

MACHINE LEARNING FOR NEWSROOMS

Jonathan Soma
Columbia University
js4571@columbia.edu
[@dangerscarf](https://twitter.com/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

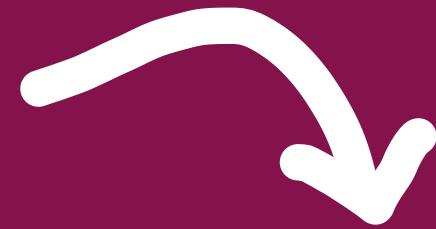
HOW M.L. WORKS

INPUT

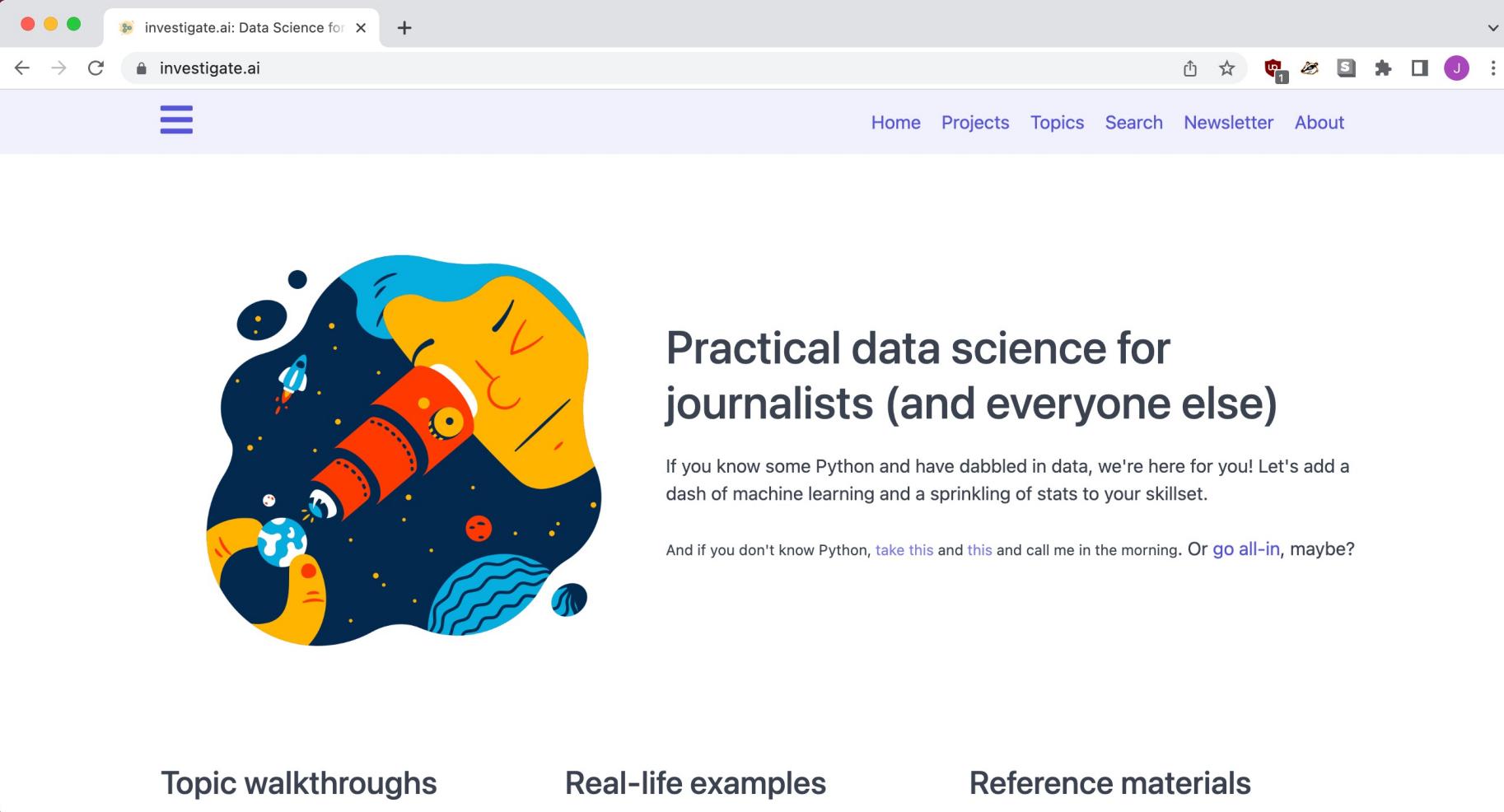


MAGIC BOX OF
MACHINE
LEARNING!

OUTPUT



INVESTIGATE.AI



The screenshot shows a web browser window with the title bar "investigate.ai: Data Science for X". The address bar contains "investigate.ai". The page itself has a light gray header with a blue three-line menu icon on the left and a navigation bar on the right with links for Home, Projects, Topics, Search, Newsletter, and About. Below the header is a large, colorful illustration of a telescope pointing towards a starry sky with planets and a rocket ship. To the right of the illustration, the text "Practical data science for journalists (and everyone else)" is displayed in a large, bold, dark font. Below this, a smaller text block reads: "If you know some Python and have dabbled in data, we're here for you! Let's add a dash of machine learning and a sprinkling of stats to your skillset." At the bottom of the page, there are three calls-to-action: "Topic walkthroughs", "Real-life examples", and "Reference materials".

Practical data science for
journalists (and everyone else)

If you know some Python and have dabbled in data, we're here for you! Let's add a dash of machine learning and a sprinkling of stats to your skillset.

And if you don't know Python, [take this](#) and [this](#) and call me in the morning. Or [go all-in](#), maybe?

