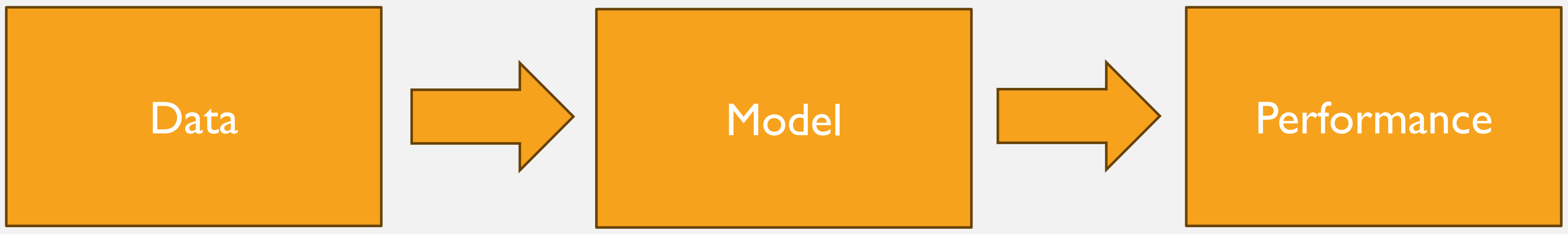
VALIDATION METHODS & PERFORMANCE METRICS

