Discovering Genomic Bio-Signatures for Monitoring and Verification with Bayesian Nonparametric Inference

Presenter: Christopher L. Dean

Advisor: John W. Fisher III

Massachusetts Institute of Technology

cdean@csail.mit.edu

Abstract:

Radiological events impact local bacterial and fungal communities. Metagenomic data analysis offers the ability to detect radiological processing or events by a bio-signature: particular patterns in these bacterial or fungal communities. Modeling bacterial and fungal communities pose significant data analysis challenges, for example, samples may contain over 15,000 distinct species and the fewer than 100 disjoint measurement sets; additional conflating factors are abundant, as weather and chemical (rather than radiological) events also significantly impact these populations.

Probabilistic models provide a natural way to address these challenges, allowing for both dimensionality reduction and jointly modeling data from different sources to identify effects. In particular, a popular class of models called Hierarchical Dirichlet processes (HDPs) express data (species abundance counts) as arising from a weighted combination of shared bio-signature contributors, or distributions over species. This approach allows for statistical strength to be pooled across groups of data, while still allowing groups of data to have different models due to the effect of idiosyncratic mixture weights. Further, HDPs do not specify the model complexity— the number of bio-signature contributors—a priori, allowing this to be determined by the data.

We develop multi-modal HDPs (mmHDPs), an extension of HDPs that can accommodate multiple types of data. In doing so, we present a model and inference method that can learn joint topics amongst both base data (relative abundance counts for bacteria and fungi) and metadata, for example, weather data that includes temperature, precipitation, and solar radiation. We investigate the capacity for mmHDPs to extract relations between base measurements and metadata.