EXPLAINABLE AI: WHAT IS IT AND WHAT CAN IT DO?

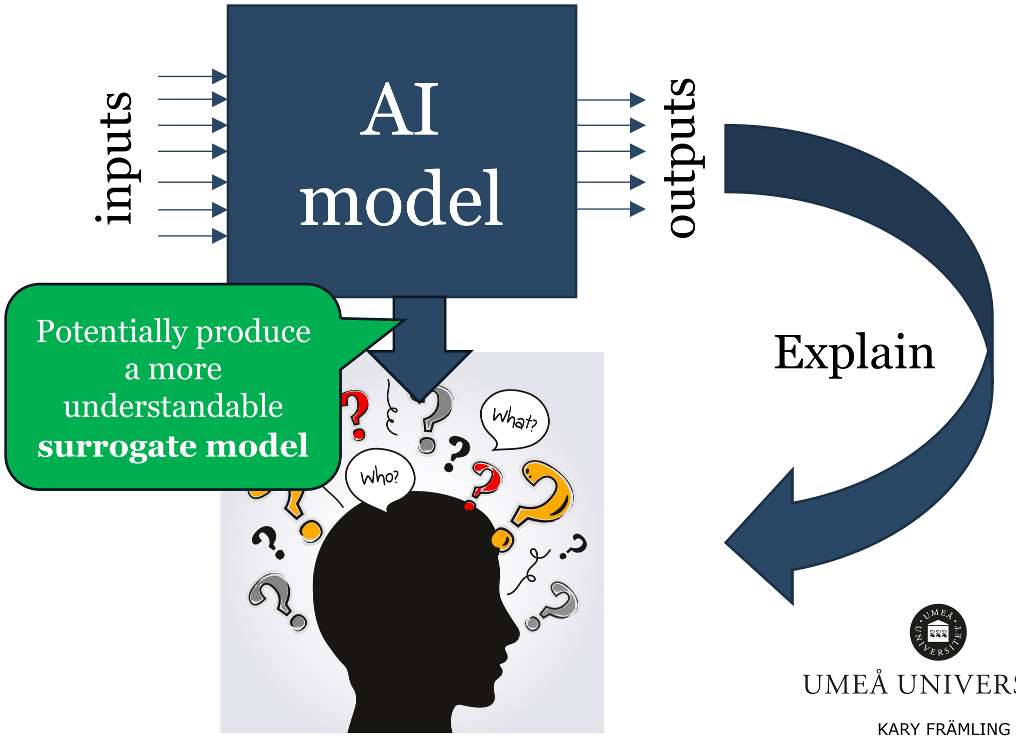
Kary Främling

Professor in Data Science

Head of Explainable AI Team

UMEÅ UNIVERSITY

WHAT IS XAI?



- **Inputs ⇒ outputs** examples:
  - Income situation ⇒ credit grading
  - Image pixels ⇒ probabilities of different objects
  - Sensor signals ⇒ probabilities of different diagnostics
  - Text, sound, …
- **AI models**: Rule base, decision tree, random forest, neural network, …
- **XAI is not only for Machine Learning models** – but the XAI community seems to think it is only for ML
- What is understandable for the **AI engineer** may not be understandable for **expert** and even less for **end-user**
- **Interpretable(/transparent?)**: can be understood, interpreted and explained by expert
- **Explainable**: can produce end-user understandable explanations
- No consensus on meaning of Interpretable/Explainable

ROUGH TIMELINE OF XAI

Kary develops Contextual Importance and Utility (CIU) method[2,3]

CIU revived

Rule-based expert systems

Neural Nets

AI rises again

MYCIN

PROLOG

Semantic nets

Black-box explainability

Explainable AI (XAI) concept[1]

~1972 ~1975 End of 1980’s 1990’s Mid-1990’s 2015 2016

[1] https://www.cc.gatech.edu/~alanwags/DLAI2016/%28Gunning%29%20IJCAI-16%20DLAI%20WS.pdf
(but XAI seems to have been proposed as a name also earlier)

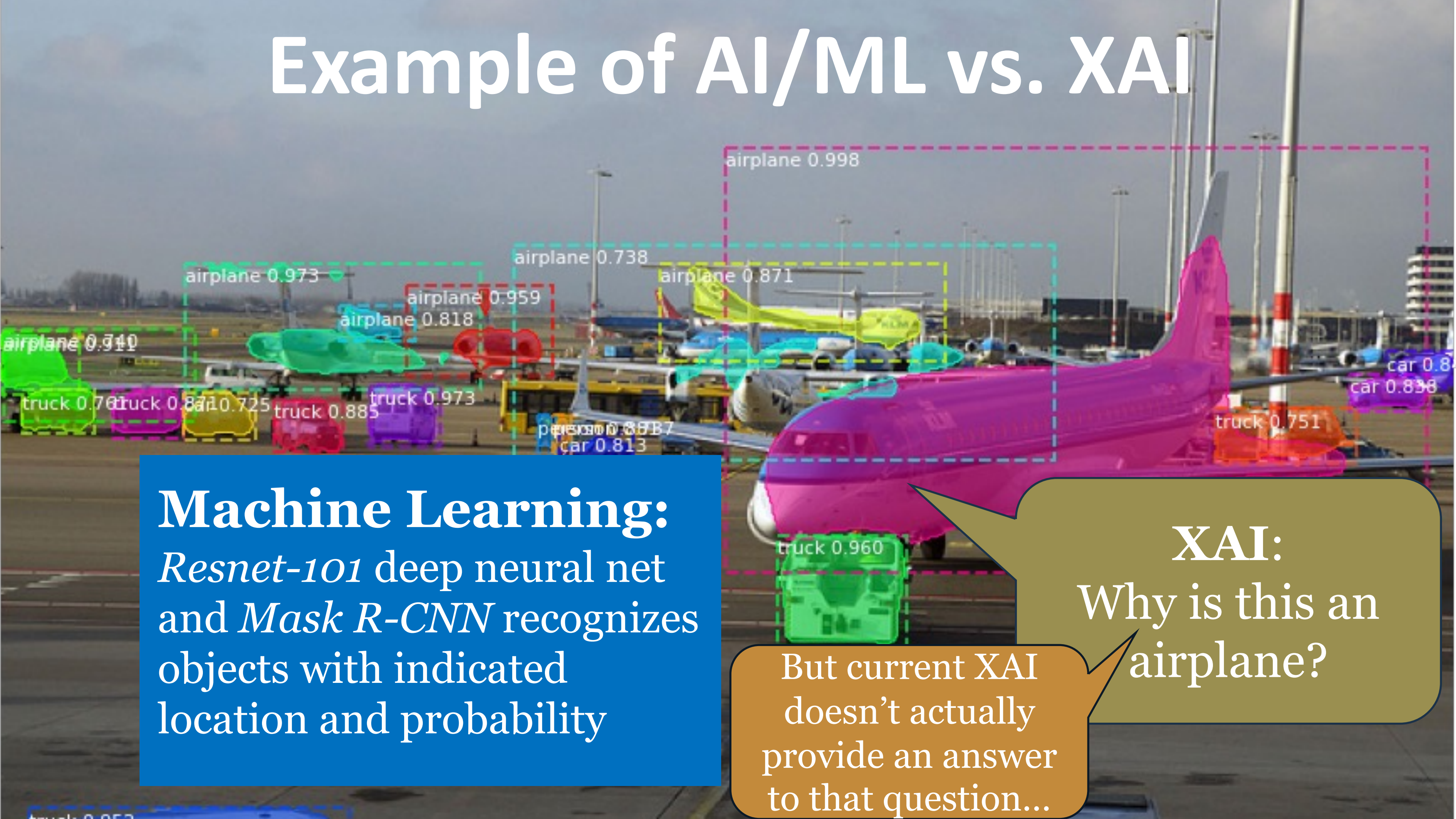
[2] *Främling, Kary. Explaining Results of Neural Networks by Contextual Importance and Utility. In: Robert Andrews and Joachim Diederich (eds.), Rules and networks: Proceedings of the Rule Extraction from Trained Artificial Neural Networks Workshop, AISB’96 conference, 1-2 April 1996. Brighton, UK, 1996.*
[3] *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.*

