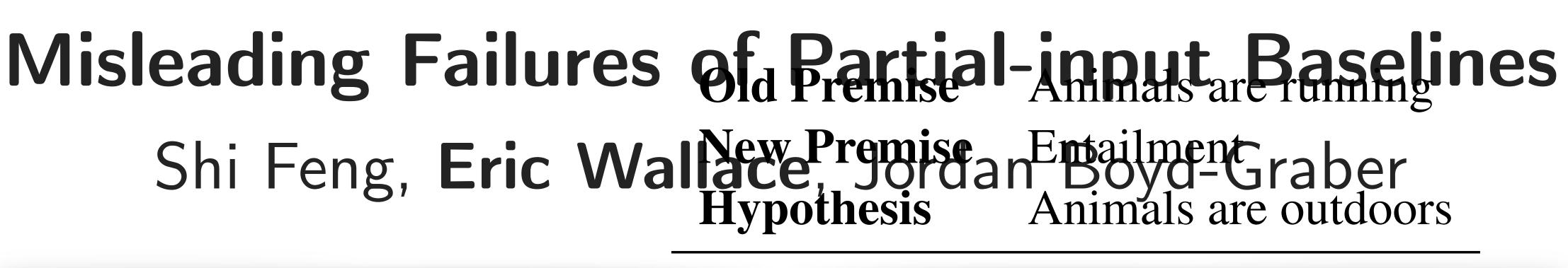


Label



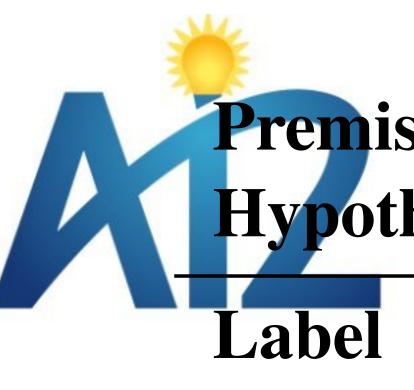


Table 3: Each

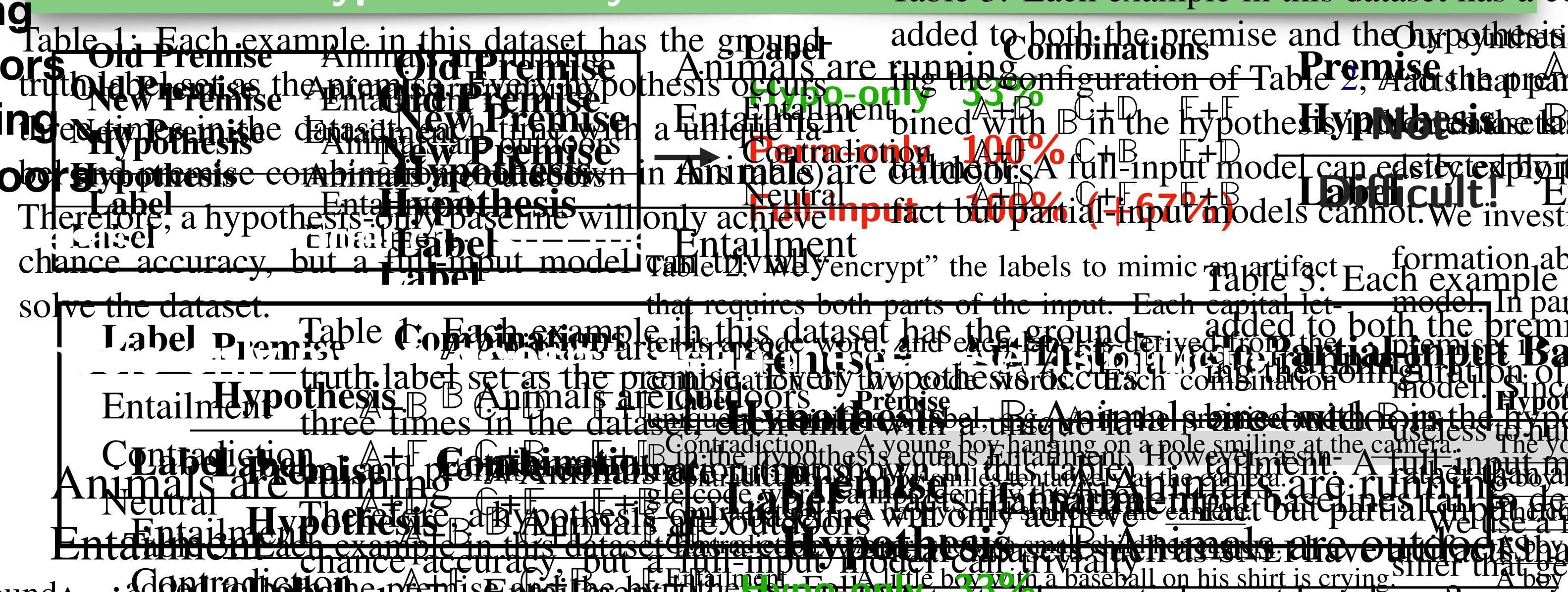
How do we measure the difficulty of a dataset?

PremiseAnimals are runningHypothesisAnimals are outdoors

Entailment

Table 1: Each example in this dataset has the groundtruth label set as the premise. Every hypothesis occurs three times inche dataset 2 such time with a unique label and premise combination (not shown in this table). Therefore, a hypothesis on **Fypethesis** on **B** a **Amive IS face puperties** thanks a ccuracy, but a full-input model can trivially solve the dataset. **Label**

Label But Entrienosithypothesis-only model doesn't mean the clatastifics Failure of hypothesis-only model doesTable 3: Each example in this dataset has a co



Dund An ihren in ander ihren in an en in an en in an en in a some toget fib y partial input baschart in attain curs in the first and the provident of the present of the p

