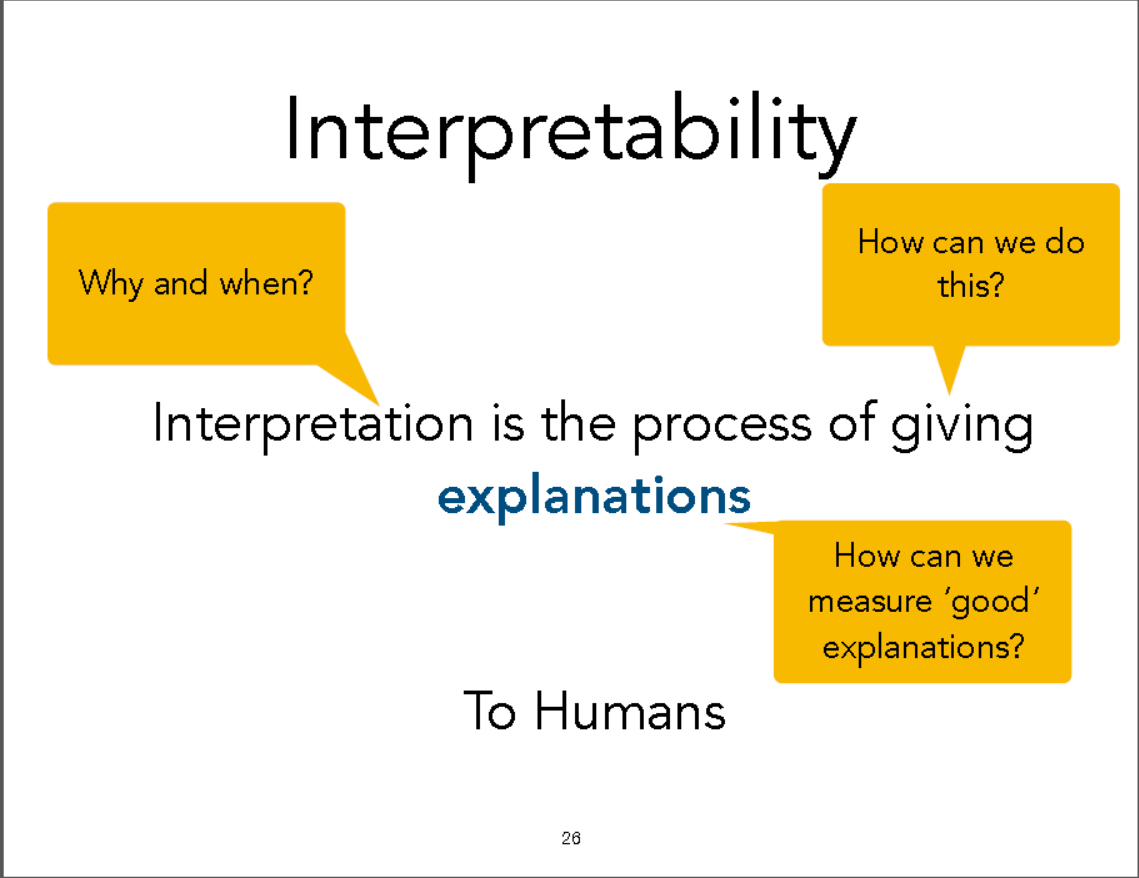
What is the interpretation?

Interpretation is the process of giving explanations to humans.

