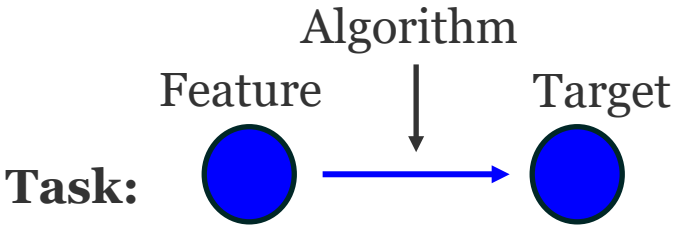
THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 10: How to Train Classifiers, Ensembles, Boosting, and Dimensionality Curse

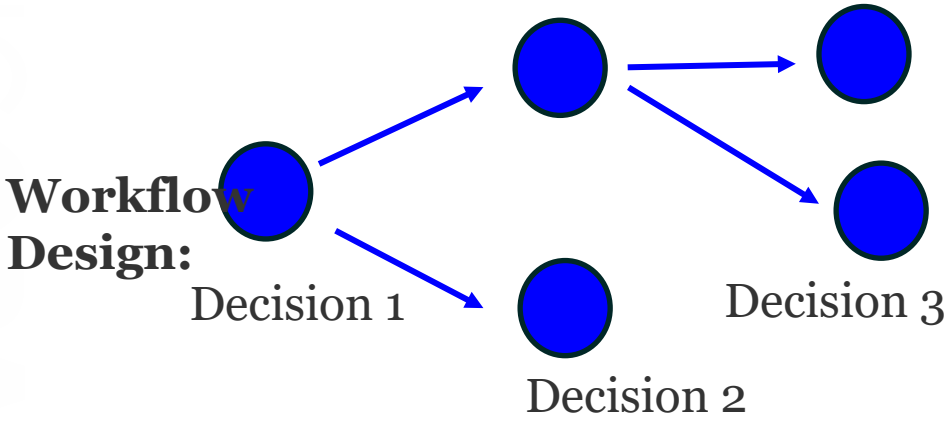
Sergei V. Kalinin

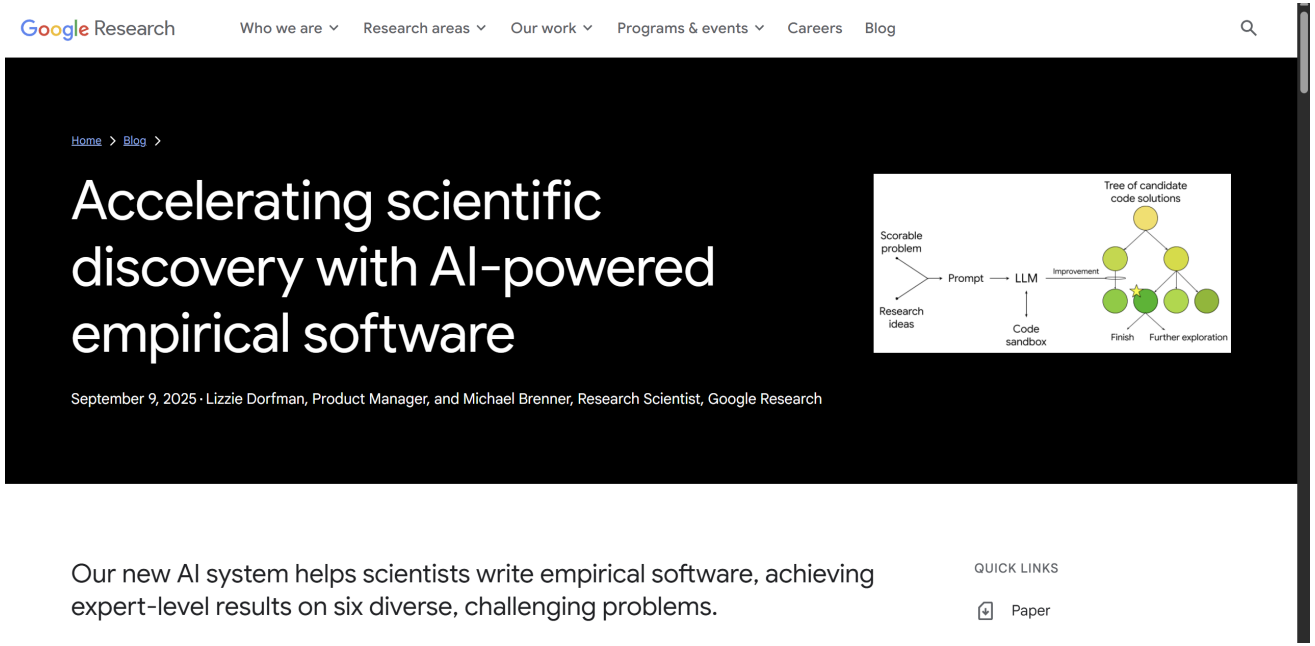
Tasks, workflows, and workflow planning



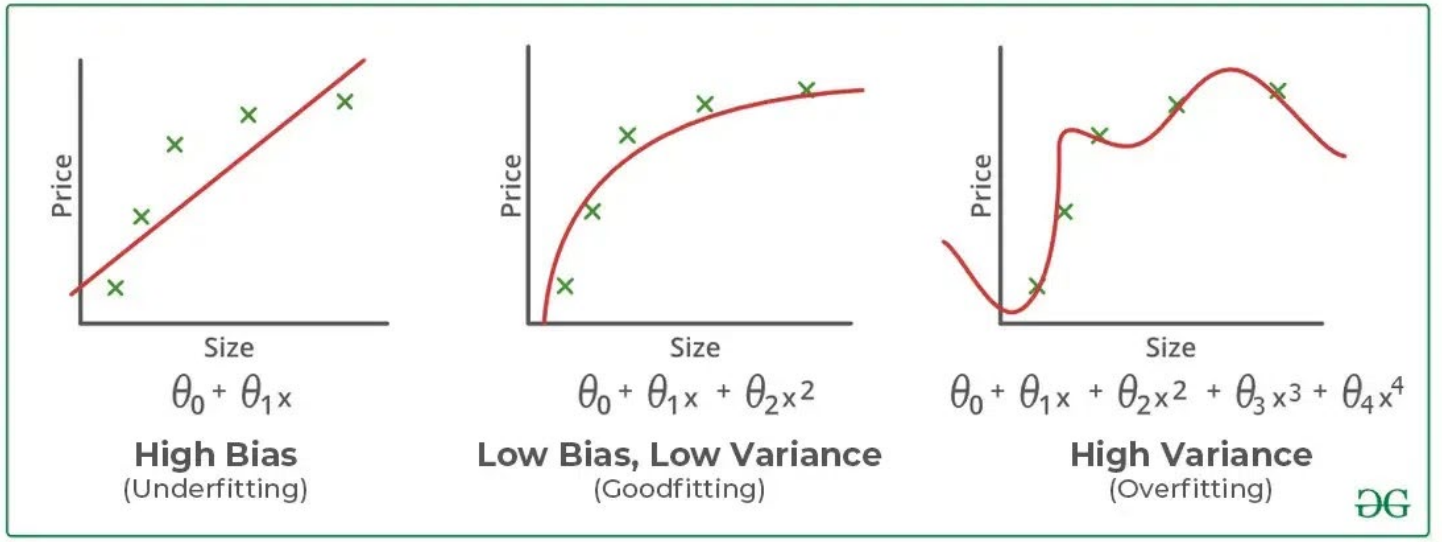
Task can be classification, regression, clustering, optimization.
Algorithm can be kNN, perceptron, DCNN, etc.

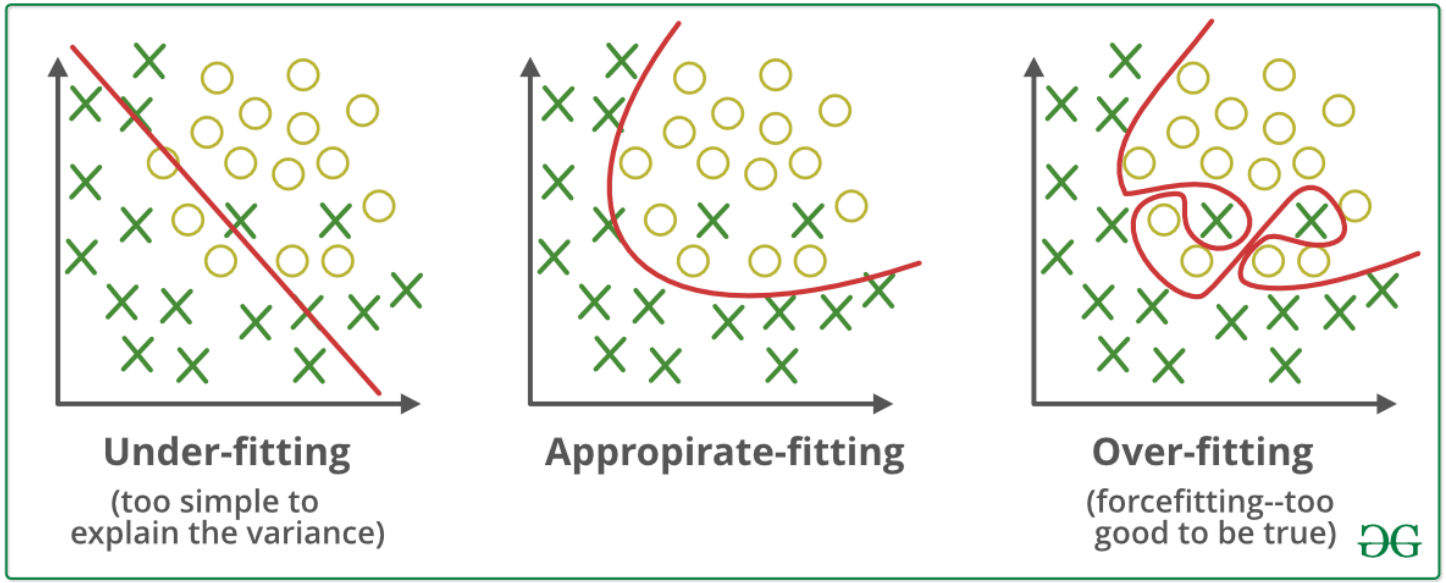
We can use complex algorithms for “simple” tasks and simple algorithms for complex tasks.





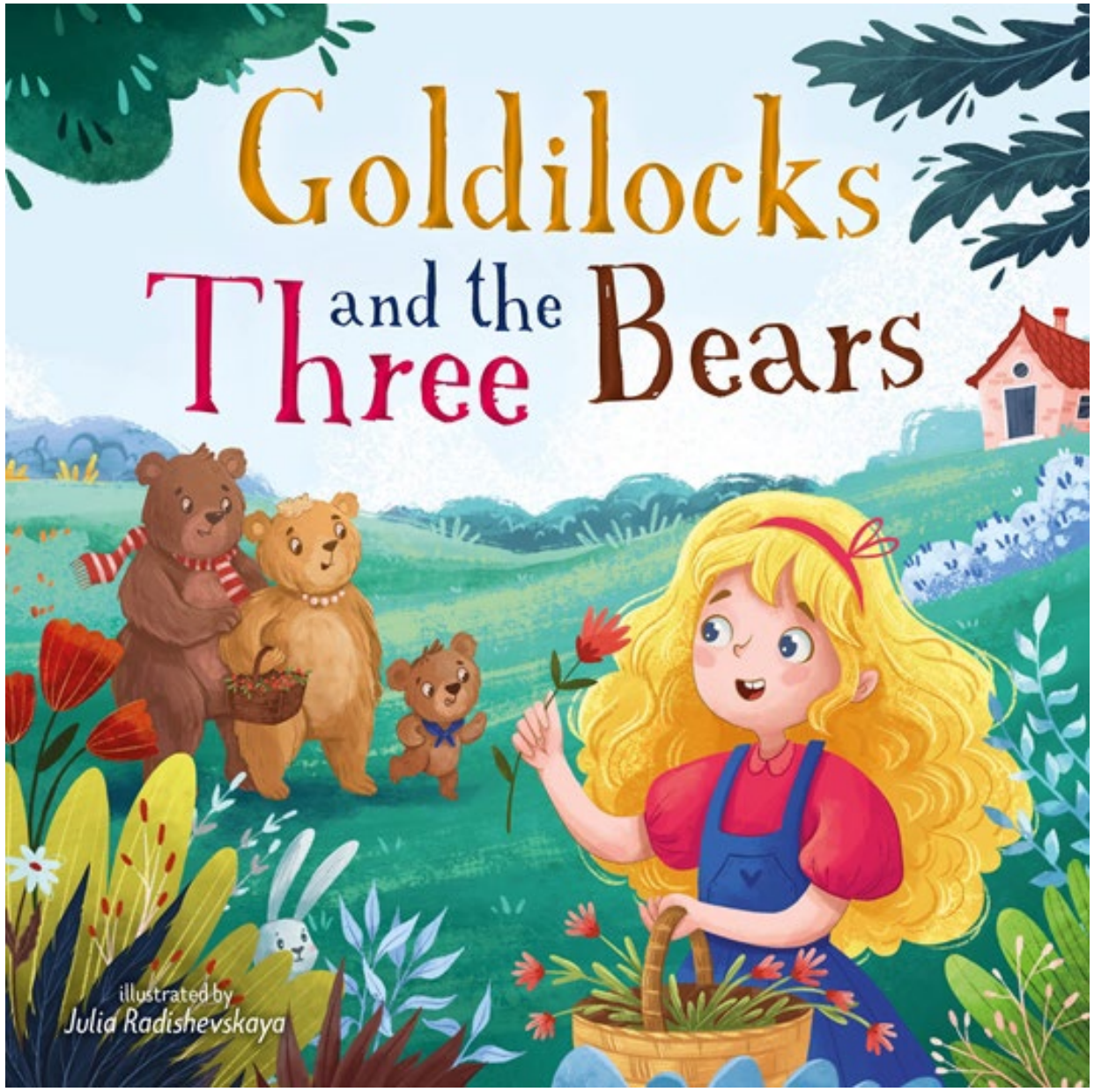
Overfitting and Underfitting





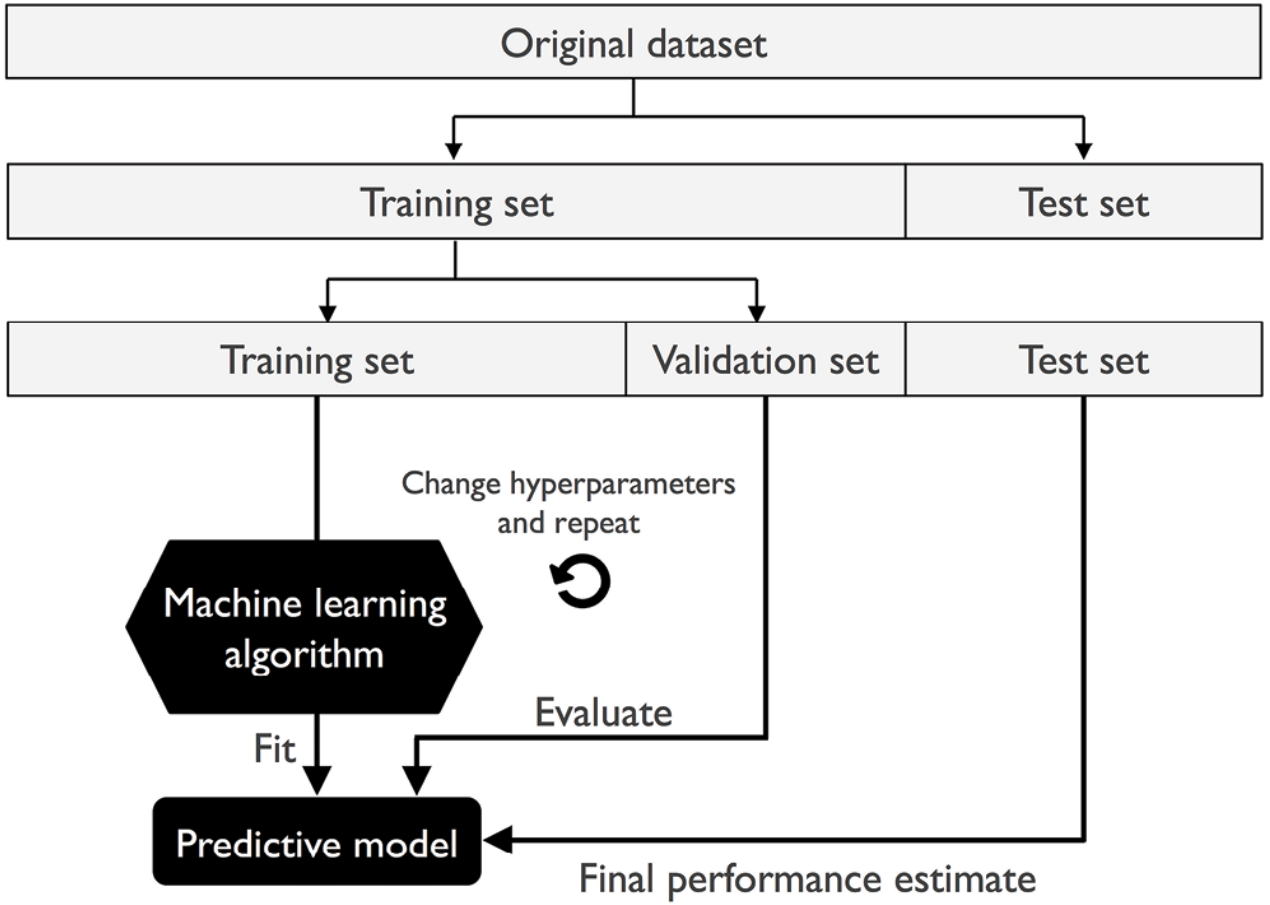
https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Overfitting and Underfitting



