Stephen H. Bach

Curriculum Vitae

sbach@cs.brown.edu
cs.brown.edu/people/sbach

Research

My research in **machine learning** focuses on improving the processes by which humans teach computers. That includes engineering training data, with methods like programmatic weak supervision, as well as learning to generalize from fewer examples, with methods like zero-shot and few-shot learning. Often, our group's methods focus on exploiting high-level, symbolic or otherwise semantically meaningful domain knowledge. Applications of our work include information extraction, image understanding, scientific discovery, and other areas of data science.

Positions

Assistant Professor, 2018–Present
Brown University
Computer Science Department

Research Scientist, 2019–Present
Snorkel AI
Menlo Park, CA

Visiting Scholar, 2018
Google
Mountain View, CA

Postdoctoral Scholar, 2015–2018
Stanford University
Computer Science Department
Advisor: Christopher Ré

Education

Ph.D., Computer Science, 2015
University of Maryland, College Park
Advisor: Lise Getoor
Dissertation: Hinge-Loss Markov Random Fields and Probabilistic Soft Logic:
A Scalable Approach to Structured Prediction
Committee: Rama Chellapa, Hal Daumé III, Larry Davis, Kevin Murphy
Larry S. Davis Doctoral Dissertation Award

B.S., Computer Science and Mathematics (double major), 2010
Georgetown University
Advisor: Mark Maloof
Magna Cum Laude

Publications

Pre-Prints

R. Smith, J. A. Fries, B. Hancock, and S. H. Bach, "Language models in the loop: Incorporating prompting into weak supervision," vol. arXiv:2205.02318 [cs.LG], 2022.

R. Esfandiarpoor, A. Pu, M. Hajabdollahi, and S. H. Bach, "Extended few-shot learning: Exploiting existing resources for novel tasks," vol. arXiv:2012.07176 [cs.LG], 2020.

Journal Papers

N. V. Nayak and S. H. Bach, "Zero-shot learning with common sense knowledge graphs," *Transactions on Machine Learning Research (TMLR)*, 2022.

A. J. Ratner, S. H. Bach, H. E. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, no. 2, pp. 709–730, 2020.

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)*, vol. 18, no. 109, pp. 1–67, 2017.

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.

Peer-Reviewed Conference Papers

N. V. Nayak*, P. Yu*, and S. H. Bach, "Learning to compose soft prompts for compositional zero-shot learning," in *International Conference on Learning Representations (ICLR)*, 2023.

V. Sanh*, A. Webson*, C. Raffel*, S. H. Bach*, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations (ICLR)*, 2022. **Selected for spotlight presentation, 5% of submitted papers (176/3391)**.

P. Yu, T. Ding, and S. H. Bach, "Learning from multiple noisy partial labelers," in *Artificial Intelligence and Statistics (AISTATS)*, 2022.

W. Piriyakulkij, C. Menghini, R. Briden, N. V. Nayak, J. Zhu, E. Raisi, and S. H. Bach, "TAGLETS: A system for automatic semi-supervised learning with auxiliary data," in *Conference on Machine Learning and Systems (MLSys)*, 2022.

J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, and H. Lakkaraju, "Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations," in *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2022.

J. A. Fries*, L. Weber*, N. Seelam*, G. Altay*, D. Datta, S. Garda, M. Kang, R. Su, W. Kusa, S. Cahyawijaya, F. Barth, S. Ott, M. Samwald, S. H. Bach, S. Biderman, M. Sänger, B. Wang, A. Callahan, D. L. Periñán, T. Gigant, P. Haller, J. Chim, J. D. Posada, J. M. Giorgi, K. R. Sivaraman, M. Pàmies, M. Nezhurina, R. Martin, M. Cullan, M. Freidank, N. Dahlberg, S. Mishra, S. Bose, N. M. Broad, Y. Labrak, S. S. Deshmukh, S. Kiblawi, A. Singh, M. C. Vu, T. Neeraj, J. Golde, A. V. del Moral, and B. Beilharz, "BigBIO: A framework for data-centric biomedical natural language processing," in *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

A. Mazzetto*, C. Menghini*, A. Yuan, E. Upfal, and S. H. Bach, "Tight lower bounds on worst-case guarantees for zero-shot learning with attributes," in *Neural Information Processing Systems (NeurIPS)*, 2022.

A. Mazzetto*, C. Cousins*, D. Sam, S. H. Bach, and E. Upfal, "Adversarial multiclass learning under weak supervision with performance guarantees," in *International Conference on Machine Learning (ICML)*, 2021.

A. Mazzetto, D. Sam, A. Park, E. Upfal, and S. H. Bach, "Semi-supervised aggregation of dependent weak supervision sources with performance guarantees," in *Artificial Intelligence and Statistics (AISTATS)*, 2021.

