Wharton
Statistics Department

Auction Modeling

Robert Stine

Statistics Department

The Wharton School, Univ of Pennsylvania

October, 2004

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

Range of Challenges

Wharton
Statistics Department

- Anticipate bankruptcy
  - Which borrowers are most likely to default soon?
- Adverse effects
  - Which patients are at risk of side effects from medication?
- Facial recognition
  - How can we train computers to find faces in images?
- Other domains…
  - Employee evaluation: Who should we hire?
  - Fraud detection: Which loan applications were made up?
  - Document classification: Can you find one like this?

Similarities

Wharton
Statistics Department

Different contexts, but some similarities too …

- Rare events
  - Few cases dominate costs.
  - Millions of accounts, thousands of defaults.
- Synergies
  - Linear models find little. Interactions work.
  - Too many combinations seem plausible.
- Wide data: possibly more features than cases
  - Interactions, transformations, categories, missing data…
  - Too many to find the best at each stage.

Data sets keep getting wider

Application	*Number of Cases*	*Number of Raw Features*
Bankruptcy	3,000,000	350
Faces	10,000	1,400
Genetics	1,000	10,000
CiteSeer	500	∞

Common Objective

Wharton
Statistics Department

- Regardless of the context
  - Anticipating default on loan
  - Identifying those at risk of disease
  - Deciding whether there's a face in the image
- Pragmatic goal remains *prediction*.
- Best model generates highest revenue
  - Asymmetry of costs, presence of rare events
- Many schemes for building a predictive model
  - Various algorithms, features, and criteria such as…

