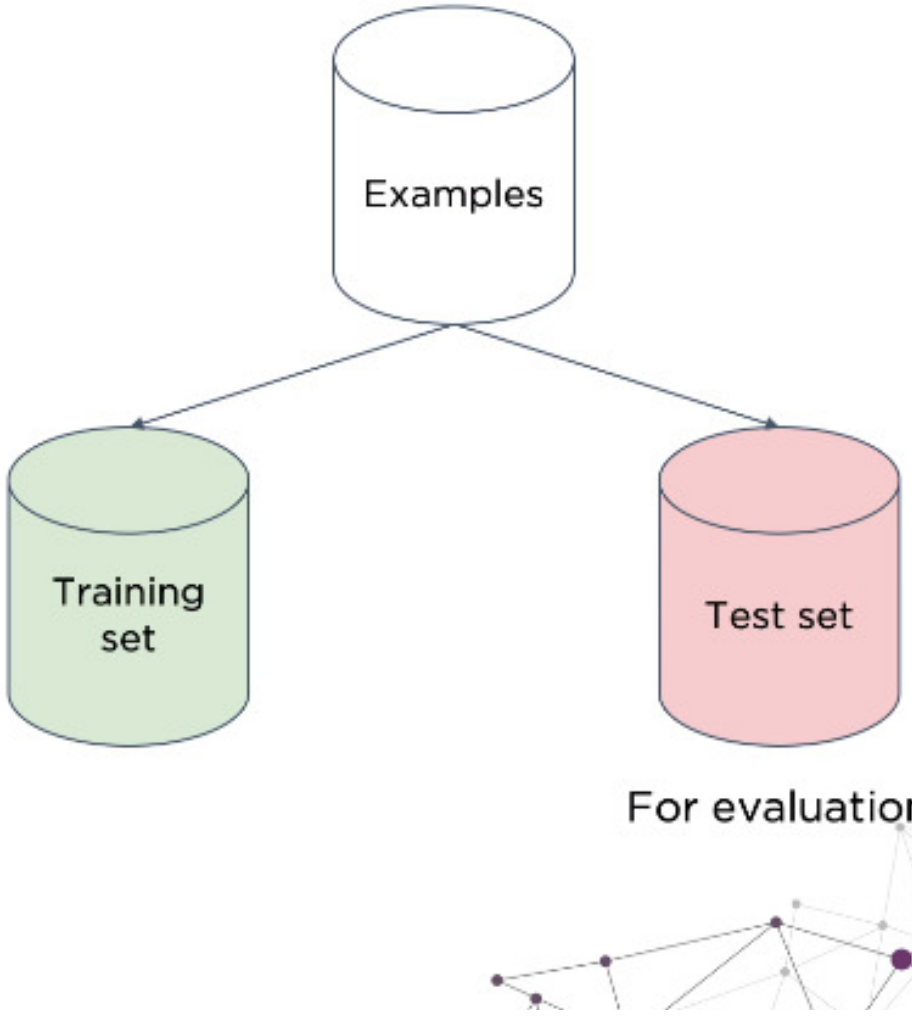
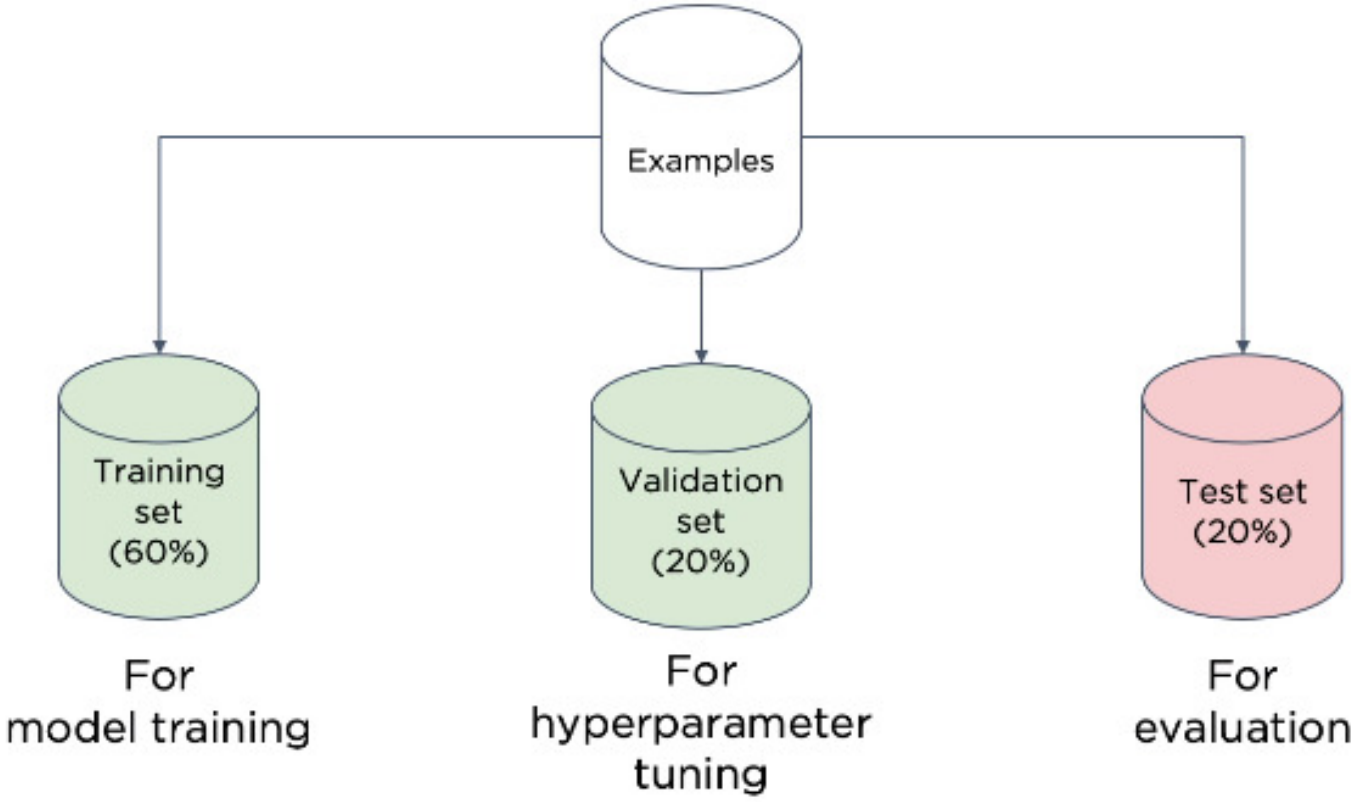
Experimental protocol

Never use a training example for evaluating an algorithm; otherwise the performance metric will be higher than in production.

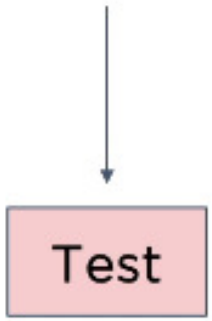


How to choose the best hyperparameters?



Cross-validation

Split	Exp. A	Exp. B	Exp. C
	Performance A1	Performance B1	Performance C1



K-fold cross-validation

Splits	Exp. A	Exp. B	Exp. C
	Performance A_1	Performance B_1	Performance C_1
	Performance A_2	Performance B_2	Performance C_2
	Performance A_K	Performance B_K	Performance C_K
	Average perf. A StdError perf. A	Average perf. B StdError perf. B	Average perf. C StdError perf. C