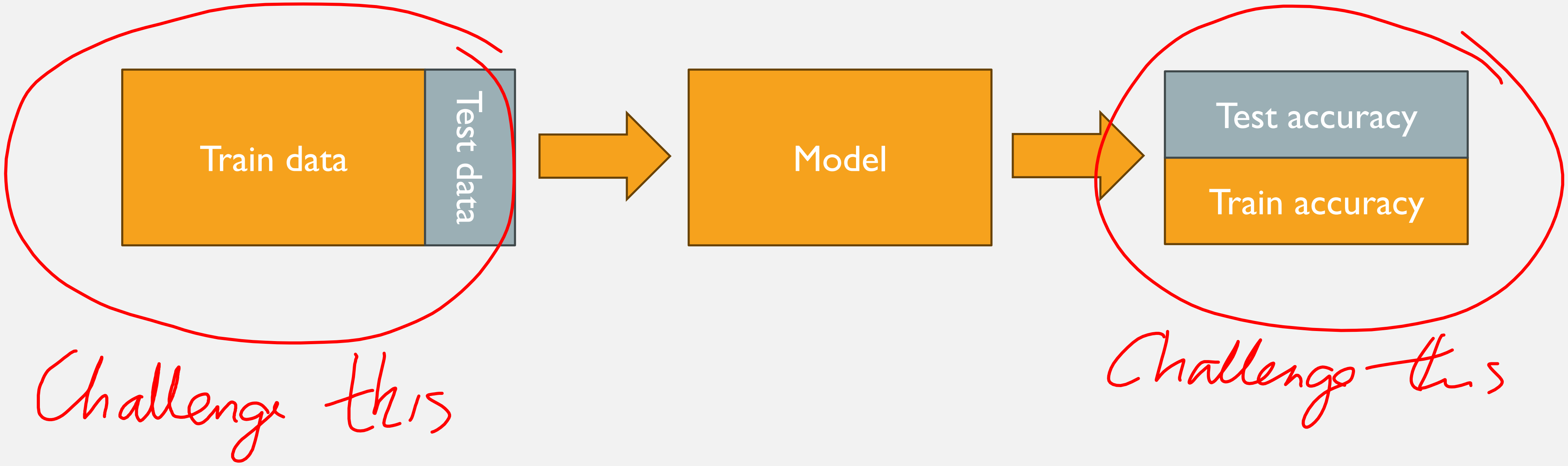
Lecture 6

MAL1, 2024

THE BIG PICTURE



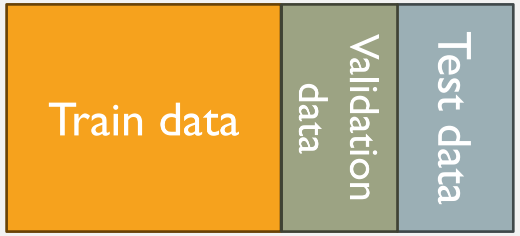
THE BIG PICTURE



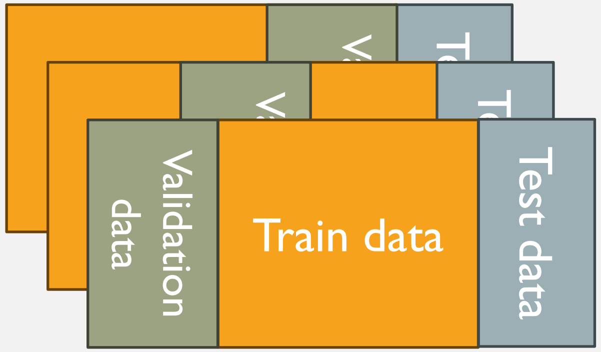
VALIDATION METHODS

VALIDATION METHODS

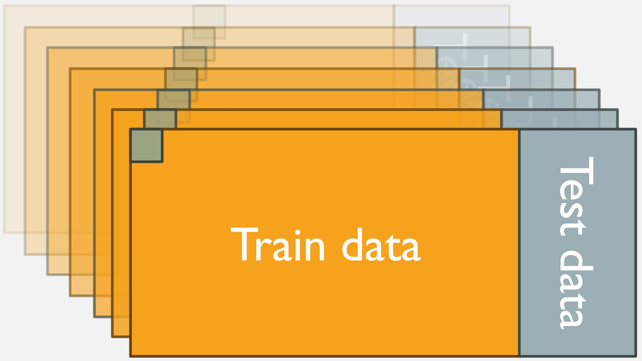
Train-test methodology



Train-val-test methodology



