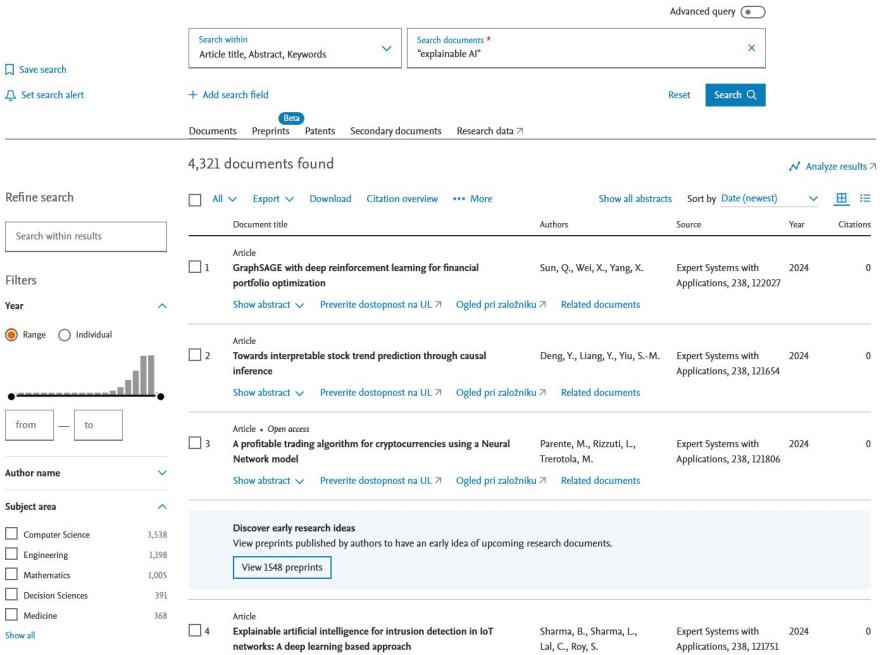
Explainable AI

A brief overview

Outline

- General state of explainable AI,
- Main problems and methods,
- Model agnostic feature importance,
- Methods for deep learning,
- Evaluating explainable AI.

The popularity of XAI



Schwalbe & Finzel (2023): A comprehensive taxonomy for explainable artificial intelligence: a systematic **survey of surveys** on methods and concepts

Explainable AI (XAI) - What is it?

Understanding is described as the human ability to recognize correlations, as well as the context of a problem and is a necessary precondition for explanations (Bruckert et al, 2020). The concept of understanding can be divided into mechanistic understanding (*"How does something work?"*) and functional understanding (*"What is its purpose?"*) (Páez, 2019).

Explicability refers to making properties of an AI model inspectable (Bruckert et al, 2020).

Explainability goes one step further than *explicability* and aims for making (a) the context of an AI system's reasoning, (b) the model, or (c) the evidence for a decision output accessible, such that they can be *understood* by a human (Bruckert et al, 2020).

Transparency is fulfilled by an AI model, if its algorithmic behaviour with respect to decision outputs or processes can be *understood* by a human *mechanistically* (Páez, 2019). Transparency will be discussed more closely in Subsubsection 5.1.2.

Explaining means utilizing *explicability* or *explainability* to allow a human to *understand* a model and its purpose (Bruckert et al, 2020; Páez, 2019).

Global explanations *explain* the model and its logic as a whole ("How was the conclusion derived?").

Local explanations *explain* individual decisions or predictions of a model ("Why was this example classified as a car?").

Interpretability means that an AI model's decision can be *explained globally* or *locally* (with respect to *mechanistic understanding*), and that the model's purpose can be *understood* by a human actor (Páez, 2019)(*i.e. functional understanding*).

Correctability means that an AI system can be adapted by a human actor in a targeted manner in order to ensure correct decisions (Kulesza et al, 2015; Teso and Kersting, 2019; Schmid and Finzel, 2020). Adaptation refers either to re-labelling of data (Teso and Kersting, 2019) or to changing of a model by constraining the learning process (Schmid and Finzel, 2020).

