S&DS

Statistics & Data Science Project Pitch

September 10, 2018 | 4:15PM to 5:30PM @ Yale Institute for Network Science

Introduction:

Daniel A. Spielman

Department Chair, Statistics and Data Science; Sterling Professor of Computer Science; Professor of Mathematics

Project Pitches:

Hongyu Zhao

Department Chair and Ira V. Hiscock Professor of Biostatistics, Genetics and Statistics & Data Science, http://zhaocenter.org/
hongyu.zhao@yale.edu

- Risk(trait) predictions using genotype and other data
- Causal inference among different traits, e.g. smoking and BMI
- Molecular/mechanistic understanding of complex diseases through data integration
- Genome interpretations

Nick Turk-Browne

Professor, Psychology, Child Study Center, Cognitive Science Program, Psychiatry and Interdepartmental Neuroscience Program, https://ntblab.yale.edu/
nicholas.turk-browne@yale.edu

- Real-time fMRI: Advanced analysis under time pressure

The combination of fMRI and real-time analysis allows for powerful new types of brain imaging experiments. This potential has yet to be realized, however, as the temporal constraints of real-time analysis have prevented the adoption of the most powerful machine learning and other methods in wide use by standard “offline" cognitive neuroscience studies. The challenge is to develop optimizations, algorithms, and systems that allow for scalable, on-demand, and mission-critical computing. The benefit will be new ways of measuring and potentially enhancing the human brain.

Molly Crockett

Assistant Professor, Department of Psychology, http://www.crockettlab.org/people/
molly.crockett@yale.edu

- Moral Outrage in the Digital Age

There is a growing concern that social media platforms threaten democracy by widening political divides and spreading fake news. This set of projects explores the hypothesis that these threats – if they indeed exist – can be at least partly explained by the tendency of social media to amplify moral outrage, an ancient human emotion that evolved to punish bad behaviors in social groups. Might the specific design of social media be changing the nature of human outrage in ways not yet understood? In this research, we will analyze the social and political behavior of social media users across a variety of platforms (Twitter, Facebook, YouTube). We are developing algorithms for detecting outrage expression and will apply these to answer questions about how social reinforcement of outrage increases its prevalence, how this might explain the spread of fake news, and whether online outrage affects culture through art and entertainment. This work will involve collecting and wrangling big data, sentiment analysis with various supervised machine learning methods, time-series data analysis, data viz and practice creating user-friendly and reproducible code. We are tackling big questions about how social media is impacting our moral and political lives during a period when political polarization and disinformation campaigns are rising at an alarming rate.

Jim Duncan

Ebenezer K. Hunt Professor of Biomedical Engineering, Radiology, Electrical Engineering and Statistics & Data Science, https://seas.yale.edu/faculty-research/faculty-directory/james-duncan
james.duncan@yale.edu

- Towards Automated Liver Cancer Assessment for Image-Guided Therapies: A Machine Learning Approach

This work addresses efforts to use information derived from multiparameter Magnetic Resonance Images (mpMRI) to develop an image-guided strategy for treating patients with hepatocellular carcinoma (HCC) (primary liver cancer). Our current approaches use a variety of image analysis and machine learning strategies (including random forests and convolutional neural nets) for targeting the tumor/lesions, registering the mpMRI-derived targets to images in the treatment environment and finding image-derived biomarkers to predict which patients will respond to the treatment. Opportunities for undergraduate research include extending our current tumor classification approaches so that little-to-no human initialization is required, improving their accuracy and working on improved outcome prediction algorithms.

Mark Gerstein

Albert L Williams Professor of Biomedical Informatics, Molecular Biophysics & Biochemistry & Computer Science, http://www.gersteinlab.org/about/

Presented by Jonathan Warrell (Associate Research Scientist, Gerstein Lab)

mark@gersteinlab.org; jonathan.warrell@yale.edu; jonathan.warrell@gmail.com (cc both)

- Deep integrated models with informative latent structure for neurogenomics

We have two broad project areas which involve relating variation at genetic and epigenetic levels in the brain to high-level human cognitive traits. The first (a) involves modeling genetic risk for psychiatric conditions such as Schizophrenia or Autism. The second (b) involves linking genetic variation (and/or inter-region epigenetic variation) to differences in connectivity and hierarchical relationships between/within brain regions using fMRI data. In both cases, we are interested in learning deep predictive models, with latent structure that is informed by intermediate-level data representing potential mediating mechanisms, e.g. gene networks in (a) and neural circuit models in (b). We are interested in using techniques in deep-learning and probabilistic programming for this purpose. This could involve using variational techniques, such as amortized inference and likelihood-free variational methods to learn models with complex latent structure. Particularly, for (a) we would use model architectures related to Boltzmann machines, variational autoencoders and hierarchical implicit models, and in (b) we would draw on amortized program inference applied to simulations of neural circuits. In both cases, we would like to formulate a variational lower-bound which links the latent structure to the partially observed intermediate-level data. We will draw on (post-mortem) genetic and molecular data from the PsychENCODE project, and in (b) also use data from the Human Connectome Project.

