

# Introduction to Subgroup Discovery

NOMAD SUMMER  
Lausanne, Sep 26 2018



**Dr. Mario Boley**  
Max Planck Institute for Informatics  
[mboley@mpi-inf.mpg.de](mailto:mboley@mpi-inf.mpg.de)

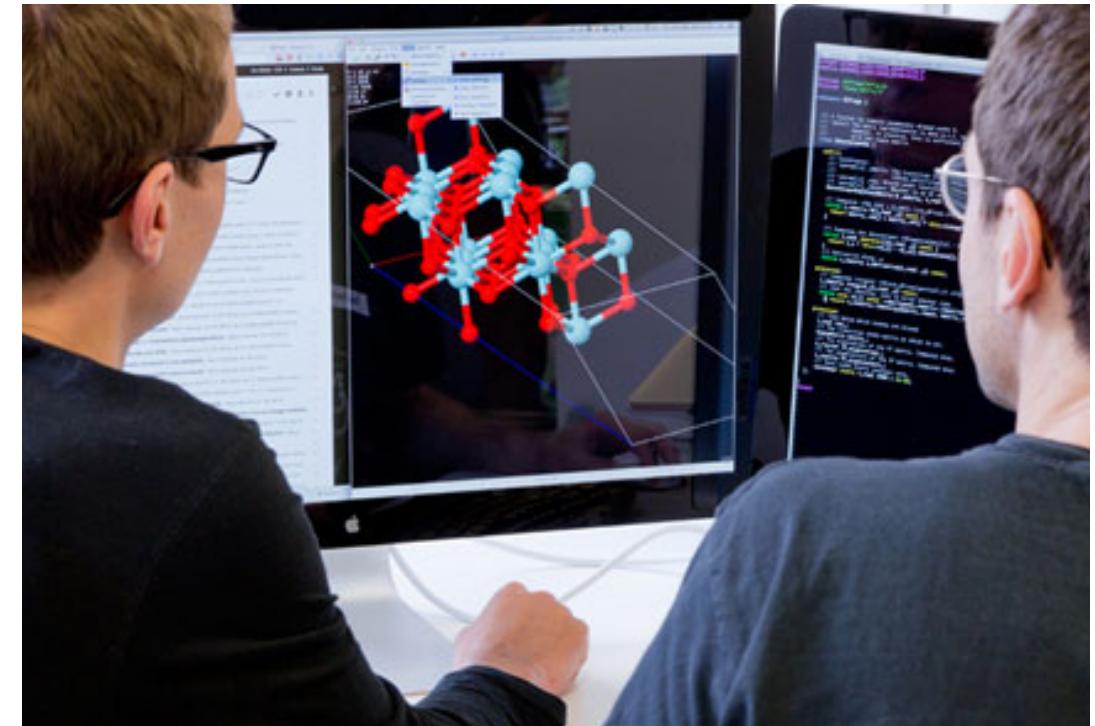


# Two flavors of data science

2



Predictive modelling



Data analytics

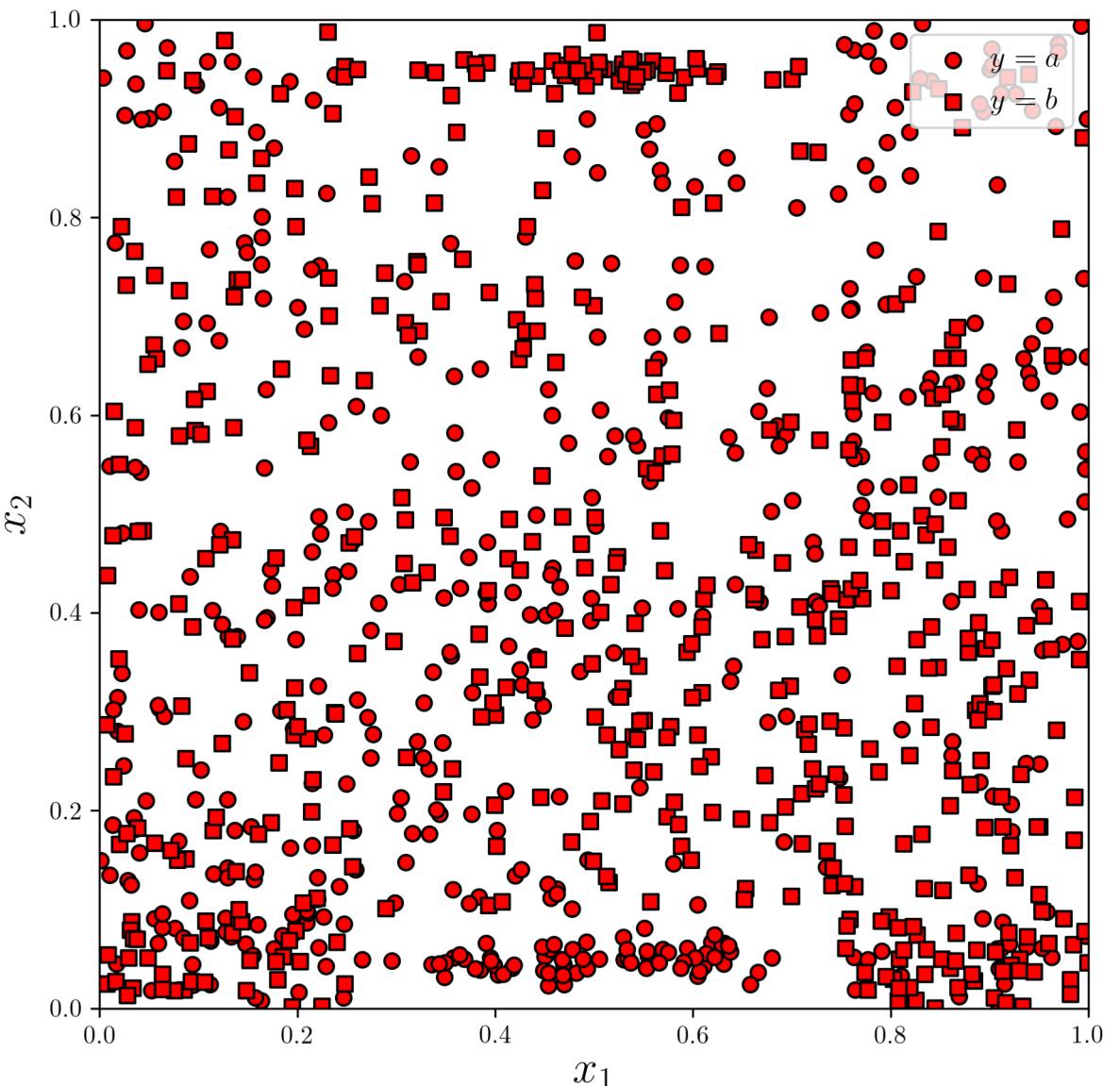
# Basic setup

**Given**

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{a, b, c, \dots\}$

Features  $x_j: P \rightarrow X_j$



# Basic setup

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{a, b, c, \dots\}$

Features  $x_j: P \rightarrow X_j$

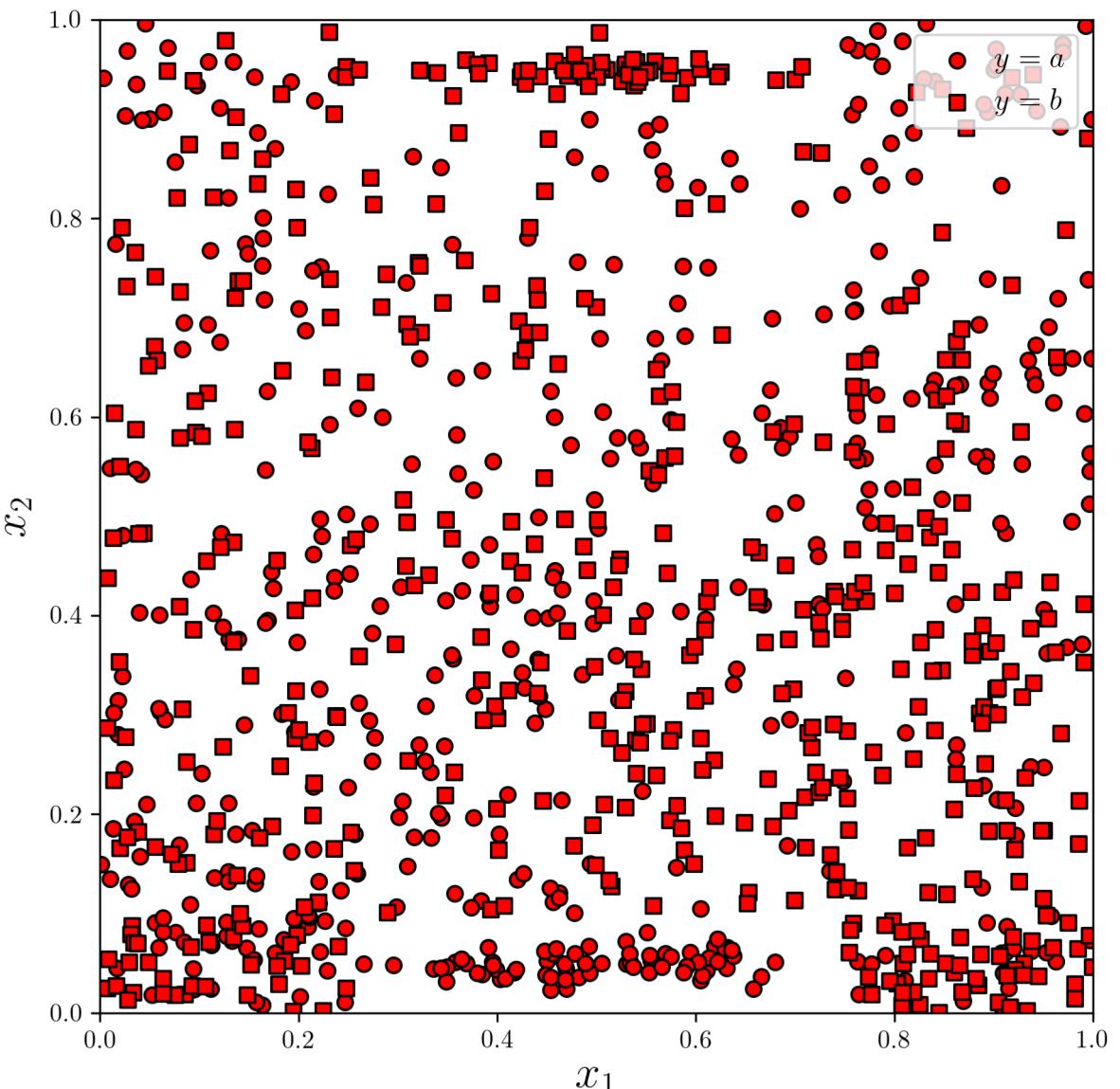
## Predictive modeling

Model class  $M \subseteq X \rightarrow Y$

Error measure  $\text{err}: Y \times Y \rightarrow \mathbb{R}_+$

## Minimize

$$f(m) = \sum_{i \in S} \text{err}(m(x(i)), y(i)) / |S| + \gamma \|m\|$$



