MACHINE LEARNING FOR NEWSROOMS

Jonathan Soma
Columbia University
js4571@columbia.edu
@dangerscarf

https://bit.ly/djc-2022-ml

MACHINE LEARNING: WHAT AND WHY?

https://bit.ly/djc-2022-ml

WHAT IS M.L.?

Machine learning

Artificial intelligence

Fancy statistics

Not-so-fancy statistics

Almost everything!

TWO USE CASES

ML for Business

Conversion rates

Article summaries

Content creation/article generation

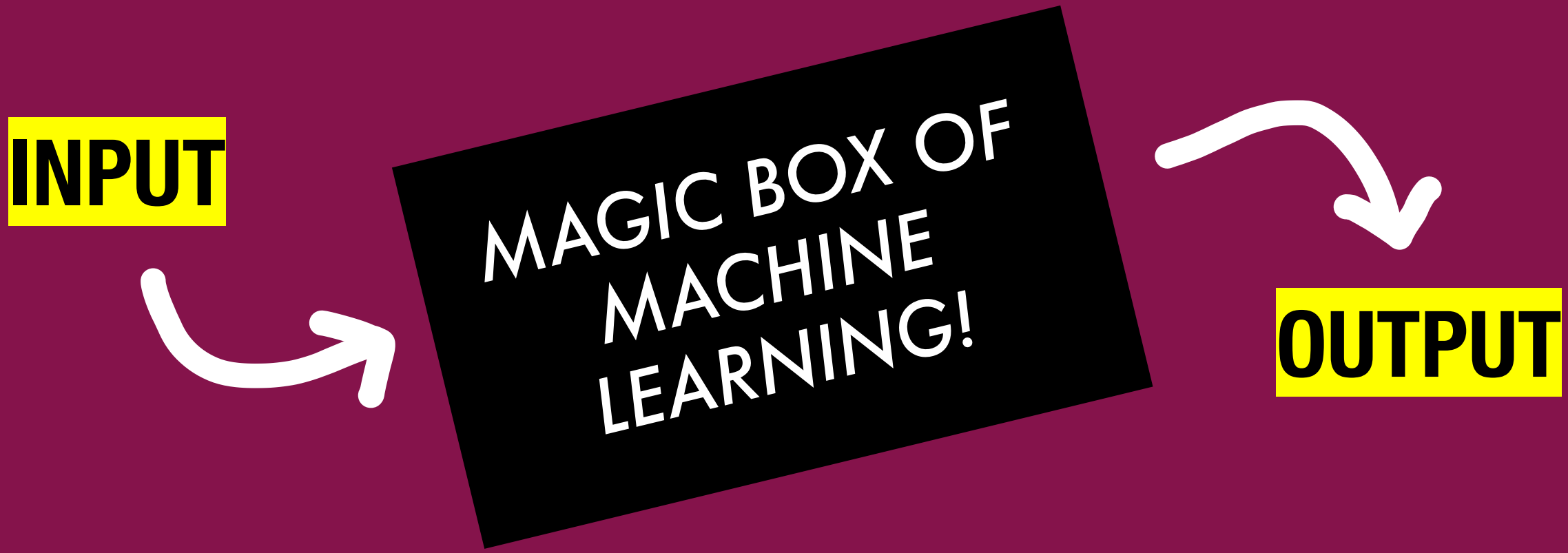
ML for Investigations

Uncovering unusual data points

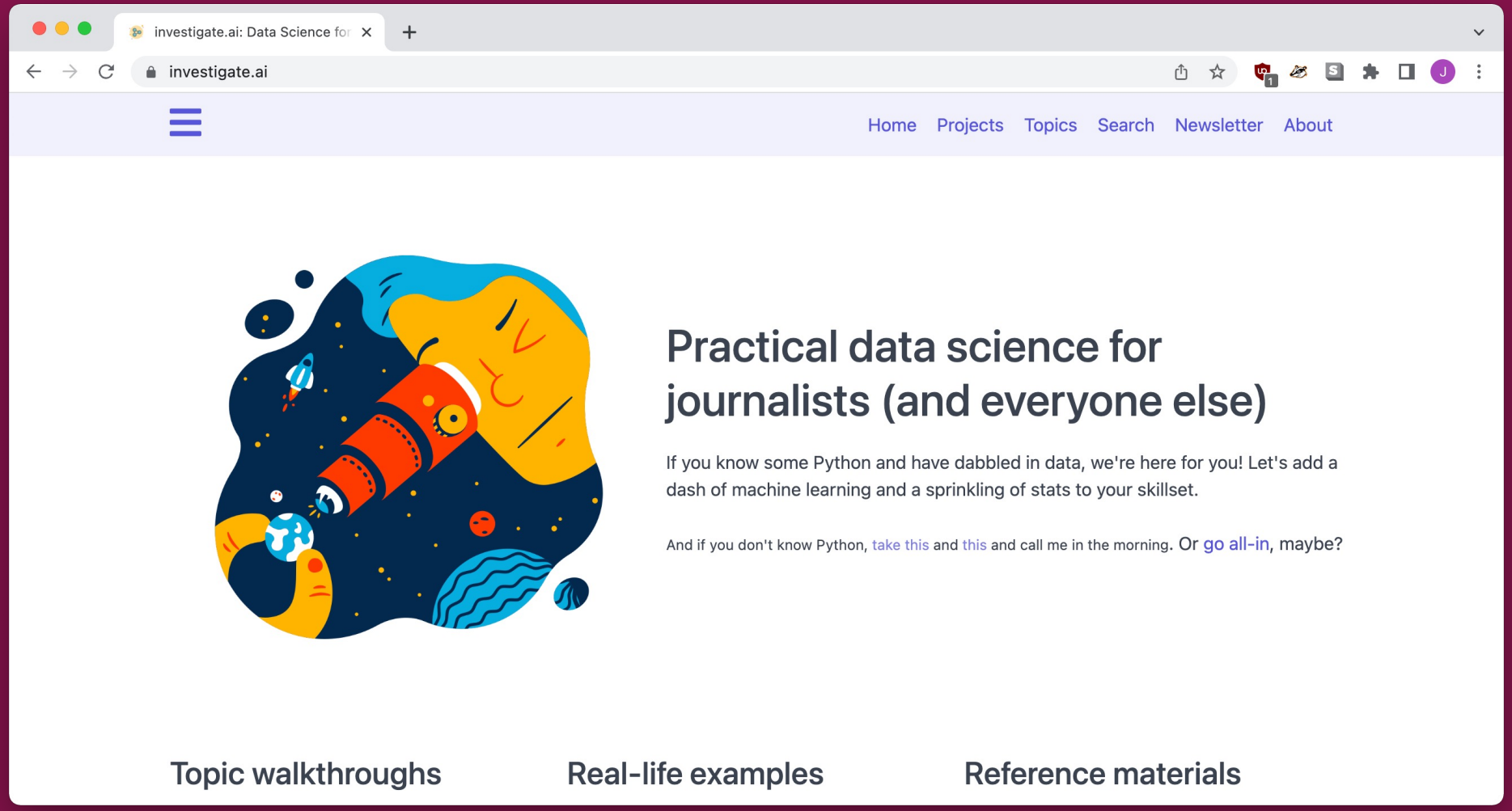
Finding bias

Doing large amounts of work, relatively quickly and easier

