Machine Learning for Chemical Mixtures in Epidemiology—Marianthi-Anna Kioumourtzoglou

Traditionally, environmental epidemiologic studies have focused on assessing risks related to a single environmental exposure at a time. This does not reflect reality, since we are constantly exposed to multiple chemicals at once. It is very important, therefore, to be able to assess exposure to chemical mixtures when conducting environmental epidemiologic analyses. Doing so is especially challenging, mainly due to the high dimension of the multi-chemical exposure matrix and because these chemicals are usually very highly correlated with each other. Machine learning (ML) techniques are being increasingly incorporated into environmental mixture research to overcome issues with traditional methods. ML approaches, however, are often developed to serve a different purpose, mostly focusing on optimizing predictive accuracy, which is not necessarily well-aligned with Public and Environmental Health. Although such methods are very useful and can help environmental health researchers deal with high-dimensional and complex data, we emphasize the importance of well-defined research questions and adaptation, and if necessary extension, of robust ML methods that can yield interpretable results instead of solely prioritizing predictive accuracy.

Computationally Leveraging the Collective: Mining Published Data and Crowdsourcing Consensus Models—Nicole Kleinstreuer

Advances in computational tools now provide unprecedented access to collecting, curating, and analyzing large datasets, and enable the construction of models that can provide predictions across the chemical universe. This talk will examine the application of machine learning methods to address problems in predictive toxicology, from identifying high-quality reference data and mechanistically informative tools to global crowdsourcing approaches to model adverse health outcomes. A manually curated database of guideline-like animal studies is being used as a training set to develop machine learning models that can be combined with natural language processing to automatically identify high quality studies in the scientific literature. Large, curated datasets have been used in international collaborations to build quantitative structure-activity relationship models to predict endpoints such as endocrine disruption and acute systemic toxicity. Consensus QSAR models leverage the strengths and compensate for the weaknesses of any individual algorithm, and provide robust predictions across a wide range of chemicals. These and other examples will highlight the way that computational toxicology is changing the environmental health landscape, not only through how new models are being built, but how data is being collected, curated, and combined.

Ethical decision making in computational systems: mechanisms for ensuring social trust—Alex John London

This talk presents an overview of ethical issues in the use of artificial intelligence in the area of health, broadly construed. Because learning is often structured by values—from the goals of inquiry, the criteria for success or failure, and constraints on how inquiry can be carried out—methodological issues surrounding the use of different learning strategies are connected to a network of ethical issues. Whether
stakeholders trust the results of inquiry often depends on the warrant for claims about the appropriateness of the question asked, the accuracy and reliability of the methods used to carry out inquiry, and mechanisms in place to ensure accountability. Issues that will be discussed include: whether data is fit for purpose, the potential for bias in data to produce biased models, the difference between systems that make actuarial predictions and those that guide intervention, the value of explainability and the challenges of ensuring that results seen in system development will be preserved in real world application.

A Large Multitask Neural Network for Chemical Hazards—model deployment and evaluation.—Thomas Luechtefeld

Chemical data is growing at an exponential pace, with thousands of new tests and millions of chemicals tested each year. However, most chemical property models focus on results from a handful of test types and fail to leverage the vast majority of available data.

Neural networks enable transfer learning where models trained on one task accelerate learning on other tasks. The rapid expansion of chemical data increases the value of transfer learning between chemical property domains. For example, models of eye irritation can benefit from models of skin irritation via transfer learning. This kind of transfer learning allows models to leverage much larger data sets.

In this presentation, we show transfer learning models result in accuracy improvements across a variety of chemical endpoints. The UL Cheminformatics Tool Kit models 74 different chemical endpoints including acute oral-, dermal- and inhalation-toxicity, dermal sensitisation, dermal- and eye-irritation, genotoxicity, and acute- and chronic-aquatic toxicity. These models are constructed in two stages. Stage (1) constructs a neural network for each endpoint from chemical structural data. Stage (2) constructs networks with inputs taken from stage 1 hidden layers.

We will discuss transfer learning, graph convolutional networks, computational validation approaches and their relevance to chemical hazard modeling.

Battle of Two Cultures: Statistics Versus(?) Data Science in Environmental Health—Bhramar Mukherjee

The title of my talk is inspired by Leo Breiman's seminal paper in 2001, "Statistical Modeling: The Two Cultures", where Breiman describes the intellectual tension between a classic stochastic modeler and an algorithmic modeler. The shock that "data science" has injected into the world of classical statistics has been palpable. I will compare and contrast these approaches specifically in the context of environmental health, in particular, with respect to analysis of multipollutant chemical mixtures. My focus will remain on searching for interaction and mediation effects with data on a large number of exposure and other molecular biomarkers. This is joint work with multiple collaborators in Biostatistics, Epidemiology and Environmental Health at the University of Michigan School of Public Health.

