Breakthroughs using Ensembles – a Committee of Models

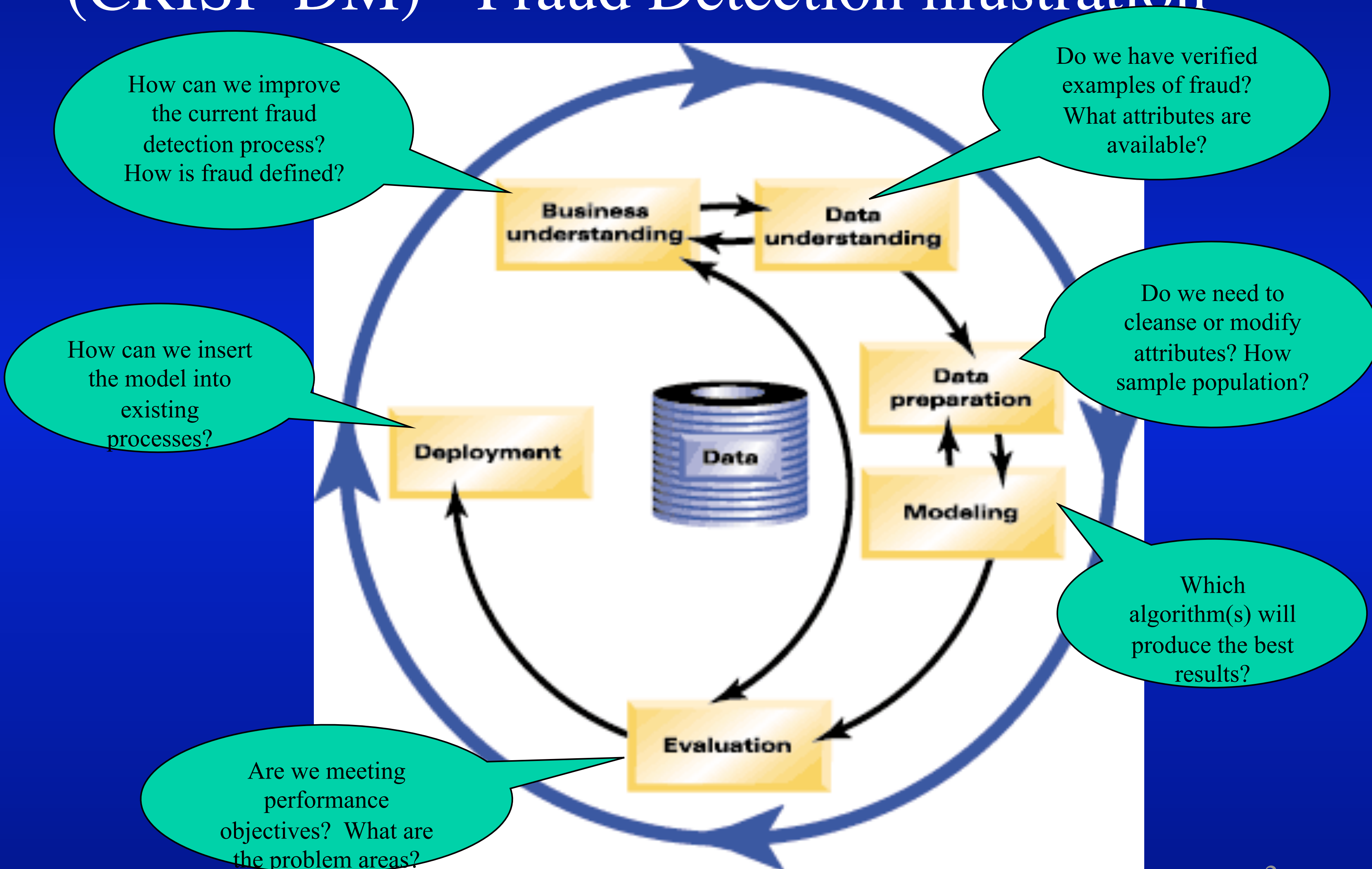
MITRE ASIAS Symposium

McLean, VA
July 27-28, 2009

Cheryl G. Howard, Ph.D.
cheryl@datamininglab.com

Elder Research, Inc.
571-216-4926
635 Berkmar Circle
Charlottesville, Virginia 22901
www.datamininglab.com

Cross Industry Standard Process for Data Mining (CRISP-DM) - Fraud Detection illustration



Properties of Algorithms

(a subjective, but empirical assessment)

Algorithm	Accurate	Scalable	Interpretable	Useable	Robust	Versatile	Fast	Hot
Classical (LR, LDA)	–	√	√–	√	–	–	√	x
Neural Networks	√	x	x	-x	–	x	xx	√
Visualization	√	xx	√	√	√√	x	xxx	√–
Decision Trees	x	√	√-	√	√	√	√–	√–
Polynomial Networks	√	–	x	–	–x	–	–x	–
K-Nearest Neighbors	x	xx	√–	–	–x	x	√	x
Kernels	√	xx	x	–x	x	x	√	x

√: good **-**: neutral **x**: bad

Why Ensembles?

- The process of selecting a model involves
  - Model class selection
    - Linear regression, decision trees, neural network
  - Variable selection
    - variable exclusion, transformation, smoothing
  - Parameter estimation
- One tends to choose the model that fits the data best as *the* model.

Empirical Comparison

Commenting (favorably) on Leo Breiman's contribution to the 11/1996 issue of *Machine Learning*, the Executive Editor revealed:

"...In some of my own papers (1995), we conducted only one run of each algorithm and then applied a test for the difference of two proportions to draw statistical conclusions. We did not consider the possibility that if the algorithms were run again on a second training set, the results could have been very different."

What's wrong with that?

- Two models may equally fit a dataset
(with repect to some loss function)
but have different predictions.
- Competing interpretable models with equivalent performance support ambigiuous conclusions.
- Model search dilutes the evidence.
"Part of the evidence is spent specifying the model."

Bayesian Model Averaging

Goal: Account for model uncertainty

Method: Use Bayes' Theorem and average the models by their posterior probabilities

$$P(M_k \mid D) = \frac{P(D \mid M_k)P(M_k)}{\sum_{l=1}^{K} P(D \mid M_l)P(M_l)}$$

M_k - model
D - data
P(D|M_k) - integrated likelihood of M_k
P(M_k) - prior model probability

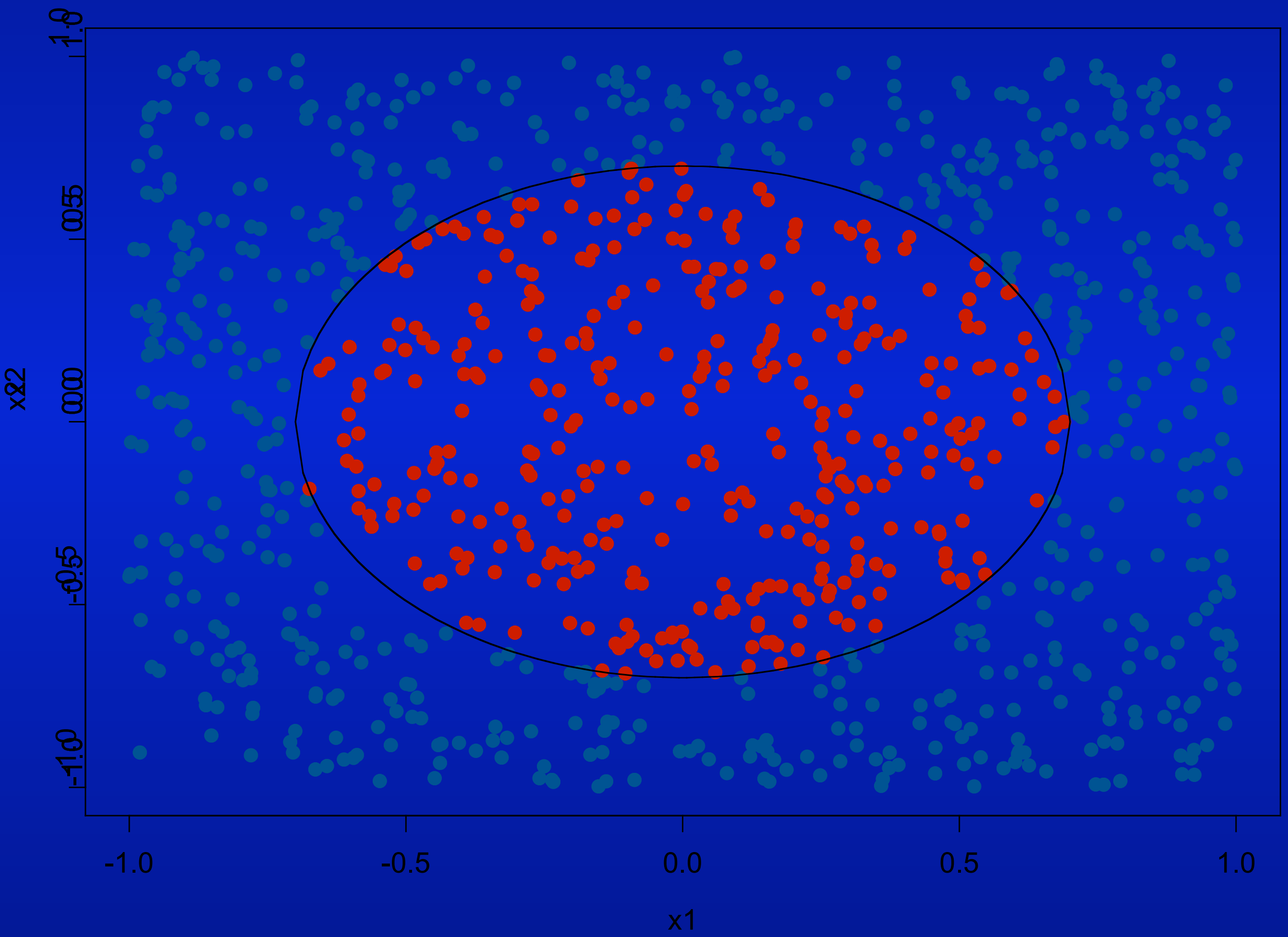
+ Improves predictive performance
+ Theoretically elegant
– Computationally costly

