Scientific Knowledge Discovery
Combining results into Knowledge— Heterogeneous Catalysis
Draft 1
July 2, 2013
Amber Boehnlein

Preface:

In the field of heterogeneous catalysis, a standard model for analyzing the activity, selectivity, and stability of catalysts has been established and to systematically screen tens of thousands of potential catalytic materials on the computer using atomic-scale simulations.

This has led to some interest in establishing large databases of catalysis relevant properties, which is closely related to the general national ambition of establishing a broader materials genome: http://www.whitehouse.gov/mgi

One thing is to map all materials and their relevant properties, however a very different objective is to enable gaining any knowledge from that data. Where as there are myriads of new initiatives to generate and curate massive amounts of data, there is little focus on storing the data in a form that is actually valuable for searching the data to recognize properties or correlations that one did not consider at the time the data was generated. Achieving this objective would require the development of a data warehouse to store results in a representation that would include an adequate amount of meta-data to be able to recalculate the data automatically. This solves two problems:

1) Computer-generated data which is curated will loose its value exponentially over time, as CPUs/GPUs get faster, and so it will typically have to be redone at short intervals with increased accuracy.

2) A data warehouse is ideal for subsequently applying data mining tools and machine learning algorithms to the data. Whereas the current approach to catalyst design builds on a lot of physical insight and intuition, it is likely that at some point soon, the catalysts and processes we consider will be so complicated that there will be no way to avoid using machine learning algorithms to recognize the relevant correlations in the data.

The design of novel catalysts are likely to require multiple types of machine

learning including ridge regression of linear models where data is correlated, heteroscedastic and undersampled (typical features of electronic structure simulations of materials properties) to using variational Bayes methods, information entropy maximization, and Markov Chain Monte Carlo methods to confidently discover predictive distributions, and neural networks and kernel-based to create fast and reasonably accurate effective potentials for simulating thermal contributions to materials properties.

Issues:

- The ability to trigger regeneration of simulated data to increase the precision and accuracy, either due to improved models or more processing power.
- Development of knowledge representation that including meta data and provenance that can enable federation across the sources of materials data and provision for attribution
- Development of fast algorithms for machine learning that can be applied to a variety of data types, analogous to current efforts in applied mathematics to develop 'fast' math libraries