Arg-XAI: a Tool for Explaining Machine Learning Results

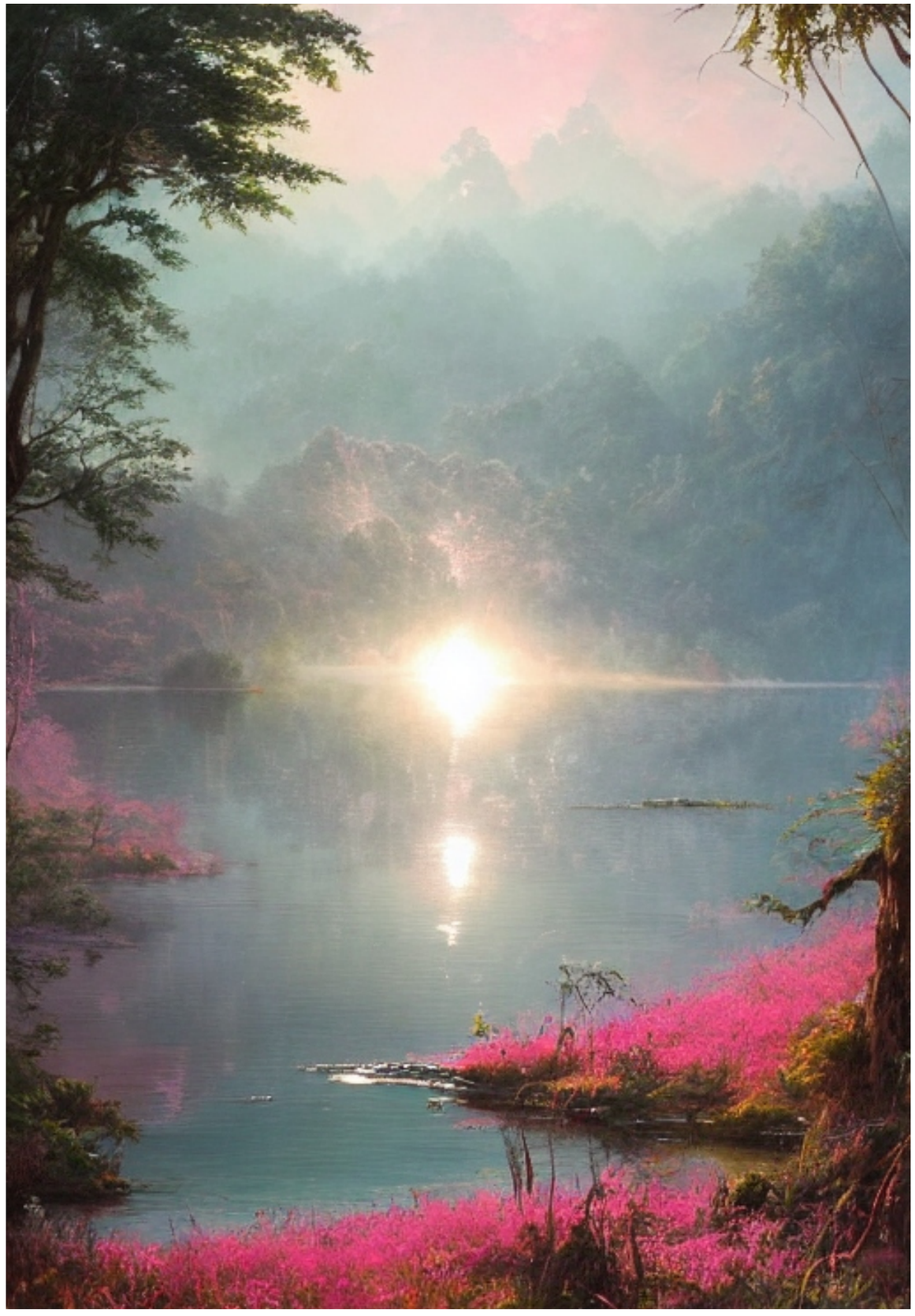
Stefano Bistarelli, Francesco Santini,
Alessio Mancinelli, Carlo Taticchi

Ital-IA 2023

A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA

Overview

- Introduction
- Arg-XAI
- Demo
- Validation
- Conclusion + FW



Titanic: a Class Prediction Example

1. Get a dataset
2. Train the model
3. ???
4. Profit

Feature	Values	Type	Description
Pclass	$1, 2, 3$	categorical	Ticket class
sex	$0, 1$	categorical	passenger gender
SibSp	$0 - 8$	categorical	# of siblings/spouses
Parch	$0 - 6$	categorical	# of parents/children
Embarked:	C, Q, S	categorical	port of embarkation
Survived:	$0, 1$	categorical	passenger survival
Age	$0.17 - 76$	numerical	passenger age
Fare	$0 - 512$	numerical	passenger fare

