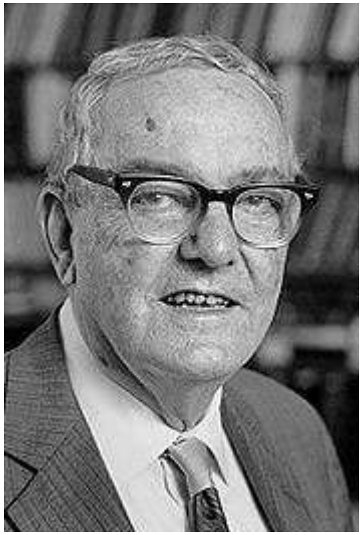
Machine Learning

- **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

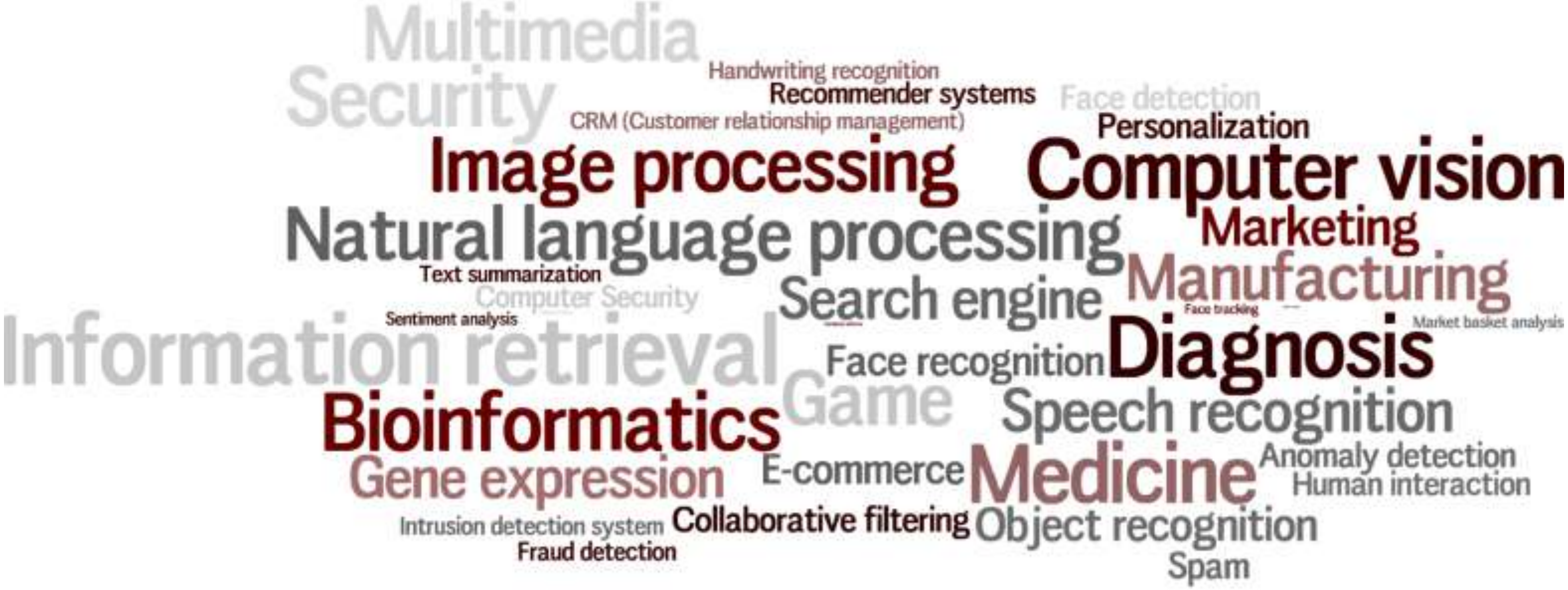
Why Machine Learning?

- 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

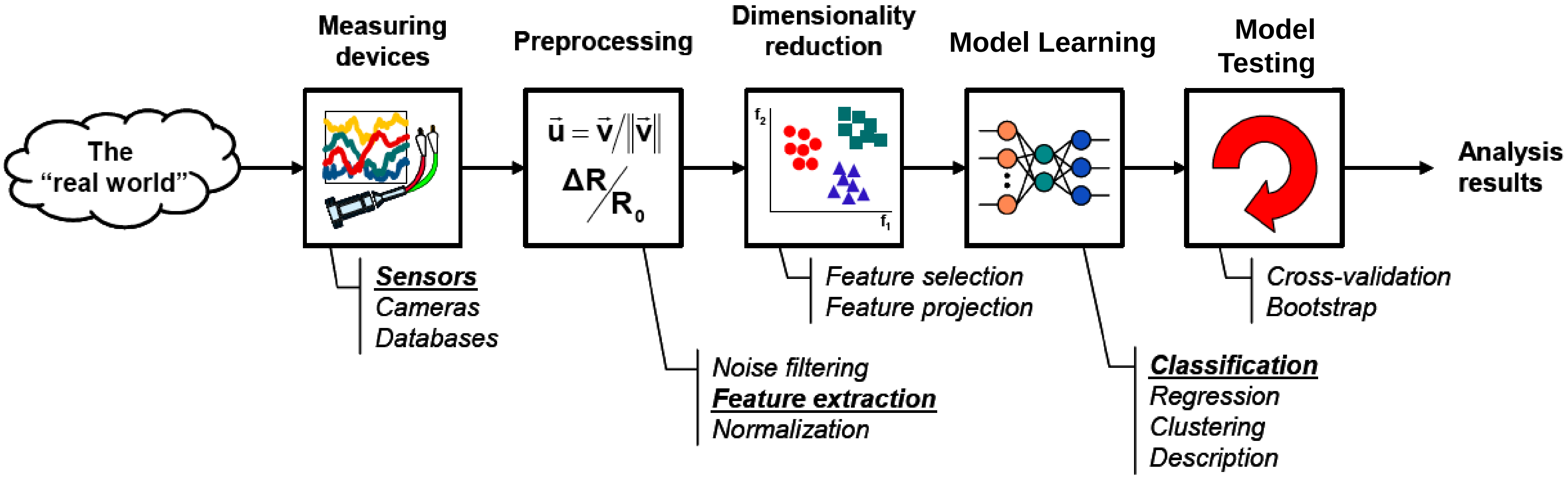
P:

% 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

The Learning Process



The "real world"

Measuring devices

<u>Sensors</u>
Cameras
Databases

Preprocessing

Noise filtering
<u>Feature extraction</u>
Normalization

Dimensionality reduction

Feature selection
Feature projection

Model Learning

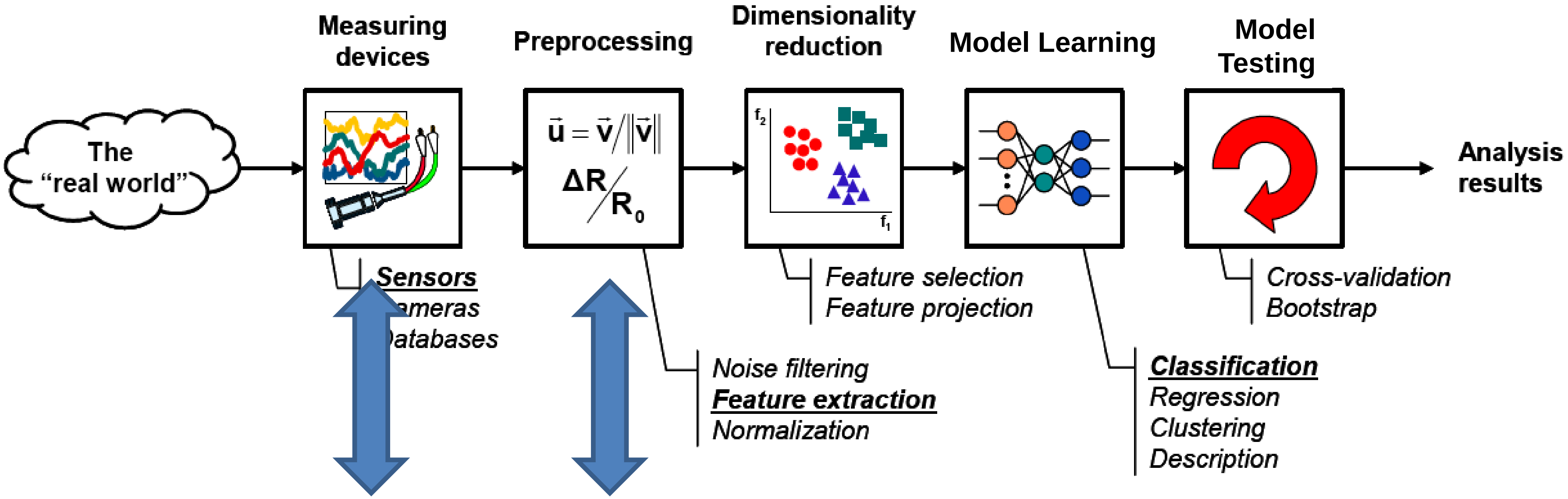
<u>Classification</u>
Regression
Clustering
Description

