Explainable AI (XAI)

A brief overview

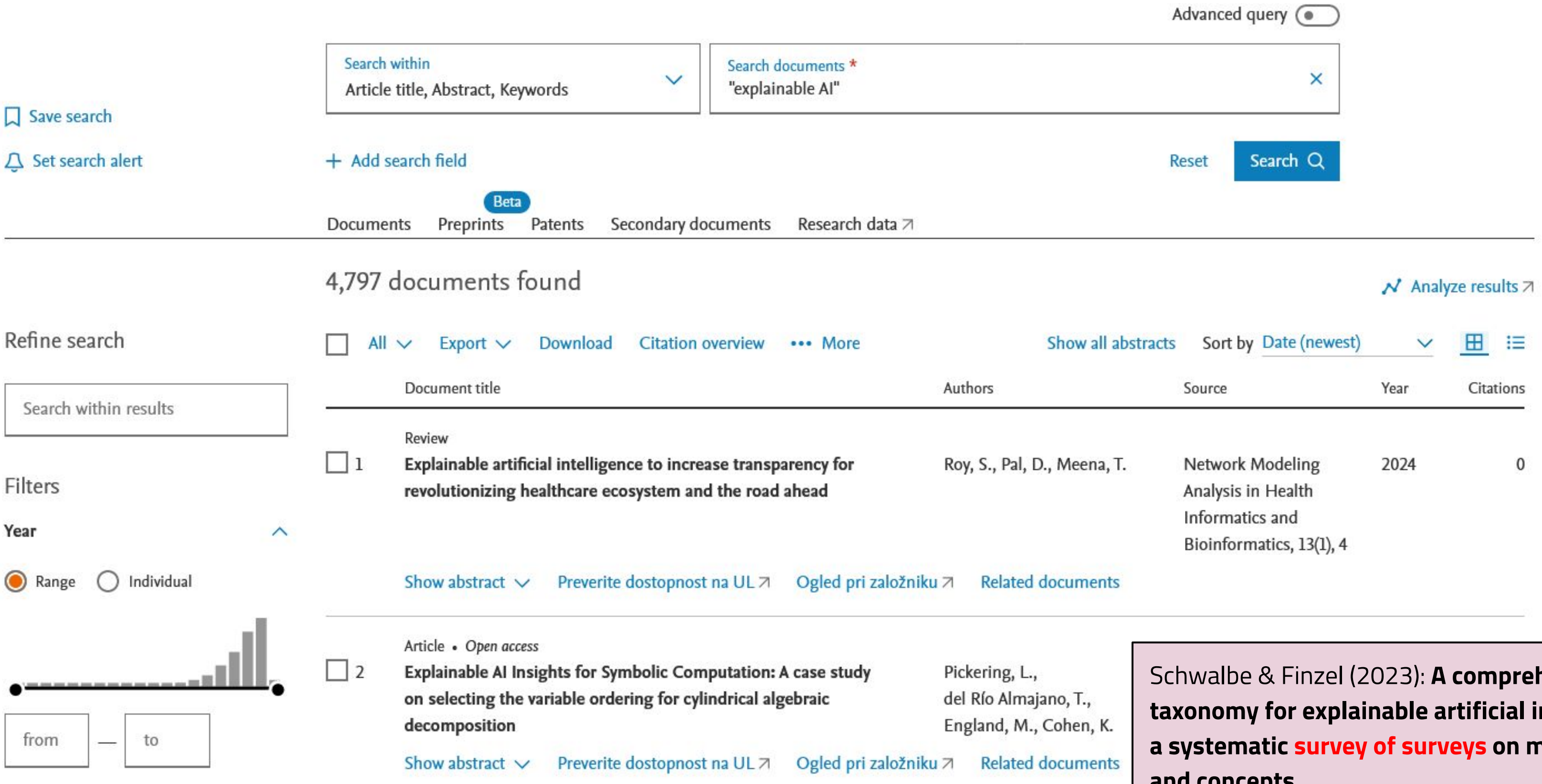
Two-part lecture:

- **Brief overview of the field of eXplainable AI (XAI).**
- **Mechanistic explanations of predictive models.**

  For the ML practitioner. Debug models and occasionally provide some insights to users.

Explainable AI is a young and relatively popular field!

Explainable AI Research in Scopus



Search within: Article title, Abstract, Keywords

Search documents *: "explainable AI"

4,797 documents found

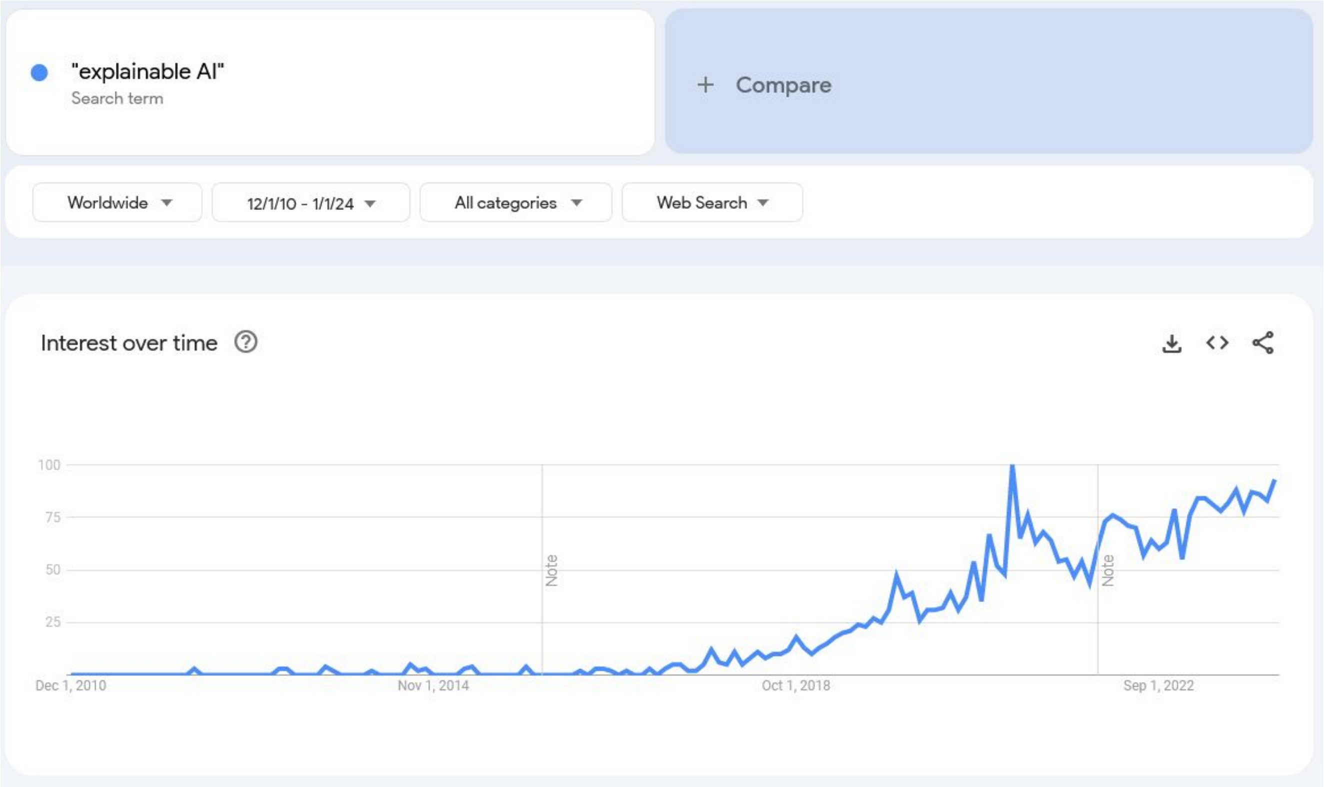
Sort by: Date (newest)

Year: ● Range ○ Individual

	Document title	Authors	Source	Year	Citations
☐ 1	Review **Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead**	Roy, S., Pal, D., Meena, T.	Network Modeling Analysis in Health Informatics and Bioinformatics, 13(1), 4	2024	0
☐ 2	Article • *Open access* **Explainable AI Insights for Symbolic Computation: A case study on selecting the variable ordering for cylindrical algebraic decomposition**	Pickering, L., del Río Almajano, T., England, M., Cohen, K.			

