JPL-Caltech Virtual Summer School

Big Data Analytics

September 2 – 12, 2014

Thomas Fuchs (JPL, Caltech)

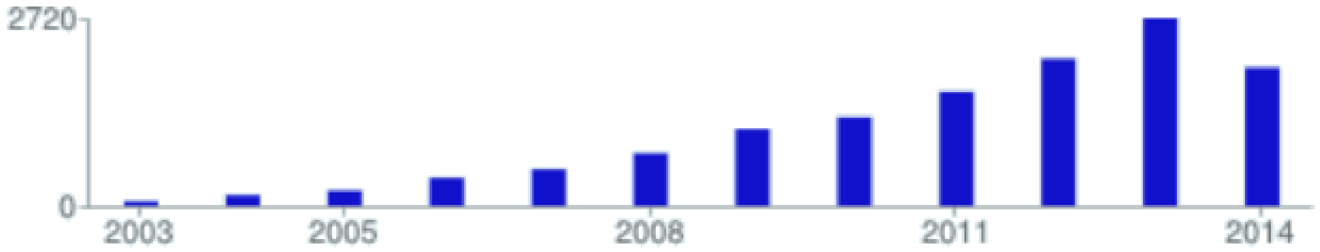
Random Forests

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Random Forests: Citations

Total citations Cited by 13154

Citations per year



Scholar articles Random forests
L Breiman - Machine learning, 2001
Cited by 13154 - Related articles - All 83 versions

Google Scholar Citations: 2014-08-21

1

RANDOM FORESTS

Leo Breiman
Statistics Department
University of California
Berkeley, CA 94720

January 2001

Abstract

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Freund and Schapire[1996]), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

History

1983	**CART**	*Breiman*
1996	**Bagging**	*Breiman*
1996	**AdaBoost**	*Freund & Schapire*
2001	**Random Forests**	*Breiman*

Leo Breiman
1928 - 2005

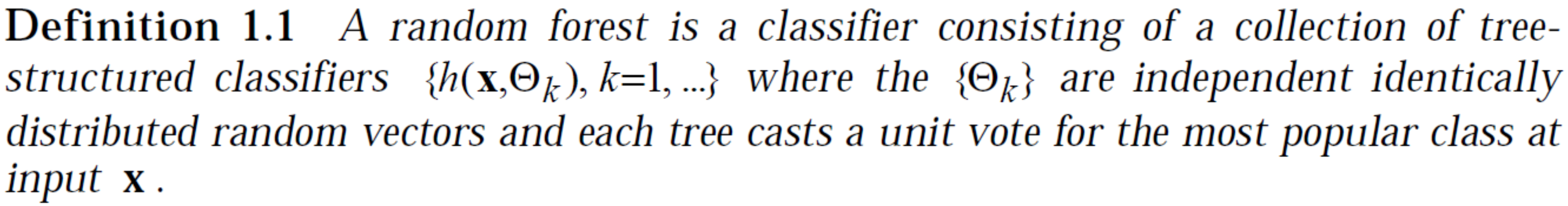
History

1983	**CART**	*Breiman*
1996	**Bagging**	*Breiman*

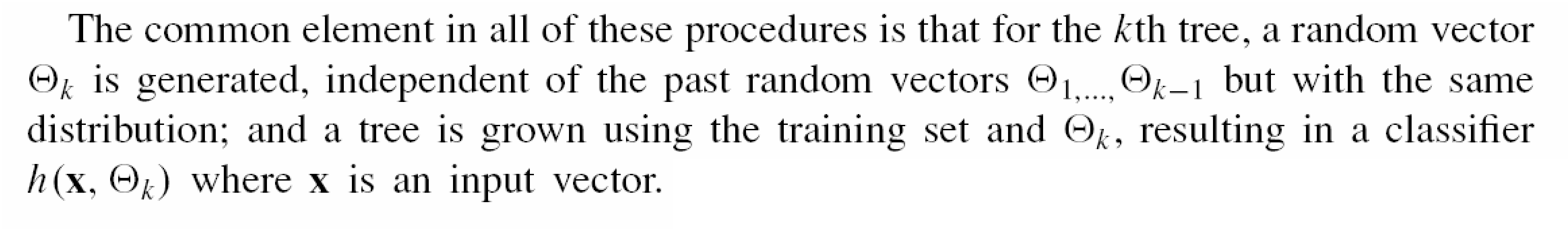
Leo Breiman
1928 - 2005

1994	**Randomized Trees (WS)**	*Amint & Geman*
1996	**AdaBoost**	*Freund & Schapire*
1997	**Randomized Trees**	*Amint & Geman*
1998	**Decision Forests**	*Ho*
1998	**Random split selection**	*Dietterich*
2001	**Random Forests**	*Breiman*

