Hidden variables: finding latent variables in bacterial communities

Abstract:
The analyses of complex biological systems often results in output that may seem just as complex, with little useful knowledge extracted as a result of the multiple layers of information. Although distances are an important component of effective data science, we will show examples where distances taken in isolation of probability measure information give spurious results. In bioinformatics for instance standard methods for identifying taxa used fixed radii at 97% similarity regardless of sequence prevalence leading to spurious results. The standard base rate neglect fallacy (Kahneman and Tversky, 1974) still prevails even in mathematics where methods such as topological data analyses still ignore relevant changes in measure.

The use of multi-scale strategies is providing useful predictions of preterm birth and a deeper understanding of resilience of the human microbiome after antibiotic perturbations.

Susan Holmes has been working in non parametric multivariate statistics applied to Biology since 1985. She has taught at MIT, Harvard and was an Associate Professor of Biometry at Cornell before moving to Stanford in 1998. She teaches the Thinking Matters class: Breaking Codes and Finding patterns and likes working on big messy data sets, mostly from the areas of Immunology, Cancer Biology and Microbial Ecology. Her theoretical interests include applied probability, MCMC (Monte Carlo Markov chains), Graph Limit Theory, Differential Geometry and the topology of the space of Phylogenetic Trees. Her honors and awards include CASBS Fellow, Center for the Advanced study of the Behavioral Sciences (2017-2018), Breiman Lecturer, NIPS (December, 2016), Fellow, Fields Institute in Mathematical Sciences, Toronto, Canada (2015), Director's Transformative Research Award, NIH (2013), John Henry Samter University Fellow in Undergraduate Education, Stanford (2012), Fellow of the Institute of Mathematical Statistics, IMS (2005)