On the Tradeoff Between Correctness and Completeness in Argumentative Explainable AI

Nico Potyka, Xiang Yin and Francesca Toni

1st International Workshop on Argumentation for eXplainable AI (ArgXAI)

European Research Council
Established by the European Commission

Imperial College London

What is a good Explanation?

Black-Box Classifier

School Bus

A **plausible** explanation
is **not necessarily**
a **correct** explanation

Black-Box Classifier

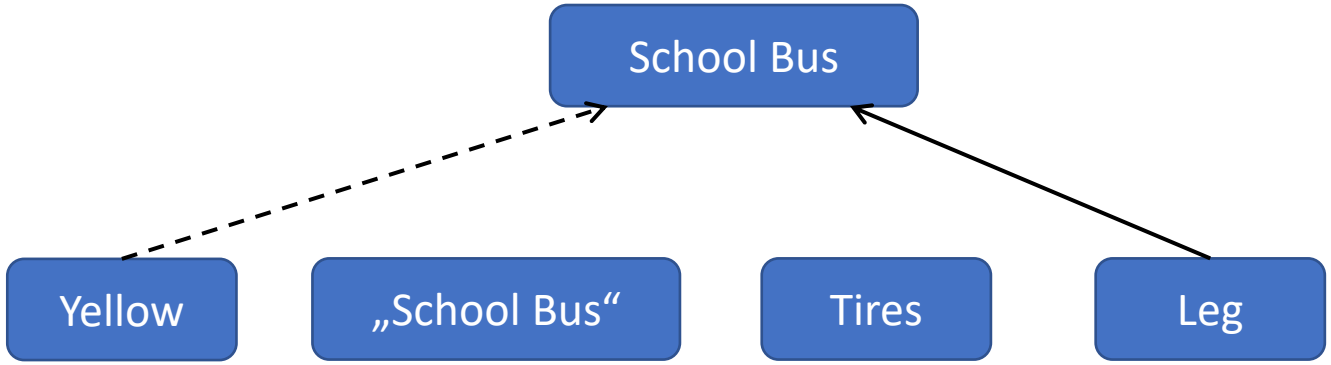
Ostrich

What makes an Explanation Trustworthy?

- **Faithfulness**: explanation explains what the model actually does

  (which ,unfortunately, is not necessarily what we want it to do)

- Instantiation for argumentative explanations: Reinforcement [1,2]
  - Supporter: should increase confidence in class
  - Attacker: should decrease confidence in class



Black-Box Classifier

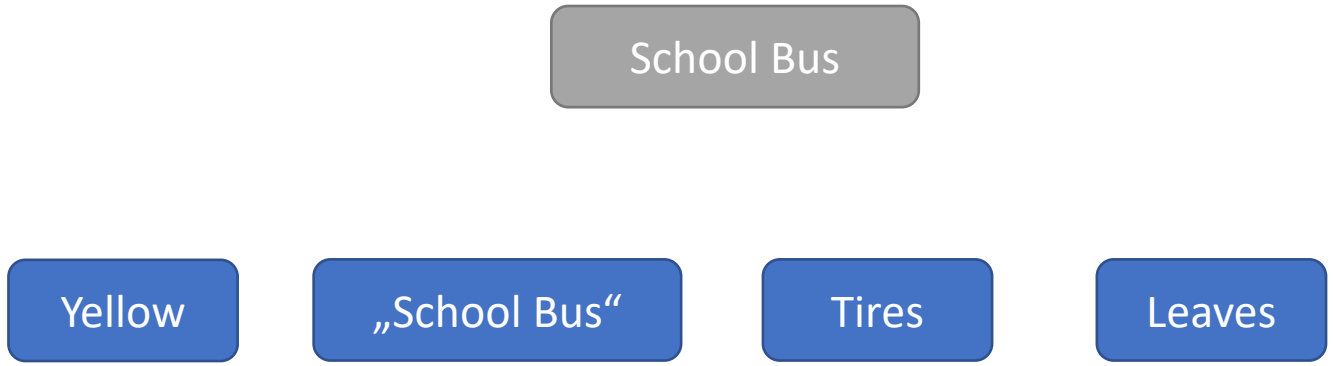
Class	P(Class)
School Bus	0.9
Ostrich	0.05
Other	0.05

1: Amgoud L, Ben-Naim J. Weighted Bipolar Argumentation Graphs: Axioms and Semantics. In: Lang J, editor. International Joint Conference on Artificial Intelligence, IJCAI. ijcai.org; 2018. p. 5194-8

2: Rago A, Baroni P, Toni F. Explaining Causal Models with Argumentation: the Case of Bi-variate Reinforcement. In: KR; 2022. p. TBA.

