A Walk through a Random Forest

Andrew Sage

ajsage@iastate.edu

Iowa State University

November 7, 2017

Big Data in Our World

- More data has been created in the last two years than in all of previous human existence.
- Data are used to . . .
  - develop personalized cancer therapies
  - create targeted advertising
  - find winning strategies in sports
  - increase student success in college
- New kinds of data require new statistical methods.

Leo Breiman

The Founder of Random Forest Methodology

http://statistics.berkeley.edu/memory/leo-breiman

Leo Breiman (1928-2005)

- Professor of Statistics, University of California, Berkeley
- Author of *Classification and Regression Trees* (1984)
- *Random Forests* (2001) paper has been cited more than 30,000 times

Iowa State STEM Early Alert

From 2011-2016, 19,081 ISU first-year students chose STEM majors.

- 13% left ISU before start of 2nd year
- 8% stayed at ISU but left STEM

We seek to ...

- identify 2017 first-year STEM majors at risk of leaving STEM.
- notify advisors so they can help these students succeed.
- identify variables that predict a student leaving STEM.

Data & Task

We use prior years' data on 38 variables including:

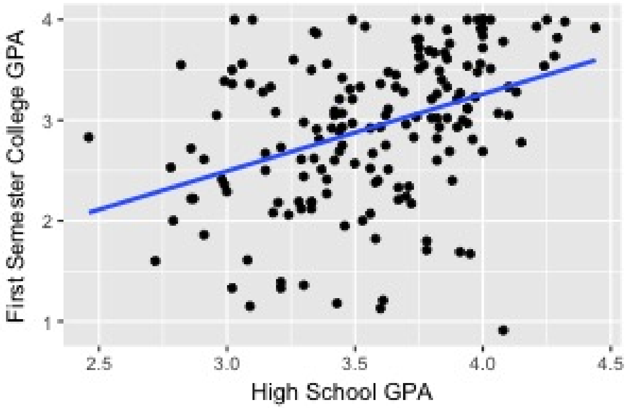
- Demographics
- High school courses, grades and standardized test scores
- Major, first-semester courses, and ISU activities
- Self efficacy and proximal environment (Mapworks® survey)
- Goals and interests (ACT Interest Survey)

For each 2017 first-year STEM student, we want to

1. Predict the student's first-semester GPA.
2. Estimate the probability of the student leaving STEM during first year at Iowa State.

Predicting College GPA

- Predict first-semester GPA for 2014 CS, math, stat majors using simple linear regression on high school GPA.



Expected Semester 1 College GPA $\approx 0.20 + 0.77 \times$ (HS GPA)

Model Assumptions

- Multiple regression can be used to account for variation explained by other explanatory variables. Requires assumptions including
  - expected GPA is a linear function of the explanatory variables
  - there are no interactions between explanatory variables unless we specify them

- Logistic regression is useful in predicting binary outcomes like leaving STEM
  - also requires assumptions about linearity and interactions

- Random forest methodology is a nonparametric, tree-based approach that does not require these assumptions

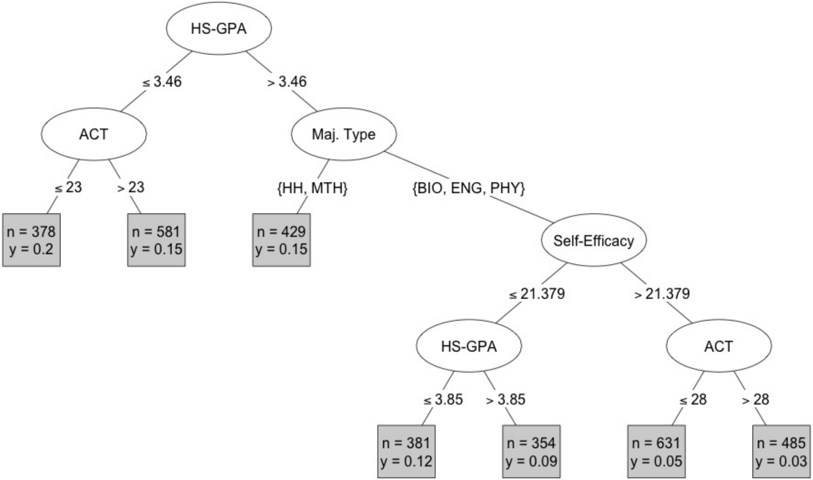
Random Forests

Random forests...

