Statistics & Data Science Project Pitch

December 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

Bonus Projects (without pitches):

[Data Science Projects in Fundamental Physics at Yale Wright Lab](https://wlab.yale.edu/sites/default/files/files/wl-data-science-2020.pdf)

At <https://wlab.yale.edu/sites/default/files/files/wl-data-science-2020.pdf>

Project Pitches:

# Vahideh Manshadi

Associate Professor of Operations, Yale School of Management

[vahideh.manshadi@yale.edu](mailto:vahideh.manshadi@yale.edu)

**Analytics for Food Rescue**

Over 60 million tons of food goes to waste in the U.S. each year, while 40 million people—including 6.5 million children—live in food-insecure households. This mismatch is driven in part by the cost of last-mile transportation required to recover perishable food from local restaurants and grocery stores. In recent years, online platforms have emerged to facilitate crowdsourcing food recovery through volunteer or paid labor. In collaboration with one such platform (Food Rescue US), we are developing predictive models and analytic tools to help make automated and better decisions. These decisions include (1) how to fairly and efficiently allocate donated food across different agencies, (2) how to engage volunteers without over-utilizing them, and (3) whether to prioritize the recruitment of volunteers or the recruitment of donors.

This project illustrates an application of data science in matching markets with direct societal impact. Even though we will work with data from Food Rescue US, the matching problems we study arise in many other applications.

# Douglas Duhaime

Developer at Digital Humanities Lab

[douglas.duhaime@yale.edu](mailto:douglas.duhaime@yale.edu)

**UMAP Parameter Optimization**

Our goal is to find UMAP parameters that reduce the dimensionality of input datasets while minimizing the distortion of pointwise distances.

# Ilker Yildirim

Assistant Professor of Psychology and Statistics and Data Science

[ilker.yildirim@yale.edu](mailto:ilker.yildirim@yale.edu)

<http://cncl.yale.edu>

**Computational basis of seeing and thinking**

Abstract: TBD

# Meg Urry

Israel Munson Professor of Physics

[meg.urry@yale.edu](mailto:meg.urry@yale.edu)

<https://urrylab.yale.edu/>

**Characterizing Black Holes and Galaxies**

(with graduate students Aritra Ghosh and Chuan Tian)

Galaxies (gravitationally bound ensemble of stars and gas) have very massive black holes at their centers. These grow to 106-1010 times the mass of our Sun by accreting matter (stars, gas) from the surrounding galaxy, probably in short episodes throughout most of cosmic history. This releases lots of energy into the galaxy, visible to us as radiation at radio to gamma-ray wavelengths, during what we call “Active Galactic Nuclei” (AGN) phases*. Do galaxy mergers trigger this growth? What impact does it have on galaxy evolution? How can we find the actively accreting black holes among the much larger population of inactive ones (i.e., how do we identify galaxies, stars and AGN among millions of surveyed sources)?*

We have two specific projects related to these questions:

1. *AGNFinder*: The goal is to use ML techniques to identify AGN in millions of objects that have measured multi-wavelength intensities of at 10-30 wavelengths. AGN should separate cleanly from stars and galaxies. (Perhaps unsupervised methods?) We can provide unlimited simulated sources for training and testing, based on theoretical models, and smaller samples of observed data that has already been validated.
2. *GaMorNet*: The goal is to measure galaxy properties like morphology (which encodes the recent merger history), size, brightness (total light emitted), mass (which can be derived from the light as a function of wavelength), and star formation rates (ditto). We can provide large samples of simulated galaxies (100,000 or more) and smaller samples of observed galaxies with measured characteristics (1000s).

Priya Panda (presented by Timothy Foldy-Porto)

Assistant Professor, Electrical Engineering

<https://intelligentcomputinglab.yale.edu/>

[priya.panda@yale.edu](mailto:priya.panda@yale.edu)

