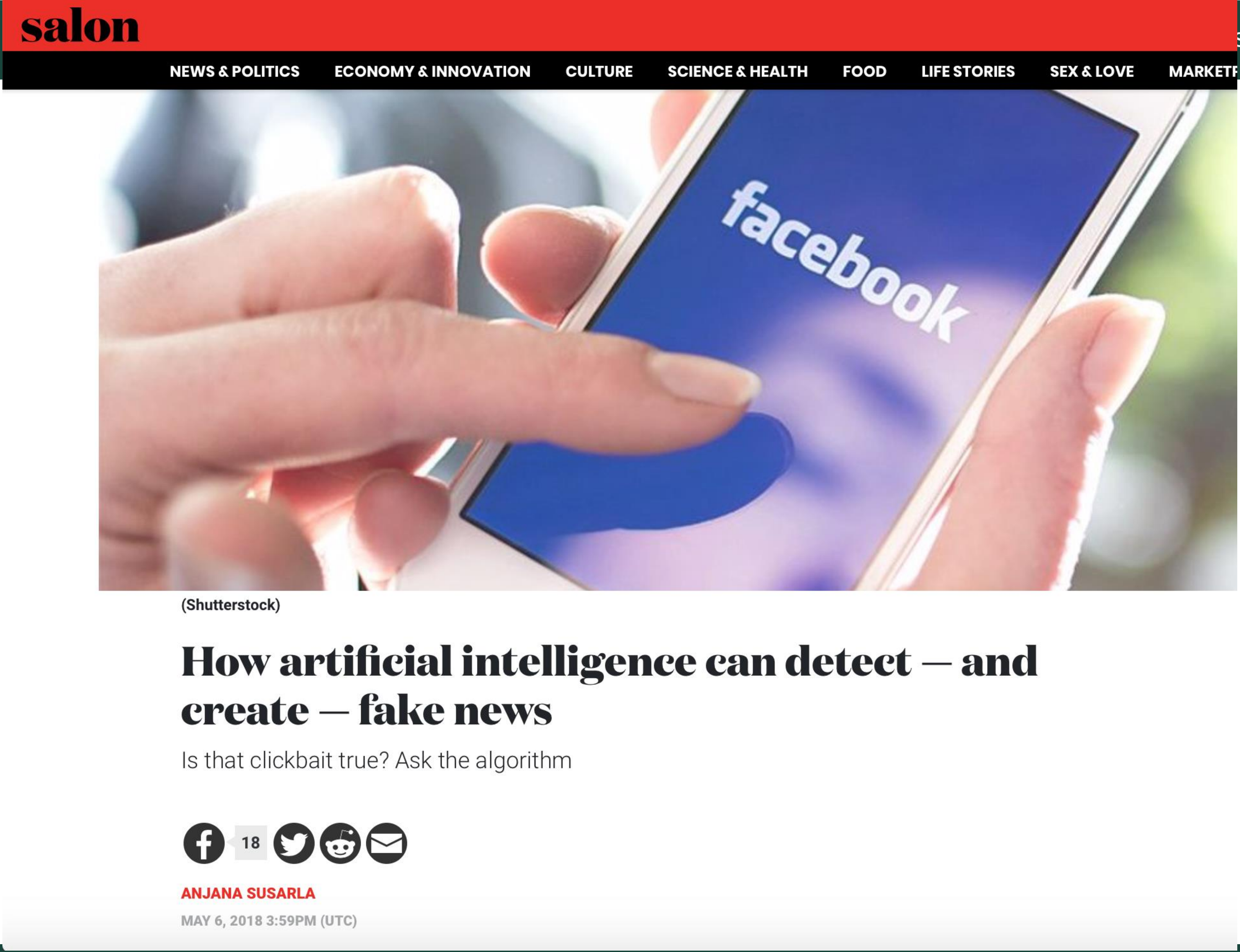
Teaching ML Without Coding

Anjana Susarla

Michigan State University

asusarla@msu.edu

Many thanks to Ravi Aron, Kartik Hosanagar, Dokyun Lee, Kiron Ravindran



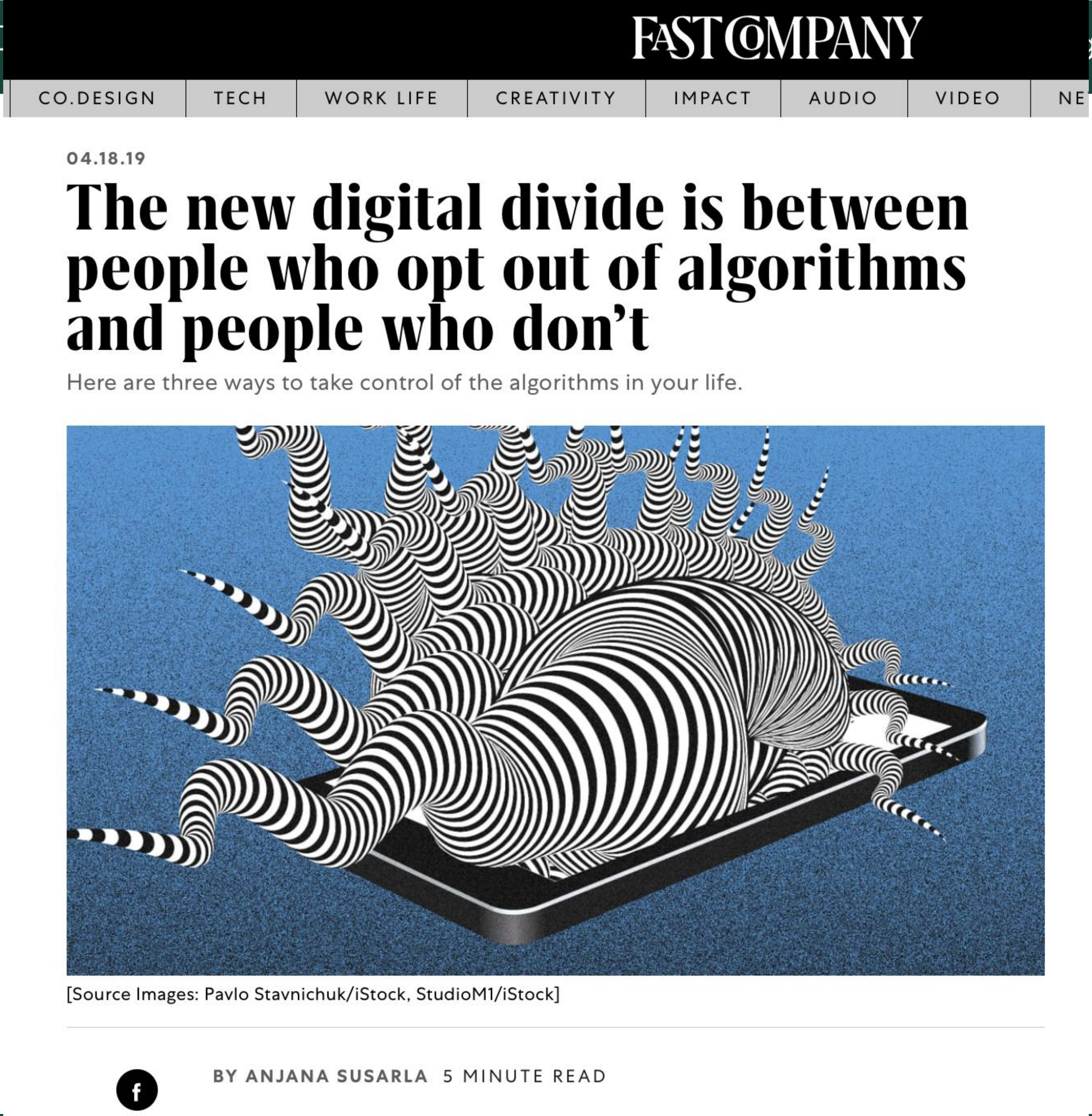
(Shutterstock)