Interactivity applies if one of the following is possible: (a) interactive explanations, meaning a human actor can incrementally explore the internal working of a model and the reasons behind its decision outcome; or (b) the human actor may adapt the AI system (*correctability*).

Comprehensibility relies, similar to *interpretability*, on local and global *explanations* and *functional understanding*. Additionally, *comprehensible* artificial intelligence fulfills *interactivity* (Bruckert et al, 2020; Schmid and Finzel, 2020). Both, *interpretable* presentation and intervention are considered as important aspects for in depth *understanding* and therefore preconditions to *comprehensibility* (see also (Gleicher, 2016)).

Human-AI system is a system that contains both algorithmic components and a human actor, which have to cooperate to achieve a goal (Schmid and Finzel, 2020). We here consider in specific **explanation systems**, *i.e.*, such human-AI systems in which the cooperation involves *explanations* about an algorithmic part of the system (the *explanandum*) by an *explanator* component, to the human interaction partner (the *explainee*) resulting in an action of the human (Bruckert et al, 2020).

Explanandum (*what is to be explained*, cf. Subsection 5.1) refers to what is to be *explained* in an *explanation system*. This usually encompasses a model (*e.g.*, a deep neural network), We here also refer to an explanandum as the object of explanation.

Explanator (*the one that explains*, cf. Subsection 5.2) is the *explanation system* component providing *explanations*.

Explainee *(the one to whom the explanandum is explained)* is the receiver of the *explanations* in the *explanation system*. Note that this often but not necessarily is a human. *Explanations* may also be used *e.g.*, in multi-agent systems for communication between the agents and without a human in the loop in most of the information exchange scenarios.

Interpretable models are defined as machine learning techniques that learn more structured representations, or that allow for tracing causal relationships. They are *inherently interpretable* (cf. definition in Subsection 5.2), *i.e.*, no additional methods need to be applied to *explain them*, unless the structured representations or relationship are too complex to be processed by a human actor at hand.

Interpretable machine learning (iML) is the area of research concerned with the creation of *interpretable* AI systems (*interpretable models*).

Model induction (also called model distillation, student-teacher approach, or reprojection (Gleicher, 2016)) is a strategy that summarizes techniques which are used to infer an approximate *explainable* model—the (*explainable*) *proxy* or *surrogate model*—by observing the input-output behaviour of a model that is *explained*.

Deep explanation refers to combining deep learning with other methods in order to create hybrid systems that produce richer representations of what a deep neural network has learned, and that enable extraction of underlying semantic concepts (Gunning and Aha, 2019).

Comprehensible artificial intelligence (cAI) is the result of a process that unites *local interpretability* based on *XAI* methods and *global interpretability* with the help of *iML* (Bruckert et al, 2020). The ultimate goal of such systems would be to reach *ultra-strong machine learning*, where machine learning helps humans to improve in their tasks. For example, (Muggleton et al, 2018) examined the *comprehensibility* of programs learned with Inductive Logic Programming, and (Schmid et al, 2016; Schmid and Finzel, 2020) showed that the *comprehensibility* of such programs could help laymen to *understand* how and why a certain prediction was derived.

Explainable artificial intelligence (XAI) is the area of research concerned with *explaining* an AI system's decision.

from Schwalbe & Finzel

Many different stakeholders and tasks

- Researchers,
- professionals,
- decision-makers,
- customers/impacted groups,
- regulatory bodies.

- Diagnostics tool,
- decision support
- regulatory compliance…
- … improve performance,
- trust,
- confidence.

