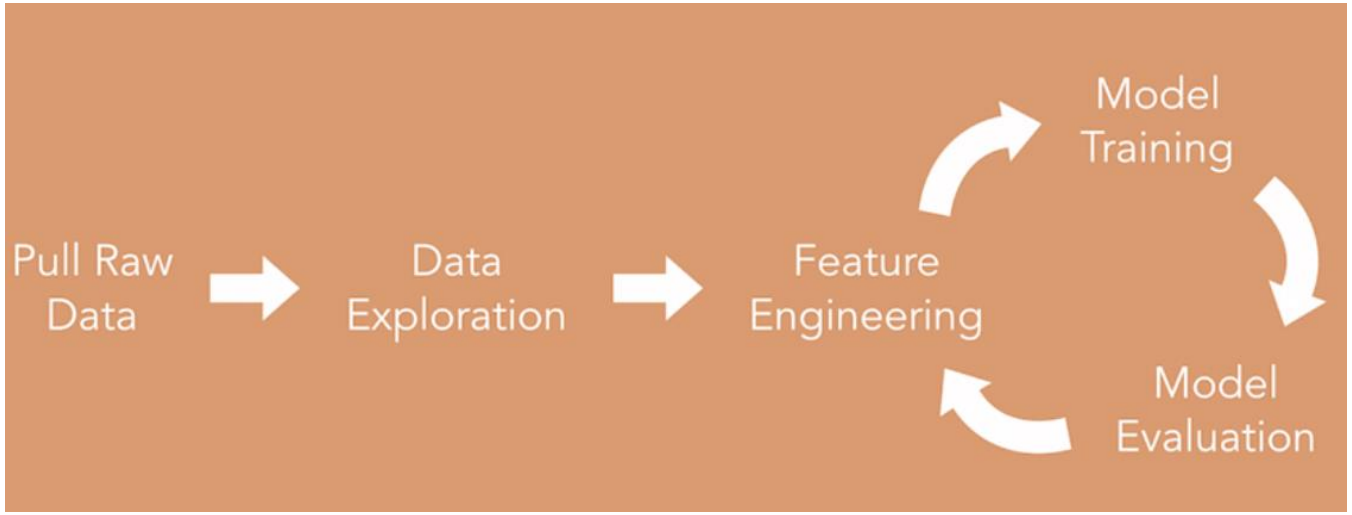
Feature Engineering

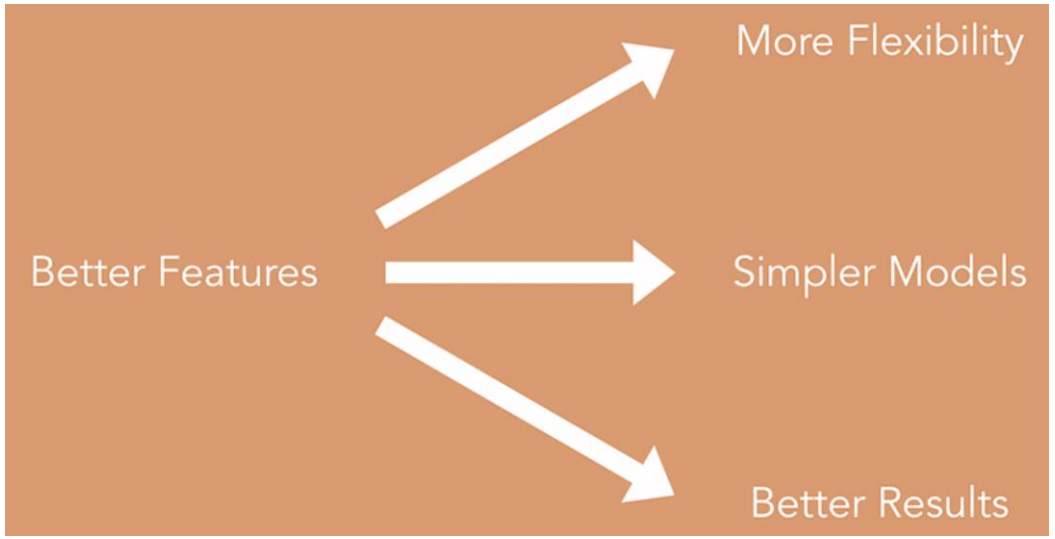
Feature engineering is the process of transforming raw data into features that better represent the underlying signal to be fed to a machine learning model, resulting in improved model accuracy on unseen data.



"Feature engineering is another topic which doesn't seem to merit any review papers or books, or even chapters in books, but it is absolutely vital to ML success. Sometimes the features are obvious; sometimes not. Much of the success of machine learning is actually success in engineering features that a learner can understand."

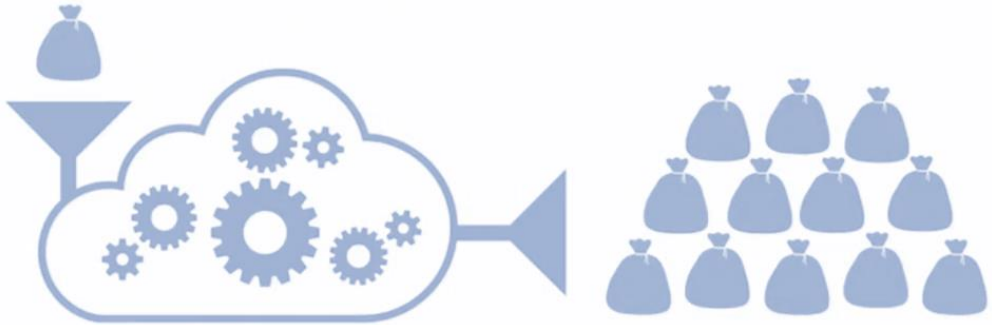
- Scott Locklin

"Neglected Machine Learning Ideas" (2014)



GIGO: Garbage In, Garbage Out

The quality of information coming out cannot be better than the quality of information that went in.



Tools in the Feature Engineering Toolbox

Common sense and domain expertise

- What factors would you expect to influence the thing you are trying to predict?

Dropping irrelevant features (feature selection)

- Including irrelevant features in a model just adds noise and makes it harder for the model to pick up on the true signal

Cleaning existing features

- Impute missing values
- Remove outliers
- Scale the data
- Transform skewed data

Tools in the Feature Engineering Toolbox

Splitting or grouping features

- Combine two related features to make one (for example, combine numbers of cats and number of dogs into number of pets)
- Split one feature into two (for example, extract day of the week from a date)

Binning or creating indicator variables

- Convert a continuous variable into simpler categorical feature (for example, whether a person has ever defaulted on a loan rather than the number of times they have defaulted on a loan)

Learning new features

- Learning word or document embeddings from raw text