How artificial intelligence can detect — and create — fake news

Is that clickbait true? Ask the algorithm

ANJANA SUSARLA

MAY 6, 2018 3:59PM (UTC)



FAST COMPANY

CO.DESIGN | TECH | WORK LIFE | CREATIVITY | IMPACT | AUDIO | VIDEO | NE

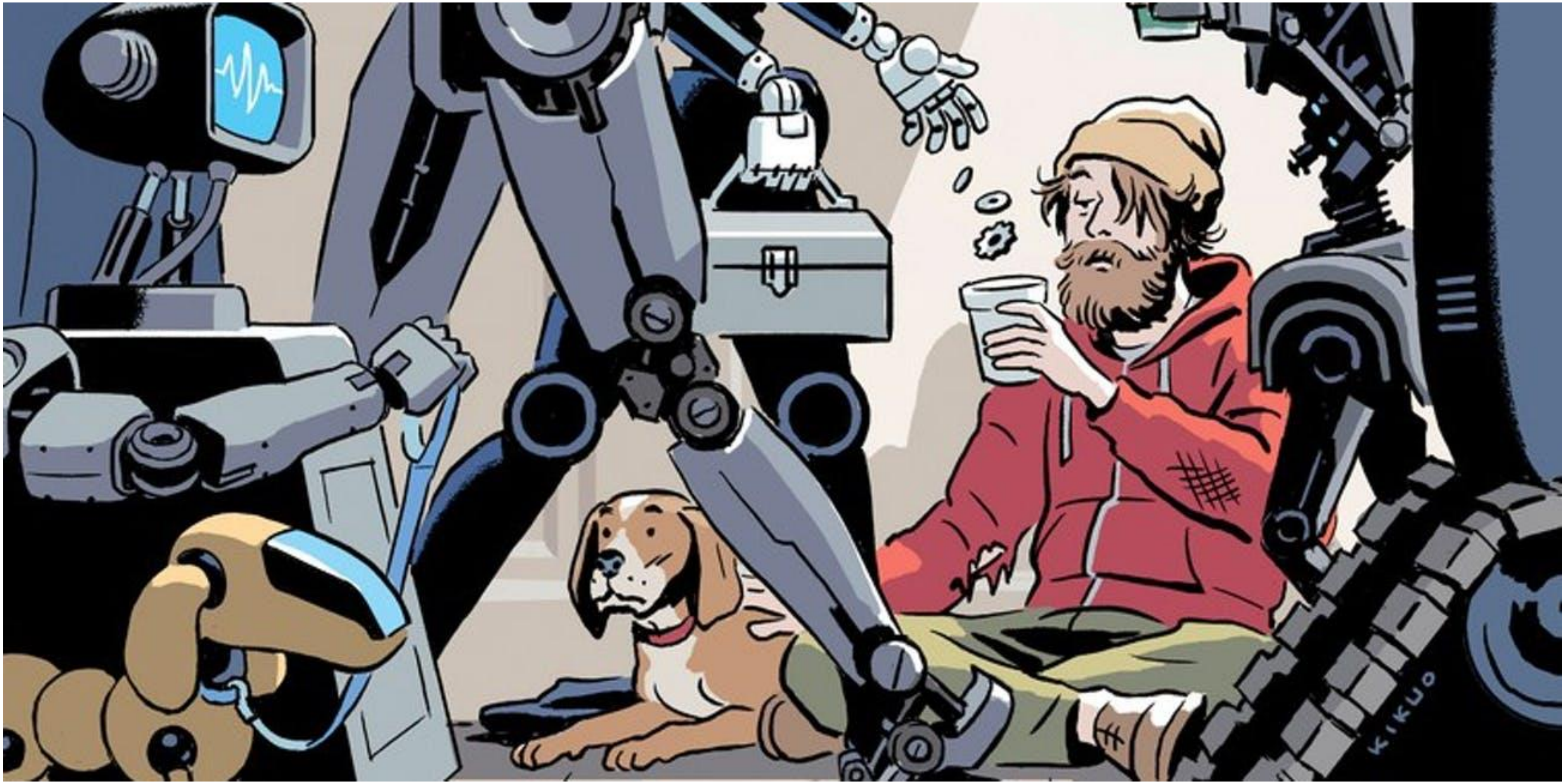
04.18.19

The new digital divide is between people who opt out of algorithms and people who don’t

Here are three ways to take control of the algorithms in your life.

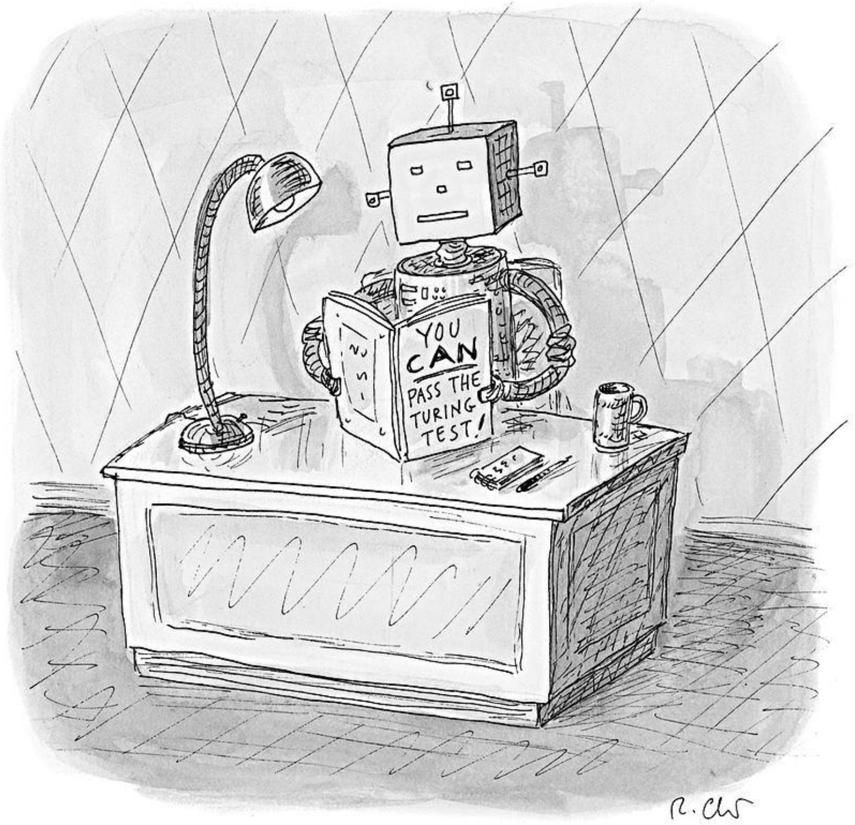
[Source Images: Pavlo Stavnichuk/iStock, StudioM1/iStock]