Bagging (*B*ootstrap *Agg*regating) algorithm (Breiman, 1996)

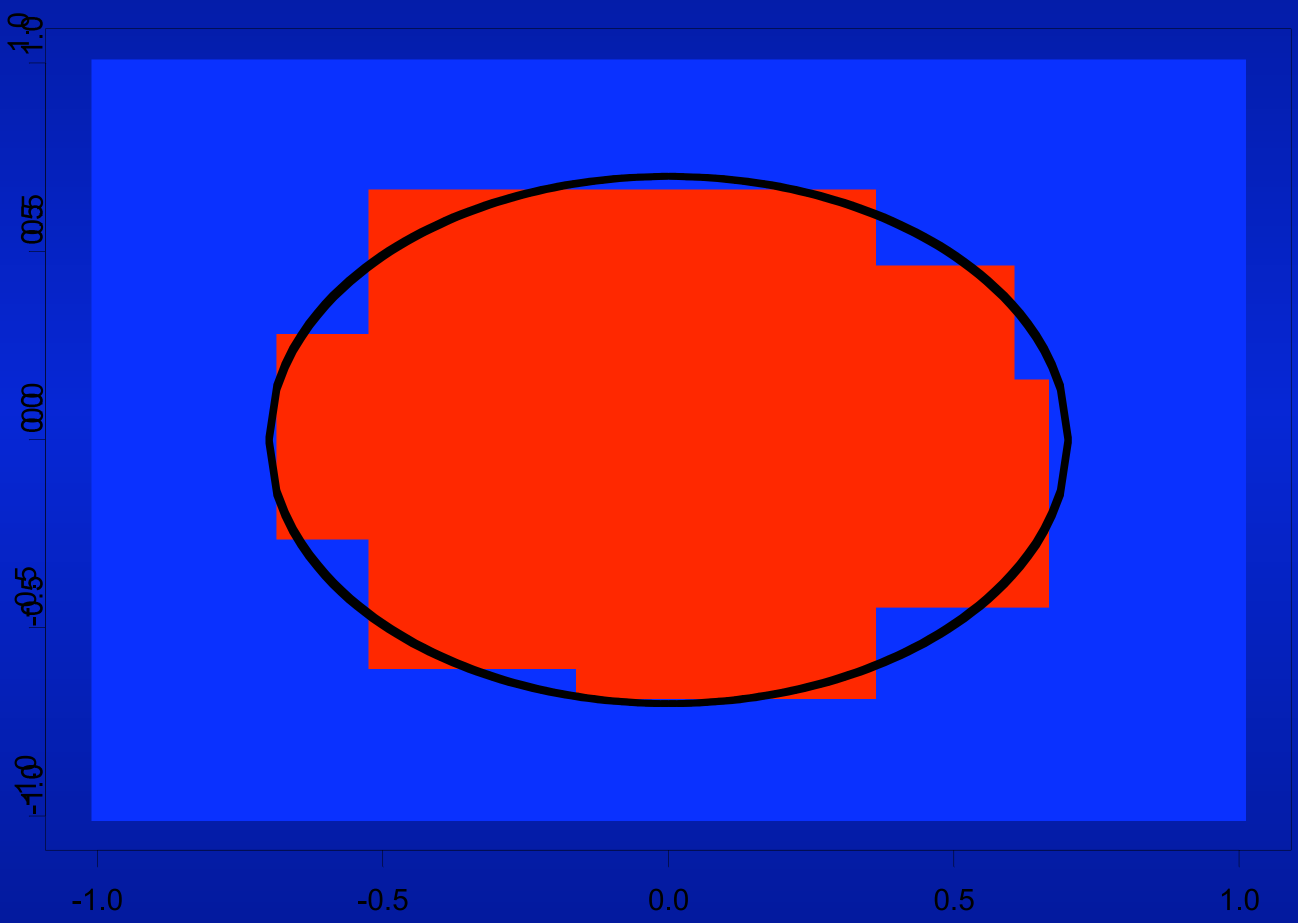
1. Create *K* bootstrap replicates of the dataset.
2. Fit a model to each of the replicates.
3. Average (or vote) the predictions of the *K* models.

Bootstrapping simulates the stream of infinite datasets in a bias-variance decomposition.

