Beyond Single-Feature Importance with ICECREAM

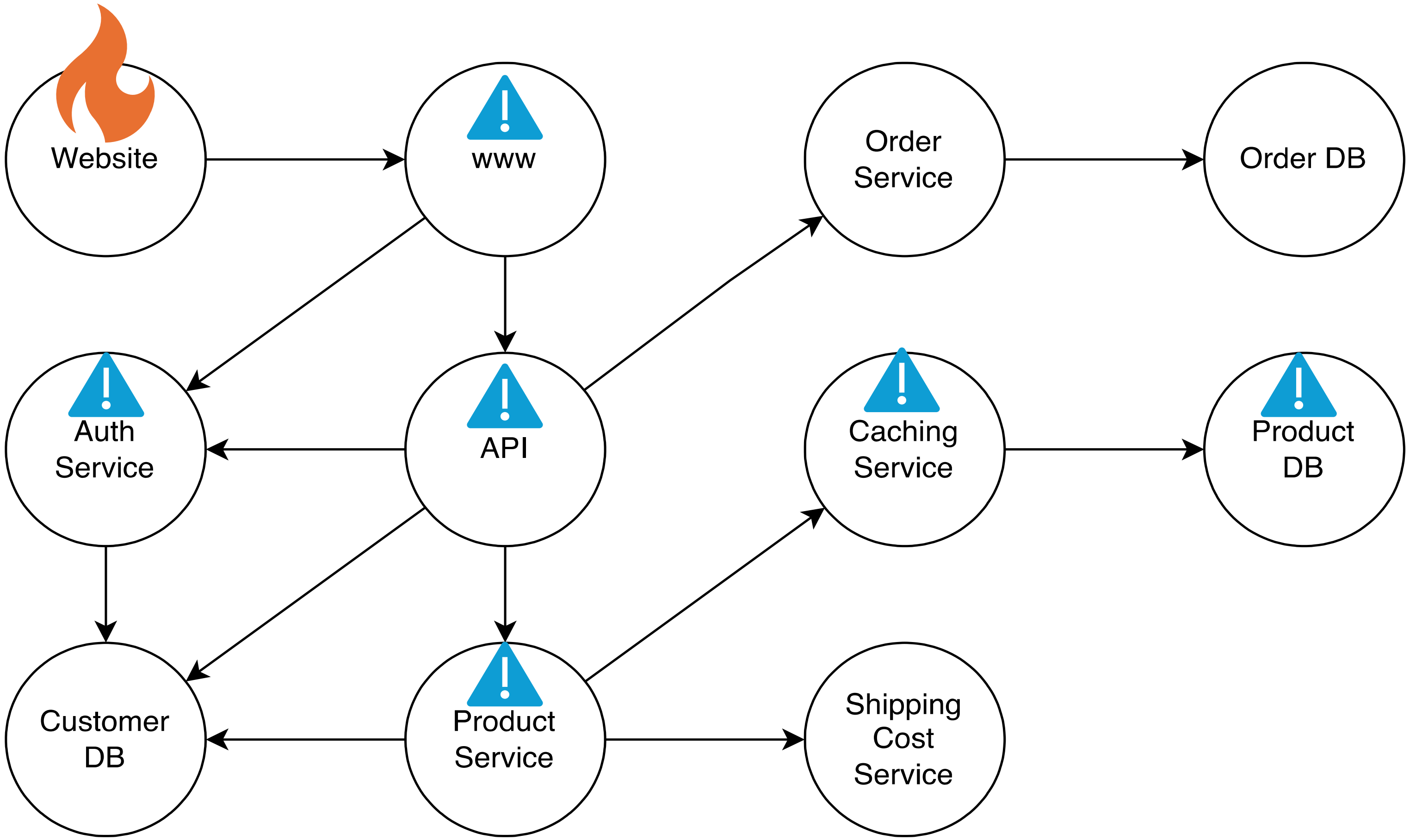
M. Oesterle, P. Blöbaum, A. A. Mastakouri, **E. Kirschbaum**

