INTERPRETABILITY AND EXPLAINABILITY

Christian Kaestner

Required reading: 🎧 Data Skeptic Podcast Episode “Black Boxes are not Required” with Cynthia Rudin (32min) or 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

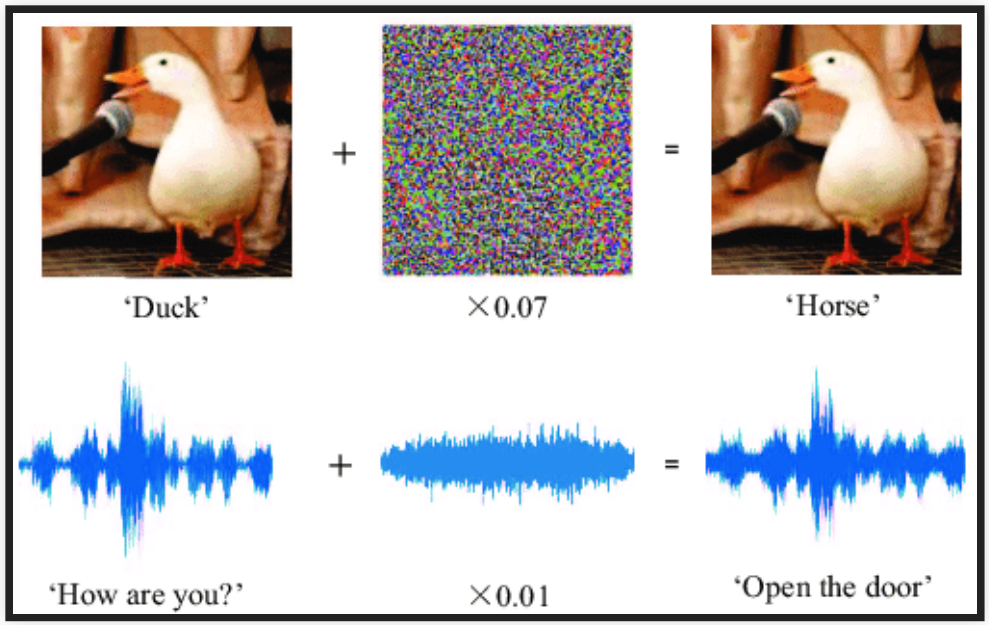
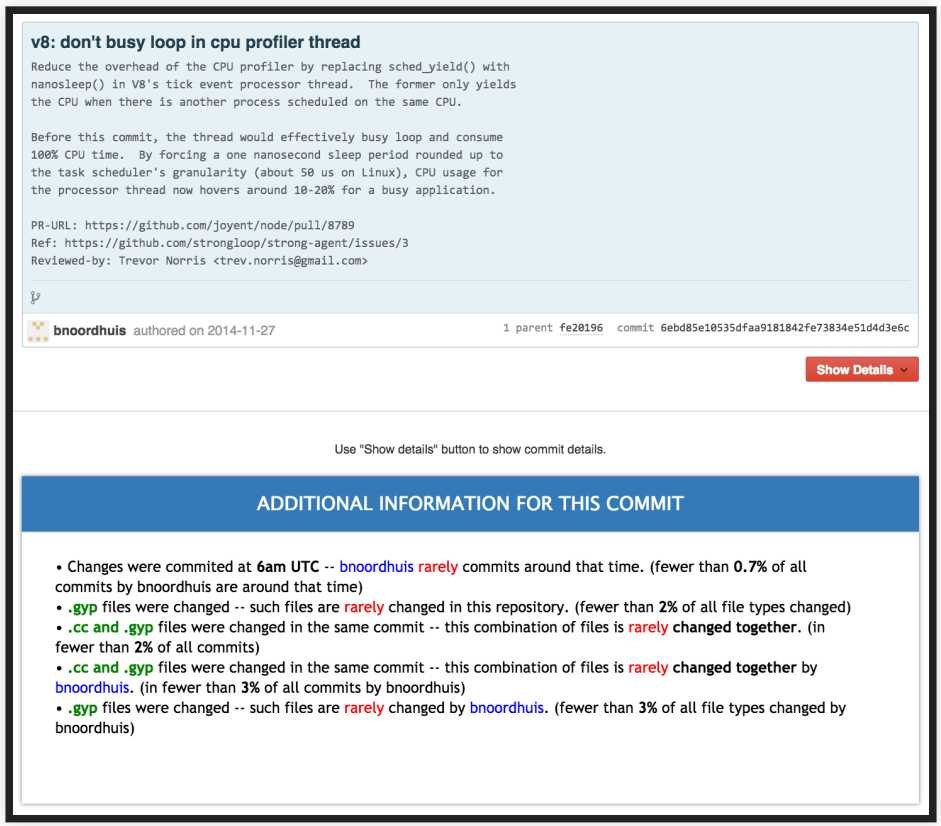


Image: Gong, Yuan, and Christian Poellabauer. "An overview of vulnerabilities of voice controlled systems." arXiv preprint arXiv:1803.09156 (2018).

DETECTING ANOMALOUS COMMITS



v8: don't busy loop in cpu profiler thread

```
Reduce the overhead of the CPU profiler by replacing sched_yield() with
nanosleep() in V8's tick event processor thread.  The former only yields
the CPU when there is another process scheduled on the same CPU.

Before this commit, the thread would effectively busy loop and consume
100% CPU time.  By forcing a one nanosecond sleep period rounded up to
the task scheduler's granularity (about 50 us on Linux), CPU usage for
the processor thread now hovers around 10-20% for a busy application.

PR-URL: https://github.com/joyent/node/pull/8789
Ref: https://github.com/strongloop/strong-agent/issues/3
Reviewed-by: Trevor Norris <trev.norris@gmail.com>
```

bnoordhuis authored on 2014-11-27 — 1 parent fe20196 commit 6ebd85e10535dfaa9181842fe73834e51d4d3e6c

Show Details

Use "Show details" button to show commit details.

ADDITIONAL INFORMATION FOR THIS COMMIT

- Changes were commited at **6am UTC** -- bnoordhuis rarely commits around that time. (fewer than **0.7%** of all commits by bnoordhuis are around that time)
- **.gyp** files were changed -- such files are rarely changed in this repository. (fewer than **2%** of all file types changed)
- **.cc and .gyp** files were changed in the same commit -- this combination of files is rarely **changed together**. (in fewer than **2%** of all commits)
- **.cc and .gyp** files were changed in the same commit -- this combination of files is rarely **changed together** by bnoordhuis. (in fewer than **3%** of all commits by bnoordhuis)
- **.gyp** files were changed -- such files are rarely changed by bnoordhuis. (fewer than **3%** of all file types changed by bnoordhuis)

Goyal, Raman, Gabriel Ferreira, Christian Kästner, and James Herbsleb. "Identifying unusual commits on GitHub." Journal of Software: Evolution and Process 30, no. 1 (2018): e1893.

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 INTERPRET THE RESULTS?

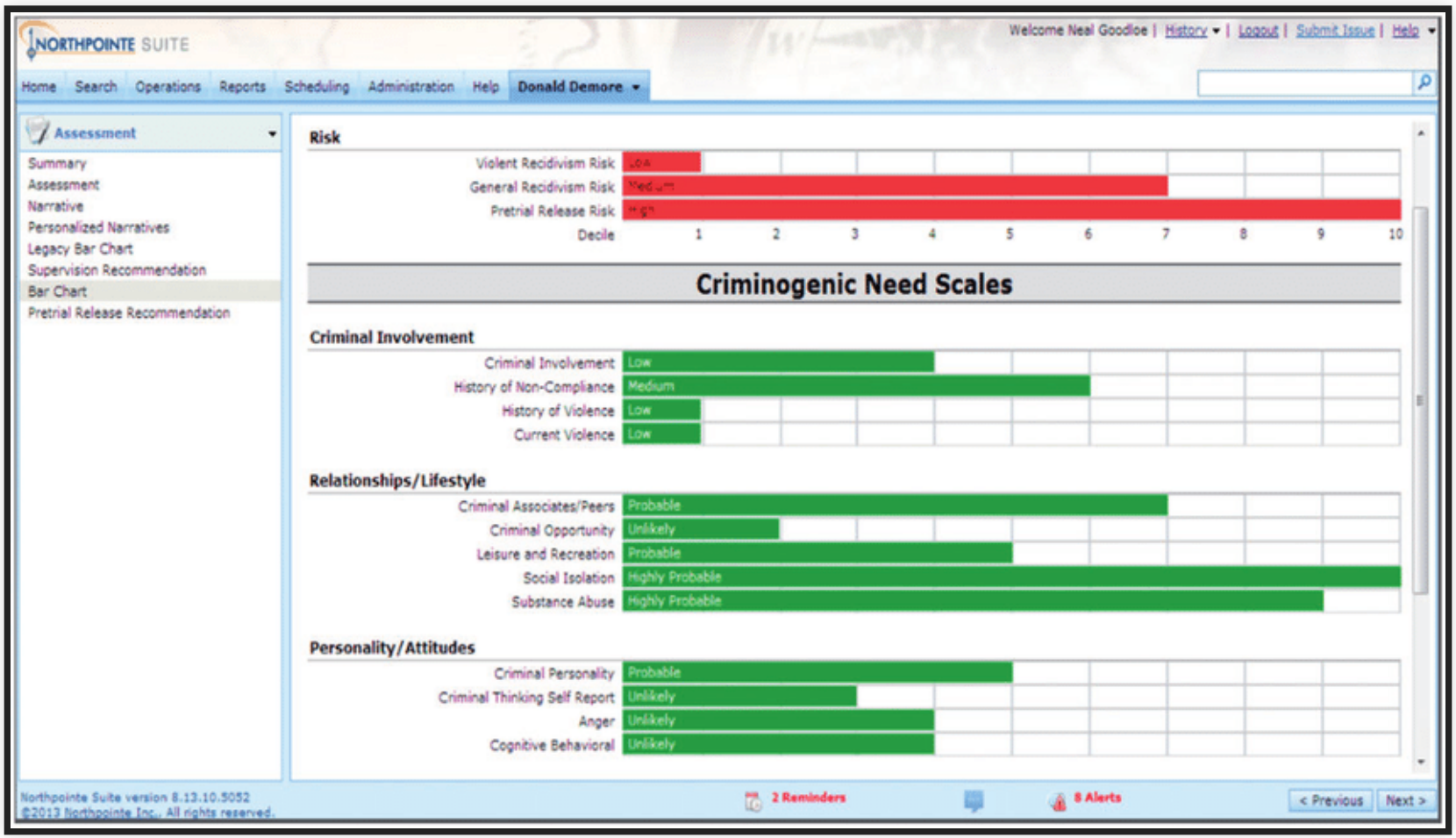


Image source (CC BY-NC-ND 4.0): Christin, Angèle. (2017). Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society. 4.

HOW TO JUDGE RELATIVE TO 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.

WHAT FACTORS GO INTO PREDICTING STROKE RISK?

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%

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?

Tweet

Tweet

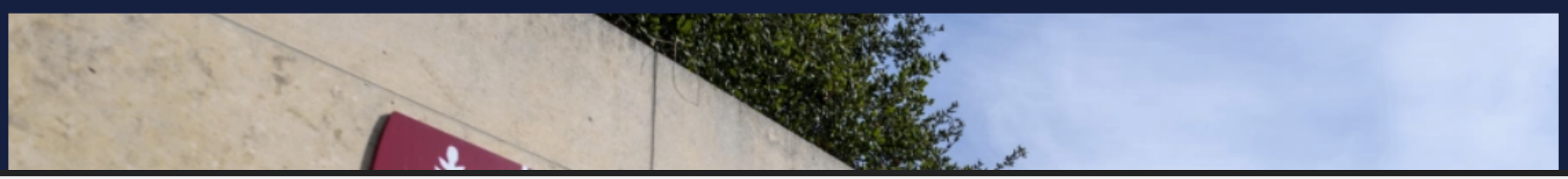
PANDEMIC TECHNOLOGY PROJECT

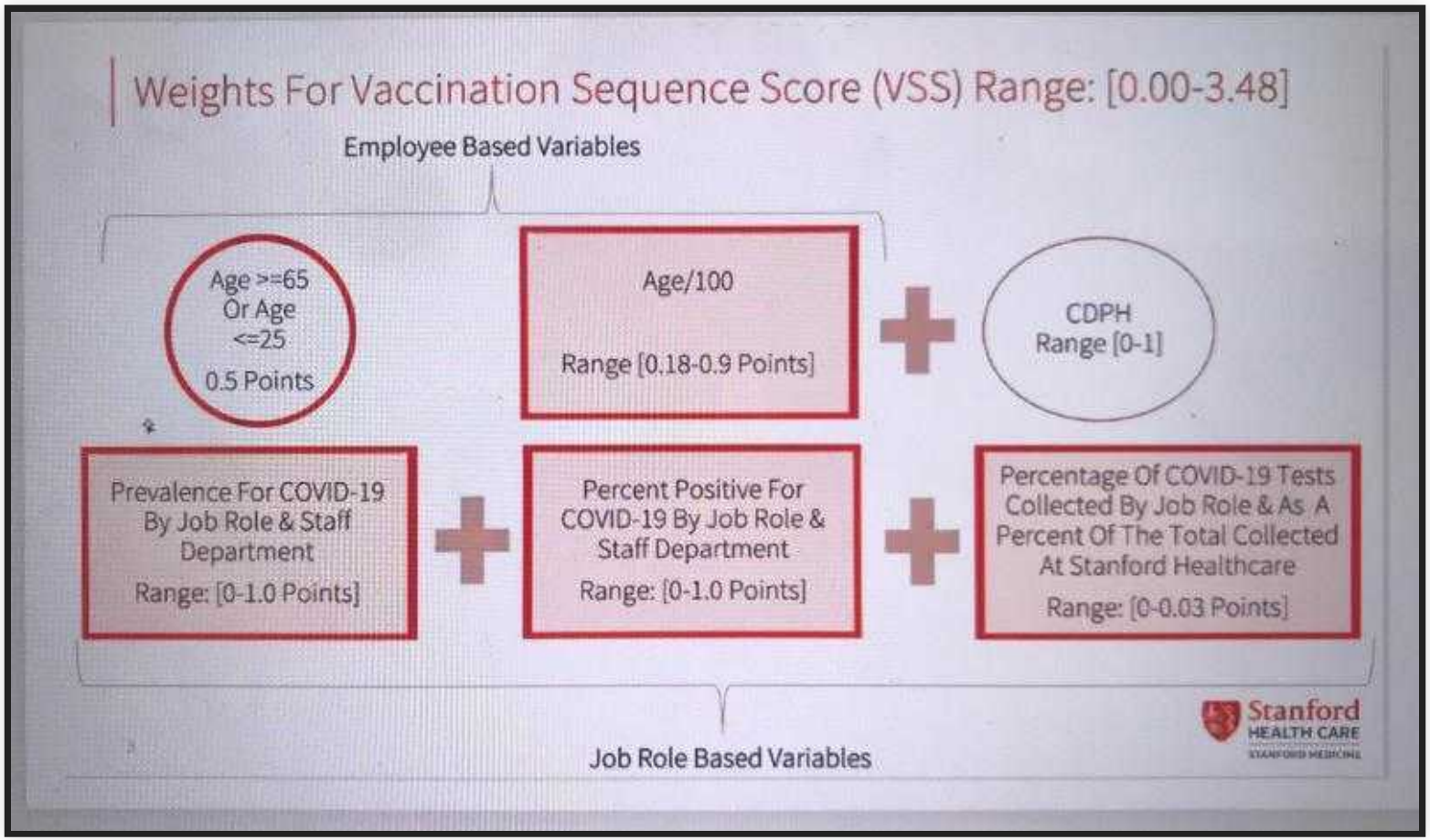
This is the Stanford vaccine algorithm that left out frontline doctors

