Novel Material Discovery

Fawzi Mohamed, Luca Ghiringhelli

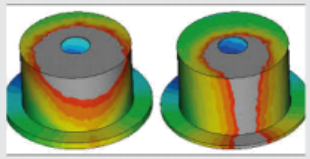
FHI Max-Planck Gesellschaft

© BBDC

Big Data Applications

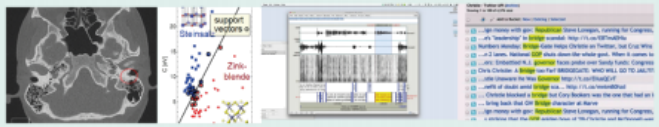
Information Marketplaces

E-Health



Materials Science

Scalable Machine Learning



Video Annotation
Text Analytics
Material Characteristics

Statistical Analysis & Machine Learning

Emma (Declarative Specification)

ML | Graph | DataBag

Compiler / Optimizer

SystemML

R - Dialect

Compiler / Optimizer

Apache Flink

Framework Testing • Development Tools • Benchmarking

Declarative, Scalable Data Analytics

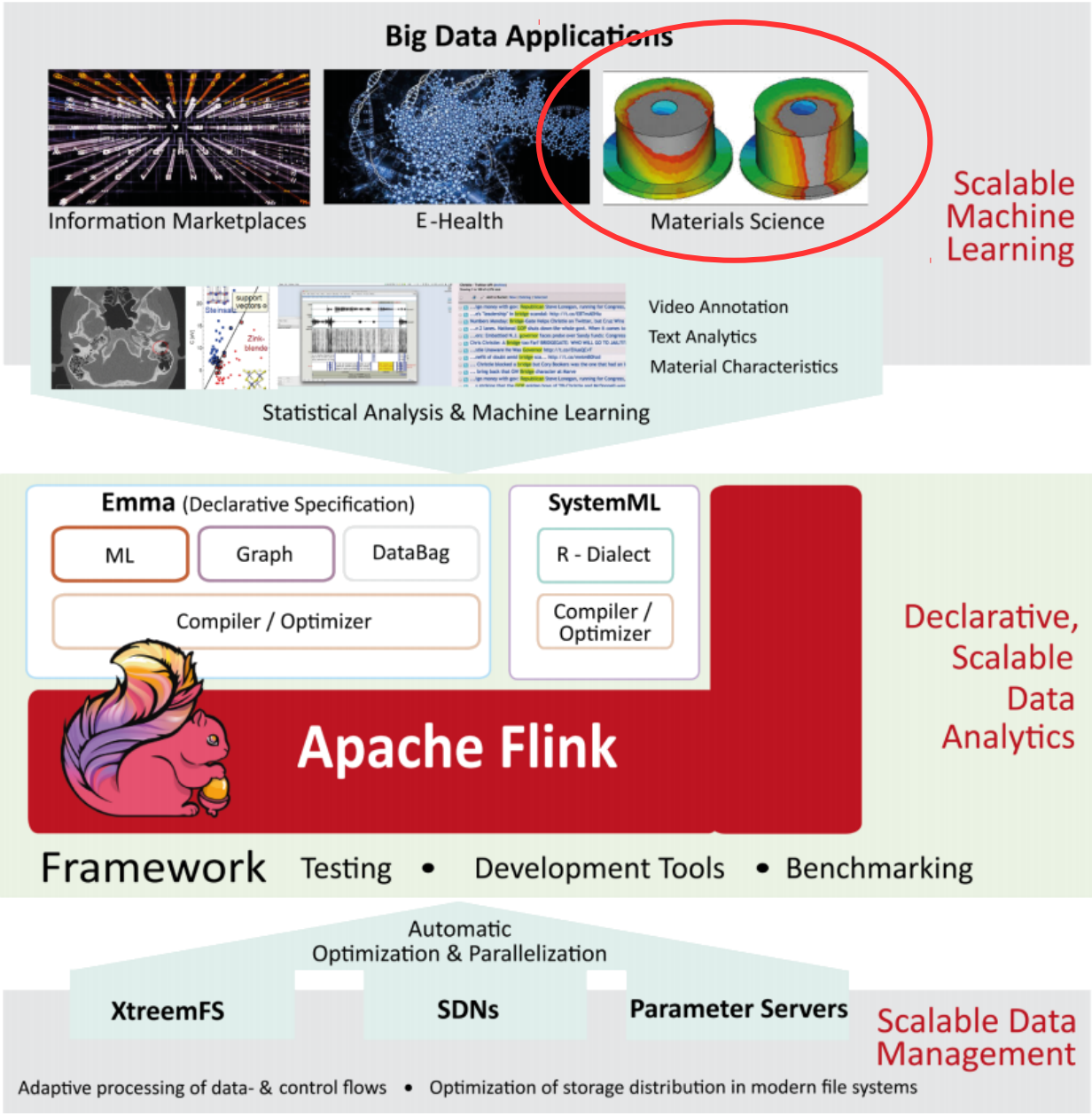
Automatic Optimization & Parallelization

XtreemFS | **SDNs** | **Parameter Servers**

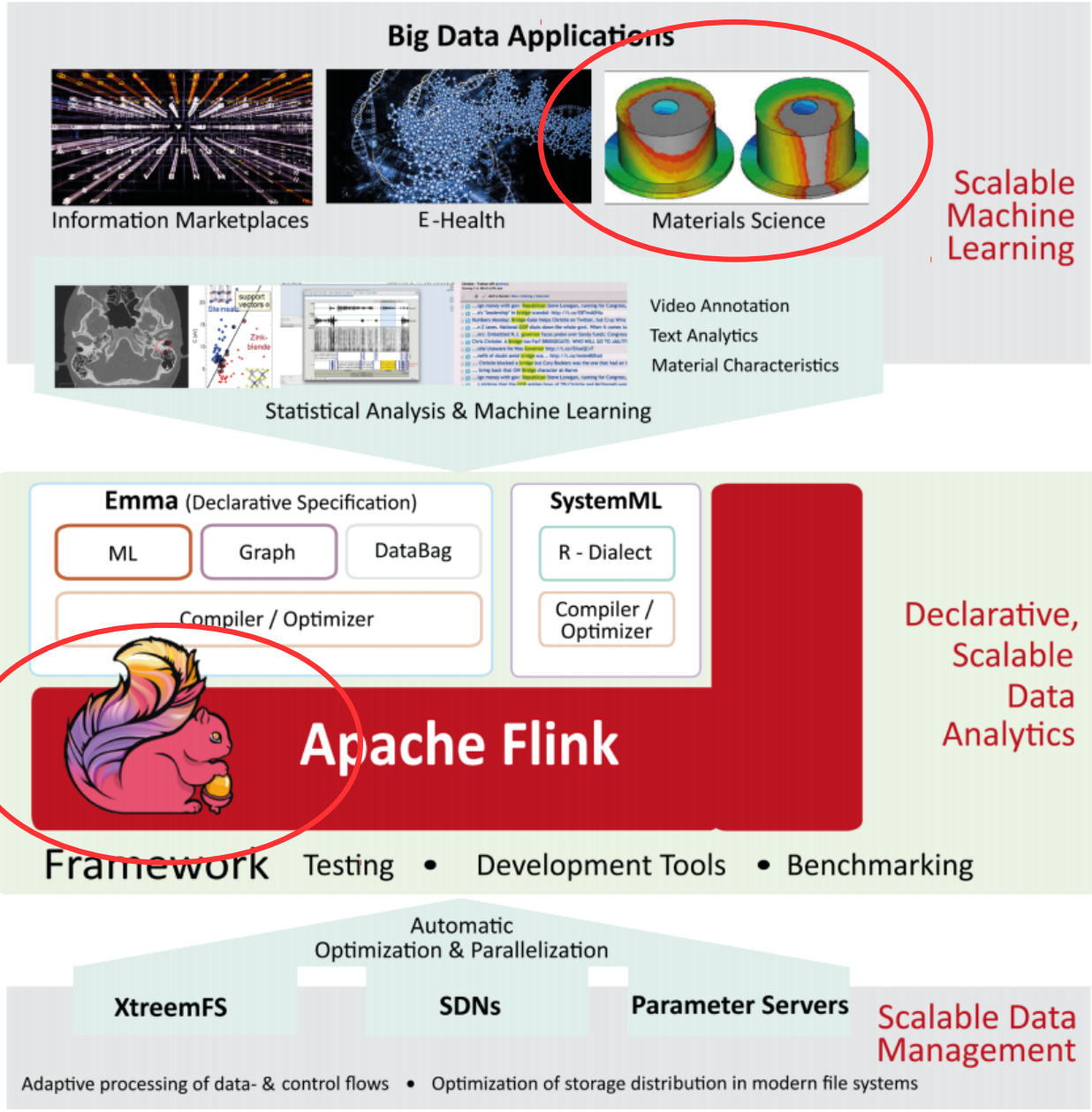
Scalable Data Management

Adaptive processing of data- & control flows • Optimization of storage distribution in modern file systems

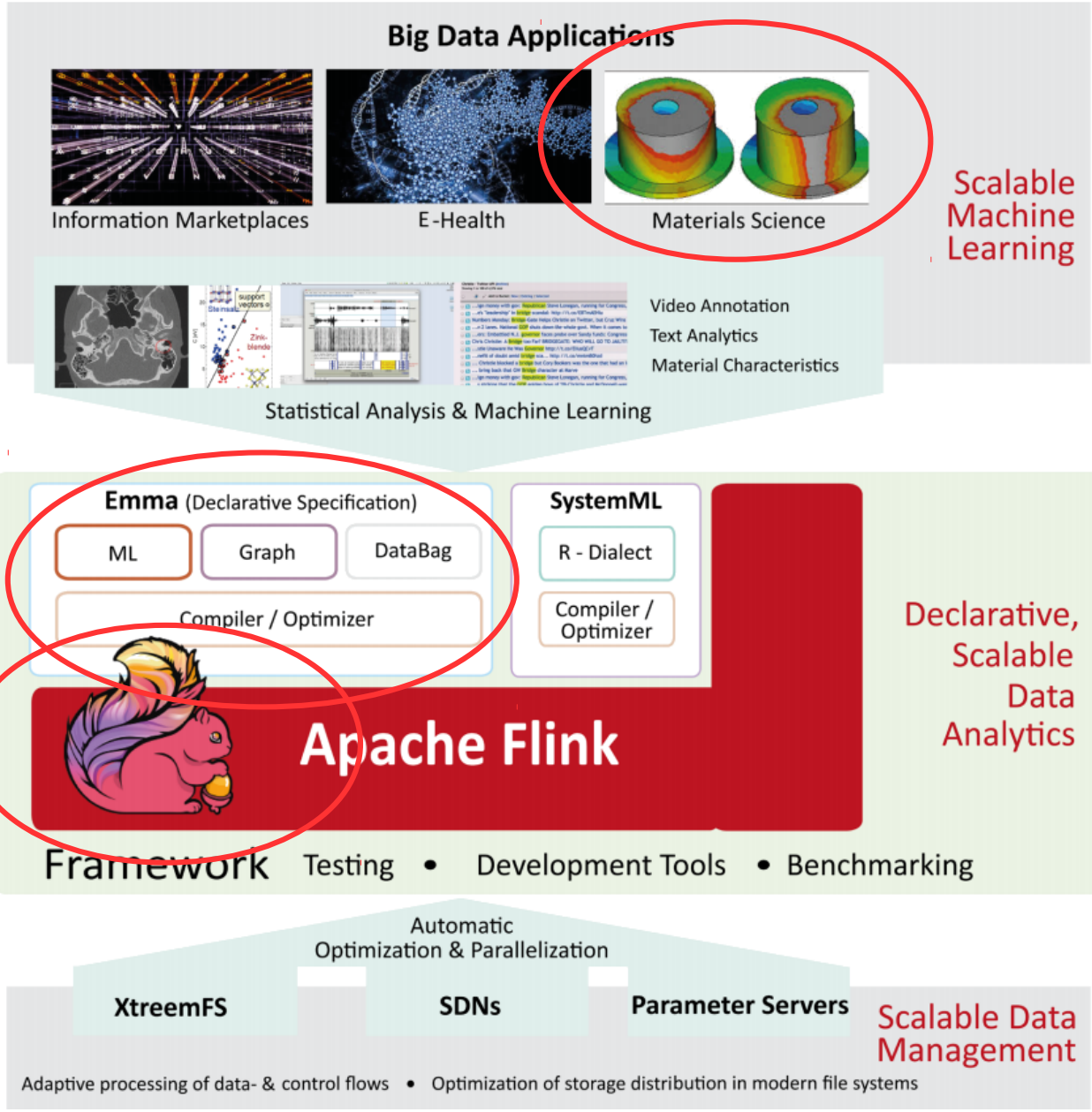
© BBDC


Big Data Applications
Information Marketplaces
E-Health
Materials Science
Scalable Machine Learning
Video Annotation
Text Analytics
Material Characteristics
Statistical Analysis & Machine Learning
Emma (Declarative Specification)
ML
Graph
DataBag
Compiler / Optimizer
SystemML
R - Dialect
Compiler / Optimizer
Apache Flink
Declarative, Scalable Data Analytics
Framework
Testing • Development Tools • Benchmarking
Automatic Optimization & Parallelization
XtreemFS
SDNs
Parameter Servers
Scalable Data Management
Adaptive processing of data- & control flows • Optimization of storage distribution in modern file systems

© BBDC


Big Data Applications
Information Marketplaces
E-Health
Materials Science
Scalable Machine Learning
Video Annotation
Text Analytics
Material Characteristics
Statistical Analysis & Machine Learning
Emma (Declarative Specification)
ML
Graph
DataBag
Compiler / Optimizer
SystemML
R - Dialect
Compiler / Optimizer
Apache Flink
Framework
Testing • Development Tools • Benchmarking
Declarative, Scalable Data Analytics
Automatic Optimization & Parallelization
XtreemFS
SDNs
Parameter Servers
Scalable Data Management
Adaptive processing of data- & control flows • Optimization of storage distribution in modern file systems

© BBDC


Big Data Applications
Information Marketplaces
E-Health
Materials Science
Scalable Machine Learning
Video Annotation
Text Analytics
Material Characteristics
Statistical Analysis & Machine Learning
Emma (Declarative Specification)
ML
Graph
DataBag
Compiler / Optimizer
SystemML
R - Dialect
Compiler / Optimizer
Apache Flink
Framework
Testing • Development Tools • Benchmarking
Declarative, Scalable Data Analytics
Automatic Optimization & Parallelization
XtreemFS
SDNs
Parameter Servers
Scalable Data Management
Adaptive processing of data- & control flows • Optimization of storage distribution in modern file systems

© BBDC

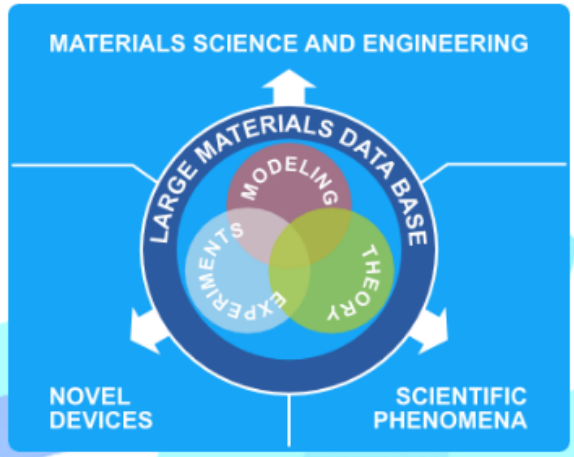
THE NOMAD LABORATORY
A EUROPEAN CENTRE OF EXCELLENCE

HOME PROJECT INDUSTRY OUTREACH TEAM CODES NEWS PRESS KIT CONTACT US

Enter Search...

The Novel Materials Discovery (NOMAD) Laboratory develops a *Materials Encyclopedia* and *Big-Data Analytics* and *Advanced Graphics Tools* for materials science and engineering.

Eight complementary computational materials science groups and four high-performance computing centers form the synergetic core of this Centre of Excellence.



Latest News

Nov 1, 2016
Industry Interview - BIOVIA

Oct 21, 2016
NOMAD Year 1 Report

Oct 7, 2016
NOMAD Year 1 Meeting with Scientific Advisory Committee

Sep 24, 2016
7th School and Workshop on Time-Dependent Density-Functional Theory (TDDFT)

Sep 19, 2016
Industry Interview - Lockheed Martin

http://nomad-coe.eu

MATERIALS ENCYCLOPEDIA | BIG-DATA ANALYTICS | ADVANCED GRAPHICS | HPC INFRASTRUCTURE | OUTREACH

© BBDC

Ziele:

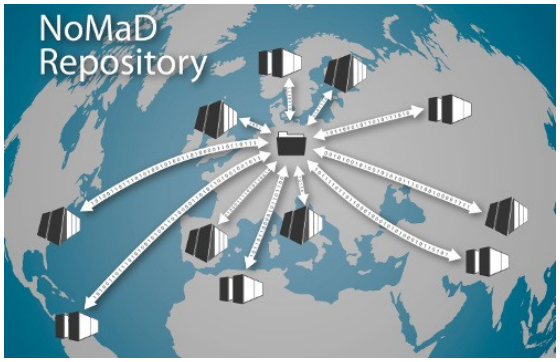
- Korrelation und Struktur in Materialien Big Data

- Materialien für spezifische Applikationen
- Materialien zu erforschen in zukünftige Studien

- Den ersten Schritt ist die **Daten Verfügbar** für die Analyse **zu machen**

- Dann interaktive Anwendungen um Trends und Unregelmäßigkeit zu erforschen.
- BBDC Kenntnisse in Skalierbare Big Data Analyse bereichern die Expertise der NOMAD CoE

© BBDC



http://nomad-repository.eu

Claudia Draxl
HUB

Matthias Scheffle
FHI

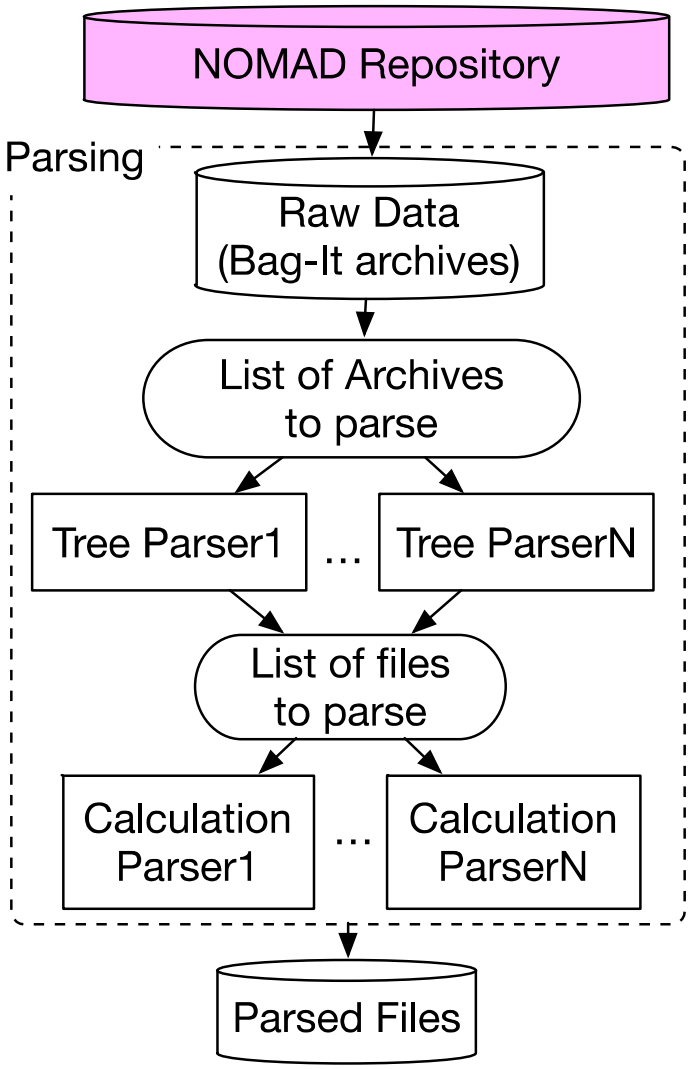
- Source of our data
- Established to host organize and share materials data
- Keeps data for at least 10 Years
- Open access and restricted data
  - >5.2M entries, >5.4M open access
- We use only open access data
- Joint effort by the groups of
  - Matthias Scheffler, FHI Berlin
  - Claudia Draxl, HU Berlin
  - Max Planck Computer & Data Facility (MPCDF), Garching, headed by Stefan Heinzel.

Jungho Shin
HUB

Lorenzo Pardin
HUB

Thomas Zastro
MPCDF

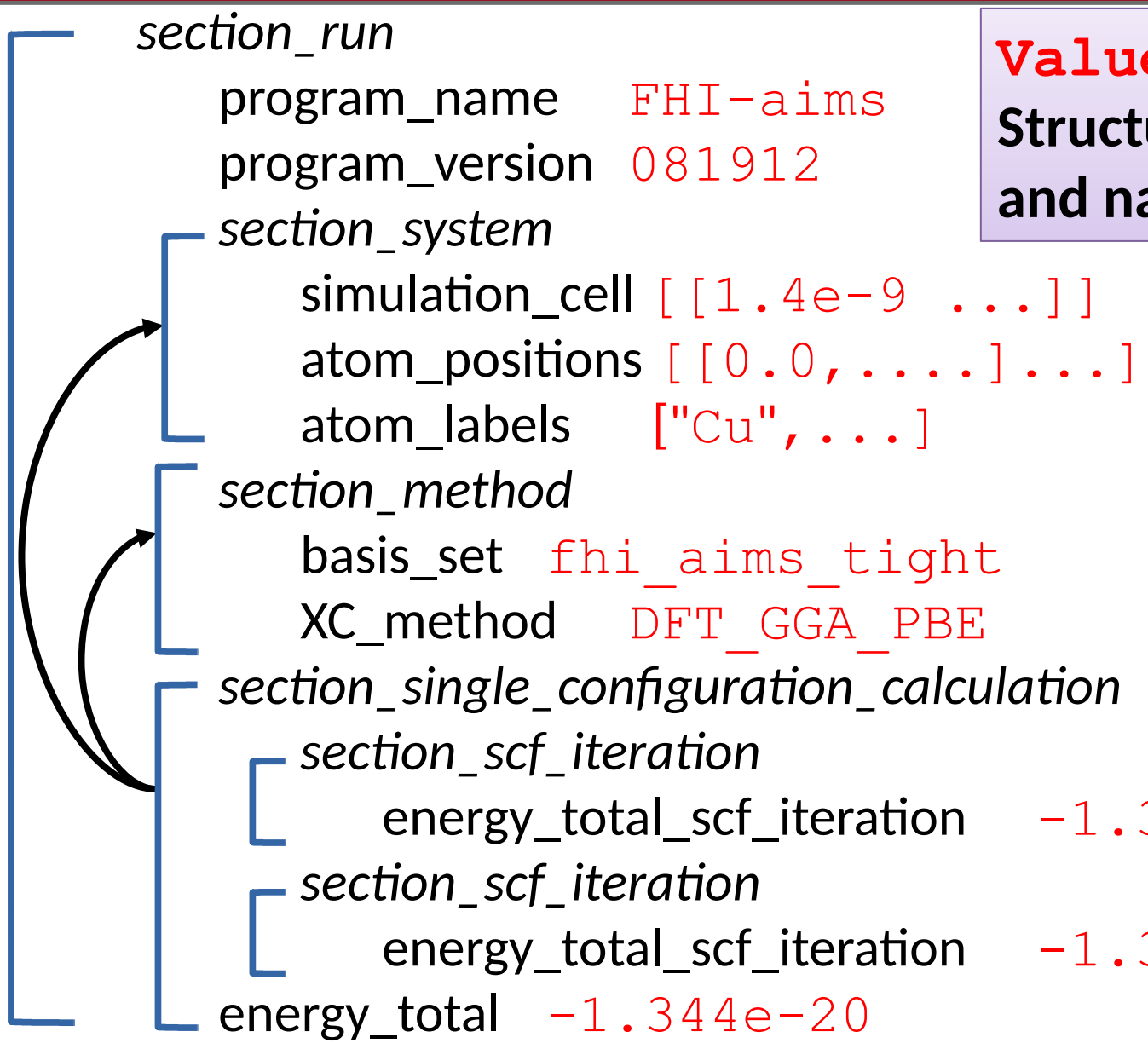
© BBDC



- 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

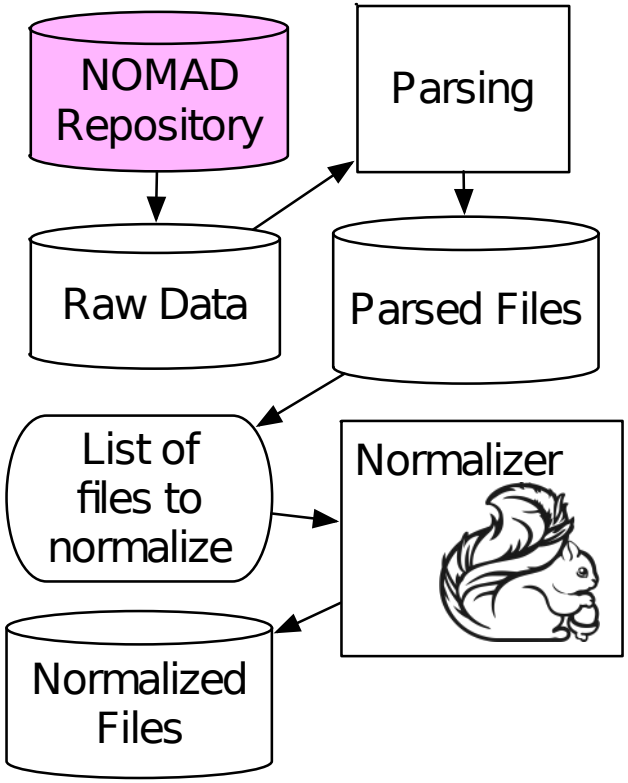
© BBDC



Values: Data
Structure
and names: Metadata

SI Units:
- lengths: m
- energies: J
- ...

© BBDC



Standardization

- Parsers standardize the data,
- avoid the loss of information

Normalization

- The work packages define **derived quantities** (normalized representations,…)
- These can be generally useful for analysis or visualization

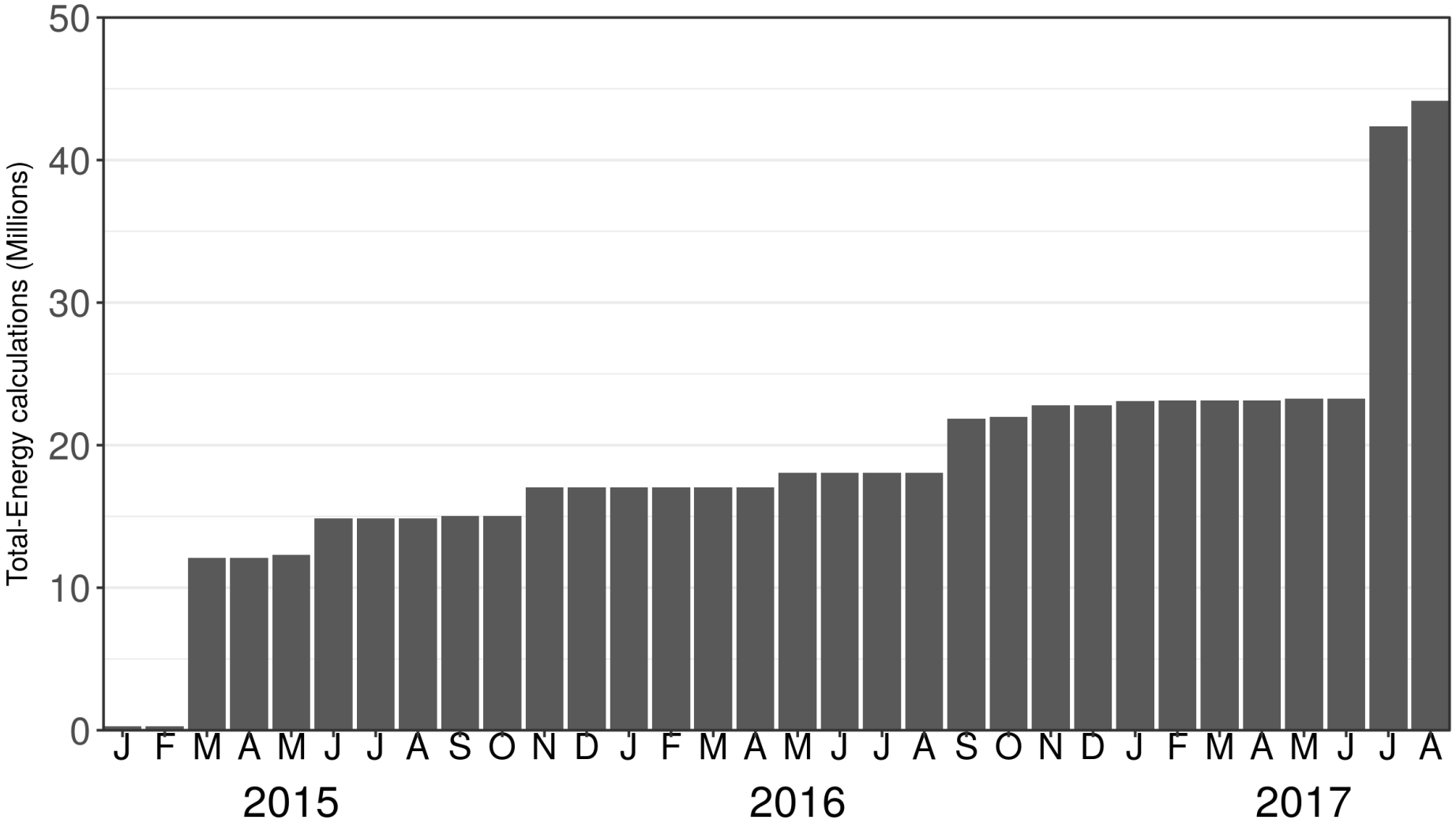
- Normalization is an infrastructure to apply automatically some transformations and store their result along with the parsed data

© BBDC

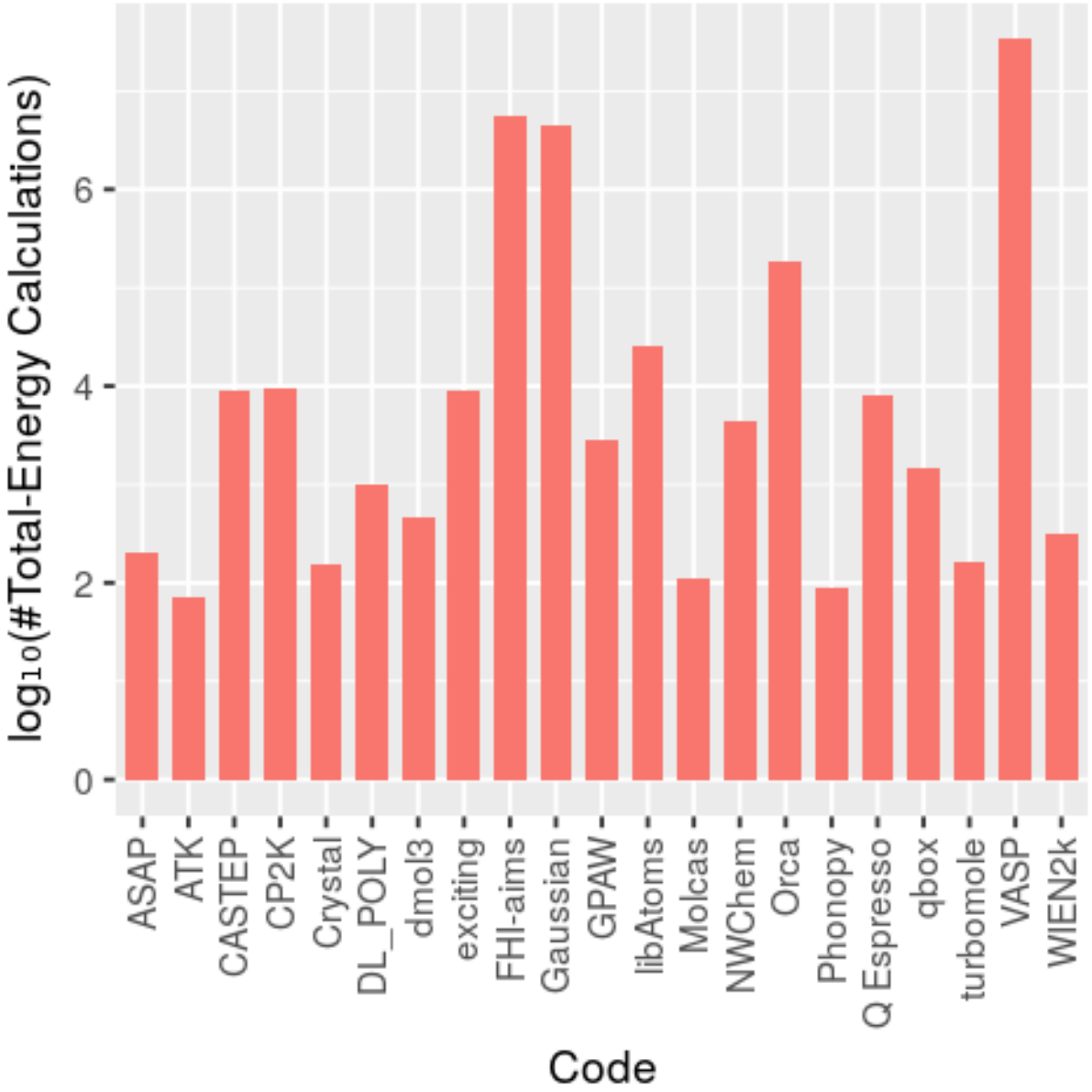
44,179,006	39,402,326	220,918	4,517,016
Total-Energy Calculations	Bulk Crystals	Surfaces	Molecules/Clusters
40,481,615	281,135	1,936,325	91
Different Geometries	Chemical Compositions	Band Structures	Phonon Calculations

- 9,274 Zip Archives for parsing: 16.5 TB of data (compressed)
- Data extracted with parsing:
  - 5.6 TB of HDF5 files, ~0.5TB Parquet
- Data classified using 168 public metadata of the NOMAD Meta Info and 2,360 code-specific metadata
- Number of parsed quantities 871,497,996

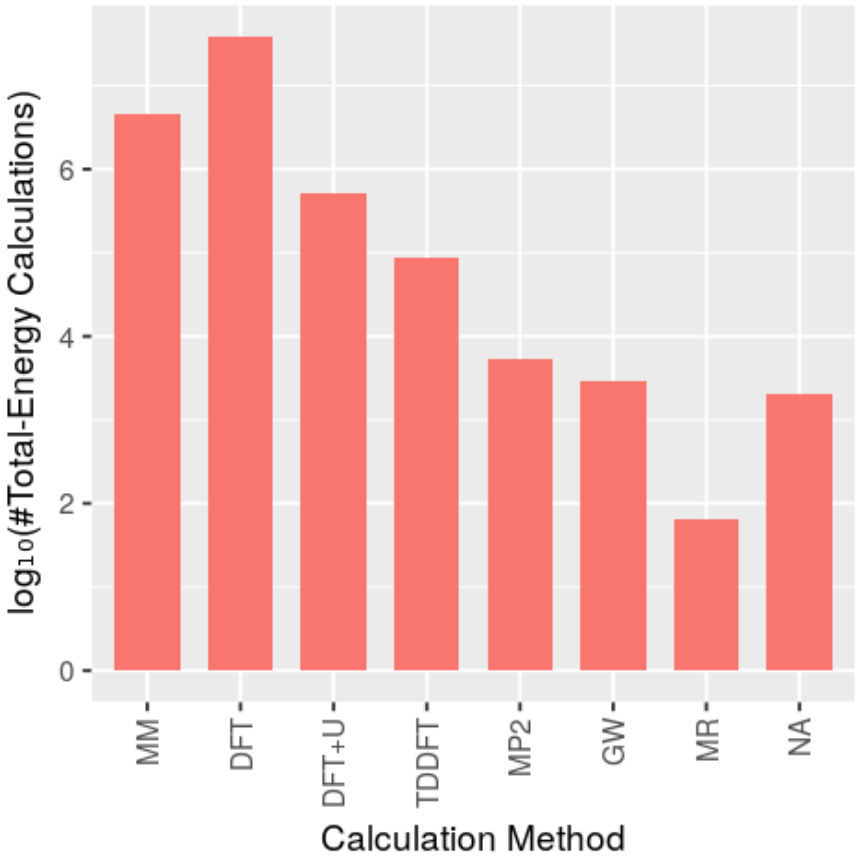
© BBDC

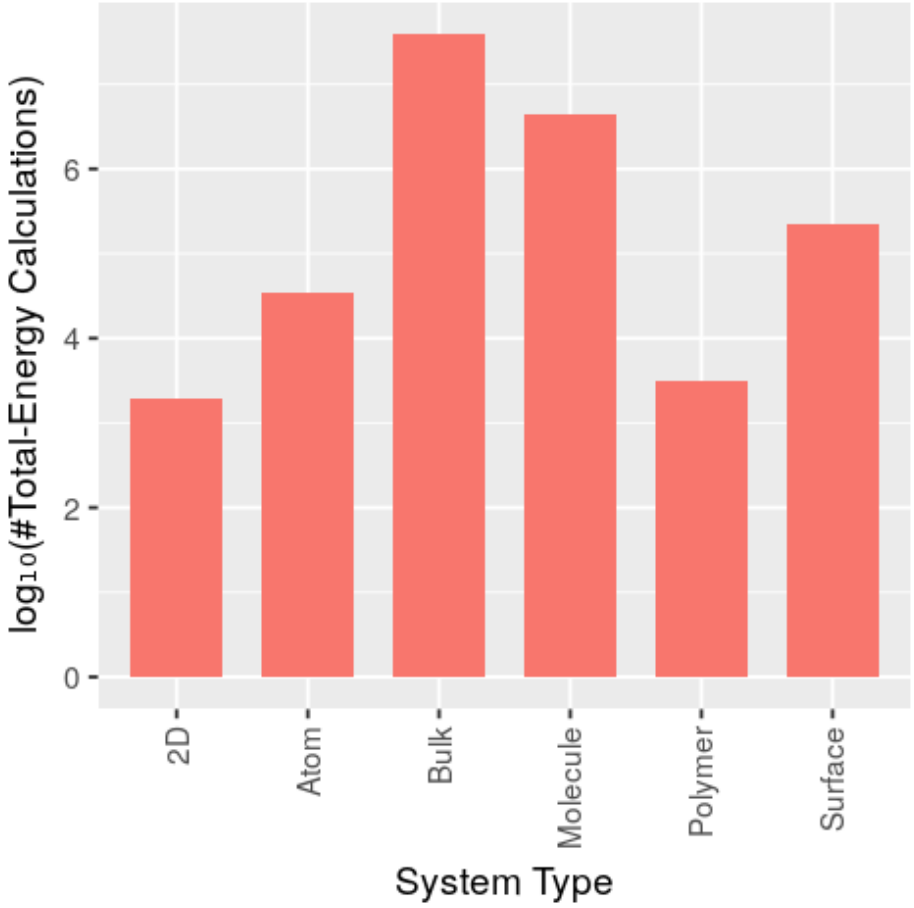

Total-Energy calculations (Millions)
50
40
30
20
10
0
J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A
2015
2016
2017

© BBDC


log10(#Total-Energy Calculations)
0
2
4
6
ASAP
ATK
CASTEP
CP2K
Crystal
DL_POLY
dmol3
exciting
FHI-aims
Gaussian
GPAW
libAtoms
Molcas
NWChem
Orca
Phonopy
Q Espresso
qbox
turbomole
VASP
WIEN2k
Code

© BBDC


log10(#Total-Energy Calculations)
0
2
4
6
MM
DFT
DFT+U
TDDFT
MP2
GW
MR
NA
Calculation Method


log10(#Total-Energy Calculations)
0
2
4
6
2D
Atom
Bulk
Molecule
Polymer
Surface
System Type

© BBDC

Kristallstrukturbestimmung

- Materialeigenschaften sind abhängig von Kristallstruktur
- Verschiedene Kristallstrukturen können geringe Energieunterschiede haben
- Man kann Kristallstrukturen Patentieren
- Kristallstruktur ist wichtig wenn mehrere Materialien kombiniert werden

- Wir versuchen klein dimensional Beschreibungen (descriptors) die robust sind zu finden (extrapolation)

© BBDC

NOMAD analytics toolkit

Tutorial example on Crystal prediction I: The case of octet-binary zincblende-vs.-rocksalt semiconductor

developed by Angelo Ziletti, Ankit Kariryaa, Emre Ahmetcik, Fawzi Mohamed, Luca Ghiringhelli, and Matthias Scheffler. [Last update September 13, 2

Introduction and motivation

Machine learning method: Compressed sensing (LASSO performed on a tailor-made feature space, followed by L0-regularized minimization).
Click here for more info on the LASSO+L0 method.

Reference: "Big Data of Materials Science: Critical Role of the Descriptor"
L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. 114, 105503 (2015) (Click here for the free access pdf)

Instructions

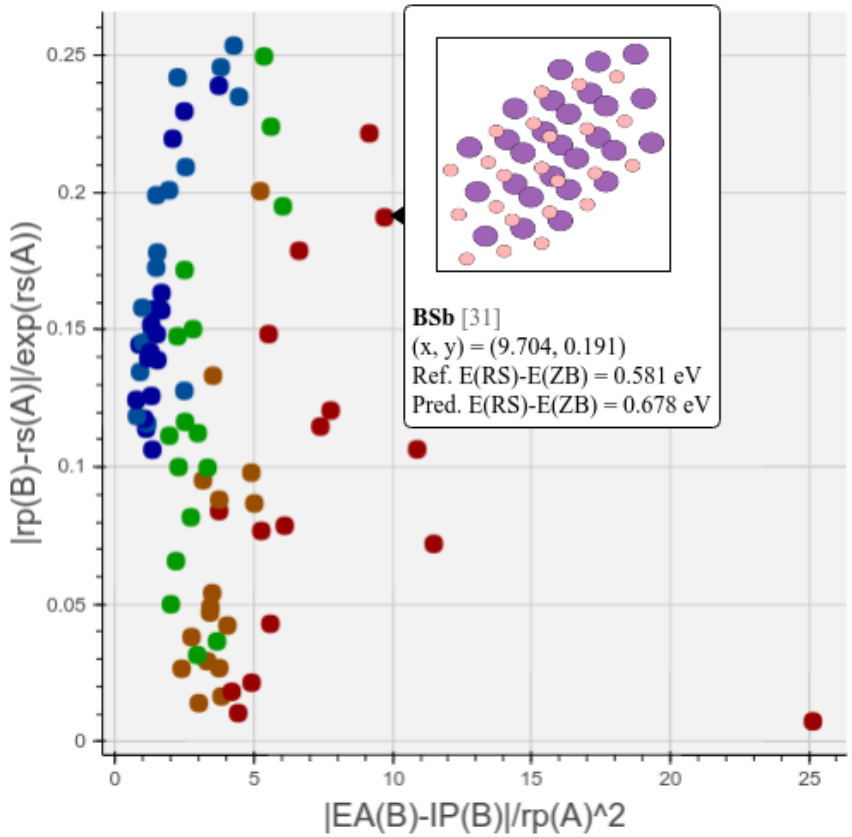
Primary features (hover the mouse pointer over the feature names to see a full description):

