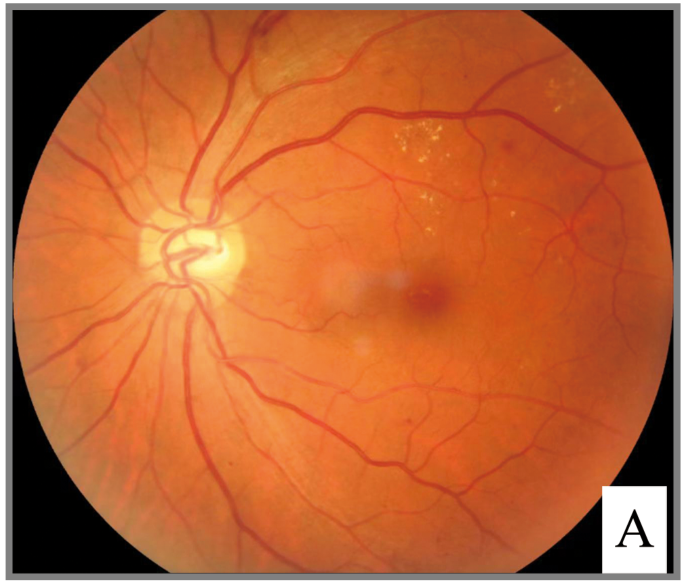
Machine Learning in Production

Explainability and Interpretability

Administrativa

Homework I4 to be released Monday; 1 week assignment, due Apr 17



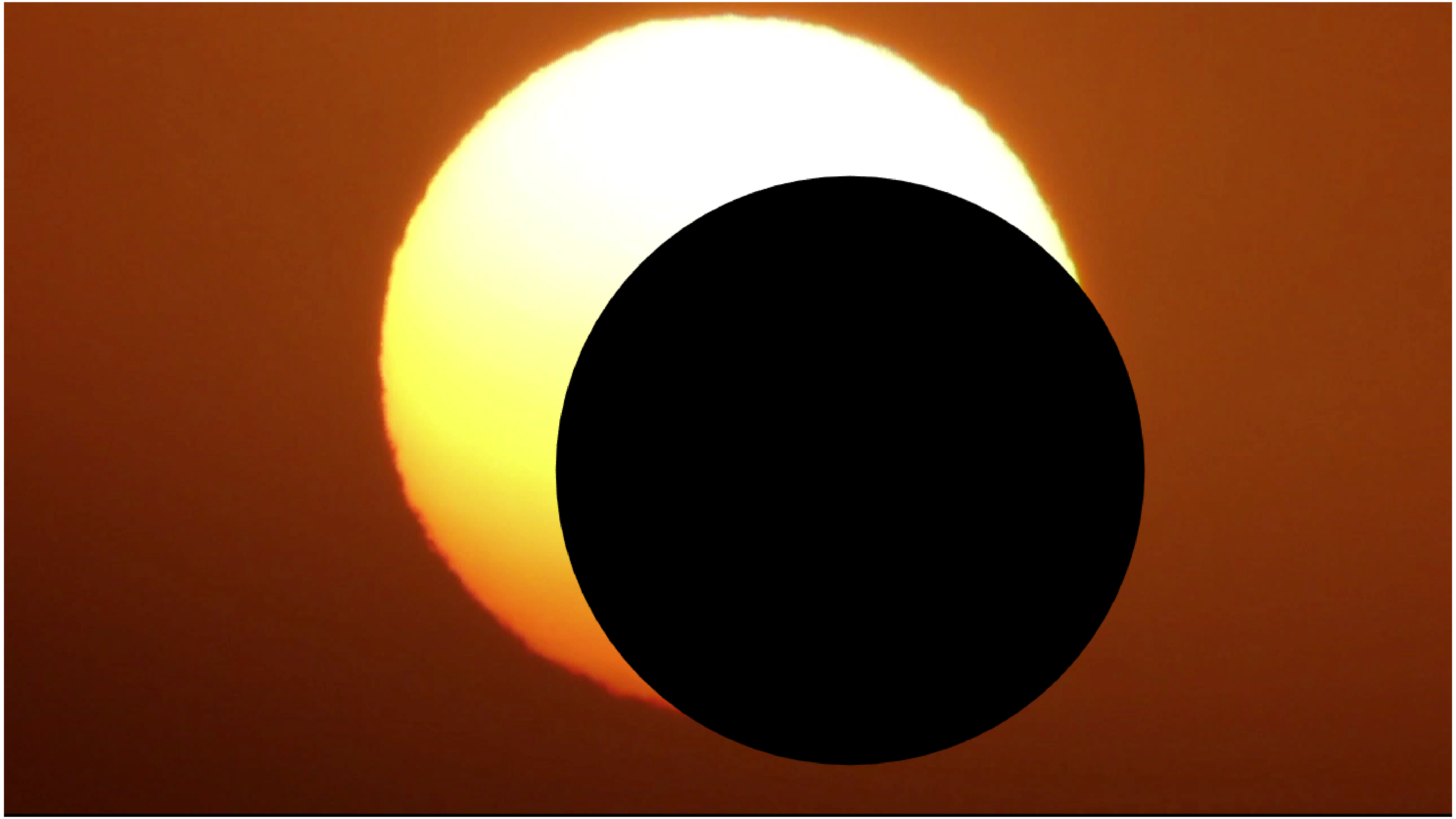
Research in this Course (in I4)

We are conducting academic research on explainability policies and evidence. This research will involve analyzing student work of this assignment. You will not be asked to do anything above and beyond the normal learning activities and assignments that are part of this course. You are free not to participate in this research, and your participation will have no influence on your grade for this course or your academic career at CMU. If you do not wish to participate, please send an email to Nadia Nahar (nadian@andrew.cmu.edu). Participants will not receive any compensation or extra credit. The data collected as part of this research will not include student grades. All analyses of data from participants' coursework will be conducted after the course is over and final grades are submitted -- instructors will not know who chooses not to participate before final grades are submitted. All data will be analyzed in de-identified form and presented in the aggregate, without any personal identifiers. If you have questions pertaining to your rights as a research participant, or to report concerns to this study, please contact Nadia Nahar (nadian@andrew.cmu.edu) or the Office of Research Integrity and Compliance at Carnegie Mellon University (irb-review@andrew.cmu.edu; phone: 412-268-4721).

Milestone 3: Availability Requirement

10pt: The recommendation service is at least 70% available in the 72 hours before the submission and the 96 hours after (i.e., max downtime of 50h), while at least two updates are performed in that time period.

5pt: Bonus points if the recommendation service is at least 99% available in the same 7-day window (max 100min downtime), while at least two updates are performed in that time period.



TAing Next Semester?

Explainability as Building Block in Responsible Engineering

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:
System and model goals
User requirements
Environment assumptions
Quality beyond accuracy
Measurement
Risk analysis
Planning for mistakes

Architecture + design:
Modeling tradeoffs
Deployment architecture
Data science pipelines
Telemetry, monitoring
Anticipating evolution
Big data processing
Human-AI design

Quality assurance:
Model testing
Data quality
QA automation
Testing in production
Infrastructure quality
Debugging

Operations:
Continuous deployment
Contin. experimentation
Configuration mgmt.
Monitoring
Versioning
Big data
DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance, versioning, reproducibility

Safety

Security and privacy

Fairness

Interpretability and explainability

Transparency and trust

Ethics, governance, regulation, compliance, organizational culture

"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

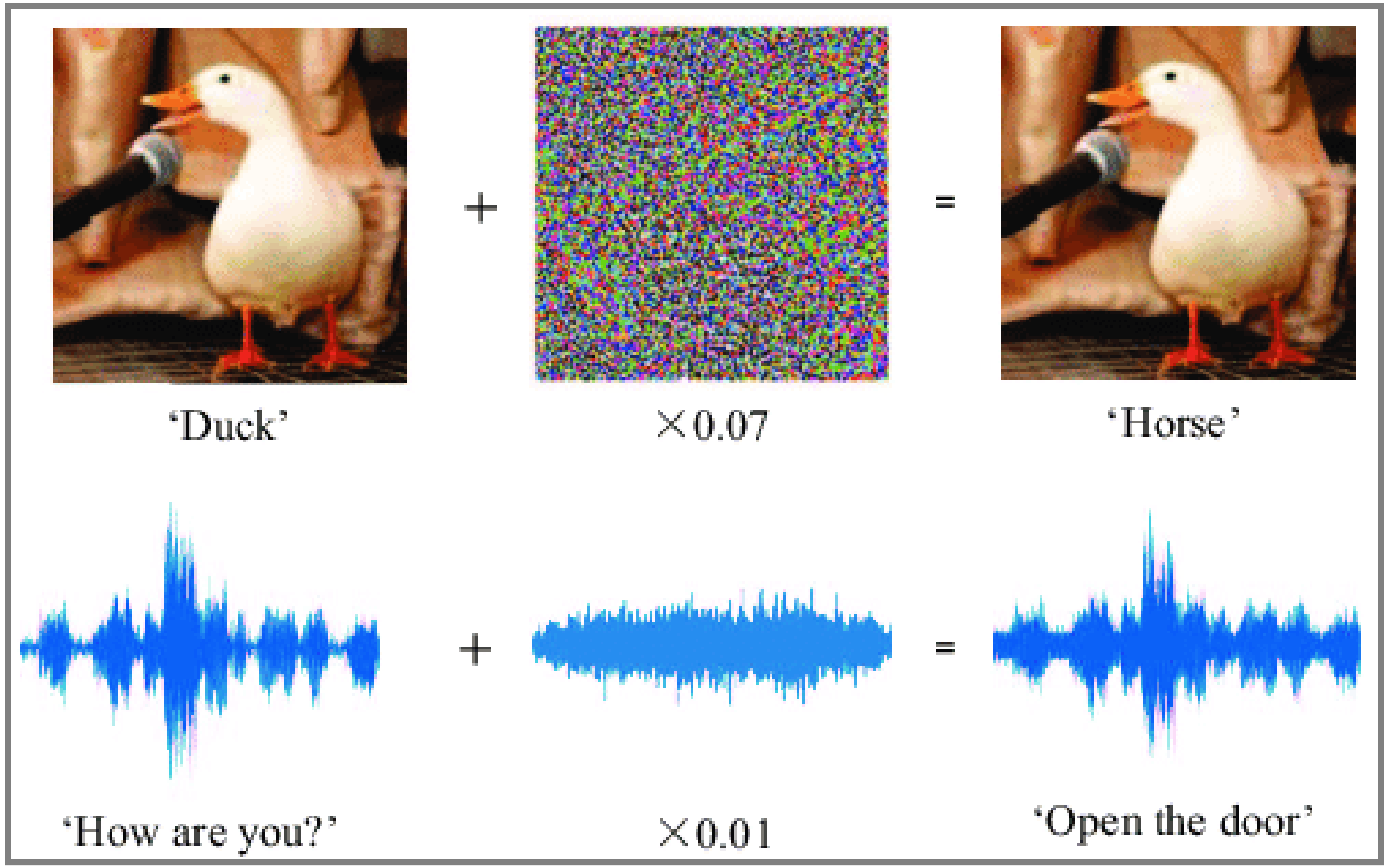
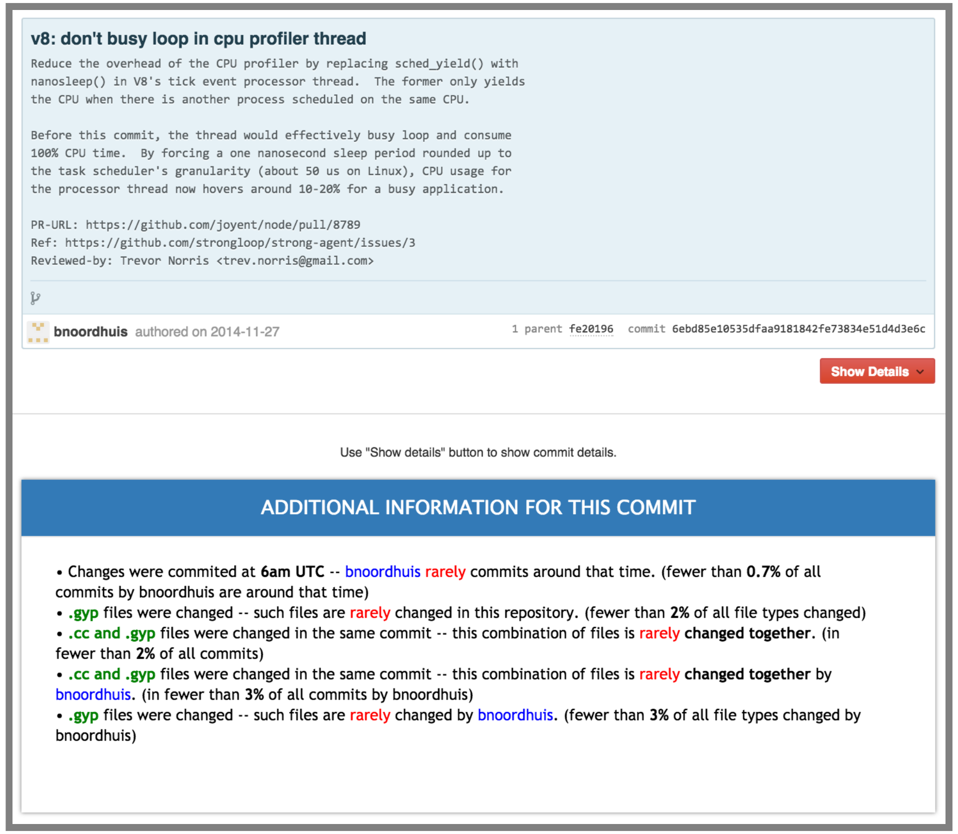


Image: Gong, Yuan, and Christian Poellabauer. "An overview of vulnerabilities of voice controlled systems." arXiv preprint arXiv:1803.09156 (2018).

