Wharton
Department of Statistics

Auctioning Predictors: Combining Domain Knowledge with Automated Search Strategies

Bob Stine

Department of Statistics

The Wharton School, Univ of Pennsylvania

September, 2003

www-stat.wharton.upenn.edu/~stine

Examples

Wharton
Department of Statistics

- Predicting the onset of bankruptcy (Credit VII)
  - 3,000,000 record database
  - ≥ 100,000 potential predictors of many types
- Estimating risk of default on underwritten loans
  - 300,000 outstanding loans
  - Many types of predictors
    - Industry characteristics, macroeconomic factors
    - Historical record, properties of a specific loan
- Predicting efficacy of new medication
  - Lab tests, doctor opinion, patient self-reported data

Predictive modeling

Wharton
Department of Statistics

- Predict characteristic
  - Personal bankruptcy, business loan default, reaction to drug
- Predict using features selected from LARGE database of possible features
  - Many types of features
  - Some substantively motivated, others just “available”
  - Some expensive to gather, some much cheaper
  - Some you have now, some you collect later
- Hard part
  - Once you have the data, which predictors to use?

Which features to use?

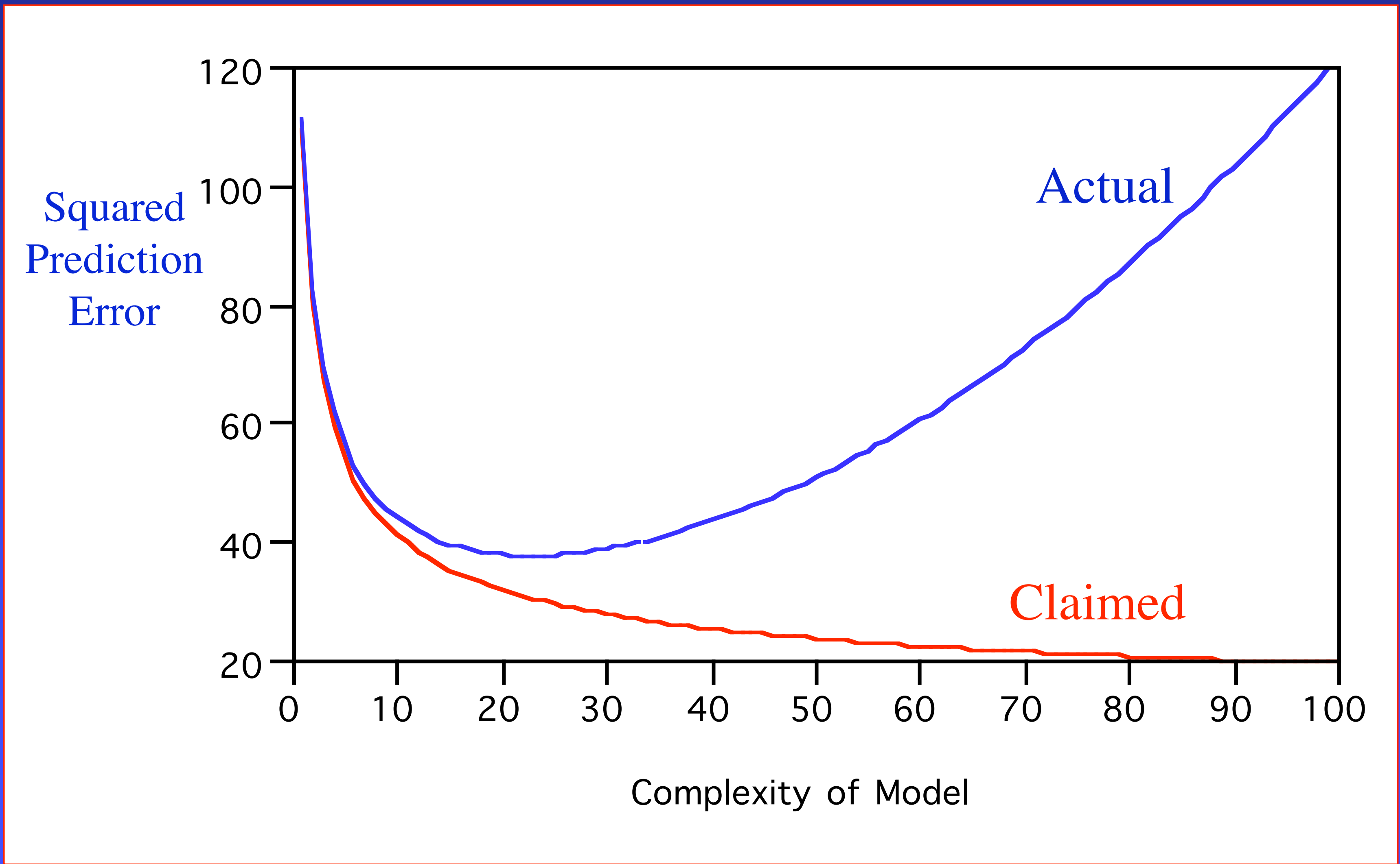
Wharton
Department of Statistics

- Substantive
  - Pick predictors “by hand”
- Advantages
  - Leverage expertise, domain knowledge
  - Easy to “explain” to customer or regulator
- Disadvantages
  - Time consuming to construct
  - Did we miss something?
  - Time consuming to *maintain*
  - Has the world changed?

- Automatic
  - Algorithmic feature selection
- Advantages
  - Scans large database quickly
  - Automatic rebuilding
  - Exploits automated data streams
- Disadvantages
  - Does not exploit domain expertise.
  - Often hard to explain or interpret.
  - Overfitting

Overfitting

Wharton
Department of Statistics



Best of both worlds?

Wharton
Department of Statistics