- are grown from decision trees that recursively partition training data so that similar cases are grouped together
- do not require specification of a model
  - no linearity assumptions
  - handles interactions automatically
  - lets the data tell the story
- allow for a large number of predictors
- can handle missing values
- provide a measure of variable importance

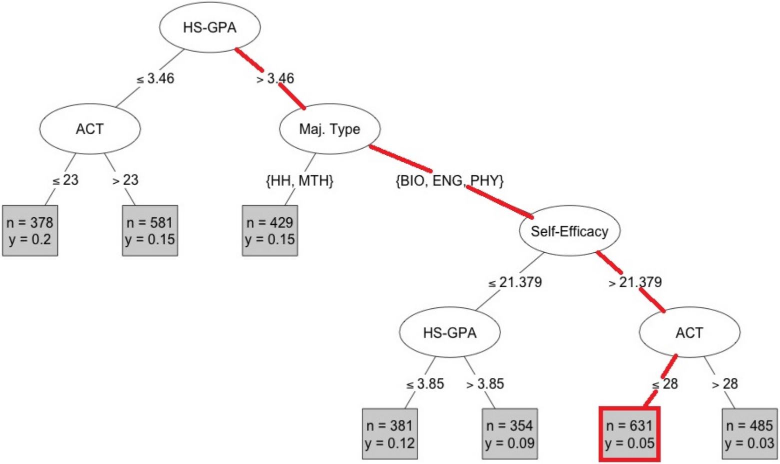
A Decision Tree

First splits in a tree predicting whether a student will leave STEM.
y indicates proportion leaving STEM.



Prediction using Trees

Estimate the leave probability of a BIO major with
HS GPA: 3.81 Self Efficacy:22 ACT:28



From a Tree to a Forest

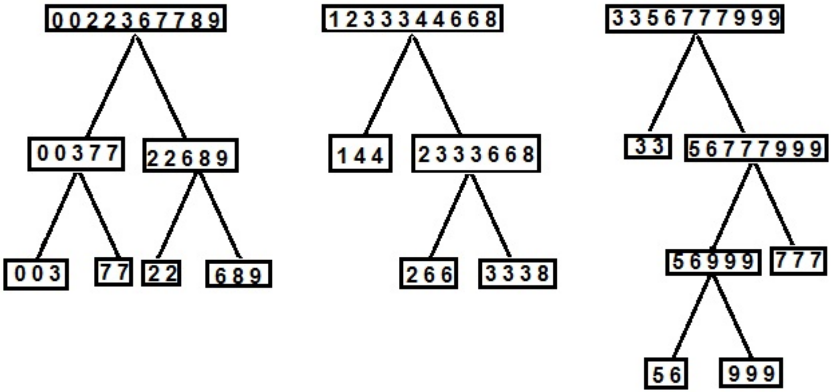
Individual decision trees are often unstable. Small changes in data can lead to large changes in estimates.

A random forest consists of many trees that differ in two ways.

- Each tree is grown using a different random sample of size equal to that of the training data. These samples are selected using replacement and are called bootstrap samples. Cases not used to grow a tree are called out-of-bag (OOB cases).
- A randomly selected subset of predictor variables is considered for each split.

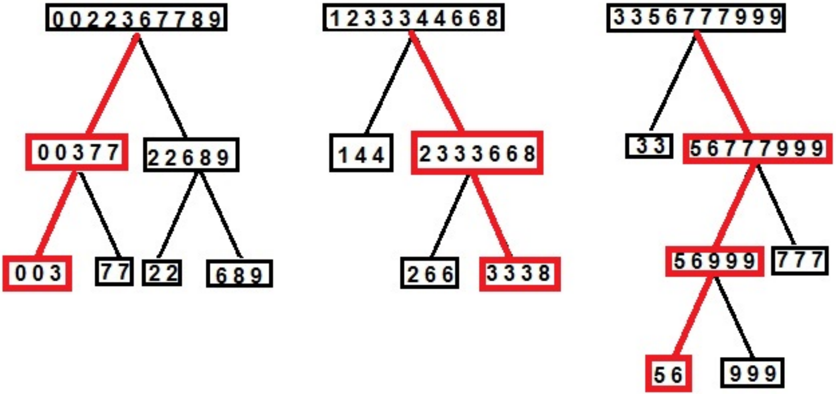
Random Forest Illustration

Consider a dataset of 10 observations with responses $\{0,1,2,3,4,5,6,7,8,9\}$. We grow a (small) random forest of 3 trees.



