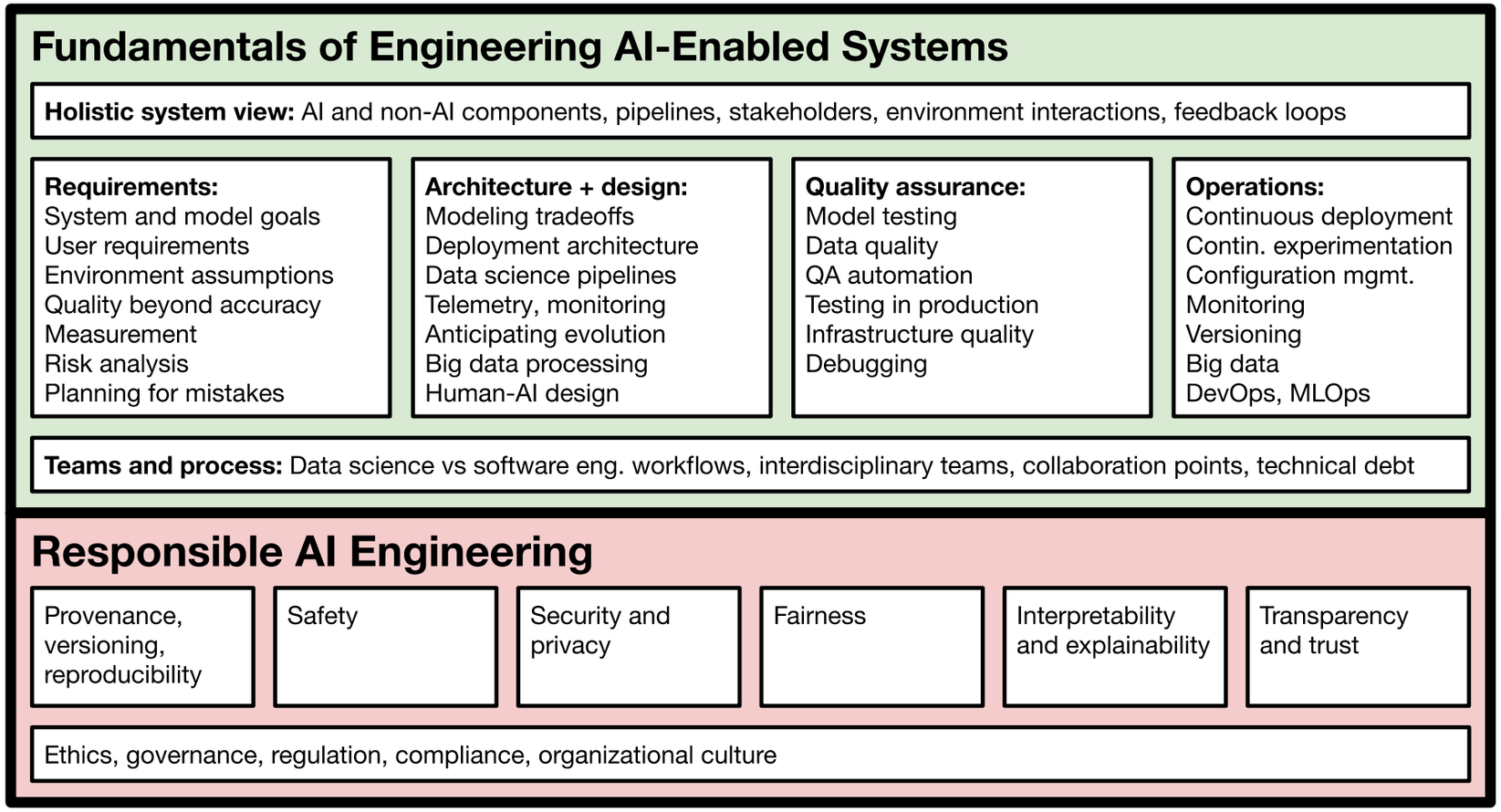

Machine Learning in Production
Explainability and Interpretability

Explainability as Building Block in Responsible Engineering



"Readings"

Required one of:

- 🎧 Data Skeptic Podcast Episode “Black Boxes are not Required” with Cynthia Rudin (32min)
- Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215.

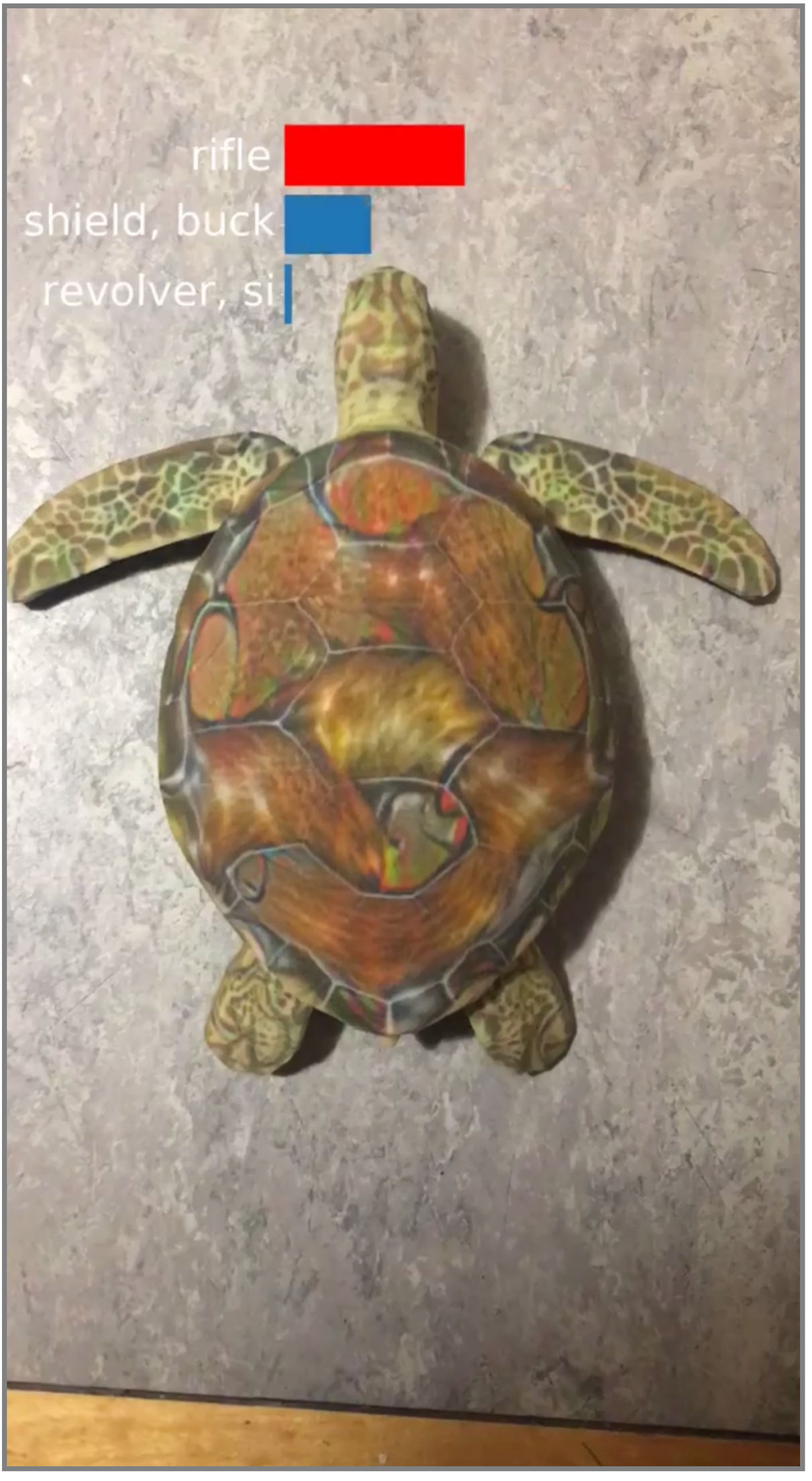
Recommended supplementary reading:

- Christoph Molnar. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2019

Learning Goals

- Understand the importance of and use cases for interpretability
- Explain the tradeoffs between inherently interpretable models and post-hoc explanations
- Measure interpretability of a model
- Select and apply techniques to debug/provide explanations for data, models and model predictions
- Eventuate when to use interpretable models rather than ex-post explanations

Motivating Examples


rifle
shield, buck
revolver, si

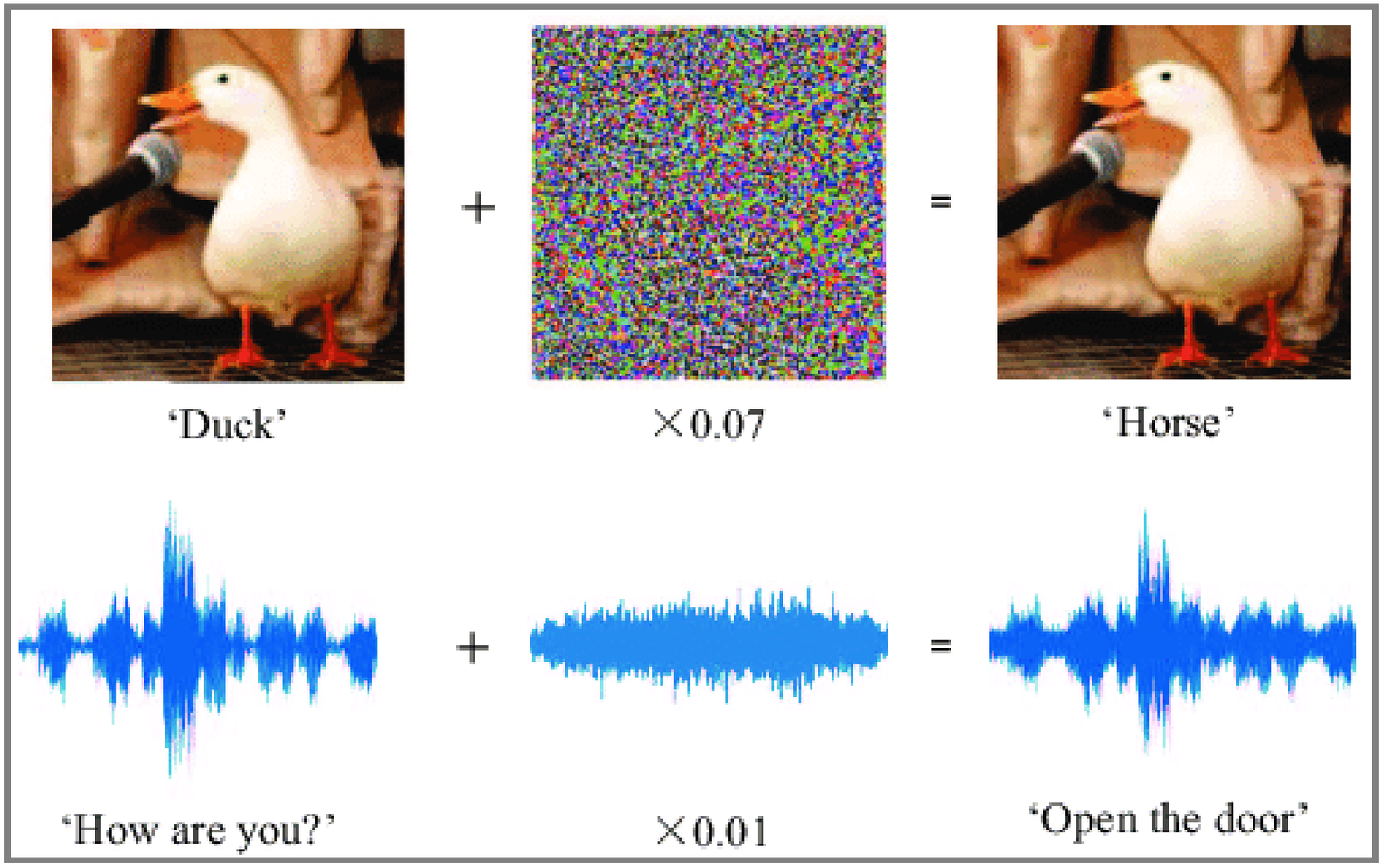


Image: Gong, Yuan, and Christian Poellabauer. "An overview of vulnerabilities of voice controlled systems." arXiv preprint arXiv:1803.09156 (2018).

Is this recidivism model fair?

```
IF age between 18–20 and sex is male THEN
  predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN
  predict arrest
ELSE IF more than three priors THEN
  predict arrest
ELSE
  predict no arrest
```

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215.

How to consider seriousness of the crime?

1.	Age at Release between 18 to 24	2 points		$\cdots$
2.	Prior Arrests ≥ 5	2 points	$+$	$\cdots$
3.	Prior Arrest for Misdemeanor	1 point	$+$	$\cdots$
4.	No Prior Arrests	-1 point	$+$	$\cdots$
5.	Age at Release ≥ 40	-1 point	$+$	$\cdots$
		SCORE	$=$	$\cdots$

PREDICT ARREST FOR ANY OFFENSE IF SCORE > 1

1.	Prior Arrests ≥ 2	1 point		$\cdots$
2.	Prior Arrests ≥ 5	1 point	$+$	$\cdots$
3.	Prior Arrests for Local Ordinance	1 point	$+$	$\cdots$
4.	Age at Release between 18 to 24	1 point	$+$	$\cdots$
5.	Age at Release ≥ 40	-1 points	$+$	$\cdots$
		SCORE	$=$	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Rudin, Cynthia, and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." Interfaces 48, no. 5 (2018): 449-466.

Is there an actual problem? How to find out?

