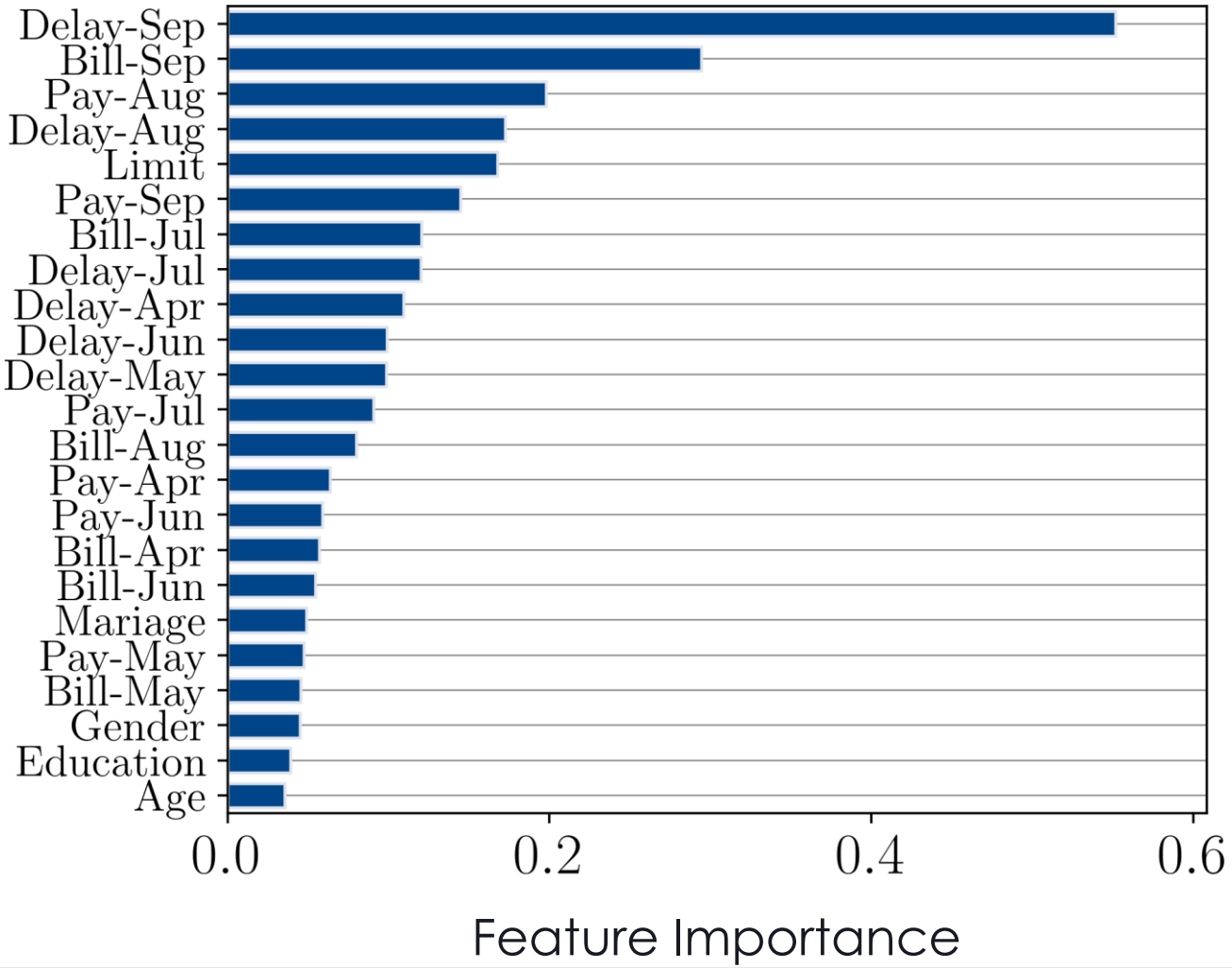
THALES
Building a future we can all trust

Tackling the XAI Disagreement Problem with Adaptive Feature Grouping

Gabriel Laberge
Ola Ahmad

www.thalesgroup.com

Feature Importance Scores (Tabular Data)



Saliency Maps (Image data)

{OPEN}

> Is the model leveraging context information (fence, branches) to make its prediction?












