# 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.

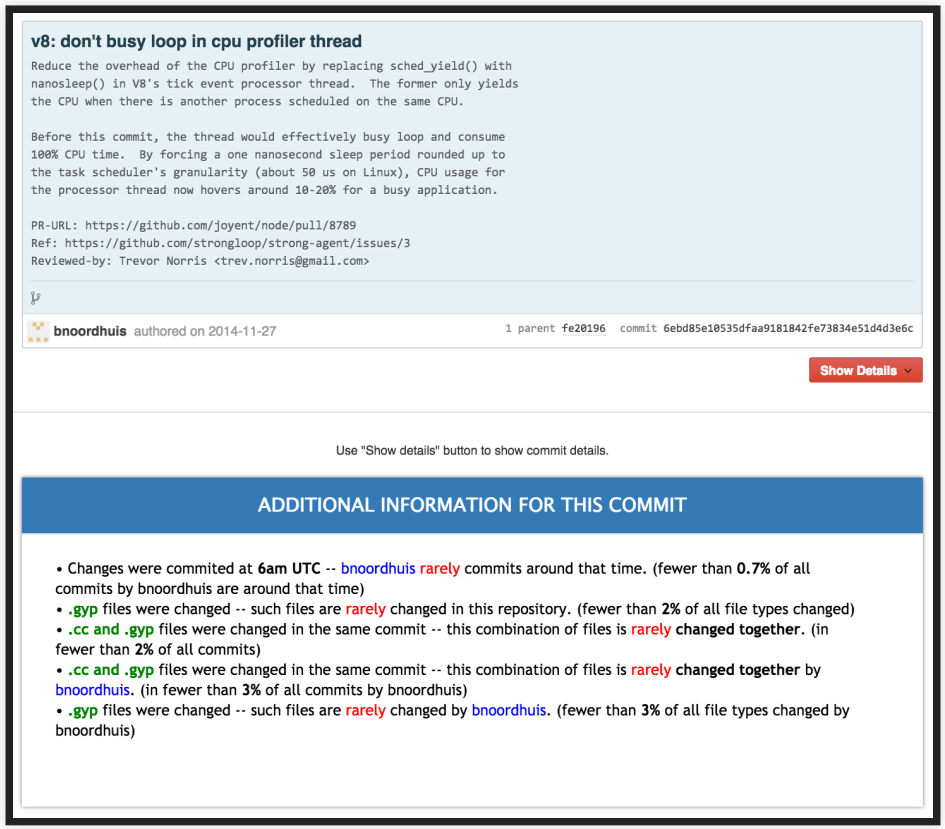
Recommended supplementary reading:  Christoph Molnar. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2019

# 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

# 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?

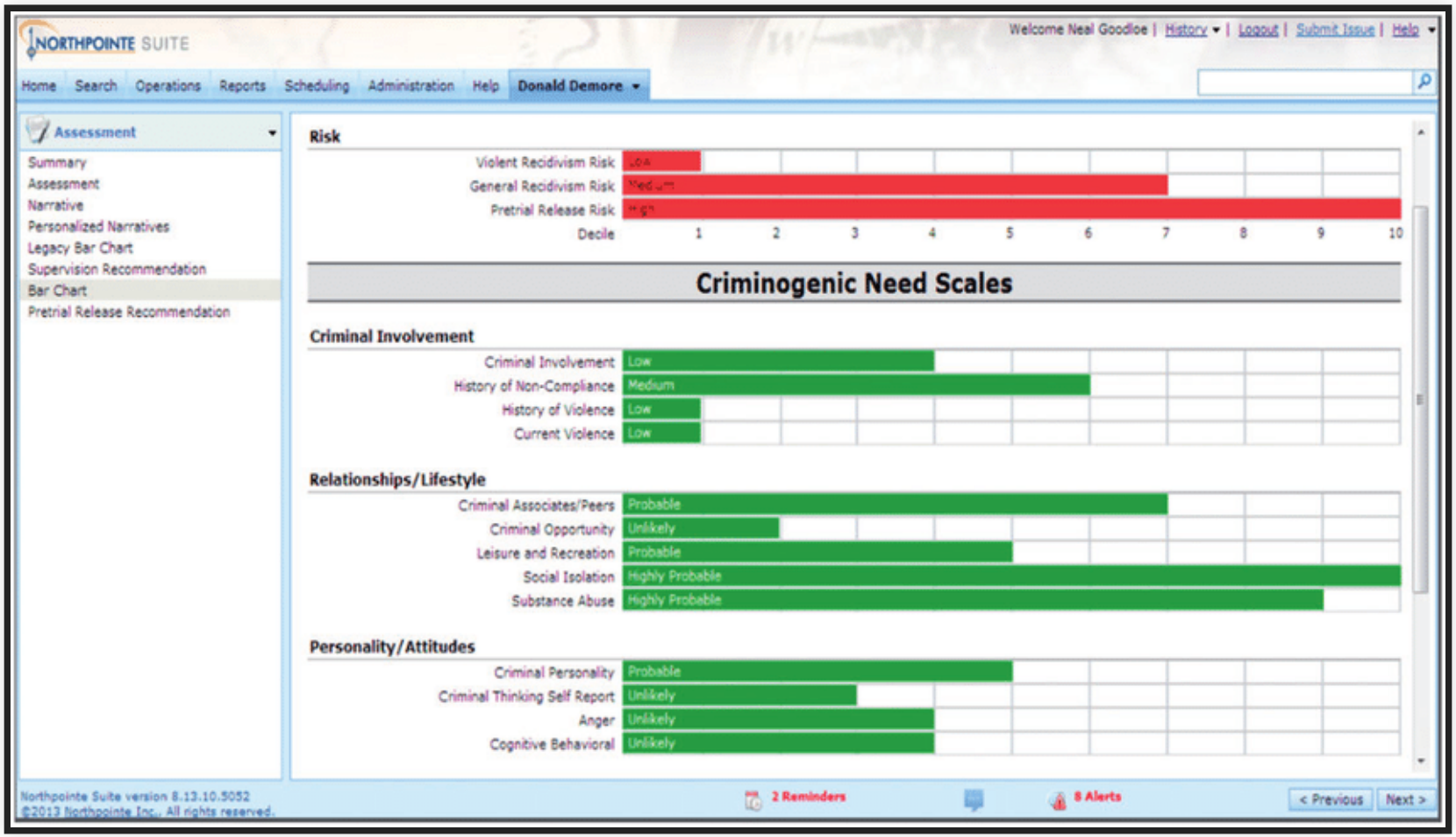



Image source (CC BY-NC-ND 4.0): Christin, Angèle. (2017). Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society. 4.

# HOW TO JUDGE RELATIVE TO SERIOUSNESS OF THE CRIME?

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

**PREDICT ARREST FOR ANY OFFENSE IF SCORE $> 1$**

| | | | |
|---|---|---|---|
| 1. | Prior Arrests $\geq 2$ | 1 point | ··· |
| 2. | Prior Arrests $\geq 5$ | 1 point | + ··· |
| 3. | Prior Arrests for Local Ordinance | 1 point | + ··· |
| 4. | Age at Release between 18 to 24 | 1 point | + ··· |
| 5. | Age at Release $\geq 40$ | -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 healthcare and criminal justice." Interfaces 48, no. 5 (2018): 449-466.