Model Testing

Cross-validation
Bootstrap

Analysis results

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

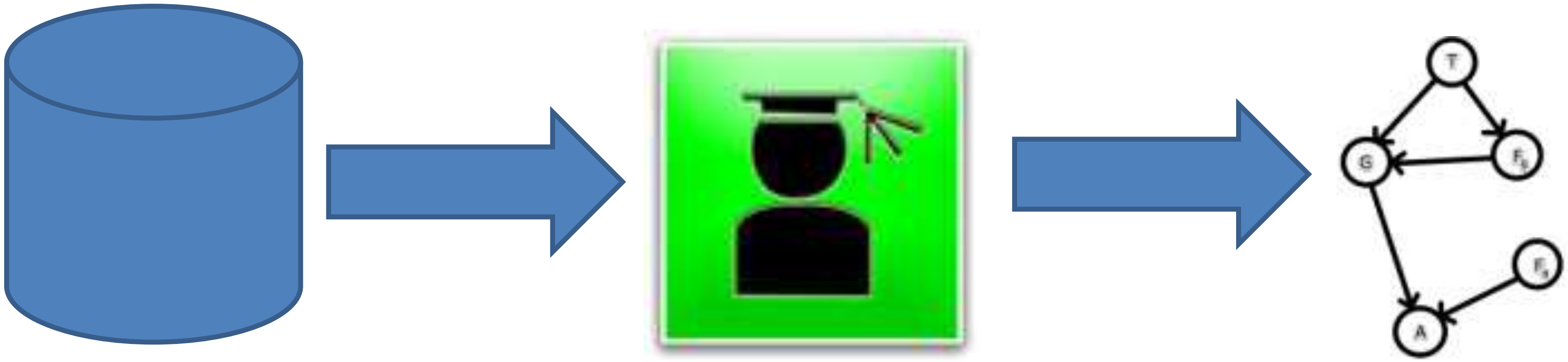
Instances

Numeric

Nominal

Ordinal

Step 4: Model Learning

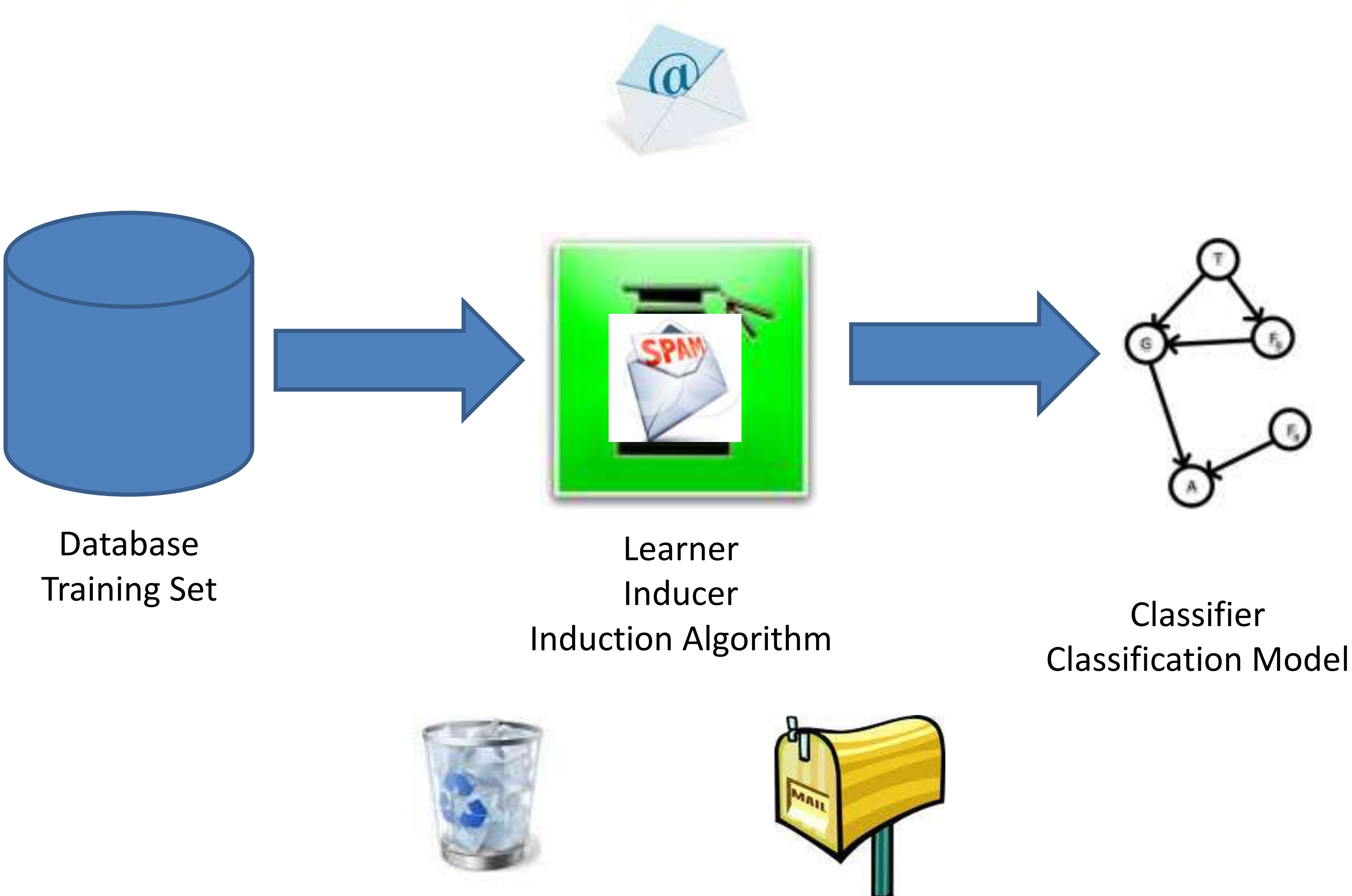


Database
Training Set

Learner
Inducer
Induction Algorithm

Classifier
Classification Model

Step 5: Model Testing

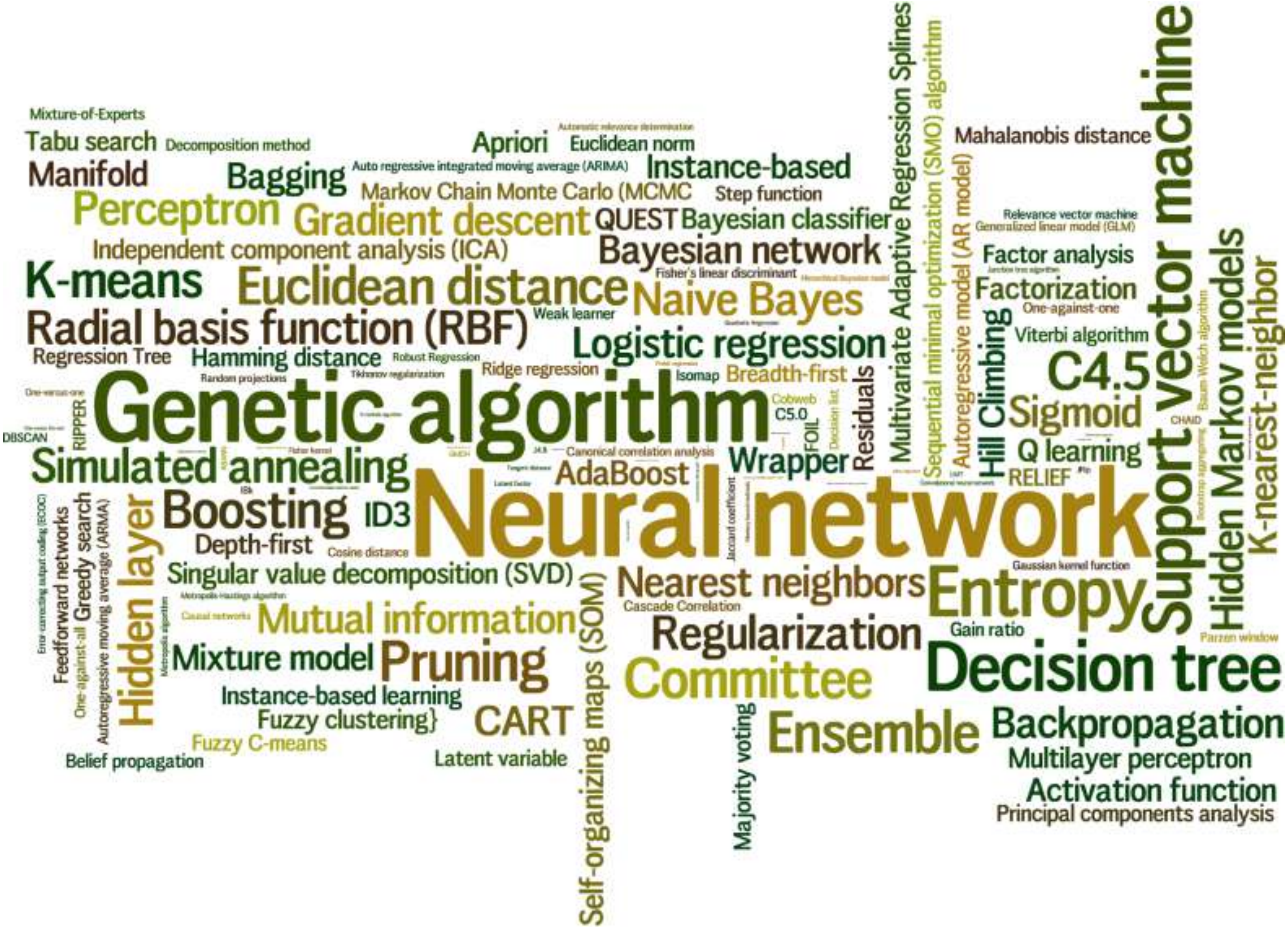


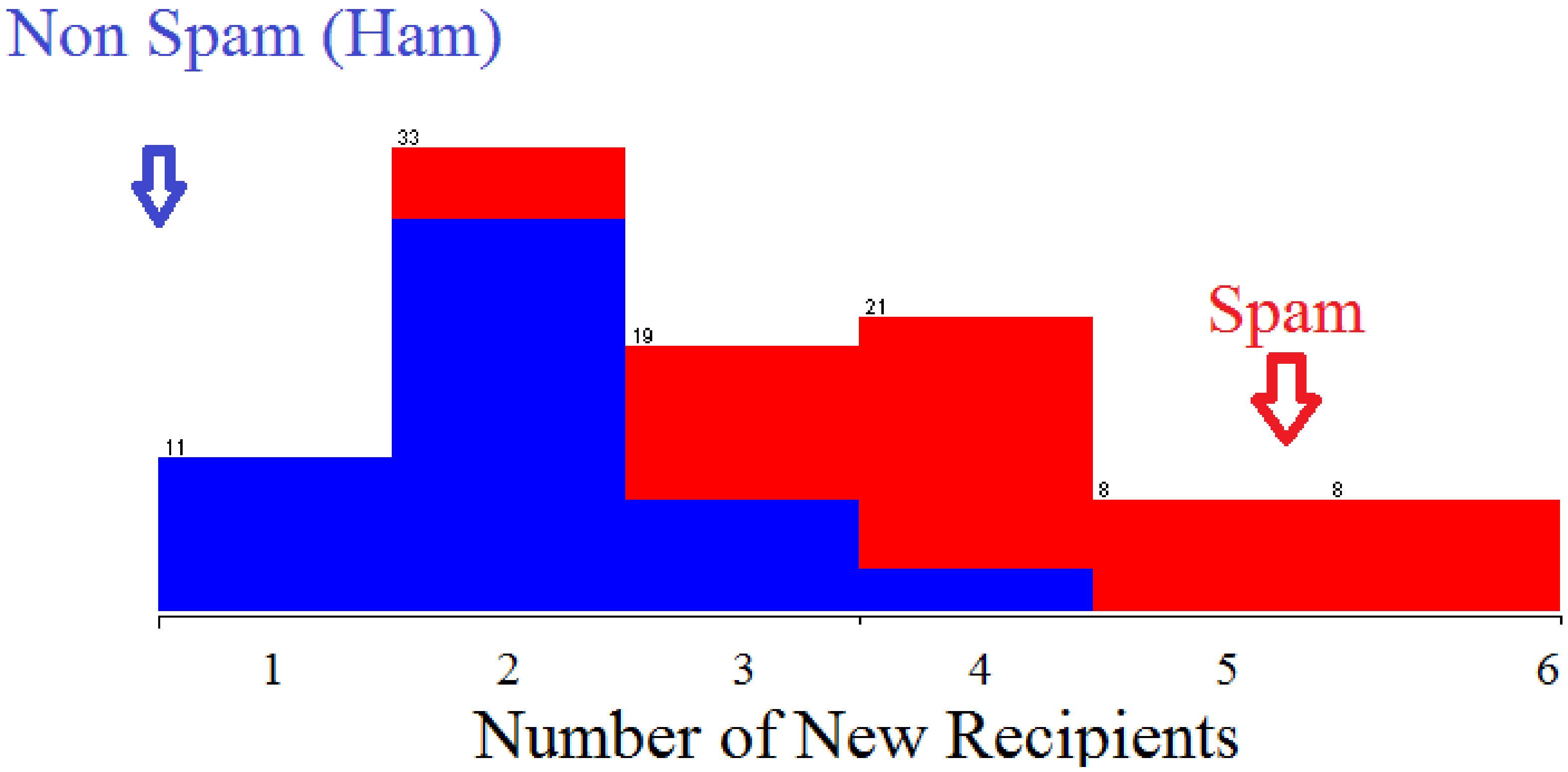
Database
Training Set

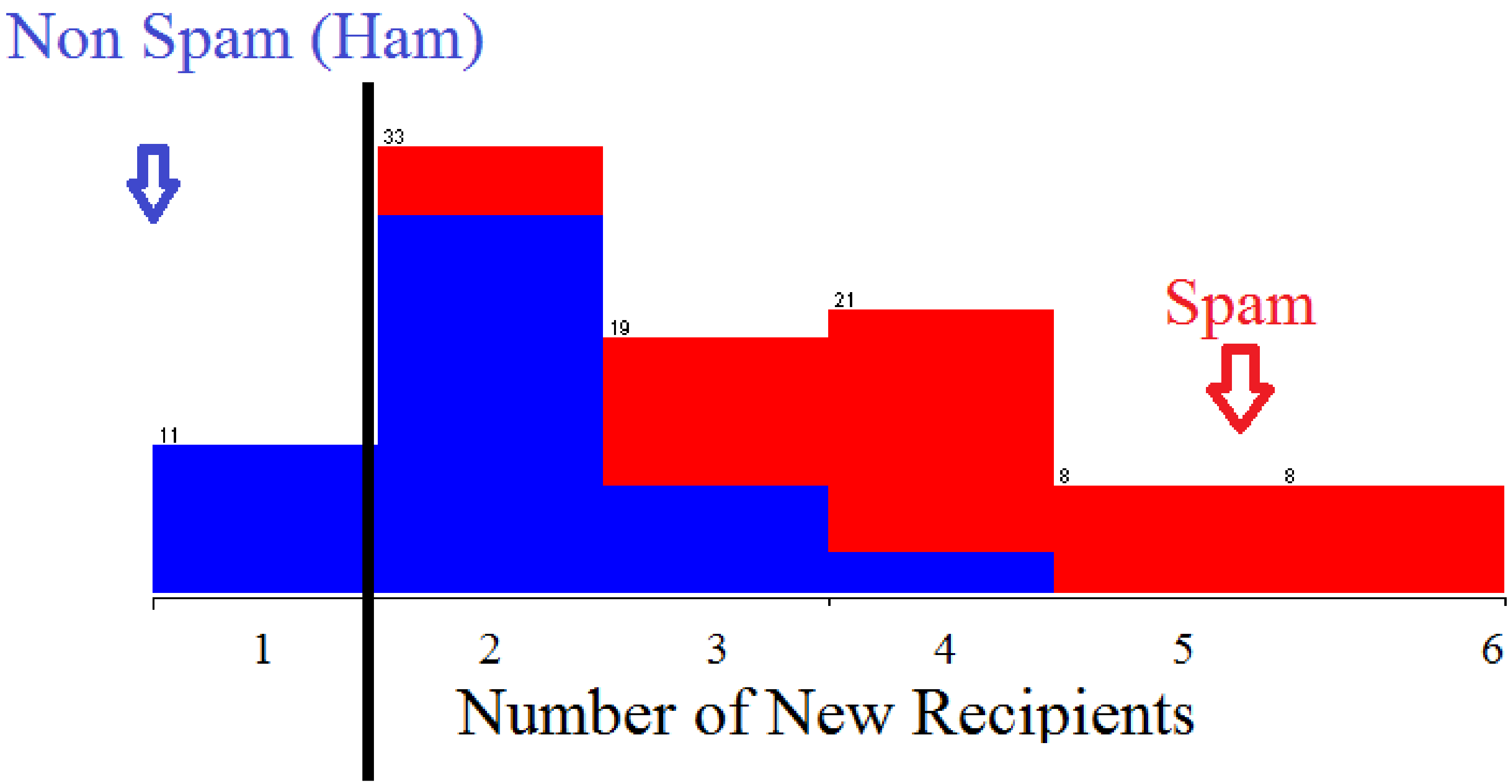
Learner
Inducer
Induction Algorithm

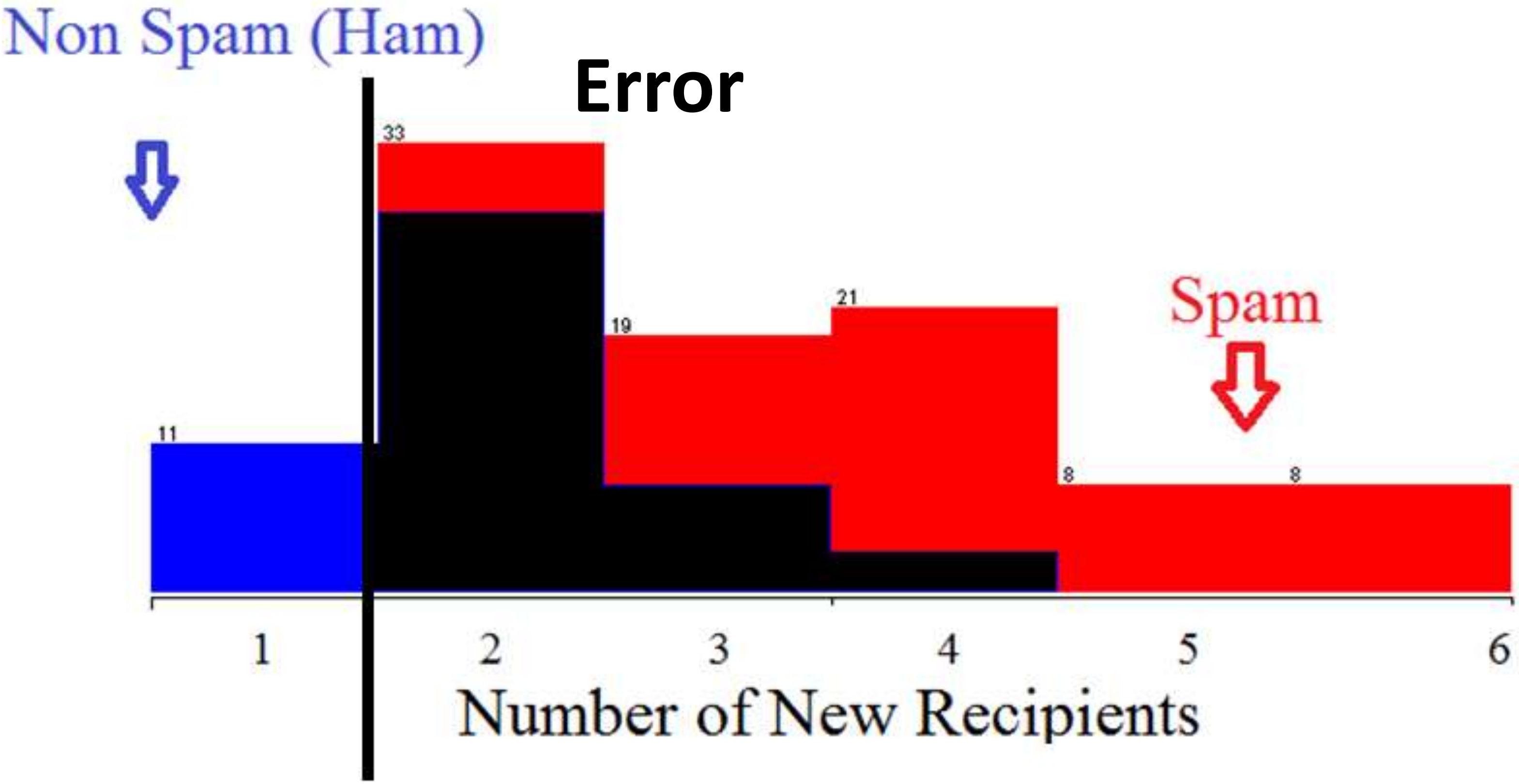
Classifier
Classification Model