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



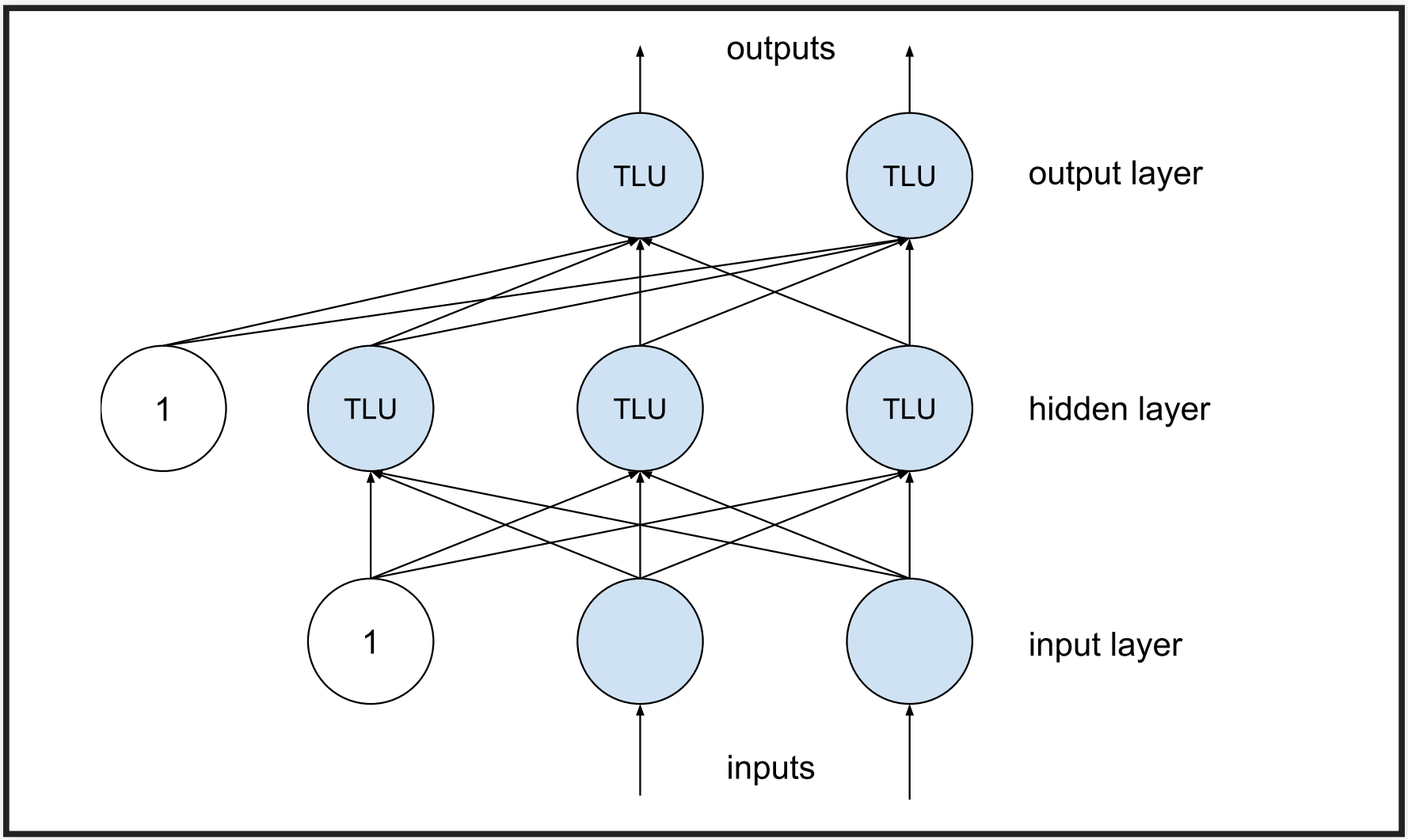

Weights For Vaccination Sequence Score (VSS) Range: [0.00-3.48]
Employee Based Variables
Age >=65 Or Age <=25
0.5 Points
Age/100
Range [0.18-0.9 Points]
CDPH
Range [0-1]
Prevalence For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]
Percent Positive For COVID-19 By Job Role & Staff Department
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

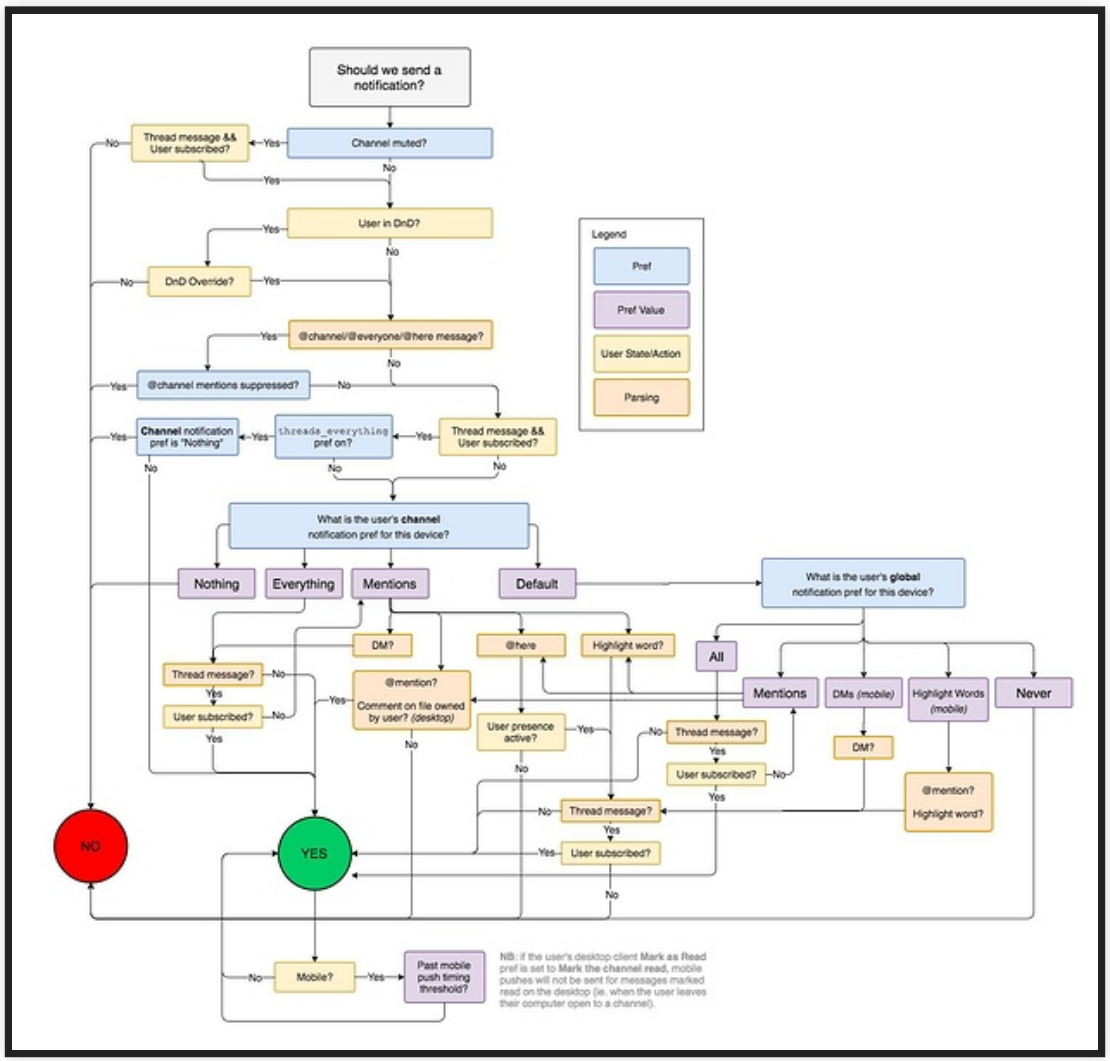
Cat? Dog? Lion?

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?

WHY EXPLAINABILITY?

DEBUGGING

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

Most common use case in practice according to recent study (Bhatt et al. "Explainable machine learning in deployment." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648-657. 2020.)

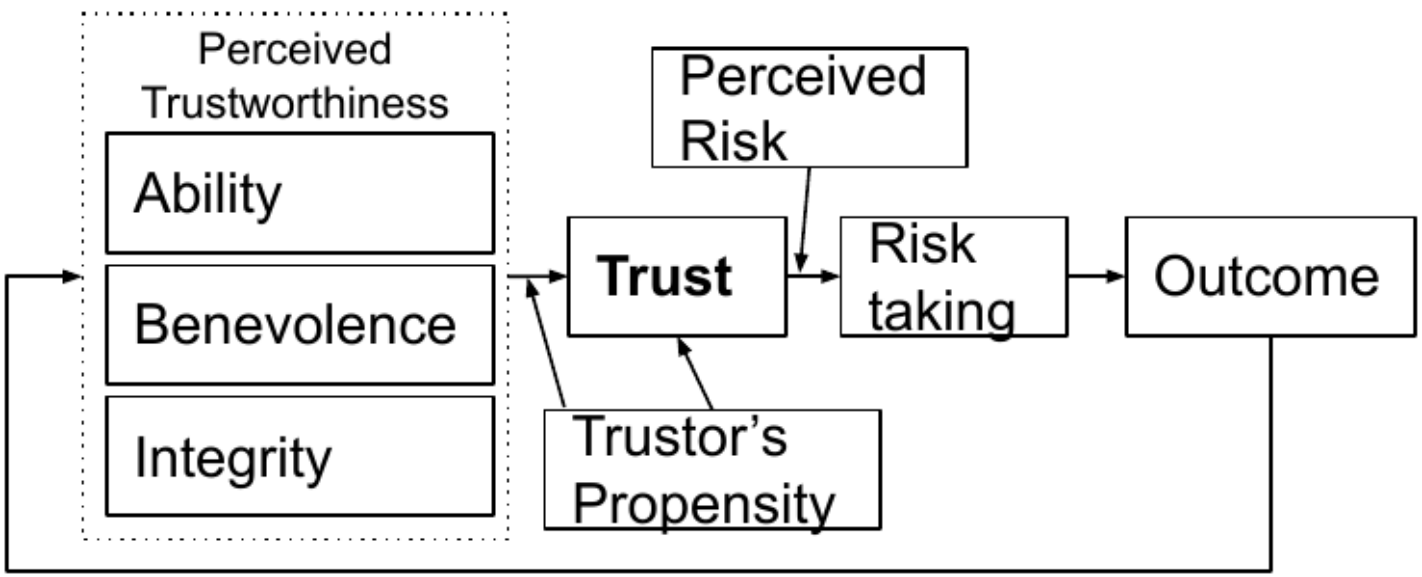
AUDITING

- Understand safety implications
- Ensure predictions are based on objective criteria and reasonable rules
- Inspect fairness properties
- Reason about biases and feedback loops
- ML as Requirements Engineering view: Validate "mined" 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 arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

TRUST

- Willing to accept a prediction more if understandable 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.

ACTIONALBLE 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 European Union General Data Protection Regulation extends the automated decision-making rights in the 1995 Data Protection Directive 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

- What drove our past hiring decisions? Who gets promoted around here?
- What factors influence cancer risk? Recidivism?
- What influences demand for bike rentals?
- Which organizations are successful at raising donations and why?

	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

SETTINGS WHERE INTERPRETABILITY IS *NOT* IMPORTANT?

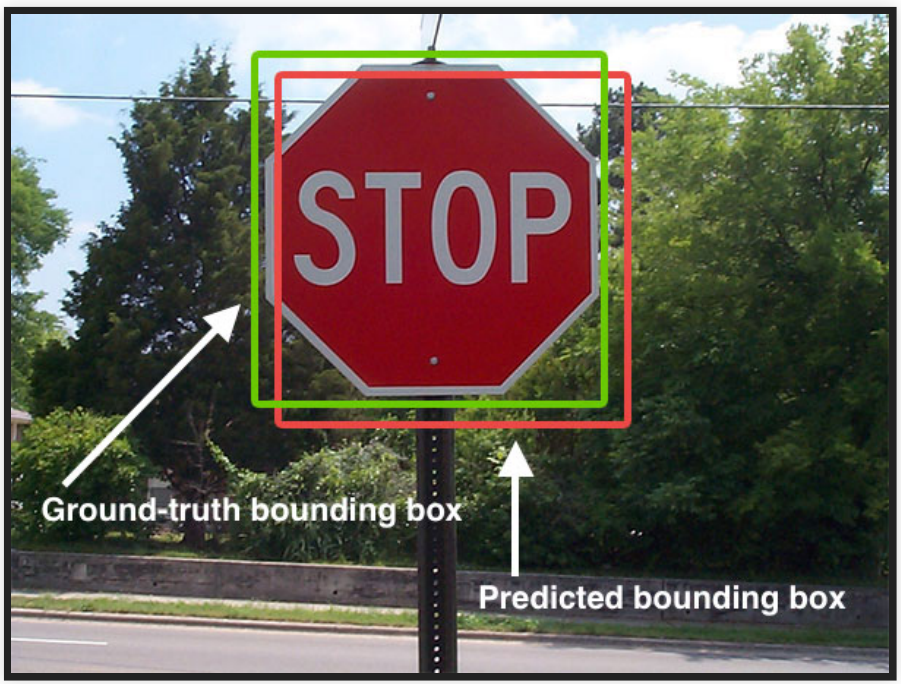
Speaker notes

- 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 5 min report back to the group.

Algorithm bad at recognizing some signs in some conditions:



Graduate application system seems to rank applicants HBCUs lowly:

DEFINING AND MEASURING INTERPRETABILITY

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

INTERPRETABILITY DEFINITIONS

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)

MEASURING INTERPRETABILITY?

Speaker notes

Experiments asking humans questions about the model, e.g., what would it predict for X, how should I change inputs to predict Y?

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, such as

- Why was the loan rejected? (justification)
- Why did the treatment not work for the patient? (debugging)
- Why is turnover higher among women? (general data science question)

MEASURING EXPLANATION QUALITY?

THREE LEVELS OF EVALUATING INTERPRETABILITY

- Functionally-grounded evaluation, proxy tasks without humans (least specific and expensive)
    - Depth of a decision tree (assuming smaller trees are easier to understand)
- Human-grounded evaluation, simple tasks with humans
    - Ask crowd-worker which explanation of a loan application they prefer
- Application-grounded evaluation, real tasks with humans (most specific and expensive)
    - Would a radiologist explain a cancer diagnosis in a similar way?

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," 2017.

