A Recommendation Mechanism For Explainable Artificial Intelligence (XAI) Methods

Master Thesis Defence

Presented by: Kaoutar Chennaf
Supervised by: Dr. Sharif Sakr, Dr. Ralf-Detlef Kutsche

September 3rd, 2020 @ Technische Universität Berlin

Technische Universität Berlin
Institut für Softwaretechnik und Theoretische Informatik

Kaoutar CHENNAF

In completion of the
Erasmus Mundus Master in Big Data Management and Analytics (BDMA)

Email: chennaf@campus.tu-berlin.de
https://www.dima.tu-berlin.de

Agenda

1. XAI in a Glimpse
2. Interpretability Review
3. User-centric XAI: Proposal
4. User-centric XAI: Empirical Study
5. Conclusion

1.
XAI in a Glimpse

Origin – Meaning – Relevance - Goals

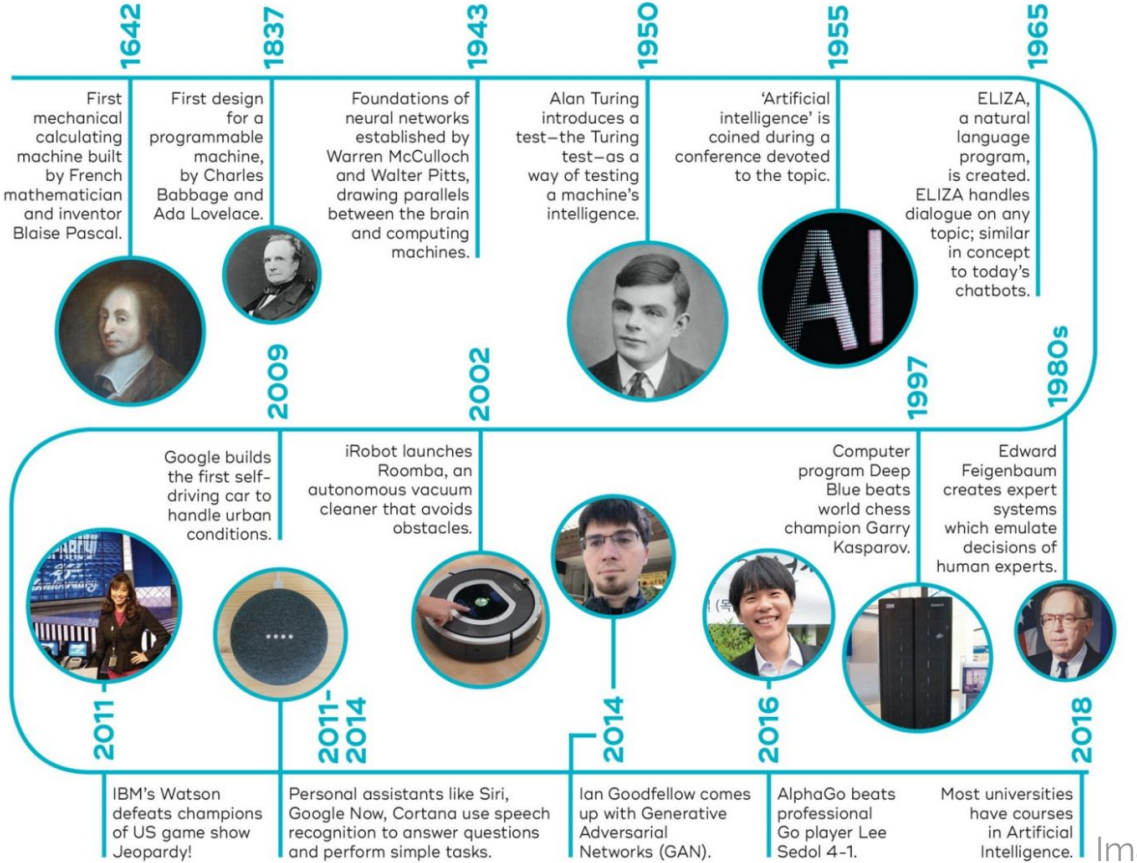


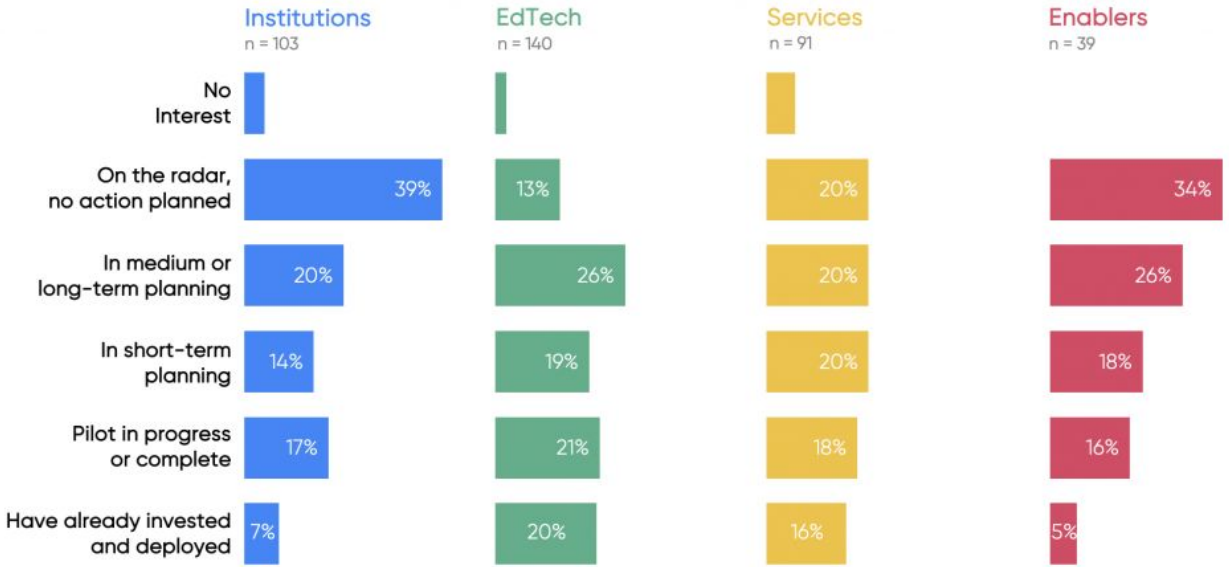
Image source [3]

GLOBAL EDUCATION ARTIFICIAL INTELLIGENCE

Holon IQ

How is adoption of AI progressing?

1 in 10 organizations have invested in and deployed artificial intelligence. 1 in 5 are conducting a pilot, and 1 in 3 have no current plans for AI.



Source: HolonIQ Global Executive Panel, April 2019. n = 377

Image source [2]

Image source [1]

"ALEXA ORDERS ALL THE DOLLS HOUSES"

"Google Home Minis spied on their owners"

"IBM Watson Medical AI Fails"

"Supercomputer Makes Awful Investment Decisions"

"Facebook allowed ads to be targeted to "Jew Haters" "

AI is NOT perfect!

"Facial Recognition Failure In China"

"Google Translate shows gender bias in Turkish-English translation"

"Facebook chatbots shut down after developing their own language"

"Alexa brings the party with her in Germany"

"Amazon's AI-Powered Recruiting Tool is Biased"

"AI World Cup 2018 Predictions Fail"

"Google Allo suggested man in turban emoji as response to a gun emoji"

"SELF DRIVING CAR CAUSES DEATH"

Source [1, 4]

“XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.” David Gunning, DARPA 2017 [5]

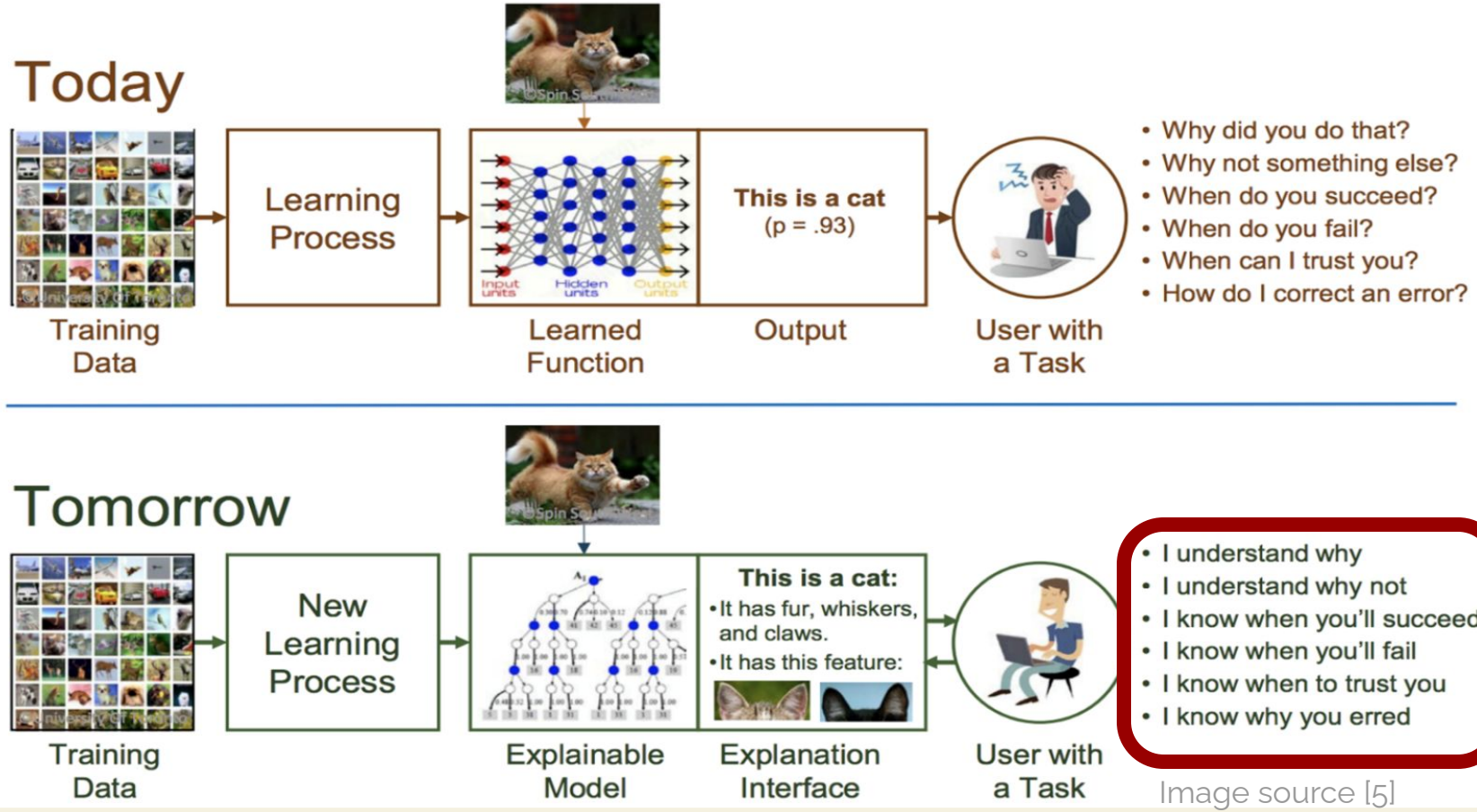


Image source [5]

XAI in a Glimpse

Origin

First coined in the 70's by Lent et al. when describing their system's ability to explain the behavior of AI entities in simulation games

Contributing Fields

- Human Science
- Cognitive Science
- Human-computer Interaction (HCI)

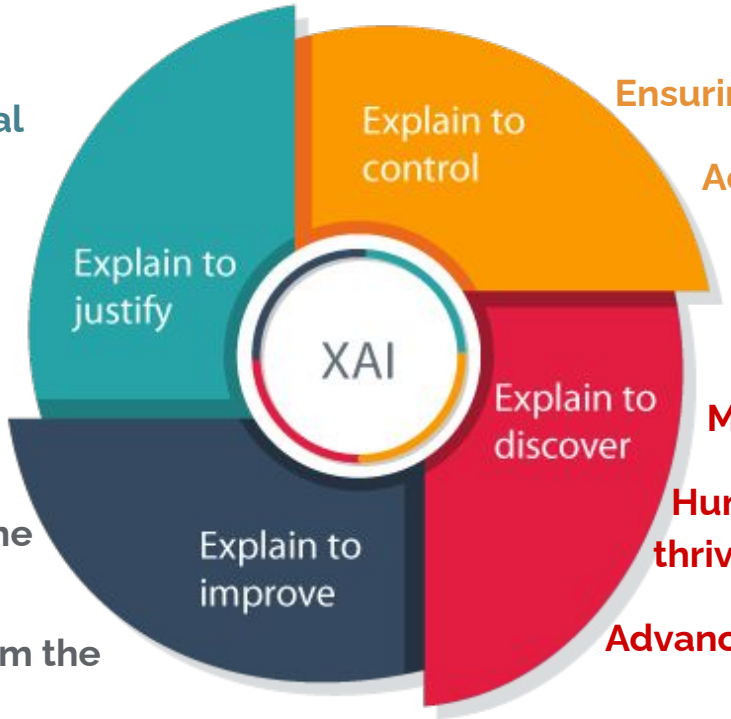
Relevance

- Worldwide investment in AI will triple between 2017 and 2022
- Global revenue from AI is forecasted to reach 105.8 billion USD by 2023
- Growingly significant impact of AI on our society
- Recently imposed regulations and legislations requiring accountability and full transparency

Goals of XAI

Increasing trust and social acceptance

Reconciling decisions



Ensuring safety measures

Adhering to regulations

Managing social interactions

Human curiosity and learning to thrive

Advancing in scientific fields

Advancing in AI and ML

Debugging and auditing of machine learning models

Correcting models and giving them the capacity to evolve

2.
Interpretability Review

Terminology - Taxonomy - Scope - Approaches - Challenges

“The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks. […] The need for interpretability arises from an incompleteness in problem formalization” Doshi-Velez and Kim, 2017 [6]

Terminology

Explanation

There is no such thing as the best explanation for all cases and for all explainees: “One explanation does not fit all”.

