Tree Depth in a Forest

Mark Segal
Center for Bioinformatics & Molecular Biostatistics
Division of Bioinformatics
Department of Epidemiology and Biostatistics
UCSF

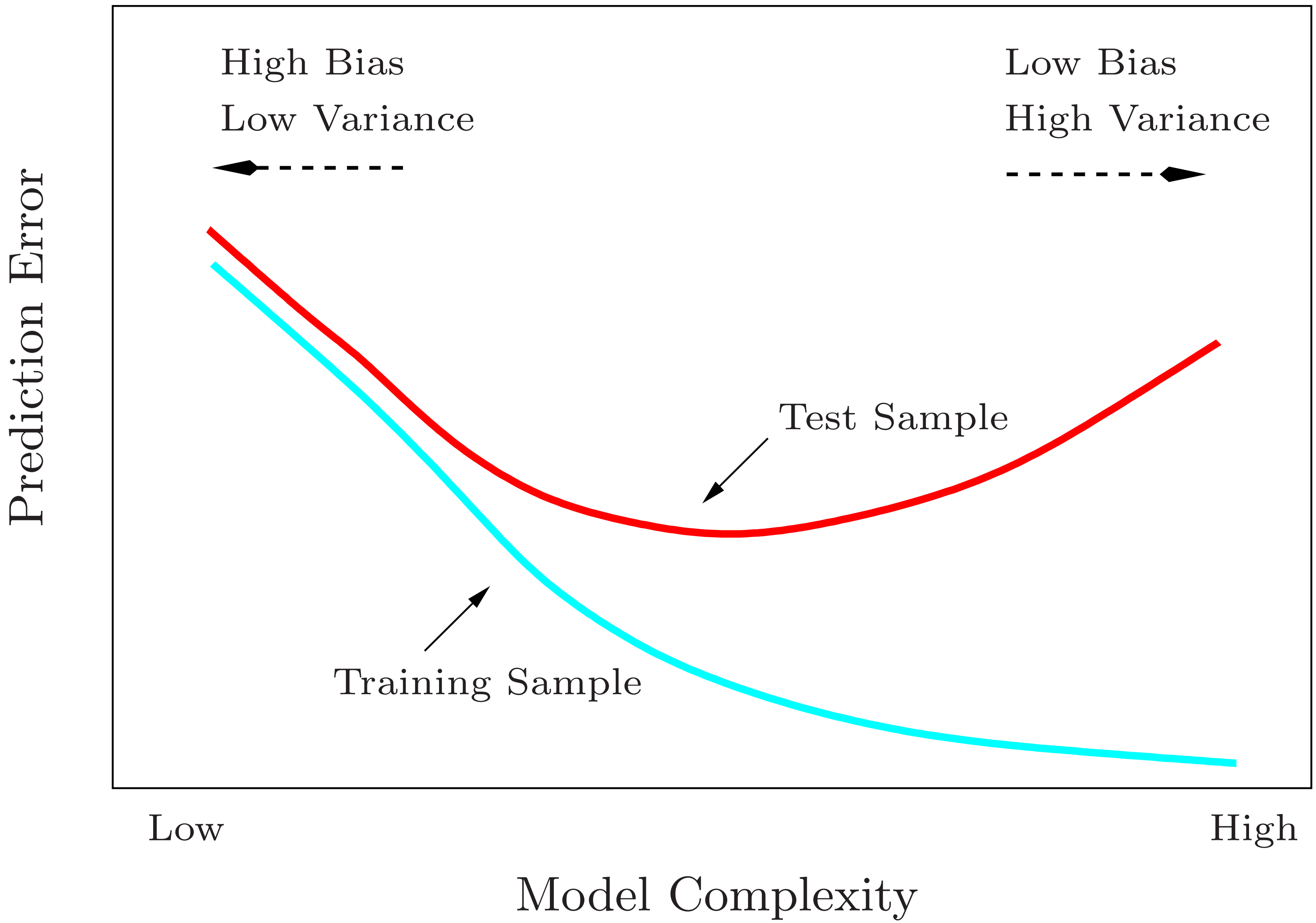
Center for
Bioinformatics
& Molecular
Biostatistics

