Statistics & Data Science Project Pitch

September 9, 2019 | 3:45PM to 5:00PM @ Yale Institute for Network Science

Introduction:

Daniel A. Spielman

Sterling Professor of Computer Science; Professor of Statistics & Data Science, and of Mathematics

Project Pitches:

Eli Fenichel

Associate Professor, Yale School of Forestry & Environmental Studies
https://environment.yale.edu/profile/eli-fenichel/
eli.fenichel@yale.edu

Harnessing data for a more livable, environmental friendly, and sustainable planet

Understanding the natural world, and how people interact with it, is imperative for ensuring a sustainable future for earth. Scientists in the School of Forestry and Environmental Studies work with data to answer a range of questions from understanding how and why trees grow the way they do in the tropics; to what drives stream chemistry and water quality; to what do people around the world think about climate change; to how do researchers from different fields work together to solve environmental problems? These projects use a range of cross-sectional, time series, and panel data along with a range of analytical techniques including Bayesian analysis, network analysis, and regression modeling. Professors in the School of Forestry and Environmental Studies are excited about having DSDS students working on these important questions.

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.

Nelson Rios

Division of Informatics, Yale Peabody Museum
nelson.rios@yale.edu

Patrick Sweeney

Division of Botany, Yale Peabody Museum
patrick.sweeney@yale.edu

Automated Classification of Leaf Shape from Herbarium Specimen Images

Understanding the underlying evolutionary significance of the variety of plant leaf shapes seen in nature has long interested botanists, and many hypotheses have been proposed to explain leaf shape diversity. To more fully explore the correlation between leaf shape and various environmental factors that might have driven leaf shape evolution, we aim to characterize leaf shape for tens of thousands of leaves from hundreds of species. This project will focus specifically on the degree of "toothiness" around the outer edge of leaves. Undergraduate students will utilize the Zooniverse platform to annotate a few thousand leaves from images of plant specimens within the Yale Peabody Museum and other collections around the world. The annotated data will be used to train and test an Image Classification Convolutional Neural Network (CNN). The CNN will then be used to classify leaf margins for the remaining unclassified data. The end result will be a large data set that can be used to more deeply examine the relationship between toothiness and environmental variables that have been postulated to play a role in the evolution of leaf shape such as temperature, light, and moisture.

Valentina Greco

Carolyn Walch Slayman Professor of Genetics
www.grecolab.org Office: SHM I 336
valentina.greco@yale.edu

Tissue regeneration captured by live imaging

The goal of my laboratory is to discover the fundamental principles of organ regeneration during homeostasis and how pathology emerges from normal tissue growth. A major challenge in the mammalian stem cell field was the inability to follow the same cells in vivo. This limitation obscured insight into the dynamics of these processes, the contributions of intercellular interactions to tissue growth, and the initial events leading to malignancy. To overcome this challenge, my lab established, for the first time, the ability to visualize skin epithelial stem cells in an intact animal by two-photon microscopy. Looking at the skin epithelium, we have studied regeneration in real time by labeling and tracking stem cells and the different tissue types that surround them. To address their functional roles, we manipulate cells and their behaviors by altering key signaling pathways and study the consequences to identify the critical cellular and signaling components for the regenerative process. Our work addresses the fundamental principles utilized by the homeostatic process, which will provide critical knowledge that will lay the foundation to later address the cellular mechanisms that go awry during disease states such as cancer. This novel live imaging approach has repeatedly led us to a significant number of breakthroughs that neither we nor the field could have anticipated, including but not limited to 1) position dictates stem cell fate, 2) a stem cell-mediated phagocytic clearance mechanism that regulates the size of the stem cell pool and 3) the unanticipated plasticity of the skin to correct aberrant tissue growths induced by mutational and non-mutational insults. While we have gained increased understanding of the process of regeneration, the complexity of the large dataset we have generated and the use of more complicated molecular sensors is stimulating us to evolve computational approaches aimed to extract deeper learning from our experiments. *Our current rotation projects* revolve around understanding skin regeneration by investigating how different tissue types within the skin maintain their homeostasis, as well as how normal tissues cope with mutant stem cells.

