Classification for Ranking in Drug Discovery: Identifying and Aggregating Relevant Subsets of Variables

Yuanyuan Wang
University of Waterloo, Canada (now with Pfizer)

Hugh Chipman
Acadia University

Will Welch
University of British Columbia, Canada
will@stat.ubc.ca

Research funded by MITACS, NSERC, GlaxoSmithKline

Outline

- Drug discovery
- National Cancer Institute (NCI) AIDS antiviral data
- Statistical methods: trees, k-nearest neighbours, etc.
- Averaging subsets of variables
- Prediction performance
- Identifying relevant variables
- Conclusions

What is a Pharmaceutical Drug?

A drug is just a (small) organic chemical compound.

e.g., Aspirin $C_9H_8O_4$.

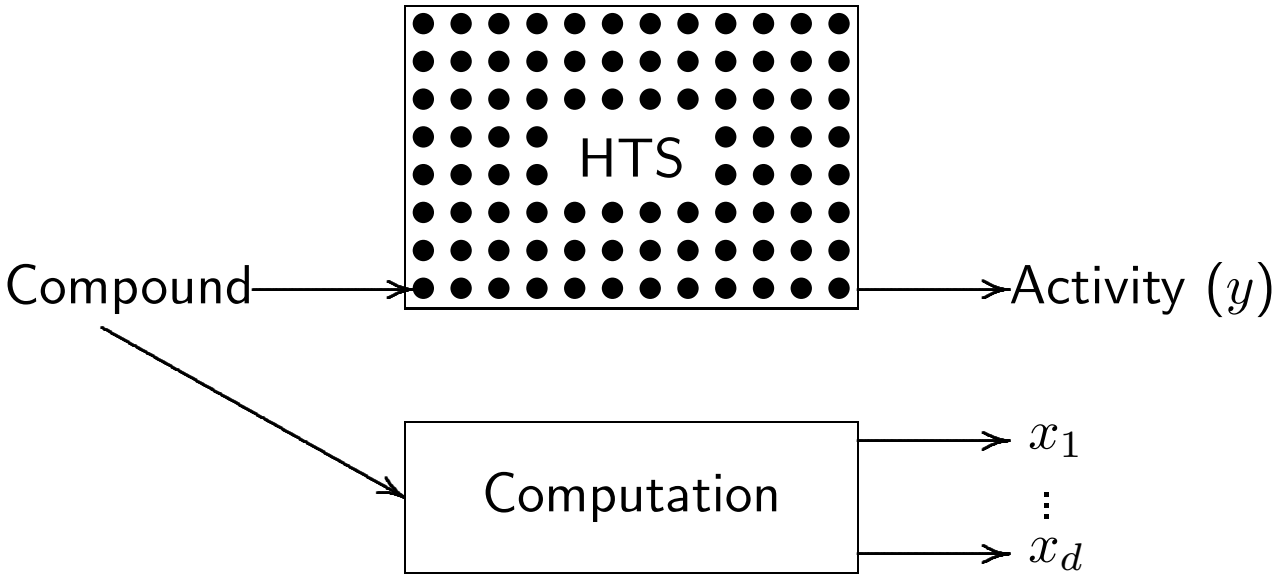
Drug Discovery's Place in Drug Development

- **Drug discovery:** find several chemical compounds that are active against a biological target (e.g., a specific protein)
- Lead optimization: atom-by-atom optimization (activity, toxicity, mutagenicity)
- Animal studies
- Clinical trials (Phase I, II, III)

High Throughput Screening Data

Response variable: Activity (% inhibition, IC50 concentration, inactive/active)

Explanatory variables: d molecular descriptors, $x_1, \ldots, x_d$



Sequential Screening

Could screen "everything" (i.e., assay the millions of compounds in a corporate collection).

Alternatively, sequential (smart) screening:

- Underlying premise: Biological activity (y) is related to chemical structure ($x_1, \ldots, x_d$)
- Assay 1000's of compounds in high throughput screening to measure y
- Model $y(x_1, \ldots, x_d)$
- Use model to predict y for millions of compounds
- Assay only the most promising compounds

Objective

Use the model to *rank* the unassayed compounds most promising to least promising:

- Build a classifier using the training data. Classifier gives scores for each compound in the test set:

  Estimated probability that $y = 1$ (active)

  (Continuous response: score is $\hat{y}$)

- Use the classifier to select relatively few compounds from the test set with the largest scores.

- Performance: How many active compounds ("hits") do we find?

Interpretation: Identify the variables important for activity and their ranges.

