On the Robustness of Interpretability Methods

1. Is this robustness given necessarily true? If I'm training parity, 1&2 are "close", but they should give different inputs.
2. Could robustness be determined by the fact that multiple models can have the same explanation

Interpreting Deep Learning Models

1. Is Mojito just explaining Deep Learning models for Entity Resolution?

"Why should I trust you"

1. What do you think is the most important thing for a classifier to report to trust it?
2. Do explanations just benefit the user from an understandability standpoint?
3. Could you clarify how instances are chosen to be explained?