Random Forests (Breiman 2001)

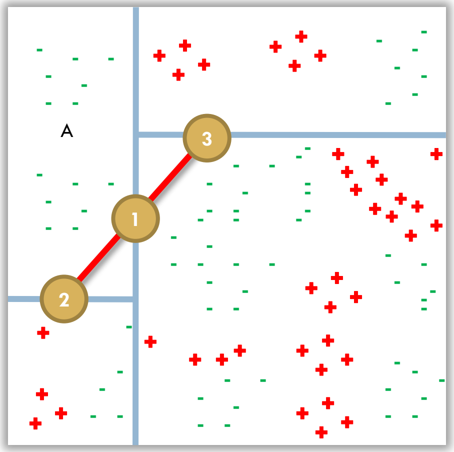


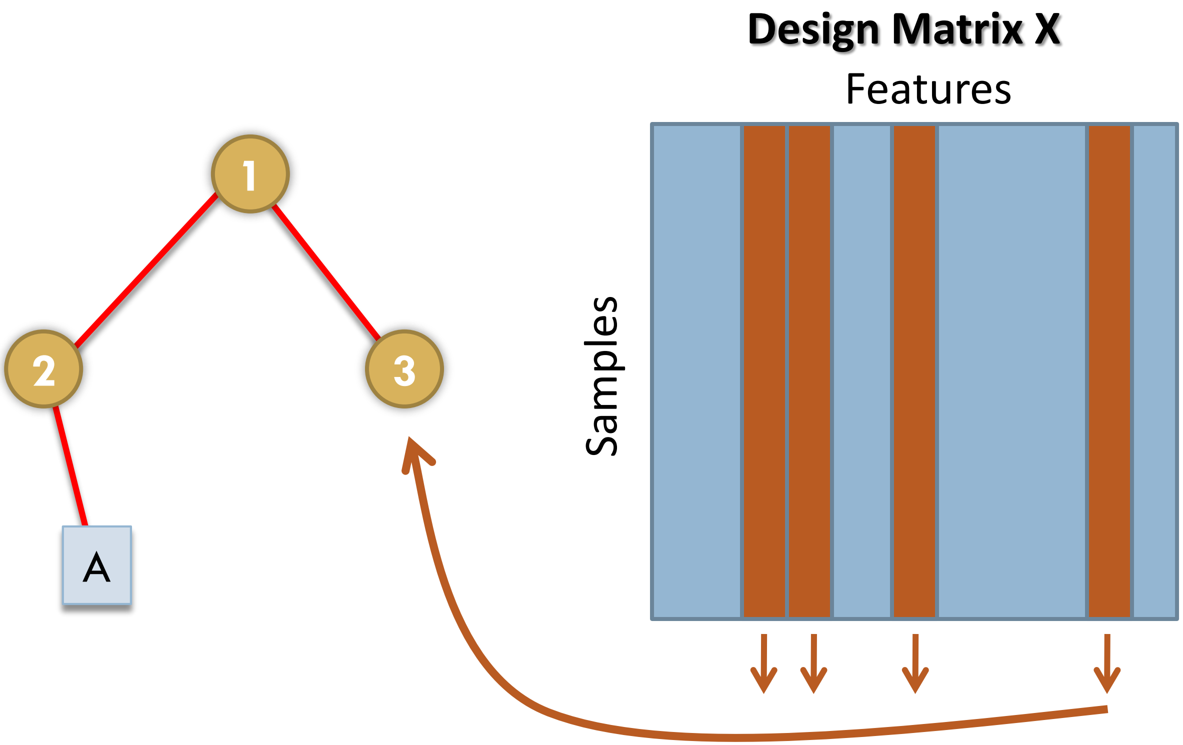
Definition 1.1 *A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(\mathbf{x},\Theta_k), k=1, ...\}$ *where the* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\mathbf{x}$.



The common element in all of these procedures is that for the kth tree, a random vector Θ_k is generated, independent of the past random vectors $\Theta_{1,\ldots,}\Theta_{k-1}$ but with the same distribution; and a tree is grown using the training set and Θ_k, resulting in a classifier $h(\mathbf{x}, \Theta_k)$ where $\mathbf{x}$ is an input vector.