CLeaR 2025

Agenda

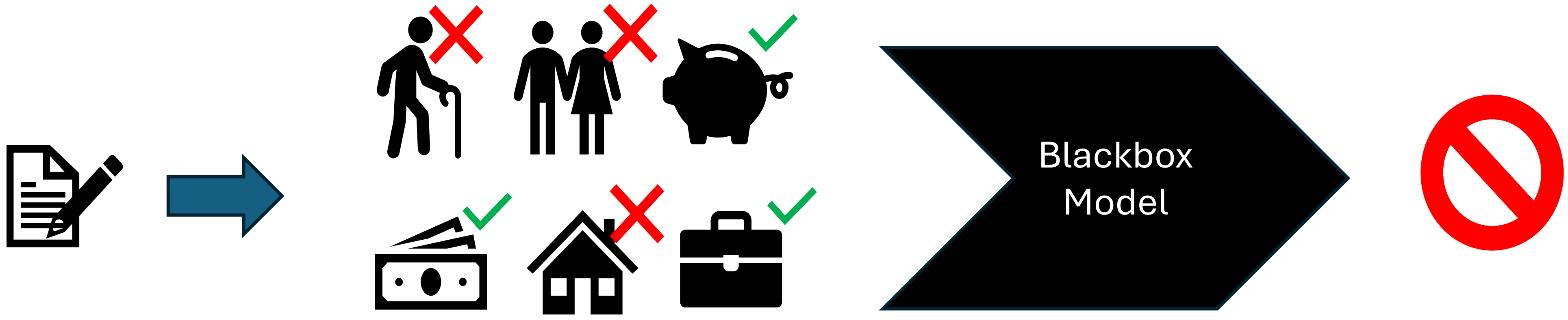
- Motivation
- Background: Graphical Causal Models
- ICECREAM
- Experiments and Results
- Conclusion

Eample 1: Cloud Application



Can we resolve the issue by fixing a ***single*** service, or do we need to fix ***multiple*** components?

Example 2: Credit Risk Prediction



Which features were respondible for the rejection of the credit application?

Was it a ***single*** feature, or the combination of ***multiple*** features?

What do existing methods do?

- Popular SHAP method (Lundberg and Lee, 2017) ranks and quantifies importance of *individual* features

    1. credit purpose
    2. housing situation
    3. personal status
    4. ...

But what combination of features was crucial for the outcome?
How many features are actually required to explain the models decision?

What does ICECREAM do instead?

- ***I***dentifying ***C***oalition-based ***E***xplanations for ***C***ommon and ***R***are ***E***vents in ***A***ny ***M***odel
- Detects combinations of variables whose interplay explain a certain outcome of a model or system
- Can be used for explainability and interpretability of any feature-target system
- Can be applied to perform root cause analysis (RCA) in any system with known causal structure

Background: Graphical Causal Models

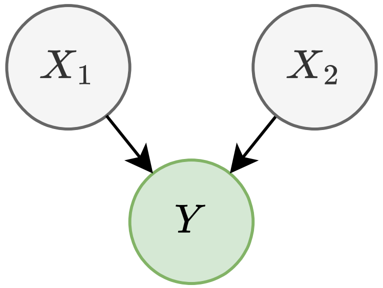
Background: Graphical Causal Models (GCM)

- Causal graph $G = (V, E)$
- Nodes = random variables $V = \{V_1, \ldots, V_N\}$
- Edges $\mathrm{V_i} \rightarrow V_j \in E$ represent causal relationships
- We split V as follows:

  Y = target variable whose value we want to explain

  X_i= observed variables

  Λ_i= unobserved variables



Background: Graphical Causal Models (GCM)

- Intervention: set variable V_i to some value v_i while all other variables keep their relationships
- Represented as cutting all incoming edges at the intervened variable
- Interventional distribution

$$\mathbb{P}\left[V_j \,\middle|\, do(V_i = v_i)\right] = \mathbb{P}\left[V_j \,\middle|\, do(V_i)\right] \neq^{*} \mathbb{P}[V_j | \, V_i = v_i]$$

* only equal if V_i has no causal parents

ICECREAM

Question:

