**Workshop: Big-Data driven Materials Science**

**Location: CECAM-HQ-EPFL, Lausanne, Switzerland**
**Dates: September 11, 2017 to September 13, 2017**
**Organizers: Luca M. Ghiringhelli and Matthias Scheffler**

## Introduction: scope of the workshop

Many, probably most, areas in the basic and applied sciences and engineering are increasingly facing the challenge of dealing with massive amounts of data, nowadays commonly addressed as „big data“. This big-data challenge is not only about storing and processing huge amounts of data, but also, and in particular, it is a chance for new methodology and understanding, as it opens qualitatively new routes for doing research.
The number of possible materials, including organic and inorganic materials, surfaces, interfaces, and nanostructures, as well as hybrids of the mentioned systems, is practically infinite. Less than 200,000 materials are “known” to exist, but only for very few of these “known” materials, the basic properties (elasticity constants, plasticity, piezoelectric tensors, conductivity, etc.) have been determined. When considering 60 commercial elements blended together, there is essentially an infinite number of compounds to be explored.
It is, therefore, highly likely that new materials with superior and up to now simply unknown property profiles exist that could help solving fundamental issues in the fields of energy, mobility, safety, information, and health.

There have not been many breakthroughs, yet, in terms of predicting new materials. The best examples may be the works on electric breakdown [1] and on thermoelectrics [2]. Other works did not use analytics, i.e., a specified dataset was scanned for optimizing a specified quantity [3-9]. The discoveries in these studies were "limited" to finding the best (formerly unknown) material optimizing the given quantity. However, in general, creating data on elastic constants and piezoelectric parameters [10] is also very important – even without the analytics.

For materials science it is already clear that in terms of properties and functions, big data are structured, and in terms of materials properties and functions, the space of all possible materials is sparsely populated. Finding this structure in the big data, e.g., asking for efficient catalysts for methane formation, good thermal barrier coatings, shape memory alloys, or thermoelectric materials for energy harvesting from temperature gradients, may be possible, even if the actuating physical mechanisms of these properties and functions are not yet understood in detail. Novel big-data analytics tools, e.g., based on machine learning and in particular compressed sensing, promise to do so.
Finding structure in big data is just one example of a promising route in big-data-driven materials science. However, at present there is a significant hype associated with the term “big data”. Often, promises are not well founded, because trustful big-data analytics tools and error bars associated with these tools are hardly established. Thus, from a science perspective, certainly for materials science and engineering, “big-data-driven science” is a just emerging field. However, there is hardly any doubt that this field will considerably affect the way science is done in the future.

We identify these two outstanding challenges in the "big-data driven materials science":
1. Developing big-data analytics to find structures and causal relationships in big data of materials that are not recognizable by "naked eye" or standard tools.
2. Assigning error bars or uncertainty tags to the data.

The aim of the workshop was to put in contact the community that develops models and methodologies for the data analytics with the continuously growing part of the materials science community that is applying those models and methodologies to relevant problems in the field.
The purpose of this cross-breeding is on one side to expose the material scientists to novel, state-of-the-art and beyond, methods; on the other side, to stimulate the theoretical data analytics and management community with practical problems whose solution may require further advance in their disciplines.
It is also worth noting that materials science presents a couple of "anomalies", compared to other disciplines where big-data analytics is routinely performed (e.g., social sciences, drug design, meteorology) with respect to the kind of data that are handled.
a) Predictions in materials science need to be unusually accurate, in order to identify the “needle in a haystack”, i.e., say, the top 100 best materials for prescribed target properties out of a pool that contains a practically infinite number of possible chemical and/or structural compounds.
b) Thanks to modern computational methodologies, the behavior of compounds that are difficult to create or handle in the lab (e.g., because poisonous, radioactive or unstable at room conditions) can be calculated. In contrast, in other disciplines, all the data are typically collected before the analysis and it is normally impossible to acquire new data under the same conditions in order to test what was found by the analysis.

These "anomalies" call for "domain tailored" data-analytics techniques that require the cooperation of developers of such techniques with the material scientists.

