Responsible Data Management

Julia Stoyanovich

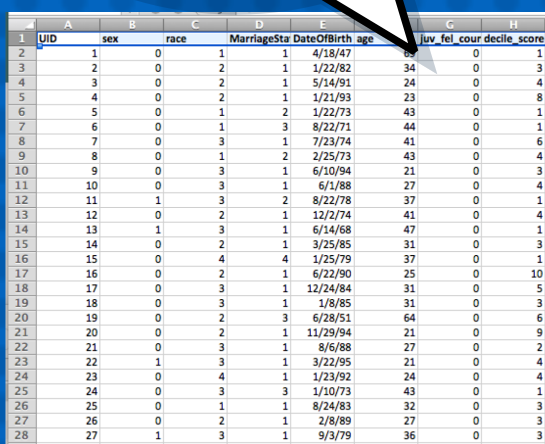
Computer Science and Engineering &
Center for Data Science
New York University, NY USA

data RESPONSIBLY

NYU

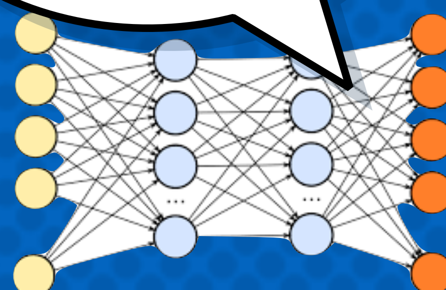
Frog’s eye view

where did the data come from?



	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

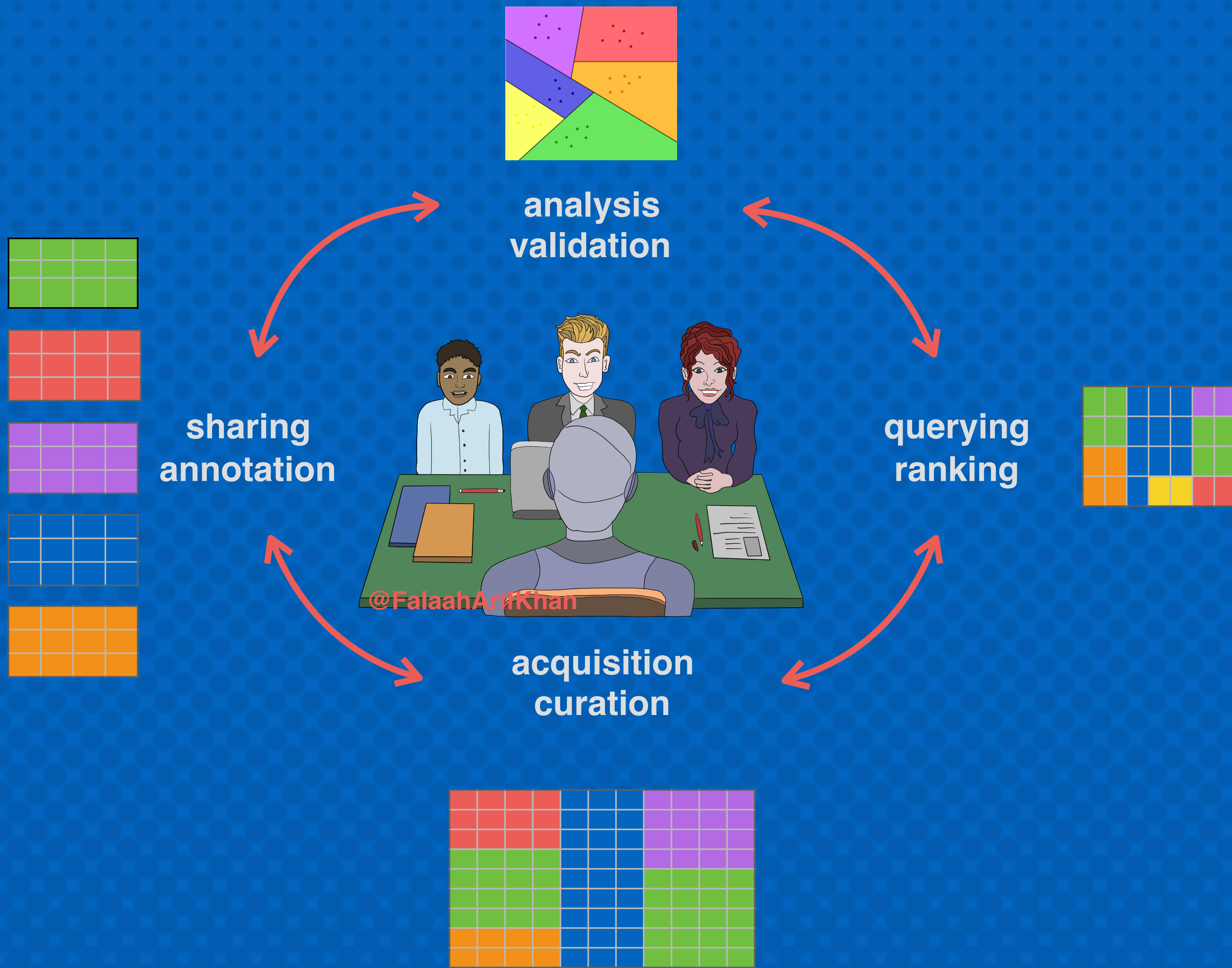
what happens inside the box?