SOME CHOICES TO MAKE IN XAI

- Use “white-box” or “black-box” models?
- Model-agnostic vs. Model-specific?
  - White box XAI methods are always model-specific (or are they?)
- Expose the inner workings of the AI system as rules or similar (“global”) or only what lead to a specific result (“local”)?
- Use surrogate models or not?
  - Then there’s a question of fidelity, accuracy, …
  - Why not use non-surrogate methods, such as CIU, Counterfactual?
- Can explanations be produced in reasonable time, and with what precision?

UMEÅ UNIVERSITY

Example of AI/ML vs. XAI



Machine Learning:
Resnet-101 deep neural net and *Mask R-CNN* recognizes objects with indicated location and probability

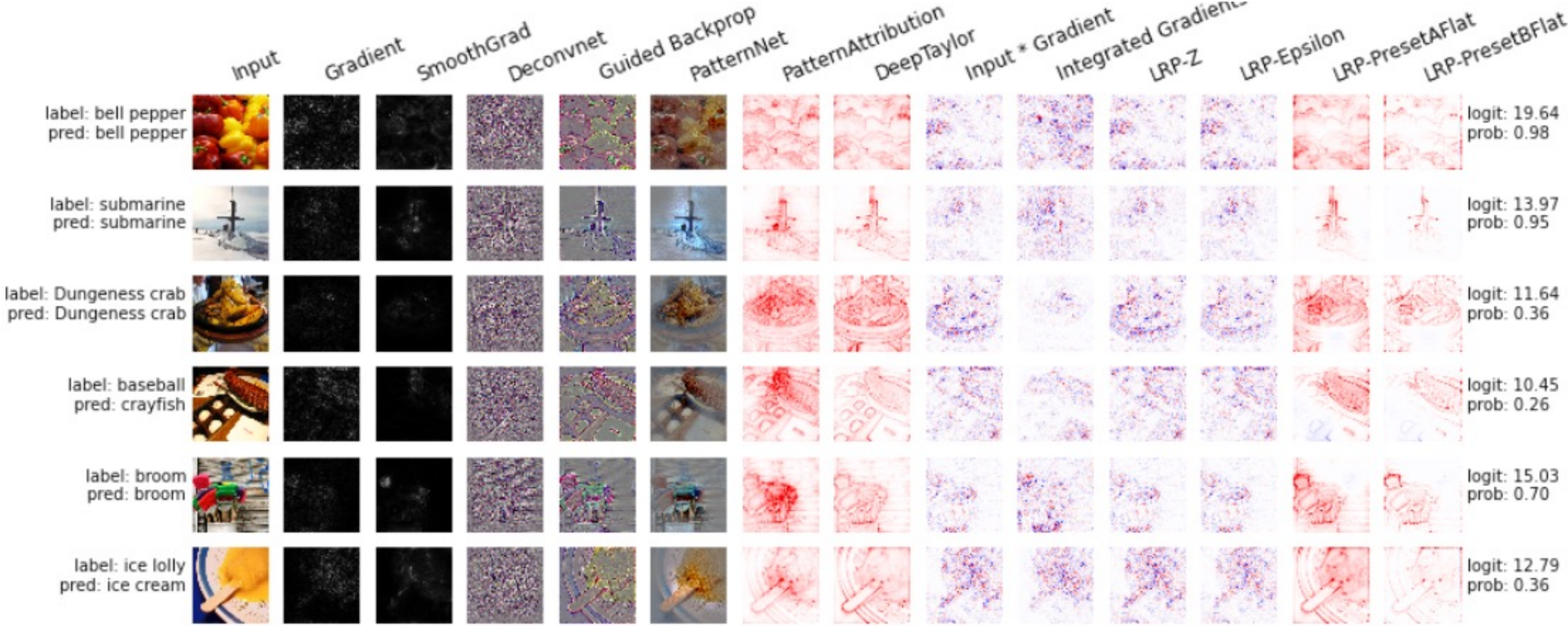
XAI:
Why is this an airplane?