Interpretability

Ability to summarize the reasons for system’s behaviour, gain the trust of users, or produce insights about the causes of decisions

Explainability

An explainable model is complete, with the capacity to defend its actions, provide relevant responses to questions, and be audited

Transparency

The degree of understandability of the model. If a model is by itself understandable, it is considered to be transparent.

Descriptive Accuracy

How well does the explanation method objectively captures the learned relationships by the model?

Explanation Methods

Post-hoc interpretability techniques/algorithms used to interpret black-box models

Taxonomy

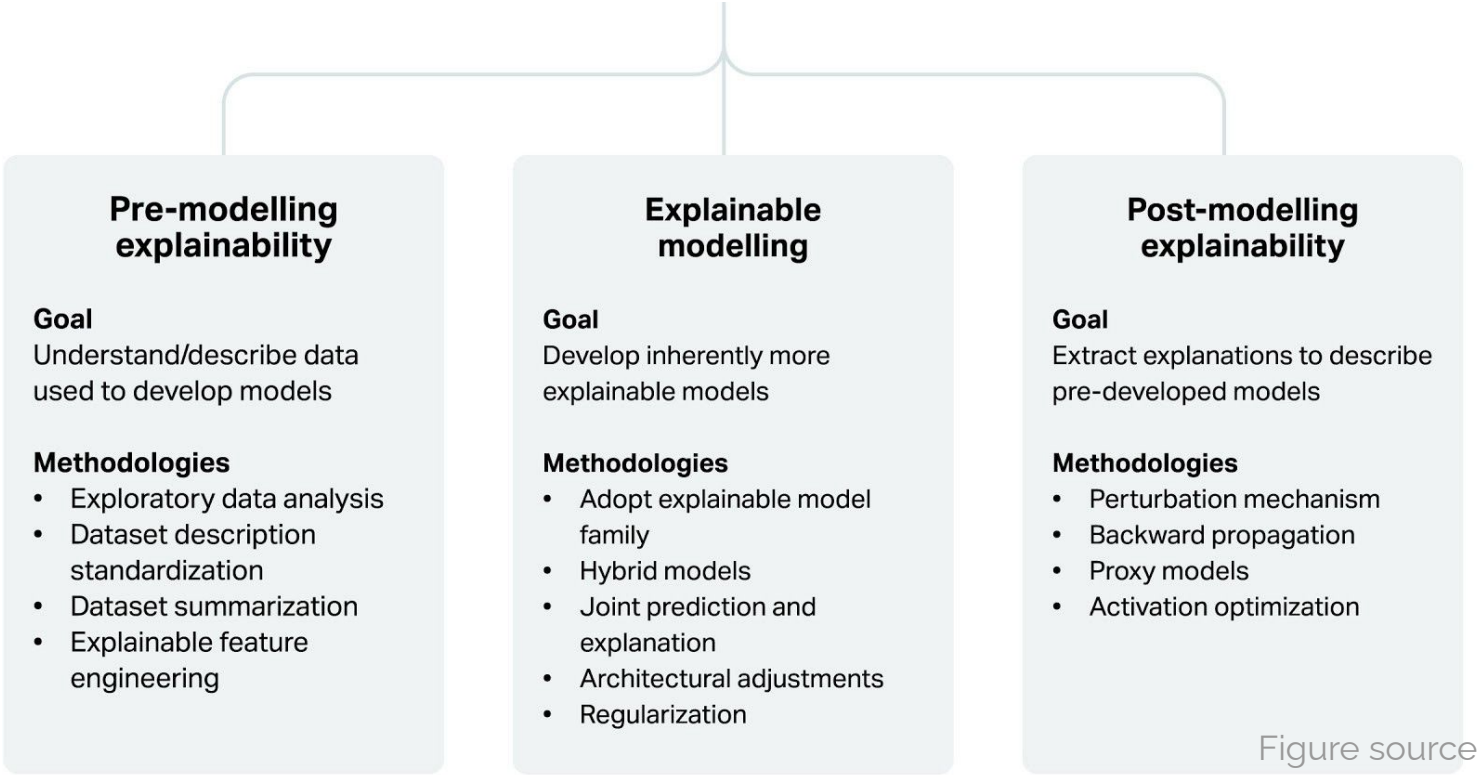


Figure source [8]

Taxonomy

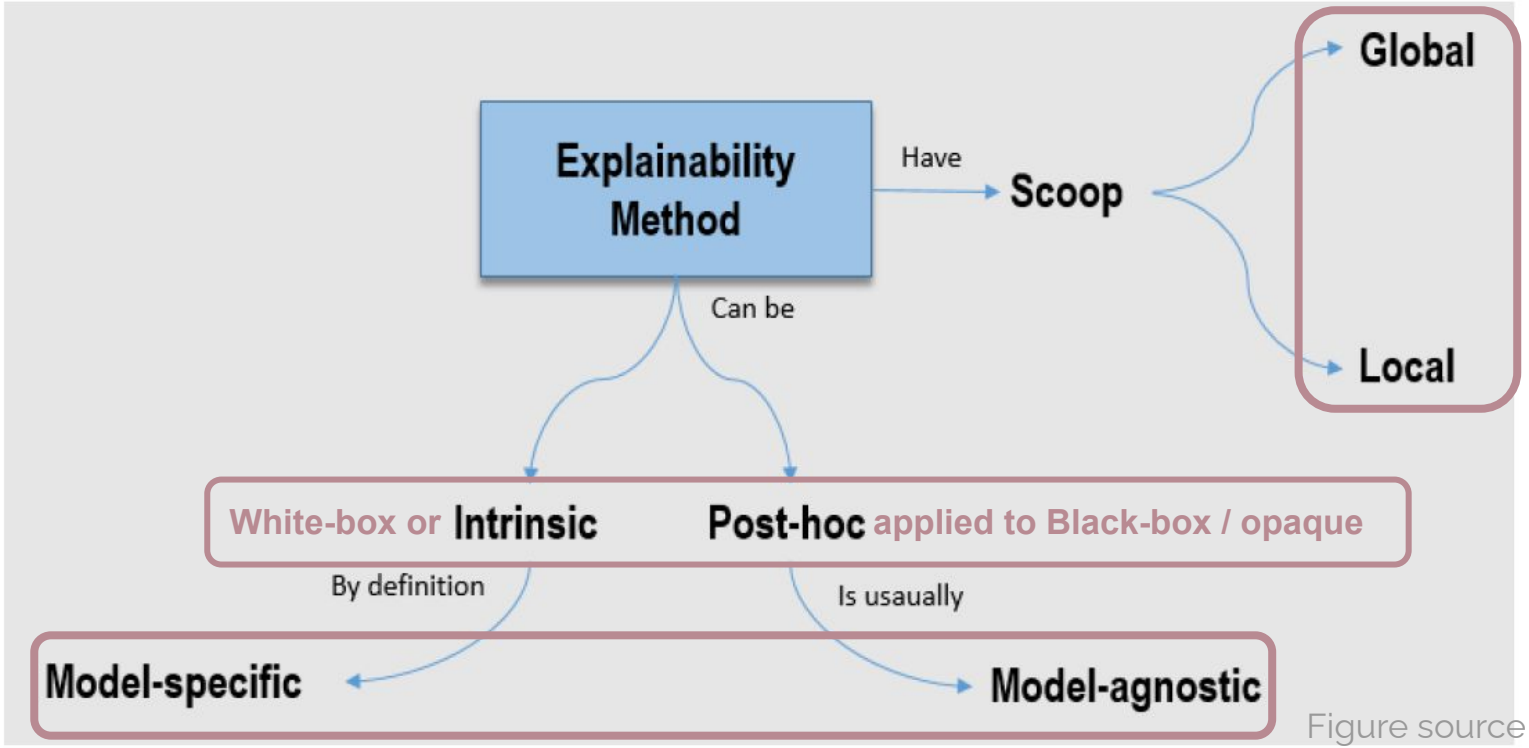


Figure source [7]

Model	Transparent ML Models		
	Simulatability	Decomposability	Algorithmic Transparency
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools

Figure source [9]

Models Classification

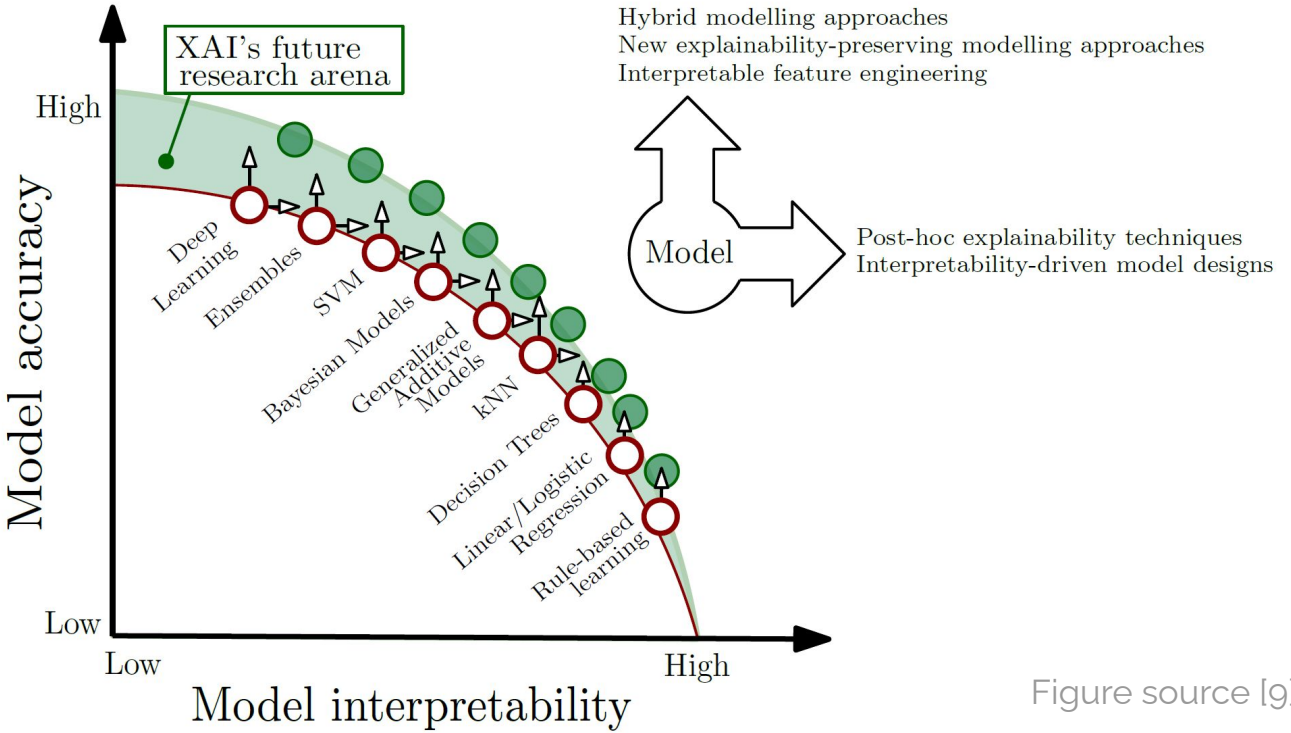


Figure source [9]