# IS THERE AN ACTUAL PROBLEM? HOW TO FIND OUT?

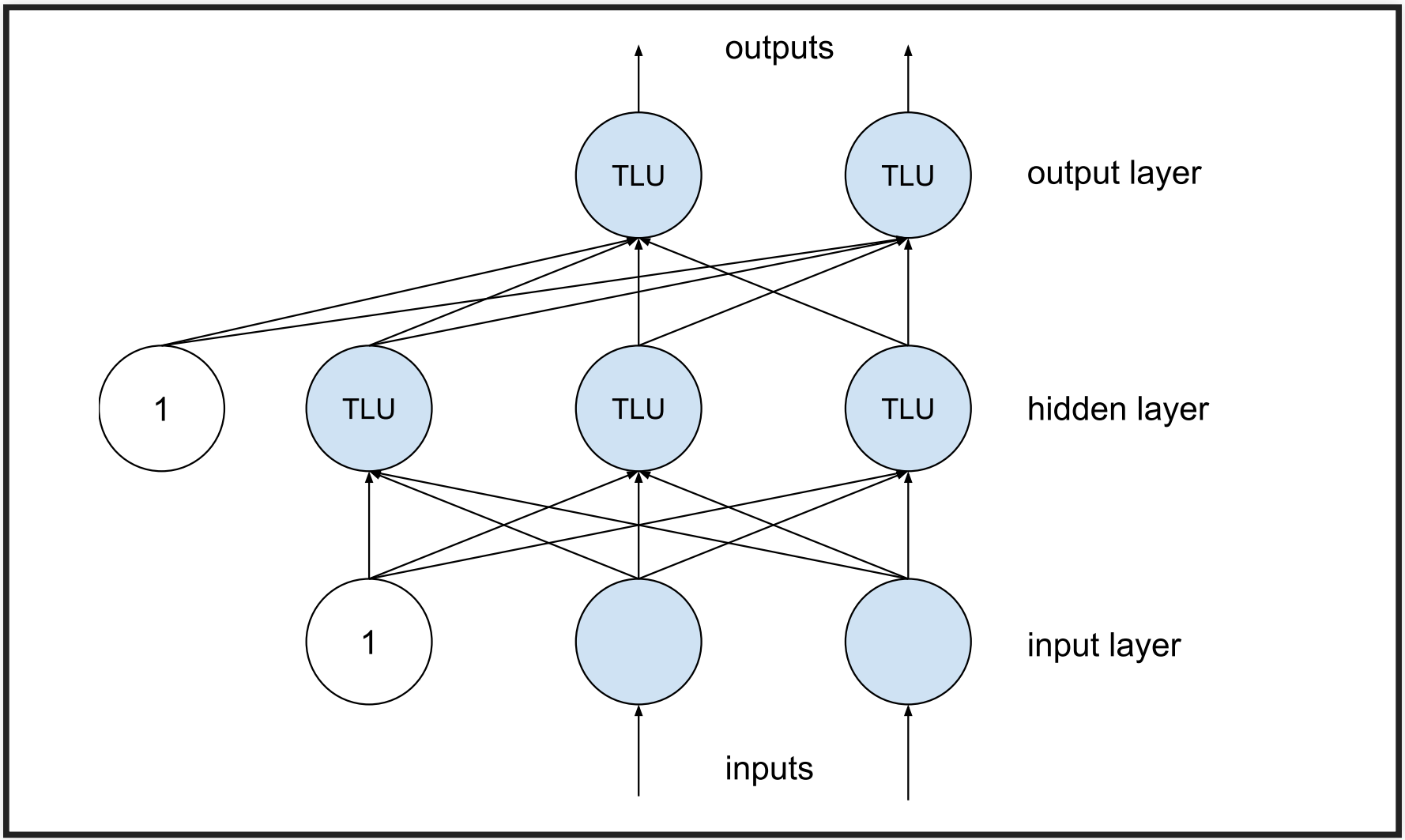
*Nothing t[illegible]lgorithm...*

# EXPLAINING DECISIONS

Cat? Dog? Lion?

Confidence? Why?

# WHAT'S HAPPENING HERE?



# EXTRACTING KNOWLEDGE FROM DATA

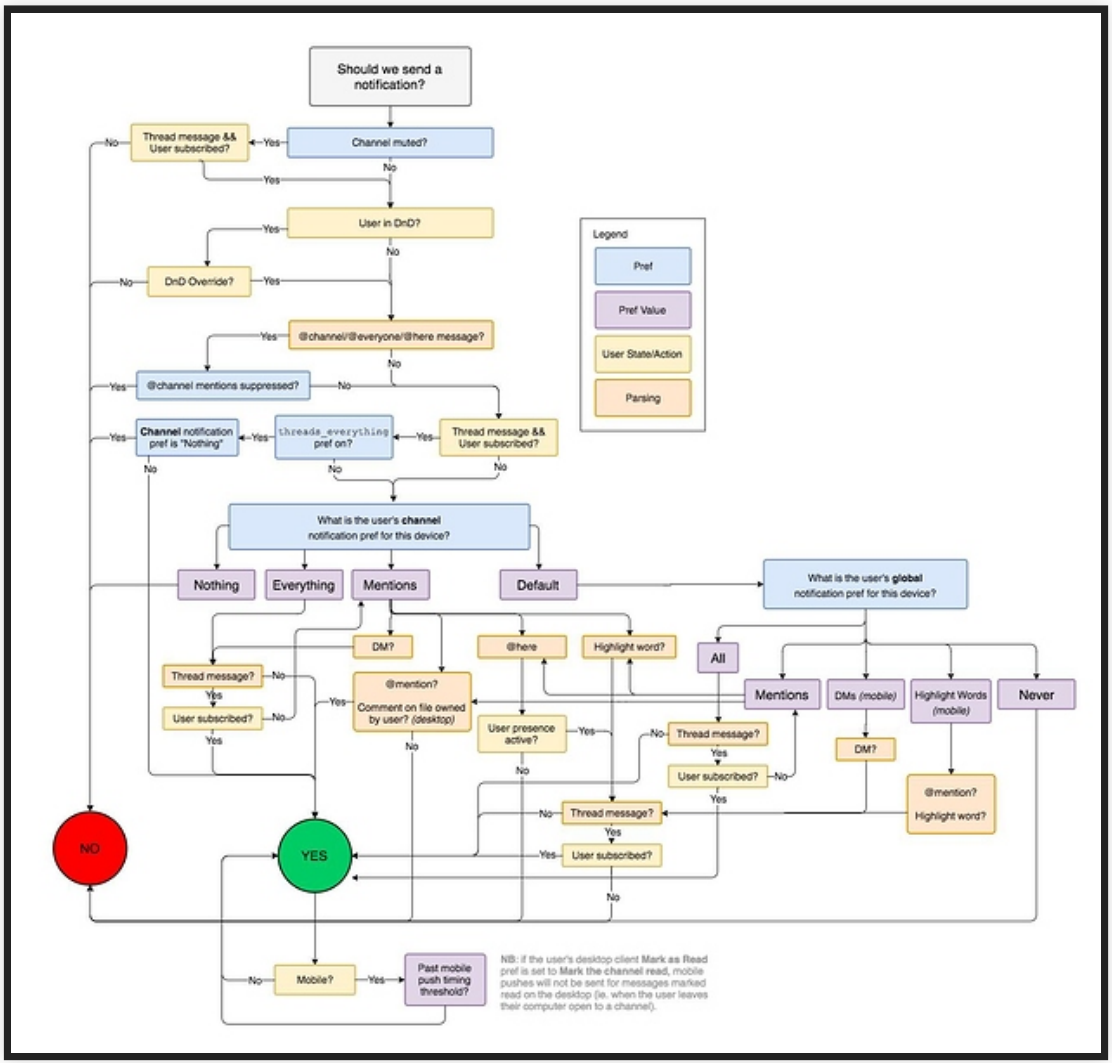
- Sale 1: Bread, Milk
- Sale 2: Bread, Diaper, Beer, Eggs
- Sale 3: Milk, Diaper, Beer, Coke
- Sale 4: Bread, Milk, Diaper, Beer
- Sale 5: Bread, Milk, Diaper, Coke

Rules

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

(see association rule mining in earlier lecture)

# EXPLAINING DECISIONS



# EXPLAINING DECISIONS

```
> parent(john, douglas).
> parent(bob, john).
> parent(ebbon, bob).

> parent(john, B)?
parent(john, douglas).

> parent(A, A)?

> ancestor(A, B) :- parent(A, B).
> ancestor(A, B) :- parent(A, C), ancestor(C, B).

> ancestor(A,B)?
ancestor(john, douglas).
ancestor(ebbon, bob).
```

# EXPLAINABILITY IN AI

- Explain how the model made a decision
    - Rules, cutoffs, reasoning?
    - What are the relevant factors?
    - Why those rules/cutoffs?
- Challenging in symbolic AI with complicated rules
- Challenging with ML because models too complex and based on data
    - Can we understand the rules?
    - Can we understand why these rules?

# WHY EXPLAINABILITY?

# 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

# HELP CUSTOMERS ACHIEVE BETTER 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?*

# 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

# 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

# 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?

# 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

# 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
- Interpretability vs explainability often used inconsistently or interchangeble

