Streaming Feature Selection

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

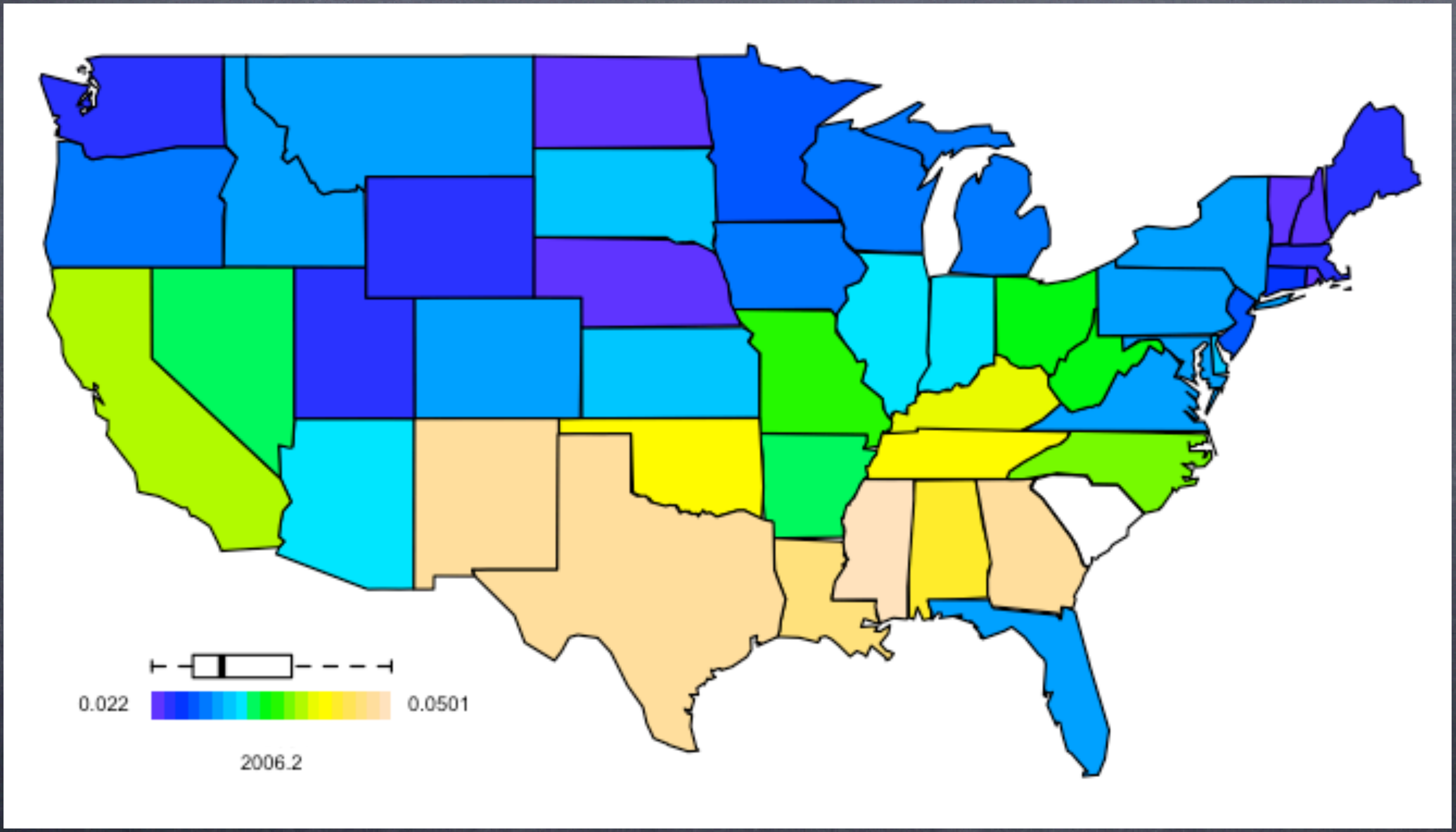
Plan

- Motivating applications: predictive models
  - Credit default rates
  - Linguistics
- Sequential testing
  - Alpha investing
- Robust standard errors
  - Sandwich estimator
- Auction framework
  - Blend several streams, strategies
- Collaborators
  - Dean Foster
  - Dongyu Lin

