Responsible Data Management

Julia Stoyanovich

Computer Science and Engineering &
Center for Data Science
New York University, NY USA

data RESPONSIBLY

NYU

The promise of "AI"

Power

unprecedented data collection

enormous computational power

ubiquity and broad acceptance

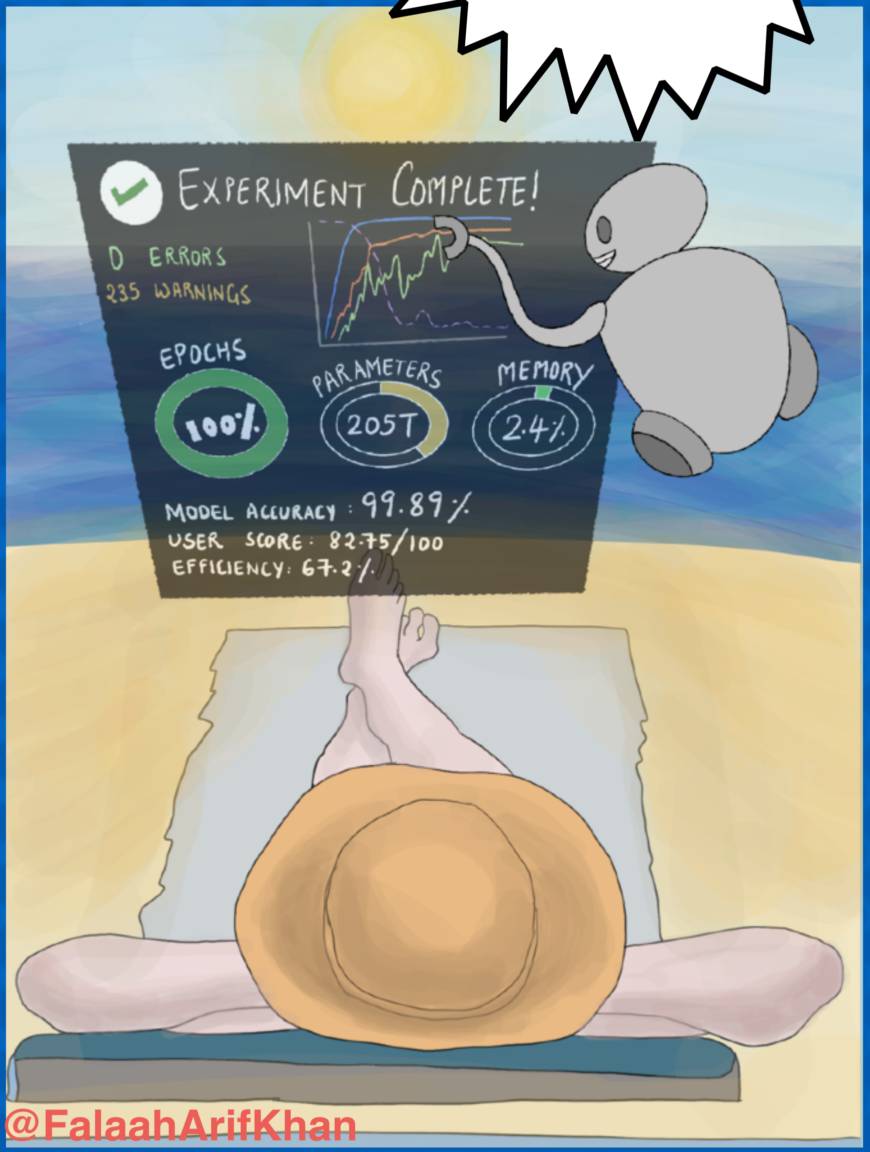
Opportunity

accelerate science

boost innovation

transform government

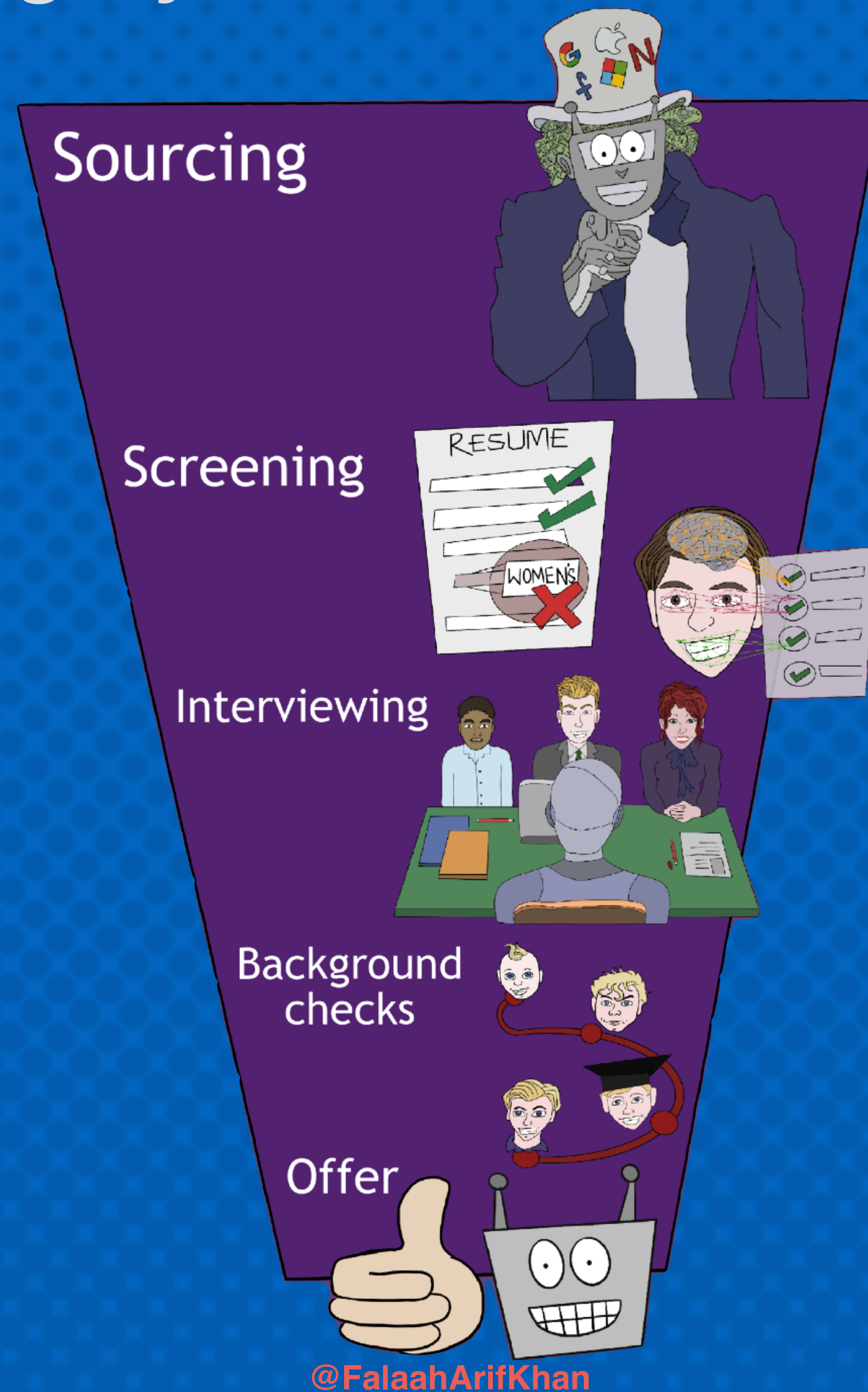
goal



Automated hiring systems

"Automated hiring systems act as modern gatekeepers to economic opportunity."

