IMPROVING THE FIT

SUPERVISED LEARNING AND RESIDUALS

IRON HACK

IRONHACK MADRID

MACHINE
LEARNNG

ERROR (RESIDUALS) METRICS

Metrics

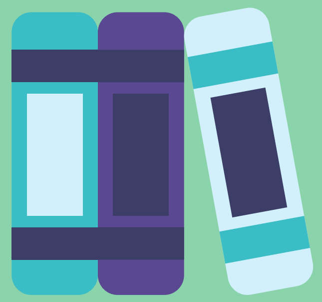
MAE, MSE, RMSE, LOGMSE

Enemies

Underfitting and Overfitting

IDENTIFYING THE PROBLEM

EACH OF THESE PROBLEMS PRESENTS ITSELF DIFFERENTLY. WE MUST TAKE CARE OF BOTH IN ORDER TO BETTER OUR MODELS AND GET THE BEST PREDICTIONS POSSIBLE!



UNDERFIT

An underfit model is a model that does not fit (match) the distribution of our data whatsoever. Therefore it is not very usefull in predicting anything. It means our model is not specific (sensible) enough.

OVERFIT

An overfit model, on the other hand, is too specific. It does match our data perfectly, but only our training data. It will fare terribly against new and unknown data. It means our model is too specific and does not generalize well.

What causes it?

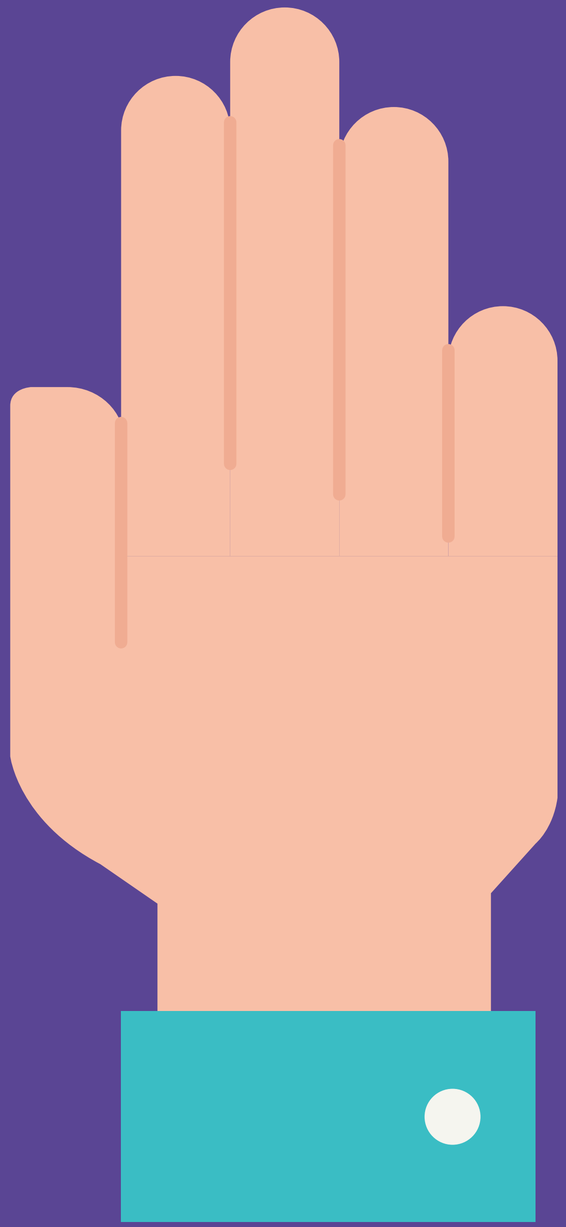
How could we minimize it?

$$\hat{y} = f(x) + \varepsilon$$

General expression for any model

$$\text{Error}(x) = (\text{Bias}[\ \hat{f}(x)\ ])^2 + \text{Var}[\ \hat{f}(x)\ ] + \sigma^2$$

DECOMPOSITION OF ERROR(X) [MSE] INTO BIAS², VARIANCE AND IRREDUCTIBLE ERROR



WHAT IS THE CATCH?