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

Explainable AI on Google Trends

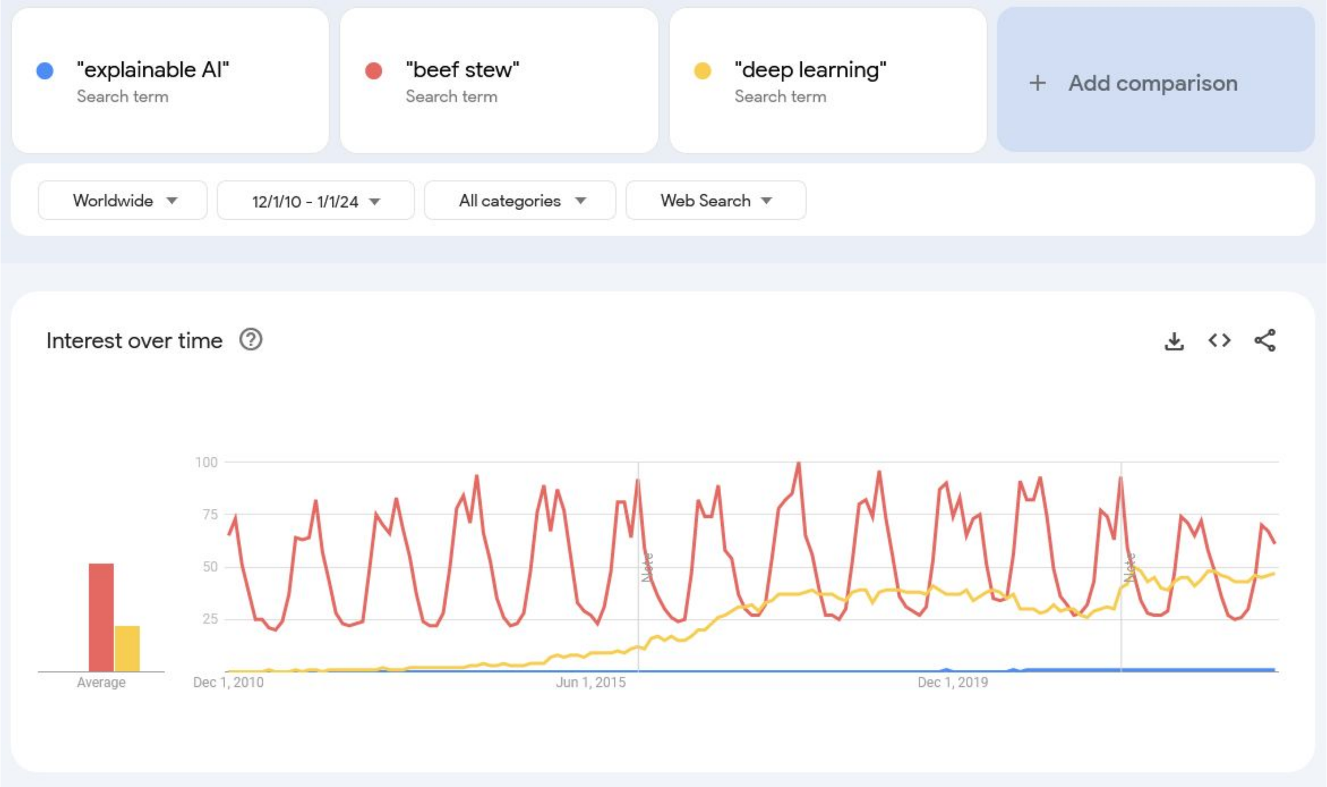


"explainable AI"

Search term

Worldwide: 12/1/10 - 1/1/24, All categories, Web Search

Explainable AI on Google Trends



● "explainable AI"
Search term

● "beef stew"
Search term

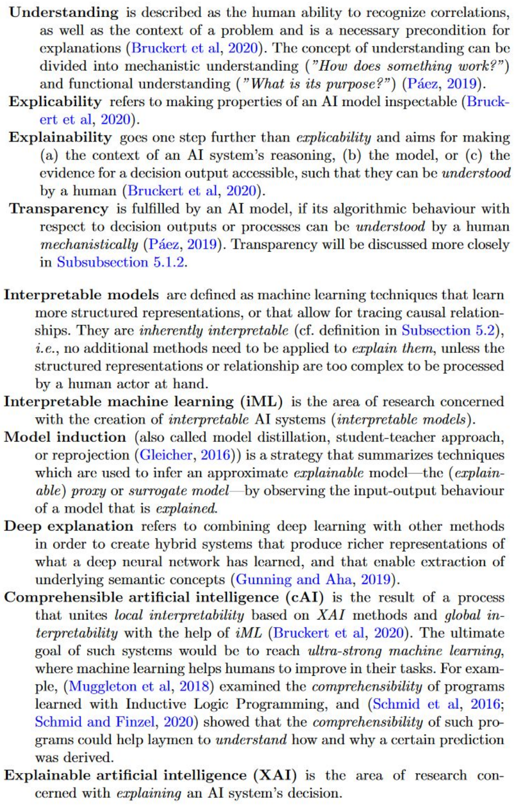
● "deep learning"
Search term

Worldwide
12/1/10 - 1/1/24
All categories
Web Search

Interest over time

What is Explainable AI?

Explainable, interpretable, and/or comprehensible?



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).

2a **Explicability** refers to making properties of an AI model inspectable (Bruckert et al, 2020).

2b **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.

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.

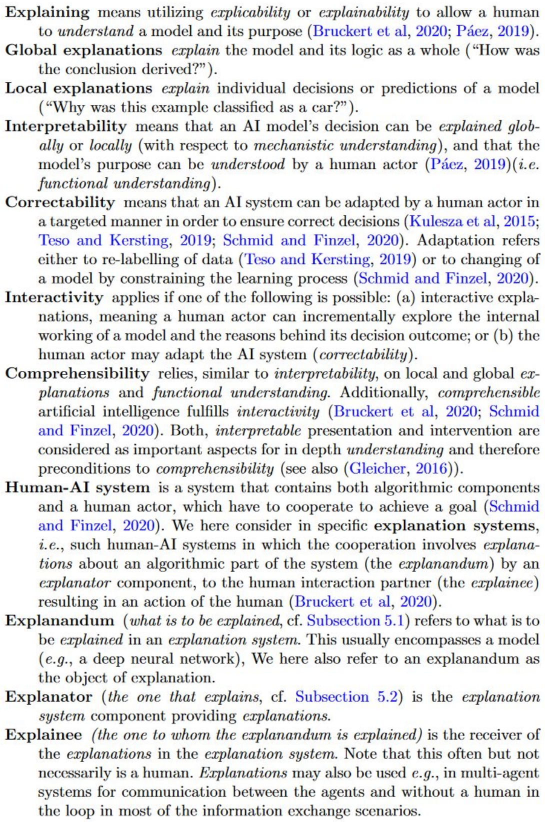
a **Interpretable machine learning (iML)** is the area of research concerned with the creation of *interpretable* AI systems (*interpretable models*).

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).

x **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.

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



1 **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?").

b **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*).

y **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*.