HOW M.L. WORKS

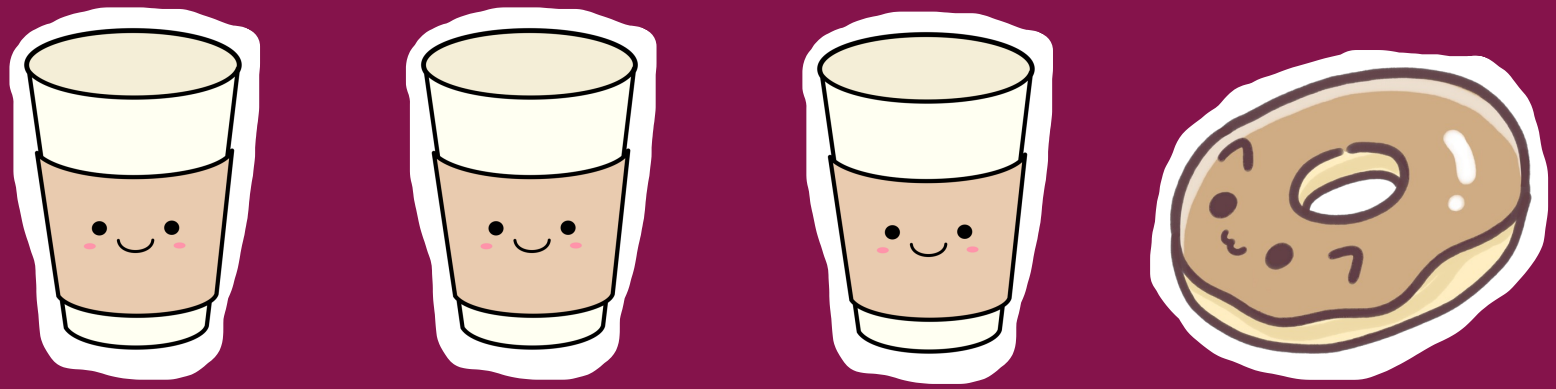


INVESTIGATE.AI



BE A JOURNALIST, NOT A STATISTICIAN OR DATA SCIENTIST

+ tip $7

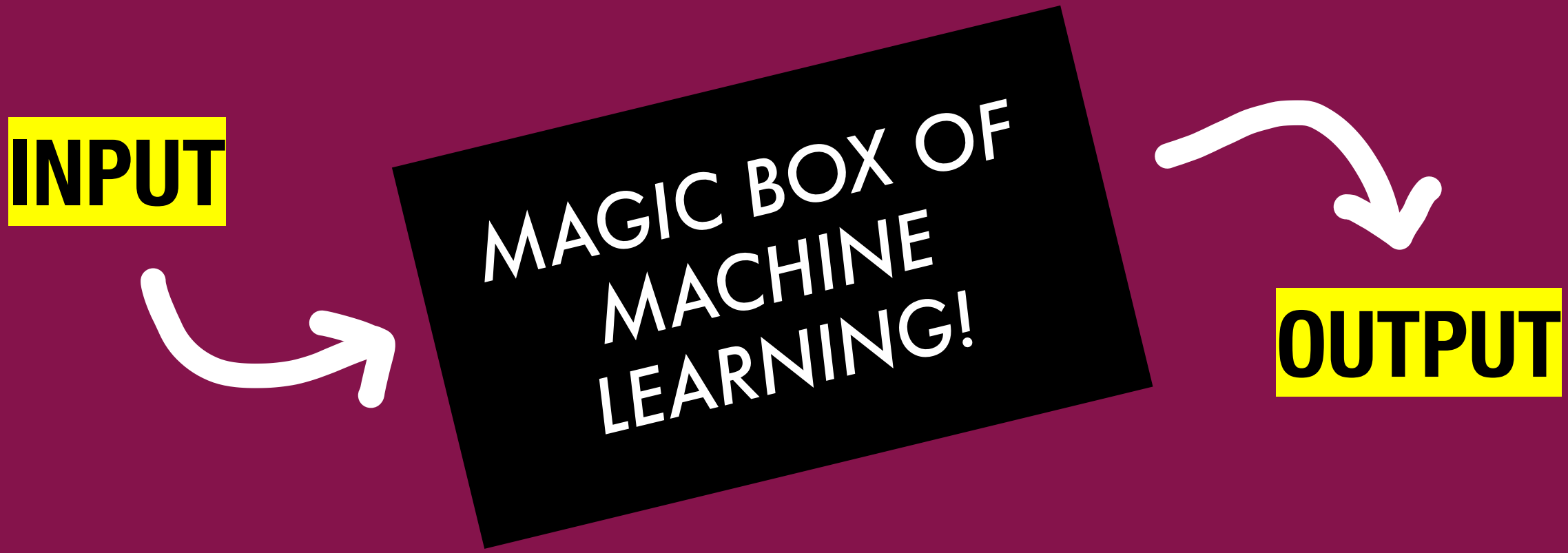


+ tip $12

+ tip $15

“TRAINING” A “MODEL”

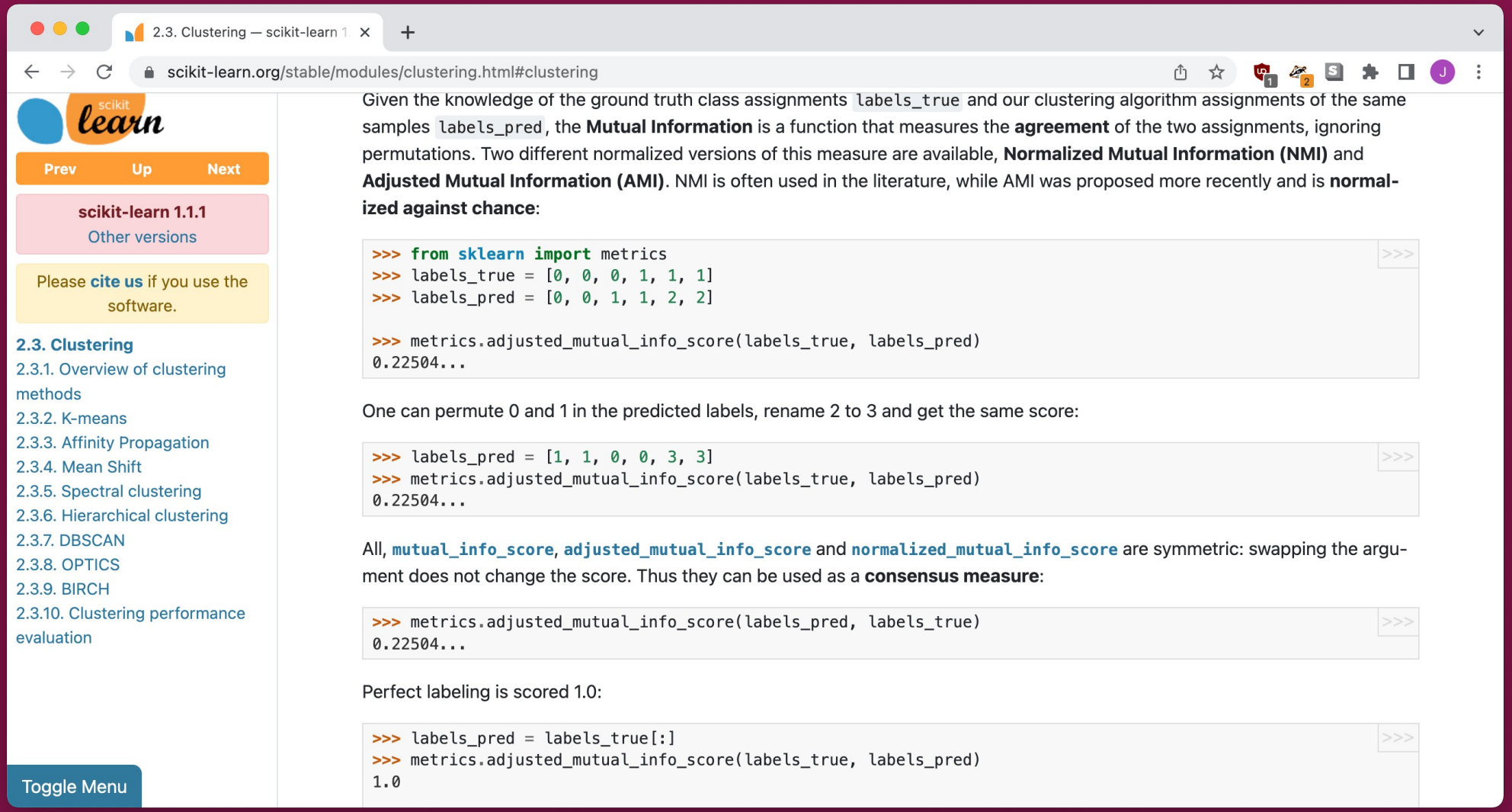
WHAT IS A "MODEL"?



