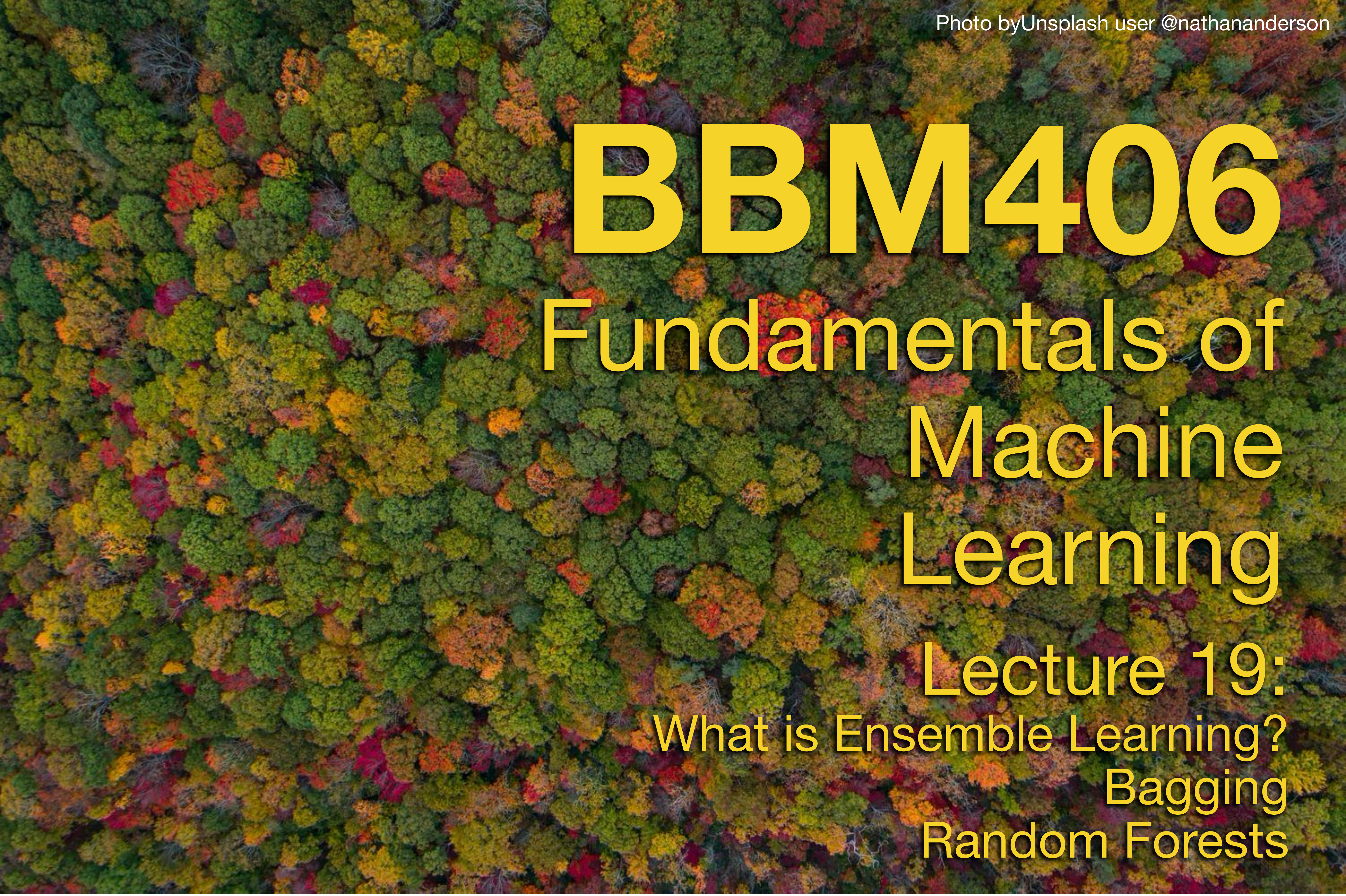


Photo byUnsplash user @nathananderson

BBM406

Fundamentals of Machine Learning

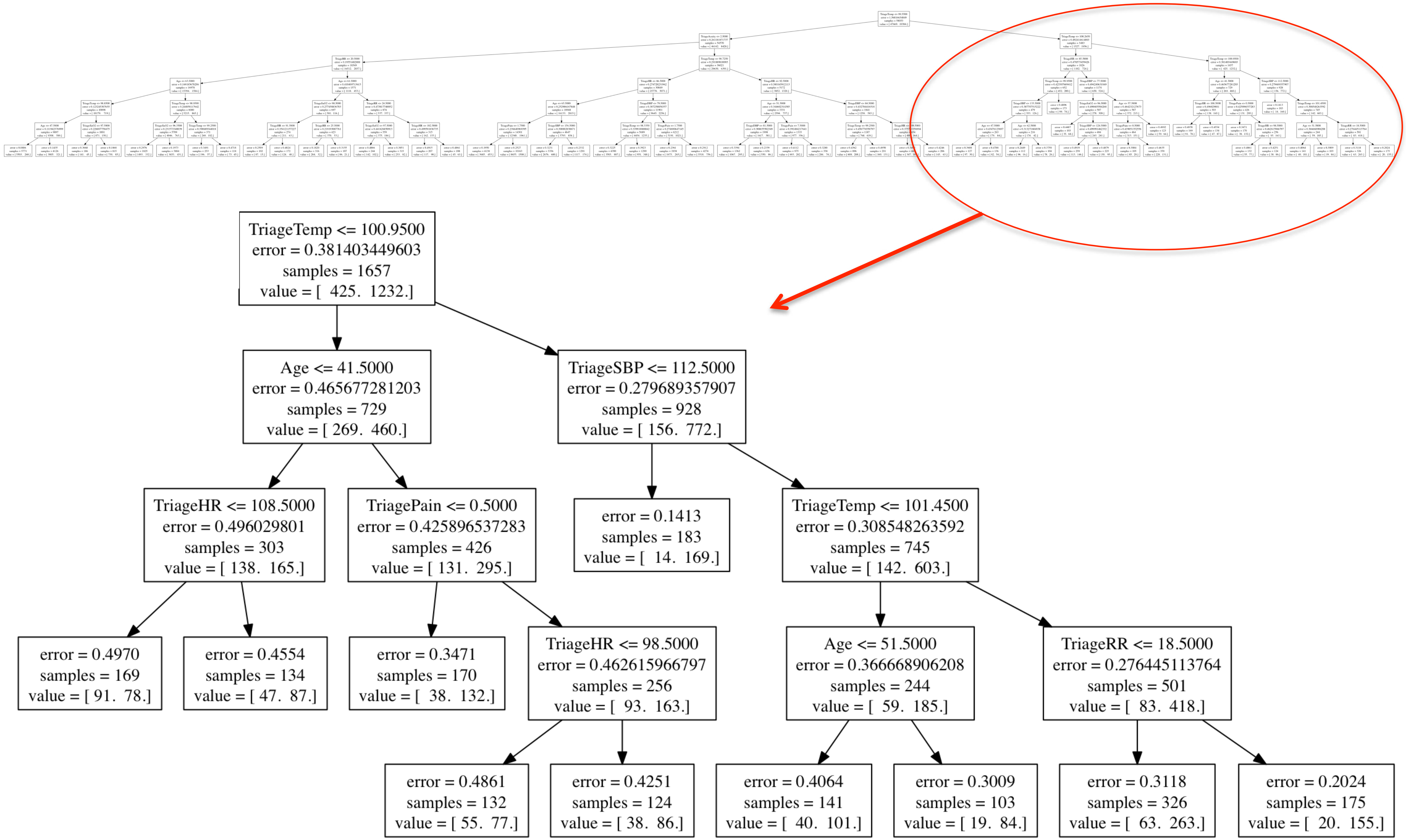
Lecture 19:

What is Ensemble Learning?
Bagging
Random Forests

HACETTEPE UNIVERSITY COMPUTER VISION LAB

Erkut Erdem // Hacettepe University // Spring 2021

Last time… **Decision Trees**



slide by David Sontag

Last time… Information Gain

- Decrease in entropy (uncertainty) after splitting

$$IG(X) = H(Y) - H(Y \mid X)$$

In our running example:

$IG(X_1) = H(Y) – H(Y|X_1)$
$= 0.65 – 0.33$

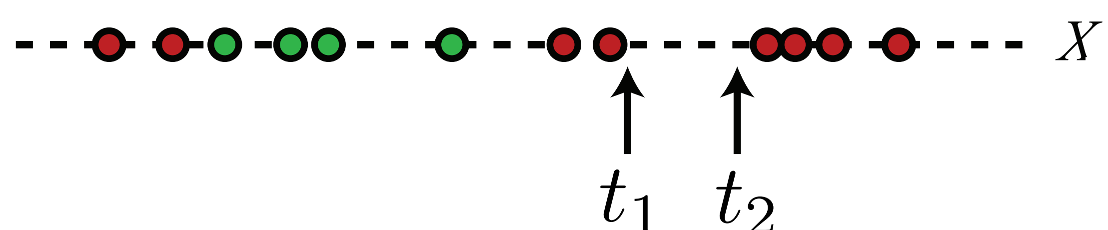
$IG(X_1) > 0 \rightarrow$ we prefer the split!

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F

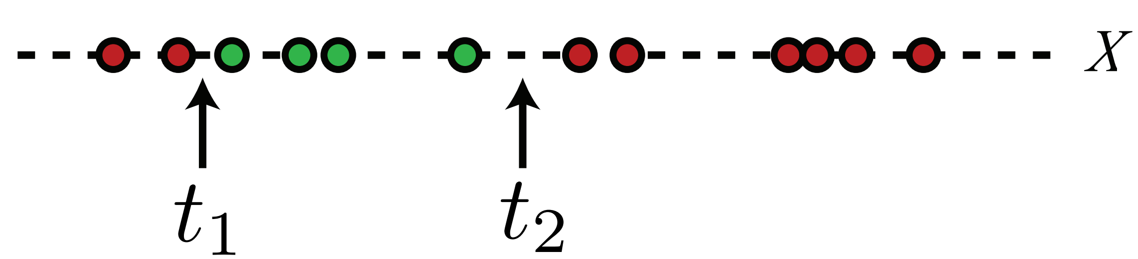
slide by David Sontag

Last time… Continuous features

- Binary tree, split on attribute X
  - One branch: $X < t$
  - Other branch: $X \geq t$
- Search through possible values of t
  - Seems hard!!!
- But only a finite number of t's are important:



- Sort data according to X into $\{x_1,\ldots,x_m\}$
- Consider split points of the form $x_i + (x_{i+1} - x_i)/2$
- Moreover, only splits between examples from different classes matter!



slide by David Sontag

Last time… **Decision trees will overfit**

- Standard decision trees have no learning bias
  - Training set error is always zero!
    - (If there is no label noise)
  - Lots of variance
  - Must introduce some bias towards simpler trees

- Many strategies for picking simpler trees
  - Fixed depth
  - Fixed number of leaves