Detecting Anomalous Commits



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?

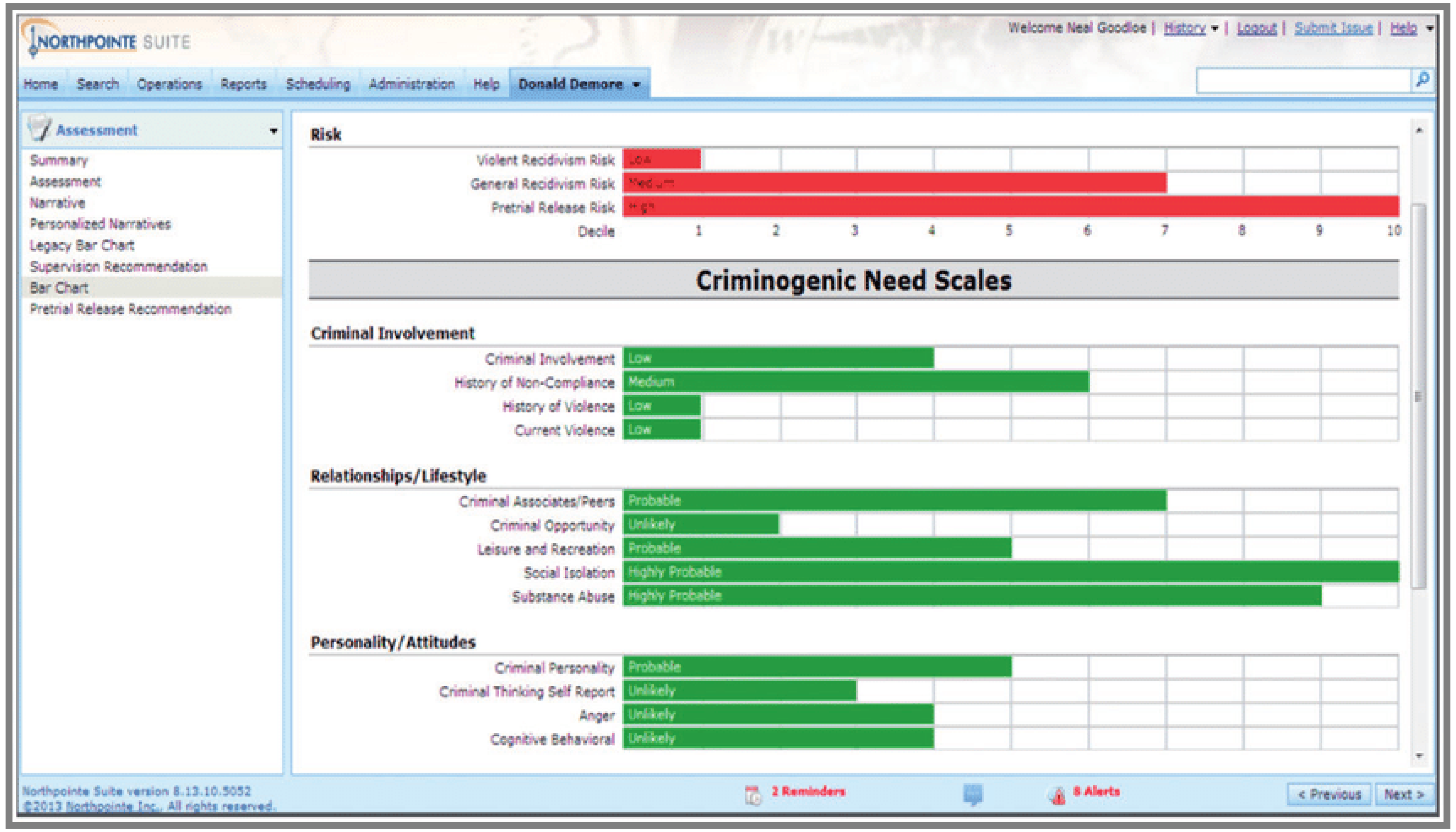


Image source (CC BY-NC-ND 4.0): Christin, Angèle. (2017). Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society. 4.

How to consider seriousness of the crime?

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

PREDICT ARREST FOR ANY OFFENSE IF SCORE $>$ 1

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

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

≡

Is there an actual problem? How to find out?

DHH
@dhh · Follow

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple’s black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

8:34 PM · Nov 7, 2019

24.5K Reply Copy link to post

Read 1.2K replies

DHH · Nov 8, 2019
@dhh · **Follow**
Replying to @dhh

I wasn't even pessimistic to expect this outcome, but here we are: @AppleCard just gave my wife the VIP bump to match my credit limit, but continued to be an utter fucking failure of a customer service experience. Let me explain...

DHH
@dhh · **Follow**

She spoke to two Apple reps. Both very nice, courteous people representing an utterly broken and reprehensible system. The first person was like "I don't know why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".

11:20 PM · Nov 8, 2019

4K Reply Copy link to post

Read 57 replies

Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

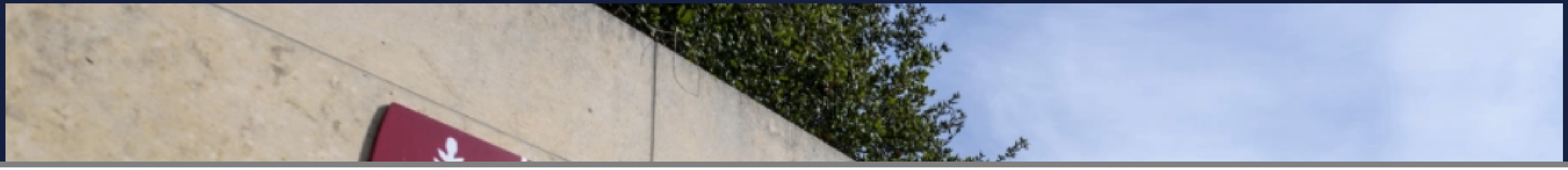
PANDEMIC TECHNOLOGY PROJECT

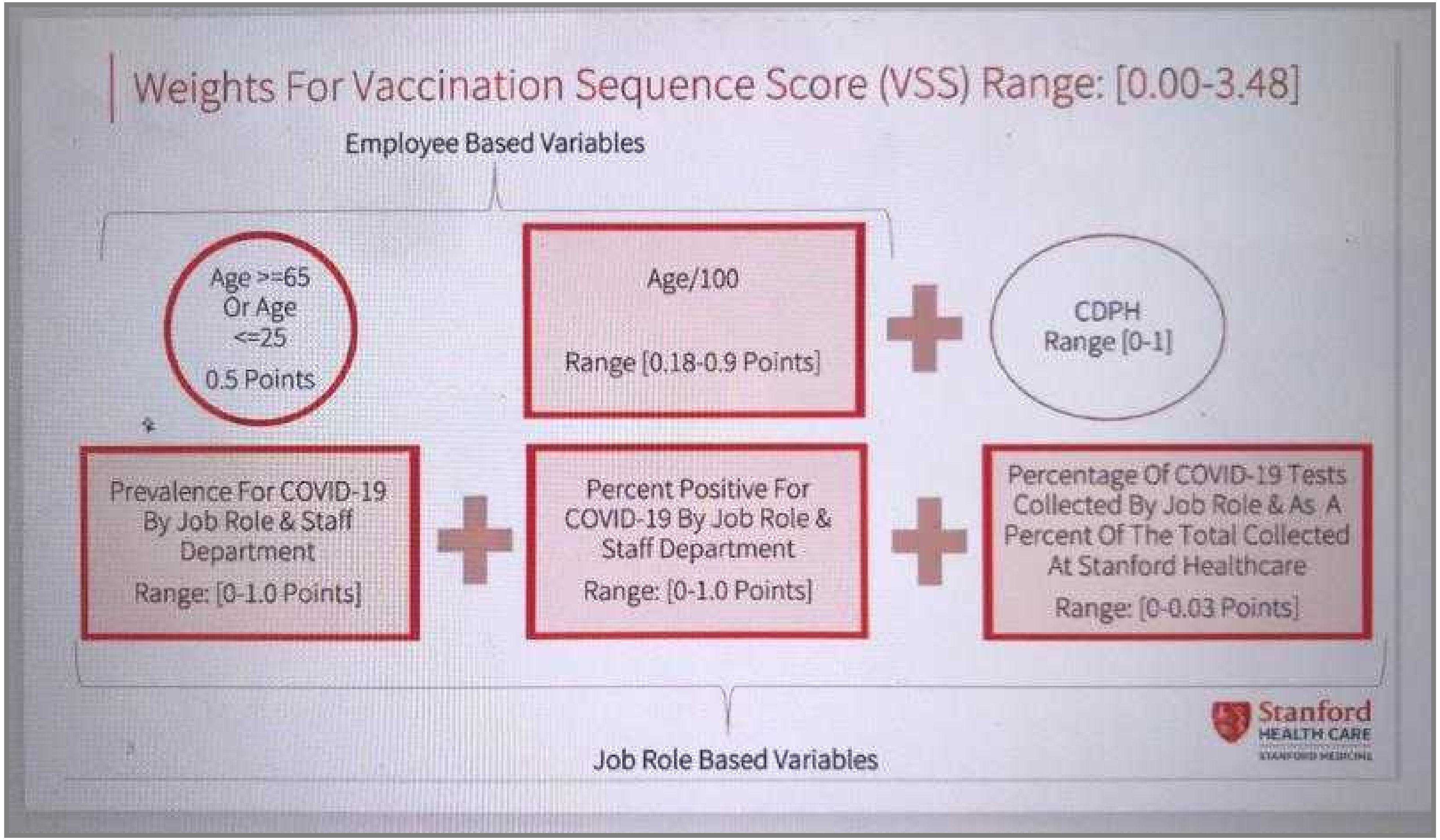
This is the Stanford vaccine algorithm that left out frontline doctors