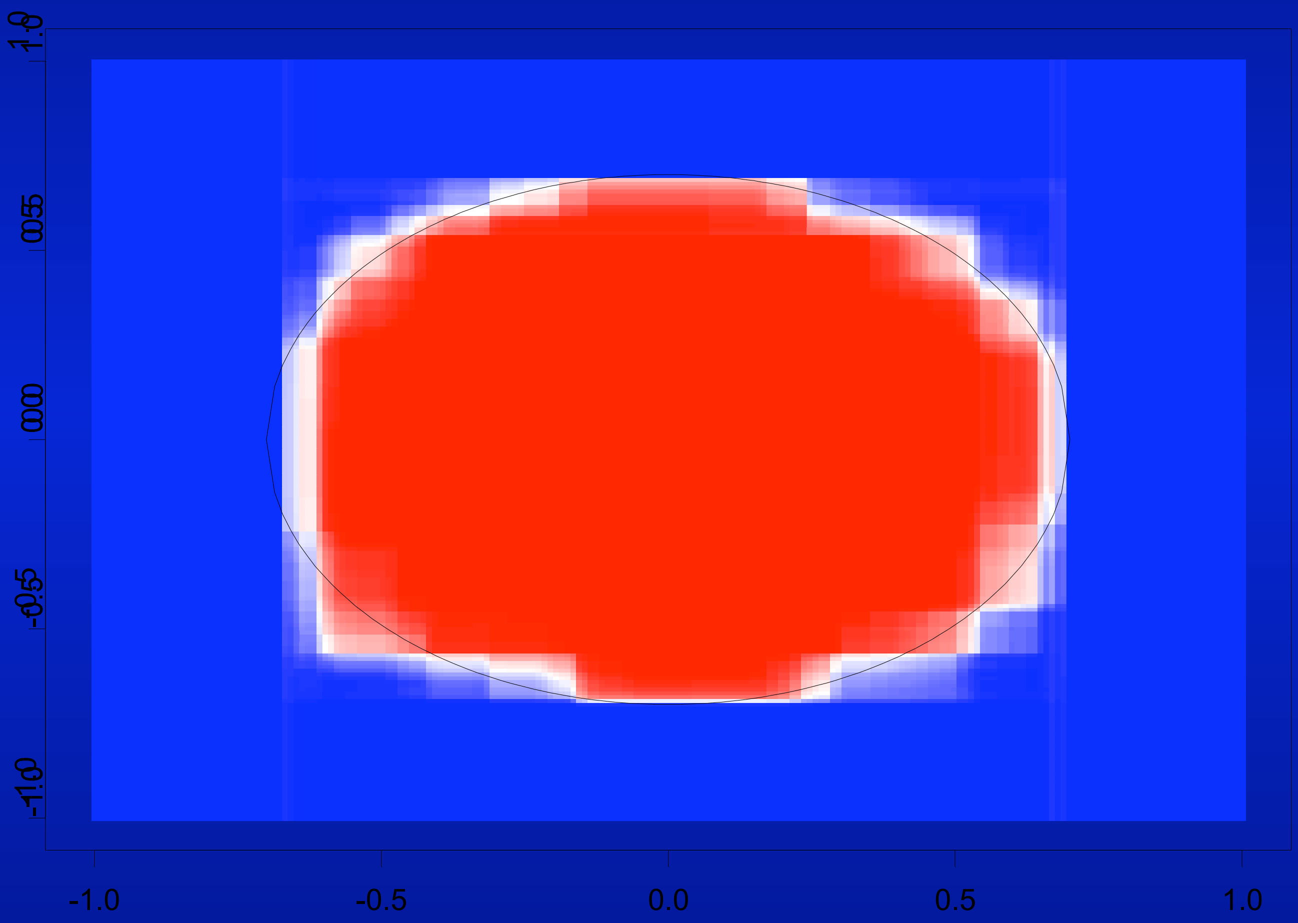
Bagging Example



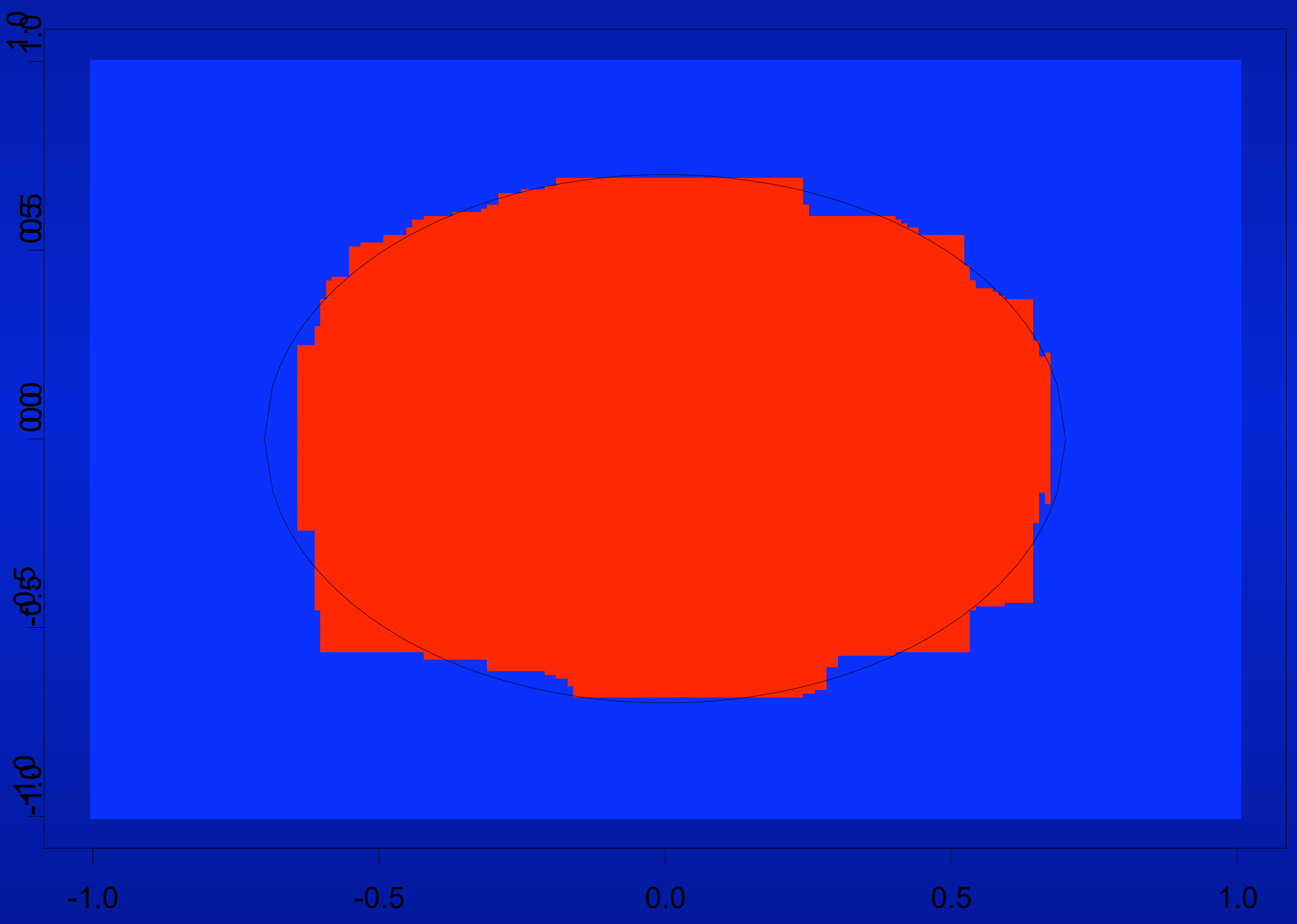
CART decision boundary



100 bagged trees

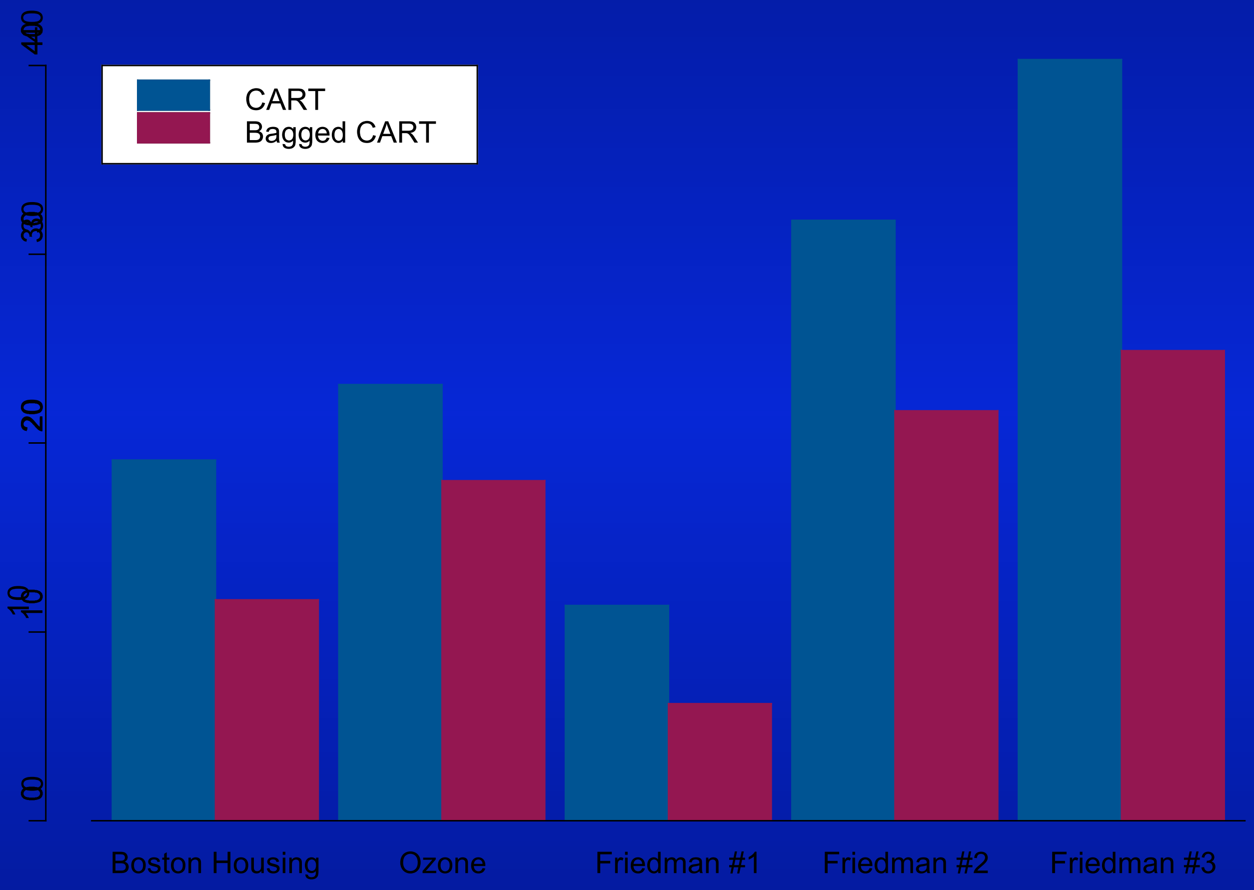


Bagged tree decision boundary



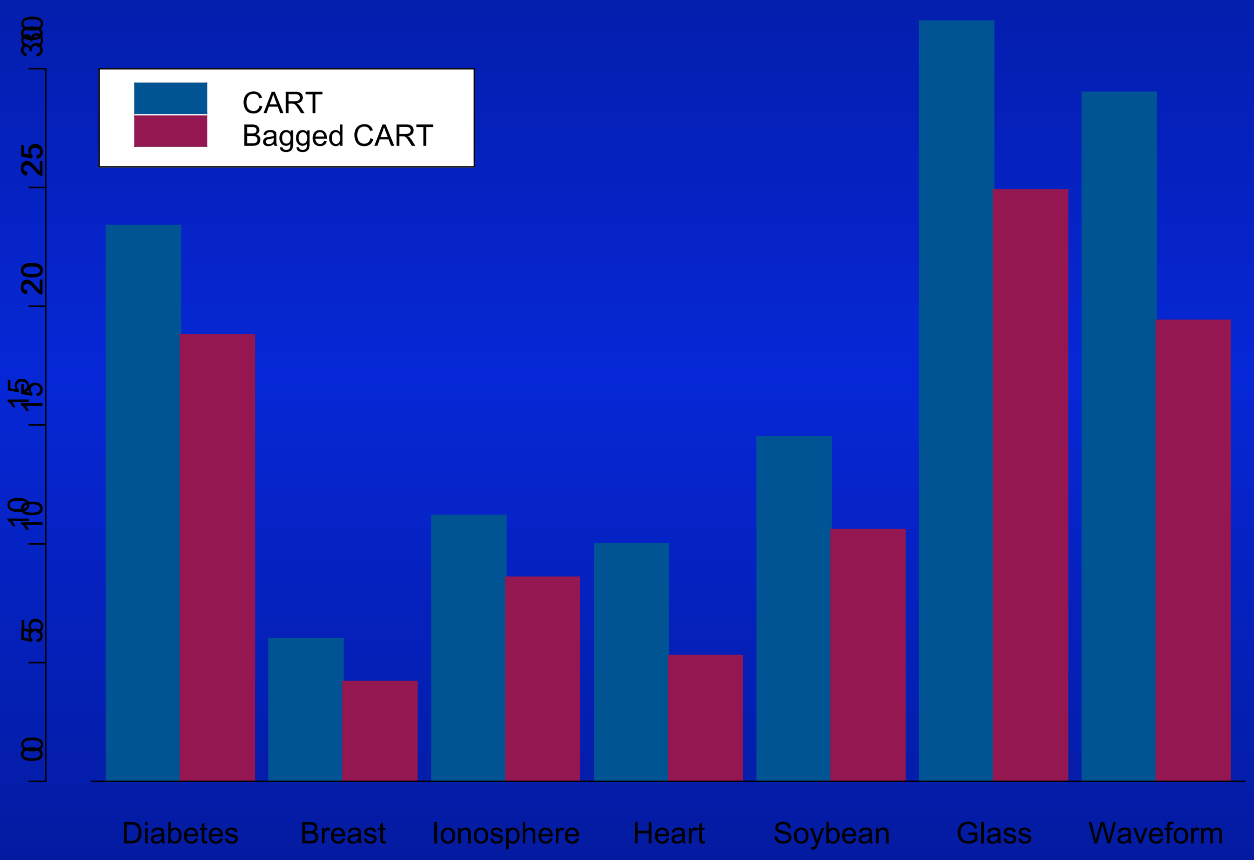
Regression results

Squared error loss



Classification results

Misclassification rates



The Significance of a type of Bundling (Boosting)