**Enabling ubiquitous machine intelligence with algorithm-hardware co-design**

Advances in deep learning have led to computers matching or surpassing human performance in several cognitive tasks. However, implementation of such neural algorithms in conventional "von Neumann" architectures are several orders of magnitude more area and power expensive than the biological brain. Hence, we need fundamentally new approaches to sustain exponential growth in performance at high energy-efficiency beyond the end of the CMOS roadmap in the era of ‘data deluge’ and emergent data-centric applications. Exploring the new paradigm of computing necessitates a multi-disciplinary approach: exploration of new learning algorithms inspired from neuroscientific principles, developing network architectures best suited for such algorithms and new hardware techniques to achieve orders of improvement in energy consumption. The **Intelligent Computing Lab** at Yale is focused on enabling enhanced learning capabilities in the next generation of artificial intelligence systems with algorithm-hardware co-design approaches. Opportunities for graduate/undergraduate research include: a) Designing novel spike-based temporal computing solutions to address the algorithm vulnerabilities (such as, adversarial attacks, non-interpretability) of today’s intelligent systems; b) Exploring bio-plausible algorithms and network architectures guided by natural intelligence (how the brain learns, the internal fabric of the brain etc.) to define the next generation of robust and efficient AI systems for beyond-vision static recognition tasks with the ability to perceive, reason, decide autonomously in real-time.

Jeffrey Park (presented by Will Frazer)

Professor of Geology and Geophysics

[jeffrey.park@yale.edu](mailto:jeffrey.park@yale.edu)

**Searching for Seismicity Under Volcanic Islands**

We work in seismology which is the study of earthquakes and the waves they generate. In particular, we are interested in time series analysis of the signals recorded by seismometers as seismic waves. A new hypothesis for the structure beneath volcanic islands (e.g. the Hawaiian Islands) suggests there may be significant mineral hydration at the crust-mantle boundary, also known as the Moho. In order for this hydration to occur, seawater must move from the ocean floor through the crust to the Moho. This movement of fluids would generate seismic tremor which can be detected by seismic instruments located on the surface; however, these signals are low in amplitude and are difficult to detect via traditional techniques due to a low signal to noise ratio. This project will consist of examining time series data from long running seismic stations from the Hawaii Volcano Observatory for previously undetected tremors using both time- and frequency-domain methods such as convolution. These analyses will utilize previously catalogued events as templates to find new tremor and different ways of ensuring robust detection will be examined. After the source location of the tremors are determined with a technique called double-difference relocation, an interpretation of their relevance to fluid movement will be made. (Rock fracture, magmatic flow, and fluid infiltration should have distinct spectral signatures.) This project will enhance the understanding of how volcanic islands form and evolve in the context of global plate tectonics.

Chris Cotsapas

Associate Professor of Neurology and Genetics, Yale School of Medicine

[chris.cotsapas@yale.edu](mailto:chris.cotsapas@yale.edu)

**Identifying cell responses to stimulus**

We are trying to understand how immune cells respond to external stimuli. As you might expect, not every cell responds to every stimulus. We know that this is - in part - genetic: variants in DNA alter signal reception and/or signal propagation through cells, and this alters the likelihood that a cell will trigger a full response to the stimulus. We have recently discovered that this set-point is

associated to diseases of the immune system, but we do not understand the details. How do quantitative changes in stimulus response affect go/no-go decisions in an individual cell? What genes drive these decisions?

To answer these questions we have profiled cell responses from 90 donors using signaling assays and genomic technologies (single-cell RNA sequencing, ATAC sequencing, and enhancer element mapping). We are offering projects in data integration, statistical modeling and causal inference to interested students, who will work both with Chris and with post-docs in the laboratory.

Emily Gilmore

Associate Professor of Neurology, Yale School of Medicine

[emily.gilmore@yale.edu](mailto:emily.gilmore@yale.edu)

**Predicting (and preventing) neurologic complications of acute brain injury**

In emergency and critical care neurology, we treat patients with a broad range of acute brain injury, including large ischemic strokes, brain bleeds, uncontrolled seizures, infections and traumatic brain injury. After their initial injuries, many patients suffer complications both during and after their hospitalization. If we had better ways of understanding and predicting when and who is at risk for developing these complications, we might be able to prevent them. We believe that prevention using individualized treatments based on one’s intrinsic physiology rather than on population level thresholds will lead to improved outcomes. At Yale, we are amassing enriched timeseries brain and body physiologic data to optimize the individualized approach for each patient. Given the challenges such robust data streams pose analytically for the bedside clinician to make real-time decisions, we want your help in applying data science and machine learning methods to develop and refine prediction models aimed at enhancing treatments, and ultimately outcomes, for our patients.

