Explainable and Usable AI

Explain Your Decisions

Positive or Negative Review?

all the amped up tony hawk style stunts and thrashing rap-metal can't disguise the fact that, really, we've been here, done that.

Dog or Cat?

Likely to get arrested again in next 2 years?

The defendant is a Male aged 38. They have been charged with: Battery. This crime is classified as a Misdemeanor. They have been convicted of 0 prior crimes. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

Sources: Review: example from https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis. Recidivism: Random sample from Broward County, Florida records from 2013 and 2014, based on data acquisition and analysis by ProPublica, displayed as in Dressel and Farid 2018 https://www.science.org/doi/10.1126/sciadv.aao5580. Image: https://www.boredpanda.com/puppy-looks-like-cat-dog-hybrid/

How did you explain it?

Main Points

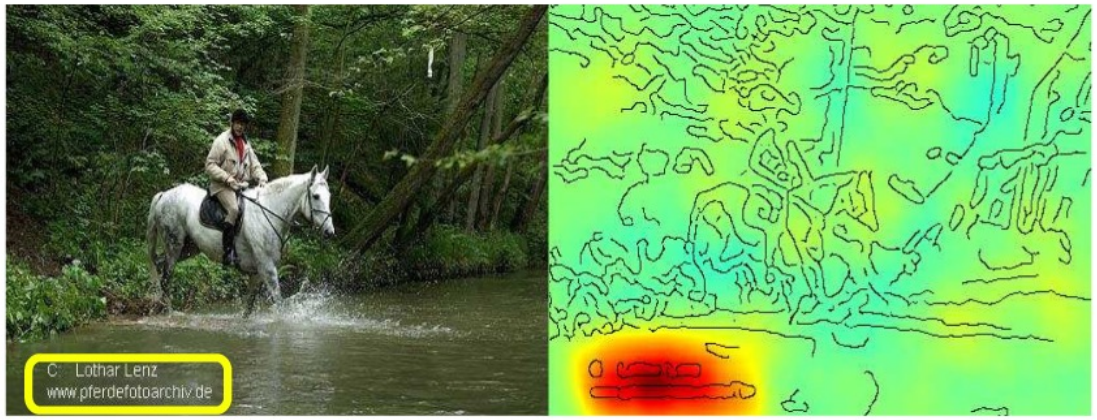
- Explainable ≠ interpretable
- Models don’t have to be black-box to be accurate.
- “Amplify, Augment, Empower, and Enhance People” (Shneiderman)

Why care?

- *harms* when systems aren't **reliable**, **safe**, **trustworthy**
- *benefits* when systems **empower** people

