Summer School of the Max-Planck-EPFL Center for Molecular Nanoscience & Technology

July 27 - 31, 2015 in Schloss Ringberg, Germany

	Monday	Tuesday	Wednesday	Thursday	Friday
8:45 - 9:55		Silke Biermann - *Electronic structure calculations using dynamical mean field theory*	Ivano Tavernelli - *Trajectory-based nonadiabatic dynamics using time-dependent density functional theory*	Cecile Hebert - *Investigation of molecules at surfaces and chemical reactions by transmission electron microscopy: is a dream becoming true ?*	
9:55 - 11:05		Matthieu Verstraete - *Ab initio approaches to electron transport*	Olle Hellman - *Phonons and anharmonics*	Andrea Cepellotti – *Thermal Transport in 2D Materials*	Alec Wodtke - *The dynamics of molecular interactions and chemical reactions at metal surfaces: Testing the foundations of theory*
11:05		Coffee break	Coffee break	Coffee break	Coffee break
11:25- 12:35		Carsten Baldauf - *Molecular dynamics of peptides in isolation and computation on physical observables*	Matthias Scheffler - *Big-Data Analytics for Materials Science: Concepts, Challenges, and Hype*	Markus Elstner -*Multiscale simulations of biological structures and processe*	Examinations
12:35		Lunch	Lunch	Lunch	Lunch
14:15 - 15:25		Tom Rizzo - *Biomolecules in isolation – challenges and benchmarks for theory*	Christian Carbogno – *Thermal Conductivities from First Principles Molecular Dynamics*		
15:25	Coffee break	Coffee break	Coffee break		

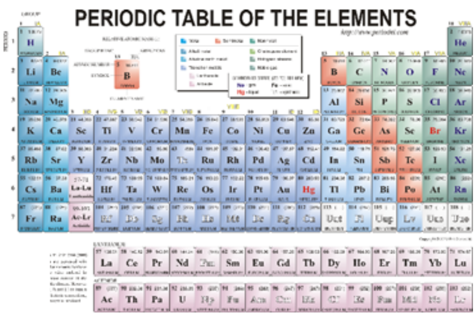
Big Data Analytics for Materials Science: Concepts, Challenges, and Hype

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.**

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

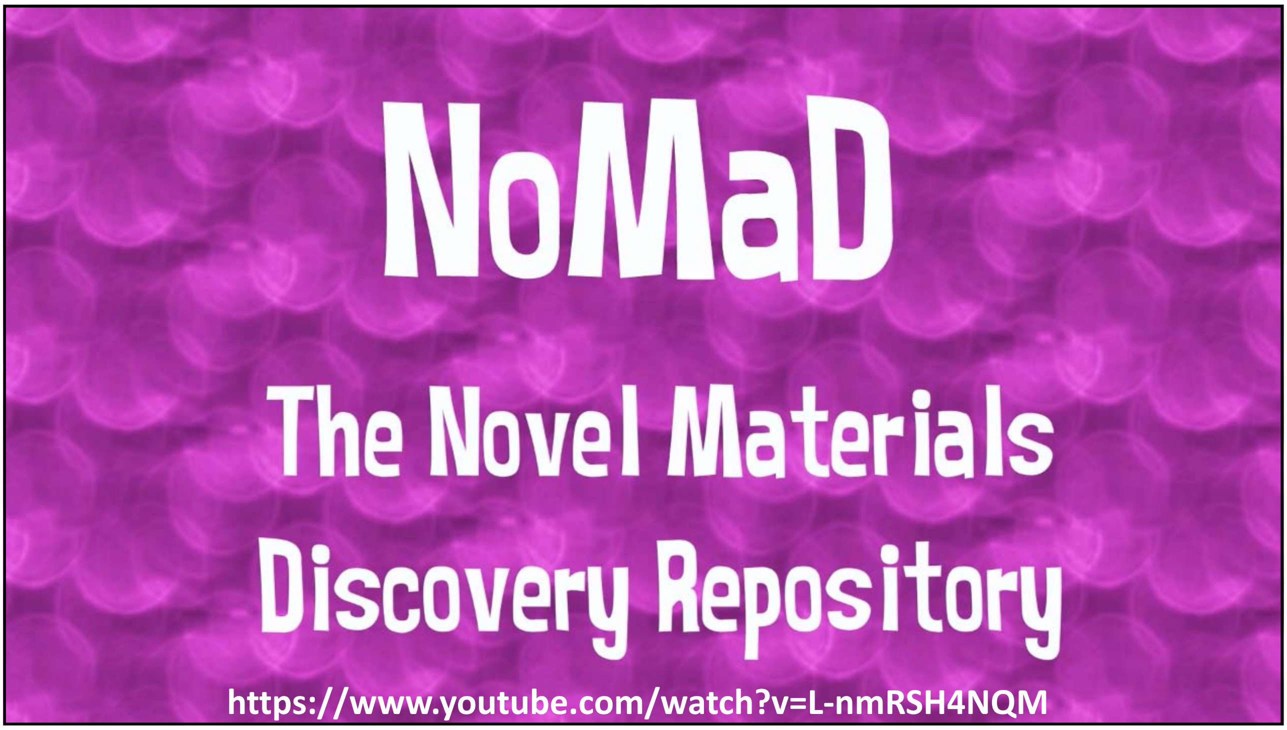
"twice as fast,
at a fraction of the cost"