{OPEN}

The Disagreement Problem

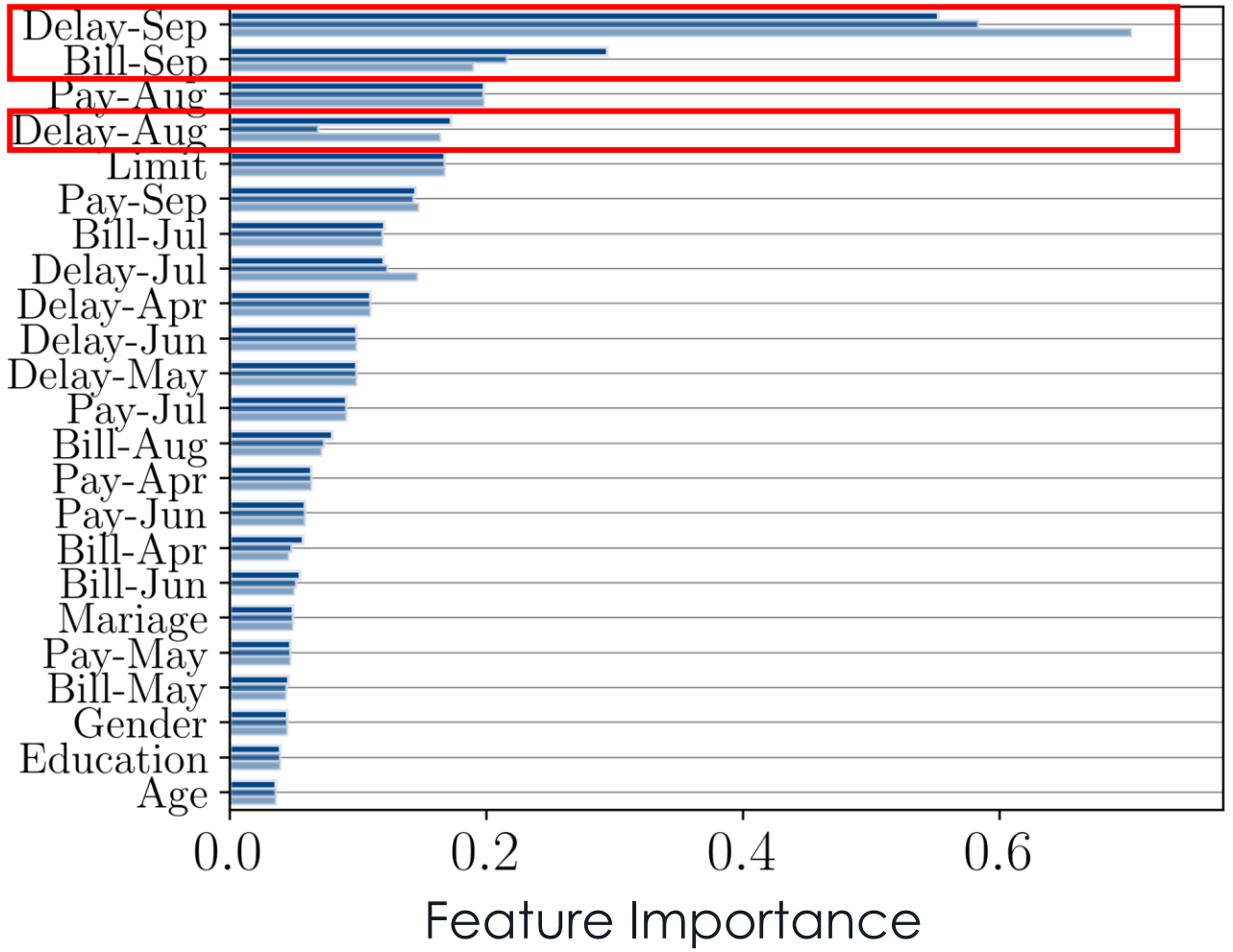
{OPEN}

> Tabular Explainability Techniques

- A plethora of techniques have been proposed to interpret predictive black-boxes on tabular data. The most notable examples are:
  - Partial Dependence Plots (PDPs) (Friedman, 2001)
  - Permutation Feature Importance (PFI) (Breiman, 2001)
  - SHAP (Lundberg et al, 2017)

