Ensemble Methods in Data Mining

Andrew Fast, Ph.D.
Chief Scientist
Elder Research, Inc.
fast@dataminininglab.com

John F. Elder, IV, Ph.D.
President and CEO
Elder Research, Inc.
elder@dataminininglab.com

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

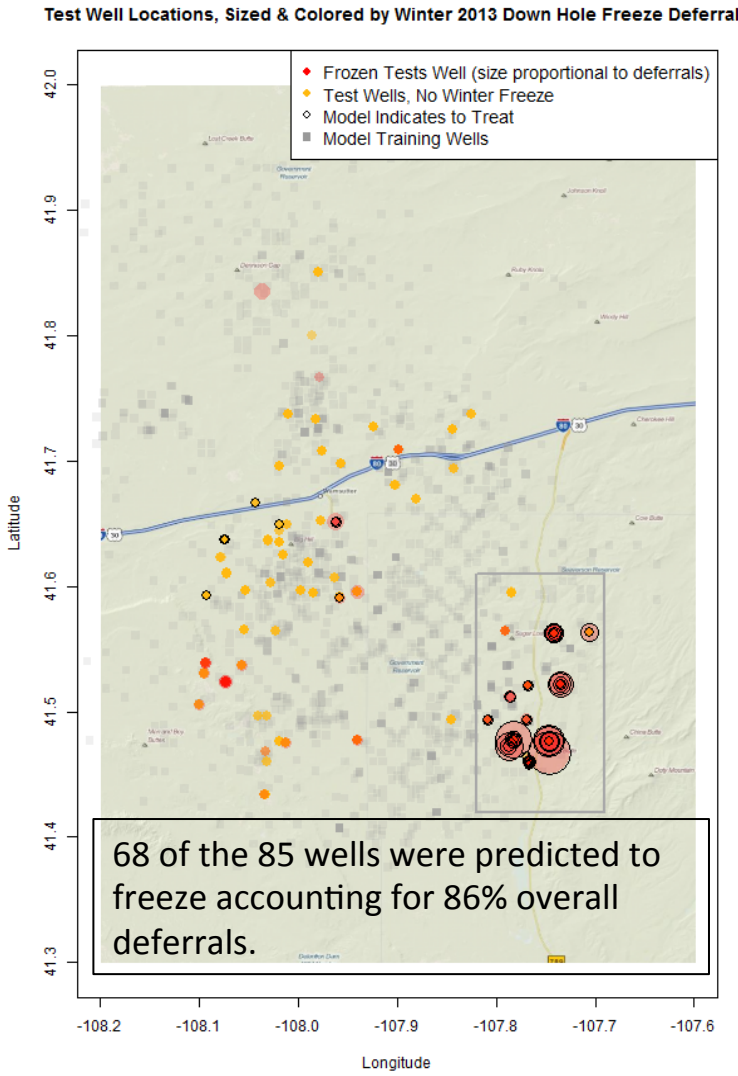
Conference on Applied Statistics in Defense (CASD)
Washington, D.C., October 2014

Occam's Razor and Induction

- "Entities should not be multiplied beyond necessity"
- Prefer parsimonious models when learning from data to increase performance on new, unseen data
- *Paradox of ensembles:* Combined models perform better than single models

[e.g., Domingos 1999]

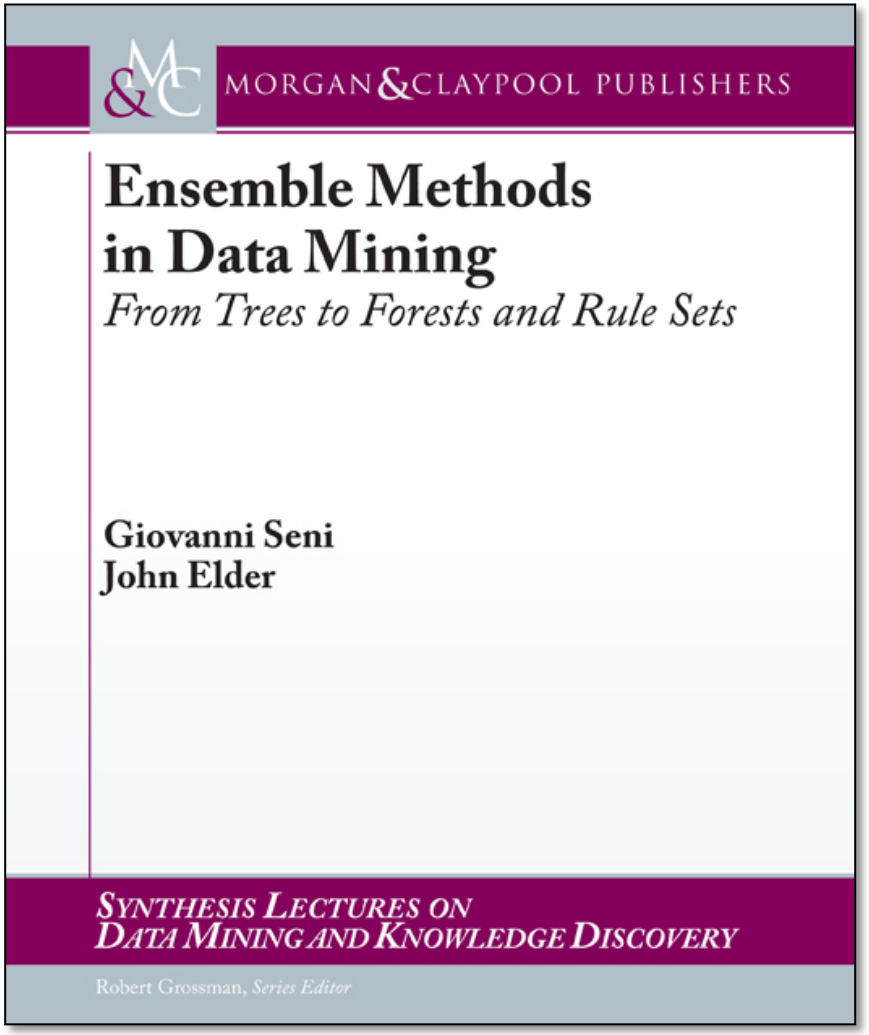
Ensembles in Action



- Work for a major petrochemical company
- Focusing on well-stoppages called "freezes"
  - Lost production called "deferrals"
- Can we provide a ranked list of wells most likely to "freeze"?

Ensemble Methods in Data Mining

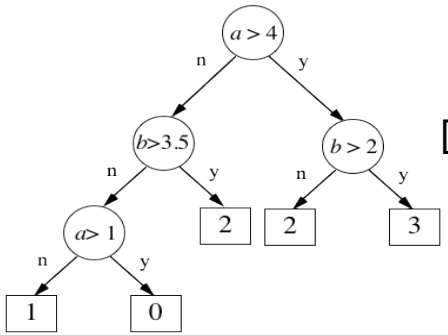
- Giovanni Seni & John Elder
- First published in 2010
  - 2nd Edition coming soon
- Complete exposition of the Importance Sampling Learning Ensembles (ISLE) approach



[Seni & Elder 2010]

Distinctives of Inductive Modeling

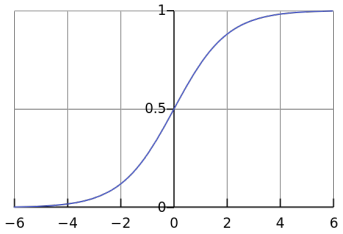
- Observational data only, no experimental control



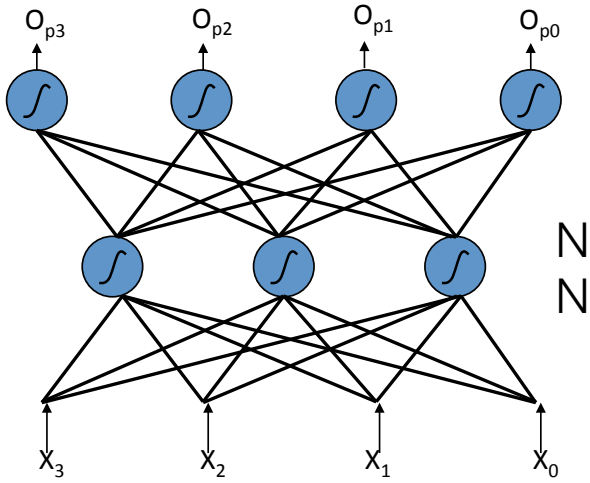
Decision Trees

- Learning *structure* of model, not just fitting parameters

Logistic Regression



- Both Classification and Regression tasks



Neural Networks

Challenge: Overfit

- Vast Search Effect
  - Many models considered
  - Can look good by chance alone
- Most inductive algorithms will learn structure from random data
- Overfit models do not generalize well to unseen data

http://www.bypolaropposite.com/whatsthat-health/2014/10/8/pareidolia-that-thing-has-a-face-i-saw-bob-marley-in-a-tree

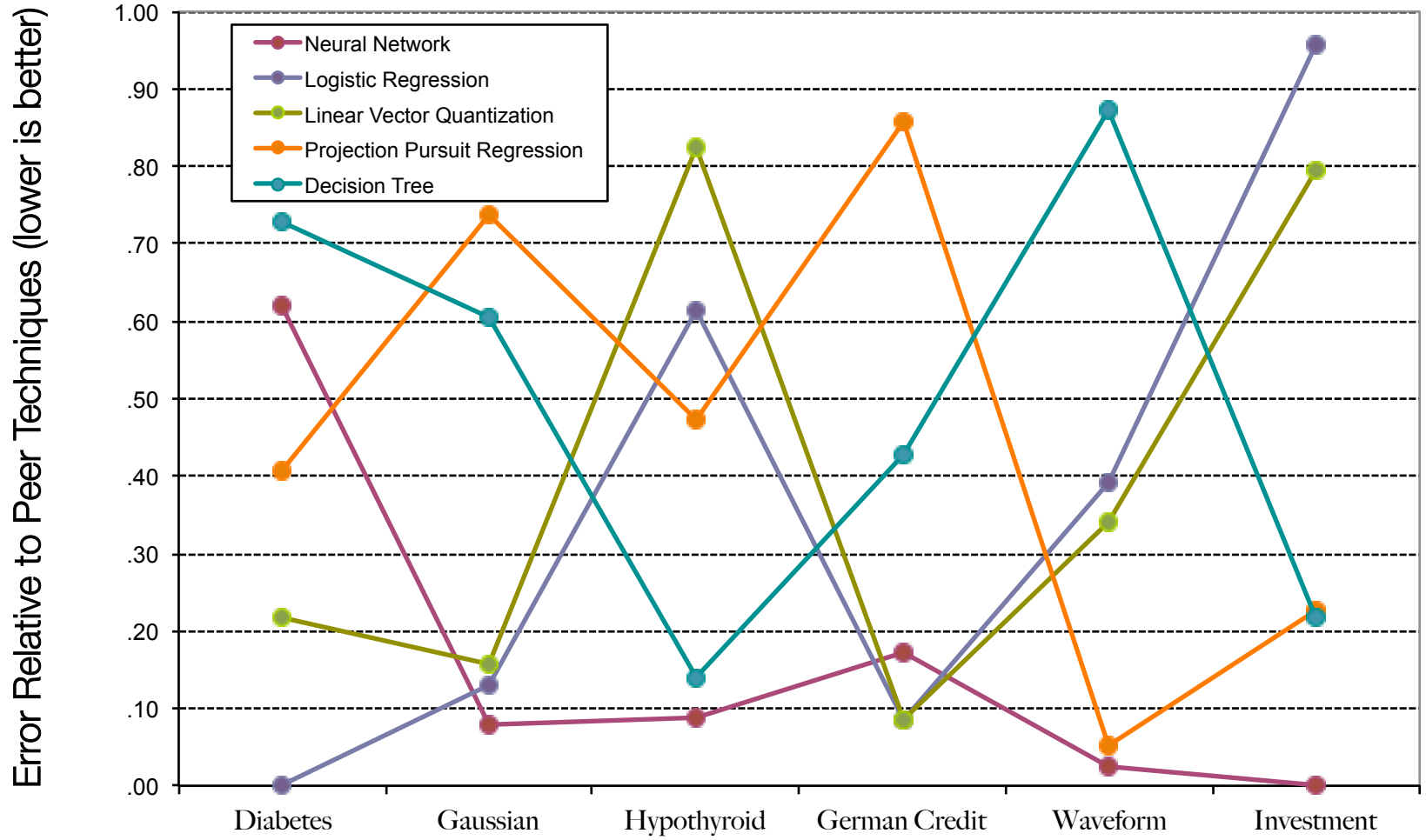
Solution: Penalize Model Complexity

1. Explicit Complexity Constraints
   - Ridge, Lasso, BIC, etc.
2. Incremental model building
   - Stepwise Regression
3. Robust Loss Functions
   - Huber Loss

Regularization is "…any part of model building which takes into account – implicitly or explicitly – the finiteness and imperfection of the data and the limited information in it, which we can term 'variance' in an abstract sense."