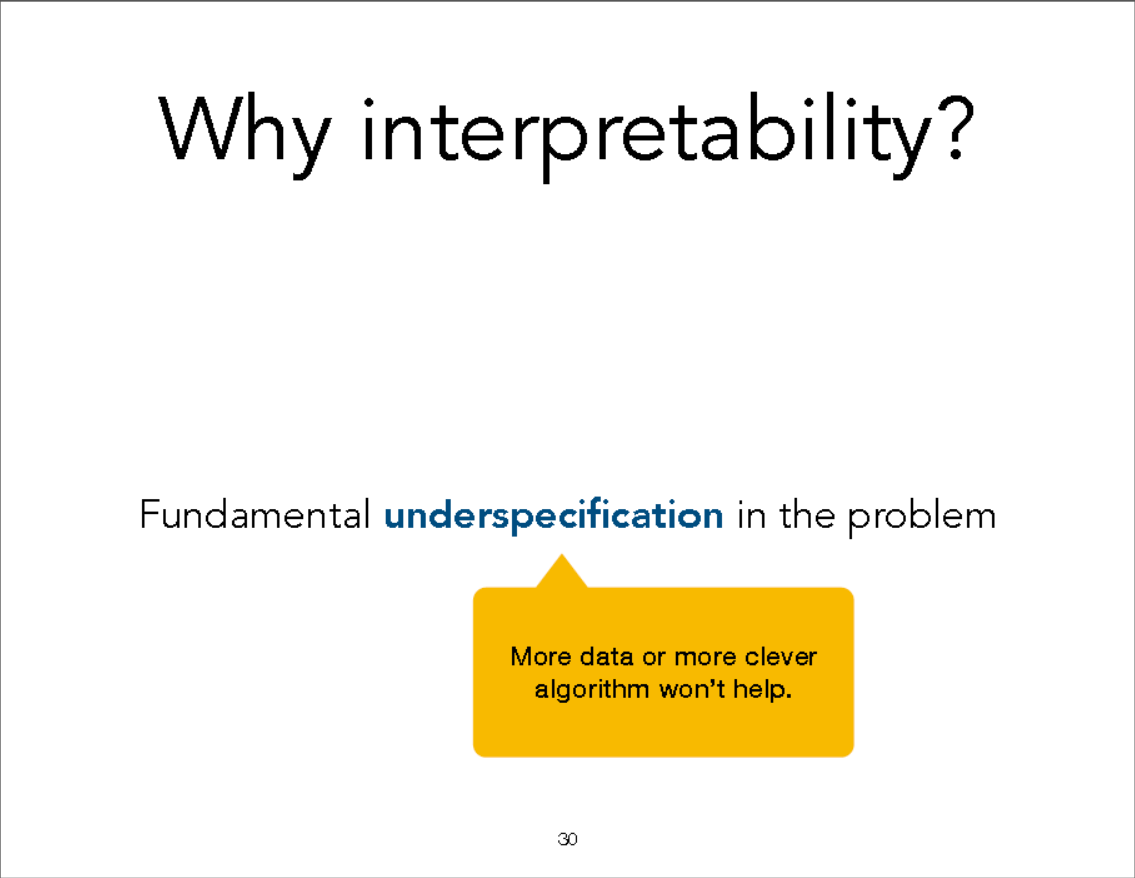


以下内容均围绕三个目标

Why and when?

How can we do this?

How can we measure ‘good’ explanations?



Under-specification and interpretability works example：

Such as predict safe problem/medical issues/debugging/ mismatched objectives and multi-objective trade-offs/science/legal/ethics

Examples of underspecification

• Safety: We want to make sure the system is making sound

decisions.

• Debugging: We want to understand why a system doesn't work,

so we can fix it.

• Science: We want to understand something new.

• Mismatched Objectives and multi-objectives trade-offs: The

system may not be optimizing the true objective.

• Legal/Ethics: We're legally required to provide an explanation

and/or we don't want to discriminate against particular groups.

+ Your case?

Fundamental underspecification in the problem

Before building

any model

• Visualization

• Exploratory data analysis

Building a new model

Types of interpretable methods

• rule-based, per-feature based

• case-based

• sparsity

• monotonicity

1. The limitation of rule-based methods.
2. It may not be as interpretable as you may think
   1. Depth/Length of the tree might be too big
   2. Complexity of rules might be high
   3. May need lots of splits to fit complex function
3. It only works if the original features are interpretable
4. Case-based

“I recommend treatment X because it worked for other patients like you...”

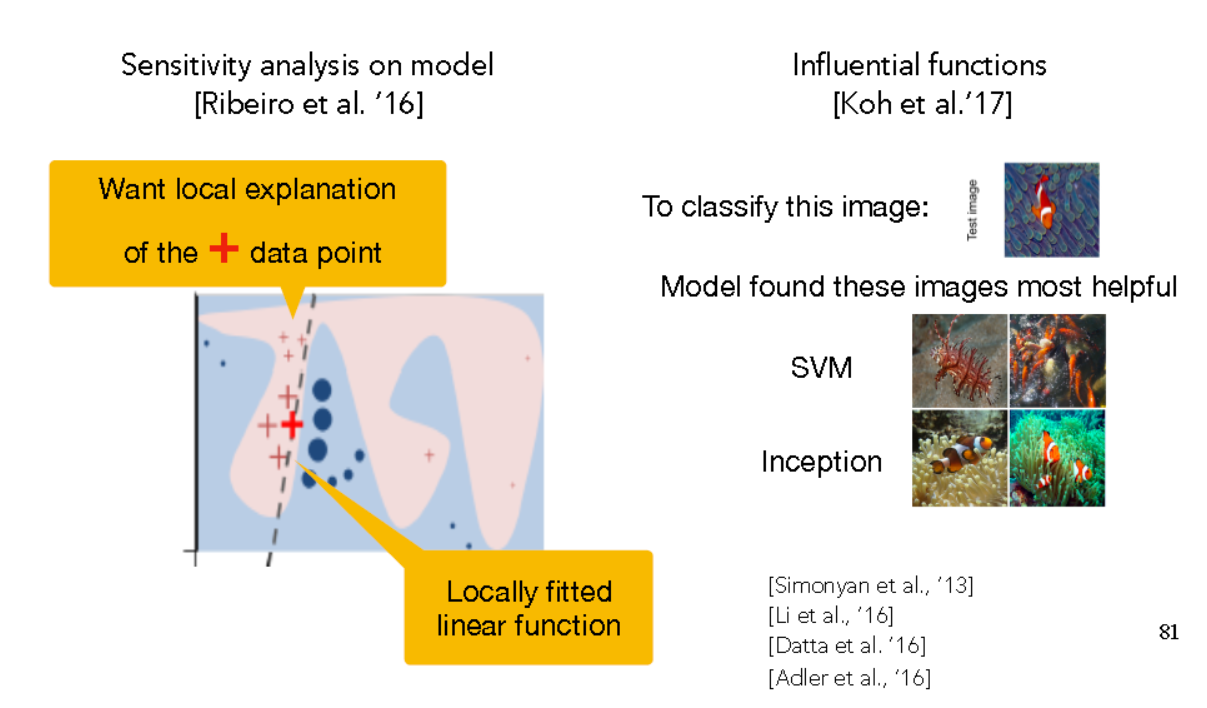
• Explain clustering results using examples (Bayesian Case Model)

