Statistical Machine Learning: Improving Biomedical Dataset Analysis with Dr. Fei Sha

Fei Sha

Dr. Fei Sha has returned to USC from an academic leave to join the USC Michelson Center for Convergent Bioscience. Dr. Sha received his Ph.D. from the University of Pennsylvania. He received a Google research award twice (in 2009 and 2016 respectively) and received a Young Investigator Award from the Army Research Office in 2012. He was selected as an Alfred P. Sloan Foundation Research Fellow in 2013.

Dr. Sha's research lies in the landscape of Statistical Machine Learning and Artificial Intelligence. He is broadly interested in theoretical and algorithmic foundations as well as their practical applications.

Dr. Sha will lead the Center of Data, Algorithms, and Systems for Health (DASH) at the Michelson Center for Convergent Bioscience. The DASH center focuses primarily on two areas of research: Theoretical and Data Sciences (TEDS) which advances fundamental research in theory, methods and (large-scale) computing for statistical machine learning, and Artificial Intelligence for Medical Science (AIMS) which applies the cutting-edge findings of statistical machine learning (including advances from TEDS), specifically to medical science.

Two factors drew Dr. Sha to the Michelson Center for Convergent Bioscience; first, he is enthusiastic to apply an "Engineering Plus Science" perspective – a term coined and popularized by Dean Yannis Yortsos from the Viterbi School of Engineering to his research. He explained it as being aligned with the vision of the Michelson Center -- to bring together scientists and engineers to conduct interdisciplinary research. Secondly, Dr. Sha is excited at the opportunity of collaborating with the best and brightest faculty who have joined (and will be joining) the Michelson Center.

> *"When I think of convergence in the context of the Michelson Center for Convergent Bioscience, I see two interwoven threads: one thread representing interdisciplinary research and the other representing the focus on innovation and advancements to improve human life." – Fei Sha*

Dr. Sha elaborated on his decision to join the Michelson Center for Convergent Bioscience: "Until now, my research has been Statistical Machine Learning, with a broad application to Artificial Intelligence. The biological and medical science research in the Michelson Center for Convergent Bioscience provides a new stage to expand my research profile. I am excited by the chance to apply algorithmic, computational and statistical models to unravel mysteries in biological organisms and to generate life-saving biomedical research outcomes. When I think of "convergence" in the context of the Michelson Center, I see two interwoven threads: one thread representing interdisciplinary research and the other representing the focus on innovation and advances to improve human life."

Dr. Sha explained that as a technologist, interdisciplinary research and collaboration almost surely leads to convergence in its best form. In addition to partnering with faculty within Dornsife College of Letters

Arts and Sciences, and Viterbi School of Engineering, he has also partnered with researchers in the Keck School of Medicine, and Lawrence Ellison Institute for Transformative Medicine. "I am a technologist—I am interested in working with anyone who believes that machine learning and artificial intelligence can and should make an impact to her or his research field, as well as society at large," Dr. Sha said.

Dr. Sha explained that he will bring the leading edge in Statistical Machine Learning and Artificial Intelligence to scientific discovery at the Michelson Center for Convergent Bioscience in order to meet the enormous demand to make sense of data from varying fields including biology, biomedical and clinical science. "Data is central to my research. Biological data is wickedly complex, highly intra- and inter- connected, extremely noisy and in many cases it is still expensive or impossible to acquire. This presents a whole new set of challenges to the current statistical methodologies, computing and algorithmic frameworks," Dr. Sha said.

Statistical machine learning creates unique opportunities in the domain of Convergent Bioscience including both theoretical and applied topics in learning for decision-making such as clinical trial, personalized treatment and precision medicine, Dr. Sha added.

Dr. Sha explained that deep learning (a branch of statistical machine learning) consists of models that are inspired by the ways that the human brain processes information. For example, deep learning models process information layer by layer (i.e. stages) — this is similar to the human brain's visual pathway for processing visual inputs from our eyes. Those models also use a large number of interconnected units, by mapping inputs to outputs nonlinearly— this is analogous to the way the human brain's neurons work. Because deep learning models typically involve millions of such units, they often require special computing and algorithmic techniques to learn from data.

So far, Dr. Sha's work has been focused on designing statistical methods and models to imbue computer systems with artificial intelligence equivalent to human-level intelligent information processing. His previous work includes improvement of computer "vision." He has developed methods for computers to robustly categorize objects in photographs, and to automatically summarize long videos while retaining information important to human browsers.

Dr. Sha explained that humans are often able to visually categorize objects with very few or zero prior exposure, "For example, suppose a child has never seen a panda bear but they have been told what pandas look like. They are black and white, look like bears, and eat bamboo. Would a computer be able to then recognize a panda in a zoo? My son recognized pandas based on that information because he had previously seen a bear, a black and white animal (dog or zebra,) and bamboo. He worked from his existing knowledge base or 'dataset' to extrapolate what a panda bear looked like instantaneously. Many computer visions systems on the other hand, would not have that capability." At this point, computer vision systems have required a lot more targeted information to reach the same conclusion as Dr. Sha's son.

Dr. Sha and his team recently developed an algorithm that improves the ability of computer vision systems to sort challenging datasets by visual category and reach the right conclusions without having to see as many examples as they have had to see previously. Dr. Sha explained that this method has outperformed existing competing approaches from well-known industry research labs. In addition to computer vision problems, Dr. Sha and his team have worked on a natural language processing program for computers to generate captions for images.

Dr. Sha is ready to apply his knowledge to convergent biomedical science, where his expertise in machine learning and artificial intelligence technologies is to bring long-lasting impact. Indeed, perhaps we will have a chance to see how an artificial intelligence scientist and engineer make scientific discoveries.