Evaluation of Explainable AI

Sermkiat Lolak , M.D

Promise

- Interpretability VS Explanability
- Evaluation of ML explanation
- Interpretability Method

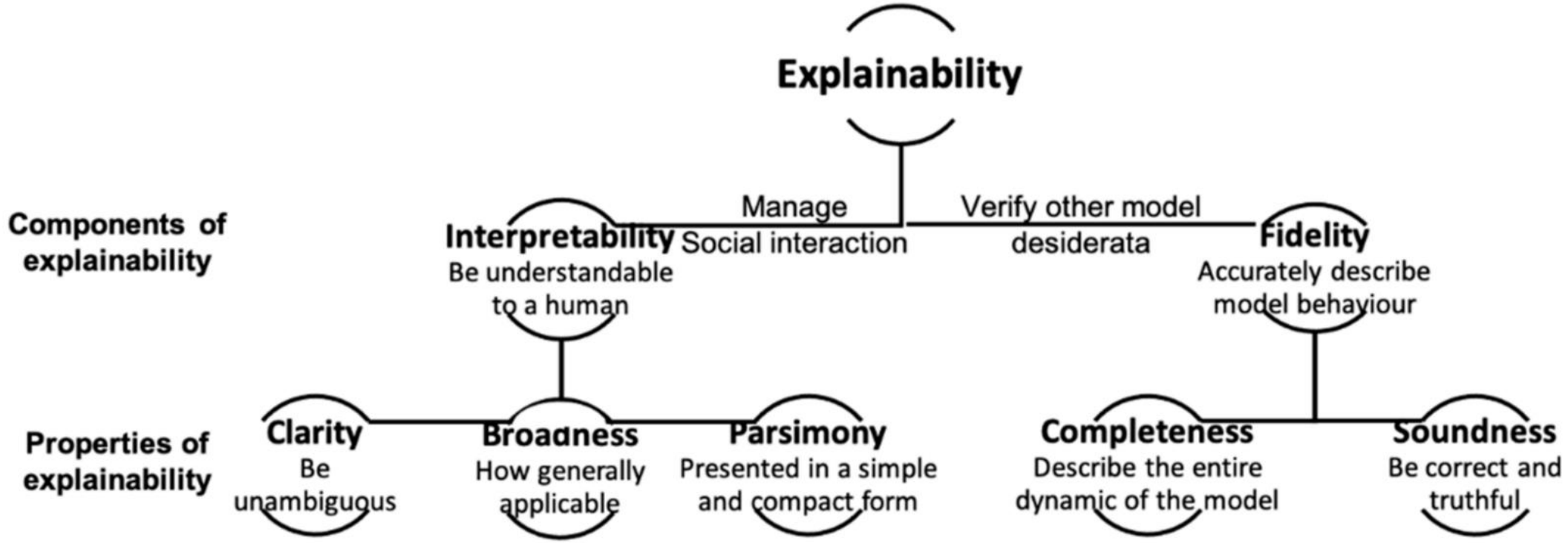
Interpretability VS Explainability

- Interpretability : the ability to explain or to present in understandable terms to a human

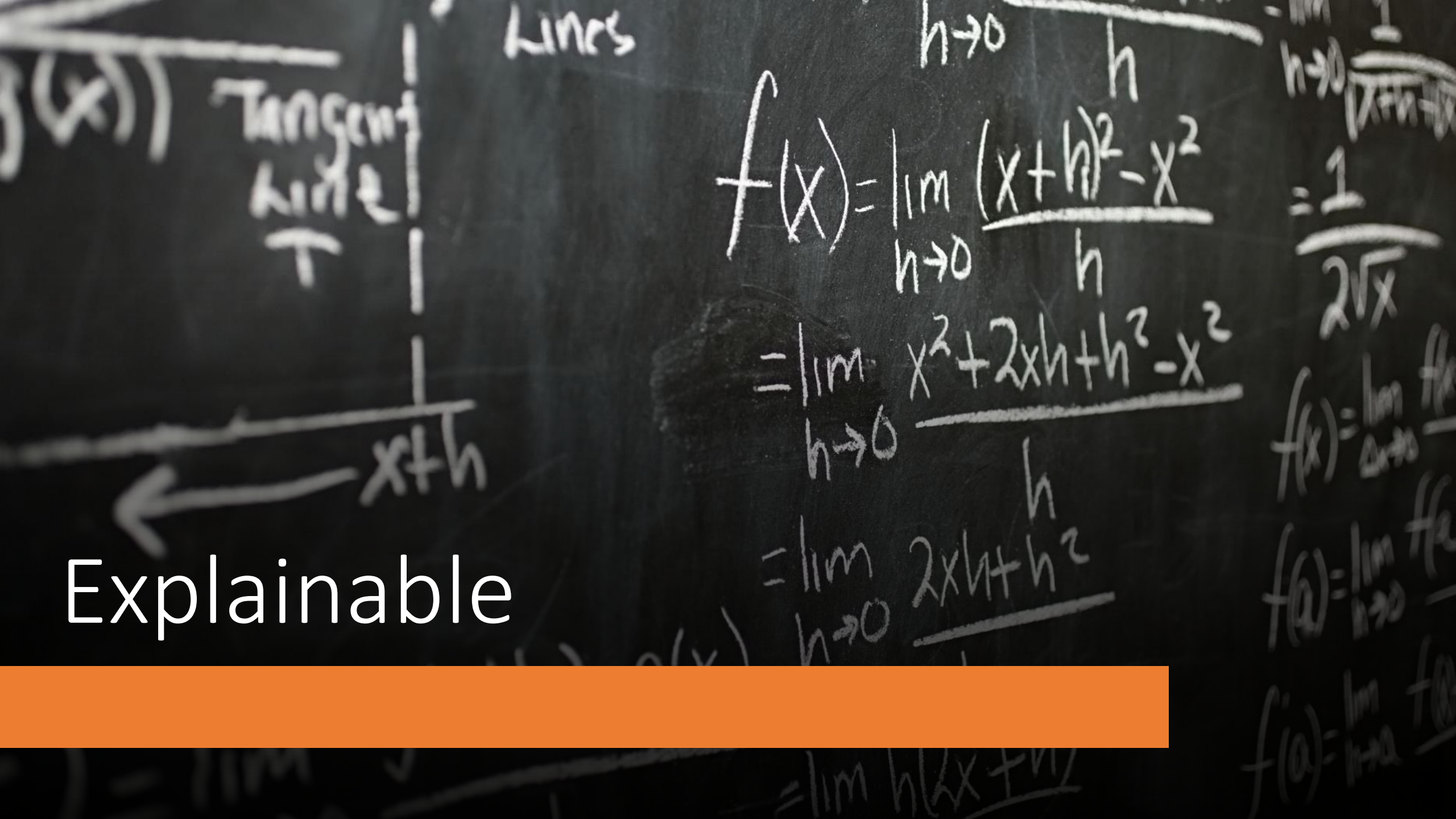
- Explainability :
  - internal logic and mechanics inside a machine learning system.