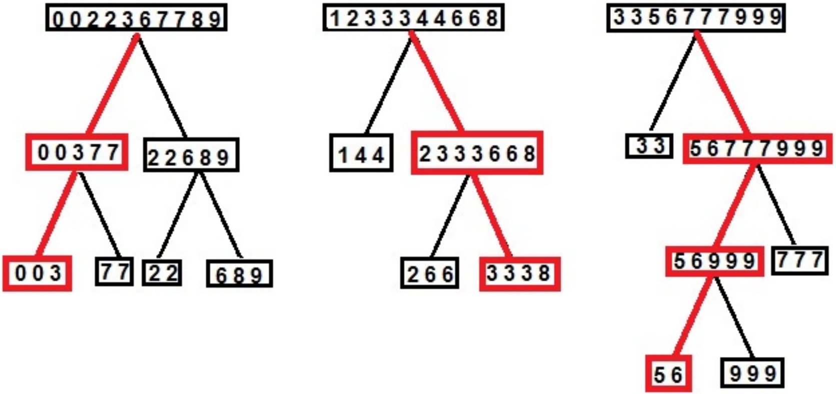
Random Forest Predictions

To make a prediction, run a new case through every tree and average the predictions.



Prediction is $\frac{1+4.25+5.5}{3} \approx 3.5833.$

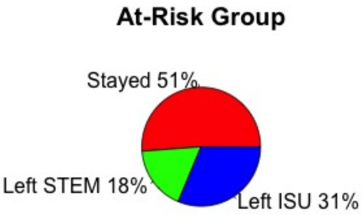
Prediction Case Weights



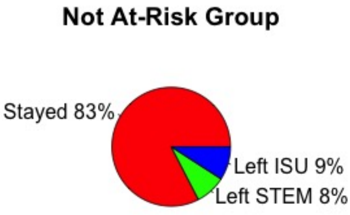
Weight on response 3: $\left(\frac{1}{3}+\frac{3}{4}+0\right)/3 \approx .361$
Weight on response 6: $\left(0+0+\frac{1}{2}\right)/3 \approx .167$

STEM Random Forest Performance

- Grew random forest of 1,000 trees using 2011-14 students.
- Estimated probability of leaving STEM for 2015 students.
- Classified 418 students as at-risk.
- Evaluated performance of model using actual results.



418 students



2,779 students

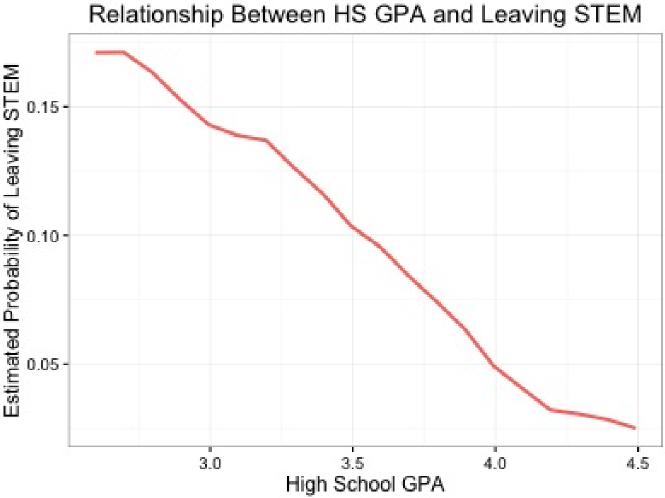
Estimating Response Curve

We can use random forests to estimate the expected response as a function of a predictor variable.

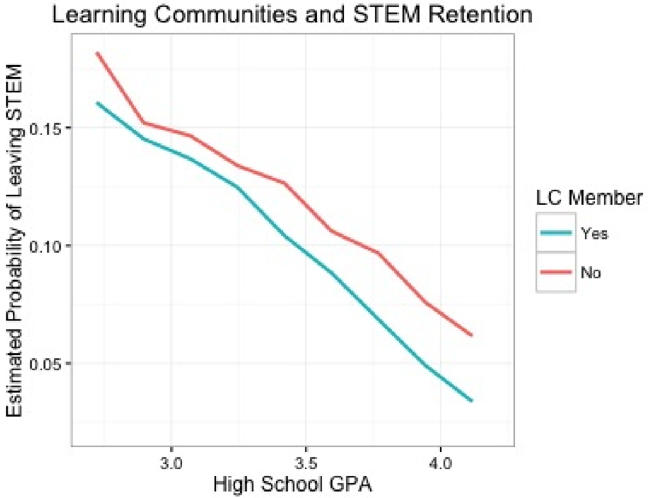
Example:
How is high school GPA related to the probability of leaving STEM?

