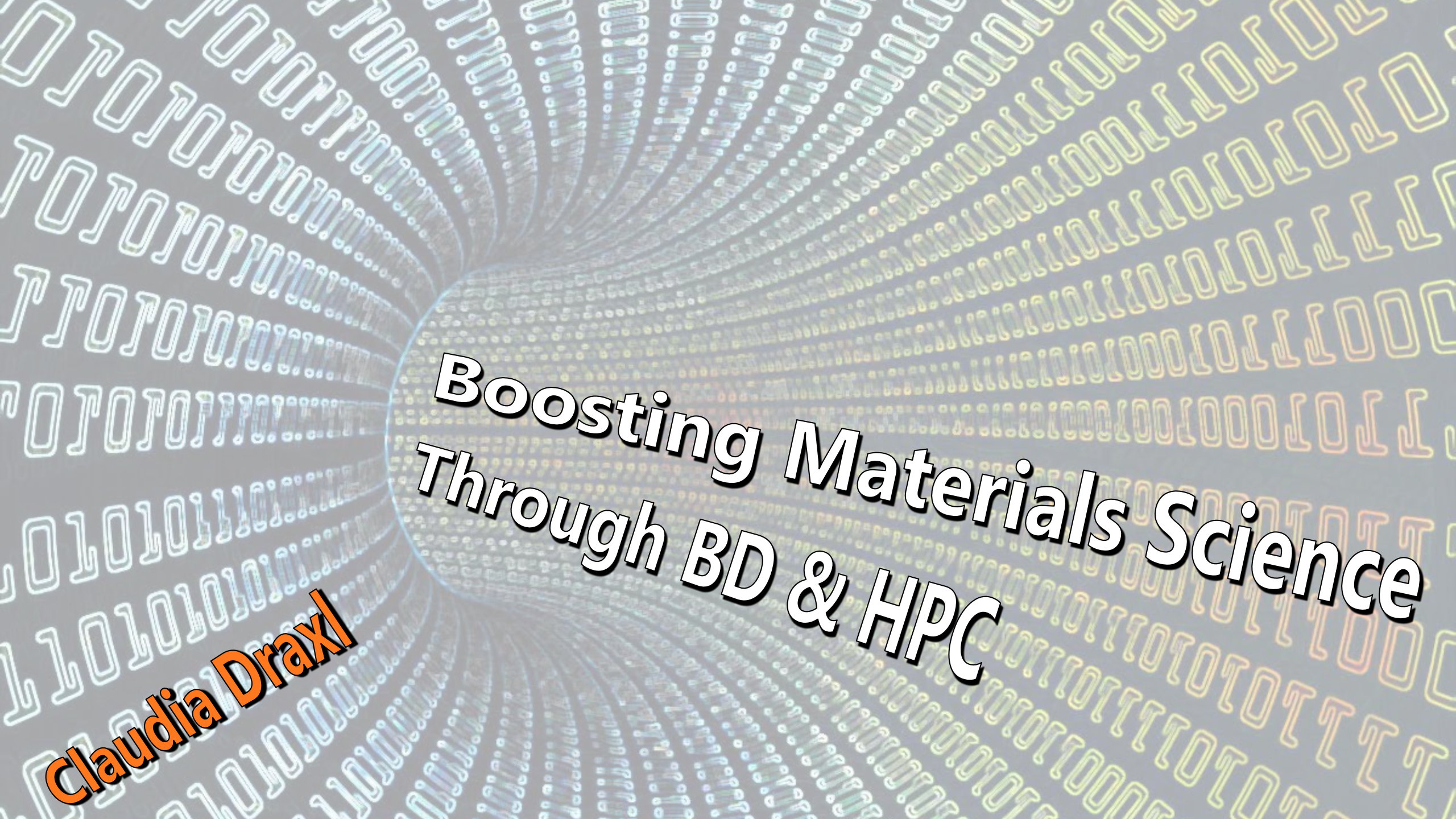
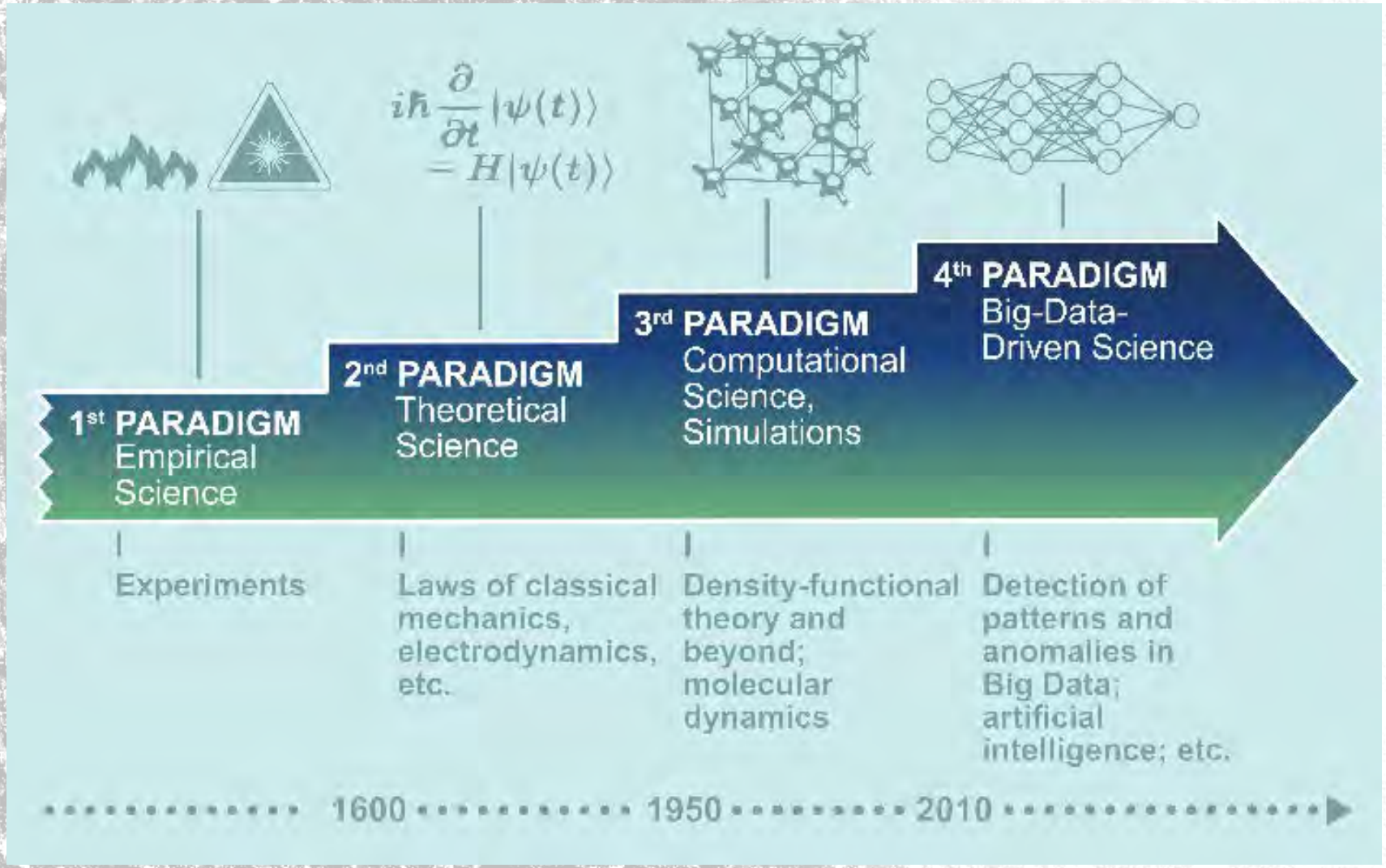
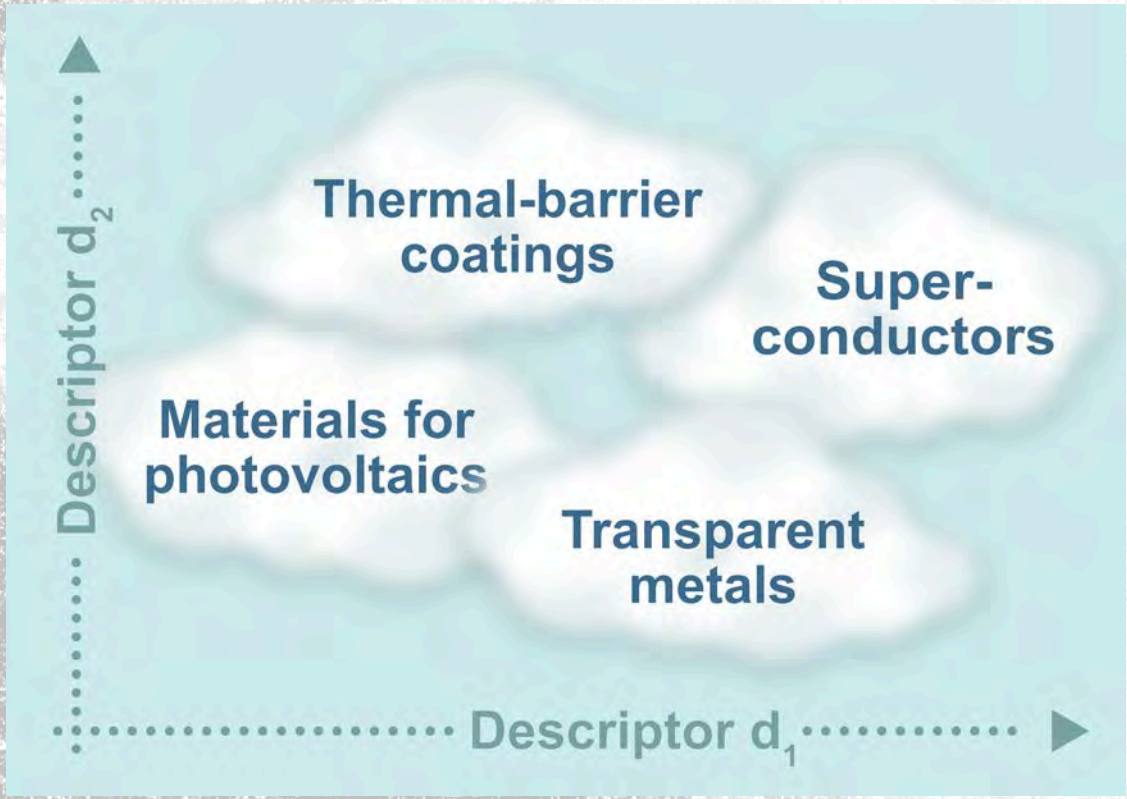

Boosting Materials Science
Through BD & HPC
Claudia Draxl

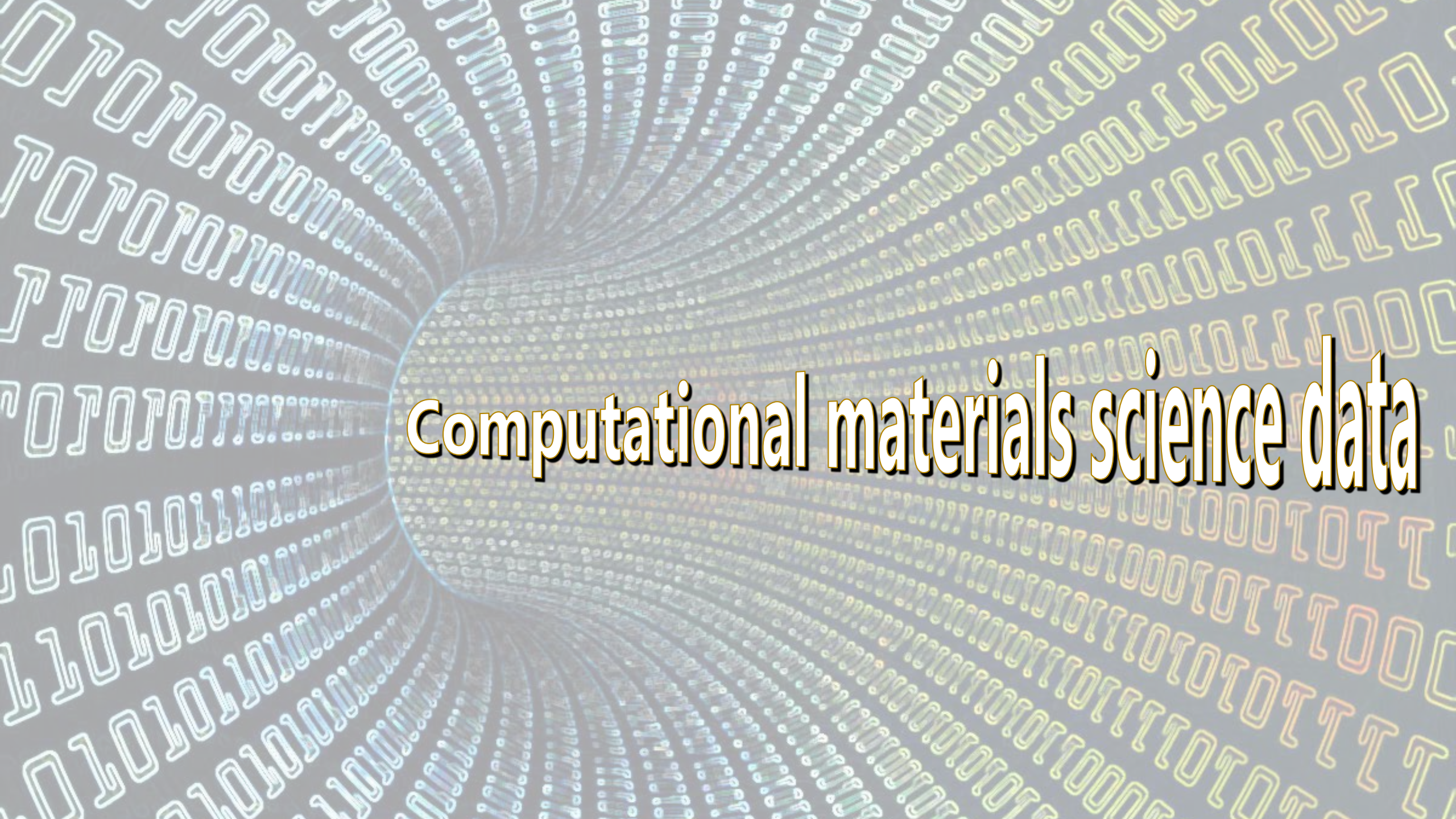
The paradigms of materials science



Our scientific vision is to draw maps

What are the actuators behind the trends and patterns that are invisible to the human eye?




Computational materials science data

What do we need to solve?

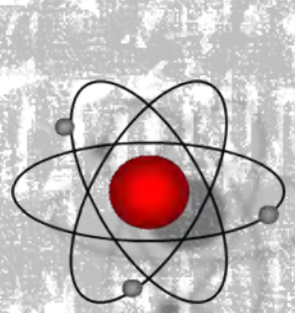
In principle, the Schrödinger equation …

$$\hat{H}\,\psi = E\,\psi$$

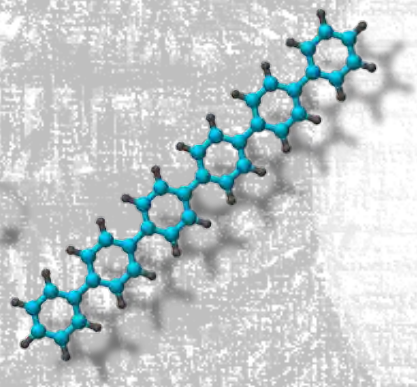
What do we need to solve?

In practice, density-functional theory and beyond …

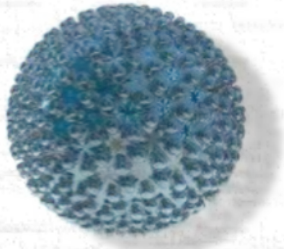
Ab-initio theory for ….



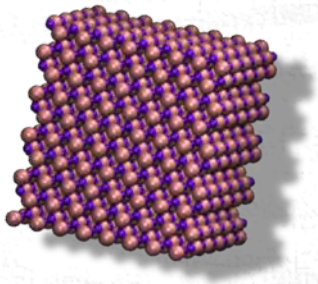
Atoms



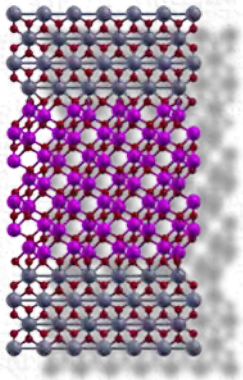
Molecules

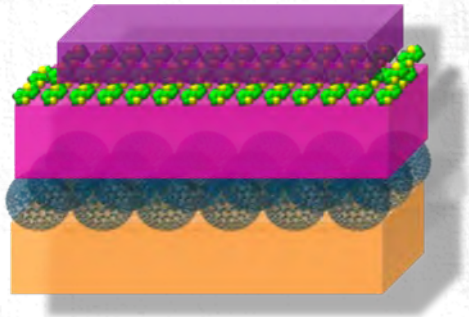


Clusters



Bulk crystals





Surfaces & interfaces

Nanostructures

Materials data and their structure

Level	Properties	Methods	Size
I	Atomic positions and nuclear charges, properties of free atoms, symmetry, tempera	Input: definition of material,	10 kB - 10 MB
II	Total en wavefur geometr		10 MB - 10 TB
III	Excitatic dielectri Coulom phonon spectra, thermal conductivity, etc.	*ab initio* MD	1 GB - 10 TB
IV	Thermoelectric figure of merit, turn-over frequency of catalyst, efficiency of solar cell, etc. as a function of T and P	Modeling, output derived from levels I-III *phenotype*	10 kB - 1 MB

The amount of materials data produced on workstations, compute clusters, and supercomputers is growing exponentially.

Most of it is thrown away ...

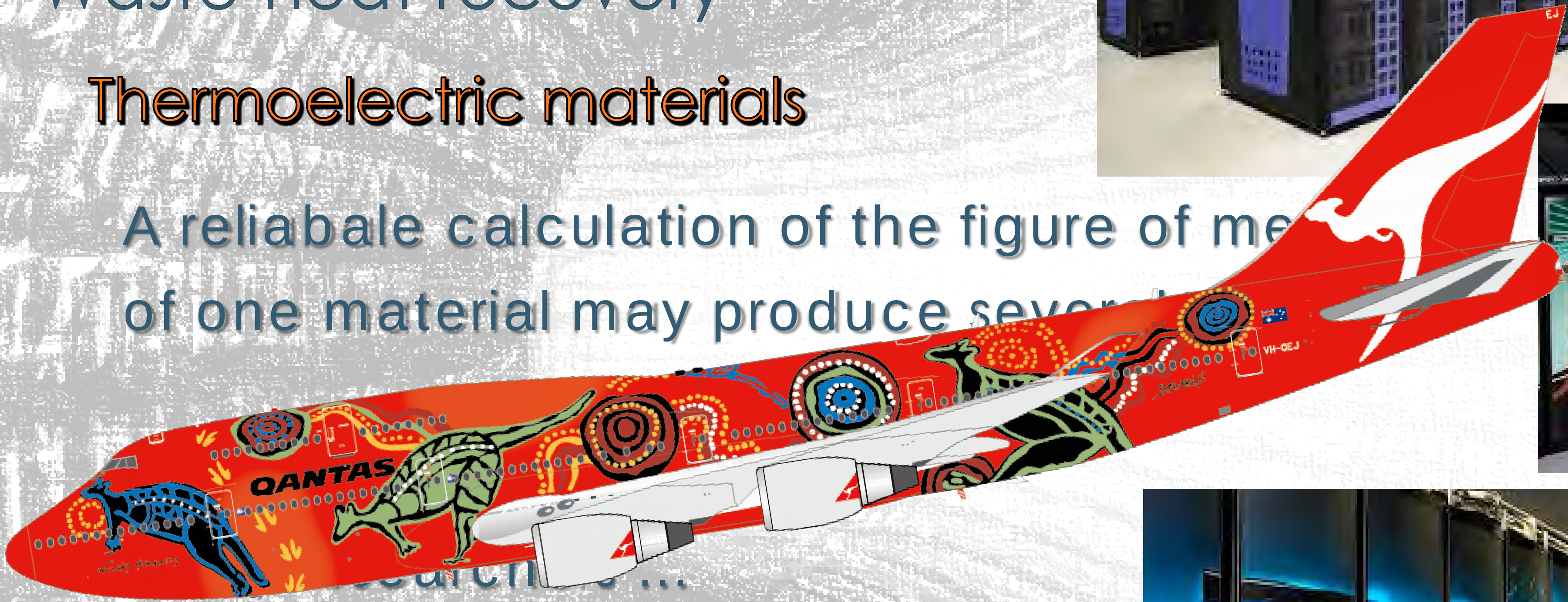
An example ...

Waste-heat recovery

Thermoelectric materials

A reliabale calculation of the figure of me

of one material may produce sever



100 MW

Materials data and their structure

Level	Properties	Methods	Size
I	Atomic positions and nuclear charges, properties of free atoms, symmetry, temperature (T), pressure (P)	Input: definition of material, *gene*	10 kB - 10 MB
II	Total energy, electron density, potential, wavefunctions, atomic forces, optimized geometry, elastic constants, etc.	Density-functional theory (DFT) and *ab initio* molecular dynamics (MD)	10 MB - 10 TB
III	Excitation energies, electrical conductivity, dielectric screening, matrix elements of Coulomb interaction, etc. optical spectra, phonon spectra, thermal conductivity, etc.	Many-body perturbation theory (MBPT), DF perturbation theory, *ab initio* MD	1 GB - 10 TB
IV	Thermoelectric figure of merit, turn-over frequency of catalyst, efficiency of solar cell, etc. as a function of T and P	Modeling, output derived from levels I-III *phenotype*	10 kB - 1 MB

NOvel MAterials Discovery

The FAIR Concept for Big-Data-Driven Materials Science

C. Draxl and M. Scheffler, MRS Bulletin, Sep. 2018

Findable
Accessible
Interoperable
Re-usable

Mark D. Wilkinson, et al.,
Sci. Data **3**, 160018 (2016).

51.786.061 open-access calculations

https://Repository.NOMAD-CoE.eu

The NOMAD Repository

- All input and output files of more than 50 million calculations
  - Raw data
  - Data quality *known*
- Handling requires only few metadata
  - Authors, code & version, upload date, …
- Hosted by Max-Planck Computing and Data Facility
  - Clones being built
- Worldwide largest collection
  - Contains data of US DBs
- Recommended by Scientific Data

The NOMAD Archive

More than 50 million calculations coming from …

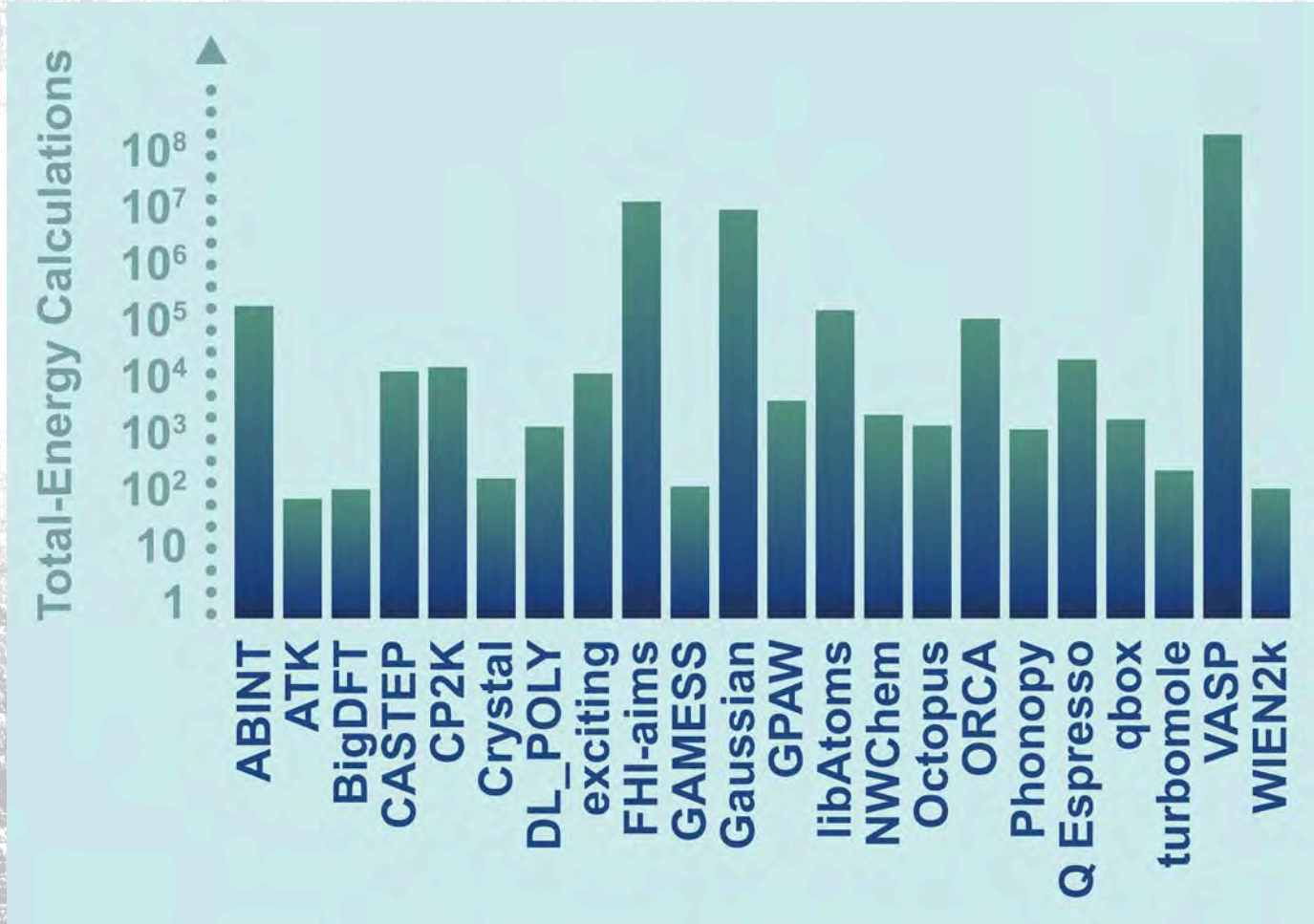
- 40 different codes

Normalized data

- Unified format, units, …

Crucially important

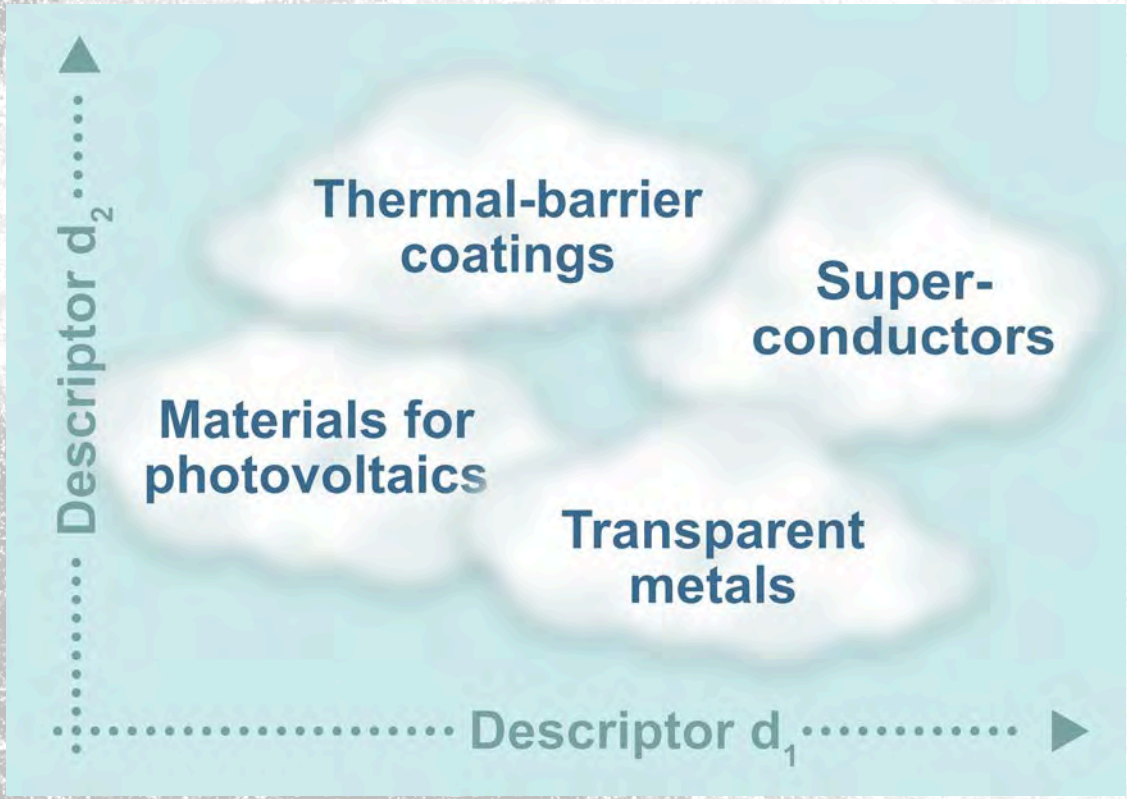
- Metadata
  - Every output is fully parsed




Back to data-driven science ...

Our scientific vision is to draw maps

What are the actuators behind the trends and patterns that are invisible to the human eye?

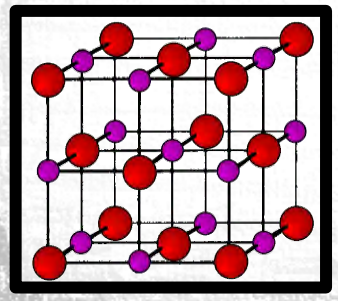


Big-Data analytics - an example

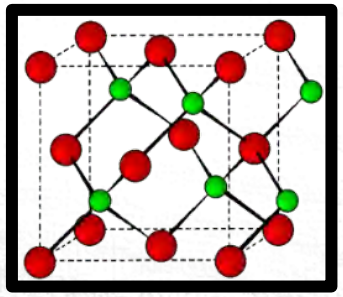
Phillips - van Vechten problem

Given atoms A and B

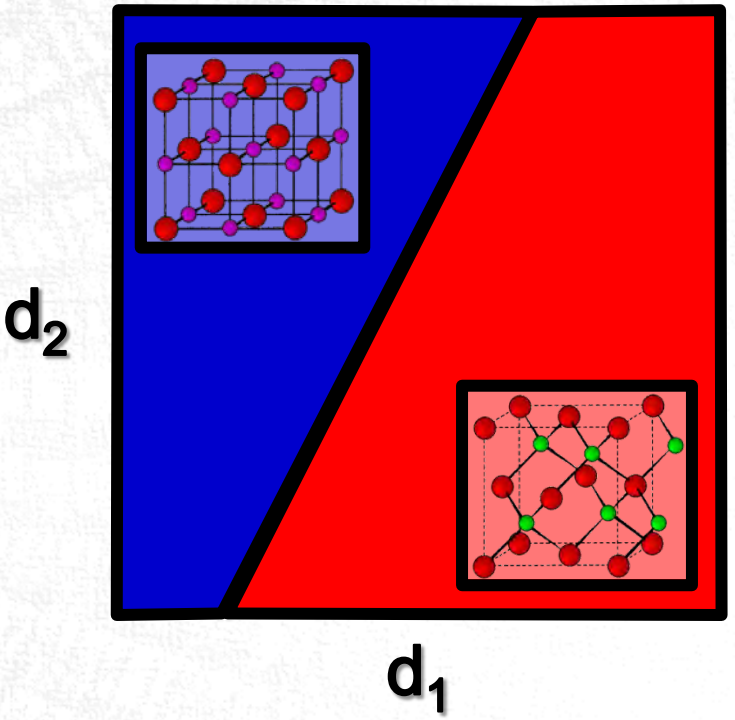
What crystal would they form?



rocksalt



zincblende



L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, CD, and M. Scheffler, PRL 114, 105503 (2015).
L.M. Ghiringhelli, et al., New J. Phys. 19, 023017 (2017).

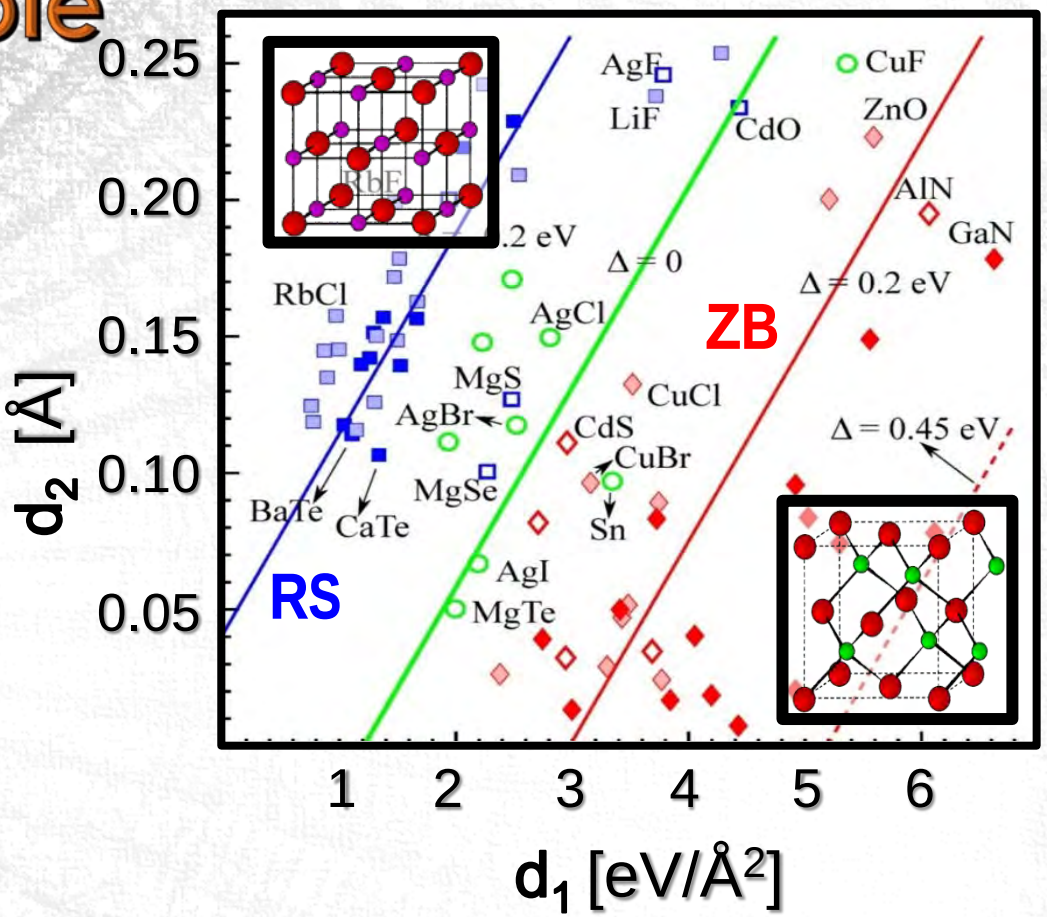
Big-Data analytics - an example

Structure map of binary semiconductors, obtained with a compressed-sensing algorithm

Predictions from free neutral atoms A and B

Results can be reenacted at the NOMAD Anlytics Toolkit

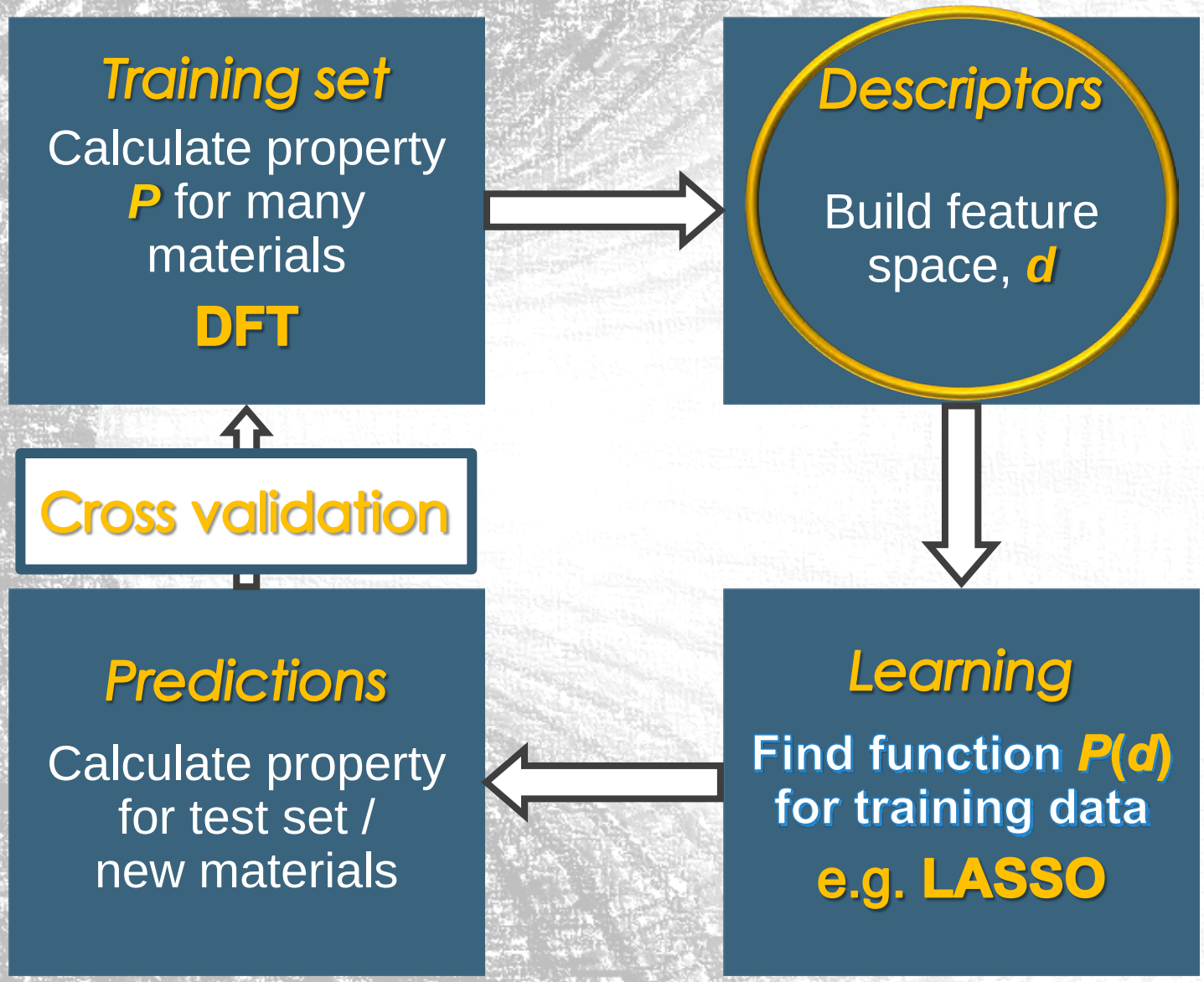
https://analytics-toolkit.nomad-coe.eu/



$$d_1 = \frac{IP(A) - EA(B)}{r_p(A)^2} \qquad d_2 = \frac{|r_s(A) - r_p(B)|}{\exp[rp(A)]}$$

L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, CD, and M. Scheffler, PRL 114, 105503 (2015).
L.M. Ghiringhelli, et al., New J. Phys. 19, 023017 (2017).

Model building



Building descriptors

Primary features

Free atoms

IP(A), IP(B)	Ionization potential
EA(A), EA(B)	Electron affinity
H(A), H(B)	Highest occupied Kohn-Sham level
L(A), L(B)	Lowest unoccupied Kohn-Sham level
r_s(A), r_s(B)	Radius at max. of *s*-like wavefunction
r_p(A), r_p(B)	Radius at max. of *p*-like wavefunction
r_d(A), r_d(B)	Radius at max. of *d*-like wavefunction

Dimers

HL(AA), HL(BB), HL(AB)	HOMO-LUMO KS gap
E_b(AA), E_b(BB), E_b(AB)	Binding energy
d(AA), d(BB), d(AB)	Equilibrium distance

Building descriptors

Full feature space

10 000 **nonlinear** combinations of primary features

+, -, *, /, ^{2}, ^{3}, √, exp, …

Linear relationship $P(d)=c\,d$

Let the machine choose most relevant descriptors

$$\min_{c \in R^M} \|P - c\,d\|^2 + \lambda\|c\|_1$$


Beyond computational materials research

Challenge - experimental (meta) data

What characterizes the sample?

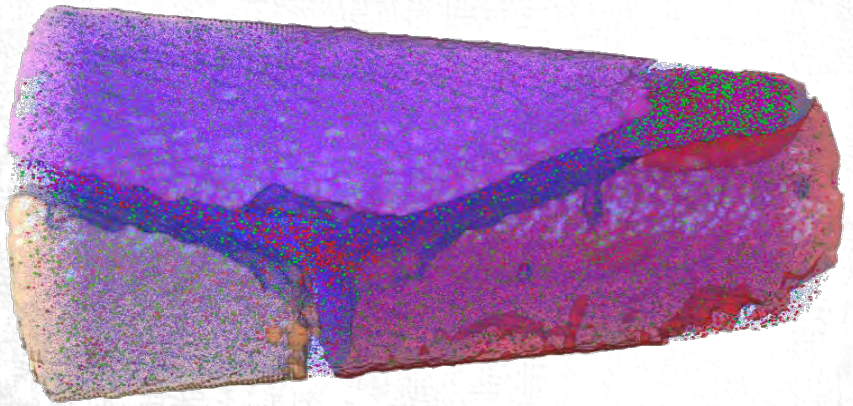
- Composition, history, treatment, ...
  - unique sample and specimen identifiers

What characterizes the apparatus?

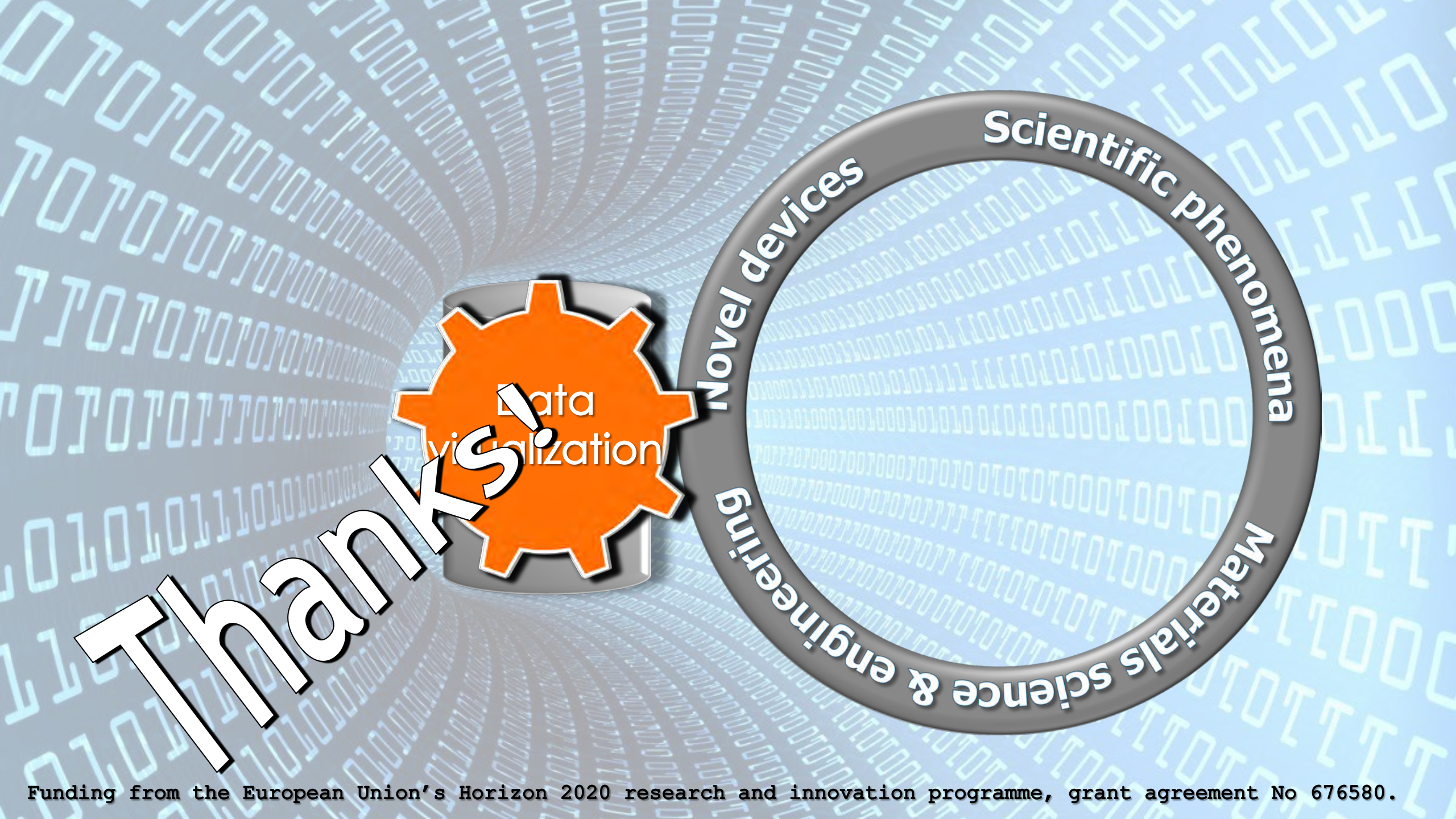
- Spatial & energetic resolution, ...

What characterizes the measurement?

- Temperature, pressure, ...



Courtesy Babtiste Gault & Dierk Raabe



Funding from the European Union's Horizon 2020 research and innovation programme, grant agreement No 676580.