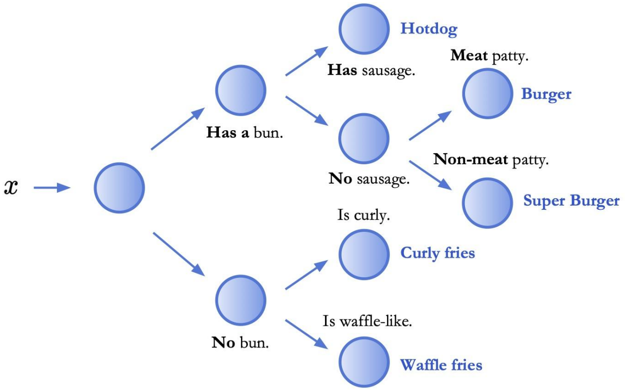
Explainee (*the one to whom the explanandum is explained*) is the receiver of the *explanations* in the *explanation system*. Note that this often but not necessarily is a human. *Explanations* may also be used *e.g.*, in multi-agent systems for communication between the agents and without a human in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

Functional understanding: Focuses on the purpose, behavior, or outcomes of a system without necessarily delving into the underlying mechanisms or processes.

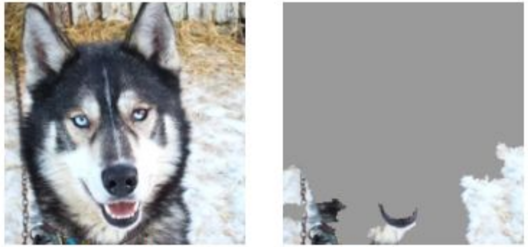
Mechanistic Understanding: Involves a deeper exploration of the inner workings, components, and processes that drive the system's functionality.

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

Examples of explainable AI



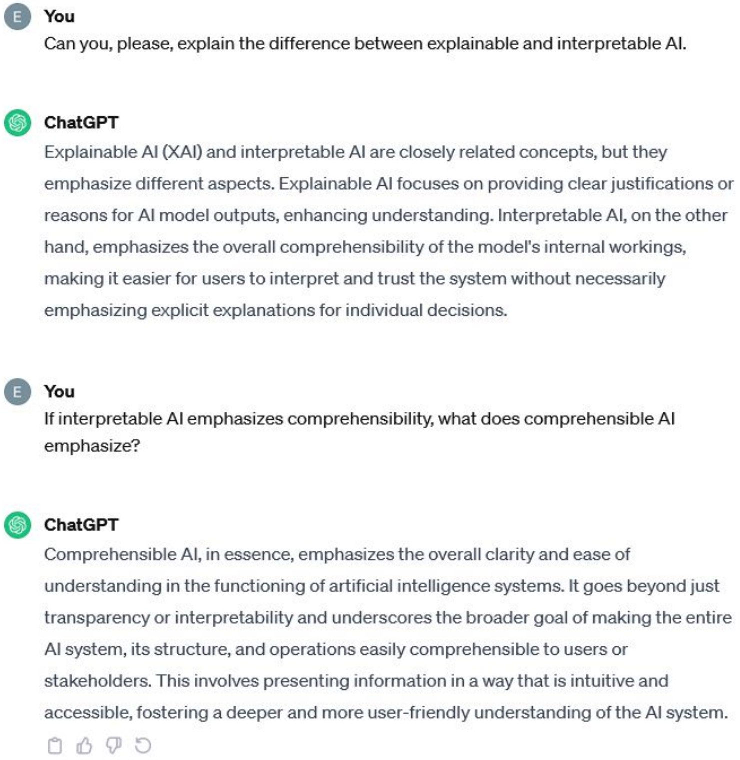
Wan et al (2021): **NBDT: Neural-Backed Decision Tree**



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Ribeiro et al. (2016): **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**



You

Can you, please, explain the difference between explainable and interpretable AI.

ChatGPT

Explainable AI (XAI) and interpretable AI are closely related concepts, but they emphasize different aspects. Explainable AI focuses on providing clear justifications or reasons for AI model outputs, enhancing understanding. Interpretable AI, on the other hand, emphasizes the overall comprehensibility of the model's internal workings, making it easier for users to interpret and trust the system without necessarily emphasizing explicit explanations for individual decisions.

You

If interpretable AI emphasizes comprehensibility, what does comprehensible AI emphasize?

ChatGPT

Comprehensible AI, in essence, emphasizes the overall clarity and ease of understanding in the functioning of artificial intelligence systems. It goes beyond just transparency or interpretability and underscores the broader goal of making the entire AI system, its structure, and operations easily comprehensible to users or stakeholders. This involves presenting information in a way that is intuitive and accessible, fostering a deeper and more user-friendly understanding of the AI system.

ChatGPT 3.5

Broad scope and implications of Explainable AI

Different stakeholders:

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

Different tasks:

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

Outline

- An overview of Explainable AI research activity,
- predictive modeling,
- global insights with feature importance and partial dependence plots,
- views on explaining predictive models,
- common types of local explanation and model-agnostic implementations:
  - Feature importance (LIME, Shapley value-based explanations),
  - Counterfactual explanations,
  - Anchors,
- briefly on computer vision and NLP approaches,
- evaluating explainable AI,
- summary.

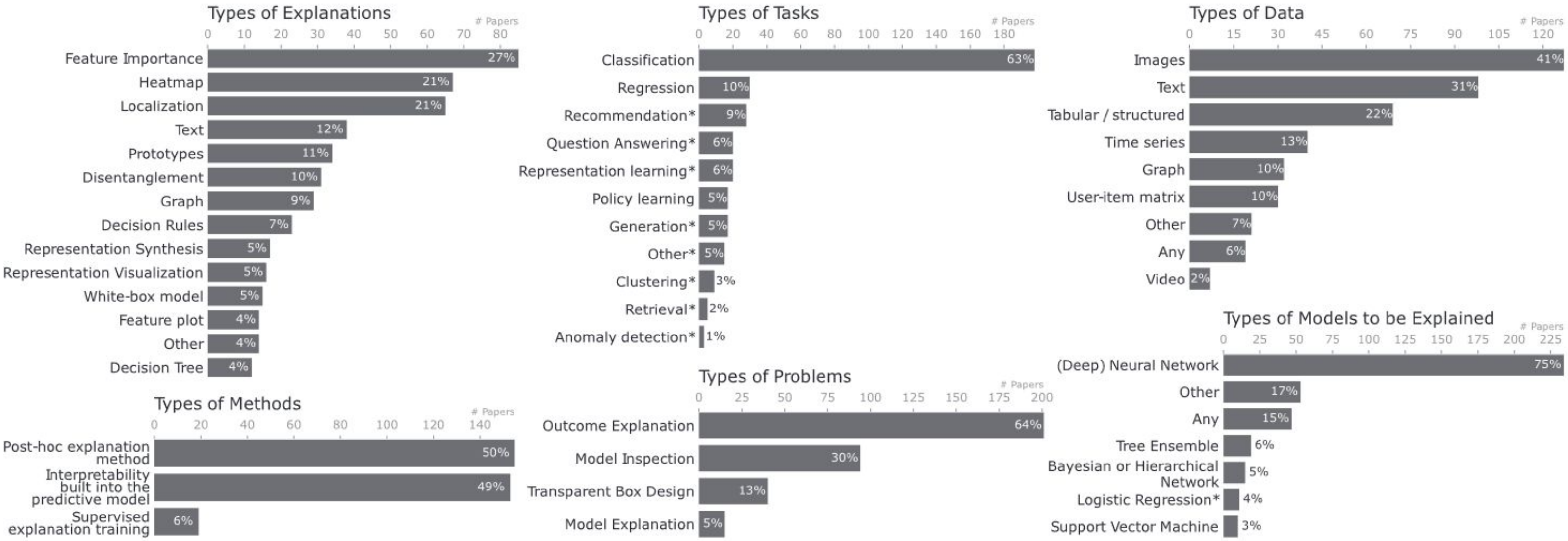
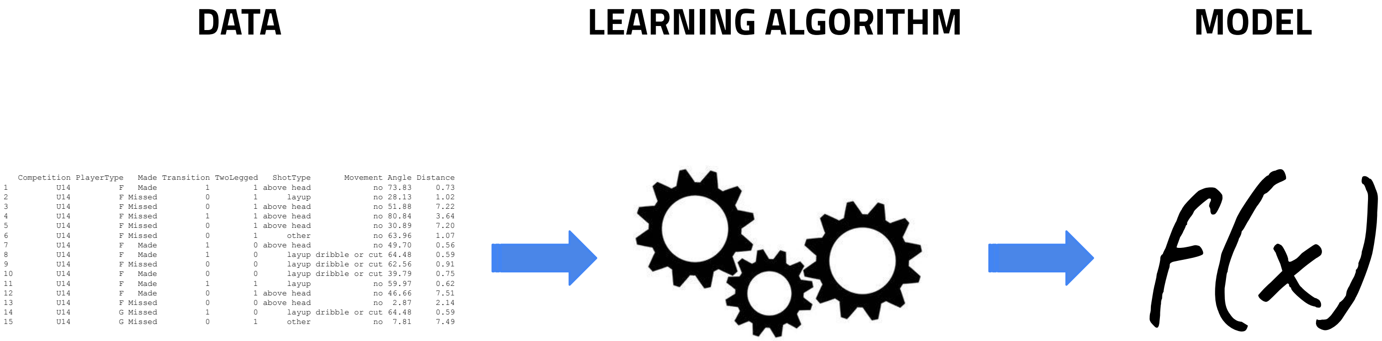