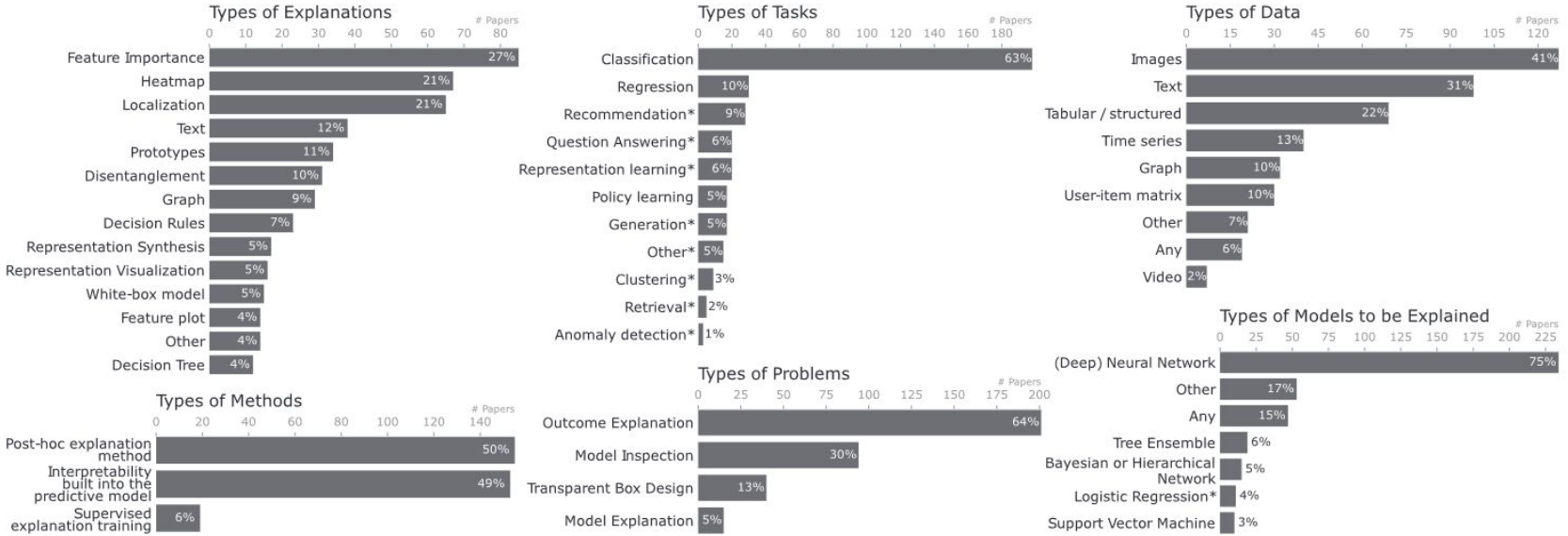
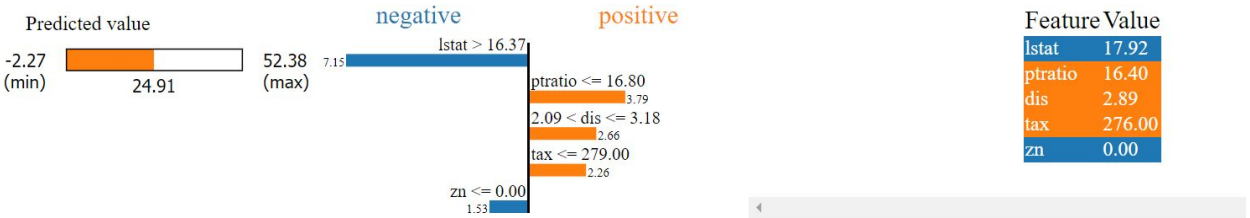


Fig. 5. Categorization of papers introducing an explainable AI method, following the six dimensions as presented in Section 3.2. Note that categories are non-exclusive, so a paper can fall into multiple categories per dimension. *: category is manually added after the reviewing process and might therefore not be complete (i.e., high precision, potentially low recall).

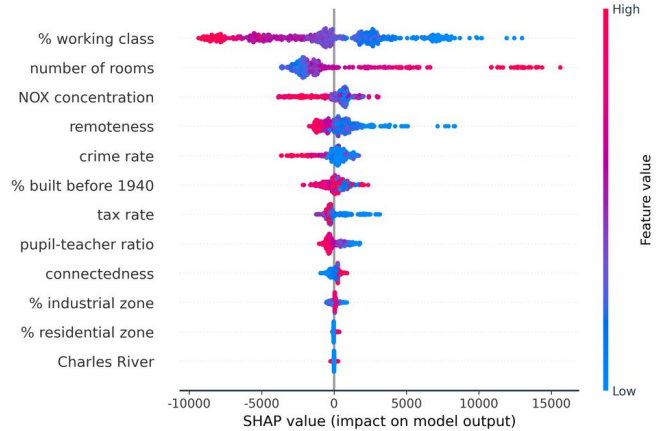
from Nauta et al.