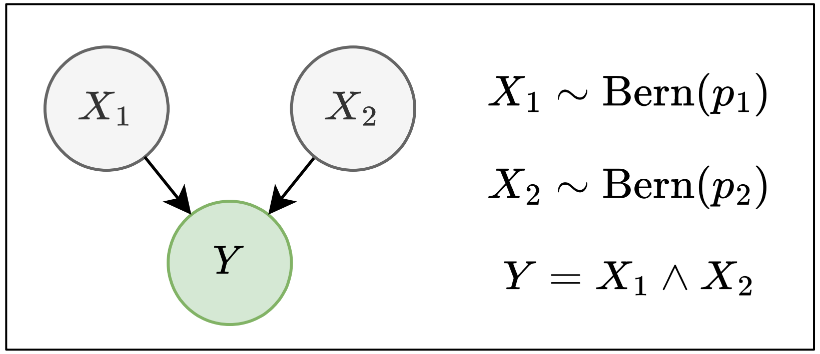
What is the smallest coalition of variables which fully explains the target value y for an observation v?

→Define an *explanation score* that quantifies the influence of a set of

variables on the target variable

Desired Properties of the Explanation Score

1. The explanation score of a coalition is not just the sum of the explanation scores of the individual variables.



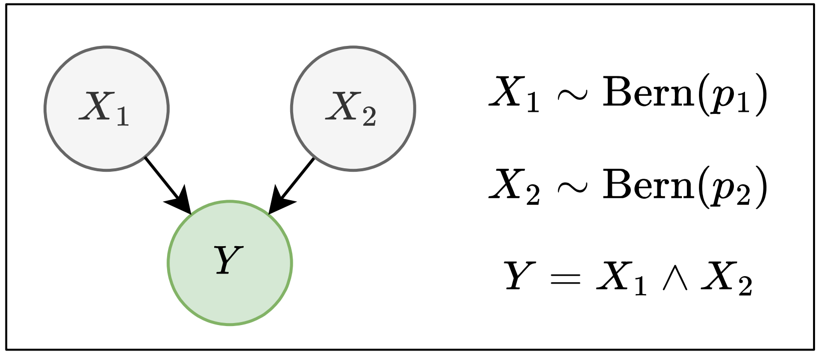
Event: $y = 0, x_1 = 0, x_2 = 0$

Already $x_1 = 0$ fully explains $y = 0$, similarly does $x_2 = 0$

→ the coalition $\{X_1, X_2\}$ should not get a higher score than $\{X_1\}$ or $\{X_2\}$

Desired Properties of the Explanation Score

2. Rare events get a higher explanation score than common events.



Event: $y = 1, x_1 = 1, x_2 = 1$

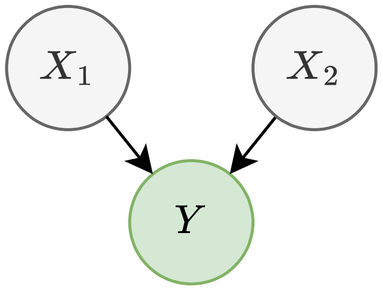
Consider $p_1 = 0.001, p_2 = 0.9$:

→ X_1 is the more interesting explanation

→ score for $\{X_1\}$ should be higher than for $\{X_2\}$

Desired Properties of the Explanation Score

3. Explanation scores in anti-causal direction are zero.



$X_1 \sim \mathrm{Bern}(p_1)$

$X_2 \sim \mathrm{Bern}(p_2)$

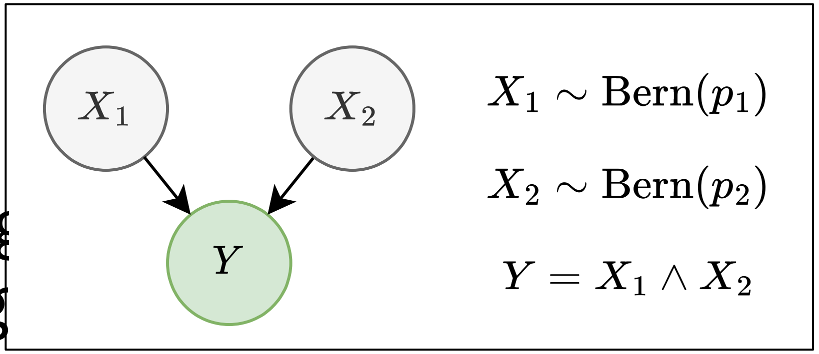
$Y = X_1 \wedge X_2$

Y cannot influence X_1 or X_2

→ effects cannot explain their causes

Desired Properties of the Explanation Score

4. The explanation score of a variable is not depend on how the other variables are grouped into coalitions.



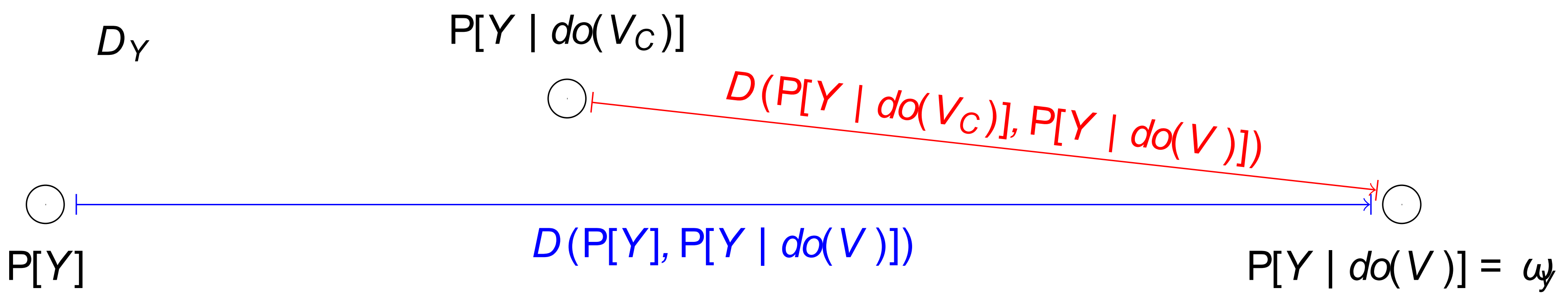
For the influence of X_1 on Y it should not matter if X_2 is actually a sing variable, or the agg variables with the same effect on Y

Formal Definition of ICECREAM's Explanation Score

Definition 1: Let $\mathcal{D}_\mathcal{Y}$ denote the set of all probability distributions over the domain $\mathcal{Y}$. Further let $D: \mathcal{D}_\mathcal{Y}^2 \to \mathbb{R}$ be a distance measure to quantify the distance between distributions in $\mathcal{D}_\mathcal{Y}$. The explanation score of the coalition $V_C \subseteq V$ with respect to the observation $v \in \mathcal{V}$ is then defined as the function $\mathcal{E}_v: 2^I \to (-\infty, 1]$ with

$$\mathcal{E}_v(V_C) = 1 - \frac{D(\mathbb{P}[Y|\, do(V_C)], \mathbb{P}[Y|\, do(V)])}{D(\mathbb{P}[Y], \mathbb{P}[Y|\, do(V)])}$$

Intuition behind ICECREAM's Explanation Score



Further Properties of the Explanation Score

- $\mathcal{E}_v(V_C) > 0$ iff fixing coalition values gets the distribution of the target closer to the point mass distribution $Y \sim \delta_y$
- $\mathcal{E}_v(V_C) < 0$ iff fixing coalition values moves the distribution of the target further away from the point mass distribution $Y \sim \delta_y$
- $\mathcal{E}_v(V_C) = 0$ iff fixing coalition values does not change the distance to the point mass distribution $Y \sim \delta_y$
- $\mathcal{E}_v(V_C) = 1$ iff $\mathbb{P}[Y \mid do(V_C)] = \delta_y$, and we say the coalition fully explains the target value

Further Properties of the Explanation Score

- If a variable V_i is irrelevant for Y, it does not change the explanation score when it is added to a coalition
- If $\mathcal{E}_v(V_C) = 1$, then all coalitons containing V_C also have explanation score 1
- For values smaller than 1, the explanation score is not monotonic since some variables can have contradictory evidence for an outcome (unless a coalition already fully explains that outcome)
- If $\mathcal{E}_v(V_C) = 1$, there exists no variable outside of V_C that could change the value of Y