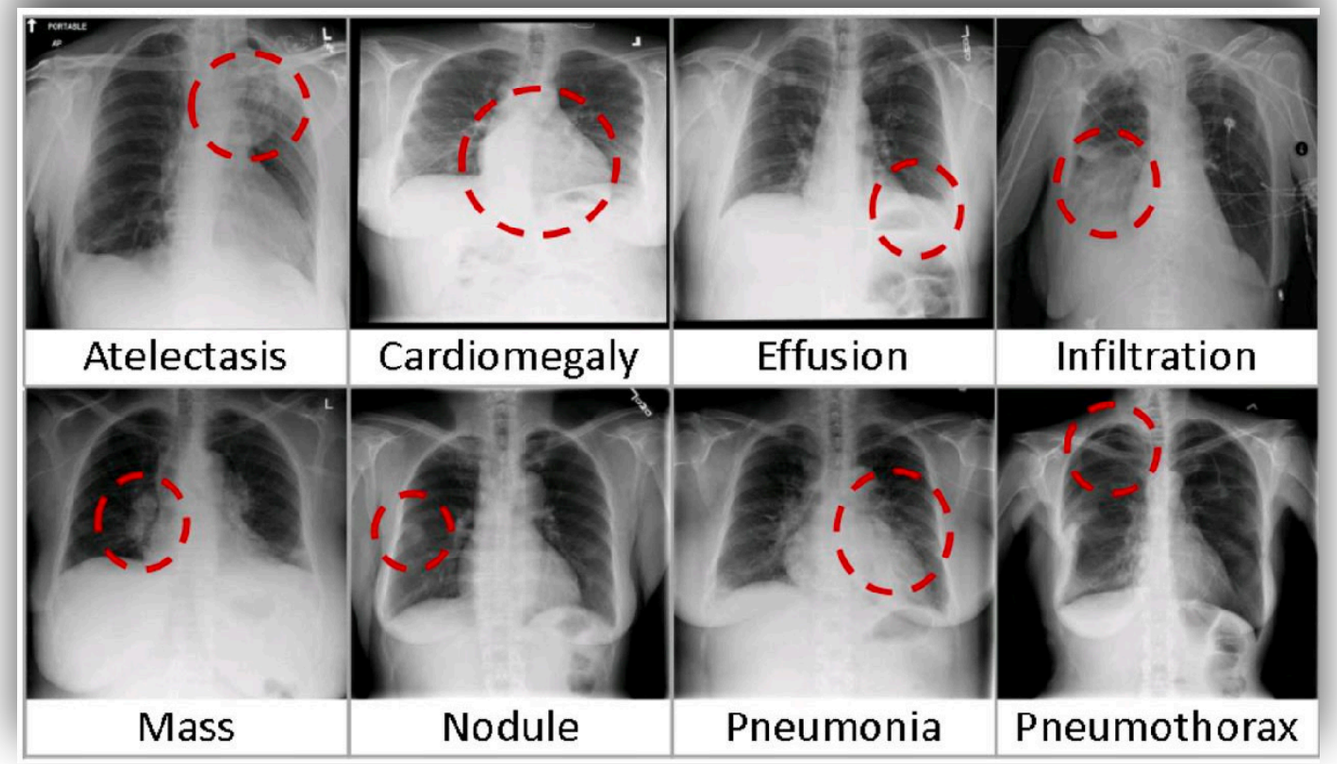
Titanic: a Class Prediction Example

1. Get a dataset
2. Train the model
3. ???
4. Profit

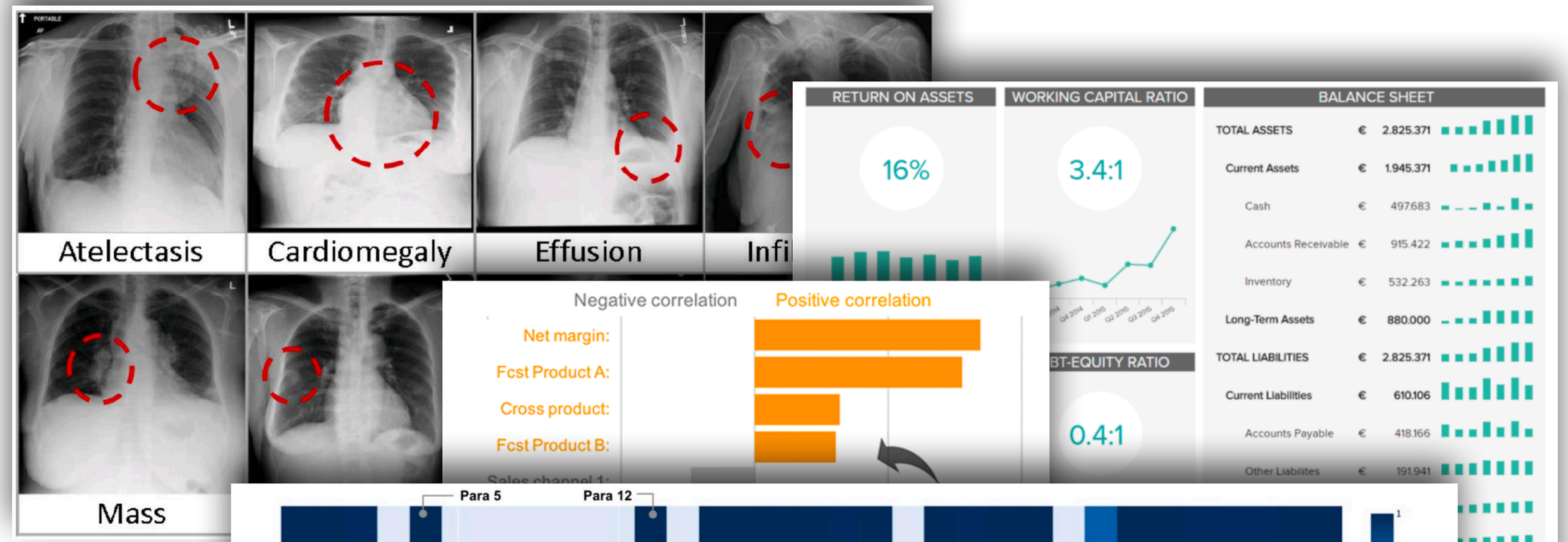
Feature	Values	Type	Description
Pclass	$1, 2, 3$	categorical	Ticket class
sex	$0, 1$	categorical	passenger gender
SibSp	$0 - 8$	categorical	# of siblings/spouses
Parch	$0 - 6$	categorical	# of parents/children
Embarked:	C, Q, S	categorical	port of embarkation
Survived:	$0, 1$	categorical	passenger survival
Age	$0.17 - 76$	numerical	passenger age
Fare	$0 - 512$	numerical	passenger fare

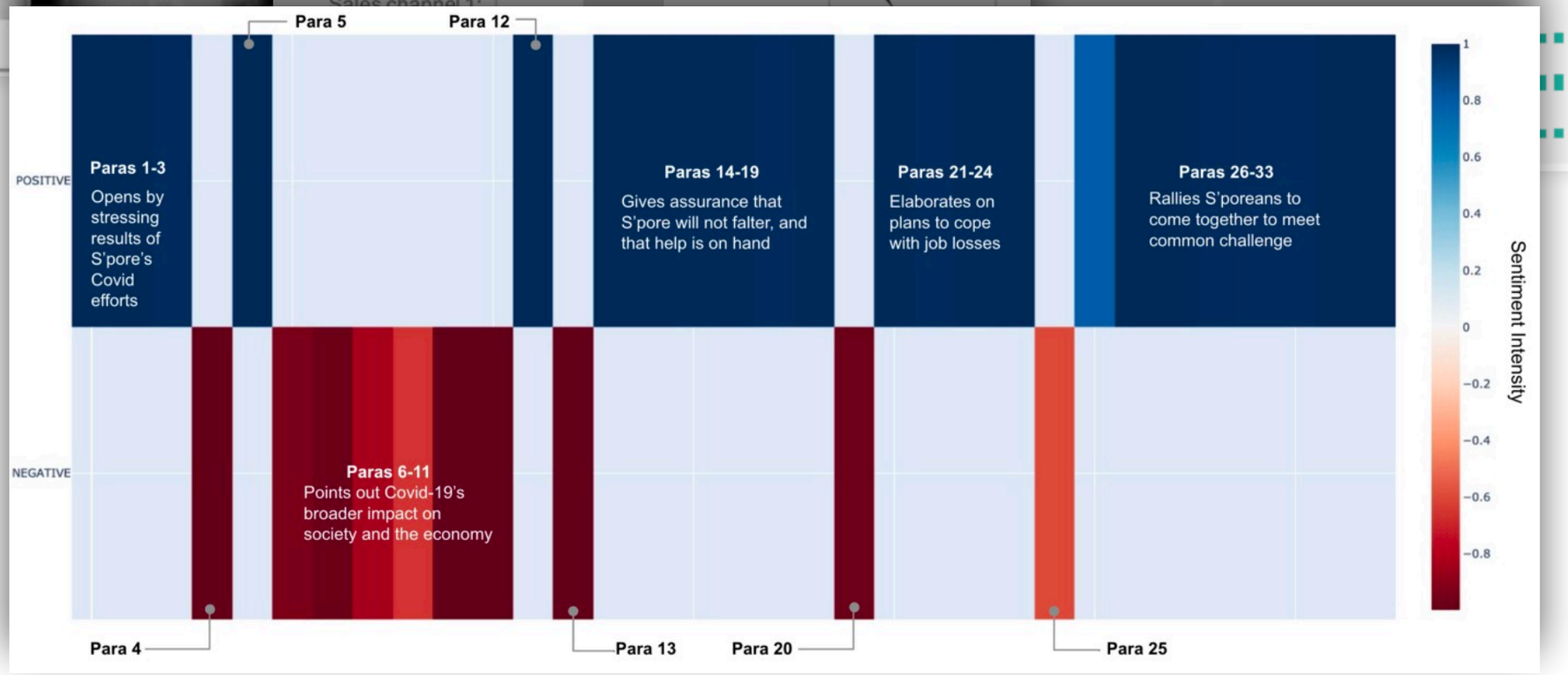
- Anna survived
- … but why?

Explanation in Critical Fields



Explanation in Critical Fields





Argumentation Theory

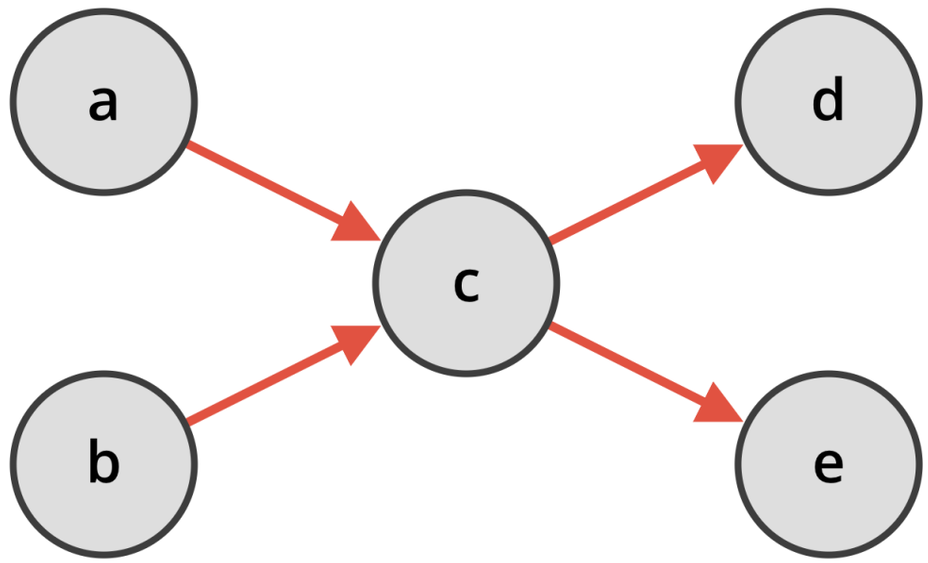
- Studies how conclusions can be **supported** or **undermined**
- Includes various forms of dialogue
  - Deliberation
  - Negotiation
  - Persuasion

Argumentation Theory

- Studies how conclusions can be **supported** or **undermined**
- Includes various forms of dialogue
  - Deliberation
  - Negotiation
  - Persuasion
- Concerned with collaborative decision-making procedures
- It naturally provides explanations