- Random forests

slide by David Sontag

Today

- Ensemble Methods
  - Bagging
    - Random Forests

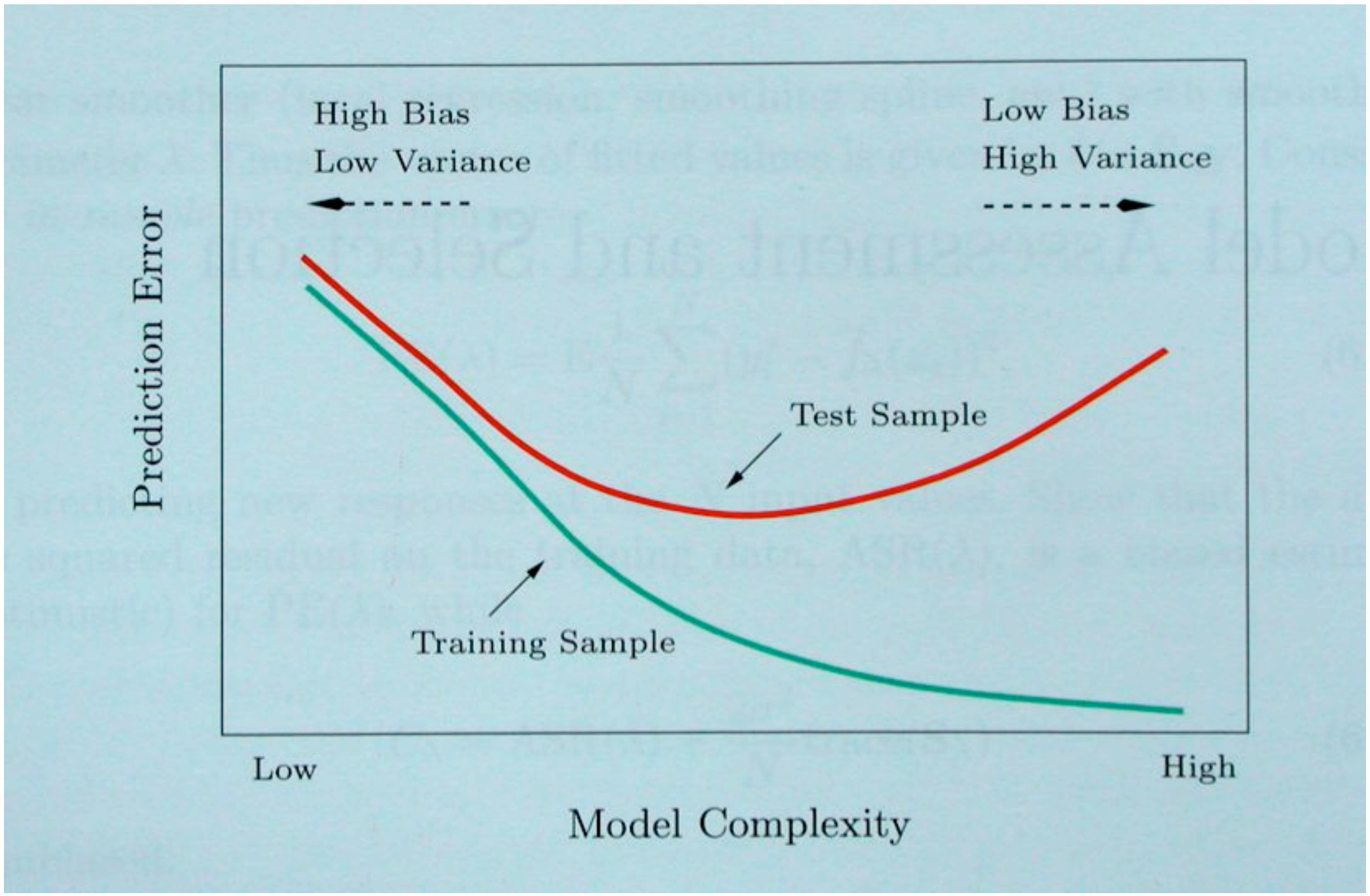
Ensemble Methods

- High level idea
  - Generate multiple hypotheses
  - Combine them to a single classifier

- Two important questions
  - How do we generate multiple hypotheses
    - we have only one sample
  - How do we combine the multiple hypotheses
    - Majority, AdaBoost, ...

slide by Yishay Mansour

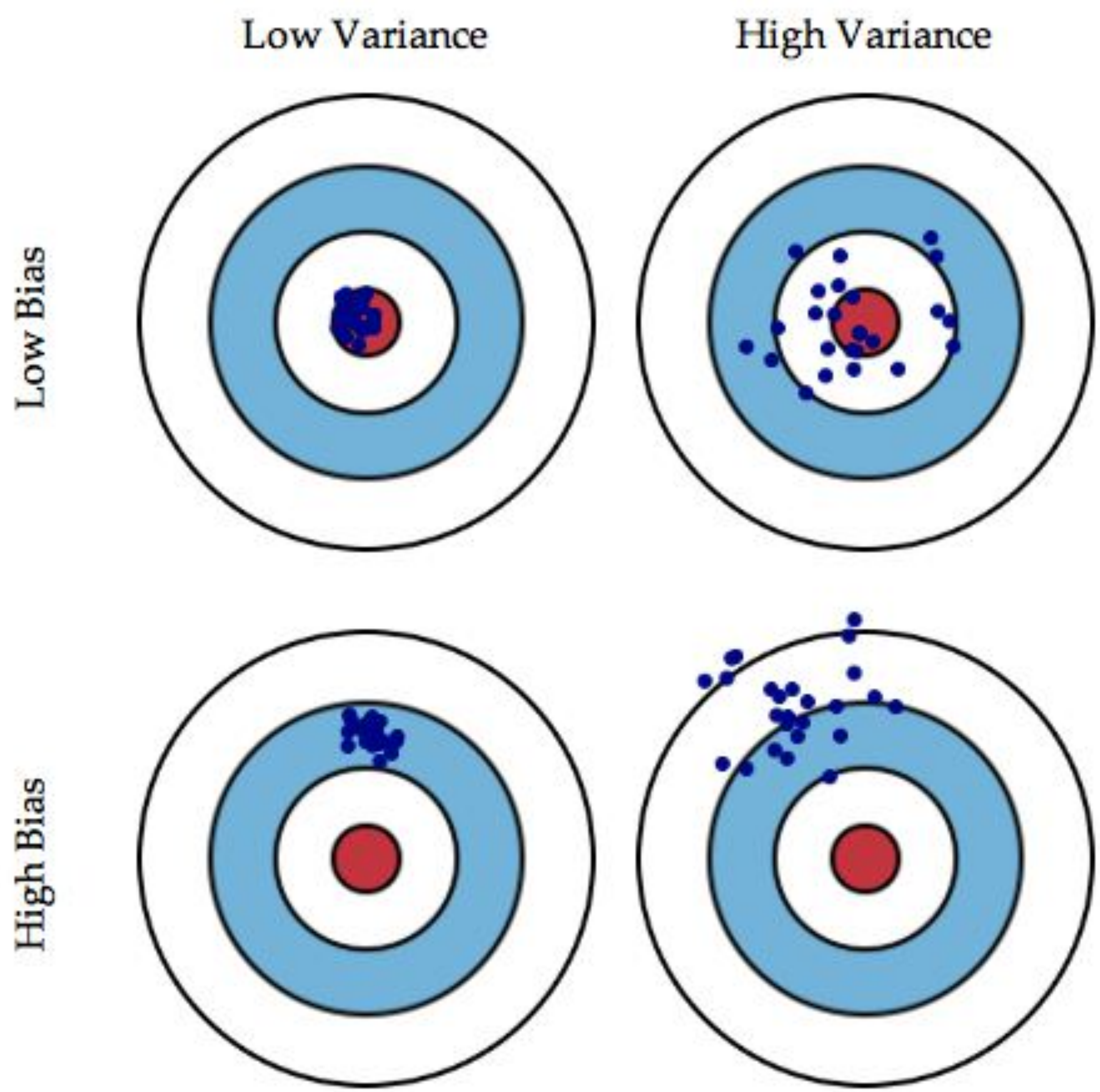
Bias/Variance Tradeoff



Hastie, Tibshirani, Friedman “Elements of Statistical Learning” 2001

slide by David Sontag

Bias/Variance Tradeoff



Graphical illustration of bias and variance.

http://scott.fortmann-roe.com/docs/BiasVariance.html

slide by David Sontag

Fighting the bias-variance tradeoff

- **Simple (a.k.a. weak) learners are good**
  - e.g., naïve Bayes, logistic regression, decision stumps (or shallow decision trees)
  - Low variance, don't usually overfit
- **Simple (a.k.a. weak) learners are bad**
  - High bias, can't solve hard learning problems

slide by Aarti Singh

Reduce Variance Without Increasing Bias

- **Averaging** reduces variance:

$$Var(\overline{X}) = \frac{Var(X)}{N}$$

(when prediction are **independent**)

- Average models to reduce model variance
- One problem:
  - Only one training set
  - Where do multiple models come from?

slide by David Sontag

Bagging (Bootstrap Aggregating)

- Leo Breiman (1994)
- Take repeated bootstrap samples from training set D.
- **Bootstrap sampling:** Given set D containing N training examples, create D′ by drawing N examples at random with replacement from D.

- Bagging:
  - Create *k* bootstrap samples D_1 ... D_k.
  - Train distinct classifier on each D_i.
  - Classify new instance by majority vote / average.

slide by David Sontag

Bagging

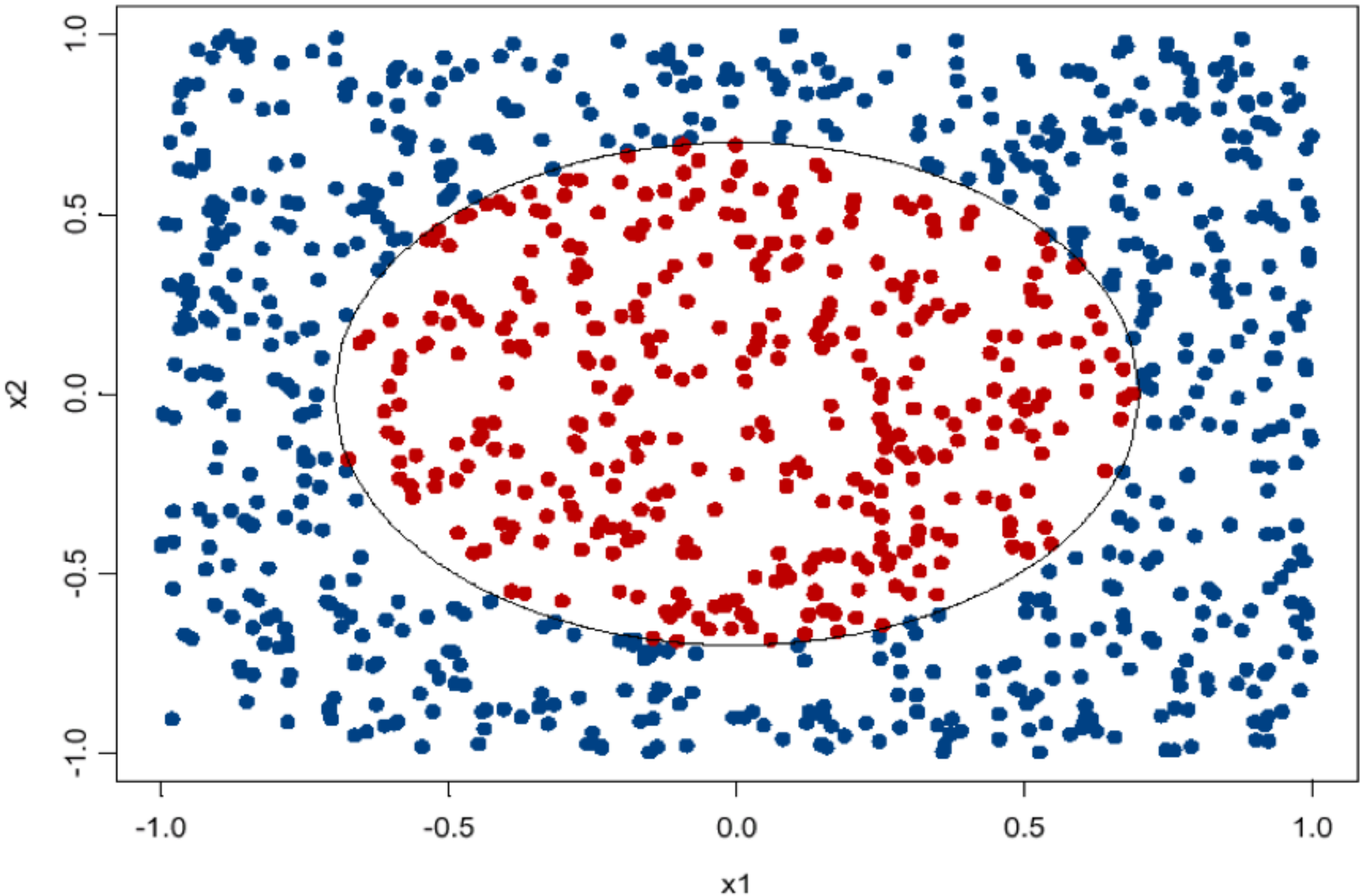
- Best case:

$$Var(Bagging(L(x, D))) = \frac{Var(L(x, D))}{N}$$

- In practice:
    - models are correlated, so reduction is smaller than 1/N
    - variance of models trained on fewer training cases usually somewhat larger

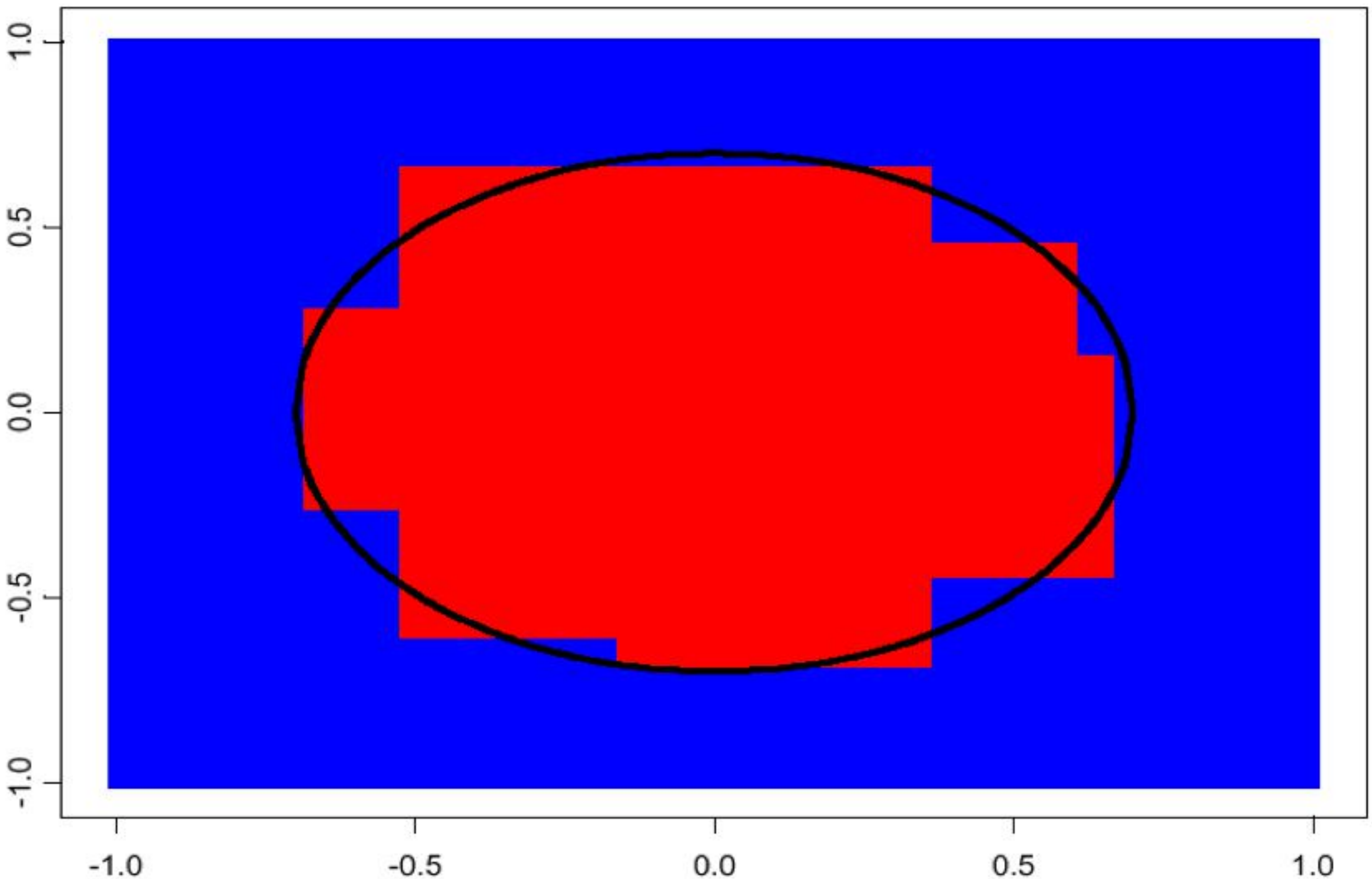
slide by David Sontag

Bagging Example



slide by David Sontag

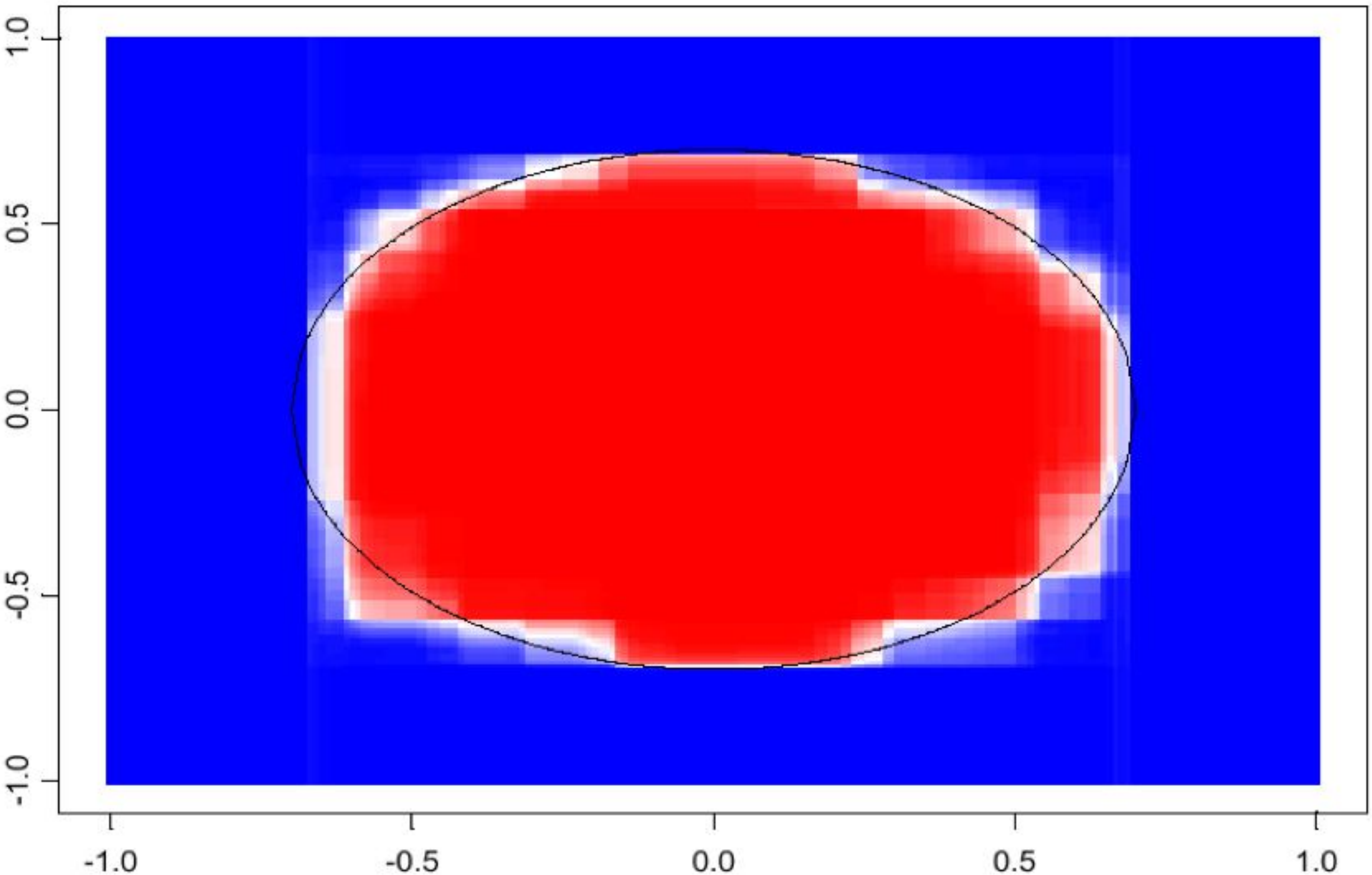
CART* decision boundary



* A decision tree learning algorithm; very similar to ID3

slide by David Sontag

100 bagged trees



slide by David Sontag

- Shades of blue/red indicate strength of vote for particular classification

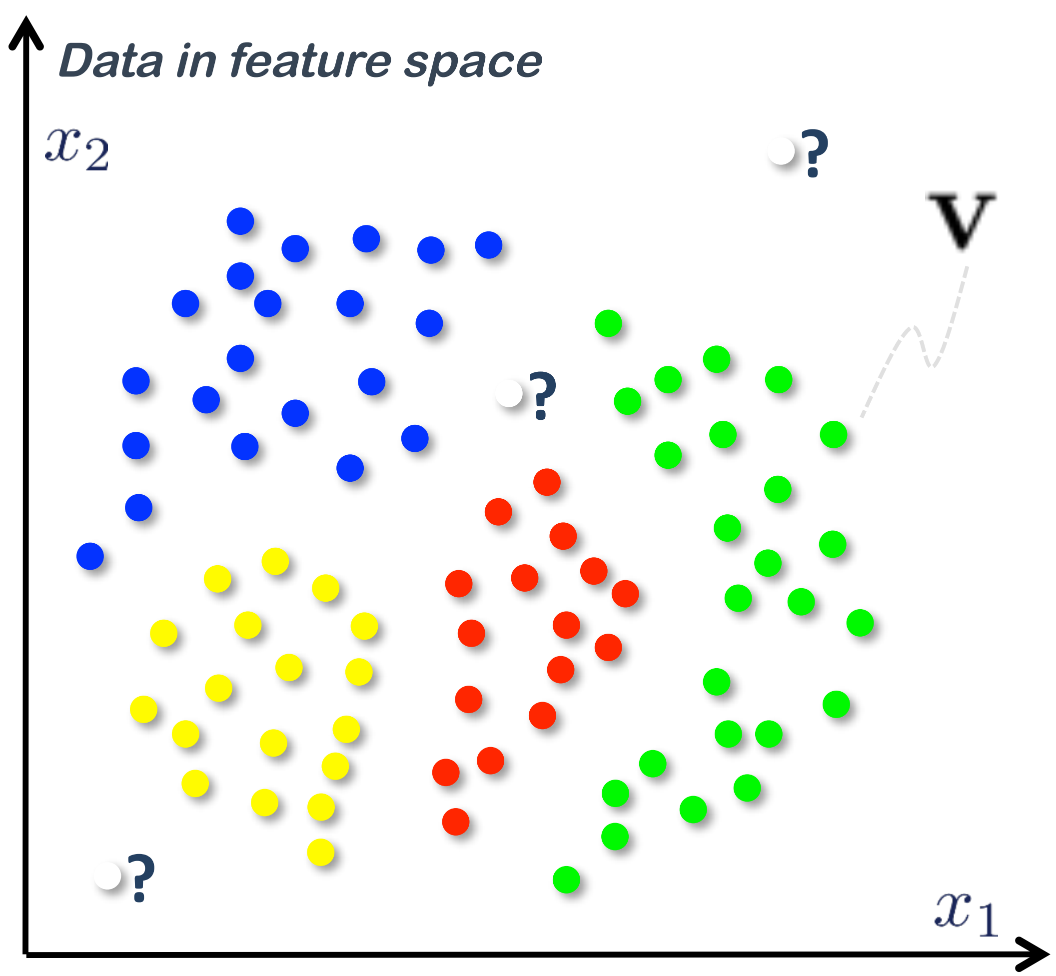
Random Forests

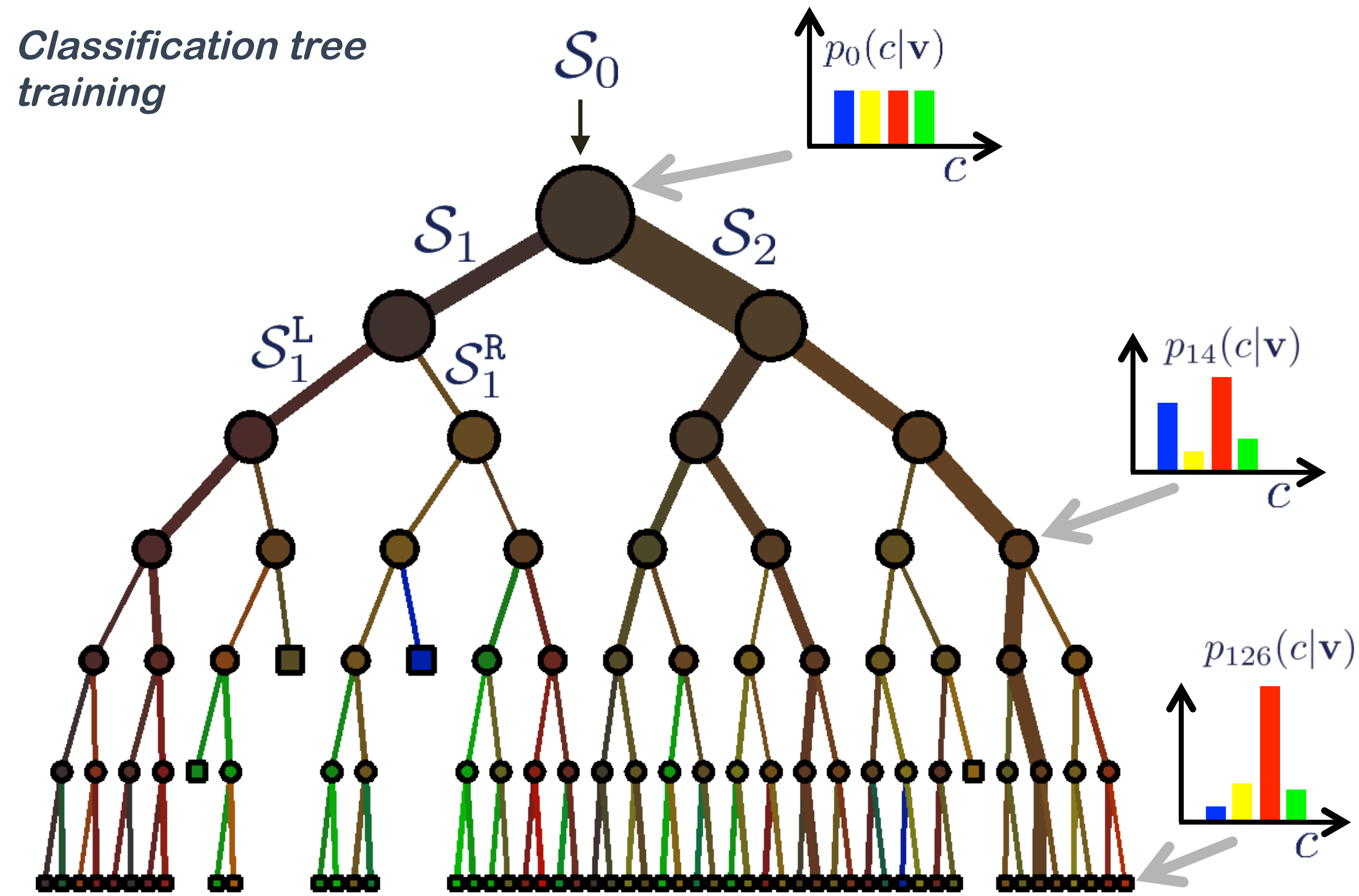
Random Forests

- Ensemble method specifically designed for decision tree classifiers
- Introduce two sources of randomness: “Bagging” and “Random input vectors”
  - **Bagging method:** each tree is grown using a bootstrap sample of training data
  - **Random vector method:** **At each node**, best split is chosen from a random sample of m attributes instead of all attributes

slide by David Sontag

Classification tree





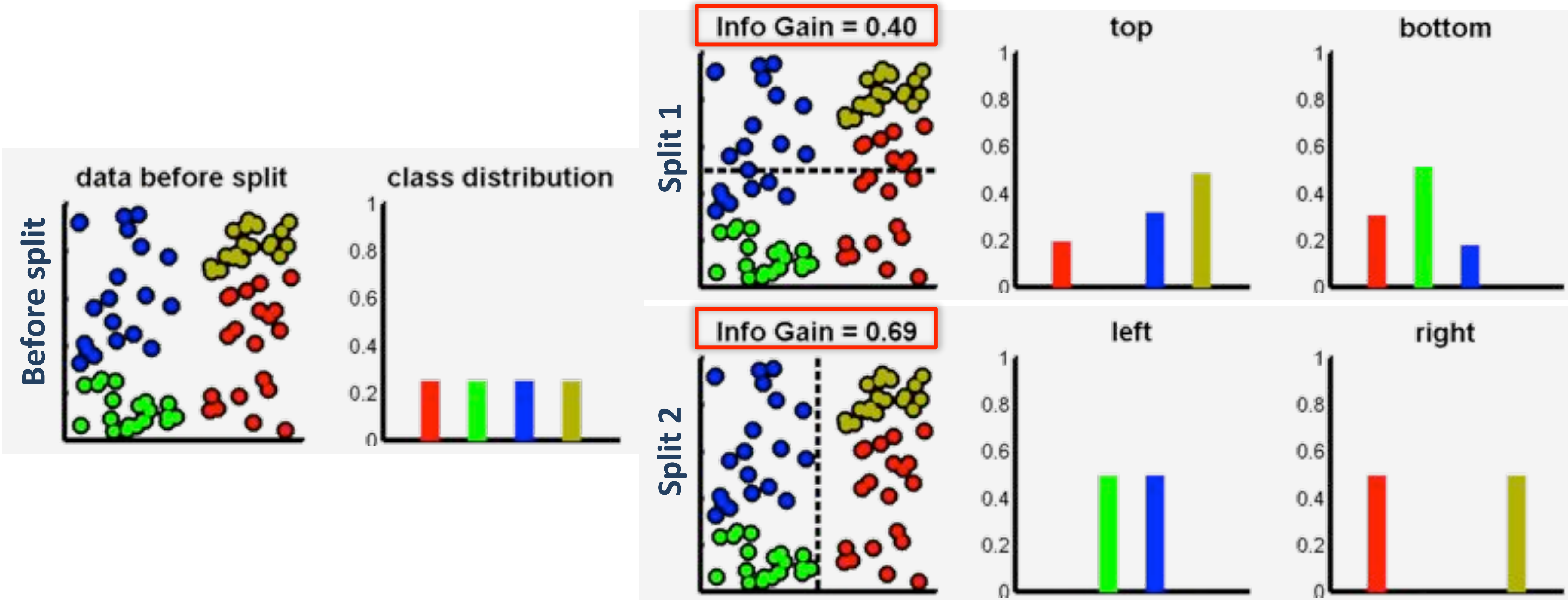
A generic data point is denoted by a vector $\mathbf{v} = (x_1, x_2, \cdots, x_d)$

$$\mathcal{S}_j = \mathcal{S}_j^{\mathrm{L}} \cup \mathcal{S}_j^{\mathrm{R}}$$

slide by Nando de Freitas

[Criminisi et al., 2011]

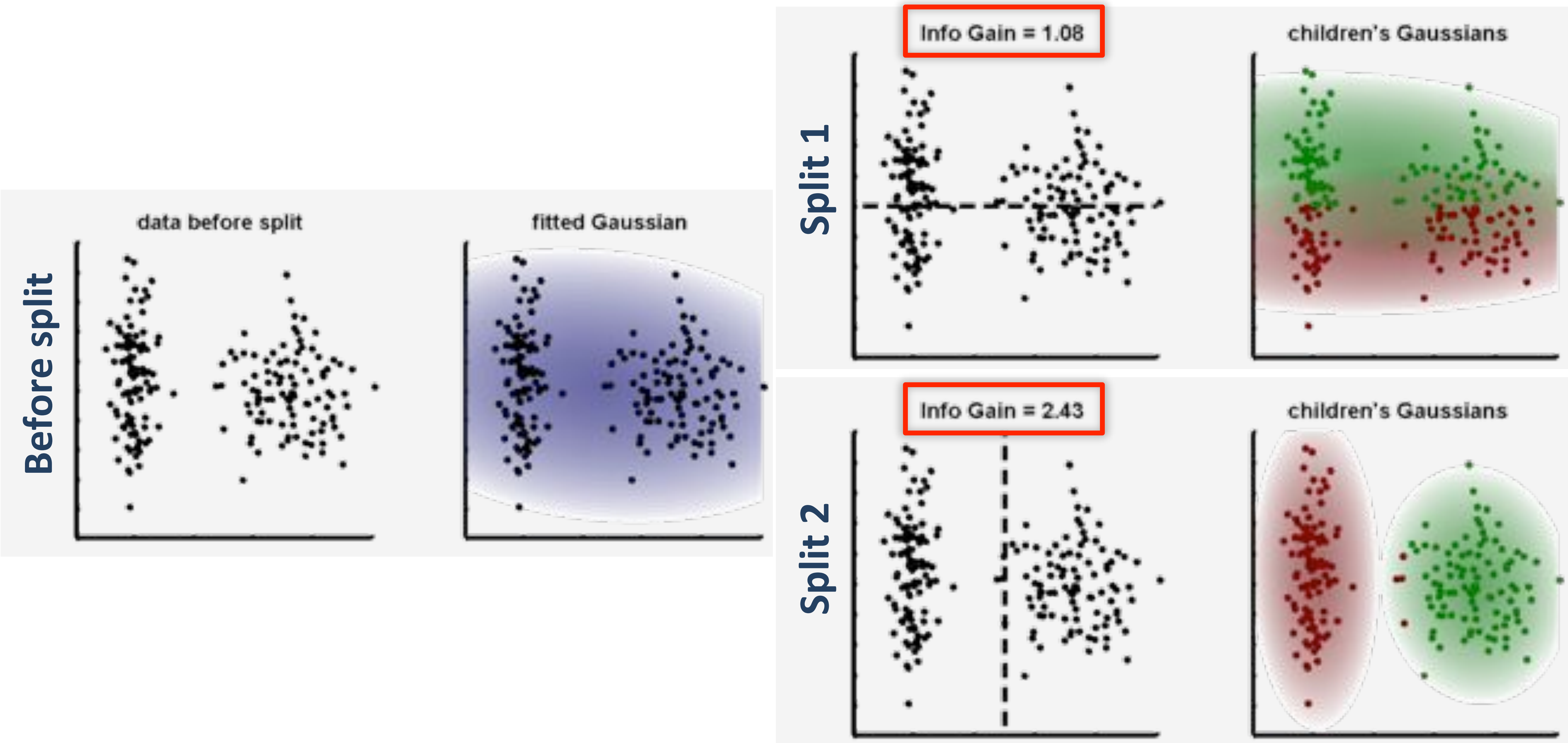
Use information gain to decide splits



$$I_j = H(\mathcal{S}_j) - \sum_{i \in \{L,R\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} H(\mathcal{S}_j^i)$$

[Criminisi et al., 2011]

Advanced: Gaussian information gain to decide splits



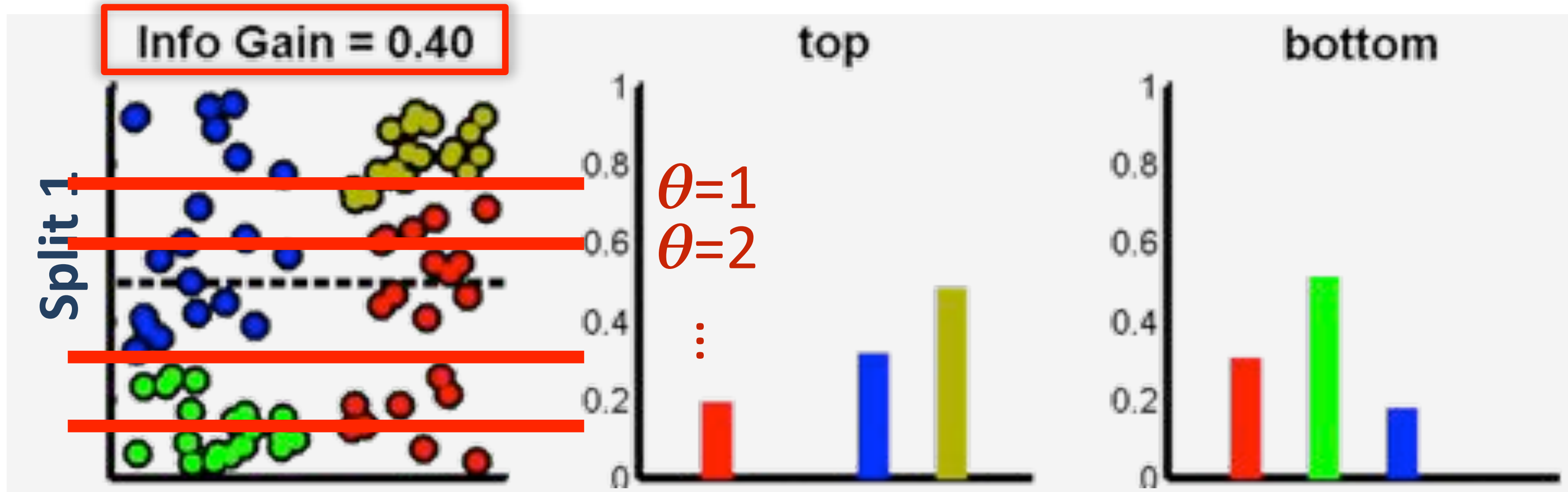
$$H(\mathcal{S}) = \frac{1}{2}\log\left((2\pi e)^d |\Lambda(\mathcal{S})|\right)$$