Horse-picture from Pascal VOC data set

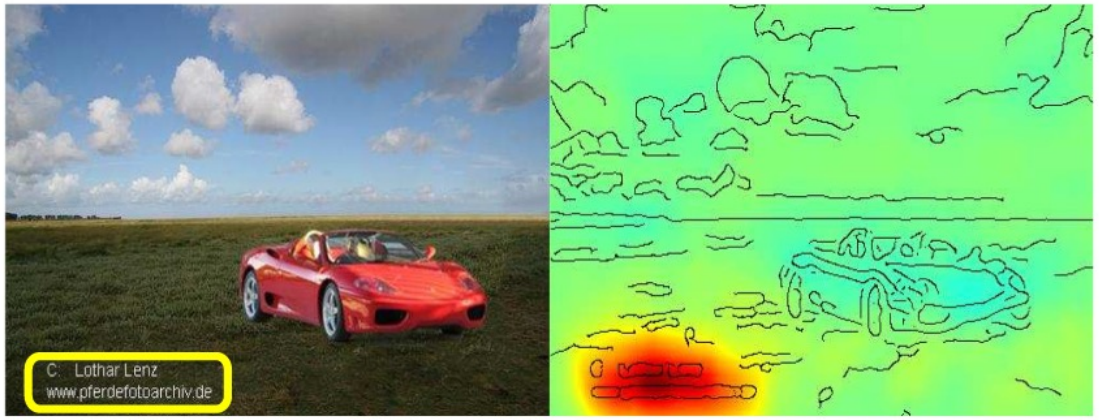
Artificial picture of a car

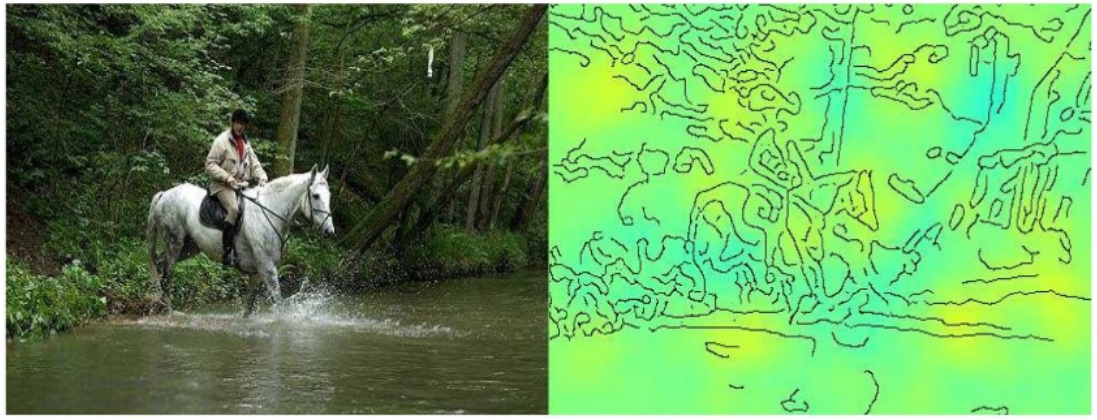


Source tag present

↓

Classified as horse





No source tag present

↓

Not classified as horse

Lapuschkin et al. 2019. [Unmasking Clever Hans predictors and assessing what machines really learn]

Some Interpretable Models

All examples from Rudin et al. 2020, Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Rule lists and scoring systems

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.	

Patient screens positive for obstructive sleep apnea if Score >1			
1.	age $\geq$ 60	4 points	. . . . .
2.	hypertension	4 points	+ . . . . .
3.	body mass index $\geq$ 30	2 points	+ . . . . .
4.	body mass index $\geq$ 40	2 points	+ . . . . .
5.	female	-6 points	+ . . . . .
	Add points from row 1-6	Score	= . . . . .

Angelino et al. 2017. Learning Certifiably Optimal Rule Lists for Categorical Data

Generalized Additive Models (GAM)

Score: risk of having diabetes.