Potential Problems

- Faithfulness/Reinforcement can be seen as correctness property
- Problem: correctness can be satisfied in trivial ways



Correct Explanation BAGs

for boolean data

Setting

- Focus on tabular data

Age	Income	Education	Approve

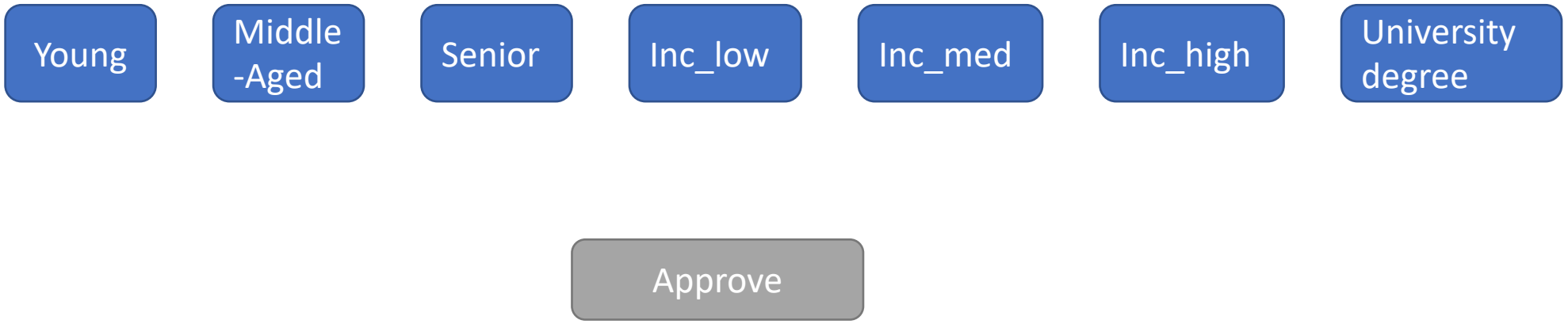
- And, for now, boolean features

Young	Middle-Aged	Senior	Inc_low	Inc_med	Inc_high	University degree	Approve

Naive Classification Arguments

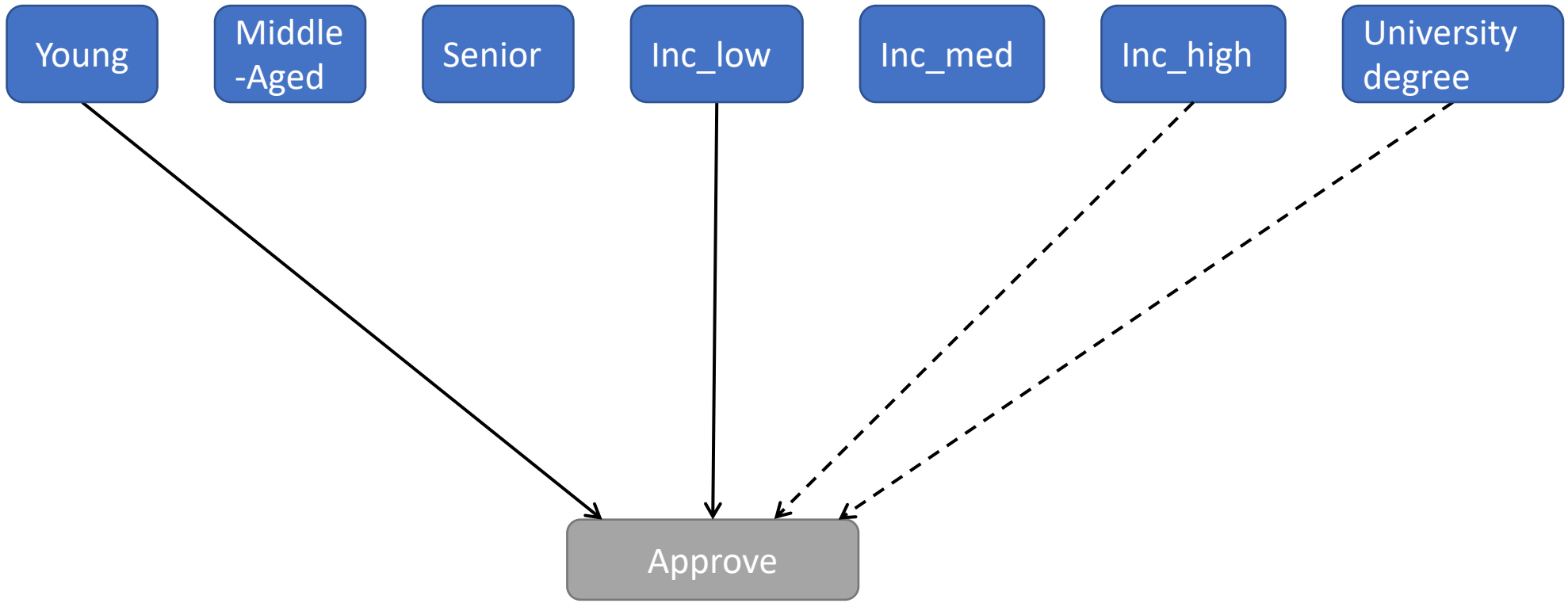
- Create one argument per feature and class