Applications

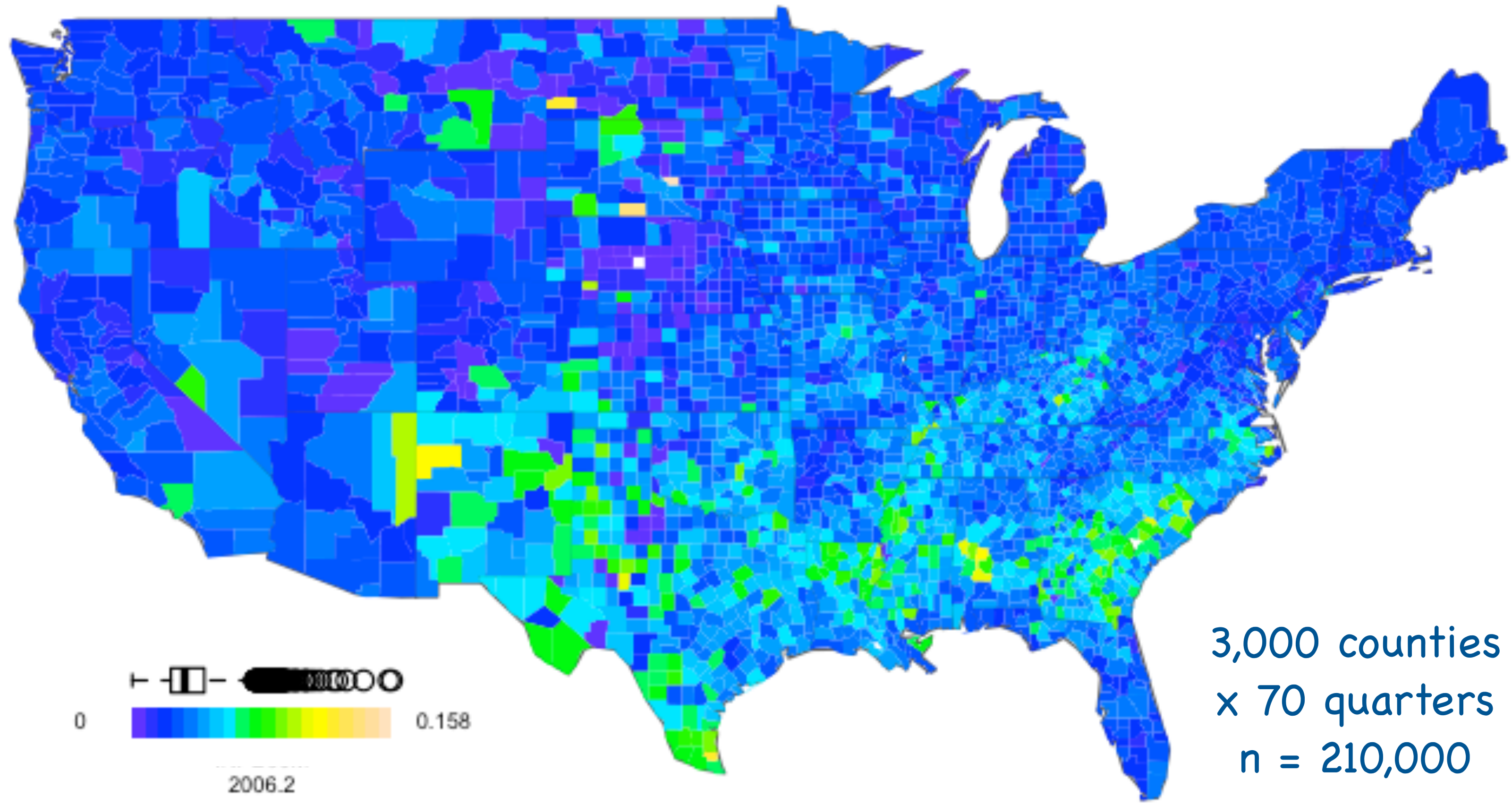
Spatial Temporal Models

- Goal
  - Predict default rates, such as in credit cards



Spatial Temporal Models

- Goal
  - Predict default rates, such as in credit cards

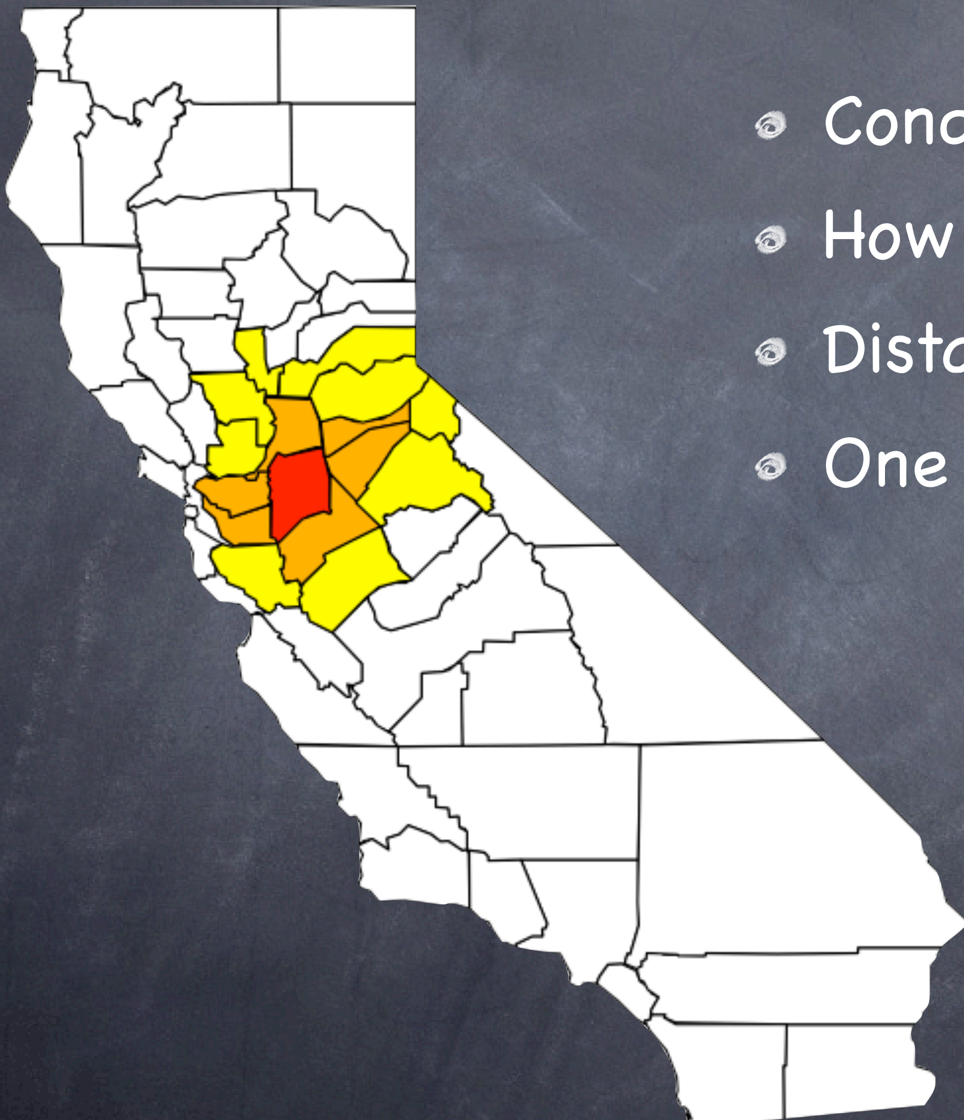


3,000 counties
x 70 quarters
n = 210,000

Plan to move to individual consumer next...