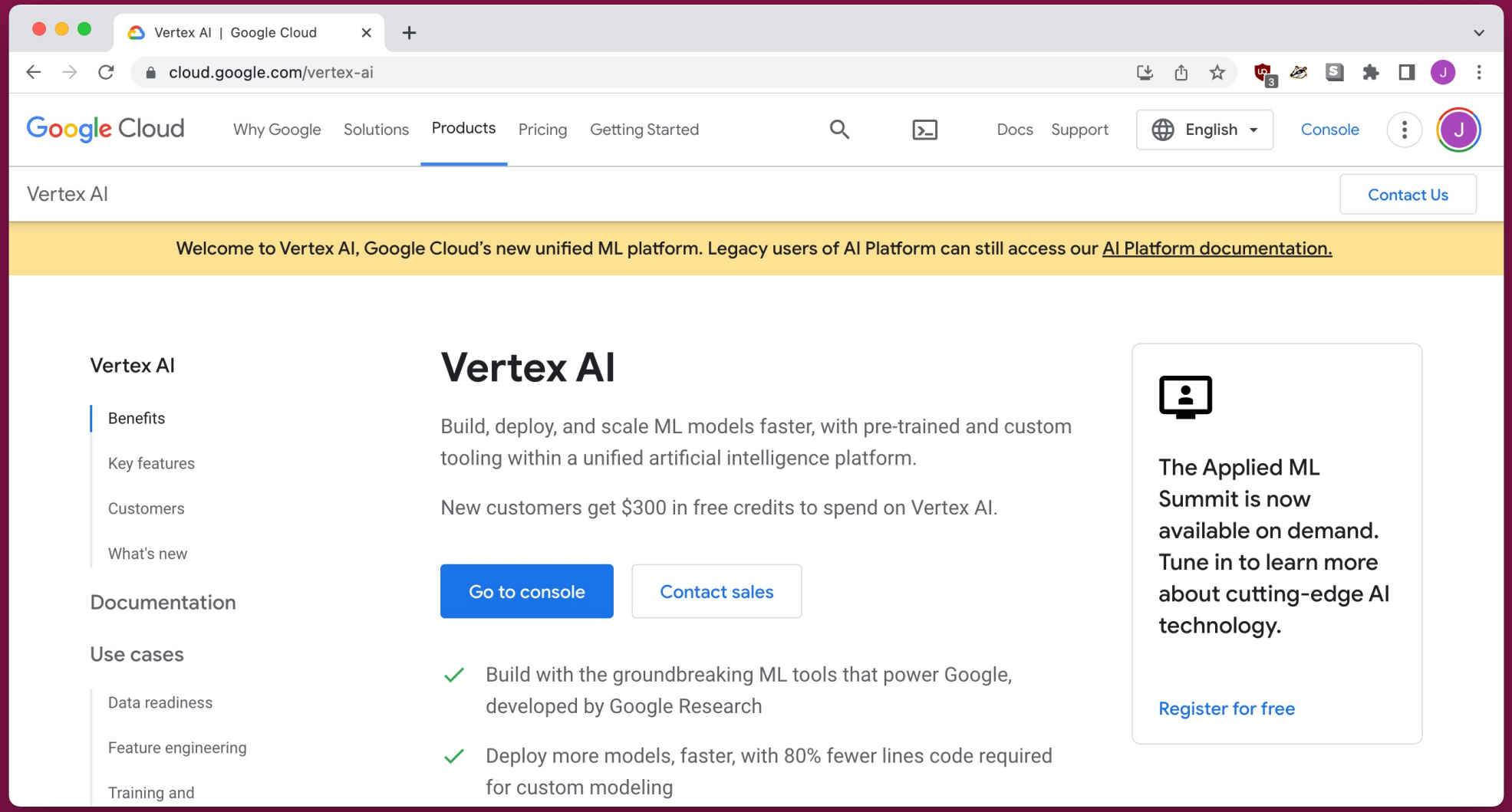
IT'S NOT THE TOOL, IT'S HOW YOU USE IT.

