Inherent Trade-Offs in Algorithmic Fairness

With Jon Kleinberg
Cornell University

Recent discussion in the public sphere about classification by algorithms has involved tension between competing notions of what it means for such a classification to be fair to different groups. We consider several of the key fairness conditions that lie at the heart of these debates and discuss recent research establishing inherent trade-offs between these conditions. We also consider a variety of methods for promoting fairness and related notions for classification and selection problems that involve sets rather than just individuals. This talk is based on joint work with Sendhil Mullainathan, Manish Raghavan, and Maithra Raghu.