National Cancer Institute AIDS Antiviral Data

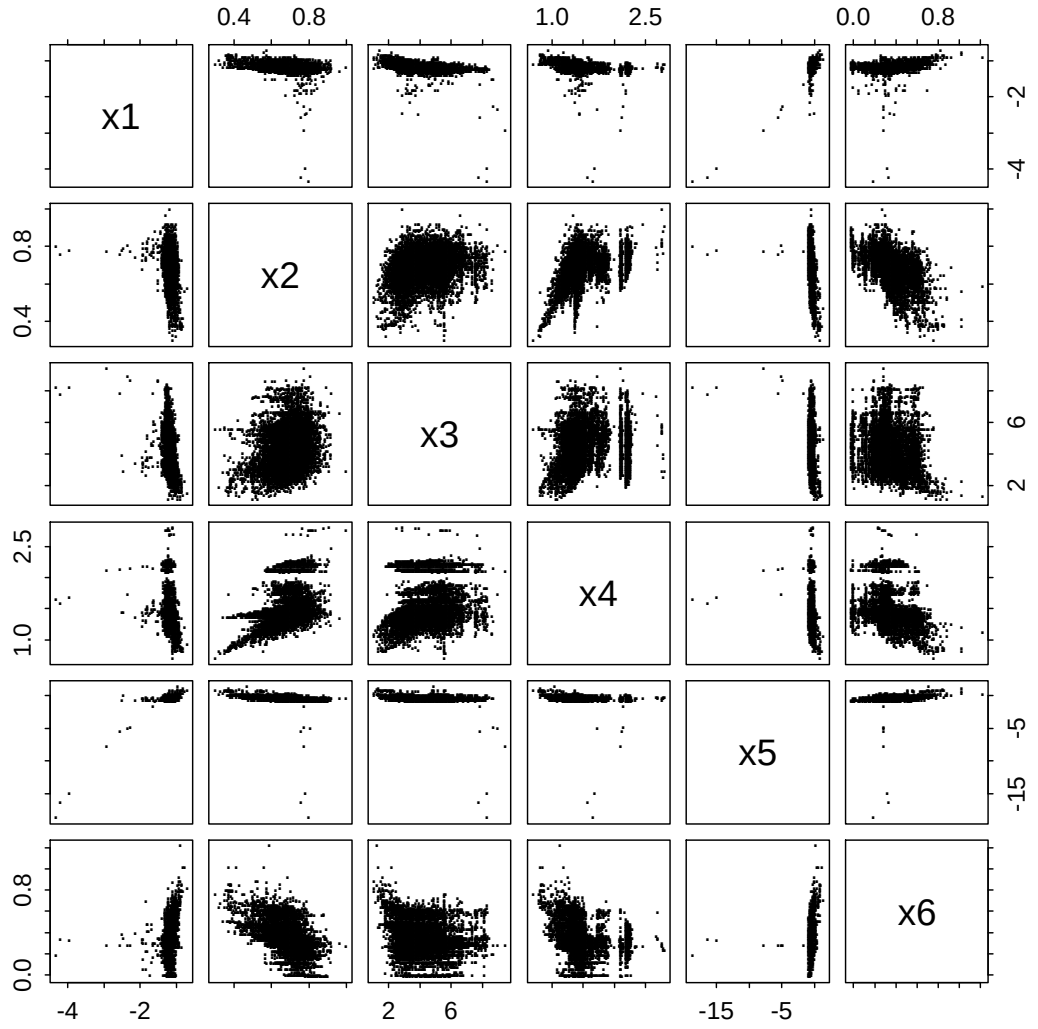
- Response: $y = 0/1$ inactive/active

	Total	Training data	Test data
Compounds	29,812	14,906	14,906
Active	608 (2%)	304	304
Inactive	29,204	14,602	14,602

- Descriptors (x's): 6 BCUT's

Use the training data to model and rank the probability of activity for the compounds in the test data.

NCI BCUT'S



Modelling Challenges

- Extreme imbalance of active/inactive compounds
- Multiple mechanisms for activity
- Highly nonlinear effects
- Systematic error in measured activity
- Quantifying a compound's structure via descriptor variables is difficult
- Some descriptor sets have highly correlated variables
- Compound collection is strongly clustered in descriptor space

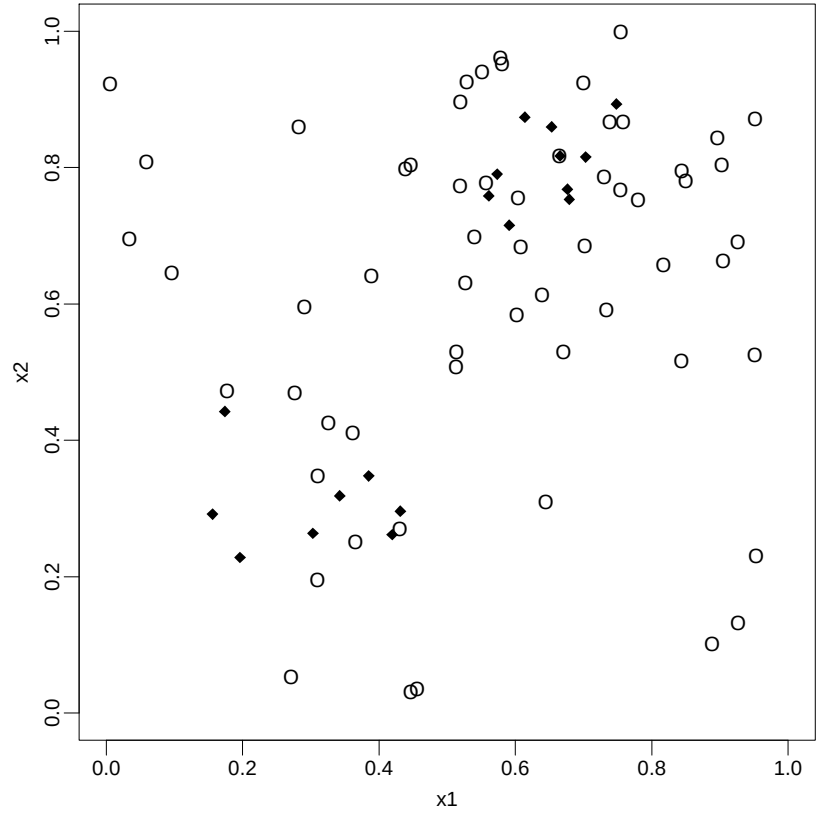
k-Nearest Neighbours (k-NN)

Extremely simple idea: Predict activity of a compound using the k nearest compounds in the training (assayed) data.