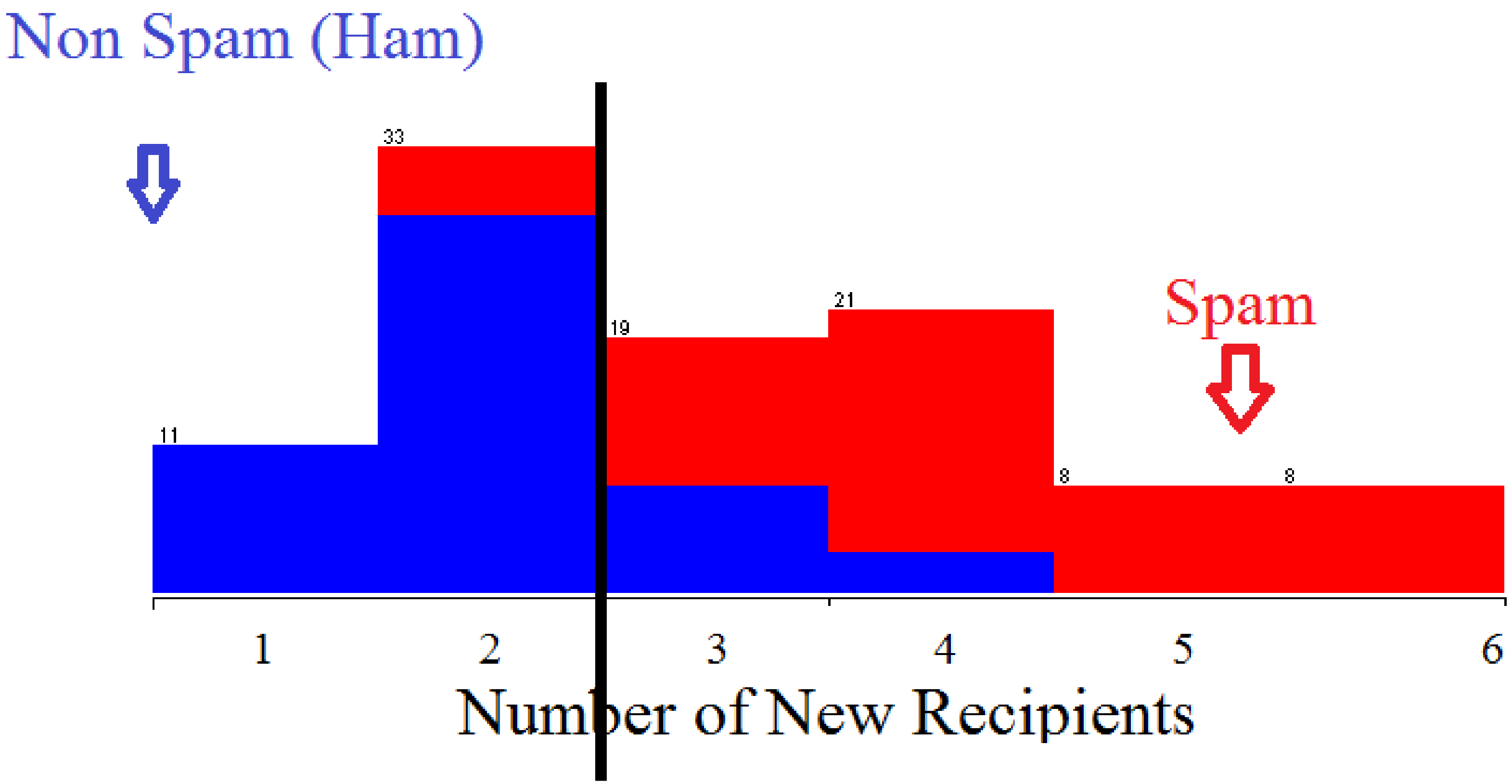
Learning Algorithms

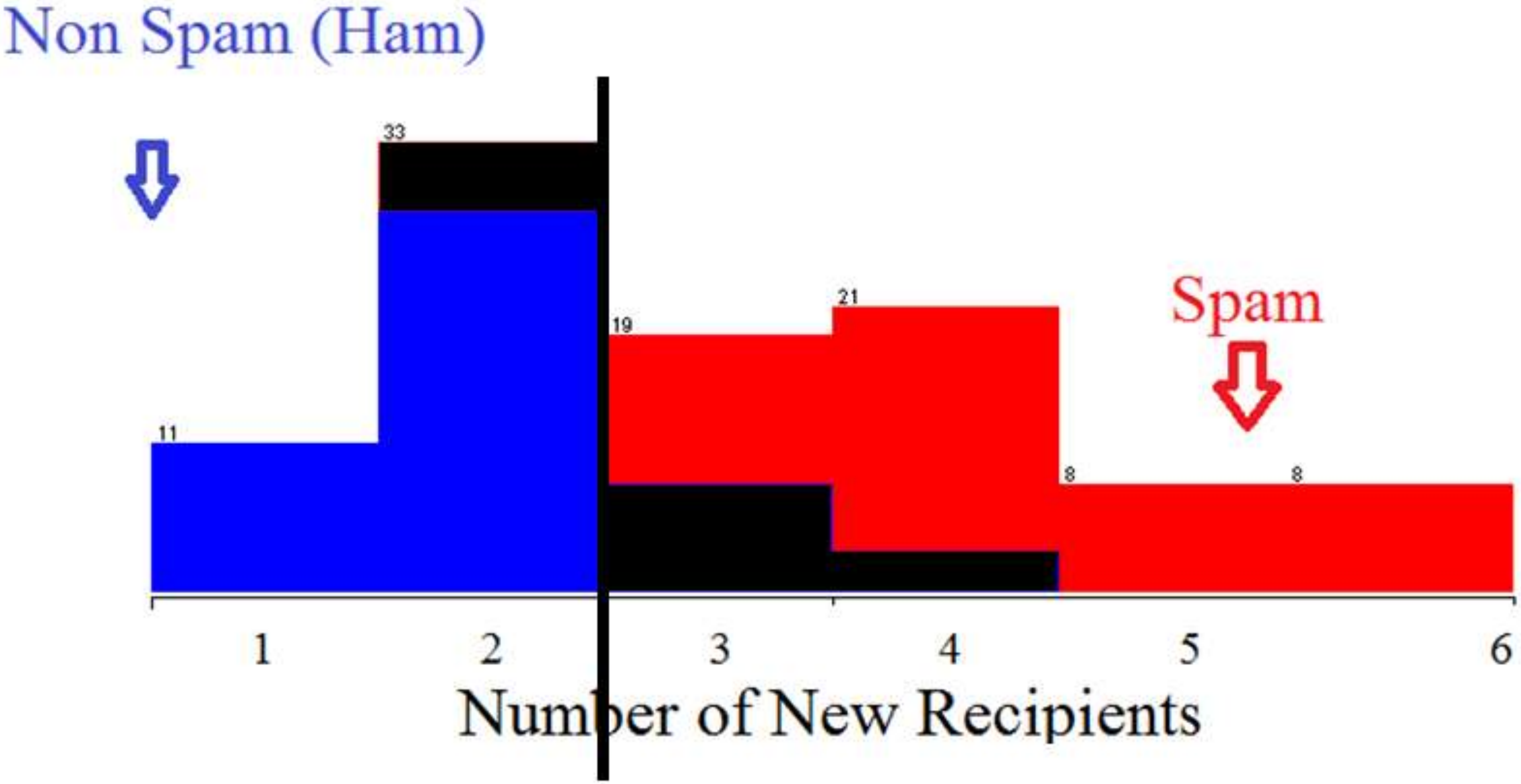


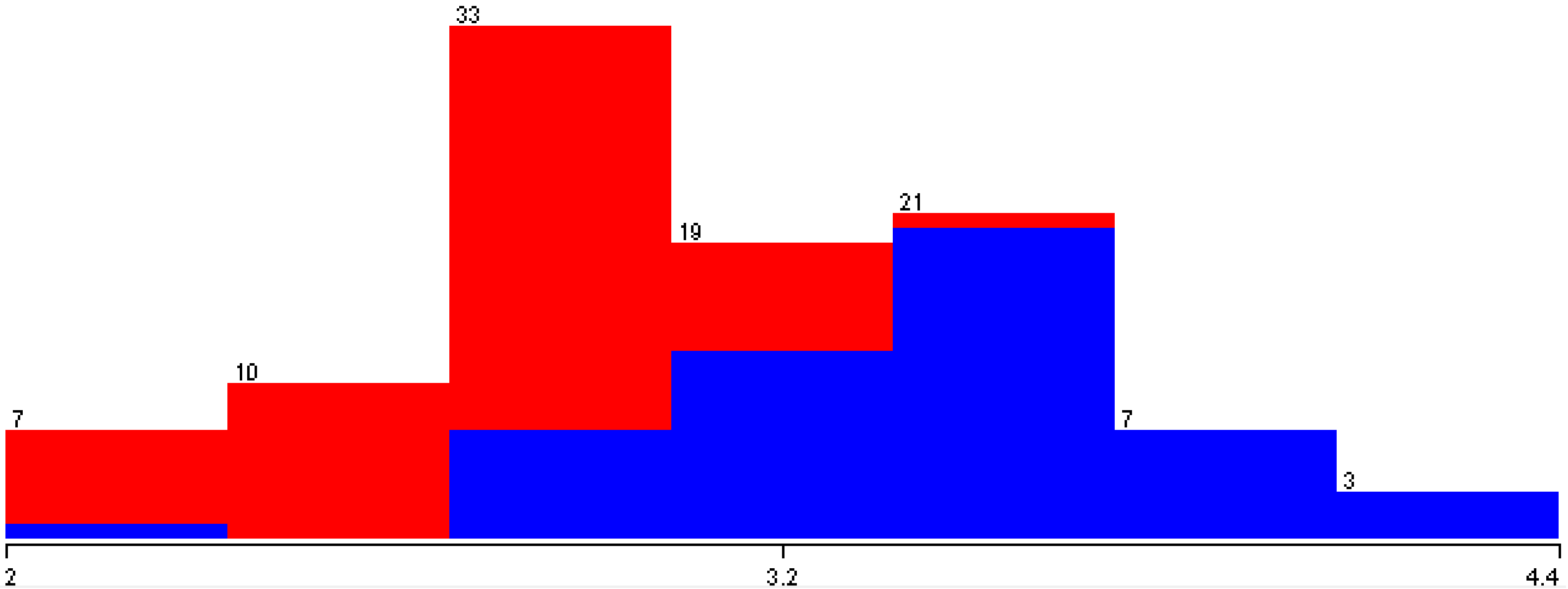

Non Spam (Ham)
33
19
21
Spam
11
8
8
1
2
3
4
5
6
Number of New Recipients


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


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

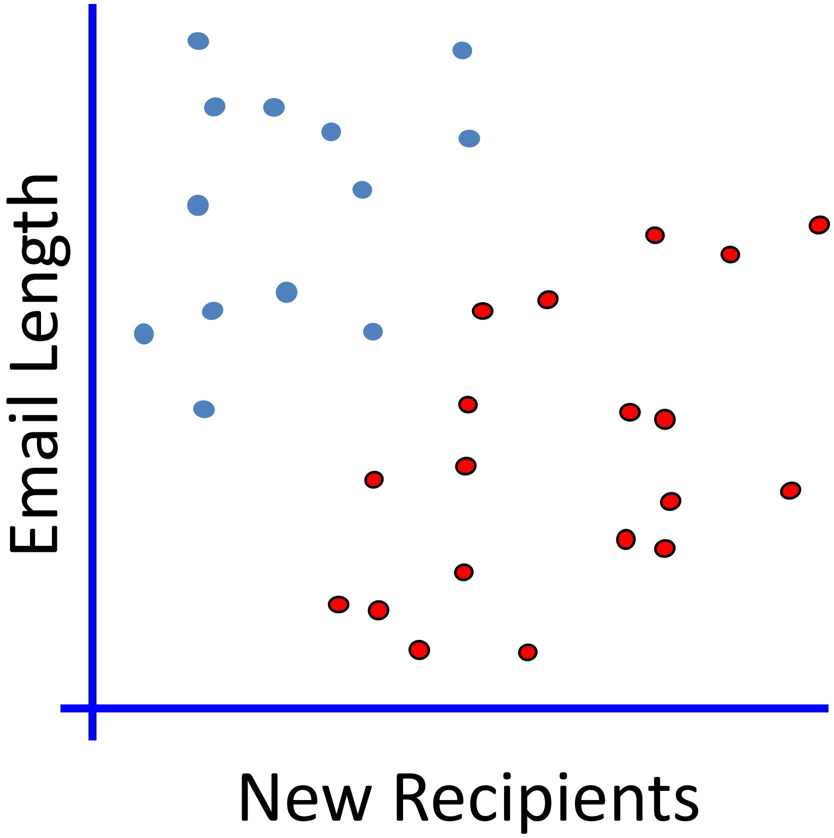

Non Spam (Ham)
33
19
21
Spam
11
8
8
1
2
3
4
5
6
Number of New Recipients


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



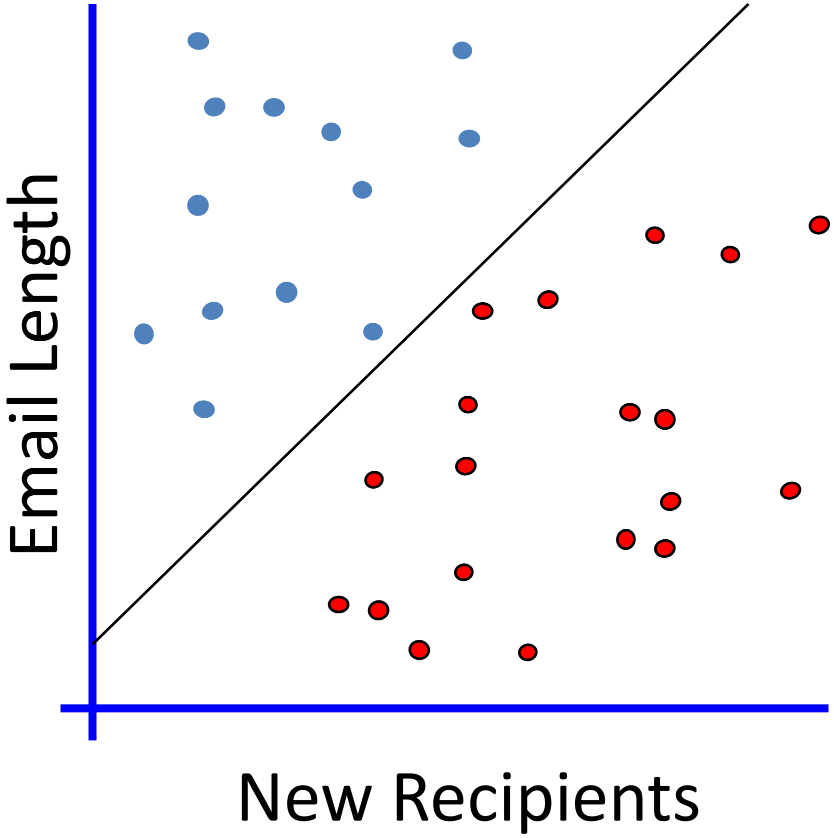
Email Length

Linear Classifiers



How would you classify this data?

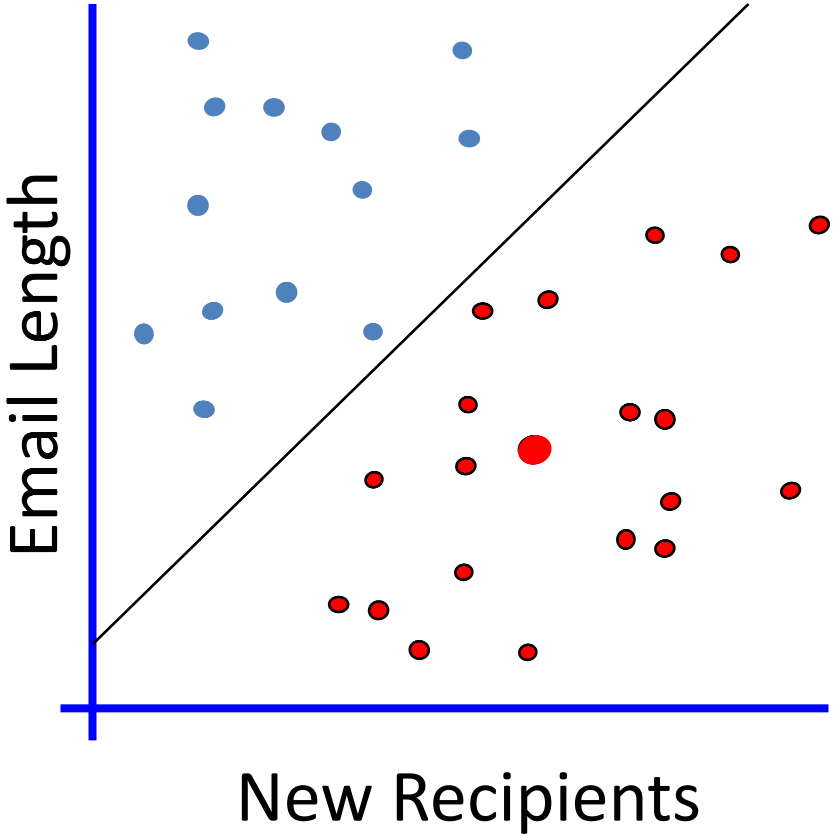
Linear Classifiers



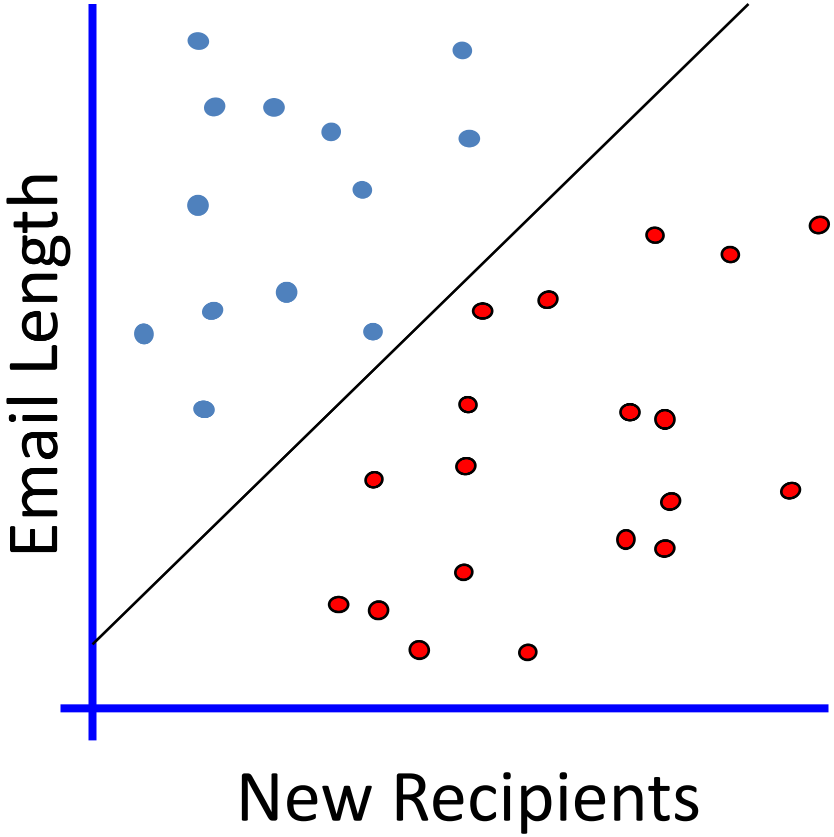
How would you classify this data?

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

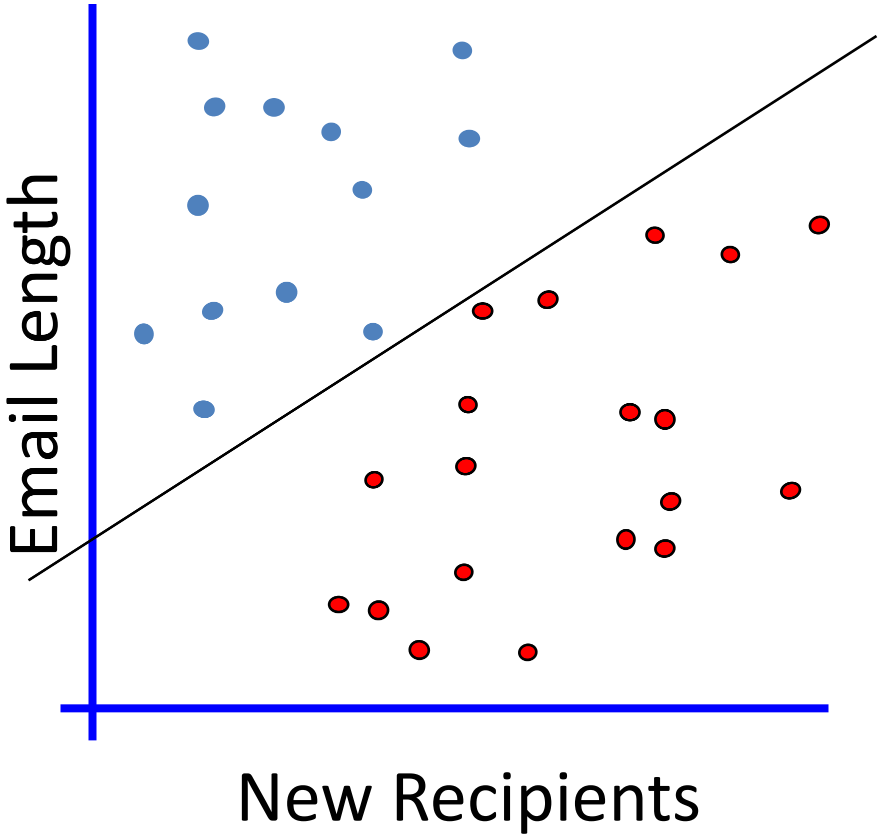


Linear Classifiers



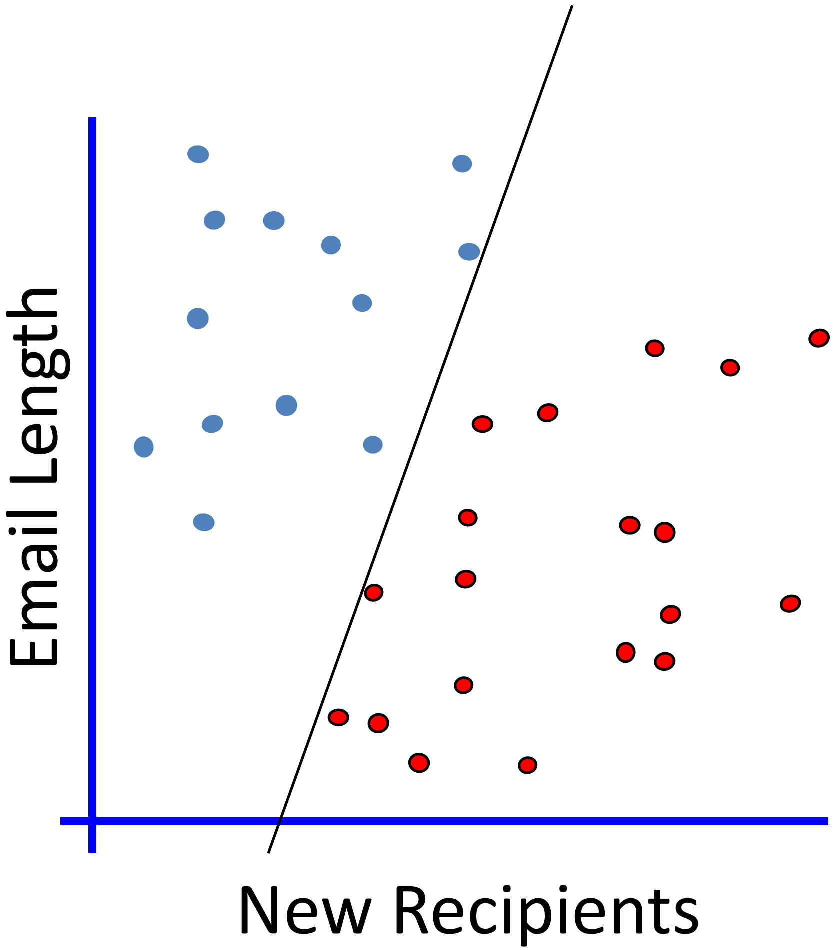
How would you classify this data?

Linear Classifiers



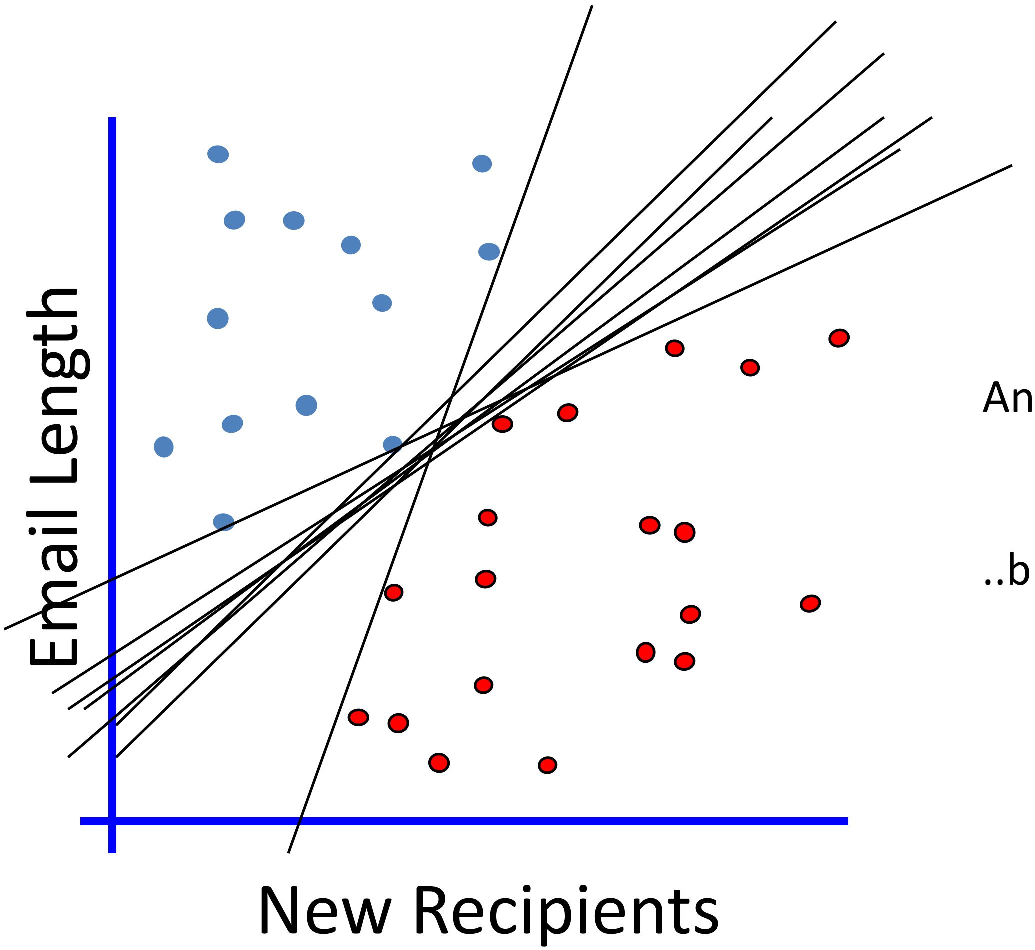
How would you classify this data?

Linear Classifiers



How would you
classify this data?

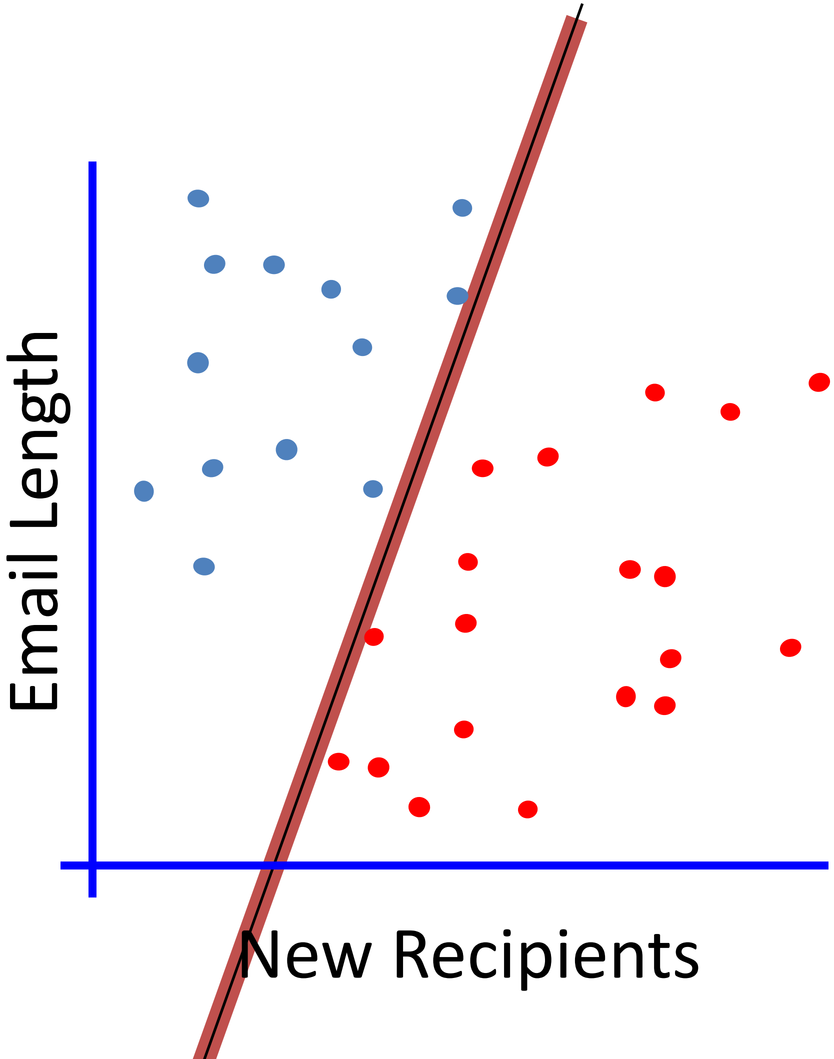
Linear Classifiers



Any of these would be fine..

