Vassar College

Data Science and Society

Data Science and Society Colloquium Series Talk

Kristina Lerman
University of Southern California
Information Sciences Institute

Biases in Data & Other Threats to
Validity of Predictive Models

Friday, April 7, 2023

4:00pm

Rockefeller Hall 310

Tea at 3:30pm (Rocky 305 - Math Lounge)

Data is often heterogeneous, generated by subgroups with different traits and behaviors. The correlations between the traits, behaviors, and how the data is collected, can bias data and confound analysis. Models trained on biased data will make invalid inferences about individuals – what's known as ecological fallacy. The inferences can also unfairly discriminate against individuals based on their membership in protected groups. I describe common sources of bias in heterogeneous data, including Simpson's paradox, survivor bias, and aggregation bias, showing that ignoring these sources of bias can dramatically alter conclusions of analysis and lead to wrong policy recommendations.

I highlight with an example of COVID-19 pandemic to show that spatial aggregation of disease statistics exaggerates estimated growth rates. Finally, I describe a mathematical framework for de-biasing data that addresses these threats to validity of predictive models. The framework creates covariates that do not depend on protected features, such as gender or race, and can be used with any model to create fairer, unbiased predictions. The framework promises to learn unbiased models even in analytically challenging data sets.

Kristina Lerman is a Principal Scientist at the University of Southern California Information Sciences Institute and holds a joint appointment as a Research Professor in the USC Computer Science Department. Trained as a physicist, she now applies network analysis and machine learning to problems in computational social science, including crowdsourcing, social network and social media analysis. Her work on modeling and understanding cognitive biases in social networks has been covered by the Washington Post, Wall Street Journal, and MIT Tech Review. She is a fellow of the AAAI.