SDSI 2016 Retreat Speakers Bios and Abstracts

Peter Bailis

Assistant Professor of Computer Science

Peter Bailis is an Assistant Professor of Computer Science at Stanford University. Peter's research in the Future Data Systems group focuses on the design and implementation of next-generation, post-database data- intensive systems. His work spans large-scale data management, distributed protocol design, and architectures for high-volume complex decision support. He is the recipient of an NSF Graduate Research Fellowship, a Berkeley Fellowship for Graduate Study, best-of-conference citations for research appearing in both SIGMOD and VLDB, and the CRA Outstanding Undergraduate Researcher Award. He received a PhD from UC Berkeley in 2015 and an A.B. from Harvard College in 2011, both in Computer Science.

Alison Callahan

Research Scientist in Biomedical Informatics

Alison Callahan is a research scientist in Nigam Shah's lab at the Stanford Center for Biomedical Informatics Research. Her research combines data mining of electronic medical records, text processing and ontology-driven data annotation to study human disease and health care. As a doctoral student, she designed and implemented a computational framework for biomedical data integration and hypothesis evaluation, using semantic web services to query and analyze large volumes of biomedical data. Her postdoctoral research extended methods for hypothesis evaluation and text processing to the domain of spinal cord injury research, enabling community- driven annotation of published literature and semi-automated evaluation of hypotheses about drug therapies that affect recovery after SCI. Her current research focuses on analyzing electronic medical records and biomedical literature using ontologies and machine learning for drug and device safety surveillance.

Title: Extracting patient-reported pain from clinical notes using data programming

Abstract: Clinical notes offer a wealth of information about patient experience and clinician practice that are not captured in the structured billing codes, lab reports and medication orders of electronic health records. Extracting information from unstructured clinical text presents a unique challenge, given the natural variation in clinical language and reporting styles, the design of EHR software, and the cost of creating labeled datasets for supervised machine learning methods. I will present our data programming-based system for scalably extracting mentions of patient-reported pain from the clinical notes of joint replacement patients. More than 1 million joint replacements are carried out annually to treat osteoarthritis (the most common cause of disability in U. S. adults), and a significant proportion of those who have the procedure do not experience an improvement in pain. Our aim is to measure patient-reported pain pre- and post-operatively, to better understand patient outcomes following joint replacement surgeries.

Justin Cheng

PhD Candidate in Computer Science

Justin Cheng is a PhD candidate in the Computer Science Department at Stanford University, where he is advised by Jure Leskovec and Michael Bernstein. His research lies at the intersection of data mining and crowdsourcing, and focuses on cascading behavior in social networks. This work has received several best paper nominations at CHI, CSCW, and ICWSM. He is also a recipient of a Microsoft Research PhD Fellowship and a Stanford Graduate Fellowship.

Title: Can anyone become a troll?

Abstract: Social media systems rely on user feedback and rating mechanisms for personalization, ranking, and content filtering. However, when users evaluate content contributed by fellow users (e.g., by liking a post or voting on a comment), these evaluations create complex social feedback effects. First, we focus on trolling in a community - identifying its causes and quantifying its effects on the community at large through a combination of experimentation and large-scale longitudinal analysis. Next, we more generally investigate the influence of feedback on future user behavior to understand how positive or negative feedback may percolate through a community.

Kevin Clark

PhD Candidate in Computer Science

Kevin Clark is a PhD candidate in the Computer Science Department at Stanford University. He works in the Natural Language Processing Group and is advised by Chris Manning. His main research interests are in deep learning and natural language understanding.

Chris De Sa

PhD Candidate in Computer Science

Christopher De Sa is a PhD candidate at Stanford University advised by Chris Ré and Kunle Olukotun. He primarily studies fast stochastic algorithms, such as Gibbs sampling and stochastic gradient descent. He takes particular interest in heuristics that improve performance on modern heterogeneous hardware and can be guaranteed to not affect statistical efficiency.

Title: Fast Stochastic Algorithms for Data Analytics on Parallel Hardware.

Abstract: Fast stochastic algorithms, such as stochastic gradient descent (SGD) and Gibbs sampling, are widely used in data analysis and machine learning. In this talk, I will describe how to improve the performance of these ubiquitous algorithms by taking advantage of the parallel compute resources available on modern hardware.

Stefano Ermon

Assistant Professor of Computer Science

Stefano Ermon is currently an Assistant Professor in the Department of Computer Science at Stanford University, where he is affiliated with the Artificial Intelligence Laboratory and the Woods Institute for the Environment. He completed his PhD in computer science at Cornell in 2015. His research interests include techniques for scalable and accurate inference in graphical models, statistical modeling of data, large-scale combinatorial optimization, and robust decision making under uncertainty, and is motivated by a range of applications, in particular ones in the emerging field of computational sustainability. Stefano has won several awards, including two Best Student Paper Awards, one Runner-Up Prize, and a McMullen Fellowship.

