hessian.AI

Philipps Universität Marburg

Can we trust XAI? Current status and challenges of evaluating XAI methods

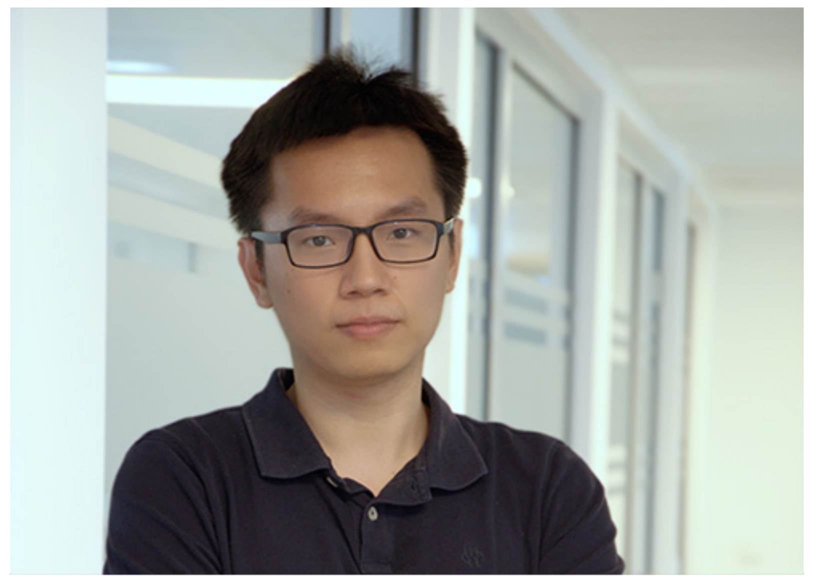
Christin Seifert

University of Marburg, Hessian.AI


Meike Nauta


Jörg Schlötterer


Shreyasi Pathak


Van Bach Nguyen


Jan Trienes


Le Phuong Qyiun

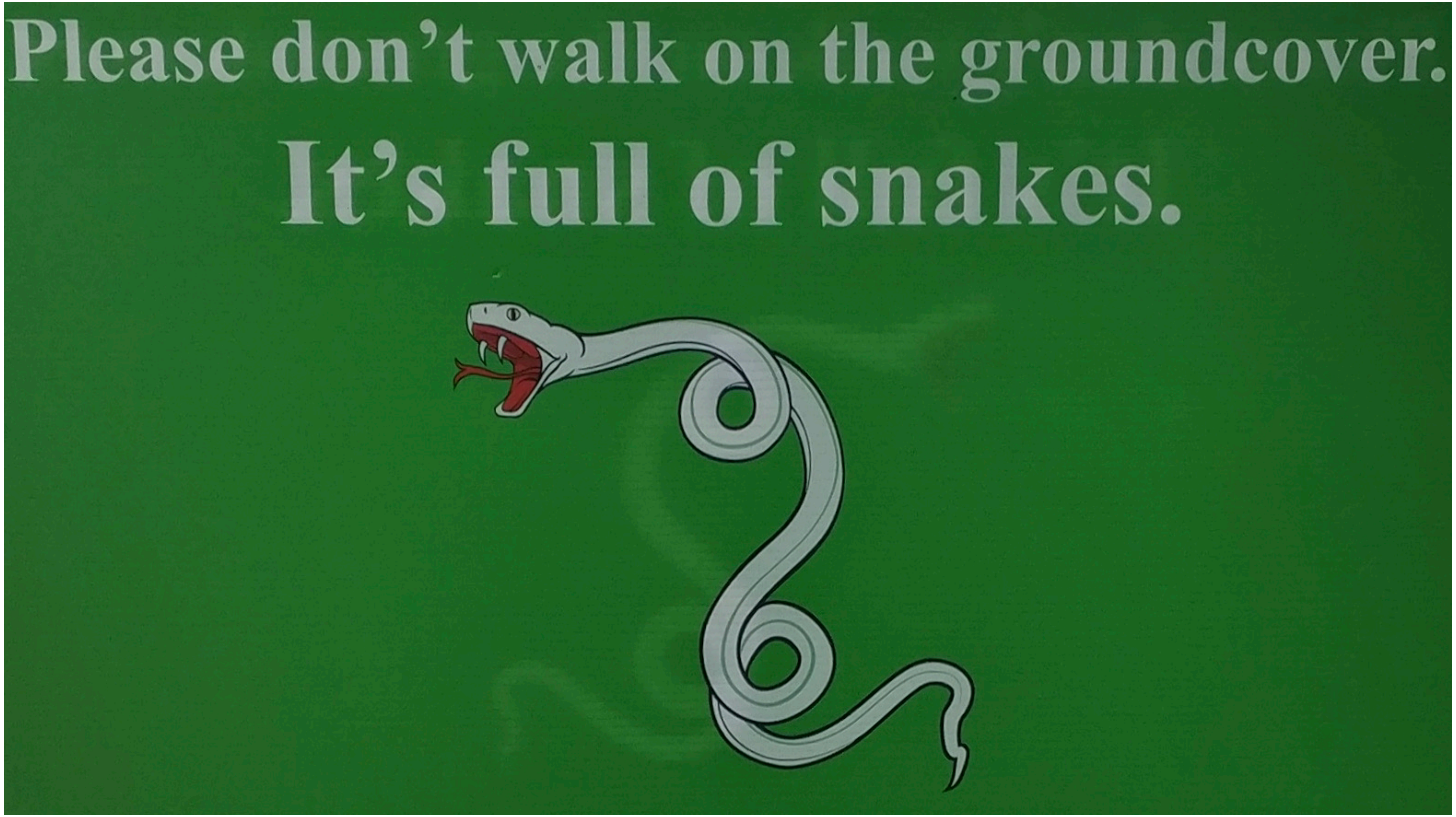
XAI@ECAI, 1.10.2023

X (AI)

Sign on the well trimmed lawn at the University of Dallas, Texas.
Walking across the lawn would be the shortest distance to coffee.
It’s 35° centigrade.

Q: Would you walk on the lawn?

X (AI)



Q: Would you walk on the lawn?

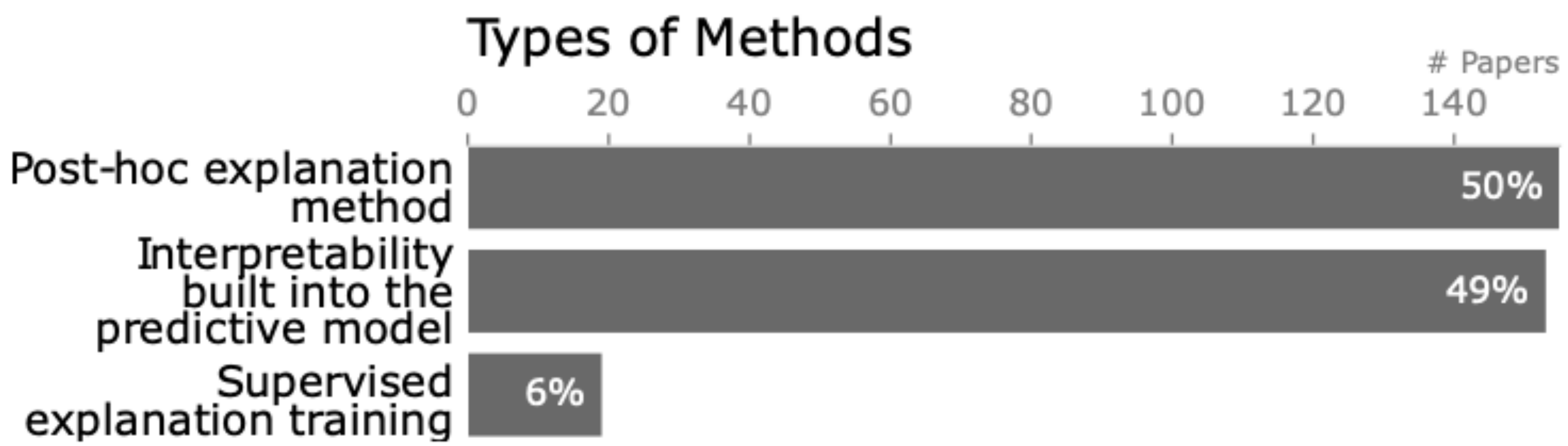
X (AI)

Q: Do you trust the explanation?

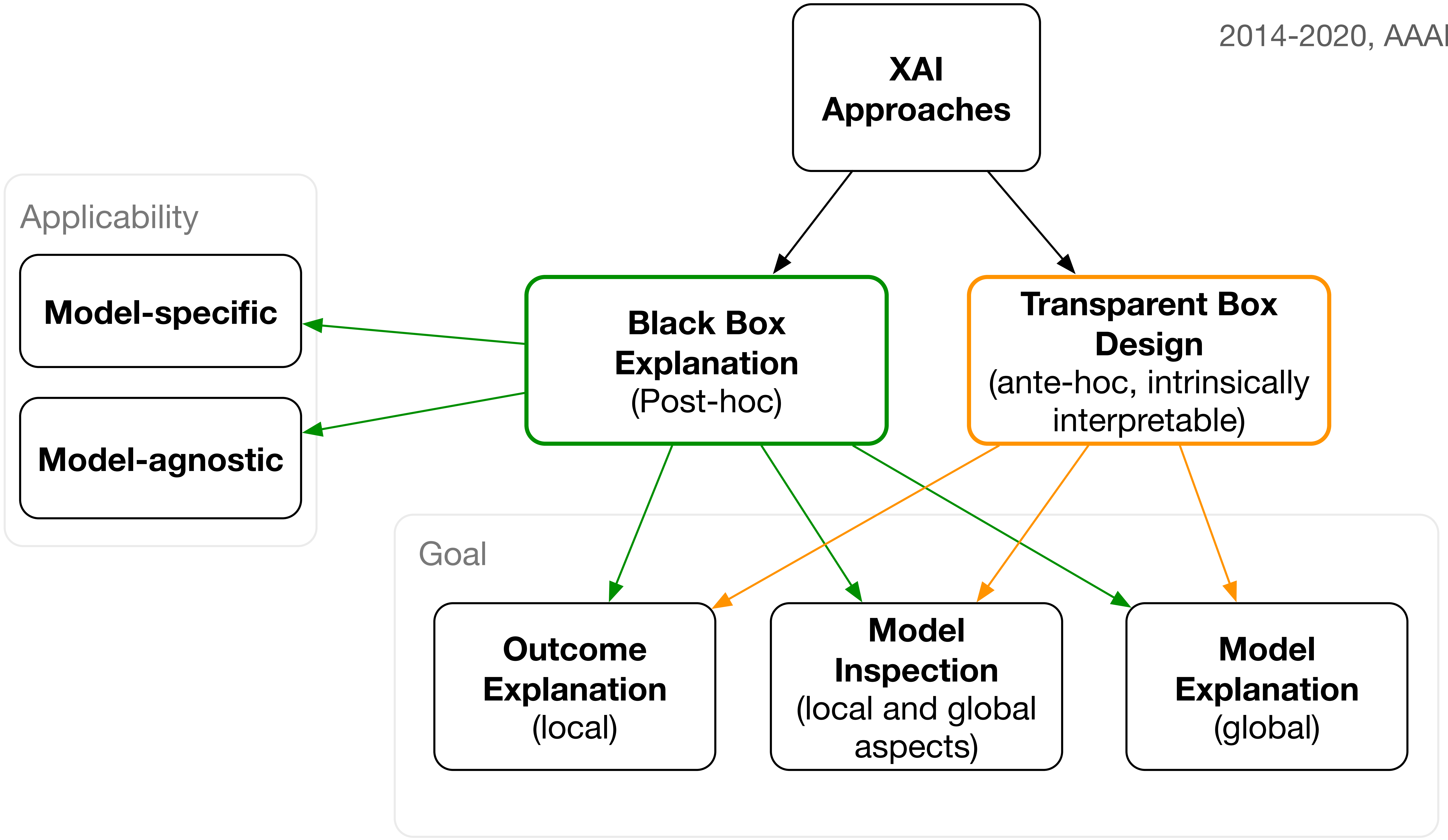
An **explanation in AI** is a presentation of (aspects of) the reasoning, functioning and/or behaviour of a machine learning model in **human-understandable terms.**