Jenny Yang



And now… some bad news

theguardian July 2015

Women less likely to be shown ads for high-paid jobs on Google, study shows

MIT Technology Review February 2013

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

REUTERS October 2018

Amazon scraps secret AI recruiting tool that showed bias against women

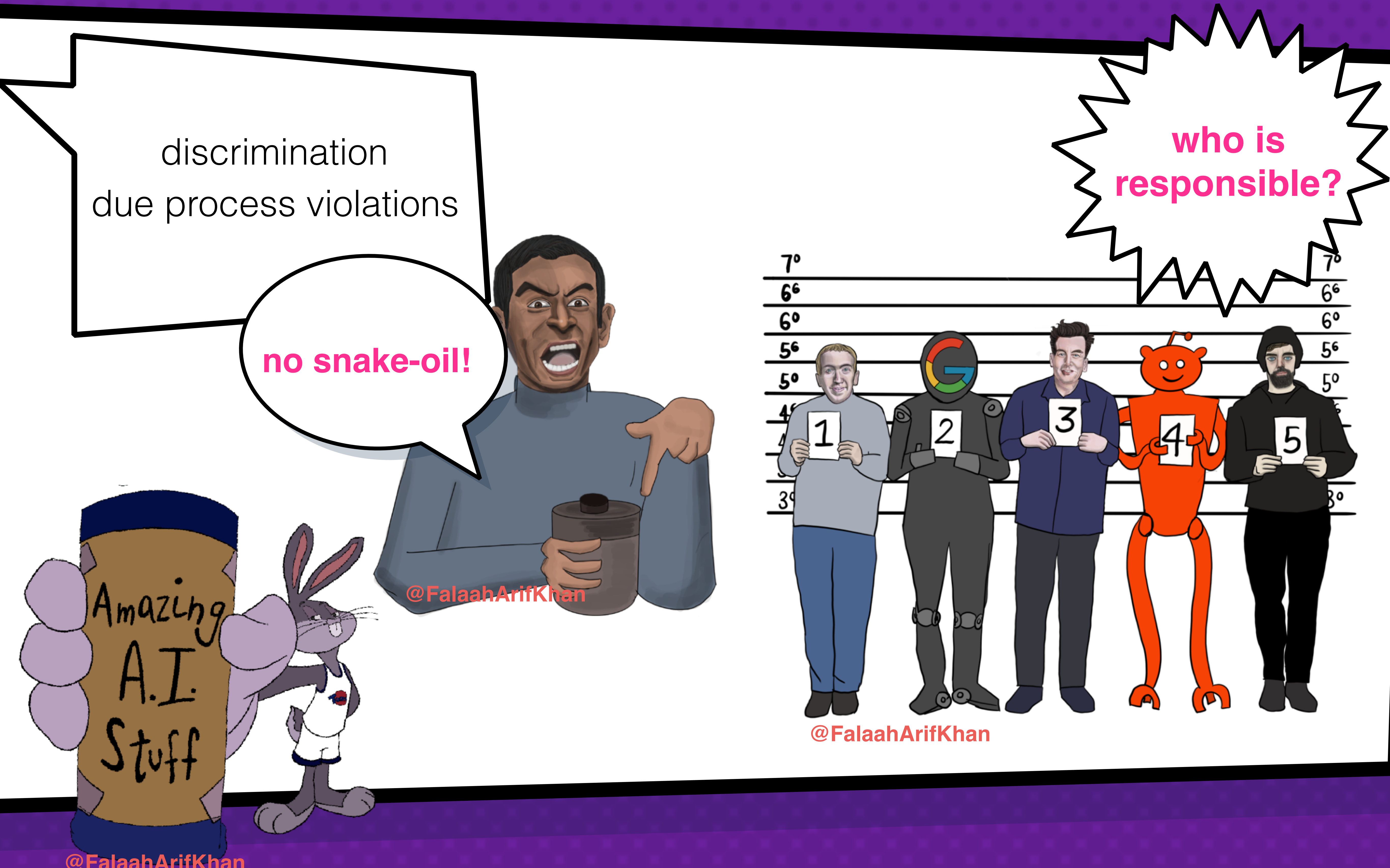
THE WALL STREET JOURNAL. September 2014

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination

@FalaahArifKhan

And now… some bad news



a push for regulation

Automated Decision Systems (ADS)

Automated Decision Systems (ADS)

process data about people

help make consequential decisions

combine human & automated decision making

aim to improve **efficiency** and promote **equity**

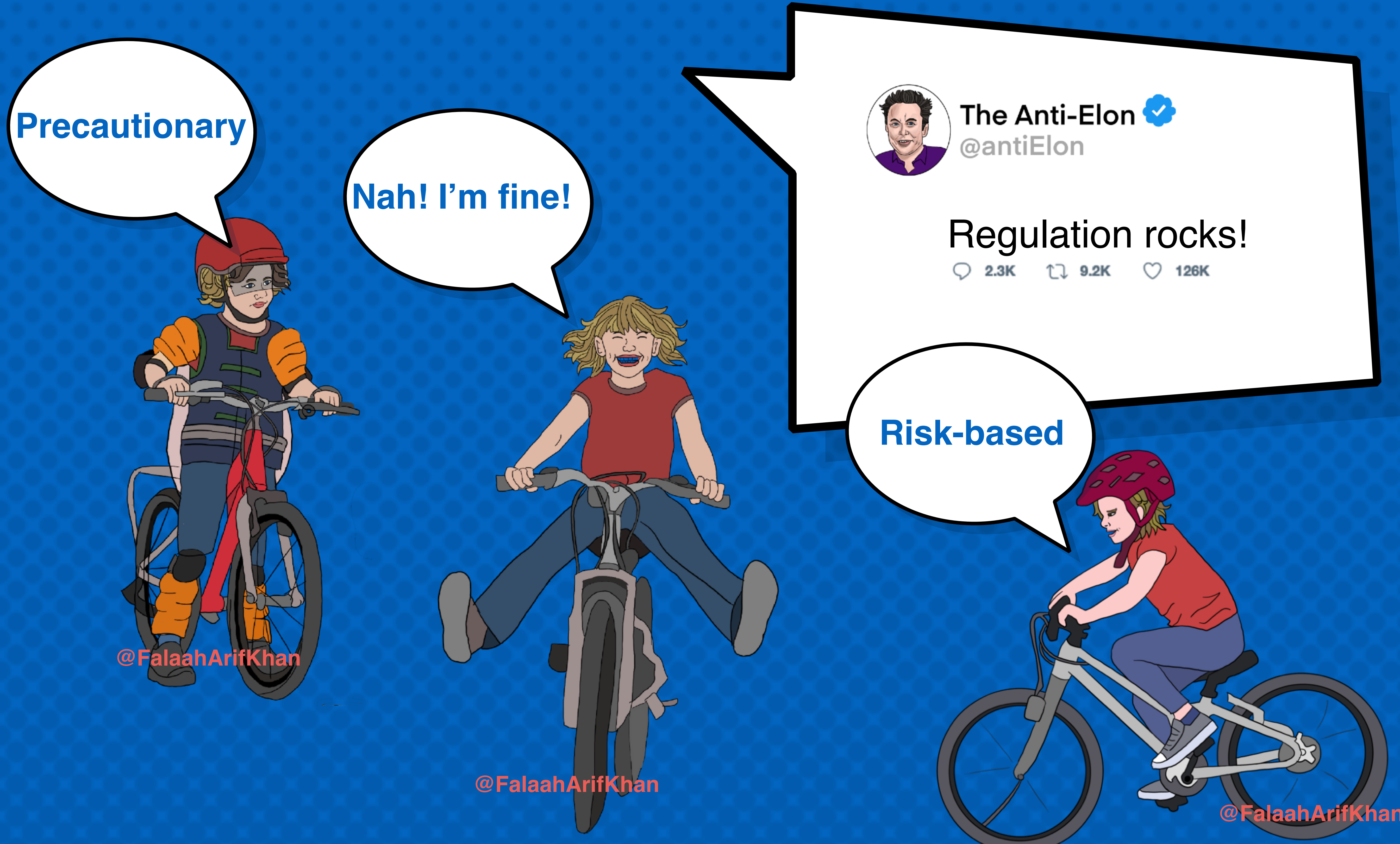
are subject to **auditing** and **public disclosure**

may or may not use AI

may or may not have autonomy

rely heavily on data

Regulating ADS?



ADS regulation in NYC: take 1

@FalaahArifKhan

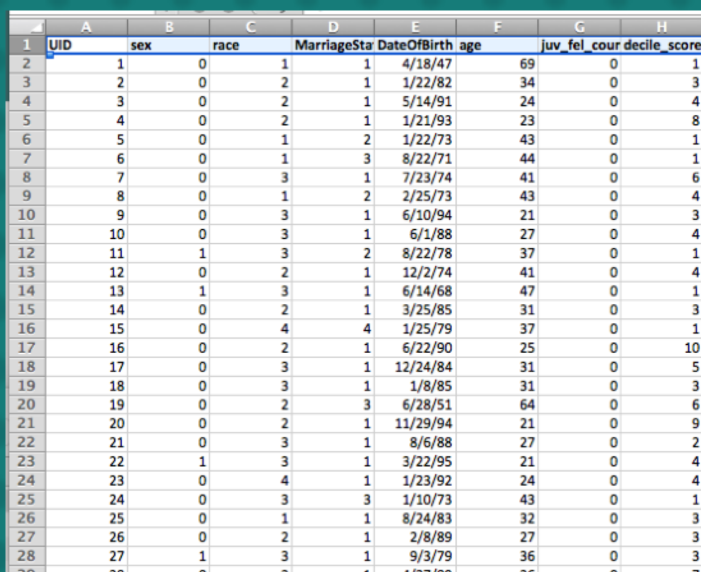
Principles

- using ADS **where** they promote innovation and efficiency in service delivery
- promoting **fairness**, **equity**, **accountability**, and **transparency** in the use of ADS
- reducing potential harm **across the lifespan** of ADS

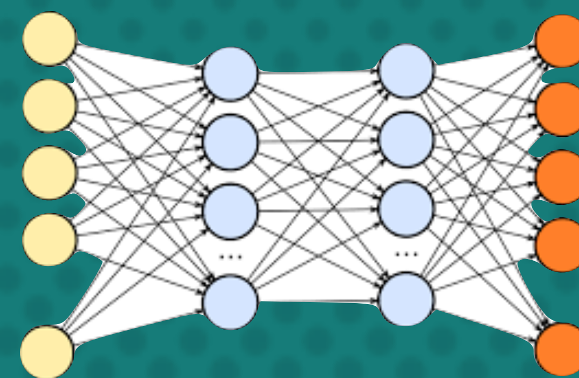

great!
now what?