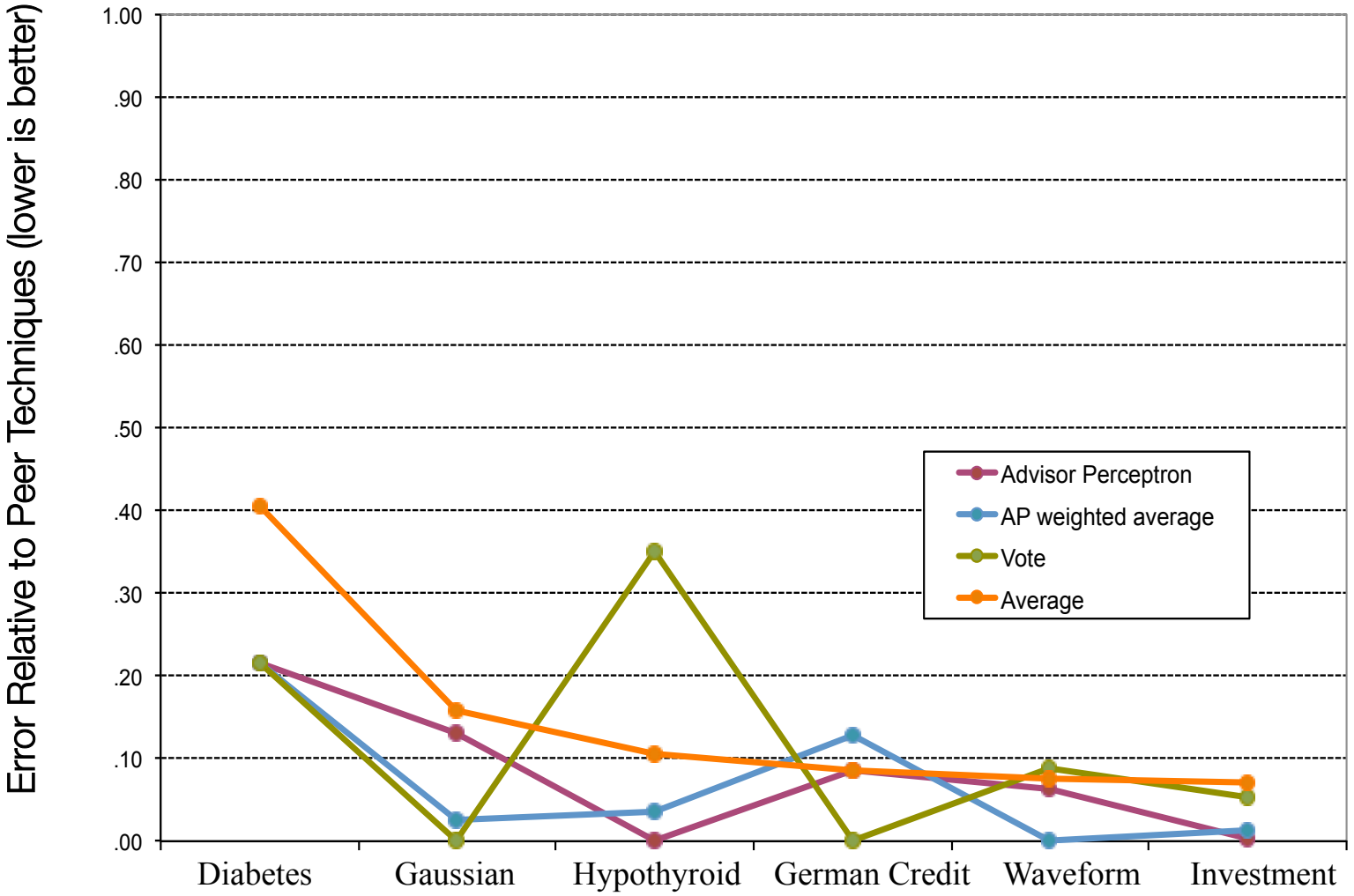
- Saharon Rosset (2003), *Ph.D. Thesis*

Single Model Performance



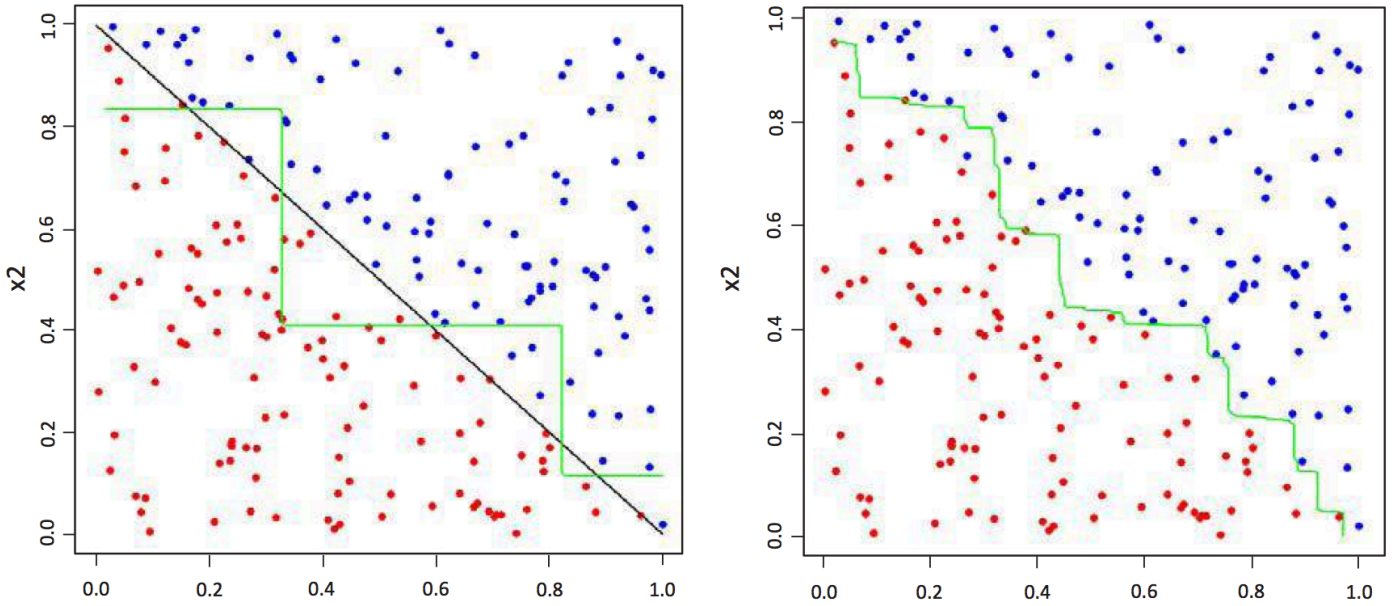
Comparison of 5 Algorithms on 6 Datasets (from Elder and Lee, 1997)

“Bundling” Models Improve Performance



Traditional Ensembles

- Combine multiple models together into a single result
- Usually leads to improved accuracy and generalization
- Techniques include:
  - Boosting
  - Bagging
  - Random Forests
  - Bayesian Model Averaging
  - Stacking



Importance Sampling Learning Ensembles

$$\left\{\hat{c}_m, \hat{p}_m\right\}_0^M = \min_{\left\{c_m, p_m\right\}_0^M} \sum_{i=1}^{N} L\left(y_i, c_0 + \sum_{m=1}^{M} c_m T_m(x; p_m)\right)$$

