Stanford MEDICINE | Biomedical Data Science

DBDS Workshop in Biostatistics

Remote Access Only:

Contact kkanagaw@stanford.edu for Zoom dial-in details.

DATE:	November 19, 2020
TIME:	2:30-3:50pm
TITLE:	Interpretability and Human Validation of Machine Learning
SPEAKER:	**Finale Doshi-Velez** Associate Professor of Computer Science Harvard University

Abstract:

As machine learning systems become ubiquitous, there is a growing interest in interpretable machine learning -- that is, systems that can provide human-interpretable rationale for their predictions and decisions. In this talk, I'll first give examples of why interpretability is needed in some of our work in machine learning for health, discussing how human input (which would be impossible without interpretability) is crucial for getting past fundamental limits of statistical validation. Next, I'll speak about some of the work we are doing to understand interpretability more broadly: what exactly is interpretability, and how can we assess it? By formalizing these notions, we can hope to identify universals of interpretability and also rigorously compare different kinds of systems for producing algorithmic explanations.

Includes joint work with Been Kim, Andrew Ross, Mike Wu, Michael Hughes, Menaka Narayanan, Sam Gershman, Emily Chen, Jeffrey He, Isaac Lage, Roy Perlis, Tom McCoy, Gabe Hope, Leah Weiner, Erik Sudderth, Sonali Parbhoo, Marzyeh Ghassemi, Pete Szolovits, Mornin Feng, Leo Celi, Nicole Brimmer, Tristan Naumann, Rohit Joshi, Anna Rumshisky, Omer Gottesman, Emma Brunskill, Yao Liu, Sonali Parbhoo, Joe Futoma, and the Berkman Klein Center.

Suggested Readings:

- "Towards A Rigorous Science of Interpretable Machine Learning"