Fig. 5. Categorization of papers introducing an explainable AI method, following the six dimensions as presented in Section 3.2. Note that categories are non-exclusive, so a paper can fall into multiple categories per dimension. *: category is manually added after the reviewing process and might therefore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

Nauta et al. (2023): **From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI**

Explaining a predictive model



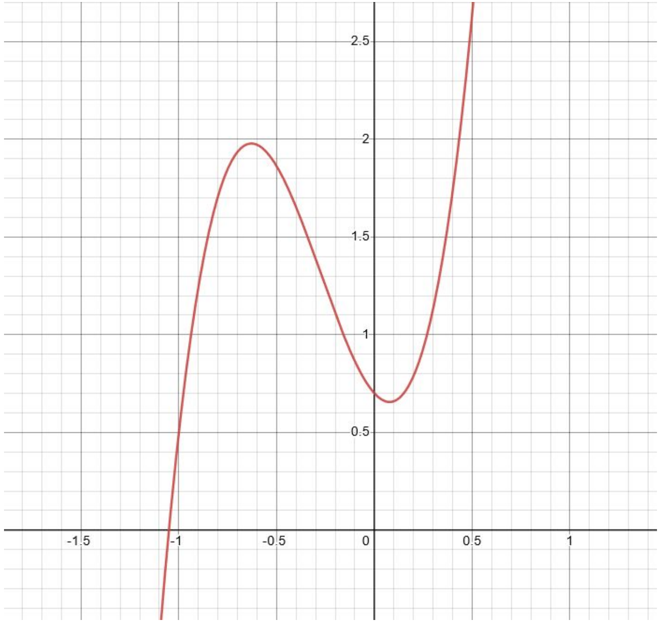
DATA

```
   Competition PlayerType   Made Transition TwoLegged   ShotType        Movement Angle Distance
1          U14          F   Made          1         1 above head              no 73.83     0.73
2          U14          F Missed          0         1      layup              no 28.13     1.02
3          U14          F Missed          0         1 above head              no 51.88     7.22
4          U14          F Missed          1         1 above head              no 80.84     3.64
5          U14          F Missed          0         1 above head              no 30.89     7.20
6          U14          F Missed          0         1      other              no 63.96     1.07
7          U14          F   Made          1         0 above head              no 49.70     0.56
8          U14          F   Made          1         0      layup dribble or cut 64.48     0.59
9          U14          F Missed          0         0      layup dribble or cut 62.56     0.91
10         U14          F   Made          0         0      layup dribble or cut 39.79     0.75
11         U14          F   Made          1         1      layup              no 59.97     0.62
12         U14          F   Made          0         1 above head              no 46.66     7.51
13         U14          F Missed          0         0 above head              no  2.87     2.14
14         U14          G Missed          1         0      layup dribble or cut 64.48     0.59
15         U14          G Missed          0         1      other              no  7.81     7.49
```

LEARNING ALGORITHM

MODEL

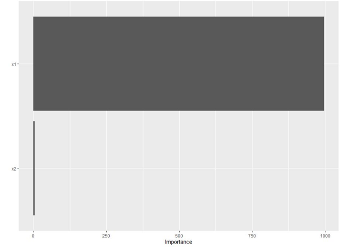
Example f



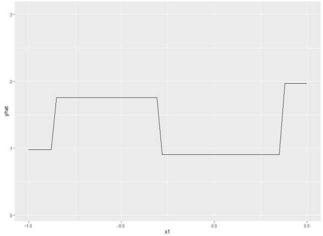
$$y = 7.5x^3 + 6.2x^2 - 1.1x + 0.7$$

Common types of global explanation:

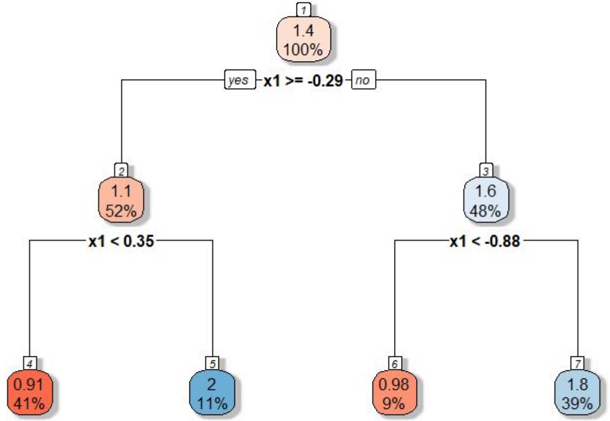
Feature importance



Partial dependence plot

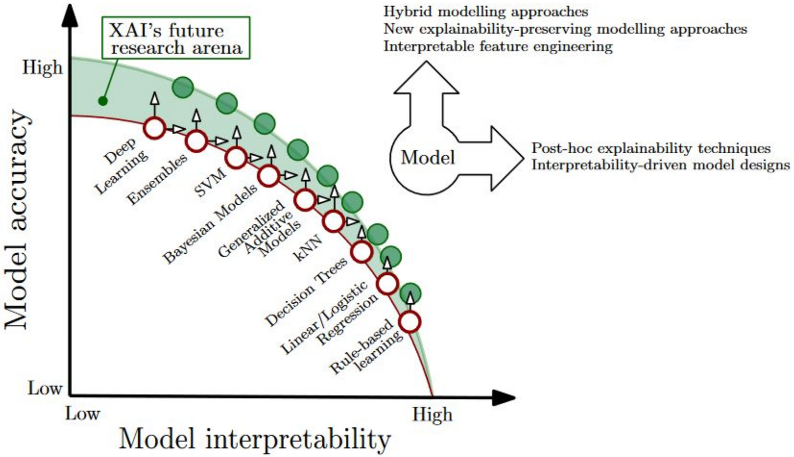


Surrogate model



Explaining predictive models

- View A: There is an Accuracy-interpretability Tradeoff and therefore a strong need for explainable AI.
- View B: The need for explainable AI is exaggerated:
  - It is a myth that there is necessarily a trade-off between accuracy and interpretability.
  - XAI provided explanations that are not faithful to the model,
  - Explanations often do not make sense, or do not provide enough detail to understand what the black box is doing.
  - Black box models are often not compatible with situations where information outside the database needs to be combined with a risk assessment.
  - Black box models with explanations can lead to an overly complicated decision pathway.