Nauta et. al, 2023

XAI Taxonomy



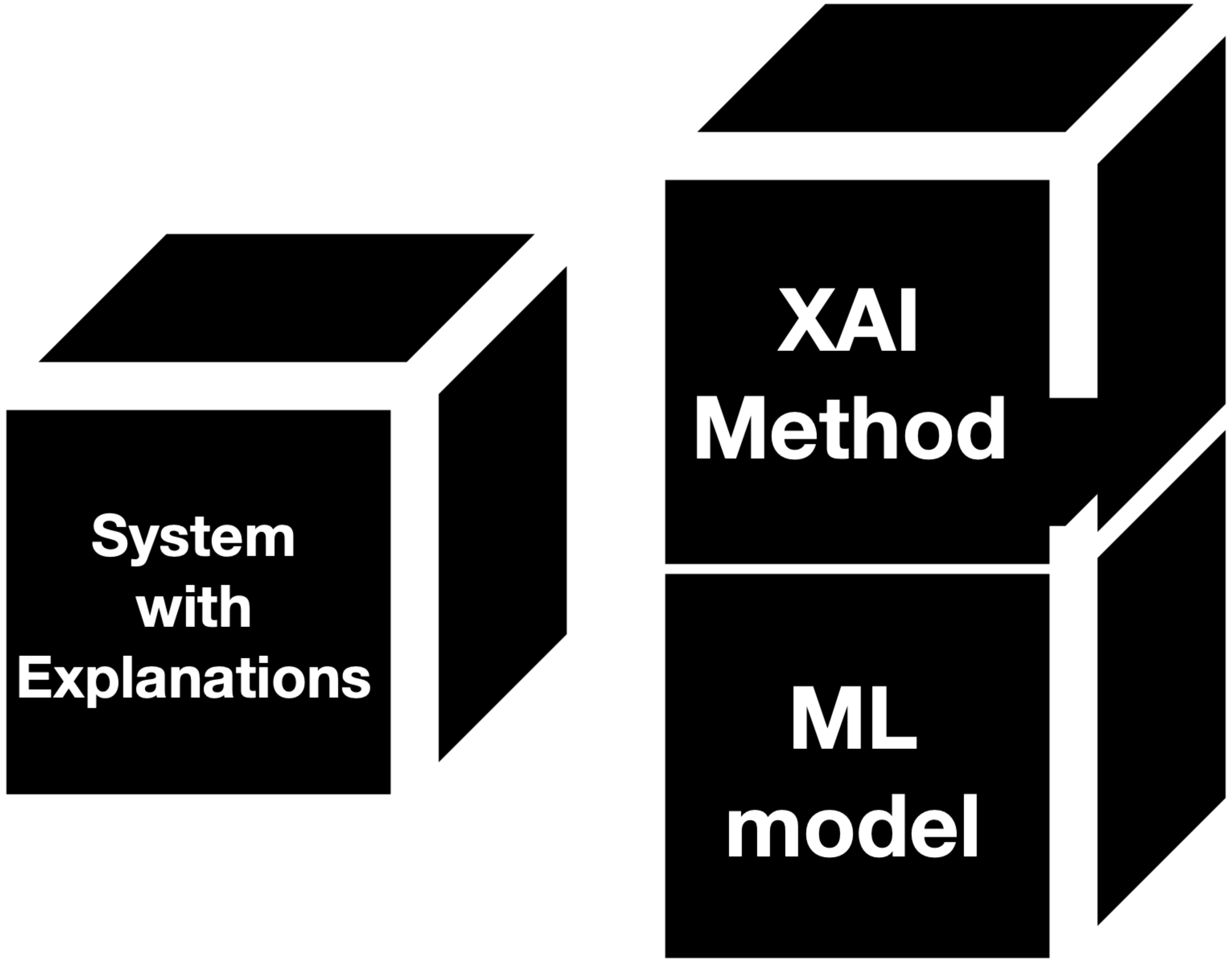
2014-2020, AAAI, IJCAI, NeurIPS, ICML, ICLR, CVPR, ICCV, ACL, WWW, ICDM, SIGKDD, SIGIR



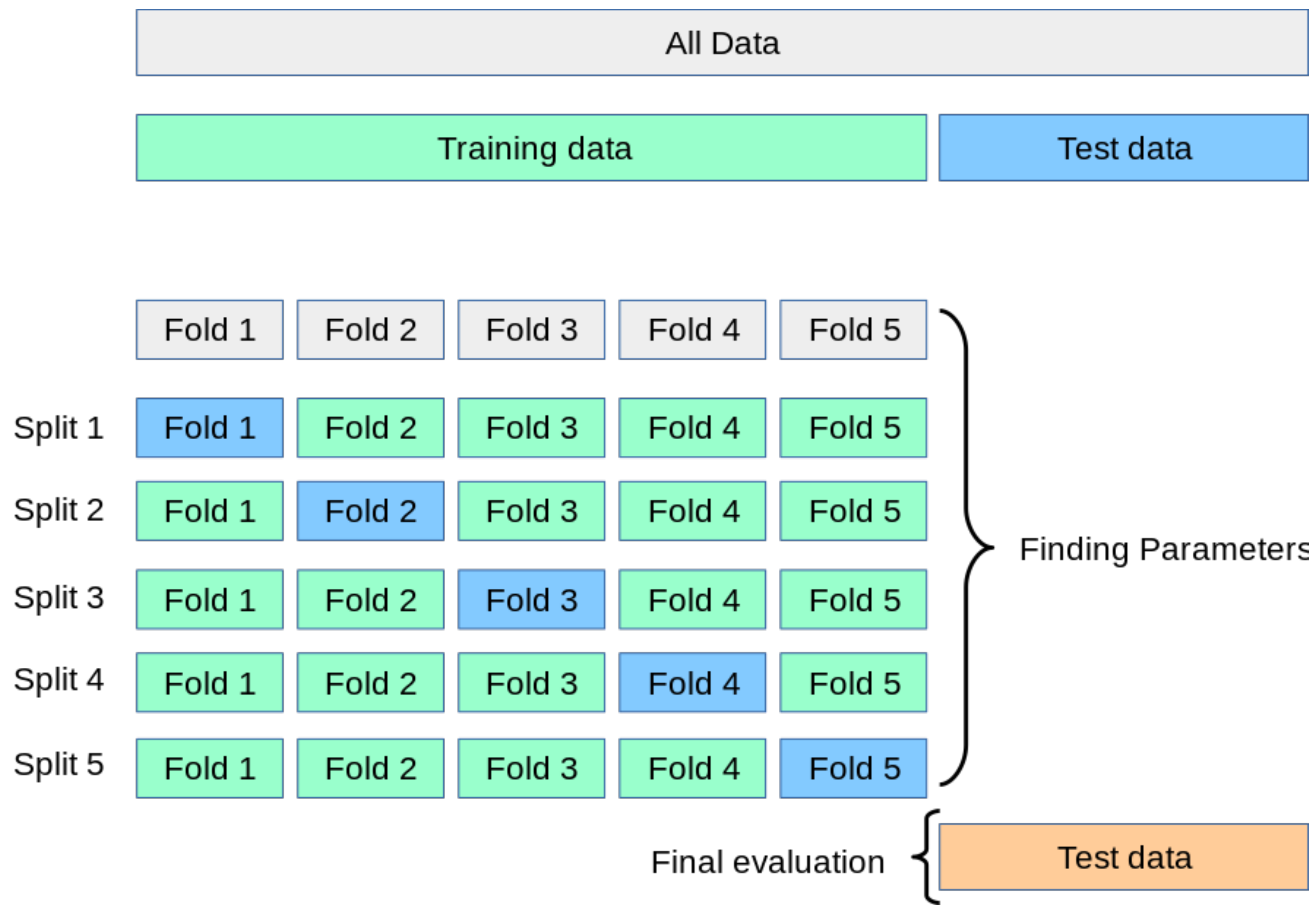
R. Guidotti, A. Monreale, S. Ruggieri,F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018, doi: [10.1145/3236009](https://doi.org/10.1145/3236009).
Meike Nauta, Jörg Schlötterer, Maurice van Keulen and Christin Seifert (2023). PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Evaluation of a System with Explanations

- If the output is bad, should we attribute this to the ML model or the XAI method?
- Notation: $f(x)$ predictive model, $g(f(x))$ explanation method
- We need to be able to measure the quality of an explanation method