Title: Random Projections for Probabilistic Modeling and Inference

Abstract: Reasoning about high-dimensional probabilistic models (i.e., with many variables) is a key computational challenge in AI and Machine Learning. In this talk, I will introduce new approaches to learn and make inferences in high-dimensional models using random projections. Intuitively, random projections are used to simplify a high-dimensional model while preserving some of its key properties. These novel randomized approaches provide provable guarantees on the accuracy, and outperform traditional methods in a range of domains. I will discuss applications in learning deep generative models and for the analysis of large scale spatio-temporal data.

Jason Fries

Postdoctoral Researcher in Computer Science

Jason Fries is a postdoctoral researcher in Computer Science at Stanford University. He works with Prof. Chris Ré and Scott Delp as part of Stanford's Mobilize Center, an NIH Big Data to Knowledge (BD2K) site of excellence that explores data science approaches to understanding diseases of human mobility. His research focuses on information extraction and predictive modeling using unstructured text and time series data from scientific literature and the electronic medical record. His most recent projects include developing weak supervision regimes for extracting named entities from text without using labeled data and modeling postoperative trajectories of pain and function after joint replacement surgery. Jason received his PhD from the University of Iowa in 2015, co-advised by Alberto Segre and Dr. Phil Polgreen, working as part of Iowa's Computational Epidemiology Research Group. His thesis explored large-scale information extraction in electronic medical record text as well as machine learning approaches to sexual health surveillance using social media.

Title: Weak supervision regimes: leveraging structured resources in data programming

Abstract: Building weakly-supervised extraction systems is heavily influenced by the availability of structured resources, which vary widely by domain and application setting. We define three weak supervision regimes -- low, medium, and high -- based on the availability of resources like labeled data and formal ontologies. Under this framing, we demonstrate how data programming unifies multiple forms of weak supervision and easily integrates with rich, curated resources like the Unified Medical Language System (UMLS) to programmatically generate and denoise training data. Using this approach in three biomedical entity extraction tasks, we can construct models that approach or match supervised learning benchmarks, but are fundamentally trained on unlabeled data. In these tasks, we find that human-provided labels are less valuable than ontologies and other noisy, structured resources. This suggests data programming is a viable way of building extraction systems for the long tail of biomedical entity types with little-to-no labeled training data.

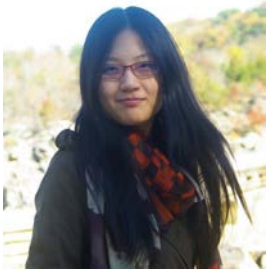
Jeff Hancock

Professor of Communication

Jeff Hancock is a Professor in the Department of Communication at Stanford University. Professor Hancock and his group work on understanding psychological and interpersonal processes in social media. The team specializes in using computational linguistics and experiments to understand how the words we use can reveal psychological and social dynamics, such as deception and trust, emotional dynamics, intimacy and relationships, and social support. Recently Professor Hancock has begun work on understanding the mental models people have about algorithms in social media, as well as working on the ethical issues associated with computational social science. Professor Hancock was a Customs Officer in Canada before earning his PhD in Psychology at Dalhousie University, Canada. He was a Professor of Information Science and Communication at Cornell prior to joining Stanford.

He He

Postdoctoral Researcher in Computer Science



He He is a postdoctoral researcher at Stanford University, working with Percy Liang. Prior to Stanford, she earned her PhD in Computer Science at the University of Maryland, College Park, advised by Hal Daumé III and Jordan Boyd-Graber (currently at CU Boulder). Her interests are at the interface between machine learning and natural language processing. She develops algorithms that acquire information dynamically and do inference incremen- tally, with an emphasis on problems in natural language processing.

David Jurgens

Postdoctoral Researcher in Computer Science

David Jurgens is a postdoctoral researcher, jointly in the the Stanford NLP and SNAP Groups under Dan Jurafsky, Jure Leskovec and Dan McFarland. His research interests span Natural Language Processing, Network Science and Computational Social Science. Before joining the NLP Group, he was a postdoctoral scholar at McGill University in the Network Dynamics group with Derek Ruths. Prior, he was a research scientist at the Linguistics Computing Laboratory at Sapienza University under Roberto Navigli and before that, was a visiting researcher at the Information and Systems Science Lab at HRL Laboratories. He received his PhD in Computer Science from the University of California, Los Angeles under Michael Dyer. He received his BA in Philosophy and Political Science and an MS in Computer Science on Computer Vision under Robert Pless from Washington University in St. Louis.