DALL-E, MIDJOURNEY

SKLEARN, TENSORFLOW

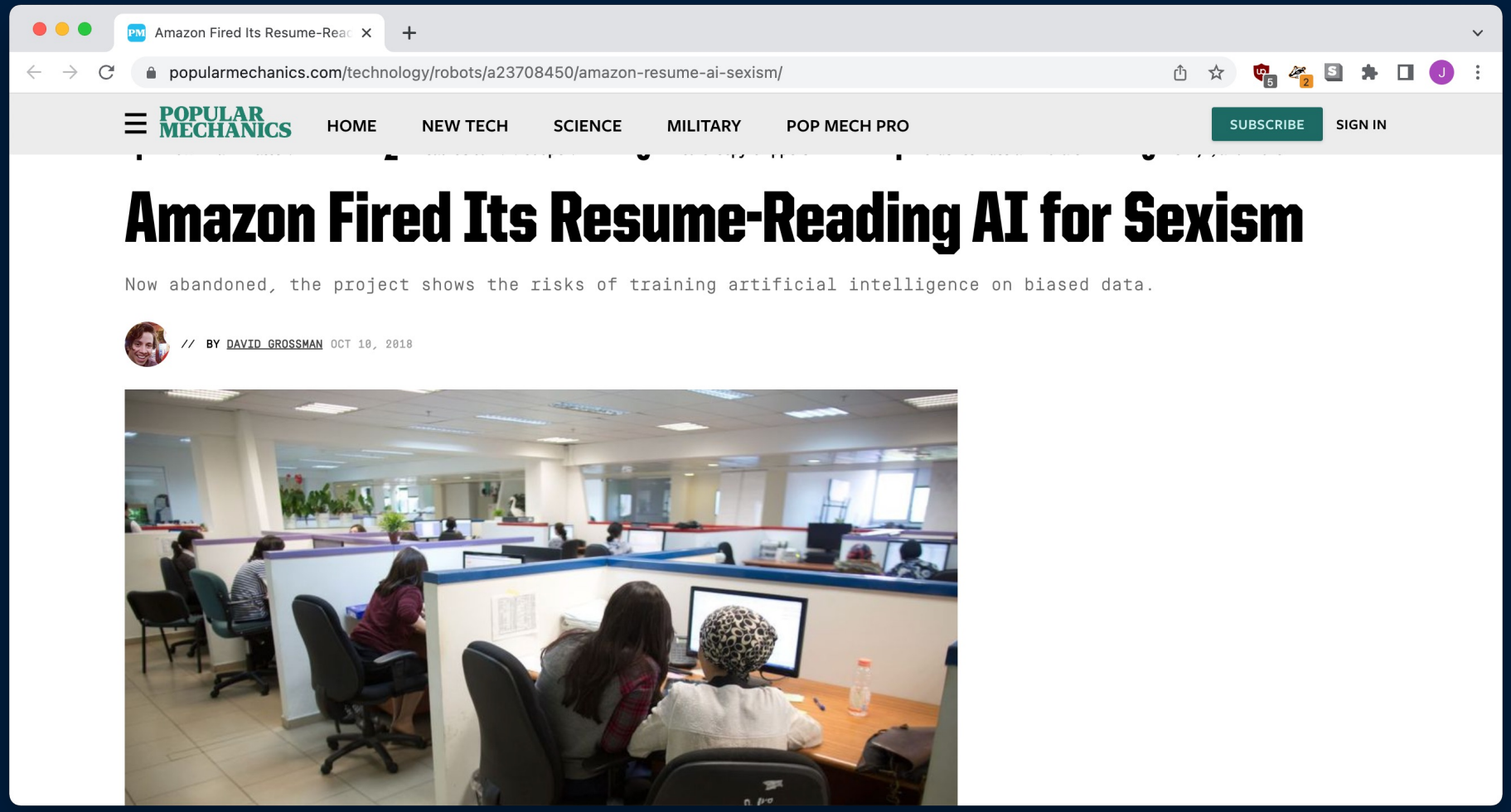


REFINERY, VERTEX AI, ETC...