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

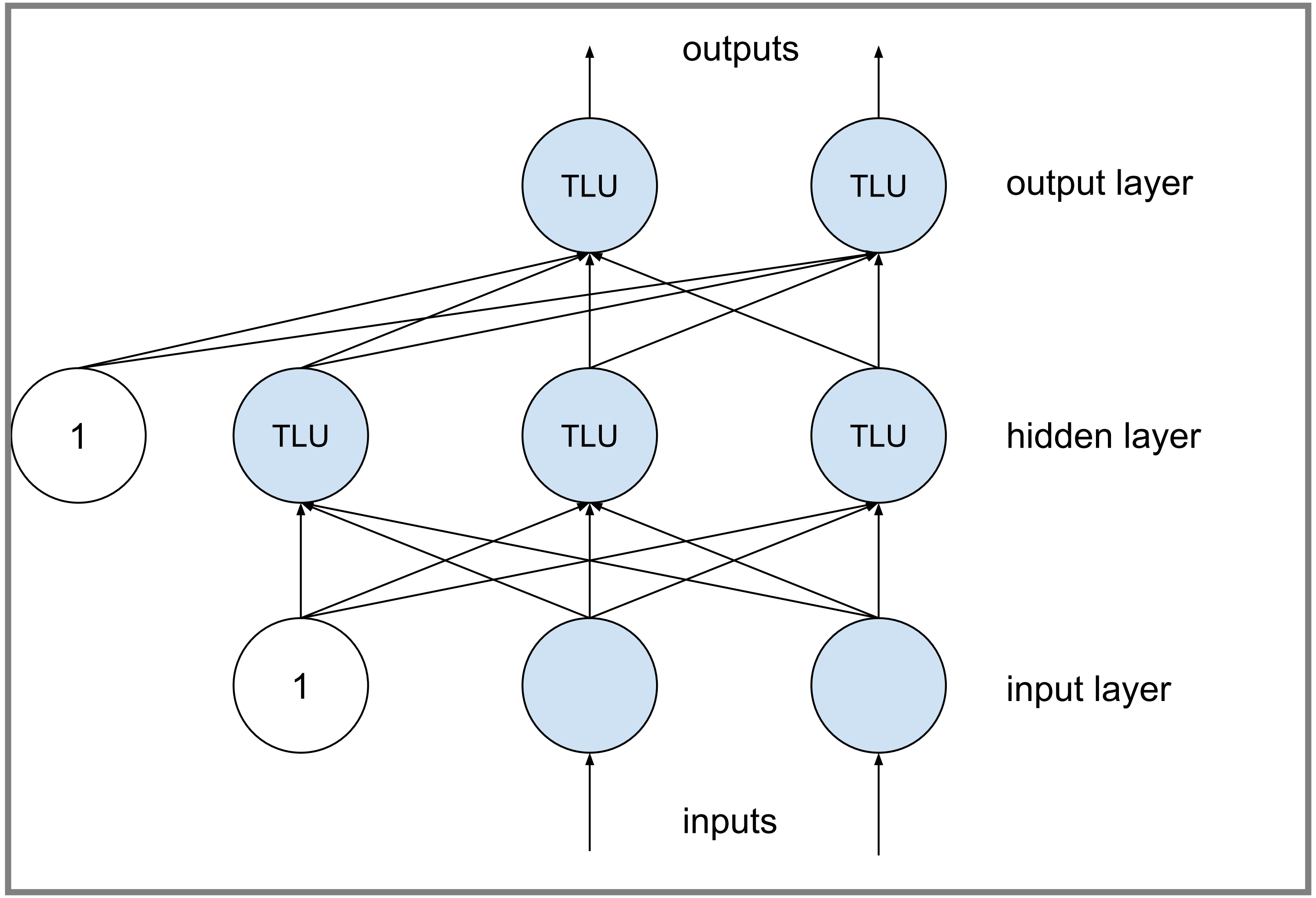



Weights For Vaccination Sequence Score (VSS) Range: [0.00-3.48]
Employee Based Variables
Age >=65 Or Age <=25
0.5 Points
Age/100
Range [0.18-0.9 Points]
CDPH Range [0-1]
Prevalence For COVID-19 By Job Role & Staff Department
Range: [0-1.0 Points]
Percent Positive For COVID-19 By Job Role & Staff Department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

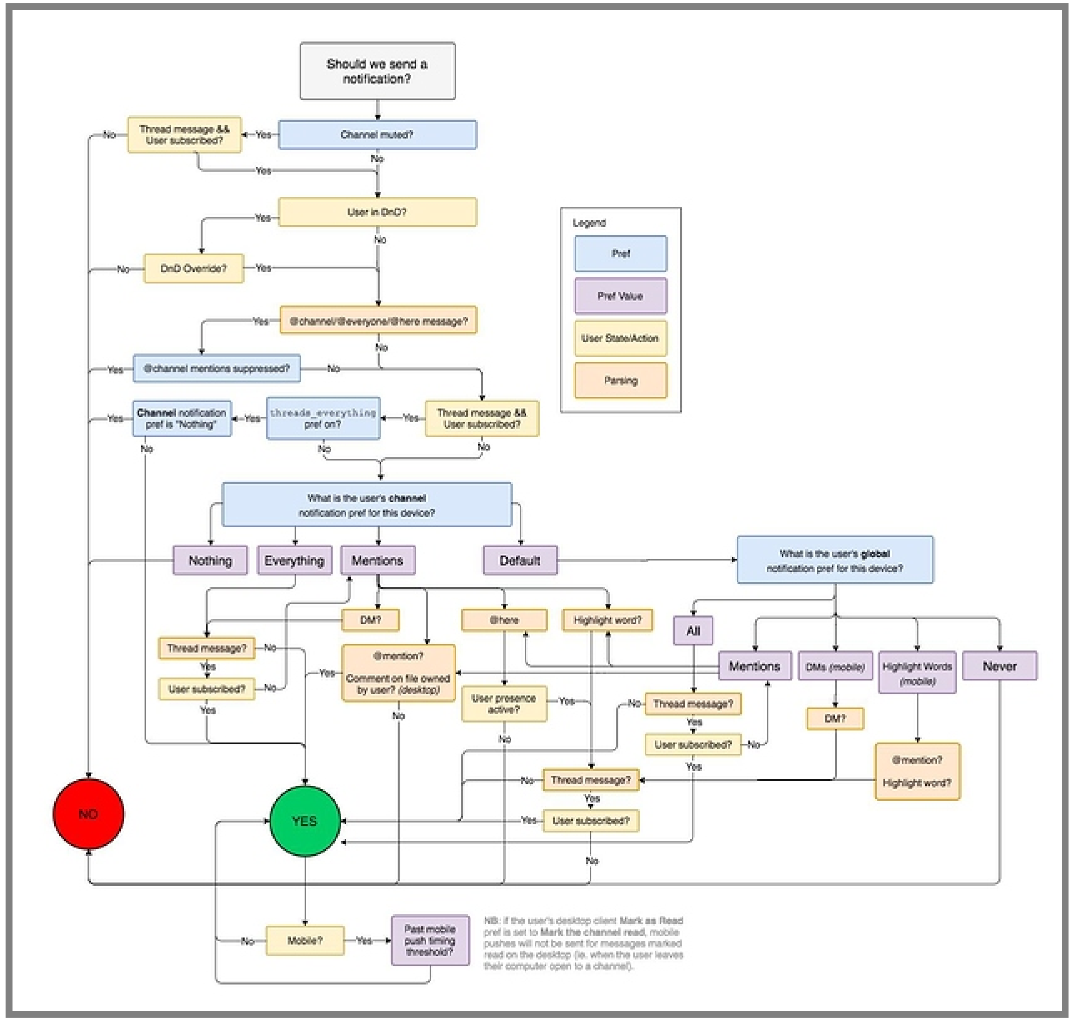
Explaining Decisions

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 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

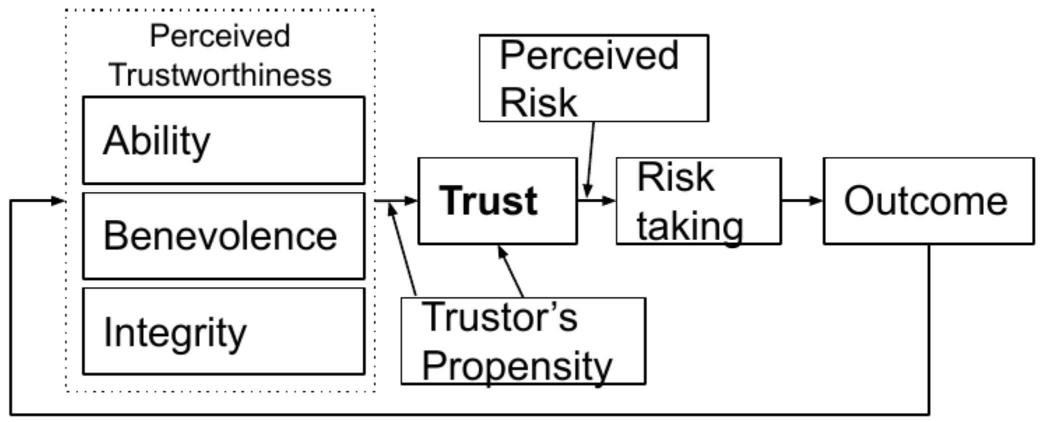
- Understand safety implications
- Ensure predictions use objective criteria and reasonable rules
- Inspect fairness properties
- Reason about biases and feedback loops
- Validate "learned specifications/requirements" with stakeholders

```
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."*

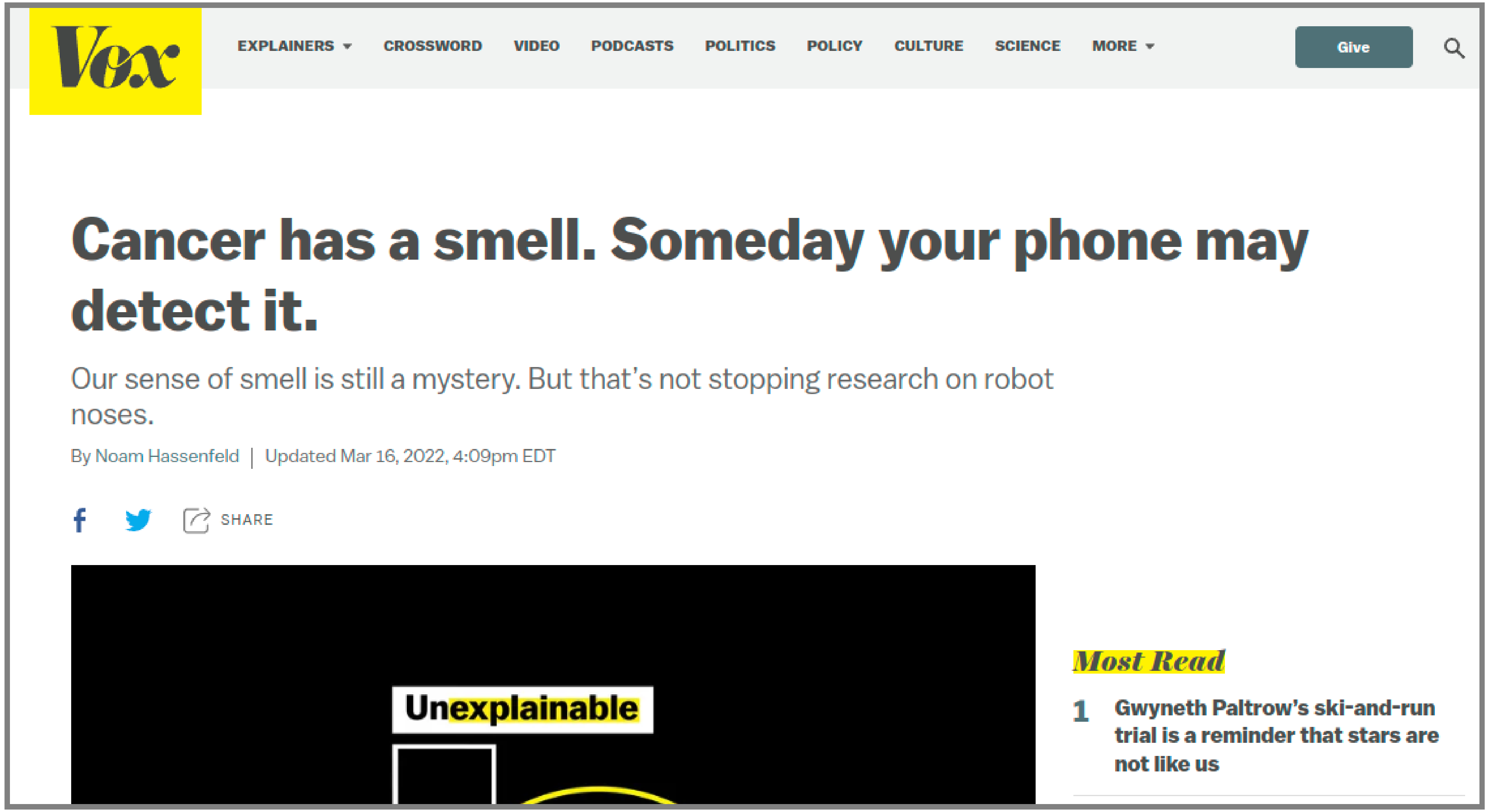
See also https://en.wikipedia.org/wiki/Right_to_explanation

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*) $R_m^2 = 0.04$, $R_c^2 = 0.35$	
	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



Vox

EXPLAINERS CROSSWORD VIDEO PODCASTS POLITICS POLICY CULTURE SCIENCE MORE

Give

Cancer has a smell. Someday your phone may detect it.

Our sense of smell is still a mystery. But that's not stopping research on robot noses.

By Noam Hassenfeld | Updated Mar 16, 2022, 4:09pm EDT

SHARE

Unexplainable

Most Read

1 Gwyneth Paltrow's ski-and-run trial is a reminder that stars are not like us

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

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

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, 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 science question)

How would you measure explanation quality?

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

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

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		⋯
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%

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
```

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.

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

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?