  - Deeper the understanding that humans achieve in terms of the internal procedures while the model is training or making decisions.

Definition of machine learning (ML) explainability and related properties

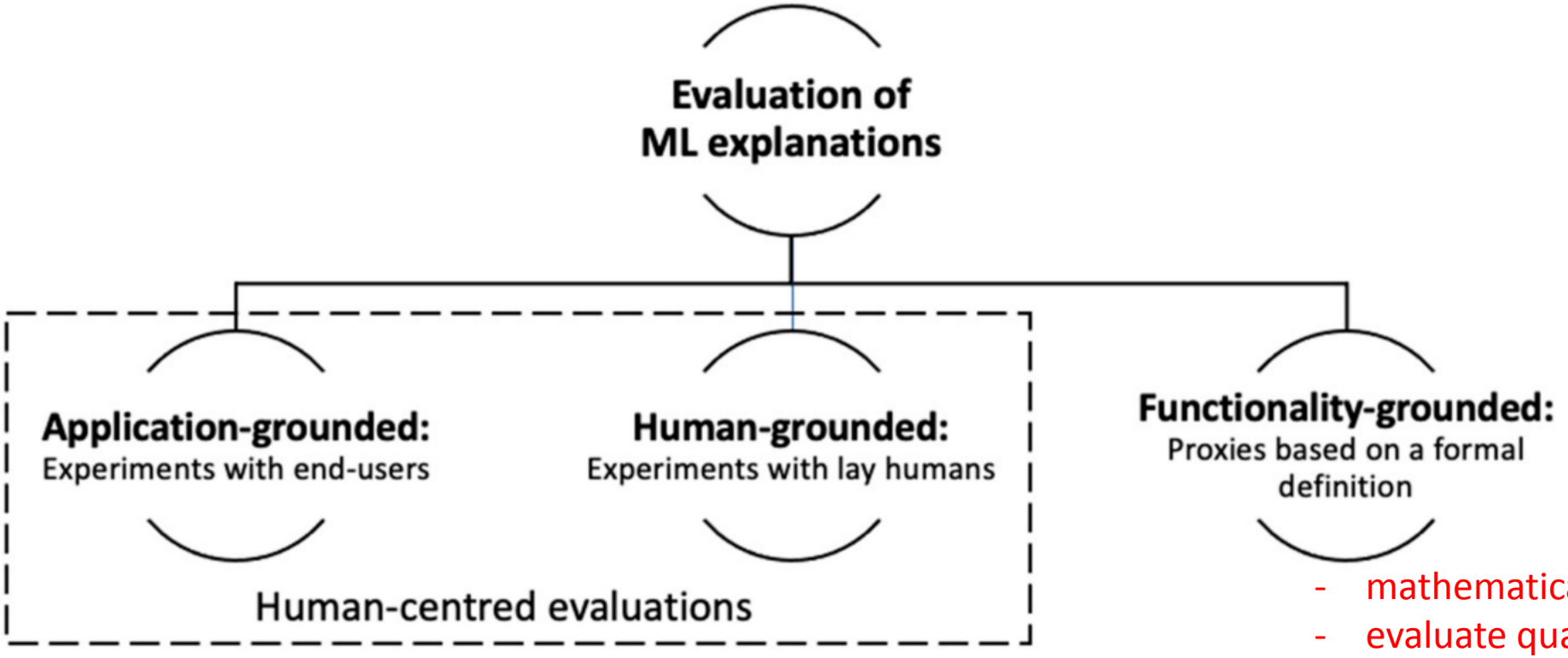


Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. arXiv 2020, arXiv:2007.15911.