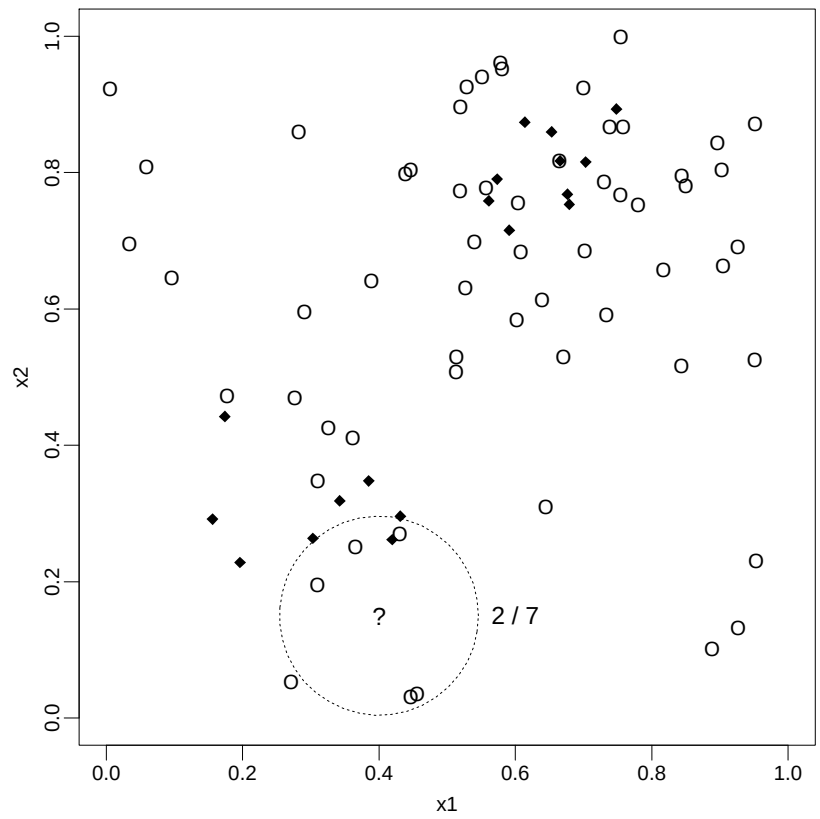
k chosen for best prediction accuracy by cross-validating the training data.

e.g., Two-Dimensional Descriptor Space

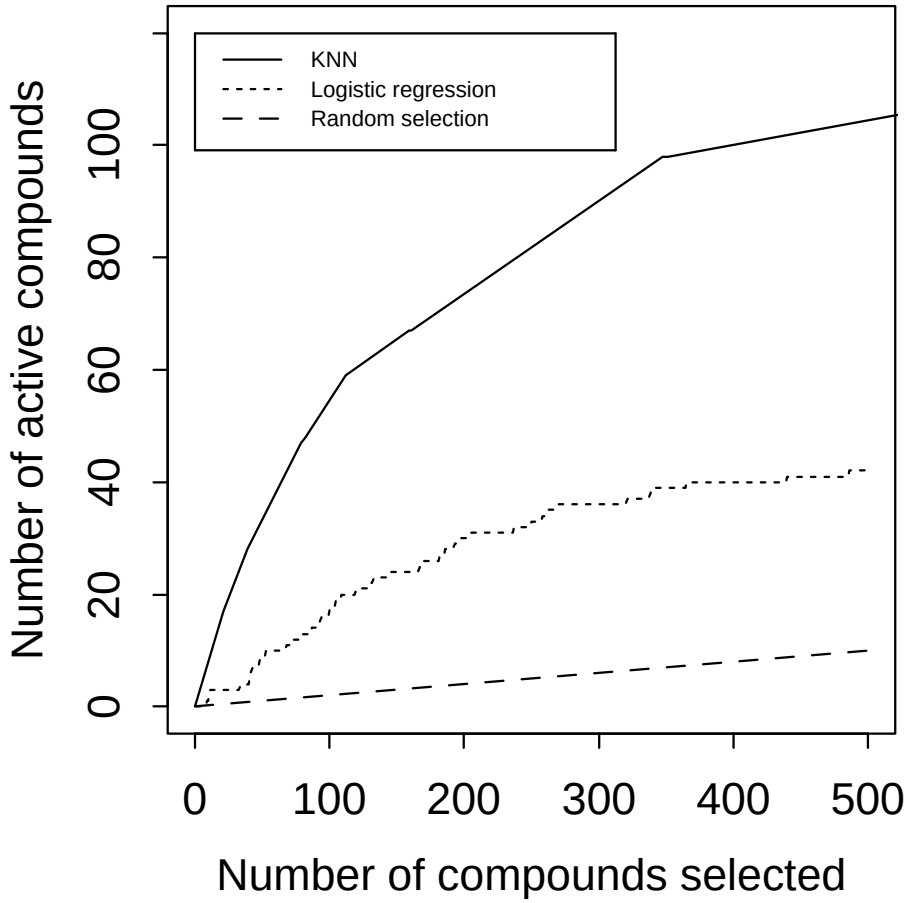
Inactives (Circles) and Actives (Diamonds)



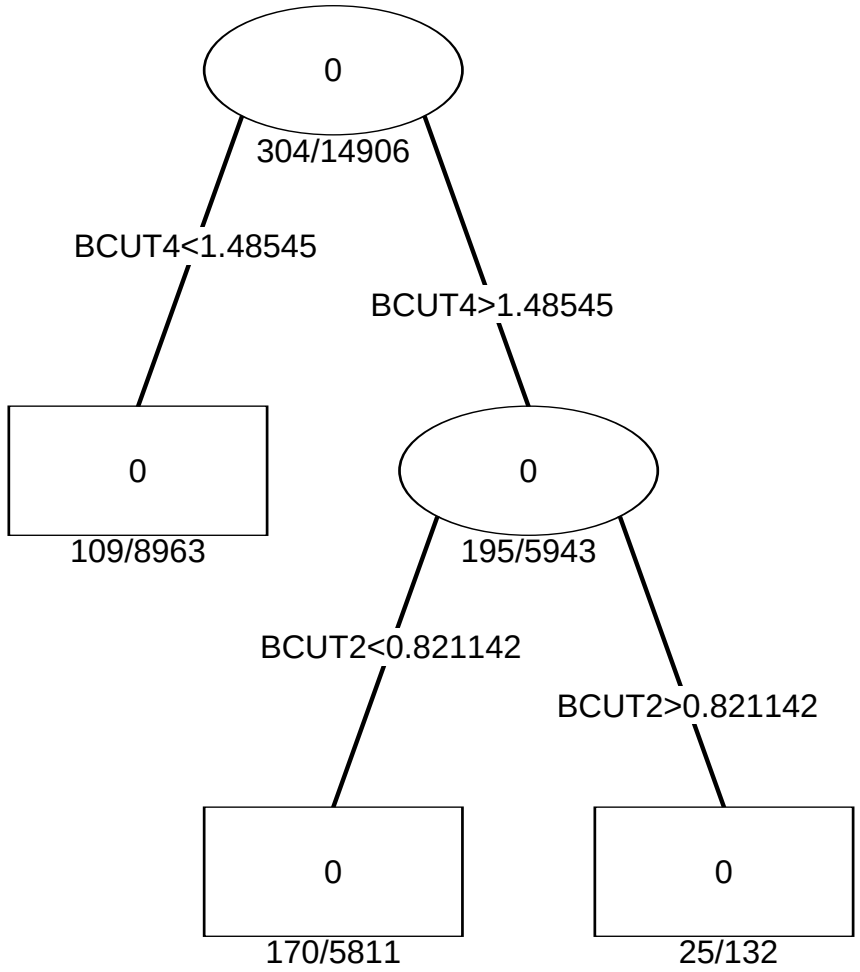
Predict at Test Point "?"



Performance Evaluation: e.g., k-NN for NCI Data



Classification and Regression Trees (CART)



(We use trees with 100's of terminal nodes.)

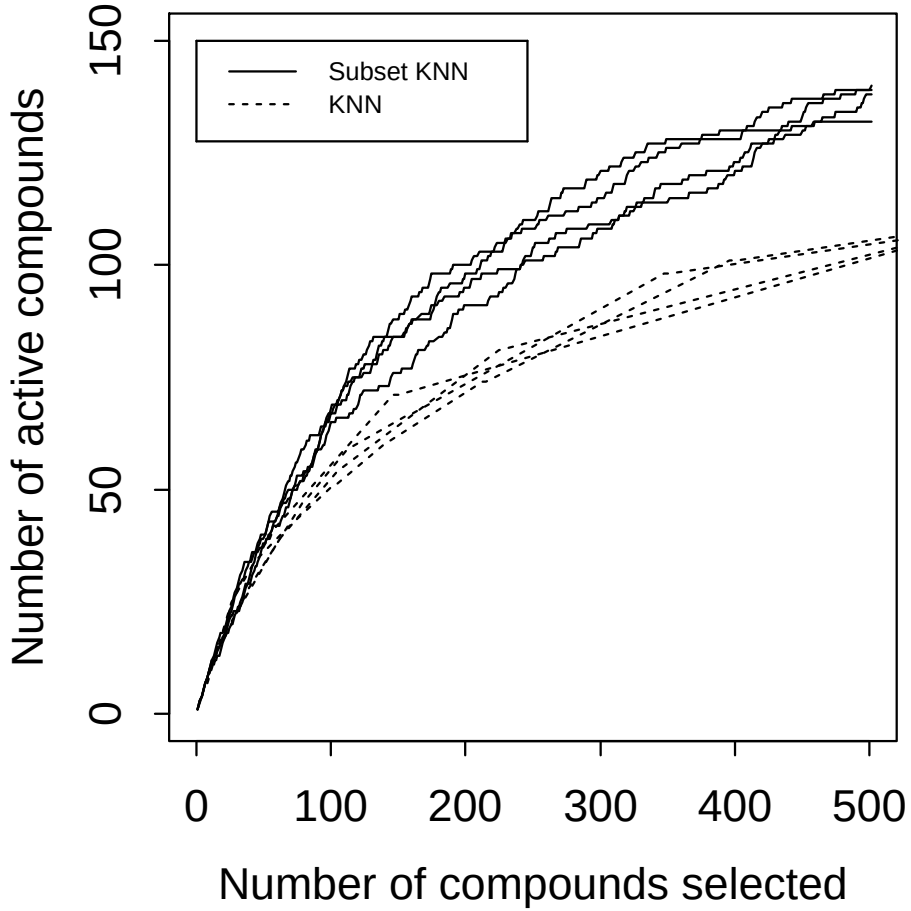
k-NN With Subset Averaging

Average k-NN predicted probabilities over subsets of descriptor variables.

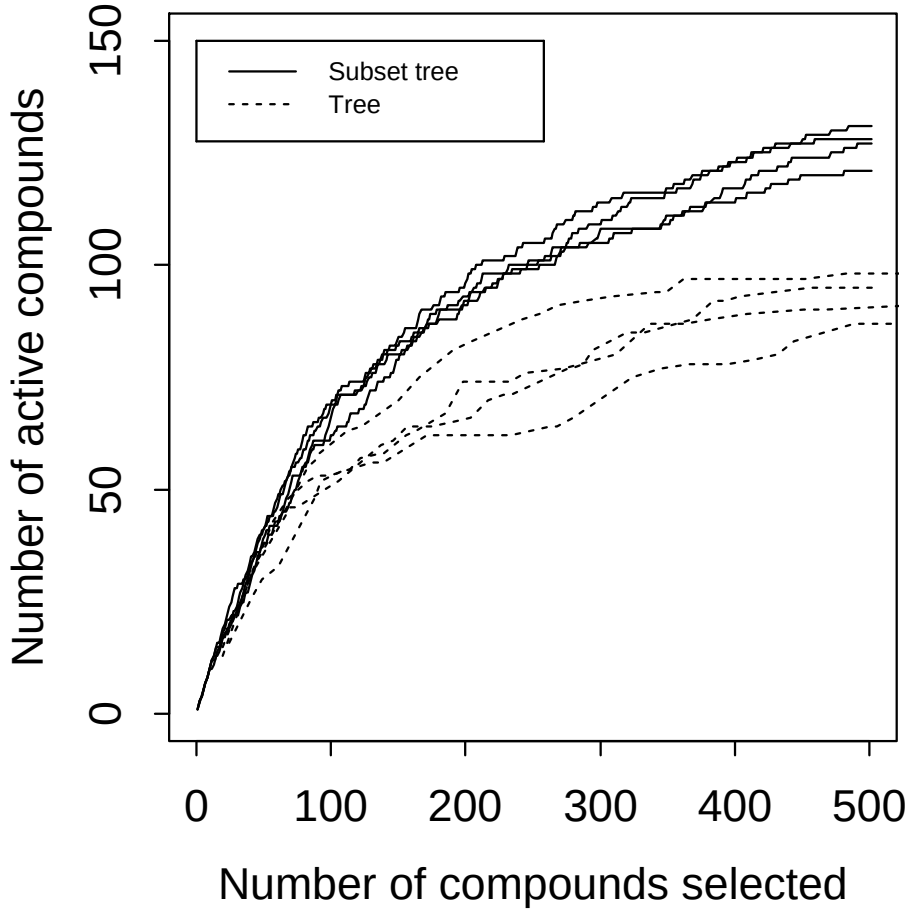
Variables used in k-NN		Predicted probability of activity for compound i
x_1	$\Rightarrow$	$\hat{p}_i^{(1)}$
$\vdots$	$\vdots$	$\vdots$
x_6	$\Rightarrow$	$\hat{p}_i^{(6)}$
x_1, x_2	$\Rightarrow$	$\hat{p}_i^{(12)}$
$\vdots$	$\vdots$	$\vdots$
x_5, x_6	$\Rightarrow$	$\hat{p}_i^{(56)}$
$\vdots$	$\vdots$	$\vdots$
$x_1, \ldots, x_6$	$\Rightarrow$	$\hat{p}_i^{(123456)}$
Average		$\hat{p}_i$

Same idea can be applied to trees.

Performance of Subset k-NN



Performance of Subset Tree



Other Averaging Methods: Bagging

Breiman (1996)

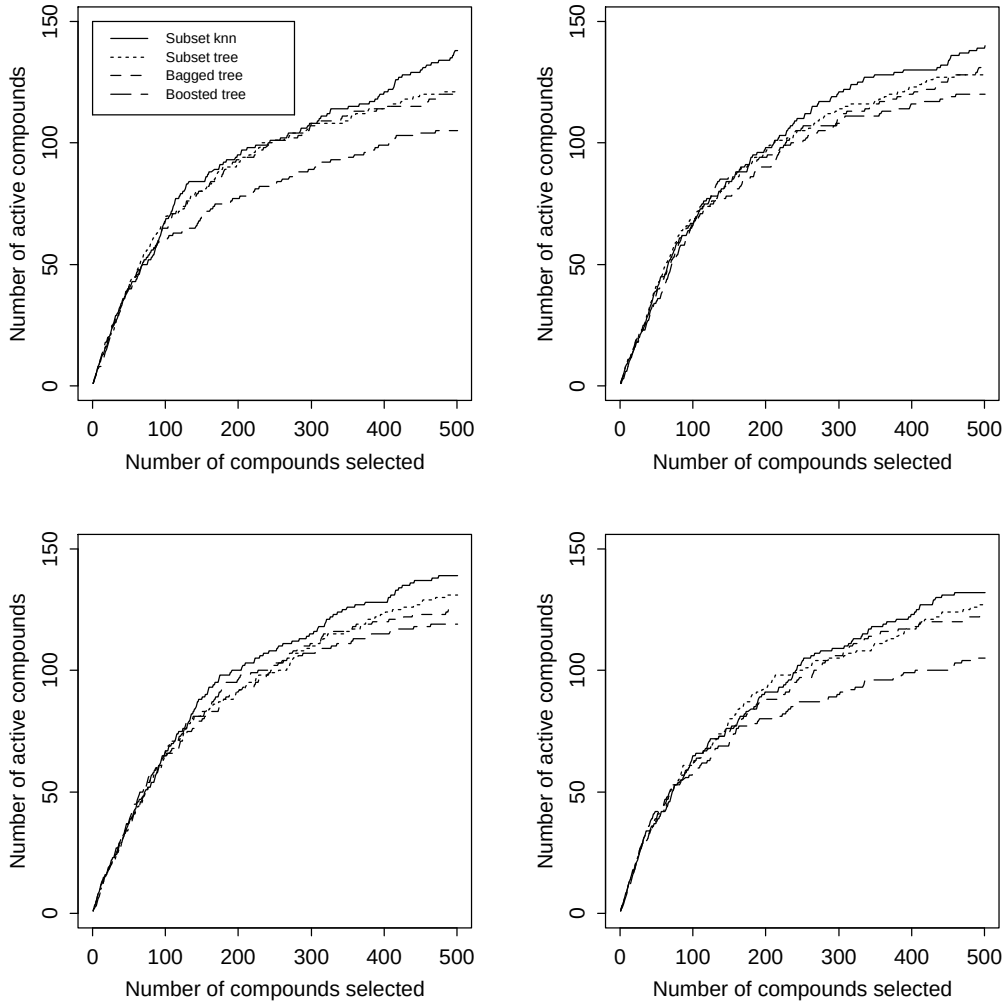
- Generate say 100 bootstrap samples from the training data.
- For each sample:
  - Build tree (or perform k-NN)
  - Get estimated probabilities of activity for the test compounds.
- For each test compound:
  - Average its 100 estimated probabilities of activity.

Other Averaging Methods: Boosted Trees

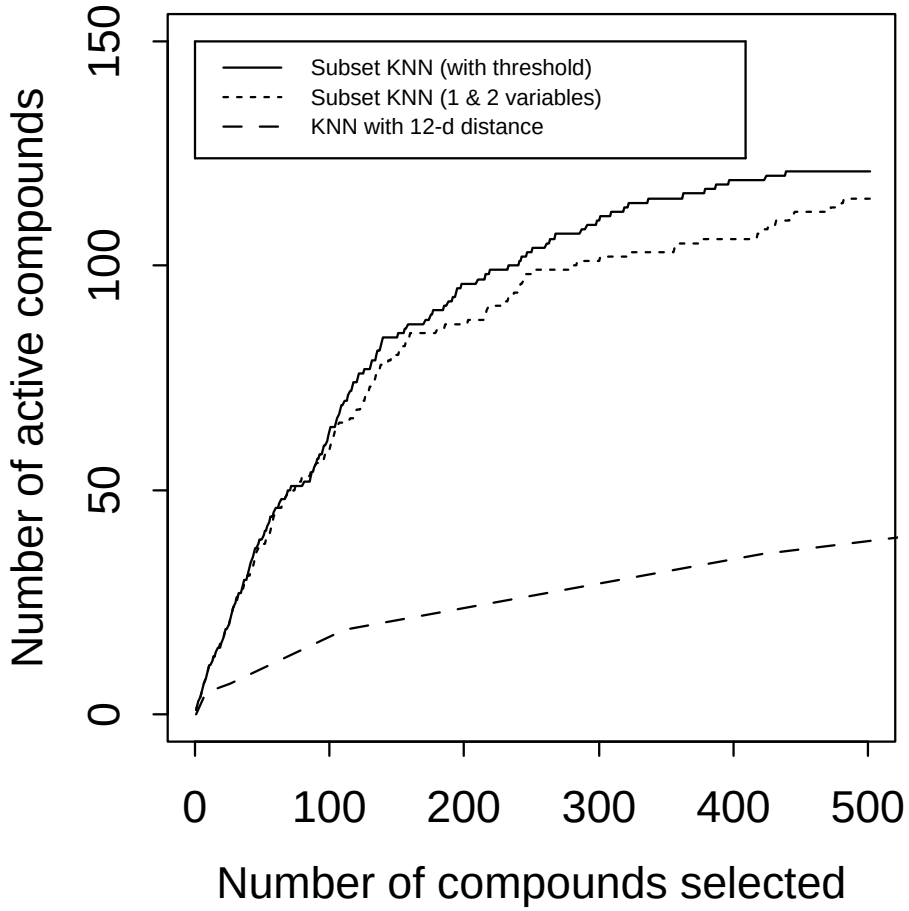
Schapire (1990), Friedman et al (2000), Friedman (2001)

- All training compound weights are equal.
- Build a (small) tree, $\mathcal{T}$ (unweighted) from the training data.
- Iterate for say 1000 trees:
  - Increase the weights of the compounds misclassified by $\mathcal{T}$.
  - Build a weighted tree, $\mathcal{T}$.
- For each test compound:
  - Average its 1000 estimated probabilities of activity from the 1000 trees (average gives more weight to trees with lower misclassification rate).

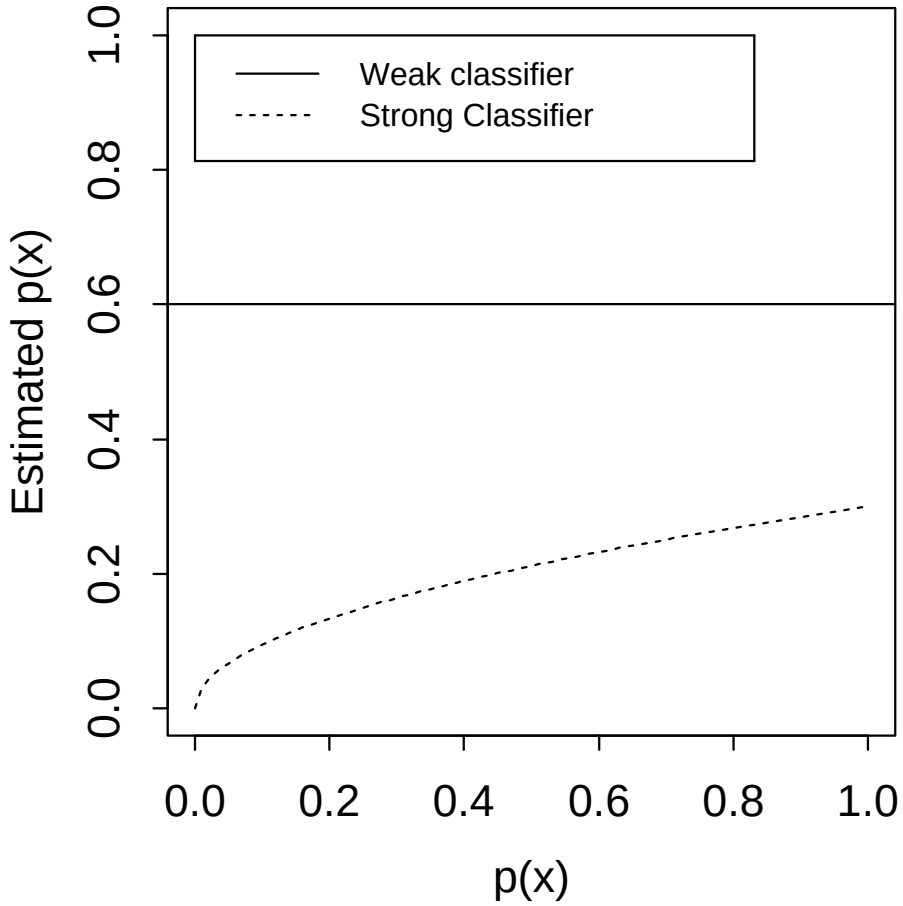
Subset Averaging Versus Bagging and Boosting



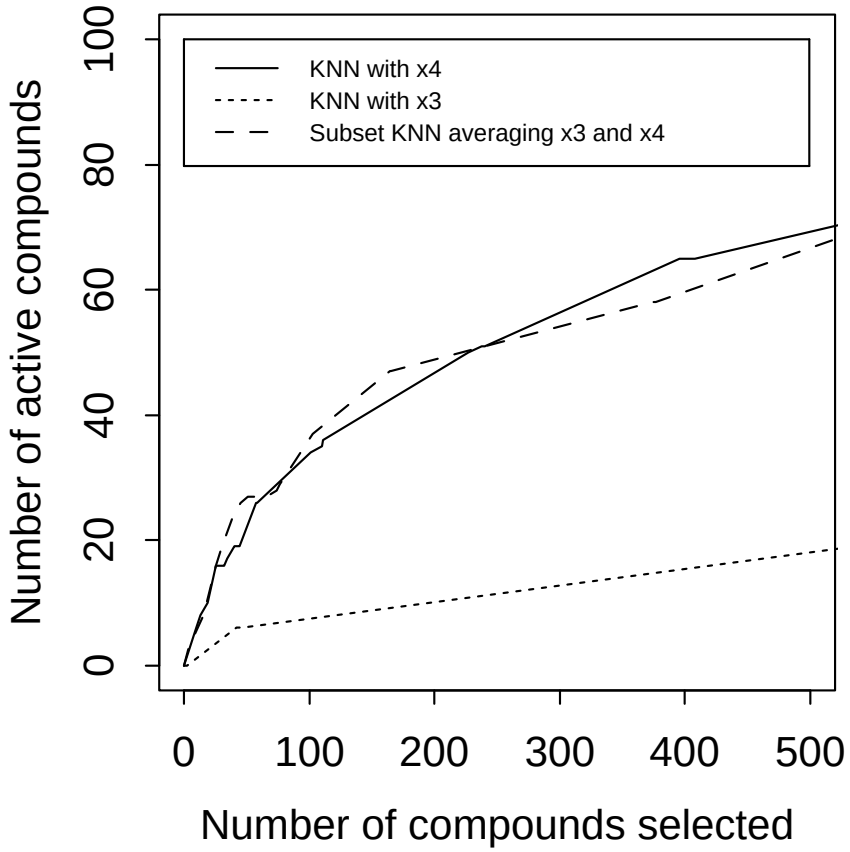
Curse of Dimensionality: Add Six "Noise" Descriptors



Why? Subset Averaging and *Ranking*



Why? NCI Data



Selecting Variables: Average Hit Rate

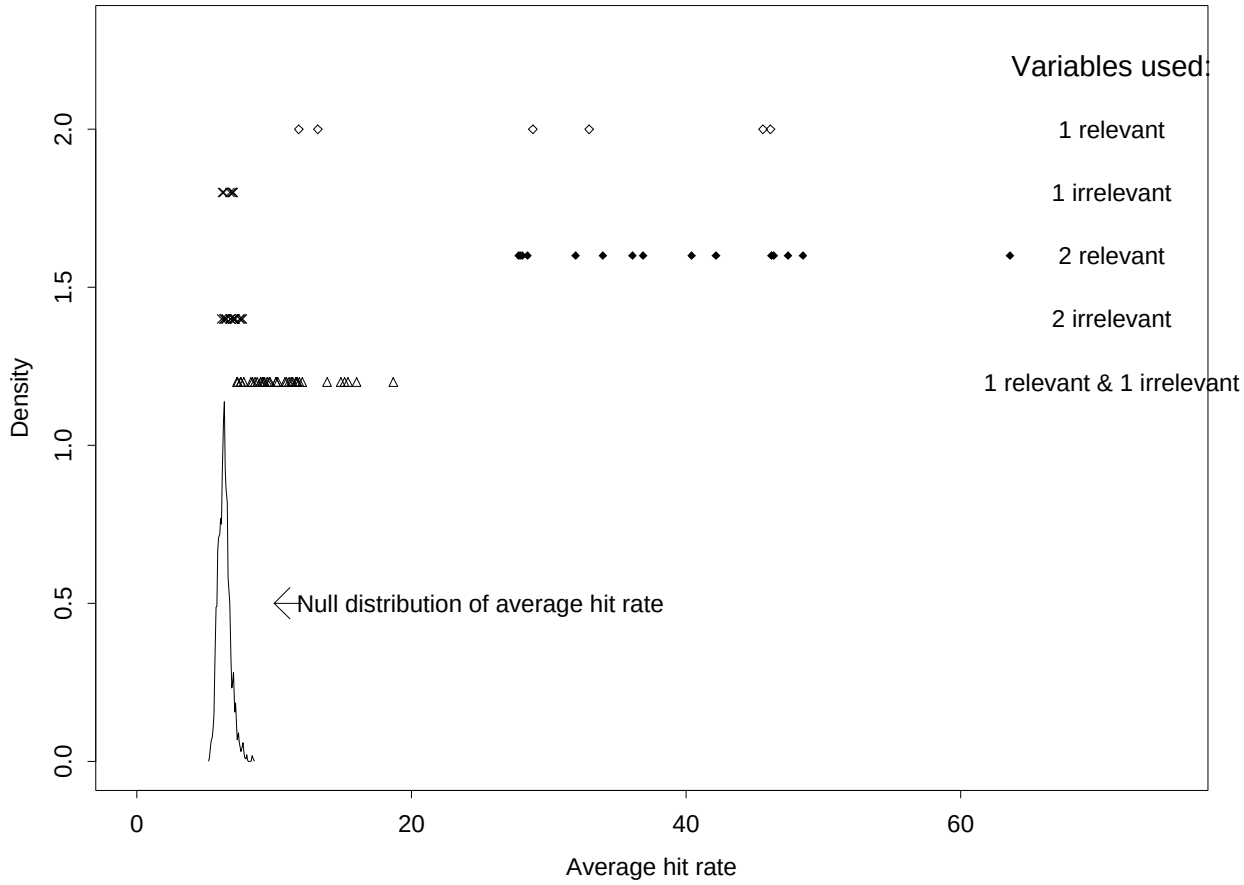
e.g, 5 test compounds are ranked in terms of estimated probability of a hit.

We find they are:
Active, Inactive, Active, Inactive, Active

Average the hit rate at points on the hit curve where each active is found:

$$\text{AHR} = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5}\right) = 0.76.$$

Selecting Variables: NCI (12 Variables)



Conclusions

- Drug discovery data are hard to model.
- Local methods work well (this is consistent with Ray Lam's thesis work).
- k-NN and trees perform well.
- Averaging helps performance.
- Subset averaging performs best here (for ranking) and is easiest to interpret.