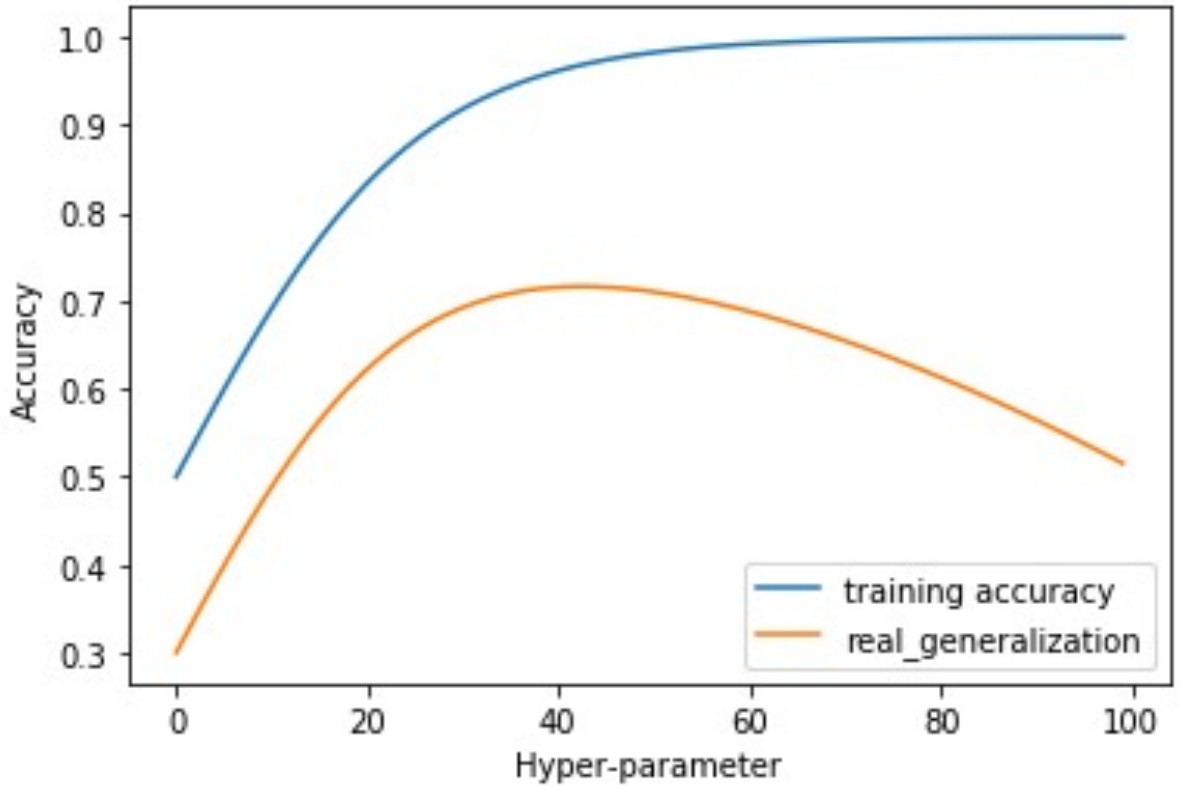
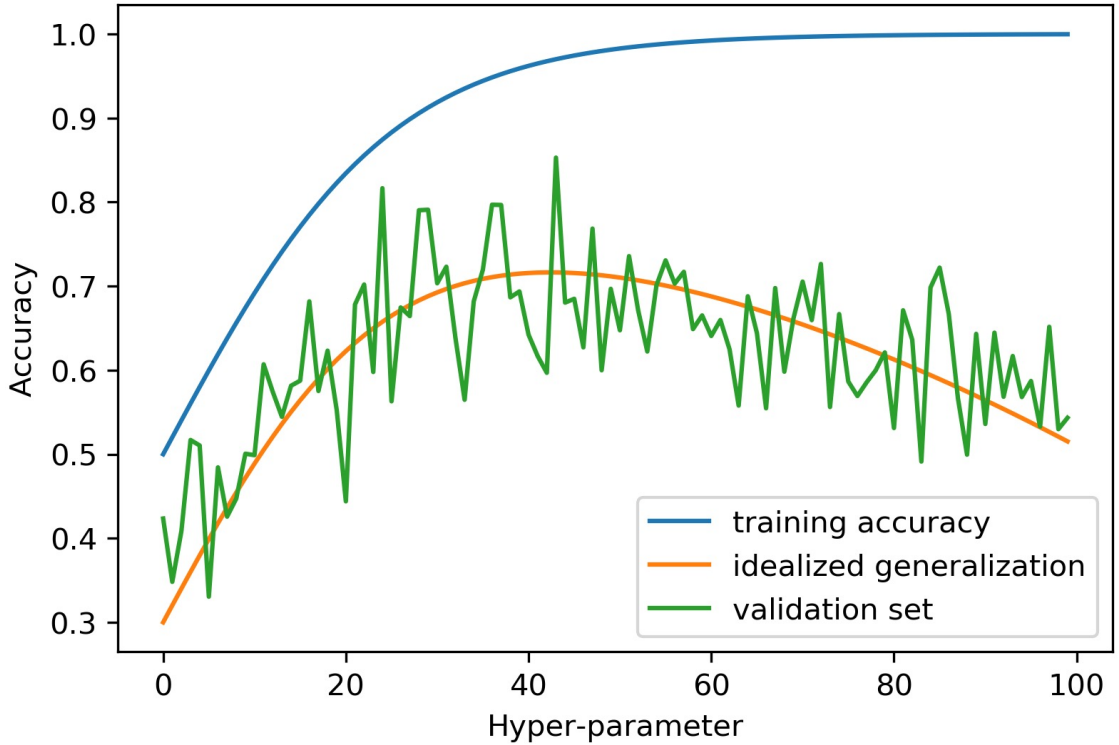
Train – Validate – Test