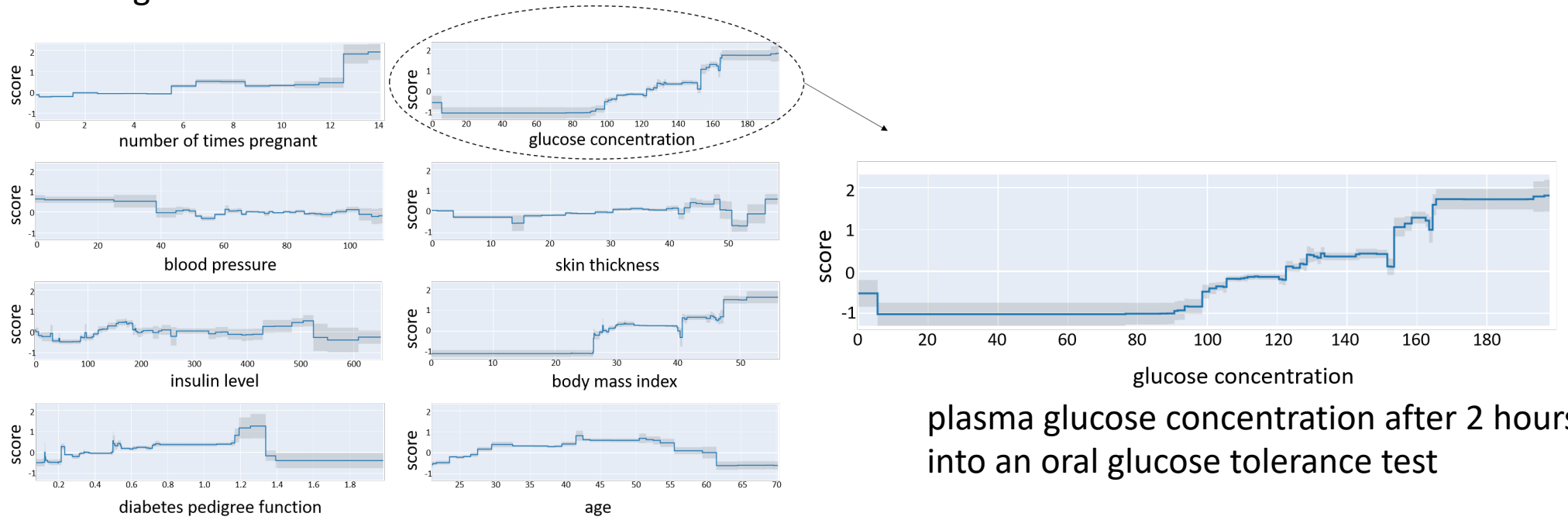
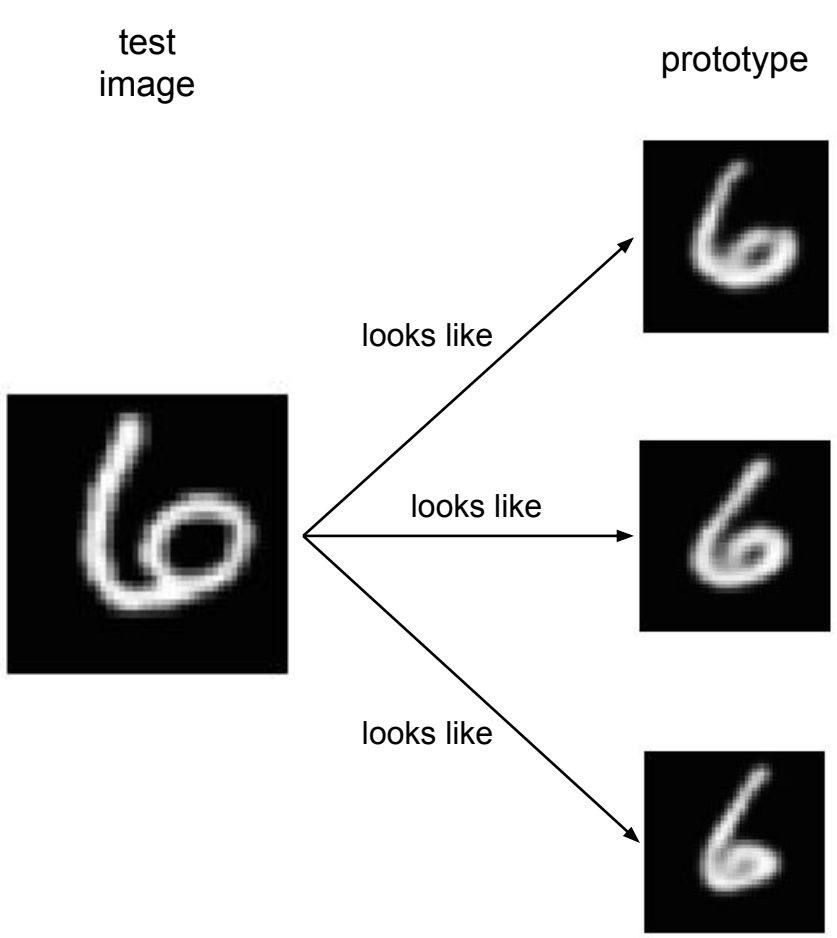