Young	Middle-Aged	Senior	Inc_low	Inc_med	Inc_high	University degree	Approve



Classification BAG

- Classification BAG is formed by adding support and attack edges

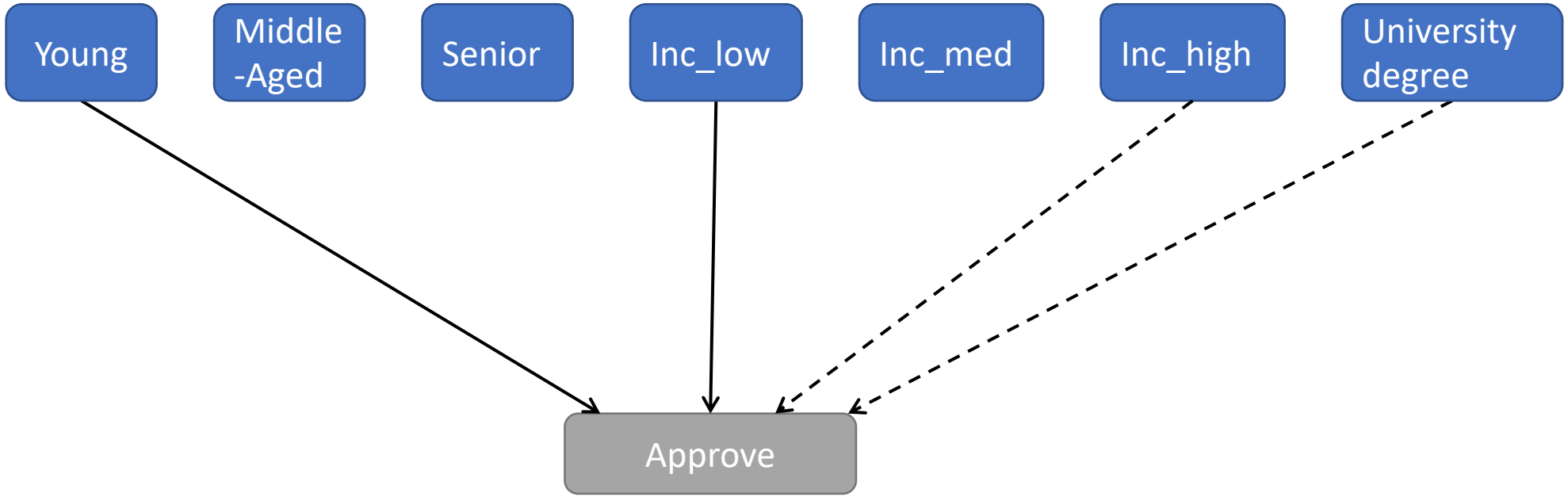


Reinforcement

- Classification BAG satisfies reinforcement if
  - Supporter increases confidence in class
  - Attacker decreases confidence in class

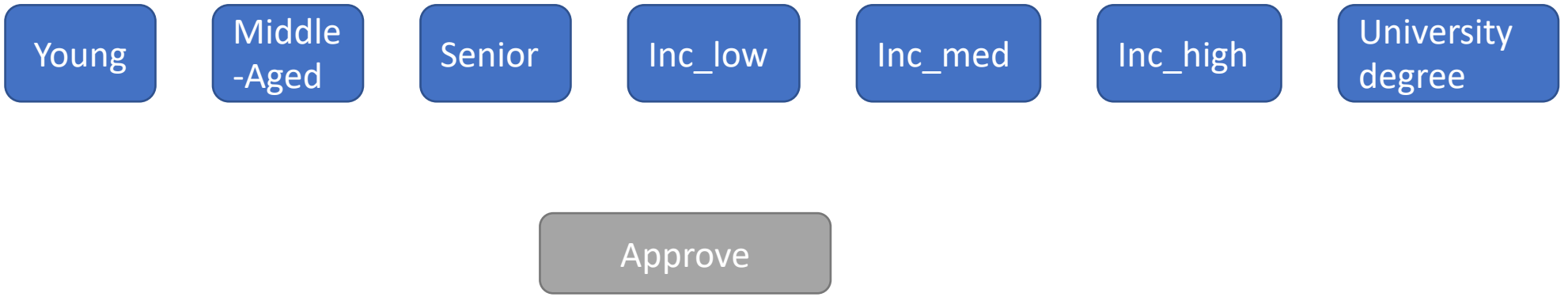
Black-Box Classifier

Class	P(Class)
School Bus	0.9
Ostrich	0.05
Other	0.05



Correctness Alone is not Meaningful

- The empty graph is correct/faithful/satisfies reinforcement



- Even when adding all edges that respect reinforcement, the graph may miss many important relationships

