Research Statement

Stephen Bach

My research investigates **machine learning methods for incorporating high-level knowledge into statistical models.** Today's practitioners hand-label massive data sets and carefully engineer features to train statistical models for new problems. Exploiting structured knowledge such as rules, databases, and information networks can drastically reduce these costs, accelerate development, and improve predictive performance; but doing so effectively is an ongoing challenge. My approach is to develop new methods that **blend statistical methods with rule-based systems** to achieve benefits from each. I investigate the **foundations of statistical machine learning** to discover how to make these methods practical to apply at scale. I also **build open-source systems** to validate my ideas and learn what new problems arise in downstream applications. Highlights of my work include:

- I advanced progress on the biggest bottleneck in machine learning today: collecting enough labeled training examples for data-hungry methods like deep neural networks. People have tried *weak supervision* methods like heuristic rules as inexpensive sources of labels, but like expert systems, these rules often miss the long tail of uncommon cases. Combining multiple sources of weak supervision might improve quality, but raises fundamental statistical problems around resolving their conflicts. I introduced the **first method for learning the structure of generative models** to denoise labeling rules without access to ground truth [6].
- Creating training data at scale from rules opens up new opportunities to put machine learning in the hands of subject matter experts. Can they train models quickly just by programming their knowledge as rules? To study whether this *code-as-supervision* paradigm is practical, I led the development of Snorkel,[1] a system for writing labeling rules and automatically denoising them to synthesize training data. I found that **Snorkel dramatically accelerates the development of statistical models**, both in controlled settings and by enabling applications that were previously impossible because of a lack of training data [10, 11].
- Another opportunity for exploiting knowledge is using first-order rules to compactly specify statistical models for structured data. For example, one can express in a single rule that all users in a social network are more likely to share opinions in common with their friends. This power to easily specify rich models has a major pitfall: users commonly specify intractable models. I investigated approximating statistical reasoning over uncertain rules with convex optimization. I discovered that **seemingly distinct methods from theoretical computer science, machine learning, and artificial intelligence are actually equivalent** [4, 5].
- Proving the equivalence of these methods enabled me to combine their benefits into a **new type of probabilistic graphical model** well-suited to reasoning with **uncertain rules** [2]. These *hinge-loss Markov random fields* admit convex inference, making them highly scalable, while still capturing rich structure. I then developed probabilistic soft logic (PSL),[2] **a probabilistic programming language** for building scalable, rule-based models over **relational data** like knowledge bases and social networks [5].

My work is published in top venues in machine learning and data science (ICML, NIPS, JMLR, Machine Learning, UAI, AISTATS, ILP, VLDB, KDD, ICDM). It is used by companies including Accenture, Adobe, Ali Baba, Chegg, Google, Microsoft, NEC, and Toshiba; to fight human trafficking as part of DARPA's MEMEX program; and in collaboration with the U.S. Department of Veterans Affairs and the U.S. Food and Drug Administration.

Synthesizing Training Data with Weak Supervision

My research shows that **expressing domain knowledge as labeling rules** has the power to **dramatically reduce the cost** of developing statistical models. Supervised machine learning traditionally uses hand-labeled data to fit a model. **Deep learning** in particular requires tens of thousands to *millions* of labeled examples to work well. Labeling this much data is **the biggest bottleneck in machine learning today**. I have worked to address this bottleneck through statistically modeling the behavior of noisy but inexpensive-to-develop *labeling rules* provided by *subject matter experts* (SMEs) such as biologists, social scientists, and intelligence analysts. By modeling the rules' behavior, it is possible to **resolve their conflicts without access to ground truth** and train state-of-the-art models like deep neural networks that **approach the performance of models trained on hand-labeled data**.