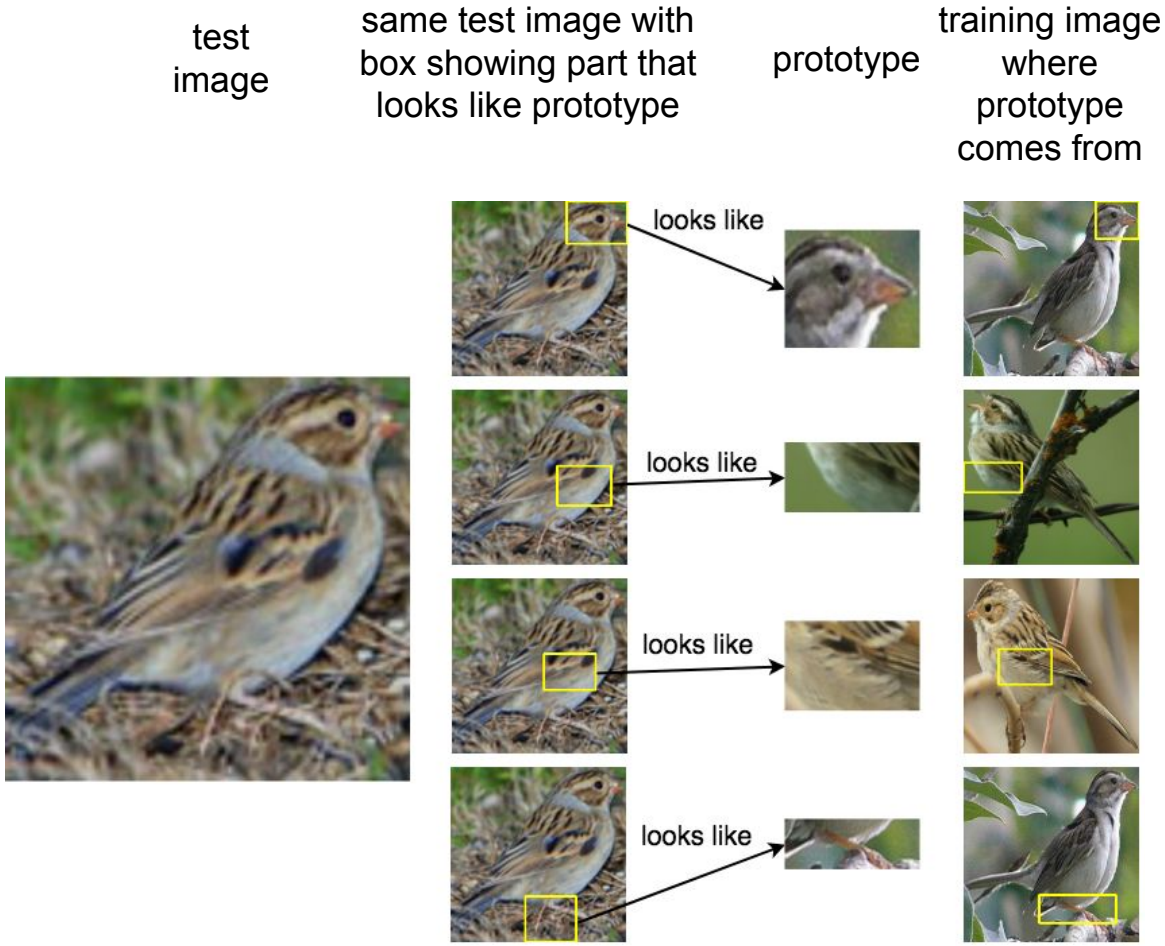


