First, a true story, from
Greenwich Connecticut, 2007

First, a true story, from Greenwich Connecticut, 2007

Financial markets were at all-time highs (this is before the Great Financial Crisis)

Mr V worked at a quant hedge fund as a trader of credit derivatives.

HOW?

Mr V was paid to build financial models, **convince the hedge fund's owner that they were good models**, and then trade them with the HFs' money

Using Backtests of course!

Using Backtests of course!

A backtest "runs" the model on recent market data,
and tells how it performed.

Easy as Pie!!

Err..wasn't the model also built using recent market data?

Err..Yes..

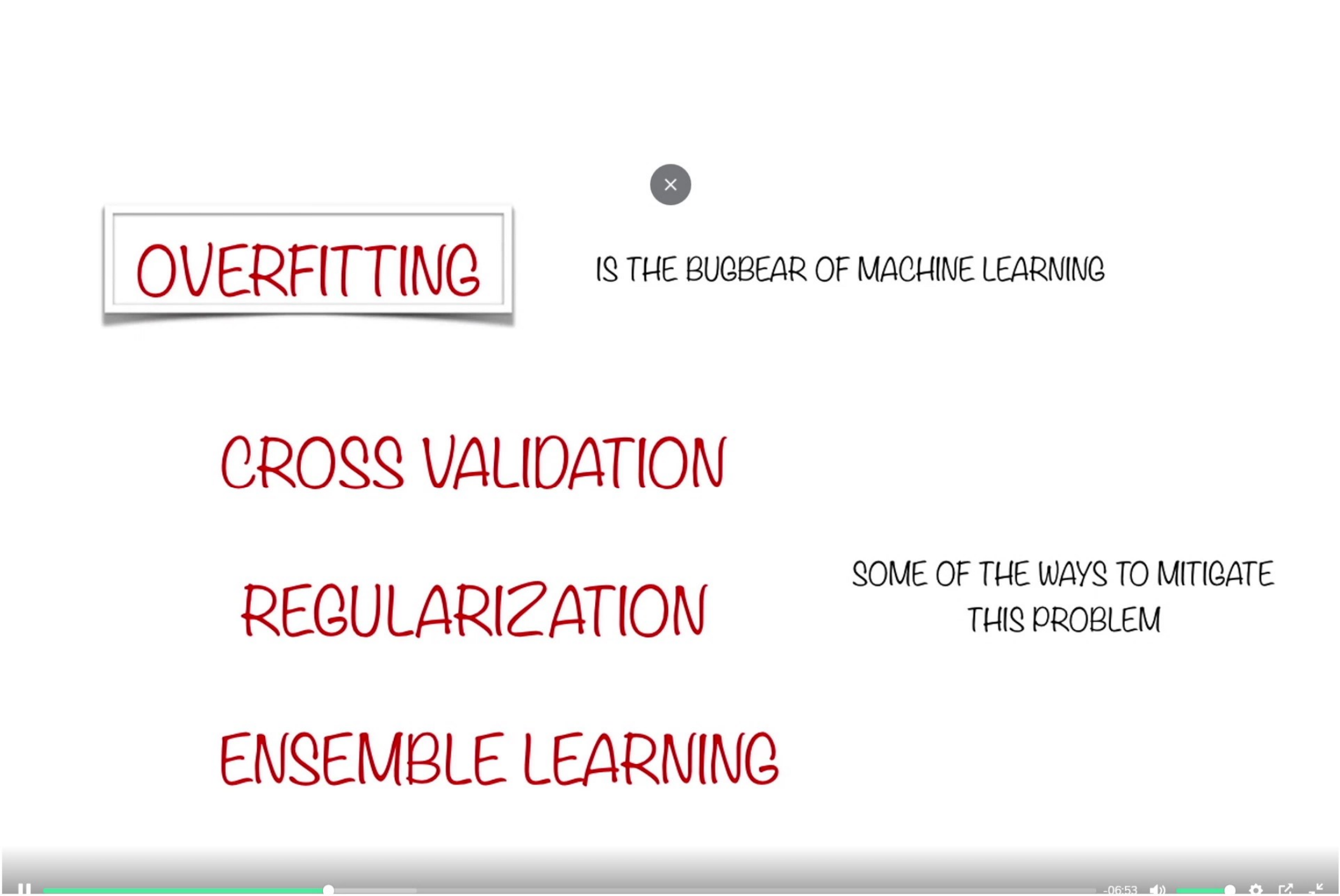
Its really not an exaggeration that Overfitting ML models directly contributed to causing the GFC.

OVERFITTING IS THE BUGBEAR OF MACHINE LEARNING

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING


OVERFITTING
IS THE BUGBEAR OF MACHINE LEARNING
CROSS VALIDATION
REGULARIZATION
ENSEMBLE LEARNING
SOME OF THE WAYS TO MITIGATE
THIS PROBLEM
-06:53

FRODO AND SAM ATE AT A RESTAURANT EVERY DAY LAST WEEK AND RATED IT ON EACH DAY

MONDAY	GOOD
TUESDAY	BAD
WEDNESDAY	GOOD
THURSDAY	GOOD
FRIDAY	GOOD
SATURDAY	BAD
SUNDAY	GOOD

AT THE END OF THE WEEK,

FRODO SAYS

THE FOOD IS GOOD AT THIS RESTAURANT

SAM SAYS

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS EXCEPT TUESDAYS AND SATURDAYS

WHICH ONE OF THEM IS RIGHT?

HOW DO WE MEASURE THIS?

WE COULD CHECK EACH OF THEIR

STATEMENTS

MODELS

AGAINST THE DATA WE ALREADY HAVE

TRAINING SET

WHICH ONE OF THEM IS RIGHT?

	TRAINING SET	FRODO'S MODEL	SAM'S MODEL
MONDAY	GOOD	GOOD	GOOD
TUESDAY	BAD	GOOD	BAD
WEDNESDAY	GOOD	GOOD	GOOD
THURSDAY	GOOD	GOOD	GOOD
FRIDAY	GOOD	GOOD	GOOD
SATURDAY	BAD	GOOD	BAD
SUNDAY	GOOD	GOOD	GOOD

WE COULD CHECK EACH OF
THEIR STATEMENTS
AGAINST THE DATA WE ALREADY HAVE

ACCURACY

WHICH ONE OF THEM IS RIGHT?

	TRAINING SET	FRODO'S MODEL	SAM'S MODEL
MONDAY	GOOD	GOOD	GOOD
TUESDAY	BAD	GOOD	BAD
WEDNESDAY	GOOD	GOOD	GOOD
THURSDAY	GOOD	GOOD	GOOD
FRIDAY	GOOD	GOOD	GOOD
SATURDAY	BAD	GOOD	BAD
SUNDAY	GOOD	GOOD	GOOD
		71%	100%

WE COULD CHECK EACH OF THEIR STATEMENTS AGAINST THE DATA WE ALREADY HAVE

ACCURACY

WHICH ONE OF THEM IS RIGHT?

			FRODO'S MODEL	SAM'S MODEL
WEEK 1	MONDAY	GOOD	GOOD	GOOD
	TUESDAY	BAD	GOOD	BAD
	WEDNESDAY	GOOD	GOOD	GOOD
	THURSDAY	GOOD	71%	100%
	FRIDAY	GOOD	GOOD	GOOD
	SATURDAY	BAD	GOOD	BAD
	SUNDAY	GOOD	GOOD	GOOD
WEEK 2	MONDAY	GOOD	GOOD	GOOD
	TUESDAY	GOOD	GOOD	BAD
	WEDNESDAY	BAD	GOOD	GOOD
	THURSDAY	GOOD	71%	42%
	FRIDAY	GOOD	GOOD	GOOD
	SATURDAY	GOOD	GOOD	BAD
	SUNDAY	BAD	GOOD	GOOD

ON THE TRAINING SET, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 100% ACCURACY

SAM AND FRODO GO BACK TO THE RESTAURANT NEXT WEEK

ON NEW DATA, FRODO'S MODEL HAS 71% ACCURACY AND SAM'S MODEL HAS 42% ACCURACY

WHICH ONE OF THEM IS RIGHT?

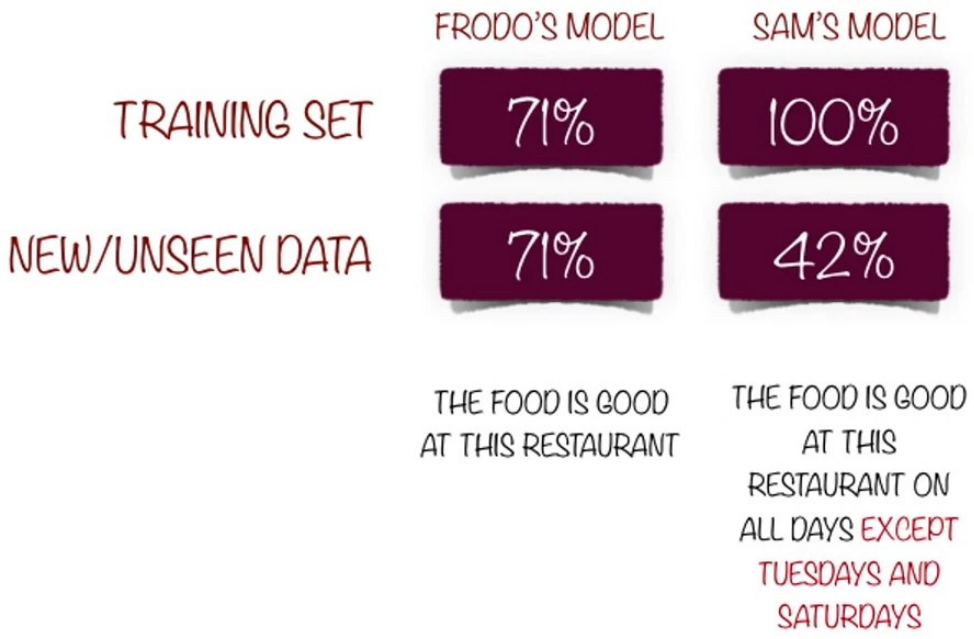
WHAT HAPPENED HERE?