# Global models classify whole space

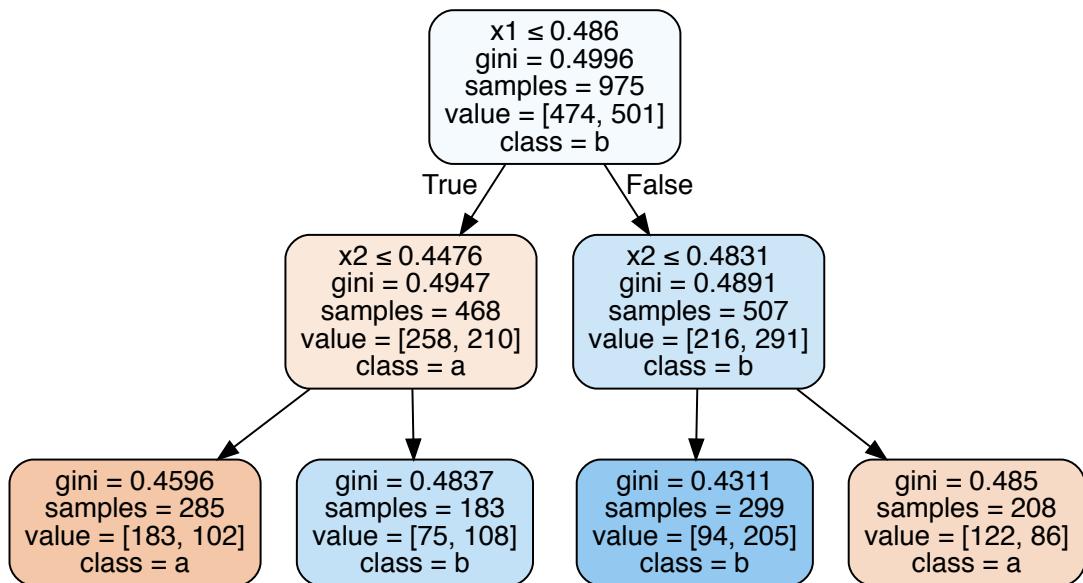
**Given**

Sample  $S \subseteq P$

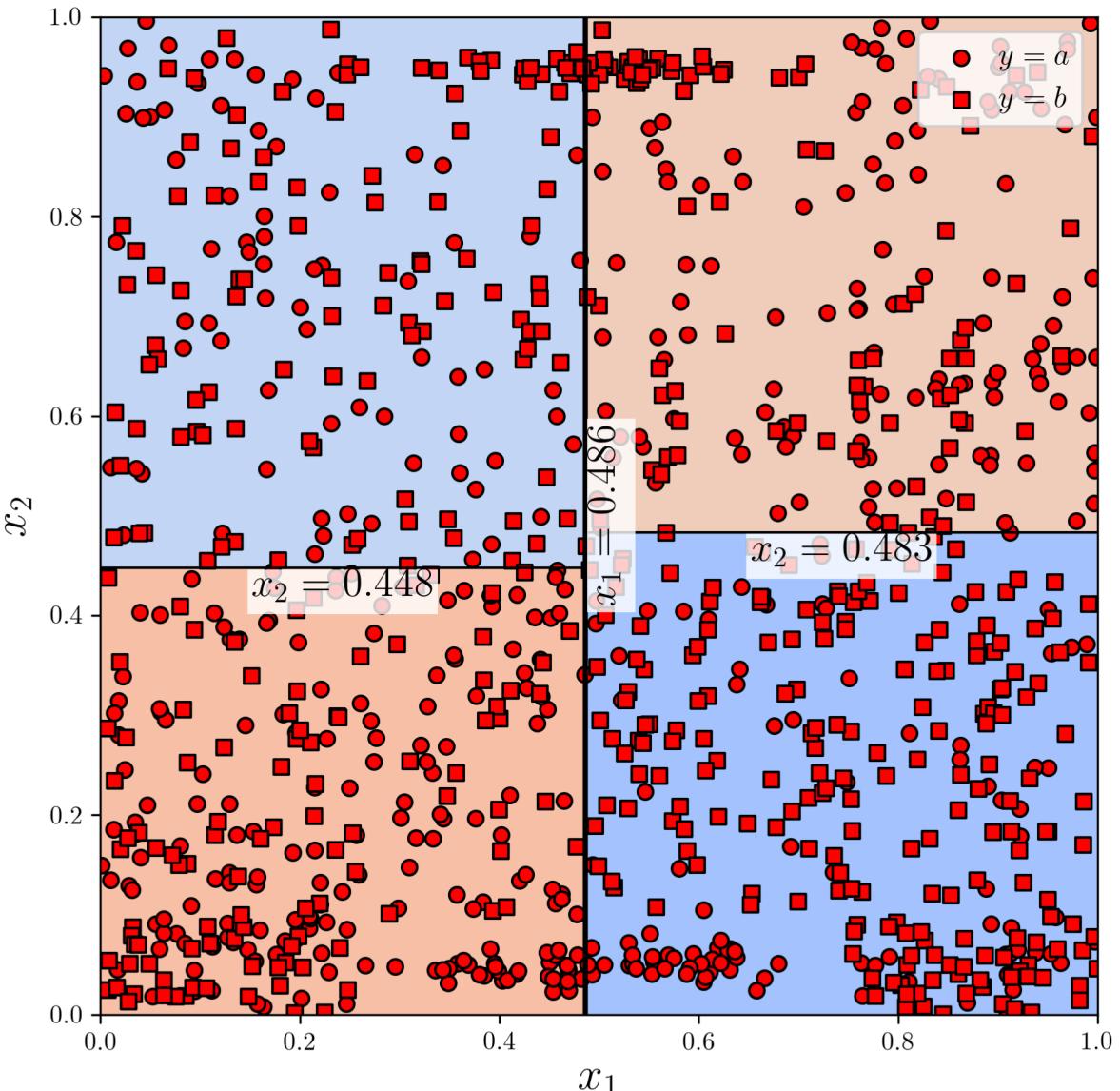
Target variable  $y: P \rightarrow \{a, b, c, \dots\}$

Features  $x_j: P \rightarrow X_j$

**Decision tree**



Misses interesting local phenomena



# Subgroup discovery focusses on local observations

6

## Given

Sample  $S \subseteq P$

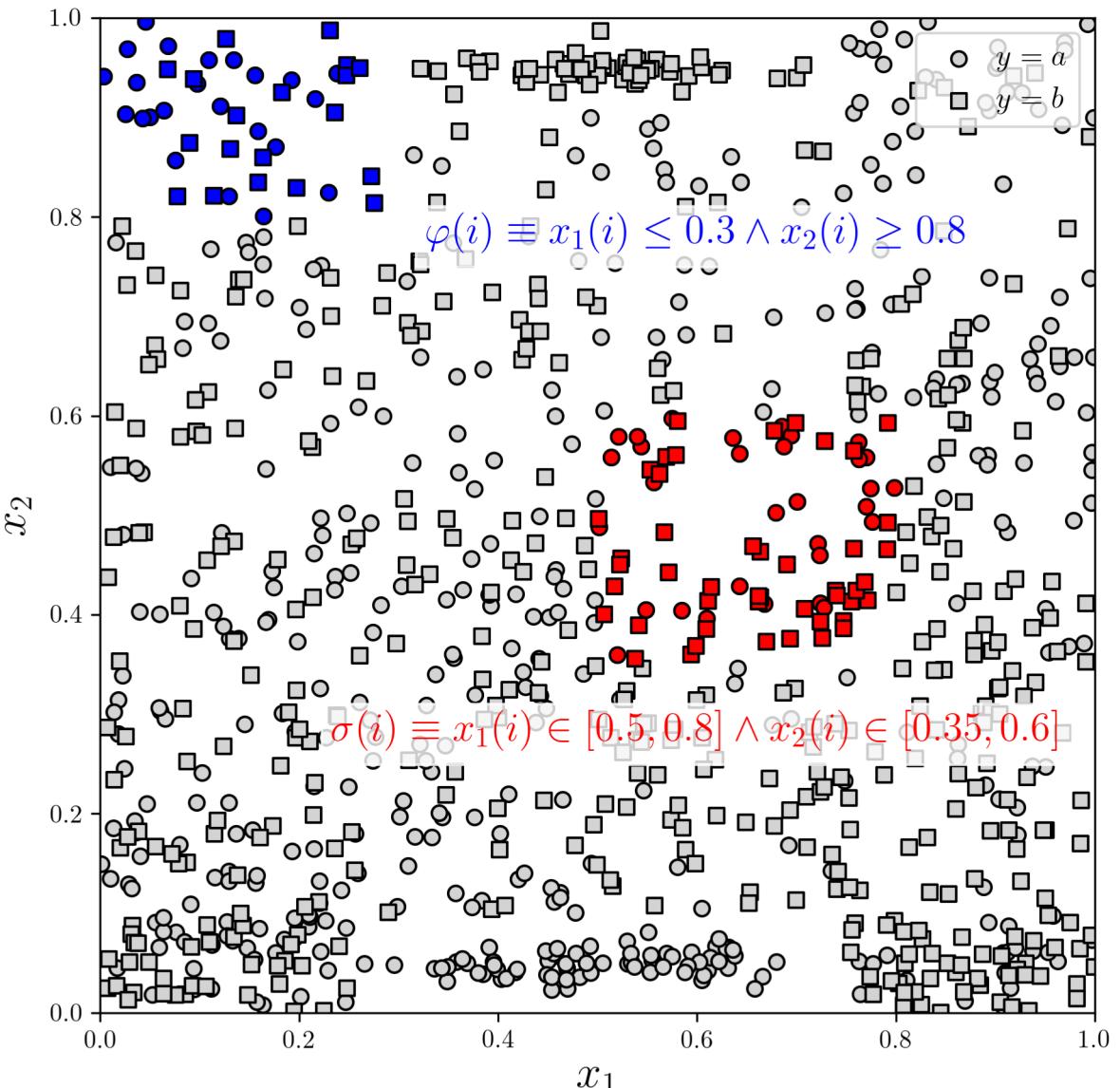
Target variable  $y: P \rightarrow \{a, b, c, \dots\}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$



# Subgroup discovery focusses on local observations

7

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{a, b, c, \dots\}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

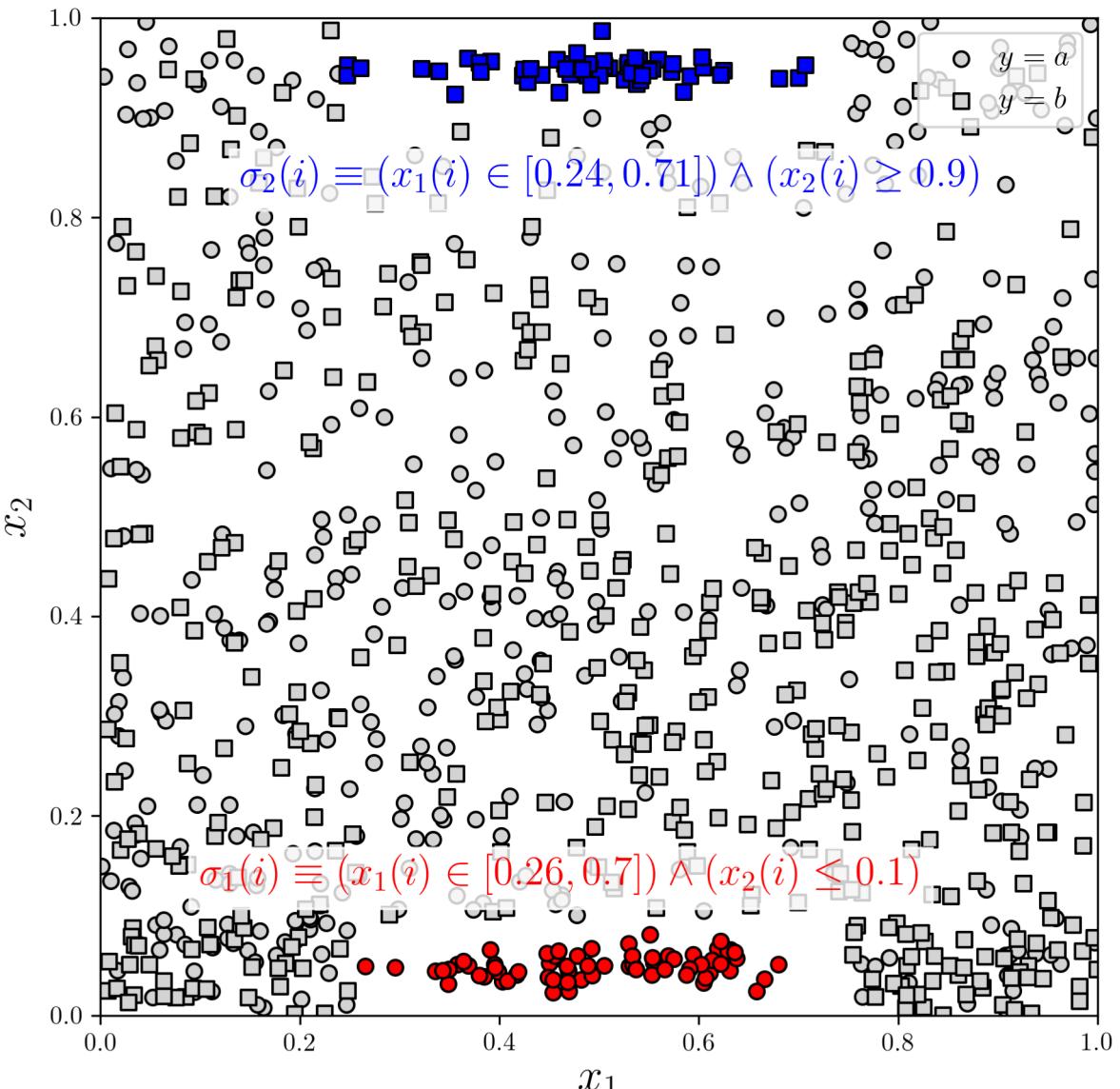
Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

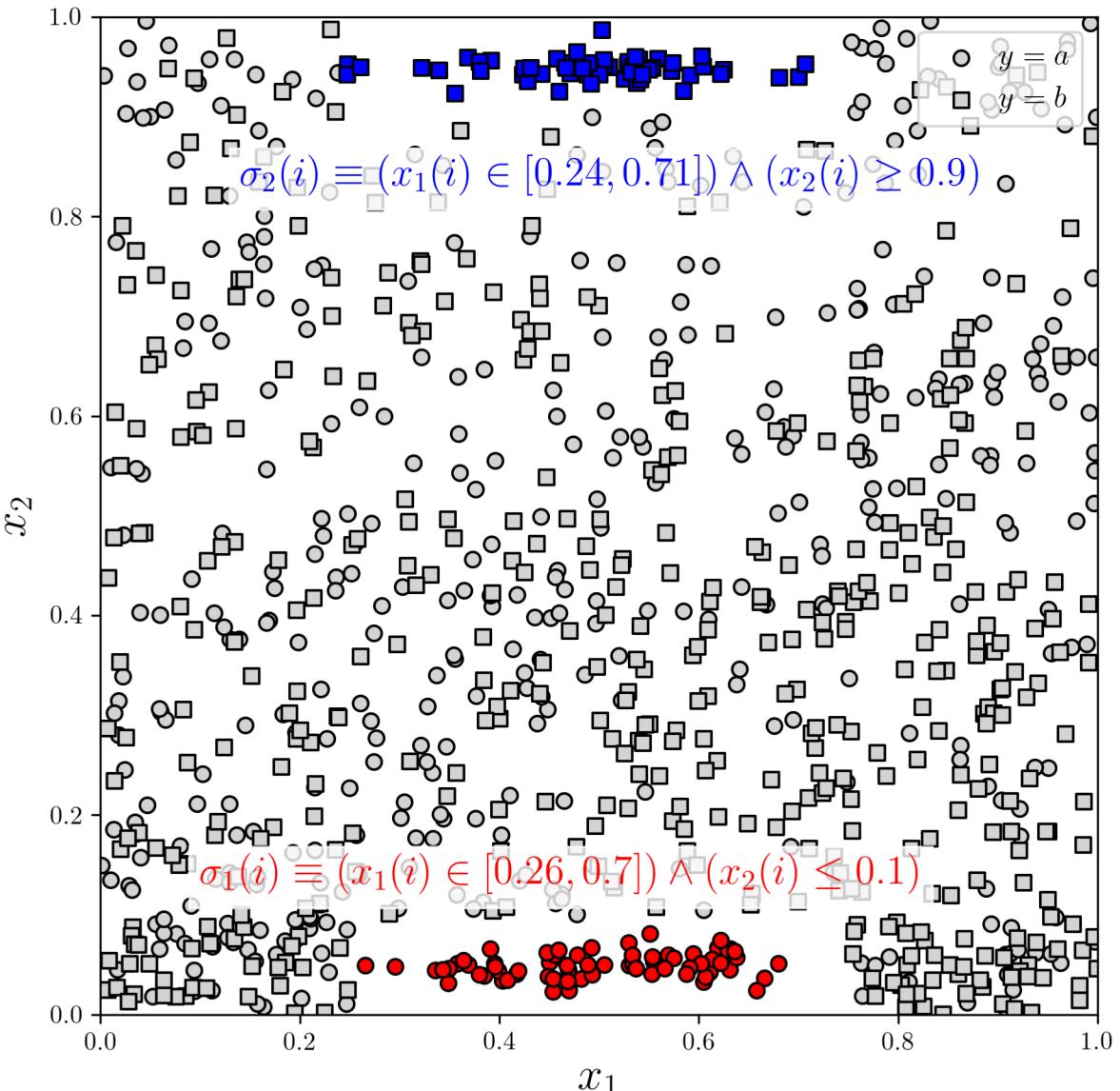
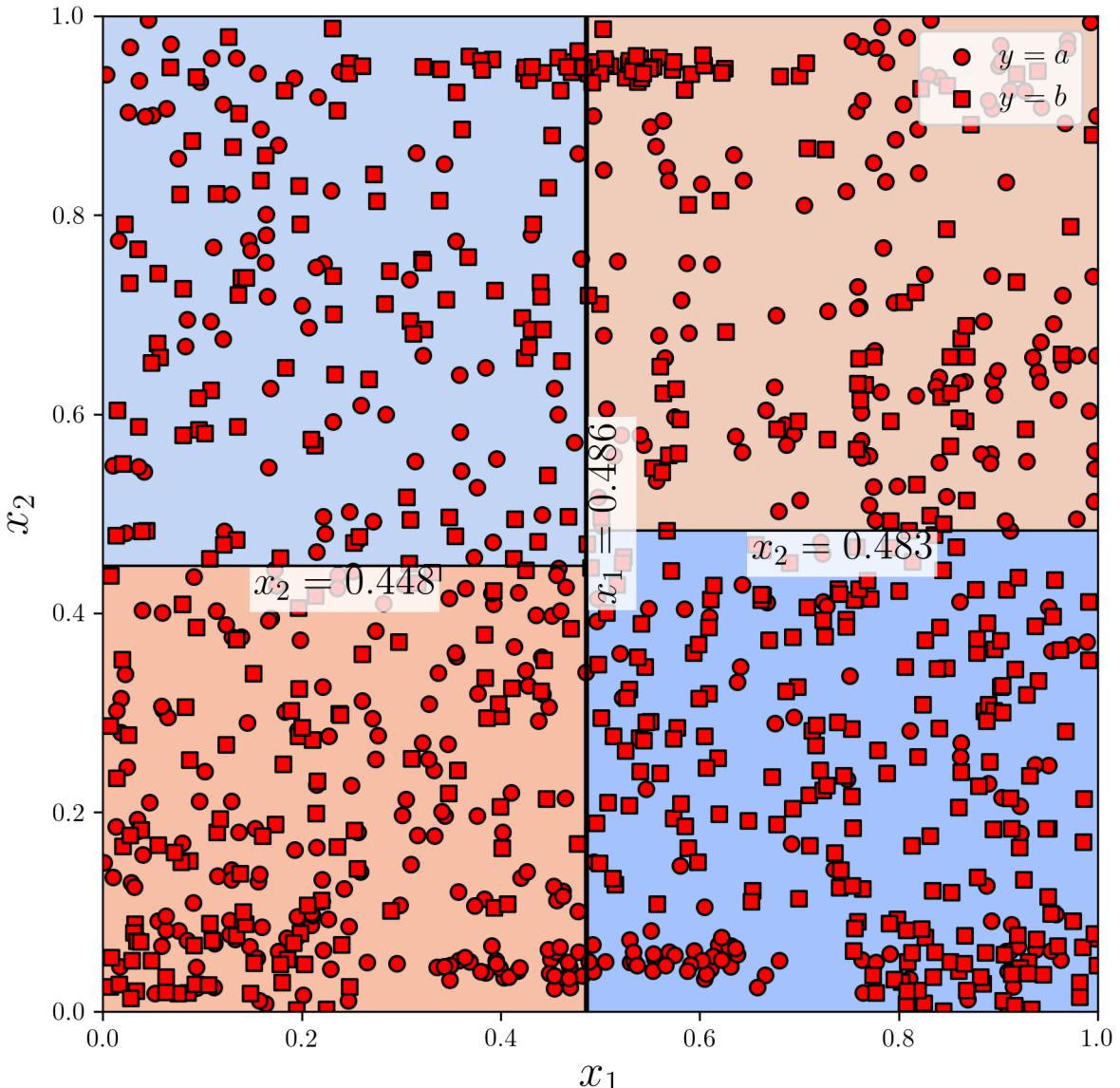
with

- $Q = \{i \in S: \sigma(i) = T\}$  extension
- $\text{cov}(Q) = |Q|/|S|$  coverage
- $\text{eff}(Q) = (H_y(S) - H_y(Q))/H_y(S)$  effect
- $H_y(Q) = -\sum_v p_Q(y=v) \log p_Q(y=v)$  entropy



# Subgroup discovery focusses on local observations

8



# Application 1: octet binary crystal structures

9

## Population

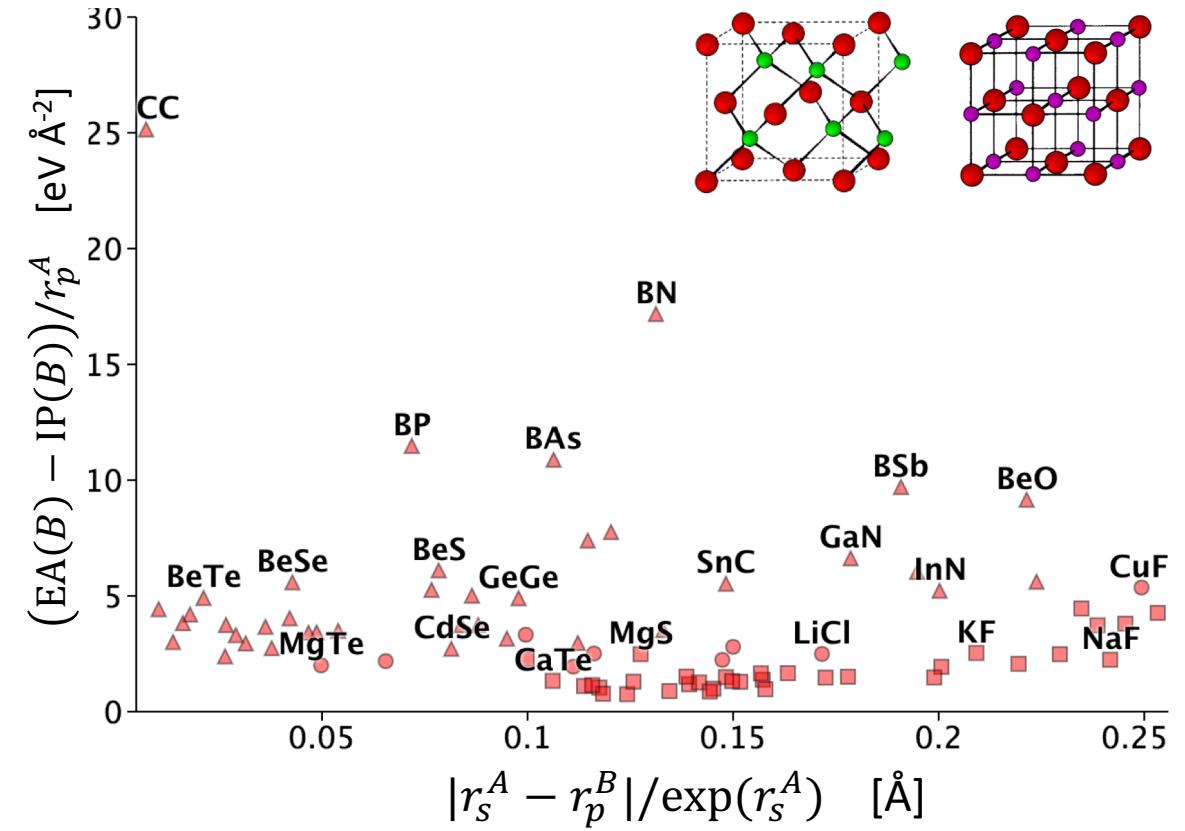
$$P = \{AB : A = \text{Ag, Al, Ba, ...} \wedge B = \text{Br, Cl, F, ...}\}$$

## Target

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

## Features

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



# Application 1: octet binary crystal structures

10

## Population

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

## Target

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

## Features

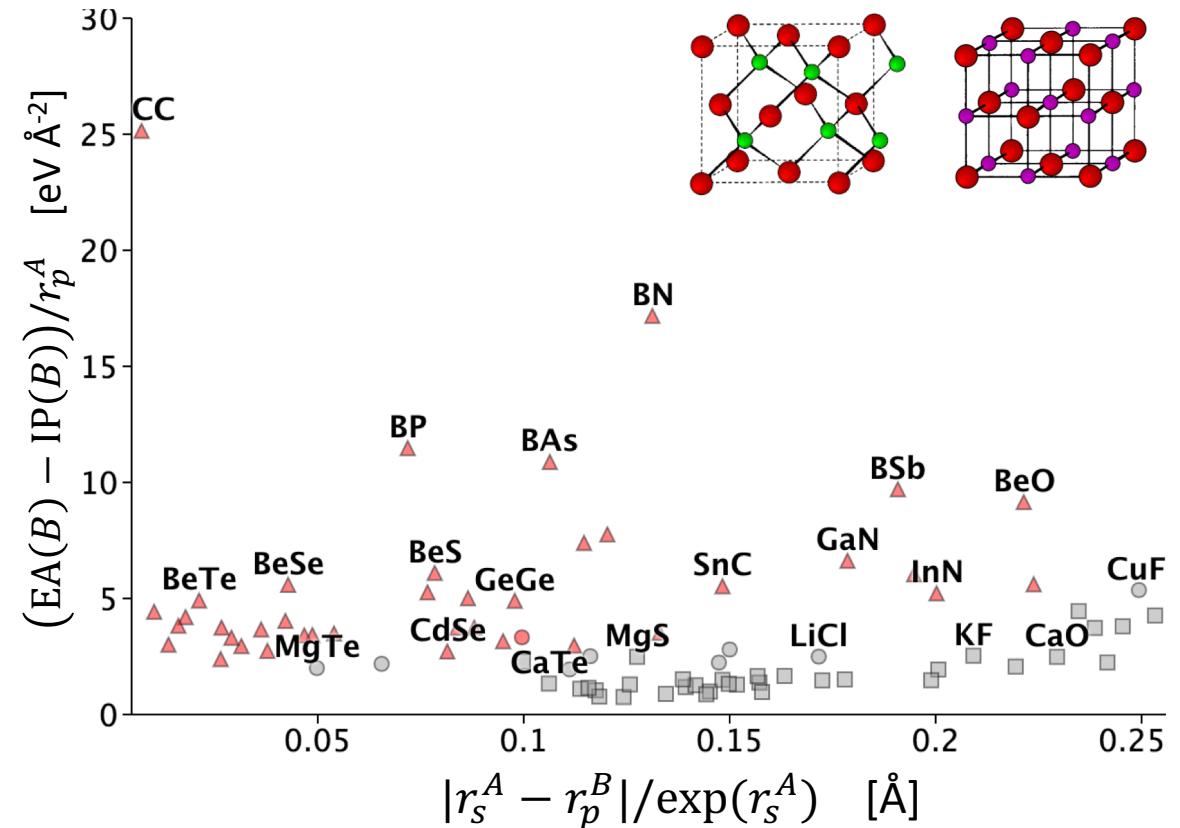
$$x \in \{\text{IP}^A, \text{EA}^A, r_s^A, r_p^A, r_d^A, \text{IP}^B, \text{EA}^B, r_s^B, r_p^B, r_d^B, \text{IP}^A - \text{IP}^B, \text{EA}^A - \text{EA}^B, |r_s^A - r_s^B|, \dots\}$$

## Selector

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

## Parameters

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



# Application 1: octet binary crystal structures

11

## Population

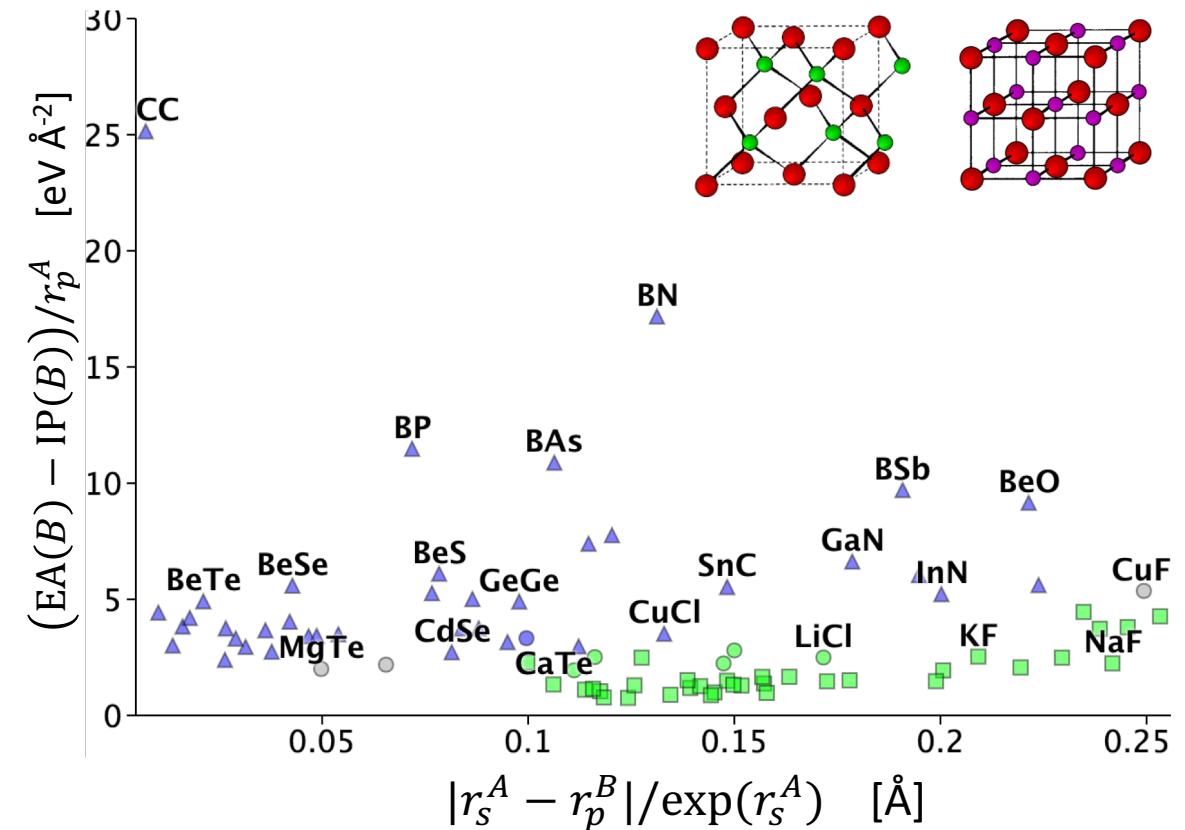
$$P = \{AB : A = \text{Ag, Al, Ba, ...} \wedge B = \text{Br, Cl, F, ...}\}$$

## Target

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

## Features

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



## Selector

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

## 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 1: octet binary crystal structures

12

## Population

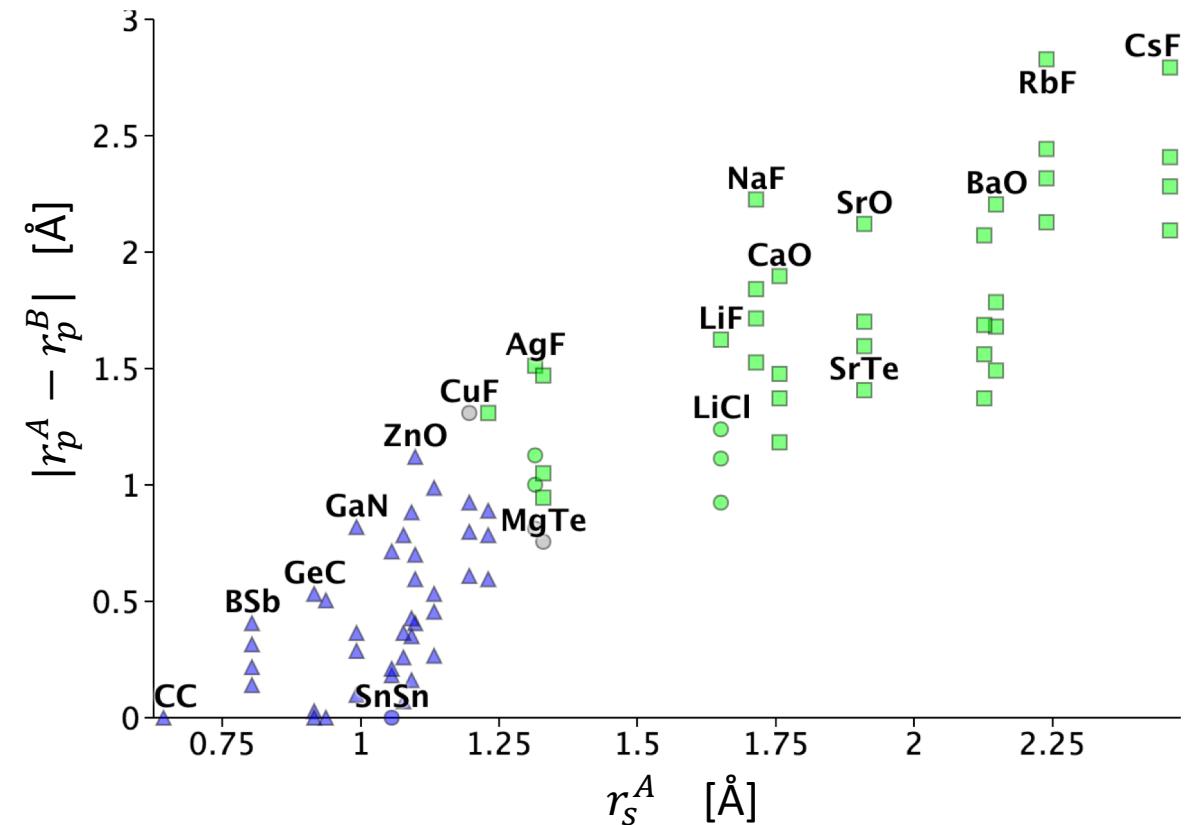
$$P = \{AB : A = \text{Ag, Al, Ba, ...} \wedge B = \text{Br, Cl, F, ...}\}$$

## Target

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

## Features

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



## Selector

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

## Parameters

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

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

# Subgroup discovery for real-valued targets

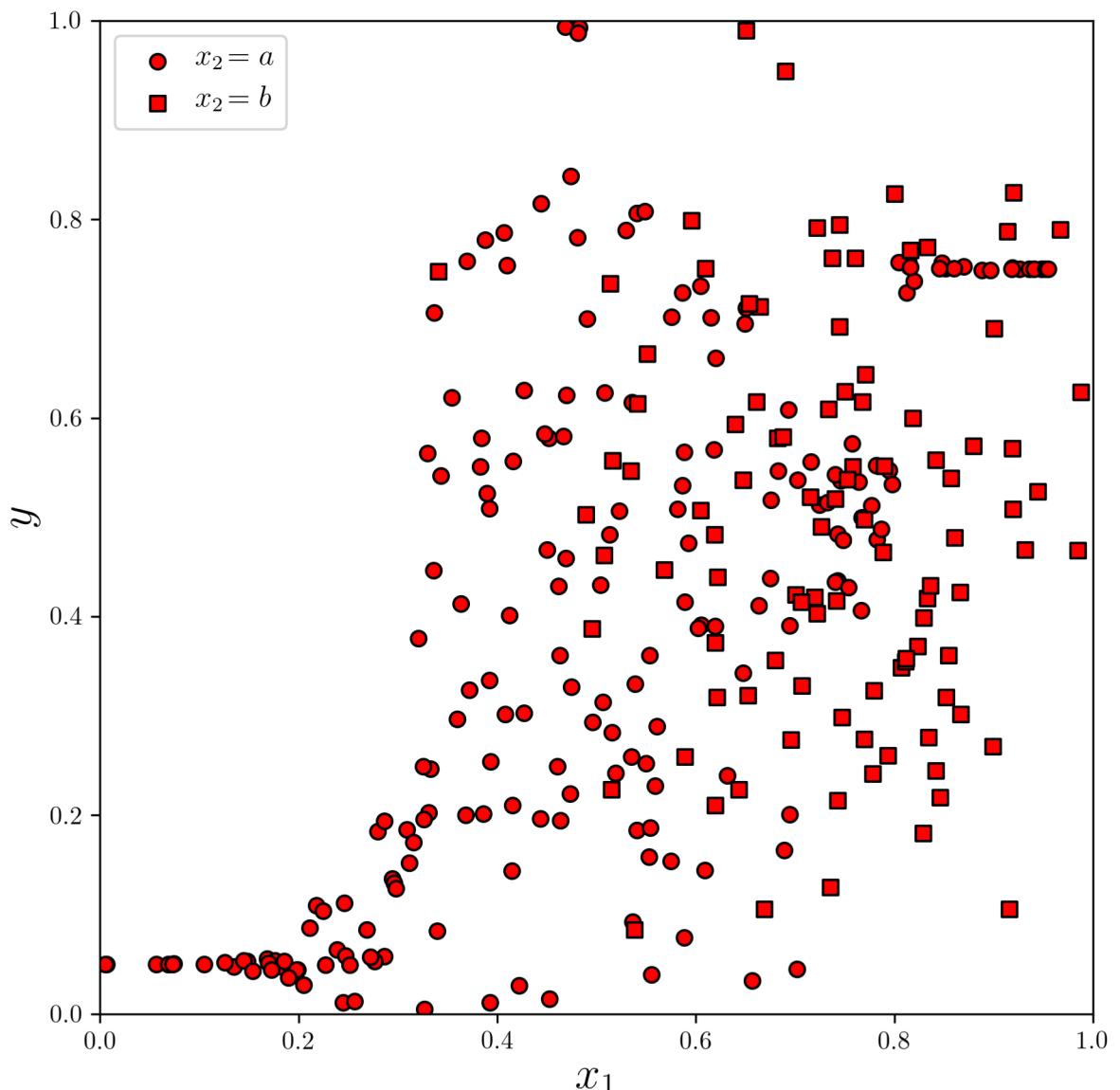
13

**Given**

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$



# Subgroup discovery for real-valued targets

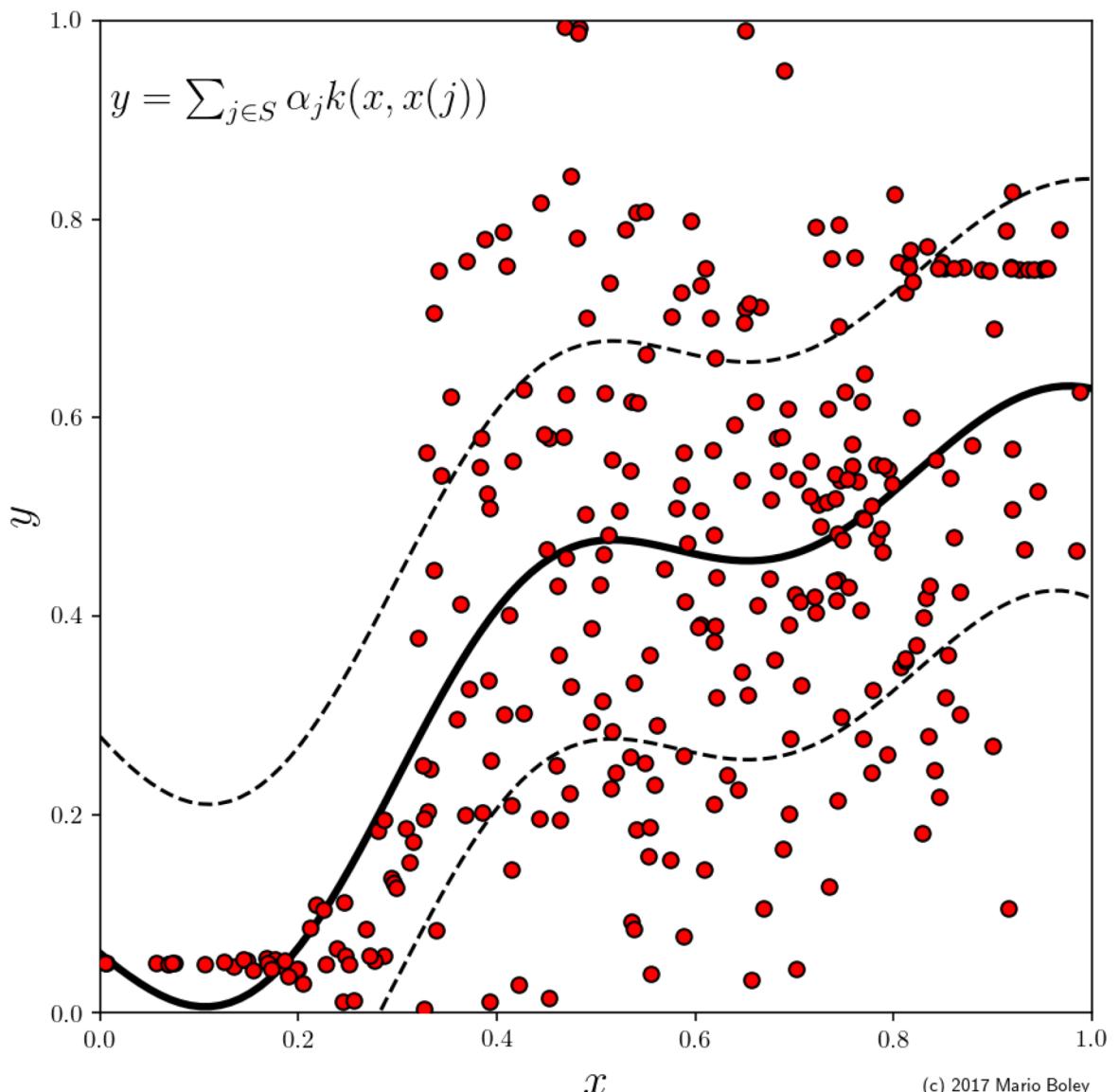
14

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$



$x$

(c) 2017 Mario Boley

©2017, Mario Boley

# Subgroup discovery for real-valued targets

15

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

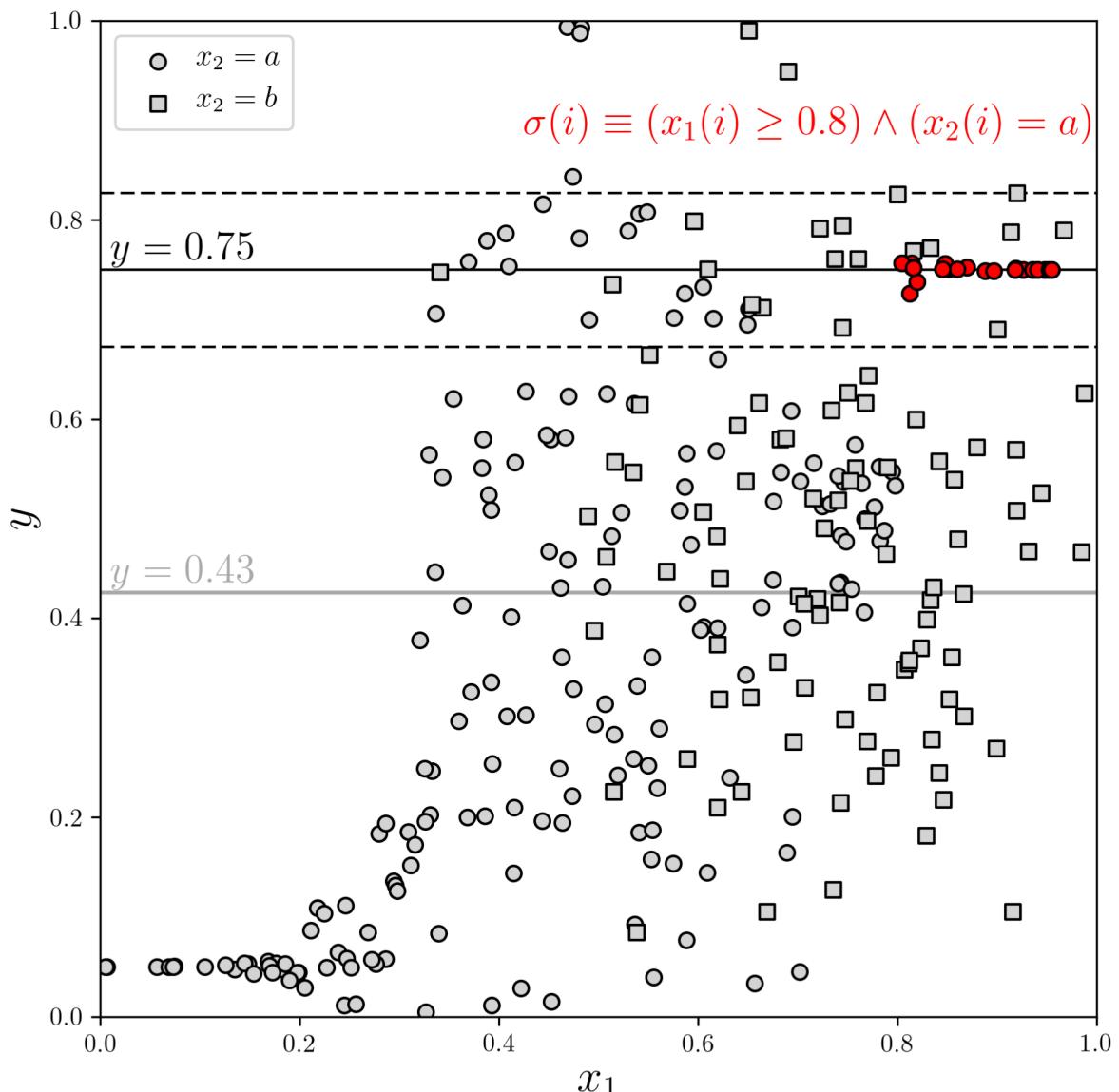
Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \{i \in S : \sigma(i) = T\}$  extension
- $\text{cov}(Q) = |Q|/|S|$  coverage
- $\text{eff}(Q) = (s_y(S) - s_y(Q)) / s_y(S)$  effect
- $s_y(Q) = \sqrt{\sum_{i \in Q} (\bar{y} - y(i))^2 / (|Q| - 1)}$  std. dev.



# Application 2: Au structure/property relationship

16

## Population

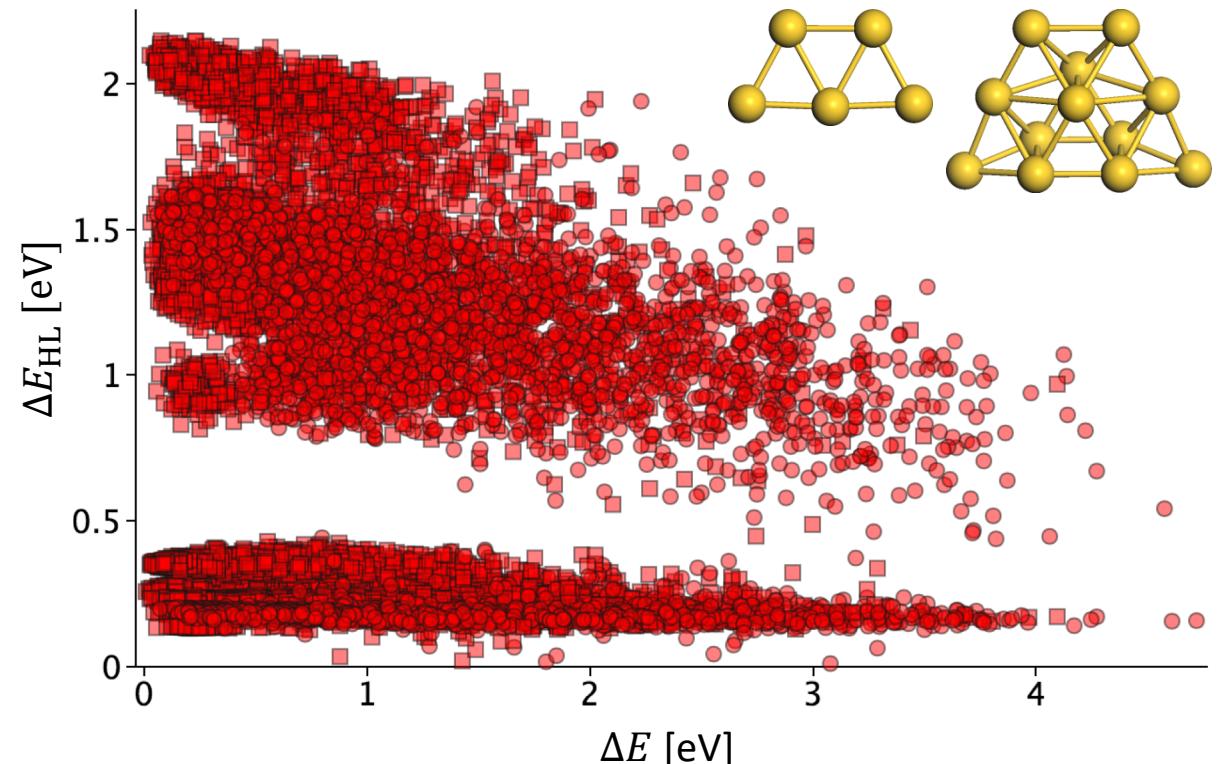
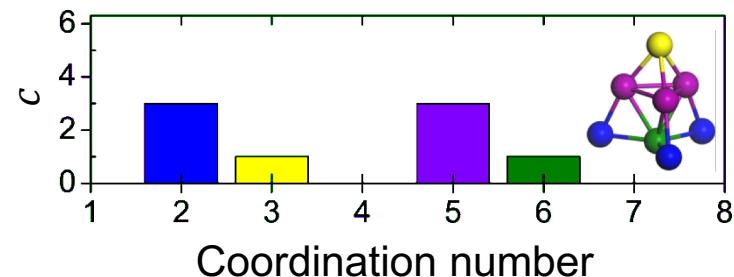
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Target

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

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



# Application 2: Au structure/property relationship

17

## Population

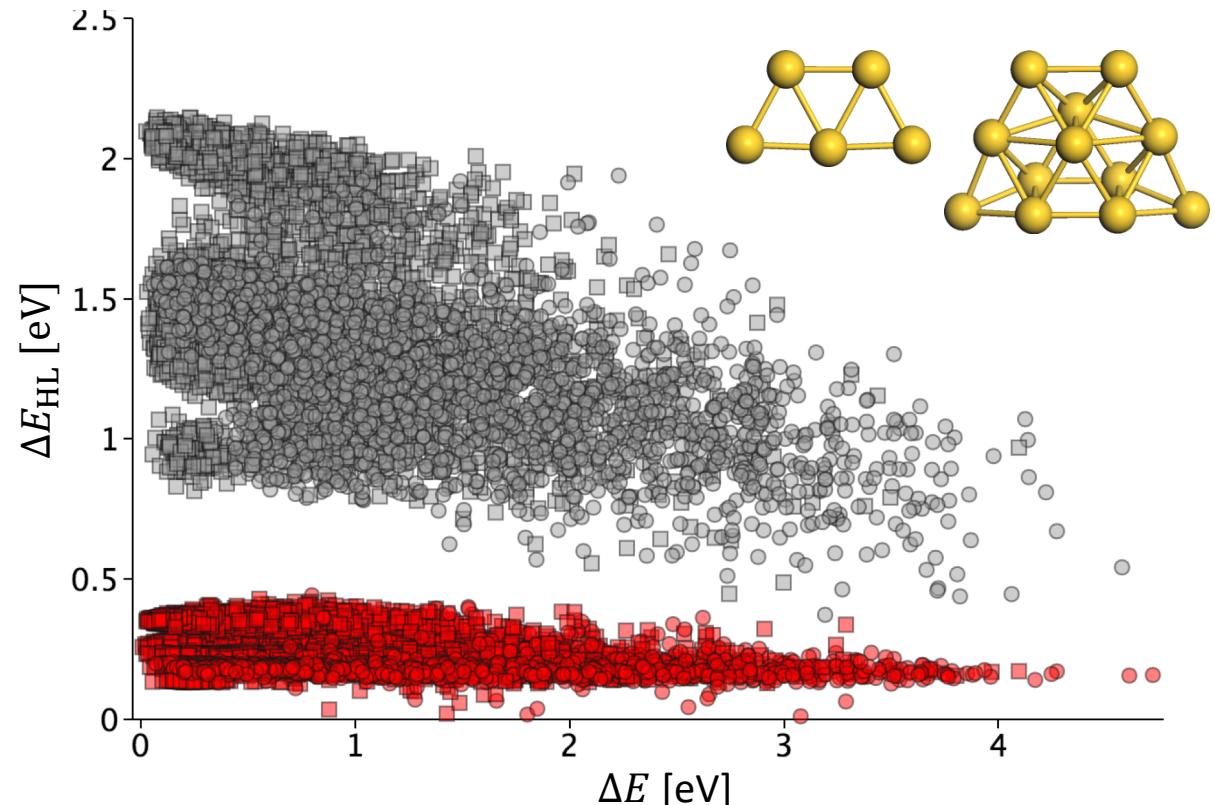
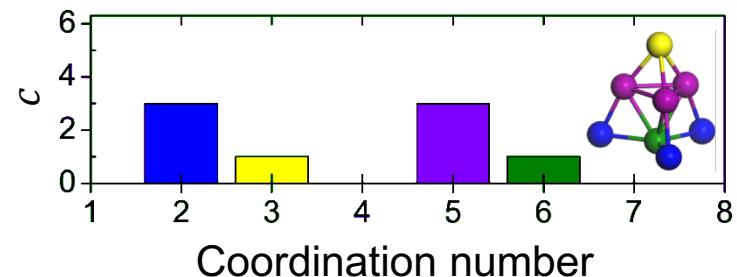
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Target

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

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



## Selector

$$\sigma(i) \equiv \text{odd}(a(i))$$

## Parameters

$$\begin{aligned} \text{cov}(\sigma) &= 0.5 & \text{eff}(\sigma) &= 0.9 & [s_y(Q) &= 0.06, s_y(S) = 0.58] \\ &&&& [\bar{y}(Q) &= 0.22, \bar{y}(S) = 0.42] \end{aligned}$$

# Application 2: Au structure/property relationship

18

## Population

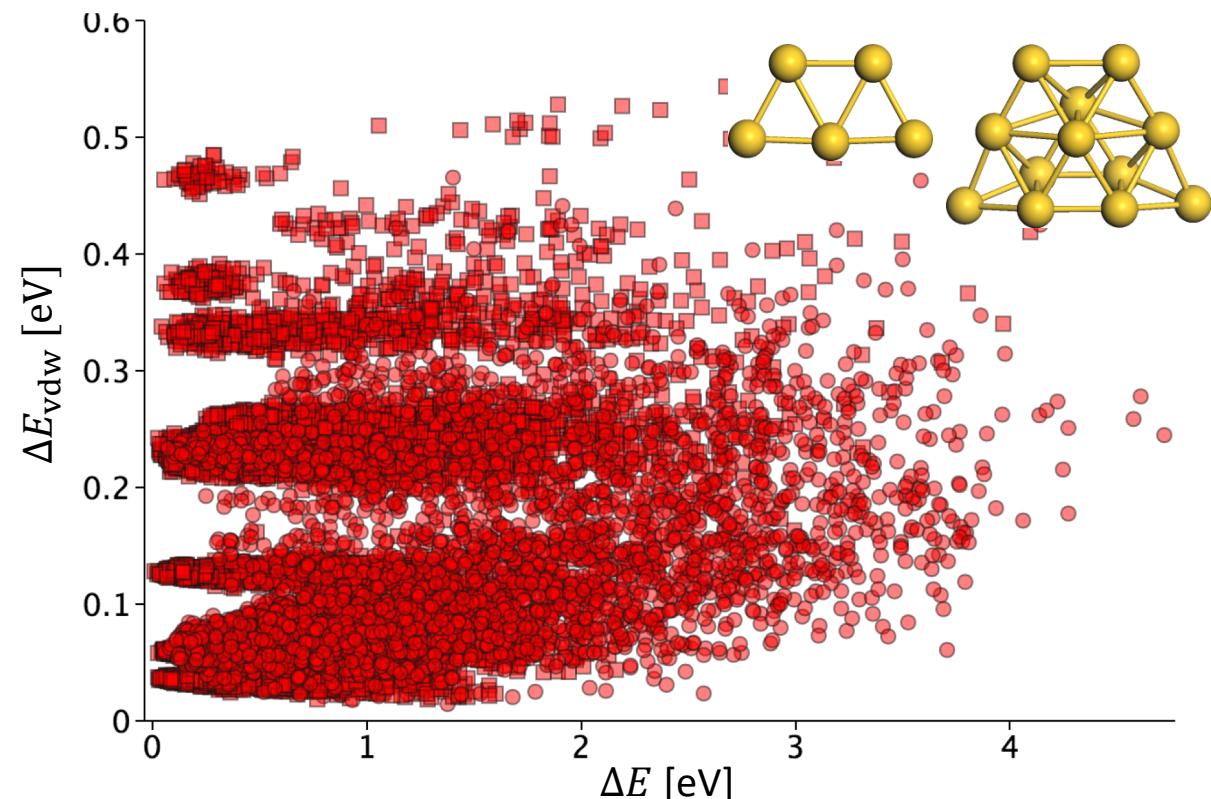
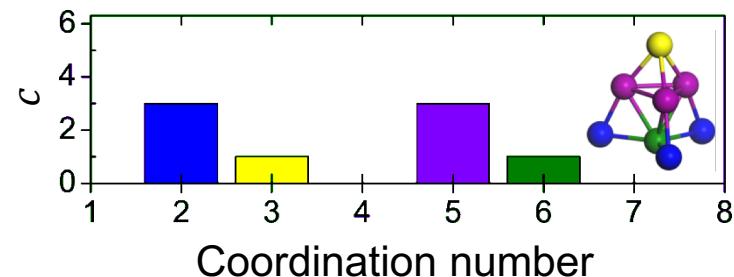
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Target

$$y = \Delta E_{\text{vdw}} \text{ van der Waals energy (ref.)}$$

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



# Application 2: Au structure/property relationship

19

## Population

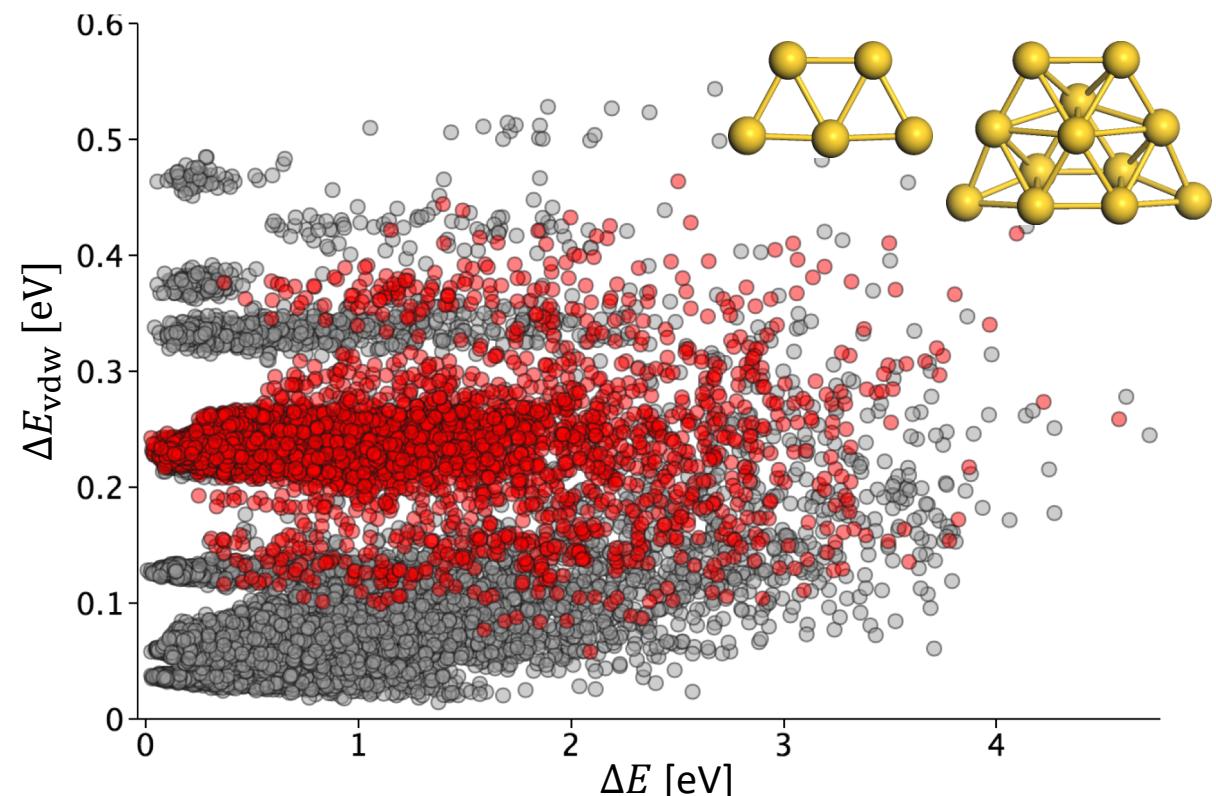
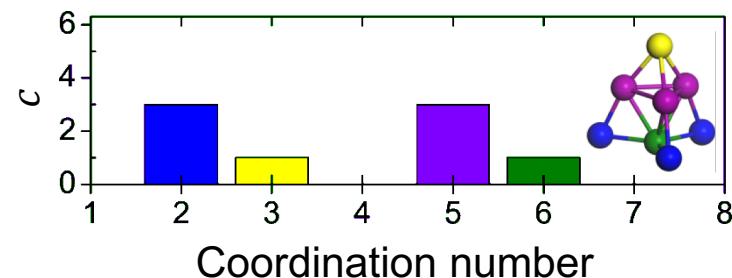
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Target

$$y = \Delta E_{\text{vdw}} \text{ van der Waals energy (ref.)}$$

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



## Selector

$$\sigma(i) \equiv a(i) \in [8,12] \wedge c_2(i) > 0.17 \wedge c_6(i) < 0.28 \wedge r(i) > 0.86$$

## Parameters

$$\begin{aligned} \text{cov}(\sigma) &= 0.2 & \text{eff}(\sigma) &= 0.68 & [s_y(Q) &= 0.03, s_y(S) = 0.09] \\ &&&& [\bar{y}(Q) &= 0.23, \bar{y}(S) = 0.13] \end{aligned}$$

# Subgroup discovery with multiple targets

20

## Given

Sample  $S \subseteq P$

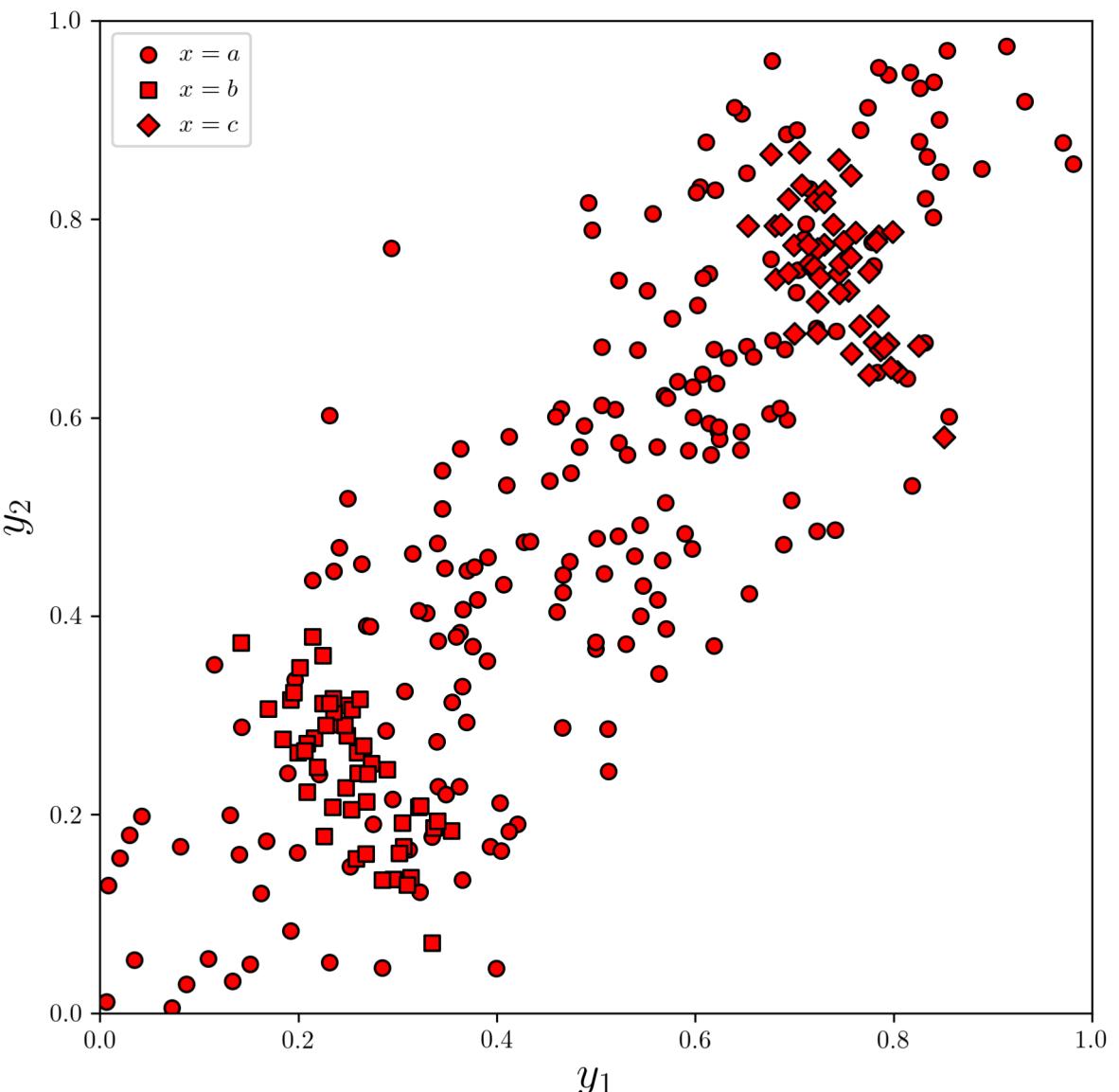
Target variable  $y_1, y_2: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$



# Subgroup discovery with multiple targets

21

## Given

Sample  $S \subseteq P$

Target variable  $y_1, y_2: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

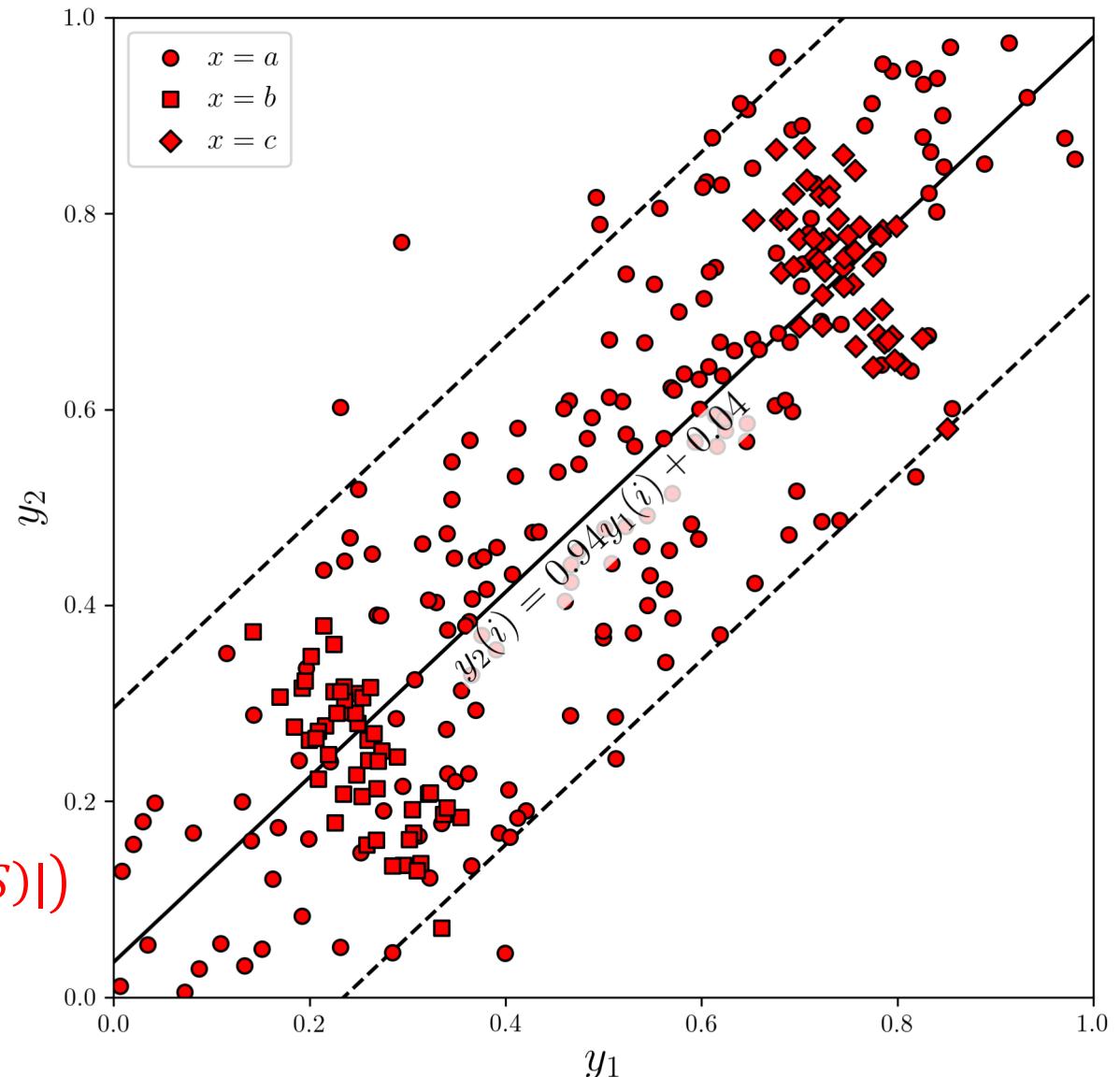
Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \{i \in S: \sigma(i) = T\}$
- $\text{cov}(Q) = |Q|/|S|$
- $\text{eff}(Q) = (|r_{y_1, y_2}(Q)| - |r_{y_1, y_2}(S)|)/(1 - |r_{y_1, y_2}(S)|)$
- $r(Q) = \frac{1}{|Q|-1} \sum_{i \in Q} \left( \frac{\bar{y}_1(Q) - y_1(i)}{s_{y_1}(Q)} \right) \left( \frac{\bar{y}_2(Q) - y_2(i)}{s_{y_2}(Q)} \right)$



# Subgroup discovery with multiple targets

22

## Given

Sample  $S \subseteq P$

Target variable  $y_1, y_2: P \rightarrow \mathbb{R}$

Features  $x_j: P \rightarrow X_j$

## Define

Propositions  $\Pi_x = \{\pi_1, \dots, \pi_k\}$

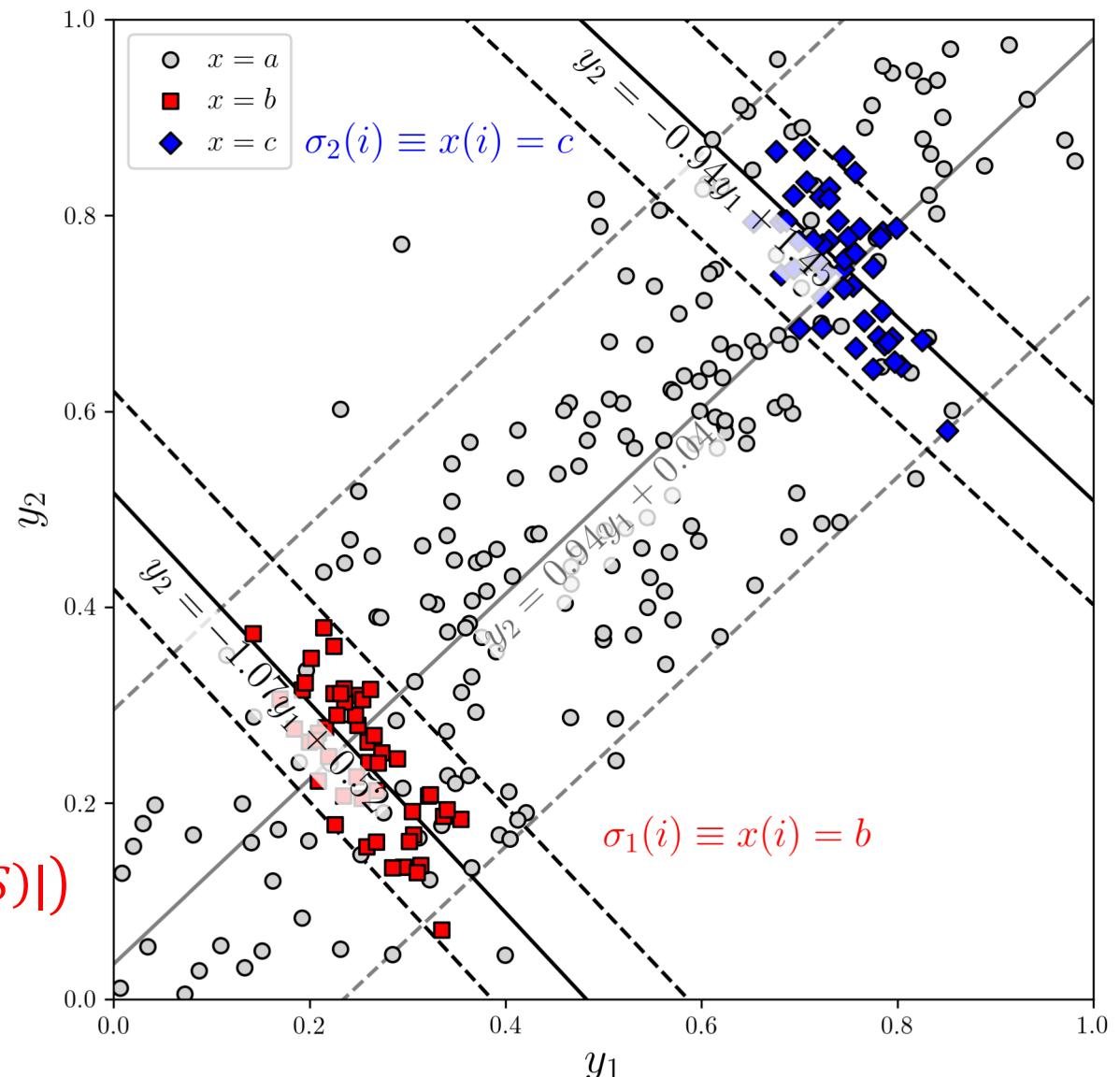
Selection language  $\mathcal{L}_x = \{\sigma(i) = \pi_{j_1}(i) \wedge \dots \wedge \pi_{j_l}(i)\}$

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \{i \in S: \sigma(i) = T\}$
- $\text{cov}(Q) = |Q|/|S|$
- $\text{eff}(Q) = (|r_{y_1, y_2}(Q)| - |r_{y_1, y_2}(S)|)/(1 - |r_{y_1, y_2}(S)|)$
- $r(Q) = \frac{1}{|Q|-1} \sum_{i \in Q} \left( \frac{\bar{y}_1(Q) - y_1(i)}{s_{y_1}(Q)} \right) \left( \frac{\bar{y}_2(Q) - y_2(i)}{s_{y_2}(Q)} \right)$



# Application 2: Au structure/property relationship

23

## Population

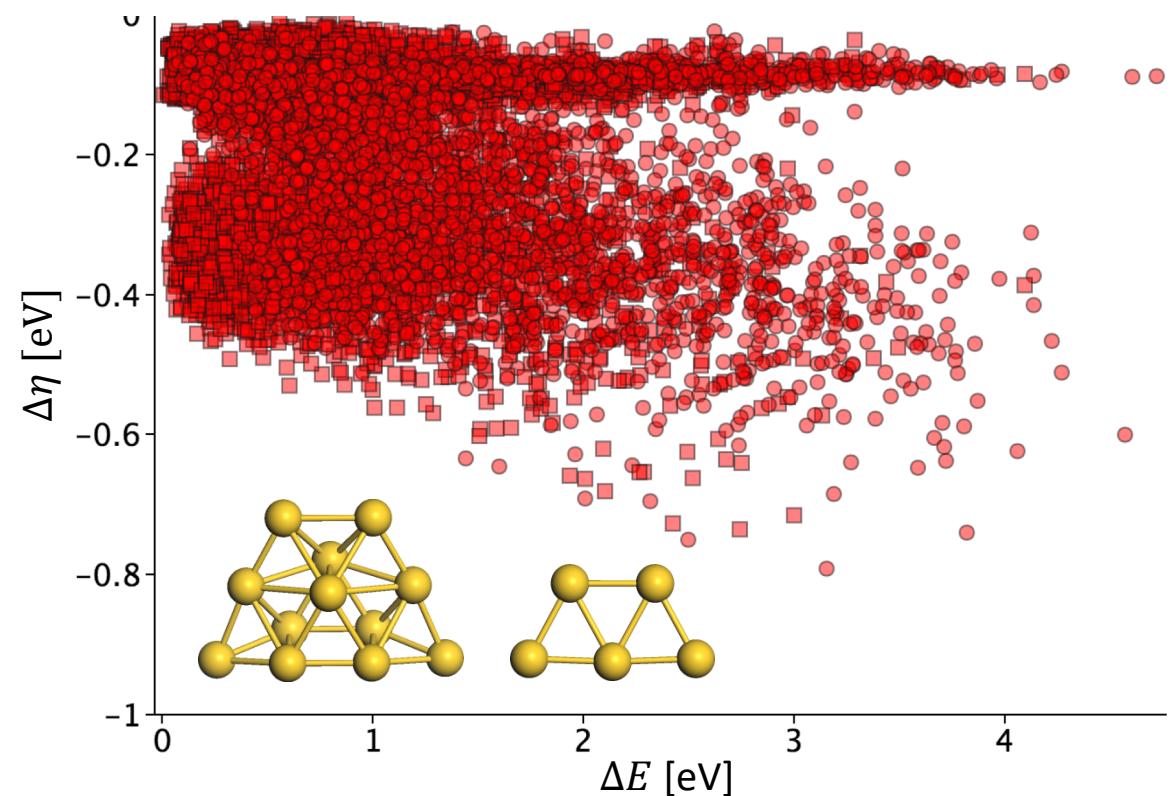
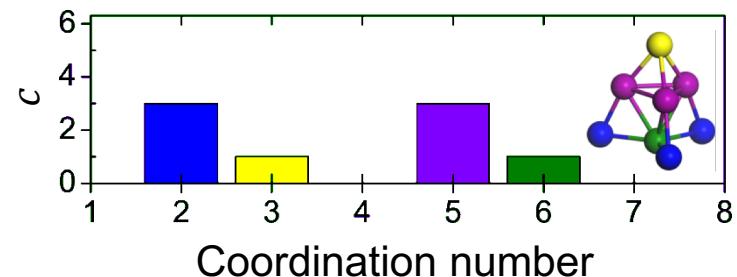
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Targets

$$y_1 = \Delta E, y_2 = \Delta \eta \text{ chem. hardness}$$

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



# Application 2: Au structure/property relationship

24

## Population

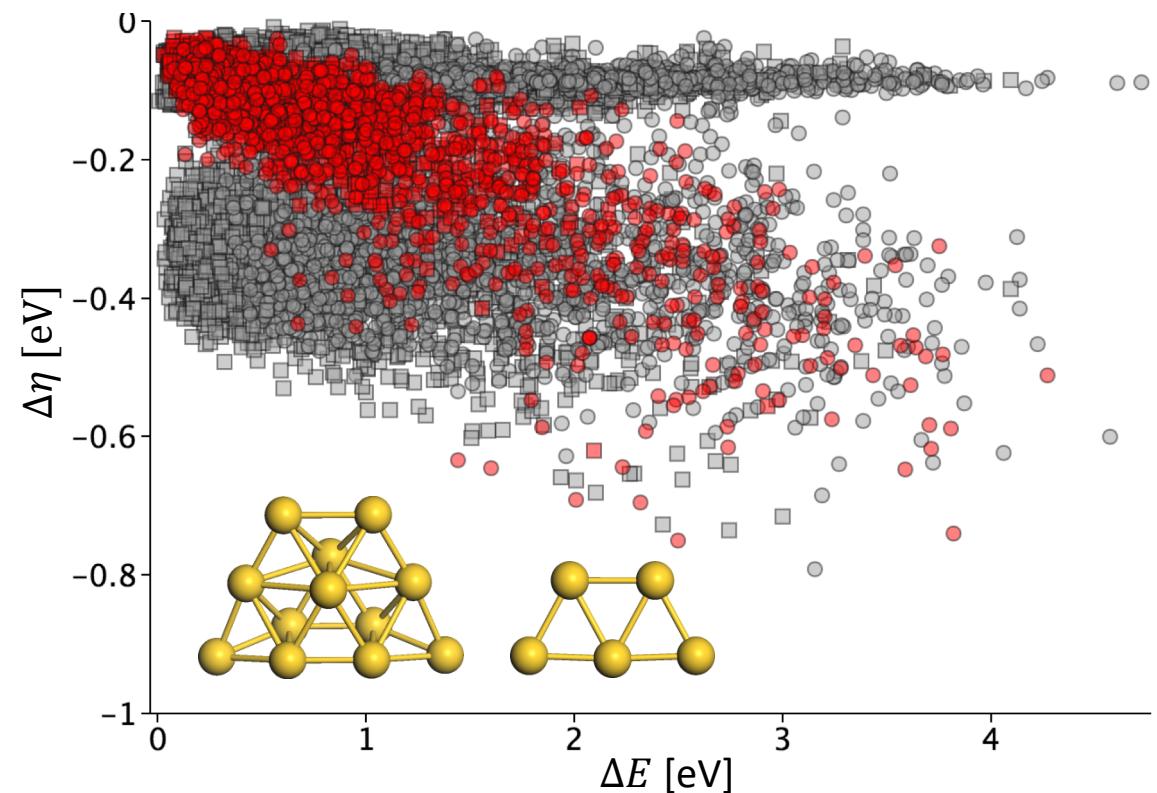
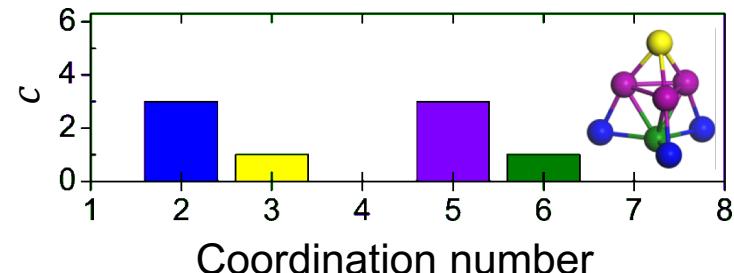
$$P = \{c: c \text{ conf. of Au5} - \text{Au14}\}$$

## Targets

$$y_1 = \Delta E, y_2 = \Delta \eta \text{ chem. hardness}$$

## Features

$$x \in \{a, c_1, c_2, c_3, c_4, c_5, c_6, r, \text{shape}, \text{Mo}_{\text{co}}, \text{Me}_{\text{co}}\}$$



## Selector

$$\sigma(i) \equiv \text{even}(a(i)) \wedge (c_5(i) \leq 0.24) \wedge (\Delta E_{\text{vdw}}(i) \leq 0.18)$$

## Parameters

$$\text{cov}(\sigma) = 0.2 \quad \text{eff}(\sigma) = 0.74 \quad [r(Q) = -0.81, r(S) = -0.27]$$

## Topics

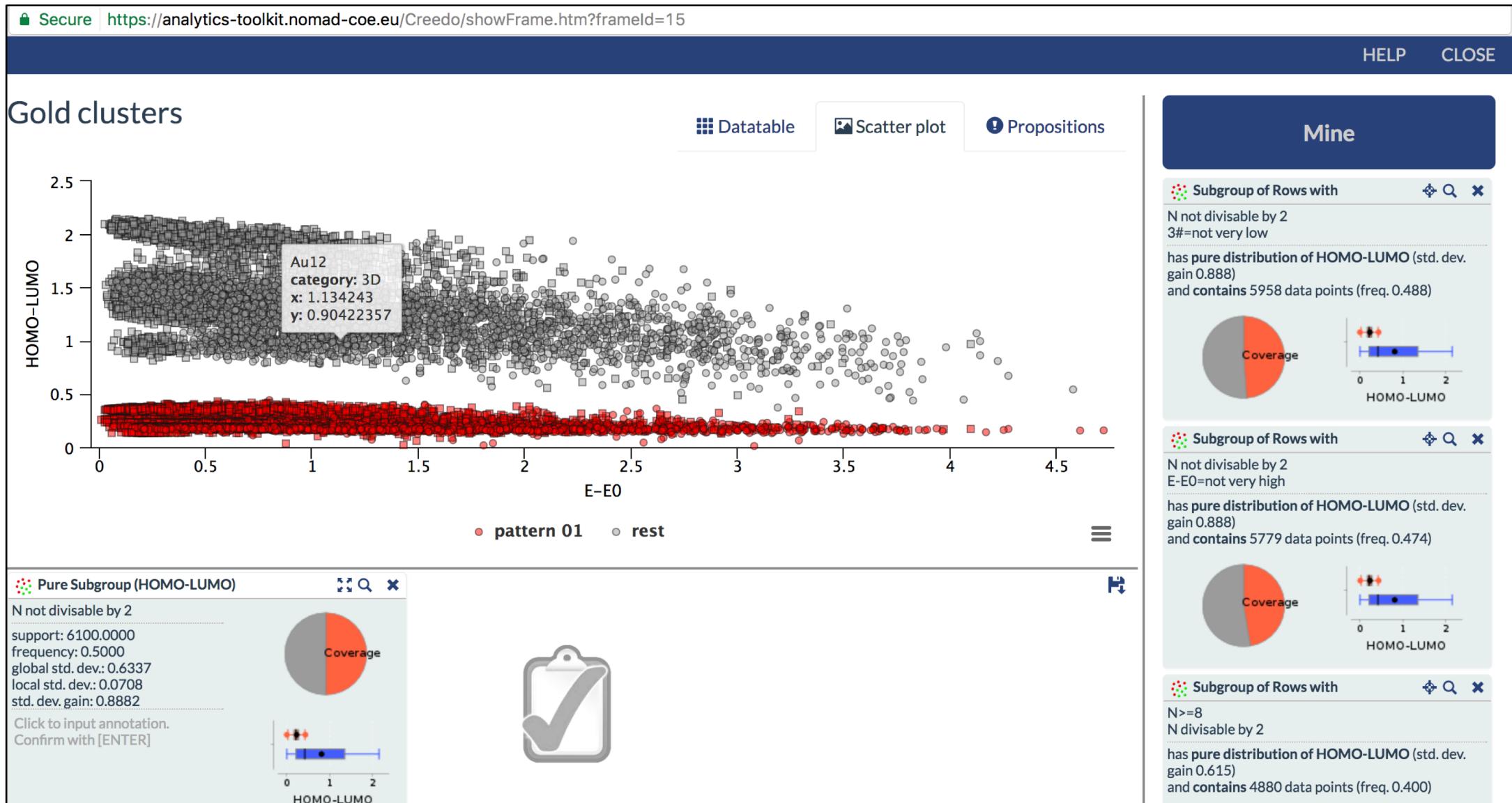
- Basic concepts: selectors, extensions, objective functions
- Application I: octet binary crystal structures
- A glimpse beyond: numeric and multiple targets
- Application II: Au structure/property relationship

## References

- **Boley, [www.realkd.org](http://www.realkd.org): The power of saying ‘I don’t know’**
- Atzmueller, WIREs Data Mining Knowl Discov, 2015: Subgroup discovery – advanced review
- Friedman and Fisher, Stat Comput, 1999: Bump hunting in high-dimensional data
- **Goldsmith et al., New J Phys, 2017: Uncovering structure-property relationships by subgroup discovery**
- Boley et al., Data Min Knowl Disc, 2017: Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery

# Try it out for yourself

26



# Try it out for yourself

27

Secure | <https://analytics-toolkit.nomad-coe.eu/Creedo/showFrame.htm?frameId=15>

HELP CLOSE

## Crystal structure prediction

On-the-fly data analysis for the NOMAD Archive

Predicting energy differences between crystal structures

Tutorial on compressed sensing for materials property prediction

Discovering simple descriptors for crystal-structure classification

Evaluating the (dis)similarity of crystalline, disordered, and molecular compounds

Building structure maps for crystal-structure classification



and contains 4880 data points (freq. 0.400)