"Quantity or Quality? Capitalizing on Small but Rich Materials Data Sets"

Professor Elizabeth A. Holm
Carnegie Mellon University, USA

The process of scientific inquiry involves observing a signal (data) and interpreting it to generate information (knowledge). Artificial intelligence (AI) – a broad term comprising data science, machine learning (ML), neural network computing, computer vision, and other technologies – opens new avenues for extracting information from materials data by uncovering high-dimensional trends that are hard to identify by conventional analysis. Thus, the key to all AI methods is data. However, for many AI applications, the quantity and quality of data required for optimal outcomes is not understood. One solution is to err on the side of data quantity, amassing large, homogeneous data sets. While this may be viable in the social media realm, it is less feasible for physical science and engineering problems where the data is expensive and often heterogeneous. Fortunately, physical data collected by scientists have several advantages: They are selected for their known relevance to the problem, bounded by a physical basis, expertly acquired, and rich in information. Using examples from microstructural characterization, we will survey the factors that should be considered when designing a materials science data set for AI analysis. We will evaluate the relative importance of data size, data type, and data quality. One encouraging observation is that the richness of materials data often enables excellent AI outcomes with surprisingly small data sets.