FRODO'S MODEL IS THE BETTER MODEL

IT GENERALIZES WELL

FRODO'S MODEL PERFORMS WELL ON BOTH TRAINING AND NEW/UNSEEN DATA

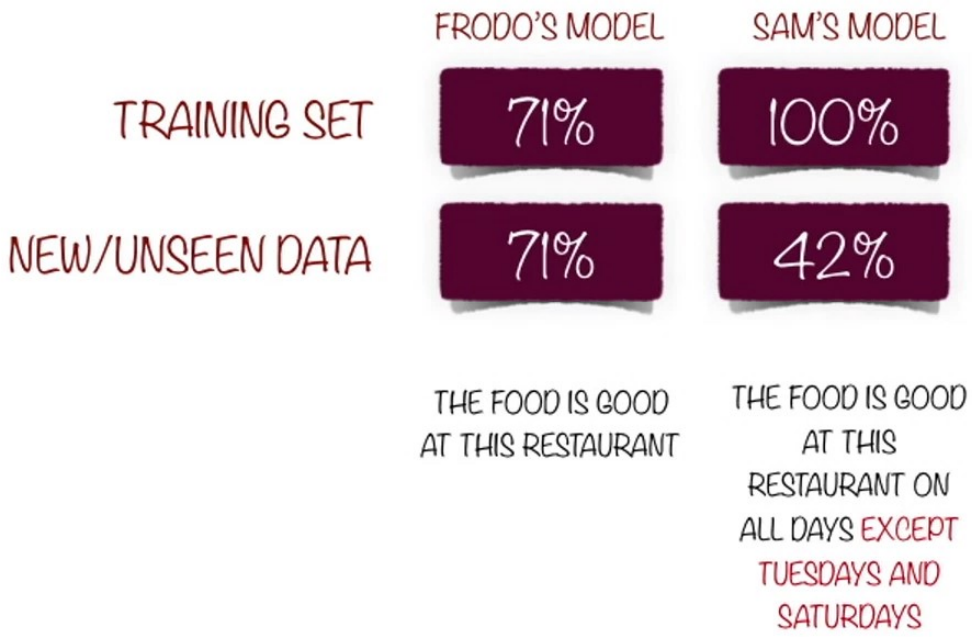
WHAT HAPPENED HERE?



FRODO'S MODEL IS SIMPLER ("DUMBER", IN FACT), YET IT PERFORMS BETTER

SAM'S MODEL IS MORE COMPLEX, AND MORE ACCURATE ON THE TRAINING SET

WHAT HAPPENED HERE?

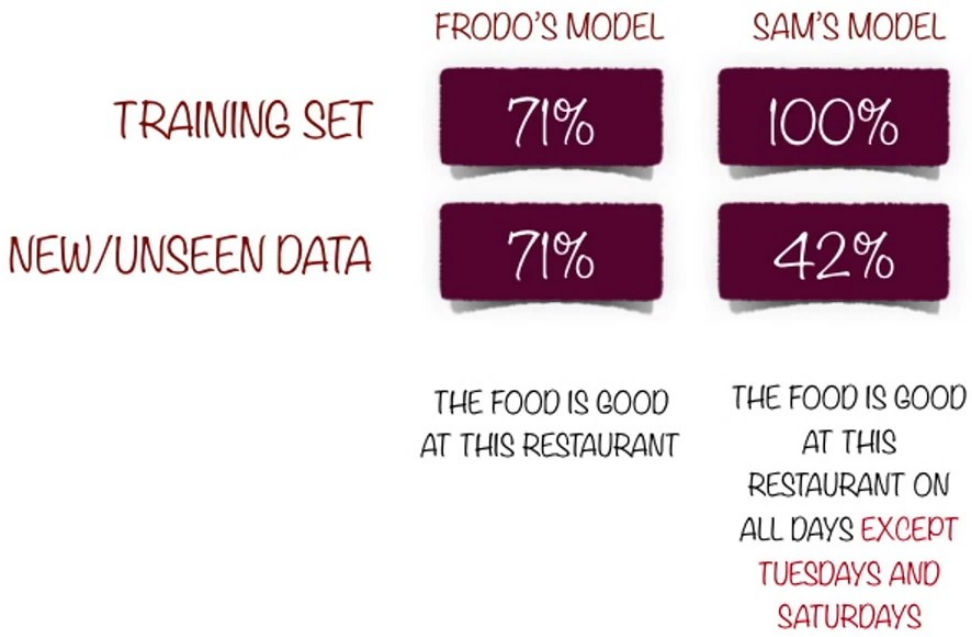


YET, IT PERFORMS BADLY ON NEW DATA

FRODO'S MODEL IS SIMPLER ("DUMBER", IN FACT), YET IT PERFORMS BETTER

SAM'S MODEL IS MORE COMPLEX, AND MORE ACCURATE ON THE TRAINING SET

WHAT HAPPENED HERE?



FRODO'S MODEL IS SIMPLER ("DUMBER", IN FACT), YET IT PERFORMS BETTER

SAM'S MODEL IS MORE COMPLEX, AND MORE ACCURATE ON THE TRAINING SET

YET, IT PERFORMS BADLY ON NEW DATA

IE, SAM'S MODEL DOES NOT GENERALIZE WELL

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS

EXCEPT TUESDAYS AND SATURDAYS

SAM'S MODEL PICKS UP ON A RELATIONSHIP BETWEEN THE WEEKDAY AND THE QUALITY OF FOOD

THIS RELATIONSHIP HOWEVER, IS SPECIFIC TO THE TRAINING SET, AND NOT TRUE IN GENERAL

SAM'S MODEL IS A PERFECT EXAMPLE OF

OVERFITTING

THE FOOD IS GOOD AT THIS RESTAURANT ON ALL DAYS
EXCEPT TUESDAYS AND SATURDAYS

SAM'S MODEL PICKS UP ON A RELATIONSHIP BETWEEN THE WEEKDAY AND THE QUALITY OF FOOD

THIS RELATIONSHIP HOWEVER, IS SPECIFIC TO THE TRAINING SET, AND NOT TRUE IN GENERAL

SAM'S MODEL IS A PERFECT EXAMPLE OF

OVERFITTING

OVERFITTING OCCURS WHEN A MODEL PICKS UP ON RANDOM PHENOMENA OR NOISE PRESENT IN THE TRAINING SET
INSTEAD OF THE UNDERLYING RELATIONSHIP BETWEEN THE INPUT AND OUTPUT

OVERFITTING

BUT WHY IS OVERFITTING SUCH A COMMON PROBLEM?

THE TRAINING SET IS ONLY PART OF A MUCH LARGER SET

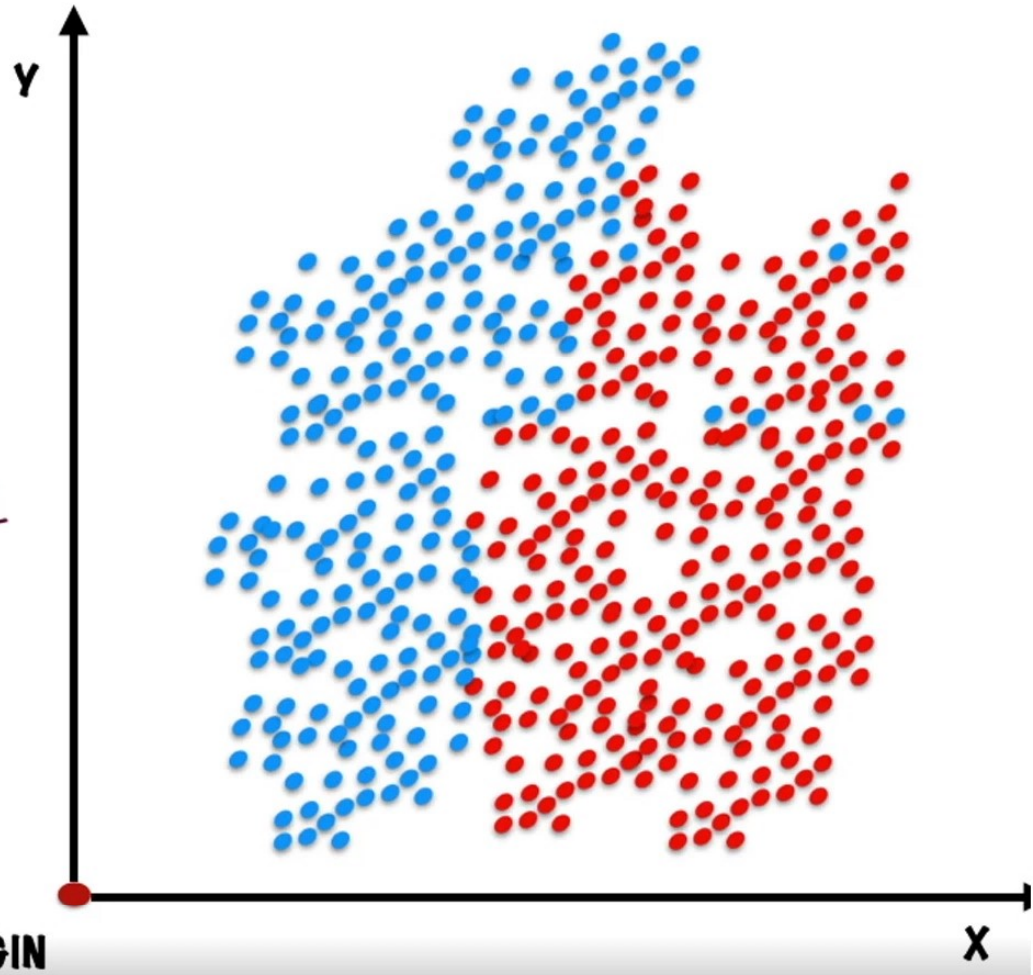
WE ARE TRYING TO FIND A MODEL, THAT DESCRIBES THIS MUCH LARGER SET

IT'S LIKE TRYING TO DESCRIBE PHOTOGRAPH, BUT YOU ARE ONLY SHOWN A SMALL, ZOOMED IN PORTION OF THE PHOTOGRAPH

OVERFITTING

YOU WANT TO CLASSIFY EMAILS AS SPAM OR HAM

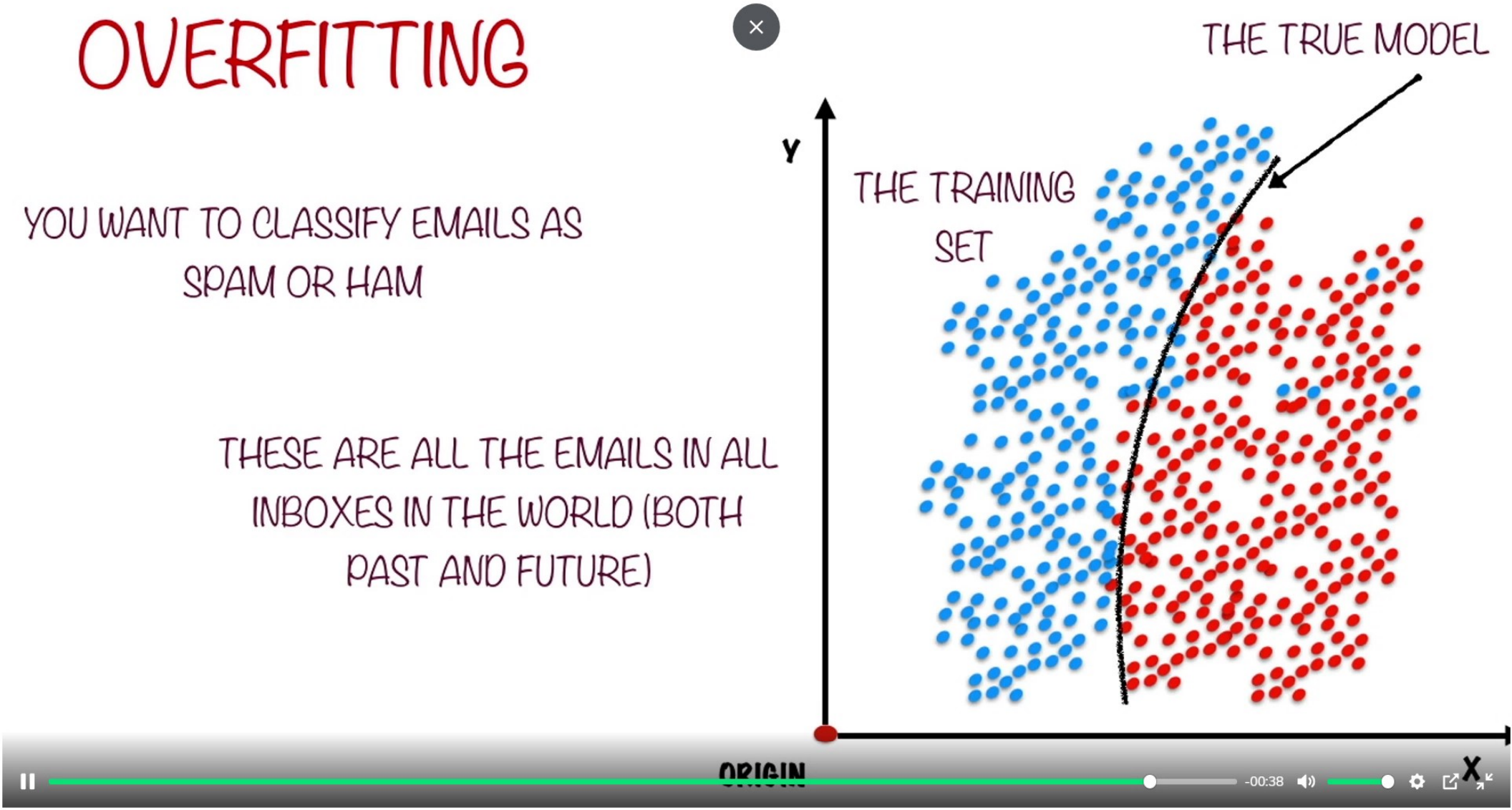
THESE ARE ALL THE EMAILS IN ALL INBOXES IN THE WORLD (BOTH PAST AND FUTURE)



OVERFITTING

YOU WANT TO CLASSIFY EMAILS AS SPAM OR HAM

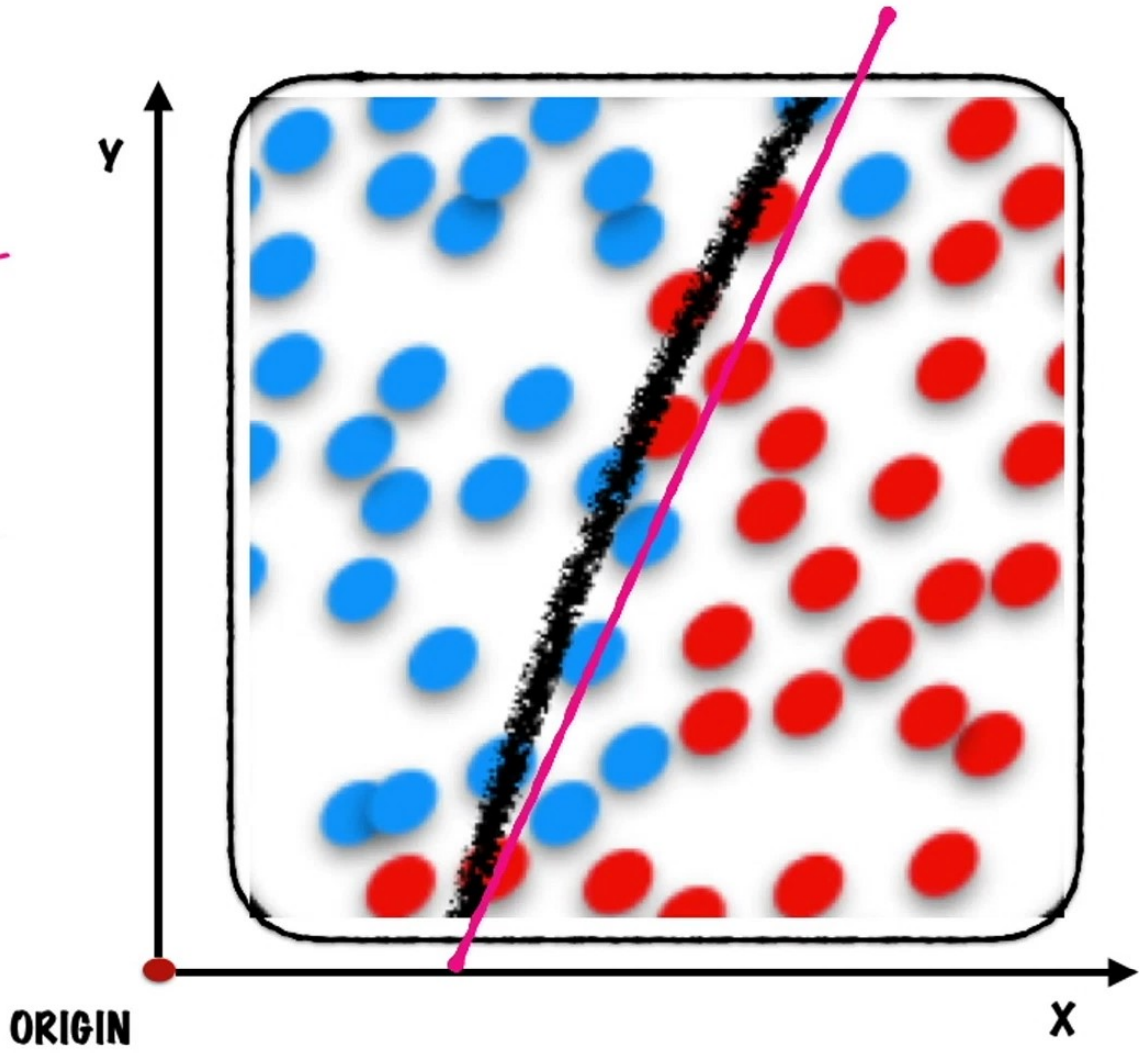
THESE ARE ALL THE EMAILS IN ALL INBOXES IN THE WORLD (BOTH PAST AND FUTURE)



OVERFITTING

1. A SIMPLE LINEAR MODEL

2. AN OVERFITTED MODEL

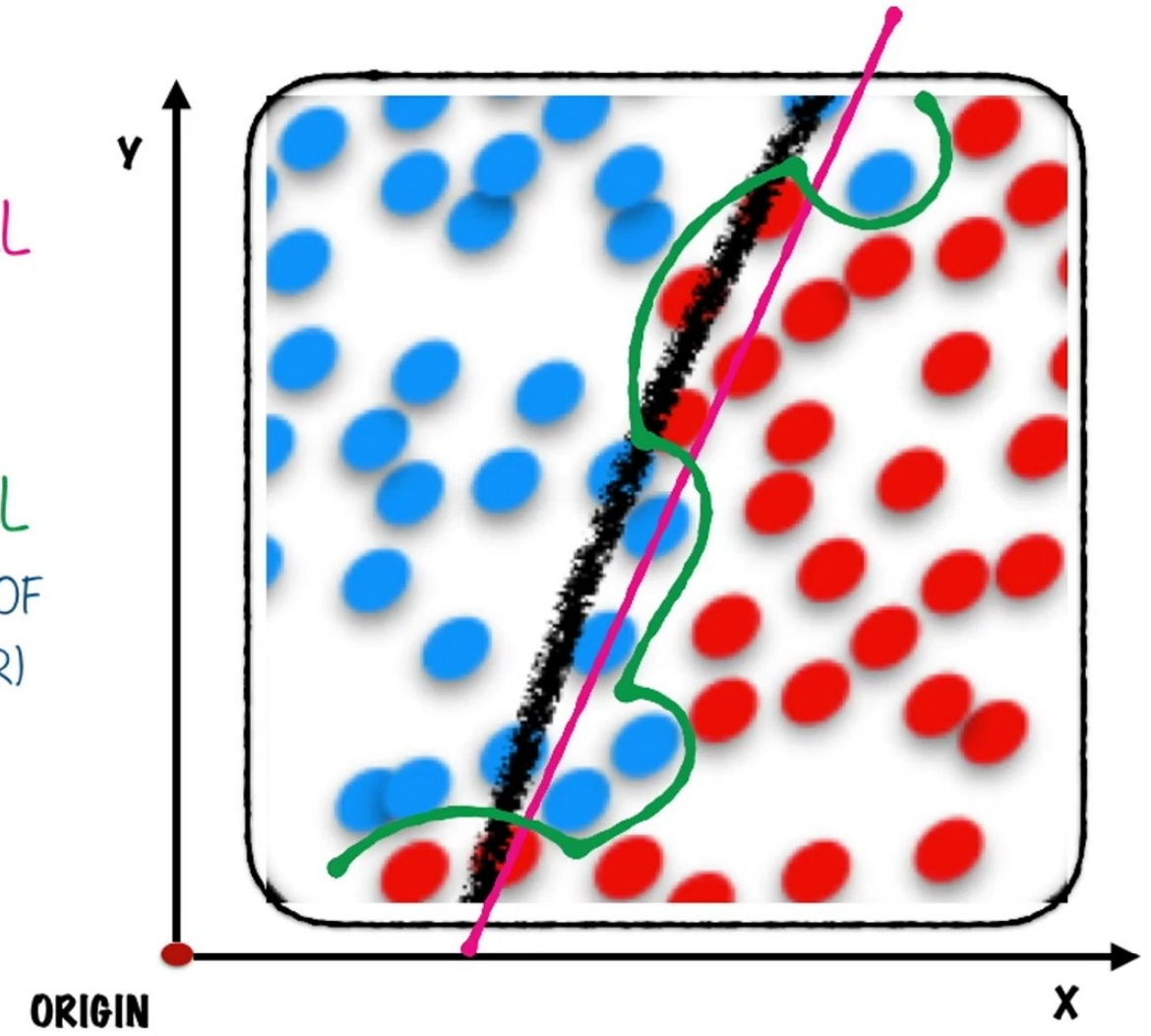


OVERFITTING

1. A SIMPLE LINEAR MODEL

2. AN OVERFITTED MODEL

(USUALLY A POLYNOMIAL OF EXTREMELY HIGH ORDER)



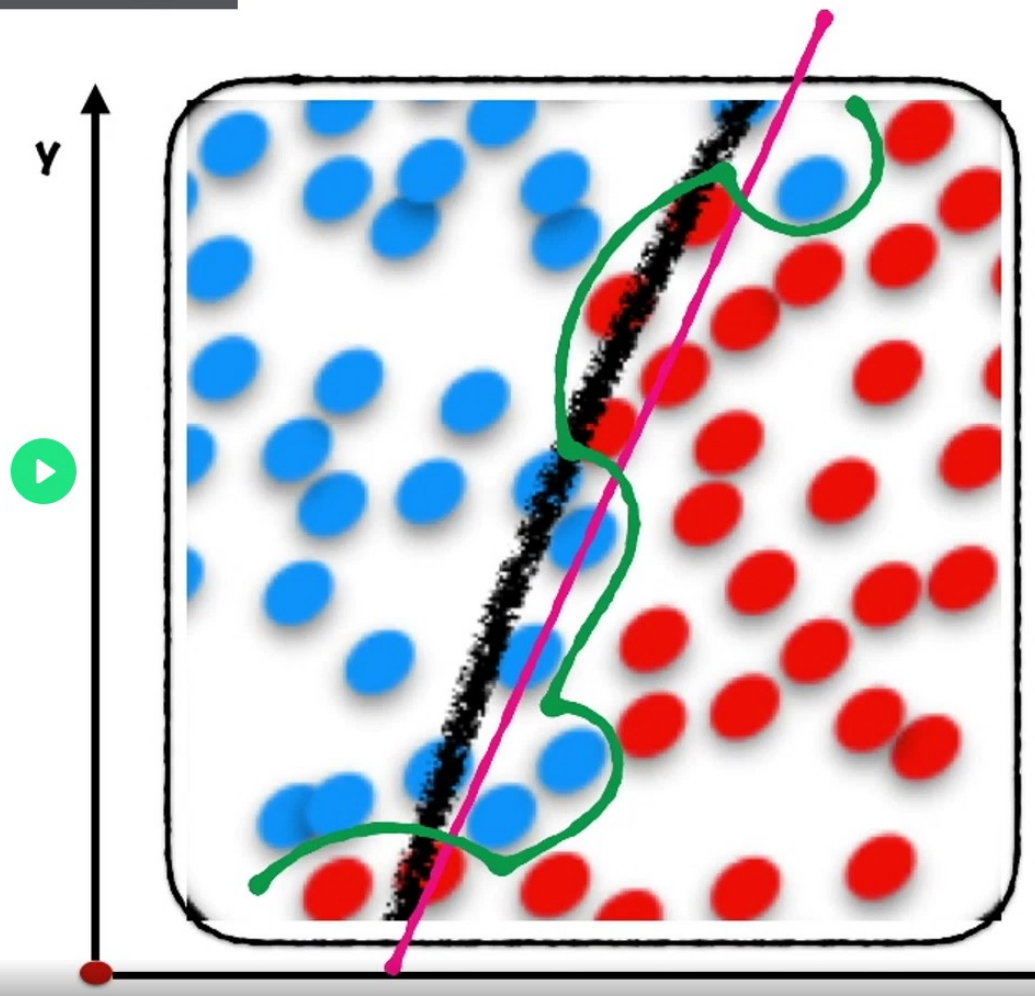
OVERFITTING

To exit full screen, press Esc

1. A SIMPLE LINEAR MODEL

2. AN OVERFITTED MODEL

(USUALLY A POLYNOMIAL OF EXTREMELY HIGH ORDER)



00:00

Key takeaways from this chapter

What is Overfitting? And Why is it a Problem?
Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably".
The root cause of overfitting is tha the training data available to us does not do a good job of representing the real world data.
Which decreses the acuracy of the model when it come to real world data.
Whereas Underfitting of the model can occur where in the model is too general.

Trying to work out a compromise between a overfitted and underfitted model is known as bias variance tradeoff in Machine learning

What is Overfitting

Options

Good performance on the training data, poor generliazation to other data

Poor performance on the training data and poor generalization to other data

What one is not the method of overfitting ?

Options

Cross-Validation

Early Stopping

Data dredging

Regularization

OVERFITTING

IS A PRETTY DIFFICULT
PROBLEM TO SOLVE

BECAUSE THE TRAINING DATA IS ONLY A PART OF THE PICTURE

WE CAN'T TELL FOR SURE WHAT IS RELEVANT AND WHAT'S NOT

OVERFITTING

IS A PRETTY DIFFICULT PROBLEM TO SOLVE

BY AVOIDING OVERFITTING, WE CAN END UP WITH THE OPPOSITE ERROR OF UNDERFITTING

THIS IS THE FAMOUS BIAS-VARIANCE TRADEOFF

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

CROSS VALIDATION

IS A TECHNIQUE FOR MODEL SELECTION

PERFORMING WELL ON TRAINING DATA IS NO GUARANTEE FOR A GOOD MODEL

IN ORDER TO TEST THE PERFORMANCE OF A MODEL, IT WOULD BE NICE IF WE CAN

A GOOD MODEL IS ONE THAT PERFORMS WELL ON DATA IT HAS NOT SEEN BEFORE → GET SOME DATA THAT WE MIGHT SEE IN THE FUTURE (SOME NEW DATA)

A GOOD MODEL DOES NOT OVERFIT → GET MULTIPLE TRAINING DATA SETS

WE CAN THEN FIND A MODEL THAT PERFORMS WELL ACROSS TRAINING DATA SETS, AND NOT JUST ON ONE TRAINING SET

IN ORDER TO TEST THE PERFORMANCE OF A MODEL, IT WOULD BE NICE IF WE CAN

GET SOME DATA THAT WE MIGHT SEE IN THE FUTURE (SOME NEW DATA)

GET MULTIPLE TRAINING DATA SETS

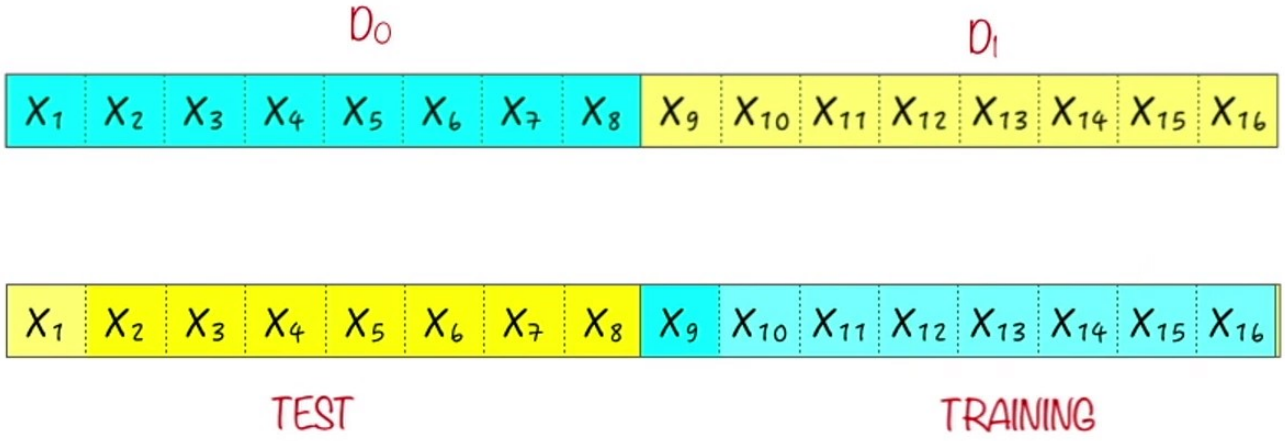
WE CAN THEN FIND A MODEL THAT PERFORMS WELL ACROSS TRAINING DATA SETS, AND NOT JUST ON ONE TRAINING SET

CROSS VALIDATION IS A COMBINATION OF THESE TWO IDEAS

THE BELOW TABLE REPRESENTS THE ENTIRE TRAINING DATA SET

1. DIVIDE THE TRAINING SET RANDOMLY INTO TWO EQUAL PARTS - D_0 AND D_1

2. USE D_0 TO TRAIN THE MODEL AND D_1 TO TEST THE PERFORMANCE



3. THEN, USE D_1 TO TRAIN THE MODEL AND D_0 TO TEST THE PERFORMANCE

THE BEST MODEL IS THE ONE WITH BEST AVERAGE PERFORMANCE

THIS TECHNIQUE IS CALLED

2-FOLD CROSS VALIDATION

WHEN DO YOU USE CROSS VALIDATION?

1. TO CHOOSE BETWEEN DIFFERENT ALGORITHMS

SUPPORT VECTOR MACHINES VS K-NEAREST NEIGHBOURS

2. TO TUNE THE PARAMETERS OF THE ALGORITHM

THE VALUE OF K IN K-NEAREST NEIGHBOURS,
THE MAX DEPTH OF A DECISION TREE

3. TO IDENTIFY THE FEATURES THAT ARE RELEVANT

IF YOU HAVE 20 FEATURES, SHOULD YOU USE ALL OF THEM? OR A SUBSET?

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

ENSEMBLE LEARNING

REGULARIZATION

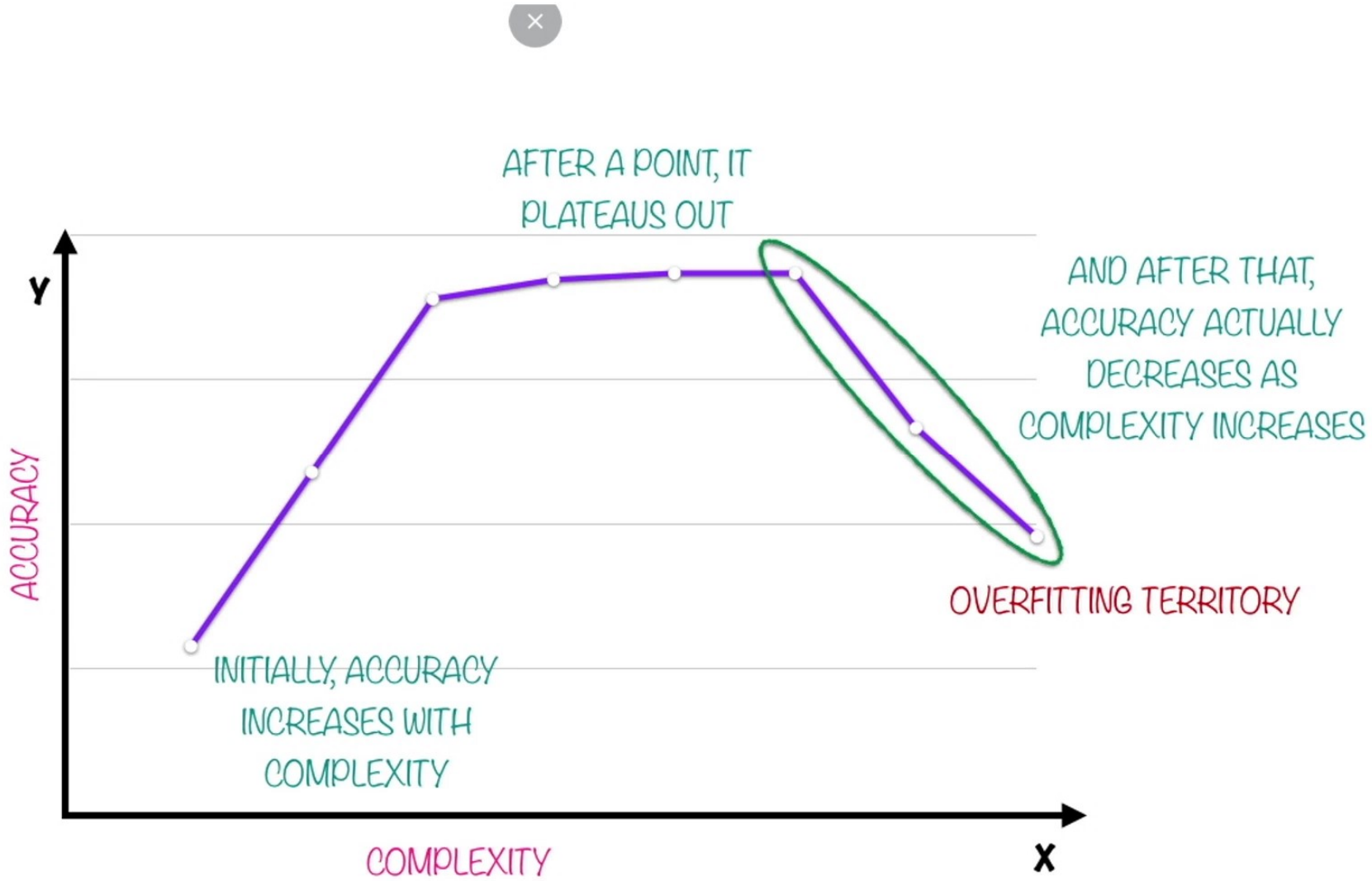
PENALIZES MODELS WHICH ARE TOO COMPLEX

OVERFITTING OCCURS BECAUSE THE MODEL HAS BECOME NEEDLESSLY COMPLEX

EXAMPLES OF COMPLEXITY MEASURES

(THE NUMBER OF BRANCHES IN A DECISION TREE (OR) THE ORDER OF THE POLYNOMIAL USED TO REPRESENT A CURVE)

LET'S SAY YOU PLOTTED COMPLEXITY OF A MODEL VS ACCURACY


AFTER A POINT, IT PLATEAUS OUT
AND AFTER THAT, ACCURACY ACTUALLY DECREASES AS COMPLEXITY INCREASES
OVERFITTING TERRITORY
INITIALLY, ACCURACY INCREASES WITH COMPLEXITY
Y
ACCURACY
COMPLEXITY
X

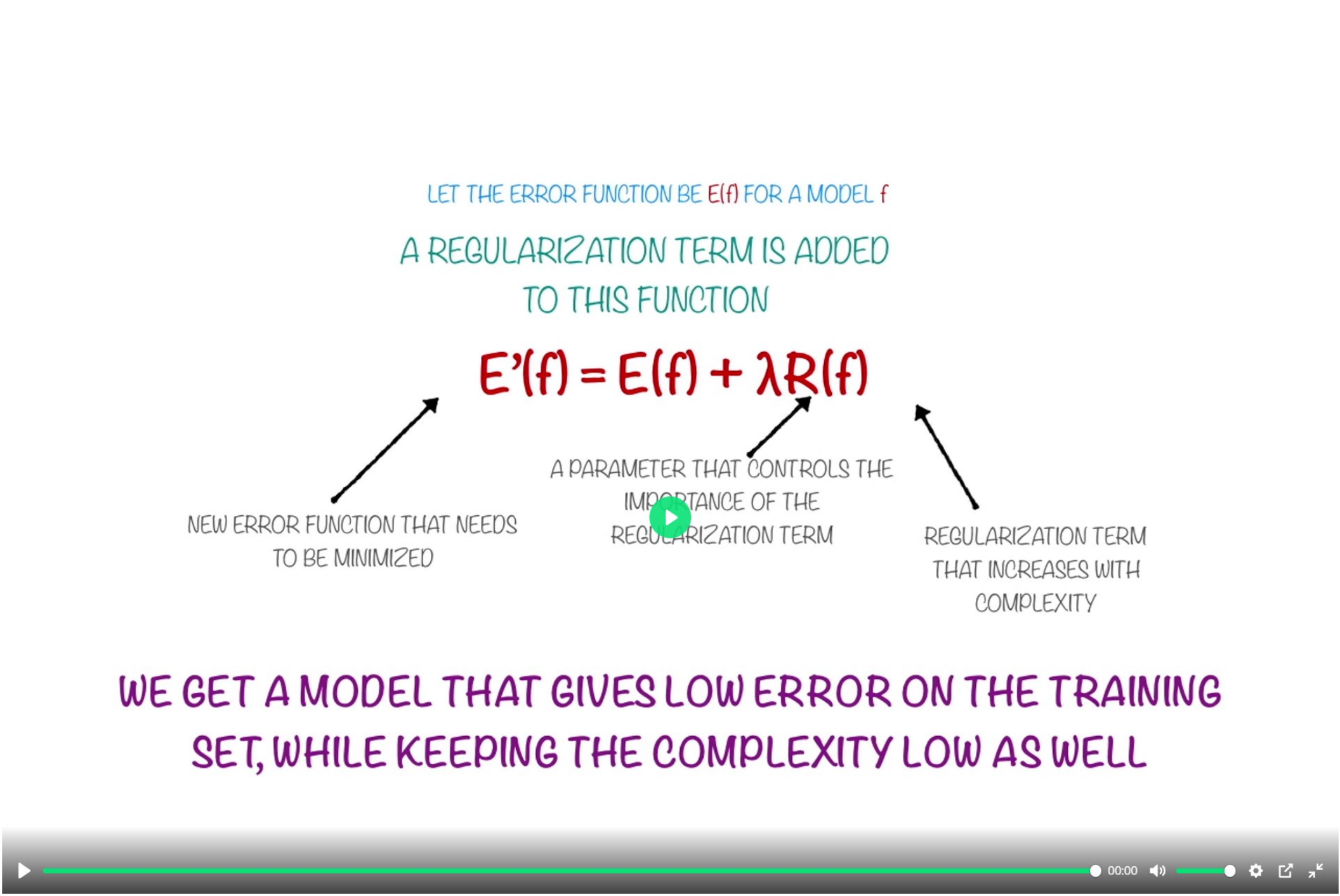
REGULARIZATION

PENALIZES MODELS WHICH ARE TOO COMPLEX

FINDING A MODEL USUALLY INVOLVES MINIMIZING AN ERROR FUNCTION

FOR EXAMPLE, THE ERROR FUNCTION COULD BE THE SUM OF SQUARES OF DISTANCES BETWEEN THE PREDICTED POINTS AND THE ACTUAL POINTS IN THE TRAINING SET

LET THE ERROR FUNCTION BE $E(f)$ FOR A MODEL f


LET THE ERROR FUNCTION BE E(f) FOR A MODEL f
A REGULARIZATION TERM IS ADDED TO THIS FUNCTION
E'(f) = E(f) + λR(f)
NEW ERROR FUNCTION THAT NEEDS TO BE MINIMIZED
A PARAMETER THAT CONTROLS THE IMPORTANCE OF THE REGULARIZATION TERM
REGULARIZATION TERM THAT INCREASES WITH COMPLEXITY
WE GET A MODEL THAT GIVES LOW ERROR ON THE TRAINING SET, WHILE KEEPING THE COMPLEXITY LOW AS WELL

Key takeaways from this chapter

Random Forest Lab: Use an Ensemble of Decision Trees to get Better Results

-Duplicate the earlier created DecisionTree notebook.
-Rename the file as you please.
-remove the attribute part of the notebook
-Import RandomForestClassifier from sklearn
-Set classifier as RandomForestClassifier()
-Check the Accuracy Score.
-use hyperparameter tuning to achieve greater accuracy
-Continue tweaking the parameter until you get a greater accuracy score for your dataset.

Which of the following options is/are true for K-fold cross-validation?

1.Increase in K will result in higher time required to cross validate the result.
2.Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.
3.If K=N, then it is called Leave one out cross validation, where N is the number of observations.

Options

1 and 1

2 and 3

1 and 3

1,2 and 3

Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6.
Now consider the points below and choose the option based on these points.
Note: All other hyper parameters are same and other factors are not affected.
Depth 4 will have high bias and low variance
Depth 4 will have low bias and low variance

Options

Only 1

Only 2

Both 1 and 2

None of the above

Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6.
Now consider the points below and choose the option based on these points.
Note: All other hyper parameters are same and other factors are not affected.
Depth 4 will have high bias and low variance
Depth 4 will have low bias and low variance

Options

Only 1

Only 2

Both 1 and 2

None of the above

For k cross validation,smaller k value implies less variance.

Options

True

False

For an image recognition problem (recognizing a cat in a photo), which architecture of neural network would be better suited to solve the problem?

Options

- Convolution Neural Network
- Recurrent Neural Network
- Multi layer Neural Network
- Perceptron

To exit full screen, press Esc

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

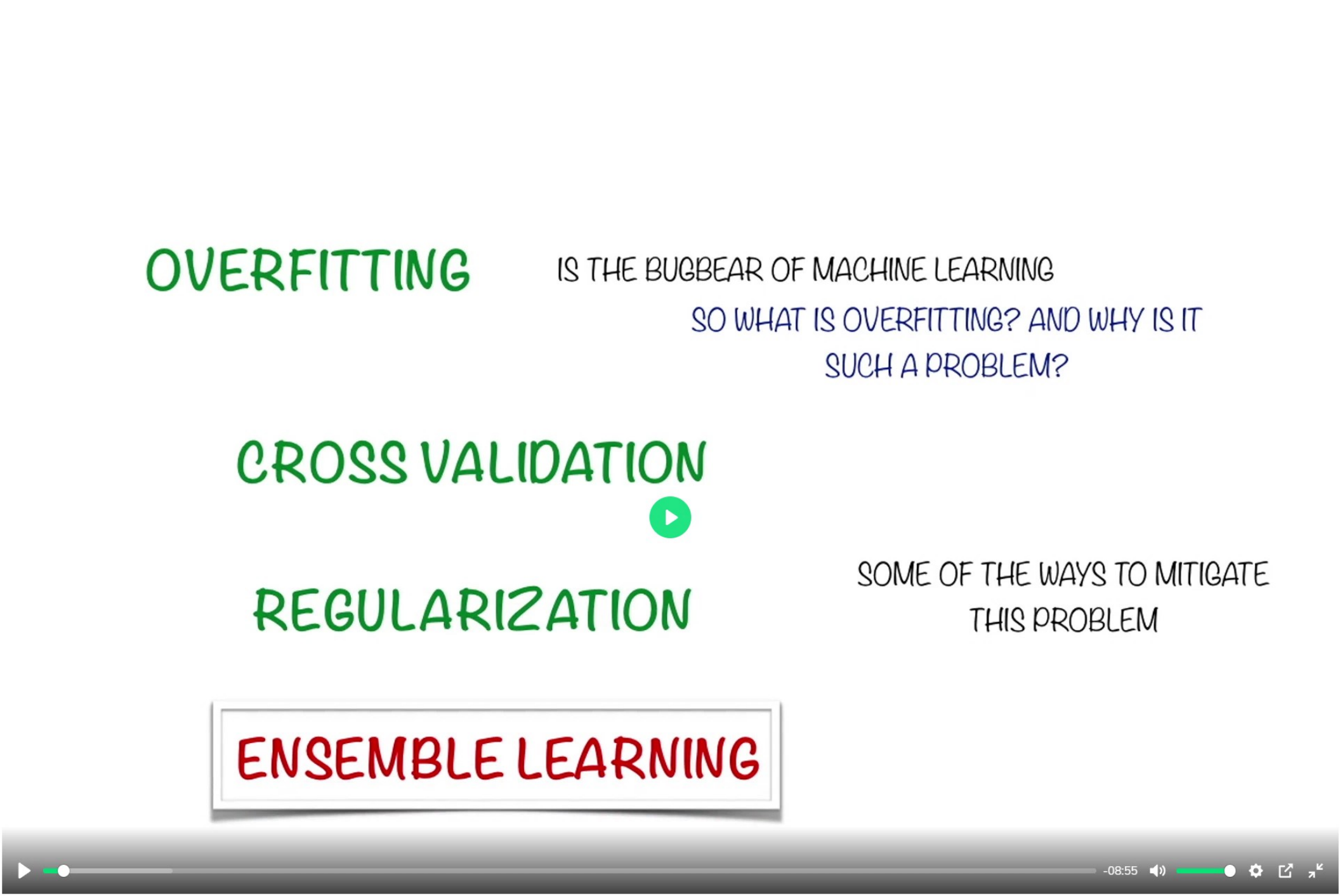
SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

REGULARIZATION

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

ENSEMBLE LEARNING


OVERFITTING
IS THE BUGBEAR OF MACHINE LEARNING
SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?
CROSS VALIDATION
REGULARIZATION
SOME OF THE WAYS TO MITIGATE THIS PROBLEM
ENSEMBLE LEARNING
-08:55

ENSEMBLE LEARNING

INVOLVES THE USE OF MULTIPLE LEARNERS AND COMBINING THEIR RESULTS

IN 2006, NETFLIX HELD AN OPEN COMPETITION FOR A MACHINE LEARNING ALGORITHM TO PREDICT A USER'S RATING OF A MOVIE

THE GRAND PRIZE WAS A COOL MILLION !

THE COMPETITION WENT ON FOR 3 YEARS, BEFORE A GRAND PRIZE WINNER WAS DECLARED

AN INTERESTING THING HAPPENED DURING THIS TIME...

THE CONTESTANTS FOUND THAT, INSTEAD OF USING 1 SINGLE MODEL, COMBINING MULTIPLE MODELS WORKED BETTER

TEAMS STARTED MERGING INTO LARGER TEAMS, THEY WOULD COMBINE THEIR MODELS TO DO BETTER

IN THE END, THE GRAND PRIZE WINNER (AND A VERY CLOSE RUNNER UP) WERE BOTH ENSEMBLES OF MORE THAN A 100 LEARNERS EACH..

AND COMBINING THEM IMPROVED THE RESULTS EVEN FURTHER!

THE IDEA OF ENSEMBLE LEARNING IS SIMPLE..

LET'S TAKE AN EXAMPLE

CLASSIFY A TWEET AS POSITIVE OR NEGATIVE SENTIMENT
(THIS IS A CLASSIFICATION PROBLEM)

METHOD 1. CHOOSE 1 TECHNIQUE

NAIVE BAYES (OR) SUPPORT VECTOR MACHINES (OR) NEURAL NETWORKS

METHOD 2. USE AN ENSEMBLE

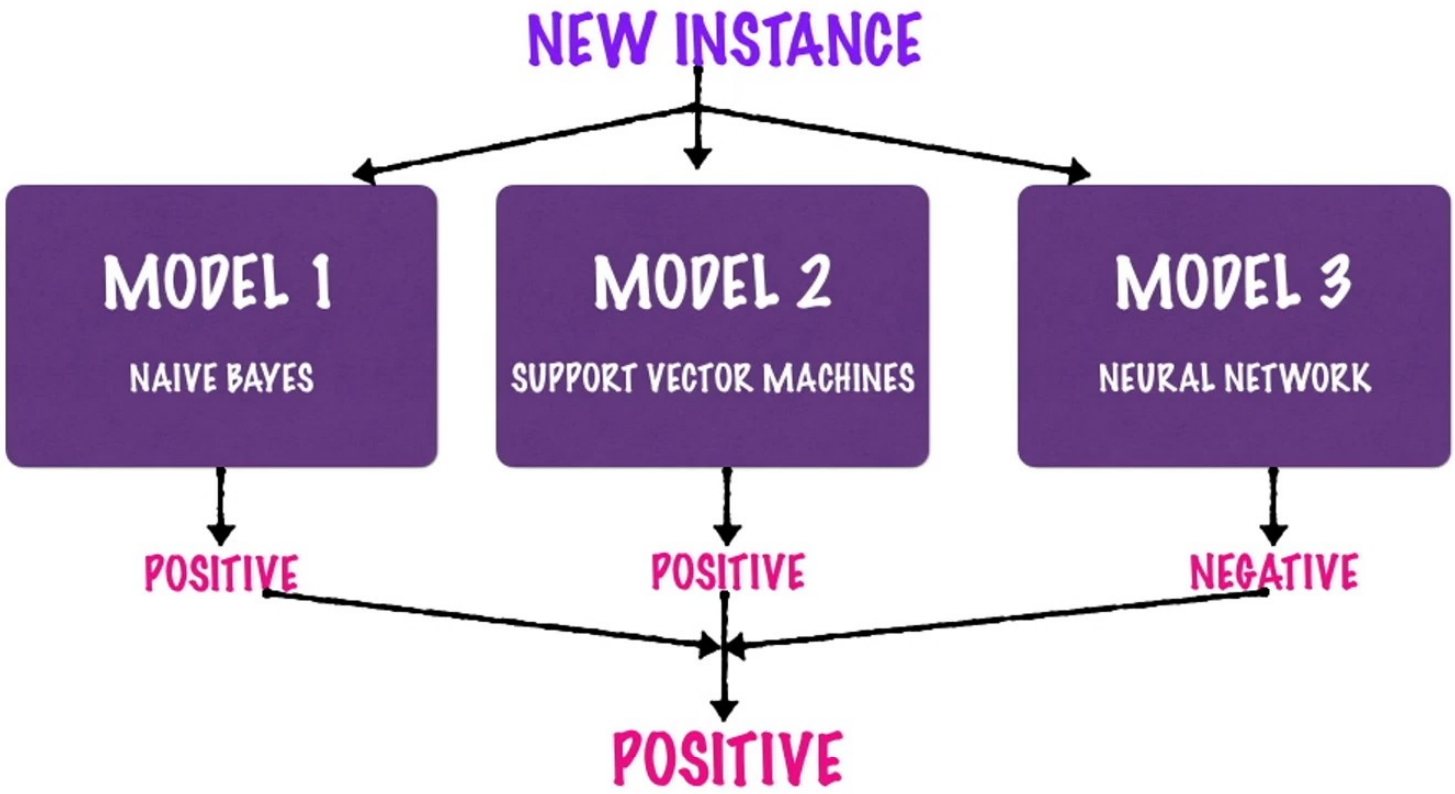
NAIVE BAYES (AND) SUPPORT VECTOR MACHINES (AND) NEURAL NETWORKS

METHOD 2. USE AN ENSEMBLE NAIVE BAYES (AND) SUPPORT VECTOR MACHINES (AND) NEURAL NETWORKS

1. TAKE THE TRAINING SET AND TRAIN EACH OF THE ABOVE CLASSIFIERS ON IT

2. WHEN A NEW INSTANCE (TWEET) COMES IN, GET THE PREDICTIONS FROM EACH OF THE MODELS

3. TAKE THE MAJORITY VOTE OF THE MODELS AND THAT WILL BE THE FINAL PREDICTION

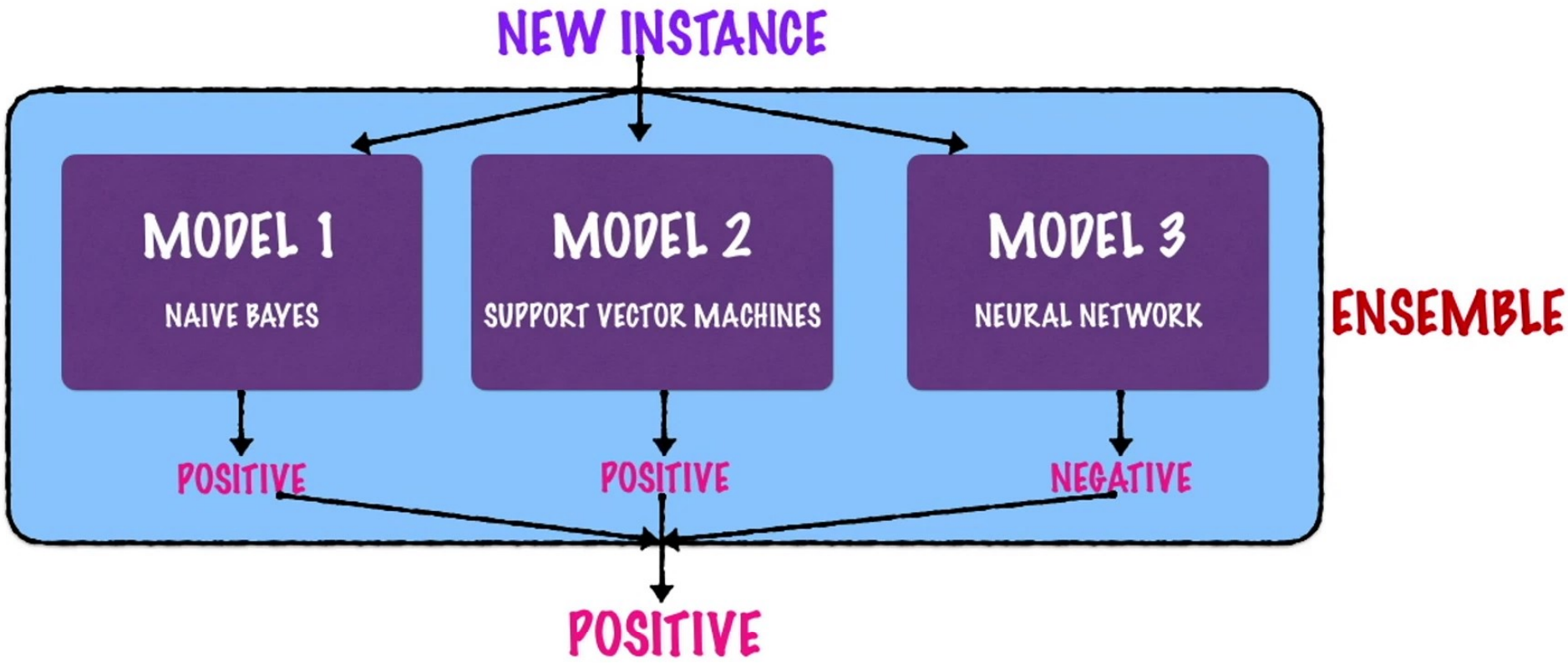


METHOD 2. USE AN ENSEMBLE NAIVE BAYES (AND) SUPPORT VECTOR MACHINES (AND) NEURAL NETWORKS

1. TAKE THE TRAINING SET AND TRAIN EACH OF THE ABOVE CLASSIFIERS ON IT

2. WHEN A NEW INSTANCE (TWEET) COMES IN, GET THE PREDICTIONS FROM EACH OF THE MODELS

3. TAKE THE MAJORITY VOTE OF THE MODELS AND THAT WILL BE THE FINAL PREDICTION



A MACHINE LEARNING ENSEMBLE IS A COLLECTION OF MODELS

THE MODELS IN THE ENSEMBLE CAN BE

BASED ON DIFFERENT TECHNIQUES

A COLLECTION WITH 1 SVM, 1 DECISION TREE, 1 NAIVE BAYES, 1 KNN

TRAINED ON DIFFERENT TRAINING SETS

USING DIFFERENT FEATURES

USING DIFFERENT VALUES OF PARAMETERS

A MACHINE LEARNING ENSEMBLE IS A COLLECTION OF MODELS

THE MODELS IN THE ENSEMBLE CAN BE

BASED ON DIFFERENT TECHNIQUES

A COLLECTION WITH 1 SVM, 1 DECISION TREE, 1 NAIVE BAYES, 1 KNN

TRAINED ON DIFFERENT TRAINING SETS

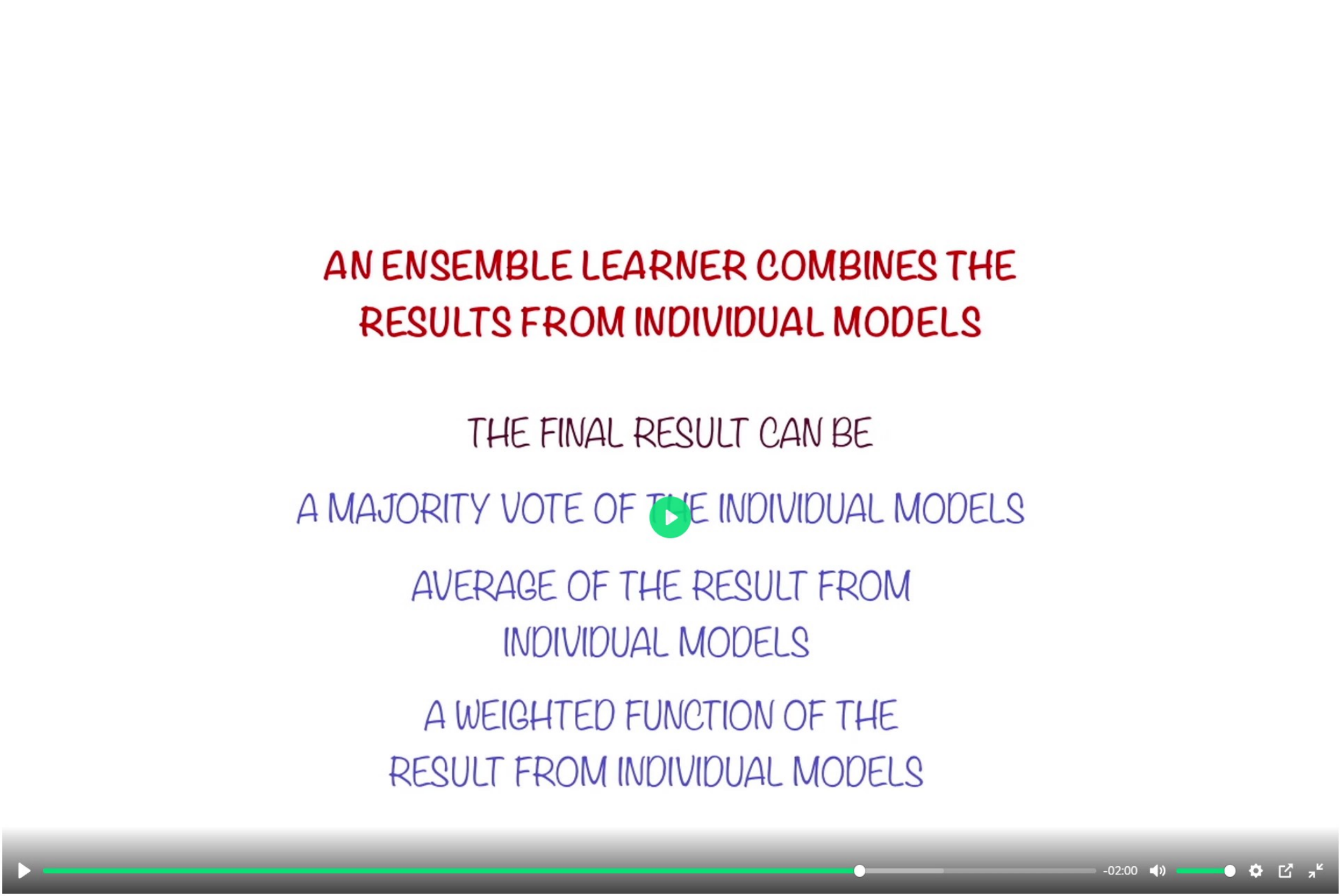
A COLLECTION OF SVMS, EACH TRAINED ON A DIFFERENT TRAINING SET

USING DIFFERENT FEATURES

A COLLECTION OF DECISION TREES, EACH GIVEN A DIFFERENT SET OF FEATURES

USING DIFFERENT VALUES OF PARAMETERS

A COLLECTION OF K-NEAREST NEIGHBOURS, EACH WITH A DIFFERENT VALUE OF K


AN ENSEMBLE LEARNER COMBINES THE RESULTS FROM INDIVIDUAL MODELS
THE FINAL RESULT CAN BE
A MAJORITY VOTE OF THE INDIVIDUAL MODELS
AVERAGE OF THE RESULT FROM INDIVIDUAL MODELS
A WEIGHTED FUNCTION OF THE RESULT FROM INDIVIDUAL MODELS
-02:00

OVERFITTING

IS THE BUGBEAR OF MACHINE LEARNING

SO WHAT IS OVERFITTING? AND WHY IS IT SUCH A PROBLEM?

CROSS VALIDATION

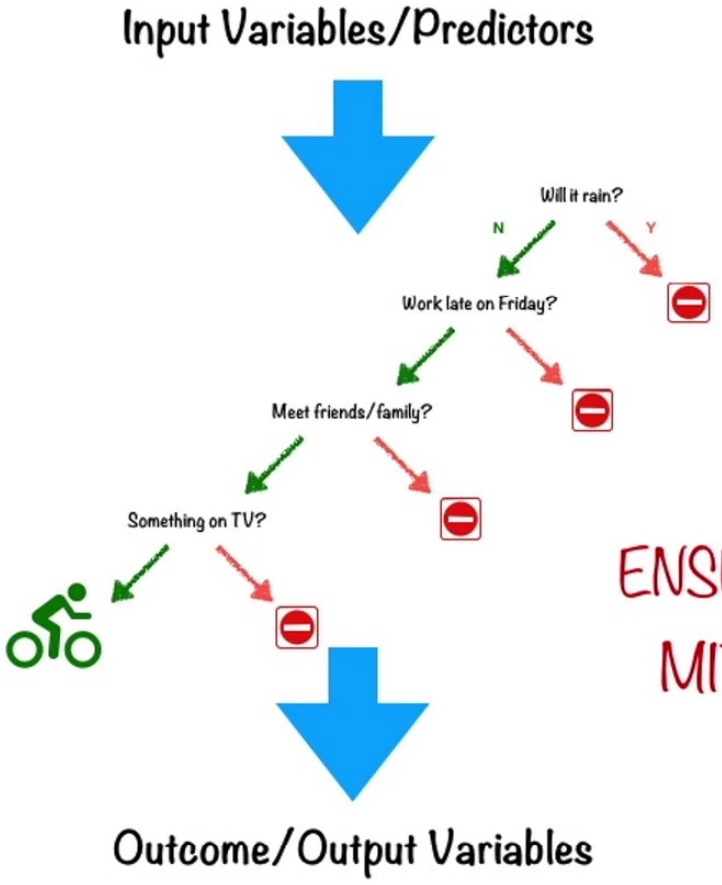
REGULARIZATION

ENSEMBLE LEARNING

SOME OF THE WAYS TO MITIGATE THIS PROBLEM

Decision Tree

DECISION TREES ARE VERY PRONE TO THE RISK OF OVERFITTING



ENSEMBLE LEARNING CAN MITIGATE THE RISK OF OVERFITTING

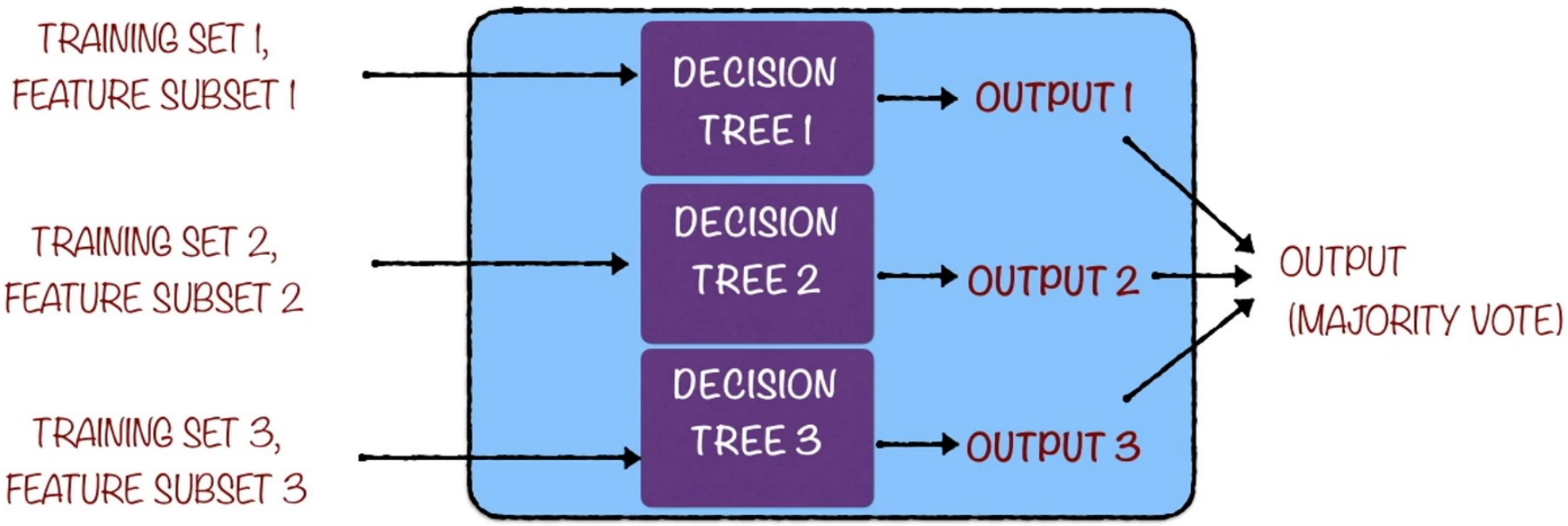
A RANDOM FOREST IS AN ENSEMBLE OF DECISION TREES

EACH DECISION TREE IN THE ENSEMBLE IS

TRAINED ON DIFFERENT TRAINING SETS

USING DIFFERENT FEATURES
(A RANDOMLY SELECTED SUBSET OF FEATURES)

Random Forest



Key takeaways from this chapter

Teamwork: How Ensembles like Random Forest Mitigate the Problem of Overfitting
Ensemble learning is just a way of combining the output of multiple models in order to get a better result. This is a cheap way of improving your model, so i would bet that most real world applications use some sort of ensemble.
This technique became popular after the Netflix Challenge where the winning teams used ensembles of a lot of models to win the competition. This challenge aimed to build a new recommender system for Netflix to propose new movies for their users.

If you use an ensemble of different base models, is it necessary to tune the hyper parameters of all base models to improve the ensemble performance?

Options

Yes

No

cant say

Which of the following algorithm is not an example of an ensemble method?

Options

Extra Tree Regressor

Random Forest

Gradient Boosting

Decision Tree

Key takeaways from this chapter

Teamwork: How Ensembles like Random Forest Mitigate the Problem of Overfitting
Ensemble learning is just a way of combining the output of multiple models in order to get a better result. This is a cheap way of improving your model, so i would bet that most real world applications use some sort of ensemble.
This technique became popular after the Netflix Challenge where the winning teams used ensembles of a lot of models to win the competition. This challenge aimed to build a new recommender system for Netflix to propose new movies for their users.

Avoiding Overfitted Models - Cross Validation and Regularization

Techinques for Mitigating Overfitting
Cross Validation:

In Machine Learning, Cross-validation is a resampling method used for model evaluation to avoid testing a model on the same dataset on which it was trained. This is a common mistake, especially that a separate testing dataset is not always available. However, this usually leads to inaccurate performance measures (as the model will have an almost perfect score since it is being tested on the same data it was trained on). To avoid this kind of mistakes, cross validation is usually preferred.
The concept of cross validation is actually simple: Instead of using the whole dataset to train and then test on same data, we could randomly divide our data into training and testing datasets.

When to use Cross Validation?
1. TO CHOOSE BETWEEN DIFFERENT ALGORITHMS
   For Eg:SUPPORT VECTOR MACHINES VS K-NEAREST NEIGHBOURS
2. TO TUNE THE PARAMETERS OF THE ALGORITHM
   For Eg:THE VALUE OF K IN K-NEAREST NEIGHBOURS,THE MAX DEPTH OF A DECISION TREE
3. TO IDENTIFY THE FEATURES THAT ARE RELEVANT
   For Eg:IF YOU HAVE 20 FEATURES, SHOULD YOU USE ALL OF
   THEM? OR A SUBSET?

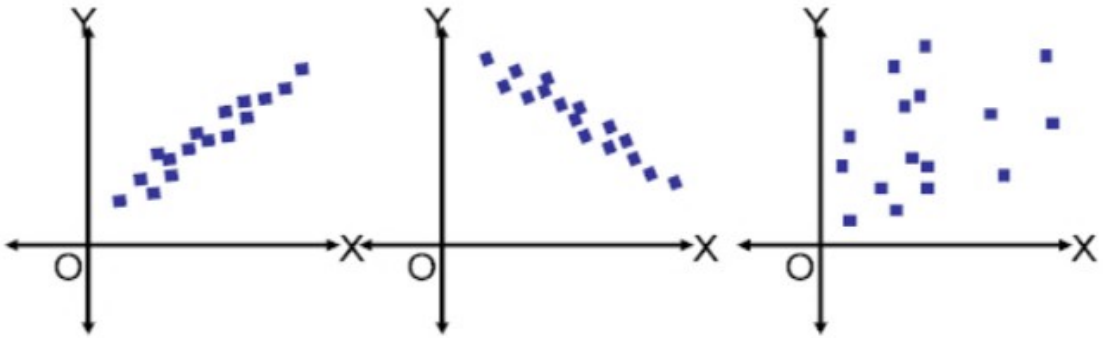
REGULARIZATION
Simply put Regularization penalizes model which are too complex.

Finding a model usually involves mimimizing an error function.
For Example, the Error Function could be the Sum of Squares of Distance Between the predicted points and the actual points in the training set. Let the Error Function be E(f) for a model f
A Regularization term is added to this function
$E'(f) = E(f) + \lambda R(f)$

NEW ERROR FUNCTION THAT NEEDS TO BE MINIMIZED
A PARAMETER THAT CONTROLS THE IMPORTANCE OF THEREGULARIZATION TERM

REGULARIZATION TERM THAT INCREASES WITH COMPLEXITY

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right)



In the above images, which of the following is/are example of multi-collinear features?

Options

Features in Image 1

Features in Image 2

Features in Image 3

Features in Image 1&2

Features in Image 1&3

1. **Accuracy**: Accuracy is calculated as the ratio of correctly predicted instances (true positives and true negatives) to the total number of instances.

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

For accuracy to be ~0.91, the number of correct predictions (TP + TN) should be about 91% of the total predictions.

2. **Misclassification rate**: This is the complement of accuracy and is calculated as the ratio of incorrect predictions (false positives and false negatives) to the total number of instances.

$$\text{Misclassification Rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN}$$

If the accuracy is ~0.91, then the misclassification rate would be ~0.09, not ~0.91.

3. **False positive rate (FPR)**: This is the ratio of false positives to the total number of actual negatives.

$$\text{FPR} = \frac{FP}{FP + TN}$$

For FPR to be ~0.95, false positives must constitute 95% of the total actual negatives, indicating a very high number of false positives compared to true negatives.

4. **True positive rate (TPR)**: Also known as sensitivity or recall, this is the ratio of true positives to the total number of actual positives.

$$\text{TPR} = \frac{TP}{TP + FN}$$

   For TPR to be ~0.95, true positives must constitute 95% of the total actual positives, indicating a high rate of correctly identifying positive cases.

Given these definitions:

- The **accuracy** being ~0.91 is plausible, depending on the values in the confusion matrix.
- The **misclassification rate** of ~0.91 is unlik[illegible] if the accuracy is ~0.91, as these two should sum to 1.

Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

1. **Accuracy is ~0.91**
2. **Misclassification rate is ~ 0.91**
3. **False positive rate is ~0.95**
4. **True positive rate is** ~0.95

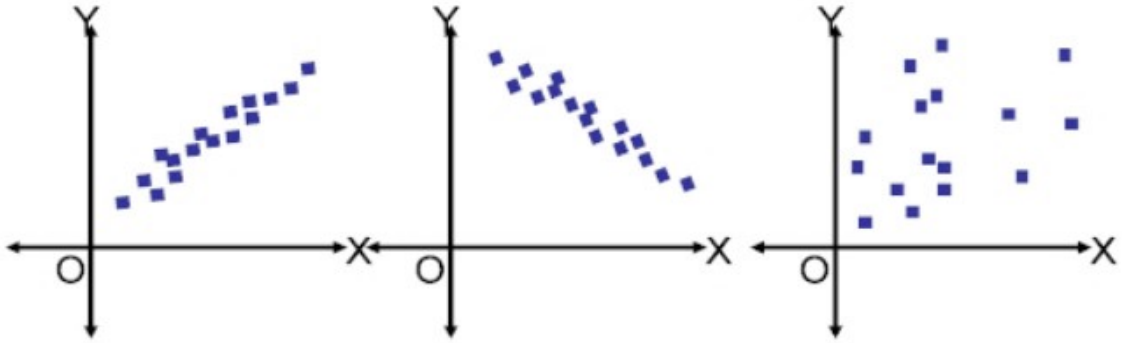
Options

1 and 3

2 and 4

1 and 4

2 and 3



suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. Remove both collinear variables.
2. Instead of removing both variables, we can remove only one variable.
3. Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.

Options

Only 1

Only 2

Only 3

Either 1 or 3

In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.

Which of the following statements is / are true for weak learners used in ensemble model?

1-They don't usually overfit.

2-They have high bias, so they cannot solve complex learning problems

3-They usually overfit.

Options

- 1 and 2
- 1 and 3
- 2 and 3
- only 1
- only 2