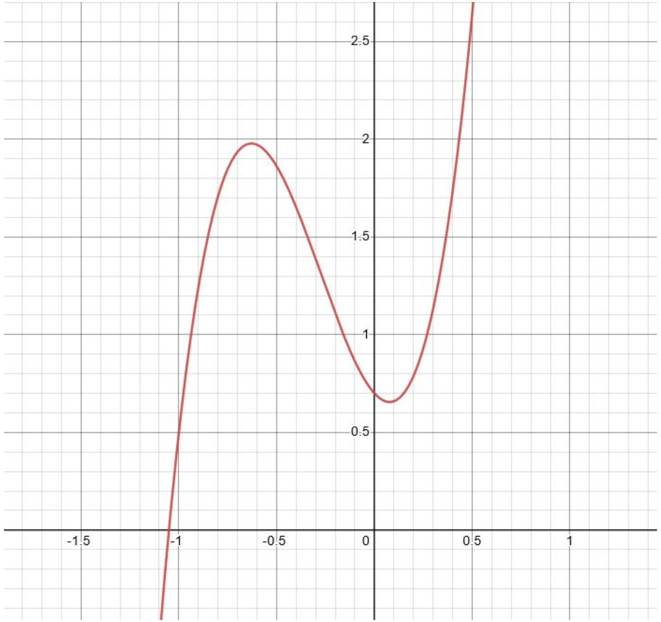

XAI's future research arena
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs
High
Model accuracy
Low
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Low
High
Model interpretability

Arrieta et al. (2020): **Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI**

Rudin (2019): **Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead**

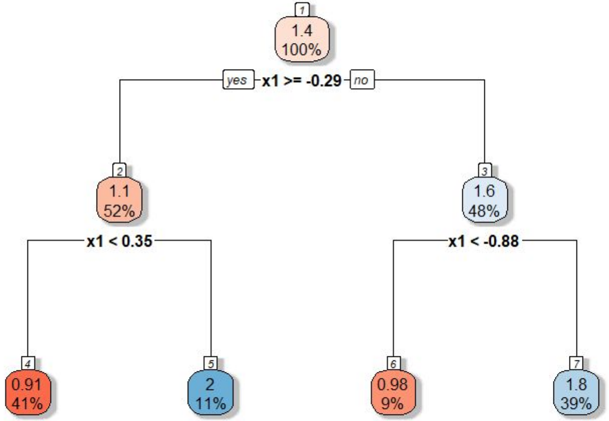
My take

- If explainability is not one of the objectives, don't pay any attention to it.
- If a "white box" model performs well-enough, use it!
- In theory, you can replace any deep neural network with a linear regression with transformed features, however, that can (a) take a lot of time and (b) your model is as most as interpretable as your features.
- In practice, there is always a tradeoff between accuracy and interpretability.
- We overestimate the interpretability of "white box" models.



$y = 7.5x^3 + 6.2x^2 - 1.1x + 0.7$

How white-box are white-box models?



Common types of local explanation and model-agnostic implementations:

- Feature importance (LIME, Shapley value-based),
- Counterfactual explanations,
- Anchors.

Molnar (2000): **Interpretable machine learning**

Local Interpretable Model-Agnostic Explanations (LIME)

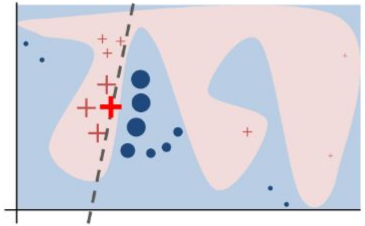
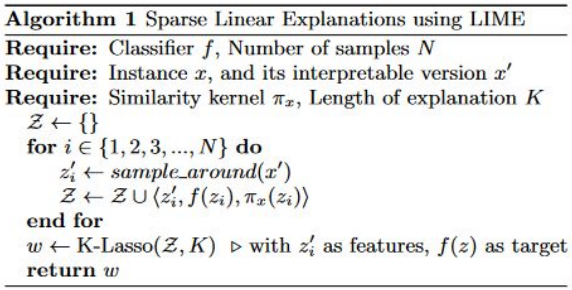
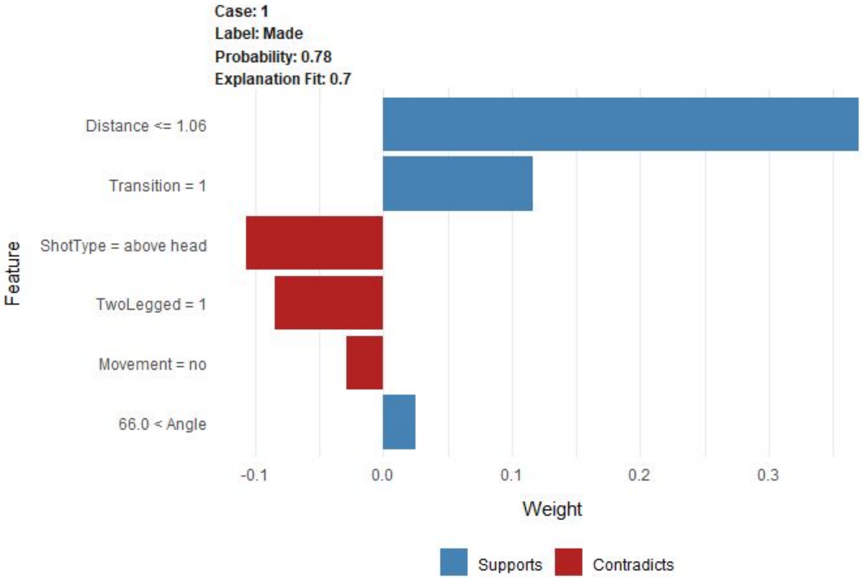


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.



```
Algorithm 1 Sparse Linear Explanations using LIME
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its interpretable version $x'$
Require: Similarity kernel $\pi_x$, Length of explanation $K$
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ do
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$  ▷ with $z'_i$ as features, $f(z)$ as target
  return $w$
```