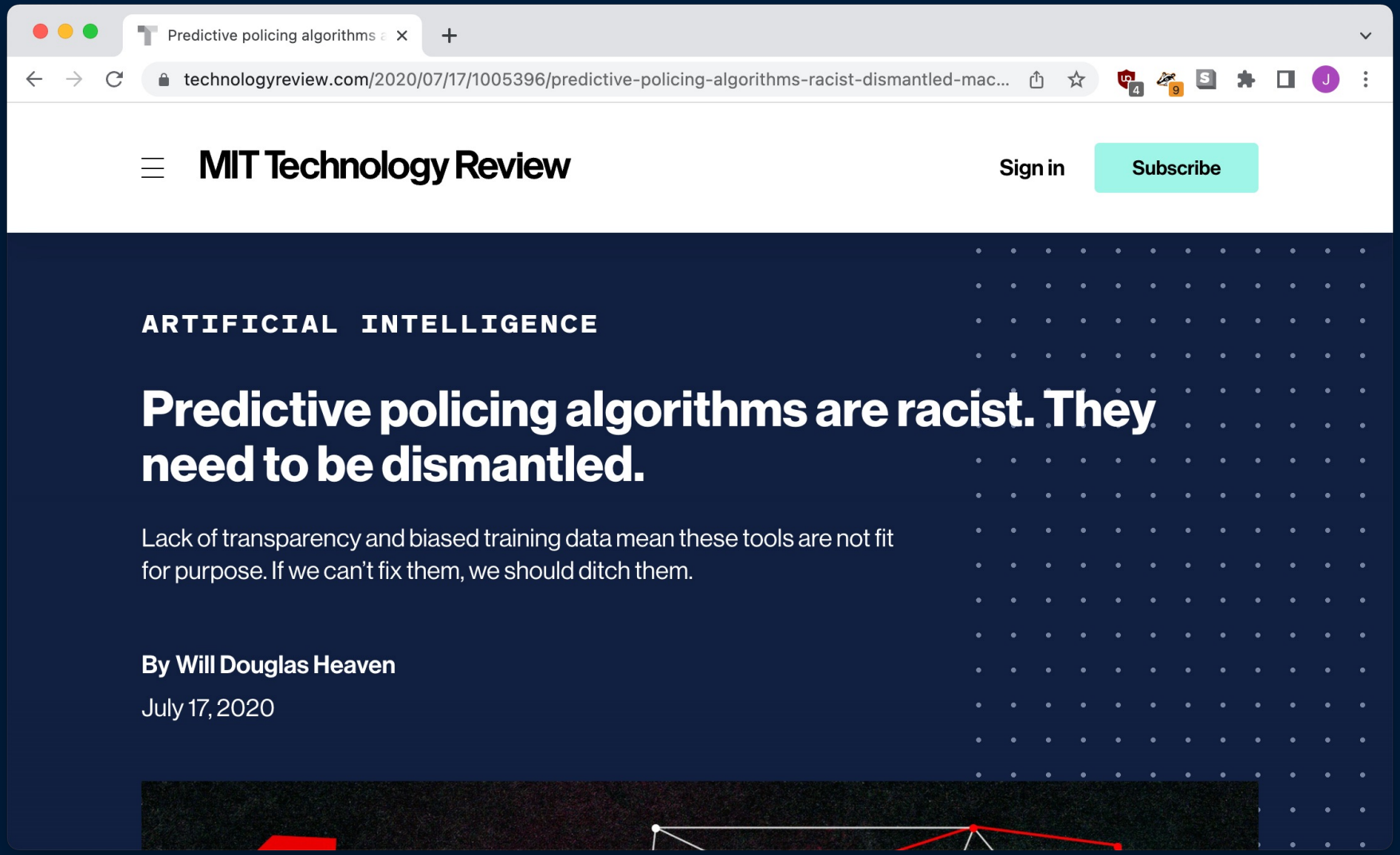


CAUTION: THE MACHINE ONLY REPRODUCES WHAT IT'S SEEN.

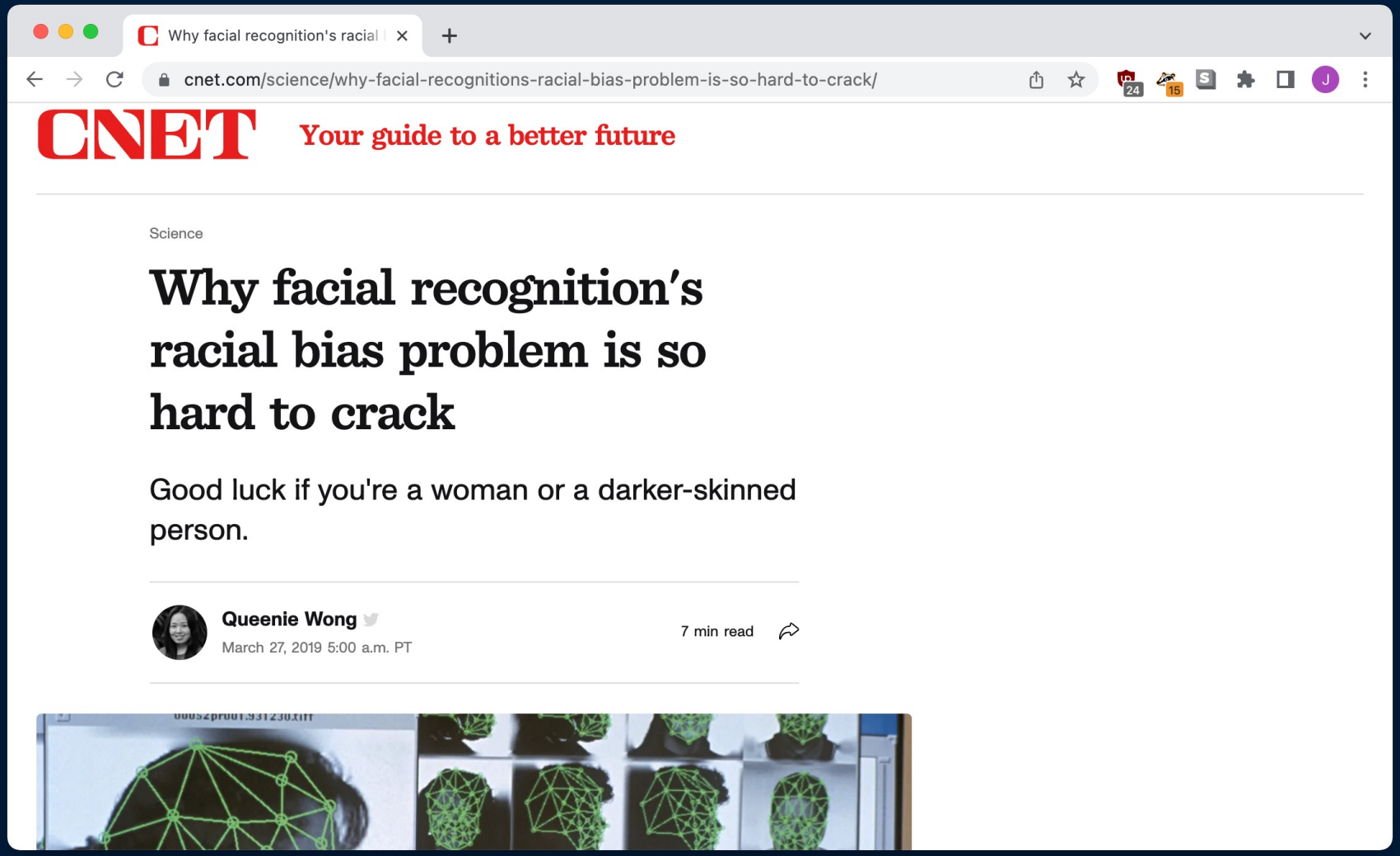
RESUME SCREENER



“BIAS LAUNDERING”



