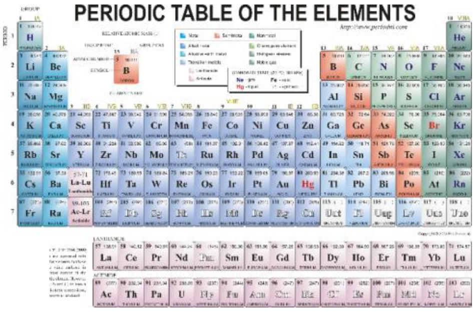
Big-Data Analytics for Materials Science: Concepts, Challenges, and Hype

Matthias Scheffler(*)
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin; http://th.fhi-berlin.mpg.de/

Dmitri Mendeleev
(1834-1907)

From *the periodic table of the elements* to *a chart (a map) of materials*: Organize materials according to their properties and functions.

- ○ figure of merit of thermoelectrics (as function of *T*)
- ○ turn-over frequency of catalytic materials (as function of *T* and *p*)
- ○ efficiency of photovoltaic systems
- ○ etc.


PERIODIC TABLE OF THE ELEMENTS

(*) Work performed in collaboration with **Luca Ghiringhelli, Jan Vybiral, Claudia Draxl, et al.**

Big-Data Analytics for Materials Science: Concepts, Challenges, and Hype

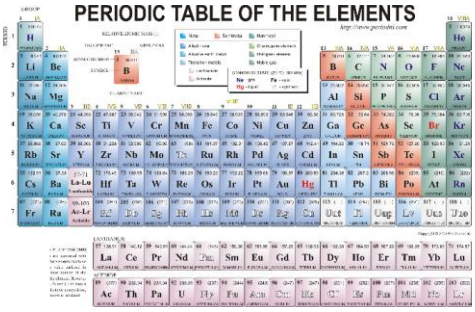
Matthias Scheffler(*)
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin; http://th.fhi-berlin.mpg.de/

Dmitri Mendeleev
(1834-1907)

From the periodic table of the elements to a chart (a map) of materials: Organize materials according to their properties and functions.

○ figure of merit of thermoelectrics (as function of T)
○ turn-over frequency of catalytic materials (as function of T and p)
○ efficiency of photovoltaic systems
○ etc.

(*) Work performed in collaboration with Luca Ghiringhelli, Jan Vybiral, Claudia Draxl, et al.


PERIODIC TABLE OF THE ELEMENTS

Materials Genome Initiative for Global Competiveness

To help business discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative. The invention of silicon circuits and lithium ion batteries made computers and iPods and iPads possible, but it took years to get those technologies from the drawing boards to the market place. We can do it faster.

President Obama
Carnegie Mellon University, June 2011

Materials Genome Initiative
for Global Competitiveness
June 2011

"twice as fast,
at a fraction of the cost"

Materials Genome Initiative for Global Competiveness

To help business discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative. The invention of silicon circuits and lithium ion batteries made computers and iPods and iPads possible, but it took years to get those technologies from the drawing boards to the market place. We can do it faster.

President Obama
Carnegie Mellon University, June 2011

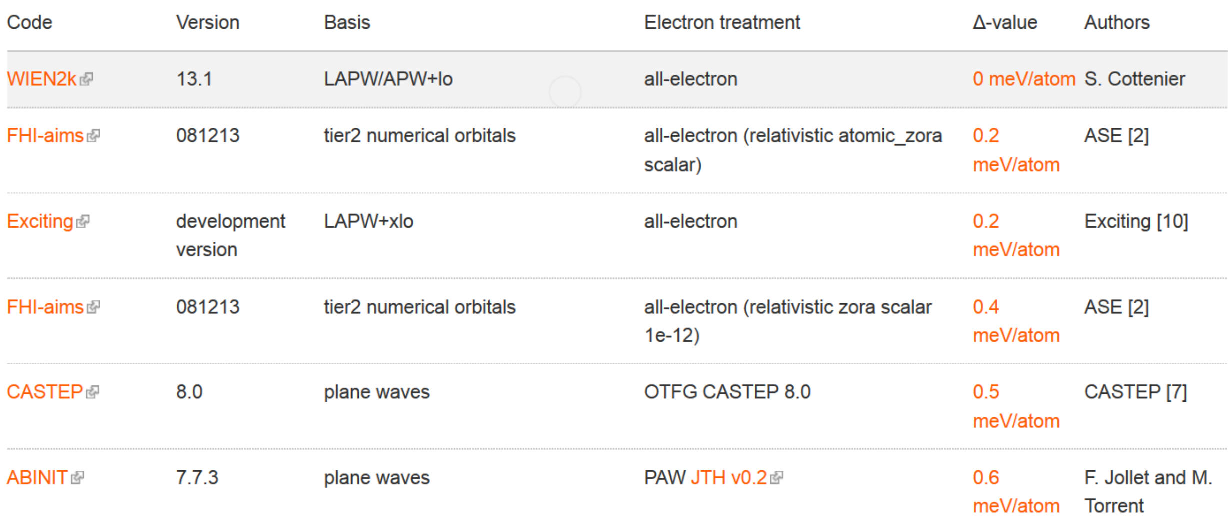
Compute or measure the basic properties („genes") of many (ten thousand) materials and disseminate that information to the materials community to enable rapid searches and design.

"twice as fast,
at a fraction of the cost"

What is “Computational Materials Science”

what is meant by “first-principles (*ab initio*) calculations”

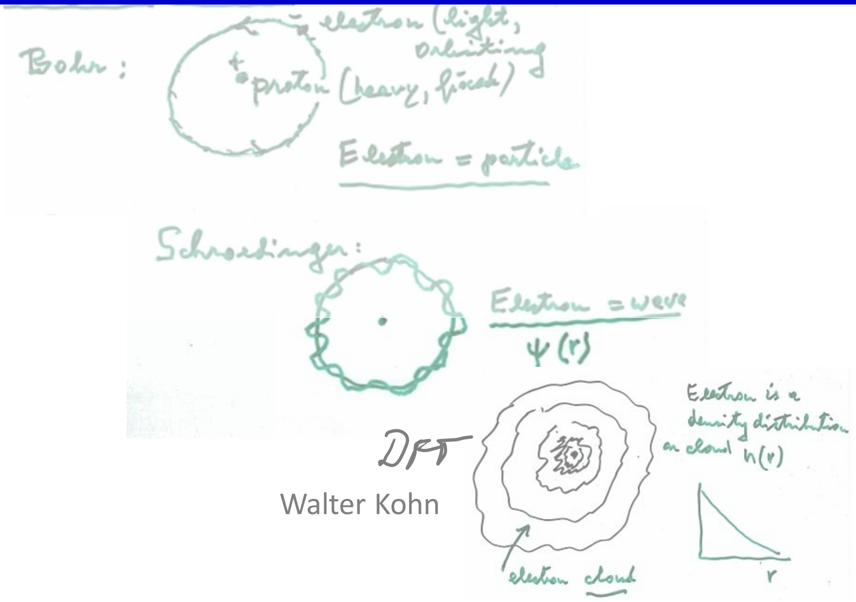
- accuracy of materials-science codes:
  - basis sets,
  - relativity,
  - pseudopotentials,
  - other numerical approximations (verification)
- accuracy of the exchange-correlation functional (validation)



Code	Version	Basis	Electron treatment	Δ-value	Authors
WIEN2k	13.1	LAPW/APW+lo	all-electron	0 meV/atom	S. Cottenier
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic atomic_zora scalar)	0.2 meV/atom	ASE [2]
Exciting	development version	LAPW+xlo	all-electron	0.2 meV/atom	Exciting [10]
FHI-aims	081213	tier2 numerical orbitals	all-electron (relativistic zora scalar 1e-12)	0.4 meV/atom	ASE [2]
CASTEP	8.0	plane waves	OTFG CASTEP 8.0	0.5 meV/atom	CASTEP [7]
ABINIT	7.7.3	plane waves	PAW JTH v0.2	0.6 meV/atom	F. Jollet and M. Torrent

K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, Crit. Rev. Solid State Mater. Sci. 39, 1-24 (2014); https://molmod.ugent.be/deltacodesdft. Reference code: WIEN2k

The Kohn-Sham Ansatz of Density-Functional Theory



Walter Kohn

The Kohn-Sham Ansatz of Density-Functional Theory

- Kohn-Sham (1965): Replace the original many-body problem by an independent electron problem **that can be solved!**

$$E_v[n] = T_s[n] + \int v(\mathbf{r})\, n(\mathbf{r})\, d^3\mathbf{r} + E^{\text{Hartree}}[n] + E^{\text{xc}}[n]$$

- With $T_s\,[n]$ the kinetic energy functional of independent electrons, and $E^{\text{xc}}[n]$ the unknown functional.
- The challenge is to find useful, approximate xc functionals.

Walter Kohn

The Kohn-Sham Ansatz of Density-Functional Theory

- Kohn-Sham (1965): Replace the original many-body problem by an independent electron problem **that can be solved!**

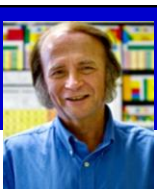
$$E_v[n] = T_s[n] + \int v(\mathbf{r})\, n(\mathbf{r})\, d^3\mathbf{r} + E^{\text{Hartree}}[n] + E^{\text{xc}}[n]$$

Approximate xc functionals have been very successful

but there are problems

- for certain bonding situations (vdW, hydrogen bonding, certain covalent bonds)
- for highly correlated situations, and
- for excited states.

Perdew’s Dream: Jacob’s Ladder in Density-Functional Theory



The exchange-correlation functional

accuracy ↑

	our favorite	
5	unoccupied $\psi_i(\mathbf{r})$,	EX + cRPA, as given by ACFD
4	occupied $\psi_i(\mathbf{r})$,	hybrids (B3LYP, PBE0, HSE, …)
3	τ $(\mathbf{r})$,	meta-GGA (e.g., TPSS)
2	$\nabla n(\mathbf{r})$,	Generalized Gradient Approximation
1	$n(\mathbf{r})$,	Local-Density Approximation

$\tau(\mathbf{r})$: Kohn-Sham kinetic-energy density

EX: exact exchange: $E_x = -\frac{1}{2}\sum^{occ}\iint d\mathbf{r}\, d\mathbf{r}' \frac{\psi_n^*(\mathbf{r})\psi_m(\mathbf{r})\psi_m^*(\mathbf{r}')\psi_n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}$

cRPA : random-phase approximation for correlation

ACFD : adiabatic connection fluctuation dissipation theorem

Bohm, Pines (1953); Gell-Mann, Brueckner (1957);
Gunnarsson, Lundqvist (1975, 1976); Langreth, Perdew (1977);
X. Ren, P. Rinke, C. Joas, and M. S., Invited Review, Mater. Sci. 47, 21 (2012)

Perdew’s Dream: Jacob’s Ladder in Density-Functional Theory

The exchange-correlation functional

accuracy ↑

our favorite

5 unoccupied $\psi_i(\mathbf{r})$,
4 occupied $\psi_i(\mathbf{r})$,
3 τ $(\mathbf{r})$,
2 $\nabla n(\mathbf{r})$,
1 $n(\mathbf{r})$,

Functionals of level 1 and 2 suffer from severe self-interaction errors.
Functionals of level 1, 2, 3, & 4 are lacking the long-range vdW tails.

With “Level 5 plus” validation (error estimation) is becoming possible.

Bohm, Pines (1953); Gell-Mann, Brueckner (1957);
Gunnarsson, Lundqvist (1975, 1976); Langreth, Perdew (1977);
X. Ren, P. Rinke, C. Joas, and M. S., Invited Review, Mater. Sci. 47, 21 (2012)

Test Sets for Materials Science and Engineering?

Chemists have shown the way. For small and light molecules they developed test sets: G2, NHTBH38, HTBH38, S22, S66 ...

We need a materials test set! We can now do renormalized second-order perturbation theory (similar to CCSD) and even full CI (*) – for certain systems.

Comparison with experiment is very important as well (adsorption energies of molecules, *e.g*. by microcalometry). However, theory-theory comparison is better defined.

(*) G. H. Booth, A. J. W. Thom, and A. Alavi, J. Chem. Phys. 131, 054106 (2009).
G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, Nature 493, 365 (2013).

Test set for materials science and engineering

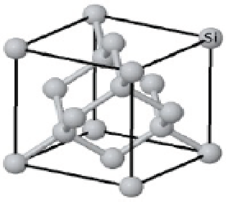
MSE
TEST SET

7 elements and 12 binaries
with cubic structure
(for the start)

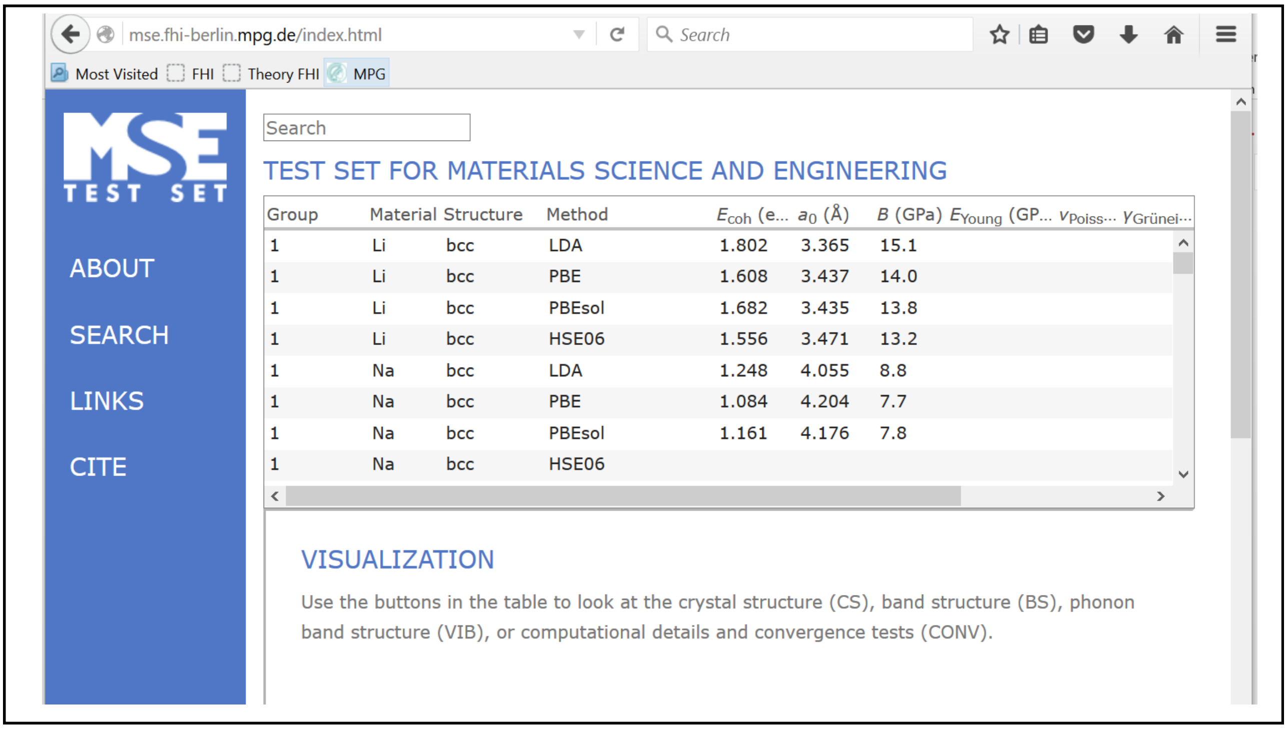
Light main group elements

H							He
Li	Be	B	C	N	O	F	Ne
Na	Mg	Al	Si	P	S	Cl	Ar
K	Ca	Ga	Ge	As	Se	Br	Kr
Rb	Sr	In	Sn	Sb	Te	I	Xe
Cs	Ba						

Ne, Ar, Al (fcc); Li, Na (bcc); C, Si (diamond);
LiH, LiF, LiCl, NaF, NaCl, MgO, MgS (rocksalt);
BeS, BP, AlP, SiC, BN (zincblende)



- **MSE properties**: cohesive, electronic, elastic and vibrational
- **Representative** for cubic metals, semiconductors, and insulators
- **Numerically accurate reference values from theory**, incl. MP2, RPA, CCSD(T)



Search

ABOUT

SEARCH

LINKS

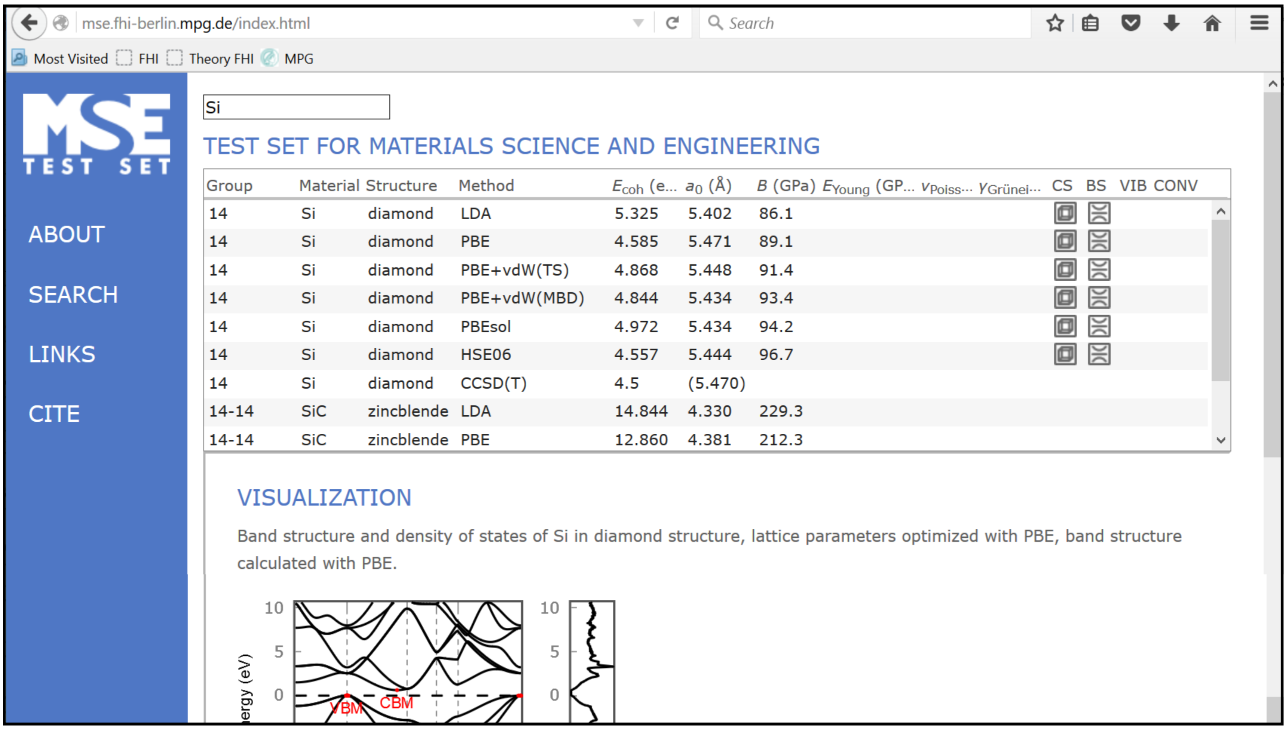
CITE

TEST SET FOR MATERIALS SCIENCE AND ENGINEERING

Group	Material	Structure	Method	E_{coh} (e...	a_0 (Å)	B (GPa)	E_{Young} (GP...	ν_{Poiss}...	$\gamma_{Grünei}$...
1	Li	bcc	LDA	1.802	3.365	15.1			
1	Li	bcc	PBE	1.608	3.437	14.0			
1	Li	bcc	PBEsol	1.682	3.435	13.8			
1	Li	bcc	HSE06	1.556	3.471	13.2			
1	Na	bcc	LDA	1.248	4.055	8.8			
1	Na	bcc	PBE	1.084	4.204	7.7			
1	Na	bcc	PBEsol	1.161	4.176	7.8			
1	Na	bcc	HSE06						

VISUALIZATION

Use the buttons in the table to look at the crystal structure (CS), band structure (BS), phonon band structure (VIB), or computational details and convergence tests (CONV).



Si

ABOUT

SEARCH

LINKS

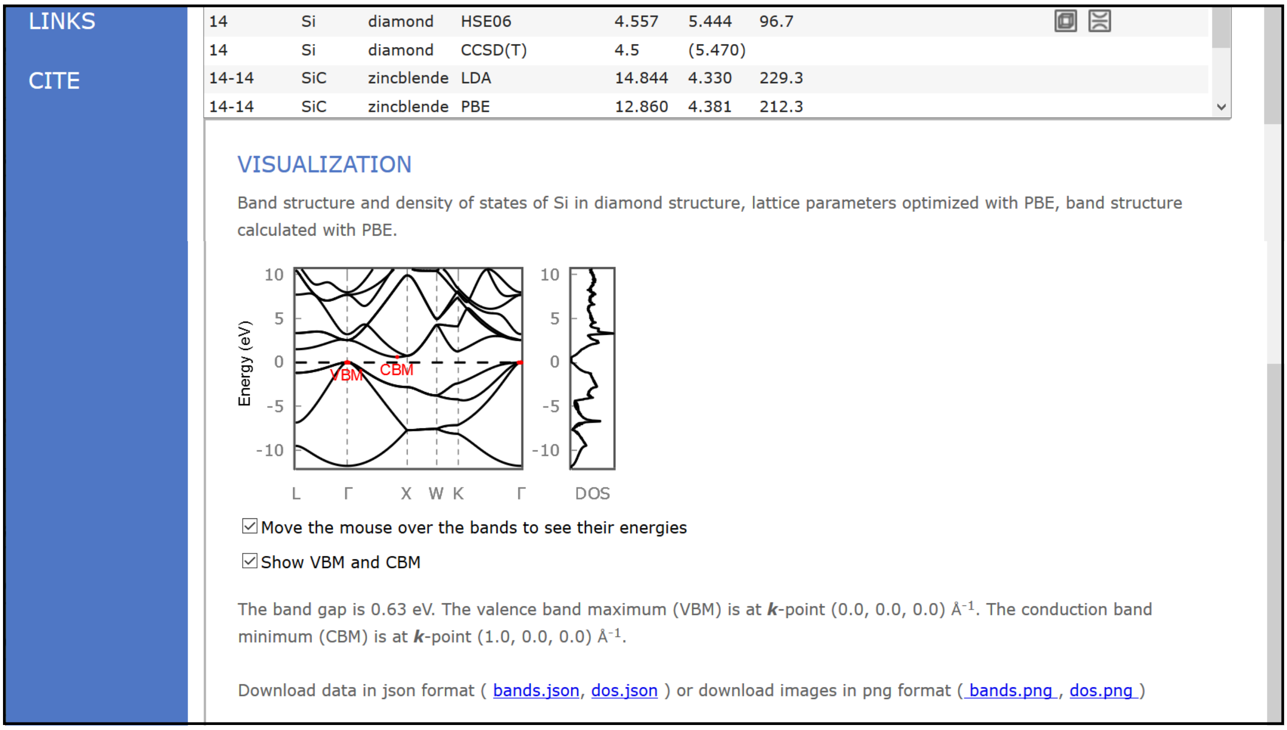
CITE

TEST SET FOR MATERIALS SCIENCE AND ENGINEERING

Group	Material	Structure	Method	E_{coh} (e...	a_0 (Å)	B (GPa)	E_{Young} (GP...	ν_{Poiss}...	$\gamma_{Grünei}$...	CS	BS	VIB	CONV
14	Si	diamond	LDA	5.325	5.402	86.1				▣	⧗		
14	Si	diamond	PBE	4.585	5.471	89.1				▣	⧗		
14	Si	diamond	PBE+vdW(TS)	4.868	5.448	91.4				▣	⧗		
14	Si	diamond	PBE+vdW(MBD)	4.844	5.434	93.4				▣	⧗		
14	Si	diamond	PBEsol	4.972	5.434	94.2				▣	⧗		
14	Si	diamond	HSE06	4.557	5.444	96.7				▣	⧗		
14	Si	diamond	CCSD(T)	4.5	(5.470)								
14-14	SiC	zincblende	LDA	14.844	4.330	229.3							
14-14	SiC	zincblende	PBE	12.860	4.381	212.3							

VISUALIZATION

Band structure and density of states of Si in diamond structure, lattice parameters optimized with PBE, band structure calculated with PBE.



14	Si	diamond	HSE06	4.557	5.444	96.7
14	Si	diamond	CCSD(T)	4.5	(5.470)	
14-14	SiC	zincblende	LDA	14.844	4.330	229.3
14-14	SiC	zincblende	PBE	12.860	4.381	212.3

VISUALIZATION

Band structure and density of states of Si in diamond structure, lattice parameters optimized with PBE, band structure calculated with PBE.

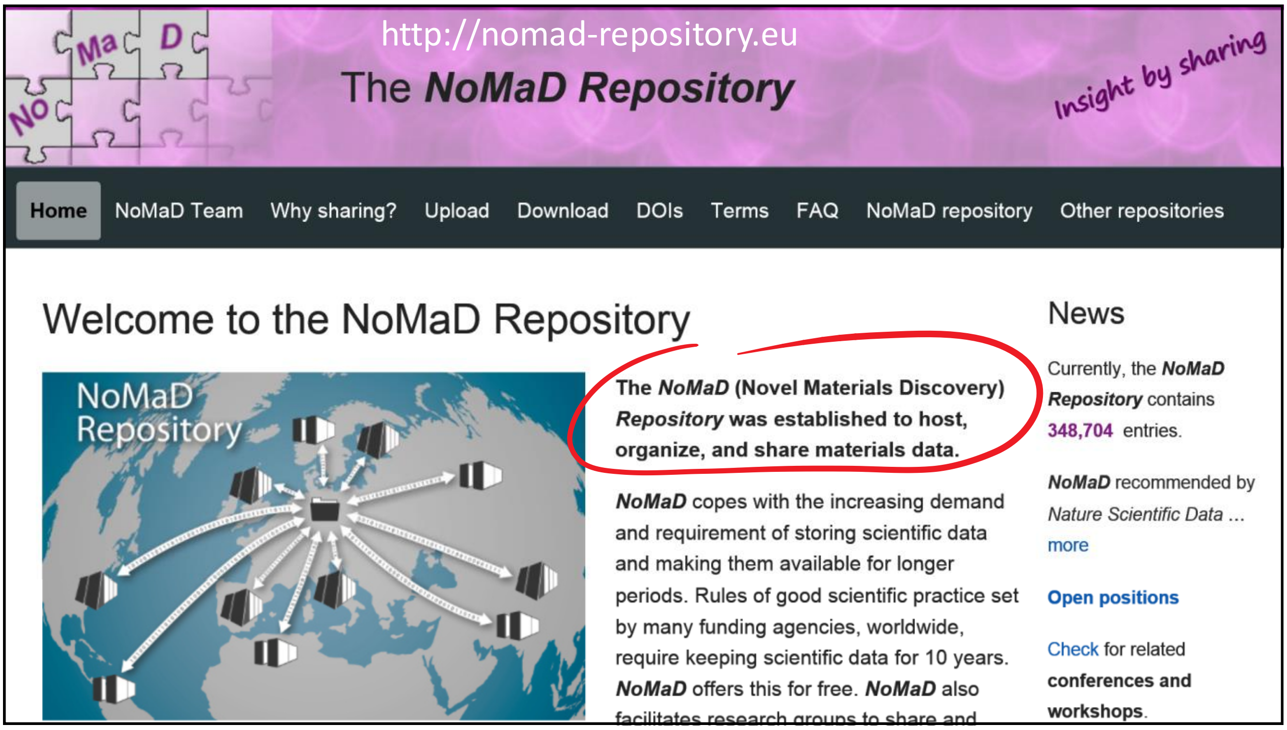
☑ Move the mouse over the bands to see their energies

☑ Show VBM and CBM

The band gap is 0.63 eV. The valence band maximum (VBM) is at ***k***-point (0.0, 0.0, 0.0) Å^{-1}. The conduction band minimum (CBM) is at ***k***-point (1.0, 0.0, 0.0) Å^{-1}.

Download data in json format (bands.json, dos.json) or download images in png format (bands.png , dos.png)

https://www.youtube.com/watch?v=L-nmRSH4NQM
http://v.youku.com/v_show/id_XMTM0NDA0NDIxMg==.html



http://nomad-repository.eu

The ***NoMaD Repository***

Insight by sharing

Welcome to the NoMaD Repository

The *NoMaD* (Novel Materials Discovery) *Repository* was established to host, organize, and share materials data.

NoMaD copes with the increasing demand and requirement of storing scientific data and making them available for longer periods. Rules of good scientific practice set by many funding agencies, worldwide, require keeping scientific data for 10 years. ***NoMaD*** offers this for free. ***NoMaD*** also facilitates research groups to share and

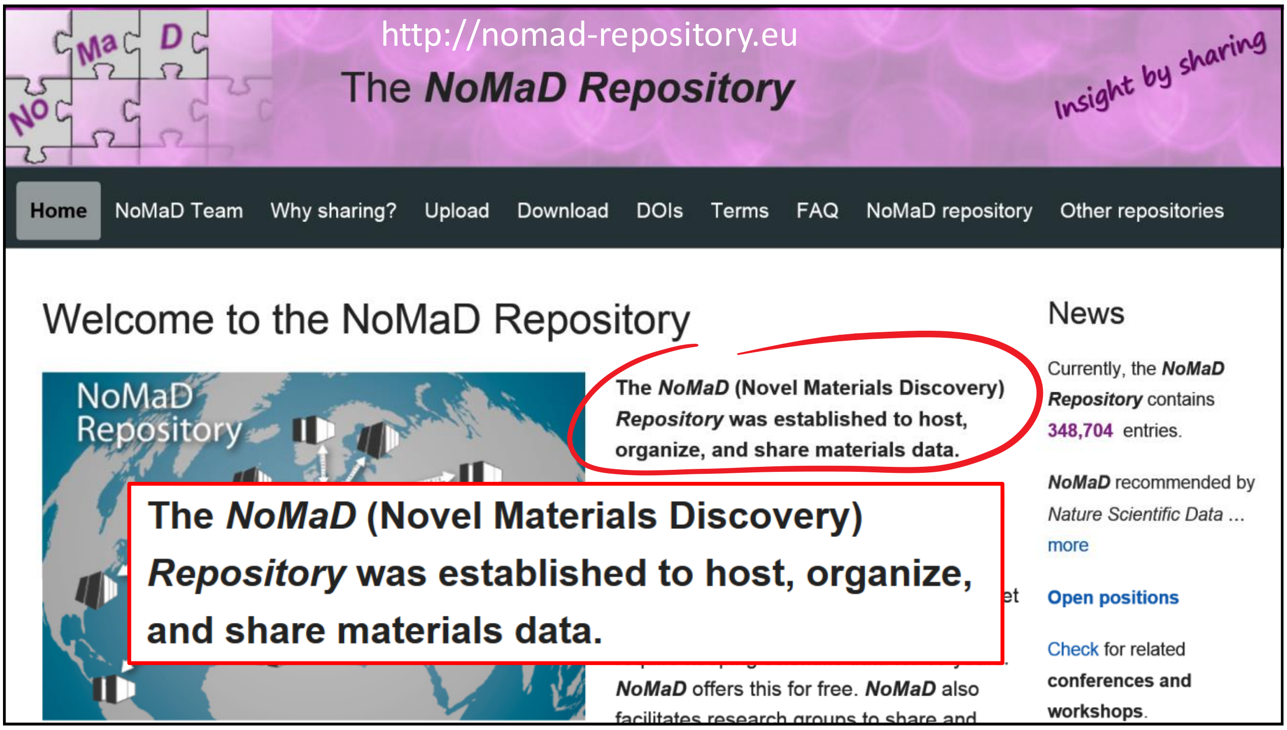
News

Currently, the ***NoMaD Repository*** contains 348,704 entries.

NoMaD recommended by *Nature Scientific Data* ... more

Open positions

Check for related **conferences and workshops**.



http://nomad-repository.eu

The ***NoMaD Repository***

Insight by sharing

Welcome to the NoMaD Repository

The *NoMaD* (Novel Materials Discovery) *Repository* was established to host, organize, and share materials data.

et

NoMaD offers this for free. ***NoMaD*** also facilitates research groups to share and

News

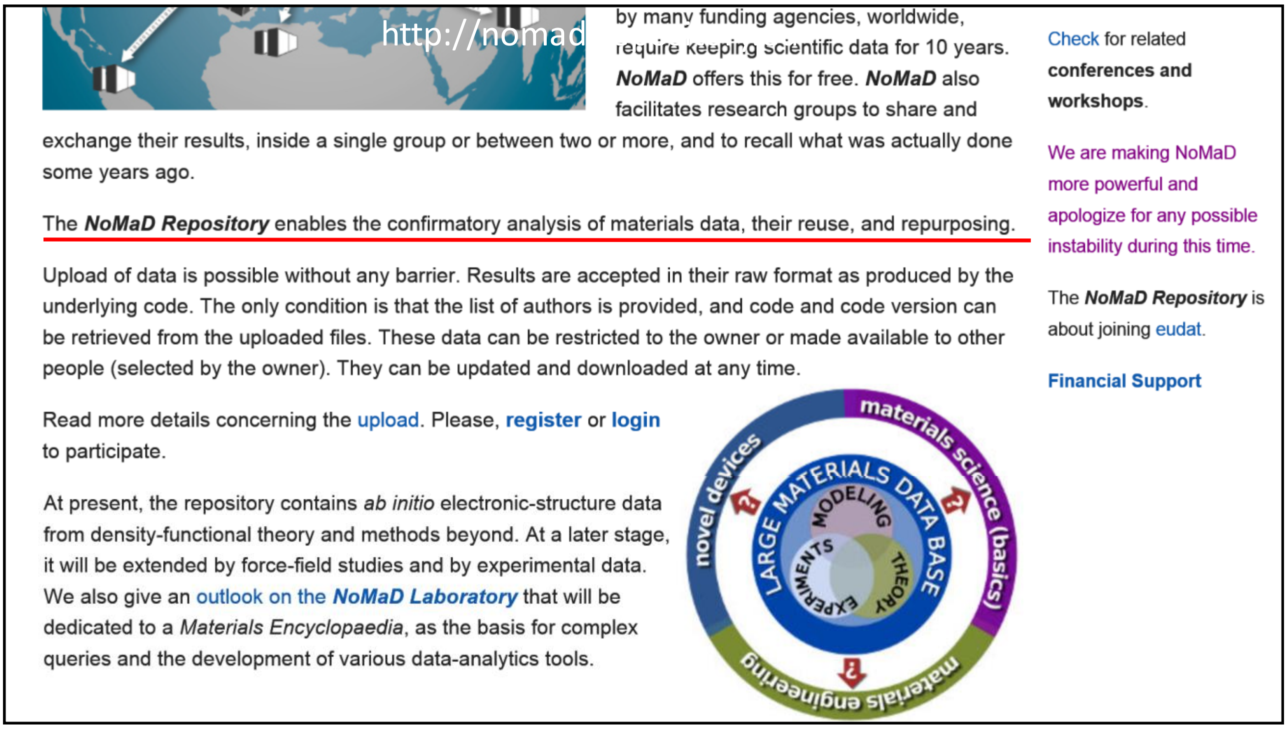
Currently, the ***NoMaD Repository*** contains 348,704 entries.

NoMaD recommended by *Nature Scientific Data* ... more

Open positions

Check for related **conferences and workshops**.

The *NoMaD* (Novel Materials Discovery) *Repository* was established to host, organize, and share materials data.



by many funding agencies, worldwide, require keeping scientific data for 10 years. ***NoMaD*** offers this for free. ***NoMaD*** also facilitates research groups to share and exchange their results, inside a single group or between two or more, and to recall what was actually done some years ago.

The ***NoMaD Repository*** enables the confirmatory analysis of materials data, their reuse, and repurposing.

Upload of data is possible without any barrier. Results are accepted in their raw format as produced by the underlying code. The only condition is that the list of authors is provided, and code and code version can be retrieved from the uploaded files. These data can be restricted to the owner or made available to other people (selected by the owner). They can be updated and downloaded at any time.

Read more details concerning the upload. Please, **register** or **login** to participate.

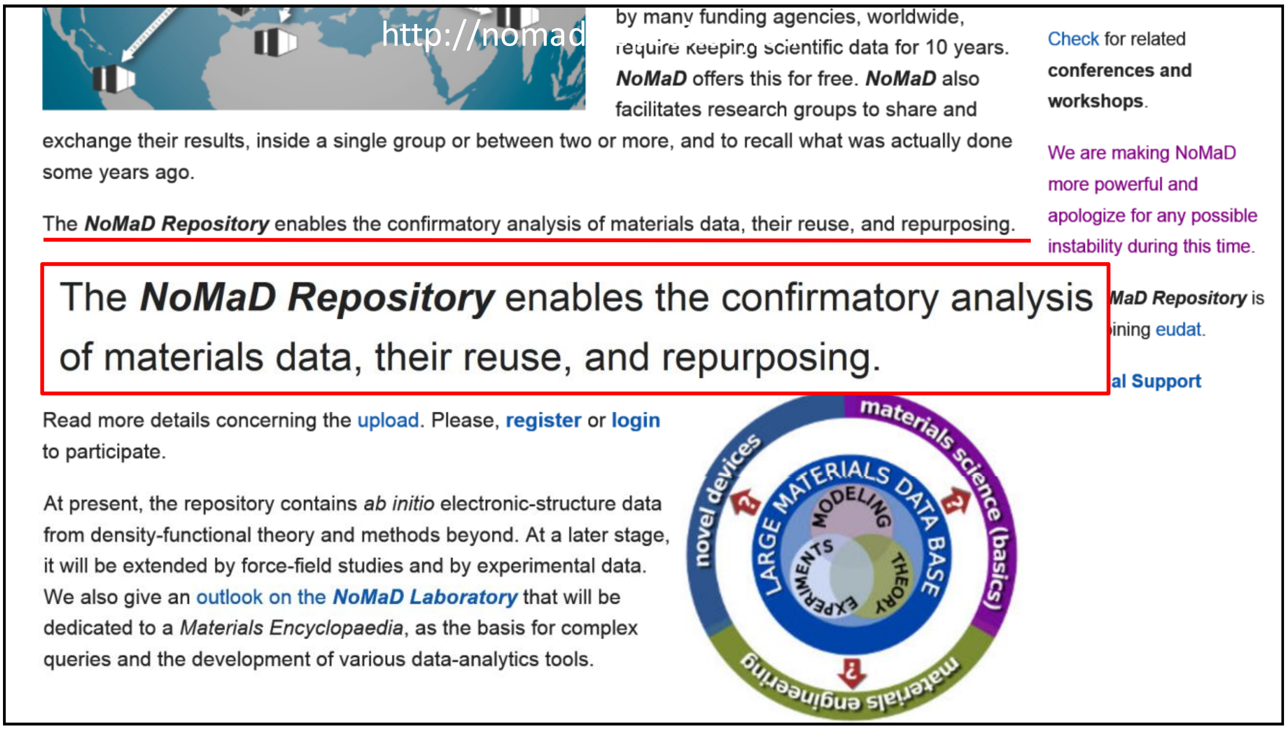
At present, the repository contains *ab initio* electronic-structure data from density-functional theory and methods beyond. At a later stage, it will be extended by force-field studies and by experimental data. We also give an outlook on the ***NoMaD Laboratory*** that will be dedicated to a *Materials Encyclopaedia*, as the basis for complex queries and the development of various data-analytics tools.

Check for related **conferences and workshops**.

We are making NoMaD more powerful and apologize for any possible instability during this time.

The ***NoMaD Repository*** is about joining eudat.

Financial Support



http://nomad

by many funding agencies, worldwide, require keeping scientific data for 10 years. ***NoMaD*** offers this for free. ***NoMaD*** also facilitates research groups to share and exchange their results, inside a single group or between two or more, and to recall what was actually done some years ago.

The ***NoMaD Repository*** enables the confirmatory analysis of materials data, their reuse, and repurposing.

The ***NoMaD Repository*** enables the confirmatory analysis of materials data, their reuse, and repurposing.

Read more details concerning the upload. Please, **register** or **login** to participate.

At present, the repository contains *ab initio* electronic-structure data from density-functional theory and methods beyond. At a later stage, it will be extended by force-field studies and by experimental data. We also give an outlook on the ***NoMaD Laboratory*** that will be dedicated to a *Materials Encyclopaedia*, as the basis for complex queries and the development of various data-analytics tools.

Check for related **conferences and workshops**.

We are making NoMaD more powerful and apologize for any possible instability during this time.

MaD Repository is ining eudat.

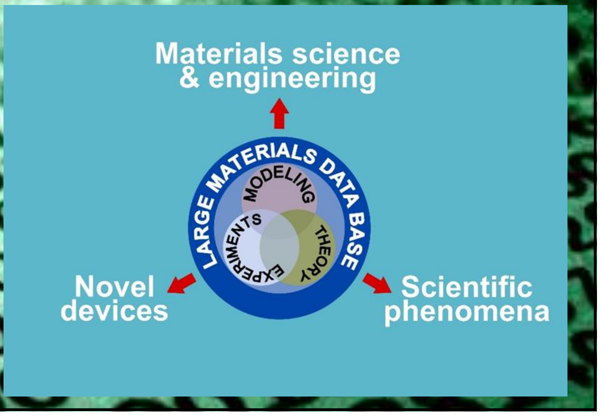
al Support

What To Do With The Data?

NoMaD
Repository

http://nomad-repository.eu

Currently, the NoMaD Repository contains **631,432** entries



The Four *V* of Big Data and an *A*

Data – data – data
(analog to Moore's law)

(so far: most data are not used and even thrown away)

The Four *V* of Big Data and an *A*

Data – data – data
(analog to Moore's law)

(so far: most data are not used and even thrown away)

Big-Data Challenge: *"four V"*:
Volume (amount of data),
Variety (heterogeneity of form and meaning of data),
Velocity at which data may change or new data arrive,
Veracity (uncertainty of quality).

The Four *V* of Big Data and an *A*

Data – data – data
(analog to Moore's law)

(so far: most data are not used and even thrown away)

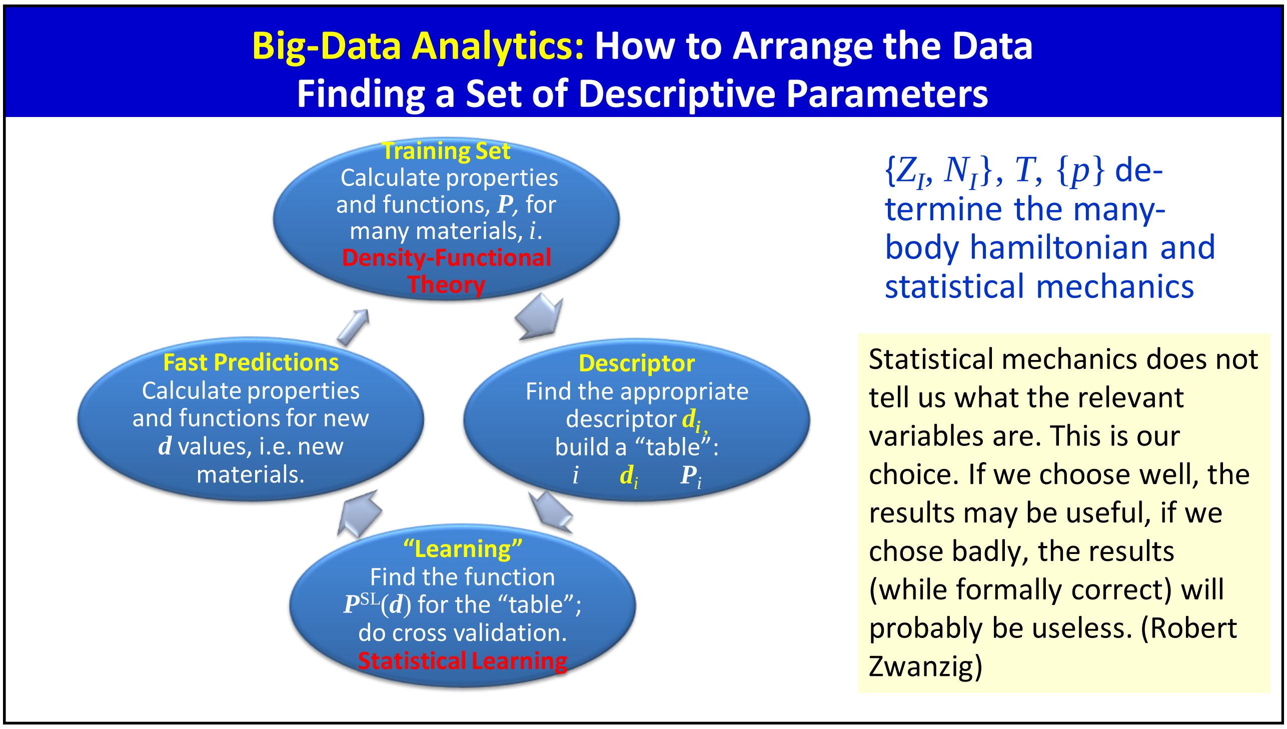
Big-Data Challenge: *"four V"*:
Volume (amount of data),
Variety (heterogeneity of form and meaning of data),
Velocity at which data may change or new data arrive,
Veracity (uncertainty of quality).

Query and read out what was stored; high-throughput screening.

Shouldn't we do more?!

The four V should be complemented by an "***A***", **Big-Data Analytics**:

- identify (so far) hidden trends,
- What is the next most promising candidate that should be studied?
- identify anomalies,
- identify the mechanisms behind a certain material property/function.



Big-Data Analytics: How to Arrange the Data
Finding a Set of Descriptive Parameters

Training Set
Calculate properties and functions, $\boldsymbol{P}$, for many materials, i.
Density-Functional Theory

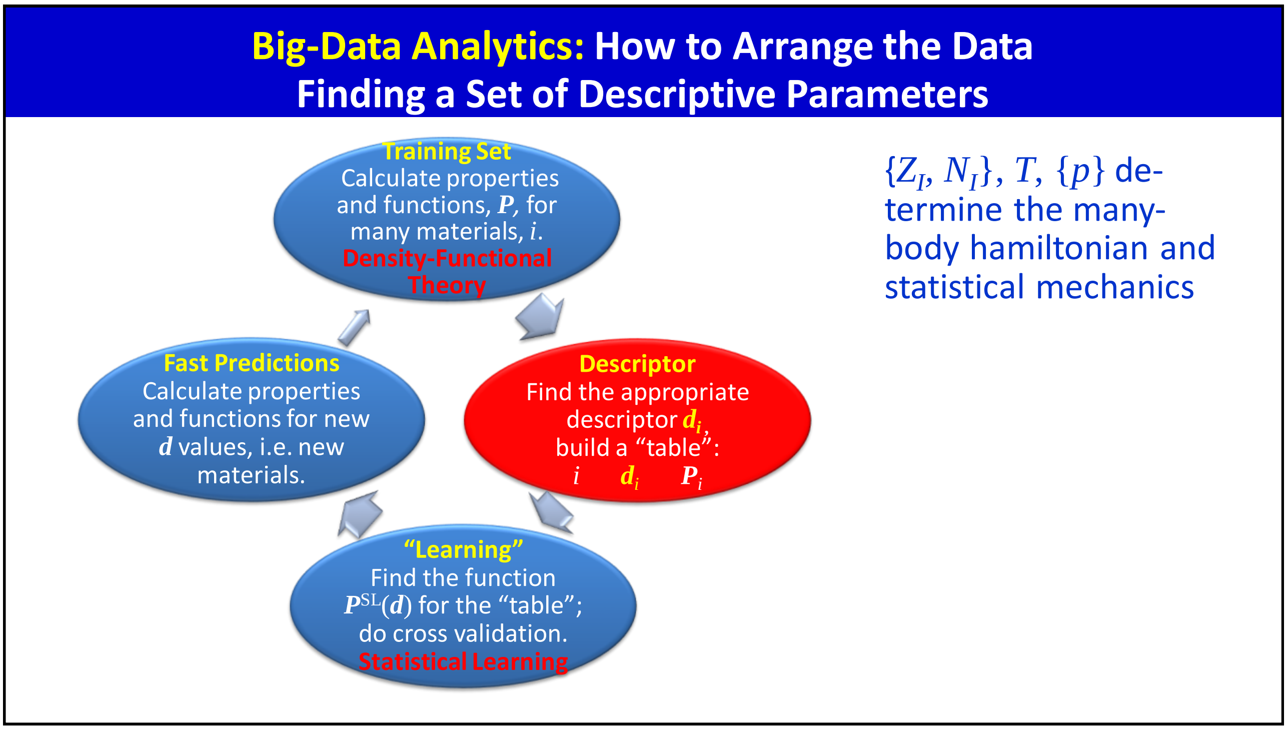
Descriptor
Find the appropriate descriptor $\boldsymbol{d}_i$, build a "table":
i $\boldsymbol{d}_i$ $\boldsymbol{P}_i$

"Learning"
Find the function $\boldsymbol{P}^{\mathrm{SL}}(\boldsymbol{d})$ for the "table"; do cross validation.
Statistical Learning

Fast Predictions
Calculate properties and functions for new $\boldsymbol{d}$ values, i.e. new materials.

$\{Z_I, N_I\}, T, \{p\}$ determine the many-body hamiltonian and statistical mechanics

Statistical mechanics does not tell us what the relevant variables are. This is our choice. If we choose well, the results may be useful, if we chose badly, the results (while formally correct) will probably be useless. (Robert Zwanzig)



Big-Data Analytics: How to Arrange the Data
Finding a Set of Descriptive Parameters

Training Set
Calculate properties and functions, $\boldsymbol{P}$, for many materials, i.
Density-Functional Theory

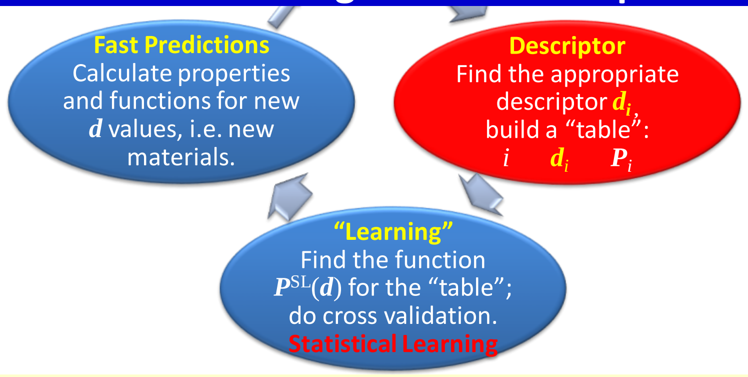
Descriptor
Find the appropriate descriptor $\boldsymbol{d}_i$, build a "table":
i $\boldsymbol{d}_i$ $\boldsymbol{P}_i$

"Learning"
Find the function $\boldsymbol{P}^{\mathrm{SL}}(\boldsymbol{d})$ for the "table"; do cross validation.
Statistical Learning

Fast Predictions
Calculate properties and functions for new $\boldsymbol{d}$ values, i.e. new materials.

$\{Z_I, N_I\}, T, \{p\}$ determine the many-body hamiltonian and statistical mechanics

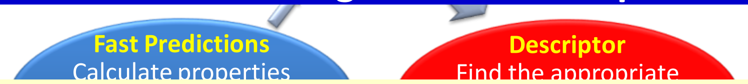
Big-Data Analytics: How to Arrange the Data
Finding a Set of Descriptive Parameters


Fast Predictions
Calculate properties and functions for new $\boldsymbol{d}$ values, i.e. new materials.
Descriptor
Find the appropriate descriptor $\boldsymbol{d}_i$, build a “table”:
$i \quad \boldsymbol{d}_i \quad \boldsymbol{P}_i$
“Learning”
Find the function $\boldsymbol{P}^{\mathrm{SL}}(\boldsymbol{d})$ for the “table”; do cross validation.
Statistical Learning

$\{Z_I, N_I\}$, T, $\{p\}$ determine the many-body hamiltonian and statistical mechanics

$\boldsymbol{d}$ characterizes the relevant mechanisms that govern the observed property/function $\boldsymbol{P}$. Identifying the descriptor $\boldsymbol{d}$ from known data $\boldsymbol{P}_i$, is an ill-posed problem (statistical-learning theory): A little error in the data $\boldsymbol{P}_i$ may suggest a different descriptor $\boldsymbol{d}$. Thus, knowledge of the accuracy of data $\boldsymbol{P}_i$ is crucial (veracity). The choice of $\boldsymbol{d}$ is not unique.

Big-Data Analytics: How to Arrange the Data
Finding a Set of Descriptive Parameters



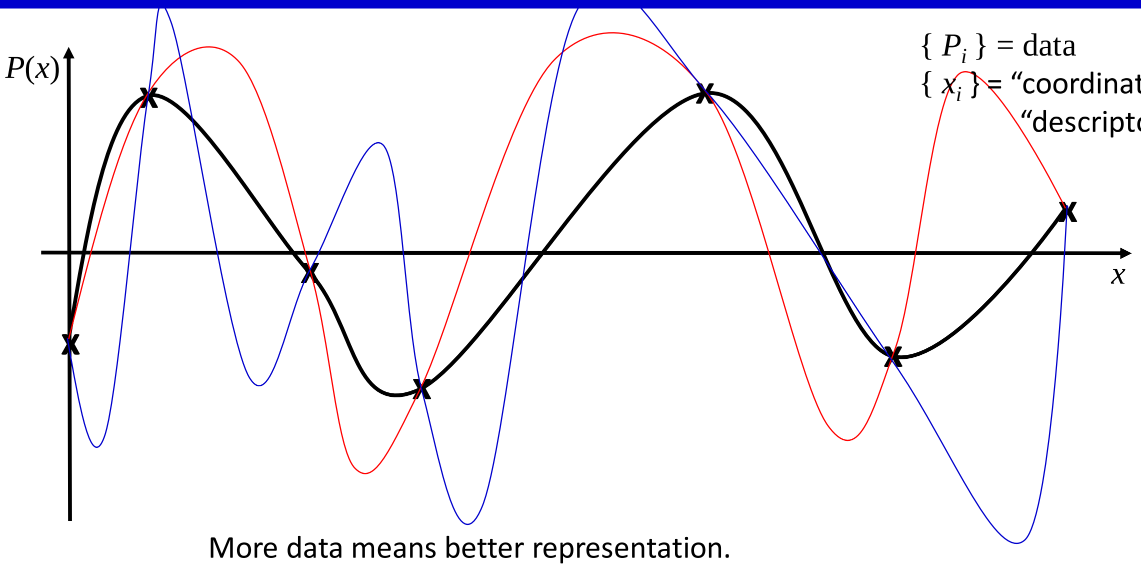
$\{Z_I, N_I\}$, T, $\{p\}$ de-

$\boldsymbol{d}$ characterizes the relevant mechanisms that govern the observed property/function $\boldsymbol{P}$. Identifying the descriptor $\boldsymbol{d}$ from known data $\boldsymbol{P}_i$, is an ill-posed problem (statistical-learning theory): A little error in the data $\boldsymbol{P}_i$ may suggest a different descriptor $\boldsymbol{d}$. Thus, knowledge of the accuracy of data $\boldsymbol{P}_i$ is crucial (veracity). The choice of $\boldsymbol{d}$ is not unique.

A) Veracity: Accuracy of state-of-the-art density-functional theory (validation and verification)

B) Descriptor: How to find it, how to understand the causality between $\boldsymbol{d}$ and $\boldsymbol{P}^{\mathrm{SL}}$?

Data Fitting, Statistical Learning, and Machine Learning



$\{\, P_i \,\}$ = data
$\{\, x_i \,\}$ = “coordinate” or “descriptor”

More data means better representation.
Think about what may be the best (meaningful) coordinate!

Kernel Regression

We have data $\{\boldsymbol{P}_i\}$ at “coordinates” $\{\mathbf{x}_i\}$ $\mathbf{x}_i$ = set of descriptive parameters (descriptor)

$$\boldsymbol{P}_i \overset{!}{=} \boldsymbol{P}^{\mathrm{SL}}(\mathbf{x}_i) = \sum_{k=1}^{N} \mathrm{c}_k K(\mathbf{x}_i, \mathbf{x}_k)$$

Linear regression: $K(\mathbf{x}_i, \mathbf{x}_k) = \mathbf{x}_i \,.\, \mathbf{x}_k$ $\boldsymbol{P}^{\mathrm{SL}}(\mathbf{x}_i) = \mathbf{x}_i \,.\, \boldsymbol{c}^*$

Polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_k) = (\, \mathbf{x}_i \,.\, \mathbf{x}_k + c \,)^{d}$

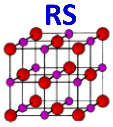
Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_k) = \exp\left(-\sum_j (\, x_i - x_k)^2 \,/\, 2\sigma_j^2 \right)$

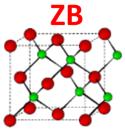
More data means better representation.
Do we “learn” anything?

For successful learning, we need a “good” descriptor: $\{\mathbf{x}_i\} \rightarrow \{\boldsymbol{d}_i\}$

Toy Model: Descriptor for the Classification “Zincblende/Wurtzite or Rocksalt?”

Can we predict not yet calculated structures from Z_A and Z_B? **Can we create a map: “The *ZB/W* community lives here and the *RS* community there?”**





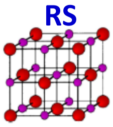
Energy differences between different structures are very small.

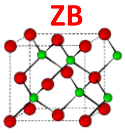
For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons.

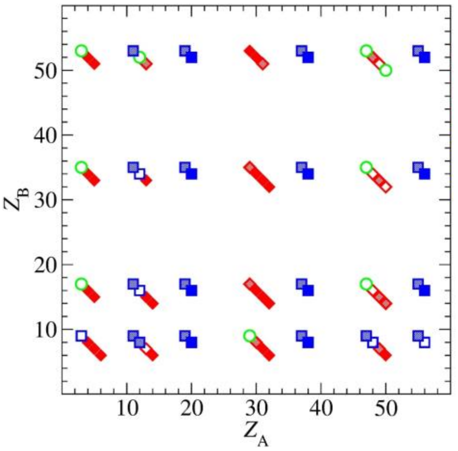
Complexity: $T_s[n]$ and E_{xc}.

Toy Model: Descriptor for the Classification “Zincblende/Wurtzite or Rocksalt?”

Can we predict not yet calculated structures from Z_A and Z_B? **Can we create a map: “The *ZB/W* community lives here and the *RS* community there?”**







$\Delta = E(\mathrm{RS}) - E(\mathrm{ZB})$

◆ ZB, $\Delta > 0.2$ eV

◆ ZB, $0.1\ \mathrm{eV} < \Delta \le 0.2$ eV

◇ ZB, $0.05\ \mathrm{eV} < \Delta \le 0.1$ eV

○ $-0.05\ \mathrm{eV} < \Delta \le 0.05$ eV

□ RS, $-0.1\ \mathrm{eV} < \Delta \le -0.05$ eV

▨ RS, $-0.2\ \mathrm{eV} < \Delta \le -0.1$ eV

■ RS, $\Delta \le -0.2$ eV

Energy differences between different structures are very small.

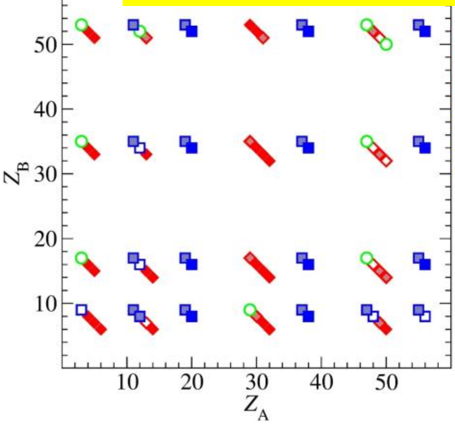
For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons.

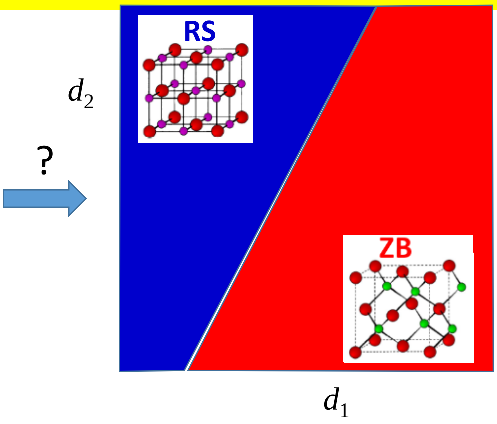
Complexity: $T_s[n]$ and E_{xc}.

Toy Model: Descriptor for the Classification "Zincblende/Wurtzite or Rocksalt?"

We need to arrange the data such that statistical learning is efficient. We need a good set of descriptive parameters.

How to find d_1, d_2?
In reality the representation will be higher than 2-dimensional.





J. A. van Vechten, Phys. Rev. 182, 891 (1969).
J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).
A. Zunger, Phys. Rev. B 22, 5839 (1980).
D. G. Pettifor, Solid State Commun. 51, 31 (1984).
Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

Statistical Learning (Machine Learning)

fit and/or interpolation of known data points $\{ P_i \}$ and building a function $P(\boldsymbol{d})$

the key scientific challenge: find a reliable, low dimensional descriptor $\boldsymbol{d}$.

kernel ridge regression	linear
$P(\mathbf{d}) = \sum_{i=1}^{N} c_i \exp\left(-\lVert \mathbf{d}_i - \mathbf{d} \rVert_2^2 / 2\sigma^2\right)$	$P(\mathbf{d}) = \mathbf{dc}$

Statistical Learning (Machine Learning)

fit and/or interpolation of known data points $\{ P_i \}$ and building a function $P(\boldsymbol{d})$

the key scientific challenge: find a reliable, low dimensional descriptor $\boldsymbol{d}$.

kernel ridge regression

$$P(\mathbf{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\mathbf{d}_i - \mathbf{d}\|_2^2/2\sigma^2\right)$$

minimize

$$\sum_{i=1}^{N}(P(\mathbf{d}_i) - P_i)^2 \quad +$$

$$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\mathbf{d}_i - \mathbf{d}_j\|_2^2/2\sigma^2\right)$$

$$\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$$

linear

R. Tibshirani, J. Royal Statist. Soc. B 58, 267 (1996)

$$P(\mathbf{d}) = \mathbf{dc}$$

$$\sum_{i=1}^{N}(P(\mathbf{d}_i) - P_i)^2 \quad +$$

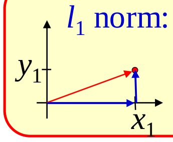
$$\lambda\|\mathbf{c}\|_1$$

$$\|\mathbf{c}\|_1 = \sum_{\alpha=1}^{M} |c_\alpha|$$

least absolute shrinkage and selection operator (LASSO) for feature selection

Statistical Learning (Machine Learning)

l_1 norm: $|x_1| + |y_1|$ Manhattan (taxi cab) distance



l_2 norm: sqrt($x_1^2 + y_1^2$)

kernel ridge regression

$$P(\mathbf{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\mathbf{d}_i - \mathbf{d}\|_2^2/2\sigma^2\right)$$

minimize

$$\sum_{i=1}^{N}(P(\mathbf{d}_i) - P_i)^2 \quad +$$

$$\lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\mathbf{d}_i - \mathbf{d}_j\|_2^2/2\sigma^2\right)$$

$$\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$$

linear

R. Tibshirani, J. Royal Statist. Soc. B 58, 267 (1996)

$$P(\mathbf{d}) = \mathbf{dc}$$

$$\sum_{i=1}^{N}(P(\mathbf{d}_i) - P_i)^2 \quad +$$

$$\lambda\|\mathbf{c}\|_1$$

$$\|\mathbf{c}\|_1 = \sum_{\alpha=1}^{M} |c_\alpha|$$

least absolute shrinkage and selection operator (LASSO) for feature selection

1) Primary Features, 2) Feature Space, 3) Descriptors

1)

ID	Description **free atoms**	Symbols	#
*A*1	Ionization Potential (IP) and Electron Affinity (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
*A*2	Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels	H(A) L(A) H(B) L(B)	4
*A*3	Radius at the max. value of *s*, *p*, and *d* valence radial radial probability density	$r_s(A)$ $r_p(A)$ $r_d(A)$ $r_s(B)$ $r_p(B)$ $r_d(B)$	6

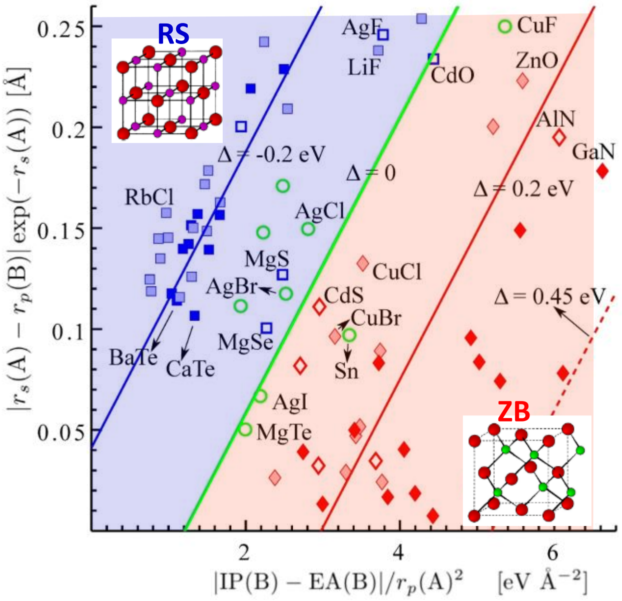
ID	Description **free dimers**	Symbols	#
*A*4	Binding energy	$E_b(AA)$ $E_b(BB)$ $E_b(AB)$	3
*A*5	HOMO-LUMO KS gap	HL(AA) HL(BB) HL(AB)	3
*A*6	Equilibrium distance	$d(AA)$ $d(BB)$ $d(AB)$	3

2) We start with 23 primary features
and build > 10,000 non linear combinations

3) LASSO finds the descriptors:

$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2},\ \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))},\ \frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}) + r_s(\mathrm{B}))}$$

"The Map"
Statistical Learning (Machine Learning): LASSO, 2-Dim. Descriptor



Δ = E(RS) − E(ZB)
◆ ZB, Δ > 0.2 eV
◆ ZB, 0.1 eV < Δ ≤ 0.2 eV
◇ ZB, 0.05 eV < Δ ≤ 0.1 eV
○ − 0.05 eV < Δ ≤ 0.05 eV
□ RS, − 0.1 eV < Δ ≤ − 0.05 eV
▨ RS, − 0.2 eV < Δ ≤ − 0.1 eV
■ RS, Δ ≤ − 0.2 eV

$$\boxed{P(\mathbf{d}) = \mathbf{d}\mathbf{c}}$$

The complexity and science is in the descriptor (identified from >10,000 features).

L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. ***114****, 105503 (2015).*

Statistical Learning (Machine Learning): Descriptor

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Statistical Learning (Machine Learning): Descriptor

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

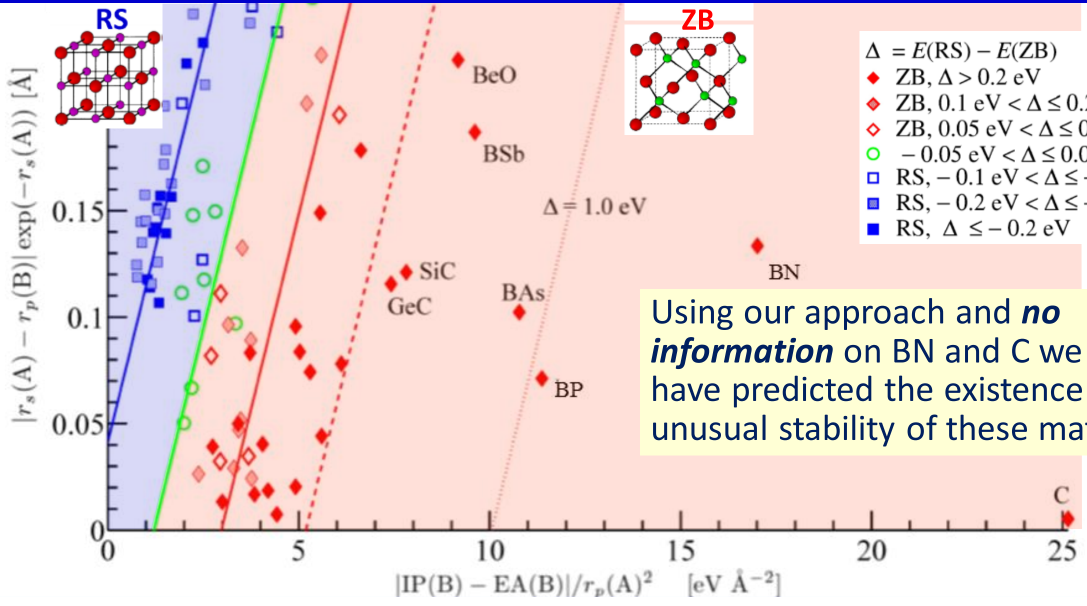
Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Statistical Learning (Machine Learning): Descriptor

Mean absolute error (MAE), and maximum absolute error (MaxAE), in eV, (first two lines) and for a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last two lines). For (Z_A*, Z_B*), each atom is identified by a string of three random numbers.

Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Statistical Learning (Machine Learning): LASSO, 2-Dim. Descriptor



Using our approach and ***no information*** on BN and C we would have predicted the existence and unusual stability of these materials.

Big-Data-Driven Science vs. Model-Driven Science

Traditional approach in the empirical sciences (e.g. physics, chemistry):

- Study a few systems
- Build a model,
- Improve the model when needed

(e.g. strength of transition metals Ti, … Fe, … Cu: d-band occupation, etc.).

The new option offered by Big-Data Analytics (and big-data-driven science):

- Find structure in big data that is probably invisible for humans.
- Offer many (thousands) of optional models, and
- employ applied math/information theory to find out which model is best (e.g. compressed sensing, statistical learning).

Drawing Causal Inference from Big Data (Scientific Insight) -- can we trust a prediction? --

Correlation between $\boldsymbol{d}$ and $\boldsymbol{P}$, i.e. $\boldsymbol{P}$ is a function of $\boldsymbol{d}$, $\boldsymbol{P}(\boldsymbol{d})$,
reflects causal inference
if it is based on sufficient information(*)

Judea Pearl

There are four possibilities (types of causality) behind $\boldsymbol{P}(\boldsymbol{d})$:

1. $\boldsymbol{d} \to \boldsymbol{P}$: $\boldsymbol{P}$ “listens” to $\boldsymbol{d}$
2. $\boldsymbol{A} \to \boldsymbol{d}$ and $\boldsymbol{A} \to \boldsymbol{P}$: There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but $\boldsymbol{d}$ and $\boldsymbol{P}$ both “listen” to a third “actuator”
3. $\boldsymbol{P} \to \boldsymbol{d}$: $\boldsymbol{d}$ “listens” to $\boldsymbol{P}$
4. There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but they have a common effect that listens to both and screams: “I occurred“ (Berkson bias; Judea Pearl)

(*) Construct $\boldsymbol{d}$ with scientific knowledge (prejudice?), or use “big data” for $\{\boldsymbol{P}_i\}$.

Drawing Causal Inference from Big Data (Scientific Insight) -- can we trust a prediction? --

LASSO has provided us with an equation for the quantitative energy difference:

$$\Delta E = 0.108\frac{\mathrm{EA(B)} - \mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.790\frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} + 3.766\frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}))} - 0.0267$$

This is an equation, but not a scientific law:

Case #2:

Nuclear numbers Z_A, Z_B ↔ our descriptor

Nuclear numbers Z_A, Z_B → total-energy differences

a mapping exists, even a physical intuition exist, but ΔE does not listen directly to the descriptor (intricate causality)

Drawing Causal Inference from Big Data (Scientific Insight) -- can we trust a prediction? --

Correlation between $\boldsymbol{d}$ and $\boldsymbol{P}$, i.e. $\boldsymbol{P}$ is a function of $\boldsymbol{d}$, $\boldsymbol{P}(\boldsymbol{d})$,
reflects causal inference
if it is based on sufficient information[(*)]

Judea Pearl

There are four possibilities (types of causality) behind $\boldsymbol{P}(\boldsymbol{d})$:

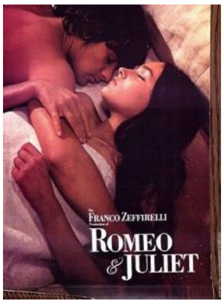
1. $\boldsymbol{d} \rightarrow \boldsymbol{P}$: $\boldsymbol{P}$ "listens" to $\boldsymbol{d}$
2. $\boldsymbol{A} \rightarrow \boldsymbol{d}$ and $\boldsymbol{A} \rightarrow \boldsymbol{P}$: There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but $\boldsymbol{d}$ and $\boldsymbol{P}$ both "listen" to a third "actuator"
3. $\boldsymbol{P} \rightarrow \boldsymbol{d}$: $\boldsymbol{d}$ "listens" to $\boldsymbol{P}$
4. There is no direct connection between $\boldsymbol{d}$ and $\boldsymbol{P}$, but they have a common effect that listens to both and screams: "I occurred" (Berkson bias; Judea Pearl)

[(*)] Construct $\boldsymbol{d}$ with scientific knowledge (prejudice?), or use "big data" for $\{\boldsymbol{P}_i\}$.

Drawing Causal Inference from Big Data (Scientific Insight) -- can we trust a prediction? --

ROMEO: “It was the lark, the bird that sings at dawn, not the nightingale. Look, my love, what are those streaks of light in the clouds parting in the east? Night is over, and day is coming. … ”

case # 3

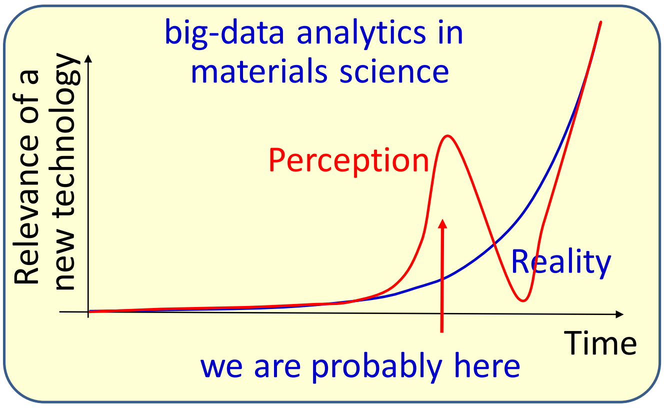


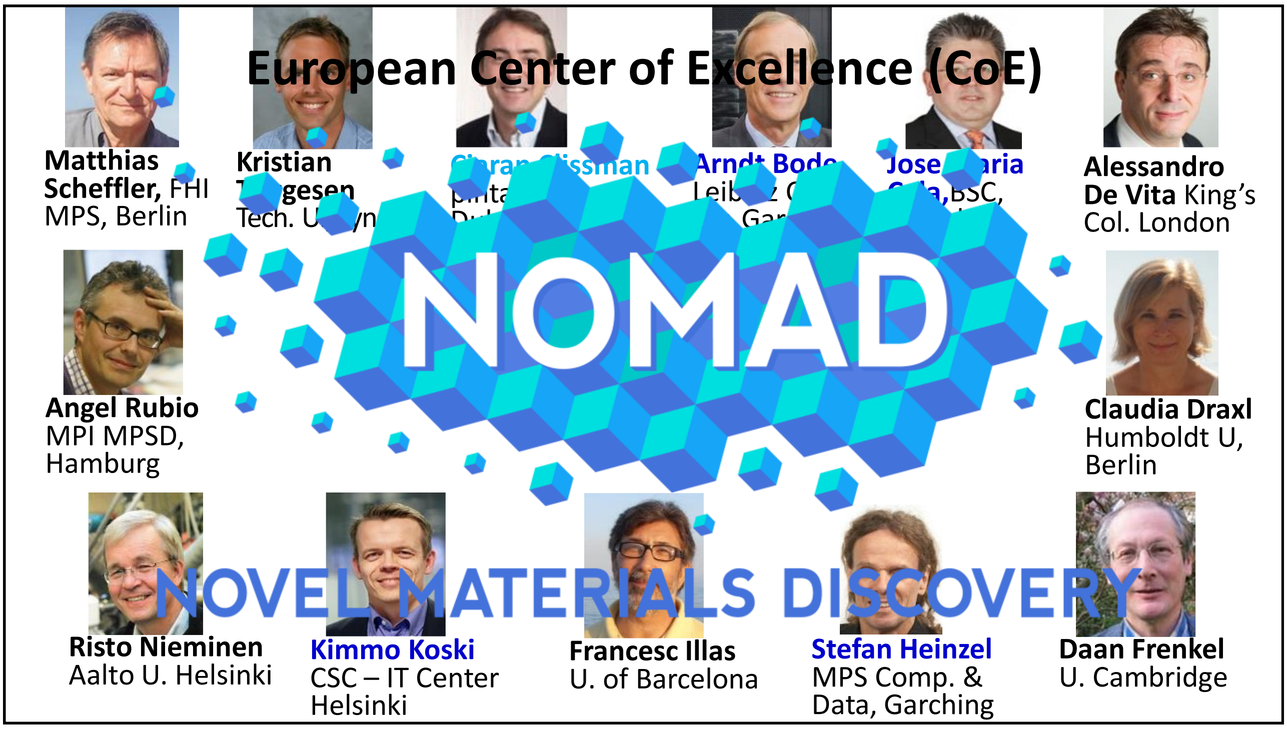
The ***singing of the lark*** is a good descriptor for “***the sun will rise soon”***.
The ***singing of the lark*** is not the actuator of (the mechanism behind) the sunrise.

Conclusion / Suggestion: Accept “larks” (not just scientific laws) to predict materials properties.

Summary and Outlook

- Machine learning ***may*** find structure in “big data” that is invisible to humans.
- Correlation reflects causal inference (if based on sufficient information).
- The causality may be intricate and complex.
- Causal models, i.e. employing ***causal descriptors***, are able to provide scientific insight and understanding.





European Center of Excellence (CoE)

Matthias Scheffler, FHI MPS, Berlin

Kristian T gesen Tech. U yn

Ciaran Clissman pinta Dul

Arndt Bode Leib z C Gar

Jose aria C a, BSC,

Alessandro De Vita King's Col. London

Angel Rubio MPI MPSD, Hamburg

Claudia Draxl Humboldt U, Berlin

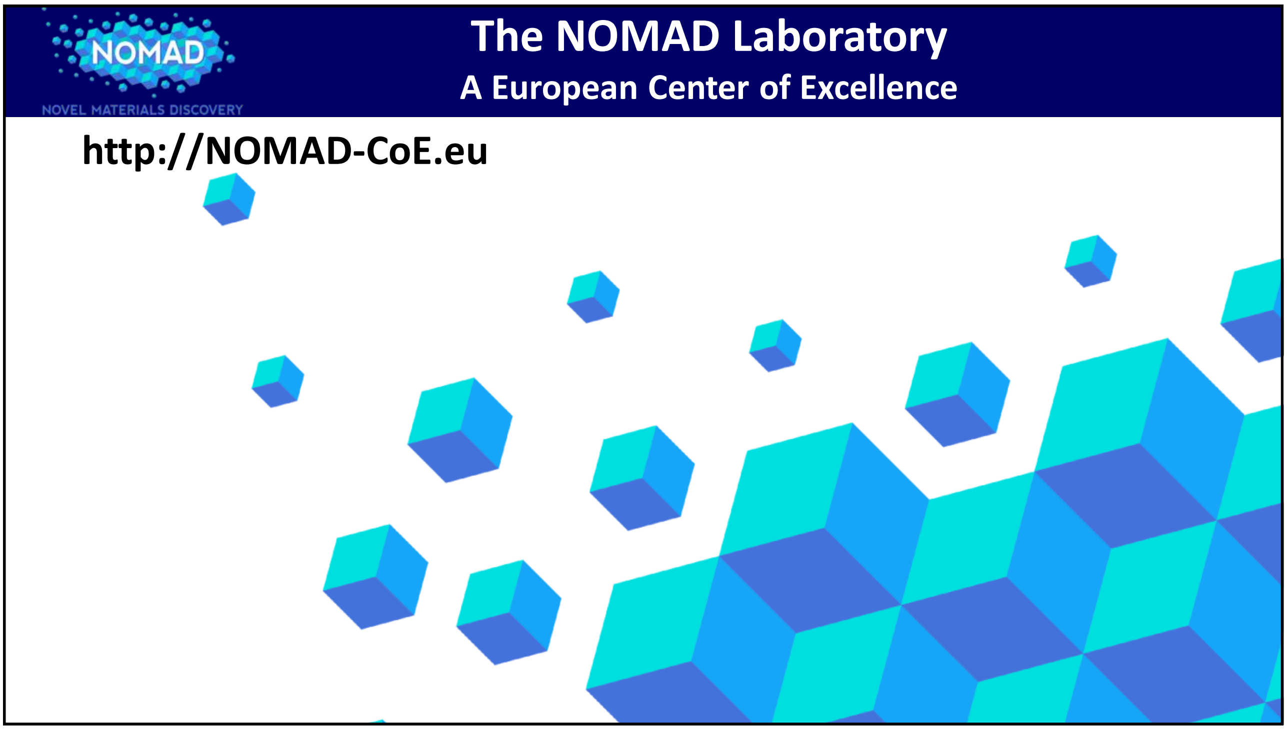
Risto Nieminen Aalto U. Helsinki

Kimmo Koski CSC – IT Center Helsinki

Francesc Illas U. of Barcelona

Stefan Heinzel MPS Comp. & Data, Garching

Daan Frenkel U. Cambridge



The NOMAD Laboratory

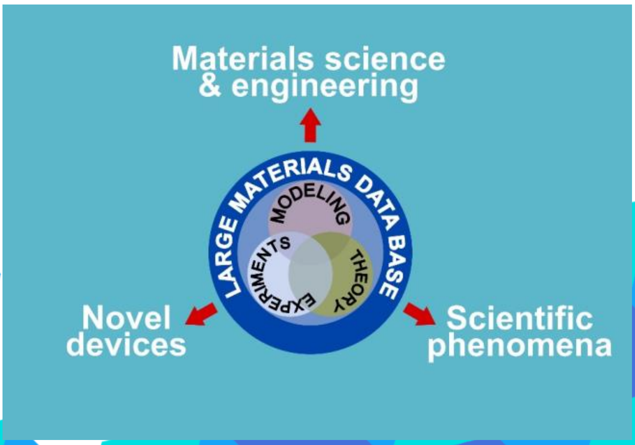
A European Center of Excellence

http://NOMAD-CoE.eu

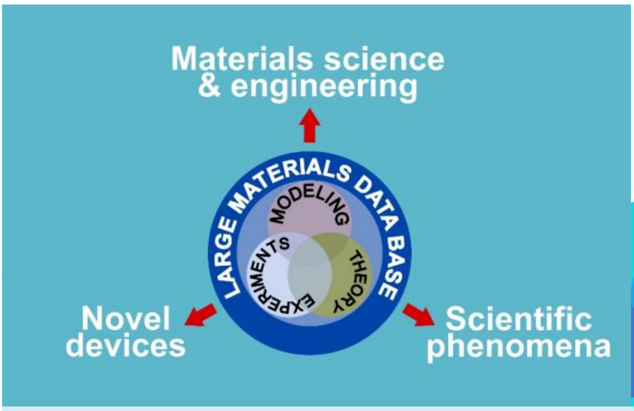
The NOMAD Laboratory

A European Center of Excellence

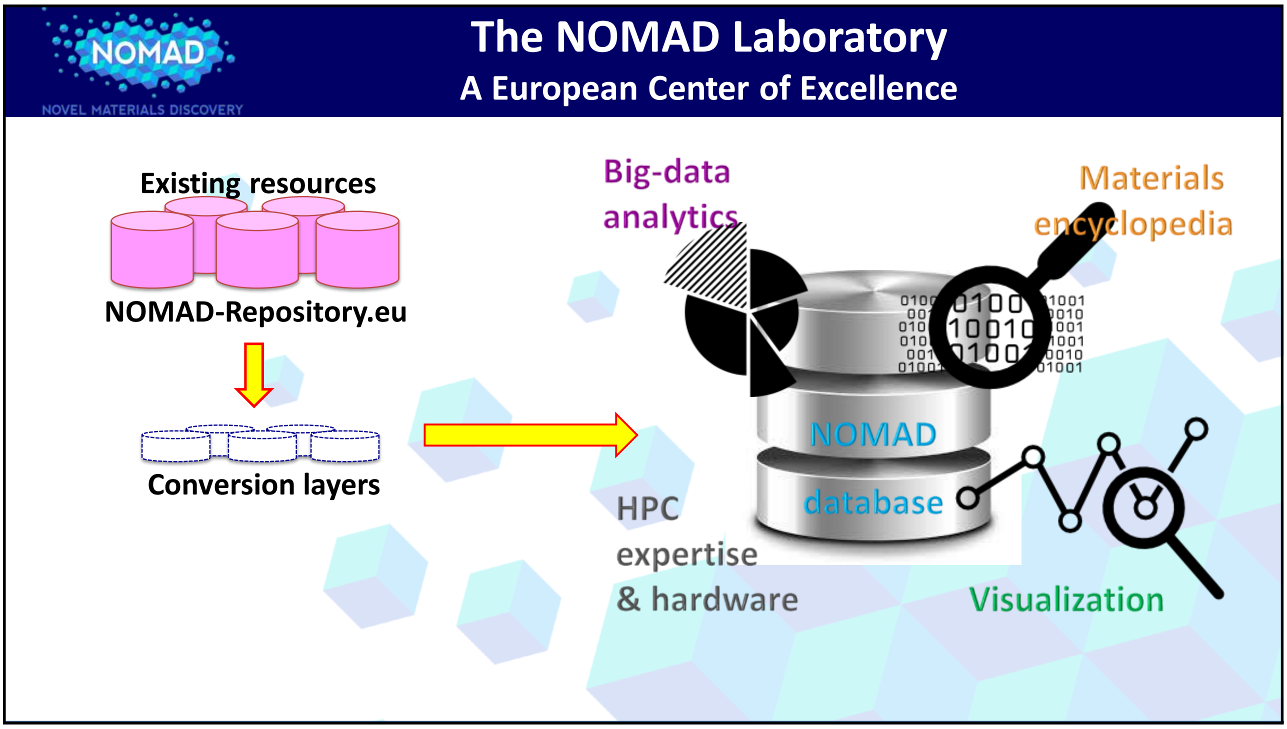
http://NOMAD-CoE.eu



The NOMAD Laboratory
A European Center of Excellence
http://NOMAD-CoE.eu



The NOMAD CoE develops a *Materials Encyclopedia* and *Big-Data Analytics* tools for materials science and engineering. Eight complementary research groups of highest scientific standing in computational materials science along with four high-performance computer centers form the synergetic core of this CoE.



The NOMAD Laboratory

A European Center of Excellence

Existing resources

NOMAD-Repository.eu

Conversion layers

Big-data analytics

Materials encyclopedia

NOMAD database

HPC expertise & hardware

Visualization