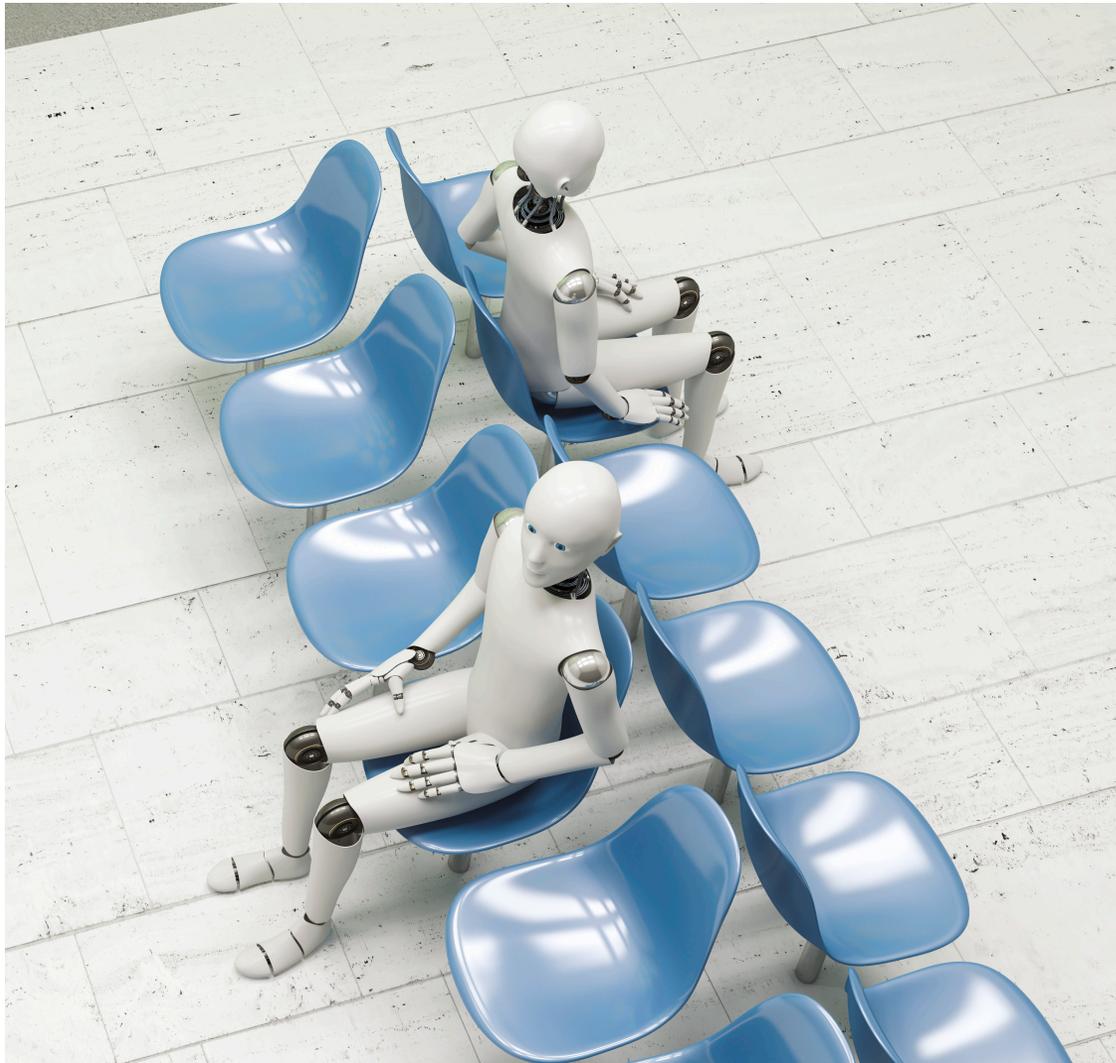


Human-AI Interaction

CS 344 2021-04-26



deep learning | ✕

- deep learning **stocks**
- deep learning **vs machine learning**
- deep learning **ai**
- deep learning **book**
-  **Deep Learning with Python**
Book by François Chollet
- deep learning **definition**
- deep learning **algorithms**
- deep learning **interview questions**
- deep learning **neural networks**
- deep learning **applications**