[1] Kim, C., Pilania, G., & Ramprasad, R. (2016). From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown. Chemistry of Materials **28**, 1304-1311.
[2] Carrete, J., *et al.* (2014). Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. Physical Review X **4**, 11019.
[3] Castelli, I. E., *et al.* (2012) Computational screening of perovskite metal oxides for optimal solar light capture. Energy & Environmental Science **5**, 5814-5819.
[4] Curtarolo, S., *et al.* (2013) The high-throughput highway to computational materials design. Nature Materials **12**, 191-201.
[5] Jain, A. et al. (2011) A high-throughput infrastructure for density functional theory calculations. Computational Materials Science **50**, 2295-2310.
[6] Kang, B., & Ceder, G. (2009). Battery materials for ultrafast charging and discharging. Nature **458**, 190-193.
[7] Potyrailo, R. *et al.* (2011) Combinatorial and high-throughput screening of materials libraries: Review of state of the art. ACS combinatorial science **13**, 579-633.
[8] Rost, C. M., *et al.* (2015) Entropy-stabilized oxides. Nature communications **6**, 8485.
[9] Yang, J., *et al.* (2010). High capacity hydrogen storage materials: attributes for automotive applications and techniques for materials discovery. Chemical Society Reviews **39**, 656-675.
[10] Armiento, R., *et al.* (2014). High-throughput screening of perovskite alloys for piezoelectric performance and thermodynamic stability. Physical Review B, **89**, 134103.

**Snapshots from the workshop**
The workshop was organized in 5 sessions, devoted to 3 main subjects in the emerging field of data-driven materials science: “The central role of the descriptor”, “Extracting information from data”, and “Accuracy and error bars” (the first two topics had a “theory” and “application” session each).

The presented methods spanned established techniques such as kernel ridge regression (e.g., von Lilienfeld, Ramprasad), “re-discovered” techniques such as deep neural network (e.g., Tkatchenko), and emerging techniques such as compressed sensing (e.g., Ghiringhelli). Also techniques from data mining such as subgroup discovery (Vreeken) have recently been adapted to materials-science problems and show promising developments. On the high-throughput side established platforms like AFLOW (Curtarolo) and Materials Project (Winston), as well as newer initiative like NOMAD (Scheffler) and Aiida (Marzari) were presented. From these contributions, it is clear that the potential for the exploration if not the discovery of new materials propelled by these initiatives is high; in some cases, actual new discoveries are already documented (one example for all: the high-entropy materials by Curtarolo and co-workers).
The session on accuracy summarized present challenges in the standardization (in view of possible comparisons and reuse) of calculation parameters, for ground-state and beyond-ground-state methods. This session ideally continued discussions ignited at the “Towards a Common Format for Computational Materials Science Data” CECAM/Psi-k workshop, held in Lausanne in January 2017, with the new contributions aiming both at providing reference, “numerical-errors free” data and at developing machine-learning models for the estimate of the expected error for calculation with not-fully-converged settings.
The extended time allotted for discussions, both at the end of each talk (15 minutes for Q&A) and, in the form of moderated discussion slots (30 minutes) at the end of each session, allowed the participants to ask detailed questions about the methods and results presented in the talks, but gave also room to perspective discussions.

Two discussion topics that involved most of the participants were:
- Possible convergence of the “configurational” vs “chemical space routes” in machine-learned models in materials science. The “configurational route” addresses the learning of, typically, energy and forces by using as input configurations represented in various way (Csanyi, Hirn, Tkatchenko). These methods, quite successful in predicting properties within the set of chemical species used in the training, even for configurations that are not “close” to the training one, are not constructed for extrapolate over new chemical species.

The “chemical space” way (Ramprasad, Pilania, Ghiringhelli, von Lilienfeld) uses as input the compositions of the materials and some minimal information on the structure, when necessary, in order to predict properties of interest (e.g., band gaps, or electric-breakdown field, or classifications such as “being topological or trivial insulator, or metal”). By construction, these methods attempt to predict properties also for unseen chemical species, in combination with those used in the training. However, they have difficulties in distinguishing between polymorphs of the same material.
In several talks (Csanyi, Ghiringhelli, Tkatchenko), it was pointed out the necessity that the two approaches learn from one another and possibly will converge to a unified approach. Tkatchenko and Csanyi have presented already ideas towards this goal, but there is more work to do.
- Interpretable vs non-interpretable models. In a seminal contribution, Joachim Buhmann, a key figure in Information Science, suggested that we, humans, may “surrender” the wish to understand *why* modern machine-learned models (in particular, coming from “deep learning”) work, as long as we can test thoroughly their validity and robustness with respect to noise. Naturally, a physicist and therefore a materials scientist would be reluctant to give up understanding and in facts several recent contributions to the field (e.g., Tkatchenko, Ramprasad, Scheffler) explicitly look for a physical “interpretability” of the model found via machine learning or other data-analytics tools. The idea is that the scientists’ understanding of the physical reason why a certain model works, can lead to its improvement or allow for predictions that are completely outside the training data, i.e., discovery of novel systems (materials classes), rather than just new (better materials – with respect to a certain property or function – but similar to known ones and so they were “just” overlooked). Clearly the two tendencies will coexist and would benefit of each other.

**Conclusion: community needs**
At the beginning of the workshop, Csanyi noted that he felt like he was participating to a “family reunion”, and the family was clearly growing and thriving. Indeed, in November 2015 Scheffler and Ghiringhelli, together with Levchenko, had organized a workshop with similar scope to the present one (“Big Data of Materials Science -- Critical Next Steps” ) with some of the same invited speakers.

It has been indeed astonishing to witness the evolution of the field in these two years, in terms of better understanding and therefore further development of "older" approaches (one example for all, Csanyi's Gaussian Approximation Potentials, leading to structural similarity recognition tools) and the flourishing of new methods and their applications.
It seems therefore highly recommended to envision a series, perhaps with a "natural" biennial frequency, of workshops on the more and more established field of "Data-driven materials science". Most of the classes of techniques used in this field (in particular neural networks and compressed sensing) are well-suited for on-going and future development of HCP infrastructure. The existence of ever growing repository of materials science data, allows also to decouple the intensive data production from the data analysis.
This said, the most important investment for the development of the field is the cross-breeding of data-analytics method developers (computer scientists, but also mathematicians) with "domain experts", i.e., the materials scientists that identify the physical relevance of classes of problems that may benefit from data-analytics approaches.
It is also recommendable that the future generation of materials scientist is exposed to (big-)data analytics early in their studies, in order to be prepared for an approach that it is easily foreseeable as a pillar of near and far future materials-discovery research.

**Invited speakers**
Alexander Tkatchenko (University of Luxembourg, Faculté des Sciences, de la Technologie et de la Communication)
Anatole von Lilienfeld (University of Basel)
Andris Gulans (Humboldt-Universität zu Berlin, Germany)
Donald Winston (Lawrence Berkeley National Laboratory, USA)
Gabor Csanyi (University of Cambridge)
Ghanshyam Pilania (Los Alamos National Laboratory)
Jilles Vreeken (Exploratory Data Analysis, Cluster of Excellence MMCI, Saarland University)
Joachim Buhmann (ETH Zurich, Institute for Machine Learning)
Kristian S. Thygesen (Technical University of Denmark, Lyngby)
Matthew Hirn (Michigan State University, Department of Computational Mathematics, Science & Engineering)
Nicola Marzari (Swiss Federal Institute of Technology Lausanne (EPFL))
Praveen Pankajakshan (Shell Technology Centre,Bangalore, India)
Rampi Ramprasad (University of Connecticut, USA)
Stefan T. Bromley (University of Barcelona, Spain)
Stefan Goedecker (University of Basel)
Stefano Curtarolo (Duke University, Durham, )
Thomas Hammerschmidt (ICAMS, Ruhr-Universität Bochum)

**More information (program, slides):**
https://www.cecam.org/workshop-1437.html