Disagreement Problem

> Comparing Partial Dependence, SHAP, Permutation Feature Importance

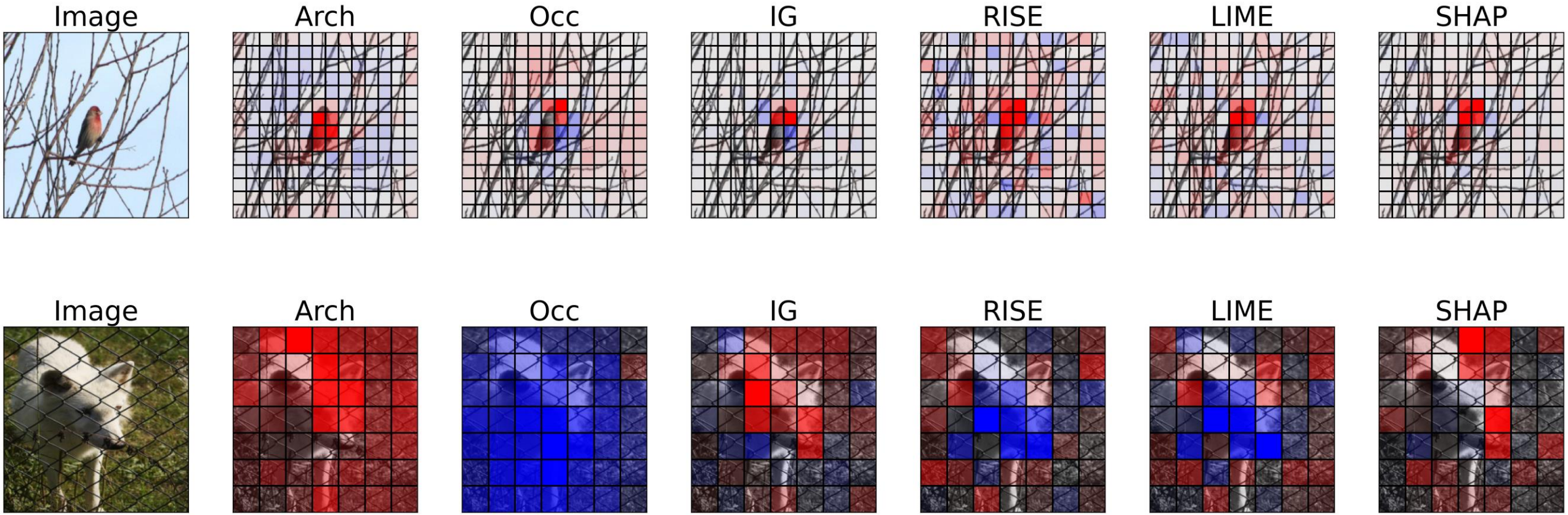


> Saliency Map Techniques

- Even more techniques have been proposed to interpret decision of image classifiers.
  - Occlusion (Zeiler et al., 2014)
  - LIME (Ribeiro et al., 2016)
  - IG (Sundararajan et al., 2017)
  - SHAP (Lundberg et al., 2017)
  - RISE (Petsiuk et al., 2018)
  - Arch Attribute (Tsang et al., 2020)