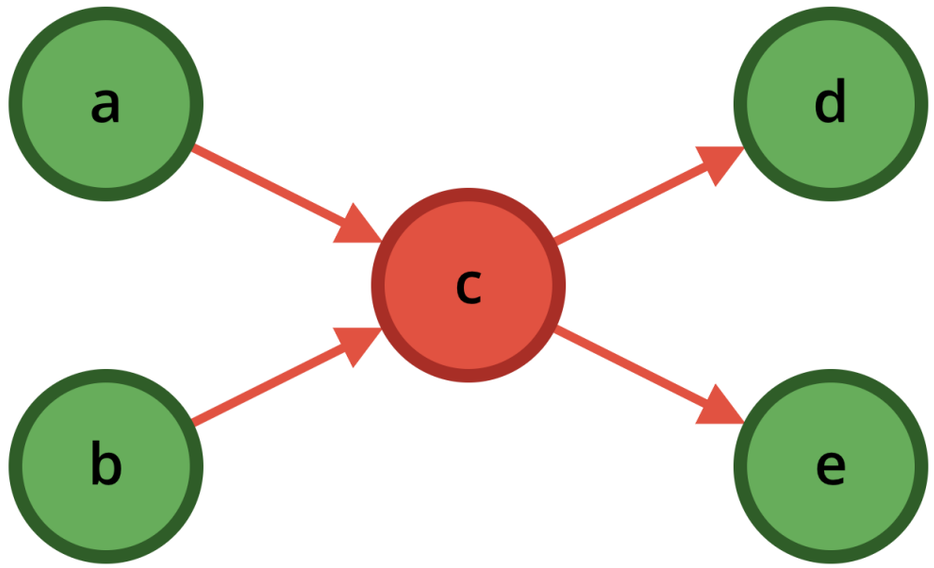
Abstract Argumentation

- Argumentation Frameworks (AFs)
- Find acceptable arguments



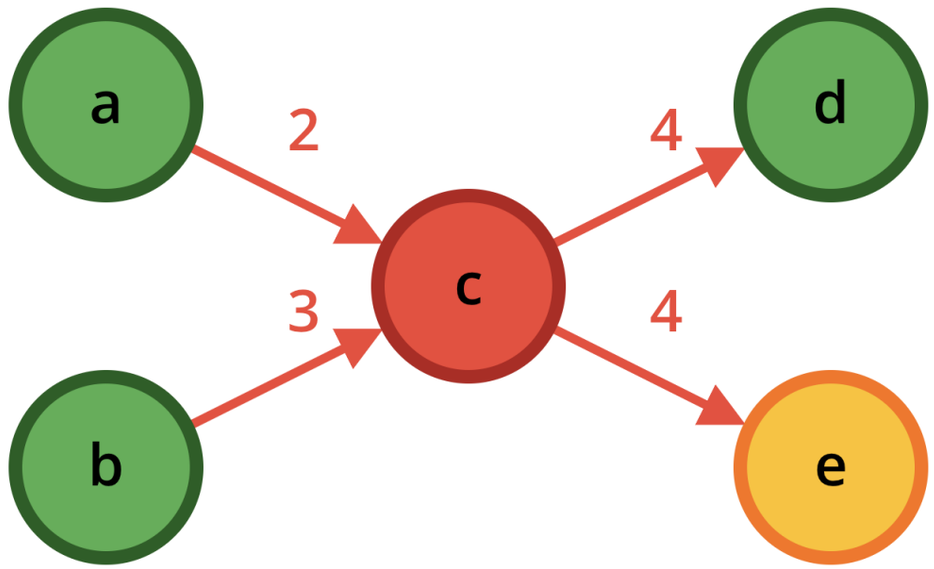
Abstract Argumentation

- Argumentation Frameworks (AFs)
- Find acceptable arguments
- Criteria: argumentation semantics



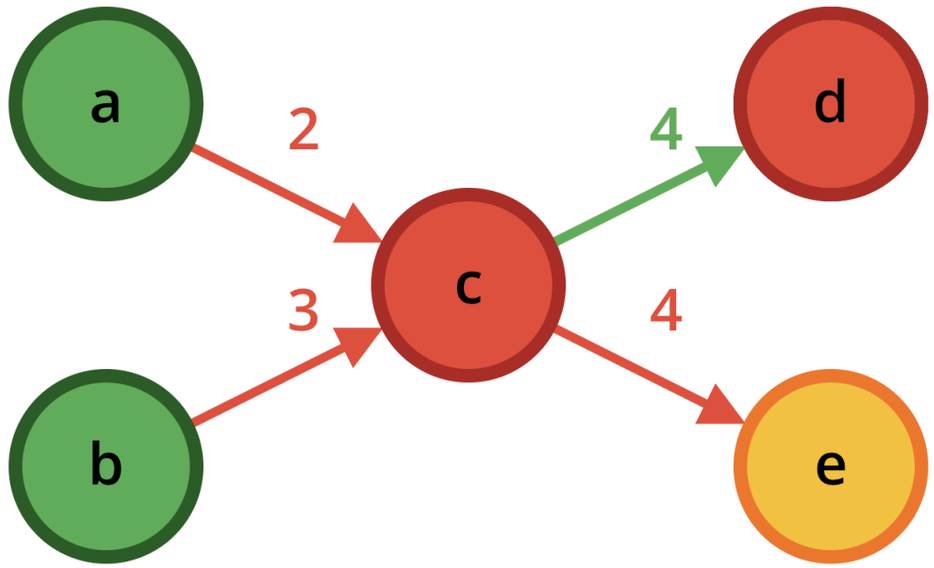
Abstract Argumentation

- Argumentation Frameworks (AFs)
- Find acceptable arguments
- Criteria: argumentation semantics
- Weights on attacks



Abstract Argumentation

- Argumentation Frameworks (AFs)
- Find acceptable arguments
- Criteria: argumentation semantics
- Weights on attacks
- Bipolar AFs



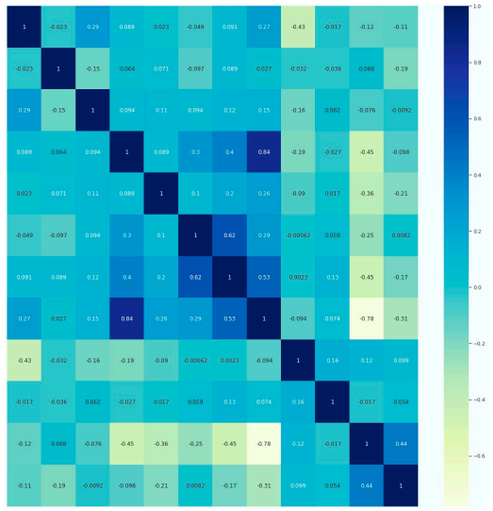
Arg-XAI

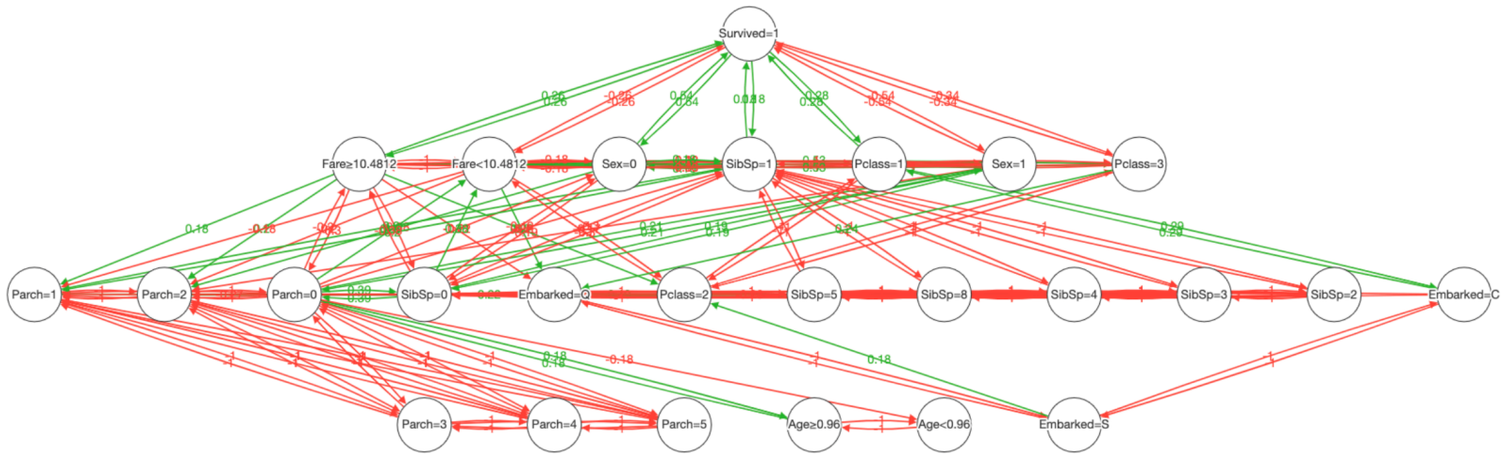
Step 1: Dataset Clustering

- We map **features** into **arguments**
- Start from the input dataset
- Split numerical features into categories (through entropy)

