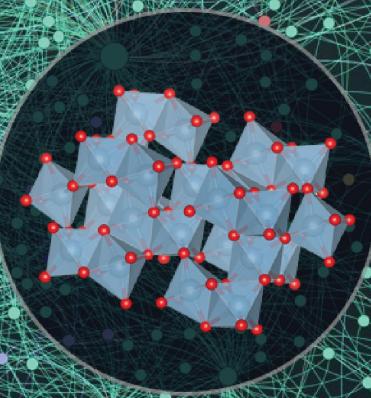
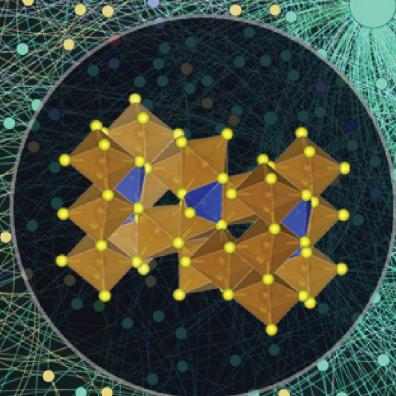


Featurization in Materials Informatics



Featurization in Materials Informatics



How will ML perform tasks if we don't provide instructions?



“The dog has been in the habit of carrying this stick behind his master. Being a heavy stick the dog has held it tightly by the middle, and the marks of his teeth are very plainly visible. The dog's jaw, as shown in the space between these marks, is too broad in my opinion for a terrier and not broad enough for a mastiff. It may have been – yes, by Jove it is a curly-haired spaniel.”

Features should be correlated with the target label

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon.

Choosing **informative**, **discriminating** and **independent** features is a crucial element of effective algorithms in pattern recognition, classification and regression.

Features should be correlated with the target label



- Shoe size (inches)
- Shoe size (cm)
- Age
- Favorite food
- Race
- Shirt Color
- Gender
- Zip Code
- Asleep/Awake
- Genome sequence
- Occupation
- Eye color
- Weight

Not all features are equally important



Shoe size (inches)

Age

Gender

Weight

Race

Genome sequence (highly complex)

Occupation (weak?)

Eye color (weak?)

Zip Code (weak?)

Shoe size (cm) (not independent)

Favorite food (no correlation)

Shirt Color (no correlation)

Asleep/Awake (no correlation)

Some machine learning techniques tell us which features matter the most



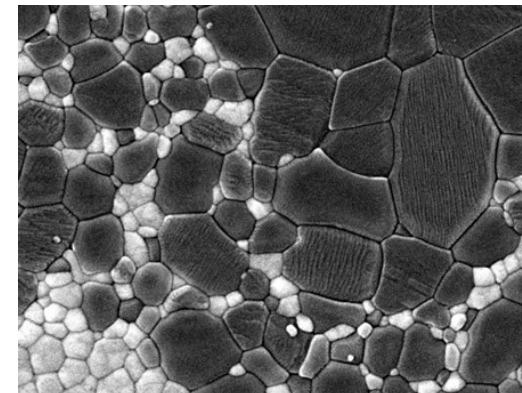
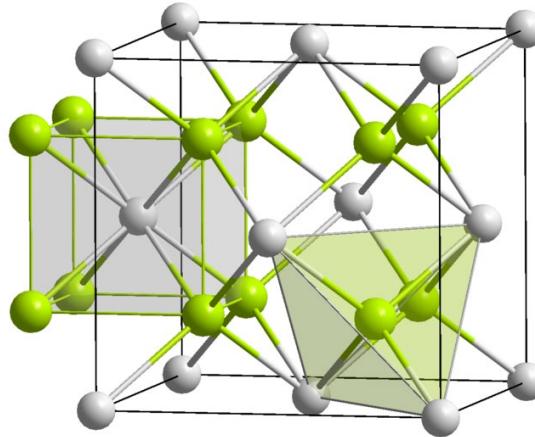
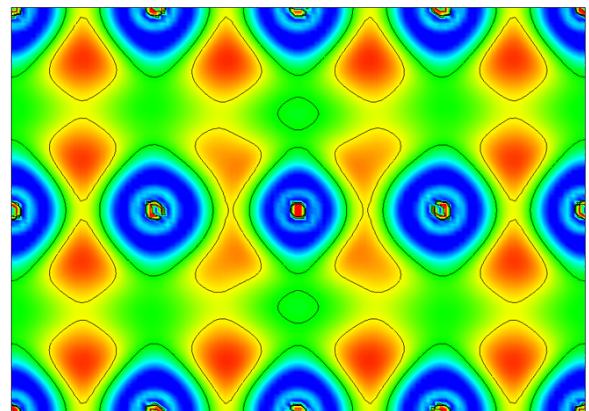
Featurization is different than feature engineering

Featurization is a way to change some form of data (text data, graph data, time-series data,...) into a numerical vector.

Featurization is different from feature engineering. Feature engineering is just transforming the numerical features somehow so that the machine learning models work well. In feature engineering, features are already in the numerical form, whereas in featurization data not need to be in the form of numerical vector.

What sort of materials features are available?

Electronic → Atomic → Microstructure → Macroscale



Composition-based feature vector