[1] `http://snorkel.stanford.edu`
[2] `http://psl.linqs.org`

Learning Generative Models to Synthesize Training Data

How can we estimate the accuracy of labeling rules without access to ground truth? A promising approach is to estimate a *probabilistic generative model*, in which the true label for a data point is a latent variable that generates the observed outputs of SME-expressed labeling rules. The basic setup is similar to models used in crowdsourcing and data integration. However, labeling rules can be statistically interdependent in complex ways. For example, two rules that express similar regular expressions will be correlated regardless of the true label. Not accounting for these correlations leads to overweighting the influence of these rules in estimating the true label. To address this problem, I introduced the **first method for learning the structure of probabilistic generative models without access to ground truth** [6]. The estimator maximizes the regularized pseudolikelihood of each of the labeling rules, which admits an efficient computation of the exact gradient. I showed that the method guarantees structure recovery with a sample complexity sublinear in the number of possible dependencies. It is also efficient in practice; it learns structures over dozens of labeling rules with tens of thousands of data points in a few seconds.

New Development Frameworks for Machine Learning

Denoising labeling rules to create training data has the potential to greatly accelerate the development of machine learning applications. To study this hypothesis, I led the development of **Snorkel, the first system for managing multiple weak supervision sources to rapidly create training data** [11]. Snorkel is a development framework that enables SMEs to express labeling rules as short Python functions and execute them over unlabeled data. The Snorkel engine then learns a probabilistic generative model of the labeling rules, estimating their accuracies and statistical dependencies. Using this model, Snorkel then infers a probabilistic training label for each of the unlabeled points. Any of a number of state-of-the-art classifiers, such as deep neural networks, can then be trained with these labels. I demonstrated that on benchmark tasks in information extraction, computer vision, and sentiment analysis, models developed with Snorkel approach the predictive performance of those trained on hand-labeled data. This is exciting because it **demonstrates the feasibility of training state-of-the-art neural networks without hand-labeling any training data**, unlocking applications in low-resource and restricted domains like electronic health records and intelligence analysis.

Real-World Impact

My research is guided by the application goals of users. To evaluate how weak supervision affects the development of machine learning applications, I have collaborated with domain scientists, technology companies, and government agencies to deploy Snorkel and study the results. Through weekly office hours held at Stanford, a large user community has grown around Snorkel's open source implementation. Snorkel is **used at major companies such as Accenture, Ali Baba, Chegg, Microsoft, NEC, and Toshiba**. Snorkel is also used in scientific research to enable projects where hand-labeled data is impractical to collect. In **collaboration with the U.S. Department of Veterans Affairs**, my team developed a model to extract mentions of chemical reactions from the scientific literature, in order to build a knowledge base to predict drug interations. In **collaboration with Stanford Hospital and Clinics and the U.S. Food and Drug Administration**, we developed a model to mine unstructured text in electronic health records. On these applications, Snorkel improved prediction quality over previous heuristic approaches by an average 110%.

Feedback from these collaborations have shaped Snorkel's design. To study the benefits to users of the resulting system, **I conducted a user study in collaboration with the NIH-funded Mobilize Center at Stanford** [10]. SME participants from universities, labs, and companies around the United States applied to attend a hands-on workshop[3] on using machine learning to construct biomedical knowledge bases from unstructured data with Snorkel. We selected half of the applicants to attend the two-day workshop. We found that SMEs built models 2.8× faster and increased model quality an average 45.5% versus seven hours of hand labeling.

Statistical Relational Models for Structured Data

I have also developed new statistical methods that capture how many interdependent predictions are related with intuitive, logical rules. Whereas the models described above use rules to relate a target prediction to observations, many other problems exhibit dependencies among predictions that are best modeled jointly. Such problems are

[3] `http://mobilize.stanford.edu/events/snorkelworkshop2017/`

the focus of **statistical relational learning** and the closely related field of **structured prediction**, and they arise in modeling structured data like knowledge bases, social networks, and biological networks. For examples, inferring the truth of one fact in a knowledge base implies many others, and friends in a social network often have correlated preferences. To better capture these dependencies, I introduced a new class of probabilistic graphical models for structured data that emphasizes scaling up to massively sized problems, and a probabilistic programming language for making them easier to construct.

