Novel Material Discovery (NOMAD)

Fawzi Mohamed, Emre Ahmetcik, Angelo Ziletti, Luca Ghiringhelli and Matthias Scheffler

Fritz Haber Institute Berlin

NOMAD: Novel Material Discovery

The Novel Materials Discovery (NOMAD) Laboratory maintains the largest *Repository* for input and output files of all important computational materials science codes.

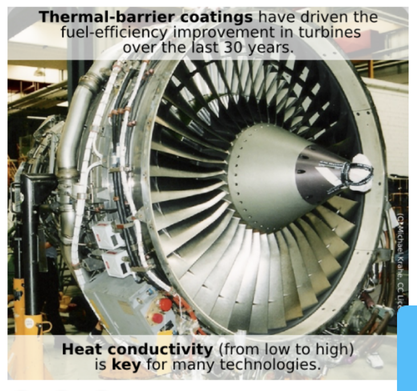
From its open-access data it builds several *Big-Data Services* helping to advance materials science and engineering.

Watch a 3-minute summary on the NOMAD Laboratory CoE (or at YOUKU in China)

http://nomad-coe.eu

NOMAD Scope and Overview

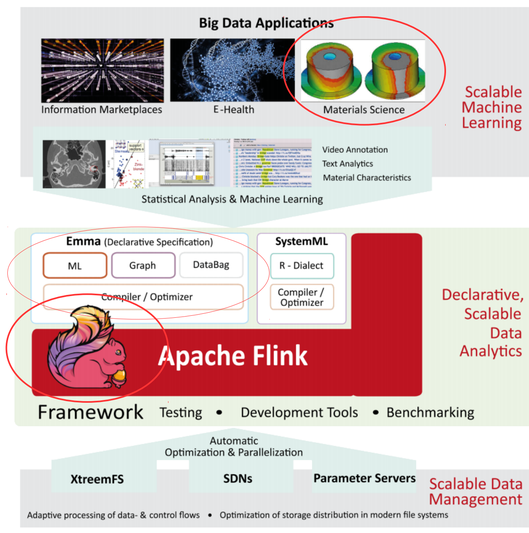
Data is a crucial raw material of the 21st century.

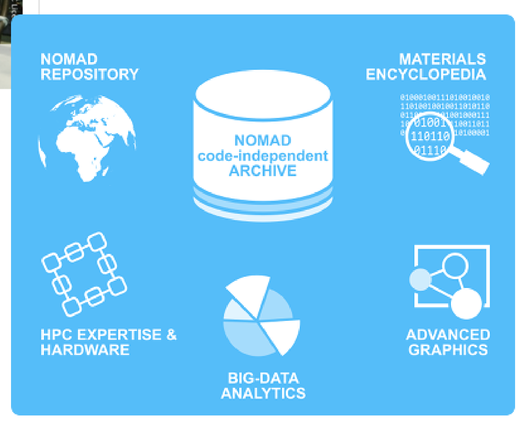


We develop and implement methods that identify correlations and structure in big data of materials. This will enable scientists and engineers to decide which materials are useful for specific applications or which new materials should be the focus of future studies.

The first step to enable this is to **make** the **data available** for analysis.

Then we want to provide interactive tools to find trends and anomalies to discover novel materials.
Here the **BBDC** knowledge in **scaling big data analysis** enriches the expertise of the NOMAD Center of Excellence

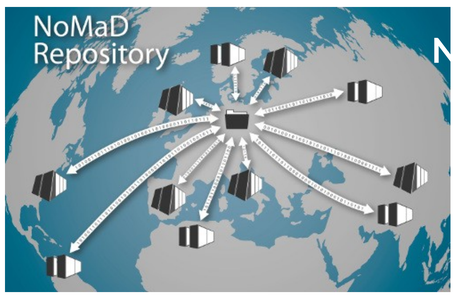




The preparation, synthesis, and characterization of new materials is a complex and costly aspect of materials design. About 200,000 materials are “known” to exist, but the basic properties (e.g., optical gap, elasticity constants, plasticity, piezoelectric tensors, conductivity, etc.) have been determined for very few of them. Considering organic and inorganic materials, surfaces, interfaces, and nanostructures, as well as inorganic/organic hybrids, the number of possible materials is practically infinite. It is therefore highly likely that new materials with superior (but currently unknown) properties exist but still have yet to be identified, which could help address fundamental issues in a number of widespread fields such as energy storage and transformation, mobility, safety, information, and health.

Despite a huge number of possible materials, we note that “the chemical compound space” is sparsely populated when the focus is on selected properties or functions. Our aim is to develop big-data analytics tools that will help to sort all of the available materials data to identify trends and anomalies.

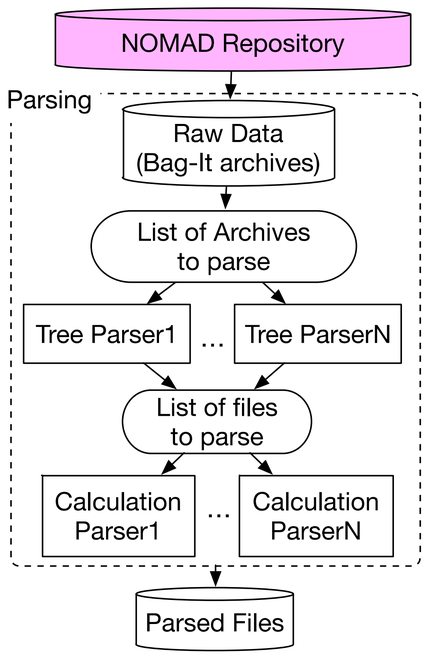
Data: Theoretical Material Science Calculations



http://nomad-repository.eu

- Source of our data
- Established to host organize and share materials data
- Keeps data for at least 10 Years
- Open access and restricted data
- Largest repository
- not limited to a single computer code or closed research group or consortium.

- We use only open access data
- Joint effort by the groups of
  Matthias Scheffler, FHI Berlin and Claudia Draxl, HU Berlin
  Max Planck Computer & Data Facility (MPCDF)



- Parsers **interpret** all calculation data
- **Organize** it according to the metadata structure
- Data not extracted is invisible
- Writing a parser cannot be automatized and requires a person with *scientific knowledge*

- **Parallel** execution
- *Tree Parser* identifies the files
- *Calculation Parser* performs the parsing and generates the parsed files
- Parsing is **pure**: the same version on the same data should give the same result

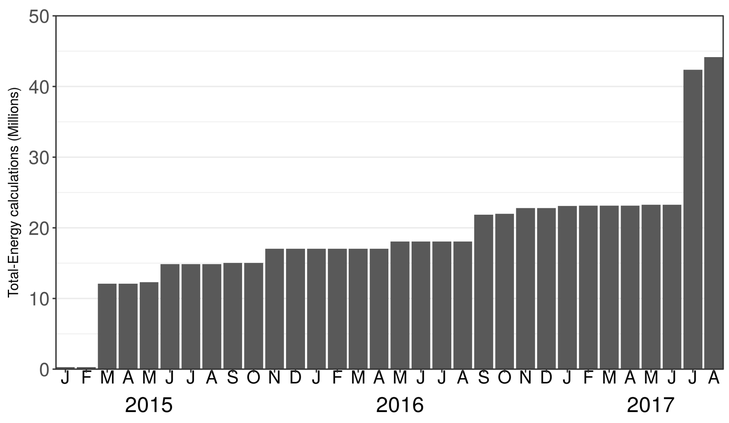
Lots of codes and formats

abinit	crystal	exciting	GPAW	octopus	SIESTA	VASP
ASAP	DFTB+	FHI-aims	GULP	onetep	Smeagol	WIEN2k
ATK	DFTB+	FLEUR	LAMMPS	ORCA	turbomole	
CASTEP	DL_POLY	FPLO	Molcas	Qbox		
cp2k	Dmol3	GAMESS	MOPAC	Quantum Espresso		
CPMD	ELK	Gaussian	NWChem	QUIP /libatoms/GAP		