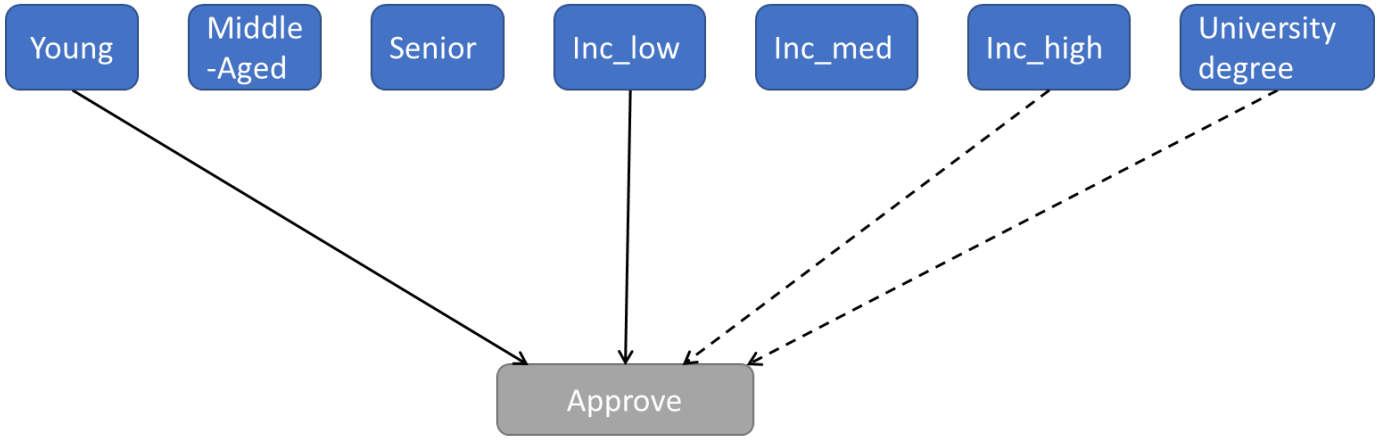
Completeness of Explanation BAGs

What does Completeness mean?

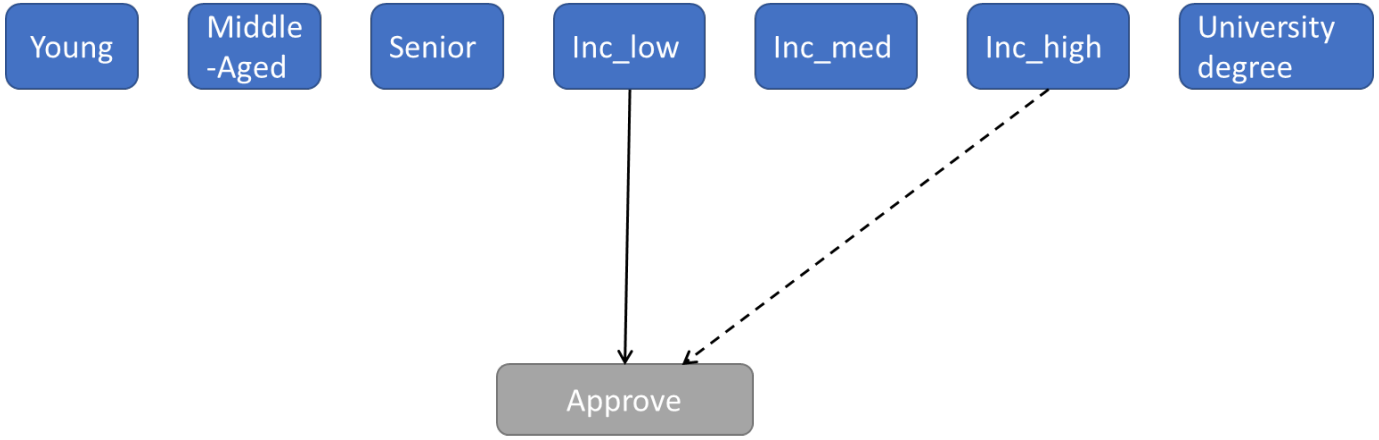
- Defining **completeness** is difficult