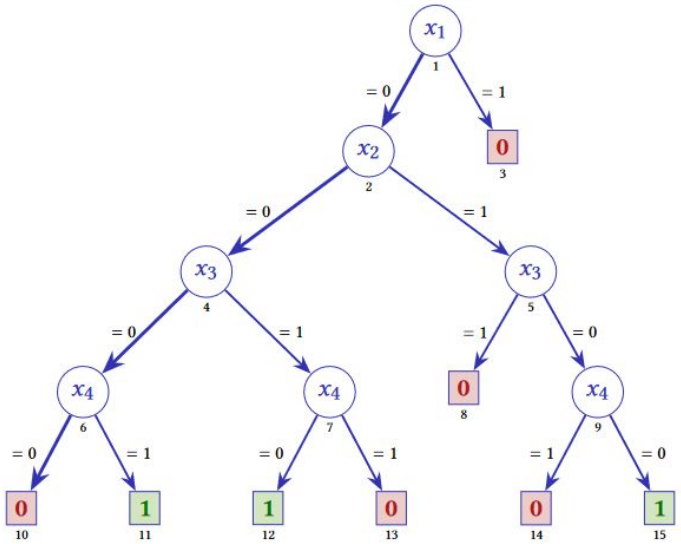
Model agnostic feature importance

LIME



SHAP





row #	x_1	x_2	x_3	x_4	$\kappa(\mathbf{x})$
1	0	0	0	0	0
2	0	0	0	1	1
3	0	0	1	0	1
4	0	0	1	1	0
5	0	1	0	0	1
6	0	1	0	1	0
7	0	1	1	0	0
8	0	1	1	1	0
9	1	0	0	0	0
10	1	0	0	1	0
11	1	0	1	0	0
12	1	0	1	1	0
13	1	1	0	0	0
14	1	1	0	1	0
15	1	1	1	0	0
16	1	1	1	1	0

Figure 1: Example classifier – decision tree and its truth table. For this classifier, we have $\mathcal{F} = \{1, 2, 3, 4\}$, $\mathcal{D}_i = \{0, 1\}, i = 1, 2, 3, 4$, $\mathbb{F} = \{0, 1\}^4$, and $\mathcal{K} = \{0, 1\}$. The classification function is given by the decision tree shown, or alternatively by the truth table. Finally, the instance considered is $((0, 0, 0, 0), 0)$, corresponding to row 1 in the truth table. The instance is consistent with path $\langle 1, 2, 4, 6, 10 \rangle$, which is highlighted in the DT. The prediction is 0, as indicated in terminal node 10.

from Marques-Silva & Huang

Counterfactual explanations

Query instance (original outcome : 0)

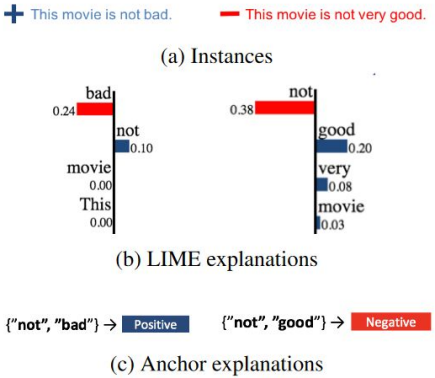
	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	22.0	Private	HS-grad	Single	Service	White	Female	45.0	0.009411

Diverse Counterfactual set (new outcome : 1)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	57.0	Private	Doctorate	Single	White-Collar	White	Female	45.0	0.724
1	36.0	Private	Prof-school	Married	Service	White	Female	37.0	0.869
2	22.0	Self-Employed	Doctorate	Married	Service	White	Female	45.0	0.755
3	43.0	Private	HS-grad	Married	White-Collar	White	Female	63.0	0.822

https://github.com/interpretml/DiCE

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

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts P(beagle) $> 90\%$

from Ribeiro et al.