[Criminisi et al., 2011]

slide by Nando de Freitas

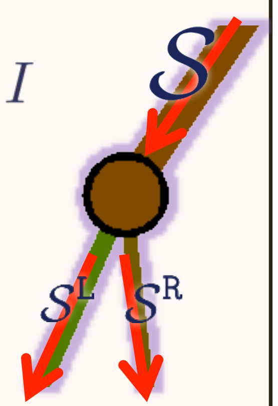
Each split node j is associated with a binary split function

$$h(\mathbf{v}, \boldsymbol{\theta}) \in \{\mathtt{true}, \mathtt{false}\}$$



Split node (train)

$$\boldsymbol{\theta}_j = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}_j} I$$

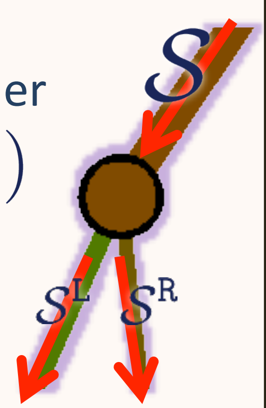


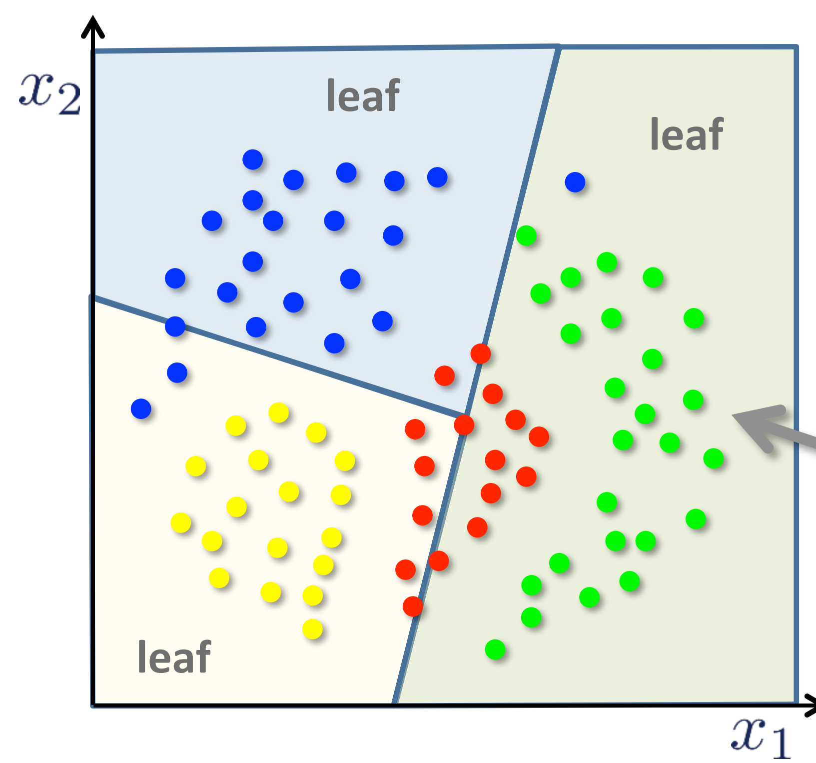
Split node (test)

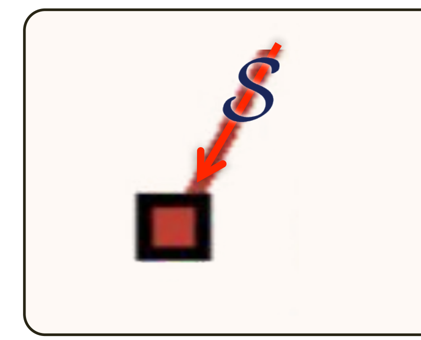
Node weak learner

$$h(\mathbf{v}, \boldsymbol{\theta}_j)$$

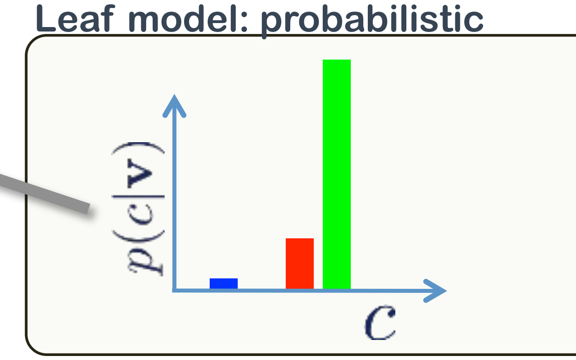
$$\boldsymbol{\theta} \in \mathcal{T}_j$$







Leaf model: probabilistic



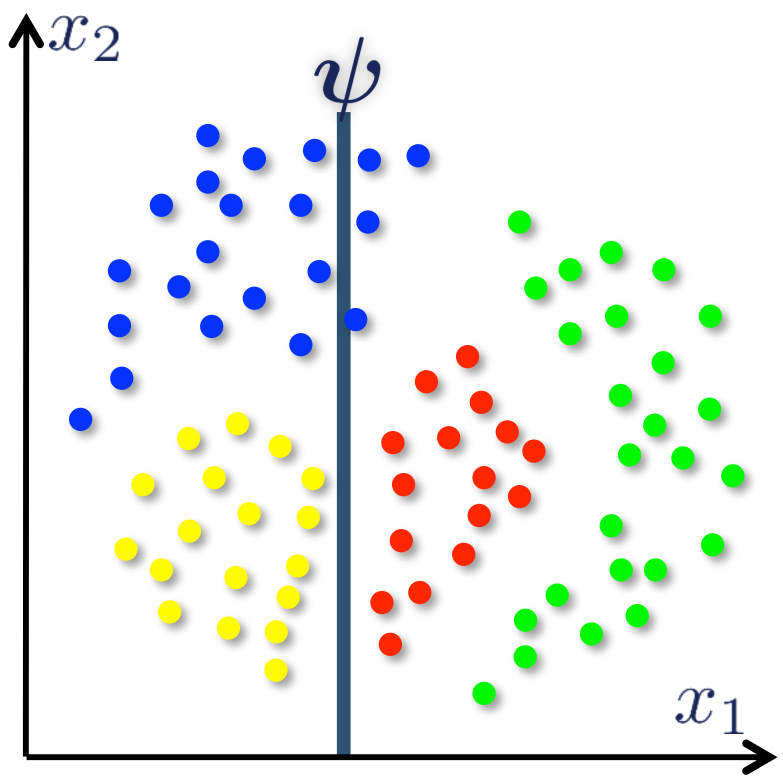
$$I_j = H(\mathcal{S}_j) - \sum_{i \in \{L,R\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} H(\mathcal{S}_j^i)$$

[Criminisi et al., 2011]

slide by Nando de Freitas

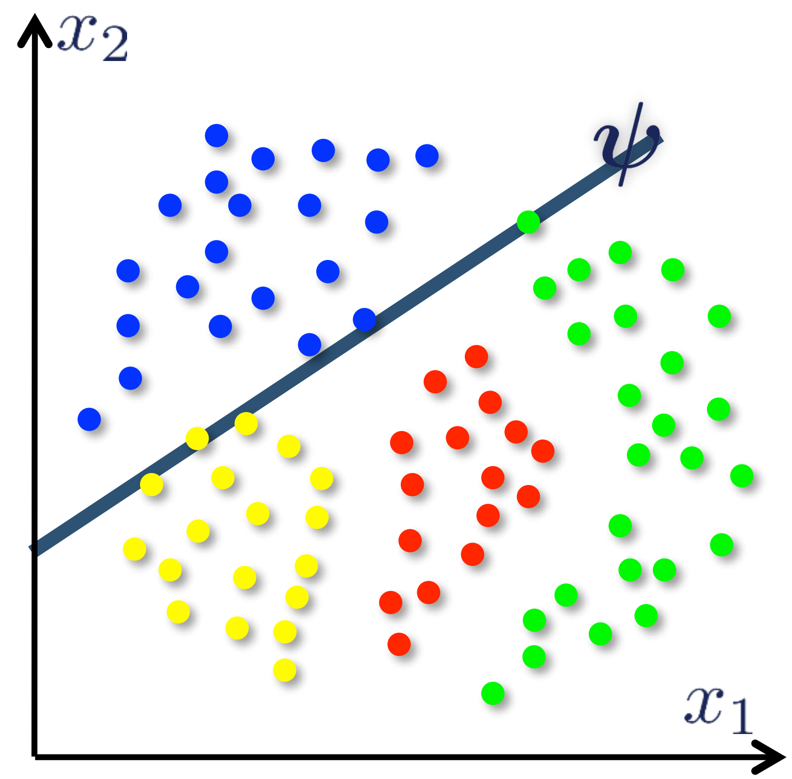
Alternative node decisions

$$\mathbf{v} = (x_1 \; x_2) \in \mathbb{R}^2 \qquad \boldsymbol{\phi}(\mathbf{v}) = (x_1 \; x_2 \; 1)^\top$$



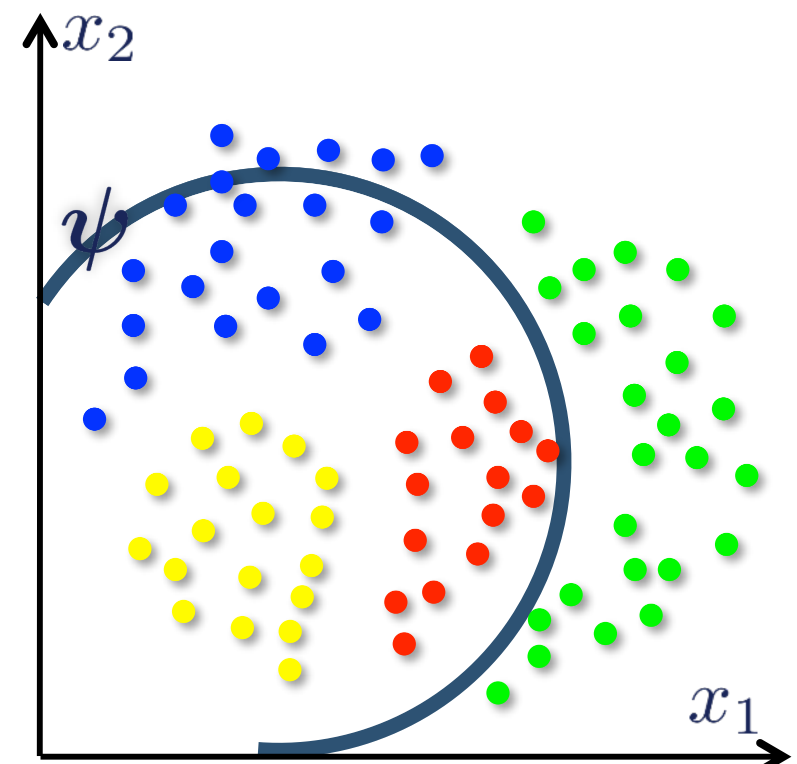
$h(\mathbf{v}, \boldsymbol{\theta}) = [\tau_1 > \boldsymbol{\phi}(\mathbf{v}) \cdot \boldsymbol{\psi} > \tau_2]$

axis aligned



$h(\mathbf{v}, \boldsymbol{\theta}) = [\tau_1 > \boldsymbol{\phi}(\mathbf{v}) \cdot \boldsymbol{\psi} > \tau_2]$

oriented line



$h(\mathbf{v}, \boldsymbol{\theta}) = \left[\tau_1 > \boldsymbol{\phi}^\top(\mathbf{v}) \; \boldsymbol{\psi} \; \boldsymbol{\phi}(\mathbf{v}) > \tau_2\right]$

conic section

examples of weak learners

slide by Nando de Freitas

Building a random tree

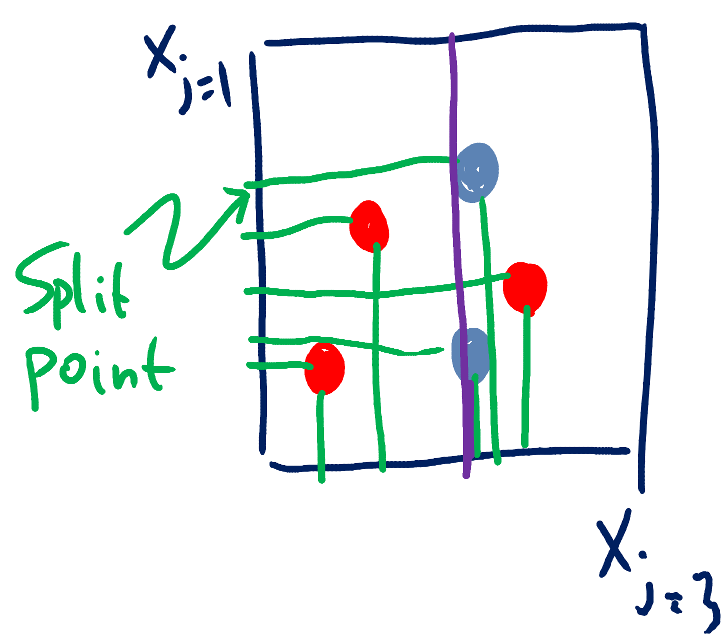
d=3 features

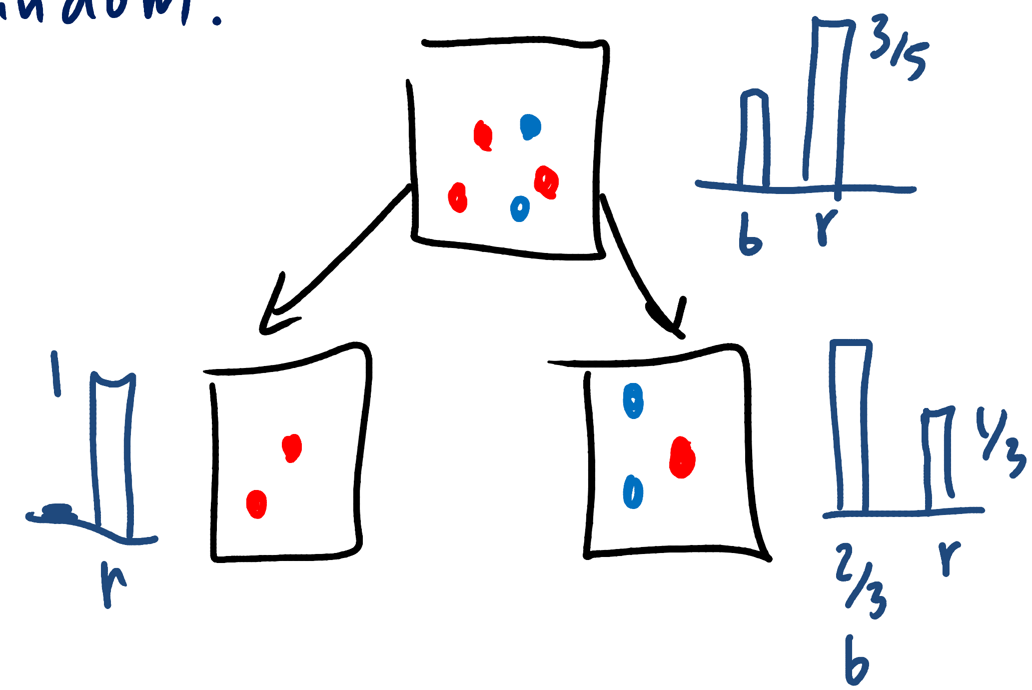
n=5 data

$$X = \begin{array}{c} \\ j{=}1 \\ j{=}2 \\ j{=}3 \end{array} \begin{array}{c} \begin{array}{ccccc} i{=}1 & i{=}2 & & & i{=}n{=}5 \end{array} \\ \left[\begin{array}{ccccc} 1 & 3 & 0 & 8 & 5 \\ 0 & 6 & 2 & 9 & 5 \\ 2 & 1 & 4 & 0 & 1 \end{array}\right] \end{array}$$

$$Y = \begin{array}{c} \begin{array}{ccccc} i{=}1 & i{=}2 & & & i{=}5 \end{array} \\ \left[\begin{array}{ccccc} 0 & 1 & 0 & 1 & 0 \end{array}\right] \end{array}$$

Pick 2 features at random.





slide by Nando de Freitas

Random Forests algorithm

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

[From the book of Hastie, Friedman and Tibshirani]

slide by Nando de Freitas

Randomization

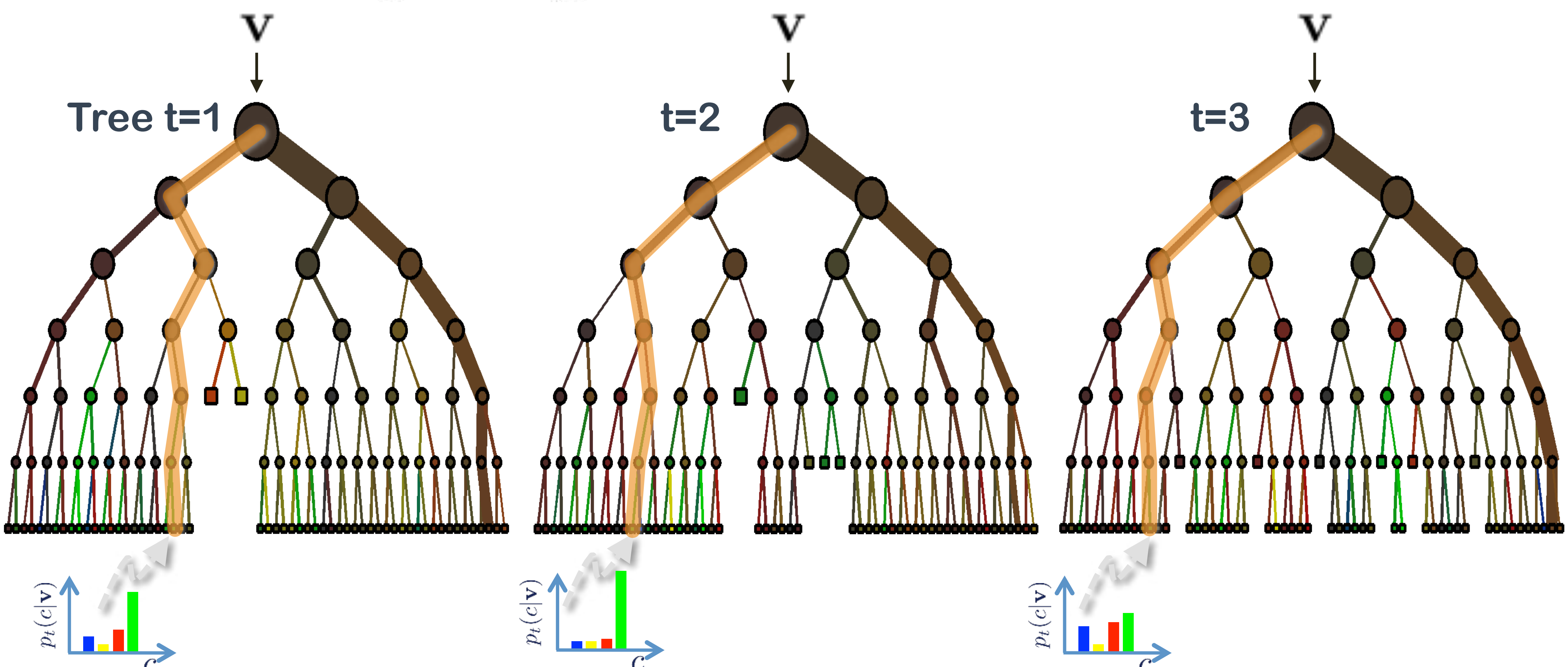
Randomized node optimization. If $\mathcal{T}$ is the entire set of all possible parameters $\boldsymbol{\theta}$ then when training the j^{th} node we only make available a small subset $\mathcal{T}_j \subset \mathcal{T}$ of such values.

$$\boldsymbol{\theta}_j^* = \arg\max_{\boldsymbol{\theta}_j \in \mathcal{T}_j} I_j.$$

slide by Nando de Freitas

Building a forest (ensemble)

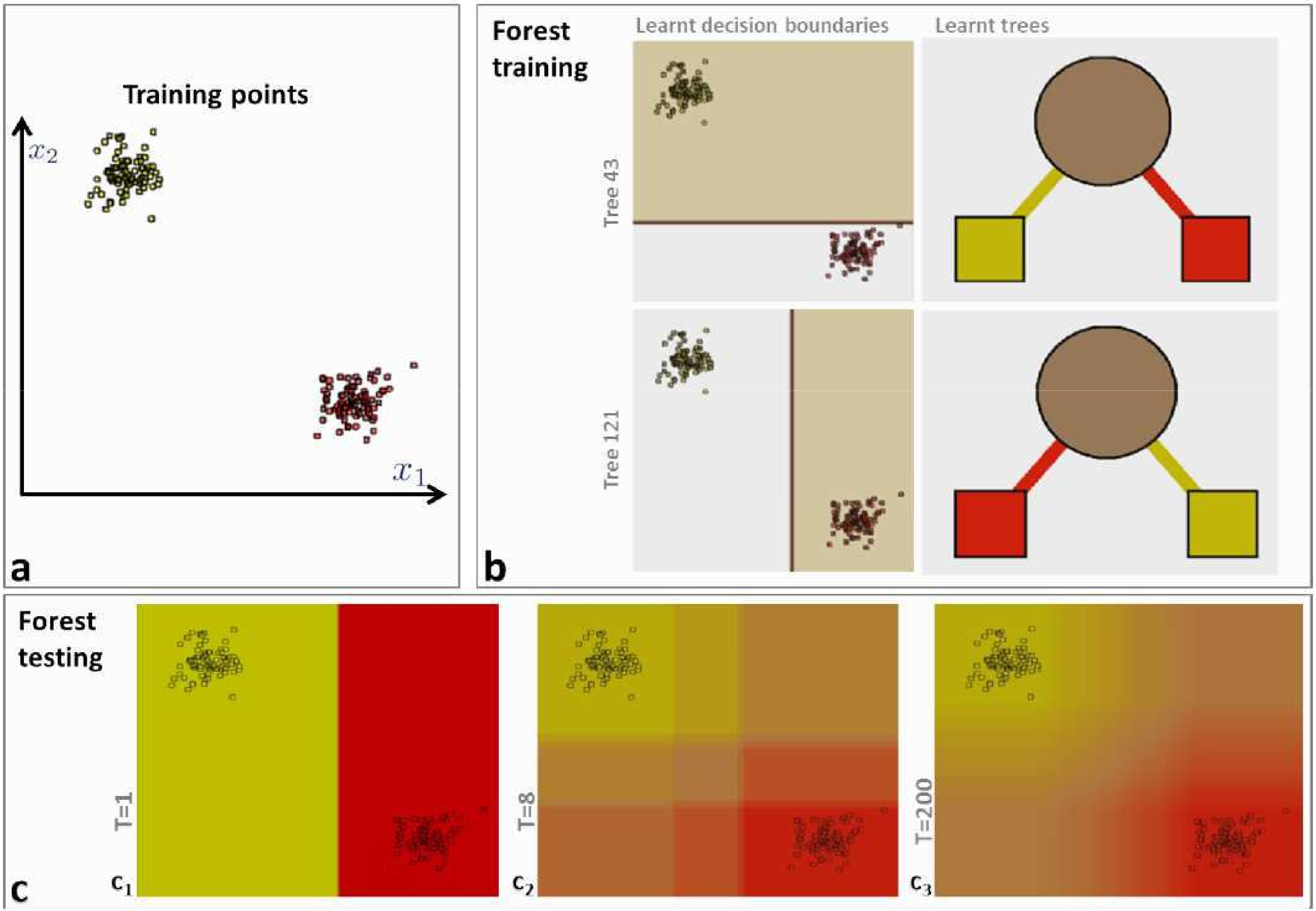
In a forest with T trees we have $t \in \{1, \cdots, T\}$. All trees are trained independently (and possibly in parallel). During testing, each test point $\mathbf{v}$ is simultaneously pushed through all trees (starting at the root) until it reaches the corresponding leaves.



$$p(c|\mathbf{v}) = \frac{1}{T} \sum_{t}^{T} p_t(c|\mathbf{v})$$

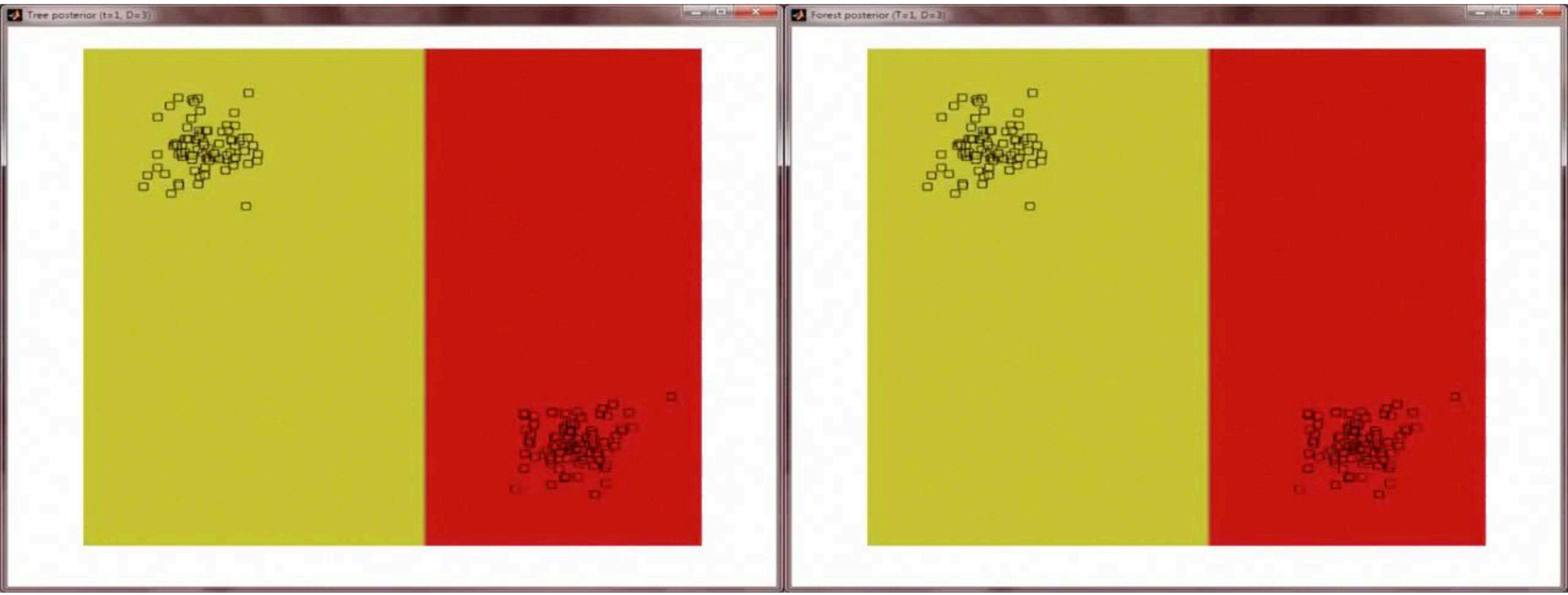
slide by Nando de Freitas

Effect of forest size

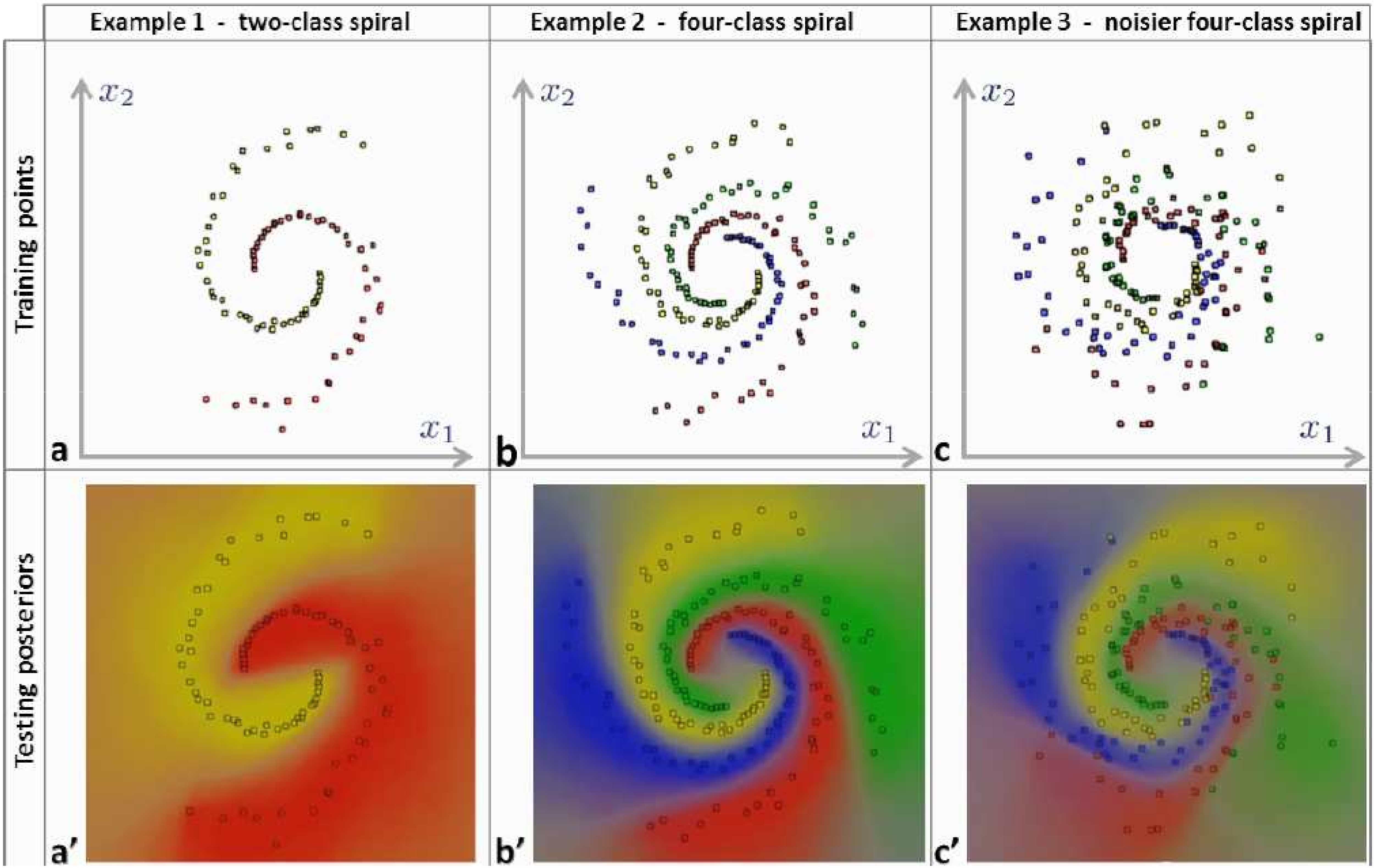


slide by Nando de Freitas

Effect of forest size

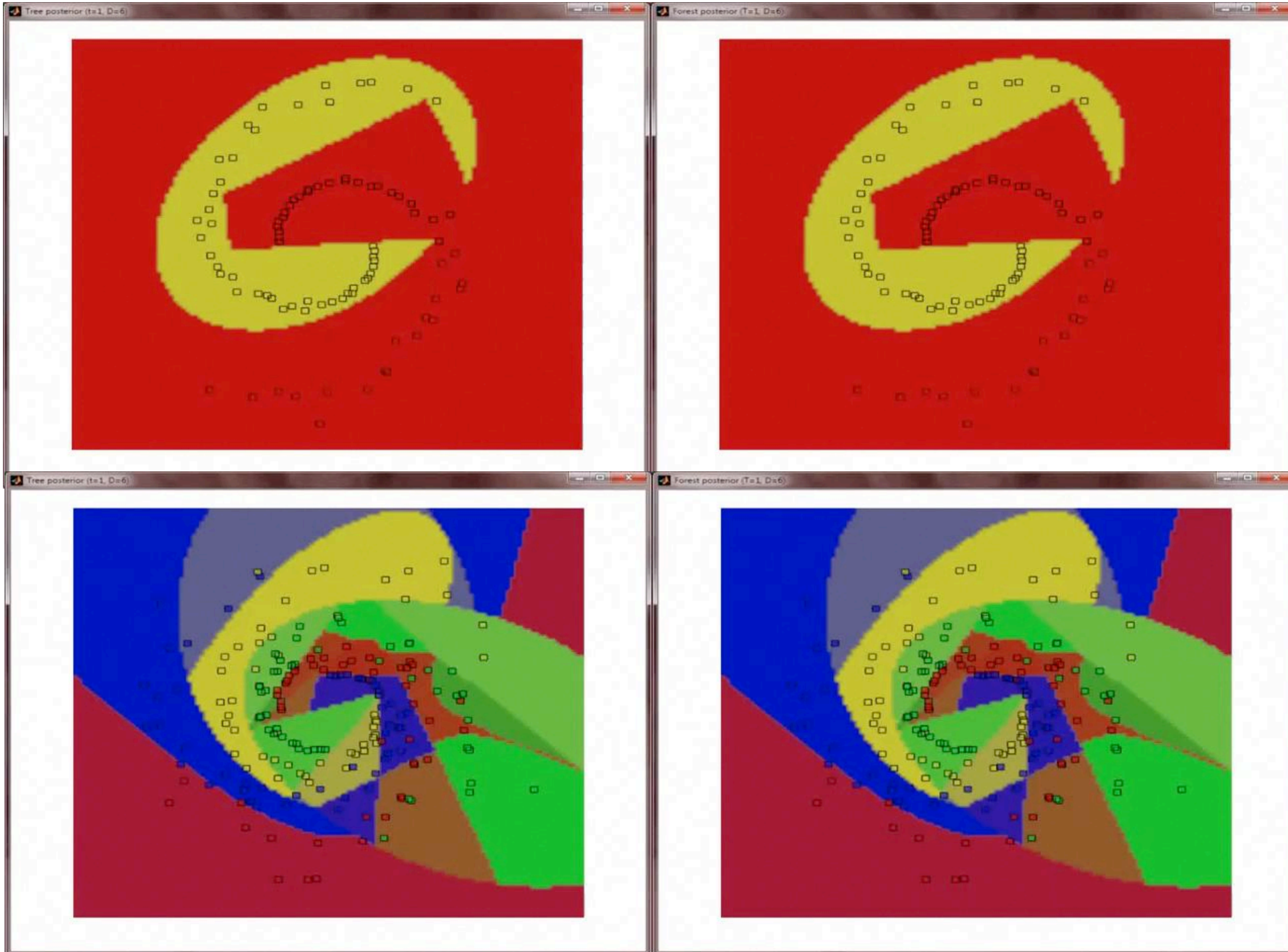


Effect of more classes and noise



slide by Nando de Freitas

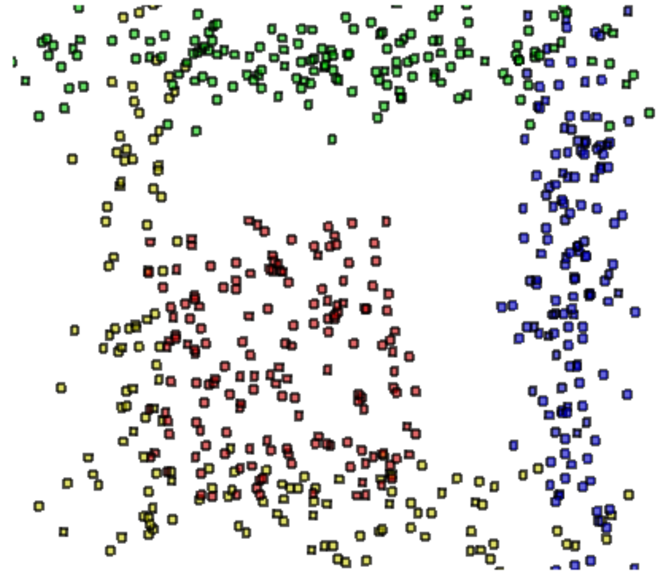
Effect of more classes and noise

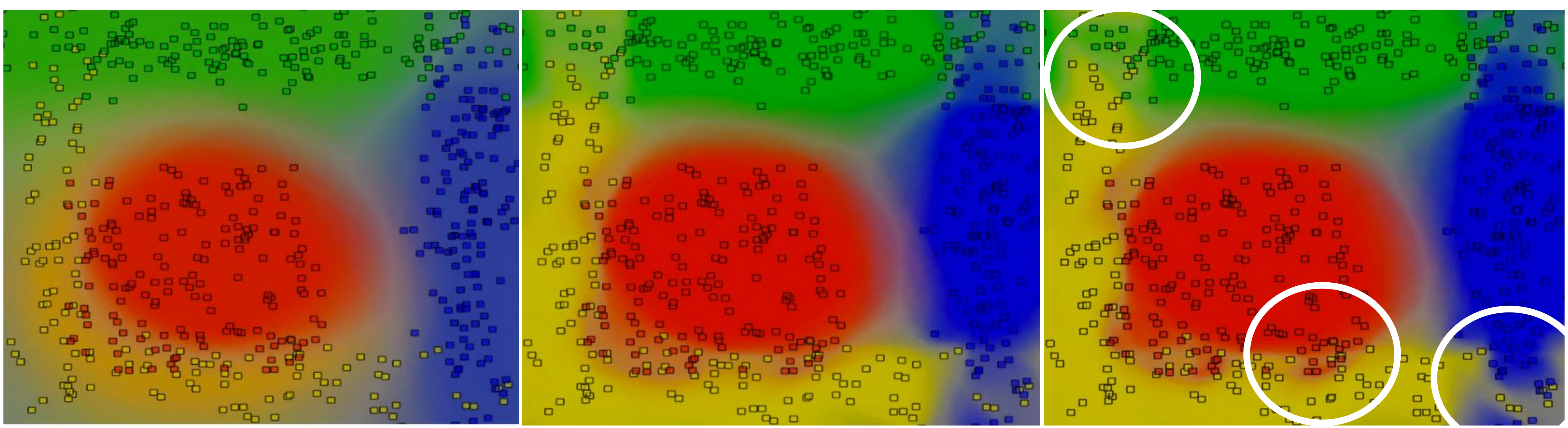


slide by Nando de Freitas

Effect of tree depth (D)

Training points: 4-class mixed

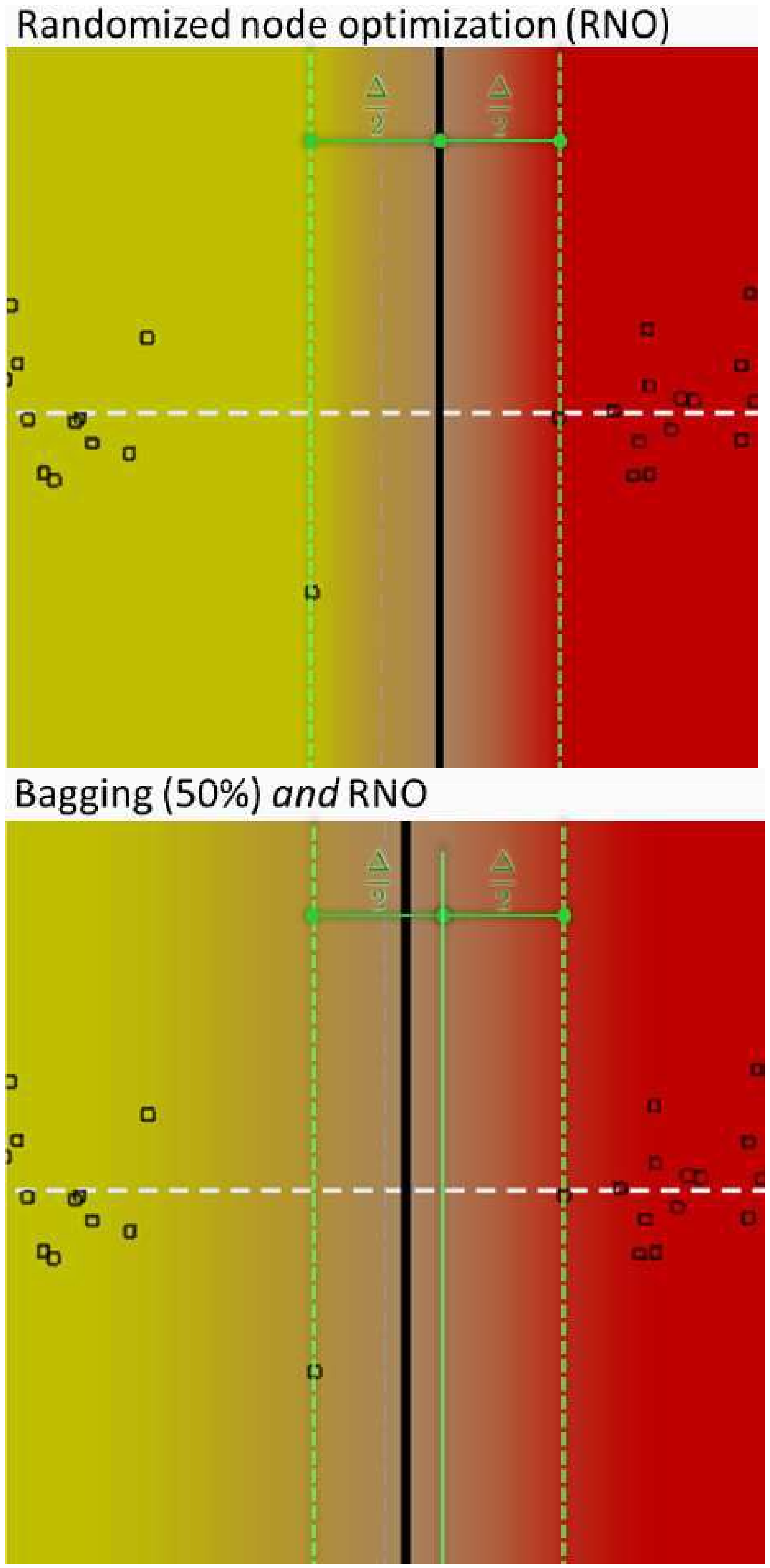




D=3 (underfitting) | D=6 | D=15 (overfitting)

slide by Nando de Freitas

Effect of bagging



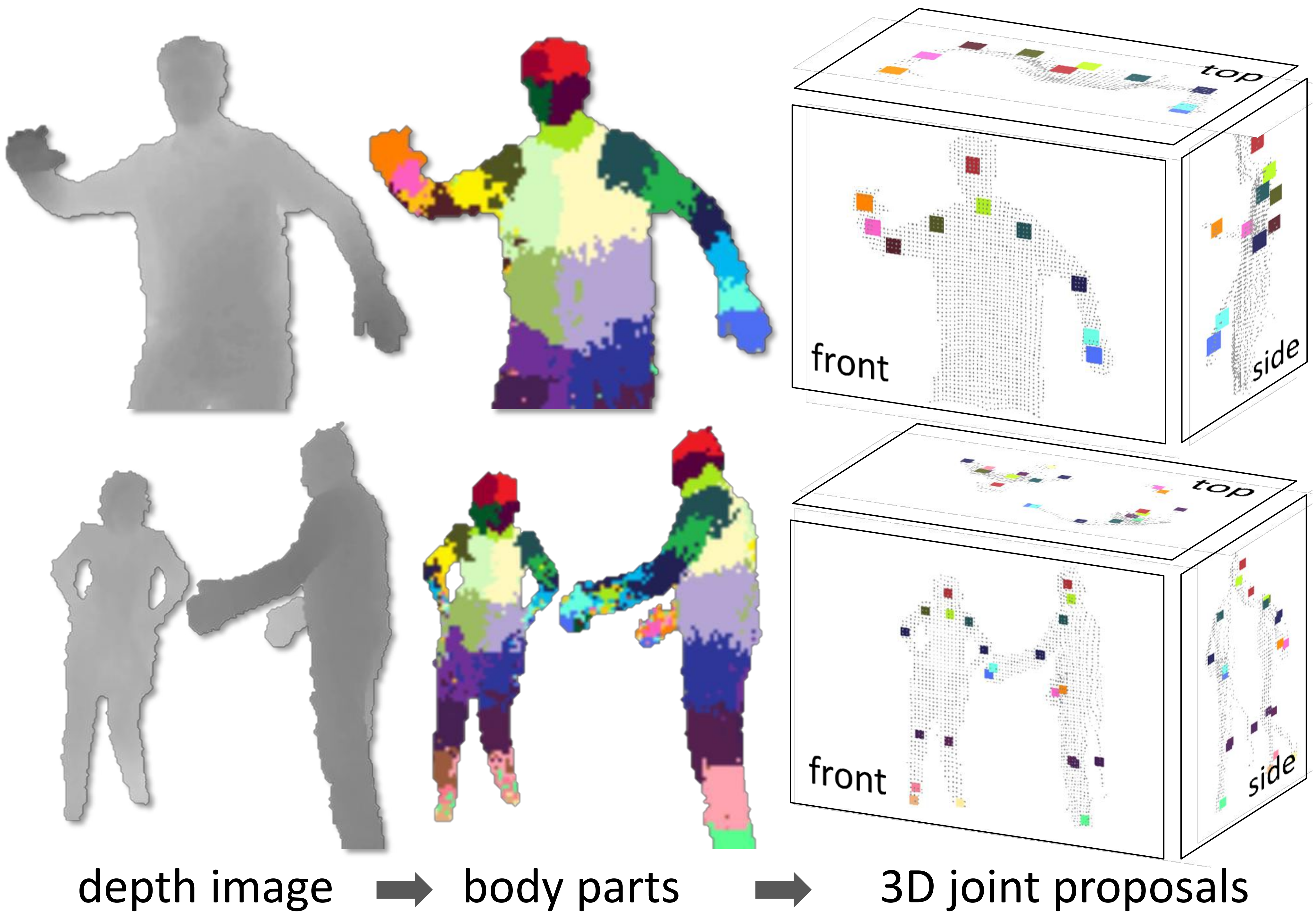
no bagging => max-margin