- Defining (and eliminating) sources of **incompleteness** is easier
  - Joint effects of features
  - Non-monotonic effects of ordinal features (supporting in some/ attacking in other regions)
  - Combinations of the two

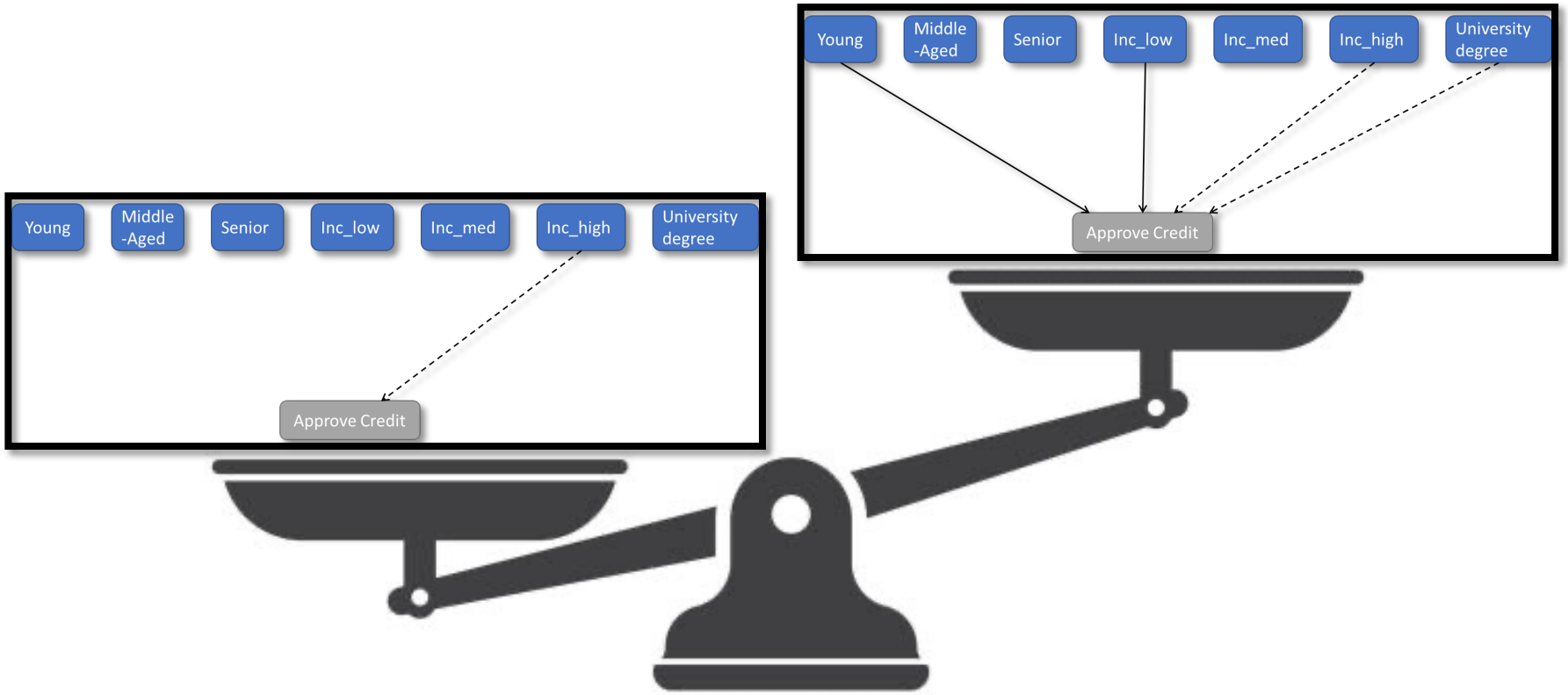
Source of Incompleteness: Joint Effects



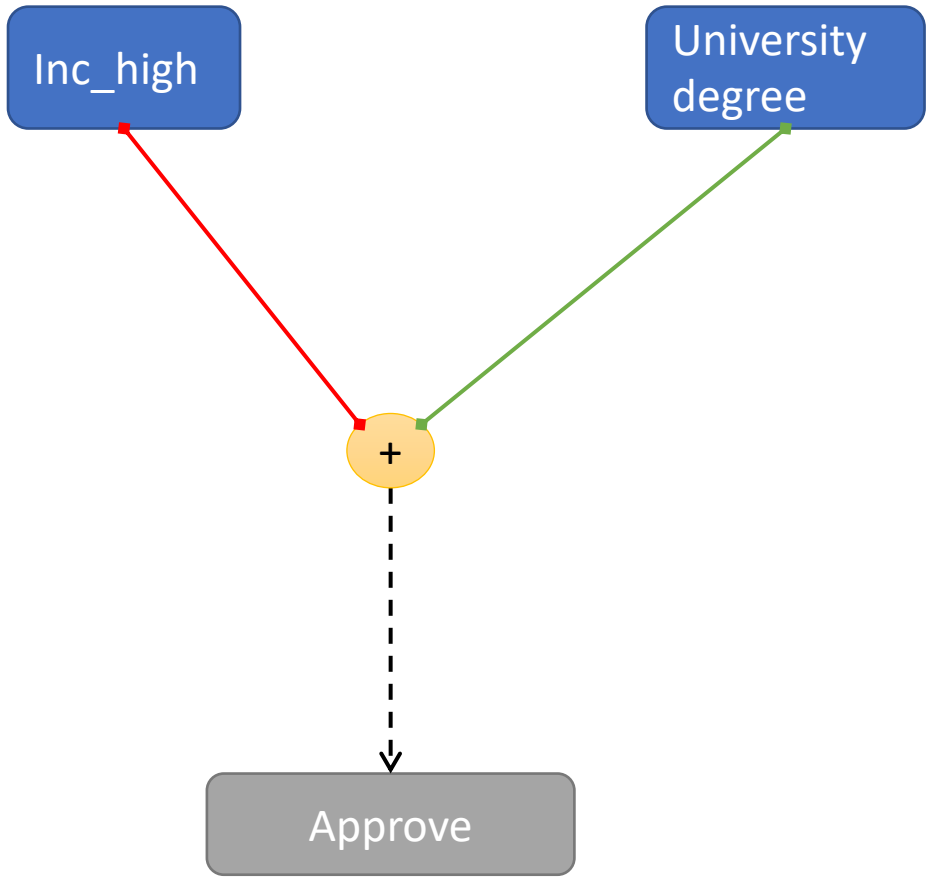
- Example: If the income is high, the other features are irrelevant



