Summary of key points (so far)

1) **General supervised learning:**

$\hat{y} = f(\vec{x}) \approx y \ldots \quad \downarrow \mathcal{L}(y, \hat{y})$

2) **Success depends on:**

quantity & quality of data & features

3) **Validation is essential:**

CV to tune hyperparameters (IID)

↳ test set to further probe generalizability.

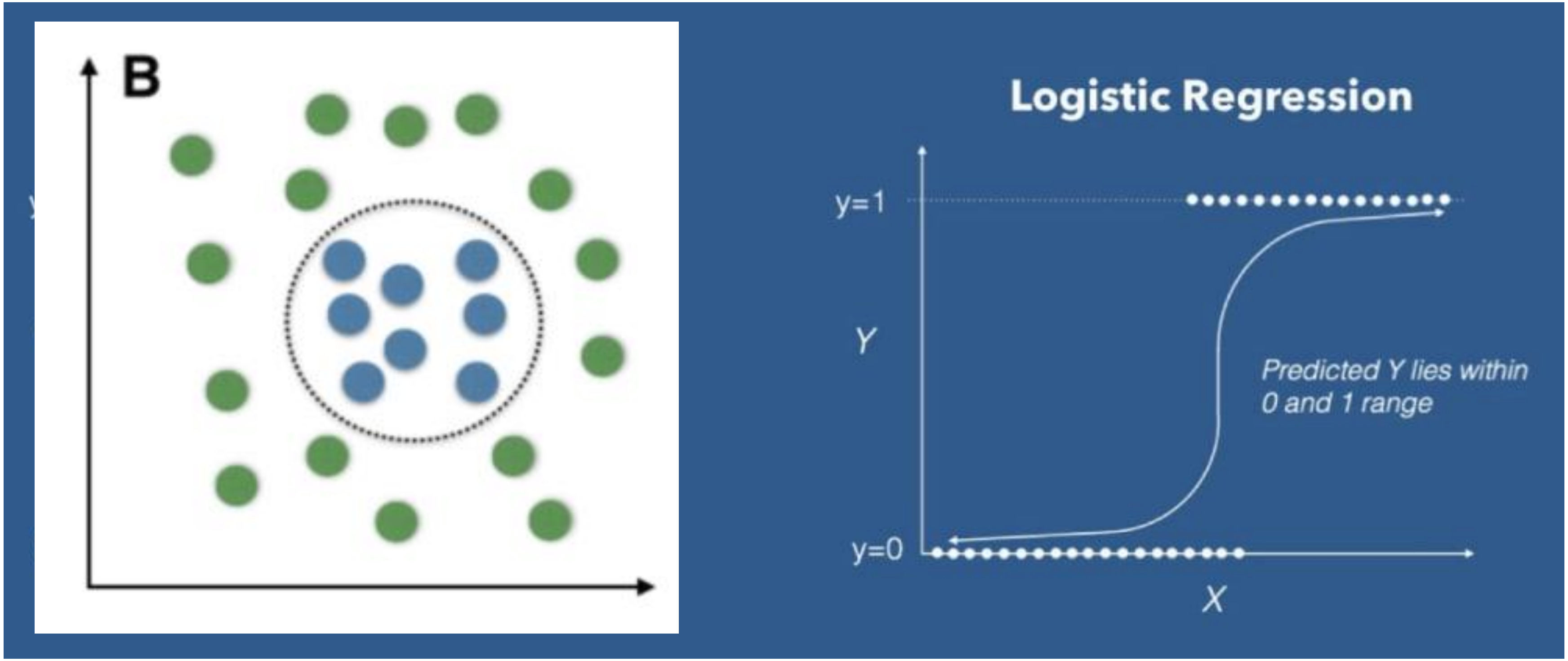
4) **Leverage your domain knowledge:**