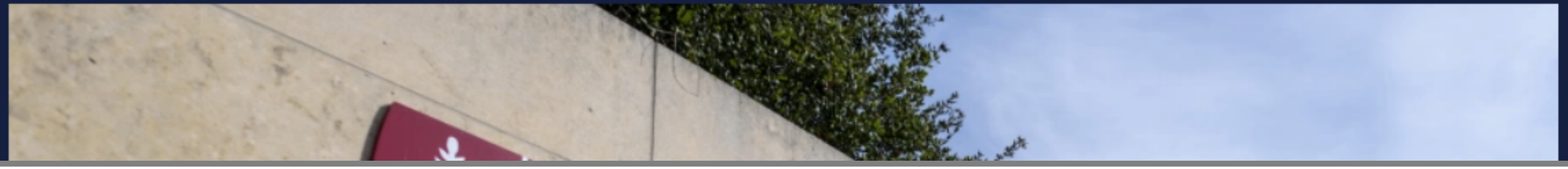
PANDEMIC TECHNOLOGY PROJECT

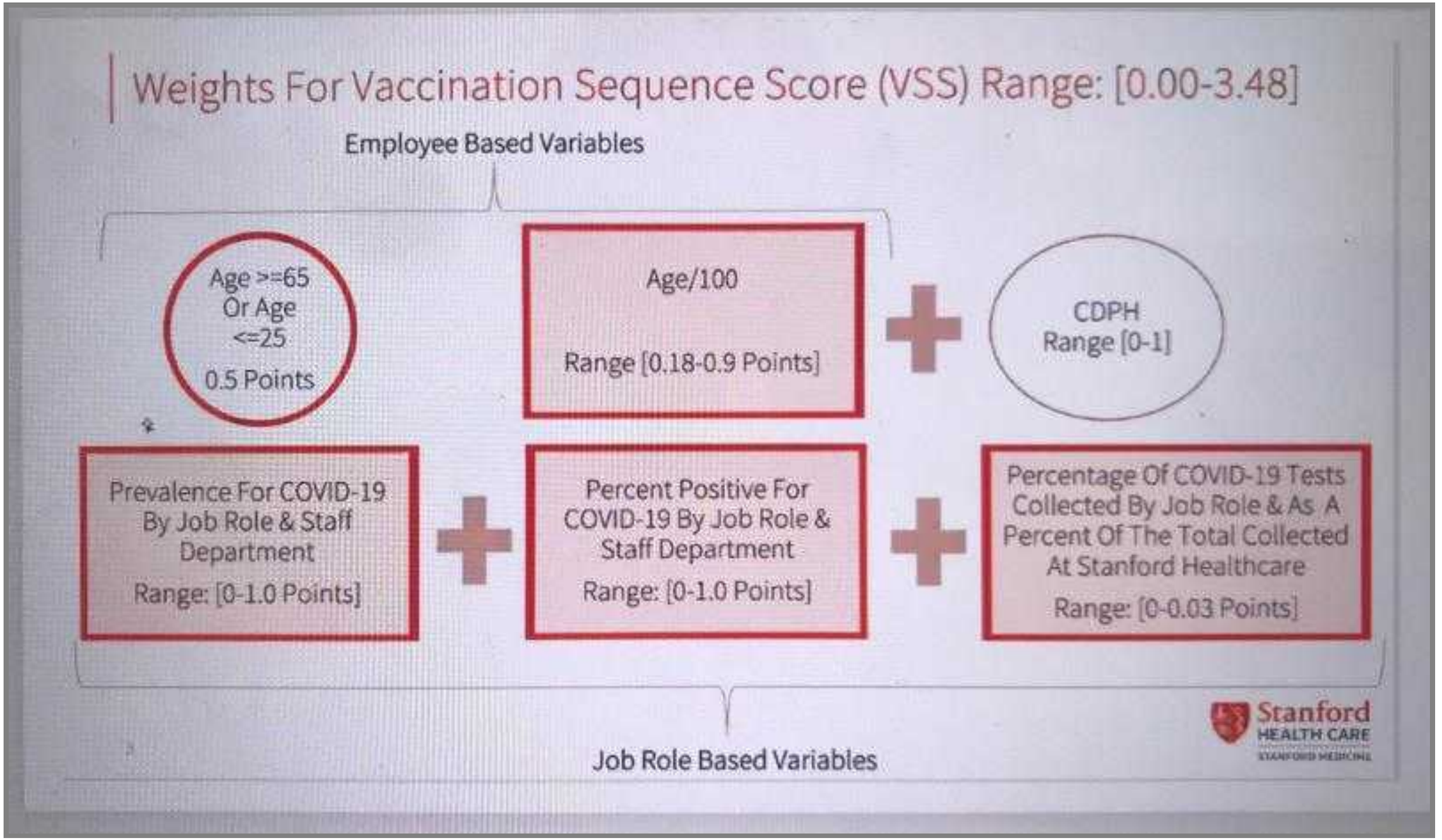
This is the Stanford vaccine algorithm that left out frontline doctors

The university hospital blamed a "very complex algorithm" for its unequal vaccine distribution plan. Here's what went wrong.

By Eileen Guo & Karen Hao

December 21, 2020

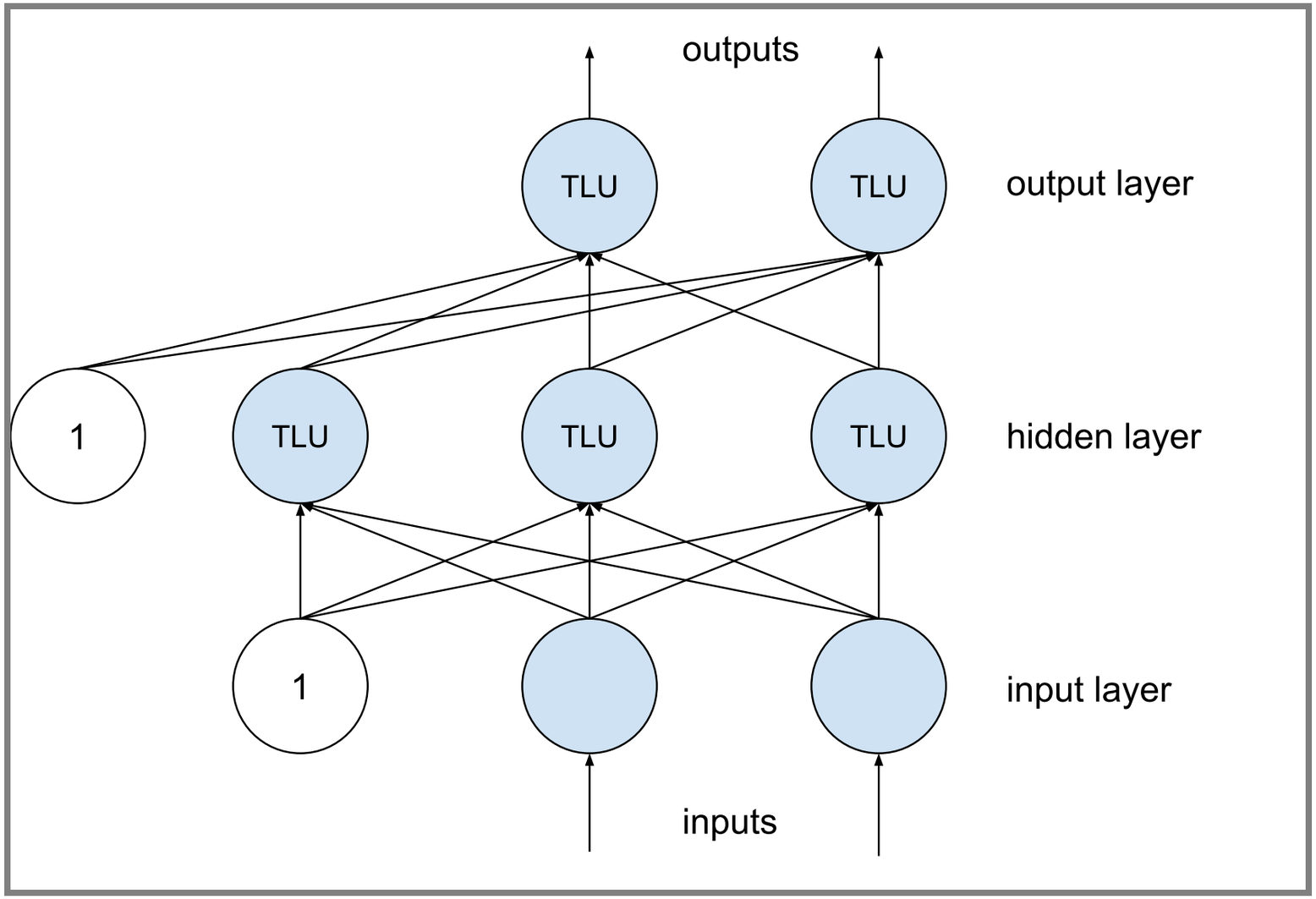



Weights For Vaccination Sequence Score (VSS) Range: [0.00-3.48]
Employee Based Variables
Age >=65 Or Age <=25
0.5 Points
Age/100
Range [0.18-0.9 Points]
CDPH
Range [0-1]
Prevalence For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]
Percent Positive For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]
Percentage Of COVID-19 Tests Collected By Job Role & As A Percent Of The Total Collected At Stanford Healthcare
Range: [0-0.03 Points]
Job Role Based Variables
Stanford
HEALTH CARE

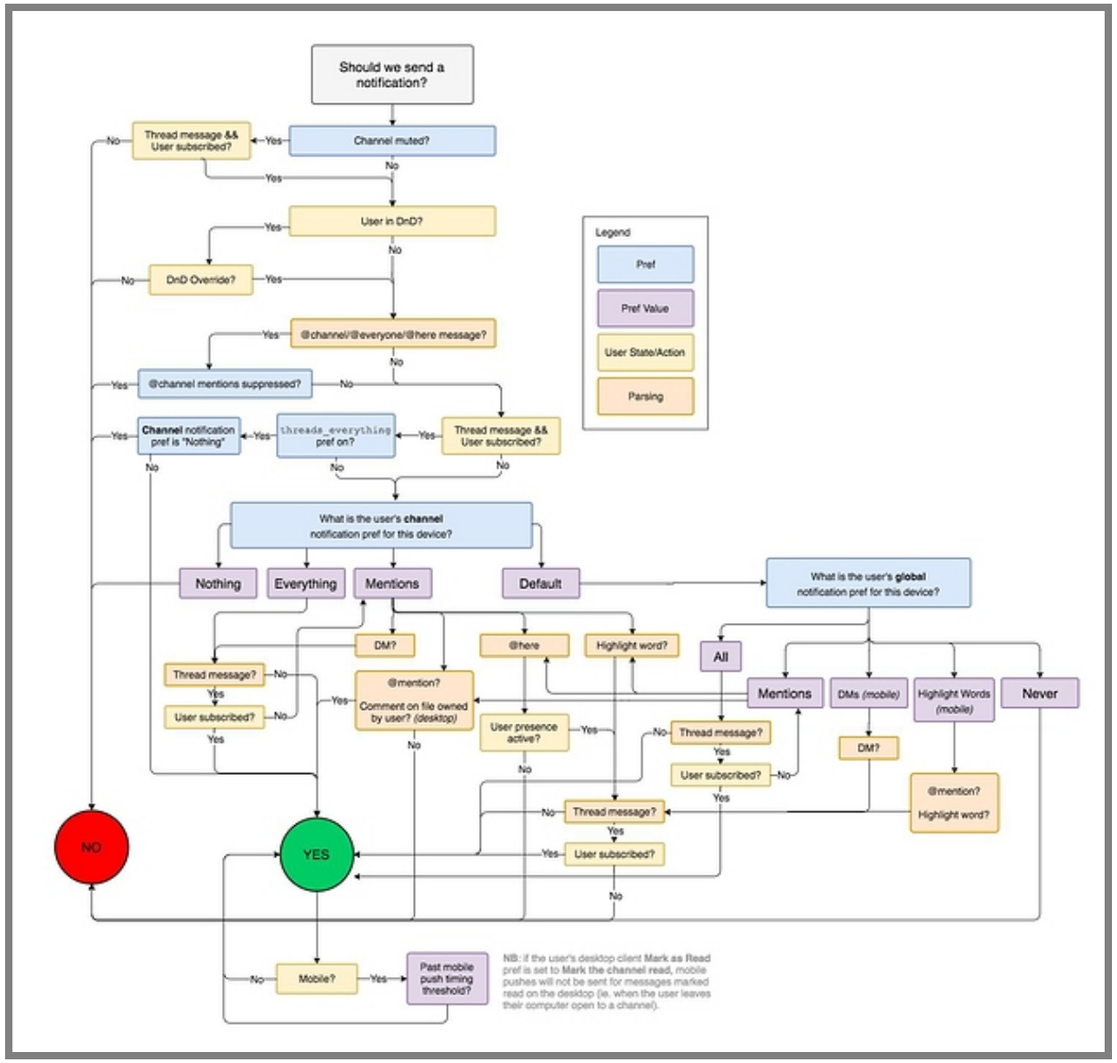
Explaining Decisions

Cat? Dog? Lion? -- Confidence? Why?

What's happening here?



Explaining Decisions



Explainability in ML

Explain how the model made a decision

- Rules, cutoffs, reasoning?
- What are the relevant factors?
- Why those rules/cutoffs?

Challenging because models too complex and based on data

- Can we understand the rules?
- Can we understand why these rules?

Why Explainability?

Debugging

- Why did the system make a wrong prediction in this case?
- What does it actually learn?
- What data makes it better?
- How reliable/robust is it?
- How much does second model rely on outputs of first?
- Understanding edge cases

Debugging is the most common use in practice (Bhatt et al. "Explainable machine learning in deployment." In Proc. FAccT. 2020.)

Auditing

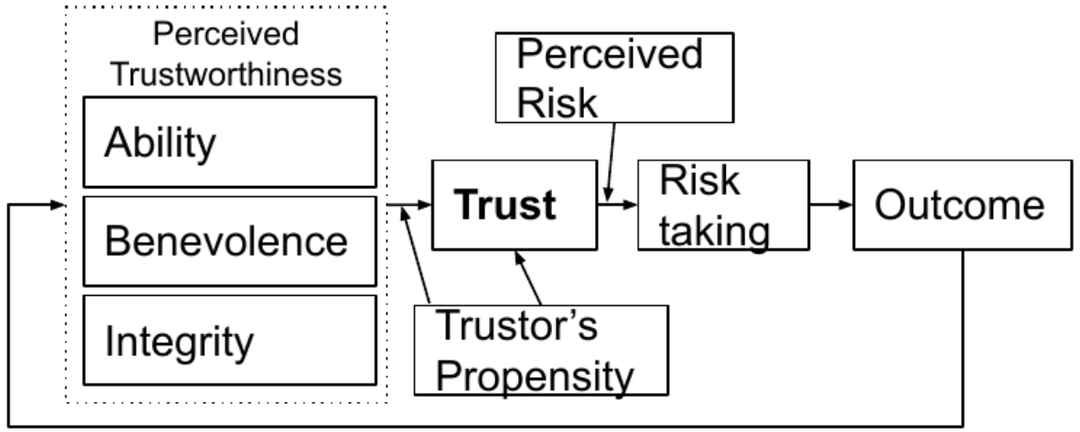
- Understand safety implications
- Ensure predictions use objective criteria and reasonable rules
- Inspect fairness properties
- Reason about biases and feedback loops
- Validate "learned specifications/requirements" with stakeholders

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Trust

More accepting a prediction if clear how it is made, e.g.,

- Model reasoning matches intuition; reasoning meets fairness criteria
- Features are difficult to manipulate
- Confidence that the model generalizes beyond target distribution



Conceptual model of trust: R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. Academy of Management Review, 20(3):709–734, July 1995.

Actionable Insights to Improve Outcomes

"What can I do to get the loan?"

"How can I change my message to get more attention on Twitter?"

"Why is my message considered as spam?"