- Substantive
  - Pick predictors "by hand"
- Advantages
  - **Leverage expertise, domain knowledge**
  - **Easy to "explain" to customer or regulator**
- Disadvantages
  - Time consuming to construct
  - Time consuming to *maintain*
  - Did we miss something?
  - Has the world changed?

- Automatic
  - **Algorithmic feature selection**
- Advantages
  - **Scans large database quickly**
  - **Automatic rebuilding**
  - **Exploits automated data streams**
- Disadvantages
  - Does not exploit domain expertise.
  - May still be hard to explain or interpret?
  - Overfitting

Auction process

Wharton
Department of Statistics

- *Predictor streams* offer predictors to consider.
- *Bidders* rate possible choices (assign a probability).
- *Auctioneer* selects predictor with highest rating.
- *Bidders* place bids on this predictor.
- *Auctioneer* tests whether predictor adds value.
  - Statistical test to see if predictor improves underlying model
- Winning *bidders* collect if predictor chosen for model.

Predictor streams

Wharton
Department of Statistics

- Differentiate domains
  - Different streams for different domains, e.g.
    - One stream for macro features
    - A second stream for individual features
- Experts order predictors
  - Expert determines the order in which stream offers its predictors to the auction
- Bidders
  - *Learn* which streams offer predictors that join model

Auction process

Wharton
Department of Statistics

- *Predictor streams* offer predictors to consider.
- ***Bidders* rate possible choices (assign a probability).**

Picking predictor for auction

Wharton
Department of Statistics

- Multiple predictor streams, but only one is tested in each round of the auction.
- Bidders assign probabilities to streams
  - Internal features, "preferences" of bidder
  - Experience with this stream of predictors
- Auctioneer sums probabilities and picks the predictor that attracts the most bidder interest.