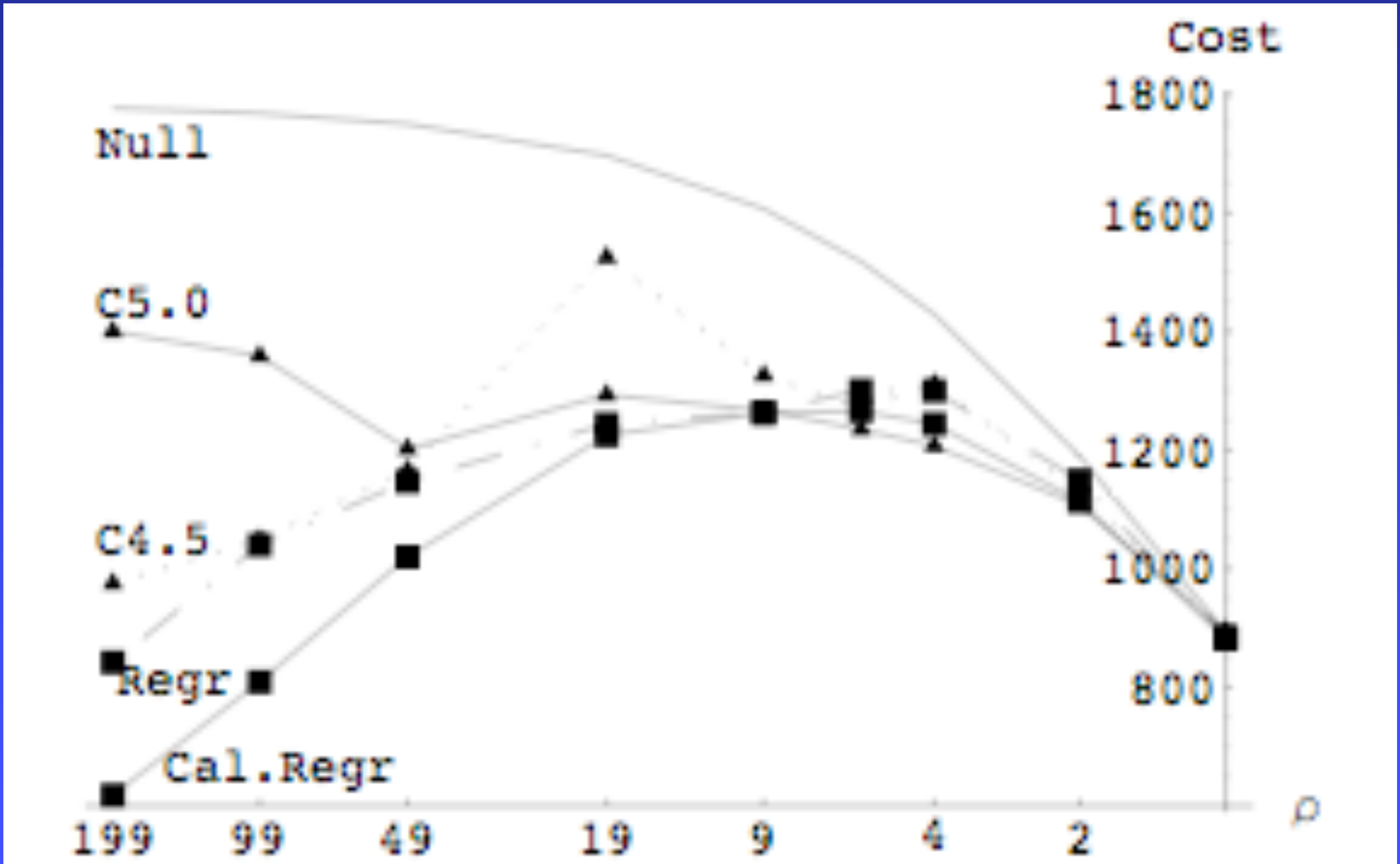
Background: Predicting BR

Wharton
Statistics Department

- Asymmetry of the costs
  - False positive (annoying a good customer): many but cheap
  - False negative (missing a bankruptcy): few but expensive
- A "slightly modified" version of stepwise regression predicts incidence of bankruptcy better than modern classification tree.
- Test results
  - Five-fold cross validation, with 600,000 cases in each fold.
  - Regression generate better decisions than using C4.5, with or without boosting.

Regression Minimizes Costs

Wharton
Statistics Department



Simple Mods to Regression

Wharton
Statistics Department

- To work well in data mining, regression needs help.
- Modified the statistics
  - Estimate standard errors using the fit computed *before* adding a predictor rather than after.
  - Bound p-values based on Bennett's inequality to control for very rare, high leverage points, then use Bonferroni.
  - Calibrate the final fit so that if the model predicts a 5% chance of BR, then we observe a 5% rate.
- Modified the computing by rearranging sweep order.
- Modified the search to consider *all* interactions.

How many predictors?

Wharton
Statistics Department

- Began with 350 predictors
  - These include categorical factors, such as region.
  - Missing data indicators
- Add all possible interactions
- Use forward stepwise regression to search the collection of