44,179,006	**39,402,326**	**220,918**	**4,517,016**
Total-Energy Calculations	Bulk Crystals	Surfaces	Molecules/Clusters
40,481,615	**281,135**	**1,936,325**	**91**
Different Geometries	Chemical Compositions	Band Structures	Phonon Calculations

- 9,274 Zip Archives for parsing: 16.5 TB of data (compressed)
- Data extracted with parsing: 5.6 TB of HDF5 files (compressed)

- Data classified using 168 public metadata of the NOMAD Meta Info and 2,360 code-specific metadata
- Number of parsed quantities 871,497,996



BBDC Fruits

More efficient Storage

Thanks to BBDC we began using **Parquet**, a columnar data format, which uses less space and can be efficiently scanned.

Optimzed query language

Querying and accessing the data is crucial. With Emma we have efficient querying, and retrival

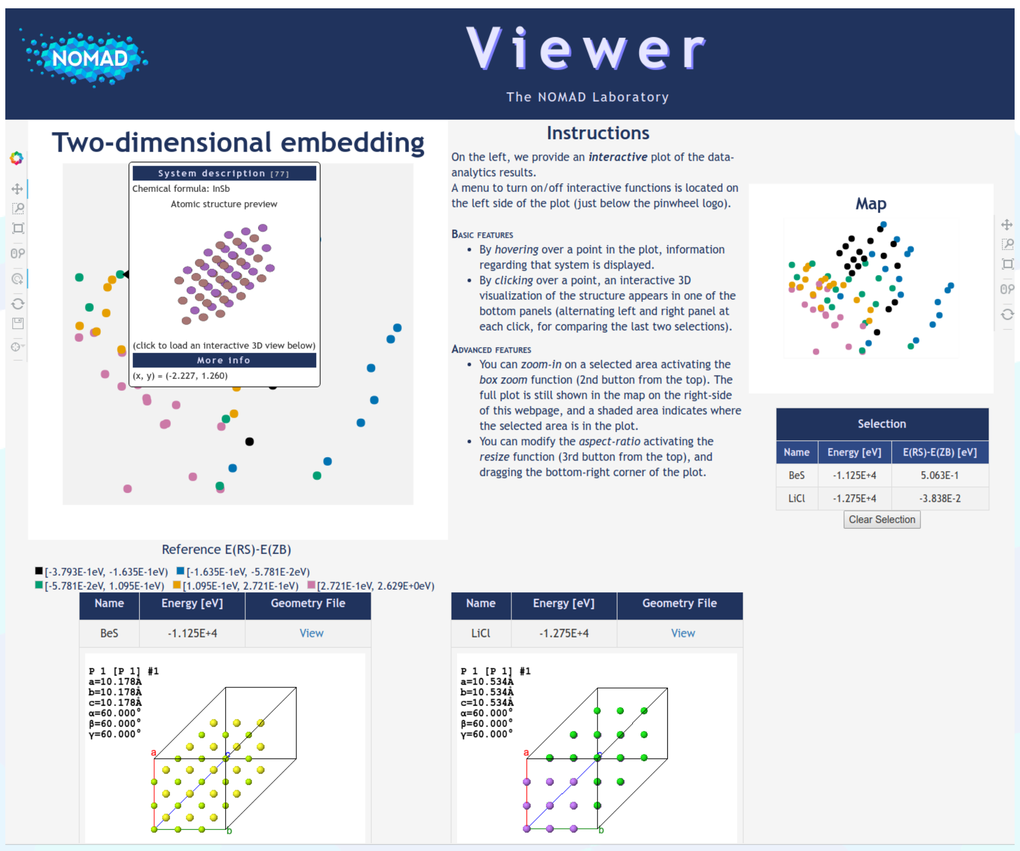
Big data analysis: Similarity & Clustering

Often we query and extract a smaller datasets that then we process in a notebook.