BY ANJANA SUSARLA 5 MINUTE READ

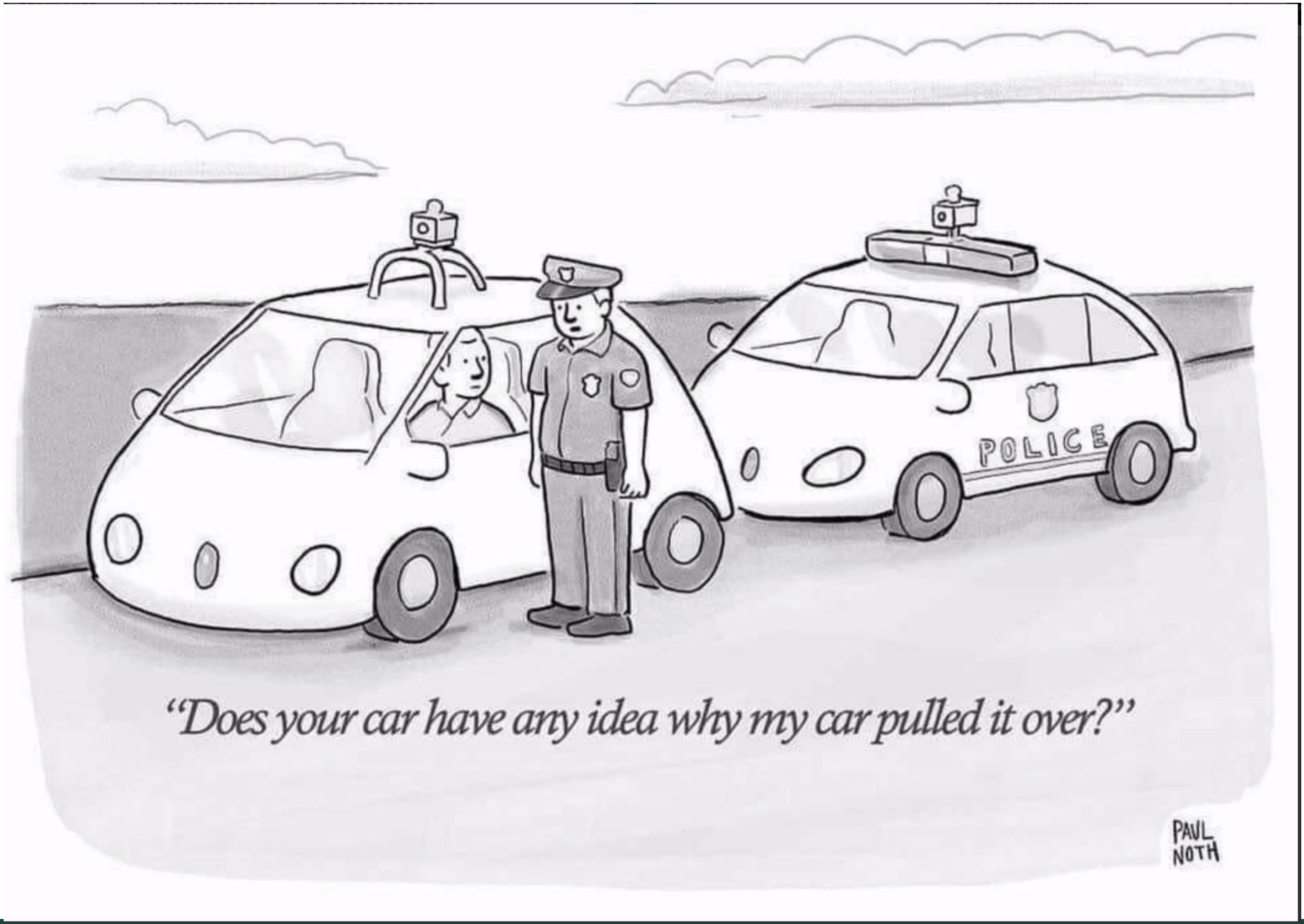


KIKUO

AI and automation

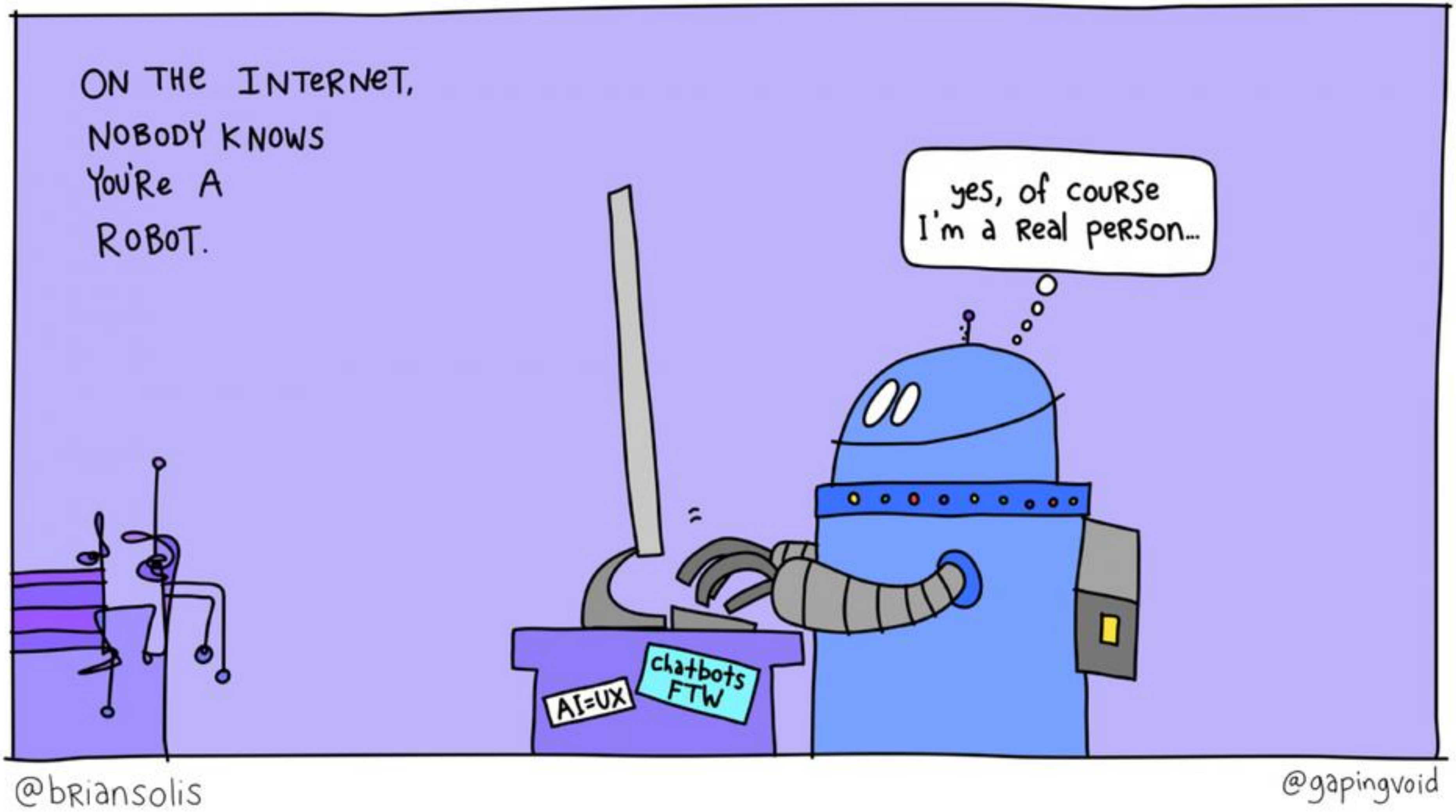


Interpretable ML?



"Does your car have any idea why my car pulled it over?"

Fairness, accountability & transparency


ON THE INTERNET, NOBODY KNOWS YOU'RE A ROBOT.
yes, of course I'm a real person...
AI=UX
chatbots FTW
@briansolis
@gapingvoid

How to teach ML without coding

- First, explain basics of decision making with data. In the domain of healthcare analytics. Key concepts:
  - Clinical impacts
  - Operational impacts
