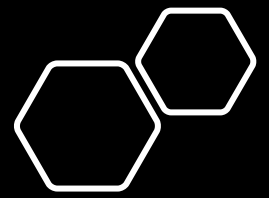
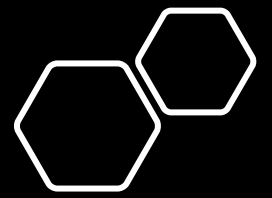
# *On the Road to Reducing Information Asymmetries from Black-Box Evidence: A Conceptual Introduction to Machine Learning for Lawyers*

Nitin Kohli
PhD Candidate, UC Berkeley School of Information
Email: nitin.kohli@ischool.berkeley.edu
February 17, 2021



# *A (very quick) conceptual introduction to supervised machine learning*

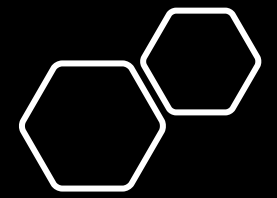


| age | gender | Top_genre | Avg_app_time | like_ke$ha |
|---|---|---|---|---|
| 13-17 | M | pop | 190 | 0 |
| 13-17 | M | rb | 12 | 1 |
| 18-24 | F | jazz | 34 | 1 |
| 55-64 | M | rb | 98 | 0 |
| 45-54 | O | punk | 12 | 1 |
| 13-17 | F | rock | 14 | 1 |
| 13-17 | F | pop | 17 | 0 |
| 18-24 | F | pop | 87 | 0 |
| 65+ | M | electronic | 91 | 0 |
| 45-54 | O | edm | 367 | 1 |
| … | … | … | … | … |

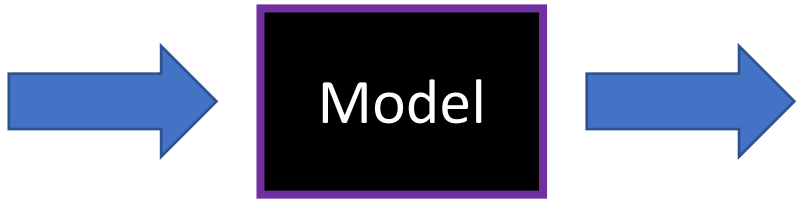
**Features** | **Targets**

*Goal*

Predict “like_ke$ha” from data



| age | gender | Top_genre | Avg_app_time |
|---|---|---|---|
| 13-17 | M | pop | 190 |
| 13-17 | M | rb | 12 |
| 18-24 | F | jazz | 34 |
| 55-64 | M | rb | 98 |
| 45-54 | O | punk | 12 |
| 13-17 | F | rock | 14 |
| 13-17 | F | pop | 17 |
| 18-24 | F | pop | 87 |
| 65+ | M | electronic | 91 |
| 45-54 | O | edm | 367 |
| … | … | … | … |




| like_ke$ha |
|---|
| 0 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 0 |
| 0 |
| 0 |
| 1 |
| … |

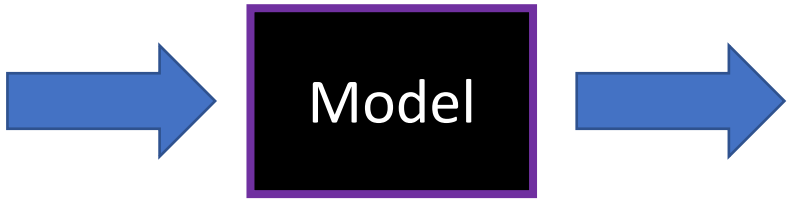
**Features**

**Targets**

*Goal*

Predict “like_ke$ha” from data

| age | gender | Top_genre | Avg_app_time |
|---|---|---|---|
| 18-24 | O | electronic | 27 |