better features, more robust data, define success

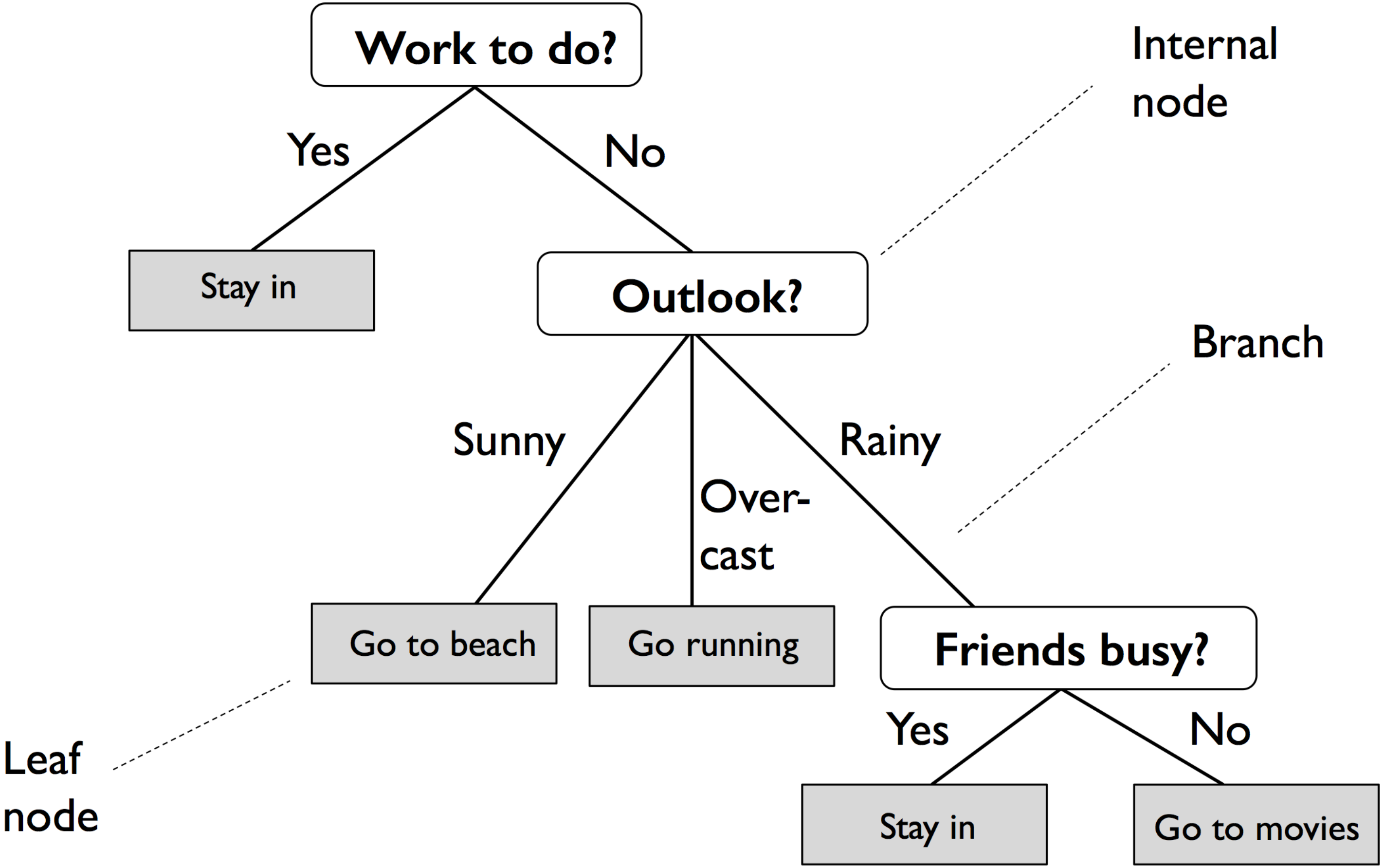
What about nonlinear classification?

For binary classification, y is no longer continuous, but binomial:

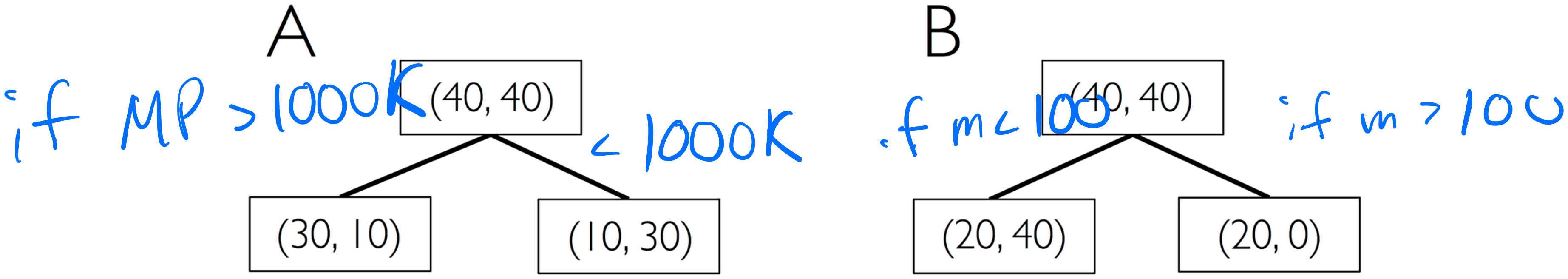
y = [1, 1, 1, -1, -1, 1, -1, -1, ...]



Brief intro to decision trees

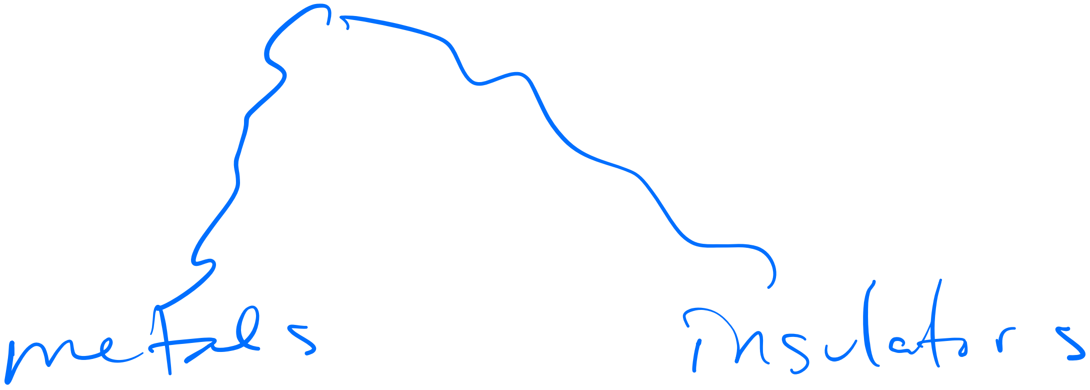


Which split is better?