Step 2: BAF Generation

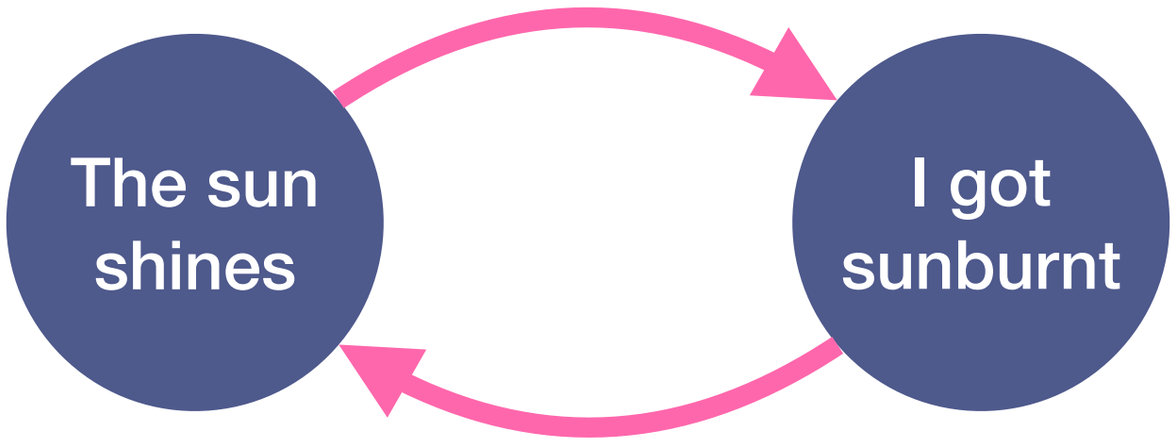
- Compute the **correlation matrix** among the features
  - Kendall, Pearson and Spearman rank coefficients
- Build a BAF based on the correlation matrix





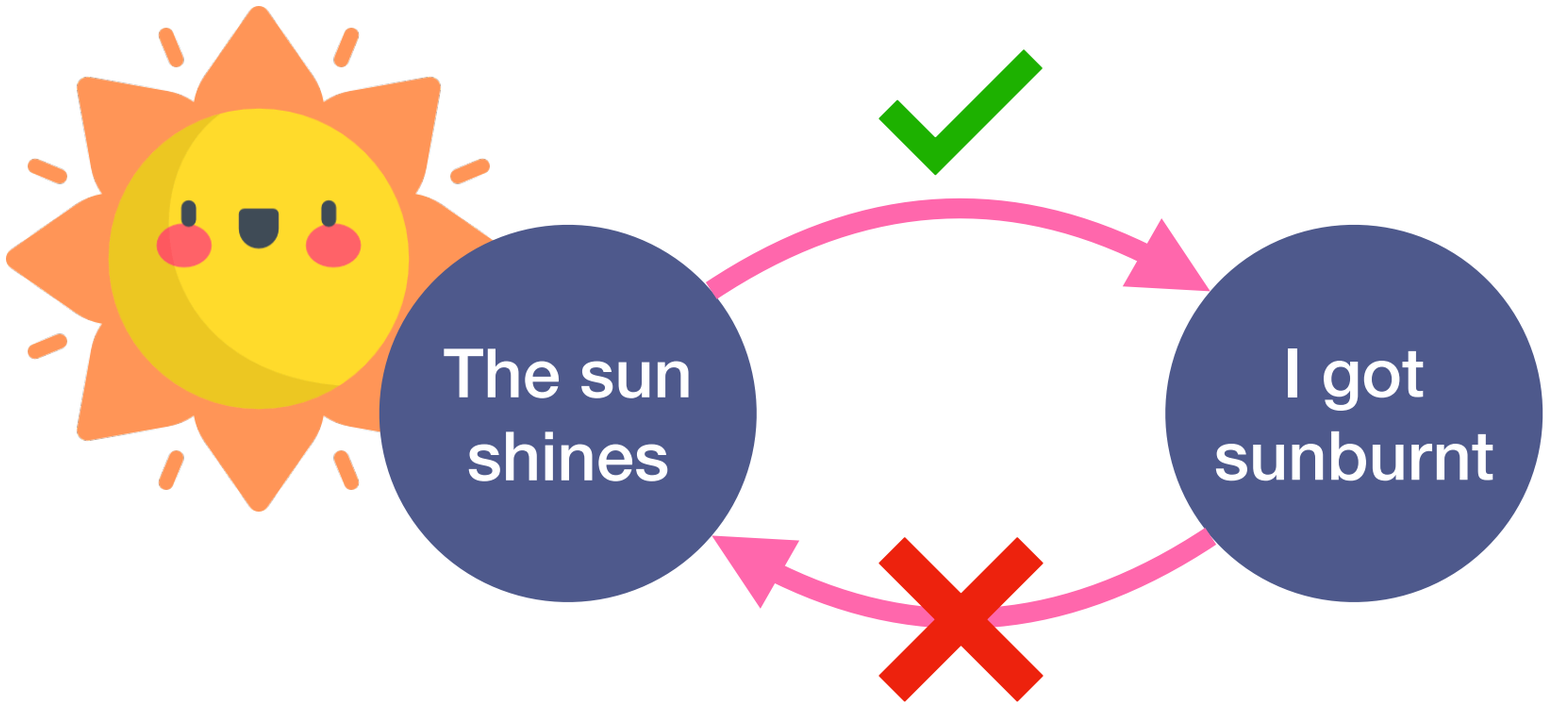
Step 3: Breaking Symmetry

- The BAF is symmetric: **no causal relationship** between features
- We use conditional probability to remove symmetric edges
  - If P (A|B) > P (B|A), then the relation from B to A is removed



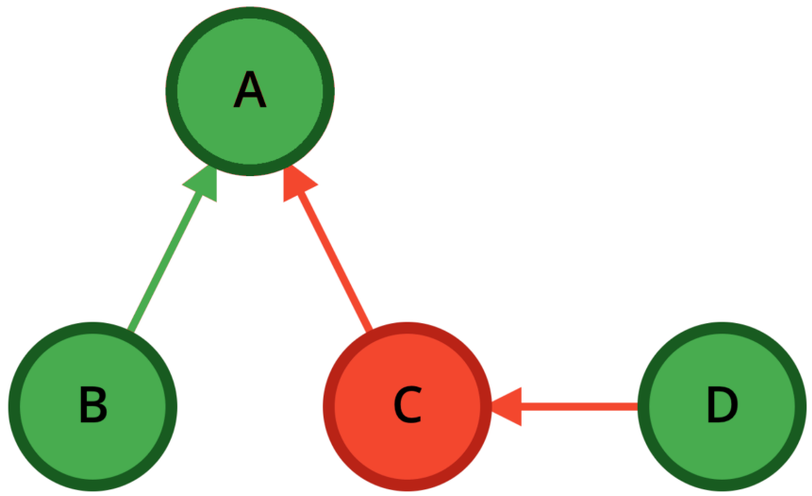
Step 3: Breaking Symmetry

- The BAF is symmetric: **no causal relationship** between features
- We use conditional probability to remove symmetric edges
  - If P (A|B) > P (B|A), then the relation from B to A is removed



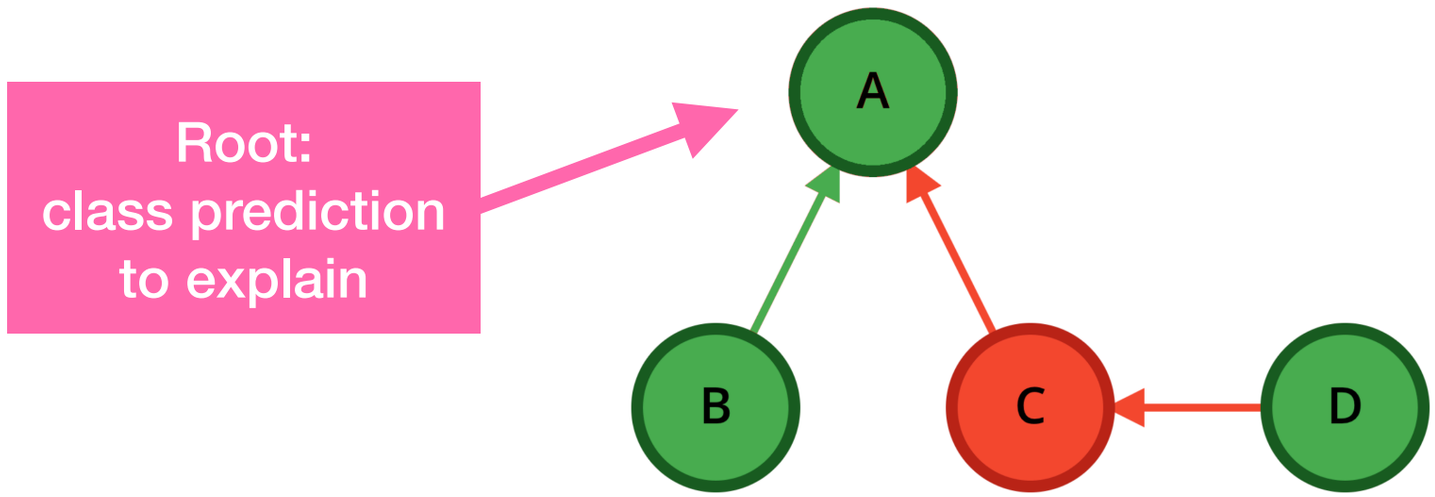
Step 4: Explanation Tree

- Compute the semi-stable extensions of the BAF
- Build the explanation tree for a selected class