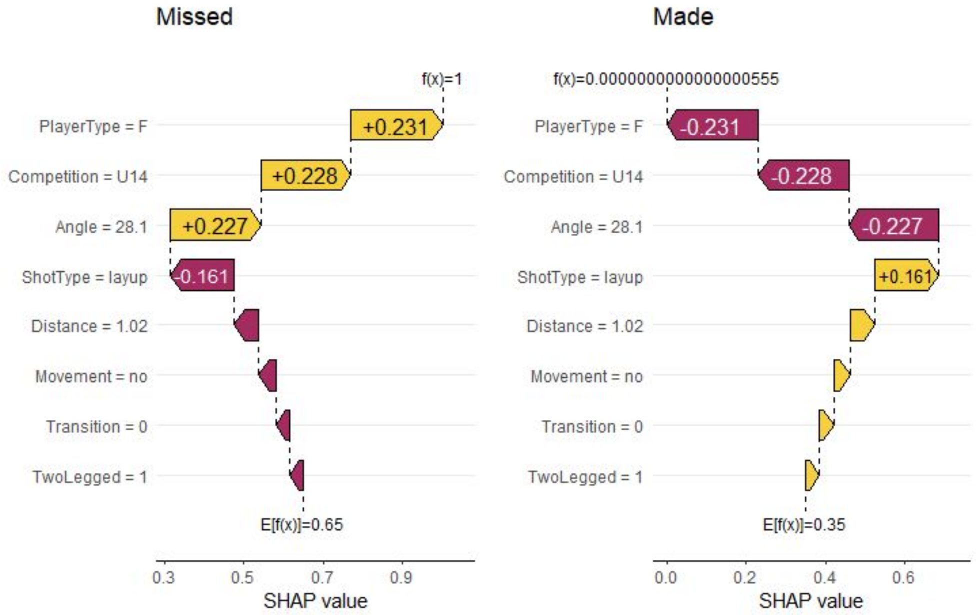


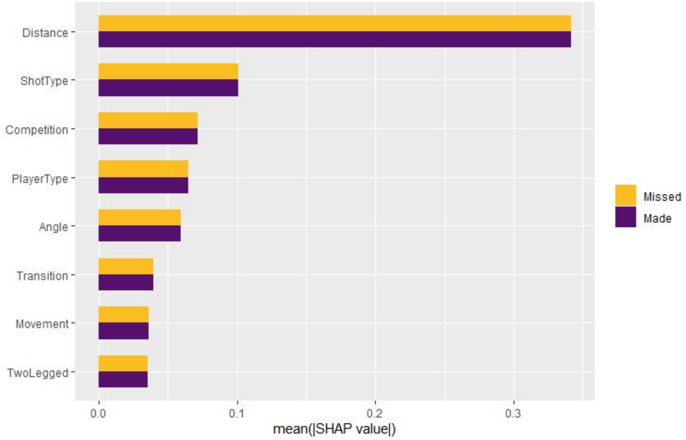
Ribeiro et al. (2016): **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**

Shapley value-based explanation & SHAP

Štrumbelj & Kononenko (2010): **An Efficient Explanation of Individual Classifications using Game Theory**

Lundberg & Lee (2017): **A unified approach to interpreting model predictions**





$$Sh_i(v) = \sum_{S \subseteq N \setminus \{i\}, s=|S|} \frac{(n-s-1)!s!}{n!} (v(S \cup \{i\}) - v(S)), \quad i = 1, \dots, n.$$

$$\varphi_i(\Delta) = \frac{1}{n!} \sum_{O \in \pi(N)} \left(\Delta(Pre^i(O) \cup \{i\}) - \Delta(Pre^i(O)) \right), \quad i = 1, \dots, n.$$

Counterfactual explanations

Example instance

```
  Competition PlayerType Made Transition TwoLegged   ShotType Movement   Angle  Distance P.Missed P.Made
          U14          F Made          1         1 above head       no   73.83      0.73    0.240  0.760
```

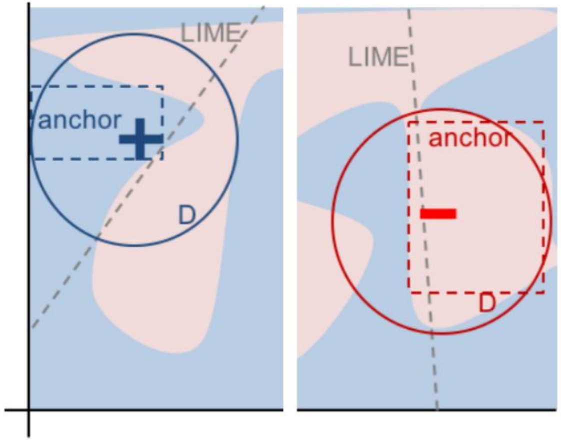
Counterfactuals with lowest P(Made |...)

```
  Competition PlayerType Made Transition TwoLegged   ShotType Movement    Angle Distance P.Missed P.Made
1         U14          F Made          0         1 above head       no 47.91121 7.328654    0.934  0.066
2         U16          F Made          0         1 above head       no 23.76989 7.328654    0.902  0.098
3         U14          F Made          0         1 above head       no 73.83000 4.628888    0.926  0.074
4         U14          F Made          0         1 above head       no 79.01073 4.722931    0.916  0.084
5         U14          F Made          0         1 above head       no 73.83000 5.546916    0.924  0.076
```

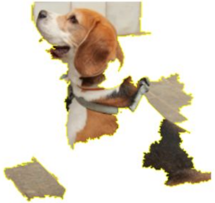
Guidotti et al. (2022): **Counterfactual explanations and how to find them: literature review and benchmarking**

Anchors (Scoped Rules)

Ribeiro et al. (2019): **Anchors: High Precision Model-Agnostic Explanations**



(a) Original image

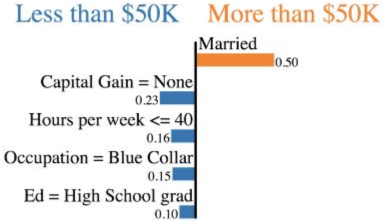


(b) Anchor for "beagle"

$28 < \text{Age} \leq 37$
Workclass = Private
Education = High School grad
Marital Status = Married
Occupation = Blue-Collar
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = Low
Hours per week ≤ 40.00
Country = United-States

$P(\textbf{Salary} > \$50\textbf{K}) = 0.57$

(a) Instance and prediction

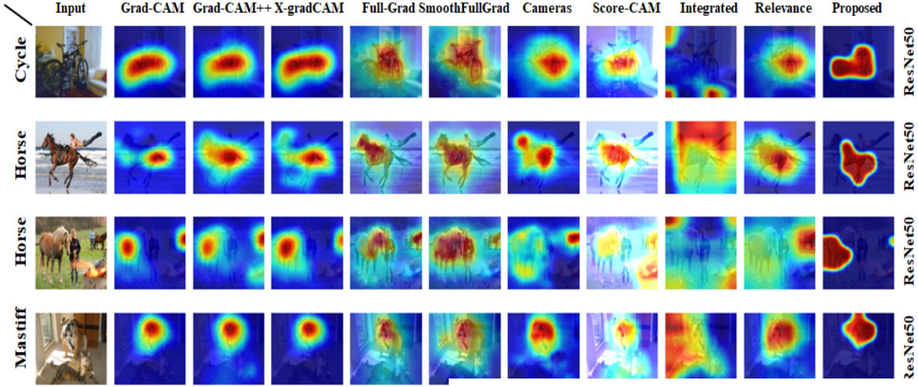


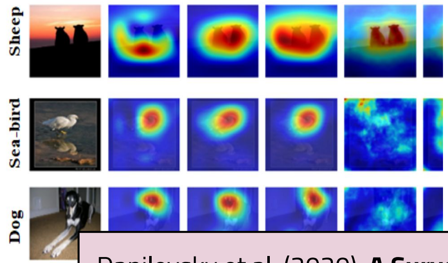
(b) LIME explanation

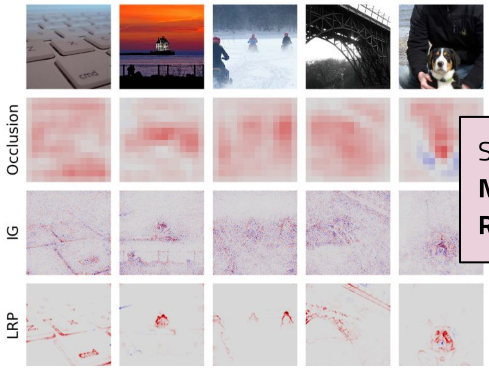
IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** $28 < \text{Age} \leq 37$
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

(c) An *anchor* explanation

Binary (localization) and non-binary (heatmap) feature importance (saliency)

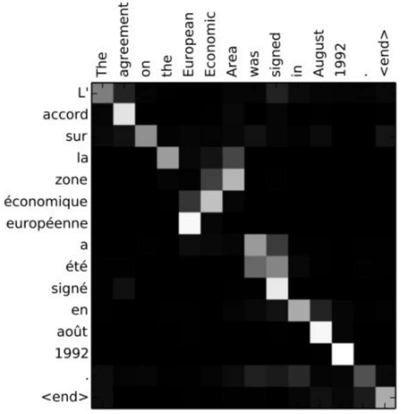




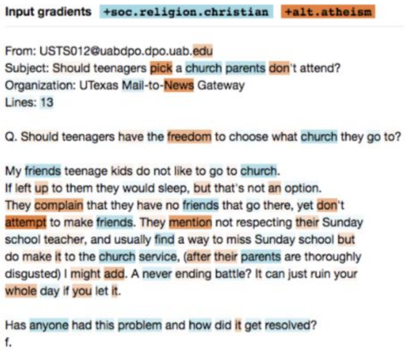


Samek et al. (2021): **Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications**

Danilevsky et al. (2020): **A Survey of the State of Explainable AI for Natural Language Processing**



(a) Saliency heatmap (Bahdanau et al., 2015)



Input gradients `+soc.religion.christian` `+alt.atheism`

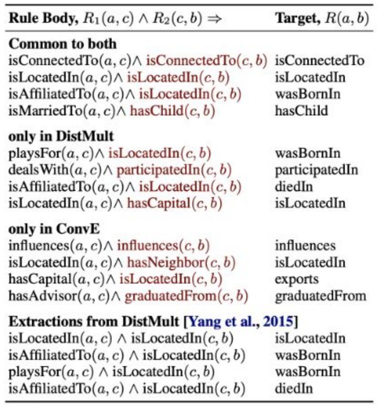
From: USTS012@uabdpo.dpo.uab.edu
Subject: Should teenagers pick a church parents don't attend?
Organization: UTexas Mail-to-News Gateway
Lines: 13

Q. Should teenagers have the freedom to choose what church they go to?

My friends teenage kids do not like to go to church.
If left up to them they would sleep, but that's not an option.
They complain that they have no friends that go there, yet don't attempt to make friends. They mention not respecting their Sunday school teacher, and usually find a way to miss Sunday school but do make it to the church service, (after their parents are thoroughly disgusted) I might add. A never ending battle? It can just ruin your whole day if you let it.

Has anyone had this problem and how did it get resolved?
f.

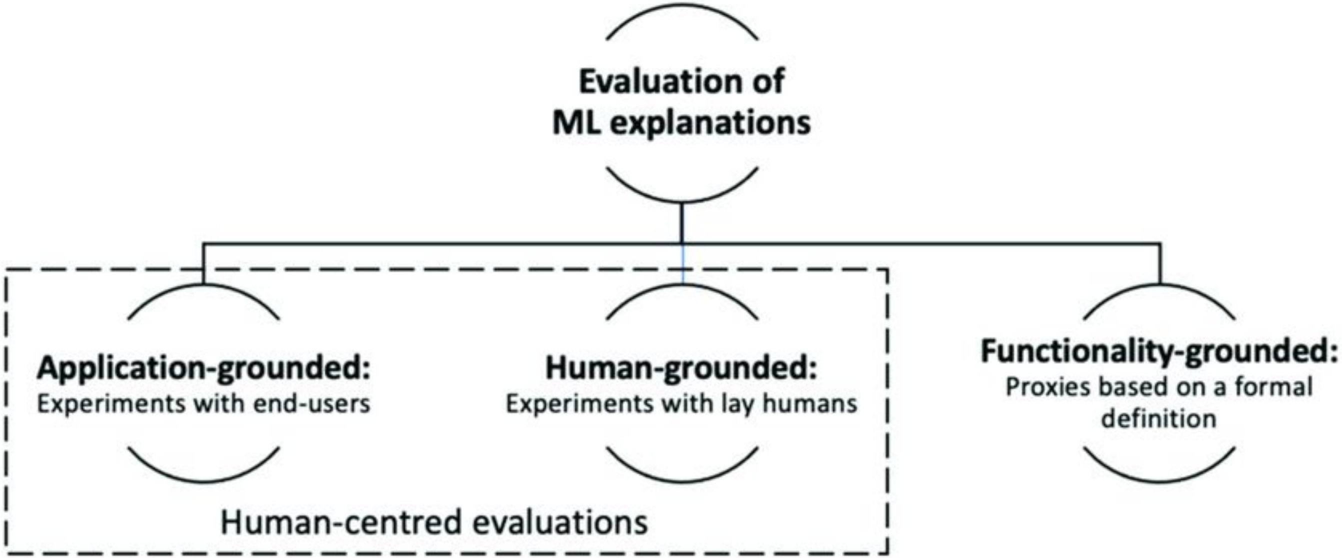
(b) Saliency highlighting (Mullenbach et al., 2018)



Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target,** $R(a,b)$
Common to both	
isConnectedTo$(a,c) \wedge$ isConnectedTo(c,b)	isConnectedTo
isLocatedIn$(a,c) \wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
isMarriedTo$(a,c) \wedge$ hasChild(c,b)	hasChild
only in DistMult	
playsFor$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
dealsWith$(a,c) \wedge$ participatedIn(c,b)	participatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	diedIn
isLocatedIn$(a,c) \wedge$ hasCapital(c,b)	isLocatedIn
only in ConvE	
influences$(a,c) \wedge$ influences(c,b)	influences
isLocatedIn$(a,c) \wedge$ hasNeighbor(c,b)	isLocatedIn
hasCapital$(a,c) \wedge$ isLocatedIn(c,b)	exports
hasAdvisor$(a,c) \wedge$ graduatedFrom(c,b)	graduatedFrom
Extractions from DistMult [Yang et al., 2015]	
isLocatedIn$(a,c) \wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
playsFor$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	diedIn

(c) Raw declarative rules (Pezeshkpour et al., 2019b)

Evaluating explainable AI



Evaluation of ML explanations

Application-grounded: Experiments with end-users

Human-grounded: Experiments with lay humans

Functionality-grounded: Proxies based on a formal definition

Human-centred evaluations

Doshi-Velez & Kim (2017): **Towards a rigorous science of interpretable machine learning**

Evaluating explainable AI

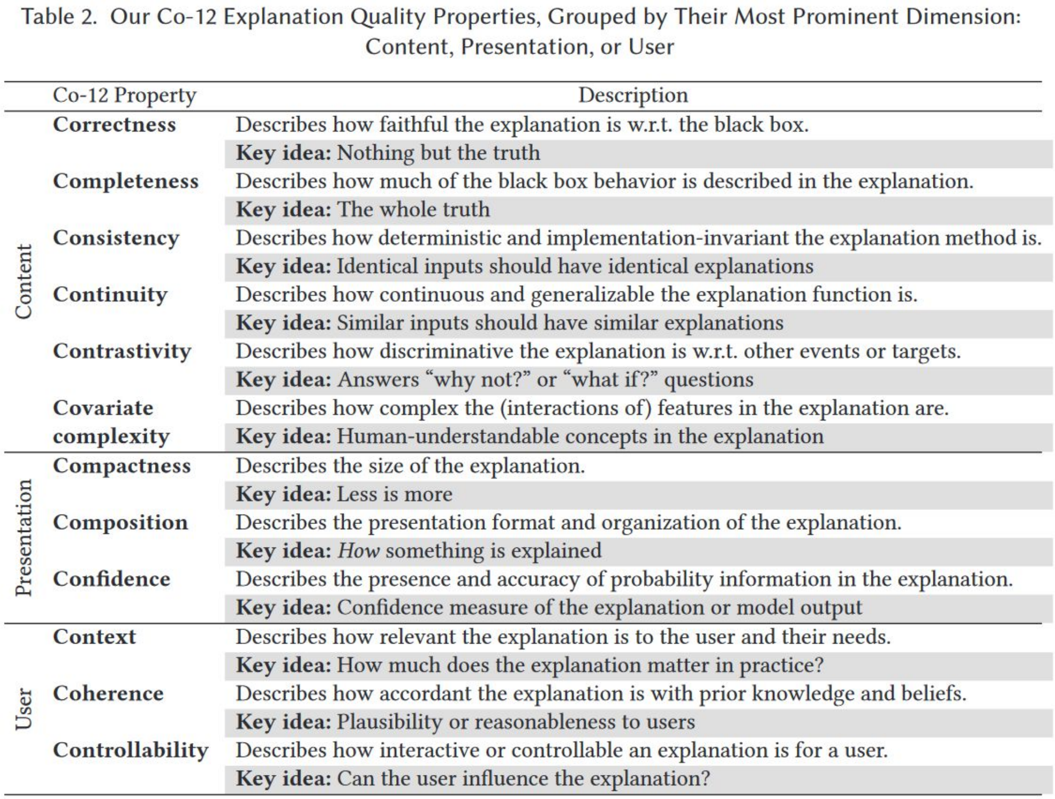
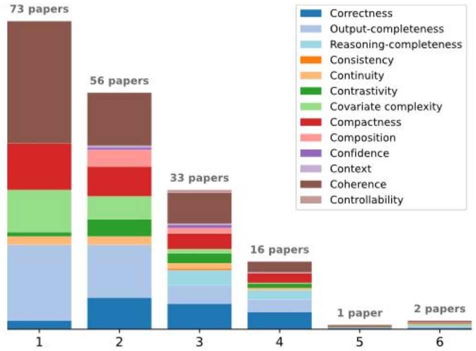


Table 2. Our Co-12 Explanation Quality Properties, Grouped by Their Most Prominent Dimension: Content, Presentation, or User

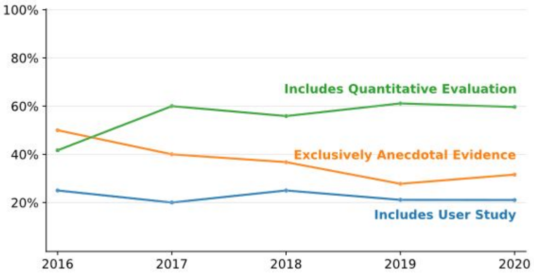
	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic 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?



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

Nauta et al. (2023): **From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AID**

Evaluating explainable AI



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

Nauta et al. (2023): **From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AID**

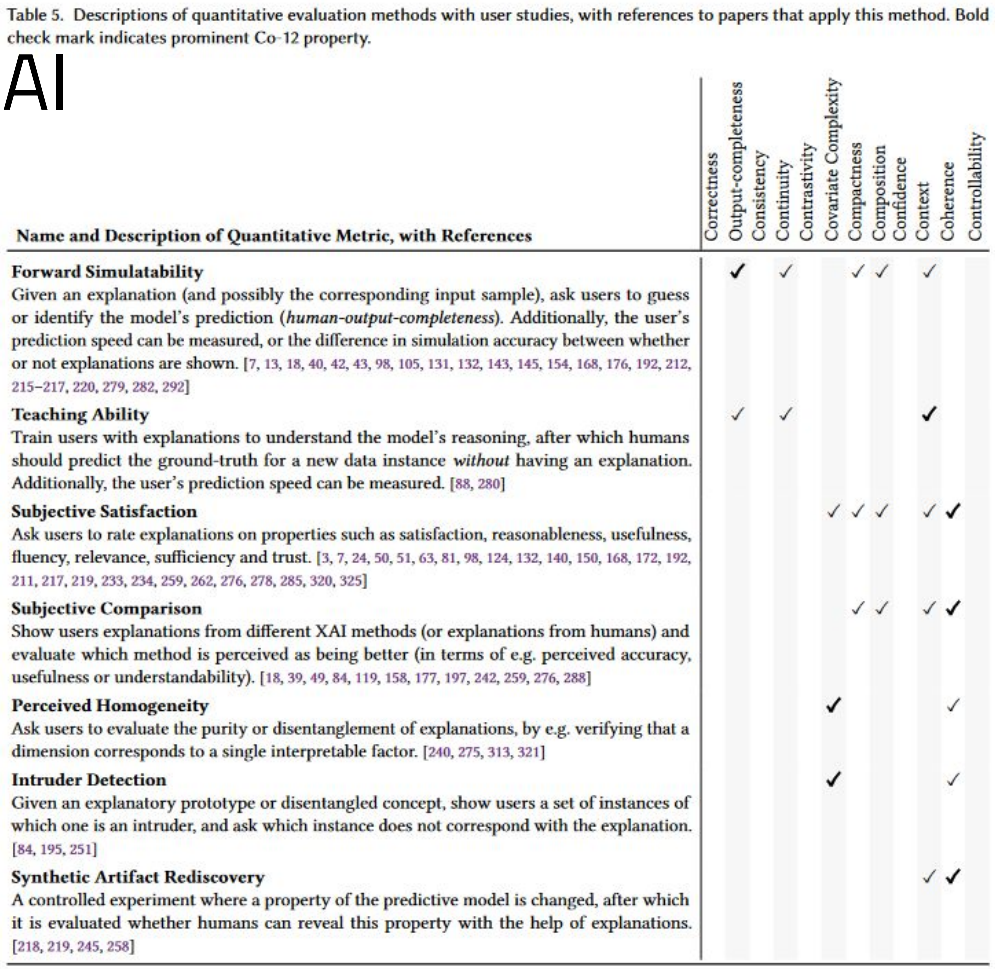


Table 5. Descriptions of quantitative evaluation methods with user studies, with references to papers that apply this method. Bold check mark indicates prominent Co-12 property.

Name and Description of Quantitative Metric, with References	Correctness	Output-completeness	Consistency	Continuity	Contrastivity	Covariate Complexity	Compactness	Composition	Confidence	Context	Coherence	Controllability
Forward Simulatability Given an explanation (and possibly the corresponding input sample), ask users to guess or identify the model's prediction (*human-output-completeness*). Additionally, the user's prediction speed can be measured, or the difference in simulation accuracy between whether or not explanations are shown. [7, 13, 18, 40, 42, 43, 98, 105, 131, 132, 143, 145, 154, 168, 176, 192, 212, 215–217, 220, 279, 282, 292]		**✓**		✓			✓	✓		✓		
Teaching Ability Train users with explanations to understand the model's reasoning, after which humans should predict the ground-truth for a new data instance *without* having an explanation. Additionally, the user's prediction speed can be measured. [88, 280]		✓		✓						**✓**		
Subjective Satisfaction Ask users to rate explanations on properties such as satisfaction, reasonableness, usefulness, fluency, relevance, sufficiency and trust. [3, 7, 24, 50, 51, 63, 81, 98, 124, 132, 140, 150, 168, 172, 192, 211, 217, 219, 233, 234, 259, 262, 276, 278, 285, 320, 325]						✓	✓	✓		✓	**✓**	
Subjective Comparison Show users explanations from different XAI methods (or explanations from humans) and evaluate which method is perceived as being better (in terms of e.g. perceived accuracy, usefulness or understandability). [18, 39, 49, 84, 119, 158, 177, 197, 242, 259, 276, 288]							✓	✓		✓	**✓**	
Perceived Homogeneity Ask users to evaluate the purity or disentanglement of explanations, by e.g. verifying that a dimension corresponds to a single interpretable factor. [240, 275, 313, 321]						**✓**					✓	
Intruder Detection Given an explanatory prototype or disentangled concept, show users a set of instances of which one is an intruder, and ask which instance does not correspond with the explanation. [84, 195, 251]						**✓**					✓	
Synthetic Artifact Rediscovery A controlled experiment where a property of the predictive model is changed, after which it is evaluated whether humans can reveal this property with the help of explanations. [218, 219, 245, 258]										✓	**✓**	

Summary

- The field is very young,
- requires interdisciplinary work
  (CS, PSY, HCI, UX),
- currently mostly ignoring the user,
- decent methods exist for the ML practitioner,
- missing a proper legal framework,
- will be interesting to see what role LLMs play,
- lots of opportunities!