E. Safranchik, S. Luo, and S. H. Bach, "Weakly supervised sequence tagging from noisy rules," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, R. Kuchhal, C. Ré, and R. Malkin, "Snorkel DryBell: A case study in deploying weak supervision at industrial scale," in *International Conference on Management of Data (SIGMOD) Industry Track*, 2019.

A. J. Ratner, S. H. Bach, H. E. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *PVLDB*, vol. 11, no. 3, pp. 269–282, 2017. **Best of VLDB 2018**.

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.

H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

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.

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. **Selected for oral presentation, 6% of submitted papers (27/442)**.

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. **Best Student Paper Award**.

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.

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.

S. H. Bach and M. A. Maloof, "A Bayesian approach to concept drift," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

S. H. Bach* and M. A. Maloof*, "Paired learners for concept drift," in *IEEE International Conference on Data Mining (ICDM)*, 2008.

Demonstrations

S. H. Bach*, V. Sanh*, Z.-X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Fevry, Z. Alyafeai, M. Dey, A. Santilli, Z. Sun, S. Ben-David, C. Xu, G. Chhablani, H. Wang, J. A. Fries, M. S. Al-shaibani, S. Sharma, U. Thakker, K. Almubarak, X. Tang, D. Radev, M. T.-J. Jiang, and A. M. Rush, "PromptSource: An integrated development environment and repository for natural language prompts," in *Meeting of the Association for Computational Linguistics (ACL)*, 2022.

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

Workshop Papers

J. Dai, S. Upadhyay, S. H. Bach, and H. Lakkaraju, "What will it take to generate fairness-preserving explanations?," in *ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, 2021.

E. Raisi and S. H. Bach, "Selecting auxiliary data using knowledge graphs for image classification with limited labels," in *CVPR Workshop on Visual Learning with Limited Labels*, 2020.

R. Patel, S. H. Bach, and E. Pavlick, "Learning visually grounded representations with sketches," in *ICML Workshop on New Tasks for Vision and Language*, 2019.

S. H. Bach, B. Huang, and L. Getoor, "Rounding guarantees for message-passing MAP inference with logical dependencies," in *NIPS Workshop on Discrete and Combinatorial Problems in Machine Learning (DISCML)*, 2014.

S. H. Bach, B. Huang, and L. Getoor, "Probabilistic soft logic for social good," in *KDD Workshop on Data Science for Social Good*, 2014.

G. Farnadi, S. H. Bach, M. Moens, L. Getoor, and M. De Cock, "Extending PSL with fuzzy quantifiers," in *International Workshop on Statistical Relational Artificial Intelligence (StaRAI)*, 2014.

S. H. Bach, B. Huang, and L. Getoor, "Large-margin structured learning for link ranking," in *NIPS Workshop on Frontiers of Network Analysis: Methods, Models, and Applications*, 2013. **Best Student Paper Award**.

S. H. Bach, B. Huang, and L. Getoor, "Learning latent groups with hinge-loss Markov random fields," in *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013.

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.

A. Kimmig, S. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "A short introduction to probabilistic soft logic," in *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

B. Huang, S. H. Bach, E. Norris, J. Pujara, and L. Getoor, "Social group modeling with probabilistic soft logic," in *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models, and Applications*, 2012.

A. Memory, A. Kimmig, S. H. Bach, L. Raschid, and L. Getoor, "Graph summarization in annotated data using probabilistic soft logic," in *Proceedings of the International Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*, 2012.

S. H. Bach, M. Broecheler, S. Kok, and L. Getoor, "Decision-driven models with probabilistic soft logic," in *NIPS Workshop on Predictive Models in Personalized Medicine*, 2010.

* Equal Contributors

Invited Talks

Data-Centric Approaches to Building on Foundation Models

Government Services Administration Community of Practice: Artificial Intelligence	Nov. 16 2022
Dept. of Homeland Security Joint IF and Software Cost Forum	Sep. 13 2022
Sharif University of Technology, International Campus, Kish Island	Dec. 20 2021
Google AdsML Summit	Nov. 10 2021

Using Knowledge Graphs to Learn with Fewer Labels

AAAI Workshop on Graphs for Learning and Reasoning (GCLR)	Feb. 9 2021
Google Tech Talk Series	Sep. 29 2020

Weakly Supervised Machine Learning at Industrial Scale

ACM/IEEE Computer Society Boston Chapter Joint Seminar Series	Jun. 20 2019
University of Massachusetts, Lowell, Computer Science Department	Oct. 16 2019

Programming Statistical Machine Learning with High-Level Knowledge

University of Chicago, Computer Science Department	Apr. 11 2018
Harvard University, Computer Science Department	Apr. 5 2018
Northeastern University, College of Computer and Info. Science	Apr. 3 2018
University of Pennsylvania, Computer Science Department	Mar. 28 2018
Rutgers University, Computer Science Department	Mar. 26 2018
Brown University, Computer Science Department	Mar. 19 2018
University of Massachusetts Amherst, College of Info. and Computer Sciences	Feb. 26 2018
Duke University, Computer Science Department	Feb. 21 2018
Dartmouth College, Computer Science Department	Feb. 16 2018
University of California, Irvine, Computer Science Department	Feb. 13 2018
Purdue University, Computer Science Department	Feb. 8 2018