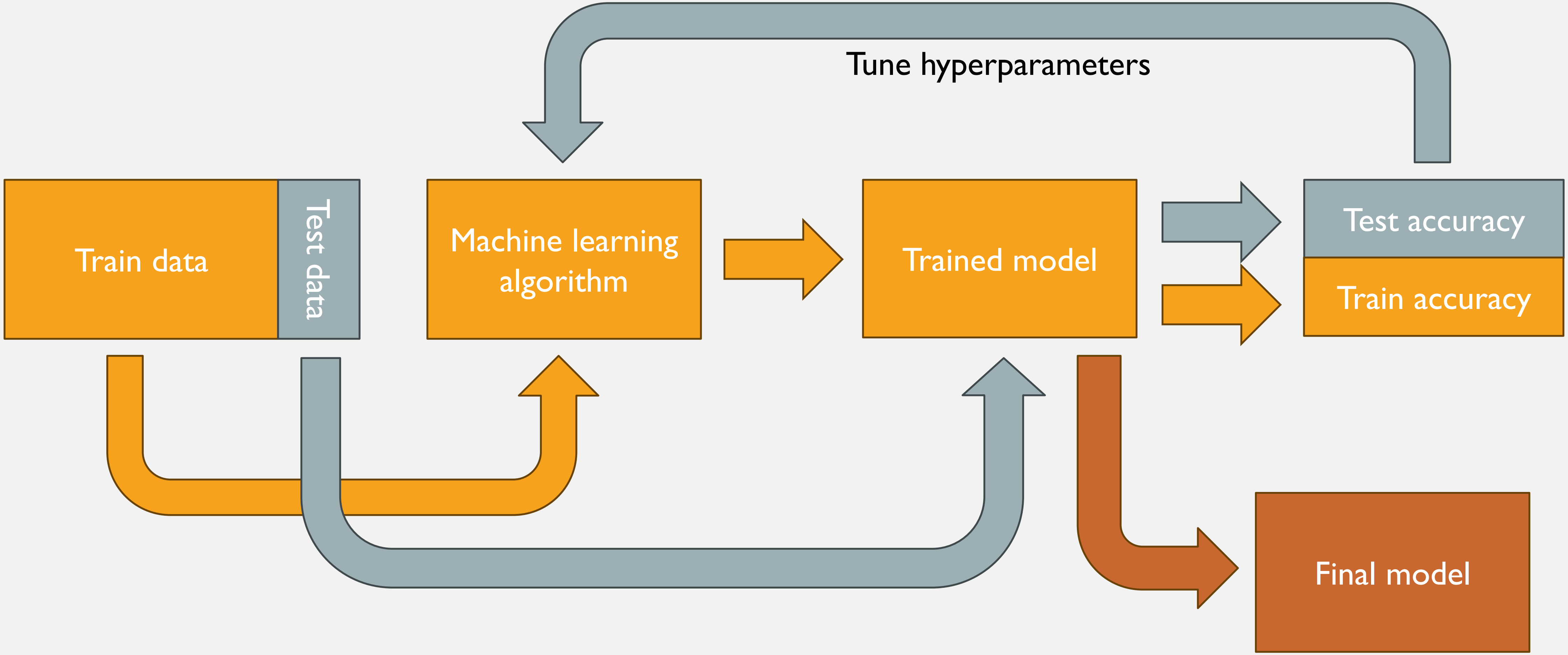
Cross-validation methodology



Leave-1-out cross validation methodology

TRAIN-TEST METHODOLOGY

Model indirectly sees test data

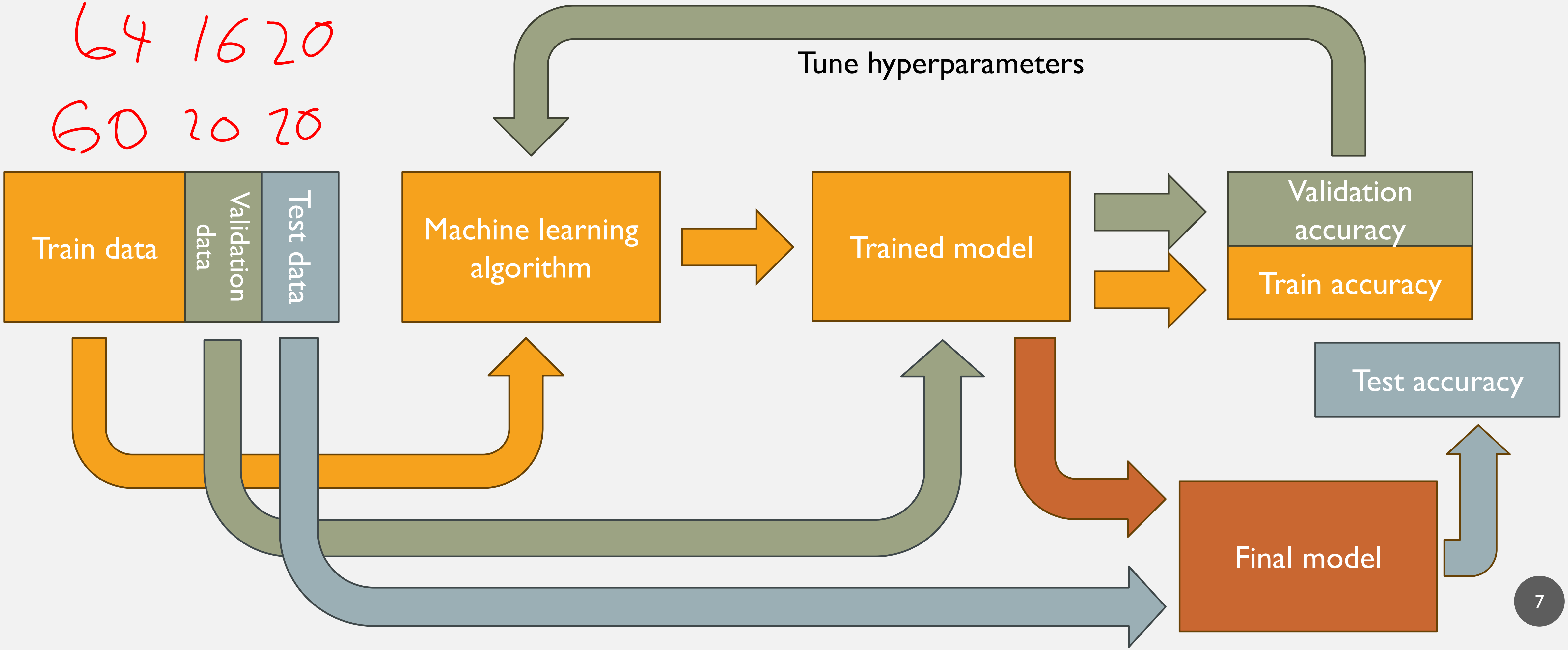


TRAIN-VAL-TEST METHODOLOGY

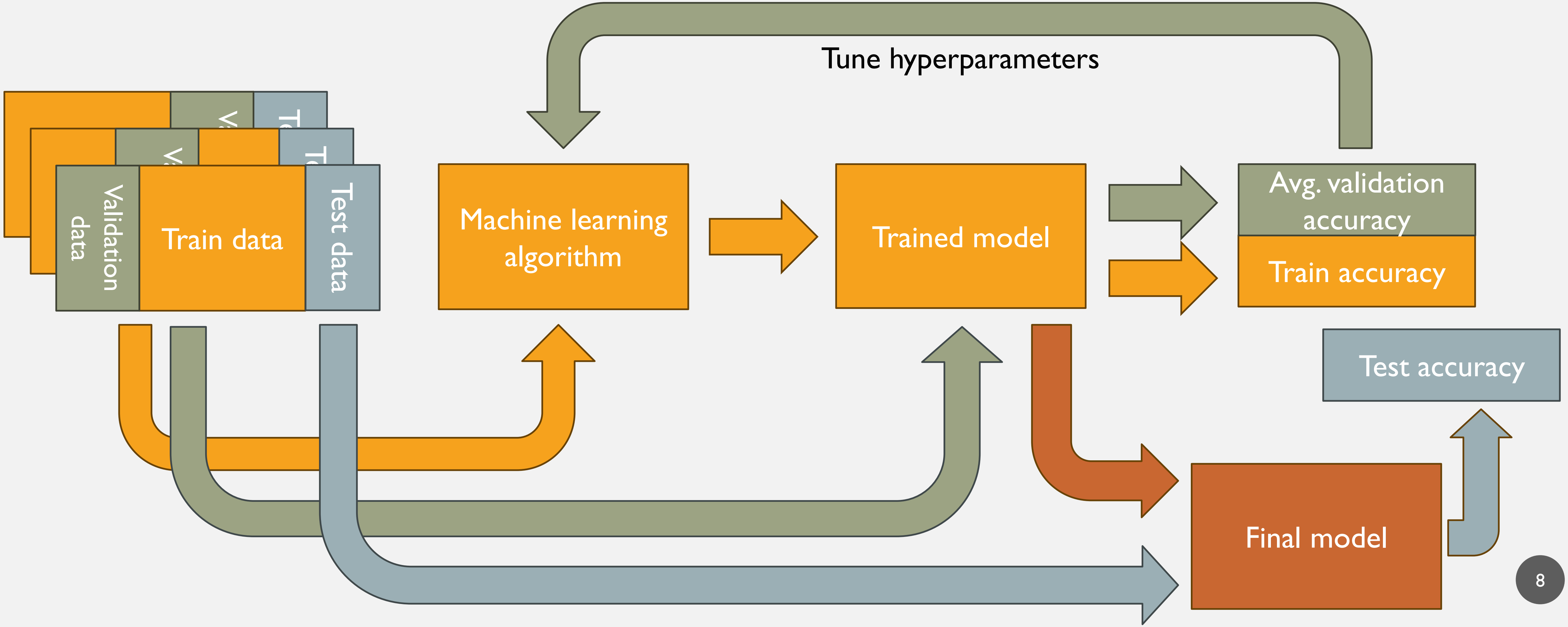
