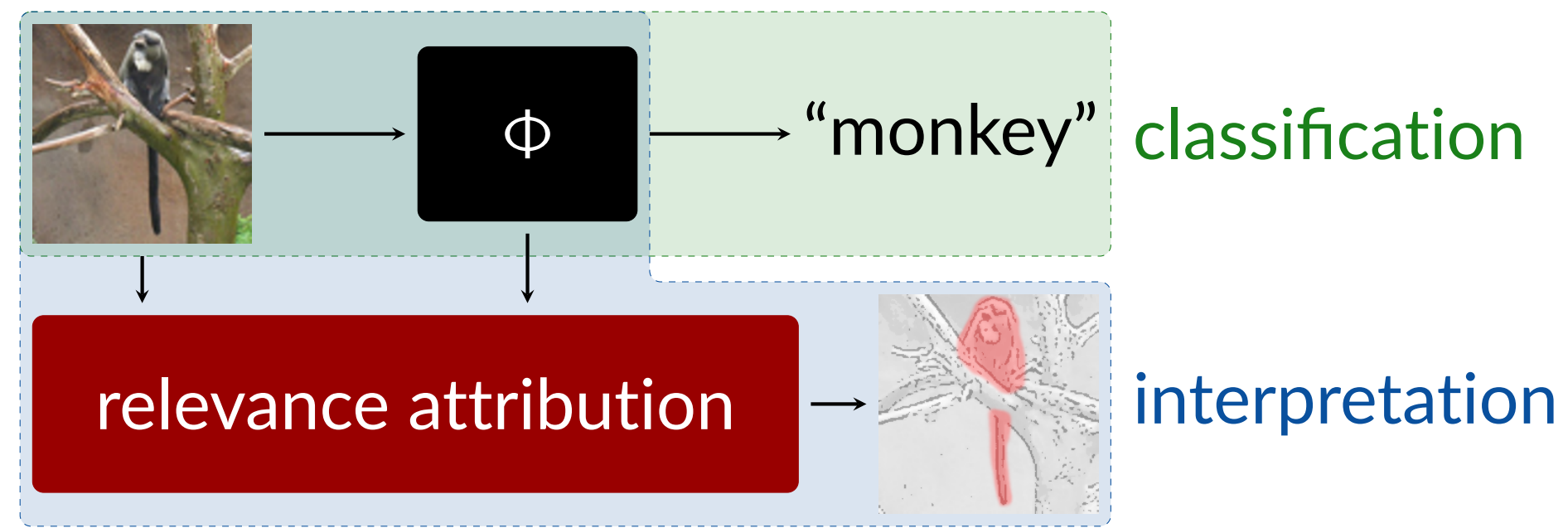


Quick Tour



Rate-Distortion Explanations

$$\|s\|_1 \text{ vs. } D(s) = \mathbb{E}_n[(\Phi(x) - \Phi(s \odot x + (1-s) \odot n))^2]$$

Rate-Constrained RDE (RC-RDE)

