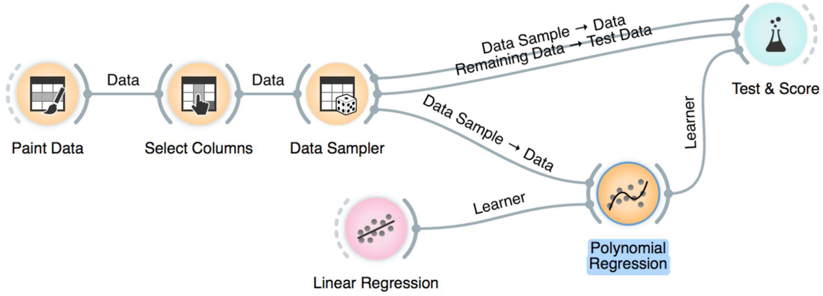
# Lesson 19: Regularization and Accuracy on Test Set

Overfitting hurts. Overfit models fit the training data well, but can perform miserably on new data. Let us observe this effect in regression. We will use hand-painted data set, split it into the training (50%) and test (50%) data set, polynomially expand the training data set to enable overfitting, build a model on it, and test the model on both the (seen) training data and the (unseen) held-out data:



Paint about 20 to 30 data instances. Use attribute y as target variable in Select Columns. Split the data 50:50 in Data Sampler. Cycle between test on train or test data in Test & Score. Use ridge regression to build linear regression model.

Now we can vary the regularization strength in Linear Regression and observe the accuracy in Test & Score. For accuracy scoring, we will use RMSE, root mean squared error, which is computed by observing the error for each data point, squaring it, averaging this across all the data instances, and taking a square root.

The core of this lesson is to compare the error on the training and test set while varying the level of regularization. Remember, regularization controls overfitting - the more we regularize, the less tightly we fit the model to the training data. So for the training set, we expect the error to drop with less regularization and more overfitting, and to increase with more regularization and less fitting. No surprises expected there. But how does this play out on the test set? Which sides minimizes the test-set error? Or is the optimal level of regularization somewhere in between? How do we estimate this level of regularization from the training data alone?

Orange is currently not equipped with parameter fitting and we need to find the optimal level of regularization manually. At this stage, it suffices to say that parameters must be found on the training data set without touching the test data.