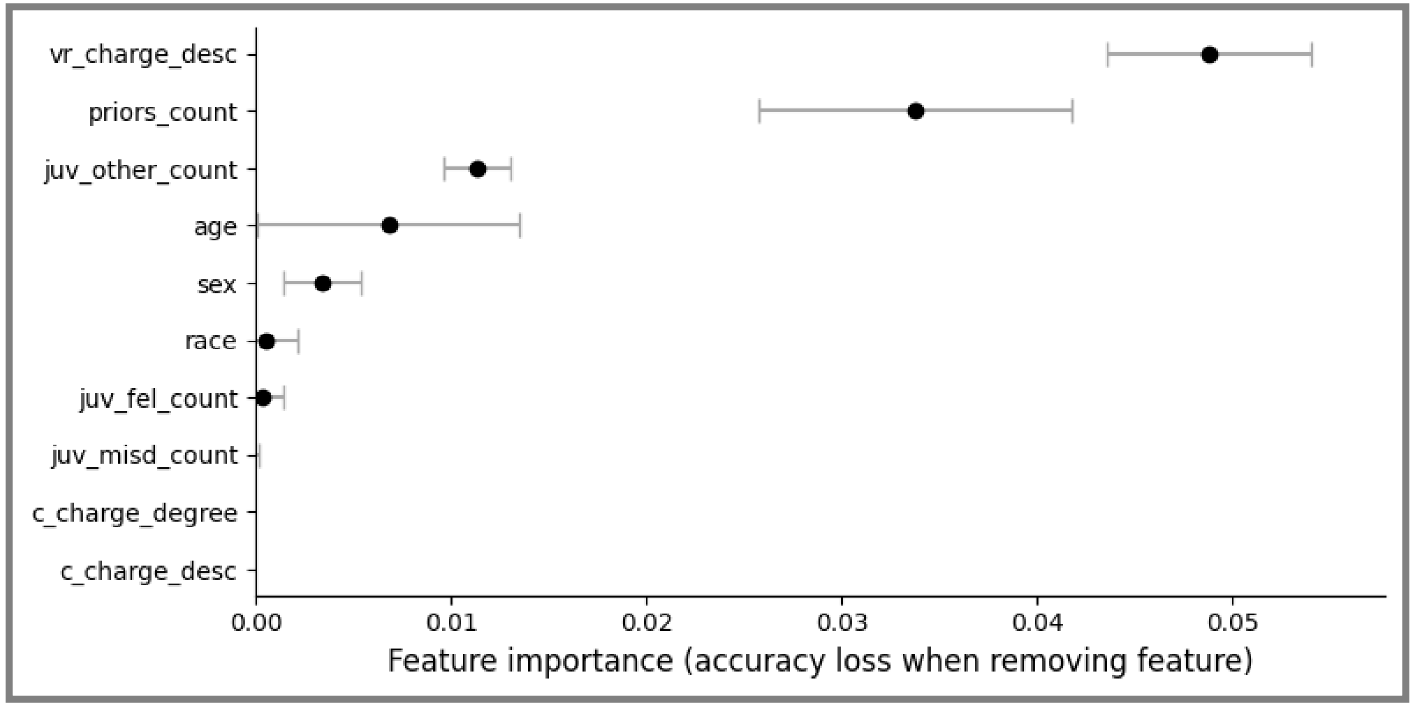
Advantages and Disadvantages of Surrogates?

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



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

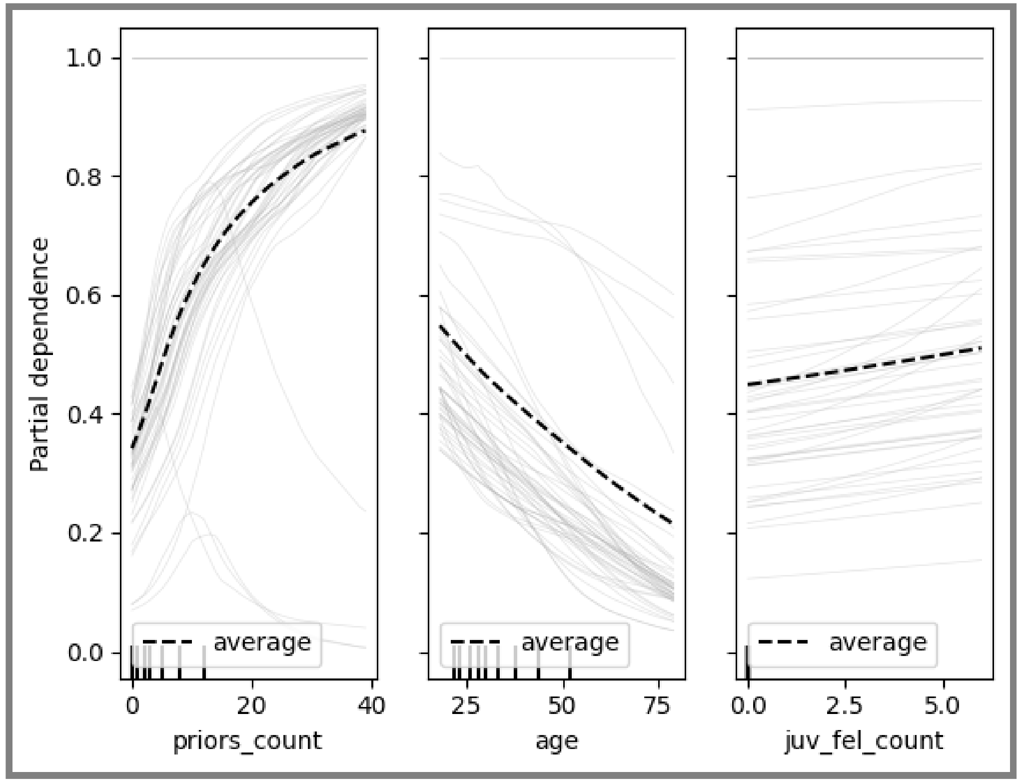
- Highly compressed, *global* insights
- Effect for feature + interactions
- Can only be computed on labeled data, depends on model accuracy, randomness from permutation
- May produce unrealistic inputs when correlations exist

(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: Partial Dependence Plot (PDP)

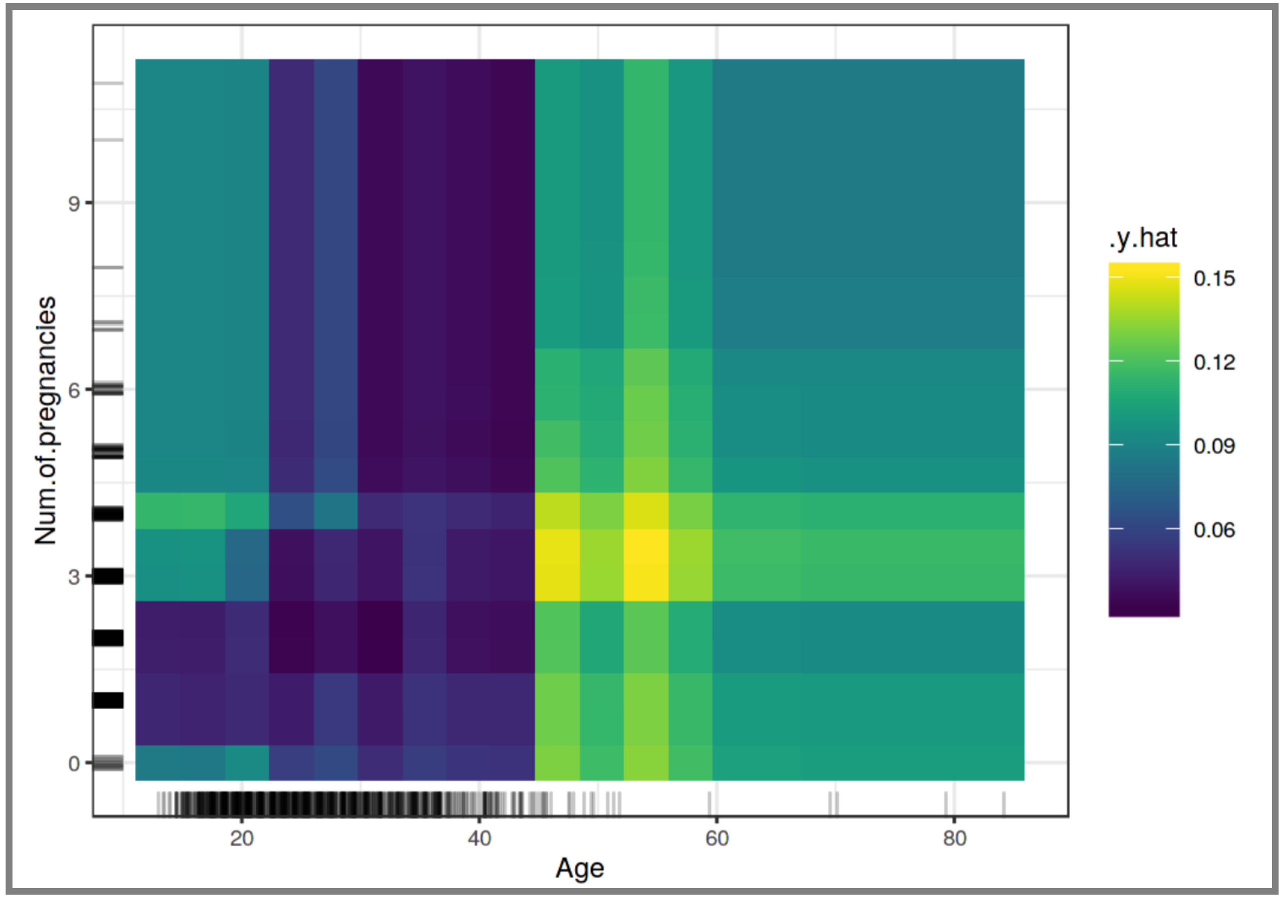


Partial Dependence Plot

- 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



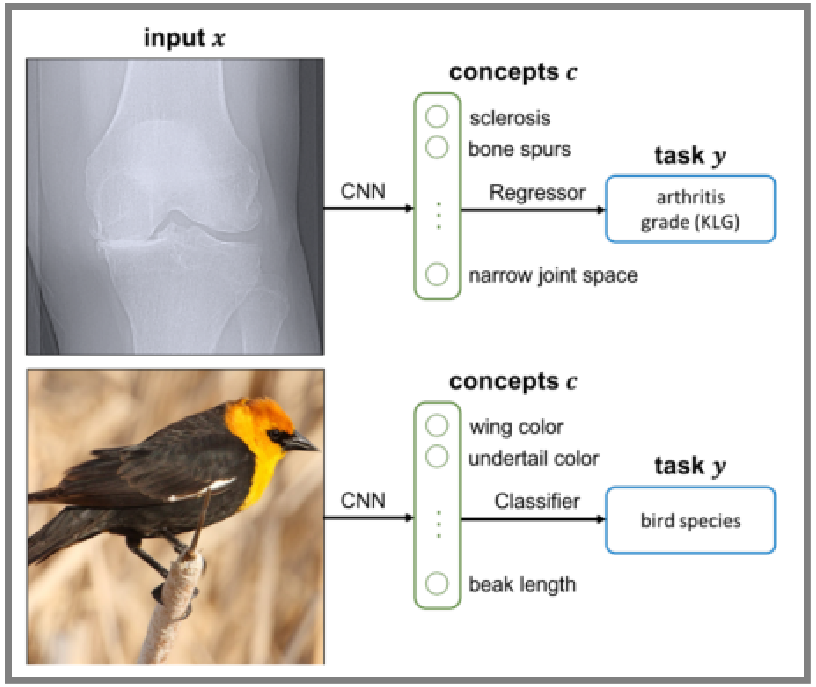
Probability of cancer; source: Christoph Molnar. "Interpretable Machine Learning." 2019

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



Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. "Concept bottleneck models." In Proc. ICML, 2020

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

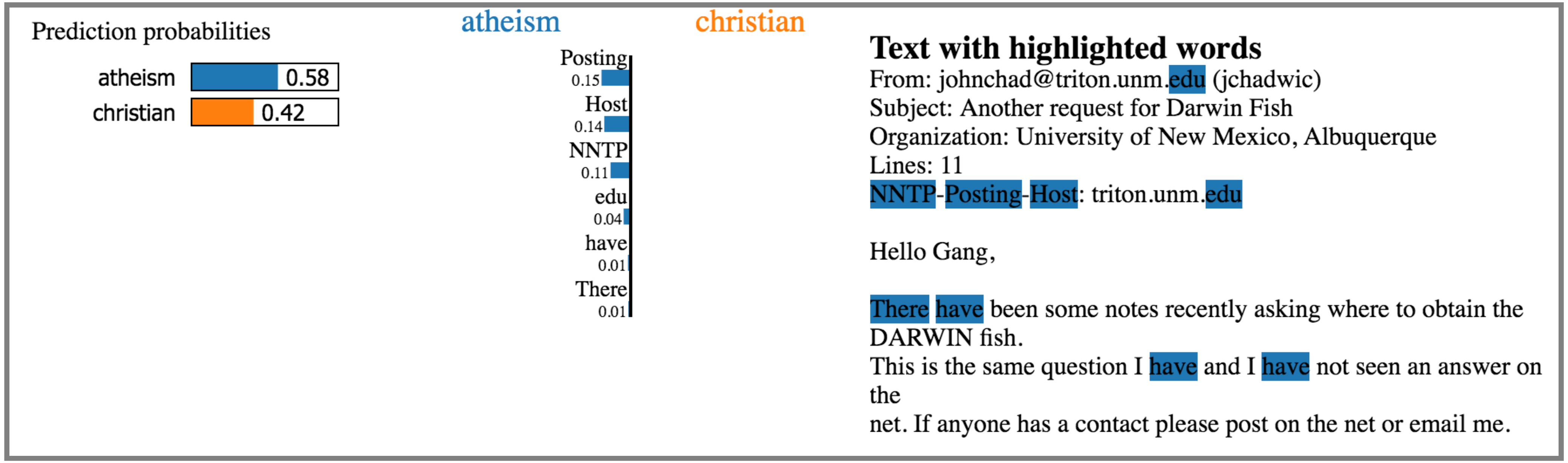
Derive contrastive counterfacturals from models

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: Feature Influences

Which features were most influential for a specific prediction?



