Decision Forest

After Twenty Years

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

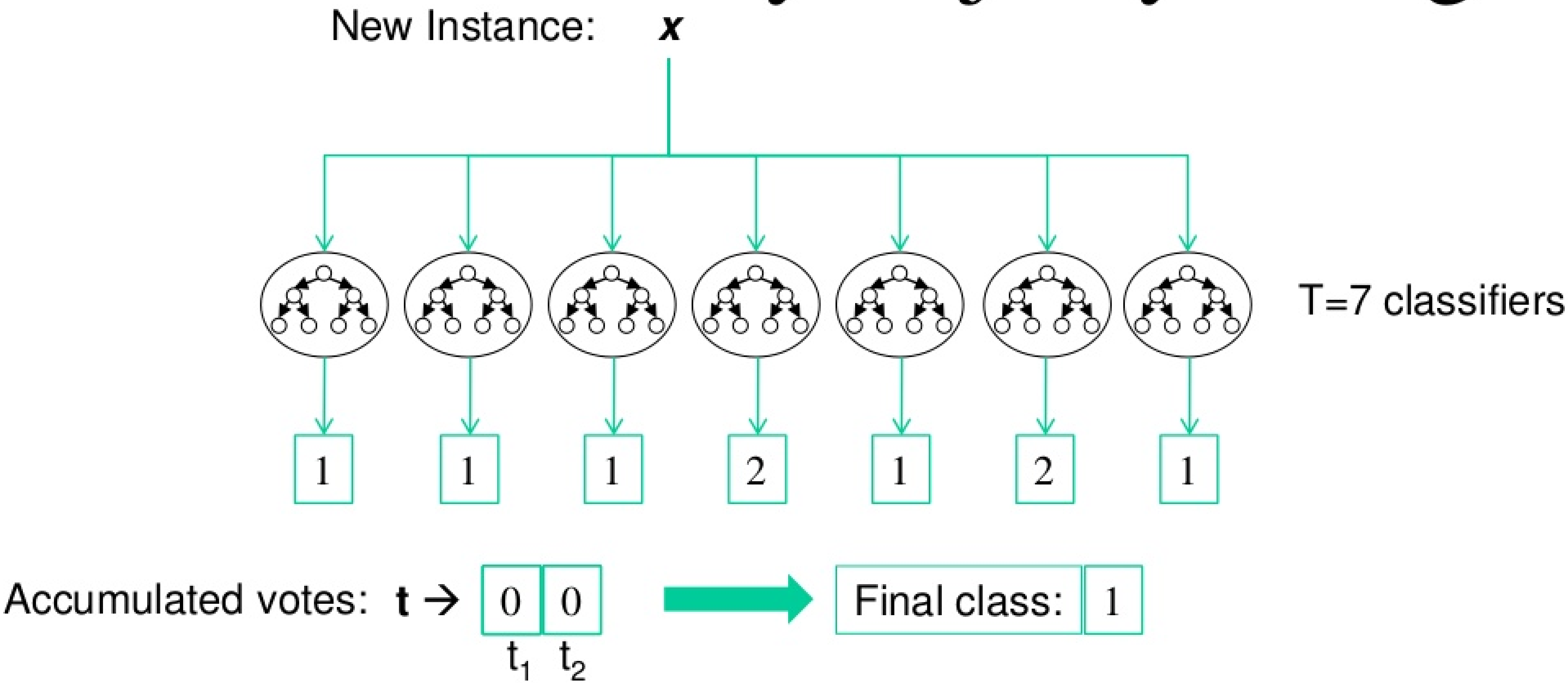
Do we need hundreds of classifiers to solve real world classification problems? (Fernández-Delgado *et al.*, 2014)

Empirically comparing

179 classification algorithms
over 121 datasets

"The classifier most likely to be the best is random forest (achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets)"

Classification by majority voting

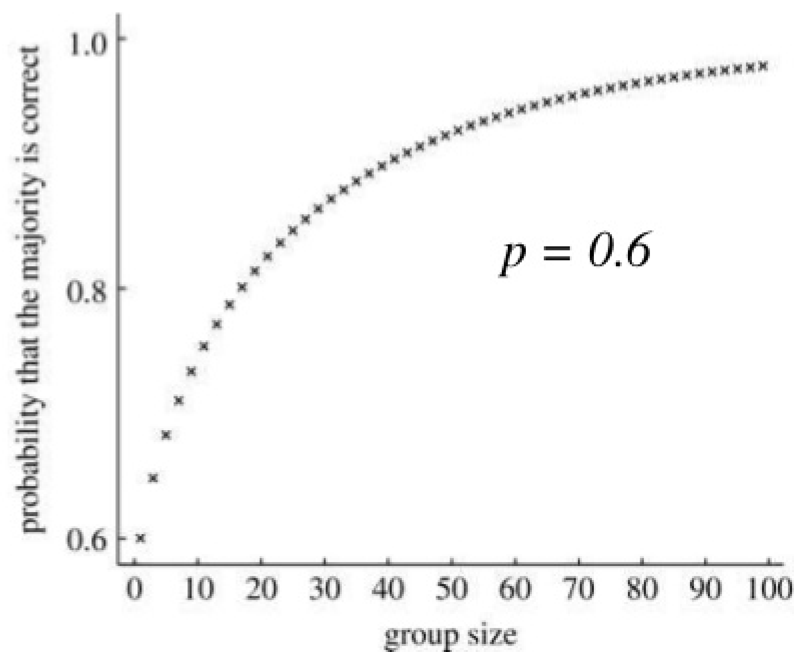


Obtained from Alberto Suárez, 2012

The Condorcet's Jury Theorem (Marquis of Condorcet, 1784)

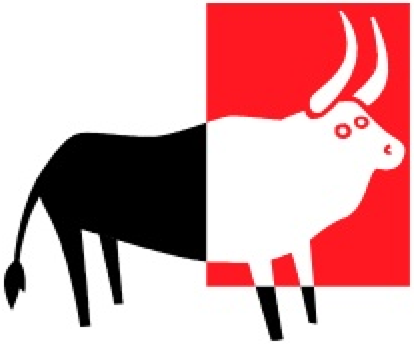
- The most basic jury theorem in social choice
- N = the number of jurors
- p = the probability of an individual juror being right
- μ = the probability that a jury gives the correct answer

$$\mu = \sum_{i=m}^{N} \left(\frac{N!}{(N-i)!\,i!} \right) (p)^{i} (1-p)^{N-i}$$



- $p > 0.5$ implies $\mu > p$.
- and $\mu \rightarrow 1$ when $N \rightarrow \infty$.