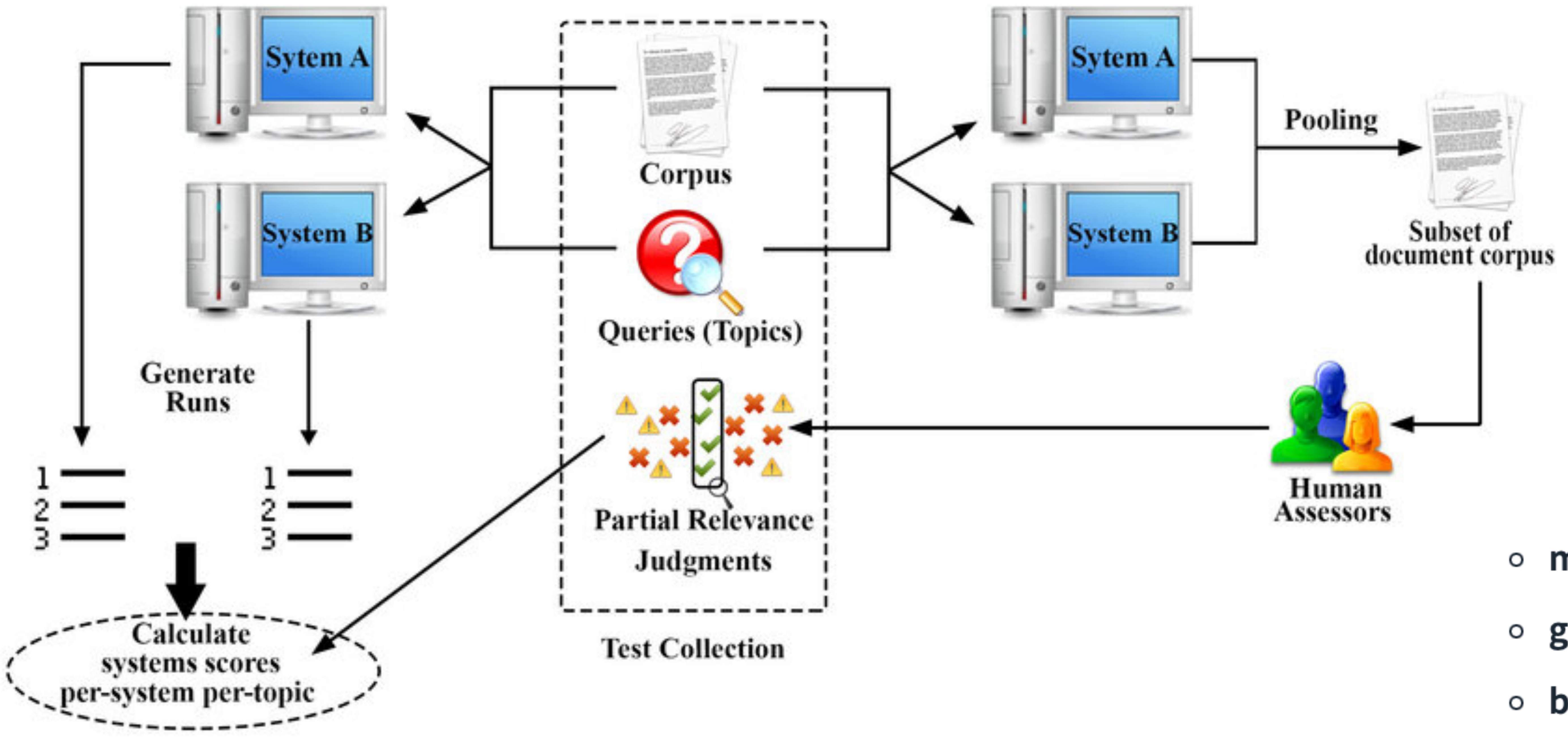
Evaluating ML Models



Scoring	Function
Classification	
'accuracy'	metrics.accuracy_score
'balanced_accuracy'	metrics.balanced_accuracy_score
'top_k_accuracy'	metrics.top_k_accuracy_score
'average_precision'	metrics.average_precision_score
'neg_brier_score'	metrics.brier_score_loss
'f1'	metrics.f1_score
'f1_micro'	metrics.f1_score
'f1_macro'	metrics.f1_score
'f1_weighted'	metrics.f1_score
'f1_samples'	metrics.f1_score
'neg_log_loss'	metrics.log_loss
'precision' etc.	metrics.precision_score
'recall' etc.	metrics.recall_score
'jaccard' etc.	metrics.jaccard_score
'roc_auc'	metrics.roc_auc_score
'roc_auc_ovr'	metrics.roc_auc_score
'roc_auc_ovo'	metrics.roc_auc_score
'roc_auc_ovr_weighted'	metrics.roc_auc_score
'roc_auc_ovo_weighted'	metrics.roc_auc_score

Image courtesy: sklearn documentation

Evaluating Information Retrieval (IR) Methods



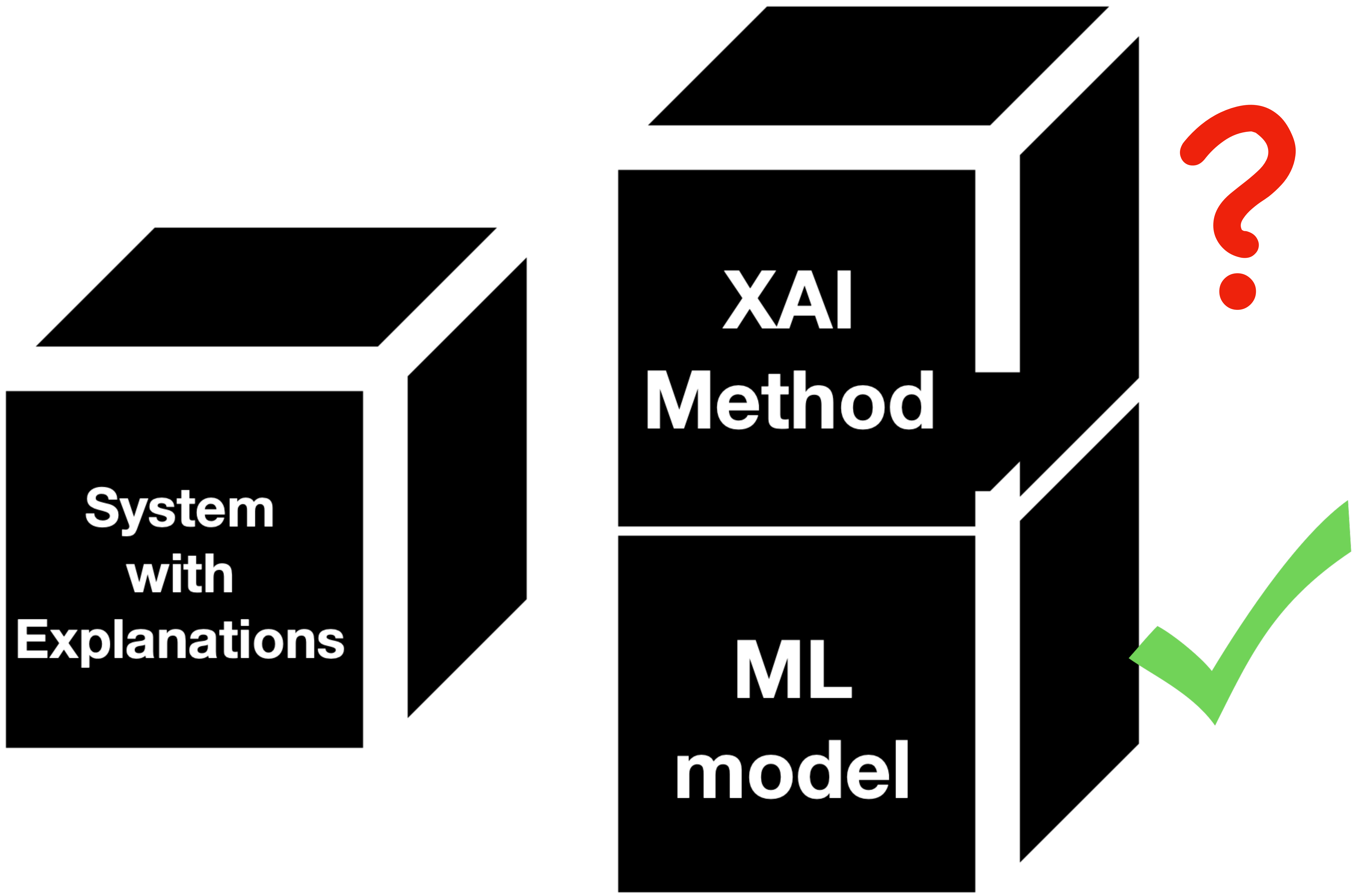
TREC Eval Paradigm

🤗 Spaces 🤗 evaluate-metric/**trec_eval**

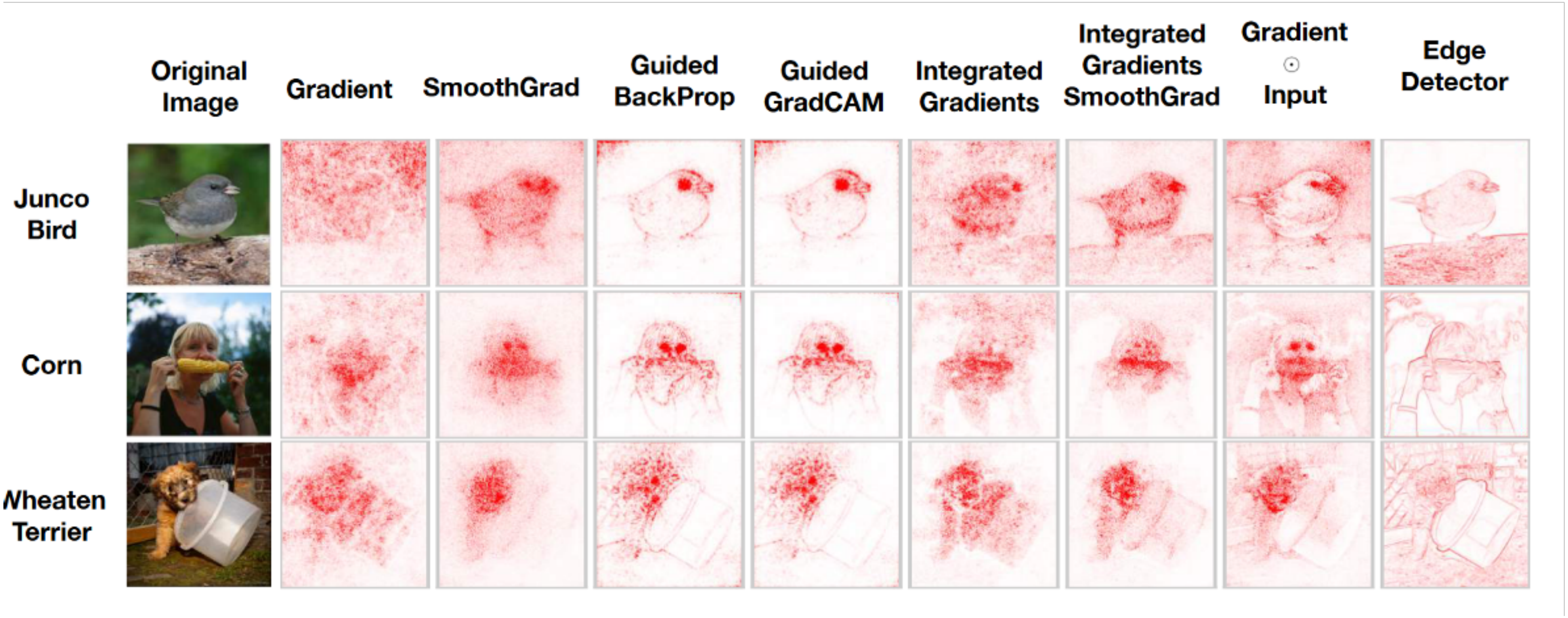
- **map** *(float): Mean average precision.*
- **gm_map** *(float): geometric mean average precision.*
- **bpref** *(float): binary preference score.*
- **Rprec** *(float): precision@R, where R is number of relevant documents.*
- **recip_rank** *(float): reciprocal rank*
- **P@k** *(float): precision@k (k in [5, 10, 15, 20, 30, 100, 200, 500, 1000]).*
- **NDCG@k** *(float): nDCG@k (k in [5, 10, 15, 20, 30, 100, 200, 500, 1000]).*