Figure 5: Left: All component functions of a GAM model trained using the `interpret` package (Nori et al., 2019) on a diabetes dataset (Dua and Graff, 2017); Right: zoom-in of component function for glucose concentration.

Prototype-based, part-based

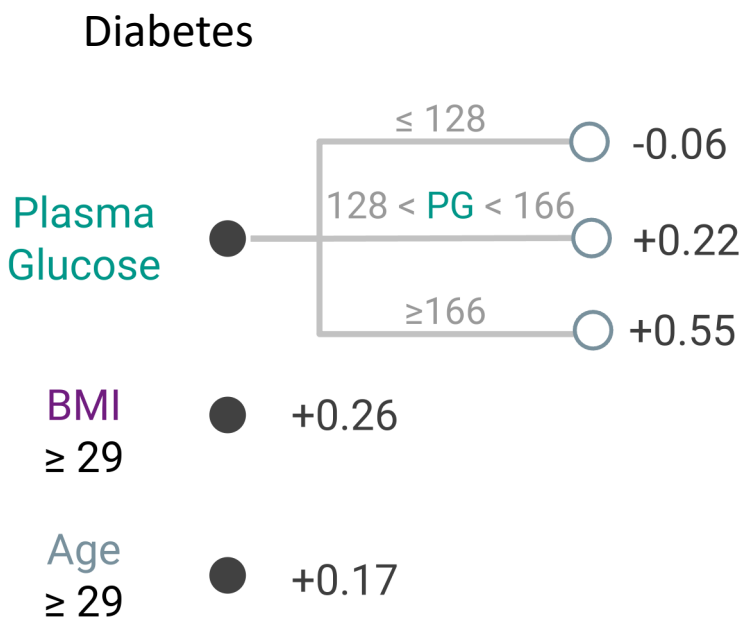




Sums of Trees

- Fast Interpretable Greedy-Tree Sums (FIGS)
- Go through each tree independently
- Sum the outputs of each tree

- Interpretable: each tree can be shallow.