Algorithm does not see test data

64 16 20

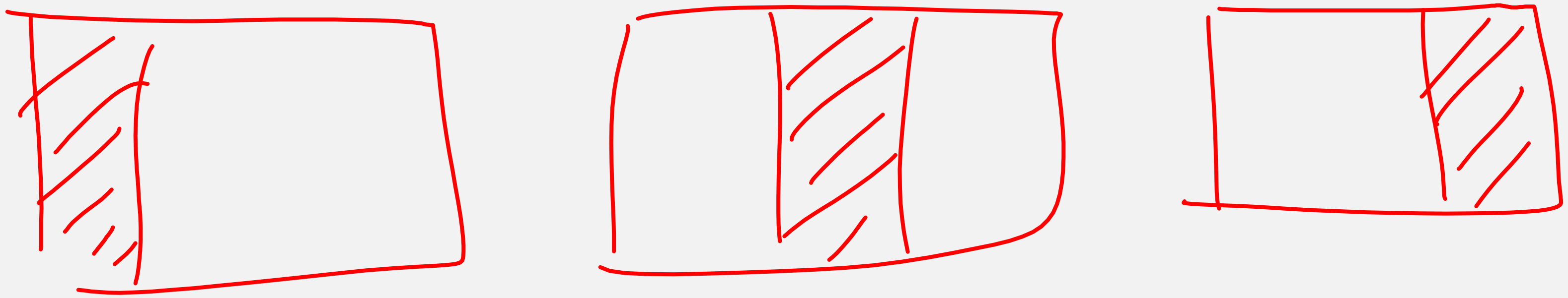
60 20 20



CROSS-VALIDATION METHODOLOGY



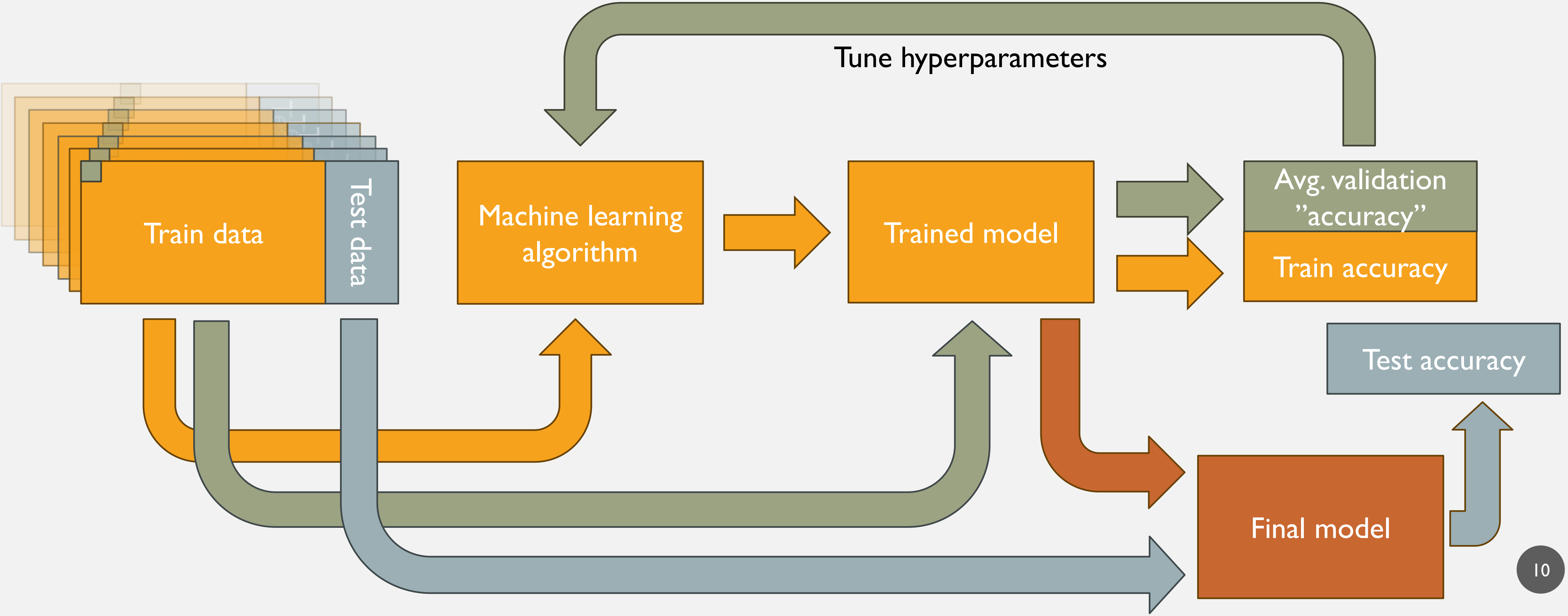
CROSS-VALIDATION METHODOLOGY



* Split data into folds (here 3, could be 5, 10)
* Train on larger set, validate on small
* Accuracy = average all models