Challenge

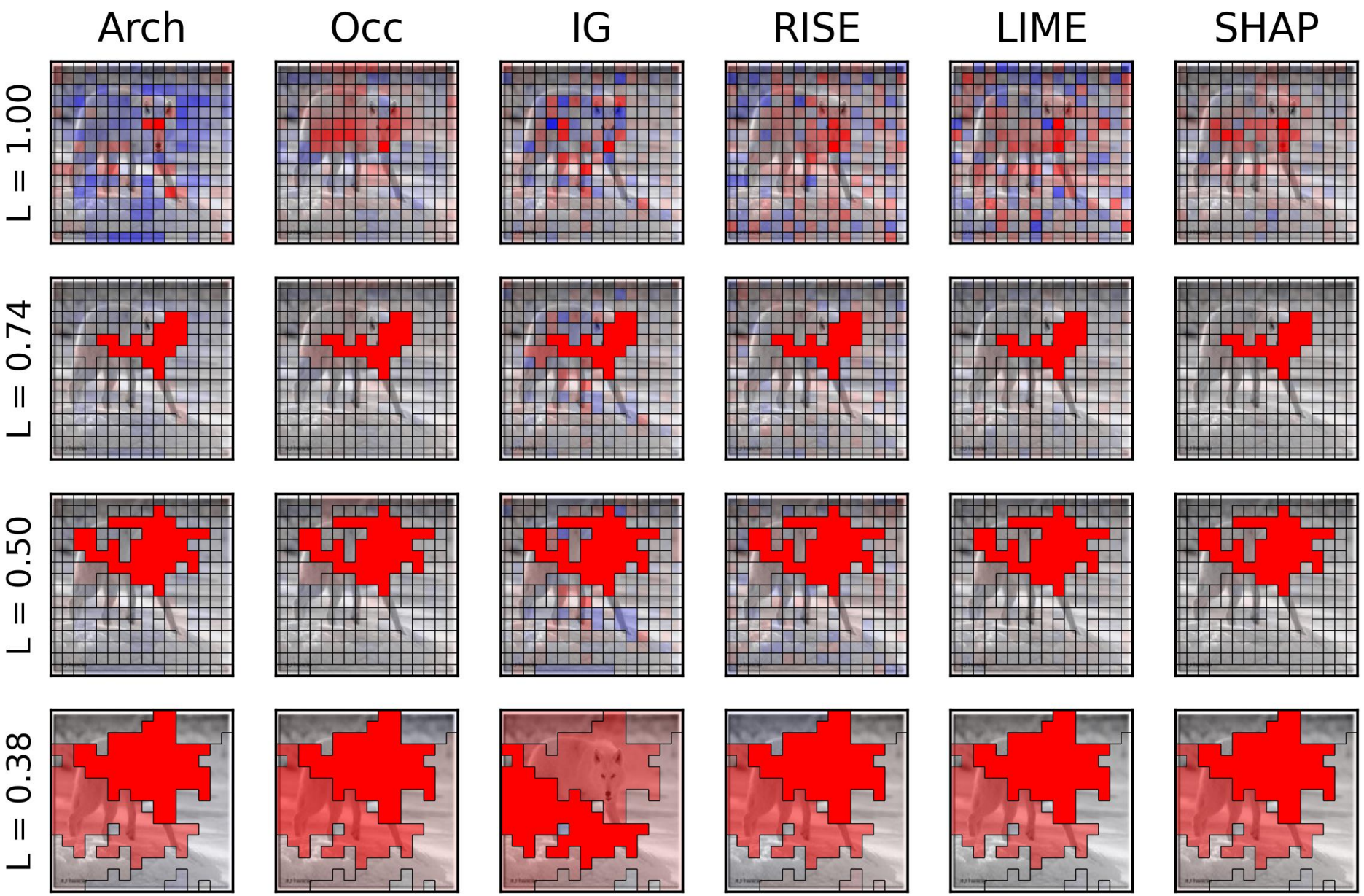
Heatmap : **red** (toward predicted class), **blue** (away-from predicted class)



