Inferring Generative Model Structure with Static Analysis
Paroma Varma, Bryan He, Payal Bajaj, Nishith Khandwala, Imon Banerjee, Daniel Rubin, Christopher Ré

**Summary**
- **Generative Models to Label Training Data**
  Use generative models to combine sources of weak supervision to assign noisy labels
- **Complex Dependencies among Sources**
  Sources of labels are rarely independent
- **Inferring Model Structure**
  Use static analysis to infer dependencies and encode in generative model structure

**Theoretical Results**
- **Learning Dependencies**
  Learning k-degree dependencies among n heuristics requires $O(n^{k-1} \log n)$ samples
- **Inferring Dependencies**
  Analyzing the code can infer the dependencies among heuristics without data
  Given dependencies, learning heuristic accuracies requires $O(n \log n)$ samples

**Experimental Results**
- Inferring dependencies outperforms learning dependencies
- Outperforms fully supervised model with additional noisy training labels

**Heuristic Structure**
- **Domain Specific Primitives (DSPs)**
  Interpretable characteristics of raw data
- **Heuristic Functions (HFs)**
  Programmatic rules that output noisy labels

**Static Analysis**
- **Shared Input** Sharing primitives as inputs leads to explicit dependencies
- **Compositions**: Primitives composed of others can lead to implicit dependencies

**Statistical Modeling**
- **HF Dependency** Represents the dependencies found using static analysis
- **DSP Similarity** Represents the learned correlations among the DSPs