$$\text{minimize } D(s) \text{ subject to } \|s\|_1 \leq k, s \in [0, 1]^n$$

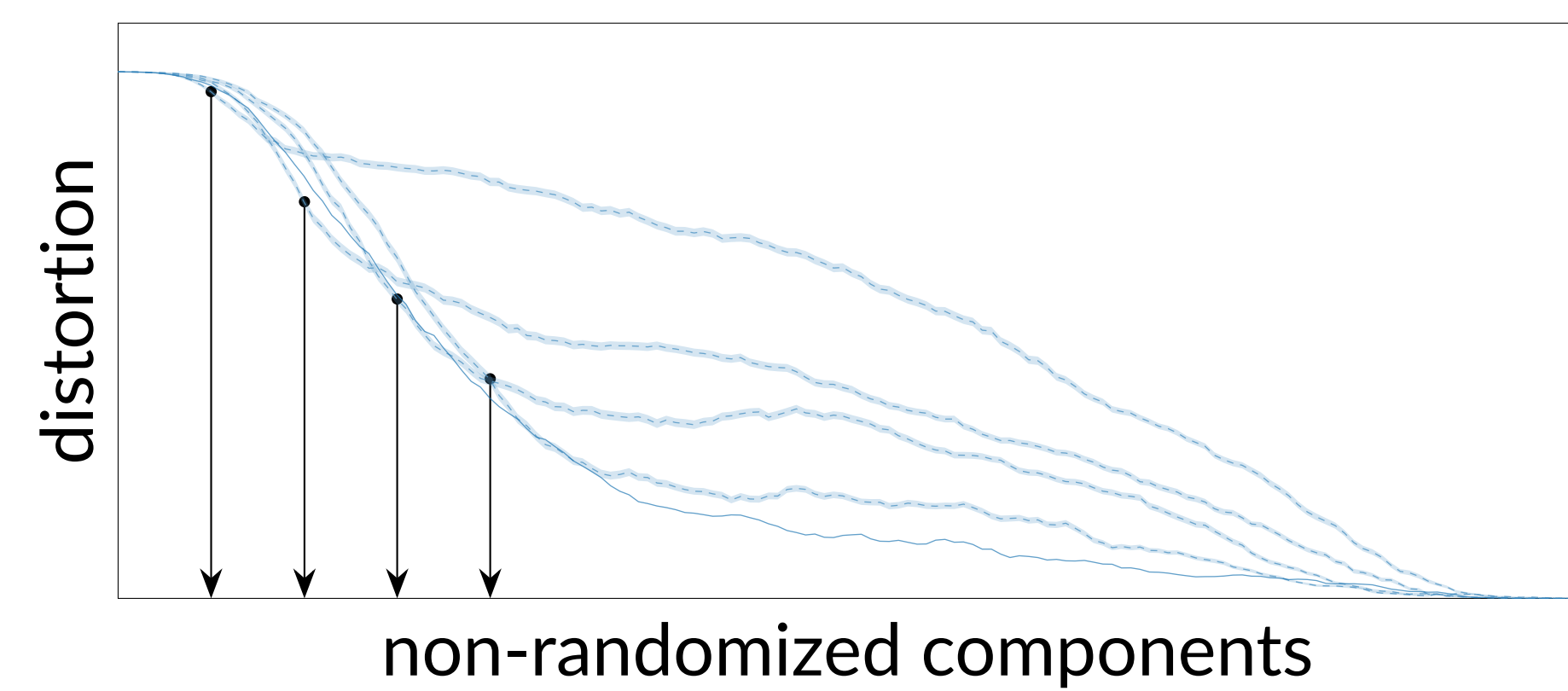
Ordering RDE (Ord-RDE)

$$\text{minimize } \sum_{k=1}^{n-1} D(\Pi p_k) \text{ subject to } \Pi \in B_n$$

- p_k vector of k ones and $n - k$ zeros
- B_n Birkhoff polytope ($n \times n$ doubly stochastic matrices)

Multi-Rate RDE (MR-RDE)

Combine multiple RC-RDE solutions at different rates k to approximate Ord-RDE.



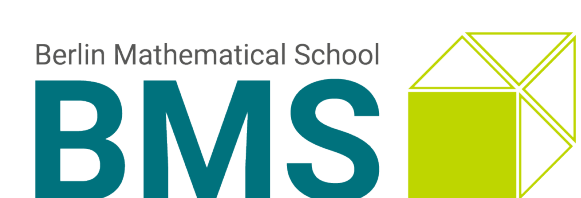
Sparse relevance maps and relevance orderings can be obtained with Frank-Wolfe algorithms.