Image courtesy: Ravana, Sri Devi & Taheri, Masumeh & Rajagopal, Prabha. (2015). Document-based approach to improve the accuracy of pairwise comparison in evaluating information retrieval systems. Aslib Journal of Information Management. 67. 408-421 10.1108/AJIM-12-2014-0171.

Evaluation of a System with Explanations

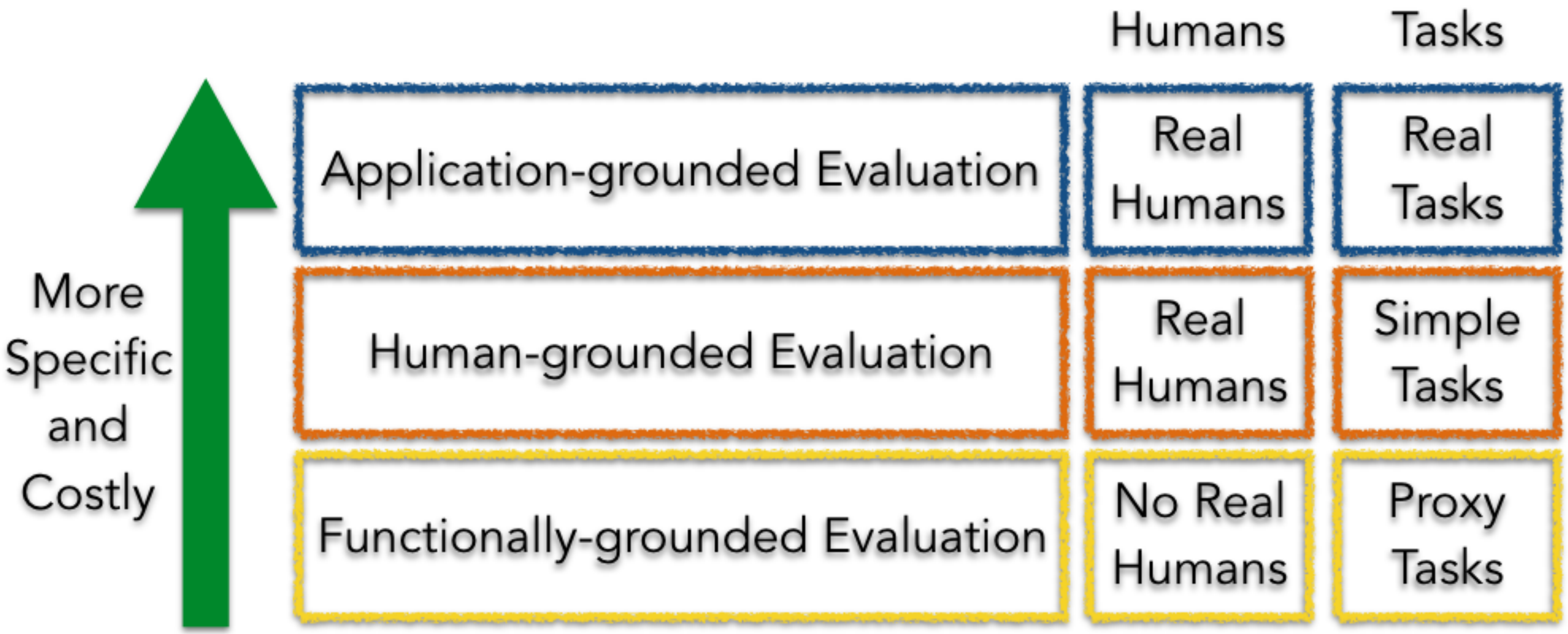


How to evaluate / compare?



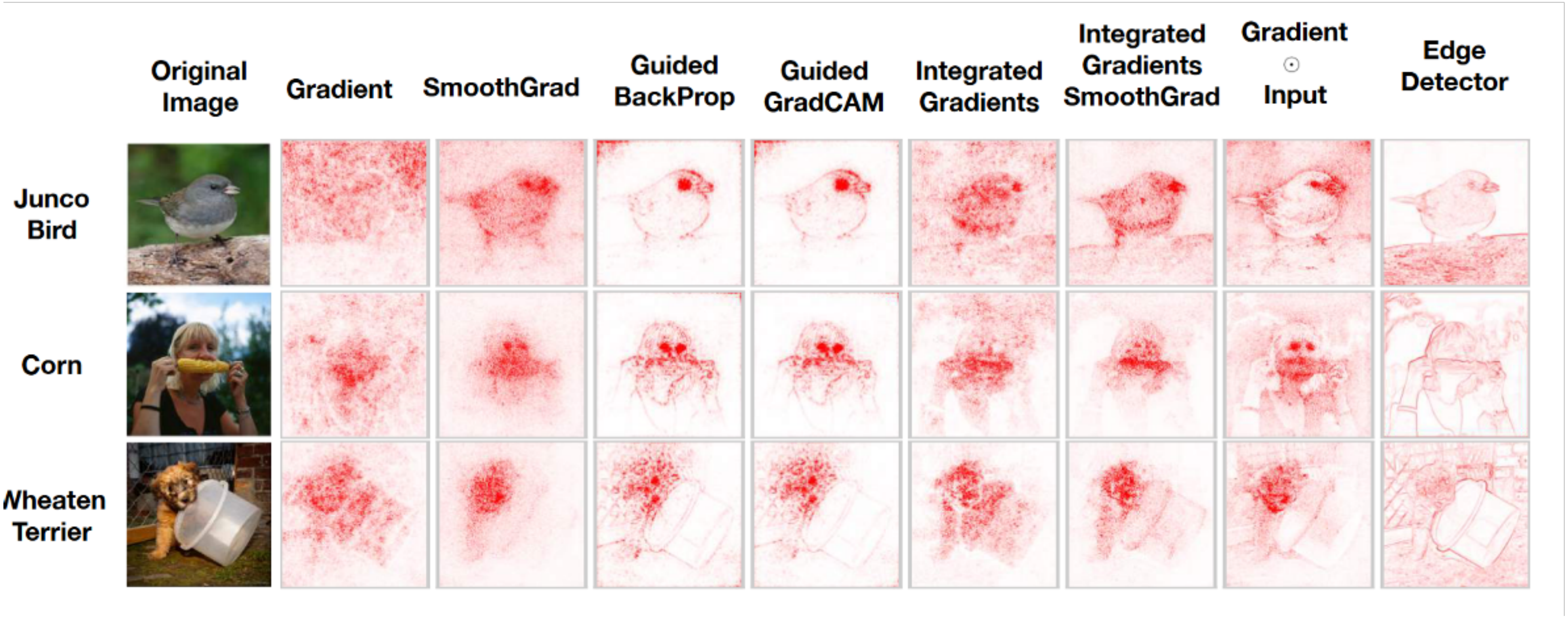
Adebayo, Julius, et al. "Sanity checks for saliency maps." Advances in neural information processing systems 31 (2018).

Evaluating XAI



F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning.” arXiv, Mar. 02, 2017. doi: 10.48550/arXiv.1702.08608.

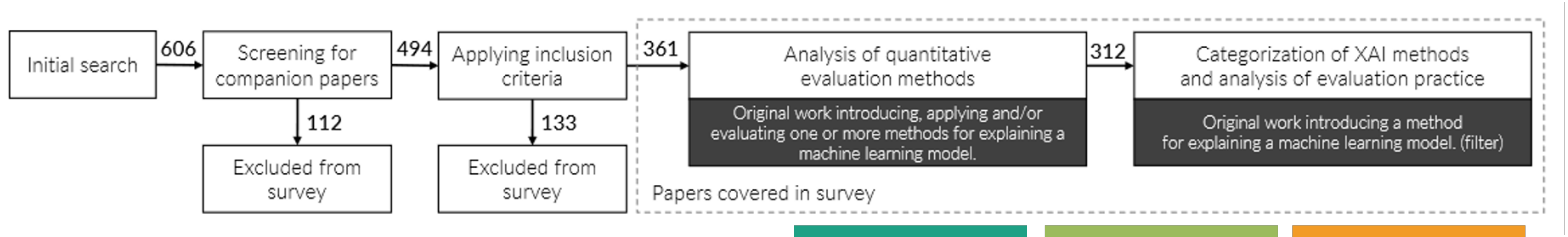
How to evaluate / compare?