Deep Learning: Heatmaps & Localization

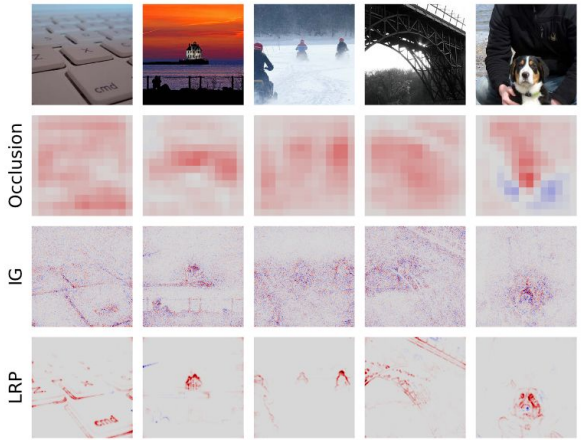
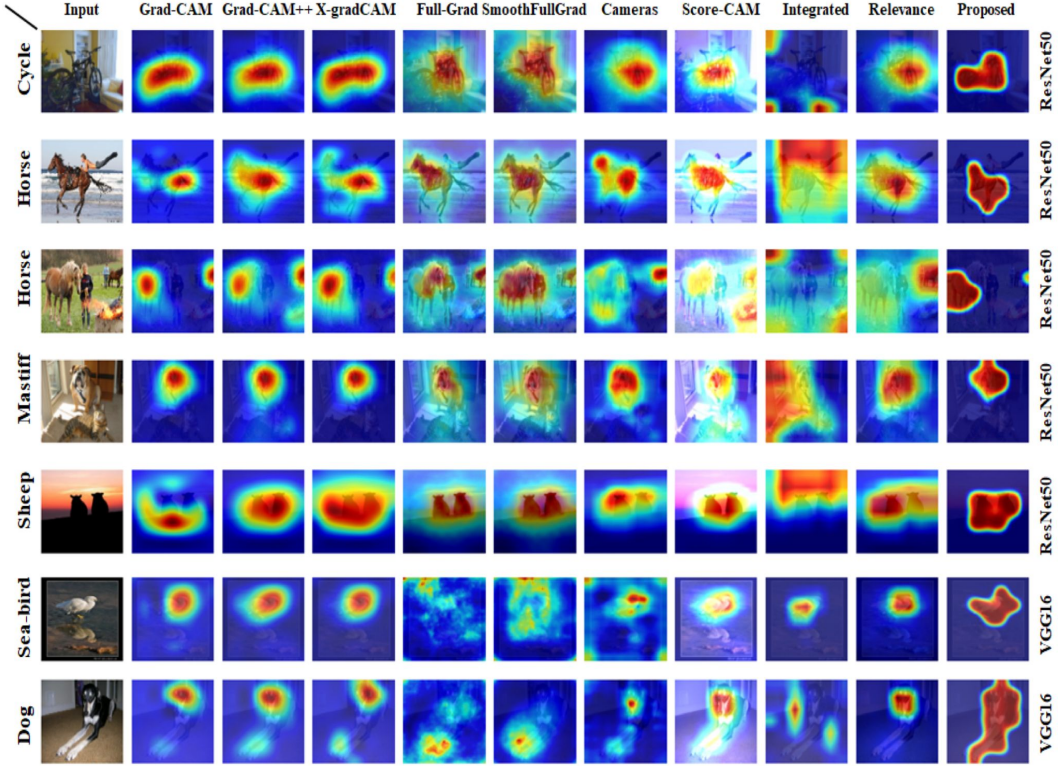


Fig. 8. *Example of two images predicted to be similar, along with a BiLRP second-order attribution of their similarity score rendered as a bipartite graph. (figure is adapted from [42]). The explanation shows that the front part of the two planes jointly contributes to the predicted similarity.*



from Samek et al.