# 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

# INHERENTLY INTERPRETABLE MODELS

# 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% |

# DECISION TREES

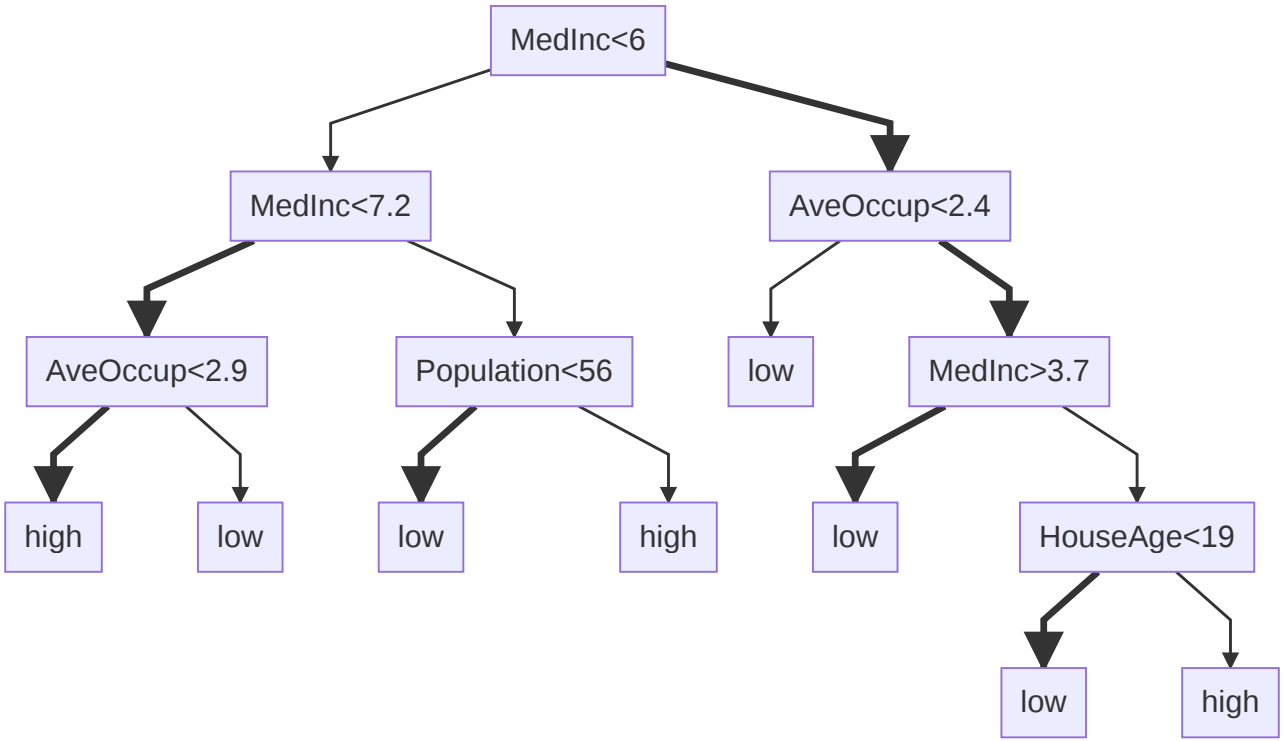
Easy to interpret up to a size

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
```

# EXAMPLE: CALIFORNIA HOUSING DATA



Speaker notes

Ask questions about specific outcomes, about common patterns, about counterfactual explanations

# DECISION RULES

*if-then rules mined from data*

easy to interpret if few and simple rules

see association rule mining, recall:

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

# K-NEAREST NEIGHBORS

- Instance-based learning
- Returns most common class among the *k* nearest training data points
- No global interpretability, because no global rules

- Interpret results by showing nearest neighbors
- Interpretation assumes understandable distance function and interpretable reference data points

*example: predict & explain car prices by showing similar sales*

# RESEARCH IN 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 EXPLANATIONS OF BLACK-BOX MODELS

(large research field, many approaches, much recent research)

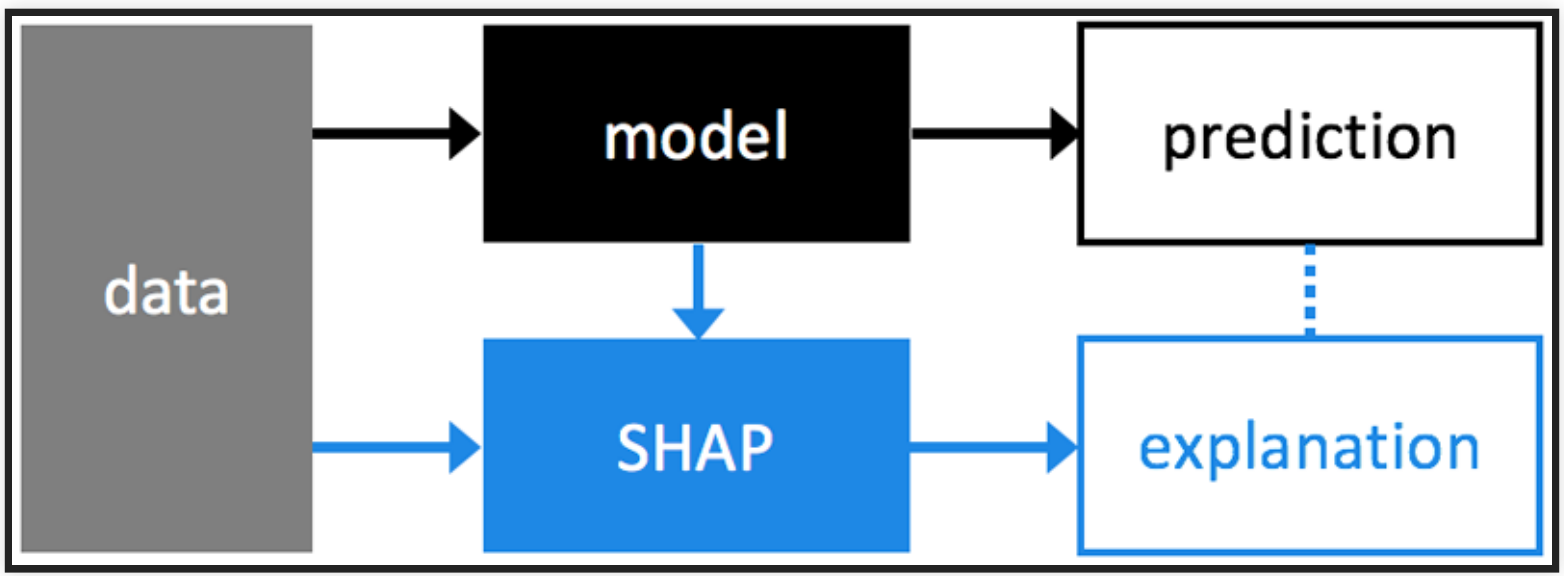



Figure: Lundberg, Scott M., and Su-In Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017.

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

# EXPLAINING BLACK-BOX MODELS

Given model $f$ observable by querying

No access to model internals or training data (e.g., own deep neural network, online prediction service, ...)