  350 base predictors

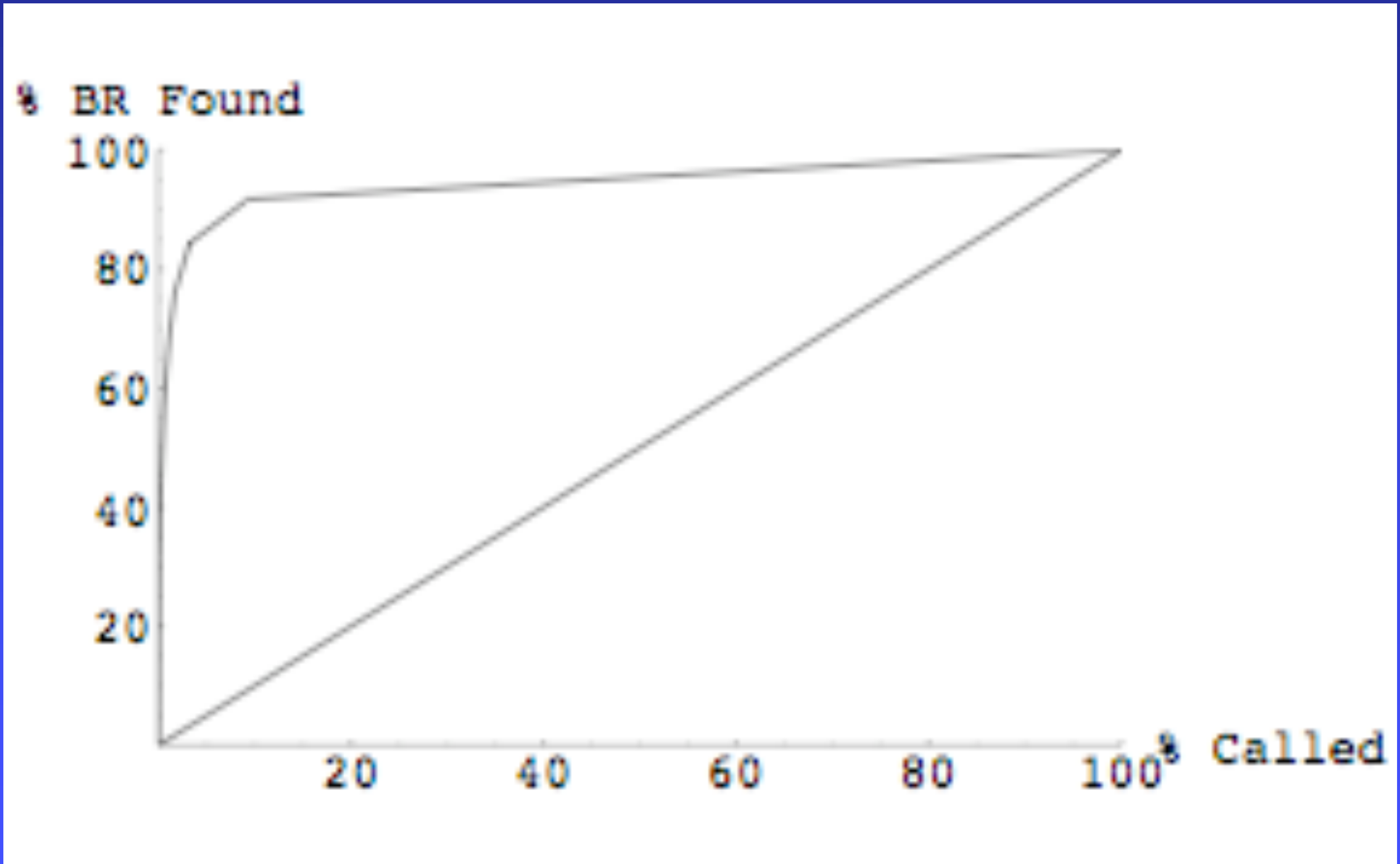
  + 350 squares of predictors

  + 350*349/2 = 66,430 interactions

  = 67,610 features

Impressive lift results

Wharton
Statistics Department



Successful, but …

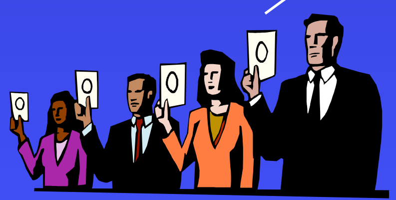
Wharton
Statistics Department

- Almost all predictors are interactions
  - Not surprising: more than 98% of the features considered in the search are interactions.
- Time consuming
  - "Breadth-first" search for next predictor
- Adding substantive features
  - Interactions represent but a few of the possible collection of features that one might want to explore.
  - If you were to talk to an expert, they could offer ideas.
  - How could you use this knowledge to find better models?

Not just one expert either…

Every domain has experts…

Which offer good advice?



How to use an expert's help?

Wharton
Statistics Department

<u>Manual</u>

Pick model "by hand"

- Advantages
  - Leverage domain knowledge
  - Can "interpret" model
- Disadvantages
  - Did we miss something?
  - Time consuming to
    - Construct
    - Maintain

<u>Automatic</u>

Computer search

- Advantages
  - Scans entire data warehouse
  - Hands-off, fast
    - Construction
    - Maintenance
- Disadvantages
  - Lost domain expertise
  - Hard to explain or interpret

Keep the good, remove the bad

Wharton
Statistics Department

Substantive

Pick model "by hand"

- Advantages
  - Leverage domain knowledge
  - Can "explain" model
- Disadvantages
  - Did we miss something?
  - Time consuming to
    - Construct
    - Maintain

