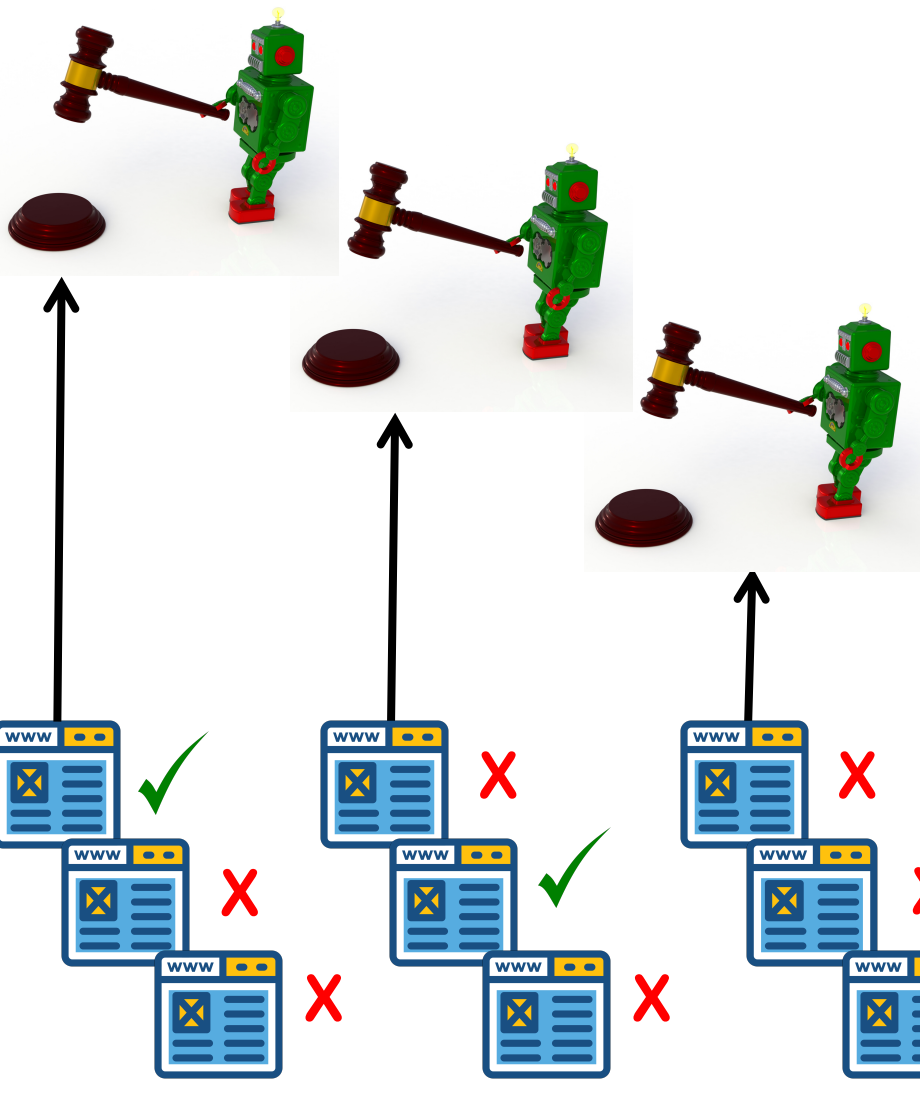


Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale



Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alexander Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Christopher Ré, Rob Malkin

Motivation

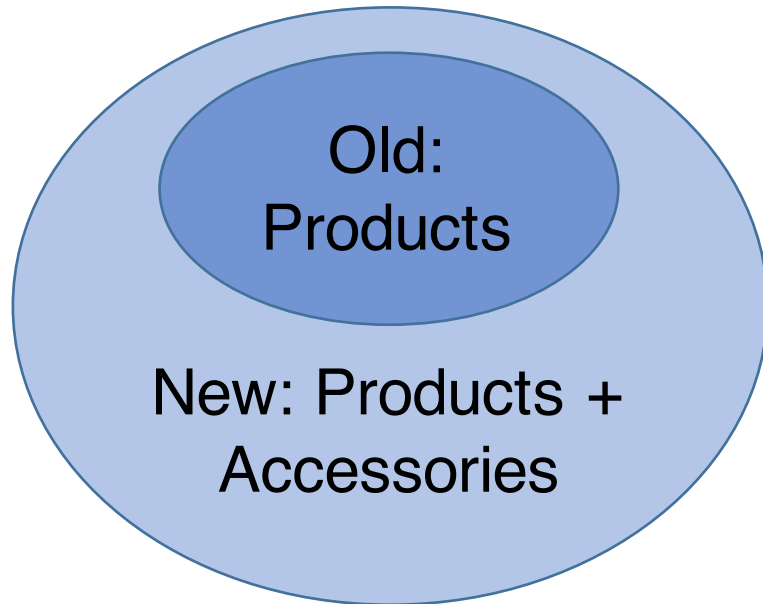


- Organizations using machine learning often **manage many different training data sets**
- These training sets are **costly to create** and **need to be updated as strategy changes**
- How can we **replace hand-labeled training data** with **existing organizational resources**?
- Can we **transfer knowledge** from resources not servable in production environments?

