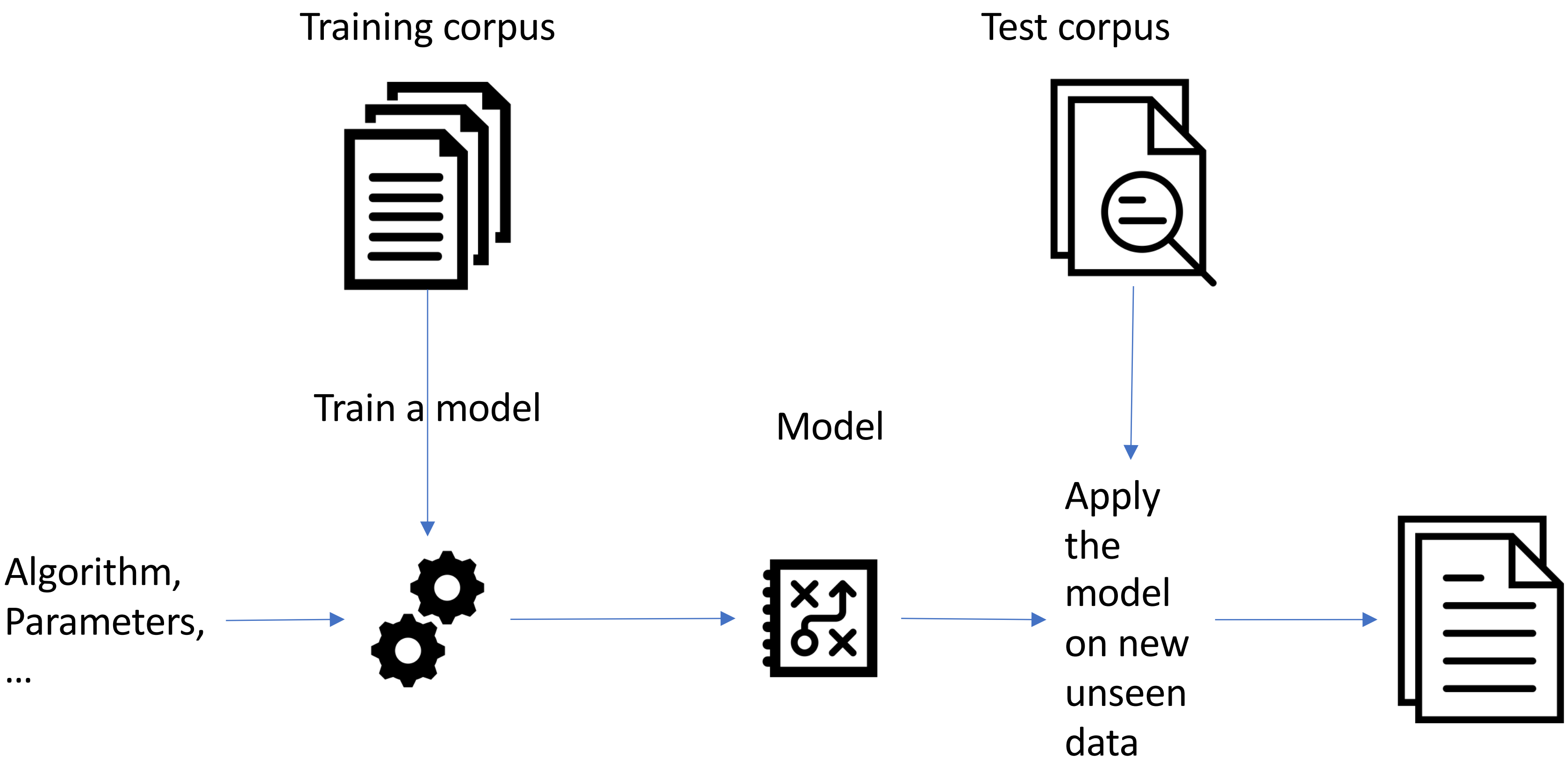
Development Cycle



Training corpus

Test corpus

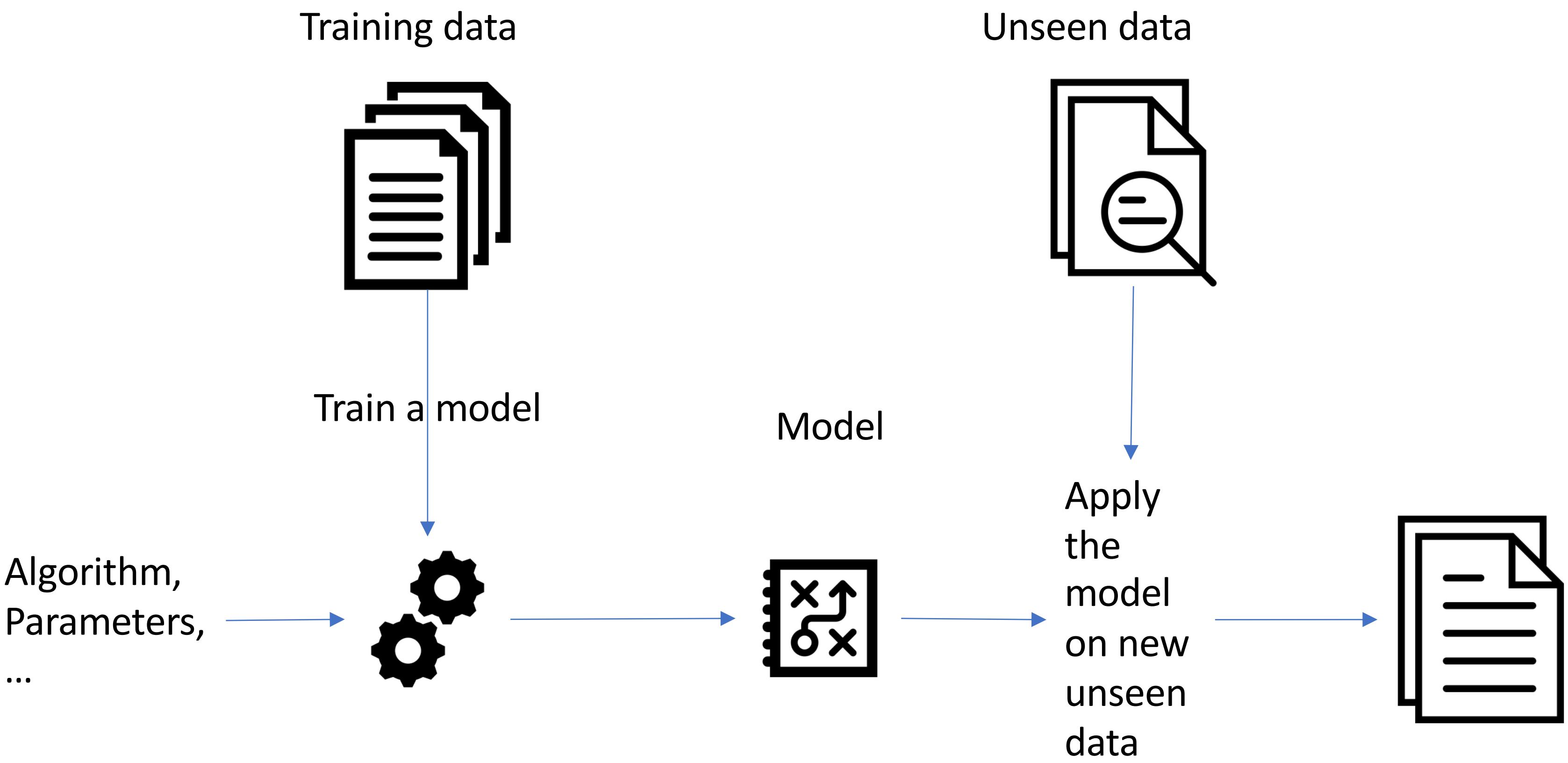
Train a model

Model

Algorithm, Parameters, …