Regulation / Legal Requirements

*The EU General Data Protection Regulation extends the automated decision-making rights [...] to provide a legally disputed form of a **right to an explanation**: "[the data subject should have] the right ... to obtain an explanation of the decision reached"*

US Equal Credit Opportunity Act requires to notify applicants of action taken with specific reasons: "The statement of reasons for adverse action required by paragraph (a)(2)(i) of this section must be specific and indicate the principal reason(s) for the adverse action."

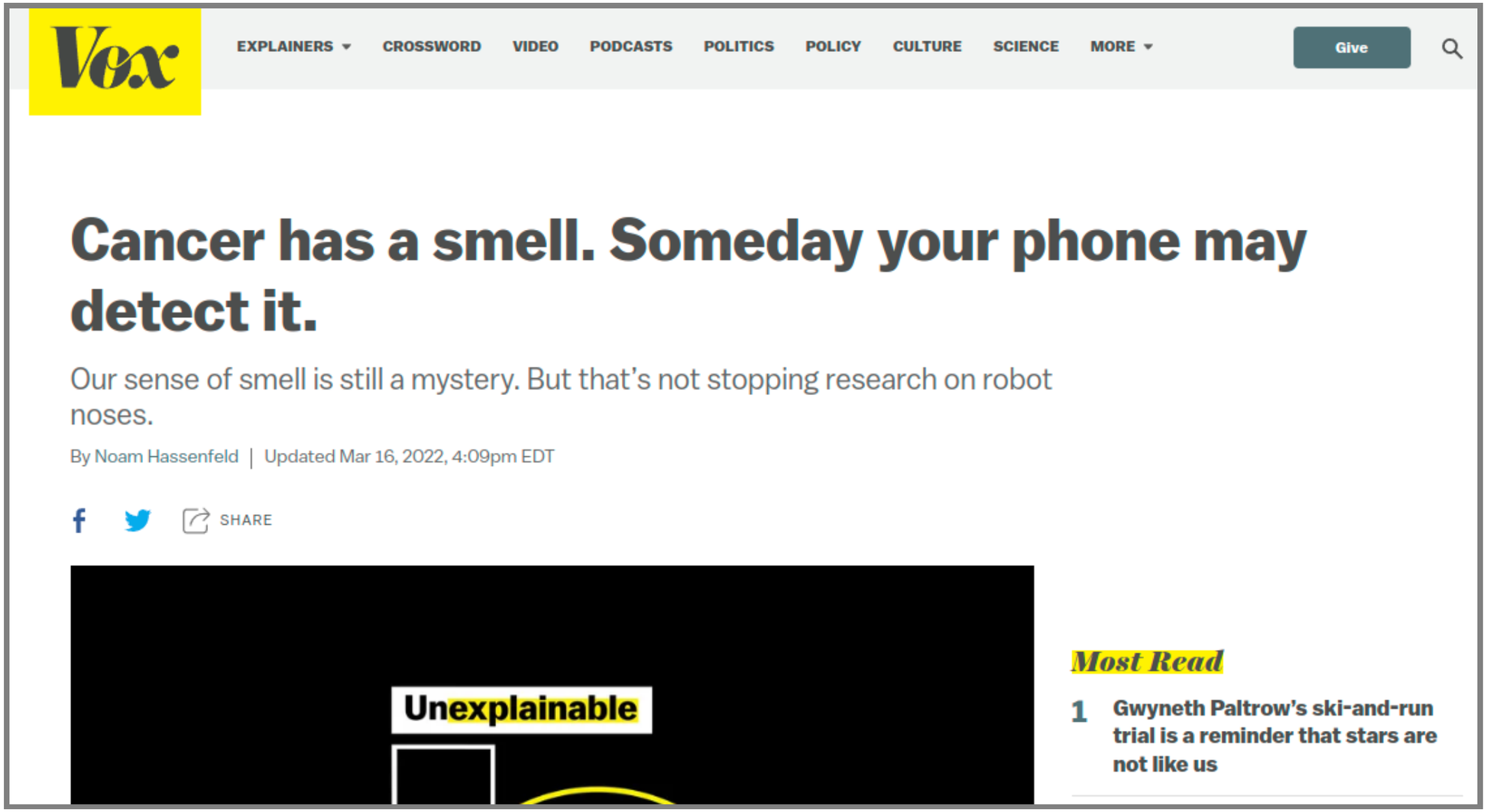
See also https://en.wikipedia.org/wiki/Right_to_explanation

Curiosity, learning, discovery, science

	Basic Model response: *freshness = 0* 17.3% deviance explained		**Full Model** response: *freshness = 0* 17.4% deviance explained		**RDD** response: log(*freshness*) $R^2_m = 0.04$, $R^2_c = 0.35$	
	Coeffs (Err.)	LR Chisq	Coeffs (Err.)	LR Chisq	Coeffs (Err.)	Sum sq.
(Interc.)	3.54 (0.03)***		3.50 (0.03)***		1.45 (0.09)***	
Dep.	−1.78 (0.01)***	32077.8***	−1.79 (0.01)***	32292.8***	−0.04 (0.02)	3.01
RDep.	0.22 (0.01)***	610.3***	0.21 (0.01)***	560.6***	−0.01 (0.02)	0.11
Stars	−0.08 (0.00)***	301.4***	−0.09 (0.00)***	311.2***	0.00 (0.01)	0.00
Contr.	−0.24 (0.01)***	500.5***	−0.25 (0.01)***	548.7***	−0.04 (0.02)*	4.39*
lastU	−0.65 (0.01)***	12080.9***	−0.64 (0.01)***	11537.9***	0.01 (0.02)	0.37
hasDM			0.24 (0.03)***	116.1***	0.45 (0.08)***	2.43
hasInf			0.11 (0.02)***	48.3***	0.04 (0.05)	0.45
hasDM:hasInf			−0.05 (0.04)	1.9	−0.32 (0.10)**	
hasOther			0.01 (0.01)			
time					0.03 (0.00)***	82.99***
intervention					−0.93 (0.03)***	1373.22***
time_after_intervention					0.11 (0.00)***	455.56***
time_after_intervention:hasDM					−0.10 (0.01)***	230.36***
time_after_intervention:hasInf					−0.00 (0.01)	1.14
time_after_intervention:hasDM:hasInf					0.03 (0.01)**	10.62**

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$;
Dep: dependencies; RDep: dependents; Contr.: contributors; lastU: time since last update; hasDM: has dependency-manager badge; hasInf: has information badge; hasOther: adopts additional badges within 15 days

Curiosity, learning, discovery, science



Settings where Interpretability is *not* Important?

Speaker notes

- Model has no significant impact (e.g., exploration, hobby)
- Problem is well studied? e.g optical character recognition
- Security by obscurity? -- avoid gaming

Exercise: Debugging a Model

Consider the following debugging challenges. In groups discuss how you would debug the problem. In 3 min report back to the class.

Algorithm bad at recognizing some signs in some conditions:

Graduate appl. system seems to rank applicants from HBCUs low:

Defining Interpretability

Interpretability Definitions

Two common approaches:

> *Interpretability is the degree to which a human can understand the cause of a decision*

> *Interpretability is the degree to which a human can consistently predict the model's result.*

(No mathematical definition)

How would you measure interpretability?

Explanation

Understanding a single prediction for a given input

> *Your loan application has been declined. If your savings account had had more than $100 your loan application would be accepted.*

Answer **why** questions based on human psychology, such as

Intrinsic interpretability vs Post-hoc explanation?

Models simple enough to understand (e.g., short decision trees, sparse linear models)

1. *Congestive Heart Failure*	1 point		⋯
2. *Hypertension*	1 point	+	⋯
3. *Age* $\geq$ *75*	1 point	+	⋯
4. *Diabetes Mellitus*	1 point	+	⋯
5. *Prior Stroke or Transient Ischemic Attack*	2 points	+	⋯
ADD POINTS FROM ROWS 1–5	**SCORE**	=	⋯

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Explanation of opaque model, local or global

Your loan application has been declined. If your savings account had more than $100 your loan application would be accepted.

On Terminology

Rudin's terminology and this lecture:

- Interpretable models: Intrinsily interpretable models
- Explainability: Post-hoc explanations

And in general:

- Interpretability: property of a model
- Explainability: ability to explain the workings/predictions of a model
- Explanation: justification of a single prediction
- Transparency: The user is aware that a model is used / how it works

Understanding a Model

Levels of explanations:

- **Understanding a model**
- Explaining a prediction
- Understanding the data

Inherently Interpretable: Sparse Linear Models

$$f(x) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n$$

- Truthful explanations; easy to understand for humans
- Easy to derive contrastive explanation and feature importance
- Requires feature selection/regularization to minimize to few important features (e.g. Lasso); possibly restricting possible parameter values

Score card: Sparse linear model with "round" coefficients

1.	*Congestive Heart Failure*	1 point		$\cdots$
2.	*Hypertension*	1 point	$+$	$\cdots$
3.	*Age* $\geq$ *75*	1 point	$+$	$\cdots$
4.	*Diabetes Mellitus*	1 point	$+$	$\cdots$
5.	*Prior Stroke or Transient Ischemic Attack*	2 points	$+$	$\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	$=$	$\cdots$

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Inherently Interpretable: Shallow Decision Trees

- Easy to interpret up to a size
- Possible to derive counterfactuals and feature importance
- Unstable with small changes to training data

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Inherently Interpretable: Decision Rules

if-then rules mined from data

easy to interpret if few and simple rules

see association rule mining:

```
{Diaper, Beer} -> Milk  (40% support, 66% confidence)
Milk -> {Diaper, Beer}  (40% support, 50% confidence)
{Diaper, Beer} -> Bread (40% support, 66% confidence)
```

Not all Linear Models and Decision Trees are Inherently Interpretable

- Models can be very big, many parameters (factors, decisions)
- Nonlinear interactions possibly hard to grasp
- Tool support can help (views)
- Random forests, ensembles no longer easily interpretable

```
173554.681081086 * root + 318523.818532818 * heuristicUnit + -103411.8707
-11816.7857142856 * heuristicVmtf + -33557.8961038976 * heuristic + -9537
3990.79729729646 * transExt * satPreproYes + -136928.416666666 * eq * heu
33925.0833333346 * eq * heuristic + -643.428571428088 * backprop * heuris
heuristicUnit + 1620.24242424222 * eq * backprop + -7205.2500000002 * eq
```

Speaker notes

Example of a performance influence model from http://www.fosd.de/SPLConqueror/ -- not the worst in terms of interpretability, but certainly not small or well formated or easy to approach.

Research in Inherently Interpretable Models

Several approaches to learn sparse constrained models (e.g., fit score cards, simple if-then-else rules)

Often heavy emphasis on feature engineering and domain-specificity

Possibly computationally expensive

Post-Hoc Model Explanation: Global Surrogates

1. Select dataset X (previous training set or new dataset from same distribution)
2. Collect model predictions for every value: $y_i = f(x_i)$
3. Train *inherently interpretable* model on (X,Y)
4. Interpret surrogate model

Can measure how well fits with common model quality measures, typically

Advantages? Disadvantages?

Speaker notes

Flexible, intuitive, easy approach, easy to compare quality of surrogate model with validation data (R^2). But: Insights not based on real model; unclear how well a good surrogate model needs to fit the original model; surrogate may not be equally good for all subsets of the data; illusion of interpretability. Why not use surrogate model to begin with?

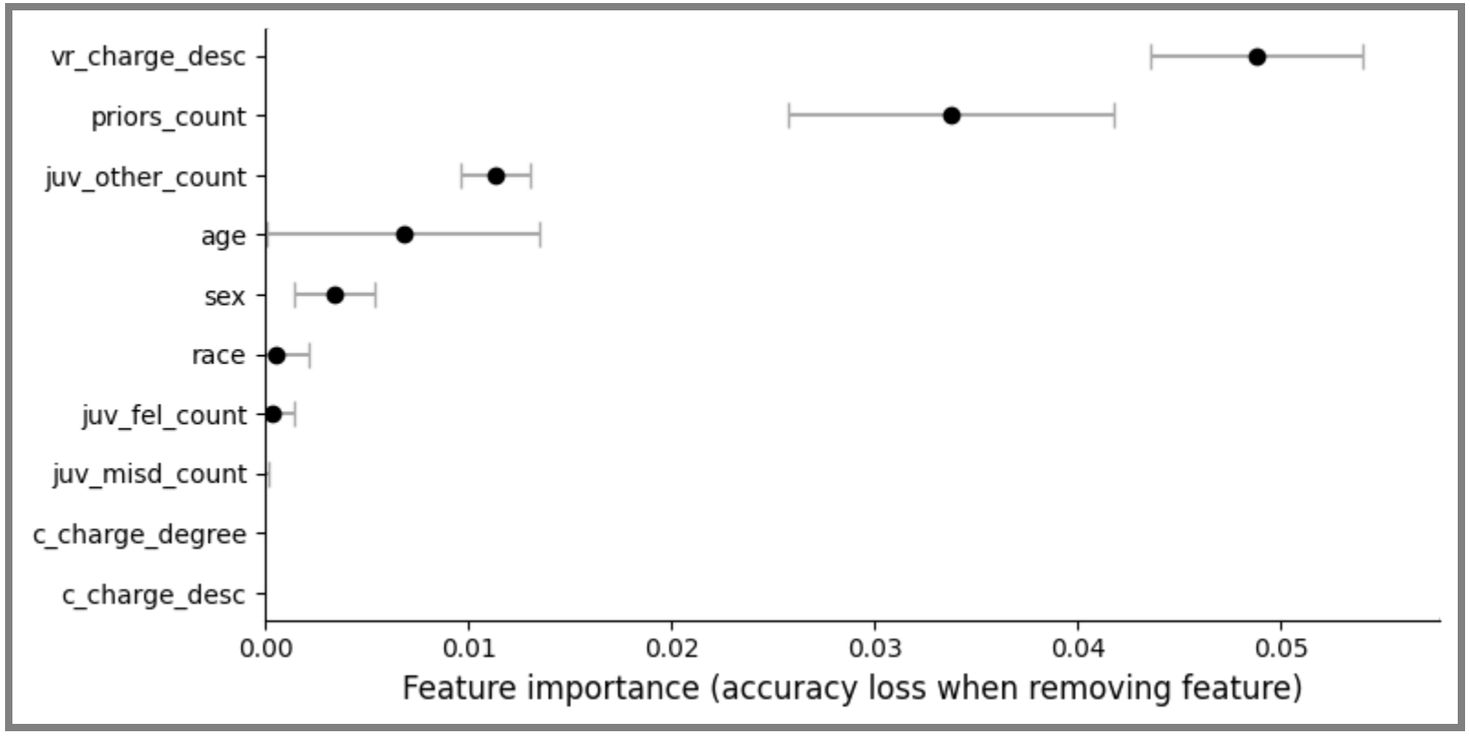
Advantages and Disadvantages of Surrogates?

Advantages and Disadvantages of Surrogates?