Deep Learning: Heatmaps & Localization

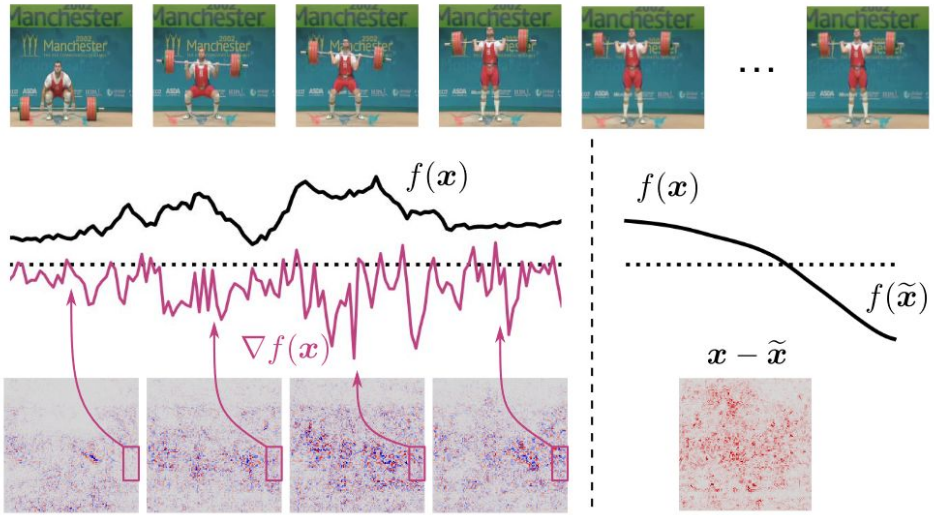
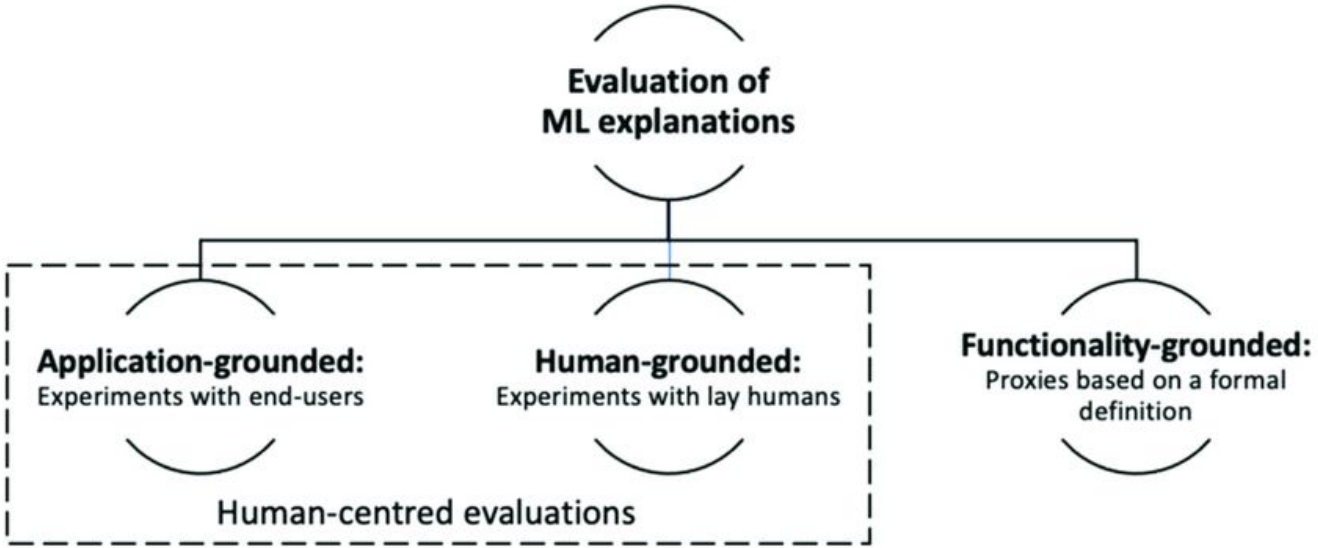