Explainable

Taxonomy of evaluation of machine learning interpretability

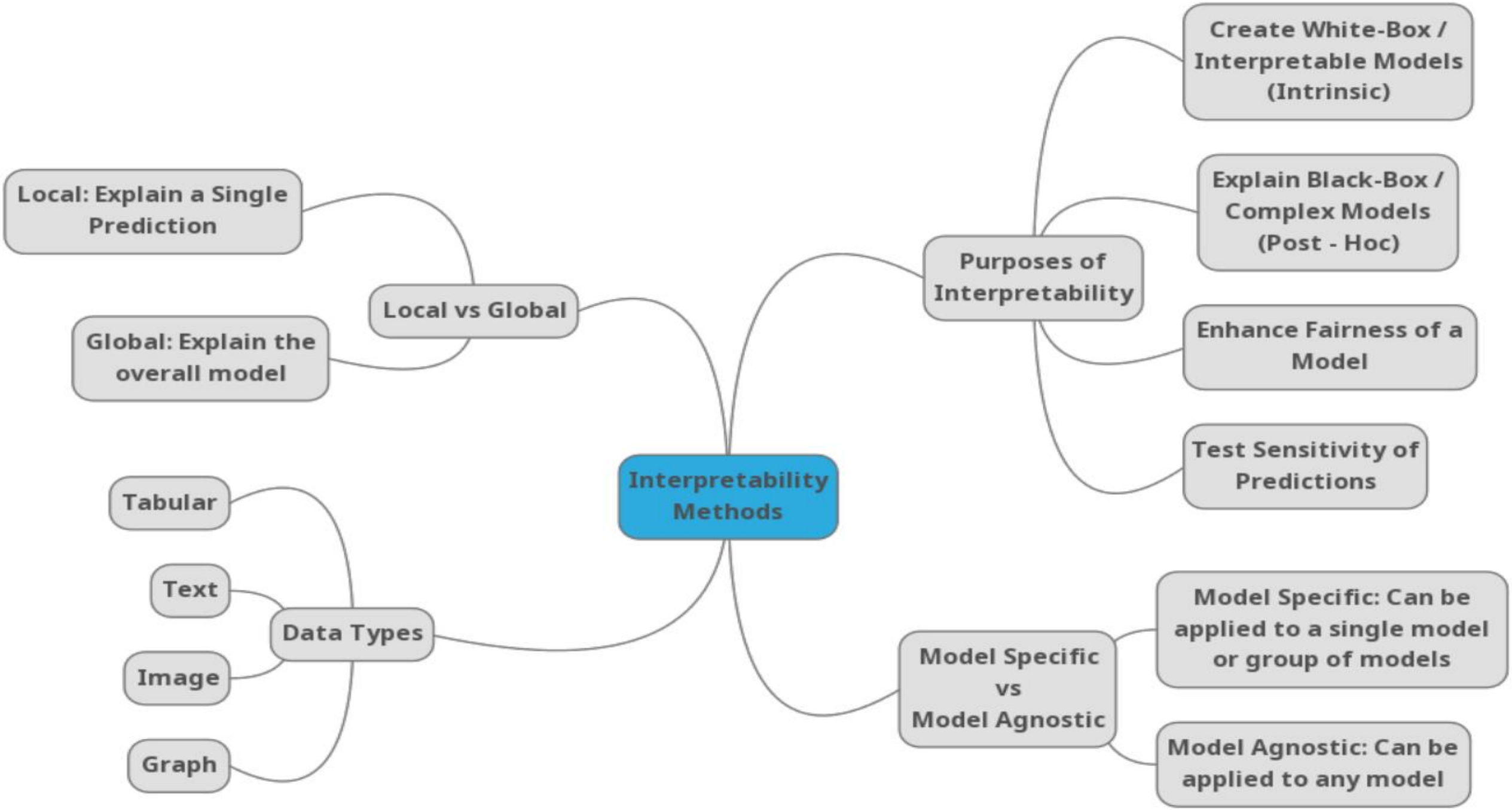


- Expert
- better identification of errors or less discrimination

- General
- test the quality of produced interpretation in a more general setting

- mathematical definitions
- evaluate quality of an interpretability method

Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. arXiv 2017, arXiv:1702.08608.


Interpretability Methods
Local vs Global
Local: Explain a Single Prediction
Global: Explain the overall model
Data Types
Tabular
Text
Image
Graph
Purposes of Interpretability
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

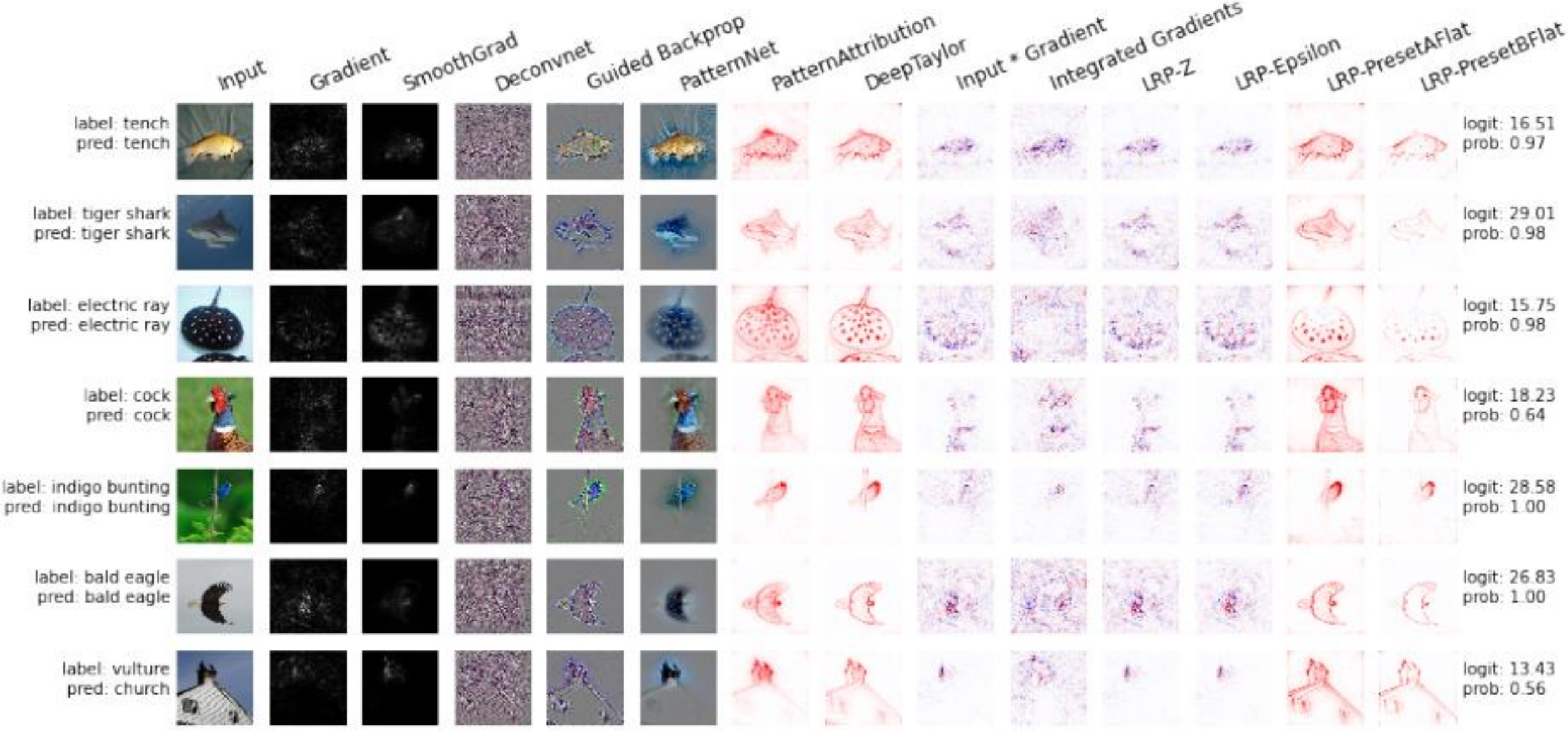
Deep Learning Interpretation

Table 1. Interpretability Methods to Explain Deep Learni

Tool	Category	Local vs. Global	Model Specific v Model Agnostic
DeepExplain iNNvestigate tf-explain	PH	L	Specific
Grad-CAM tf-explain	PH	L	Specific
CAM	PH	L	Specific
iNNvestigate	PH	L	Specific
DeepExplain iNNvestigate tf-explain	PH	L	Specific
DeepExplain iNNvestigate Integrated Gradients tf-explain alibi Skater	PH	L	Specific
Deep Visualization Toolbox	PH	L	Specific