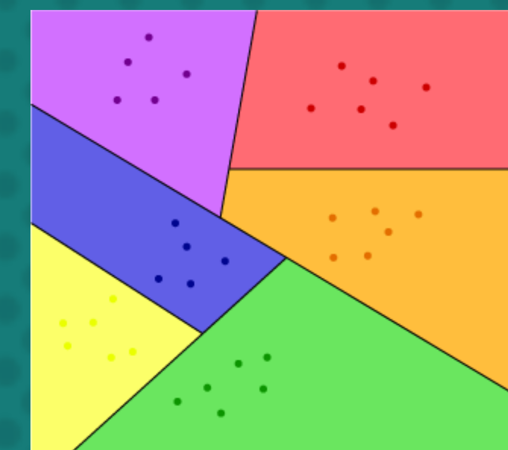
Framing technical solutions

“Bias” in predictive analytics



UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
1	0	1	1	4/18/47	69	0	1
2	0	2	1	1/22/82	34	0	3
3	0	2	1	5/14/91	24	0	4
4	0	2	1	1/21/93	23	0	8
5	0	1	2	1/22/73	43	0	1
6	0	1	3	8/22/71	44	0	1
7	0	3	1	7/23/74	41	0	6
8	0	1	2	2/25/73	43	0	4
9	0	3	1	6/10/94	21	0	3
10	0	3	1	6/1/88	27	0	4
11	1	3	2	8/22/78	37	0	1
12	0	2	1	12/2/74	41	0	4
13	1	3	1	6/14/68	47	0	1
14	0	2	1	3/25/85	31	0	3
15	0	4	4	1/25/79	37	0	1
16	0	2	1	6/22/90	25	0	10
17	0	3	1	12/24/84	31	0	5
18	0	3	1	1/8/85	31	0	3
19	0	2	3	6/28/51	64	0	6
20	0	2	1	11/29/94	21	0	9
21	0	3	1	8/6/88	27	0	2
22	1	3	1	3/22/95	21	0	4
23	0	4	1	1/23/92	24	0	4
24	0	3	3	1/10/73	43	0	1
25	0	1	1	8/24/83	32	0	3
26	0	2	1	2/8/89	27	0	3
27	1	3	1	9/3/79	36	0	3