New Graphical Models for Large-Scale Structured Data

A major challenge when modeling structured data is scalability. Traditional graphical models use discrete random variables to model relationships, leading to a combinatorial explosion of possible configurations that makes inference and learning intractable. To enable modeling complex structures like transitive relationships in data, I studied approaches that relax combinatorial inference to convex optimization. **I proved that three seemingly distinct methods for inference in logic-based models from the theoretical computer science, machine learning, and artificial intelligence communities are actually equivalent** [4]. This equivalence enables the combination of representations, algorithms, and inference-quality guarantees developed across communities.

To capture these benefits in a single formalism, **I introduced hinge-loss Markov random fields (MRFs), a new class of probabilistic graphical models** [2]. These models also generalize previous approaches by incorporating quadratic dependencies, which increases expressivity and often improves predictive performance. As part of this work, I developed new algorithms to support inference and learning with hinge-loss MRFs. I introduced **new optimization techniques that take advantage of the sparsity common in probabilistic graphical models** to perform up to 1000$\times$ faster than alternatives like interior point methods [1]. I also introduced a method for learning these models when they contain latent variables that trains $10\times$ faster, and matches or beats the predictions of traditional methods like expectation maximization [3]. This technique enabled applications like social media analysis and latent group discovery for IARPA's Open Source Indicators program and predicting outcomes in massive open online courses (MOOCs).

Probabilistic Programming with Logical Rules

One of the most powerful aspects of modeling structured data with probabilistic graphical models is defining rules that describe how recurring motifs in the data should be represented. For example, one rule might express that two people who communicate regularly are more likely to share political preferences. In this way, complex models can be constructed over massive data sets without hand-specifying each dependency. To make hinge-loss MRFs easy to construct, **I introduced probabilistic soft logic (PSL), a probabilistic programming language for relational data** like knowledge bases and social networks. Researchers around the world have used PSL for bioinformatics, computational social science, natural language processing, information extraction, and computer vision [5, and references therein]. PSL is also **used at technology companies like Adobe and Google**. I have also studied how to incorporate intuitive aggregates like "many" and "most" into probabilistic programming [7, 8], which enables SMEs to use these concepts to define their models.

Future Research

I believe that new supervision paradigms based on high-level knowledge will drive the future of machine learning research and practice. As deep learning and other pattern recognition techniques mature to super-human performance, **progress will increasingly be blocked by our ability to efficiently teach**, not a model's ability to learn. If we are to reach an age of machine learning for the masses—where any organization can automate repetitive, knowledge-intensive tasks—then we must address this bottleneck. My aim is to lead a machine learning research group that will study fundamental open questions about expressing and learning from high-level or otherwise imprecise supervision, including the following.

Multi-Modal Weak Supervision and Discovering Supervision Vocabularies

An outstanding challenge is extending the code-as-supervision paradigm embodied in Snorkel to more fully support **domains beyond natural language**. My work on training deep convolutional neural networks with Snorkel for diagnosis from radiology images [10] shows that high-level supervision can work in such domains, but the

process is still challenging for SMEs. One major reason is that there are no ready substitutes for tools like regular expressions when expressing heuristics over images, video, time-series, etc. I plan to address this limitation by leading the development of methods for automatically extracting *vocabularies of primitives* from multi-modal data sets with little to no supervision. These primitives are pattern detectors with which users can express supervision rules. For example, consider training a classifier for activity recognition in video, such as people waiting in line versus talking [9]. Natural labeling rules for this problem refer to persons' poses, such as the direction they are facing. We will investigate how to automatically suggest primitives for such distinctions to users and enable users to refine them as they express their knowledge.

This research direction is key to **empowering SMEs to rapidly develop models for understanding multi-modal unstructured data**. It also introduces challenges that are distinct from traditional unsupervised learning, because we can interleave primitive discovery with the iterative process of writing labeling rules and training a supervised model. As users refine their classifiers, the proposed framework will have access to information about the downstream task. How best to exploit this information is an exciting topic for research.

Beyond Label Synthesis: Weakly Supervised Structured Prediction

A limitation of current code-as-supervision methods is that they train classifiers to predict independent variables, i.e., no information is shared across related decisions at prediction time. While independent predictions are adequate for many tasks, a large body of work on structured prediction has shown that modeling dependencies among predictions can significantly improve performance. While learning from imprecise labels[4] has been studied, **learning structured predictors from high-level supervision** like labeling rules remains unexplored. I will address this gap by leading research on probabilistic models for integrating heuristic rules and other knowledge like ontologies and grammars to synthesize entire target structures for training, combining my expertise in weak supervision and structured prediction.