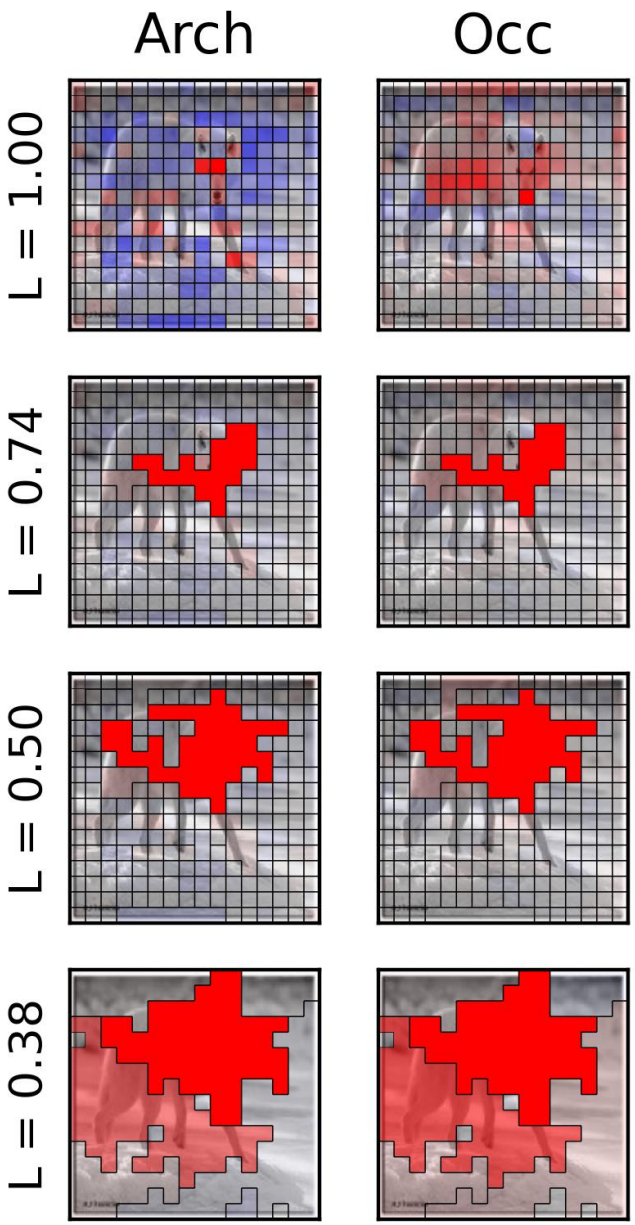
{OPEN}

AGREED

{OPEN}

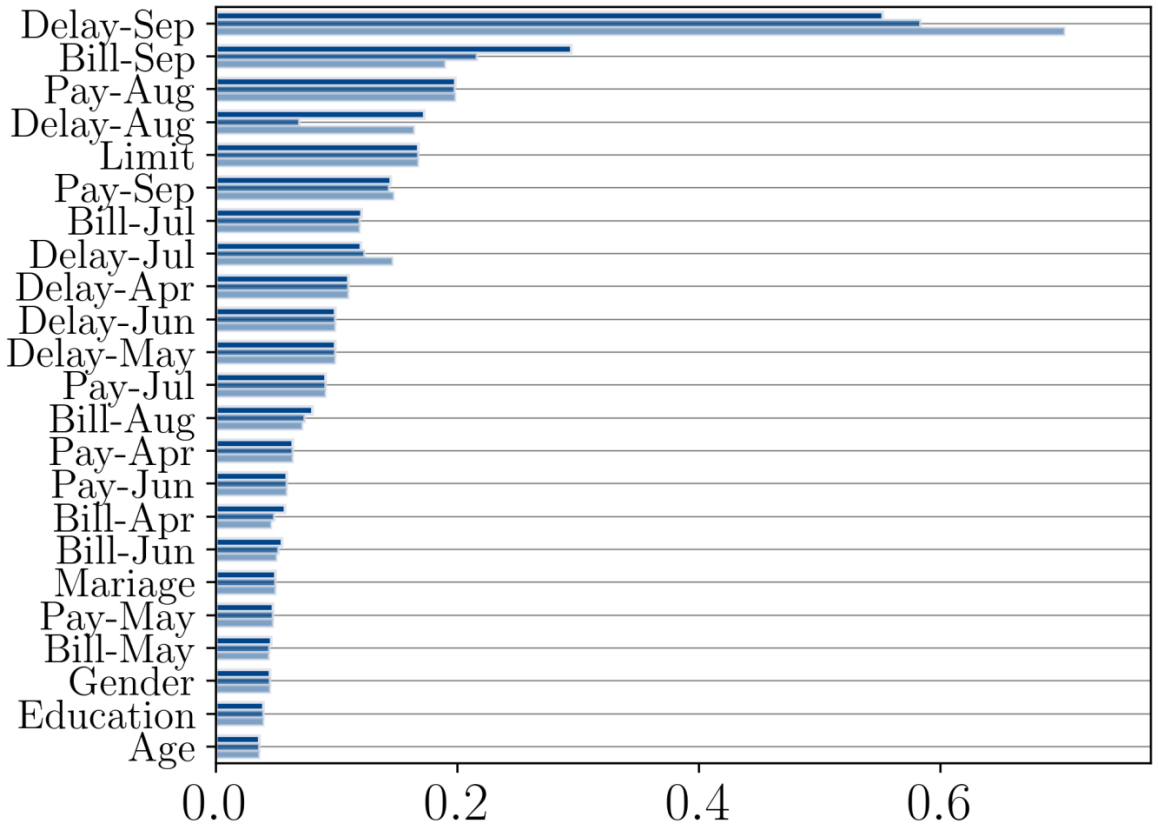