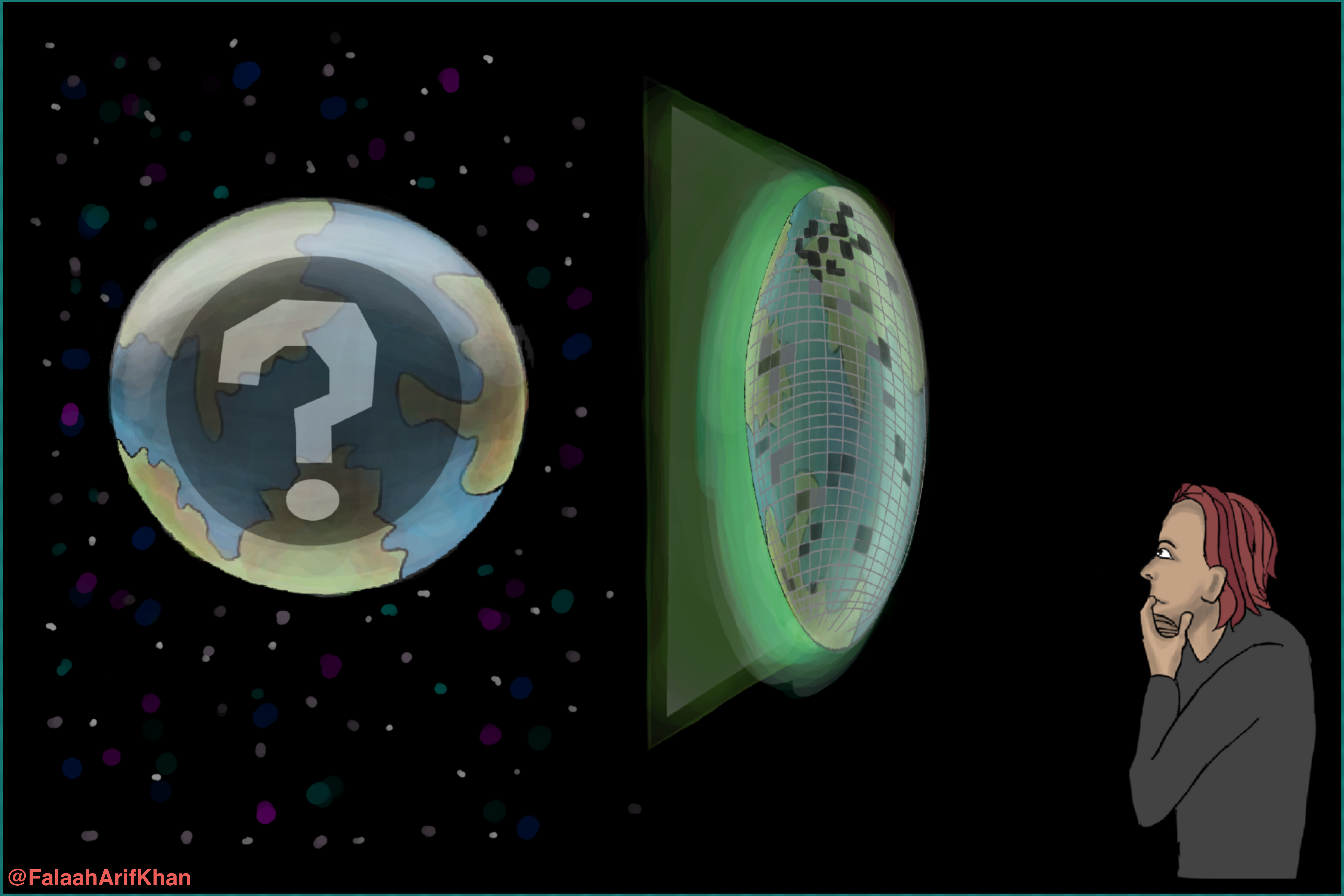
Statistical

model does not summarize the data correctly

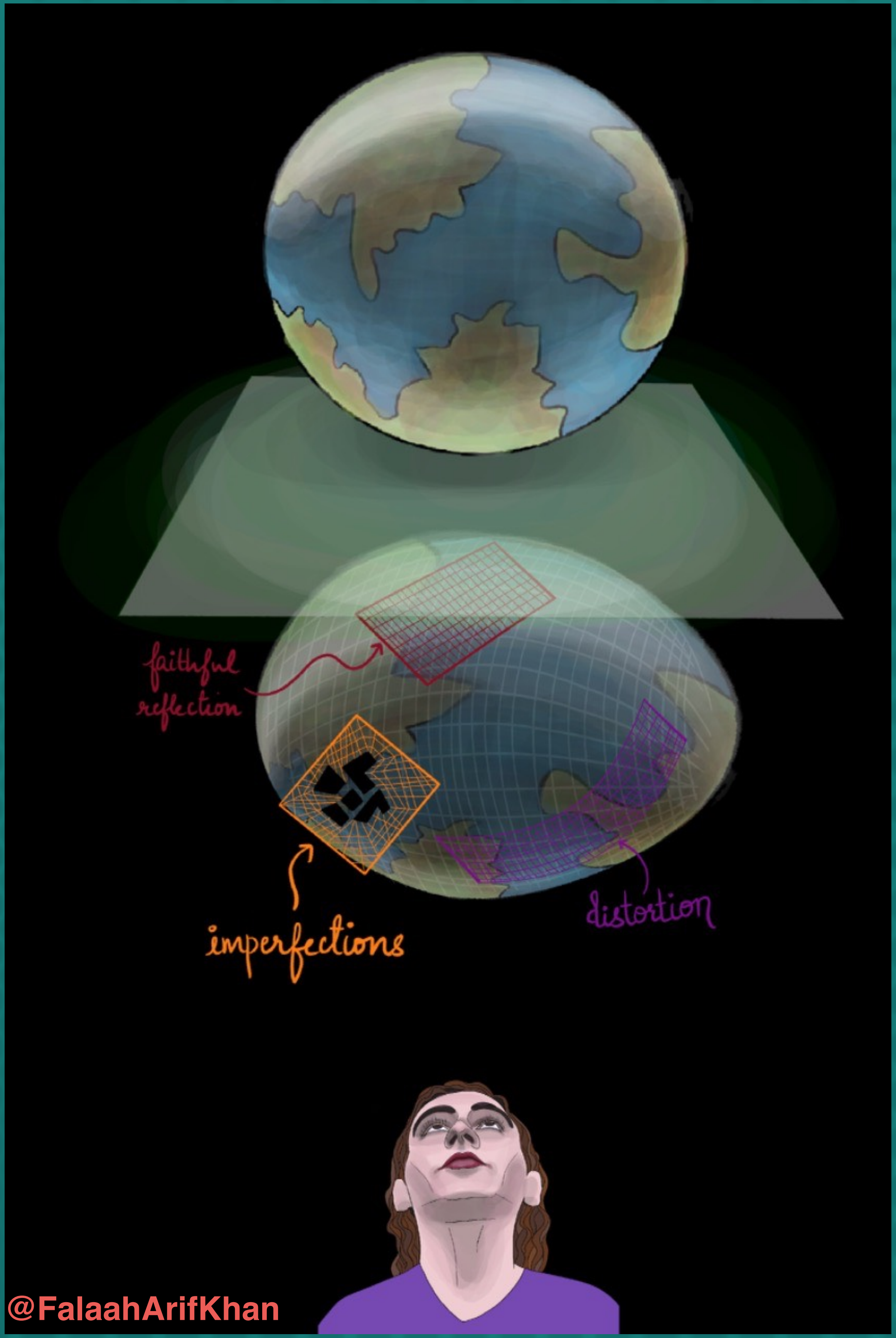
Societal

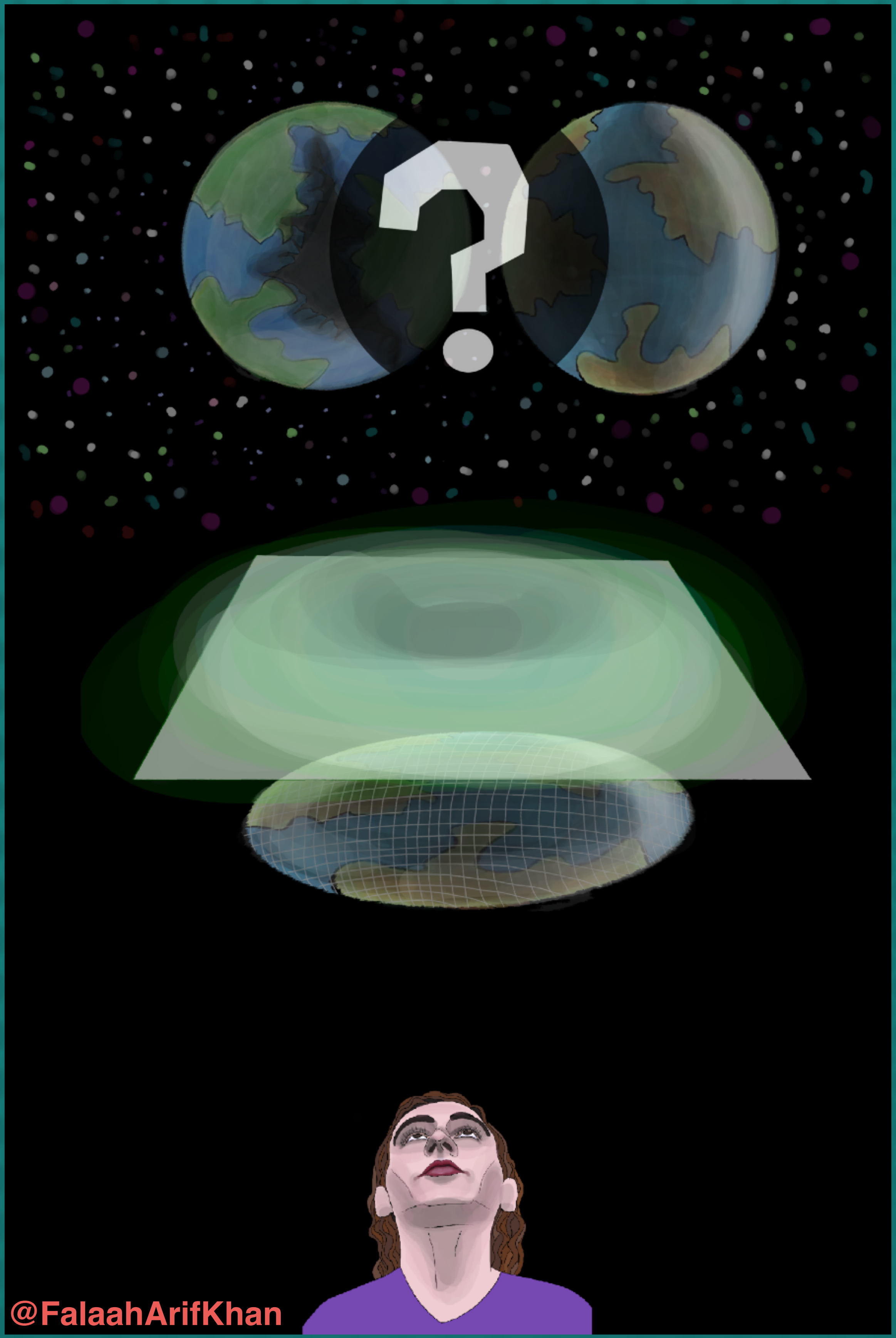
data does not represent the world correctly

Data, a reflection of the world

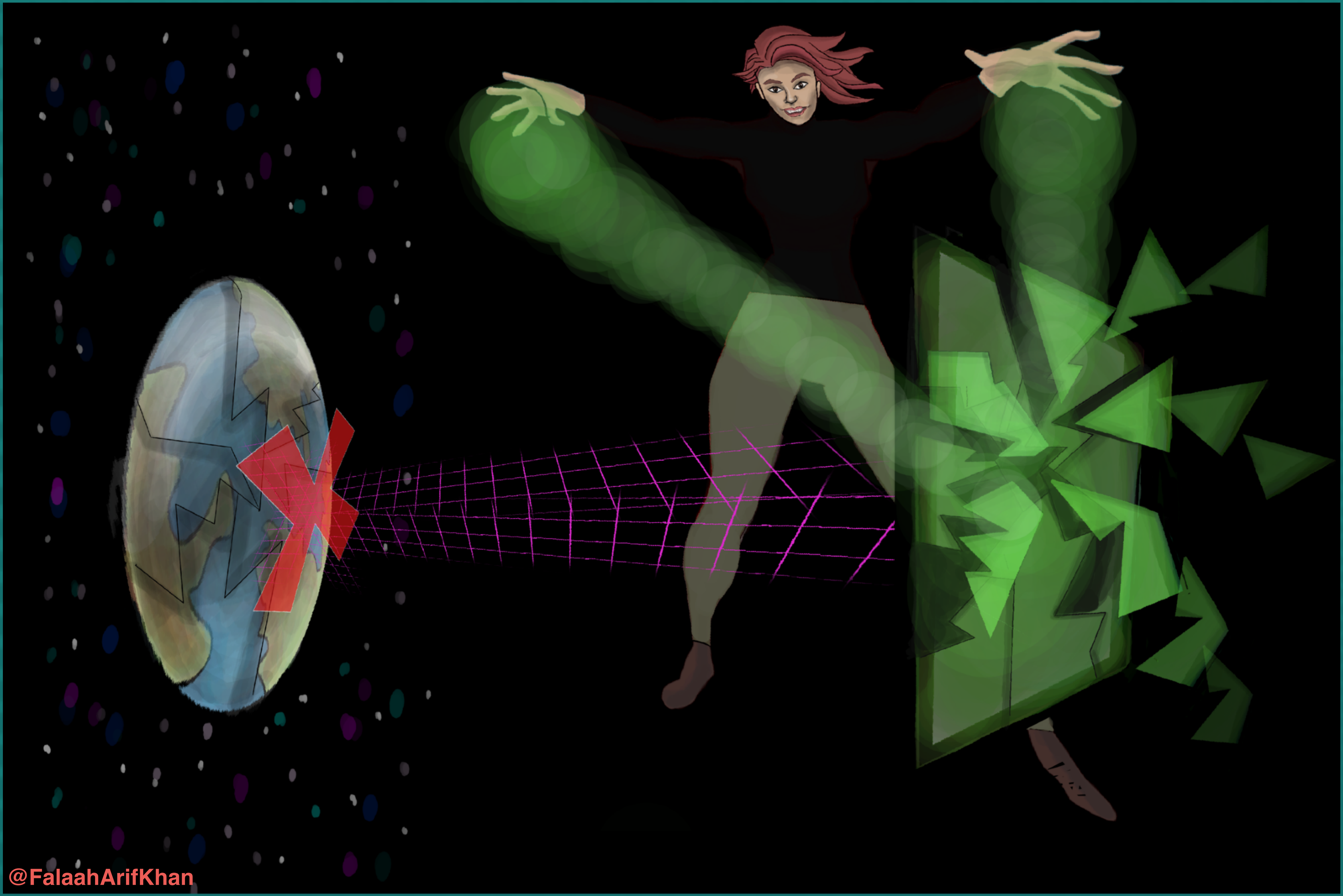


Data, a reflection of the world





Changing the reflection won't change the world



fairness & diversity

state beliefs & assumptions

cannot fully automate responsibility!

Goals and trade-offs

Goals

diversity: pick k=4 candidates, including 2 of each gender, and at least one per race

utility: maximize the total score of selected candidates

score = 373

score = 372

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

Problem

fairness: picked the best White and male candidates (A, B) but did not pick the best Black (E, F), Asian (I, J), or female (C, D) candidates

Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**

@FalaahArifKhan

[Yang, Gkatzelis, Stoyanovich (2019)]

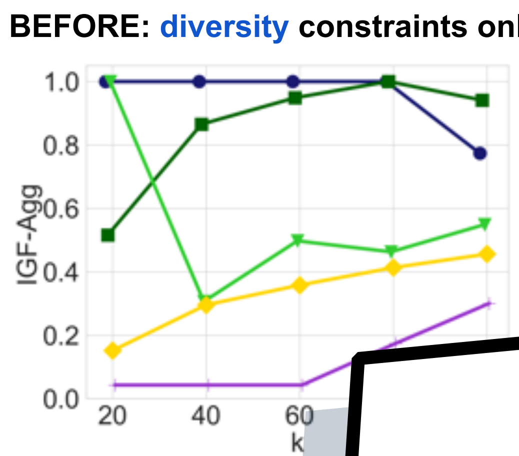
From beliefs to interventions

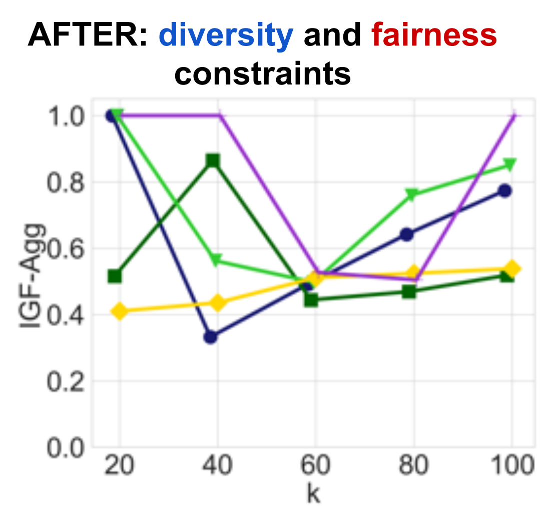
Fairness for female candidates **83 / 95 = 0.91**

