Material Subgroups

Jilles Vreeken

9 September 2019

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Question of the day

How can we find
named conditions
under which materials
behave differently
than normal?

Data

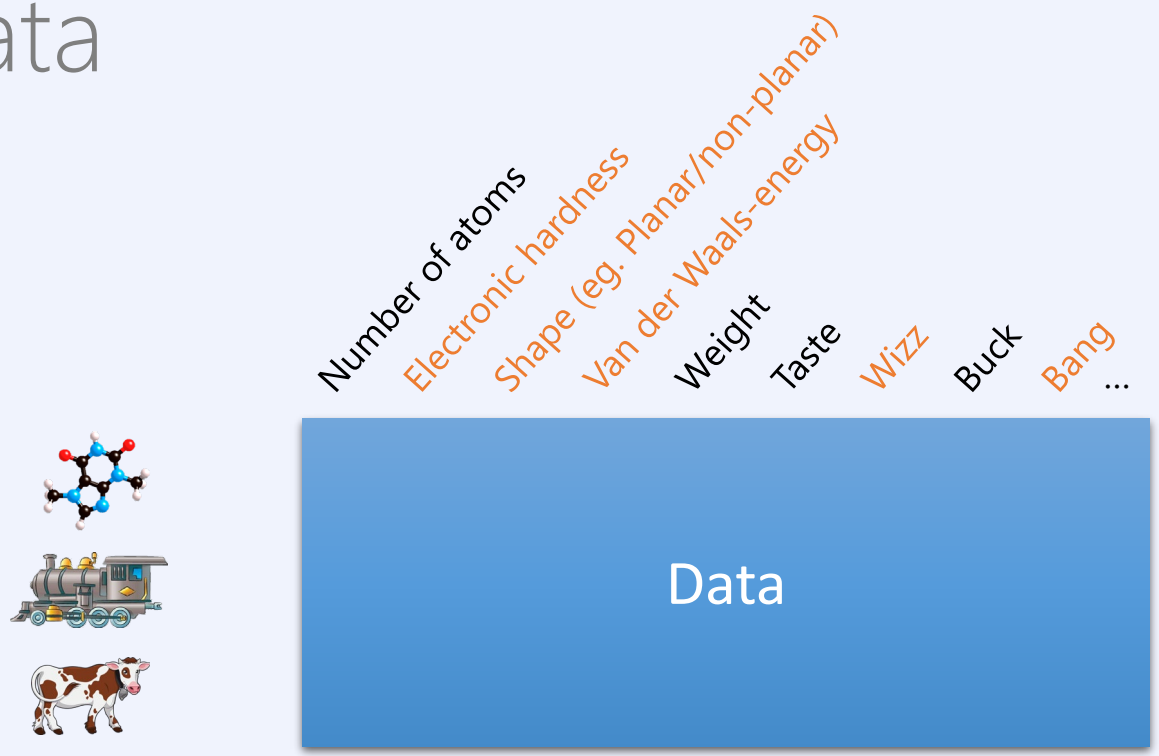
Number of atoms
Electronic hardness
Shape (eg. Planar/non-planar)
Van der Waals-energy
Weight
Taste
Wizz
Buck
Bang
...

Data

population

attributes

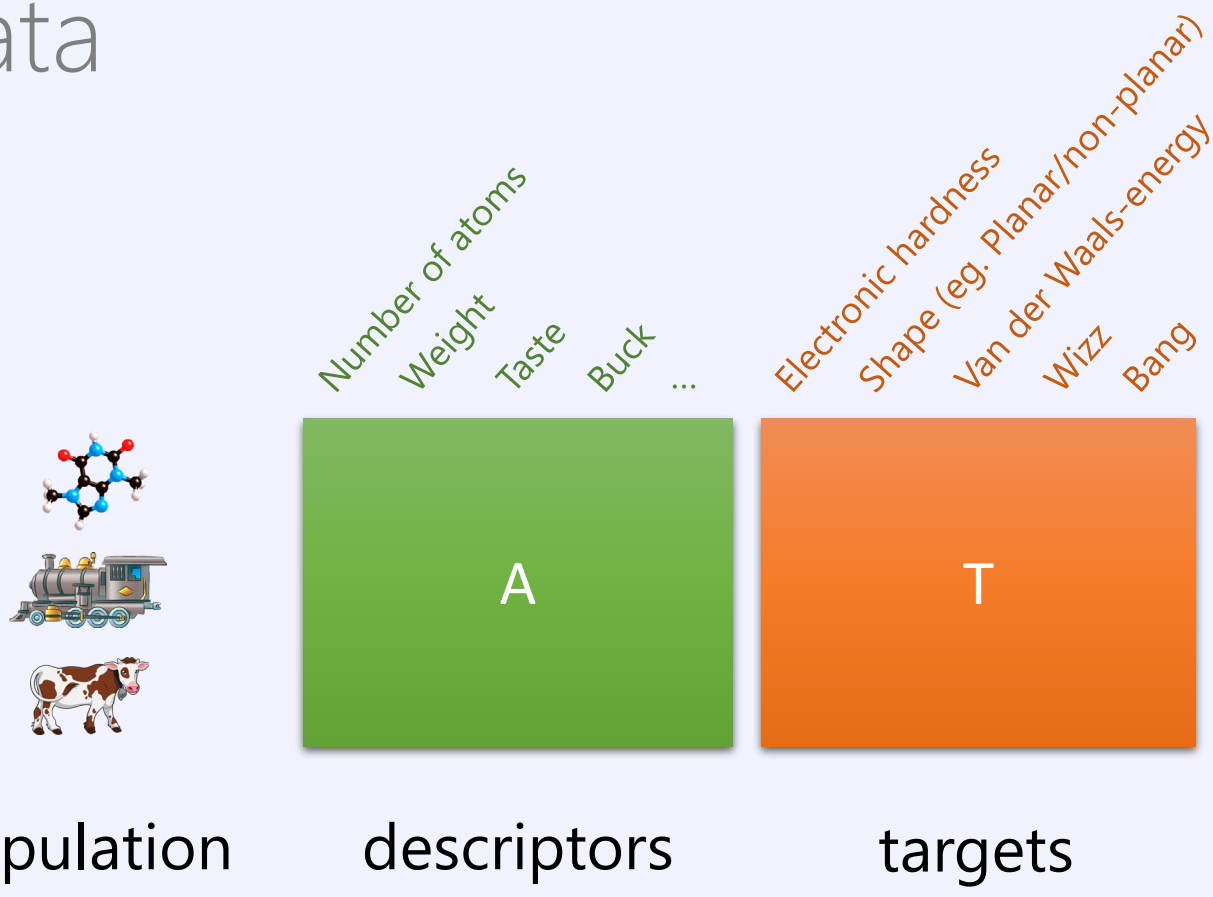
Data



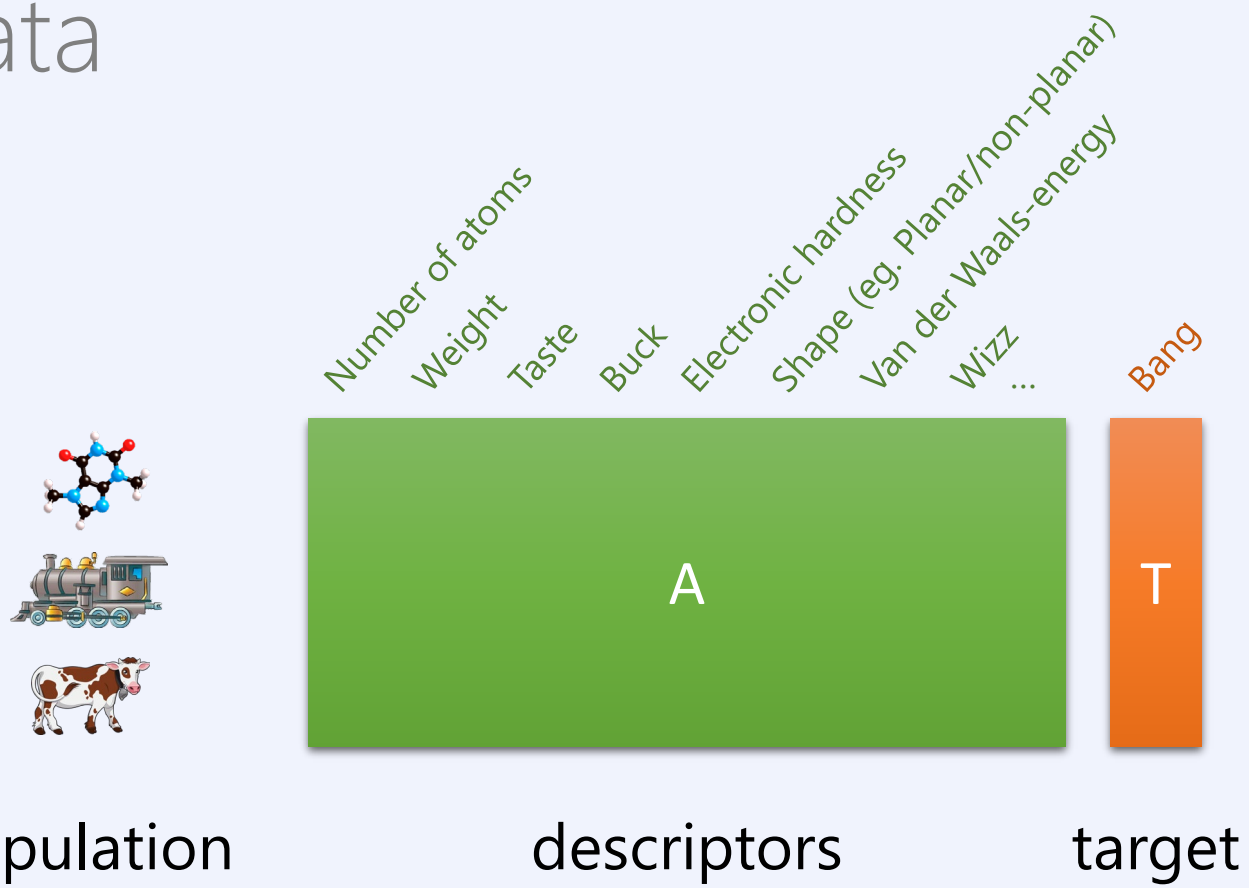
population

attributes

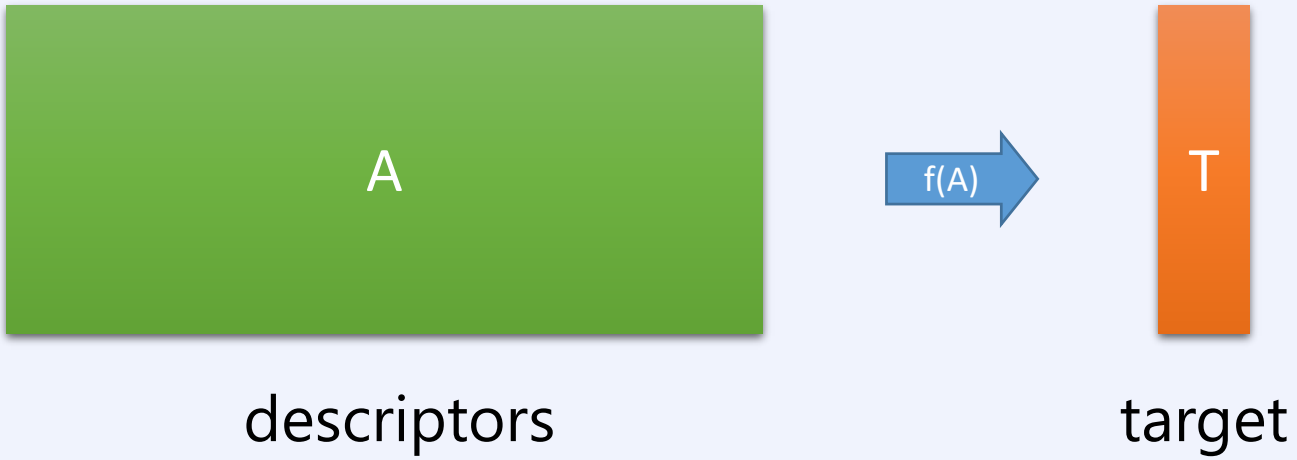
Data


Number of atoms
Weight
Taste
Buck
...
Electronic hardness
Shape (eg. Planar/non-planar)
Van der Waals-energy
Wizz
Bang
A
T
population
descriptors
targets

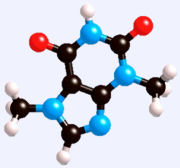
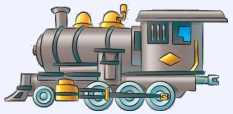
Data

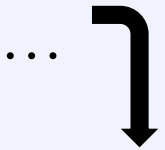

Number of atoms
Weight
Taste
Buck
Electronic hardness
Shape (eg. Planar/non-planar)
Van der Waals-energy
Wizz
...
Bang
A
T
population
descriptors
target

Global modelling



The setup

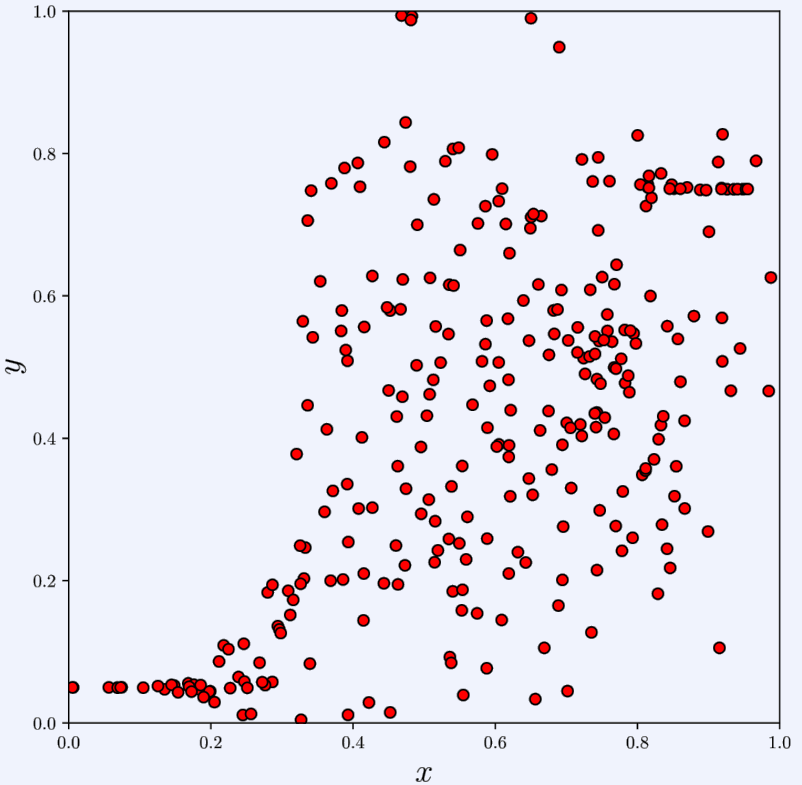


X

N	$\frac{E}{-E_0}$	T	#1	#2	#3	#4	#5	#6	$\frac{rg}{rg_0}$	#/N
10	..	..	0	0	6	2	0	2	..	1
9	..	..	0	4	0	1	4	0	..	9
..	..	..	..	..	..	..	..	..	..	..

Y

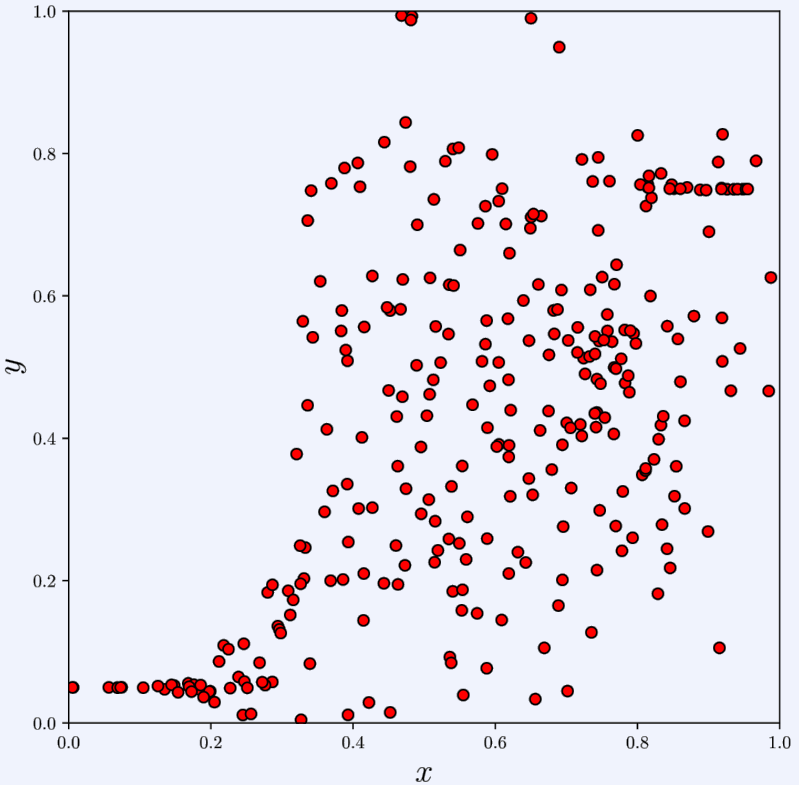
Bang
..
..
..



The setup

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

What can we tell about y?
(in terms of x)



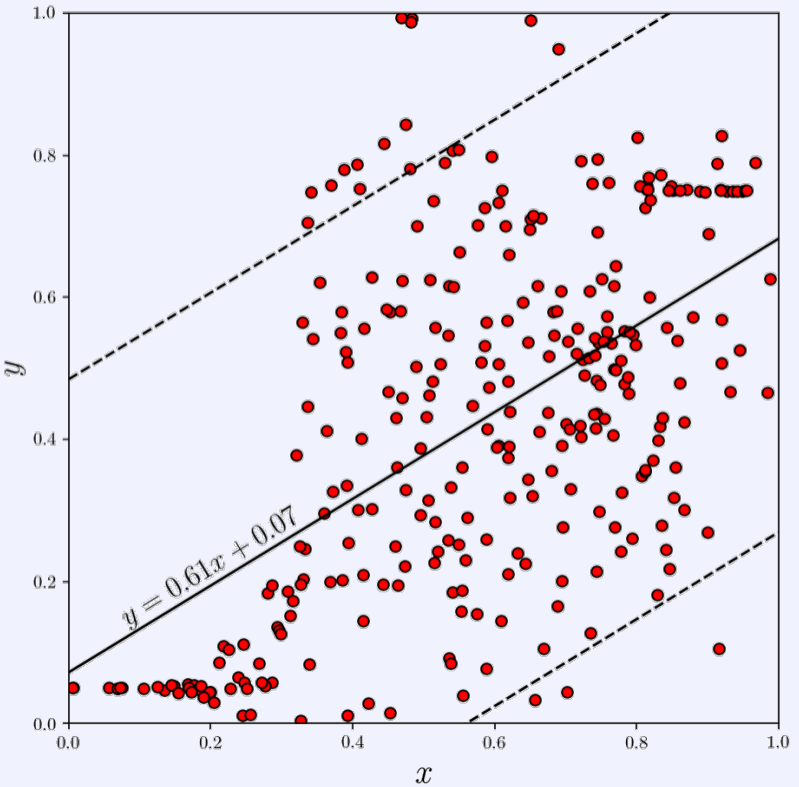
Simple model explicable but inaccurate

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

Find
Coefficients $\alpha, \beta \in \mathbb{R}$ such that objective

$$f(\alpha, \beta) = \frac{1}{n} \sum_{i \in S} (\alpha x(i) + \beta - y(i))^2 + \lambda \|\alpha, \beta\|_1$$

is minimal



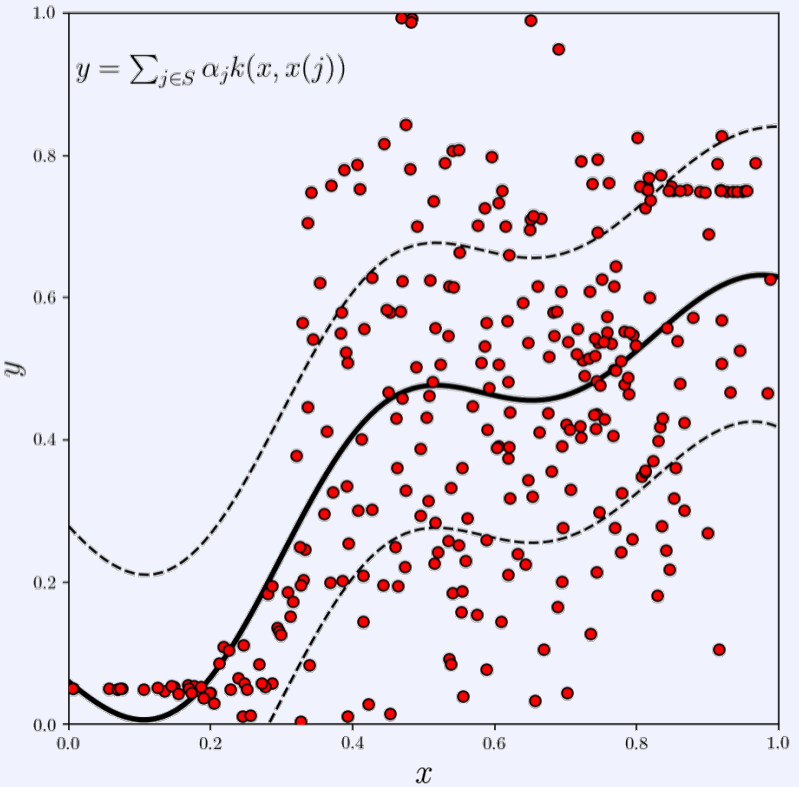
Standard global model fitting

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

Find
Coefficients $\alpha \in \mathbb{R}$ such that objective

$$f(\alpha) = \sum_{i \in S} \left(\alpha_i k\big(x, x(i)\big) - y(i) \right)^2$$

is minimal



… results in model of some accuracy

solution $\hat{f}(x) = \sum_{i=1}^{101} \alpha_i k(x_i, x)$ achieves $\mathrm{rmse}(\hat{f}, f) = 0.481$

i	α	N	$E - E_0$	T	#1	#2	#3	#4	#5	#6	rg/rg_0	#/N
1	0.46383	12	1.129305	465.24	0	0.00506593	0.49725	0.247684	0.00144502	0.24856	1.00242	3.99120
2	0.18535	9	0.505654	200	4.00E-08	0.223098	0.332524	0.222582	0.112373	0.109428	0.99251	…
3	1.39603	8	1.126513	600	0.0461463	0.328876	0.125	0.00844487	0.452526	…	…	…
4	1.04601	12	0.780292	266	0	0.00176119	0.501645	…	…	…	…	…
5	1.58057	8	0.549743	150	1.49E-09	…	…	…	…	…	…	…
6	0.26604	9	1.809864	…	…	…	…	…	…	…	…	…
7	0.32886	11	…	…	…	…	…	…	…	…	…	…
8	0.37825	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…

Global modelling

A single model over the **whole** population

- minimizes the overall prediction error
- therefore, disregards **local details**

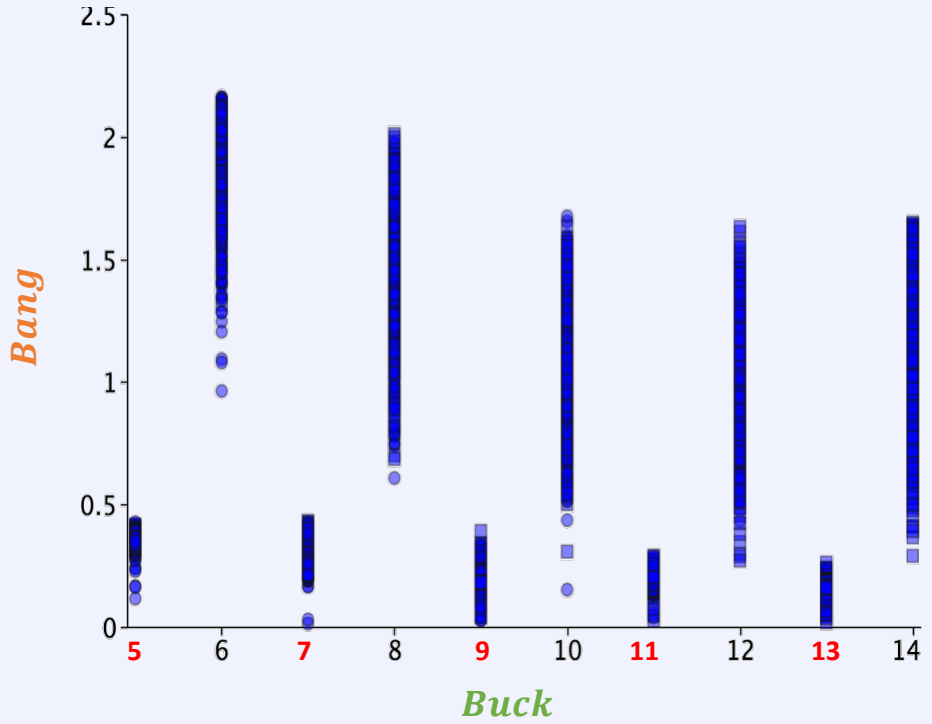
Well-known examples include

- support vector machines
- deep learning
- decision trees
- Lasso

Global models are **great** if **optimal prediction** is the goal

- however, for **insight** different tools are needed

Real **insight** looks different



"Samples with **odd *Buck*** tend to have a **small *Bang***."

When does it happen?

Most methods consider the population as **homogenous**

- only one score per projection
- 'Buck is somewhat predictive of Bang'

There may be **subpopulations** that stand out

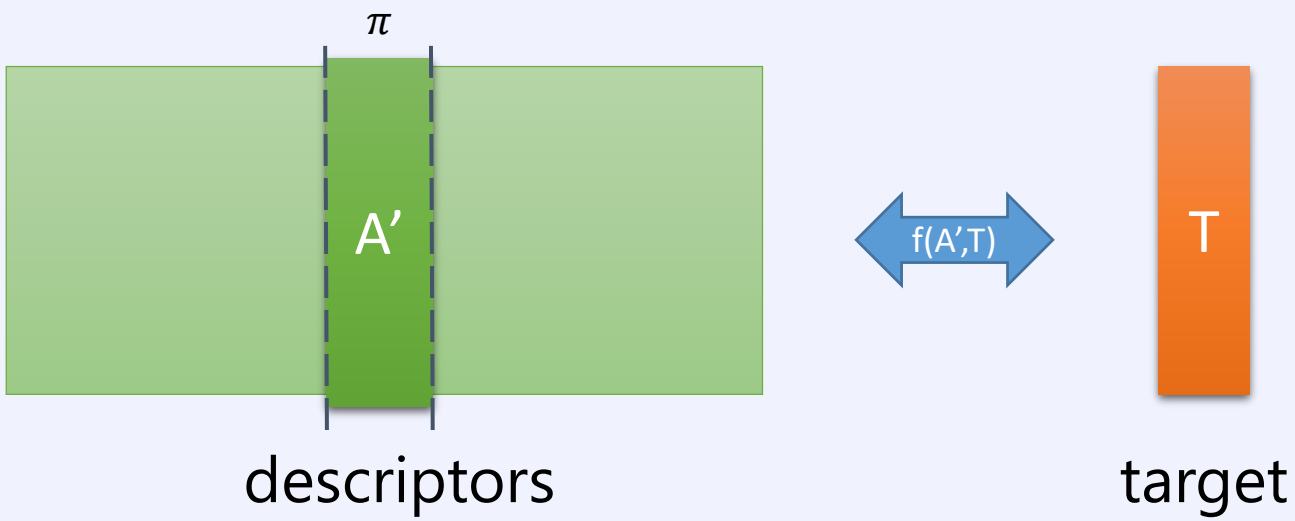
- characterising these subpopulations provides **insight**
- '**odd values of** Buck are highly predictive of Bang'

These characterisations we refer to as **patterns**

- require an **interpretable vocabulary** to be meaningful

Projections

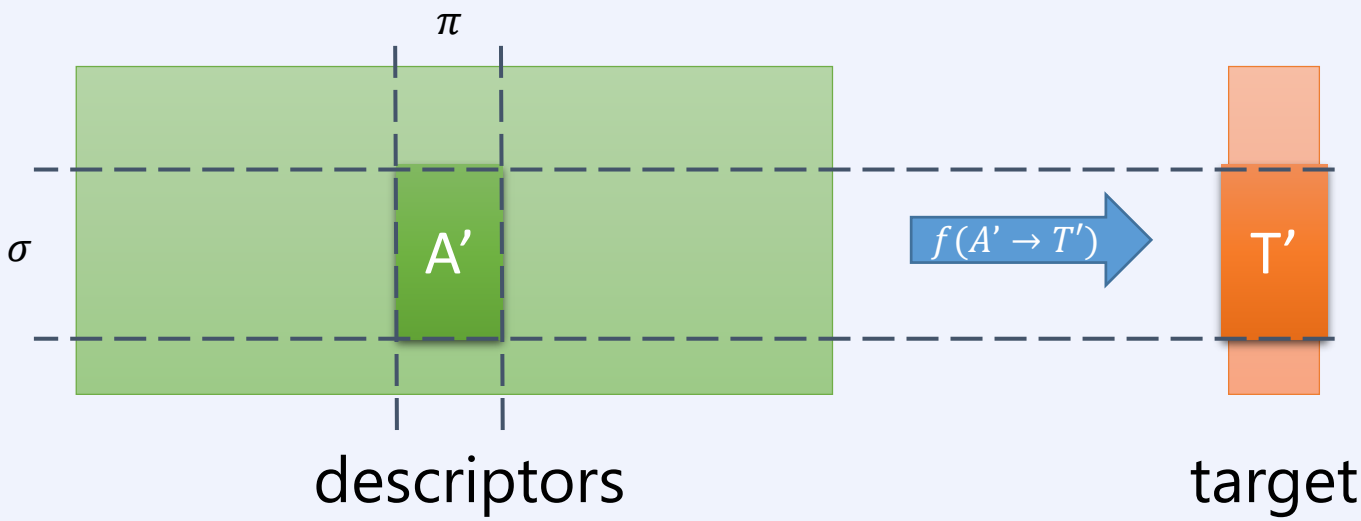
Find those projections $A' = \pi(A)$ with good $f(A', T)$



E.g. find those A' that correlate/interact strongly with T

Selection

Find selectors σ such that $A' = \sigma(\pi(A))$ has good $f(A' \rightarrow T')$



E.g. find those σ and π on A for which T' stands out

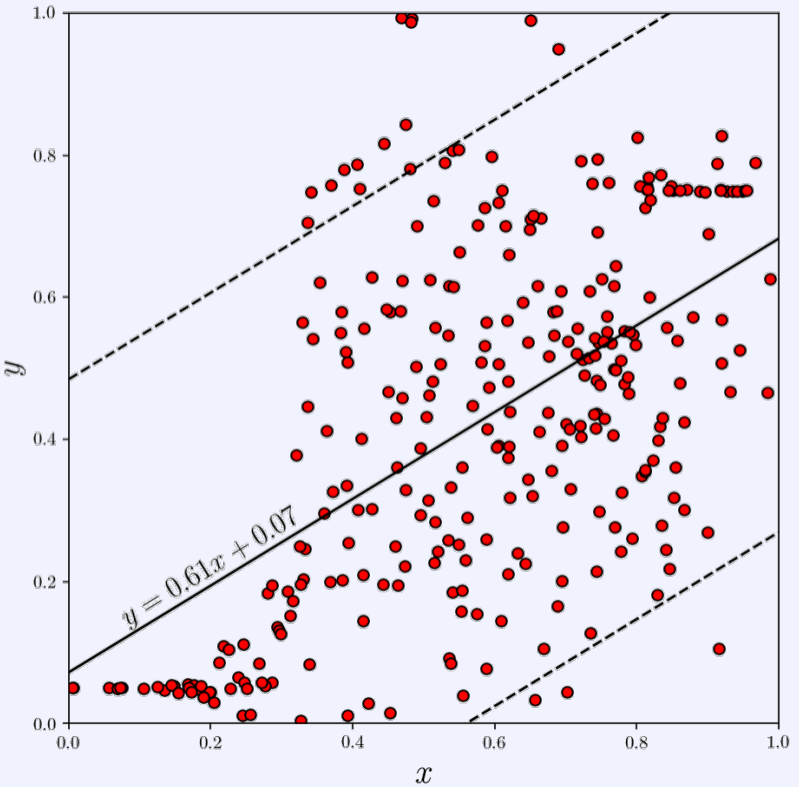
Simple model explicable but inaccurate

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

Find
Coefficients $\alpha, \beta \in \mathbb{R}$ such that objective

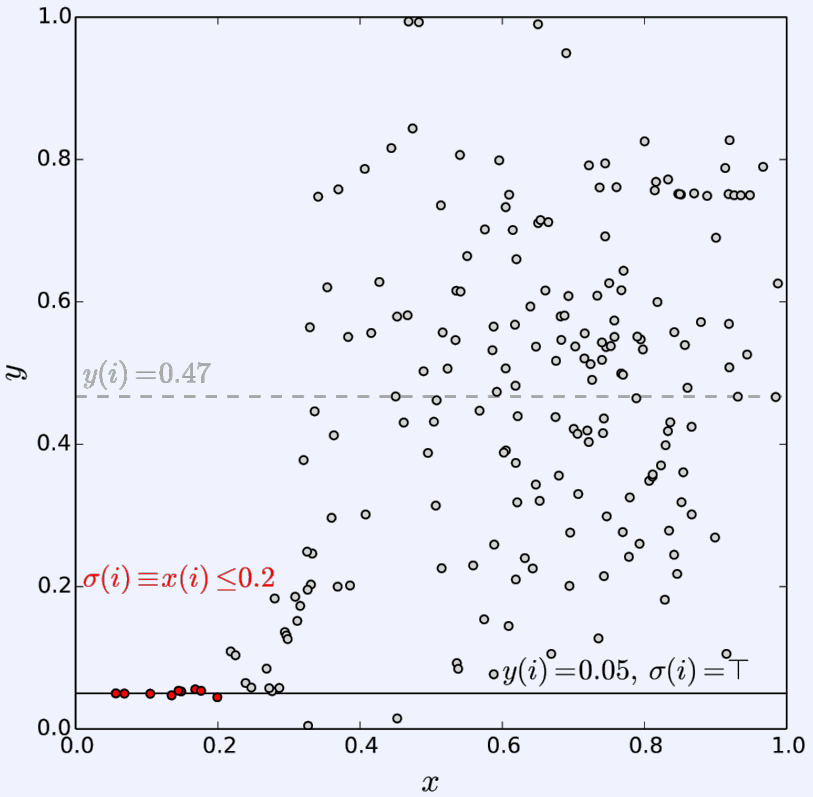
$$f(\alpha, \beta) = \frac{1}{n} \sum_{i \in S} (\alpha x(i) + \beta - y(i))^2 + \lambda \|\alpha, \beta\|_1$$

is minimal



Local model can be simple and accurate

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$



Example description language

Basic propositions

$$\Pi_x = \bigcup_{j\in[m]} \{\pi_1^{(j)}, \ldots \pi_{k_j}^{(j)}\}$$

where for categorical $x_j : P \to \{v_1, \ldots, v_s\}$

$$\pi_l^{(j)}(i) \equiv x(i) = v_l$$

e.g., 'parity=odd'

for ordinal $x_j : P \to \{v_1, \ldots, v_s\}$ with $v_1 \leq \cdots \leq v_s$

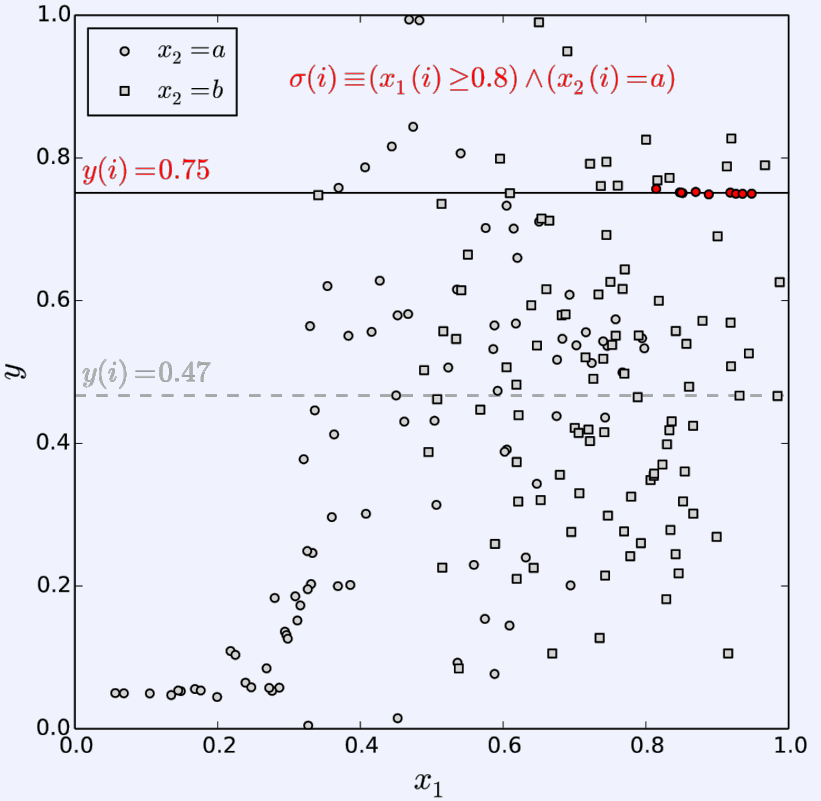
$$\pi_{2l-1}^{(j)}(i) \equiv x(i) \leq v_l$$

$$\pi_{2l}^{(j)}(i) \equiv x(i) > v_l$$

e.g., 'ionization potential ≤ 2'

Conjunctive selectors

$\sigma \in \mathcal{L}_x$ of form $\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)$



Basic subgroup discovery

Given
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

Define
Propositions $\Pi_x = \{\pi_1, \ldots, \pi_k\}$
Selection language $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \cdots \wedge \pi_{j_l}(i)\}$

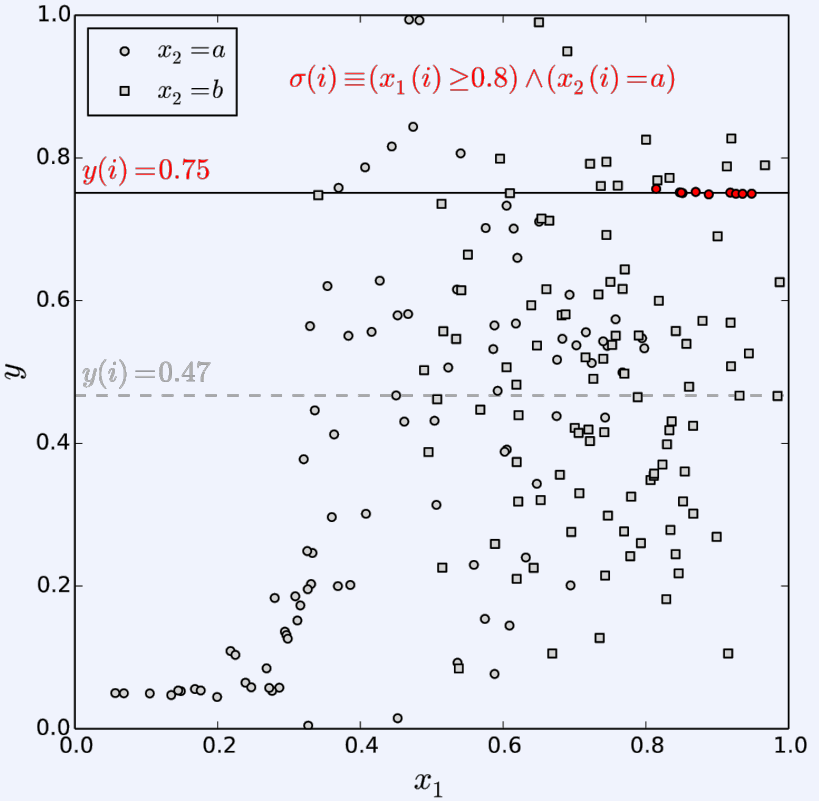
Optimize

$$f(Q) = \mathrm{cvr}(Q)^\gamma \mathrm{eff}(Q)_+$$

with

- $\mathrm{Q} = \mathrm{ext}(\sigma) = \{i \in S: \sigma(i) = \top\}$ — extension
- $\mathrm{cvr}(Q) = |Q|/|S|$ — coverage
- $\mathrm{eff}(Q) = \tilde{y}(Q) - \tilde{y}(S)$ — effect

For continuous-valued data, $\tilde{y}(Q)$ is a typically a measure of central tendency (mean, median,…)



Subgroup discovery framework

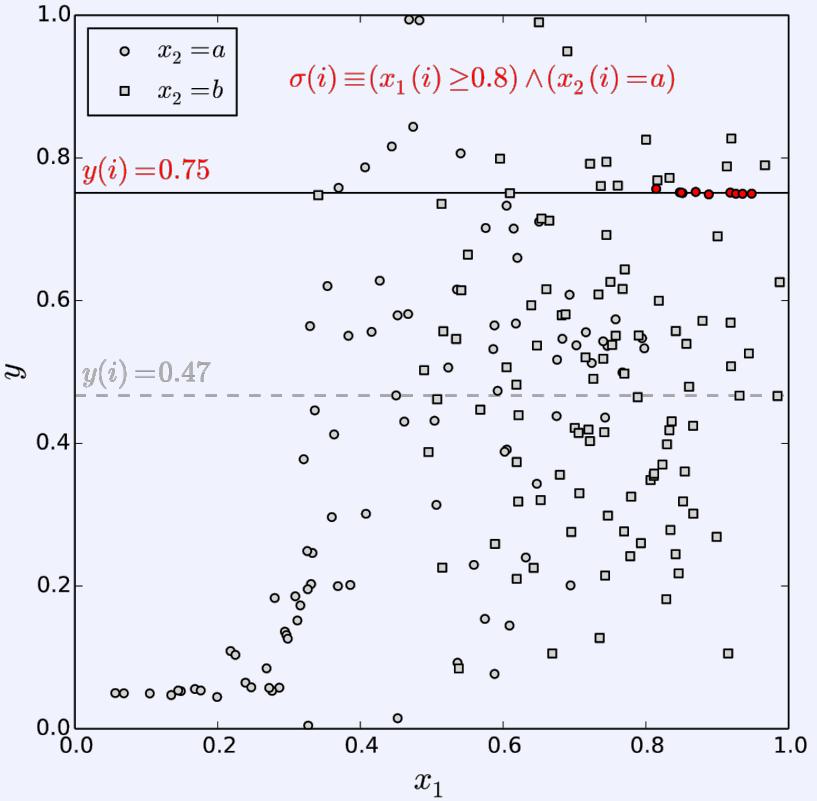
Given
Sample $S = \{1, \ldots, n\} \subseteq P$
Target variable $y: P \to Y$
Description variables $x_j: P \to X_j$, $j \in \{1, \ldots, m\}$

Define
Description language $\mathcal{L}_x \subseteq \{\bot, \top\}^P$
(for $\sigma \in \mathcal{L}_x$ write $\text{ext}(\sigma) = \{i \in P: \sigma(i) = \top\}$)
Objective function $f_y: 2^P \to \mathbb{R}$

Find
Goal: the top-k subgroups with highest $f(Q_i)$

Or, more formally, solution set $\mathcal{S} \subseteq \mathcal{L}_x$ with $|\mathcal{S}| = k$ such that for $\sigma \in \mathcal{S}$ and $\varphi \in \mathcal{L}_x \setminus \mathcal{S}$, $f_y(\text{ext}(\sigma)) \geq f_y(\text{ext}(\varphi))$

"top-k constraint-free formulation"



Application 1: structure-property analysis

Population

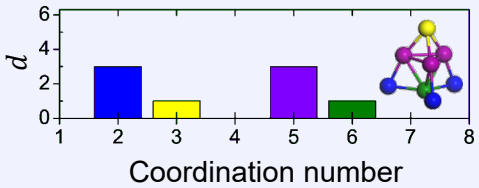
$$P = \{c : c \text{ conf. of } \mathrm{Au5} - \mathrm{Au14}\}$$

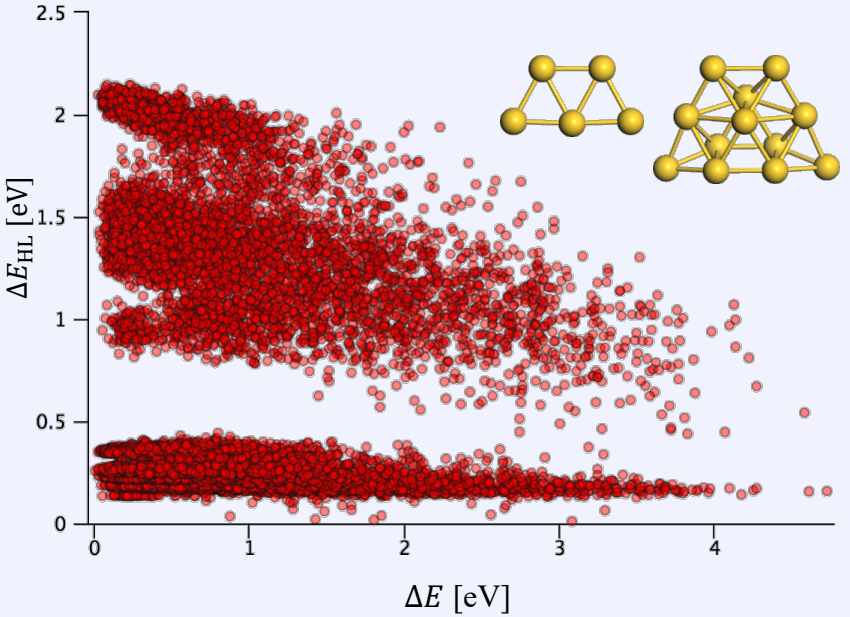
Target

$y = \Delta E_{\mathrm{HL}}$ HOMO-LUMO energy gap

Features

$x \in \{n, d_1, d_2, d_3, d_4, d_5, d_6, r, \mathrm{shape}, \mathrm{Mo_{co}}, \mathrm{Me_{co}}\}$





Application 1: structure-property analysis

Population

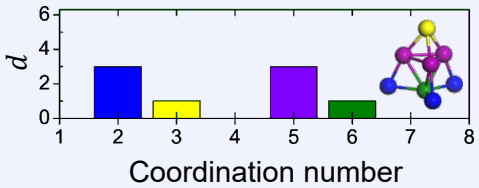
$$P = \{c : c \text{ conf. of } \mathrm{Au5} - \mathrm{Au14}\}$$

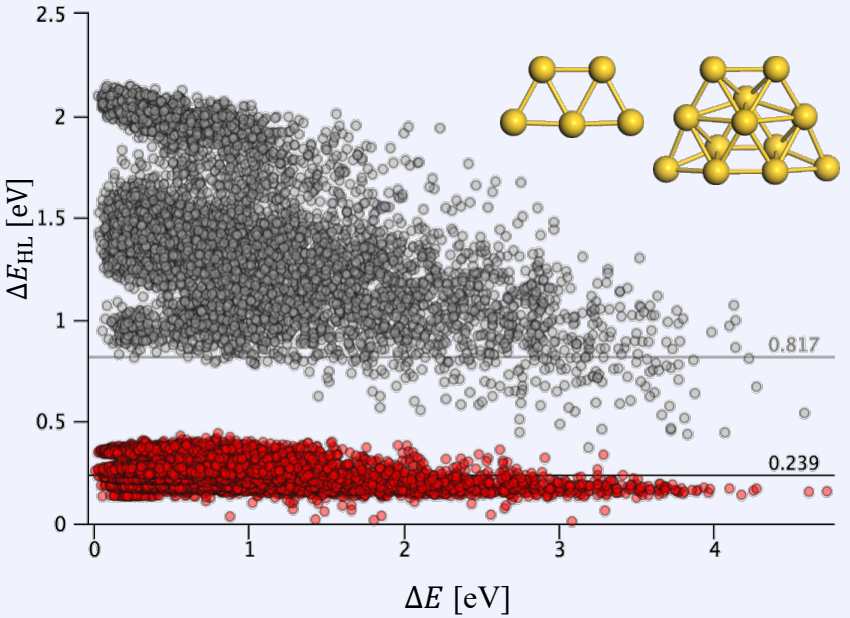
Target

$y = \Delta E_{\mathrm{HL}}$ HOMO-LUMO energy gap

Features

$$x \in \{n, d_1, d_2, d_3, d_4, d_5, d_6, r, \mathrm{shape}, \mathrm{Mo_{co}}, \mathrm{Me_{co}}\}$$



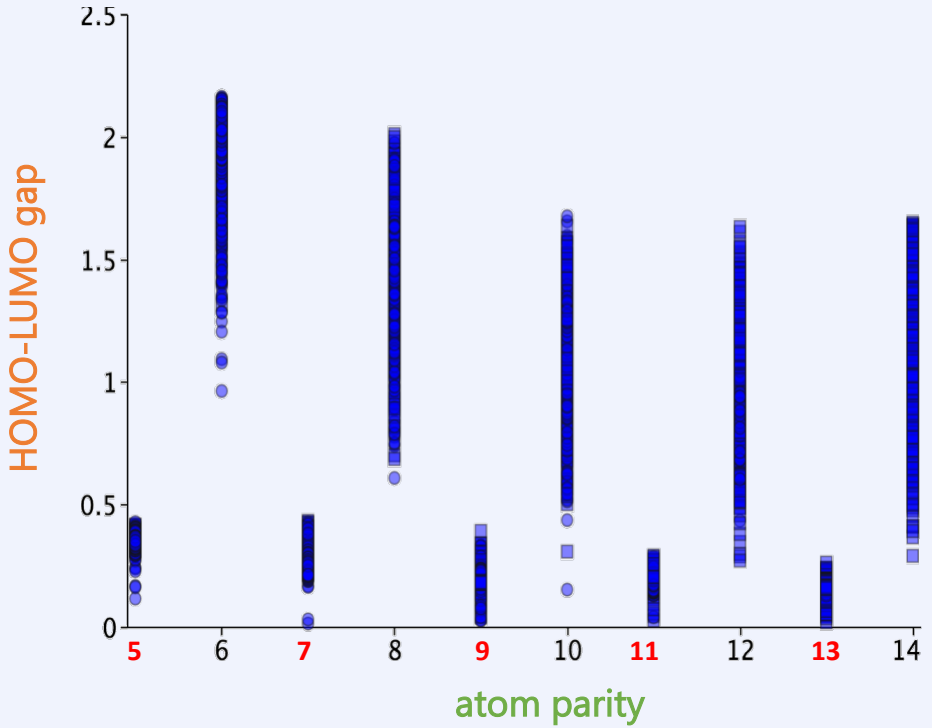


Selector $\sigma(i) \equiv \mathrm{odd}(n(i))$

Parameters $\mathrm{cvr}(\sigma) = 0.5$ $\quad \mathrm{eff}(\sigma) = 0.71$ $\quad [\bar{y}(Q) = 0.24,\ \bar{y}(S) = 0.82]$

(Goldsmith et al, 2017)

Real **insight** looks different



"Gold nano-clusters with **odd *atom parity*** tend to have a **small *HOMO-LUMO gap***."

Application 2: semi-conducting crystals

Population

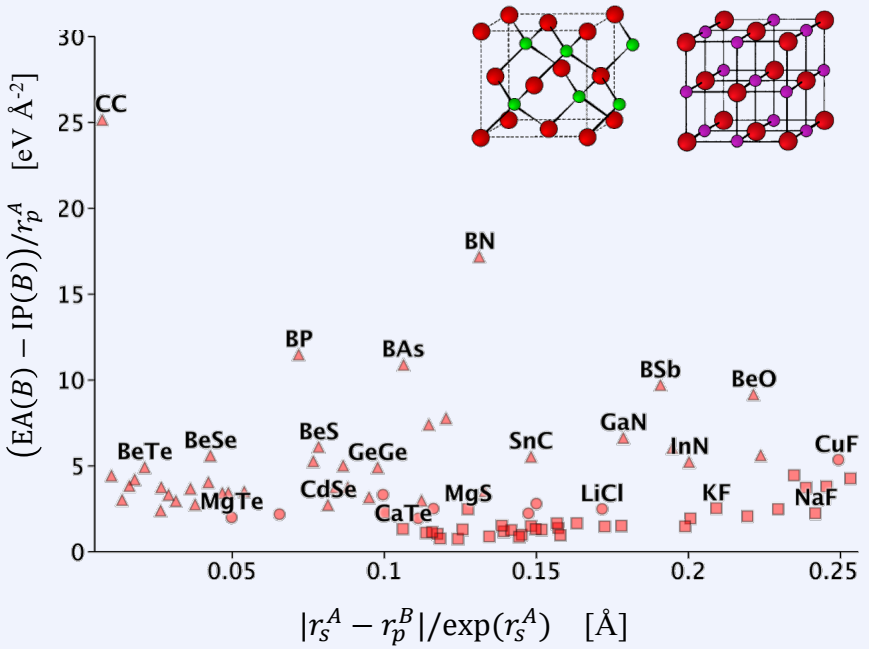
$$P = \{AB : A = \text{Ag, Al, Ba, } \ldots \wedge B = \text{Br, Cl, F, } \ldots\}$$

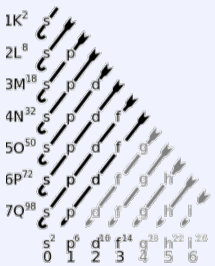
Target

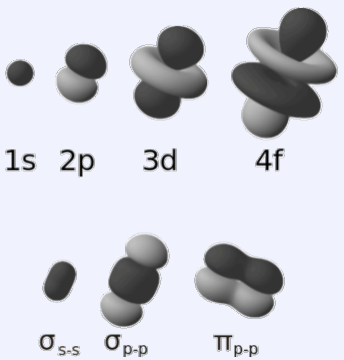
$$y = \text{sign}(\Delta E) \text{ where } \Delta_E = E_{\text{RS}} - E_{\text{ZB}}$$

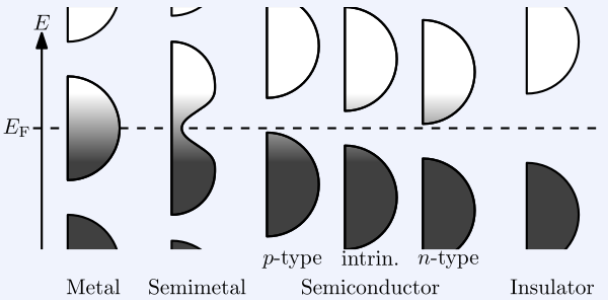
Features

$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B,$$
$$\text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \ldots\}$$









Application 2: semi-conducting crystals

Population

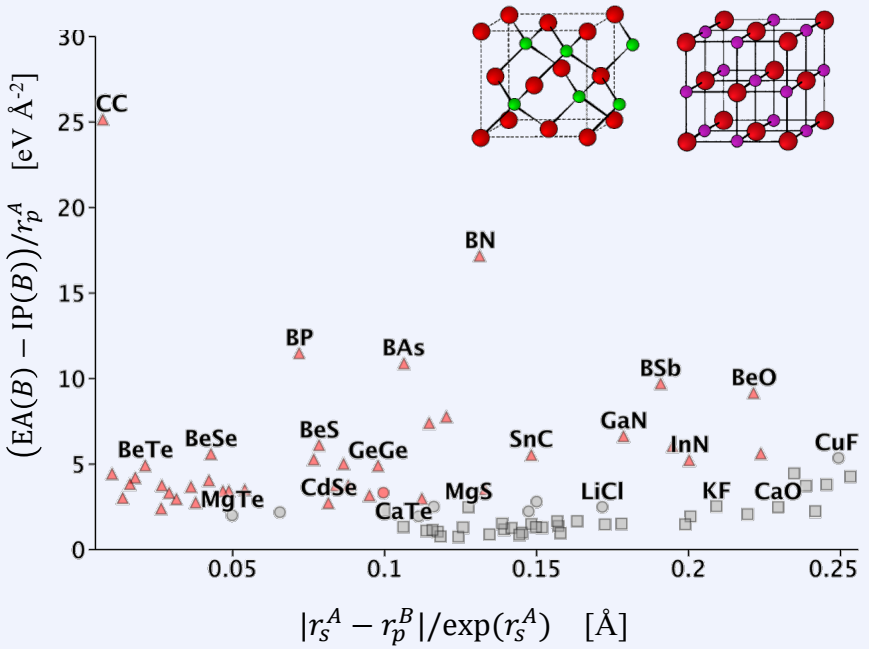
$$P = \{AB : A = \text{Ag, Al, Ba, } \ldots \wedge B = \text{Br, Cl, F, } \ldots\}$$

Target

$$y = \text{sign}(\Delta E) \text{ where } \Delta_E = E_{\text{RS}} - E_{\text{ZB}}$$

Features

$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B,$$
$$\text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \ldots\}$$



Selector $\sigma_{\text{ZB}} \equiv \left(\left|r_p^A - r_p^B\right| \leq 1.15\right) \wedge \left(r_s^A \leq 1.27\right)$

Parameters $\text{cov} = 40/82$ $\text{eff} = 1$ $[H_y(\sigma_{\text{ZB}}) = 0, H_y(P) = 1]$

Application 2: semi-conducting crystals

Population

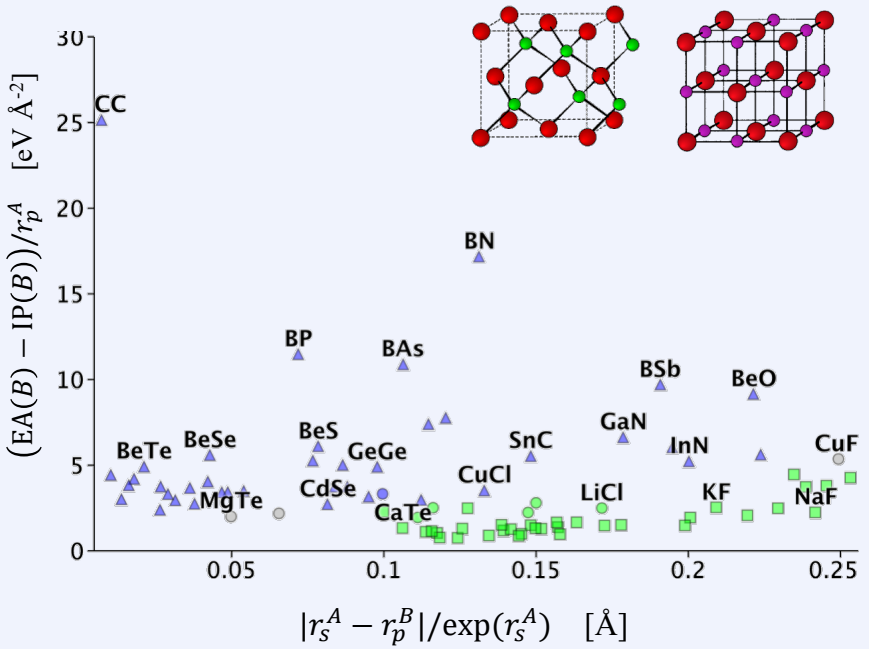
$$P = \{AB : A = \text{Ag, Al, Ba, } \ldots \wedge B = \text{Br, Cl, F, } \ldots\}$$

Target

$$y = \text{sign}(\Delta E) \text{ where } \Delta_E = E_{\text{RS}} - E_{\text{ZB}}$$

Features

$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B,$$
$$\text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \ldots\}$$



Selector

$$\sigma_{\text{ZB}} \equiv \left(\left|r_p^A - r_p^B\right| \leq 1.15\right) \wedge \left(r_s^A \leq 1.27\right)$$
$$\sigma_{\text{RS}} \equiv \left(\left|r_p^A - r_p^B\right| \geq 0.91\right) \wedge \left(r_s^A \geq 1.22\right)$$

Parameters

$\text{cov} = 40/82$ $\quad \text{eff} = 1$ $\quad [H_y(\sigma_{\text{ZB}}) = 0, H_y(P) = 1]$

$\text{cov} = 39/82$ $\quad \text{eff} = 1$ $\quad [H_y(\sigma_{\text{RS}}) = 0, H_y(P) = 1]$

Application 2: semi-conducting crystals

Population

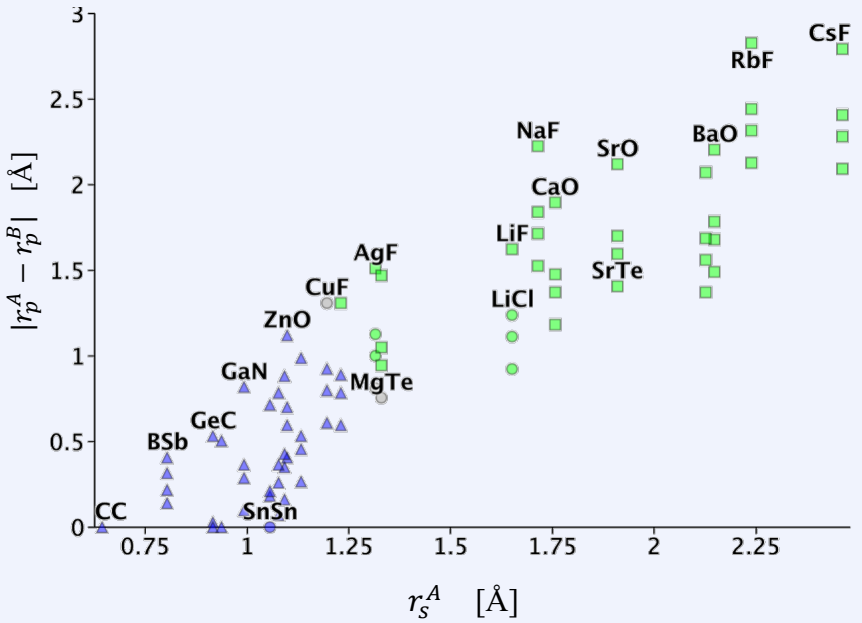
$$P = \{AB : A = \text{Ag, Al, Ba, } \ldots \wedge B = \text{Br, Cl, F, } \ldots\}$$

Target

$$y = \text{sign}(\Delta E) \text{ where } \Delta_E = E_{\text{RS}} - E_{\text{ZB}}$$

Features

$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B, \\ \text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \ldots\}$$



Selector

$$\sigma_{\text{ZB}} \equiv \left(\left|r_p^A - r_p^B\right| \leq 1.15\right) \wedge \left(r_s^A \leq 1.27\right)$$

$$\sigma_{\text{RS}} \equiv \left(\left|r_p^A - r_p^B\right| \geq 0.91\right) \wedge \left(r_s^A \geq 1.22\right)$$

Together, these two subgroups cover almost all instances, and hence form an 'almost global model', with 100% accuracy!

Parameters

$\text{cov} = 40/82$ $\quad \text{eff} = 1$ $\quad [H_y(\sigma_{\text{ZB}}) = 0, H_y(P) = 1]$

$\text{cov} = 39/82$ $\quad \text{eff} = 1$ $\quad [H_y(\sigma_{\text{RS}}) = 0, H_y(P) = 1]$

Branch & Bound Optimization

Branch

$$\boldsymbol{r}: \mathcal{L}_x \to 2^{\mathcal{L}_x}$$

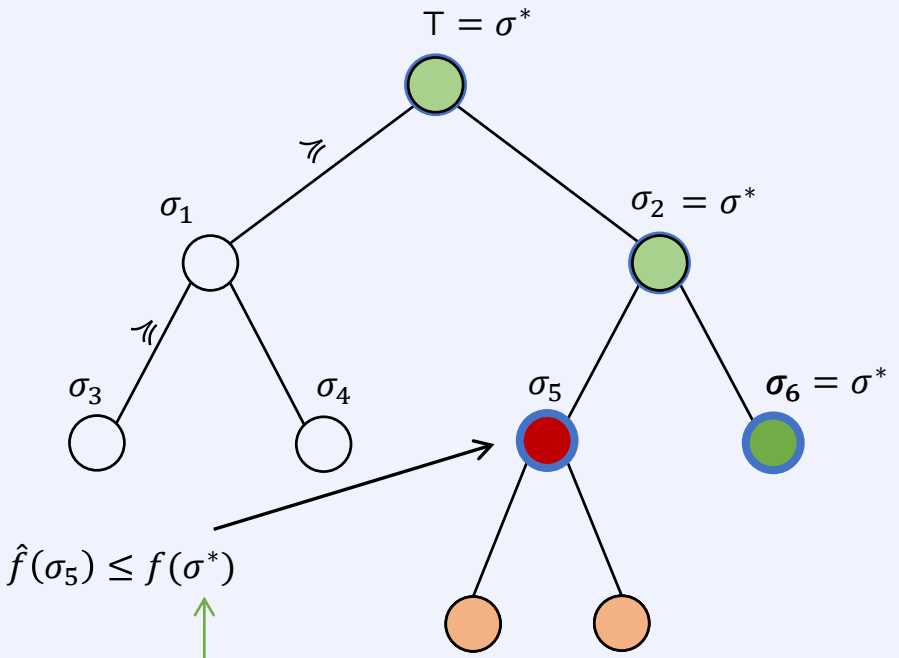
$$\varphi \in \boldsymbol{r}(\sigma) \Rightarrow \sigma \preccurlyeq \varphi$$
$$\Rightarrow \text{ext}(\sigma) \supseteq \text{ext}(\varphi)$$

Bound

$$\hat{f}(\sigma) = \max\{\boldsymbol{f}(\boldsymbol{R}): \boldsymbol{R} \subseteq \text{ext}(\boldsymbol{\sigma})\}$$
$$\geq \max\{\boldsymbol{f}(\boldsymbol{\varphi}): \boldsymbol{\varphi} \succcurlyeq \boldsymbol{\sigma}\}$$

optimistic estimator

tight optimistic estimator



optimistic estimate for σ_5 is worse than score of best subgroup σ^* found so far: we can safely prune the whole branch

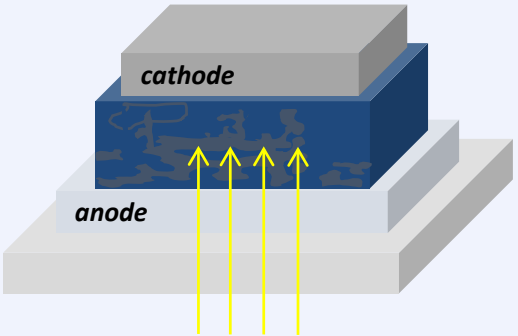
(Wrobel 1997; Grosskreutz et al. 2008; Boley and Grosskreutz 2009)

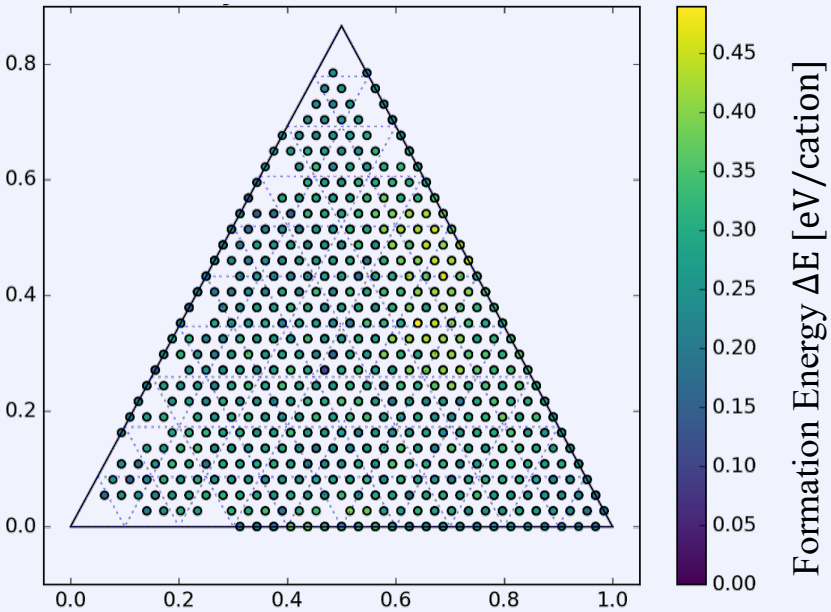
Application 3: model diagnostics

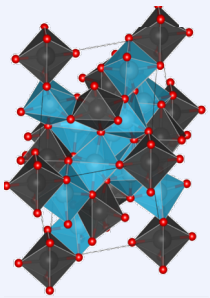
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \ \ \mathrm{a}_y\mathrm{Al}_y\right)_2\mathrm{O}_3\right\}$$

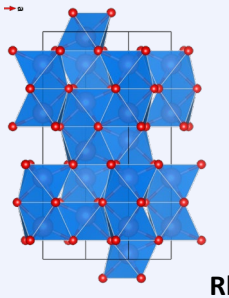
Photovoltaics



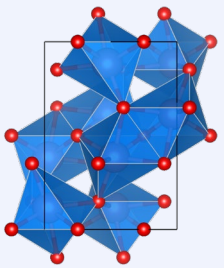




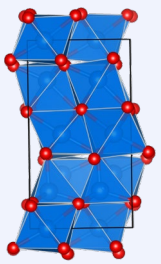
c-phase (Ia3)



α-phase (R3c)



Rh_2O_3-II-phase (Pbcn)



e-phase (Pna21)

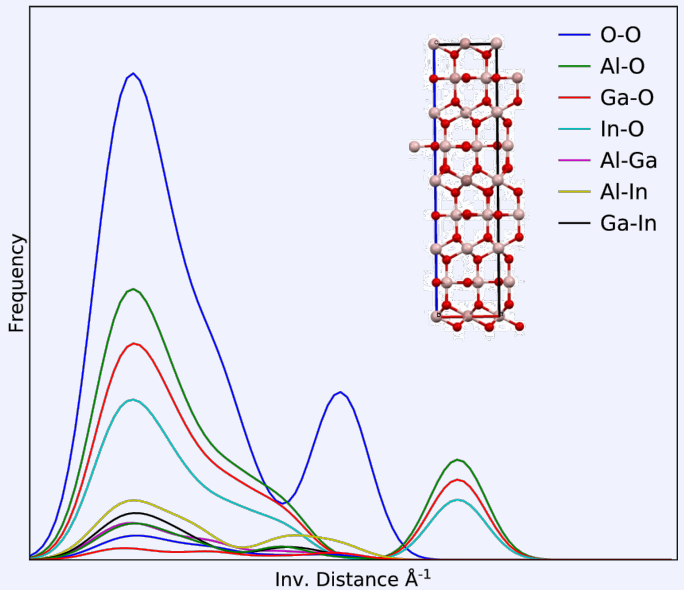
Application 3: model diagnostics

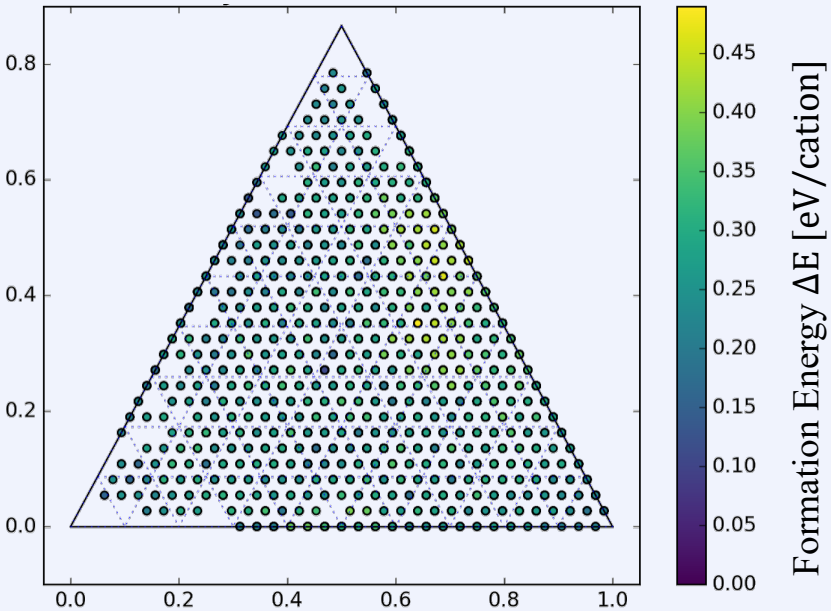
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \;\; \mathrm{a}_y\mathrm{Al}_y\right)_2\mathrm{O}_3\right\}$$

Target

$y = |\Delta E - \tilde{f}_{\Delta E}(c)|$ of regression model $\tilde{f}_{\Delta E}$





Arbitrary model that you want to use for predictions

(Hua and Rupp, 2018)

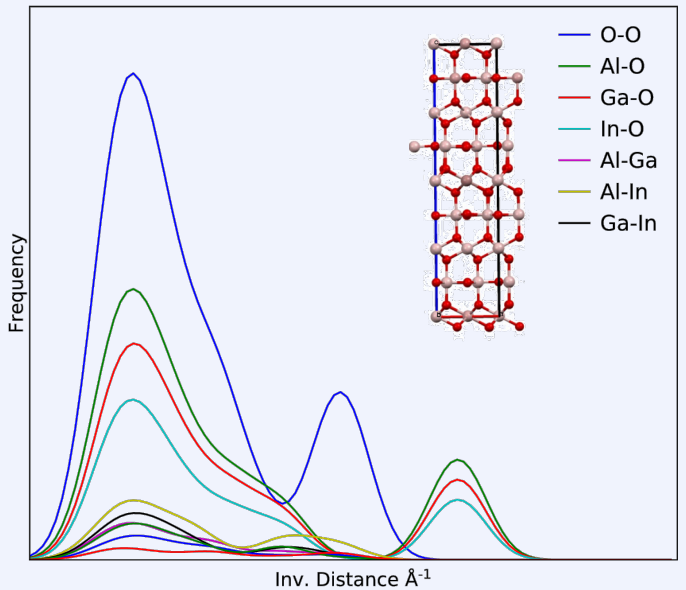
Application 3: model diagnostics

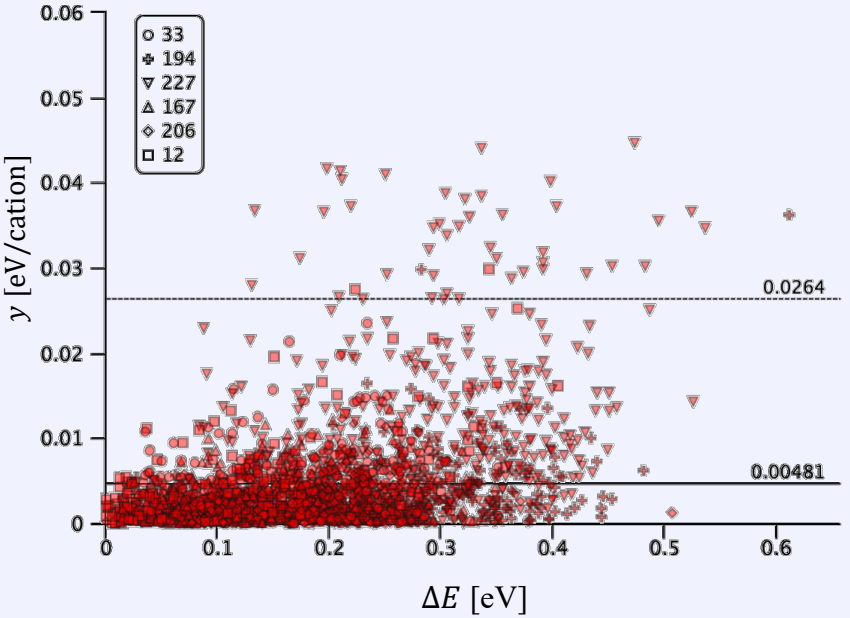
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \;\; \mathrm{a}_y\mathrm{Al}_y\right)_2\mathrm{O}_3\right\}$$

Target

$y = |\Delta E - \tilde{f}_{\Delta E}(c)|$ of regression model $\tilde{f}_{\Delta E}$





Model has low average error, but predictions are generally unreliable

(Hua and Rupp, 2018)

Application 3: model diagnostics

Population

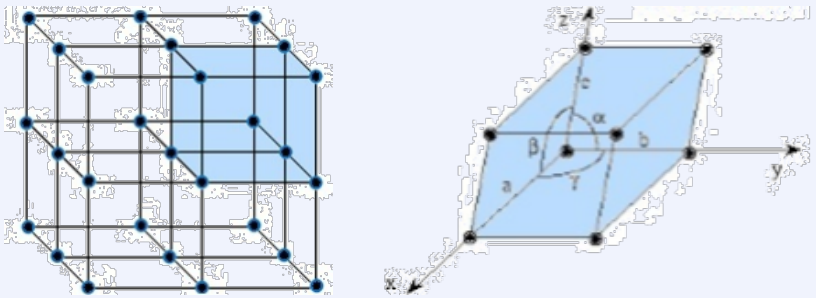
$$P = \left\{c : c \in \left(\mathrm{In}_x \ \ \mathrm{a}_y\mathrm{Al}_y\right)_2\mathrm{O}_3\right\}$$

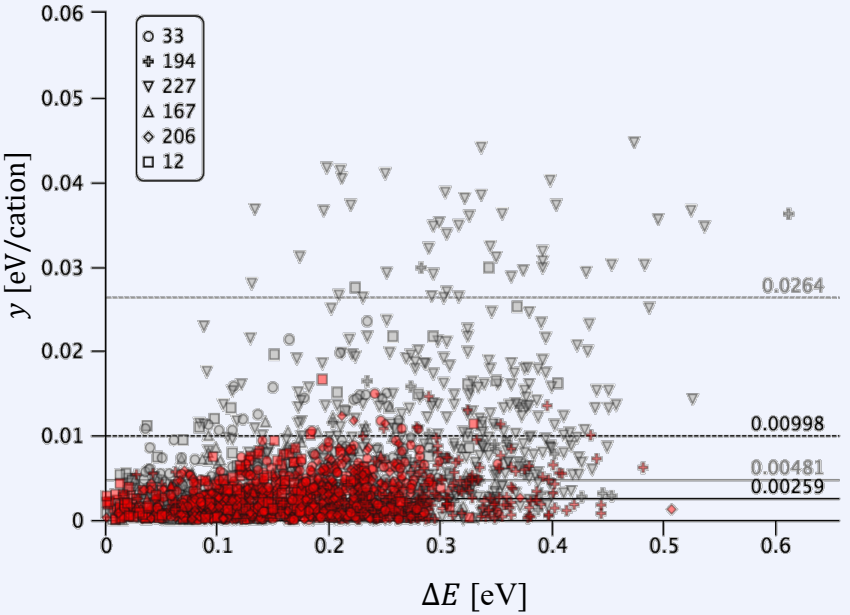
Target

$y = |\Delta E - \tilde{f}_{\Delta E}(c)|$ of regression model $\tilde{f}_{\Delta E}$

Features

$$x \in \{a, b, c, \alpha, \beta, \gamma, n, \ldots\}$$





Model is very reliable under these conditions

Selector $\sigma(c) \equiv n(c) \geq 60 \wedge \gamma(c) \leq 95 \wedge \alpha(c) \leq 121$

Parameters $\mathrm{cov}(\sigma) = 0.6$ $\mathrm{eff}(\sigma) = 0.3$ $[\overline{y}(Q) = 0.003 \pm 0.002,\ \overline{y}(S) = 0.005 \pm 0.007]$

(Hua and Rupp, 2018)

Conclusions

Subgroup discovery is great to explain phenomenae

- very useful in both science and industry

Three choices

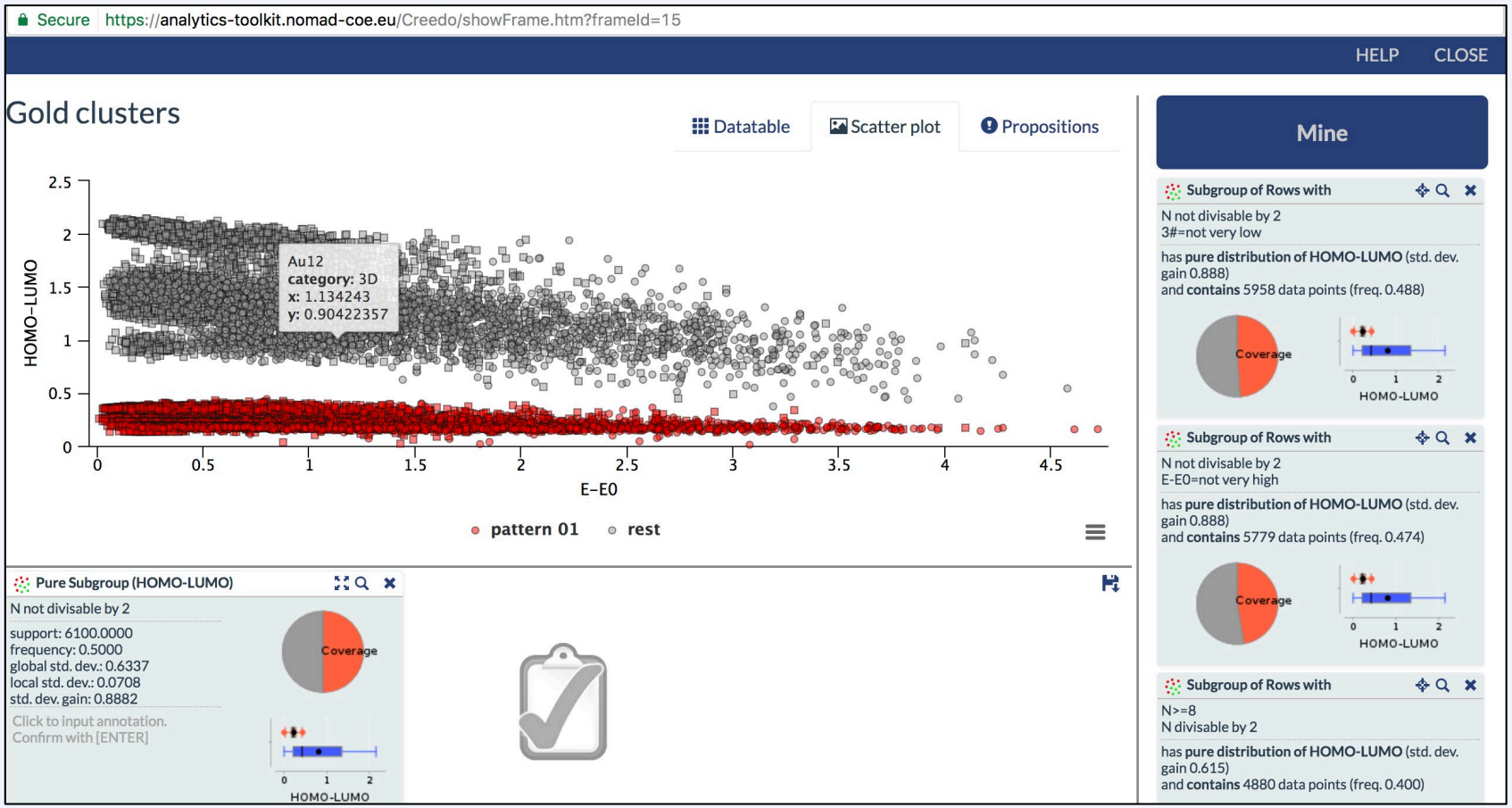
- the target(s) of interest
- how to measure 'exceptionality', score $f(Q)$
- how to select data points, the propositions

Mining subgroups is NP-hard

- greedy or beam search don't give any guarantees
- branch & bound is exact, with guarantees upon early termination
  - the true algorithmic fun is in defining (efficient) (tight) optimistic estimators

Thank you!

Try it yourself!



https://analytics-toolkit.nomad-coe.eu/Creedo/index.htm