Beyond improving predictive performance, this direction will introduce a new capability for users: the ability to **express supervision in terms of predictions that have not yet been made**. Many tasks, from named entity recognition in text to activity recognition in video, have natural dependencies across predictions. The semantics of a text token depend on the semantics of those before and after it. Likewise, the interpretation of a frame of video depends on one's beliefs about adjacent frames. The training data produced by generative models should accord with these beliefs, but it requires integrating user-expressed dependencies with labeling rules in a sound statistical framework. I will lead the development of this framework, along with scalable learning and inference techniques. The end goal is to enable users to express high-level supervision that captures how multiple predictions are related as easily as how they depend on observations.

Intelligent Model Ecosystems and Composable AI

Current research generally treats model development as happening in a vacuum, but in reality, developers have access to models for related tasks, as well as background knowledge, supervision sources, and data that can be shared among models. How can available imperfect and imprecise resources be automatically mapped to tasks, in order to accelerate model development and automate refinement? I envision **intelligent networks of models and domain knowledge** that can automatically distribute newly provided or discovered knowledge to all models that can benefit. We can go further and close the loop when some of these models are trained to extract facts from unstructured data. As they discover new knowledge, it will be used to update related models with better supervision.

What abstractions will make it easier to automatically connect models, domain knowledge, and data in these networks? My work on programmatic supervision shows that labeling rules are composable, reusable units that could form the basis for such a framework. Since they can be executed over large amounts of unlabeled data without human intervention, labeling rules are a viable mechanism for **rapidly composing new models using domain knowledge and statistical inference**. My work on learning statistical dependencies among labeling rules will serve as a foundation for automatically composing supervision sources for new tasks. If successful, this research will enable users to snap together domain knowledge for training as easily as today's developers can snap together layers of a deep network. Only then will we be able to reach the goal of machine learning for the masses.

[4] An example is learning to segment images from annotations that specify an object but not a location.

References

[1] S. H. Bach, M. Broecheler, L. Getoor, and D. P. O'Leary. Scaling MPE inference for constrained continuous Markov random fields with consensus optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[2] S. H. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

[3] S. H. Bach, B. Huang, J. Boyd-Graber, and L. Getoor. Paired-dual learning for fast training of latent variable hinge-loss MRFs. In *International Conference on Machine Learning (ICML)*, 2015.

[4] S. H. Bach, B. Huang, and L. Getoor. Unifying local consistency and MAX SAT relaxations for scalable inference with rounding guarantees. In *Artificial Intelligence and Statistics (AISTATS)*, 2015.

[5] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 18(109):1–67, 2017.

[6] S. H. Bach, B. He, A. J. Ratner, and C. Ré. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning (ICML)*, 2017.

[7] G. Farnadi, S. H. Bach, M. Blondeel, M.-F. Moens, L. Getoor, and M. De Cock. Statistical relational learning with soft quantifiers. In *International Conference on Inductive Logic Programming (ILP)*, 2015.

[8] G. Farnadi, S. H. Bach, M. Blondeel, M.-F. Moens, L. Getoor, and M. De Cock. Soft quantification in statistical relational learning. *Machine Learning*, 2017.

[9] B. London, S. Khamis, S. H. Bach, B. Huang, L. Getoor, and L. Davis. Collective activity detection using hinge-loss Markov random fields. In *CVPR Workshop on Structured Prediction: Tractability, Learning and Inference*, 2013.

[10] A. J. Ratner, S. H. Bach, H. E. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment (PVLDB)*, 11(3):269–282, 2017.

[11] A. J. Ratner, S. H. Bach, H. E. Ehrenberg, and C. Ré. Snorkel: Fast training set generation for information extraction. ACM SIGMOD Conference on Management of Data (SIGMOD) Demonstration, 2017.