C	D	G	H	K	L
95	95	90	86	83	83

highest-scoring skipped

lowest-scoring selected





Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**

[Yang, Gkatzelis, Stoyanovich (2019)]

Goals and trade-offs

Goals

diversity: pick k=3 candidates, including at least 1 of each gender

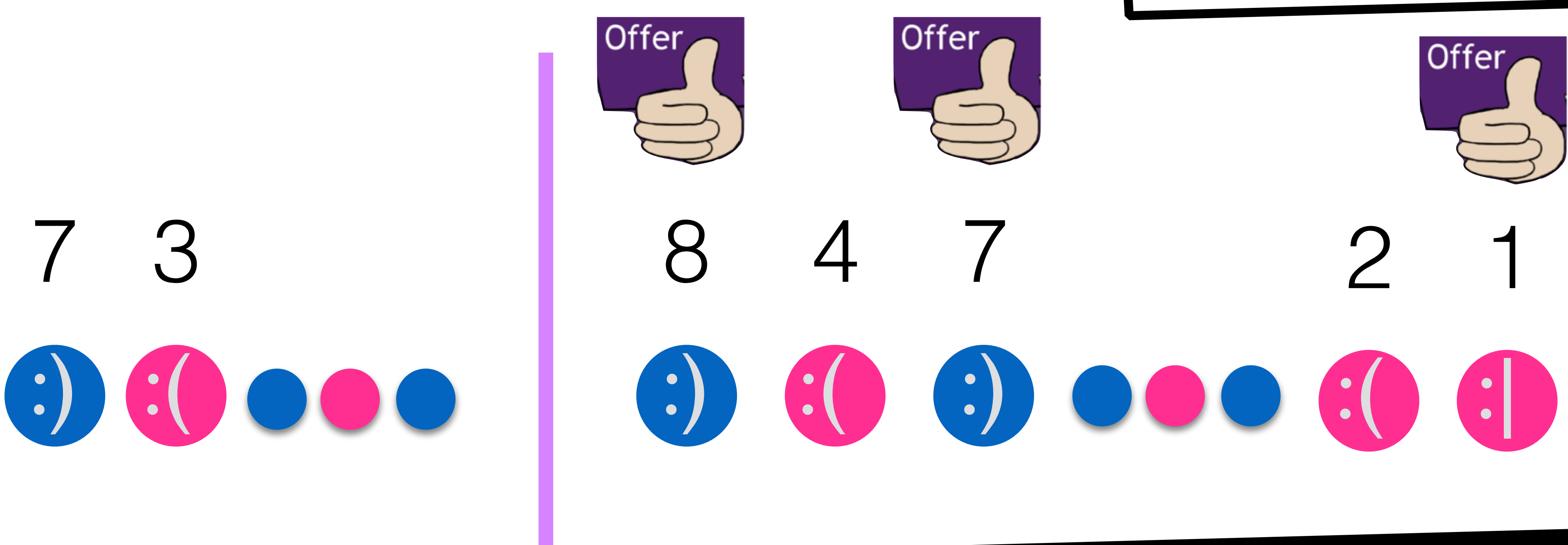
utility: maximize the total score of the selected candidates

the twist: utility revealed upon interview, must decide on the spot whether to hire a candidate

Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**



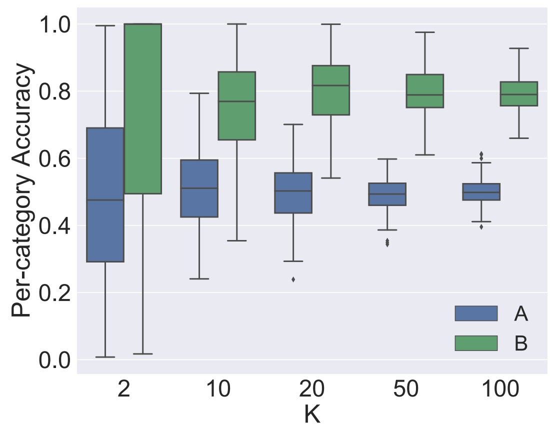
[Stoyanovich, Yang, Jagadish (2018)]

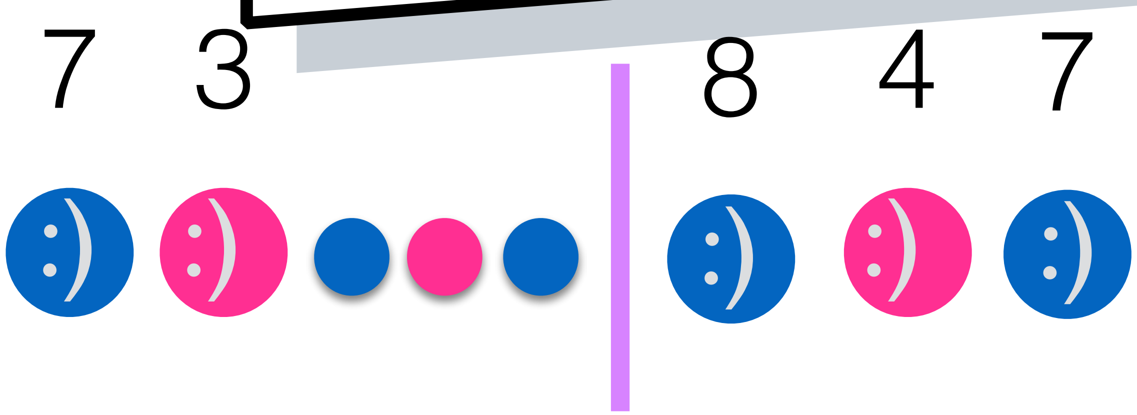
From beliefs to interventions

Idea: diverse k-choice secretary

learn what a good candidate looks separately for each category

Common warm-up period



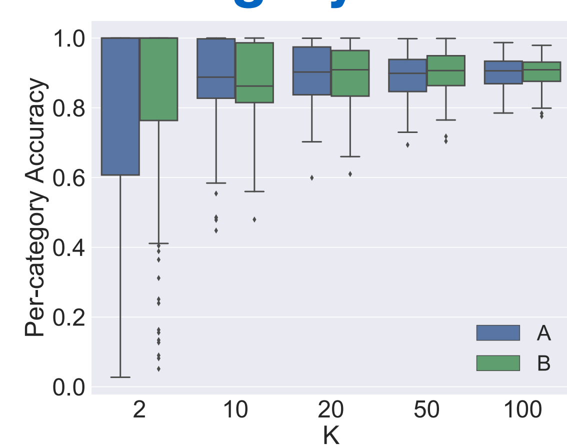


Beliefs

scores are more informative within a group than across groups - **effort is relative to circumstance**

it is important to **reward effort**

Per-category warm-up



[Stoyanovich, Yang, Jagadish (2018)]

Responsible Data Science course

“*Mirror Mirror*”. Data, Responsibly Comics, Volume 1 (2020)

#RDSComic

Thank you!

dataresponsibly.github.io
/courses
/comics

data RESPONSIBLY

NYU

Responsible Data Management

Julia Stoyanovich

Computer Science and Engineering &
Center for Data Science
New York University, NY USA

data RESPONSIBLY

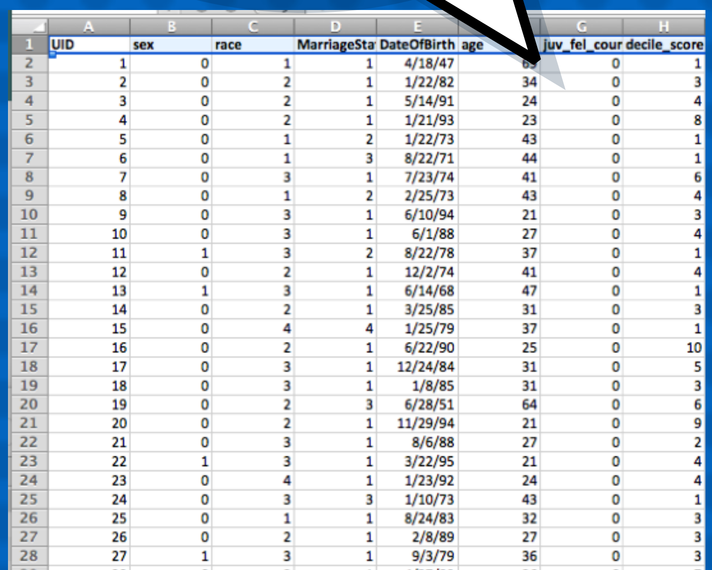
NYU

Frog’s eye view

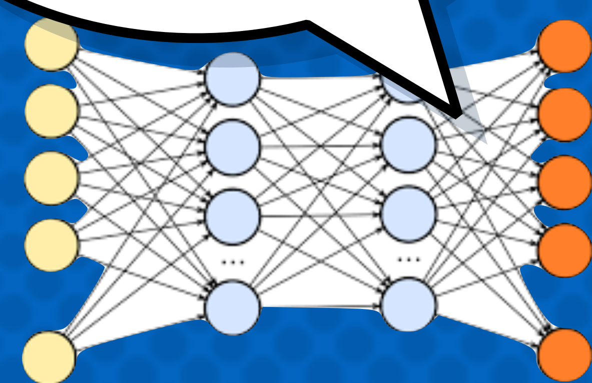
where did the data come from?