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 black-box models, local or global

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

Load applications are always declined if the savings account has less than $50.

ON TERMINOLOGY

- Rudin's terminology and this lecture:
    - Interpretable models: Intrinsily interpretable models
    - Explainability: Post-hoc explanations
- 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
- These terms are often used inconsistently or interchangeble

UNDERSTANDING A MODEL

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

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 ar
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest
```

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 understandable ("average over multiple interpretations"?)

```
173554,681081086 * root + 318523,818532818 * heuristicUnit +
-103411,870761673 * eq + -24600,5000000002 * heuristicVsids +
-11816,7857142856 * heuristicVmtf + -33557,8961038976 *
heuristic + -95375,3513513509 * heuristicUnit * satPreproYes +
3990,79729729646 * transExt * satPreproYes + -136928,416666666
* eq * heuristicUnit + 12309,4990990994 * eq * satPreproYes +
33925,0833333346 * eq * heuristic + -643,428571428088 *
backprop * heuristicVsids + -11876,2857142853 * backprop *
heuristicUnit + 1620,24242424222 * eq * backprop +
-7205,2500000002 * eq * heuristicBerkmin + -2 * Num1 * Num2 +
10 * Num3 * Num4
```

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.

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)

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

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.

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 g on (X,Y)
4. Interpret surrogate model g

Can measure how well g fits f with common model quality measures, typically R^2

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: FEATURE IMPORTANCE

- Permute a features value in training or validation set to not use it for prediction
- Measure influence on accuracy
- i.e. evaluate feature effect without retraining the model

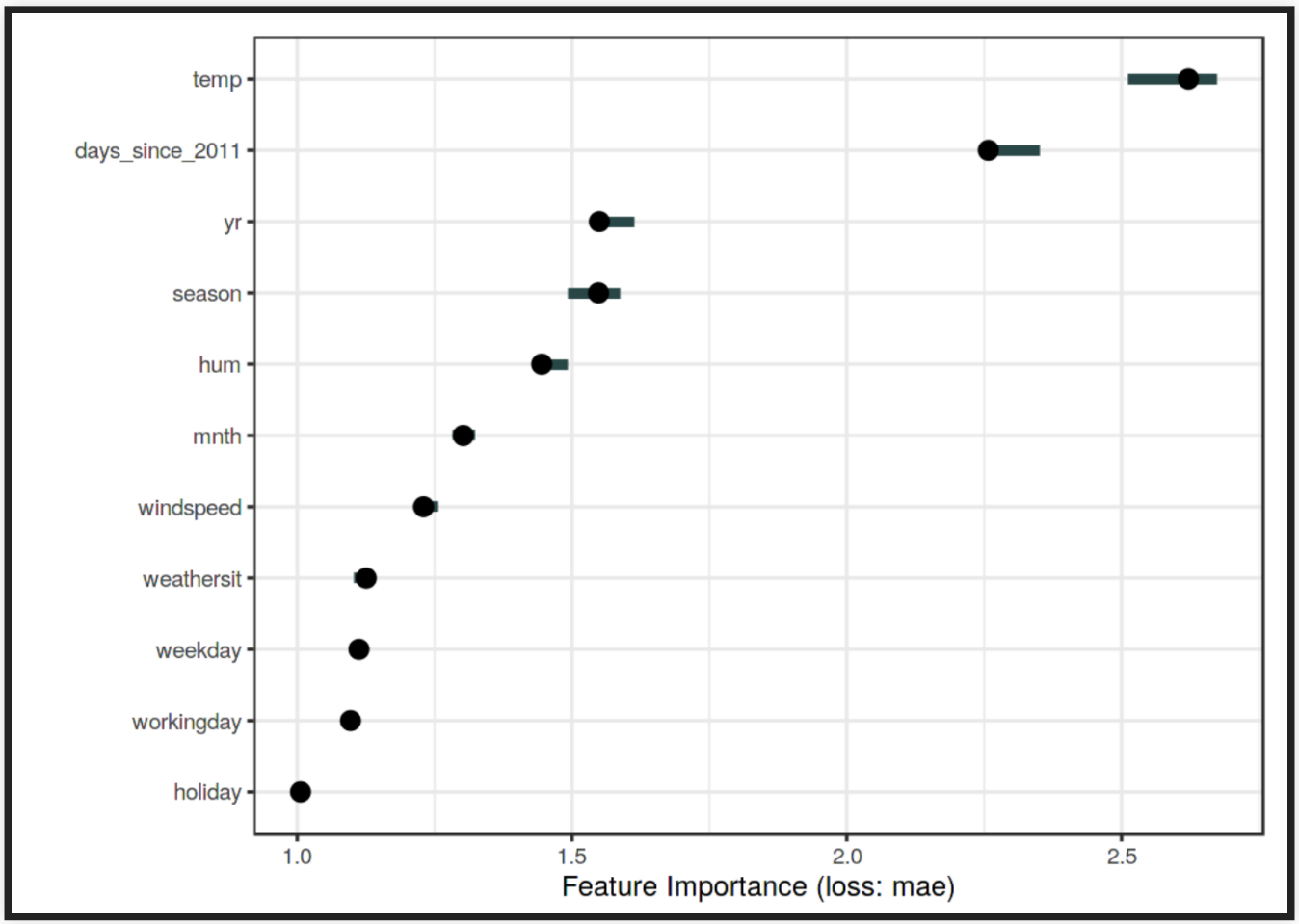
- Highly compressed, *global* insights
- Effect for feature + interactions
- Can only be computed on labeled data, depends on model accuracy, randomness from permutation
- May produce unrealistic inputs when correlations exist

Feature importance on training or 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.

FEATURE IMPORTANCE EXAMPLE



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

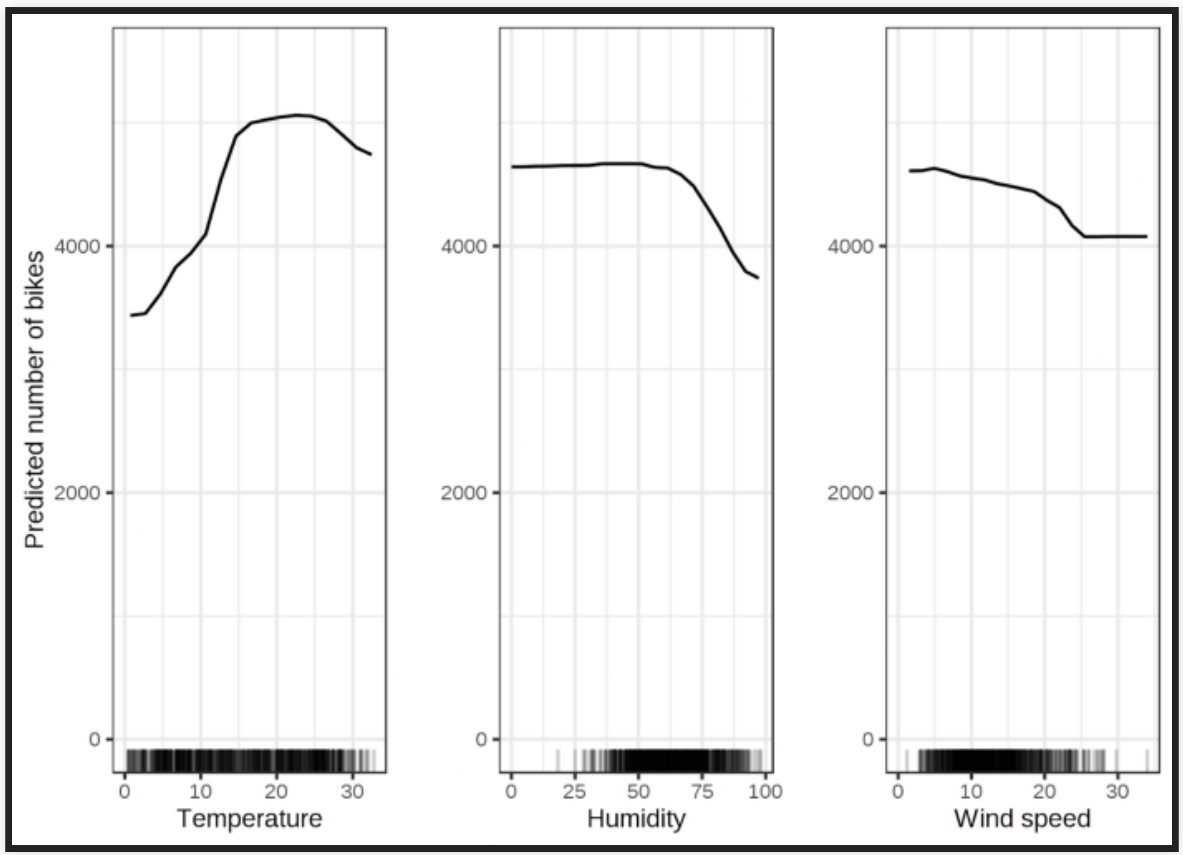
POST-HOC MODEL EXPLANATION: PARTIAL DEPENDENCE PLOT (PDP)

- 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 EXAMPLE

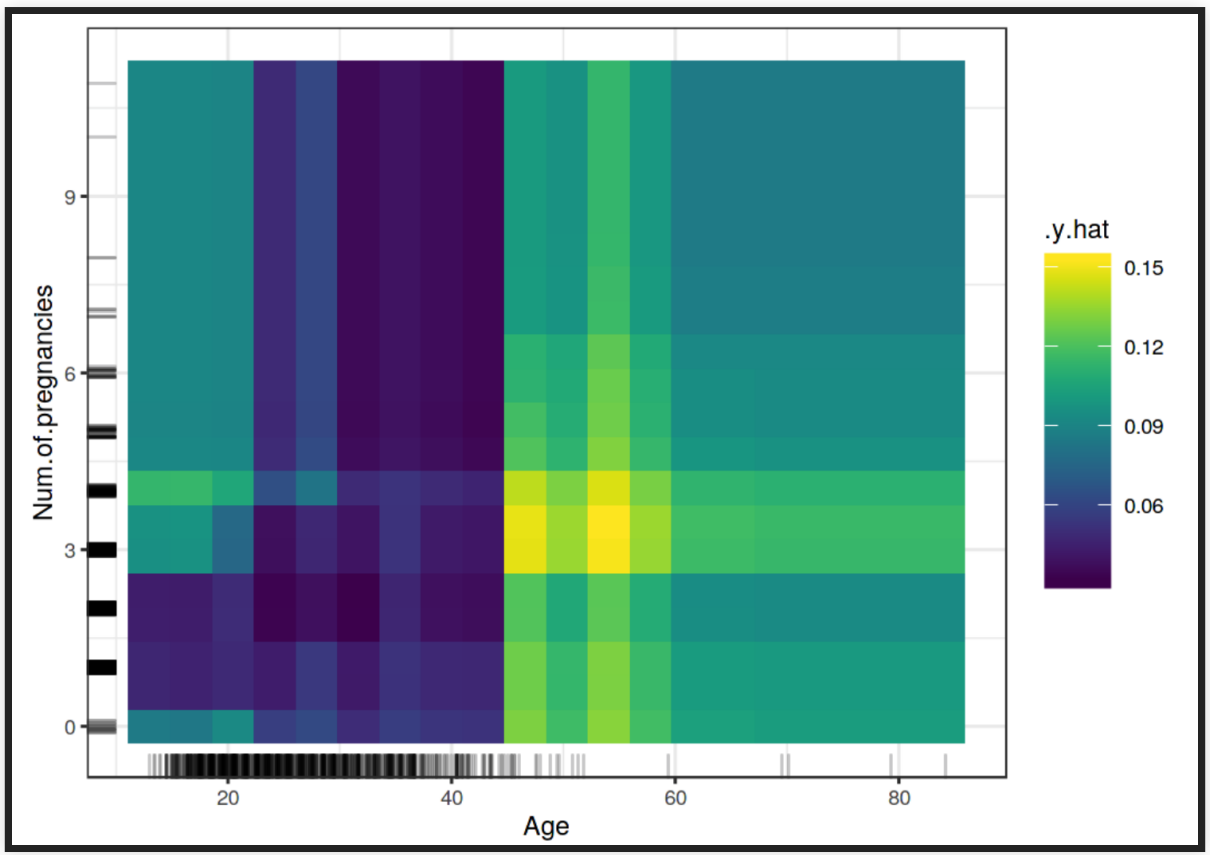
Bike rental in DC



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