What if we need a black box model?

Can we “explain” such a model?

Dimensionality Reduction

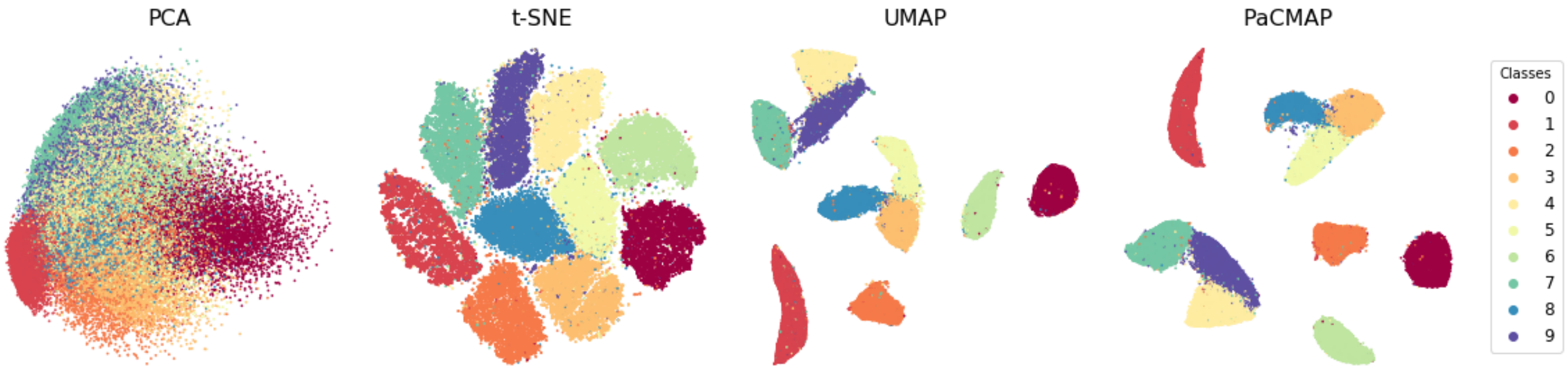
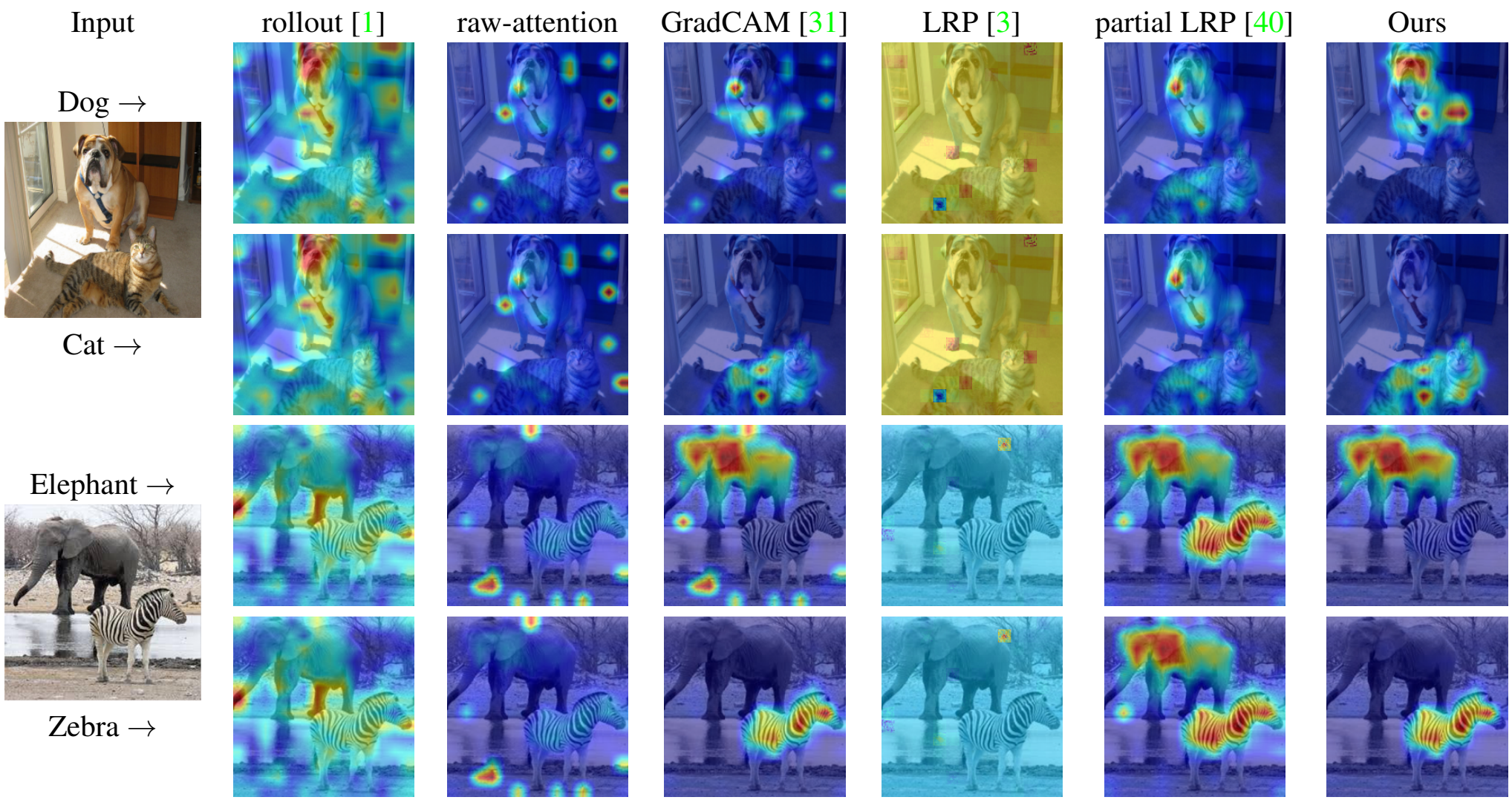