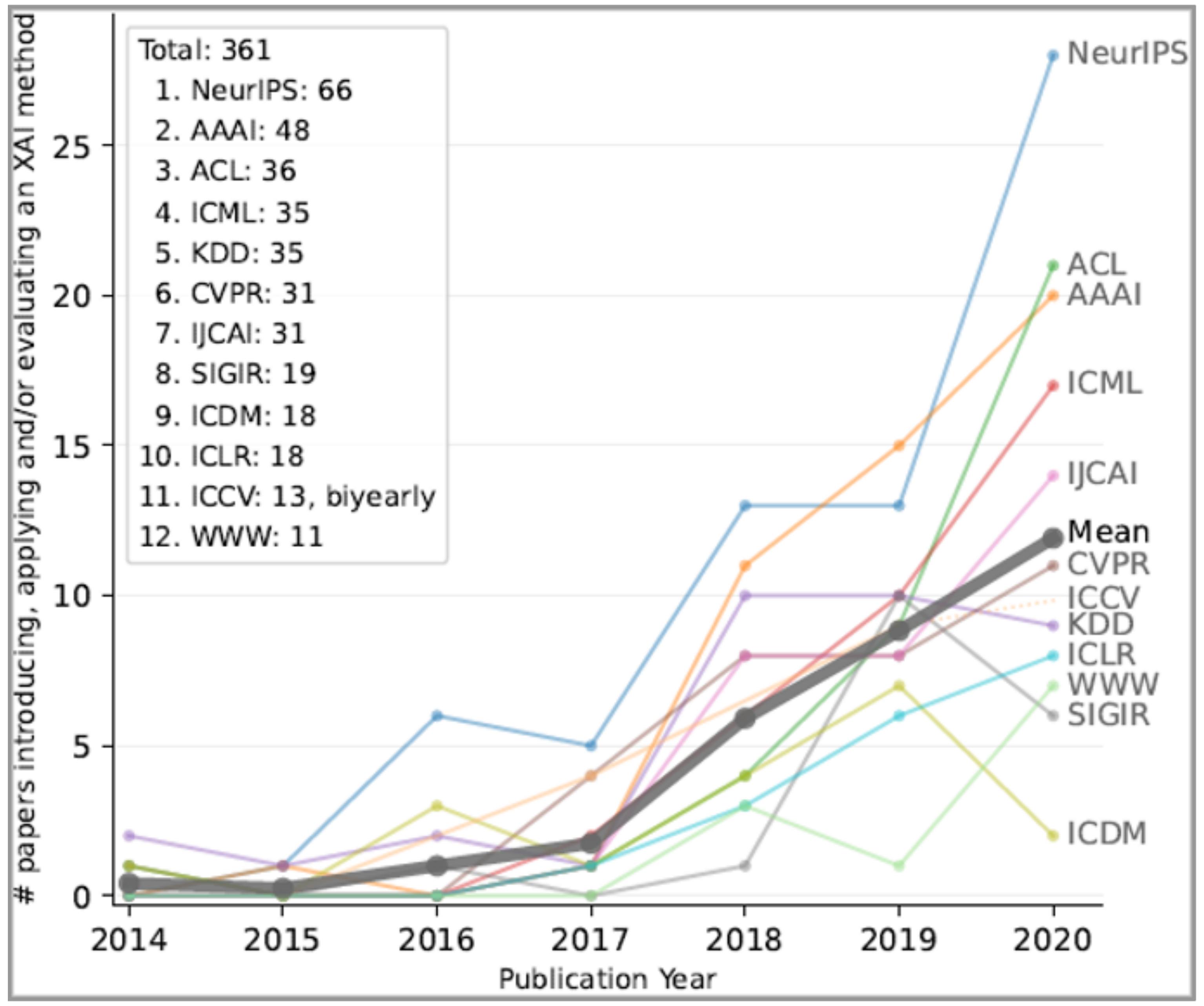


We didn't know…

Adebayo, Julius, et al. "Sanity checks for saliency maps." Advances in neural information processing systems 31 (2018).

Evaluating XAI





1 in 3

Evaluate by showing examples only (anecdotal evidence)

1 in 5

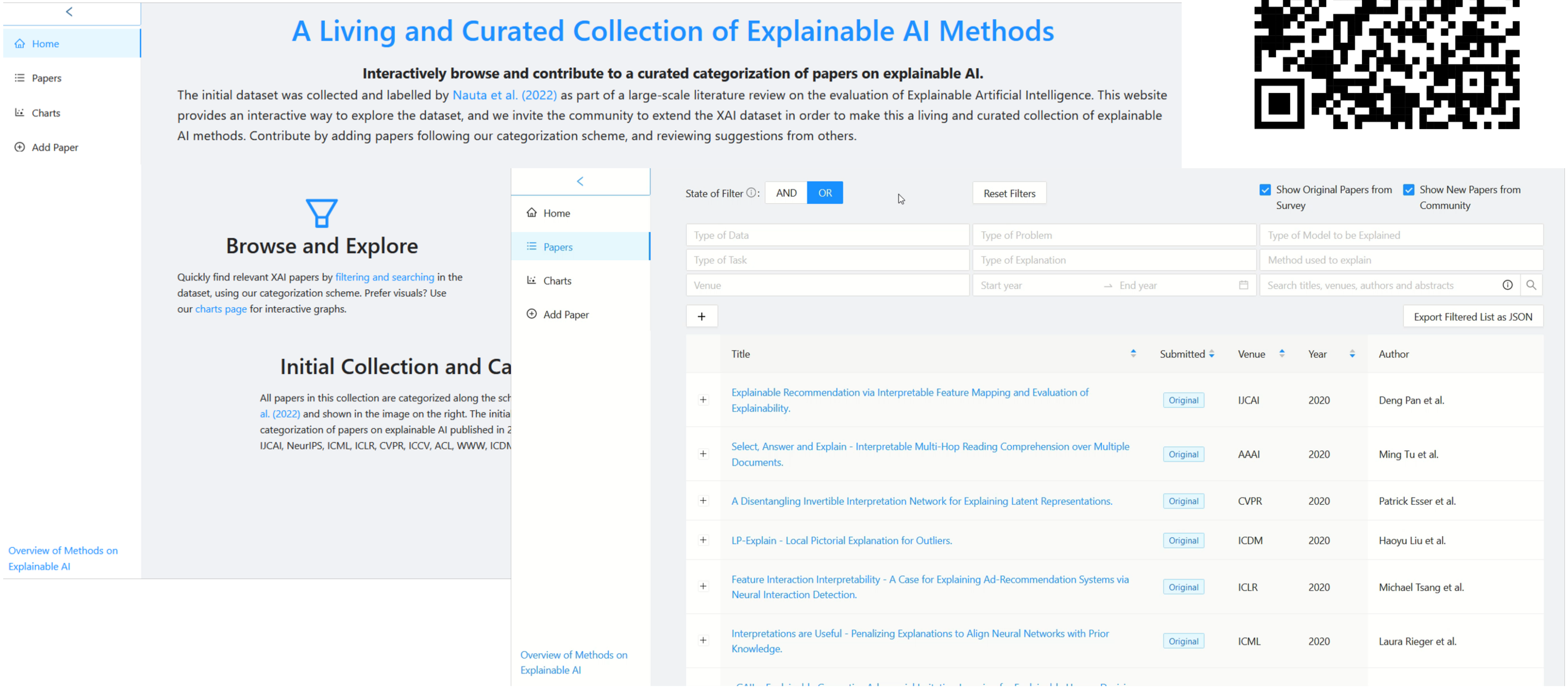
Evaluate with human subjects in a user study

58%

Evaluate with *quantitative* evaluation methods

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen and Christin SeifertFrom Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI ACM Comput. Surv., Associaton for Computing Machinery, 2023

Evaluating XAI



https://utwente-dmb.github.io/xai-papers

Co-12 Properties

Correctness
Match between model and explanation.

Completeness
How much of the model is explained?

Consistency
Robustness to small changes in model and implementation.

$g(x) = g(x)$

Continuity
Robustness to small changes input.

$g(x) = g(x')$

Contrastivity
Discriminative to other events or targets?

$g(x|Cat) \,!\!= g(x|Dog)$

Covariate Complexity
Complexity of features in the explanation

Compactness
Size of the explanation

Composition
Presentation format

Confidence
Probability information available?

$p = ?$

Context
Useful for users?

Coherence
Match with domain knowledge.

$g(x) =$

Controllability
Can user influence explanation?

$g(x)$

Explanation / Model / User

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen and Christin SeifertFrom Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI ACM Comput. Surv., Associaton for Computing Machinery, 2023

Co-12 Correctness

How faithful the explanation is w.r.t. the black box. “Nothing but the truth.“

- **Randomization Check:** Randomly perturb the predictive model → Explanation should change.
- **Whitebox Check:** Apply the explanation method to an interpretable whitebox. → Explanation should match the whitebox’ reasoning.
- **Single Deletion:** Delete or perturb single features. → Observe model output and measure correlation with explanation’s importance score.
- **Incremental Deletion / Addition:** Delete or add features in order of importance. → see Single Deletion, and can compare with addition/deletion in random order as baselines.

Co-12 Completeness

How much of the black box behaviour is described by the explanation?
“The whole truth”.

- **Preservation Check** Calculate the model’s output for the explanation (instead of the full datapoint). → Model output should be the same.

- **Deletion Check** Calculate output on the datapoint with relevant features removed. → Model output should be different.

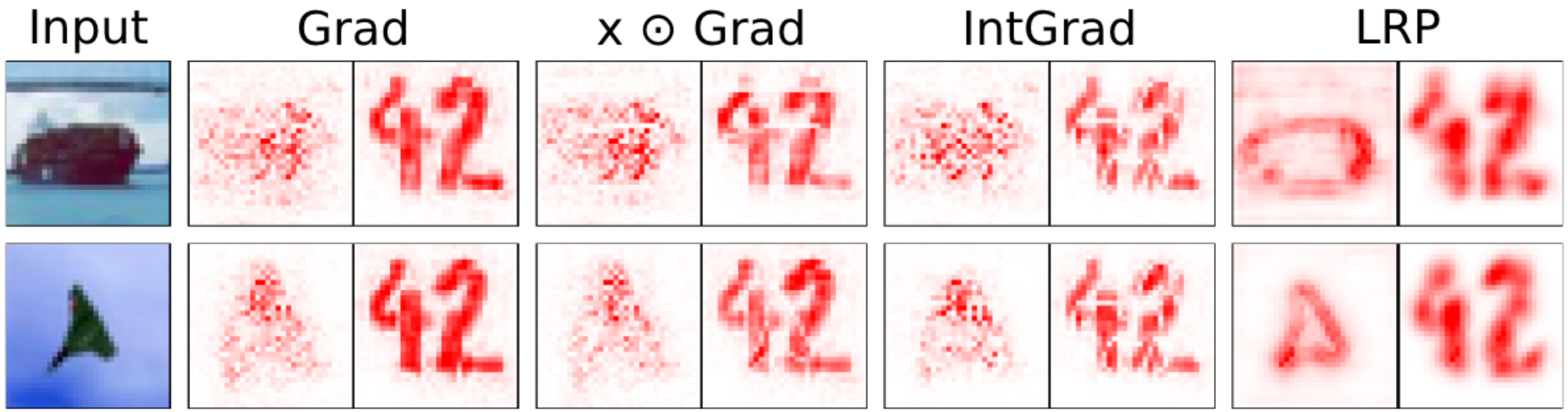
- **Fidelity** (for explanation methods that are themselves predictive models) Calculate agreement between the model output and explanation output for the same sample. → Outputs should be similar.
  A decision tree trained as surrogate model for a neural network. Calculate accuracy of the decision tree w.r.t. to the model output (not the groundtruth.)