But current XAI doesn’t actually provide an answer to that question…

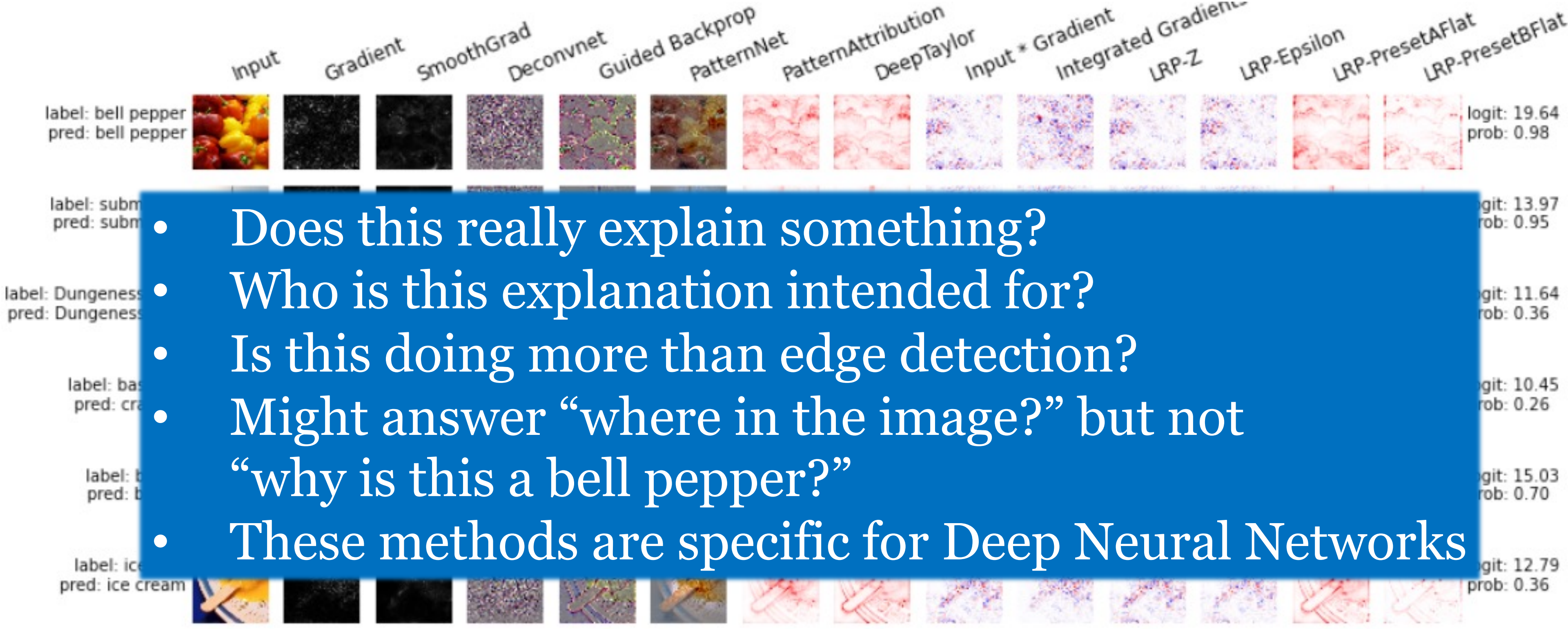
CURRENT STATE-OF-THE-ART IN XAI

UMEÅ UNIVERSITY

XAI FOR EXPLAINING IMAGE CLASSIFICATION



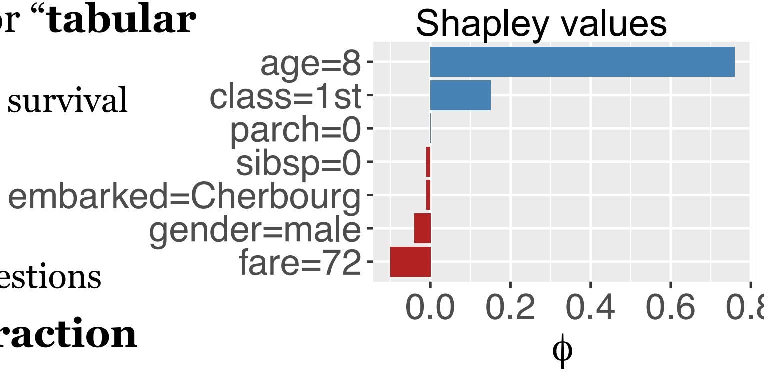
XAI FOR EXPLAINING IMAGE CLASSIFICATION



- Does this really explain something?
- Who is this explanation intended for?
- Is this doing more than edge detection?
- Might answer "where in the image?" but not "why is this a bell pepper?"
- These methods are specific for Deep Neural Networks

STATE OF THE ART IN XAI (SIMPLIFIED)

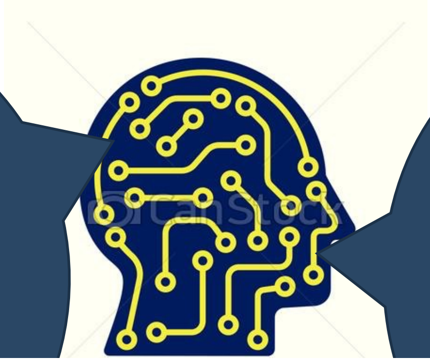
- XAI for **Image classification**:
  - GradCAM, LRP, LIME, SHAP, CIU etc.
  - Mainly answer “where?” question, rather than “why?”
- **Feature Influence** (rather than ”importance”) methods are presumably the most popular ones for the moment for “**tabular data**”
  - Shapley values, LIME etc. Example: Explain 63.6% survival probability of 8-year old boy ‘Johnny D’ on Titanic
- New trends:
  - Generate **rules** (actually the oldest XAI approach)
  - **Counterfactual**: answer “what-if?” and ”how-to?” questions
- Explanations tend to be given as such, **no user interaction**



WHO ARE EXPLANATIONS INTENDED FOR?

- End-user (patient) versus expert (doctor) perspective (versus others…)
- Different vocabularies, levels of detail for different targets

The primary source in PA is the capsule of the glenohumeral mechanism; the treatment intervention would be to improve extensibility of the capsule, whereas in case of impingement, the source lies in the scapula-thoracic mechanism altering the scapular mechanics.



Black-box AI

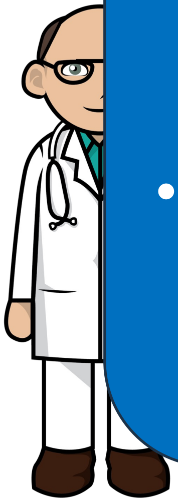
You'll be ok after 3 days of rest, no permanent damage

Interpretable

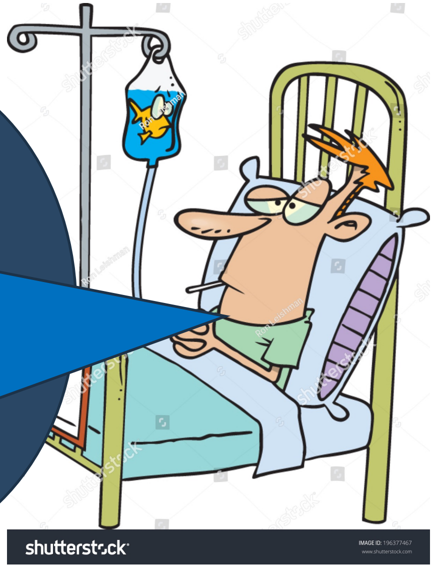
Explainable

WHO ARE EXPLANATIONS INTENDED FOR?