- short, contrastive explanations possible
- useful for debugging
- easy to use; works on lots of different problems
- explanations may use different features than original model

- explanation not necessarily truthful
- explanations may be unstable
- likely not sufficient for compliance scenario

Post-Hoc Model Explanation #1: Feature Importance



Feature Importance

- Permute a feature's values in validation data -> hide it for prediction
- Measure influence on accuracy
- -> This evaluates feature's influence without retraining the model

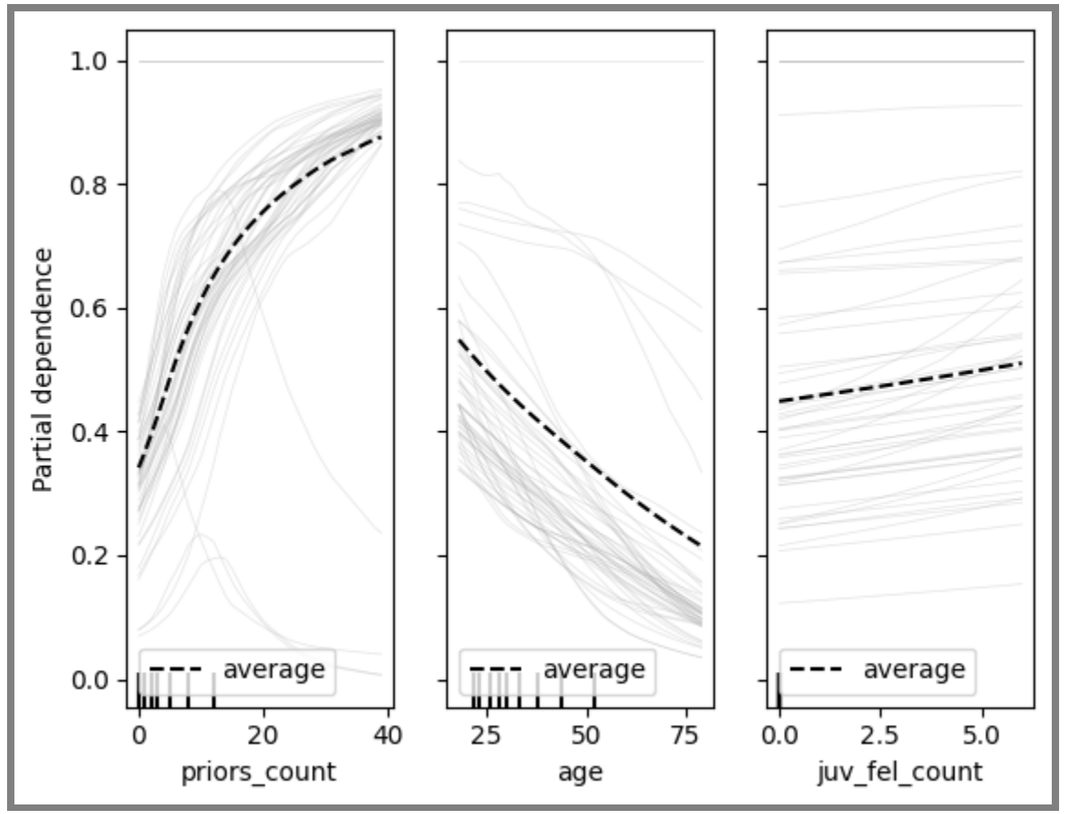
- Highly compressed, *global* insights
- Effect for feature + interactions
- Can only be computed on labeled data, depends on model accuracy, randomness from permutation
- May produce unrealistic inputs when correlations exist

(Can be evaluated both on training and validation data)

Speaker notes

Training vs validation is not an obvious answer and both cases can be made, see Molnar's book. Feature importance on the training data indicates which features the model has learned to use for predictions.

Post-Hoc Model Explanation #2: Partial Dependence Plot (PDP)

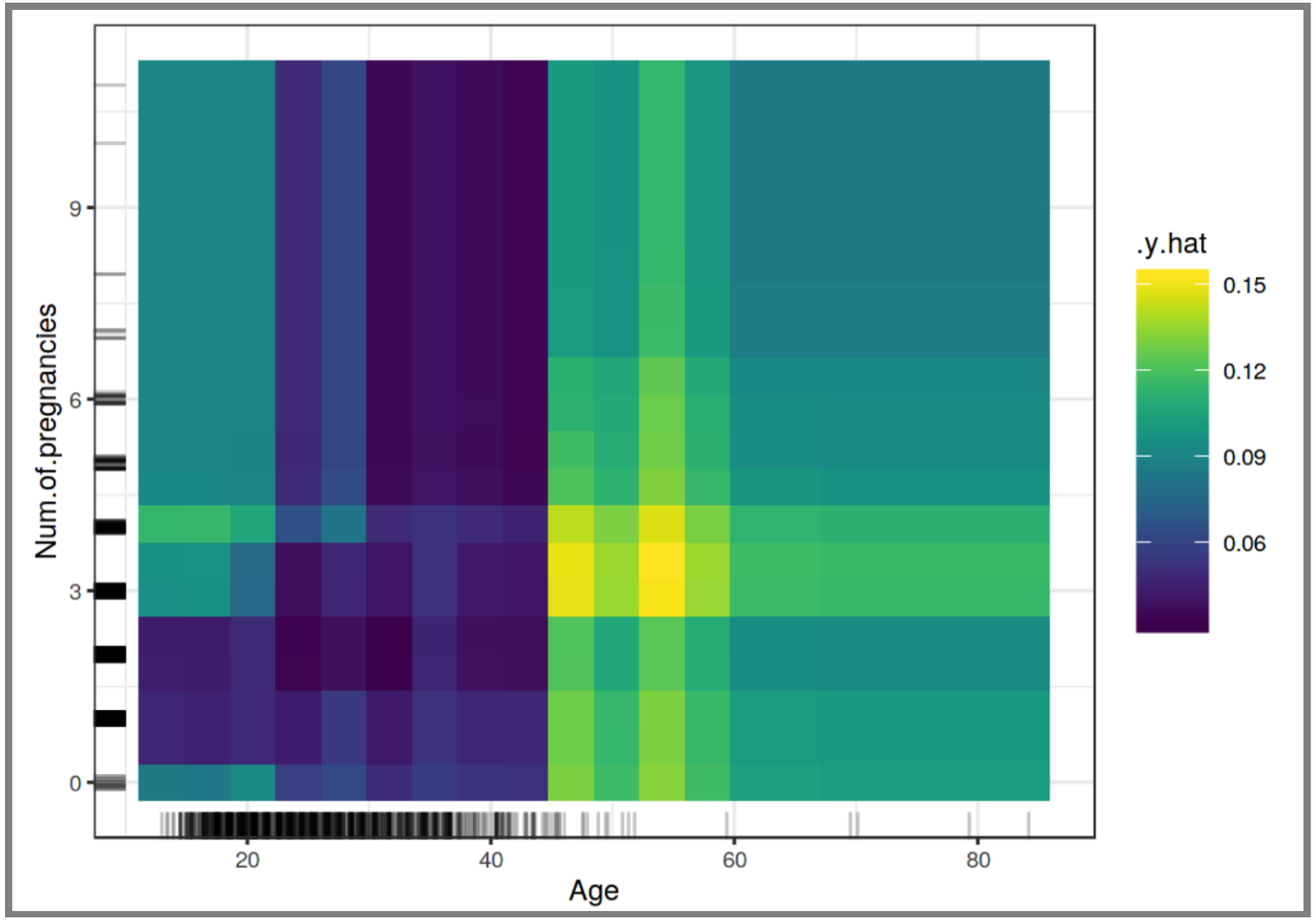


Partial Dependence Plot

- Computes marginal effect of feature on predicted outcome
- Identifies relationship between feature and outcome (linear, monotonous, complex, ...)

- Intuitive, easy interpretation
- Assumes no correlation among features

Partial Dependence Plot for Interactions



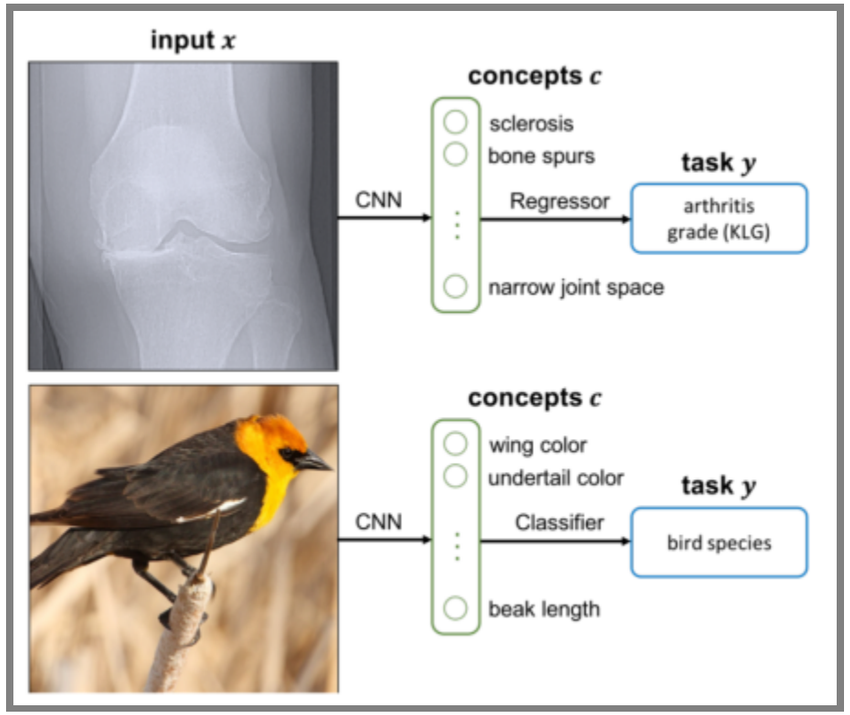
Probability of cancer; source: Christoph Molnar. "Interpretable Machine Learning." 2019

Post-Hoc Model Explanation #3: Concept Bottleneck Models

Hybrid/partially interpretable model

Force models to learn features, not final predictions. Use inherently interpretable model on those features

Requries to label features in training data



Summary: Understanding a Model

Understanding of the whole model, not individual predictions!

Some models inherently interpretable:

- Sparse linear models
- Shallow decision trees

Ex-post explanations for opaque models:

- Global surrogate models
- Feature importance, partial dependence plots
- Many more in the literature

Explaining a Prediction

Levels of explanations:

- Understanding a model
- **Explaining a prediction**
- Understanding the data

Understanding Predictions from Inherently Interpretable Models is easy

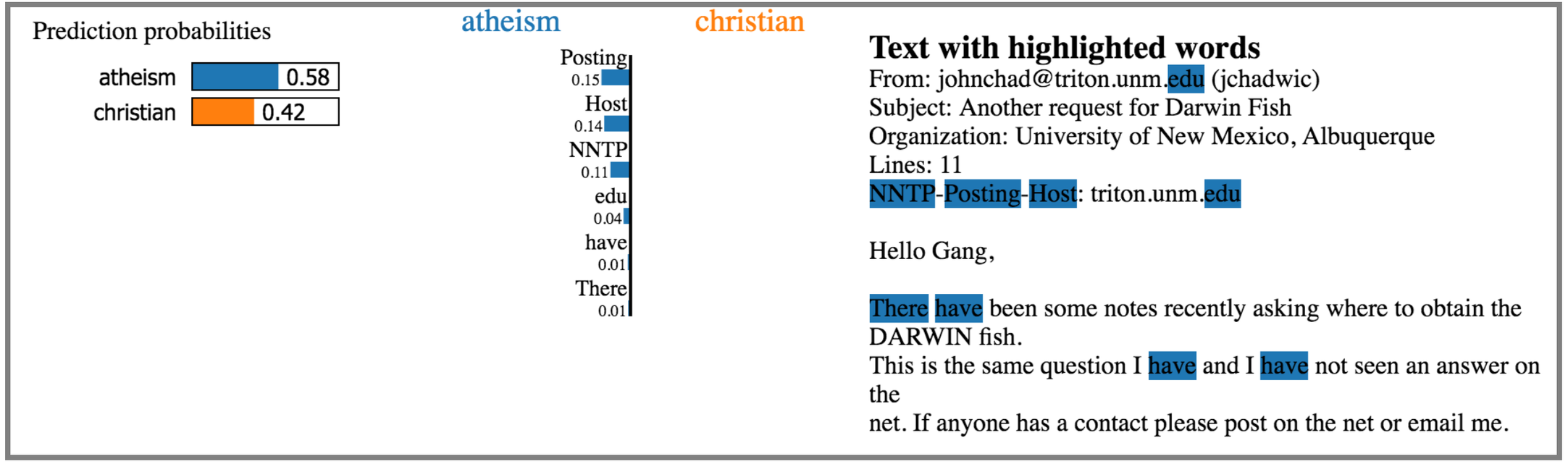
- Derive key influence factors or decisions from model parameters
- Derive contrastive counterfacturals from models

Examples: Predict arrest for 18 year old male with 1 prior:

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Posthoc Prediction Explanation #1: Feature Influences

Which features were most influential for a specific prediction?