Figure 12: Visualization of the MNIST dataset (LeCun et al., 2010) using different kinds of DR methods: PCA (Pearson, 1901), t-SNE (van der Maaten and Hinton, 2008; Linderman et al., 2019; Poličar et al., 2019), UMAP (McInnes et al., 2018), and PaCMAP (Wang et al., 2020b). The axes are not quantified because these are projections into an abstract 2D space.

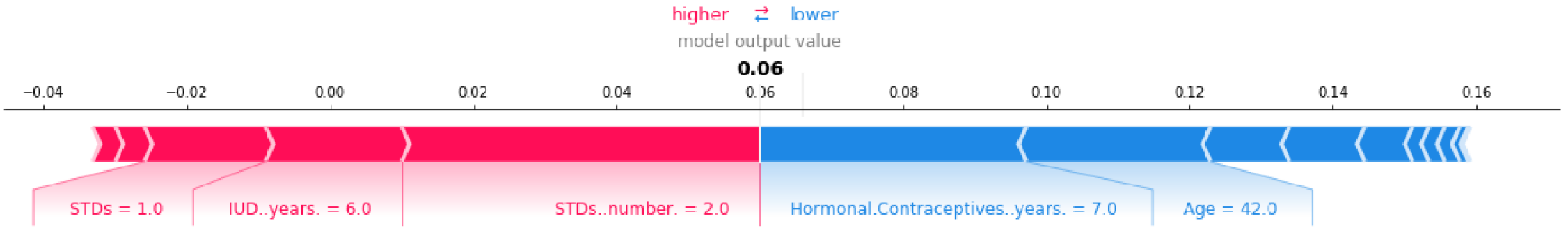


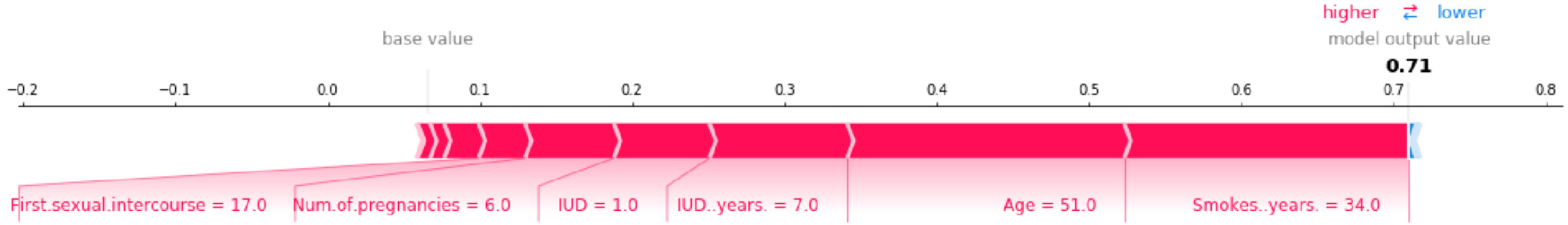
Chefer et al. 2021, [Transformer Interpretability Beyond Attention Visualization]

Other ways to interpret

Shapley Values for Explaining Predictions

Intuition: average effect of having that feature vs leaving it out





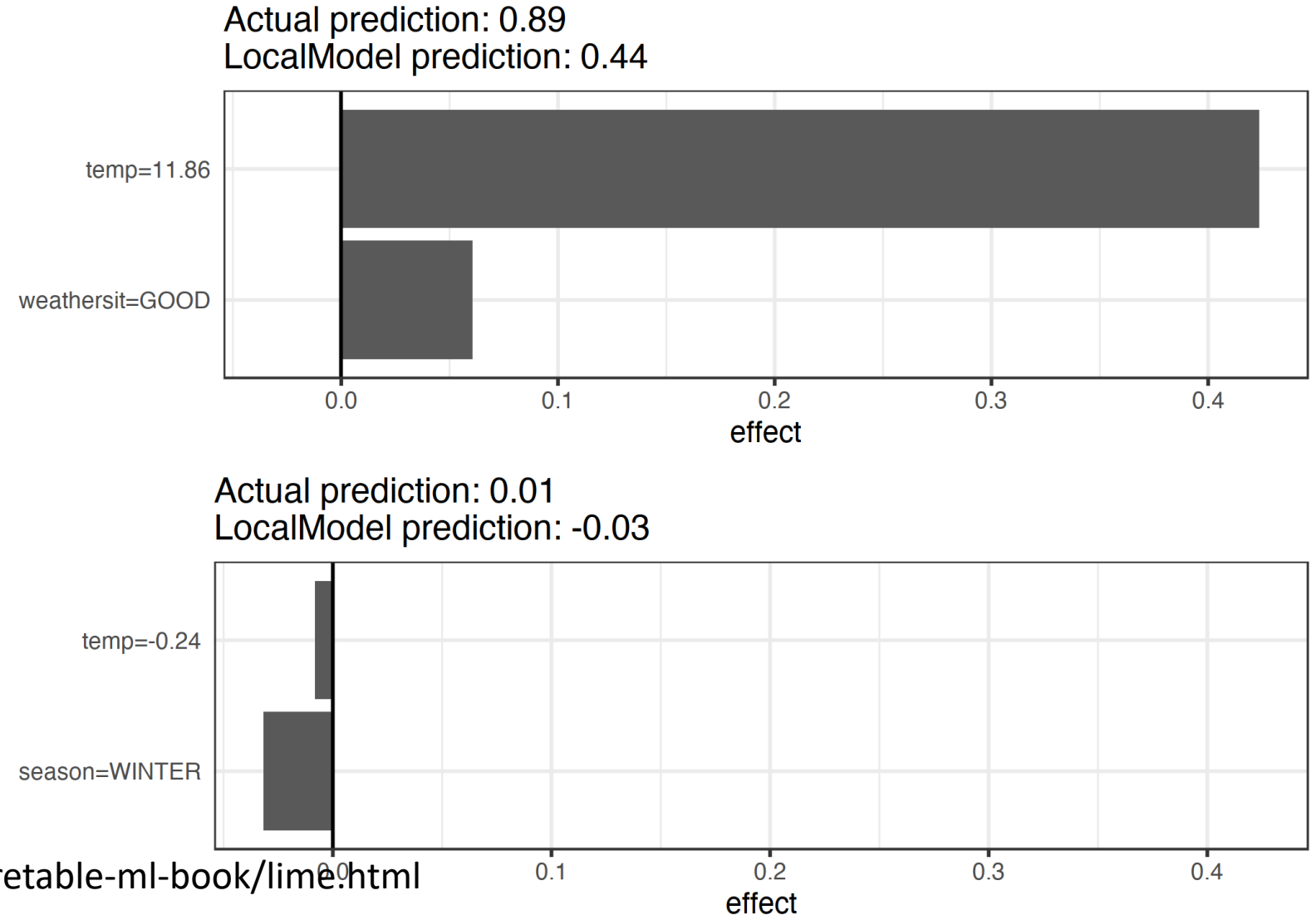
SHAP values to explain the predicted cervical cancer probabilities of two individuals