Training, testing, and validating

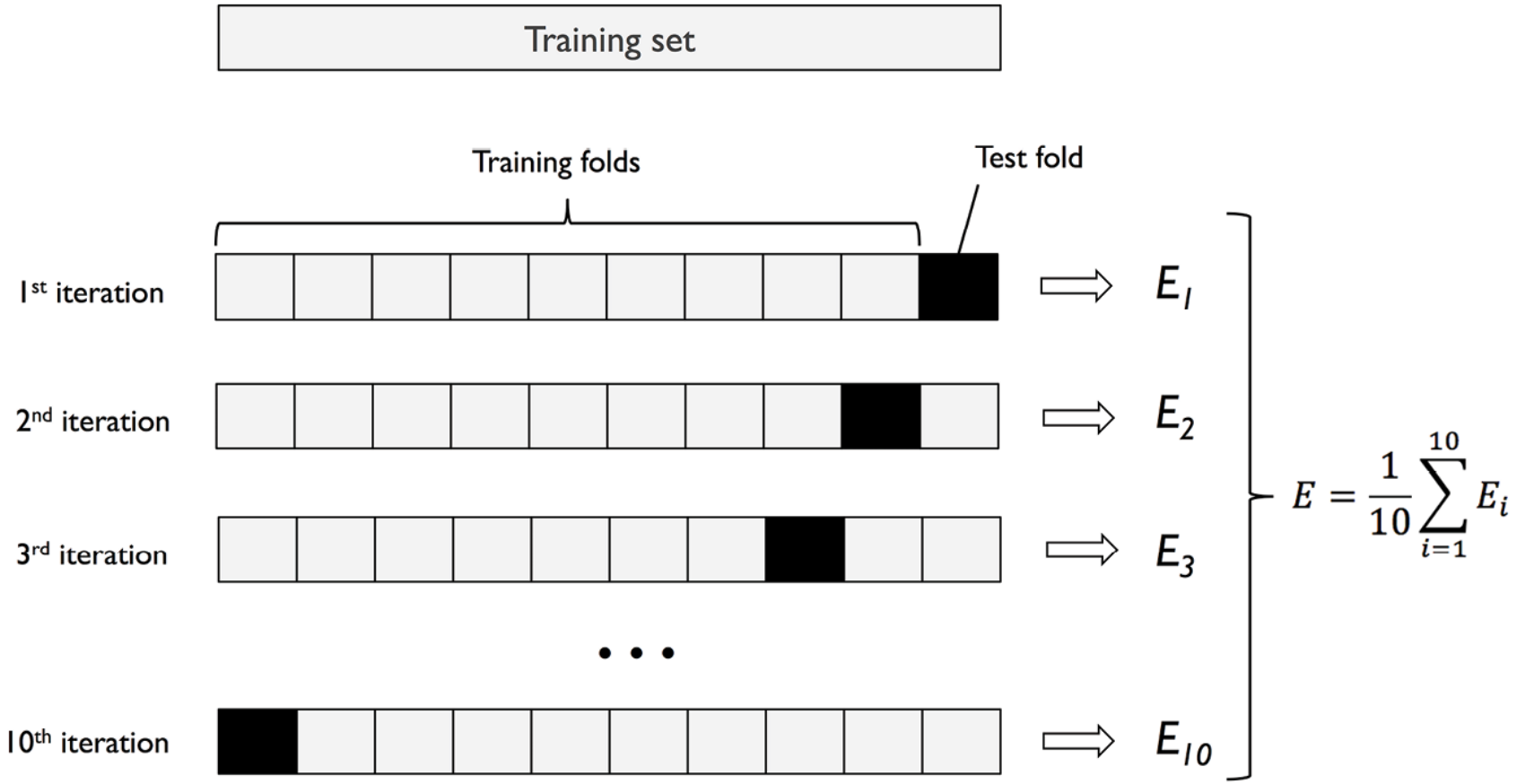
How can we be certain that model that is trained on data we have will perform well in production?



- In real world, we make decisions based on:
  - prior knowledge,
  - intuition,
  - simulations,
  - data,
  - advice
  - ....
- Now imagine we have data only!

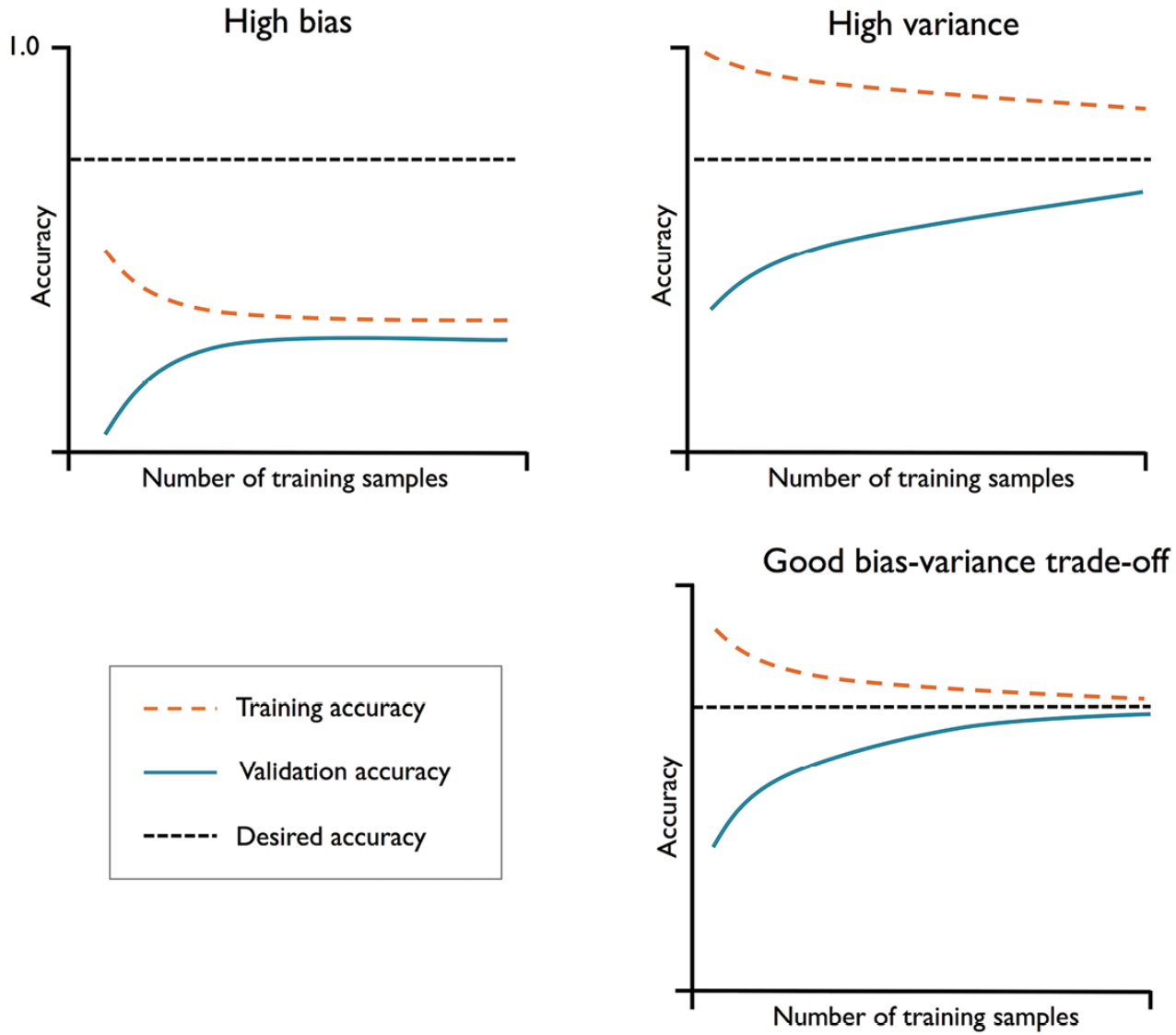
From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

k-Fold cross-validation



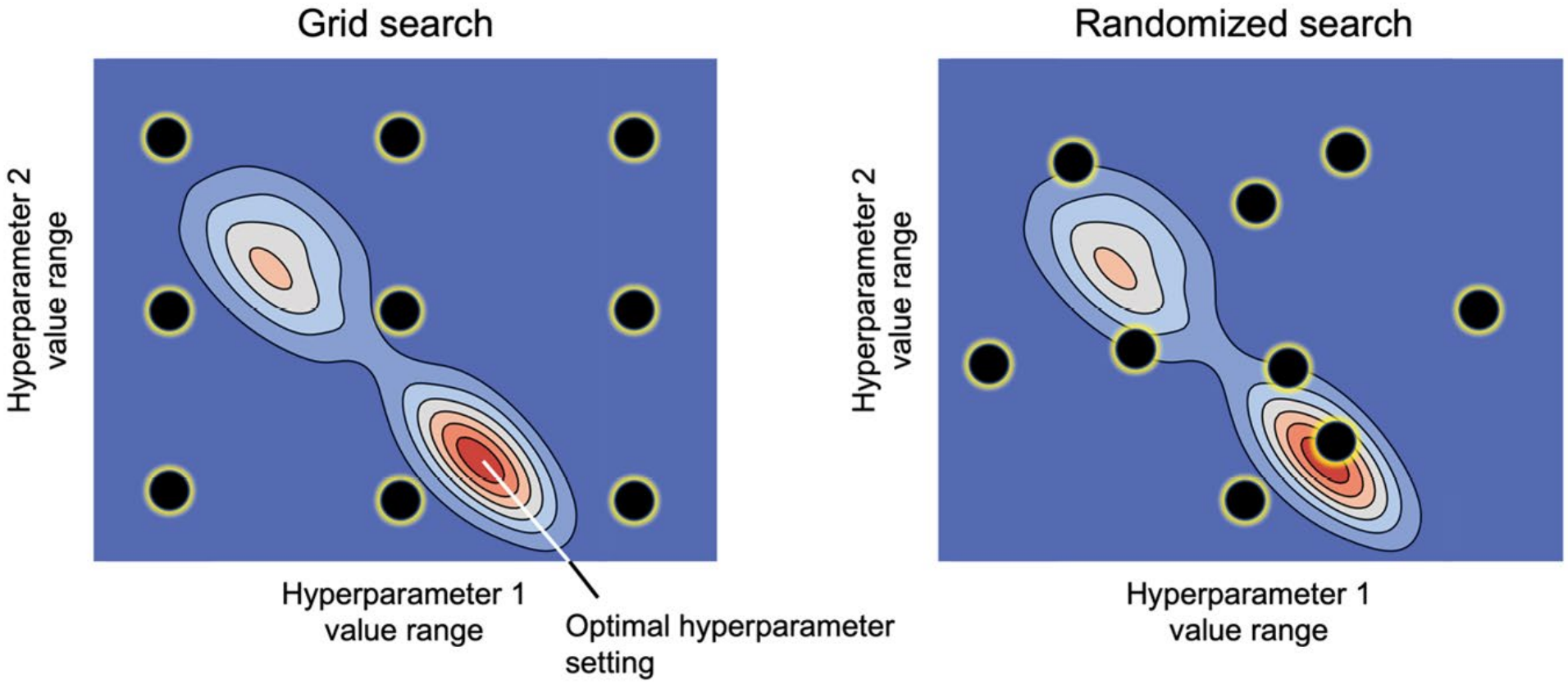
From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Bias-variance trade-off



From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Finding the right hyperparameters



From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Automated tuning hyperparameters

Parameter scape halving:

1.Draw a large set of candidate configurations via random sampling
2. Train the models with limited resources, for example, a small subset of the training data (as opposed to using the entire training set)
3. Discard the bottom 50 percent based on predictive performance
4. Go back to *step 2* with an increased amount of available resources

Hyperopt:

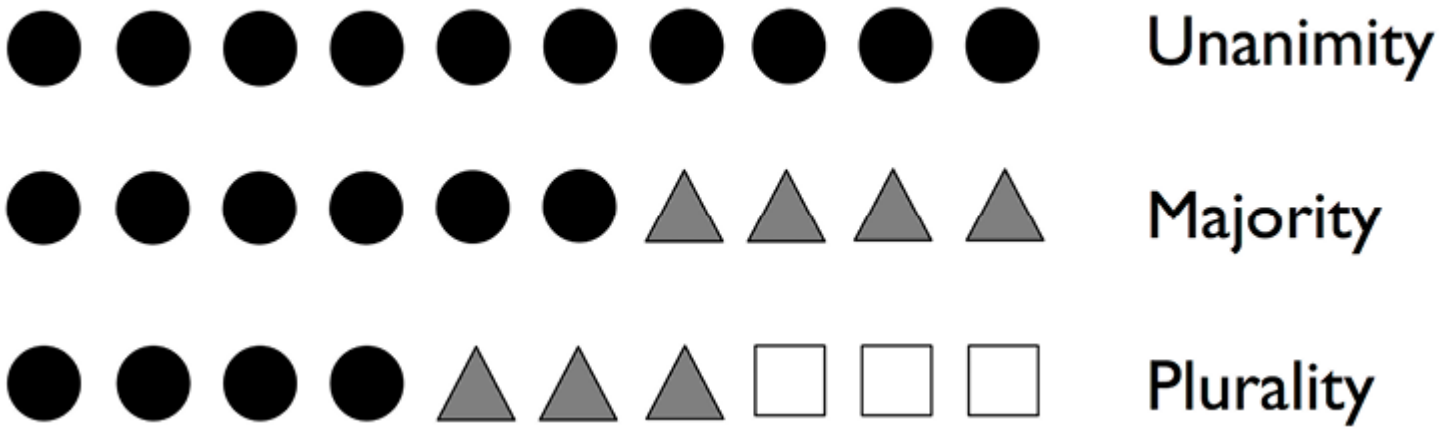
Hyperopt (https://github.com/hyperopt/hyperopt), implements several different methods for hyperparameter optimization, including randomized search and the **Tree-structured Parzen Estimators** (**TPE**) method. TPE is a Bayesian optimization method based on a probabilistic model that is continuously updated based on past hyperparameter evaluations and the associated performance scores instead of regarding these evaluations as independent events.