Fig. 3. ***Two difficulties encountered when explaining DNNs. Left: shattered gradient effect causing gradients to be highly varying and too noisy to be used for explanation. Right: pathological minima in the function, making it difficult to search for meaningful reference points.***

Deep Learning: Global explanation

Evaluating explainable AI

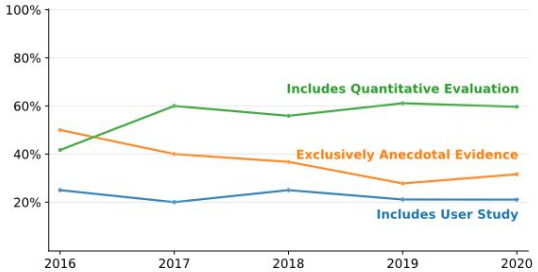


from Doshi-Velez & Kim

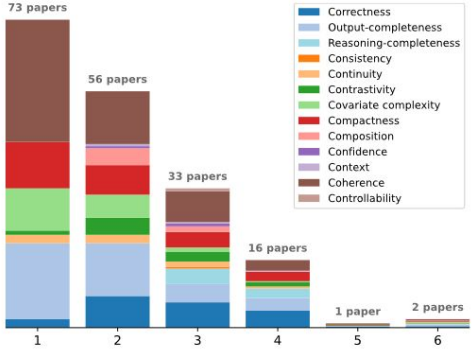
Evaluating explainable AI

Table 2. Our Co-12 Explanation Quality Properties, Grouped by Their Most Prominent Dimension: Content, Presentation, or User

	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?



(a) Evaluation practices of the 312 papers that introduce a method for explaining a machine learning model.

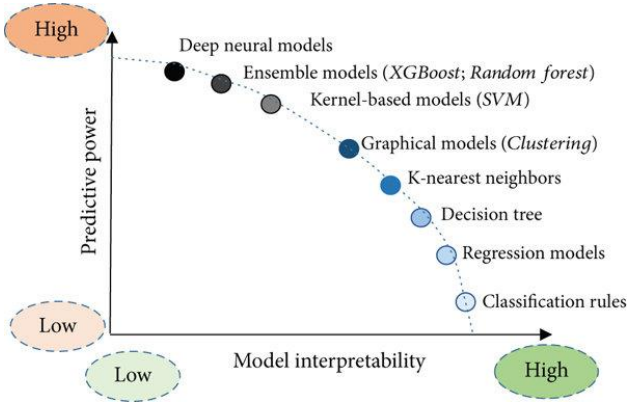


(b) Total number of unique Co-12 properties quantitatively evaluated in a paper that introduces an XAI method.

from Nauta et al.

Summary

- The field is very young,
- requires interdisciplinary work,
- currently mostly ignoring the user,
- how do we evaluate human explanations?,
- missing a proper legal framework,
- lots of opportunities!



Kumar et al. (2021): Explainable Artificial Intelligence for Sarcasm Detection in Dialogues

References

- **Doshi-Velez & Kim (2017):** Towards a rigorous science of interpretable machine learning,
- **Huang & Marques-Silva** (2023, unpublished): The Inadequacy of Shapley Values for Explainability,
- **Lundberg & Lee** (2017): A unified approach to interpreting model predictions,
- **Mase et al.** (2019, unpublished): Explaining black box decisions by Shapley cohort refinement,
- **Marques-Silva** (2023, unpublished): Disproving XAI Myths with Formal Methods -- Initial Results,
- **Marques-Silva & Huang** (2023, unpublished): Explainability is NOT a Game,
- **Mothilal et al.** (2020): Explaining machine learning classifiers through diverse counterfactual explanations,
- **Nauta et al.** (2023): From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI,
- **Ribeiro et al.** (2016): "Why Should I Trust You?": Explaining the Predictions of Any Classifier,
- **Ribeiro et al.** (2019): Anchors: High Precision Model-Agnostic Explanations,
- **Saeed & Omlin** (2022): Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities,
- **Samek et al.** (2021): Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications
- **Schwalbe & Finzel** (2023): A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts.