**Nanoinformatics and the big data challenges for the science of small things**

*Amanda S BarnardA*

AData61, CSIRO, Docklands, Australia

**Introduction**

A fundamental aim of nanomaterials research is to identify features of nanomaterials that can be tuned to control how the nanomaterial performs under specific application conditions. Dealing with the complexity of real world conditions while retaining a practical level of abstraction is challenging for a variety of reasons, and is often beyond the capabilities of conventional scientific methods. The combination of computational nanoscience with machine learning (ML) provides a powerful new way of relating nanomaterial structural features with target properties. However, the use of ML in materials and nanoscience is subject to a range of technical challenges that can outweigh the benefits and invalidate the results if not taken into consideration.

**Discussion of Methods**



To understand when to use nanoinformatics one must first recognise the conceptual differences between conventional scientific research tools, such as experiment, theory and simulation, and data-driven scientific research tools such as ML. Each scientific approach has unique advantages and is valuable because it delivers results the others cannot. Each scientific approach also comes with a unique set of disadvantages. ML, for example, was designed to work with large numbers of unique samples, small numbers of features (low dimensionality in the feature space), features of the same physical dimension, with low and consistent variance, that span all possible alternatives. Nanomaterials data sets typically contain a small number of samples relative to fields such as computer vision, with high dimensionality in the feature space, high and inconsistent variance, and are subject to numerous destructive biases. These represent a series of constraints that severely limit the type of models that can be applied and makes the appropriate use of ML in this domain unclear. There are ways of overcoming these challenges, provided one is alerted to them and attention is paid to the data science that precedes the application of learning algorithms.[1]

**Conclusion**

In this presentation we will discuss various technical constraints on data-driven nanomaterials design, and explore the differences between nanomaterials simulation and nanoinformatics that can be leveraged for greater impact, ranging from combinatorial screening to predicting multi-structure/multi-property relationships and inverse design. We will consider issues such as imbalance, bias, model selection and generalizability, and the interpretability of results to guide decision making or further research.

**References**

1. Barnard, A. S., Motevalli, B., Parker, A. J., Fischer, J. M., Feigl, C. A. & Opletal, G. (2019). Nanoinformatics, and the big challenges for the science of small things. Nanoscale, 237, DOI:10.1039/C9NR05912A