**Data Science Project Match
January 20, 2022**

# Introduction by Daniel Spielman

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

# **Projects**

Jun Deng

Professor of Therapeutic Radiology

jun.deng@yale.edu

<https://medicine.yale.edu/lab/deng/>

**Bridge to Artificial Intelligence Program and Digital Twins Grand Challenge**

Rapid advancements in artificial intelligence (AI) and machine learning (ML) in recent years have shown great promise in discovering new insights in biomedical and behavioral research, feeding upon large amounts of data of various modalities. Yet, these multimodal datasets are scattered and disconnected, not amenable to AI/ML analysis at scale. Furthermore, there is the potential for unethical handling in the data collection, algorithm development, and system training, which may contribute to health inequities, social harms, and structural injustices. To address these challenges, the National Institutes of Health (NIH) has recently initiated a brand-new Bridge to Artificial Intelligence (Bridge2AI) Program, aiming to generate flagship datasets and best practices for the collection and preparation of AI/ML-ready data. In response, we have assembled a multi-institutional team with multi-domain expertise to address the Grand Challenge of digital twins for predictive health. Via this Grand Challenge, we aim to: (1) generate new biomedical and behavioral datasets that are ethically sourced, trustworthy, and amenable to AI/ML; (2) develop software to standardize data attributes across multiple data sources and types; (3) build automated tools to assist the creation of FAIR and ethically sourced datasets; (4) share the resources by disseminating data, and providing ethical guidance, tools and best practices; and (5) construct cross-training materials and activities for workforce development.

# Andrew Taylor

Associate Professor of Emergency Medicine and Biostatistics

ED Director of Clinical Informatics and Analytics

YCMI Associate Director of Informatics and Data Science Research Yale School of Medicine

richard.taylor@yale.edu

<https://ysph.yale.edu/profile/richard_taylor/>

**Leveraging Graphs and Deep Learning to Enhance Data Quality in Patient-Centered Outcomes Research**

Research networks using large quantities of data from electronic health records (EHRs) hold considerable promise for improving patient health, but these transaction-oriented systems were not designed for research, which contributes to inaccurate findings. Developing automated methods to identify and resolve data quality issues within and across multiple sites is, therefore, essential to the validity of EHR-based clinical research. This project is to develop and expand methods for data quality curation through filtering and denoising strategies. We will leverage a variety of state-of-the-art methods that use graph-based methods and deep learning to denoise and filter the patient data.

# Hyojung Seo

Assistant Professor of Psychiatry, Yale School of Medicine

hyojung.seo@yale.edu

<https://medicine.yale.edu/profile/hyojung_seo>

**Understanding how brain generates intelligent behavior via complex neural networks**

Cognitive behavior is generated by coordinated activity across networks of neurons, but it remains poorly understood how complex spatiotemporal features of network activity mediate diverse elements of cognition. Exploiting recent advances in neurotechnology and computer science, we are interested in exploring new tools to analyze and model neural activity underlying cognition. First, several statistical methods have been proposed to decompose/analyze high-dimensional population neural activity recorded simultaneously from many neurons. The project aims to explore and assess the proof-of-concept methods by applying them to neural activity recorded from diverse brain areas and under different behavioral contexts. Second, the project aims to use artificial neural network and deep learning to model how cognition such as theory of mind can emerge from neural network dynamics.

Luke Sanford

Assistant Professor of Environmental Policy and Governance, Yale School of the Environment

luke.sanford@yale.edu

<https://lcsanford.github.io/>

**Satellite imagery and deep learning for improving estimates of carbon sequestration**

This project uses machine learning methods and satellite imagery to evaluate the impact of environmental protection. We combine state of the art remote sensing methods and statistical techniques in a double machine learning framework to estimate the net carbon impacts of protected areas and carbon sequestration schemes. Traditionally the difficulty with this type of evaluation lies in predicting the counterfactual outcome—what would have happened on the land in the absence of the protection or carbon offset program. The method being developed uses the wealth of information stored in the satellite record to improve the accuracy and precision of these counterfactual estimates. We are interested in applying deep learning methods that make use of the spatial (convolutional) and temporal (recurrent) information stored in the satellite imagery.

Meg Urry

Israel Munson Professor of Physics

Director, Yale Center for Astronomy and Astrophysics

meg.urry@yale.edu

<https://urrylab.yale.edu>

**Searching for Rapidly Growing Black Holes in Large Multi-Wavelength Data Bases**

Galaxies (gravitationally bound ensemble of stars and gas) have very massive black holes at their centers, each growing up to a billion times as massive as our Sun as it accretes matter (stars, gas) from the surrounding galaxy in short growth spurts 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. We know that most of these AGN are shrouded by gas and dust, which can make them hard to find. Using data from large multi-wavelength surveys, we want to develop a tool, provisionally called AGNFinder, that will be able to differentiate the multi-wavelength emission of AGN in the optical, infrared, and x-ray bands from other astronomical objects such as stars and galaxies. Furthermore, in addition to separating AGN from inactive galaxies, we hope to identify and classify sub-types of AGN and improve our understanding of how types of AGN physically differ from one another. We have access to a database of multiwavelength photometry for several million AGN, which we can easily supplement with similar data for galaxies and stars; we can also provide synthetic data (based on models) for use in training AGNFinder.