Walter Jetz

Professor of Ecology & Evolutionary Biology, Forestry & Environmental Studies, https://jetzlab.yale.edu/ Yale Center for Biodiversity and Global Change
Max Planck-Yale Center for Biodiversity Movement and Global Change
walter.jetz@yale.edu

- Predicting the distribution of animals in space and time – from single individuals to global ranges
- Animal-deployed sensors to measure the pulse of our planet
- Uncovering mechanisms of animal migration through GPS tracking

Work in our group addresses the patterns and processes behind the distribution of species and their traits in space and time. We are particularly interested in the scale-dependence of both evidence and mechanism in biodiversity science and how environment, ecological, and macroevolutionary mechanisms combine to determine the co-occurrence of species and the structure of assemblages. We aim to use this as basis for assessing the fate of biodiversity and its functions under global change.

Yuval Kluger

Professor of Pathology, Applied Mathematics, ttps://medicine.yale.edu/lab/kluger/
yuval.kluger@yale.edu

- Unsupervised Ensemble Learning

Our research interests include development of data mining, network, and applied mathematics techniques for analyzing high dimensional data generated in genomics, proteomics and image analysis studies. Our lab combines independent methodological research with practical solutions to analytical tasks emerging in our collaborative projects with mathematicians, and with basic, translational and clinical biomedical researchers. Our lab specializes in development and implementation of applied mathematics and deep learning approaches to mathematically challenging bioinformatics problems. The tools and capabilities that we develop are important for the fledgling field of precision medicine, as they address anticipated challenges in analyzing very large genomics and other databases that fall into the category of Big Data.

Specific projects that do not require biological knowledge are in the following areas: (a) Deep learning approaches for analyzing genomics and imaging data, (b) Unsupervised ensemble learning and crowd-sourcing, (c) Computational tools for analyzing single cell RNA-seq and mass cytometry data, (d) Nonlinear dimensional reduction approaches for detecting genomic patterns, (e) Phylogenetics in cancer and immunobiology, (f) Statistical biomarker models

Smita Krishnaswamy

Assistant Professor of Genetics at the Yale School of Medicine and Computer Science in the Yale School of Applied Science and Engineering, https://www.krishnaswamylab.org/

Presented by Scott Gigante (Ph.D. Candidate, Krishnaswamy Lab)

smita.krishnaswamy@yale.edu

- Using Manifold-Learning to Analyze Biological Systems with Single-Cell Data

In recent years, single-cell resolution analysis of biological systems has become feasible and widespread; however, the resulting high-dimensional, high-throughput and heterogeneous data introduce numerous technical challenges. The Krishnaswamy Lab applies manifold learning through graph signal processing and deep learning to the analysis of a wide range of biological systems to achieve denoising, visualization, clustering, batch correction, anomaly detection, perturbation combination prediction, and more.

Karla M Neugebauer

Professor of Molecular Biophysics and Biochemistry and of Cell Biology, https://neugebauerlab.yale.edu/

Presented by Tara Alpert (Ph.D. Candidate, Neugebauer Lab)

karla.neugebauer@yale.edu

- Quantifying changes in splicing kinetics

Our lab is interested in studying the kinetics of splicing, a basic RNA processing step required for gene expression. Splicing is the removal of intronic sequences of RNA and subsequent ligation of surrounding exons. Since nearly all genes in humans are spliced, mistakes in splicing can be far reaching and are implicated in a wide-range of diseases from Spinal Muscular Atrophy to many cancers. Over the past few years, our lab has a developed method for precisely measuring the rate of splicing in budding yeast called Single Molecule Intron Tracking (SMIT). This technique produces high density data for the fraction of RNA molecules spliced at a certain moment in time. We plot this data as kinetic curves for each gene, and we observe a large amount of variation in the shape of each curve. Our biological questions rely on making mutations and observing how splicing kinetics change, but we currently have no reliable way of quantifying changes among these curves. Your project would be to find parameters or models to classify the curves from these data sets so we can interpret how splicing is affected.

Steven W. Zucker

David and Lucile Packard Professor of Computer Science and Biomedical Engineering, http://www.cs.yale.edu/homes/vision/zucker/steve.html

steven.zucker@yale.edu