Possibly many queries of $f$

# 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?

# LOCAL SURROGATES (LIME)

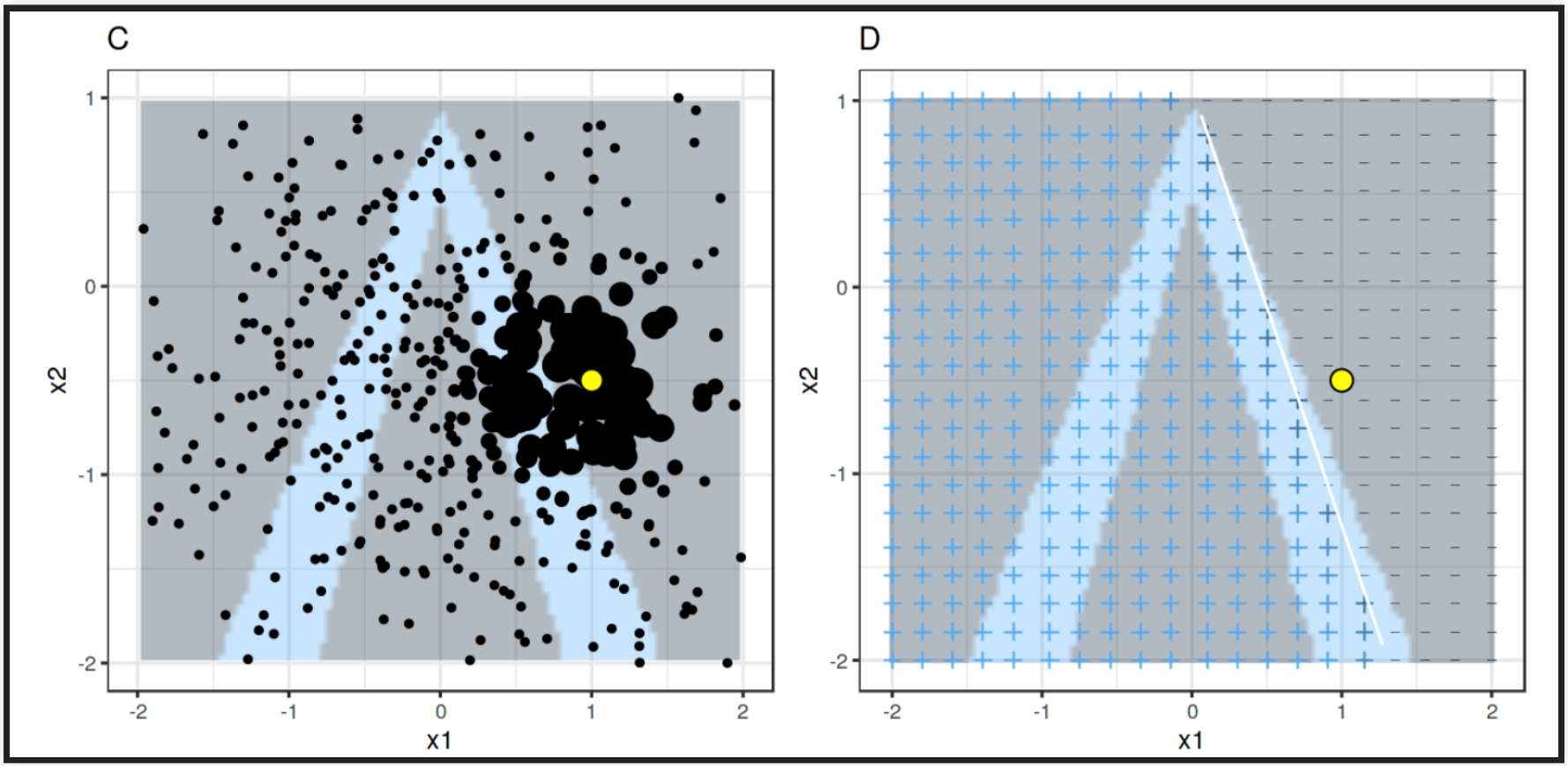
*Create an inherently interpretable model (e.g. sparse linear model) for the area around a prediction*

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

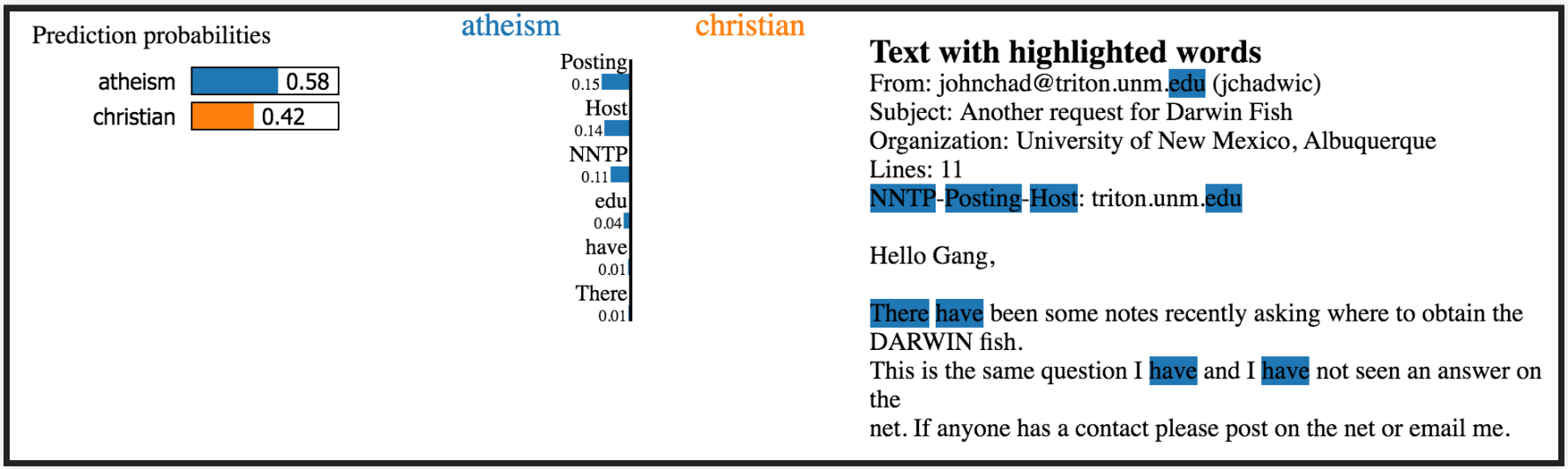



Source: Christoph Molnar. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2019

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




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

# LIME EXAMPLE

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

# 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?

# ADVANTAGES AND DISADVANTAGES OF (LOCAL) 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

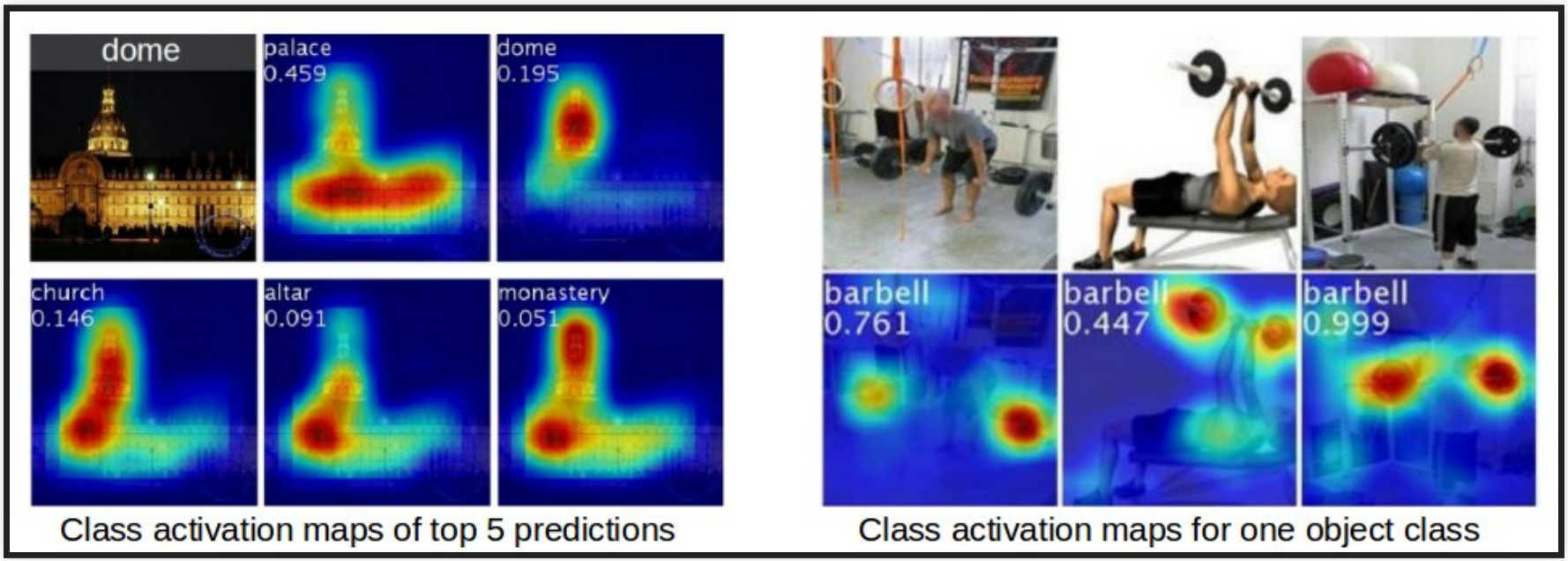
- partial local explanation not sufficient for compliance scenario where full explanation is needed
- explanations may be unstable

# SHARPLEY 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

# 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

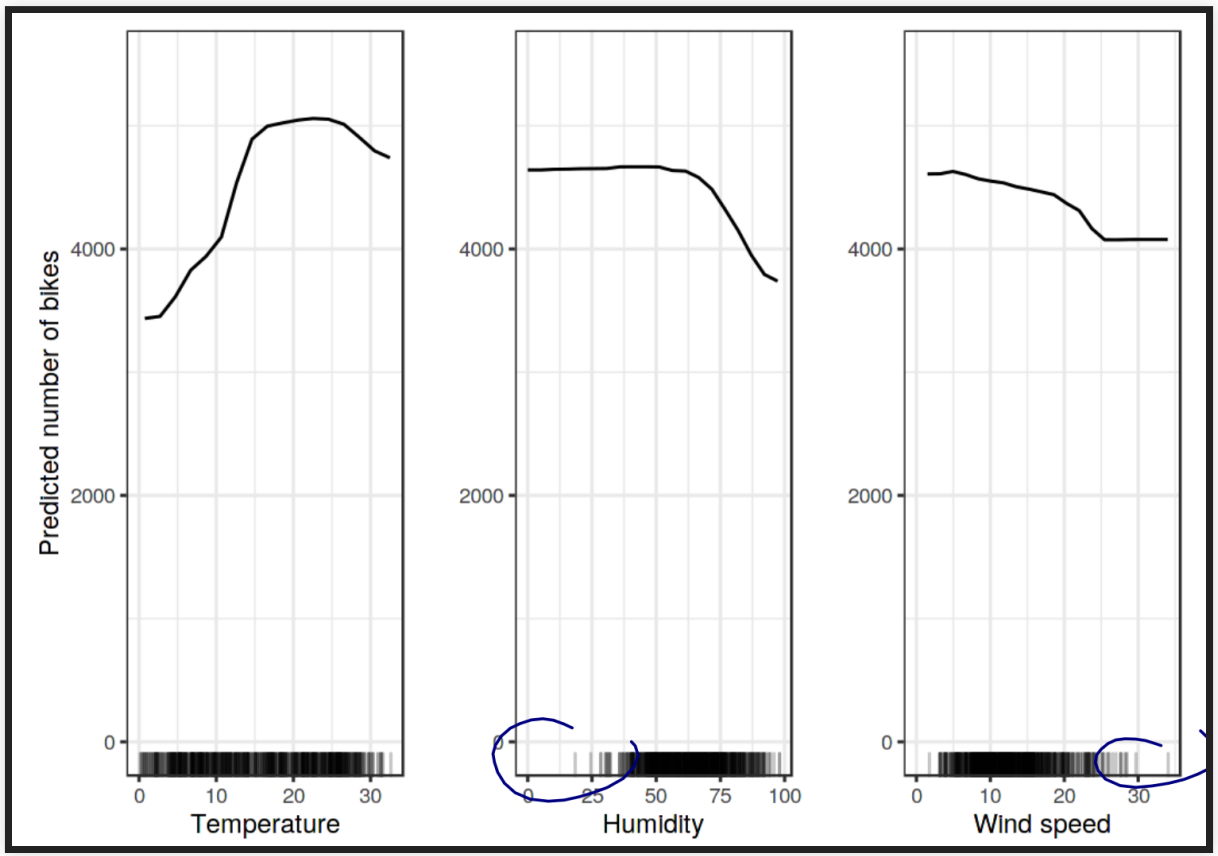
# PARTIAL DEPENDENCE PLOT (PDP)

- 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

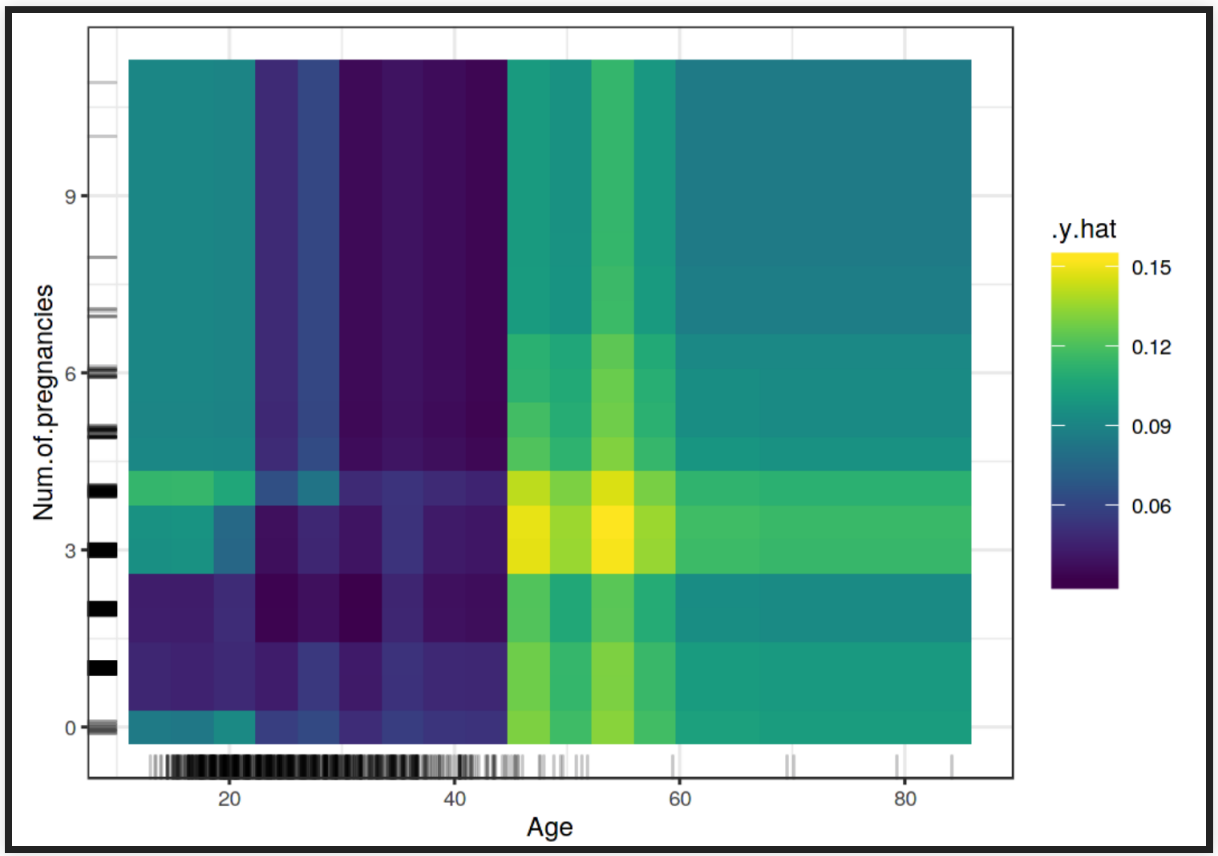
*Bike rental in DC*




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

# PARTIAL DEPENDENCE PLOT EXAMPLE

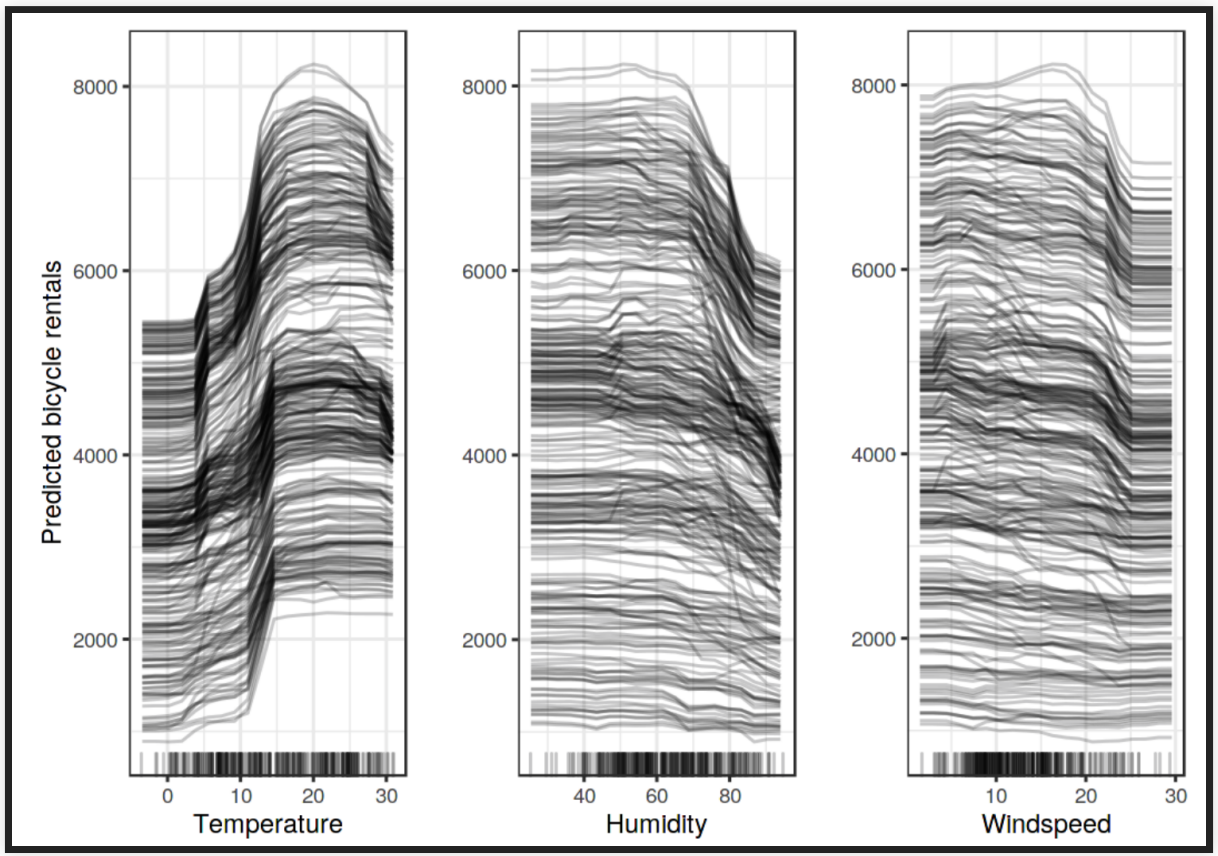
*Probability of cancer*




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

# INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

*Similar to PDP, but not averaged; may provide insights into interactions*




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

# 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

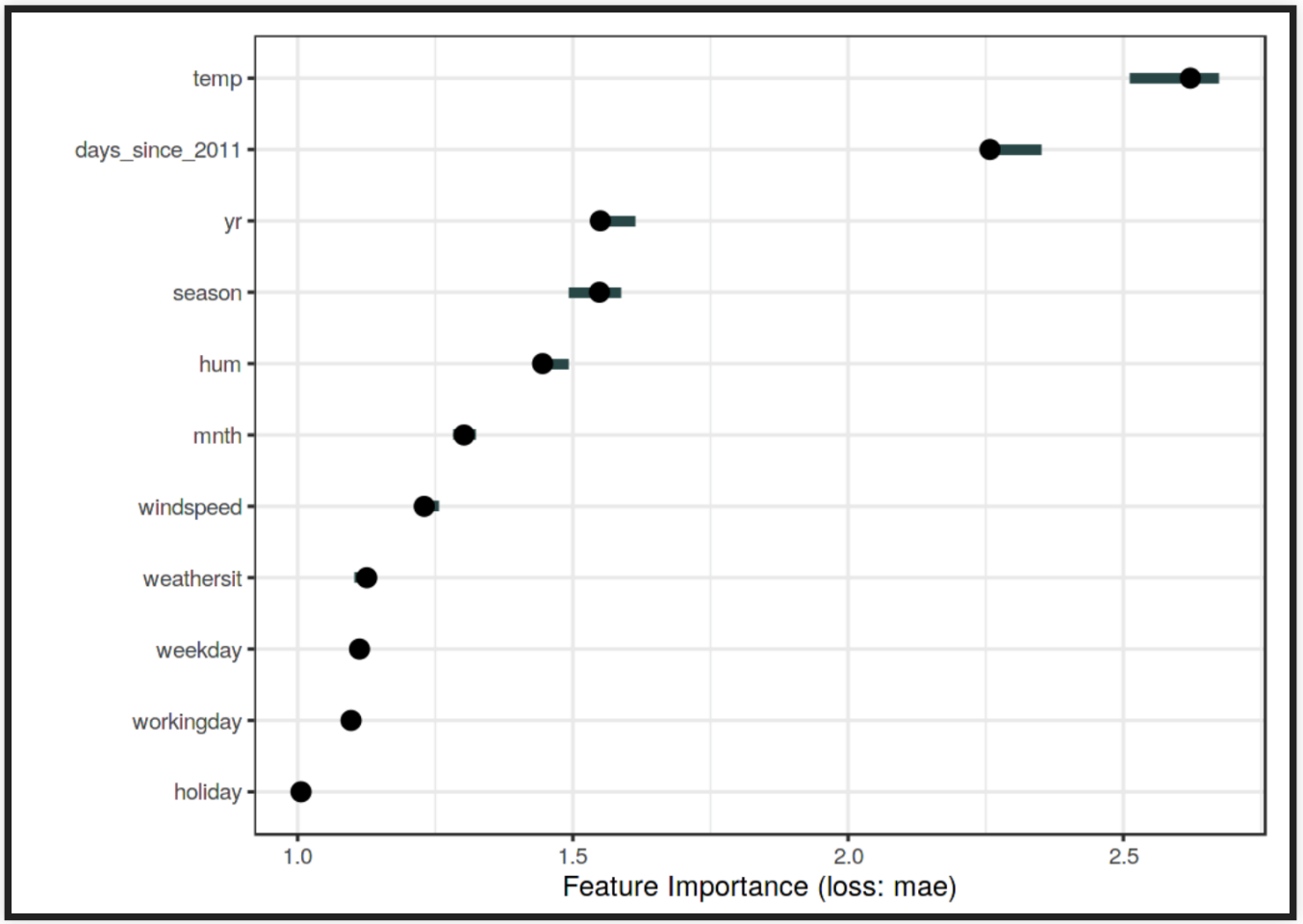
- Highly compressed, global insights
- Effect for feature + interactions
- Can only be computed on labeled data, depends on model accuracy, randomness from permutation
- May produce unrealistic inputs when correlations exist

**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

# INVARIANTS AND 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 mined rules

- Compare association rule mining and specification mining
- Rules mined from many observed examples

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

Ernst, Michael D., Jake Cockrell, William G. Griswold, and David Notkin. "Dynamically discovering likely program invariants to support program evolution." IEEE Transactions on Software Engineering 27, no. 2 (2001): 99-123.

# EXCURSION: DAIKON FOR DYNAMIC DETECTION OF LIKELY INVARIANTS

- Software engineering technique to find invariants
  - e.g., `i>0`, `a==x`, `this.stack != null`, `db.query() after db.prepare()`
  - Pre- and post-conditions of functions, local variables
- Uses for documentation, avoiding bugs, debugging, testing, verification, repair
- Idea: Observe many executions (instrument code), log variable values, look for relationships (test many possible invariants)

# DAIKON EXAMPLE