CAN'T CALCULATE

We can't calculate the real values for Bias and Variance because we don't know the true distribution function for the data.

WE CAN ESTIMATE IT, THOUGH

Check this awesome function out.

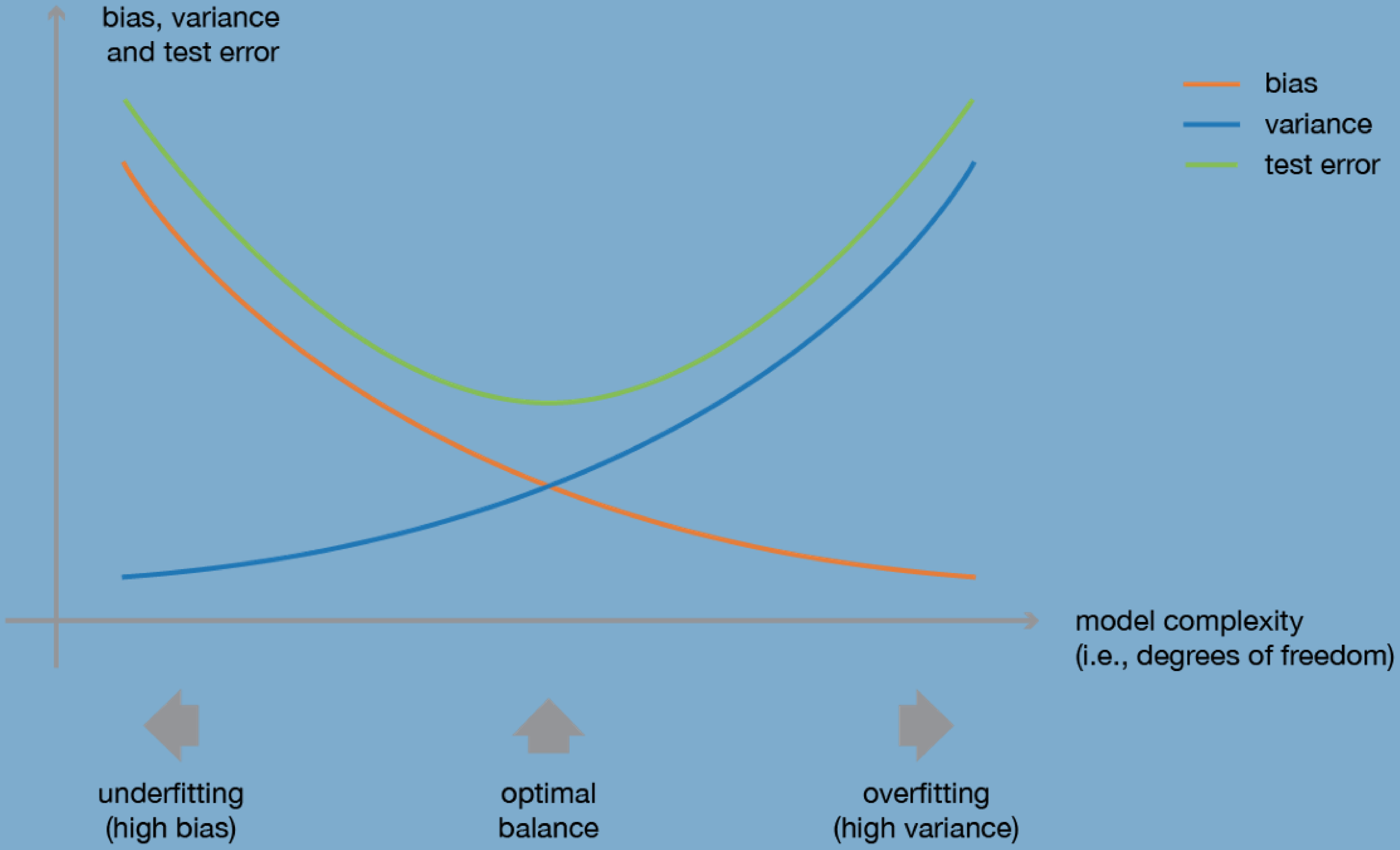
