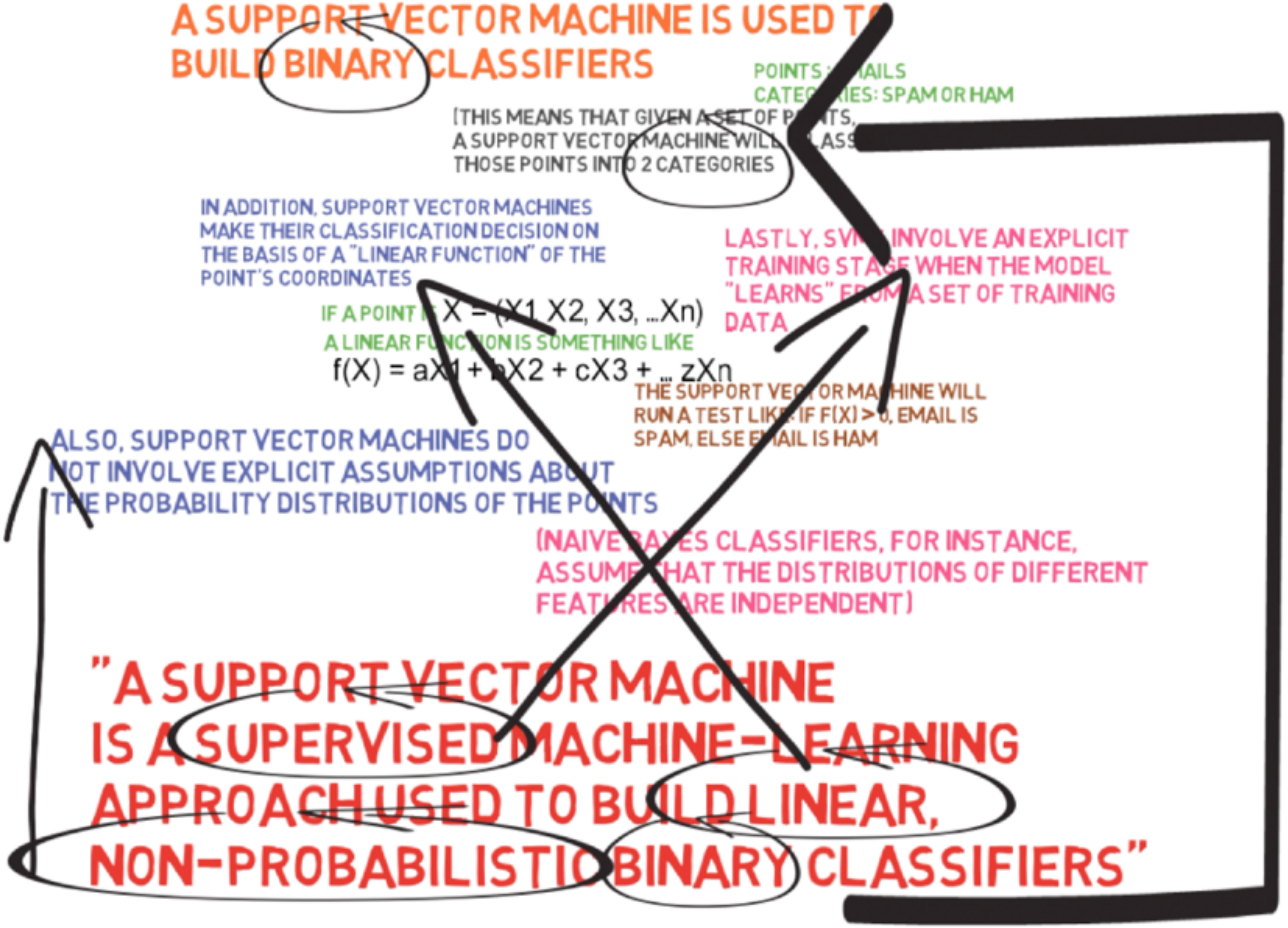
SUPPORT VECTOR MACHINES



A SUPPORT VECTOR MACHINE IS USED TO BUILD BINARY CLASSIFIERS

POINTS : EMAILS
CATEGORIES: SPAM OR HAM

(THIS MEANS THAT GIVEN A SET OF POINTS, A SUPPORT VECTOR MACHINE WILL CLASSIFY THOSE POINTS INTO 2 CATEGORIES

IN ADDITION, SUPPORT VECTOR MACHINES MAKE THEIR CLASSIFICATION DECISION ON THE BASIS OF A "LINEAR FUNCTION" OF THE POINT'S COORDINATES

IF A POINT IS X = (X1 X2, X3, ...Xn)

A LINEAR FUNCTION IS SOMETHING LIKE

f(X) = aX1 + bX2 + cX3 + ... zXn

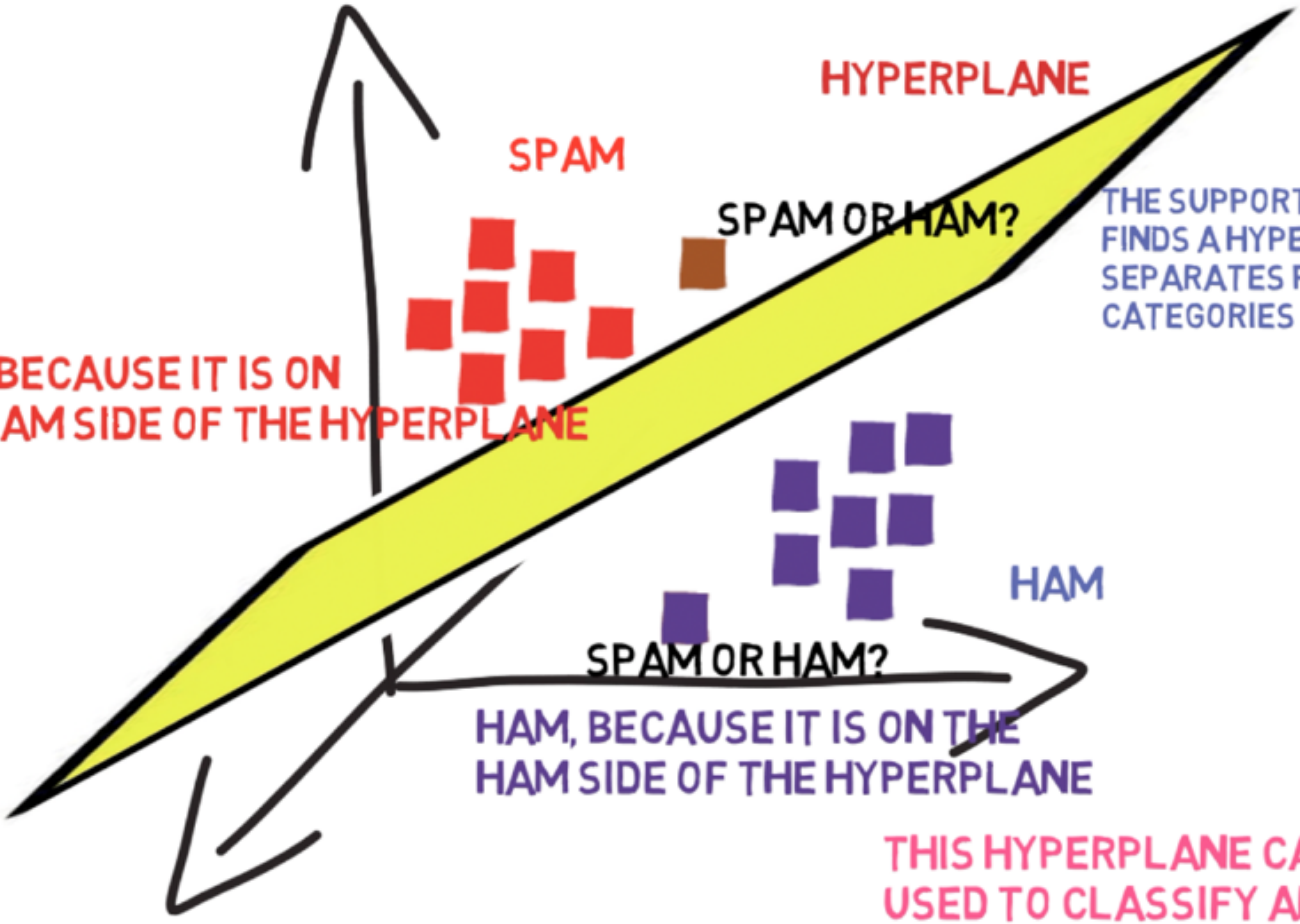
LASTLY, SVMS INVOLVE AN EXPLICIT TRAINING STAGE WHEN THE MODEL "LEARNS" FROM A SET OF TRAINING DATA

THE SUPPORT VECTOR MACHINE WILL RUN A TEST LIKE: IF F(X) > 0, EMAIL IS SPAM, ELSE EMAIL IS HAM

ALSO, SUPPORT VECTOR MACHINES DO NOT INVOLVE EXPLICIT ASSUMPTIONS ABOUT THE PROBABILITY DISTRIBUTIONS OF THE POINTS

(NAIVE BAYES CLASSIFIERS, FOR INSTANCE, ASSUME THAT THE DISTRIBUTIONS OF DIFFERENT FEATURES ARE INDEPENDENT)

"A SUPPORT VECTOR MACHINE IS A SUPERVISED MACHINE-LEARNING APPROACH USED TO BUILD LINEAR, NON-PROBABILISTIC BINARY CLASSIFIERS"



THE SUPPORT VECTOR MACHINE FINDS A HYPERPLANE THAT NEATLY SEPARATES POINTS OF THE TWO CATEGORIES

SPAM, BECAUSE IT IS ON THE SPAM SIDE OF THE HYPERPLANE

HAM, BECAUSE IT IS ON THE HAM SIDE OF THE HYPERPLANE

THIS HYPERPLANE CAN THEN BE USED TO CLASSIFY ANY NEW POINTS THAT NEED CLASSIFICATION

FIRST OFF - WHAT IS A HYPERPLANE?

IN A VECTOR SPACE OF N DIMENSIONS, A HYPERPLANE IS A GEOMETRIC SHAPE I.E. A SET OF POINTS - WITH (N-1) DIMENSIONS AND 0 THICKNESS IN ONE DIMENSION

THE EQUATION OF THE SET OF POINTS DEFINING THE HYPERPLANE IS ALWAYS "LINEAR"

$$Ax + By + Cz = D$$

IS THE EQUATION OF A HYPERPLANE IN 3D SPACE (I.E. A USUAL PLANE OF THE SORT WE JUST DREW)

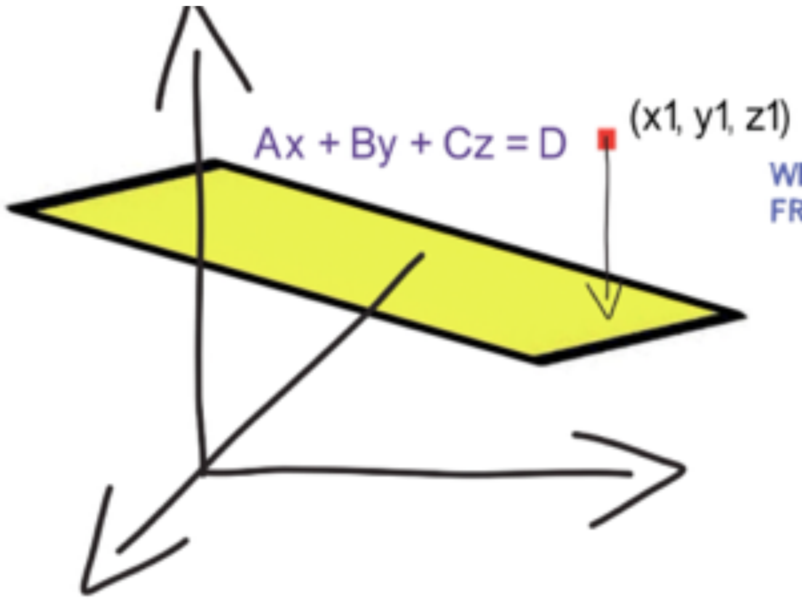
ALL POINTS ON THE PLANE WILL SATISFY THIS EQUATION

ALL POINTS ON ONE SIDE OF THE PLANE WILL SATISFY THE CONDITION

$$Ax + By + Cz > D$$

AND ALL POINTS ON THE OTHER SIDE WILL SATISFY $Ax + By + Cz < D$

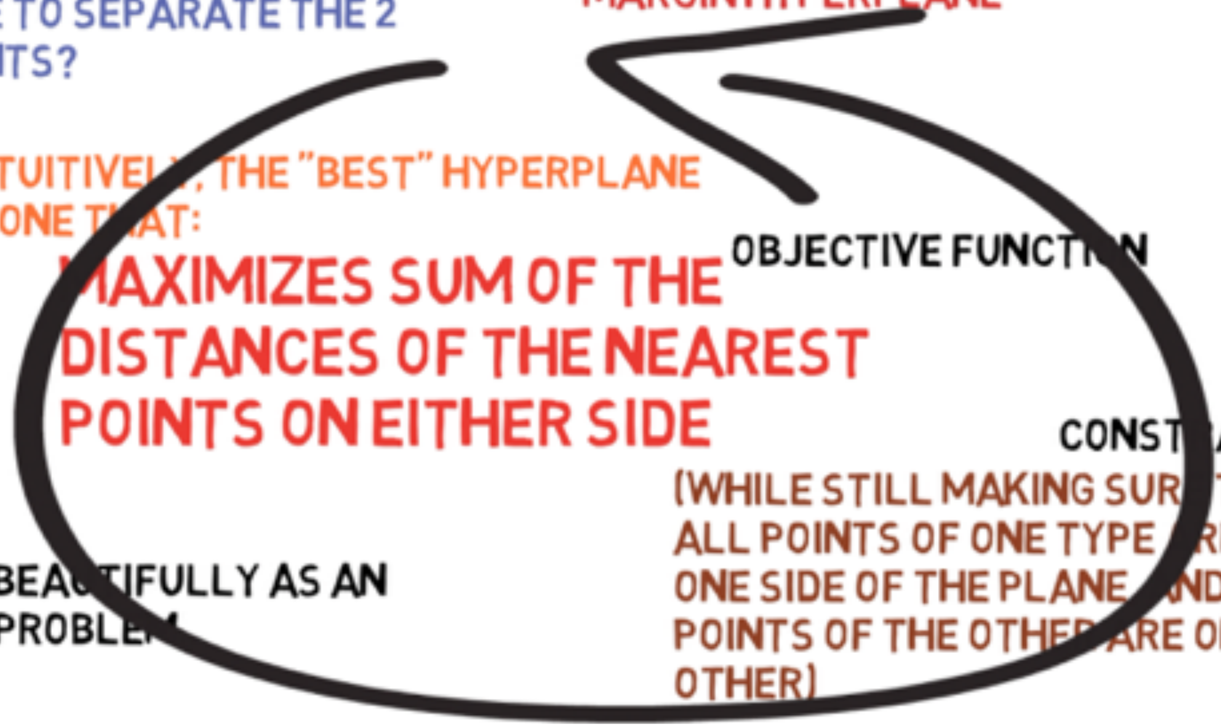
THIS IS THE LINEAR EQUATION THAT THE SVM USES TO CLASSIFY POINTS - WHICH IS WHY THE SVM IS A LINEAR CLASSIFIER



WHAT IS THE DISTANCE OF THE POINT FROM THE PLANE?

$$\frac{Ax1 + By1 + Cz1 - D}{[A^2 + B^2 + C^2]^{1/2}}$$

NOW COMING BACK TO OUR BASIC PROBLEM – HOW DOES THE SUPPORT VECTOR MACHINE FIND THE "BEST" HYPERPLANE TO SEPARATE THE 2 SETS OF POINTS?



INTUITIVELY, THE "BEST" HYPERPLANE IS ONE THAT:

MAXIMIZES SUM OF THE DISTANCES OF THE NEAREST POINTS ON EITHER SIDE

OBJECTIVE FUNCTION

CONSTRAINTS

(WHILE STILL MAKING SURE THAT ALL POINTS OF ONE TYPE ARE ON ONE SIDE OF THE PLANE AND ALL POINTS OF THE OTHER ARE ON THE OTHER)

THIS IS SET UP BEAUTIFULLY AS AN OPTIMIZATION PROBLEM

THE SOLUTION IS CALLED

THE MAXIMUM MARGIN HYPERPLANE

WE WON'T GO INTO THE DETAILS OF HOW EXACTLY THAT OPTIMIZATION PROBLEM IS FRAMED MATHEMATICALLY OR SOLVED –

BUT SUFFICE IT TO SAY THAT IT CAN BE CONVERTED INTO A FAIRLY STANDARD QUADRATIC PROGRAMMING PROBLEM FOR WHICH STANDARD SOLUTION TECHNIQUES EXIST

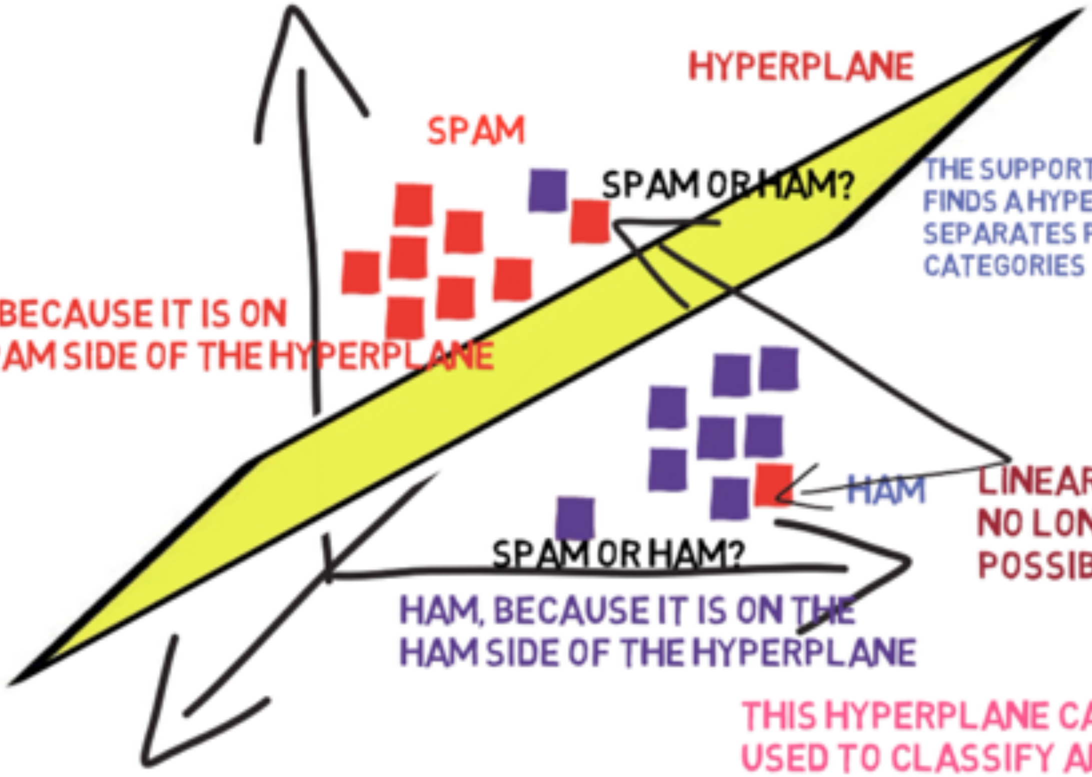
(IT TURNS OUT THE MAXIMUM MARGIN HYPERPLANE IS A FUNCTION OF THE SUPPORT VECTORS ALONE)

THE MAXIMUM MARGIN HYPERPLANE

IS FOUND – AND BTW THE "SUPPORT VECTORS" ARE SIMPLY THE "NEAREST POINTS" ON EACH SIDE – WHICH "SUPPORT" THE HYPERPLANE

THERE IS A CATCH THOUGH -
WHAT IF THE POINTS ARE NOT

LINEARLY SEPARABLE?



THE SUPPORT VECTOR MACHINE FINDS A HYPERPLANE THAT NEATLY SEPARATES POINTS OF THE TWO CATEGORIES

SPAM, BECAUSE IT IS ON THE SPAM SIDE OF THE HYPERPLANE

LINEAR SEPARATION NO LONGER POSSIBLE!!

HAM, BECAUSE IT IS ON THE HAM SIDE OF THE HYPERPLANE

THIS HYPERPLANE CAN THEN BE USED TO CLASSIFY ANY NEW POINTS THAT NEED CLASSIFICATION

THIS REQUIRES MORE MATHEMATICAL HEAVY LIFTING, BUT SOME GREAT SOUL HAS FOUND A WAY AROUND THIS:

THE SOFT MARGIN METHOD

WHICH FINDS A HYPERPLANE THAT DOES "AS CLEAN A SEPARATION AS POSSIBLE"

THIS METHOD ALSO ALLOWS MEASUREMENT OF THE DEGREE OF MISCLASSIFICATION IN THE TRAINING DATA

BUT THERE IS ACTUALLY WAY TO USE SUPPORT VECTOR MACHINES TO PERFORM NON-LINEAR CLASSIFICATION USING SOMETHING CALLED

THE KERNEL TRICK

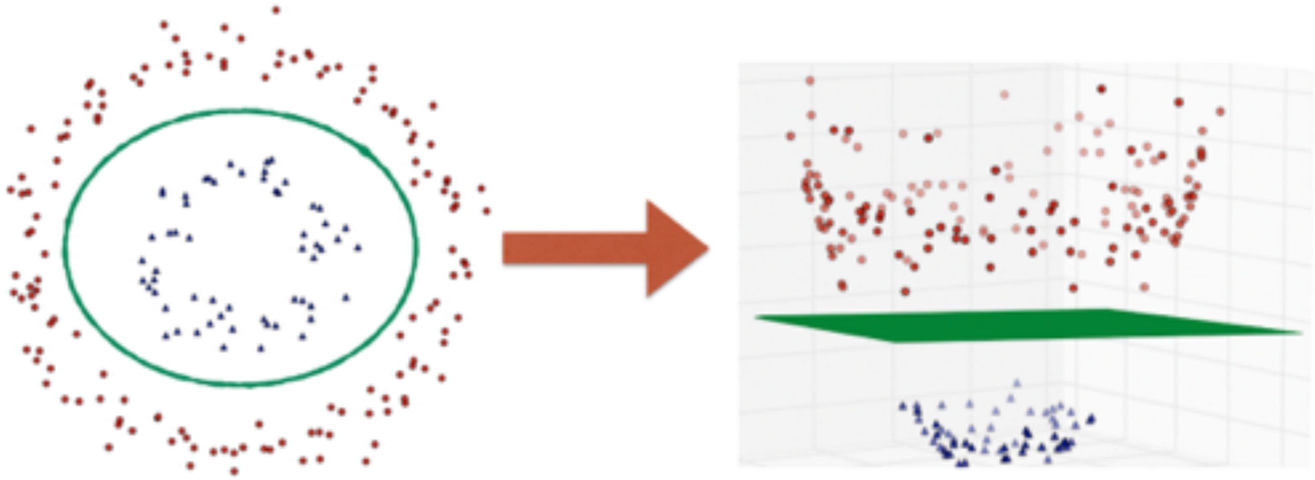
WHY DOES THIS MATTER TO US?

BECAUSE THE KERNEL TRICK IS A WAY IN WHICH SUPPORT VECTOR MACHINES CAN DO THEIR THING IN VERY VERY HIGH DIMENSIONALITY SPACES

(THE KERNEL TRICK HELPS GET AROUND THE CURSE OF DIMENSIONALITY)

THE PROBLEM WITH THE LINEAR CLASSIFICATION SETUP IS THE NEED TO CALCULATE DOT PRODUCTS OF VECTORS WITH HUGE NUMBERS OF ELEMENTS

A KERNEL IS A FUNCTION THAT ALLOWS THE CLASSIFICATION TO BE DONE IMPLICITLY, I.E. WITHOUT ACTUALLY DEALING WITH THE FEATURE VECTOR IN ELEMENT-WISE OPERATIONS



THE PROBLEM WITH THE LINEAR CLASSIFICATION SETUP IS THE NEED TO CALCULATE DOT PRODUCTS OF VECTORS WITH HUGE NUMBERS OF ELEMENTS

A KERNEL IS A FUNCTION THAT ALLOWS THE CLASSIFICATION TO BE DONE IMPLICITLY, I.E. WITHOUT ACTUALLY DEALING WITH THE FEATURE VECTOR IN ELEMENT-WISE OPERATIONS

THE KERNEL IS SIMPLY A (NON-LINEAR) FUNCTION THAT OPERATES ON TWO POINTS - THIS KERNEL FUNCTION IS USED INSTEAD OF A DOT PRODUCT

NOTE THAT KERNEL FUNCTIONS OPERATE IN A TRANSFORMED FEATURE SPACE, WHICH CAN BE OF FAR HIGHER DIMENSIONALITY THAN THE ORIGINAL FEATURE SPACE

EVEN THOUGH THE DIMENSIONALITY IN WHICH THEY OPERATE IS HIGHER, THEY WILL FIND THE MAXIMUM MARGIN HYPERPLANE MORE EASILY BECAUSE OF THE KERNEL TRICK

THIS MAXIMUM MARGIN HYPERPLANE IS IN THE TRANSFORMED FEATURE SPACE, AND IS LINEAR IN THAT SPACE..

..ALTHOUGH IT MIGHT BE NON-LINEAR IN THE ORIGINAL FEATURE SPACE

THUS THE KERNEL TRICK ALSO ALLOWS A WAY TO SOLVE PROBLEMS WHERE THE DATA IS NOT LINEARLY SEPARABLE - BY PROJECTING SUCH DATA INTO HIGHER DIMENSION SPACE