# Models for Millions

Bob Stine
Department of Statistics
The Wharton School, University of Pennsylvania
stat.wharton.upenn.edu/~stine

34th NJ ASA Spring Symposium
June 7, 2013

# Introduction

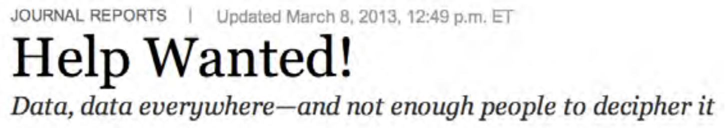
# Statistics in the News

- Hot topics
  - Big Data
  - Business Analytics
  - Data Science
- Are the authors talking about statistics?
  - Or about …

    information systems?

    database technology?

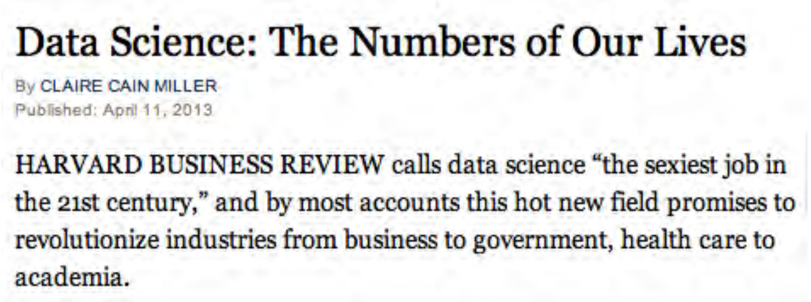
    visualization, eye candy?



JOURNAL REPORTS | Updated March 8, 2013, 12:49 p.m. ET

**Help Wanted!**

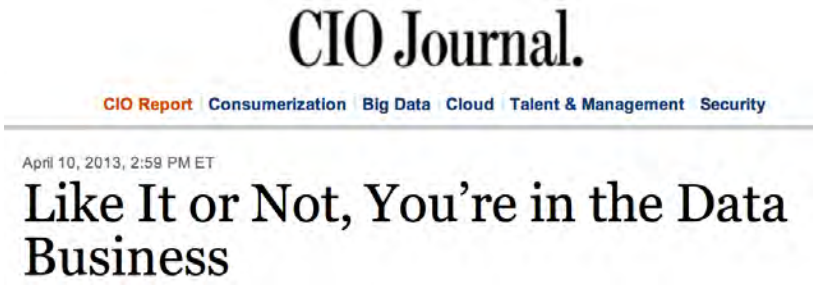
*Data, data everywhere—and not enough people to decipher it*



**Data Science: The Numbers of Our Lives**

By CLAIRE CAIN MILLER
Published: April 11, 2013

HARVARD BUSINESS REVIEW calls data science “the sexiest job in the 21st century,” and by most accounts this hot new field promises to revolutionize industries from business to government, health care to academia.



**CIO Journal.**

CIO Report | Consumerization | Big Data | Cloud | Talent & Management | Security

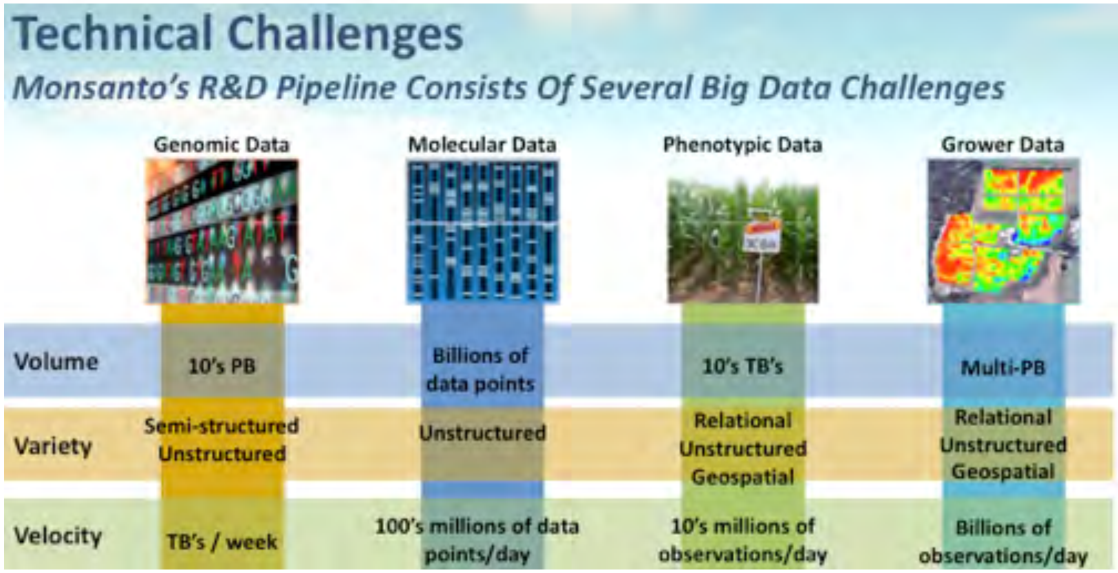
April 10, 2013, 2:59 PM ET

**Like It or Not, You’re in the Data Business**

# Even Farming…

## How B.I. and Data Make a More Efficient Farm

by David Strom | September 17, 2012



### Technical Challenges

*Monsanto's R&D Pipeline Consists Of Several Big Data Challenges*

| | Genomic Data | Molecular Data | Phenotypic Data | Grower Data |
|---|---|---|---|---|
| Volume | 10's PB | Billions of data points | 10's TB's | Multi-PB |
| Variety | Semi-structured Unstructured | Unstructured | Relational Unstructured Geospatial | Relational Unstructured Geospatial |
| Velocity | TB's / week | 100's millions of data points/day | 10's millions of observations/day | Billions of observations/day |

### ALPRO™ – Milking

From milking point control and monitoring milking performance, to checking that your milking protocols are adhered to, ALPRO delivers the information you need to fine-tune your milk production.

Business intelligence: it's not just for big-city businesses anymore.

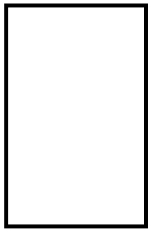
# Big Data

Notation

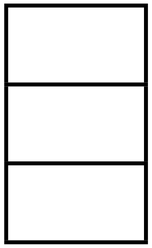
n = # rows of X

p = #columns of X

- Recent modeling projects
- Credit scoring
  - 75,000 cases
  - 15,000+ possible explanatory variables

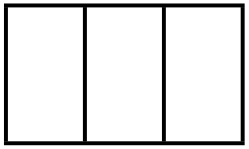


- Spatial time series
  - 3,000 locations
  - 100 time points
  - 20+ features at each location and time



- Text
  - Real estate listings
    - 6,000 prices, millions of possible descriptions

  - Tagging
    - 1.2 million words, 60,000+ ‘explanatory variables’