PARTIAL DEPENDENCE PLOT EXAMPLE

Probability of cancer



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

EXPLAINING A PREDICTION

PREDICTIONS FROM INHERENTLY INTERPRETABLE MODELS

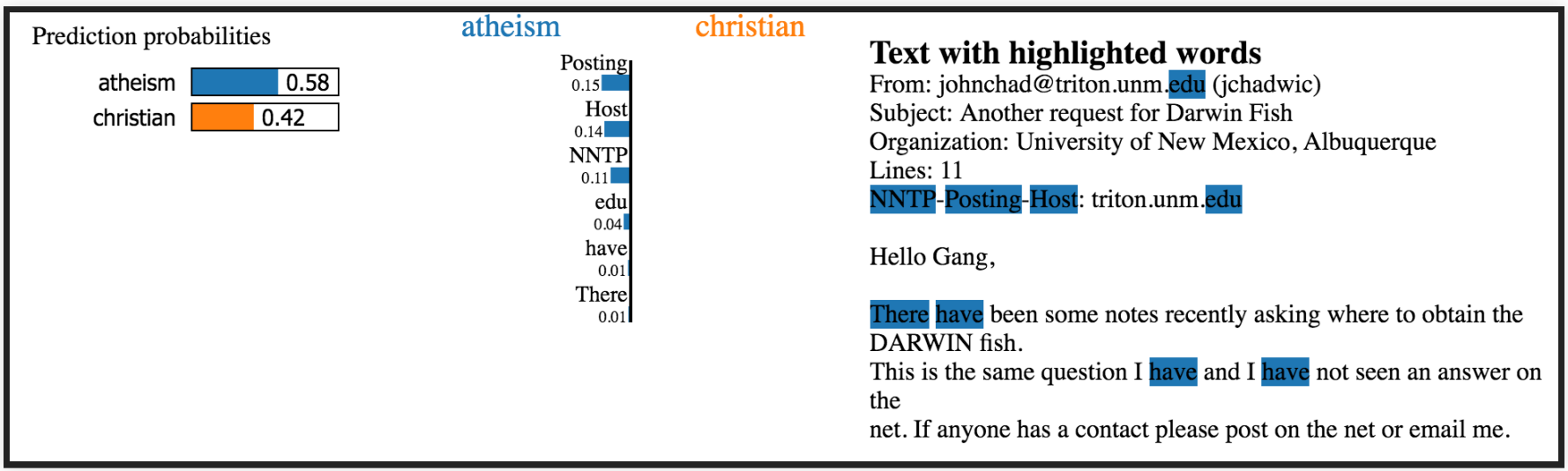
- Derive key influence factors or decisions from model parameters
- Derive contrastive counterfacturals from models

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 arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

POSTHOC PREDICTION EXPLANATION: FEATURE INFLUENCES

- Which features were most influencial for a specific prediction



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

POSTHOC PREDICTION EXPLANATION: FEATURE INFLUENCES

- Which features were most influencial for a specific prediction

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

FEATURE INFLUENCE WITH LOCAL SURROGATES (LIME)

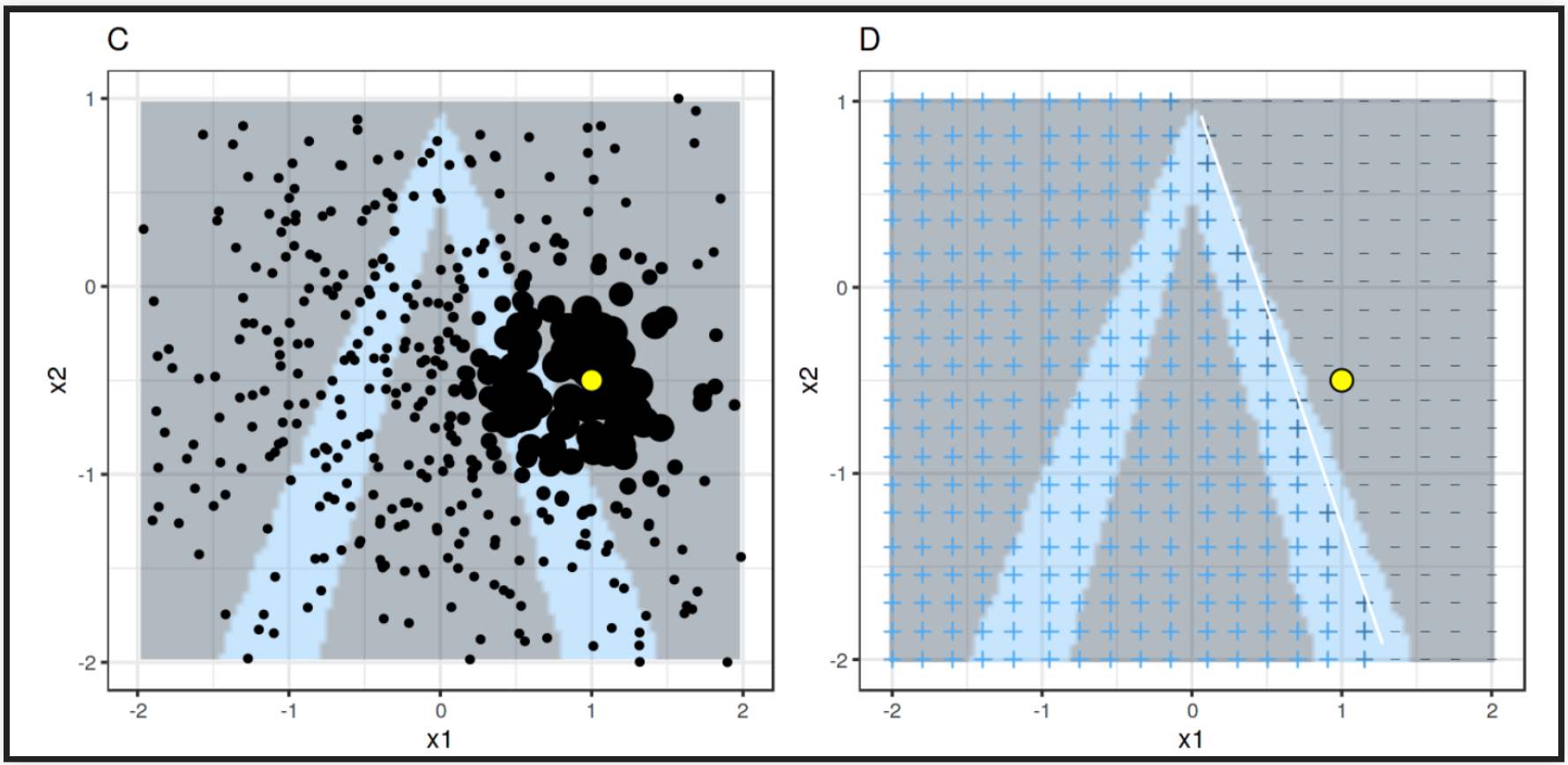
Create an inherently interpretable model (e.g. sparse linear model) for the area around a prediction

Lime approach:

- Create random samples in the area around the data point of interest
- Collect model predictions with f for each sample
- Learn surrogate model g, weighing samples by distance
- Interpret surrogate model g

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier." In Proc International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144. 2016.

LIME EXAMPLE



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

Speaker notes

Model distinguishes blue from gray area. Surrogate model learns only a while line for the nearest decision boundary, which may be good enough for local explanations.

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 International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144. 2016.

ADVANTAGES AND DISADVANTAGES OF LOCAL SURROGATES?

POSTHOC PREDICTION EXPLANATION: SHAPLEY VALUES

- Game-theoretic foundation for local explanations (1953)
- Explains contribution of each feature, over predictions with different subsets of features
    - "The Shapley value is the average marginal contribution of a feature value across all possible coalitions"

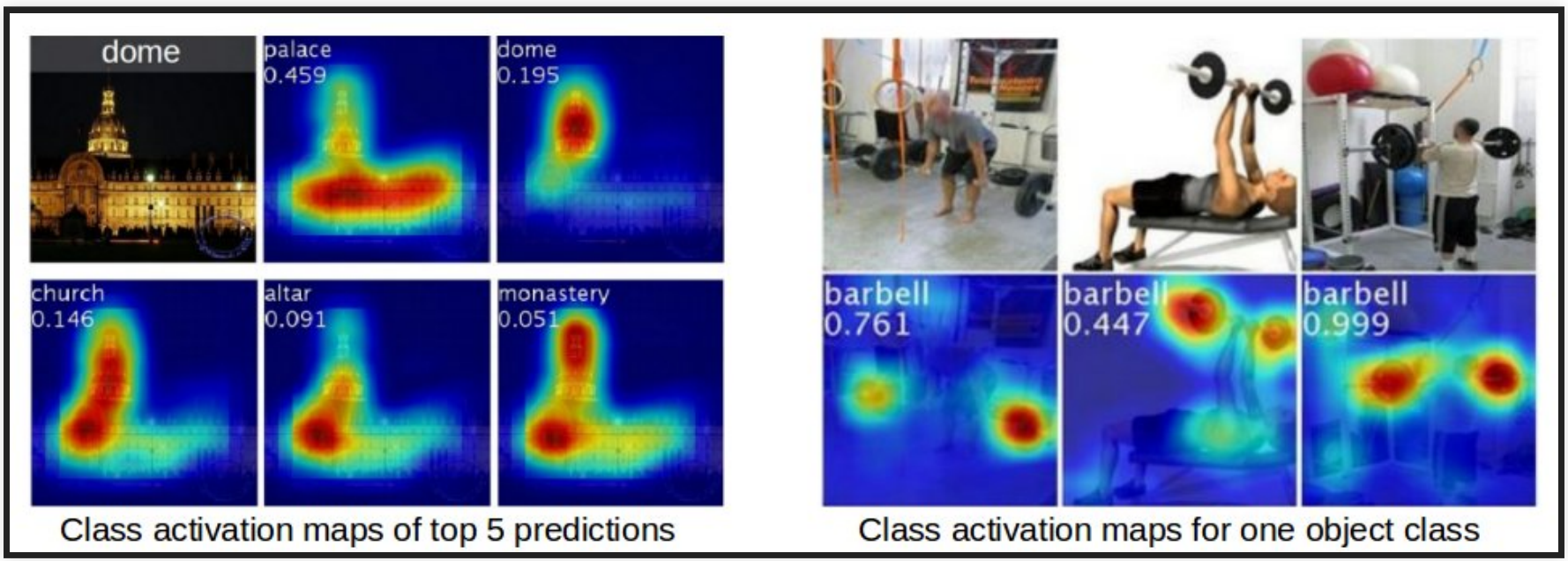
- Solid theory ensures fair mapping of influence to features
- Requires heavy computation, usually only approximations feasible
- Explanations contain all features (ie. not sparse)
- Influence, not counterfactuals

- Currently, most common local method used in practice

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." In Advances in neural information processing systems, pp. 4765-4774. 2017.

Bhatt et al. "Explainable machine learning in deployment." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648-657

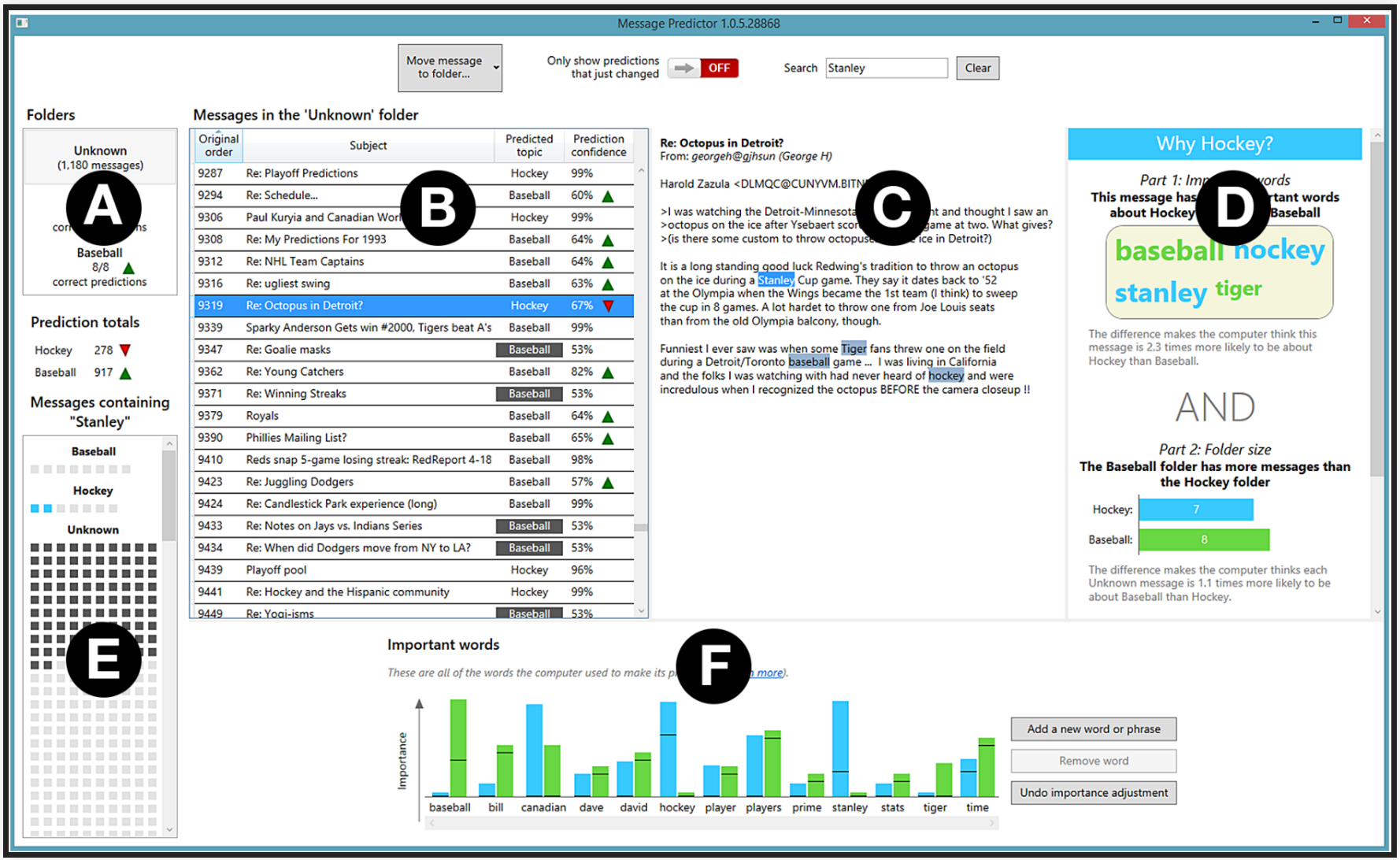
POSTHOC PREDICTION EXPLANATION: ATTENTION MAPS



Identifies which parts of the input lead to decisions