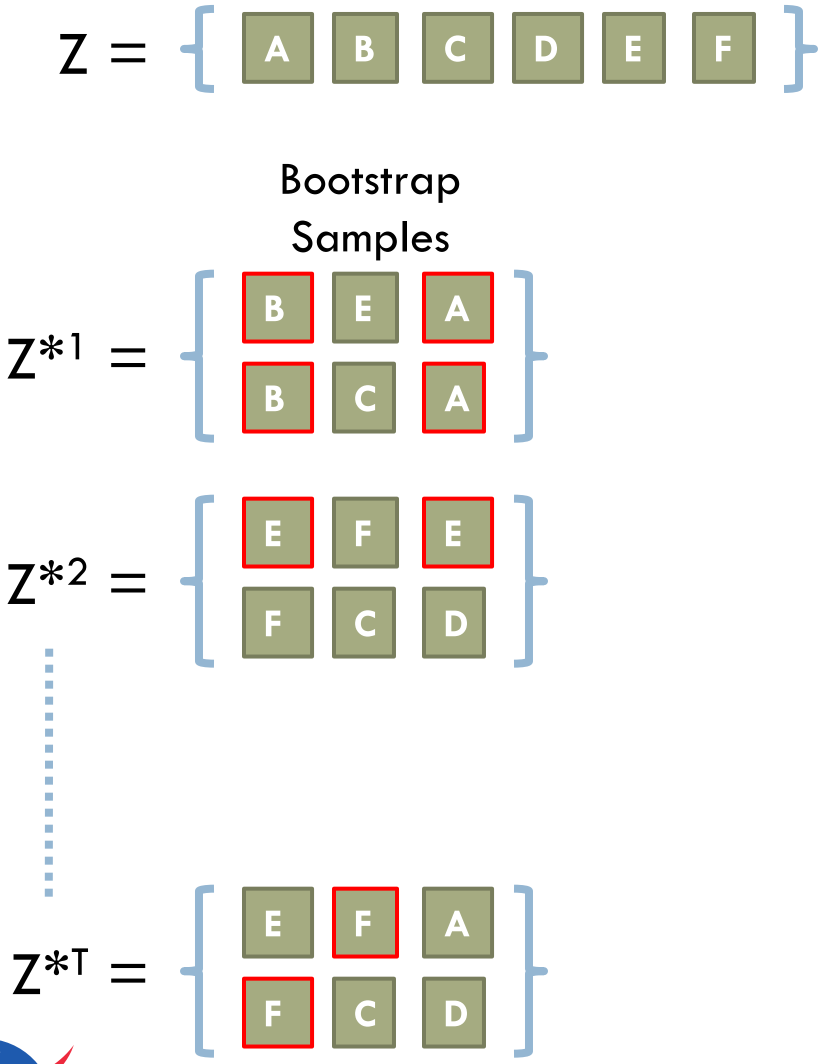
Randomized Tree Learning





At each node only a random subset of features is considered to choose the best split. Common splitting criteria are Entropy, Gini Index and misclassification rate.