Co-12 Consistency

How deterministic and implementation invariant is the explanation?

- **Implementation Invariance** Calculate agreement between model variants, e.g., hyperparameters or random initialization (but the same predictive performance). → Explanations should not change (much).

- **Robustness to Model Changes** Change the model slightly. → Explanations should not change (much).



LEFT: explanation of original model
RIGHT: explanation of slightly changed model

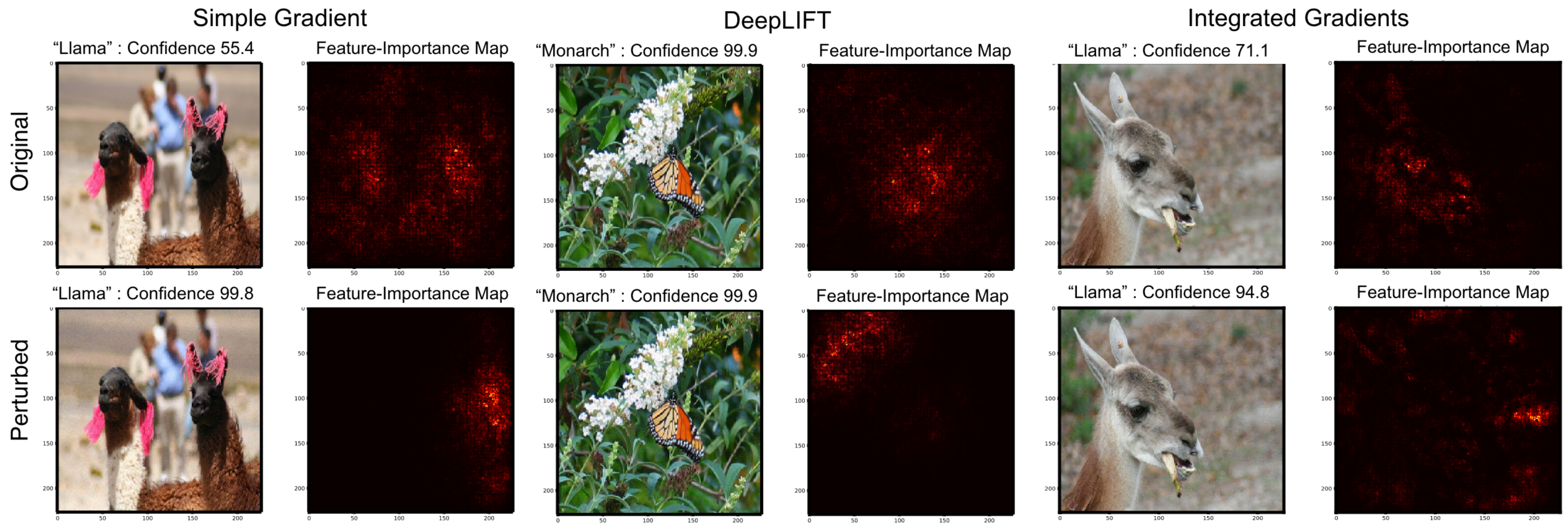
Christopher Anders et al. “Fairwashing Explanations with Off-Manifold Detergent”. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, Nov. 2020, pp. 314–323.

Co-12 Continuity

How sensitive is the explanation to small input changes?

- **Connectedness** Measure similarity of counterfactuals to real samples. → Should not be an outlier.

- **Stability for Slight Variations** Measure the difference between explanations for two similar examples (input features and model output). → Small changes in the input should not result in very different explanations.



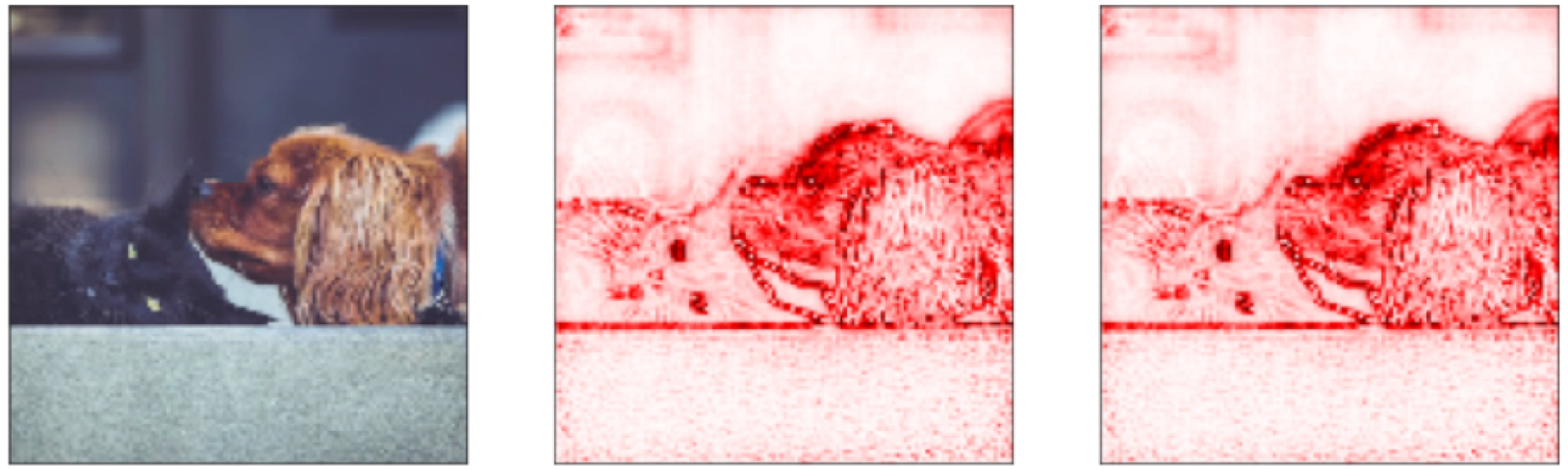
TOP: original sample
BOTTOM: slightly perturbed sample

Amirata Ghorbani, Abubakar Abid, and James Zou. "Interpretation of Neural Networks Is Fragile". In: Proceedings of the AAAI Conference on Artificial Intelligence 33.01 (July 2019), pp. 3681–3688.

Co-12 Contrastivity

How discriminative is the explanation w.r.t. other events?

- **Target Discriminativeness** Train classifier on explanations for different targets. → Should have high accuracy.
- **Data Randomization Check** Randomize labels in training data. Train a second model on this randomized data. Get explanations for a data samples for both models. → Explanations should be different.
- **Target Sensitivity** Calculate explanation for different target labels. → Should be different.



(b) Image (c) Expl. Cat (d) Expl. Dog

Leon Sixt, Maximilian Granz, and Tim Landgraf. “When Explanations Lie: Why Many Modified BP Attributions Fail”. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, Nov. 2020, pp. 9046–9057.

Co-12 Compactness

Size of the explanation.

- **Size** Depends on explanation type. E.g. number of rules in a decision set, height of a decision tree, number of pixels in an heatmap.

- **Redundancy** Not only the amount, but also the uniqueness of features is relevant. E.g. amount of learned prototypes that are very similar and represents the same concept.

- **Counterfactual Compactness** For counterfactual explanations. Measure how much is changed to explain a different outcome.

Original (positive): I liked this movie very much.
Explanation 1 (negative): I did not like this movie.
Explanation 2 (negative): This movie was one of the worst ideas ever!

Name, Description and *Main Explanation Types*	**References**
CONTINUITY (Section 6.4)	
Stability for Slight Variations *Feature importance, Heatmap, Graph, Text, Localization, Decision Rules, White-box model* Measure the similarity between explanations for two slightly different samples. Small variations in the input, for which the model response is nearly identical, should not lead to large changes in the explanation.	[8, 27, 31, 52, 60, 78, 78, 95, 136, 144, 145, 191–193, 198, 230, 240, 247, 257, 284]
Fidelity for Slight Variations – *Decision Rules, White-box model* Measure the agreement between interpretable predictions for original and slightly different samples: an explanation for original input x should accurately predict the model's output for a slightly different sample x'.	[136, 192]
Connectedness – *Prototypes, Representation Synthesis* Measure how connected a counterfactual explanation is to samples in the training data: ideally, the counterfactual is not an outlier, and there is a continuous path between a generated counterfactual and a training sample.	[120, 140, 187]
CONTRASTIVITY (Section 6.5)	
Target Sensitivity – *Heatmap* The explanation for a particular target or model output (e.g. class) should be different from an explanation for another target.	[176, 195, 232, 237, 261, 264]
Target Discriminativeness – *Disentanglement, Representation Synthesis, Text* The explanation should be target-discriminative such that *another model* can predict the right target (e.g. class label) from the explanation, in either a supervised or unsupervised fashion.	[30, 71, 113, 129, 231, 256, 259, 271, 278]
Data Randomization Check – *Feature importance, Heatmap, Localization* Randomly change labels in a copy of the training dataset, train a model on this randomized dataset and check that the explanations for this model on a test set are different from the explanations for the model trained on the original training data.	[3, 145, 209]

Summary

- Many metrics have been proposed, some only differ slightly. **Co-12** gives an overview of **general criteria**.

- There is a **tradeoff**. E.g. smaller explanations are easier to understand but less correct (correctness vs. compactness). Explanations that are more coherent (do more align with user knowledge and expectations) are not necessarily correct.