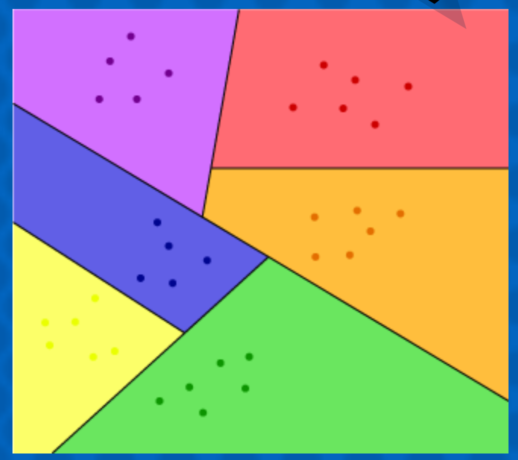
what happens inside the box?

how are results used?



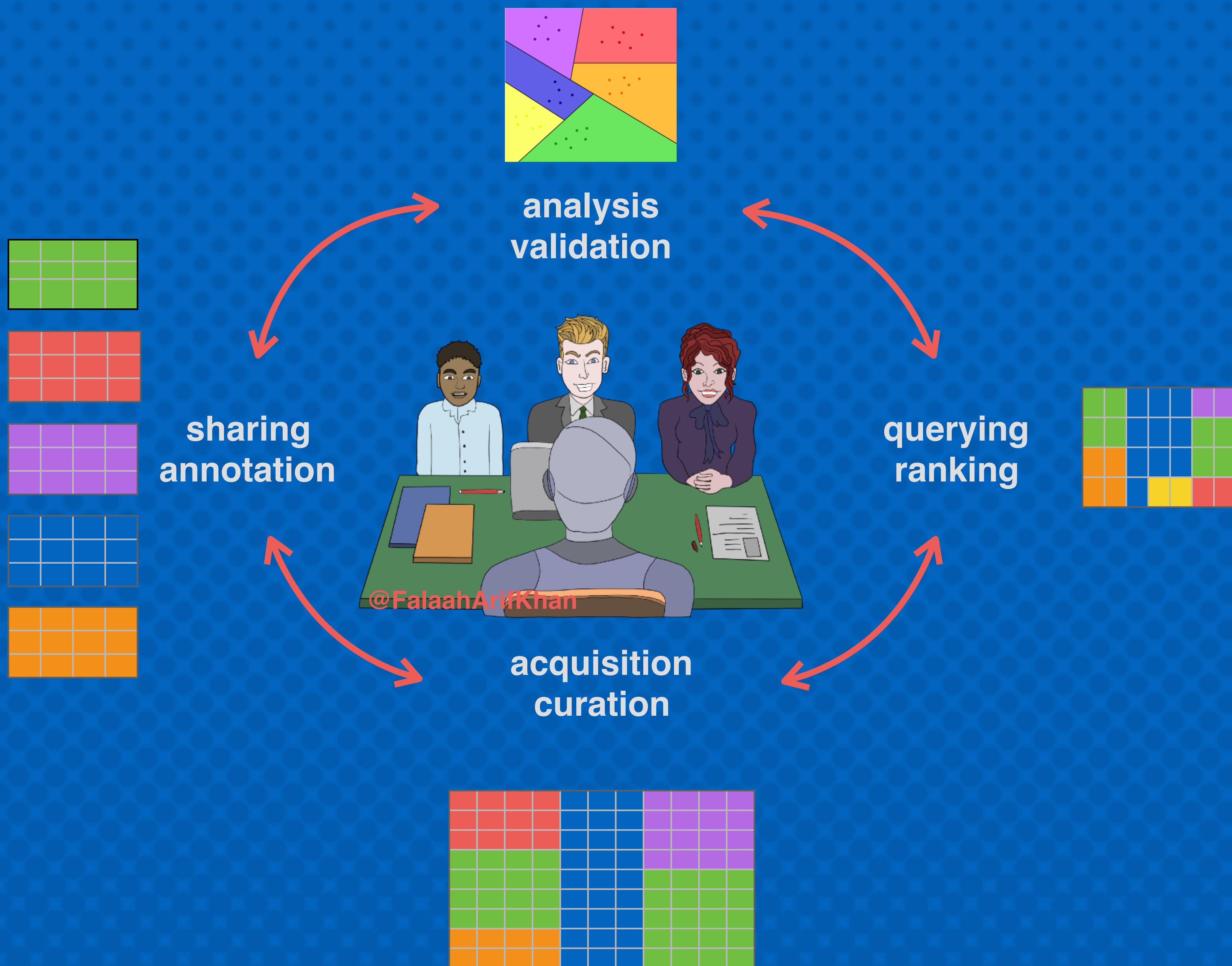
	A	B	C	D	E		G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	[illegible]	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3





a lifecycle view of ADS

Data lifecycle of an ADS



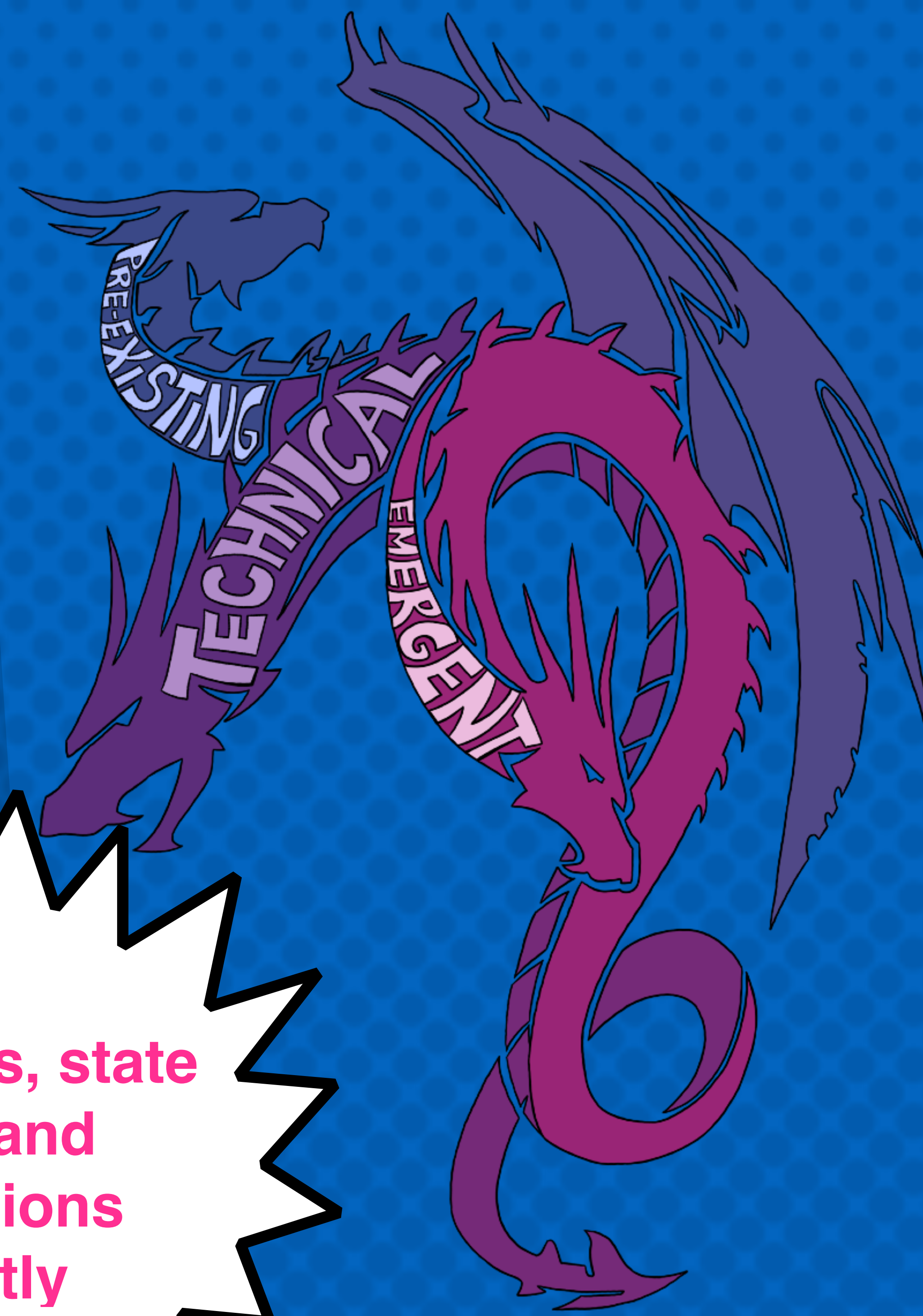
Bias in ADS, revisited

Pre-existing: exists independently of algorithm, has origins in society

Technical: introduced or exacerbated by the technical properties of an ADS

Emergent: arises due to context of use

to fight bias, state beliefs and assumptions explicitly



@FalaahArifKhan

taming technical bias

**we break it —
we fix it!**

Model development lifecycle

Goal

design a model to predict an appropriate level of compensation for job applicants

Problem

accuracy is lower for middle-aged women - **a fairness concern!**

now what?

demographics

employment

split

preprocess

interpolate missing

select model

tune & validate

[Schelter, He, Khilnani, Stoyanovich (2020)]