Part II

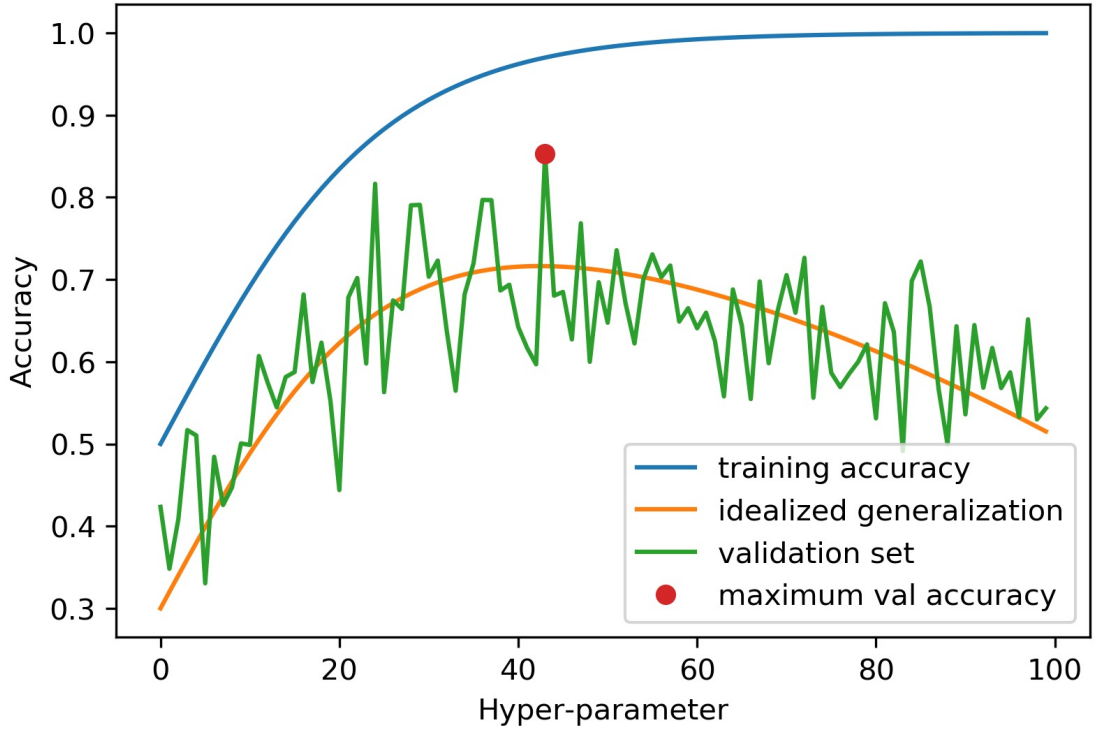
Overfitting the Validation set: Part I



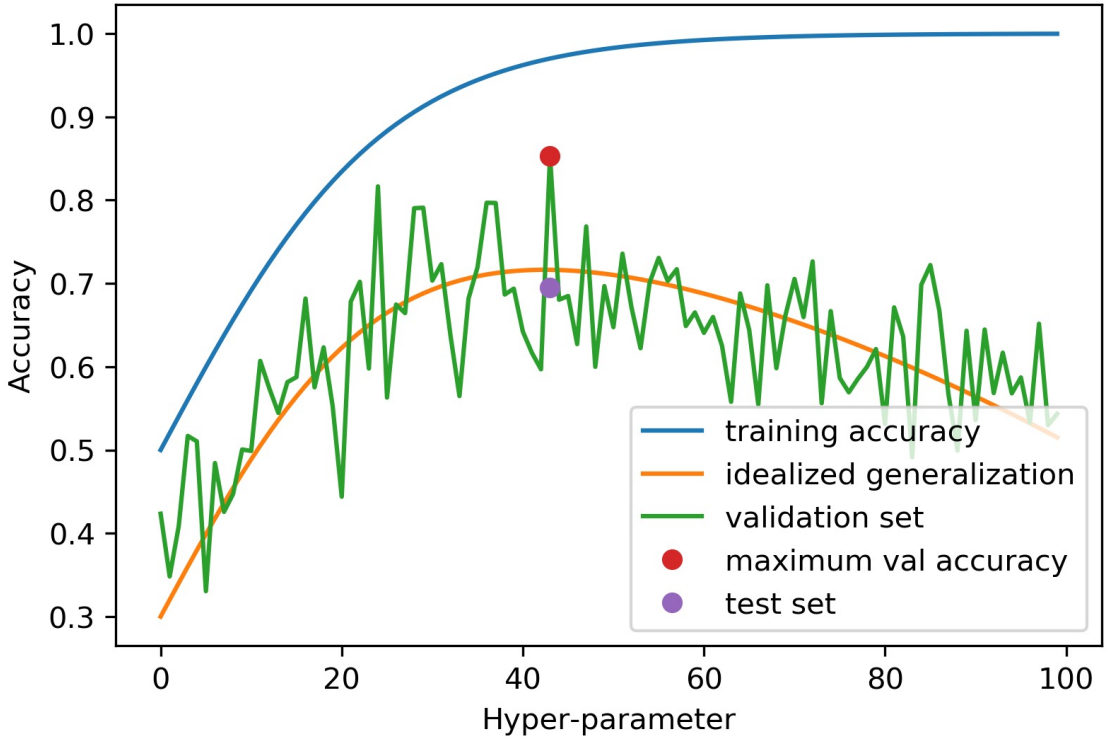