```
int ABS(int x) {
    if (x>0) return x;
    else return (x*(-1));
}
int main () {
    int i=0;
    int abs_i;
    for (i=-5000;i<5000;i++)
        abs_i=ABS(i);
}
```

Invariants found

```
std.ABS(int;):::EXIT1
x == return

std.ABS(int;):::EXIT2
return == - x

std.ABS(int;):::EXIT
x == orig(x)
x <= return
```

Speaker notes

many examples in https://www.cs.cmu.edu/~aldrich/courses/654-sp07/tools/kim-daikon-02.pdf

# 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 |

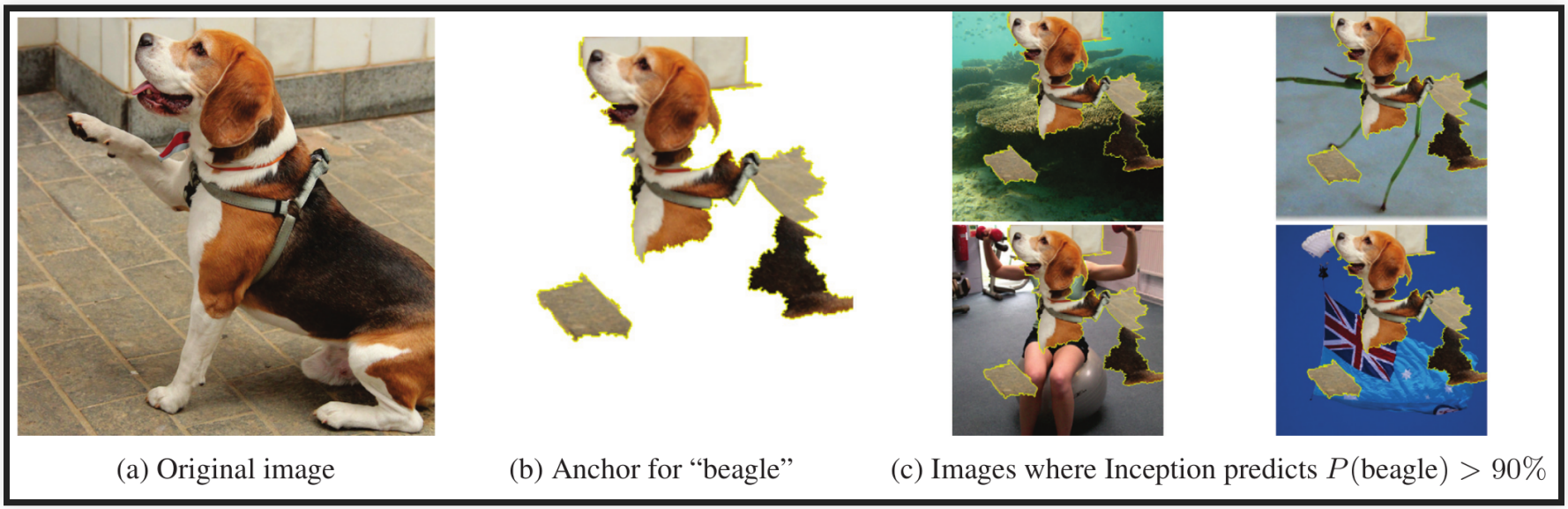
Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

# 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.

# DISCUSSION: ANCHORS AND INVARIANTS

- Anchors provide only partial explanations
- Help check/debug functioning of system
- Anchors usually probabilistic, not guarantees

# EXAMPLE-BASED EXPLANATIONS

(thinking in analogies and contrasts)

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

# 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

# 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 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

# GAMING/ATTACKING THE MODEL WITH EXPLANATIONS?

*Does providing an explanation allow customers to 'hack' the system?*

- Loan applications?
- Apple FaceID?
- Recidivism?
- Auto grading?
- Cancer diagnosis?
- Spam detection?

# GAMING THE MODEL WITH EXPLANATIONS?

# GAMING THE MODEL WITH EXPLANATIONS?

- A model prone to gaming uses weak proxy features
- Protections requires to make the model hard to observe (e.g., expensive to query predictions)
- Protecting models akin to "security by obscurity"

- Good models rely on hard facts that are hard to game and relate causally to the outcome