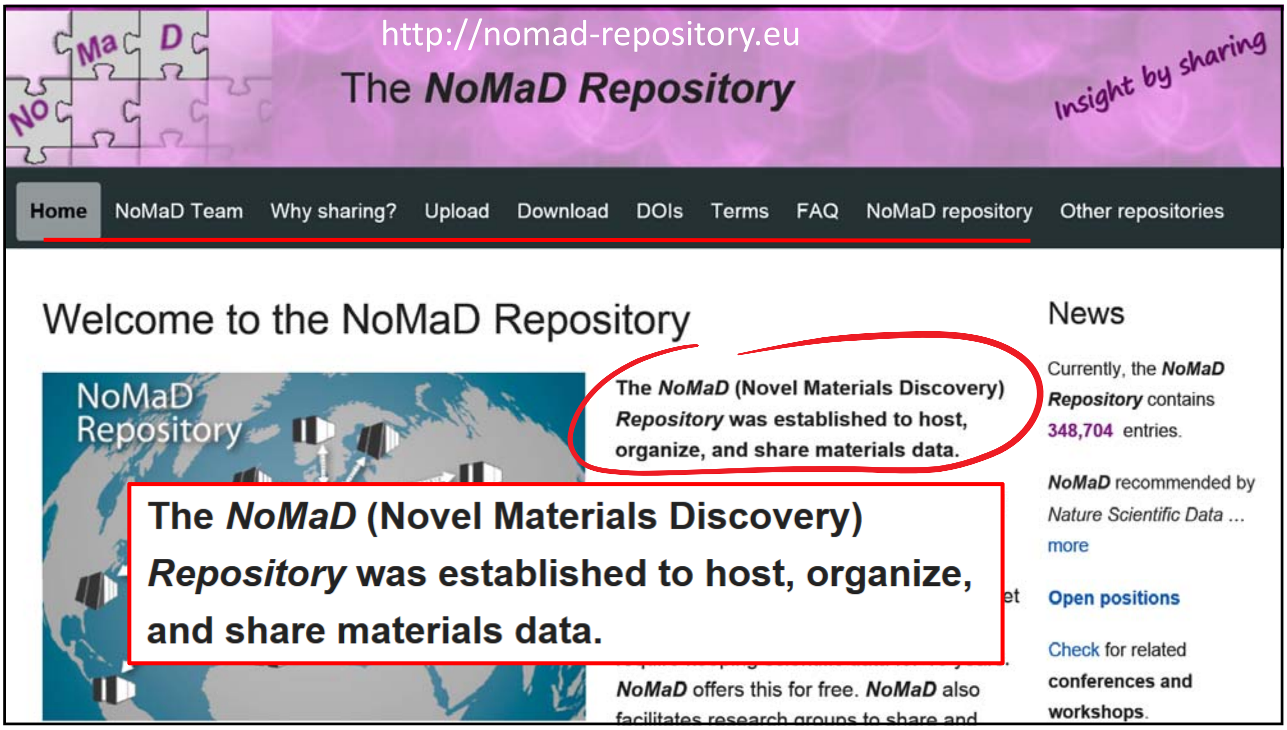
Materials Genome Initiative for Global Competiveness

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"


NoMaD
The Novel Materials
Discovery Repository
https://www.youtube.com/watch?v=L-nmRSH4NQM


http://nomad-repository.eu
The NoMaD Repository
Insight by sharing
Home
NoMaD Team
Why sharing?
Upload
Download
DOIs
Terms
FAQ
NoMaD repository
Other repositories
Welcome to the NoMaD Repository
NoMaD Repository
The NoMaD (Novel Materials Discovery) Repository was established to host, organize, and share materials data.
The NoMaD (Novel Materials Discovery) Repository was established to host, organize, and share materials data.
NoMaD offers this for free. NoMaD also
News
Currently, the NoMaD Repository contains 348,704 entries.
NoMaD recommended by Nature Scientific Data ... more
Open positions
Check for related conferences and workshops.

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.

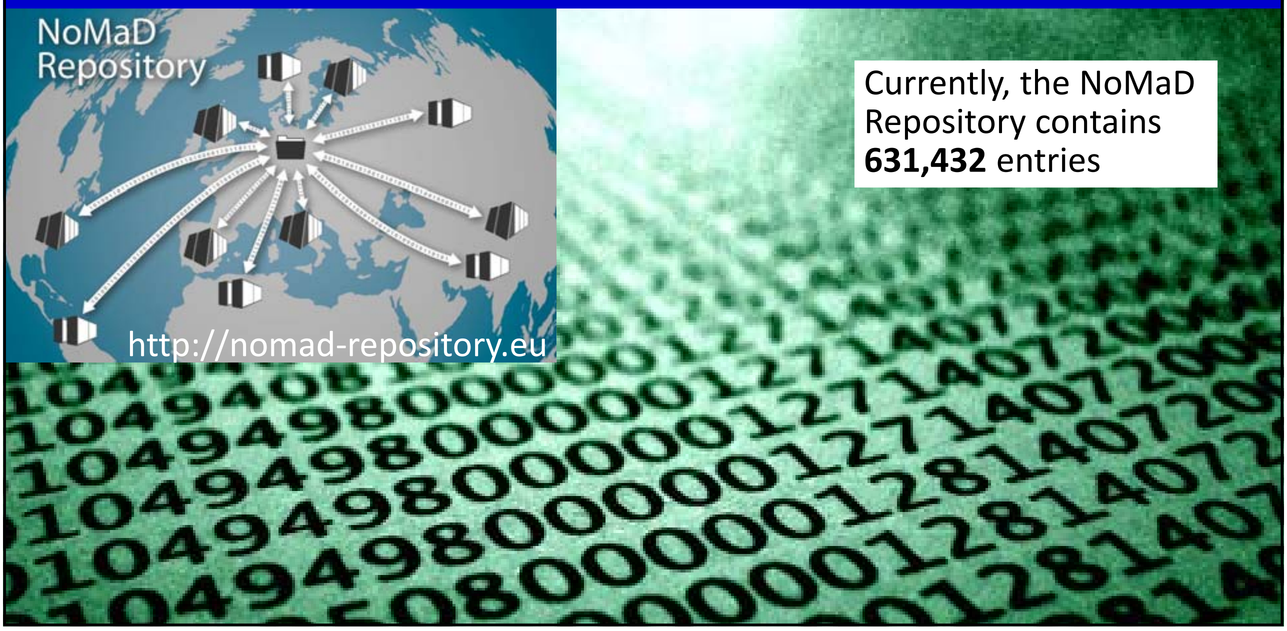
Check for related **conferences and workshops**.

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?



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)

Query and read out what was stored; high-throughput screening.
Shouldn't we do more?!

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)

Query and read out what was stored; high-throughput screening.
Shouldn't we do more?!

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

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.

The Next Step Will Start November 1, 2015: "Novel Materials Discovery (NoMaD) A European Center of Excellence"

A. Bode (Leibniz-Rechenzentrum, Garching)

C. Draxl (HU, Berlin)

D. Frenkel (U. Cambridge)

S. Heinzel (Rechenzentrum Garching MPS)

F. Illas (U. of Barcelona)

K. Koski (CSC – IT Center for Scientific Computing, Helsinki)

J. M. Cela (Barcelona Supercomputing Center)

R. Nieminen (Aalto University, Helsinki)

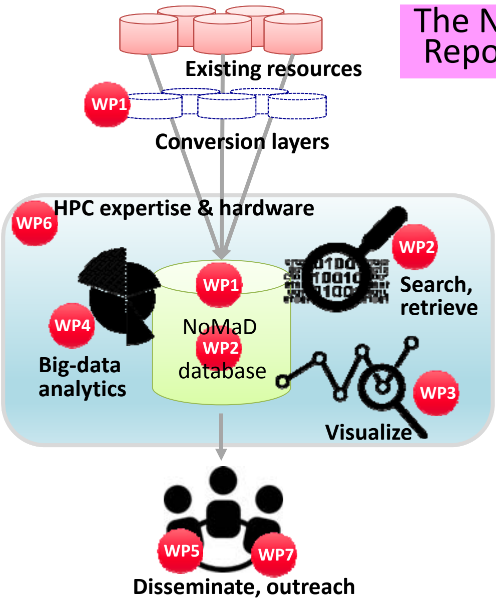
A. Rubio (MPI MPSD, Hamburg)

M. Scheffler (FHI of MPS, Berlin, project coordinator)

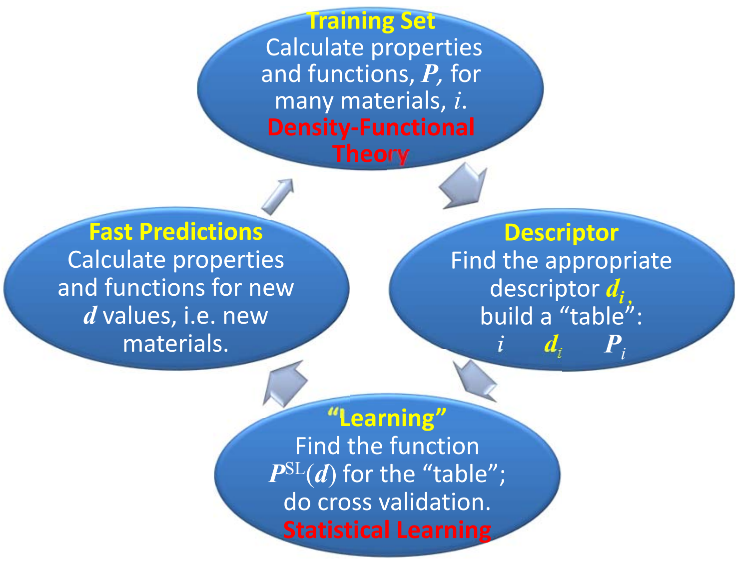
K. Thygesen (Tech. U. Denmark, Lyngby)

A. De Vita (King's College London)

The NoMaD Repository



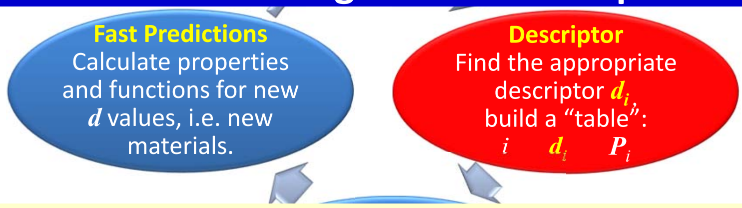
Big-Data Analytics: How to Arrange the Data Finding a Set of Descriptive Parameters



$\{Z_I, N_I\}$, 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



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

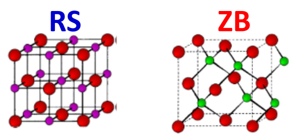
d characterizes the relevant mechanisms that govern the observed property/function P. Identifying the descriptor d from known data P_i, is an ill-posed problem (statistical-learning theory): A little error in the data P_i may suggest a different descriptor d. Thus, knowledge of the accuracy of data P_i is crucial (veracity). The choice of 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 d and P^{SL}?

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?"**



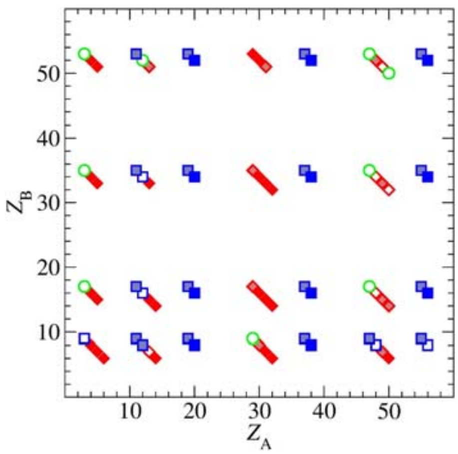
Energy differences between different structures are very small.

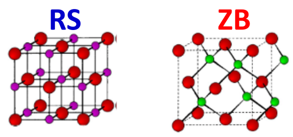
For Si: 0.01% of the energy of a Si atom, or 0.1% of the 4 valence electrons.

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 eV $< \Delta \leq 0.2$ eV
- ZB, 0.05 eV $< \Delta \leq 0.1$ eV
- -0.05 eV $< \Delta \leq 0.05$ eV
- RS, -0.1 eV $< \Delta \leq -0.05$ eV
- RS, -0.2 eV $< \Delta \leq -0.1$ eV
- RS, $\Delta \leq -0.2$ 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.

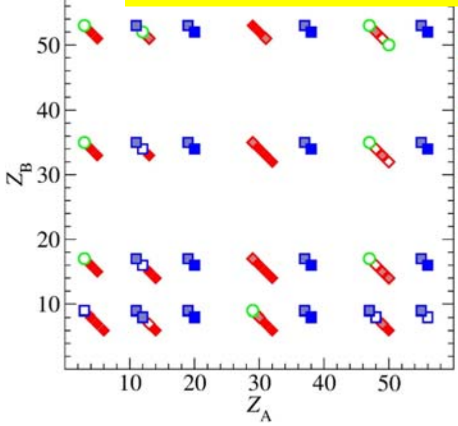
Complexity: $T_s[n]$ and E_{xc}.

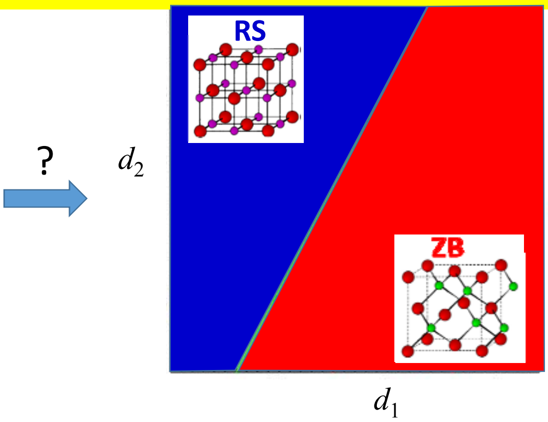
Machine learning can fit the $P(Z_A, Z_B)$ data well, but fails completely in predictions.

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

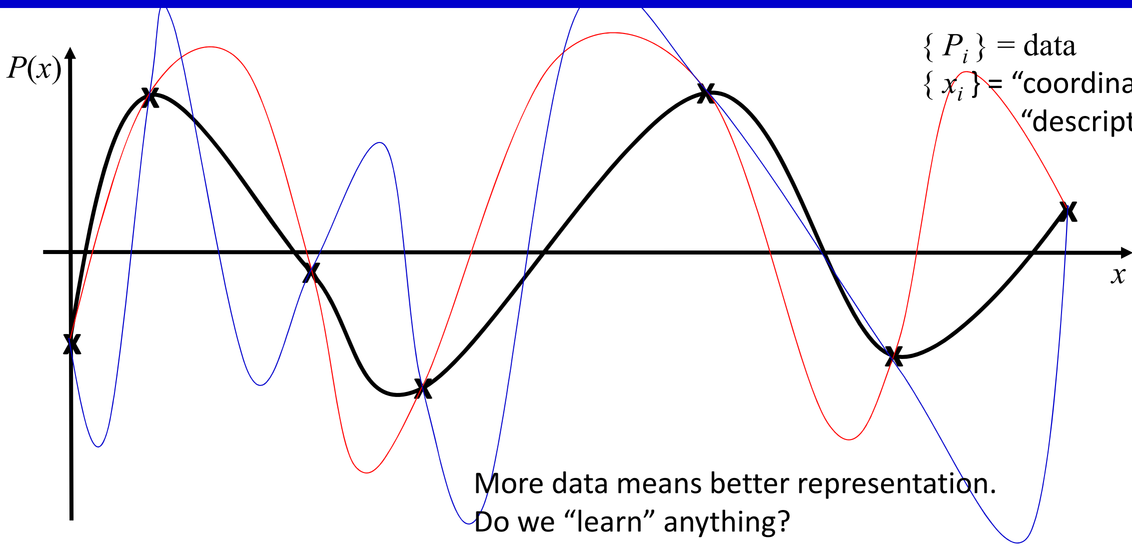
We need to arrange the data such that statistical learning is efficient. We need a good set of descriptive parameters.

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





Data Fitting and Machine Learning



Kernel Regression

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

$$\boldsymbol{P}_i = \boldsymbol{P}(\boldsymbol{x}_i) = \Sigma_{k=1}^{N} \mathrm{c}_k K(\boldsymbol{x}_i, \boldsymbol{x}_k)$$

Linear regression:	$K(\boldsymbol{x}_i, \boldsymbol{x}_k) = \boldsymbol{x}_i \, . \, \boldsymbol{x}_k$	$\boldsymbol{P}(\boldsymbol{x}_i) = \boldsymbol{x}_i \, . \, \boldsymbol{c}^*$
Polynomial kernel	$K(\boldsymbol{x}_i, \boldsymbol{x}_k) = (\boldsymbol{x}_i \, . \, \boldsymbol{x}_k + c)^d$	
Gaussian kernel	$K(\boldsymbol{x}_i, \boldsymbol{x}_k) = \exp \left(- \Sigma_j (\boldsymbol{x}_i - \boldsymbol{x}_k)^2 / 2\sigma_j^2 \right)$	

More data means better representation.
Do we "learn" anything?
For successful learning, we need a "good" descriptor: $\boldsymbol{P}(\boldsymbol{x}_i) \rightarrow \boldsymbol{P}(\boldsymbol{d}_i)$

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 ***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

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

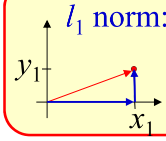
$$\|\mathbf{d}_i - \mathbf{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$$

linear

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

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

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

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

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

linear

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

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

$$\sum_{i=1}^{N}(P(\mathbf{d}_i) - P_i)^2 \; + \; \lambda\|\mathbf{c}\|_1$$

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

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

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

1)

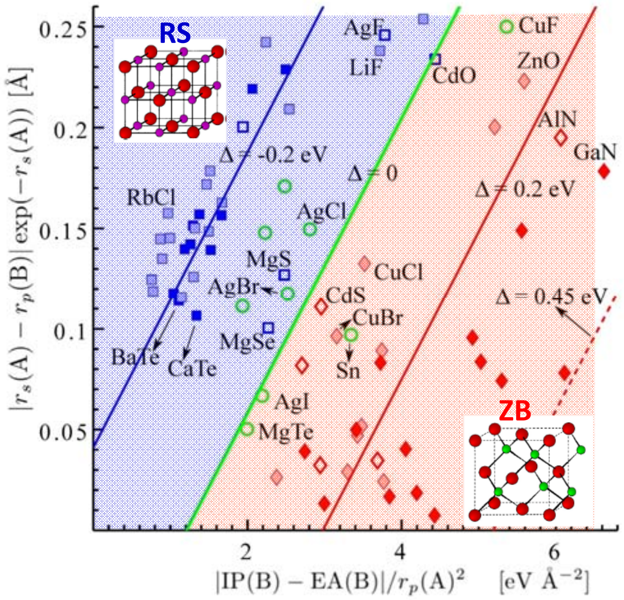
ID	Description **free atoms**	Symbols	#
A1	Ionization Potential (IP) and Electron Affinity (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
A2	Highest occupied (H) and lowest unoccupied (L) Kohn-Sham levels	H(A) L(A) H(B) L(B)	4
A3	Radius at the max. value of s, p, and d valence radial radial probability density	r_s(A) r_p(A) r_d(A) r_s(B) r_p(B) r_d(B)	6

ID	Description **free dimers**	Symbols	#
A4	Binding energy	E_b(AA) E_b(BB) E_b(AB)	3
A5	HOMO-LUMO KS gap	HL(AA) HL(BB) HL(AB)	3
A6	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



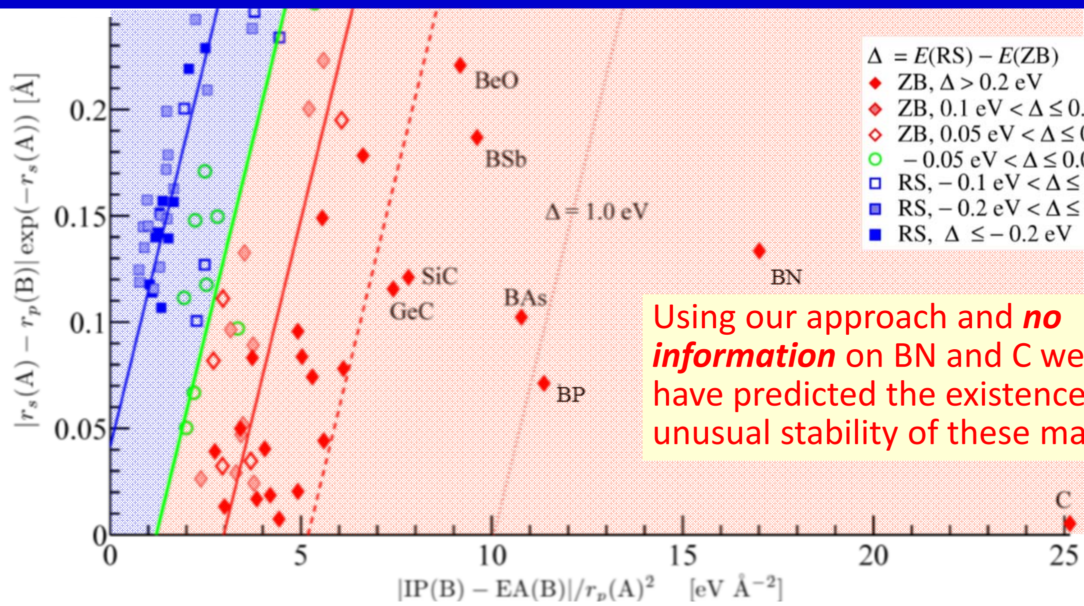
Δ = E(RS) − E(ZB)
ZB, Δ > 0.2 eV
ZB, 0.1 eV < Δ ≤ 0.2 eV
ZB, 0.05 eV < Δ ≤ 0.1 eV
− 0.05 eV < Δ ≤ 0.05 eV
RS, − 0.1 eV < Δ ≤ − 0.05 eV
RS, − 0.2 eV < Δ ≤ − 0.1 eV
RS, Δ ≤ − 0.2 eV

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

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): 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.

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? --

Example:
The probability of childhood leukemia is higher for people living close to electricity power lines.

There is no direct connection between leukemia and the electromagnetic field.

Living close to electric power lines is not a desired residence. People living near power lines tend to be poorer than the control group, and there is a relationship between poverty and cancer.

Poverty → higher probability for living close to power lines

Poverty → higher chances for cancer

? ?

correlation
no direct relation;
intricate causality

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

There is no direct connection between the structure difference and the LASSO-identified descriptor $\frac{\mathrm{IP(B)}-\mathrm{EA(B)}}{r_p(\mathrm{A})^2}, \frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}, \frac{|r_p(\mathrm{B})-r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A})+r_s(\mathrm{B}))}$

Case #2:
Nuclear numbers Z_A, Z_B $\leftrightarrow$ our descriptor

Nuclear numbers Z_A, Z_B $\rightarrow$ 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 ***d*** and ***P*** , i.e. ***P*** is a function of ***d***, ***P***(***d***), reflects causal inference
if it is based on sufficient information(*)

Judea Pearl

There are four possibilities (types of causality) behind ***P***(***d***):

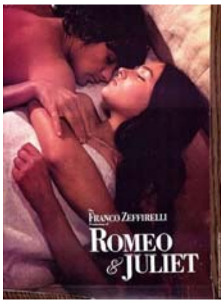
1. $d \rightarrow P$: ***P*** "listens" to ***d***
2. $A \rightarrow d$ and $A \rightarrow P$: There is no direct connection between ***d*** and ***P***, but ***d*** and ***P*** both "listen" to a third "actuator"
3. $P \rightarrow d$: ***d*** "listens" to ***P***
4. There is no direct connection between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

(*) Construct ***d*** with scientific knowledge (prejudice?), or use "big data" for $\{P_i\}$.

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.