Models and assumptions

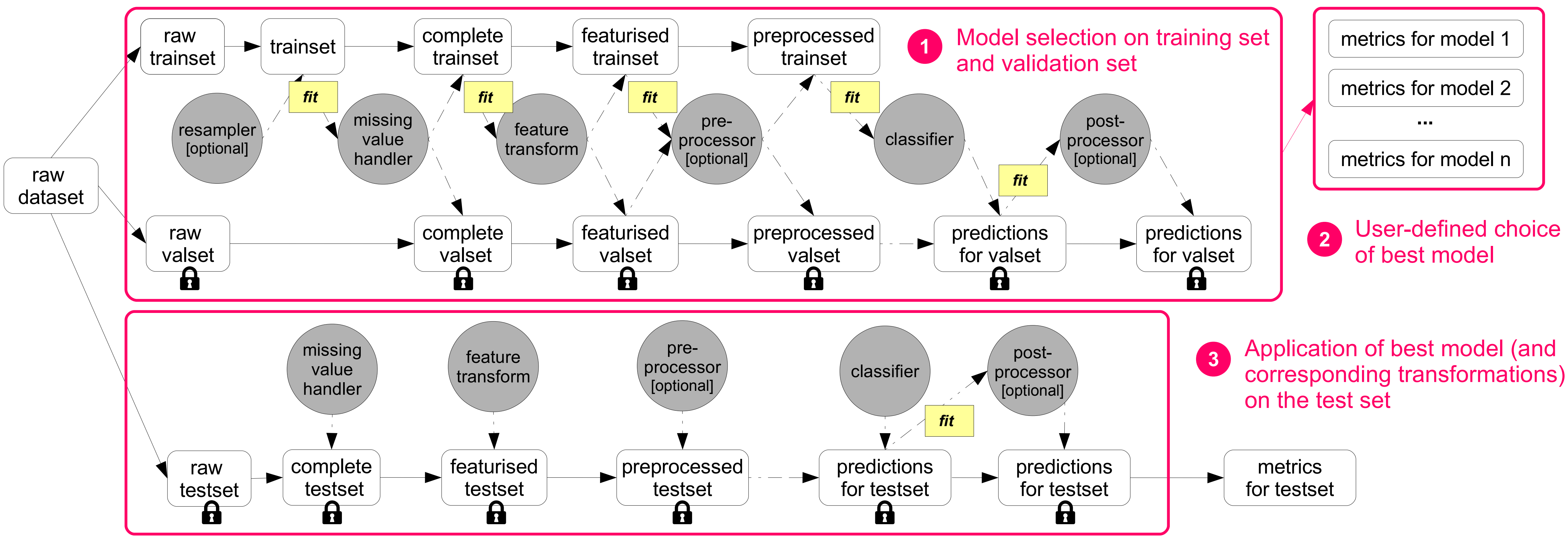
imputing age

imputing gender

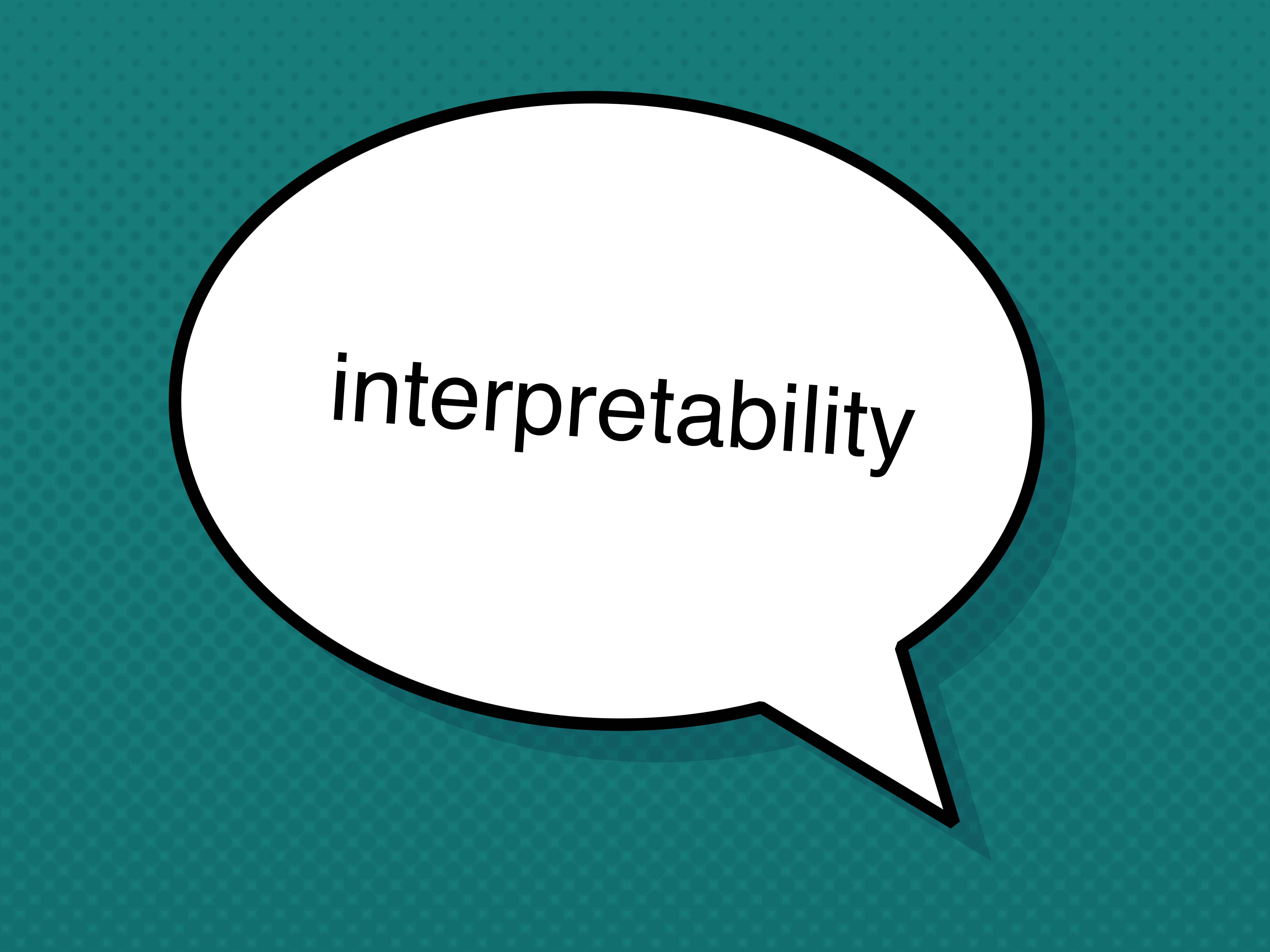
address of a homeless person?

non-binary gender

FairPrep: a holistic view of the pipeline



[Schelter, He, Khilnani, Stoyanovich (2020)]


interpretability

Ranking Facts

Recipe

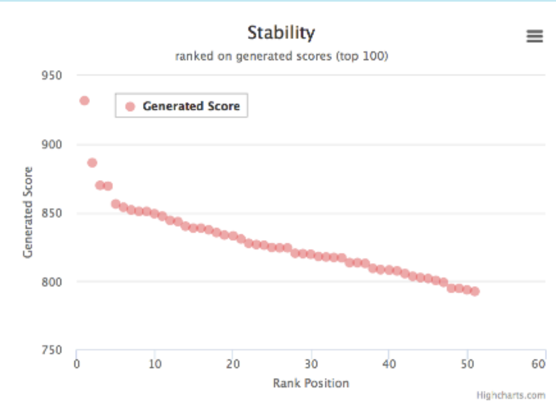
Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
Faculty	122	52.5	45
GRE	800.0	796.3	771.9

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
Faculty	122	32.0	14
GRE	800.0	790.0	757.8

Stability



Slope at top-10: -6.91. Slope overall: -1.61.

Unstable when absolute value of slope of fit line in scatter plot <= 0.25 (slope threshold). Otherwise it is stable.

Ranking Facts

Recipe

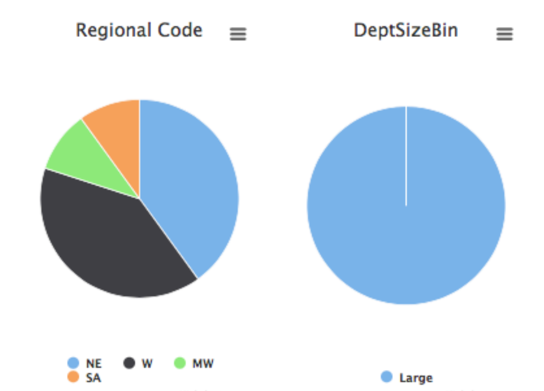
Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

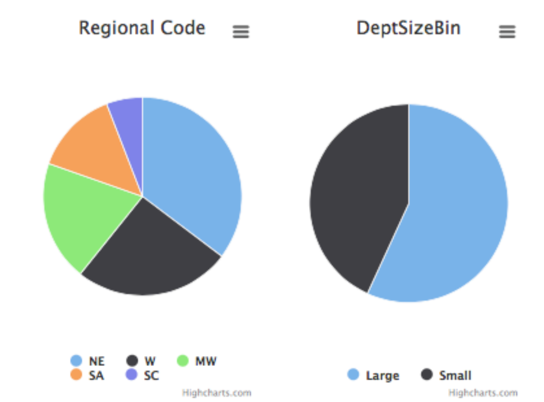
Attribute	Correlation
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Correlation strength is based on its absolute value. Correlation over 0.75 is high, between 0.25 and 0.75 is medium, under 0.25 is low.

Diversity at top-10



Diversity overall



Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

Unfair when p-value of corresponding statistical test <= 0.05.