Case Studies at Google

Product Classification

- Existing classifier used to detect products in a category of interest
- Goal: move accessories to positive class
- Instant depreciation of investment in labels!

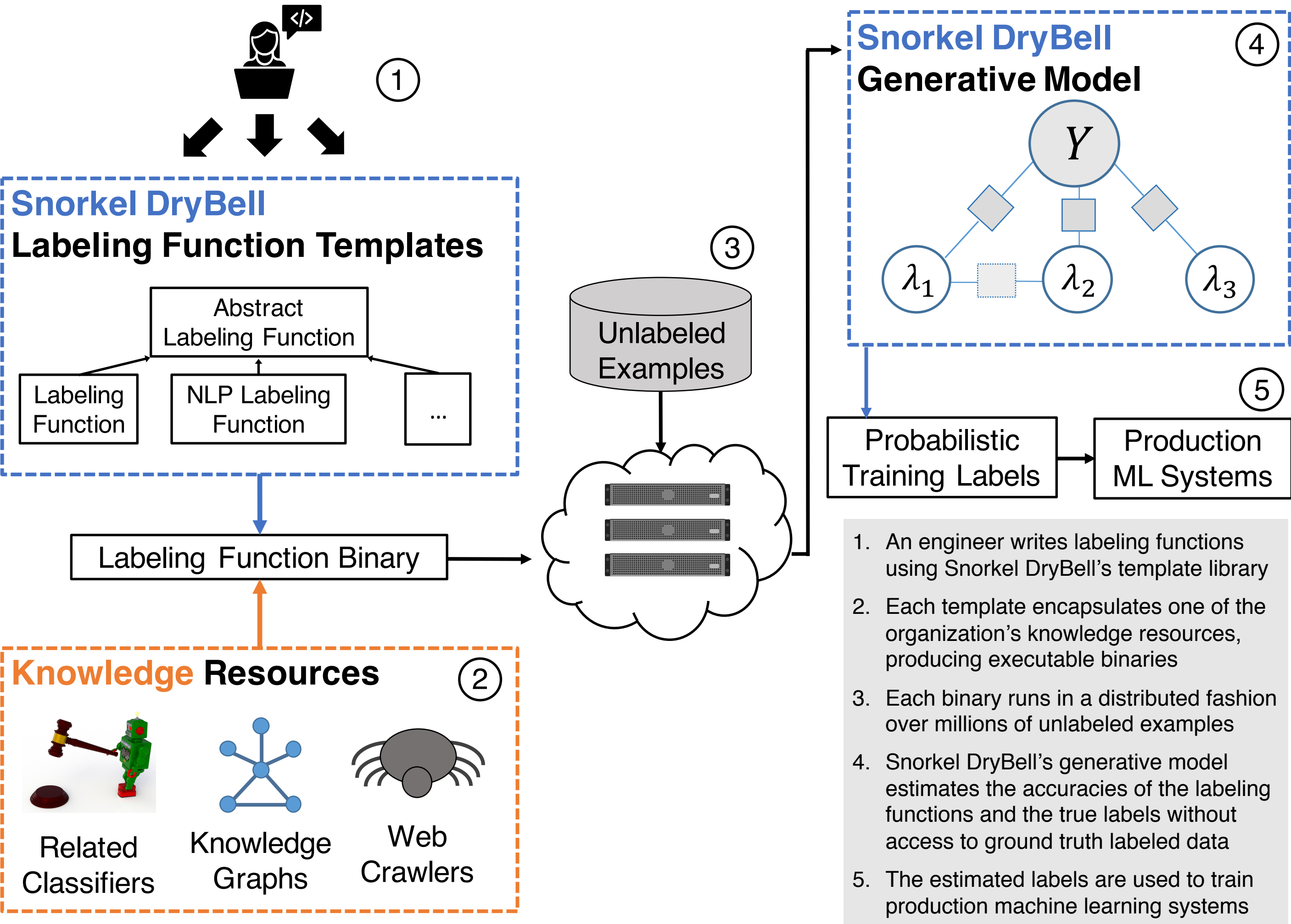


Topic Classification

- Emerging topic of interest in Web content
- Goal: develop new classifier to identify topic
- Default procedure is to collect hundreds of thousands of labels for new topic!