# Statistics in Text Mining

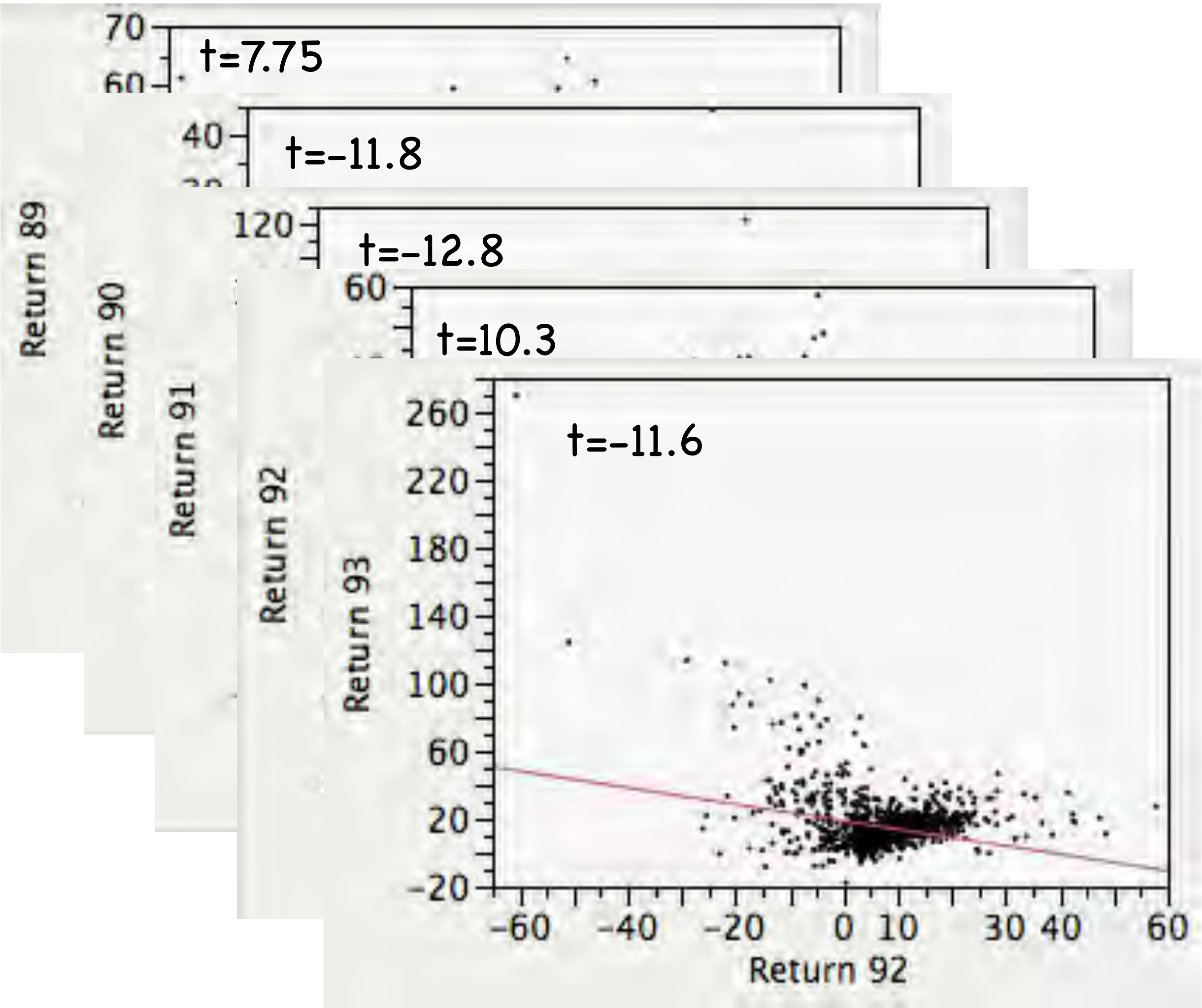
- Key tool
  - Statistics plays big role in various algorithms
- Part-of-speech (POS) tagging
  - Identify usage of ‘bank’: noun, verb, pronoun…
  - I made a deposit at the bank.
  - The pilot needed to bank the plane.
  - Stanford tagger uses bidirectional logistic regr
  - Hidden Markov models are common
- Sentence parsing
  - ‘Diagram’ sentence, recognizing
    subject, predicate, object
  - Maltparse has plug-in statistical engine

# Is Big Data Really So Big?

- Not so large as they first seem
  - Repeated measurement ≠ more degrees of freedom
  - What is the relevant source of variation?
- Transfer learning problem
  - Machine learning
  - Build model for structure of text on corpus such as the New York Times
  - What transfers from that model to
    Washington Post?
    Richmond Times-Dispatch?
- Implications for estimates of standard error

# Example of Dependence

- Predict returns on mutual funds
  - Do funds that do well in one year anticipate doing well (or poorly) the next year?



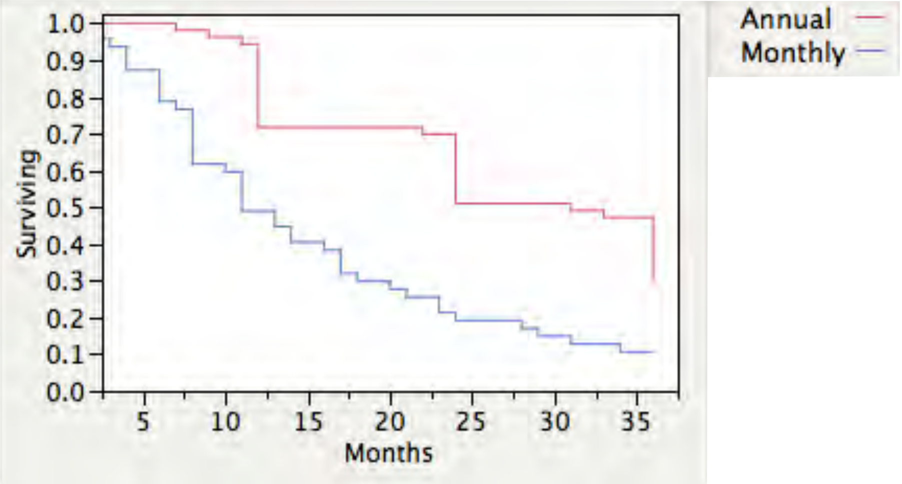
# Does Big Data Imply Big Models?

- Perhaps all one needs is a very simple analysis
  - Google
  - Massive hardware
  - Extensive data
- Text modeling
  - Hard problem: predict next word in sentence

    I took a walk _____
  - Tabulation of all 5-grams (5 word sequence)
  - Replace modeling with frequency table
- Web page design
  - Continuous experimentation
  - Randomized, two-sample t-test

# Simple Models Can Be Better

- Association rules
  - Low tech...
    Build tables
  - Identify association
  - Low-tech ≠ low impact...
    grab low-hanging fruit

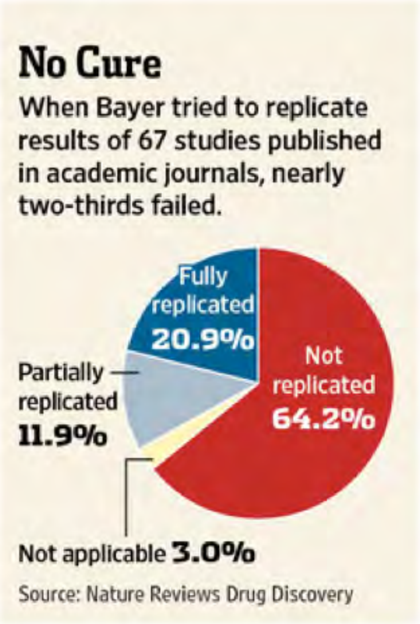


- Predictive modeling via support vector machine
  - High tech...
    Locate separating hyperplanes in kernel space
  - Identify predictive features
  - High-tech ≠ high impact...
    Complexity vs communication

Wharton
Department of Statistics

# Simple might be right!

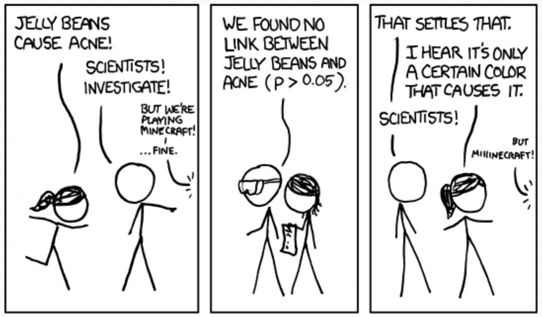
- Recent WSJ story on reproducibility and proliferation of research…



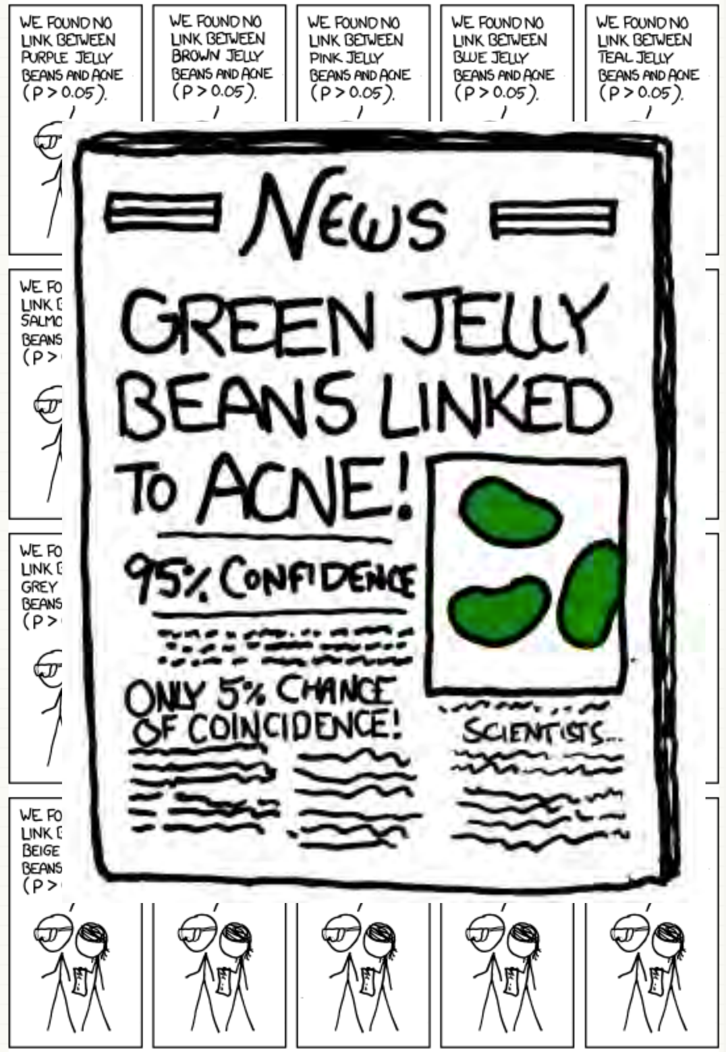
**No Cure**

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discovery



JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!



WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FO LINK SALMO BEANS (P >
WE FO LINK GREY BEANS (P >
WE FO LINK BEIGE BEANS (P >
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

# Attractive Misconceptions*

- Outliers don't matter with millions of cases
  - Central limit theorem
  - Corollary: estimators are normally distributed.
- Thinking the true predictor is in my data rather than running an experiment
  - Reject inference and white cars
  - Training: we give students the data
- I can treat methods as black boxes
  - Lasso is popular, so it's best for my application.
- Cross-validation keeps me out of trouble
  - As long as the model validates well out-of-sample, the predictions are reliable.

*ie, Lessons I have learned the hard way.

# Plan

- Familiar context
  - Fit LS regression of continuous Y to large collection of possible explanatory variables
- Two themes
  - Reducing dimensions
    - Columns: Random projections
    - Row: Subsampling
  - Streaming
    - Sequential from rows
    - Sequential from columns
  - Mixtures of the two (VIF regression)
- Comments
  - Regularization (shrinkage) can be added
  - Where are the Bayesian models?

# Dimension Reduction

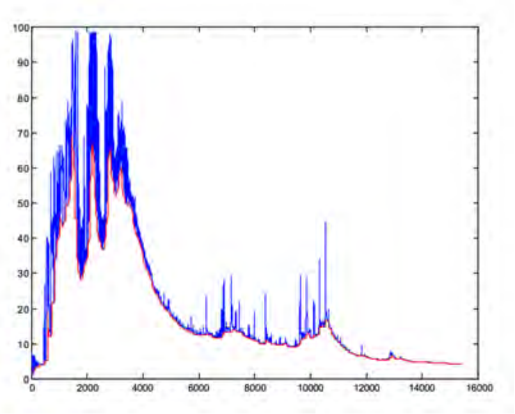
# Reducing Columns

- Context
  - PCA, common column scales
  - Huge $p \gg n$
- Random projection
  - Methods based on random projection have brought a rebirth of PCA
- Idea
  - Use random projections to reduce the data matrix to a size amenable to calculation.
  - Explanatory variables in $n \times p$ matrix $X$
  - Pick $d \ll p$
  - Multiply $X$ by a $p \times d$ matrix of random numbers $\Omega$ so that resulting dimension is $n \times d$.

# Arcene Example

- Automation
  - Automated data collection produces extensive measurements, here p=10,000 features
  - Only n=200 cases

- Arcene example from UCI
  - Mass spectrometer measurements
  - Origin: Separate normal cells from cancerous cells (prostate, ovarian)
  - Make into a regression problem
    - Use continuous response, not the 0/1 indicator in res



- Complications galore...
  - Collinear from sampling smooth function
  - Problem of too many 'perfect' solutions
  - Hard to test out-of-sample because few cases

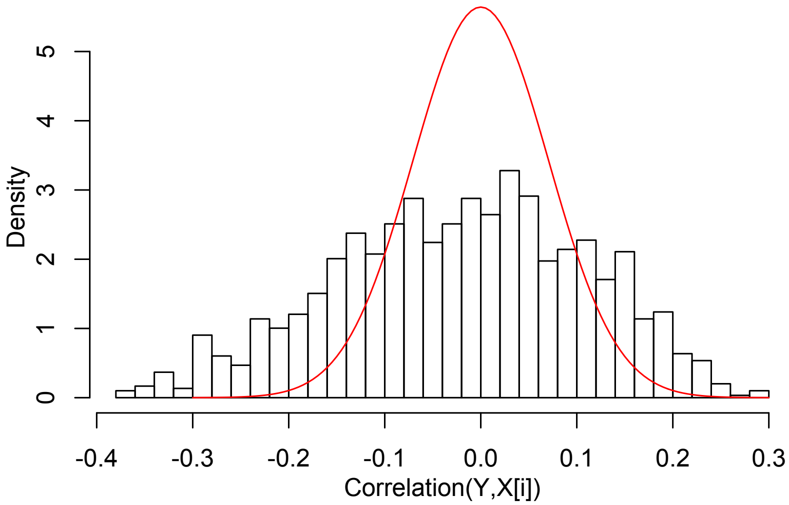
Wharton
Department of Statistics

UCI = Univ of Ca Irving ML databases, http://archive.ics.uci.edu

# Marginal Analysis

- Marginal correlations ($X_i$,Y) show signal
  - Deviate from distribution of random noise (red)
- But: weakly spread over many coordinates
  - Multiple regression finds weak effects
  - $R^2 = 0.19$ is larger than might expect

Expect p/n
$R^2 = 10/200 = 0.05$



```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.059631   1.301089   0.814  0.41643
x1          -0.004818   0.005775  -0.834  0.40512
x2          -0.007273   0.003887  -1.871  0.06288 .
x3           0.004149   0.003295   1.259  0.20954
x4          -0.003342   0.001279  -2.614  0.00967 **
x5          -0.007191   0.006658  -1.080  0.28153
x6           0.002474   0.002162   1.144  0.25401
x7          -0.001173   0.001457  -0.805  0.42188
x8           0.001113   0.009964   0.112  0.91116
x9          -0.008695   0.004328  -2.009  0.04599 *
x10          0.000841   0.002593   0.324  0.74604
```

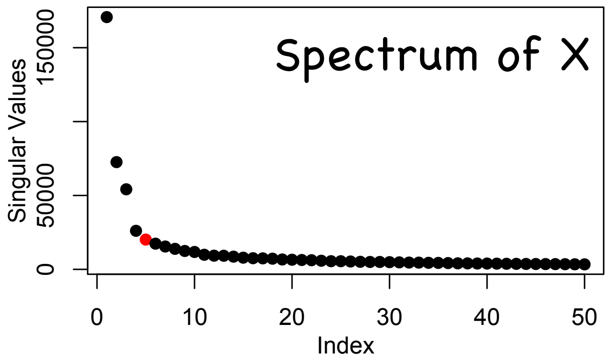
$R^2 = 0.19$

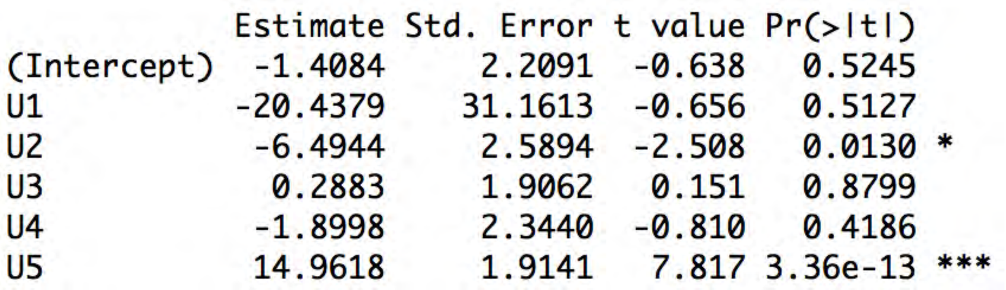
# PCA Analysis

- Compute singular value decomposition

$$X = U\,D\,V'$$

  - Columns of U, V are orthonormal
  - D is a diagonal matrix of singular values (spectrum of X)
- Doable in R if X is 200×10,000 matrix
  - Regression finds clear, strong effect in $U_5$



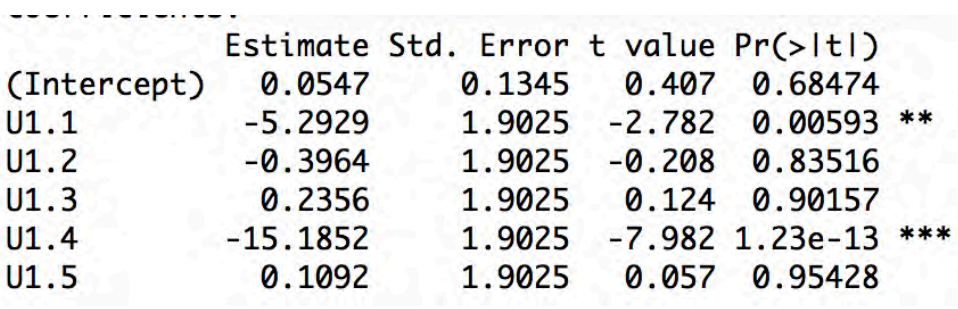


```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4084     2.2091  -0.638   0.5245
U1          -20.4379    31.1613  -0.656   0.5127
U2           -6.4944     2.5894  -2.508   0.0130 *
U3            0.2883     1.9062   0.151   0.8799
U4           -1.8998     2.3440  -0.810   0.4186
U5           14.9618     1.9141   7.817 3.36e-13 ***
```