Random Forest Learning



$Z = \{$ A B C D E F $\}$

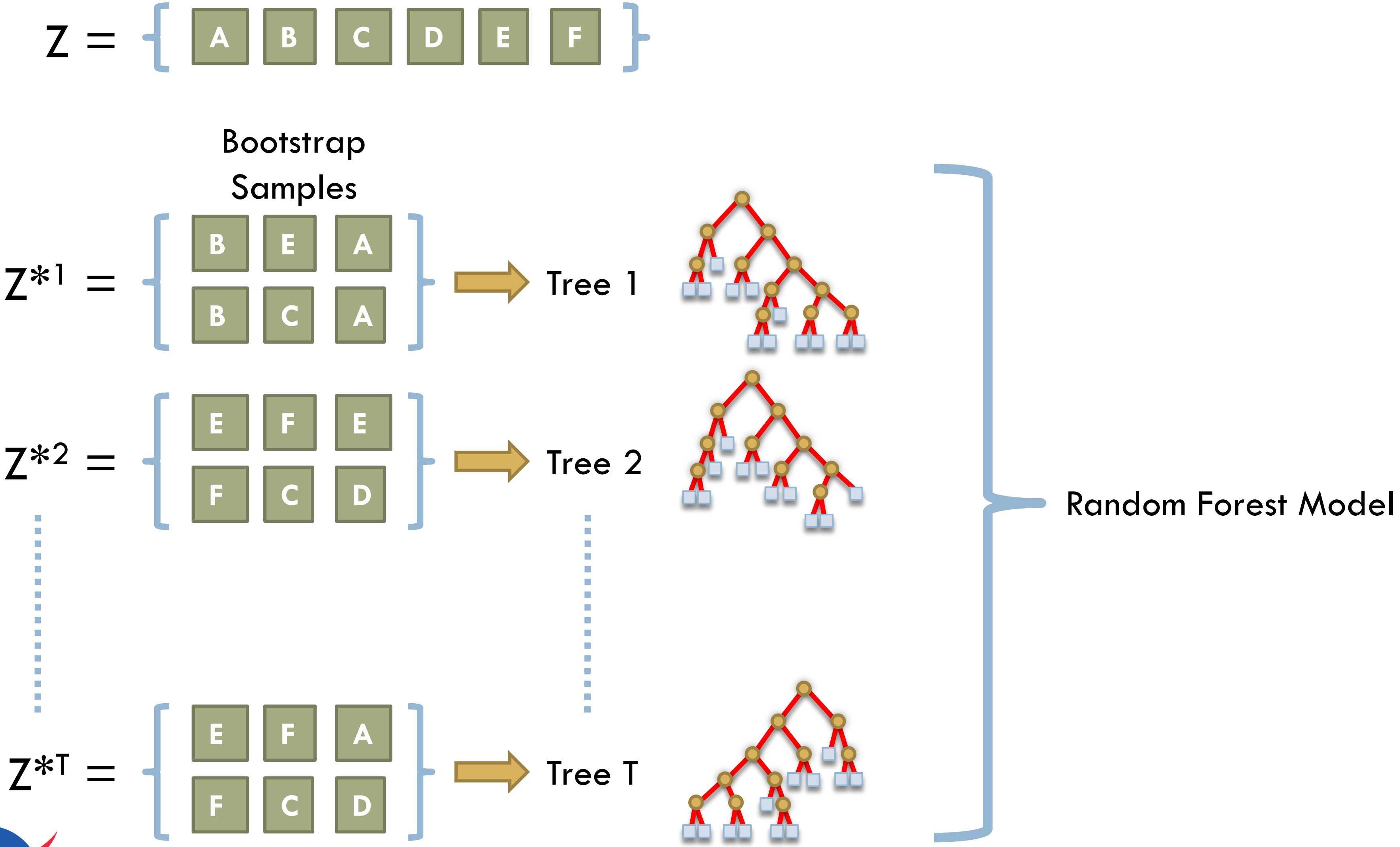
Bootstrap Samples

$Z^{*1} = \{$ B E A / B C A $\}$

$Z^{*2} = \{$ E F E / F C D $\}$

$Z^{*T} = \{$ E F A / F C D $\}$

Random Forest Learning



$Z = \{$ A B C D E F $\}$

Bootstrap Samples

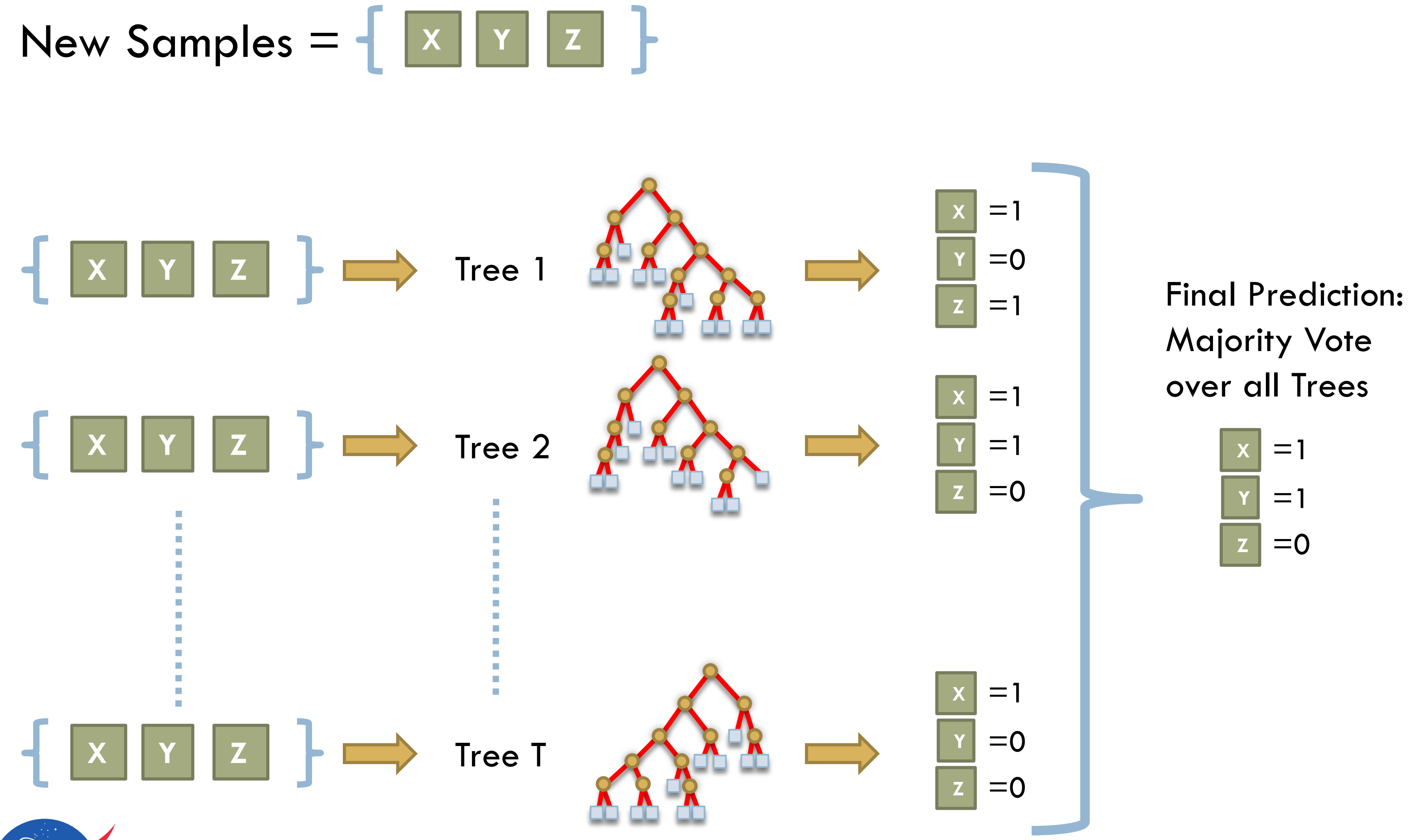
$Z^{*1} = \{$ B E A / B C A $\}$ → Tree 1