40 metals

40 insulators



Brief intro to decision trees

Determine splits by **maximizing information gain (IG)**

minimizing weighted impurity, I

$$I(n) = -\sum_{i=0}^{c} p(i|n)\ln[p(i|n)]$$

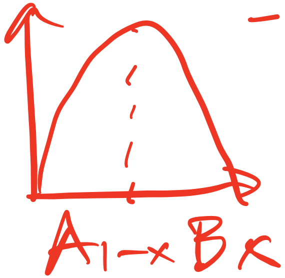
$c=0 \quad c=1$

Consider $n=0 \quad (40, 40)$

$$I_0 = -\left[\left(\frac{1}{2}\ln\frac{1}{2}\right) + \left(\frac{1}{2}\ln\frac{1}{2}\right)\right]$$

$$= 0.69$$

$\frac{S}{k_B}$



$$-\frac{S}{k_B} = x\ln x + (1-x)\ln(1-x)$$

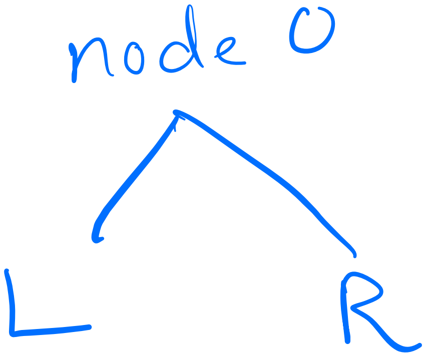
Brief intro to decision trees

Determine splits by **maximizing information gain (IG)**

minimizing weighted impurity, I

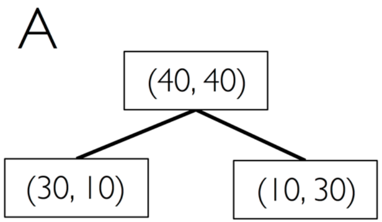
IG = how effectively we ↓ entropy

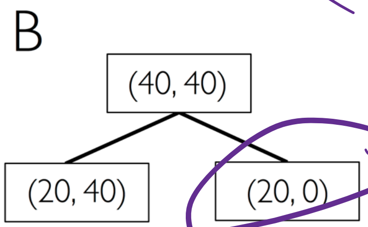
$$IG = I_0 - \left(\frac{N_L}{N} I_L + \frac{N_R}{N} I_R\right)$$



Brief intro to decision trees

$I_0 = 0.69$





$\sim [(1,0)]$

A) $I_L = -\left[\frac{3}{4}\ln\frac{3}{4} + \frac{1}{4}\ln\frac{1}{4}\right] = 0.56$

$I_R = I_L = 0.56$

$IG = 0.69 - \frac{1}{2}0.56 - \frac{1}{2}0.56 = 0.13$

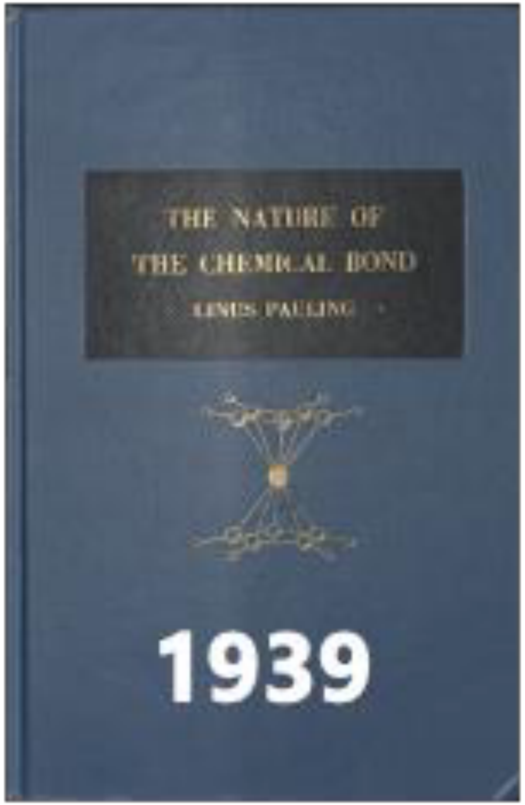
B) $I_L = -\left[\frac{1}{3}\ln\frac{1}{3} + \frac{2}{3}\ln\frac{2}{3}\right] = 0.64$