LEAVE-1-OUT CROSS-VALIDATION METHODOLOGY

Special case of CV with folds of size 1



CODE EXAMPLE

Jupyter Notebook **Validation methods**

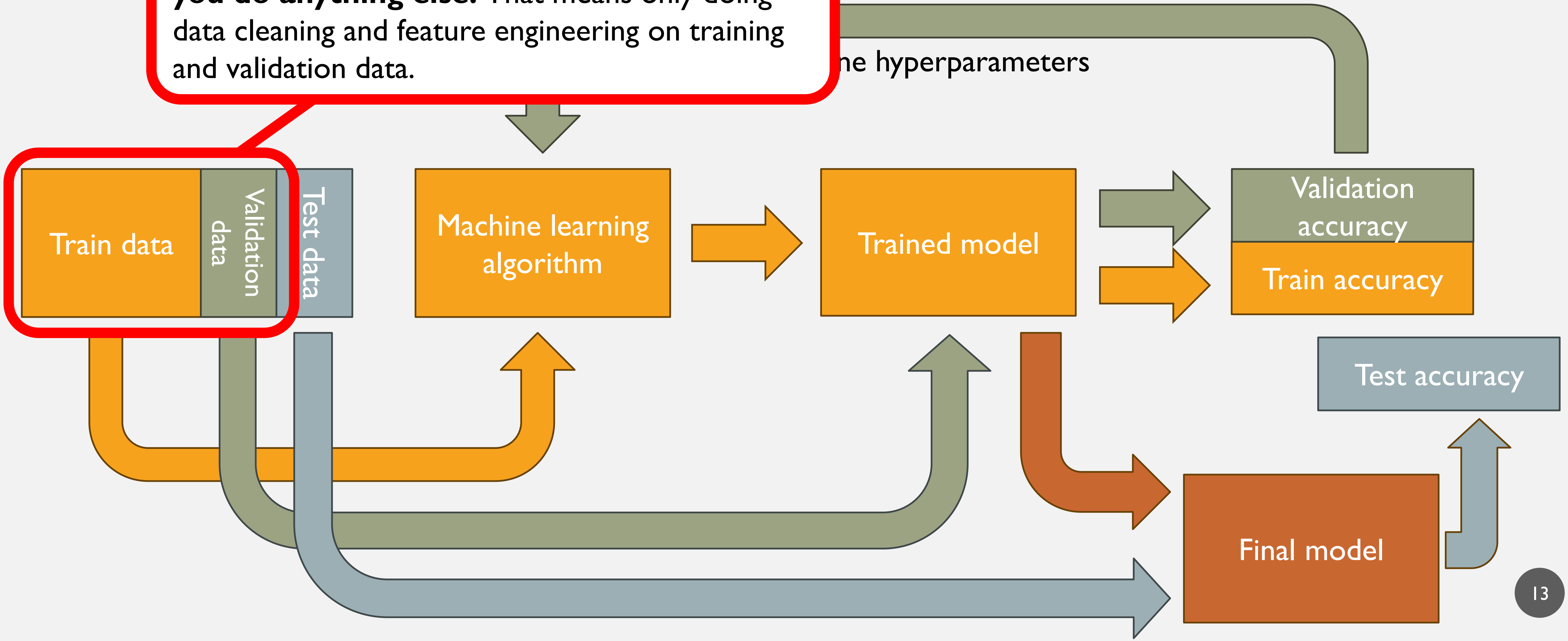
A NOTE ON PREPROCESSING

- Deal with outliers
- Deal with missing values
- Normalize/scale data
- One-hot encoding
- Representing text data
- …

but when should you do this?

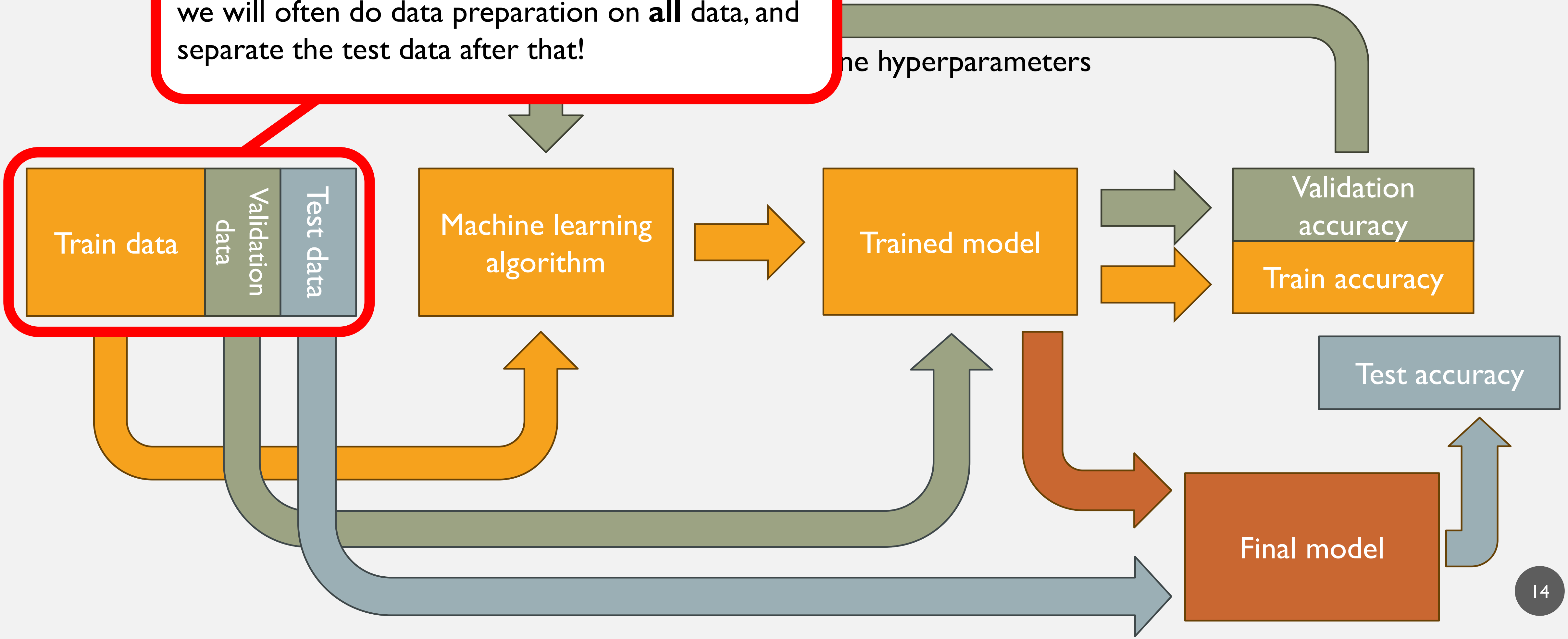
A NOTE ON PREPROCESSING

Ideally you should isolate you testing data **before you do anything else!** That means only doing data cleaning and feature engineering on training and validation data.