**NUS / IMS Workshop on
Classification and Regression Trees**

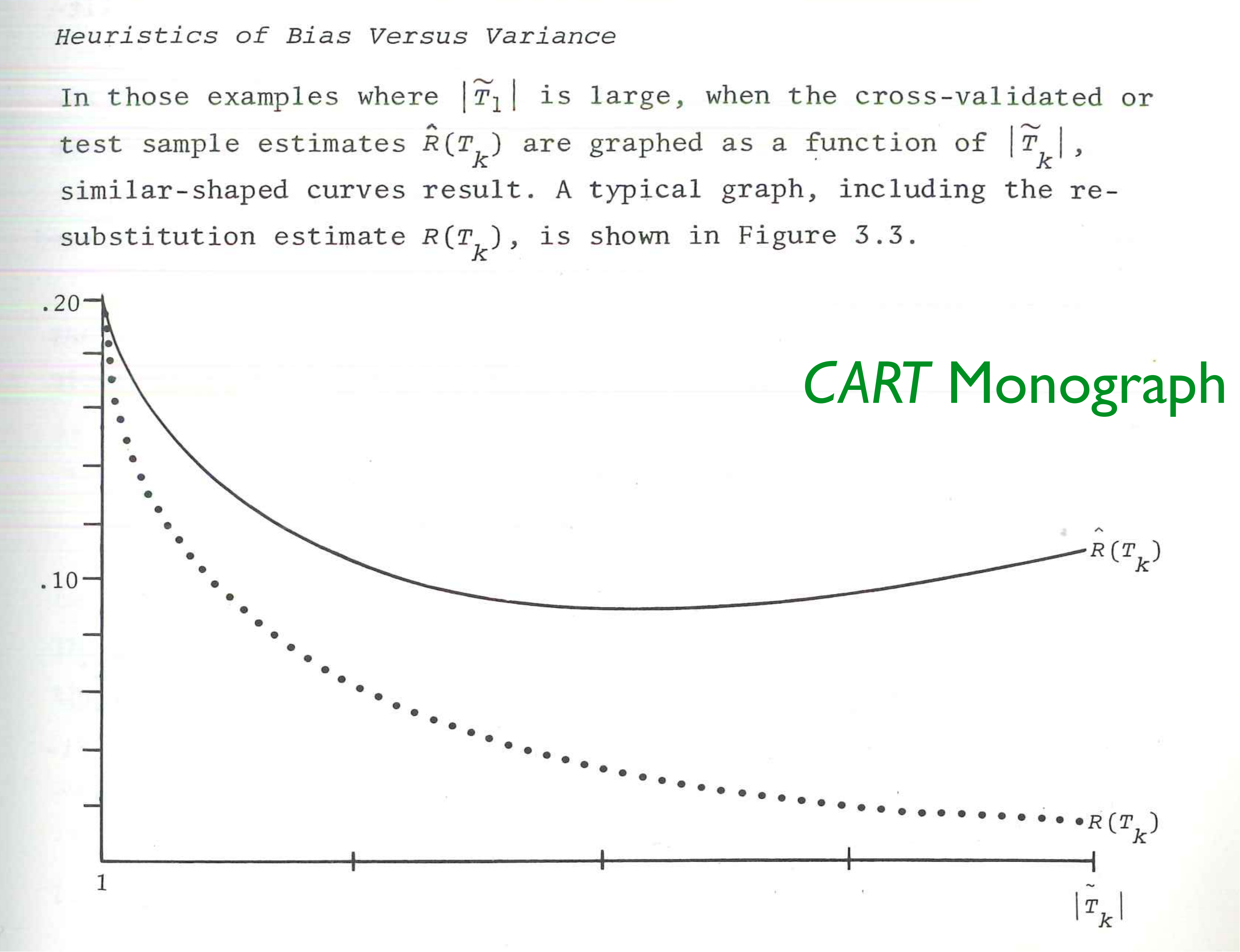
CART

- Breiman, Friedman, Olshen, Stone (1984)
- Popularized tree-structured techniques
- Primary distinction with earlier approaches?
  - Means for determining tree size
    - Grow large / maximal initial tree
      - capture all potential action
    - Cost-complexity pruning
    - Cross-validation based selection
- Size determination critical consideration
  - Why??

Predictive Performance



Predictive Performance



Heuristics of Bias Versus Variance

In those examples where $|\tilde{T}_1|$ is large, when the cross-validated or test sample estimates $\hat{R}(T_k)$ are graphed as a function of $|\tilde{T}_k|$, similar-shaped curves result. A typical graph, including the resubstitution estimate $R(T_k)$, is shown in Figure 3.3.

CART Monograph

- CART lived happily ever after
  - widespread uptake in diverse fields
  - many methodological refinements
  - this workshop (thanks Wei-Yin!)

- But, what about predictive performance??

Breiman Mantra

- Better the model fits, the more sound the inference
- Conventional models and CART tend to fit very poorly
- Fit measured by prediction error (*PE*)
- Substantial gains in *PE* can be achieved by using ensembles of (weak) predictors
  - in particular, individual trees

Random Forests

- Breiman (2001a,b)
- Have become a forefront prediction technique
- Notable gains in prediction performance over individual trees
  - PE *variance* reduced by averaging over the randomness-injected ensemble
    - Individual trees grown to large / maximal depth
      - Major departure from CART paradigm
- *Seemingly*, averaging over the ensemble *more than* compensates for increased individual tree variability

A RF is a collection of tree predictors $h(\mathbf{x};\boldsymbol{\theta}_t), t = 1,\ldots,T$; $\boldsymbol{\theta}_t$ *iid* random vectors For regression, the forest prediction is the unweighted average over the collection: $\bar{h}(\mathbf{x})$

As $t \rightarrow \infty$ the Law of Large Numbers ensures $E_{\mathbf{X},Y}(Y - \bar{h}(X))^2 \rightarrow E_{\mathbf{X},Y}(Y - E_{\boldsymbol{\theta}} h(\mathbf{X};\boldsymbol{\theta}))^2$ $\equiv PE_f^*$ the forest prediction error

Convergence implies forests *don't* overfit

Average prediction error for a single tree is

$PE_t^* = E_{\boldsymbol{\theta}} E_{\mathbf{X},Y} (Y - h(\mathbf{X}; \boldsymbol{\theta}))^2$

Assume $EY = E_{\mathbf{X}} h(\mathbf{x}; \boldsymbol{\theta}) \; \forall \boldsymbol{\theta}$

Then $PE_f^* \leq \bar{\rho} PE_t^*$ where $\bar{\rho}$ is weighted corr^n between residuals for independent $\boldsymbol{\theta}^{'}, \boldsymbol{\theta}^{''}$

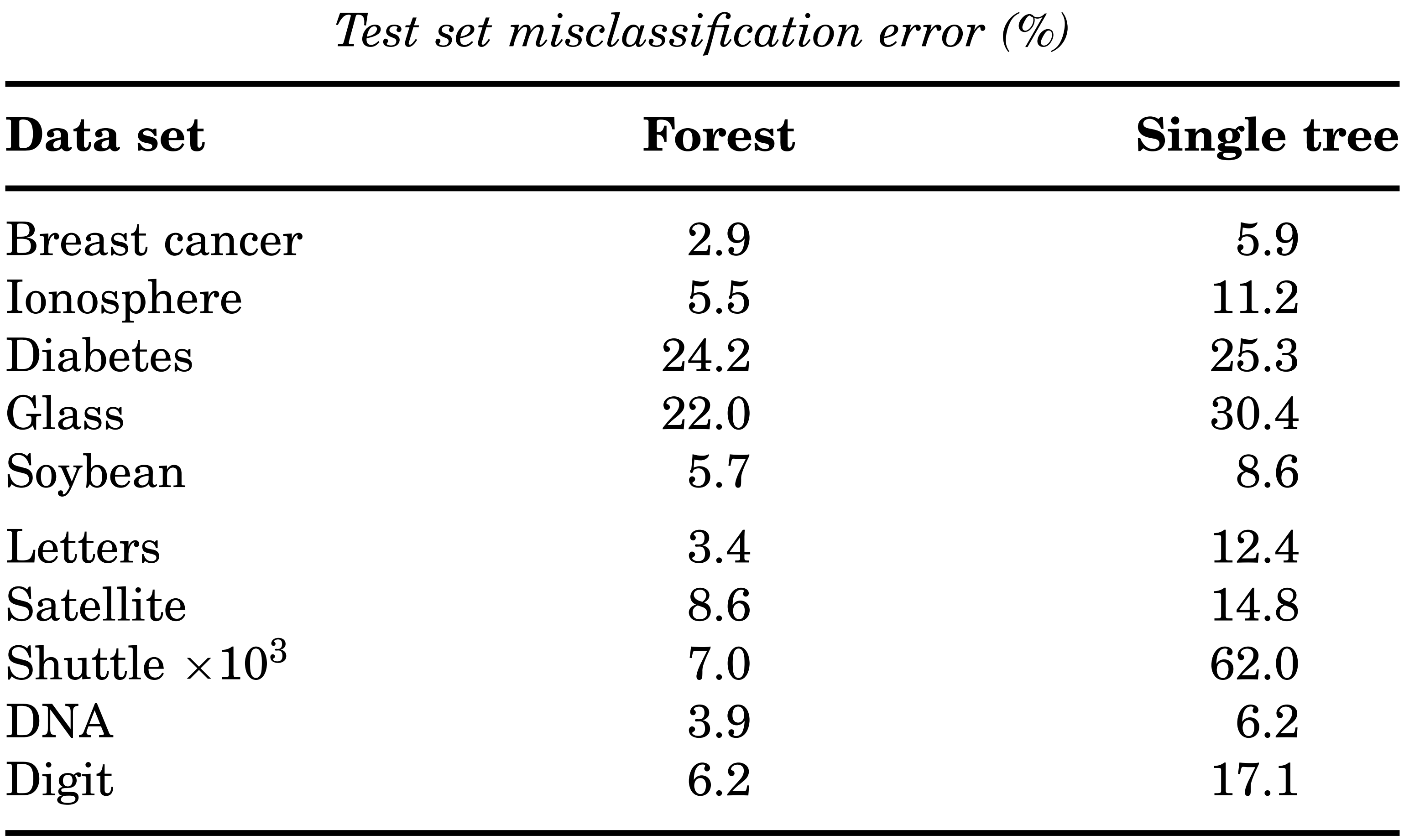
Inequality pinpoints needs for accurate RF: low residual corr^n; low PE for individual trees

Low corr^n sought via injected randomization

But what about low PE_t^*?

- Growing trees to maximal depth minimizes bias
  - But potentially incurs prediction variance cost
  - Averaging over ensemble putatively handles this
- But how was it established that such averaging (more than) compensates for increased individual tree variability??
  - Hard to address theoretically (will try later)
- Breiman (2001a,b) addressed empirically using
  - UCI Irvine machine learning benchmark datasets
    - Includes classification and regression problems
    - Simulated and (predominantly) real data
    - Exported to R mlbench library

Some classification results from UCI Irvine machine learning benchmark datasets:



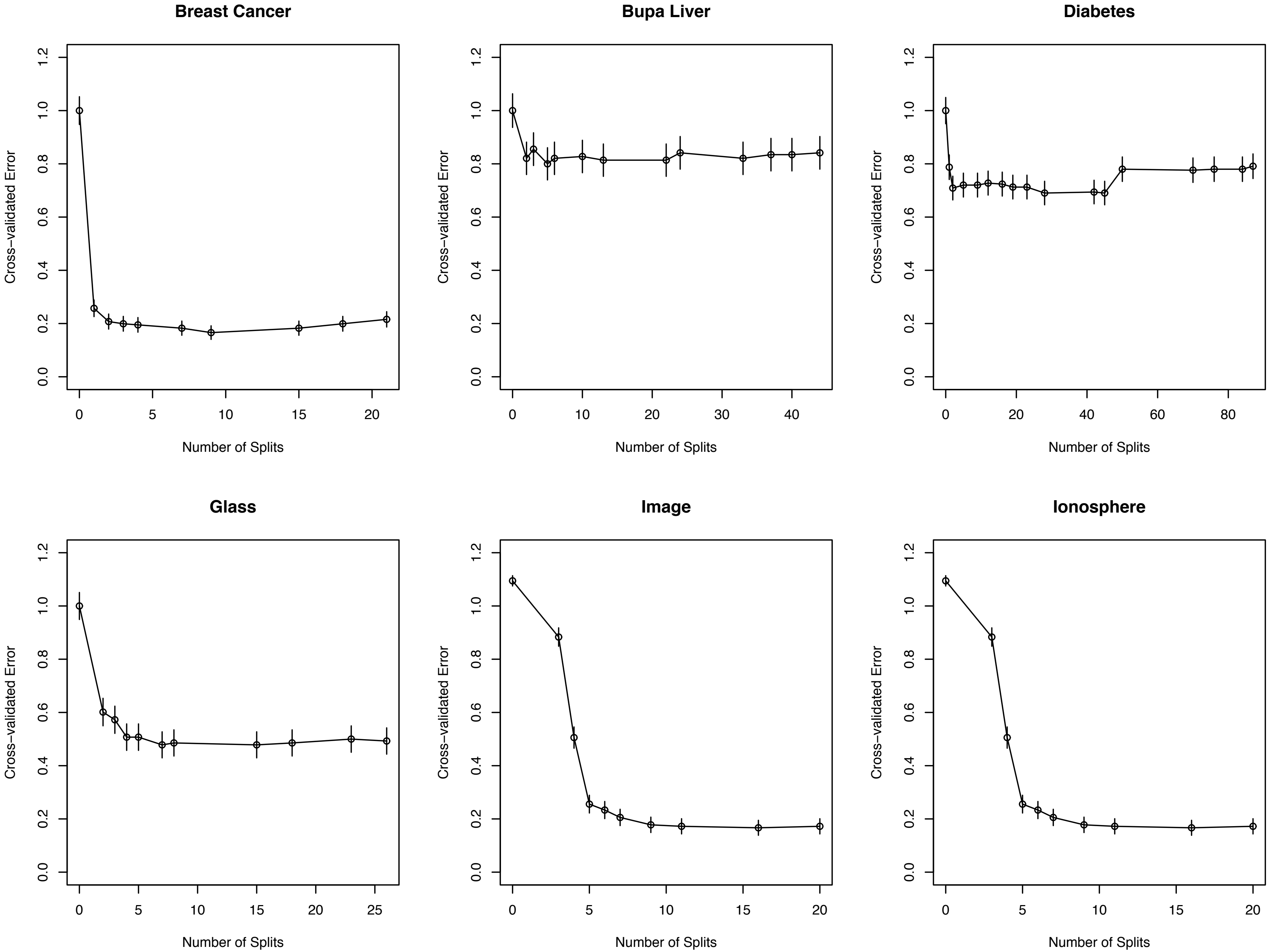
Test set misclassification error (%)

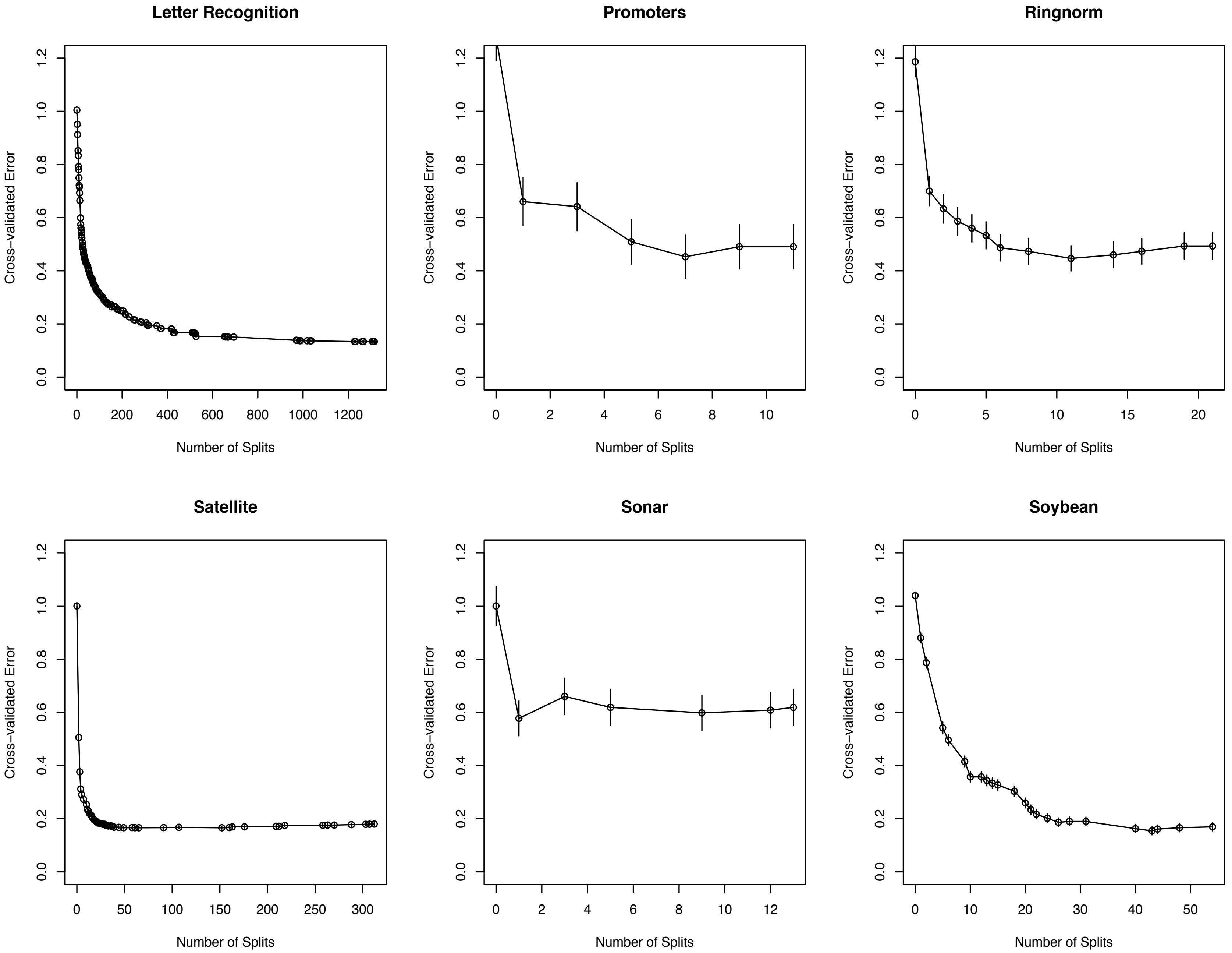
Data set	**Forest**	**Single tree**
Breast cancer	2.9	5.9
Ionosphere	5.5	11.2
Diabetes	24.2	25.3
Glass	22.0	30.4
Soybean	5.7	8.6
Letters	3.4	12.4
Satellite	8.6	14.8
Shuttle $\times 10^3$	7.0	62.0
DNA	3.9	6.2
Digit	6.2	17.1

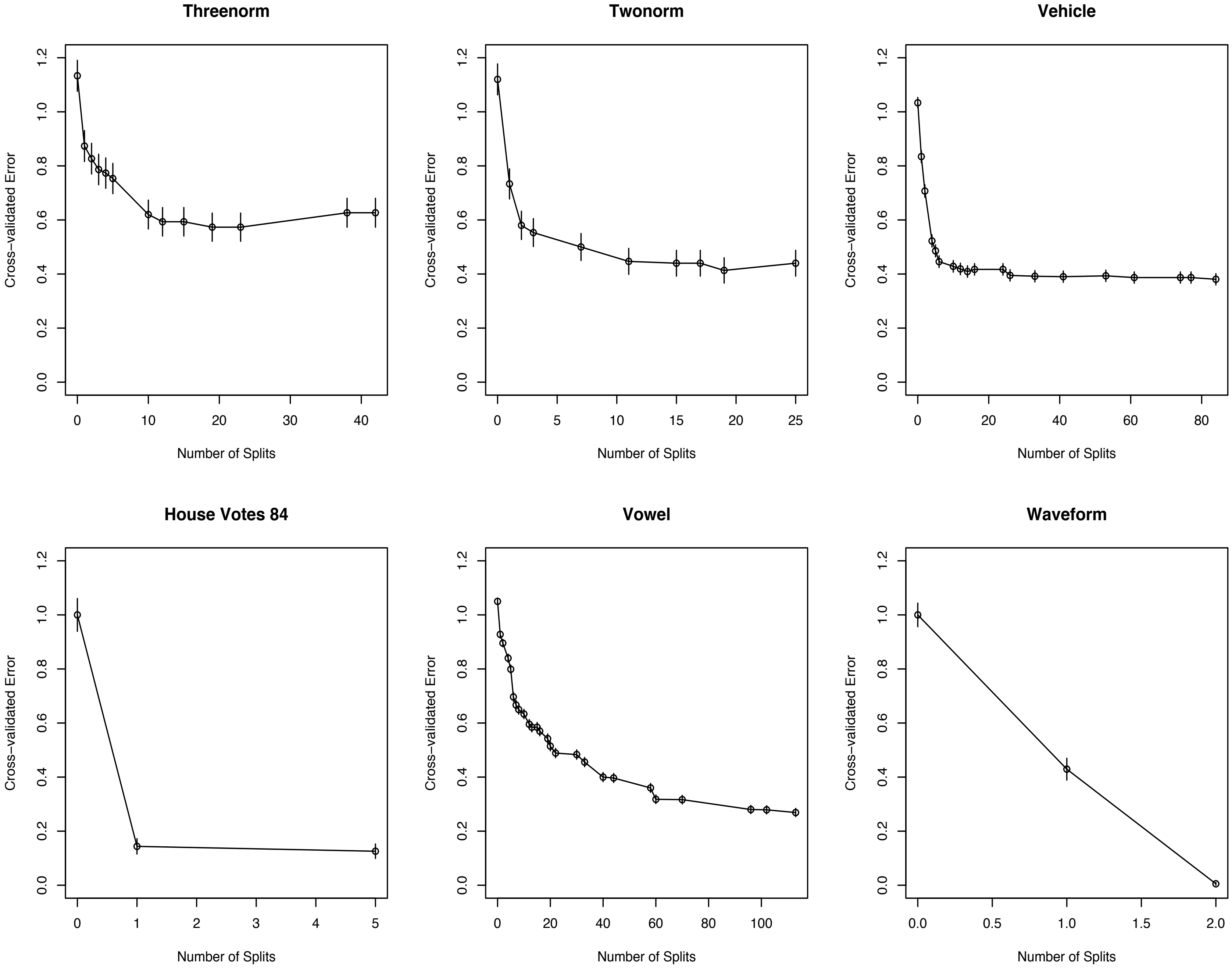
Breiman (2001a,b)

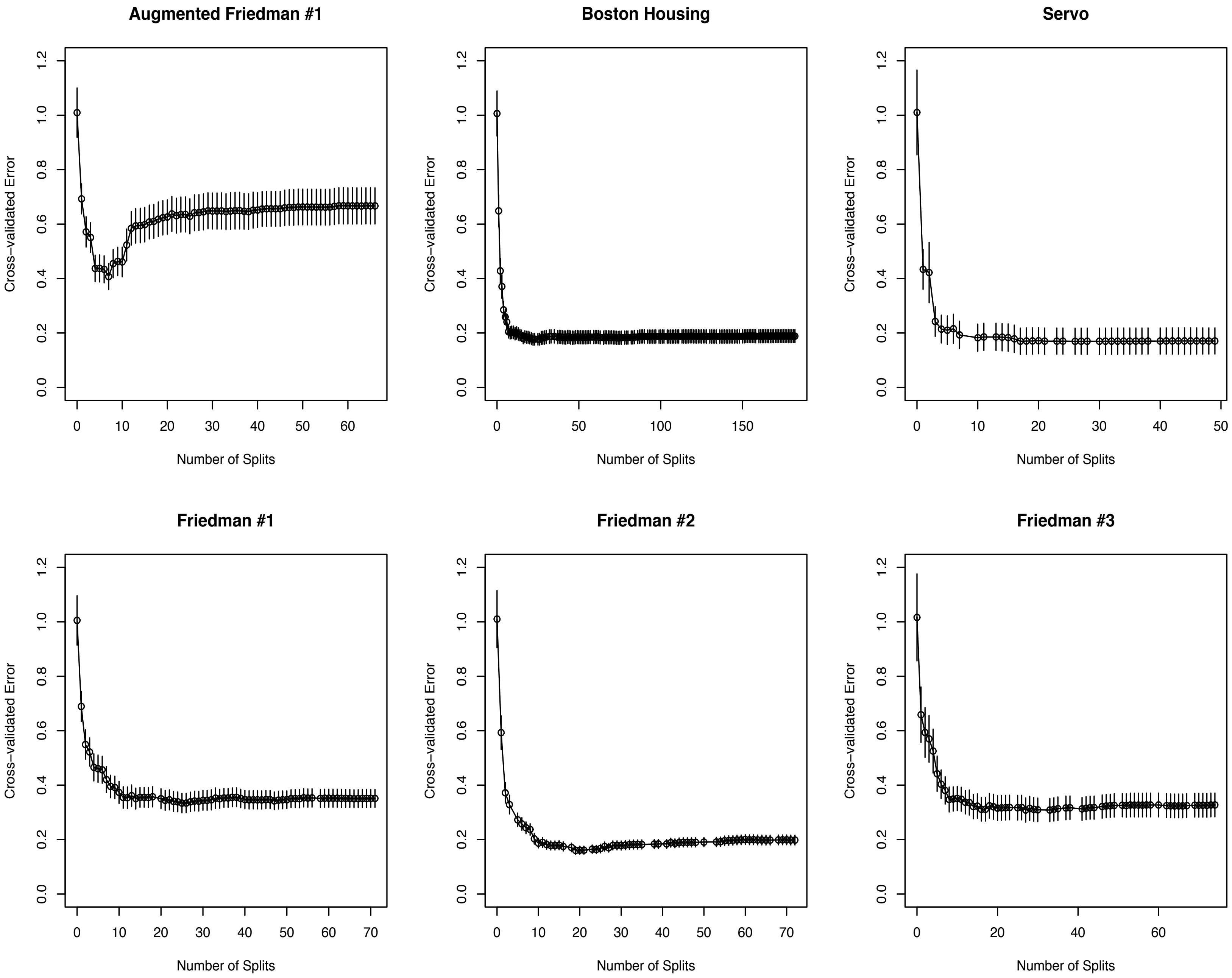
- Many further comparisons using the UCI Irvine / mlbench repository datasets:
    - several modeling / prediction frameworks:
        - CART, ANNs, LDA, QDA, kNNs...
    - regression and classification problems
- Conclusion: “Random Forests are A+ predictors”
- Discussion (Efron): Lots of knobs (tuning parameters)
- Rejoinder (Breiman): Essentially only one (*mtry*)

- Random Forests have lived happily ever after
- But, lets take a closer look at the UCI Irvine / mlbench repository datasets

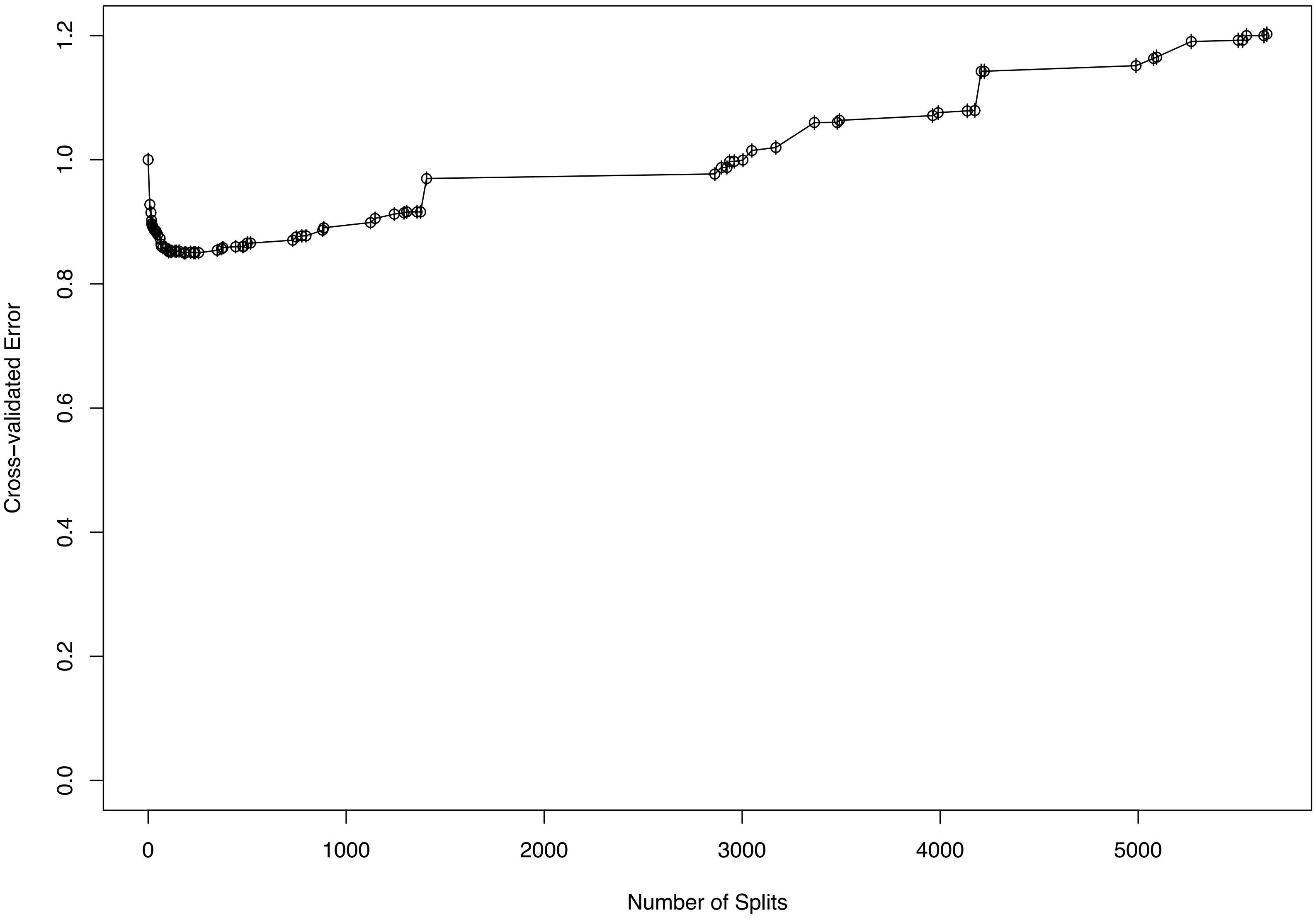

Breast Cancer
Bupa Liver
Diabetes
Glass
Image
Ionosphere
Cross-validated Error
Number of Splits
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
40
60
80


Letter Recognition
Promoters
Ringnorm
Satellite
Sonar
Soybean
Cross−validated Error
Number of Splits
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
200
400
600
800
1000
1200
0
2
4
6
8
10
12
0
5
10
15
20
0
50
100
150
200
250
300
0
10
20
30
40
50


Threenorm
Twonorm
Vehicle
House Votes 84
Vowel
Waveform
Cross−validated Error
Number of Splits
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
10
20
30
40
0
5
10
15
20
25
0
20
40
60
80
0
1
2
3
4
5
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0


Augmented Friedman #1
Boston Housing
Servo
Friedman #1
Friedman #2
Friedman #3
Cross−validated Error
Number of Splits
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
10
20
30
40
50
60
70
100
150

- Almost all UCI Irvine machine learning benchmark datasets exhibit this behaviour:
    - they are hard to *overfit* {not just with trees}
- This will make the Random Forest strategy of growing trees to maximal depth look good
- “Benchmarks” are not representative of what is at least thought to be prototypic
    - Will next showcase such an example
    - Then offer some theory and characterizations


Cross-validated Error
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1000
2000
3000
4000
5000
Number of Splits

Basal Splicing Signals

- Pre-messenger RNA splicing - responsible for precise removal of introns - is an essential step in expression of most genes
- Exons defined by short, degenerate splice site sequences at intron/exon boundaries: 5’ splice site (5’ss, donor); 3’ss, acceptor
- Each ss has a consensus sequence motif: essential nucleotides plus base usage preferences in flanking positions

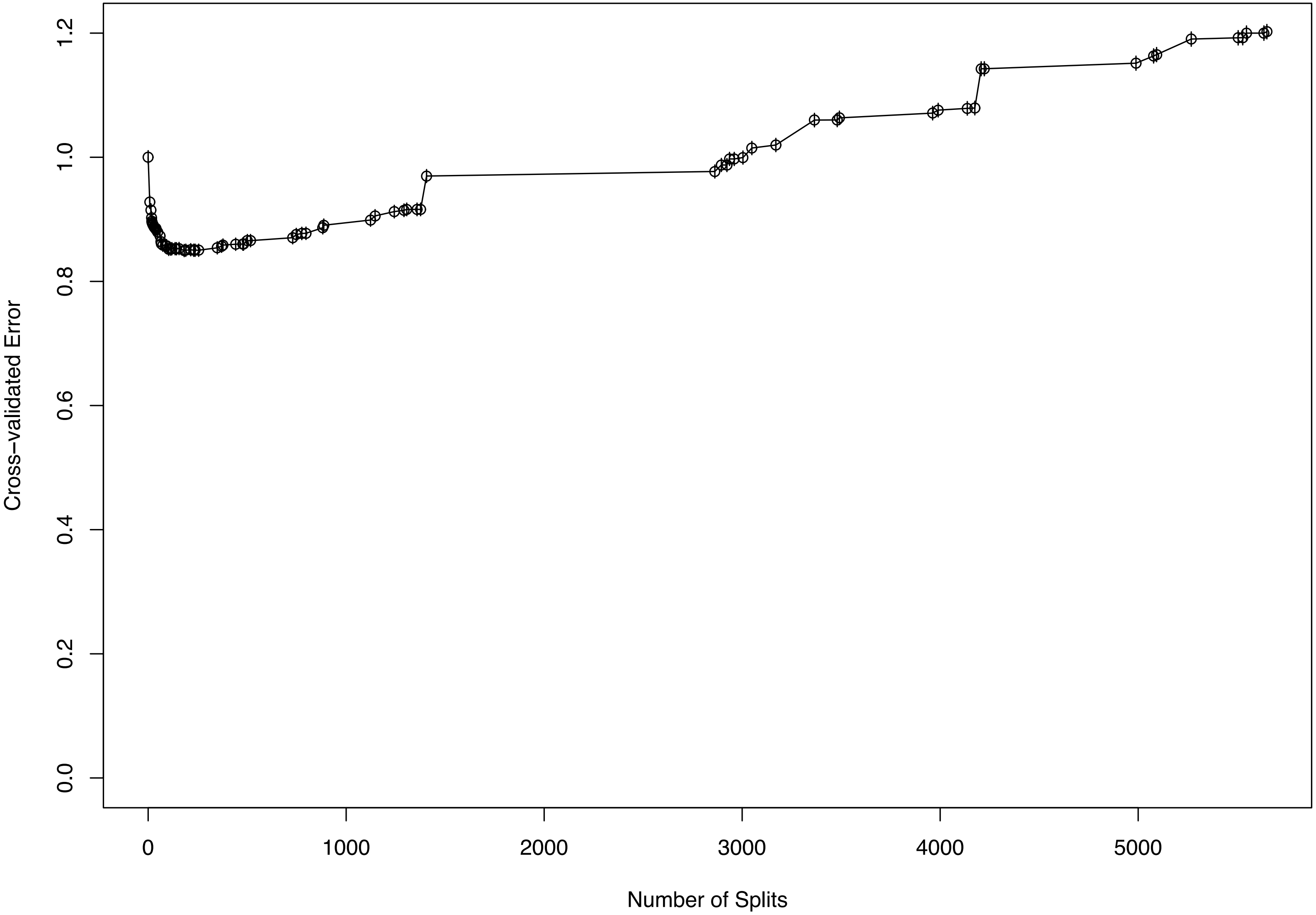
- Despite requirement for accurate splicing, human ss only moderately conserved
  - Implies an abundance of decoy ss
- Further, strong and complex dependencies between ss nucleotides exist
- Improved understanding of basal ss is important for exon recognition and, ultimately, disease impact of splicing defects
- Approach as a classification problem -- real vs decoy ss -- using large database

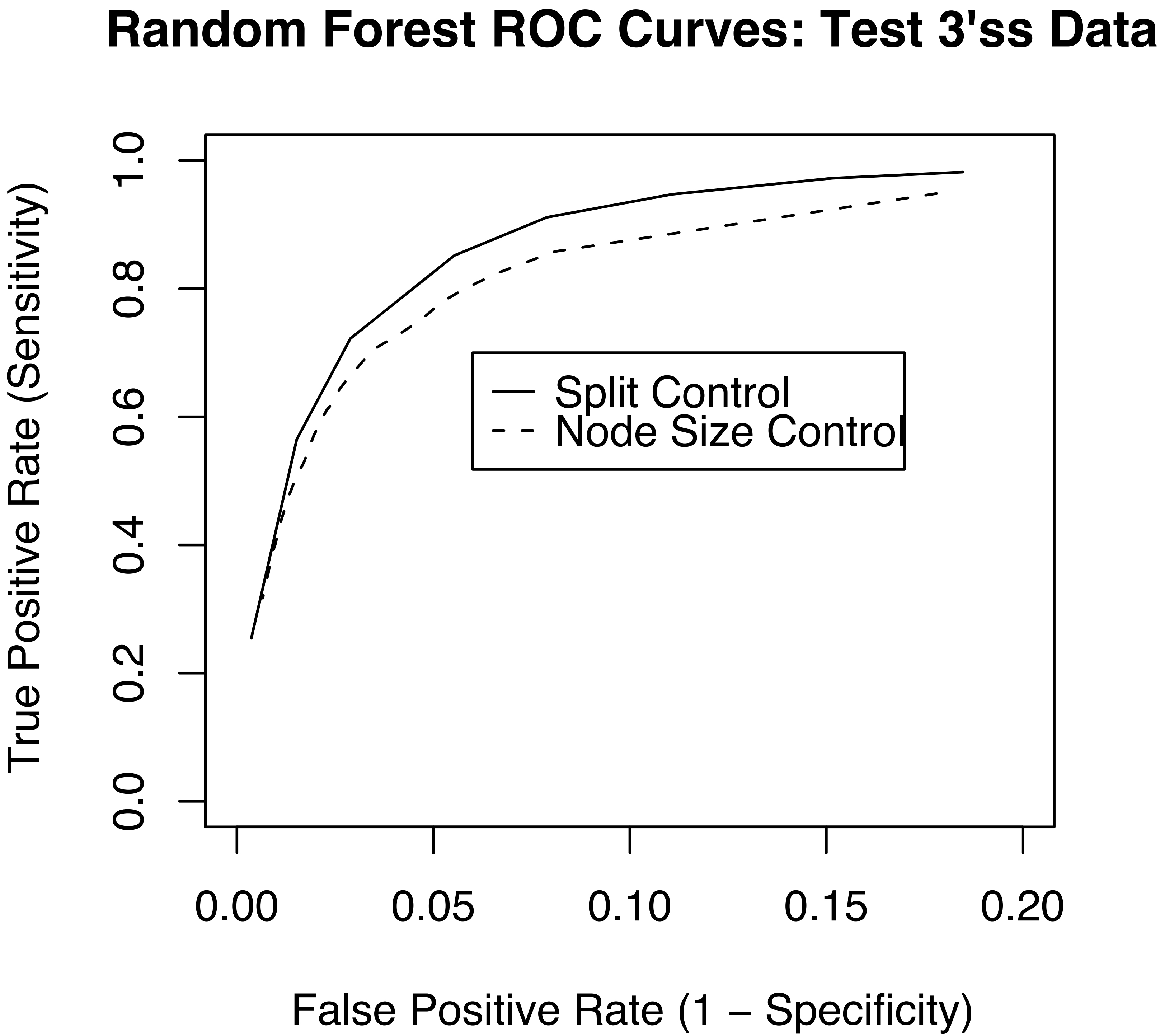
- Objective: predict 3’ splice site sequences
- Large *n*, small *p* datasets:
  - training 8465 real; 180957 decoy
  - test 4233 real; 90494 decoy

ATTCTTACAAGTCCAATAAGGTT	real
GAATCGCTTGAACCTGGGAGGTG	real
CTGAAATGTCTCATCTGCAGTAC	decoy
ATTTTATTTTTAAATTGCAGGTA	decoy

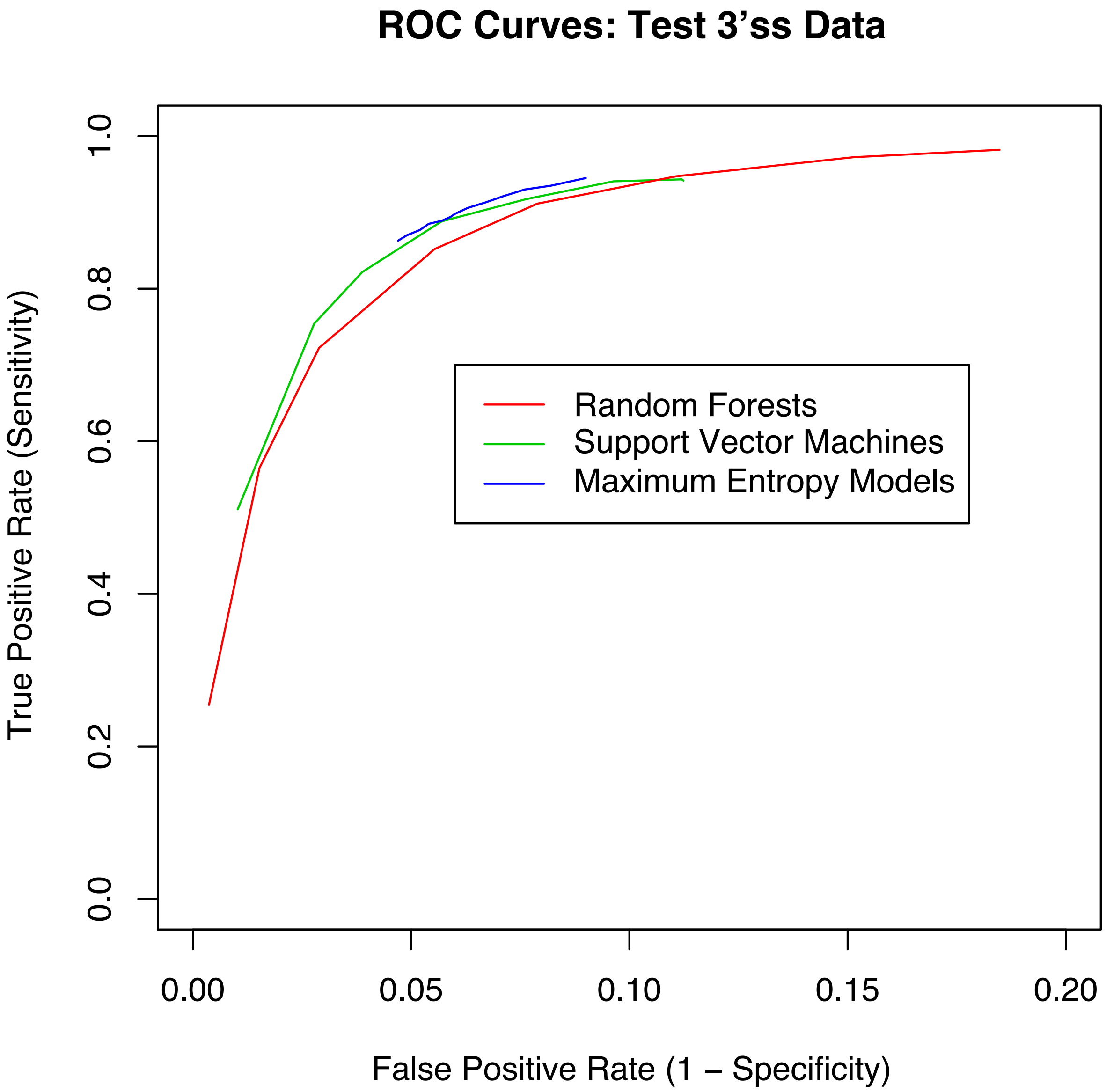
  - each (non-degenerate, aligned) position constitutes an unordered covariate (p = 21)
- data generation: Yeo and Burge (2003).

3’ss: CV error for a single tree




Random Forest ROC Curves: Test 3'ss Data
True Positive Rate (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
Split Control
Node Size Control
0.00
0.05
0.10
0.15
0.20
False Positive Rate (1 – Specificity)

{Aside: comparisons}



Tree Depth in a Forest

- Individual tree size determined by *inter-related* tuning parameters that govern (terminal) node size, number of splits, depth, split improvement
- A priori regulation via node size specifications problematic in large *n* situations
- Guidelines, rules-of-thumb as function of *n* are lacking (*cf* defaults for *m*)
- Leekasso

Potential Nearest Neighbours

- Lin and Jeong (2006, *JASA*)
- Develop construct of *k*-PNNs
- Establish connections between Random Forests and *k*-PNNs where *k* is terminal node size
  - *k* = 1 for trees grown to maximal depth
- Enables analysis of role of tree depth

RF grown on original training data $\{(\mathbf{x}_i, y_i)\}_1^n$
Prediction from tree t at target $\mathbf{x_0}$: $\sum W_{it} y_i$
$W_{it} = 1/k$ if $\mathbf{x}_i$ is among the k points in terminal node containing $\mathbf{x_0}$; zero otherwise.
Averaging over T trees the RF prediction at $\mathbf{x_0}$ is $\sum_{i=1}^n \bar{W}_i y_i$ with $\bar{W}_i = 1/T \sum_{t=1}^T W_{it}$.

RF is a weighted average of y_i's with weights depending on training data and $\boldsymbol{\theta}_t$.

Clearly $\bar{W}_i = 0$ for most sample points i.
Points with $\bar{W}_i > 0$ are called voting points.
In general, voting points are *not* NNs of $\mathbf{x_0}$ for any single distance metric.
However, voting points *are* k potential NNs: there exists a distance under which they are among the k closest sample points to $\mathbf{x_0}$ (from hyper-rectangular partitioning of RFs).

Thus RFs are a weighted k PNN method.

Under simplifying assumptions Lin and Jeon show that a lower bound on the rate of convergence of RF MSE is $k^{-1}(\log n)^{-(p-1)}$. Much inferior to standard rate $n^{-2d/(2d+p)}$ (where d is degree of target smoothness) attained by many nonparametric methods. To achieve competitiveness terminal node size k should increase with sample size n.

Intuitively: largest trees use 1-PNNs at $\mathbf{x_0}$ #1-PNNs $\sim O_p[(\log n)^{p-1}]$ which is too small.

Lin and Jeon: “growing large trees (k small) does not always give the best performance”

But, asymptotics require $n \gg p$ and even when seemingly applicable may not pertain. Consider $p = 10, d = 2, n = 100000$. Then $(\log n)^{p-1}/(p-1)! = 9793 \gg 27 = n^{2d/(2d+p)}$ Even more so the case for larger p, smaller n.

So, for high dimensional problems growing largest individual trees is often desirable.

Conclusions / Future Work

- UCI / mlbench data repositories are inadequate as representative testbeds

- *k*-PNNs provide a theoretic framework for (crudely) evaluating tree depth considerations

- In large sample settings (Big Data) growing the individual tree components of a Random Forest ensemble to maximal depth can be undesirable

- Approaches to developing guidelines, defaults, parameterizations, tuning strategies to address tree depth are yet to be developed

Acknowledgements

- Eugene Yeo
- Leo Breiman