Interpretable Neural Networks with Frank-Wolfe: Sparse Relevance Maps and Relevance Orderings

Jan Macdonald
Technische Universität Berlin

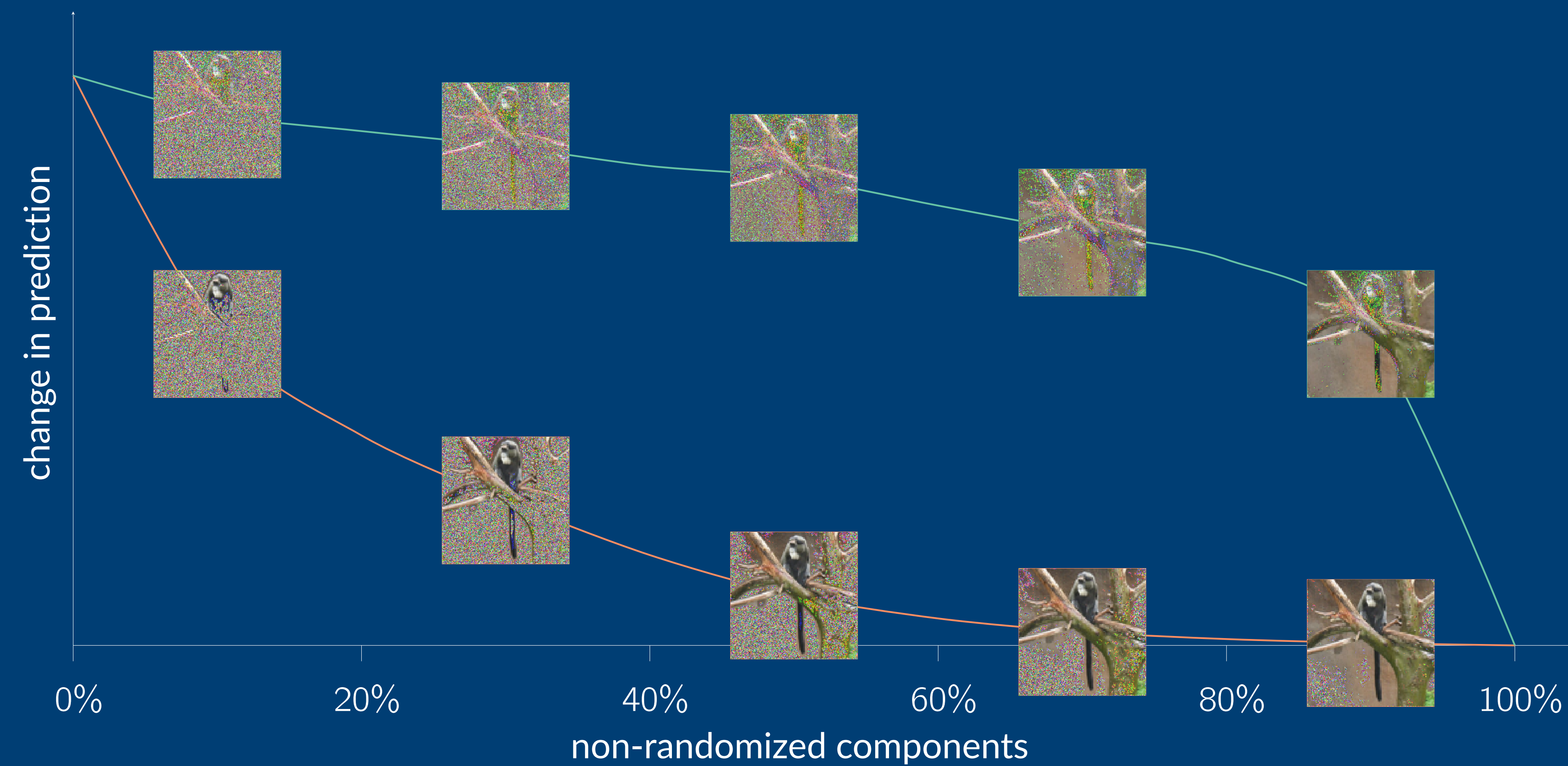
Mathieu Besançon
Zuse Institute Berlin

Sebastian Pokutta
Technische Universität Berlin
Zuse Institute Berlin

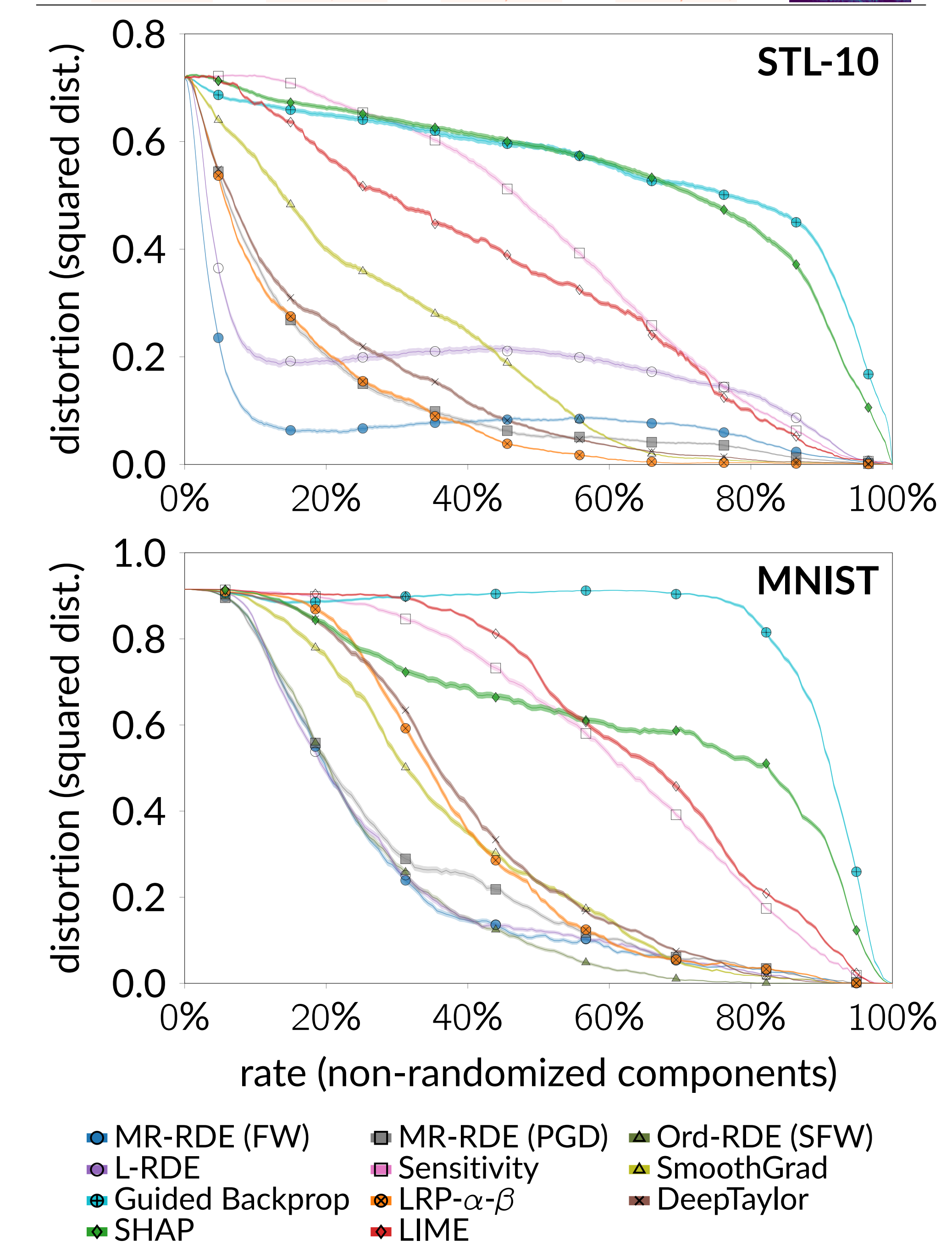