1. Partition high school GPA into intervals of predetermined size.
2. For each student, estimate probability of leaving STEM using trees where that student was OOB.
3. Average probability estimates for all students in each interval.
4. Plot average probability estimate against average high school GPA in each interval. Connect using line segments, or curve smoothing.

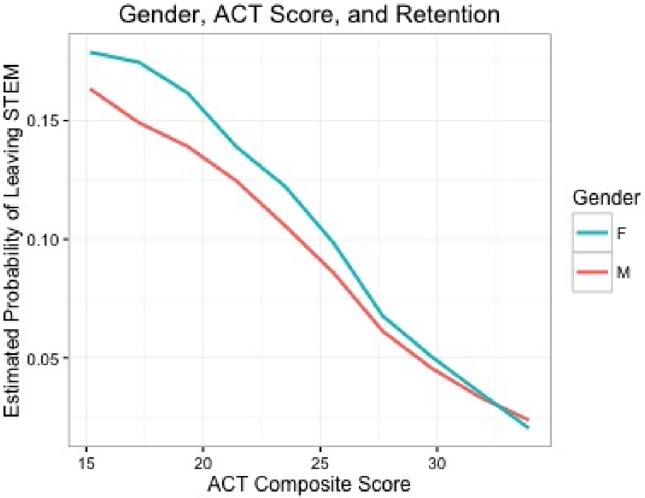
Response Curve for HS GPA



Learning Communities and Retention



Response Curve for ACT and Gender



Measuring Variable Importance

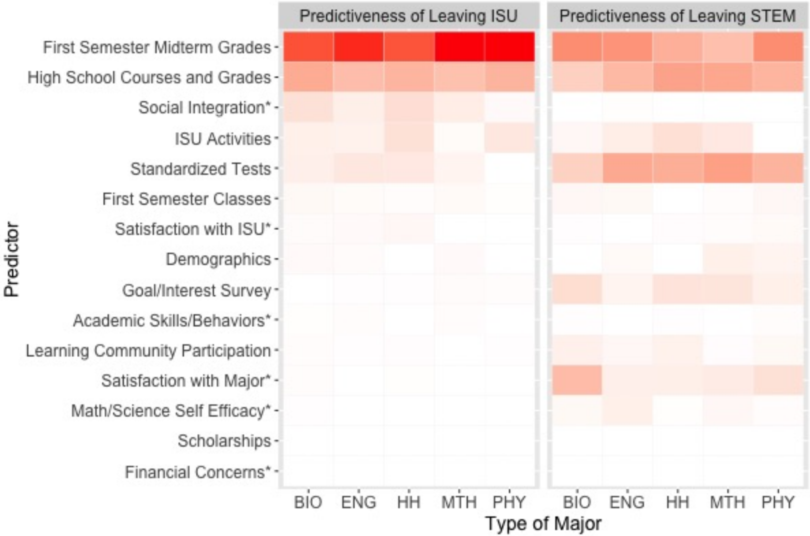
Random forests can be used to determine which variables are most important in a prediction.

To calculate the importance for variable X_j :

1. Make predictions for OOB cases in each tree.
2. Compute mean square error (MSE). Average across trees.
3. Randomly permute values for X_j.
4. Repeat steps 1 and 2.
5. Calculate change in MSE after permutation.

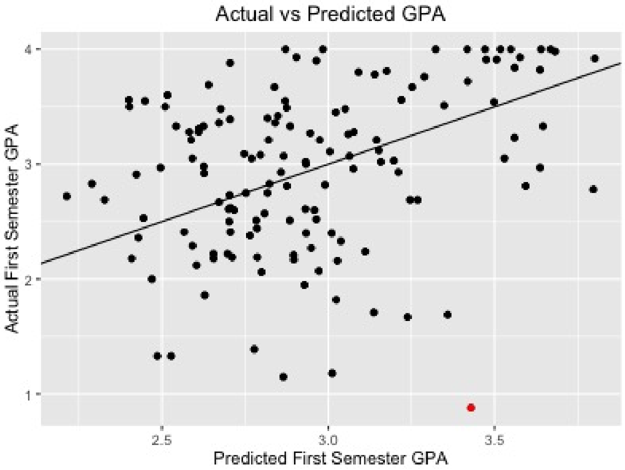
Large increase in MSE $\implies X_j$ important.
Little change in MSE $\implies X_j$ unimportant.