"*Boosting* (Freund & Shapiro 1996, Schapiro & Singer 1998) *is one of the most important recent developments in classification methodology.*"

Friedman, Hastie, and Tibsharani (1998), "Additive Logistic Regression: A Statistical View of Boosting", Technical Report, Stanford University.

Boosting algorithm (after Freund & Schapire [1996])

Equally weight the observations $(y,\boldsymbol{x})_i$

For t in $1,\ldots,T$

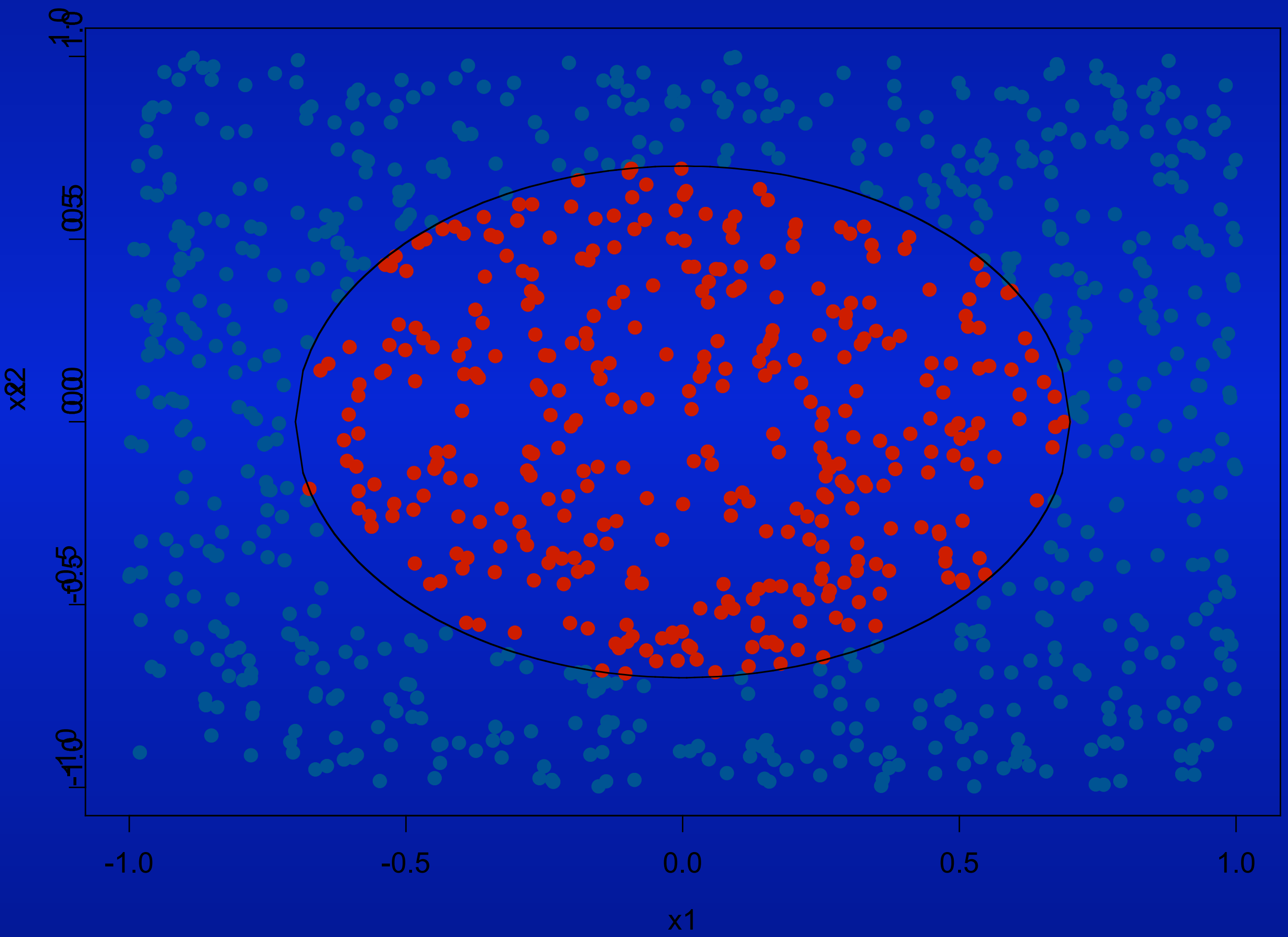
Using the weights, fit a classifier $f_t(\boldsymbol{x}) \rightarrow y$