{OPEN}

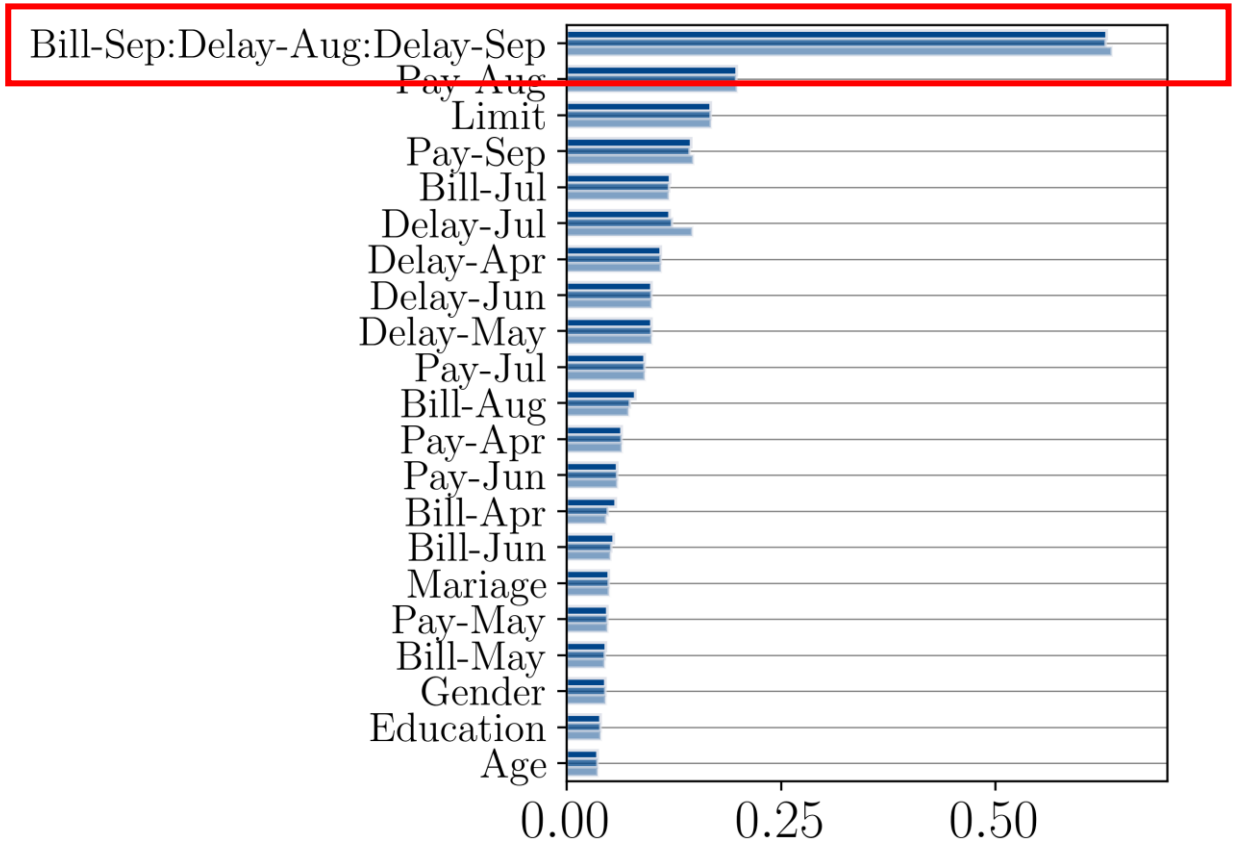


Unseen when finding the groups

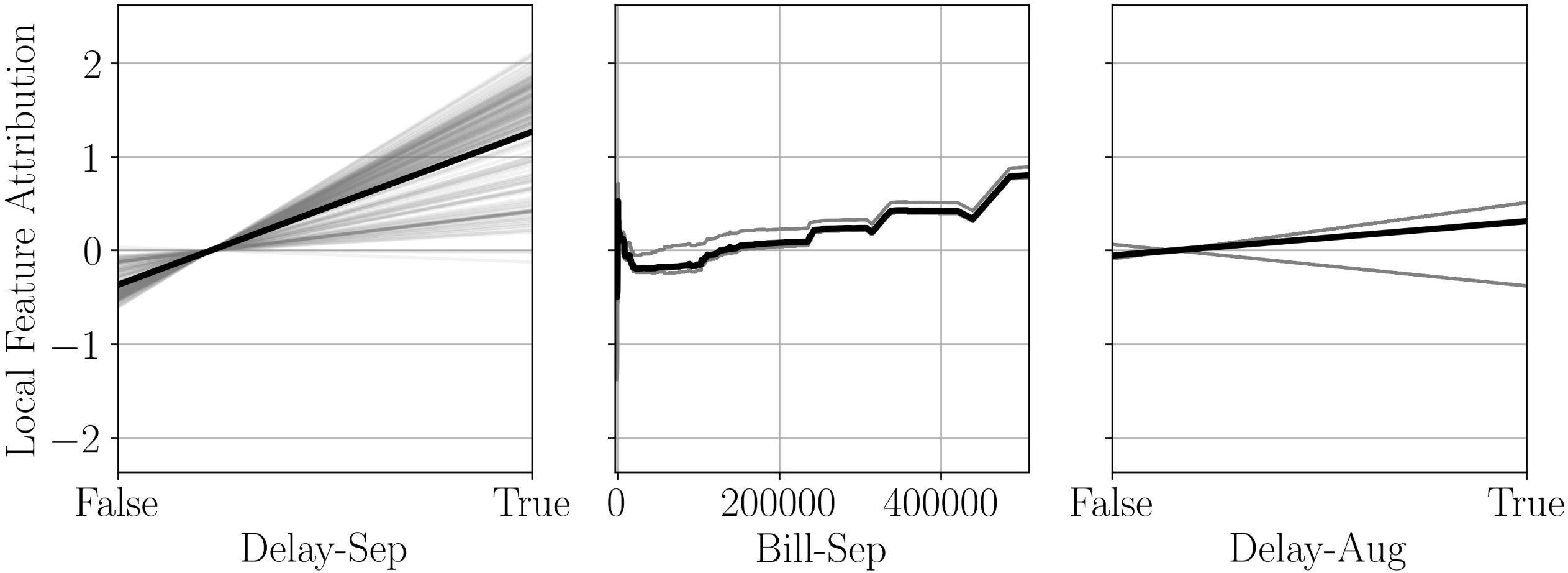