Practical Application of ICECREAM's Explanation Score

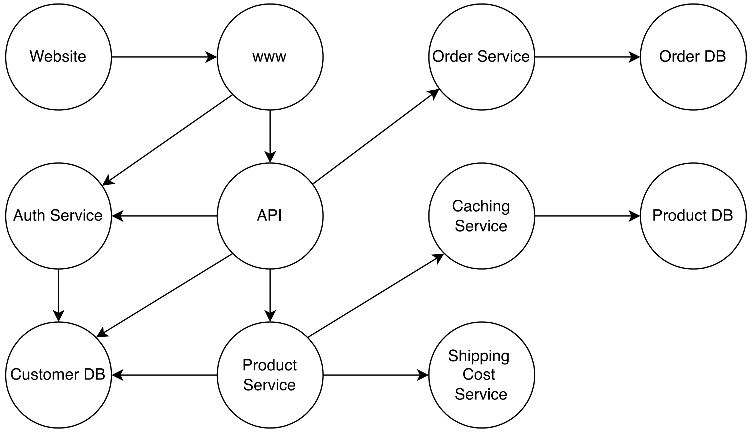
- Use KL-divergence as distance measure, simplifies explanation score

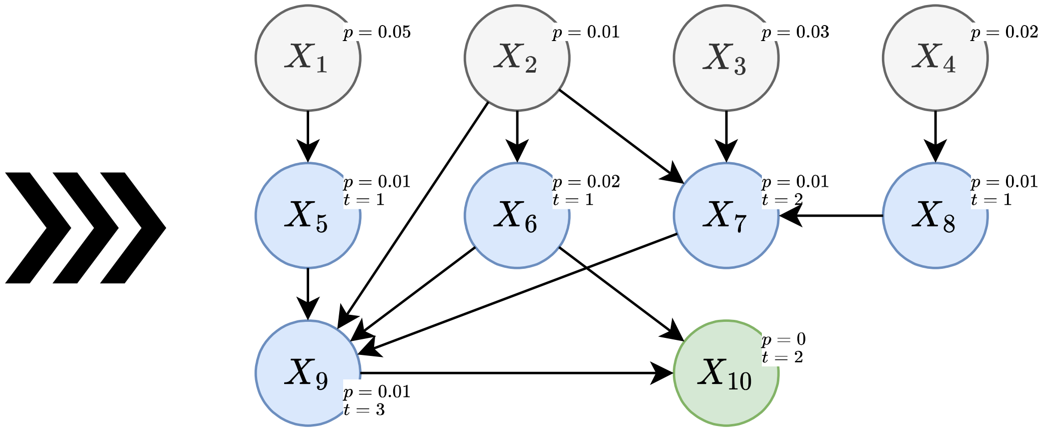
$$\mathcal{E}_v(V_C) = 1 - \frac{log(\mathbb{P}[Y \mid do(V_C)])}{log(\mathbb{P}[Y])}$$

- Loop through all possible coalitions, starting with smallest
- Stop when explanation score reaches a threshold α for some value $\alpha \in [0, 1]$ (close to 1)