how are results used?

a lifecycle view of ADS

Data lifecycle of an ADS



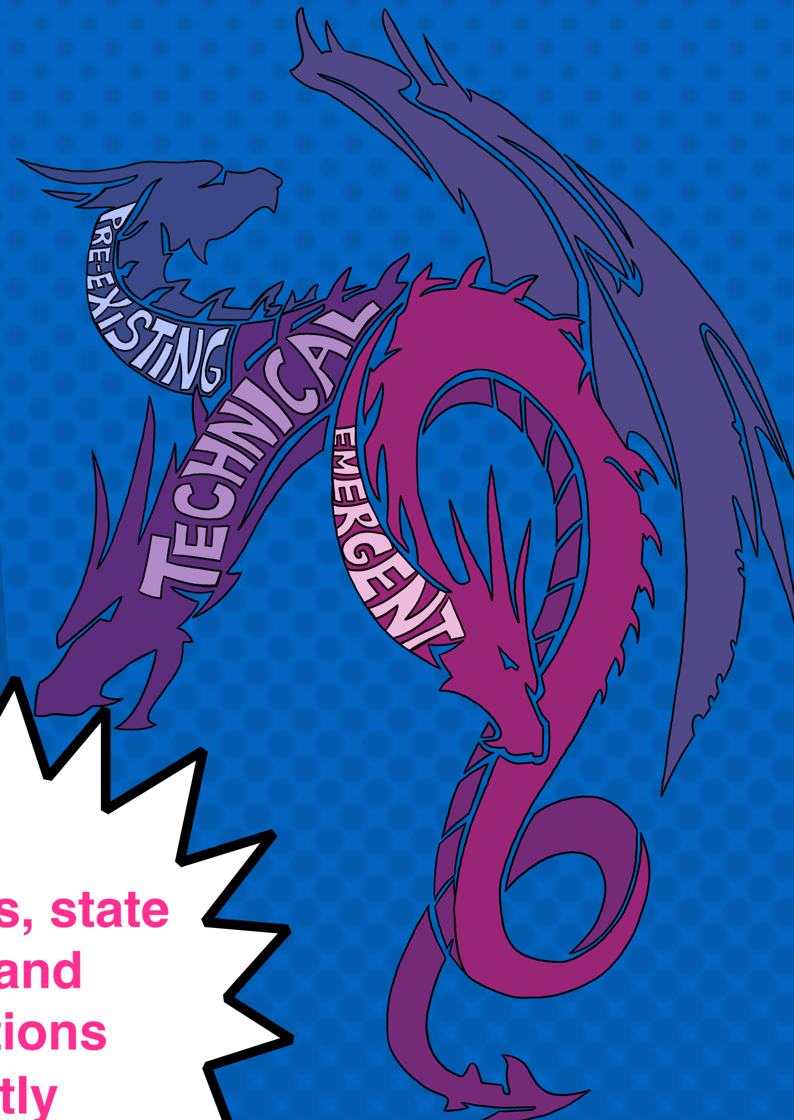
Bias in ADS, revisited

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