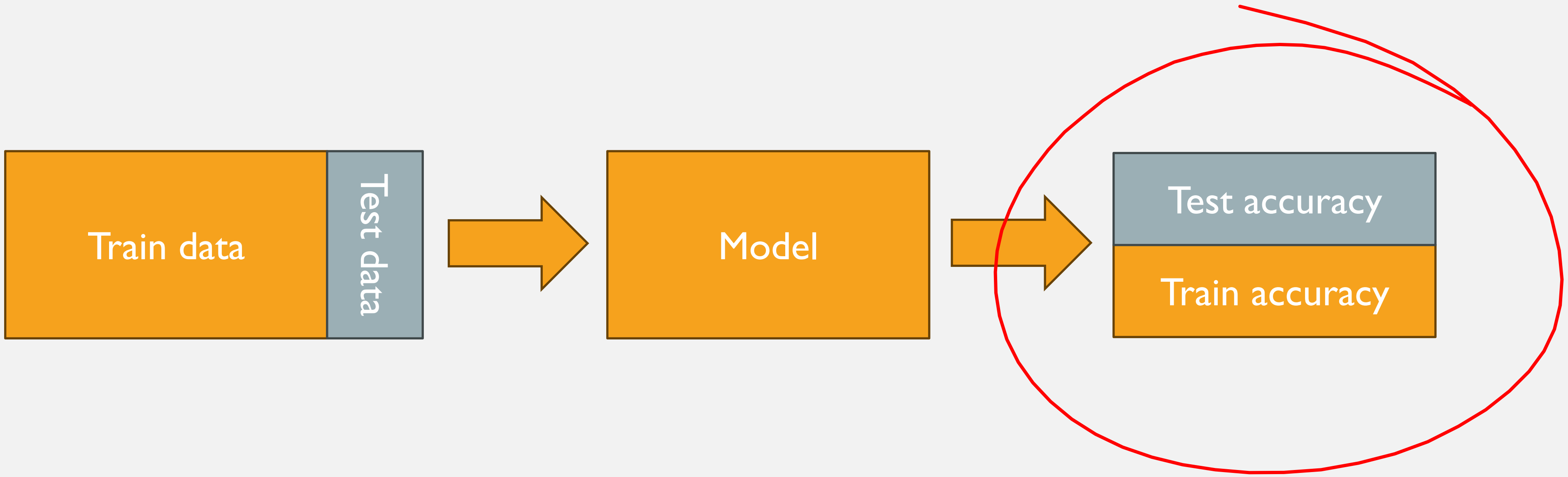
A NOTE ON PREPROCESSING

But often this is quite impractical. So for simplicity, we will often do data preparation on **all** data, and separate the test data after that!



PERFORMANCE METRICS

THE BIG PICTURE



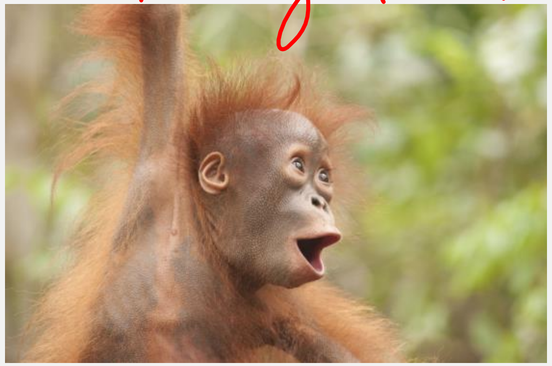
Then performance metrics

TYPES OF ERRORS

is orangutan

is not orangutan

predict orangutan



"This is an orangutan"

"This is an orangutan"

predict not orangutan

"This is not an orangutan"

"This is not an orangutan"

TYPES OF ERRORS

true class

positive | negative

predicted class

positive

"This is an orangutan"

"This is an orangutan"

negative

"This is not an orangutan"

"This is not an orangutan"

TYPES OF ERRORS

”orangutan”

”not orangutan”

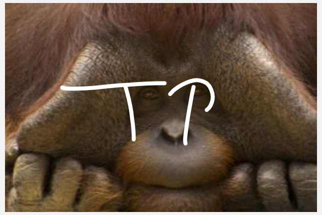
”orangutan”

”not orangutan”

”orangutan”

”not orangutan”

”orangutan”



”orangutan”

”orangutan”

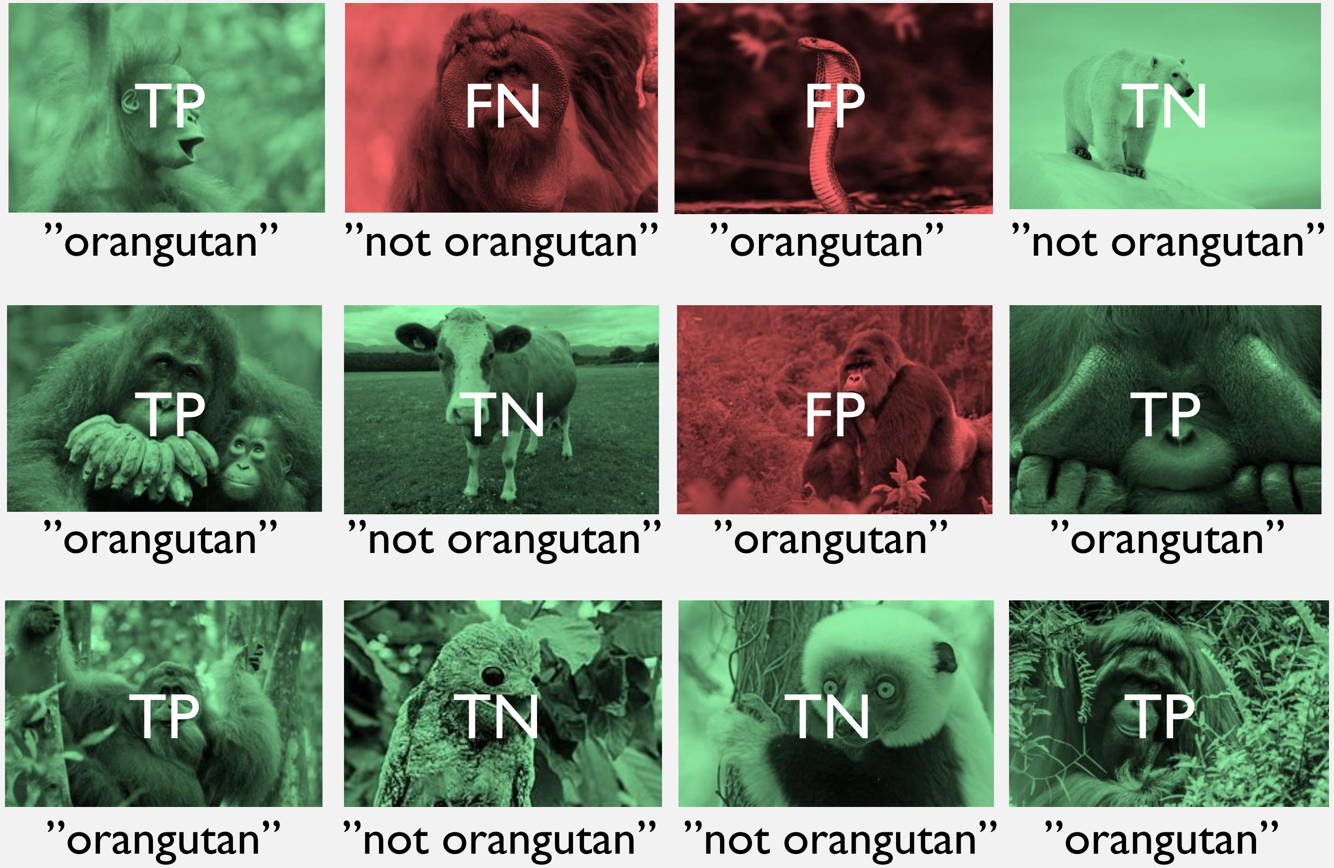
”not orangutan”

