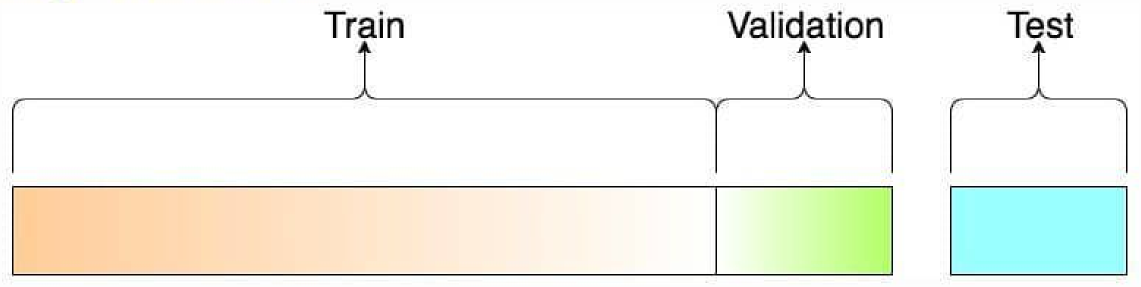
Test, train and validation datasets

- We usually split our data into training, validation and test datasets(for supervised algorithms).
- Training data:- The actual dataset that we use to train the model (weights and biases in the case of Neural Network). The model sees and learns from this data.
- Validation Dataset:- The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as a skill on the validation dataset is incorporated into the model configuration.
- Test Dataset:- The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.



@learn.machinelearning

Splitting Test, train and validation datasets

- Splitting data completely depends on the size of data and the model you are using.
- If Model with very few hyperparameters will be easy to validate and tune, so you can probably reduce the size of your validation set.
- if your model has many hyperparameters, you would want to have a large validation set as well.
- If a model with no hyperparameters or ones that cannot be easily tuned, you probably don't need a validation set too.
- Thumb rule is to use 80%-20% or 70%-30% for train and validation and 80% -20% for validation and test datasets.
- If you have more than million data points then splitting will be different like 98%-2%.