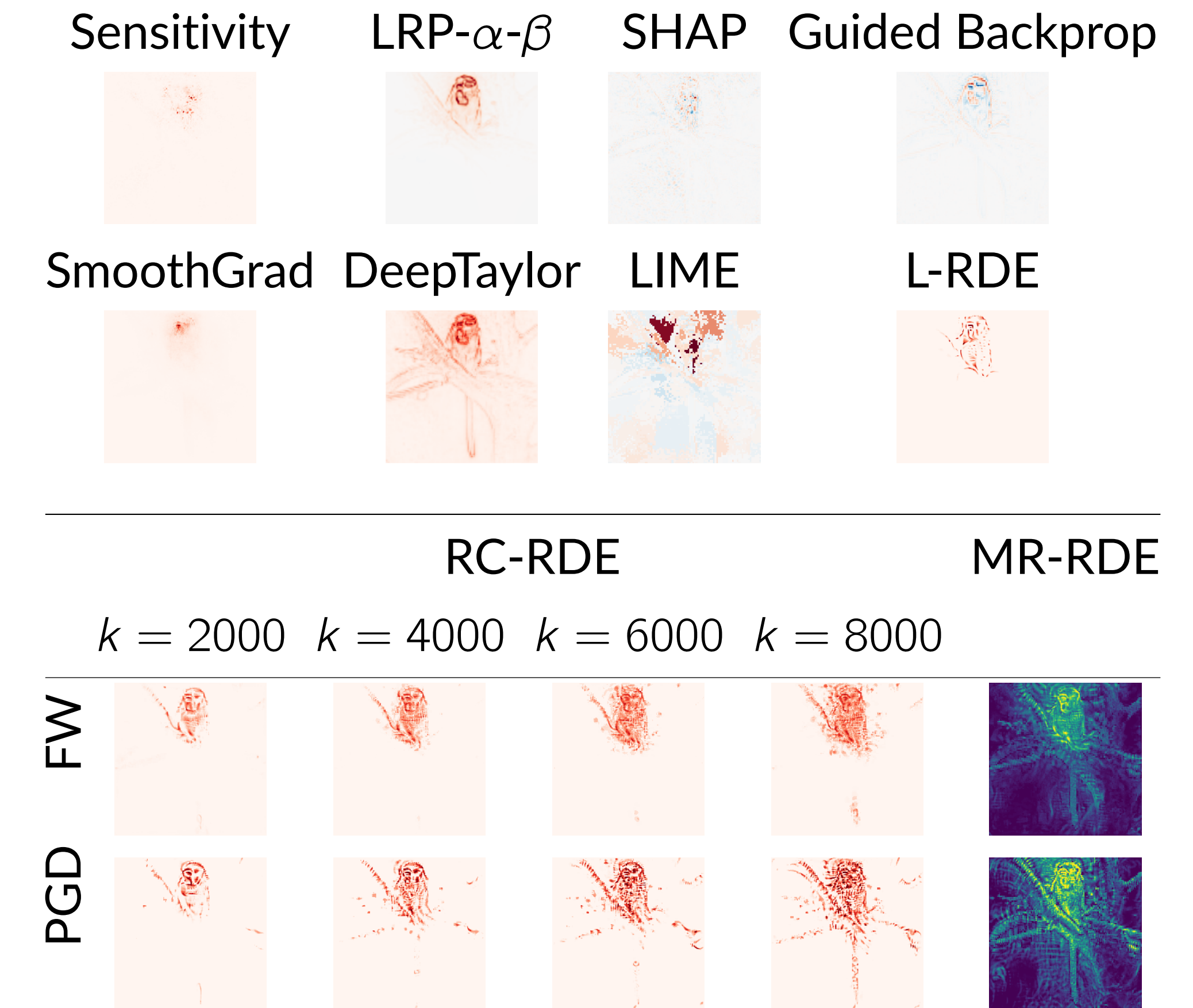


What do relevance attribution scores tell us?

The induced relevance ordering is more meaningful and useful for evaluating and comparing methods.



Results



RDE identifies the relevant components and achieves a steep decrease in the distortion.

Sensitivity [Simonyan et al. 2013]
Smoothgrad [Smilkov et al. 2017]
Guided Backprop [Springenberg et al. 2015]
LRP- α - β [Bach et al. 2015]
DeepTaylor [Montavon et al. 2018]
SHAP [Lundberg and Lee 2017]
LIME [Ribeiro et al. 2016]
L-RDE [Macdonald et al. 2019]



Scan to view the full paper on arXiv.
Code is available on Github.