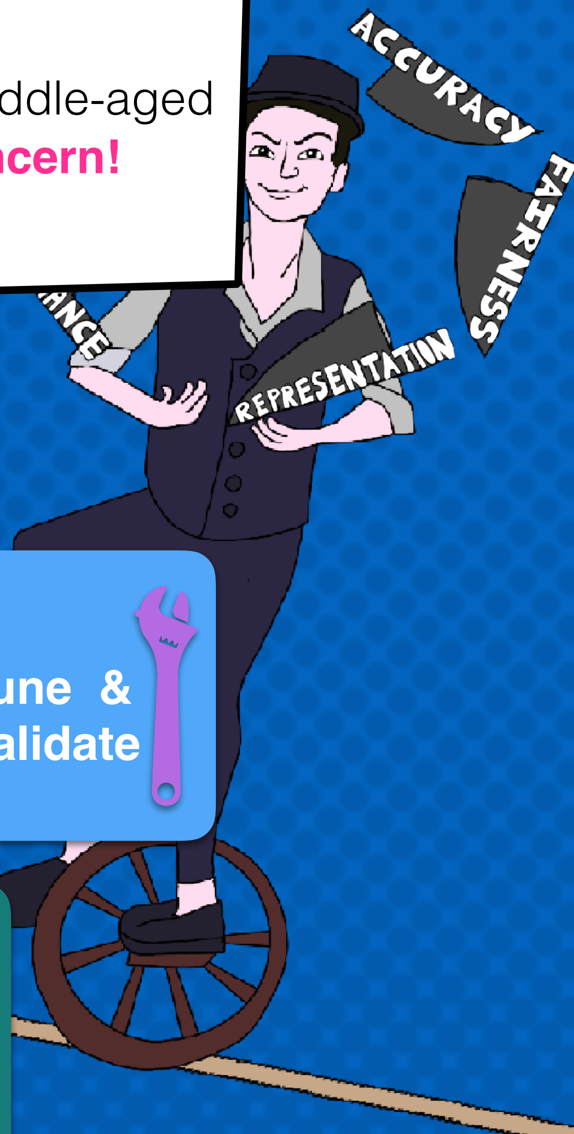
split

preprocess

interpolate missing

select model

tune & validate



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