$R^2$=0.27

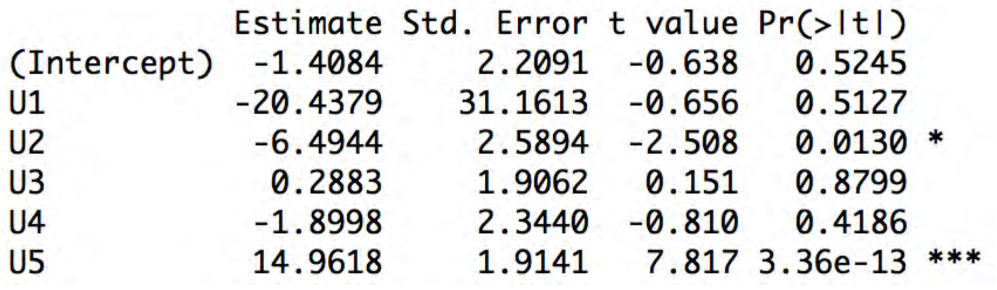
# Random Projection

- Project down to smaller size
  - Example with d=100
  - Compare random projections to exact from R
- Procedure
  - $P_0 = X\ \Omega$, $\Omega$ is 10,000×d random matrix
  - $P_1 = XX'\ P_0$ is one step of power method
  - Take first few columns of U from SVD of $P_j$
- Compare to fit with exact SVD



```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0547     0.1345   0.407  0.68474
U1.1          -5.2929     1.9025  -2.782  0.00593 **
U1.2          -0.3964     1.9025  -0.208  0.83516
U1.3           0.2356     1.9025   0.124  0.90157
U1.4         -15.1852     1.9025  -7.982 1.23e-13 ***
U1.5           0.1092     1.9025   0.057  0.95428
```

Random Projection
one iteration

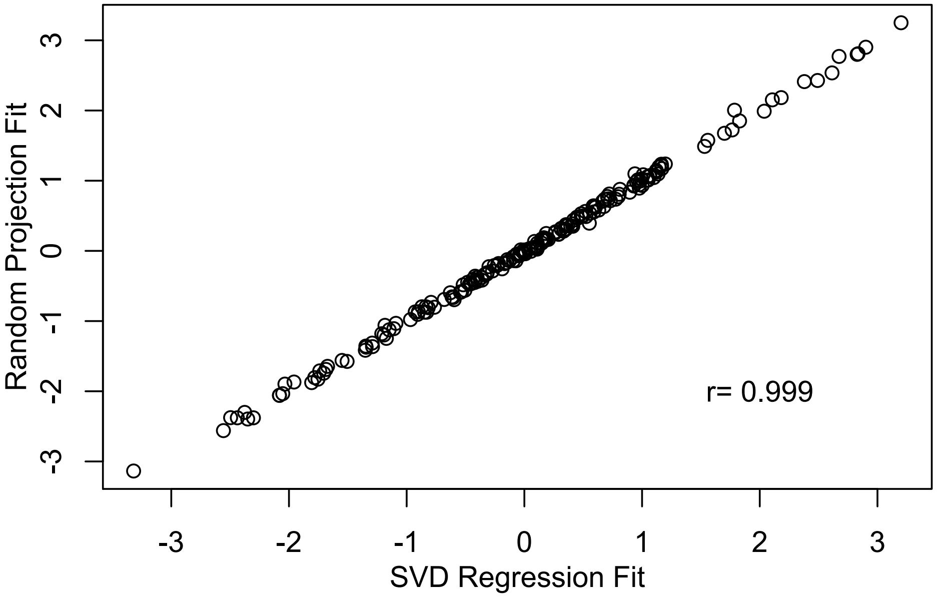


```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4084     2.2091  -0.638   0.5245
U1          -20.4379    31.1613  -0.656   0.5127
U2           -6.4944     2.5894  -2.508   0.0130 *
U3            0.2883     1.9062   0.151   0.8799
U4           -1.8998     2.3440  -0.810   0.4186
U5           14.9618     1.9141   7.817 3.36e-13 ***
```

Exact

# Comparison of Fits

- Reconstruction
  - Random projection preserves subspace holding range of matrix, but not necessarily in the same coordinates.
  - Eg: different components appear in regression
- Comparison of fits shows same subspace



After one iteration

Wharton
Department of Statistics

# A really big X matrix?

- Arcene example is 'small': we can do do the exact SVD quickly in R.
- Suppose X had more columns, say

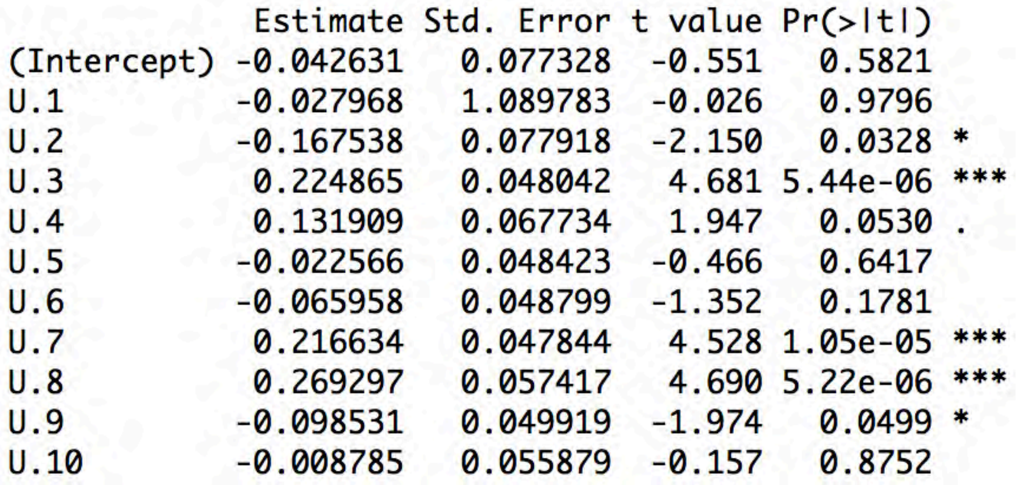
  $$10{,}000^2 = 100{,}000{,}000$$

  such as from the interaction space of X.

Okay, half that

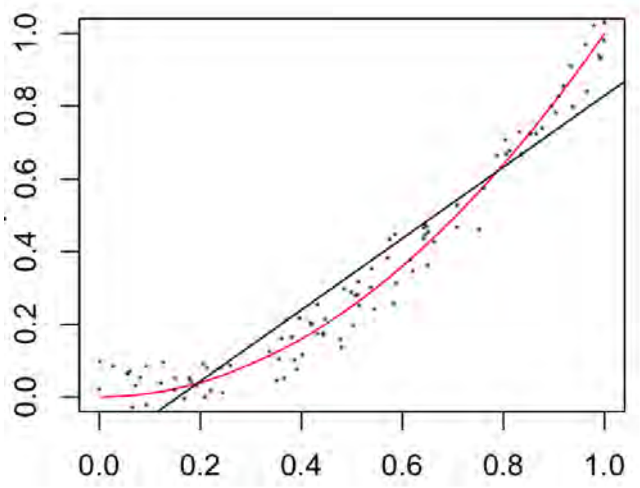
- Linear models often approximate non-linear structure…

first 10 PCs of X



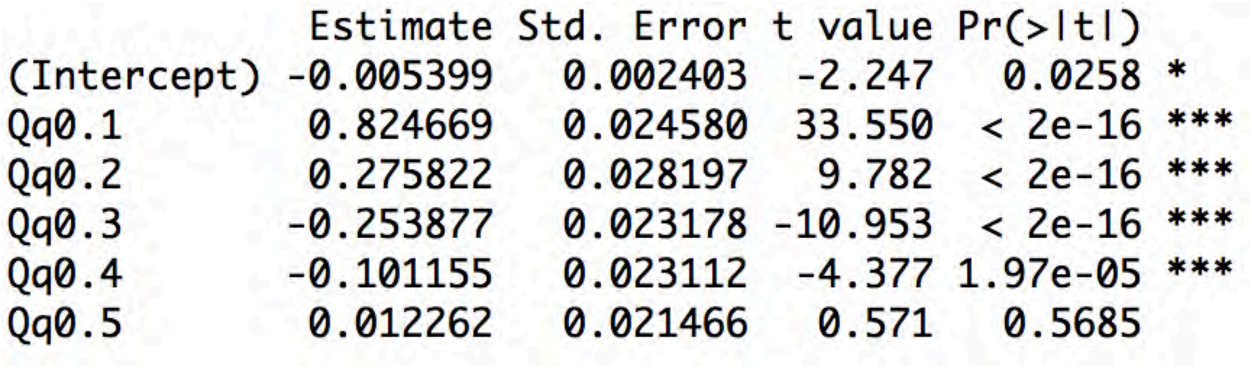
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.042631   0.077328  -0.551   0.5821
U.1         -0.027968   1.089783  -0.026   0.9796
U.2         -0.167538   0.077918  -2.150   0.0328 *
U.3          0.224865   0.048042   4.681 5.44e-06 ***
U.4          0.131909   0.067734   1.947   0.0530 .
U.5         -0.022566   0.048423  -0.466   0.6417
U.6         -0.065958   0.048799  -1.352   0.1781
U.7          0.216634   0.047844   4.528 1.05e-05 ***
U.8          0.269297   0.057417   4.690 5.22e-06 ***
U.9         -0.098531   0.049919  -1.974   0.0499 *
U.10        -0.008785   0.055879  -0.157   0.8752
```

$R^2$=0.35



# Random Projection

- Do the random projection with 50,000,000 explanatory variables ($X_j$ $X_k$)
  - Cannot compare to the exact solution for this one
  - Runs ‘quickly’ unless you want to try the power method to get a better solution!
- Fitted model on 5 elements of projection of the quadratic X’s