$I_R = 0$

$IG = 0.69 - \boxed{\frac{3}{4}0.64} - \frac{1}{4}0 = 0.21$ ✓

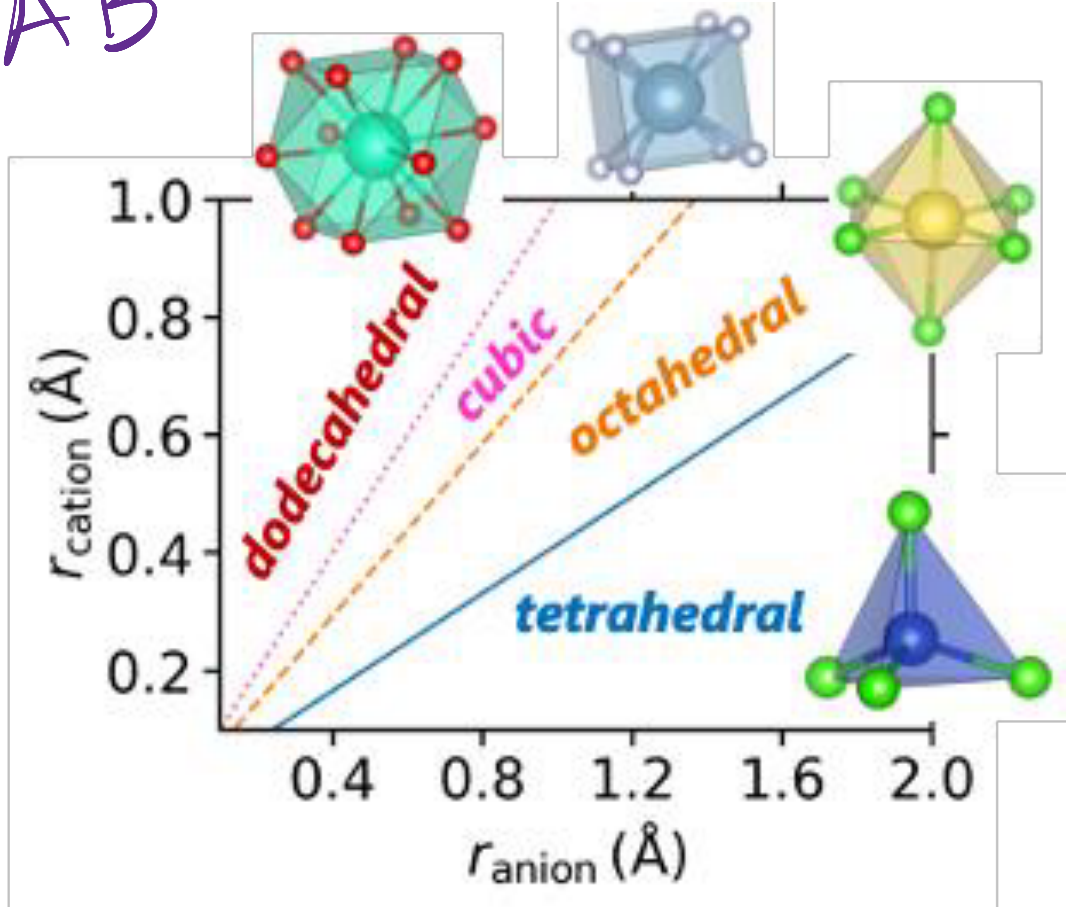
Classifying crystal structures

A B

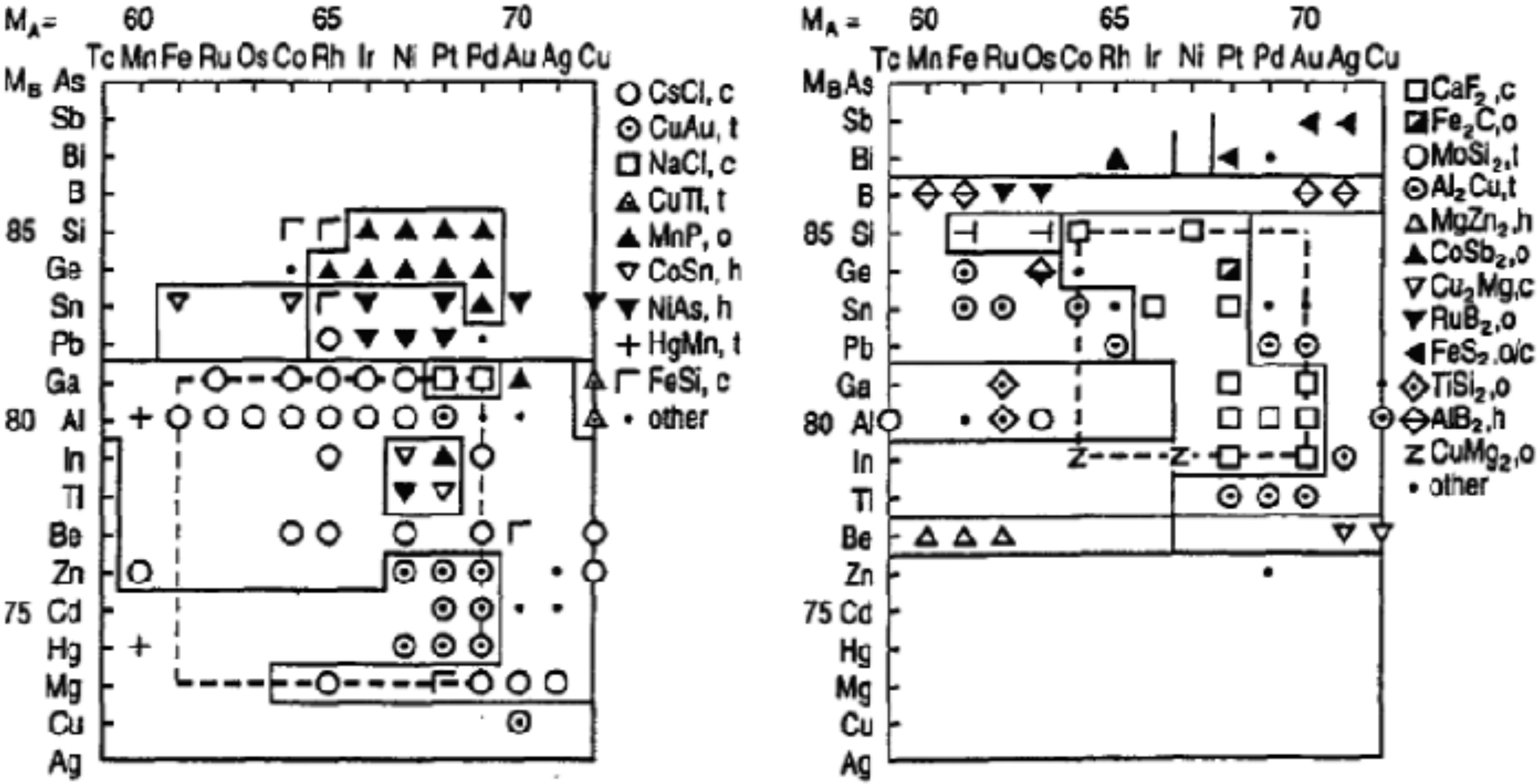


THE NATURE OF
THE CHEMICAL BOND
LINUS PAULING

1939



Classifying crystal structures

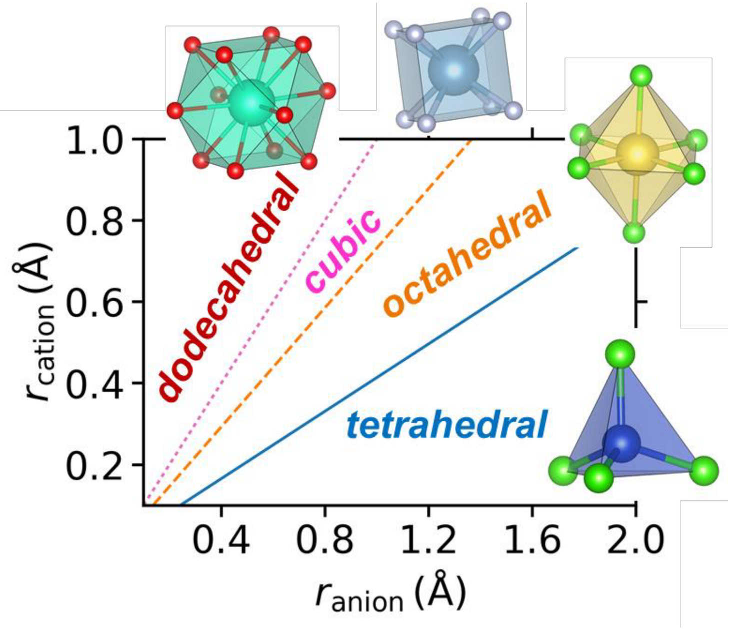


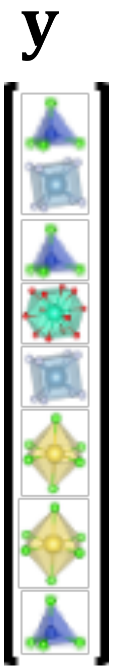
Pettifor, 1980

Finding simple models w/ supervised ML

y – *target property (observable)*

y *– data you find or generate*