Snorkel: Creating Noisy Training Data to Overcome Machine Learning's Biggest Bottleneck

AAAI Symposium: Towards AI for Collaborative Open Science	Mar. 26 2019
University of California, Berkeley, Berkeley AI Research Lab	Jul. 10 2017
SIGMOD Workshop on Data Management for End-to-End Machine Learning (DEEM)	May 14 2017

Probabilistic Soft Logic: Scaling Up Logical Reasoning in Statistical Machine Learning

CSLI Workshop on Logic, Rationality, and Intelligent Interaction	Jun. 2 2018
University of California, Santa Cruz, Computer Science Department	Mar. 4 2016
Stanford University, InfoLab, Computer Science Department	Feb. 11 2015

University of California, San Diego, San Diego Supercomputer Center Feb. 9 2015
Charles River Analytics, Cambridge, MA Aug. 20 2014

Invited Panels

Learning in the Presence of Label Scarcity. NorthEast Computational Health Summit Apr. 26 2019

Honors and Awards

Larry S. Davis Doctoral Dissertation Award
2015, Computer Science Department, University of Maryland, College Park

Teaching

Machine Learning
CSCI 1420, Brown University
Spring 2019–2023

Learning with Limited Labeled Data
CSCI 2952-C, Brown University
Fall 2018, 2020, 2022

Professional Activities

Workshop Organizer
2019 ICLR Workshop on Learning with Limited Labeled Data: Weak Supervision and Beyond
2017 NIPS Workshop on Learning with Limited Labeled Data: Weak Supervision and Beyond

National Science Foundation Peer Review Panelist
Division of Information and Intelligent Systems (IIS)

Conference Area Chair / Senior Program Committee / Meta-Reviewer
Artificial Intelligence and Statistics (AISTATS)
International Conference on Learning Representations (ICLR)
International Conference on Machine Learning (ICML)
International Joint Conference on Artificial Intelligence (IJCAI)
Neural Information Processing Systems (NeurIPS)

Conference Program Committee / Reviewer
Automatic Knowledge Base Construction (AKBC)
International Conference on Learning Representations (ICLR)
International Conference on Machine Learning (ICML)
International Joint Conference on Artificial Intelligence (IJCAI)
International World Wide Web Conference (WWW)
Neural Information Processing Systems (NeruIPS)
SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
Uncertainty in Artificial Intelligence (UAI)

Journal Reviewer
ACM Transactions on Knowledge Discovery in Data (TKDD)
Data Mining and Knowledge Discovery
IEEE Transactions on Knowledge and Data Engineering (TKDE)
IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
Journal of Machine Learning Research (JMLR)
Science Advances
Statistical Analysis and Data Mining

Workshop Program Committee
Artificial Intelligence for Humanitarian Assistance and Disaster Response
Data Driven Discovery of Models (D3M)
Declarative Learning Based Programming (DeLBP)

Knowledge Base Construction, Reasoning and Mining (KBCOM)
Mining and Learning with Graphs (MLG)

Selected Department and University Service

Committee Member

2018–2023 Computer Science Ph.D. Admissions
2018–2022 University Goldwater Scholarship Committee
2022– Co-Chair, Data Science Initiative Campus Advisory Board

Grants

Tasks Algorithmically Given Labels Established via Transferred Symbols
DARPA, $2,800,000
Co-PI, Approximate Usage: 50%
7/2019–6/2023

Weakly Supervised Security Event Detection
Cisco, Inc., $300,000
PI, Approximate Usage: 100%
8/2020–7/2023

EAGER: SaTC-EDU: Adversarial Thinking Early and Often
NSF, $297,881
Co-PI, Approximate Usage: 5%
9/2020–8/2023

Social Media, Violence, and Social Isolation Among At-Risk Adolescents: Exploring Ground Truth
NIH, $3,991,091
Co-PI, Approximate Usage: 10%
9/2020–8/2023

Collaborative Research: 21 cm Reionization Science with the MWA
NSF, $534,990
Co-PI, Approximate Usage: 10%
9/2021–8/2024

Collaborative Research: SWIFT-SAT: RFI Detection Across Six Orders of Magnitude in Intensity: A Unifying Framework with Weakly Supervised Machine Learning
NSF, $470,030
PI, Approximate Usage: 90%
9/2022–8/2025

Advising

Current Ph.D. Students

Reza Esfandiarpoor
Nihal Nayak
Jasper Solt (Physics Ph.D. co-advised with Jonathan Pober)
Zheng-Xin Yong
Peilin Yu

Current Post-Doctoral Researchers

Cristina Menghini

January 31, 2023