Models and assumptions

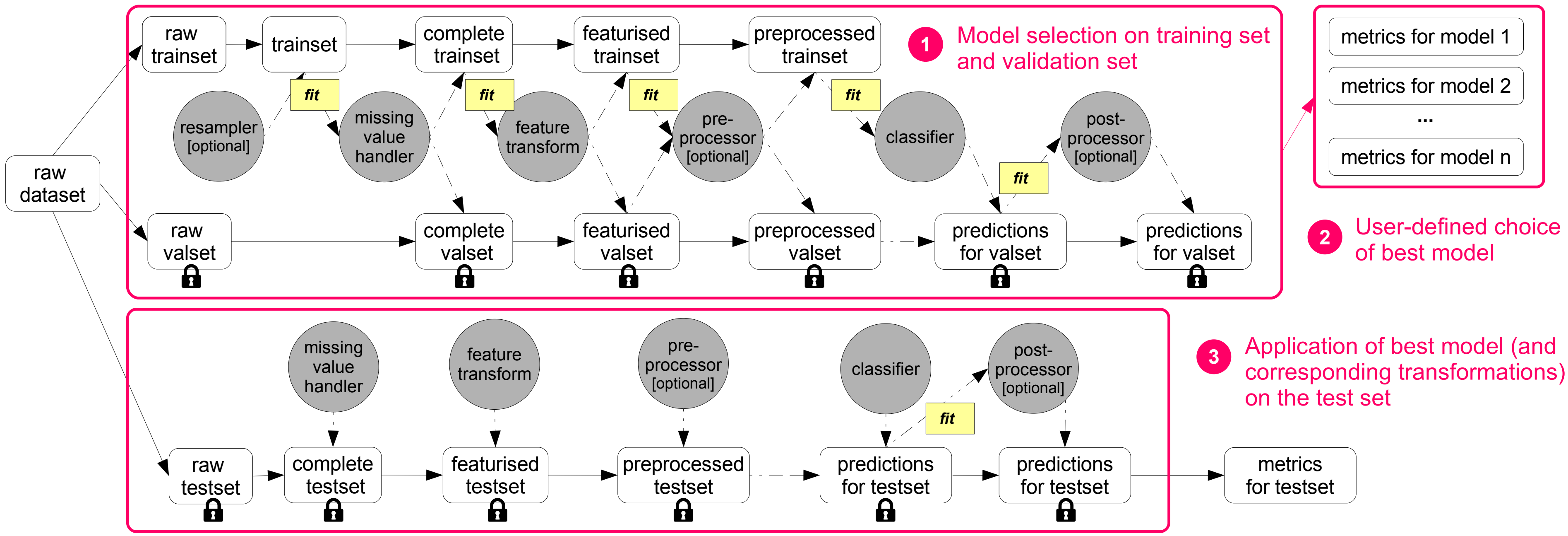
imputing age

imputing gender

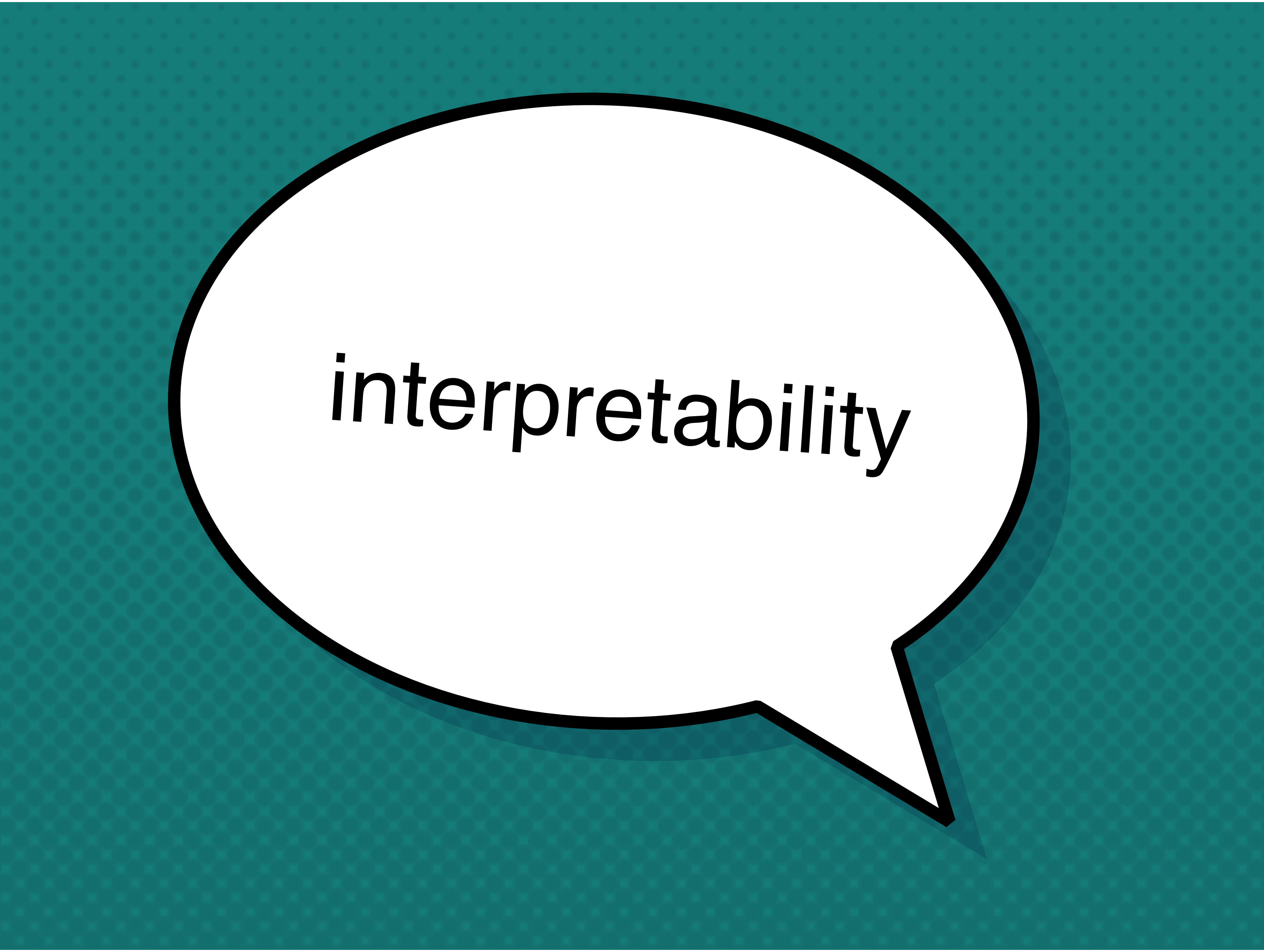
address of a homeless person?