Enter amounts to request mortgage:

Mortgage amount requested	<input type="text" value="375000"/>
Household monthly income	<input type="text" value="7000"/>
Liquid assets	<input type="text" value="48000"/>

We're sorry, your mortgage loan was not approved. You might be approved if you reduce the Mortgage amount requested, increase your Household monthly income, or increase your Liquid assets.

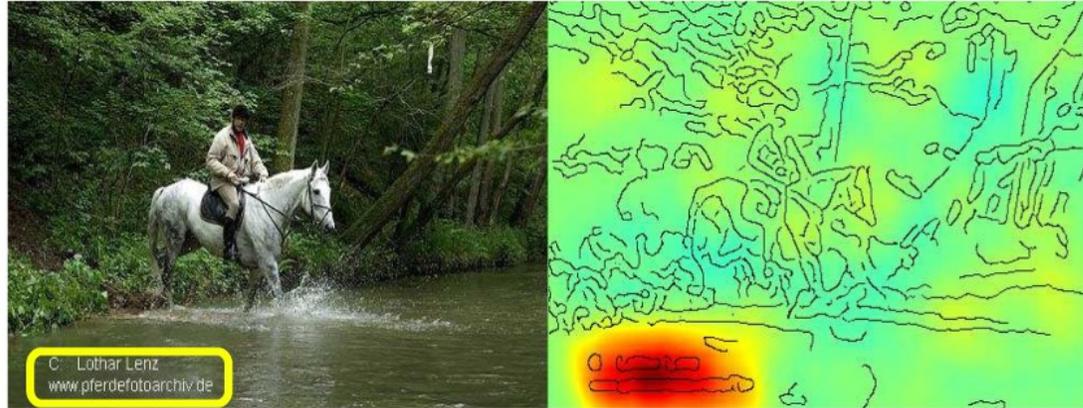
Main Points

- Explainable \neq 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

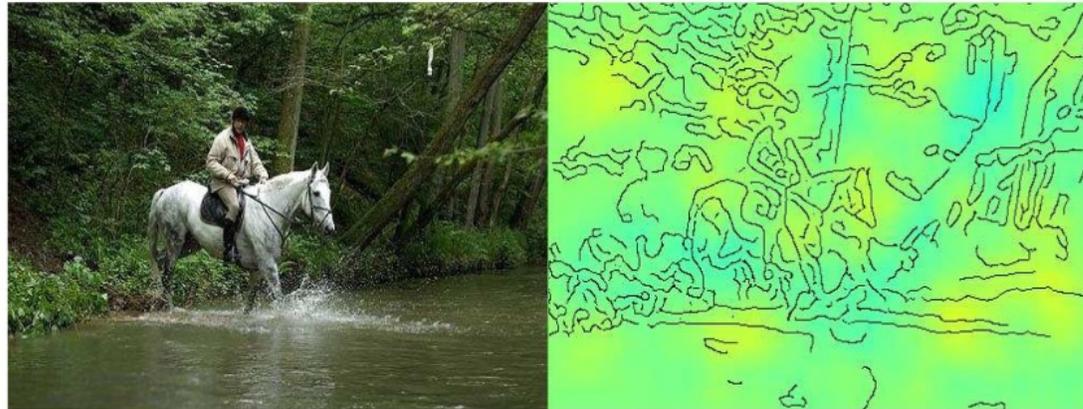
Horse-picture from Pascal VOC data set



Source tag
present



Classified
as horse



No source
tag present



Not classified
as horse

Artificial picture of a car



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	=.....