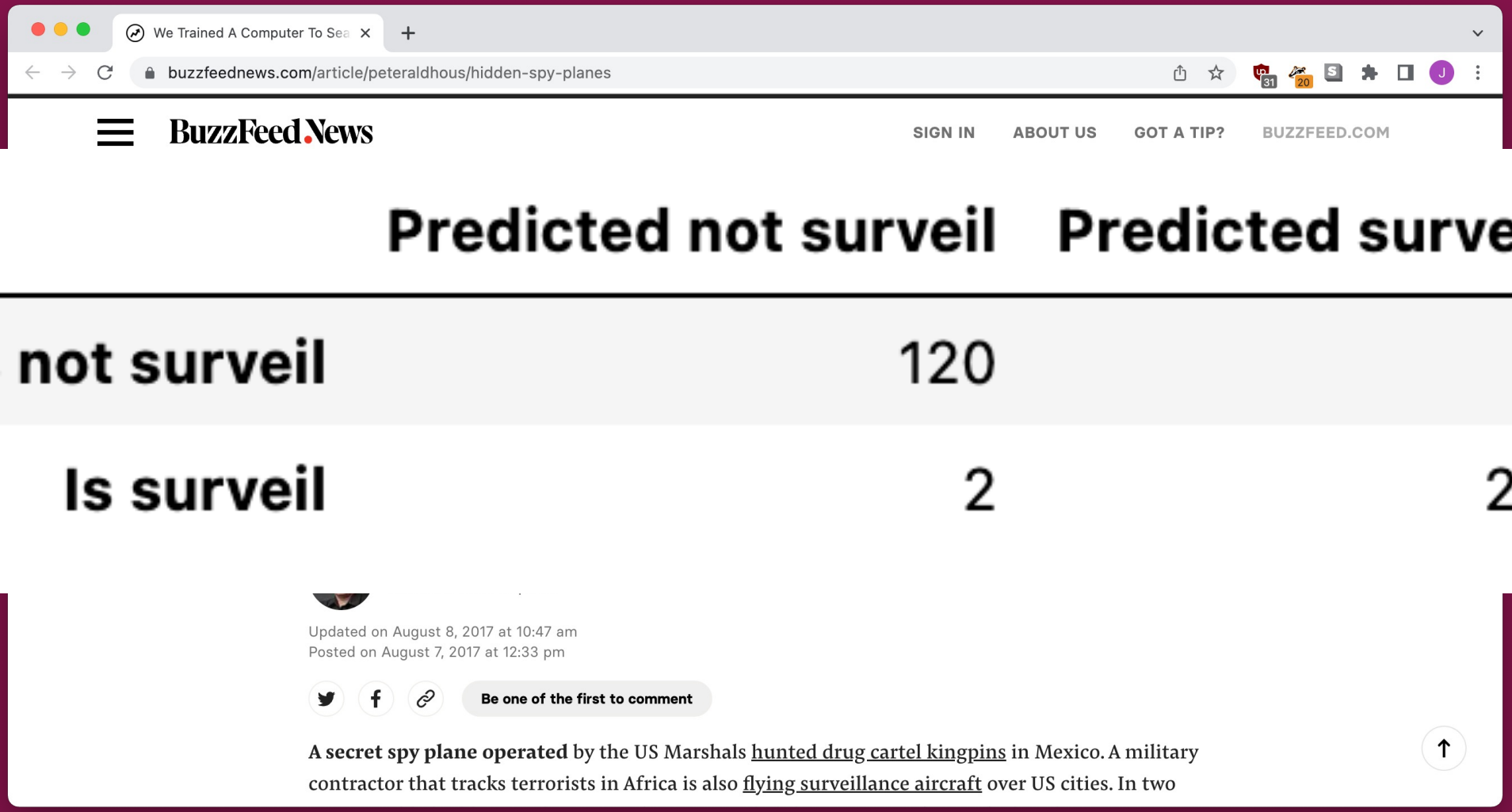
IS YOUR DATA COMPLETE?



WHERE'S YOUR DATA FROM?