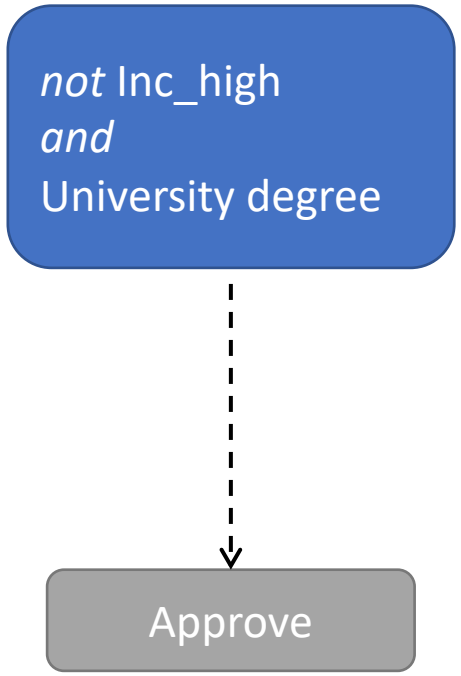
Correctness/Completeness Tradeoff



Tackling Joint Effects: Joint Relations

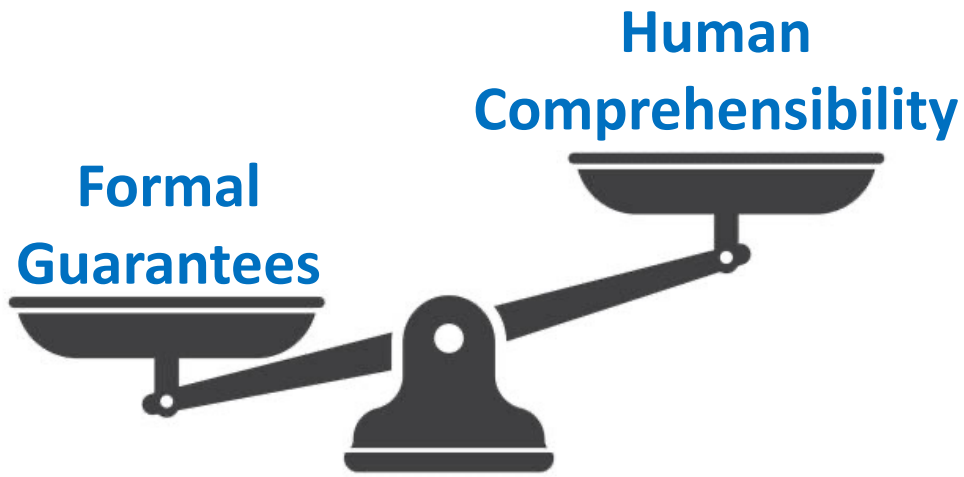


Tackling Joint Effects: Joint Arguments



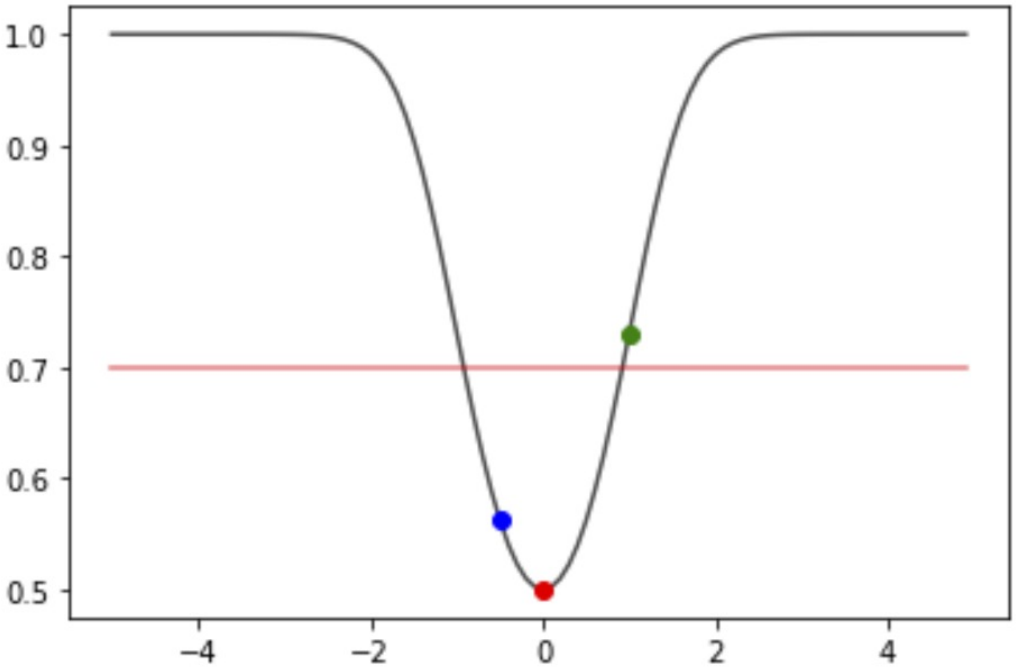