Deep Learning Interpretation

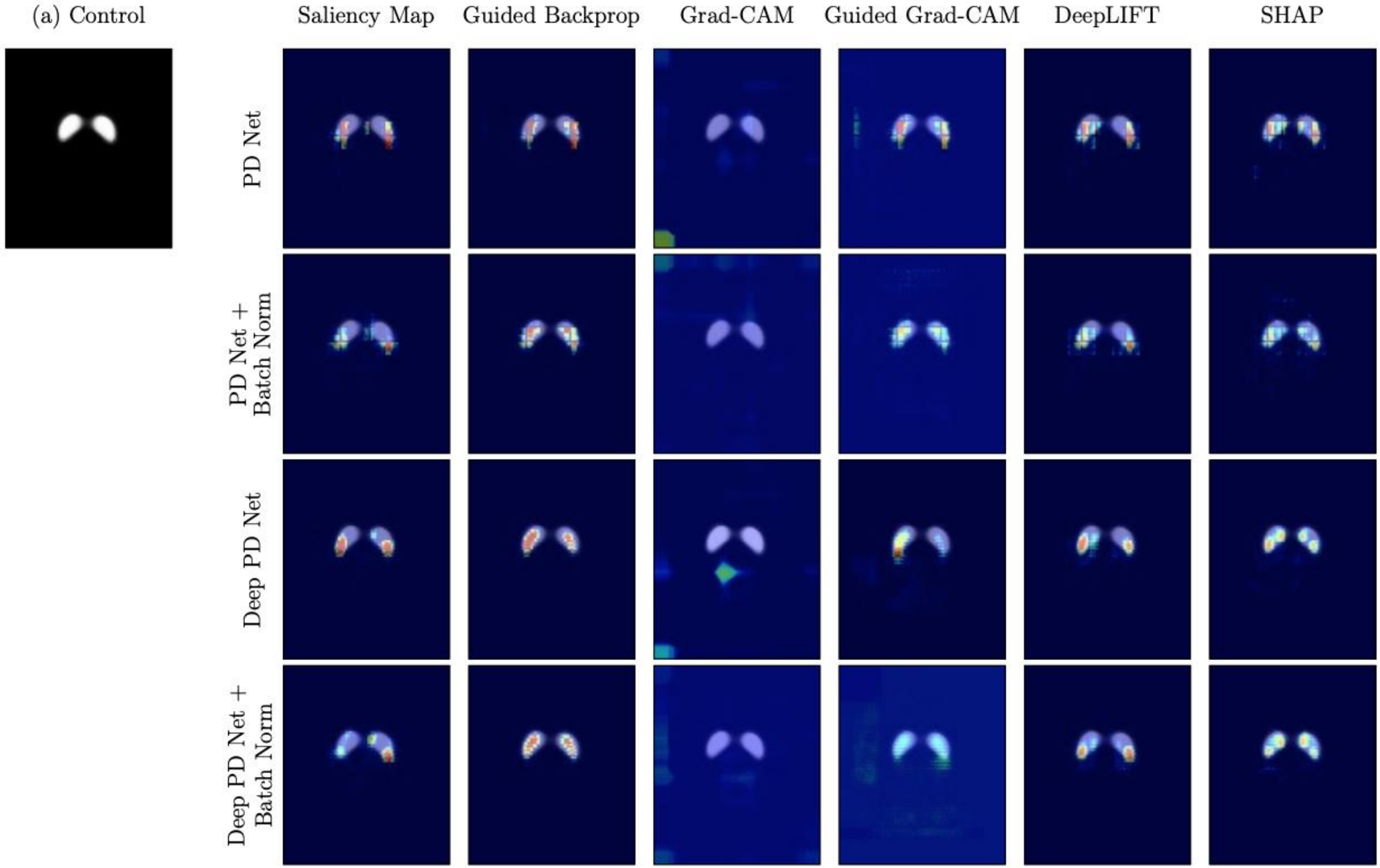
ated Gradients tf-explain alibi Skater	PH	L	Specific
Deep Visualization Toolbox	PH	L	Specific
DeepExplain iNNvestigate The LRP Toolbox Skater	PH	L	Specific
DeepExplain DeepLift iNNvestigate tf-explain Skater	PH	L	Specific
iNNvestigate	PH	L	Specific
iNNvestigate tf-explain	PH	L	Specific
tcav	PH	L	Specific
rationale	PH	L	Specific
Grad-CAM++	PH	L	Specific