From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Combining weak learners

https://en.wikipedia.org/wiki/Lemmings_(video_game)

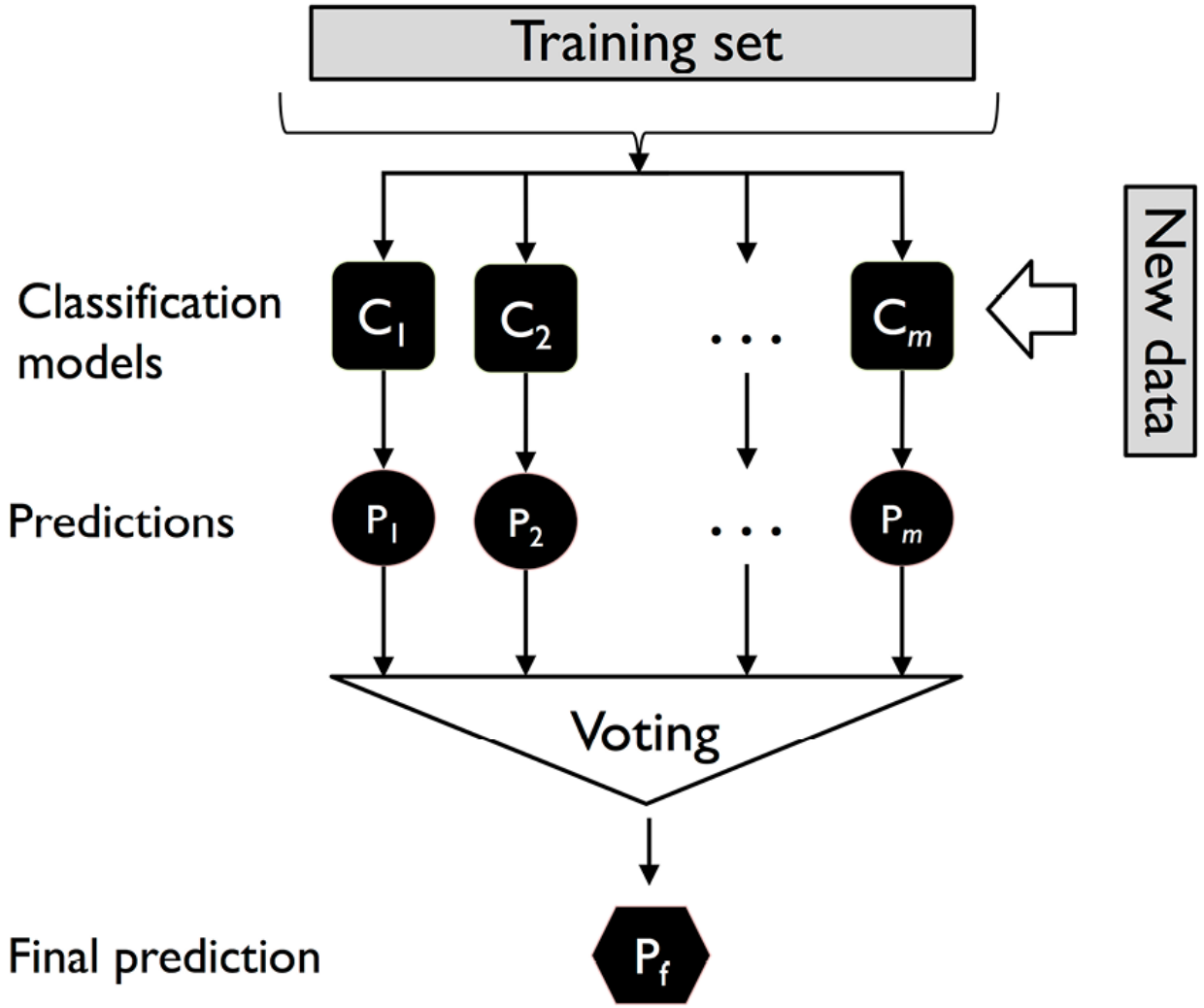
Combining weak learners



We can combine multiple weak learners into a strong learner

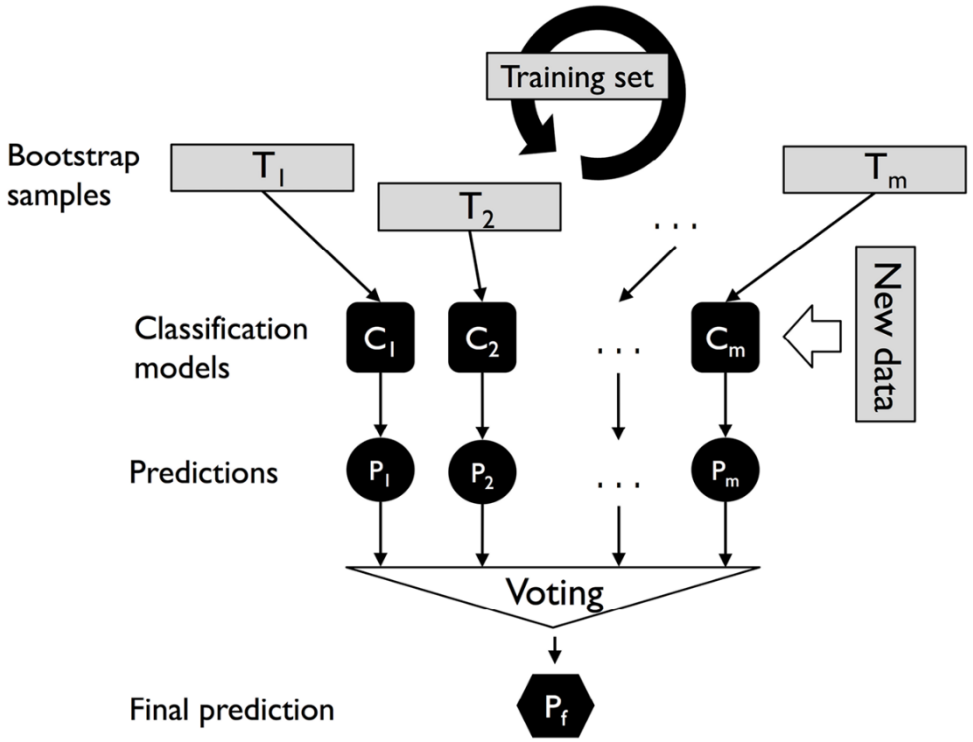
From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Combining weak learners



From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

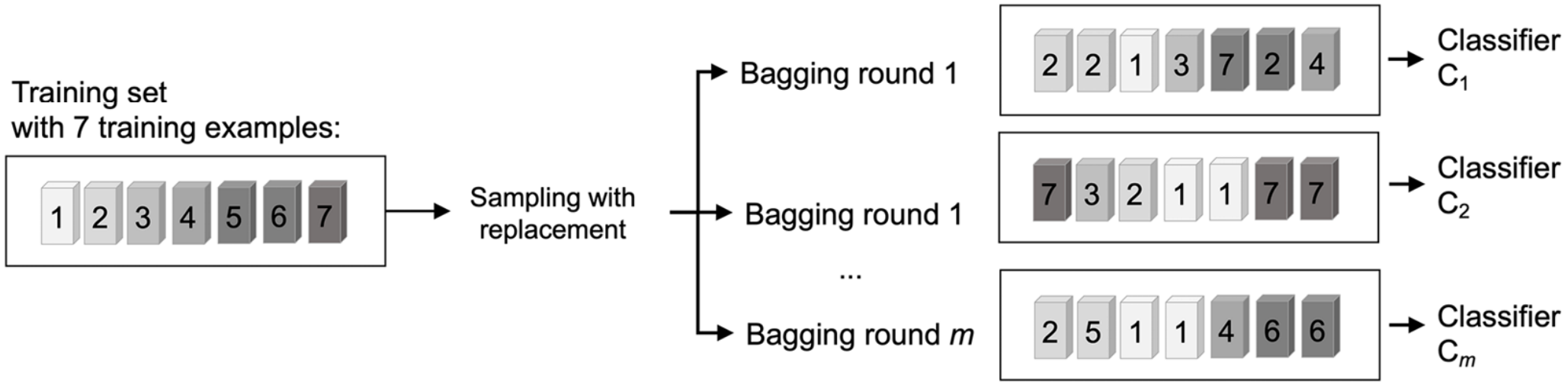
Bagging



- Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class
- The multiple versions are formed by making bootstrap replicates (samples of the data, with repetition) of the learning set and using these as new learning sets.
- Bagging can give substantial gains in accuracy.
- Bagging can improve accuracy if prediction is unstable

From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

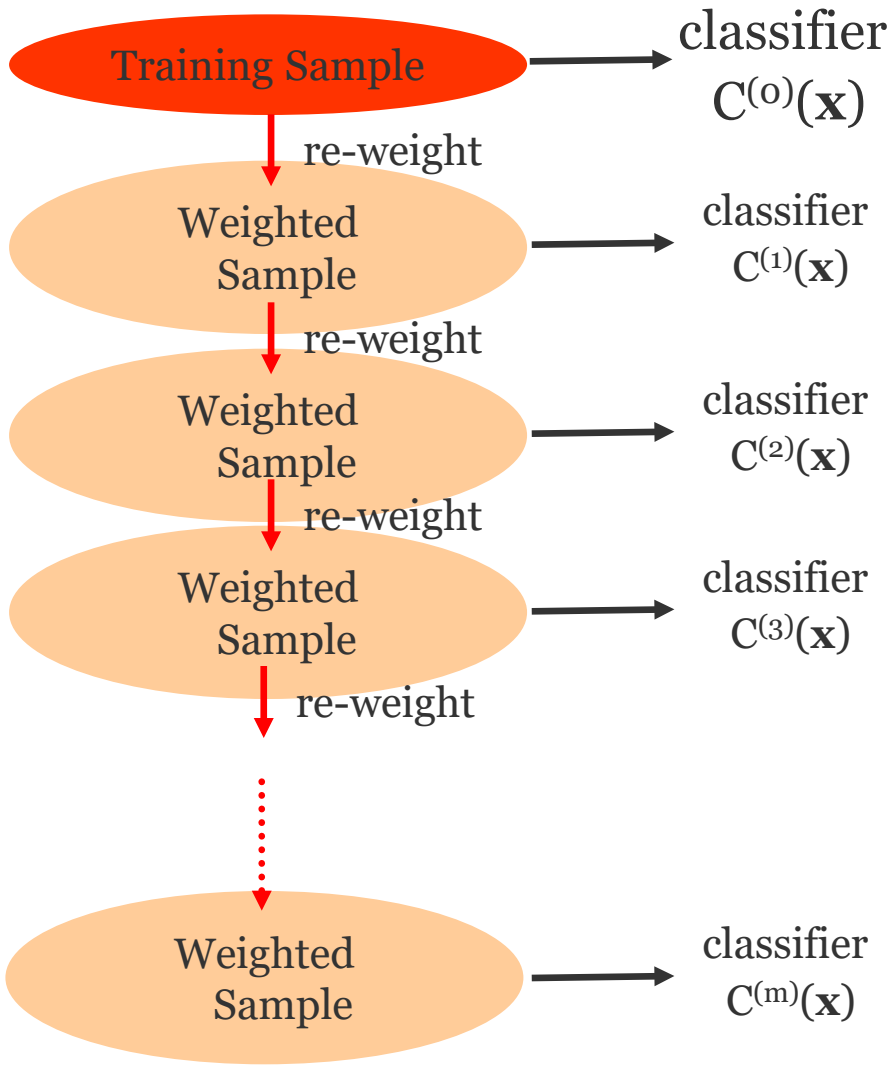
Bagging



- Draw 100 bootstrap samples of data
- Train trees on each sample → 100 trees
- Average prediction of trees on out-of-bag samples

From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

Adaptive Boosting (AdaBoost)



AdaBoost re-weights events misclassified by previous classifier by:

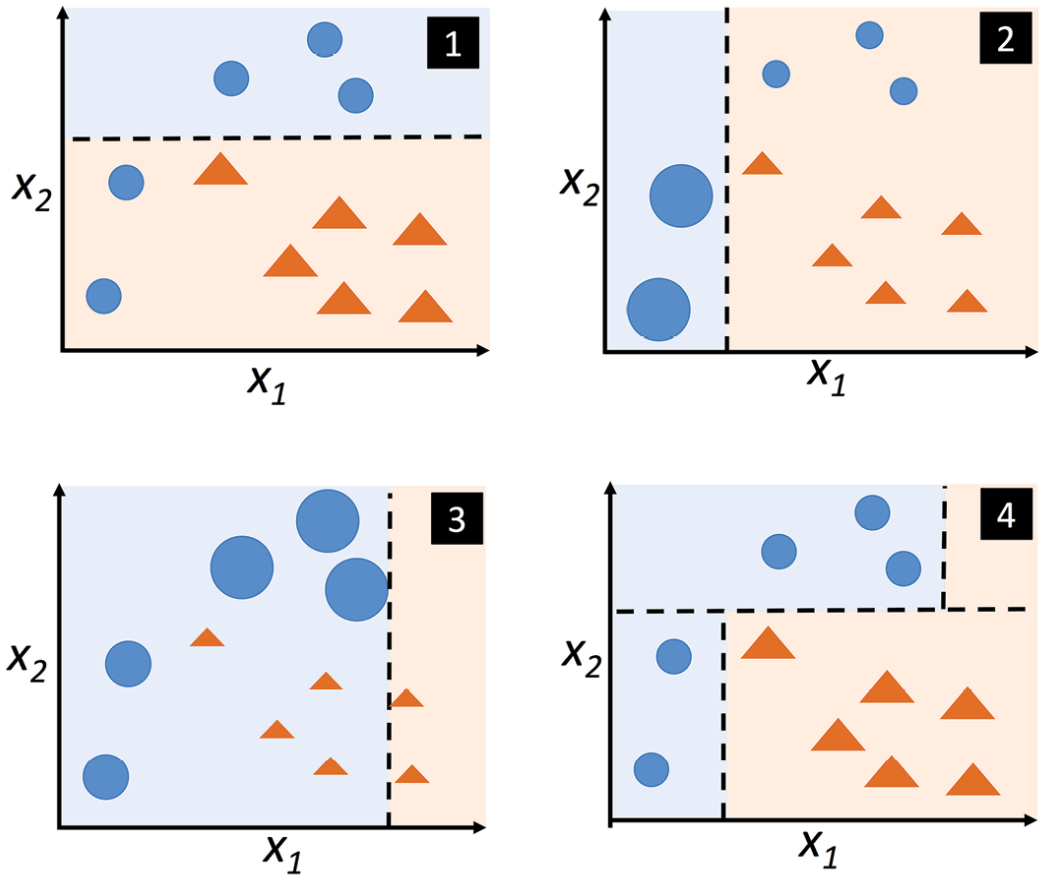
$$\frac{1-f_{err}}{f_{err}} \quad \text{with :}$$

$$f_{err} = \frac{\text{misclassified events}}{\text{all events}}$$

AdaBoost weights the classifiers also using the error rate of the individual classifier according to:

$$y(x) = \sum_{i}^{N_{Classifier}} \log\left(\frac{1-f_{err}^{(i)}}{f_{err}^{(i)}}\right) C^{(i)}(x)$$

AdaBoost



1. Draw a random subset (sample) of training examples, d_1, without replacement from the training dataset, D, to train a weak learner, C_1.

2. Draw a second random training subset, d_2, without replacement from the training dataset and add 50 percent of the examples that were previously misclassified to train a weak learner, C_2.

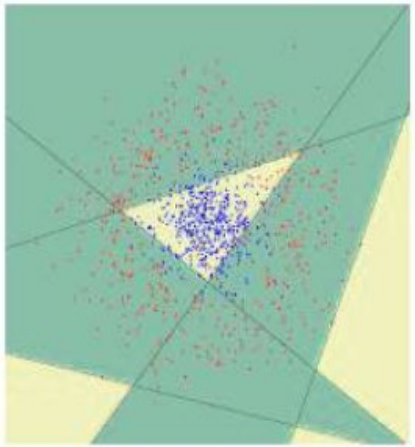
3. Find the training examples, d_3, in the training dataset, D, which C_1 and C_2 disagree upon, to train a third weak learner, C_3.

4. Combine the weak learners C_1, C_2, and C_3 via majority voting.

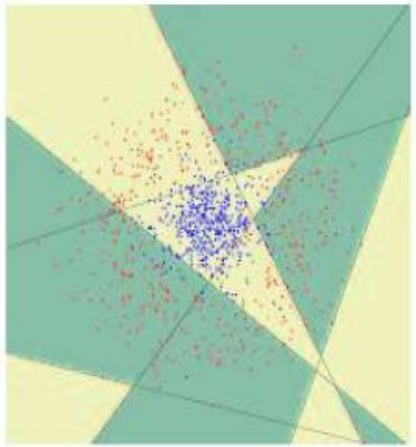
From S. Raschka, Machine Learning with PyTorch and Scikit-Learn

AdaBoost On a linear Classifier

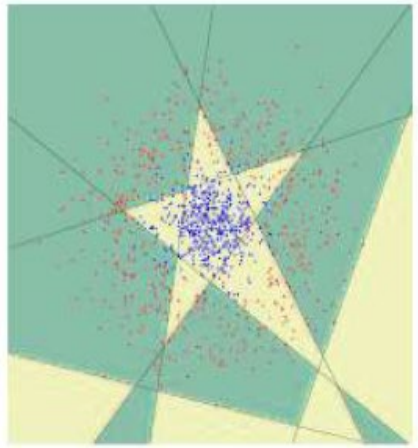
$t = 5$



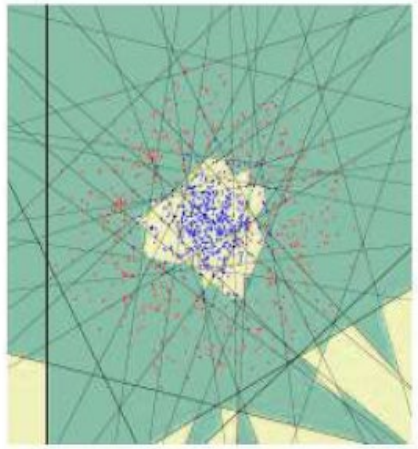
$t = 6$

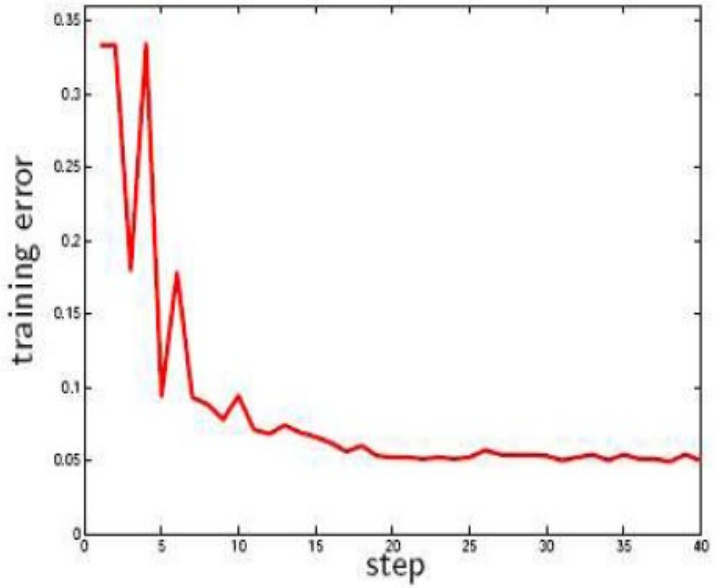


$t = 7$



$t = 40$





Boosted classifiers

1. Give events that are "difficult to categorize" more weight and average afterwards the results of all classifiers that were obtained with different weights

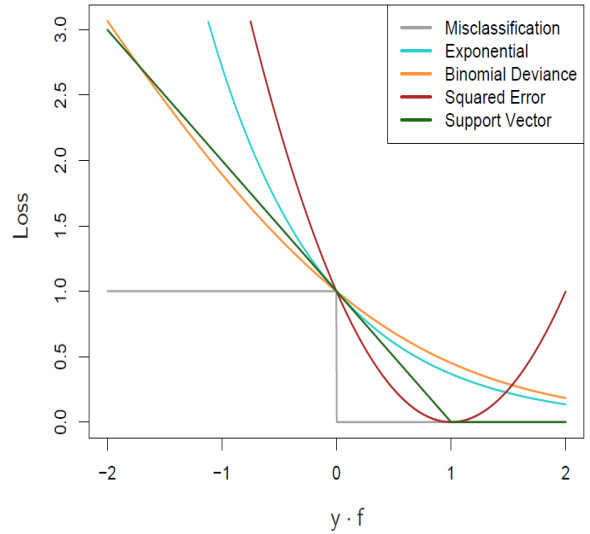
2. See each Tree as a "basis function" of a possible classifier:
   1. boosting or bagging is just a mean to generate a set of "basis functions"
   2. linear combination of basis functions gives final classifier

3. Every "boosting" algorithm can be interpreted as optimizing the loss function in a "greedy stagewise" manner, *i.e.* from the current point in the optimization – *e.g. building of the decision tree forest-* chooses the parameters for the next boost step (weights) such that one moves a long the steepest gradient of the loss function

AdaBoost: Exponential loss: $\exp(-y_0 y(\alpha,x))$

- theoretically sensitive to outliers

Binomial log-likelihood loss: $\ln(1 + \exp(-2y_0 y(\alpha,x))$ -
more well-behaved loss function



Note of warning: you cannot do better then data

How materials are discovered?

Corning Ware glass was accidentally discovered via a furnace mishap

"The temperature gauge was stuck on 900 degrees,
and I thought I had ruined the furnace …

I grabbed some tongs to get it out as fast as I could,
but the glass slipped out of the tongs and fell to the floor.

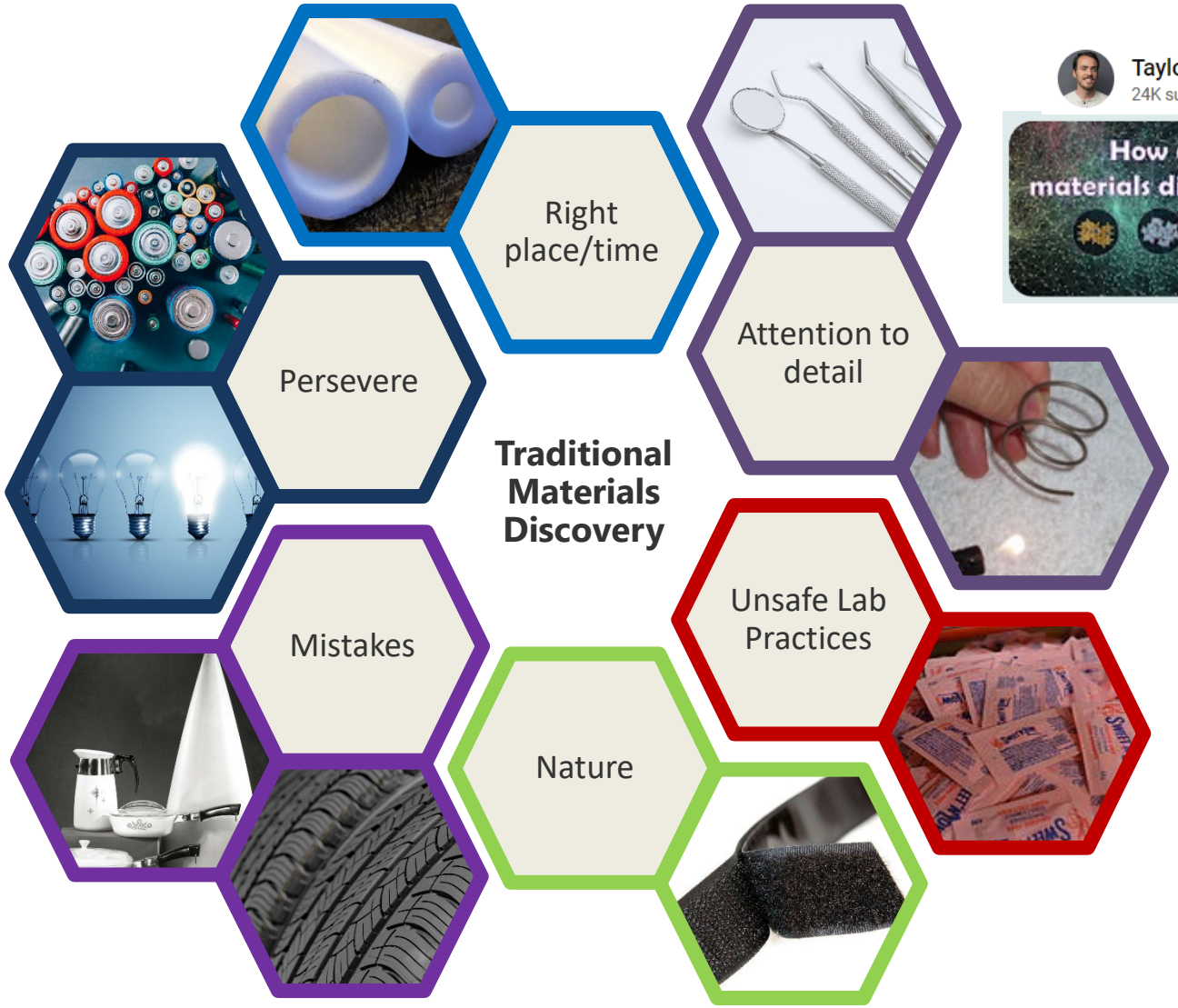
The thing bounced and didn't break."

Donald Stookey (1915-2014)

https://cen.acs.org/articles/92/web/2014/12/Donald-StookeyGuy-Gave-Us-Corning.html
https://www.nytimes.com/2014/11/07/business/s-donald-stookey-inventor-of-corningware-dies-at-99.html

Slide by Sterling Baird

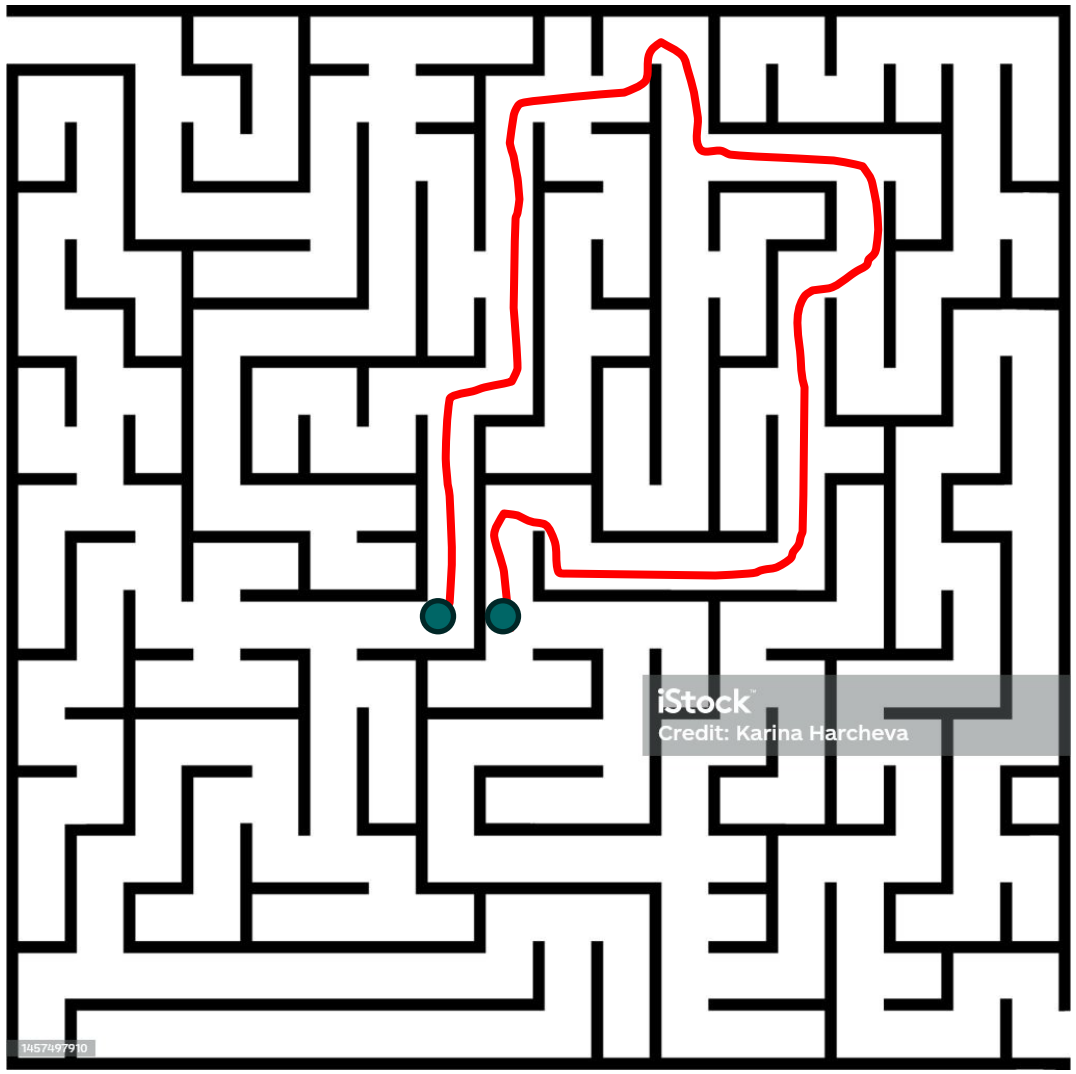
How materials are discovered?



"How are materials discovered?" by Taylor Sparks: https://www.youtube.com/watch?v=RRNcqJSJ6vc

Slide by Sterling Baird

How materials are discovered?



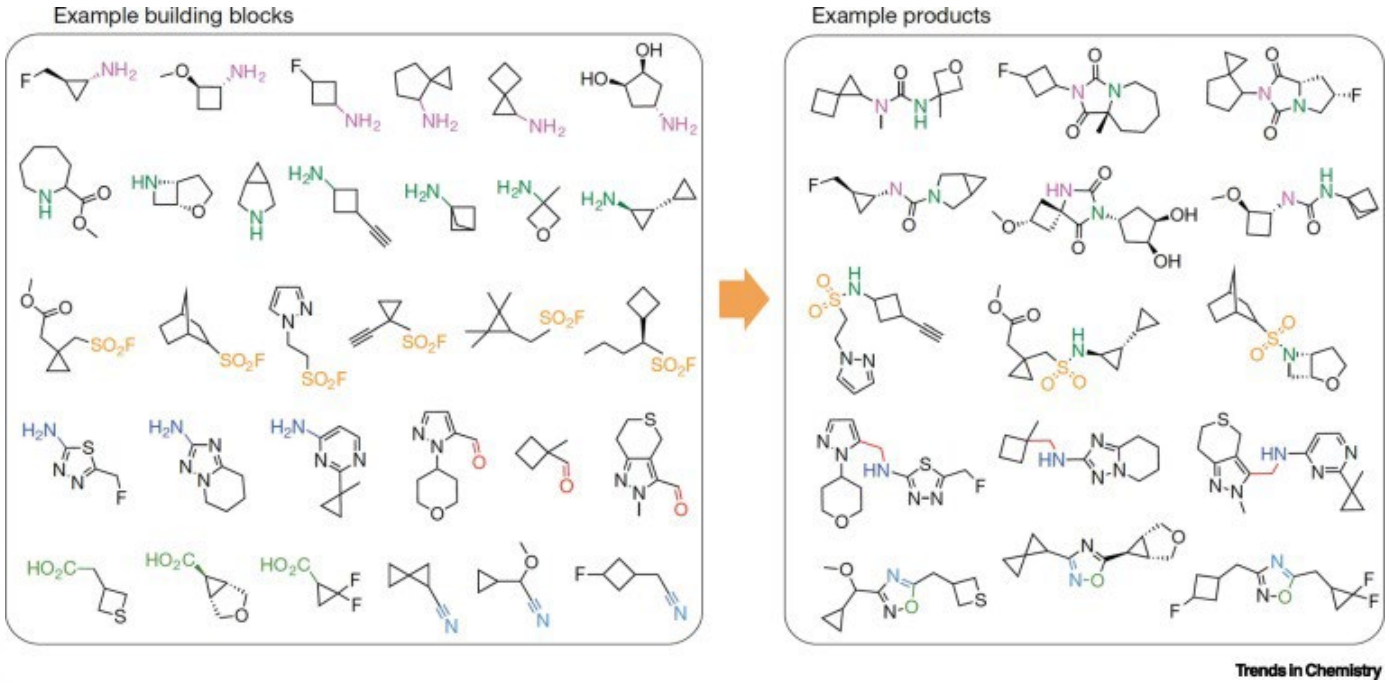
https://en.wikipedia.org/wiki/LK-99

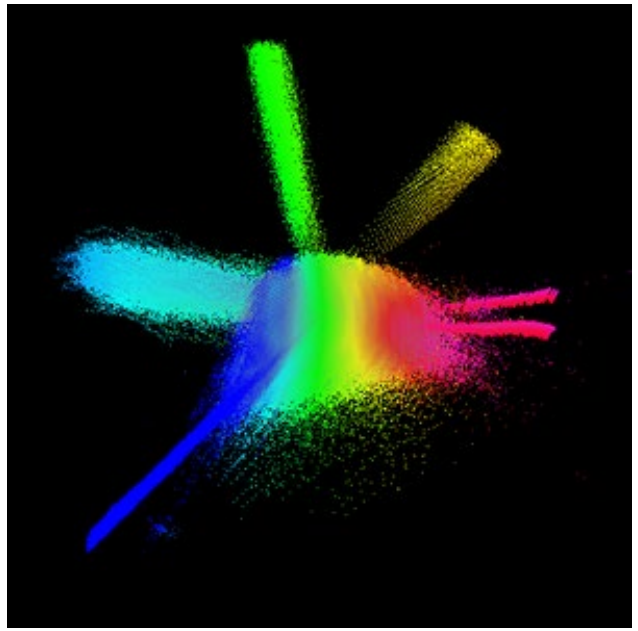
1986 – $YBa_2Cu_3O_7$. Gave rise to multiple families of Cu and Hg superconductors

2001 – MgB_2. Point compound

2006 - Layered iron pnictides. Gave rise to multiple families of superconductors

How many molecules are there?



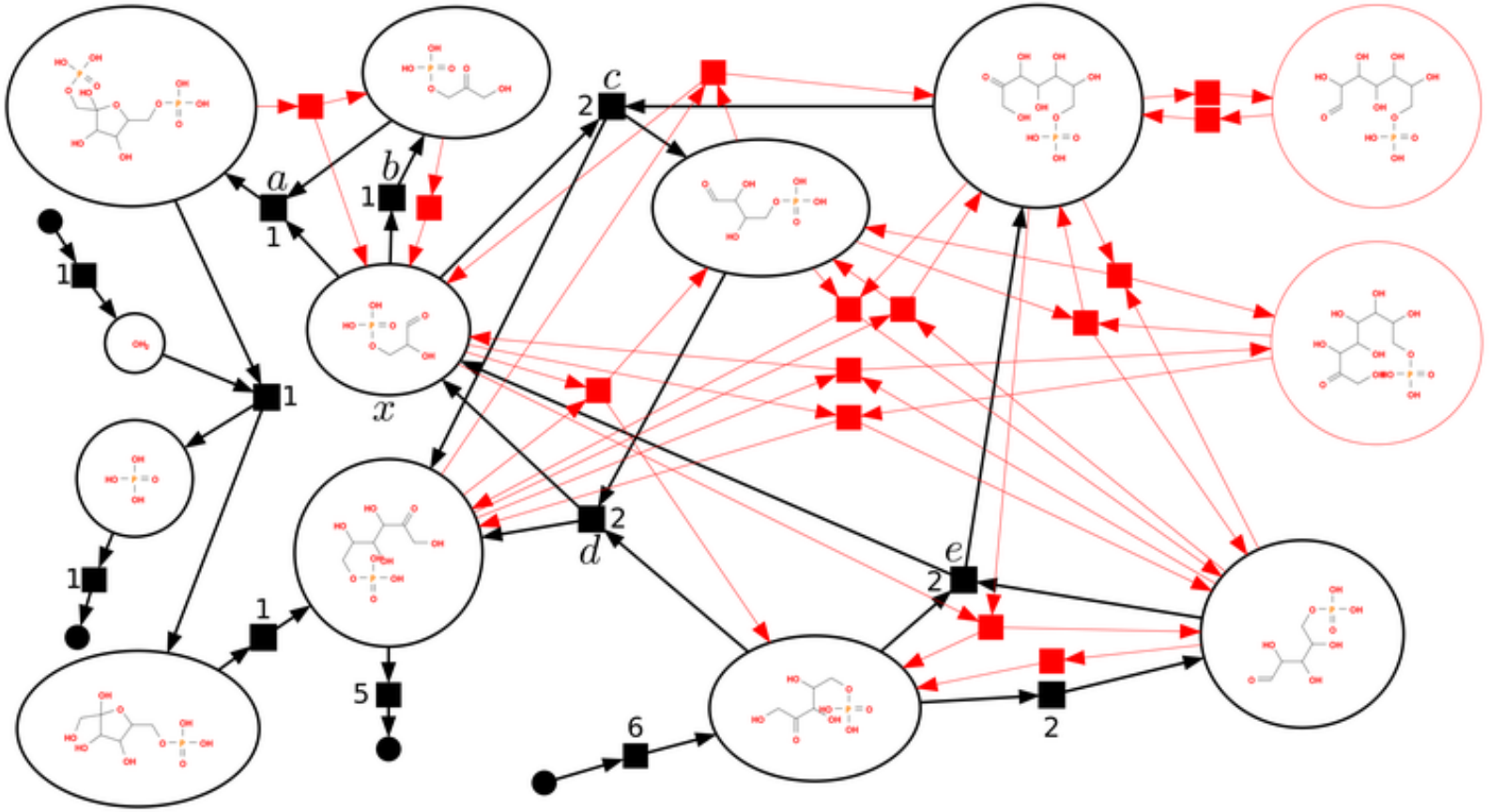


A chemical space often referred to in cheminformatics is that of potential biologically active molecules. Its size is estimated to be in the order of 10^{60} molecules. The estimate restricts the chemical elements used to be C, H, O, N and S. It further makes the assumption of a maximum of 30 atoms to stay below 500 Daltons, allows for branching and a maximum of 4 rings and arrives at an estimate of 10^{63}.

https://www.cell.com/trends/chemistry/fulltext/S2589-5974%2820%2930288-4

https://en.wikipedia.org/wiki/Chemical_space

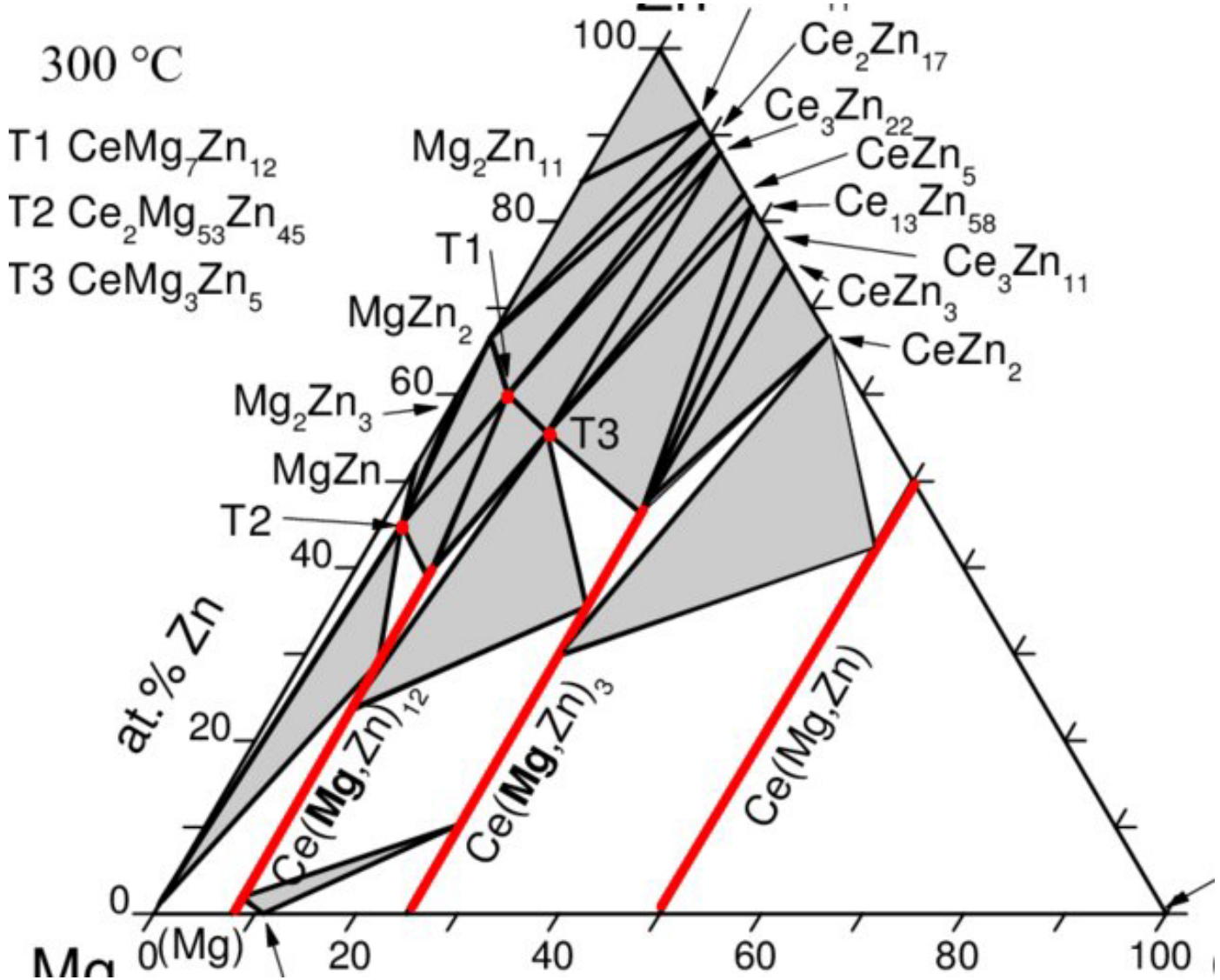
Chemical reactions networks:



- Molecular property predictions: are they **likely** to be useful?
- Synthesizability scores: what would it **probably** take to make them
- Reaction network mining and retrosynthesis: can we identify **possible** synthetic pathways?
- Optimization of specific reaction conditions and pathways: myopic and non-myopic

https://www.mis.mpg.de/stadler/research/chemical-reaction-networks.html

Why dimensionality is a problem?

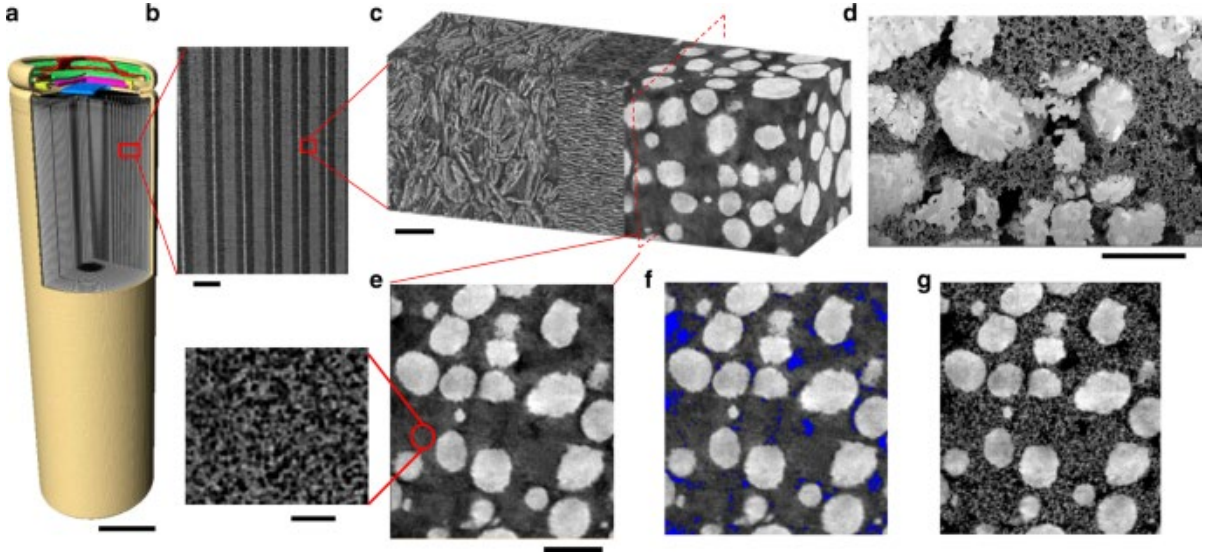


Let's think about it as a search problem:

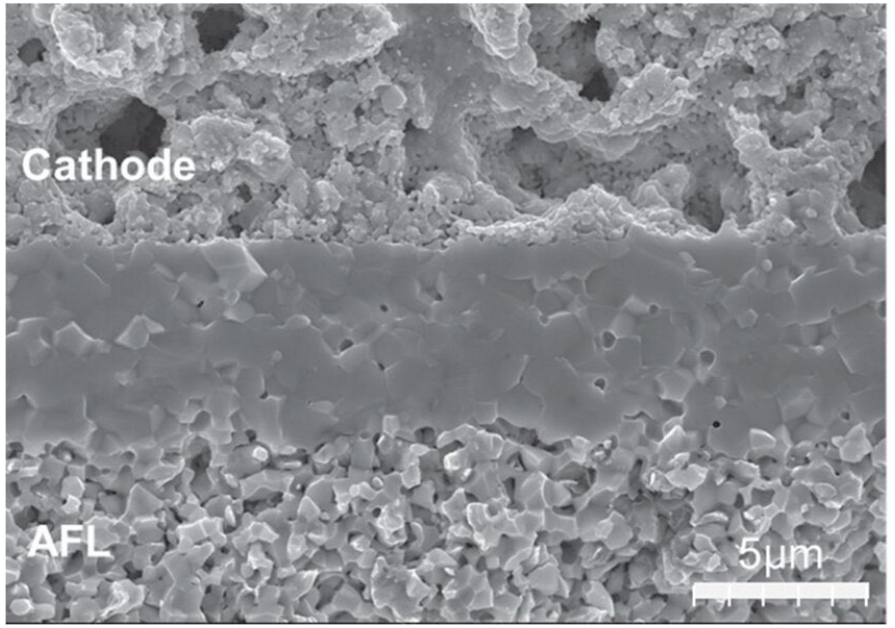
- **Alloying:** need maintain composition ~1%
- **Doping:** need maintain composition ~ 10^{-6}
- Grid search is out for D > 3 (experiment)

Microstructures

https://manyeats.com/damascus-steel-and-its-modern-attempts/

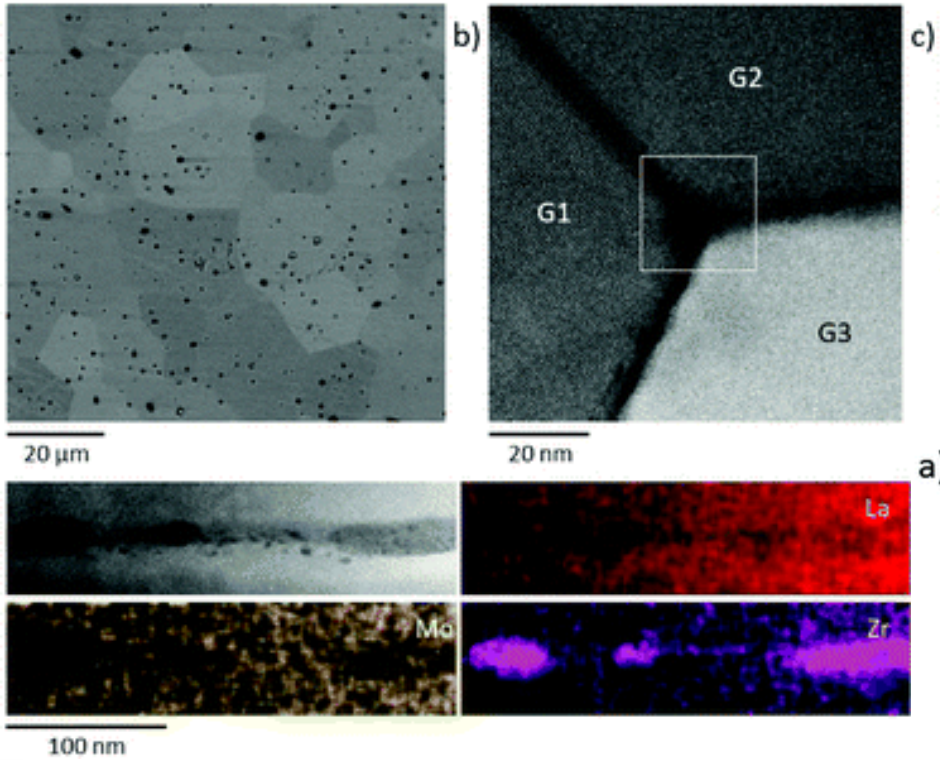


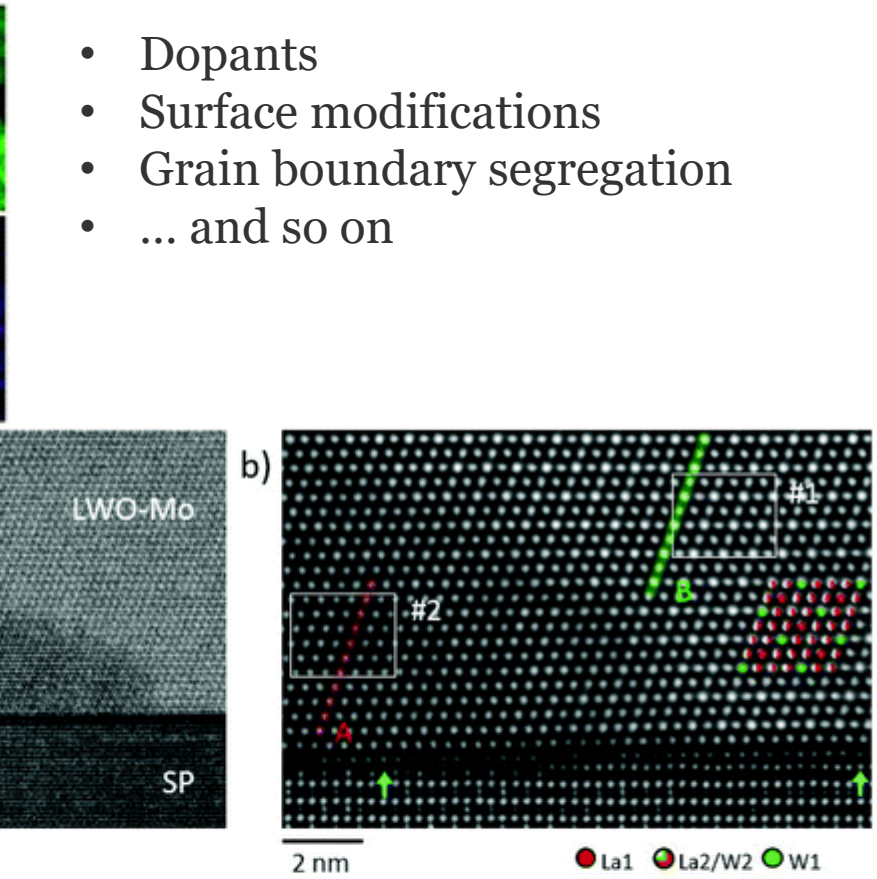
https://www.nature.com/articles/s41467-020-15811-x



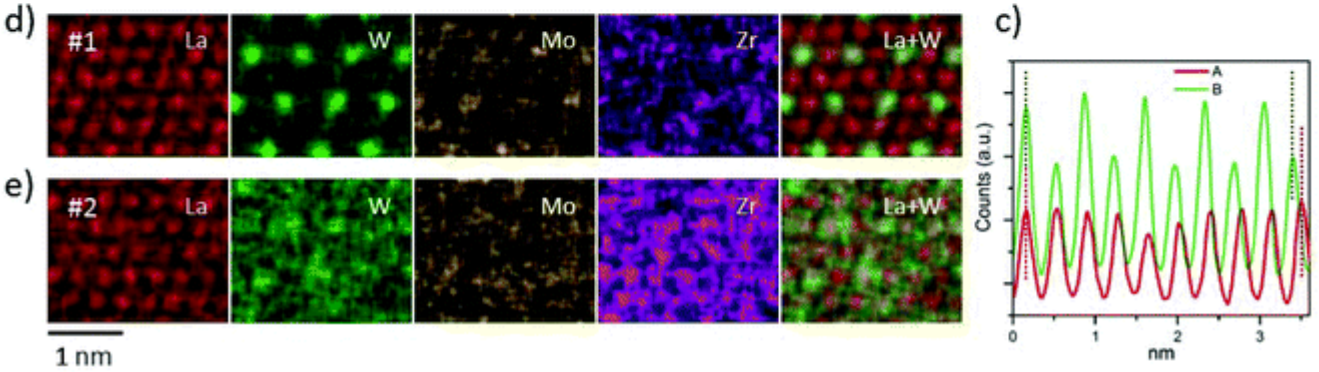
https://www.researchgate.net/figure/Microstructure-of-the-electrolyte-electrode-interface-region-in-the-fuel-cell-with-

Impurities Matter



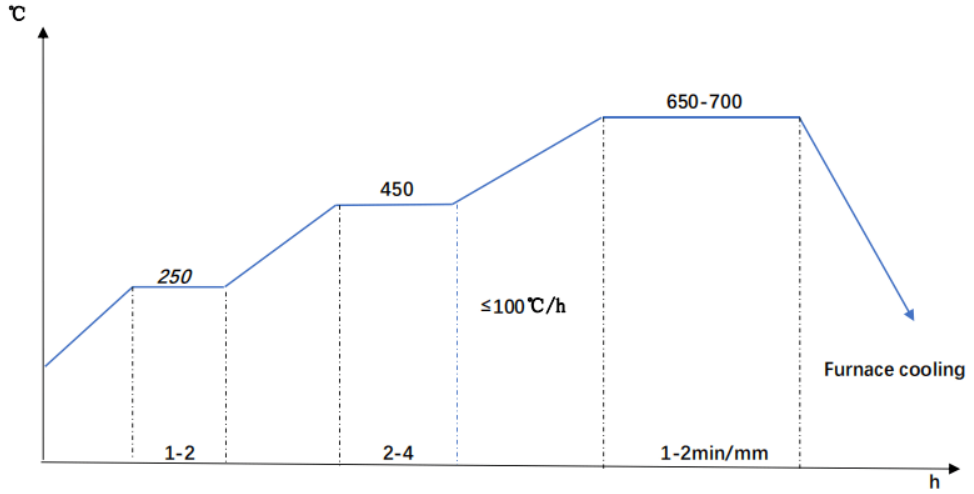


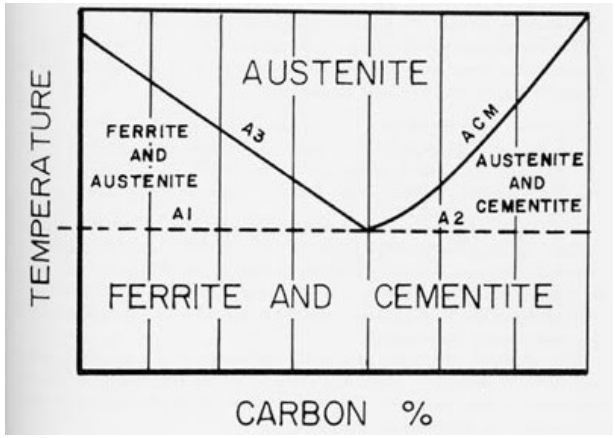
- Dopants
- Surface modifications
- Grain boundary segregation
- … and so on

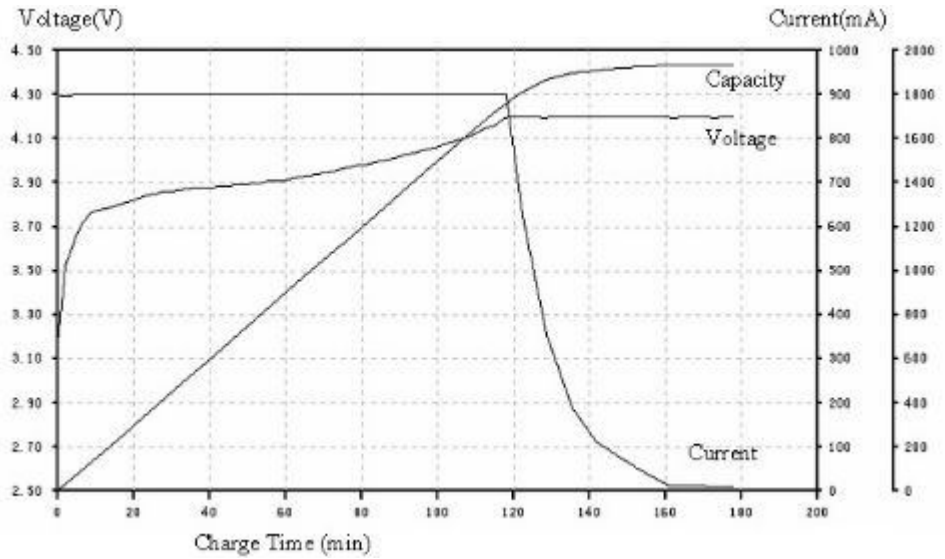


https://pubs.rsc.org/en/content/articlelanding/2020/nr/d0nr04396f

Making materials: process trajectories







- Making steel: complicated and took a lot of time optimize
- Charging battery: obvious economic impact
- Manufacturing: Annealing hybrid perovskite thin films
- Poling ferroelectric

How do we optimize trajectories if we have (a) only limited or no mechanistic information, (b) our experimental budgets are limited, but (c) we have some access to domain expertise?

Why dimensionality is a problem?

- Suppose that we have data for 1000 students' performance (discretized scores of 0; 25; 50; 75; 100)% in 2 courses c1 and c2. Then in total there are 5 x 5 = 25 different grade combinations.

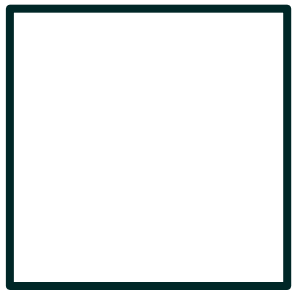
- If the 1000 students are randomly distributed among each grade combination, then on average there are 40 students with each possible grade combination, which is a good enough sample to draw conclusions such as if, for a student, grade(c1) 50 and grade(c2) 75, then that student is likely to be a Math major.

- Now suppose there are 4 courses, then the number of possible grades combination is 54 = 625, and an average number of students per combination is 1:6. For 10 courses, this number reduces to 0:0001024. This means that almost all possible combinations are never observed.

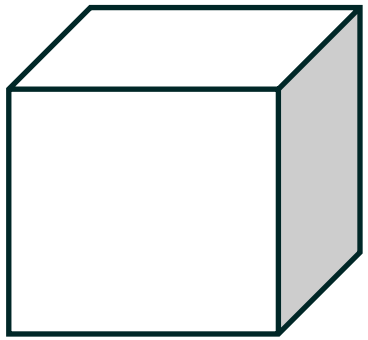
From "Curse of Dimensionality", Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

- Suppose *n* points in X are chosen uniformly at random from $[0; 1]^m$ (m-cube). For the query point q grow a hypercube around *q* to contain *f* fraction of points ($k = f\ n$) in X. This cube (the search space for q) grows very large (covering almost the whole input space) in large dimension.

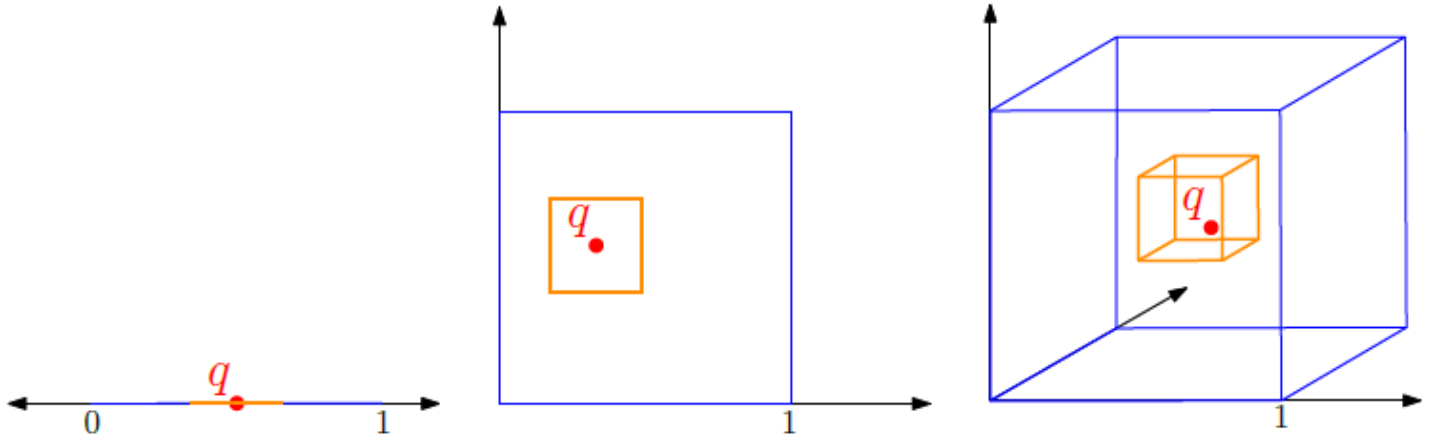


- The expected length of the edge of the search cube $E_m(f) = f^{1/m}$, i.e. in 10d to get 10% points around q need cube with edge length 0.8 (which is 80% of the whole cube, the input space). Similarly, to get only 1% points one needs to extend the search cube by 0:63 units along each dimension



From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

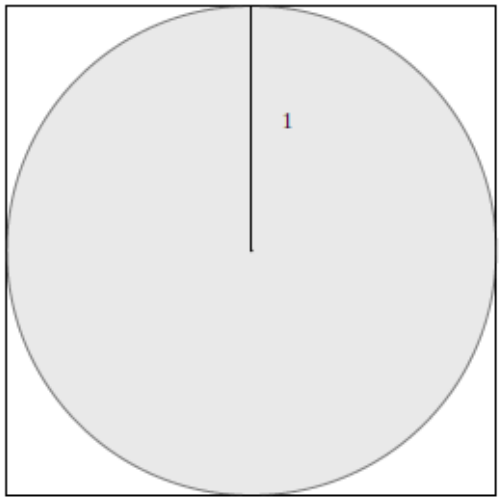


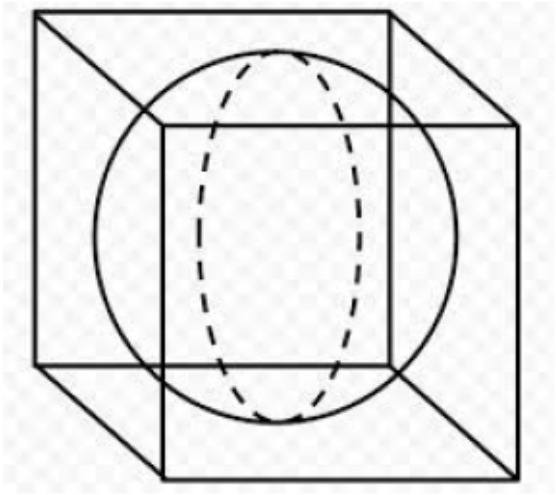
Suppose we have 5000 points:

- In 1d we have to explore 0.001 on average to capture 5 NN
- In 2d, on average we must explore 0:031 units along both dimensions to get 5 nearest neighbors points (about 3% of the whole cube).
- In 3d, on average we must go 10% of the total (unit) length in each of the 3 dimensions
- In 4d, we must explore 17:7% of unit length
- In 10d, we must go 50.1% of unit length along each dimension

From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?





dim m	volume of m-ball	volume of m-cube	ratio
2	π	2^2	~ 0.785
3	$4/3\pi$	2^3	~ 0.523
4	$\pi^2/2$	2^4	~ 0.308
6	$\pi^3/6$	2^6	~ 0.080
m	$\frac{\pi^{m/2}}{m/2!}$	2^m	$\to 0$

From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Why dimensionality is a problem?

However if a dataset exhibit this phenomenon that the issue has be overcome by getting a larger training set (exponential in m). One way to look at this is as follows.

To cover $[-1,1]^m$ with $B_{m,1}$'s, the number of balls n must be

$$n \geq \frac{2^m}{V_m(1)} = \frac{2^m}{\pi^{m/2}/m/2!} = \frac{m/2!\, 2^m}{\pi^{m/2}} \overset{m\to\infty}{\sim} \sqrt{m\pi}\left(\frac{m2^{m/2}}{2\pi e}\right)^{m/2}$$

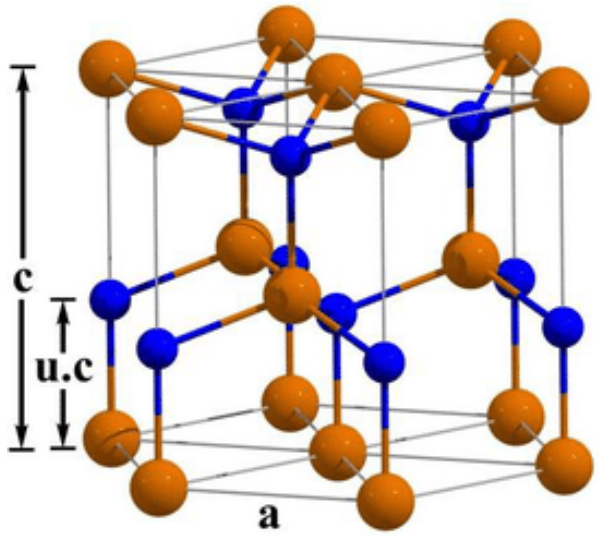
For $m = 16$ (a very small number) this n is substantially larger than 2^{58}

- In higher dimensions all the volume is in `corners'
- Points in high dimensional spaces are isolated (empty surrounding)
- The probability that a randomly generated point is within r radius of q approaches 0 as dimensionality increases
- The probability of a close nearest neighbor in a data set is very small

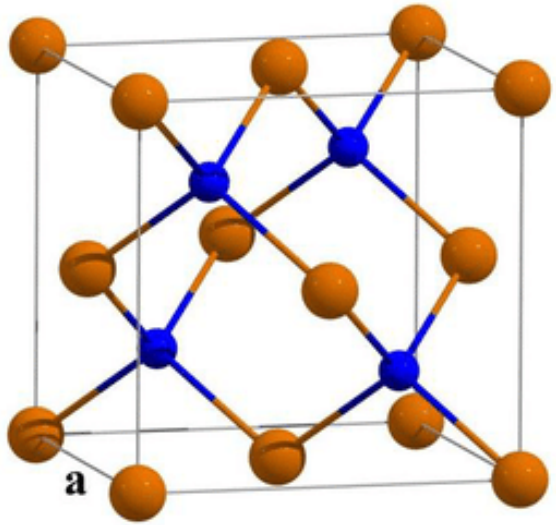
From “Curse of Dimensionality”, Lecture Notes for Big Data Analytics, Faizad Ullah, February 2019

Binary Octet Compounds

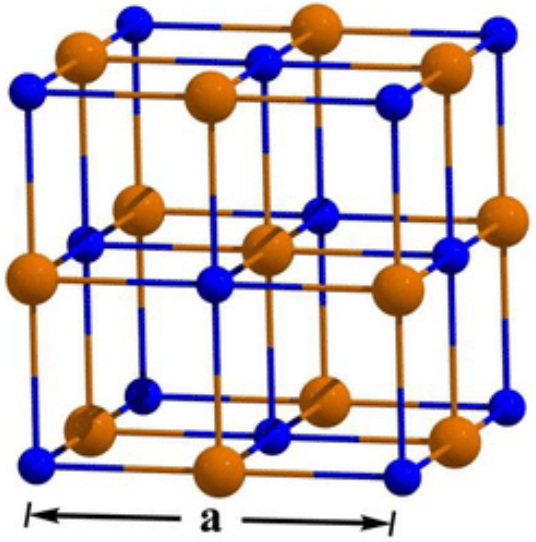
- NaCl, LiI, BeO, AlN, ….
- Can exist in zincblende (ZB), wurtzite (WZ), rocksalt (RS), cesium chloride (CsCl), and diamond cubic (DC) crystal structures



(a) wurtzite



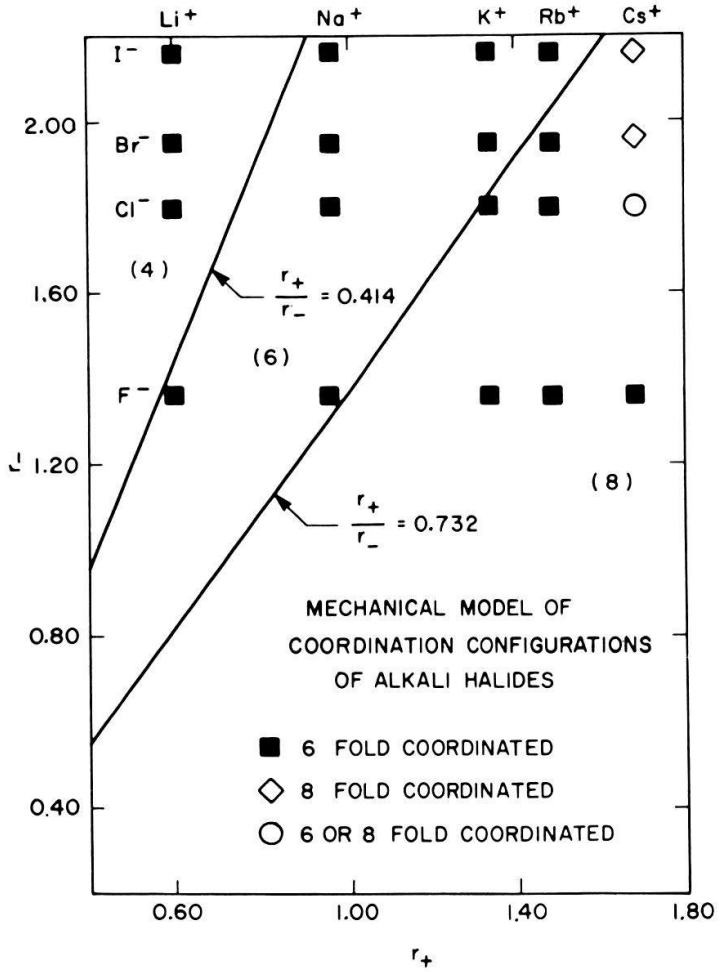
(b) zinc-blende



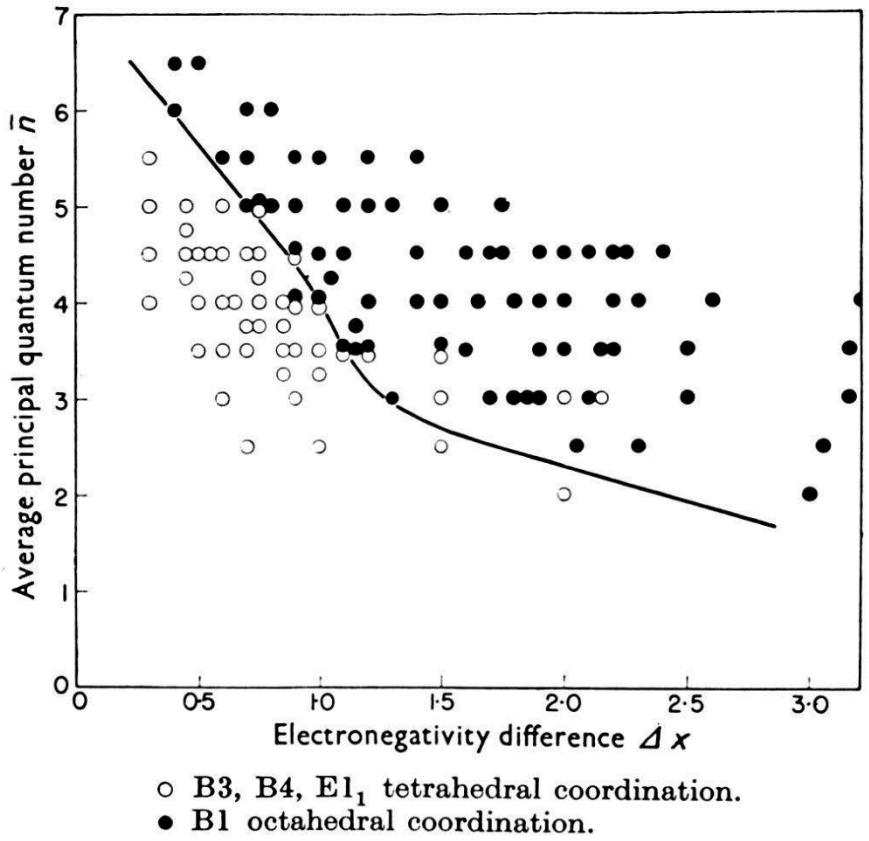
(c) rock-salt

T. Wonglakhon and D. Zahn, Interaction Potentials for modelling GaN precipitation and solid state polymorphism, Journal of Physics Condensed Matter 32(20), DOI:10.1088/1361-648X/ab6cbe

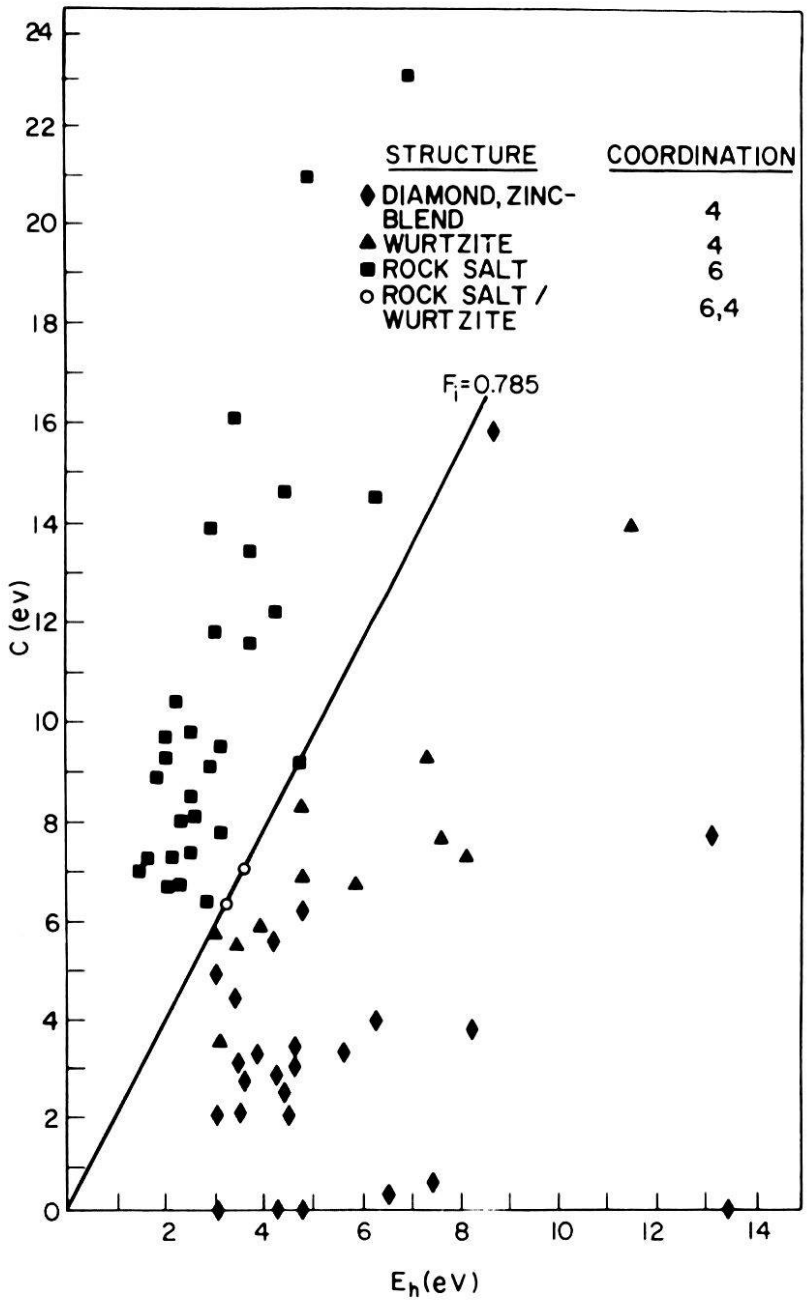
Can we predict the structure from composition?



Mooser-Pearson plots, 1959



J. C. Phillips, Structure and Properties: Mooser-Pearson plots, Helvetica Physica Acta, Vol. 58 (1985)



This average energy gap E_g was separated into covalent and ionic components, E_h and C respectively, by a Hückel relation $E_g^2 = E_h^2 + C^2$. One could then determine E_h and C separately by scaling the former with the bond length d and obtain E_g and C from ϵ. In this model the transformation from tetrahedral to octahedral coordination depends on the fraction of ionic character in the chemical bond given by $f_i = C^2/E_g^2$.

The Phillips-Van Vechten plot for AB valence compounds utilizing 'symmetric' energy-gap coordinates Eh and C. The use of quantum-mechanically defined coordinates, together with the restriction to valence compounds and exclusion of transition-metal compounds, leads to an exact separation with a straight line corresponding to constant critical ionicity.

J. C. Phillips, Structure and Properties: Mooser-Pearson plots, Helvetica Physica Acta, Vol. 58 (1985)

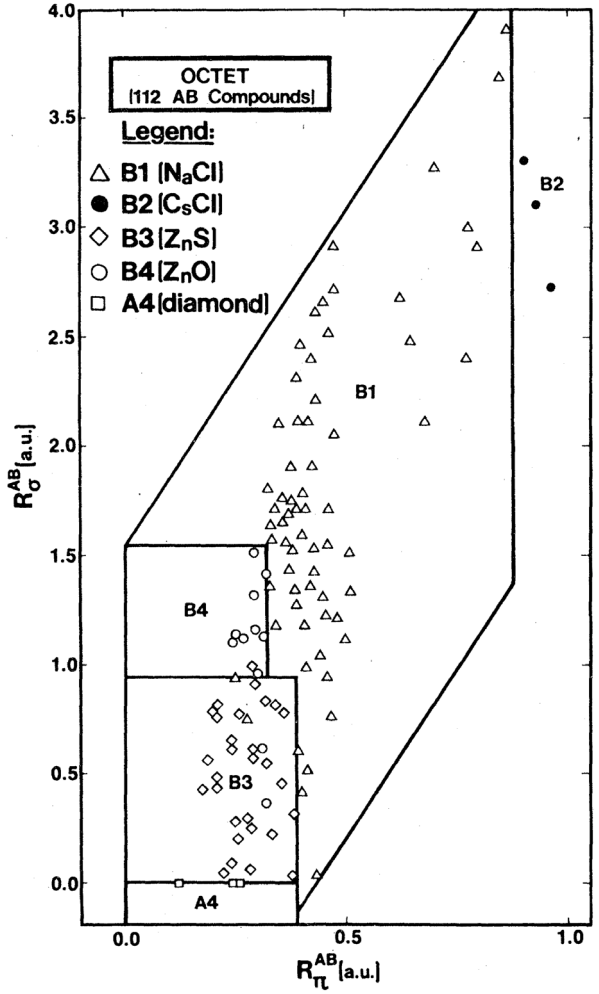


FIG. 19. Structural separation plot for the 112 binary octet compounds $A^N B^{(8-N)}$, obtained with the density-functional orbital radii, with

$$R_\sigma^{AB} = |\,(r_p^A + r_s^A) - (r_p^B + r_s^B)\,|,$$

$$R_\pi^{AB} = |\,r_p^A - r_s^A\,| + |\,r_p^B - r_s^B\,|.$$

Villars diagrams

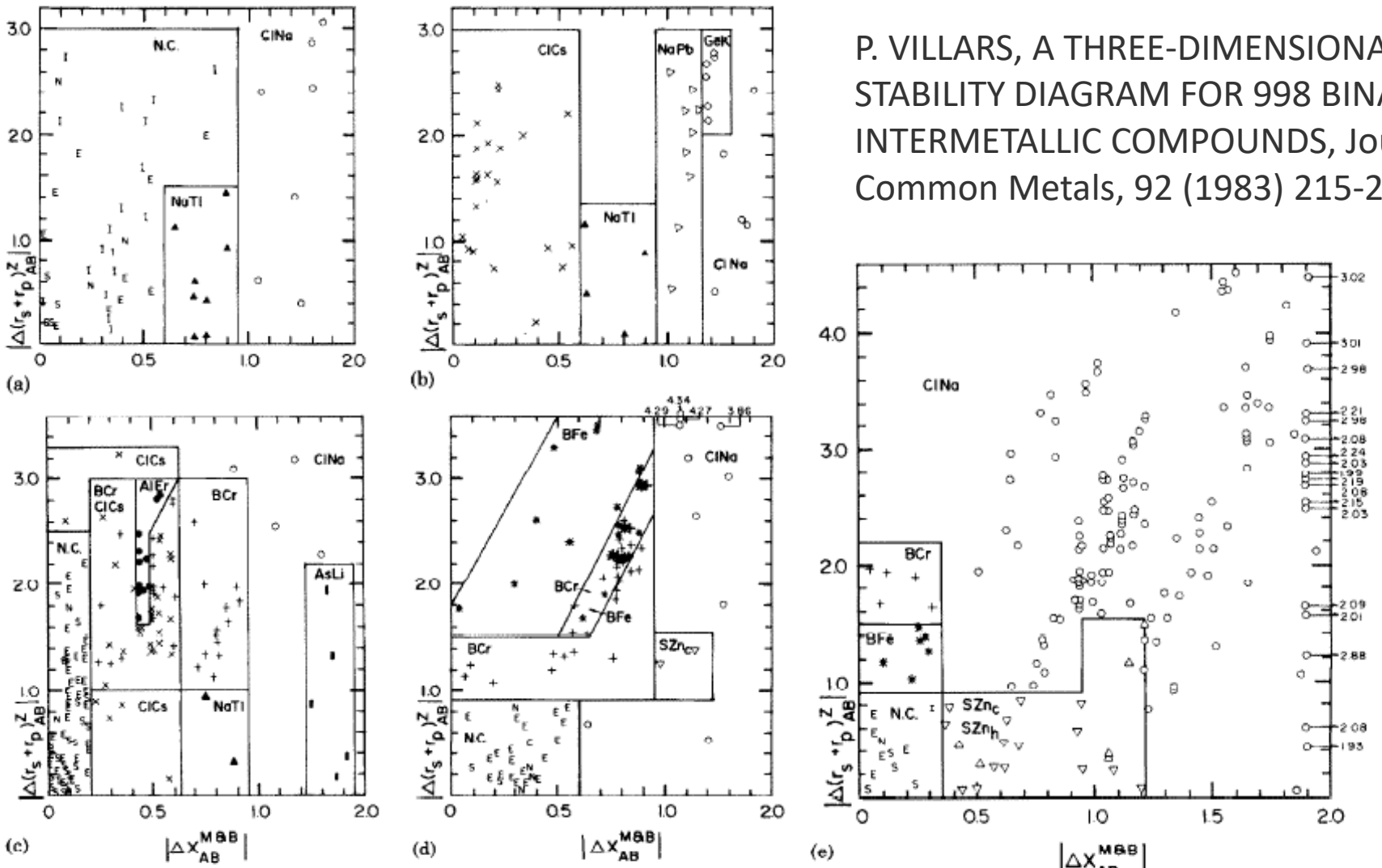
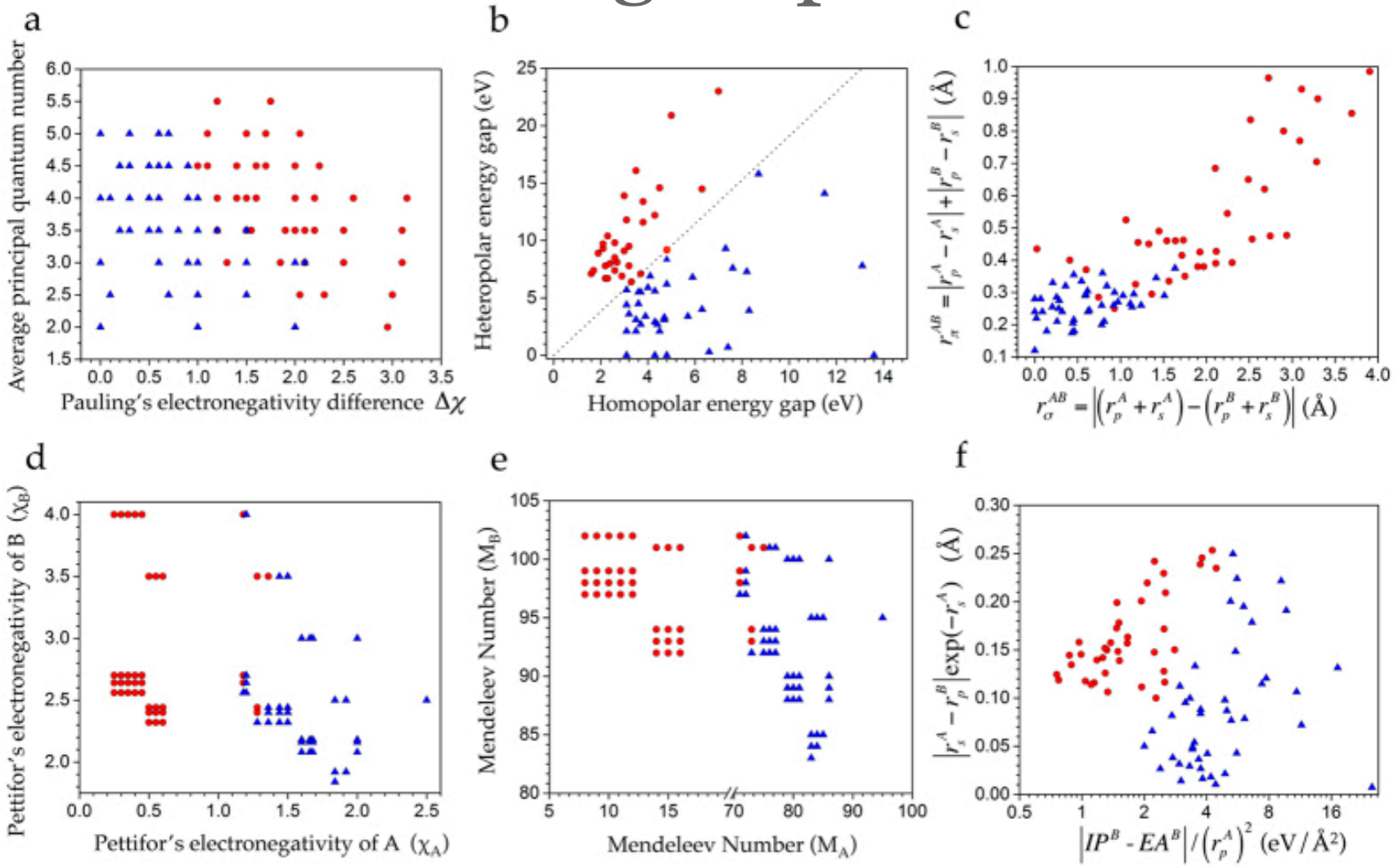


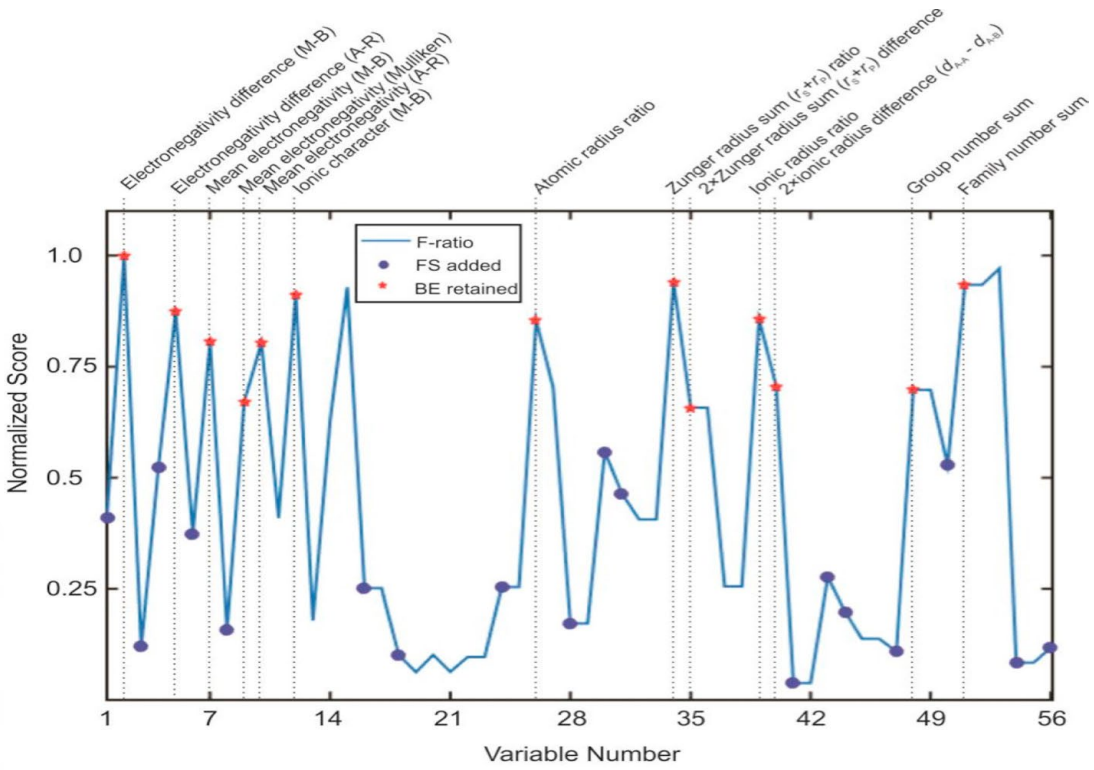
Fig. 3 *(continued).*

P. VILLARS, A THREE-DIMENSIONAL STRUCTURAL STABILITY DIAGRAM FOR 998 BINARY AB INTERMETALLIC COMPOUNDS, Journal of the Less-Common Metals, 92 (1983) 215-238 215

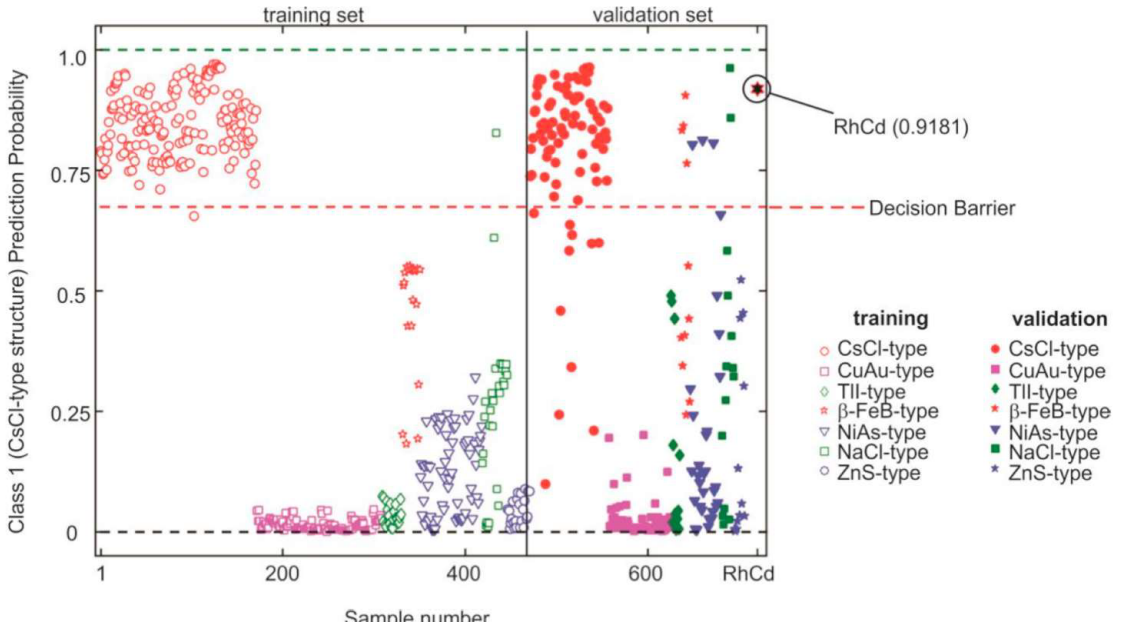
Can machine learning help?



G. Pilania, J. E. Gubernatis, and T. Lookman, Classification of octet AB-type binary compounds using dynamical charges: A materials informatics perspective, Sci Rep. 2015; 5: 17504.



1. ●Electronegativity difference (Pauling scale)	13. Ionic character (Gordy scale)	37. Ionic radius sum (d_{A-B})
2. ★Electronegativity difference (Martynov-Batsanov scale)	14. Ionic character (Mulliken scale)	38. Mean ionic radius
3. ●Electronegativity difference (Gordy scale)	15. Ionic character (Allred-Rochow scale)	39. ★Ionic radius ratio
4. ●Electronegativity difference (Mulliken scale)	16. ●Sum of valence electrons	40. ★2×ionic radius difference (d_{A-A} - d_{A-B})
5. ★Electronegativity difference (Allred-Rochow scale)	17. Mean number of electrons	41. ●Crystal radius sum (d_{A-B})
6. ●Mean electronegativity (Pauling scale)	18. ●Atomic number sum	42. Mean crystal radius
7. ★Mean electronegativity (Martynov-Batsanov scale)	19. Atomic number difference	43. ●Crystal radius ratio
8. ●Mean electronegativity (Gordy scale)	20. Mean atomic number	44. ●2×crystal radius difference (d_{A-A} - d_{A-B})
9. ★Mean electronegativity (Mulliken scale)	21. Atomic weight difference	45. Period number sum
10. ★Mean electronegativity (Allred-Rochow scale)	22. Mean atomic weight	46. Mean period number
11. Ionic character (Pauling scale)	23. Atomic weight sum	47. ●Period number difference
12. ★Ionic character (Martynov-Batsanov scale)	24. ●Atomic radius sum (d_{A-B})	48. ★Group number sum
	25. Mean atomic radius	49. Mean group number
	26. ★Atomic radius ratio	50. ●Group number difference
	27. 2×atomic radius difference (d_{A-A} - d_{A-B})	51. ★Family number sum
	28. ●Covalent radius sum (d_{A-B})	52. Mean Family number
	29. Mean covalent radius	53. Family number difference
	30. ●Covalent radius ratio	54. ●Quantum number (l) sum
	31. ●2×covalent radius difference (d_{A-A} - d_{A-B})	55. Mean quantum number (l) mean
	32. Zunger radius sum (r_S+r_P) sum	56. ●Quantum number (l) difference
	33. Mean Zunger radius sum (r_S+r_P)	
	34. ★Zunger radius sum (r_S+r_P) ratio	
	35. ★2×Zunger radius sum (r_S+r_P) difference	
	36. Zunger radius sum (r_S+r_P) difference	



A. O. Oliynyk, L.A. Adutwum, J.J. Harynuk, and A. Mar, Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis, Chem. Mater. 2016, 28, 18, 6672–6681 (2016)

Feature engineering with machine learning

For instance, the starting point $\mathbf{\Phi}_0$ may comprise readily available and relevant properties, such as atomic radii, ionization energies, valences, bond distances, and so on. The operators set is defined as

$$\hat{H}^{(m)} \equiv \{I, +, -, \times, /, \exp, \log, |-|, \sqrt{}, ^{-1}, ^{2}, ^{3}\}[\phi_1, \phi_2],$$

- Start with available physical descriptors
- Create dimensionally-consistent combinations via allowed operations
- Choose the ones that give best classification

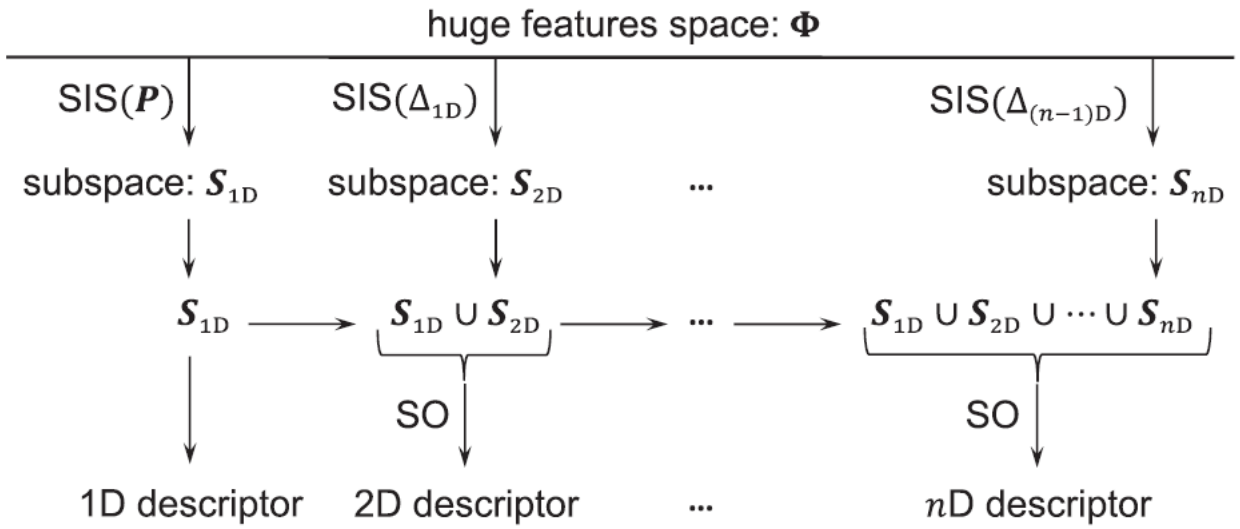


FIG. 1. The method SISSO combines unified subspaces having the largest correlation with residual errors $\mathbf{\Delta}$ (or $\boldsymbol{P}$) generated by sure independence screening (SIS) with sparsifying operator (SO) to further extract the best descriptor.

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)

Feature engineering with machine learning

TABLE I. Dependence of the metal-insulator classification descriptors on the prototypes of training binary materials.

prototypes	#materials	primary features	descriptor	classification accuracy
NaCl	132	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, EA_A, EA_B, v_A, v_B, d_{AB}	$d_1 := \frac{IE_A IE_B (d_{AB} - r_{\text{covA}})}{\exp(\chi_A)\sqrt{r_{\text{covB}}}}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si	217	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{IE_B d_{AB}^2}{\chi_A r_{\text{covA}}^2 \sqrt{CN_B}}$, $d_2 := \frac{IE_A^2 r_{\text{covB}} \log(IE_A) \lvert r_{\text{covA}} - r_{\text{covB}} \rvert}{CN_B}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{d_{AB}/r_{\text{covA}} - \chi_A/\chi_B}{\exp(CN_B/IE_B)}$, $d_2 := \frac{r_{\text{covA}}^3 d_{AB} IE_B}{\lvert \chi_B/\chi_A - \lvert CN_B - CN_A \rvert \rvert}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{V_{\text{cell}}}{\sum V_{\text{atom}}} \frac{\sqrt{\chi_B}}{\chi_A}$, $d_2 := \frac{IE_A IE_B}{\exp(V_{\text{cell}}/\sum V_{\text{atom}})}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs, Al_2O_3, La_2O_3, Th_3P_4, ReO_3, ThH_2	299	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{x_B}{\sum V_{\text{atom}}/V_{\text{cell}}} \frac{IE_B \sqrt{\chi_B}}{\chi_A}$, $d_2 := \chi_A^2 \lvert \lvert 1 - 2x_A \rvert - x_A^2 \frac{\chi_B}{\chi_A} \rvert$	99.0%[b]

[a]One entry misclassified: YP compound in NaCl prototype.

[b]Three entry misclassified: YP compound in NaCl prototype; Th_3As_4 and La_3Te_4 compounds in Th_3P_4 prototype.

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)

Feature engineering with machine learning

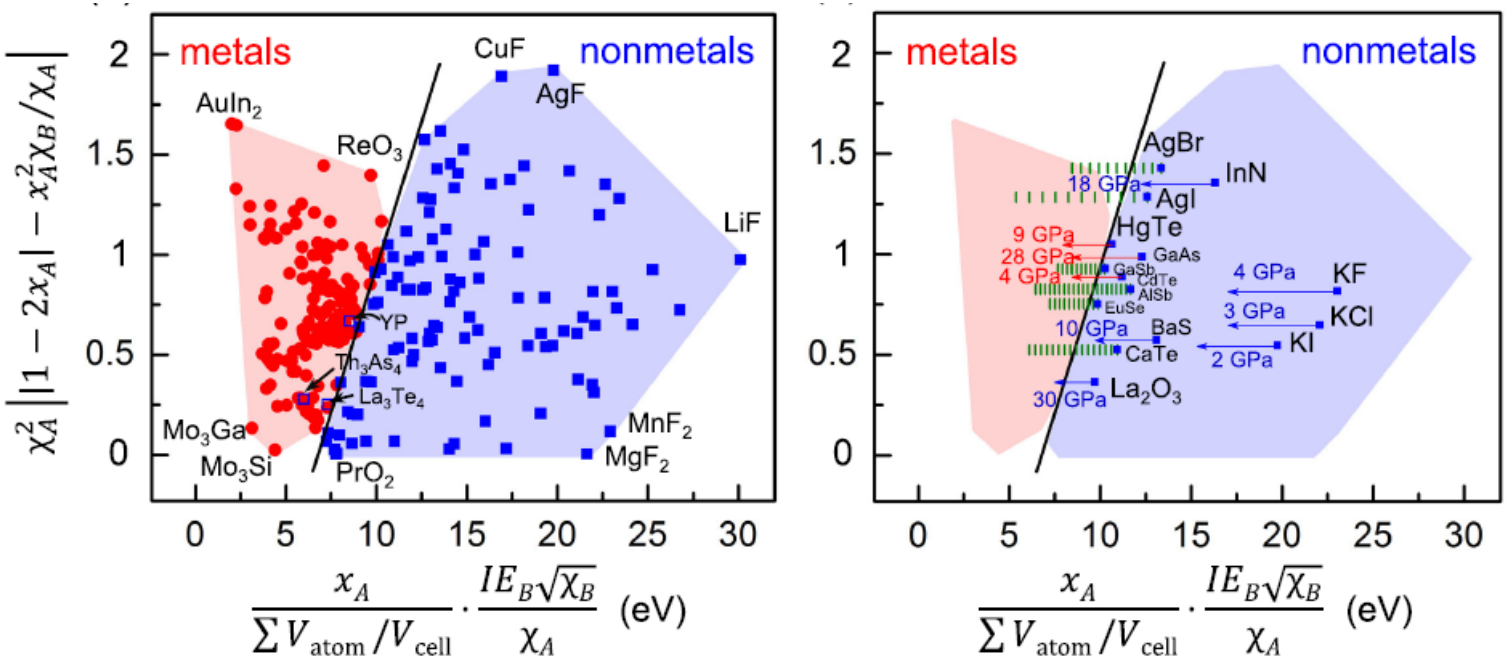


FIG. 4. SISSO for classification. (a) An almost perfect classification (99%) of metal/nonmetal for 299 materials. Symbols: χ, Pauling electronegativity; IE, ionization energy; x, atomic composition; $\sum V_{\text{atom}}/V_{\text{cell}}$, packing fraction. Red circles, blue squares, and open blue squares represent metals, nonmetals, and the three erroneously characterized nonmetals, respectively. (c) Reproduction of pressure-induced insulatormetals transitions (red arrows), of materials that remain insulators upon compression (blue arrows), and computational predictions at step of 1 GPa (green bars).

R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, PHYSICAL REVIEW MATERIALS 2, 083802 (2018)