Source: B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR'16

USER INTERFACE DESIGN



Kulesza, T., Burnett, M., Wong, W-K. & Stumpf, S. (2015). Principles of s Explanatory Debugging to personalize interactive machine learning. In: Proc. International Conference on Intelligent User Interfaces. (pp. 126-137)

ANCHORS

- Identify partial conditions that are sufficient for a prediction
- e.g. "*when income < X loan is always rejected*"
- For some models, many predictions can be explained with few rules

- Easy to derive from decision trees, probabilistic search in black-box models
- Compare to association rule mining

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

EXAMPLE: ANCHORS

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
rcdv	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5, 400 $\leq$ loan amount $\leq$ \$10, 000	Good Loan

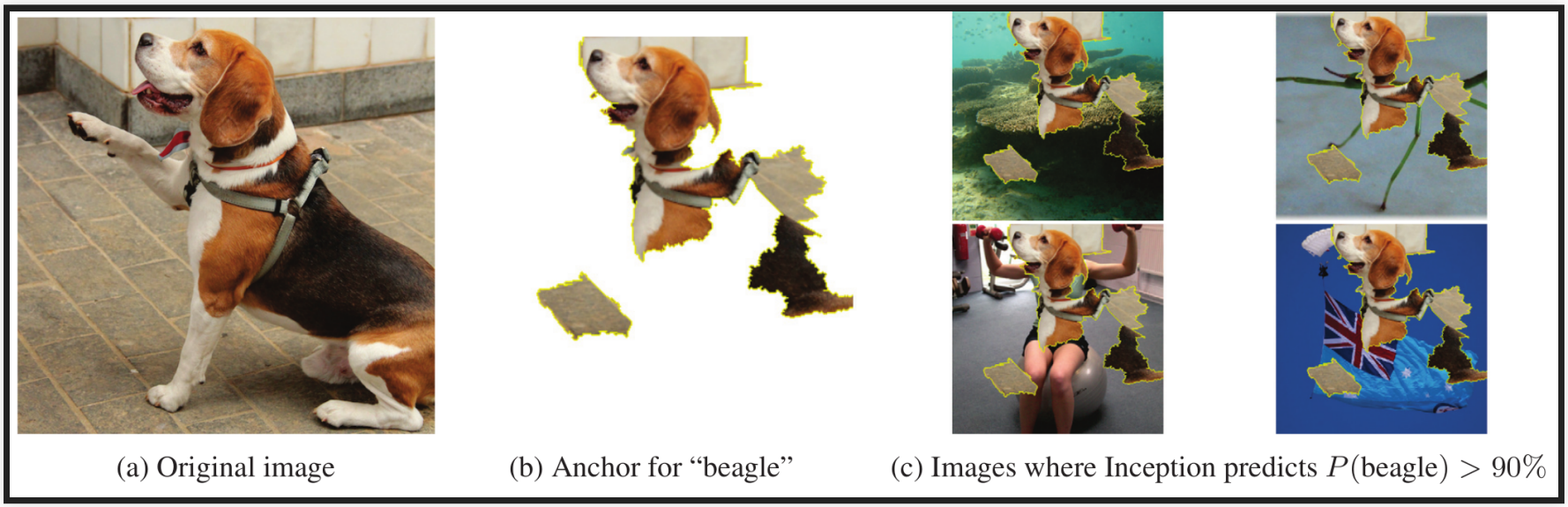
Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

EXAMPLE: ANCHORS

Instance	If	Predict
I want to play(V) ball.	previous word is PARTICLE	play is VERB.
I went to a play(N) yesterday.	previous word is DETERMINER	play is NOUN.
I play(V) ball on Mondays.	previous word is PRONOUN	play is VERB.

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

EXAMPLE: ANCHORS



(a) Original image (b) Anchor for “beagle” (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

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

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)

(Rashomon effect)

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

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}\,(x') - y')^2 + d(x, x')$$

(similar to finding adversarial examples)

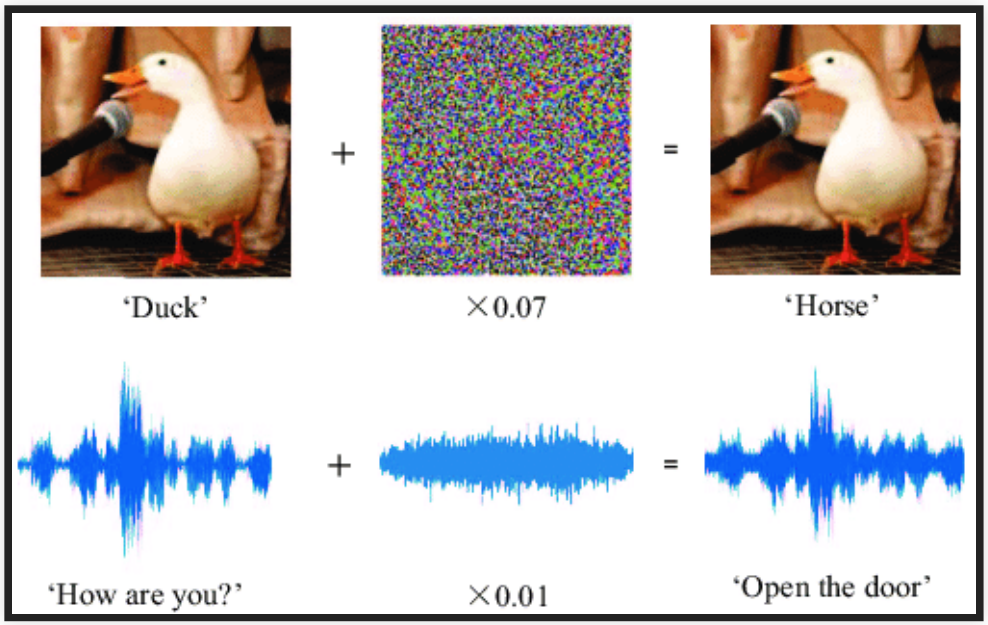
EXAMPLE COUNTERFACTUALS

redicted risk of diabetes with 3-layer neural network

Which feature values must be changed to increase or decrease the risk score of diabetes to 0.5?

- Person 1: If your 2-hour serum insulin level was 154.3, you would have a score of 0.51
- Person 2: If your 2-hour serum insulin level was 169.5, you would have a score of 0.51
- Person 3: If your Plasma glucose concentration was 158.3 and your 2-hour serum insulin level was 160.5, you would have a score of 0.51

DISCUSSION: COUNTERFACTUALS


'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

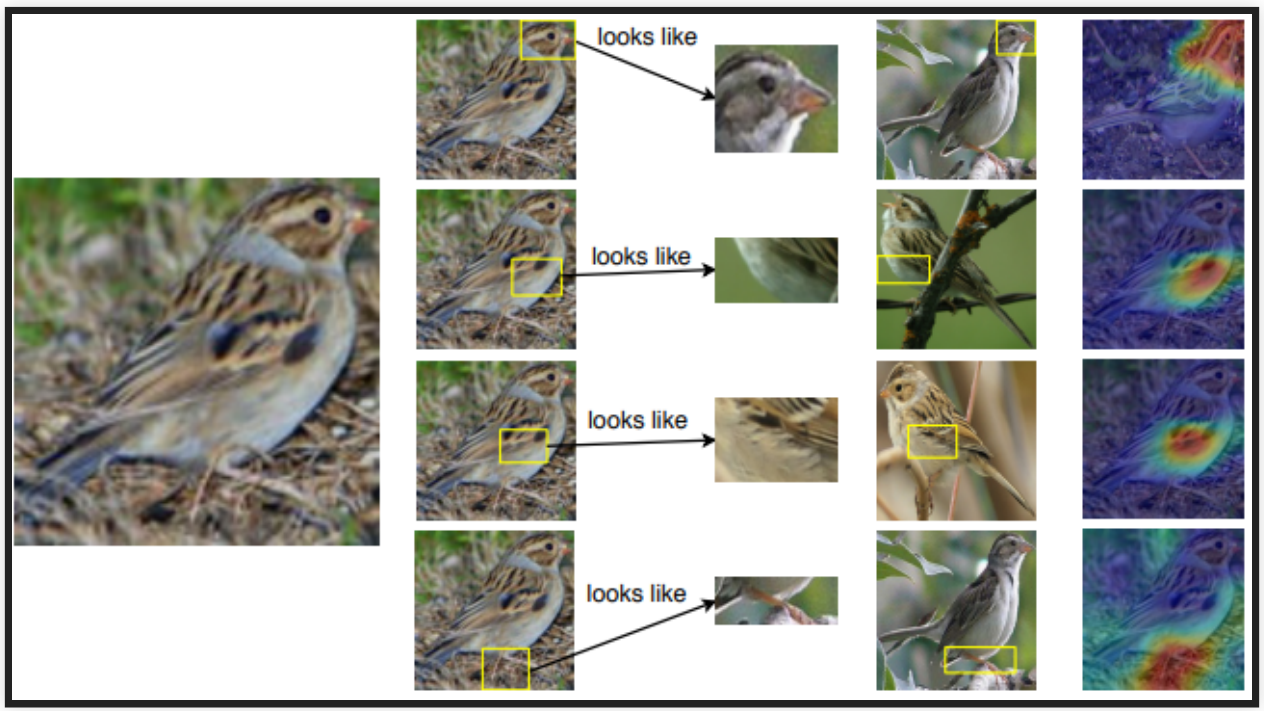
ACTIONABLE COUNTERFACTUALS

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

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." Advances in neural information processing systems 32 (2019).

UNDERSTANDING THE DATA

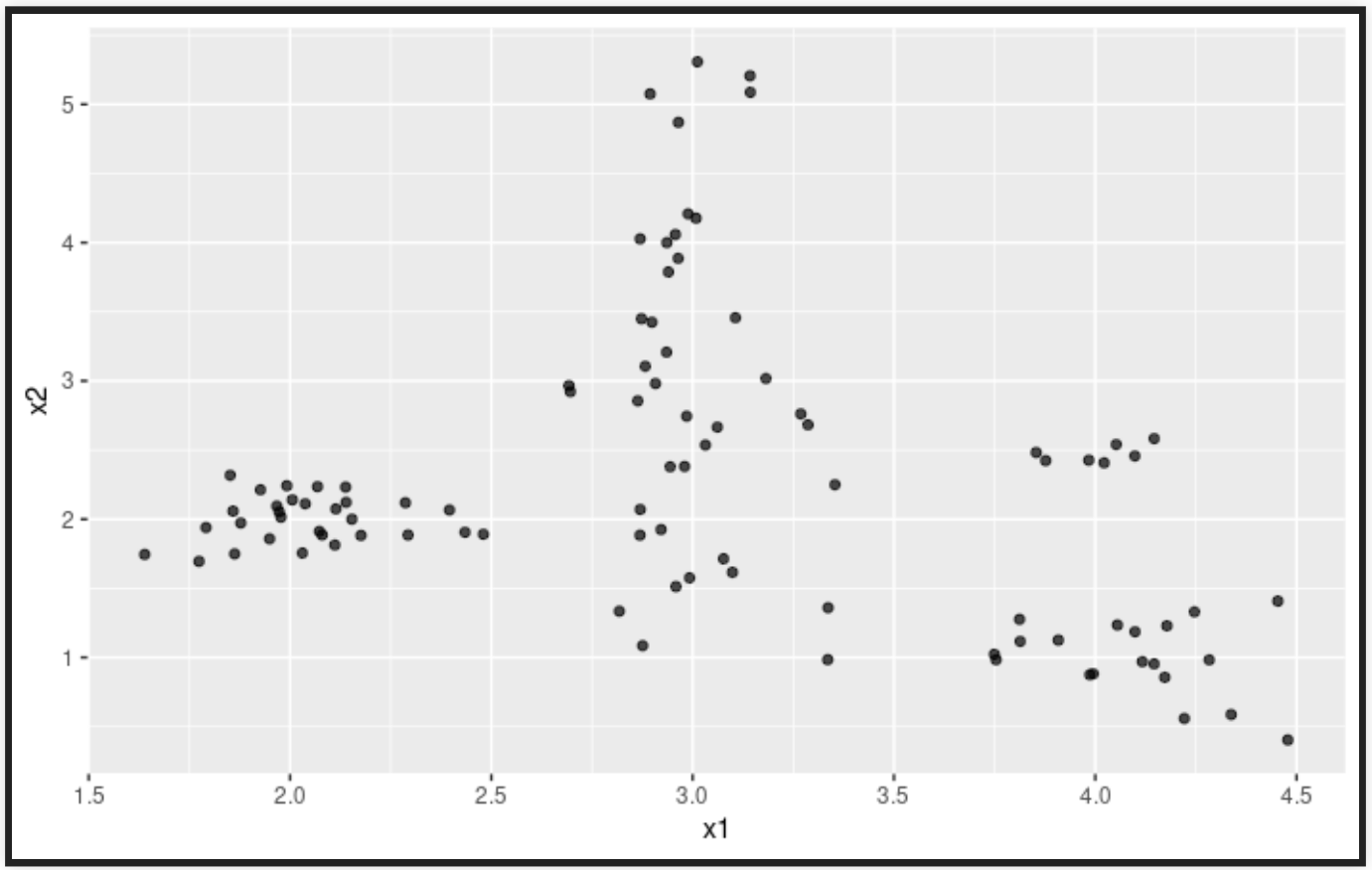
PROTOTYPES AND CRITICISMS

A prototype is a data instance that is representative of all the data.

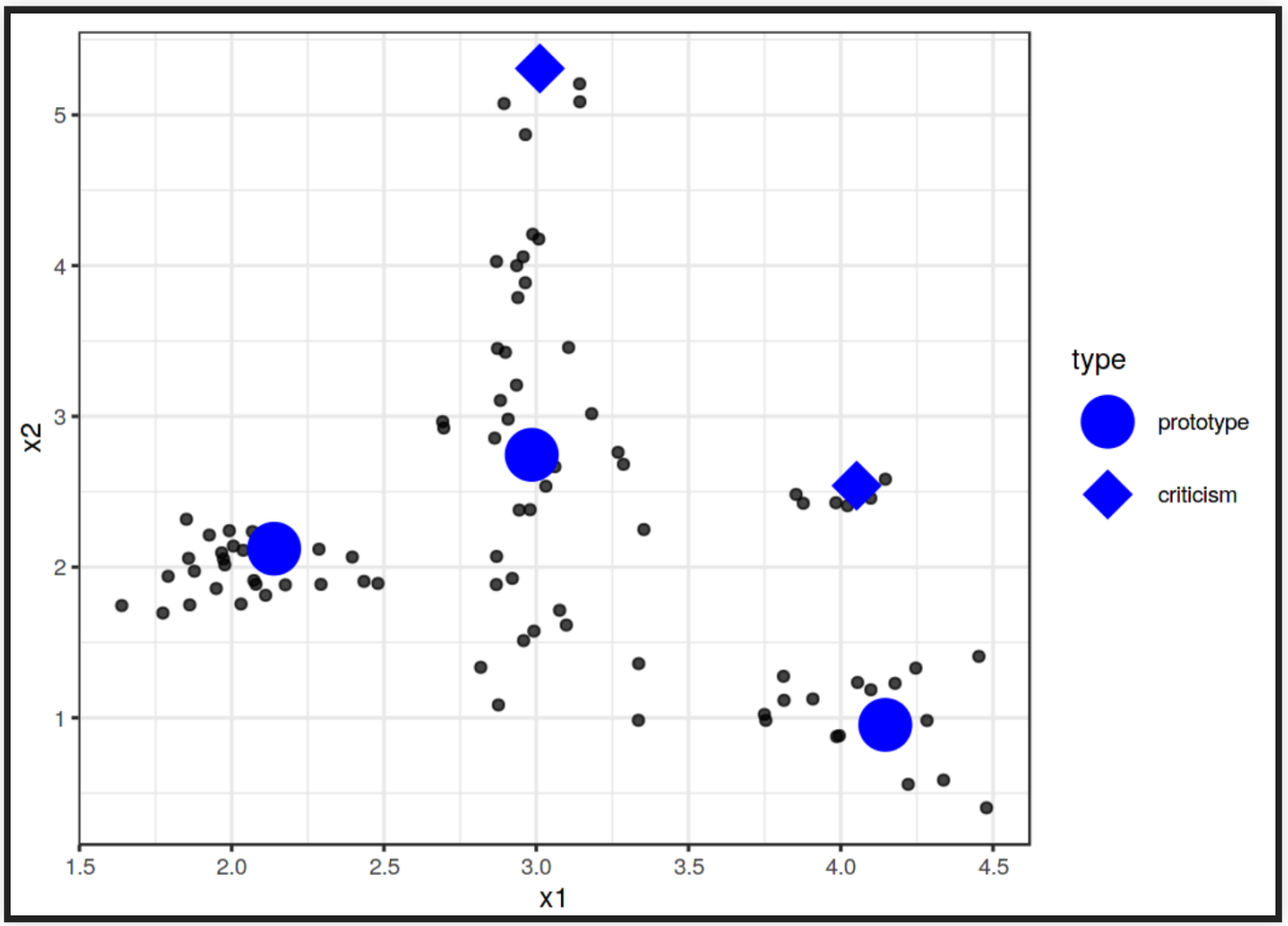
A criticism is a data instance that is not well represented by the set of prototypes.