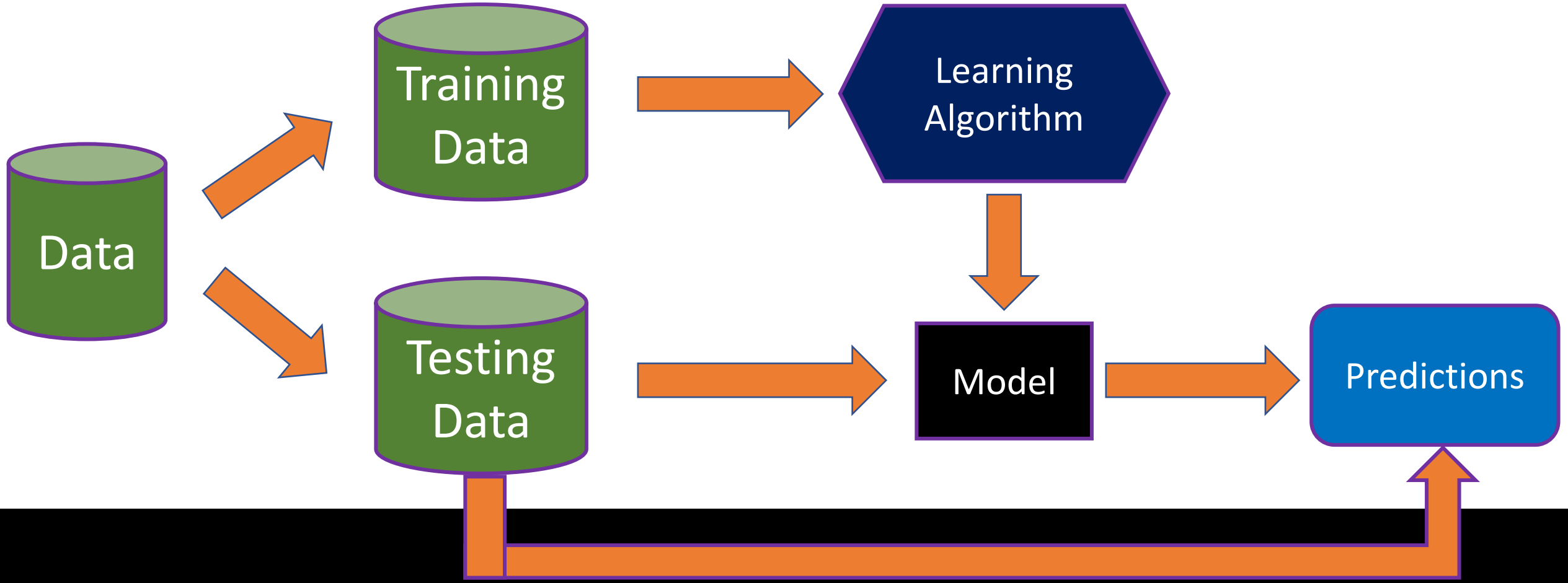

| like_ke$ha |
|---|
| 0 |

**Unknown Example**

**Prediction**

*Desiderata*

Needs to “work well” for any input (previously seen or not)




*Supervised ML Pipeline*

Two Algorithms here:

1. Learning Algorithm
2. Model = Learned Algorithm

# *Problem Formulation*

# Construct Validity and Poor Proxies

*Operationalizing a concept with a proxy may not faithfully measure the phenomenon of interest.*

## How We Analyzed the COMPAS Recidivism Algorithm

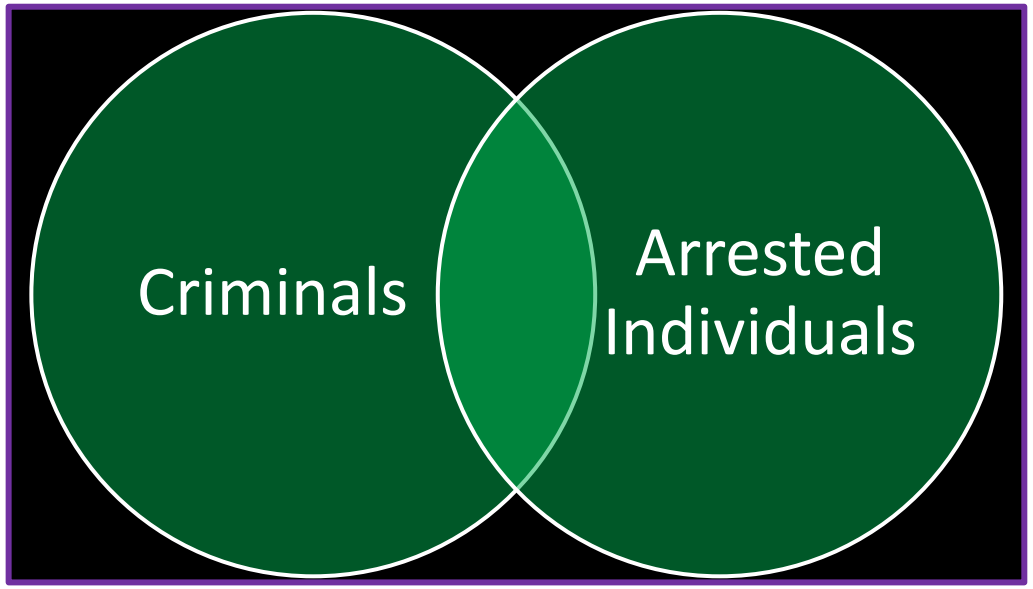
*by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*
May 23, 2016

TECHNOLOGY

## A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant's risk of committing more crimes, but a new study puts its usefulness into perspective.