Topic walkthroughs Real-life examples Reference materials

**BE A JOURNALIST,
NOT A
STATISTICIAN OR
DATA SCIENTIST**



+ tip

\$7



+ tip

\$12



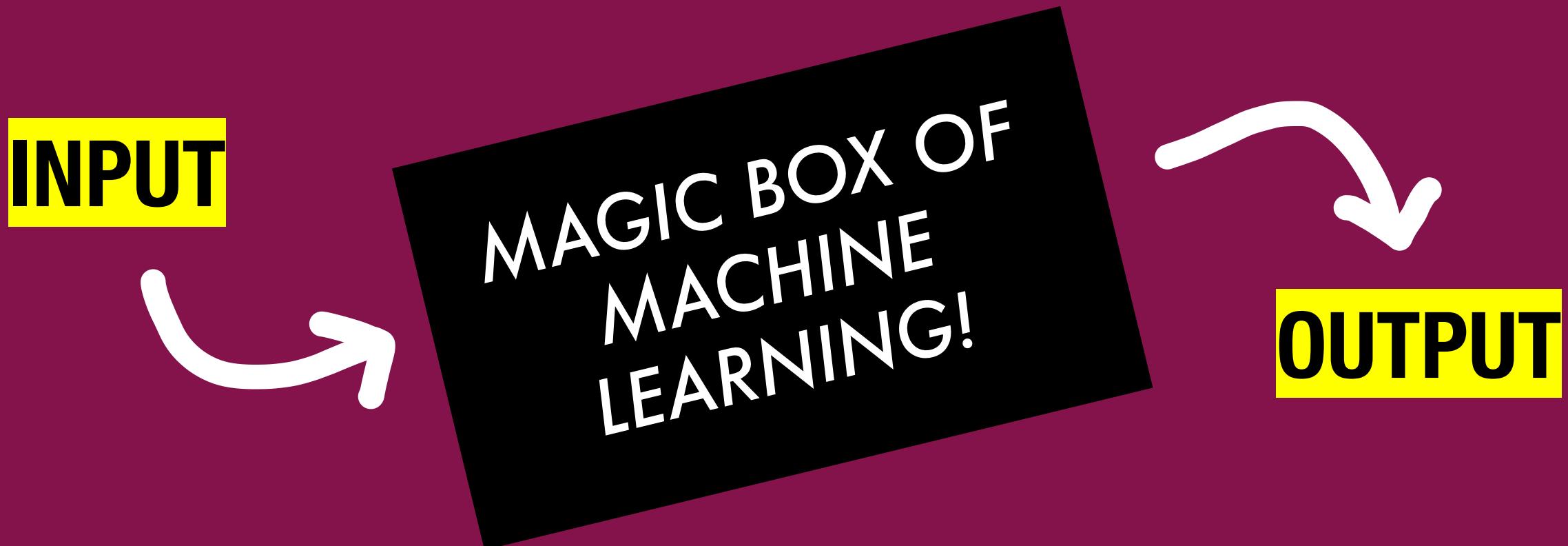
+ tip

\$15

“TRAINING” A “MODEL”



WHAT IS A “MODEL”?



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

DALL-E, MIDJOURNEY

Hybridized orange and durian



SKLEARN, TENSORFLOW

The screenshot shows a web browser window displaying the scikit-learn documentation for clustering. The page title is "2.3. Clustering — scikit-learn 1.x". The left sidebar contains navigation links for "scikit-learn 1.1.1", "Other versions", and a note to "cite us". The main content area discusses Mutual Information, Normalized Mutual Information (NMI), and Adjusted Mutual Information (AMI). It includes code snippets demonstrating the use of the `adjusted_mutual_info_score` function from the `sklearn.metrics` module. The browser interface includes standard navigation buttons, a search bar, and a tab bar.

Given the knowledge of the ground truth class assignments `labels_true` and our clustering algorithm assignments of the same samples `labels_pred`, the **Mutual Information** is a function that measures the **agreement** of the two assignments, ignoring permutations. Two different normalized versions of this measure are available, **Normalized Mutual Information (NMI)** and **Adjusted Mutual Information (AMI)**. NMI is often used in the literature, while AMI was proposed more recently and is **normalized against chance**:

```
>>> from sklearn import metrics
>>> labels_true = [0, 0, 0, 1, 1, 1]
>>> labels_pred = [0, 0, 1, 1, 2, 2]

>>> metrics.adjusted_mutual_info_score(labels_true, labels_pred)
0.22504...
```

One can permute 0 and 1 in the predicted labels, rename 2 to 3 and get the same score:

```
>>> labels_pred = [1, 1, 0, 0, 3, 3]
>>> metrics.adjusted_mutual_info_score(labels_true, labels_pred)
0.22504...
```