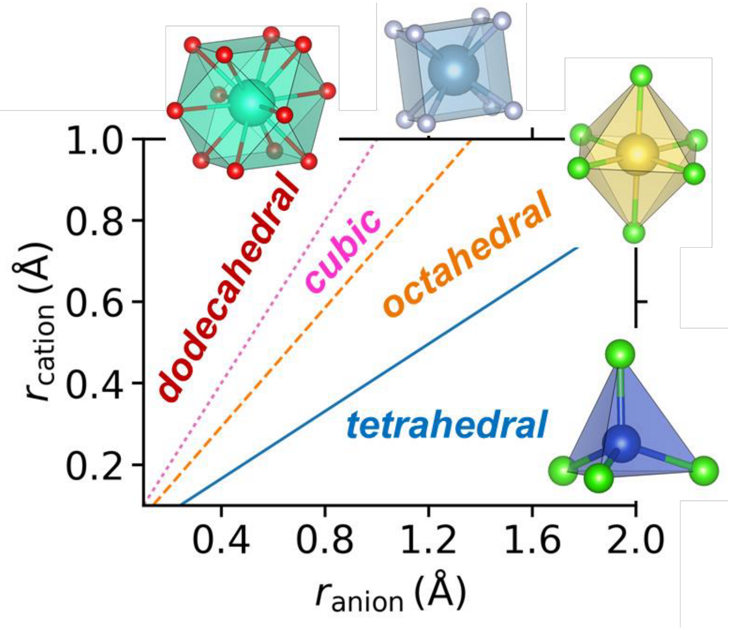




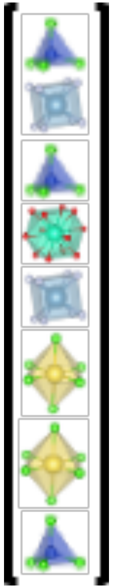
Finding simple models w/ supervised ML

y* – *target property (observable)
X* – *feature space (representation)

y *– data you find or generate*
X *– stuff you hope relates to* ***y***



y



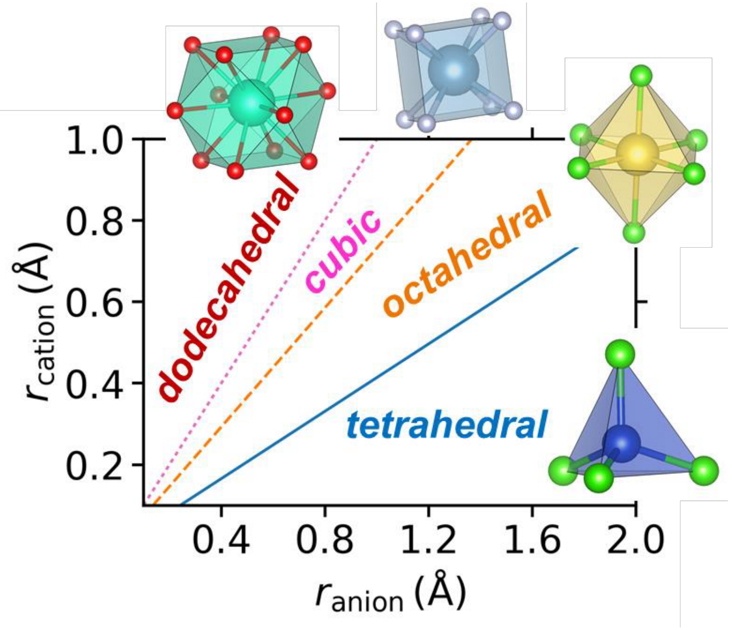
X

$$\begin{array}{cc} r_{cation} & r_{anion} \\ \left[\begin{array}{cc} 0.5 & 1.7 \\ 0.7 & 1.1 \\ 0.3 & 1.2 \\ \ldots & \ldots \end{array}\right] \end{array}$$