”not orangutan”

”orangutan”

predicted class \ true class	positive	negative
positive	TRUE POSITIVE	FALSE POSITIVE
negative	FALSE NEGATIVE	TRUE NEGATIVE

TYPES OF ERRORS



confusion matrix

		true class: positive	true class: negative
predicted class	positive	TP = 5	FP = 2
predicted class	negative	FN = 1	TN = 4

ACCURACY

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{5+4}{5+4+2+1} = 0.75$$

"How often do we get the right answer?"

		true class: positive	true class: negative
predicted class	positive	TP = 5	FP = 2
predicted class	negative	FN = 1	TN = 4

WHY ACCURACY IS NOT GOOD ENOUGH

A model to predict whether or not someone is a terrorist:

*Everyone is **not** a terrorist.*

$$\text{accuracy} = \frac{0+9999}{0+9999+0+1} = 0.9999$$

but the model is useless

particularly bad for skewed dataset

		true class: positive	true class: negative
predicted class	positive	TP = 0	FP = 0
predicted class	negative	FN = 1	TN = 9999

PERFORMANCE METRICS

- $\text{accuracy} = \dfrac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \dfrac{\text{correct predictions}}{\text{all predictions}}$

- $\text{precision} = \dfrac{\text{TP}}{\text{TP+FP}} = \dfrac{\text{correct positive predictions}}{\text{all positive predictions}}$

"how often is a positive answer correct?"

- $\text{recall} = \dfrac{\text{TP}}{\text{TP+FN}} = \dfrac{\text{correct positive predictions}}{\text{all positive instances}}$

"how often is a positive instance correctly identified?"

USING RECALL INSTEAD

A model to predict whether or not someone is a terrorist:

*Everyone is **not** a terrorist.*

$$\text{recall} = \frac{TP}{TP+FN} = \frac{0}{0+1} = 0$$

terrible model

with accuracy = 99,99%

		true class: positive	true class: negative
predicted class	positive	TP = 0	FP = 0
predicted class	negative	FN = 1	TN = 9999

SOME EXAMPLES

- Determine whether someone is a terrorist
  - avoid false negatives – use recall!
- Determine whether you have COVID-19 during the pandemic
  - avoid false negatives – use recall!
- Determine whether a video is suitable for children to watch
  - avoid false positives – use precision!
- Determine whether someone should be sentenced to life in prison
  - "Innocent until proven guilty" means avoid false positives – use precision!

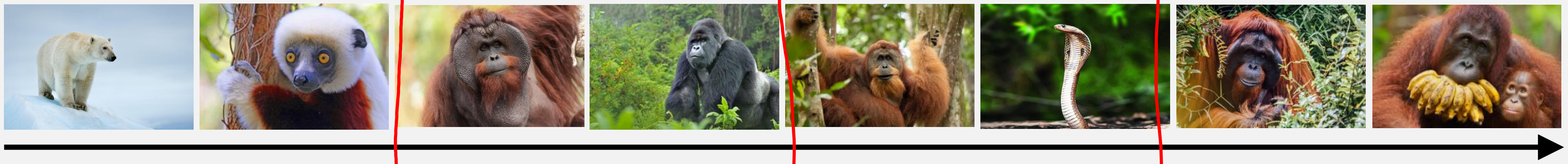
To optimize hyperparameters on a particular metric:

```
GridSearchCV(clf, parameters, scoring="recall")
GridSearchCV(clf, parameters, scoring="precision")
```

