# What makes an Ensemble (Un) Interpretable?

## Motivation


***Why?***

Can we provide a *mathematical* answer?

We take many explanation types and study the ***computational complexity*** of computing them for ***individual models*** vs. ***ensembles.***

## Classic complexity results

SHAP Explanations
Minimum Contrastive/Counterfactual Explanations
Minimum Sufficient Explanations
Probabilistic Explanations
Robust notions of Explanations

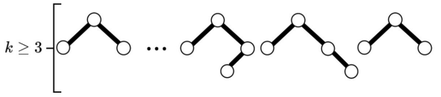
| **Explaining base models** | VS | **Explaining ensembles** |
|---|---|---|
| Decision trees/ Linear models | | ***Ensembles*** of Decision trees/Linear models |
| Polynomial Time/ Pseudo-Polynomial Time | | NP, coNP, #P, $\Sigma_2^P$ -Complete |
| Neural Networks | | ***Ensembles*** of Neural Networks |

*Provably* the same complexity.

## ***Parameterized*** complexity results

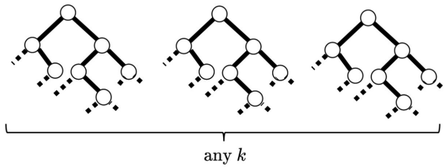
An ensemble with **many tiny models** is ***still computationally hard to explain!***

**NP-Hard (at least...)**



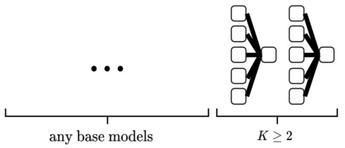

In ensembles with **a low number of decision trees**, ***explanations can be computed efficiently!***

**Polynomial!**




Ensembles with **even two linear models**, ***are already computationally hard to explain!!***

**NP-Hard (at least...)**




Shahaf Bassan, Guy Amir, Meirav Zehavi, Guy Katz; Hebrew University of Jerusalem, Cornell University, Ben Gurion University of the Negev