non-binary gender

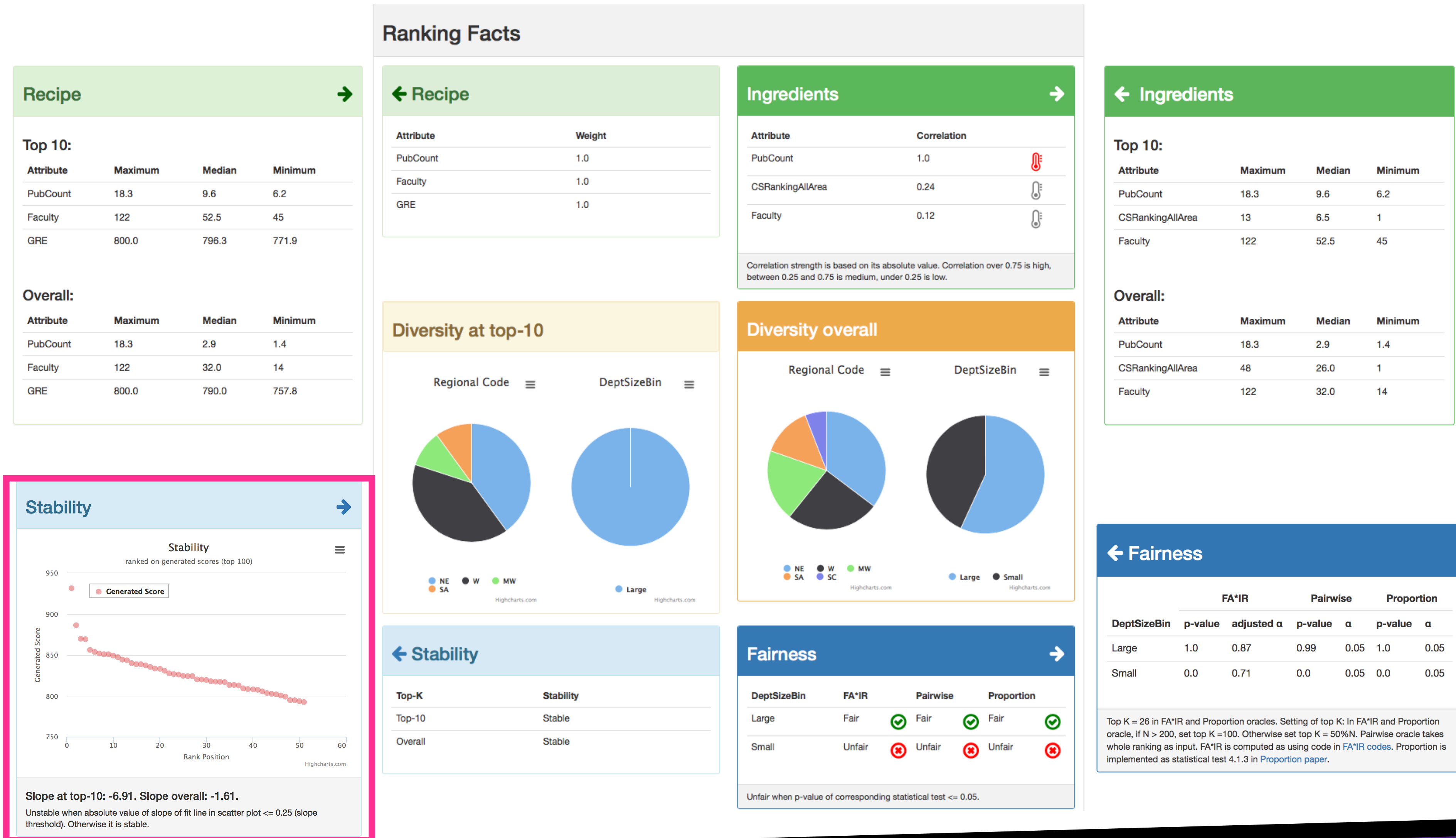
FairPrep: a holistic view of the pipeline