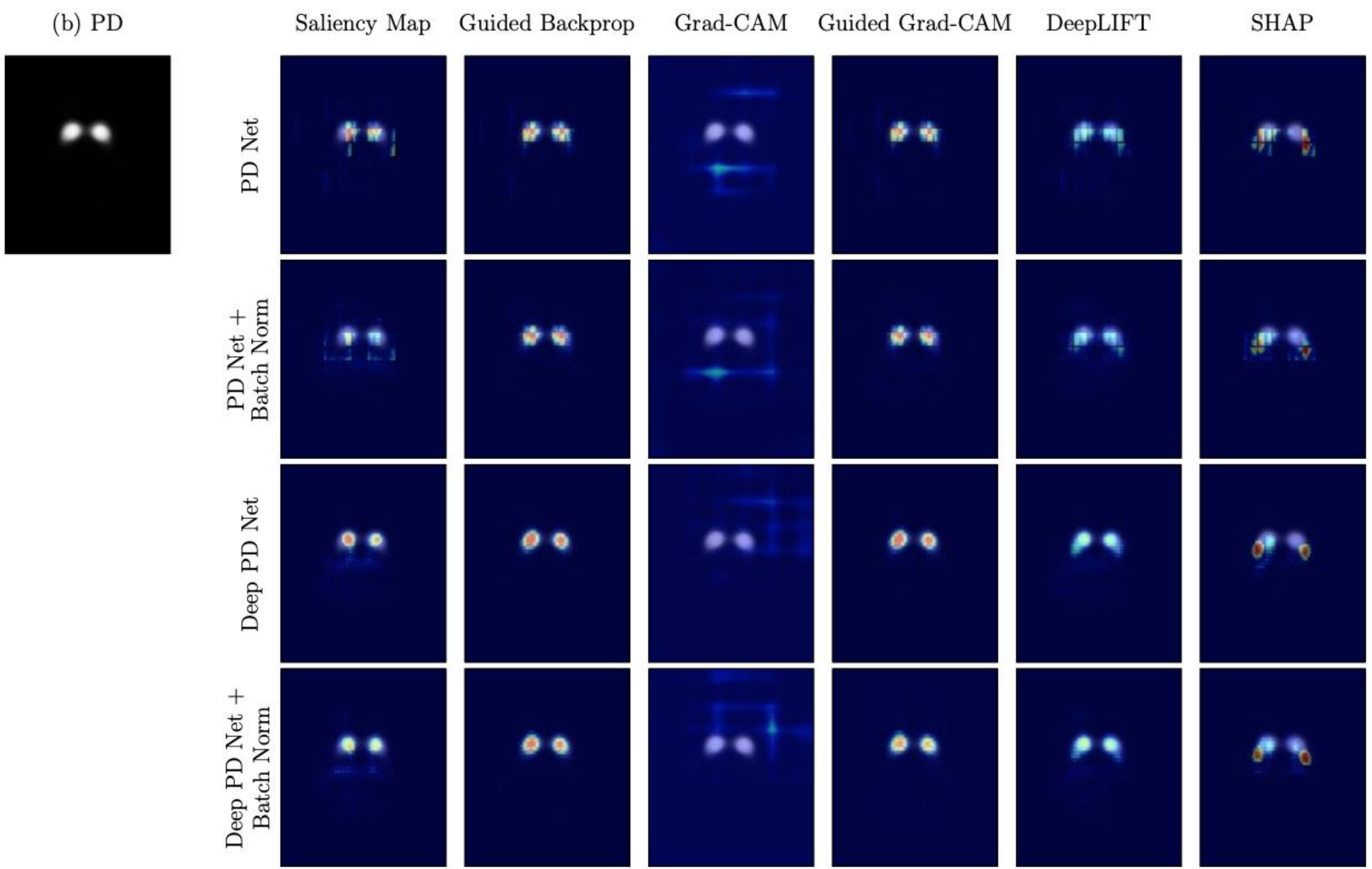

Input
Gradient
SmoothGrad
Deconvnet
Guided Backprop
PatternNet
PatternAttribution
DeepTaylor
Input * Gradient
Integrated Gradients
LRP-Z
LRP-Epsilon
LRP-PresetAFlat
LRP-PresetBFlat
label: tench
pred: tench
logit: 16.51
prob: 0.97
label: tiger shark
pred: tiger shark
logit: 29.01
prob: 0.98
label: electric ray
pred: electric ray
logit: 15.75
prob: 0.98
label: cock
pred: cock
logit: 18.23
prob: 0.64
label: indigo bunting
pred: indigo bunting
logit: 28.58
prob: 1.00
label: bald eagle
pred: bald eagle
logit: 26.83
prob: 1.00
label: vulture
pred: church
logit: 13.43
prob: 0.56

Interpretability Methods to Explain any Black-Box Mode

Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/ Year	Year
lime Eli5 InterpretML AIX360 Skater	PH	L	Agnostic	img txt tab	845.6	2016
PDPbox InterpretML Skater	PH	G	Agnostic	tab	589.2	2001
shap alibi AIX360 InterpretML	PH	L & G	Agnostic	img txt tab	504.5	2017
alibi Anchor	PH	L	Agnostic	img txt tab	158.3	2018
alibi	PH	L	Agnostic	tab img	124.5	2017

Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/ Year	Year
PyCEbox	PH	L & G	Agnostic	tab	53.3	2015
L2X	PH	L	Agnostic	img txt tab	50.3	2018
Eli5	PH	G	Agnostic	tab	41.5	2010
alibi AIX360	PH	L	Agnostic	tab img	34.3	2018
Alibi	PH	G	Agnostic	tab	23.2	2016
alibi	PH	L	Agnostic	tab img	17	2019
pyBreakDown	PH	L	Agnostic	tab	8.3	2018
pyBreakDown	PH	G	Agnostic	tab	8.3	2018
DLIME	PH	L	Agnostic	img txt tab	7.5	2019
AIX360	PH	L	Agnostic	tab	7	2019
AIX360	PH	L	Agnostic	tab img	3	2019

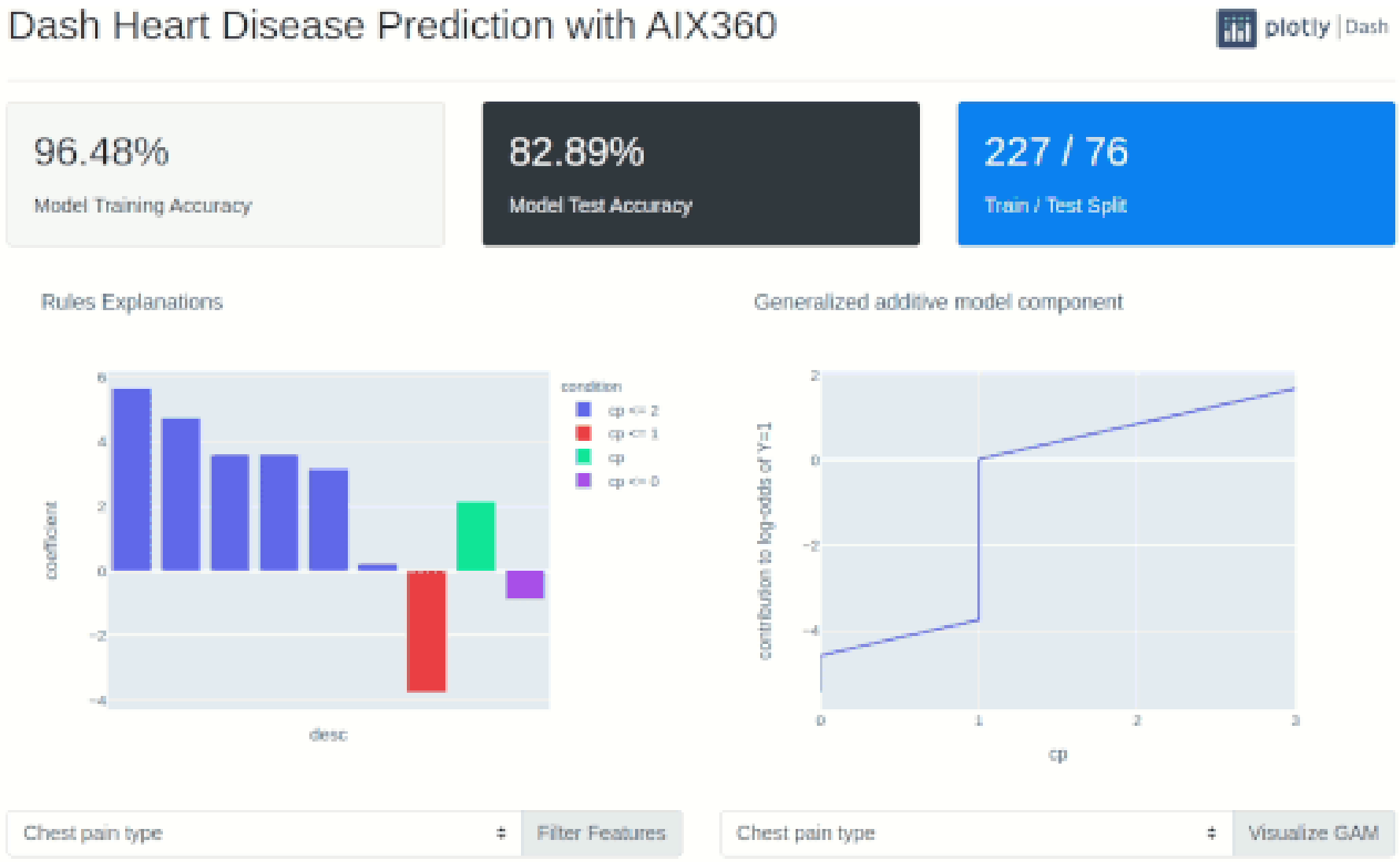

(a) Control
Saliency Map
Guided Backprop
Grad-CAM
Guided Grad-CAM
DeepLIFT
SHAP
PD Net
PD Net + Batch Norm
Deep PD Net
Deep PD Net + Batch Norm