Source: https://github.com/marcotcr/lime

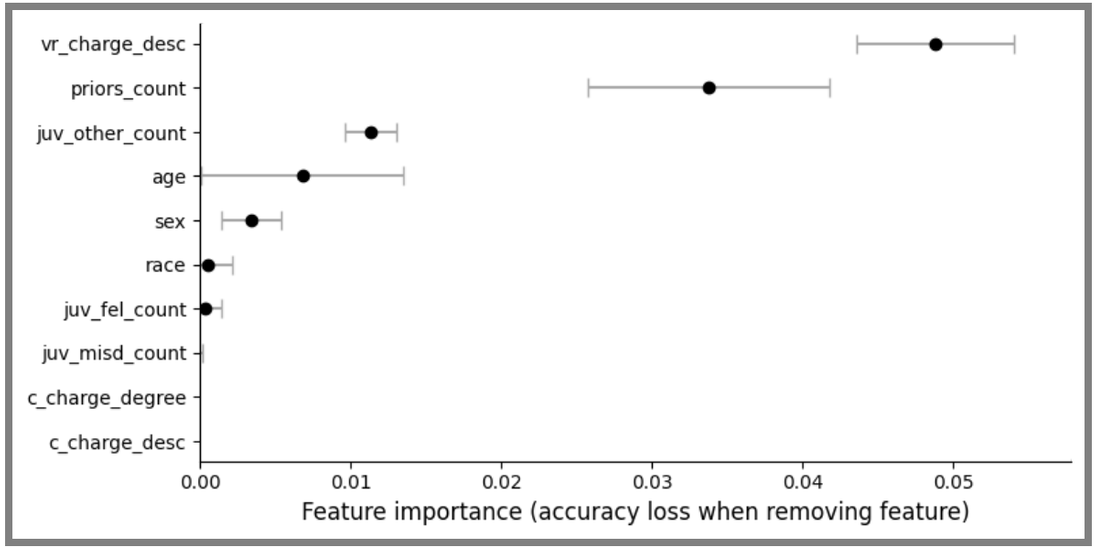
Feature Influences in Images

Source: https://github.com/marcotcr/lime

Feature Importance vs Feature Influence

Feature importance is global for the entire model (all predictions)

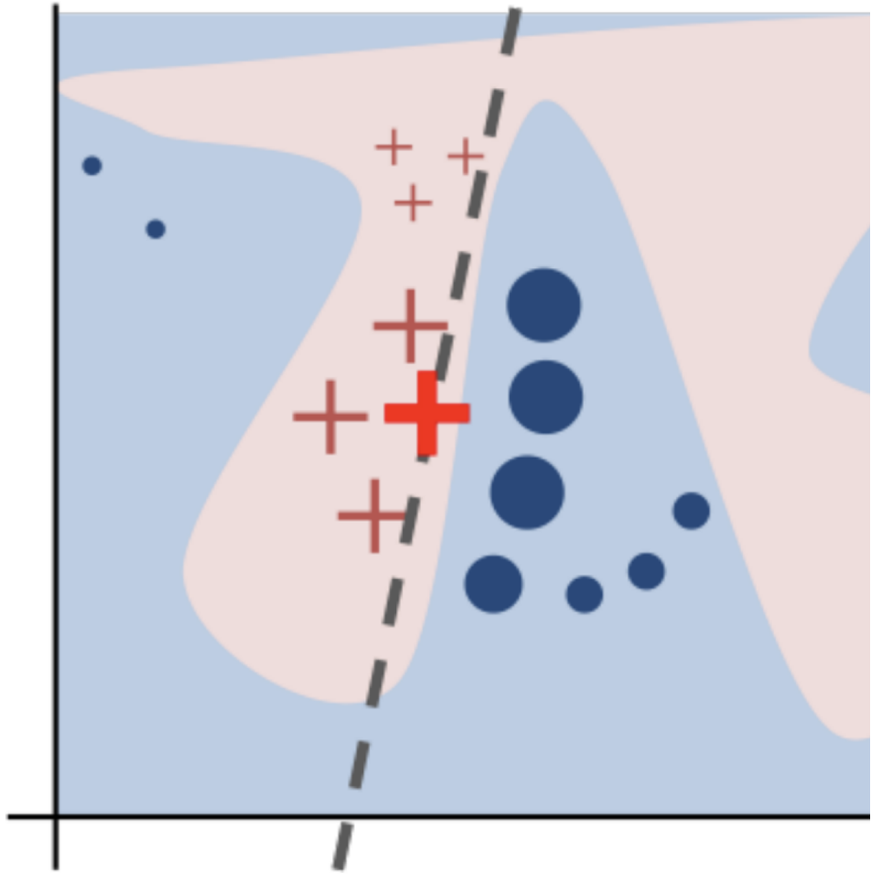
Feature influence is for a single prediction



Feature Infl. with Local Surrogates (LIME)

Create an inherently interpretable model (e.g. sparse linear model) for the area around a prediction

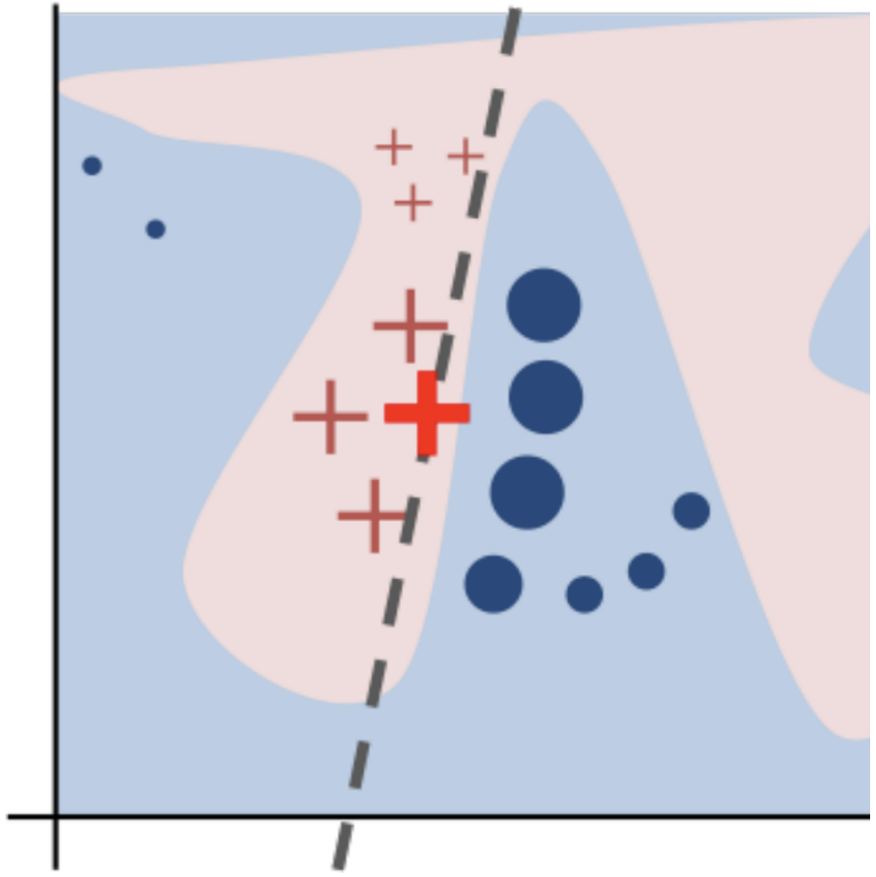
> *“An explanation is a local linear approximation of the model's behaviour. While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance. “*



Feature Infl. with Local Surrogates (LIME)

For *each input data point*...

- Look at model's predictions for a bunch of nearby inputs.
- Closer points are more important than further points.
- Fit a linear model. Its weights are the feature importances.



LIME Example

(a) Husky classified as wolf

(b) Explanation

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier." In Proc. KDD. 2016.

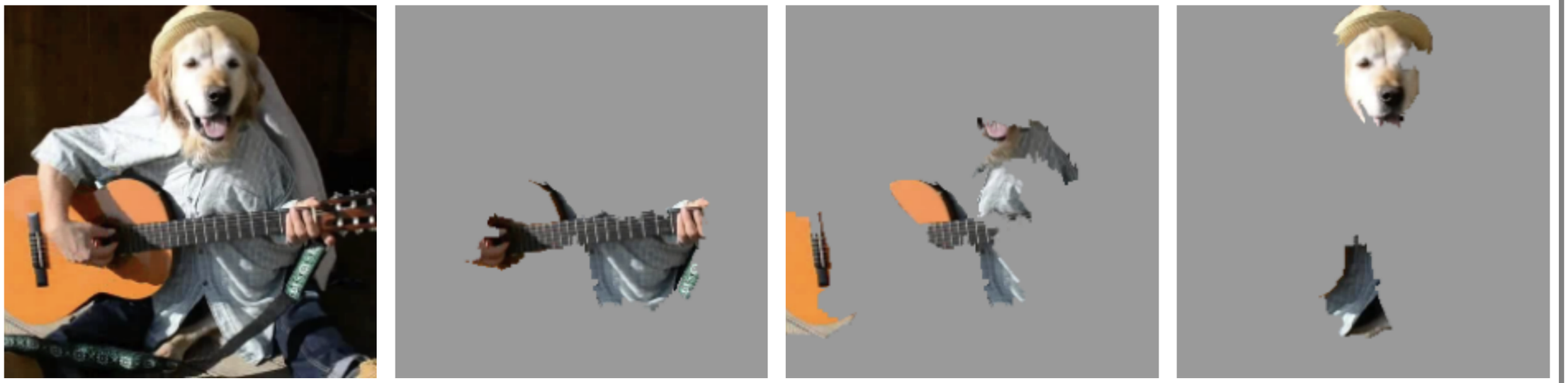
LIME Example

(a) Husky classified as wolf

(b) Explanation

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier." In Proc. KDD. 2016.

LIME Example



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

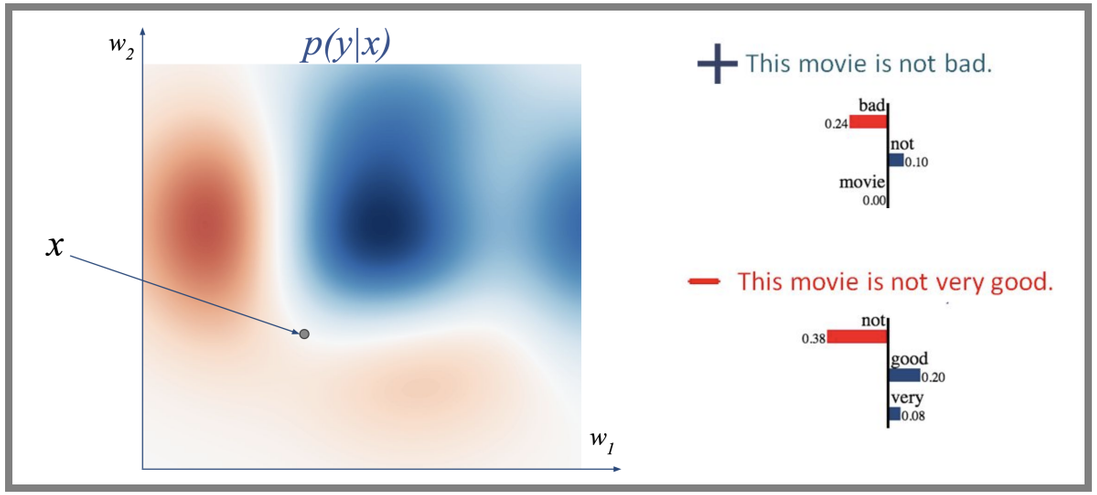
Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier." In Proc. KDD. 2016.

Advantages and Disadvantages of Local Surrogates?

Discussion on of Local Surrogates?

- Quite easy to build!
- Proven useful for debugging

- Unstable, it's training a small model based on points you select
- Too localized and you lose non-linear combinations
- Usually visualized via feature importance & the same heatmap might link to many different underlying explanation mechanisms so could end up being confusing

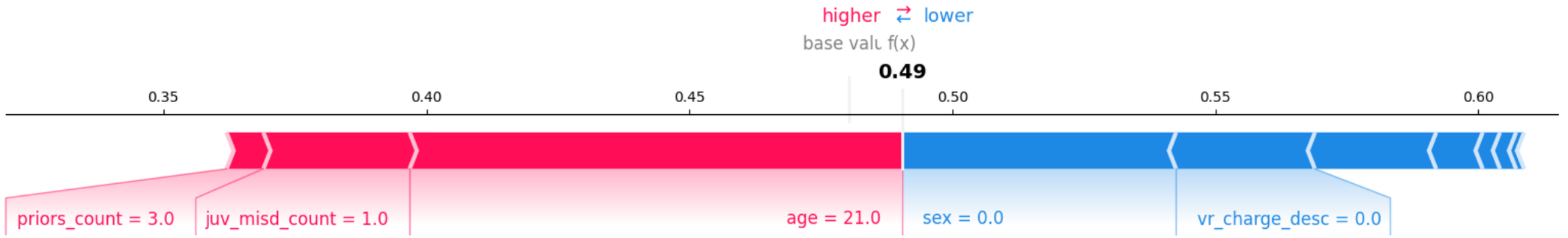


Feature Influence w/ Shapley Values / SHAP

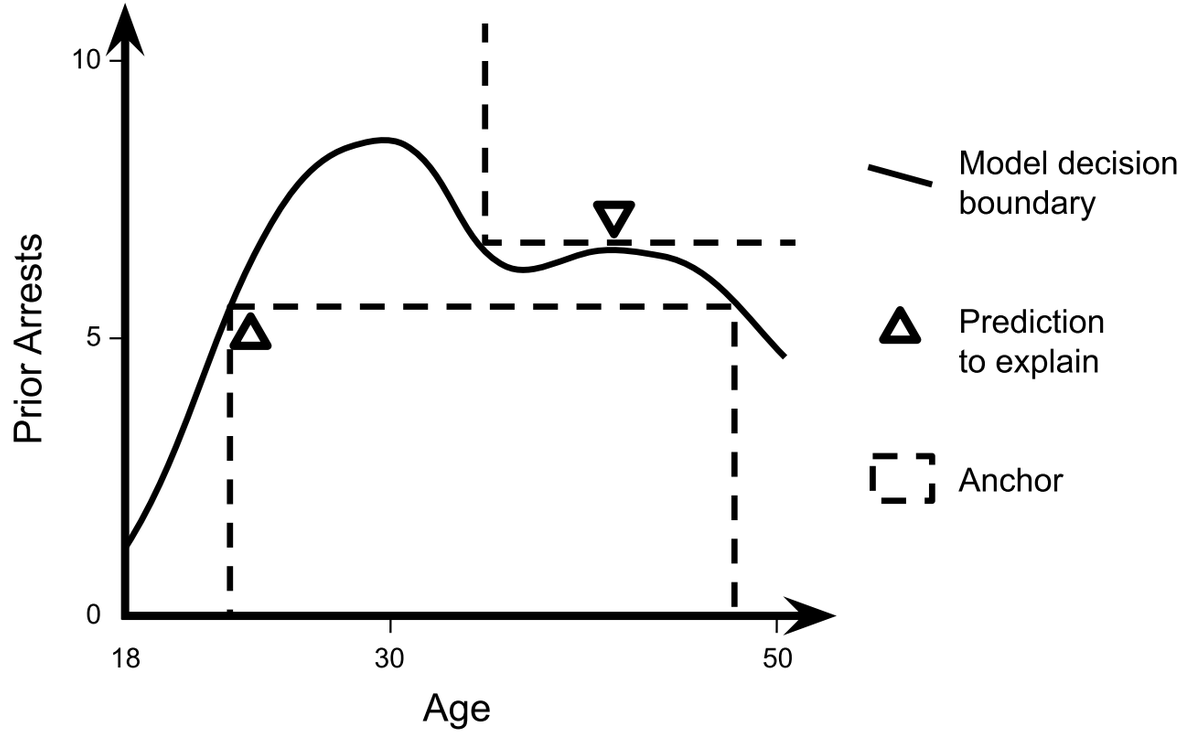
- Game-theoretic foundation for local explanations (1953)
- Explains contribution of feature, over predictions with different feature subsets
  - *"The Shapley value is the average marginal contribution of a feature value across all possible coalitions"*
- Solid theory ensures fair mapping of influence to features
- Requires heavy computation, usually only approximations feasible
- Explanations contain all features (ie. not sparse)

Currently, most common local method used in practice

SHAP Force Plot



Posthoc Prediction Explanation #2: Anchors



Object detected: Steam Locomotive

Posthoc Prediction Explanation #3: Counterfactual Explanations

if X had not occured, Y would not have happened

> *Your loan application has been declined. If your savings account had had more than $100 your loan application would be accepted.*

