EXTENDED ABSTRACT ONLY

Big decisions need more than just big data, they need big models too

B. Kramer[a], E. Qian[a], R. Swischuk[a] and K. Willcox[b]

[a] *Department of Aeronautics and Astronautics, Massachusetts Institute of Technology*
[b] *Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin*
Email: kwillcox@oden.utexas.edu

Abstract: The field of data science has exploded in the last decade, not just in the realm of the internet and social media, but also for physical systems across science, engineering and medicine. This explosion of the field is fueled in large part by the explosion in volumes of data that are being produced. But it is also fueled by the availability of computing power and the tremendous progress in algorithms. We now have the ability to collect massive amounts of data, and we also have the ability to analyze it. The central question is: *How do we extract knowledge, insight and decisions from all of these data?*

Recent years have seen incredible success of machine learning methods in recommendation systems, social media, speech recognition, and more. But when it comes to high-consequence decisions in engineering, science and medicine, we need more than just the data. These decisions are almost always based on predictions that go beyond the available data. We often need to make predictions about a future state – about the future state of a patient's illness, about the states that an engineering system may find itself experiencing in operation, or about the future state of the Earth's climate in the decades to come. In these settings, there are multiple reasons that pure-data machine learning and statistical approaches will struggle to generalize with high confidence:

- The applications are characterized by complex multiscale multiphysics dynamics, so that small changes in parameters can lead to large changes in system behavior.
- The parameter space is very high dimensional. Many parameters of interest are fields (infinite dimensional). Without the constraints of physics, the solution space is so vast that driving decisions with data alone is doomed to failure.
- Data are sparse and typically rely on physical sensing infrastructure, making them expensive to acquire. Data may be large in volume, but they provide only limited peeks into the underlying high-dimensional parameter space.
- Uncertainty quantification of predictions must provide quantified confidence in the recommended decisions. This is especially challenging but especially important as we extrapolate beyond the data to issue predictions about future states.

This talk will introduce the notion of *Predictive Data Science*, which employs a synergistic combination of data and physics-based models. Learning from data through the lens of physics-based models is a way to bring structure to an otherwise intractable problem: it is a way to respect physical constraints, to embed domain knowledge, to bring interpretability to results, and to endow the resulting predictions with quantified uncertainties.

As one specific example, I will present “Lift & Learn”, a method that combines the perspectives of physics-based model reduction and machine learning, in order to derive low-dimensional approximate models that can be used for design and control. Model reduction brings in the physics of the problem, constraining the reduced model predictions to lie on a subspace defined by the governing equations. The machine learning perspective brings the flexibility of data-driven learning – in particular, flexibility in the choice of the physical variables that define the low-dimensional subspace. Combining the two perspectives, the proposed approach identifies a set of transformed physical variables that expose quadratic structure in the physical governing equations and then learns a quadratic ROM from transformed snapshot data. This learning does not require access to or interface with the high-fidelity model implementation, which is often cumbersome for complex engineering codes.

Keywords: *Predictive data science, scientific machine learning, model reduction, surrogate model*