“I’m sorry, Dave. I can’t do that.” Social and ethical issues in training and education.—Lance A. Waller

The Statistical Society of London was founded in 1834 with the motto *Aliis exterendum*: “Let the others thrash it out”, reflecting a desire to provide objective quantitative results where the meaning of the calculations could be determined by others. Discussions of the need for ethics in statistical and computational training can be animated, controversial, and challenging. In today’s data rich society, the lines between assumption and bias, calculation and manipulation, and interpretation and “spin” are often blurred when data are scraped from the internet, preprocessed, passed through a complex computational
pipeline, and summarized in lavish visualizations. We briefly review guidelines for ethical statistical practice; cultural frameworks for ethics training in data science; how to incorporate data science accomplishments in tenure and promotion cases; how everything is conditional, especially probabilities; and how a homework problem on Bayes’ Theorem illustrates why we shouldn’t be surprised that facial recognition can lead to false arrests.

**Machine learning and artificial intelligence for biomarker discovery—Katrina Waters**

Biomarkers are measurable indicators of the presence or severity of disease, infection or exposure. They are often identified as biological molecules found in blood or other tissues that coincide with clinical symptoms or endpoints, which are themselves surrogate indicators of disease. The data sets used for biomarker identification are often complex, large in volume, and disparate in structure, making it challenging to apply traditional machine learning approaches. For example, issues surrounding quality, dynamic range, missing data and uncertainty quantification for different ‘omics data streams make it difficult and unlikely that a single model will be appropriate for the development of a biomarker classifier. Furthermore, individual variability of biomolecular signatures in humans based on their genetic background, diet, and lifestyle make biomarker discovery a precision medicine problem as well. Therefore, the future of biomarker implementation will ultimately require deployable sensors to monitor an individual’s state over time and artificial intelligence approaches to sense, reason, act and adapt to real-time information. This talk will illustrate the challenges and opportunities for machine learning and artificial intelligence approaches to biomarker discovery with examples from diabetes, ebola viral disease, and first responders.

**Predicting Chemical Exposure Pathways—John Wambaugh**

Centers for Disease Control and Prevention data shows that how we use a chemical is related to how much of that chemical can be found within us. On average, some chemicals present in household products occur at levels hundreds of times higher than chemicals that are only used outside the home. While databases exist describing how chemicals are used and models exist for predicting the human exposures that might occur, none of these cover all of the hundreds of thousands of chemicals listed in EPA’s CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard). EPA’s Office of Research and Development uses machine learning to fill these gaps and make predictions of public exposure to chemicals. *This abstract may not reflect U.S. EPA policy.*

**A Picture Tells a Thousand…Exposures—Scott Weichenthal**

Environmental pollution is a global health concern with economic impacts measured in billions of dollars each year. However, it remains difficult to estimate spatial and temporal variations in environmental exposures over broad spatial scales and characterizing the health impacts of multiple spatially-correlated exposures remains a challenge. In most regions of the world, environmental exposure data is missing entirely. New methods are needed to provide cost-effect estimates of environmental exposures to support policy interventions and evaluate health risks.

This lecture will discuss the promise, challenges, and probable data sources needed to apply AI in the fields of exposure science and environmental health. In particular, we will focus on the use of deep convolutional neural networks to estimate environmental exposures using images and other complementary data sources. As an illustrative example, we will discuss our recently developed IMAGE-PM2.5 model that can estimate global variations in long-term average outdoor fine particulate air pollution concentrations using only satellite images. We will also illustrate how deep convolutional neural
networks can be used to extend the spatial scale of traditional approaches to environmental exposure assessment and present several ongoing research projects related to the use of deep learning in environmental health research.

**Integrating AI Across the Spectrum of Environmental Health Sciences—Richard Woychik**

The National Institute of Environmental Health Sciences (NIEHS) recognizes the importance of utilizing artificial intelligence to advance the Institute’s strategic priorities. The 2018-2023 NIEHS Strategic Plan streamlines the Institute’s research priorities into three interrelated themes: advancing environmental health sciences, translating those advances from data to knowledge to action, and supporting the training, processes, partnerships and infrastructure to make the mission possible. Specifically, the strategic plan includes the development of innovative data science and data-driven approaches, including data sharing platforms, integration, and analytics, as integral components of environmental health science research and successful health initiatives. This talk will explore the challenges and opportunities of using machine learning and artificial intelligence to develop tools to analyze and integrate large-scale -omics datasets and elucidate molecular mechanisms to understand genetic susceptibility to environmental exposures. Dr. Woychik will also discuss NIEHS collaborations with external stakeholders and grantees and future training opportunities for the integration of artificial intelligence into environmental health science research.