How would you use this? (e.g., credit rating, cancer detection)

EXAMPLE: PROTOTYPES AND CRITICISMS?



EXAMPLE: PROTOTYPES AND CRITICISMS

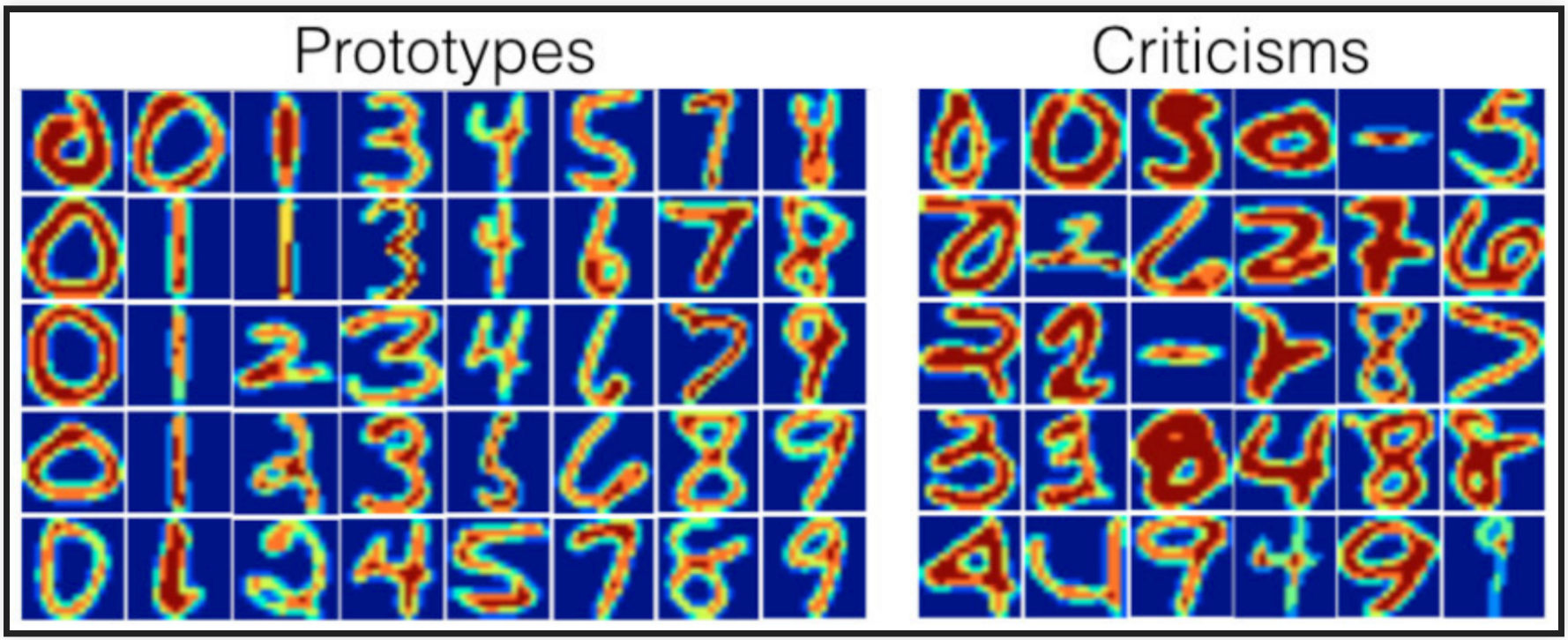


Source: Christoph Molnar. "Interpretable Machine Learning." 2019

EXAMPLE: PROTOTYPES AND CRITICISMS

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

EXAMPLE: PROTOTYPES AND CRITICISMS



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

- Usually identify number of prototypes and criticisms upfront
- 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

INFLUENTIAL INSTANCES

Data debugging!

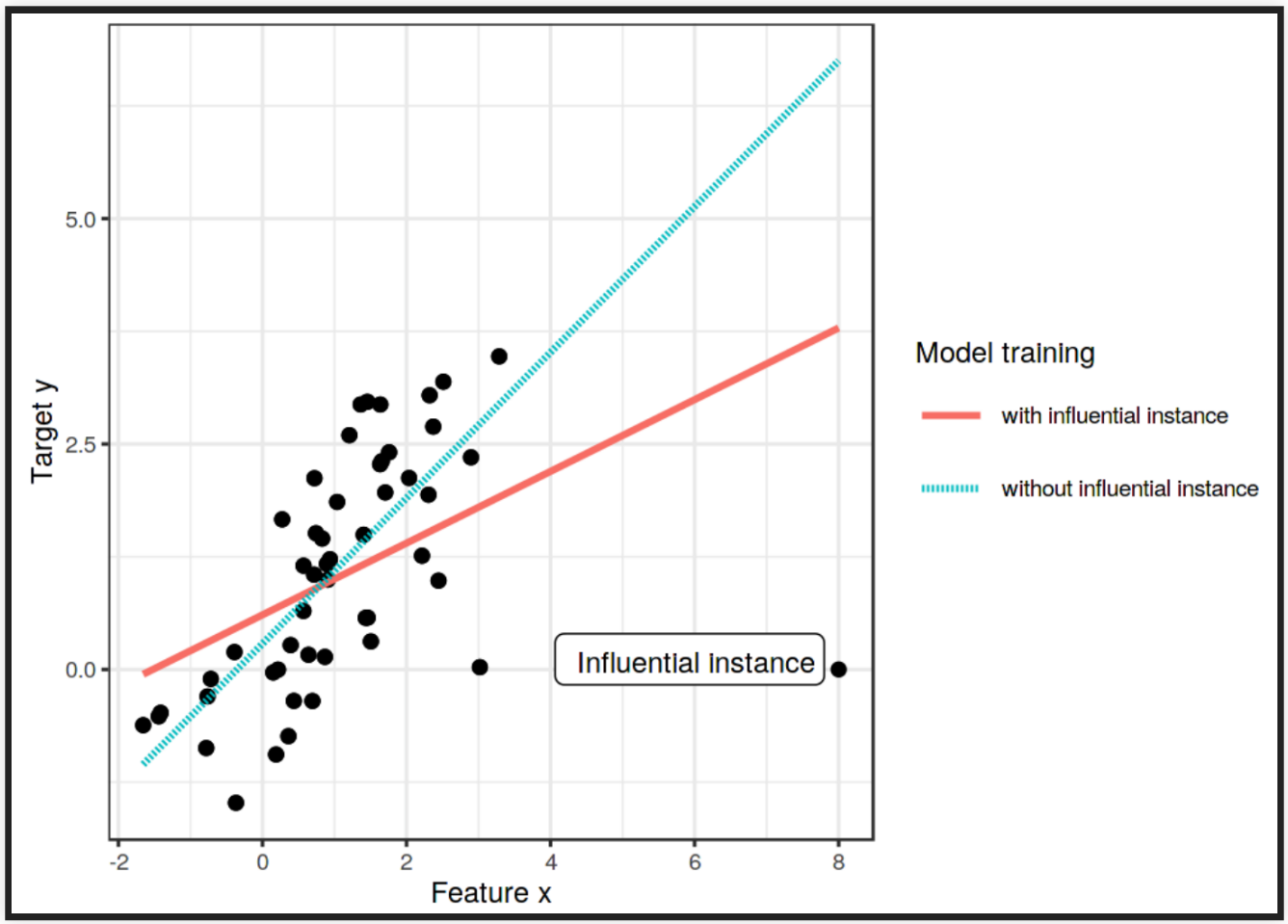
What data most influenced the training? Is the model skewed by few outliers?

- Training data with n instances
- Train model f with all n instances
- Train model g with $n - 1$ instances
- If f and g 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

EXAMPLE: INFLUENTIAL INSTANCE



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

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

- Useful model-agnostic debugging tool for models and data

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

WHAT MAKES GOOD EXPLANATIONS?

GOOD EXPLANATIONS ARE CONTRASTIVE

Counterfactuals. *Why this, rather than a different prediction?*

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

Partial explanations often sufficient in practice if contrastive

EXPLANATIONS ARE SELECTIVE

Often long or multiple explanations; parts are often sufficient

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

Your loan application has been declined. If your lived in Ohio your loan application would be accepted.

(Rashomon effect)

GOOD EXPLANATIONS ARE SOCIAL

Different audiences might benefit from different explanations

Accepted vs rejected loan applications?

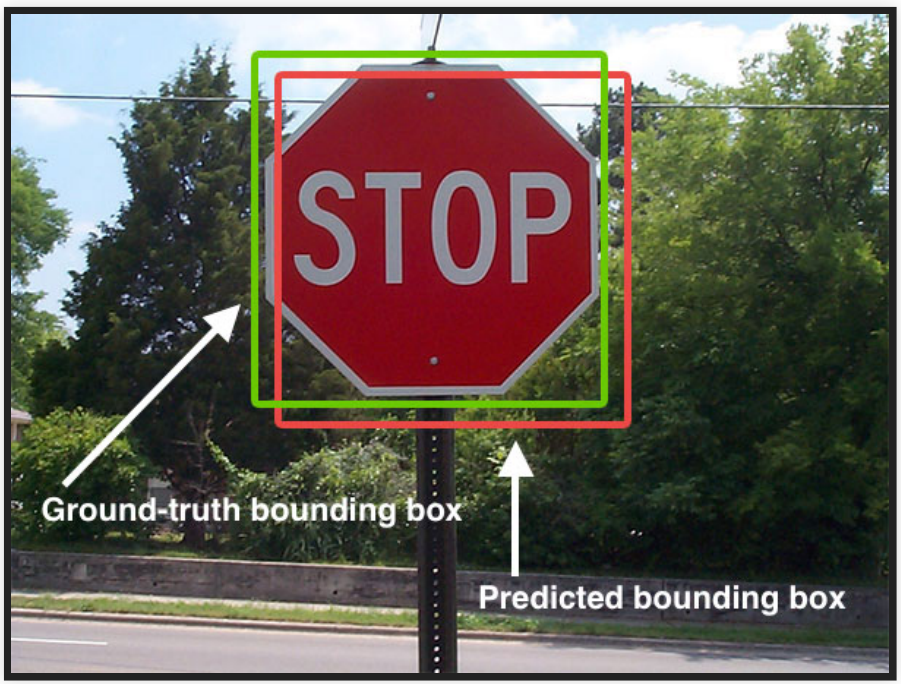
Explanation to customer or hotline support?

Consistent with prior belief of the explainee

EXERCISE: DEBUGGING A MODEL

Consider the following debugging challenges. In groups discuss which explainability tools may help and why. In 10 min report back to the group.

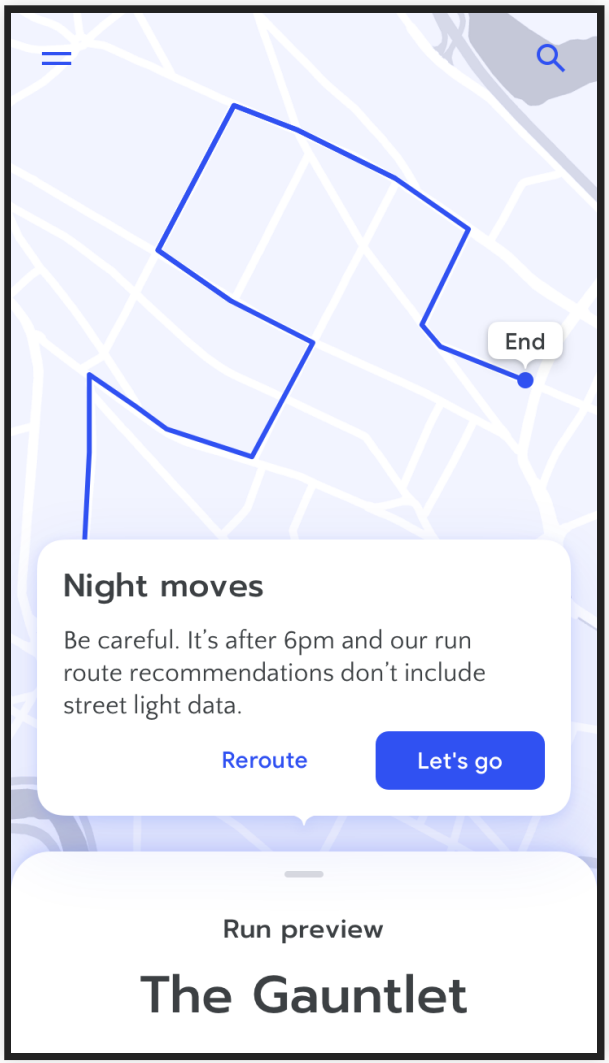
Algorithm bad at recognizing some signs in some conditions:

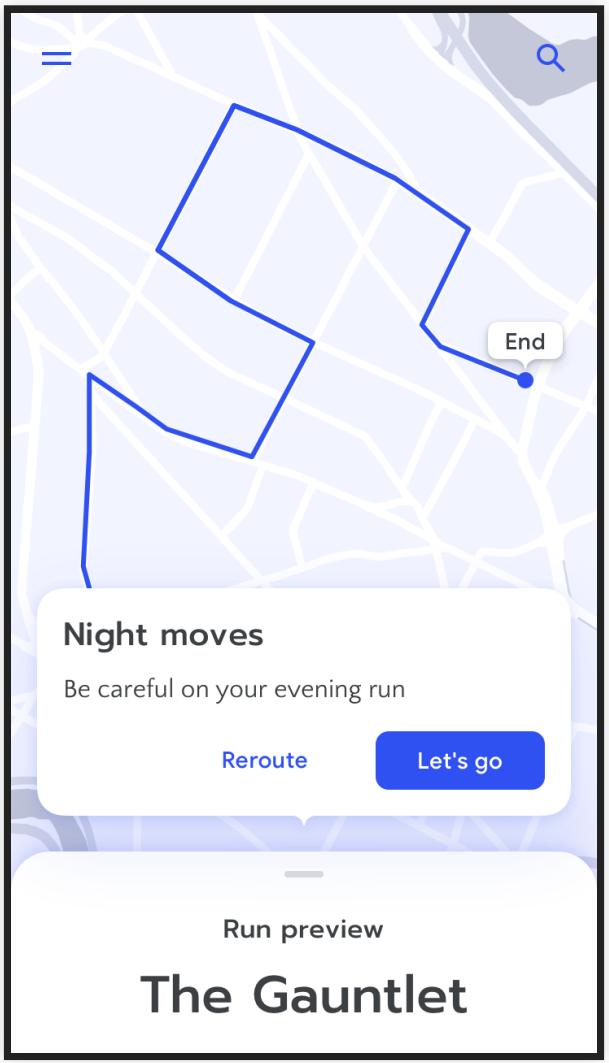


Graduate application system seems to rank applicants HBCUs lowly:

EXPLANATIONS AND USER INTERACTION DESIGN

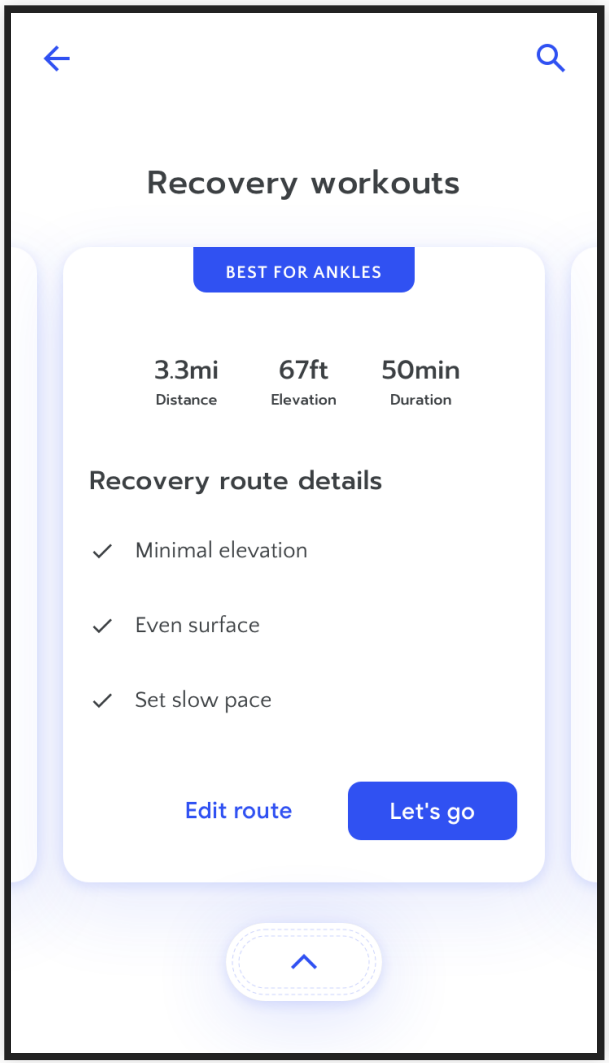
People + AI Guidebook, Google

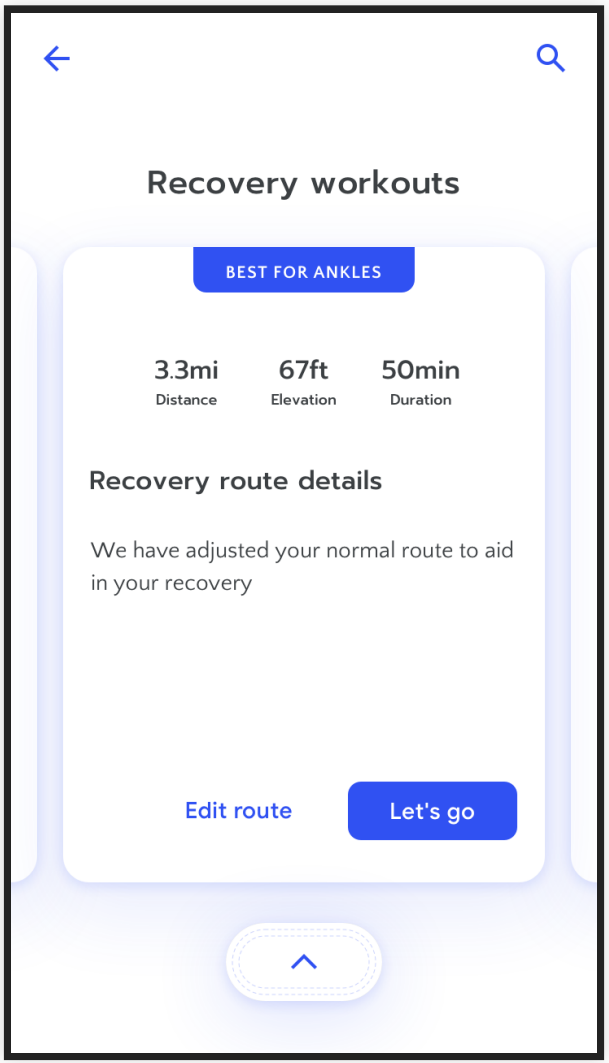




Tell the user when a lack of data might mean they'll need to use their own judgment. Don't be afraid to admit when a lack of data could affect the quality of the AI recommendations.

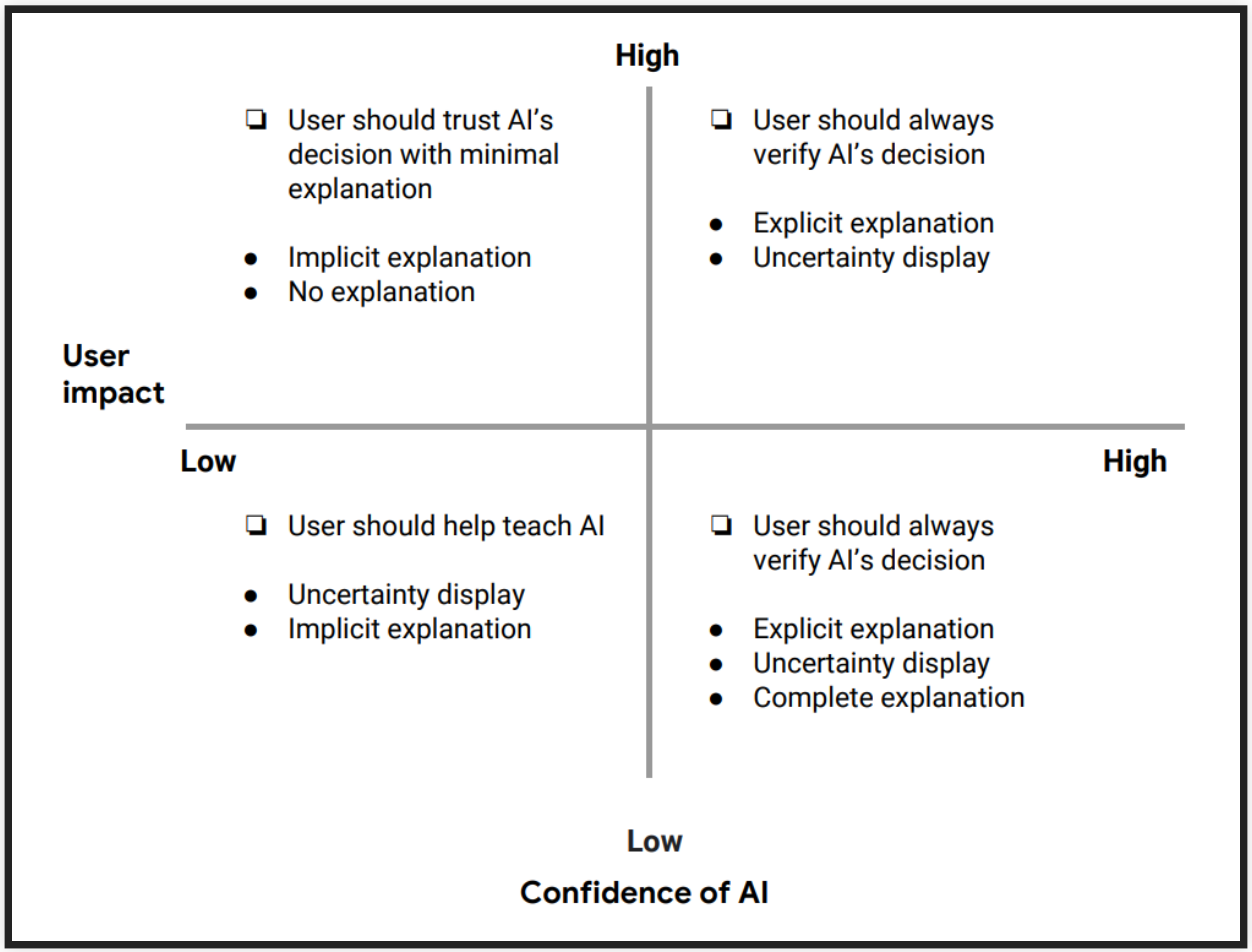
Source: People + AI Guidebook, Google