Upweight the poorly predicted observations

Downweight the well-predicted observations

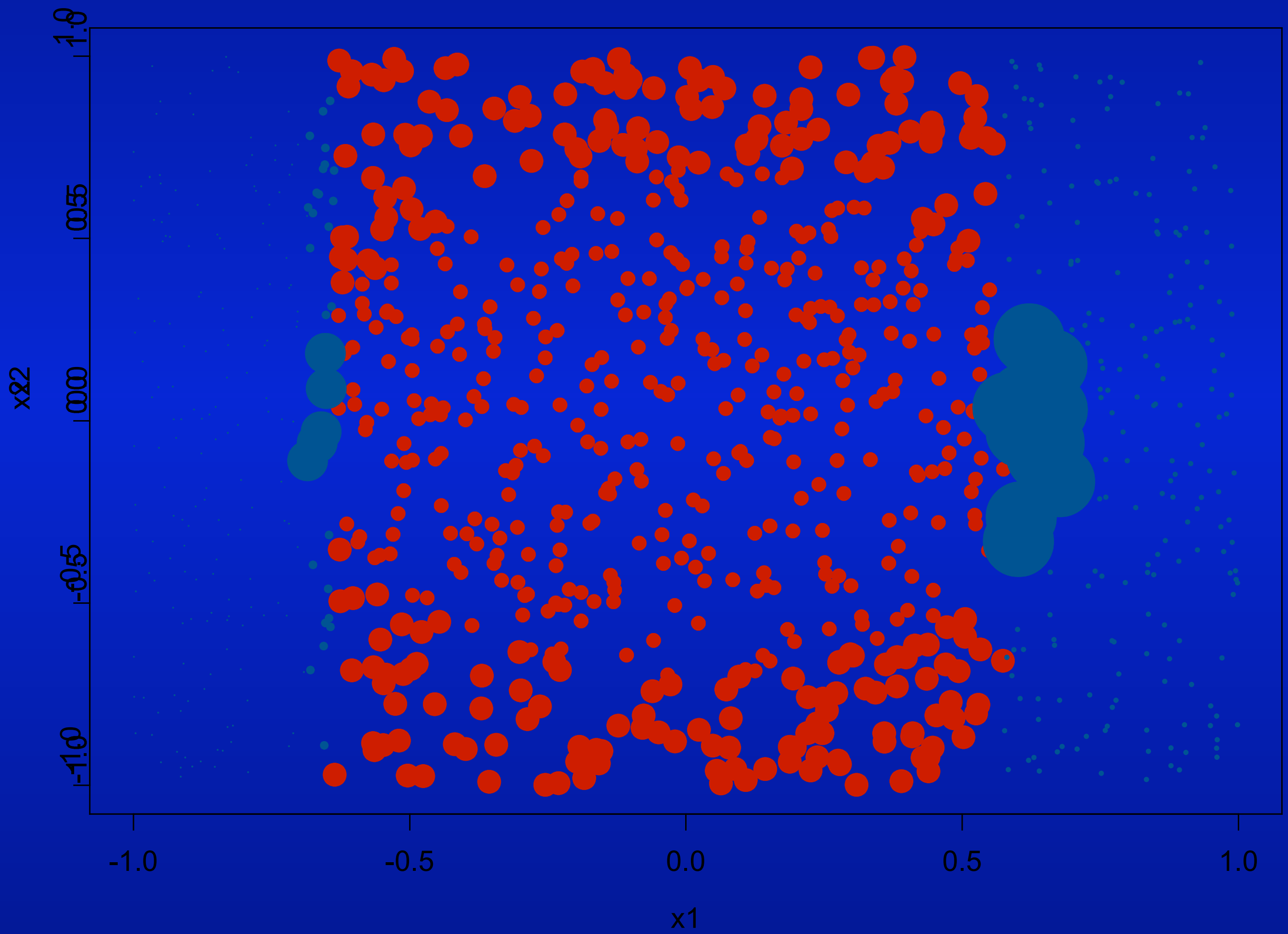
Merge $f_1,\ldots,f_T$ to form the boosted classifier

Boosting Example

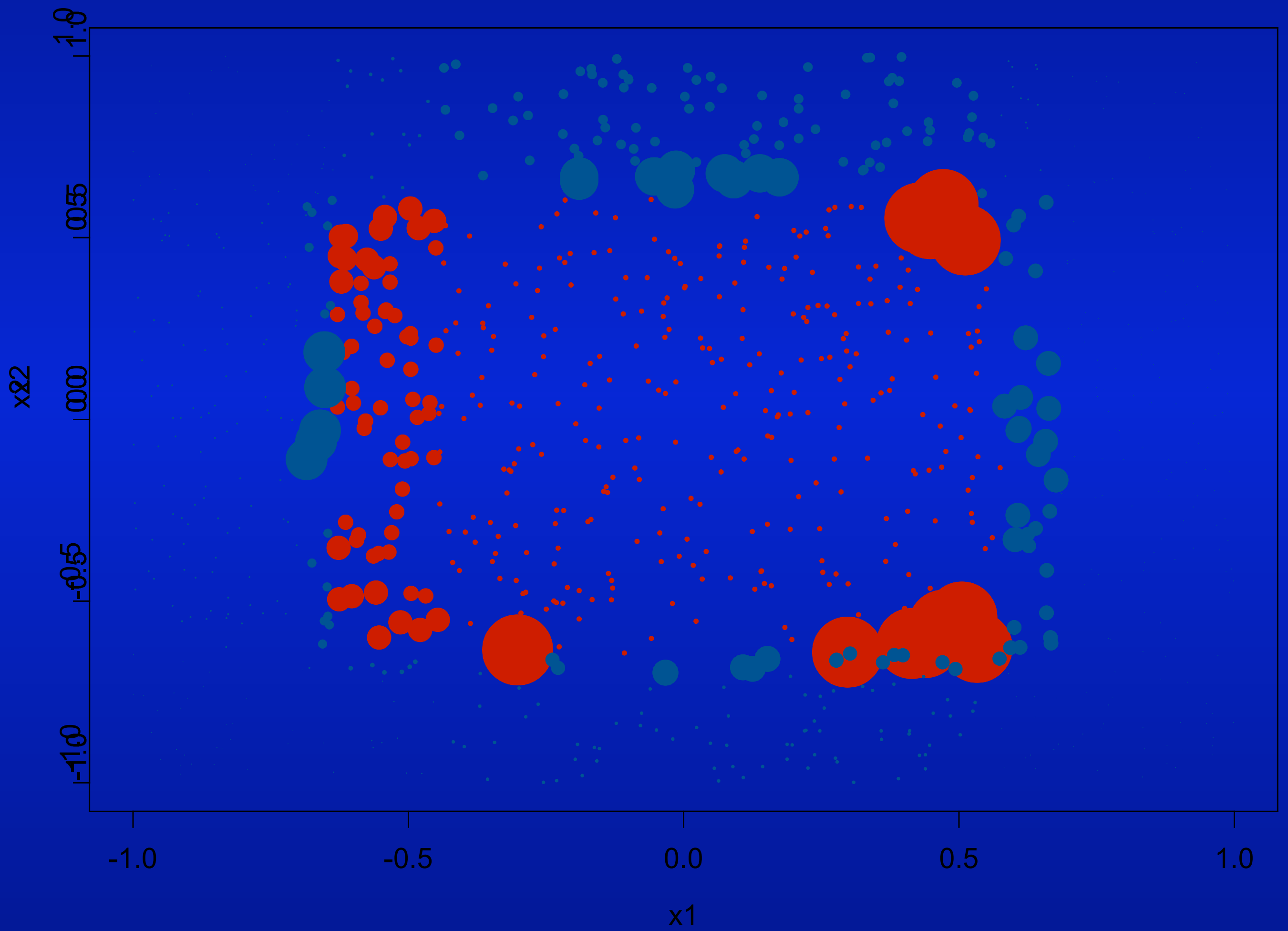


After one iteration

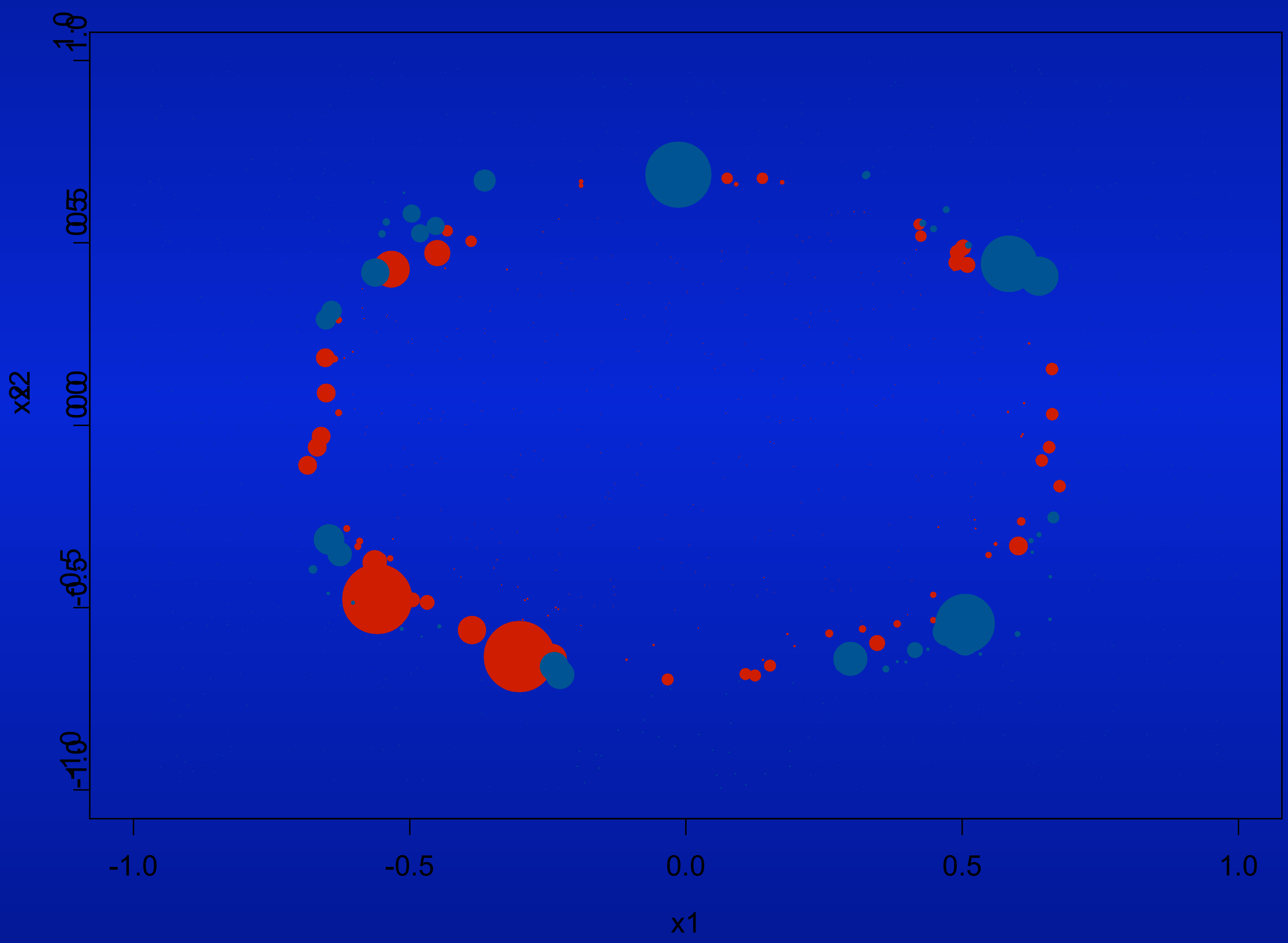
CART splits, larger points have great weight