Architecture



Example Code

In this example, we write a labeling function for classifying whether Web content is related to celebrity news, using a natural language processing (NLP) classifier as a knowledge resource.

```
string GetText(const Example& x) {
    return StrCat(x.title, " ", x.body);
}

LFVVote GetValue(const Example& x,
                 const NLPResult& nlp) {
    if (nlp.entities.people.size() == 0) {
        return NEGATIVE;
    }
    else { return ABSTAIN; }
}

int main(int argc, char *argv[]) {
    Init(argc, argv);
    NLPLabelingFunction<&GetText, &GetValue> lf;
    lf.Run();
}
```

- Defines text to analyze
- "If the text doesn't mention any people, vote negative"
- Launches MapReduce pipeline, starts NLP classifier server on each worker, and saves the results

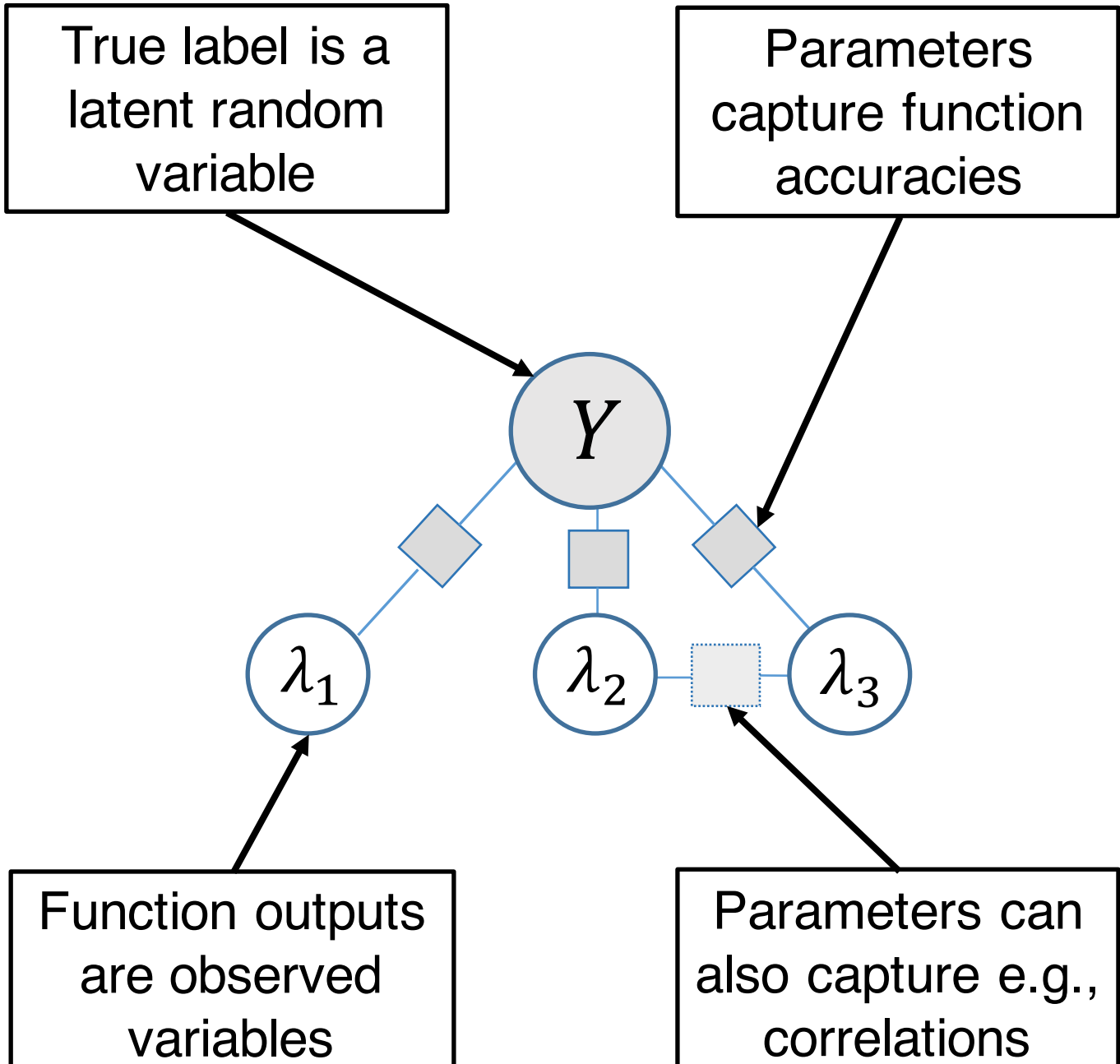
Generative Modeling

Snorkel DryBell combines the votes of labeling functions using the modeling framework of Snorkel (Ratner et al., VLDB 2017).

We estimate the statistical properties (e.g., accuracy) of the labeling functions by maximizing the log likelihood of their observed outputs:

$$\arg \max_{\theta} \log p_{\theta}(\lambda)$$
$$= \arg \max_{\theta} \log \sum_Y p_{\theta}(Y, \lambda)$$

More info.: <http://snorkel.stanford.edu>



Conclusions

- Useful resources for weak supervision are abundant.** Further, many open-source analogs exist, suggesting directions for future work.
- Rethink systems for large-scale weak supervision.** Snorkel was originally designed for novice users, but we find that expert users need more flexibility and modularity.
- Cross-feature transfer is essential.** Many resources that can be used for background knowledge are not servable. Using them as supervision enables transfer to classifiers.