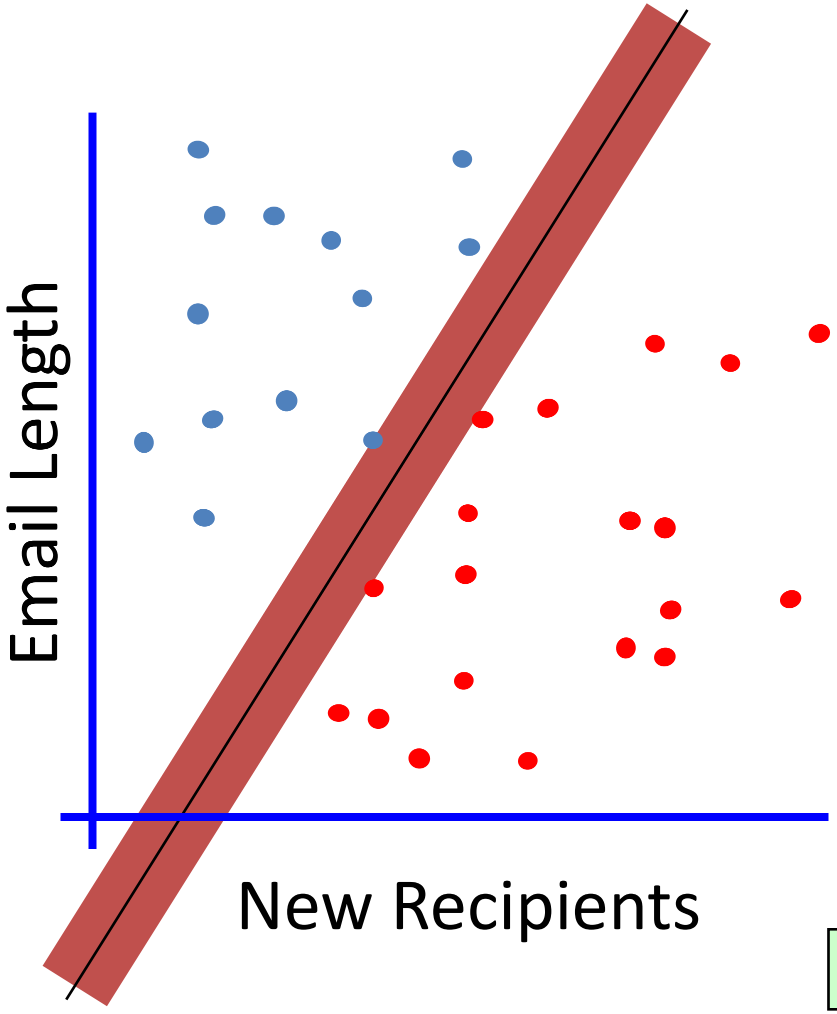
..but which is best?

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

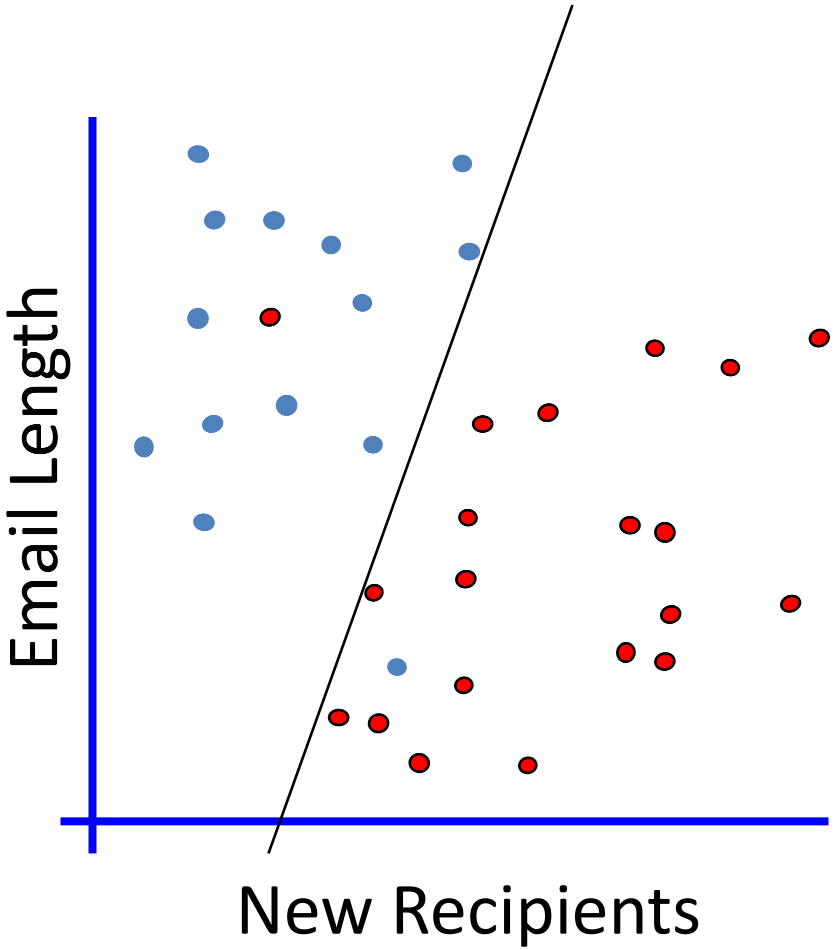


The maximum margin linear classifier is the linear classifier with the, maximum margin.

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

Linear SVM

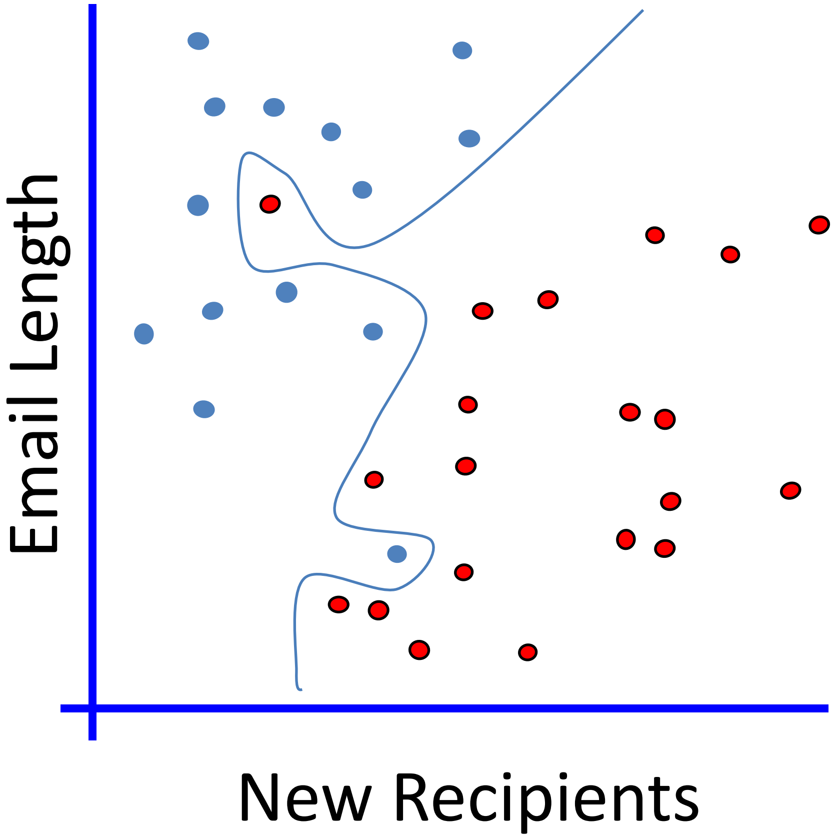
No Linear Classifier can cover all instances



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

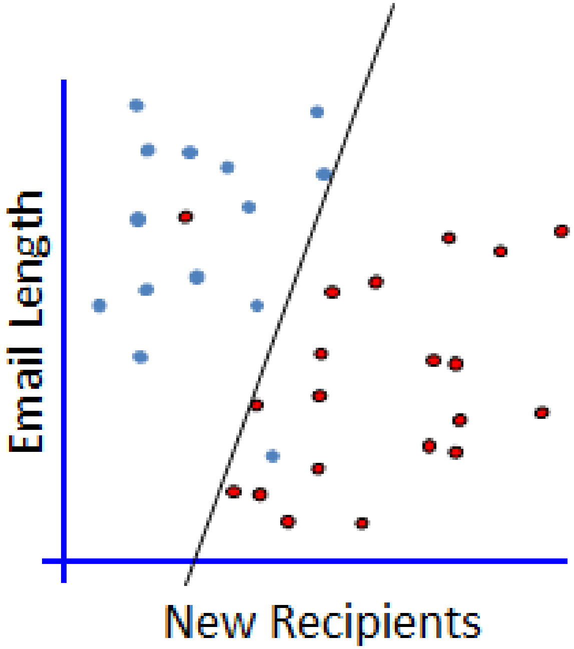


- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

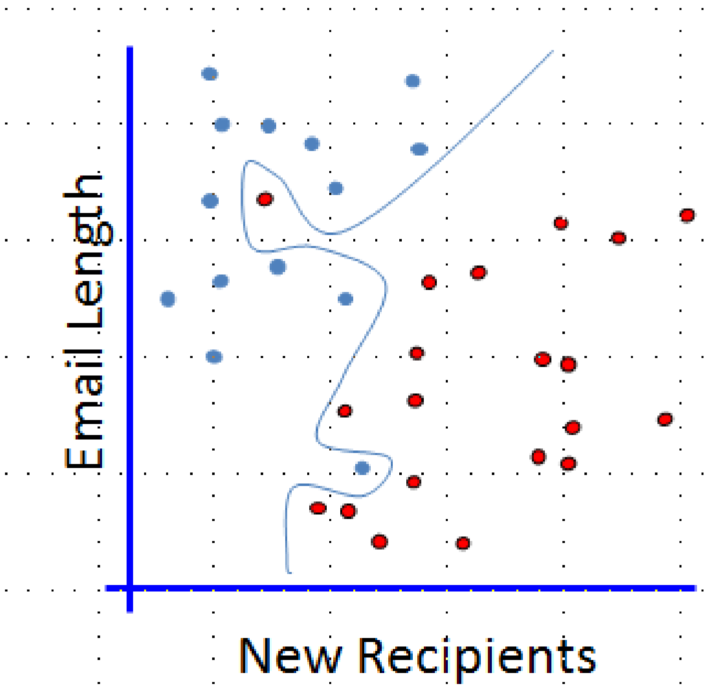
↓

Issue of generalization!

Which one?



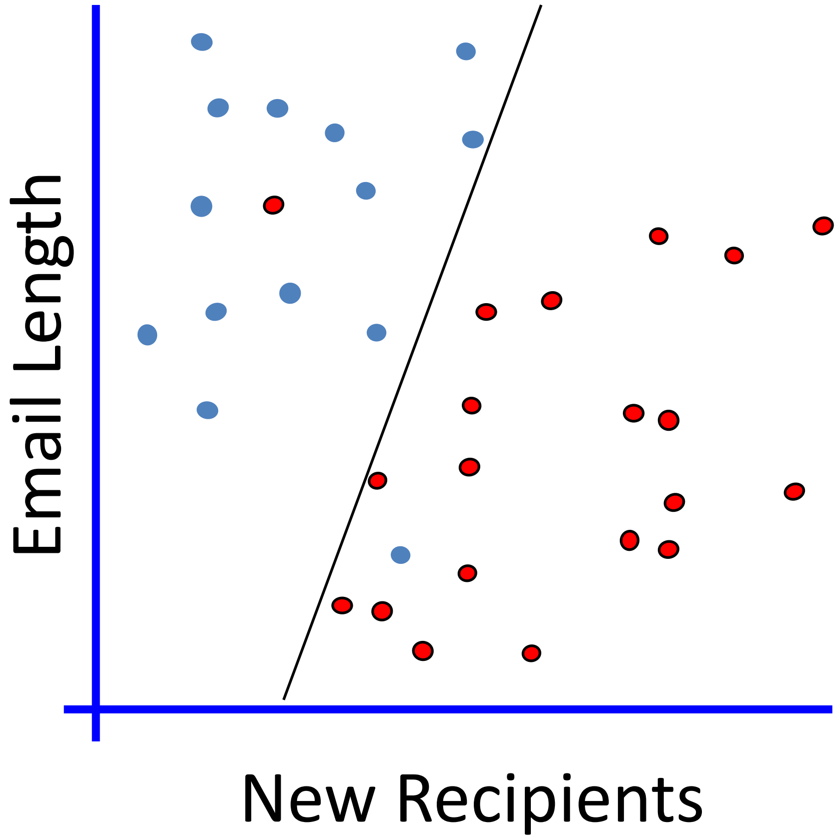
2 Errors
Simple model



0 Errors
Complicated model

Evaluating What’s Been Learned

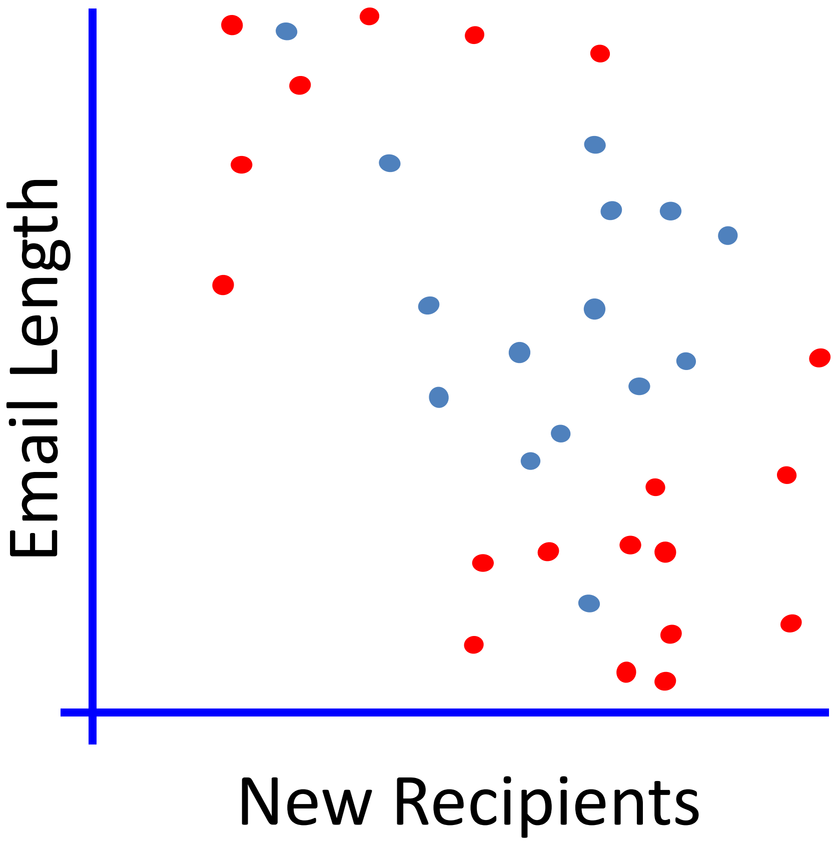
1. We randomly select a portion of the data to be used for training (the training set)
2. Train the model on the training set.
3. Once the model is trained, we run the model on the remaining instances (the test set) to see how it performs



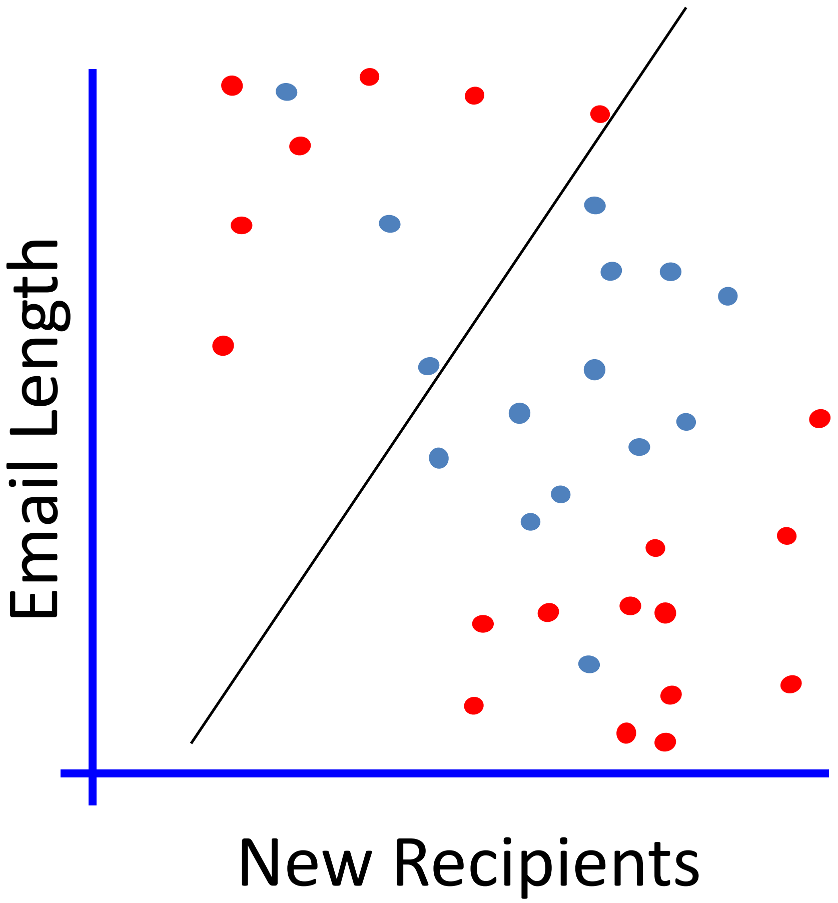
Confusion Matrix

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

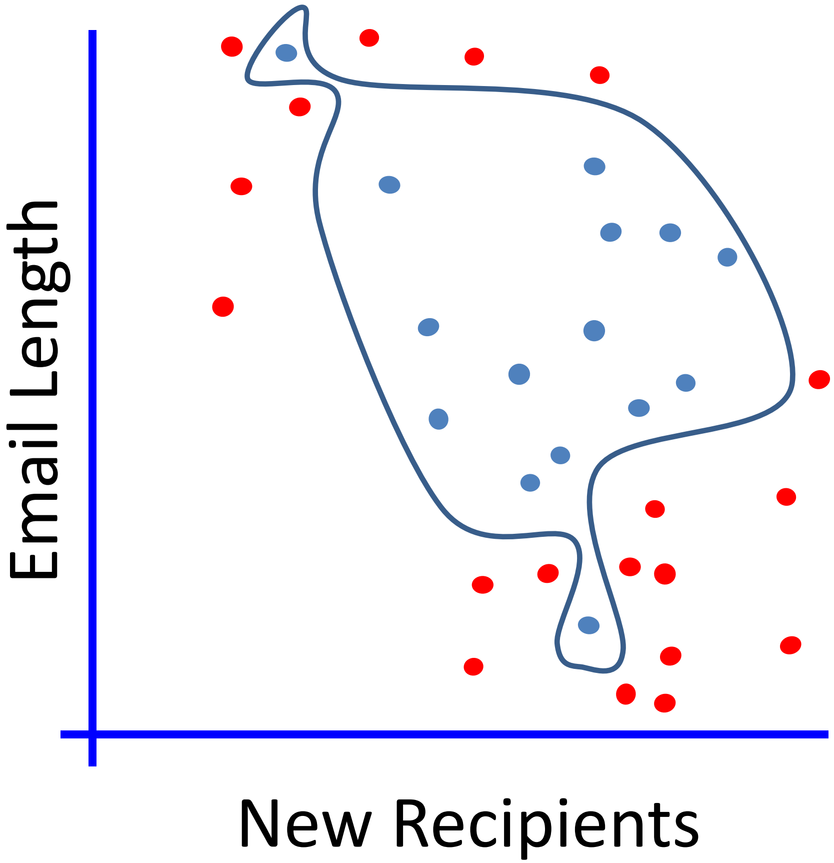
The Non-linearly separable case



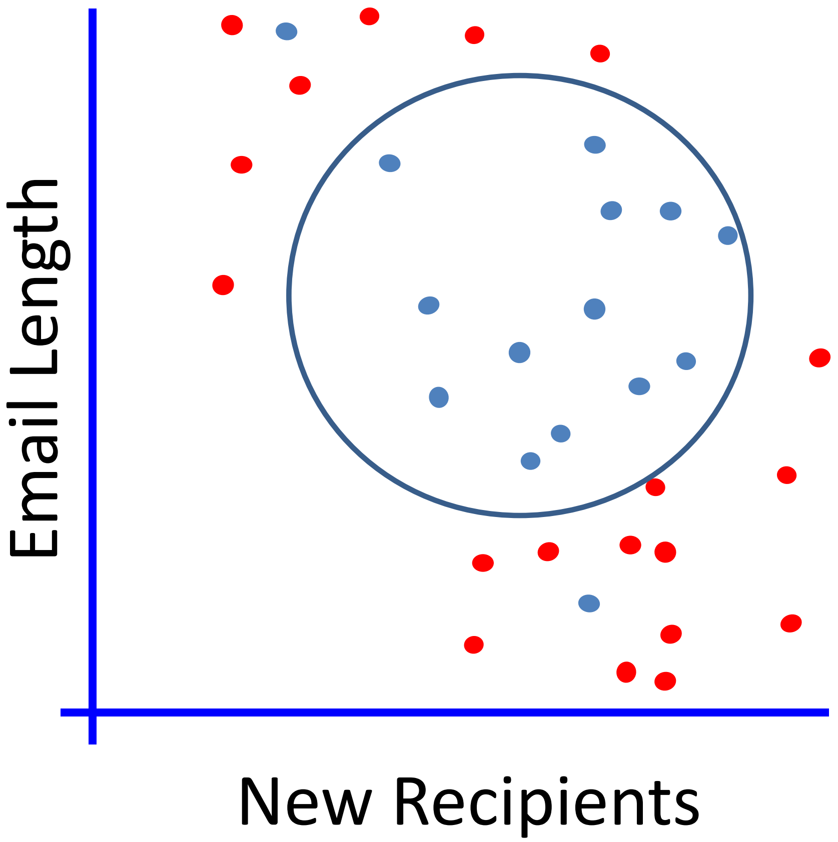
The Non-linearly separable case



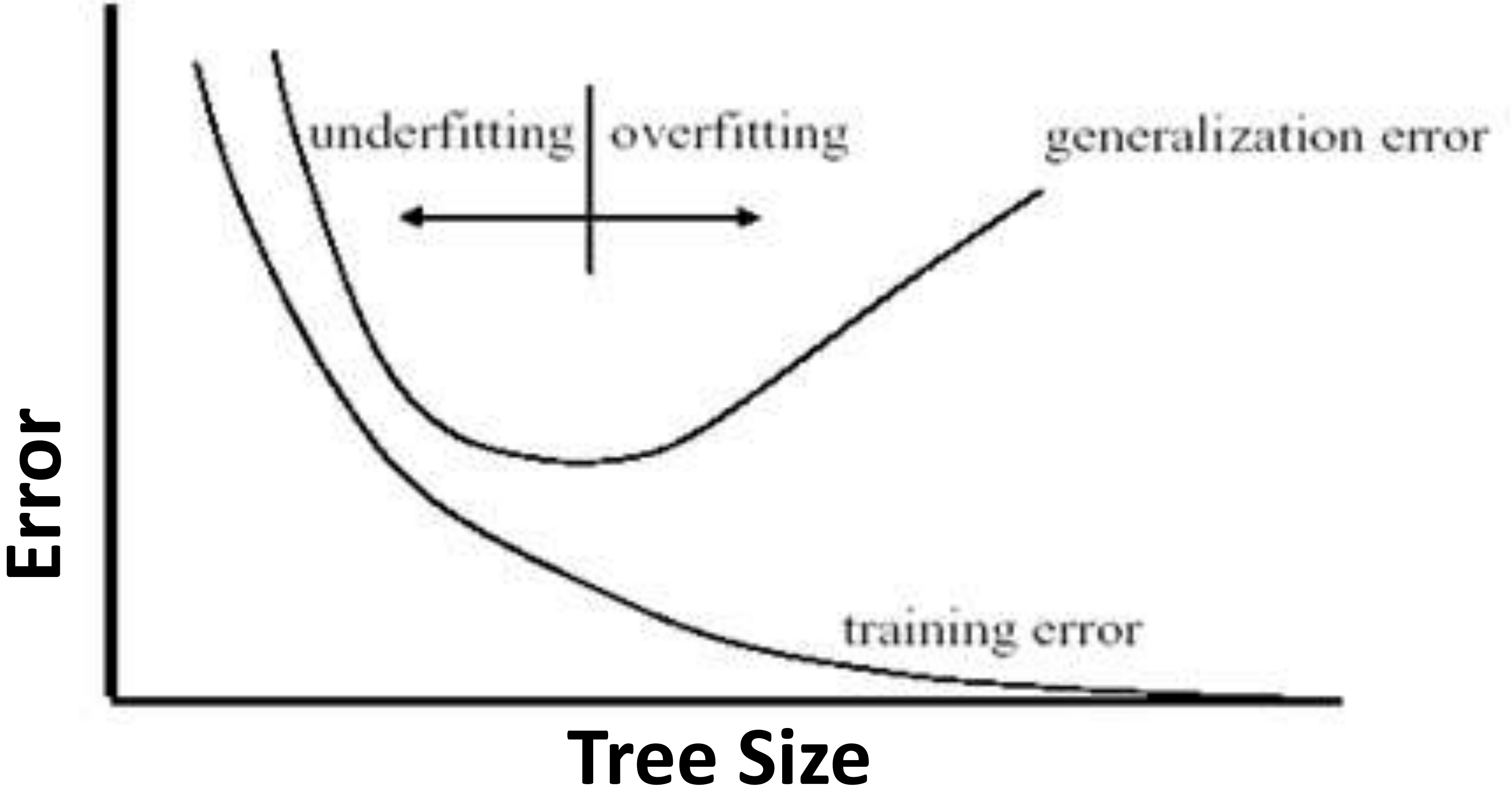
The Non-linearly separable case



The Non-linearly separable case



Overfitting and underfitting



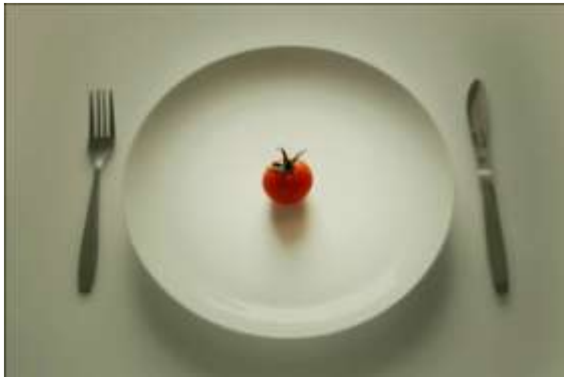
Overtraining: means that it learns the training set too well – it overfits to the training set such that it performs poorly on the test set.

Underfitting: when model is too simple, both training and test errors are large

Main Principles

No Free Lunch Theorem in Machine Learning (Wolpert, 2001)

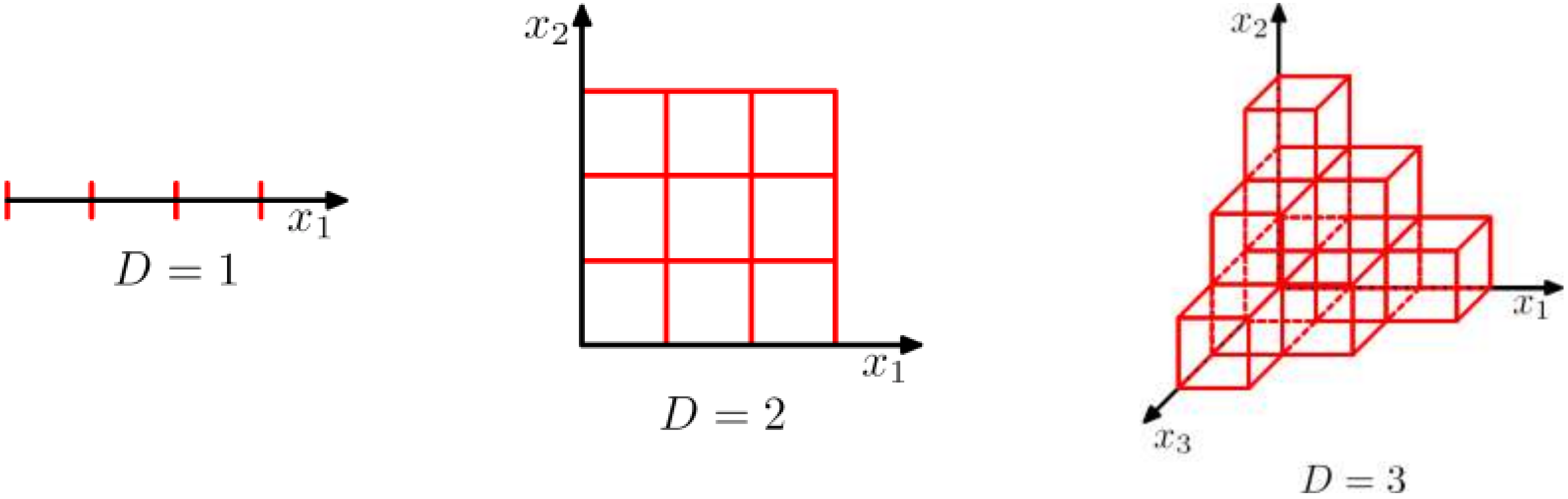
- *“For any two learning algorithms, there are just as many situations (appropriately weighted) in which algorithm one is superior to algorithm two as vice versa, according to any of the measures of "superiority"*

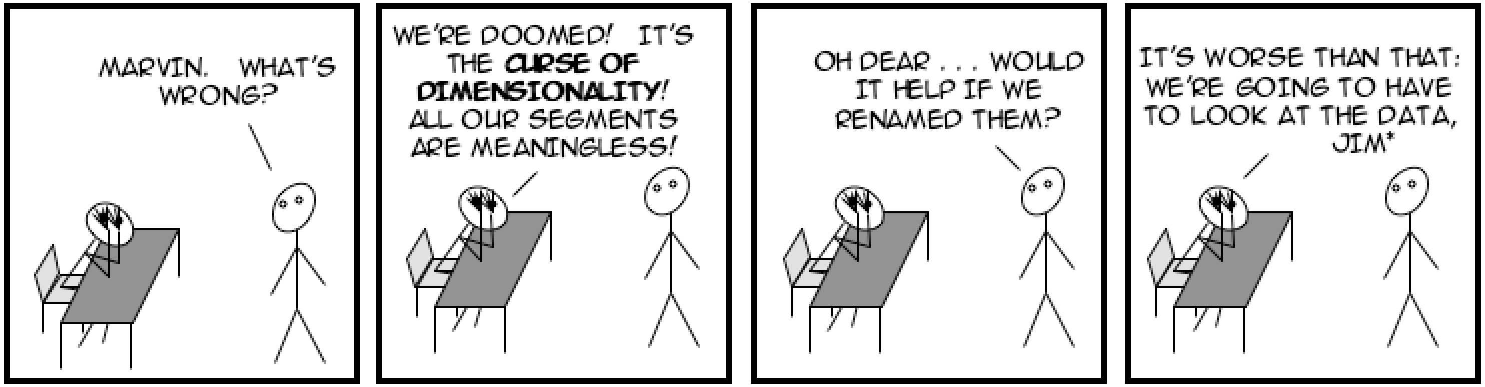


So why developing new algorithms?

- 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)



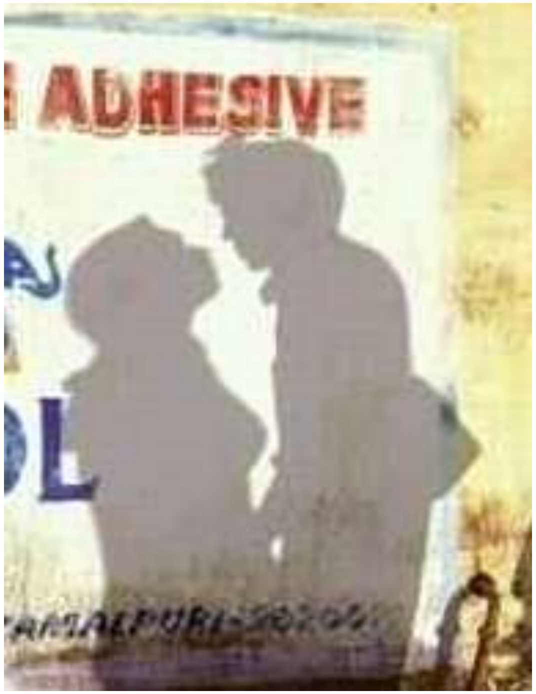

MARVIN. WHAT'S WRONG?
WE'RE DOOMED! IT'S THE CURSE OF DIMENSIONALITY! ALL OUR SEGMENTS ARE MEANINGLESS!
OH DEAR . . . WOULD IT HELP IF WE RENAMED THEM?
IT'S WORSE THAN THAT: WE'RE GOING TO HAVE TO LOOK AT THE DATA, JIM*

HTTP://SCIENTIFICMARKETER.COM

* WITH APOLOGIES TO MR SPOCK & STAR TREK.

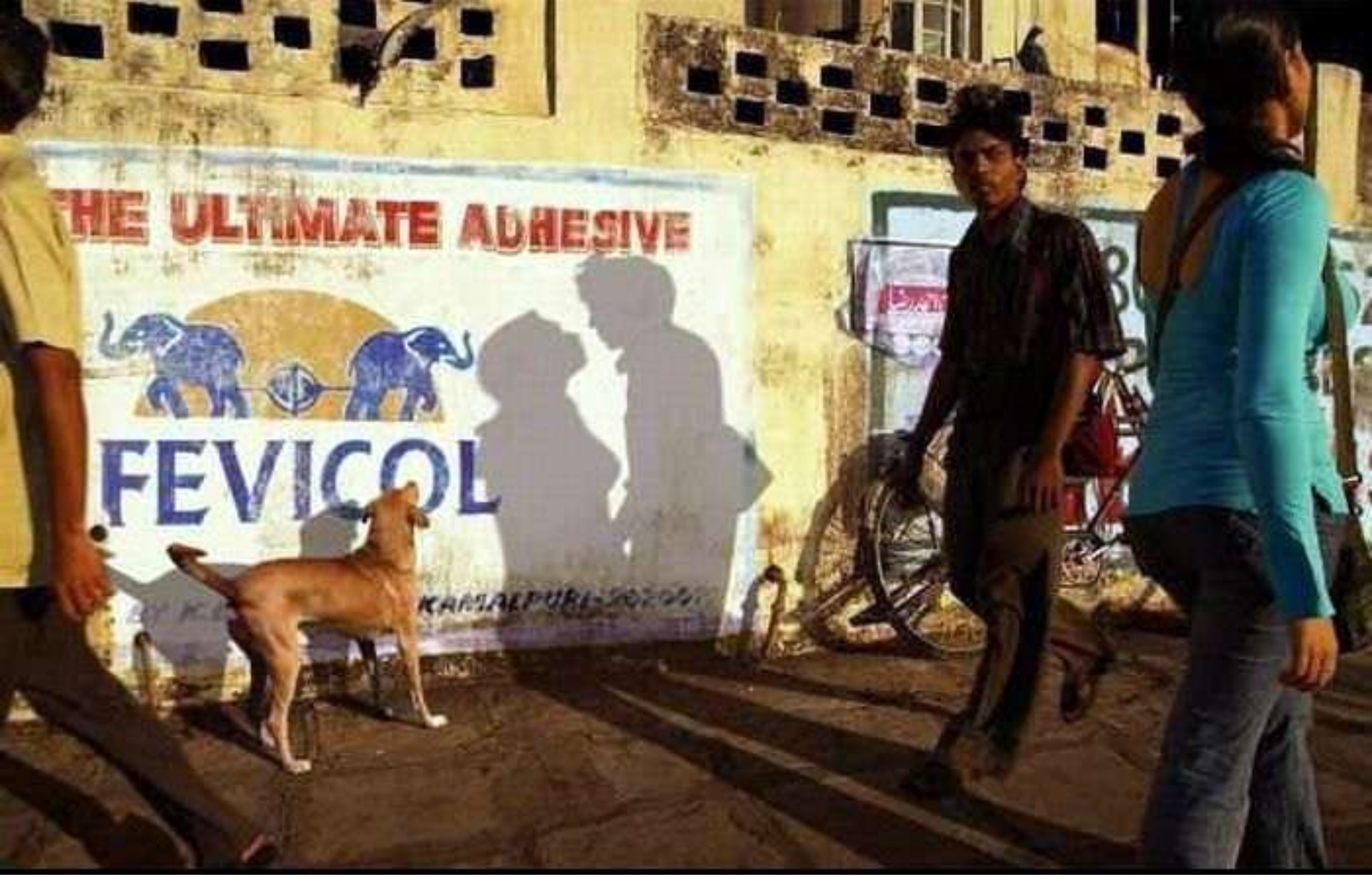
Less is More
The Curse of Dimensionality

- 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….



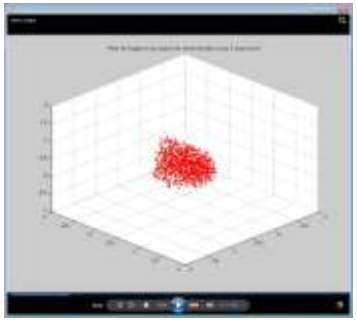
Original photographer unknown

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

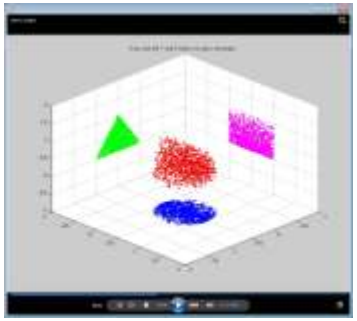
(c) eamonn keogh

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

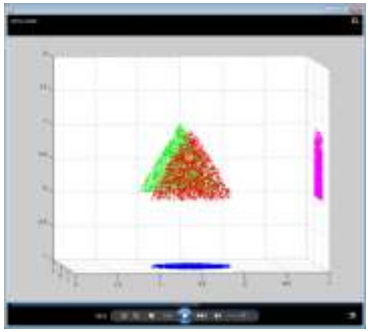
A cloud of points in 3D



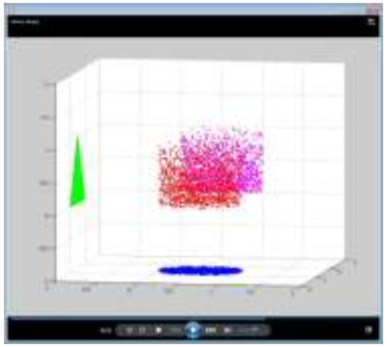
Can be projected into 2D
XY or XZ or YZ



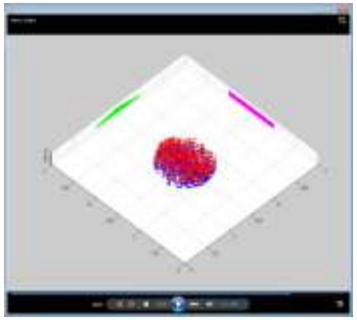
In 2D XZ we see
a triangle



In 2D YZ we see
a square



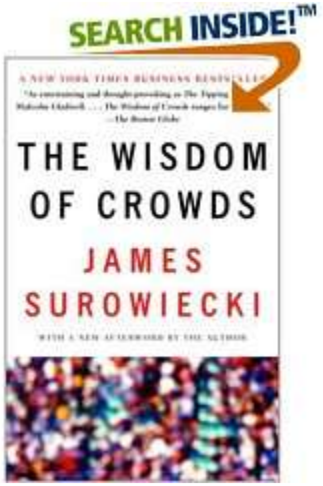
In 2D XY we see
a circle



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"

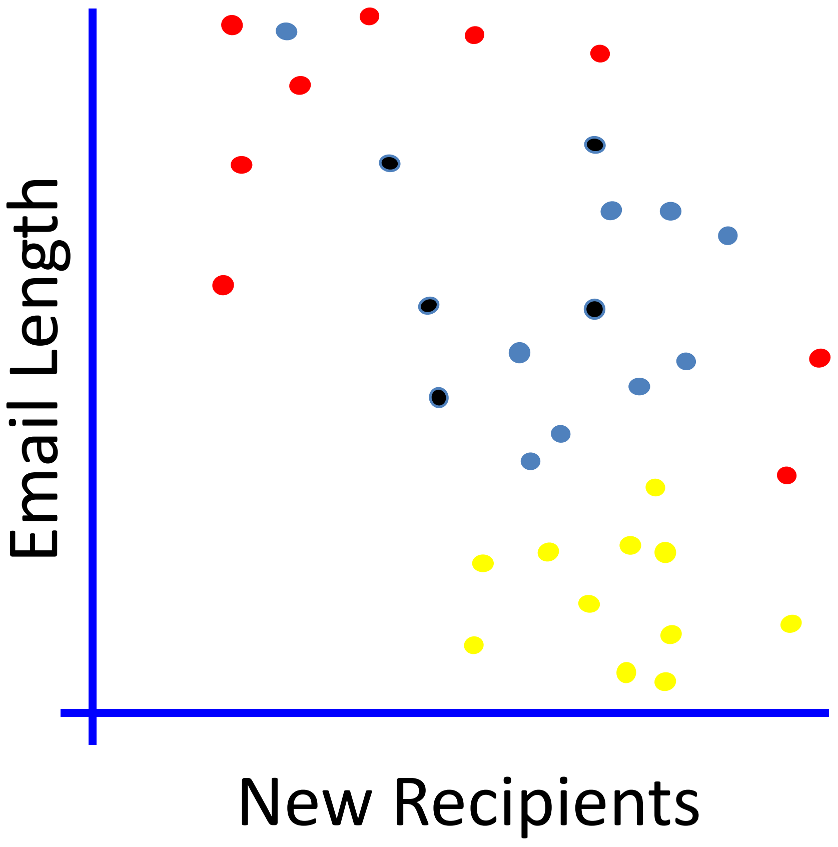
- There is not much point in setting up a committee of experts from such a group - such a committee will not improve on the judgment of an individual.

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

Other Learning Tasks

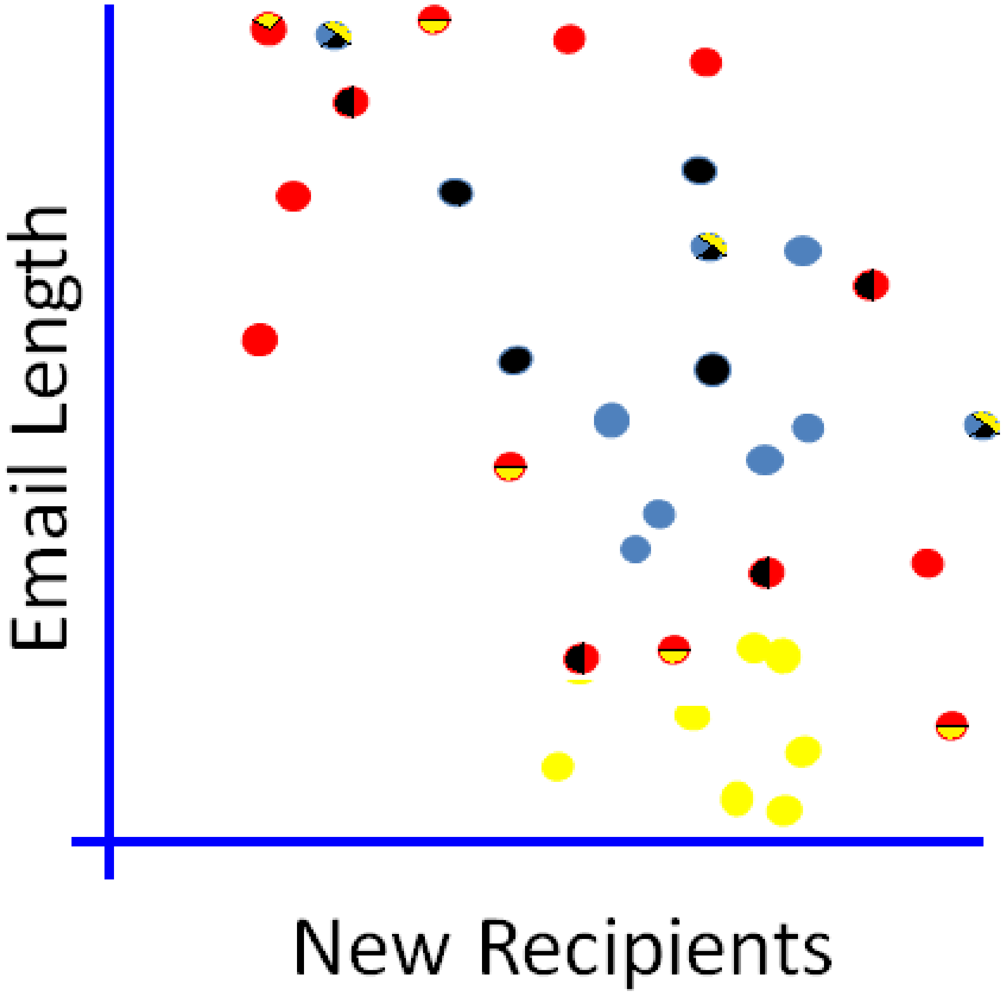
Classification
Regression
Text mining
Sparsity
Linear model
Sequence learning
Kernel
Classification rules
Link analysis
Multi-label
Transduction
Batch learning
Explanation-based learning
Exploitation
Sequential learning
Meta learning
Supervised learning
Concept learning
Eager lerner
Text categorization
Transfer Learning
Relational learning
Learning to rank
Sequence mining
Online learning
Unsupervised learning
Linear regression
Co-training
Text classification
Lazy lerner
Linearly separable
Exploration
Sparse data
Sequential data
Spatial data mining
Novelty detection
Concept drift
Incremental learning
Statistical learning
Gaussian processes
Multi-class
Expectation-maximization
Imbalanced dataset
Structural patterns
Temporal data mining
Cost-sensitive
Temporal Difference Learning
Discriminant analysis
Clustering
Association rule
Time series
Semi-supervised learning

Supervised Learning - Multi Class



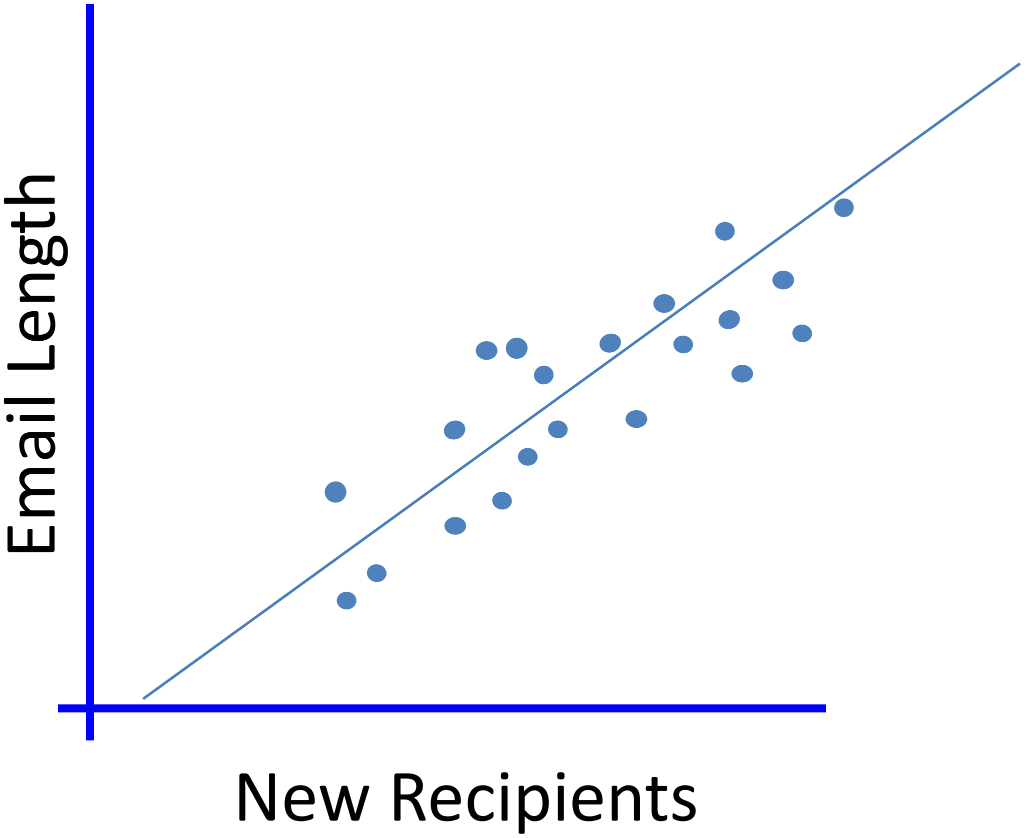
Supervised Learning - Multi Label

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



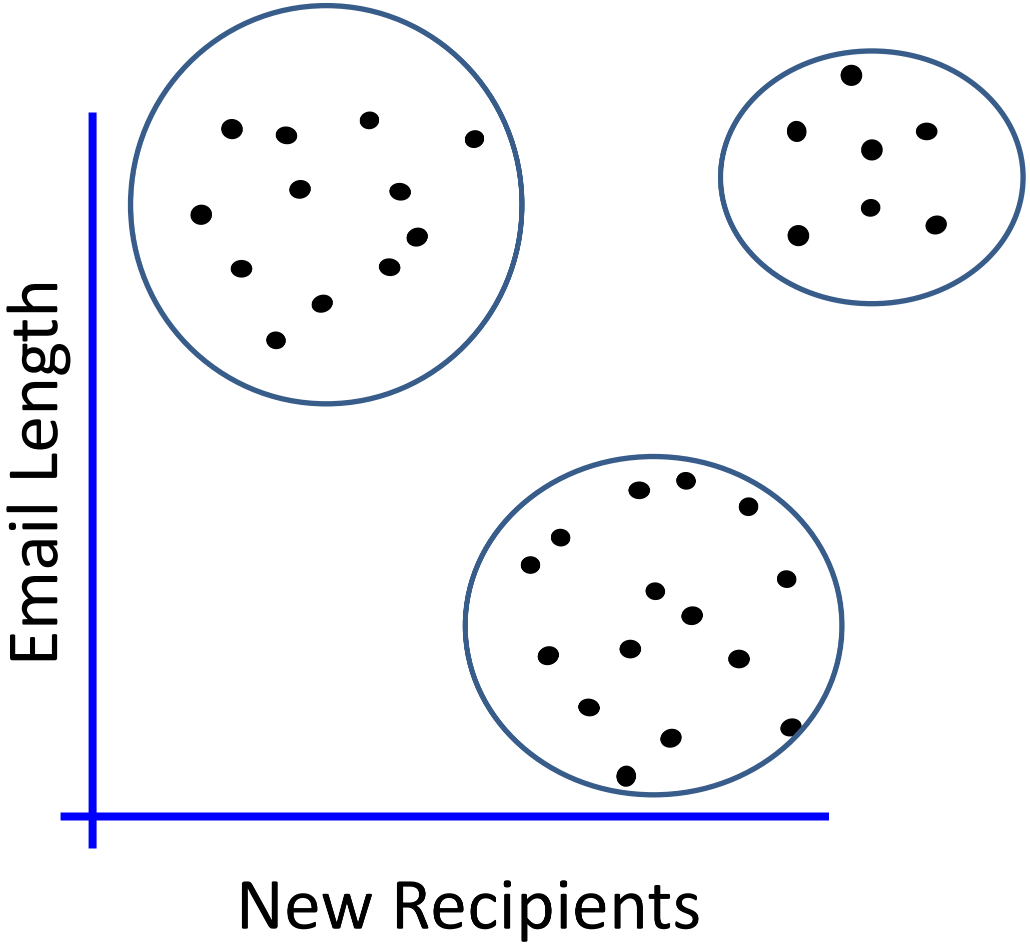
Supervised Learning - Regression

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



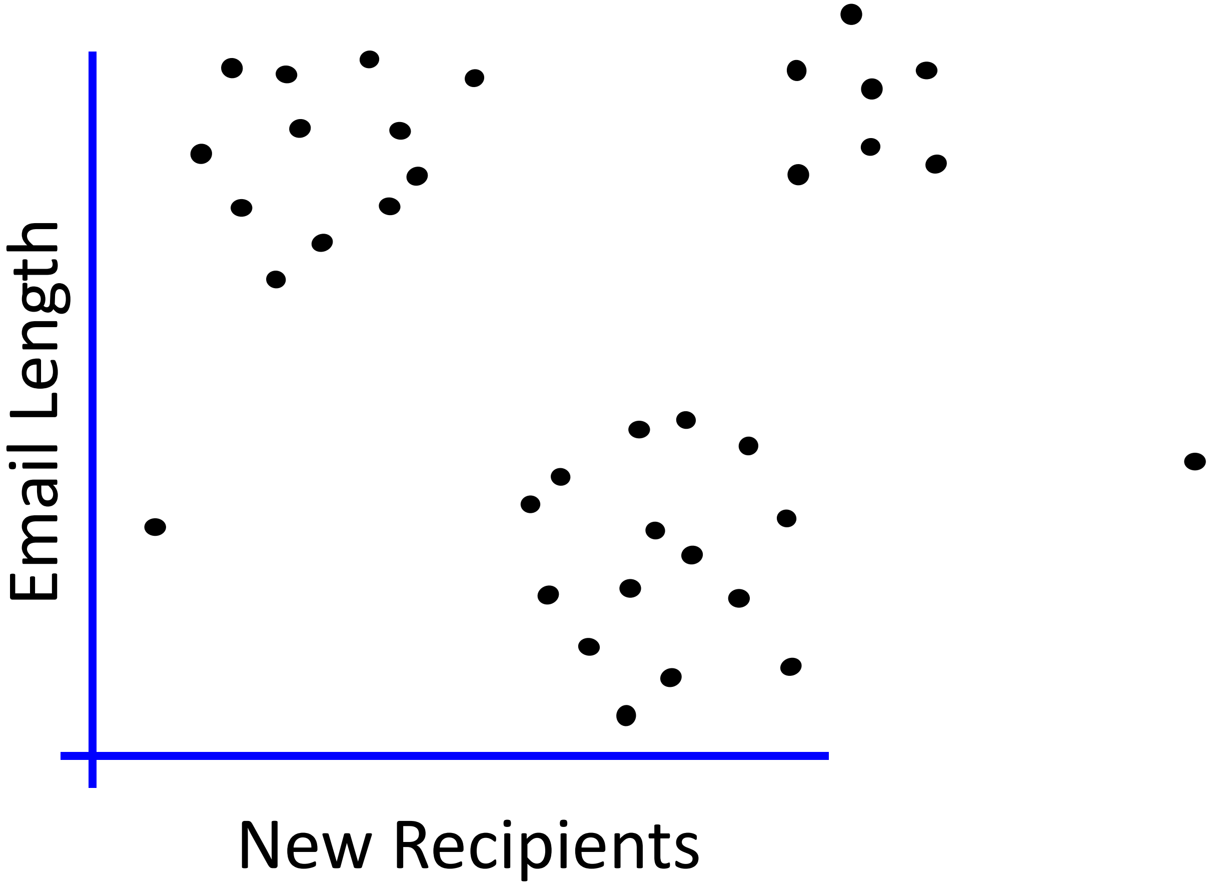
Unsupervised Learning - Clustering

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.



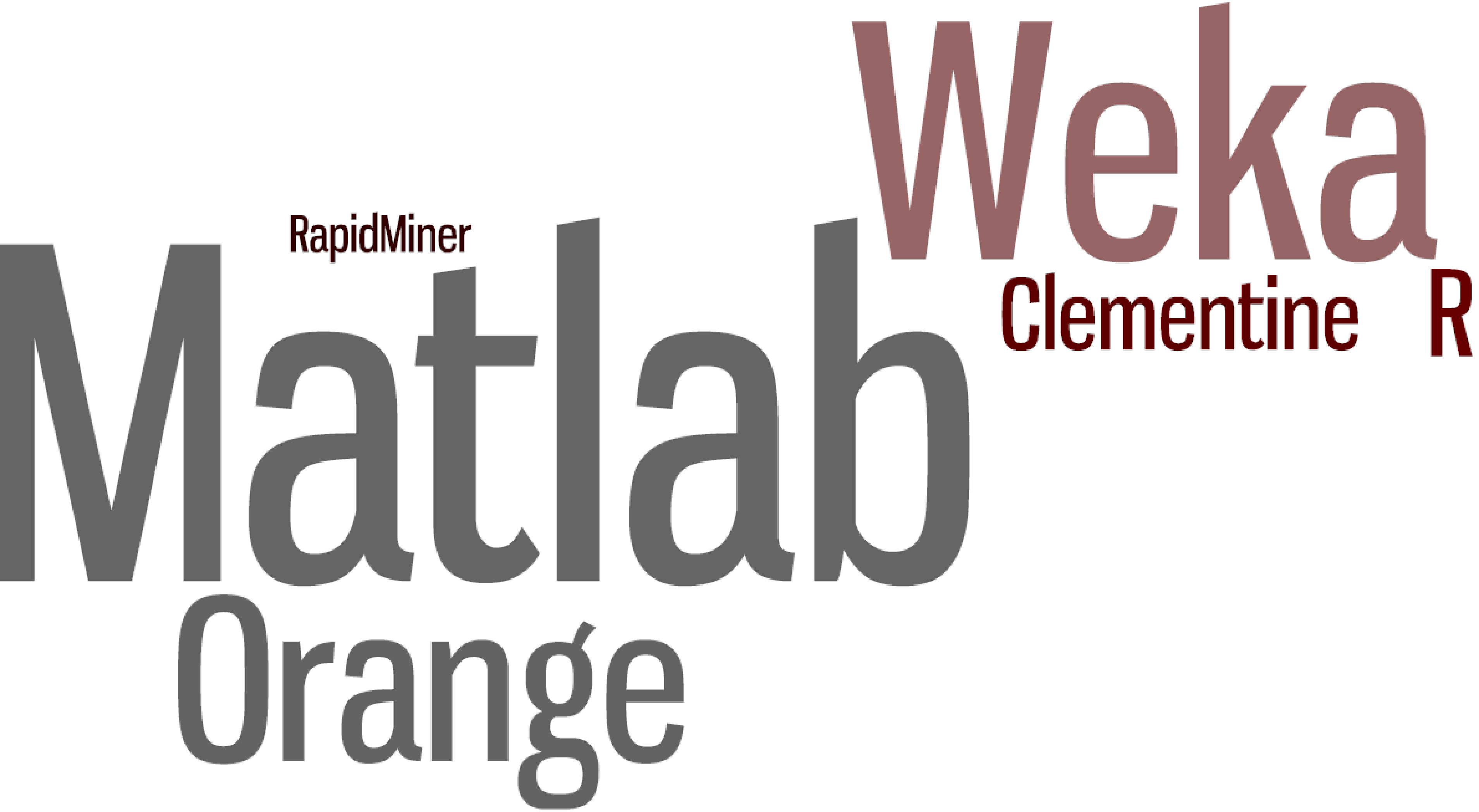
Source of Training Data

- 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

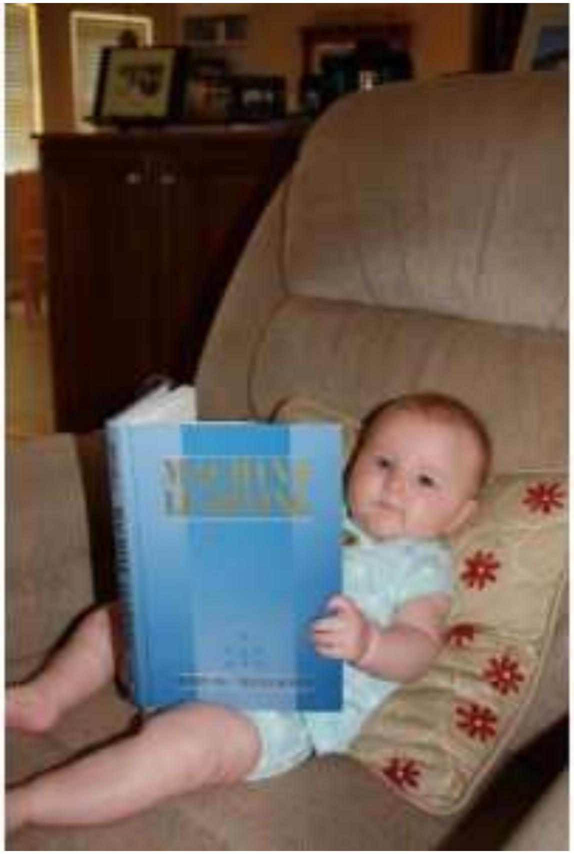
Other Learning Tasks

- **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



Want to Learn More?

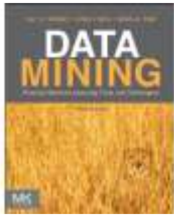


Want to Learn More?

- 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.