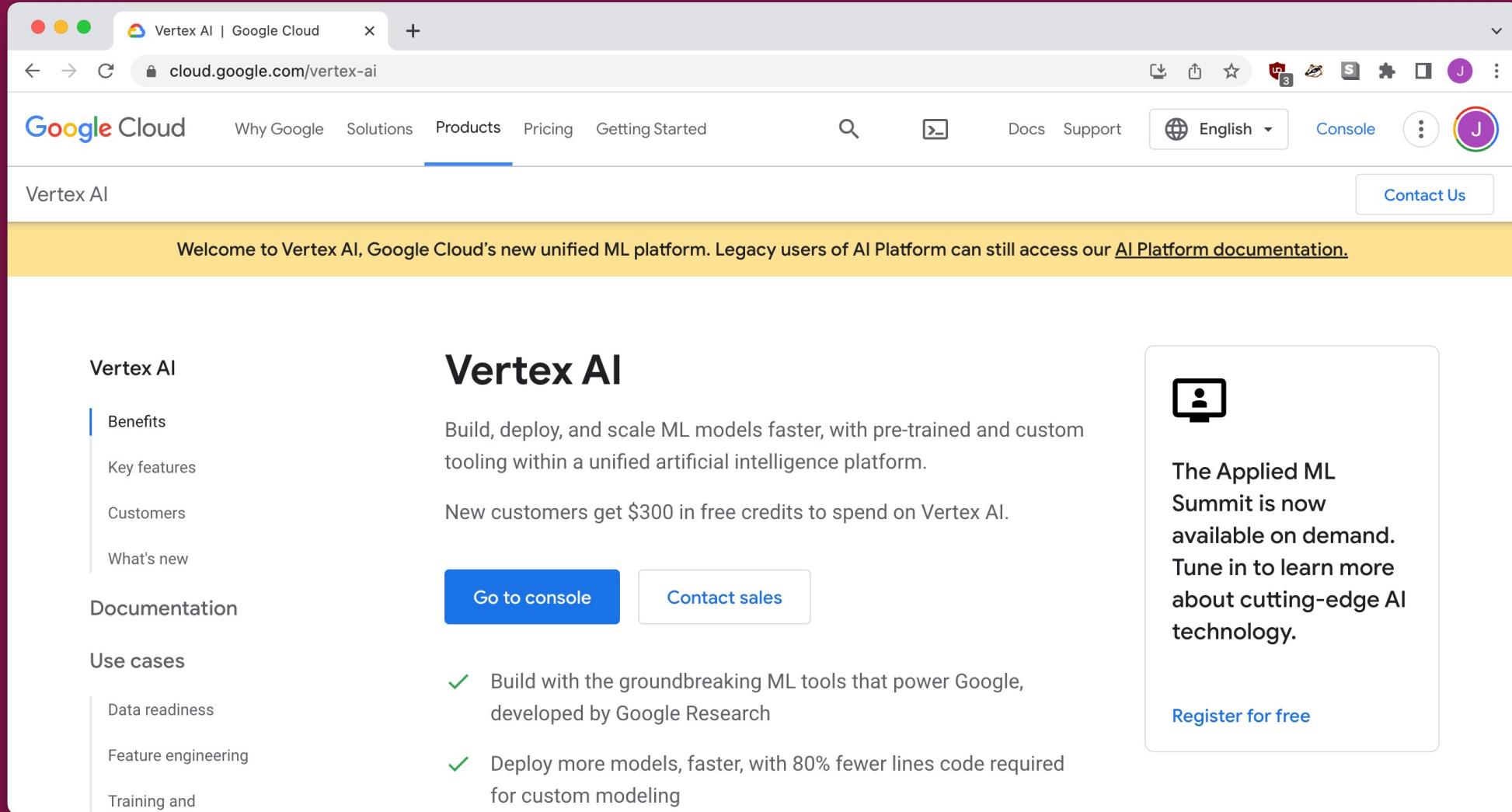
All, `mutual_info_score`, `adjusted_mutual_info_score` and `normalized_mutual_info_score` are symmetric: swapping the argument does not change the score. Thus they can be used as a **consensus measure**:

```
>>> metrics.adjusted_mutual_info_score(labels_pred, labels_true)
0.22504...
```

Perfect labeling is scored 1.0:

```
>>> labels_pred = labels_true[:]
>>> metrics.adjusted_mutual_info_score(labels_true, labels_pred)
1.0
```

REFINERY, VERTEX AI, ETC...



A screenshot of a web browser displaying the Vertex AI page on Google Cloud. The page features a navigation bar with links for Google Cloud, Why Google, Solutions, Products (which is underlined), Pricing, Getting Started, Docs, Support, English, Console, and Contact Us. A yellow banner at the top states, "Welcome to Vertex AI, Google Cloud's new unified ML platform. Legacy users of AI Platform can still access our [AI Platform documentation](#)." The main content area has a heading "Vertex AI" and a sub-section "Benefits" which includes links for Key features, Customers, and What's new. It also features a "Documentation" section with links for Data readiness, Feature engineering, and Training and. To the right, there is a large callout box with a person icon and text about the Applied ML Summit being available on demand, along with a "Register for free" button.