- [x] IP^{A}
- [x] EA^{B}
- [x] E^{A}_{LUMO}
- [x] r_s^{B}
- [x] r_d^{A}
- [ ] d^{BB}
- [ ] E^{AA}_{b}
- [ ] Z^{B}
- [ ] n^{A}_{period}
- [ ] r_{π}
- [ ] E^{AB}_{b}
- [x] IP^{B}
- [x] E^{A}_{HOMO}
- [x] E^{B}_{LUMO}
- [x] r_p^{A}
- [x] r_d^{B}
- [ ] ΔE^{AA}_{HL}
- [ ] E^{BB}_{b}
- [ ] Z^{A}_{val}
- [ ] n^{B}_{period}
- [ ] d^{AB}
- [x] EA^{A}
- [x] E^{B}_{HOMO}
- [x] r_s^{A}
- [x] r_p^{B}
- [ ] d^{AA}
- [ ] ΔE^{BB}_{HL}
- [ ] Z^{A}
- [ ] Z^{B}_{val}
- [ ] r_{σ}
- [ ] ΔE^{AB}_{HL}

Allowed operations:

Given features x and y, apply these operations:

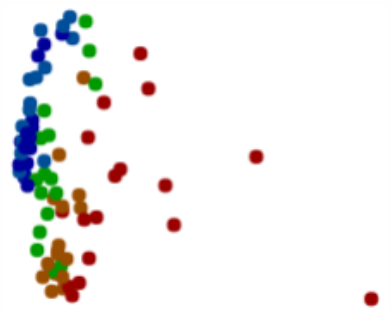
LASSO+L0 structure map



Reference E(RS)-E(ZB)

[-0.379eV, -0.164eV) [-0.164eV, -0.058eV) [-0.058eV, 0.11eV) [0.11eV, 0.272eV) [0.272eV, 2.629eV)

Map

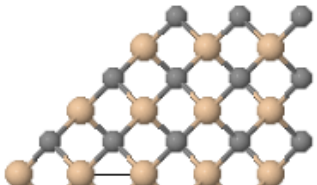


Selection		
Name	Reference E(RS)-E(ZB) [eV]	Predicted E(RS)-E(ZB) [eV]
BN	1.712	1.618
SiC	0.669	0.562

Clear Selection

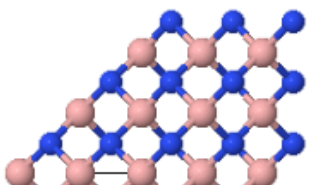
Name	Energy [eV]	Geometry File
SiC	-8903.9092	View

P 1 [P 1] #1
a=3.062Å
b=3.062Å
c=3.062Å
α=60.000°
β=60.000°
γ=60.000°



Name	Energy [eV]	Geometry File
BN	-2153.491	View

P 1 [P 1] #1
a=2.534Å
b=2.534Å
c=2.534Å
α=60.000°
β=60.000°
γ=60.000°



Andere Projekte

- Subgroup discovery
  - „spezielle“ Untergruppen Identifizieren
- Structural Similarity
  - Aktuelles Projekt
  - Kollaboration mit Shinichi Nakajima über ein Similaritätmaß basierend auf die paarweise Korrelation

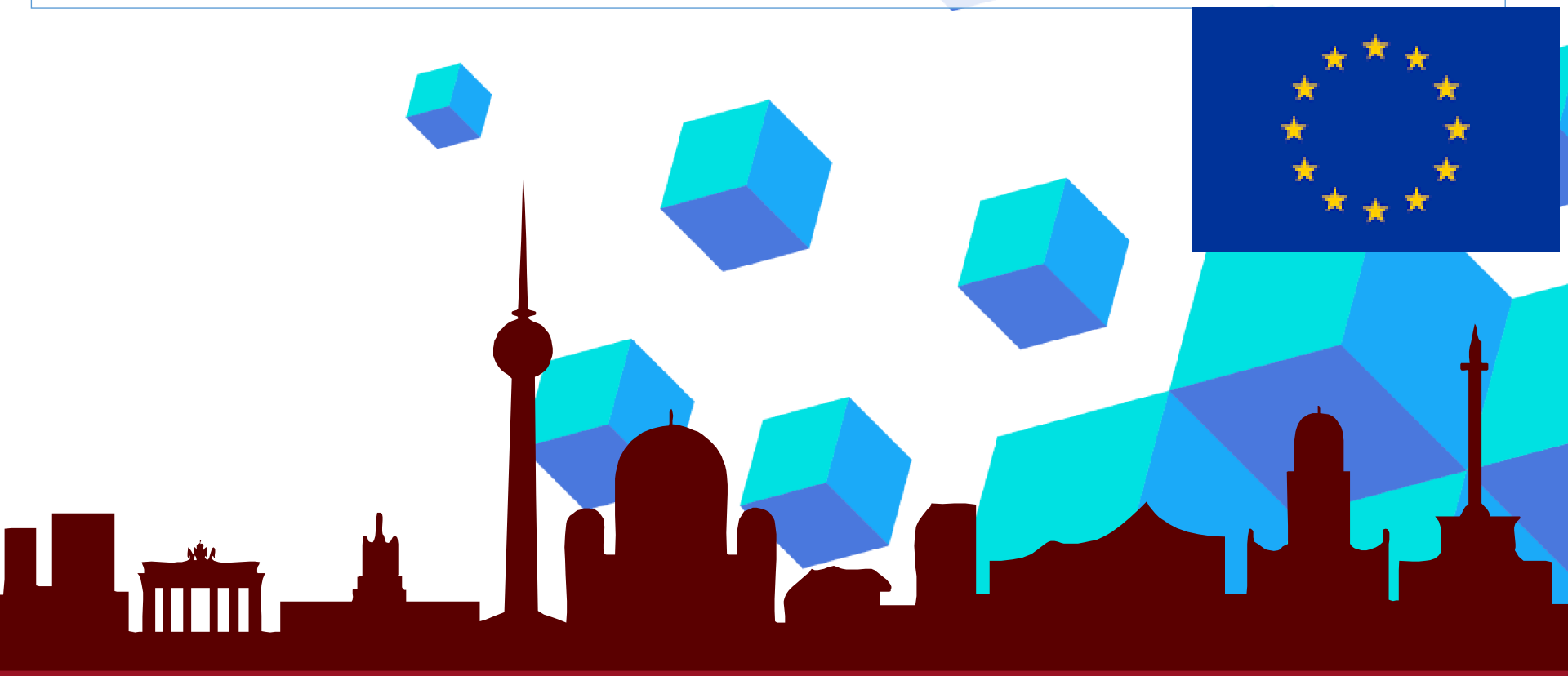
© BBDC

Danksagung

- Angelo Ziletti
- Ankit Kariryaa
- Emre Ahmetcik
- Matthias Scheffler
- Den Ganzen NOMAD Team
- BBDC

© BBDC

NOMAD has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676580.



© BBDC