- En… (versus others…)
- …s



- Current XAI methods and research seem to ignore this person (and usually also ignore the expert)
- XAI sometimes seems to be about ML researchers developing debugging tools for ML researchers



SOCIAL EXPLAINABLE AI (SXAI)

UMEÅ UNIVERSITY

WHAT IS SOCIAL XAI?

- Seminar "Social Explainable AI: Designing Multimodal and Interactive Communication to Tailor Human–AI Collaborations" in Shonan, Japan, September 2023
  - An attempt to allow communication between sociologists, psychologists, philosophers, computer scientists, AI researchers, …
  - https://shonan.nii.ac.jp/seminars/200/
- XAI 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 digestable "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

EXAMPLE: 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

EXAMPLE: KARY JUSTIFIES CAR CHOICE TO HIS 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 . . .

UMEÅ UNIVERSITY

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.
Why not?	*Wouldn't car B be better?*
Why Not answer	It's expensive and consumes a lot of fuel .
Refusal	*But I like car B more*
Question	Why do you like car B more?
(Human) answer	*I like its shape and I preferred driving with it*
New proposal	Ok, I took that into account. Then car C might actually be even better
Reaction	*Indeed, I had forgotten car C*
Why this and not that?	*So why would car C be better than car B?*
Contrastive answer	Because . . .
<. . . >	<interaction continues>

UMEÅ UNIVERSITY

SOME TYPICAL QUESTIONS ASKED BY HUMANS

XAI

- **Why?** Get a justification for the current output or result of an AI system.
- **Why not?** A “why not?”” explanation presumably emphasizes features with a negative influence
- **Why is this feature important (or not)?** Answering this question requires that the XAI method is not a black box itself
- **What if?** A “counterfactual” question, i.e. what would happen if the values of one or more features change?
- **Why A and not B?** Contrastive question where the answer emphasizes the most important features that differentiate A and B.
- **How confident are you about your outcome (and explanation)?** Again a valid question for Social XAI…
- **How?** The explainee might want a more extensive answer about the model's training, the data set used etc.
- **I don't agree with the outcome, nor the justification or explanation provided! How can I correct that?** This goes beyond the capabilities of current XAI methods but is relevant for Social (X)AI.

SOME TYPICAL QUESTIONS ASKED BY HUMANS

Social XAI

- **Why?** Get a justification for the current output or result of an AI system.
- **Why not?** A “why not?”” explanation presumably emphasizes features with a negative influence
- **Why is this feature important (or not)?** Answering this question requires that the XAI method is not a black box itself
- **What if?** This is the basic counterfactual question, i.e. what would happen if the values of one or more features change?
- **Why A and not B?** Contrastive question where the answer emphasizes the most important features that differentiate A and B.
- **How confident are you about your outcome (and explanation)?** Again a valid question for Social XAI...
- **How?** The explainee might want a more extensive answer about the model's training, the data set used etc.
- **I don't agree with the outcome, nor the justification or explanation provided! How can I correct that?** This goes beyond the capabilities of current XAI methods but is relevant for Social (X)AI.

SOCIAL XAI: WHAT ELSE TO CONSIDER?

- **Interaction**. Explainees should 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.
- Explainer needs to have a (trainable) **partner model** of the explainee in many real-world cases in order to have a successful interaction.
- The explainer has to have a **model of the context** for any explanation that goes beyond highlighting the pixels in an image.
- Etc.

HOW CAN WE MAKE IT HAPPEN?

… by using Contextual Importance and Utility (CIU)

UMEÅ UNIVERSITY

INSPIRATION BEHIND CIU: HUMAN BRAIN IS BIGGEST BLACK BOX IN THE WORLD…

- … but we still manage to explain our “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
- Why wouldn’t we be able to do the same with XAI?

UMEÅ UNIVERSITY

CONTEXTUAL IMPORTANCE AND UTILITY (CIU)

- Developed by Kary Främling during his Diplôme d'Études Approfondies[1] (D.E.A., 1991-1992) and Ph.D. Thesis[2] (1992-1996), both in French
- Papers in English:
  - FRÄMLING, Kary. *Explaining Results of Neural Networks by Contextual Importance and Utility*. In: Robert Andrews and Joachim Diederich (eds.), Rules and networks: Proceedings of the Rule Extraction from Trained Artificial Neural Networks Workshop, AISB'96 conference, 1-2 April 1996. Brighton, UK, 1996. Download as PDF.
  - FRÄMLING, Kary, GRAILLOT, Didier. *Extracting Explanations from Neural Networks*. ICANN'95 proceedings, Vol. 1, Paris, France, 9-13 October, 1995. Paris: EC2 & Cie, 1995. pp. 163-168.
- Break from 1996 to 2017, then received AI professorship at Umeå University, Sweden
  - During the break: Internet of Things, Digital Twins, Intelligent Products, Smart houses, Smart Cities, …
- Result: "nobody" remembered or found CIU when XAI became fashionable in around 2017
- Coincided with Kary's professorship in Data Science at Umeå University in November 2017
- The time was right, Kary had the time and the XAI approaches proposed were far from satisfactory (according to Kary), so CIU was revived

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

UMEÅ UNIVERSITY

WHAT IS "FEATURE IMPORTANCE"?

Common-sense (not specific to ML or XAI) definition by HuggingChat*:

- *Feature importance can be understood as the significance or relevance of a particular characteristic or attribute* ***in a given context****.*
- *Feature importance refers to the degree to which a specific characteristic or property of something contributes to its overall value, performance, or outcome. It highlights the significance or influence of that particular feature* ***in relation to the whole****.*
- Importance is typically indicated in the range [0,1]

* *meta-llama/Meta-Llama-3.1-70B-Instruct*

WHAT IS "FEATURE INFLUENCE"?