Ziad Ganim

Assistant Professor of Chemistry

ziad.ganim@yale.edu

**High-Throughput Single Molecule Chemistry with Optical Tweezers**

We would like to develop a side-loaded convolutional neural network that is capable of classifying single molecule force extension data into categories such as: 1 worm-like chain, >1 worm-like chains, junk. Side-channels include single molecule fluorescence. Our raw data is sampled at 50 kHz (~1 Mb/s) to measure the XYZ position of two optically trapped microspheres with nm precision. We would like to train the model using a ~3D Langevin equation with the goal of extracting an analytical expression that describes this statistical data set. Creative contributions to data visualization and scientific communication are welcome and appreciated.

Jun Deng

Professor of Therapeutic Radiology, Yale School of Medicine

[jun.deng@yale.edu](mailto:jun.deng@yale.edu)

[http://radonc.yale.edu/research/radiological/people/jun\_deng-1.profile](https://nam05.safelinks.protection.outlook.com/?url=http%3A%2F%2Fradonc.yale.edu%2Fresearch%2Fradiological%2Fpeople%2Fjun_deng-1.profile&data=02%7C01%7Cjun.deng%40yale.edu%7Cac702d1ba9e5435434d308d760e329f0%7Cdd8cbebb21394df8b4114e3e87abeb5c%7C0%7C0%7C637084402446724336&sdata=WlUpFBgodoDq9zrkEjfETKIHeSHEEUJuY2sU5XGDWmM%3D&reserved=0)

[http://medicine.yale.edu/lab/deng/index.aspx](https://nam05.safelinks.protection.outlook.com/?url=http%3A%2F%2Fmedicine.yale.edu%2Flab%2Fdeng%2Findex.aspx&data=02%7C01%7Cjun.deng%40yale.edu%7Cac702d1ba9e5435434d308d760e329f0%7Cdd8cbebb21394df8b4114e3e87abeb5c%7C0%7C0%7C637084402446734330&sdata=%2BlhtA5KfC3S9gPZAJyfiHgad%2FfOgMnwCvVmI%2Fp1lVDI%3D&reserved=0)

**Statistical Biopsy of Personal Health Data for Cancer Early Detection**

Statistical biopsy is a new approach proposed by our group that mines personal health data for early cancer detection with statistical modeling. The basic idea is that the trove of personal health data can be used to train and validate deep learning models to generate a holistic profile of one’s risks for a variety of cancers simultaneously prior to disease onset. Analogous to tissue biopsy that evaluates cells from a tissue specimen and liquid biopsy that evaluates circulating tumor DNA from a fluid sample, statistical biopsy mines personal health data from individuals for early cancer detection. What is different is that statistical biopsy seeks to decipher the invisible correlations and inter-connectivity between multiple medical conditions and health parameters via sophisticated statistical modeling. It offers a cost-effective and safe approach to cancer screening in real time, informing preventive interventions and screening decisions.

Edward J. Vytlacil

Professor of Economics

**Ranking Treatments from Multi-Arm RCTs while Imposing Alternative Monotonicity Restrictions**

A pervasive problem in multi-arm randomized controlled trials (RCTs) is how to evaluate the relative effectiveness of alternative treatment arms in the presence of non-compliance. We are developing a general framework for inference on average treatment effect parameters and the rankings of treatments effects in multi-arm RCTs while allowing for non-compliance. Our general framework allows the researcher to impose a priori monotonicity restrictions. For example, the researcher might be willing to impose that assigning a subject to a given treatment will never cause a subject who otherwise would have taken the treatment not to do so, or might be willing to impose that a certain treatment cannot harm the subject. We are using a linear programming formulation to present a flexible framework and to develop general results for characterizing the testable restrictions and the sharp identification of treatment effect parameters and the rankings of treatments in terms of these parameters that exploit the structure of the RCT while additionally imposing alternative monotonicity restrictions on how treatment uptake depends on random assignment and how the outcomes depend on the treatments. The resulting testable restrictions and partial identification results are given by intersections of moment inequalities. We are developing inference methods by exploiting and extending the literature on inference for moment inequalities. We wish to apply our methodology to RCTs conducted in various fields of micro-economics. For example, we wish to use our methodology to investigate the relative effectiveness of public vs private job search assistance; to evaluate the effects of cash incentives and therapy on reducing crime in Liberia; and to study the effects of alternative early childhood interventions.