Steven H. Kleinstein
Professor, Department of Pathology, Department of Immunobiology

Co-DGS, Interdepartmental Program in Computational Biology & Bioinformatics Human and Translational Immunology Program

Project presented by Anna Konstorum

Bioinformatics Scientist in the Department of Pathology, Yale School of Medicine

anna.konstorum@yale.edu

<https://medicine.yale.edu/lab/kleinstein/>

**Improving non-negative tensor decomposition methods for immunology time-course data**

John Tsang

(Incoming) Professor of Immunobiology

(Incoming) Director of Center for Systems and Engineering Immunology

john.tsang@yale.edu

jt18@hey.com

<https://www.niaid.nih.gov/research/john-tsang-phd>

**Multi-scale modeling and prediction of immune responses**

Immune response to perturbations such as infections and vaccination (e.g., SARS-CoV-2) involves complex molecular and cellular events occurring across space and time. They have been productively studied for several decades at the level of individual molecules, cells, and interactions, yet our ability to predict immune response quality (e.g., whether someone is protected against a pathogen) and quantity (e.g., how much antibodies will be generated after vaccination) remains poor. Human immune responses are also incredibly diverse at the individual level as exemplified by the enormous range of disease manifestations in COVID-19; similar variability has been observed in other scenarios including responses to vaccination and cancer. Thus, being able to predict immune response outcomes based on individual immune profiles and statuses can help advance translational applications, including more precise and individually tailored vaccination strategies. Our lab integrates systems biology, computational and experimental approaches to quantitatively model and predict immune response behavior. Here I highlight two project areas that involve the integration of data science, machine learning, and dynamical systems modeling.

1. Integration of real-time and pseudo-time using multimodal single cell data to predict immune cell dynamics in response to COVID-19 and vaccination
2. Building interpretable ML models of multi-immune cell dynamical interactions in autoimmunity: can we predict autoimmune escape?

Mark Gerstein

Albert L Williams Professor of Biomedical Informatics

Professor of Molecular Biophysics & Biochemistry, of Computer Science, and of Statistics & Data Science

Project presented by Jonathan Warrell

Associate Research Scientist, Gerstein Lab

mark@gersteinlab.org, jonathan.warrell@yale.edu

<http://gersteinlab.org>

**Quantum machine learning for problems in genomics and neuroscience**

We are currently investigating how quantum machine learning models can be applied to problems in genomics and neuroscience. We are interested in two classes of quantum model, the first being types of quantum Boltzmann machine which can be applied to model gene regulatory networks, and the second being quantum neural networks, which can be applied to learn predictive models of genetic traits. Some of the questions we are interested in are: What representational advantages do quantum network models have over matched classical models? Can we integrate techniques such as path integral Monte Carlo with variational machine learning methods to train quantum models? Can we develop methods of model interpretation to map information learned by quantum models back to the biological domain? Do quantum neural networks have distinct properties in terms of their capacity for generalization compared to classical neural networks? In each case, we will be interested in using psychiatric genomics data as a test bed within which to investigate such questions. We will have access to genetic and functional genomics (transcriptomics, epigenomics and single-cell) data from the PsychENCODE consortium, which has generated multiple sources of data to investigate the genetic determinants of psychiatric conditions such as Schizophrenia or Autism, and has provided us support for investigating the potential of quantum machine learning in the context of neuroscience.

Jay Emerson

Professor of Statistics (Adjunct)

Director of Graduate Studies Department of Statistics and Data Science

john.emerson@yale.edu

<http://www.stat.yale.edu/~jay>

Also presented by Martin Wolf

Postdoctoral Associate, Yale Center for Environmental Law & Policy

**Topics in Environmental Performance and Sustainability**

With Martin Wolf (YSE), Dan Esty (YSE), and Jay Emerson (S&DS)

The Yale Center for Environmental Law and Policy has produced the Environmental Performance Index (http://epi.yale.edu; previously the Environmental Sustainability Index) for more than 20 years, seeking to provide the best data-driven summary of environmental sustainability at the national level. There are a range of possible topics which could be appropriate for a Senior Essay, including a sensitivity analysis, a study of drivers of environmental performance, and/or further research to improve our ability to assess "progress towards 2050" on the complex topic of climate change.

Smita Krishnaswamy

Associate Professor of Genetics, Yale School of Medicine and Department of Computer Science, Yale School of Applied Science and Engineering

smita.krishnaswamy@yale.edu

<https://www.krishnaswamylab.org>

**Exploring the conversion from statistical ML to logic networks for decision-making in healthcare**

This project involves inferring logic from machine learning models that interpret healthcare decisions. Potential approaches could be to use random forests to learn decisions and then convert that to a logical chain of inference, a second approach would be to use the recently proposed logical neural network and infer interpretable logic from that. The end goal is to understand key health factors involved in diagnostics or treatment decisions and attribute them to original features.