Apply the model on new unseen data

Development Cycle

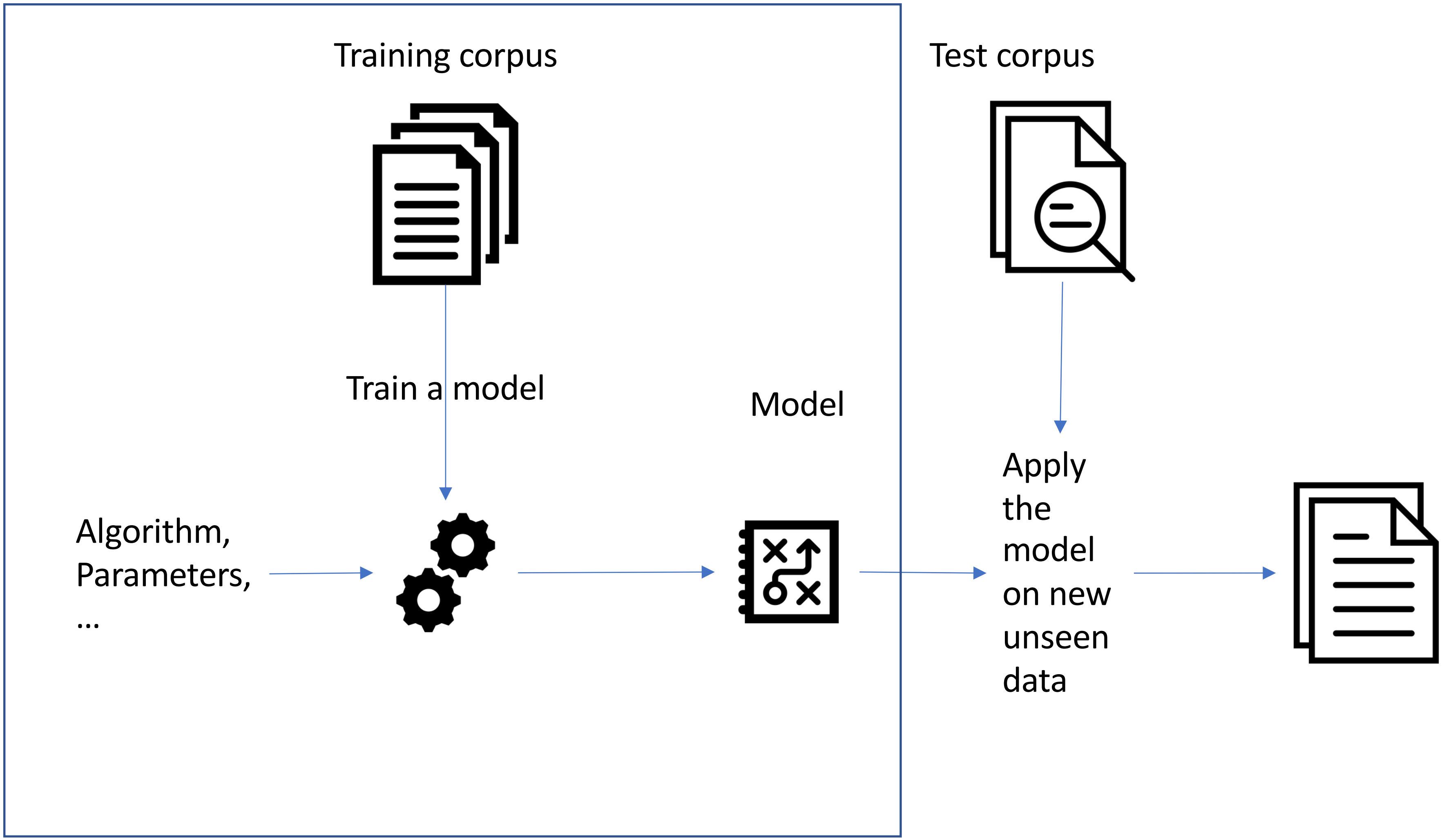


How do we know how good the model is?

Tuning

- Supervised machine learning algorithms require
  - parameters
  - different features and combinations
- How can we develop a model that we believe will work well on unseen data?