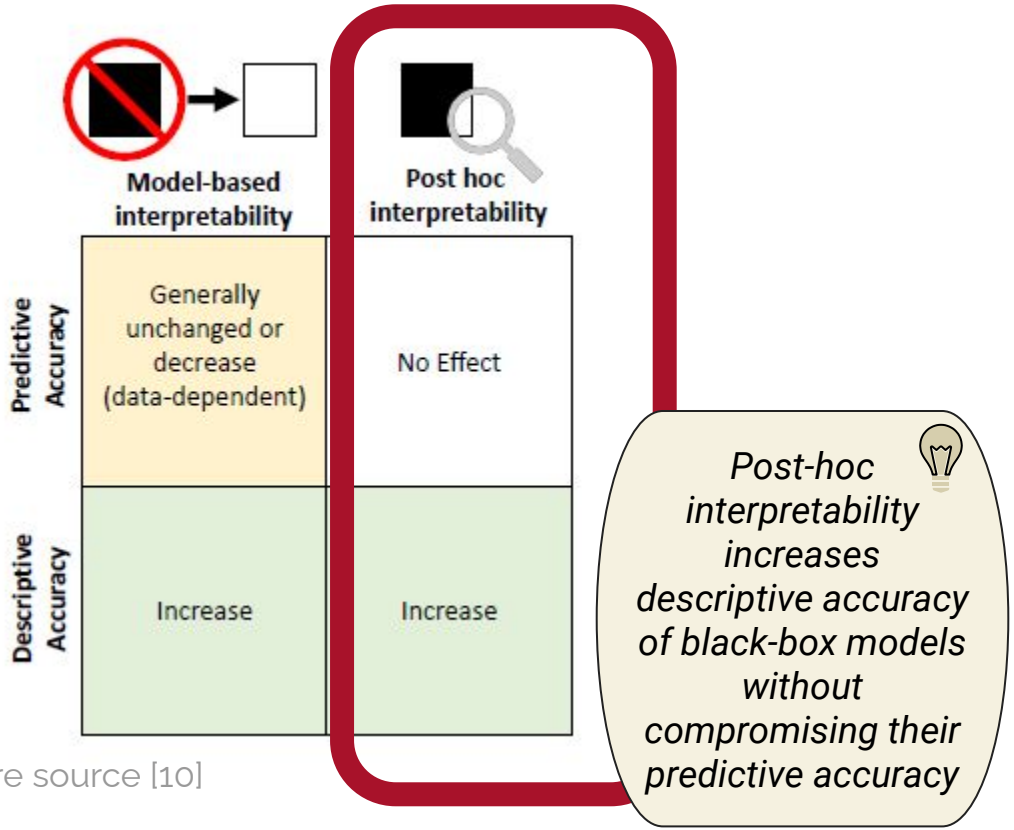


Figure source [10]

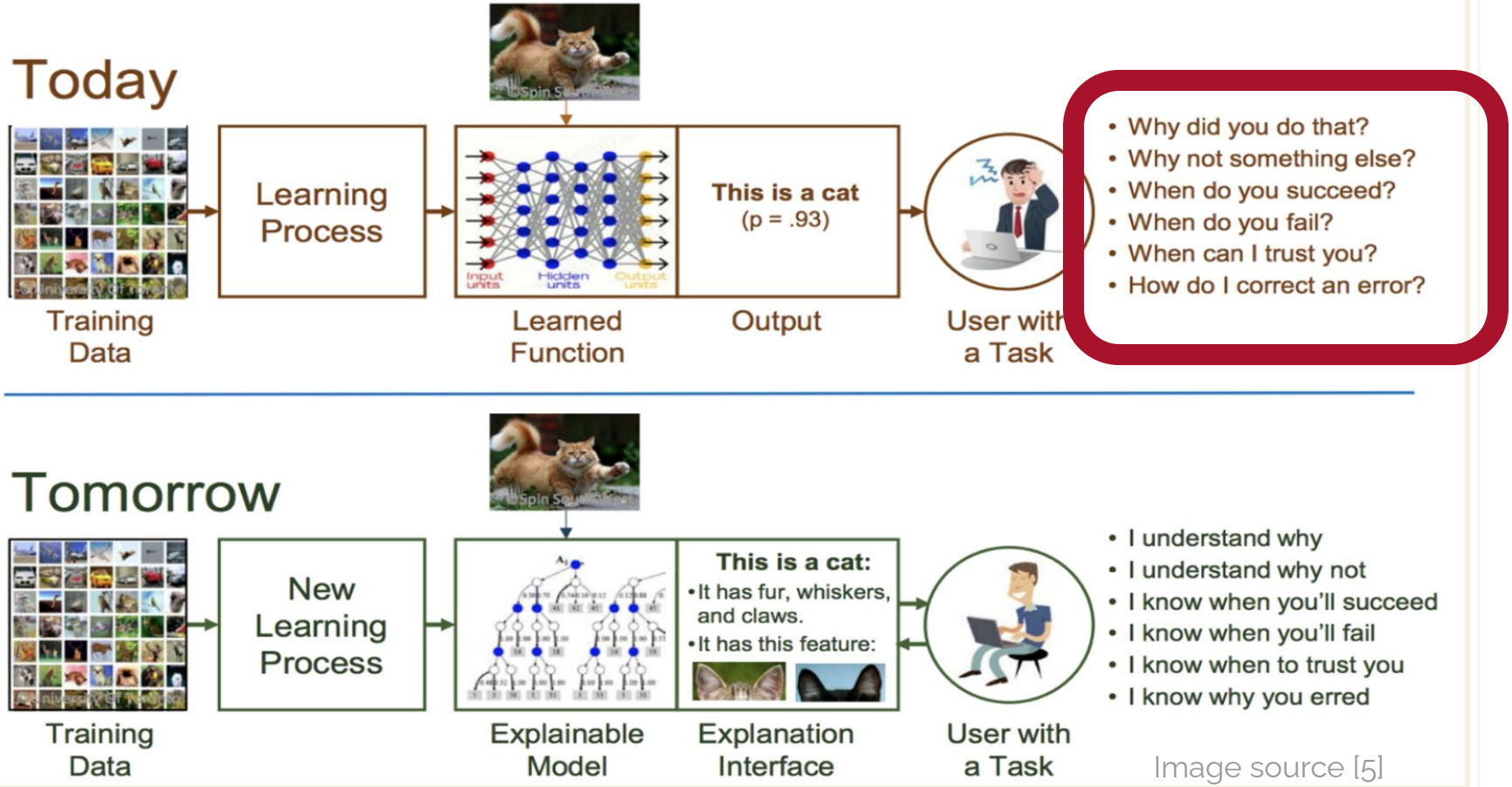


Image source [5]

Approaches

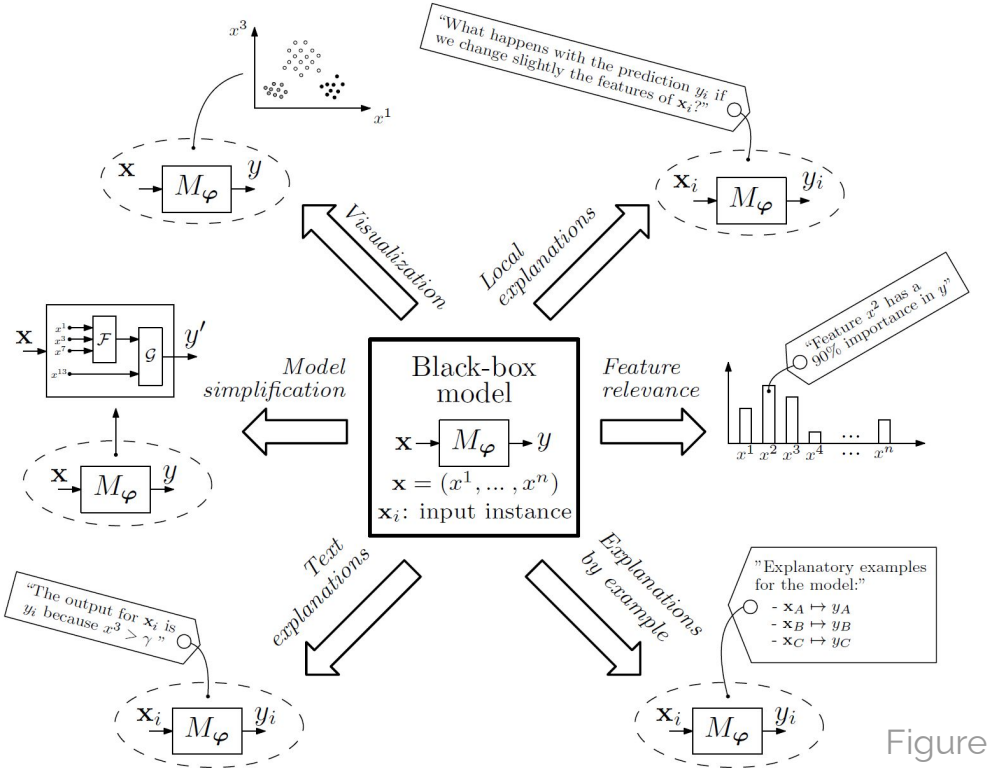


Figure source [9]

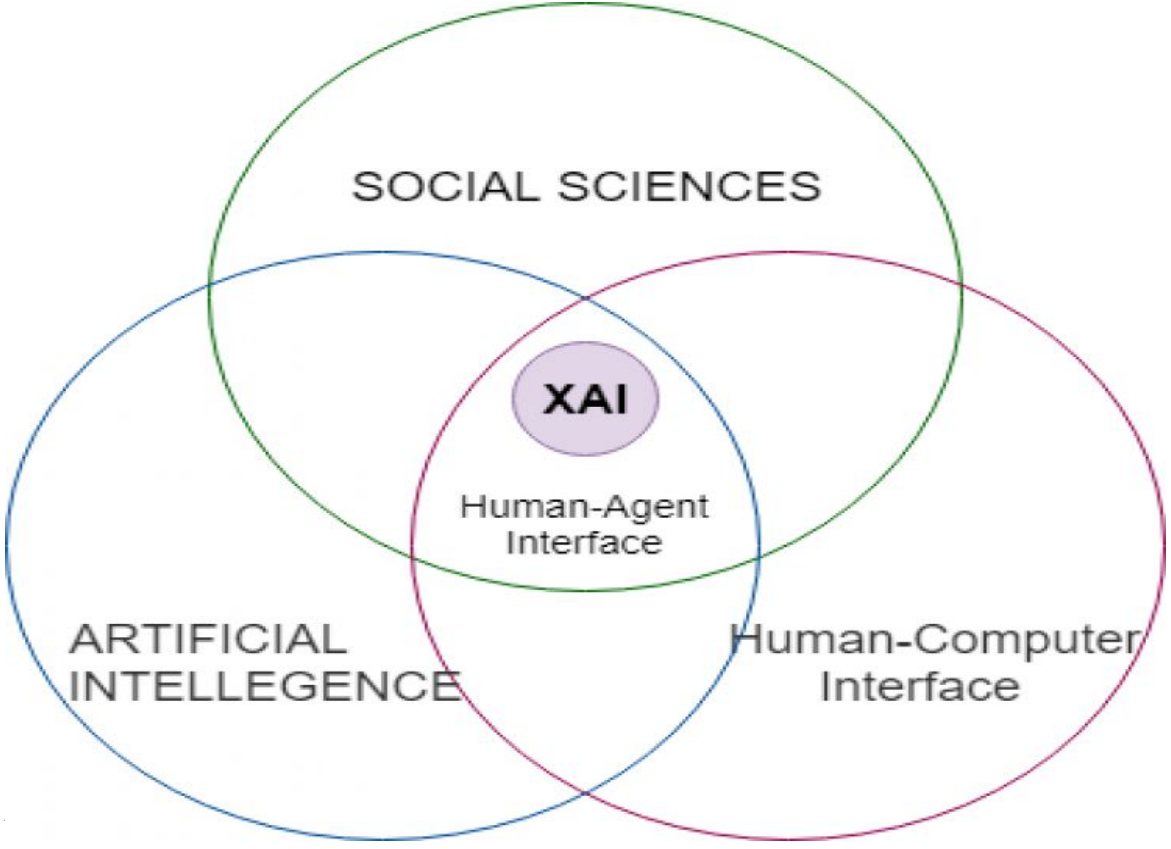
Challenges

- There remains a considerable dichotomy between explanations in ML and in other explanation sciences
- A consensus on an agreed upon definition of explainability or interpretability has not been reached yet.
- There is a growing need for validating, comparing, and evaluating explanation methods. There isn't any standard metric, benchmark or terminology among the community.
- Forging the link between explanations and real-world is still a challenge.
- "All models are wrong, but some are useful".