```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.005399   0.002403  -2.247   0.0258 *  
Qq0.1        0.824669   0.024580  33.550  < 2e-16 ***
Qq0.2        0.275822   0.028197   9.782  < 2e-16 ***
Qq0.3       -0.253877   0.023178 -10.953  < 2e-16 ***
Qq0.4       -0.101155   0.023112  -4.377 1.97e-05 ***
Qq0.5        0.012262   0.021466   0.571   0.5685    
```

$R^2=0.35 \rightarrow R^2=0.87$

# Postscript...

- What's the response in that regression?
  - What Y variable lives in the quadratic space?
- Short answer: Kernel trick
  - Compute the quadratic kernel of the data
  - Find the SVD
  - Let Y be one of the singular vectors
- Story for another day ...

# Reducing Rows

- Context
  - Very large n >> moderate p
  - Again, less interested in selecting specific Xs
- Common sense
  - Don't need to fit a model more precisely than needed for statistical precision/selection.
  - However…

    More data reveals a more interesting model, one with subtle effects
- Speed of OLS
  - $b = (X'X)^{-1}X'Y$
  - Slow part if n >> p is computing X'X $O(np^2)$

# Case Sampling

Not sampling on the response!

- Exploit familiar property of regression
  - Precision of slope is maximized by finding cases with large variation in Xs
  - Task becomes finding cases with high leverage
- Machine learning has developed methods to seek high-leverage points
  - Hard to find sequentially
- Simple improvement
  - Sample m << n cases to estimate X'X
  - Use all n cases to estimate X'Y
- Leverage points however may not be your friends in modeling large data sets...

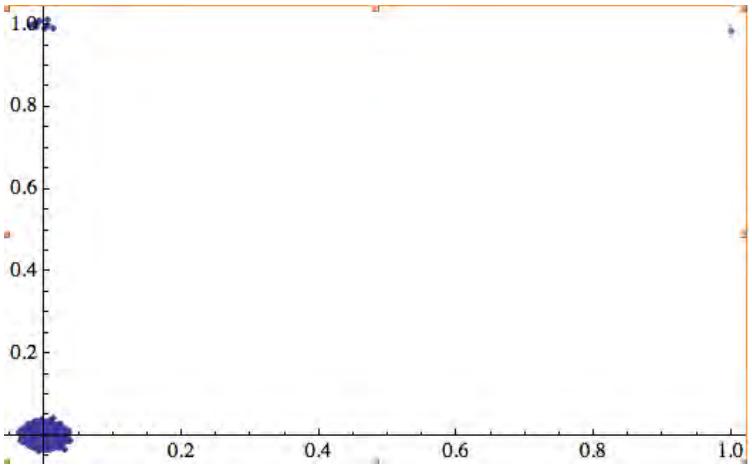
$b=(X'X)^{-1}X'Y$

Wharton
Department of Statistics

Misconceptions

# Outliers in Big Data

- Sparse data
  - n=10,000
  - X ≈ 0
    - Y=0 for 9,999, Y=1 for 10
  - X ≈ 1
    - Y=1



- What's the appropriate p-value?
- Classical OLS
  - Use residual after fit slope, as if right model
  - t ≈ 10, pick your level of significance!
- Common sense
  - p = 1/1000 more sensible p-value

Wharton
Department of Statistics

# Streaming Methods

Cases

Variables

Combined

# Streaming Cases

- Context
  - Huge number of cases, more than memory holds
- Idea
  - Compute estimates as read in data so do not have to store all data
  - Calculations can be split over network
- Different take on OLS
  - OLS estimate for n-1 cases

$$b_{n-1} = (X'X)^{-1}X'Y$$

  - The estimate for n cases is

$$\begin{aligned} b_n &= b_{n-1} + (X'X)^{-1}x_n(y_n - x_n' b_{n-1})/(1+h_n) \\ &= b_{n-1} + \boxed{[(1+h_n)(X'X)]^{-1}}\, x_n\, e \end{aligned}$$

where the leverage $h_n = x_n'(X'X)^{-1}x_n$. slow step

# Stochastic Gradient

- Build up normal equations and solutions by randomly sampling cases
- Stochastic gradient
  - Robbins & Monro
  - To minimize $(y_i - x_i'b)^2$ w.r.t. b, step in the direction of the negative gradient,

$$x_i(y_i - x_i'b) = x_i\ e_i$$

- Full least squares solution uses X'X

$$b_n = b_{n-1} + [(1+h_n)(X'X)]^{-1}\ x_n\ e$$