-> Smallest change to feature values that result in given output

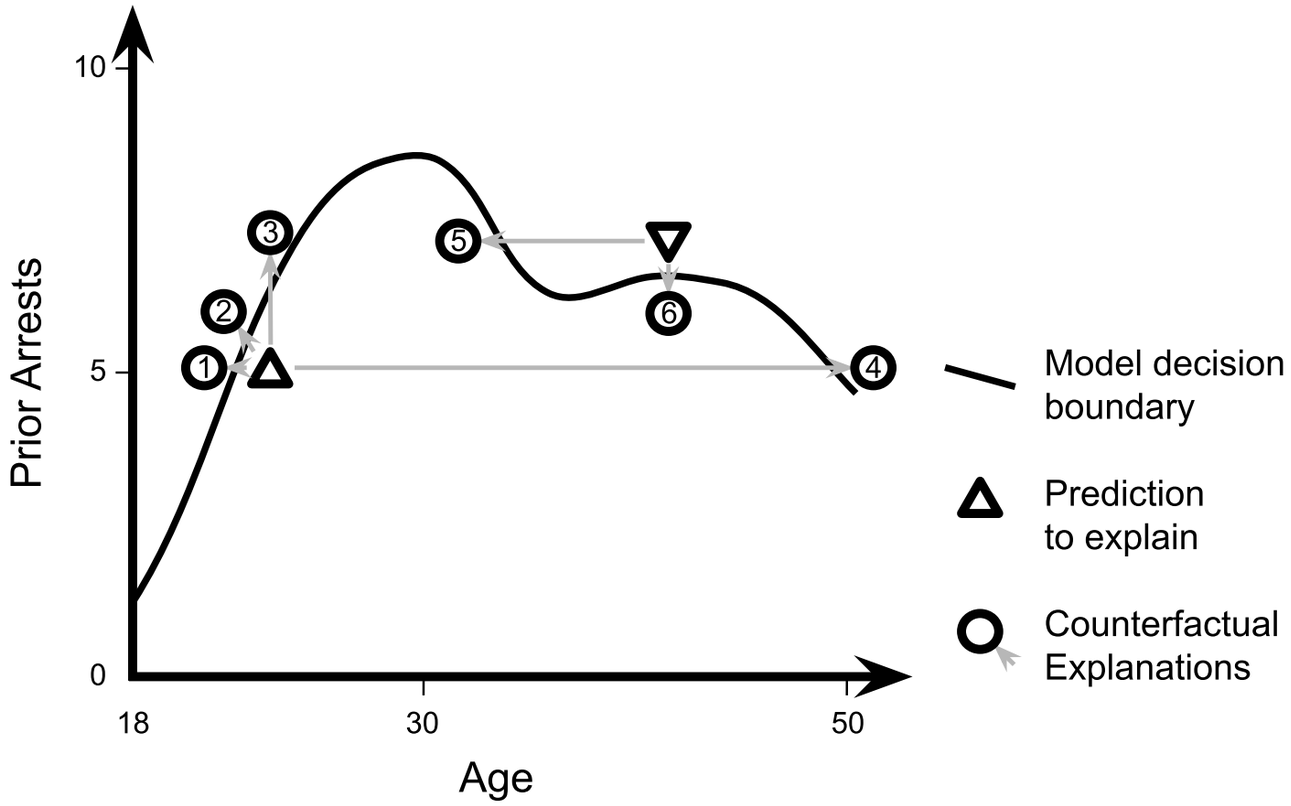
Multiple Counterfactuals

Often long or multiple explanations *(Rashomon effect)*

> *Your loan application has been declined.*
> *If your savings account ...*

> *Your loan application has been declined.*
> *If your lived in ...*

Report all or select "best" (e.g. shortest, most actionable, likely values)



Explanations for △:

1: Predict *arrest* if 3 years younger
2: Predict *arrest* if 2 years younger and one more prior arrest
3: Predict *arrest* if 3 more prior arrests
4: Predict *arrest* if 28 years older

Explanations for ▽:

5: Predict *no arrest* if 10 years younger
6: Predict *no arrest* if 2 fewer prior arr.

Searching for Counterfactuals?

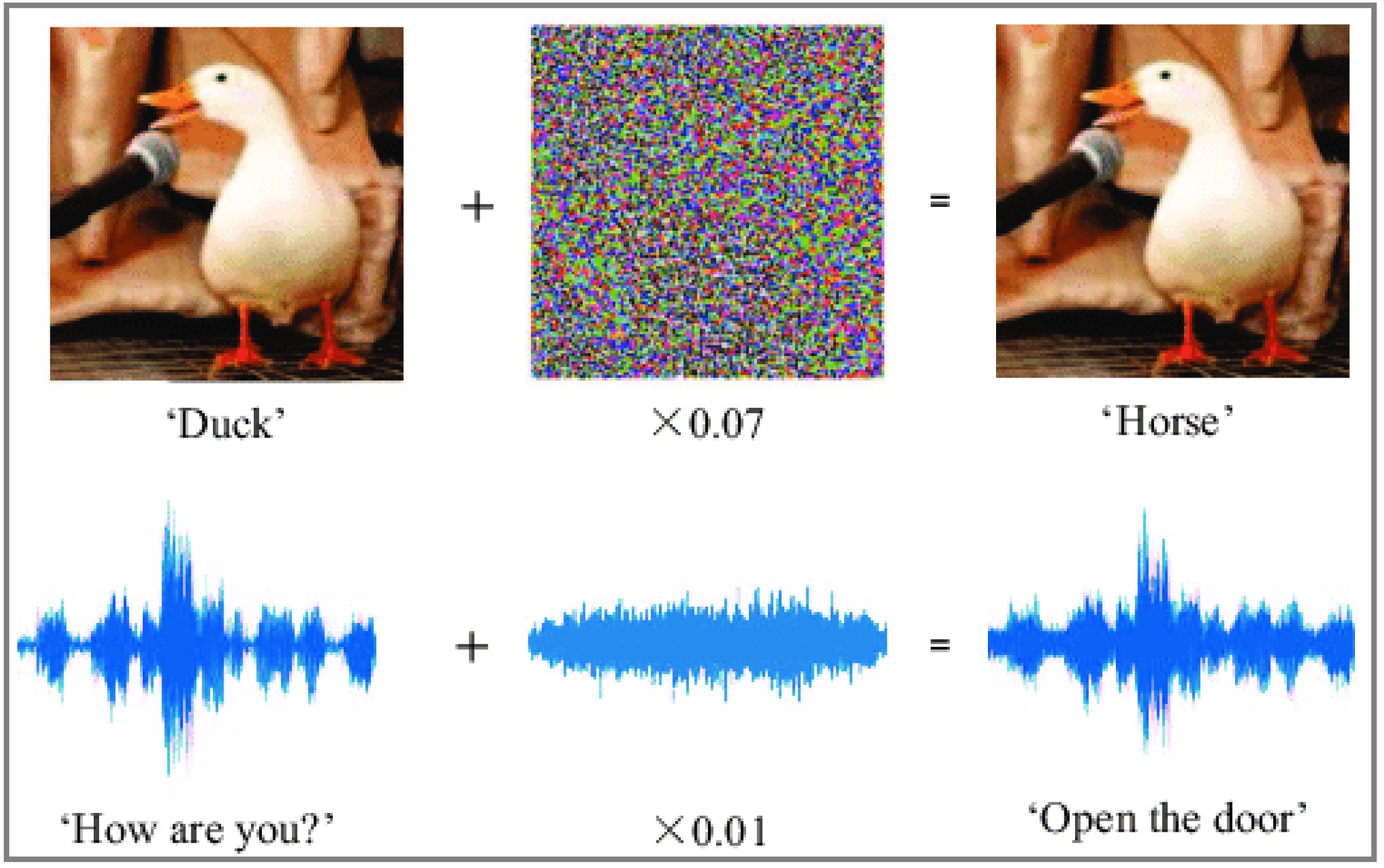
Searching for Counterfactuals

Random search (with growing distance) possible, but inefficient

Many search heuristics, e.g. hill climbing or Nelder–Mead, may use gradient of model if available

Can incorporate distance in loss function

(similar to finding adversarial examples)


‘Duck’
×0.07
‘Horse’
‘How are you?’
×0.01
‘Open the door’

Discussion: Counterfactuals

- Easy interpretation, can report both alternative instance or required change
- No access to model or data required, easy to implement

- Often many possible explanations (Rashomon effect), requires selection/ranking
- May require changes to many features, not all feasible
- May not find counterfactual within given distance
- Large search spaces, especially with high-cardinality categorical features

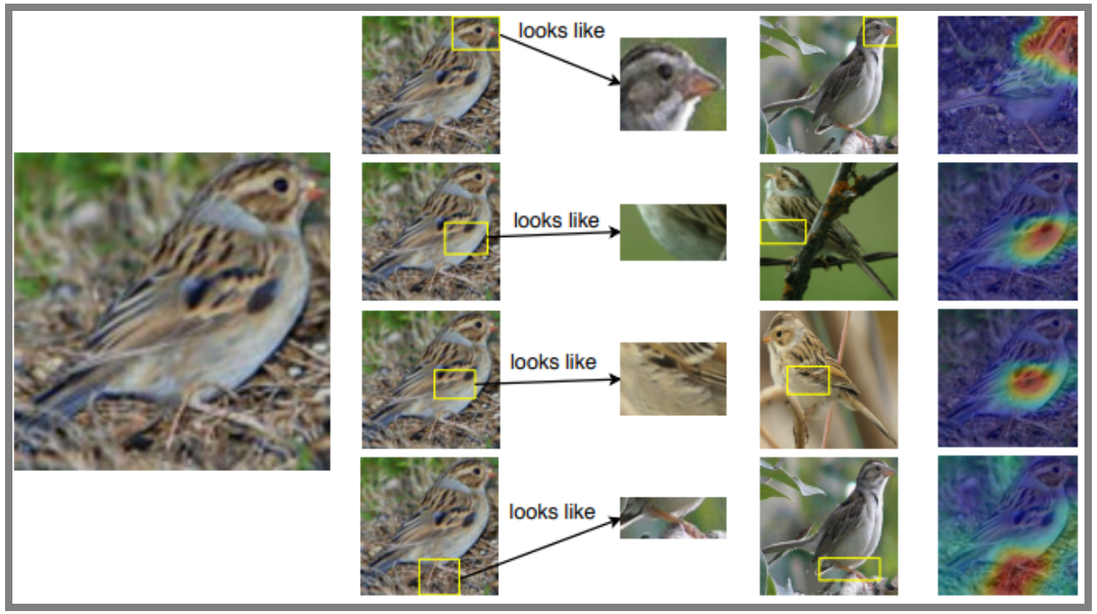
Actionable Counterfactuals

Example: Denied loan application

- Customer wants feedback of how to get the loan approved
- Some suggestions are more actionable than others, e.g.,
  - Easier to change income than gender
  - Cannot change past, but can wait
- In distance function, not all features may be weighted equally

Posthoc Prediction Explanation #4: Similarity

- k-Nearest Neighbors inherently interpretable (assuming intutive distance function)
- Attempts to build inherently interpretable image classification models based on similarity of fragments



Chen, Chaofan, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. "This looks like that: deep learning for interpretable image recognition." In NeurIPS (2019).

Summary: Understanding a Prediction

Understanding a single predictions, not the model as a whole

Explaining influences, providing counterfactuals and sufficient conditions, showing similar instances

Easy on inherently interpretable models

Ex-post explanations for opaque models:

- Feature influences (LIME, SHAP, attention maps)
- Searching for Cunterfactuals
- Similarity, knn

Understanding the Data

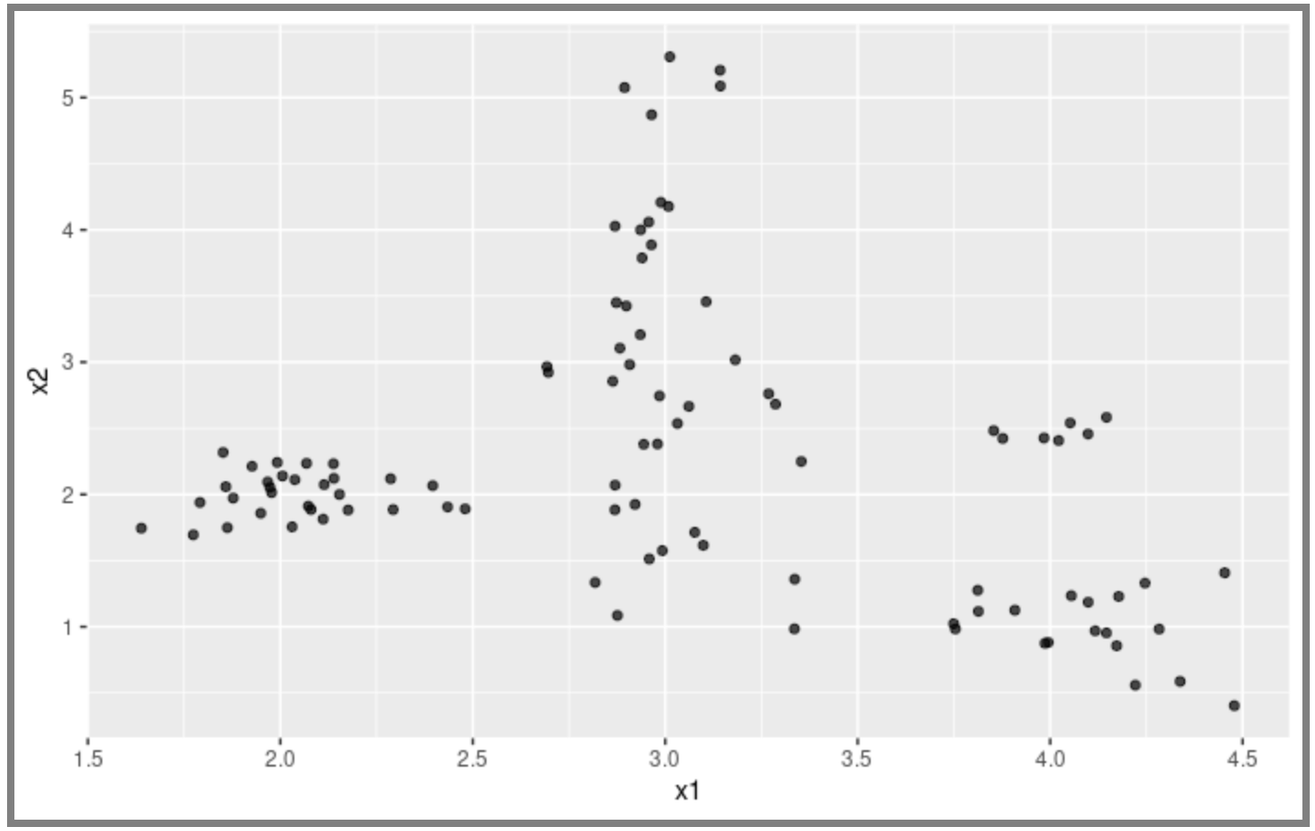
Levels of explanations:

- Understanding a model
- Explaining a prediction
- **Understanding the data**

Data Explanation #1: Prototypes and Criticisms

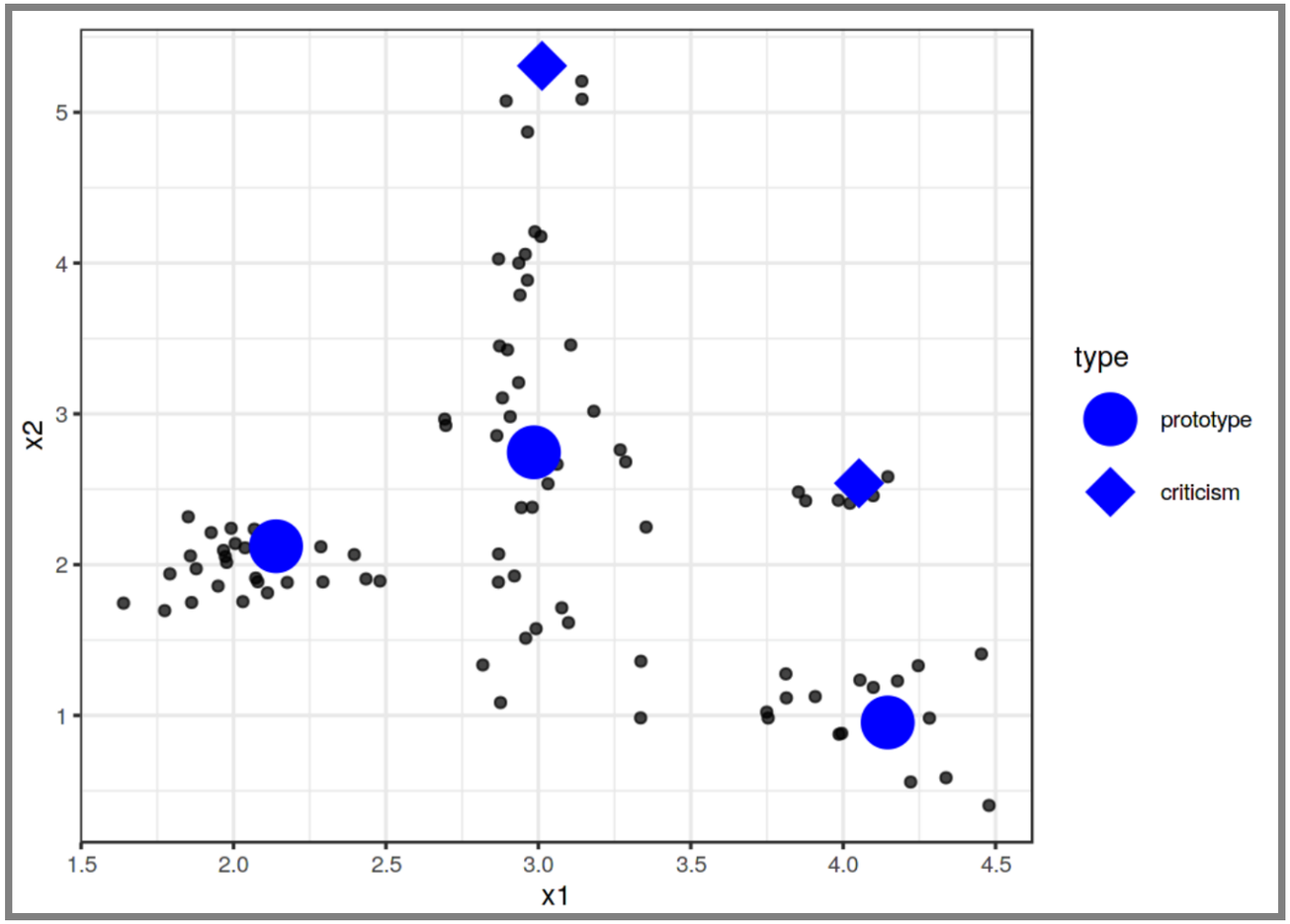
- *Prototype* is a data instance that is representative of all the data
- *Criticism* is a data instance not well represented by the prototypes

Example: Prototypes and Criticisms?



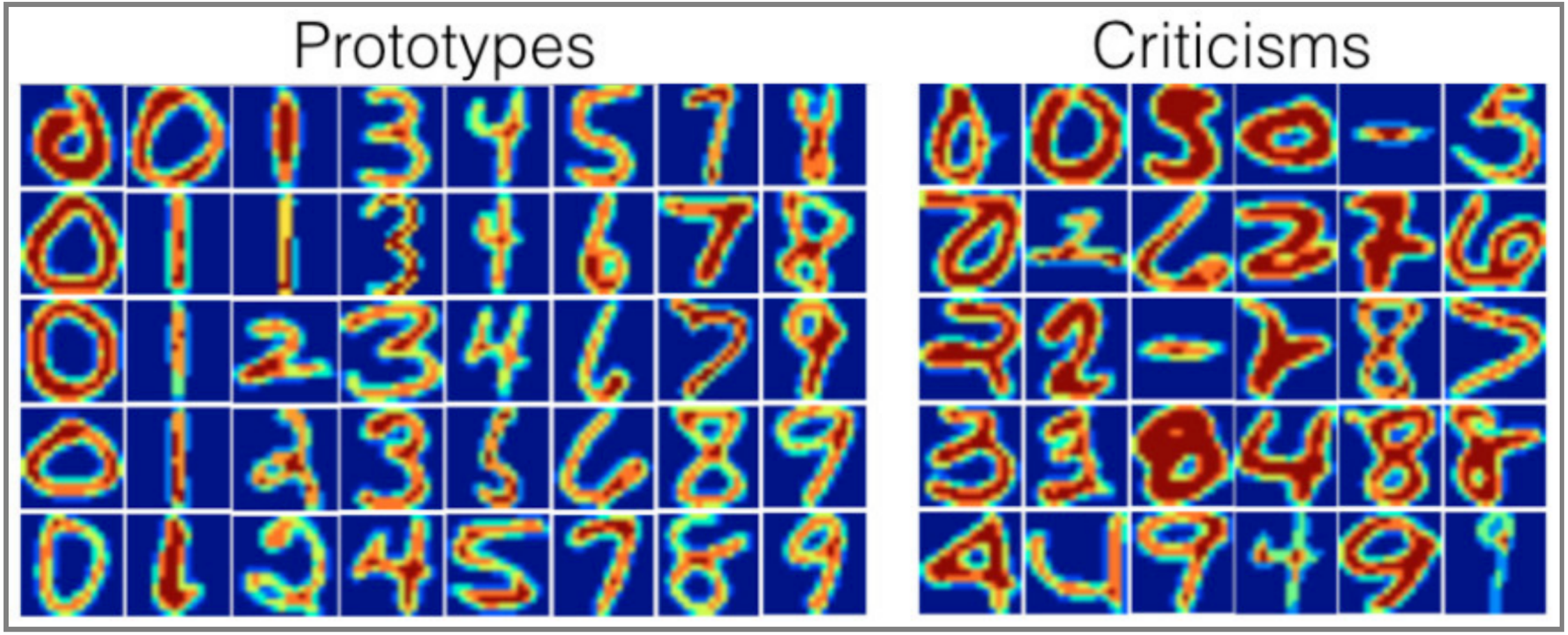
Source: Christoph Molnar. "Interpretable Machine Learning." 2019

Example: Prototypes and Criticisms



Source: Christoph Molnar. "Interpretable Machine Learning." 2019

Example: Prototypes and Criticisms



Source: Christoph Molnar. "Interpretable Machine Learning." 2019

Speaker notes

The number of digits is different in each set since the search was conducted globally, not per group.

Methods: Prototypes and Criticisms

Clustering of data (ala k-means)

- k-medoids returns actual instances as centers for each cluster
- MMD-critic identifies both prototypes and criticisms
- see book for details

Identify globally or per class

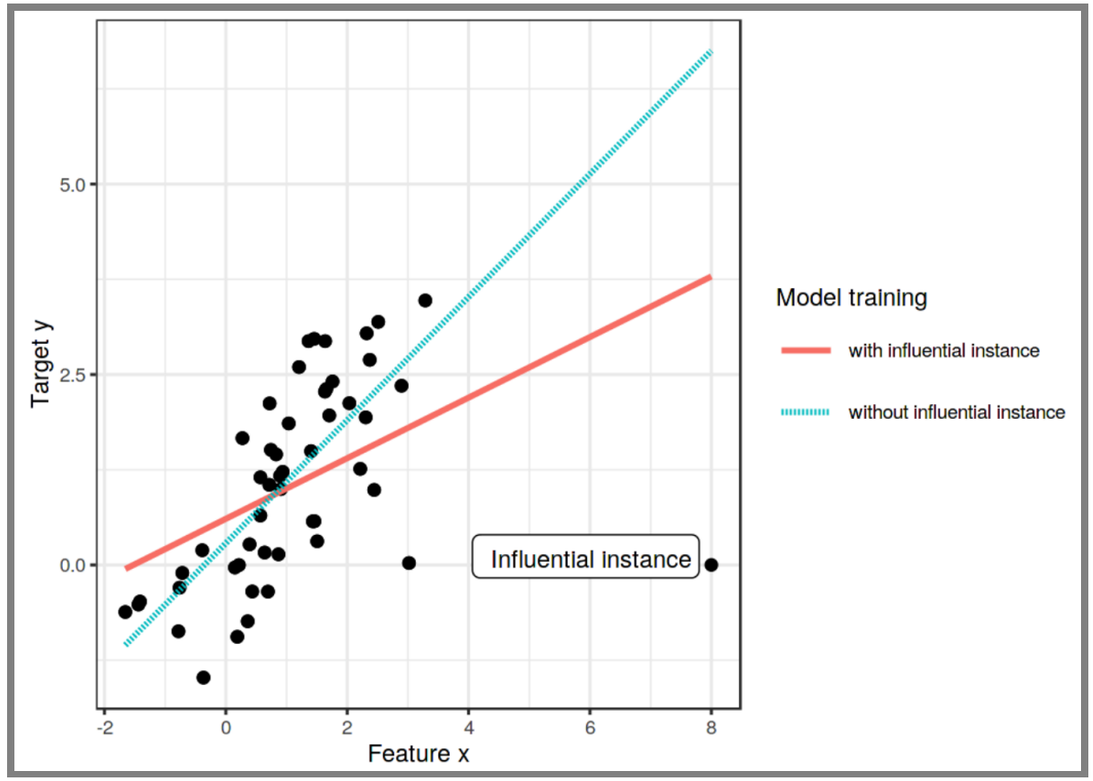
Discussion: Prototypes and Criticisms

- Easy to inspect data, useful for debugging outliers
- Generalizes to different kinds of data and problems
- Easy to implement algorithm

- Need to choose number of prototypes and criticism upfront
- Uses all features, not just features important for prediction

Data Explanation #2: Influential Instance

Data debugging: *What data most influenced the training?*



Source: Christoph Molnar. "Interpretable Machine Learning." 2019

Influential Instances

Data debugging: *What data most influenced the training? Is the model skewed by few outliers?*

Approach:

- Given training data with instances…
- … train model with all instances
- … train model with instances
- If and differ significantly, omitted instance was influential
  - Difference can be measured e.g. in accuracy or difference in parameters

Speaker notes

Instead of understanding a single model, comparing multiple models trained on different data

Influential Instances Discussion

- Retraining for every data point is simple but expensive
- For some class of models, influence of data points can be computed without retraining (e.g., logistic regression), see book for details
- Hard to generalize to taking out multiple instances together (need to take out groups of data together)
- Useful model-agnostic debugging tool for models and data

Three Concepts

Feature importance: How much does the model rely on a feature, across all predictions?

Feature influence: How much does a specific prediction rely on a feature?

Influential instance: How much does the model rely on a single training data instance?

Summary: Understanding the Data

Understand the characteristics of the data used to train the model

Many data exploration and data debugging techniques:

- Criticisms and prototypes
- Influential instances
- many others...

Breakout: Debugging with Explanations

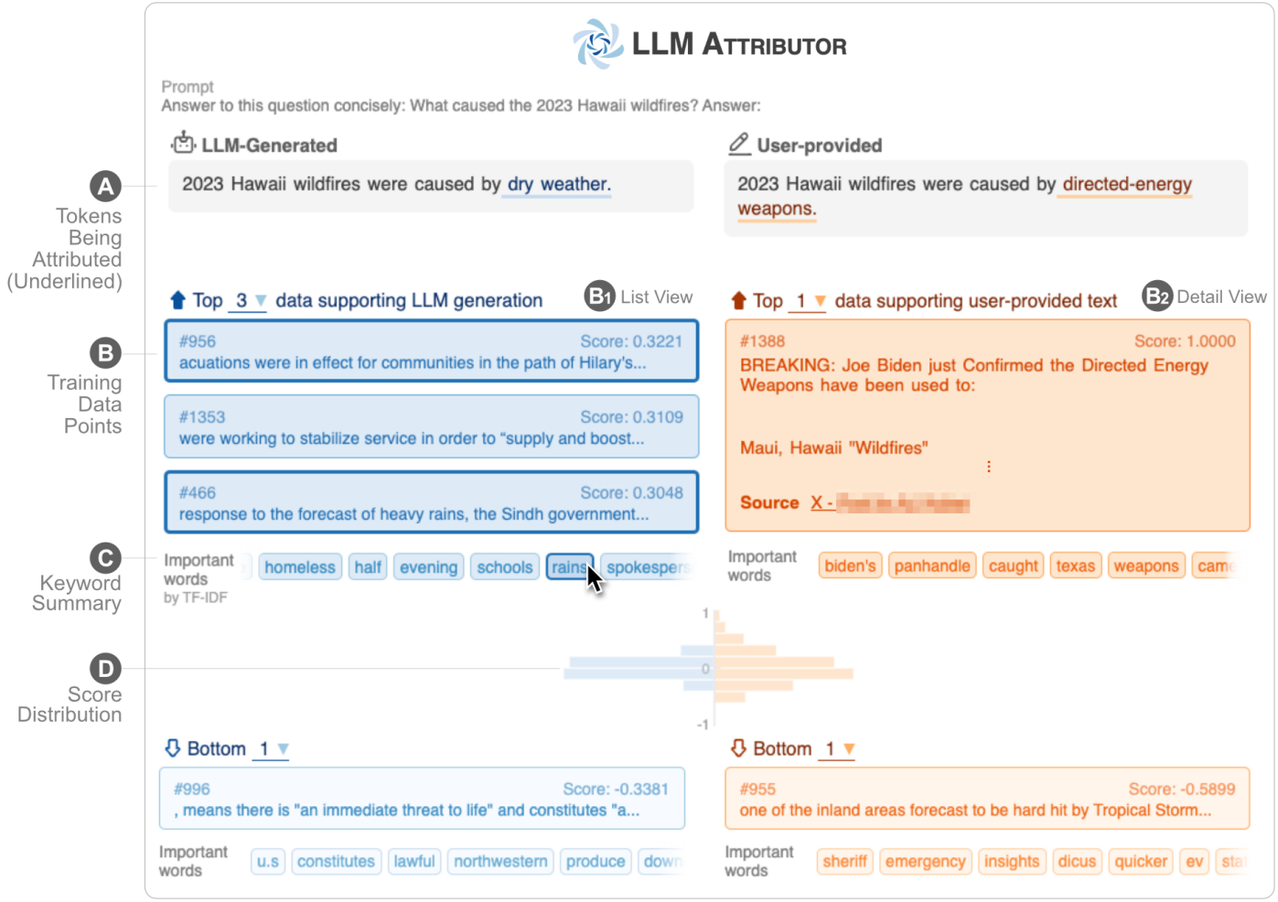
In groups, discuss which explainability approaches may help and why. Tagging group members, write to `#lecture`.