Finding simple models w/ supervised ML

y – *target property (observable)*
X – *feature space (representation)*
***f*(*X*)** – *model (descriptor)*
ŷ – *prediction (model output)*

y – *data you find or generate*
X – *stuff you hope relates to **y***
f – *the learned mapping of **X** to **y***



y

$$\boldsymbol{X} \quad \left[\begin{array}{cc} r_{cation} & r_{anion} \\ 0.5 & 1.7 \\ 0.7 & 1.1 \\ 0.3 & 1.2 \\ \dots & \dots \end{array}\right]$$

***f*(*X*): min(y-ŷ)**

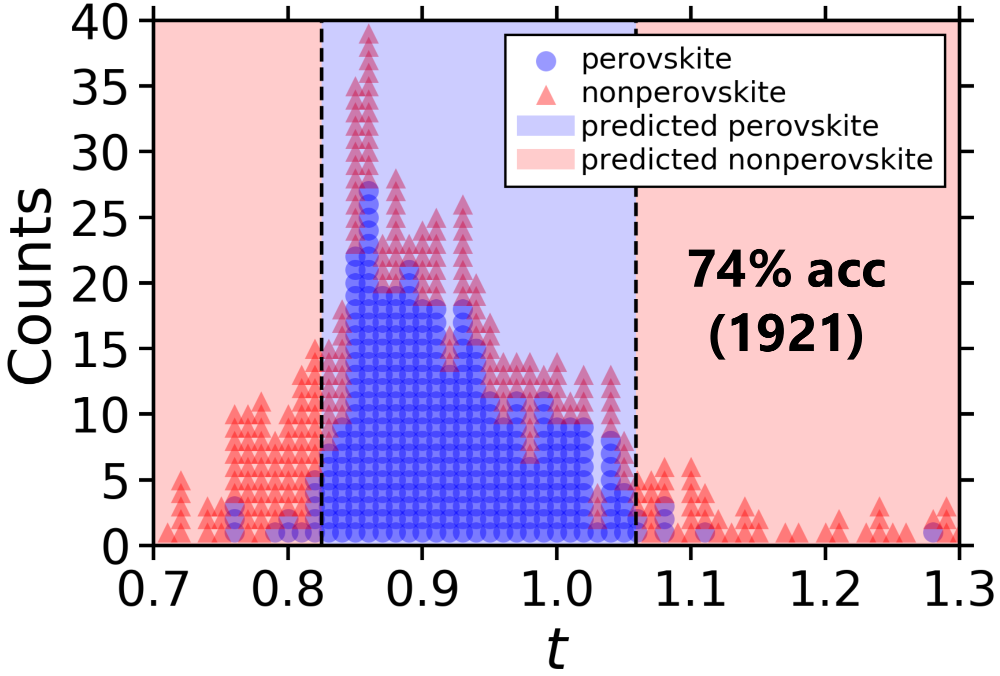
ŷ

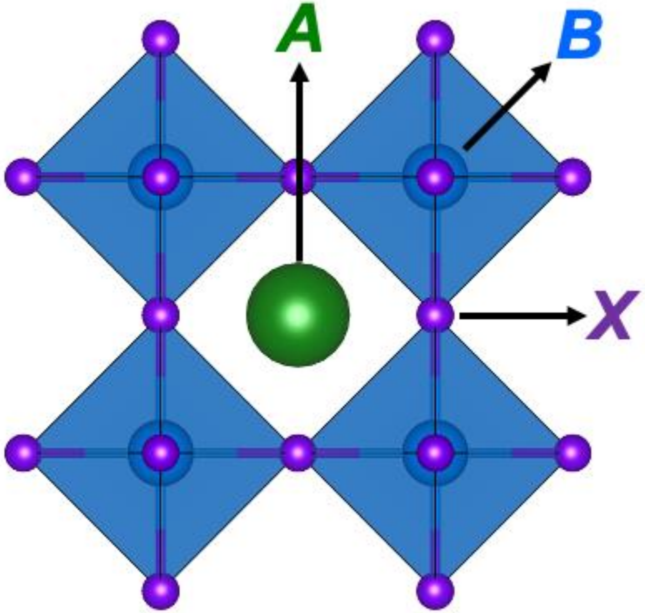
Goldschmidt’s tolerance factor for perovskite stability