The Wisdom of Crowds

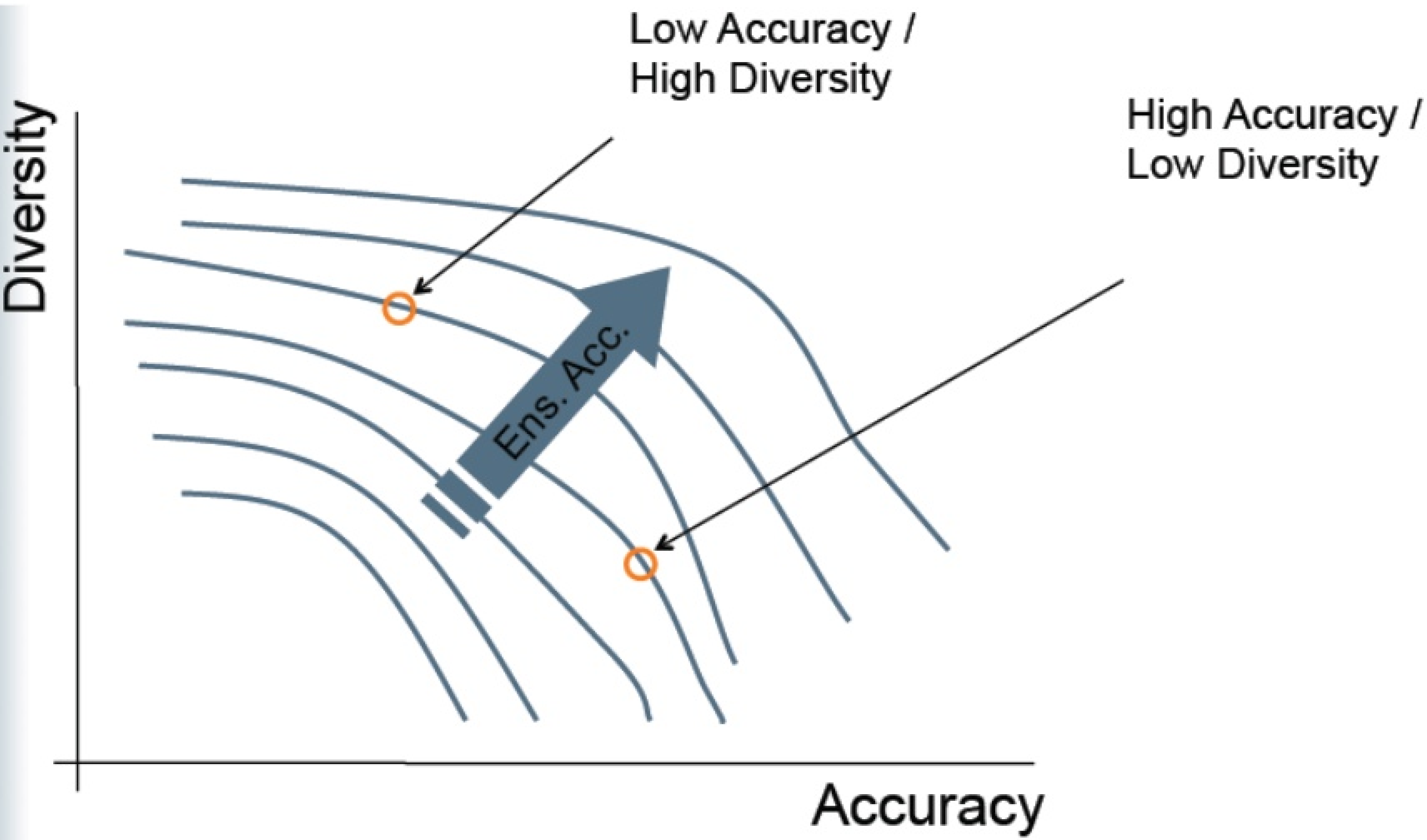


- Francis Galton promoted statistics and invented the concept of correlation.
- In 1906 Galton visited a livestock fair and stumbled upon an intriguing contest.
- An ox was on display, and the villagers were invited to guess the animal's weight.
- Nearly 800 gave it a go and, not surprisingly, not one hit the exact mark: 1,198 pounds.
- Astonishingly, however, the average of those 800 guesses came close - very close indeed. It was 1,197 pounds.

Key Criteria for Crowd to be Wise

- Diversity of opinion
  - Each person should have private information even if it's just an eccentric interpretation of the known facts.
- Independence
  - People's opinions aren't determined by the opinions of those around them.
- Decentralization
  - People are able to specialize and draw on local knowledge.
- Aggregation
  - Some mechanism exists for turning private judgments into a collective decision.