THE PRECISION/RECALL TRADE-OFF

High recall usually means low precision – and vice versa

decision thresholds



Precision $\frac{4}{4+2} = 0.67$ $\frac{3}{3+1} = 0.75$ $\frac{2}{2+0} = 1.0$

Recall $\frac{4}{4+0} = 1.0$ $\frac{3}{3+1} = 0.75$ $\frac{2}{2+2} = 0.5$

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

THE PRECISION/RECALL TRADE-OFF

High recall usually means low precision – and vice versa

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

high prec
⇒ avoid FP

high recall
⇒ avoid FN

$$F_1\text{-score} = \frac{TP}{TP + \frac{FP+FN}{2}}$$

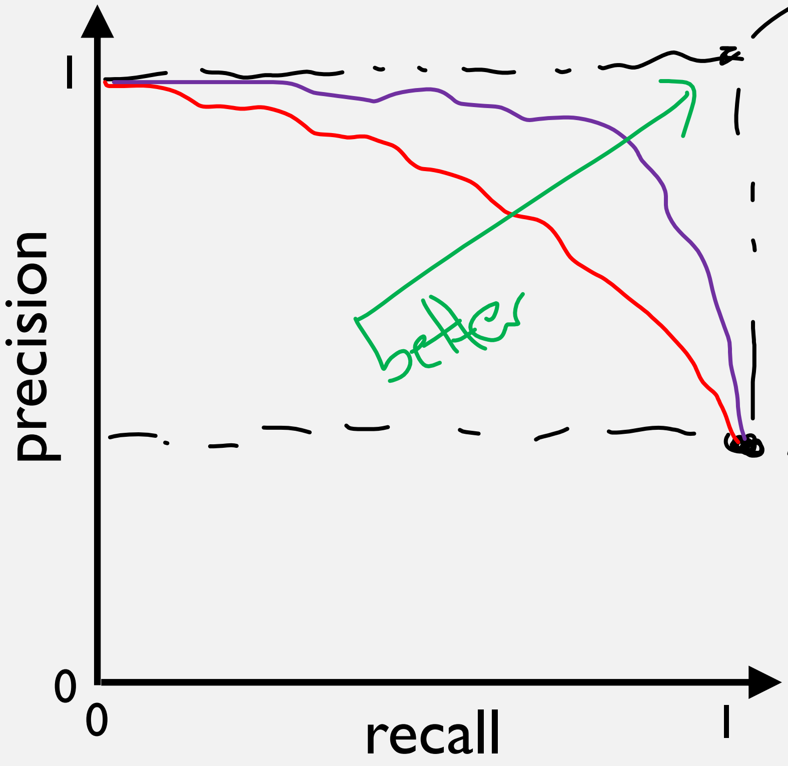
tries to avoid both

THE PRECISION-RECALL CURVE

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

AUC = area under curve

max 1



ideal classifier
rec = prec = 1

baseline all predictions are positive
recall = 1
prec: % positive

CODE EXAMPLE

Jupyter Notebook **Performance metrics**

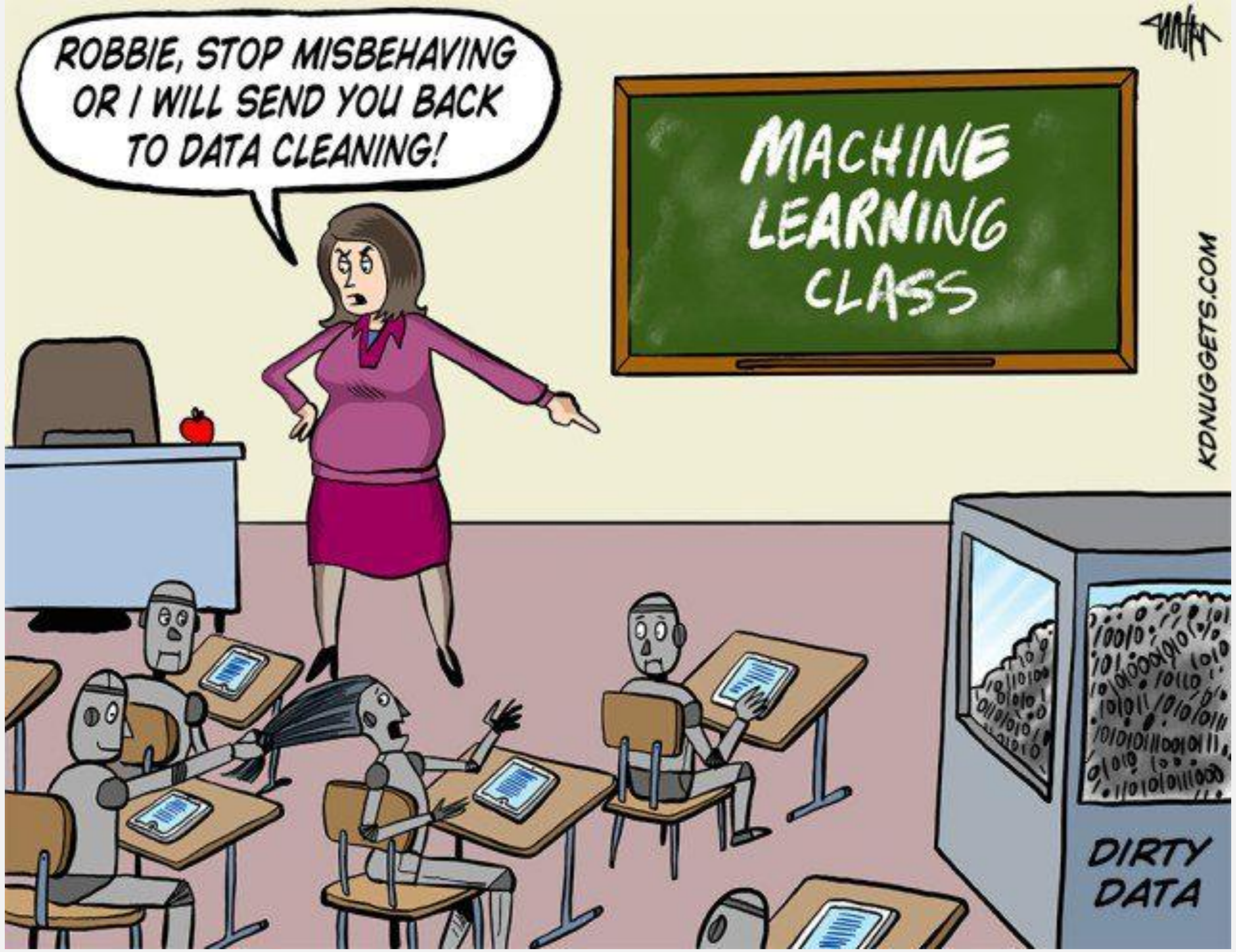
METRICS IN MULTICLASS PROBLEMS

Confusion matrix

True label \ Predicted label	dog	snake	horse	elephant	bear	cat	orangutan	tiger	sloth
dog	544	26	2	0	24	0	18	31	5
snake	37	92	0	0	8	0	1	6	6
horse	3	3	69	34	0	7	9	25	0
elephant	0	0	11	76	0	3	2	8	0
bear	34	13	0	0	111	0	0	1	41
cat	0	0	4	1	0	93	0	27	0
orangutan	17	3	6	1	0	0	80	18	0
tiger	35	5	26	5	0	34	17	253	0
sloth	5	14	0	0	22	0	1	0	158

You can calculate all metrics for all classes, but

the confusion matrix gives you all the information you need


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA