Snorkel: A System for Lightweight Extraction

Alexander Ratner Jason Fries Sen Wu
Henry Ehrenberg Stephen H. Bach Christopher Ré
Stanford University, InfoLab
{ajratner, jfries, senwu, henryre, shbach, 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, images, and diagrams–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 cross-reference against 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 using only unlabeled data, integrate 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 to this task, 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*: these systems will observe users using standard tools, and use machine learning techniques "behind the scenes" to improve their performance. In several 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.

Light-weight 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 methods 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. Examples that do not fit the old paradigm include researchers looking for previously unnoticed drug interactions in electronic health records, government agencies and NGOs responding to disaster, and financial analysts poring over a trove of newly released earning reports. In these scenarios, users have on the order of a week to write high-quality dark data extraction programs. Snorkel empowers them to write programs that are radically more robust and produce radically higher quality data than even finely tuned regular expressions or Python scripts.

Snorkel's Model The Snorkel user interface is centered around writing *labeling functions*, pieces of code that heuristically label data according to the users' desired output. Their output is obviously noisy, and Snorkel automatically denoises them using statistical techniques. The resulting labeled data set is still imperfect, but is used to train a final model with automatically generated features, e.g., LSTM-based embeddings for text. *There is no traditional training data, and the user does not engineer features.* The tooling in Snorkel is focused on iteratively improving and adding new labeling functions. In an upcoming NIPS 2016 paper, we describe the theory behind learning without traditional labeled data, and show that it provides substantial bumps in quality with far less user input than previous techniques.

- **Labeling Functions** In Snorkel, a user's sole programming task is to create a large, noisy training set by writing a set of labeling functions. Each labeling function takes in a *candidate* extraction–for example in a disease tagging task, a contiguous sequence of words–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 set they create. In automatically correcting for the errors the labeling functions make, we can free the developer from debugging previous ones and enable them to instead focus on adding new information. We then utilize this denoised training set to train one of many popular machine learning models.
- **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 weaker supervision–e.g. higher-level, less precise labeling functions–as critical to enabling users with less programming expertise and in lower-information resource domains.
- **Images, Video, and Sensor Data** We plan to extend our techniques to other forms of data beyond text, including images, video, and sensor data. The massive growth of data in these areas will necessitate new techniques, which we believe will be expressible in the general framework outlined here.

[1] https://github.com/HazyResearch/snorkel