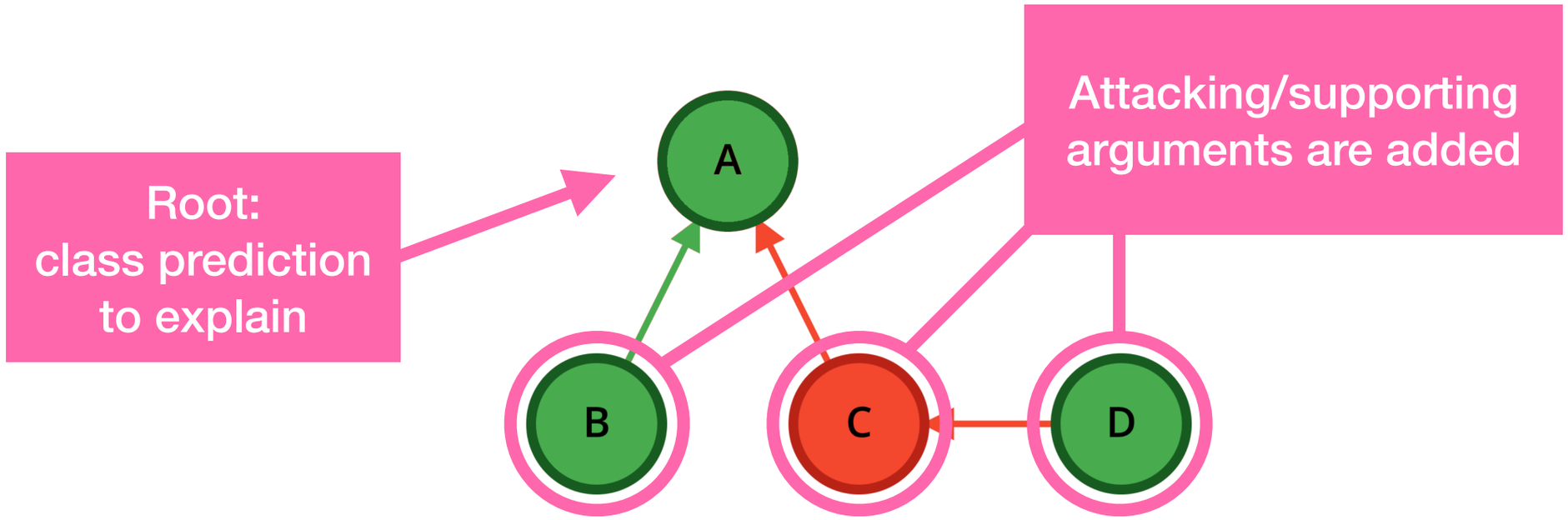
Step 4: Explanation Tree

- Compute the semi-stable extensions of the BAF
- Build the explanation tree for a selected class



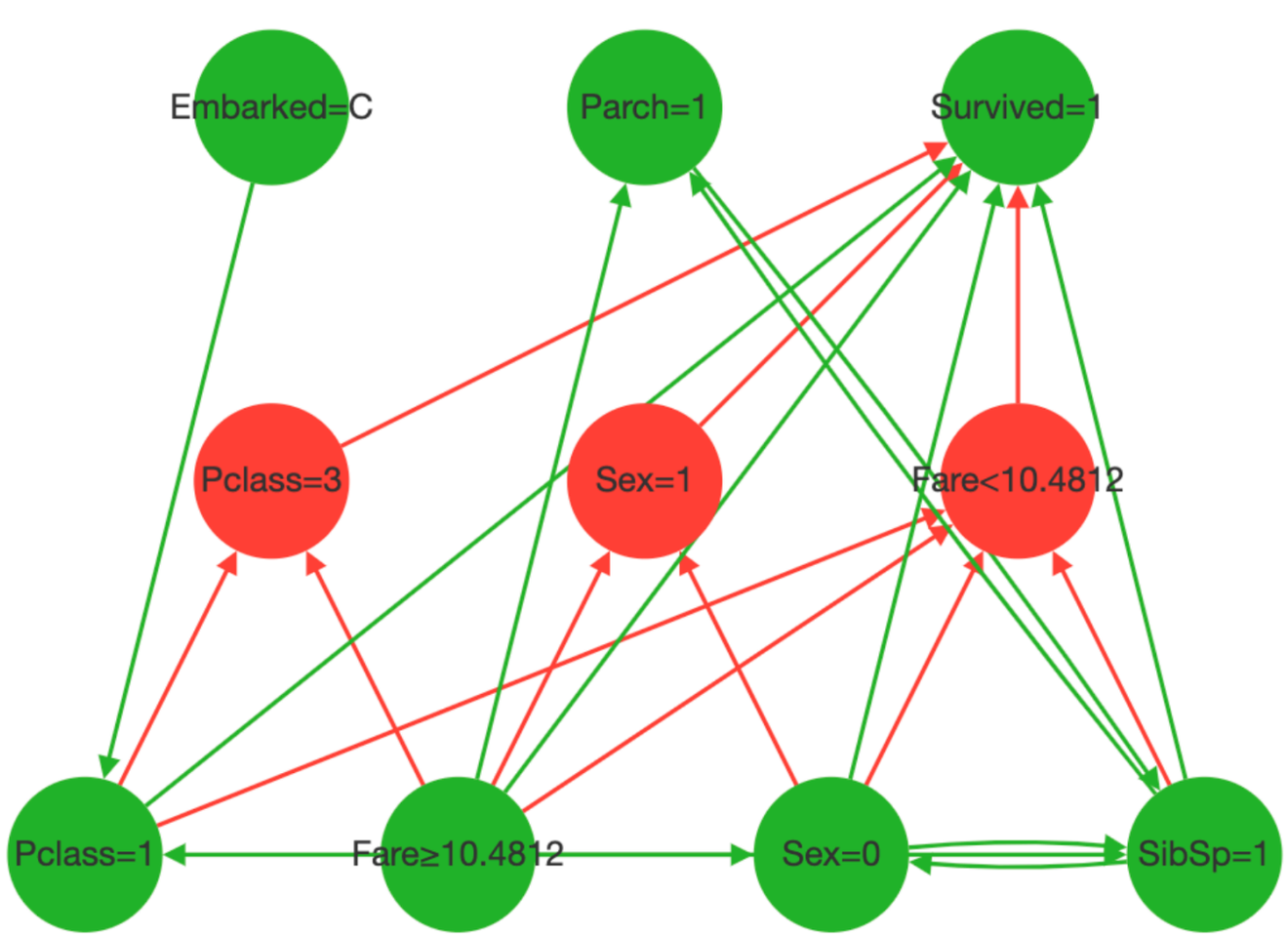
Step 4: Explanation Tree

- Compute the semi-stable extensions of the BAF
- Build the explanation tree for a selected class



Step 4: Explanation Tree

- Compute the semi-stable extensions of the BAF
- Build the explanation tree for a selected class