$Z^{*2} = \{$ E F E / F C D $\}$ → Tree 2

$Z^{*T} = \{$ E F A / F C D $\}$ → Tree T

Random Forest Model

Random Forest Classification



New Samples = { X Y Z }

{ X Y Z } → Tree 1 → X =1 Y =0 Z =1

{ X Y Z } → Tree 2 → X =1 Y =1 Z =0

{ X Y Z } → Tree T → X =1 Y =0 Z =0

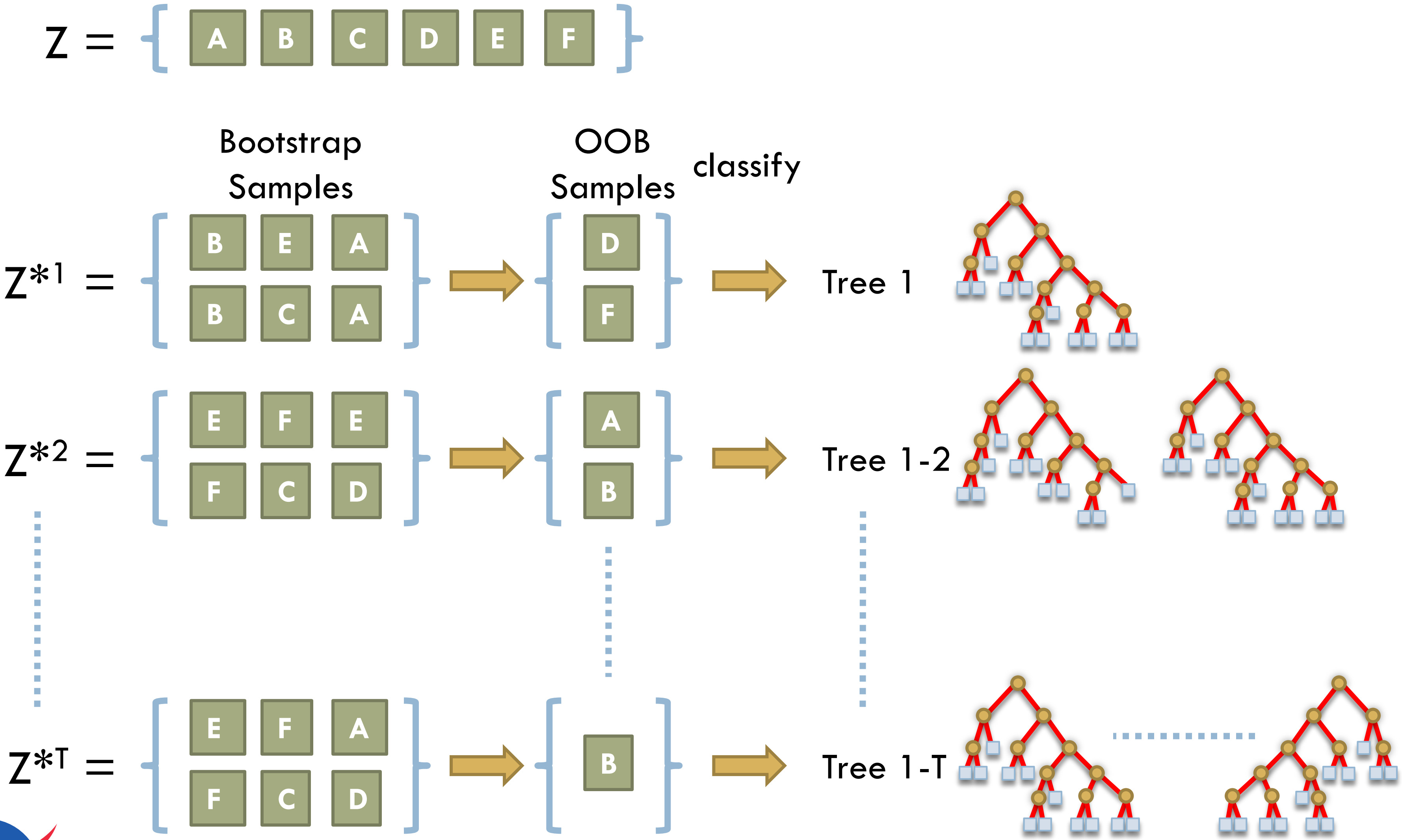
Final Prediction: Majority Vote over all Trees