Minimize the loss function , *L*

Model Weights

Collection of base learners (basis functions)

- "The Ensemble is a linear model in a (very) high dimensional space of derived variables" (Seni & Elder, pg 39)
- Must sample from space of all possible learners
  - Need to balance breadth and quality

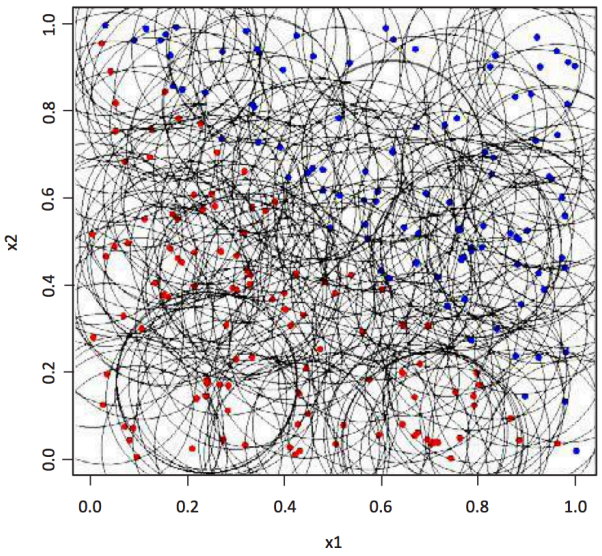
- Difficult Optimization problem!
- Use Perturbation Sampling to find a solution
  - Perturb Data distribution
  - Perturb Loss Function
  - Perturb Search algorithm

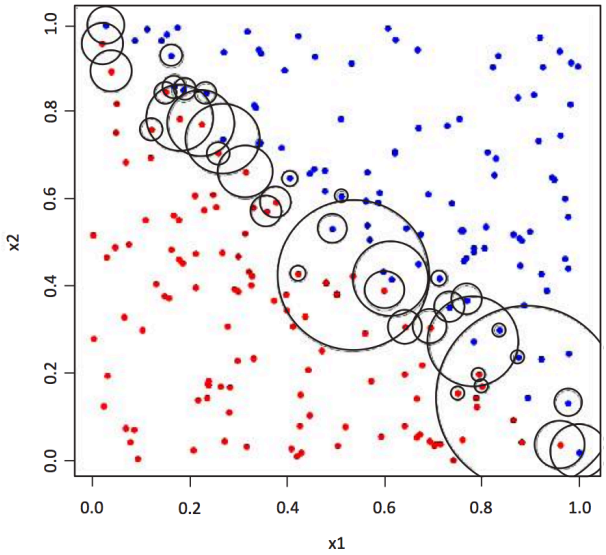
[Friedman and Popescu, 2003]

Example: Gradient Boosting

- Boosting reweights the data inputs based on error (higher error means more weight)
- Generalization of AdaBoost to any differential loss function
- Gradient Boosting is doing shrinkage implicitly with a Lasso-type penalty (Seni & Elder, pg 61)

Before: All instances have the same weight



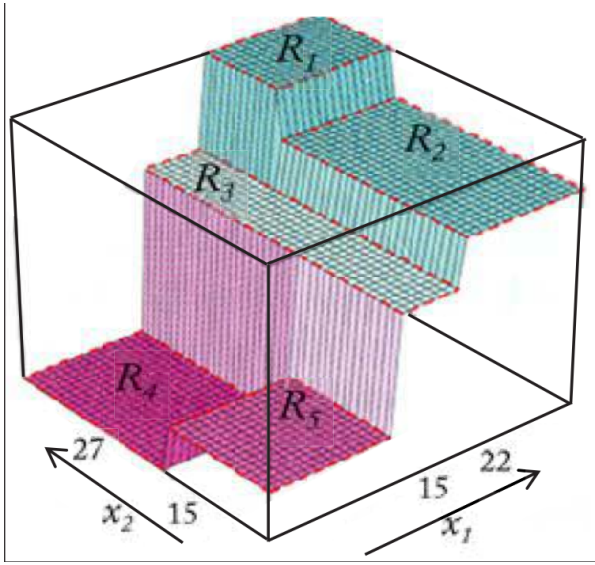


After: Only instances along the decision boundary have non-zero weight

[J. Friedman 2001]

Example: Rule Ensembles

- Ensembles lose the explainability of a single model
- Rule Ensembles provide improved explainability but with accuracy of the ensemble
- Convert trees to conjunctive rules
- Can be combined with other base learner functions
  - e.g., add linear terms



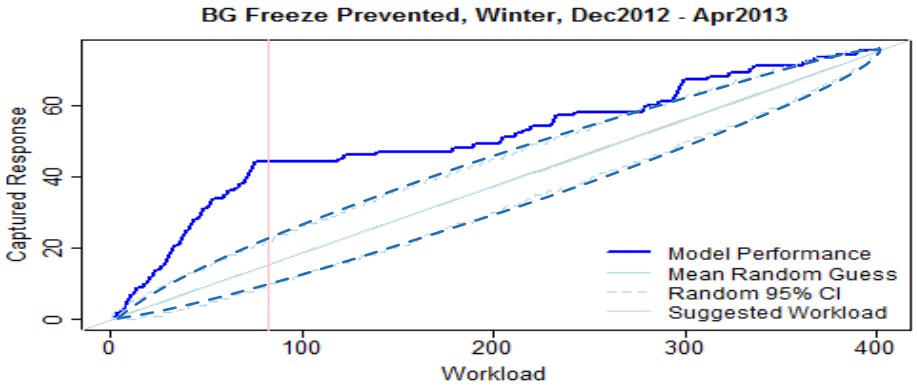
Region	Rule
R_1	$r_1(\mathrm{x}) = I(x_1 > 22) \cdot I(x_2 > 27)$
R_2	$r_2(\mathrm{x}) = I(x_1 > 22) \cdot I(0 \leq x_2 \leq 27)$
R_3	$r_3(\mathrm{x}) = I(15 < x_1 \leq 22) \cdot I(0 \leq x_2)$
R_4	$r_4(\mathrm{x}) = I(0 \leq x_1 \leq 15) \cdot I(x_2 > 15)$
R_5	$r_5(\mathrm{x}) = I(0 \leq x_1 \leq 15) \cdot I(0 \leq x_2 \leq 15)$

[Friedman and Popescu, 2005; Friedman and Bogdan, 2008]

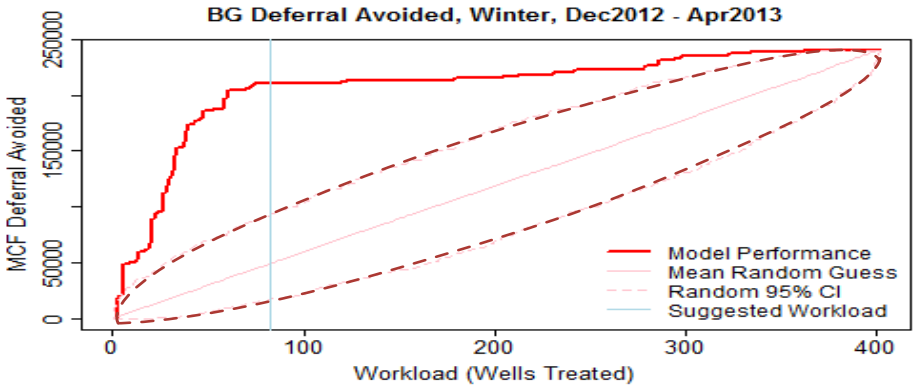
Graphic adapted from Hastie et al. , 2001

Well Maintenance Results

- Used Gradient Boosted Trees to predict which wells would require maintenance

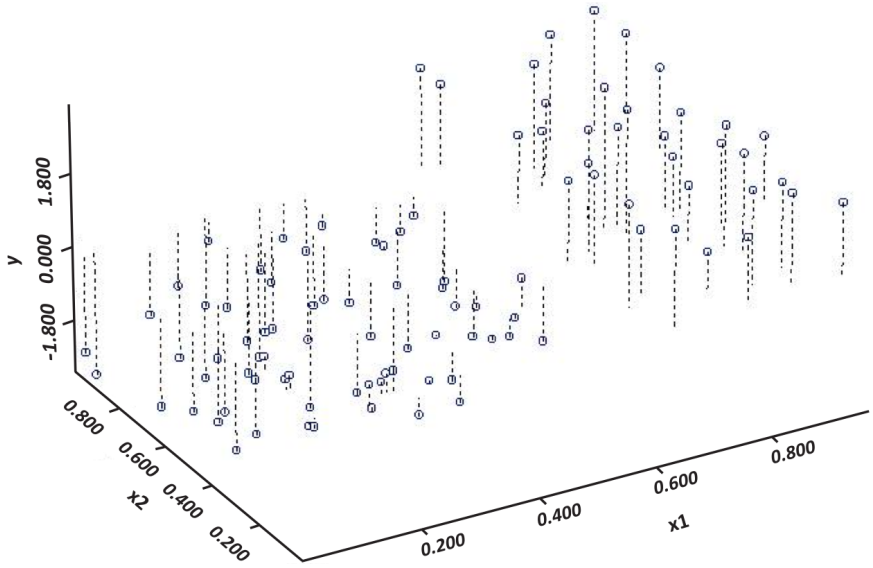


- Dotted lines indicate the 95% percent confidence interval using permutation testing

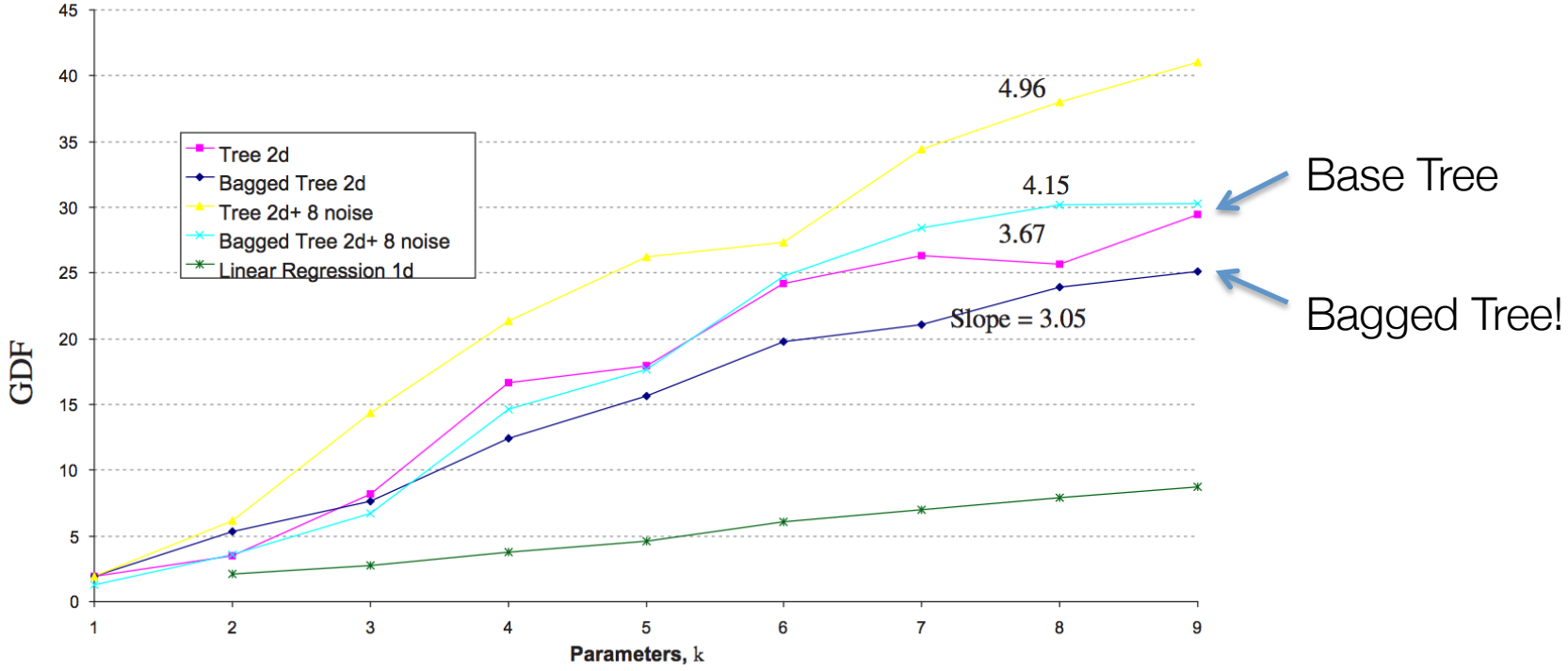


Ensembles and Complexity

- Traditionally measured by model size (number of model parameters).
  - Clearly this fails for Ensembles
- Generalized Degrees of Freedom (Ye 1998)
  - Empirical measure of the flexibility of the entire modeling process.
- True degrees of freedom can be greater than the number of parameters.



Generalized Degrees of Freedom



- Bagged Tree has fewer degrees of freedom than a single tree with the same number of splits

Summary

- Ensembles are a powerful tool for obtaining accurate results on real world problems.
- Despite multiple models, *flexibility* of the modeling process is actually lower in ensembles than a single model due to an implicit regularization process
- R packages and examples available in the book.
  - Increased adoption in COTS software, too

Contact Information

Andrew Fast, Ph.D.
Chief Scientist

fast@datamininglab.com
(434) 973-7673
www.datamininglab.com

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Andrew Fast
Chief Scientist, Elder Research, Inc.

Dr. Andrew Fast leads research in Text Mining and Social Network Analysis at Elder Research, the nation's leading data mining consultancy. ERI was founded in 1995 and has offices in Charlottesville VA and Washington DC, (*www.datamininglab.com*). ERI focuses on Federal, commercial, investment, and security applications of advanced analytics, including stock selection, image recognition, biometrics, process optimization, cross-selling, drug efficacy, credit scoring, risk management, and fraud detection.

Dr. Fast graduated Magna Cum Laude from Bethel University and earned Master's and Ph.D. degrees in Computer Science from the University of Massachusetts Amherst. There, his research focused on causal data mining and mining complex relational data such as social networks. At ERI, Andrew leads the development of new tools and algorithms for data and text mining for applications of capabilities assessment, fraud detection, and national security.

Dr. Fast has published on an array of applications including detecting securities fraud using the social network among brokers, and understanding the structure of criminal and violent groups. Other publications cover modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (work featured on ESPN.com). With John Elder and other co-authors, Andrew has written a book on Practical Text Mining, that was awarded the prose Award for Computing and Information Science in 2012.

John F. Elder IV
Founder & CEO, Elder Research, Inc.

Dr. John Elder heads the USA's largest and most experienced data mining consulting team. Founded in 1995, Elder Research, Inc. has offices in Charlottesville, Virginia, Washington DC, and Baltimore Maryland (*www.datamininglab.com*). ERI focuses on Federal, commercial, and investment applications of advanced analytics, including text mining, credit scoring, process optimization, cross-selling, drug efficacy, market timing, and fraud detection.

John earned Electrical Engineering degrees from Rice University, and a PhD in Systems Engineering from the University of Virginia, where he' s an adjunct professor teaching Optimization or Data Mining. Prior to 19 years at ERI, he spent 5 years in aerospace defense consulting, 4 heading research at an investment management firm, and 2 in Rice's *Computational & Applied Mathematics* department.

Dr. Elder has authored innovative data mining tools, is a frequent keynote speaker, and has chaired International Analytics conferences. John was honored to serve for 5 years on a panel appointed by President Bush to guide technology for National Security. His book with Bob Nisbet and Gary Miner, *Handbook of Statistical Analysis & Data Mining Applications*, won the PROSE award for top book in Mathematics for 2009. His book with Giovanni Seni, *Ensemble Methods in Data Mining*, was published in 2010, and his book with colleague Andrew Fast and 4 others on *Practical Text Mining* won the 2012 PROSE award for Computer Science.

John is grateful to be a follower of Christ and father of 5.