Julian Jara-Ettinger

Assistant Professor, Psychology, Computer Science, and Cognitive Science Program
http://compdevlab.yale.edu
julian.jara-ettinger@yale.edu

Understanding other people's behavior through inverse optimal control

Sarah Demers

Horace D. Taft Associate Professor of Physics
sarah.demers@yale.edu

Observing the Higgs Boson

The Higgs boson, discovered in 2012 by the ATLAS and CMS experiments at CERN's Large Hadron Collider, is key to our understanding of the acquisition of mass by fundamental particles. The Higgs itself is a massive particle, and its mass is central to its identification. My group on the ATLAS experiment is working on a measurement of the Higgs in a production and decay mode that has not yet been observed. This project is to use machine learning techniques to improve the way we pick the signal out from the background with a strong reliance on a mass reconstruction technique. The primary experimental challenge is that the decay mode of the Higgs that we are hunting for involves the production of neutrinos, and our detector is not sensitive to neutrinos so they must be inferred in their absence. This project is to investigate the current performance of the technique used and to explore other methods of reconstructing the mass.

John Wettlaufer

A.M. Bateman Professor of Geophysics, Professor of Mathematics and Physics
Director of Undergraduate Studies, Applied Mathematics
https://gauss.math.yale.edu/~jw378/John_Wettlaufer/JSW.html
john.wettlaufer@yale.edu

Predicting Rare Events in Multiscale Dynamical Systems using Machine Learning

The dynamics of many systems in nature are nonlinear, multiscale and noisy, making both the the- oretical and numerical modeling and prediction of their states challenging. Of particular interest are those dynamics that often lead to rare transition events. Namely, the system under study spends very long periods of time in various metastable states and only very rarely and at seemingly random times, it hops between states. Understanding the dynamics of such systems requires us to study the ensemble of transition paths between the different metastable states. Using the same approach in observations, such as is found in climate data, is an essential aspect of mathematical climatology.

Hemant D. Tagare

Professor of Radiology and Biomedical Imaging
hemant.tagare@yale.edu

Can Machine Learning Help Understand Molecular Structure?

A mini scientific revolution is underway in biology - a relatively new technique called cryogenic electron microscopy (cryo-EM) has made it possible to reconstruct 3D shapes of protein molecules at near atomic resolution (the shapes of the molecules explain their function). At the heart of this method lies a set of mathematical and computational problems that my lab works on. We are interested in theoretical issues (proving theorems), in creating new algorithms, and in releasing software to the cryo-EM community. The computational problems in cryo-EM are extremely challenging - over 10,000 images are used for a single reconstruction, the images are extremely noisy, and the algorithms are slow. We are looking for students who would like to work on the theoretical or on the algorithmic problems in cryo-EM.

Maureen Long

Professor and Director of Graduate Studies, Geology and Geophysics
maureen.long@yale.edu

Machine learning approaches to denoising seismic data

Separating signal from noise is crucial in a host of scientific applications. For seismologists who use seismic waves to study the structure of the deep Earth, discriminating the signal of a seismic wave arrival from noise (both natural and anthropogenic) represents a key challenge. This is traditionally done by applying a digital filter, but this approach is problematic when the signal and the noise overlap in their frequency content. Recent work on applying machine learning approaches to seismology has led to more sophisticated “denoising” algorithms. These approaches have been applied to the problem of assembling earthquake catalogs for specific regions (such as Southern California), but they have not yet been applied to the study of the deep Earth. This project will apply a deep neural network approach to the problem of denoising seismic data in order to facilitate study of deep Earth structure and dynamics.

Tauhid Zaman
Associate Professor of Operations Management
Tauhid.zaman@yale.edu

Assessing the Impact of Bots in Social Networks

We are influenced by what we see on social networks. This fact can be exploited to run coordinated influence campaigns using automated accounts, known as bots. The effectiveness of these campaigns is determined by who the bots target, the overall network structure, and the activity level of the bots. In this project, the goal is to develop methods to assess the impact of the bots on opinions in social networks. Our team has developed algorithms for bot detection and impact assessment. The assessment algorithm includes a deep neural network which measures the opinion of text with respect to a specific topic. The goal of this project is to apply the bot detection and assessment algorithms to new datasets in Twitter, specifically those concerning the upcoming U.S. presidential election. Student responsibilities on the project include building these datasets by crawling Twitter to collect tweets and follower networks for different events (we have code for this), identifying the bots (we have code for this), and then assessing their impact (we have some code for this too). As part of this project, the student will gain experience with Python, the Twitter API, neural networks, and network opinion dynamics models.