X =1

Y =1

Z =0

Out Of Bag (OOB) Error



$Z = \{ A \; B \; C \; D \; E \; F \}$

Bootstrap Samples

OOB Samples

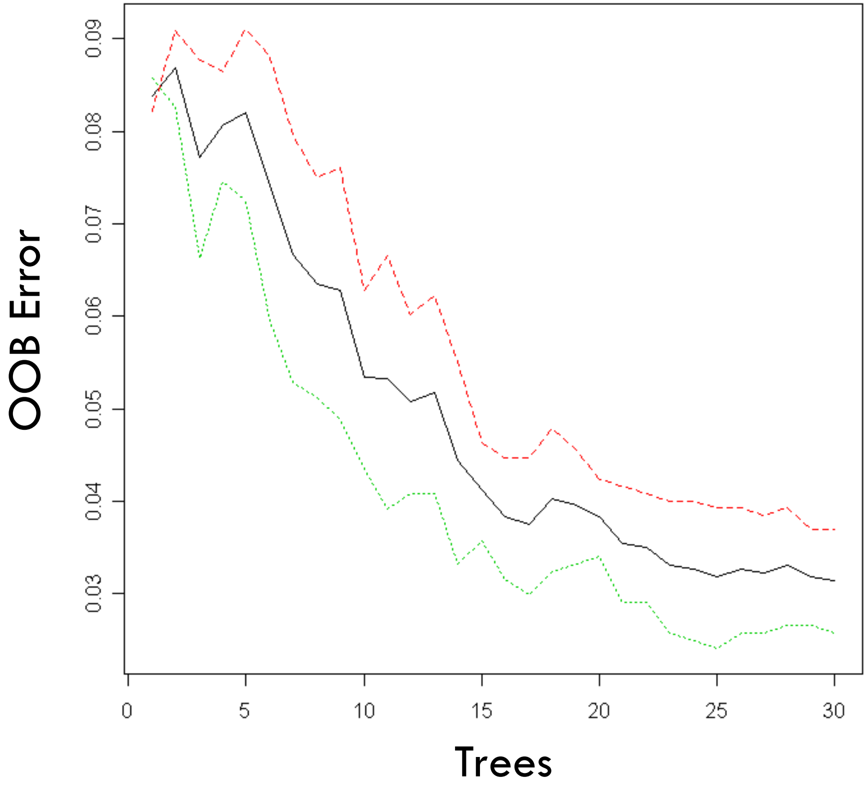
classify

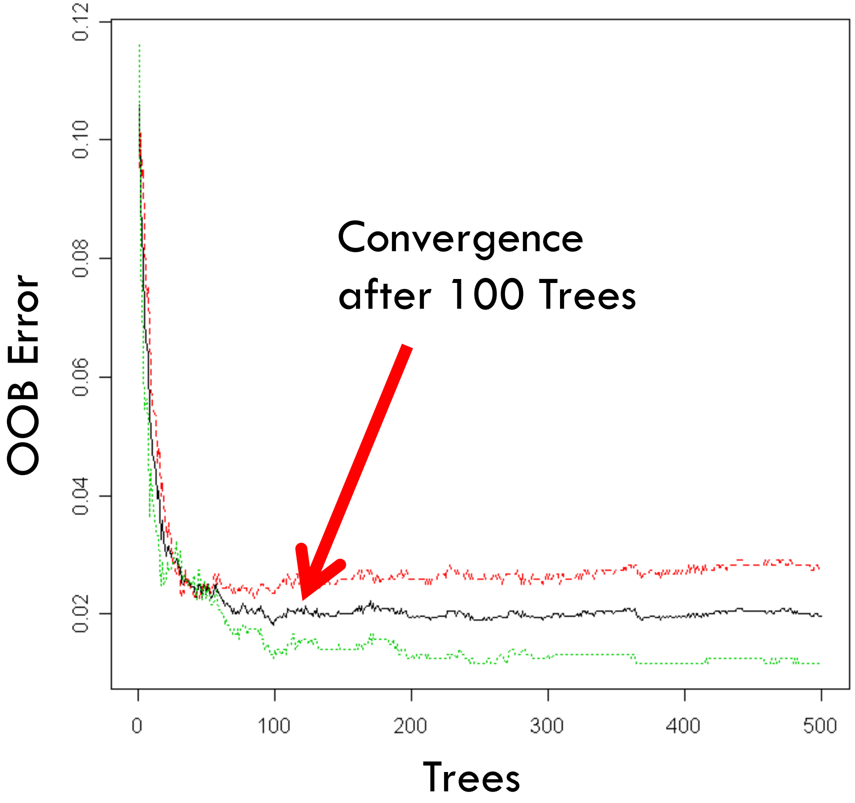
$Z^{*1} = \{ B \; E \; A \; B \; C \; A \}$ → $\{ D \; F \}$ → Tree 1

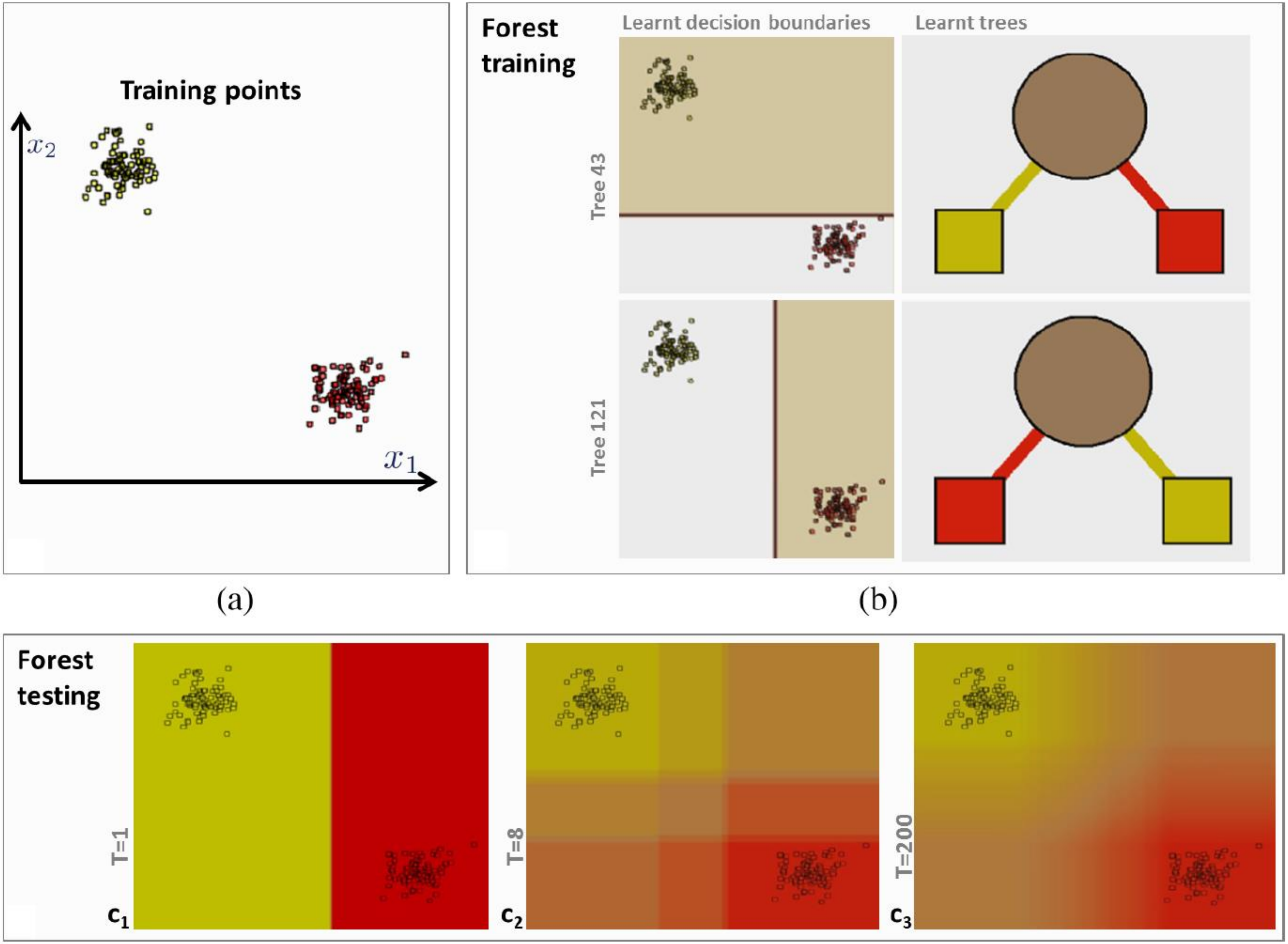
$Z^{*2} = \{ E \; F \; E \; F \; C \; D \}$ → $\{ A \; B \}$ → Tree 1-2