STEM Variable Importance



Random Forests and Outliers

2014 Math, Stat, CS majors:



Examining the Outlier

Student Number 39:

ACT	HS GPA	Gender	Greek Life	Age	LC Member	Math Course	Ed. Goal	Major	Pred. GPA	GPA
28	3.98	M	Yes	19	No	Calc. 1	MS	Math	3.43	0.88

We want to predict a new student's first semester GPA.

New Student:

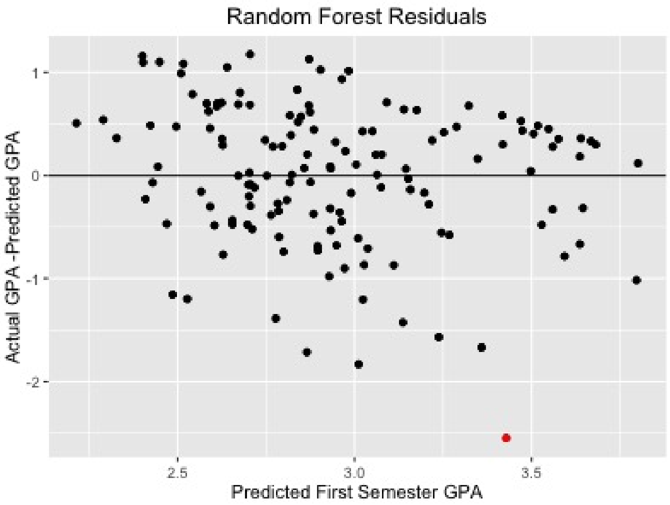
ACT	HS GPA	Gender	Greek Life	Age	LC Member	Math Course	Ed. Goal	Major
27	3.92	M	Yes	18	Yes	Calc. 2	MS	Math

RF Prediction Weights

No.	Sem. 1 GPA	Pred. Weight
39	0.88	.189
90	4.00	.108
116	4.00	.037
34	3.21	.037
82	4.00	.035
69	2.16	.035
53	3.23	.027
146	3.92	.027
56	2.81	.026
110	4.00	.026
153	3.54	.021
50	3.72	.019
⋮	⋮	⋮

$$\text{Predicted GPA} = \sum \text{Sem.1 GPA} \times \text{Pred. Weight}$$
$$= 2.95$$

Residual Plot



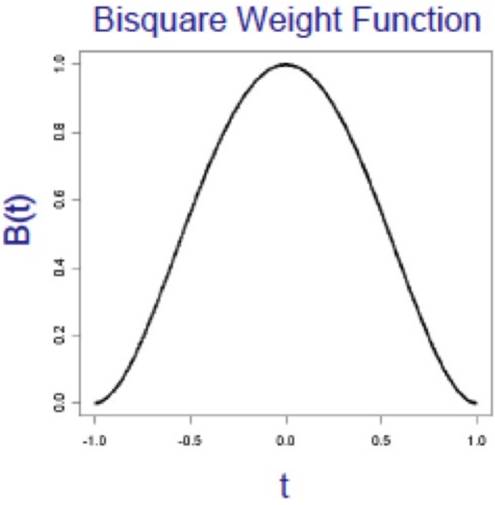
Robust RF Algorithm

Motivated by Cleveland (1979)

1. Calculate all residuals, e_k.
2. Let $M = \text{Median}(|e_k|)$.
3. $\delta_k = B\left(\frac{e_k}{\alpha M}\right)$
   where
   $$B(t) = \begin{cases} (1-t^2)^2 & \text{if } |t| < 1 \\ 0 & \text{if } |t| \geq 1 \end{cases}$$
4. Replace weight w_k with $w_k\delta_k$.
5. Rescale so weights add to 1.
6. Repeat (1)-(5) iteratively.