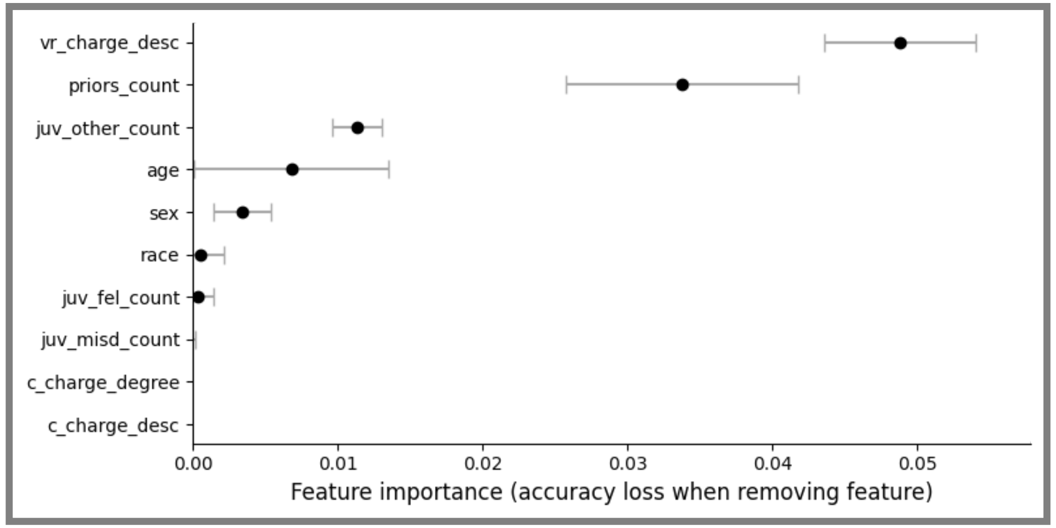
Source: https://github.com/marcotcr/lime

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

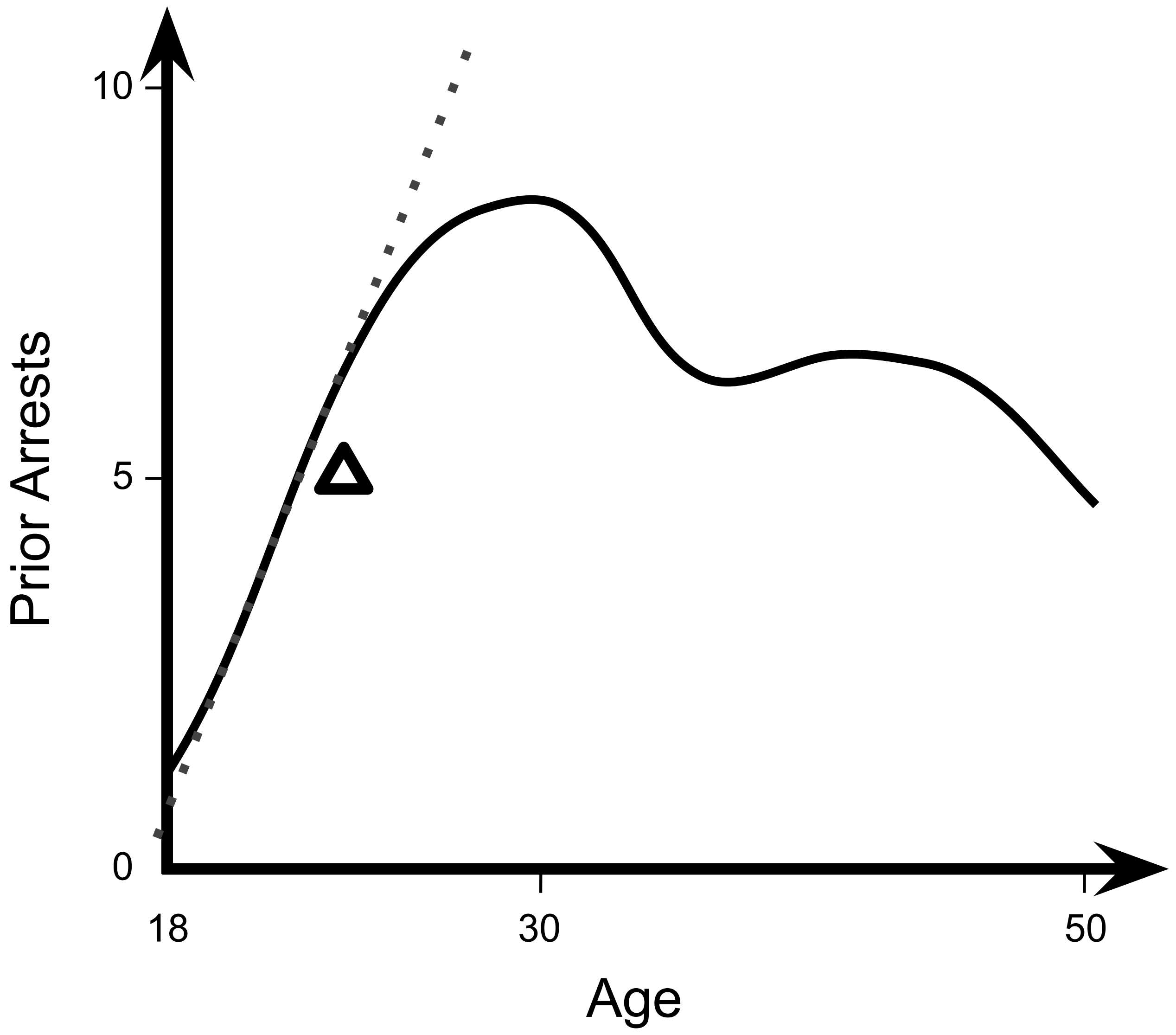
Feature Infl. with Local Surrogates (LIME)

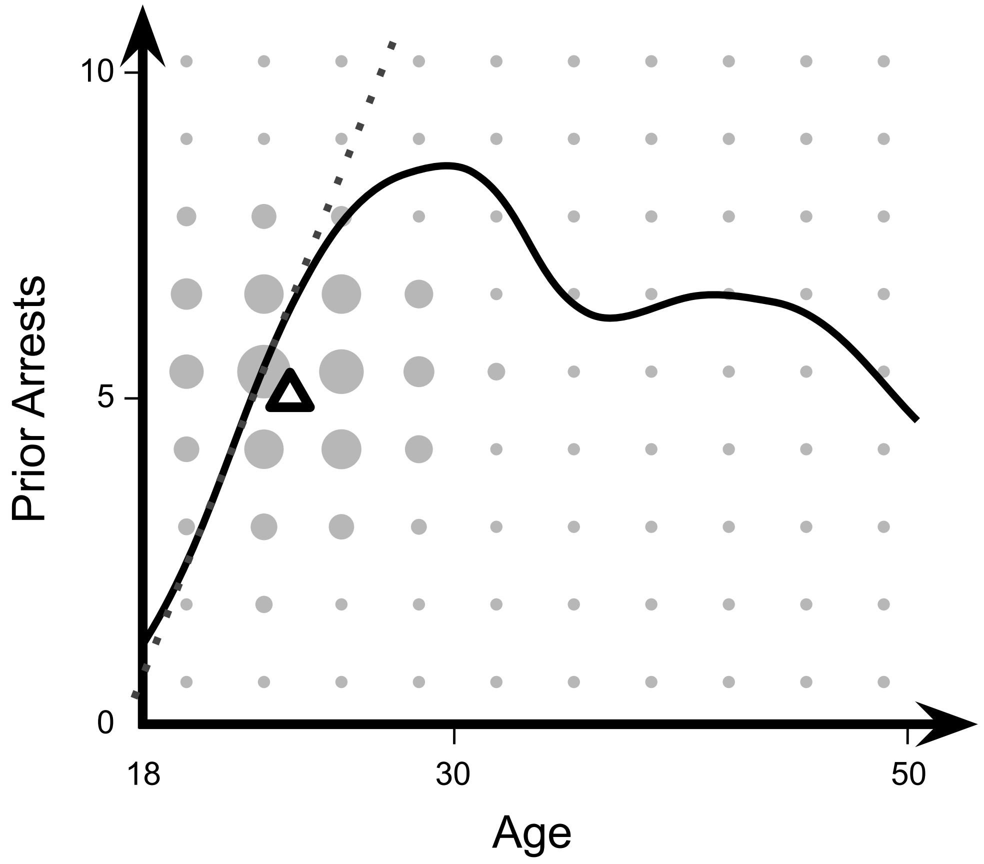
Create an inherently interpretable model (e.g. sparse linear model) for the area around a prediction

Lime approach:

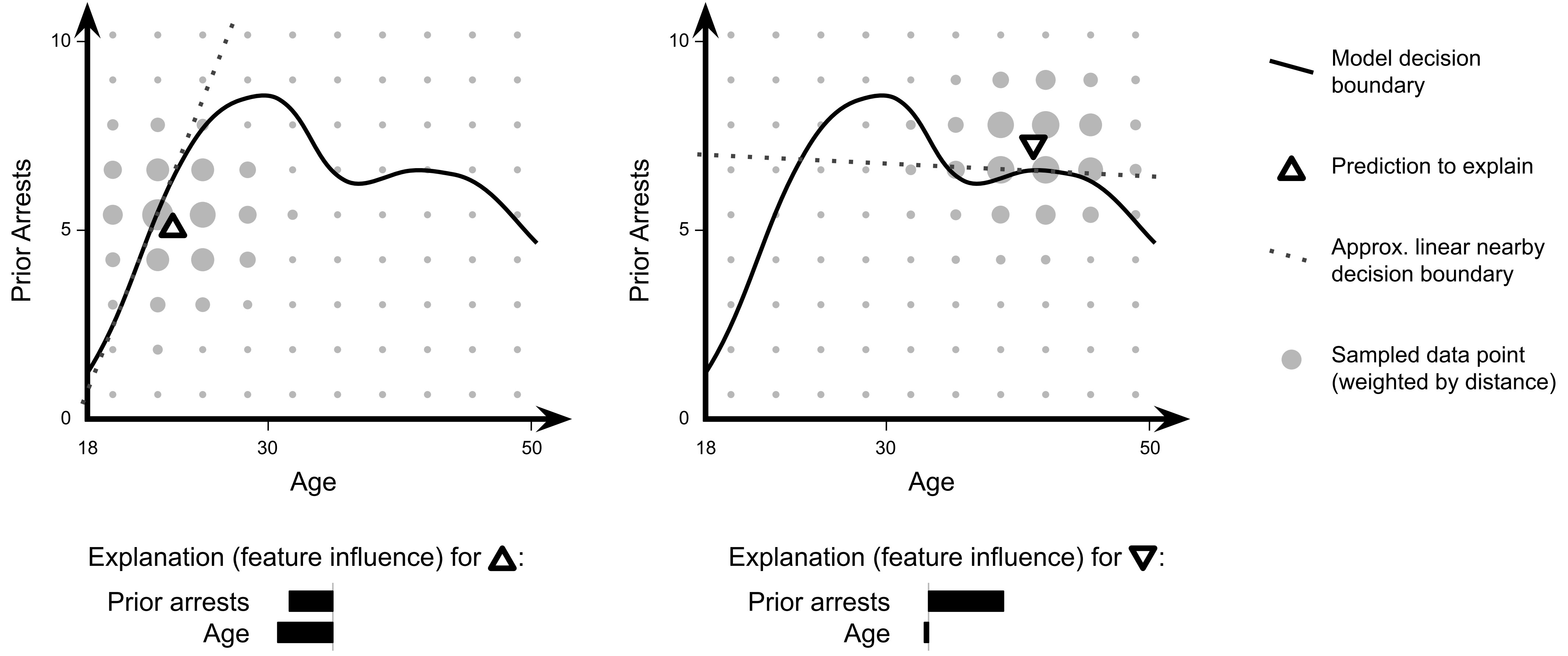
- Create random samples in the area around the data point of interest
- Collect model predictions with f for each sample
- Learn surrogate model g, weighing samples by distance
- Interpret surrogate model g

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.