Potential Limitation

- Additional structure may improve overall correctness/completeness, but can result in less comprehensible explanations

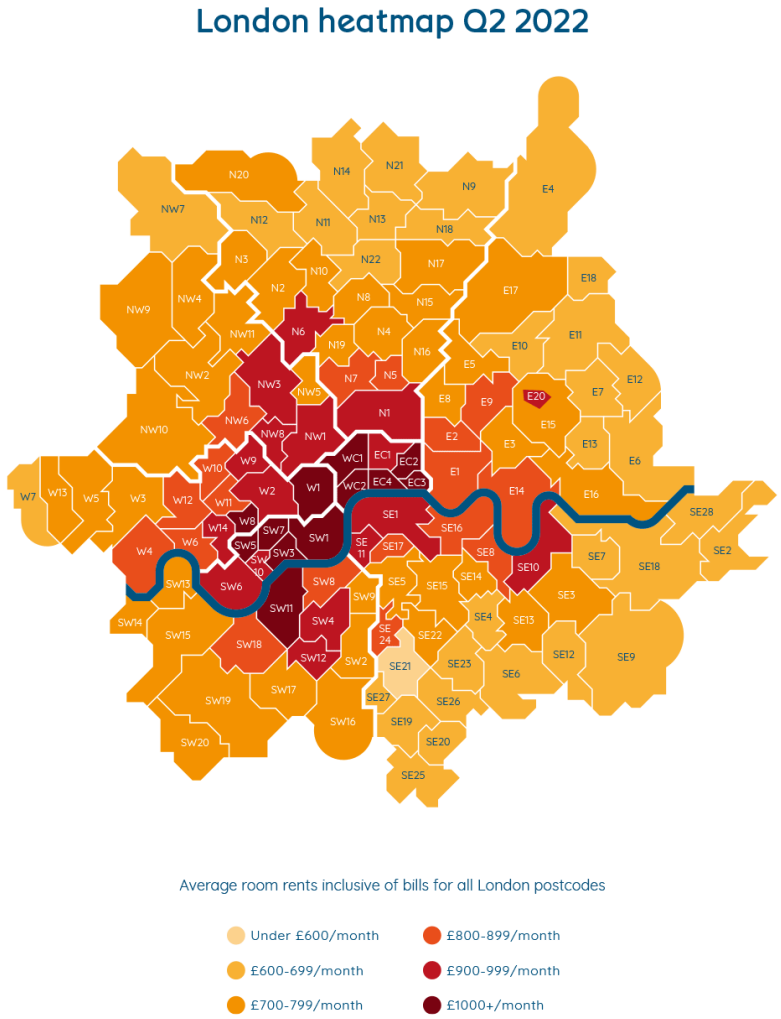


Source of Incompleteness: Non-Monotonicity

P(Anomaly)

