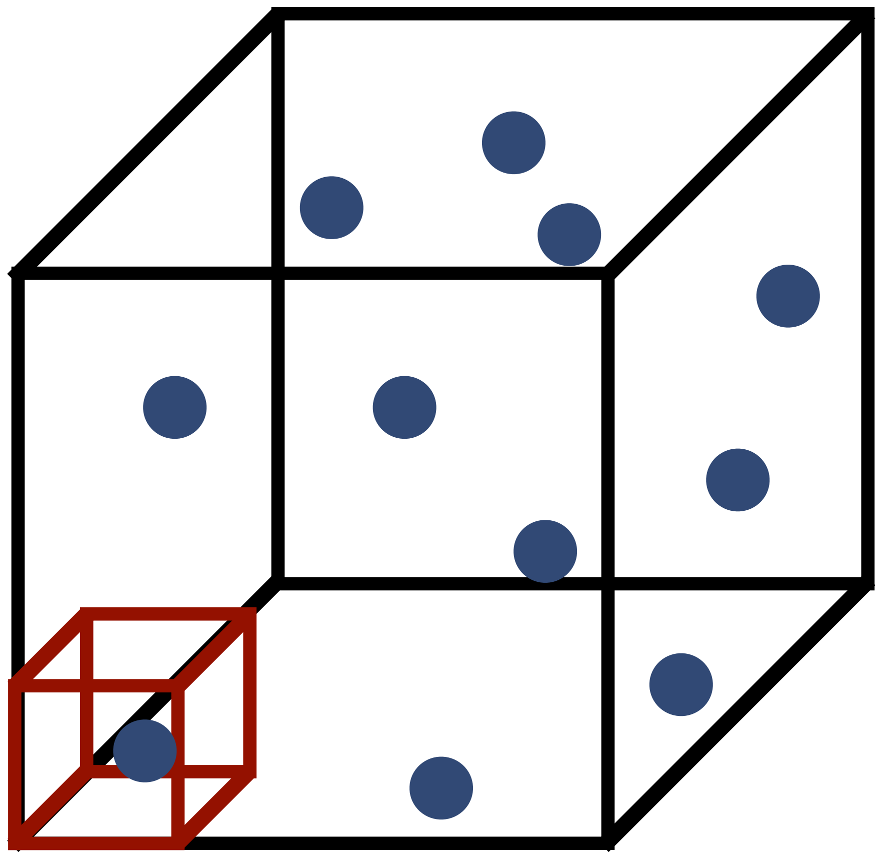
Feature Selection

Key points:

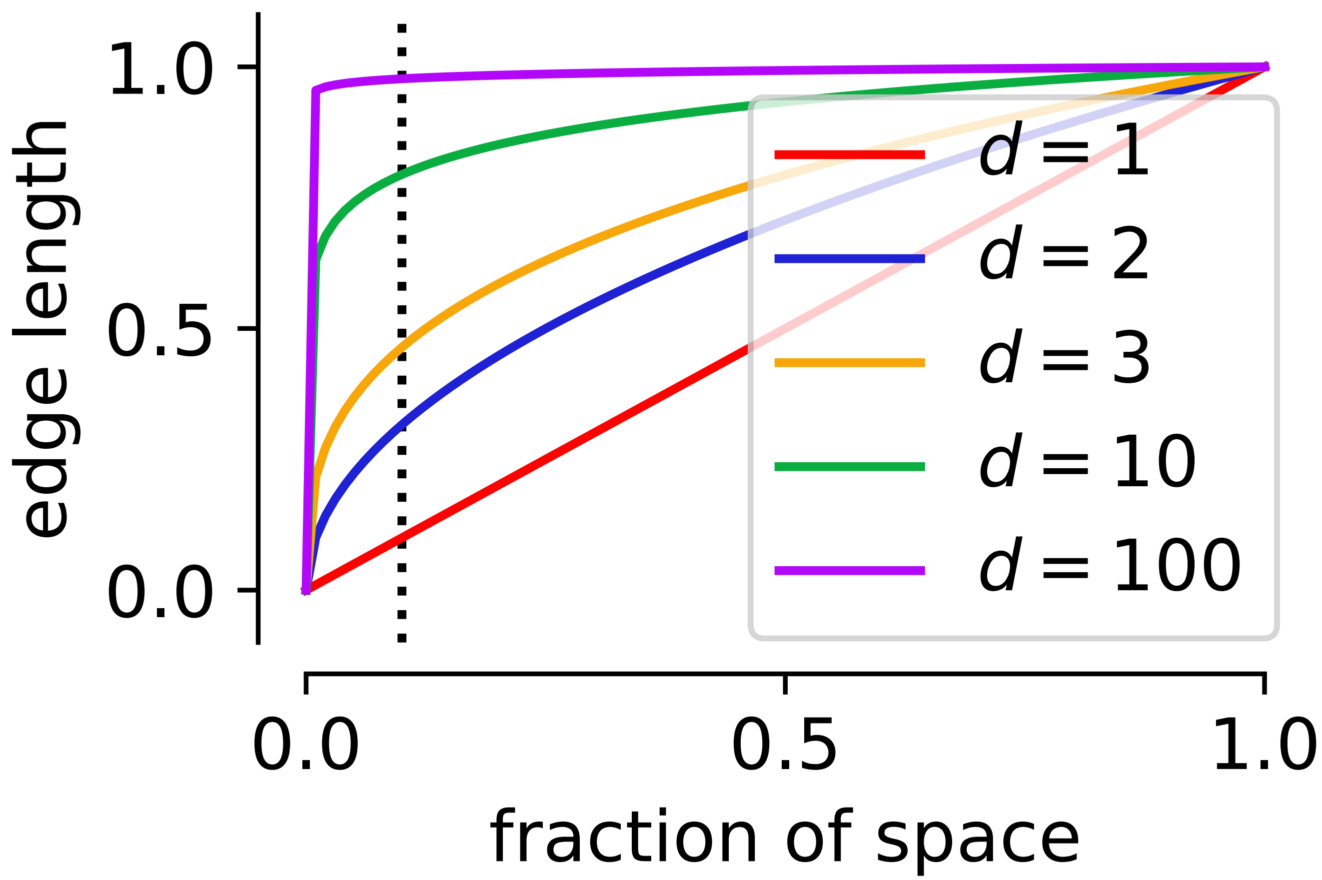
- Why do we do this?
- Some techniques that we can do for that

The Curse of Dimensionality - No Locality in High Dimensions

$$\text{edgelength} = \text{fraction of space}^{1/\text{dimensionality}}$$



unit cube



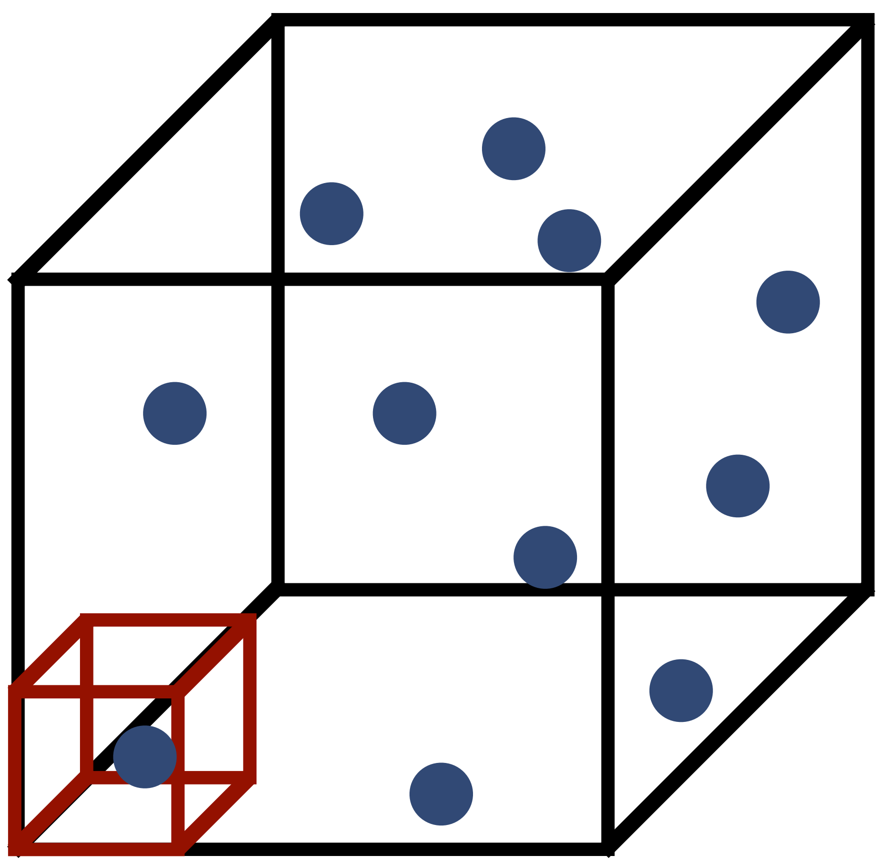
Methods that are based on similarity (kNN/KRR) might fail in high dimensional spaces!