The Diversity Tradeoff

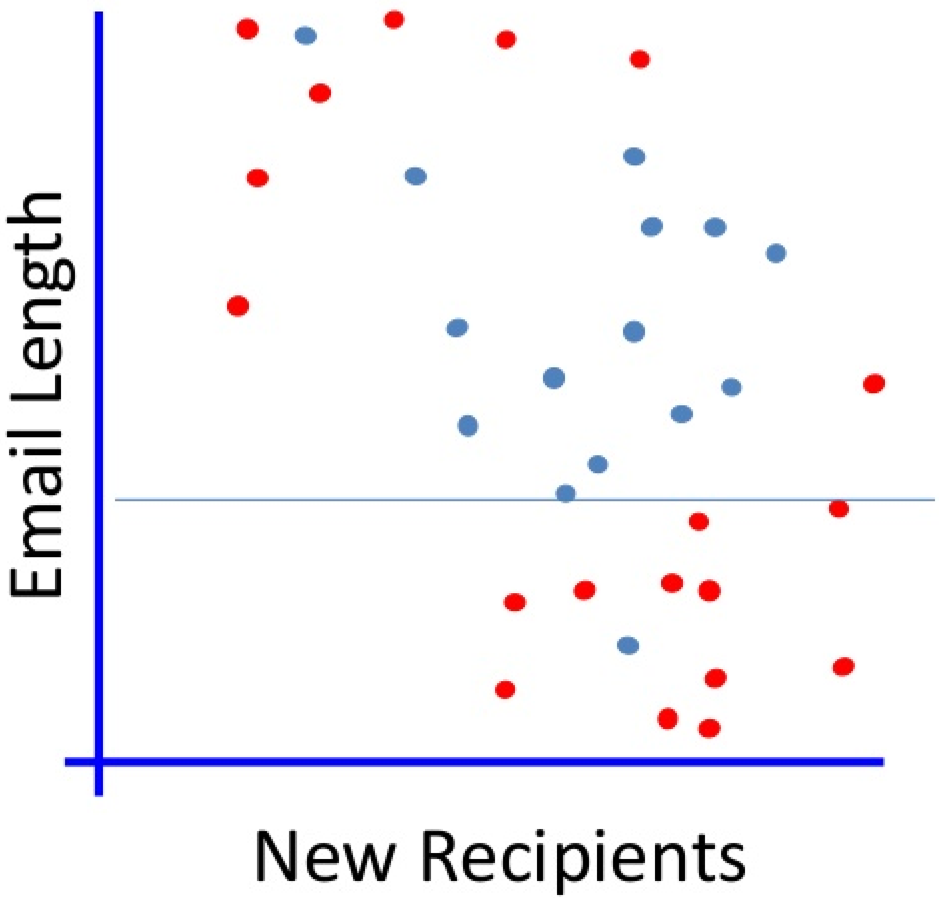


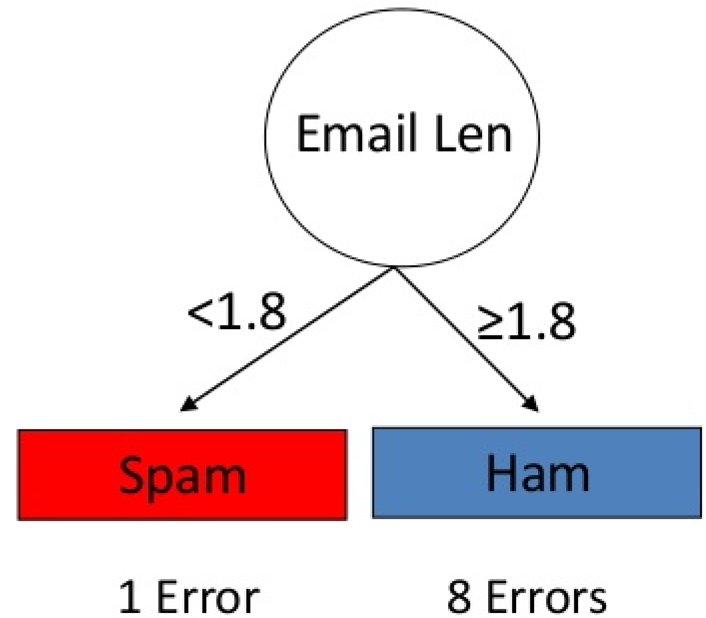
of individual trees

There’s no Real Tradeoff…

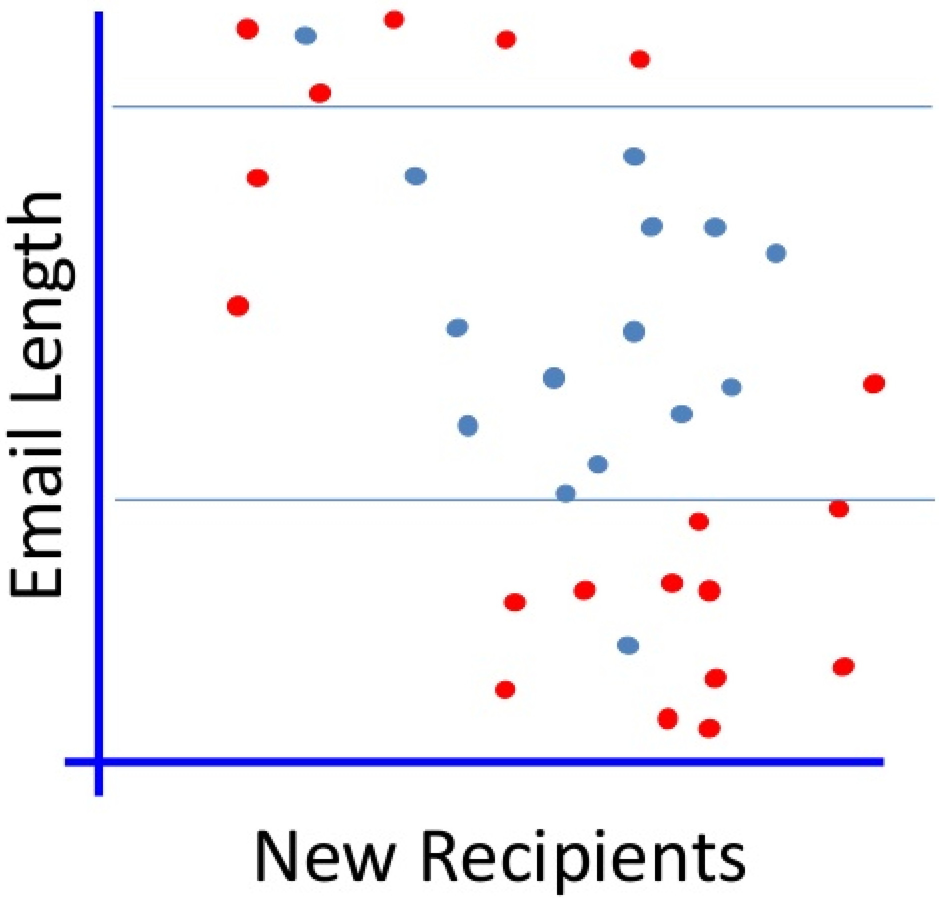
- Ideally, all trees would be right about everything!
- If not, they should be wrong about different cases.

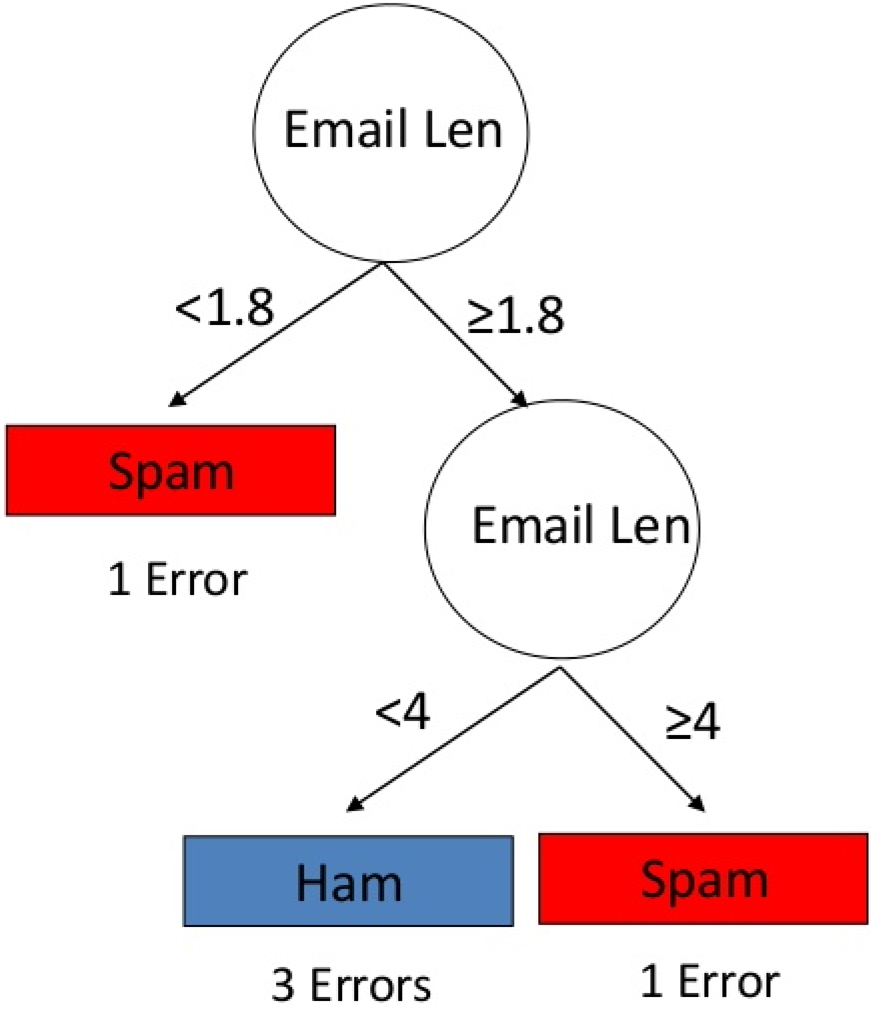
Top Down Induction of Decision Trees



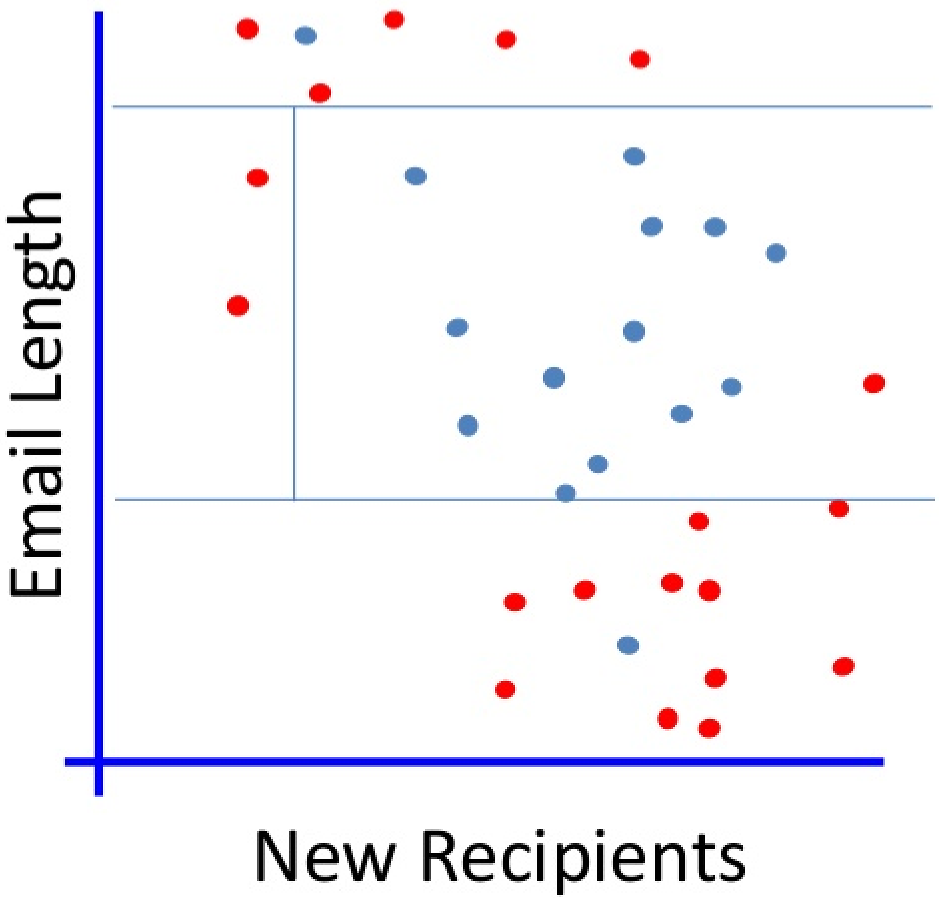


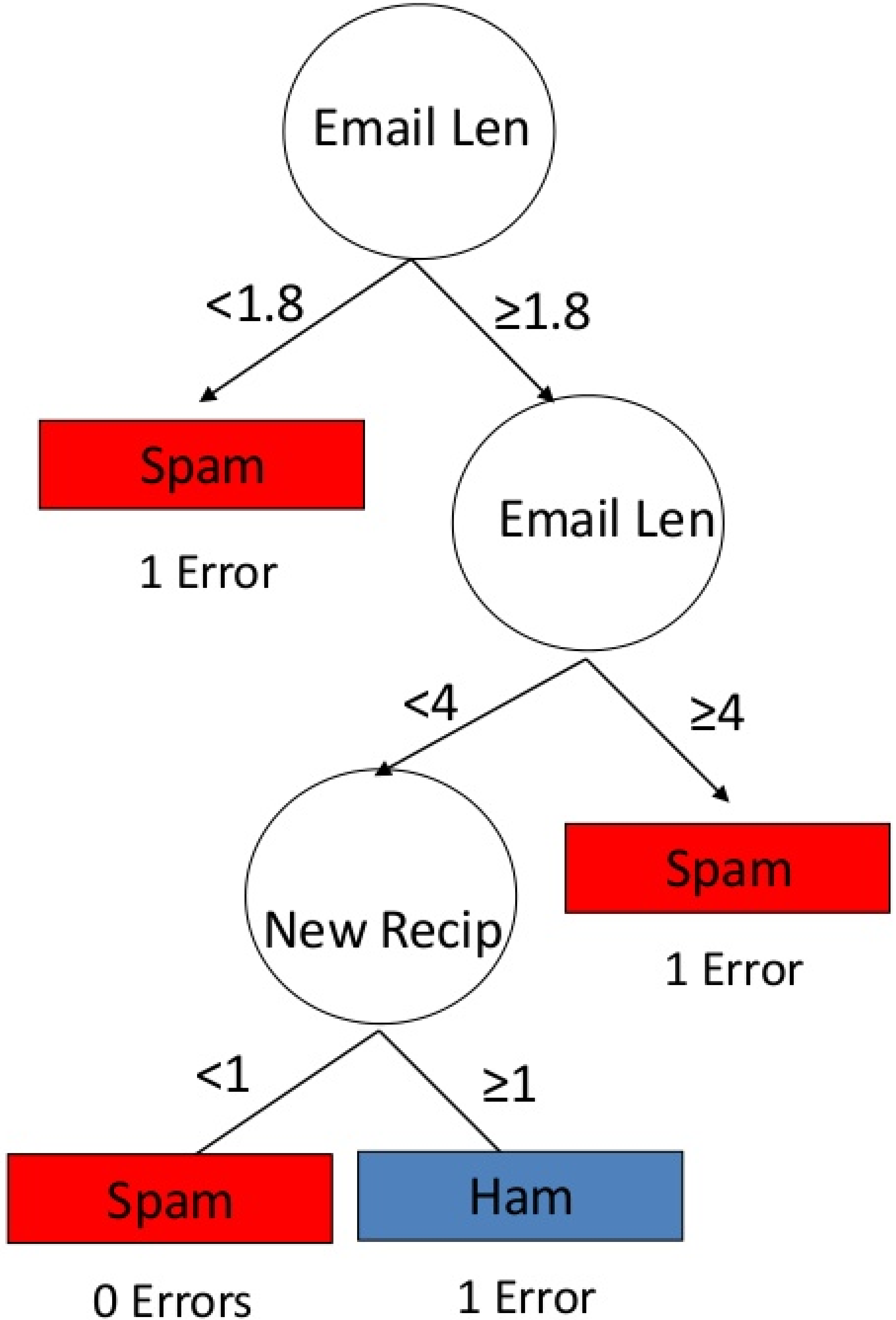
Top Down Induction of Decision Trees



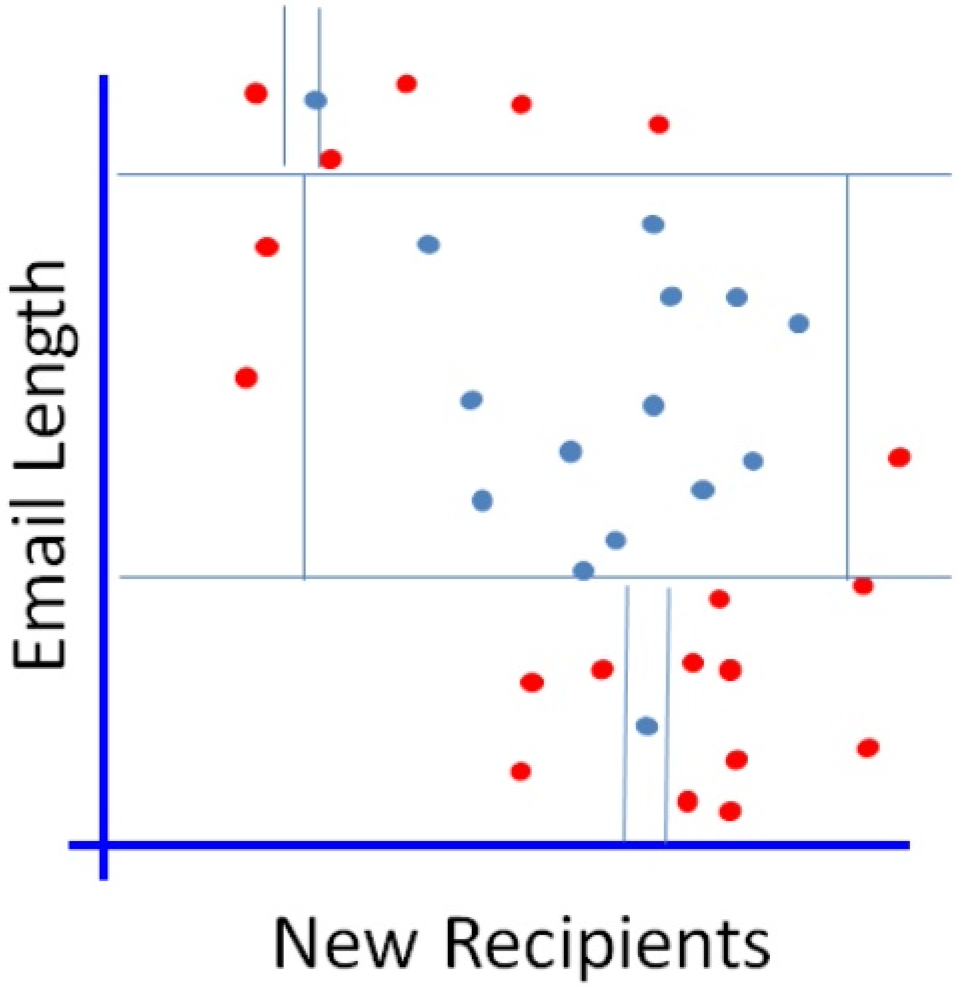


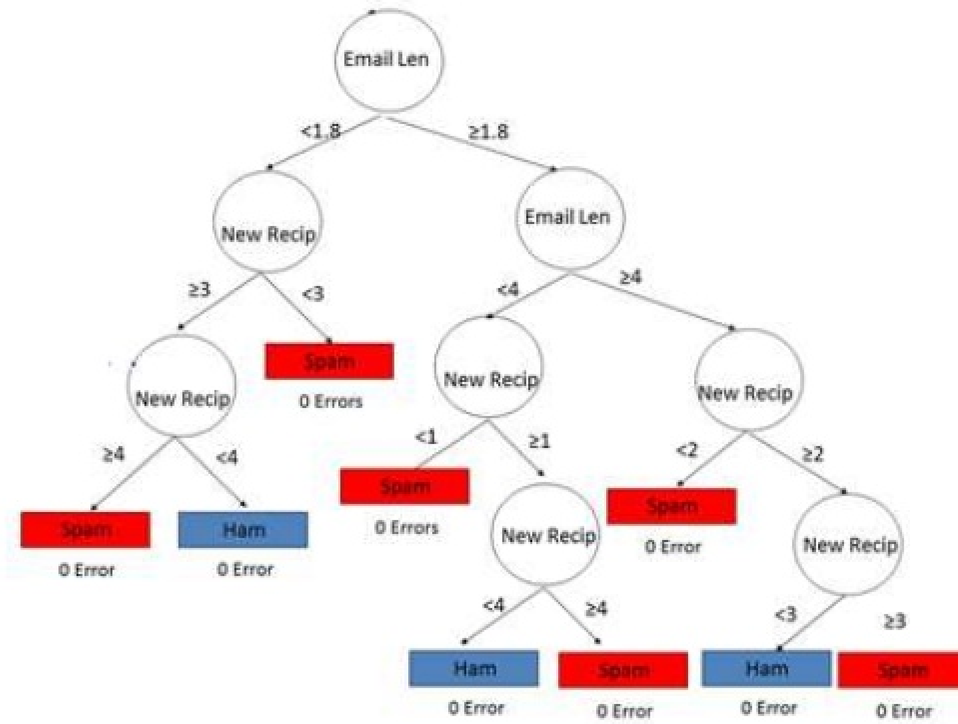
Top Down Induction of Decision Trees





Top Down Induction of Decision Trees





Why Does Decision Forest Work?

- Local minima
- Lack of sufficient data
- Limited Representation

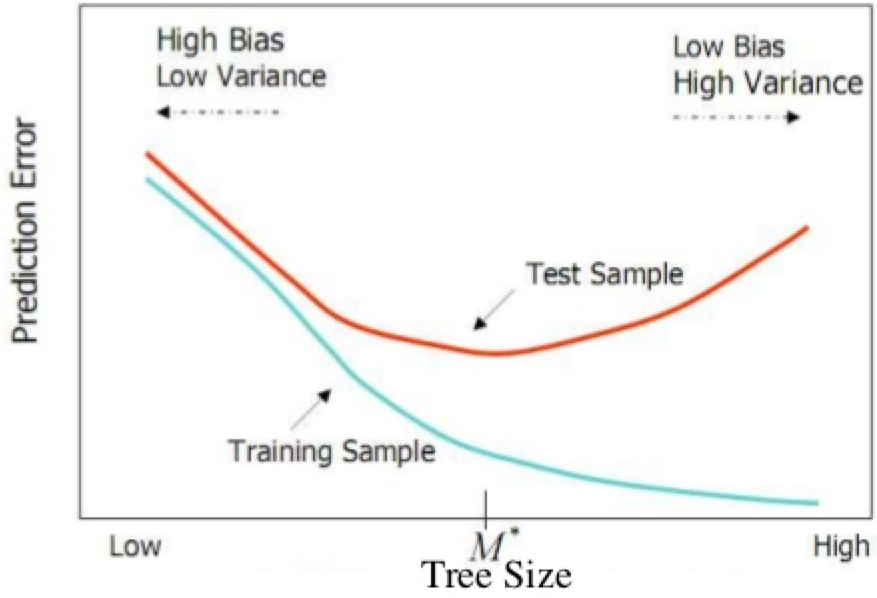
Bias and Variance Decomposition

<u>Bias</u>

- The tendency to consistently learn the same wrong thing because the hypothesis space considered by the learning algorithm does not include sufficient hypotheses

<u>Variance</u>

- The tendency to learn random things irrespective of the real signal due to the particular training set used



It all started about two years ago …

<u>Iterative Methods</u>

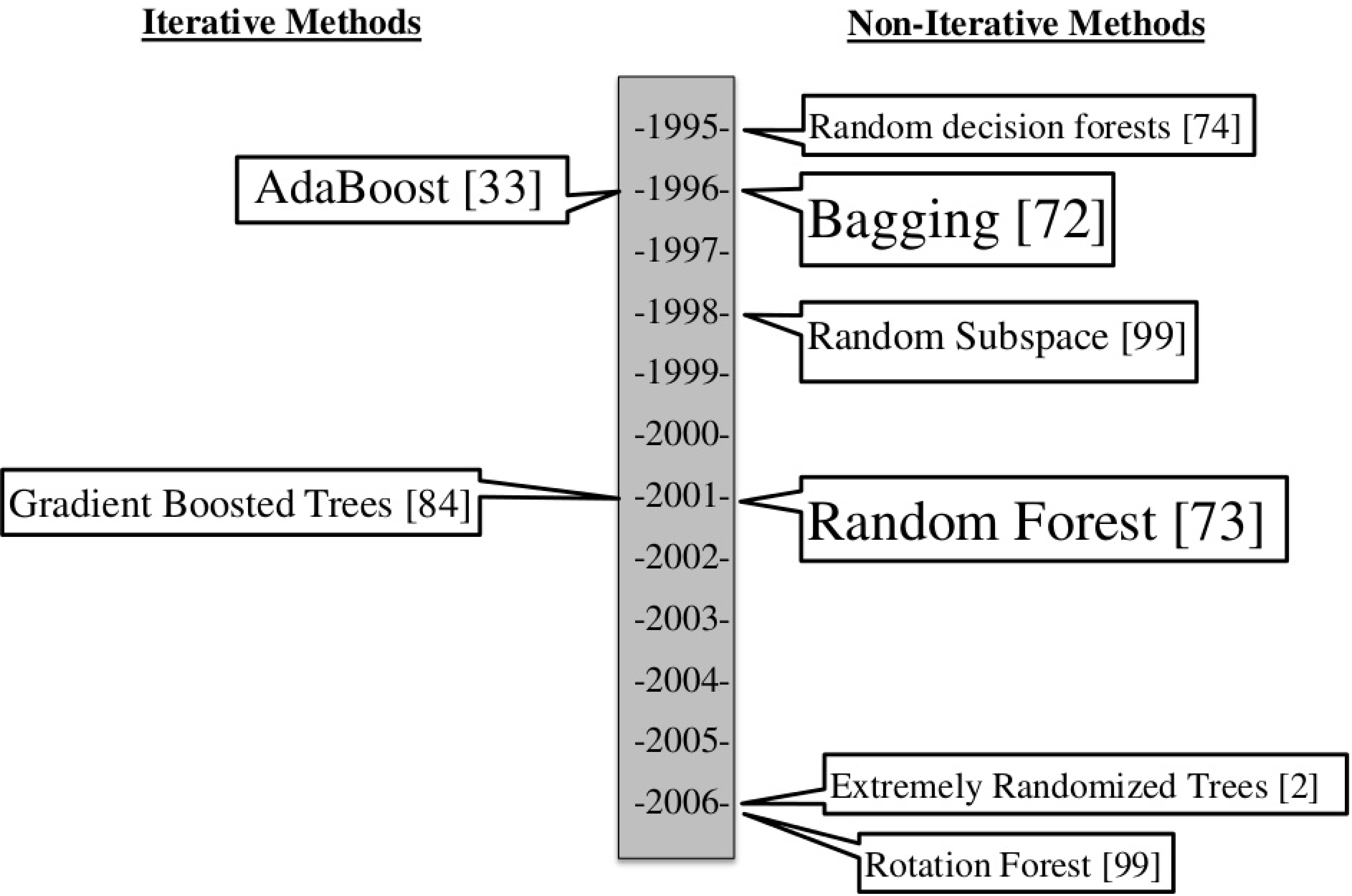
- Reduce both Bias and Variance errors
- Hard to parallelize

- AdaBoost (Freund & Schapire, 1996)
- Gradient Boosted Trees (Friedman, 1999)
- Feature-based Partitioned Trees (Rokach, 2008)
- Stochastic gradient boosted distributed decision trees (Ye et al., 2009)
- Parallel Boosted Regression Trees (Tyree et al., 2011)

<u>Non-Iterative Methods</u>

- Mainly reduce variance error
- Embarrassingly parallel

- Random decision forests (Ho, 1995)
- Bagging (Bootstrap aggregating) (Breiman, 1996)
- Random Subspace Decision Forest (Ho, 1998)
- Randomized Tree (Dietterich, 2000)
- Random Forest (Breiman, 2001)
- Switching Classes (Martínez-Muñoz and Suárez, 2005)
- Rotation Forest (Rodríguez et al., 2006)
- Extremely Randomized Trees (Geurts et al., 2006)
- Randomly Projected Trees (Schclar and Rokach, 2009)

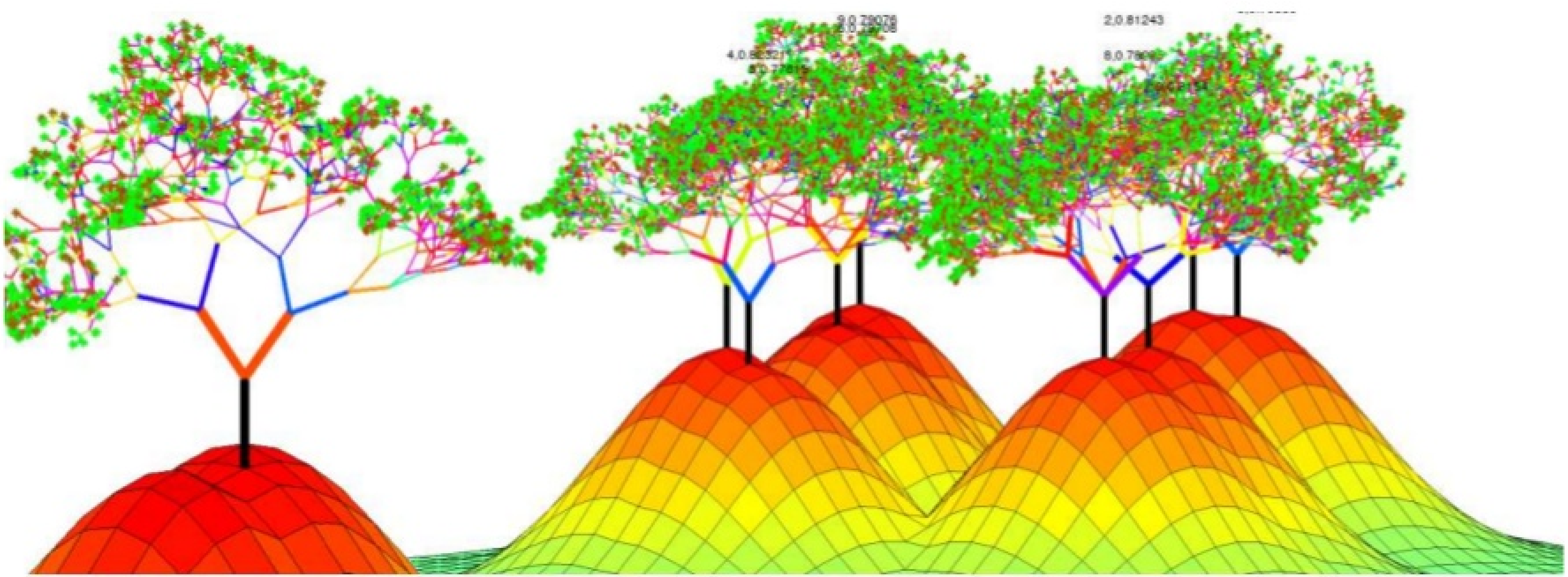


Iterative Methods
Non-Iterative Methods
-1995-
-1996-
-1997-
-1998-
-1999-
-2000-
-2001-
-2002-
-2003-
-2004-
-2005-
-2006-
Random decision forests [74]
AdaBoost [33]
Bagging [72]
Random Subspace [99]
Gradient Boosted Trees [84]
Random Forest [73]
Extremely Randomized Trees [2]
Rotation Forest [99]

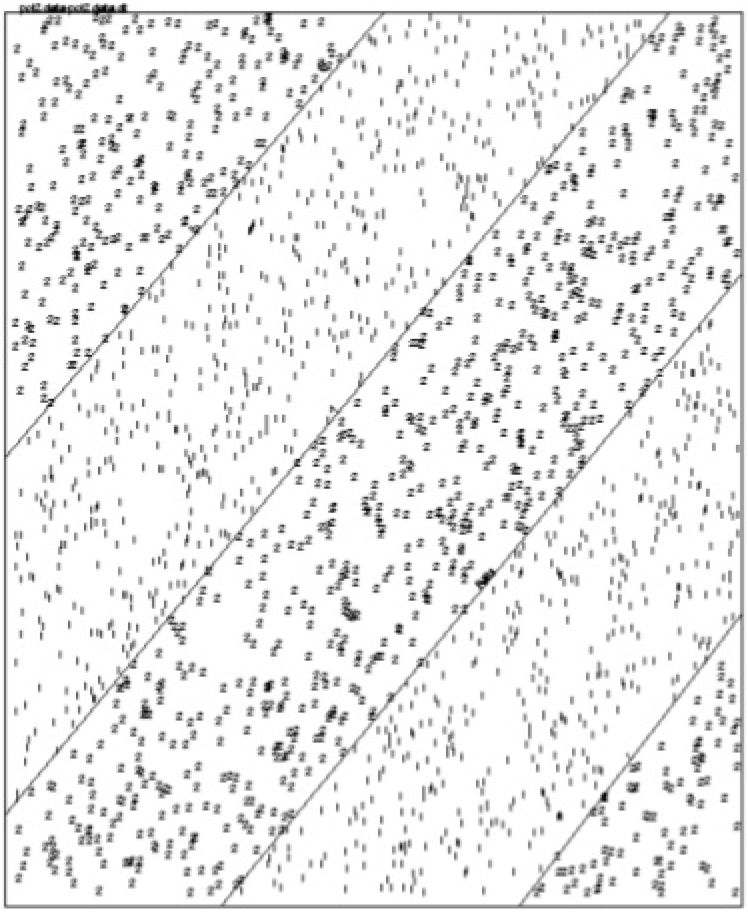
Random Forests

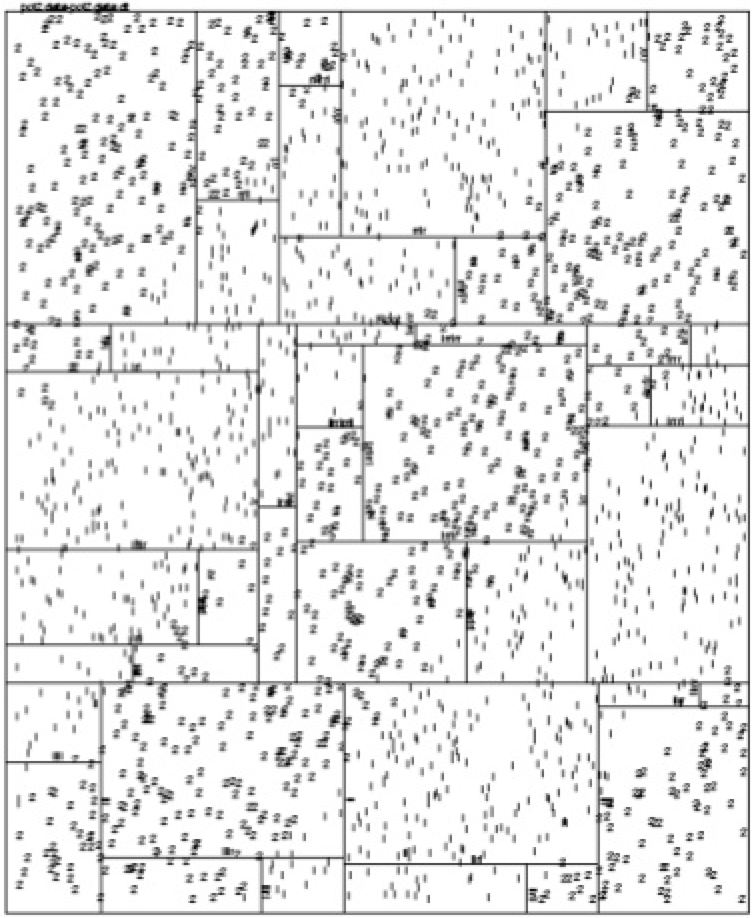
(Breiman, 2001)

1. A bootstrap ***random*** sample of size n sampled from training set ***with replacement***
2. Evaluate a node split on a ***random*** subset of variables
3. No pruning.



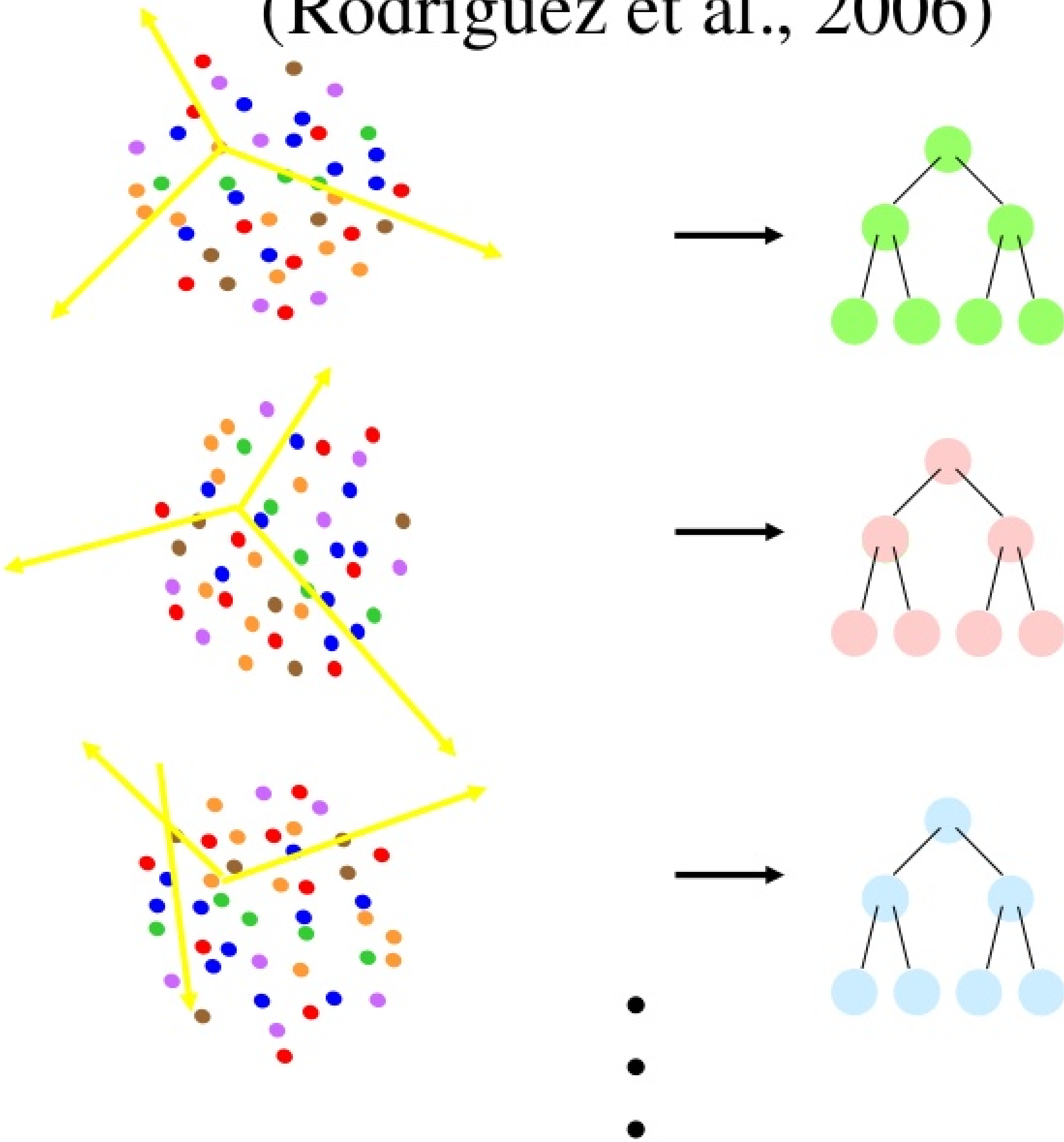
Limited Representation





Rotation Forest

(Rodríguez et al., 2006)



AdaBoost

(Freund & Schapire, 1996)

"Best off-the-shelf classifier in the world" – Breiman (1996)