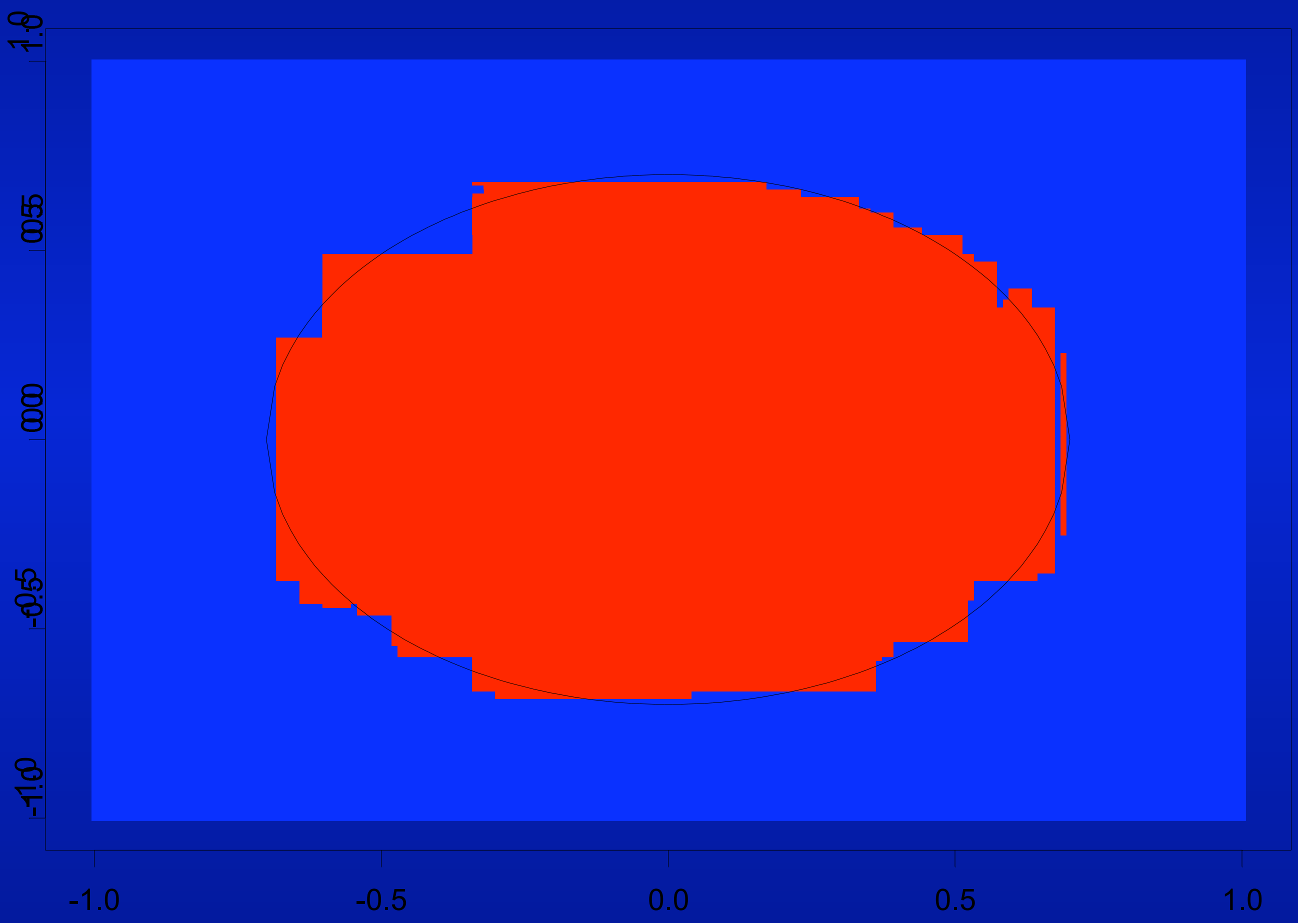
After 3 iterations



After 20 iterations



Decision boundary after 100 iterations



“Bundling” estimators consists of two steps:

1) Construct varied models, and
2) Combine their estimates

Generate component models by varying:

- Case Weights
- Data Values
- Guiding Parameters
- Variable Subsets

Combine estimates using:

- Estimator Weights
- Voting
- Advisor Perceptrons
- Partitions of Design Space, *X*

Other Bundling Techniques

We've Examined:

- ***Bayesian Model Averaging***: sum estimates of possible models, weighted by posterior evidence
- ***Bagging*** (Breiman 96) (*b*ootstrap *agg*regating) -- bootstrap data (to build trees mostly); take majority vote or average
- ***Boosting*** (Freund & Shapire 96) -- weight error cases by $\beta_t = (1\text{-}e(t))/e(t)$, iteratively re-model; average, weighing model *t* by $\ln(\beta_t)$

Additional Example Techniques:

- ***GMDH*** (Ivakhenko 68) -- multiple layers of quadratic polynomials, using two inputs each, fit by Linear Regression
- ***Stacking*** (Wolpert 92) -- train a 2nd-level (LR) model using leave-1-out estimates of 1st-level (neural net) models
- ***ARCing*** (Breiman 96) (Adaptive Resampling and Combining) -- Bagging with reweighting of error cases; superset of boosting
- ***Bumping*** (Tibshirani 97) -- bootstrap, select single best
- ***Crumpling*** (Anderson & Elder 98) -- average cross-validations
- ***Born-Again*** (Breiman 98) -- invent new X data...

Reasons to combine estimators

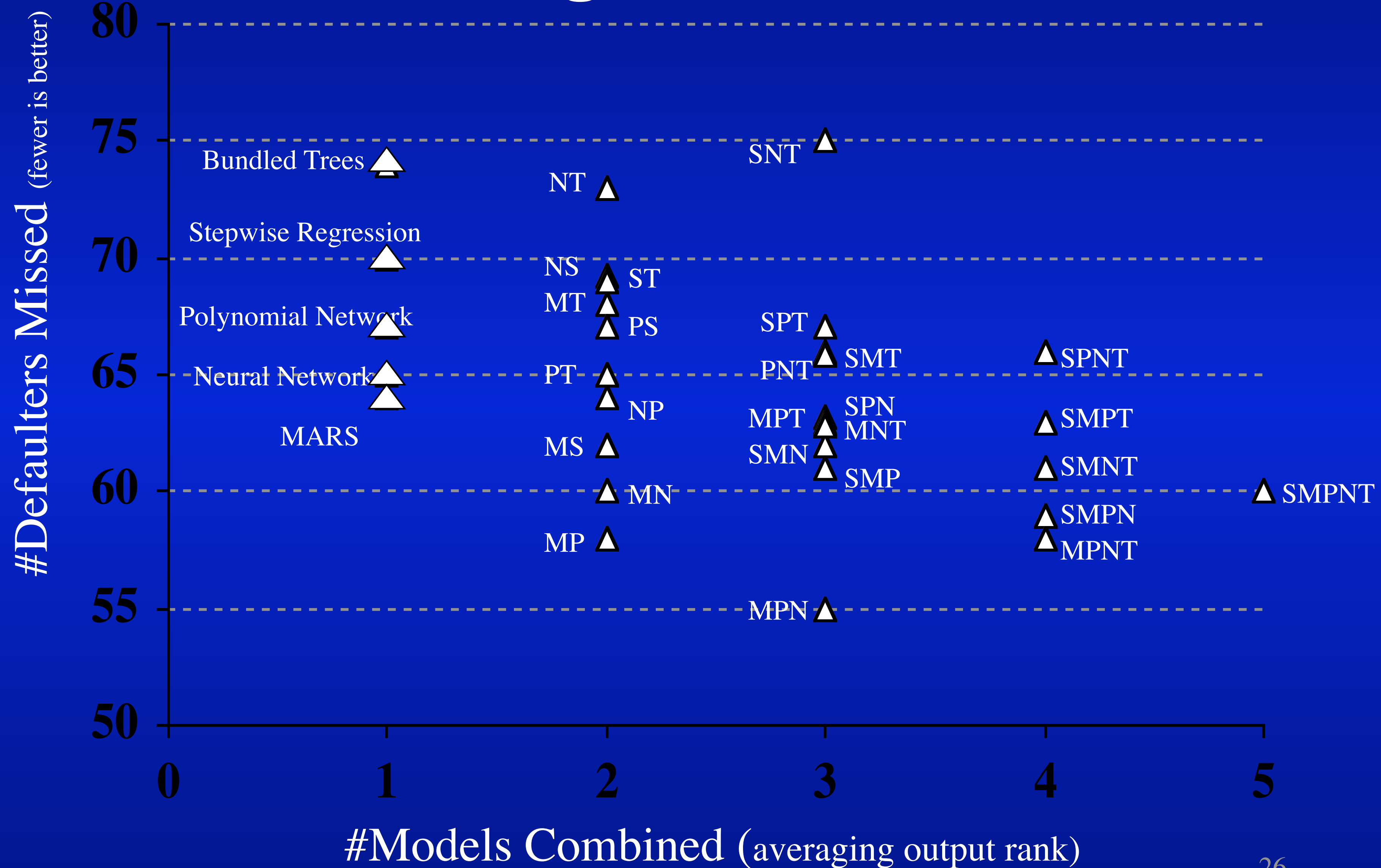
- Decreases variability in the predictions.
- Accounts for uncertainty in the model class.

