Finding patterns, correlations, and descriptors in materials-science data using subgroup discovery MARIO BOLEY, BRYAN R. GOLD-SMITH, Fritz Haber Institute of the Max Planck Society, JILLES VREEKEN, Max Planck Institute for Informatics, LUCA M. GHIRINGHELLI, MATTHIAS SCHEFFLER, Fritz Haber Institute of the Max Planck Society, NOVEL MATERI-ALS DISCOVERY LABORATORY (NOMAD) TEAM — Data analytics applied to materials-science data often focuses on the inference of a global prediction model for some physical or chemical property of interest for a given class of materials, such as activation barriers or binding energies. However, the underlying mechanism for some target property could differ for different materials within a large pool of materials-science data. Consequently, a global model fitted to the entire dataset may be difficult to interpret and may well hide or incorrectly describe the actuating physical mechanisms. In these situations, local models would be advantageous to global models. Subgroup discovery (SGD) is presented here as a data-mining approach to find interpretable local models of a target property in materials-science data. We first demonstrate that SGD can identify physically meaningful models that classify the crystal structures of 82 octet binary semiconductors as either rocksalt or zincblende. The SGD framework is subsequently applied to 24 400 configurations of neutral gas-phase gold clusters with 5 to 14 atoms to discern general patterns between geometrical and physicochemical properties.

Bryan Goldsmith
Fritz Haber Institute of the Max Planck Society