DATA-INSPIRE Workshop:

***Making predictions with big data: fitting, overfitting and***

***physics-aided learning***

Sponsored by the TRIPODS DATA-INSPIRE Institute, a joint collaboration of DIMACS and the Rutgers Departments of Computer Science, Mathematics, and Statistics http://robotics.cs.rutgers.edu/data-inspire/

Date: *November 20, 2020,*

 *2:00 pm – 4:00 pm (Eastern Time)*

Zoom Link: <https://rutgers.zoom.us/j/95969939274?pwd=WEk5UngvbklCd2RxN1hDQUUram01Zz09>

Program

2:00-2:50: Speaker: ***Peter L. Bartlett*** (University of California at Berkeley)

 Title: Benign Overfitting

2:50-3:00: Q&A

3:00-3:50: Speaker:***C. F. Jeff Wu***(Georgia Institute of Technology)

 Title: Navier-Stokes, spatial-temporal kriging and combustion stability: a prominent

 example of physics-based analytics

3:50-4:00: Q&A

**Abstracts**

**Peter L. Bartlett** (University of California at Berkeley)

Title: Benign Overfitting

Abstract: Classical theory that guides the design of nonparametric prediction methods like deep neural networks involves a tradeoff between the fit to the training data and the complexity of the prediction rule. Deep learning seems to operate outside the regime where these results are informative, since deep networks can perform well even with a perfect fit to noisy training data. We investigate this phenomenon of ‘benign overfitting’ in the simplest setting, that of linear prediction. We give a characterization of linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of effective rank of the data covariance. It shows that overparameterization is essential: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size.  It also shows an important role for finite-dimensional data: benign overfitting occurs for a much narrower range of properties of the data distribution when the data lies in an infinite dimensional space versus when it lies in a finite dimensional space whose dimension grows faster than the sample size. We discuss implications for deep networks, for robustness to adversarial examples, and for the rich variety of possible behaviors of excess risk as a function of dimension, and we describe extensions to ridge regression and barriers to analyzing benign overfitting based on model-dependent generalization bounds.

Joint work with Phil Long, Gábor Lugosi, and Alex Tsigler.

**C. F. Jeff Wu** (Georgia Institute of Technology)

Title: Navier-Stokes, spatial-temporal kriging and combustion stability: a prominent example of physics-based analytics

Abstract: Most “learning” in big data is driven by the data alone. Some people may believe this is sufficient because of the sheer data size. If the physical world is involved, this approach is often insufficient. In this talk I will give a recent study to illustrate how physics and data are used jointly to learn about the “truth” of the physical world. It also serves as an example of physics-based analytics, which in itself has many forms and meanings. In an attempt to understand the turbulence behavior of an injector, a new design methodology is needed which combines engineering physics, computer simulations and statistical modeling. There are two key challenges: the simulation of high-fidelity spatial-temporal flows (using the Navier-Stokes equations) is computationally expensive, and the analysis and modeling of this data requires physical insights and statistical tools. A surrogate model is presented for efficient flow prediction in injectors with varying geometries, devices commonly used in many engineering applications. The novelty lies in incorporating properties of the fluid flow as simplifying model assumptions, which allows for quick emulation in practical turnaround times, and also reveals interesting flow physics which can guide further investigations.