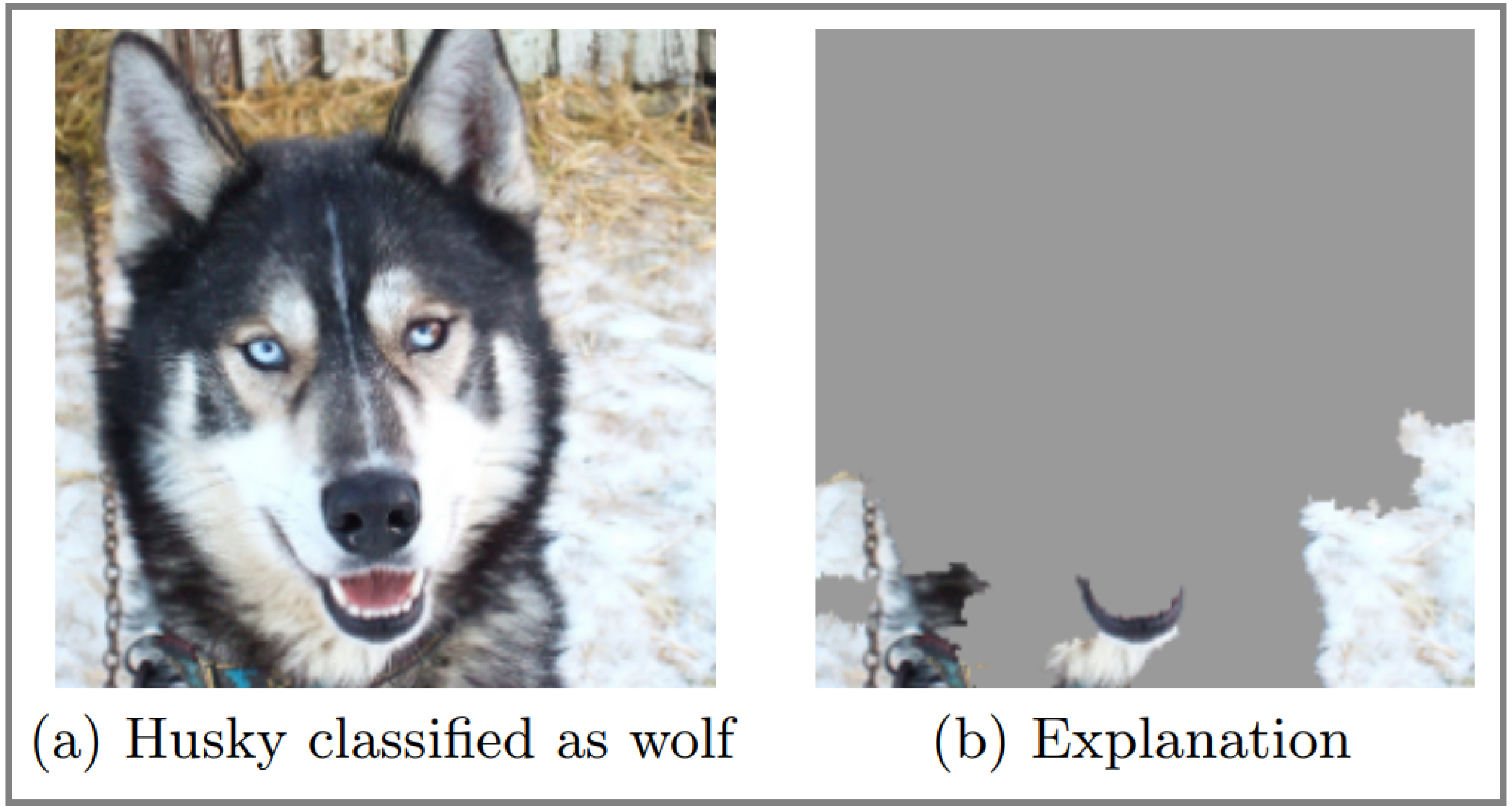

10
5
0
Prior Arrests
18
30
50
Age



Explanation (feature influence) for △:


10
5
0
Prior Arrests
18
30
50
Age
10
5
0
Prior Arrests
18
30
50
Age
Model decision boundary
Prediction to explain
Approx. linear nearby decision boundary
Sampled data point (weighted by distance)
Explanation (feature influence) for △:
Prior arrests
Age
Explanation (feature influence) for ▽:
Prior arrests
Age

LIME Example



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?

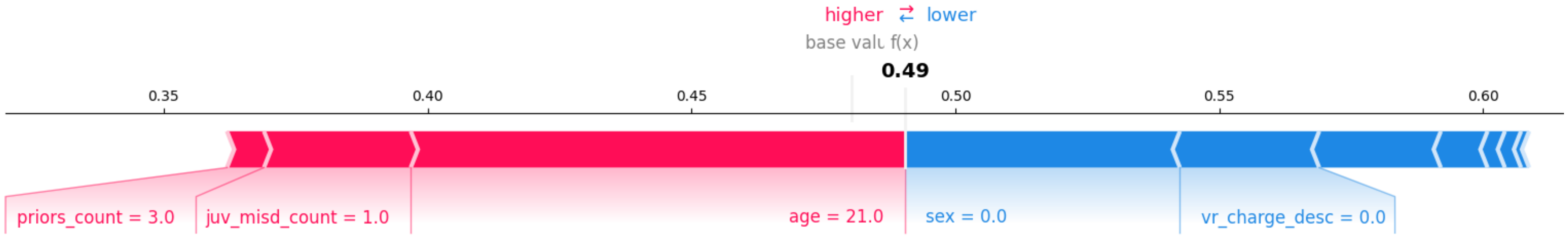
Posthoc Prediction Explanation: Shapley Values / SHAP

- 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)

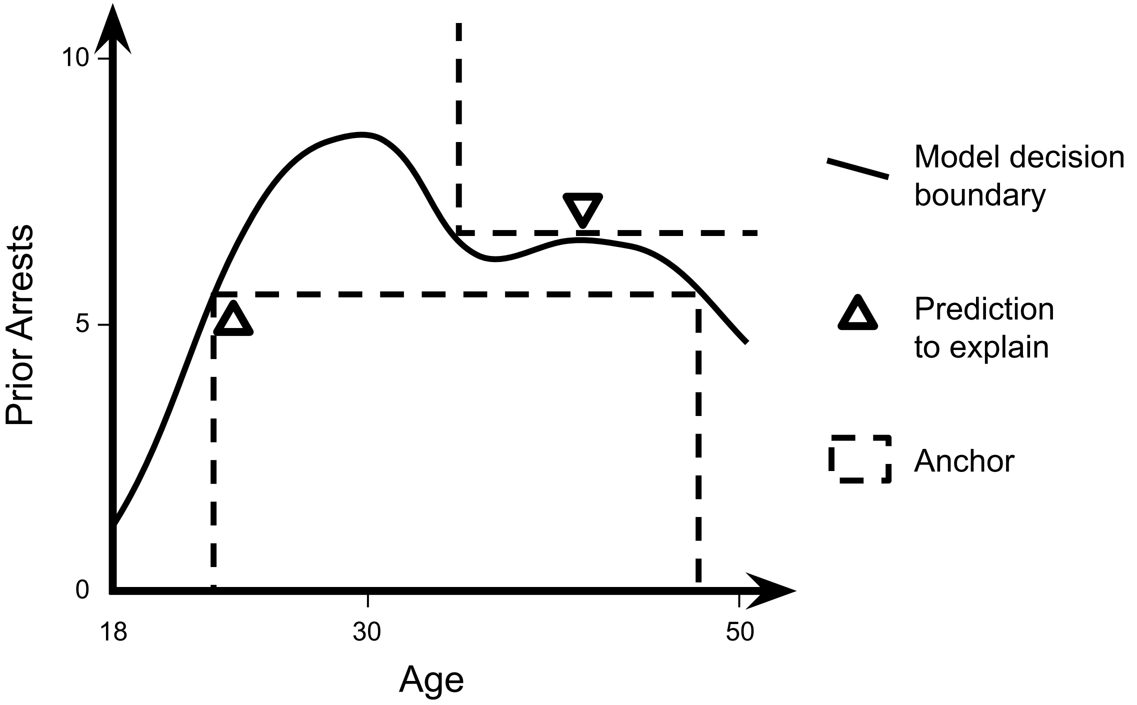
Currently, most common local method used in practice

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.