Common-sense (not specific to ML or XAI) definition by HuggingChat:

- *Feature influence refers to the causal or explanatory effect that a specific characteristic or property has on an outcome, behavior, or phenomenon. It highlights how **changes or variations** in a particular feature can impact or influence the resulting outcome.*
- This implies having some *"reference instance"* that changes/variations are applied to
- LIME method produces influence values by indicating whether changes to current instance values increase or decrease the output value. LIME values are relative.
- SHAP method's influence values indicate whether the current instance's feature values increase or decrease the output compared to a reference instance. SHAP values are relative but their sum is known.
- Influence can be positive, negative or zero.

UMEÅ UNIVERSITY

WHAT IS “VALUE UTILITY”?

- Explanation given by HuggingChat:
  - *Value utility refers to the **perceived usefulness**, benefits, or advantages that an entity, such as a product, service, idea, or action, provides **to its users, consumers, or stakeholders**. It represents the extent to which something satisfies needs, fulfills desires, or delivers value, making it desirable or beneficial to those who interact with or possess it.*
- Examples of values:
  - Temperature = 23
  - Engine horse powers is 350
  - Apartment price is $250 000
- Values in context and their 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 typically indicated in the range [0,1]

UMEÅ UNIVERSITY

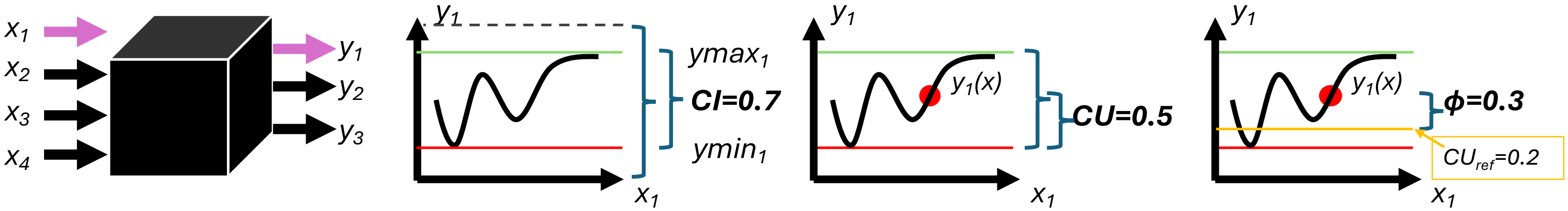
CIU: CORE CONCEPTS

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)?

Black box | Contextual Importance | Contextual Utility | Contextual Influence



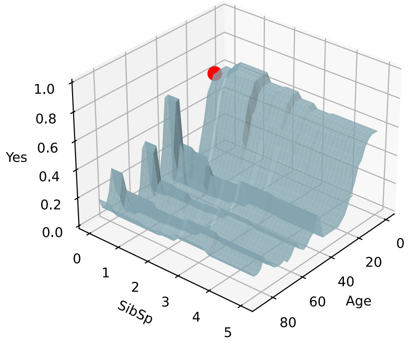
$$CI_j(x,\{i\},\{I\}) = \frac{ymax_j(x,\{i\}) - ymin_j(x,\{i\})}{ymax_j(x,\{I\}) - ymin_j(x,\{I\})} \quad CU_j(x,\{i\}) = \left|\frac{y_j(c) - yumin_j(x,\{i\})}{ymax_j(x,\{i\}) - ymin_j(x,\{i\})}\right| \quad \phi = CI \times (CU - CU_{ref})$$

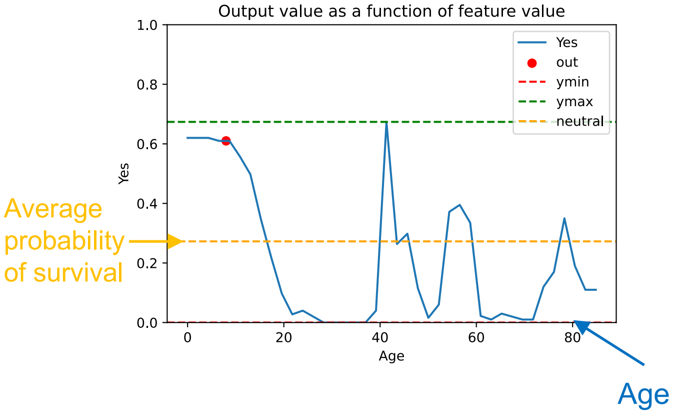
We skip the mathematics for now…

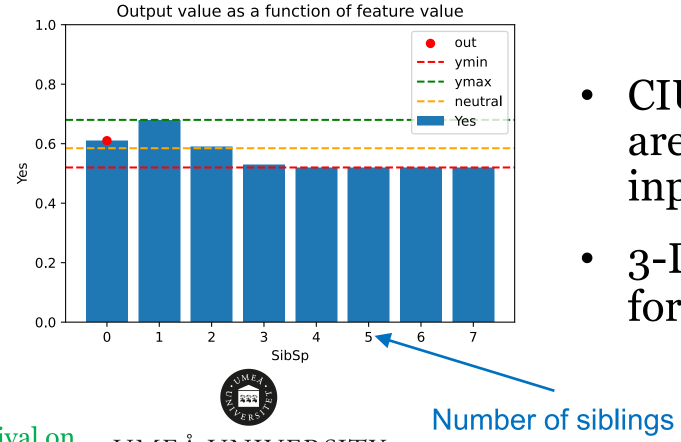
UMEÅ UNIVERSITY

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 $x_{\neg\{i\}}$





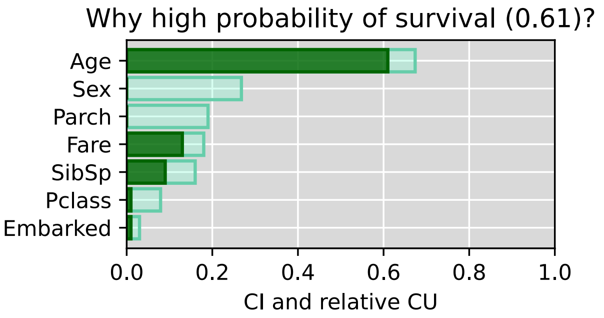


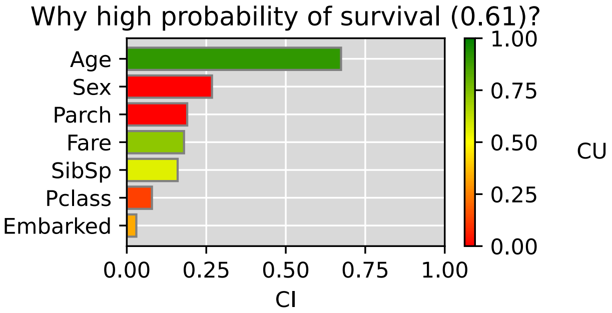
- CIU values for one feature are directly "readable" from input-output (IO) plot
- 3-D plot shows output value for two features jointly

* Learning to estimate the probability of survival on the Titanic based on the actual data is a classical Machine Learning benchmark task.