Experiments and Results

RCA in Cloud Application

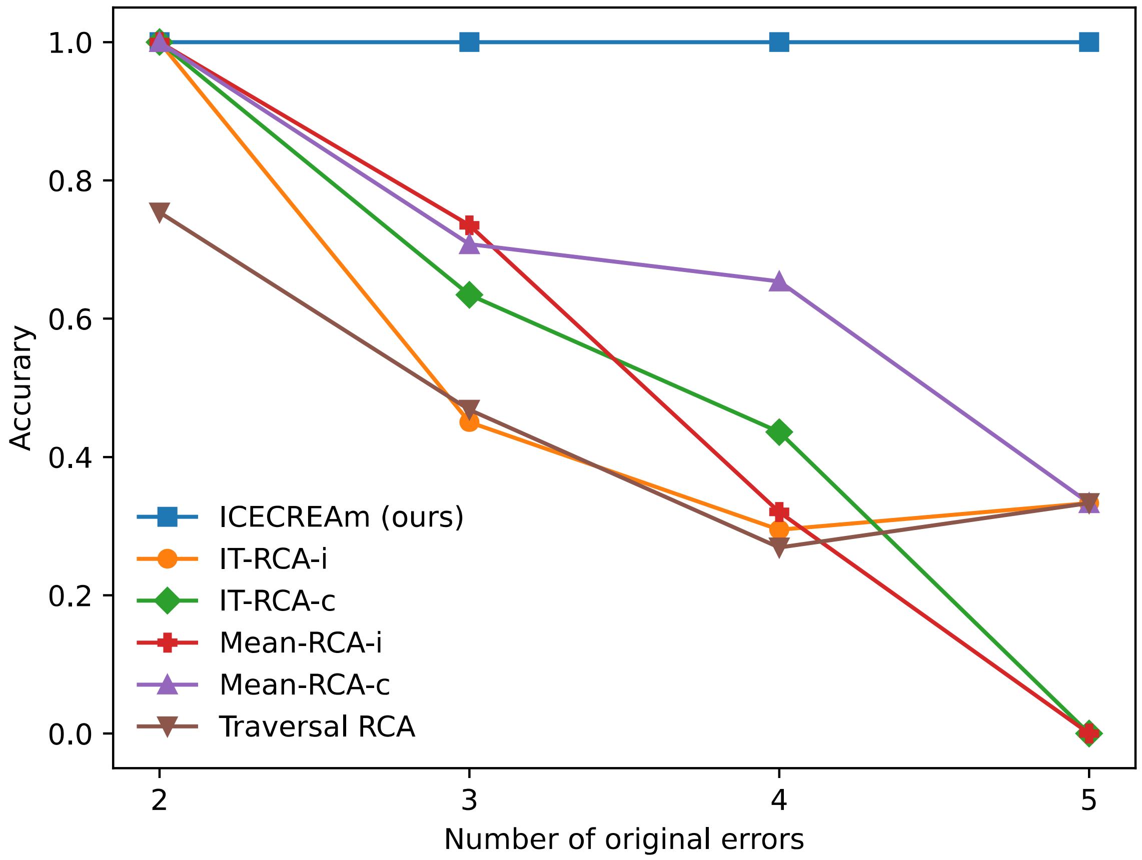




X_1 = order_db
X_2 = customer_db
X_3 = shipping
X_4 = product_db
X_5 = order
X_6 = auth
X_7 = product
X_8 = caching
X_9 = api
X_{10} = www

- System architecture is known (domain knowledge)
- Causal graph is inverted system topology
- Services produce errors with different probabilities
- Errors from parents are propagated if there are many enough