SHAP Force Plot



Anchors



Object detected: Steam Locomotive

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

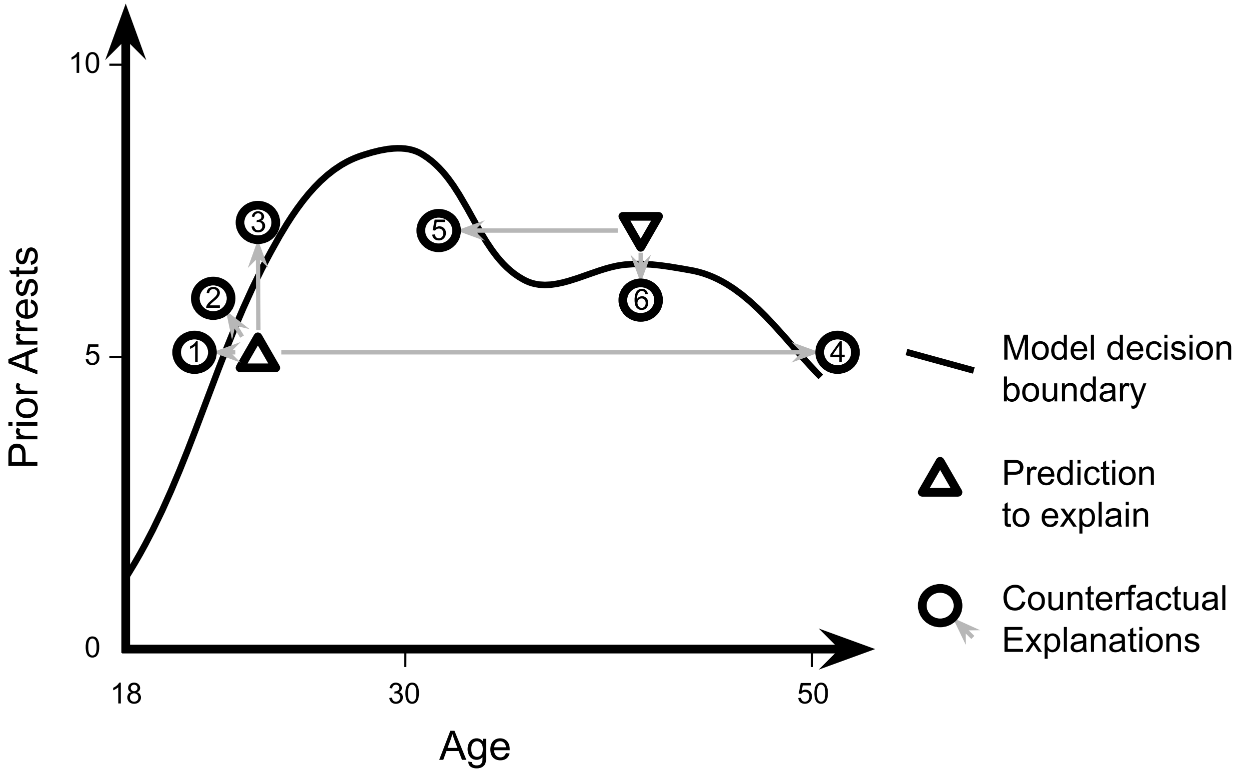
Often long or multiple explanations

(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

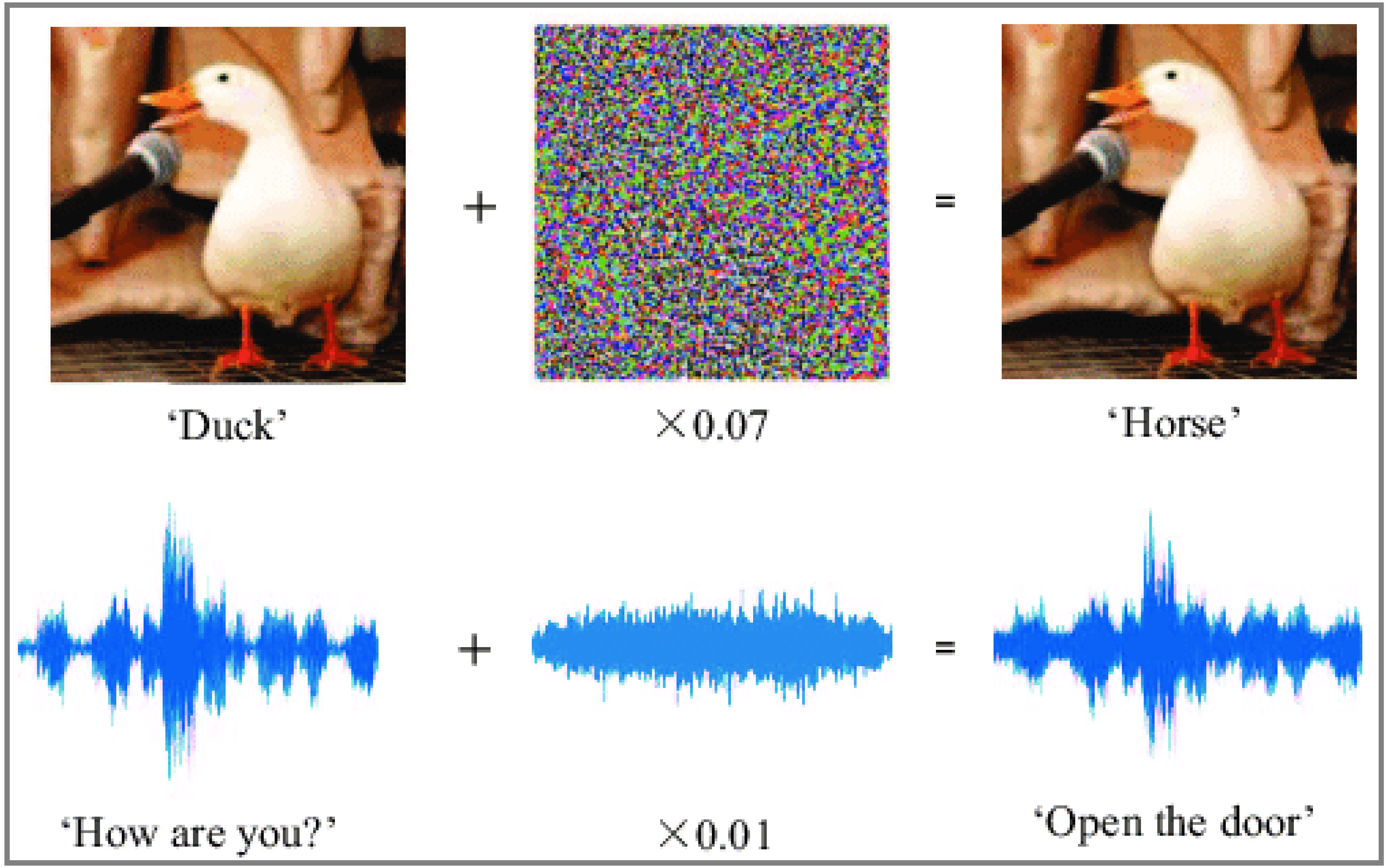
Random search (with growing distance) possible, but inefficient

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

Can incorporate distance in loss function

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

(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

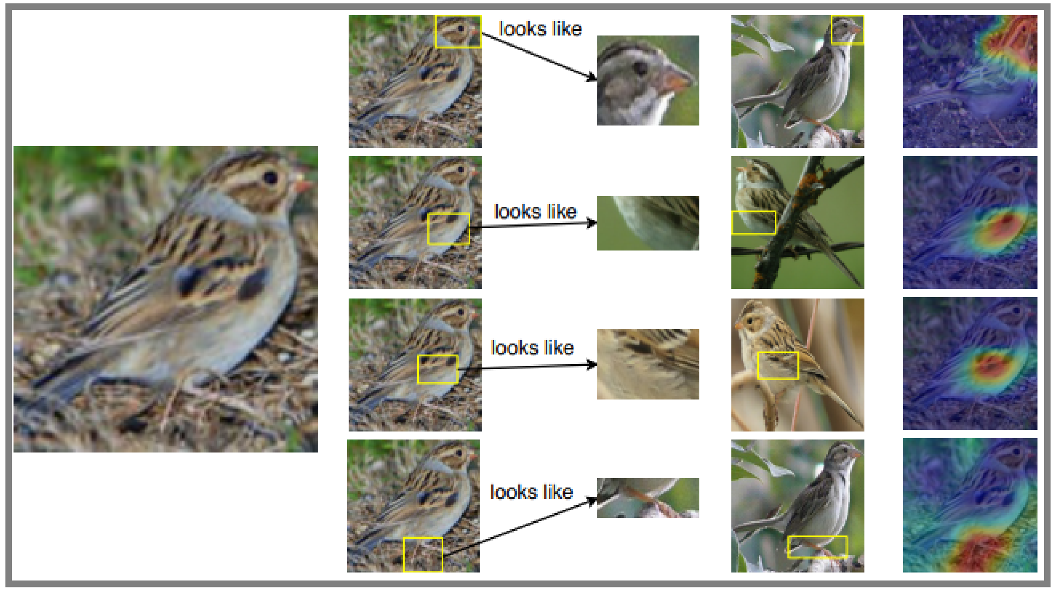
Actionable Counterfactuals

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

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:

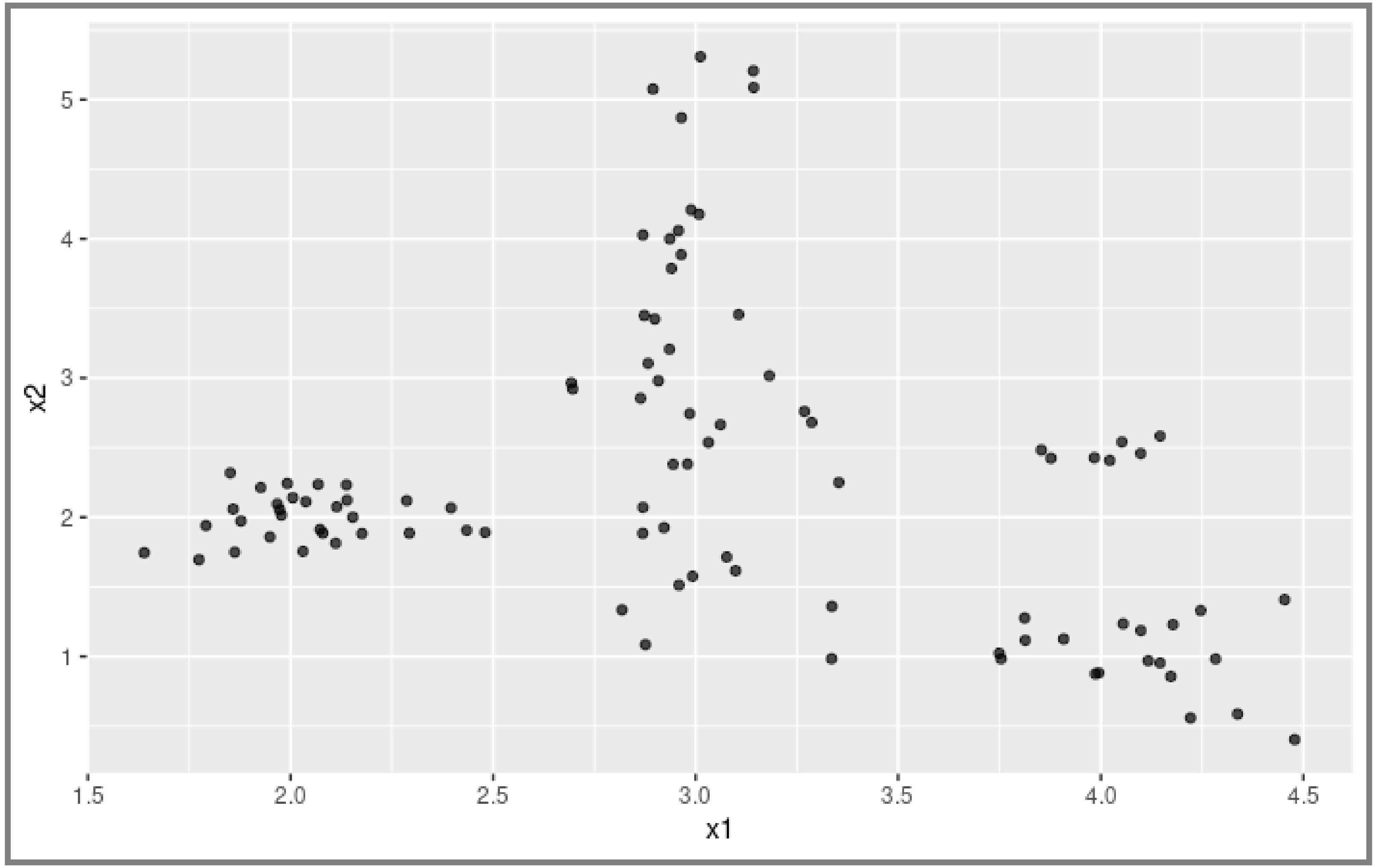
- Understanding a model
- Explaining a prediction
- **Understanding the data**

Prototypes and Criticisms

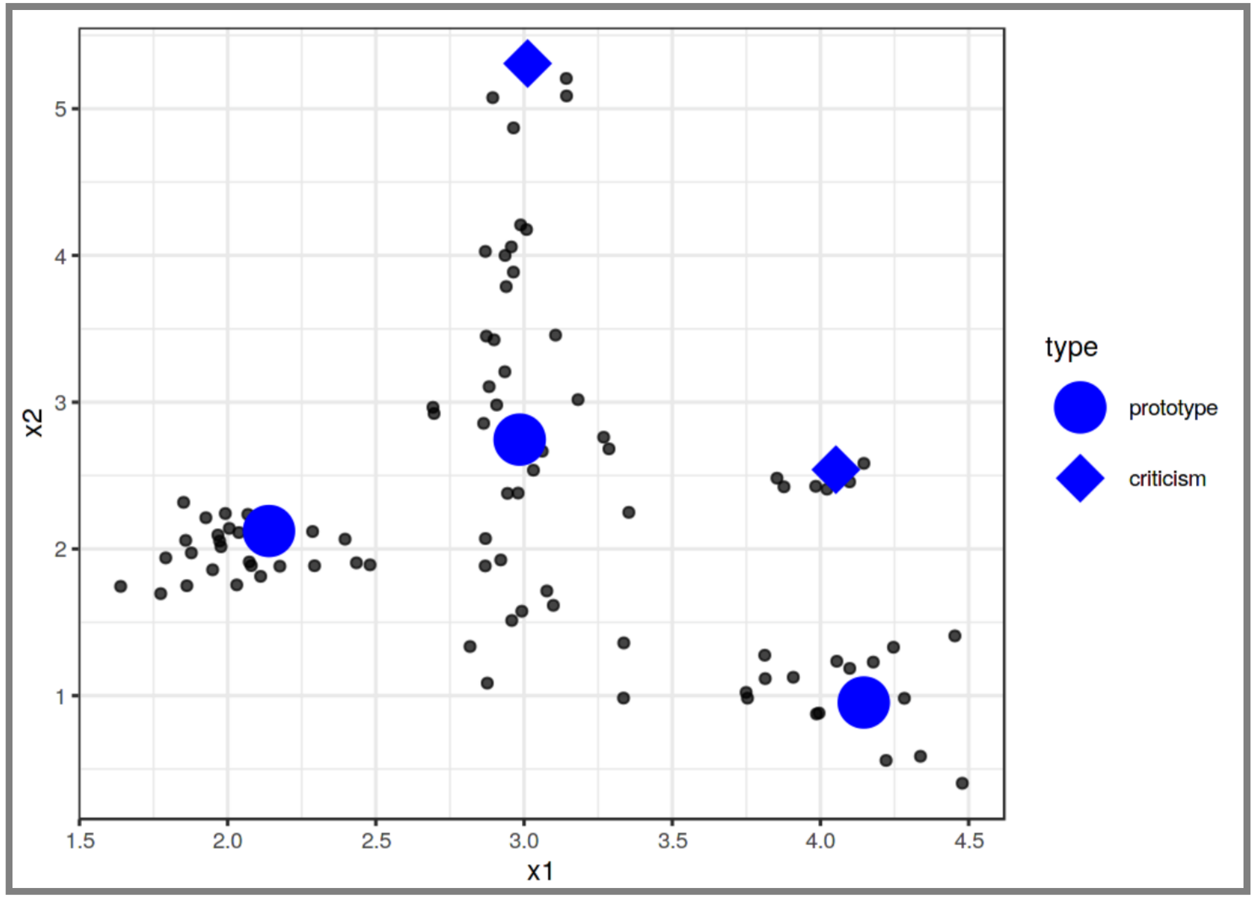
- *Prototype* is a data instance that is representative of all the data
- *Criticism* is a data instance not well represented by the prototypes

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

Example: Prototypes and Criticisms?