Welcome to Vertex AI, Google Cloud's new unified ML platform. Legacy users of AI Platform can still access our [AI Platform documentation](#).

Vertex AI

Build, deploy, and scale ML models faster, with pre-trained and custom tooling within a unified artificial intelligence platform.

New customers get \$300 in free credits to spend on Vertex AI.

[Go to console](#) [Contact sales](#)

- ✓ Build with the groundbreaking ML tools that power Google, developed by Google Research
- ✓ Deploy more models, faster, with 80% fewer lines code required for custom modeling

The Applied ML Summit is now available on demand. Tune in to learn more about cutting-edge AI technology.

[Register for free](#)

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

RESUME SCREENER

Amazon Fired Its Resume-Reading AI for Sexism

Now abandoned, the project shows the risks of training artificial intelligence on biased data.

BY DAVID GROSSMAN OCT 10, 2018



“BIAS LAUNDERING”

A screenshot of a web browser displaying an article from MIT Technology Review. The article title is "Predictive policing algorithms are racist. They need to be dismantled." by Will Douglas Heaven, published on July 17, 2020. The background of the article page features a dark blue gradient with a subtle grid pattern. The MIT Technology Review logo is at the top left, and a "Subscribe" button is at the top right. The browser's address bar shows the URL: technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-mac... .

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

By Will Douglas Heaven

July 17, 2020

IS YOUR DATA COMPLETE?

Why facial recognition's racial bias problem is so hard to crack

cnet.com/science/why-facial-recognition's-racial-bias-problem-is-so-hard-to-crack/

CNET Your guide to a better future

Science

Why facial recognition's racial bias problem is so hard to crack

Good luck if you're a woman or a darker-skinned person.

Queenie Wong 
March 27, 2019 5:00 a.m. PT

7 min read 



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

The screenshot shows a web browser window with the title bar "We Trained A Computer To See" and the URL "buzzfeednews.com/article/peteraldhous/hidden-spy-planes". The page header includes the BuzzFeed logo, navigation links for "SIGN IN", "ABOUT US", "GOT A TIP?", and "BUZZFEED.COM", and various user interface icons.

Predicted not surveil Predicted surveil

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

Updated on August 8, 2017 at 10:47 am
Posted on August 7, 2017 at 12:33 pm

Be one of the first to comment

A secret spy plane operated by the US Marshals hunted drug cartel kingpins in Mexico. A military contractor that tracks terrorists in Africa is also flying surveillance aircraft over US cities. In two

“TESTING” YOUR MODEL

A screenshot of a web browser displaying an article from the Los Angeles Times. The title of the article is "Times Investigation: LAPD misclassified nearly 1,200 violent crimes as minor offenses". The article is by Ben Poston and Joel Rubin, published on August 9, 2014, at 6:04 PM PT. The text discusses a case where police had Nathan Hunter in handcuffs but tended to his wife, who was covered in blood. The article continues to describe the investigation into LAPD's misclassification of violent crimes as minor offenses.

Times Investigation: LAPD misclassified nearly 1,200 violent crimes as minor offenses

BY BEN POSTON, JOEL RUBIN
AUG. 9, 2014 6:04 PM PT

Once police had Nathan Hunter in handcuffs, they tended to his wife. She was covered in blood. She told the officers Hunter flew into a rage that night in February 2013 because she hadn't bought him a Valentine's Day gift. He beat and choked her before stabbing her in the face with a screwdriver and throwing her down a flight of stairs at their apartment in South L.A., according to police and court records.

Hunter, 55, was convicted of felony spousal abuse and sentenced to six

CALIFORNIA

SUBSCRIBERS ARE READING >

CALIFORNIA

FOR SUBSCRIBERS

How two L.A. COVID swindlers dodged the FBI and joined the European jet set

IMAGE

Mike Davis is still a damn good storyteller

CALIFORNIA

California drought official quits, blasting Newsom for ‘gut wrenching’ inaction

**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://twitter.com/dangerscarf)

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