(b) PD
Saliency Map
Guided Backprop
Grad-CAM
Guided Grad-CAM
DeepLIFT
SHAP
PD Net
PD Net + Batch Norm
Deep PD Net
Deep PD Net + Batch Norm

Interpretability Methods to Create White-Box Models.

Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/Year
InterpretML	W	G	Specific	tab	129.5
Slim	W	G	Specific	tab	35.2
AIX360	W	G	Specific	tab	12.3
AIX360	W	L	Specific	tab	12
AIX360	W	G	Specific	tab	5

AIX 360


Dash Heart Disease Prediction with AIX360
plotly | Dash
96.48%
Model Training Accuracy
82.89%
Model Test Accuracy
227 / 76
Train / Test Split
Rules Explanations
condition
cp <= 2
cp <= 1
cp
cp <= 0
coefficient
desc
Generalized additive model component
contribution to log-odds of Y=1
cp
Chest pain type
Filter Features
Chest pain type
Visualize GAM

InterpretML

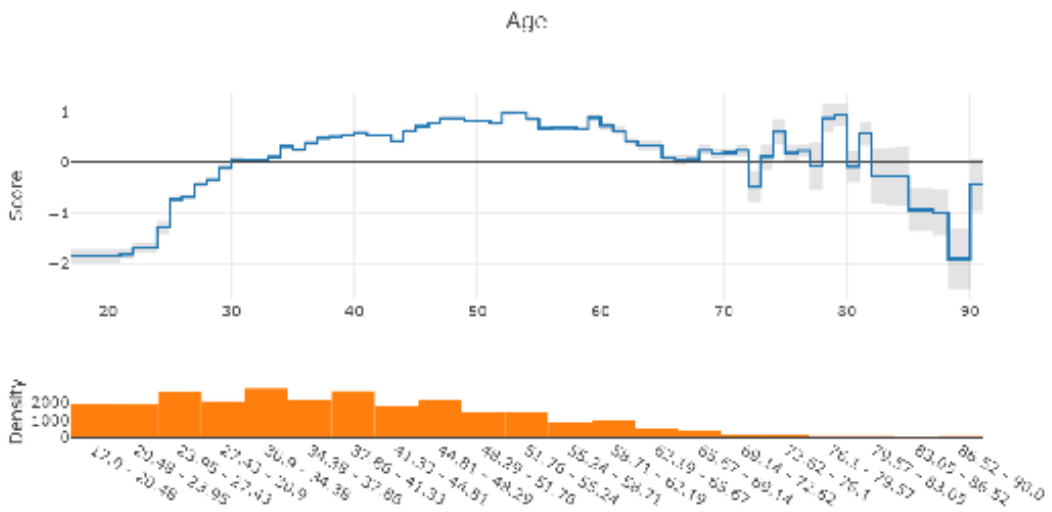
Select Component to Graph

0 : Name (Age) | Type (continuous)

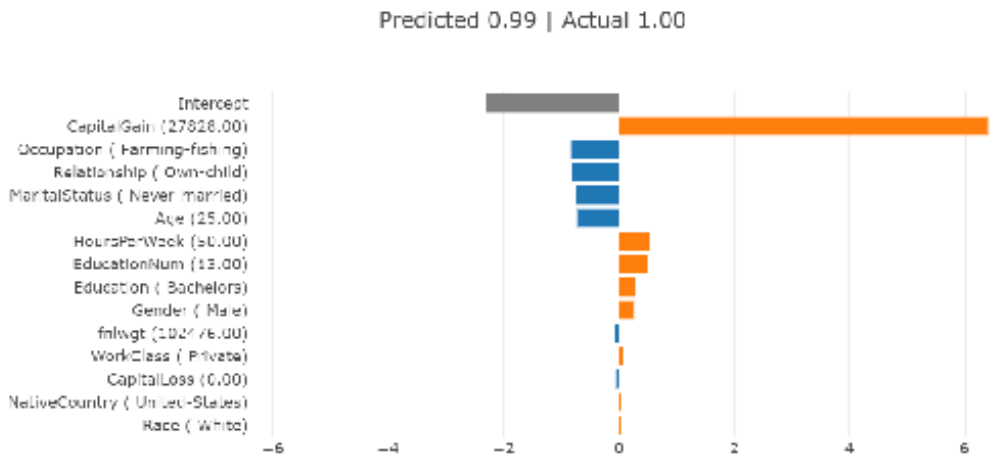
Select Component to Graph

2 : Predicted (0.994) | Actual (1.0)

EBM [0]



EBM [2]



Interpretability Methods to Restrict Discrimination and Enhance Fairness in Machine Learning Models.

