Do’s and don’ts when splitting data

**Do**

* find a good balance between test and training data.
* Reduce the dimension of your training data using a dimensionality reduction algorithm

**Don’t**

* **Overfitting the test data**
  + the model works well on the test data, but does not generalize well.
    - This happens if the training set is too small or too noisy.
* **Underfitting the test data**
  + this happens when the model is too simple to learn the underlying structure of data.
    - Linear models are prone to underfit. Most of the time, reality is more complicated than a linear model can handle.
      * can be fixed by selecting more powerful models, feeding better features to the learning algorithm or reducing the constraints on the model.