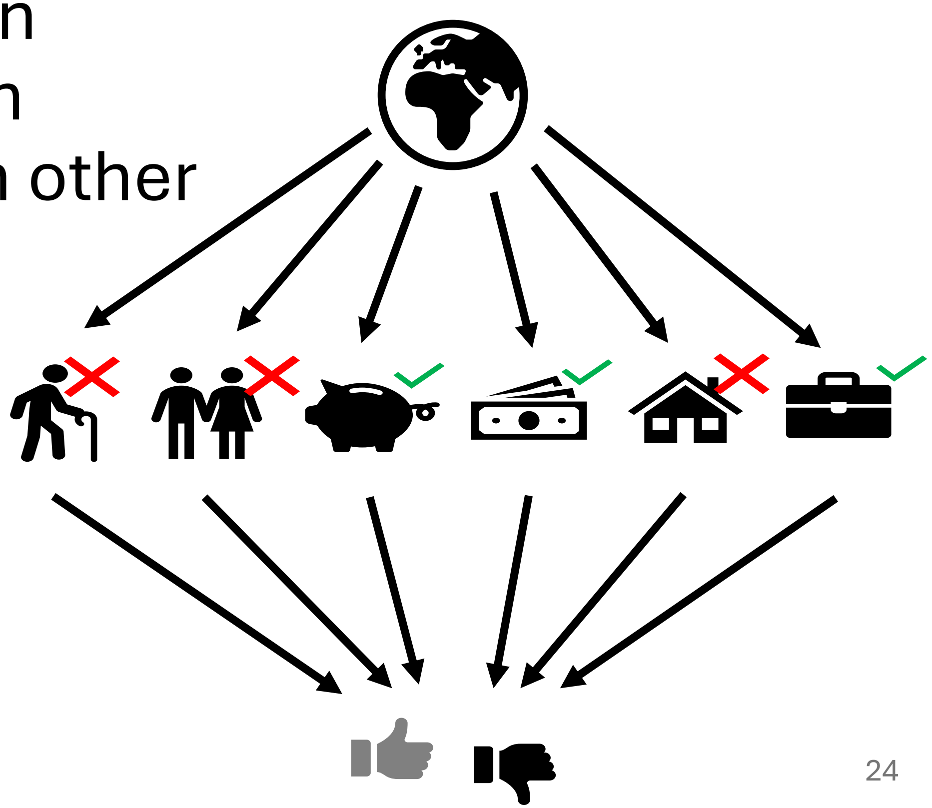
RCA in Cloud Application



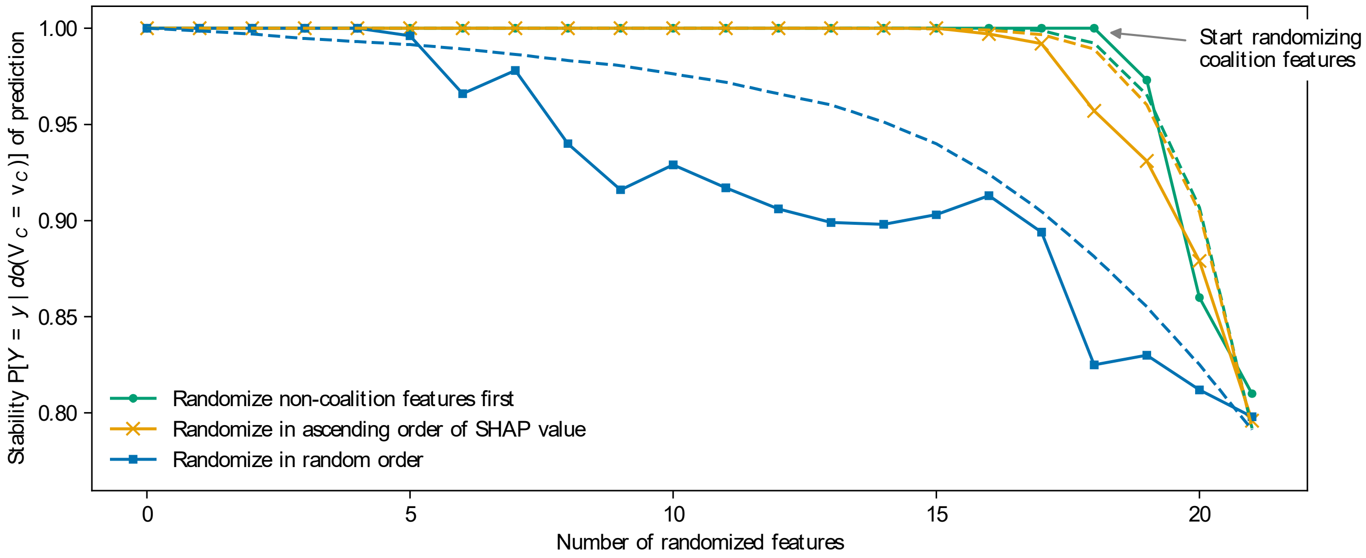
- Compared to RCA method by Budhathoki et al. (2022) and a simple traversal algorithm
- Only ICECREAM identifies the correct set of root causes if more than 1 issue is happening in the system

Explain ML Model’s Credit Risk Prediction

- Use the South German Credit data set (Grömping 2019) to train an ML model to predict if a credit application has high or low risk
- Consider the “real world” as a common cause to the input features, which then do not directly causally influence each other
- Compare ICECREAM to SHAP method



Explain ML Model’s Credit Risk Prediction



- When fixing the values of the coalition identified with ICECREAM, the prediction remains stable
- For the top features identified by SHAP the prediction remains also stable, but not as long as for ICECREAMs coalition

Conclusion

- Novel approach for explainability and RCA
- Able to identify minimal coalitions explaining a target value
- Novel explanation score satisfying all desired properties
- Evaluated on different data sets and with different use cases
  - Explain ML model: comparable performance to popular SHAP method
  - RCA in cloud application: outperforms state-of-the art methods when multiple root causes are present
- For practical applications, a more efficient algorithm needs to be developed, particularly for identifying optimal interventions

References

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In International Conference on Machine Learning, pages 2357–2369. PMLR, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NeurIPS'17, pages 4768–4777, Red Hook, NY, USA, 2017.

Ulrike Grömping. South German Credit Data: Correcting a Widely Used Data Set, November 2019.

Questions?