Wharton
Department of Statistics

Spatial Dependence



- Conditional AR? (Markovian)
- How many layers?
- Distance measure?
- One shoe fits all?

Spatial Temporal Models

- Refined goal: compare to benchmark
  - Predict default rates better than possible using only the local history of default.
  - Implications for bank's data needs
- Possible predictors
  - Macroeconomic factors
  - Default trends in nearby counties
  - Non-linear effects, interactions
  - Spatial variation in model structure
- Modeling issues
  - Dependence (spatial, temporal)
  - Heterogeneity among counties
  - Population drift: EBay patterns, hiring model

Computational Linguistics

- Variety of applications…
  - Word disambiguation
    Does "Georgia" refer to a person, US state, or perhaps to a Nation?
  - Speech tagging
    Identifying noun, verb, adjective…
  - Cloze (predicting the next word)
    "…in the midst of modern life the greatest, ____"
- Huge corpus of data
  - x,000,000 cases
  - novels, news feeds, web pages
  - text of Wikipedia used to seem huge

Challenges in Text

- Cloze
  - Is the next word "the" or "her"?
    "...in the midst of modern life the greatest, ____"
  - Balanced training data with 50/50 rate
- Possible predictors
  - Word frequencies (bag of words)
  - Neighboring sentences/words
  - Parts of speech, tree banks, stem words, synonyms
- Transfer learning
  - Do predictors based on Washington Post work for text from NY Times?
  - Dependence, unobserved latent structure

Similarities

Text	Credit
Predict word in new documents, different authors	Predict rates in same locations, but changing economic conditions
Latent structure associated with corpus	Latent temporal changes as economy evolves
Neighborhoods: nearby words	Neighborhoods: nearby locations, time periods
Vast possible corpus	Only 3,000 counties but possible to drill lower
Sparse	May be sparse

Methods

Modeling Challenge

- We like regression models
  - Familiar, interpretable, good diagnostics
- Regression models have worked well
  - Predicting rare events, such as bankruptcy
    Competitive with random forest
  - Function estimation, using wavelets and variations on thresholding
- Extend to rich environments
  - Spatial-temporal data
    Retail credit default — MRF, MCMC
  - Linguistics, text mining
    Word disambiguation, cloze — TF-IDF
- Avoid overfitting...

TF-IDF: term frequency-inverse document frequency
frequency in document relative to frequency in corpus

MRF: Markov random fields

Recent news

June 12, 2010

The New York Times Reprints

A Decade Later, Genetic Map Yields Few New Cures

By NICHOLAS WADE

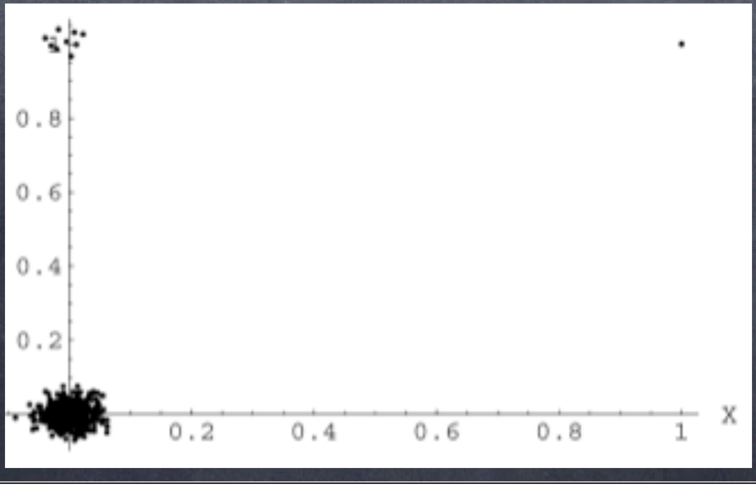
Ten years after President Bill Clinton announced that the first draft of the human genome was complete, medicine has yet to see any large part of the promised benefits.

For biologists, the genome has yielded one insightful surprise after another. But the primary goal of the $3 billion Human Genome Project — to ferret out the genetic roots of common diseases like cancer and Alzheimer's and then generate treatments — remains largely elusive. Indeed, after 10 years of effort, geneticists are almost back to square one in knowing where to look for the roots of common disease.

One sign of the genome's limited use for medicine so far was a recent test of genetic predictions for heart disease. A medical team led by Nina P. Paynter of Brigham and Women's Hospital in Boston collected 101 genetic variants that had been statistically linked to heart disease in various genome-scanning studies. But the variants turned out to have no value in forecasting disease among 19,000 women who had been followed for 12 years.

Lessons from Prior Modeling

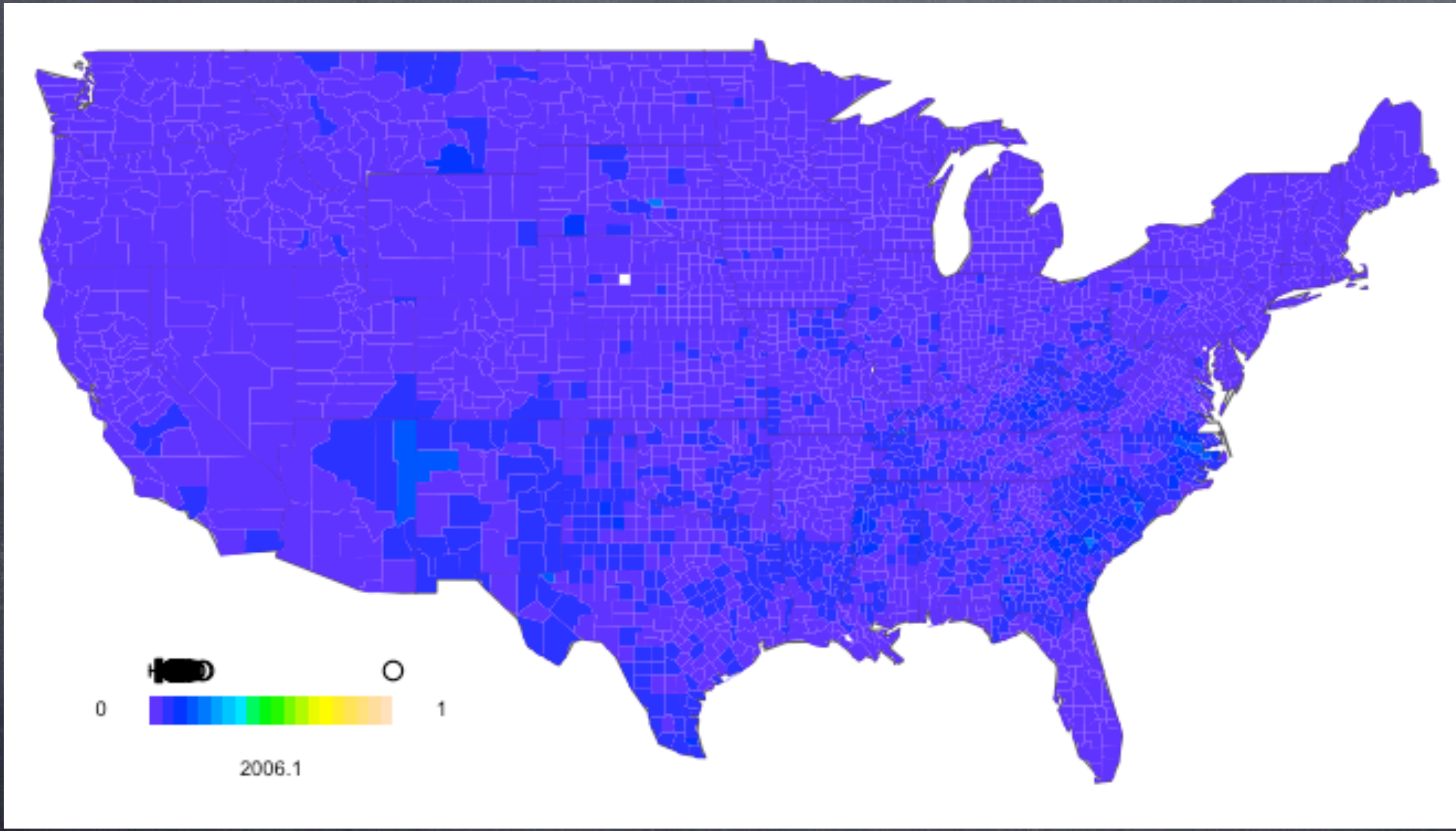
- Bankruptcy: n=500,000, p=60,000, 450 events
- "Breadth-first" search for best features
  - Slow, memory hog
  - Severe penalty on largest z-score, sqrt(2 log p)
- If tested features are mostly interactions, then selected features are mostly interactions
  - Example
    $\mu \gg 0$ and β_1, $\beta_2 \neq 0$, then $X_1 * X_2 \Rightarrow c + \beta_1 X_1 + \beta_2 X_2$
- Outliers cause problems even with large n



Real p-value ≈ 1/1000,
but
usual t-statistic ≈ 10

Wharton
Department of Statistics

Spatial Outliers Happen



Reaction to Lessons

- Breadth-first becomes streaming selection
  - Sequence of possible features
  - Examining each is very fast
  - Over-fitting? Multiplicity adjustments?
- Fixed significance levels replaced by levels that vary with the type of the variable
  - Heuristic: Revised Bonferroni (ie, hard threshold)
    Divide α level equally between linear & interactions
    p linear: test each at level $\alpha/(2p)$
    p^2 interactions: test at level $\alpha/(2p^2)$
- Rather than trust model to obtain standard errors, use a robust estimate.

Methods Overview

- "Linear" regression

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots$$

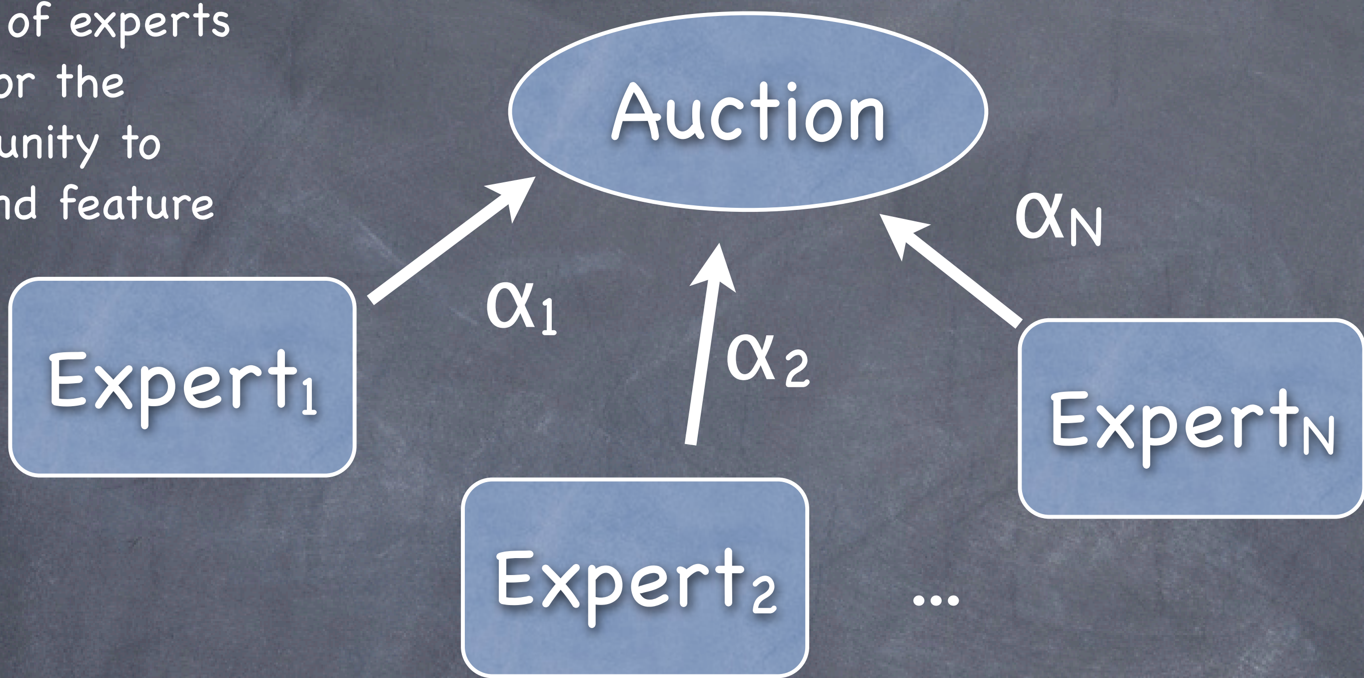
- Auction selection from multiple "experts"
  - Explore expansive feature space, including interactions and nonlinear subspaces
  - Exploit exogenous information
- Robust standard errors and p-values
  - Accommodate dependence and heterogeneity
- Alpha investing
  - Control over-fitting adaptively

Feature Auction

Collection of experts bid for the opportunity to recommend feature

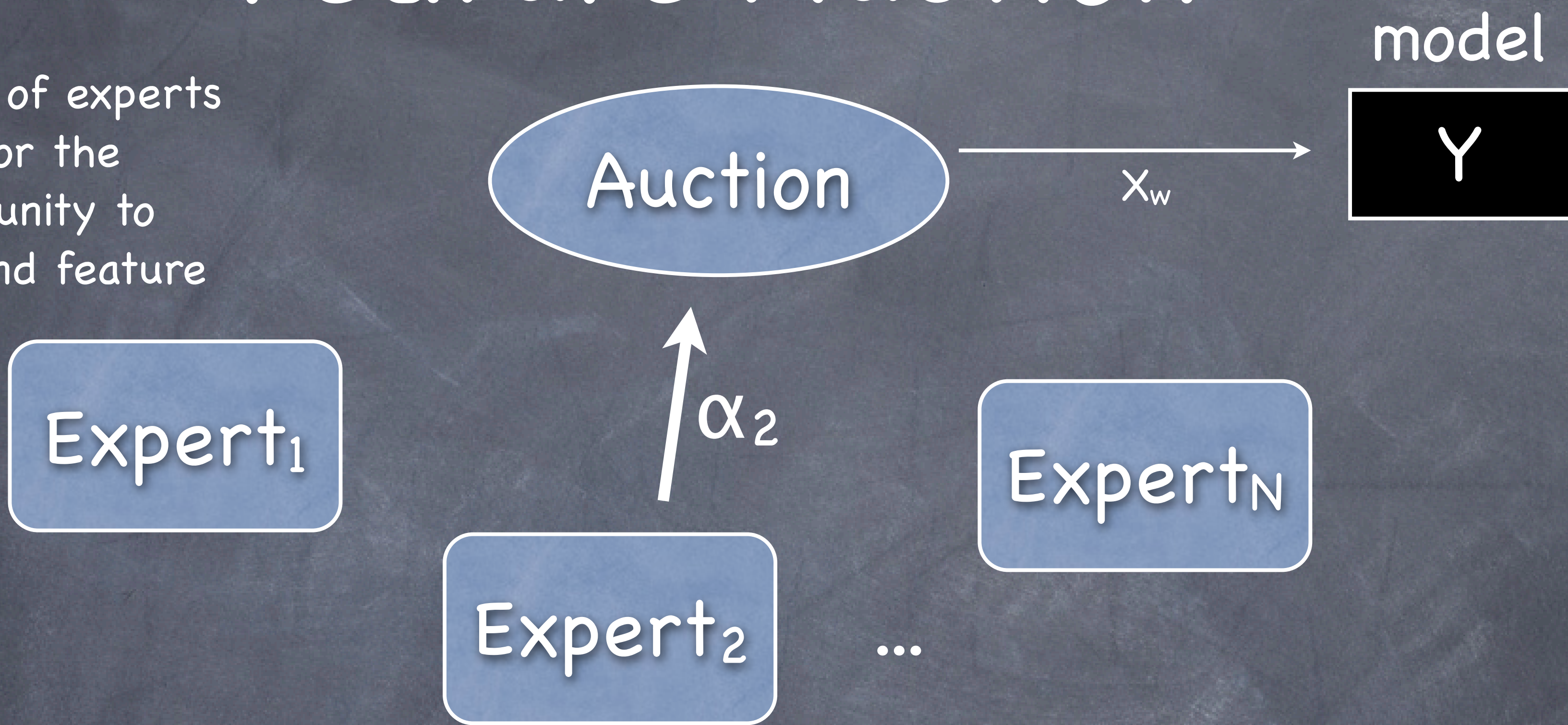
model

Y



Feature Auction

Collection of experts bid for the opportunity to recommend feature

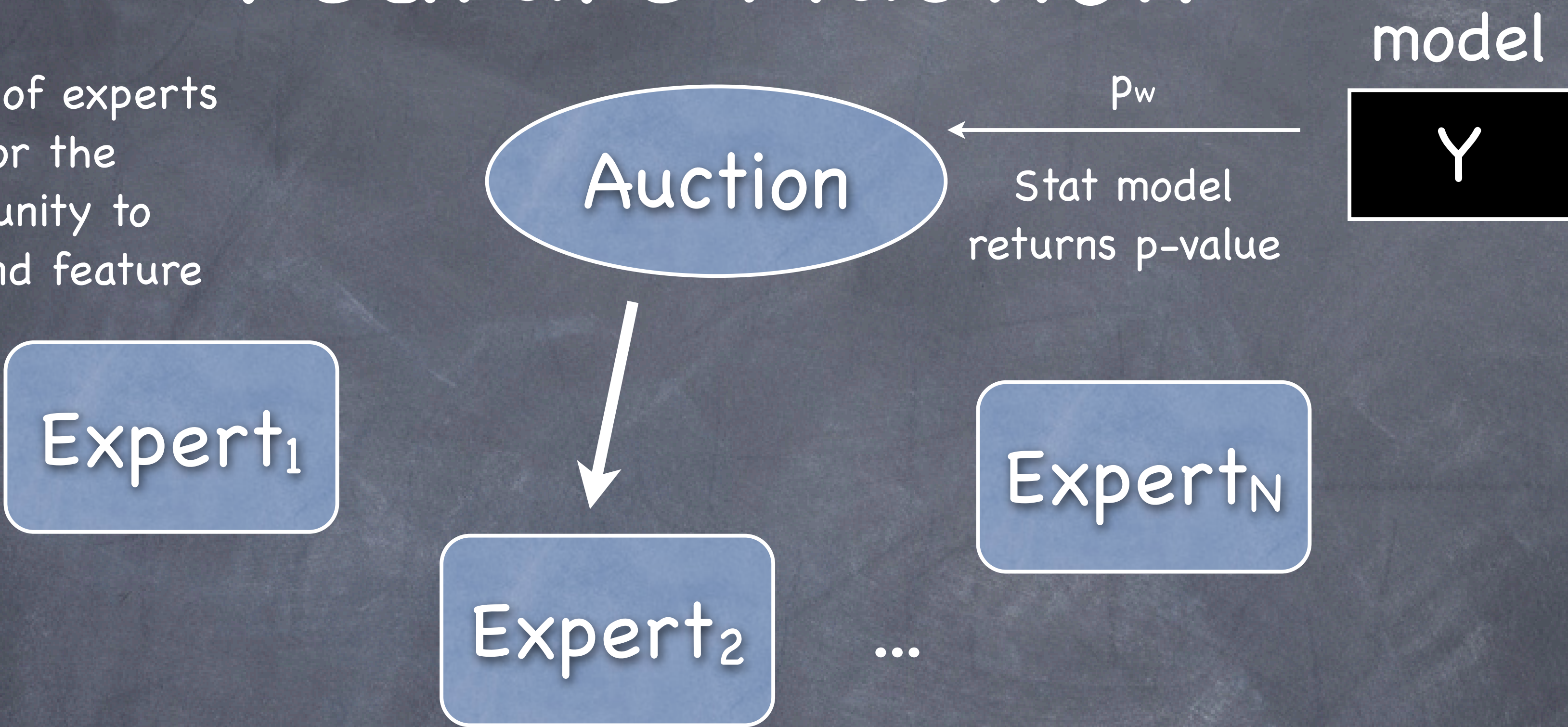


Auction collects winning bid α_2

Expert supplies recommended feature X_w

Feature Auction

Collection of experts bid for the opportunity to recommend feature



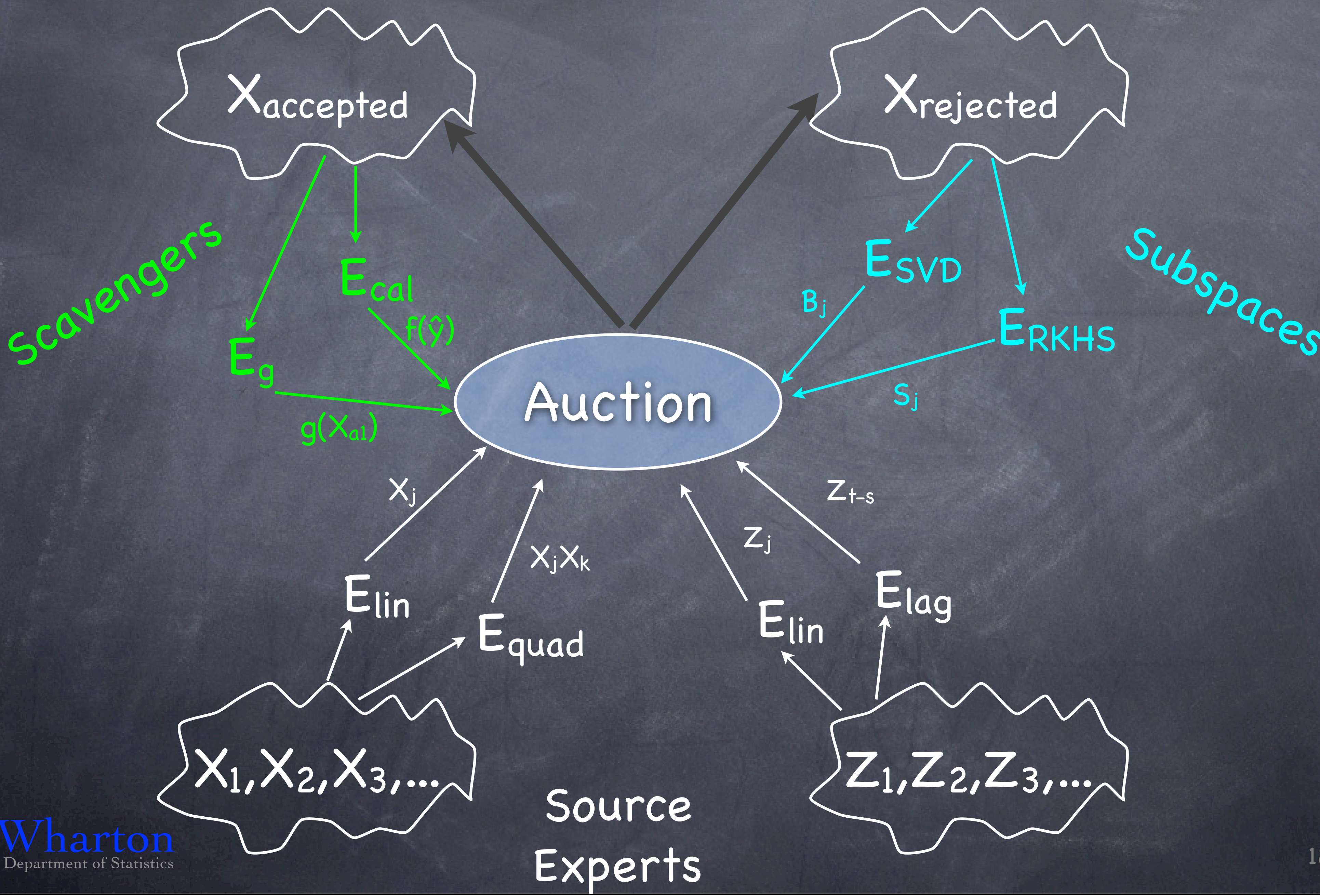
Auction collects winning bid α_2

Expert supplies recommended feature X_w

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

Experts learn if the bid was accepted, not the effect size or p_w.

Experts



$X_{accepted}$
$X_{rejected}$
Scavengers
E_{cal}
E_g
$f(\hat{y})$
$g(X_{a1})$
E_{SVD}
E_{RKHS}
Subspaces
B_j
S_j
Auction
X_j
X_jX_k
Z_j
Z_{t-s}
E_{lin}
E_{quad}
E_{lin}
E_{lag}
$X_1,X_2,X_3,...$
$Z_1,Z_2,Z_3,...$
Source
Experts

Experts

- Expert
  Strategy for creating list of features. Experts embody domain knowledge, science of application.
- Source experts
  - A collection of measurements (eg, synonyms, clusters)
  - Subspace basis (PCA, RKHS)
  - Lags of a time series
- Parasitic experts, scavengers
  - 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

Expert Wealth

- Expert is rewarded if feature accepted
  - 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, the auction
  - Rewards experts that offer useful features. These then can win later bids and recommend more X's
  - Eliminates experts whose features are not accepted.
- Taxes fund parasites and scavengers
  - Continue control overall FDR
- Critical
  - control multiplicity in a sequence of hypotheses
  - p-values determine useful features

Standard Errors

Robust Standard Errors

- p-values depend on many things
  - p-value = f(effect size, std error, prob dist)
  - Error structure likely heteroscedastic
  - Observations frequently dependent
- Dependence
  - Spatial time series at multiple locations
  - Documents from various news feeds
  - Transfer learning
    When train on observations from selected regions or document sources, what can you infer to others?
- What are the right degrees of freedom?
  - Tukey story

Sandwich Estimator

- Usual OLS estimate of variance
  - Assume your model is true

$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)\,(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

- Sandwich estimators
  - Robust to deviations from assumptions

heteroscedasticity

$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= (X'X)^{-1}\; X'D^2X\; (X'X)^{-1} \end{aligned}$$

diagonal

dependence

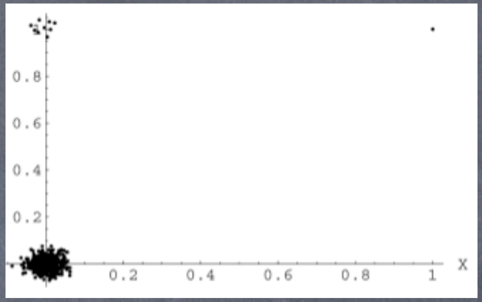
$$\begin{aligned} var(b) &= (X'X)^{-1}X'E(ee')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\; X'BX\; (X'X)^{-1} \end{aligned}$$

block diagonal

Essentially the "Tukey method"

Flashback...

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



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

Sequential Testing

Wharton
Department of Statistics

Alpha Investing

- Context
  - Test possibly infinite sequence of m hypotheses

    $H_1, H_2, H_3, \dots H_m \dots$

    obtaining p-values $p_1, p_2, \dots$
  - Order of tests can depend prior outcomes
- Procedure
  - Start with an initial alpha wealth $W_0 = \alpha$
  - Invest wealth $0 \le \alpha_j \le W_j$ in the test of H_j
  - Change in wealth depends on test outcome
  - $\omega \le \alpha$ denotes the payout earned by rejecting

$$W_j - W_{j-1} = \begin{cases} \omega & \text{if } p_j \le \alpha_j \\ -\alpha_j/(1-\alpha_j) & \text{if } p_j > \alpha_j \end{cases}$$

Wharton
Department of Statistics

Alpha Investing Martingale

- Provides uniform control of the expected false discovery rate. At any stopping time during testing, martingale argument shows

$$\sup_{\theta} \frac{E(\#\text{false rejects})}{E(\#\text{rejects})+1} \leq \alpha$$

- Flexibility in choice of how to invest alpha-wealth in test of each hypothesis
  - Invest more when just reject if suspect that significant results cluster.
  - Universal strategies
- Avoids need to compute p-values in advance

Connections

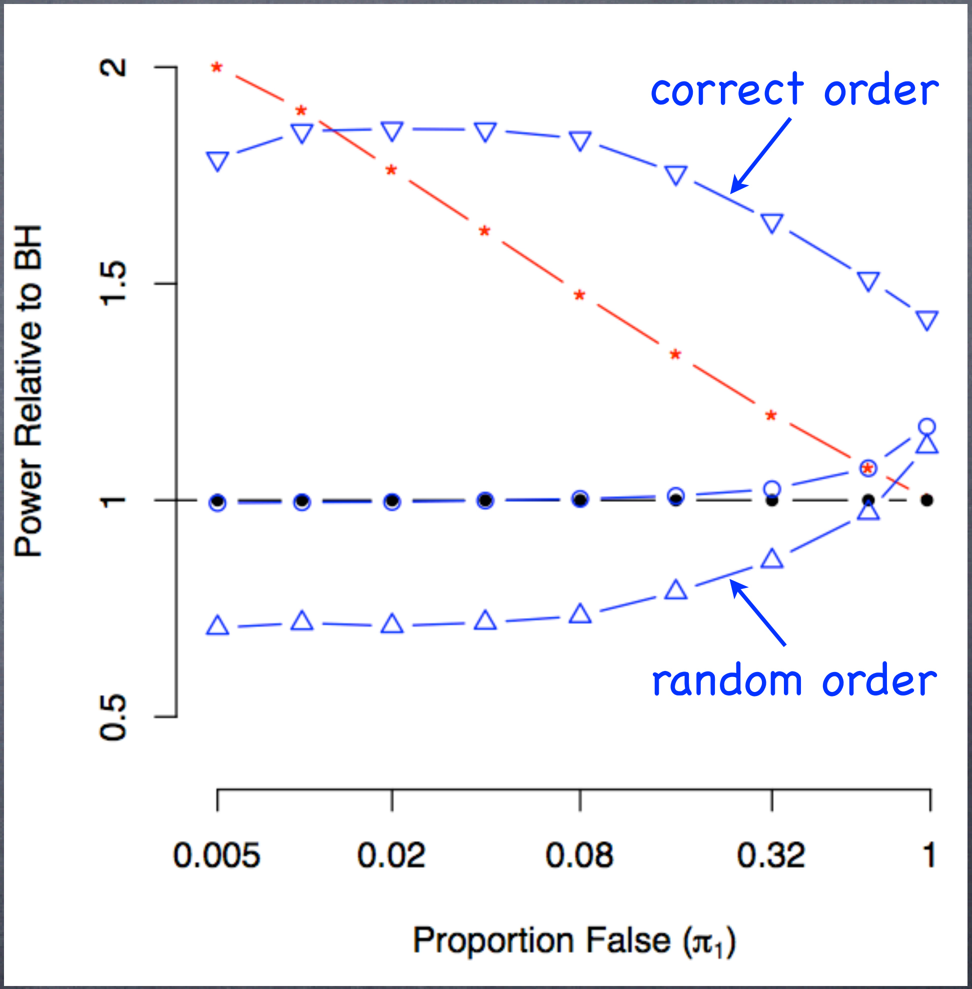
- Other methods of controlling false positives are special cases
- Bonferroni test of $H_1,\ldots,H_m$
  - Set $W_0 = \alpha$ and reward $\omega = 0$
  - Bid $\alpha_j = \alpha/m$
- Step-down test of Benjamini & Hochberg
  - Set $W_0 = \alpha$ and reward $\omega = \alpha$
  - Test all m at level α/m
  - If none are significant, done
  - If one is significant, earn α back
    - Test remaining m-1 conditional on $p_j > \alpha/m$

Benefits of Knowledge

- Simulation
- Test m = 200 hypotheses
- Compare power to Benjami-Hochberg
- Signal from spike and slab prior

Oracle BH

Alpha investing



Wharton
Department of Statistics

Next Steps

- Replace the martingale that controls alpha wealth by one that controls expected loss.
- Improved experts: more features
  - Neighborhood structure is an important method to create new types of features
    - geographical
    - temporal

    Both are links to other rows.
- Better software
  - Front end
  - Back end
  - Get some of that faster matrix code

References

- Feature auction
  - www-stat.wharton.upenn.edu/~stine
- Alpha investing
  - "α-investing: a procedure for sequential control of expected false discoveries", JRSSB, 2006
- Early improved stepwise regression
  - "Variable selection in data mining: Building a predictive model for bankruptcy", JASA, 2004
- Robust standard errors
  - "Variable selection in models with blockwise dependence", Lin and Foster.

Thanks!