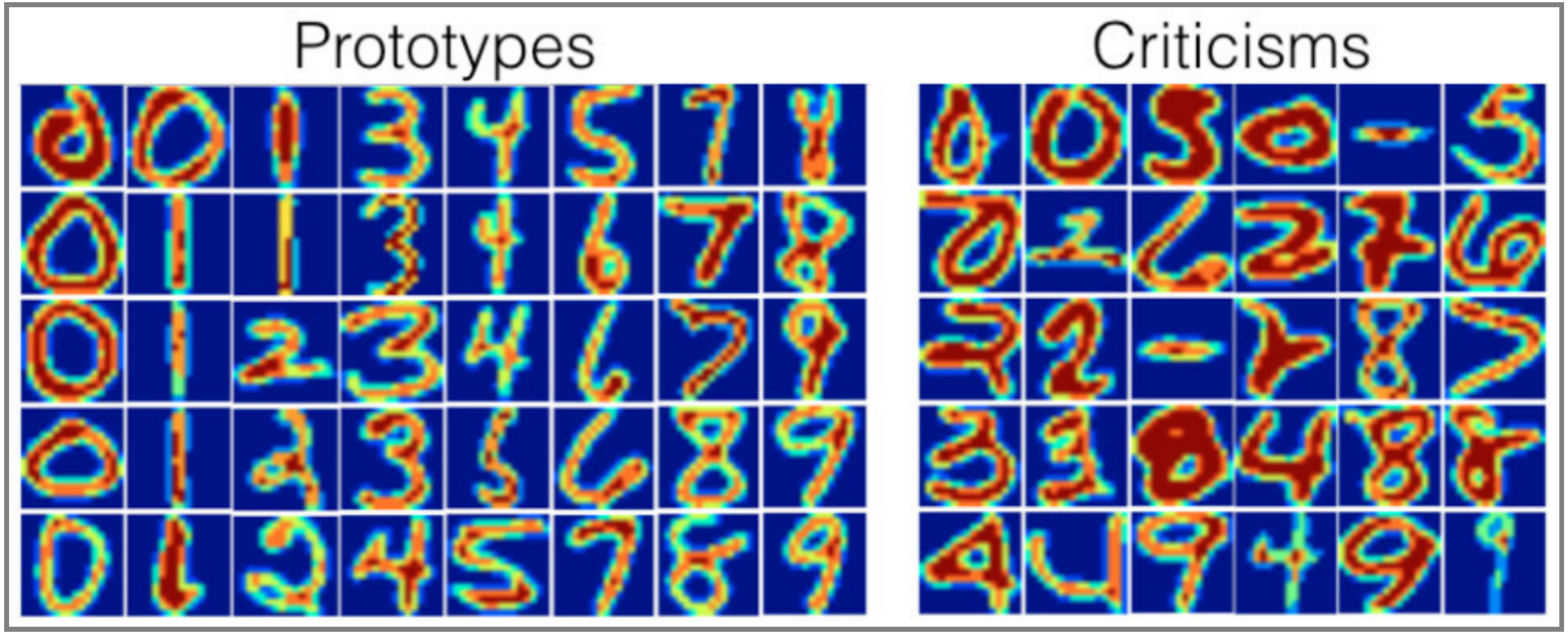


Example: Prototypes and Criticisms



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

Example: Prototypes and Criticisms



Source: Christoph Molnar. "Interpretable Machine Learning." 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

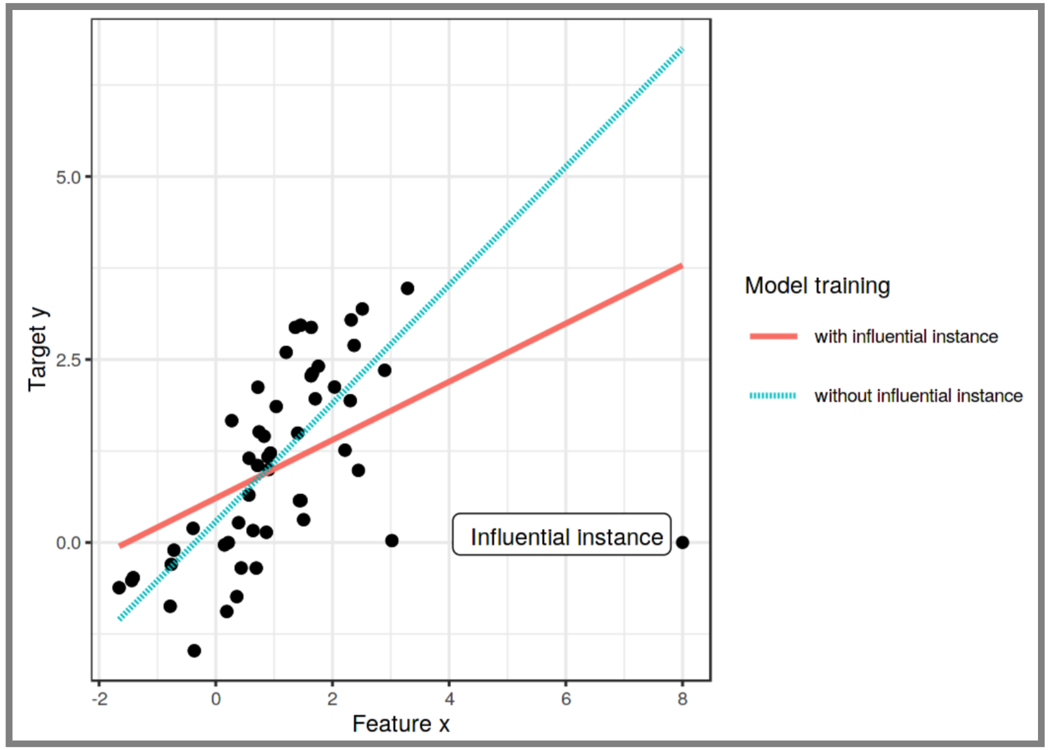
Discussion: Prototypes and Criticisms

- 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

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 n instances...
- ... train model 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

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

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

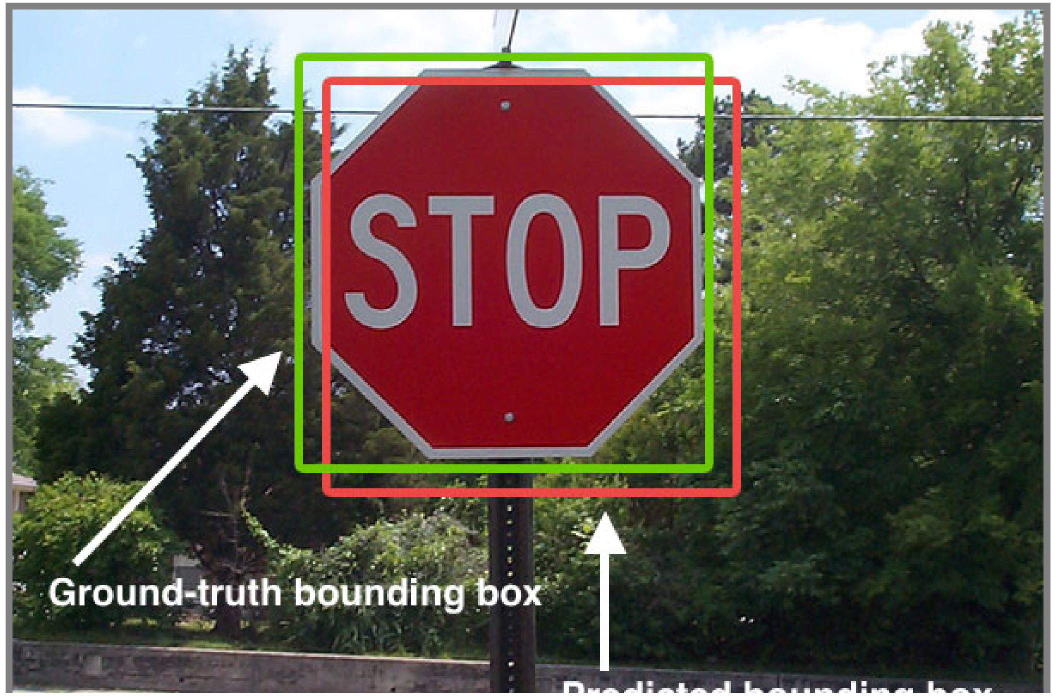
Many data exploration and data debugging techniques:

- 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`.

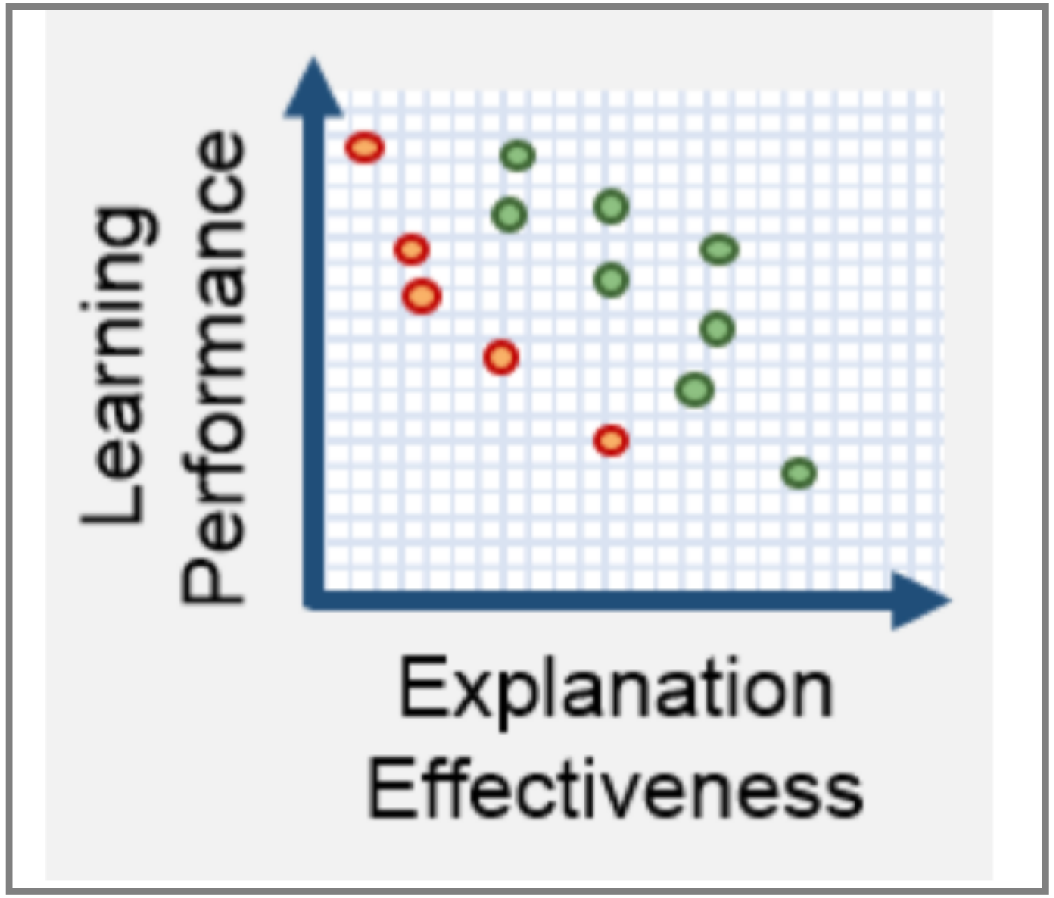
Algorithm bad at recognizing some signs in some conditions:



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

"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

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and

"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

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.