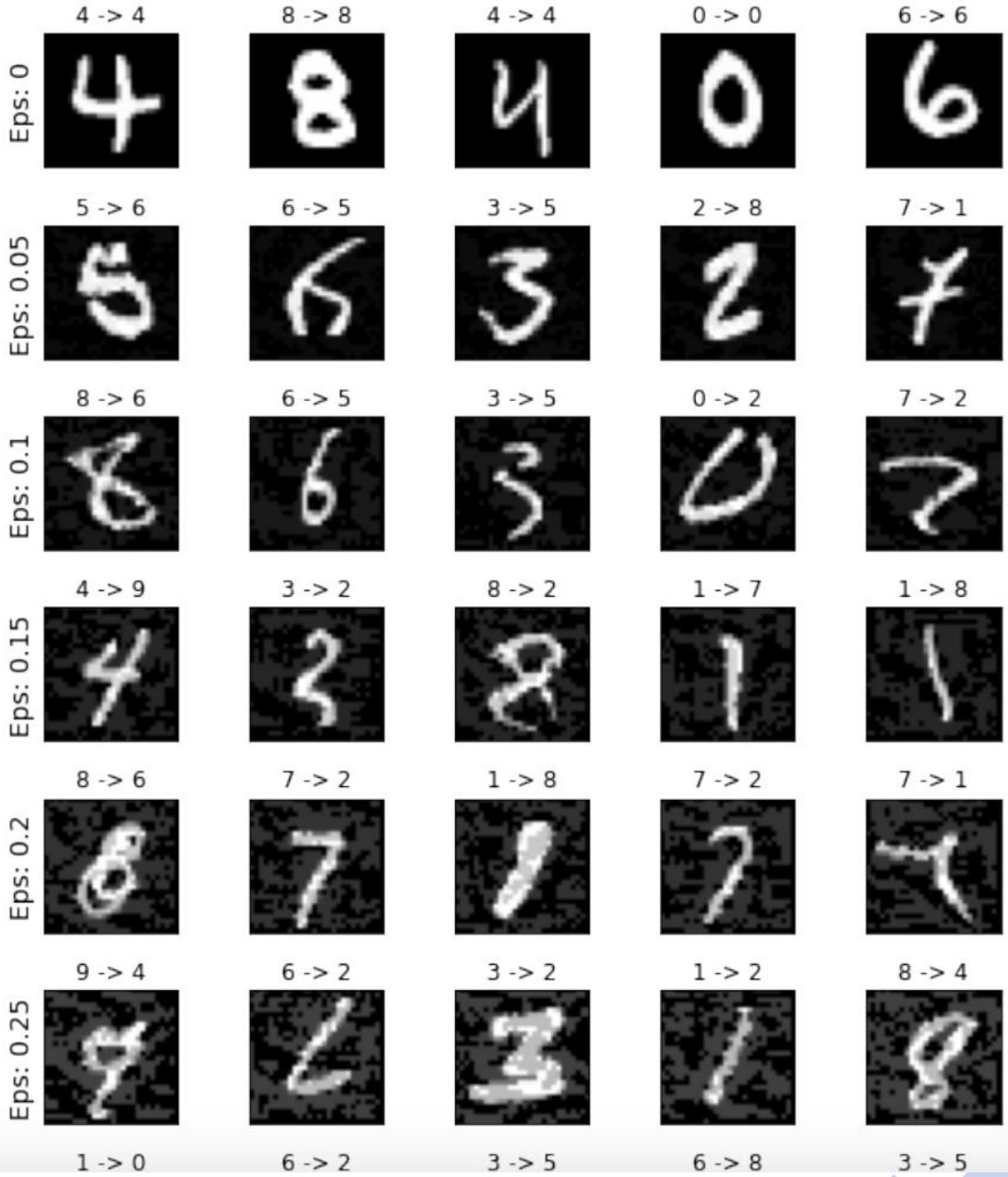
Table 4. Interpretability Methods to Restrict Discrimination and Enhance Fairness in Machine Learning Models.

Ref	Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/ Year	Year
[92]	equalized_odds_and_calibration fairlearn AIF360	F	G	Agnostic	tab	242.2	2016
[85]	debiaswe	F	L	Specific	txt	216.8	2016
[88]	fairness	F	L	Agnostic	tab	133.4	2012
[72]	Aequitas AIF360 themis-ml	F	G	Agnostic	tab	124.5	2015
[93]	fair-classification	F	G	Agnostic	tab	117.8	2017
[84]	fairness-in-ml	F	L	Agnostic	tab	115.5	2017
[94]	fair-classification	F	G	Agnostic	tab	110.8	2017
[86]	AIF360	F	L & G	Agnostic	tab	94.6	2013
[95]	fairlearn	F	G	Agnostic	tab	94	2018
[77]	AIF360	F	L & G	Agnostic	tab	92.3	2018
[96]	AIF360 GerryFair	F	G	Agnostic	tab	76	2018

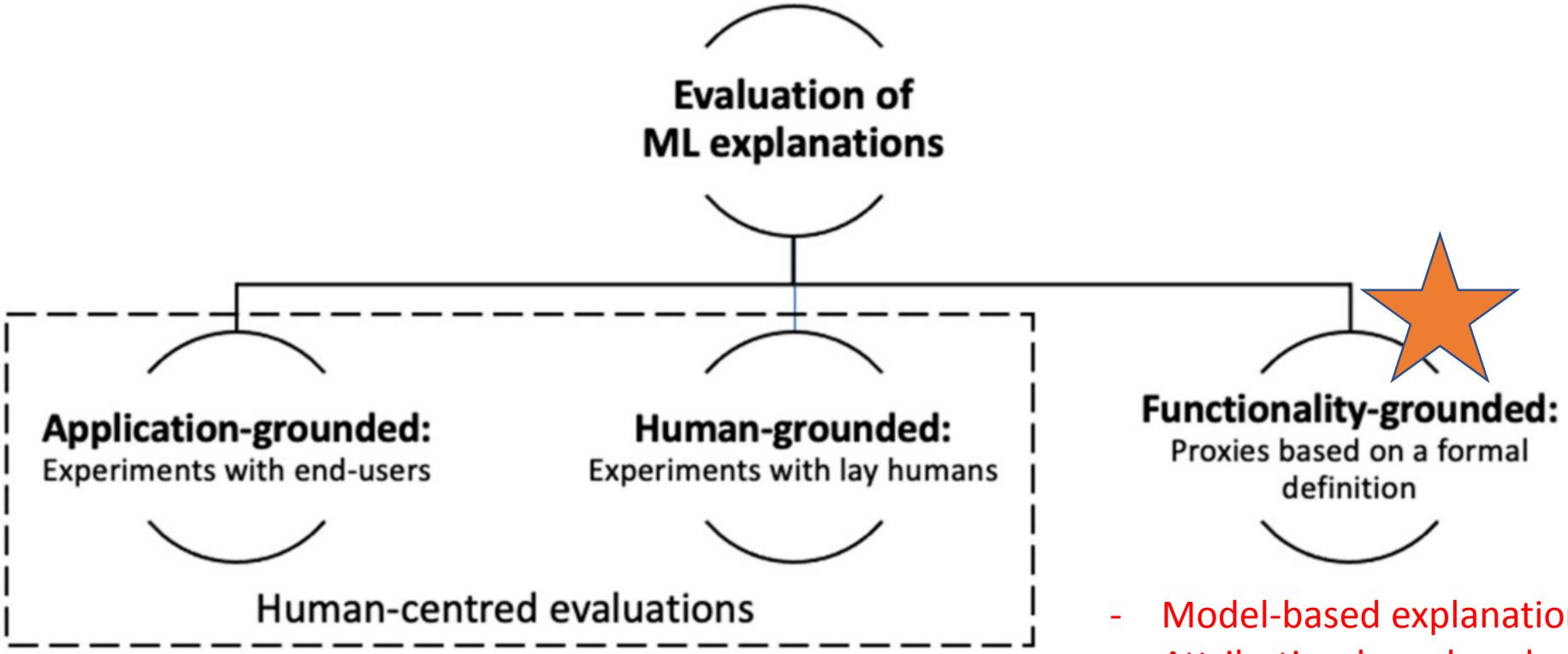
Adversarial Example-based Sensitivity Analysis

Table 6. Adversarial Example-based Sensitivity Analysis.

Ref	Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/ Year	Year
[116]	cleverhans foolbox	S	L & G	Agnostic	img	876.4	2014
[115]	cleverhans foolbox	S	L & G	Agnostic	img	727.4	2013
[123]	cleverhans nn_robust_attacks	S	L & G	Agnostic	img	716	2017
[120]	cleverhans foolbox	S	L & G	Agnostic	img	429	2016
[127]	one-pixel-attack-keras	S	L & G	Agnostic	img	409	2019
[117]	cleverhans foolbox	S	L & G	Agnostic	img	392	2016
[119]	cleverhans foolbox	S	L & G	Agnostic	img	381.2	2016
[134]	cleverhans	S	L & G	Agnostic	img	378.8	2017
[137]	influence-release	S	L & G	Agnostic	img	224	2017
[121]	cleverhans	S	L & G	Agnostic	img	181.7	2018
[152]	adversarial-squad	S	L & G	Specific	txt	162	2017
[131]	transferability-advdnn-pub	S	L & G	Agnostic	img	148.6	2016


Eps: 0
4 -> 4
8 -> 8
4 -> 4
0 -> 0
6 -> 6
Eps: 0.05
5 -> 6
6 -> 5
3 -> 5
2 -> 8
7 -> 1
Eps: 0.1
8 -> 6
6 -> 5
3 -> 5
0 -> 2
7 -> 2
Eps: 0.15
4 -> 9
3 -> 2
8 -> 2
1 -> 7
1 -> 8
Eps: 0.2
8 -> 6
7 -> 2
1 -> 8
7 -> 2
7 -> 1
Eps: 0.25
9 -> 4
6 -> 2
3 -> 2
1 -> 2
8 -> 4
1 -> 0
6 -> 2
3 -> 5
6 -> 8
3 -> 5

Taxonomy of evaluation of machine learning interpretability



- Model-based explanations
- Attribution-based explanations.
- Example-based explanations

Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. arXiv 2017, arXiv:1702.08608.

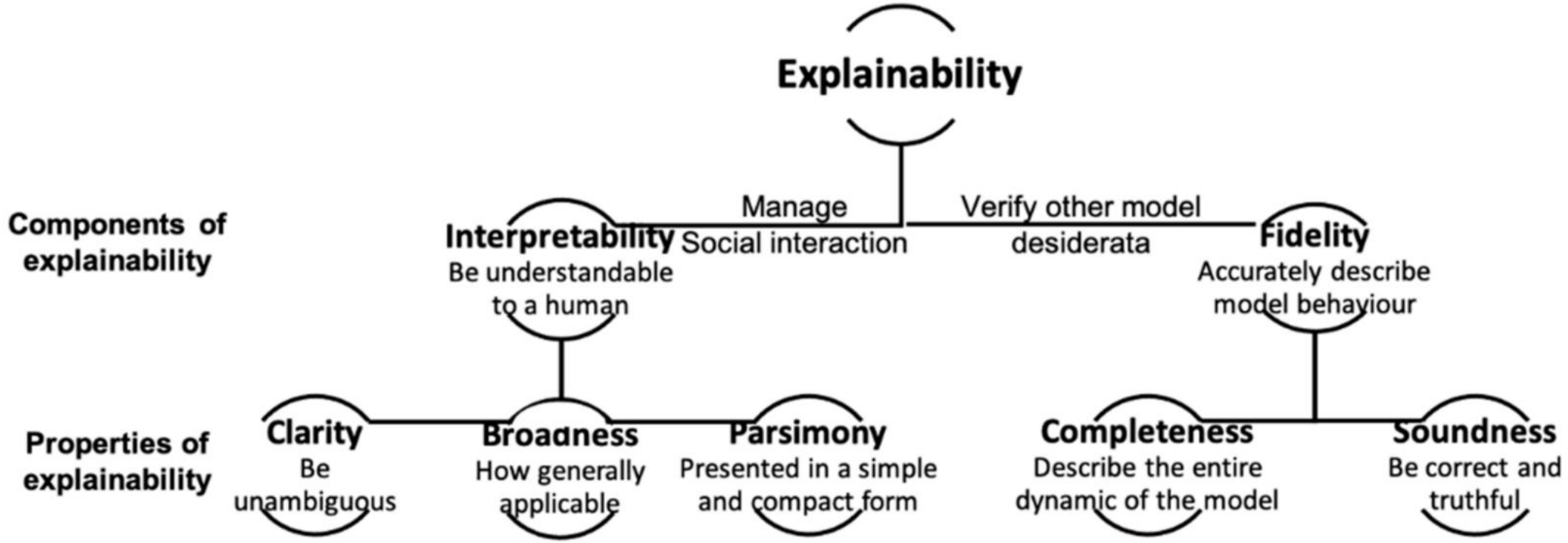
Quantitative metrics for machine learning (ML) explanations

Explanation Types	Quantitative Metrics	Properties of Explainability				
		Interpretability			Fidelity	
		Clarity	Broadness	Parsimony/ Simplicity	Completeness	Soundness
Model-based explanations	Model size [14,46]			✓		
	Runtime operation counts [81]			✓		
	Interaction strength [23,46]			✓		
	Main effect complexity [46]			✓		
	Level of (dis)agreement [82]	✓				✓

Quantitative metrics for machine learning (ML) explanations

Attribution- based explanations	Monotonicity [80]					✓
	Non-sensitivity [80,83], Sensitivity [84,85]					✓
	Effective complexity [80]		✓	✓		
	Remove and retrain [86]					✓
	Recall of important features [33]					✓
	Implementation invariance [85]					✓
	Selectivity [87]					✓
	Continuity [87]	✓				
	Sensitivity-n [88]					✓
	Mutual information [80]		✓	✓		✓
Example-based explanations	Non-representativeness [80]			✓	✓	
	Diversity [80]			✓		

Definition of machine learning (ML) explainability and related properties



Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. arXiv 2020, arXiv:2007.15911.

Contribution

- Selection of interpretability method

- Choose the method to evaluate the ML explanation based on the explanation types and property