3.
User-centric XAI: Proposal

Motivation - Related Work - Problem Statement - Lines of Work

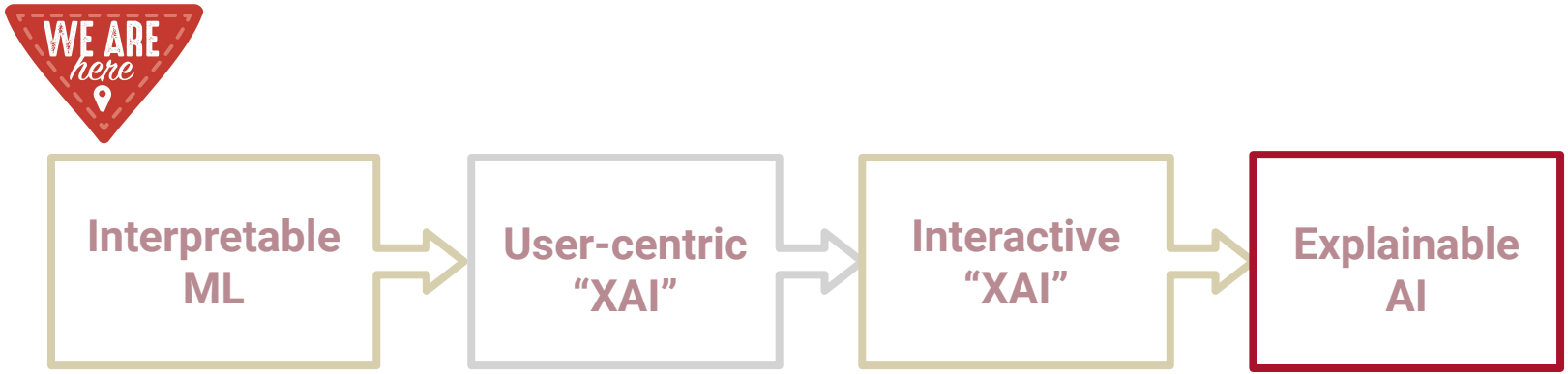
Motivation


SOCIAL SCIENCES
XAI
Human-Agent
Interface
ARTIFICIAL
INTELLEGENCE
Human-Computer
Interface

Motivation

- Explanations are multi-faceted and subjective
- Explanations are a series of interactions between the explainer and explainee
- For XAI to truly reach all its potential, it needs to embrace this nature of explanations

The Path Towards XAI



Current Status

Toolkit	Data Explanations	Directly Interpretable	Local Post-Hoc	Global Post-Hoc	Persona-Specific Explanations	Metrics
AIX360	✓	✓	✓	✓	✓	✓
Alibi [1]			✓			
Skater [7]		✓	✓	✓		
H2O [4]		✓	✓	✓		
InterpretML [6]		✓	✓	✓		
EthicalML-XAI [3]				✓		
DALEX [2]			✓	✓		
tf-explain [8]			✓	✓		
iNNvestigate [5]			✓			

Table source [11]

Current Status

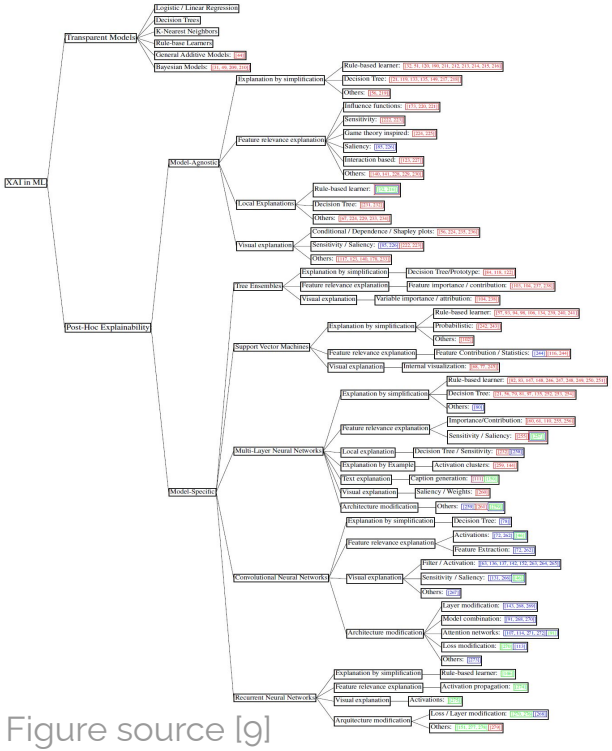
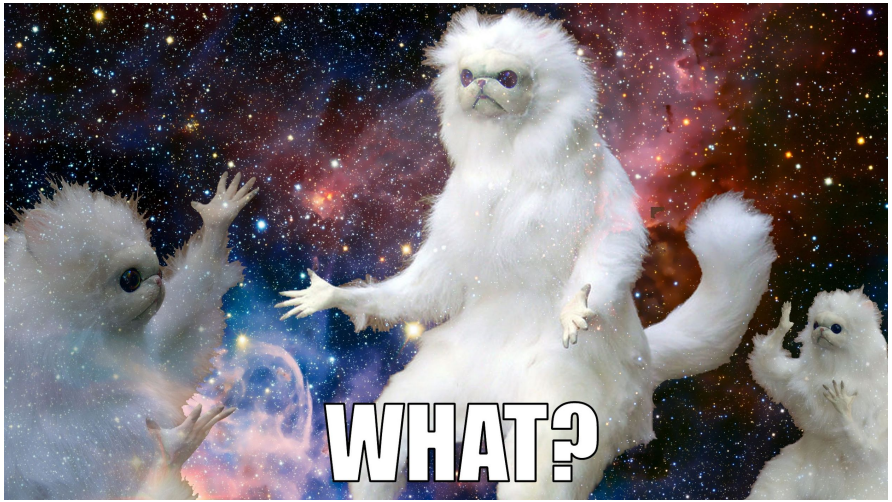


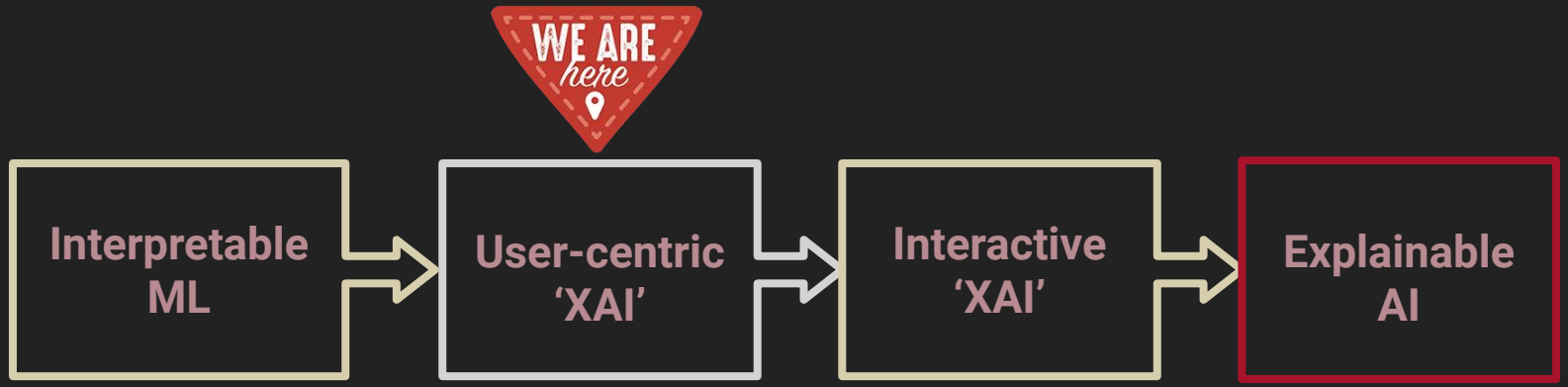
Figure source [9]



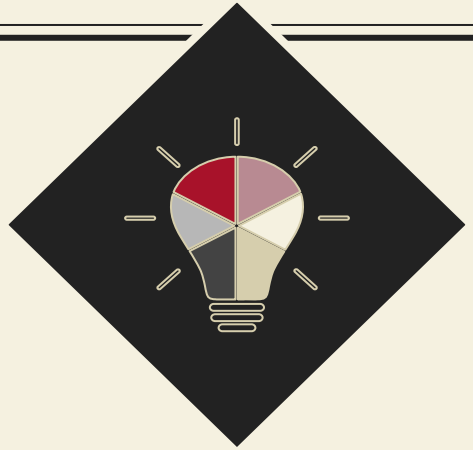
Current Status

- No agreed upon definition of Explainability
- XAI is too static
- Proliferation of explanation methods
- The user is rarely considered in the explanation production process
- Interpretability assessment remains an open challenge

What Do We Need? 😐




KEEP
CALM
AND
DIVIDE &
CONQUER



Categorization

of use cases, users and their goals and needs for XAI, their desired properties in explanations, and the needed assessment metrics for each case.

Related Work

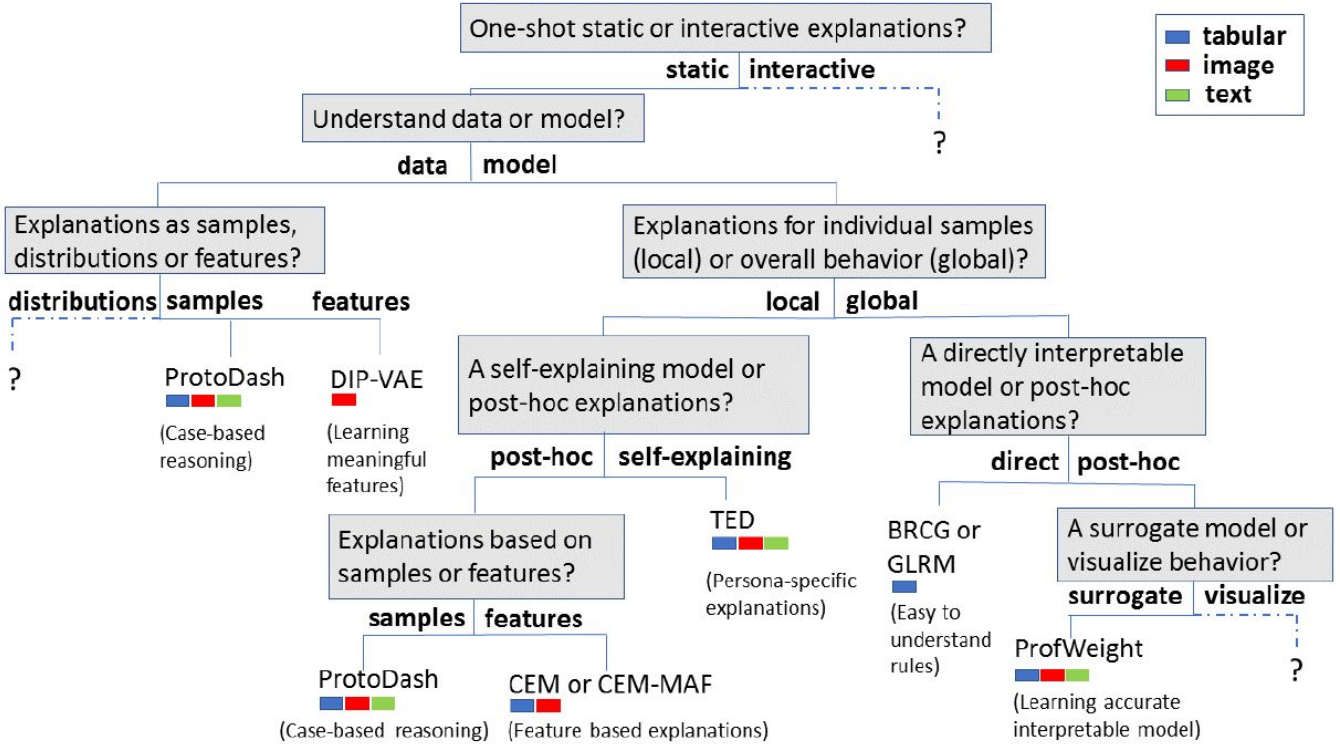


Figure source [11]

Related Work

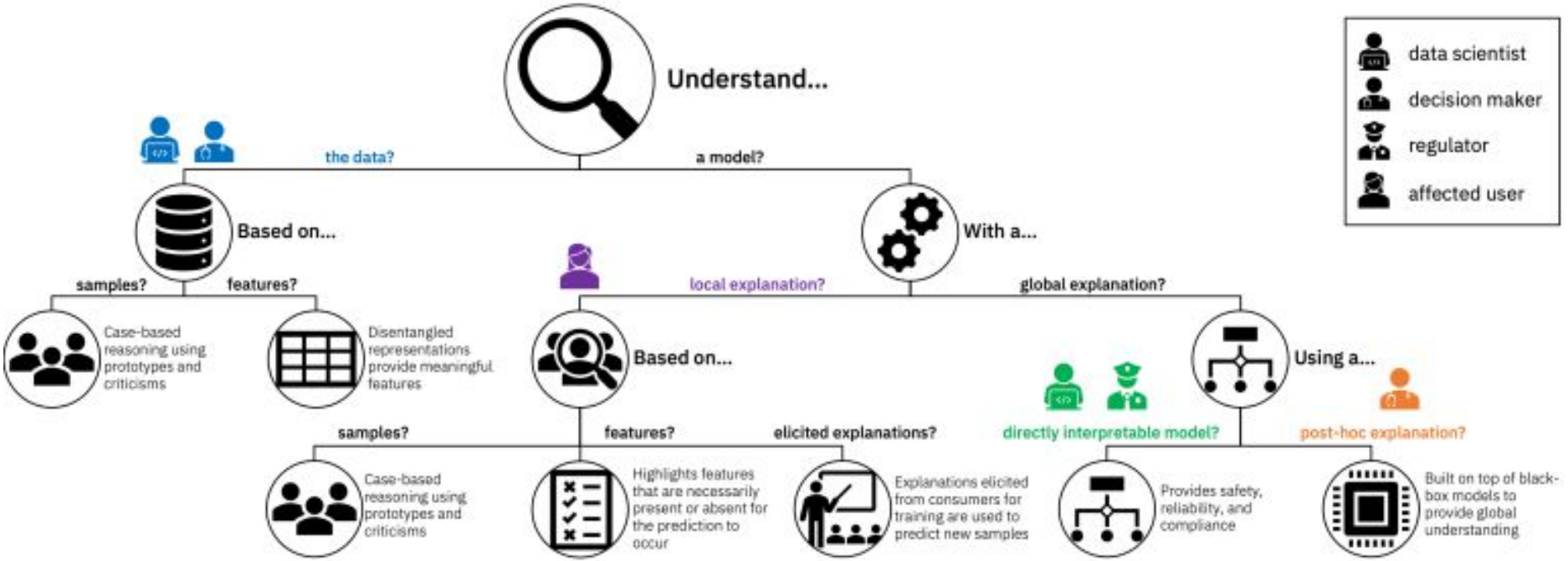


Figure source [11]

Related Work

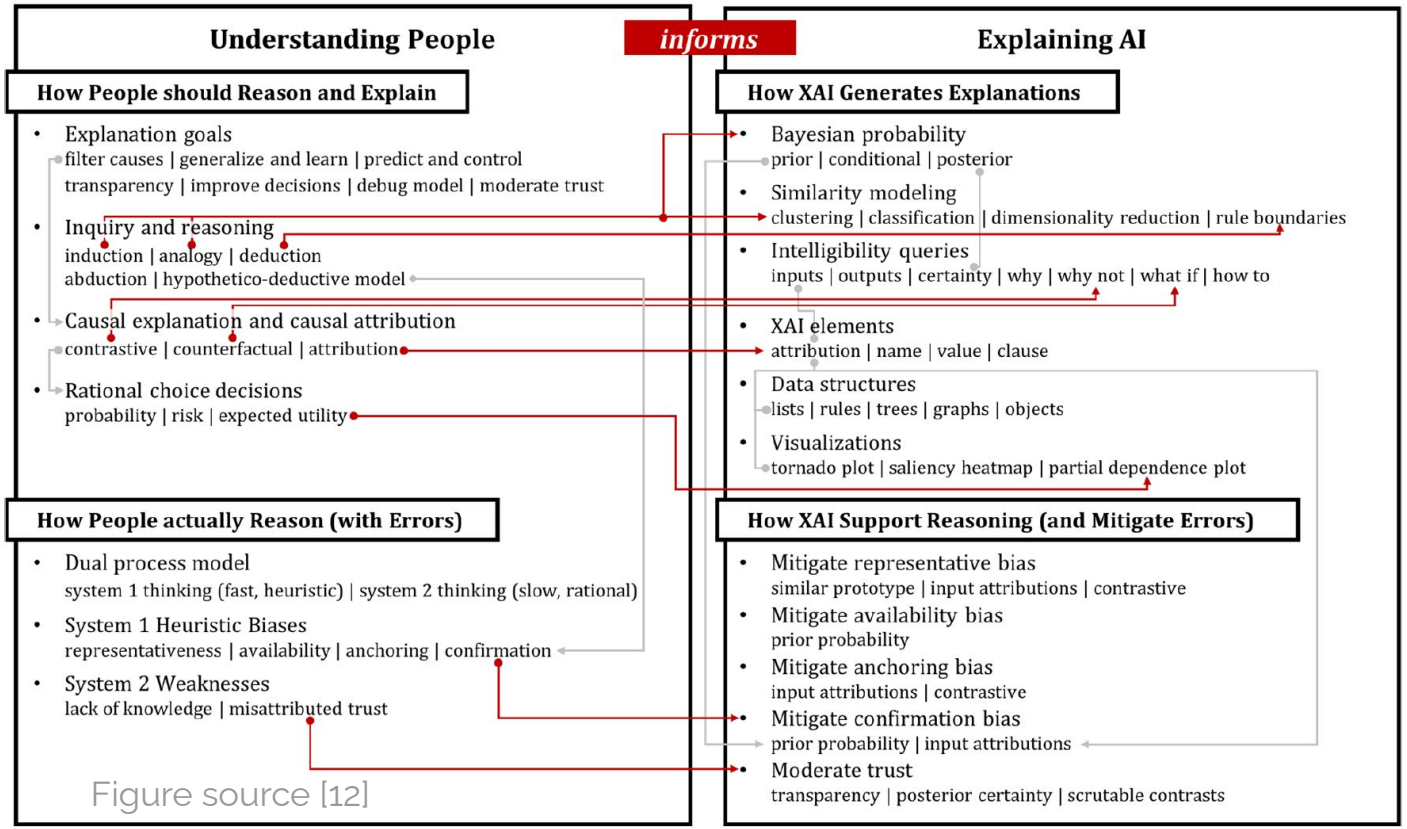


Figure source [12]

Related Work

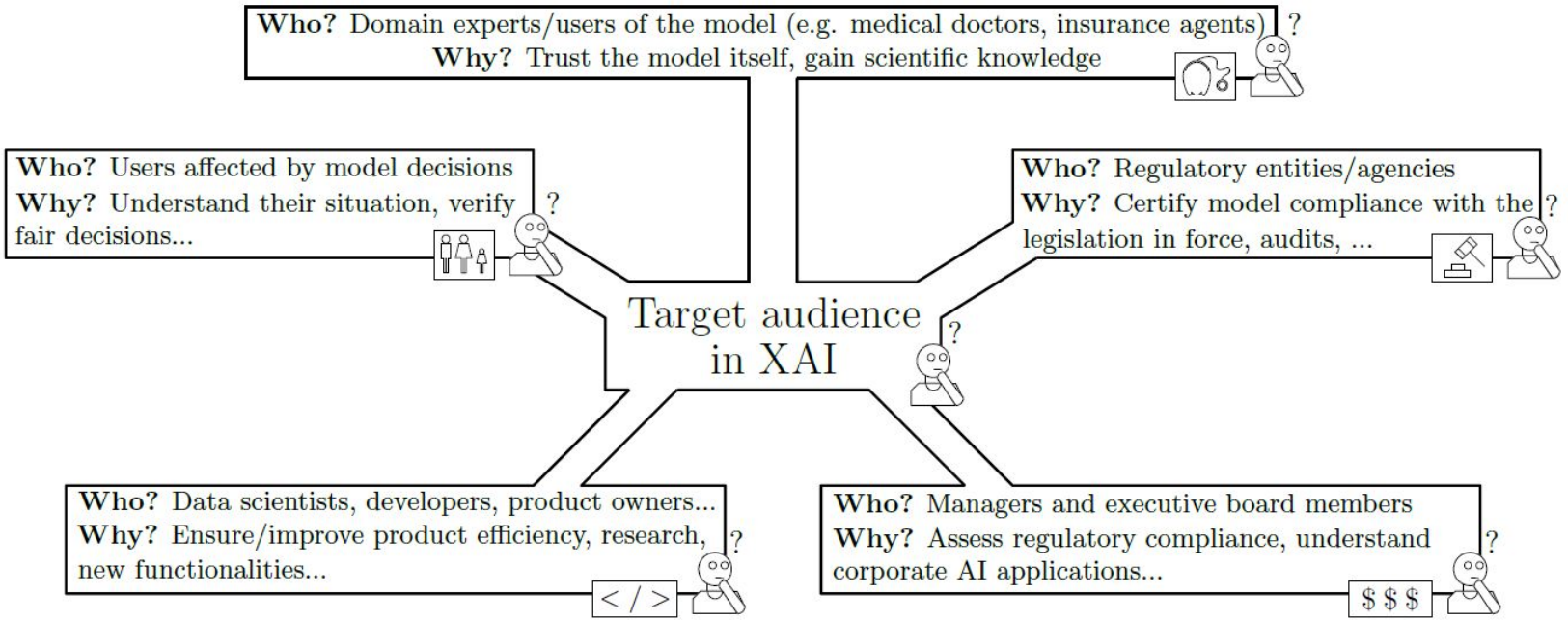
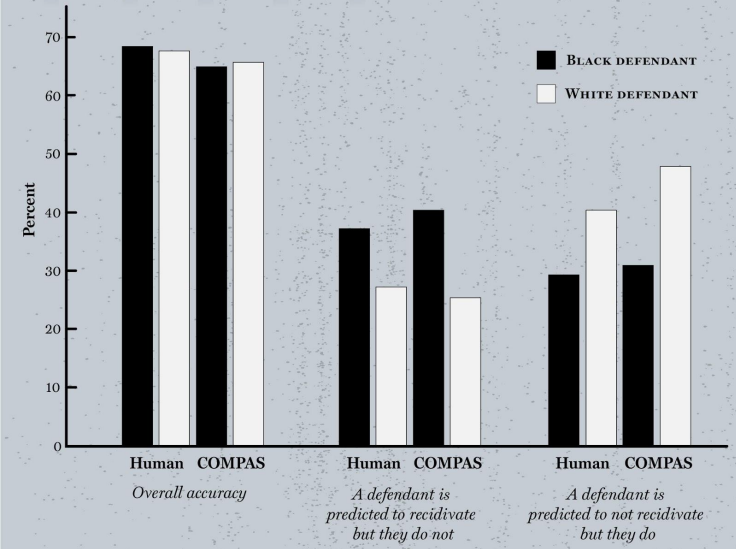


Figure source [9]

Questions To Be Considered

- What is the meaning of explainability?
- What defines a use-case in XAI?
- How many types of users exist? What are their characteristics?
- What makes an explanation suitable for each type?
- How to evaluate the “goodness” of an explanation for a given user?

Commercial Software No More Accurate Than Untrained People In Predicting Recidivism



Participants saw a description of a defendant that did not include their race and predicted whether each individual would recidivate within 2 years of their most recent crime.

Here, human predictions are compared to COMPAS algorithmic predictions. Human participants responding to an online survey, presumably none of them criminal justice experts, were approximately as accurate as COMPAS, the new *Science Advances* study reveals.

Dressel *et al.*, *Science Advances* (2018)

ScienceAdvances | AAAS

Carla Schaffer/AAAS

“In a given domain problem and use case, explainability is the process of providing suitable explanations to a recipient until clearing all his doubts about the system that is explained, in order to make both -system and recipient- reasonings more transparent”

Use-case and types of users

Determines the list of potential methods to use

Use case

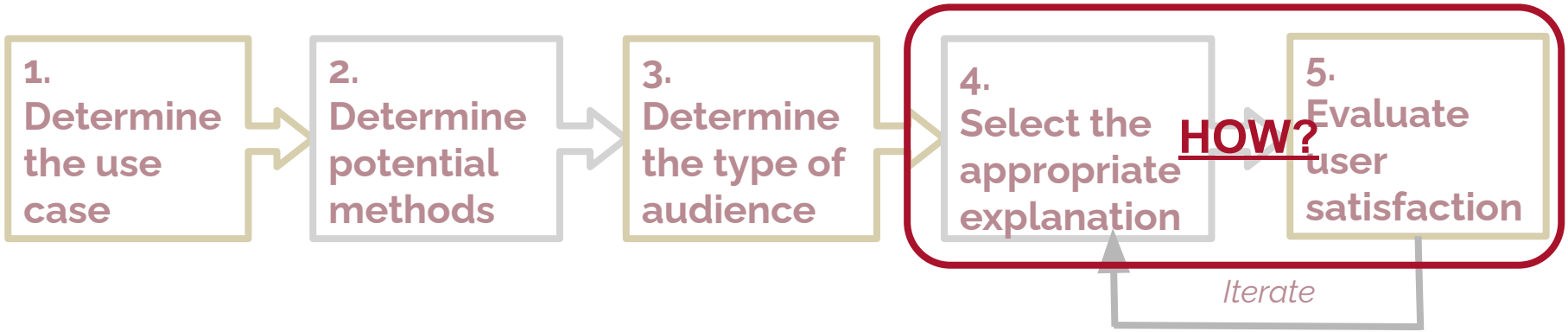
- Problem domain (i.e. credit risk, healthcare)
- Data type (i.e. tabular, images)
- Model type (i.e. NN, tree ensemble)
- Desired explanation type (global or local)
- Desired explanation scope (e.g. feature relevance)

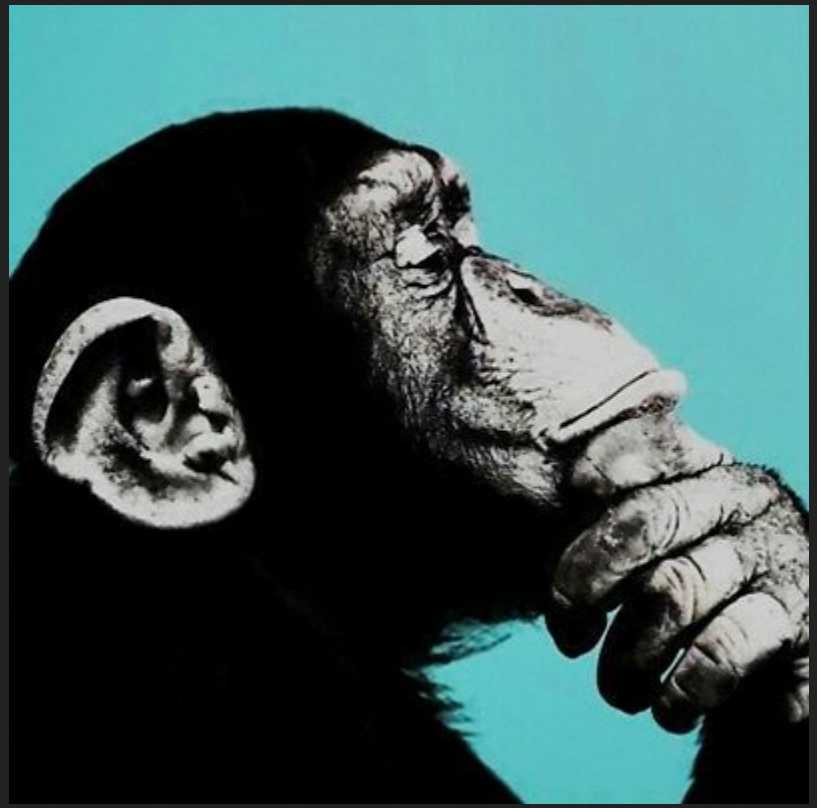
User types in technical fields

- Lay users (i.e. patients, applicants)
- Managers / Executives
- Entry-level (Junior) professionals
- Domain experts (senior) and researchers

Determines the quality of the explanation

User-centric XAI Process





Which explanation method is the most suitable for a specific user in a given use case?

Human-Friendly Explanations

Social

Explanations are essentially part of a series of interactions between the explainer and the explainee

Selected

Humans generally prefer to select some of these causes as being 'THE' explanation.

Contrastive

Humans tend to think in counterfactual cases

Consistent

Confirmation bias: "humans tend to discard information -and explanations- that are not consistent with their prior beliefs"

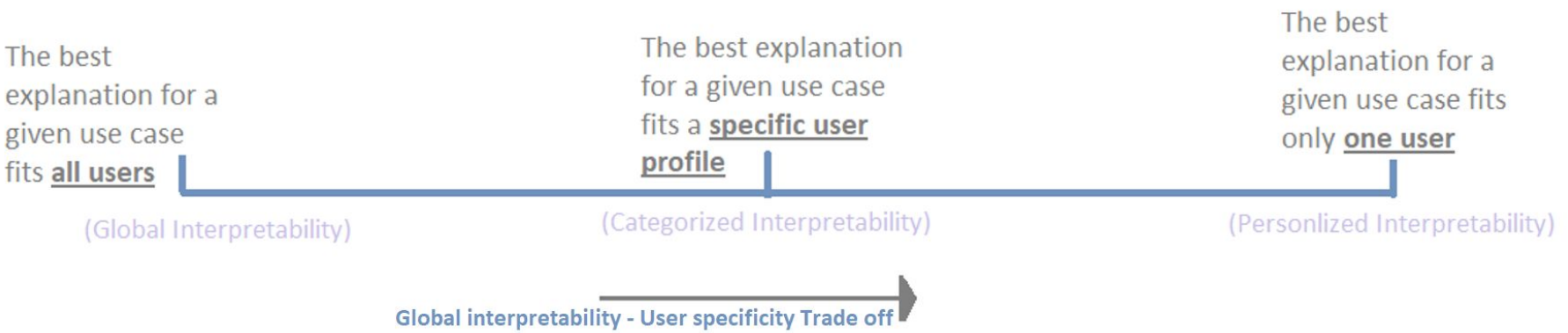
Abnormal

Humans tend to focus more on abnormal causes to explain and understand events

Truthful & General

Good explanations need to be true in other situations, which means that they need to explain other events

Lines of Work



4.
User-centric XAI: Empirical Study

Scope - Methodology - Implementation - Results

Interpretability Evaluation Levels

	Accuracy	Cost	Domain Considered
Functionally-grounded	**Low**	**Low**	**No**
Human-grounded	**Average**	**Average**	**No**
Application-grounded	**High**	**High**	**Yes**

Interpretability Indicators

Qualitative

- Sensitivity
- Implementation invariance
- Stability
- Identity
- Separability
- Completeness
- Correctness
- Compactness

Quantitative

- Form of the explanation's basic units
- Number of the explanation's basic units
- Compositionality
- Monotonicity and other types of relationships between the units
- Stochasticity and uncertainty

Scientific Process

1. **Observation** : interpretability is subjective
2. **Question**: 'Given a user profile, what is the most suitable explanation method(s) for it?' → '**Given an explanation method, which user would it be suitable for the most?**'
3. **Hypothesis**: Categorizing users depending on their level of expertise is a good trade-off between global and personalized interpretability
4. **Experiment:** Survey real users having different levels of expertise and evaluate the existence of a pattern between them
5. **Analysis:** Analyze survey results and conclude about the validity of the assumption

Scope & Methodology

1. Make a list of explanation methods (model-specific or model-agnostic) that can be applied to the most opaque but most accurate models: **Neural Networks**
2. **Cluster** them depending on their properties
3. **Apply** them to the same dataset and neural network
4. Extract the **properties** of the resulting explanations
5. Conduct a combination of human-grounded and application-grounded **evaluation**
6. **Interpret** results and hypothesis' validity

Explanations Desired Properties

- Relevancy
- Certainty
- Comprehensibility
- Representativeness
- Degree of importance
- Novelty
- Accuracy
- Fidelity
- Consistency
- Stability
- Scalability
- Simplicity
- Robustness
- Conciseness
- Monotonicity
- Interactive
- Generality
- Intelligibility

Literature Review

#↓	Name	Type	Sur/T	Data(tab	Examples	Code	Datas	Features	Expl Type	Proxy/output mo	Yea	Link	Reference
1	RxREN	NN	Tax+Sur	Tab	Medical/credi	-	Yes	Accuracy - Fi	Explanation by simplification	Rule based learner /	2011	https://wv	M. G. Augasta, T.
2	REFNE	NN	Tax+Sur	Tab	voting	-	Medical/c	-	Explanation by simplification	Rule based learner /	2003	https://wv	Z.-H. Zhou, Y. Jia
3	Using sampling and queries	NN	Tax+Sur	Tab	-	-	-	-	Explanation by simplification	Rule based learner/D	1994	http://cite	M. W. Craven, J.
4	Using genetic algorithms	NN	Tax	Tab/TXT	Monk	-	Yes	Fidelity - com	Explanation by simplification	Rule based learner	1997	https://d1	A. D. Arbatli, H. L
5	Rule generation	NN	Tax	Tab/TXT	Flower	-	-	Robustness -	Explanation by simplification	Rule based learner	1994	https://ie	L. Fu, Rule gener
6	Extracting refined rules	NN	Tax	Tab/TXT	DNA+Monk	-	Yes	Bad and long	Explanation by simplification	Rule based learner	1993	https://lin	G. G. Towell, J. W
7	Extracting Rules from distributed	NN	Tax	Tab/TXT	arm robot	-	Yes		Explanation by simplification	Rule based learner	1995	http://pag	S. Thrun, Extracti
8	FERNN	NN	Tax	Tab/TXT	Monk3	-	Yes	Fidelity - expr	Explanation by simplification	Rule based learner	2000	http://cite	R. Setiono, W. K.
9	Symbolic Interpretation	NN	Tax	Tab/TXT	Medical/Iris	Yes	Yes	Completenes	Explanation by simplification	Rule based learner	1999	http://cite	I. A. Taha, J. Gho
10	Extracting Rules from Trained NI	NN	Tax	Tab/TXT	-	-	-	-	Explanation by simplification	Rule based learner	2000	https://ie	H. Tsukimoto, Ext
11	Extracting comprehensible mode	NN	Tax	Tab/TXT	-	-	Yes	Comprehensi	Explanation by simplification	Decision Tree	1996	https://mi	M. W. Craven, Ex
12	ICU Outcome Prediction	NN	Tax	Tab/TXT	Medical	-	Yes	performance-	Explanation by simplification	Decision Tree	2016	https://wv	Z. Che, S. Purush
13	Tree Regularization	NN	Tax	Tab/TXT	Medical	-	Yes	Accuracy - Si	Explanation by simplification	Decision Tree	2018	https://ar	M. Wu, M. C. Hug
14	Distilling a Neural Network	NN	Tax	Any	MNIST	Yes	Yes	Accuracy - Si	Explanation by simplification	Decision Tree	2017	https://ar	N. Frosst, G. Hint
15	Extracting decision trees	NN	Tax+Sur	Tab	Medical/Iris	-	Yes	Accuracy - cc	Explanation by simplification	Decision Tree	1999	https://d1	R. Krishnan, G. S
16	Feature-space partitioning	NN	Tax+Sur	Any	Bricks	-	Yes	Accuracy	Explanation by simplification	Decision Tree	2016	https://ar	J. J. Thiagarajan,
17	Deepred–rule extraction	NN	Tax	Tab/TXT	-	-	-	-	Explanation by simplification	Decision Tree	2016	https://m	J. R. Zilke, E. L. M
18	ANN-DT	NN	Tax	Tab/TXT	Sin/cos	-	Yes	-	Explanation by simplification	Decision Tree	1999	https://ie	G. P. J. Schmitz,
19	Decision tree induction	NN	Tax	Tab/TXT	Credit	-	Yes	Accuracy - ge	Explanation by simplification	Decision Tree	2001	https://ie	M. Sato, H. Tsuki
20	Distilling the Knowledge	NN	Tax	IMG/TXT	-	-	-	Accuracy	Explanation by simplification	Other	2015	https://ar	G. Hinton, O. Viny
21	Deep Taylor decomposition	NN	Tax	IMG	Animals/MNI	-	Yes	Stability - per	Feature relevance explanatic	Importance/Contribut	2017	https://re	G. Montavon, S. L
22	PATTERNNET	NN	Tax	Tab	imageNet			Stability - per	Feature relevance explanatic	Importance/Contribut	2017	https://ar	P.-J. Kindermans,
23	Propagating Activation Difference	NN/DNN	Tax+Sur	Any	DNA/MNIST	Yes	-	-	Feature relevance explanatic	Importance/Contribut	2016	https://ar	A. Shrikumar, P. G
24	Explain neural network classifica	NN	Tax	Tab	voting/calls	-	Yes	-	Feature relevance explanatic	Sensitivity / Saliency	2002	https://d1	R. F´eraud, F. Cl´
25	Axiomatic Attribution	NN/DNN	Tax+Sur	Any	imageNet/me	-	Yes	Accuracy - vi	Feature relevance explanatic	Sensitivity / Saliency	2017	https://ar	M. Sundararajan,
26	Iterative Debugging	NN	Tax	Any	data analyst	-	Yes	Accuracy - Sc	Local Explanation-DecisionT	DecisionTree / Sensi	2017	https://dl	S. Krishnan, E. W
27	Iterative Debugging	AG	Tax+Sur	Any	Yes	-	Yes		Local Explanations - Decisio	Decision Tree	2017	https://dl	S. Krishnan, E. W
28	Local Explanations	NN	Tax	Any	data analyst	-	Yes	Visualization	Local Explanation-DecisionT	DecisionTree / Sensi	2018	https://ar	J. Adebayo, J. Gil
29	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Architecture modification	other	2018	https://ar	N. Papernot, P. M
30	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Local Explanations-Activatio	Activation clusters	2018	https://ar	N. Papernot, P. M
31	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Architecture modification	other	2018	https://ar	N. Papernot, P. M
32	Concept Activation Vectors	NN	Tax	IMG	analyst/medic	-	Yes	accuracy - cr	Local Explanations-Activatio	Activation clusters	2018	https://ar	B. Kim, M. Watter
33	Semantic Information	NN	Tax	Video	Random	-	Yes	captions -	Text explanation-Caption ger	Caption generation	2017	https://op	Y. Dong, H. Su, J.

85 explanation methods found

63 of them didn't include their code

17 didn't include an example explanation

53 resulting methods

Resulting List

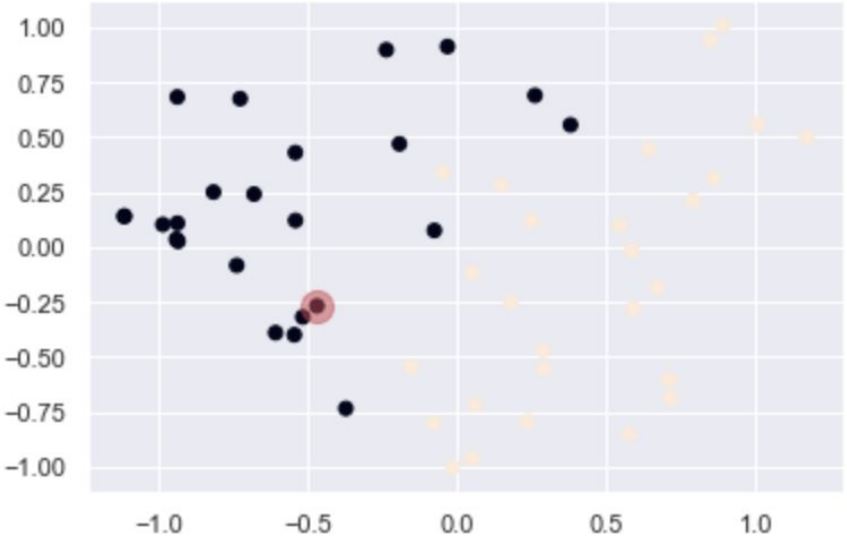
Name	Features	Explanation_Output
Concept Activation Vectors	accuracy-robustness-expressivity	Activation clusters
Deep k-Nearest Neighbors	robustness-detecting bias	Activation clusters
Generate Reviews	efficiency-scalability	Activation Maximization
IP	generality-scalability-robustness	Activation Maximization
Semantic Information	expressivity	Caption generation
Rationalizing Neural Predictions	accuracy-visualization	Caption generation/Saliency
A Unified Approach	accuracy-scalability-consistency	Conditional/Dependence/Shapley plots
ICU Outcome Prediction	accuracy-scalability-fidelity-visualization-co...	Conditional/Dependence/Shapley plots
Visualizing the Feature Importance	comparative	Conditional/Dependence/Shapley plots
Identifying Prediction Invariance (aLime)	accuracy-simplicity-fidelity	DR
Rule generation	robustness-soundnes-completeness	DR
Using genetic algorithms	accuracy-fidelity-comprehensibility	DR
Symbolic Interpretation	accuracy-completeness-comprehensibility-certai...	DR
FERNN	accuracy-fidelity-expressivity-concise	DR

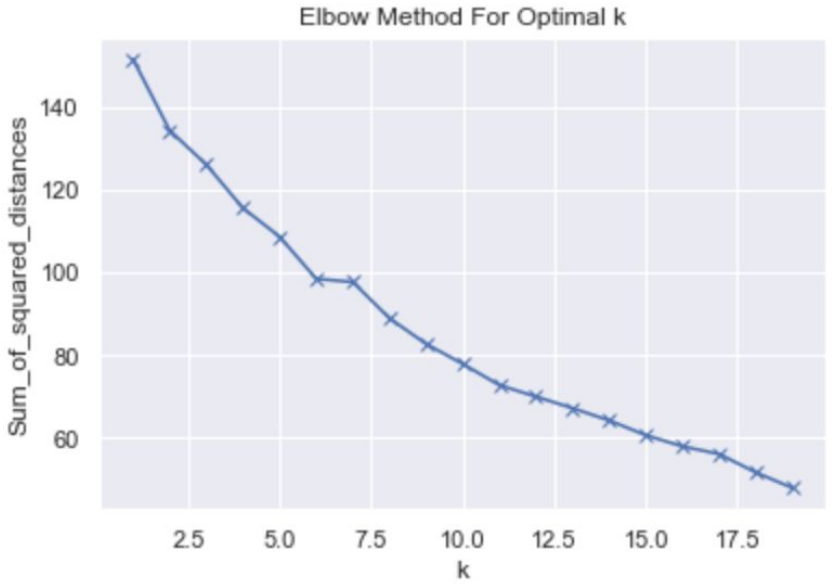
Resulting List

	F_accuracy	F_certainty factor	F_comparative	F_completeness	F_comprehensibility	F_concise	F_consistency	F_debugging	F_detecting bias
A Unified Approach	1	0	0	0	0	0	1	0	0
Anchors	0	0	0	0	0	0	0	0	0
Attribute Interactions in Datasets	0	0	0	0	0	0	0	0	0
Axiomatic Attribution	1	0	0	0	0	0	0	0	0
Boolean rule extraction	0	0	0	0	0	0	0	0	0
Concept Activation Vectors	1	0	0	0	0	0	0	0	0
Decision tree induction	1	0	0	0	0	0	0	0	0
Deep Taylor decomposition	0	0	0	0	0	0	0	0	0

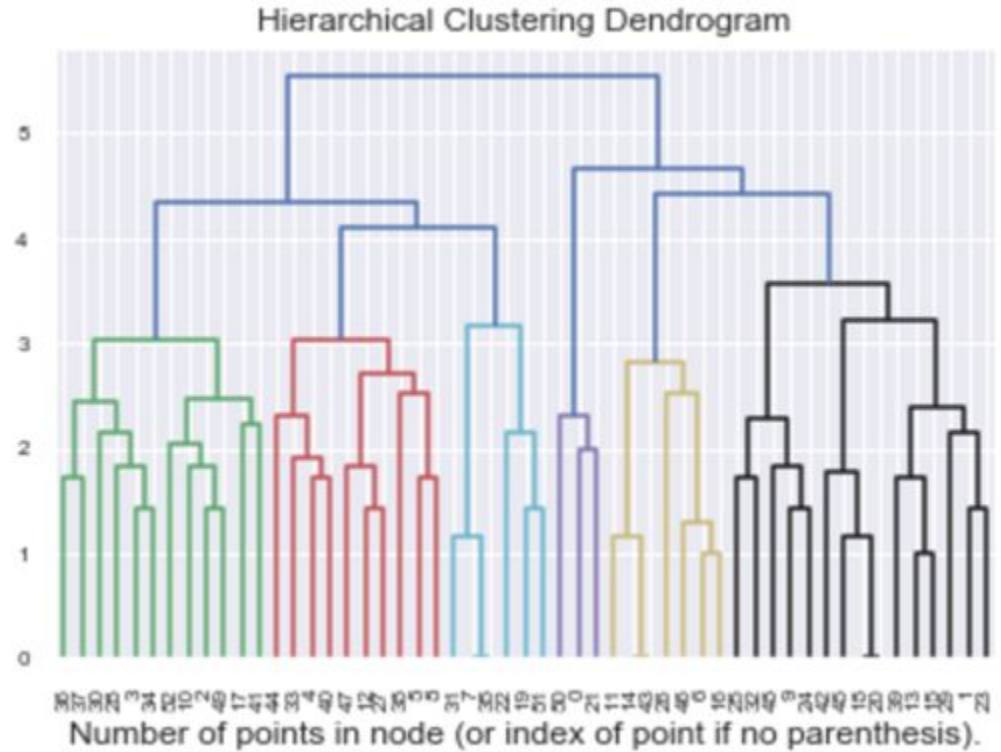
53 rows × 41 columns

Clustering


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.0
−0.5
0.0
0.5
1.0


Elbow Method For Optimal k
Sum_of_squared_distances
140
120
100
80
60
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k

Clustering


Hierarchical Clustering Dendrogram
5
4
3
2
1
0
Number of points in node (or index of point if no parenthesis).

Explainable Artificial Intelligence Survey

Intelligent systems are everywhere around us. They take many forms and their decisions are starting to have more serious impact in our lives, from credit risk evaluation, to medical diagnosis, or even self-driving cars.

A new field has emerged, that allows us -HUMANS- to extract explanations from these systems, in order to understand their reasoning, learn from it, control it, and correct it in the case of wrong interpretations.

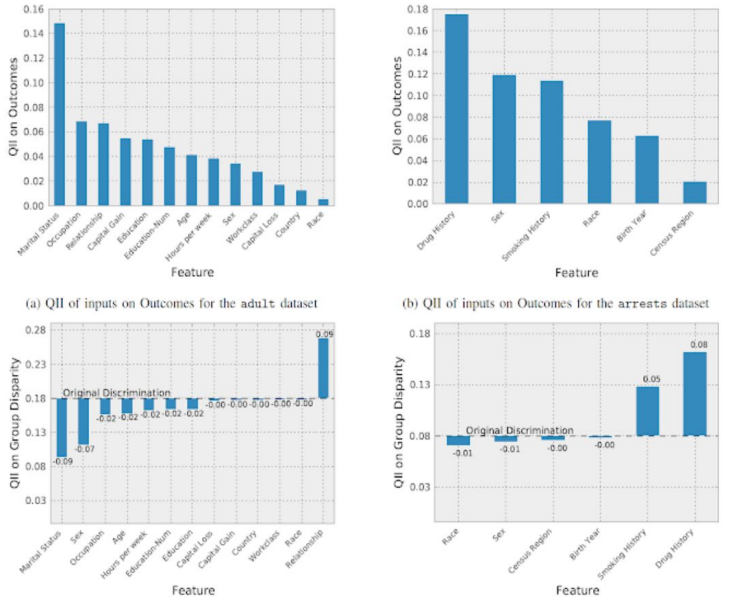
This field is called Explainable Artificial Intelligence (XAI) and it mainly tries to give answers to 'Why' questions: 'Why am I sick?' or 'Why didn't my application get accepted?'.

This survey is supposed to evaluate the usefulness of explanations that can be provided by intelligent systems to justify their predictions.

The survey will take 15 min of your time. Please answer all questions honestly.

But Why?

Question 7: How much do you rate your understanding of this explanation?



(a) QII of inputs on Outcomes for the adult dataset (b) QII of inputs on Outcomes for the arrests dataset

(c) QII of Inputs on Group Disparity by Sex in the adult dataset (d) Influence on Group Disparity by Race in the arrests dataset

Fig. 2: QII measures for the adult and arrests datasets

1 2 3 4 5 6 7 8 9 10

I am very confused and don't understand anything! 😵

All is pretty clear and well-explained! 🤓

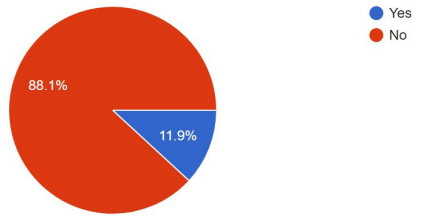
Survey Extracts

Survey Statistics

Preliminary Questions

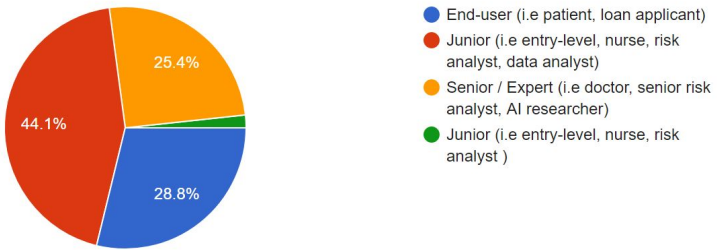
Do you know about the existence of a field that allows humans to ask machines for justifications?

159 responses



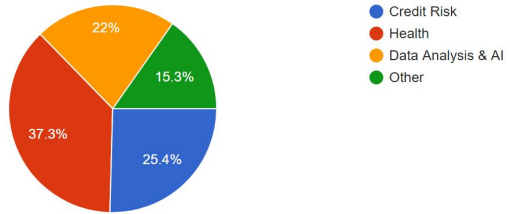
Rate your level expertise :

159 responses



Choose your field of interest/expertise :

159 responses



Results

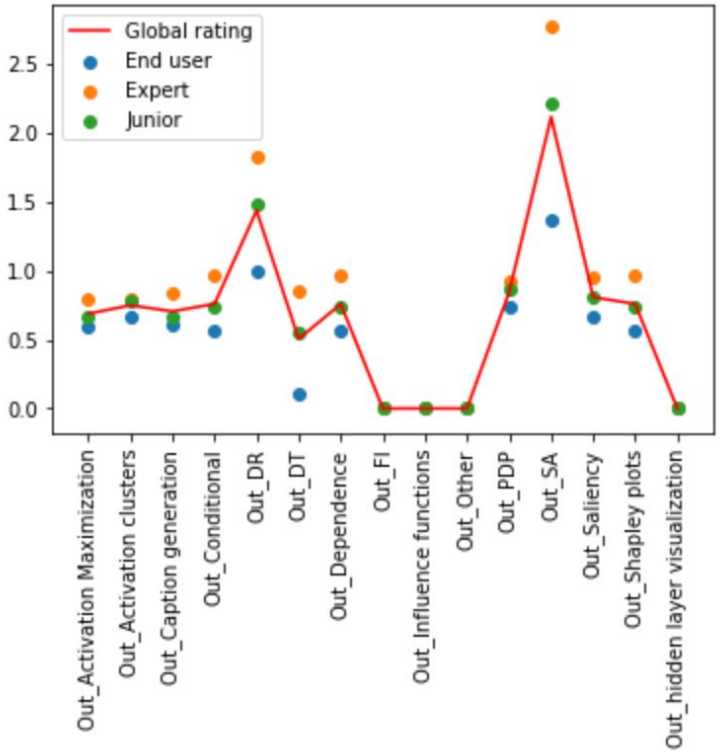
Profile	0	1	2	3	4	5	6	7	8	9
End_user	1.125000	6.000000	6.083333	6.625000	5.62500	6.083333	5.916667	6.625	7.375000	3.916667
Expert	8.562500	7.937500	8.687500	8.000000	9.71875	8.343750	9.625000	9.500	9.250000	9.625000
Junior	5.450000	6.675000	6.825000	7.825000	7.40000	6.725000	8.550000	8.100	8.625000	7.950000
mean_all	5.045833	6.870833	7.198611	7.483333	7.58125	7.050694	8.030556	8.075	8.416667	7.163889

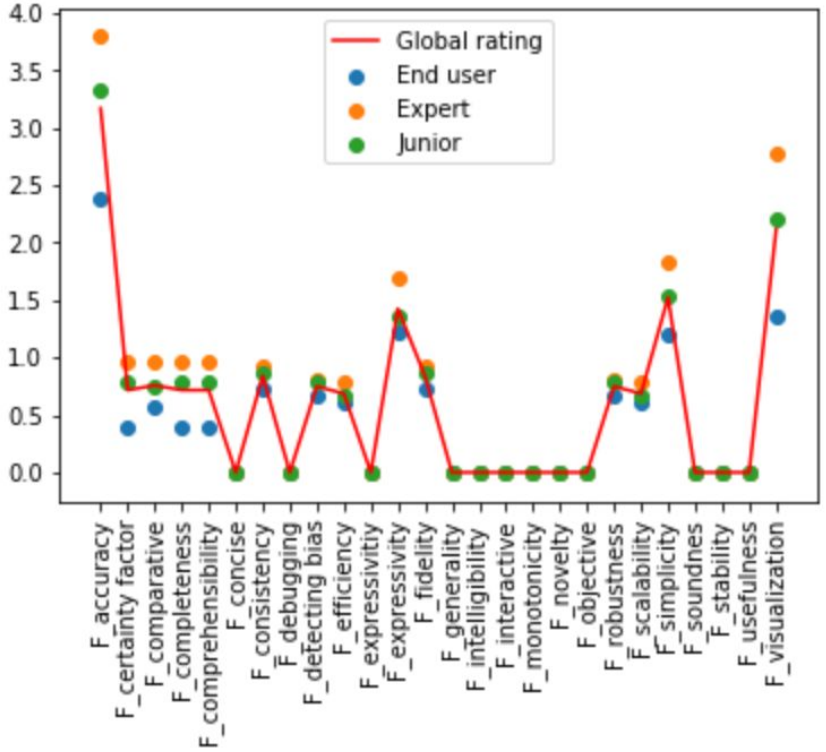
	F_accuracy	F_certainty factor	F_comparative	F_completeness	F_comprehensibility	F_concise	F_consistency	F_debugging	F_detecting bias	F_efficiency
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	0
4	0	0	1	0	0	0	0	0	0	0

Results

	F_accuracy	F_certainty factor	F_comparative	F_completeness	F_comprehensibility	F_concise	F_consistency	F_debugging	F_detecting bias	F_efficiency	...	(
0	2.383333	0.391667	0.562500	0.391667	0.391667	0.0	0.737500	0.0	0.662500	0.600000	...	0.
1	3.800000	0.962500	0.971875	0.962500	0.962500	0.0	0.925000	0.0	0.800000	0.793750	...	0.
2	3.322500	0.795000	0.740000	0.795000	0.795000	0.0	0.862500	0.0	0.782500	0.667500	...	0.
3	3.168611	0.716389	0.758125	0.716389	0.716389	0.0	0.841667	0.0	0.748333	0.687083	...	0.

Results


Global rating
End user
Expert
Junior
2.5
2.0
1.5
1.0
0.5
0.0
Out_Activation Maximization
Out_Activation clusters
Out_Caption generation
Out_Conditional
Out_DR
Out_DT
Out_Dependence
Out_FI
Out_Influence functions
Out_Other
Out_PDP
Out_SA
Out_Saliency
Out_Shapley plots
Out_hidden layer visualization


Global rating
End user
Expert
Junior
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
F_accuracy
F_certainty factor
F_comparative
F_completeness
F_comprehensibility
F_concise
F_consistency
F_debugging
F_detecting bias
F_efficiency
F_expressivity
F_expressivity
F_fidelity
F_generality
F_intelligibility
F_interactive
F_monotonicity
F_novelty
F_objective
F_robustness
F_scalability
F_simplicity
F_soundnes
F_stability
F_usefulness
F_visualization

5.
Conclusion

Recap – Summary – Lessons Learned – Future Work

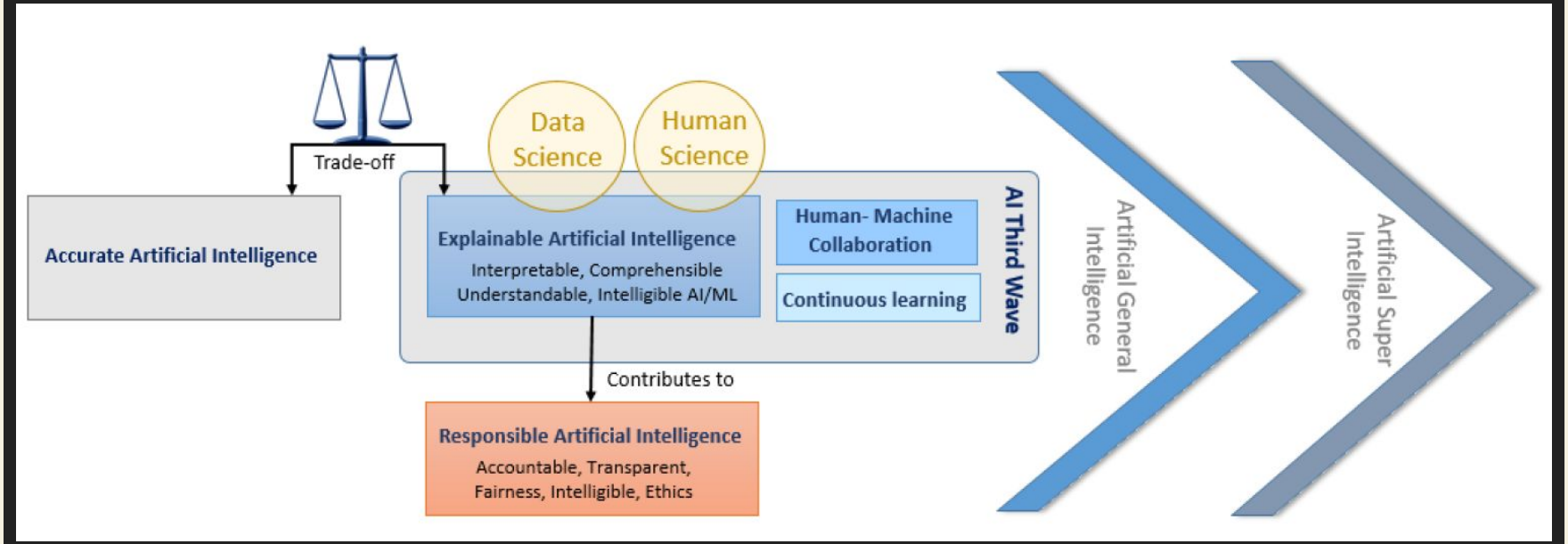


Figure source [7]

Contributions

- We presented a review of the state of the art of XAI and Interpretable ML
- We proposed a pragmatic definition of explainability
- We distinguished between Interpretable ML and XAI
- We defined three theories of explainability
- We differentiated between users depending on their level of expertise
- We suggested a method to determine desired properties of explanations for each user profile

Discussion

- The absence of an agreed upon definition of explainability confirms that it is a subjective concept
- Many papers are still reluctant to admit the importance of considering the user in the explanation process
- Interpretability assessment is still at its embryonic stage
- The vast majority of the proposed methods do not provide access to their code
- Many of the desired properties of explanations CANNOT be formulated mathematically

Lessons Learned

- Distinguish contribution opportunities in a completely new field
- Critical thinking
- Define the central problem
- Creativity in thinking about potential solutions
- Follow the Scientific process: Hypothesis → Experiment → Conclusion
- Rely on data science tools to solve the problem
- Self-supervision and self-organization
- Seek help when needed

Future Work

- This work should be extended to other fields and other profiles such as managers
- The same survey can be done differently with open questions
- We need an XAI assessment framework that allows to compare different explanations based on proxy tasks
- XAI should include social sciences and HCI in the research to truly thrive
- XAI should be part of data science education programs

Thank You For Your Attention!

ANY QUESTION OR SUGGESTION?

Long Life To XAI!

References

1. https://syncedreview.com/2019/12/19/2019-in-review-10-ai-failures/
2. https://sportswizard.com/2019/04/05/ai-adoption-and-barriers/
3. https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence
4. https://analyticsindiamag.com/top-8-funniest-and-shocking-ai-failures-of-all-time/
5. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf
6. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).
7. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018, 6, 52138–52160. [CrossRef]
8. https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f
9. Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.
10. Murdoch, W. James, et al. "Interpretable machine learning: definitions, methods, and applications." *arXiv preprint arXiv:1901.04592* (2019).
11. Arya, Vijay, et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques." arXiv preprint arXiv:1909.03012 (2019).
12. Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019.

Appendix