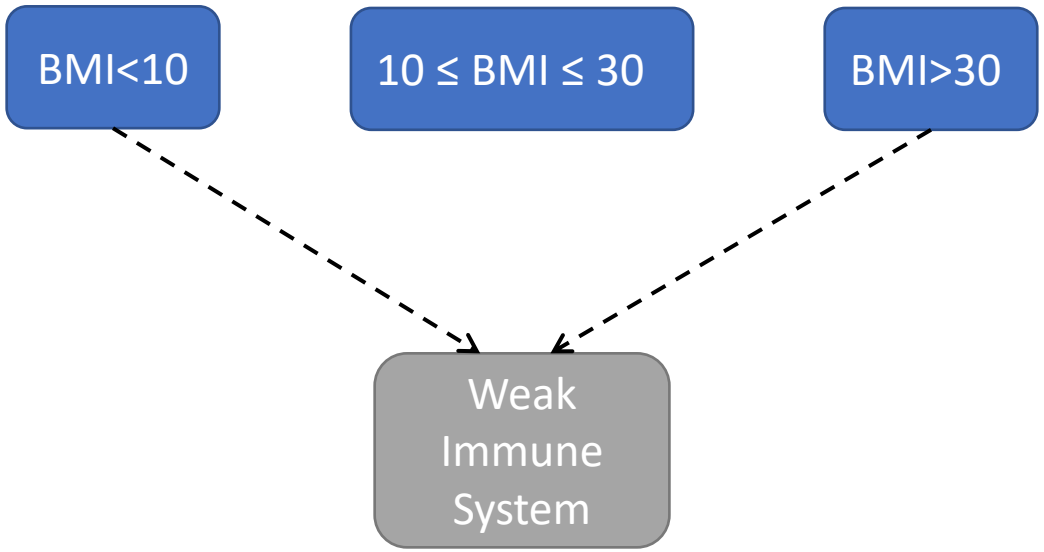


Feature Deviation from Mean



Tackling Non-Monotonicity: Binning

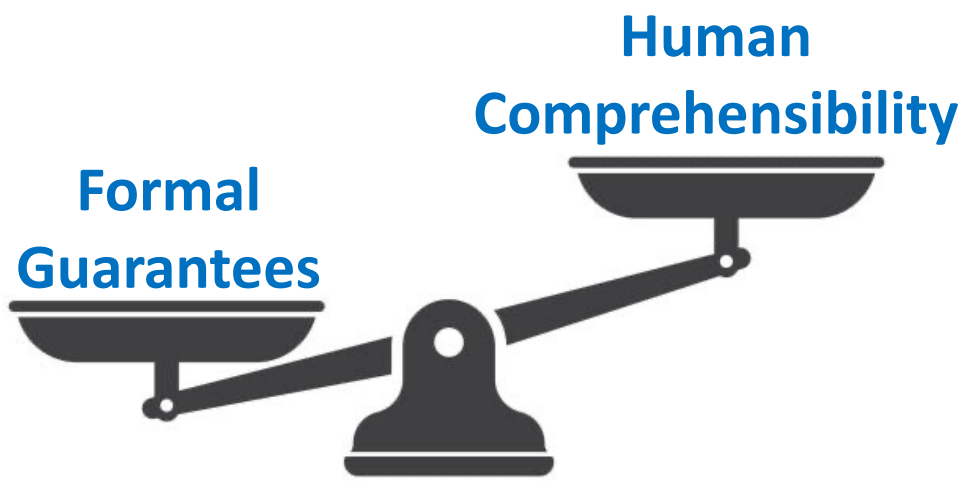
- Refining arguments (binning) can again help to improve correctness/completeness tradeoff



Conclusions and Future Work

Conclusions

- Focussing on correctness (reinforcement/faithfulness) alone does not seem sufficient for explainable AI
- More structure can help improving the Correctness/Completeness tradeoff…
- …but too much details may result in incomprehensibility



Some Interesting Questions

- Can we characterize which classifiers can be correctly and completely explained by which argumentative explanation models?
  - Conjecture: „naive“ explanation BAG can satisfy both “correctness“ and „completeness“ if and only if the classifier is „strongly monotonic“
- For which classifiers and argumentative explanation models, can we quantify correctness/completeness (efficiently)?
- Which building blocks are „most comprehensible to humans“ and which are „most effective in improving“ correctness/completeness?