Evaluation Toolkits

Toolkit	Usability	Stars	ML	Data Types	Expl. Type	Co-12 Coverage
XAI EVALUATION TOOLKITS						
Ablation (2022)[a]	5-4-5	8	P	S	FI	■□□□□□□□□□□□
CompareXAI (2022)[b]	4-2-3	7	S	I, S, X	FI	■□□□□□□□□□■□
ExPMRC (2022)[c]	0-1-4	57	n.a.[d]	X	LC	□□□□□□□□□□■□
GraphXAI (2022)[e]	4-2-4	57	P	G	FI	□■□■□□□□□□■□
OpenXAI (2022)[f]	4-3-4	121	P	S	FI, HM	■□□■□□□□□□■□
Quantus (2022)[g]	5-4-5	271	PT	I, S, T	FI, HM, LC, PT, DT	■■□■■□■□□□■□
Safari (2022)[h]	2-0-2	2	P	I, S, X	HM	□□□■□□□□□□□□
Eval XAI (2021)[i]	4-0-1	5	P	I	FI, HM	■□□□□□□□□□□□
PhE-Eval (2021)[j]	2-0-4	1	ST	I, S	FI, HM	■□□□□□□□□□□□
XAI-Bench(2021)[k]	2-2-4	32	S	I	FI, HM	■□□■□□□□□□□□
XAI-Eval(2021)[l]	0-0-1	2	K	I	HM, LC	□□■□□□□□□□■□
BAM (2019)[m]	2-2-4	44	T	I	HM, PT	■□□■■□□□□□□□

P - PyTorch / T - Tensorflow

Phuong Quynh Le, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, Christin Seifert “Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges” Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence Survey Track. Pages 6665-6673

Evaluation Datasets

Toolkit	Dataset	Description	Task	Size	B
ExPMRC	Squad	Span extraction from Wikipedia (English)	MRC	1003 (Q), 632 (P)	✓
	CMRC	Span-extraction (Chinese)	MRC	1015 (Q), 768 (P)	✓
	$Race^{+}$	Multiple-choice exams (English)	MRC	1125 (Q), 335 (P)	✓
	C^3	Multiple-choice exams (Chinese)	MRC	1005 (Q), 517 (P)	✓
GraphXAI	MUTAG	Nitroaromatic compounds, mutagenicity prediction	GC	1768 (G)	
	Benzene	Molecules, with our without benzene ring	GC	12000 (G)	
	Fluoride-carbonyl	Molecules, with or without fluoride and carbonyl	GC	8671 (G)	
	Alkanyl-carbonyl	Molecules, with or without alkane and carbonyl	GC	4326 (G)	
	SG-X	4 datasets of synthetic graphs with varying properties	NC	>13000 (N)	
BAM	Obj, Scene, Scene_only	3 datasets combining MSCOCO and MiniPlaces, labels are objects or scene labels	C	100 k (I)	
XAI-Bench	Synthetic	(Mixtures) of probability distributions	R/C	n.a. (S)	✓
OpenXAI	Synthetic	20 continuous features from Gaussian distribution	C	5000 (S)	✓

Table 2: XAI evaluation datasets with explanation ground truth available in the analysed toolkits. (B) indicates whether there is a benchmark available. Tasks: machine reading comprehension (MRC), graph-level classification (GC), node classification (NC), classification (C), regression (R). Size (Number of): questions (Q), passages (P), graphs (G), nodes (N), images (I), structured data (S). n.a. – information not available, neither in the publication nor in the GitHub repository.

Benchmarks

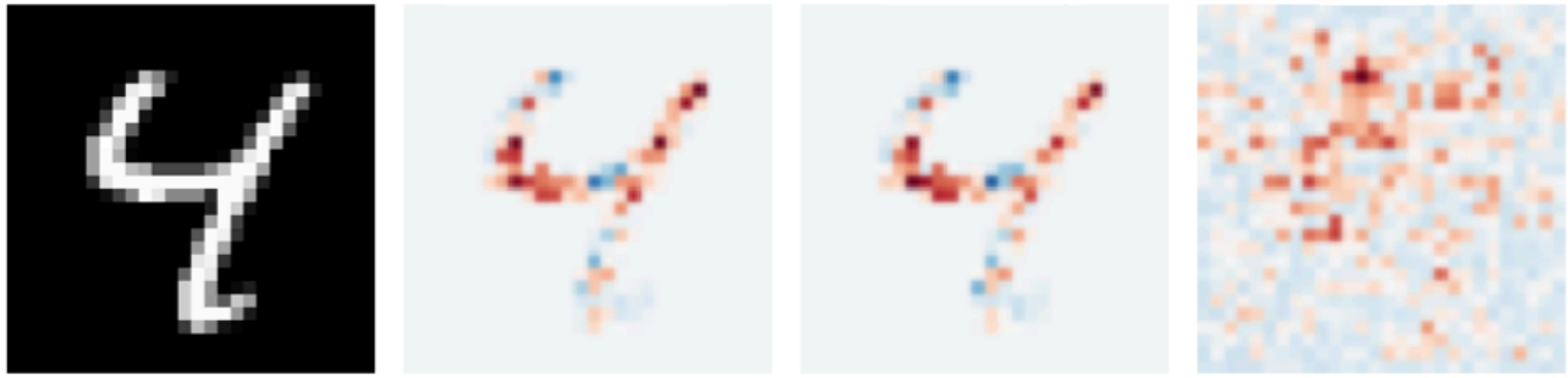
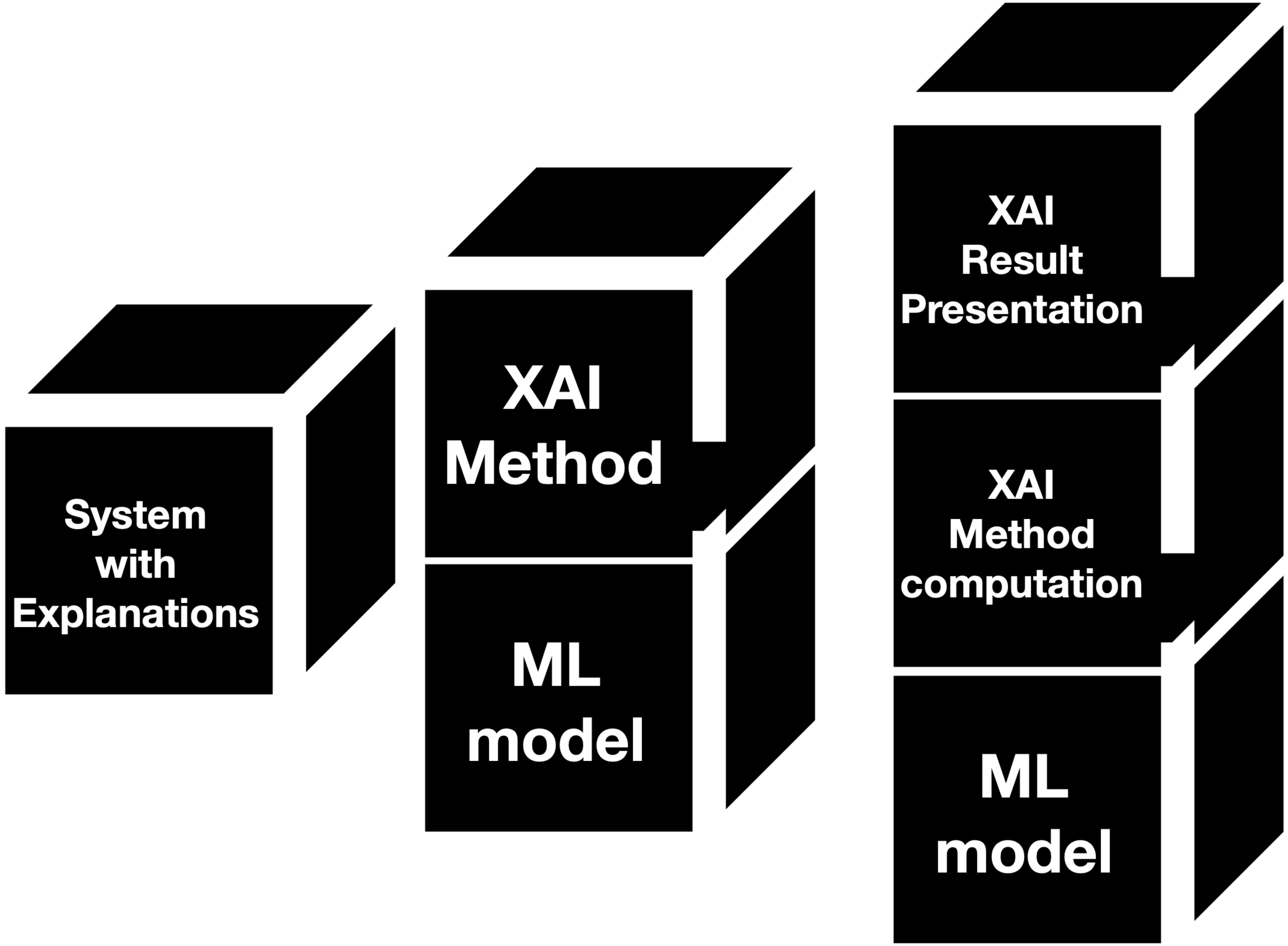


Figure 2: Original image and explanations from Intergrated Gradients, GradientShap and Saliency methods (left to right).