  - Population Health impacts
- Second, introduce students to different flavors of analytics
  - Key concept: Prescriptive, predictive and descriptive analytics
- Third, introduce students to the ML toolkit
  - Basic stats
  - Regression, Decision trees
  - Incorporate a few exercises with XLMiner
  - Key concepts: Supervised vs. unsupervised learning

How to teach stats without any formulae

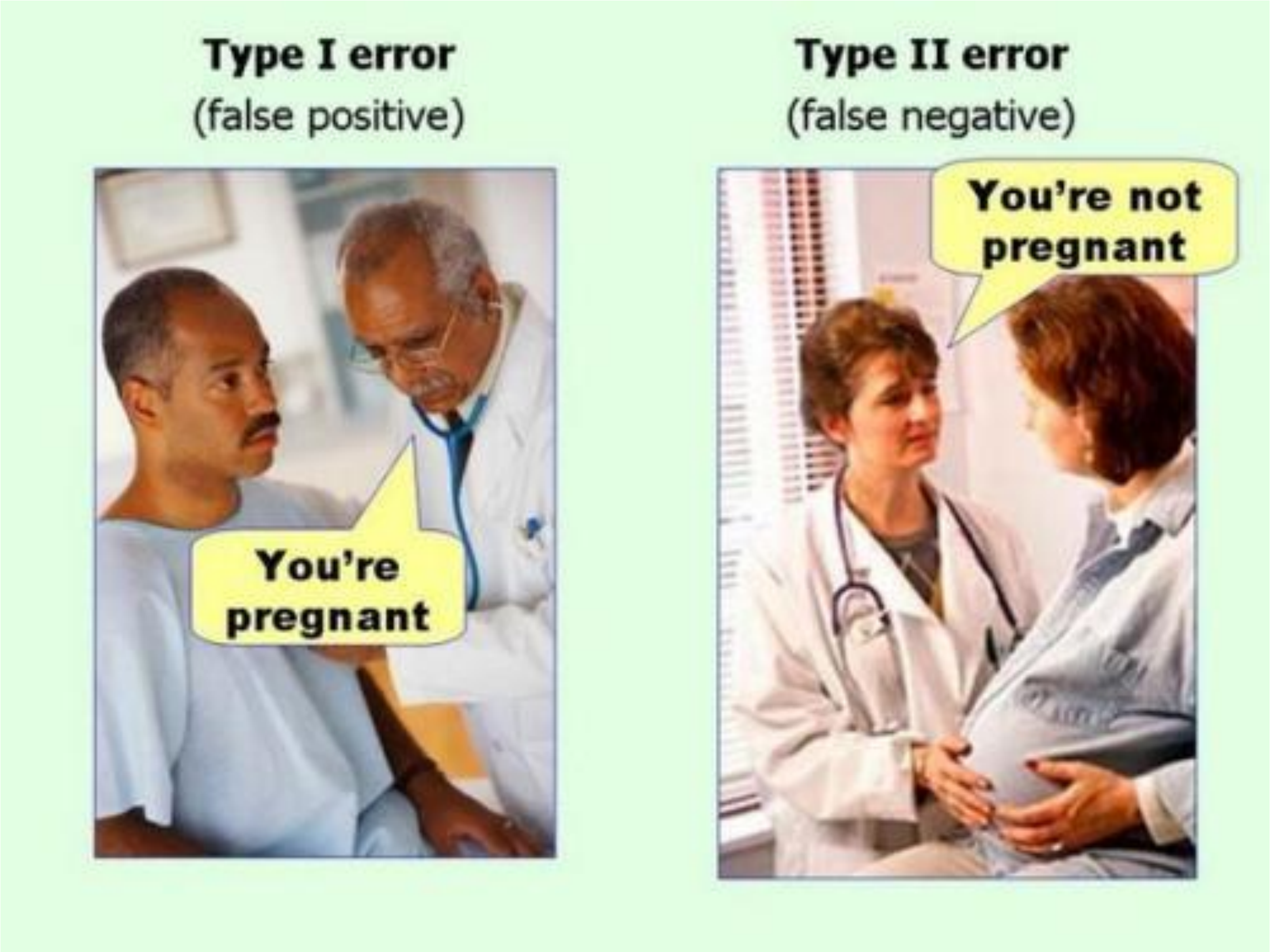
Correlation

- The degree to which two variables have a tendency to vary together
- Can be positive or negative
- Range: -1 to +1, 0 means no correlation

Examples:

- The more you study, the better you do on the exam
- The more I talk about correlation, the less you want to be here

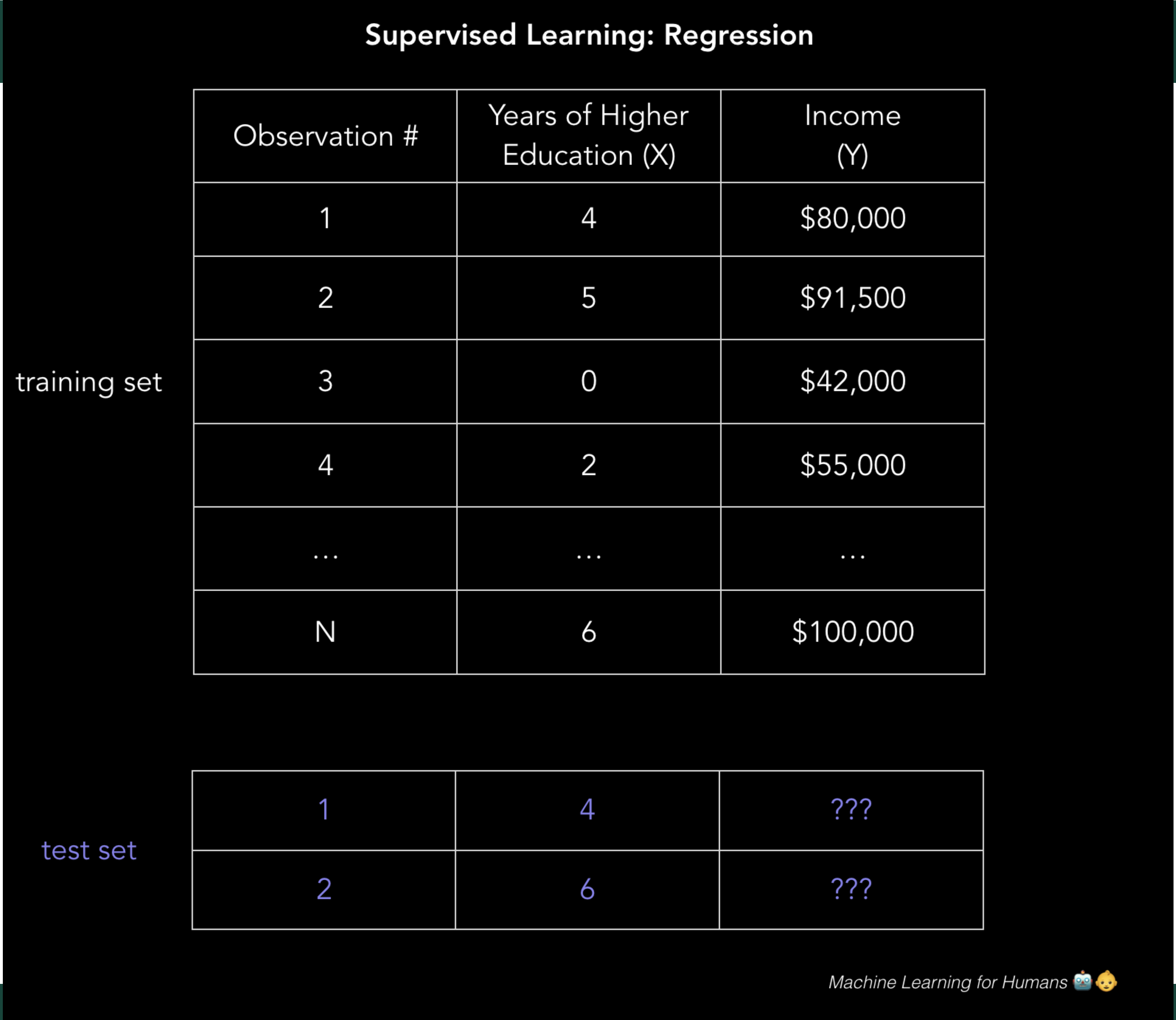
Type I vs. Type II Errors



Supervised vs. Unsupervised learning

What is the difference between Supervised and Unsupervised Learning

Supervised learning is done using a **ground truth**, or in other words, we have prior knowledge of what the output values for our samples should be. Therefore, the goal of supervised learning is to learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable i the data. Unsupervised learning, on the other hand, does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points.



Supervised Learning: Regression

training set

Observation #	Years of Higher Education (X)	Income (Y)
1	4	$80,000
2	5	$91,500
3	0	$42,000
4	2	$55,000
…	…	…
N	6	$100,000

test set

1	4	???
2	6	???

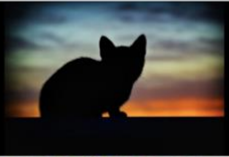
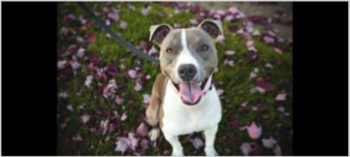
Machine Learning for Humans 🤖👶

Supervised Learning: Classification

training set

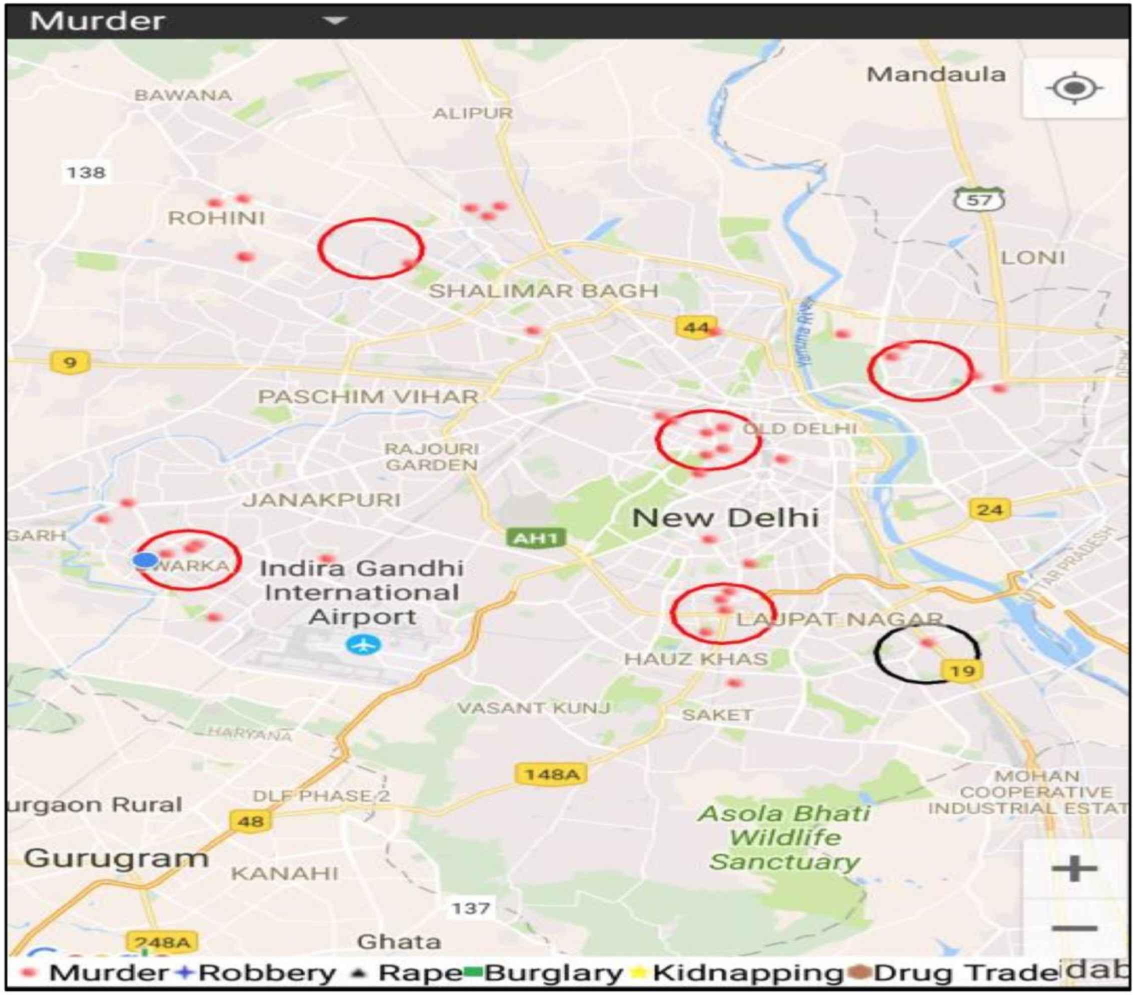
Observation #	Input image (X)	Label (Y)
1		“dog”
2		“cat”
3		“dog”
…	…	…
N		“dog”

test set

1		???
2		???

Machine Learning for Humans 🤖👶

A NETFLIX ORIGINAL SERIES
DELHI CRIME
ALL EPISODES MARCH 22
NETFLIX
D2


Murder
Mandaula
BAWANA
ALIPUR
138
ROHINI
57
LONI
SHALIMAR BAGH
44
9
PASCHIM VIHAR
OLD DELHI
RAJOURI GARDEN
JANAKPURI
New Delhi
24
GARH
AH1
DWARKA
Indira Gandhi International Airport
LAJPAT NAGAR
HAUZ KHAS
19
VASANT KUNJ
SAKET
HARYANA
148A
MOHAN COOPERATIVE INDUSTRIAL ESTAT
Gurgaon Rural
DLF PHASE 2
48
Asola Bhati Wildlife Sanctuary
Gurugram
KANAHI
137
248A
Ghata
● Murder ✚ Robbery ▲ Rape 🟩 Burglary ● Kidnapping ● Drug Trade

Performance of ML methods

Predicting Probabilities

- In a classification problem, we may decide to predict the class values directly.
- Alternately, it can be more flexible to predict the probabilities for each class instead. For example, a default might be to use a threshold of 0.5, meaning that a probability in [0.0, 0.49] is a negative outcome (0) and a probability in [0.5, 1.0] is a positive outcome (1).

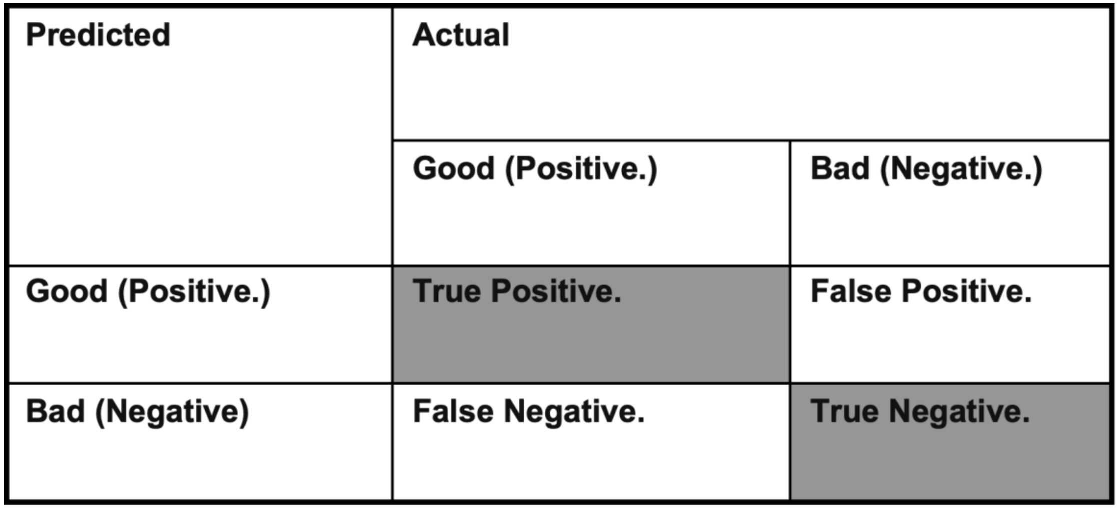
When making a prediction for a binary or two-class classification problem, there are two types of errors that we could make.

- **False Positive**. Predict an event when there was no event.
- **False Negative**. Predict no event when in fact there was an event.

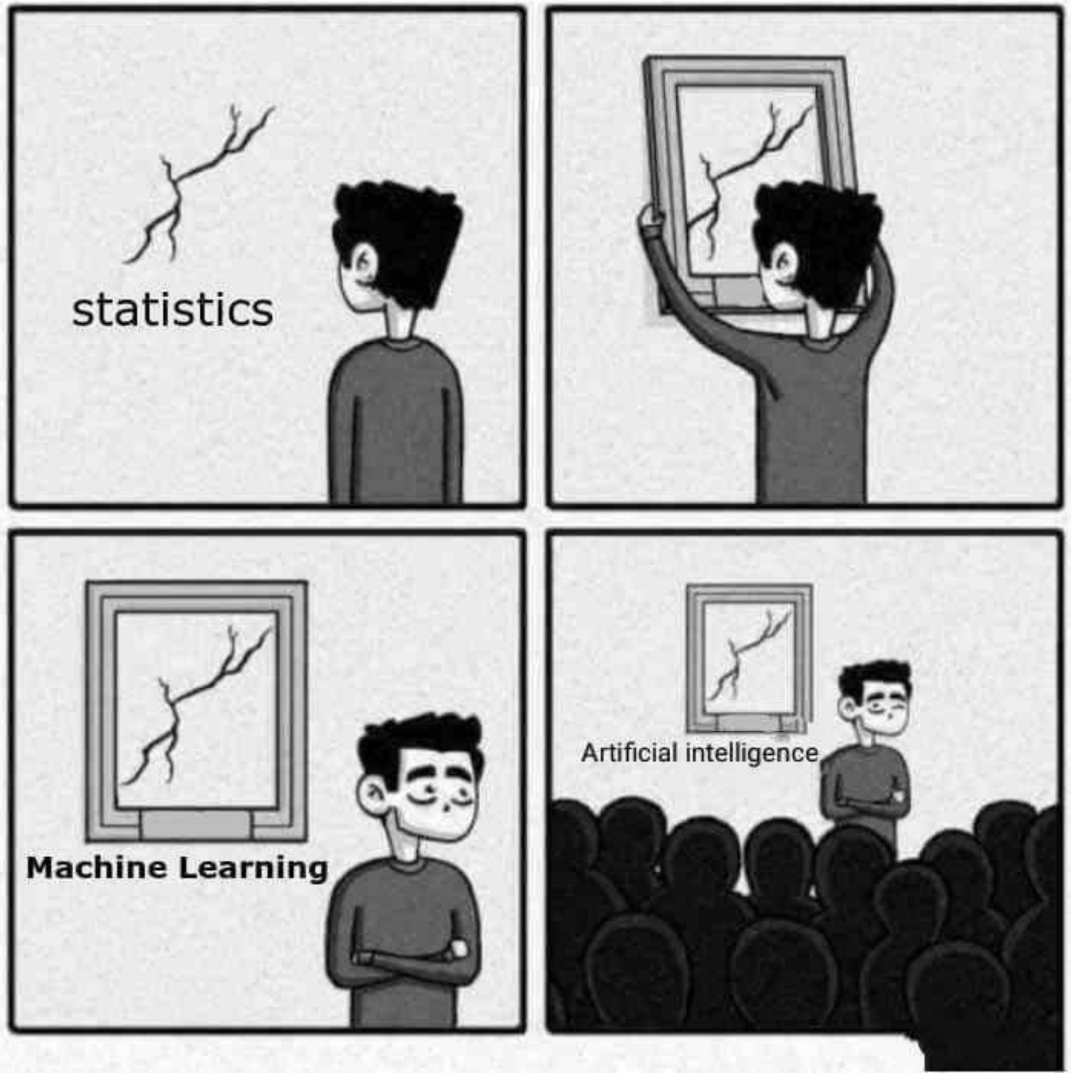
If you are an epidemiologist working an Ebola outbreak, then you don't want to have false negatives that end up being sent home to infect others. You want that number low. Do you care about false positives? Well, maybe not if the therapy won't kill someone, or maybe you do if a positive test means being put into a ward with people who are sick.

What about pregnancy tests to take at home? You probably don't worry too much about false negatives (pregnant women who test negative) because those women will still be pregnant and probably take the test again if they continue to miss their period or feel other signs/symptoms of pregnancy.

Performance



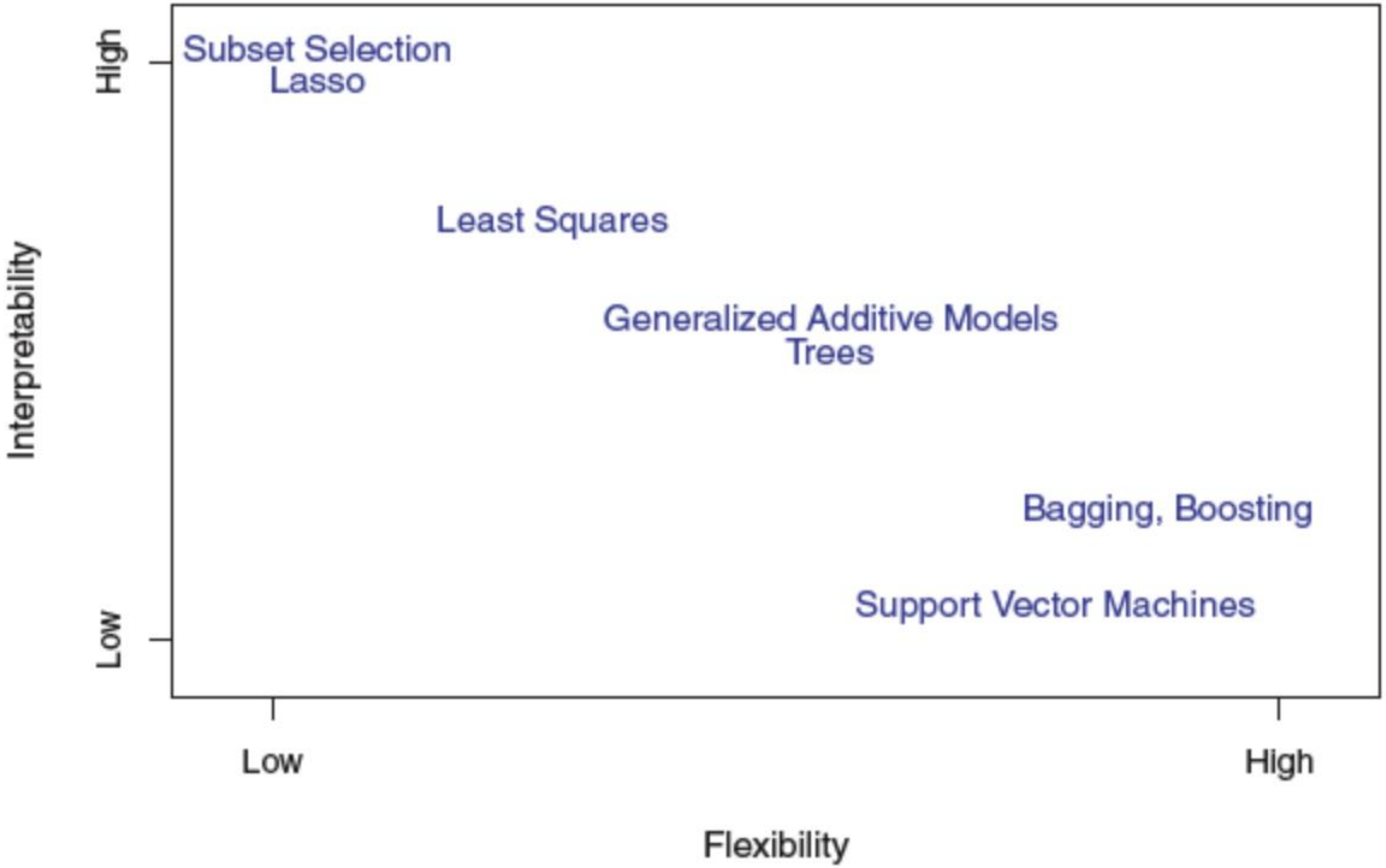
Predicted	Actual	
	Good (Positive.)	Bad (Negative.)
Good (Positive.)	True Positive.	False Positive.
Bad (Negative)	False Negative.	True Negative.