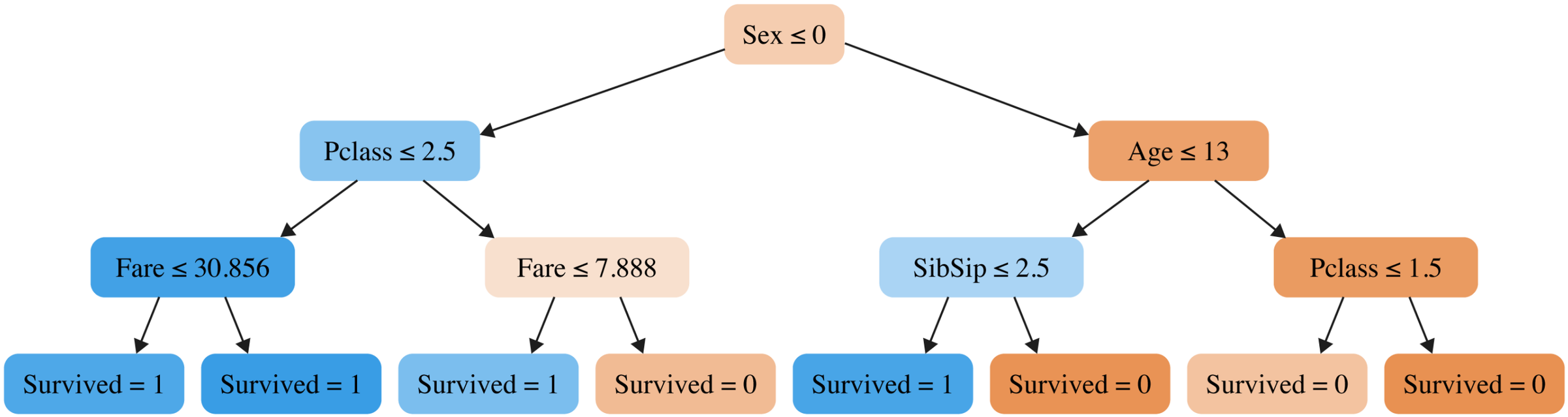
Demo

arg-xai.dmi.unipg.it

Validation A: Decision Tree

- Semi-stable extension:

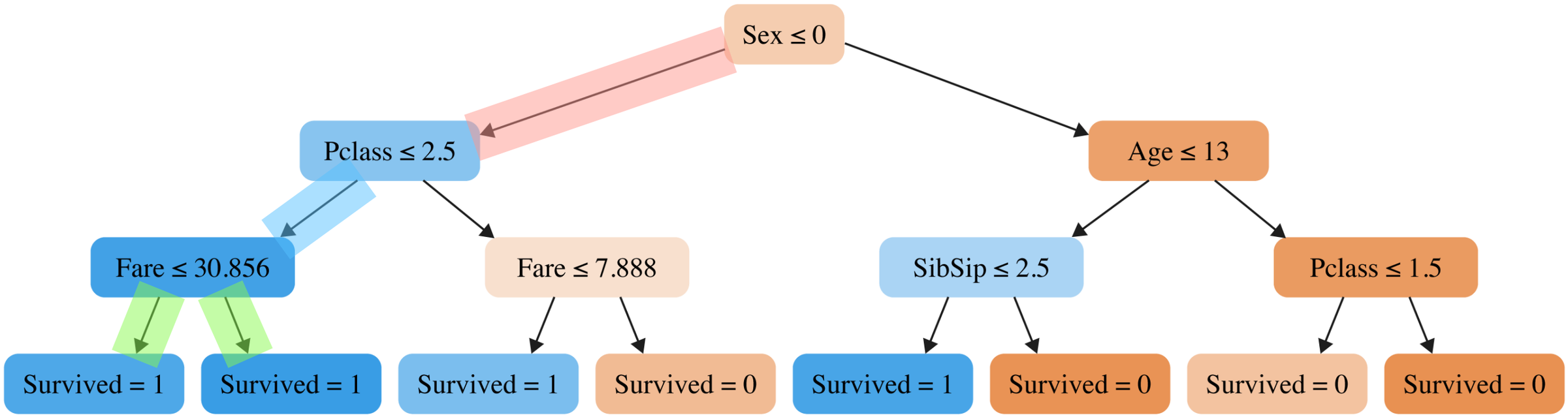
Sex=0, Pclass=1, Fare≥10.4812, Survived=1,
Age<0.96, Embarked=C, Parch=1, SibSp=1



Validation A: Decision Tree

- Semi-stable extension:

Sex=0, **Pclass=1**, **Fare≥10.4812**, **Survived=1**,
Age<0.96, Embarked=C, Parch=1, SibSp=1



Validation B: Rule-Based Classifier

- Semi-stable extension:

  thall=2, caa=0, slp=2, exng=0, sex=0, cp=2, chol<245.5, oldpeak<1.7, thalach≥147.5, trtbps<107, age<54.5, restecg=1, **Heart Attack=1**, fbs=1

- We use RIPPER to derive a set of rules for **Heart Attack=1**

  thall=2 ∧ caa=0 ∧ slp=2, exng=0 ∧ caa=0 ∧ sex=0, exng=0 ∧ thall=2 ∧ cp=2, caa=0 ∧ thall=2 ∧ sex=1, trtbps=130.0-138.0 ∧ chol=187.0-207.0

Validation B: Rule-Based Classifier

- Semi-stable extension:

thall=2, caa=0, slp=2, exng=0, sex=0, cp=2, chol<245.5, oldpeak<1.7, thalach≥147.5, trtbps<107, age<54.5, restecg=1, **Heart Attack=1**, fbs=1

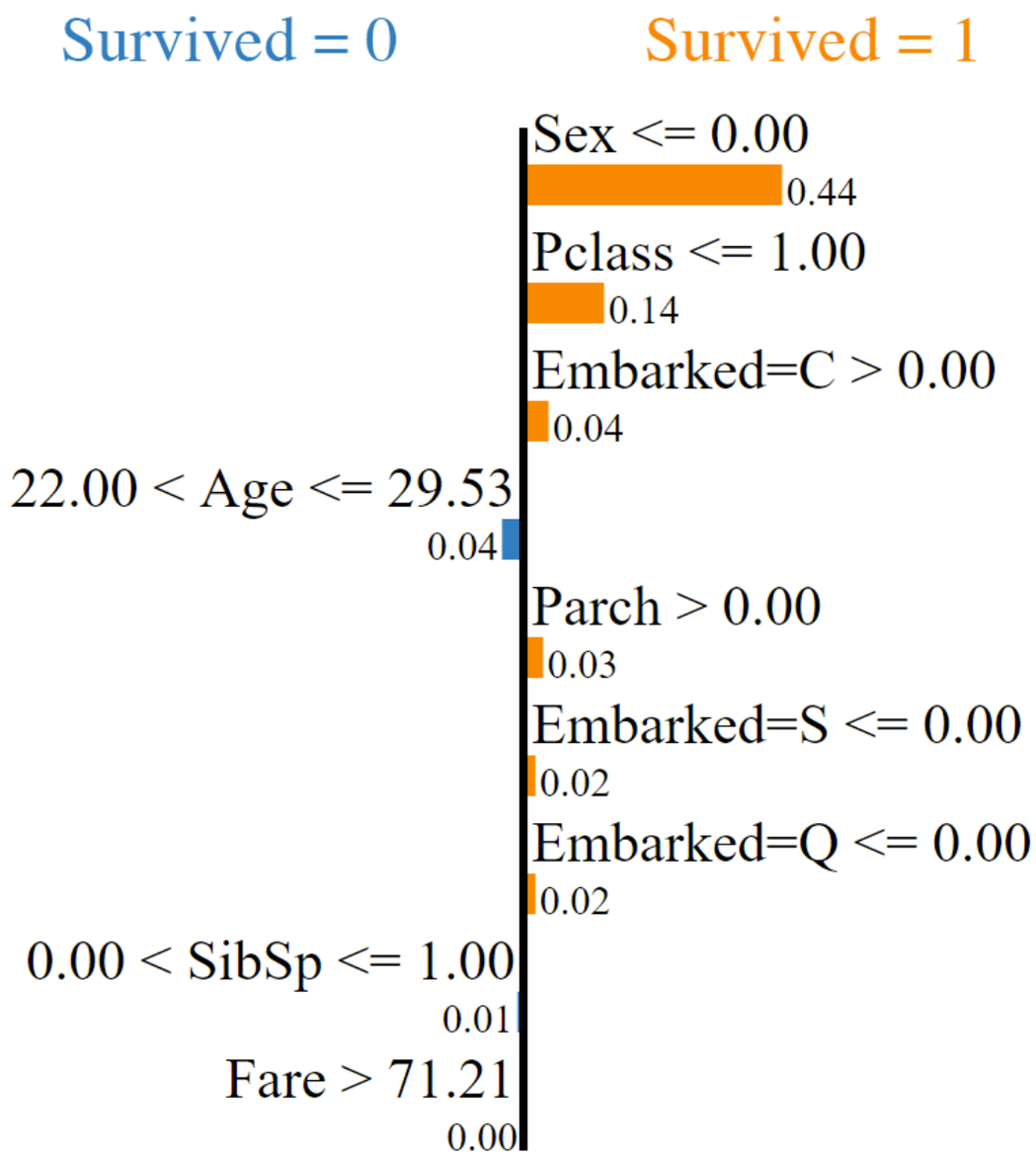
- We use RIPPER to derive a set of rules for **Heart Attack=1**

thall=2 ∧ caa=0 ∧ slp=2, exng=0 ∧ caa=0 ∧ sex=0, exng=0 ∧ thall=2 ∧ cp=2, caa=0 ∧ thall=2 ∧ sex=1, trtbps=130.0-138.0 ∧ chol=187.0-207.0