Generalized Additive Models (GAM)

Score: risk of having diabetes.

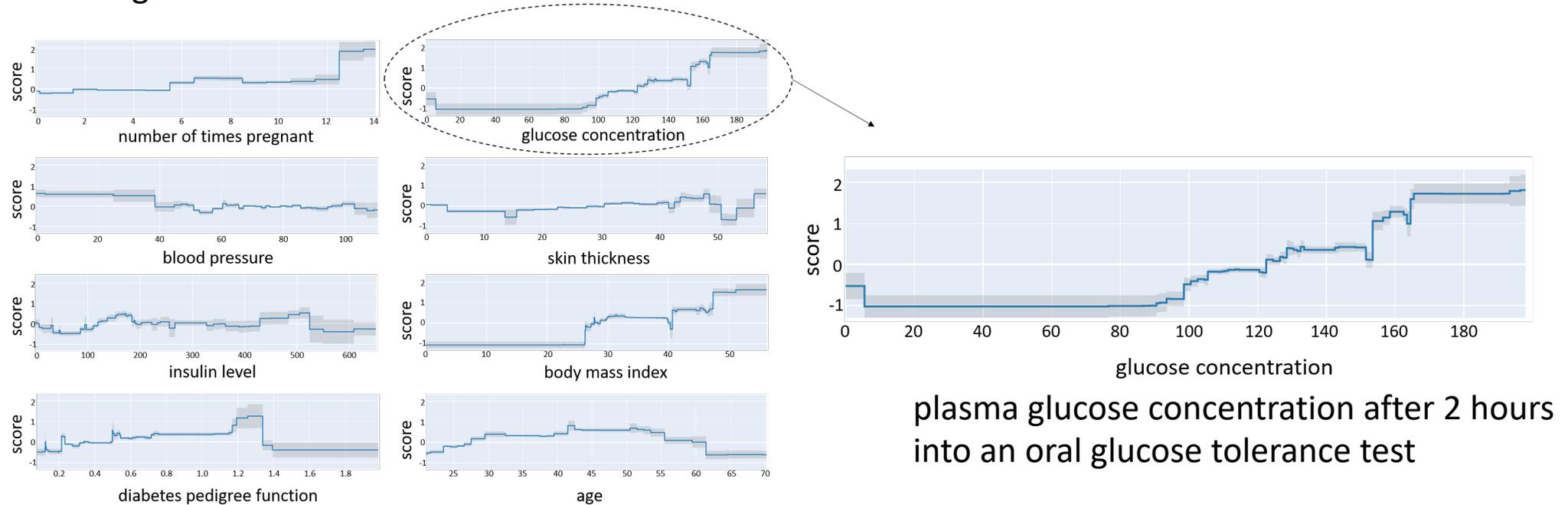
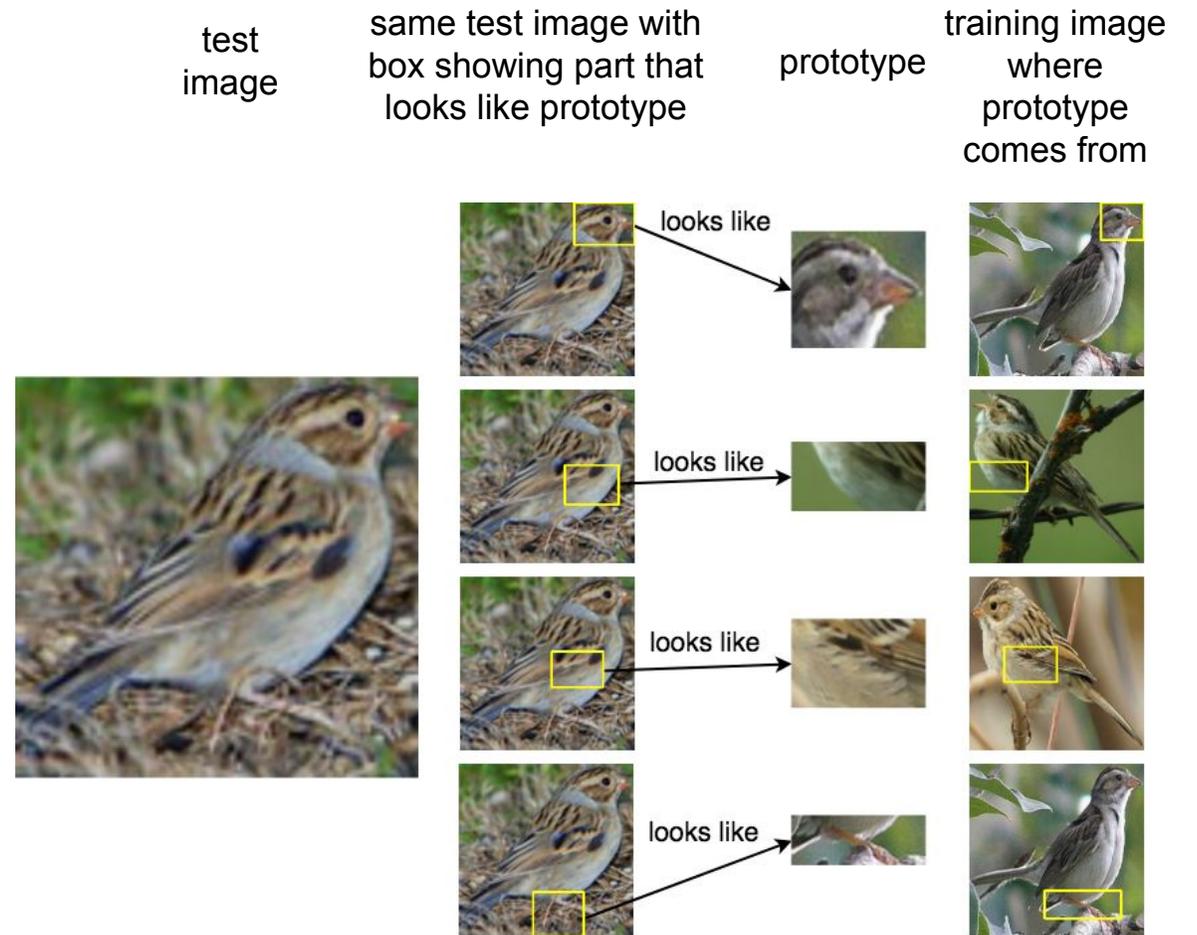
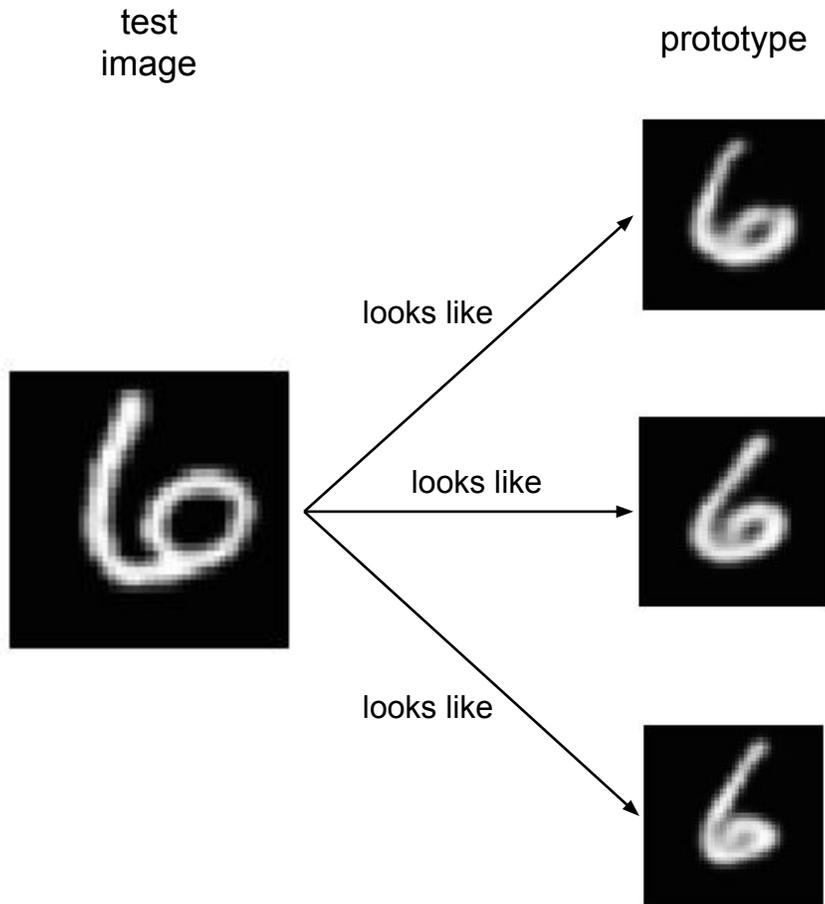


Figure 5: Left: All component functions of a GAM model trained using the `interpret` package (Nori et al., 2019) on a diabetes dataset (Dua and Graff, 2017); Right: zoom-in of component function for glucose concentration.

Prototype-based, part-based



What if we need a black box model?

Can we “explain” such a model?

Dimensionality Reduction

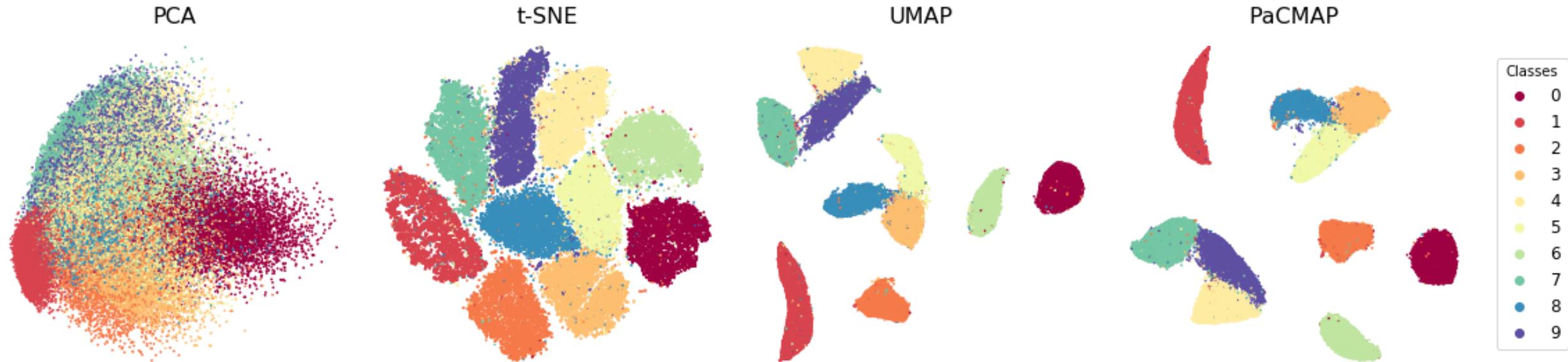
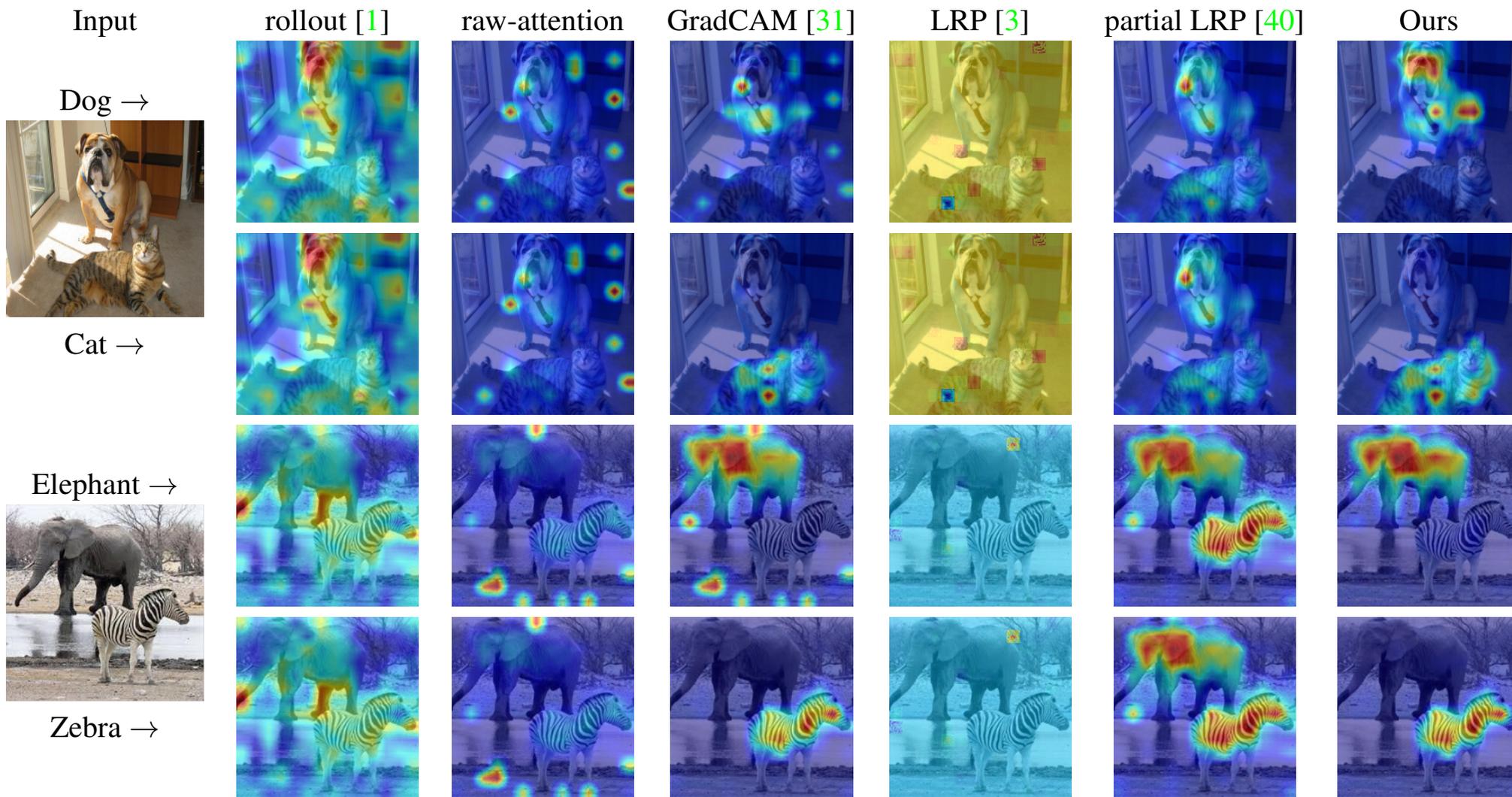


Figure 12: Visualization of the MNIST dataset (LeCun et al., 2010) using different kinds of DR methods: PCA (Pearson, 1901), t-SNE (van der Maaten and Hinton, 2008; Linderman et al., 2019; Poličar et al., 2019), UMAP (McInnes et al., 2018), and PaCMAP (Wang et al., 2020b). The axes are not quantified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



Explainable
vs
Interpretable

Explainable: why did the black box model give 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