slide by Nando de Freitas

Random Forests and the Kinect

slide by Nando de Freitas

Random Forests and the Kinect

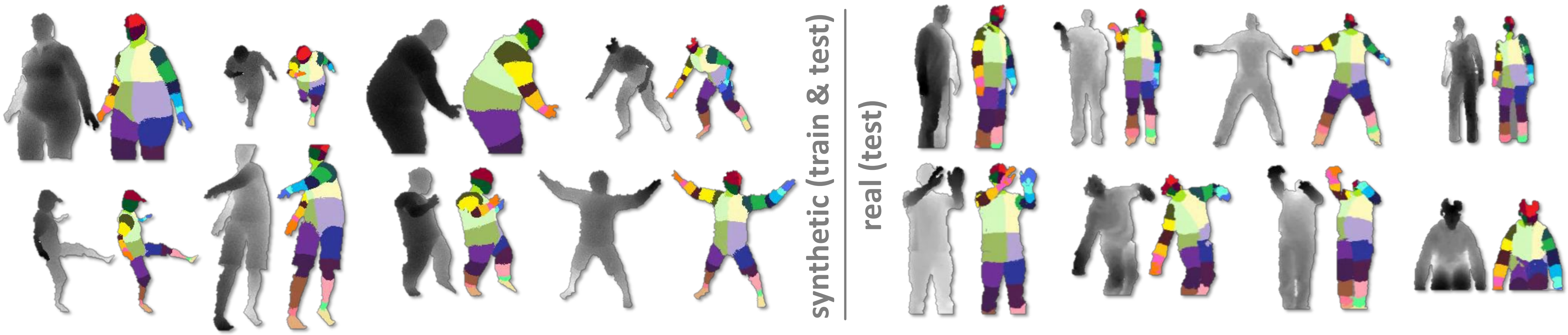


[Jamie Shotton et al., 2011]

adopted from Nando de Freitas

Random Forests and the Kinect

- Use computer graphics to generate plenty of data



Real-Time Human Pose Recognition in Parts from Single Depth Images

CVPR 2011

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, Andrew Blake
Microsoft Research Cambridge & Xbox Incubation

adopted from Nando de Freitas

[Jamie Shotton et al., 2011]

Reduce Bias^2 and Decrease Variance?

- Bagging reduces variance by averaging
- Bagging has little effect on bias
- Can we average *and* reduce bias?
- Yes: **Boosting**

slide by David Sontag

Next Lecture:

Boosting