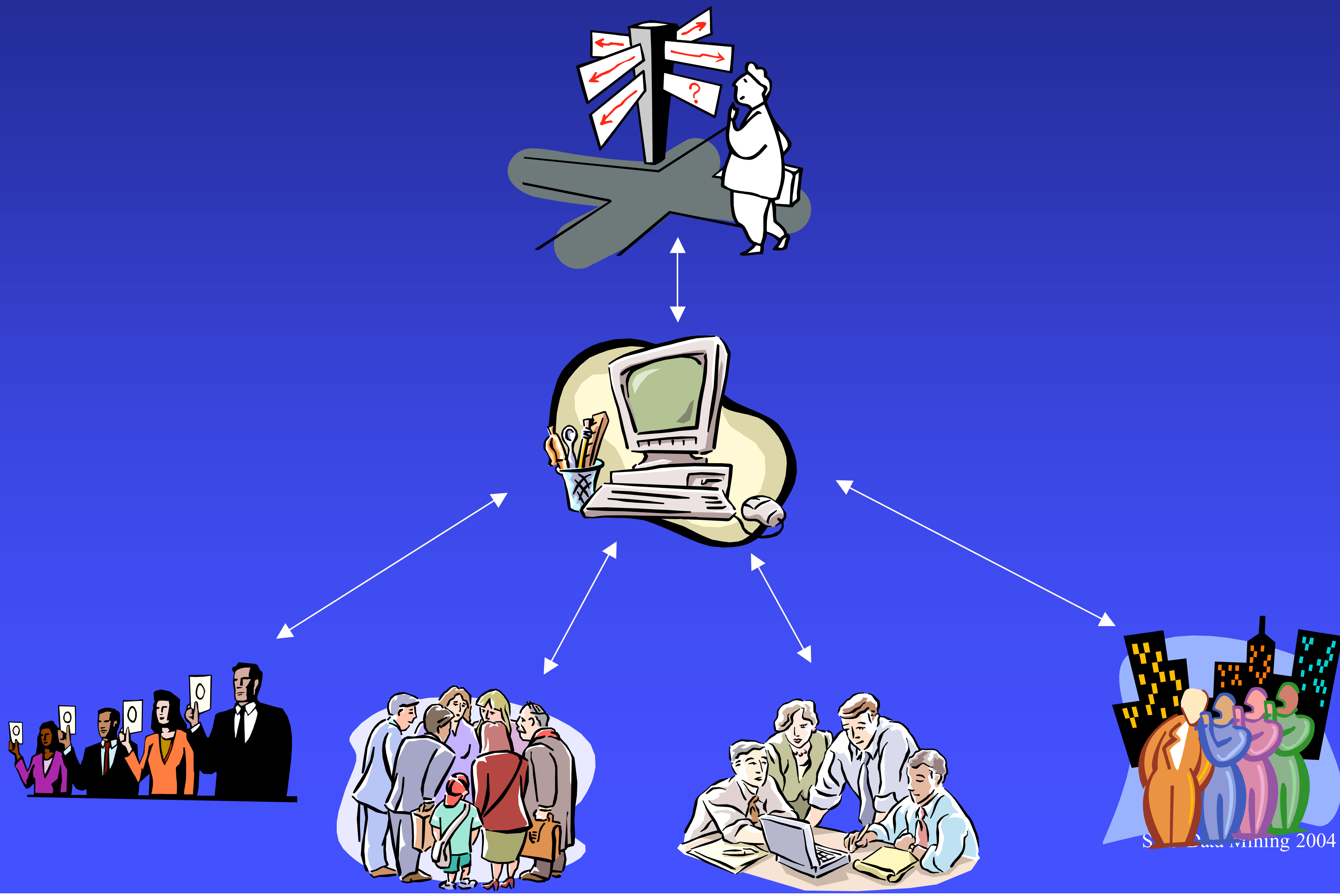
Automatic

Computer search

- Advantages
  - Scans entire data warehouse
  - Hands-off
    - Construction
    - Maintenance
- Disadvantages
  - Lost domain expertise
  - Hard to explain or interpret

Best of Both

Wharton
Statistics Department



Auction = Experts + Model

Predictive Model

Feature Auction

Domain Expert

Domain Expert

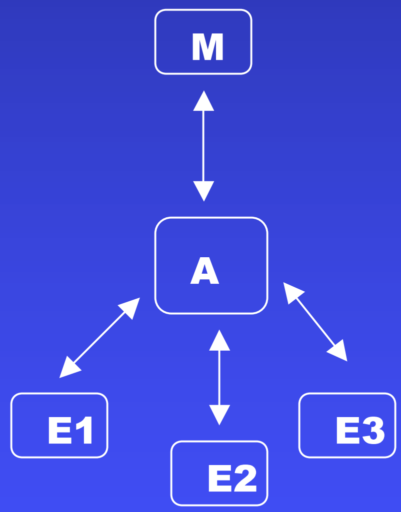
Domain Expert

Domain Expert

AWKTION Modeling

Wharton
Statistics Department

- *Experts* recommend features based on context.
- *Auction* identifies feature with highest bid.
- *Statistical model* tests this feature.
  - Bid determines p-value threshold
  - Accepts significant predictors, rejects others
- *Auction* passes results back to experts.
  - Winning bids earn wealth for expert.
  - Losing bids reduce wealth.
- *Information* flows both ways.



Experts

Wharton
Statistics Department

- Experts recommend predictive features
- *Substantive* experts order features
  - Domain knowledge of specific area
  - Offer a list of features to consider
  - Scheme/strategy to generate "next" predictors
- *Automatic* experts
  - Interactions based on other experts
  - Transformations
    - Segments, nearest-neighbor, principal components
    - Nonlinearity

Auction is sequential

- Each expert offers a predictor to the auction.
  - Each expert has wealth as allowed Type 1 error rate.
  - Experts offer a bid with each predictor.
  - The bid is a p-to-enter threshold.
- Auction takes the predictor with the highest total bid.
  - It collects the bids on this feature from the experts.
- Auction passes the chosen predictor to model.
  - Model assigns p-value to feature.
  - If p-value < bid, add the feature and “pay” bidders.
- Continue

Auction addresses concerns

- More types of features get used
  - One expert recommends raw predictors.
  - Second expert recommends interactions.
  - Second expert has to spread wealth over more possibilities
- Each step of the search is fast
  - "Depth-first" searching is fast. Just need p-value, not best.
  - The only game in town if the list of features is endless.
- Experts capture knowledge
  - Recommend features from substantive knowledge
  - Recommend features from state of the current model

Theory: Sequential selection

Wharton
Statistics Department

- Evaluate each feature as offered rather than finding the best feature available.
  - Essential when the choice of the next feature depends on what has worked so far, as in CiteSeer application.
- Fast, even when experts are dumb.
- SDR: the sequential discovery rate
  - Resembles an alpha-spending rule as used in clinical trials
  - Works like FDR, but allows an infinite sequence of tests.
- Variable selection
  - Ordering captures prior information on size of effects

Sequential vs. Batch Selection

Wharton
Statistics Department

Sequential

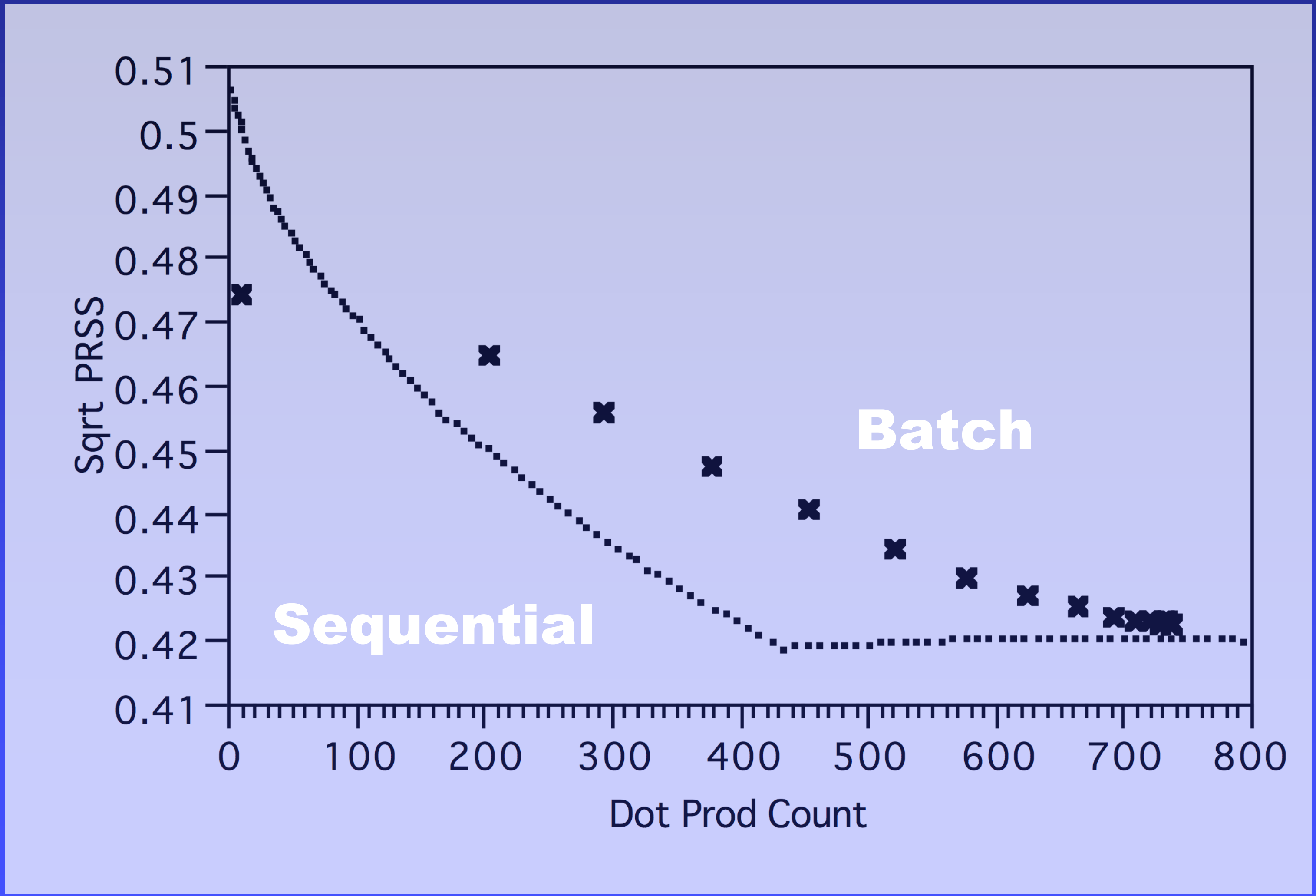
- Search features in order identified by domain expert
- Allows an infinite stream of features.
- Adapts search to successful domains.
- Reduces calculations to a sequence of simple fits.

Batch

- Search "all possible" features to find the best one.
- Needs all possible features before starts.
- Constrains search to those available at start.
- Requires onerous array manipulations.

Sequential works…

Wharton
Statistics Department



Theory: Bidding strategy

- Auction prevents “strategic betting”
  - Experts offer honest estimate of value of the predictor.
- Multiple bidders represent each expert
  - Geometric bidder: Spend λ% of current wealth on next bid.
  - Use mixture of bidders with varying λ.
- Auction adaptively discovers smart experts
  - Auction rewards the bidder/expert with the right rate
  - Wipes out the others.
- Universal bidding strategies (universal Bayes prior)

Statistical Model

Wharton
Statistics Department

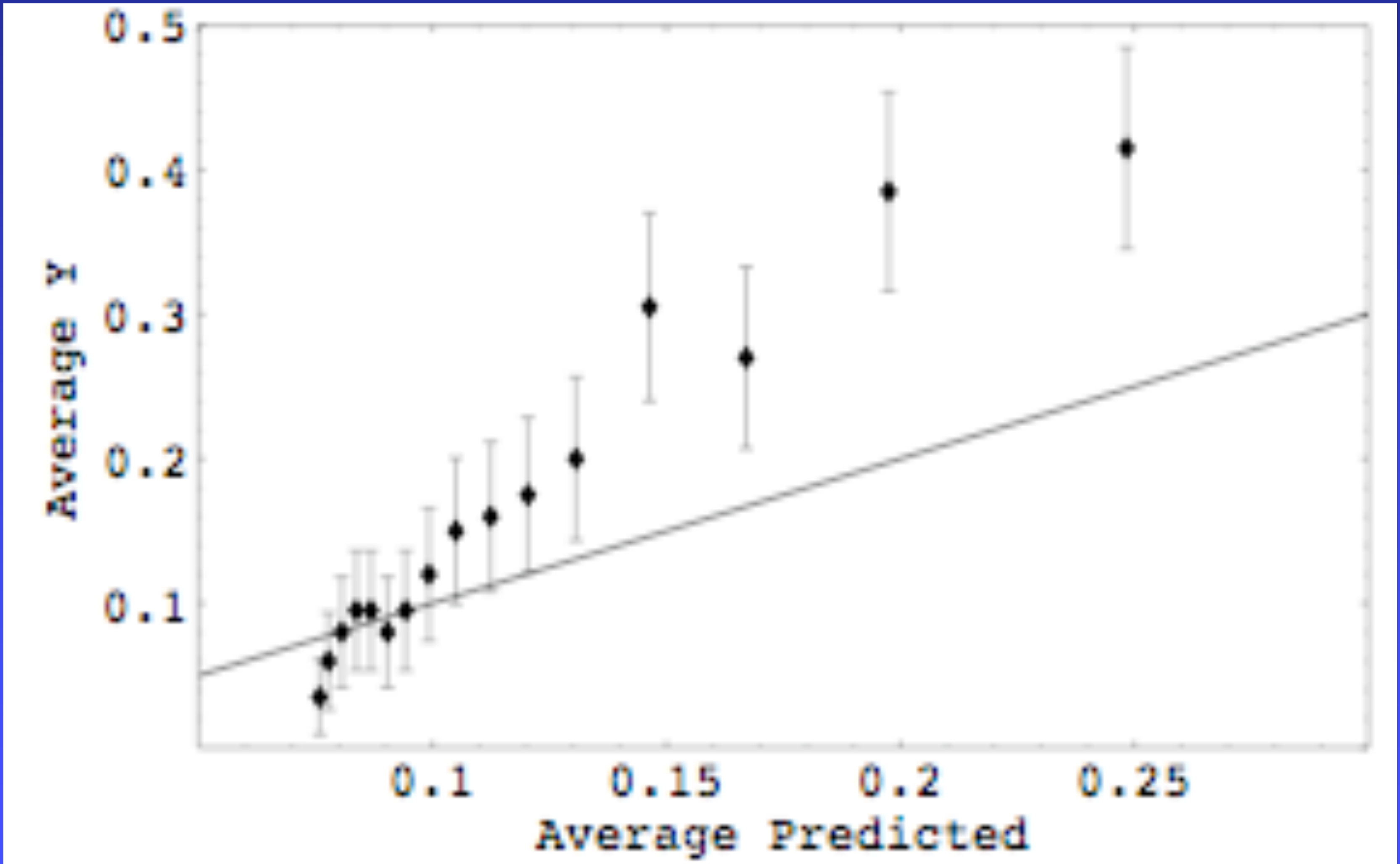
- Calibrated logistic regression
- Logistic regression
  - Well matched to classification
  - Allows over-sampling on the response
  - Simple calculations for scoring predictors
- Calibration
  - First-order calibration

$$E\left(Y \middle| \hat{Y}\right) = \hat{Y}$$

  - Build a calibrator using a smoothing spline to avoid predictors that only serve to calibrate the model.

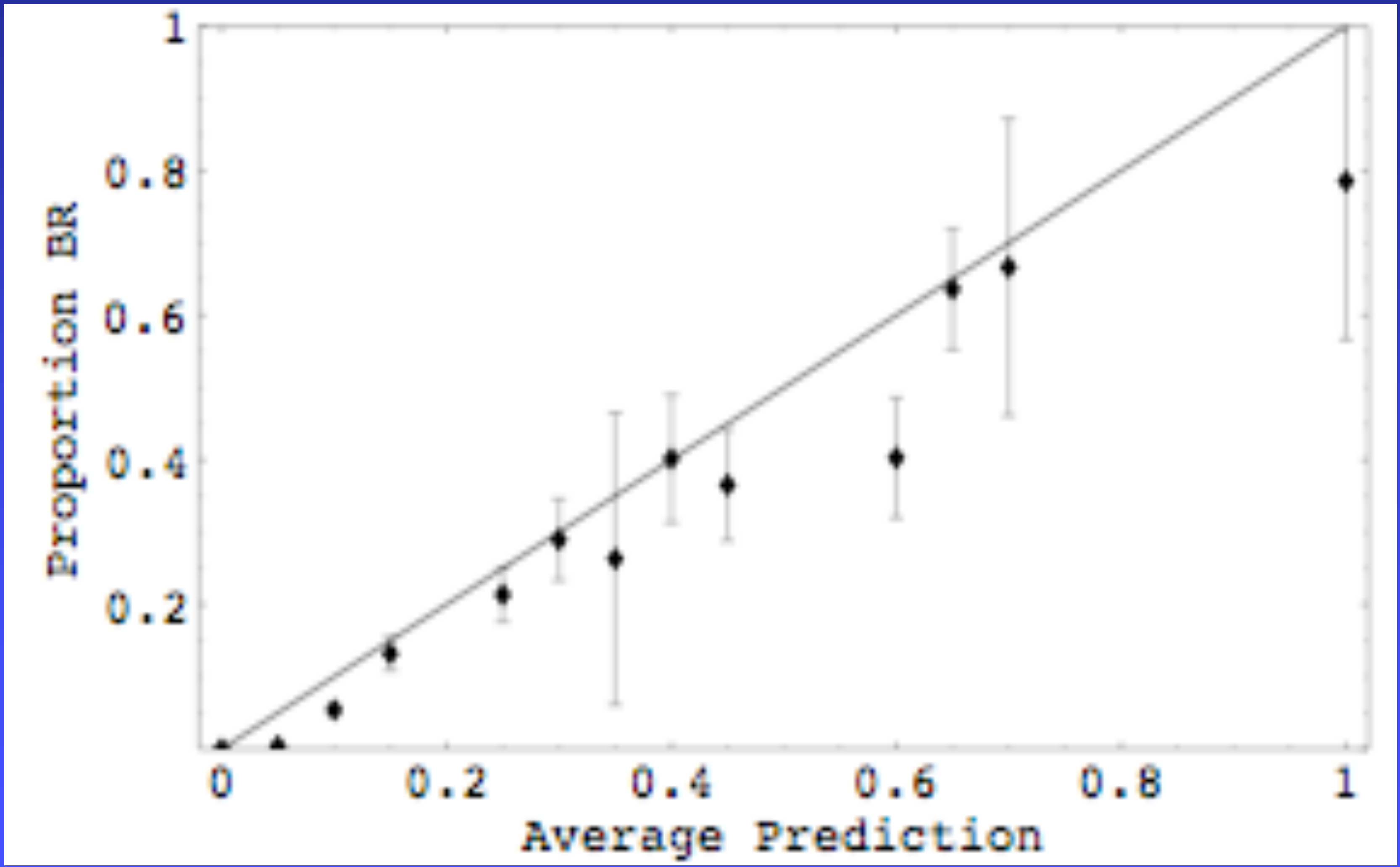
Calibration plot, before

Wharton
Statistics Department



Calibration plot, after

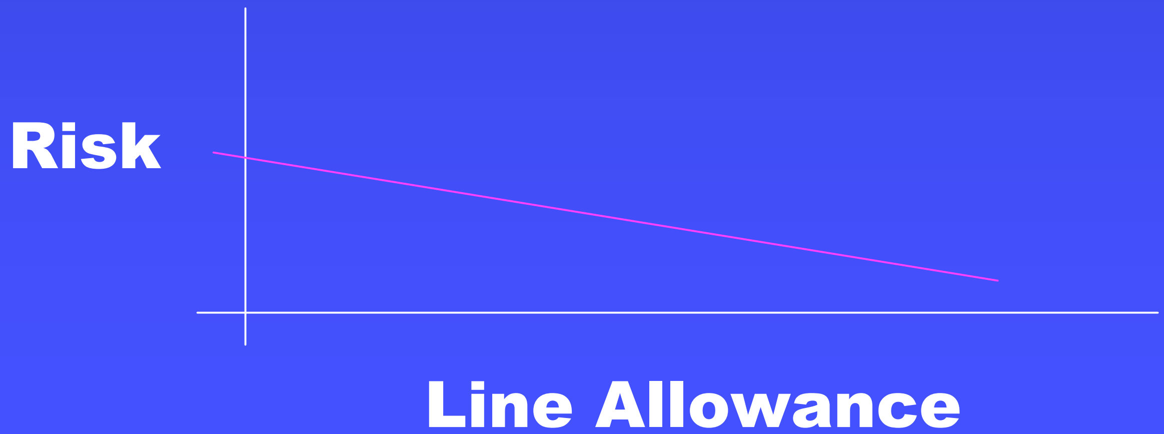
Wharton
Statistics Department



Stylized Example

Wharton
Statistics Department

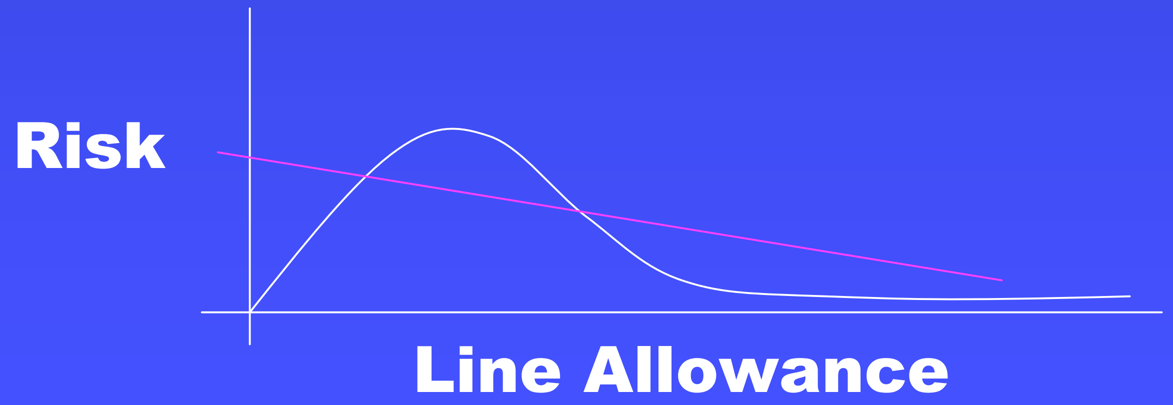
- Predicting default
  - Logistic regression model
  - 15,000 cases, 67,000 possible features (most interactions).
- Standard search finds linear predictor
  - Higher risk with lower line allowance.
  - Statistically significant



Discovers nonlinear pattern

Wharton
Statistics Department

- Auction model
  - Experts recommendations based on state of model.
  - Look for combinations of extant predictors.
- Discovers nonlinear effect
  - Nonlinear effect for size of credit line
  - Statistically significant "bump" in risk



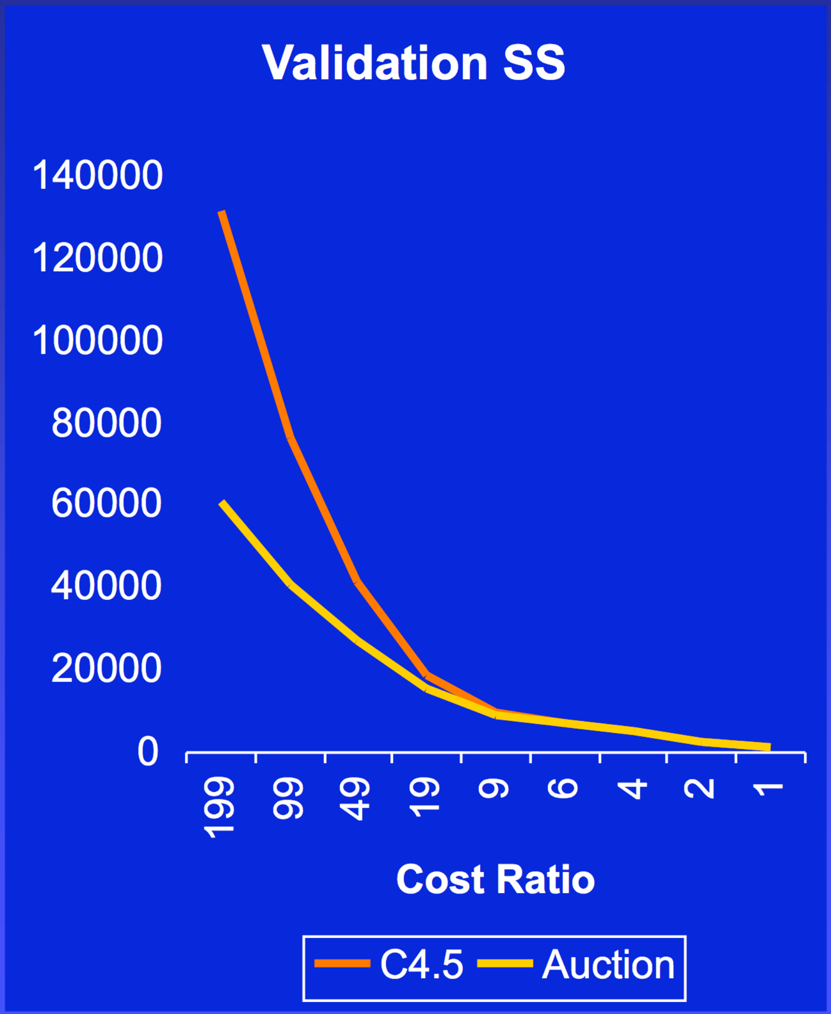
Cross-validation comparison

Wharton
Statistics Department

- Rare events data
- Five-fold "reversed" cross-validation
  - 100,000 cases per fold
  - Fit on one fold, predict other 4 folds
- Methods
  - C 4.5 with boosting
  - Auction with calibrated logistic regression and multiple geometric experts using SDR to spend alpha rate.
- Goal: Minimize costs of classification errors in the validation data.

Cross-validation comparison

Wharton
Statistics Department



- At higher cost ratios, auction produces much lower error costs.
- If the two errors have equal costs, either method does well.
- For each fold, auction builds one model for all cost ratios.
- C4.5 uses a new tree for each fold and for each cost ratio within a fold.

Want to try?

- Statistics should have (or use) a repository of test data sets like those used in computer science.
  - UC Irvine repository
- Can get this data from my web page.
  - Sanitized version of the bankruptcy data used in our study of data mining with regression.
  - Hidden the variable names and standardized the columns.
  - Reduced the data to 100,000 cases per fold.
- Only ask that you let us know how it turns out.

Computing comments

Wharton
Statistics Department

- Prior code
  - Monolithic C program
- Auction
  - Written in C++, using objects and standard libraries
  - Modular design
    - Templates (e.g., can swap in different type of model)
    - Runs as a unix command-line task
    - Separate commands for data processing, modeling, and validation
    - Adopt C4.5 file layout convention

Summary

Wharton
Statistics Department

- Auction modeling combines
  - Domain knowledge
  - Automatic search procedures
- Offers
  - Fast, guided search over complex domains
  - Ability to handle very wide data sets
  - Use of any model that can provide p-value
- More information…

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