```
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
```

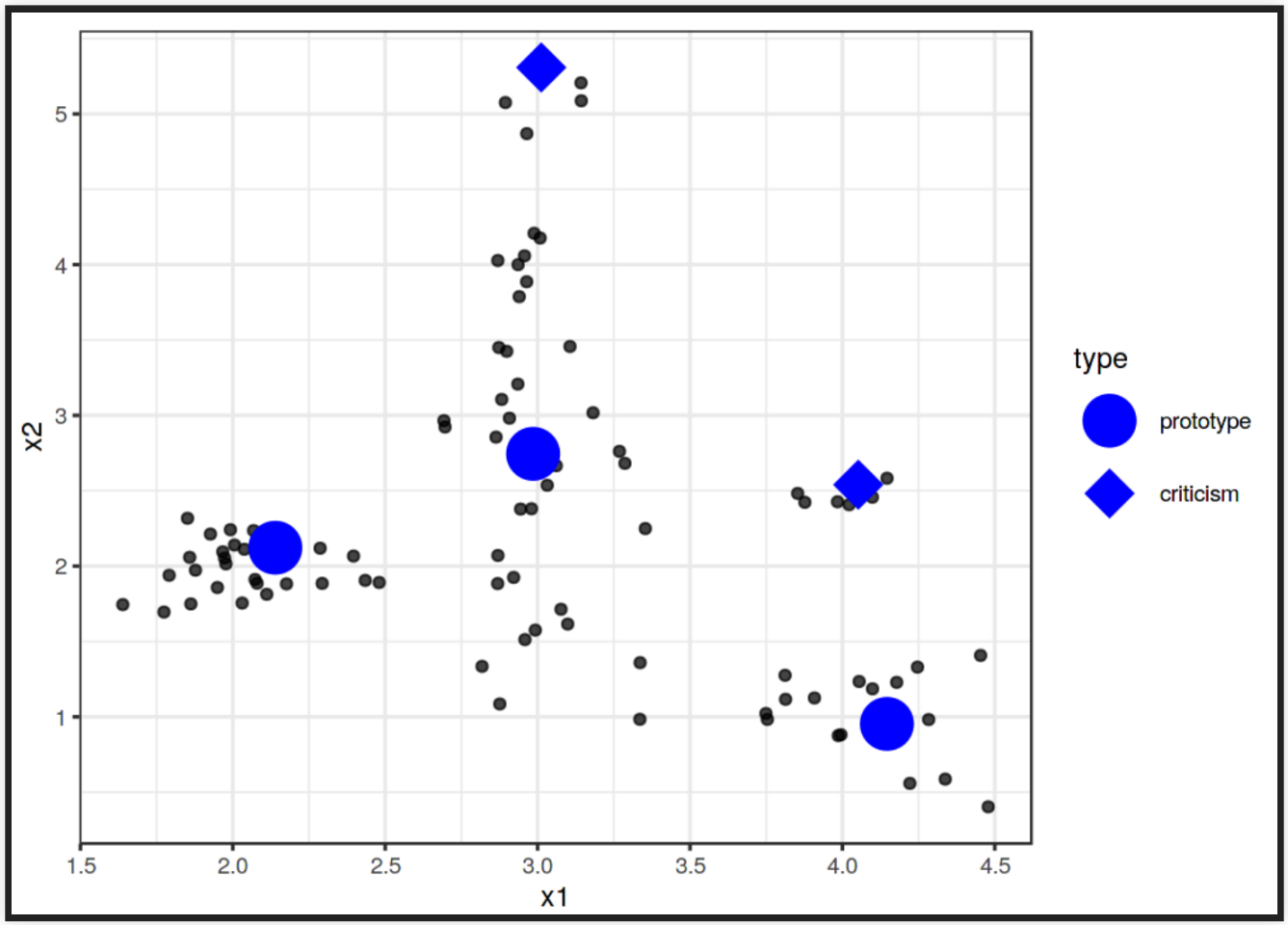
# PROTOTYPES AND CRITICISMS

*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

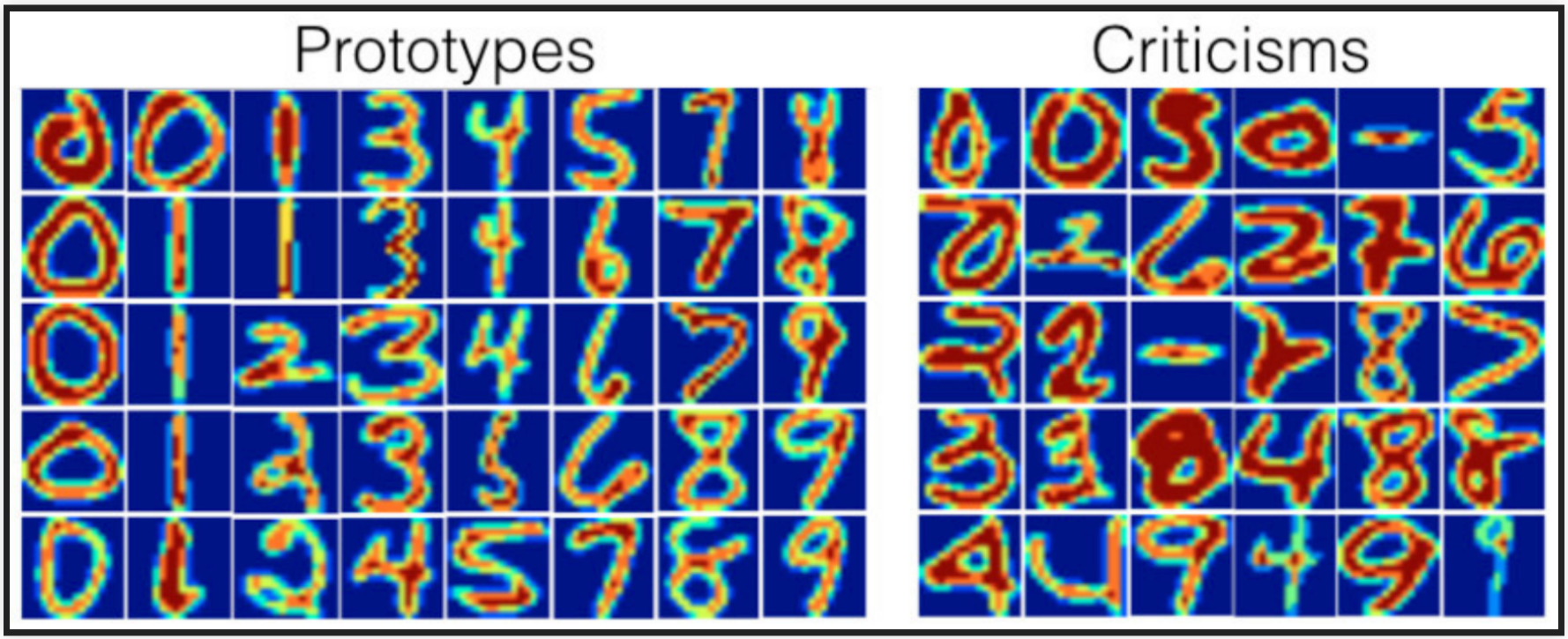



Source: Christoph Molnar. "Interpretable Machine Learning." 2019

# 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!**

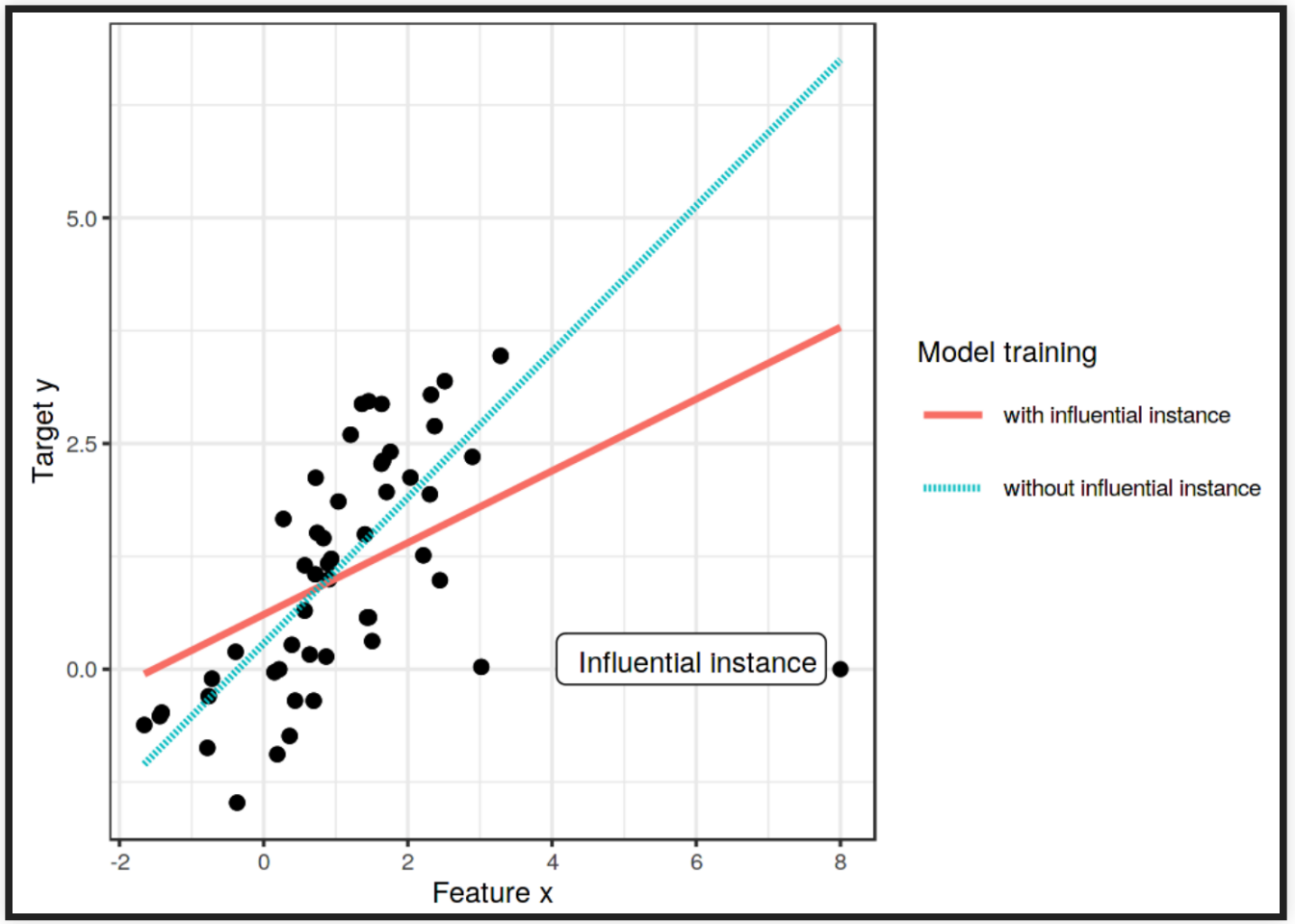
*What data most influenced the training? Is the model skewed by few outliers?*

- 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

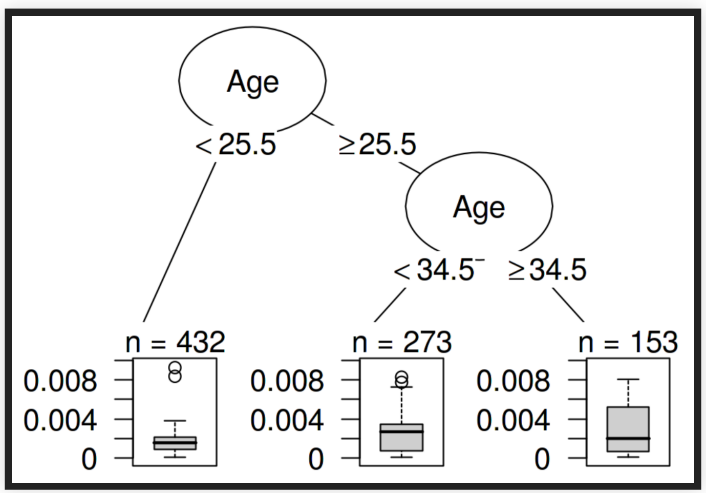
# EXAMPLE: INFLUENTIAL INSTANCE




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

# WHAT DISTINGUISHES AN INFLUENTIAL INSTANCE FROM A NON-INFLUENTIAL INSTANCE?

Compute influence of every data point and create new model to explain influence in terms of feature values




(cancer prediction example)

*Which features have a strong influence but little support in the training data?*

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

Speaker notes

Example from cancer prediction. The influence analysis tells us that the model becomes increasingly unstable when predicting cancer for higher ages. This means that errors in these instances can have a strong effect on the model.

# DEBUGGING DRIFT WITH INFLUENTIAL INSTANCES

- Which data points on the training data influenced the model to work less on newer production data?
- Identify influential training instances on recent production misclassification
- Example: Cancer prediction model built in one hospital but works less well in other hospital
    - Is there training data that causes poor generalization? What are the characteristics of that data (e.g., different age groups)? Are differences due to concept or data drift?