Development Cycle



Training corpus

Train a model

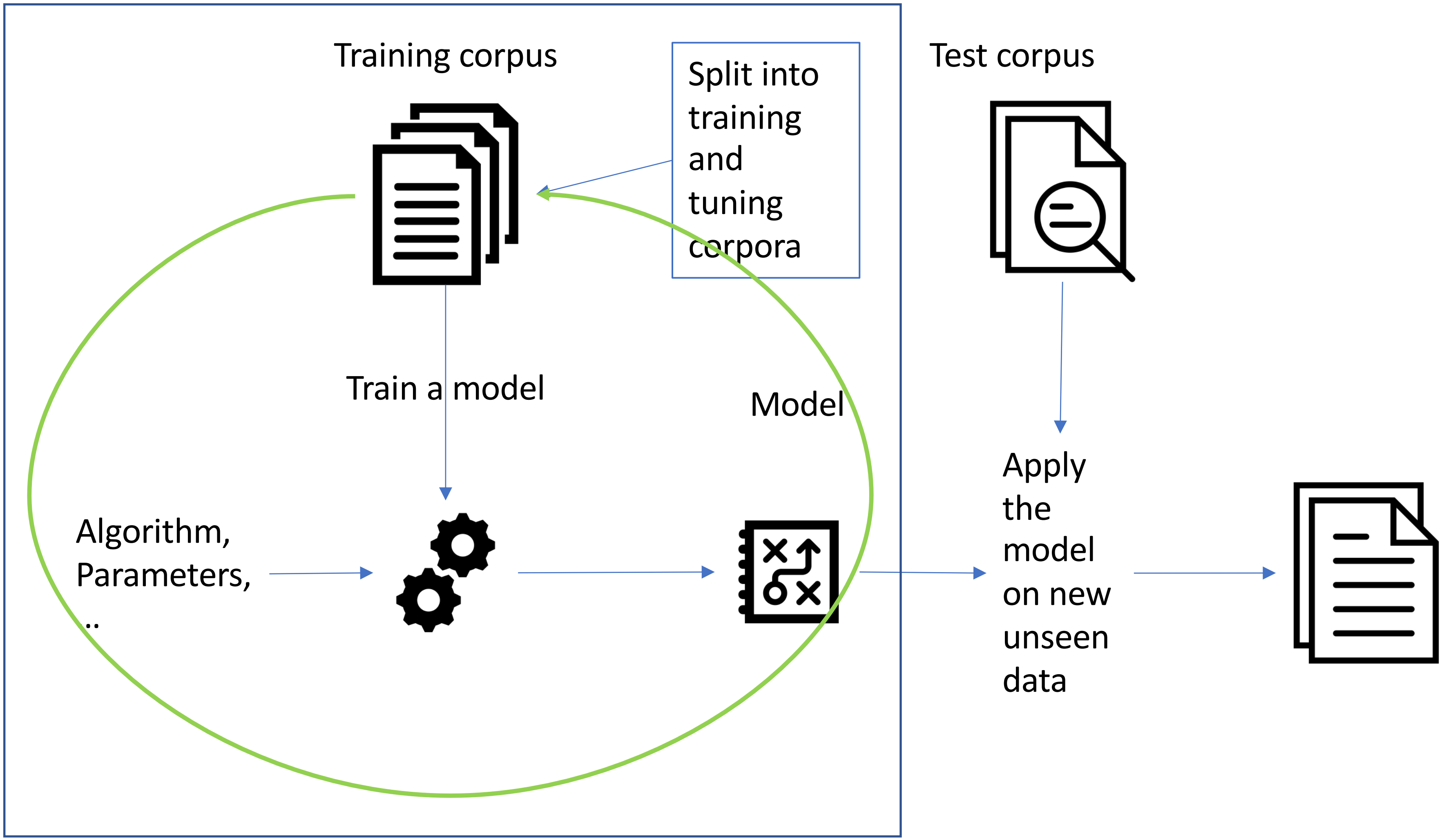
Model

Algorithm, Parameters, …

Test corpus

Apply the model on new unseen data

