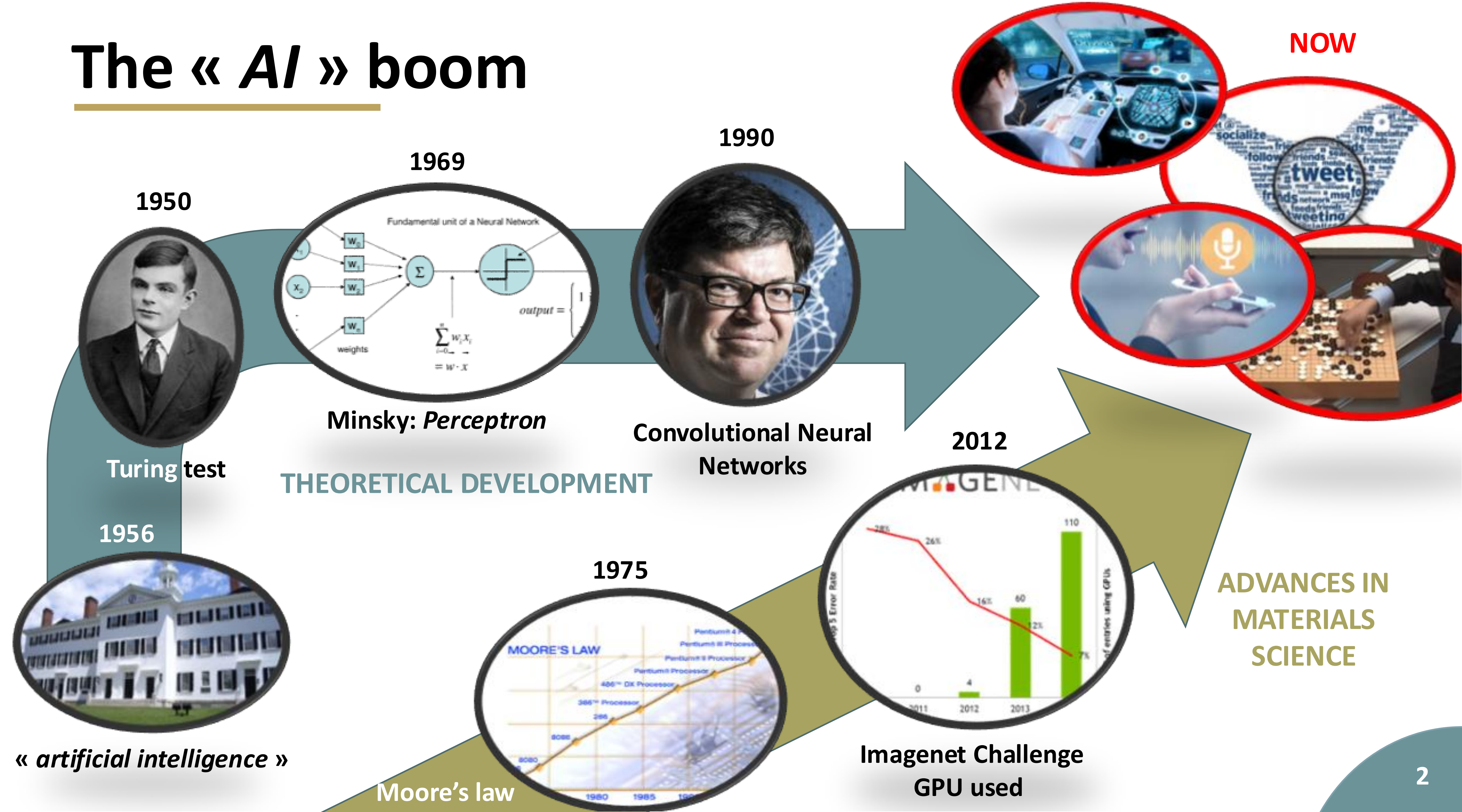
Initiation à l'apprentissage automatique en science des matériaux

1. Introduction

J.-C. Crivello, LINK : jean-claude.crivello@cnrs.fr
C. Barreteau, ICMPE : celine.barreteau@cnrs.fr
S. Junier, ICMPE : sebastien.junier@cnrs.fr

https://link.cnrs.fr/ML/

The « *AI* » boom

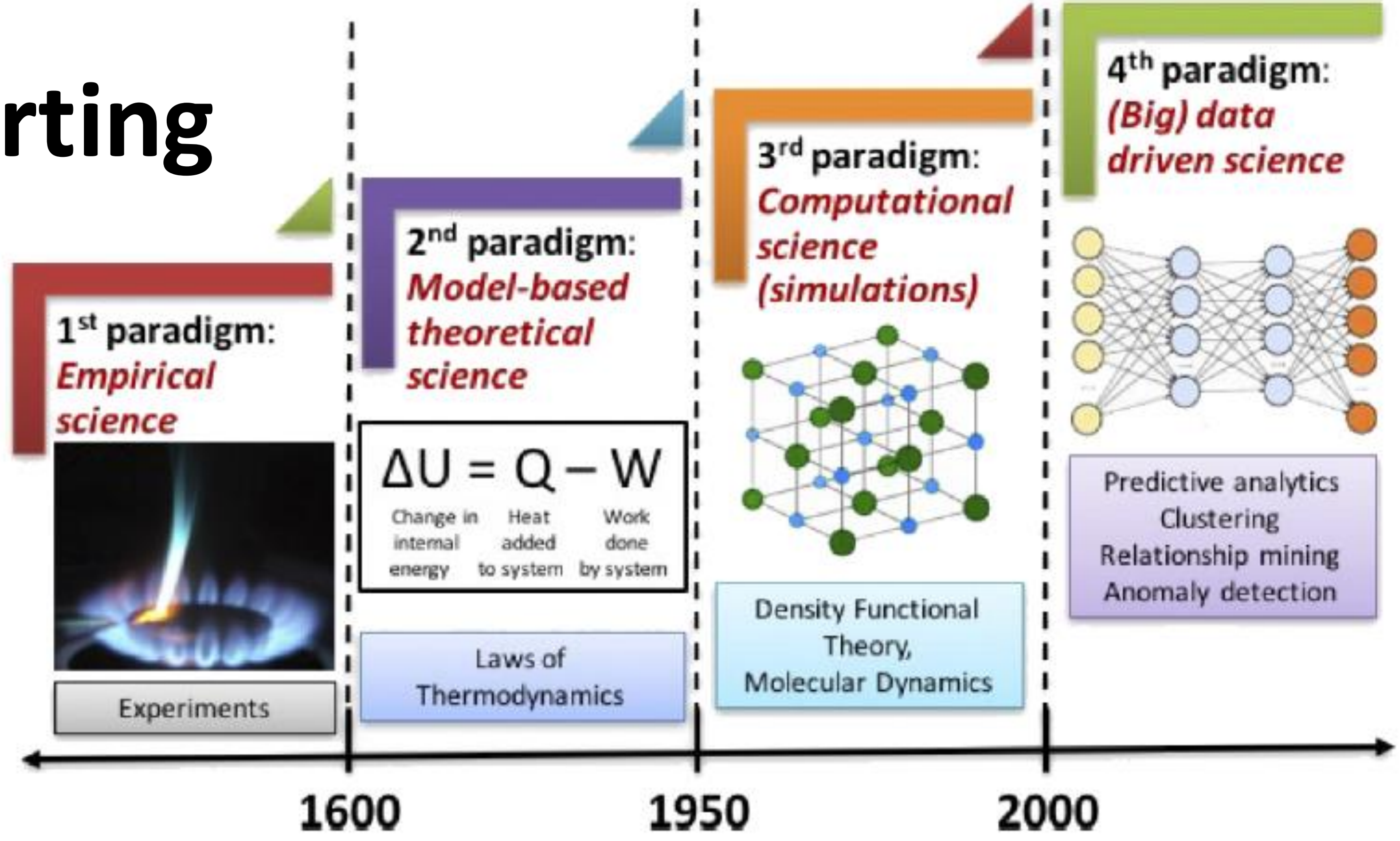

NOW
1950
Turing test
1956
« artificial intelligence »
1969
Fundamental unit of a Neural Network
Minsky: Perceptron
1990
Convolutional Neural Networks
THEORETICAL DEVELOPMENT
1975
MOORE'S LAW
Moore's law
2012
Imagenet Challenge
GPU used
ADVANCES IN MATERIALS SCIENCE

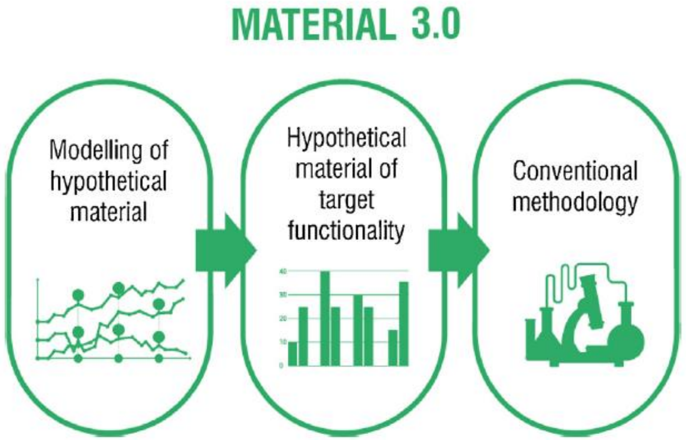
A revolution is starting

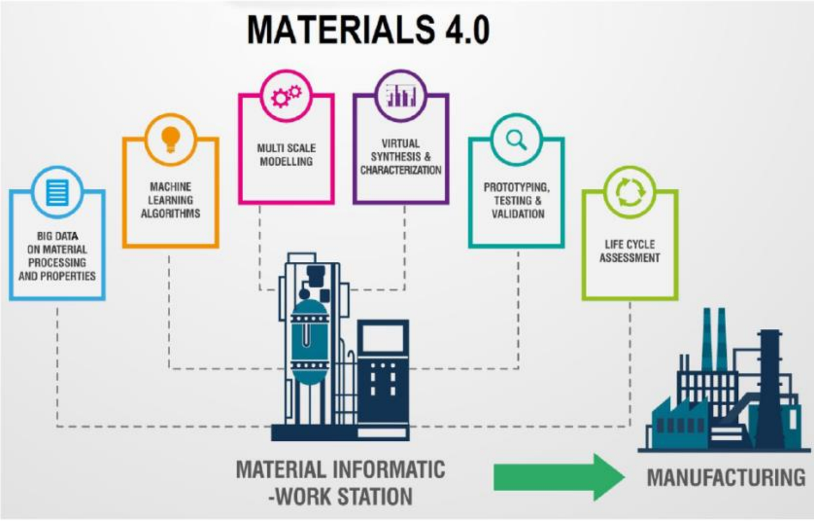
J. Gray (2007)
4th paradigm is Science

Agrawal *et al*. APL Materials (2016)
Perspective: Materials informatics and big data: realization of the « fourth paradigm » of science in materials science

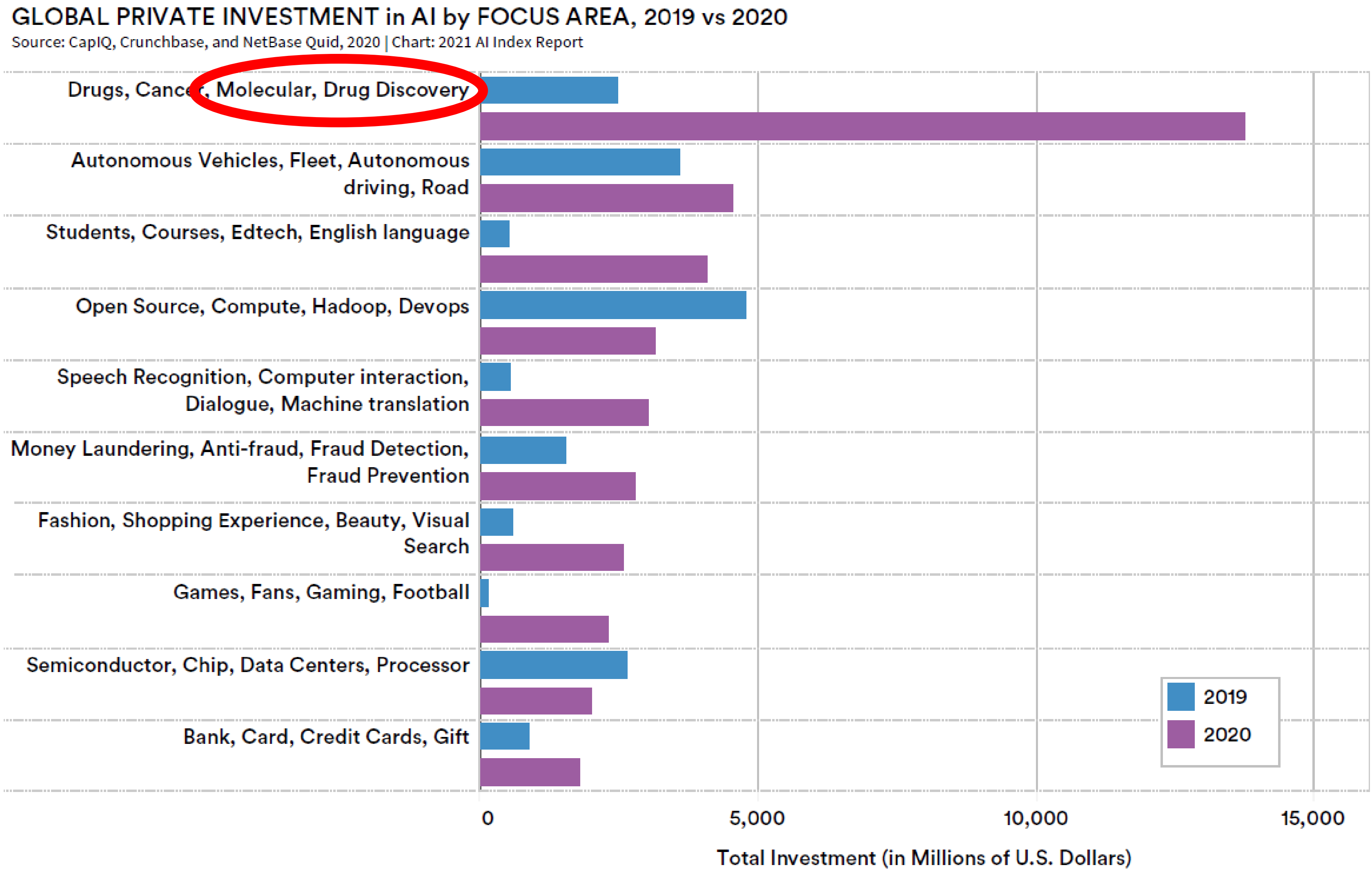
Jose *et al*. Applied materials today (2018)
Materials 4.0: Materials big data enabled materials discovery







AI in chemistry ?

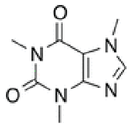


OECD

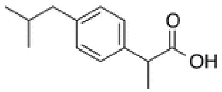
Artificial Intelligence Index Report 2021

AI in chemistry

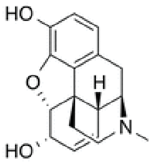
Segler *et al*. ACS Central Science (2018)
Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks



Caffeine
CN1c2ncn(C)c2C(=O)N(C)C1=O

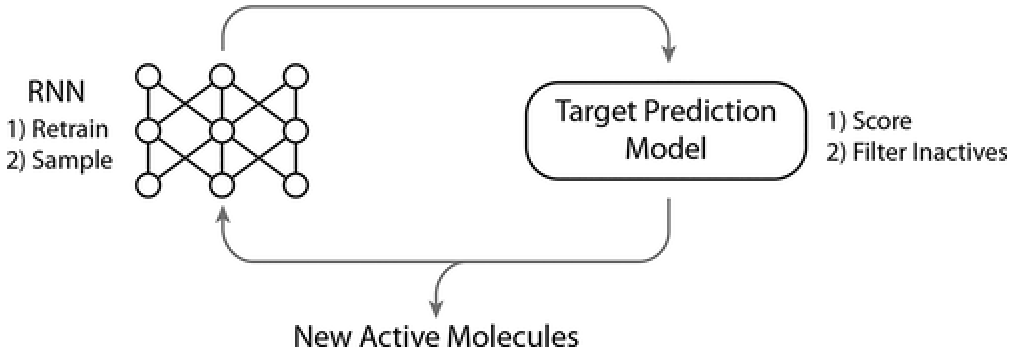


Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(O)=O

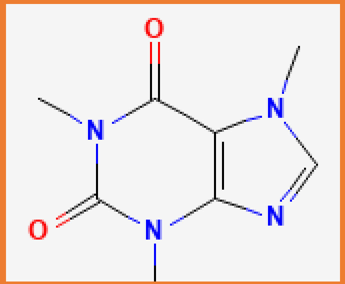


Morphine
[H][C@]12C=C[C@H](O)[C@@H]3Oc4c5c(C[C@H]1N(C)CC[C@@]235)ccc4O

Batch	Generated Example	valid
0	Oc.BK5i%ur+7oAFc7L3T=F8B5e=n)CS6RCTAR((OVCp1CApb)	no
1000	OF=CCC2OCCCC)C2)C1CNC2CCCCCCCCCCCCCCCCCCCCCCC	no
2000	O=C(N)C(=O)N(c1occc1OC)c2ccccc2OC	yes
3000	O=C1C=2N(c3cc(ccc3OC2CCC1)CCCc4cn(c5c(Cl)cccc54)C)C	yes



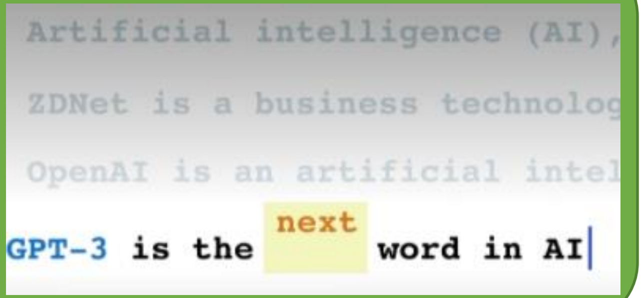
SMILE Format



CN1C=NC2=C1C(=O)N(C(=O)N2C)C

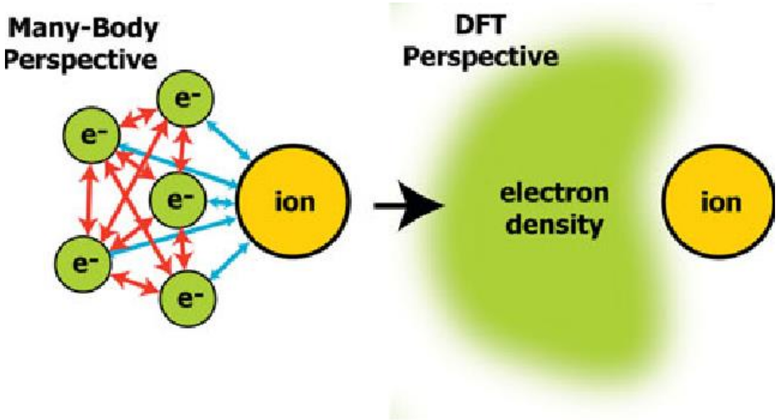
Natural Language Processing

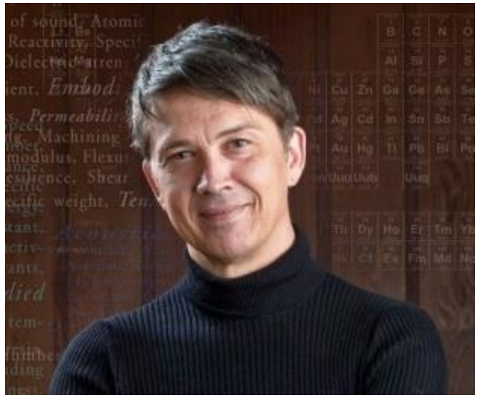
GPT-3 Generative algorithm



USA : « computational material science »

N. Nosenko. Nature (2016)
The material code





2003 : G. Ceder

2011 : Materials Genome Initiative

2011-2016 : $500 million

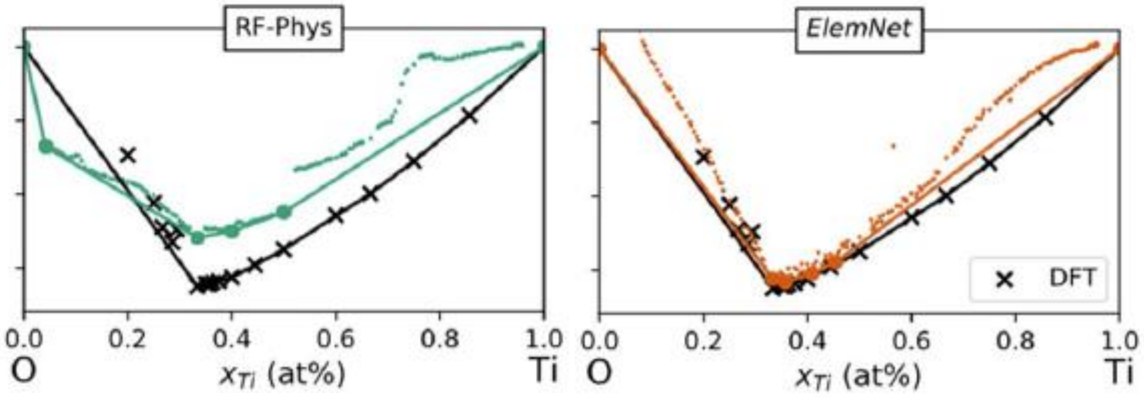
From *big data* to *machine learning*

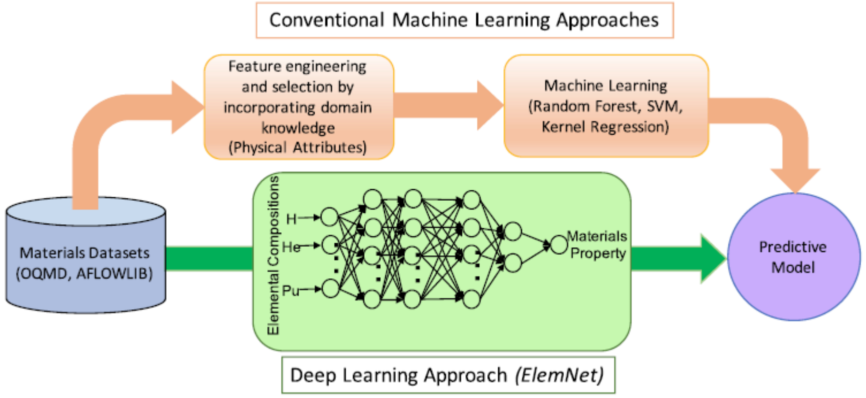
K. Butler *et al.* Nature (2018)
Machine learning for molecular and materials science

Table 3 | Publicly accessible structure and property databases for molecules and solids

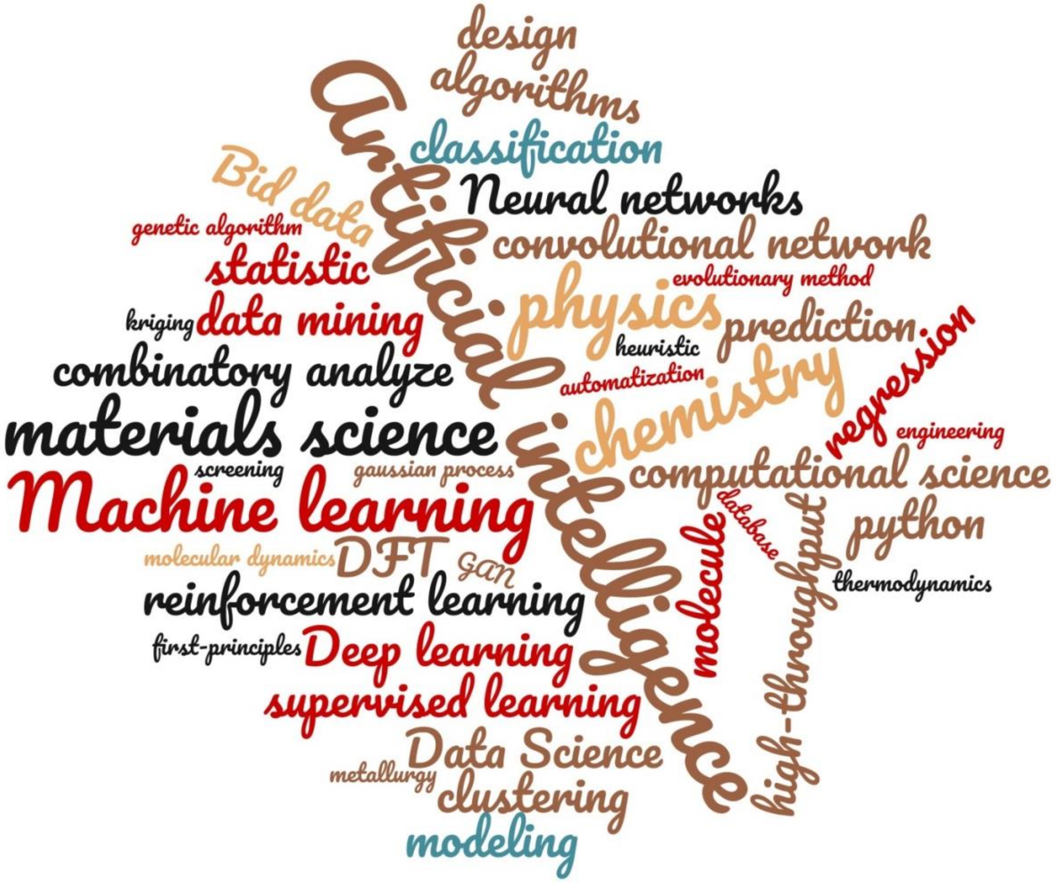
Name	Description	URL
Computed structures and properties		
AFLOWLIB	Structure and property repository from high-throughput ab initio calculations of inorganic materials	http://aflowlib.org
Computational Materials Repository	Infrastructure to enable collection, storage, retrieval and analysis of data from electronic-structure codes	https://cmr.fysik.dtu.dk
GDB	Databases of hypothetical small organic molecules	http://gdb.unibe.ch/downloads
Harvard Clean Energy Project	Computed properties of candidate organic solar absorber materials	https://cepdb.molecularspace.org
Materials Project	Computed properties of known and hypothetical materials carried out using a standard calculation scheme	https://materialsproject.org
NOMAD	Input and output files from calculations using a wide variety of electronic-structure codes	https://nomad-repository.eu
Open Quantum Materials Database	Computed properties of mostly hypothetical structures carried out using a standard calculation scheme	http://oqmd.org
NREL Materials Database	Computed properties of materials for renewable-energy applications	https://materials.nrel.gov
TEDesignLab	Experimental and computed properties to aid the design of new thermo-electric materials	http://tedesignlab.org
ZINC	Commercially available organic molecules in 2D and 3D formats	https://zinc15.docking.org
Experimental structures and properties		
ChEMBL	Bioactive molecules with drug-like properties	https://www.ebi.ac.uk/chembl
ChemSpider	Royal Society of Chemistry's structure database, featuring calculated and experimental properties from a range of sources	https://chemspider.com
Citrination	Computed and experimental properties of materials	https://citrination.com
Crystallography Open Database	Structures of organic, inorganic, metal–organic compounds and minerals	http://crystallography.net
CSD	Repository for small-molecule organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk
ICSD	Inorganic Crystal Structure Database	https://icsd.fiz-karlsruhe.de
MatNavi	Multiple databases targeting properties such as superconductivity and thermal conductance	http://mits.nims.go.jp
MatWeb	Datasheets for various engineering materials, including thermoplastics, semi-conductors and fibres	http://matweb.com
NIST Chemistry WebBook	High-accuracy gas-phase thermochemistry and spectroscopic data	https://webbook.nist.gov/chemistry
NIST Materials Data Repository	Repository to upload materials data associated with specific publications	https://materialsdata.nist.gov
PubChem	Biological activities of small molecules	https://pubchem.ncbi.nlm.nih.gov

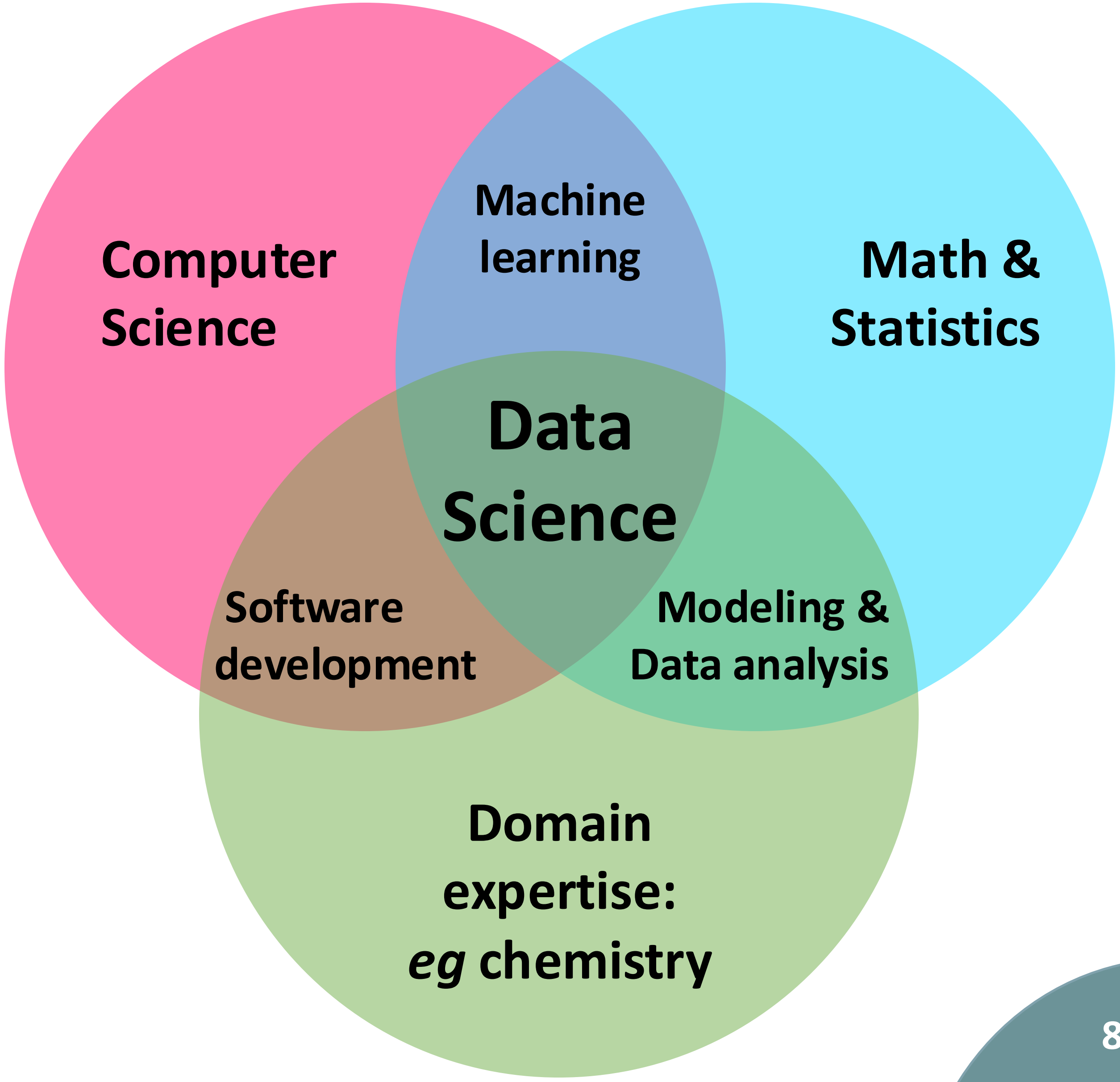
Jha, Wolverton, Agrawal *et al*. Scientific Report (2018)
ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition



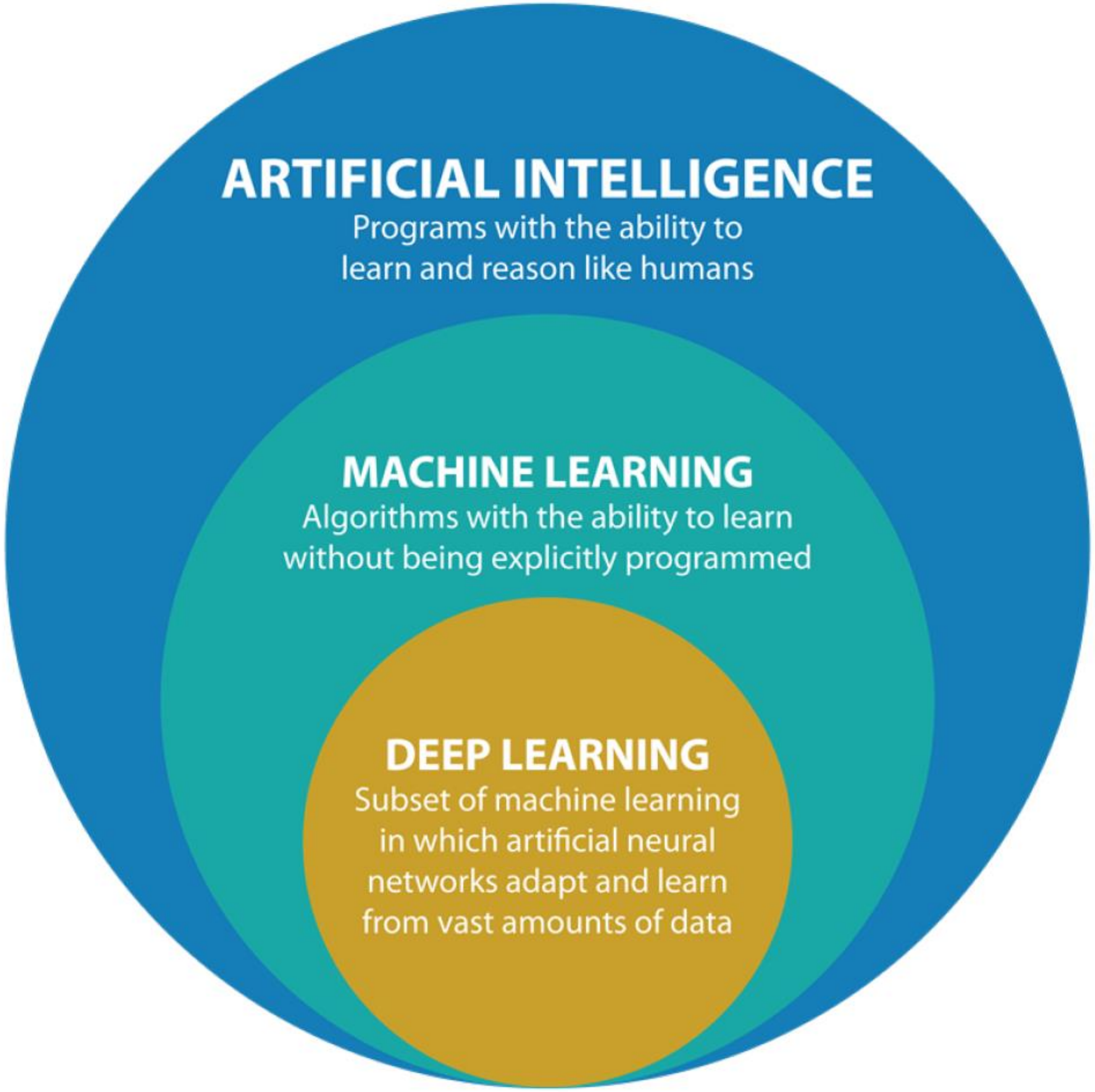


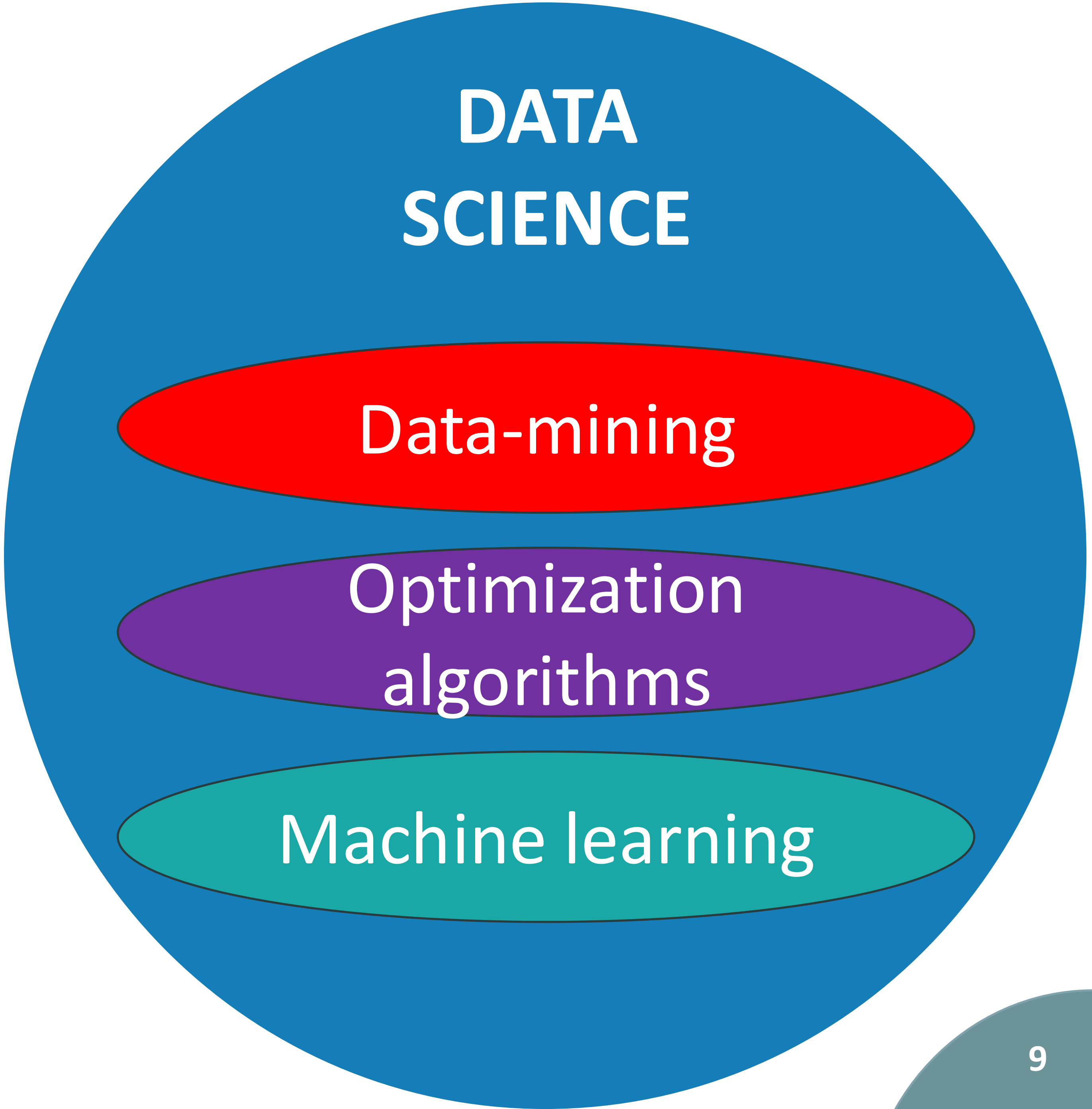
Data science: an interdisciplinary field



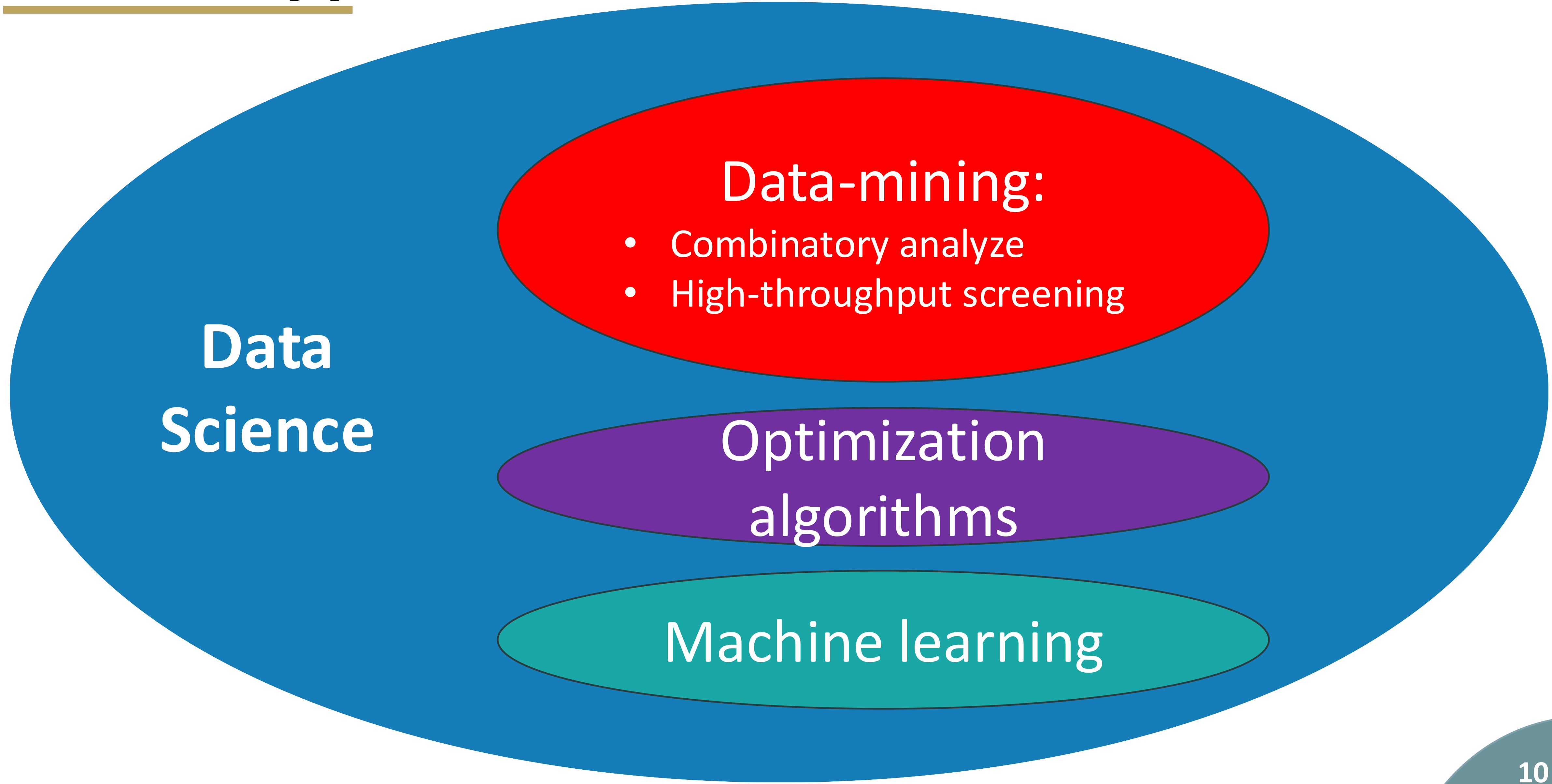


General Pictures





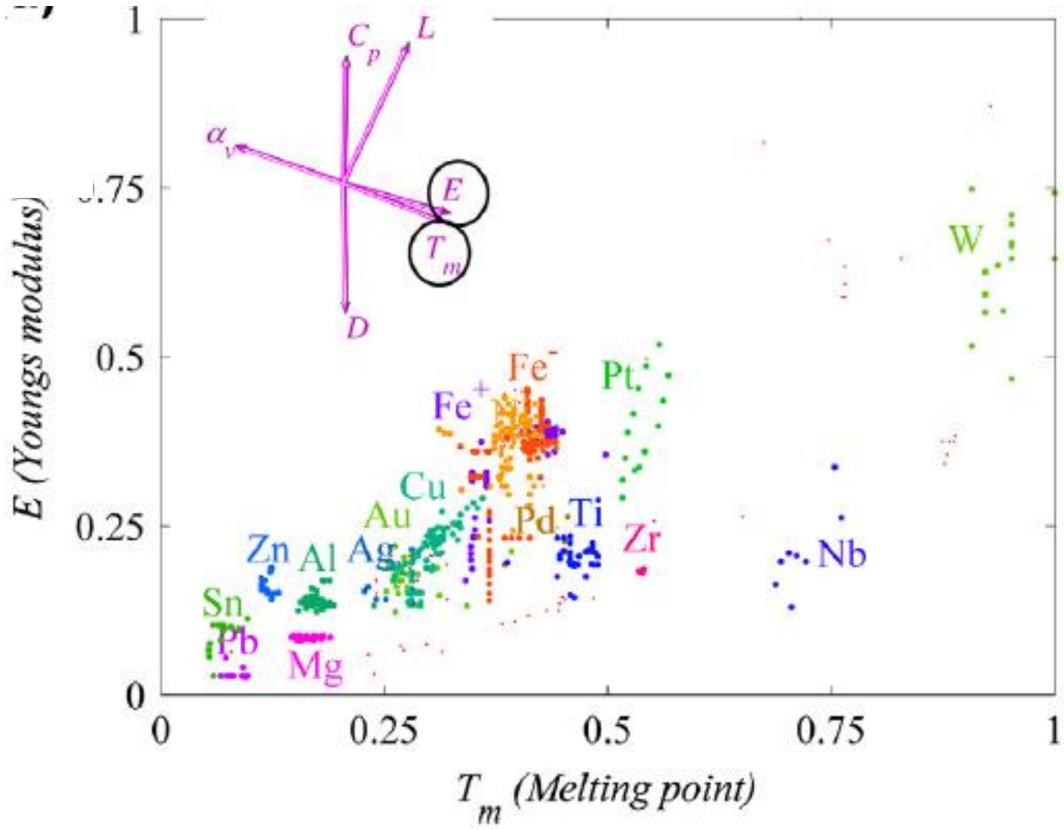
Several approaches

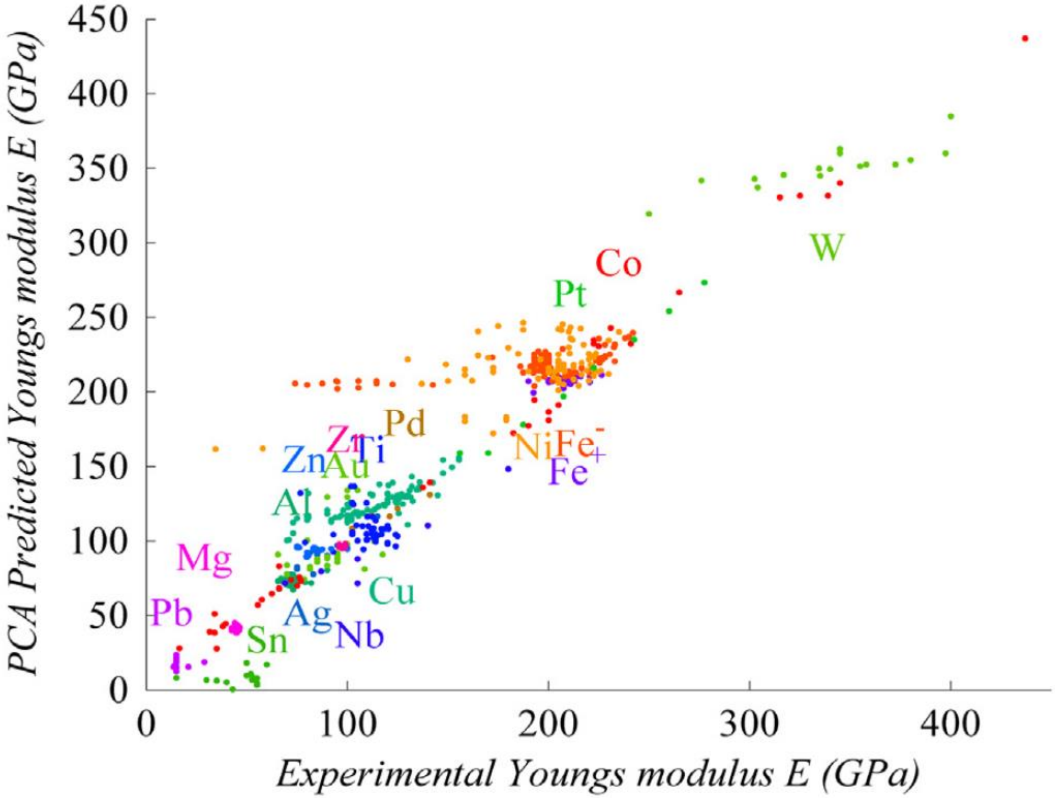


Data-mining algorithms

> Combinatory analyze

Toda-Caraballo *et al.*
JALCOM (2013)
Unravelling the materials genome: Symmetry relationships

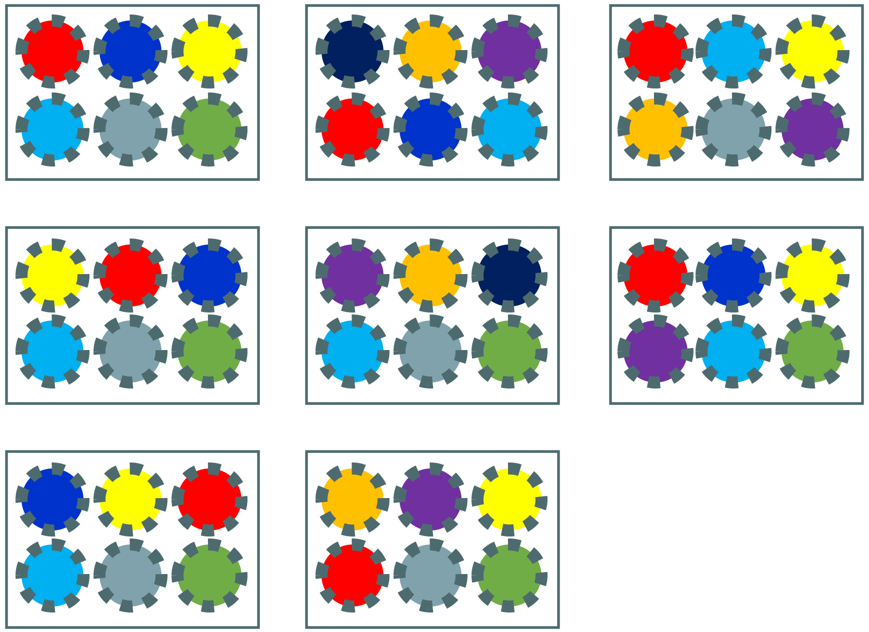




e.g. **Principal Component Analysis (PCA)**

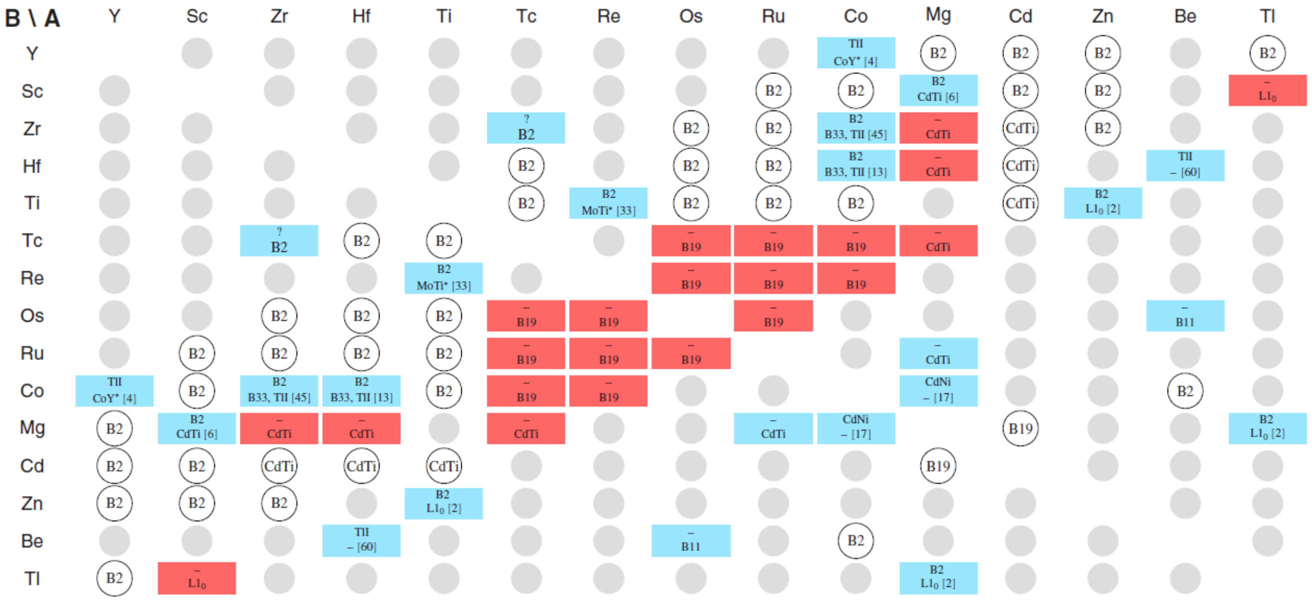
Data-mining algorithms

> High-throughput screening



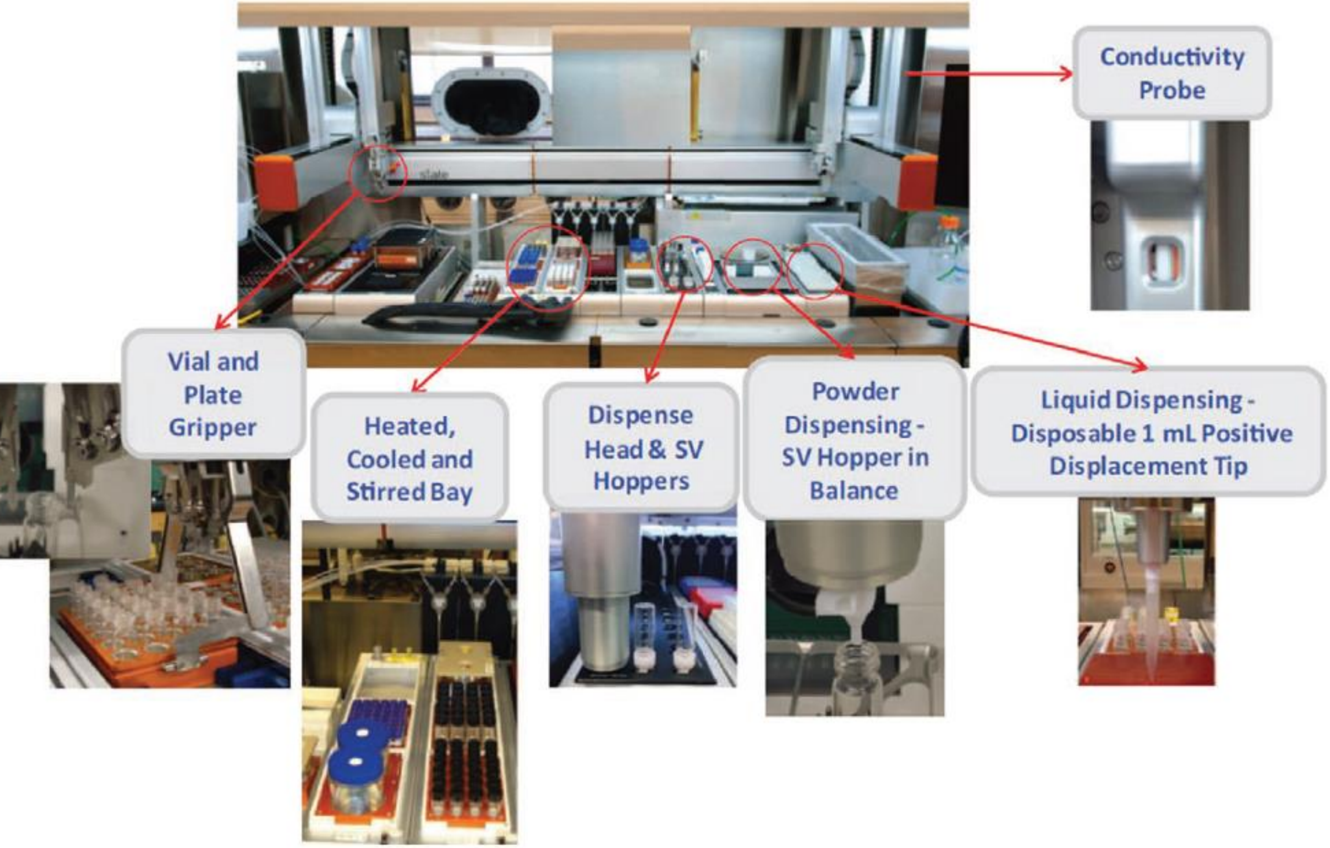
Levy, Hart, Curtarolo.
Phys. Rev. B (2010)
Structure maps for hcp metals from first-principles calculations

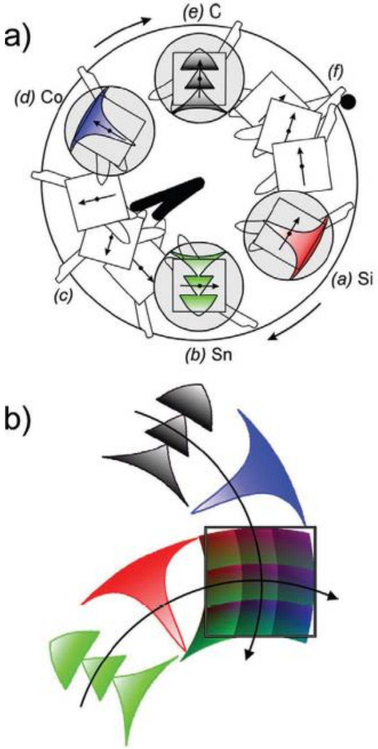
http://aflowlib.org/

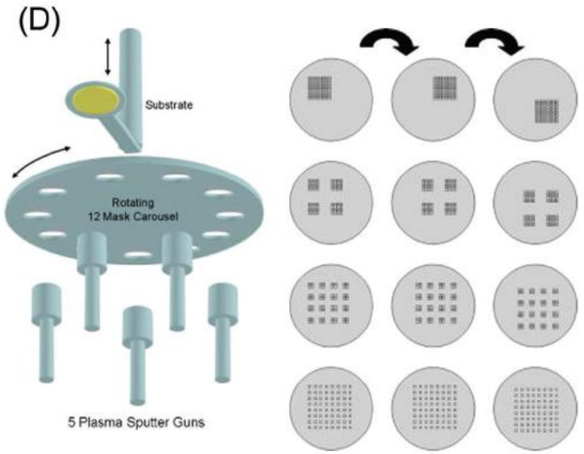


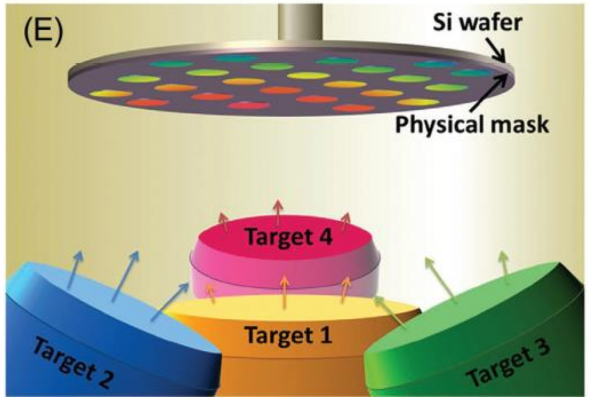
Data-mining algorithms

> Experimental screening by robotic/automatized platforms







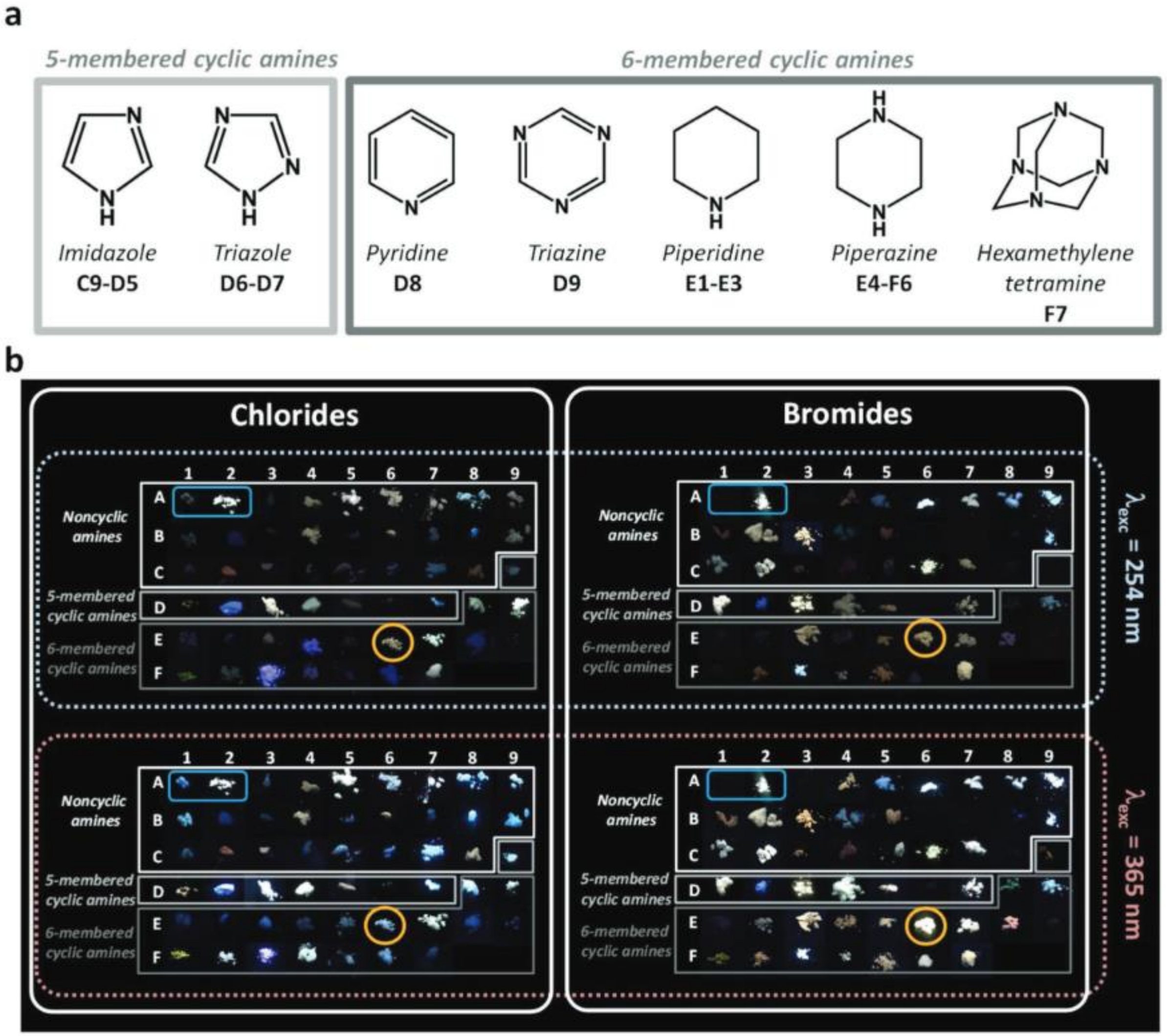


Adv. Ener. Mater. (2021)

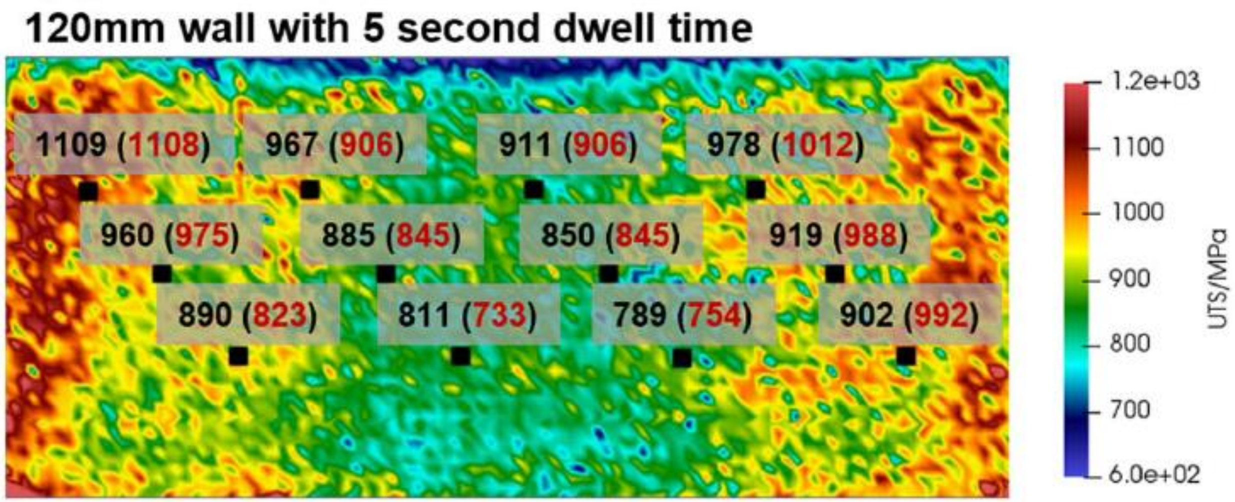
High-Throughput Experimentation and Computational Freeway Lanes for Accelerated Battery Electrolyte and Interface Development Research

Data-mining algorithms

> Experimental screening by Fast characterization / Additive manufacturing

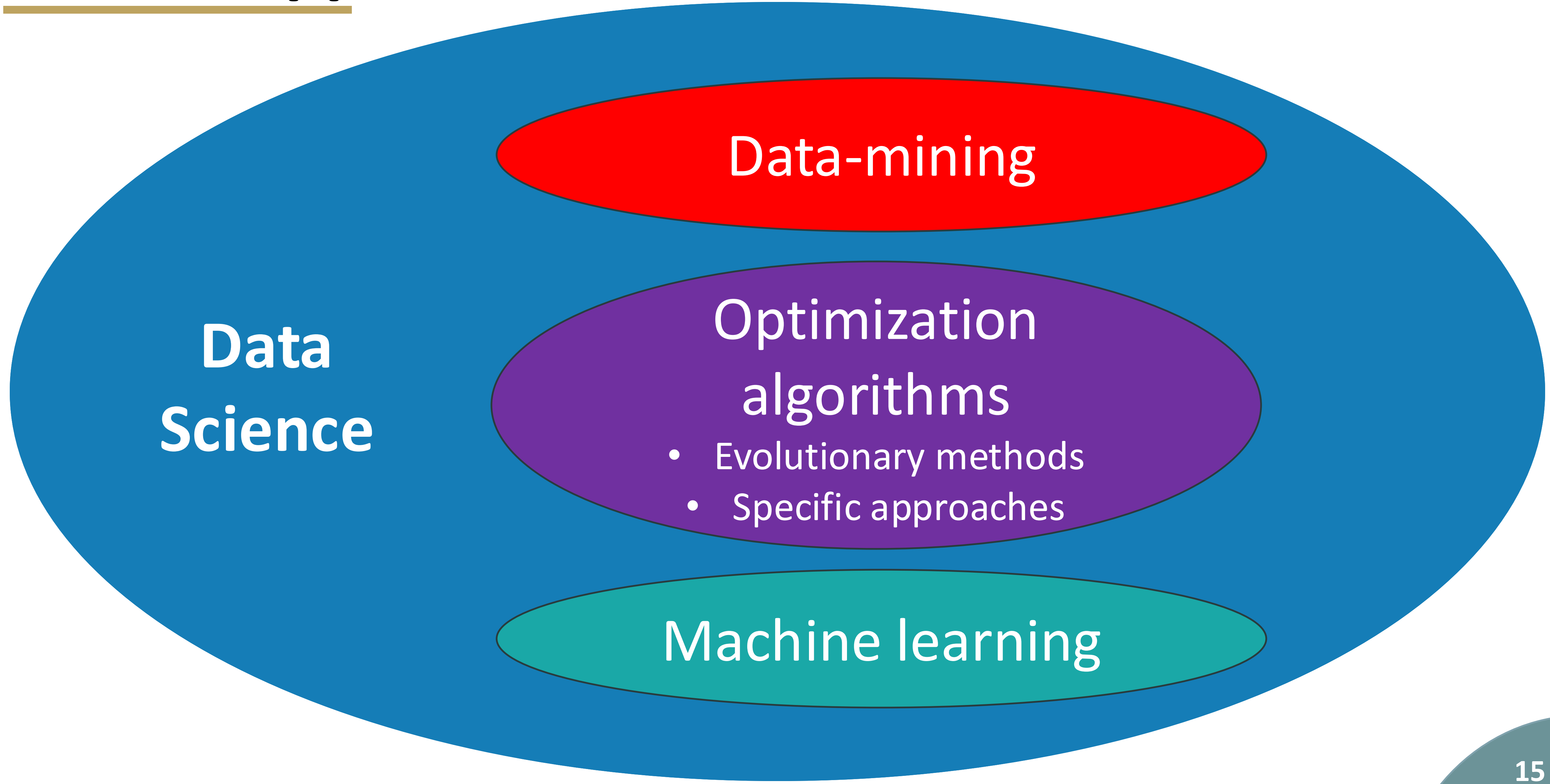


Brochard-Garnier et al. Adv. Funct. Mater. (2019)
Screening Approach for the Discovery of New Hybrid Perovskites with Efficient Photoemission



Xie *et al*. Npj Comp Mater (2021)
Mechanistic data-driven prediction of as-built mechanical properties in metal additive manufacturing

Several approaches



Optimization algorithms

> Evolutionary methods, genetic algorithm, …

Initial population → Selection → Reproduction → Mutation → Evaluate new candidates

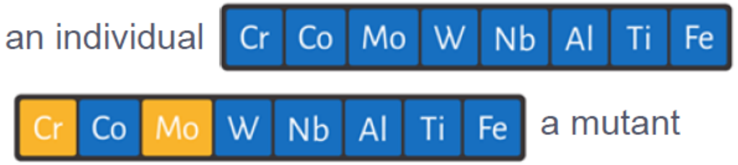
1 **alloying element** = 1 **gene**

1 **alloy** = 1 **individual** = 1 group of genes

1 **population** = 1 group of **alloys**

Reproduction

Mutation

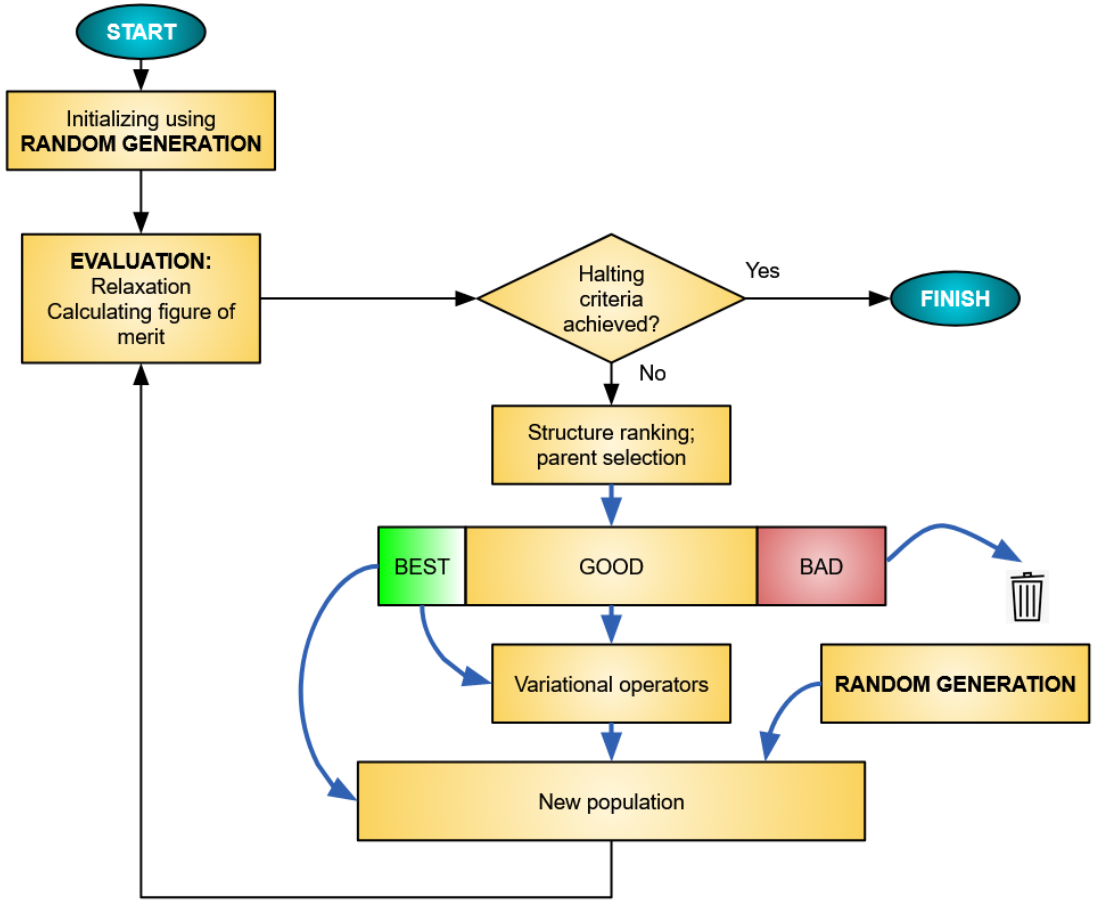


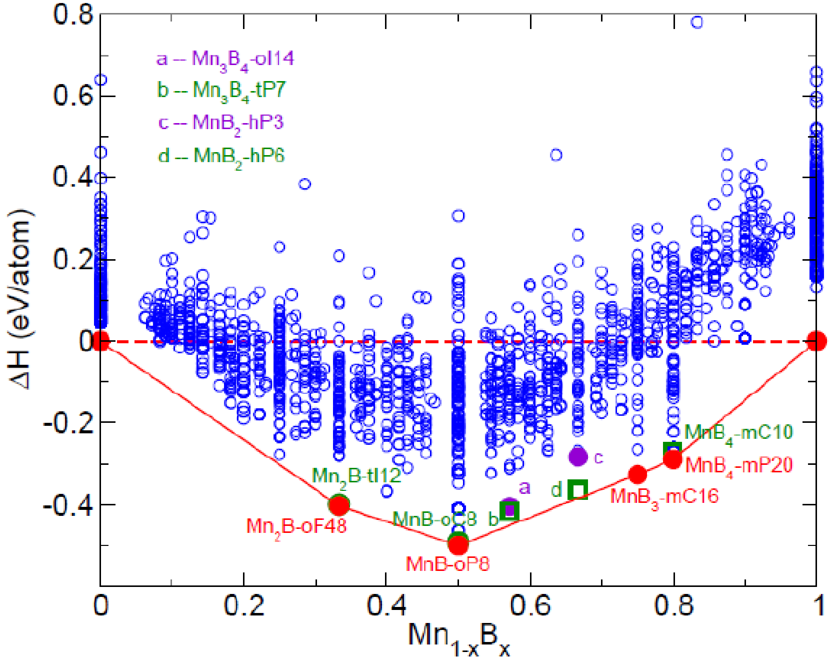
Menou, Tancret *et al.* Scripta Mat. (2018)
Computational design of light and strong high entropy alloys (HEA)

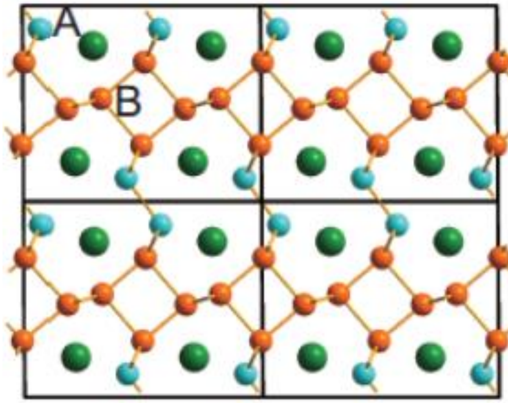
SAFRAN

Optimization algorithms

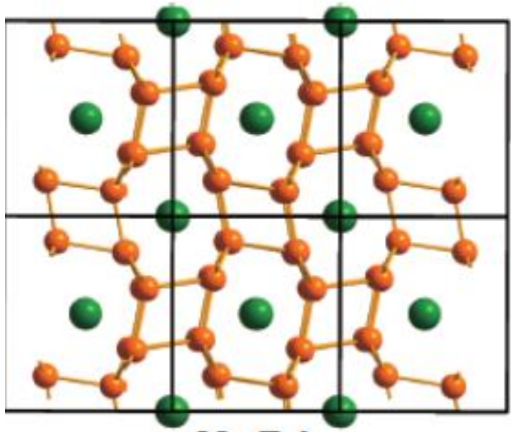
> Evolutionary methods for crystal design: USPEX







MnB3



MnB4

Courtesy of V. Baturin.

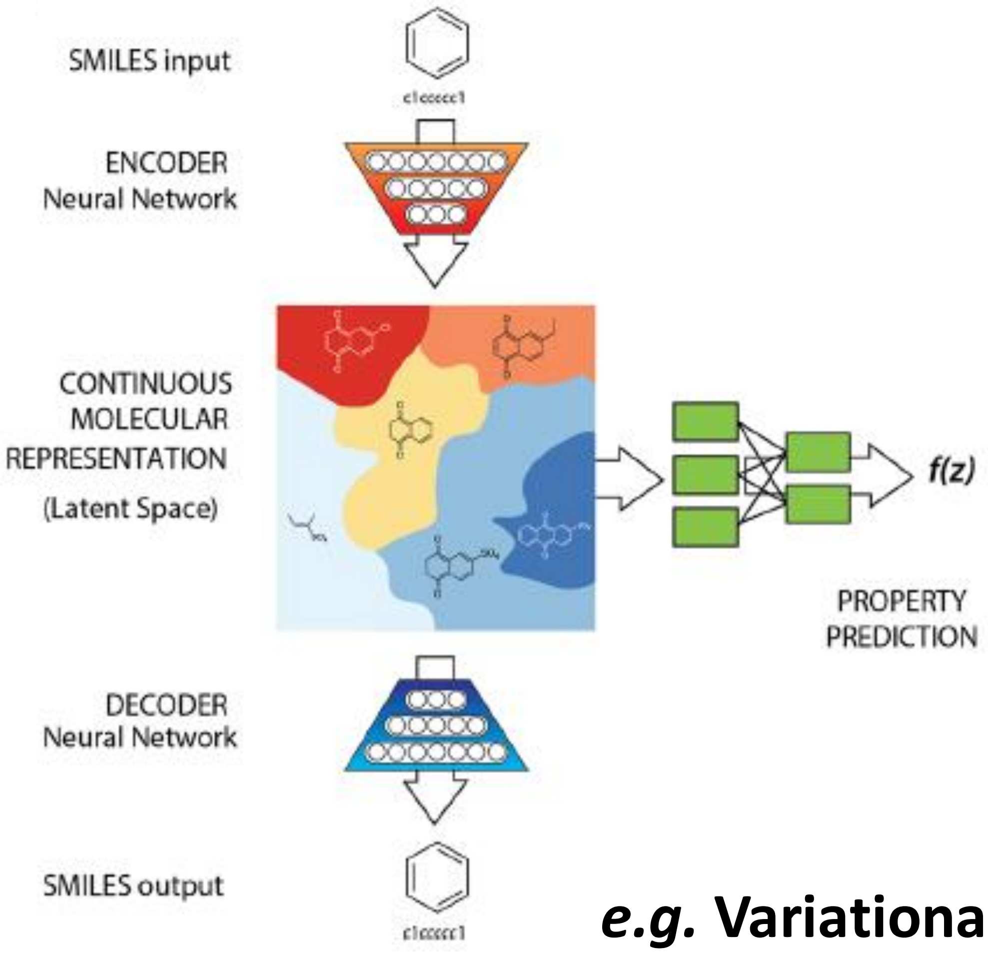
Oganov *et al*. Nature (2009)

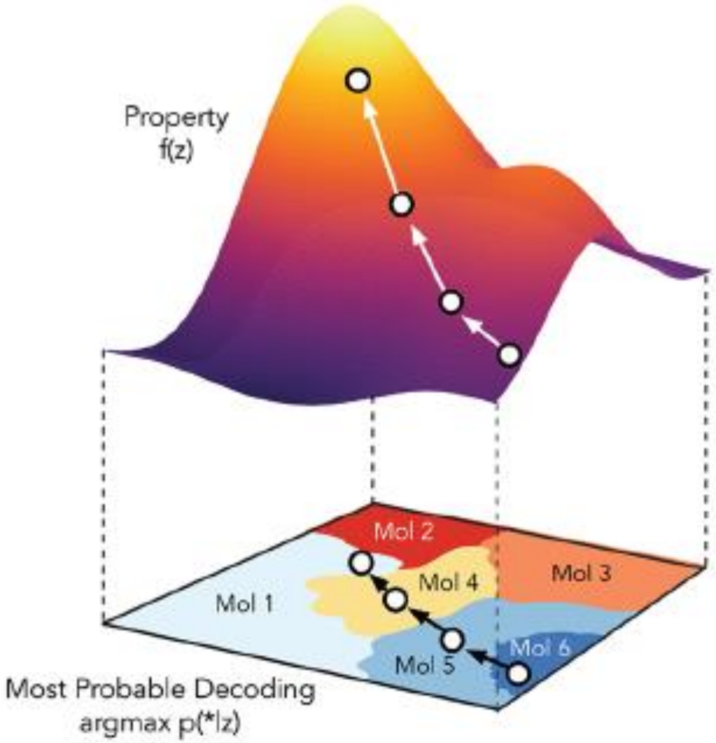
New super hard structure of Boron

Optimization algorithms

> Generative methods

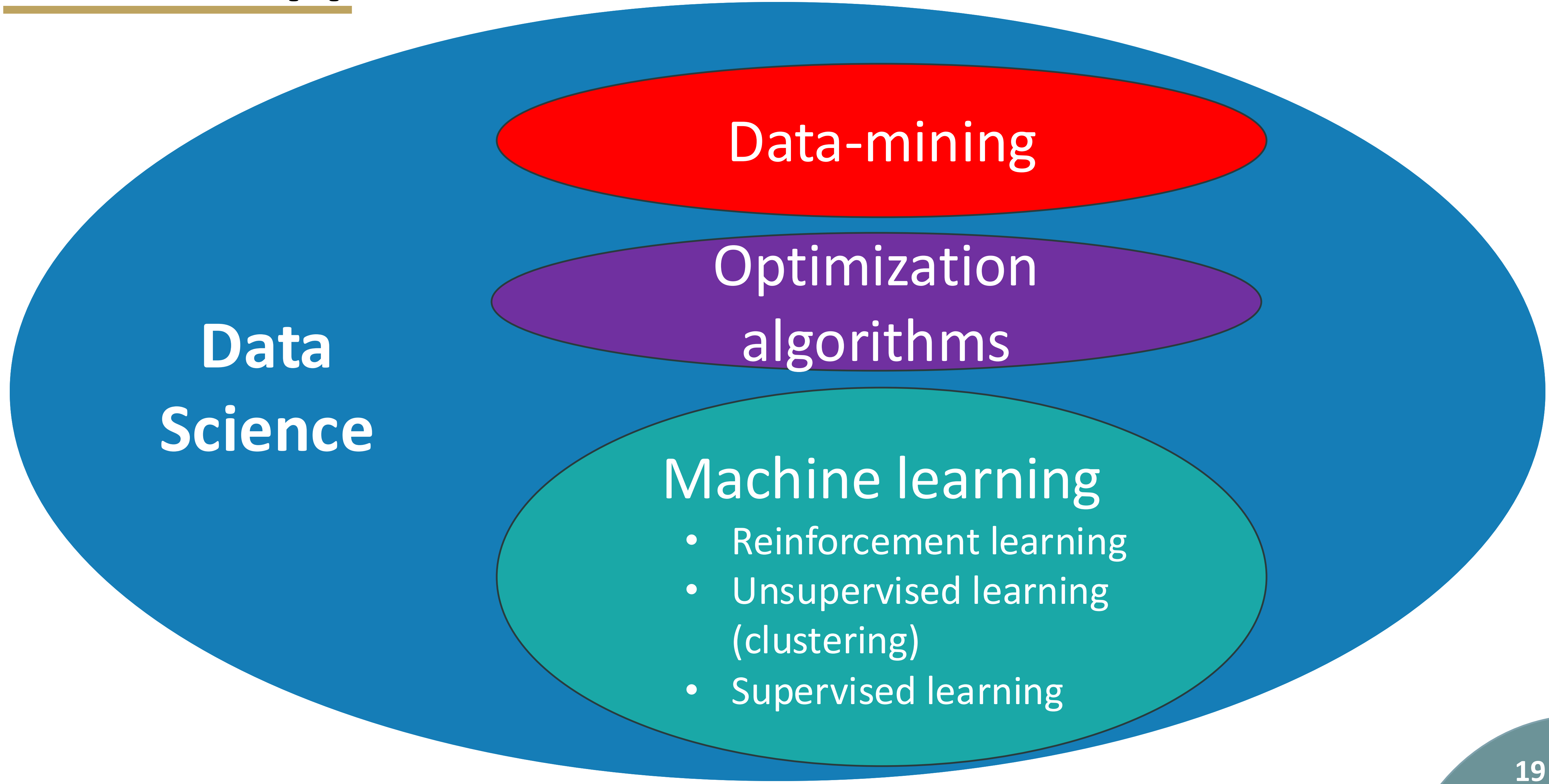
Gomez-Bombarelli *et al*.
ACS Central Science (2018)
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules





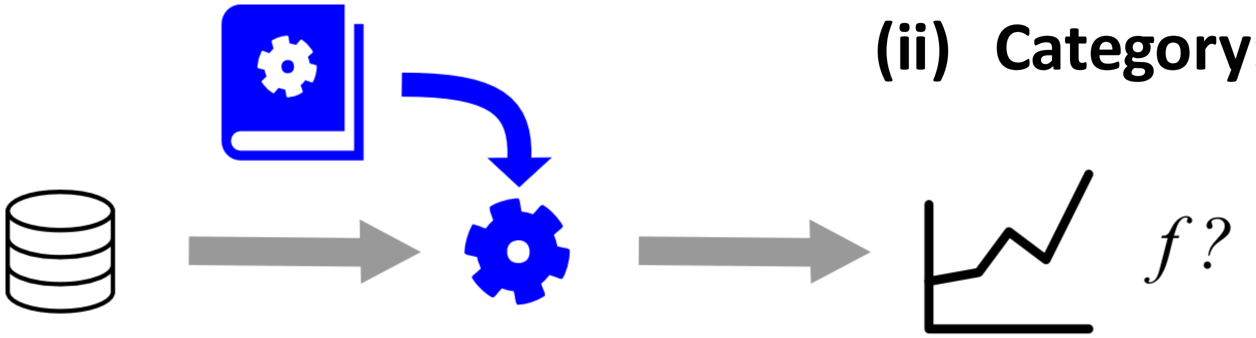
e.g. **Variational Auto Encoder (VAE)**

Several approaches



Machine learning algorithms (ML)

> Supervised learning



The target value *f*(*x*) is a:

(i) Numerical value: Regression

(ii) Category: Classification

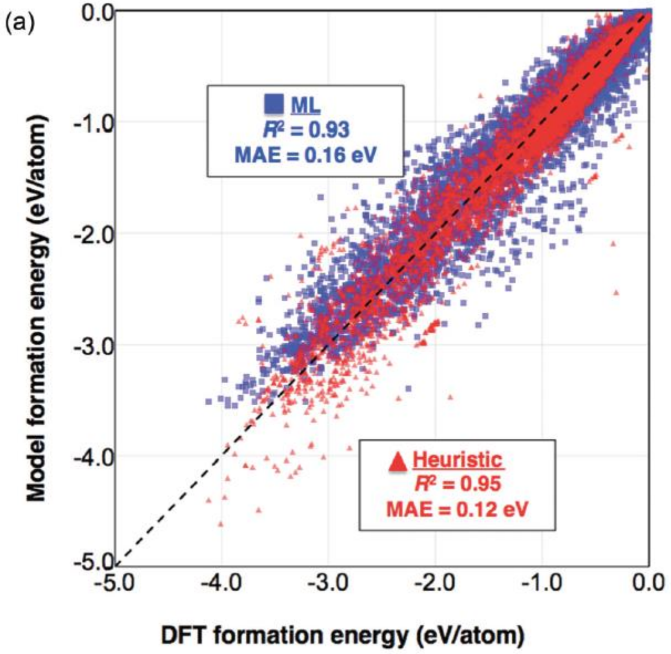
ML > Supervised learning

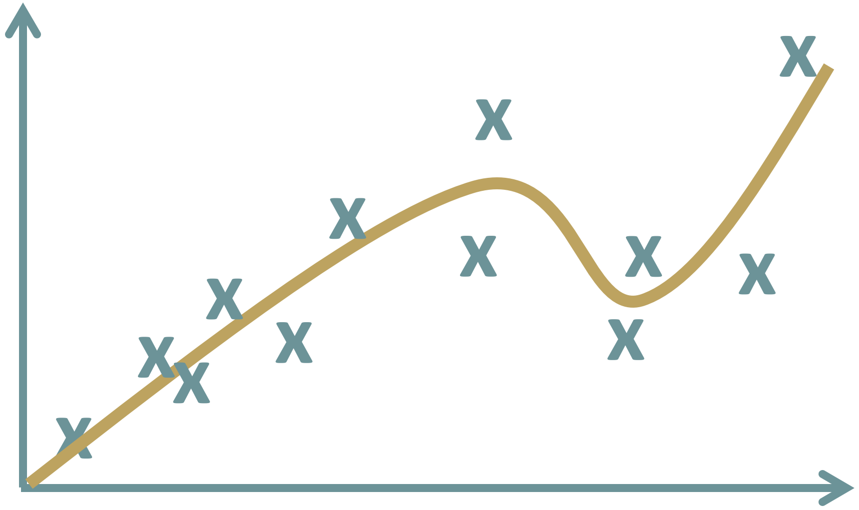
(i) *f*(*x*) = numerical value

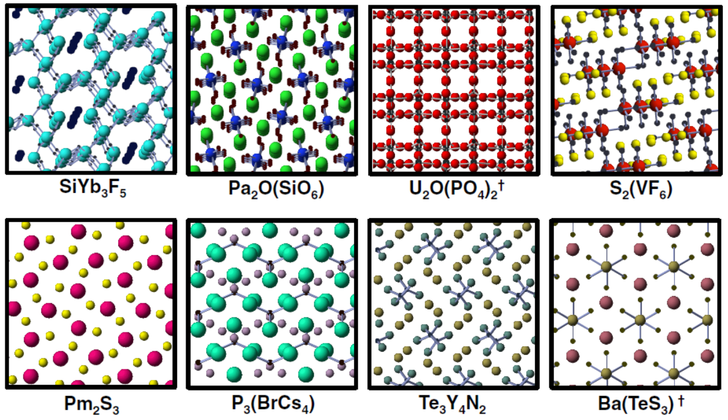
→ Regression algorithms

- Linear, non-linear

Meredig, Wolverton *et al*. Phys. Rev. B (2014)
Combinatorial screening for new materials in unconstrained composition space with machine learning





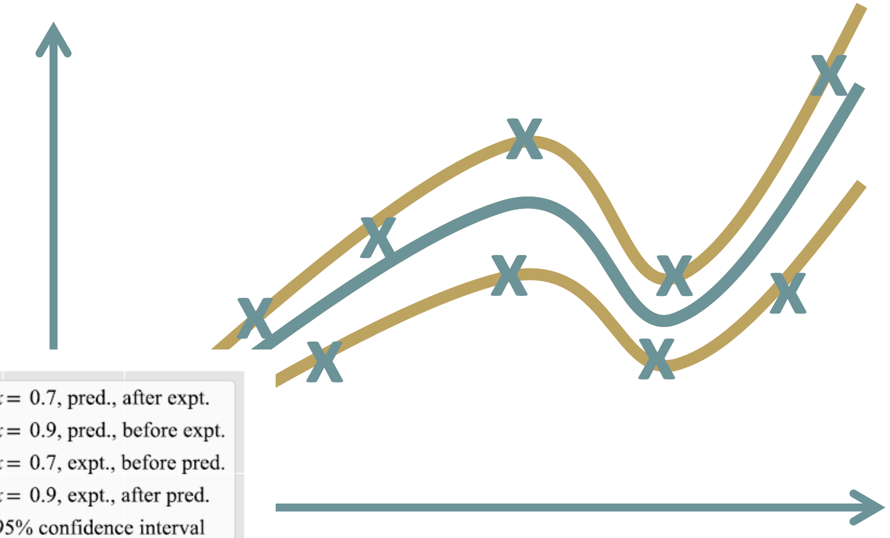


ML > Supervised learning

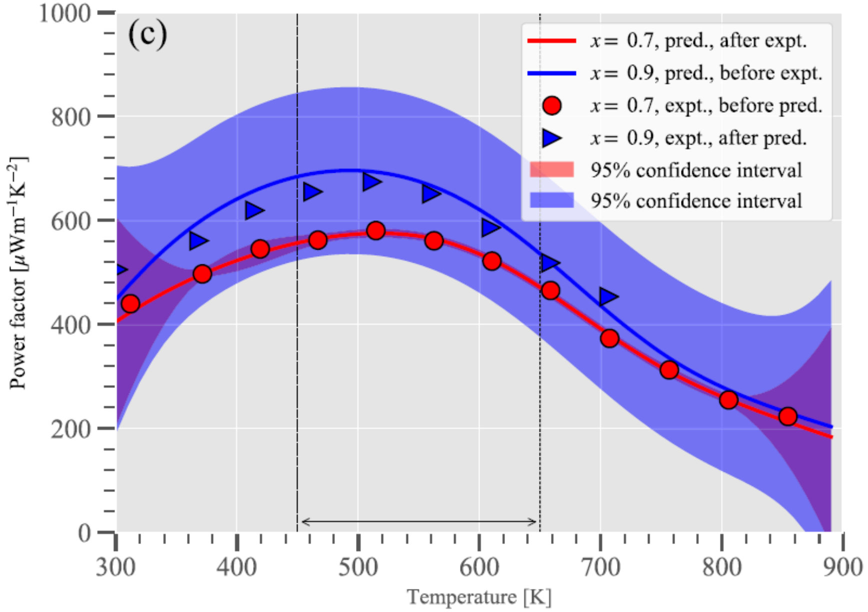
(i) *f*(*x*) = numerical value

→ Regression algorithms

- Linear, non-linear
- **Gaussian process regression** (kriging)



Hou, Shinohara *et al*. Applied Materials & Interfaces (2019)
Machine-Learning-Assisted Development and Theoretical Consideration for the $Al_2Fe_3Si_3$ Thermoelectric Material



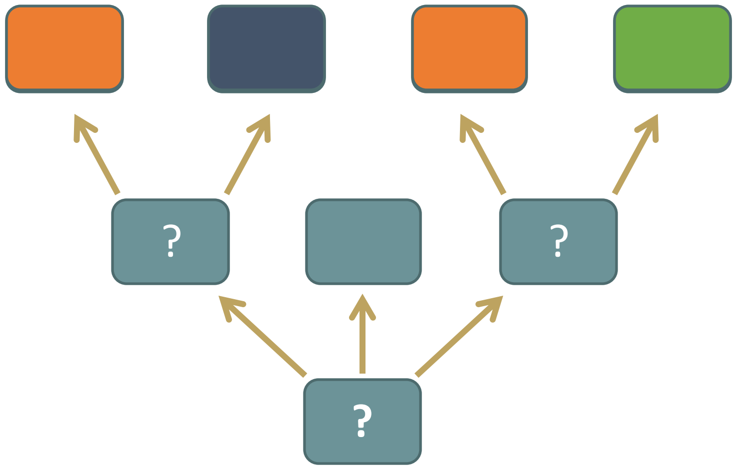
$Al_{23.5+x}Fe_{36.5}Si_{40-x}$

ML > Supervised learning

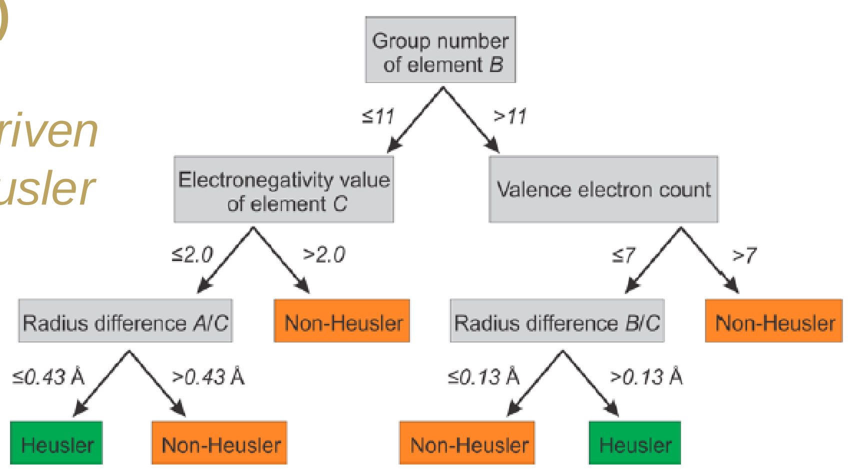
(ii) $f(x)$ = Category

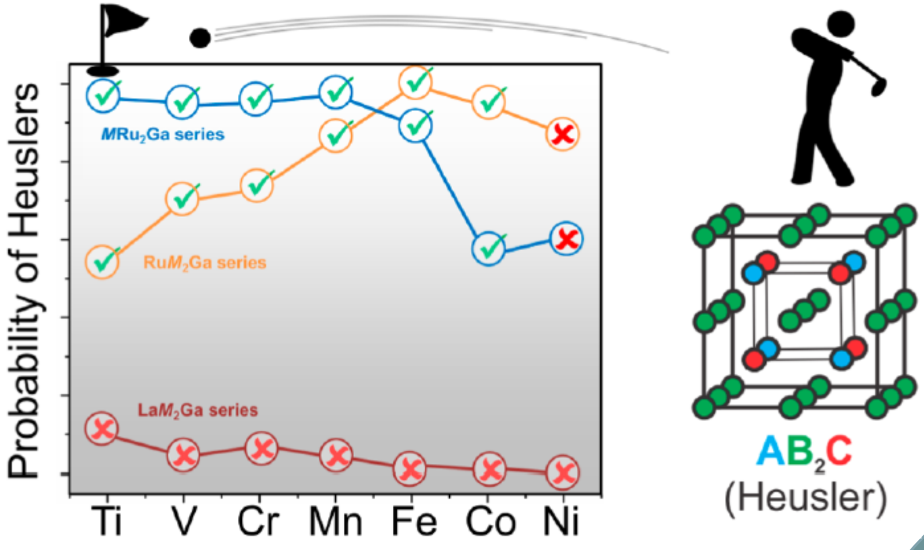
→ Classification algorithms

- Decision tree



Oliyny, Sparks *et al*.
Chem. Mater. (2016)
High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds



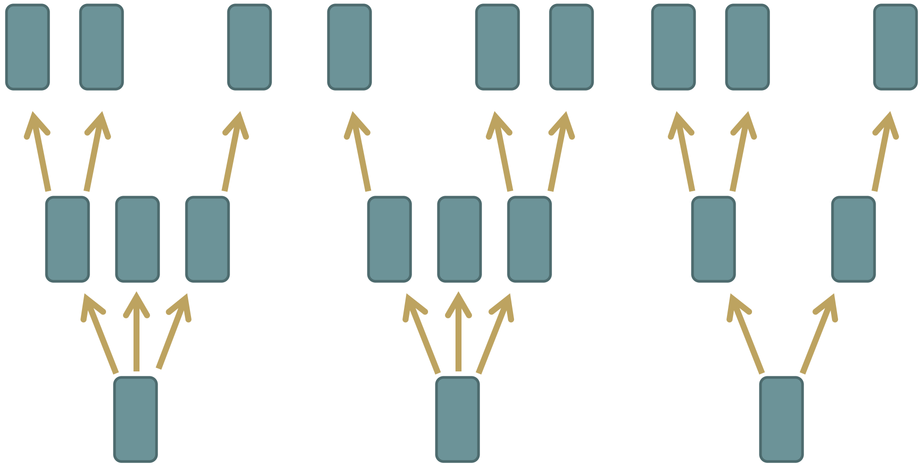


ML > Supervised learning

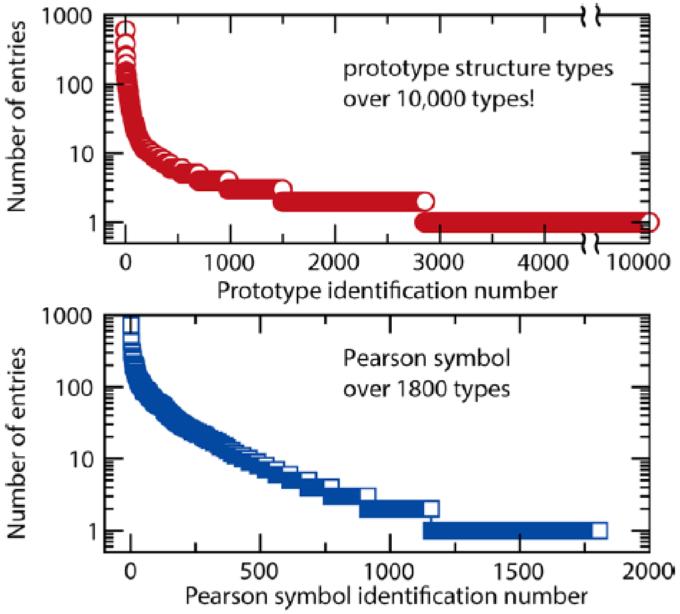
(ii) $f(x)$ = Category

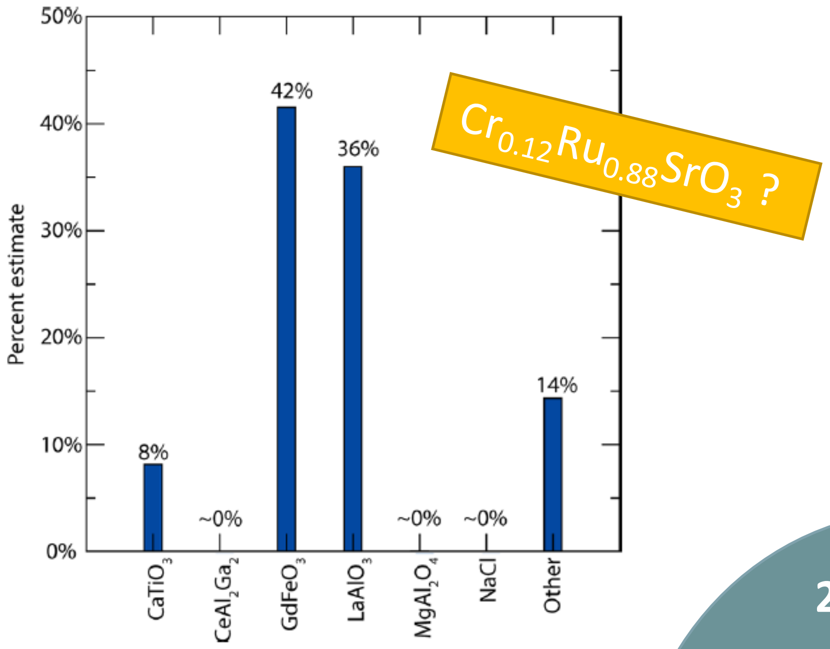
→ Classification algorithms

- Decision tree, **random forests**



Graser, Sparks *et al.* Chem. Mater. (2018)
Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons



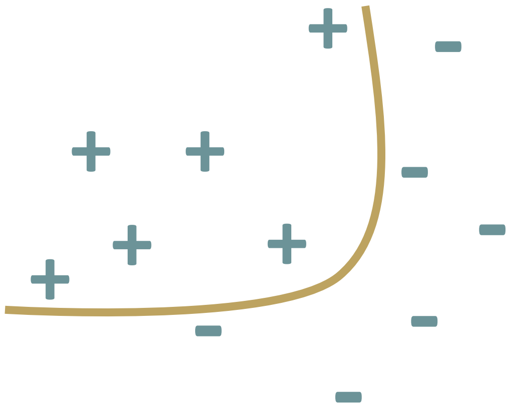


ML > Supervised learning

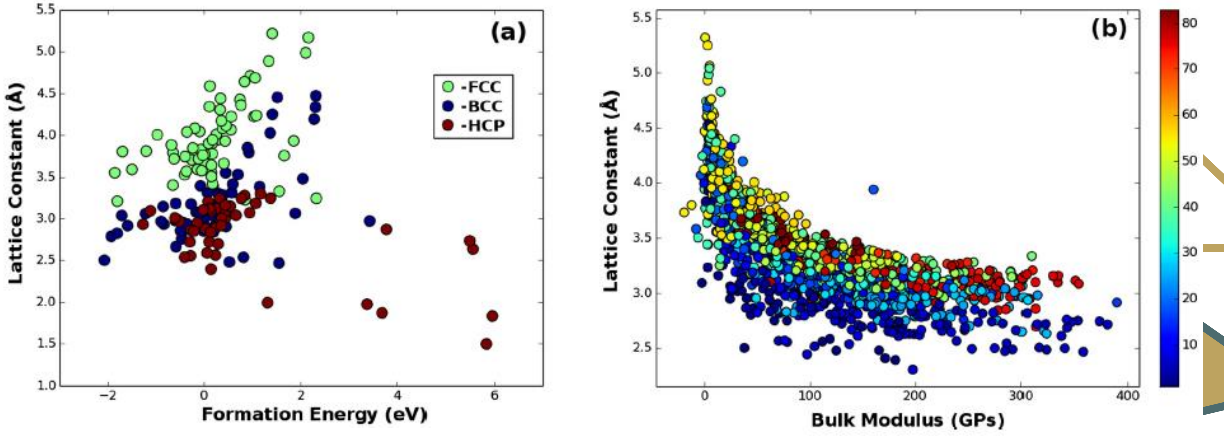
(ii) *f*(*x*) = Category

→ Classification algorithms

- Decision tree, random forests
- **Support vector machine (SVM)**



Takahashi, Tanaka.
Comput. Mater. Sci. (2016)
Material synthesis and design from first principle calculations and machine learning



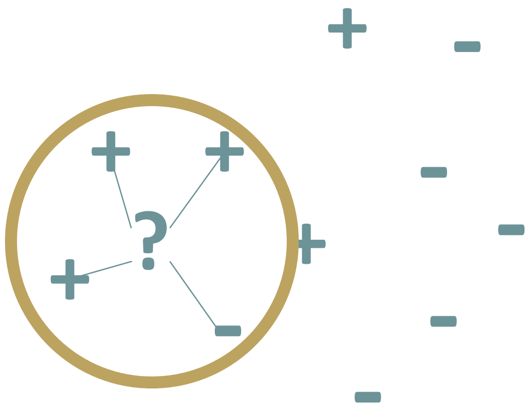
	Structure	Lattice	Formation
FeAl	BCC (BCC [13])	2.5–3.0 (2.91 [13])	Exothermic (Exothermic [14])
FeNi	FCC (FCC [16])	3–3.5 (3.57 [16])	Endothermic (Endothermic [17])
FeTi	BCC (BCC [20])	3–3.5 (2.98 [20])	Exothermic (Exothermic [21])

ML > Supervised learning

(ii) *f*(*x*) = Category

→ Classification algorithms

- Decision tree, random forests
- Support vector machine (SVM)
- **K-nearest neighbor** (KNN), Naive Bayes classifier



Naik *et al*.
Metals (2019)
Texture-Based Metallurgical Phase Identification in Structural Steels

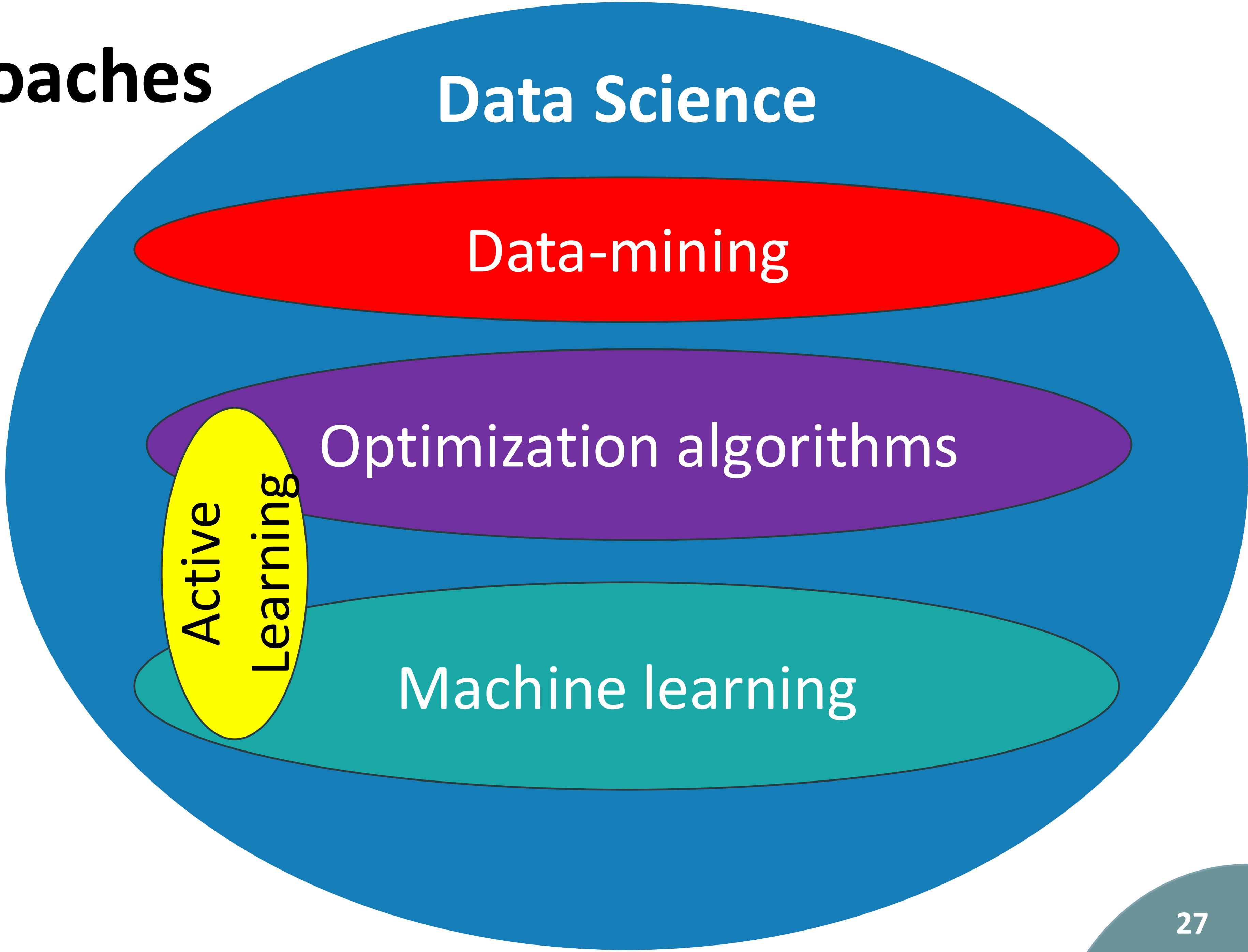
F-Ferrite
P-Pearlite
M-Martensite

Test Image (500X500) (From Original Image)

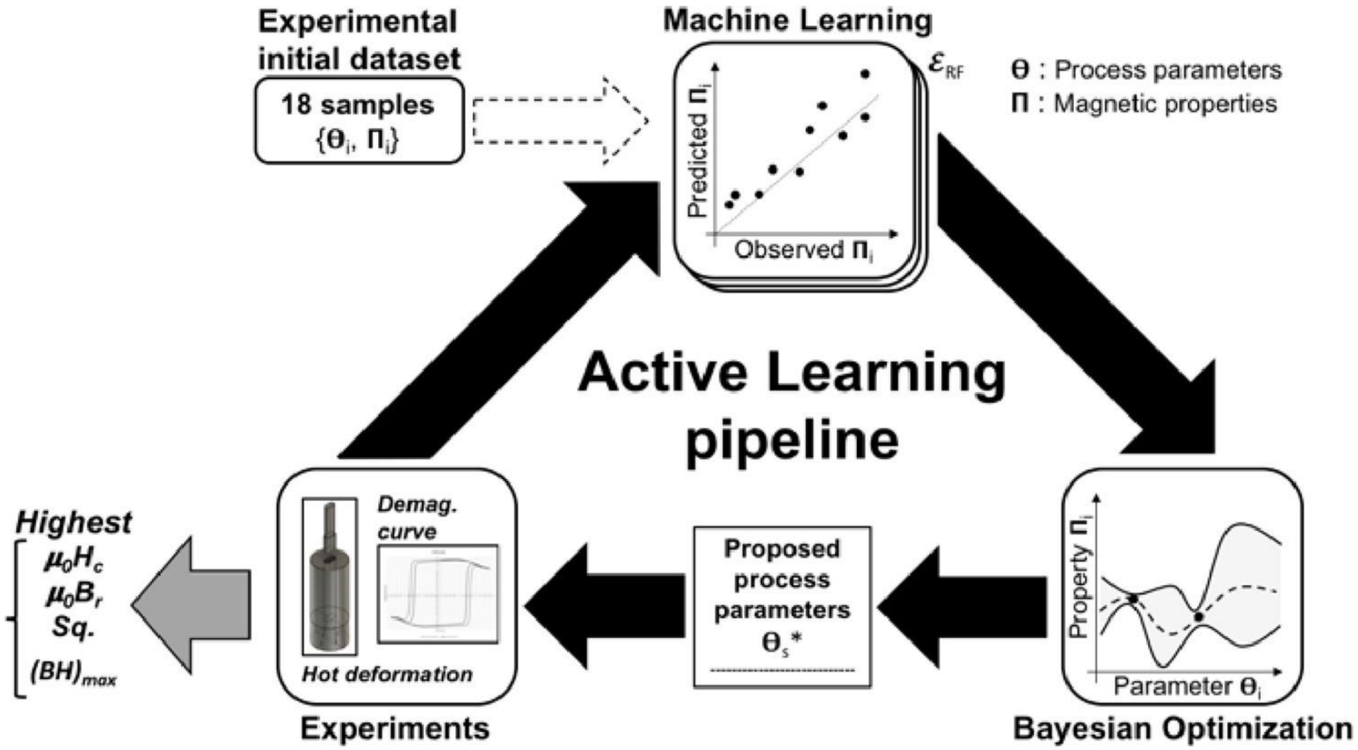
Naïve Bayes F/P=77/23

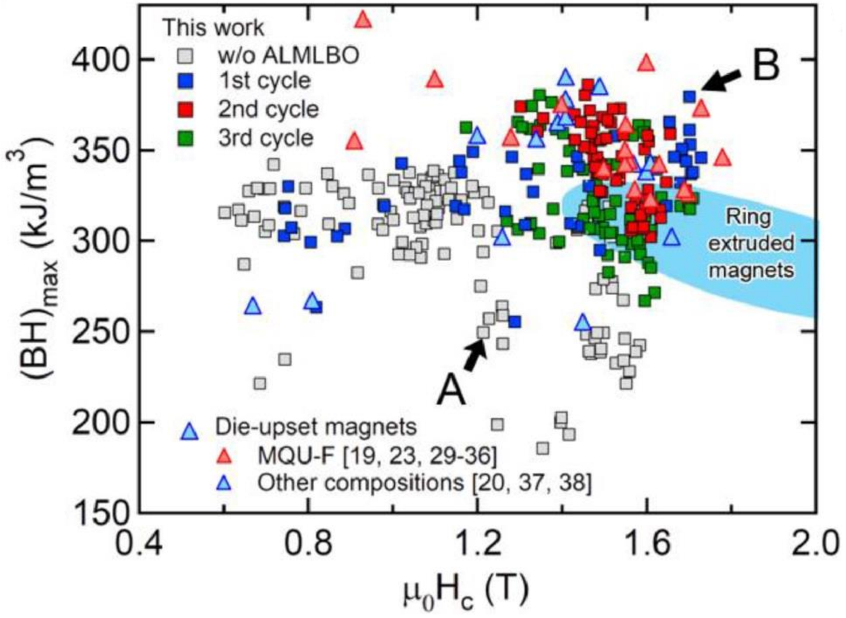
K-Nearest Neighbor F/P-76.9/20.6

Others approaches



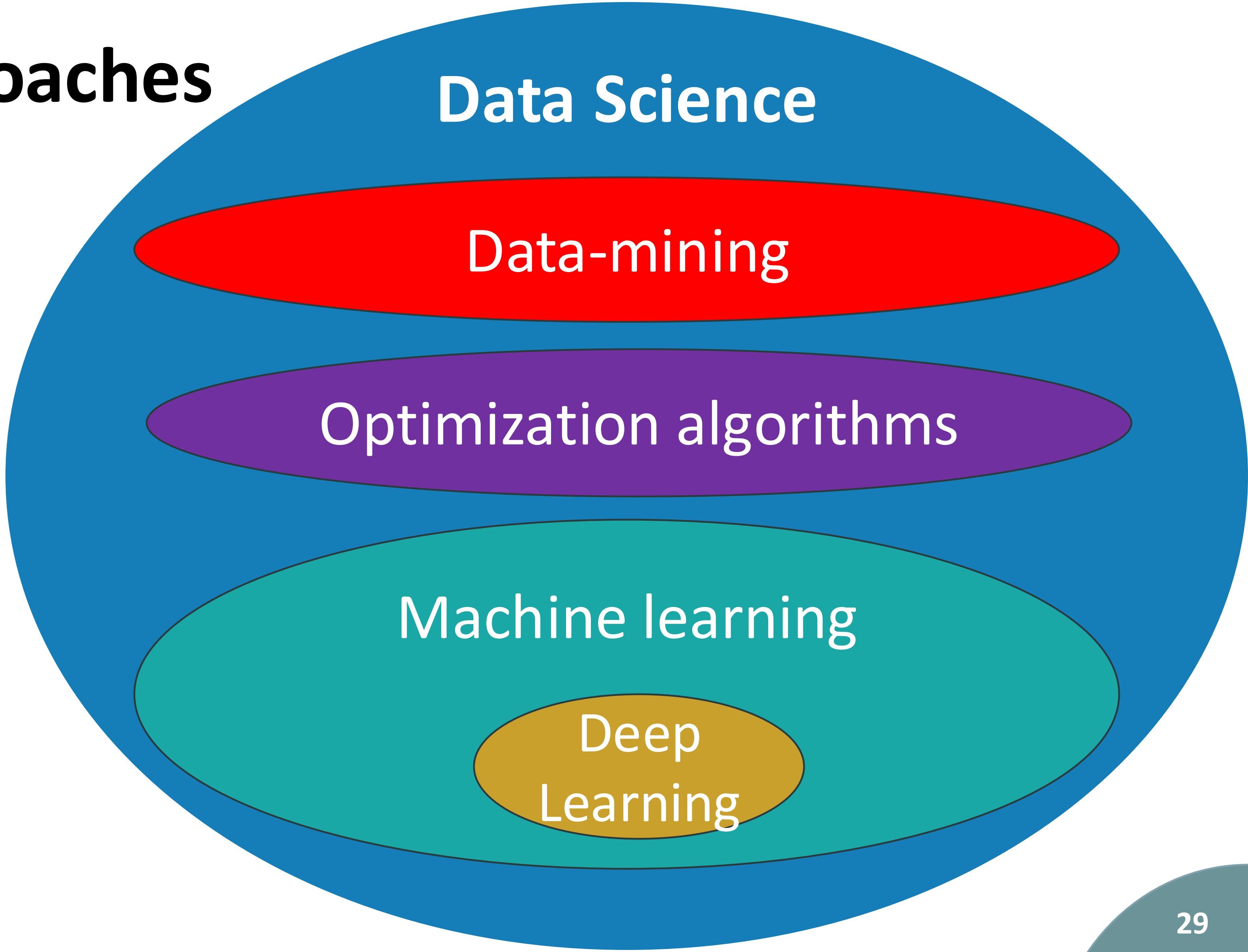
Active learning





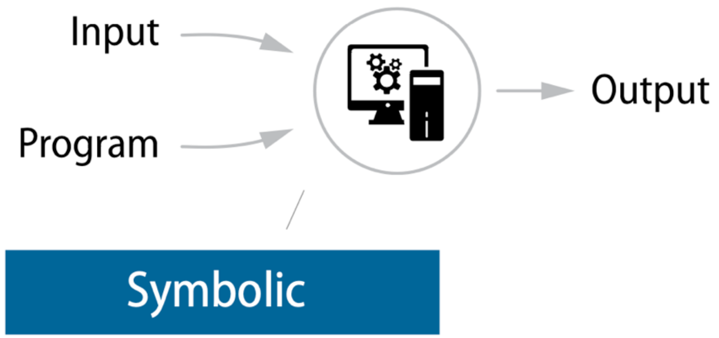
G. Lambard *et al.* Scripta Materialia (2022)
Optimization of direct extrusion process for Nd-Fe-B magnets using active learning assisted by machine learning and Bayesian optimization

Others approaches



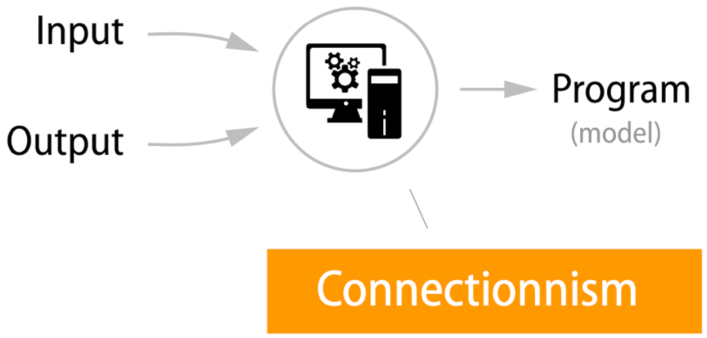
The Big Controversy

Deductive approach



V.S.

Inductive approach

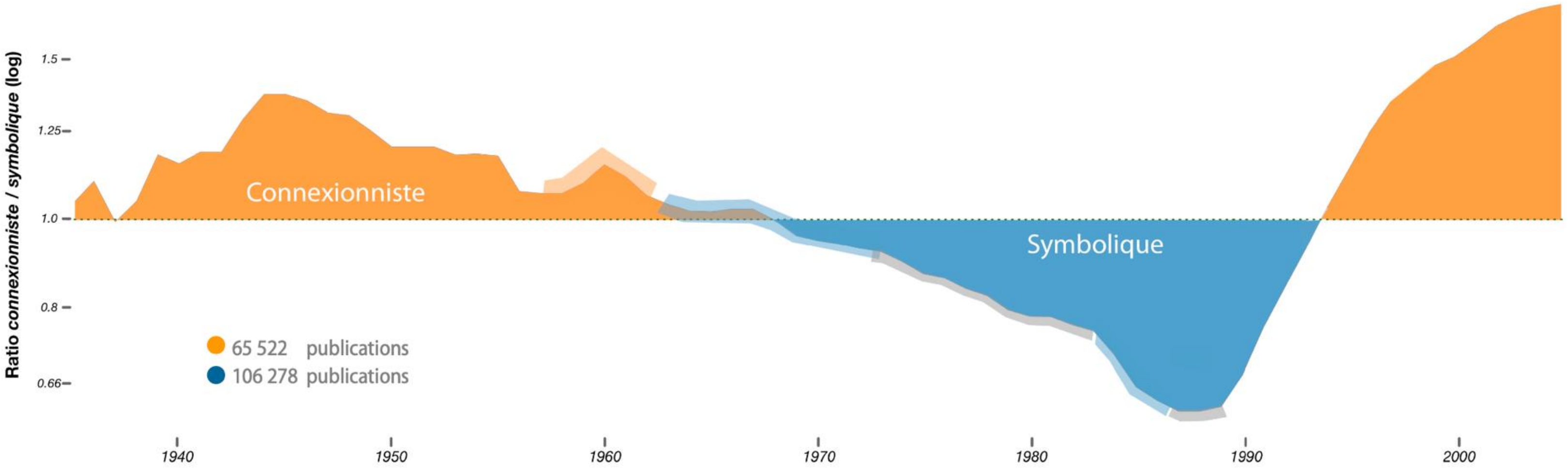


Expert

Rules and laws ▶ Special case

Facts ▶ Rules and laws

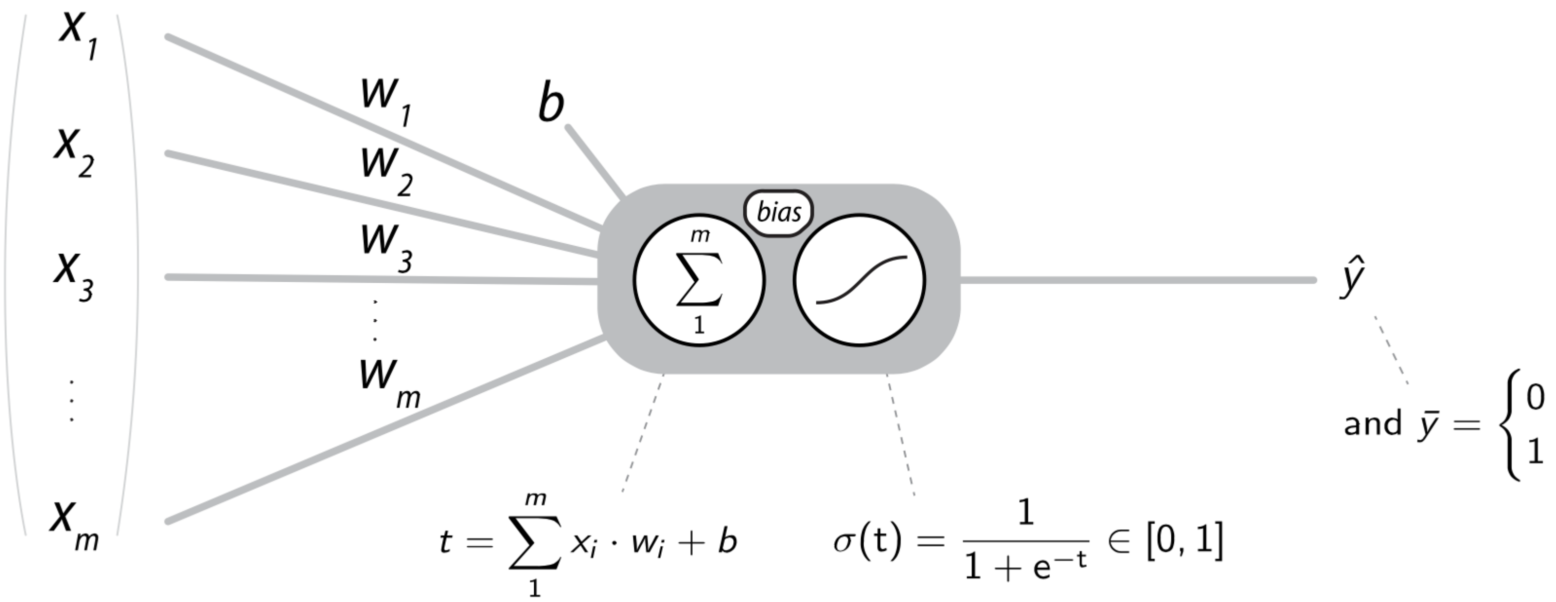
Model



D. Cardon *et al.* Réseaux (2018)
La revanche des neuronnes

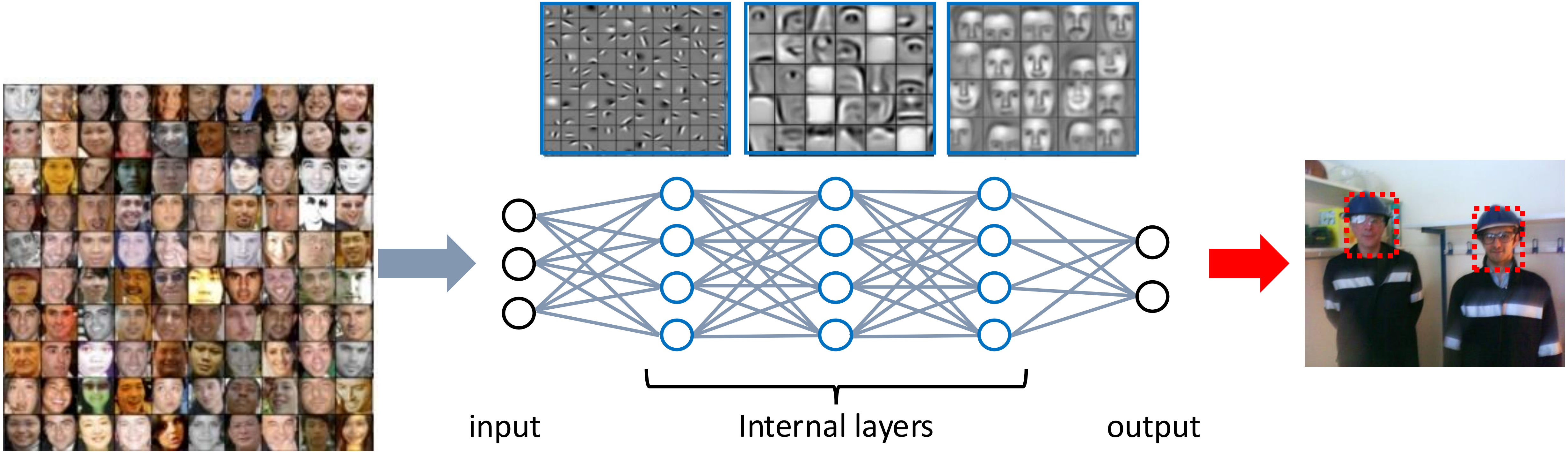
The artificial neuron

$$\hat{y} = \sigma(\Theta^T \cdot X + b)$$



Input	Bias / Weight	Activation function	Output
X	Θ	$\sigma(t)$	$\hat{y}$

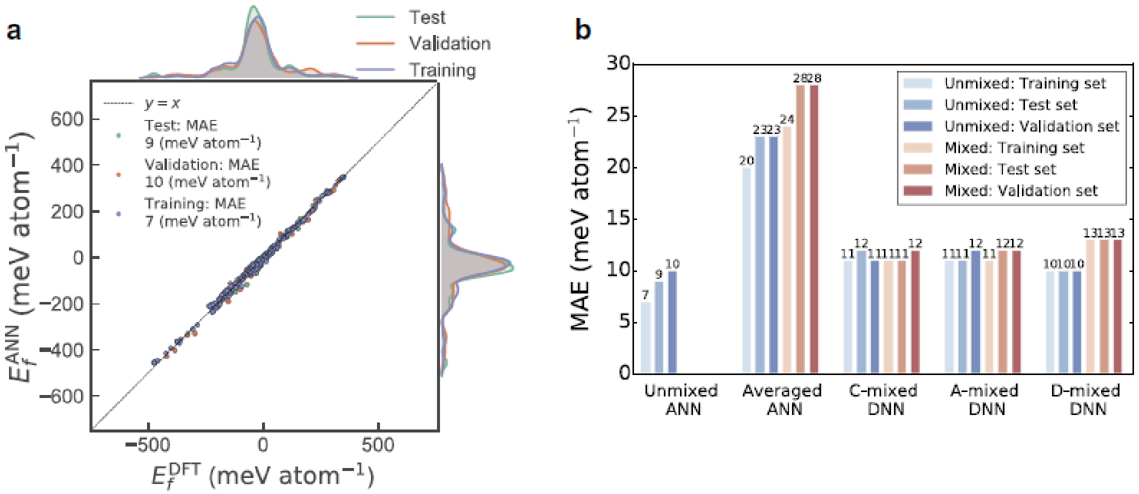
The artificial neural network (ANN)

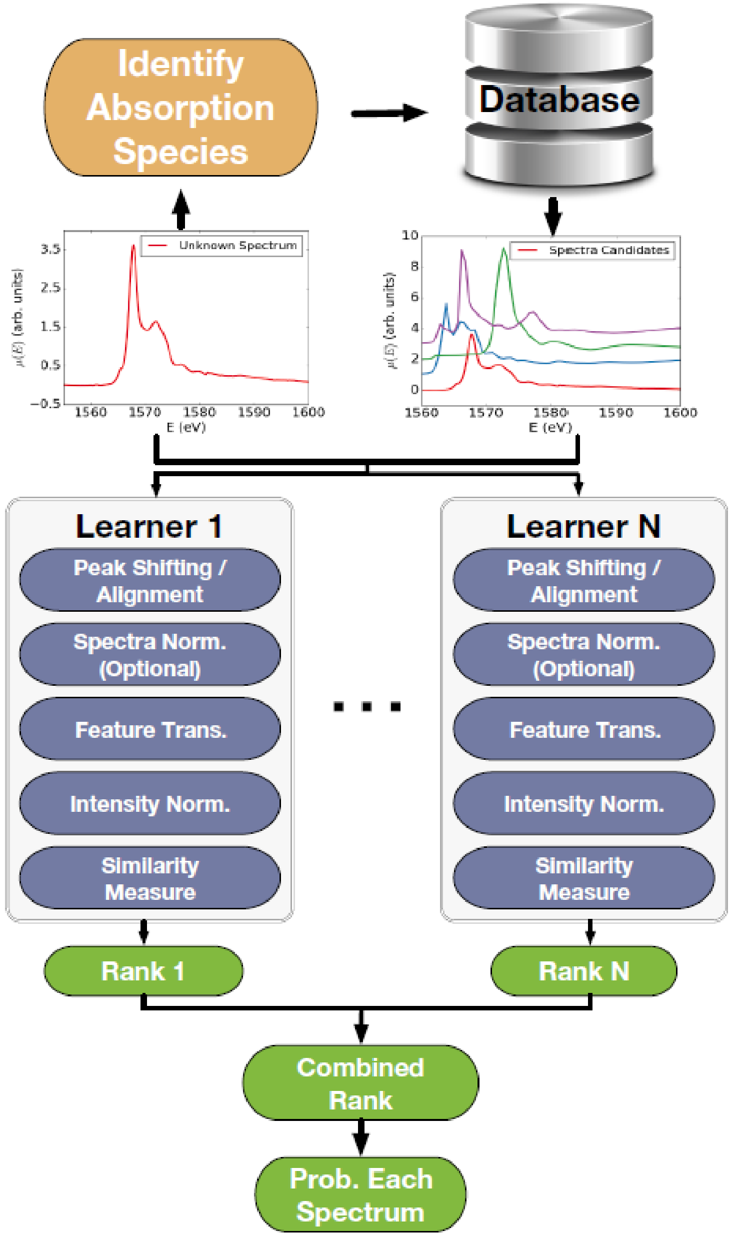


Somes examples…

Ong, Comput. Mater. Sci. (2019)

Accelerating materials science with high-throughput computations and machine learning

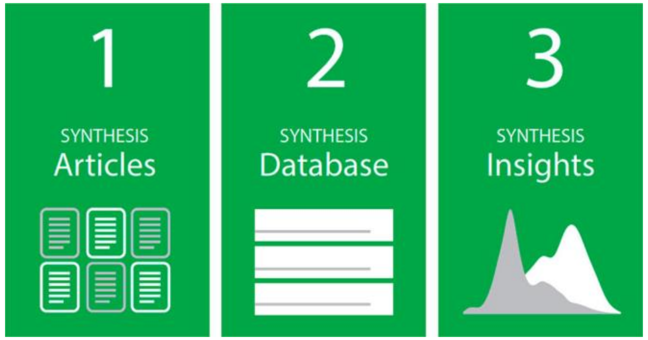


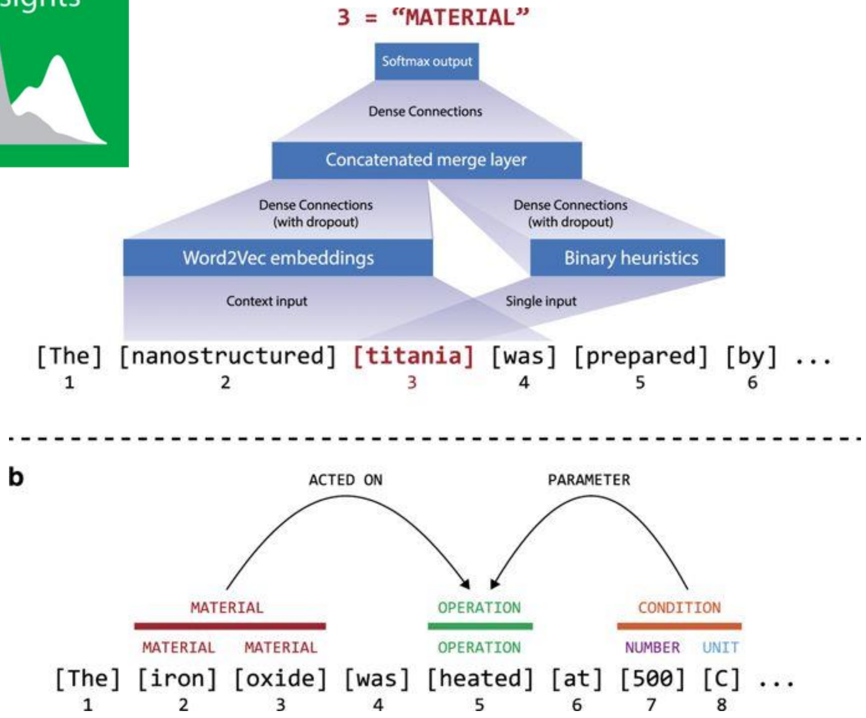


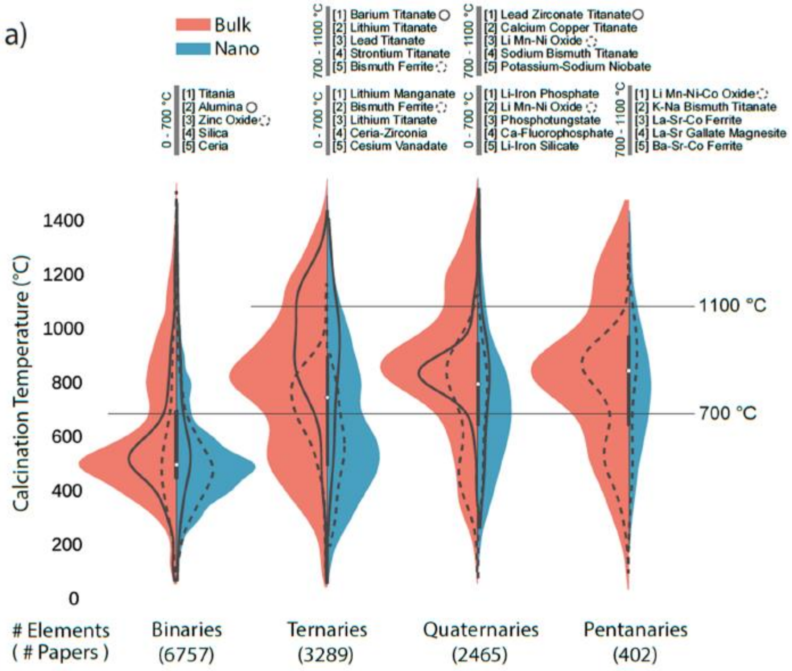
Text mining

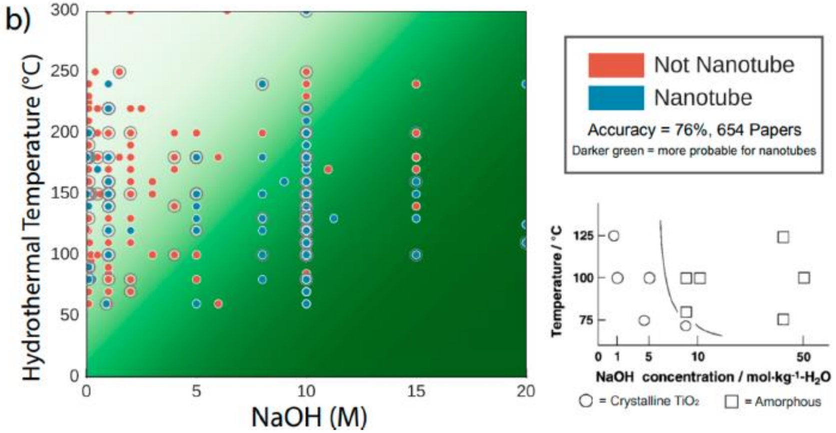
Kim, Saunders, Ceder, Olivetti *et al*. Chemistry of materials (2017)

Materials synthesis insights from scientific literature via text extraction and machine learning









Timeline

2000

Data Mining - Evolutionary algorithms - DFT High-throughput

2010

Open Database - Neural Network - Supervised ML

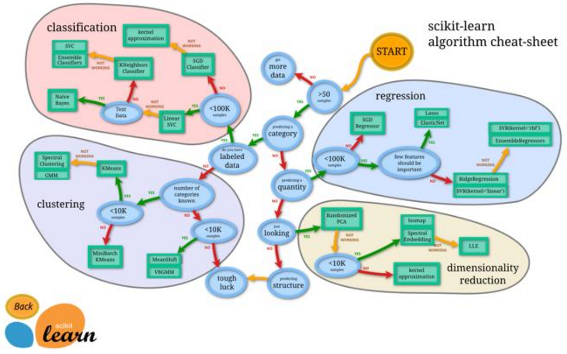
2020

Generalized descriptors - Generative algorithms

NOMAD

AFLOW
Automatic - FLOW for Materials Discovery

OPTIMADE
Open Databases Integration for Materials Design

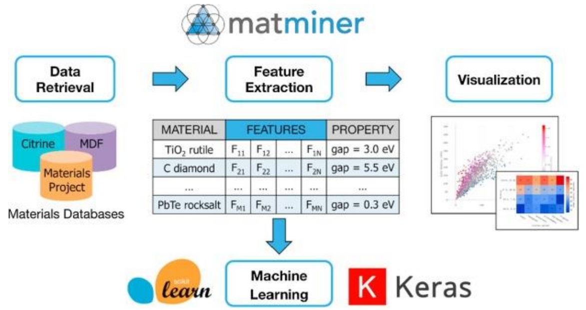


ASE

mendeleev

Mv

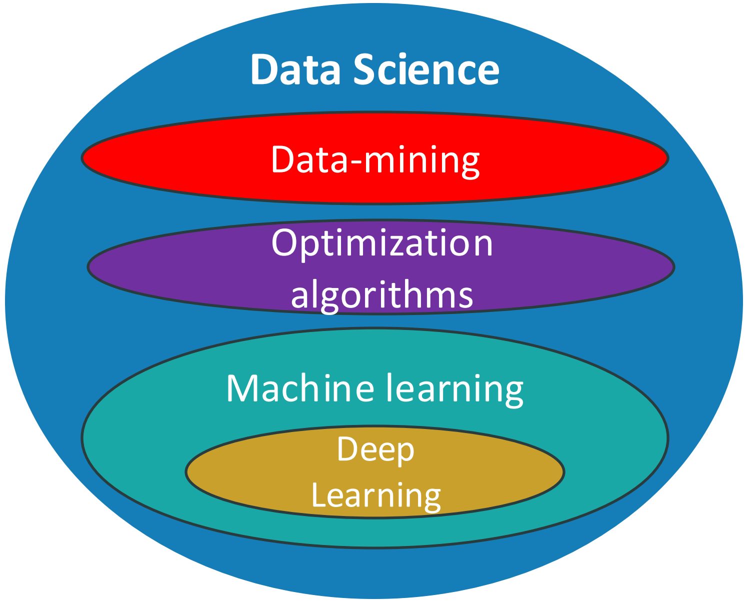
pymatgen



?

ChatGPT

Conclusions



- Artificial intelligence is not able to think instead of human : IA is just efficient for a dedicate learning

- We need (good) data!