$Z^{*T} = \{ E \; F \; A \; F \; C \; D \}$ → $\{ B \}$ → Tree 1-T

Out Of Bag Error




Convergence after 100 Trees

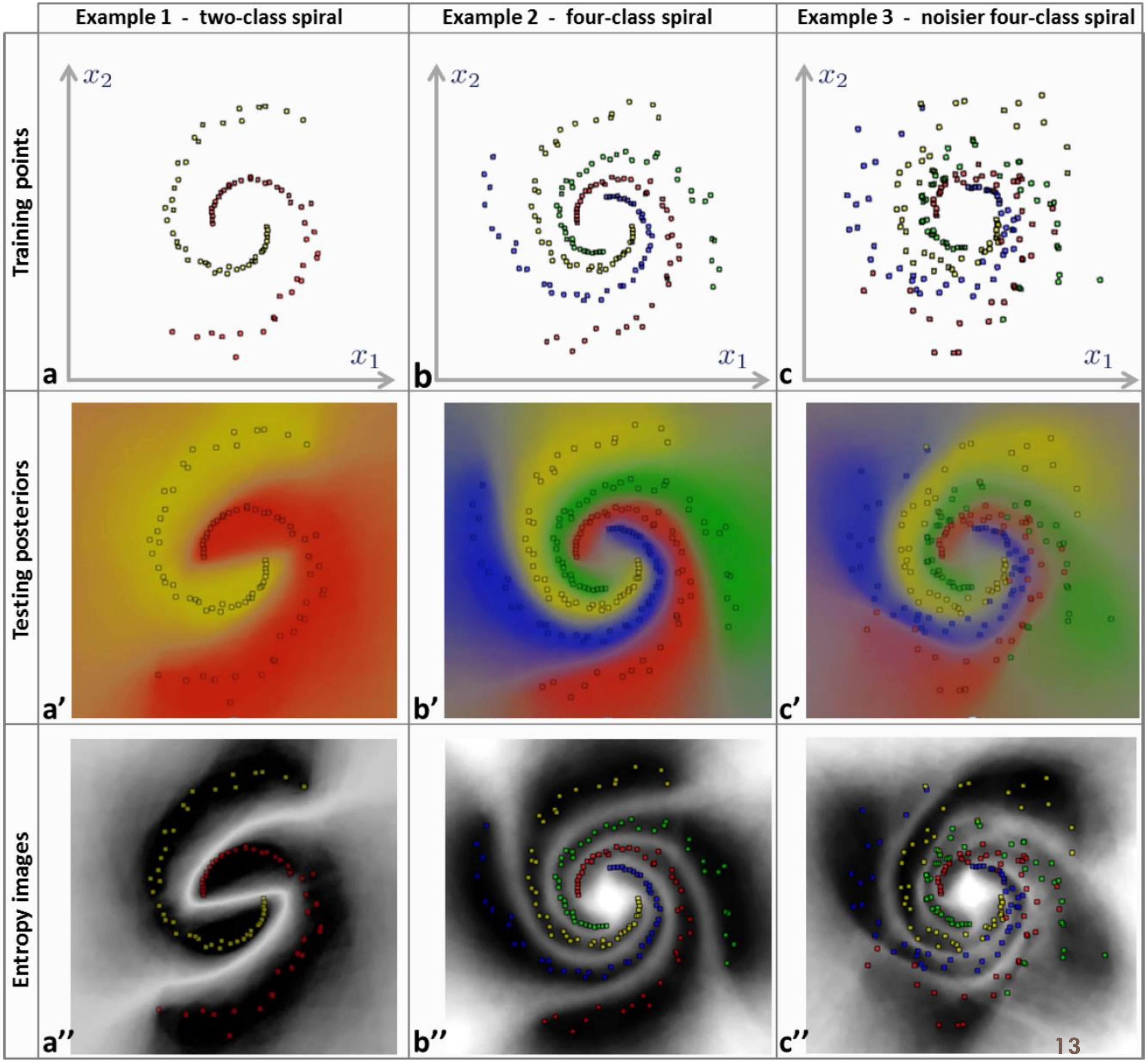


(a)

(b)

(c)

Plot from [Criminisi et al. 2012]



Plot from [Criminisi et al. 2012]