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

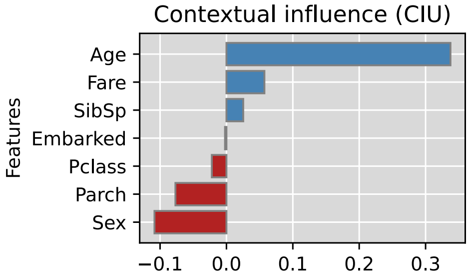
TEXTUAL EXPLANATION

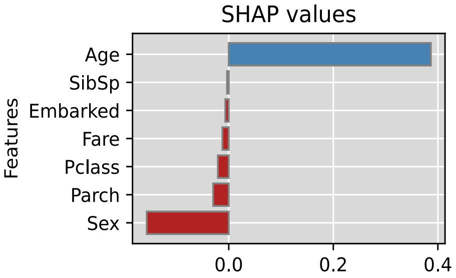
The explained value is ***Yes*** with the value 0.61 (CU=0.61), which is **higher than average utility**.
Feature *Pclass* has **very low importance (CI=0.08)** and has value(s) 1.0, which is **low utility (CU=0.13)**
Feature *Sex* has **low importance (CI=0.27)** and has value(s) 1.0, which is **low utility (CU=0.00)**
Feature *Age* has **high importance (CI=0.67)** and has value(s) 8.0, which is **high utility (CU=0.91)**
Feature *SibSp* has **very low importance (CI=0.16)** and has value(s) 0.0, which is **higher than average utility (CU=0.56)**
Feature *Parch* has **very low importance (CI=0.19)** and has value(s) 0.0, which is **low utility (CU=0.00)**
Feature *Fare* has **very low importance (CI=0.18)** and has value(s) 72.0, which is **higher than average utility (CU=0.72)**
Feature *Embarked* has **very low importance (CI=0.03)** and has value(s) 3.0, which is **lower than average utility (CU=0.33)**

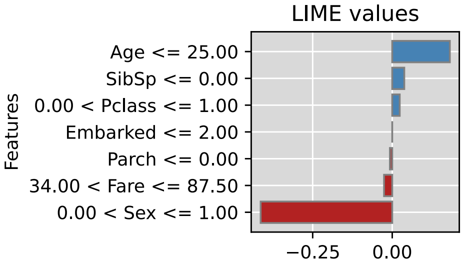
- Since CI $\in [0,1]$ and CU $\in [0,1]$ their values are meaningful as such
- Current CIU implementations use a simple template-based approach
- The use of textual explanations has not been elaborated or compared with other modalities, even though CIU initially produced textual explanations

UMEÅ UNIVERSITY

INFLUENCE-BASED EXPLANATIONS





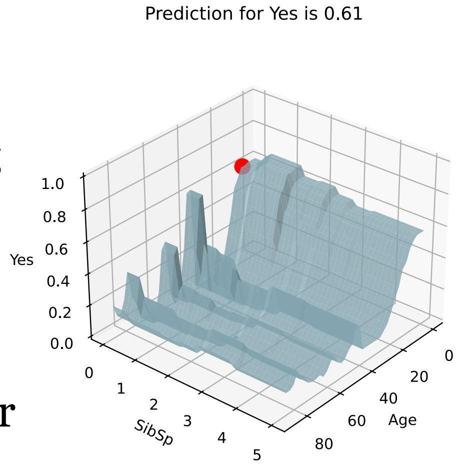


- Somewhat similar for the three XAI methods but not identical, which one to trust?
  - At least Contextual influence can be understood and validated from IO-plots and is consistent with PI plots, textual explanations etc.
- Even important features may have small or even zero influence, which can be misleading
  - For instance, the number of siblings (SibSp) and number of accompanying parent/children (Parch) give an impression of being quite insignificant, even though the PI-plot shows that they are quite important
  - Features with values that are close to the reference instance values will have small influence, by definition!

UMEÅ UNIVERSITY

INTERMEDIATE CONCEPTS

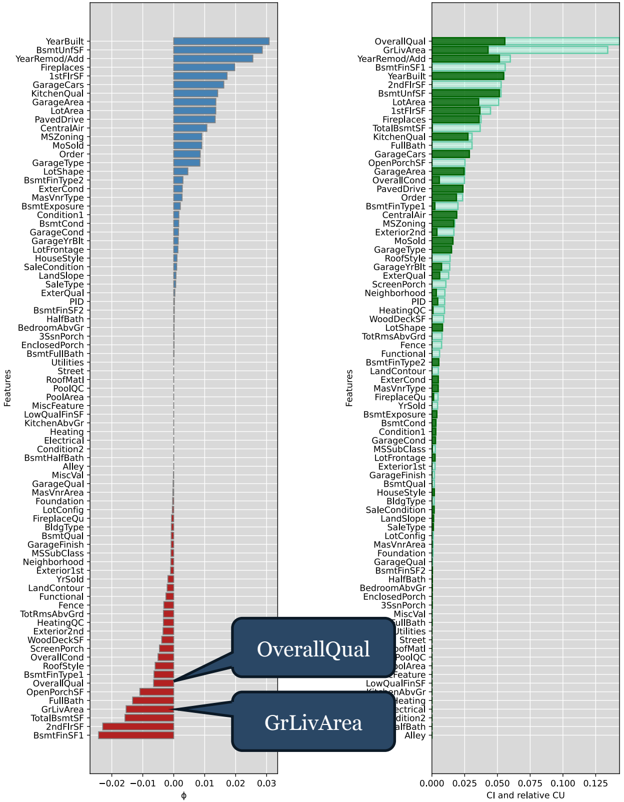
- An Intermediate Concept (IC) is a **named coalition of features**
  - ICs are a **cornerstone concept of CIU** since the beginning
- Features in coalitions can be **dependent** in many ways
  - Fuzzy/binary AND, OR, XOR or “whatever” for ML systems
- **CIU is “well-defined”** also for dependent features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



UMEÅ UNIVERSITY

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

EXAMPLE VOCABULARY FOR HOUSE EXPLANATIONS (AMES)

- House
  - House condition
    - Overall quality
    - Year built
    - Overall condition
    - Year of remodeling
    - External quality
    - External condition
  - Basement
    - Total basement surface
    - Surface of unfinished basement
    - Surface of finished basement Type 1
    - Surface of finished basement Type 2
    - Presence of Type 1 finished basement surface
    - Presence of Type 2 finished basement surface
    - Quality of basement
    - Condition of basement
    - Full bathroom in basement
    - Half bathroom in basement
    - Basement exposure
  - Garage
    - ...
  - Lot
    - ...
  - Access
    - ...
  - House type
    - ...
  - House aesthetics
    - ...
  - First floor surface
  - Above ground living area

Meronomy used for explanations

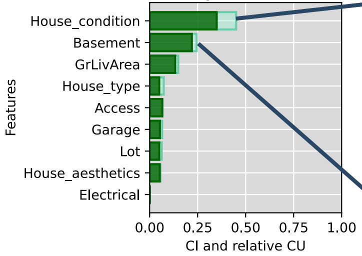
Same in Python code

```
ames_voc = {
  "Garage":[c for c in df.columns if ‘Garage' in c],
  "Basement":[c for c in df.columns if ‘Bsmt' in c],
  "Lot":list(df.columns[[3,4,7,8,9,10,11]]),
  "Access":list(df.columns[[13,14]]),
  "House_type":list(df.columns[[1,15,16,21]]),
  "House_aesthetics":list(df.columns[[
   22,23,24,25,26]]),
  "House_condition":list(df.columns
   [[20,18,21,28,19,29]]),
  "First_floor_surface":list(df.columns[[43]]),
  "Above_ground_living area":
   [c for c in df.columns if 'GrLivArea' in c]
}
```

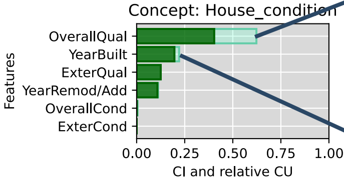
UMEÅ UNIVERSITY

"SOCIAL" XAI

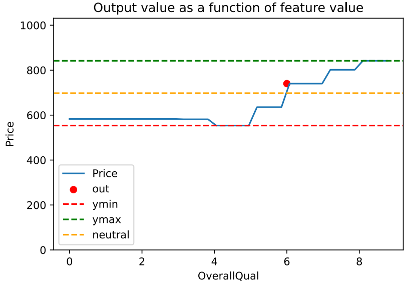
Why is this house expensive ($740 222)?



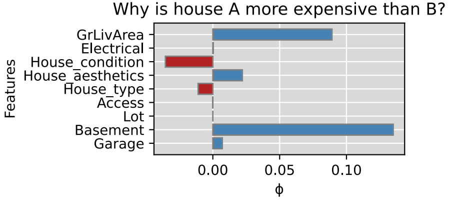
Why is house condition important and good?



Why is Overall Quality important and average??



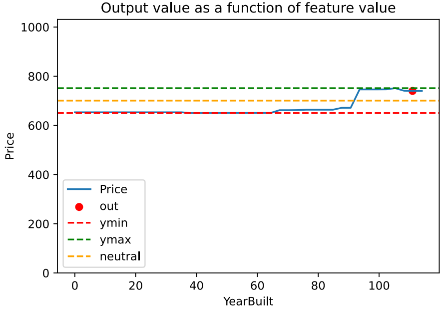
Why is house A more expensive than house B ($568 000)?



Why is Basement important and good?

The explained value is ***Basement*** for output ***Price***.
Feature *BsmtQual* has **very low importance (CI=0.06)** and has value(s) 0.0, which is **high utility (CU=1.00)**
Feature *BsmtCond* has **very low importance (CI=0.02)** and has value(s) 5.0, which is **high utility (CU=1.00)**
Feature *BsmtExposure* has **very low importance (CI=0.02)** and has value(s) 1.0, which is **high utility (CU=1.00)**
Feature *BsmtFinType1* has **very low importance (CI=0.05)** and has value(s) 2.0, which is **high utility (CU=1.00)**
Feature *BsmtFinSF1* has **normal importance (CI=0.45)** and has value(s) 941.0, which is **high utility (CU=0.94)**
Feature *BsmtFinType2* has **very low importance (CI=0.02)** and has value(s) 6.0, which is **low utility (CU=0.00)**
Feature *BsmtFinSF2* has **very low importance (CI=0.01)** and has value(s) 0.0, which is **lower than average utility (CU=0.32)**
Feature *BsmtUnfSF* has **very low importance (CI=0.15)** and has value(s) 81.0, which is **higher than average utility (CU=0.68)**
Feature *TotalBsmtSF* has **low importance (CI=0.39)** and has value(s) 847.0, which is **high utility (CU=0.93)**
Feature *BsmtFullBath* has **very low importance (CI=0.19)** and has value(s) 1.0, which is **high utility (CU=1.00)**
Feature *BsmtHalfBath* has **very low importance (CI=0.00)** and has value(s) 1.0, which is **low utility (CU=0.00)**

Why is Year Built important and good??

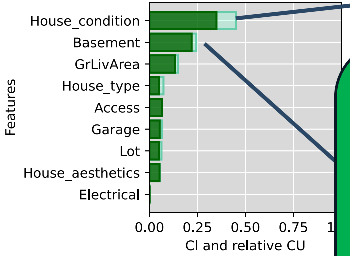


UMEÅ UNIVERSITY

"SOCIAL" XAI

How is "Overall Quality" defined?
(from meta-data, not CIU)

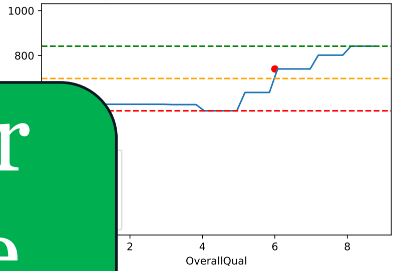
Why is this house expensive ($740 222)?



Why is house condition important and good?

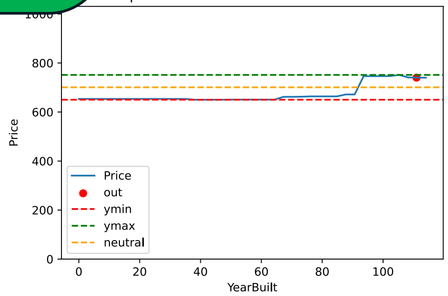
Concept: House_condition

Why is Overall Quality important and average??