ED YONG JAN 17, 2018

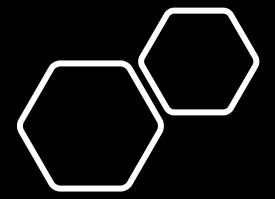


## TECHNICAL FLAWS OF PRETRIAL RISK ASSESSMENTS RAISE GRAVE CONCERNS

Some risk assessments define public safety risk more narrowly as the risk that a person will be arrested for a violent crime while on pretrial release. But because pretrial violence is exceedingly rare, it is challenging to statistically predict. Risk assessments cannot identify people who are more likely than not to commit a violent crime. The fact is, the vast majority of even the highest risk individuals will not go on to be arrested for a violent crime while awaiting trial. Consider the dataset used to build the Public Safety Assessment (PSA): 92% of the people who were flagged for pretrial violence did not get arrested for a violent crime and 98% of the people who were not flagged did not get arrested for a violent crime.[4] If these tools were calibrated to be as accurate as possible, then they would predict that

*Problem Formulation*

Construct Validity and Poor Proxies



# *Input Data*

## Systematic Issues in Underlying Data Collection

$$Total\ Crime$$

$$=$$

$$Observed + Unobserved\ Crime$$

$$\geq$$

$$Observed\ Crime$$

$$\geq$$

$$Reported\ \&\ Observed\ Crime$$

# *Input Data*

# Systematic Issues in Underlying Data Collection

*This is math, not magic.*

"Data science tools" cannot allow us to generalize to this level (absent major additional assumptions)


Real Datasets Live Here

**Rachel Thomas**
@math_rachel

Follow

If we don't want the future to look like the past, we can't just unthinkingly apply machine learning. – Nitin Kohli

2:26 PM - 7 Mar 2019

**28** Retweets **102** Likes

1 28 102

# *Evaluation Metrics*

How do we assess the "correctness" of a model?

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | 50 | 50 |
| | 0 | 50 | 850 |

# *Evaluation Metrics*

How do we assess the "correctness" of a model?

Predicted

| Actual | | 1 | 0 |
|---|---|---|---|
| | 1 | 50 | 50 |
| | 0 | 50 | 850 |

Accuracy = Percentage correct
= (50 + 850) / (1000)
= 90%

# *Evaluation Metrics*

How do we assess the "correctness" of a model?

Predicted

| Actual | | 1 | 0 |
|---|---|---|---|
| | 1 | 0 | 100 |
| | 0 | 0 | 900 |

# *Evaluation Metrics*

How do we assess the "correctness" of a model?

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | 0 | 100 |
| | 0 | 0 | 900 |

Accuracy = Percentage correct
= (0 + 900) / (1000)
= 90%

# *Evaluation Metrics*

How do we assess the "correctness" of a model?

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | 0 | 100 |
| | 0 | 0 | 900 |

But this model is practically useless because it never predicts 1!
It always just predicts 0

# *Evaluation Metrics*

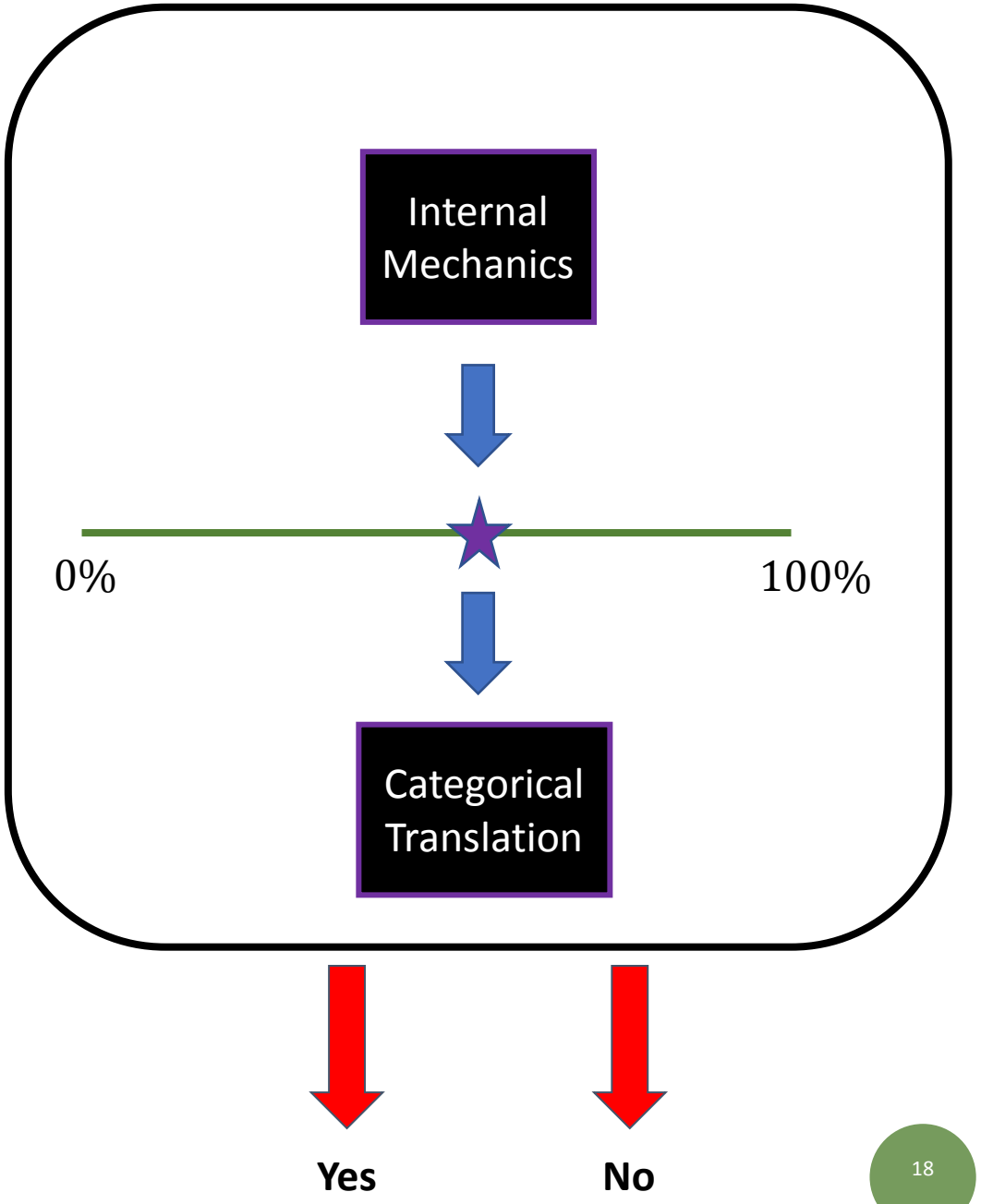
How do we assess the "correctness" of a model?

*Error types matter! Accuracy alone can paint with too broad a brush.*

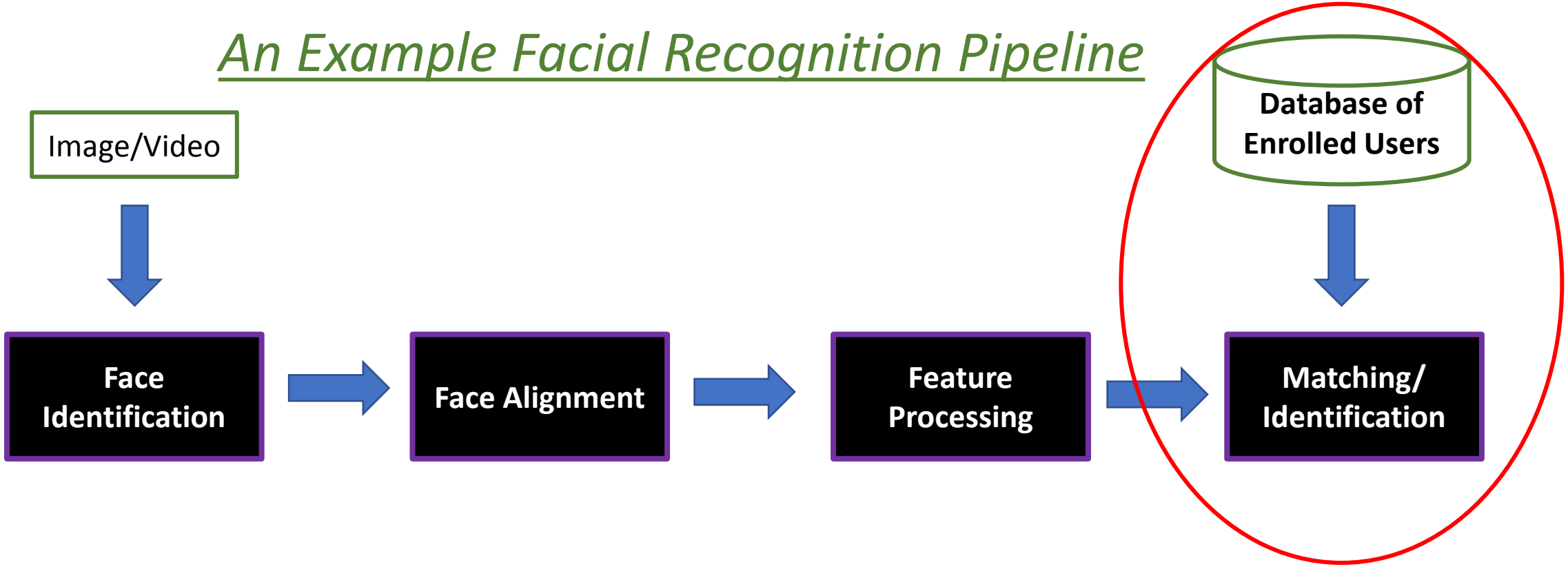
| | | Predicted: 1 | Predicted: 0 |
|---|---|---|---|
| Actual | 1 | TP | FN |
| | 0 | FP | TN |

# *The Role of Thresholds*

For certain technologies, thresholds are needed to make categorical decisions

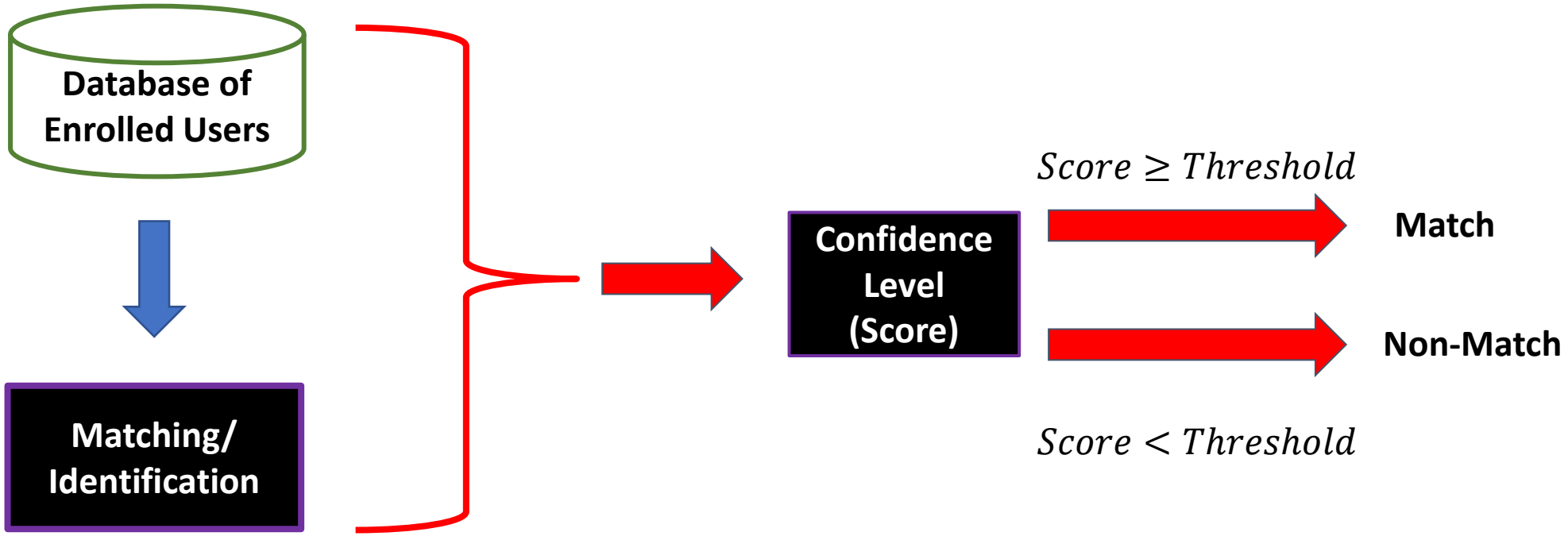


# An Example Facial Recognition Pipeline




## The Role of Thresholds

For certain technologies, thresholds are needed to make categorical decisions




## *The Role of Thresholds*

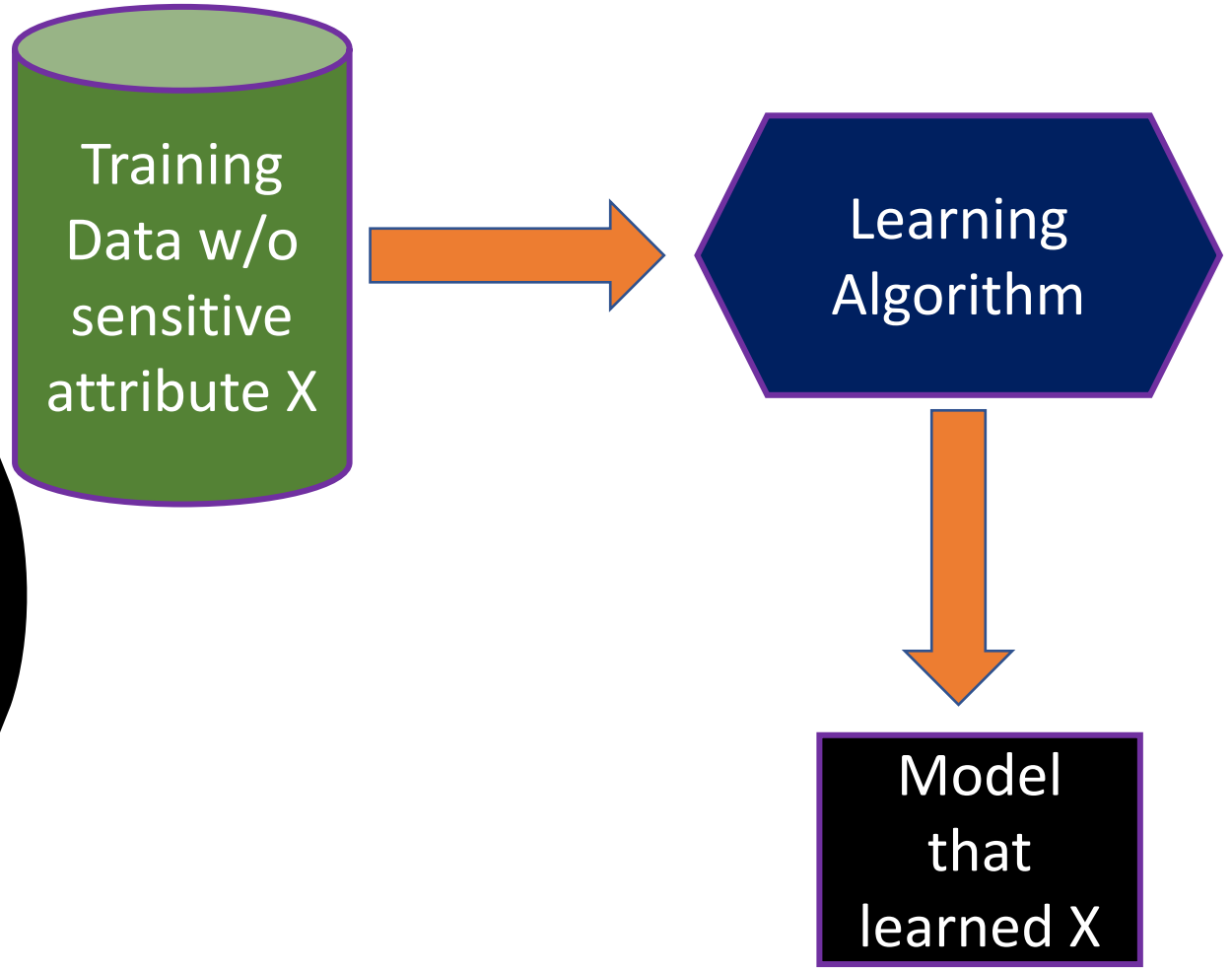
For certain technologies, thresholds are needed to make categorical decisions

*These thresholds are policy decisions that tradeoff error types – they do not make the tech any smarter or dumber.*

# *Discrimination Without Explicit Programming*

The fallacy of "fairness through unawareness"

*Ignoring a sensitive attribute does not guarantee a model won't learn it through correlated features.*



## *Moral of the Story*

Machine learning systems are fragile representations of the world they model