Methods for detecting cyber attacks Introduction To ML

Dr. Ran Dubin

Lecture slides were taken from Prof. Lior Rokach and Prof. Asaf Shabtai

Machine Learning vs. Statistics

The two cultures

Machine Learning	Statistics
Uses algorithmic models and treats the data mechanism as unknown	Assumes that the data are generated by a given stochastic data model
Complicated data	Relatively simple data
Model validation. Measured by predictive accuracy.	*Model validation*. *Question* Yes–no Using goodness-of-fit tests (how well it fits a set of observations) and residual examination (predicted value – expected value).

Leo Breiman, Statistical Modeling: The Two Cultures, Statistical Science, 2001, Vol. 16, No. 3, 199–231

Machine Learning

- Herbert Alexander Simon:

  "Learning is any process by which a system improves performance from experience."

- "Machine Learning is concerned with computer programs that automatically improve their performance through experience."

Herbert Simon
Turing Award 1975
Nobel Prize in Economics 1978

Why Machine Learning Now?

- Flood of available data (Internet)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers (open source projects)
- Increasing support from industries

The Concept of Learning In a ML System

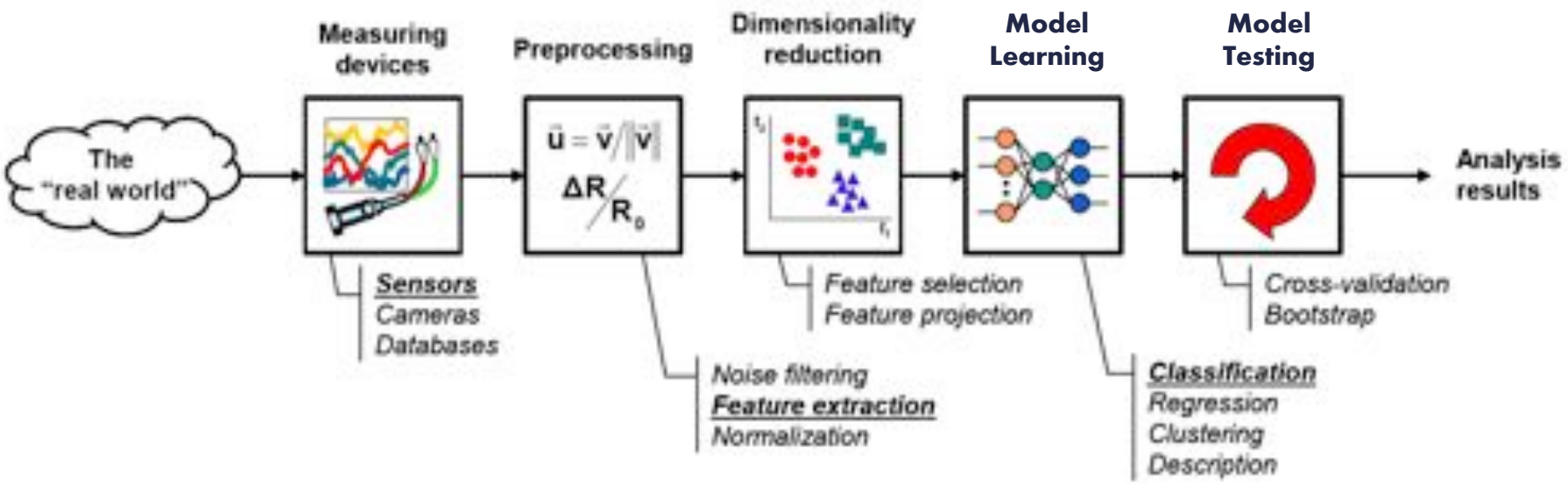
- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*

SPAM Use Case

- Spam - an email that the user does not want to receive and has not asked to receive
  - ***Improve*** *task T*: Identify Spam Emails
  - ***Performance*** *metric P*:
    - % of spam emails that were filtered
    - % of ham (non-spam) emails that were incorrectly filtered-out
  - *Based on* ***experience*** *E*:
    a database of emails that were labelled by users/experts

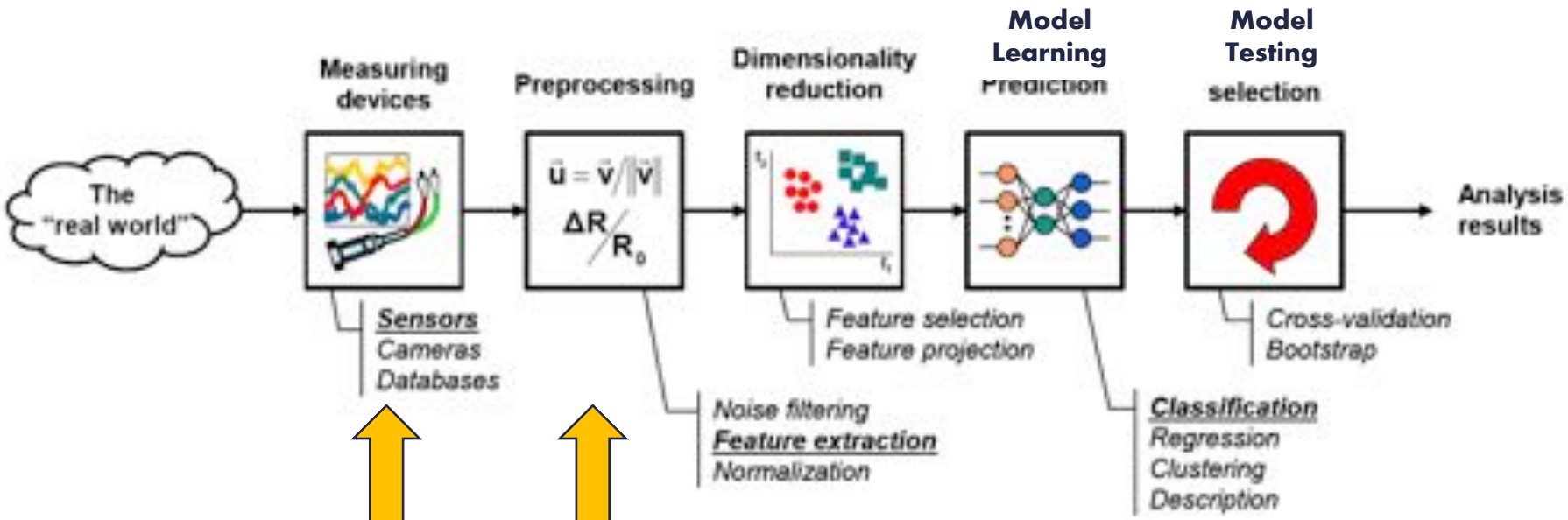
The learning process

The learning process



The learning process

SPAM filtering example



Email Server

- Number of recipients
- Size of message
- Number of attachments
- Number of "re's" in the subject line
- …

Dataset

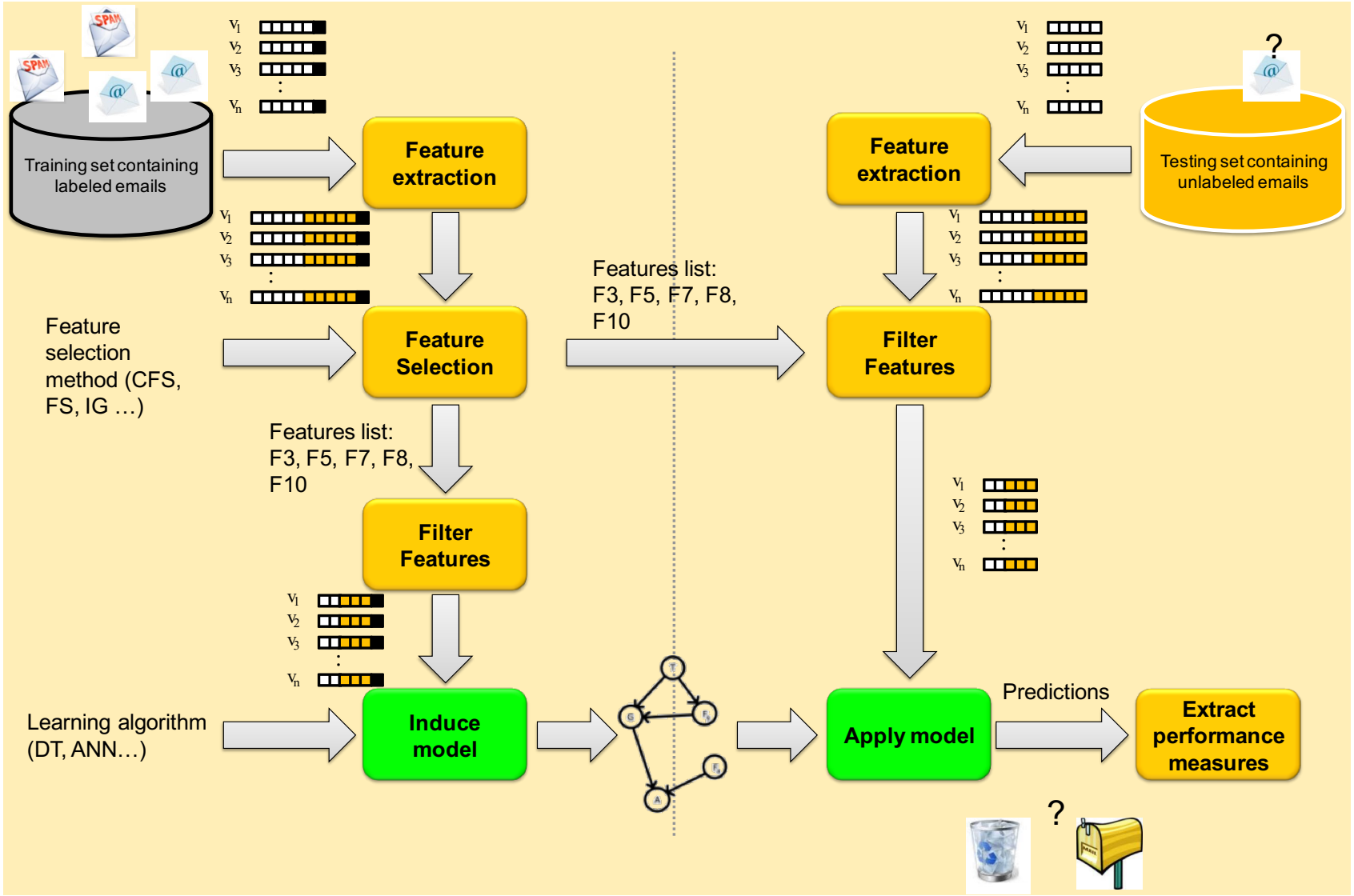
Input Attributes

Target Attribute

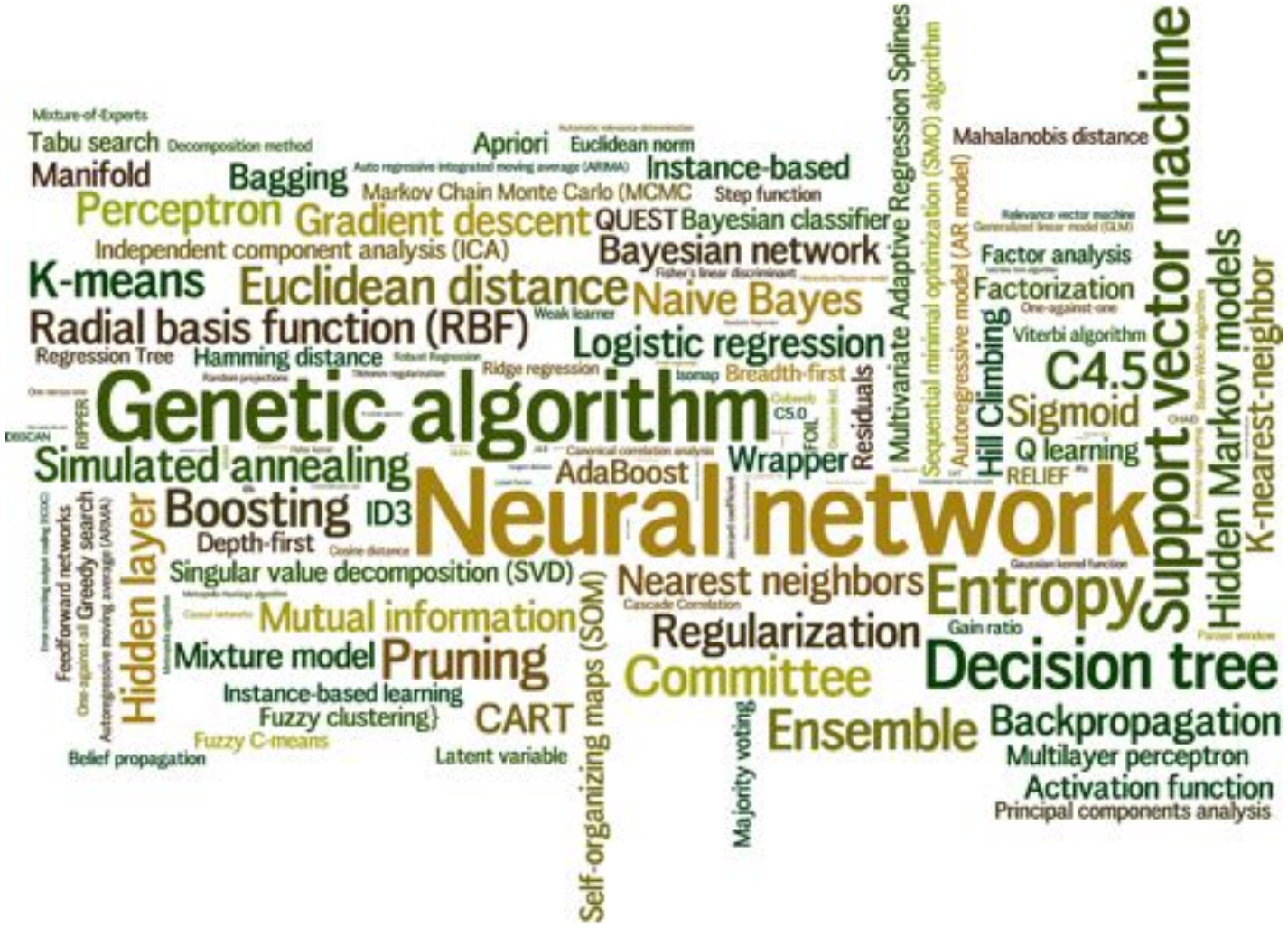
Instances

Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

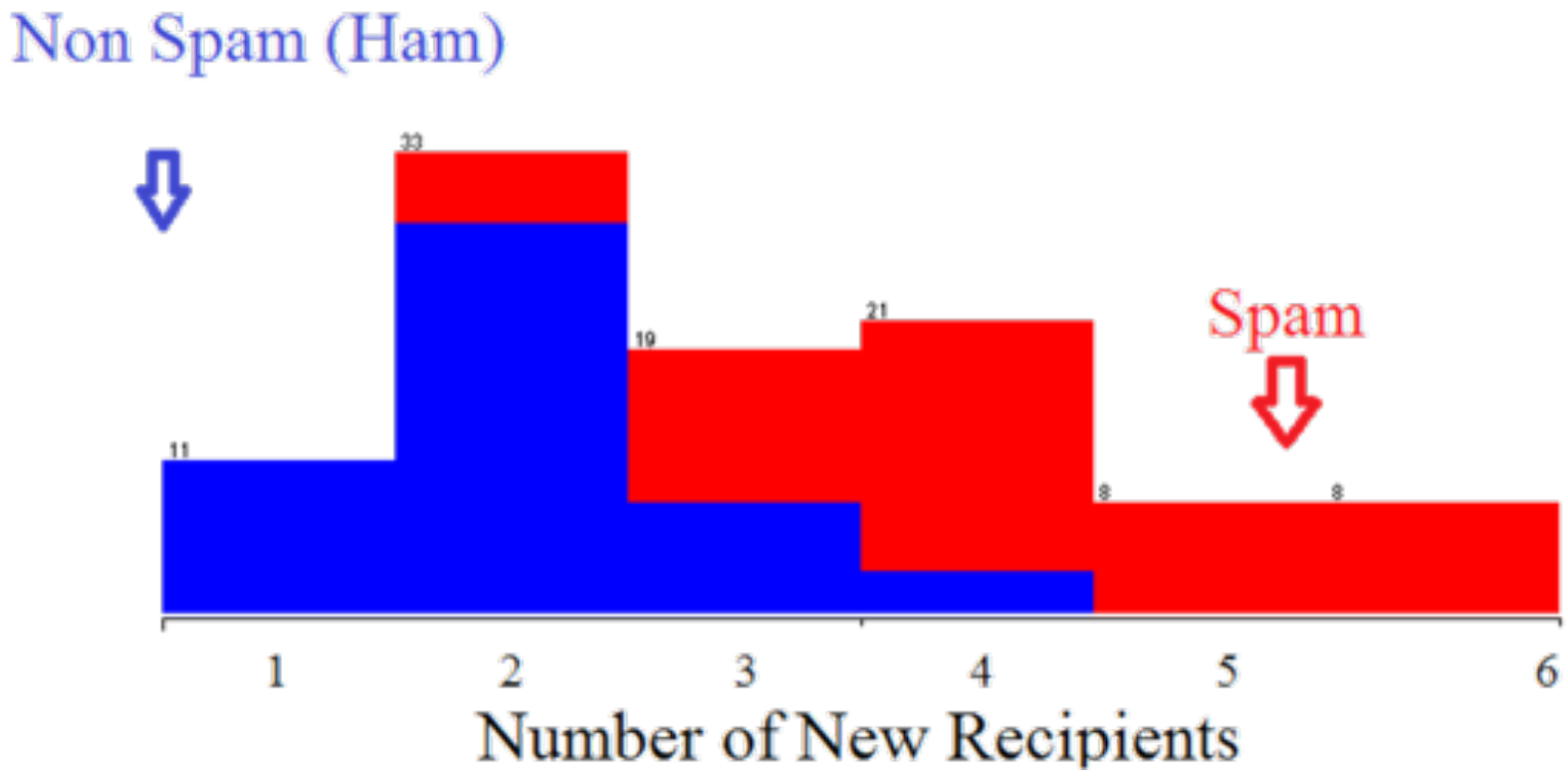
Model learning and testing



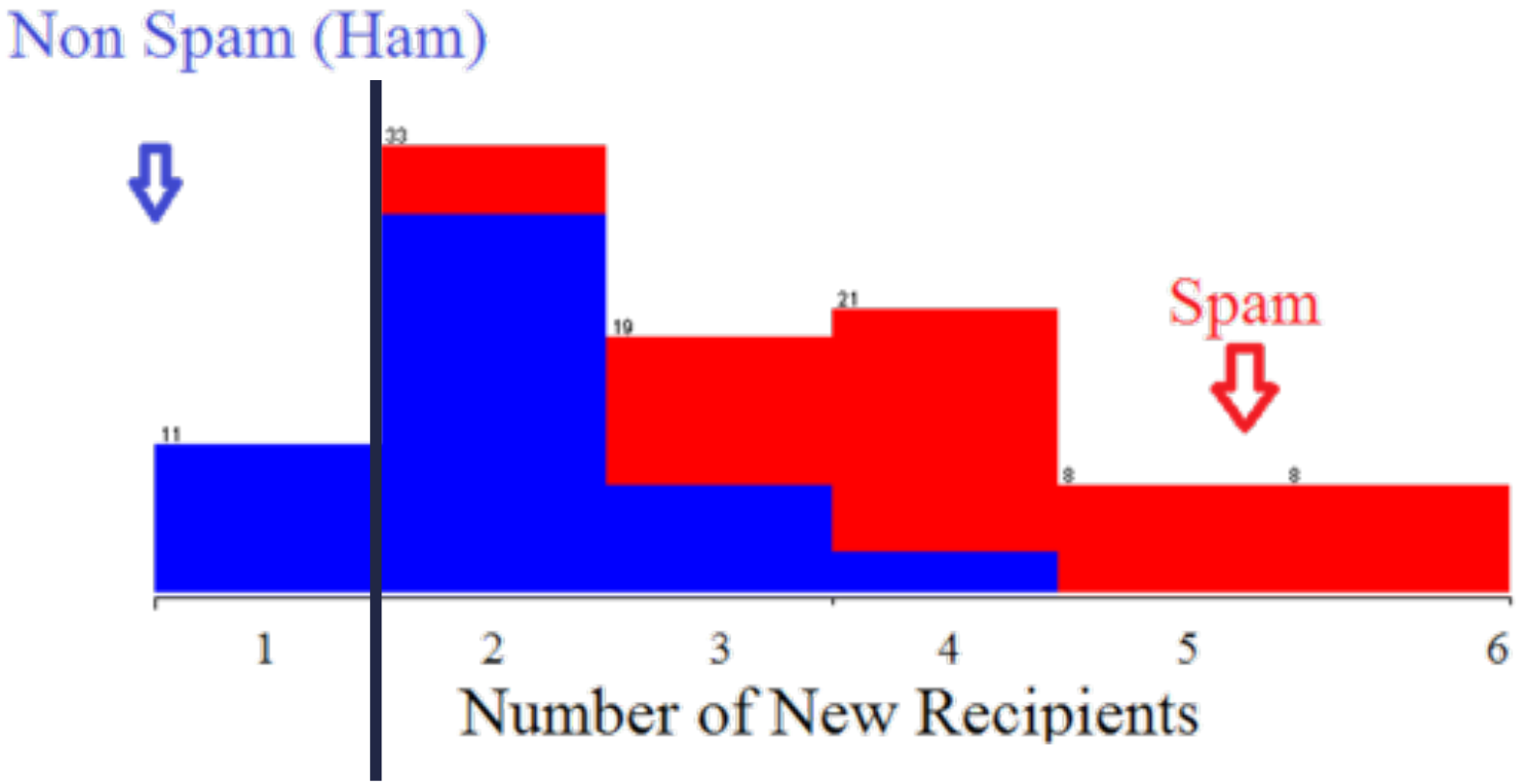
Learning algorithms



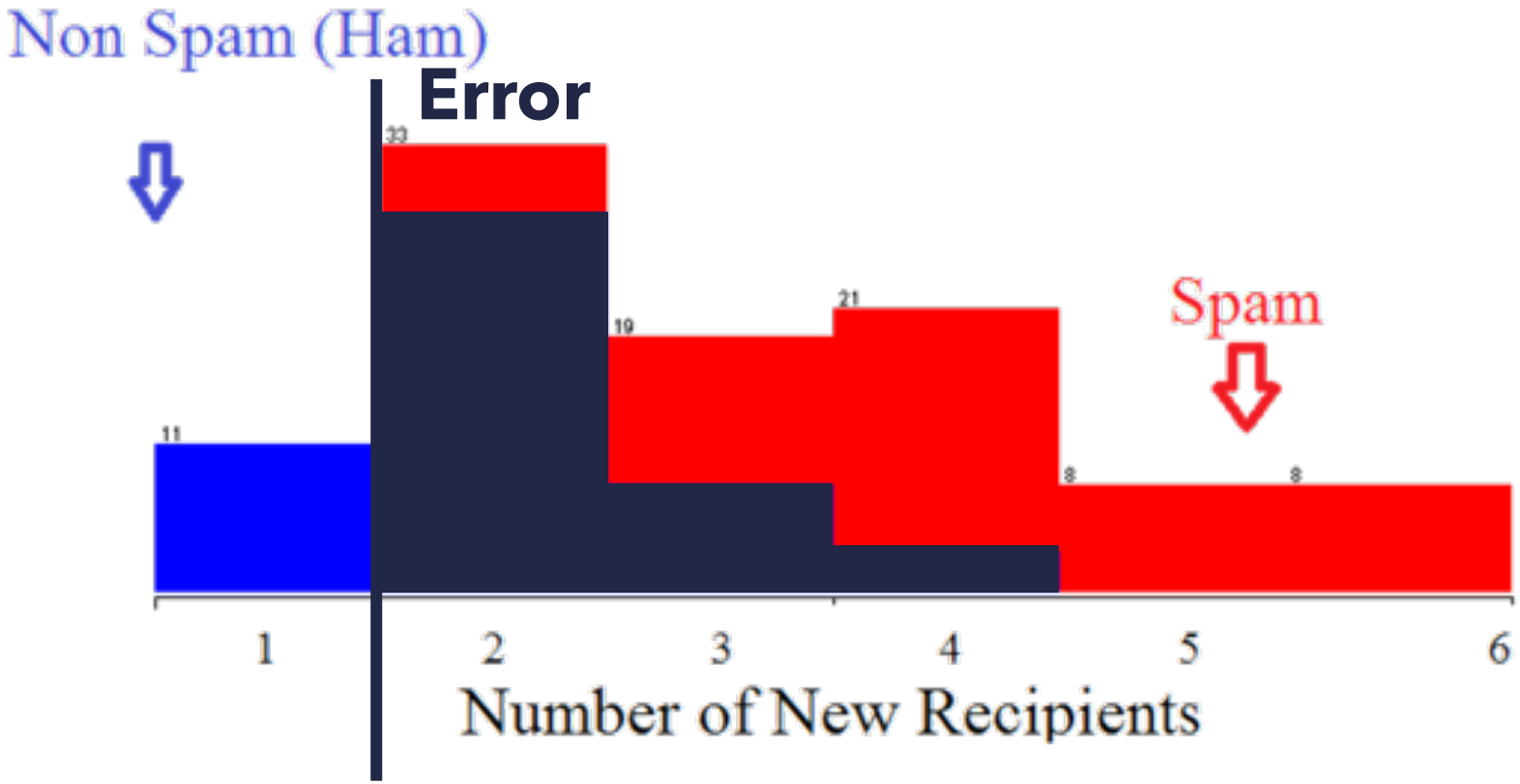
SPAM / HAM ? Data Exploration



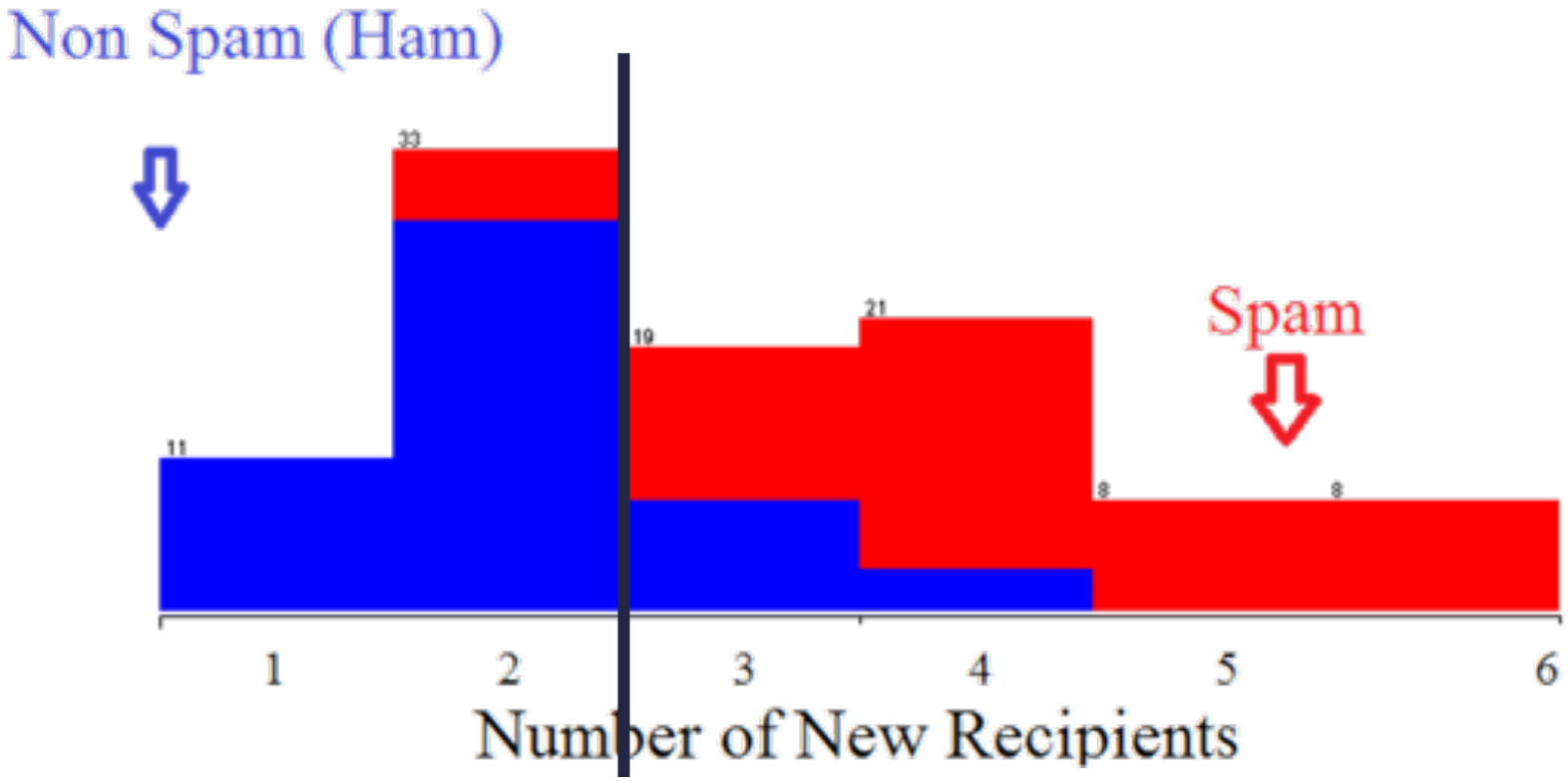
SPAM / HAM ? Data Exploration



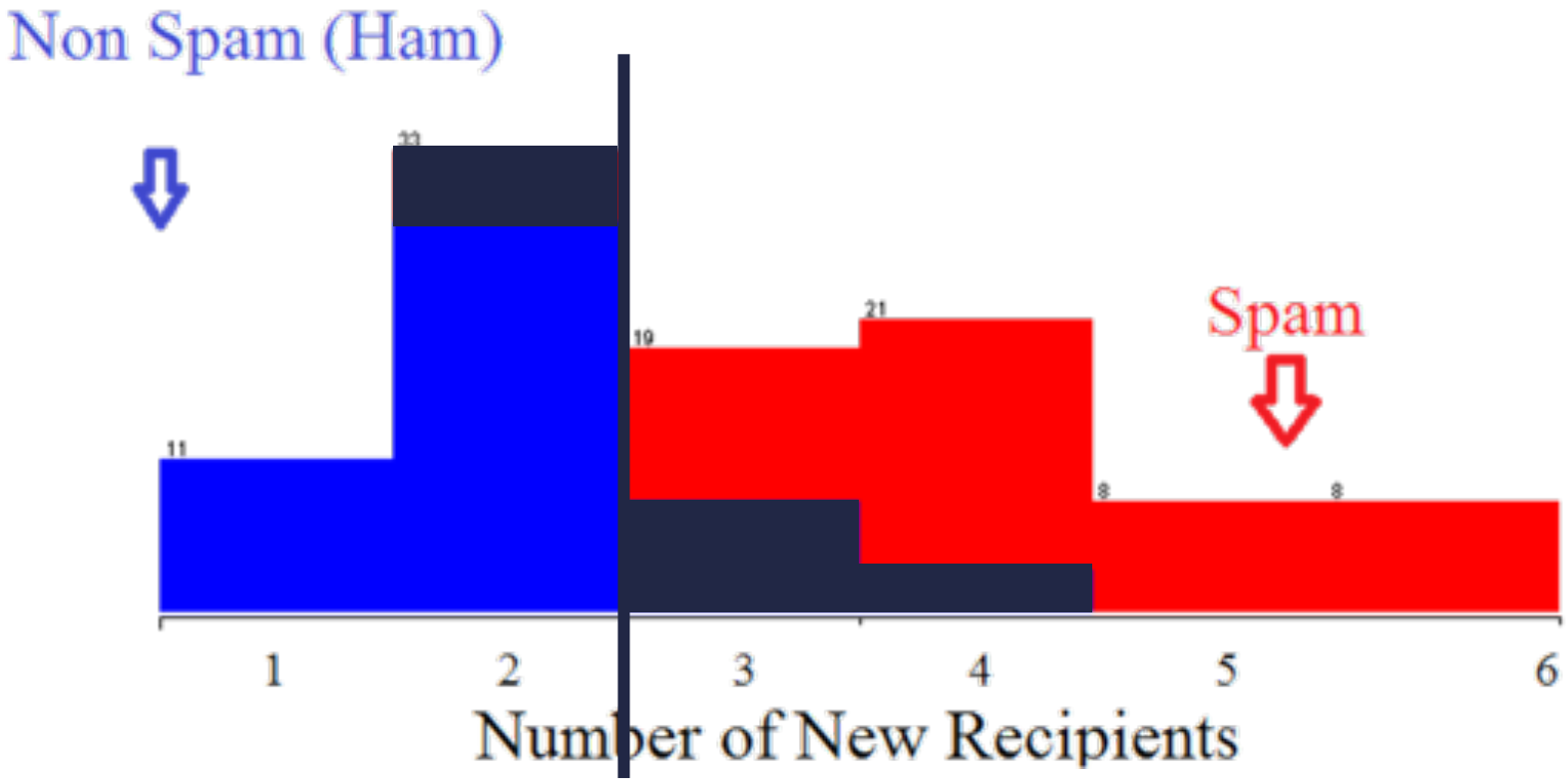
SPAM / HAM ? Data Exploration



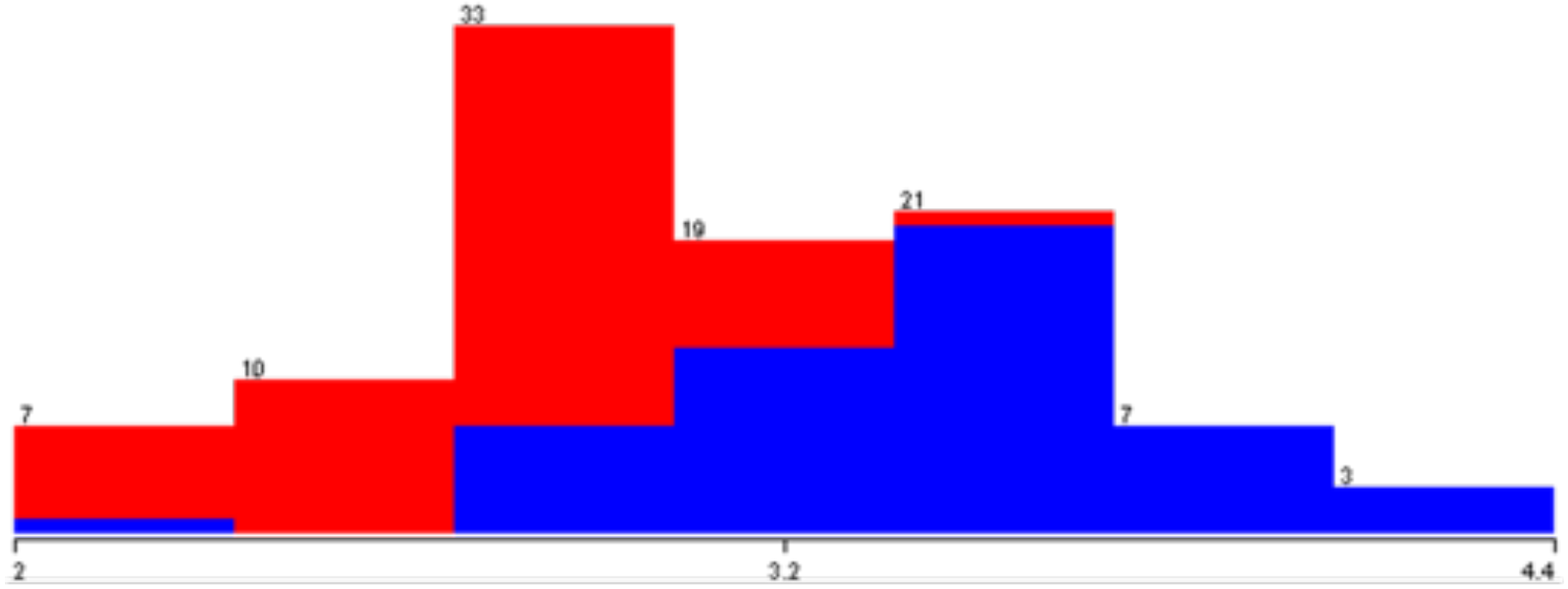
SPAM / HAM ? Data Exploration



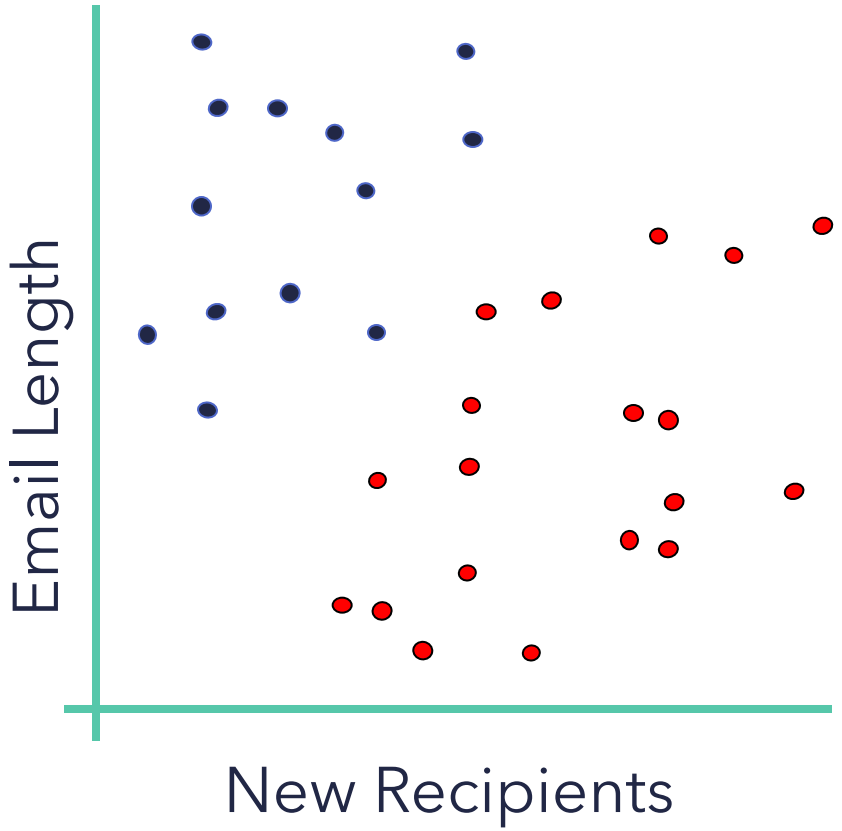
SPAM / HAM ? Data Exploration



SPAM / HAM ? Data Exploration

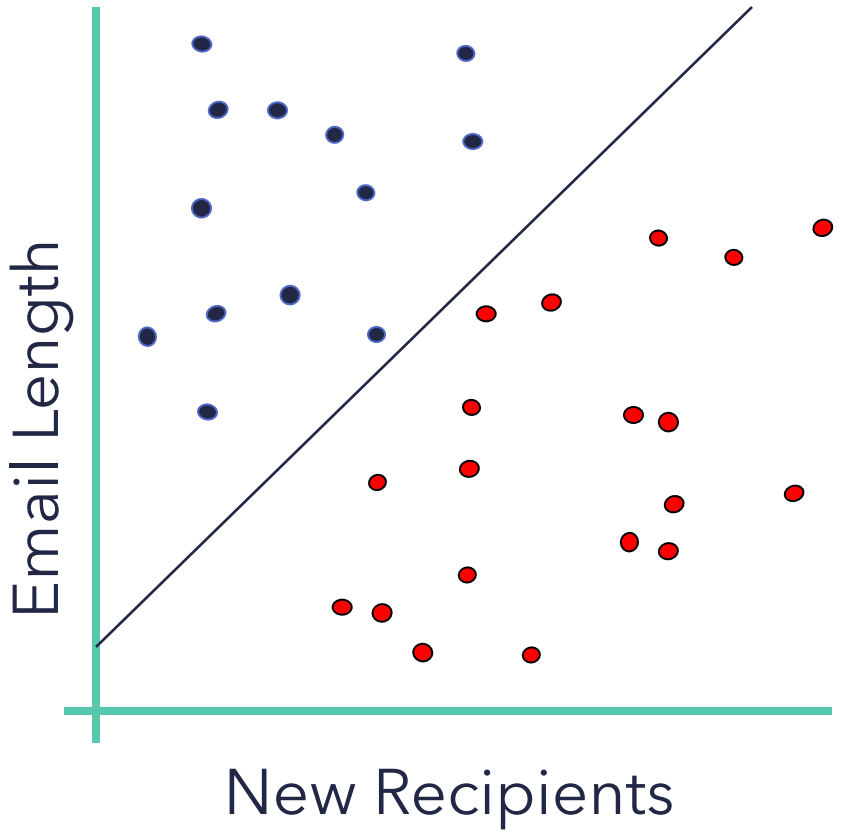


Linear Classifiers



How would you classify this data?

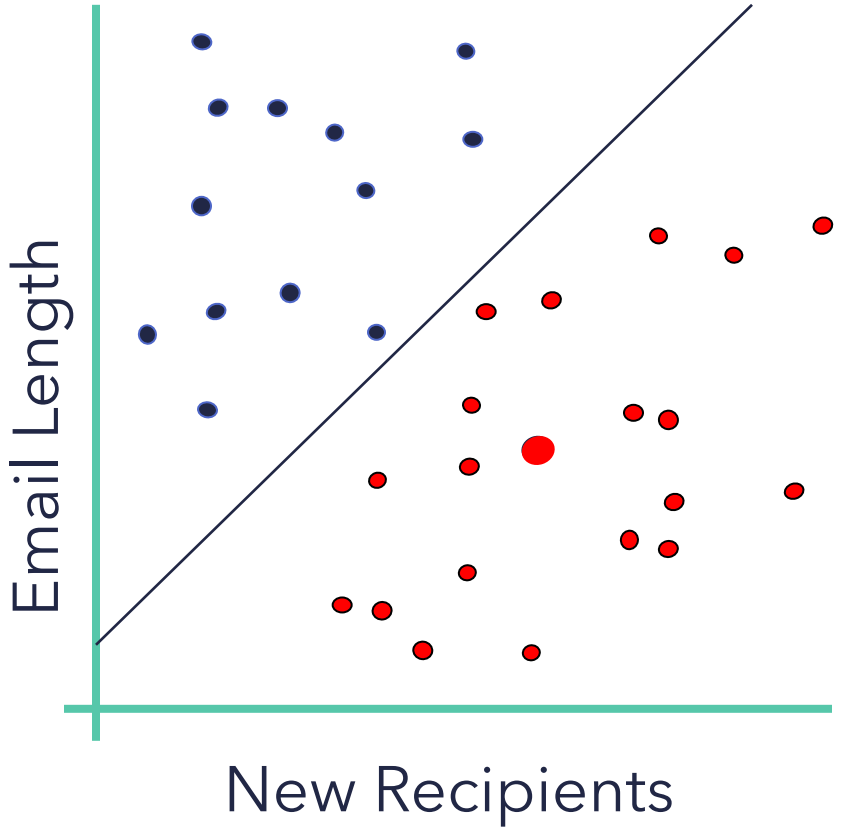
Linear Classifiers



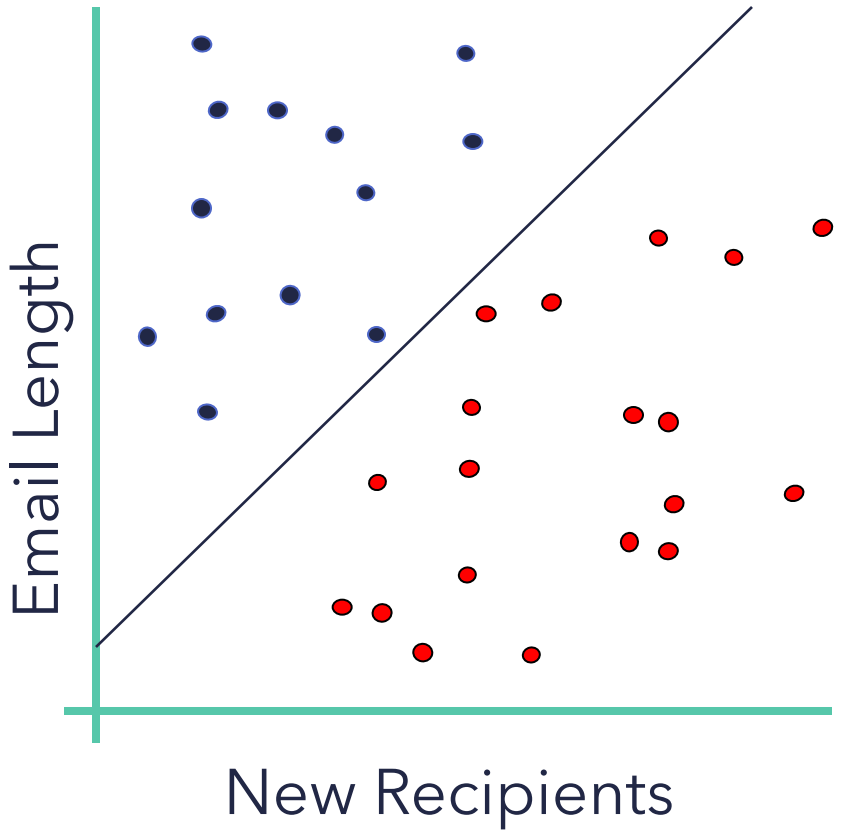
How would you classify this data?

When a New Email Is Sent

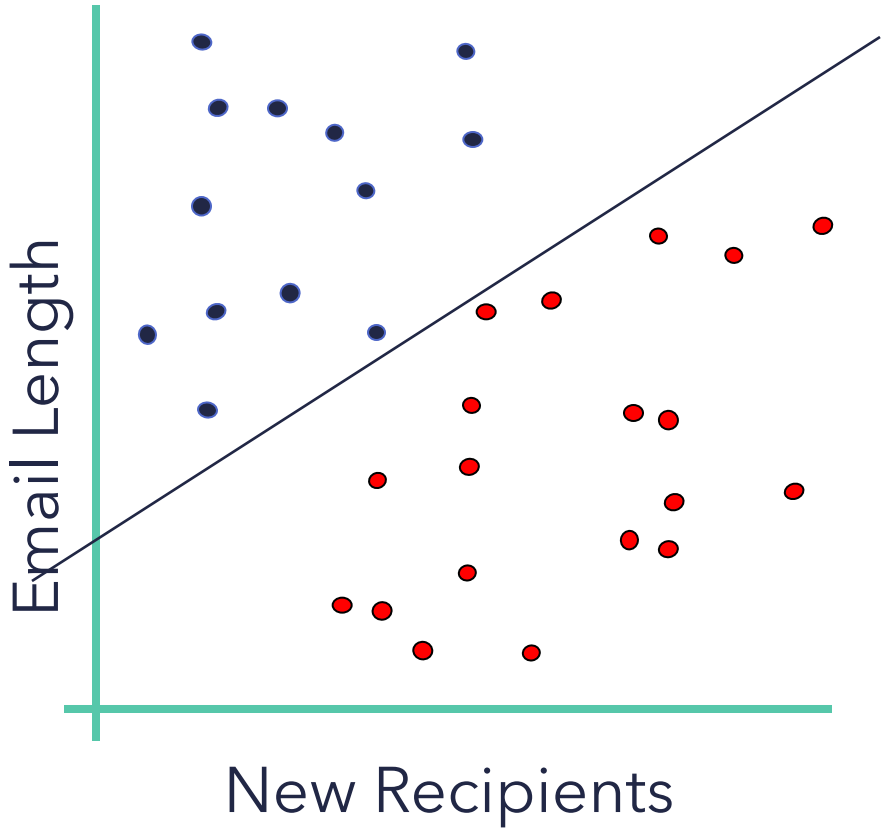
1. We first place the new email in the space
2. Classify it according to the subspace in which it resides



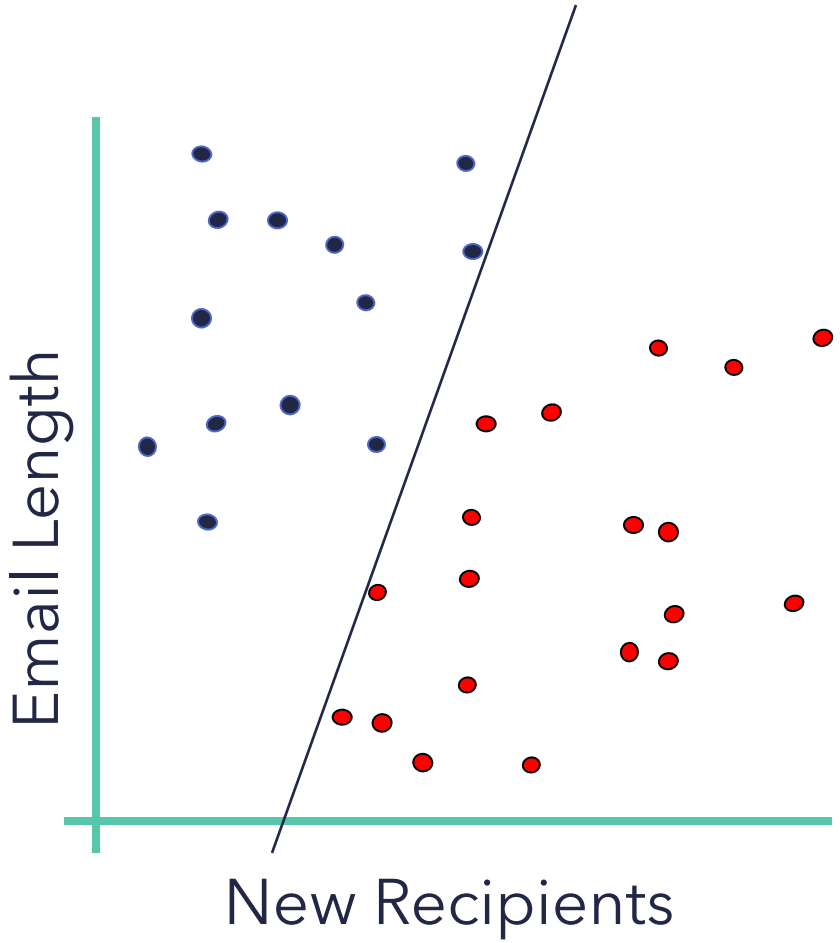
Linear classifiers



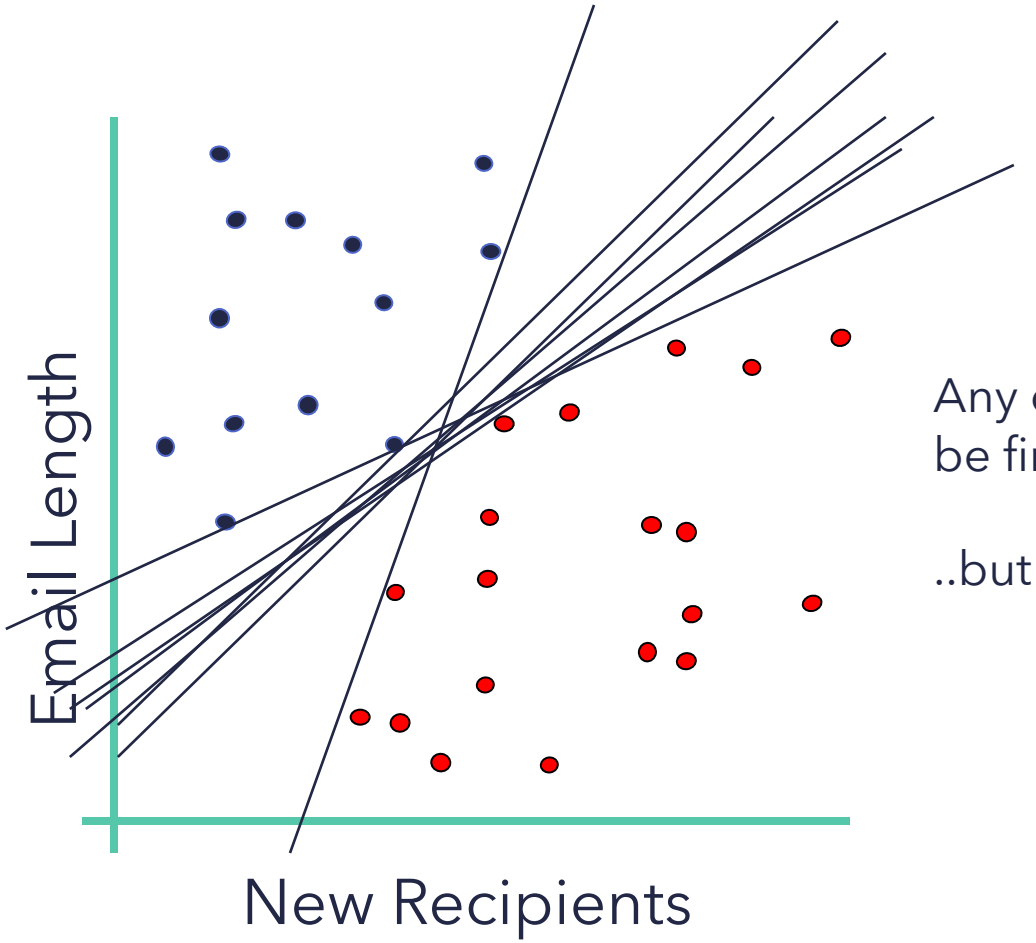
Linear classifiers



Linear classifiers



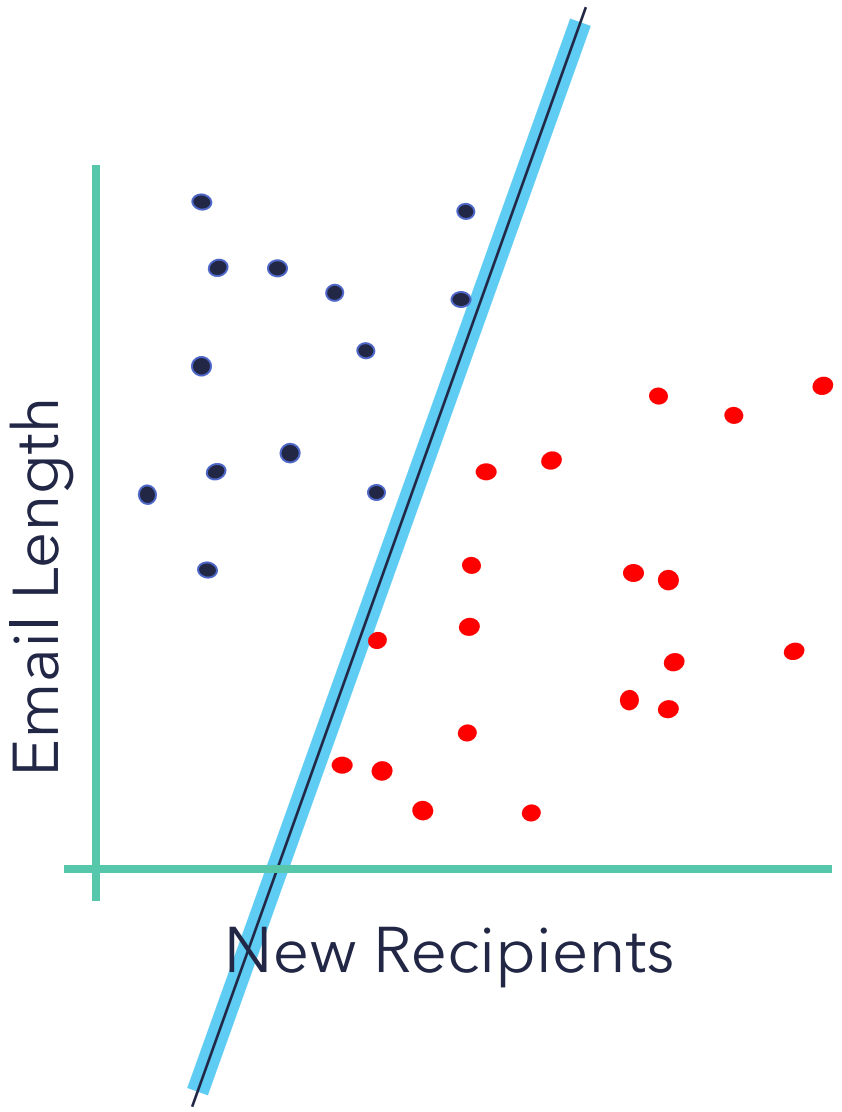
Linear classifiers



Any of these would be fine..

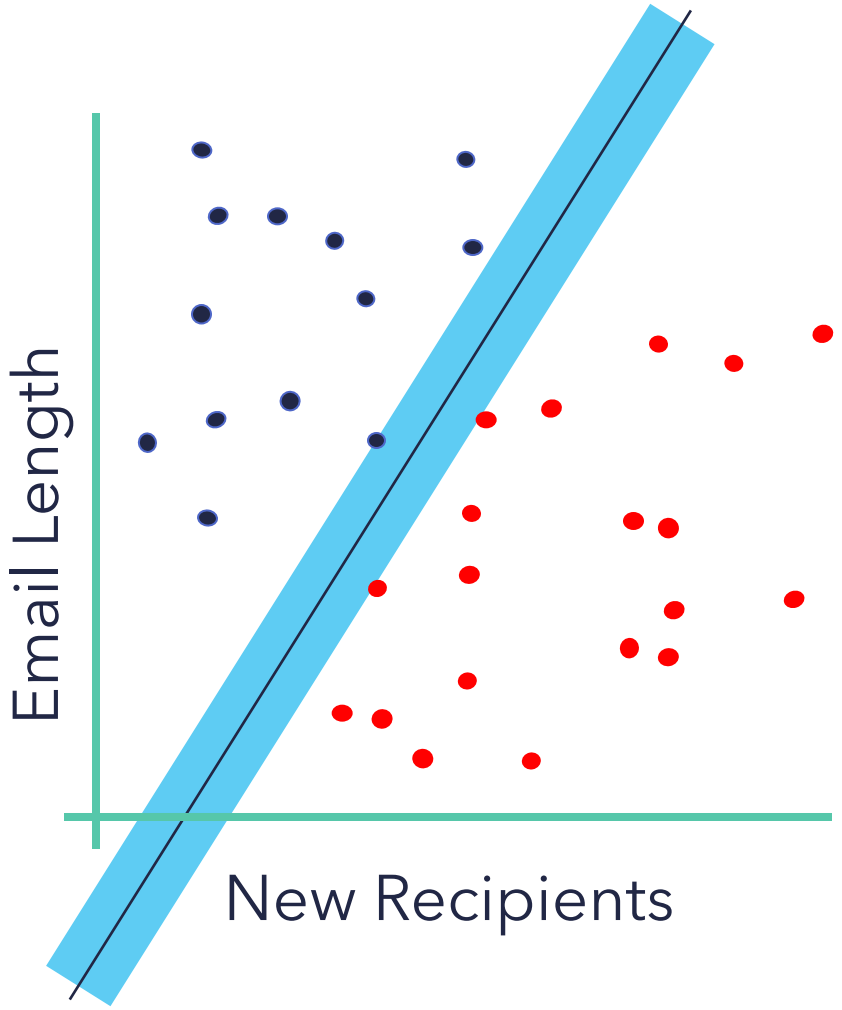
..but which is best?

Classifier margin



Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a data point

Maximum margin



The maximum margin linear classifier is the linear classifier with the, maximum margin.
This is the simplest kind of SVM (called an LSVM).

Maximum Margin

<u>Question</u>: What is the best separating hyperplane?

<u>Answer</u>: The one that maximizes the distance to the closest data points from both classes. We say it is the hyperplane with <u>maximum margin</u>.

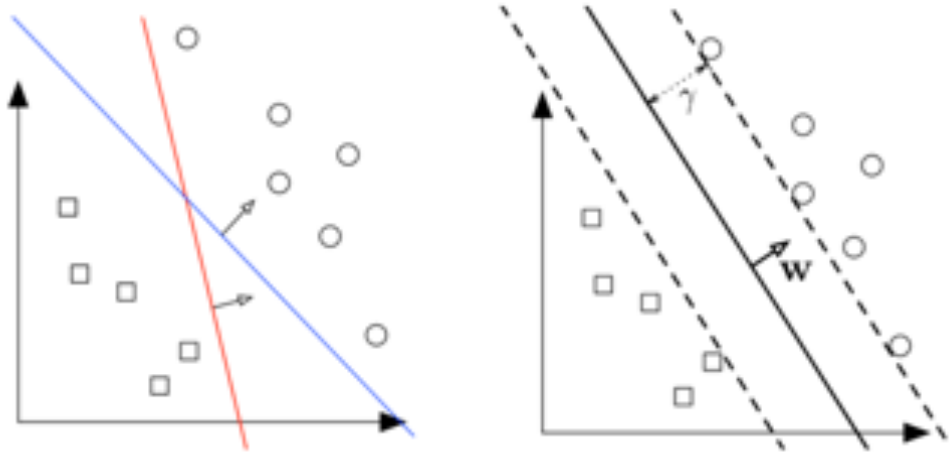
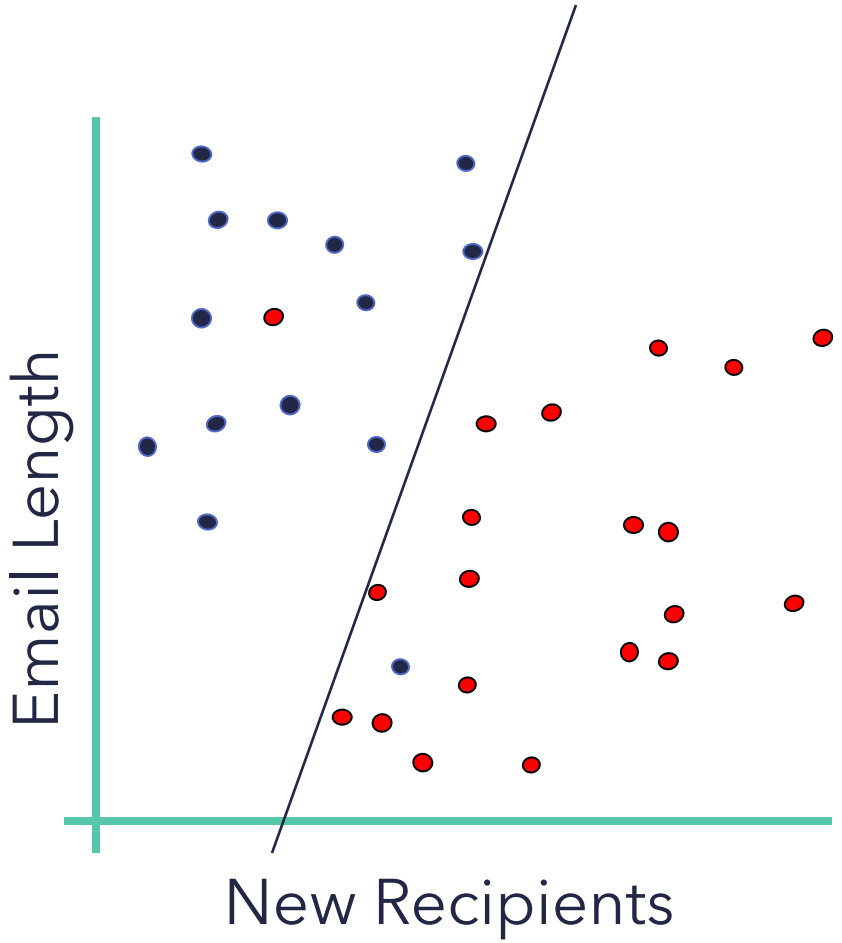


Figure 1: (Left:) Two different separating hyperplanes for the same data set. (Right:) The maximum margin hyperplane. The margin, γ, is the distance from the hyperplane (solid line) to the closest points in either class (which touch the parallel dotted lines).

No linear classifier can cover all instances

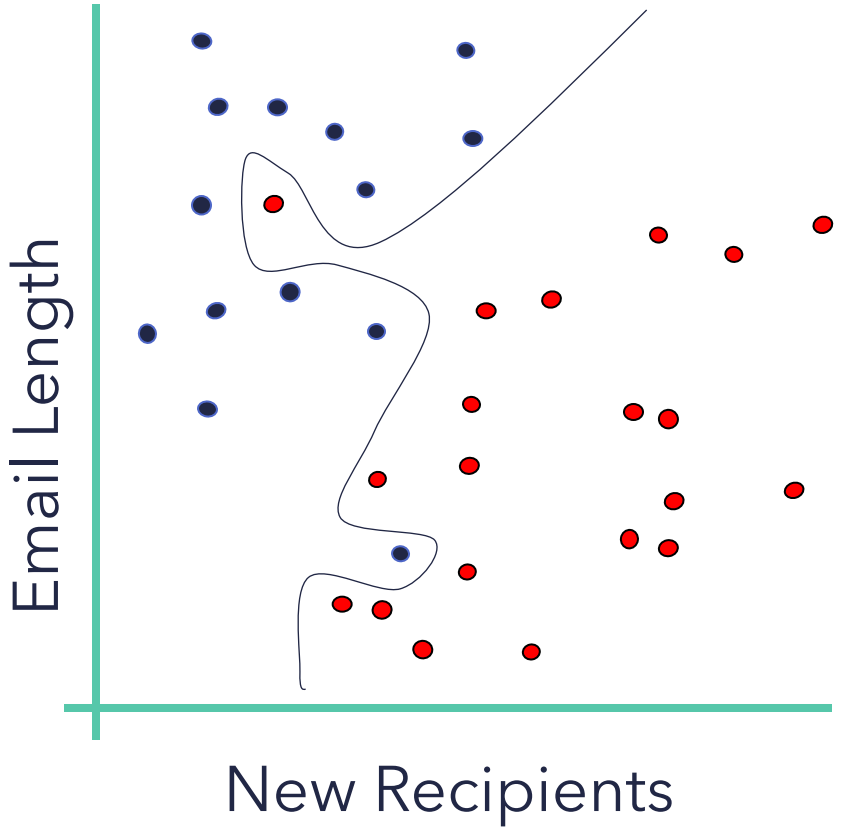


How would you classify this data?

No linear classifier can cover all instances

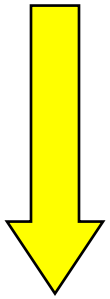
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure

No linear classifier can cover all instances



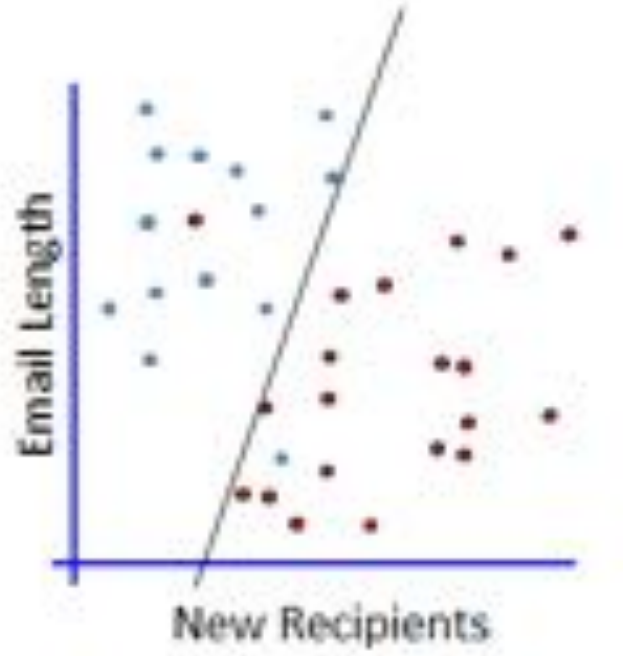
No linear classifier can cover all instances

- However, our satisfaction is premature because the main goal of designing a classifier is to correctly classify novel input

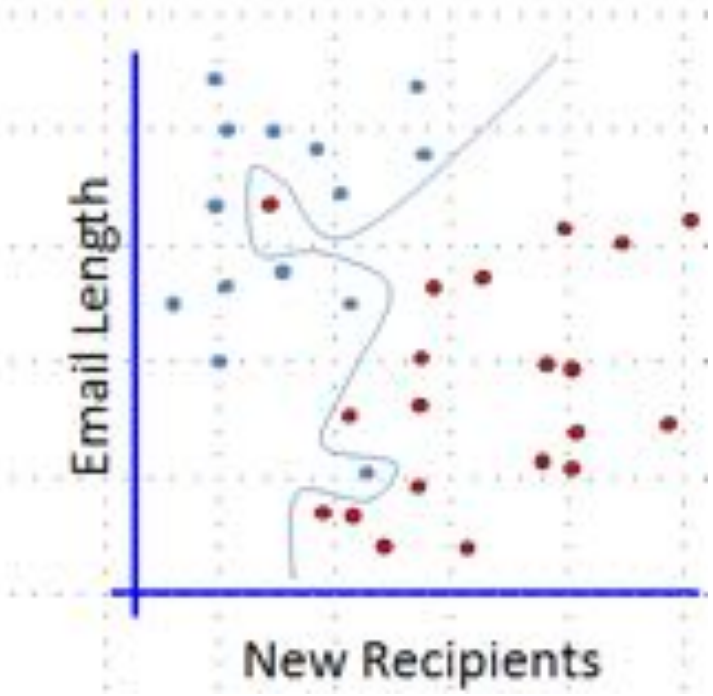


Issue of generalization!

Which one?



2 errors
Simple model



0 errors
Complicated model

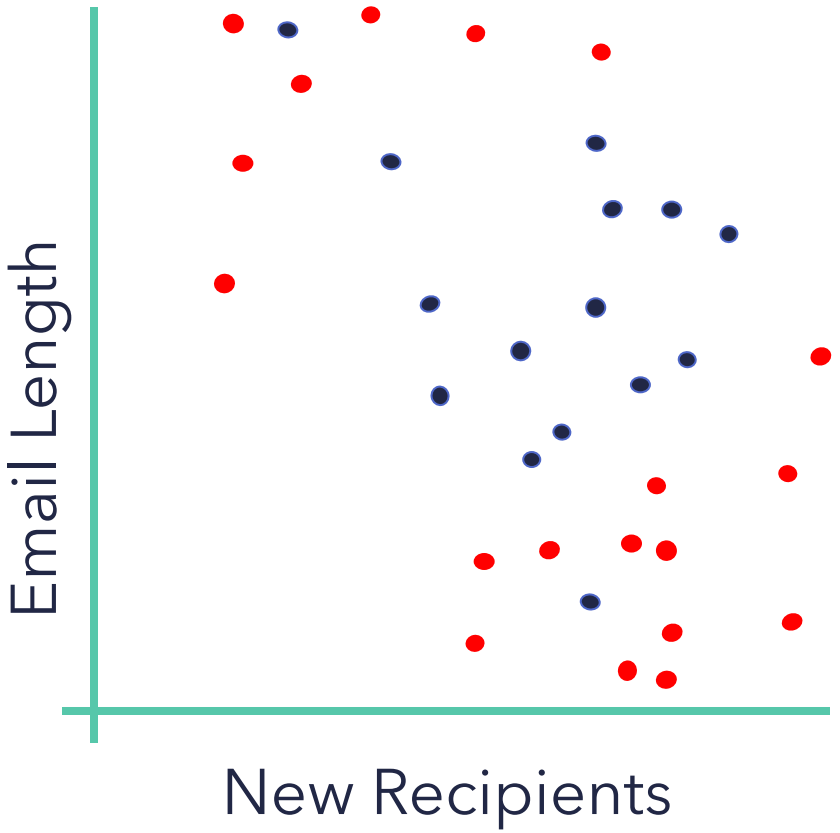
Evaluating what has been learned

1. We randomly select a portion of the data to be used for training (the training set)
2. Train the model on the training set
3. Once the model is trained, we run the model on the remaining instances (the test set) to see how it performs

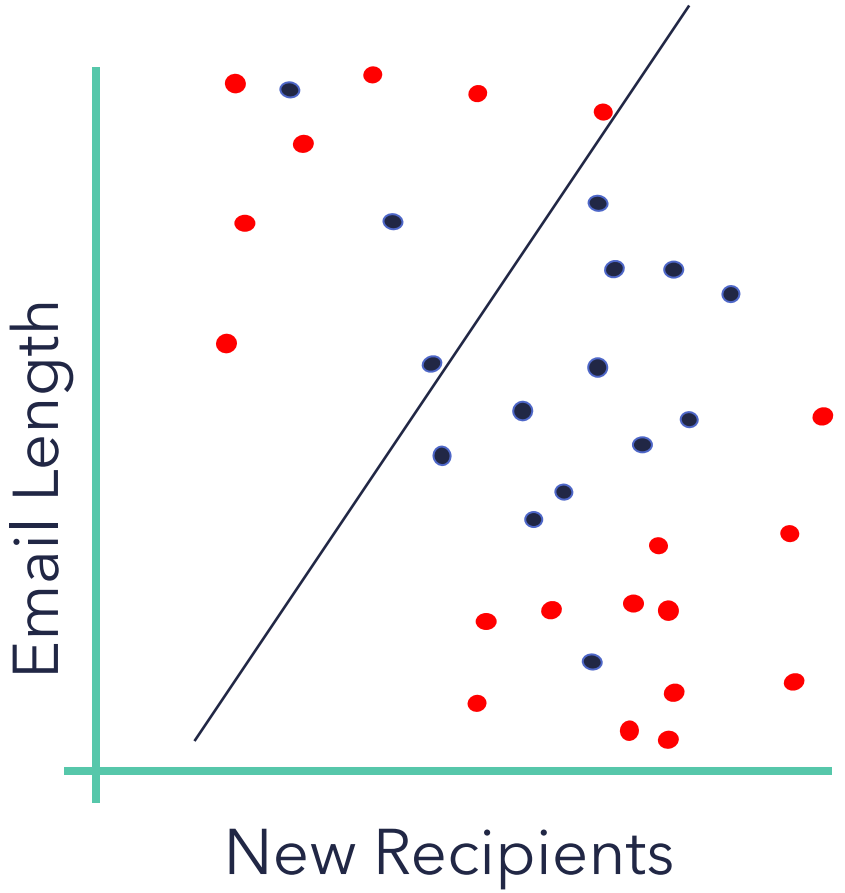
Confusion Matrix

		Classified as: **Black**	Classified as: **Red**
Actual	**Black**	7	1
Actual	**Red**	0	5

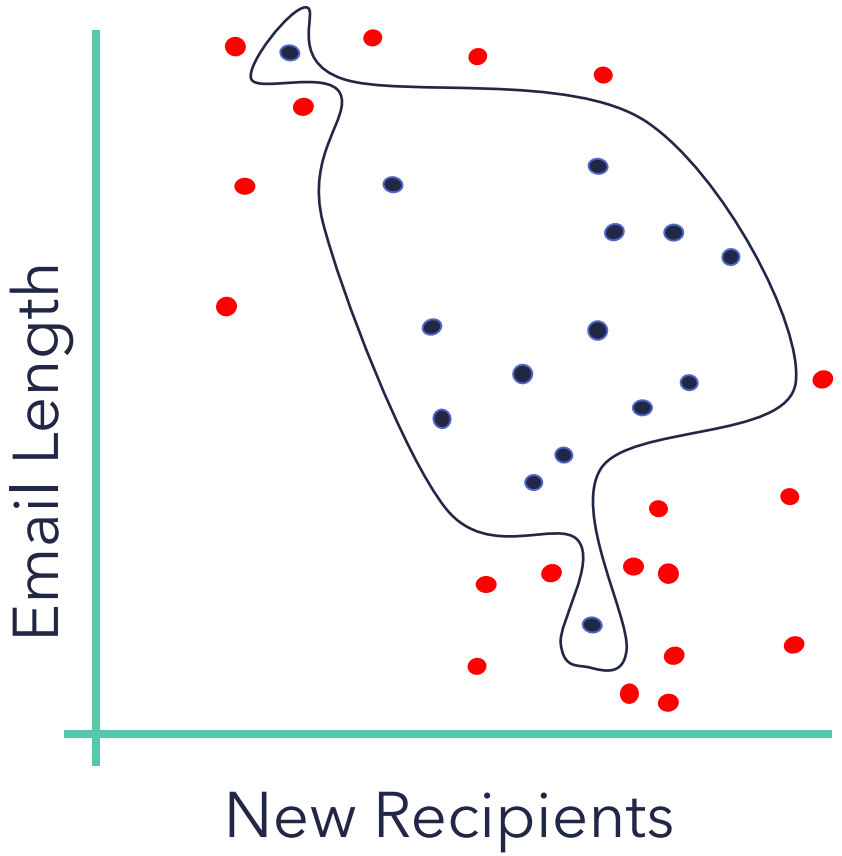
The non-linearly separable case



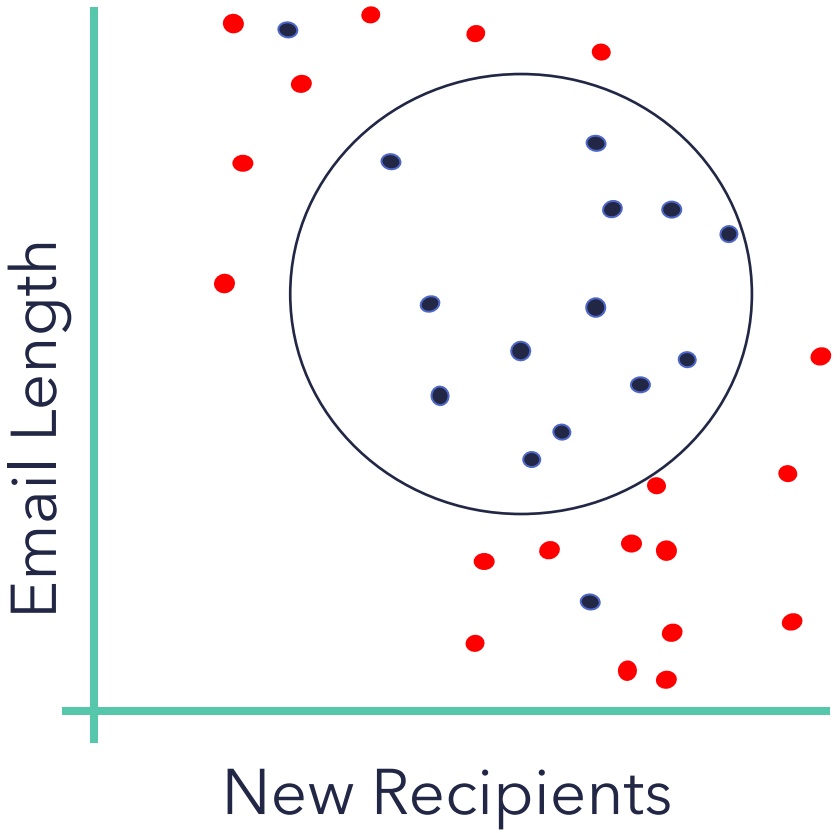
The non-linearly separable case



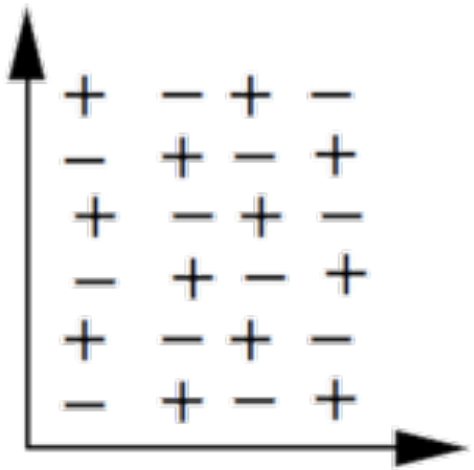
The non-linearly separable case



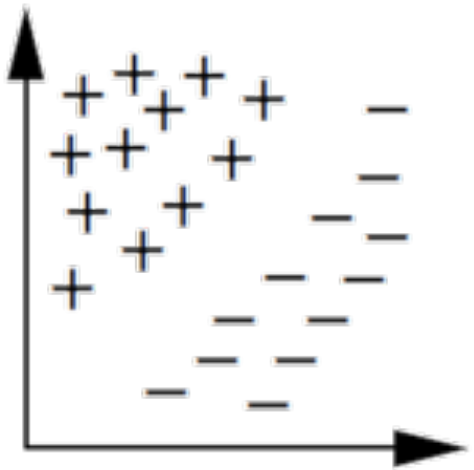
The non-linearly separable case



How good the features are?



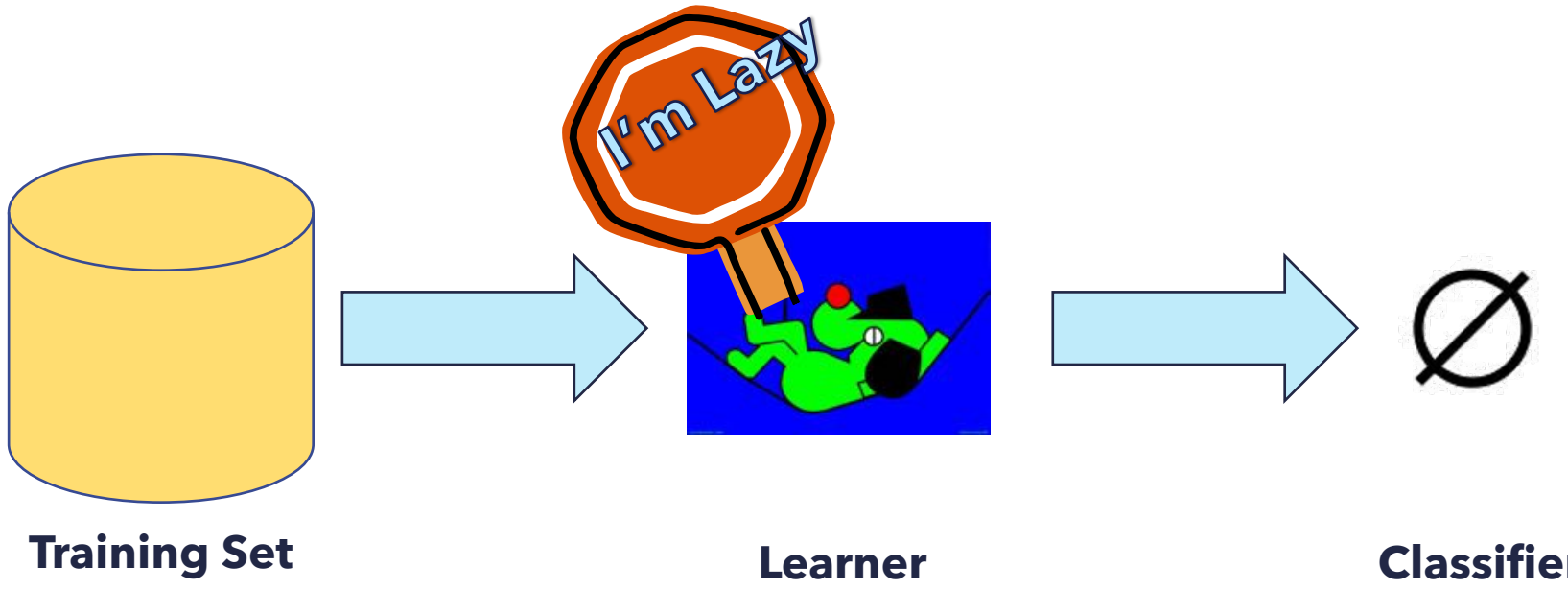
(a)



(b)

Lazy learners

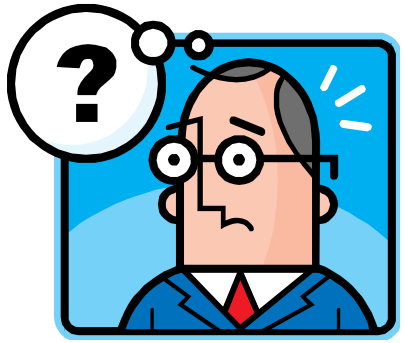
- Generalization beyond the training data is delayed until a new instance is provided to the system



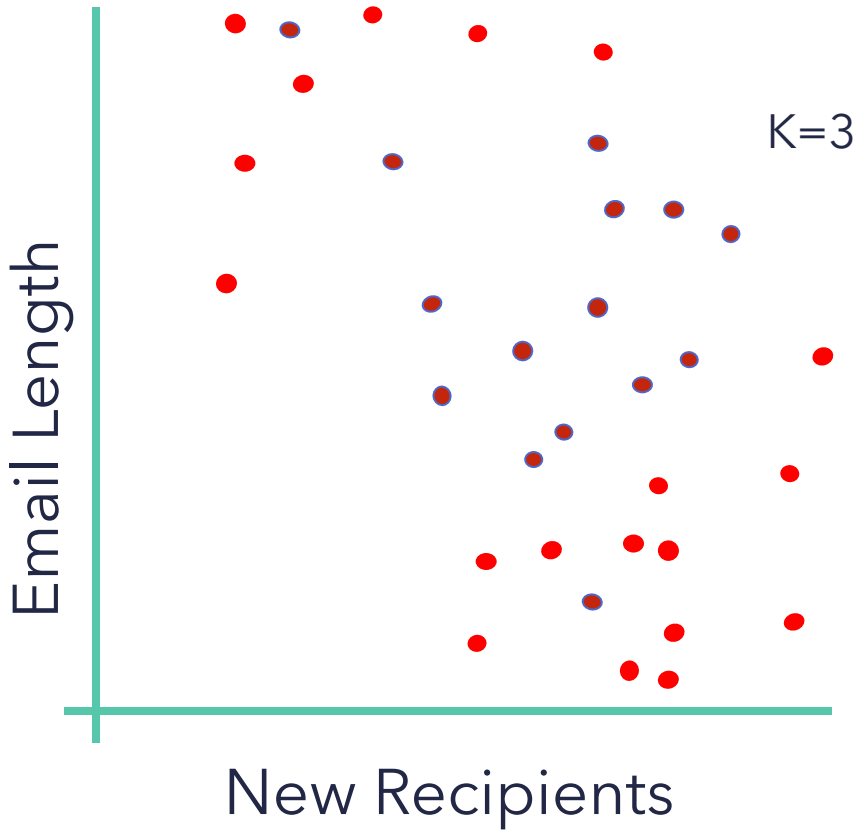
Lazy learners

Instance-based learning

Training Set

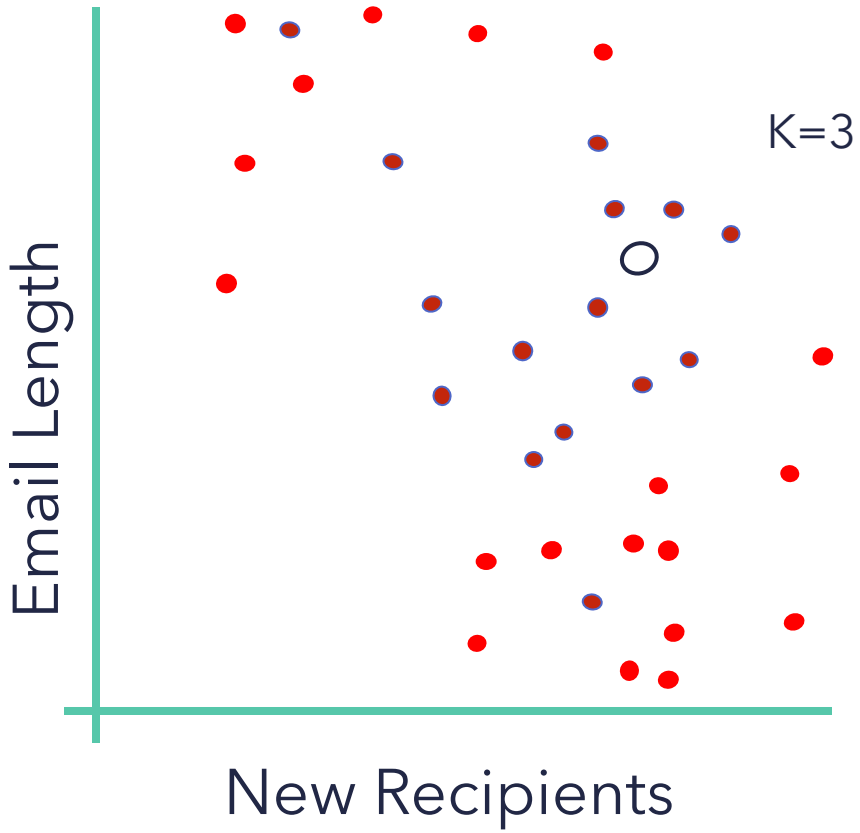


Lazy learner: k-Nearest Neighbors



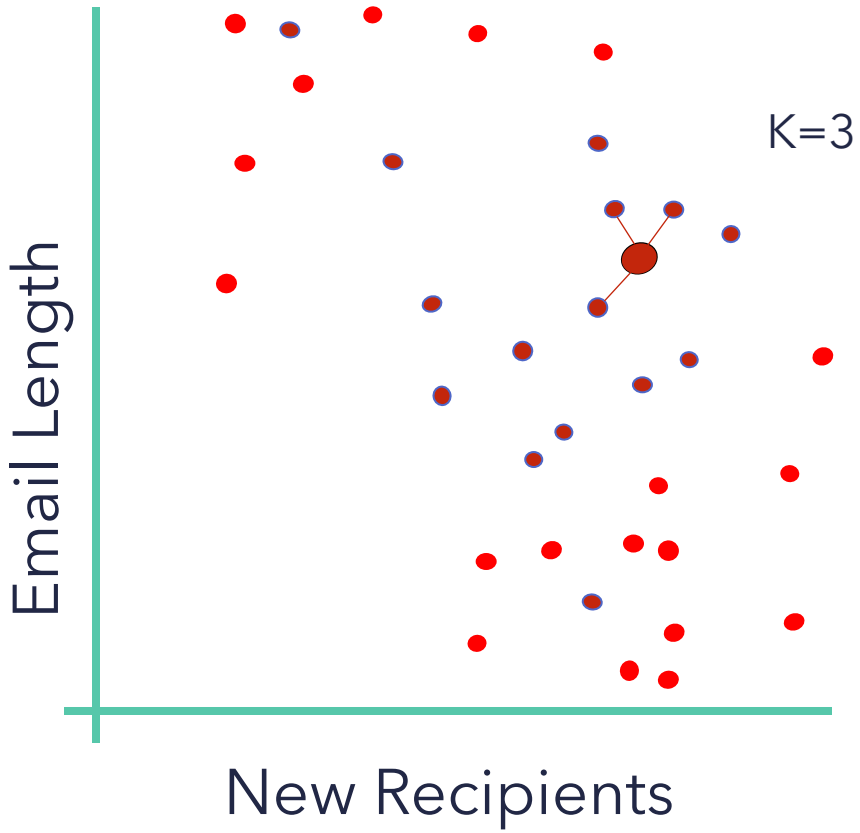
- What should be k?
- Which distance measure should be used?

Lazy learner: k-Nearest Neighbors



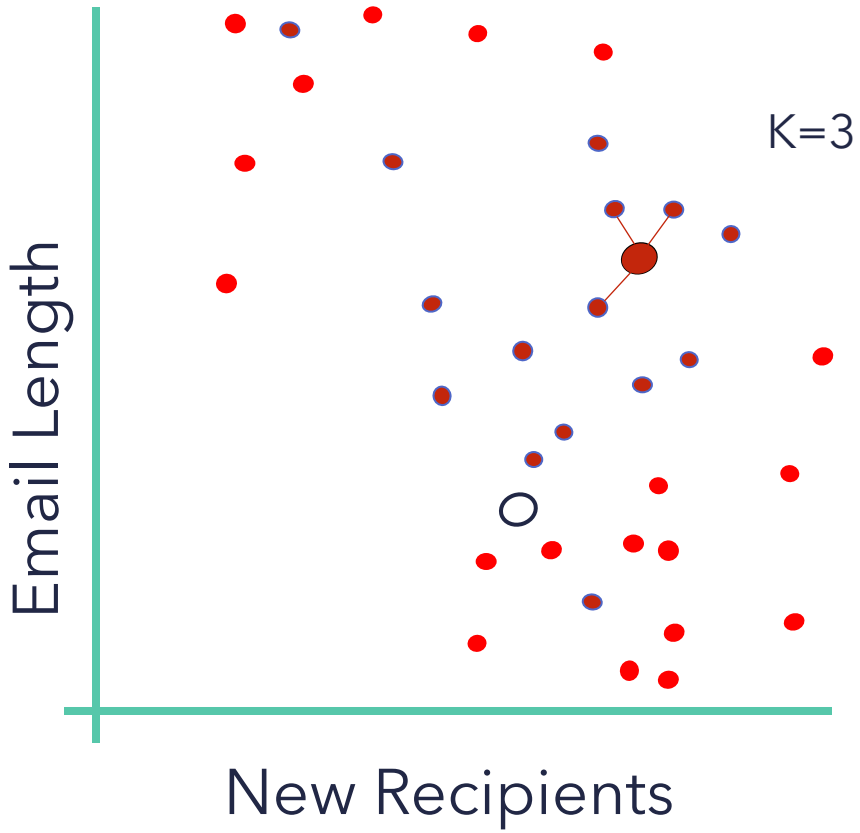
- What should be k?
- Which distance measure should be used?

Lazy learner: k-Nearest Neighbors



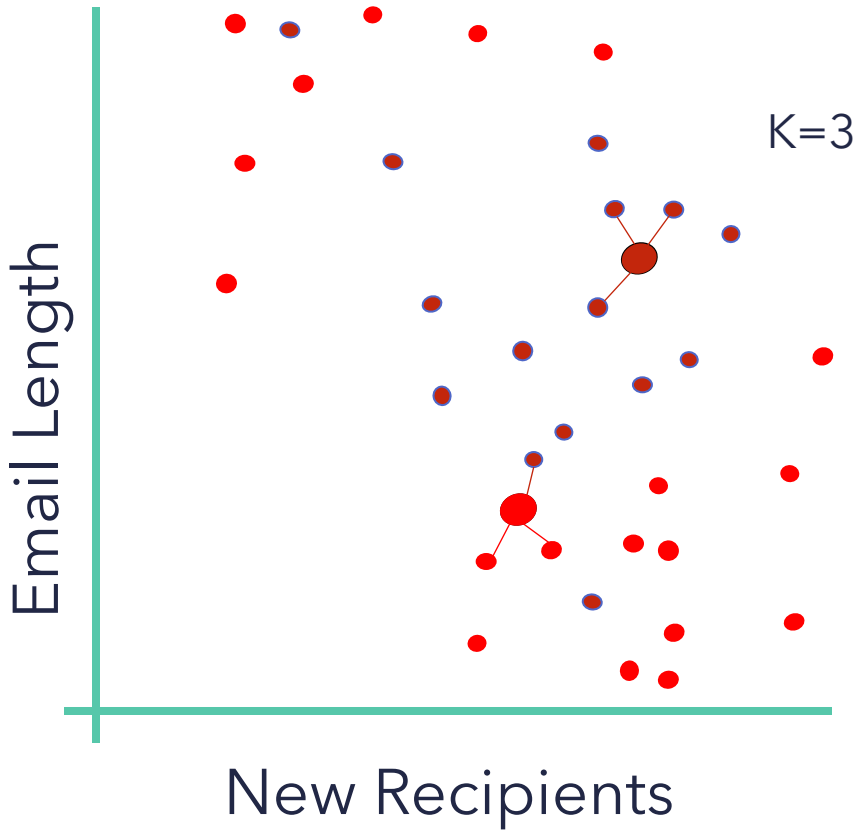
- What should be k?
- Which distance measure should be used?

Lazy learner: k-Nearest Neighbors



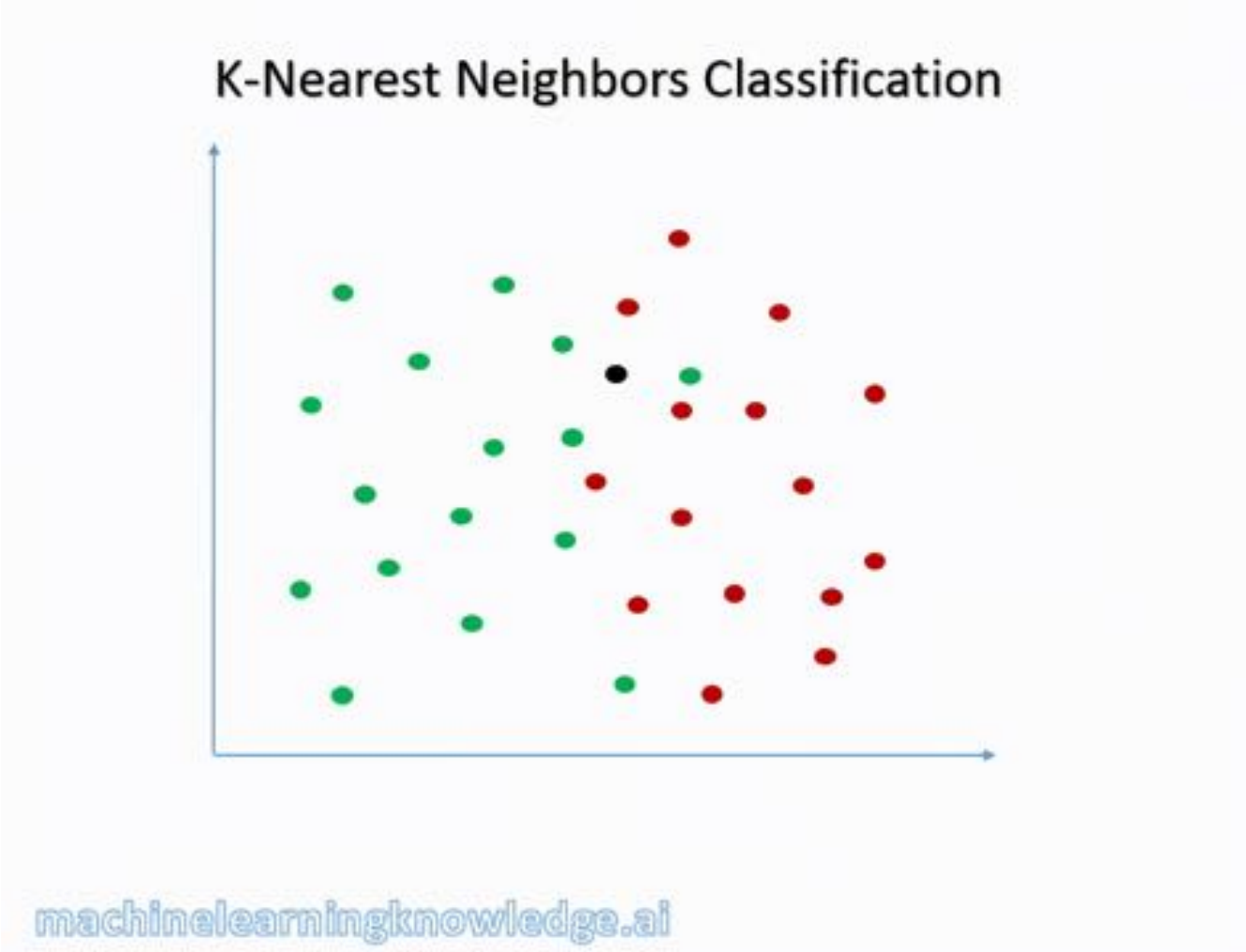
- What should be k?
- Which distance measure should b
  used?

Lazy learner: k-Nearest Neighbors



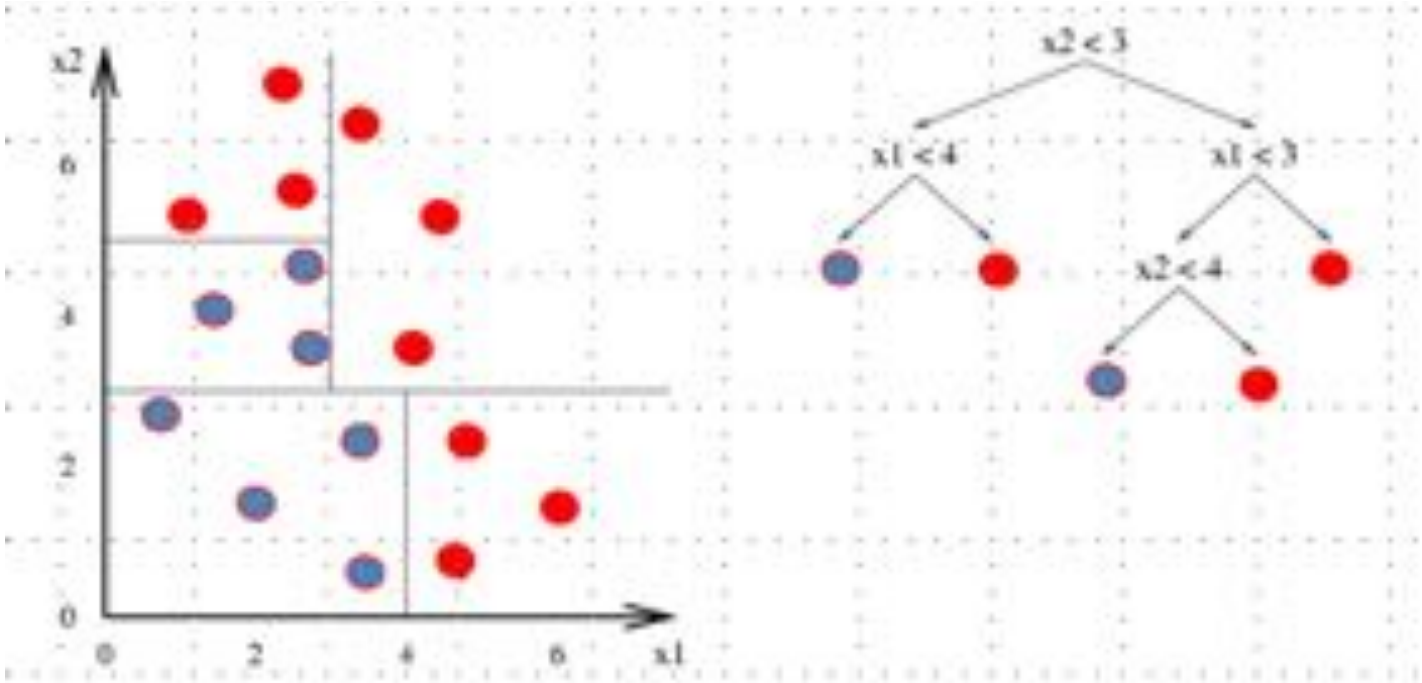
- What should be k?
- Which distance measure shou
  used?

Lazy learner: k-Nearest Neighbors

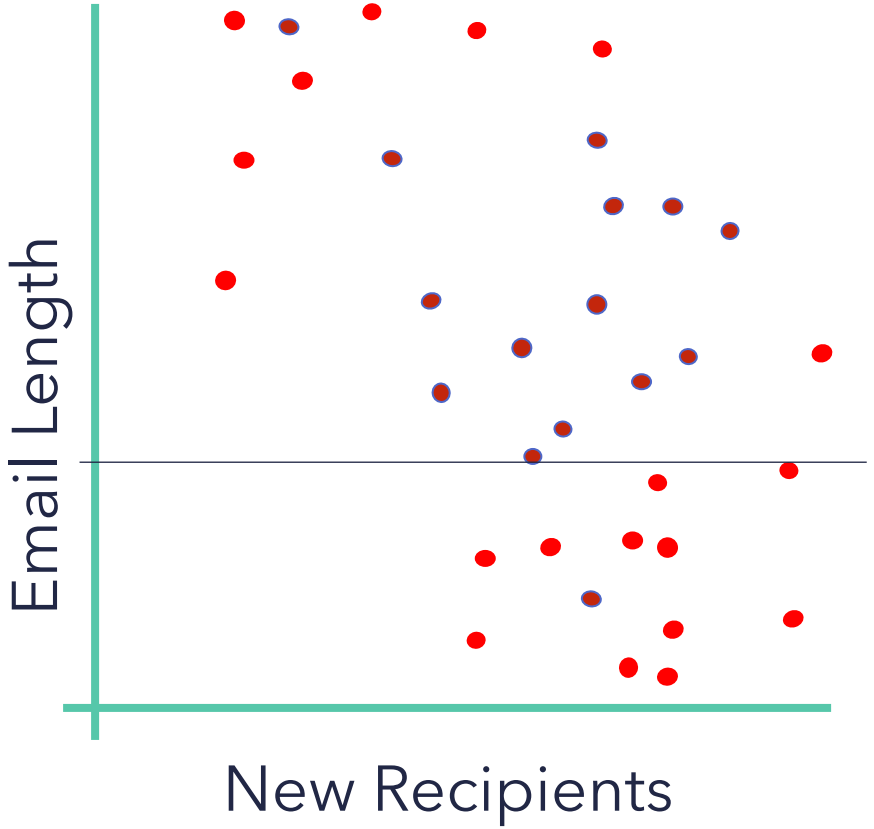


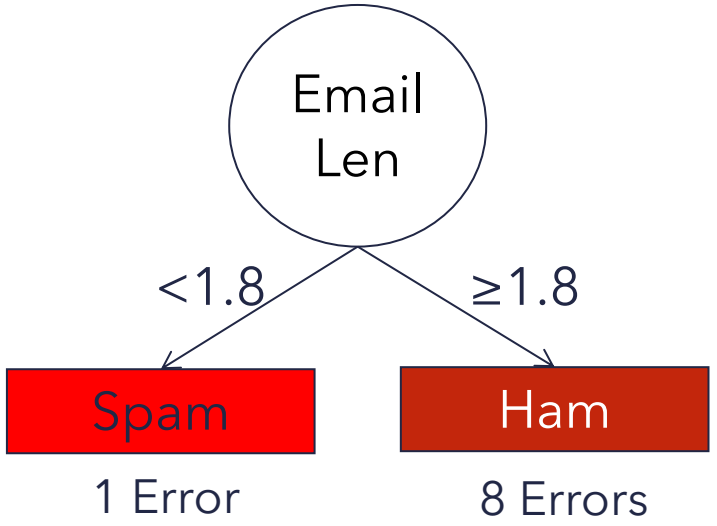
Decision tree

- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution



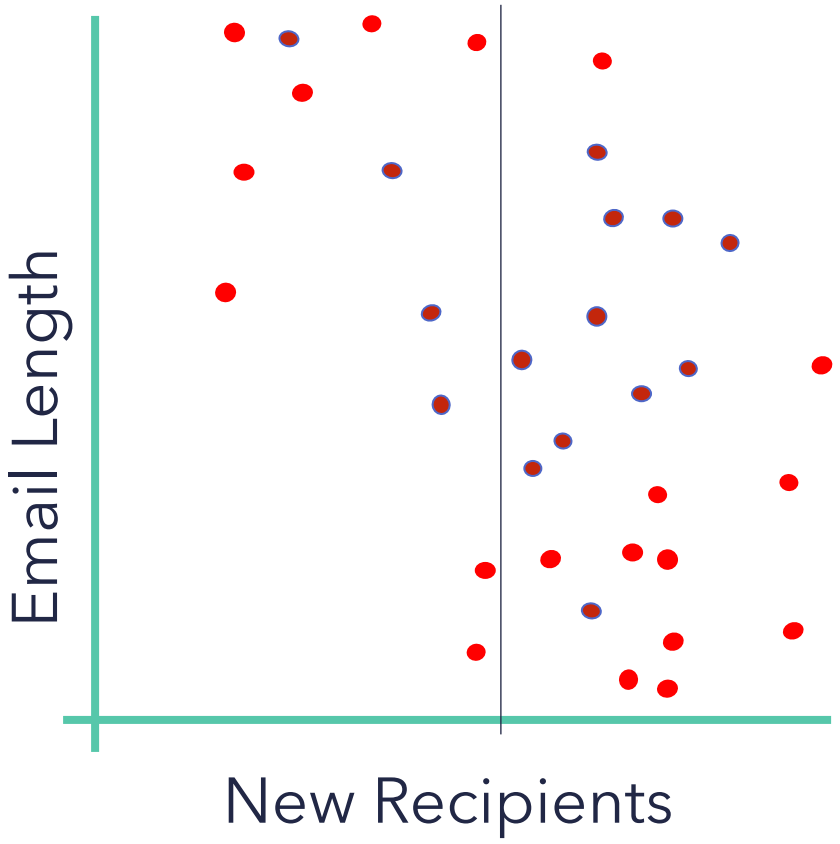
Top down induction of decision trees

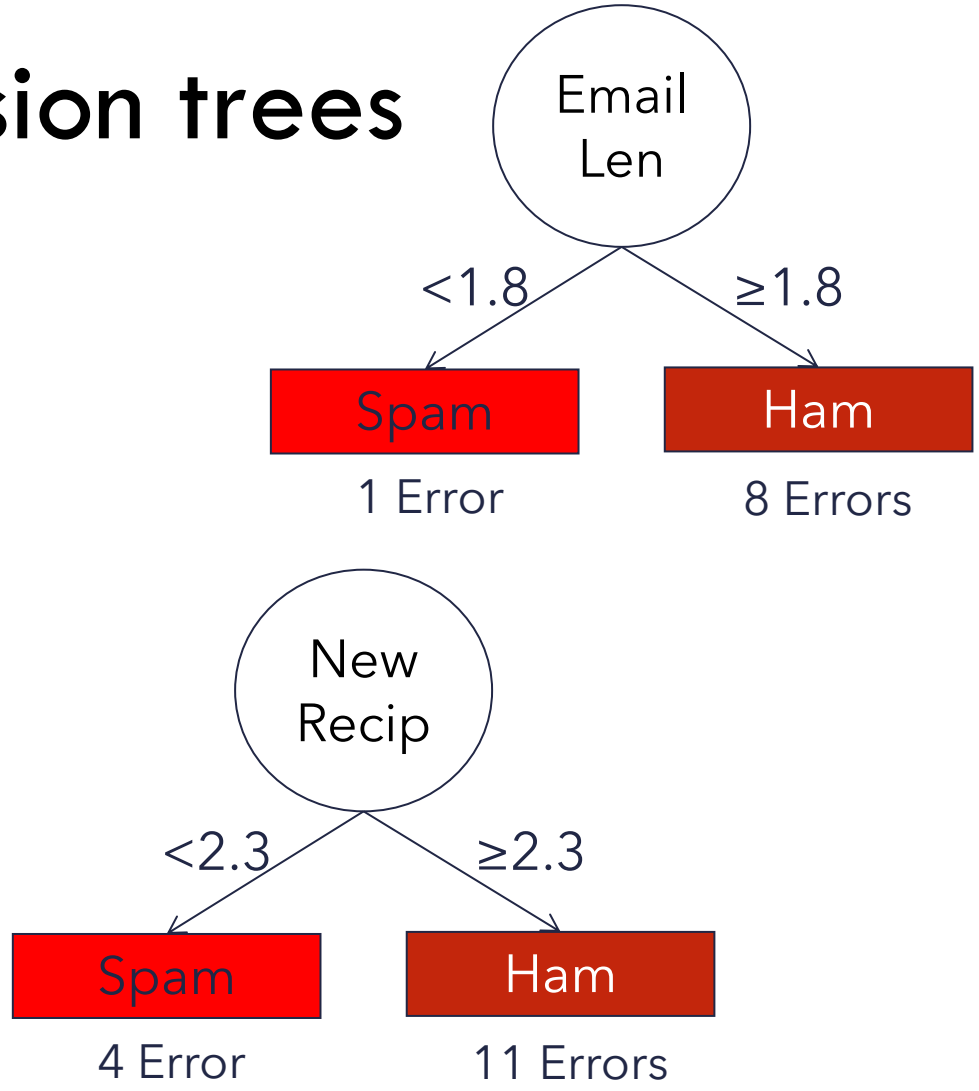




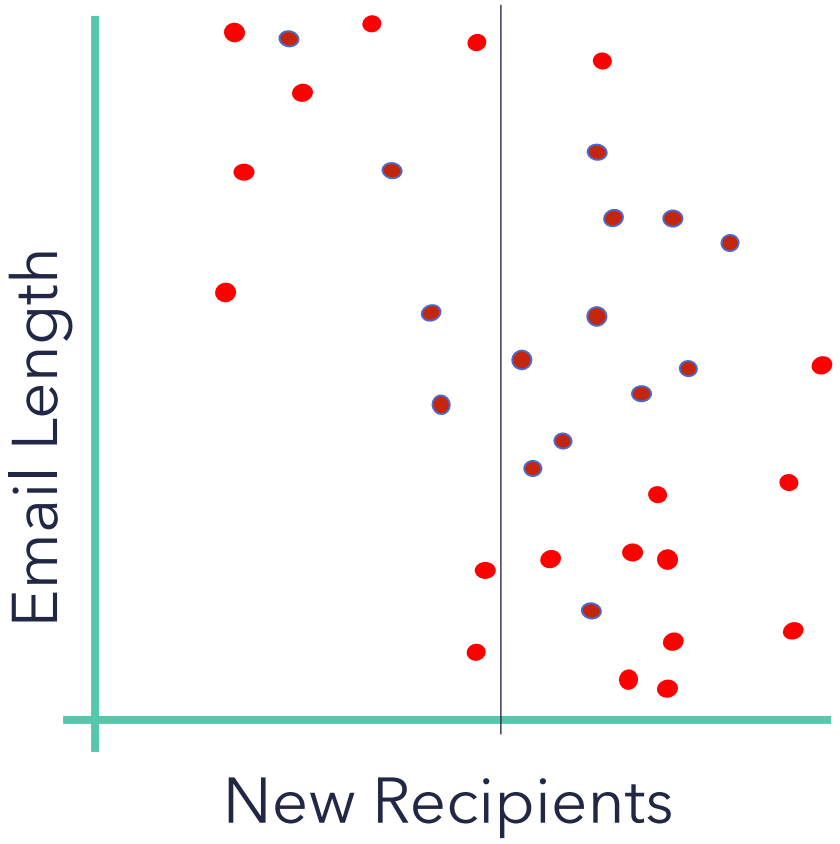
A single level decision tree is also known as Decision Stump

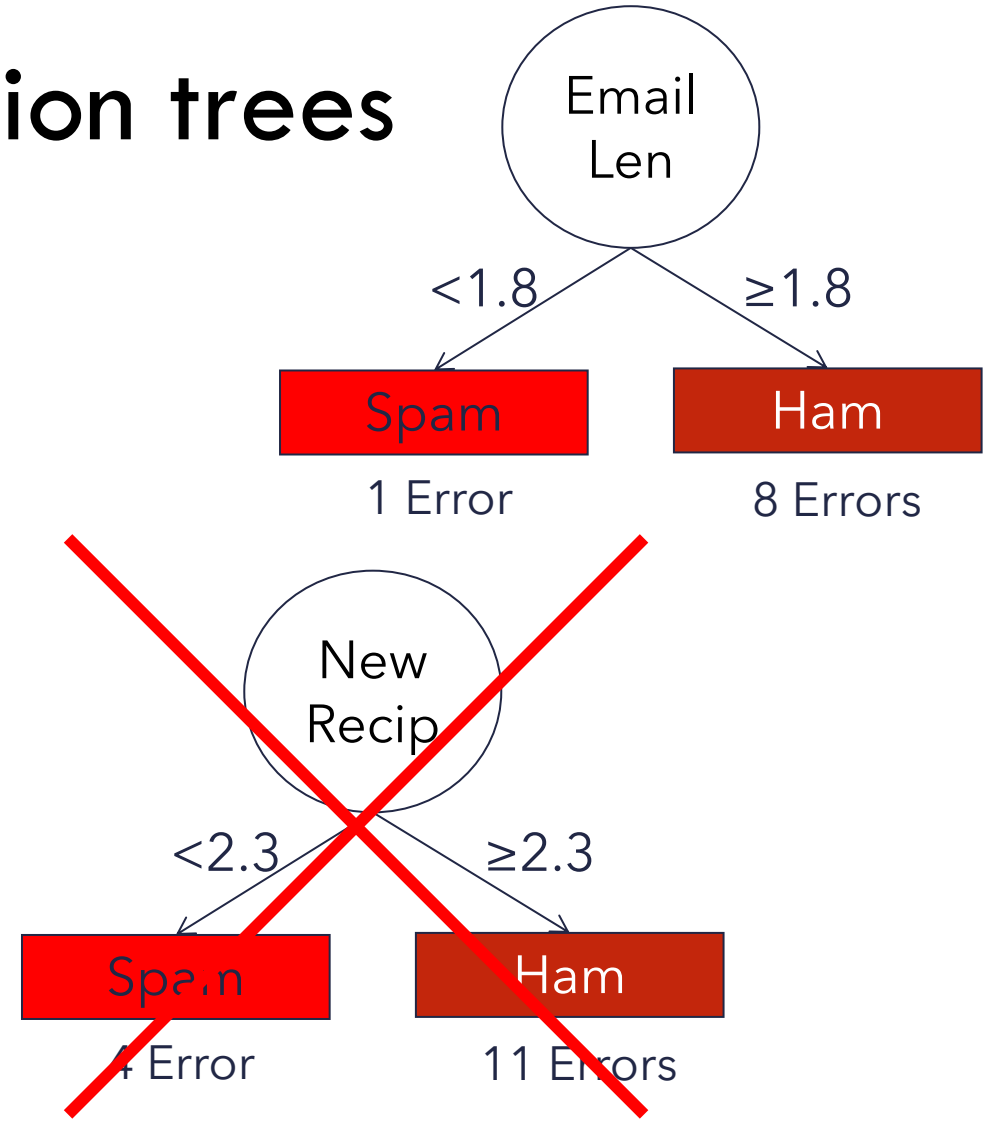
Top down induction of decision trees



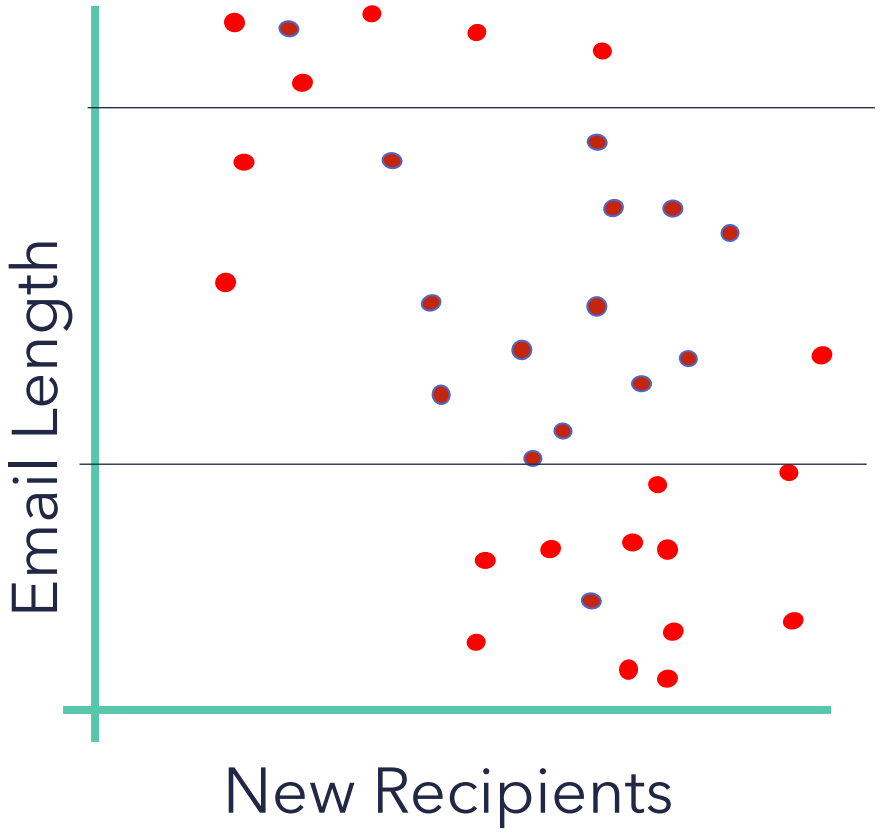


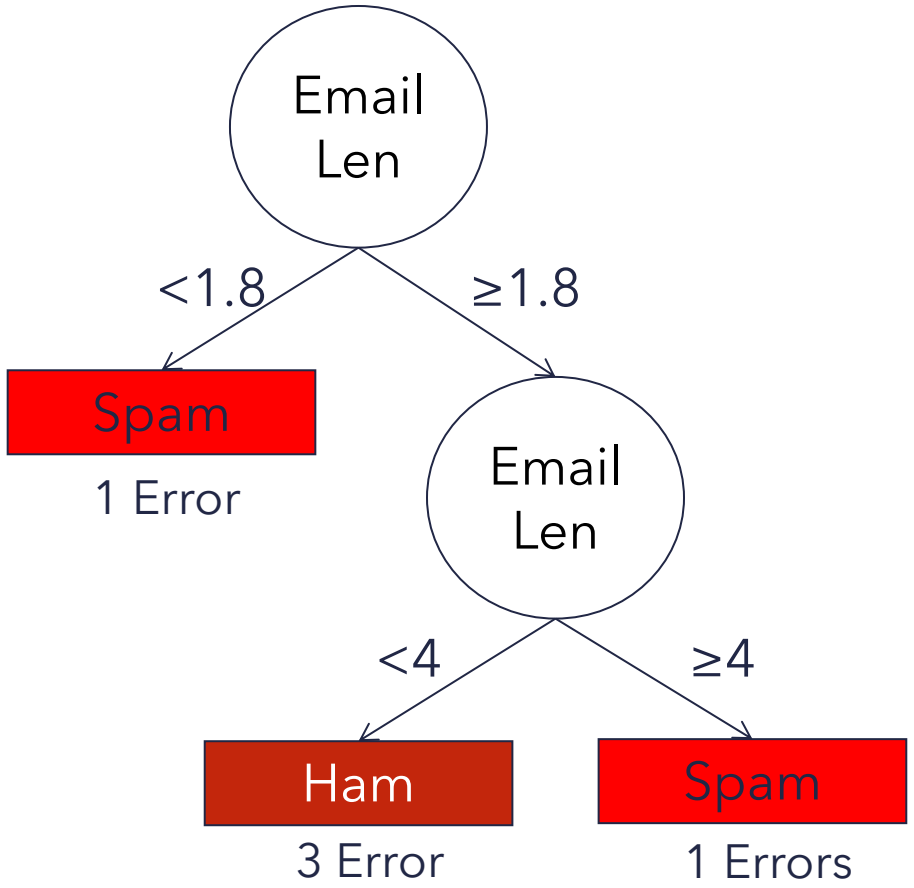
Top down induction of decision trees



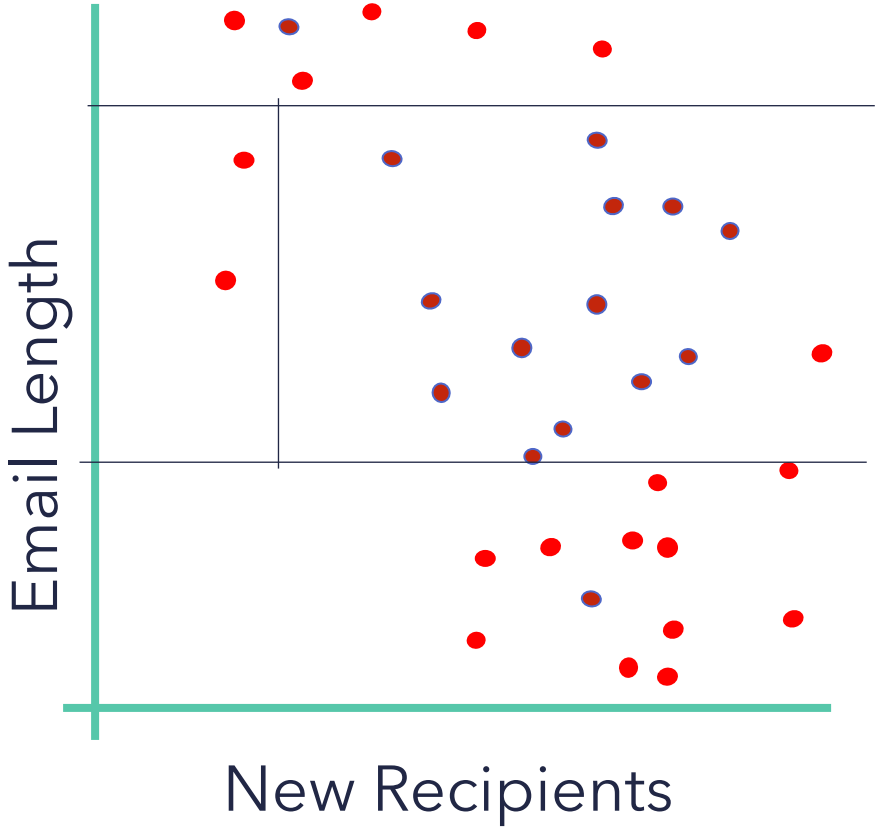


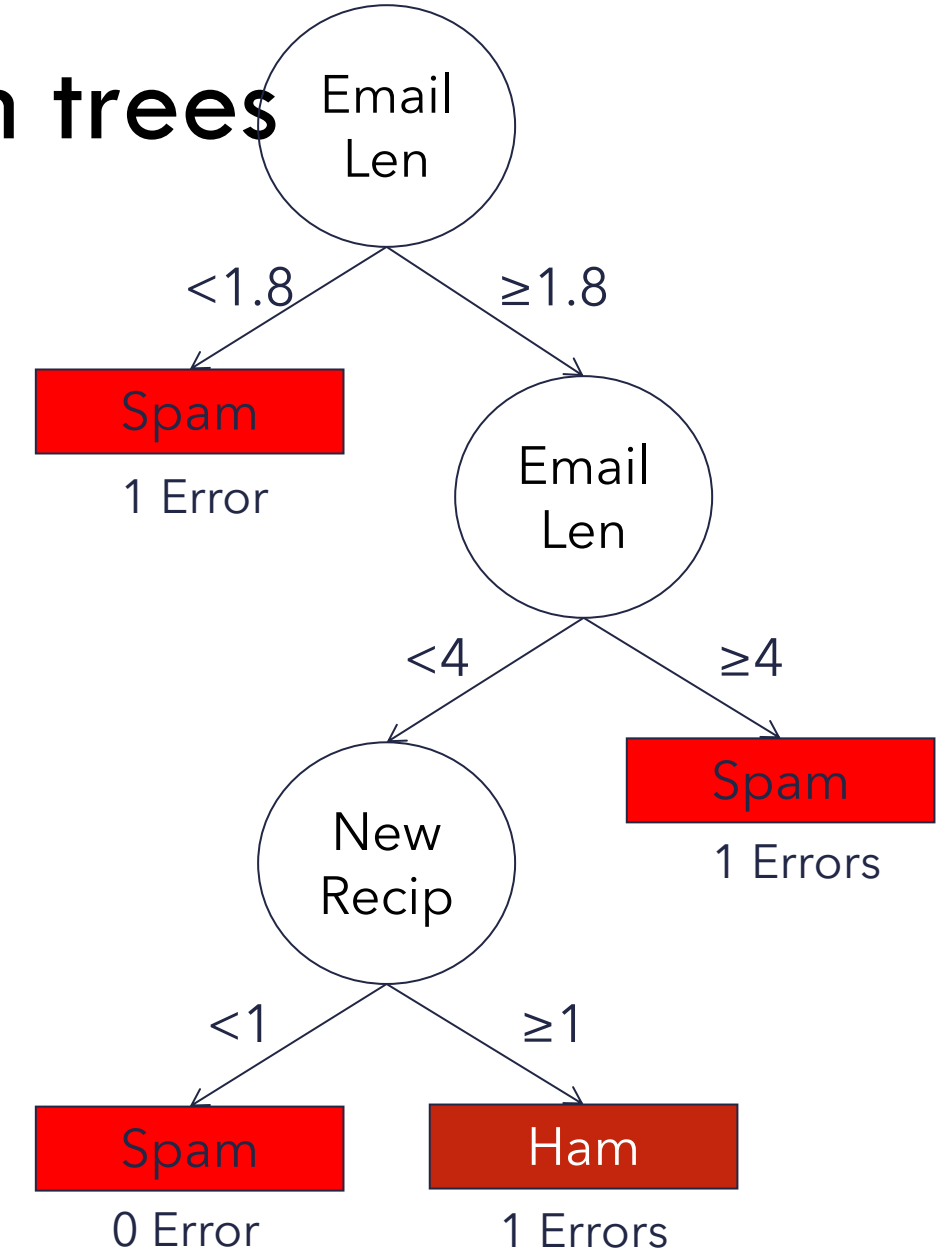
Top down induction of decision trees



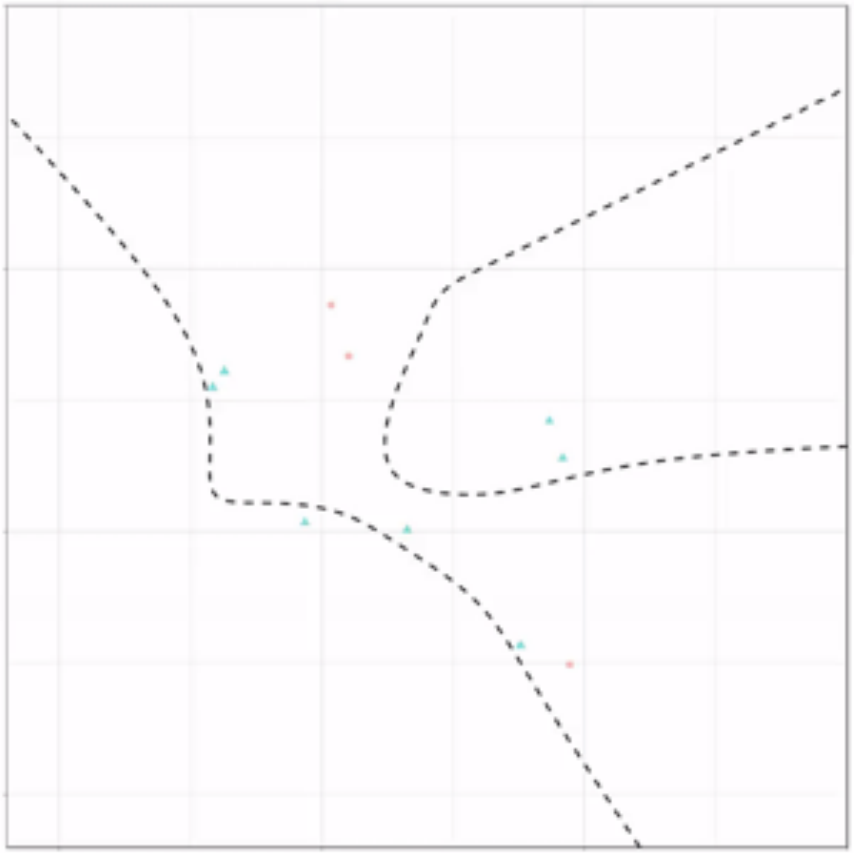


Top down induction of decision trees

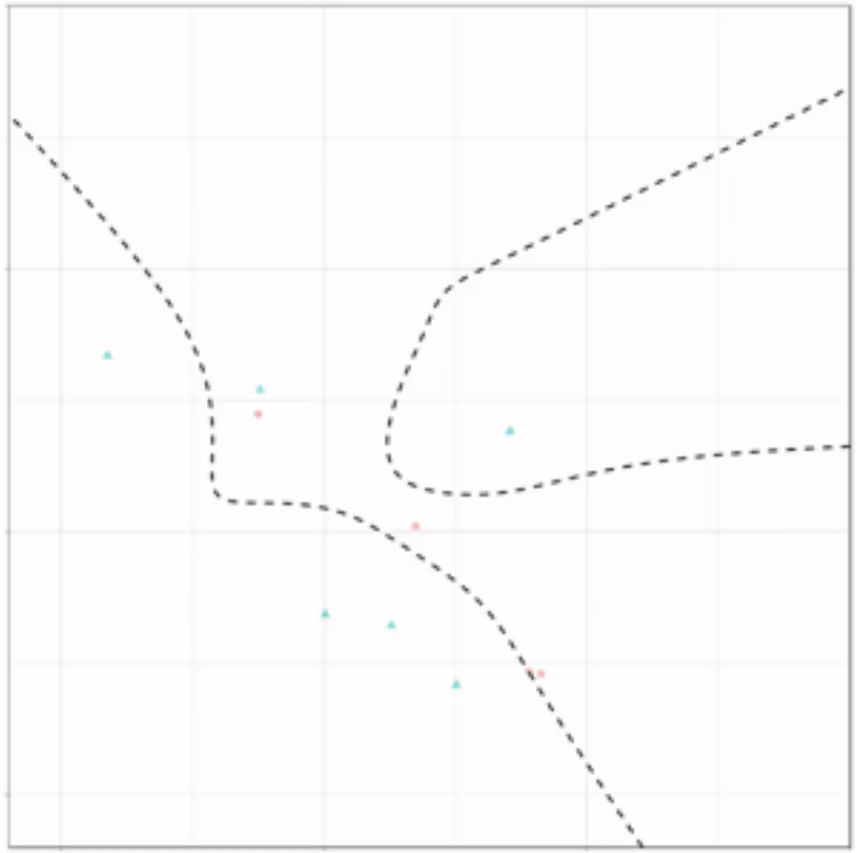




Decision Tree Training

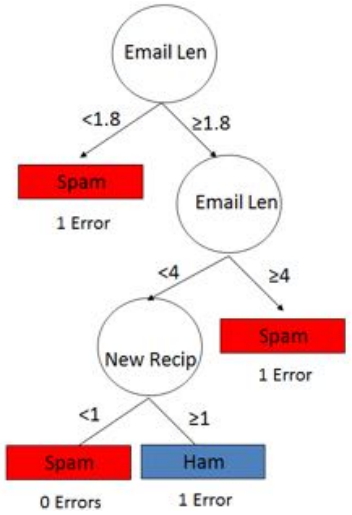


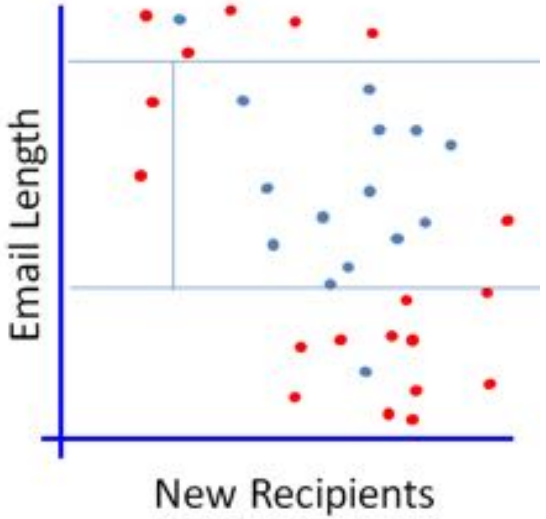
Random Forest, which uses hundreds of trees in the back end and thus results in a more flexible boundary
Random Forest gif by Ryan Holbrook
Link

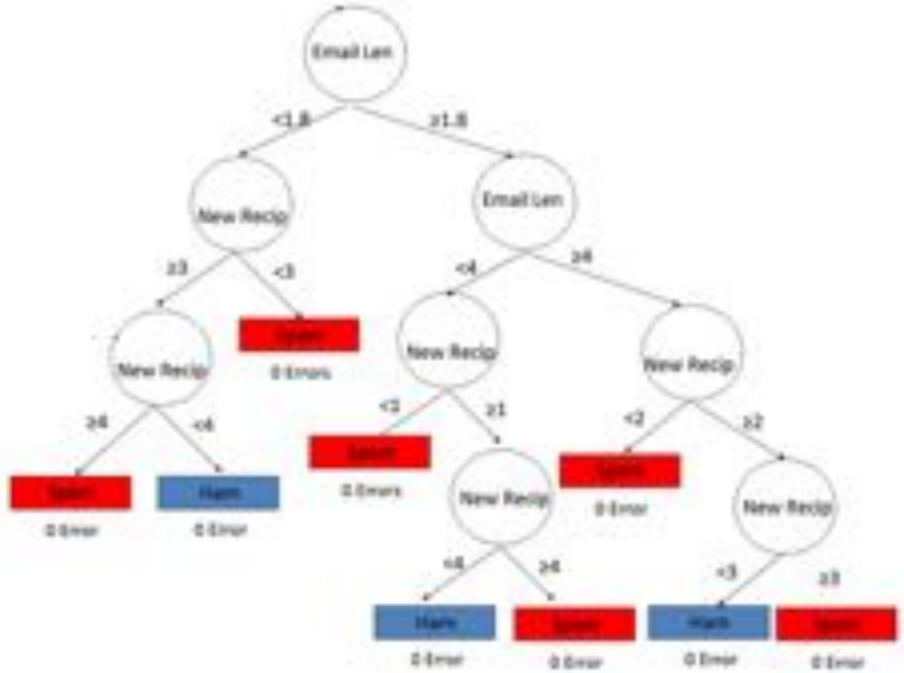


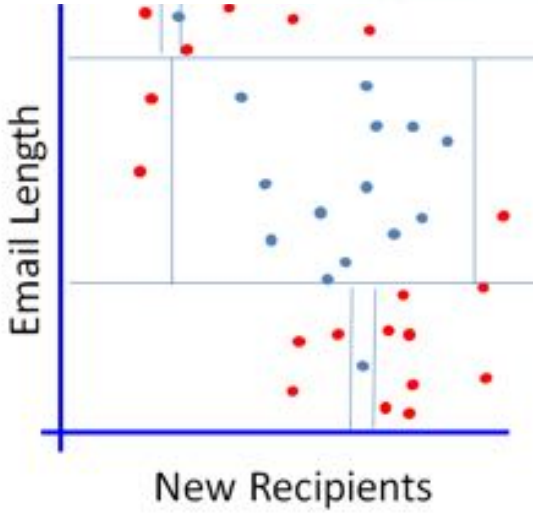
A decision tree classifies data based on multiple, sequential, binary splits.
Decision tree gif by Ryan Holbrook
Link

Which one?


Email Len
<1.8
≥1.8
Spam
1 Error
Email Len
<4
≥4
New Recip
Spam
1 Error
<1
≥1
Spam
Ham
0 Errors
1 Error

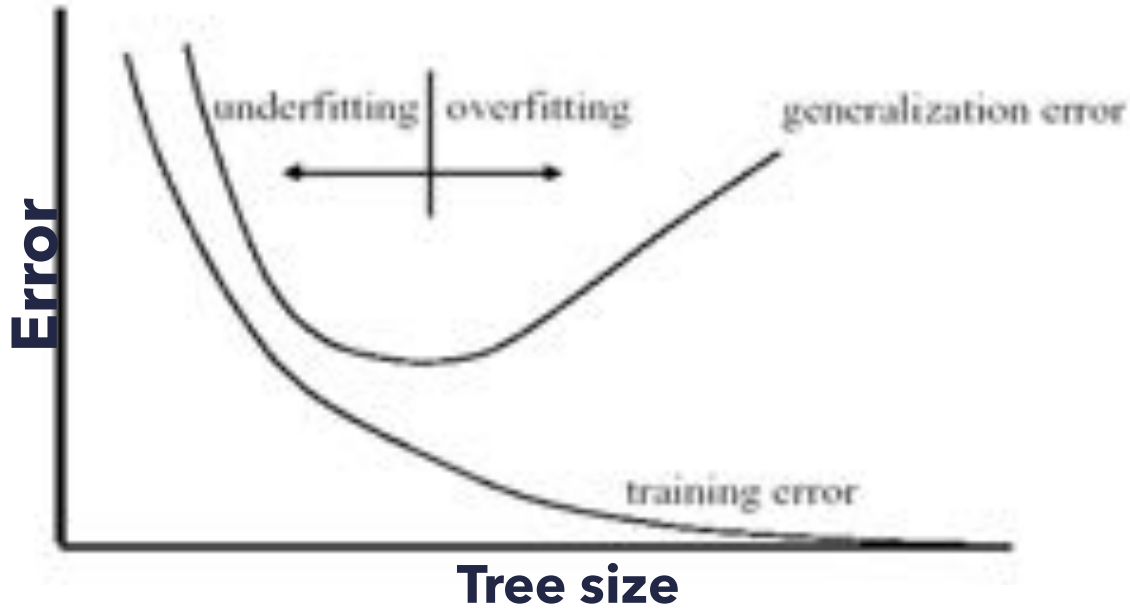

Email Length
New Recipients




Email Length
New Recipients

Overfitting and underfitting

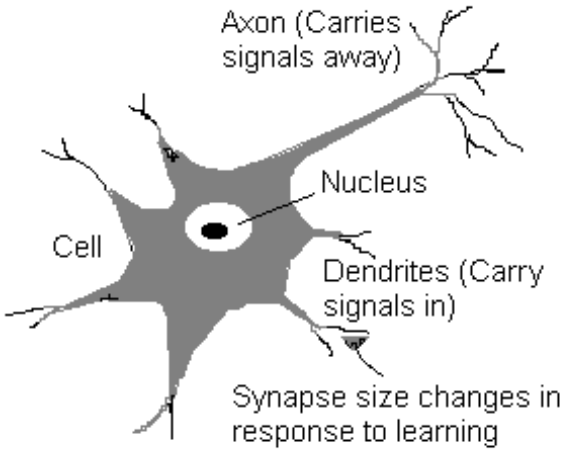
- **Overtraining/Overfitting** - learning the training set too well – it overfits to the training set such that it performs poorly on the test set
- **Underfitting** - when model is too simple, both training and test errors are large



Artificial neural network (ANN)

Inspiration from Neurobiology

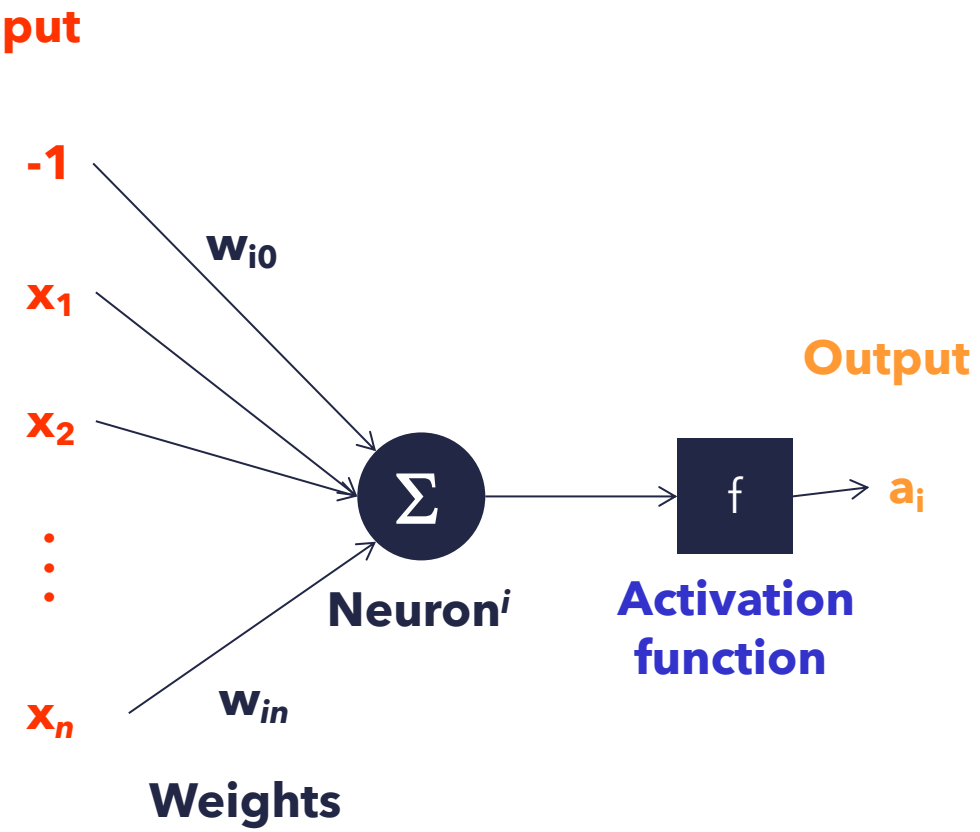
- A neuron has many inputs and one output
- Incoming signals from other neurons determine if the neuron shall excite ("fire")
- Output subject to attenuation in the synapses, which are junction parts of the neuron



Artificial neural network (ANN)

Artificial Neuron Model

- Non linear, parameterized function with restricted output range



Artificial neural network (ANN)

Artificial Neuron Model

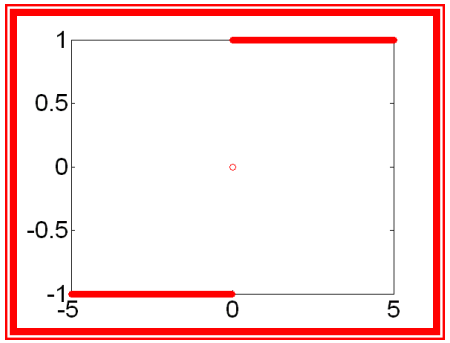
- Neuron *i* computes its activation level a_i:

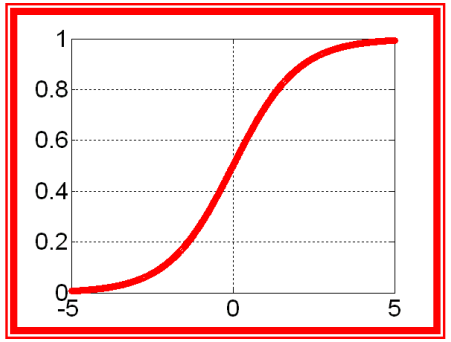
$$a_i = f(n_i) = f(\sum_{i=0}^{n} w_{ij}\, x_j)$$

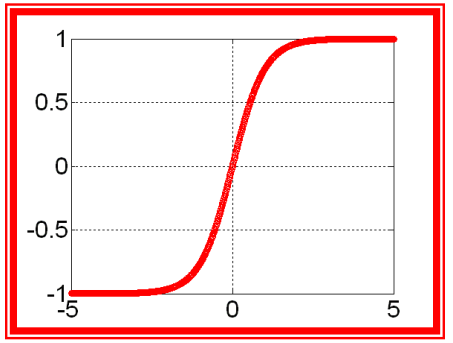
- where *f* is one of the several possible activation functions (normalize the value 0-1)

Activation functions

Name	Input/Output Relation	Icon	MATLAB Function
Hard Limit	$a = 0 \quad n < 0$ $a = 1 \quad n \geq 0$		hardlim
Symmetrical Hard Limit	$a = -1 \quad n < 0$ $a = +1 \quad n \geq 0$		hardlims
Linear	$a = n$		purelin
Saturating Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n \leq 1$ $a = 1 \quad n > 1$		satlin
Symmetric Saturating Linear	$a = -1 \quad n < -1$ $a = n \quad -1 \leq n \leq 1$ $a = 1 \quad n > 1$		satlins
Log-Sigmoid	$a = \frac{1}{1+e^{-n}}$		logsig
Hyperbolic Tangent Sigmoid	$a = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}}$		tansig
Positive Linear	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n$		poslin
Competitive	$a = 1$ neuron with max n $a = 0$ all other neurons	C	compet







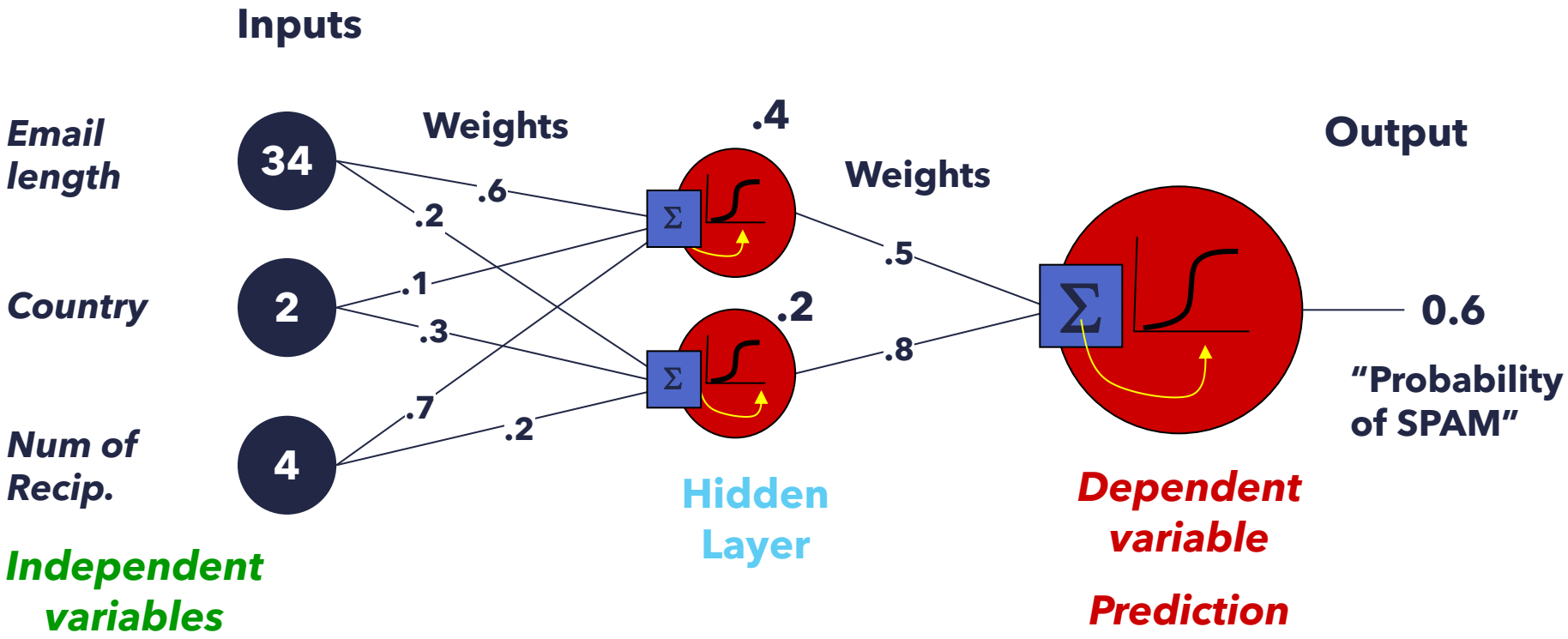
Artificial neural network (ANN)

Training procedure

- The process of determining the values for *W* on the basis of the data is called training or learning

- The training Here is like any Non-Linear Regression

Neural Network Model



Neural Network Model

Back propagation

- Find the set of weights that minimize:

$$\sum_i (y_i - out(x_j))^2$$

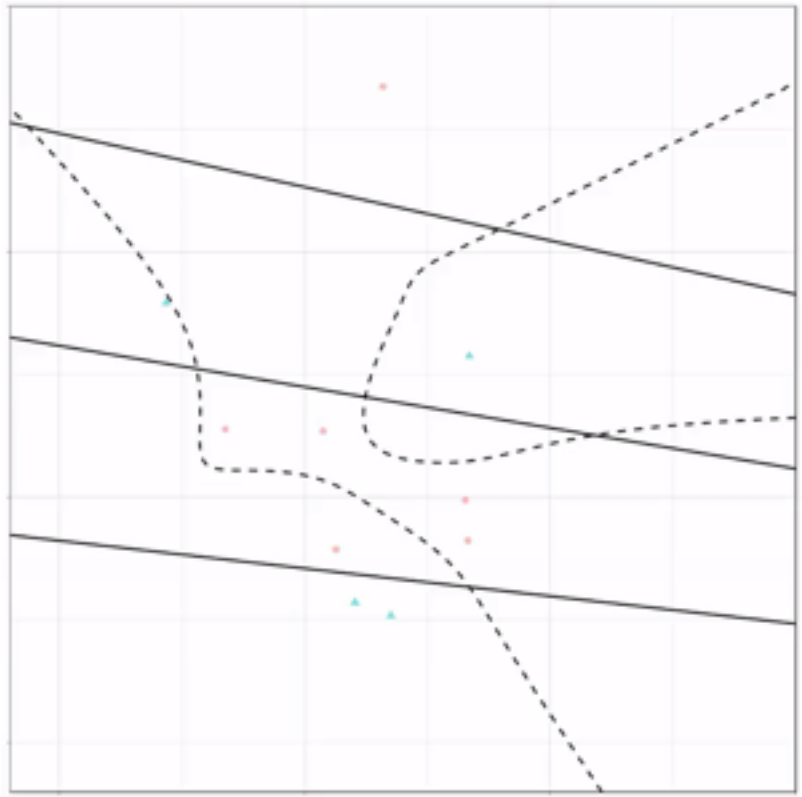
  - calculate first the changes for the weights of the output neuron;
  - calculate the changes backward starting from layer p-1, and propagate backward the local error terms

Neural Network Model

Back propagation

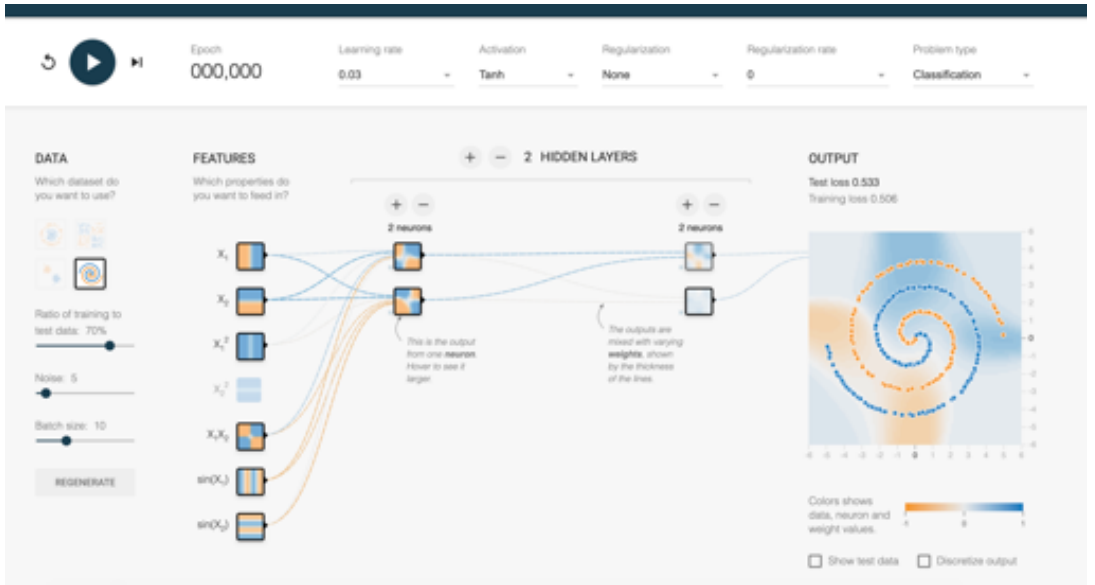
- Convergence to a global minimum is not guaranteed
- Tweaking to find the right number of hidden units, or a useful learning rate
- Learning rate
  - too small: can take days instead of minutes to converge
  - too large: diverges (MSE gets larger and larger while the weights increase and usually oscillate)
- Usually, it is enough to have a single layer of nonlinear neurons in a neural network in order to learn to approximate a nonlinear function
- In such case general optimisation may be applied without too much difficulty

ANN Training



ANN gif by Ryan Holbrook
Link

TRY THIS SIMULATOR : LINK

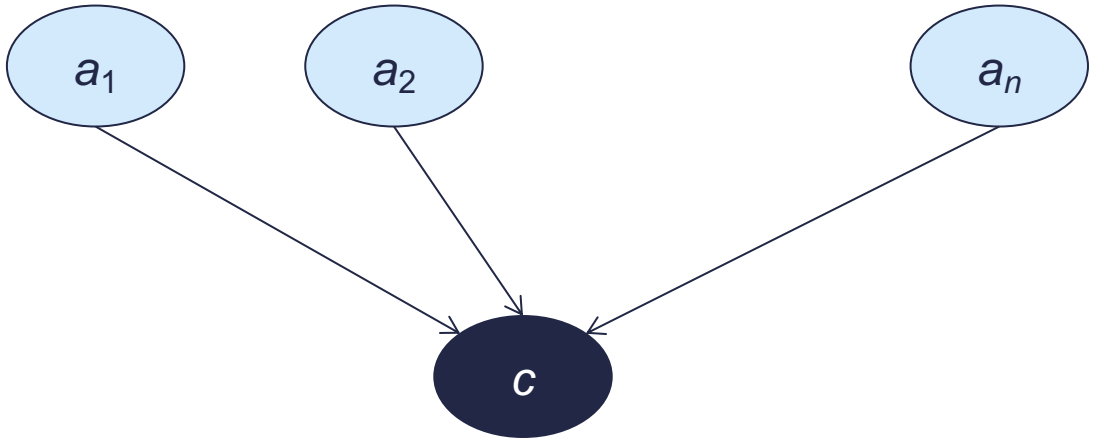


Naïve Bayes classifier

- **Naïve Bayes assumption**: attributes that describe data instances are conditionally independent given the classification hypothesis

$$P(a_1, a_2 \ldots a_n|c) = \prod_i P(a_i|c)$$

$$P(A,B|C) = P(A|C) * P(B|C)$$



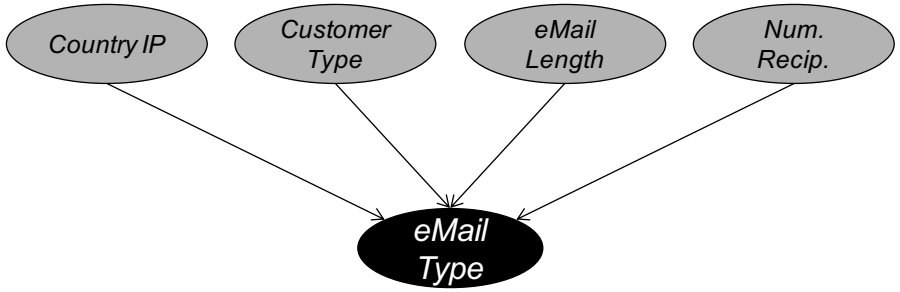
Example

Data

Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
10	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Russia	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
10	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Example

Learning phase



CountryIP	Spam	Ham
Germany	0/3	3/3
Russia	2/2	0/2
USA	1/2	1/2

Customer Type	Spam	Ham
Gold	0/2	2/2
Silver	1/2	1/2
Bronze	2/3	1/3

Length	Spam	Ham
≤2	2/4	2/4
>2	1/3	2/3

Recip.	Spam	Ham
≤4	2/4	2/4
>4	1/3	2/3

$P(\text{Spam}) = 3/7$

$P(\text{Ham}) = 4/7$

Example

Test phase

- Given a new instance

x=(Country=*USA*, *Customer*=*Silver*, Length=3, Recip.=6)

P(Country=USA|Spam) = 1/2

P(Customer=Silver|Spam) = 1/2

P(Length>2|Spam) = 1/3

P(Recip.>4|*Spam*) = 1/3

P(*Spam*) = 3/7

P(Country=USA|Ham) = 1/2

P(Customer=Silver|Ham) = 1/2

P(Length>2|Ham) = 2/3

P(Recip.>4|*Ham*) = 2/3

P(*Ham*) = 4/7

$$\underset{j}{\operatorname{argmax}}\, P(c_j)\, P(\boldsymbol{x}|c_j) = \underset{j}{\operatorname{argmax}}\, P(c_j) \prod_i P(a_i|c_j)$$

Given the fact P(*Spam*|**x**) < P(*Ham*|**x**), we label **x** to be "*Ham*"

Relevant issues

$$P(a_1, a_2 \ldots a_n | c) \neq \prod_i P(a_i | c)$$

- Violation of independence assumption
  - The assumption of independence means that your data isn't connected in any way (at least, in ways that you haven't accounted for in your model).
  - No example contains the attribute value
- Continuous attributes
- Positive : Training and testing are very easy and fast

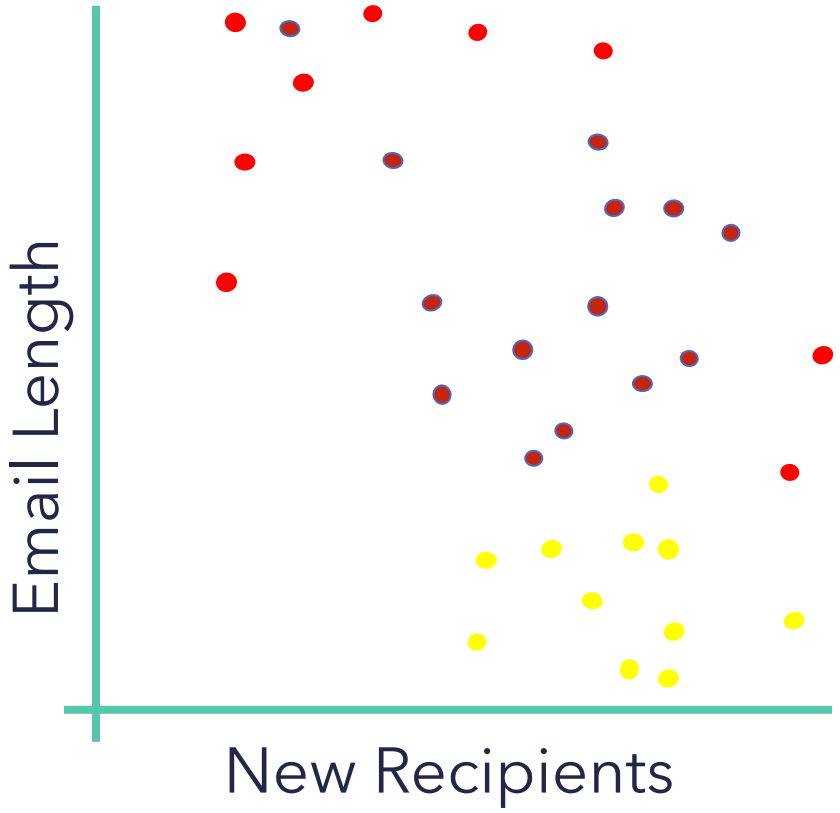
Ensemble learning

- The idea is to use multiple models to obtain better predictive performance than could be obtained from any of the constituent models

- Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified

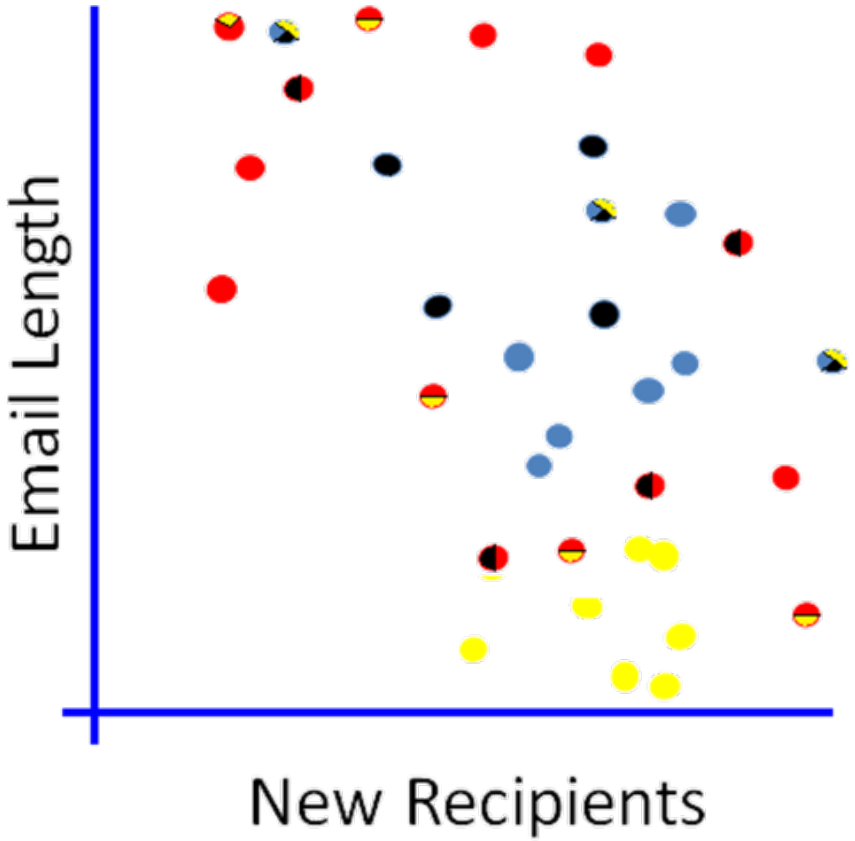
Other learning tasks

Supervised learning - multi class



Supervised learning - multi label

- Classification problem where each example can be assigned to multiple class labels simultaneously

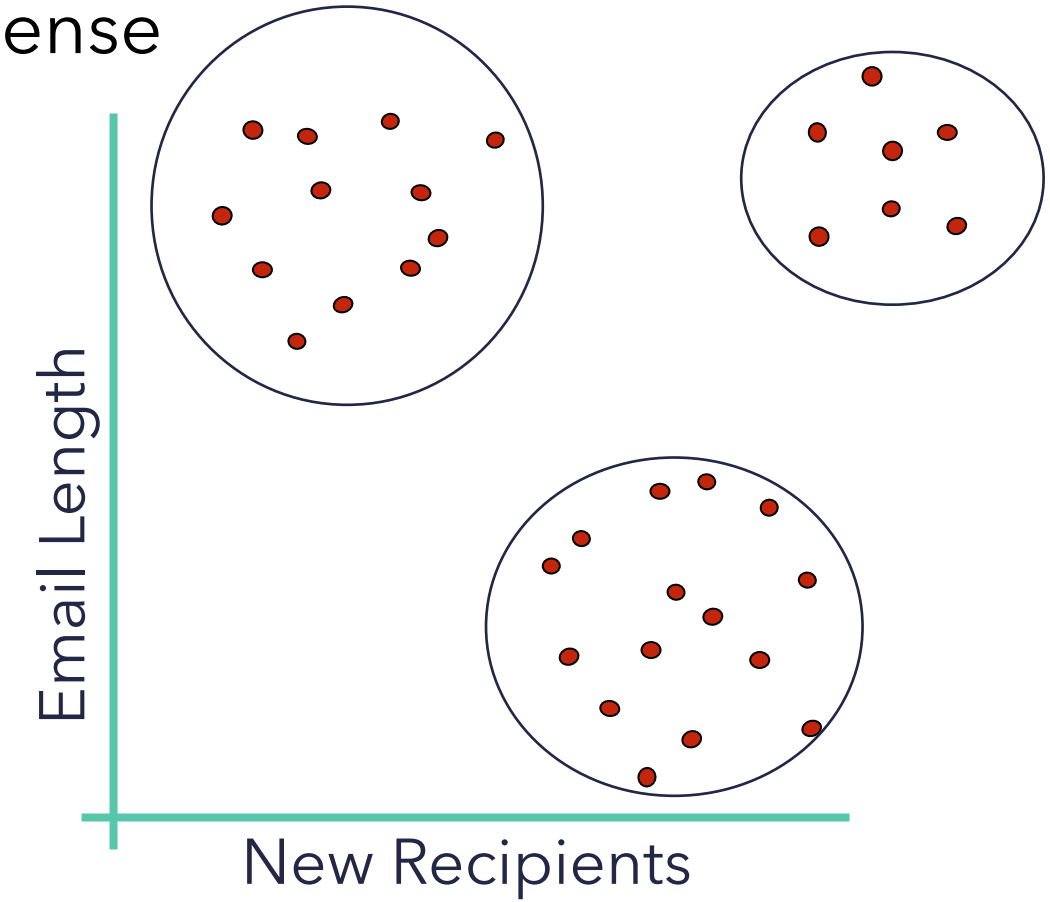


Supervised learning - regression

- Find a relationship between a **numeric** dependent variable and one or more independent variables

Unsupervised learning - clustering

- **Clustering** is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense



Unsupervised learning– anomaly detection

- Detecting patterns in a given data set that do not conform to an established normal behavior

Software – PYTHON PLEASE

Main principles

Occam's razor

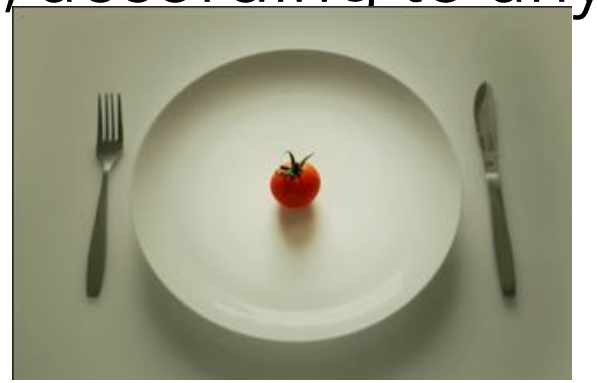
(14th-century)

- In Latin “lex parsimoniae”, translating to “law of parsimony”
- The explanation of any phenomenon should make as **few assumptions** as possible, **eliminating those** that **make no difference in the observable predictions** of the explanatory hypothesis or theory
- The Occam Dilemma: unfortunately, in ML, accuracy and simplicity (interpretability) are in conflict

No free lunch theorem in Machine Learning

(Wolpert, 2001)

- “For any two learning algorithms, there are just as many situations (appropriately weighted) in which algorithm one is superior to algorithm two as vice versa, according to any of the measures of "superiority"

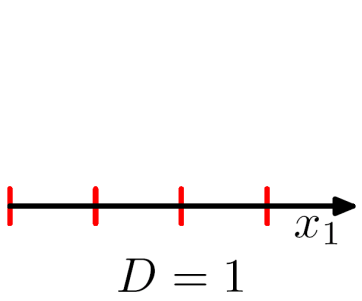


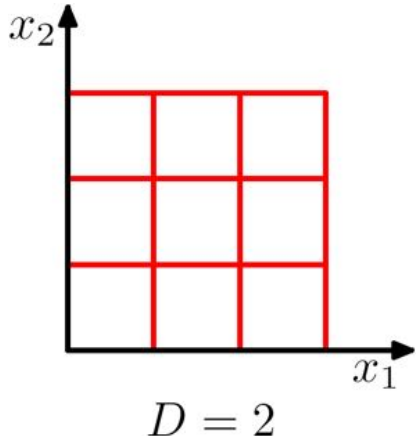
So why developing new algorithms?

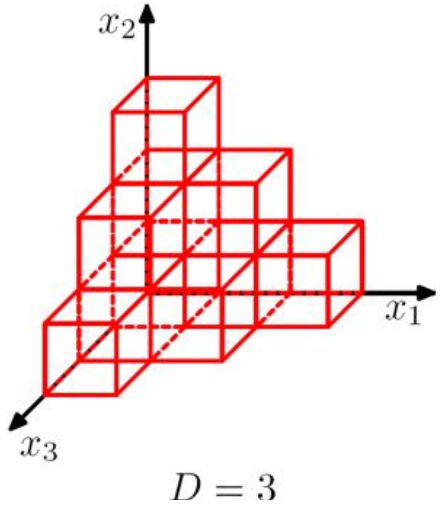
- Practitioner are mostly concerned with choosing the most appropriate algorithm for the **problem at hand**.
- This requires some a priori knowledge – data distribution, prior probabilities, complexity of the problem, the physics of the underlying phenomenon, etc.
- The No Free Lunch theorem tells us that – unless we have some a priori knowledge – simple classifiers (or complex ones for that matter) are not necessarily better than others. However, given some a priori information, certain classifiers may better **MATCH** the characteristics of certain type of problems.
- The main challenge of the practitioner is then, to identify the correct match between the problem and the classifier! …which is yet another reason to arm yourself with a diverse set of learner arsenal !

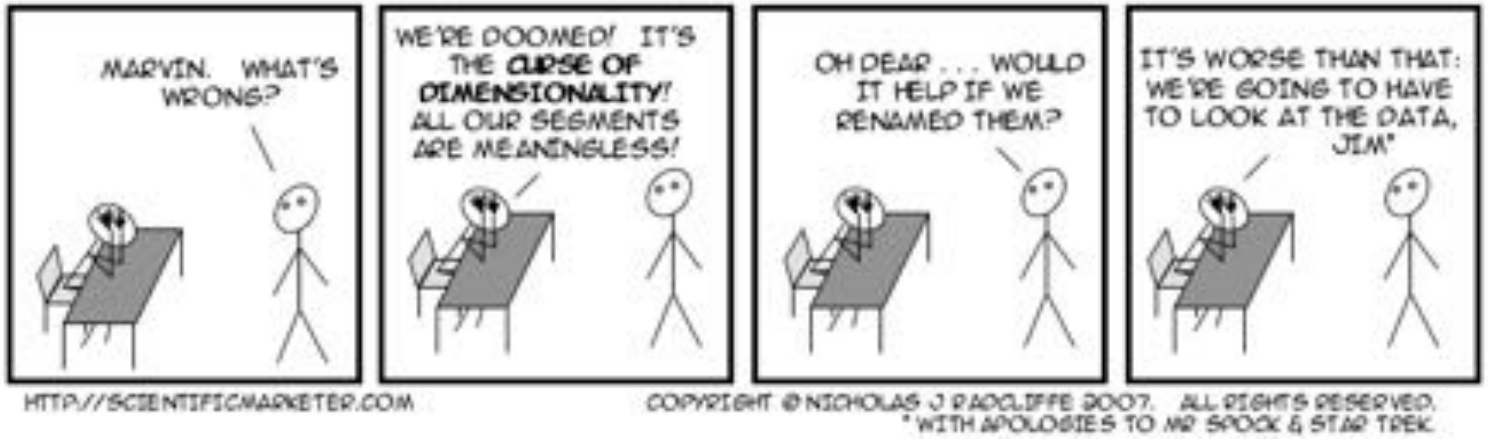
Less is more?

The curse of dimensionality (Bellman, 1961)









Less is more?

The curse of dimensionality (Bellman, 1961)

- Learning from a high-dimensional feature space requires an enormous amount of training to ensure that there are several samples with each combination of values
- With a fixed number of training instances, the predictive power reduces as the dimensionality increases
- As a counter-measure, many dimensionality reduction techniques have been proposed, and it has been shown that when done properly, the properties or structures of the objects can be well preserved even in the lower dimensions
- Nevertheless, naively applying dimensionality reduction can lead to pathological results

Less is more?

The curse of dimensionality (Bellman, 1961)

- While **dimensionality reduction** is an important tool in machine learning/data mining, we must always be aware that it can distort the data in misleading ways
- Above is a two dimensional projection of an intrinsically three dimensional world…

Original photographer unknown
See also www.cs.gmu.edu/~jessica/DimReducDanger.htm

Less is more?

The curse of dimensionality (Bellman, 1961)

- In the past the published advice was that high dimensionality is dangerous
- But, reducing dimensionality reduces the amount of information available for prediction
- Today try going in the opposite direction: instead of reducing dimensionality, increase it by adding many functions of the predictor variables
- The higher the dimensionality of the set of features, the more likely it is that separation occurs

Meaningfulness of answers

- A big data-mining risk is that you will “discover” patterns that are meaningless
- Statisticians call it **Bonferroni’s principle**: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Examples of Bonferroni's Principle

- Track terrorists
- The **Rhine Paradox**: a great example of how not to conduct scientific research

Why tracking terrorists is (almost) impossible!

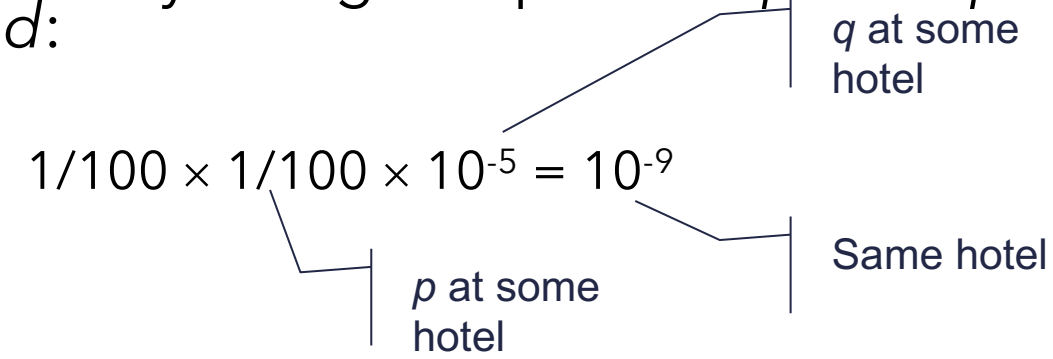
- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil
- We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day

The details

- 109 people being tracked
- 1000 days
- Each person stays in a hotel 1% of the time (10 days out of 1000)
- Hotels hold 100 people (so 105 hotels)
- If everyone behaves randomly (i.e., no evil-doers) will the data mining detect anything suspicious?

Calculations – (1)

Probability that given persons p and q will be at the same hotel on given day d:



- Probability that p and q will be at the same hotel on given days d_1 and d_2:

  $10^{-9} \times 10^{-9} = 10^{-18}$

- Pairs of days:
  - 5×10^5

Calculations – (2)

- Probability that *p* and *q* will be at the same hotel on some two days:
  - $5\times10^{5} \times 10^{-18} = 5\times10^{-13}$
- Pairs of people:
  - 5×10^{17}
- Expected number of "suspicious" pairs of people:
  - $5\times10^{17} \times 5\times10^{-13} = 250{,}000$

Conclusion

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice

- Analysts have to sift through 250,010 candidates to find the 10 real cases
  - Not gonna happen
  - But how can we improve the scheme?

Moral

- When looking for a property (e.g., “two people stayed at the same hotel twice”), make sure that the property does not allow so many possibilities that random data will surely produce facts “of interest.”

Rhine Paradox – (1)

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception (ESP)
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Rhine Paradox – (2)

- He told these people they had ESP and called them in for another test of the same type
- Alas, he discovered that almost all of them had lost their ESP
- What did he conclude?
  - Answer on next slide

Rhine Paradox – (3)

- He concluded that you shouldn’t tell people they have ESP; it causes them to lose it

Moral

- Understanding Bonferroni's Principle will help you look a little less stupid than a parapsychologist

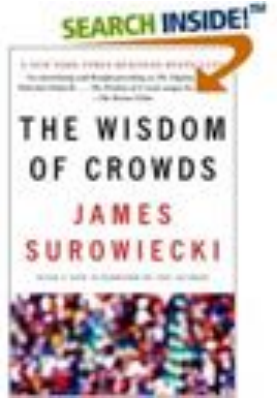
Instability and the Rashomon Effect

- Rashomon is a Japanese movie in which four people, from different vantage points, witness a criminal incident When they come to testify in court, they all report the same facts, but their stories of what happened are very different
- The Rashomon effect is the effect of the subjectivity of perception on recollection
- The Rashomon Effect in ML is that there is often a multitude of classifiers of giving about the same minimum error rate
- For example in decision trees, if the training set is perturbed only slightly, we can get a tree quite different from the original but with almost the same test set error

The Wisdom of Crowds

Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations

- Under certain controlled conditions, the aggregation of information in groups, resulting in decisions that are often superior to those that can been made by any single - even experts
- Imitates our second nature to seek several opini before making any crucial decision
- We weigh the individual opinions, and combine them to reach a final decision



Committees of experts

> " … a medical school that has the objective that all students, given a problem, come up with an identical solution"

- There is not much point in setting up a committee of experts from such a group - such a committee will not imp ent of an individual

- Consider:
  - There needs to be disagreement for the committee to have the potential to be better than an individual