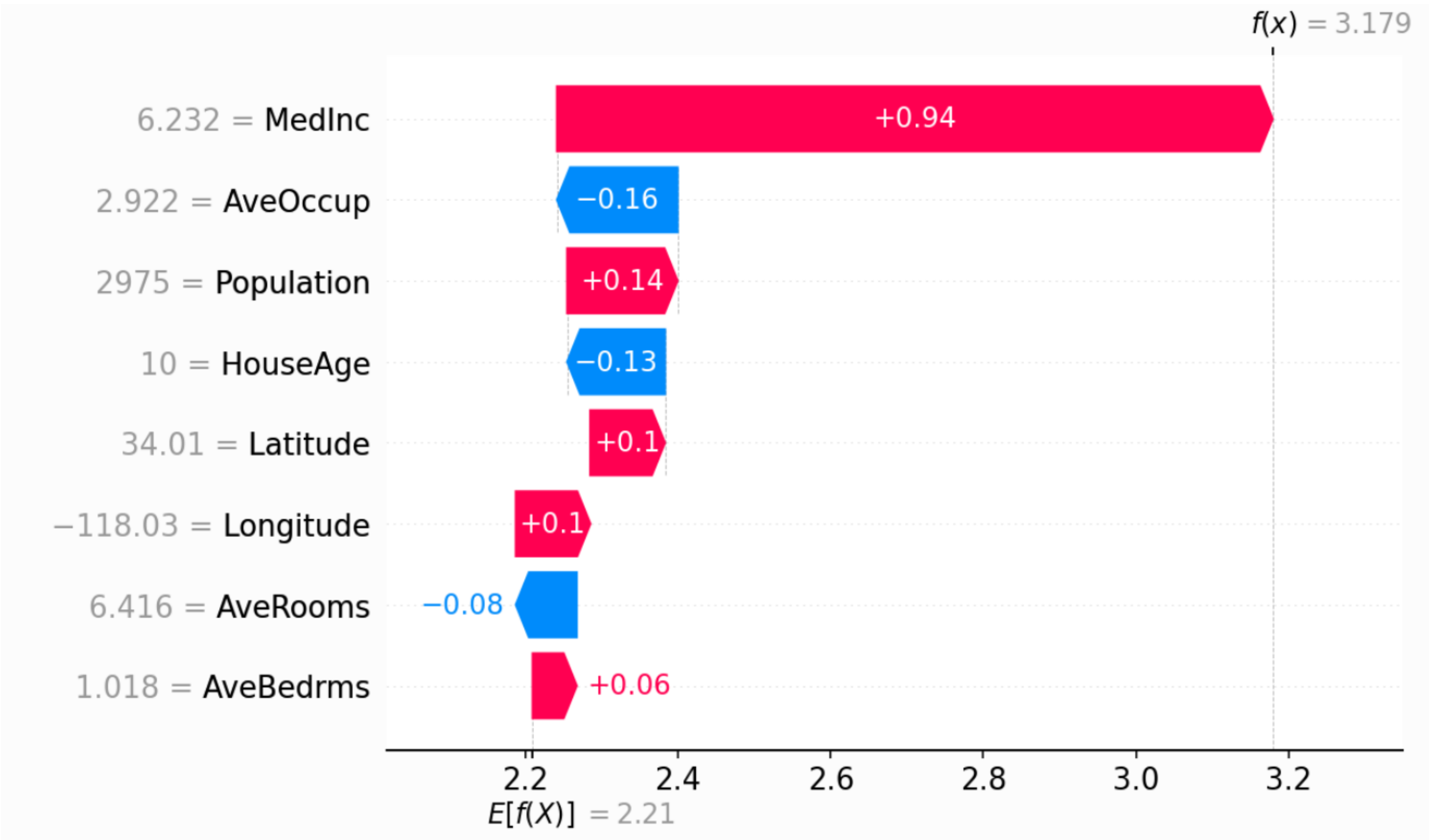
Toolkit	XAI method		
	IG	GradientShap	Saliency
Original	1.21	1.56	10.02
Quantus	24780	25635	5356752
Captum	5735	7098	7423
InterpretDL	2.36	3.19	13.81

Infidelity measure proposed in a paper as measured by different XAI evaluation libraries

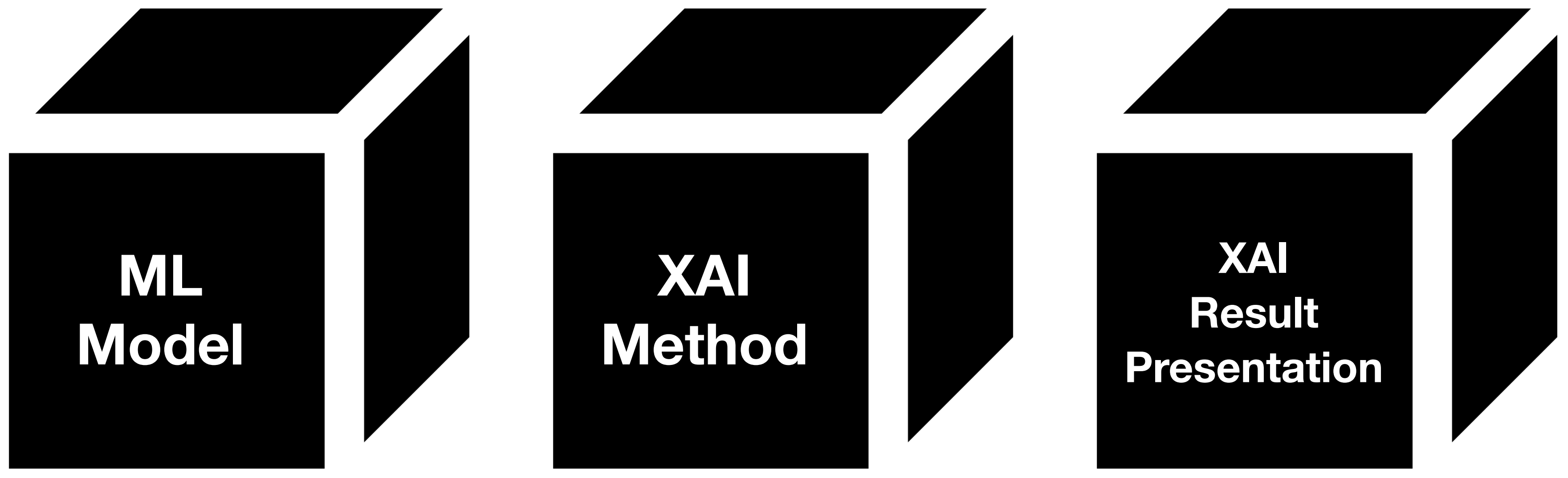
It's even more complex…



	Feature	Average	Attribution
Model Null			22.92
Crime per Capita	3.85	4.06	1.75
Residental Zoning %	0.00	13.23	-0.02
% Industrial Zoning	18.10	10.40	0.02
House on the River?	1.00	0.09	0.57
Nitric Oxides PPM	0.77	0.54	0.20
# Rooms	6.39	6.26	-1.23
% Houses Older than 1940	91.00	62.66	-0.30
Distance to Employment Hubs	2.51	3.94	0.11
Highway Accessibility	24.00	8.74	0.05
Property Tax	666.00	393.96	-0.23
Pupil-Teacher Ratio	20.20	18.19	-0.57
% Lower Status	13.27	11.99	-1.19



Conclusion



- Evaluating XAI methods is **necessary**.
- There is **no one-fits-all** measure / metric.
- Evaluation is **multifactorial**, which factors depends on application.
- XAI evaluation toolkits are available, but do not report consistent results -> use eval toolkits **AND** report which toolkit was used.
- Some evaluation schemes measure both, **evaluation metric calculation AND presentation** (including human perception of colours etc.)

TODO: Unified evaluation paradigm.

Correctness
Match between model and explanation.

Completeness
How much of the model is explained?

Consistency
Robustness to small changes in model and implementation.
$g(x) = g(x)$

Continuity
Robustness to small changes input.
$g(x) = g(x')$

Contrastivity
Discriminative to other events or targets?
g(x|Cat) != g(x|Dog)

Covariate Complexity
Complexity of features in the explanation

Compactness
Size of the explanation

Composition
Presentation format

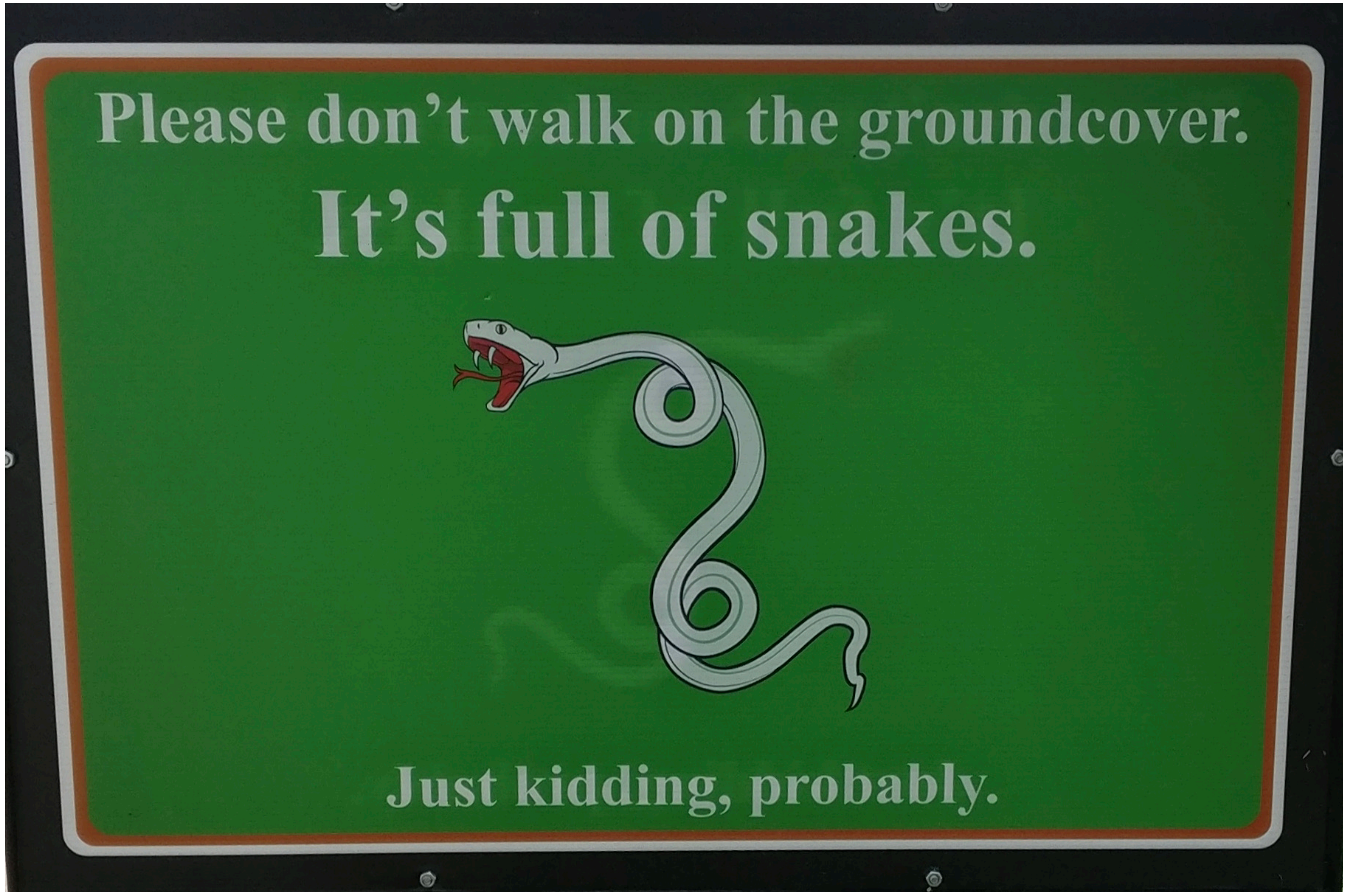
Confidence
Probability information available?
$p = ?$

Context
Useful for users?

Coherence
Match with domain knowledge.
$g(x) =$

Controllability
Can user influence explanation?
$g(x)$

Explanation / Model / User



Christin.Seifert@uni-marburg.de