Source: https://christophm.github.io/interpretable-ml-book/shap.html

Local Interpretable Model-agnostic Explanations

Intuition: fit a simple model in the “neighborhood” of an example

- Example: predict if a day will have more or fewer bike rentals than average
- Main model: random forest
- Surrogate model: logistic reg, 2 features



https://christophm.github.io/interpretable-ml-book/lime.html

Explainable
vs
Interpretable

Explainable: why did the black box model gave us this answer?

Interpretable: the model isn't a black box

HCAI Attributes that Are Candidates for Assessment

General virtues of the system itself

- **Trustworthy**: Can users trust the system to perform correctly?
- **Responsible/Humane**: Has the system been designed, developed, and tested in a responsible way?
- **Ethical Design**: Were stakeholders involved in the design?
- **Ethical Data**: Was the data collected in an ethical manner?
- **Ethical Use**: Will the system's outcome be used in an ethical manner?
- **Well-being/Benevolence**: Does the system support human health, comfort, and values?
- **Secure**: How vulnerable is the system to attack?
- **Private**: Does the system protect a person's identity and data?

Performs well in practice

- **Robust/Agile**: Does the system perform well when inputs change?
- **Reliable/Dependable**: Does the system do the right thing?
- **Available**: Is the system running when needed?
- **Resilient/Adaptive**: Can the system recover from disruptions?
- **Testable/Verifiable/Validatable/Certifiable:** Can be tested to verify adherence to requirements?
- **Safe**: Does the system have a history of safe use?

Clarity to stakeholders

- **Accurate**: Does the system deliver correct results on test cases and real world cases?
- **Fair/Unbiased**: Are the system's biases understood and reported?
- **Accountable/Liable**: Who or what is responsible for the system's outcome?
- **Transparent**: Is it clear to an external observer how the system's outcome was produced?
- **Interpretable/Explainable/Intelligible/Explicable:** Can the system explain the outcome?
- **Usable**: Can a human use it easily?

Enables independent oversight

- **Auditable**: Can the system be audited by others for retrospective forensic analysis of failures?
- **Trackable:** Does the system display status and next steps so human intervention is possible?
- **Traceable:** Is the system designed to allow tracing back from an outcome to the root cause?
- **Redressable**: Is there a process for those harmed to request review and compensation?
- **Insurable**: Does the design permit insurance companies to offer policies?
- **Recorded**: Does the system record activity for retrospective forensic review?
- **Open**: Is code and data publicly available for others to review?
- **Certifiable**: Can it be certified and approved for use?

Complies with accepted practices

- **Compliant with standards**: Does the system comply with relevant standards, e.g. IEEE P7000 series?
- **Compliant with accepted software engineering workflows:** Was a trusted process used?

Design Guidelines

Eight Golden Rules
1. Strive for consistency
2. Seek universal usability
3. Offer informative feedback
4. Design dialogs to yield closure
5. Prevent errors
6. Permit easy reversal of actions
7. Keep users in control
8. Reduce short-term memory load

Eight Silver Slogans for HCAI Systems
1. Store rich data from powerful sensors
2. Design information abundant displays
3. Provide interactive information visualization
4. Make predictive models visual
5. Smooth human-to-human communication
6. Create clear control panels
7. Implement audit trails
8. Develop incident reporting websites

https://www.cs.umd.edu/~ben/goldenrules.html