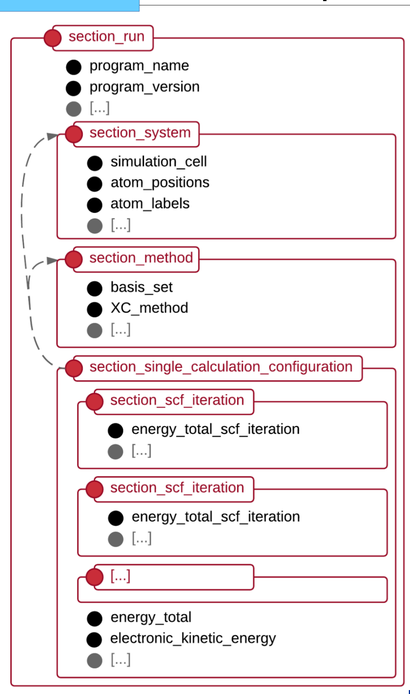
Structure similarity and clustering is an important building block for classification (naming,…) and analysis. Here being able to do across the whole data gives an obvious advantage. For this reason we worked in the past with Shinichi (Machine Learning, TU) at improving our similarity measures, and now also with Alexander (DIMA) to evaluate them efficiently.

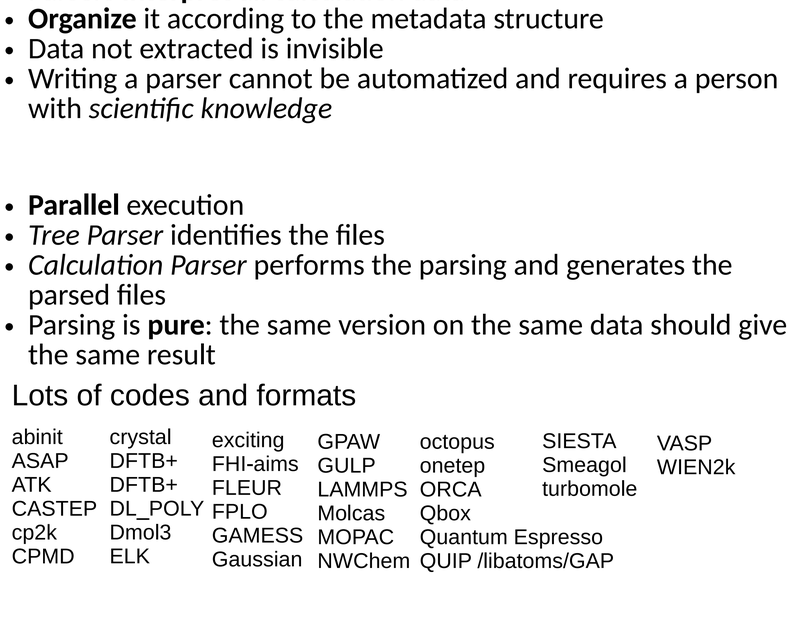
Tool: Two-dimensional Embedding

A web-based implementation (via notebook) of a data-analysis tool for the recognition of the similarity among crystal structures and to highlight a quantity like the energy difference in formation energy among them. The tool gathers the data for the analysis by a query to the NOMAD Archive, that contains several millions of crystal configurations.
The similarity-recognition algorithm, based on the radial distribution function can then projected in two dimensions using several algorithms: i.e. Principal Component Analysis (PCA), and a selection of non-linear embedding methods.



Metadata: our conceptual model





H2020 NOMAD
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 676580. The material presented and views expressed here are the responsibility of the author(s) only. The EU Commission takes no responsibility for any use made of the information set out.