Give the user details about why a prediction was made in a high stakes scenario. Here, the user is exercising after an injury and needs confidence in the app's recommendation. Don't say "what" without saying "why" in a high stakes scenario.

Source: People + AI Guidebook, Google



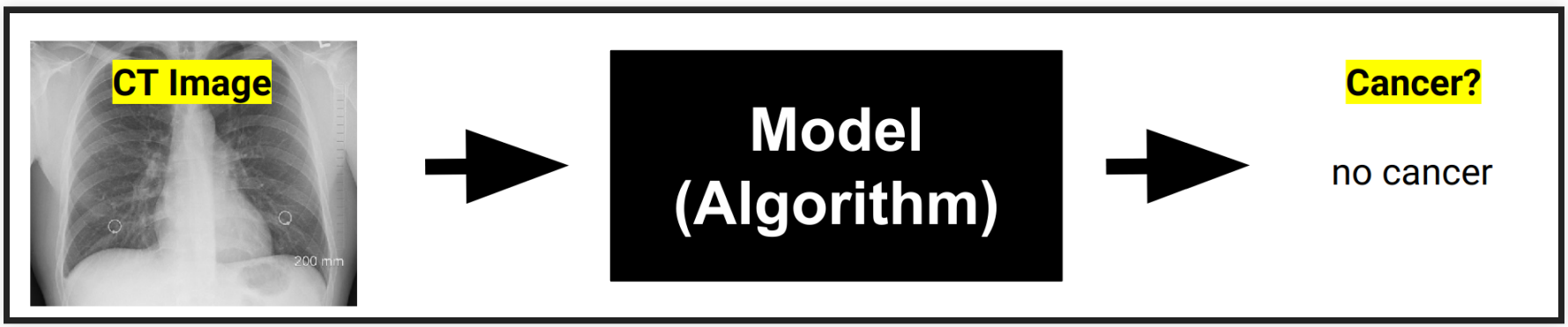
Example each?

Source: People + AI Guidebook, Google

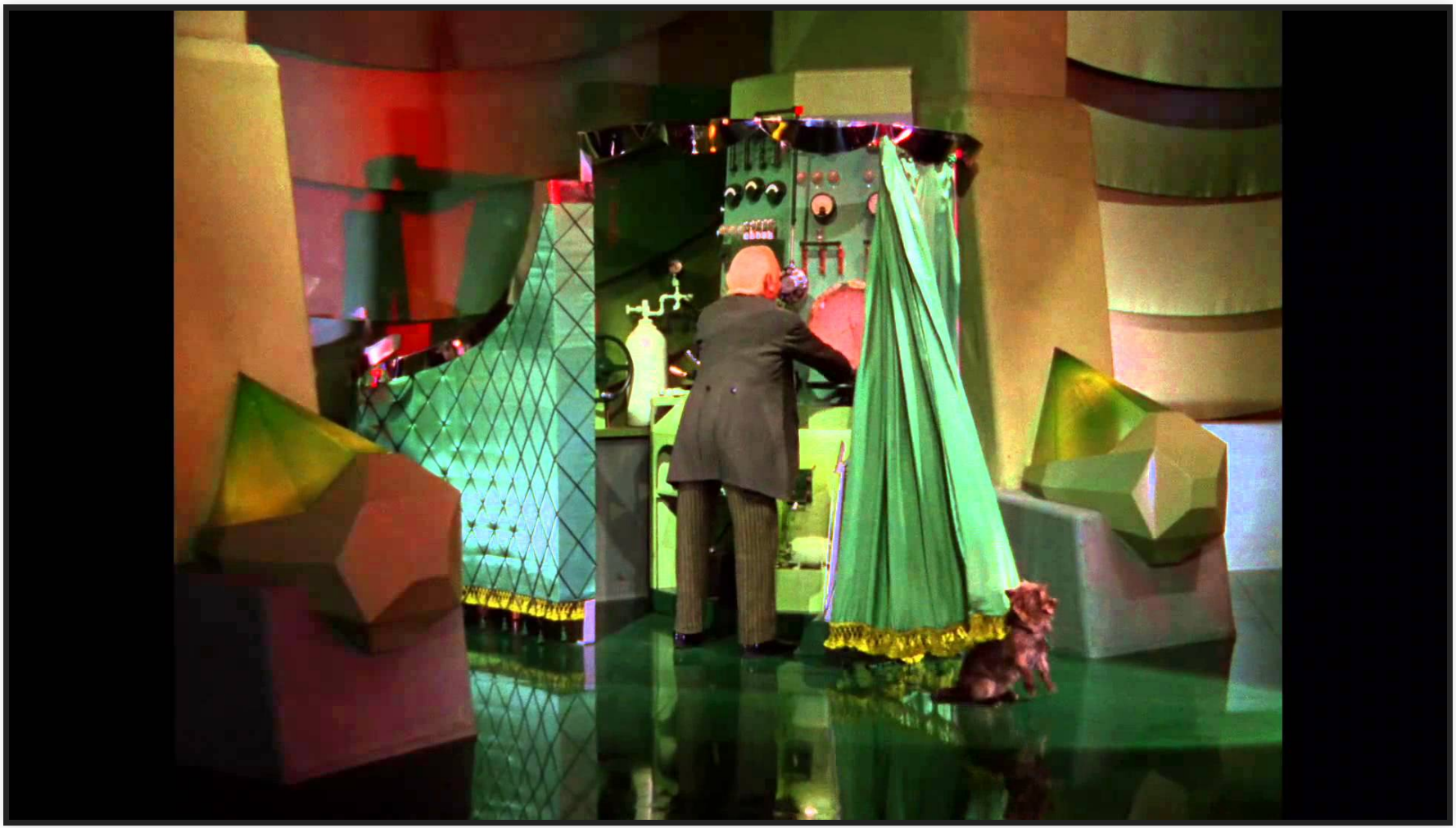
BEYOND "JUST" EXPLAINING THE MODEL

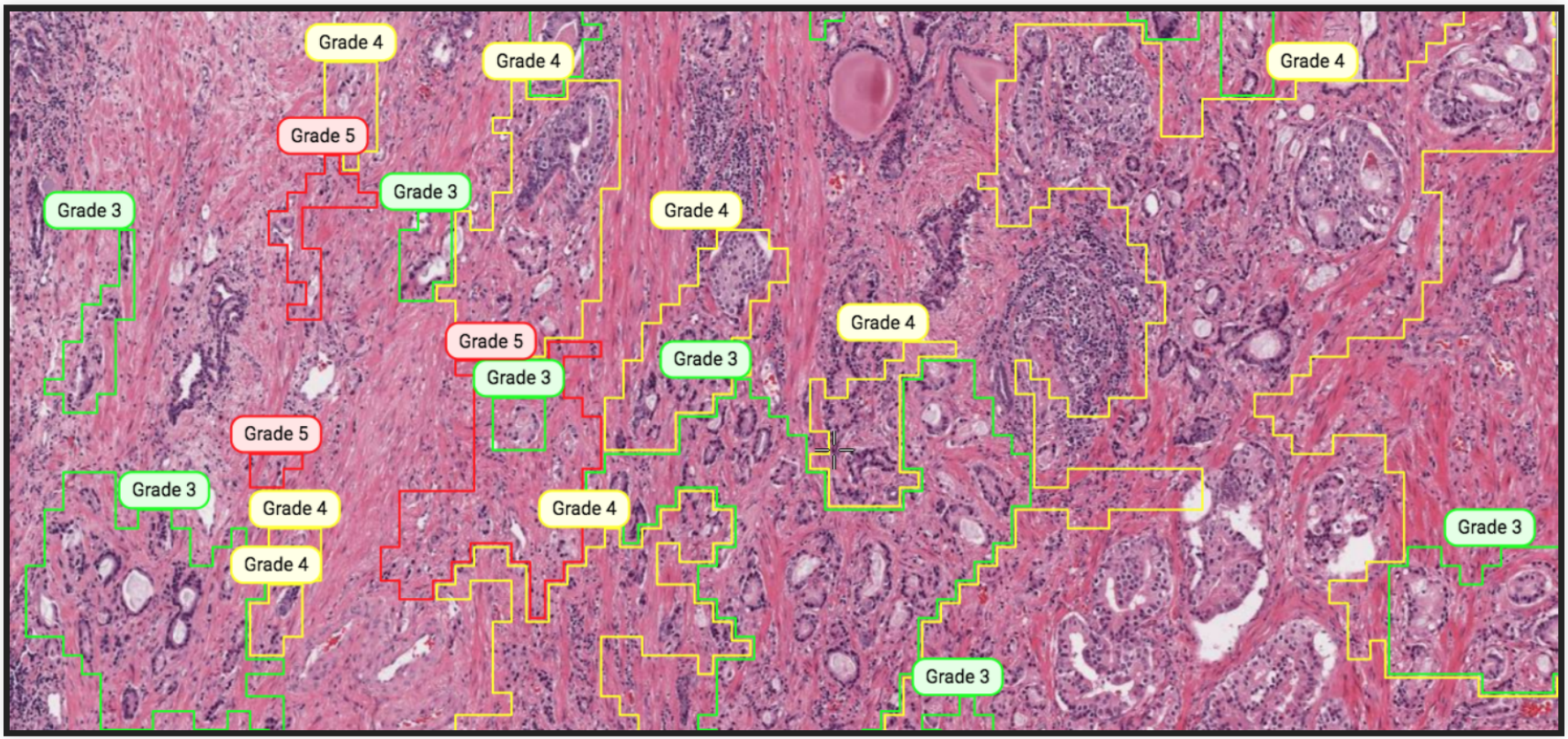
Cai, Carrie J., Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. ""Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making." Proceedings of the ACM on Human-computer Interaction 3, no. CSCW (2019): 1-24.

SETTING CANCER IMAGING -- WHAT EXPLANATIONS DO RADIOLOGISTS WANT?



- *Past attempts often not successful at bringing tools into production. Radiologists do not trust them. Why?*
- Wizard of oz study to elicit requirements




Grade 4
Grade 4
Grade 4
Grade 5
Grade 3
Grade 3
Grade 4
Grade 5
Grade 3
Grade 4
Grade 3
Grade 5
Grade 3
Grade 4
Grade 4
Grade 4
Grade 3
Grade 3

RADIOLOGISTS' QUESTIONS

- How does it perform compared to human experts?
- "What is difficult for the AI to know? Where is it too sensitive? What criteria is it good at recognizing or not good at recognizing?"
- What data (volume, types, diversity) was the model trained on?
- "Does the AI assistant have access to information that I don’t have? Does it have access to any ancillary studies?" Is all used data shown in the user interface?
- What kind of things is the AI looking for? What is it capable of learning? ("Maybe light and dark? Maybe colors? Maybe shapes, lines?", "Does it take into consideration the relationship between gland and stroma? Nuclear relationship?")
- "Does it have a bias a certain way?" (compared to colleagues)

RADIOLOGISTS' QUESTIONS

- Capabilities and limitations: performance, strength, limitations; e.g. how does it handle well-known edge cases
- Functionality: What data used for predictions, how much context, how data is used
- Medical point-of-view: calibration, how liberal/conservative when grading cancer severity
- Design objectives: Designed for few false positives or false negatives? Tuned to compensate for human error?
- Other considerations: legal liability, impact on workflow, cost of use

Paper, Tab 1

INSIGHTS

- AI literacy important for trust
- Be transparent about data used
- Describe training data and capabilities
- Give mental model, examples, human-relatable test cases
- Communicate the AI’s point-of-view and design goal

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.

THE DARK SIDE OF EXPLANATIONS

MANY EXPLANATIONS ARE WRONG

- Approximations of black-box models, often unstable
- Explanations necessarily partial, social
- Often multiple explanations possible (Rashomon effect)