CEMS Chemical Engineering & Materials Science

For 576 experimentally characterized ABX_3 compounds

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





Bartel et al., *Science Advances*, 2019, 5, eaav0693

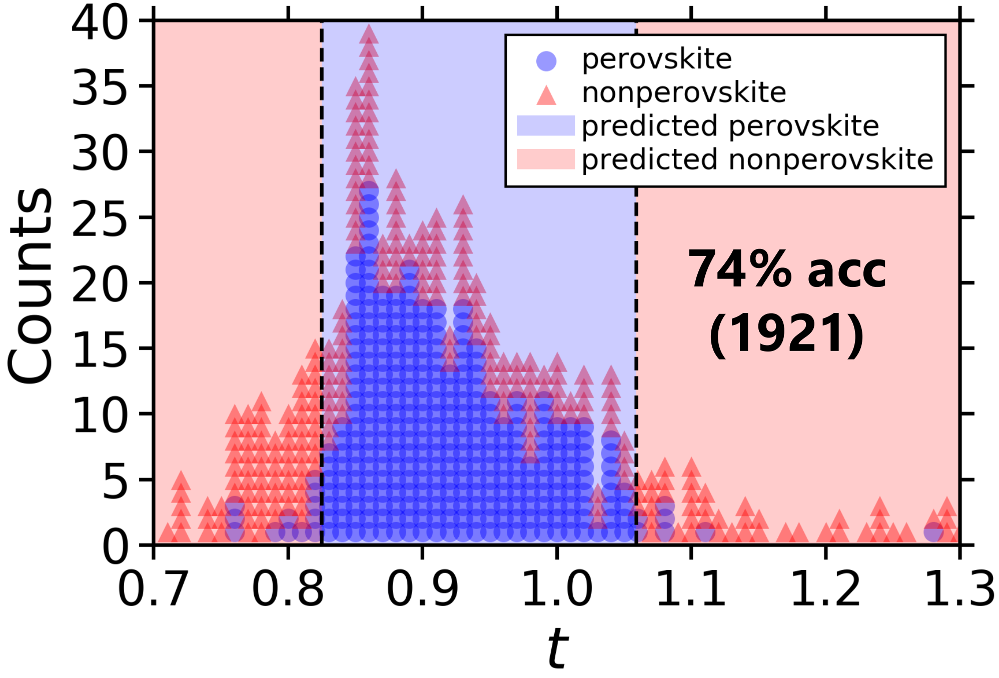
New tolerance factor!

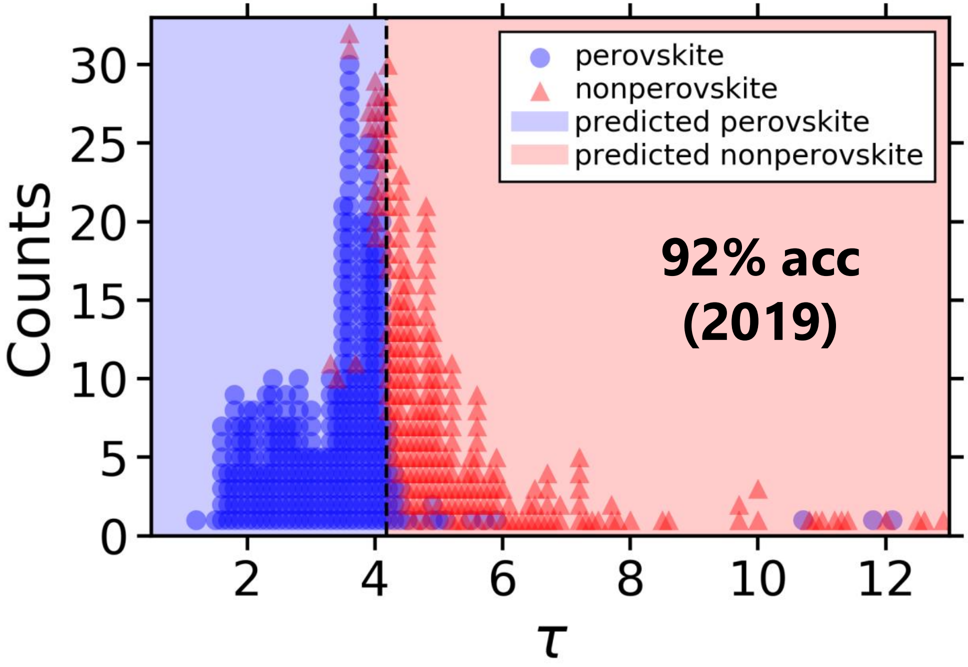
For 576 experimentally characterized *ABX*$_3$ compounds

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

SISSO

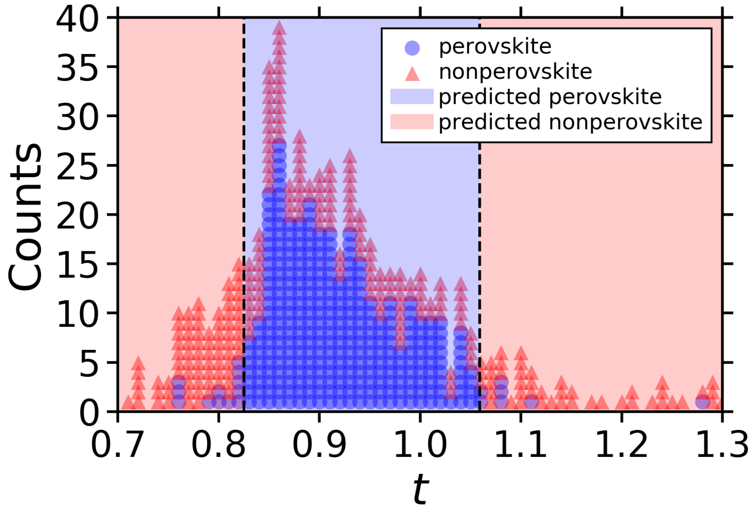
$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln r_A/r_B}\right)$$