Dan Jurafsky

Professor and Chair of Linguistics; Professor of Computer Science

Dan Jurafsky is Professor and Chair of Linguistics and Professor of Computer Science at Stanford University. He is the recipient of a 2002 MacArthur Fellowship, is the co-author with Jim Martin of the widely-used textbook "Speech and Language Processing", and co-created with Chris Manning one of the first massively open online courses, Stanford's course in Natural Language Processing. His new trade book "The Language of Food: A Lin- guist Reads the Menu" came out on September 15, 2014, and was a finalist for the 2015 James Beard Award. Dan received a BA in Linguistics in 1983 and a PhD in Computer Science in 1992 from the University of California at Berkeley, was a postdoc 1992-1995 at the International Computer Science Institute, and was on the faculty of the University of Colorado, Boulder until moving to Stanford in 2003. His research ranges widely across com- putational linguistics; special interests include natural language understanding, machine translation, spoken language and conversation and the relationship between human and machine processing.

Urvashi Khandelwal

PhD Candidate in Computer Science

Urvashi Khandelwal is a second year PhD candidate at the Stanford NLP Group where she is advised by Prof. Dan Jurafsky. Her research work includes exploring machine learning models for natural language generation, most recently using deep learning models for text summarization. She received her BS in Computer Science at the University of Illinois Urbana-Champaign.

Michal Kosinski

Assistant Professor of Organizational Behavior at Stanford Graduate School of Business

Michal Kosinski is an Assistant Professor in Organizational Behavior at Stanford University Graduate School of Business. His research focuses on humans in a digital environment and employs cutting-edge computational methods and Big Data mining. Michal's work had a significant impact on both academia and the industry. His findings featured in *The Economist*, inspired two TED talks, and prompted a discussion in the EU Parliament. Three of his papers were placed among Altmetrics' "Top 100 Papers That Most Caught the Public Imagination" and he was listed among the 50 most influential people in Big Data by DataIQ and IBM. Michal received his PhD in Psychology from the University of Cambridge (UK) in 2014. Prior to his current appointment, he was as a Post- Doctoral Scholar at Stanford's Computer Science Department and a researcher at Microsoft Research.

Jure Leskovec

Associate Professor of Computer Science

Jure Leskovec is an Associate Professor of Computer Science at Stanford University where he is a member of the InfoLab and the AI lab. He joined the department in September 2009. He is also working as Chief Scientist at Pinterest, where he focuses on machine learning problems. He is co-founder of a machine learning startup Ko- sei, which was acquired by Pinterest. In 2008/09 he was a postdoctoral researcher at Cornell University working with Jon Kleinberg and Dan Huttenlocher. He completed his PhD in the Machine Learning Department, School of Computer Science at Carnegie Mellon University under the supervision of Christos Faloutsos in 2008.

Percy Liang

Assistant Professor of Computer Science

Percy Liang is an Assistant Professor of Computer Science at Stanford University (B.S. from MIT, 2004; PhD from UC Berkeley, 2011). His research interests include modeling natural language semantics and developing machine learning methods that infer rich latent structures from limited supervision. His awards include the IJCAI Computers and Thought Award (2016), an NSF CAREER Award (2016), a Sloan Research Fellowship (2015), a Microsoft Research Faculty Fellowship (2014), and the best student paper at the International Conference on Machine Learning (2008).

Title: Provenance and Contracts in Machine Learning

Abstract: This talk poses two questions. The first question is: Why did the model make a certain prediction? I will discuss the importance of making a prediction via the correct means, which not only provides human interpretability but also more robust generalization. For example, a question answering system should not only be able to answer the question but to justify the answer with the proper provenance. The second question is: How should we reason about a model's behavior? The implicit contract in machine learning is that if the training data looks like the test data, then we will get good generalization. But this contract is often broken in practice. We discuss two alternative contracts: one based on the ability to say "don't know", which allows us to obtain 100% precision when the model is well-specified, and the other based on leveraging conditional independence structure, which allows us to perform unsupervised risk estimation.

Emily Mallory

PhD Candidate in Biomedical Informatics

Emily Mallory is a 5th year PhD candidate in the Biomedical Informatics training program and is advised by Russ Altman in the Departments of Bioengineering, Genetics, and Medicine. She is interested in both the extraction of biomedical information from text and using this text-derived information in machine learning models for predicting protein druggability and drug repurposing.

Christopher Manning

Professor of Linguistics; Professor of Computer Science

Christopher Manning is a professor of computer science and linguistics at Stanford University. His PhD is from Stanford in 1995, and he held faculty positions at Carnegie Mellon University and the University of Sydney before returning to Stanford. His research goal is computers that can intelligently process, understand, and generate human language material. Manning concentrates on machine learning approaches to computational linguistic problems, including syntactic parsing, computational semantics and pragmatics, textual inference, machine translation, and deep learning for NLP. He is an ACM Fellow, a AAAI Fellow, and an ACL Fellow, and has coau- thored leading textbooks on statistical natural language processing and information retrieval. He is a member of the Stanford NLP group (@stanfordnlp).