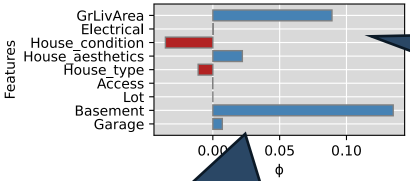


Enables "social" or at least interactive XAI

Why is Year Built important and good??



Why is house A more expensive than house B ($568 000)?



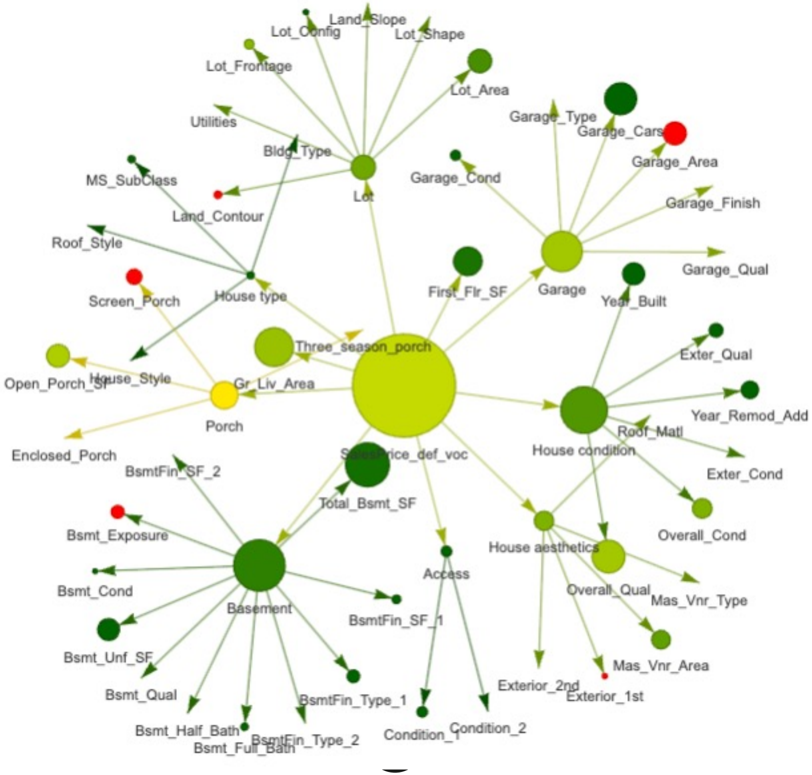
Feature *BsmtFinType1* has **very low importance (CI=0.05)** and has value(s) 2.0, which is **high utility (CU=1.00)**
Feature *BsmtFinSF1* has **normal importance (CI=0.45)** and has value(s) 941.0, which is **high utility (CU=0.94)**
Feature *BsmtFinType2* has **very low importance (CI=0.02)** and has value(s) 6.0, which is **low utility (CU=0.00)**
Feature *BsmtFinSF2* has **very low importance (CI=0.01)** and has value(s) 0.0, which is **lower than average utility (CU=0.32)**
Fe... **...han average utility (CU=0.68)**
F... **(CU=0.93)**
... **(CU=1.00)**
... **(CU=0.00)**

Alternative: How does house A compare to average price house?

Contrastive

UMEÅ UNIVERSITY

DEMONSTRATION



UMEÅ UNIVERSITY

WHY SHOULD YOU FORGET WHAT YOU JUST SAW?

UMEÅ UNIVERSITY

CIU DOES NOT EXIST IN THE "XAI WORLD"

- Tens of XAI survey papers have been published over the last years but hardly any even mention CIU, even though CIU has presumably been around the longest (by far)
- Why could that be? No clue but some theories:
    - XAI researchers do not want a "new kid in town" (or a "dark horse"?) that might disrupt current SOTA too much (and maybe the status of current XAI "thought leaders")?
    - "Being scientific" is about complex mathematics, rather than solving real problems?
    - More comfortable to find challenges with current approaches, rather than solving them?
    - (X)AI research prefers using benchmark tasks with a small number of features and avoid involving "real users"?
    - XAI researchers don't want humans to mix up their mathematics?
    - General effect of path dependency?

HOW ABOUT CIU IN INDUSTRY?

- Collected comments:
  - We only use SOTA, no new methods
  - Old saying from 1980's: "Nobody has ever been fired for buying IBM", i.e. "do like everyone else and you are safe"
  - Real question asked: "Everyone else is using SHAP (LIME, …), why are you doing something different?"

UMEÅ UNIVERSITY

RECENT PAPERS ON CIU

- 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 Corea, 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/dnT09

- 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

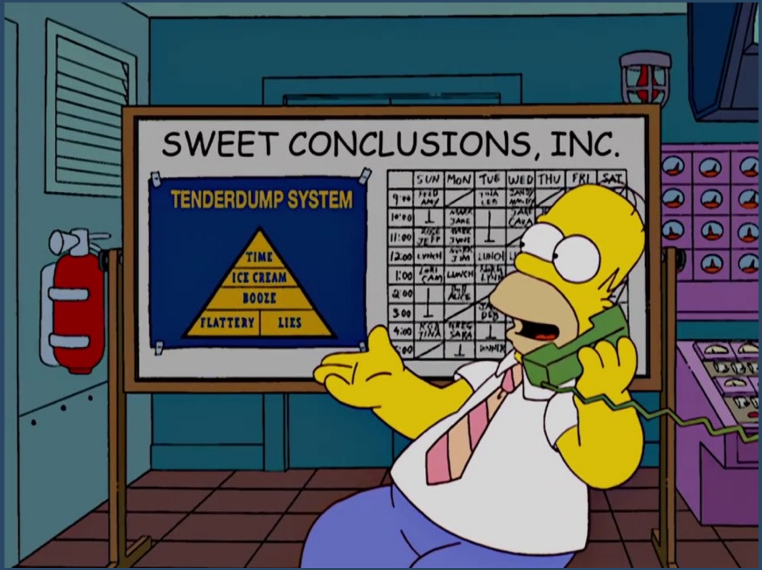
UMEÅ UNIVERSITY

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, …
- Book project "Social Explainable AI" to be published in 2025
  - https://link.springer.com/book/9789819652891
  - Extensive book written by tens of authors from different disciplines
  - Among other things, describes how partner models, context models, dialogs etc. can be implemented based on a combination of CIU, weighted Knowledge Graphs and adaptive learning of user preferences (which was shown in the demonstration)

UMEÅ UNIVERSITY

THANK YOU!



UMEÅ UNIVERSITY