1. $e_{39} = 0.88 - 3.43 = -2.55$
2. $M = 0.47$.
3. Using $\alpha = 6$,
   $$\delta_k = B\left(\frac{-2.55}{6(0.47)}\right) = 0.03411$$
4. $$w_k\delta_k = (0.189)(0.03411) = 0.0064$$
5. New weight is 0.009.

Tukey's Bisquare Function



Robust RF Prediction Weights

No.	Sem. 1 GPA	Pred. Weight	Residual	Adj. Pred Weight
39	0.88	.189	-2.55	.009
90	4.00	.108	1.02	.113
116	4.00	.037	0.58	.047
34	3.21	.037	0.24	.050
82	4.00	.035	0.67	.043
69	2.16	.035	-0.87	.039
53	3.23	.027	-0.33	.036
146	3.92	.027	0.12	.036
56	2.81	.026	-0.78	.030
110	4.00	.026	0.53	.033
153	3.54	.021	0.04	.029
50	3.72	.019	0.40	.025
⋮	⋮	⋮	⋮	⋮

$$\text{Predicted GPA} = \sum \text{Sem.1 GPA} \times \text{Adj. Pred. Weight}$$
$$= 3.44$$

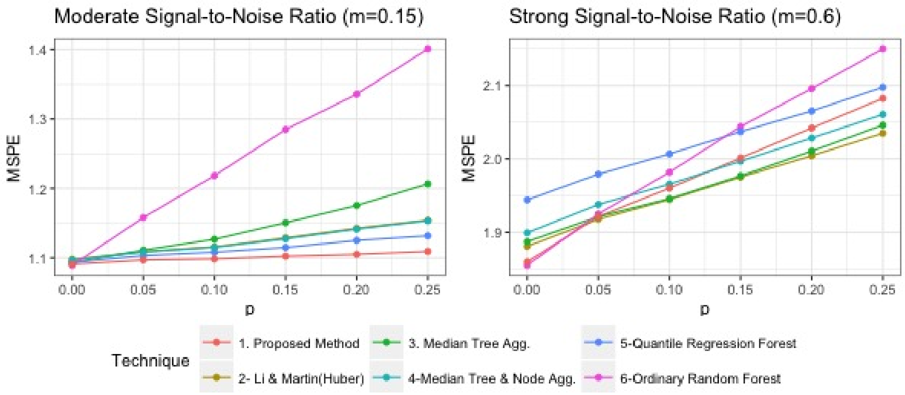
Simulation Study

Tested approach using simulation studies from Roy & Larocque (2013)

$$
\begin{aligned}
Y_i = & m \times \Big(X_{1i} + 0.707X_{2i}^2 + \mathbb{I}(X_{3i} > 0) + 0.873\text{log}(|X_{1i}|)X_3 \\
& + 0.894X_{2i}X_{4i} + 2\mathbb{I}(X_{5i} > 0) + 0.464\text{exp}(X_{6i}) \Big) \\
& + \epsilon_i\mathbb{I}(r_i > p) + \delta_i\mathbb{I}(r_i < p)
\end{aligned}
$$

- $X_i \sim \mathcal{N}(0,1)$, $\epsilon_i \sim \mathcal{N}(0,1)$, $\delta_i \sim \mathcal{N}(0,5)$, $r_i \sim$ Uniform(0,1)
- $p = \%$ of training data from contaminating distribution
- small $m \implies$ noisy data, large $m \implies$ strong signal
- Contamination occurs in training data but not test data
- Simulated 500 repetitions consisting of 500 training cases and 1,000 test cases

Simulation Results



- Proposed method outperforms others for moderate signal-to-noise
- Consistently beats original random forest and competitive with other robust approaches

Future Research

- Random forests might be used to estimate quantities such as
  - risk of disease
  - likelihood of mortgage default
  - probability of winning a sporting event or election

- Methodological research might
  - continue to enhance predictive performance
  - make random forests more interpretable
  - find ways to optimally combine random forest predictions with other methods

Acknowledgements

I would like to acknowledge...

- my major professors Ulrike Genschel and Dan Nettleton
- STEM retention collaborators Cinzia Cervato and Craig Ogilvie
- The Howard Hughes Medical Institute, whose grant to Iowa State University in part supports the STEM retention research

Concluding Remarks

http://statistics.berkeley.edu/memory/leo-breiman

Remember that the great adventure of statistics is in gathering and using data to solve interesting and important real world problems.
-Leo Breiman