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


interpretability

Ranking Facts



[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

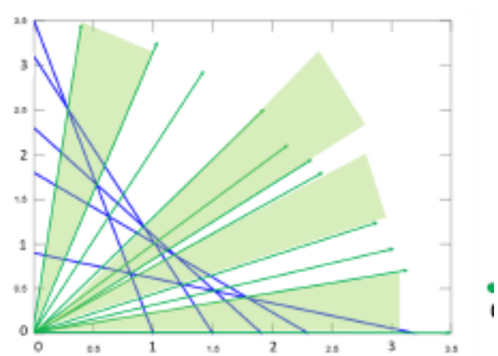
utility: with similar weights as what the human decision-maker has in mind

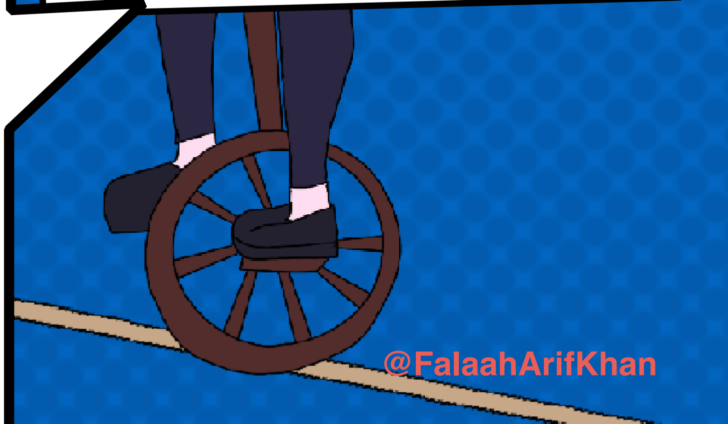
stability: so that the resulting ranking doesn't reshuffle when weights are changed slightly

Belief

stable rankings are more trustworthy

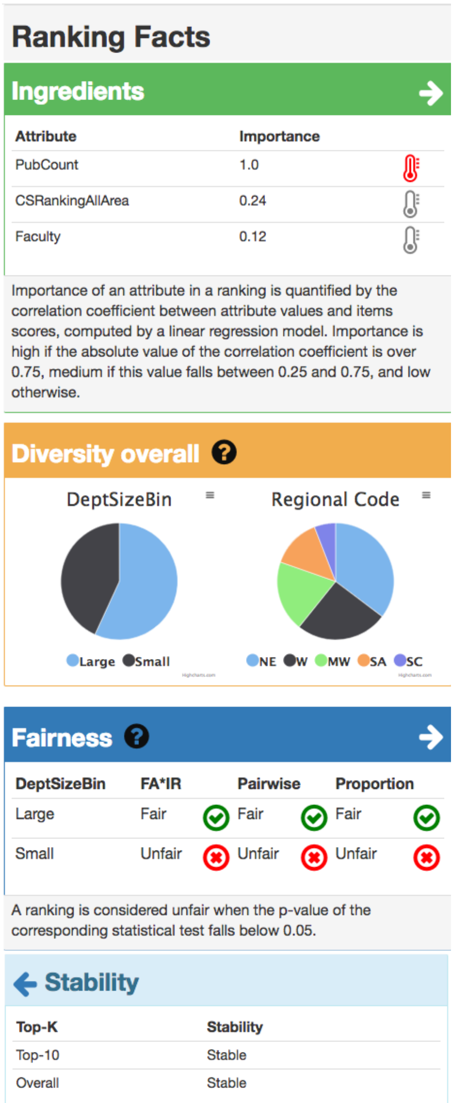
$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4