Overfitting the Validation set: Part II



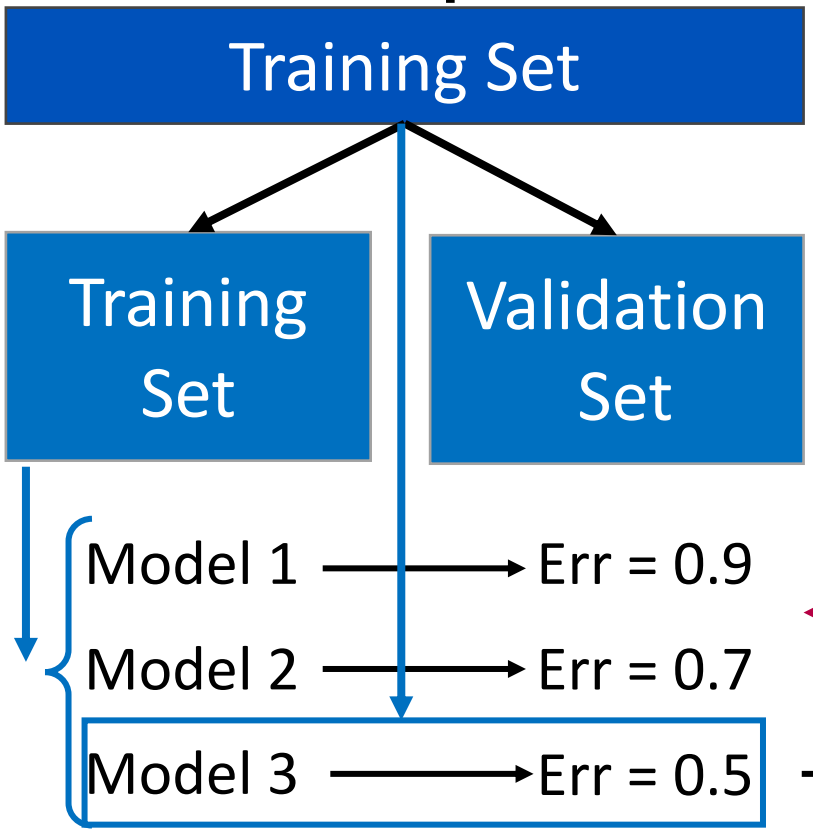
Overfitting the Validation set: Part III



Overfitting the Validation set: Part IV



Threefold Split



public

Testing Set

private

Testing Set

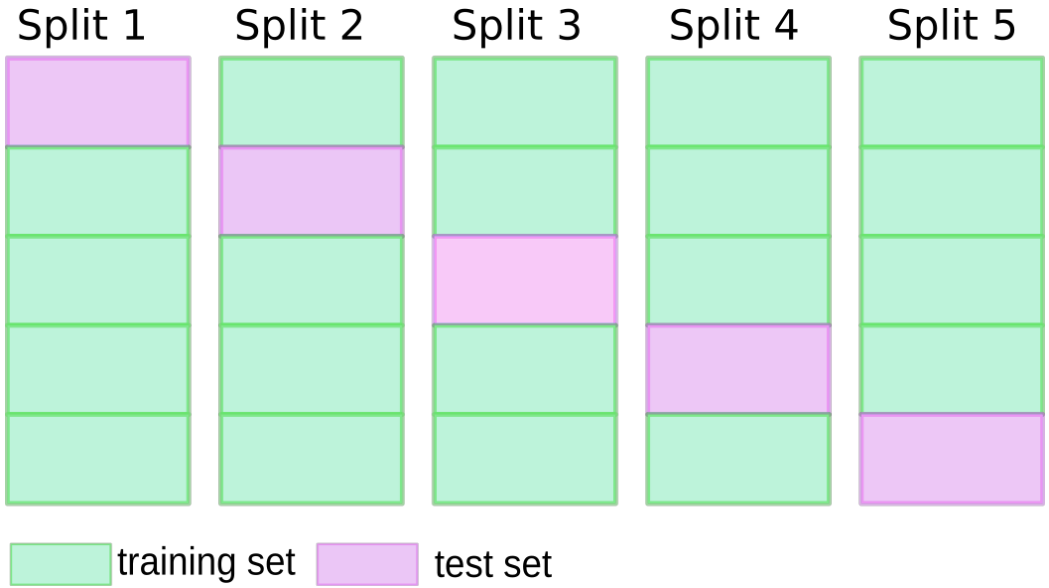
Using the results of public testing data to tune your model.

Here, you are making the public set better than private set.

Not recommended

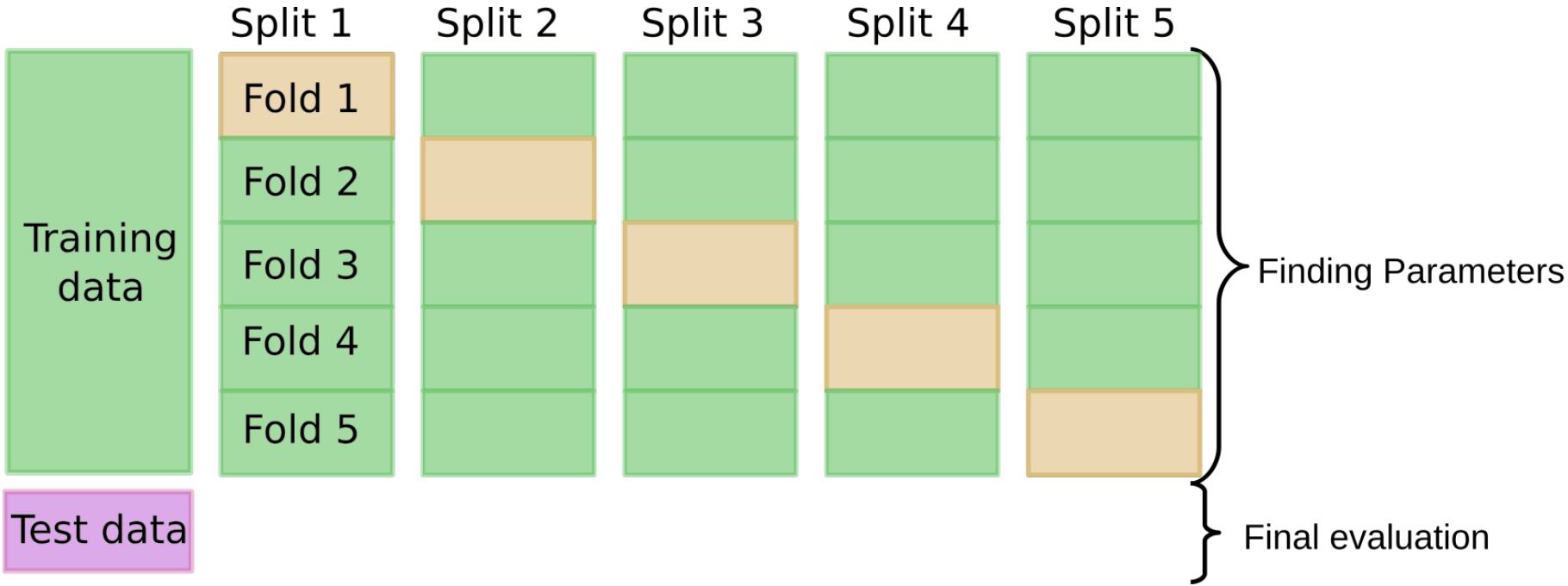
Err > 0.5 → Err > 0.5

Cross Validation

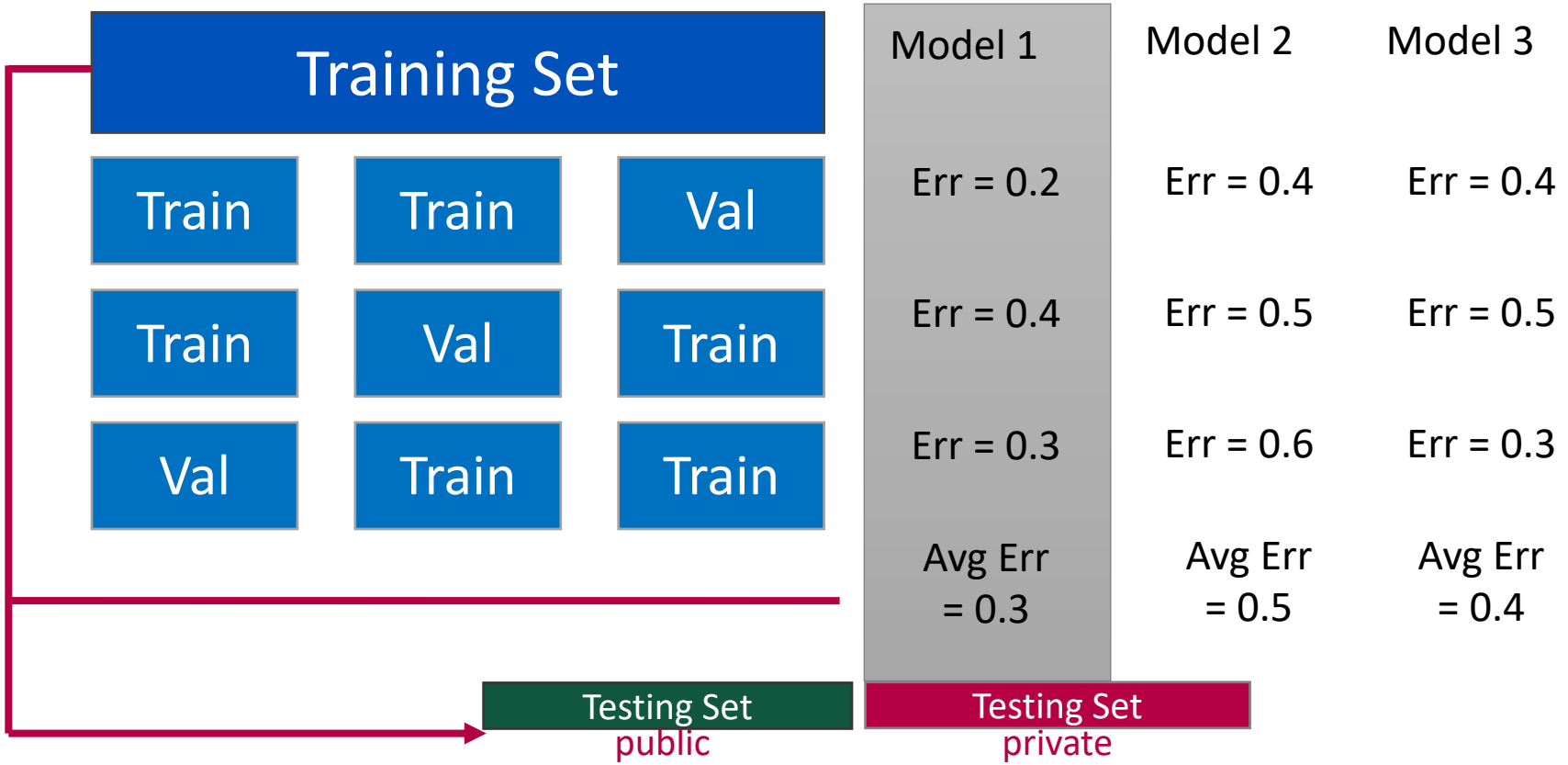


Pro: More stable, more data
Con: Slower

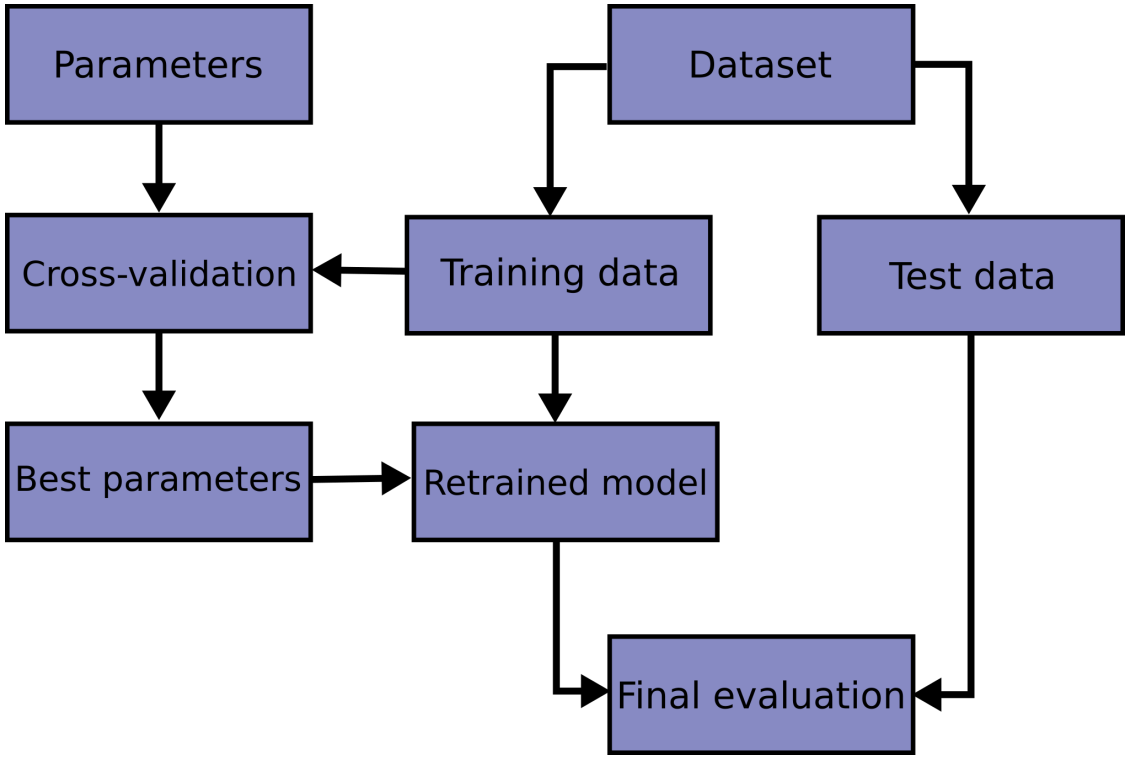
Cross Validation + Test Set



N-Fold Cross Validation



Machine Learning Workflow: Part II



Machine Learning Workflow: Part III

1. Split labeled data into **training, validation,** and **test sets.**
2. Repeat steps below until happy with performance on validation set:
   a. Build and revise your feature extraction methodology.
   b. Choose a machine learning algorithm.
   c. Train machine learning model with various hyperparameter settings.
   d. Evaluate prediction functions on validation set.
3. Retrain model (train + validation)
4. Evaluate performance on test set. Report this number to product manager.
5. Retrain on all labeled data (training + validation + test).
6. Deploy resulting prediction function.