Snorkel: Rapid Training Set Creation with Weak Supervision
Alex Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Chris Ré

The Traditional ML Pipeline has a New Bottleneck

- Collect Raw Data
- Create a Training Set
- Engineer Features
- Fit a Model

Now we use deep learning, but it’s data hungry!

Snorkel + Data Programming: A Unifying Framework for Weak Supervision

1. Users write labeling functions to generate noisy labels
2. We model the labeling functions’ behavior to denoise them
3. We use the estimated probabilistic labels to train a model

Generating Noisy Labels with Labeling Functions

In Snorkel, users write labeling functions (LFs), which are just scripts that noisily label subsets of the data. Ex: Labeling relations in text based on an existing knowledge base:

```
def if1(x):
    cid = (x.gene, x.pheno)
    if cid in KB:
        return Label=TRUE
    else:
        return Label=FALSE
```

We provide an optimizer with the following options:

- Base learners (e.g., decision trees, SVMs)
- Distant supervision
- Crowdsourcing
- Domain expert heuristics

Snorkel Workshop: User Study
71% of first-time, non-expert users beat 7 hours of hand-labeling in 2hrs. with Snorkel, with 45.5% avg. improvement

Computational Tradeoffs Modeling Weak Supervision

- New time-accuracy tradeoff space for modeling weak supervision:
  - When to model the source accuracies: We provide an optimizer with theoretical guarantees that can speed up initial dev. cycles
  - How much structure to model: Can speed up modeling by up to 10x!

Results on Real-World Text & Image Tasks

- With tens of LFs, we improve on DS/heuristic baselines, and come within points of expensive, hand-labeled datasets

Open-source code and tutorials: snorkel.stanford.edu