Snorkel: A System for Lightweight Extraction

Alexander Ratner Stephen H. Bach Henry Ehrenberg
Jason Fries Sen Wu Christopher Ré
Stanford University, InfoLab
{ajratner, shbach, henryre, jfries, senwu, chrismre}@stanford.edu

We describe a vision and an initial prototype system for extracting structured data from unstructured or *dark* input sources–such as text, embedded tables, and images–called SNORKEL[1], in which users write traditional extraction scripts which are automatically enhanced by machine learning techniques. The key technical idea is to view the user's actions with standard tools as implicitly defining a statistical model. For example, to extract mentions of supplier-purchaser relations in SEC filings, a user of SNORKEL might write several scripts that reference lists of company names, known supplier-purchaser relations, or specific textual patterns. SNORKEL is able to automatically assess each script's reliability for the task, combine their outputs together in a statistically sound way, and use the combined signals to train a machine learning model with automatically generated features to perform the task more accurately and broadly. Compared to current machine learning approaches, SNORKEL is our attempt to make an end-run around two major pain points: *hand-labeling training data* and *feature engineering*. More broadly, SNORKEL is a first step toward our vision of a new generation of data systems that are *observational*: systems that observe users working with standard tools, and use machine learning techniques "behind the scenes" to improve performance. In preliminary hacakathons, non-expert users from the biomedical domain have quickly neared or exceeded competition benchmarks, and SNORKEL is now in use by a handful of technology companies, government organizations, and scientists.

Lightweight Macroscopic Analysis SNORKEL is intended for tasks in which users' time and technical skills are limited, and the output schema is unknown or rapidly changing. Typically, dark data systems are deployed only in large corporations and government agencies due to their expense and high technical barrier to entry. Moreover, they are only deployed in situations in which a fixed, high-value schema is known in advance. In many scenarios, however, users may only have on the order of a week to write high-quality extractors with new and evolving schemas—for example, researchers looking for new types of anomalous drug interactions in electronic health records, or financial analysts examining newly released earning reports. SNORKEL empowers users to quickly write programs that are radically more robust and produce radically higher quality data than even finely tuned regular expressions or scripts.

Snorkel's Model User interaction with SNORKEL is centered around writing *labeling functions*, pieces of code that heuristically label data. Their output is noisy, and SNORKEL automatically denoises and combines them using statistical techniques. The resulting labeled data set is used to train a final model with automatically generated features, e.g., LSTM-based embeddings for text. *There is no traditional training label set, and the user does not engineer features.* The tooling in SNORKEL is focused on iteratively improving and adding new labeling functions. In a NIPS 2016 paper, we describe the theory behind this new approach, and show that it provides increased quality with far less user input.

- **Labeling Functions** In SNORKEL, a user's sole programming task is to create a large, noisy training label set by writing a set of labeling functions. Each labeling function takes in a *candidate* extraction and returns a label. This enables users to easily and flexibly express domain heuristics using standard scripting languages. For example, a labeling function could be a Python function which utilizes regular expressions, external knowledge bases or ontologies, or any other expressible heuristic.
- **Data Programming** We treat the user's labeling functions as implicitly describing a generative model of the true labels. Essentially, by examining the overlapping and conflicting labels they emit, we can estimate their relative accuracies and denoise the large training label set they create. In automatically adjusting for the errors the labeling functions make, we can free the developer from debugging previous ones and allow them to focus on adding new heuristics. We then use this denoised training set to train the user's machine learning model of choice.
- **Automatic Features** There is massive interest in methods like deep learning that automatically create features. However, they require large labeled training sets to work well. SNORKEL makes it easy to create large, labeled training sets quickly, allowing us to bypass feature engineering by leveraging these automatic methods.

Future Directions We see two immediate research steps toward our overarching vision of observational ML systems:

- **Weaker Supervision** We see the ability to use higher-level, less precise labeling functions as critical to enabling users with less programming expertise.
- **Images, Video, and Sensor Data** We plan to extend our techniques to other forms of data beyond text, including images, video, and sensor data.

[1] snorkel.stanford.edu

8th Biennial Conference on Innovative Data Systems Research (CIDR '17) January 8-11, 2017 , Chaminade, California, USA.