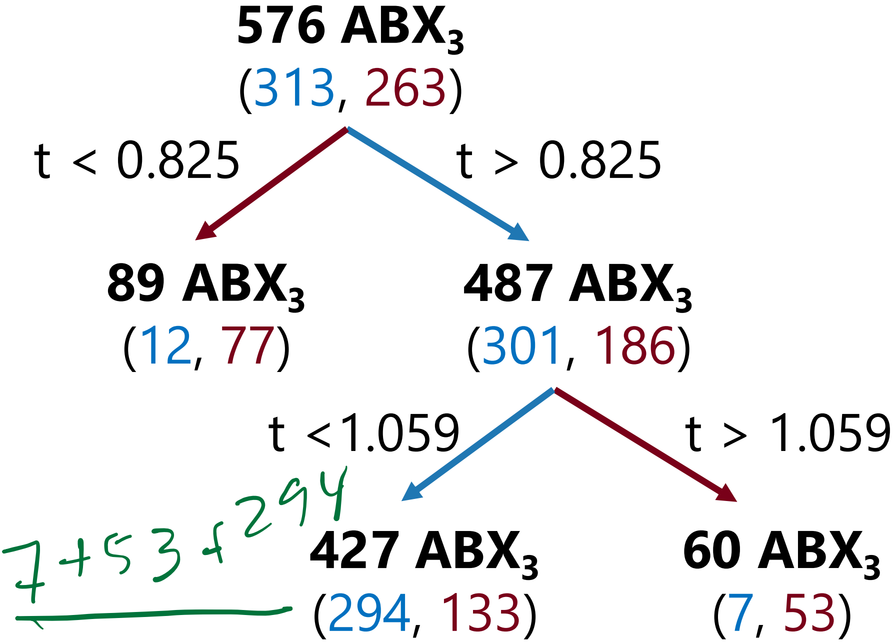




Bartel et al., *Science Advances*, 2019, 5, eaav0693

Decision trees w/ Goldschmidt's t





576 ABX_3
(313, 263)

t < 0.825

t > 0.825

89 ABX_3
(12, 77)

487 ABX_3
(301, 186)

t <1.059

t > 1.059

427 ABX_3
(294, 133)

60 ABX_3
(7, 53)

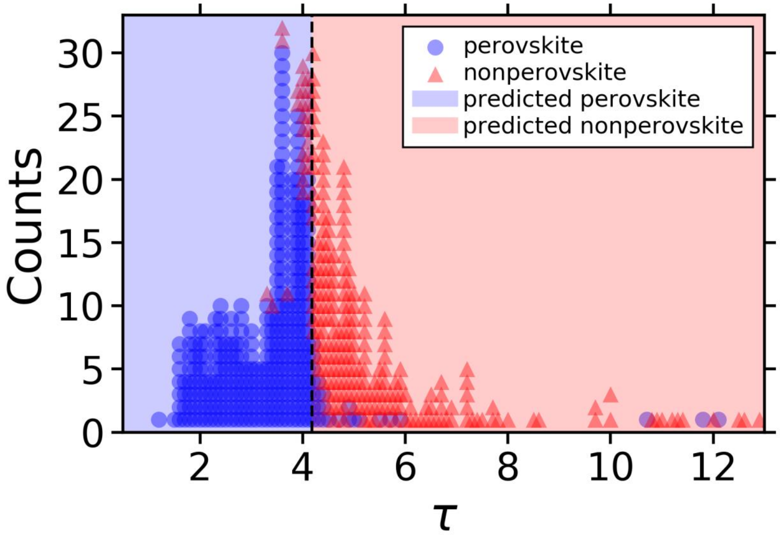
$$
\mathbf{y} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ \ldots \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} t = \dfrac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \end{bmatrix}
$$

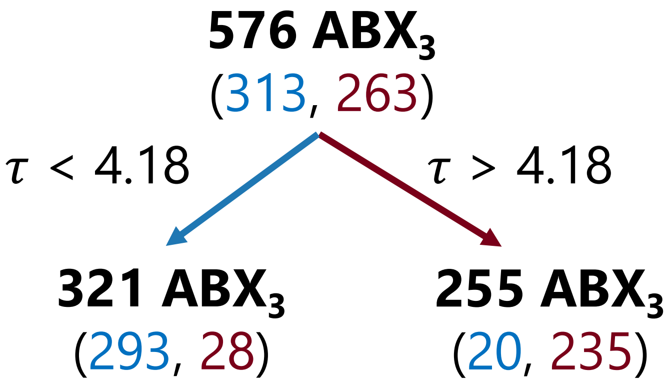
$$
Acc = \frac{77+53+294}{576} = 0.74
$$

$$
FPR = \frac{133}{263} \approx 50\%
$$

Bartel et al., *Science Advances*, 2019, 5, eaav0693

Decision trees w/ τ




576 ABX_3
(313, 263)

τ < 4.18

τ > 4.18

321 ABX_3
(293, 28)

255 ABX_3
(20, 235)

$$Acc = \frac{293+235}{576} \approx 0.92$$

y

X

$$\begin{bmatrix} 1 \\ -1 \\ 1 \\ \cdots \end{bmatrix} \quad \begin{bmatrix} \tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln r_A/r_B}\right) \\ \\ \\ \end{bmatrix}$$

$$FPR = \frac{28}{263} \approx 10\%$$

Bartel et al., *Science Advances*, 2019, 5, eaav0693