Christopher Potts

Professor of Linguistics

Christopher Potts is Professor of Linguistics and, by courtesy, of Computer Science, at Stanford, and Director of the Center for the Study of Language and Information (CSLI) at Stanford. In his research, he uses computational methods to explore how emotion is expressed in language and how linguistic production and interpretation are influenced by the context of utterance. He is the author of the 2005 book “The Logic of Conventional Implica- tures” as well as numerous scholarly papers in computational and theoretical linguistics.

Balaji Prabhakar

Professor of Electrical Engineering; Professor of Computer Science

Balaji Prabhakar is a faculty member in the Departments of Electrical Engineering and Computer Science at Stanford University. His research interests are in computer networks; notably, in designing algorithms for the Internet and for Data Centers. Recently, he has been interested in Societal Networks: networks vital for society’s functioning, such as transportation, electricity and recycling systems. He has been involved in developing and deploying incentive mechanisms to move commuters to off-peak times so that congestion, fuel and pollution costs are reduced. He has been a Terman Fellow at Stanford University and a Fellow of the Alfred P. Sloan Foundation. He has received the CAREER award from the U.S. National Science Foundation, the Erlang Prize, the Rollo Davidson Prize, and delivered the Lunteren Lectures. He is the recipient of the inaugural IEEE Innovation in Societal Infrastructure Award which recognizes “significant technological achievements and contributions to the establishment, development and proliferation of innovative societal infrastructure systems.”

Title: An Expert System for the Cloud Infrastructure

Abstract: Over the past decade, the users and operators of large cloud platforms and campus networks have desired a much more programmable network infrastructure so as to configure it to the needs of different applications and reduce the friction they can cause to each other. This has culminated in the SDN paradigm, initiated at Stanford, and now widely adopted. But it is hard to program what you do not understand: the volume, velocity and richness of network applications and traffic seem beyond the ability of direct human comprehension. What is needed is an expert system that can observe the data emitted by a network during the course

of its operation, continually learn the best responses to rapidly-changing load and operating conditions, and help the network adapt to them in real-time. In this talk we describe initial work in our group towards developing such an expert system.

Alex Ratner

PhD Candidate in Computer Science

Alex Ratner is a 3rd-year PhD candidate in Chris Re's lab, where he researches new machine learning paradigms for settings where limited or no hand-labeled training data is available, in particular for information extraction applications in domains like genomics, clinical diagnostics, and political science. He helps lead development of the Snorkel framework for lightweight information extraction.

Christopher Ré

Assistant Professor of Computer Science

Christopher (Chris) Ré's work goal is to enable users and developers to build applications that more deeply understand and exploit data. He spent four wonderful years on the faculty of the University of Wisconsin, Madison, before moving to Stanford in 2013. He helped discover the first join algorithm with worst-case optimal running time, which won the best paper at PODS 2012. He also helped develop a framework for feature engineering that won the best paper at SIGMOD 2014. He also helped understand the fundamental limits of asychrony for Gibbs Sampling, which won best paper at ICML 2016. In addition, work from his group has been incorporated into scientific efforts including the IceCube neutrino detector and PaleoDeepDive, and into Cloudera's Impala and products from Oracle, Pivotal, and Microsoft's Adam. He received an SIGMOD Dissertation Award in 2010, NSF CAREER Award in 2011, an Alfred P. Sloan Fellowship in 2013, a Moore Data Driven Investigator Award in 2014, the VLDB early Career Award in 2015, the MacArthur Foundation Fellowship in 2015, and an Okawa Research Grant in 2016.

Gregory Valiant

Assistant Professor of Computer Science

Gregory Valiant is an Assistant Professor in Stanford's Computer Science Department. Some of his recent projects focus on developing techniques for learning and estimation that are robust to the presence of significant fractions of biased or adversarial data, and designing algorithms that accurately infer information about complex distributions, when given surprisingly little data. More broadly, his research spans algorithms, learning, applied probability, and statistics. Prior to joining Stanford, he was a postdoc at Microsoft Research, New England, and received his PhD from Berkeley in Computer Science, and BA in Math from Harvard.

Matei Zaharia

Assistant Professor of Computer Science

Matei Zaharia is an assistant professor in Stanford's Computer Science Department. He works on computer systems and big data. He is also co-founder and Chief Technologist of Databricks, the big data company commercializing Apache Spark. Prior to joining Stanford, he was an Assistant Professor of Computer Science at MIT.