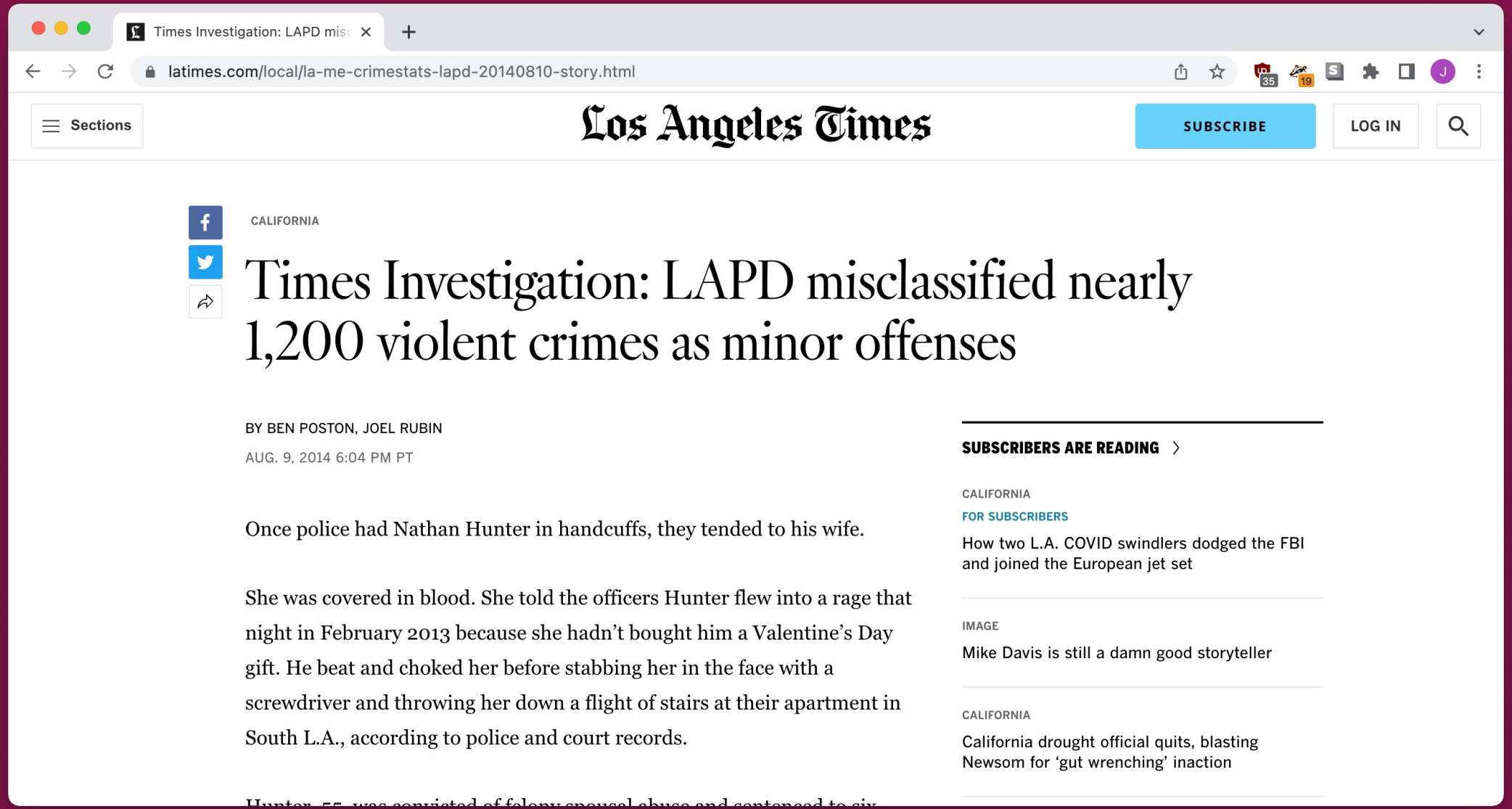
	content	textblob	textblob_bayes	nltk
0	I love love love love this kitten	0.5	-0.0879325	0.9571
1	I hate hate hate hate this keyboard	-0.8	-0.214151	-0.9413
2	I'm not sure how I feel about toast	-0.25	0.394659	-0.2411
3	Did you see the baseball game yesterday?	-0.4	0.61305	0
4	The package was delivered late and the contents were broken	-0.35	-0.57427	-0.4767
5	Trashy television shows are some of my favorites	0	0.0400757	0.4215
6	I'm seeing a Kubrick film tomorrow, I hear not so great things about it.	0.8	0.717875	-0.6296
7	I find chirping birds irritating, but I know I'm not the only one	-0.2	0.257148	-0.25

“TESTING” YOUR MODEL



	Predicted not surveil	Predicted surveil
Is not surveil	120	4
Is surveil	2	24

“TESTING” YOUR MODEL



LET'S PRODUCE SOME STORIES!

FINDING UNUSUAL DATA

Does reality match what you'd expect?

1. Collect your data
2. Train your model
3. Make predictions: what outcomes don't match?

Click for LINEAR REGRESSION

FINDING BIAS

What *really* affects an outcome?

1. Find a situation where the outcome might be biased
2. Decide what might affect the outcome
3. Collect your data
4. Train your model
5. What inputs are most important to the outcome?

Click for LOGISTIC REGRESSION

AUTOMATING WORKFLOWS

Tell the computer: “do what I just did, but faster”

1. Collect the data you’d like to organize or sort
2. Label some of the items
3. Train your model
4. Predict labels for the unlabeled data

Click for CLASSIFICATION

MACHINE LEARNING FOR NEWSROOMS

Jonathan Soma
Columbia University
js4571@columbia.edu
@dangerscarf

https://bit.ly/djc-2022-ml