☆-> Improved accuracy on new data.

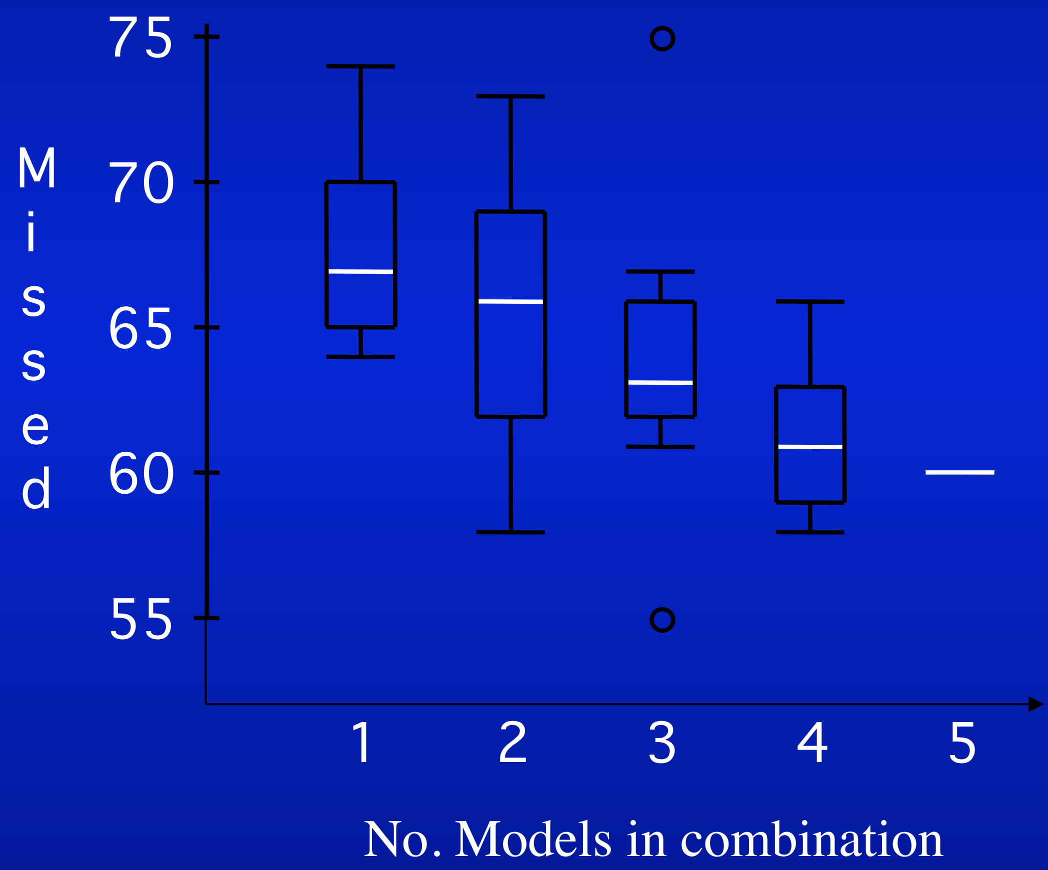
Application Example: Credit Scoring
(Elder Research 1996-1998)

- After 2 years experience, label credit accounts:
  0 (good), 1 (*default* = 90 days late at least once).
- Create models to forecast this outcome
  using only information known at time of credit application.
- Use several (here, 5) different algorithms,
  all employing the same candidate model inputs.
- Rank-order accounts:
  - Give highest-risk value a rank of 1, second highest 2, etc.
  - For bundling, combine model ranks (not estimates) into a new consensus estimate (which is again ranked).
- Report number of defaulting accounts missed (in top portion).