No Feature Groups

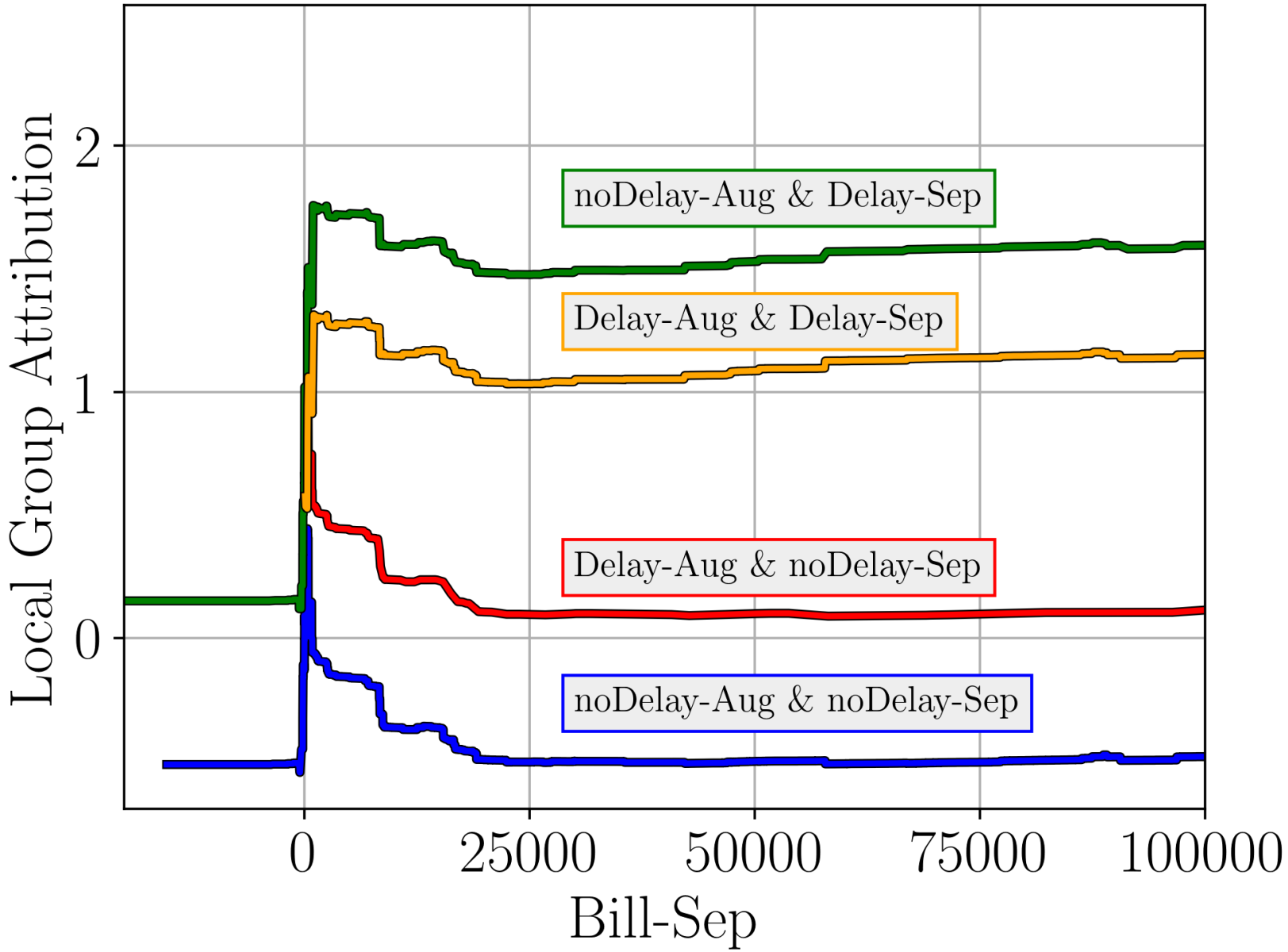


With Feature Groups



How do we interpret the contribution of a group???





THALES
Building a future we can all trust

Thank you

www.thalesgroup.com