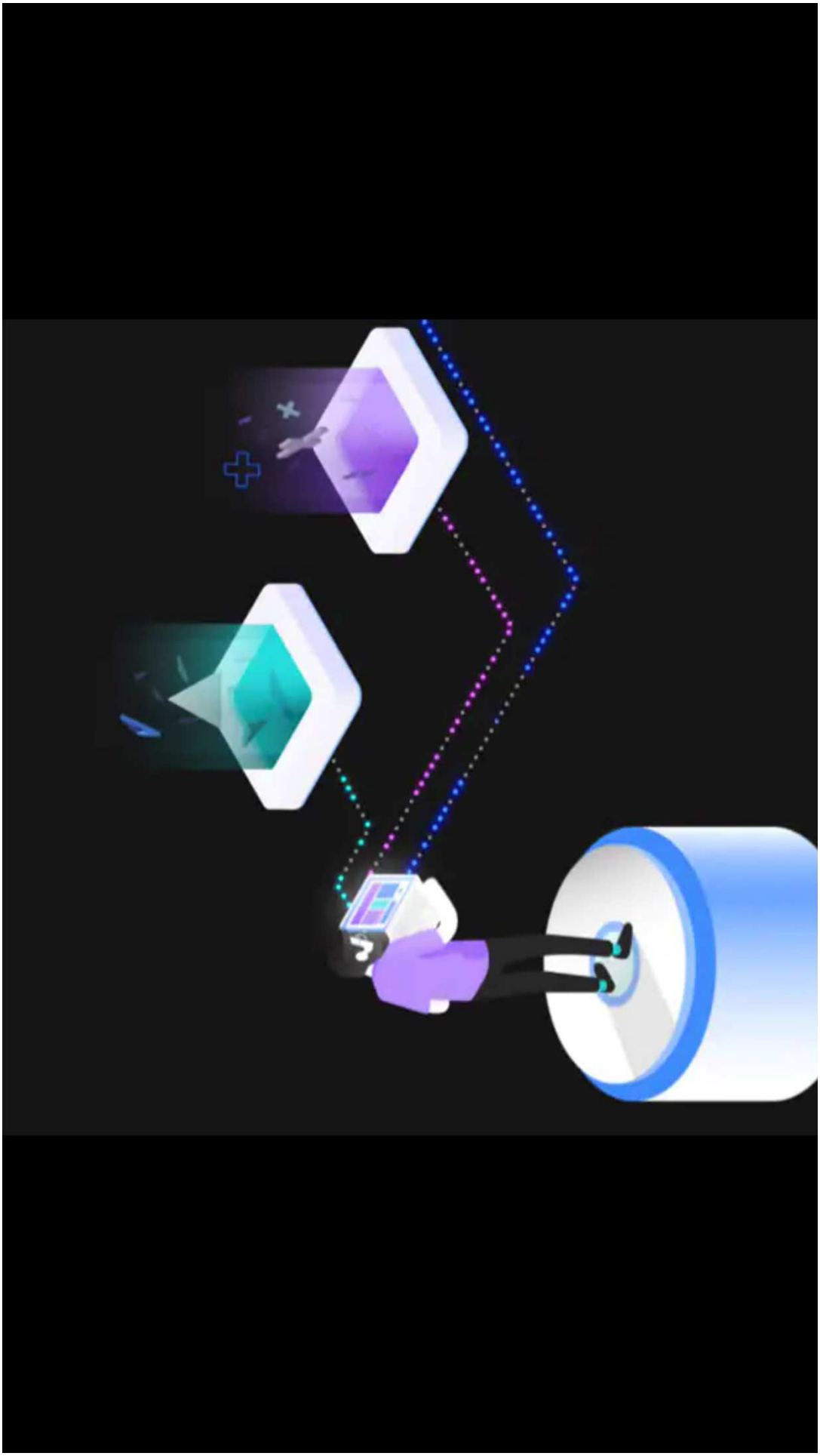
AI FACTSHEET

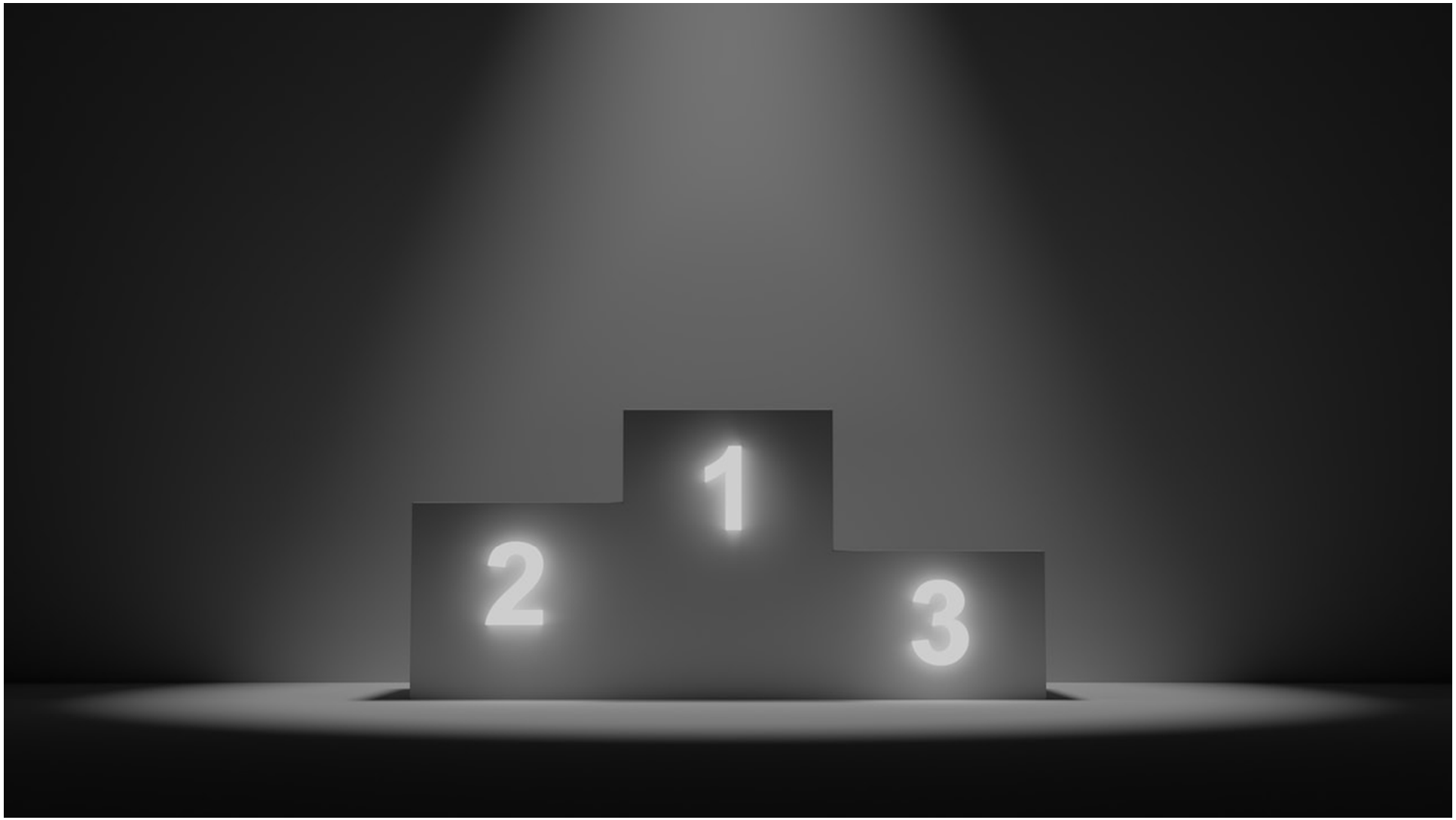
Model Name	Mortgage Evaluator Privacy
Target Audience	Model Risk Officer
Overview	This evaluation tests the vulnerability of the model to select inference attacks and measures certain risk factors that are known to be strongly correlated with privacy risk.
Evaluation Details	• **Membership inference** 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

data privacy




2
1
3

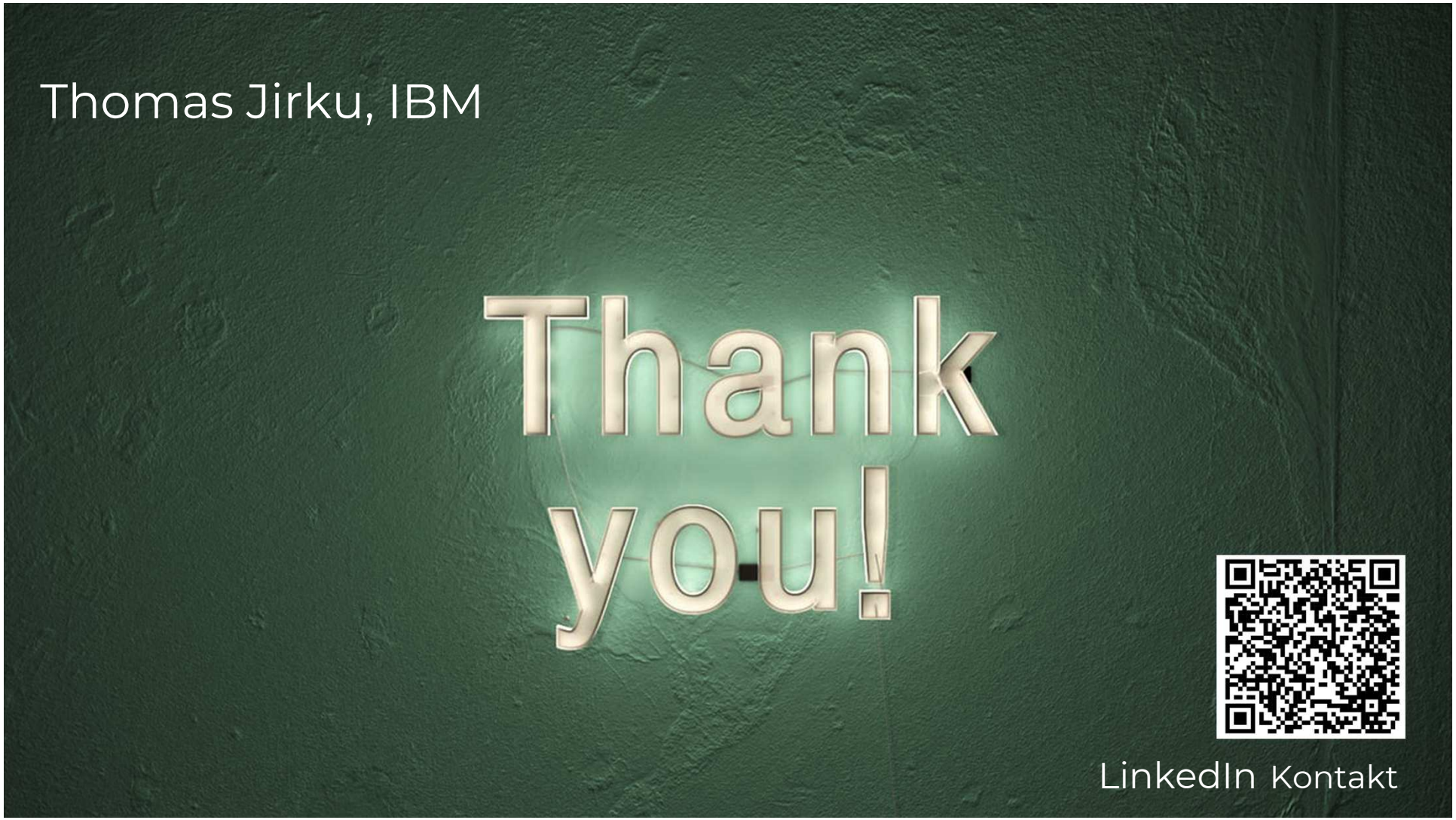
DENIED

literature:

- **angriff der algorithmen**, cathy o`neil
- **weapons of math destruction**, cathy o`neil

- anleitung zum unglücklich sein, paul watzlawick

Thomas Jirku, IBM



LinkedIn Kontakt