- Pretend X'X is diagonal, and life moves by faster

$$b^*_n \;= b^*_{n-1} + \delta_n\ D^{-1}\ x_n\ e^*$$

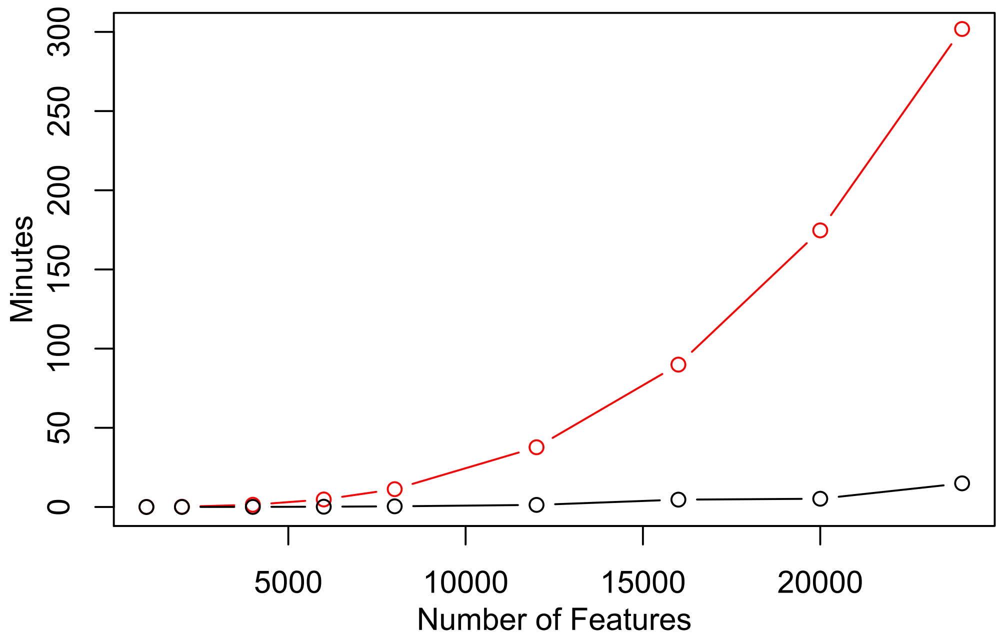
with D = diagonal (X'X) and $\delta_n$ is a learning rate.

# How fast is it?

- Goal in stochastic gradient is to run as fast as you can read data!

| n | p | OLS | SG |
|---|---|---|---|
| 2,500 | 500 | <0.1 | <0.1 |
| 5,000 | 1,000 | 0.7 | 0.2 |
| 10,000 | 2,000 | 9.5 | 0.9 |
| 20,000 | 4,000 | 84 | 3.7 |
| 40,000 | 8,000 | 675 | 25 |
| 80,000 | 16,000 | 5394 | 276 |
| 100,000 | 20,000 | 10480 | 312 |

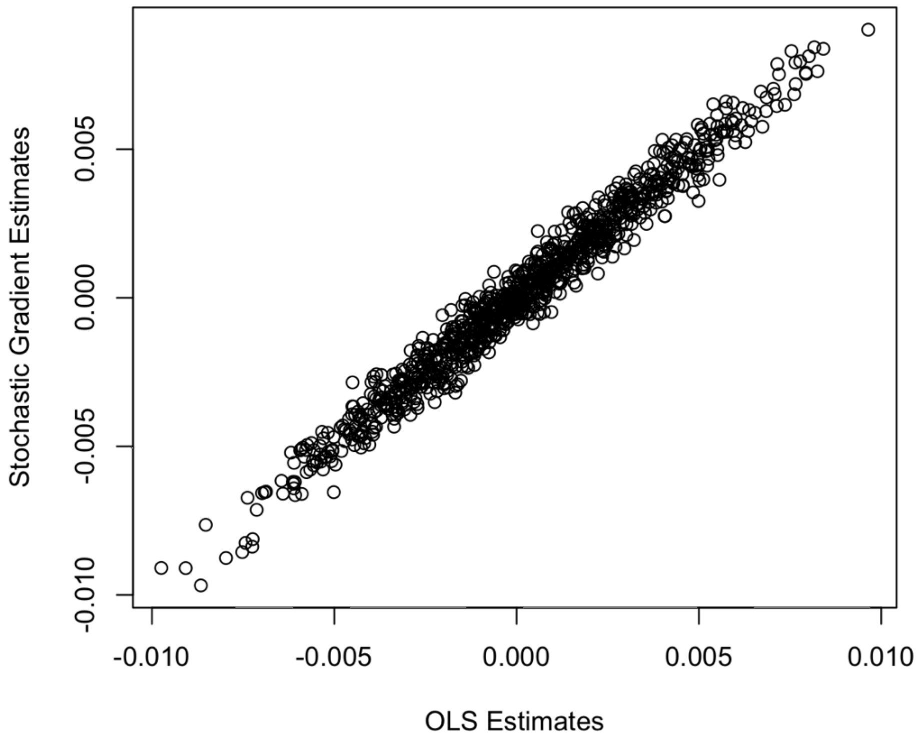
n=5p seconds



# How good are estimates?

- Graph plots estimated coefficients from one-pass of stochastic gradient versus exact OLS
- Deviation from OLS below standard error
  - Small error relative to variation in estimates

p=1000



At least when there is not much collinearity!

Wharton
Department of Statistics

# Statistical Significance?

- Don’t have X’X so don’t have usual SE
  - How to evaluate modeling?
- Cross-validation
  - Less sensitive to modeling assumptions
  - Split data

    Training data: Fit model on part of the data

    Test data: Reserved data
  - Compare fit in two datasets
- Three way split becoming necessary
  - Training data
  - Tuning data...
    - Set tuning parameters, such as level of shrinkage
  - Testing data

Misconceptions

# Population Drift

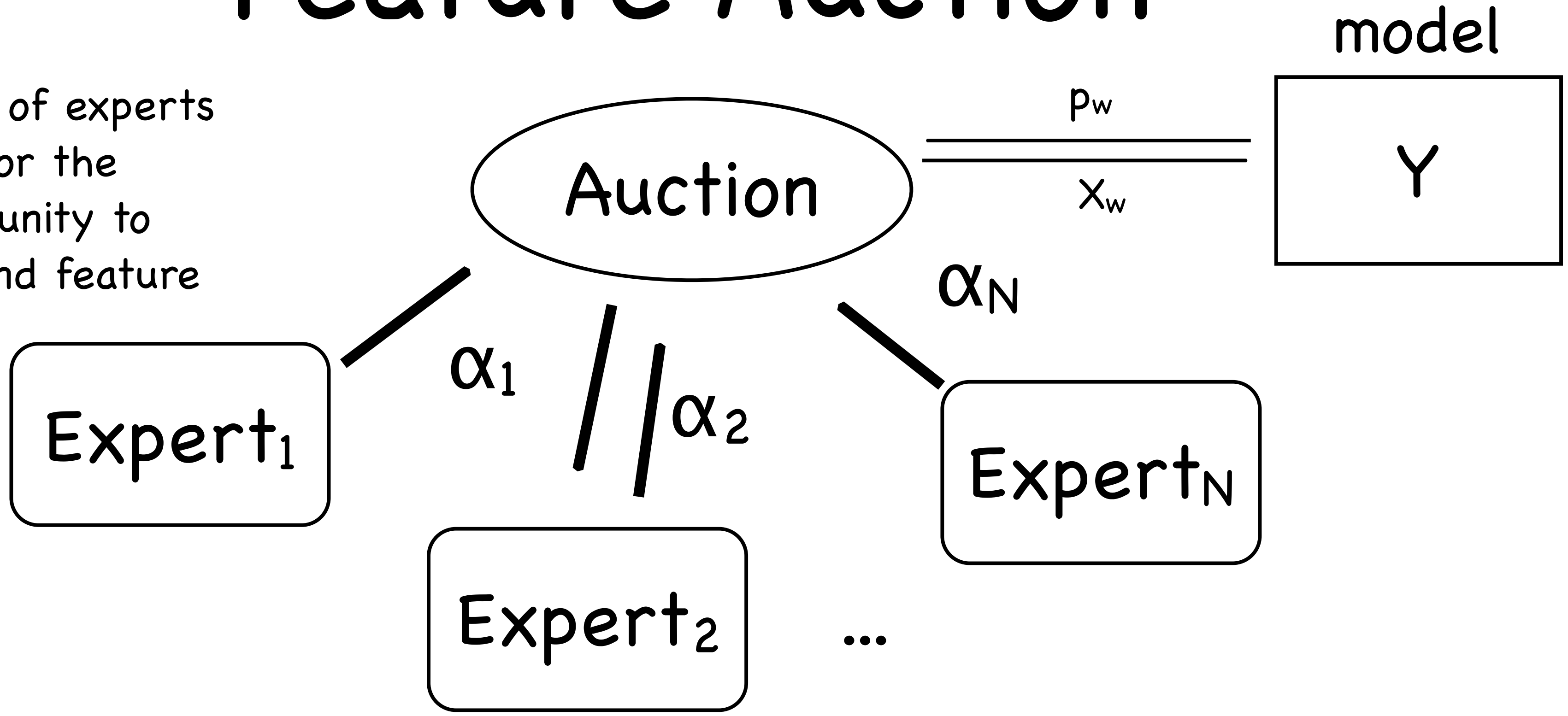
- Cross-validation is an optimistic assessment
  - One of few places when have random sample
- Credit scoring
  - Predict performance of applicants
  - Cross-validation shows model spot on
- Data collection is a long process
  - Gather data over 1-2 years
  - Takes 1-2 more years to find the response
- The world changed!
  - Booming economy during data collection
  - Collapsing recession when implemented
  - No way CV could see this problem

More issues …
Variation?
How to allocate?

# Streaming Variables

- Context
  - Huge number of variables
  - Want to preserve scales
- Idea
  - Stepwise search pays a large cost for searching
    Bonferroni p-value threshold 0.05/millions
  - Streaming: Examine features one at a time
    - Resembles forward stepwise, but
      without sorting/ordering based on p-values
- Exploit context
  - "Scientist" orders variables, defines search strategy
  - Adaptive:Build interactions as features added

# Feature Auction



Auction collects winning bid $\alpha_2$

Expert supplies values of recommended feature $X_w$

Expert receives payoff $\omega$ if $p_w \le \alpha_2$

Experts only learn if the bid was accepted, not the value of b or the p-value.

Wharton
Department of Statistics

# Experts

- Expert
  Strategy for creating list of features. Experts embody domain knowledge, science of application.
- Source experts
  - A collection of measurements (eg, synonyms, clusters)
  - Components of a subspace basis (PCA, RKHS)
  - Lags of a time series
- Scavenger experts
  - Interactions
    - among features accepted into model
    - among features rejected by model
    - between those accepted with those rejected
  - Transformations
    - segmenting, as in scatterplot smoothing
    - polynomial transformations

# Winning Experts

- Expert is rewarded if correct
  - Experts have alpha-wealth
  - If recommended feature is accepted in the model, expert earns ω additional wealth
  - If recommended feature is refused, expert loses bid
- As auction proceeds, it...
  - Rewards experts that offer useful features.
  - Eliminates experts whose features are not accepted.
- Taxes fund scavenger experts
  - Ensure that continue to control overall FDR
- Critical
  - Adjust for multiplicity
  - p-values determine useful features

# Robust Standard Errors

- p-values are critical, but…
  - Error structure often heteroscedastic
  - Observations frequently dependent
- Dependence
  - "Observations"
    - Spatial time series at multiple locations
    - Documents from various news feeds
  - Transfer learning problem
- Examples
  - Use sandwich-type estimate of standard error

heteroscedasticity

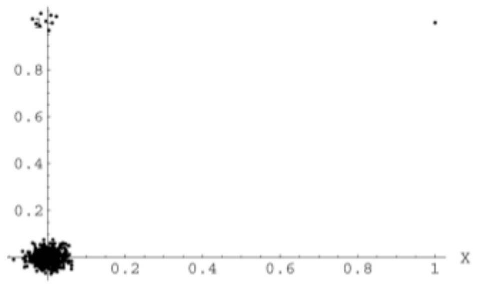
$$\mathrm{var}(b) = (X'X)^{-1}X'E(ee')X(X'X)^{-1}$$
$$= (X'X)^{-1}\, X'D^2X\, (X'X)^{-1}$$

dependence

$$\mathrm{var}(b) = (X'X)^{-1}X'E(ee')X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}\, X'BX\, (X'X)^{-1}$$

# Flashback...

- Heteroscedastic error
  - Estimate standard error with outlier
  - Sandwich estimator allowing heteroscedastic error variances gives a t-stat ≈ 1, not 10.



- Dependent error
  - Even more important need for accurate SE
  - Netflix example
    Bonferroni (or hard thresholding) overfits due to dependence in responses.
  - Spatial modeling
    Everything seems significant unless incorporate dependence into the calculation of the SE

Wharton
Department of Statistics

# Control for Over Fitting

- Alpha investing
  - Test possibly infinite sequence of m hypotheses

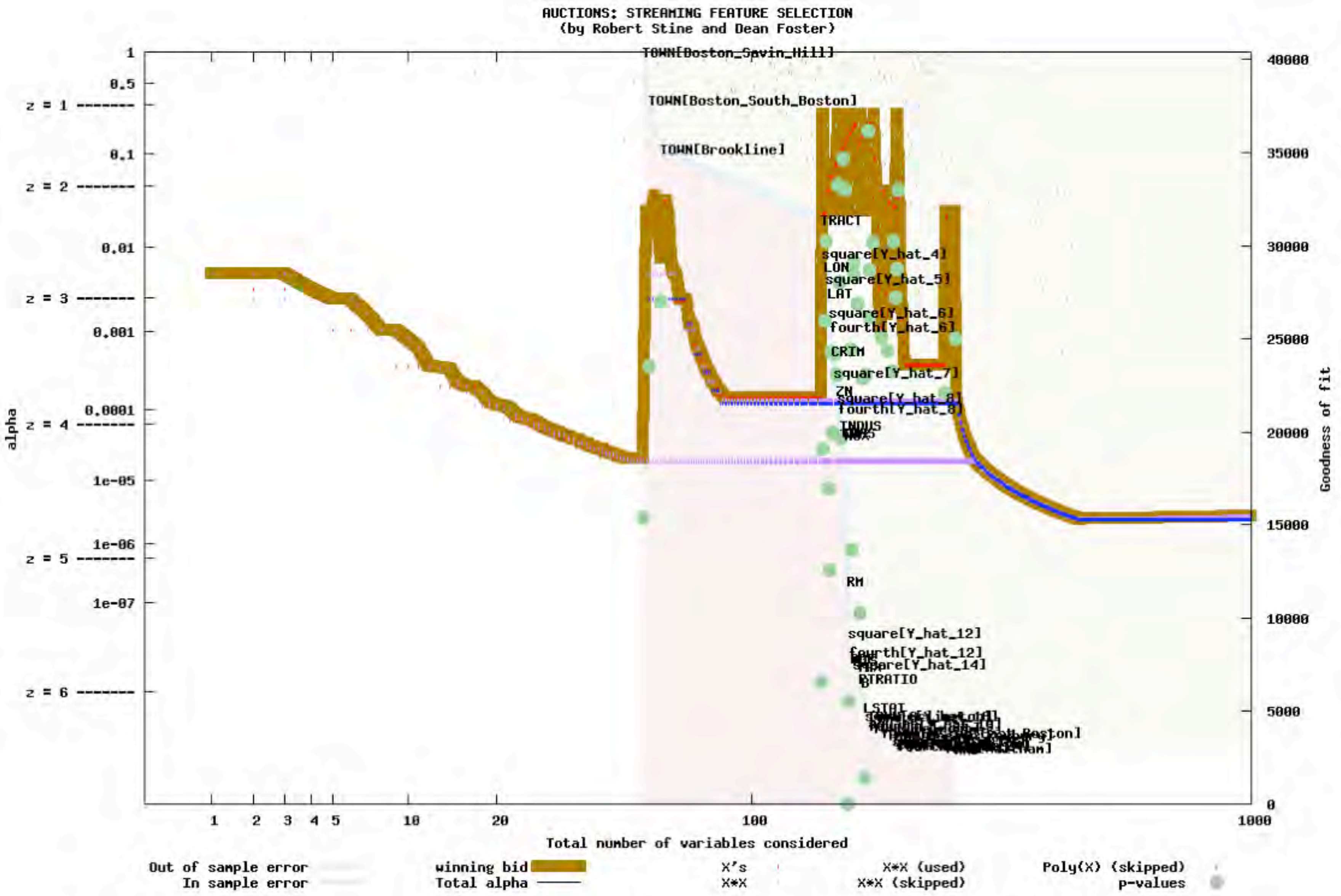
    $H_1, H_2, H_3, \ldots H_m \ldots$

    obtaining the p-values $p_1, p_2, \ldots$
- Procedure
  - Start with an initial alpha wealth $W_0$
  - Invest wealth $0 \le \alpha_j \le W_j$ in the test of Hj
  - Change in wealth depends on test outcome

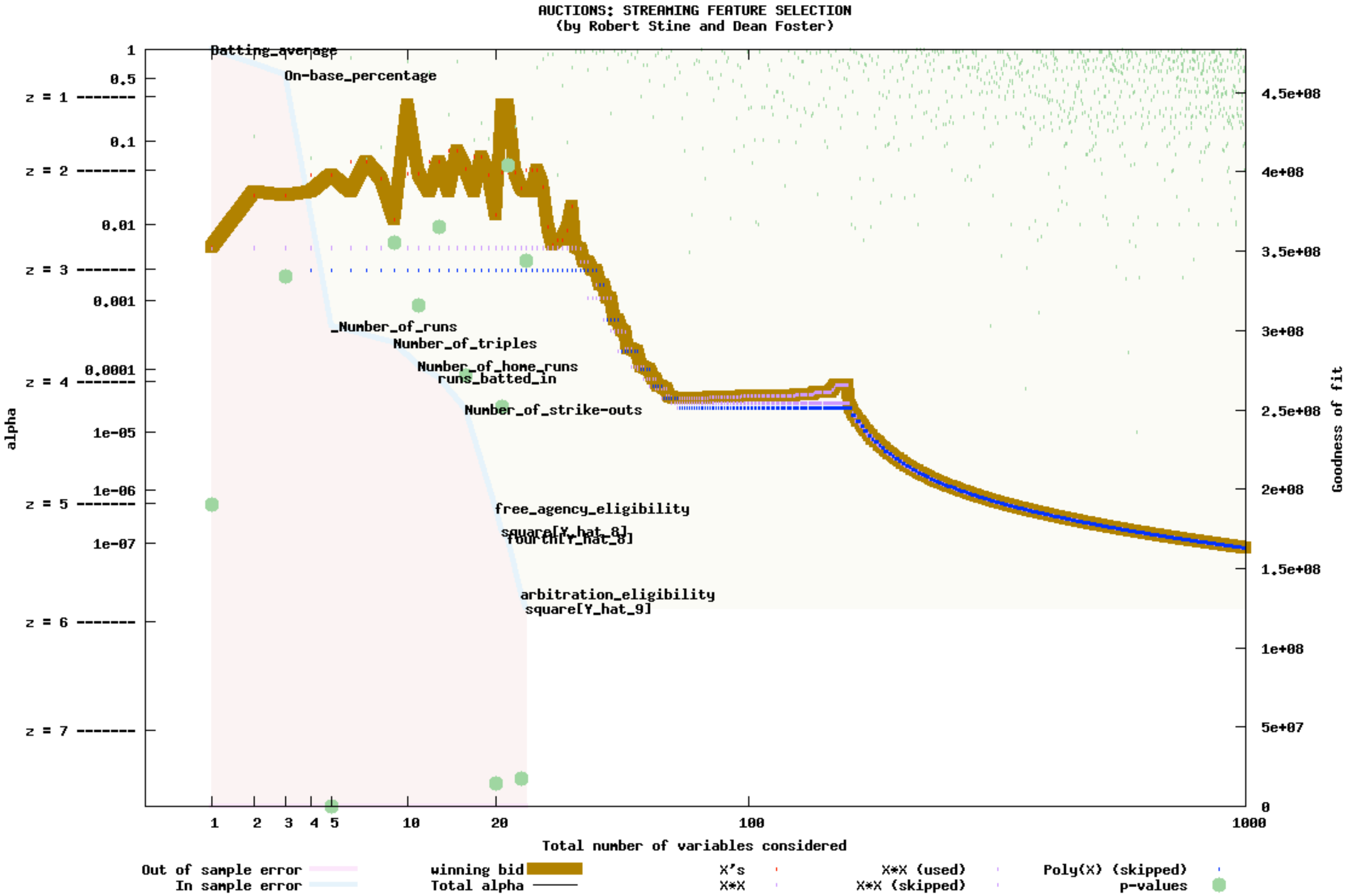
    If reject, wealth goes up by payout $\omega-\alpha_j$

    If don't reject, wealth goes down by $\alpha_j$
- Properties
  - Controls expected false discovery rate
  - Can reproduce Bonferroni or FDR methods

# Boston Housing



# Baseball


AUCTIONS: STREAMING FEATURE SELECTION
(by Robert Stine and Dean Foster)
1
0.5
z = 1 --------
0.1
z = 2 --------
0.01
z = 3 --------
0.001
0.0001
z = 4 --------
1e-05
1e-06
z = 5 --------
1e-07
z = 6 --------
z = 7 --------
alpha
Batting_average
On-base_percentage
_Number_of_runs
Number_of_triples
Number_of_home_runs
runs_batted_in
Number_of_strike-outs
free_agency_eligibility