• Joint inference on prototypes, prototypes subspaces cluster labels

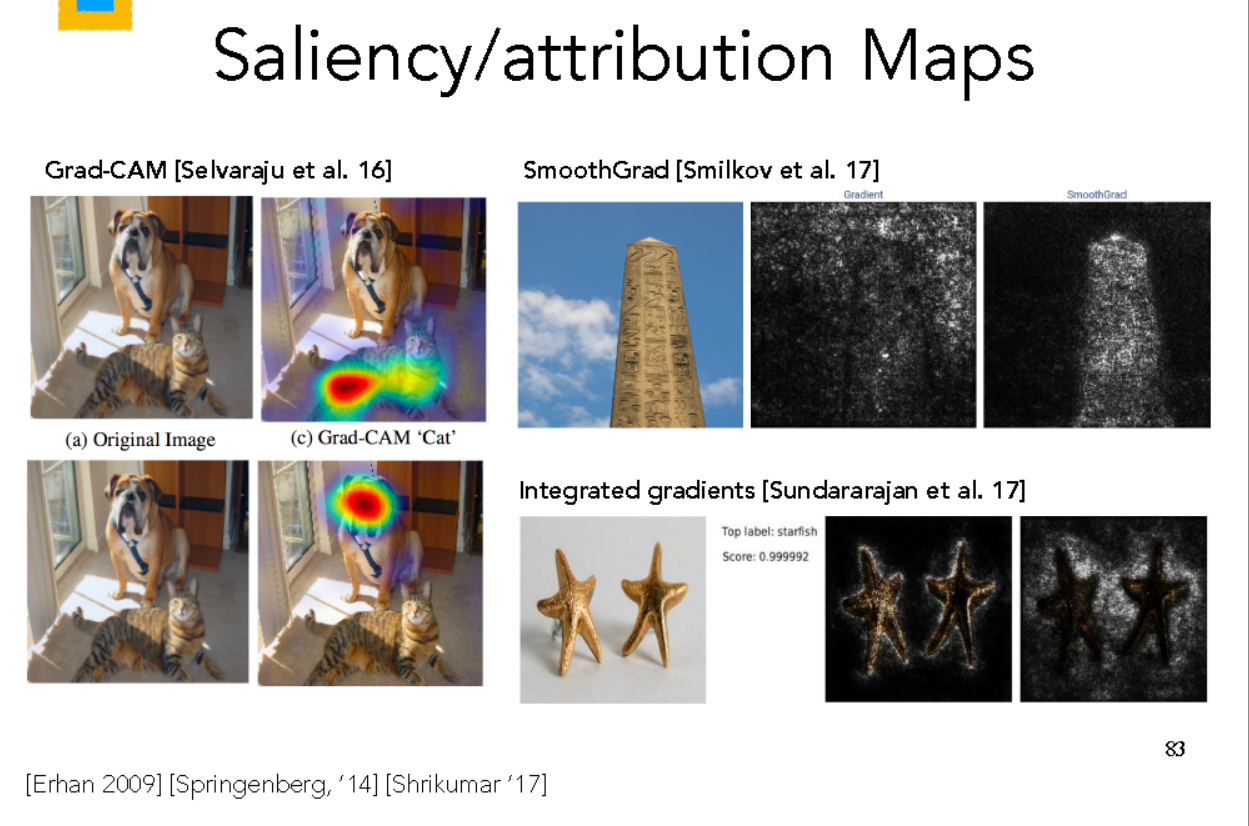
After building a new model

Types of interpretable methods

• Sensitivity analysis, gradient-based methods



• mimic/surrogate models (Saliency/attribution Maps)



• Model compression or distillation [Bucila et al. ’06, Ba et al. ’14,

Hinton et al. ’15]

• Visual explanations [Hendricks et al. ’16]

• Investigation on hidden layers