Credit Scoring Model Performance

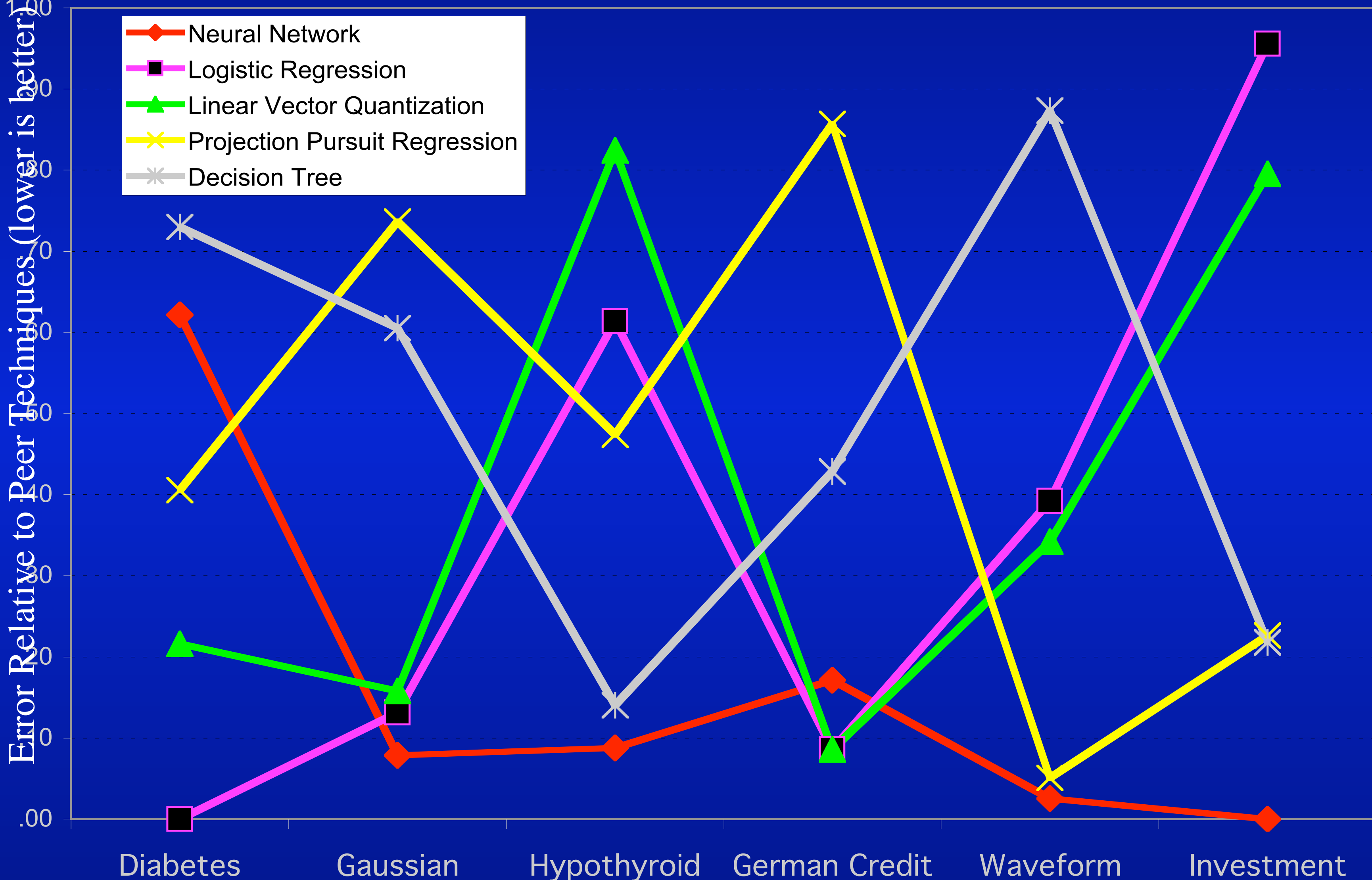


Median (and Mean) Error Reduced with each Stage of Combination

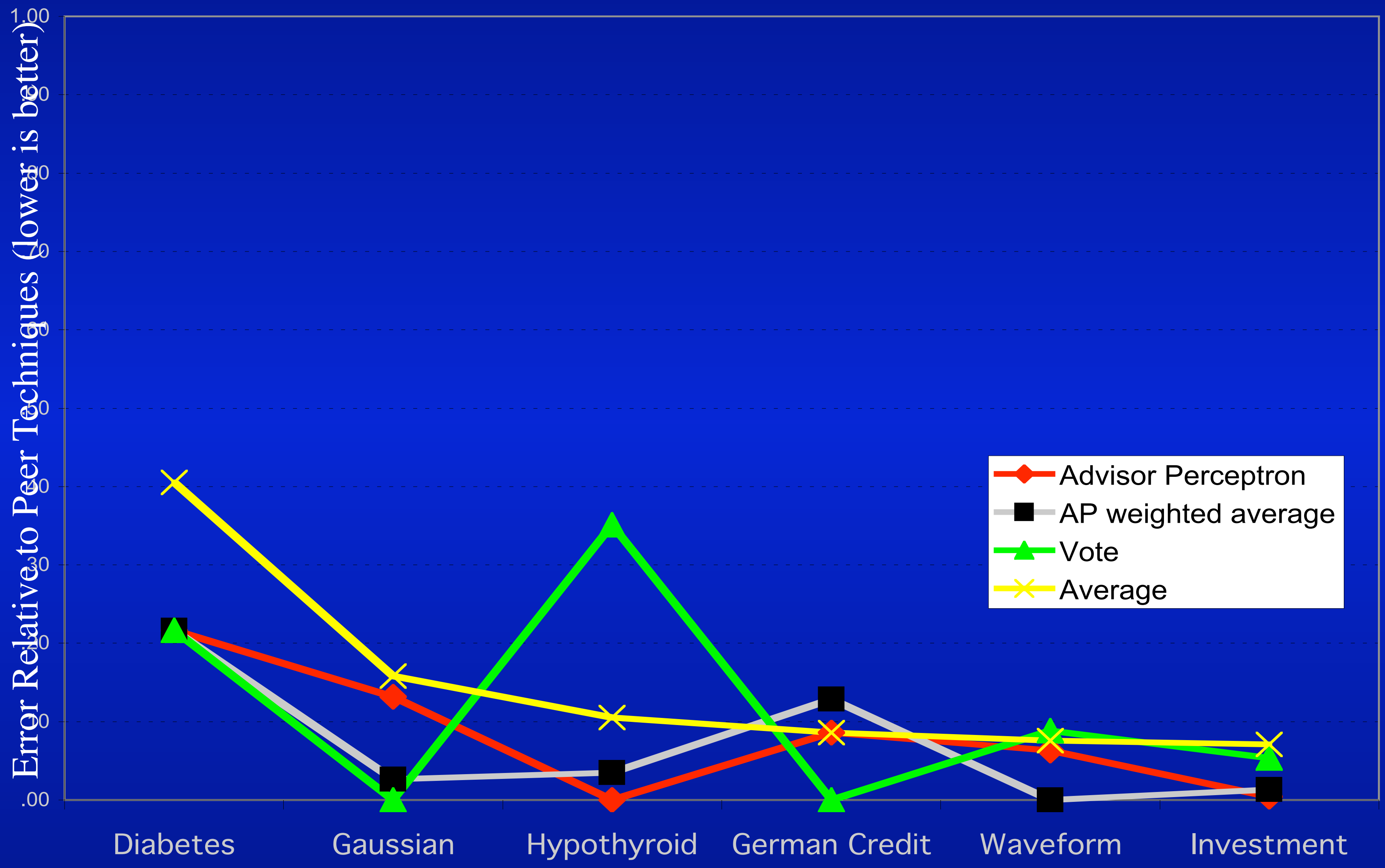


Relative Performance Examples: 5 Algorithms on 6 Datasets

(John Elder, Elder Research & Stephen Lee, U. Idaho, 1997)

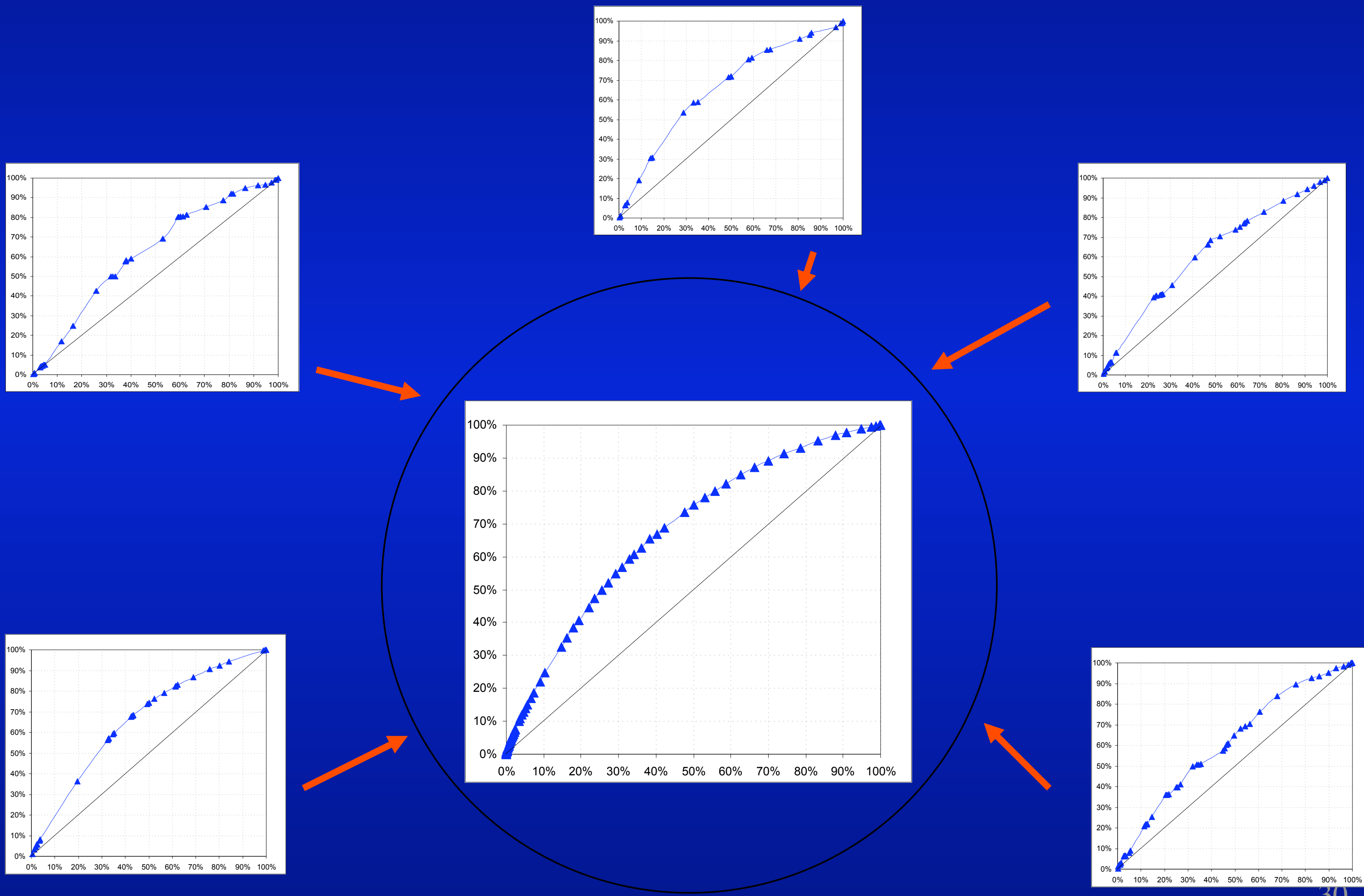


Essentially every Bundling method improves performance



Bundling 5 Trees

Improves lift, smoothness, and possible decision points



Interpreting why Bundling works

- (semi-) Independent Estimators
- Bayes Rule - weighing evidence
- Shrinking (ex: stepwise LR)
- Smoothing (ex: decision trees)
- Additive modeling and maximum likelihood (Friedman, Hastie, & Tibshirani 8/20/98)

… Open research area.
Meanwhile, we recommend bundling competing candidate models both within, and between, model families.

Ensemble Summary

- At very least, compare your method to a conventional one (linear regression say, or linear discriminant analysis).
- The use of multiple approaches can also serve as a useful verification tool. E.g., if one approach used
- Not checking other methods leads to blaming the *algorithm* for the results. But, it's somewhat unusual for the particular modeling technique to make a big difference, and when it will is hard to predict.
- Best: use a handful of good tools. (Each adds only 5-10% effort.)