square[Y_hat_8]
fourth[Y_hat_8]
arbitration_eligibility
square[Y_hat_9]
4.5e+08
4e+08
3.5e+08
3e+08
2.5e+08
2e+08
1.5e+08
1e+08
5e+07
0
Goodness of fit
1
2
3
4
5
10
20
100
1000
Total number of variables considered
Out of sample error
In sample error
winning bid
Total alpha
X's
X*X
X*X (used)
X*X (skipped)
Poly(X) (skipped)
p-values

# Streaming Cases & Variables

- Background
  - A variance inflation factor (VIF) is a diagnostic for collinearity in regression
- VIF compares variances of slope estimates
  - Variance of $b_k$ were it uncorrelated with others

    $$var(b_x) = s^2/(x_k'x_k)$$

  - Actual variance is larger due to collinearity

    $$var(b_k) \approx VIF_k\ s^2/(x_k'x_k)$$

    where $1 \leq VIF_k = 1/(1-R^2_{k|rest})$
- Handy interpretation
  - Is $x_k$ not significant because

    It is not useful?

    Redundant?

Wharton
Department of Statistics

# VIF Regression

- Idea
  - Speed up the slow step in forward stepwise
- Usual selection
  - Has variables X and residual

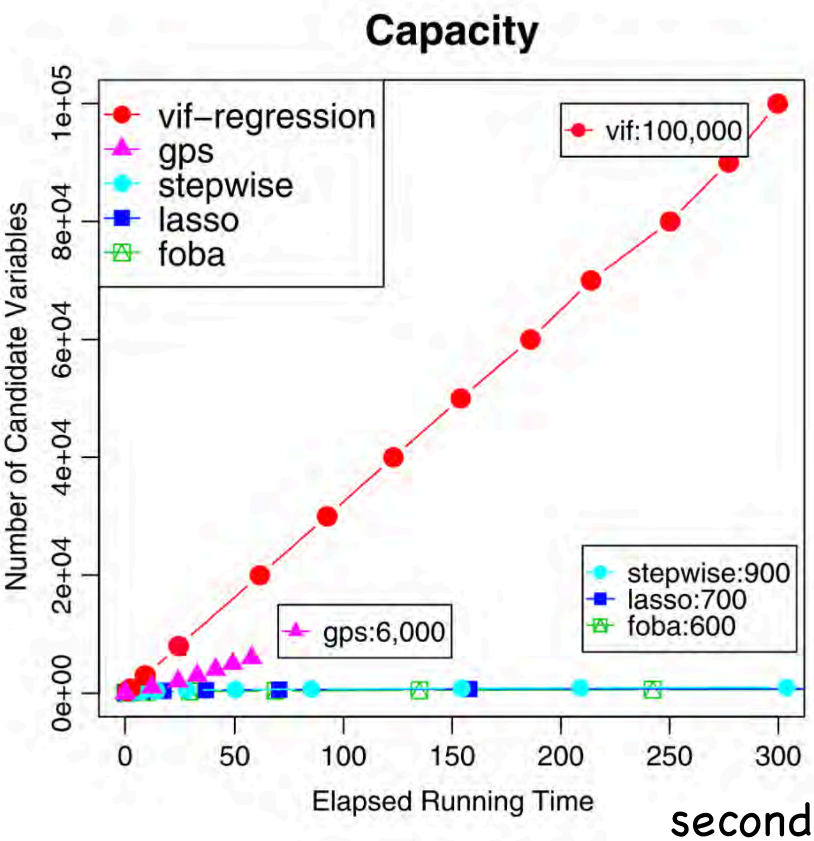
    $e = (I - X(X'X)^{-1}X')\ y = (I - H)\ y$
  - Partial t-statistic for testing another variable z with partial regression $\boxed{z^*=(I-H)z}$ $O(np^2)$ given $(X'X)^{-1}$

    $t^2 = (z^{*\prime}e)^2/(s^2\, z^{*\prime}z^*)$
- Re-express t-statistic using VIF

  $t^2 = (z'e)^2/(s^2\, z'z\ \mathrm{VIF}_k)$
- Conservatively estimate $\mathrm{VIF}_k$ from sample

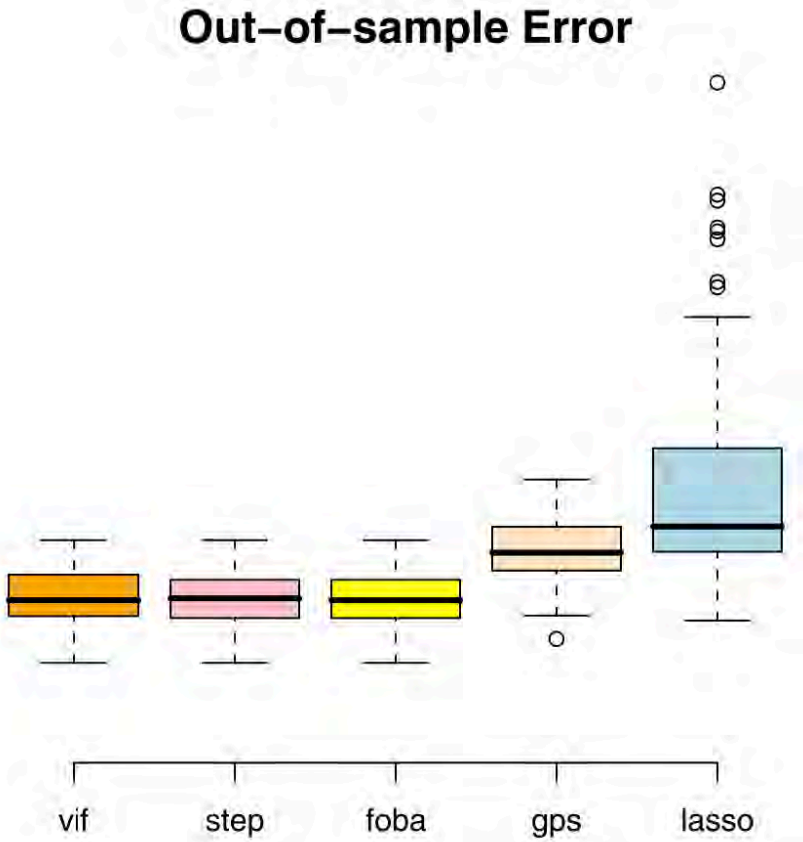
# Performance

- Faster than rivals
  - Plus smaller out-of-sample error



seconds

n=1000



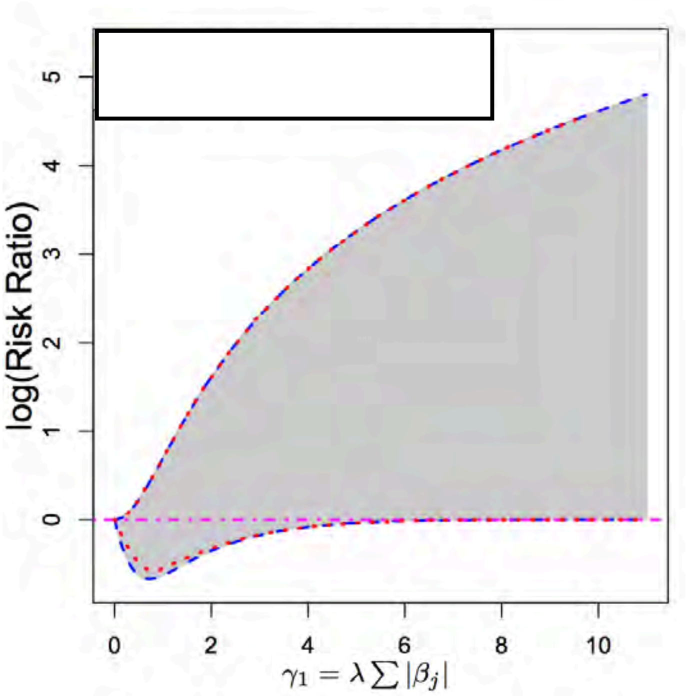
n=1000, p=500

Wharton
Department of Statistics

Misconceptions

# Comment on $L_1$

- Success of lasso depends on nature of underlying model
- Risk comparison
  - Compare the risk of the model identified by subset selection to the model identified by lasso ($L_1$).
  - Grey region in plot represent possible model datasets
- Take-away
  - In models for which lasso identifies high penalty, $L_0$ has better performance.



Wharton
Department of Statistics

# Wrap-Up

- Dimension reduction
  - Random projection
  - Subsampling
- Streaming
  - VIF regression
  - Alpha investing, auction models
- Issues
  - Importance of substantive insight
  - Prediction/association vs causation
  - Dependence, population drift

# References

- Stochastic Gradient
  - Papers of John Langford, Microsoft Research
- Random projection
  - Halko, Martinsson, and Tropp, SIAM Review, 2011
- VIF Regression
  - "VIF Regression: A Fast Regression Algorithm for Large Data", JASA, 2011, Lin, Foster and Ungar
- Alpha investing
  - "α-investing: a procedure for sequential control of expected false discoveries", JRSSB, 2006
- Improved stepwise regression
  - "Variable selection in data mining: Building a predictive model for bankruptcy", JASA, 2004
- Streaming feature selection
  - "Streamwise feature selection", JMLR, 2006, with Foster, Ungar, and Zhou.

Thanks!