Auction process

Wharton
Department of Statistics

- *Predictor streams* offer predictors to consider.
- *Bidders* rate possible choices (assign a probability).
- *Auctioneer* selects predictor with highest rating.
- ***Bidders* place bids on this predictor.**

Bidding

- Each bidder has “wealth”
  - Wealth = cumulative *rate* of picking predictors that fail, i.e. each bidder has its own □ rate.
  - Wealth initially allocated to bidders by the auctioneer.
- Bidders bid on the offered predictor
  - Share of current wealth
    - Probability that this predictor will join model
    - Bayesian schemes, exploiting risk aversion
  - Successful bidders have more to bid
- Auctioneer collects bid from each bidder

Auction process

Wharton
Department of Statistics

- *Predictor streams* offer predictors to consider.
- *Bidders* rate possible choices (assign a probability).
- *Auctioneer* selects predictor with highest rating.
- *Bidders* place bids on this predictor.
- ***Auctioneer* tests whether predictor adds value.**
  - Statistical test to see if predictor improves underlying model

Evaluating a predictor

- Thresholding
  - Compare p-value of predictor to threshold
  - Bankruptcy analysis discusses optimal thresholding
  - Variety of schemes for setting threshold (AIC, BIC,…)
- For the auction, cumulative bid sets threshold
  - Bidders spend □ for error as they bid on predictors.
- Related ideas in multiple testing
  - □-spending rules in sequential clinical trials
  - Family-wide error rates
  - Step-up/step-down testing

Auction process

Wharton
Department of Statistics

- *Predictor streams* offer predictors to consider.
- *Bidders* rate possible choices (assign a probability).
- *Auctioneer* selects predictor with highest rating.
- *Bidders* place bids on this predictor.
- *Auctioneer* sees whether predictor adds value.
  - Statistical test to see if predictor improves underlying model
- ***Bidders* collect if predictor chosen for model.**

Paying off the bidders

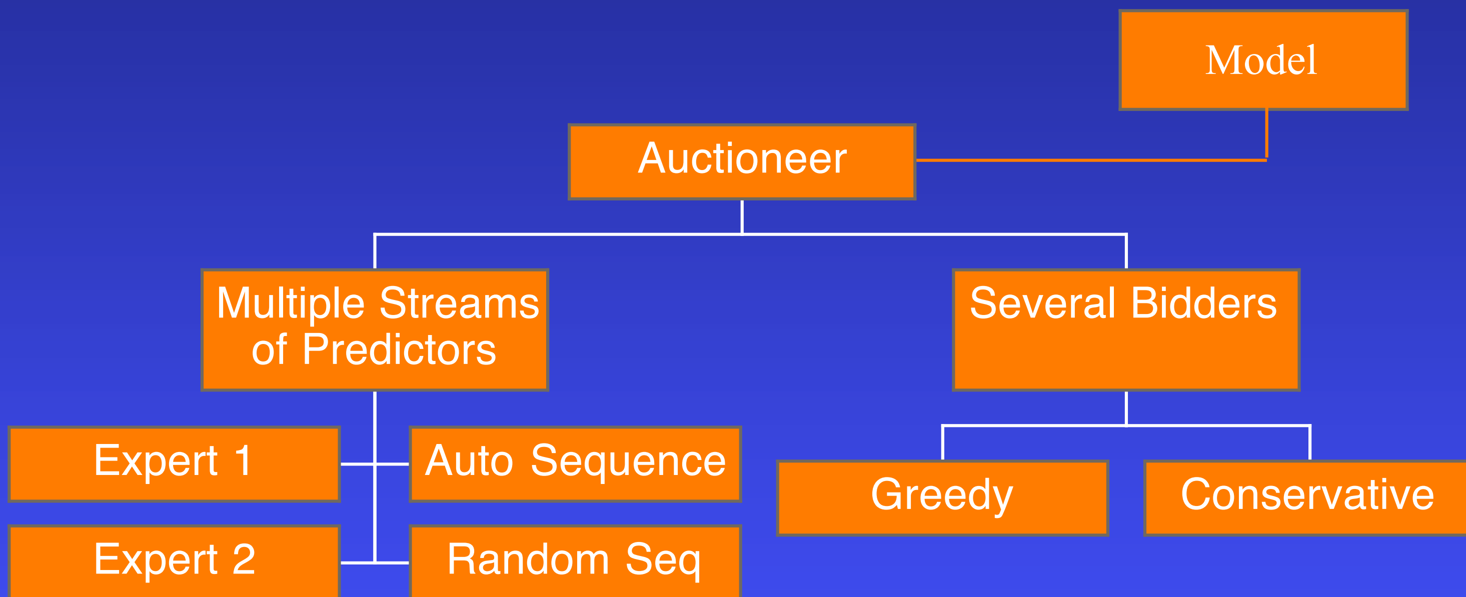
- Auction begins with an allowed probability for error
  - Total □ for the auction controls the *rate* of false positives.
  - Tuning parameter, typically set total □ to 0.05.
- Each predictor added to model earns the auction more chance for error in considering rest of predictors
  - Finding good predictor increases the total □ for the auction by 0.05.
  - Auctioneer distributes this added "wealth" to the bidders proportional to bid.
  - Bidders who bid more on good predictors accumulate "wealth" and have more to bet in future rounds.

Auction process

- *Predictor streams* offer predictors to consider.
- *Bidders* rate possible choices (assign a probability).
- *Auctioneer* selects predictor with highest rating.
- *Bidders* place bids on this predictor.
- *Auctioneer* sees whether predictor adds value.
  - Statistical test to see if predictor improves underlying model
- *Bidders* collect if predictor chosen for model.
- **Auction continues to the next round.**

Predictor auction schematic

Wharton
Department of Statistics



Sequential feature selection

Wharton
Department of Statistics

- Auction considers predictors offered by streams *sequentially*, rather than "all at once" (batch).
- Can you really find features one-at-a-time?
- Good substantive knowledge, domain expertise
  - Predictor stream offers best conjectures first
  - Order of predictors is key to optimal predictor selection
- Weak (or no) substantive knowledge
  - Traditional automatic feature selection is batch
    - E.g., stepwise regression
  - Sequential selection works just as well and can be faster!

Sequential vs. Batch Selection

Wharton
Department of Statistics

Sequential	Batch
Try the predictors in the order offered.	Search through "all possible" predictors to find the best predictor out there.
Allows an infinite stream of possible predictors.	Must have all of the predictors there for search.
Can direct search in reaction to successful domains.	Need to identify all possible predictors from the start.
Calculations are just a sequence of simple fits.	Array manipulations can be onerous in large problems.

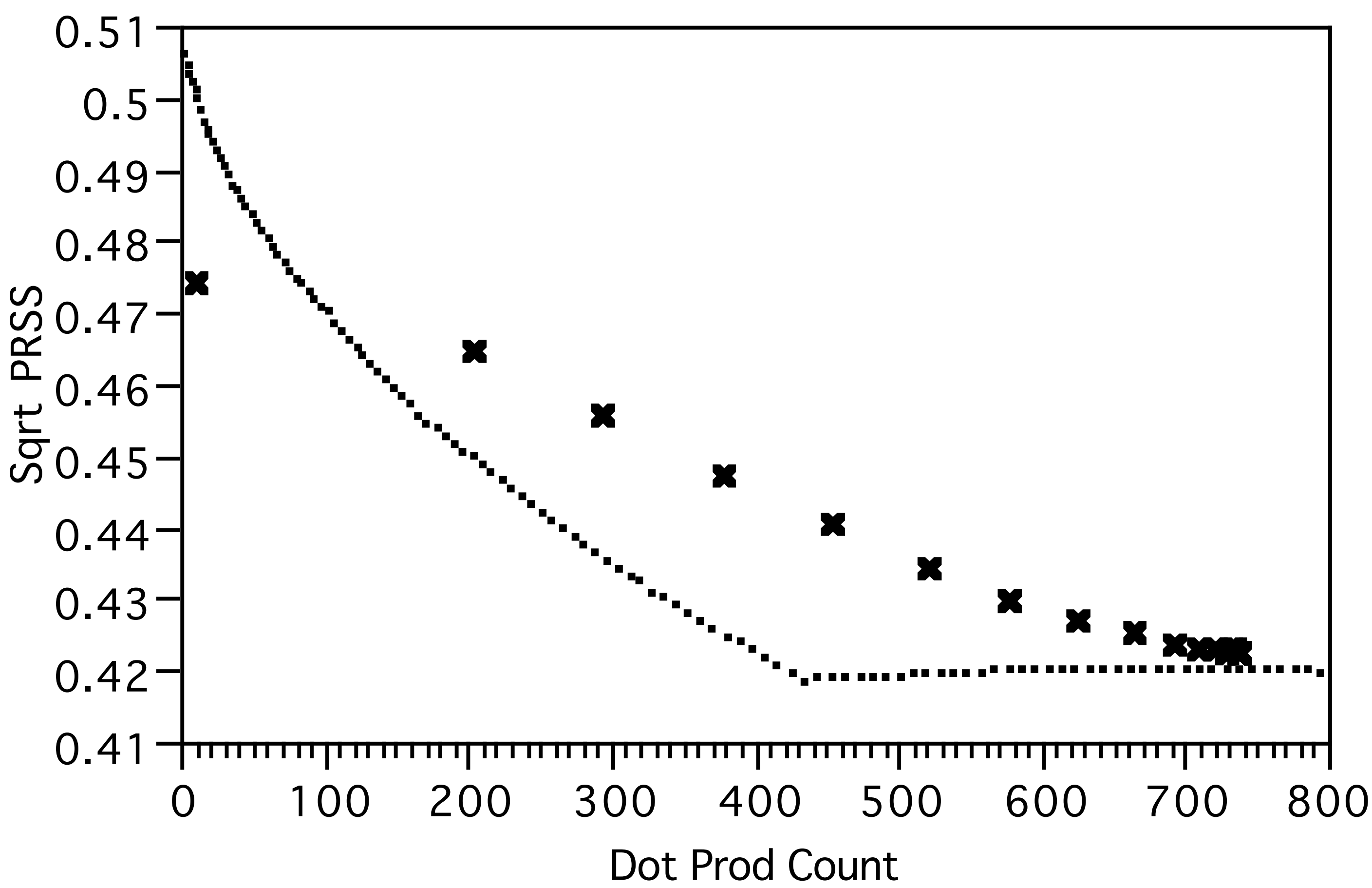
Does it work with collinear data?

Wharton
Department of Statistics

- Yes!
- Next slide shows results of small simulation.
  - Plot shows out-of-sample error on number of n-fold dot-products required to achieve the fit.
- Model distributes fit over many coefficients, so large unexplained variation obscures useful predictors.
- Sequential searches predictors in random order
  - i.e., no useful domain knowledge is being used.
- Batch search is usual greedy stepwise solution

Sequential works…

Wharton
Department of Statistics



Test problem for auction

Wharton
Department of Statistics

- Dataset of 2000 persons
  - All had been accepted.
  - Roughly half turn out "good" and half "bad".
- Want to predict outcome status.
  - Have ~ 100 application characteristics
- Prior models
  - Search of linear effects: lots of features in model
  - Allowing interactions: so many possible, they control the fit.
- Mixture
  - Want to have a model that gives linear terms more weight, while still reserving chance for interactions.

Auction analysis

Wharton
Department of Statistics

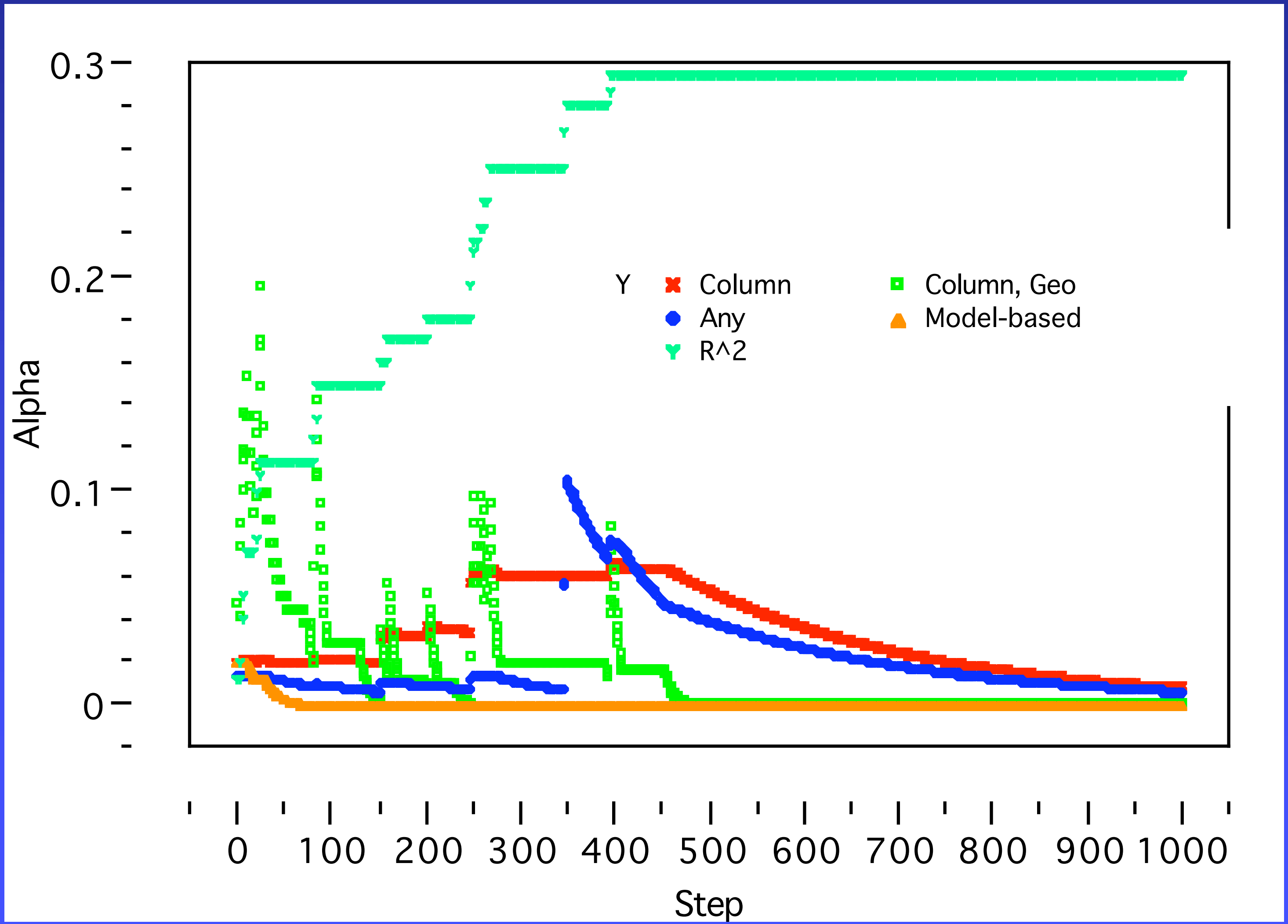
- Three streams of predictors
  - One linear, one quadratic
    - Use ordering of predictors from data file
  - One model based
    - Forms interactions from terms picked in model
- Four bidders
  - Two want linear; another wants anything
  - One looks for terms that expand current model
  - Mix of constant rate and “conservative optimist”
- Auction begins with overall error rate α = 0.05.

Auction analysis

- Prior fits
  - Linear experts: fit accuracy 26%
  - Quadratic (pure interaction): fit accuracy 23%
- Auction model
  - Linear bidders win heavily initially, increasing their wealth
- Final auction fit
  - Add interaction of terms in model quadratic
  - Improves fit accuracy to 29%
- Next slide shows components of auction error rate.
  - Division between linear and quadratic bidders.

Auction progress: alpha & R^2

Wharton
Department of Statistics



Summary and discussion

Wharton
Department of Statistics

- Auction modeling exploits both
  - Domain knowledge
  - Automatic search procedures
- Auctions possess well-developed foundations
  - Family wide error rate and step-up/step-down test
    - Powerful heuristic motivates use
  - Optimality of thresholding methods
    - Automatically generated threshold
  - Analogies to economic modeling, game theory
  - Machine learning theory