

WHY CURRENT STATE-OF-THE-ART EXPLAINABLE AI METHODS ARE INADEQUATE

Kary Främling

Professor in Data Science

Head of Explainable AI Team



UMEÅ UNIVERSITY

KARY'S XAI VISION

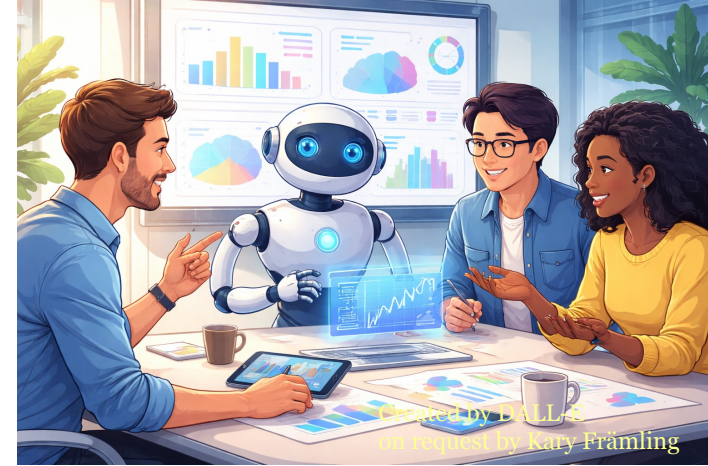
- The Human brain is (still) the biggest "black box" in the world - but humans manage to explain their "high-level" actions and decisions without knowing anything about the black-box internals
- Humans can even adapt explanations to the target explainee(s)
 - Vocabulary
 - Level of detail/abstraction
 - Visualize, illustrate in different ways
 - Interactive dialogs
- This is the idea and ambition that led to the Contextual Importance and Utility (CIU) method in Kary's doctoral thesis¹⁾

¹⁾ Främling, Kary. *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. PhD. thesis : Institut National de Sciences Appliquées de Lyon, Ecole Nationale Supérieure des Mines de Saint-Etienne, France, 1996. 209 p.



SOCIAL XAI VISION

- Based on Shonan seminar in September 2023
- Book published in spring 2026¹⁾
- Social XAI (sXAI) systems should communicate in similar ways as humans when justifying or explaining decisions, actions, plans, ...
 - Adapt the used vocabulary, levels of details etc. to the explainee
 - Present information in digestible “chunks”, go into details only if needed (incrementality)
 - Use images, text, gestures as appropriate (multimodality)
 - Be “co-constructive”, i.e. have explainer and explainee gain mutual understanding throughout the interaction



Katharina J. Rohlfling · Kary Främling · Brian Lim · Suzana Alpsancar · Kirsten Thommes *Editors*

Social Explainable AI

Communications of NII Shonan Meeting

This open access book introduces social aspects relevant to research and developments of explainable AI (XAI). The new surge of XAI responds to the societal challenge that many algorithmic approaches (such as machine learning or autonomous intelligent systems) are rapidly increasing in complexity, making justified use of their recommendations difficult for the users. A large body of approaches now exists with many ideas of how algorithms should be explainable or even be able to explain their output. However, few of them consider the users' perspective, and even less address the social aspects of using XAI. To fill the gap, the book offers a conceptualization of explainability as a social practice, a framework for contextual factors, and an operationalization of users' involvement in creating relevant explanations.

For this, scholars across disciplines gathered at the Shonan meeting to account for how explanation generation can be tailored to diverse users and their heterogeneous goals when interacting with XAI. As a result, social interaction is the key to the involvement of the users. Accordingly, we define sXAI (social eXplainable AI) as systems that interact with the users in such a way that an incremental adaptation of explaining to the users is possible, along with the unfolding context of interaction, to yield a relevant explanation at the interface with both active partners—human and AI. To encourage novel interdisciplinary research, we propose to account for the following dimensions:

- **Patternedness:** XAI should account for different contexts that yield different social roles impacting the construction of explanations.
- **Incrementality:** XAI should build on the contribution of the involved partners who adapt to each other.
- **Multimodality:** XAI needs to use different communication modalities (e.g., visual, verbal, and auditory).

This book also addresses how to evaluate social XAI systems and what ethical aspects must be considered when employing sXAI. Together, the book pushes forward the building of a community interested in sXAI. To increase the readability across disciplines, each chapter offers a rapid access to its content.

Except where otherwise noted, this book is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



ISBN 978-981-96-5289-1



▶ springer.com

ELLIIT, HALMST,
SEPTEMBER 202

Rohlfling · Främling · Lim
Alpsancar · Thommes *Eds.*



Social Explainable AI

Katharina J. Rohlfling · Kary Främling
Brian Lim · Suzana Alpsancar
Kirsten Thommes *Editors*

Social Explainable AI

Communications
of NII Shonan Meeting



OPEN ACCESS



Springer

EXAMPLE FROM SXAI BOOK: KARY JUSTIFIES CAR CHOICE TO HIS WIFE

Question/Answer	Explanatory move
Why?	Why should we buy car A?
Why answer	It's safe, spacious and not too expensive.
Why not?	But car B looks quite similar and it's cheaper.
WhyNot answer	Yes but it consumes a lot of fuel and maintenance costs are high.
Contrastive	But car C is bigger than car A and has the same price, so why not car C instead?
Contrastive answer	Yes, but then . . .
What if?	But what if you would add this extra option to car C?
Counterfactual answer	Yes, then it would become better but the price . . .



UMEÅ UNIVERSITY

KARY FRÄMLING

Actual example used by Kary since the 1990's

EXAMPLE FROM SXAI BOOK: KARY JUSTIFIES CAR CHOICE TO TECHIE FRIENDS

Question/Answer	Explanatory move
Why?	Why did you buy car A?
Why answer	It's quick, fun, big enough and not too expensive.
Why not?	Why didn't you buy car B instead, which has 4-wheel drive?
WhyNot answer	It's expensive and consumes a lot of fuel .
Contrastive	But then car C has more power and same price?
Contrastive answer	Yes, but then . . .
What if?	But what if you would remove this extra option from car C?
Counterfactual answer	Then it would become cheaper but . . .

Explaining same choice
but different vocabulary
and emphasis!



UMEÅ UNIVERSITY

KARY FRÄMLING

Actual example used by Kary since the 1990's

XAI SYSTEM JUSTIFIES CAR RECOMMENDATION TO HUMAN

Proposal/Reaction/Question/Answer	Explanatory move
<...>	<initial interaction before AI system gives first proposal>
Proposal	I suggest you buy car A
Why?	<i>Why should I buy car A?</i>
Why answer	It's safe, economic and within your price range.
Why not?	<i>Wouldn't car B be better?</i>
Why Not answer	It's expensive and consumes a lot of fuel .
Refusal	<i>But I like car B more</i>
Question	Why do you like car B more?
(Human) answer	<i>I like its shape and I preferred driving with it</i>
New proposal	Ok, I took that into account. Then car C might actually be even better
Reaction	<i>Indeed, I had forgotten car C</i>
Why this and not that?	<i>So why would car C be better than car B?</i>
Contrastive answer	Because ...
<...>	<interaction continues>



UMEÅ UNIVERSITY

KARY FRÄMLING

XAI SYSTEM JUSTIFIES CAR RECOMMENDATION TO HUMAN

Proposal/Reaction/Question/Answer	Explanatory move
<...>	<initial interaction before AI system gives first proposal>
Proposal	I suggest you buy car A
Why?	Why should I buy car A?
Why answer	It's safe, economic and within your price range.
W	Interactive and co-constructive dialogue that involves many different kinds of questions!
W	
R	
Q	
(E	
N	
R	
Why this and not that:	So why would car C be better than car B?
Contrastive answer	Because ...
<...>	<interaction continues>



UMEÅ UNIVERSITY

KARY FRÄMLING

XAI: STATE-OF-THE-ART

**... and Some Reasons for Why it is
Inadequate**



UMEÅ UNIVERSITY

KARY FRÄMLING

WHAT QUESTIONS DO XAI METHODS ANSWER?

- **SHAP, LIME** (Additive Feature Attribution - AFA): “Changing the value of which features would influence the result more than others?”
 - Partial answer to “Why?” and “Why not?” questions
- **Counterfactuals**: “How to?” change outcome from one class to another
- **Rules, decision trees**: “What is the internal logic of the black box?”
- **Some unanswered questions**: Simple “Why?”/“Why not?”, “What if?”, “Why A and not B?”, “Why is this feature important (or not)?”



UMEÅ UNIVERSITY

KARY FRÄMLING

AFA METHODS DO NOT REALLY ANSWER “WHY?” OR “WHY NOT?”

- AFA methods calculate **influence** values
- Influence values are **contrastive** in the sense that they indicate the difference between the current instance and a *reference instance*
- Definition: A **Reference Instance (RI)** is an instance that has zero influence for the result, as well as for all features
- RI is typically not a real instance in AFA methods
- AFA methods actually answer a question like “How do the differences in feature values between the current instance and the reference instance influence the outcome?”
 - Where “outcome” is actually the outcome for the current instance versus the outcome of the reference instance



UMEÅ UNIVERSITY

KARY FRÄMLING

AFA AND THE NULL EXPLANATION

- If the current instance is identical to the reference instance, then we get a null explanation: All influence values are zero
- Real-life example that illustrates why this can lead to unexpected explanations that can easily be misinterpreted:

An AI system predicts a person's wealth from three inputs: salary, savings, and debt. For a person whose salary equals the population mean, SHAP assigns zero attribution to salary. The explanation suggests that salary is irrelevant to this person's wealth. Yet salary is plainly the most structurally important feature: if it were different, wealth would change substantially.

- **Importance and influence are two distinct and different concepts – and that is a crucial distinction to make!**



UMEÅ UNIVERSITY

KARY FRÄMLING

COUNTERFACTUALS ARE NOT EXPLANATIONS

- A counterfactual instance is one that modifies the current instance in a way that changes the output into a target output
 - The target output is typically a different class than the current one
- Finding counterfactuals is an optimization problem that can give an infinite number of solutions
- A list of counterfactual instances is not an explanation as such
- Answering the question “Why A not B?” would be an explanation: “You can get the loan if you take these actions”
- **Is there a XAI method that can actually answer the fundamental question “Why A and not B?” (and its inverse) in a consistent way?**



UMEÅ UNIVERSITY

KARY FRÄMLING

CONTEXTUAL IMPORTANCE AND UTILITY (CIU)

**... and How CIU Answers All the Questions
in a Mathematically Consistent Way –
While Enabling User-Adapted sXAI**



UMEÅ UNIVERSITY

KARY FRÄMLING

CIU ORIGINS

- Initial project for PhD thesis¹: Waste disposal site selection, 1991
- Three different decision support methods were tested: Rules, AHP, Electre (non-ML)
- ML models avoid expressing decider preferences
 - Neural Networks
- Necessary to explain to: citizens, inhabitants, decision makers, ...
- Requirement: XAI method has to be truly model-agnostic (not limited to ML models)
- First version of Contextual Importance and Utility presented in 1992²

1. Främling, K. *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. PhD. Thesis, 1996. 209 p.
2. Främling, K. *Les réseaux de neurones comme outils d'aide à la décision floue*. D.E.A. thesis, 1992. 55 p.

ELLIIT, HALMSTAD,
24 SEPTEMBER 2026



UMEÅ UNIVERSITY

KARY FRÄMLING



CIU JUSTIFIES SITE SELECTION (1992)

- Local XAI method:
Justifies result for one instance (or “context”)
- Textual explanations
- The actual explanation could be configured to use different styles (rule/weight), sort according to feature importance or value utility, etc.

Le Secteur 1222.12 est bon, valeur de préférence 63.7 sur 100, parce que:

Les critères ayant une très bonne valeur sont:

- Longueur du parcours en agglomération, qui est *assez important* (0.10).
- Valeur écologique, qui est *peu important* (0.10).
- Longueur parcours RN-Secteur, qui est *peu important* (0.05).
- Longueur de ruisseaux, qui est *très peu important* (0.05).
- Proximité des fournisseurs de DS, qui est *très peu important* (0.05).
- Longueur à construire ou à élargir, qui est *très peu important* (0.04).
- Proximité d'un utilisateur de marnes, qui est *très peu important* (0.04).
- Nombre de sources privées, qui est *très peu important* (0.03).

Les critères ayant une bonne valeur sont:

- Nombre de localités traversées, qui est *assez important* (0.13).
- Géologie globale, qui est *assez important* (0.12).

Les critères ayant une très mauvaise valeur sont:

- Aptitude agricole, qui est *assez important* (0.15).
- Proximité de chemin de fer, qui est *très peu important* (0.05).

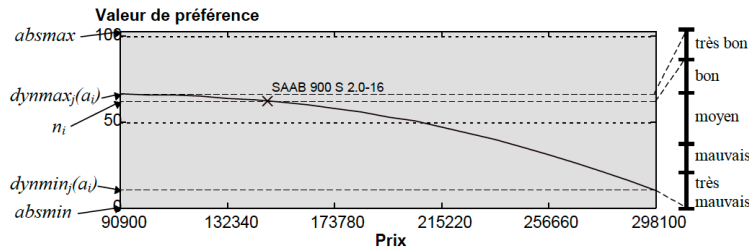


UMEÅ UNIVERSITY

KARY FRÄMLING

XAI ELEMENTS OF CIU FOR CAR SELECTION (1996)

- **Model:** Neural network that estimates a preference value for different cars, based on their characteristics
- **Intermediate Concepts (IC)** enable more or less detailed explanations
- **Input-Output (IO) plots** show output value changes when modifying one feature value, while other feature values are fixed
- Core CIU building blocks still remain the same but **new capabilities added since 2018**



La voiture SAAB 900 S 2.0-16 est moyenne, valeur de préférence 57.3 sur 103, parce que:

- Prix, qui est *presque déterminant* (0.58) a une très bonne valeur (147800).
- Consommation, qui est *extrêmement important* (0.39) a une très bonne valeur (10.2).
- Performances, qui est *important* (0.33) a une mauvaise valeur.
- Equipement, qui est *assez important* (0.10) a une moyenne valeur (50/83).
- Transmission, qui est *peu important* (0.06) a une très bonne valeur (manuelle).
- Motrices, qui est *peu important* (0.05) a une très mauvaise valeur (traction avant).
- Dimension, qui est *peu important* (0.06) a une bonne valeur (11.4).
- Coffre, qui est *peu important* (0.06) a une bonne valeur (451).
- Esthétique, qui est *peu important* (0.08) a une bonne valeur (belle).
- Image marque, qui est *très peu important* (0.05) a une très bonne valeur (très bon).

Performances, qui est un critère *important* (0.33) a une mauvaise valeur.

La valeur de ce critère dépend des valeurs des critères:

- Puissance, qui est *important* (0.33) a une mauvaise valeur (130).
- Accélération, qui est *important* (0.30) a une moyenne valeur (31.8).
- Vitesse, qui est *important* (0.30) a une mauvaise valeur (200).
- Poids, qui est *assez important* (0.12) a une bonne valeur (1285).



UMEÅ UNIVERSITY

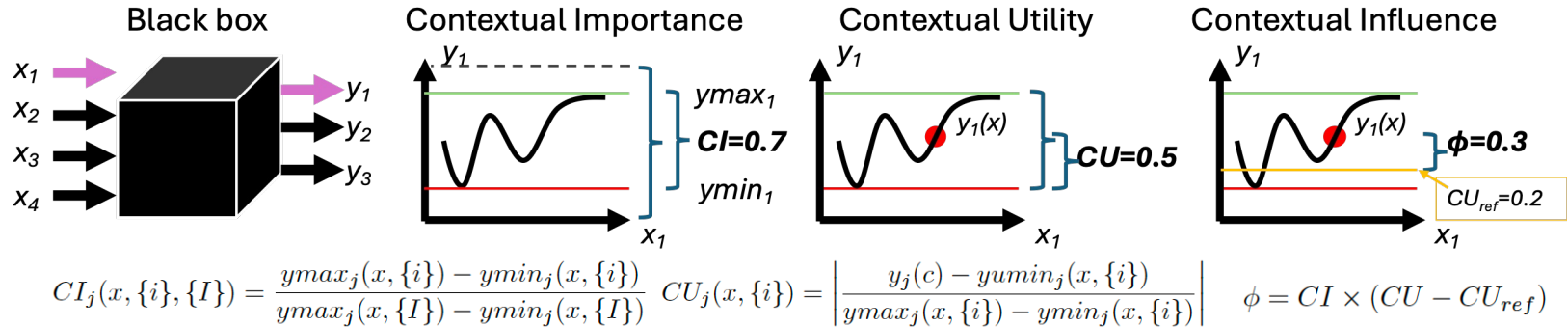
KARY FRÄMLING

CIU: CORE CONCEPTS

How much can the result change by modifying the value(s) of one or more features (jointly)?

How favorable are the current feature value(s) for getting a high-utility output?

How positive or negative is the effect of one or more features for the result/output compared to a reference utility value (or a reference instance)?



We skip detailed mathematics for now;
formal definitions are in “utility space”



UMEÅ UNIVERSITY

KARY FRÄMLING

WHAT IS “FEATURE IMPORTANCE”?

- Common-sense (not specific to ML or XAI) wording by ChatGPT:
 - *Feature importance describes how strongly a feature affects the model’s outputs*
 - *Example: Salary is typically important for an individual’s wealth*
- Contextual Importance: Same thing but for a specific context.
 - *Example: This person’s savings are huge and no debts, so salary is not important for the person’s wealth*
- Importance is in range $[0,1]$
 - Zero signifies no importance
 - 1 (100%) signifies that the result can be changed “completely” (within the possible range)
 - Assumption: The joint set (coalition) of all features has importance 1



UMEÅ UNIVERSITY

KARY FRÄMLING

WHAT IS "VALUE UTILITY"?

- Wording by ChatGPT:
 - *Value utility tells how good or useful a particular feature value is for achieving the outcome.*
 - *Example: A salary of €30,000/month is very useful for increasing wealth, while €1000/month has low utility (is a bad value).*
 - *Key intuition: How favorable is this value for a high output value (or actually a high-utility output value)?*
- Examples of values:
 - Temperature = 23
 - Engine horse powers is 350
 - Apartment price is \$250 000
- Values in context and their utility (Contextual Utility):
 - +23° Celsius is a **good** indoor temperature but **bad** in a freezer
 - 350 hp is **a lot of** power for a car but **not so much** for a truck
 - \$250 000 can be **high** or **low** in different contexts but also for different users, e.g. seller vs. buyer
- Utility is in range [0,1]



UMEÅ UNIVERSITY

KARY FRÄMLING

WHAT IS "FEATURE INFLUENCE"?

- Common-sense (not specific to ML or XAI) wording by ChatGPT:
 - *Feature influence describes whether a feature increases or decreases the prediction for a specific instance*
 - *Example: A high income may increase the predicted probability of loan approval for one applicant*
 - *Key intuition: How does this feature affect this specific prediction?*
- Influence can be positive, negative or zero.
- Influence tends to be Contextual by definition (depends on the instance)
- **Always comparative/contrastive against a reference instance**
- Contextual influence is in range $[-1,1]$

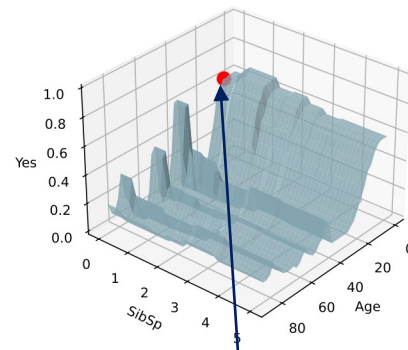
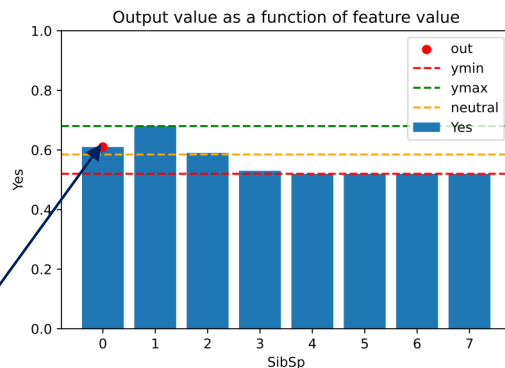
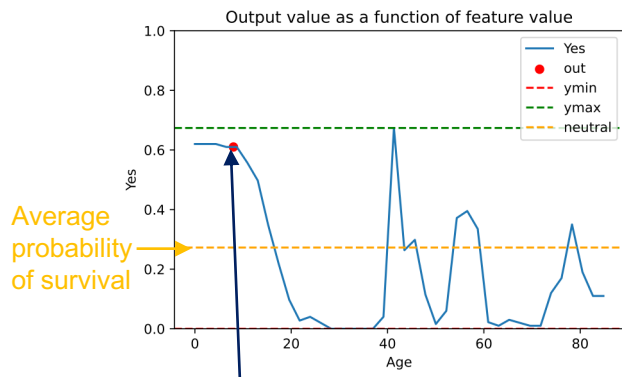


UMEÅ UNIVERSITY

KARY FRÄMLING

CIU FOR EXPLAINING THE PROBABILITY OF SURVIVAL ON THE TITANIC*

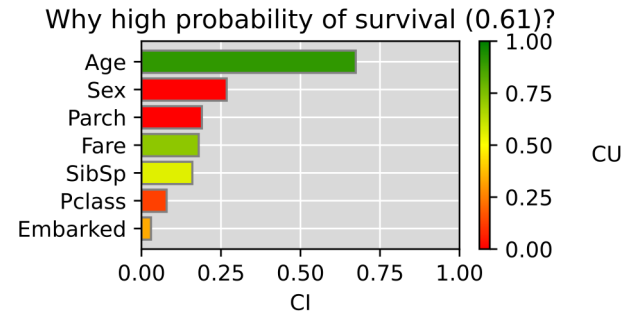
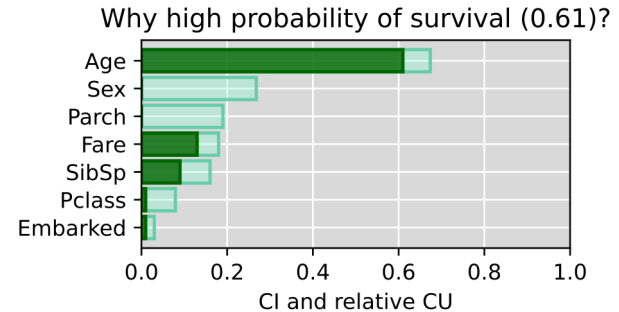
- Example: Random Forest model predicts 61.0% probability of survival on Titanic for “JohnnyD”, an 8-year old boy who travels alone
- Plot output value y_j as a function of one (or two) inputs $x_{\{i\}}$ while keeping the values of all other inputs set to $x_{-\{i\}}$



- CIU values for one feature are directly “readable” from input-output (IO) plot
- 3-D plot shows output value for two features jointly

THE POTENTIAL INFLUENCE (PI) PLOT

- Uses only importance (CI) and utility (CU):
 - Importance is indicated with a transparent bar
 - Solid bar indicates how “good” the current value is
- “Potential influence” because a big transparent area indicates potential improvement by changing the feature value
- Solid area indicates how much worse the output could become by changing feature value
- Example of use: *“What renovation of my house would increase its value the most?”*
- Utility can also be shown using colors

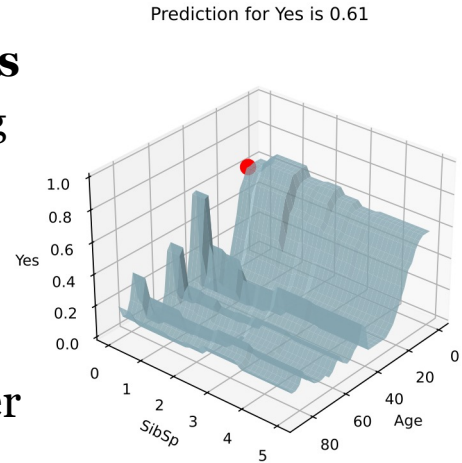


UMEÅ UNIVERSITY

KARY FRÄMLING

INTERMEDIATE CONCEPTS

- An Intermediate Concept (IC) is a **named set of features**
 - ICs are a **cornerstone concept of CIU** since the beginning
- Features in sets can be **dependent** in many ways
 - Fuzzy/binary AND, OR, XOR or “whatever” for ML systems
- **CIU is well defined** also for dependent features
- ICs can be interrelated as meronomies, hierarchies or other kinds of semantic structures, i.e. **vocabularies**
- Vocabularies can be **adapted** to context, explaineer, interaction, ...

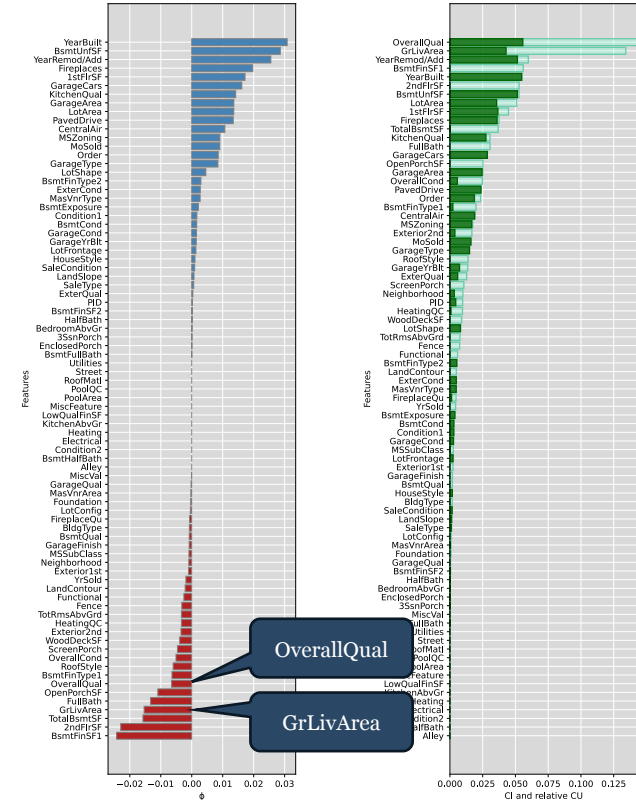


UMEÅ UNIVERSITY

KARY FRÄMLING

EXAMPLE: AMES HOUSING

- Ames housing is a data set with 2930 houses described by 81 features
- Gradient boosting model trained to predict the sale price based on the 80 other features
- With 80 features, “classical” bar plot explanations become unreadable
- Influence values may mislead the explainees: The most important features may have even zero influence!
- Many features are dependent, which causes misleading explanations because individual features have a small importance, whereas the joint importance can be significant
- IC-based vocabulary can be adapted to the context, the explainee, ...
- Simple, common-sense vocabulary used as example

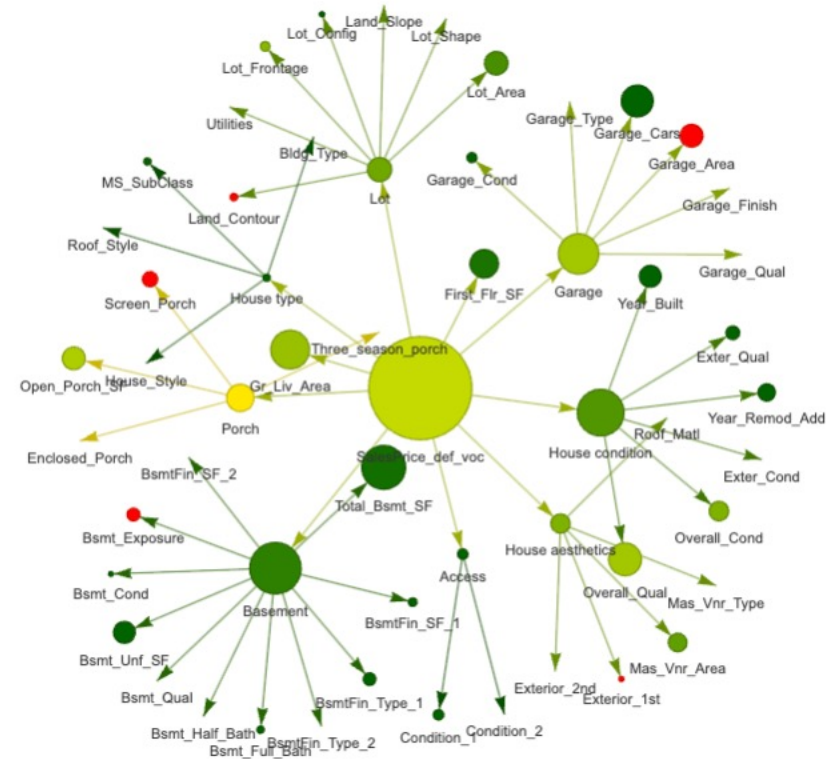


UMEÅ UNIVERSITY

KARY FRÄMLING

DEMONSTRATION

- Illustration on the right: The **Contextual Attribution Tree (CAT)** shows the concept hierarchy and the importance/utility values for all concepts in the tree
 - Circle area corresponds to CI value
 - Color illustrates CU value
 - Can also use influence values
- Implemented in **R** using the **Shiny** framework
- **Status:** Ready for release as new version of the CIU package at CRAN (The Comprehensive R Archive Network)



CONCLUDING REMARKS



UMEÅ UNIVERSITY

SOCIAL XAI: CIU ENABLES CORE CAPABILITIES

- **Interaction.** Explainees have the possibility to guide the dialog by choosing the question to ask (“Why?”, “Why not?”, “What if?”, . . .)
- Capability of **structuring explanations into appropriate chunks** that provide the necessary amount of information but not more
- Capability to **adjust the modality, abstraction level etc.** used in different explanatory moves
- Adjust explanations depending on the **context, explainee model and other information** that the AI systems might discover or learn during the interaction
- Implement a (trainable) **partner model** of the explainee
- CIU mathematics works for any input modality, even for **mixed modality**
- Etc.



UMEÅ UNIVERSITY

KARY FRÄMLING

SOME PAPERS ON CIU

- Främling, Kary. Social Explainable AI: What Is It and How to Make It Happen with CIU? In: Calvaresi, D., et al. Explainable, Trustworthy, and Responsible AI and Multi-Agent Systems. *7th International Workshop EXTRAAMAS 2025*. Lecture Notes in Computer Science, vol 15936. Springer, Cham. pp. 58–63. https://doi.org/10.1007/978-3-032-01399-6_4
- Främling, Kary. Contextual importance and utility in python: new functionality and insights with the py-ciu package. In: *XAI 2024 Workshop of 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, Jeju, South Korea, 2024. <https://arxiv.org/abs/2408.09957>
- Främling, Kary. Feature Importance versus Feature Influence and What It Signifies for Explainable AI. In: *Longo, L. (eds) xAI 2023: Explainable Artificial Intelligence. Communications in Computer and Information Science book series (CCIS, volume 1901)*. Springer, Cham. pp. 241–259. <https://arxiv.org/abs/2308.03589>
- Främling, Kary. Counterfactual, Contrastive, and Hierarchical Explanations with Contextual Importance and Utility. In: *Calvaresi, D., et al. Explainable and Transparent AI and Multi-Agent Systems. EXTRAAMAS 2023. Lecture Notes in Computer Science, vol 14127*. Springer, Cham. pp. 180–184. <https://rdcu.be/dnTo9>
- Främling, Kary. Contextual Importance and Utility: a Theoretical Foundation. In: Long G., Yu X., Wang S. (eds) *AI 2021: Advances in Artificial Intelligence. AI 2022. Lecture Notes in Computer Science, vol 13151*. Springer, Cham. pp. 117–128. https://link.springer.com/content/pdf/10.1007/978-3-030-97546-3_10.pdf?pdf=inline%20link
- Knapic, Samanta, Malhi, Avleen, Saluja, Rohit, Främling, Kary. Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. *Machine Learning & Knowledge Extraction*, Vol. 3, Issue 3, 2021.pp. 740–770. <https://doi.org/10.3390/make3030037>
- Främling, Kary. Decision Theory Meets Explainable AI. In: *Calvaresi D., Najjar A., Winikoff M., Främling K. (Eds.): EXTRAAMAS 2020, LNAI 12175*, pp. 57–74, 2020. Springer Nature Switzerland AG. https://link.springer.com/content/pdf/10.1007/978-3-030-51924-7_4.pdf



UMEÅ UNIVERSITY

KARY FRÄMLING

TRY OUT FOR YOURSELF

- Open-source implementations published on GitHub:
 - <https://github.com/KaryFramling/py-ciu> (Python, tabular data)
 - <https://github.com/KaryFramling/ciu> (R, tabular data)
 - https://github.com/KaryFramling/py_ciu_image (Python, images)
 - <https://github.com/KaryFramling/ciu.image> (R, images; obsolete)
- Ongoing for other data types such as time series, natural language, ...



UMEÅ UNIVERSITY

KARY FRÄMLING

THANK YOU!



UMEÅ UNIVERSITY