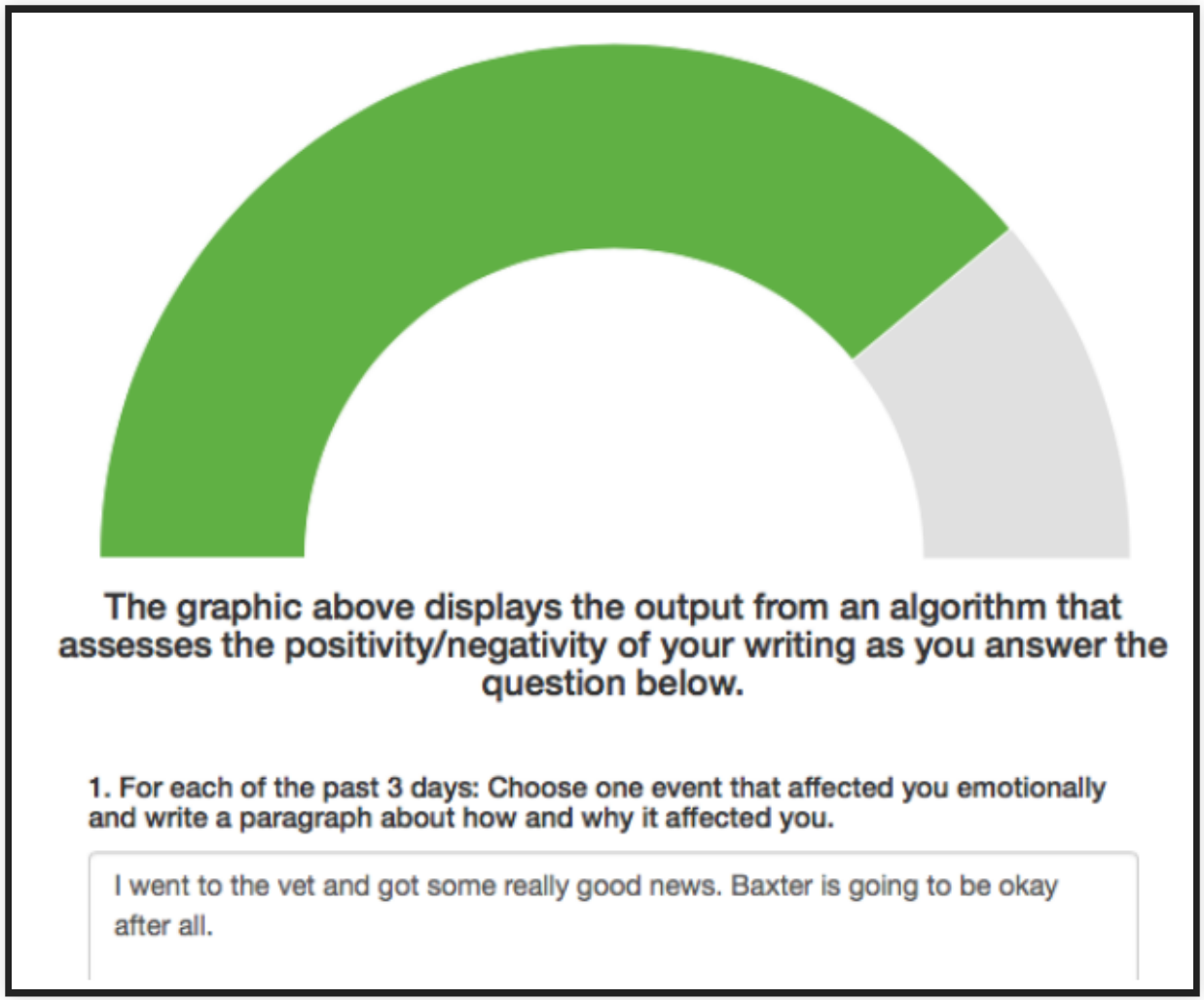
- Possible to use inherently interpretable models instead?
- When explanation desired/required: What quality standard acceptable?

EXPLANATIONS FOSTER TRUST

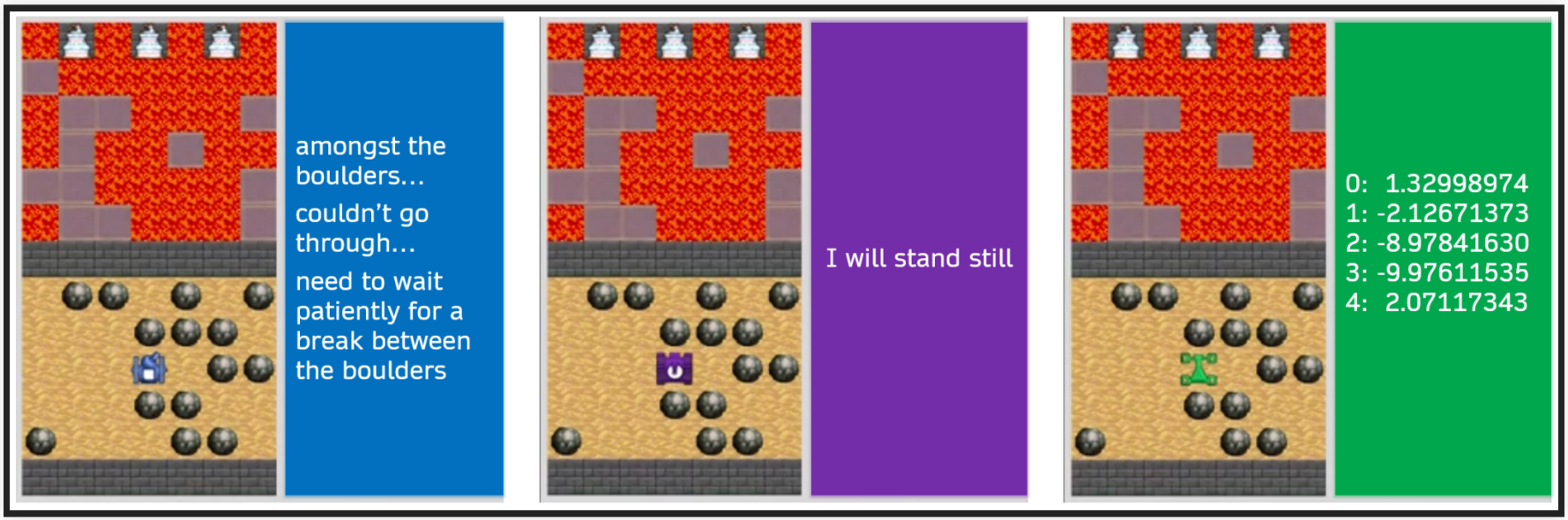
- Users are less likely to question the model
- Even if explanations are unreliable
- Even if explanations are nonsensical/incomprehensible

Danger of overtrust and intentional manipulation

Stumpf, Simone, Adrian Bussone, and Dympna O'sullivan. "Explanations considered harmful? user interactions with machine learning systems." In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI). 2016.



Springer, Aaron, Victoria Hollis, and Steve Whittaker. "Dice in the black box: User experiences with an inscrutable algorithm." In 2017 AAAI Spring Symposium Series. 2017.



(a) Rationale, (b) Stating the prediction, (c) Numerical internal values

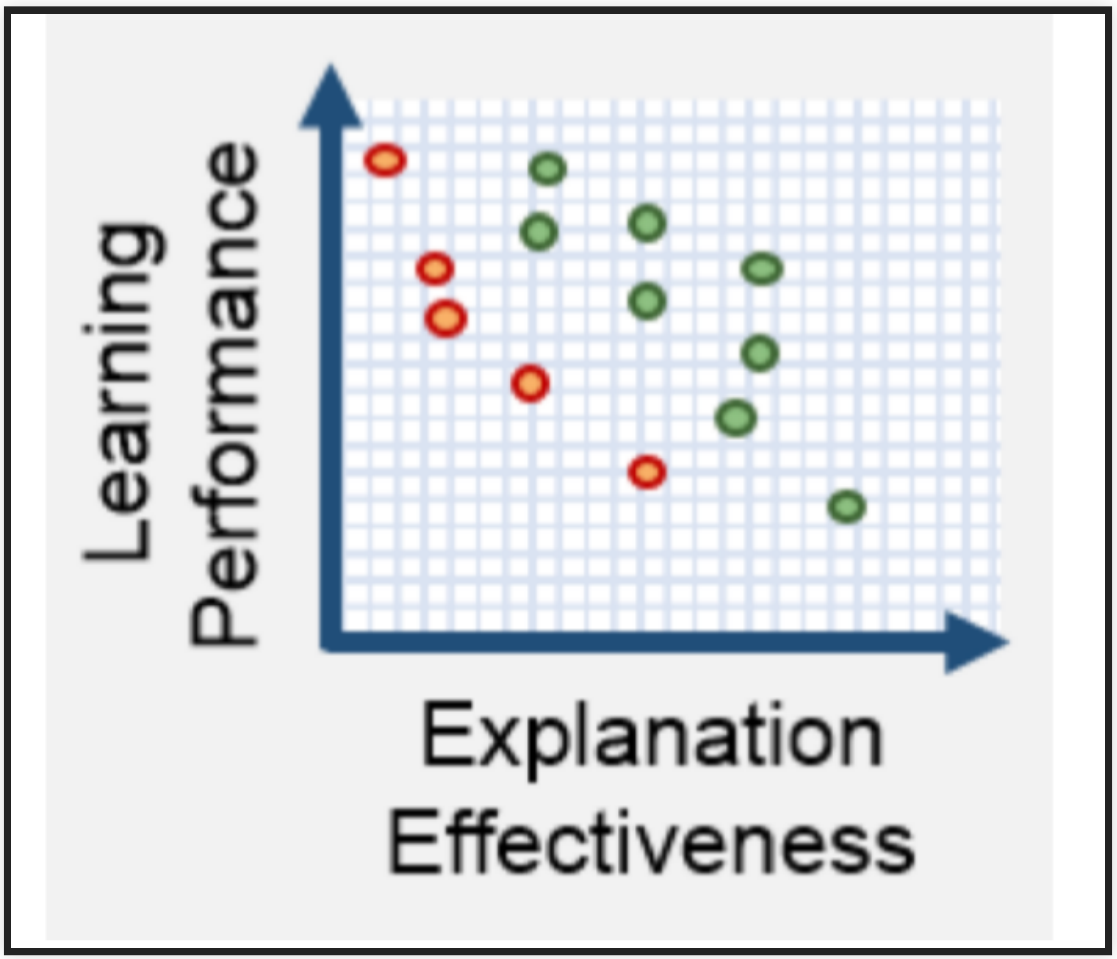
Observation: Both experts and non-experts overtrust numerical explanations, even when inscrutable.

Ehsan, Upol, Samir Passi, Q. Vera Liao, Larry Chan, I. Lee, Michael Muller, and Mark O. Riedl. "The who in explainable AI: how AI background shapes perceptions of AI explanations." arXiv preprint arXiv:2107.13509 (2021).

"STOP EXPLAINING BLACK BOX MACHINE LEARNING MODELS FOR HIGH STAKES DECISIONS AND USE INTERPRETABLE MODELS INSTEAD."

Cynthia Rudin (32min) or 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.

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 arrest
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 BLACK BOX MACHINE LEARNING MODELS FOR HIGH STAKES DECISIONS AND USE INTERPRETABLE MODELS INSTEAD"

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)

INTERPRETABLE MODELS VS POST-HOC EXPLANATIONS

- High-stakes decisions
  - interpretable models provide faithful explanations
  - post-hoc explanations may provide limited insights or illusion of understanding
  - interpretable models can be audited
- In many cases similar accuracy
- Larger focus on feature engineering, but insights into when and *why* the model works
  - exploratory data analysis, plots, association rule mining
  - more effort for building interpretable models (especially beyond well structured tabular data)
- Less research on interpretable models and some methods computationally expensive
  - additional constraints on model form for interpretability limit degrees of freedom: sparseness, parameters with easy to read weights, ...

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 blackbox models
    - global and local surrogates
    - dependence plots and feature importance
    - anchors
    - counter-factual explanations
- Data debugging with prototypes, criticisms, and influential instances
- Consider implications on user interface design
- Gaming and manipulation with explanations