We want to **reduce the dimensionality** of our feature matrix!

Domingos, P. A Few Useful Things to Know about Machine Learning. *Communications of the ACM* **2012**, *55* (10), 78..

The Curse of Dimensionality - No Locality in High Dimensions

$$\text{edgelength} = \text{fraction of space}^{1/\text{dimensionality}}$$



unit cube

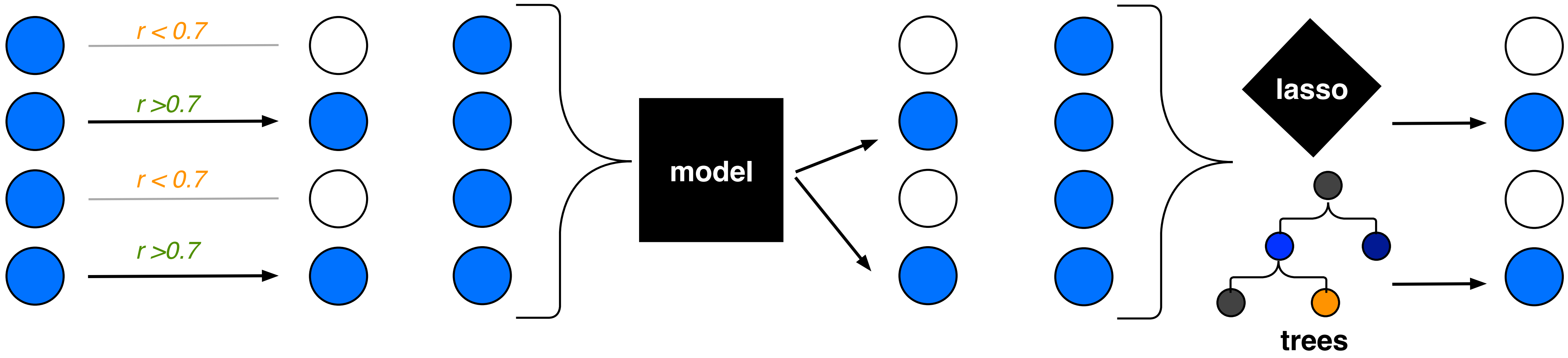
Methods that are based on similarity (*k*NN/KRR) might fail in high dimensional spaces!

We want to **reduce the dimensionality** of our feature matrix!

Domingos, P. A Few Useful Things to Know about Machine Learning. *Communications of the ACM* **2012**, *55* (10), 78..

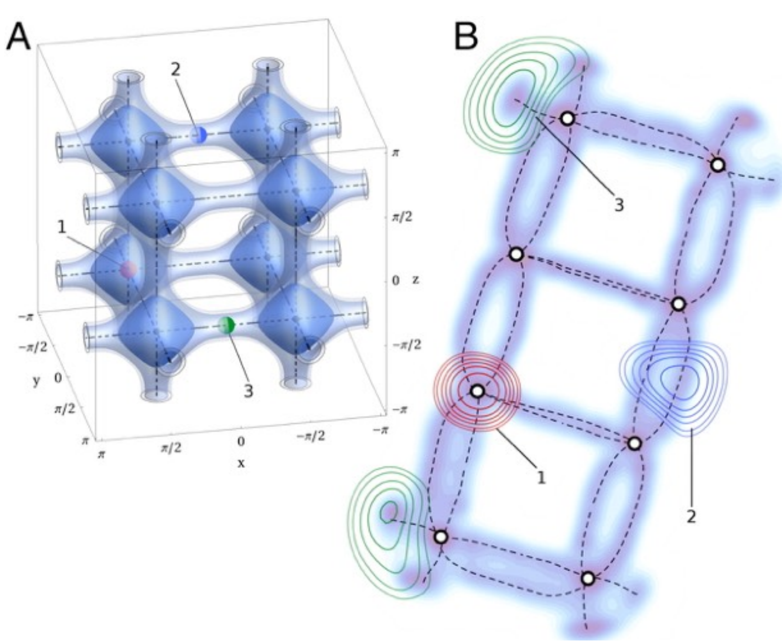
Feature Projection and Feature Selection

Reduce dimensionality of feature space by feature selection (compression)

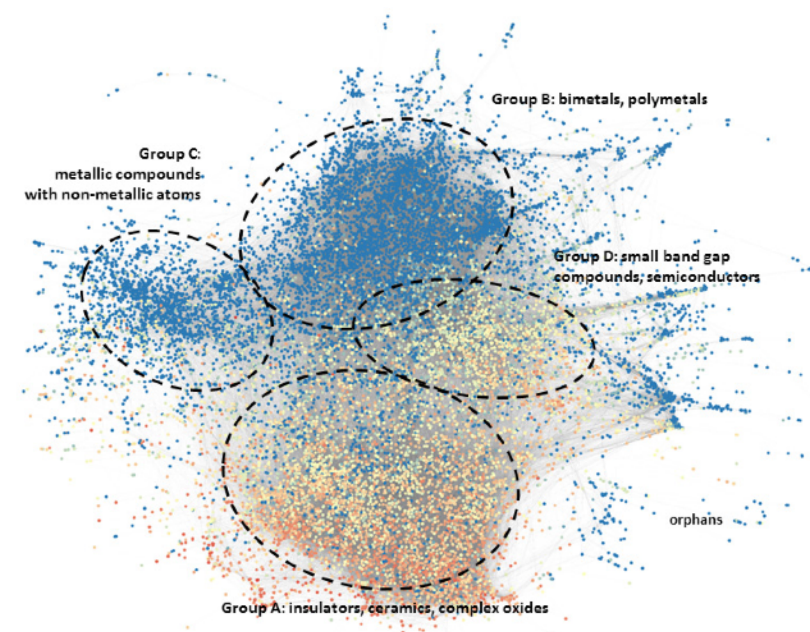


a Univariate filters. **b** Wrapper methods. **c** Shrinkage or direct.

Reduce size of feature space by dimensionality reduction (feature projection)



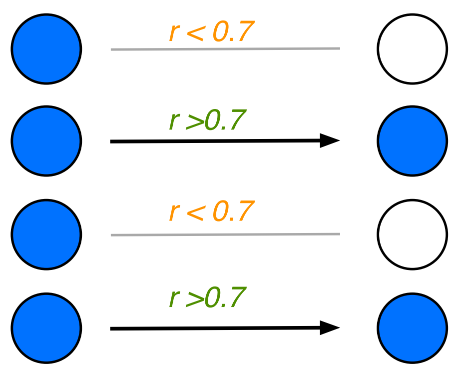
Visualize data and Materials Cartography



Janet, J. P.; Kulik, H. J. *J. Phys. Chem. A* **2017**, *121* (46), 8939–8954.

Tribello, G. A.; Ceriotti, M.; Parrinello, M. *PNAS* **2012**, *109* (14), 5196–5201.

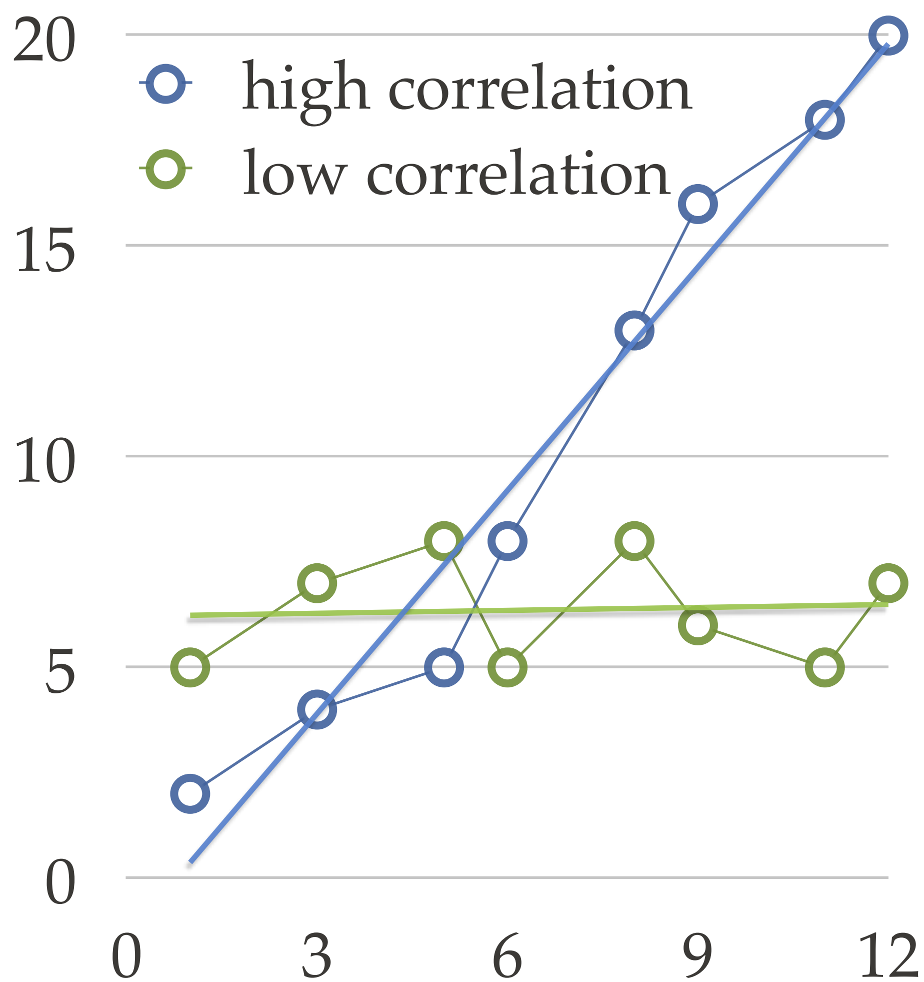
Feature Selection: Filter Methods



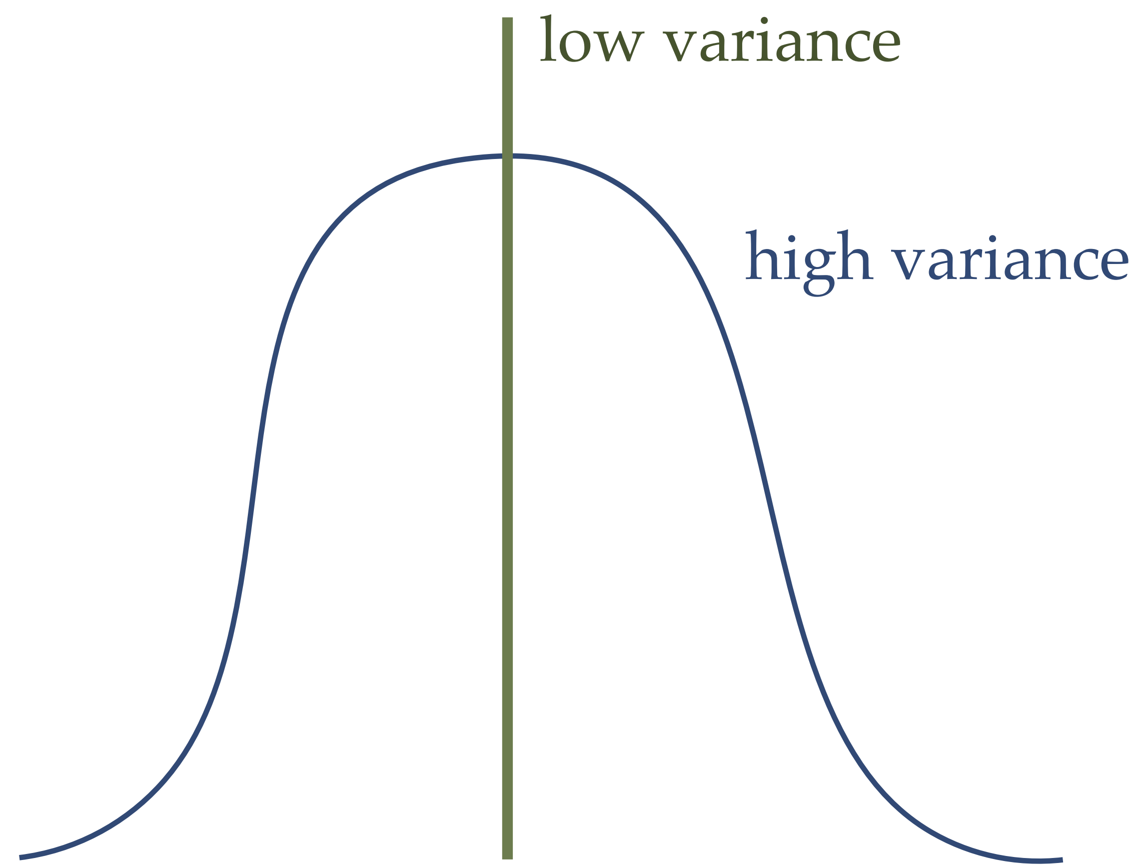
Easy and cheap

Interaction effects are not considered

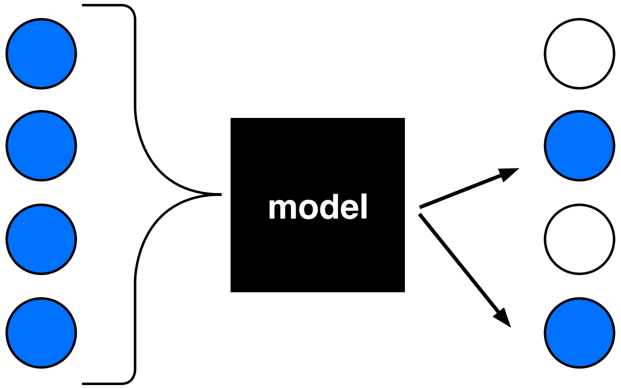
correlation threshold



variance threshold



Feature Selection: Wrapper Methods



Uses a good surrogate of the real objective

Expensive

For example recursive feature addition or elimination

$$\begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \cdots & x_1^p \\ x_2^1 & x_2^2 & x_2^3 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & x_n^3 & \cdots & x_n^p \end{pmatrix} \xrightarrow{\text{subset generation}} \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 \\ x_2^1 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & x_n^3 \end{pmatrix} \rightarrow \boxed{\text{subset evaluation}}$$

full design matrix

not fulfilled

$\boxed{\text{stopping criterion}}$

Feature Selection: Just Relax Best Subset Selection

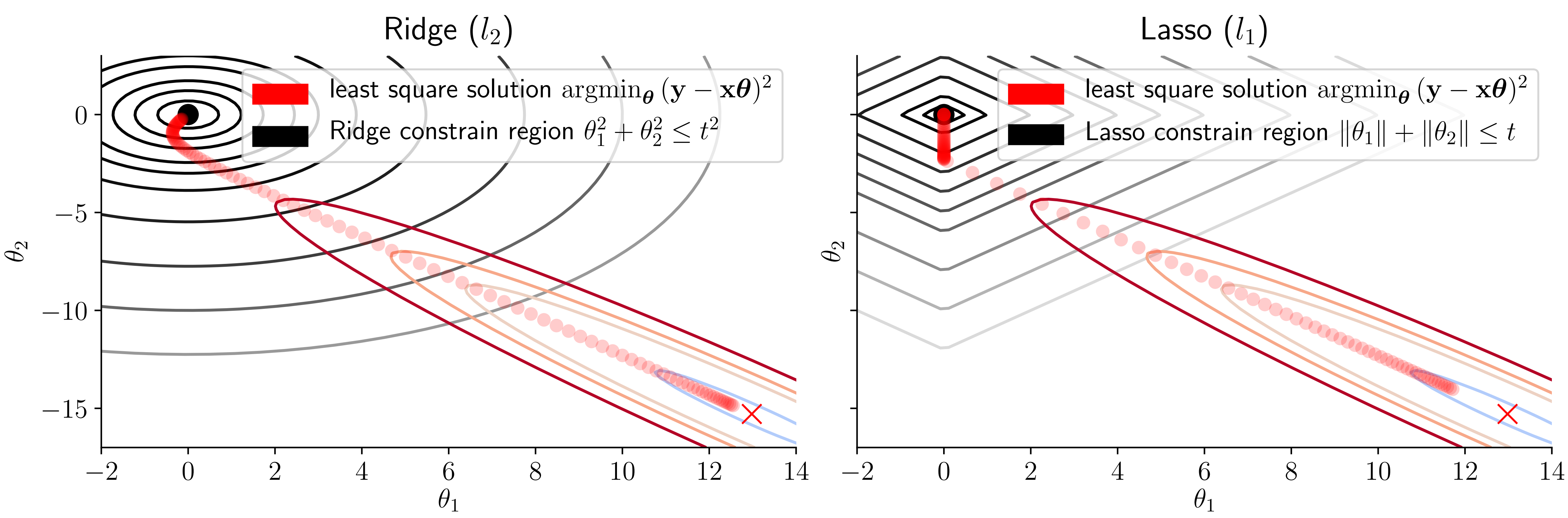
The basic problem: Best subset selection

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|^2 \text{ subject to } \quad \|\beta\|_0 \leq k \qquad \|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$$

But this is our hard problem (NP hard) …

… hence we relax the constraint to have a problem that is convex

… the Lasso gives use sparsity as the most feasible approximation to l_0



Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; Monographs on statistics and applied probability; CRC Press, Taylor & Francis Group: Boca Raton, 2015.

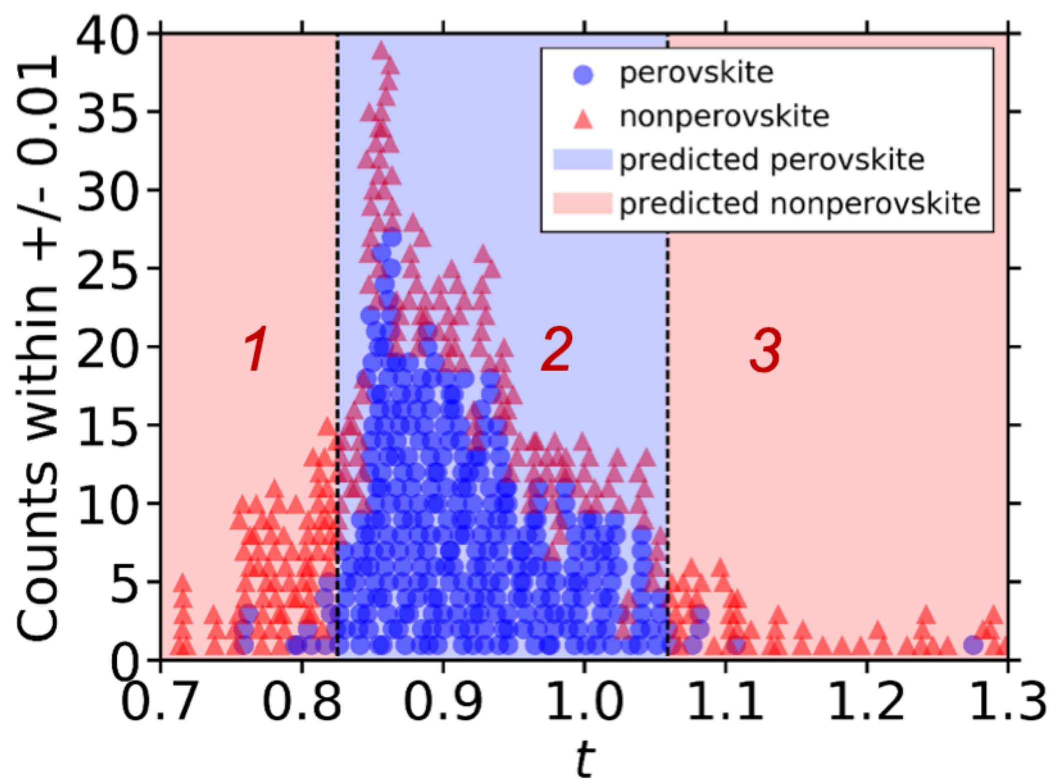
Lasso in Practice: Finding New Tolerance Factors For Perovskites (Developing Causal Models)

primary features $\xrightarrow{\hat{R}}$ **billions** of feature candidates $\xrightarrow{\text{(SI) +Lasso}}$ subset of features

$$\text{primary features} = \{r_A, r_B, n_a, n_b, \dots\} \quad \hat{R} = \{+, -, \cdot, \exp, \lg, {}^{-1}, {}^{2}, {}^{3}, \sqrt{\ }, \|\cdot\|\}$$

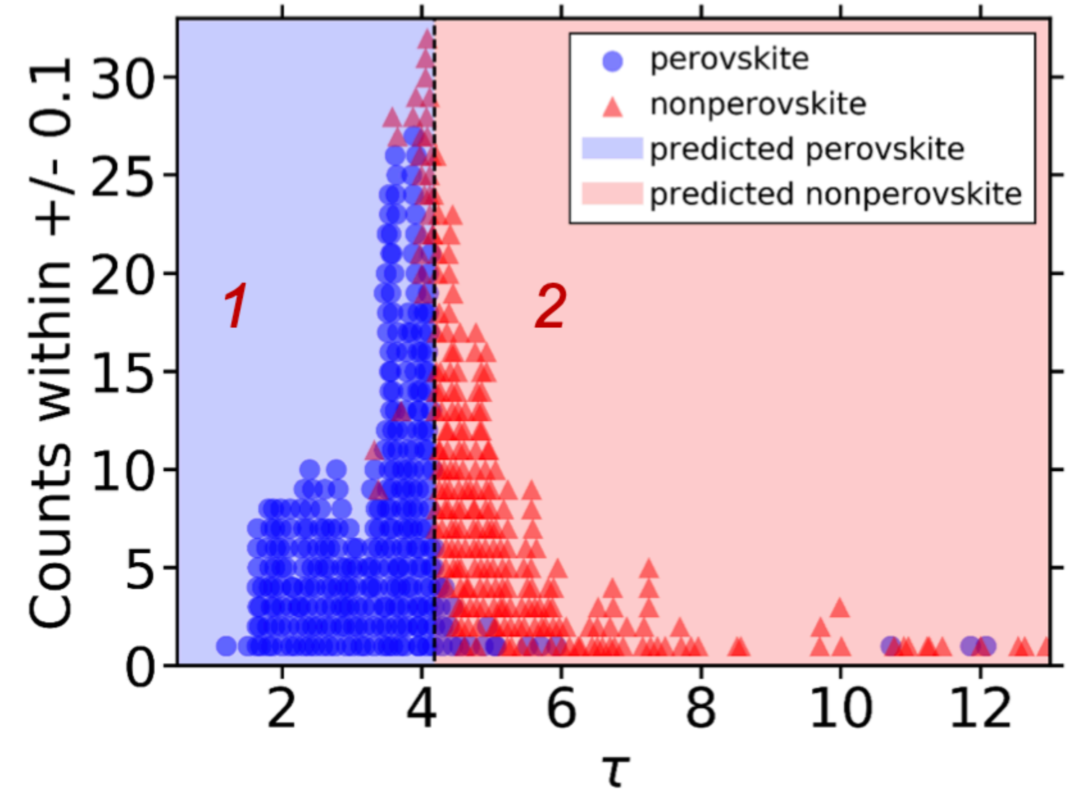
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}$$

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln r_A/r_B}\right)$$



0.825 < t < 1.059 → perovskite

74% accuracy



τ < 4.18 → perovskite

92% accuracy

Ghiringhelli, *et al*. *Phys*. *Rev*. *Lett*. **2015**, *114* (10), 105503.
Ouyang, R.; *et al*. *Phys*. *Rev*. *Materials* **2018**, *2* (8), 083802.
Bartel, C.; *et al*., M. *Sci*. *Adv*. **2019**, *5* (2), eaav0693.

Feature Selection

Key points:

- Why do we do this?
  - *Curse of dimensionality*
- Some techniques that we can use for that
  - *Filter, Wrapper, Direct*
  - *Difference between selection and projection*