No, ML is not Stats!

Machine Learning = Representation + Evaluation + Optimization

Prediction vs. Explanation



Other Key Learning Objectives

- Prediction accuracy vs. model interpretability
- Model complexity and bias-variance
- Cross-validation
- Developing a model as a search problem

Mini Assignments (non-coding)

Reading: Bruce, P. & Bruce, A. (2017). Practical Statistics for Data Scientists. Sebastopol, CA: O’Reilly Media, Chapters 4 and 5

- Summarize and give an example of Predictive Analytics
- How is predictive analytics different from descriptive analytics?

Decision-Making Scenario 1: Clinical

Whitefield Health Services operates a chain of clinical care practices. Using machine learning, they were able to better manage the population of patients with chronic kidney diseases (CKD). Early diagnosis and effective disease management is a serious challenge especially given that several of patients with kidney disease also tend to exhibit complex clinical histories. A major goal of clinicians is to pro-actively manage the disease including early diagnosis and thereby determining the appropriate medical treatments at different stages of disease progression.

Whitefield collected detailed patient data for a sample of over 500 patients at each visit including patient’s demographic data, weight, blood pressure, blood sample test variables such as serum creatinine levels, fasting plasma glucose levels, lipid profile, calcium, phosphorus, hemoglobin, and other parameters. In addition, they collected detailed patient profiles and medical history to understand comorbidities and correlated health conditions. Using machine learning methods, the clinical practice was able to quantify and predict variations in the glomerular filtration rate (GFR), which is a reliable parameter of the renal function and progression of CKD. CKD is a progressive disease; thus, ongoing management and intervention is critical. Thus, analytics offers a way for clinics to learn the most probable clinical pathways and predicting future states associated with temporal patterns of biochemical measurements and patient subgroups. This in turn helps them manage future patient visits better and pro-actively manage disease progression.

Decision Making scenario -2

Hydrangea runs a retirement community for active seniors including rehabilitation and long term care after hospital stay. Patients may go to rehabilitation facilities to recover from hospital stay due to conditions such as a stroke, injury, or recent surgery. Rehab facilities generally require that patients be able to undergo at least three hours of physical and occupational therapy per day, five days a week. Patients at these facilities are presumed to be healthier than patients in a more typical hospital or a nursing home. But sometimes the care makes things worse. It has been estimated that almost 29 percent of patients in rehab facilities suffered a medication error, bedsore, infection or some other type of harm as a result of the care they received. The challenge is that the discharge-planning process in hospitals is inherently complex and needs information from several sources, such as patient-specific information as well as information about available resources.

The two major forces influencing the discharge-planning process over the last two decades are the Medicare prospective payment system (PPS) and the rise of managed care, both of which have created incentives to reduce the length of hospital stays.

The incentives enabled by PPS actually disadvantages Medicare patients by encouraging the early discharge of patients into post-acute care (PAC). Using cohorts of patients from Medicare, it has been found that patient debility accounts for 10% of inpatient rehabilitation cases among Medicare beneficiaries. Debility has the highest 30-day readmission rate among 6 impairment groups most commonly admitted to inpatient rehabilitation.

Hydrangea used descriptive analytics of very granular data from Medicare and combined with hospital data to develop detailed profiles of patients including comorbidities significantly associated with higher hazard of readmission such as: CMG comorbidity tier and chronic pulmonary disease (up to 30 days), congestive heart failure, fluid/electrolyte disorders, renal failure, peripheral vascular disease, weight loss, solid tumor without metastasis, coagulopathy, metastatic cancer, lymphoma, and liver diseases. Such detailed segmentation of patients allows for active monitoring and management of rehabilitation stay, and reduces the likelihood of readmissions.

Decision Making Scenario 3 (Population Health)

A chronic condition has been defined as “a physical or mental health condition that lasts more than one year and causes functional restrictions or requires ongoing monitoring or treatment”. Watson is an insurance provider that recognizes that focusing solely on reducing medical risks through preventive and chronic care management is not enough. For instance, consider the following four patients: Abdul, Bill, Mike and Sam. All of them were diagnosed with diabetes at the age of 47. By the time they turn 51, Abdul and Mike have their blood sugar under control and actively manage their own condition, including using social media to communicate with other patients on diet and exercise. However, Bill and Sam develop other complications and in the case of Sam, developed kidney disease from untreated diabetes. As four of them have comparable clinical care. However, Bill and Sam are exposed to more stress from their social, economic and environmental issues. Such social determinants of health have been shown to lead to adverse health and medical outcomes.

Health plans have traditionally used claims data to segment and stratify patients to identify those at higher-risk for chronic care management, by looking at metrics such as the frequency of visits to the emergency department (ED). It is considerably more challenging to identify those that are currently healthy but who might experience a major medical event that would significantly change their consumption of resources. Identifying users with a high likelihood of having a major medical event (and those that would significantly use medical resources) enables the healthcare providers and stakeholders to intervene and change their health trajectory. This is extremely important not only in terms of improving a patient’s quality of life and health outcomes but also improves societal outcomes by lowering costs and ensuring that resources are targeted in an equitable manner. The challenge is that developing such detailed population profiles of patients cannot be done with claims data alone, since claims data does not indicate those at elevated risks to health conditions from social and population level determinants of health.

If only claims data was used, all four patients would have been identified as needing chronic care management since they were all diagnosed with diabetes. It is important to consider additional data that highlights the specific risk factors associated with socioeconomic conditions faced by Bill and Sam. Relevant data could include:

- US census data, including the average income and the density of housing.
- Marketing or consumer data, including credit card transactions.
- Community rankings, such as the Gallup-Healthways Well-Being Index, that provide additional information about the neighborhood a member lives in.

Watson used machine learning based models create detailed profiles of patients including medical and behavioral factors, social determinants of health, etc. to create a risk score to each member of the insurance company, which helps identify what type of services are necessary to enable ongoing care management. For instance, marketing data may reveal that Sam does not own a car and also lives far from his health care provider, which may mean the member has transportation challenges. Watson was able to use population health analytics to usher in personalized medicine by recommending lifestyle changes along with clinical management of symptoms.

Mini Assignments (Non-Coding)

Aligned reading:

McKinsey Report on Healthcare.

Ginsburg, P. B., Loera-Brust, A., Brandt, C., & Durak, A. (2018). The opportunities and challenges of data analytics in health care. Retrieved from https://www.brookings.edu/

- Identify and explain two hypothetical scenarios where analytics can benefit the hospital.
- How can descriptive analytics and predictive analytics be used in healthcare? Does one offer more benefits than the other? Explain.
- The Brookings report refers to the slow pace of innovation in healthcare. What are some reasons why that is the case? Explain one industry outside healthcare that has moved to adopt data analytics at a faster pace?
- How would the adoption of payment for care policies impact the adoption of data analytics in healthcare?