# SELECTIVELY CHECKING DATA QUALITY WITH INFLUENTIAL INSTANCES

- Labeled data comes in different qualities (see data programming lecture)
- Double check labels of influential instances; lower quality labels may be sufficient for less influential instances

# 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

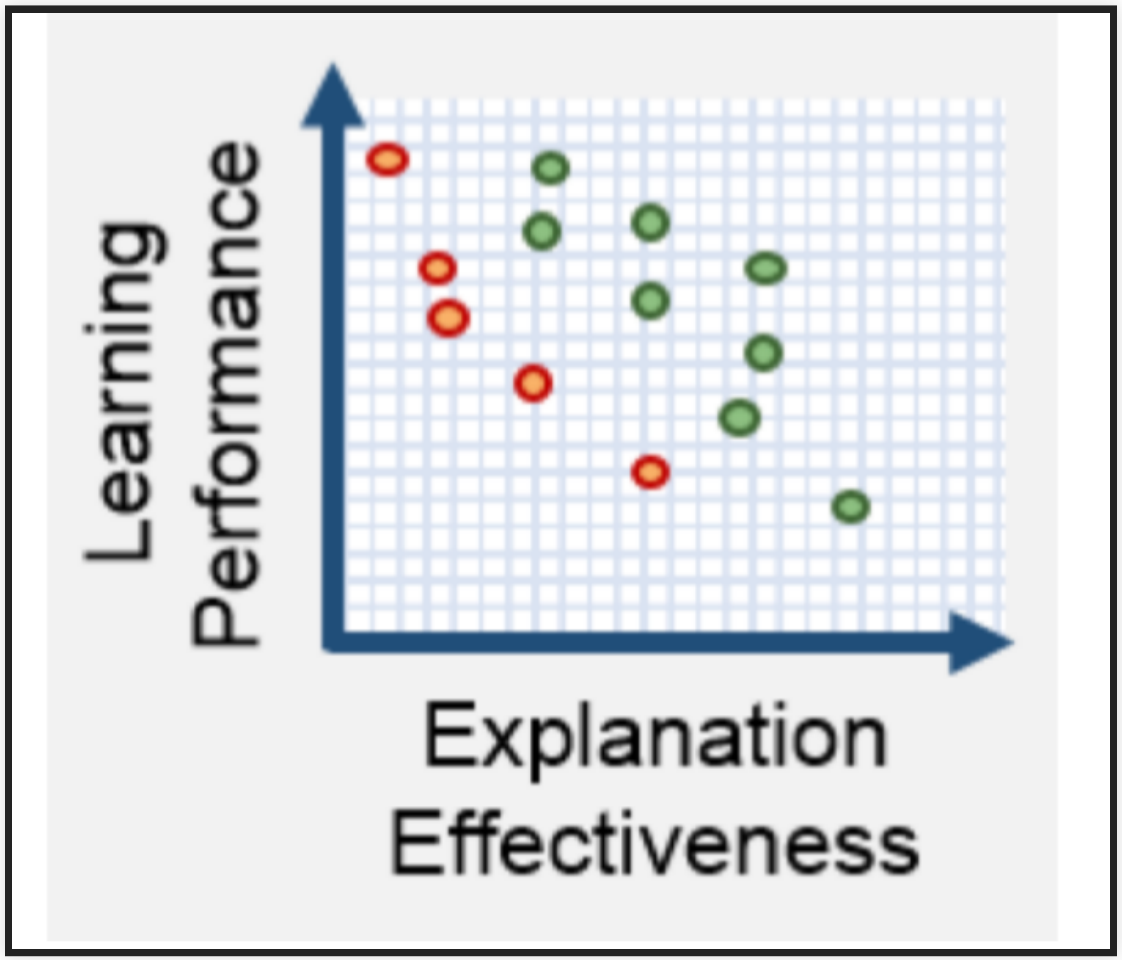
- Useful model-agnostic debugging tool for models and data

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

# "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 (within 2 years) |
| 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

# CALL FOR TRANSPARENT AND AUDITED MODELS

> *"no black box should be deployed when there exists an interpretable model with the same level of performance"*

- High-stakes decisions with government involvement (recidivism, policing, city planning, ...)
- High-stakes decisions in medicine
- High-stakes decisions with discrimination concerns (hiring, loans, housing, ...)
- Decisions that influence society and discourse? (content curation on Facebook, targeted advertisement, ...)

*Regulate possible conflict: Intellectual property vs public health/welfare*

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)

# SUMMARY

- Interpretability useful for many scenarios: user feedback, debugging, fairness audits, science, ...
- Defining and measuring interpretability
- Inherently interpretable models: sparse regressions, shallow decision trees, ...
- Providing ex-post explanations of blackbox models
    - global and local surrogates
    - dependence plots and feature importance
    - invariants (anchors)
    - counter-factual explanations
- Data debugging with prototypes, criticisms, and influential instances
- Considerations for high-stakes decisions