Ingredients

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.99	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

Top K = 26 in FA*IR and Proportion oracles. Setting of top K: In FA*IR and Proportion oracle, if N > 200, set top K =100. Otherwise set top K = 50%N. Pairwise oracle takes whole ranking as input. FA*IR is computed as using code in FA*IR codes. Proportion is implemented as statistical test 4.1.3 in Proportion paper.

[Yang, Stoyanovich, Asudeh, Howe, Jagadish, Miklau (2020)]

Stability in ranking

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

1. Chevrolet Corvette 205
2. Lotus Evora 195
3. Porsche Cayman 195

1. Lotus Evora 205
2. Porsche Cayman 198
3. Chevrolet Corvette 192

1. Porsche Cayman 193
2. Chevrolet Corvette 186
3. Lotus Evora 182

Rankings are not benign. They enshrine very particular **ideologies**, and, at a time when American higher education is facing a crisis of accessibility and affordability, we have adopted **a de-facto standard** of college quality that is uninterested in both of those factors. And why? Because a group of magazine analysts in an office building in Washington, D.C., decided twenty years ago to **value selectivity over efficacy**, to **use proxies** that scarcely relate to what they're meant to be proxies for, and to **pretend that they can compare** a large, diverse, low-cost land-grant university in rural Pennsylvania with a small, expensive, private Jewish university on two campuses in Manhattan.

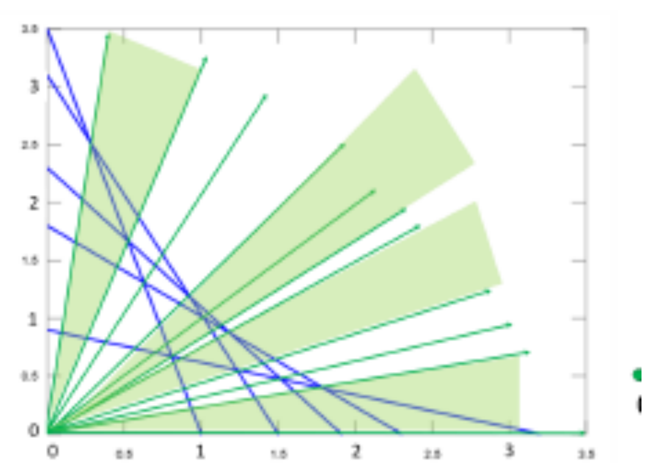
Designing stable rankers

Goal find a scoring function to rank applicants

utility: with similar weights as what the human decision-maker has in mind

stability: so that the resulting ranking doesn't reshuffle when weights are changed slightly

$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4

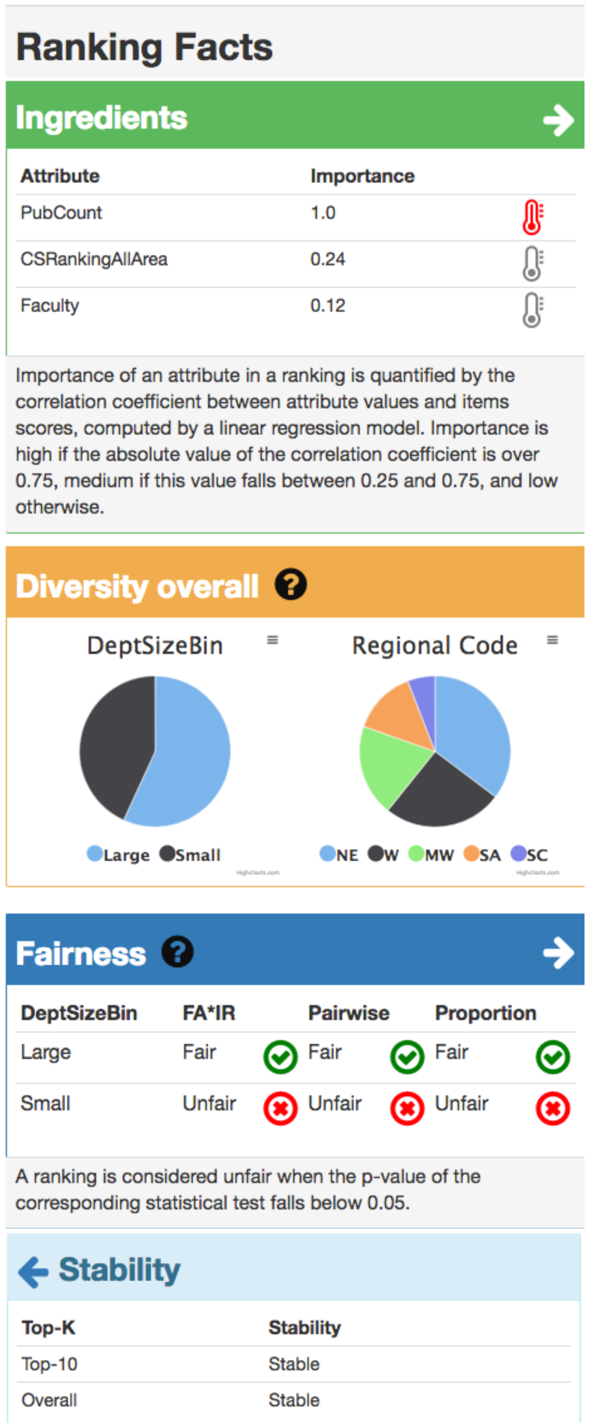


Belief

stable rankings are more trustworthy

[Asudeh, Jagadish, Miklau, Stoyanovich (2018)]

Back to nutritional labels



comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

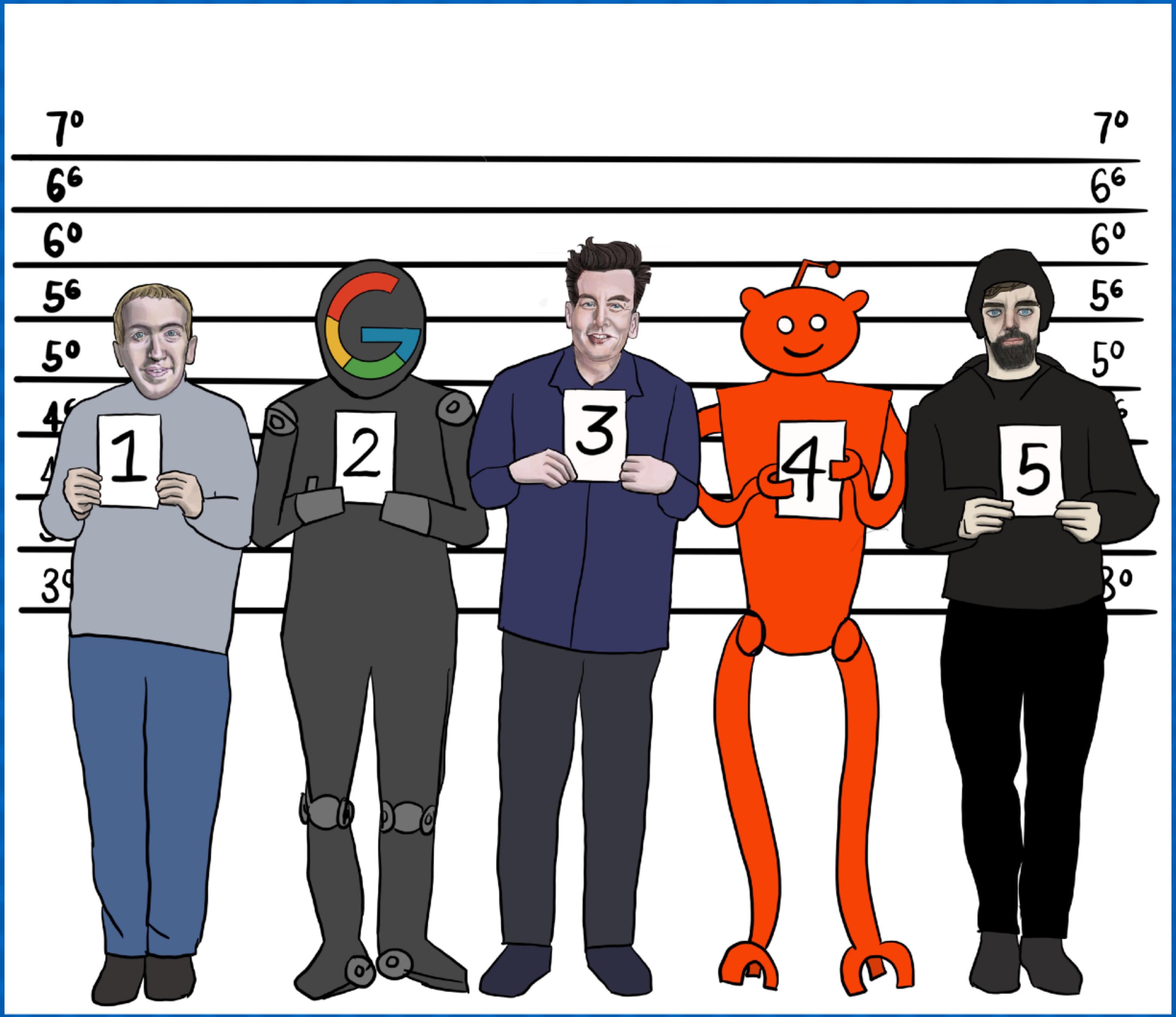
computable: incrementally constructed

[Stoyanovich, Howe (2019)]

take-aways

Framing technical solutions

We all are responsible



@FalaahArifKhan

Tech rooted in people

Responsible Data Science course

"*Mirror Mirror*". Data, Responsibly Comics, Volume 1 (2020)

#RDSComic

Thank you!

dataresponsibly.github.io
/courses
/comics