IRREDUCTIBLE

The irreductible part of the error is just.... ireductible by modeling

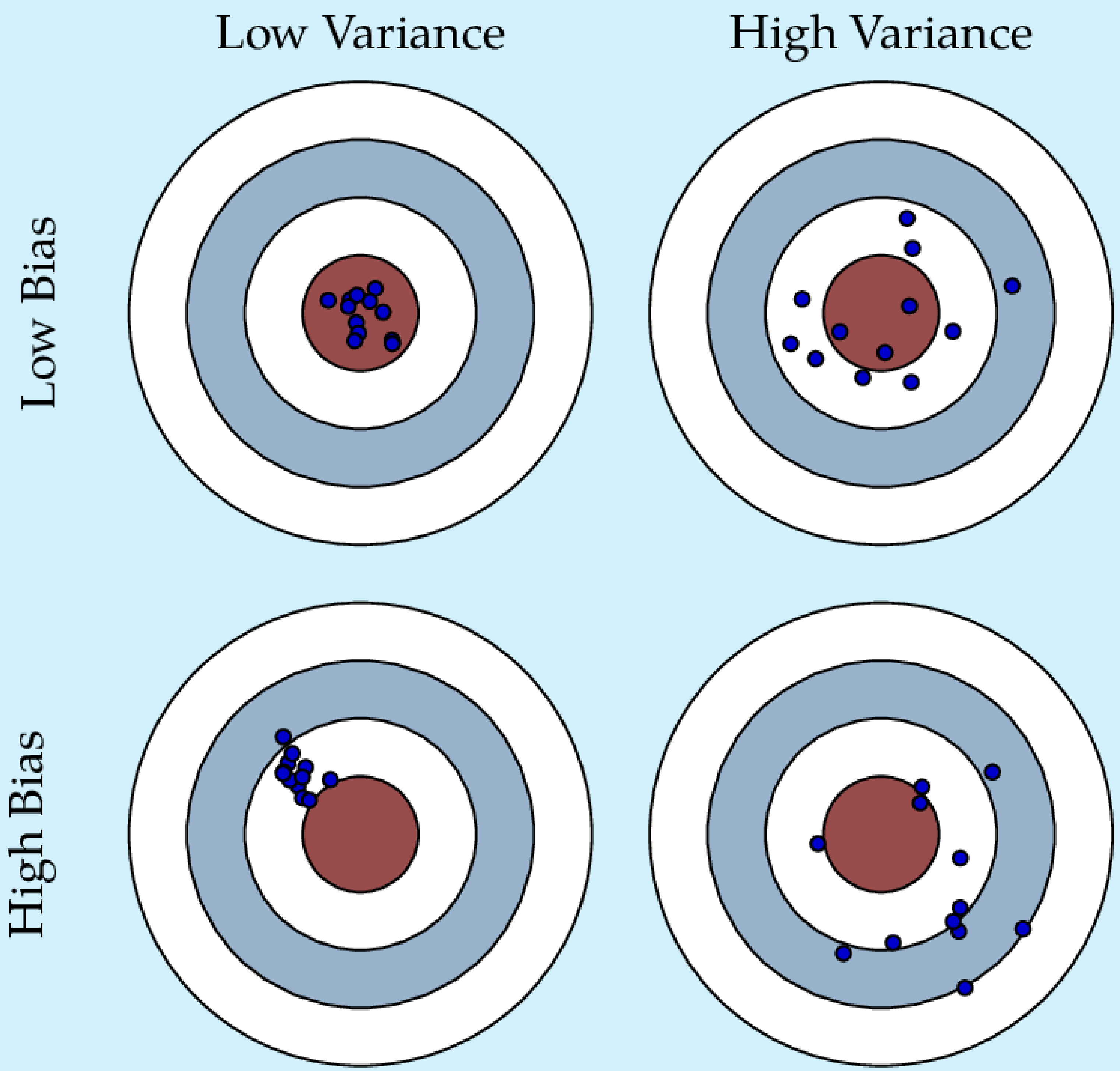
IT IS A TRADEOFF

The lower the bias, higher the variance.
The lower the variance, higher the bias.
Both an vice-verse

BIAS VARIANCE TRADEOFF



Taking our models out to the gun range



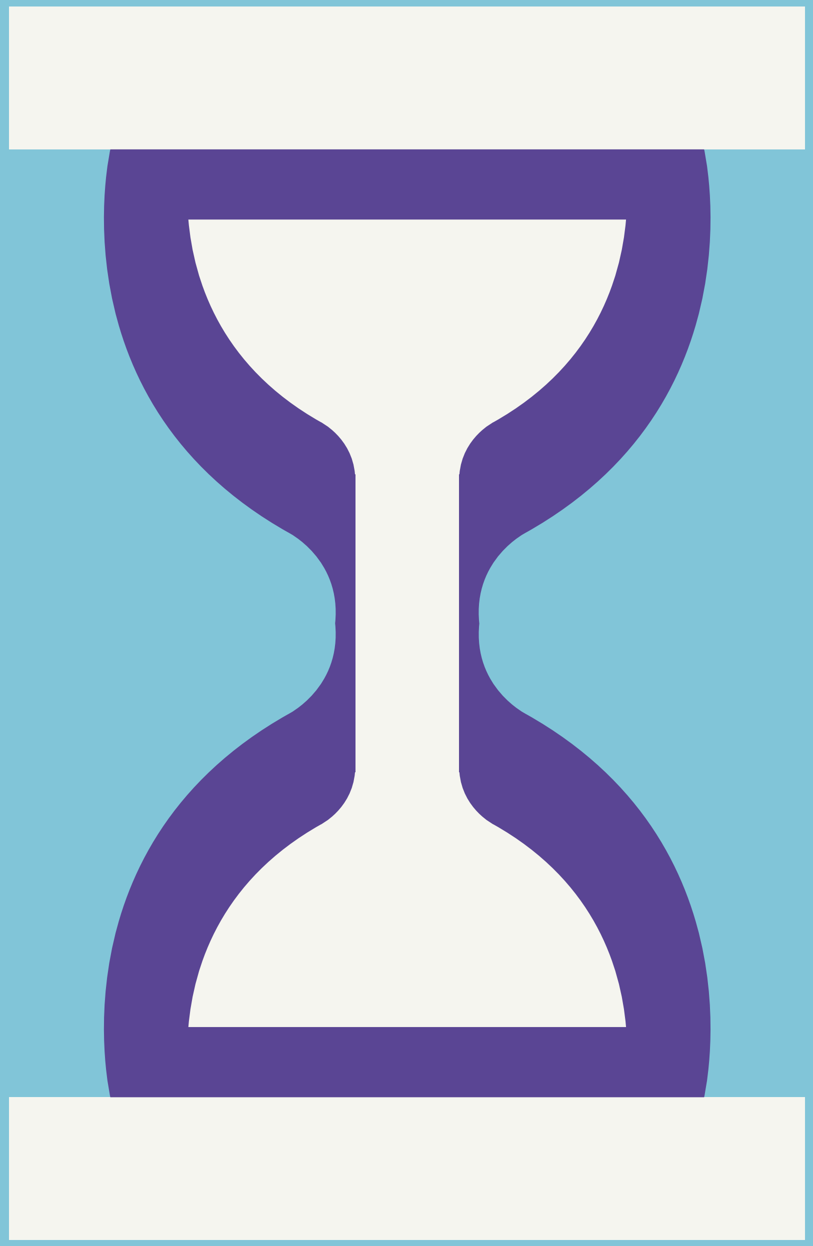
REGULARIZATION

One way to try and avoid overfitting

Penalize large coefficients in your model

Check out more!!!

- L1 regularization
  LASSO
- L2 regularization
  Ridge



$$J(\theta) = MSE(\theta) + \alpha\sum|\theta_i|$$

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

- L1 penalizes sum of absolute value of weights.
- L1 has a sparse solution.
- L1 generates model that are simple and interpretable but cannot learn complex patterns.
- L1 is robust to outliers

RIDGE REGRESSION

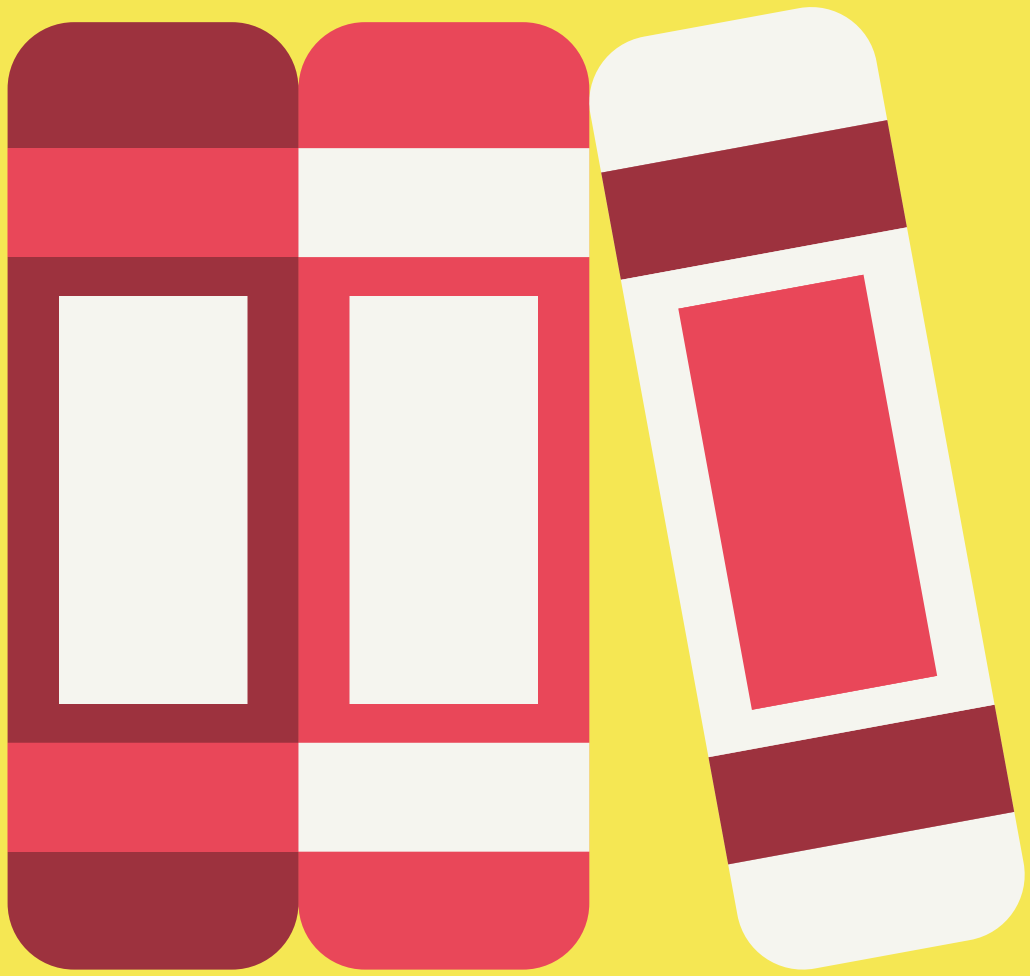
- L2 regularization penalizes sum of square weights.
- L2 has a non sparse solution.
- L2 regularization is able to learn complex data patterns.
- L2 has no feature selection.
- L2 is not robust to outliers

$$J(\theta) = MSE(\theta) + \alpha\sum_i \theta_i^2$$

TRAIN TEST SPLITTING

LET'S HIDE SOME OF THE DATA AND PRETEND IT'S NEW AND UNKNOWN.

TRAIN TEST SPLITTING OUR DATASET ALLOWS US TO GET A METRIC FOR OUR MODEL, NOT ONLY FOR HOW IT WORKS WITH DATA IT IS FAMILIAR WITH, BUT ALSO WITH UNFAMILIAR DATA POINTS

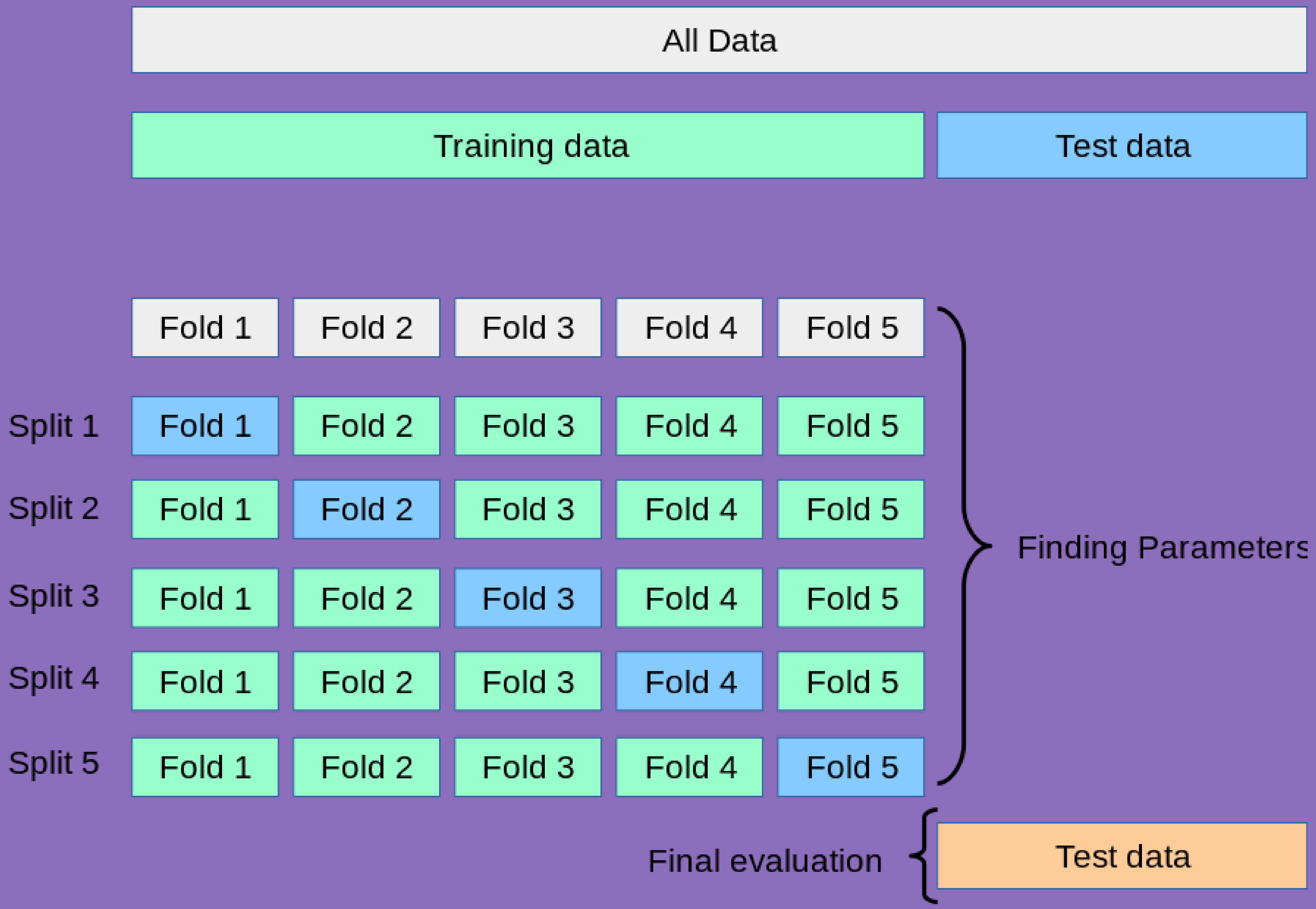


WHAT IF...?

HOW CAN WE BE SURE THAT NO IMPORTANT OR SIGNIFICANT DATA POINT IS LEFT OUT OF EITHER TRAIN OR TEST SET?

JUST DO IT OVER AND OVER AGAIN!

CROSS VALIDATION


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters
Final evaluation
Test data