Development Cycle



Training corpus

Split into training and tuning corpora

Test corpus

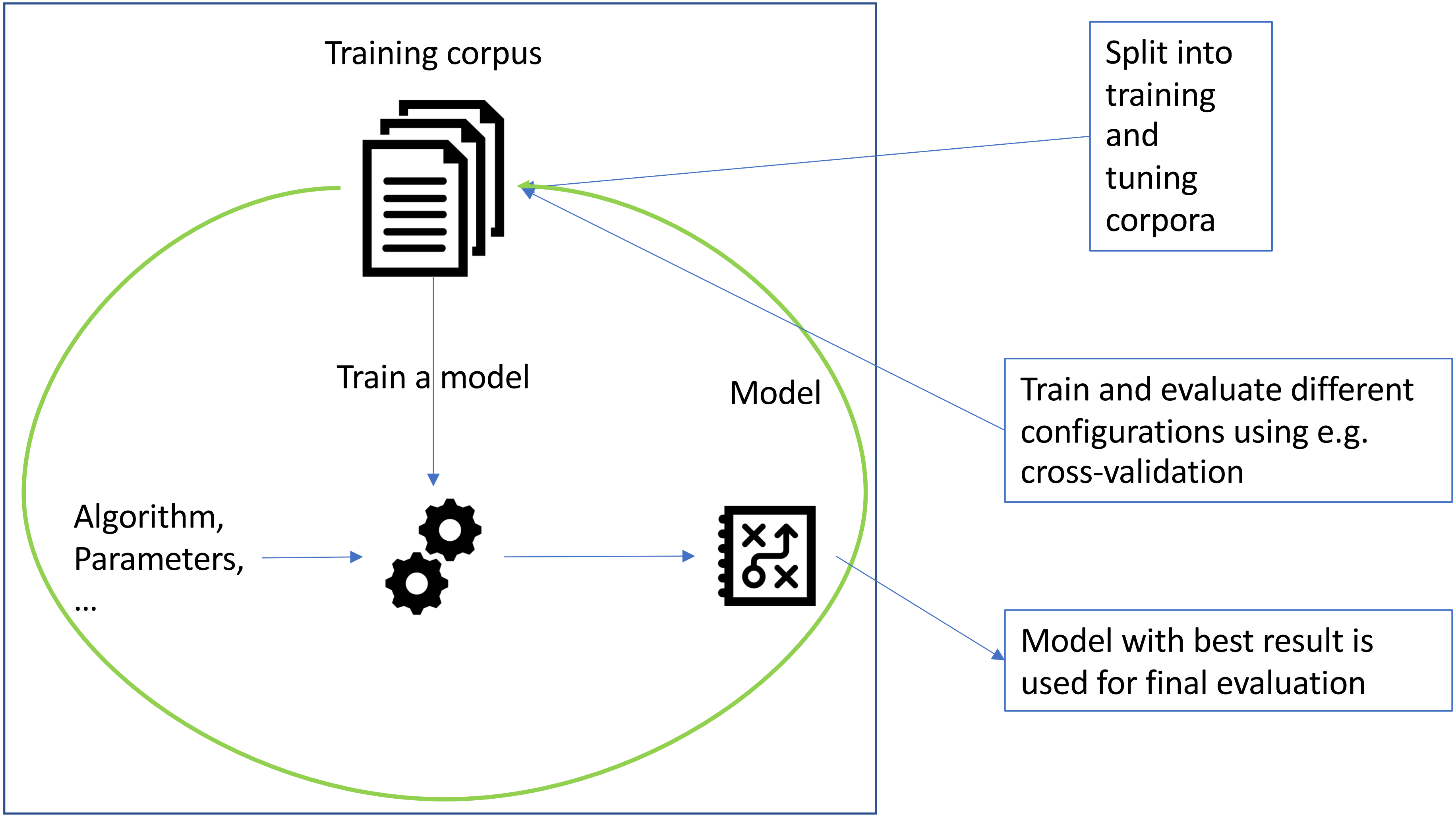
Train a model

Model

Algorithm, Parameters, ..

Apply the model on new unseen data

Development Cycle



Training corpus

Train a model

Model

Algorithm, Parameters, …

Split into training and tuning corpora

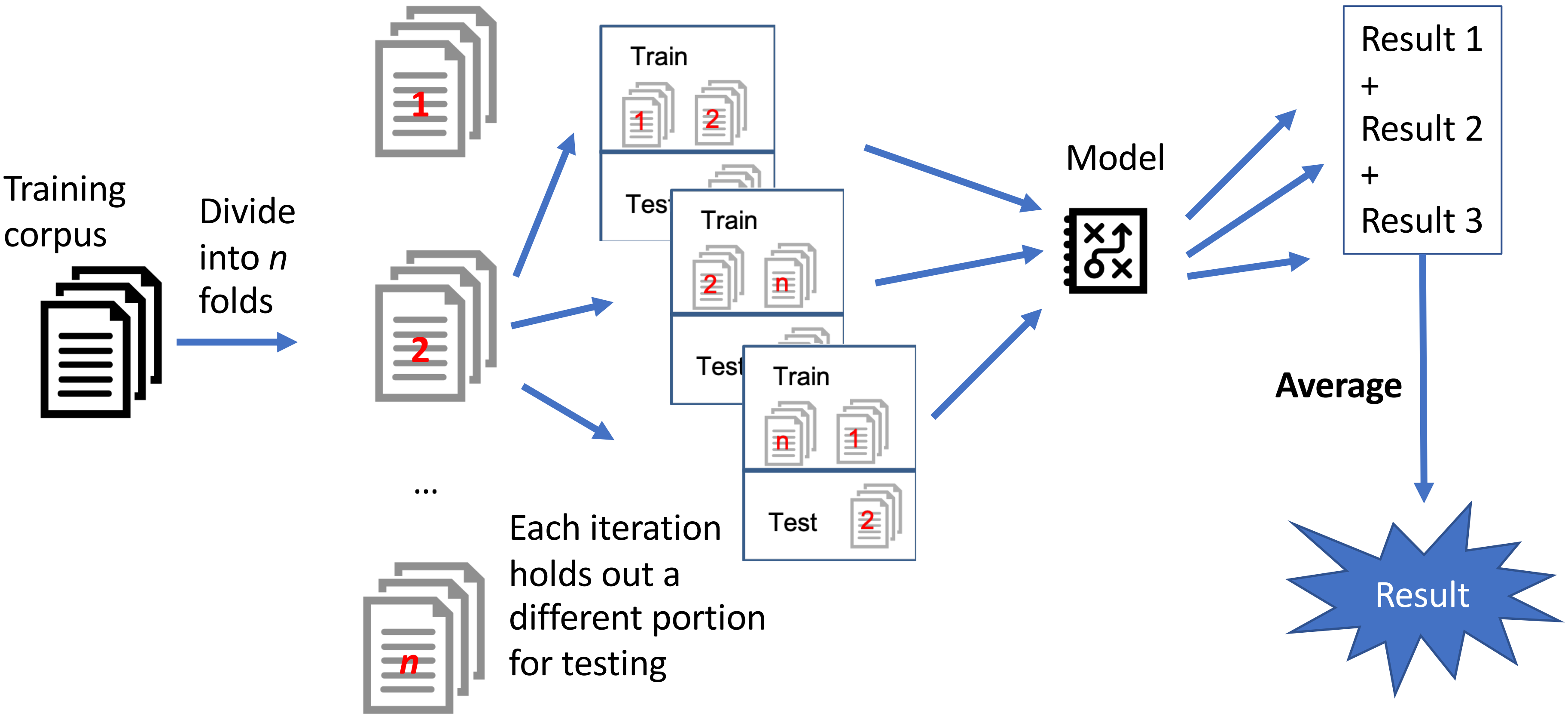
Train and evaluate different configurations using e.g. cross-validation

Model with best result is used for final evaluation

Cross-validation

- Splits the training data into n portions
- Each iteration, one nth is used for testing, the rest for training
- All results averaged

Cross-validation



Training corpus

Divide into *n* folds

...