Algorithm bad at recognizing some signs in some conditions:

Graduate appl. system seems to rank applicants from HBCUs low:

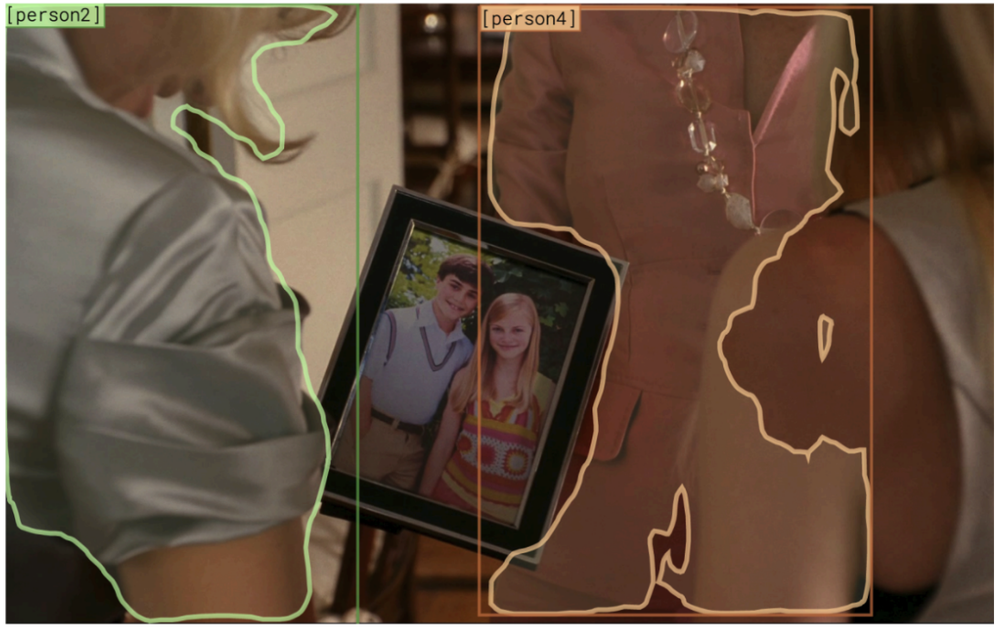
Bonus: Explanations for Generative Models

Attribution reuses existing methods



LLM Attributor: Attribute LLM's Generated Text to Training Data

We can do Natural Language Explanation



Question: What is going to happen next?

Answer: [person2] holding the photo will tell [person4] how cute their children are.

Free-text explanation:

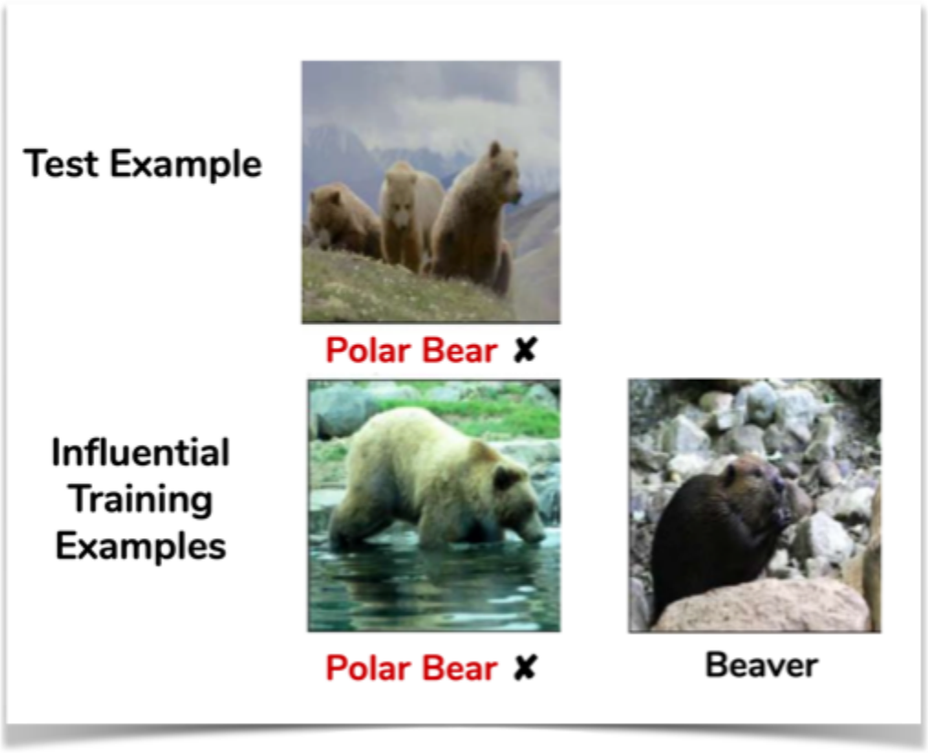
[person4] is showing the photo to [person2] [person2] will want to be polite

We cannot highlight this in the input!

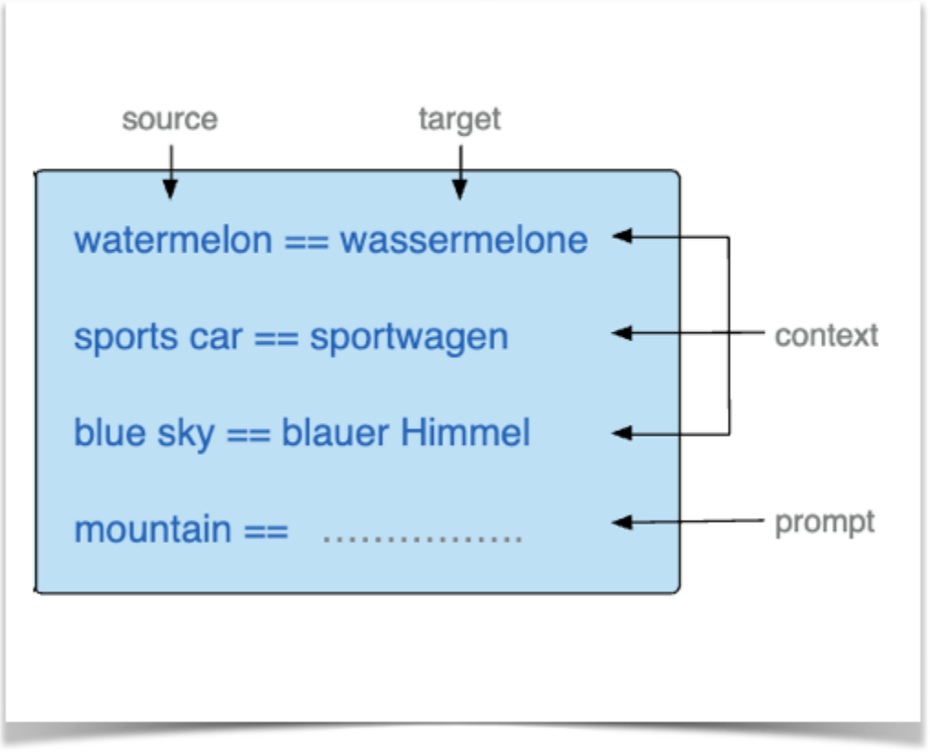
Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." CVPR 2019

Prompt vs. Explanation

Data influence: for a given test prediction, identify the most influential training points



Few-shot selection in LLM prompting has similar format and content!!



Prompt vs. Explanation

NL explanation: Describe why the model make a prediction in plain English

Question:
Where is a frisbee in play likely to be?
Answer Choices:
outside park roof tree air

Free-Flow (FF) Explanation
A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game, so while in play it is most likely to be in the air. A frisbee can be outside or in a park anytime, and other options are possible only after play.

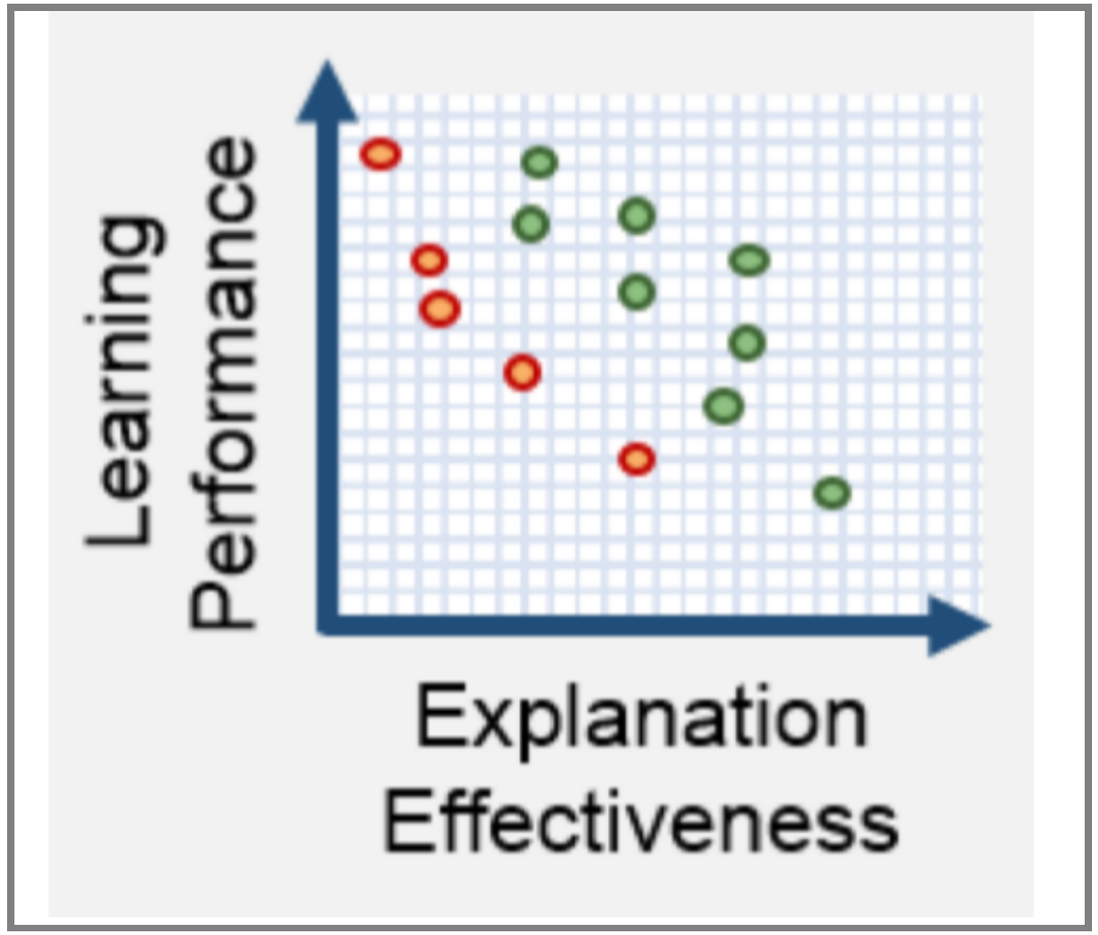
Chain-of-Thought in LLM prompting has similar format and content!!

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

"Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead."

Accuracy vs Explainability Conflict?



Graphic from the DARPA XAI BAA (Explainable Artificial Intelligence)

Faithfulness of Ex-Post Explanations

CORELS' model for recidivism risk prediction

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Simple, interpretable model with comparable accuracy to proprietary COMPAS model

"Stop explaining ..."

Hypotheses:

- It is a myth that there is necessarily a trade-off between accuracy and interpretability (when having meaningful features)
- Explainable ML methods provide explanations that are not faithful to what the original model computes
- Explanations often do not make sense, or do not provide enough detail to understand what the black box is doing
- Black box models are often not compatible with situations where information outside the database needs to be combined with a risk assessment
- Black box models with explanations can lead to an overly complicated decision pathway that is ripe for human error

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215. (Preprint)

Prefer Interpretable Models over Post-Hoc Explanations

- Interpretable models provide faithful explanations
  - post-hoc explanations may provide limited insights or illusion of understanding
  - interpretable models can be audited
- Inherently interpretable models in many cases have similar accuracy
- Larger focus on feature engineering, more effort, but insights into when and *why* the model works
- Less research on interpretable models and some methods computationally expensive

ProPublica Controversy

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Speaker notes

"ProPublica’s linear model was not truly an “explanation” for COMPAS, and they should not have concluded that their explanation model uses the same important features as the black box it was approximating."

ProPublica Controversy

```
IF age between 18–20 and sex is male THEN
  predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN
  predict arrest
ELSE IF more than three priors THEN
  predict arrest
ELSE
  predict no arrest
```

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215.

Drawbacks of Interpretable Models

Intellectual property protection harder

- may need to sell model, not license as service
- who owns the models and who is responsible for their mistakes?

Gaming possible; "security by obscurity" not a defense

Expensive to build (feature engineering effort, debugging, computational costs)

Limited to fewer factors, may discover fewer patterns, lower accuracy

Summary

- Interpretability useful for many scenarios: user feedback, debugging, fairness audits, science, ...
- Defining and measuring interpretability
  - Explaining the model
  - Explaining predictions
  - Understanding the data
- Inherently interpretable models: sparse regressions, shallow decision trees
- Providing ex-post explanations of opaque models: global and local surrogates, dependence plots and feature importance, anchors, counterfactual explanations, criticisms, and influential instances
- Consider implications on user interface design
- Gaming and manipulation with explanations

Further Readings

- Christoph Molnar. “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.” 2019
- Google PAIR. People + AI Guidebook. 2019.
- Cai, Carrie J., Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. “”Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making.” Proceedings of the ACM on Human-computer Interaction 3, no. CSCW (2019): 1–24.
- Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. “Principles of explanatory debugging to personalize interactive machine learning.” In Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 126–137. 2015.
- Amershi, Saleema, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. “Modeltracker: Redesigning performance analysis tools for machine learning.” In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 337–346. 2015.