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

Back to nutritional labels



comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

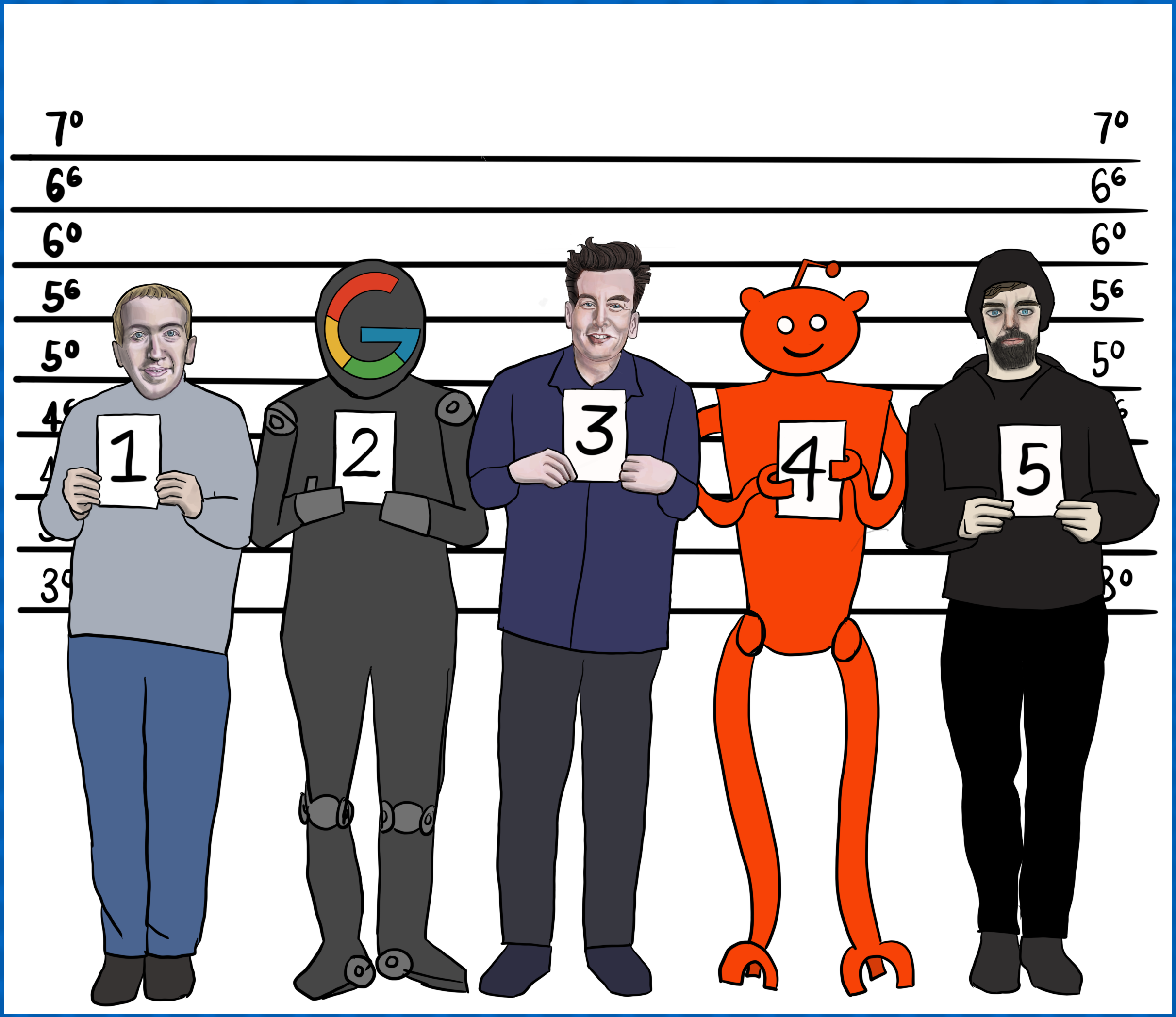
computable: incrementally constructed

[Stoyanovich, Howe (2019)]


take-aways

Framing technical solutions

We all are responsible



@FalaahArifKhan

Tech rooted in people

Responsible Data Science course

NSF

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

#RDSComic

Thank you!

datarensponsibly.github.io
/courses
/comics

data RESPONSIBLY

NYU