Each iteration holds out a different portion for testing

Train

Test

Train

Test

Train

Test

Model

Result 1
+
Result 2
+
Result 3

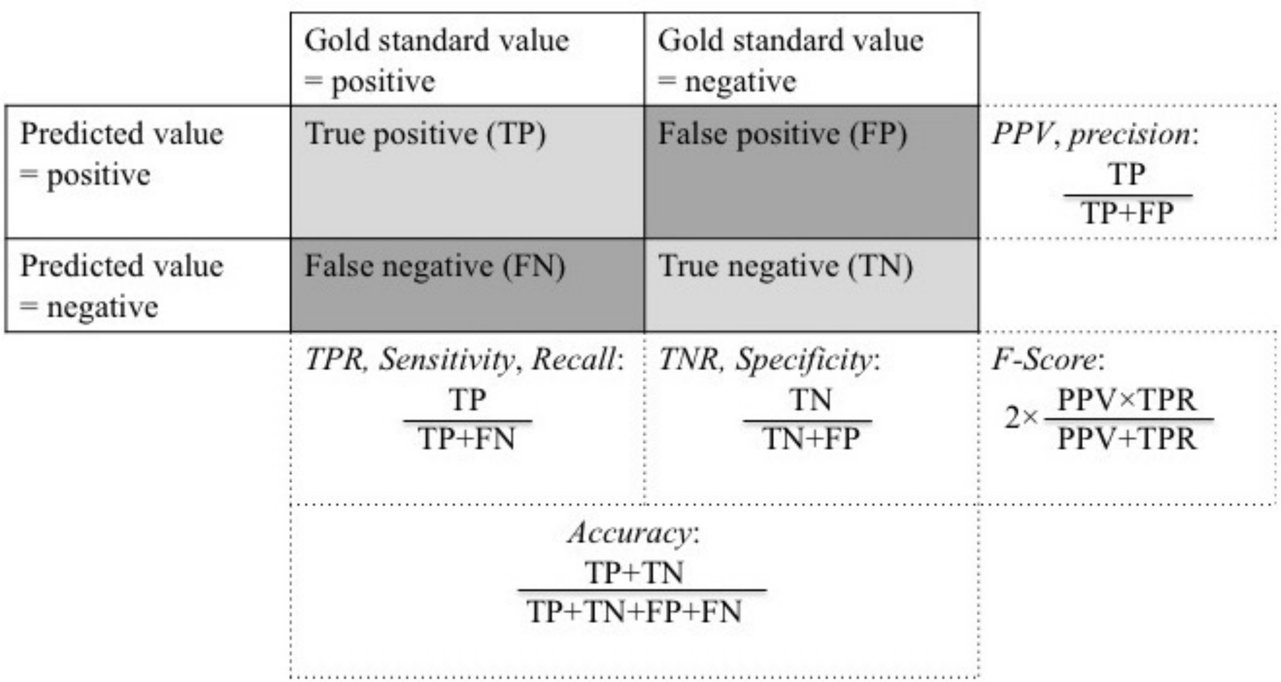
Average

Result

Final evaluation

- Gold standard
  - To evaluate model performance, we need test data with the ‘right answers’
  - This has to be different data than the training data!

Intrinsic evaluation



	Gold standard value = positive	Gold standard value = negative	
Predicted value = positive	True positive (TP)	False positive (FP)	*PPV*, *precision*: $\frac{\text{TP}}{\text{TP+FP}}$
Predicted value = negative	False negative (FN)	True negative (TN)	
	TPR, Sensitivity, Recall: $\frac{\text{TP}}{\text{TP+FN}}$	*TNR, Specificity*: $\frac{\text{TN}}{\text{TN+FP}}$	*F-Score*: $2\times\frac{\text{PPV}\times\text{TPR}}{\text{PPV+TPR}}$
	Accuracy: $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$		

Intrinsic evaluation

- Micro and macro average
  - Micro average is computed on all instances
  - Macro average is computed independently for each class and then averaged
    - Problematic if there is big class imbalance

Practicals

- We will use sklearn, nltk, spacy and jupyter notebooks today
  - You can try different machine learning algorithms, and you will work on evaluation in different ways.
- Other tools you can try
  - GATE has support for most supervised learning algorithms and allows for easy experimentation with other language features
  - Weka
  - …

Thank you!

sumithra.velupillai@kcl.ac.uk