Validation C: LIME

- Semi-stable extension:

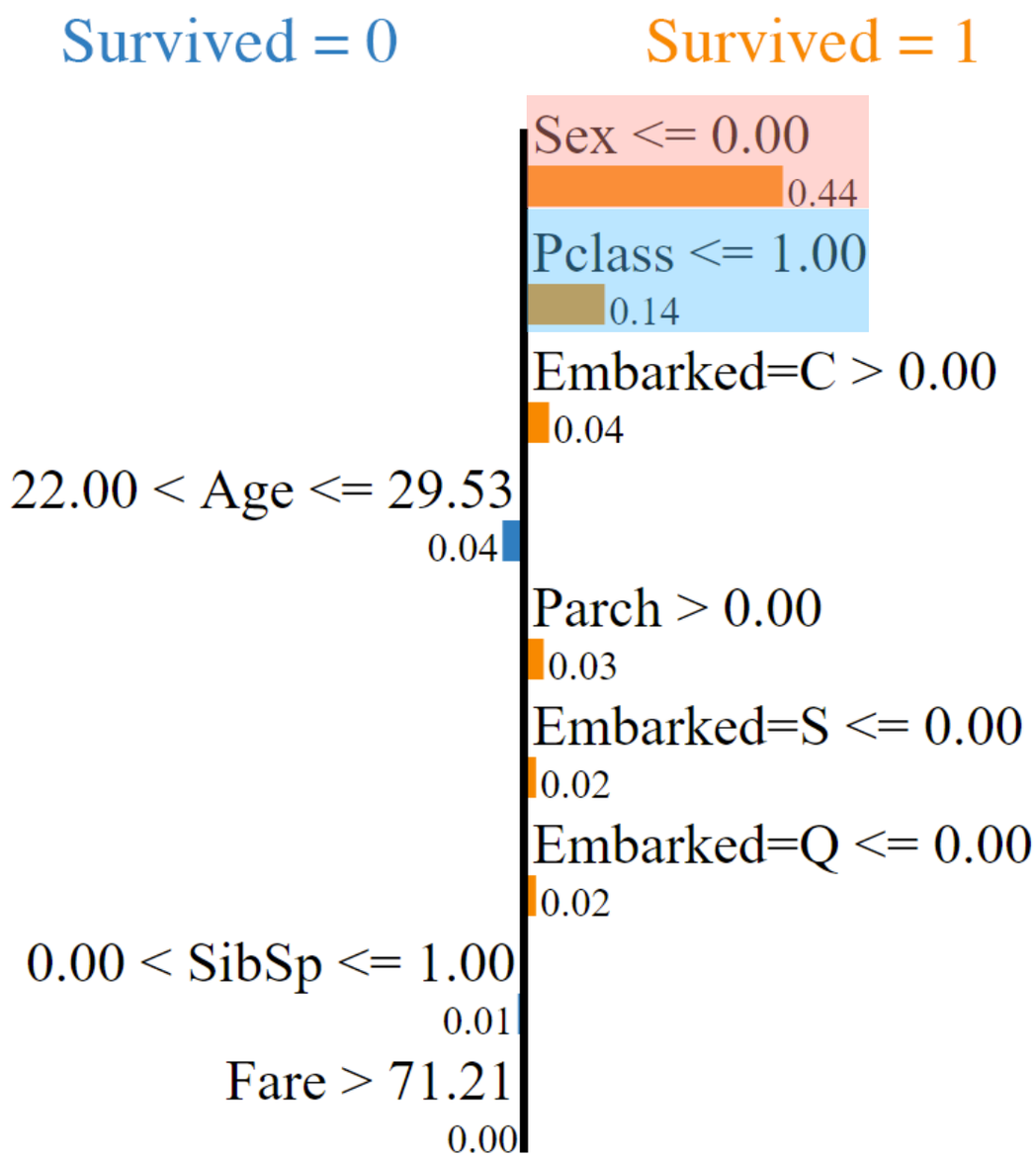
Sex=0,
Pclass=1,
Fare≥10.4812,
Survived=1,
Age<0.96,
Embarked=C,
Parch=1,
SibSp=1



Validation C: LIME

- Semi-stable extension:

Sex=0,
Pclass=1,
Fare≥10.4812,
Survived=1,
Age<0.96,
Embarked=C,
Parch=1,
SibSp=1



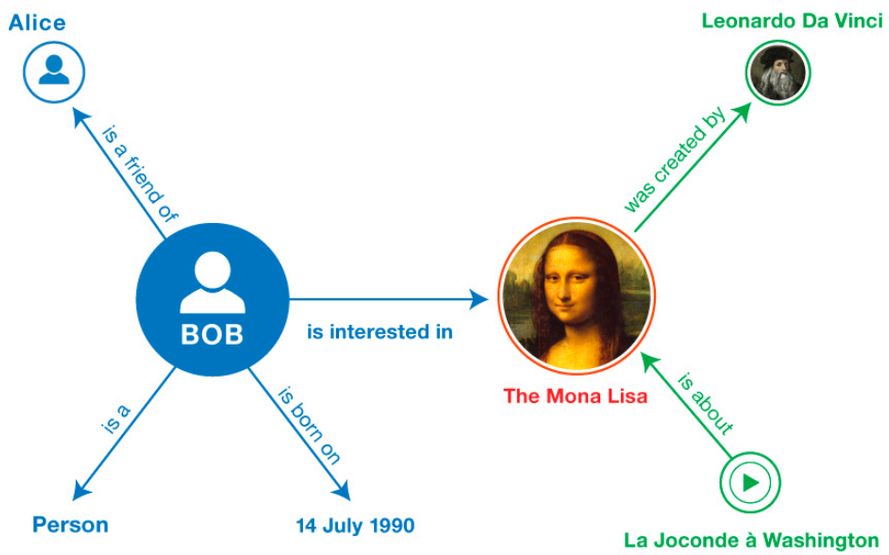
Future research lines

Issue #1: relate the features

- Problem: correlation is symmetric

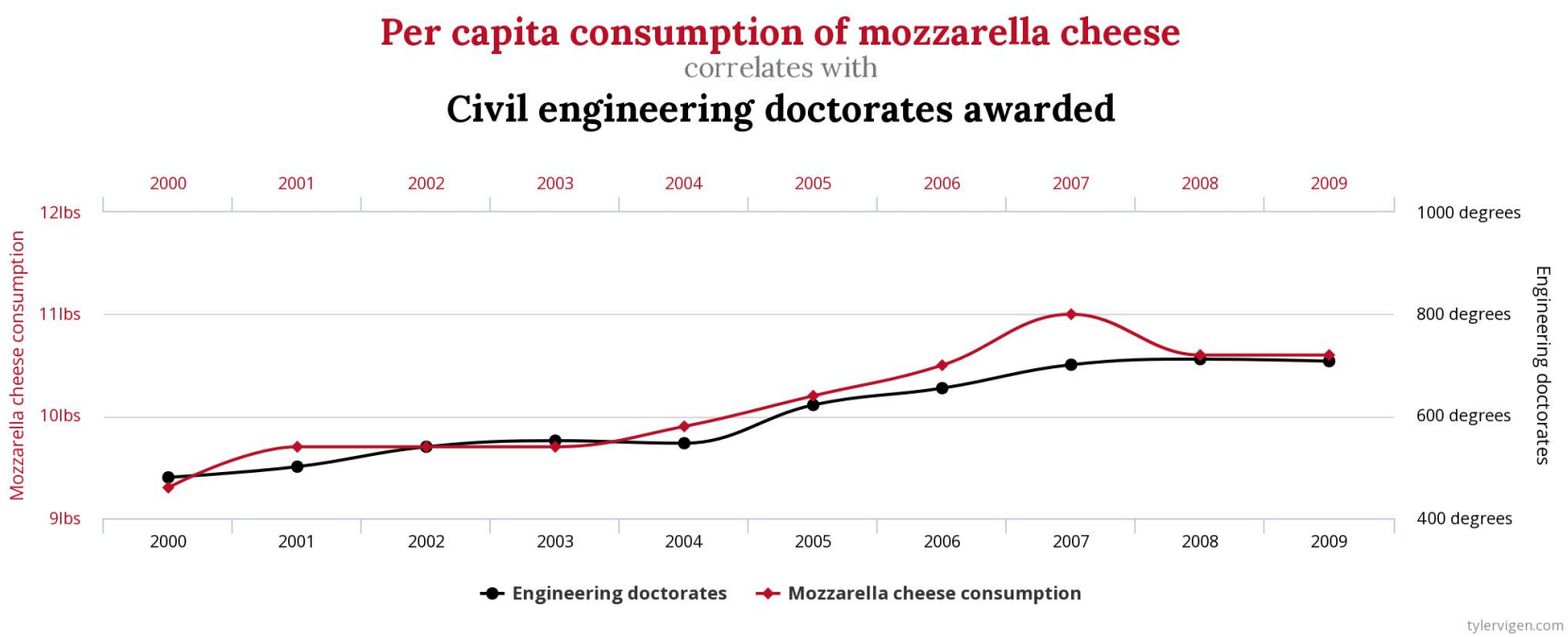
Issue #1: relate the features

- Problem: correlation is symmetric
- Proposal: derive **causality** between features using external, additional knowledge (semantic web?)



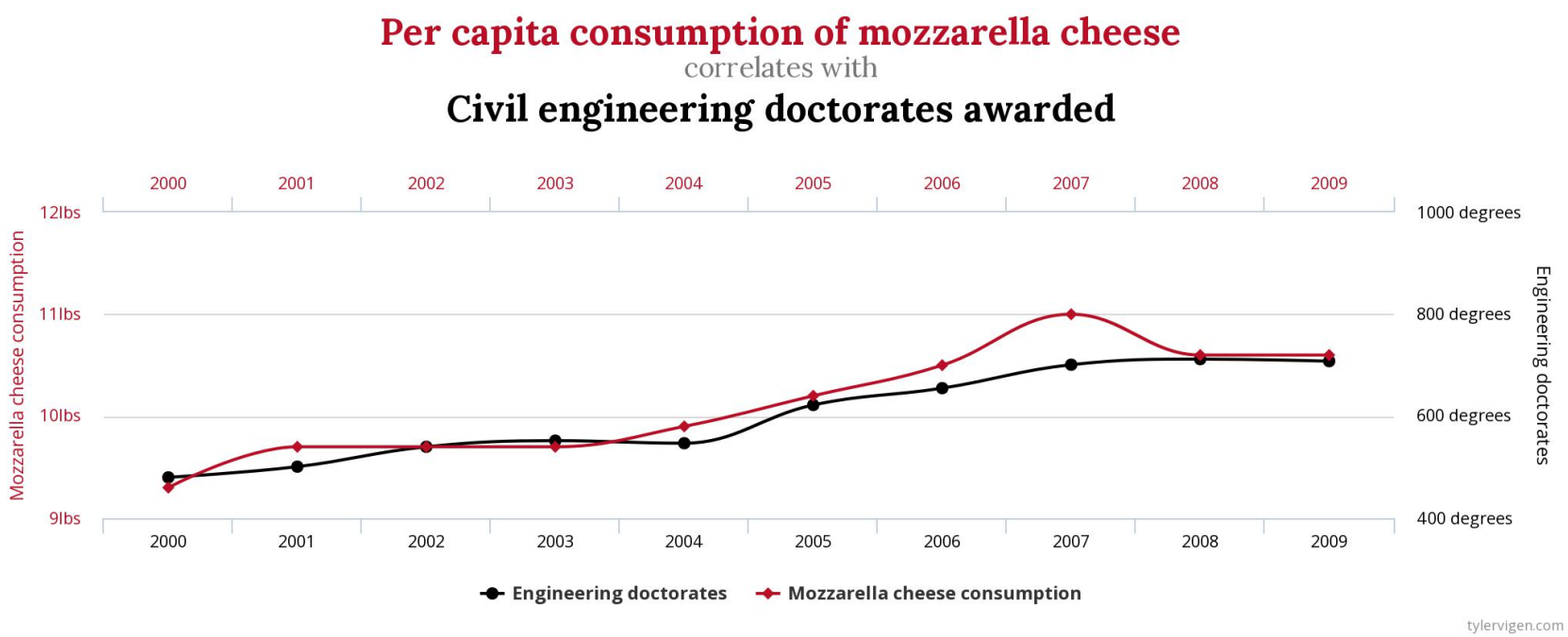
Issue #2: spurious correlation

- Relations in which events are associated but not causally related



Issue #2: spurious correlation

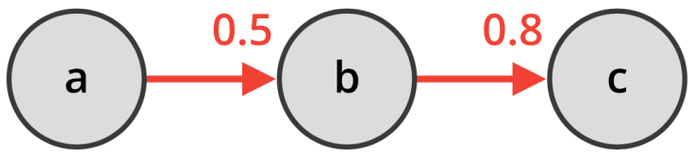
- Relations in which events are associated but not causally related



- Proposed solution: reasoning

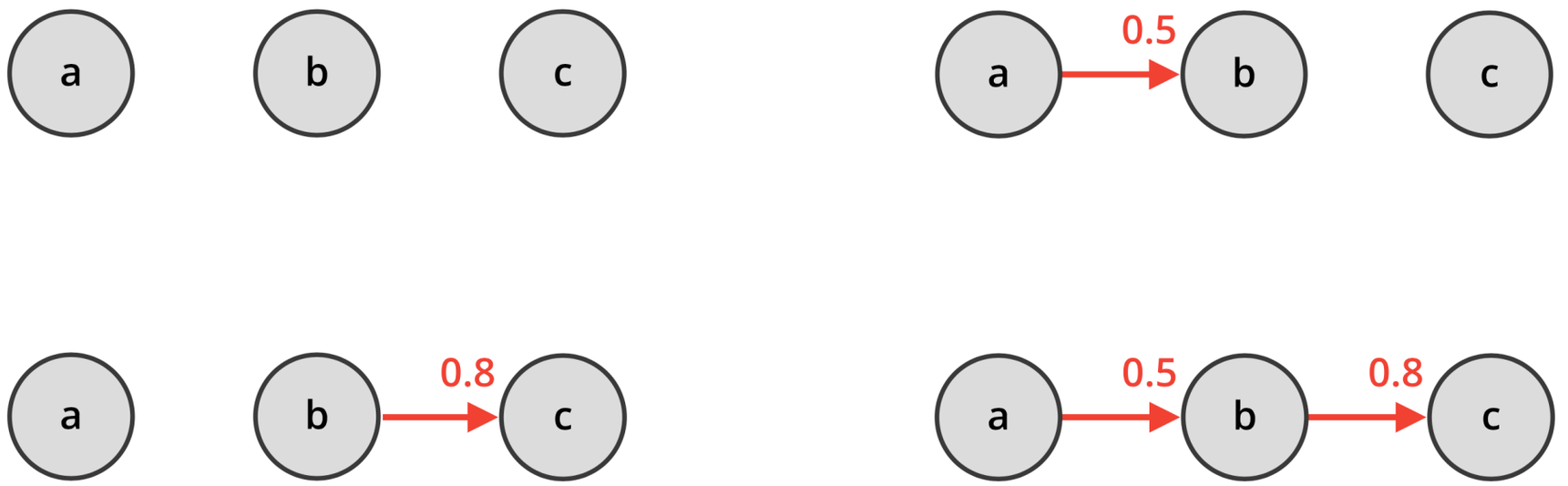
Issue #3: computational complexity

- An attack (a,b) exists with a probability P(a,b)



Issue #3: computational complexity

- An attack (a,b) exists with a probability P(a,b)
- **Constellation** approach: generate all possible worlds
- Computing the grounded semantics is intractable [1]



1. Fazzinga, B., Flesca, S., Parisi, F.: On the Complexity of Probabilistic Abstract Argumentation Frameworks. ACM Trans. Comput. Log. 16(3): 22:1-22:39 (2015)

Issue #3: computational complexity

- An attack (a,b) exists with a probability P(a,b)
- **Constellation** approach: generate all possible worlds
- Computing the grounded semantics is intractable [1]

- Possible solution: restrict to particular classes of graphs
- Example: the problem becomes treatable when the underlying structure is a polytree [2]
- Trees are relevant in modelling real-world scenarios through Argumentation techniques

1. Fazzinga, B., Flesca, S., Parisi, F.: On the Complexity of Probabilistic Abstract Argumentation Frameworks. ACM Trans. Comput. Log. 16(3): 22:1-22:39 (2015)
2. Bistarelli, S., David, V., Santini, F., Taticchi, C.: Computing Grounded Semantics of Uncontroversial Acyclic Constellation Probabilistic Argumentation in Linear Time. AI³@AI*IA 2022

Take-home message

- Arg-XAI: a tool for explaining the outcomes of a classifier
- Abstract argumentation to obtain dialectical explanations
- Build an explanation tree showing why a class is assigned

Take-home message

- Arg-XAI: a tool for explaining the outcomes of a classifier
- Abstract argumentation to obtain dialectical explanations
- Build an explanation tree showing why a class is assigned
- What's next?
  - Derive causality
  - Detect spurious correlations
  - Devise tractable approaches

THANK YOU FOR YOUR ATTENTION

Arg-XAI: a Tool for Explaining Machine Learning Results

Stefano Bistarelli, Francesco Santini, Alessio Mancinelli, Carlo Taticchi

Ital-IA 2023

A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA