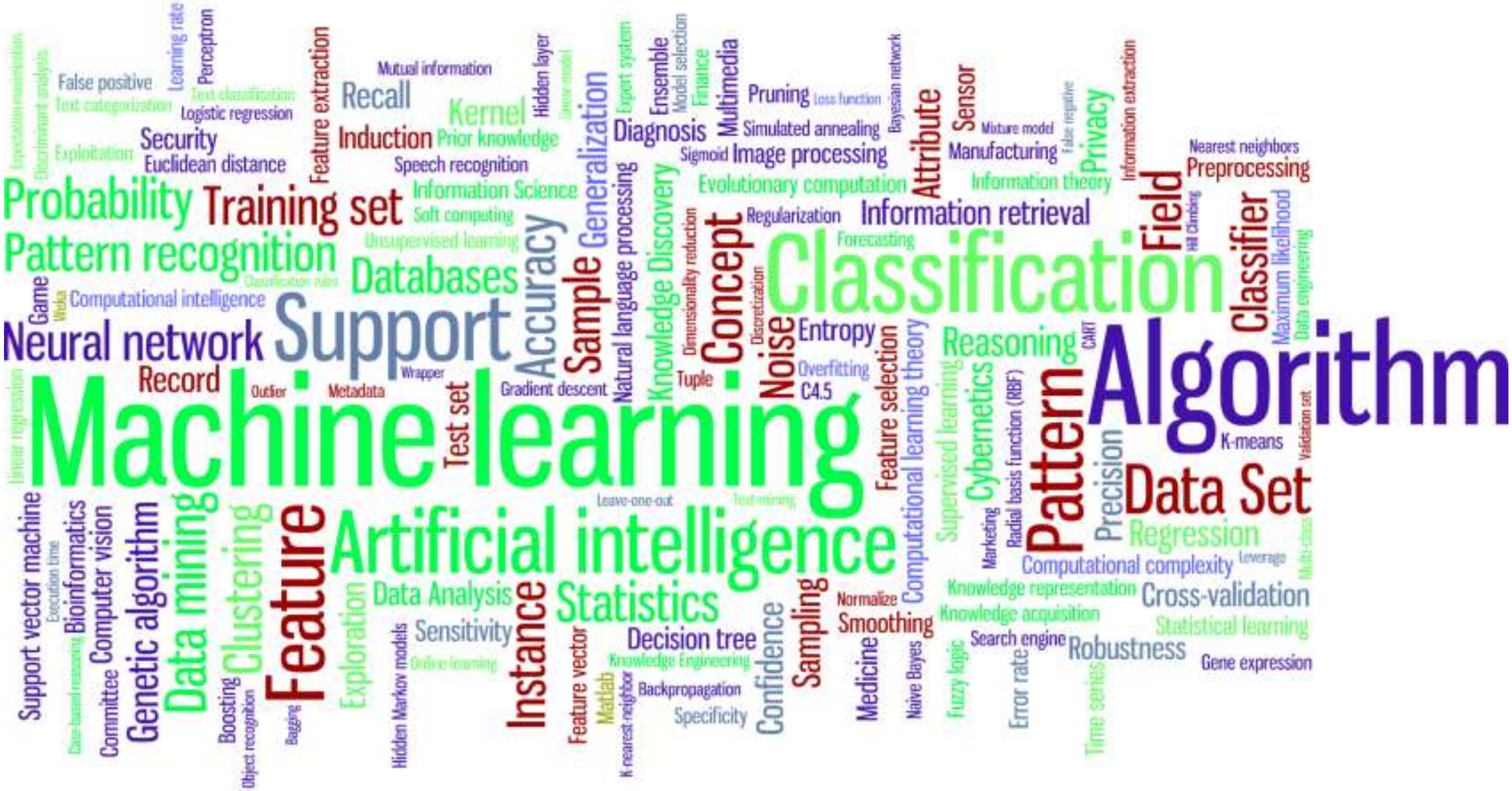
# Introduction to Machine Learning



Lior Rokach
Department of Information Systems Engineering
Ben-Gurion University of the Negev

# About Me

Prof. Lior Rokach

Department of Information Systems Engineering

Faculty of Engineering Sciences

Head of the Machine Learning Lab

Ben-Gurion University of the Negev

Email: liorrk@bgu.ac.il

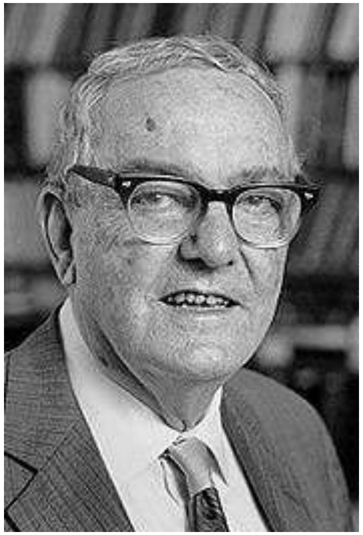
http://www.ise.bgu.ac.il/faculty/liorr/

PhD (2004) from Tel Aviv University

PATTERN CLASSIFICATION USING ENSEMBLE METHODS
DATA MINING WITH DECISION TREES
Theory and Applications
Lior Rokach • Oded Maimon
Soft Computing for Knowledge Discovery and Data Mining
DECOMPOSITION METHODOLOGY FOR KNOWLEDGE DISCOVERY AND DATA MINING
RECOMMENDER SYSTEMS HANDBOOK
The Data Mining and Knowledge Discovery Handbook
A Complete Guide for Research Scientists and Practitioners
DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK
A Survey of Data Leakage Detection and Prevention Solutions

# 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

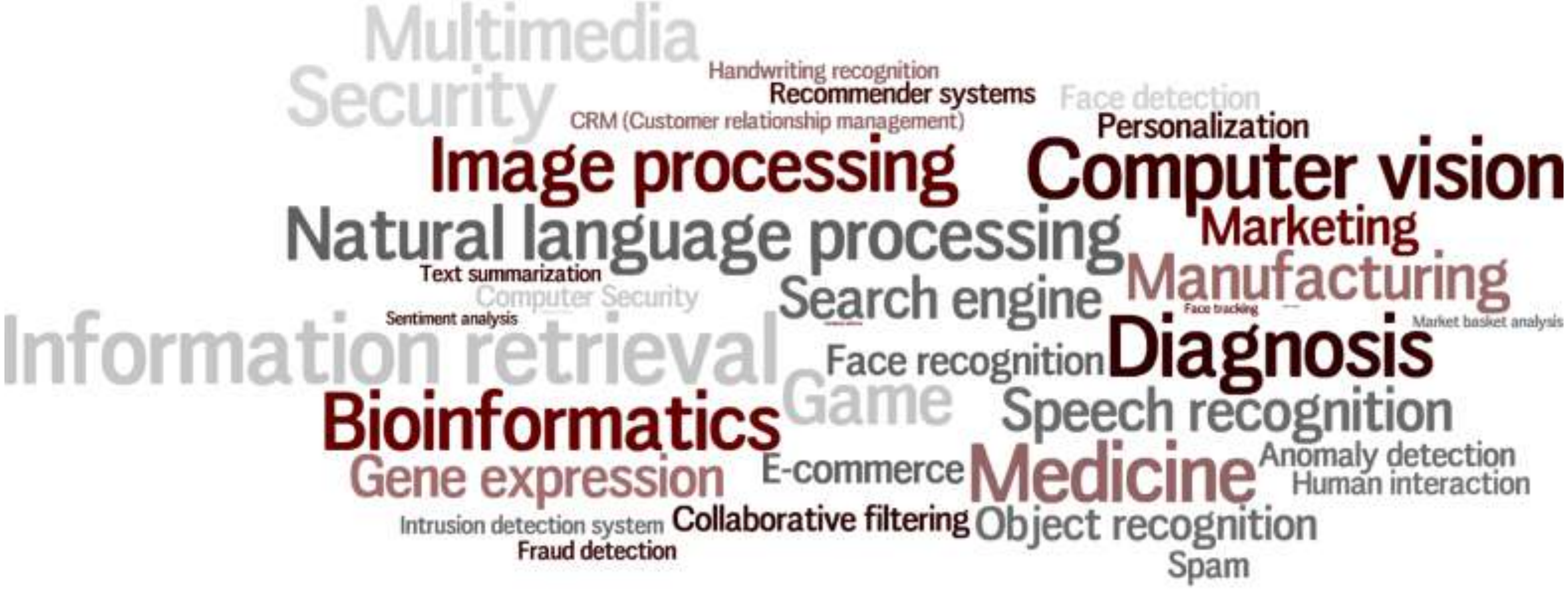
# Why Machine Learning?

- Develop systems that can automatically adapt and customize themselves to individual users.
  - Personalized news or mail filter
- Discover new knowledge from large databases (***data mining***).
  - Market basket analysis (e.g. diapers and beer)
- Ability to mimic human and replace certain monotonous tasks - which require some intelligence.
    - like recognizing handwritten characters
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (knowledge engineering bottleneck).

# Why now?

- Flood of available data (especially with the advent of the Internet)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries

# ML Applications



# 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*.

# Motivating Example
# Learning to Filter Spam

**Example**: Spam Filtering

Spam - is all email the user does not want to receive and has not asked to receive

*T*: Identify Spam Emails

*P*:

% of spam emails that were filtered

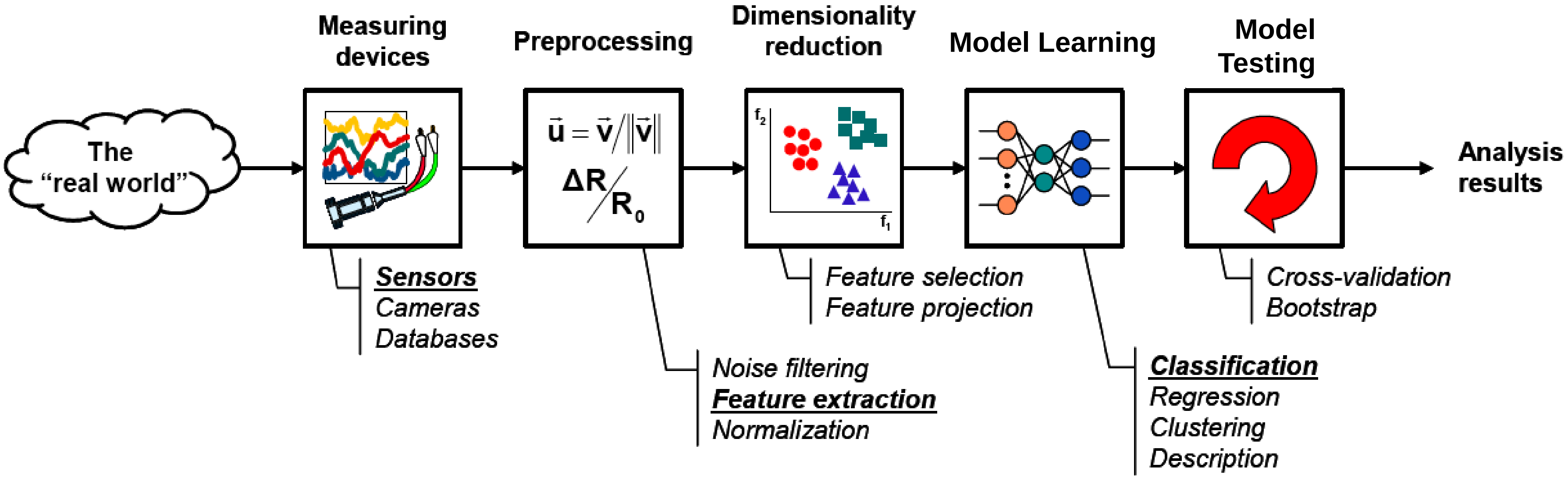
% of ham/ (non-spam) emails that were incorrectly filtered-out

*E*: a database of emails that were labelled by users

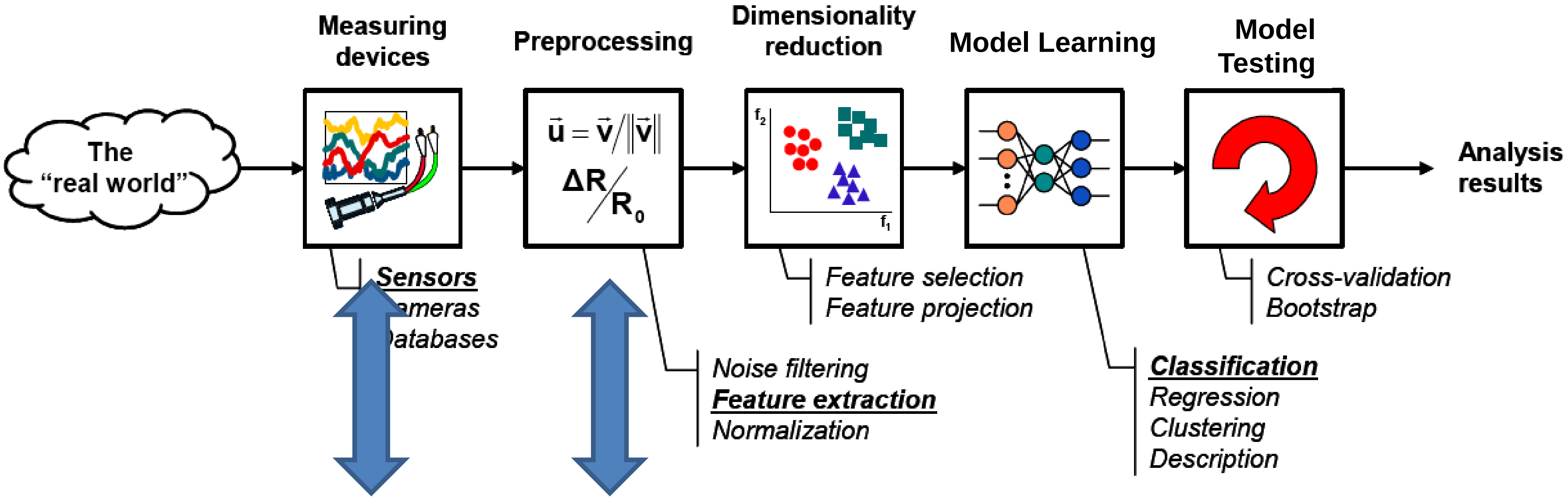
Preprocessing
Classifier
Training set
Data set
Validation set
Data cleaning
Feature vector
Wavelet transform
Discrete wavelet transform (DWT)
Information extraction
Information fusion
Induction
Test set
Concept
Feature
Feature selection
Relational data
Categorical attribute
Metadata
Class attribute
Outlier
Binning
Discretization
Field
Scaling factor
Nominal attribute
Numeric attribute
Normalize
Inducer
Data transformation
Validation data
Data preparation
Data reduction
Record
Multidimensional scaling
Unlabeled data
Attribute selection
Missing value
Sensor
Dimensionality reduction
Standardizing
Outlier detection
Data cleansing
Pattern
Tuple
Noise
Smoothing
Sample
Induction algorithm
Discrete fourier transform (DFT)
Continuous attribute
Instance
Sampling
Feature extraction
Attribute

ess

# The Learning Process



# The Learning Process in our Example



Email Server

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

…

# Data Set

Input Attributes

Target Attribute

| 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 |

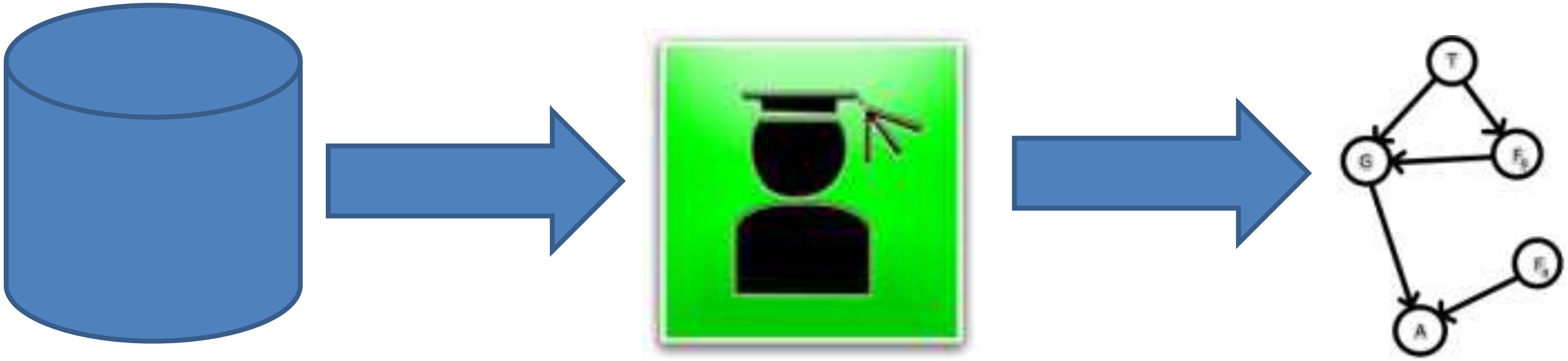
Instances

Numeric

Nominal

Ordinal

# Step 4: Model Learning

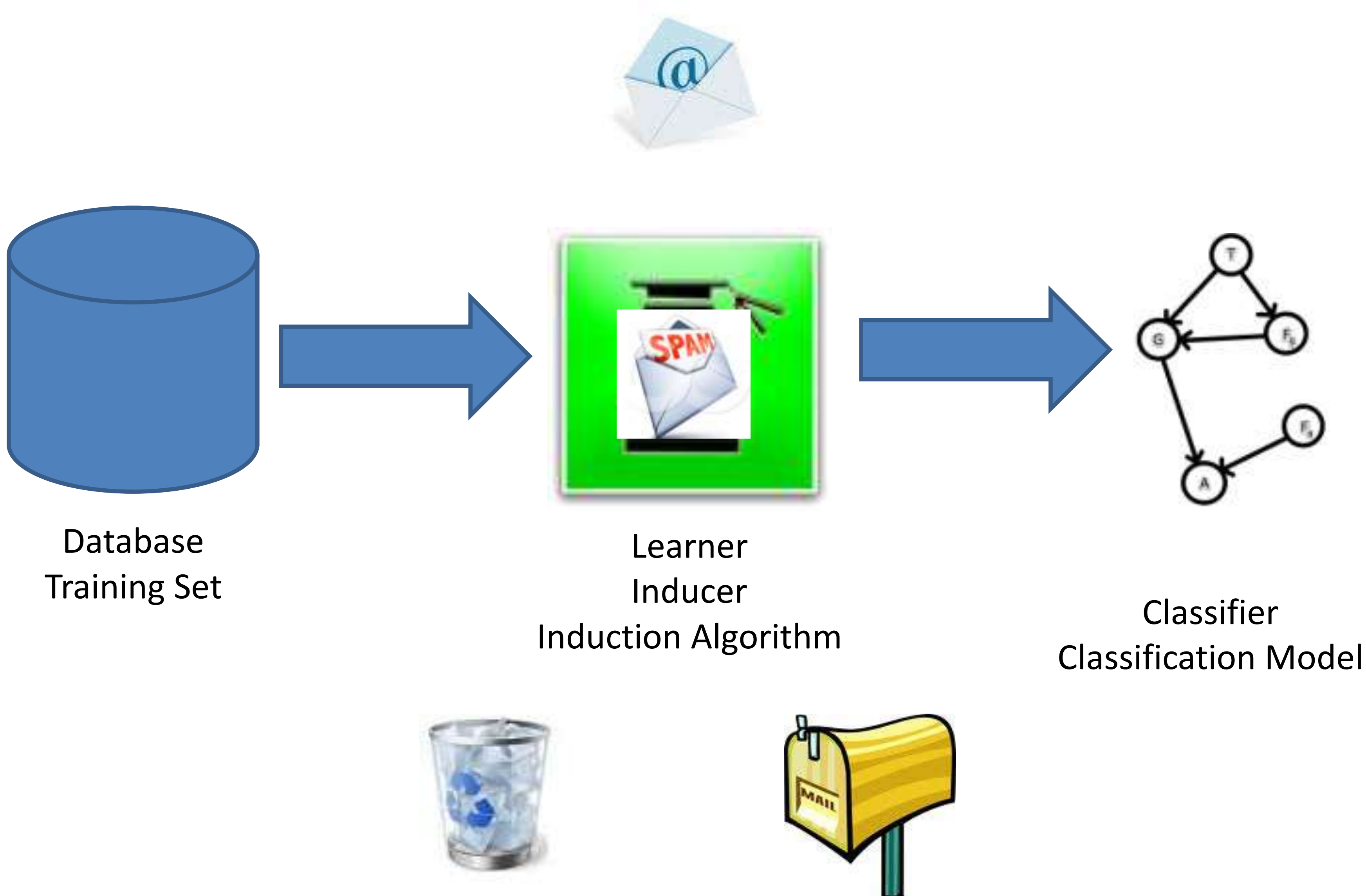


Database
Training Set

Learner
Inducer
Induction Algorithm

Classifier
Classification Model

# Step 5: Model Testing

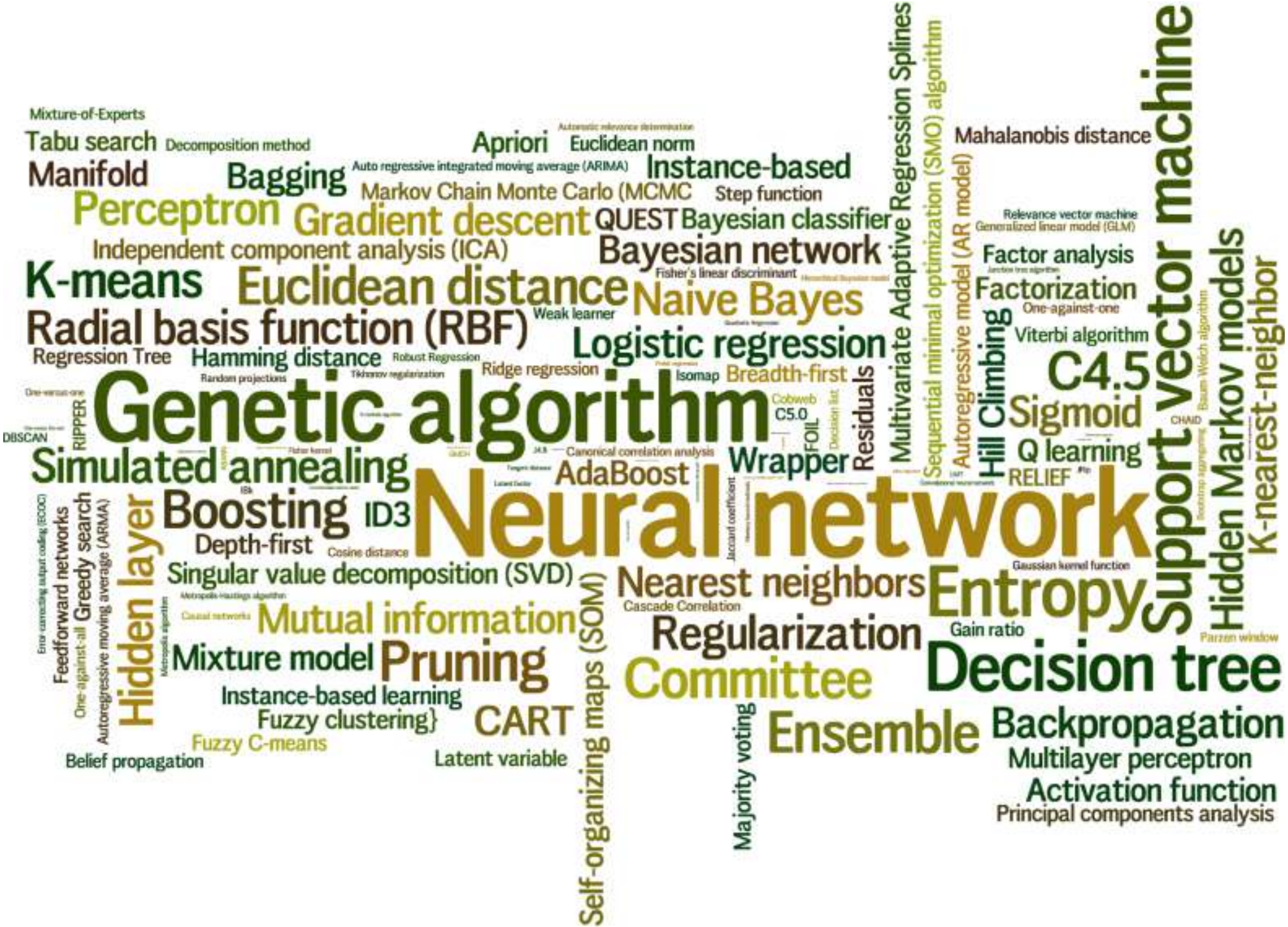


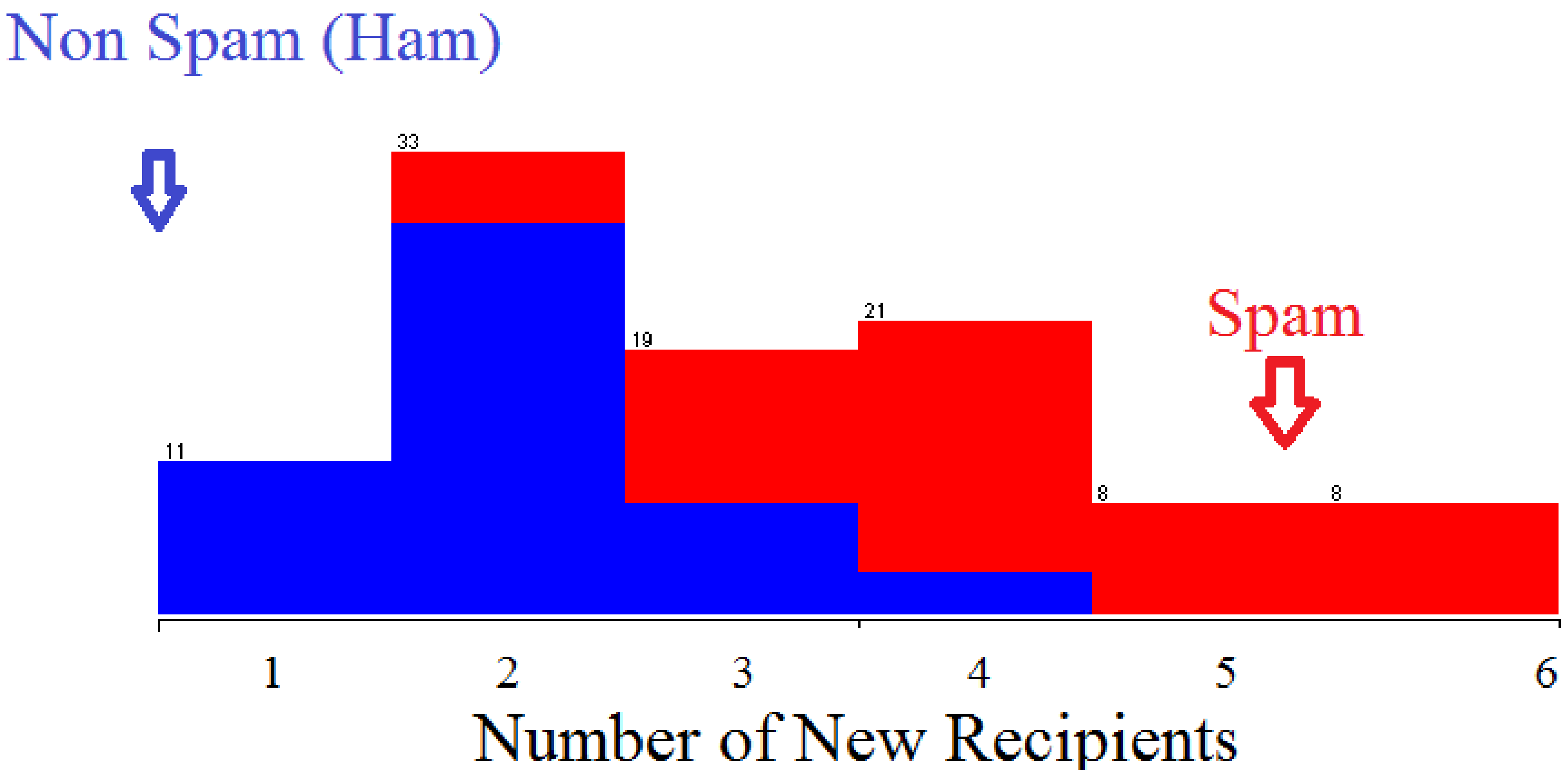
Database
Training Set

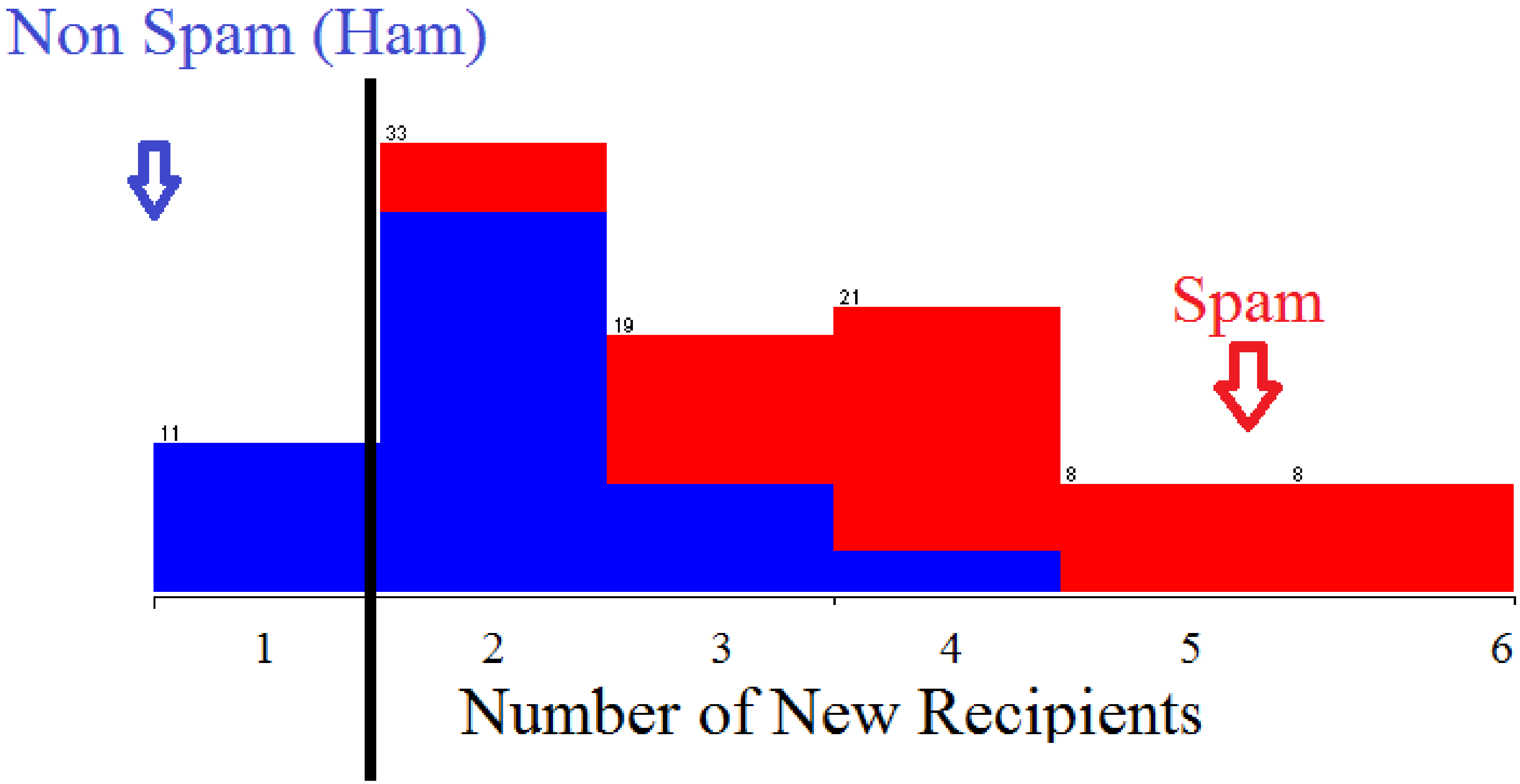
Learner
Inducer
Induction Algorithm

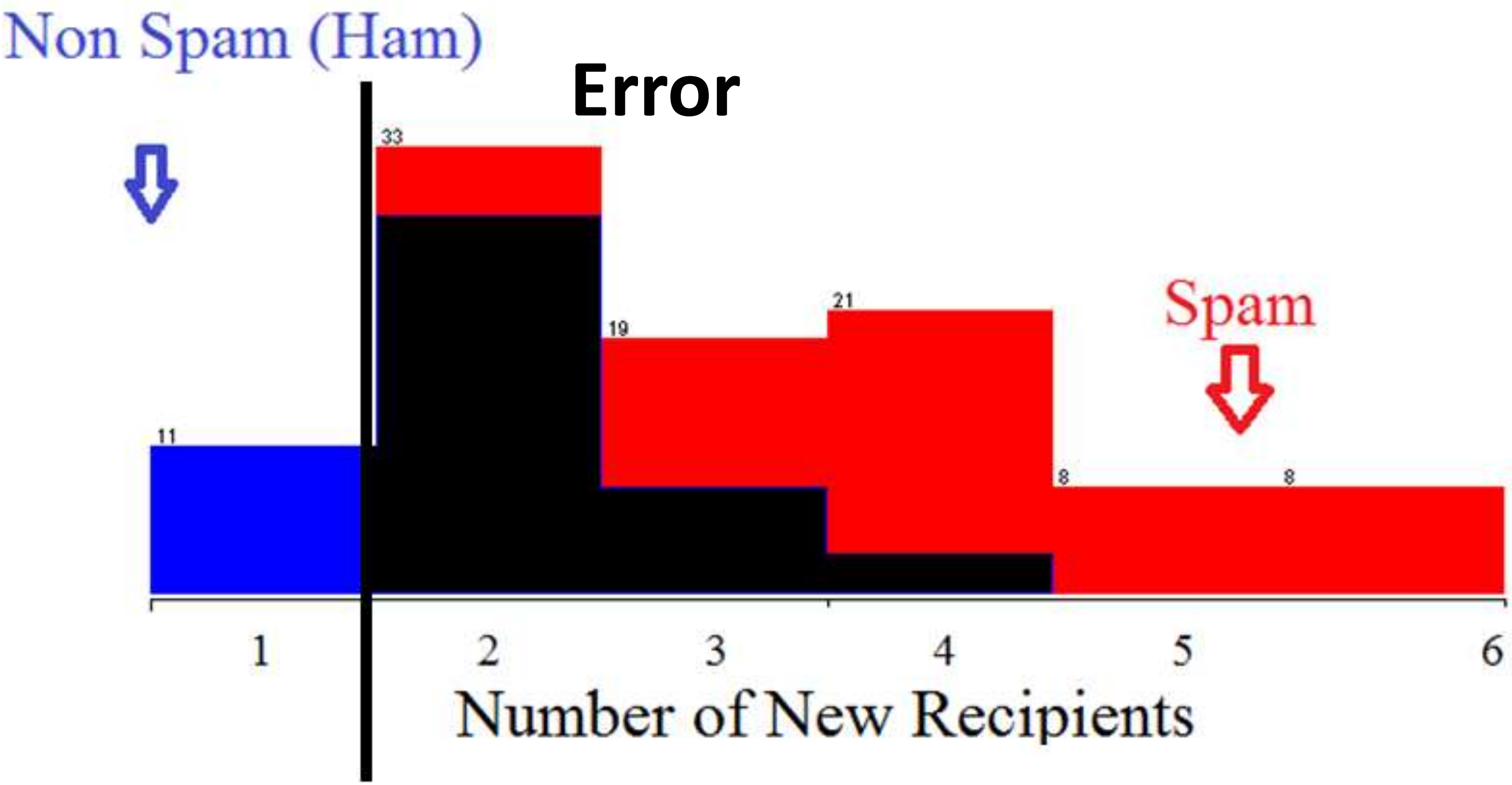
Classifier
Classification Model

# Learning Algorithms



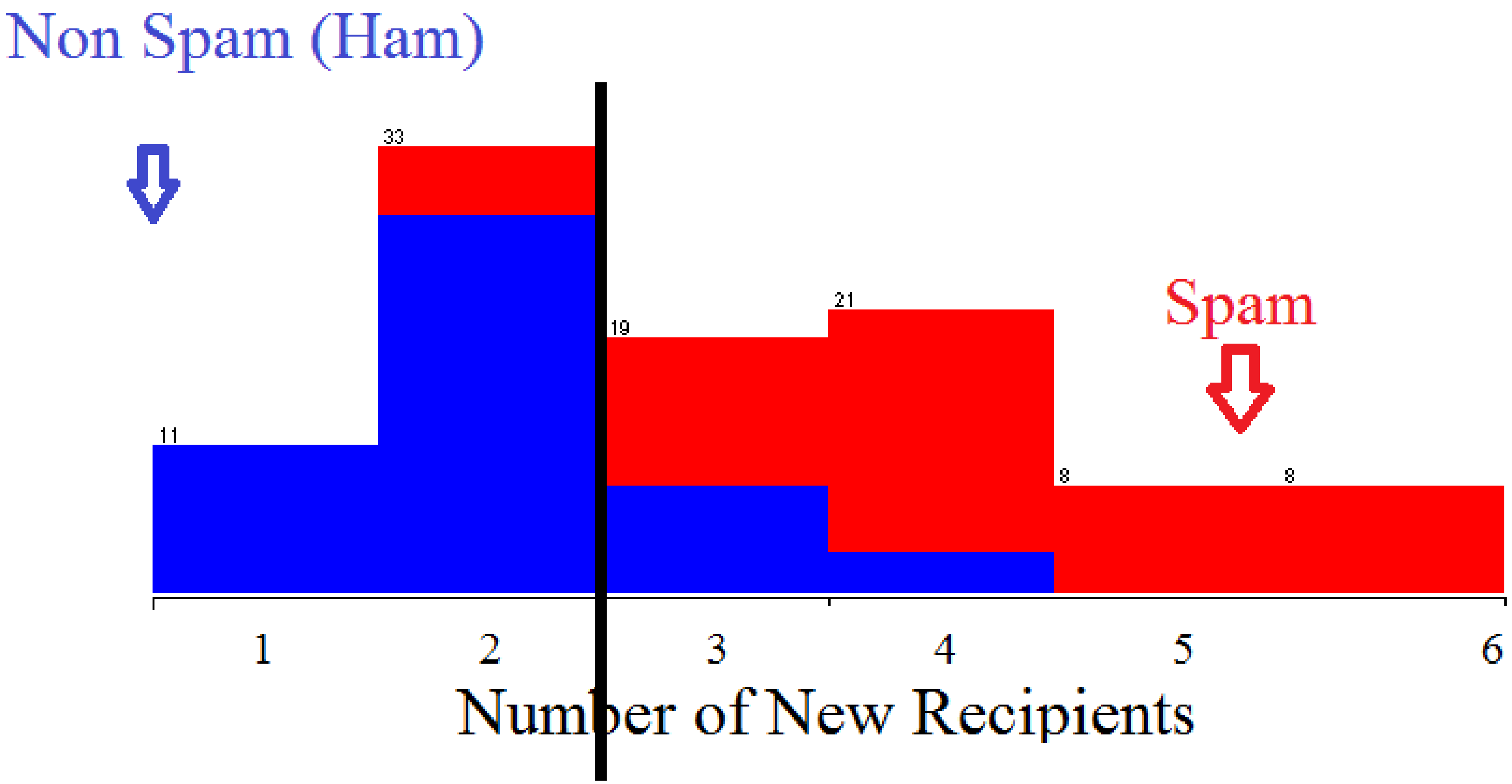

Non Spam (Ham)
Spam
11
33
19
21
8
8
1
2
3
4
5
6
Number of New Recipients

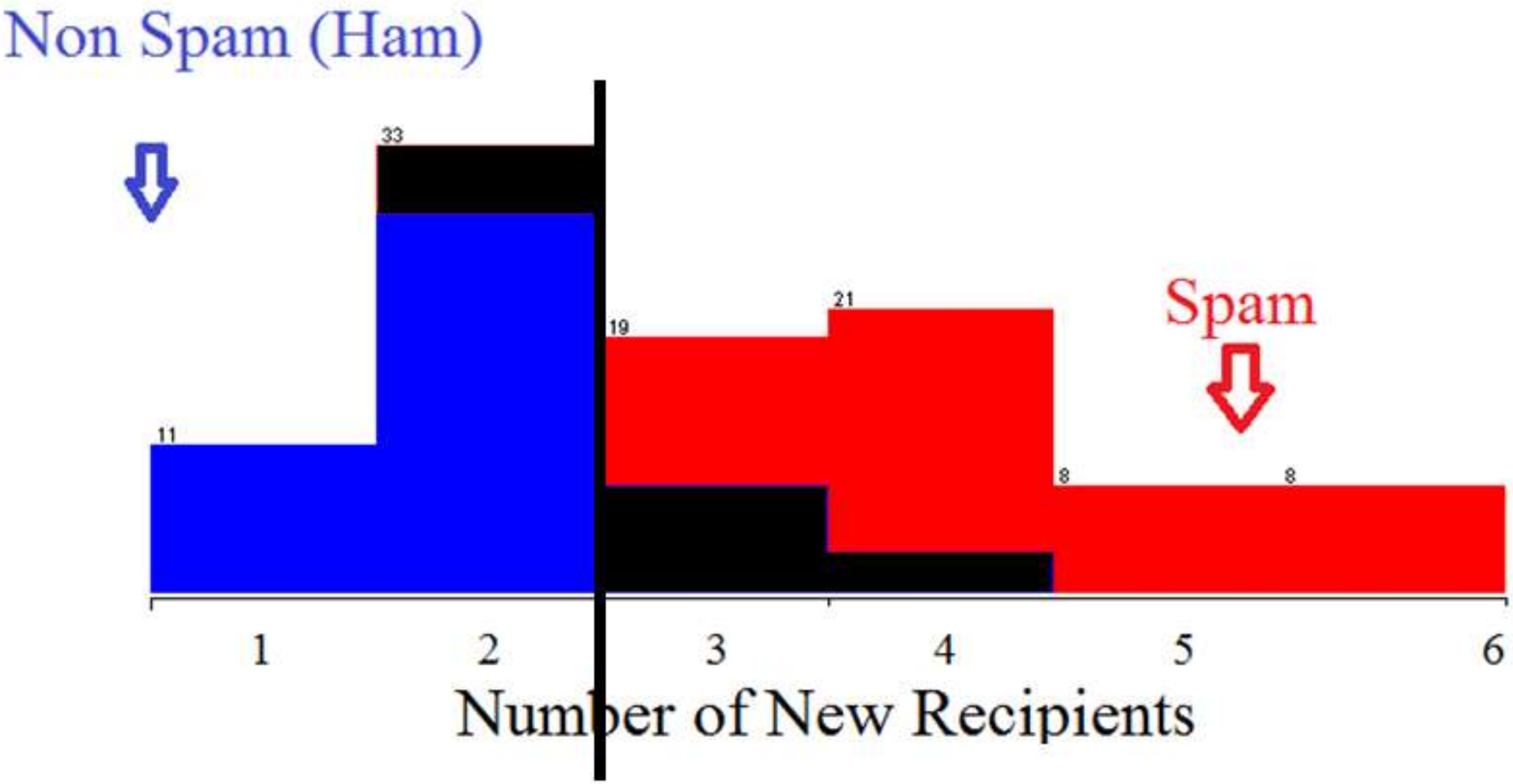

Non Spam (Ham)
Spam
11
33
19
21
8
8
1
2
3
4
5
6
Number of New Recipients

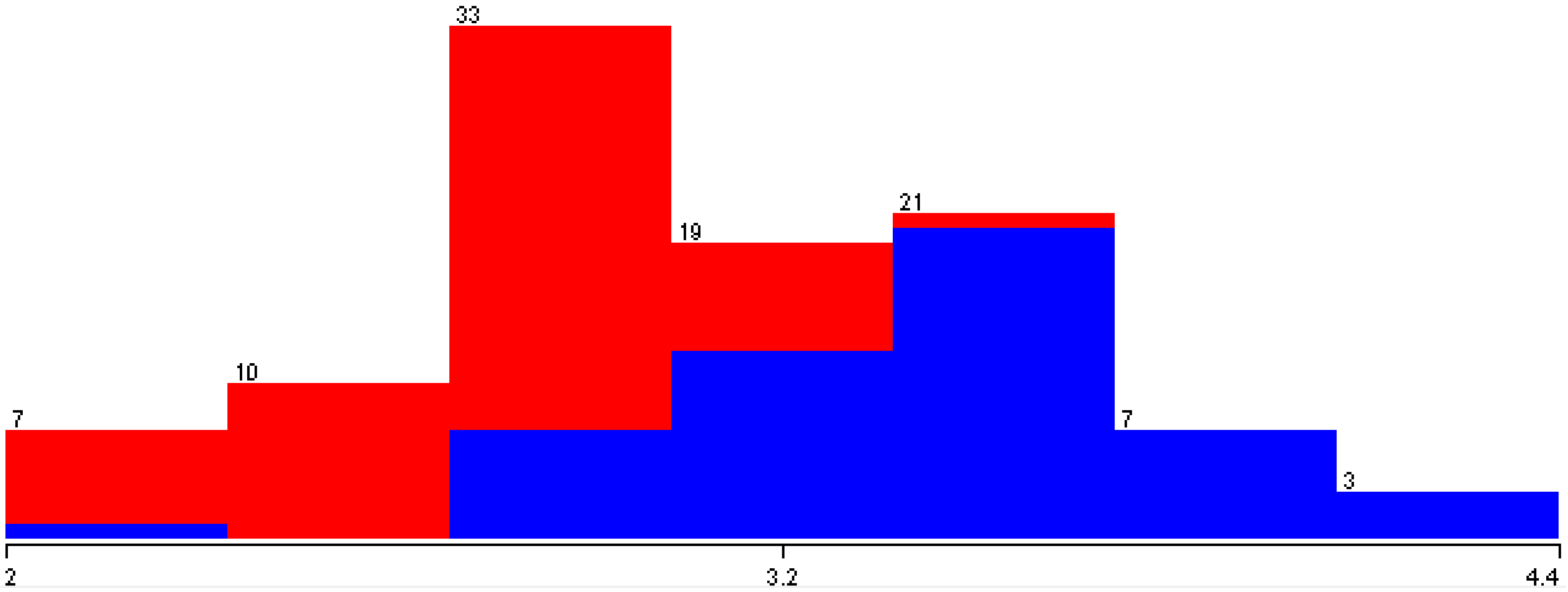


Error

Non Spam (Ham)
Spam
33
19
21
11
8
8
1
2
3
4
5
6
Number of New Recipients

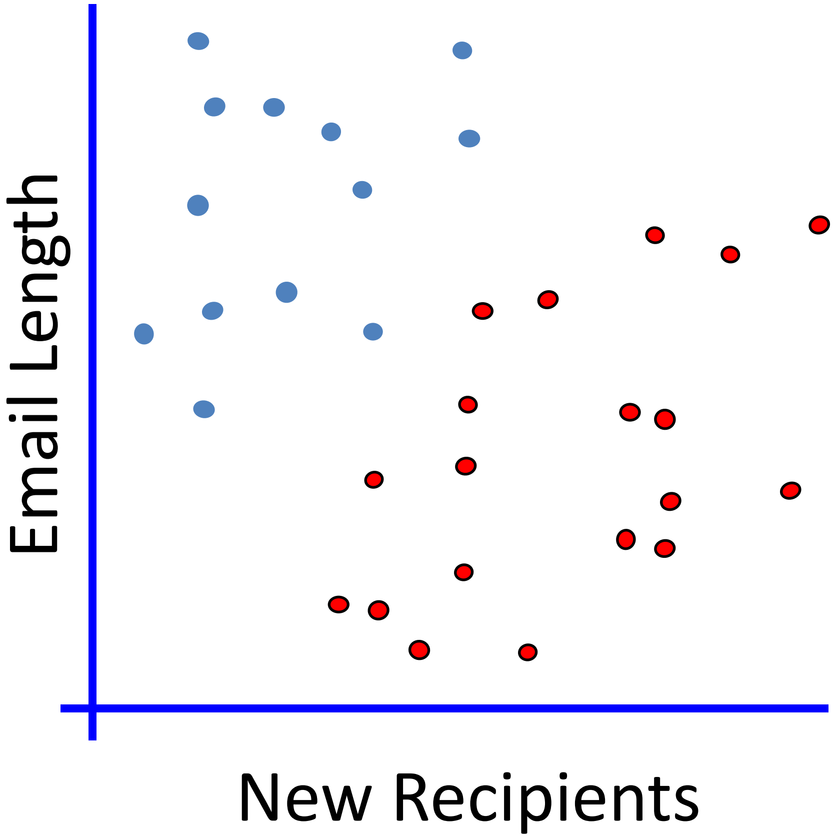

Non Spam (Ham)
33
11
19
21
Spam
8
8
1
2
3
4
5
6
Number of New Recipients


Non Spam (Ham)
Spam
11
33
19
21
8
8
1
2
3
4
5
6
Number of New Recipients



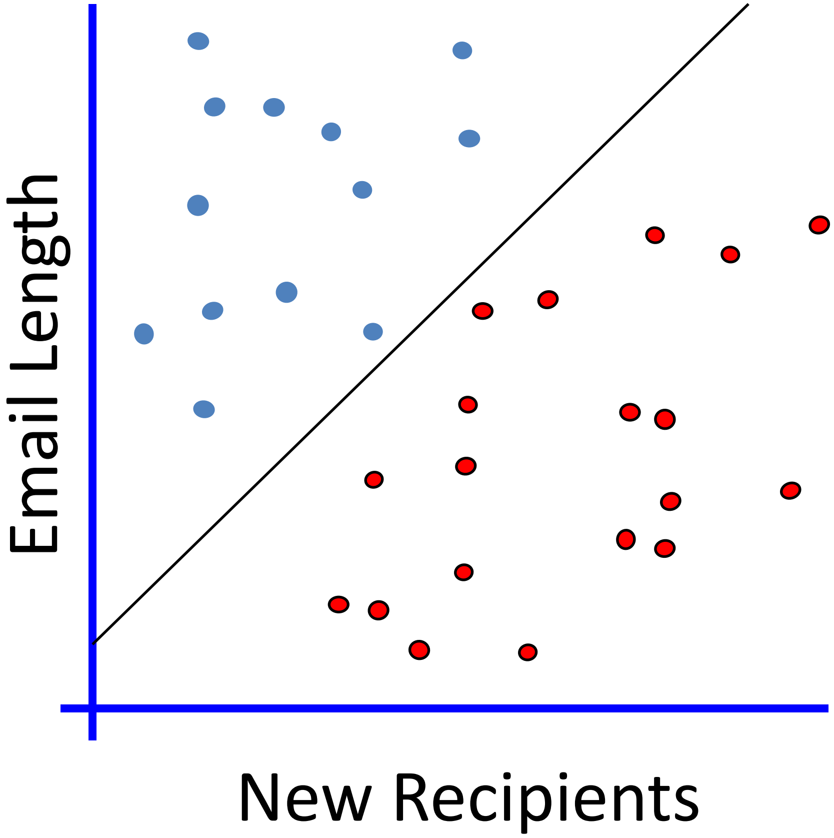
Email Length

# Linear Classifiers



How would you classify this data?

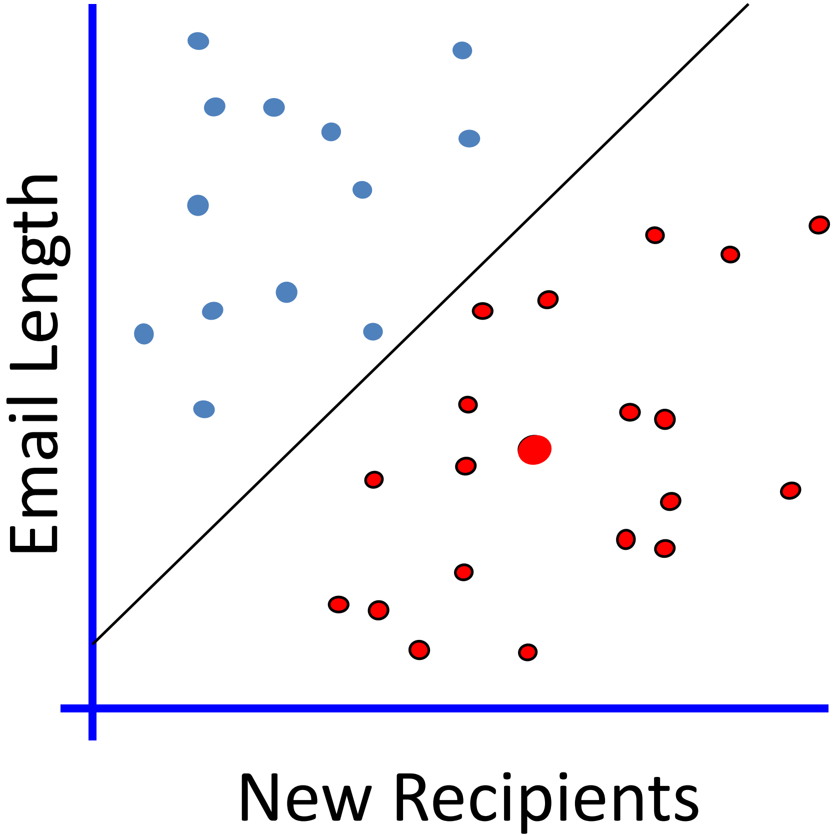
# Linear Classifiers



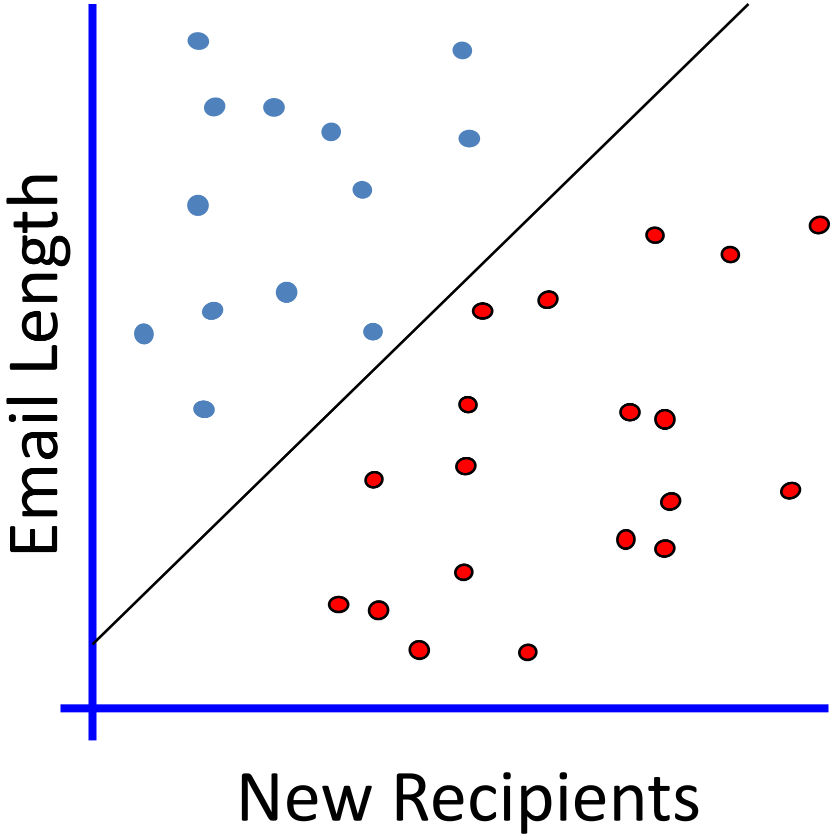
How would you classify this data?

# 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

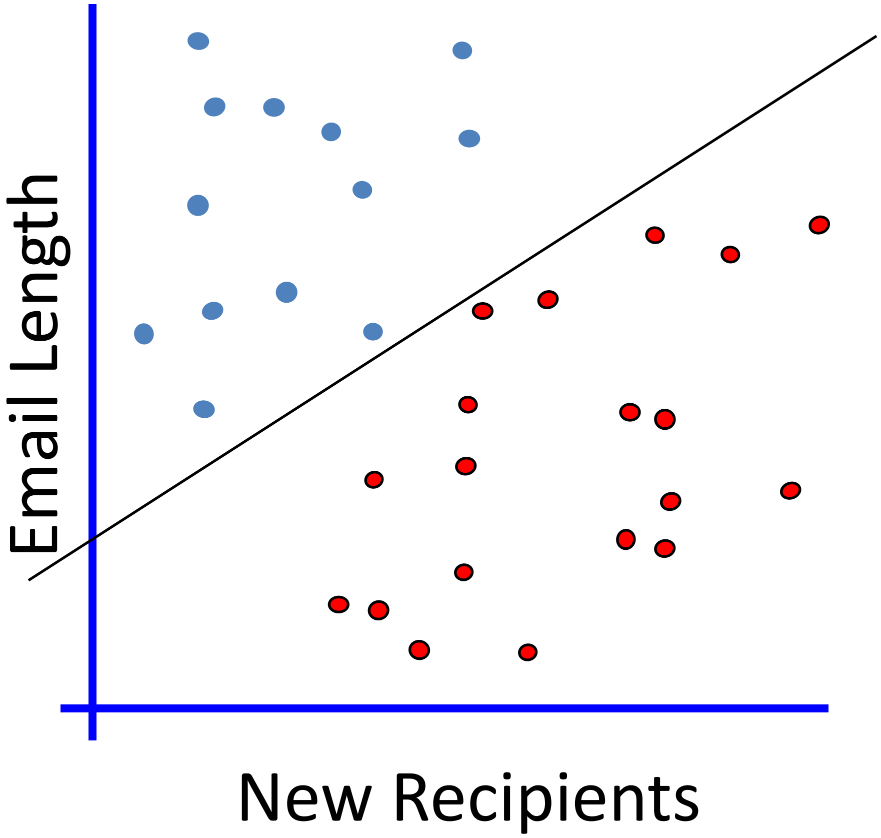


# Linear Classifiers



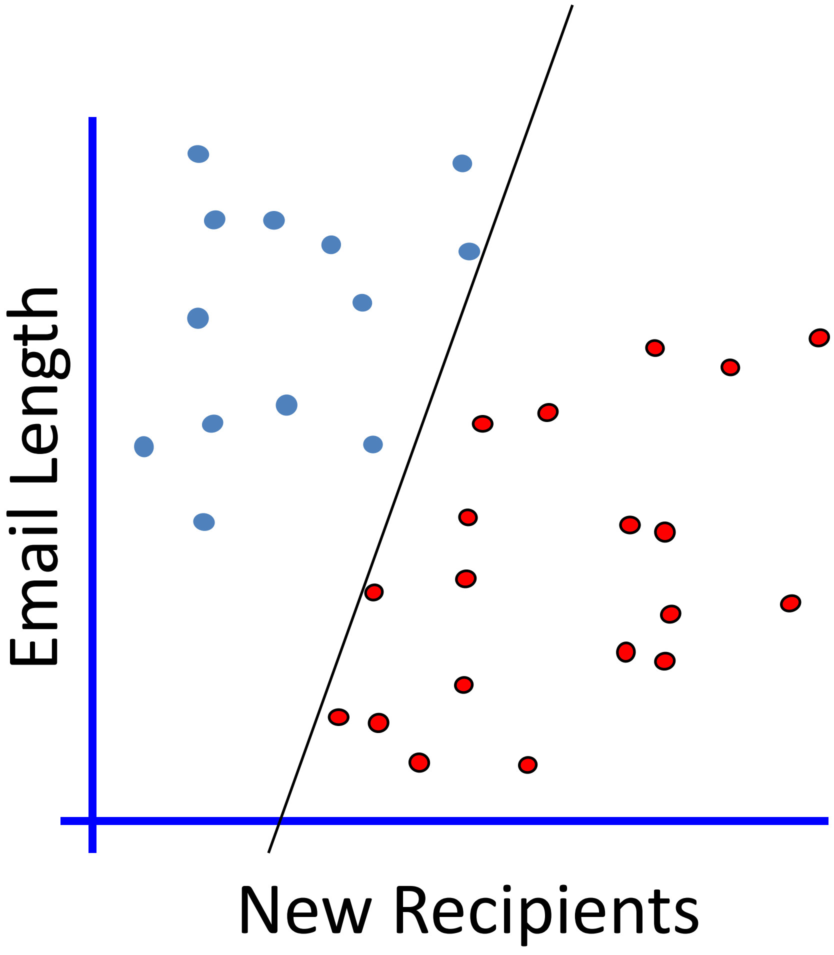
How would you
classify this data?

# Linear Classifiers



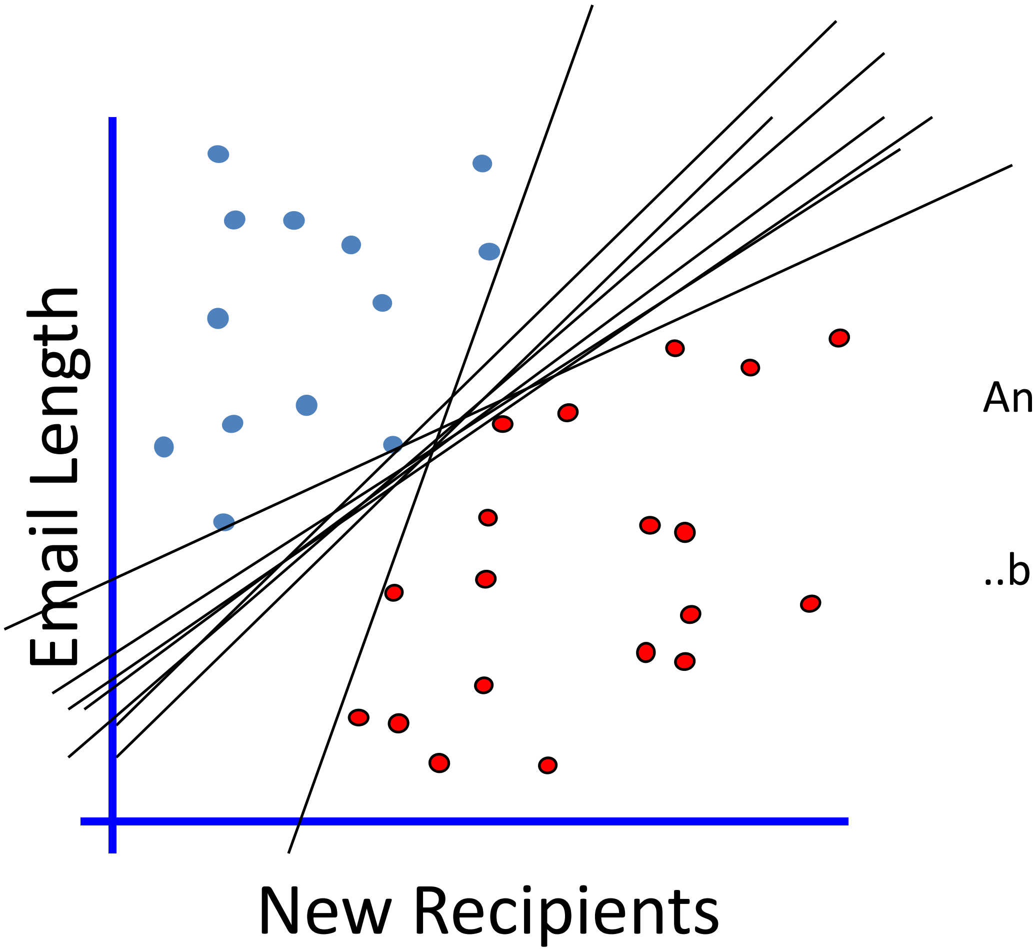
How would you
classify this data?

# Linear Classifiers



How would you classify this data?

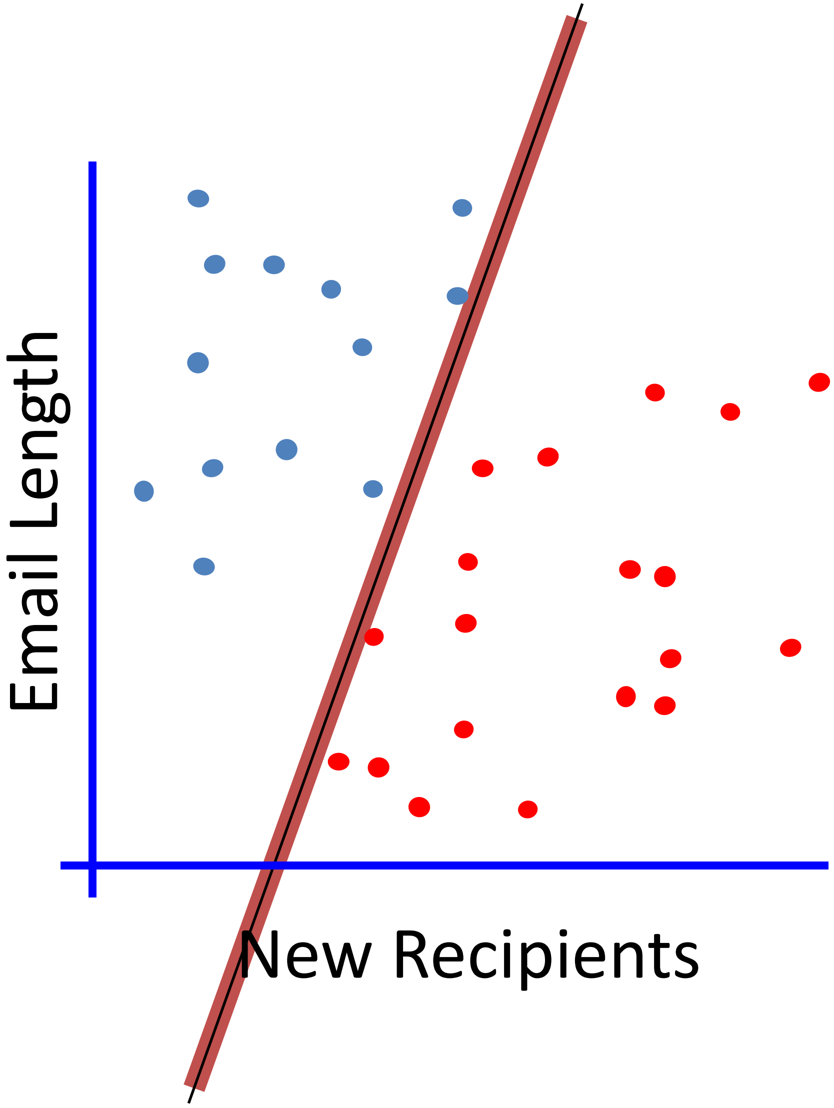
# Linear Classifiers



Any of these would be fine..

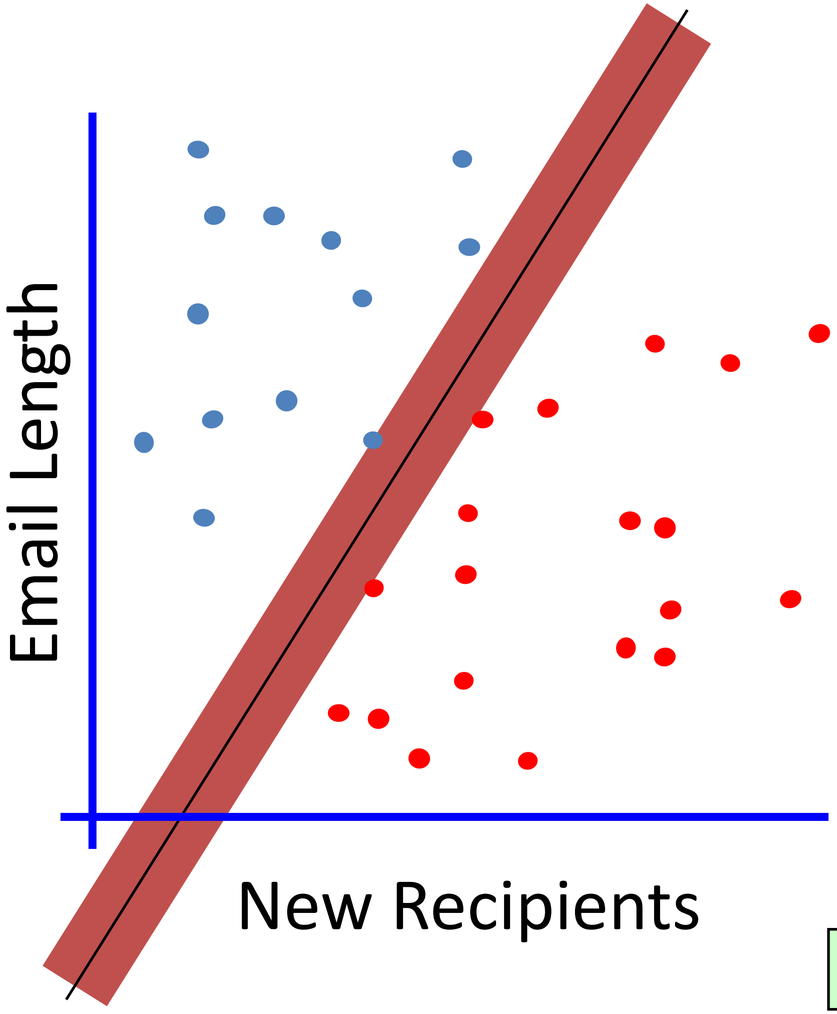
..but which is best?

# Classifier Margin



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

# Maximum Margin

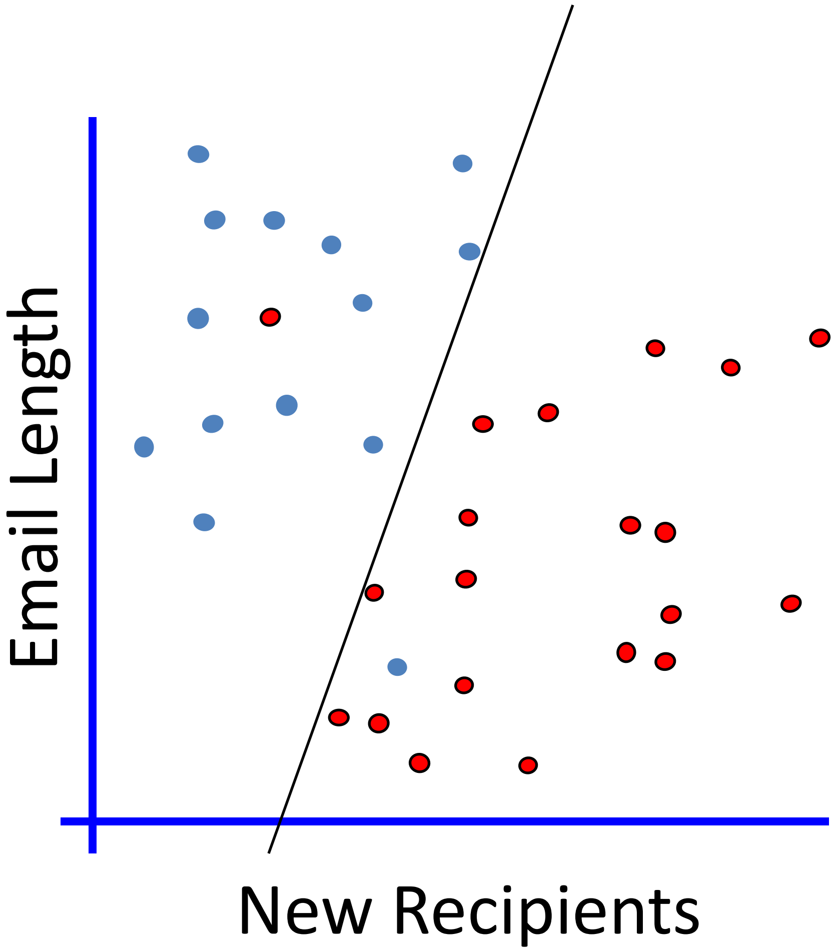


The maximum margin linear classifier is the linear classifier with the, maximum margin.

This is the simplest kind of SVM (Called an LSVM)

Linear SVM

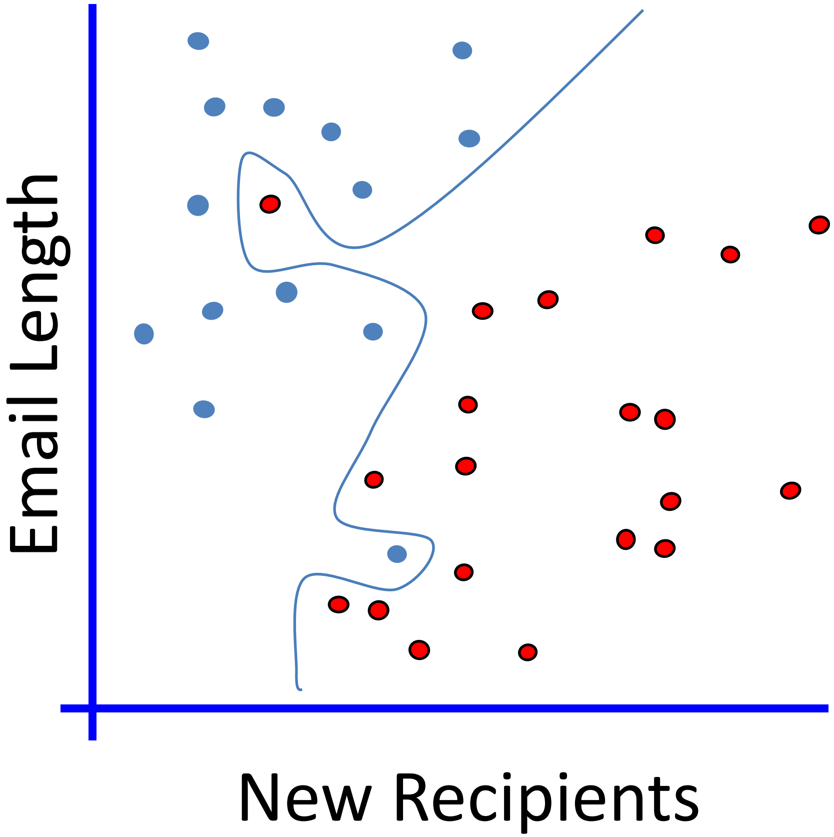
# No Linear Classifier can cover all instances



How would you classify this data?

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

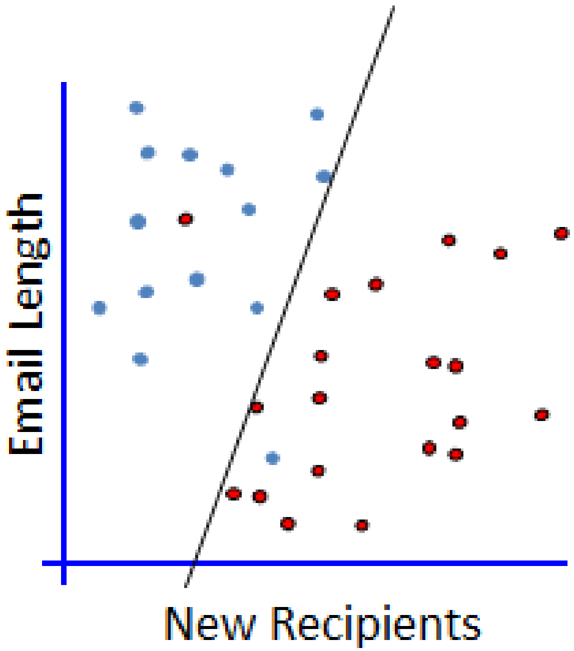
# No Linear Classifier can cover all instances



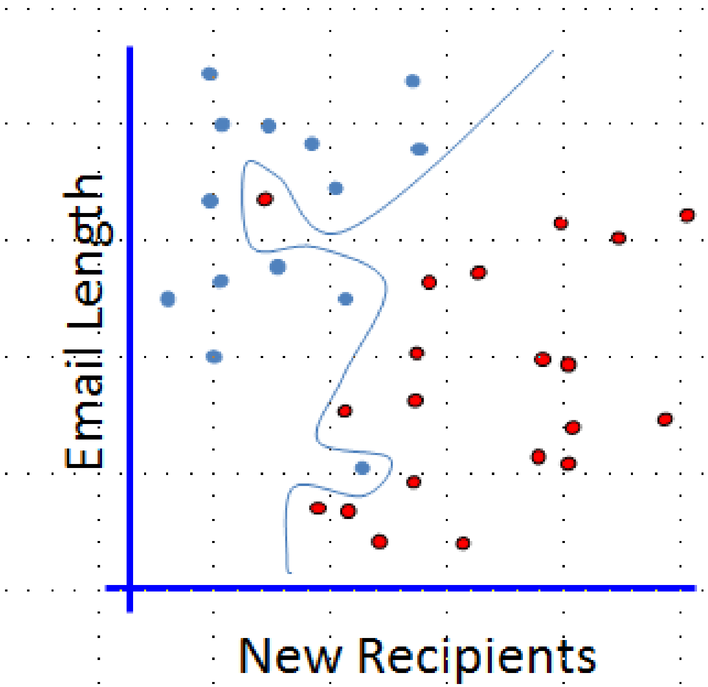
- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

Issue of generalization!

# Which one?



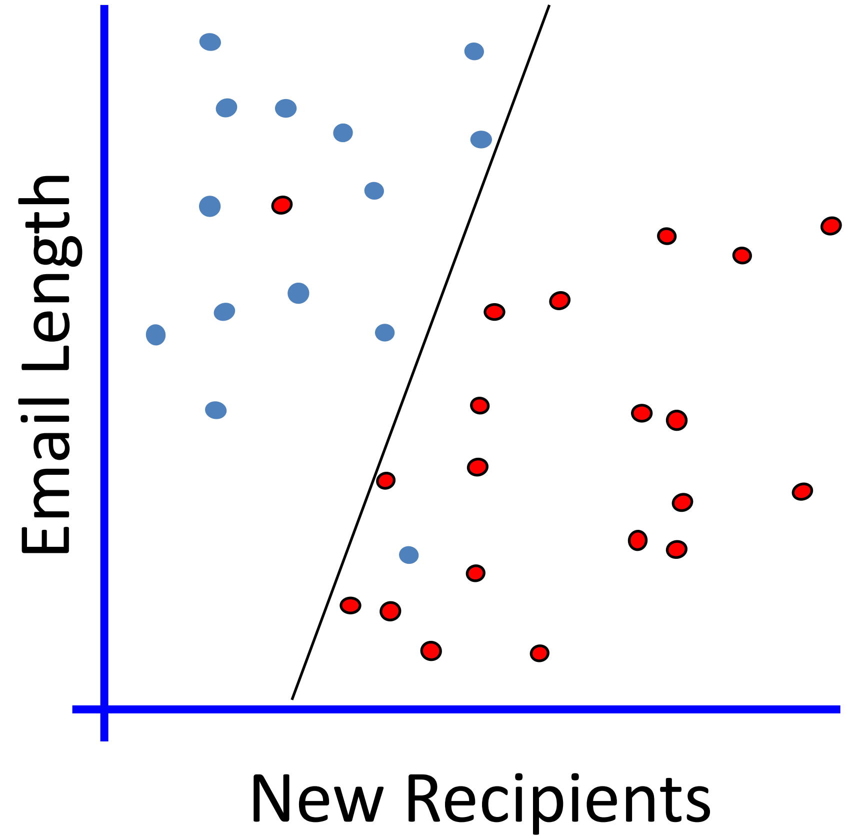
2 Errors
Simple model



0 Errors
Complicated model

# Evaluating What’s Been Learned

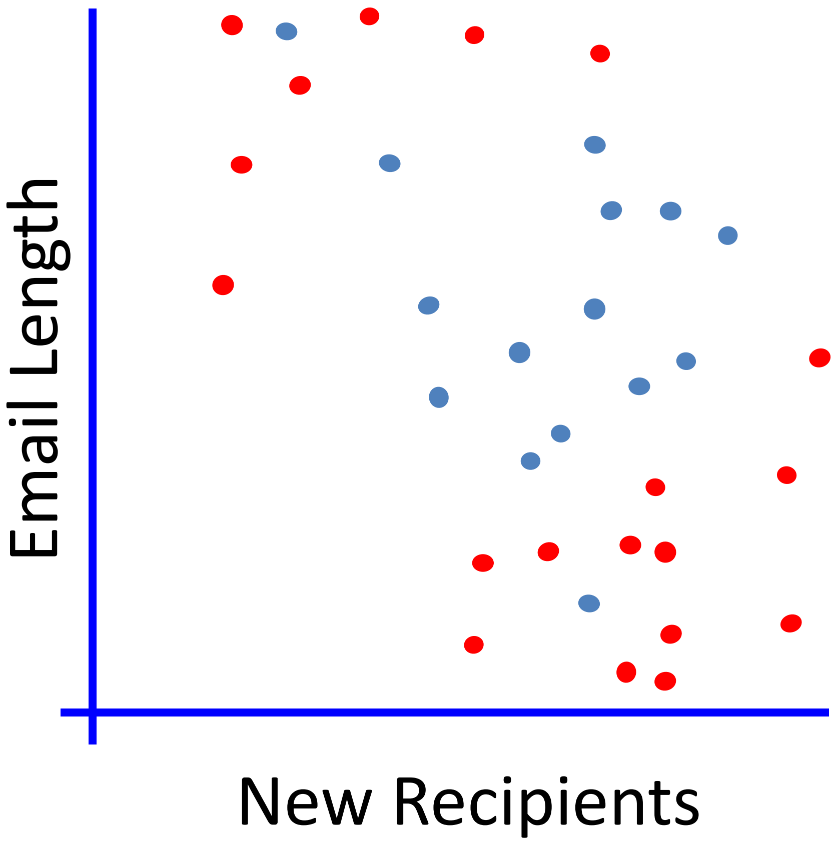
1. We randomly select a portion of the data to be used for training (the training set)
2. Train the model on the training set.
3. Once the model is trained, we run the model on the remaining instances (the test set) to see how it performs



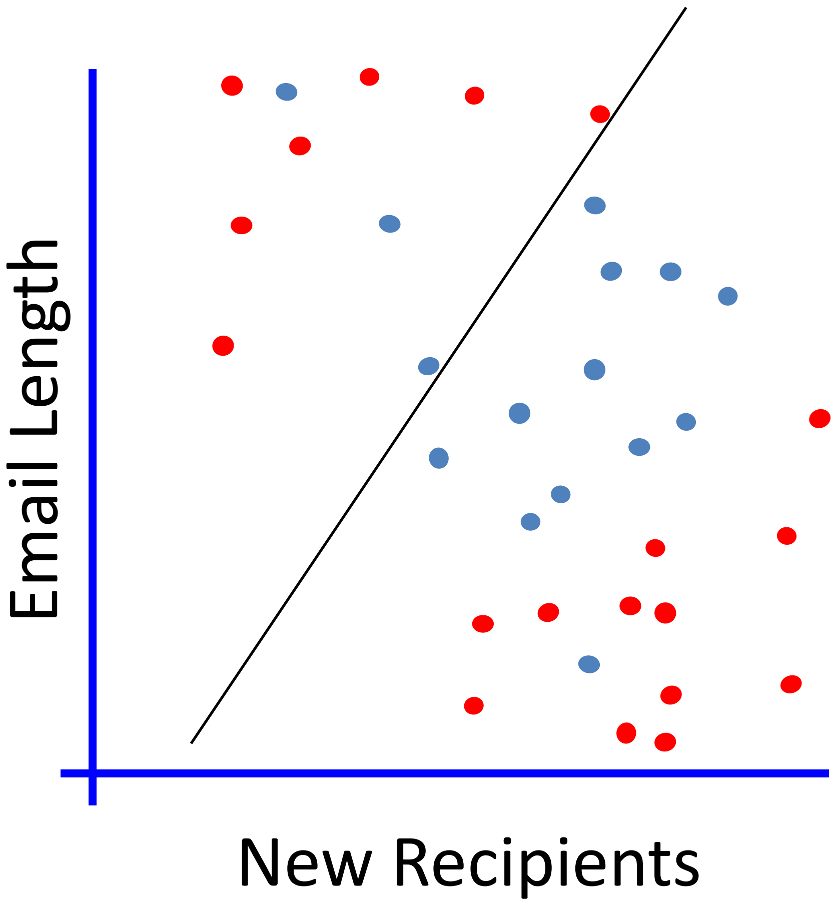
**Confusion Matrix**

|  |  | Classified As |  |
|---|---|---|---|
|  |  | **Blue** | **Red** |
| Actual | **Blue** | 7 | 1 |
|  | **Red** | 0 | 5 |

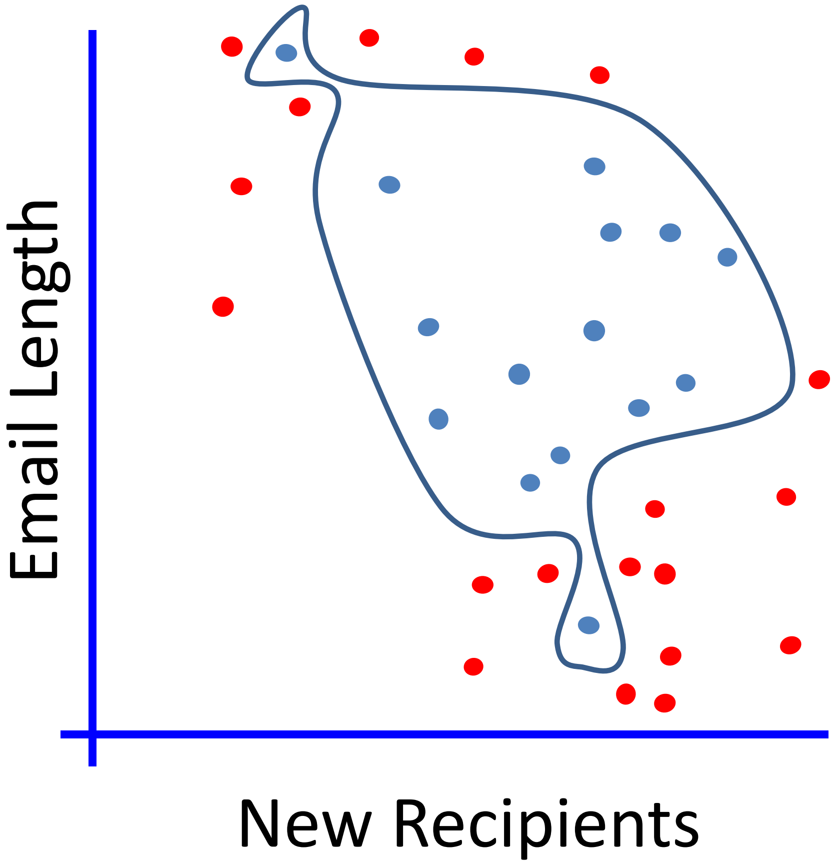
# The Non-linearly separable case



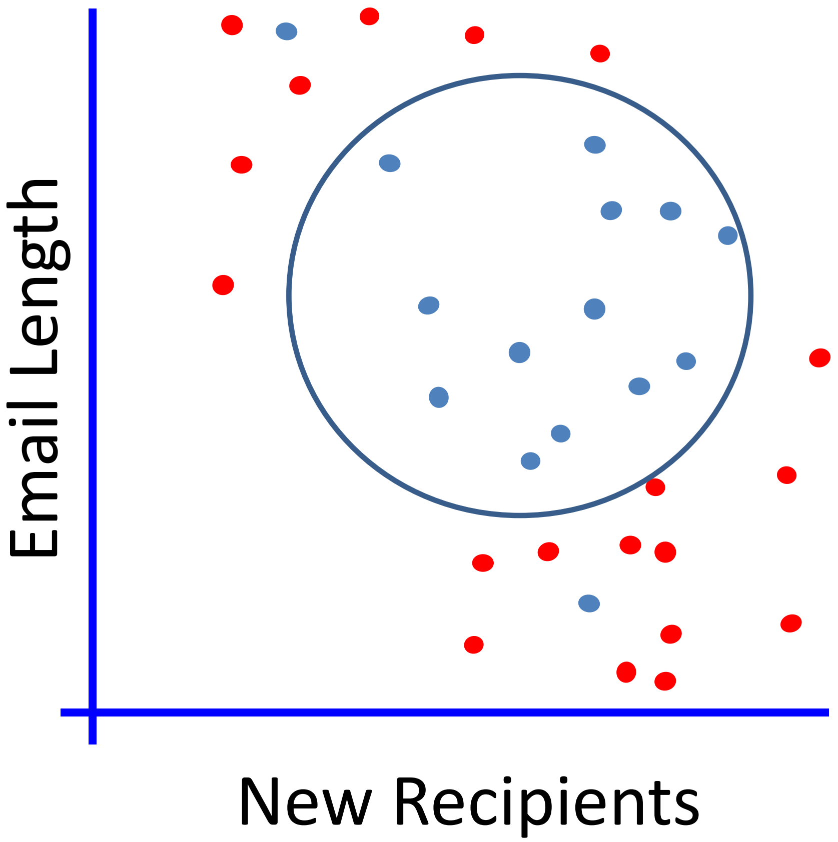
# The Non-linearly separable case



# The Non-linearly separable case

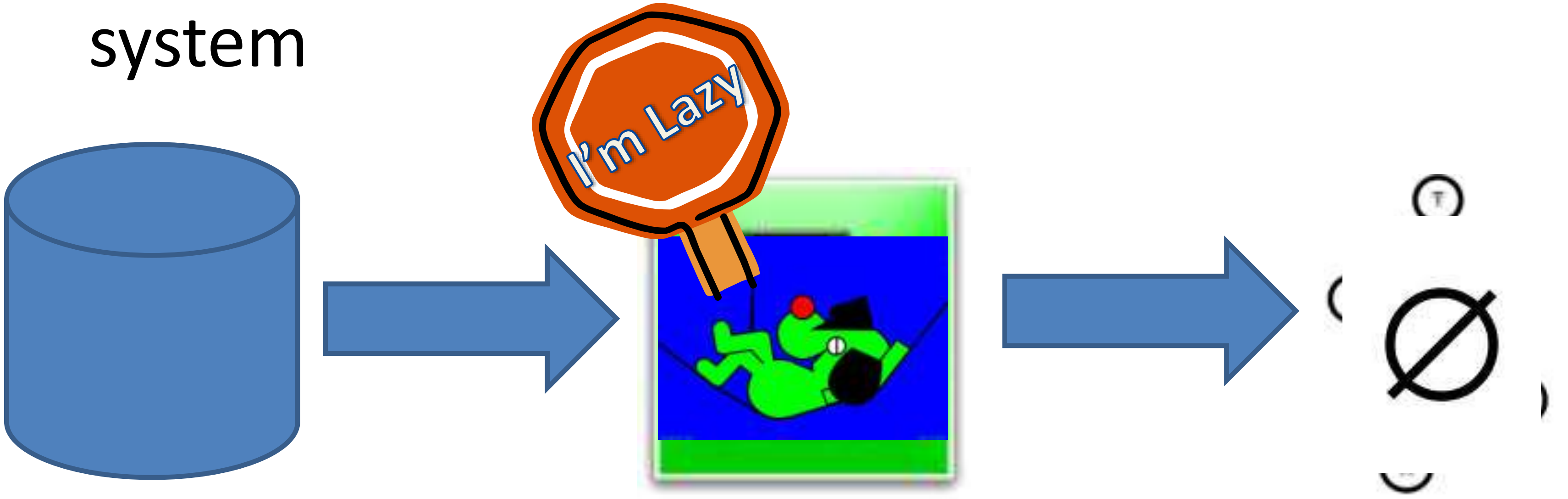


# The Non-linearly separable case



# Lazy Learners

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



Training Set

Learner

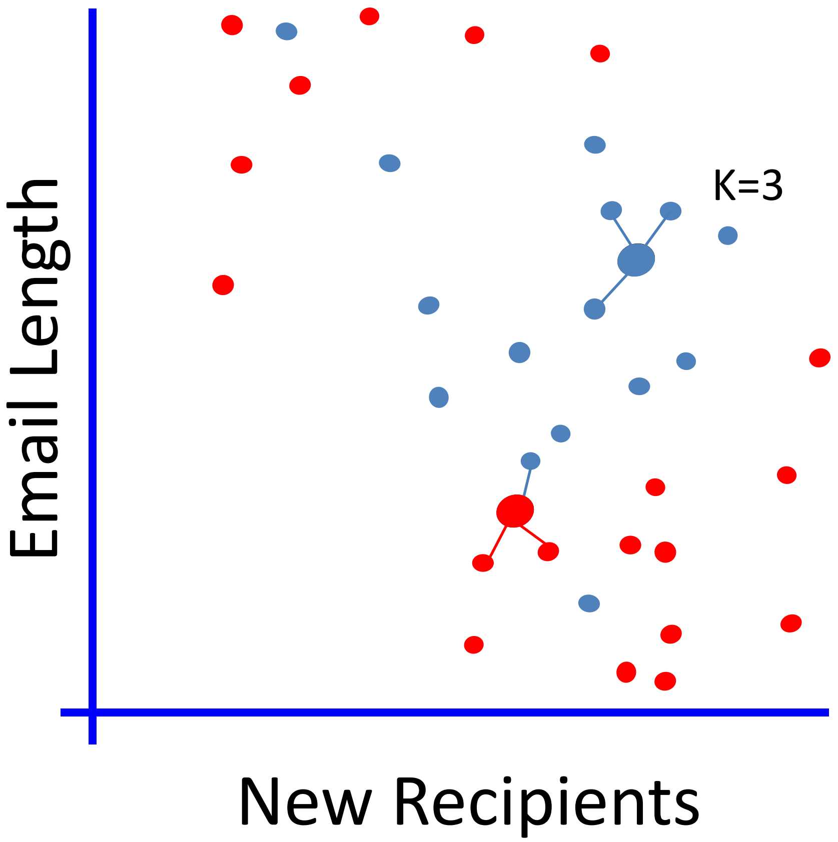
Classifier

# Lazy Learners

## Instance-based learning

# Lazy Learner: k-Nearest Neighbors

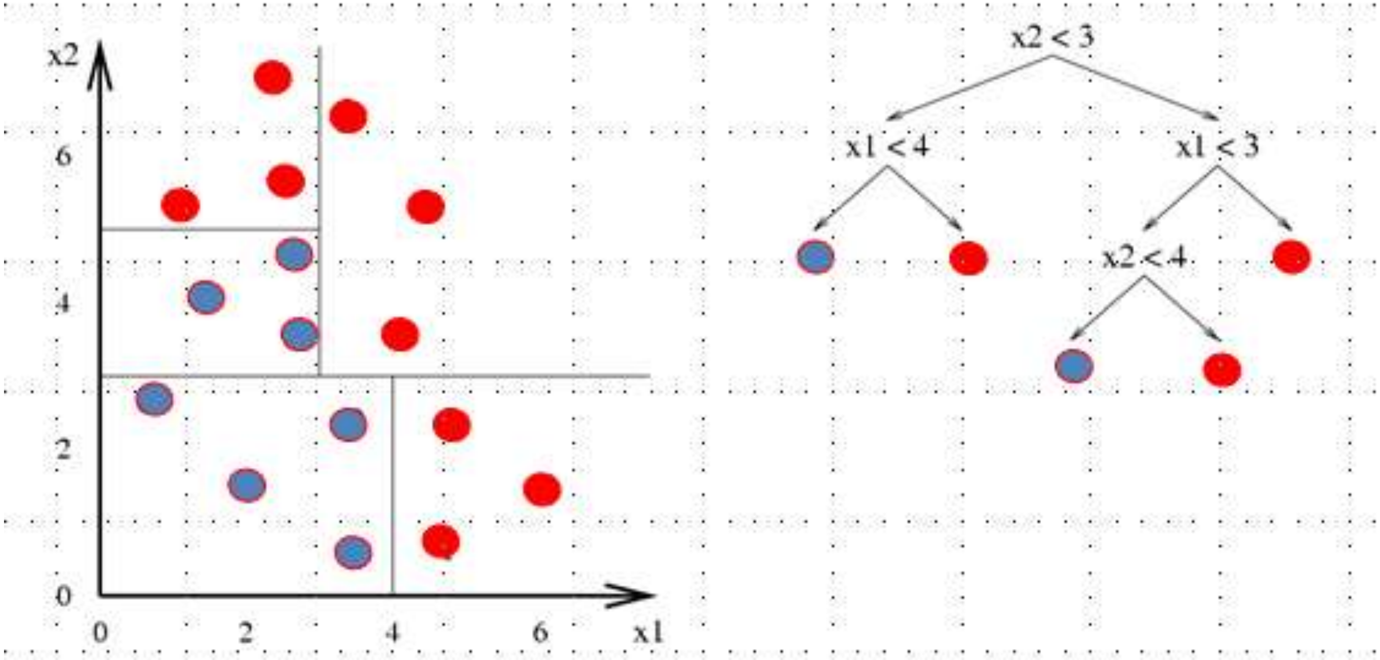
- What should be k?
- Which distance measure should be used?



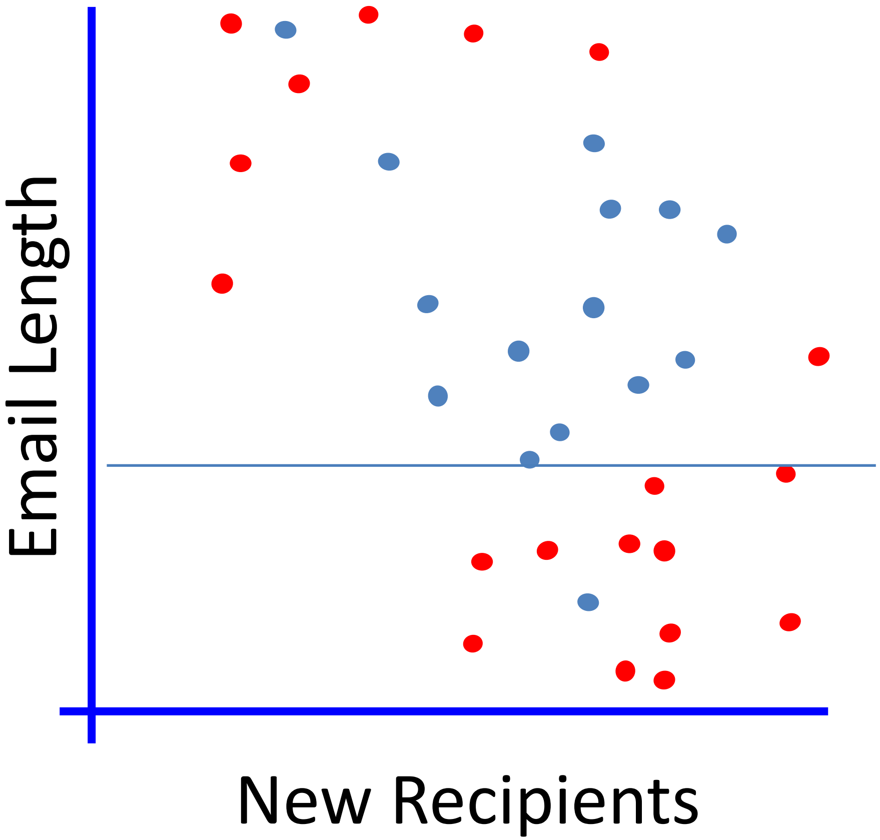
# Decision tree

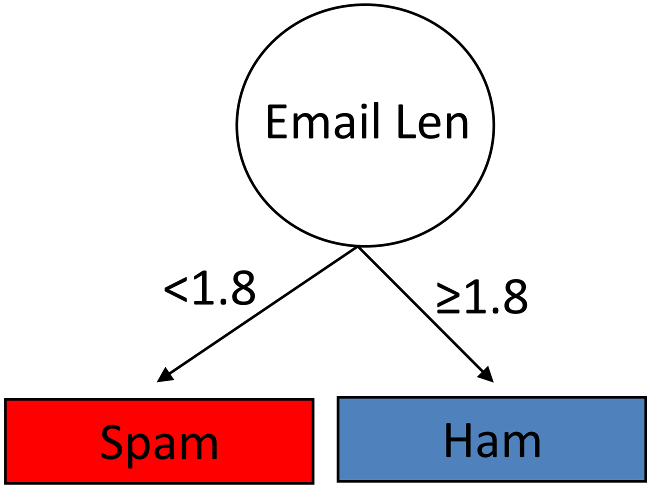
- A flow-chart-like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution

Decision trees divide the feature space into axis-parallel rectangles, and label each rectangle with one of the $K$ classes.



# Top Down Induction of Decision Trees





Email Len

<1.8

≥1.8

Spam

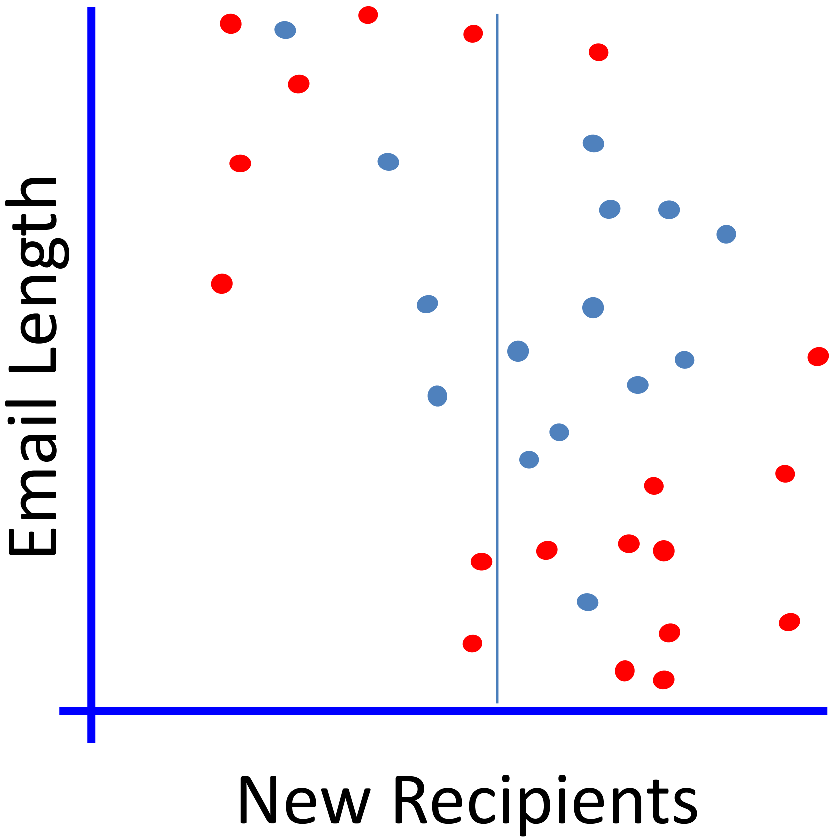
Ham

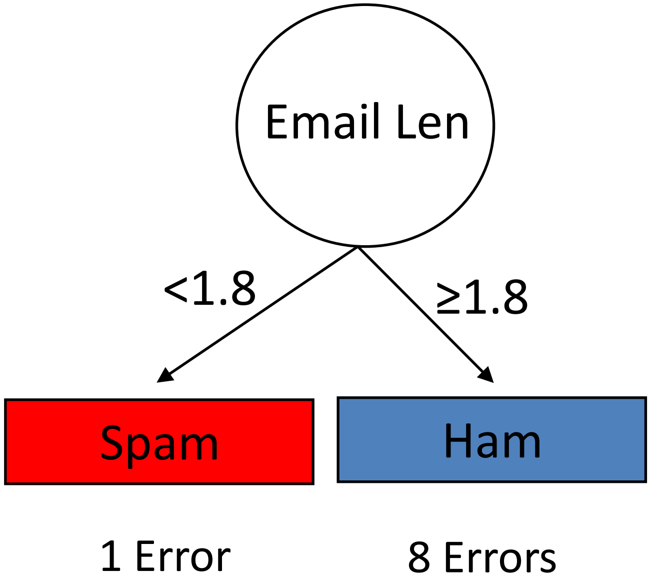
1 Error

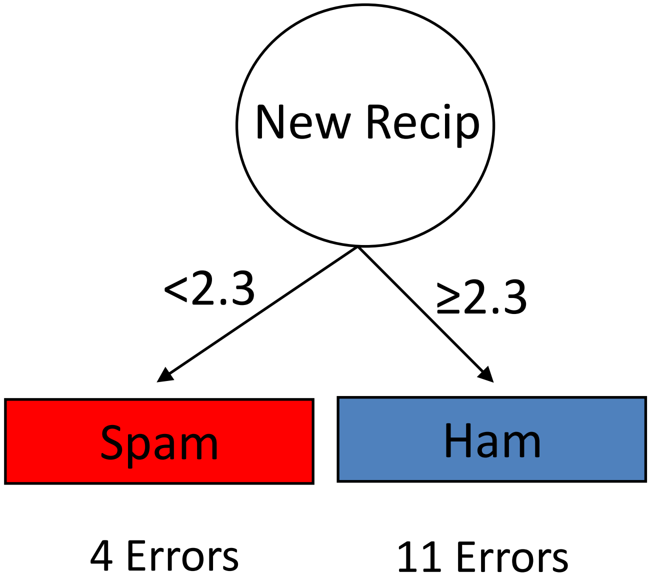
8 Errors

A single level decision tree is also known as Decision Stump

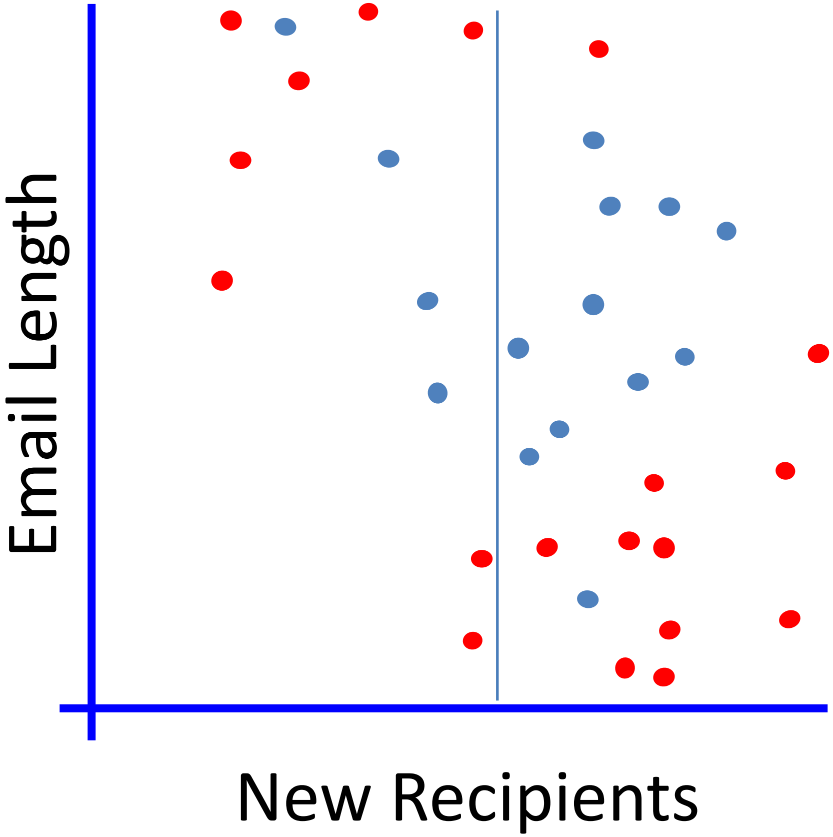
# Top Down Induction of Decision Trees

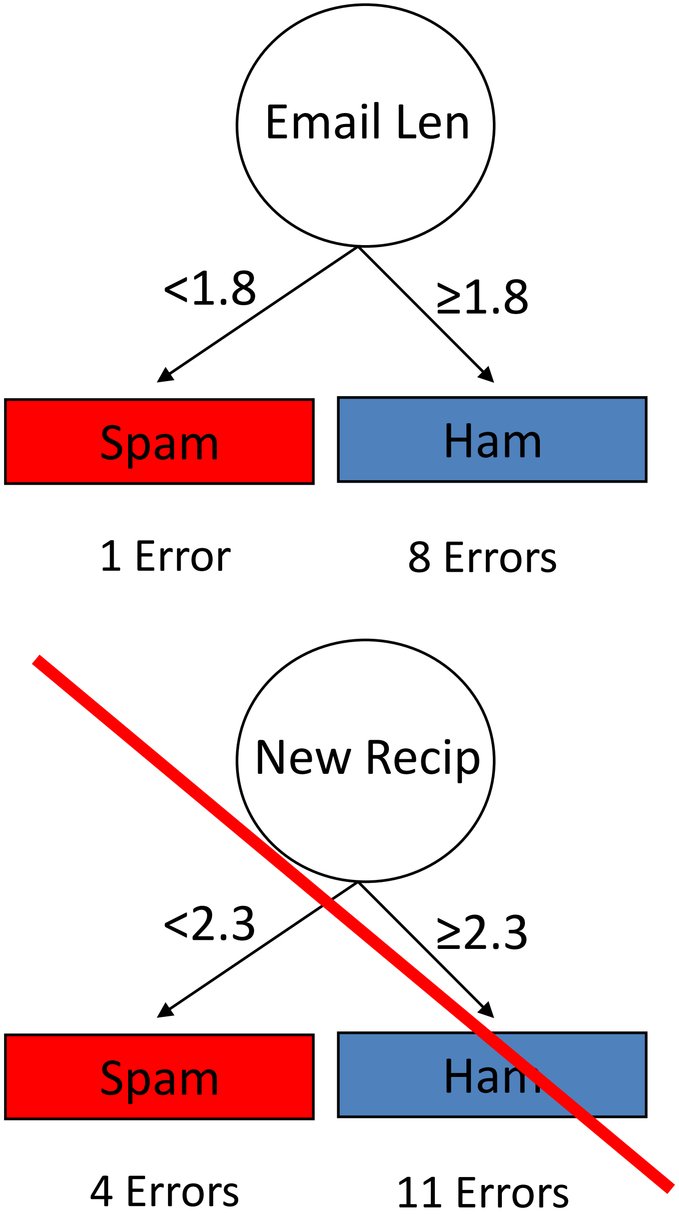




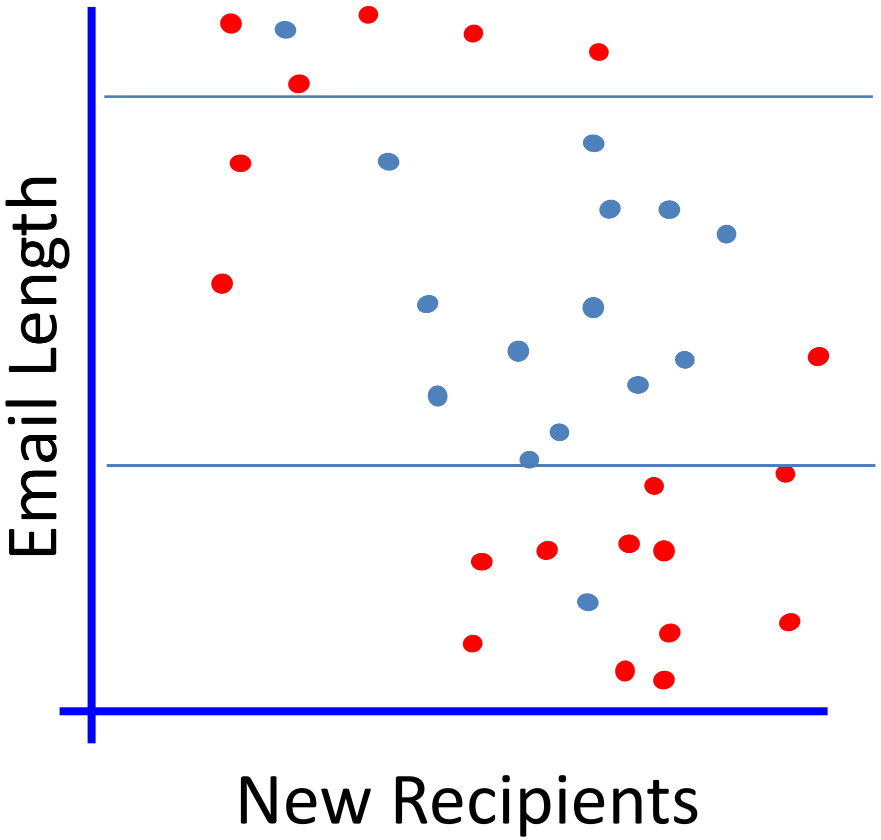


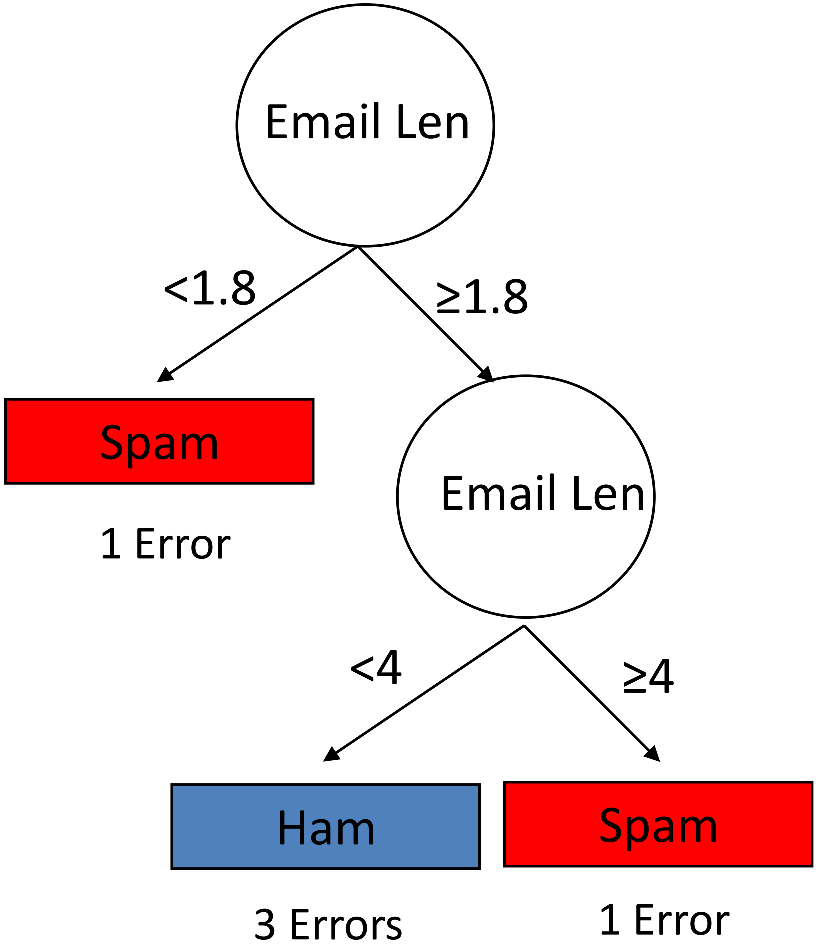
# Top Down Induction of Decision Trees





# Top Down Induction of Decision Trees





Email Len

<1.8

≥1.8

Spam

1 Error

Email Len

<4

≥4

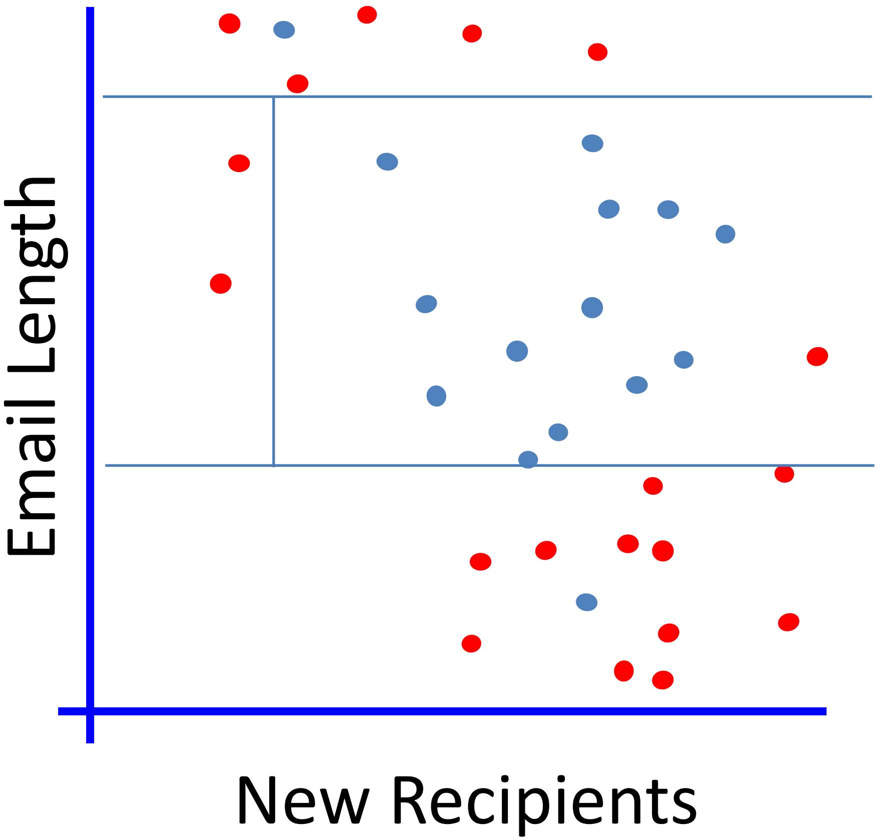
Ham

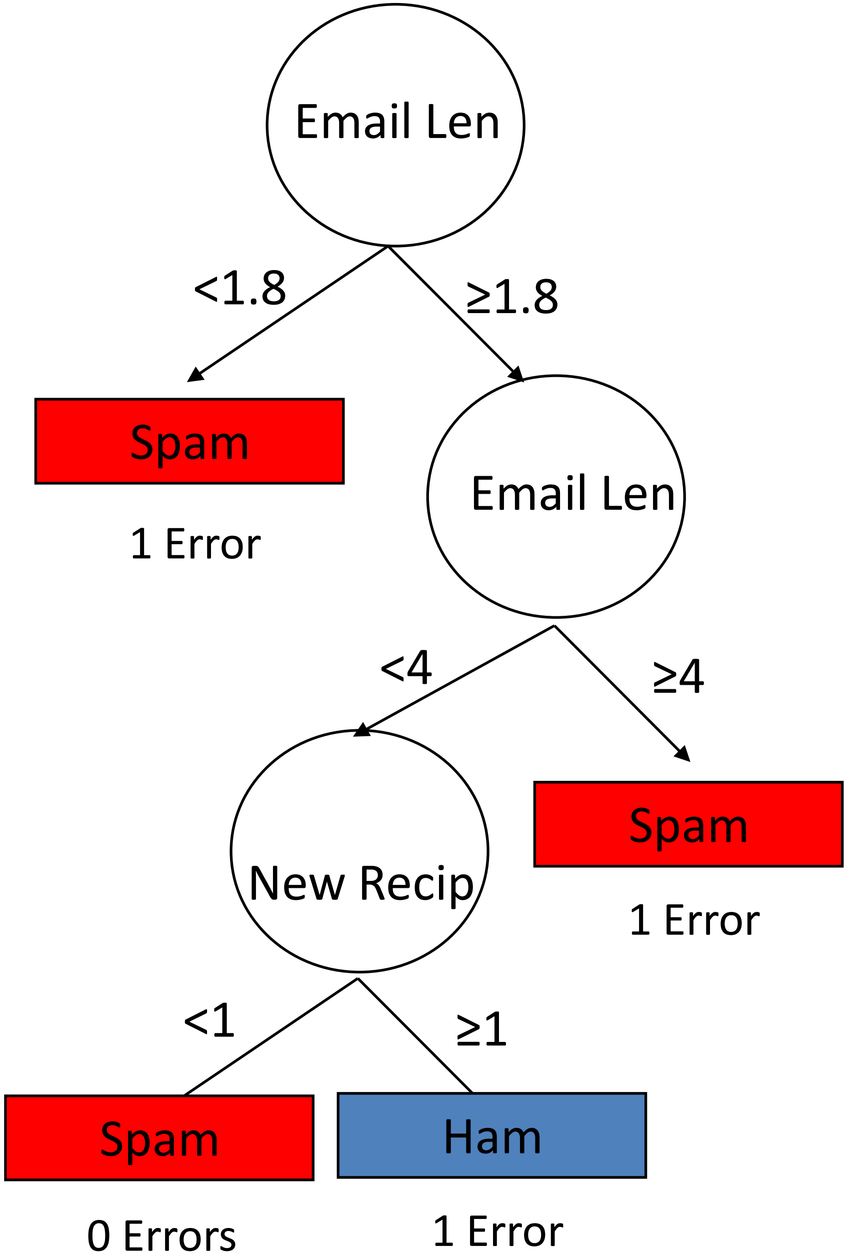
Spam

3 Errors

1 Error

# Top Down Induction of Decision Trees





Email Len

<1.8

≥1.8

Spam

1 Error

Email Len

<4

≥4

Spam

1 Error

New Recip

<1

≥1

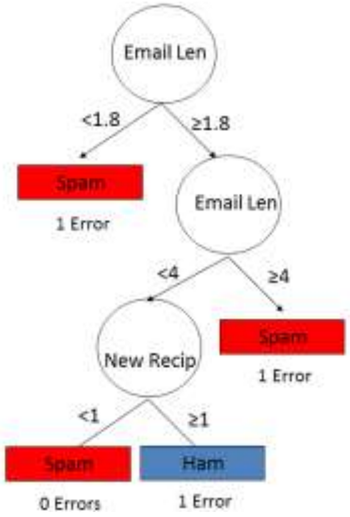
Spam

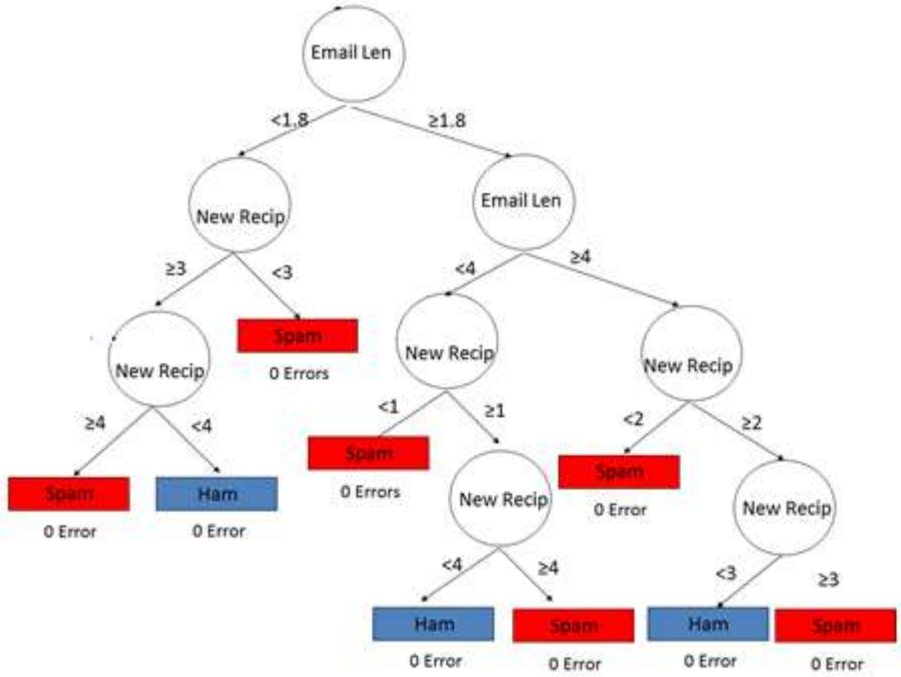
0 Errors

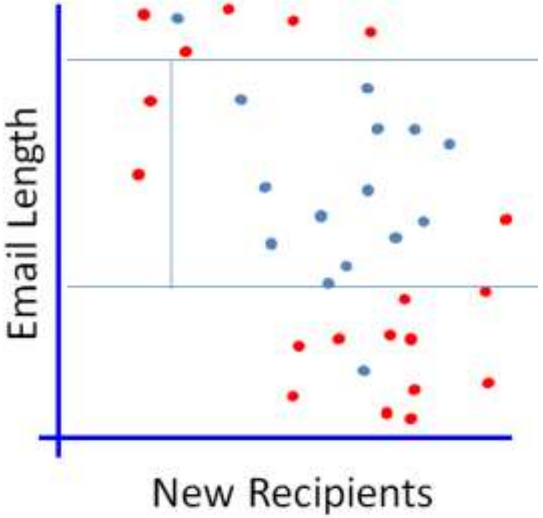
Ham

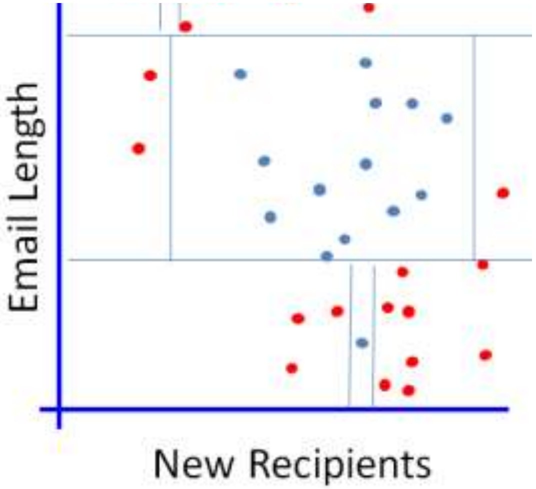
1 Error

# Which One?

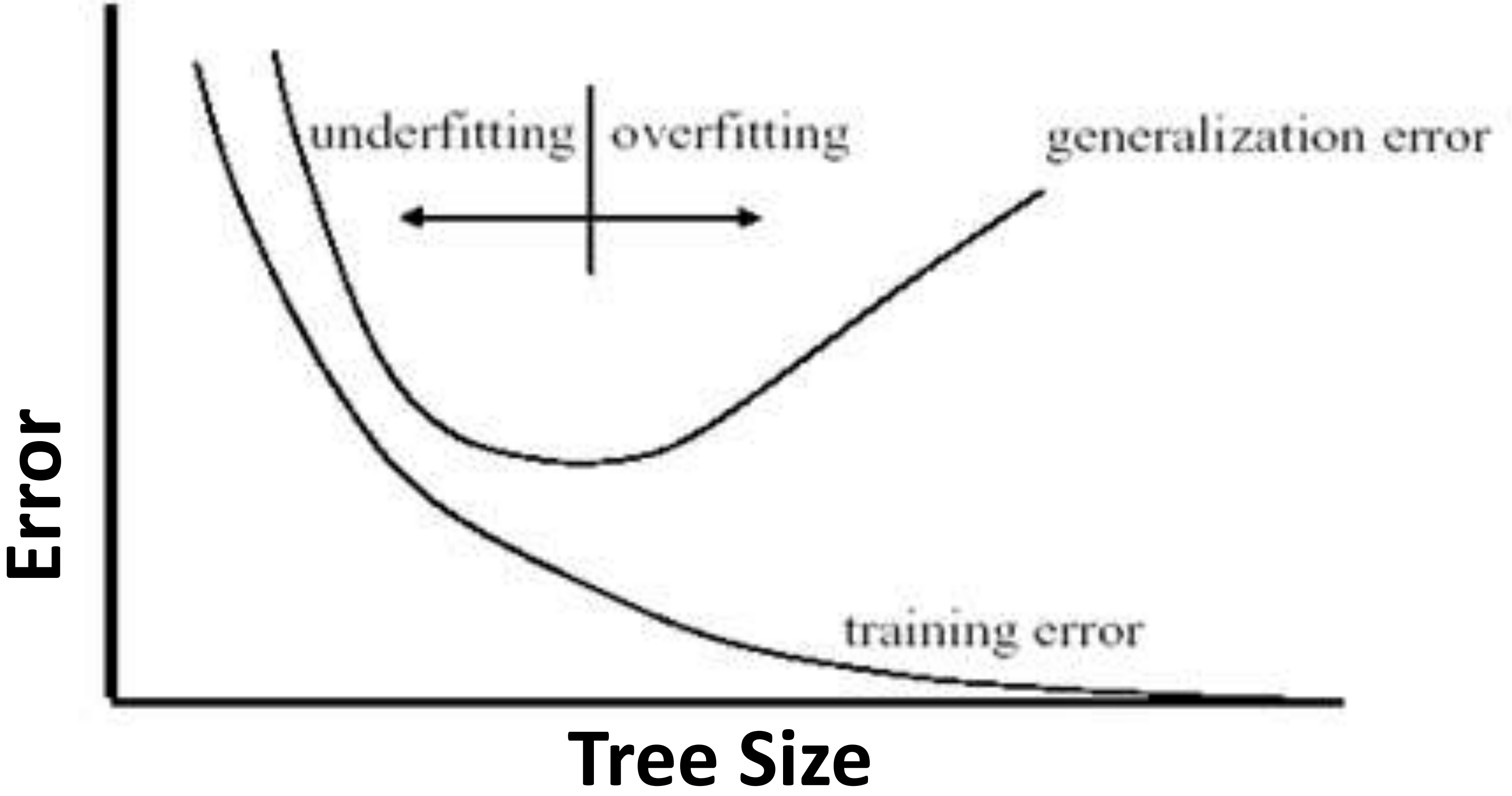

Email Len
<1.8
≥1.8
Spam
1 Error
Email Len
<4
≥4
New Recip
Spam
1 Error
<1
≥1
Spam
Ham
0 Errors
1 Error


Email Len
<1.8
≥1.8
New Recip
Email Len
≥3
<3
<4
≥4
New Recip
Spam
0 Errors
New Recip
New Recip
≥4
<4
<1
≥1
<2
≥2
Spam
Ham
Spam
New Recip
Spam
New Recip
0 Error
0 Error
0 Errors
0 Error
<4
≥4
<3
≥3
Ham
Spam
Ham
Spam
0 Error
0 Error
0 Error
0 Error


Email Length
New Recipients


Email Length
New Recipients

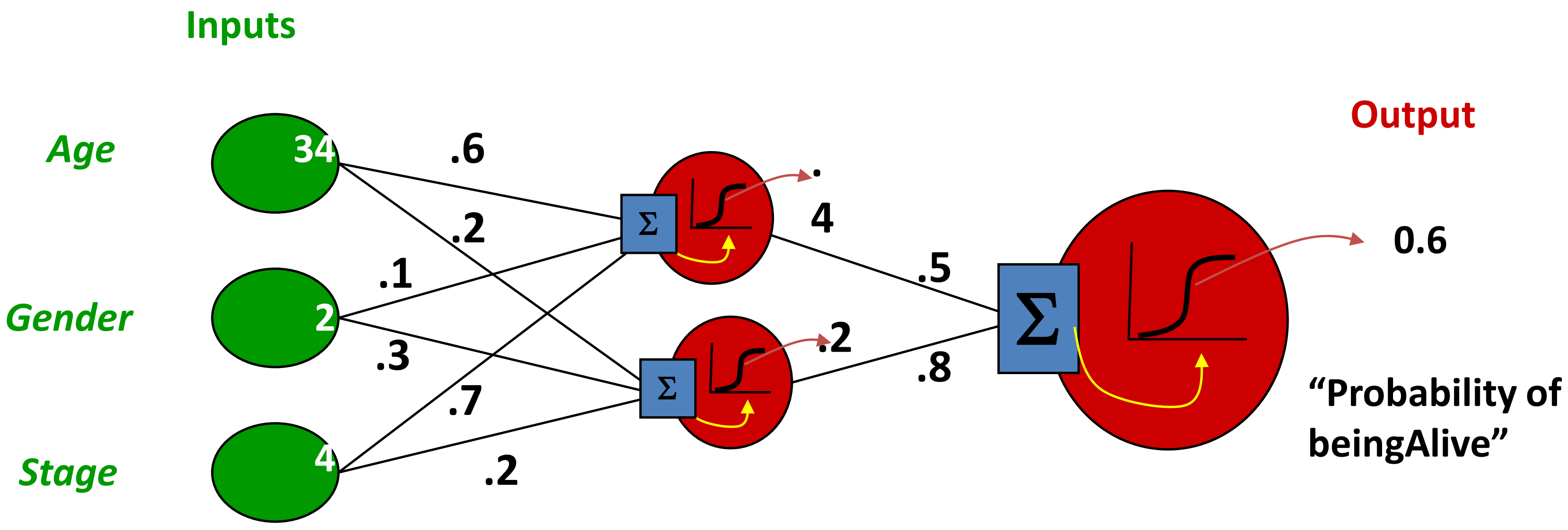
# Overfitting and underfitting



**Overtraining:** means that it learns 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

# Neural Network Model



Inputs

Output

Age

34

Gender

2

Stage

4

.6

.2

.1

.3

.7

.2

.

4

.2

.5

.8

0.6

"Probability of beingAlive"

*Independent variables*

Weights

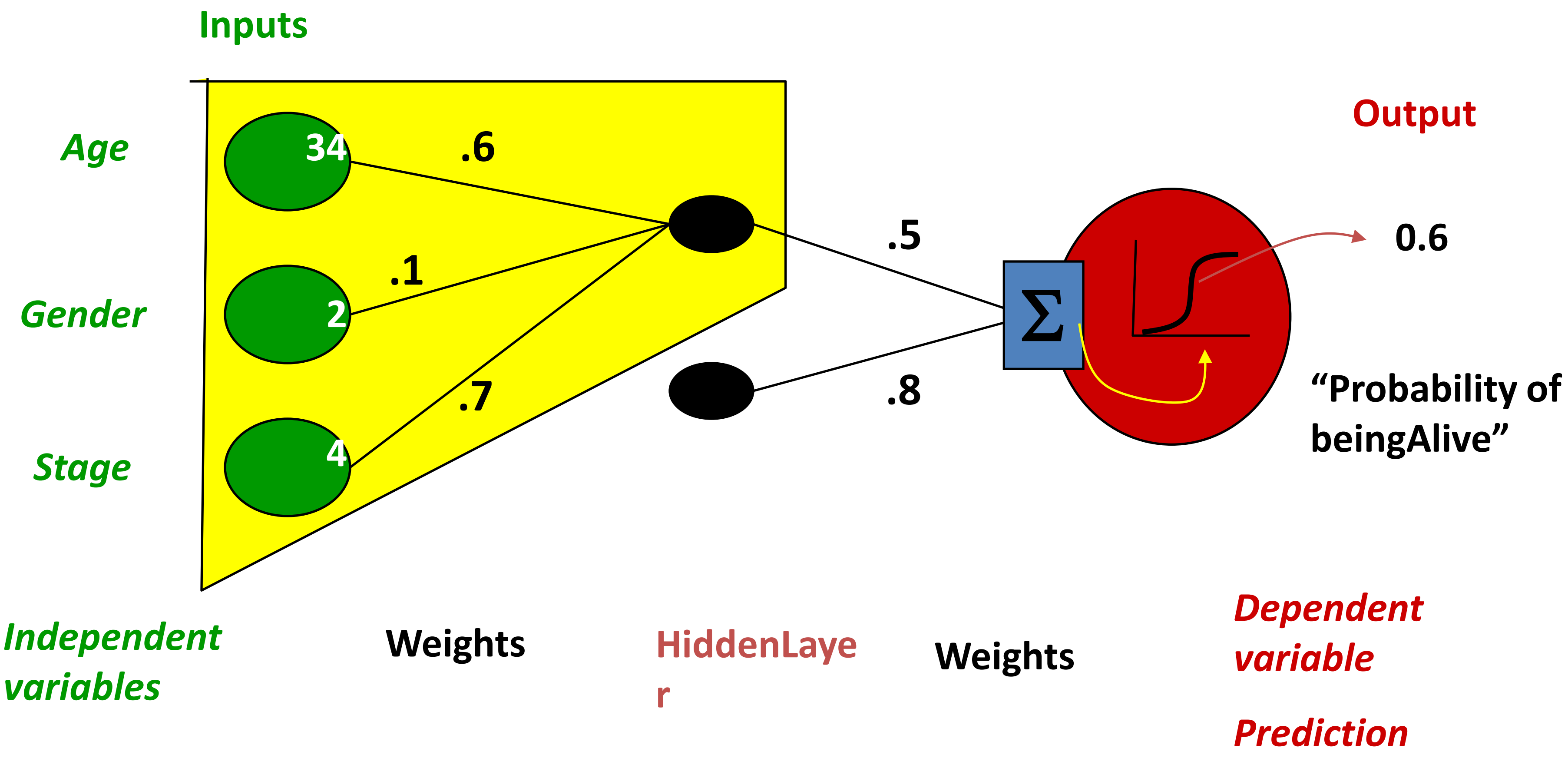
HiddenLaye
r

Weights

*Dependent variable*

*Prediction*

# “Combined logistic models”



Inputs

*Age*

*Gender*

*Stage*

Output

“Probability of beingAlive”

*Independent variables*

Weights

HiddenLayer

Weights

*Dependent variable*

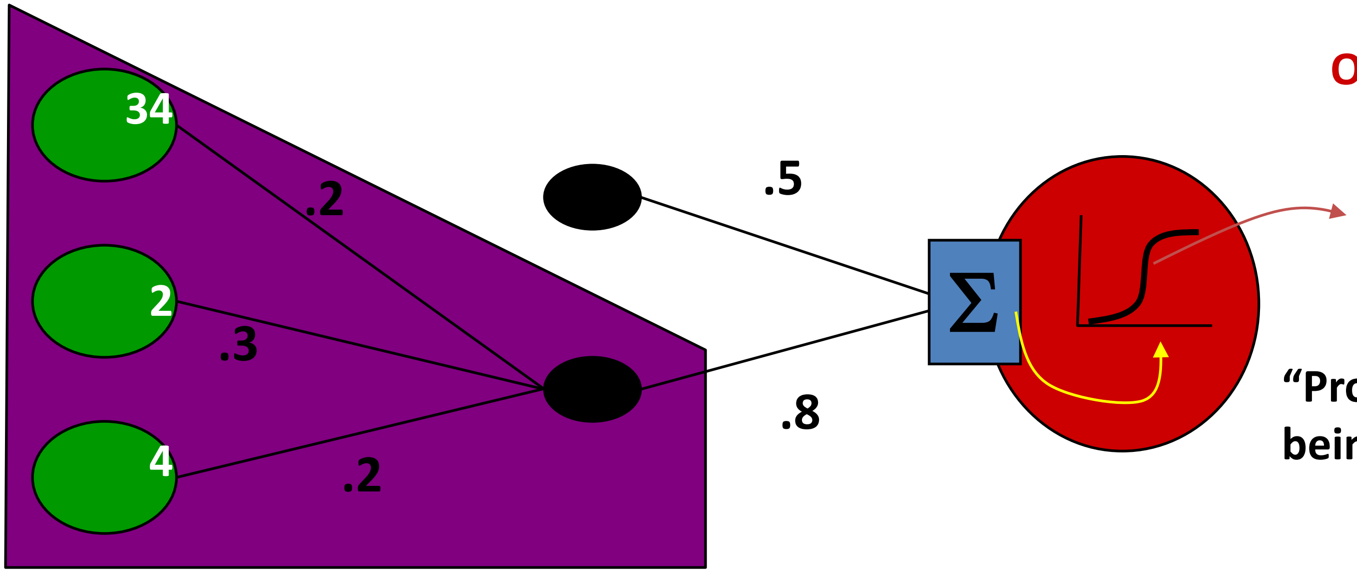
*Prediction*

Inputs

*Age*

*Gender*

*Stage*



**Output**

0.6

“Probability of beingAlive”

*Independent variables*

Weights

HiddenLaye r

Weights

*Dependent variable*

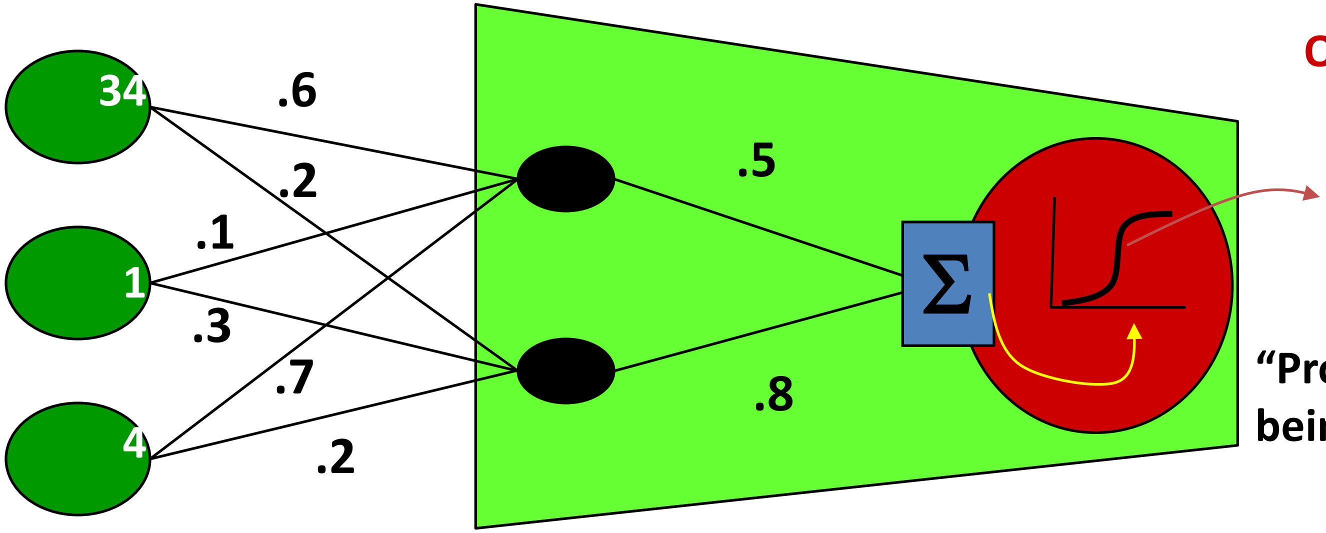
*Prediction*

**Inputs**

*Age*

*Gender*

*Stage*



**Output**

**0.6**

**“Probability of beingAlive”**

*Independent variables*

**Weights**

**HiddenLayer**

**Weights**

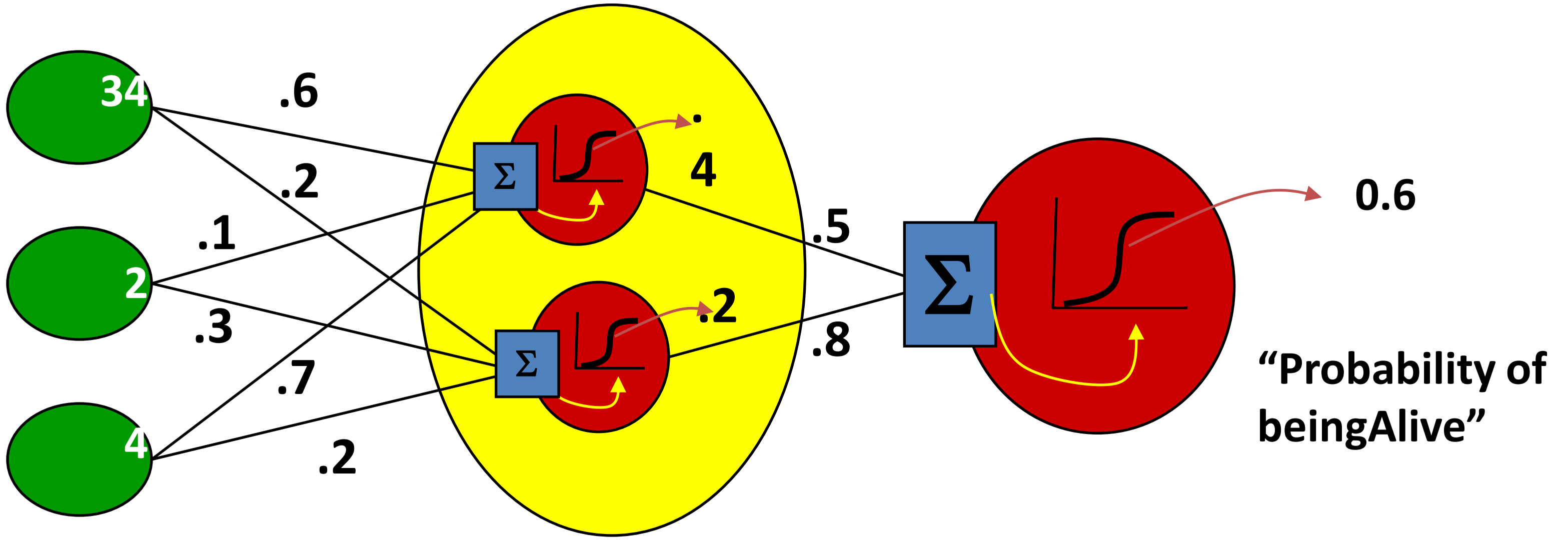
*Dependent variable*

*Prediction*

*Age*

*Gender*

*Stage*



*Independent variables*

**Weights**

**HiddenLaye r**

**Weights**

***Dependent variable***

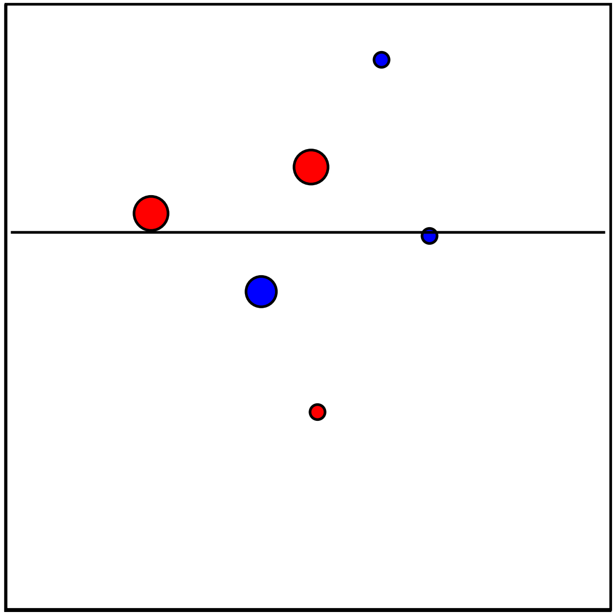
***Prediction***

# Ensemble Learning

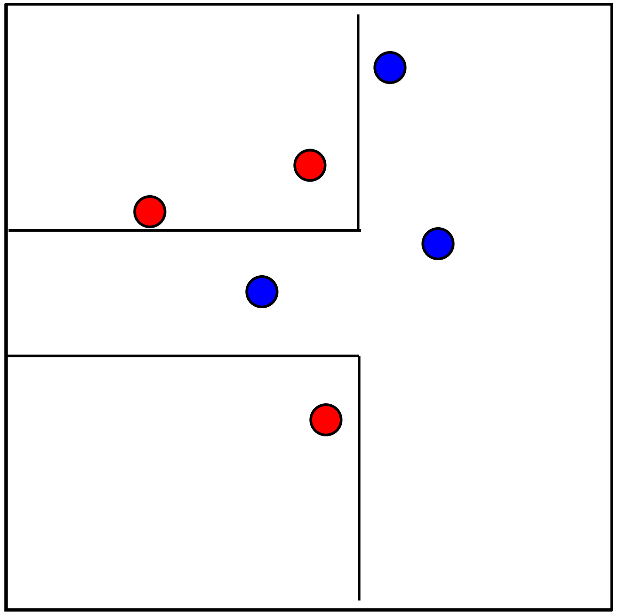
- The idea is to use multiple models to obtain better predictive performance than could be obtained from any of the constituent models.

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

# Example of Ensemble of Weak Classifiers



Training



Combined classifier

# Main Principles

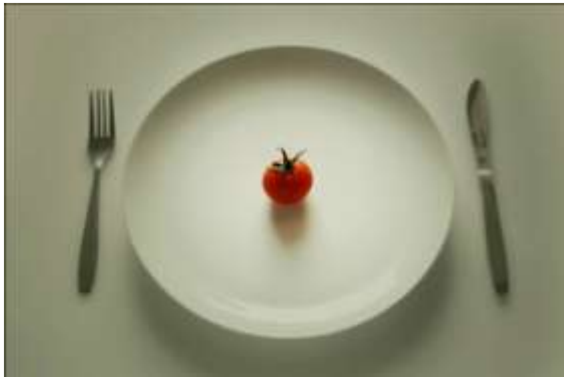
Curse of Dimensionality
Goodness of fit
Vapnik-Chervonenkis dimension
Learning rate
Generative model
Empirical risk minimization
Bayesian information criterion (BIC)
Bayes optimal
Underfitting
Empirical Bayes
Maximum posterior (MAP)
Maximum margin
Occam's razor
Structural risk minimization
Generalization
Computational intelligence
Kernel density estimator
0/1 loss
Kernel trick
Maximum likelihood
Computational learning theory
Bias-variance tradeoff
Probably approximately correct (PAC)
No free lunch theorem
Inductive bias
Loss function
Minimum message length (MML)
Computational complexity
Overfitting
Minimum description length (MDL)
Hypothesis space
Akaike information criterion (AIC)

# **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)

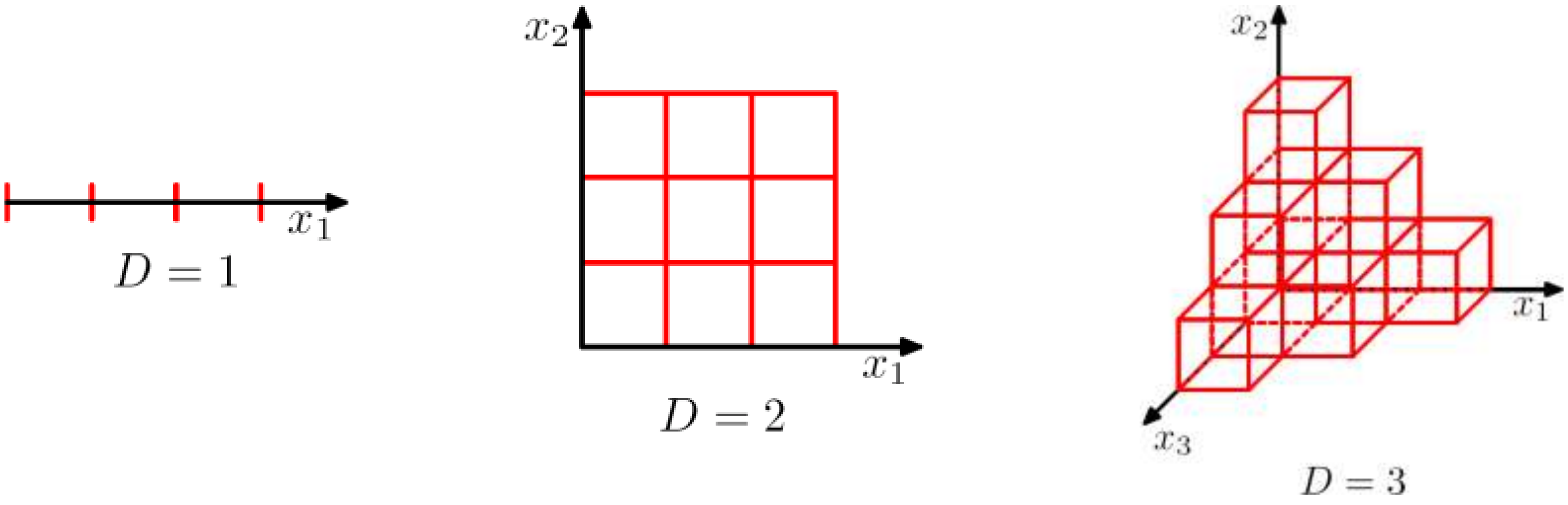
- *“For any two learning algorithms, there are just as many situations (appropriately weighted) in which algorithm one is superior to algorithm two as vice versa, according to any of the measures of "superiority"*

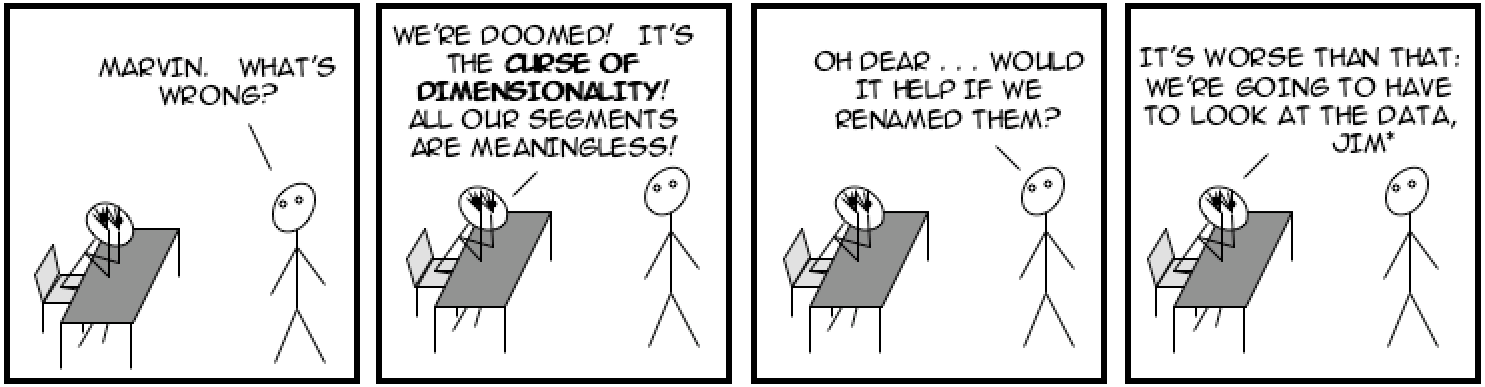


# So why developing new algorithms?

- 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)



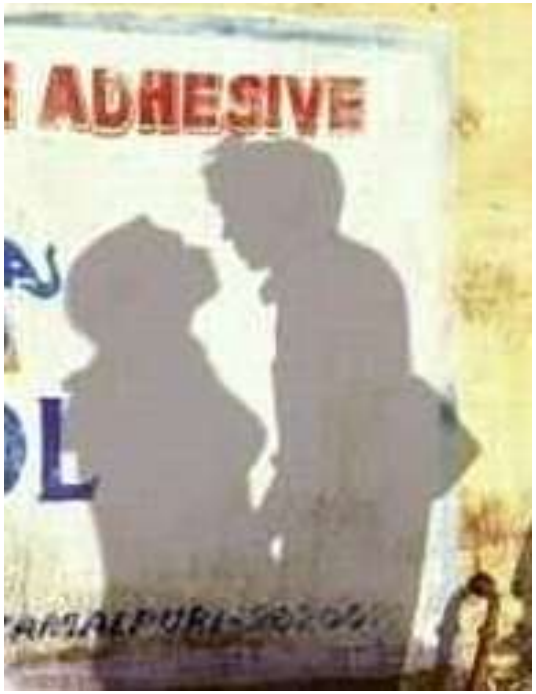

MARVIN. WHAT'S WRONG?
WE'RE DOOMED! IT'S THE CURSE OF DIMENSIONALITY! ALL OUR SEGMENTS ARE MEANINGLESS!
OH DEAR . . . WOULD IT HELP IF WE RENAMED THEM?
IT'S WORSE THAN THAT: WE'RE GOING TO HAVE TO LOOK AT THE DATA, JIM*

HTTP://SCIENTIFICMARKETER.COM

* WITH APOLOGIES TO MR SPOCK & STAR TREK.

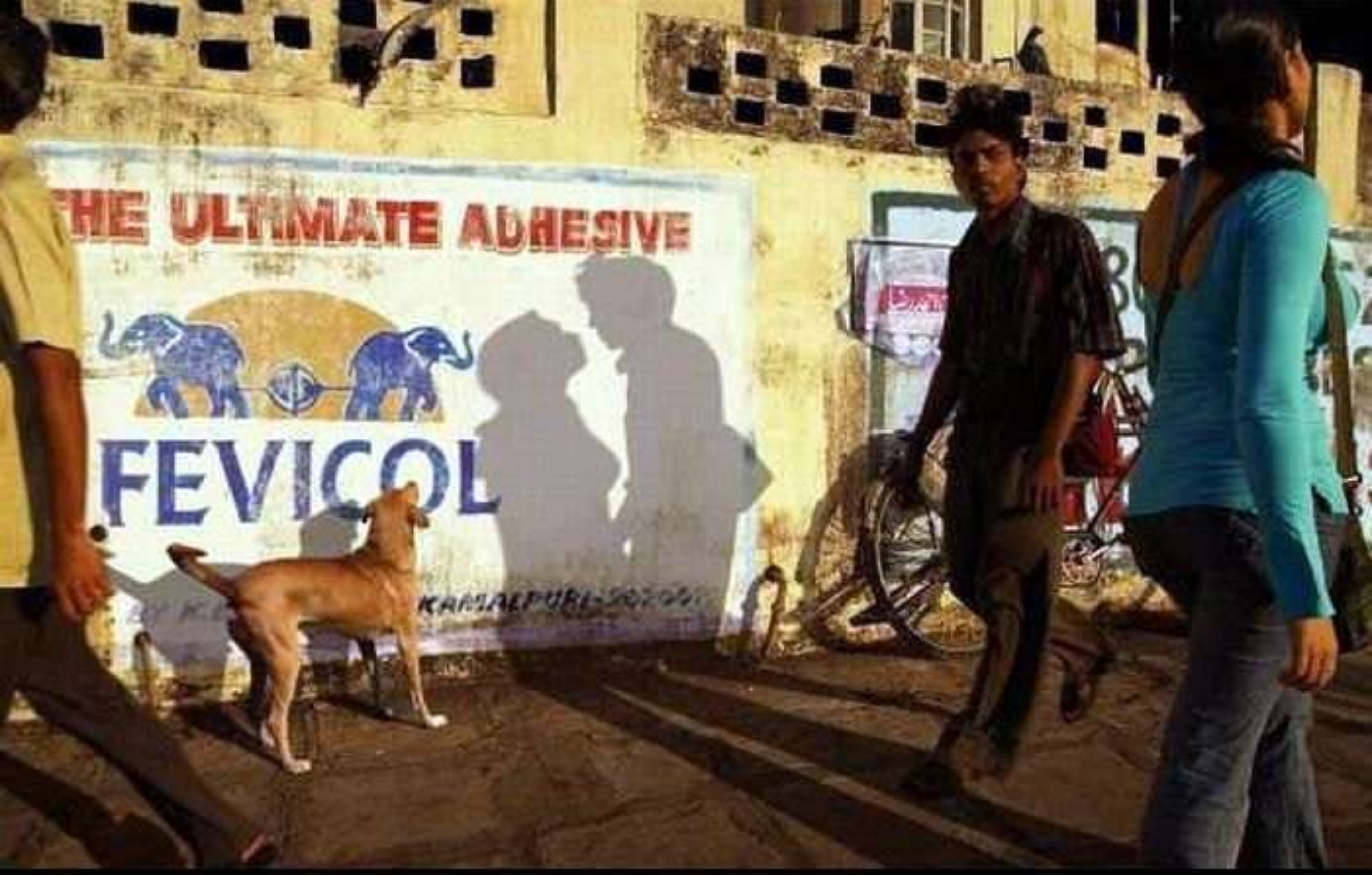
# Less is More
# The Curse of Dimensionality

- 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.



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.

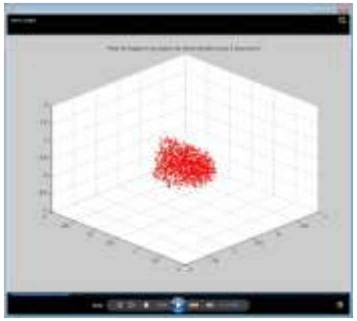
Above is a two dimensional projection of an intrinsically three dimensional world….



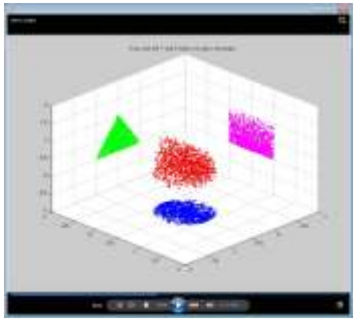
*Original photographer unknown*
See also www.cs.gmu.edu/~jessica/DimReducDanger.htm (c) eamonn keogh

Screen dumps of a short video from www.cs.gmu.edu/~jessica/DimReducDanger.htm
I recommend you imbed the original video instead

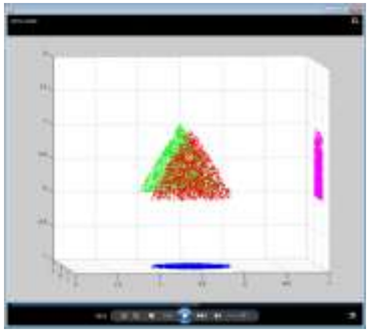
A cloud of points in 3D



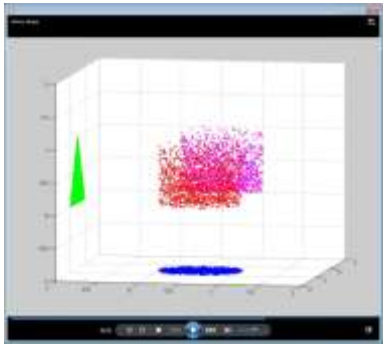
Can be projected into 2D
XY or XZ or YZ



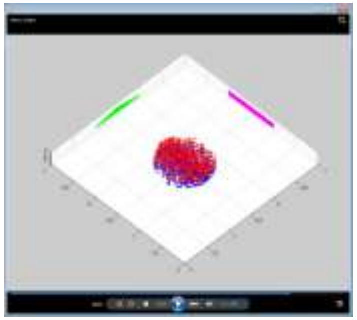
In 2D XZ we see
a triangle



In 2D YZ we see
a square



In 2D XY we see
a circle



# Less is More?

- 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

1. Track Terrorists
2. 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

- $10^9$ 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 $10^5$ 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* :
  - $\overset{p \text{ at some hotel}}{\boxed{1/100}} \times \overset{q \text{ at some hotel}}{\boxed{1/100}} \times \overset{\text{Same hotel}}{\boxed{10^{-5}}} = 10^{-9}$.
- 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:
  - $5 \times 10^5$.

# 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:
  - $5\times10^{17}$.
- 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.
- 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

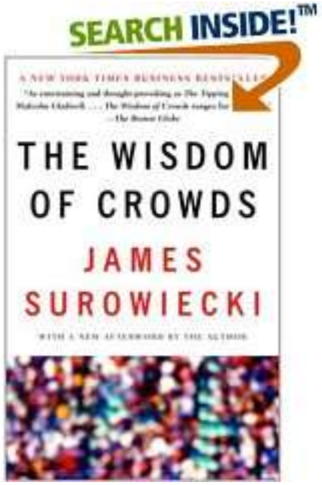
- Rashomon is a Japanese movie in which four people, from different vantage points, witness a criminal incident. When they come to testify in court, they all report the same facts, but their stories of what happened are very different.
- The Rashomon effect is the effect of the subjectivity of perception 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 decition trees, if the training set is perturbed only slightly, I can get a tree quite different from the original but with almost the same test set error.

日本最初のベニス国際映画祭グランプリ アカデミー外国映画賞受賞作品
羅生門
むせかえる真夏の草いきれの中で繰り展げられる盗賊と美女とその夫の、息詰るような愛慾絵巻！
黒沢 明 監督作品
三船敏郎 京マチ子 志村喬 森雅之
上田吉二郎 加東大介 千秋実 本間文子
原作 芥川龍之介
脚本 黒沢明 橋本忍
撮影 宮川一夫
大映映画

# *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 opinions before making any crucial decision. We weigh the individual opinions, and combine them to reach a final decision

# Committees of Experts

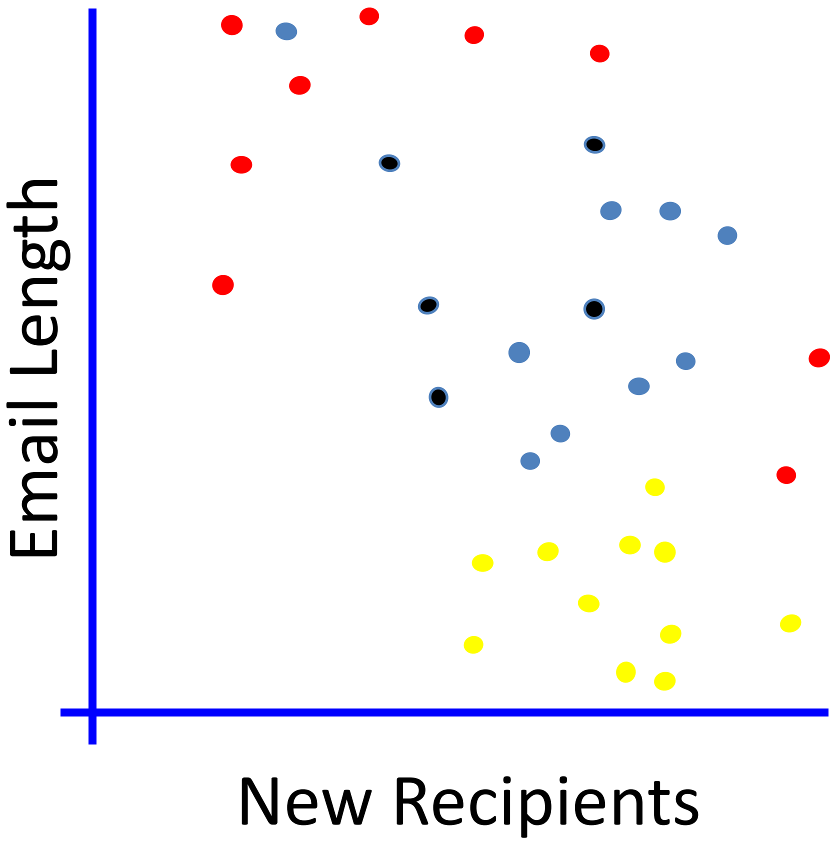
  - " … a medical school that has the objective that all students, given a problem, come up with an identical solution"

- There is not much point in setting up a committee of experts from such a group - such a committee will not improve on the judgment of an individual.

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

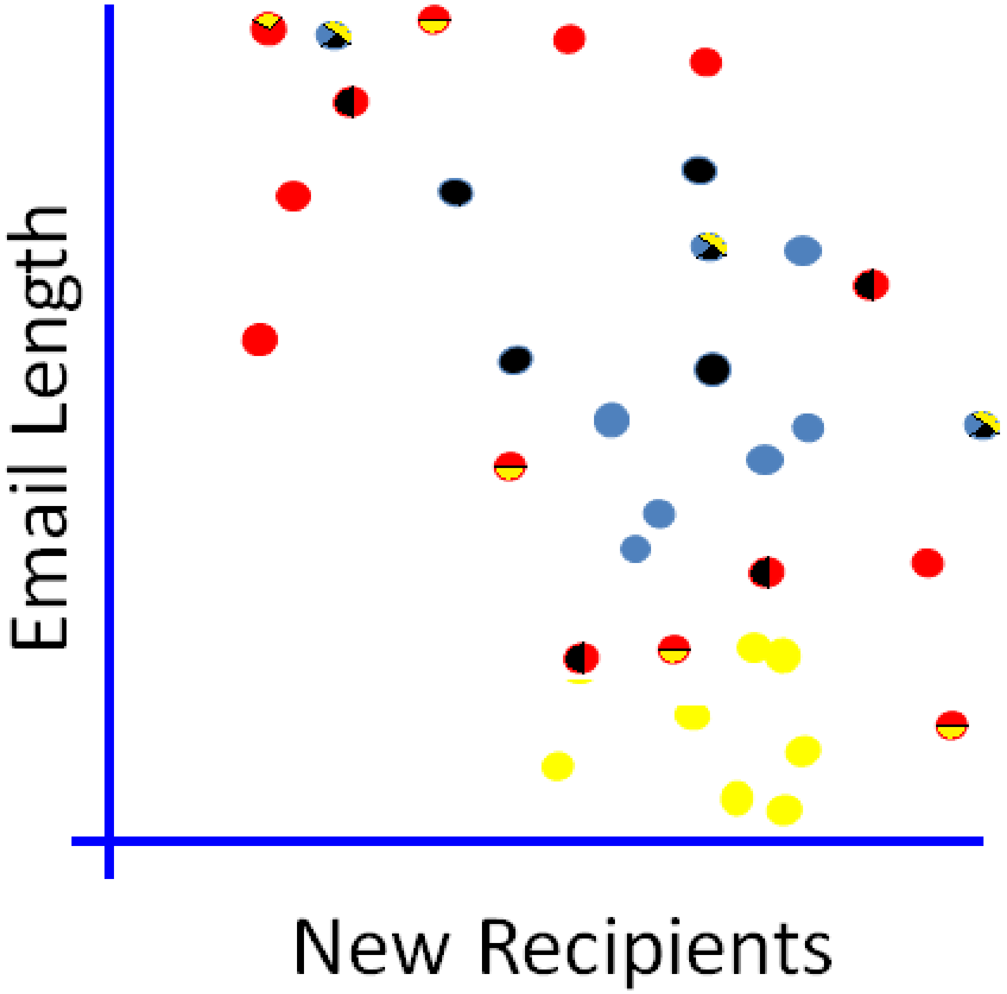
# Other Learning Tasks

# Supervised Learning - Multi Class



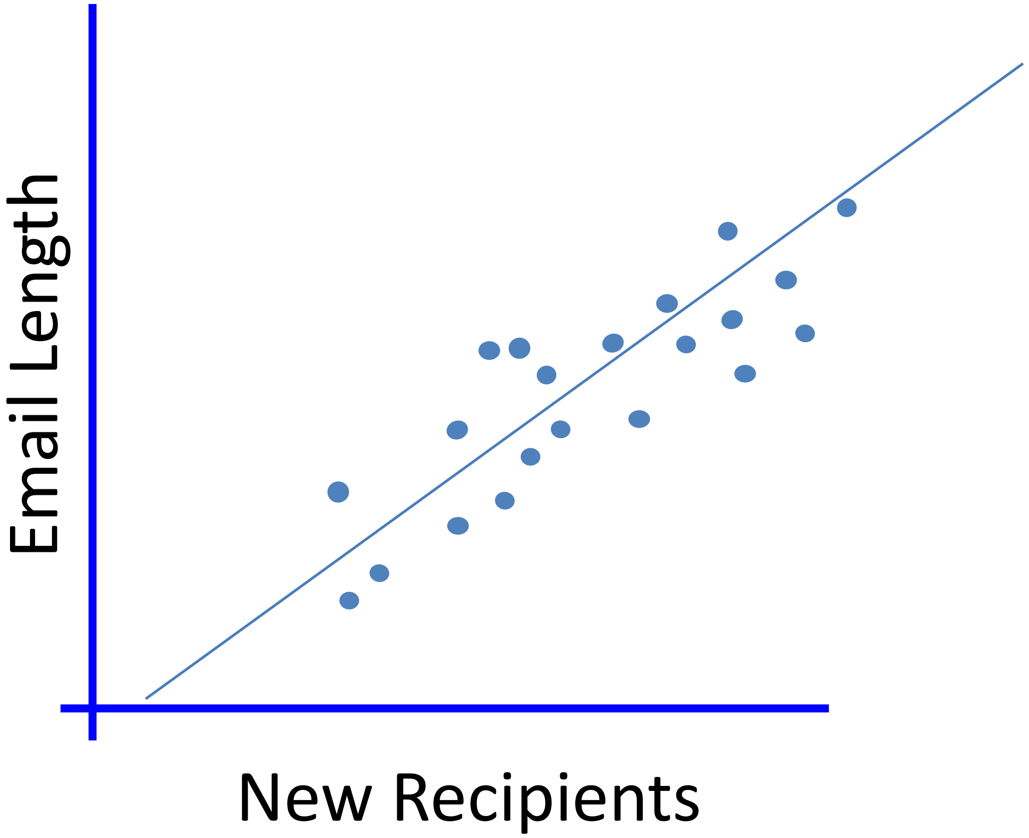
# Supervised Learning - Multi Label

*Multi-label learning* refers to the classification problem where each example can be assigned to multiple class labels simultaneously



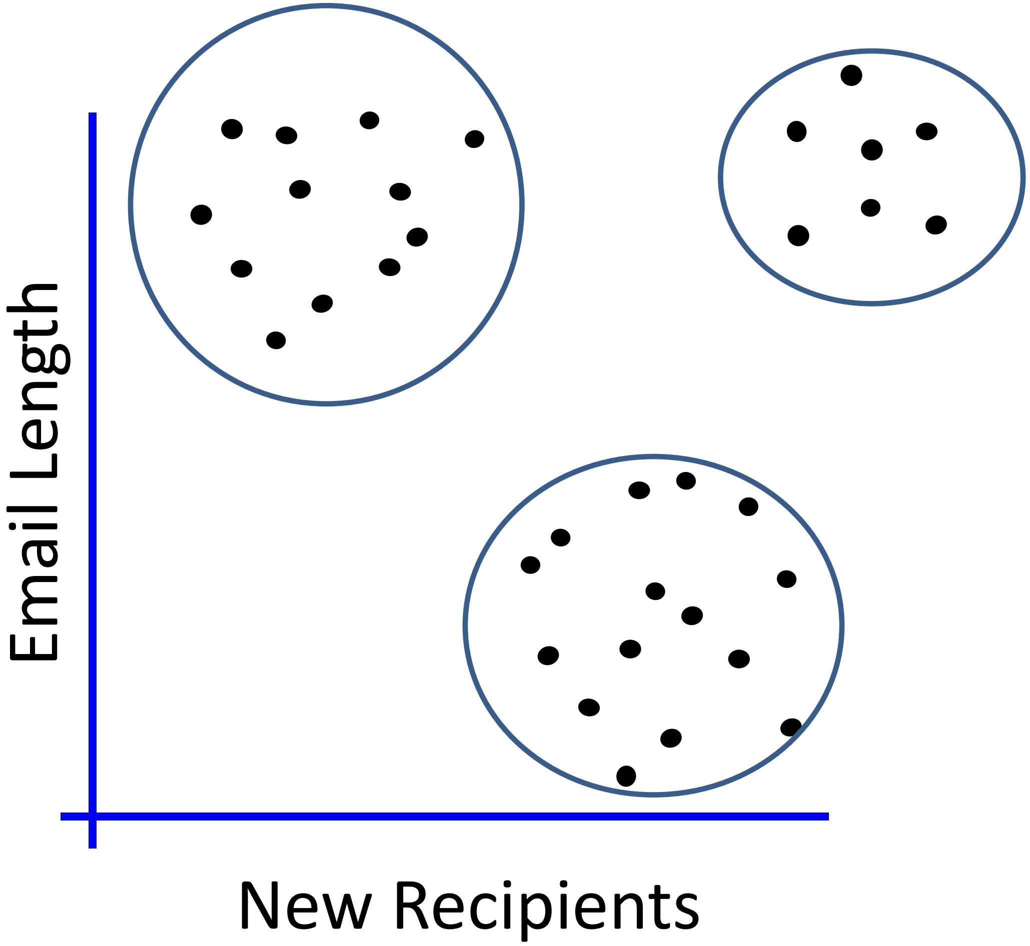
# Supervised Learning - Regression

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



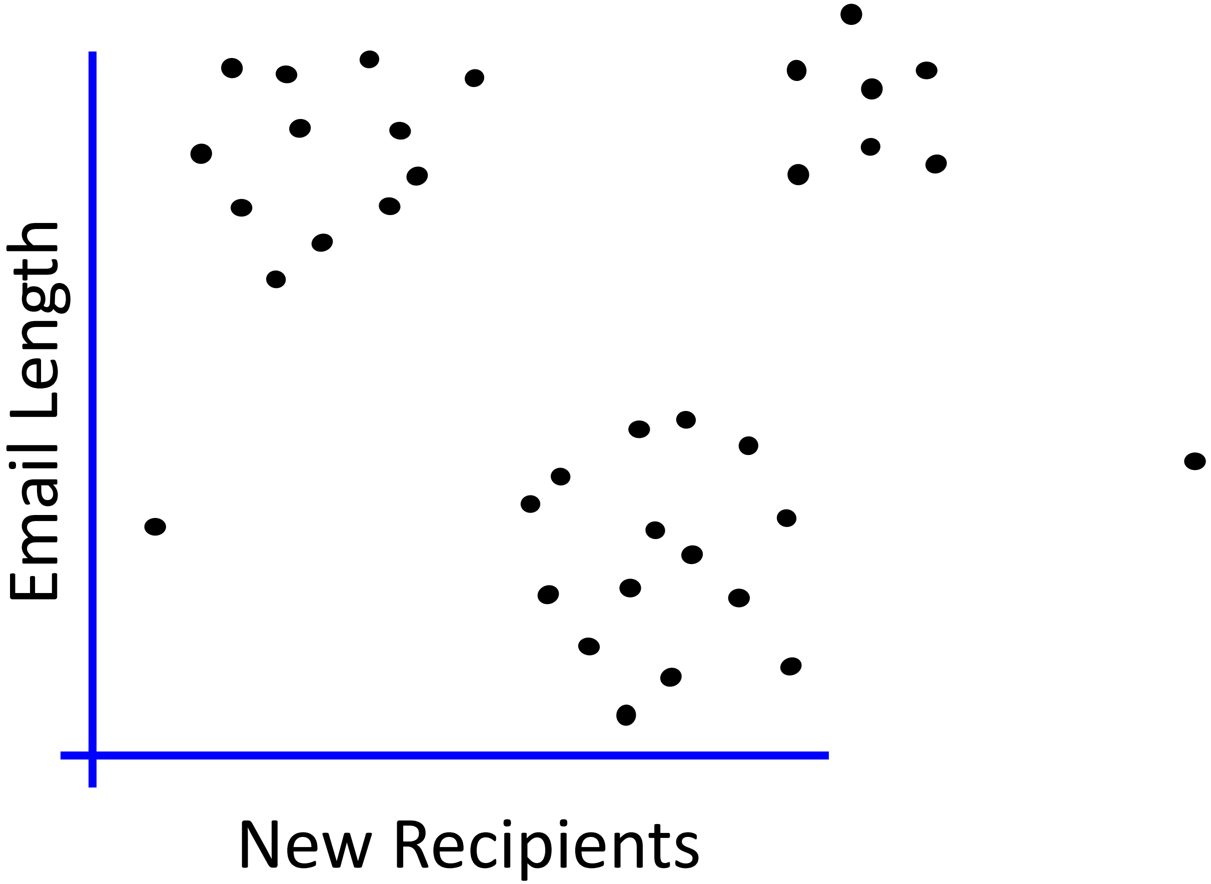
# Unsupervised Learning - Clustering

**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.



# Source of Training Data

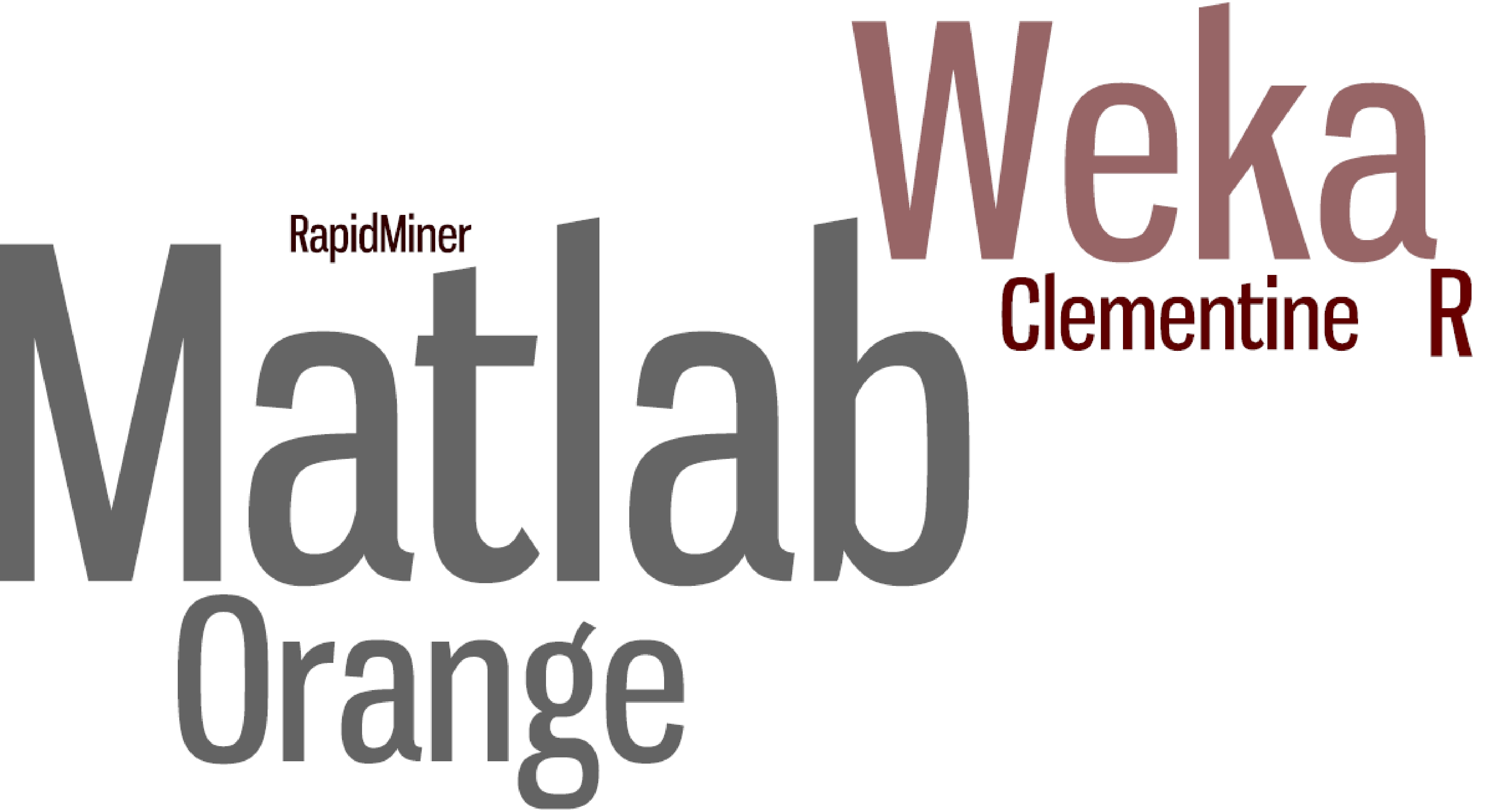
- Provided random examples outside of the learner’s control.
  - Passive Learning
  - Negative examples available or only positive? Semi-Supervised Learning
  - Imbalanced
- Good training examples selected by a “benevolent teacher.”
  - “Near miss” examples
- Learner can query an oracle about class of an unlabeled example in the environment.
  - Active Learning
- Learner can construct an arbitrary example and query an oracle for its label.
- Learner can  run directly in the environment without any human guidance and obtain feedback.
  - Reinforcement Learning
- There is no existing class concept
  - A form of discovery
  - Unsupervised Learning
    - Clustering
    - Association Rules

    -

# Other Learning Tasks

- **Other Supervised Learning Settings**
  - Multi-Class Classification
  - Multi-Label Classification
  - Semi-supervised classification – make use of labeled and unlabeled data
  - One Class Classification – only instances from one label are given
- **Ranking and Preference Learning**
- **Sequence labeling**
- **Cost-sensitive Learning**
- **Online learning  and Incremental Learning-  Learns one instance at a time.**
- **Concept Drift**
- **Multi-Task and Transfer Learning**
- **Collective classification – When instances are dependent!**

# Software



Weka

RapidMiner

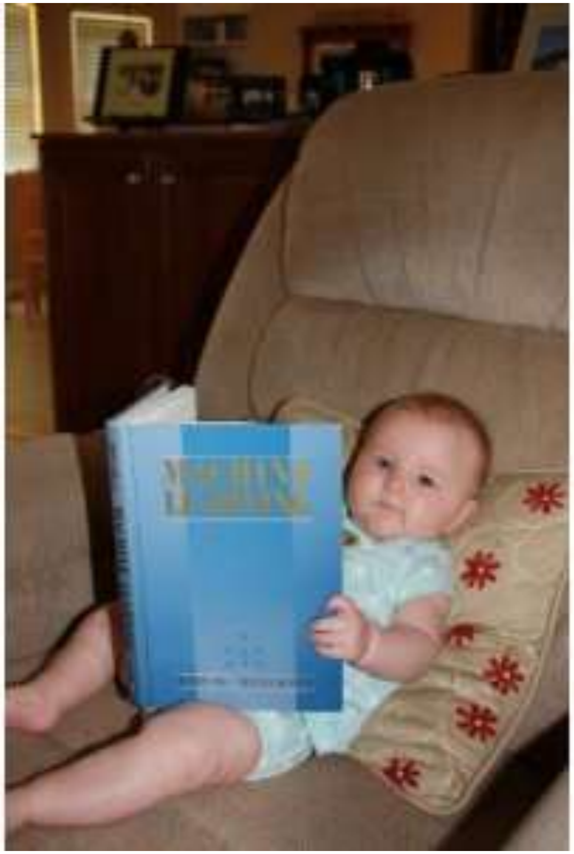
Matlab

Clementine

R

Orange

# Want to Learn More?

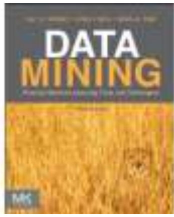


# Want to Learn More?

- Thomas Mitchell (1997), Machine Learning, Mcgraw-Hill.

- *R. Duda, P. Hart, and D. Stork (2000),* Pattern Classification, Second Edition.

- *Ian H. Witten, Eibe Frank, Mark A. Hall (2011), Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, The Morgan Kaufmann*



- *Oded Maimon, Lior Rokach (2010),* Data Mining and Knowledge Discovery Handbook, Second Edition, Springer.