Strata+Hadoop WORLD

MAKE DATA WORK
MARCH 28–29, 2016: TRAINING
MARCH 29–31, 2016: CONFERENCE
SAN JOSE, CA

Validating models in R

- Tuesday, 03/29/2016 11:00 AM – 12:30 PM Location: LL20C

- Nina Zumel and John Mount, Win-Vector LLC.

- We demonstrate a number of techniques, R packages, and R code for validating predictive models, using example code, data, and live demonstrations and exercises. Learn how to determine if there is usable signal in your data, select variables, and choose models using R and R graphics (ggplot2). Increase your statistical efficiency and squeeze more signal out of your data.

- Materials: https://github.com/WinVector/ValidatingModelsInR

Win-Vector LLC

“Essentially, all models are wrong, but some are useful.”

– George Box

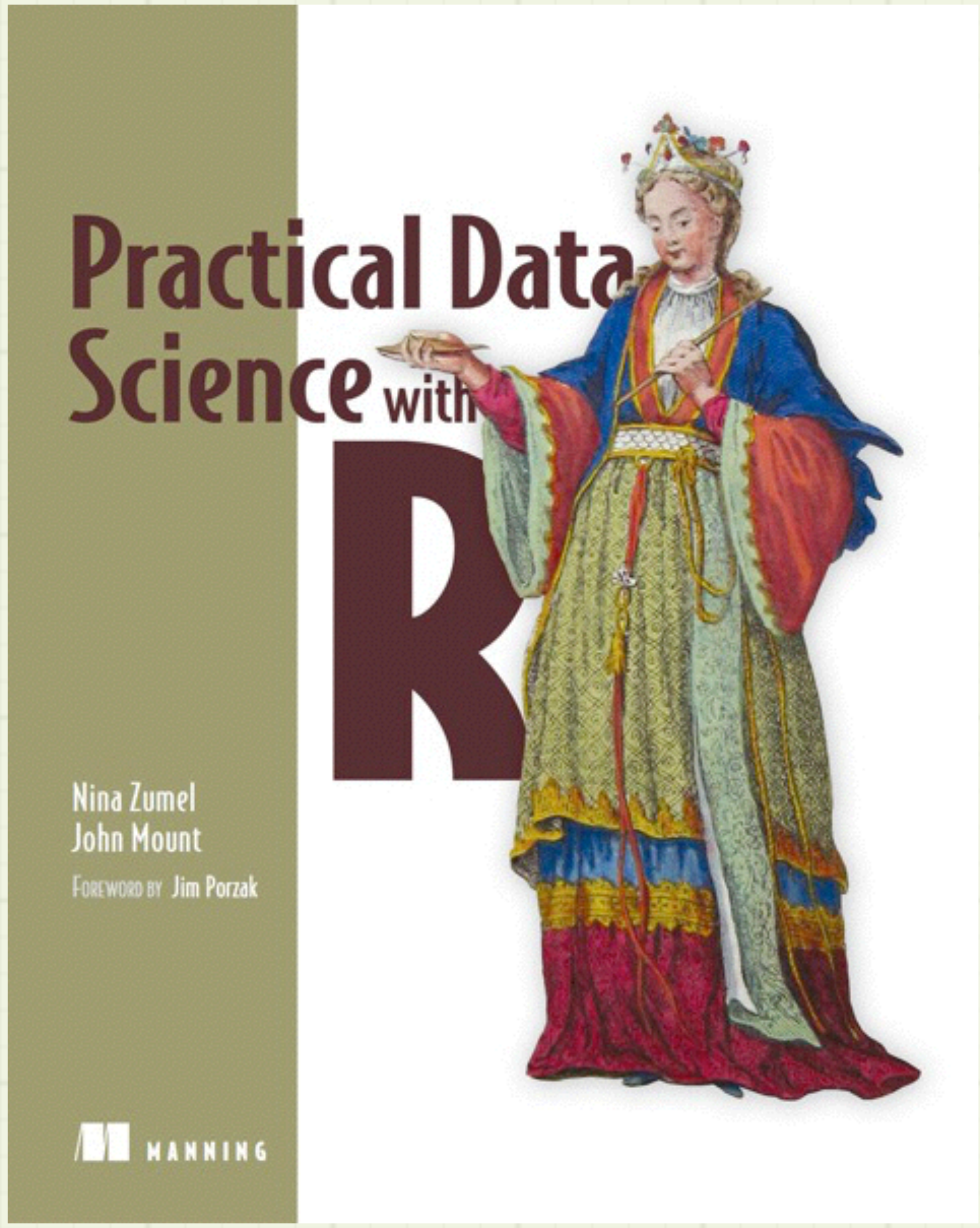
Goals of this Tutorial

- Give you a sophisticated tool box of model quality measures that are:
    - Statistically well founded.
    - Business motivated!
    - Organized by a taxonomy of needs.
    - With ready to go R code and graphs.

Biography

Nina Zumel
Win-Vector LLC
Dr. **Nina Zumel** is a principal consultant and founder at Win-Vector LLC a San Francisco data science consultancy and training company. Nina started her advanced education with an EE degree from UC Berkeley and holds a Ph.D. in Robotics from Carnegie Mellon University. Nina has worked as research scientist as SRI and developed revenue optimization platforms. She frequently writes and speaks on statistics and machine learning.

Nina is also the coauthor of the popular book of *Practical Data Science with R* (Manning Publications, 2014).



John Mount
Win Vector LLC
Dr. **John Mount** is a principal consultant and founder at Win-Vector LLC a San Francisco data science consultancy and training company. John has worked as a computational scientist in biotechnology and a stock-trading algorithm designer and has managed a research team for Shopping.com (now an eBay company). John started his advanced education in mathematics at UC Berkeley and holds a PhD in computer science from Carnegie Mellon.

John is also the coauthor of *Practical Data Science with R* (Manning Publications, 2014).

Please contact contact@win-vector.com for projects and collaborations. **http://win-vector.com/** Twitter: @WinVectorLLC .

Win-Vector LLC

Why are we giving a statistics talk during R Day of a data science conference?

- With R classic statistical advice becomes immediately actionable.
- With "big data" you can remove unwanted assumptions and directly estimate model quality.
- If you see something new we can influence how you work. If nothing is new, maybe we can influence how you teach.
- A *great* excuse to use R Markdown, ggplot2, and other packages.

Win-Vector LLC

The two issues

- Choosing a convincing measure.
    - Part 1 of this session.
- Confirming you have a decisive measurement.
    - Part 2 of this session.

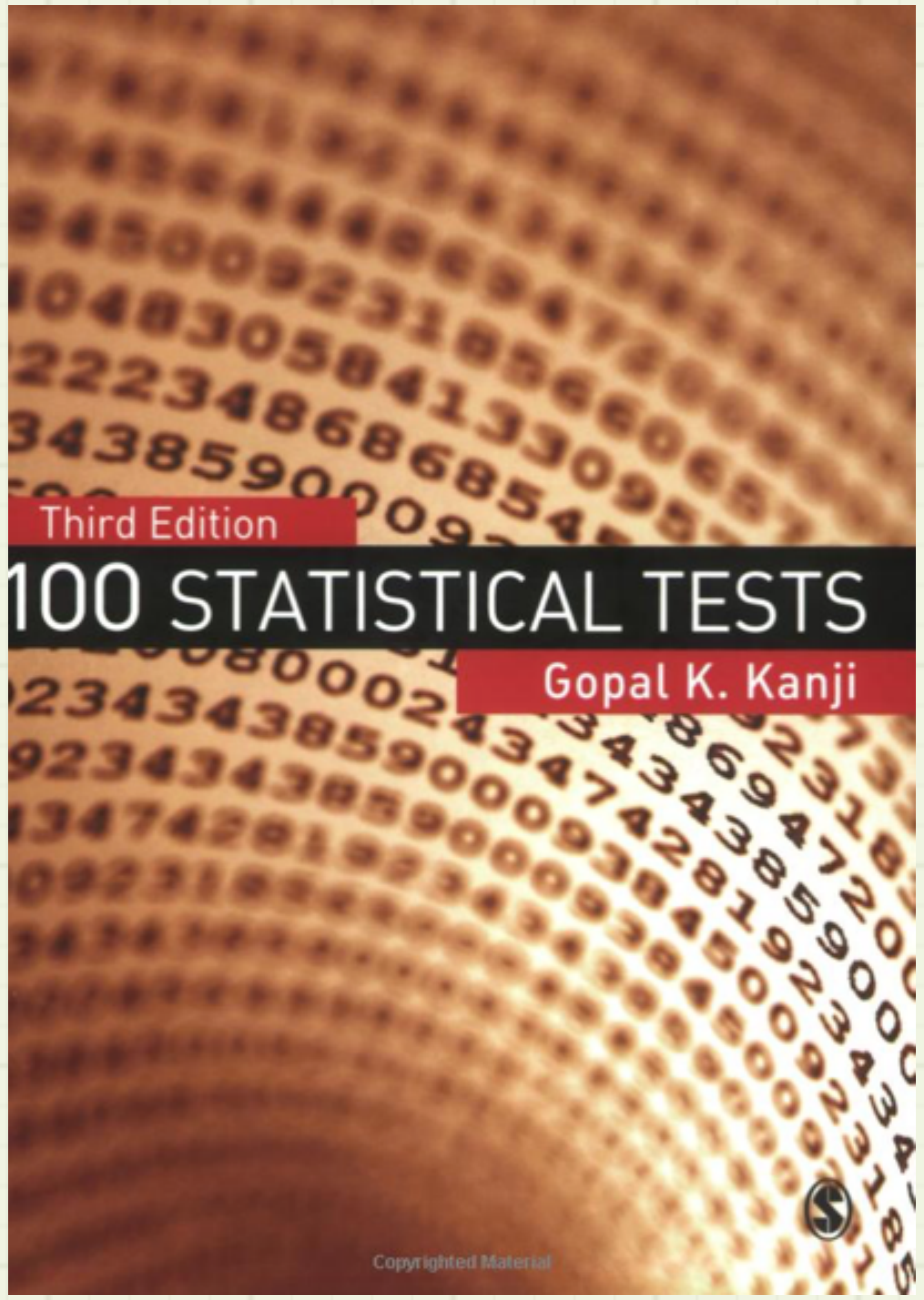
Things are simpler when you

Start here

- Separate what to measure from how to confirm significance or estimate posteriors.

- You rely on a programing environment for composable, reusable methods, simulation and visualization.

Not here

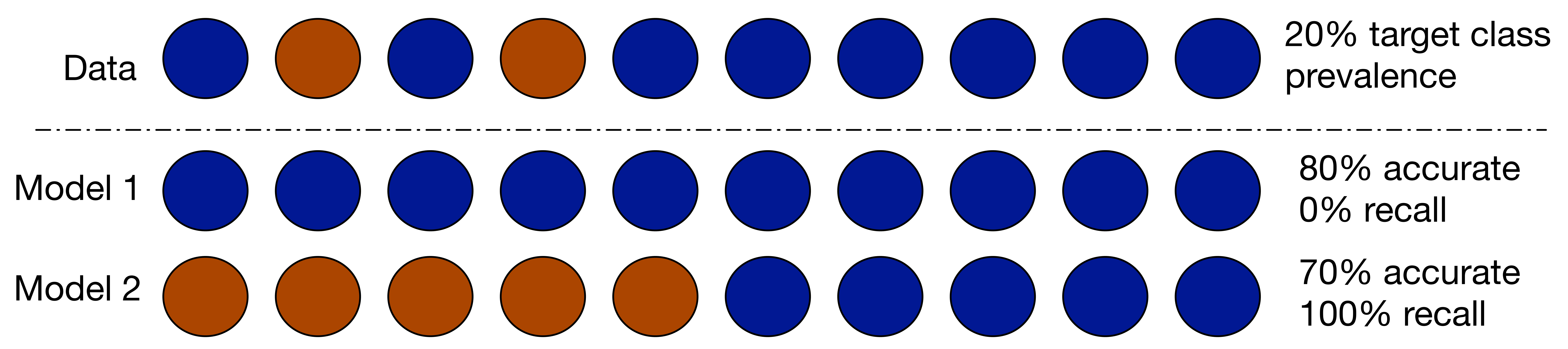


(reserve this to improve solutions later)

What is a Good Model?
Performance Metrics

How do you measure model performance?

- Hint: accuracy is not usually the right way



How are you going to use your model?

- Decision Procedure
  - Predict customer churn; predict home sales price
  - Correct answers are important
- Sorting or prioritization
  - Target at-risk customers for intervention; identify most valuable homes
  - Correct comparisons are important

Metrics for Classifiers

Technical Metrics

- AUC (ROC), deviance
- Good metrics for data scientists and between data scientists
- Useful proxy measures for comparing candidate models
- Not always easily translatable to business goals

Domain Metrics

- Precision, Recall, Sensitivity, Specificity
- Good metrics for business

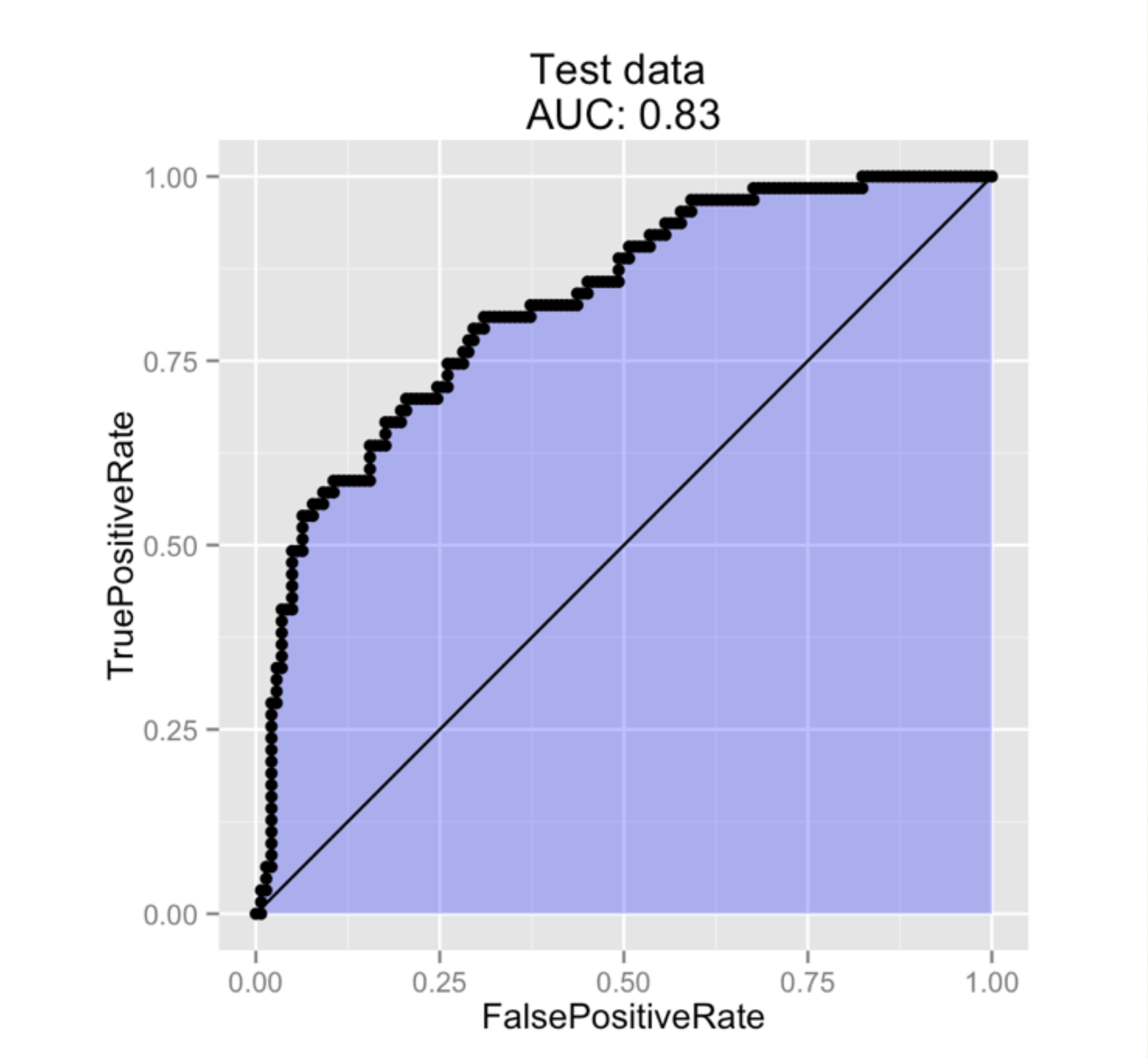
ROC/AUC

- Trade-off between true positive and false positive rates as labeling threshold T is varied

- AUC: area under the curve
  - Probability that a randomly chosen positive example will score higher than a randomly chosen negative example (with appropriate tie-breaking).

- Invariant to monotonic transformations of scoring function

- Independent of target class prevalence



Win-Vector LLC

Deviance

- Deviance penalizes “mis-matches” between prediction and true class label
- Analogous to variance
  - smaller is better
- Dependent on dataset size
  - Don’t compare unnormalized deviance across evaluation sets of different size

deviance = -2*[sum(y*log(py)+(1-y)*log(1-py))]

y=1 and P(class = 1) -> 1 contribution -> 1 * log(1) -> 0

y=0 and P(class = 1) -> 0 contribution -> (1-0) * log(1-0) -> 0

y=1 and P(class = 1) -> 0 contribution -> 1 * log(0) -> large negative value

y=0 and P(class = 1) -> 1 contribution -> (1-0) * log(1-1) -> large negative value

When Correct Answers are Important

- Confusion matrix
- Recall/Precision
- Sensitivity/Specificity
- Pricing of False Positives, False Negatives

Confusion Matrix

False Positives

	Prediction FALSE	Prediction TRUE	
Diabetic FALSE	434	66	500
Diabetic TRUE	110	158	268
	544	224	768

Outcome sums

Prediction sums

False Negatives

Win-Vector LLC

Confusion matrix is paramount

- Most common classifier score follow from:
- The confusion matrix
- Or scaled summaries of it such as
  - tpr, fpr, tnr, fnr

False Positive Rate, False Negative Rate:
False Positive Rate

$$\frac{\text{False Positives}}{\text{All Negatives}}$$

66/(434+66)
= 0.132

“The fraction of non-diabetics misdiagnosed as diabetic.”

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

False Positives

All Negatives

Win-Vector LLC

False Positive Rate, False Negative Rate: False Negative Rate

False Negatives

———————————-

All Positives

110/(110+158)
= 0.41

"The fraction of diabetics misdiagnosed as non-diabetic."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

False Negatives

All Positives

True Positive Rate, True Negative Rate: True Positive Rate

True Positives
———————————-
All Positives

158/(158+110)
= 0.589

"The fraction of diabetics correctly identified as such."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Positives

All Positives

True Positive Rate, True Negative Rate: True Negative Rate

$$\frac{\text{True Negative}}{\text{All Negatives}}$$

434/(434+66)
= 0.868

"The fraction of non-diabetics correctly identified as such."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Negatives

All Negatives

What about accuracy?

- Almost all business partners will ask for “accuracy.”
- This is only because it is likely the only score that has been explained to them in any detail.

Accuracy

∑diagonals/∑entries

(434+158)/768
= 0.77

"The fraction of patients correctly diagnosed."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

Common derived measures

Sensitivity, Specificity: Sensitivity

$$\frac{\text{True Positives}}{\text{All Positives}}$$

158/(158+110)
= 0.589

"The fraction of diabetics correctly identified as such."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Positives

All Positives

Win-Vector LLC

Sensitivity, Specificity: Specificity

$$\frac{\text{True Negatives}}{\text{All Negatives}}$$

434/(434+66)
= 0.868

"The fraction of non-diabetics correctly identified as such."
(Or 1-FPR)

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Negatives

All Negatives

Precision, Recall: Precision

$$\frac{\text{True Positives}}{\text{Predicted Positives}}$$

158/(158+66)
= 0.705

> "The fraction of patients diagnosed as diabetic who really are."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Positives

Predicted Positives

Win-Vector LLC

Precision, Recall: Recall

True Positives
———————————
All Positives

158/(158+110)
= 0.589

> "The fraction of diabetics correctly identified as such."

Diabetic \ Prediction	FALSE	TRUE
FALSE	434	66
TRUE	110	158

True Positives

All Positives

labs/ Lab01ScoringClassifiers

Bringing it all together

Which Metrics Are Appropriate?

Question	Metric	Example
Is it important that a positive classification is correct?	Precision	If the test comes back positive, is the patient really diabetic?
Is it important we find all positive cases?	Recall Sensitivity	Do we miss any diabetics through this test?
Are false positives expensive?	Precision Specificity	Diagnoses that lead to costly treatment
Are false negatives expensive?	Recall Sensitivity	Diagnosing conditions that are costly if untreated
Is it important to get everything right?	Accuracy	

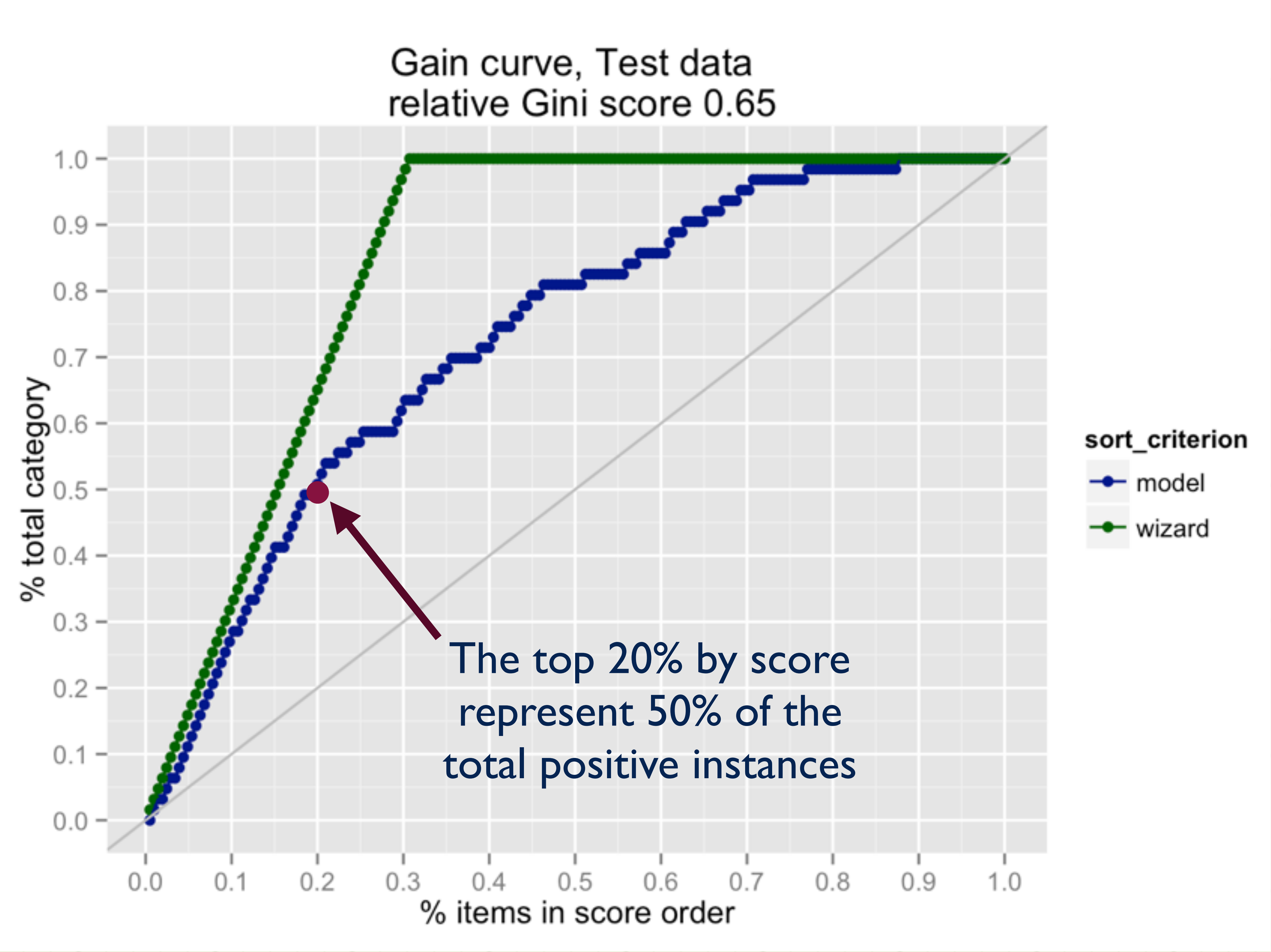
When Sorting is Important

- Gain Curve
  - Applies for both probability models and general regression models
- Gini Coefficient

Win-Vector LLC

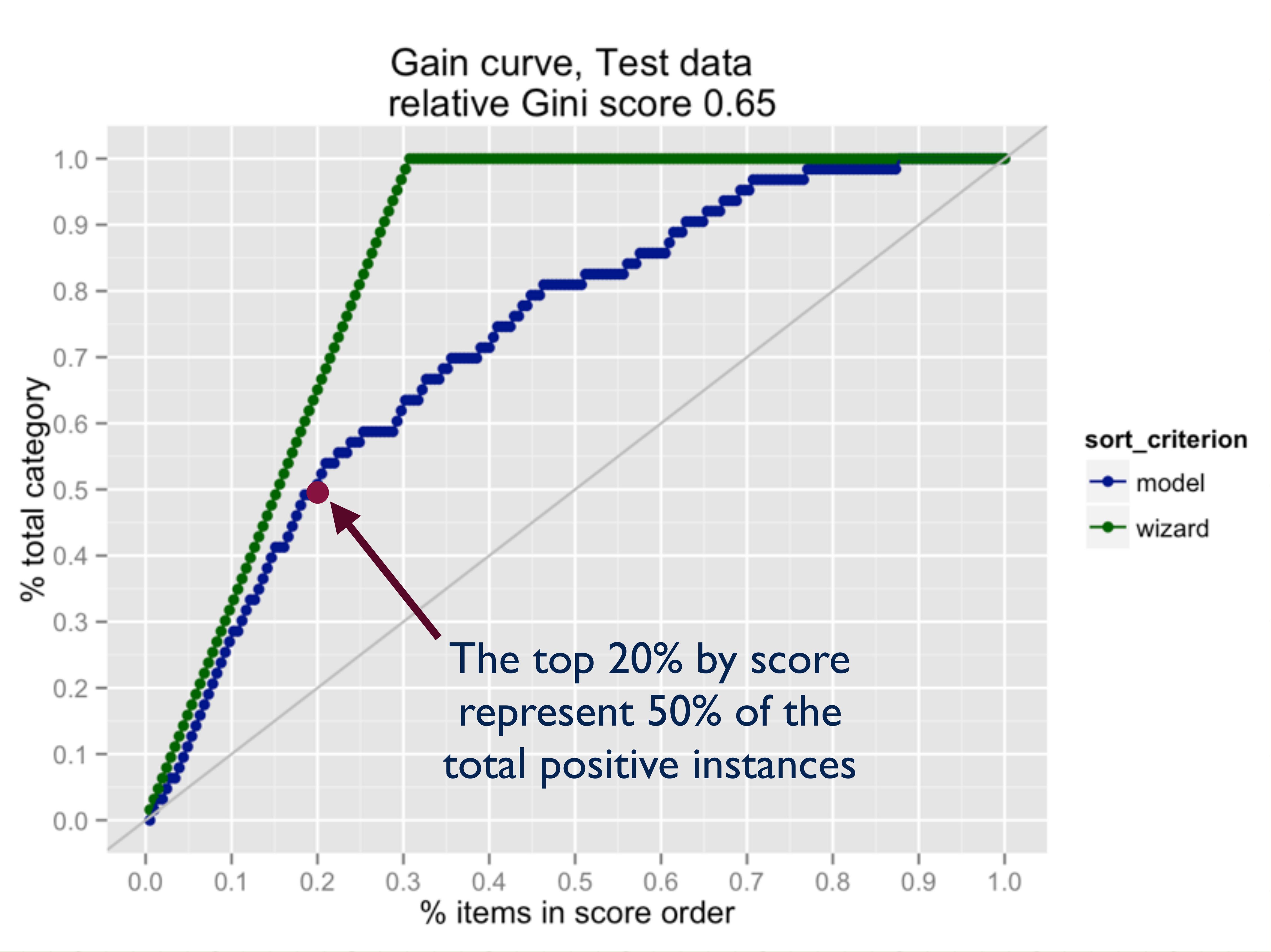
Gain/Lift Curve

- Sort by probability score, descending
- Measures how quickly you identify all positive instances (by sort order)
  - Good when score is used for prioritizing items to act upon
    - E.g. Probability of fraud



Gain Curve (cont.)

- Gini score: 2 * area between the curve and x=y line
- Relative Gini score: ratio of model Gini to ideal Gini score
  - “Wizard” in plot
- **Gain Curve is sensitive to target class prevalence**



Win-Vector LLC

labs/Lab02GainCurve

Metrics for Regression

Technical Metrics

- R^2
  - $1 - \sum(yi - pi)2/\sum(yi - E[y])2$
  - Fraction of variance “explained” by the model
- Again, useful for comparing candidate models, not always a clear connection to business goals

Win-Vector LLC

Domain Metrics

- Root mean squared error
- Mean absolute error
- Mean relative error

Root Mean Squared Error

$$\mathrm{rmse}(y,\ p) = \sqrt{(\textstyle\sum(y_i - p_i)^2/n)}$$

- Estimates the "average" error between outcome and prediction
    - Large differences dominate score

- Related to what linear regression minimizes to fit a model
    - Optimizing RMSE gets expectations and totals correct

Mean Absolute Error

$$\text{mae}(y,\ p) = \left(\sum |y_i - p_i| / n\right)$$

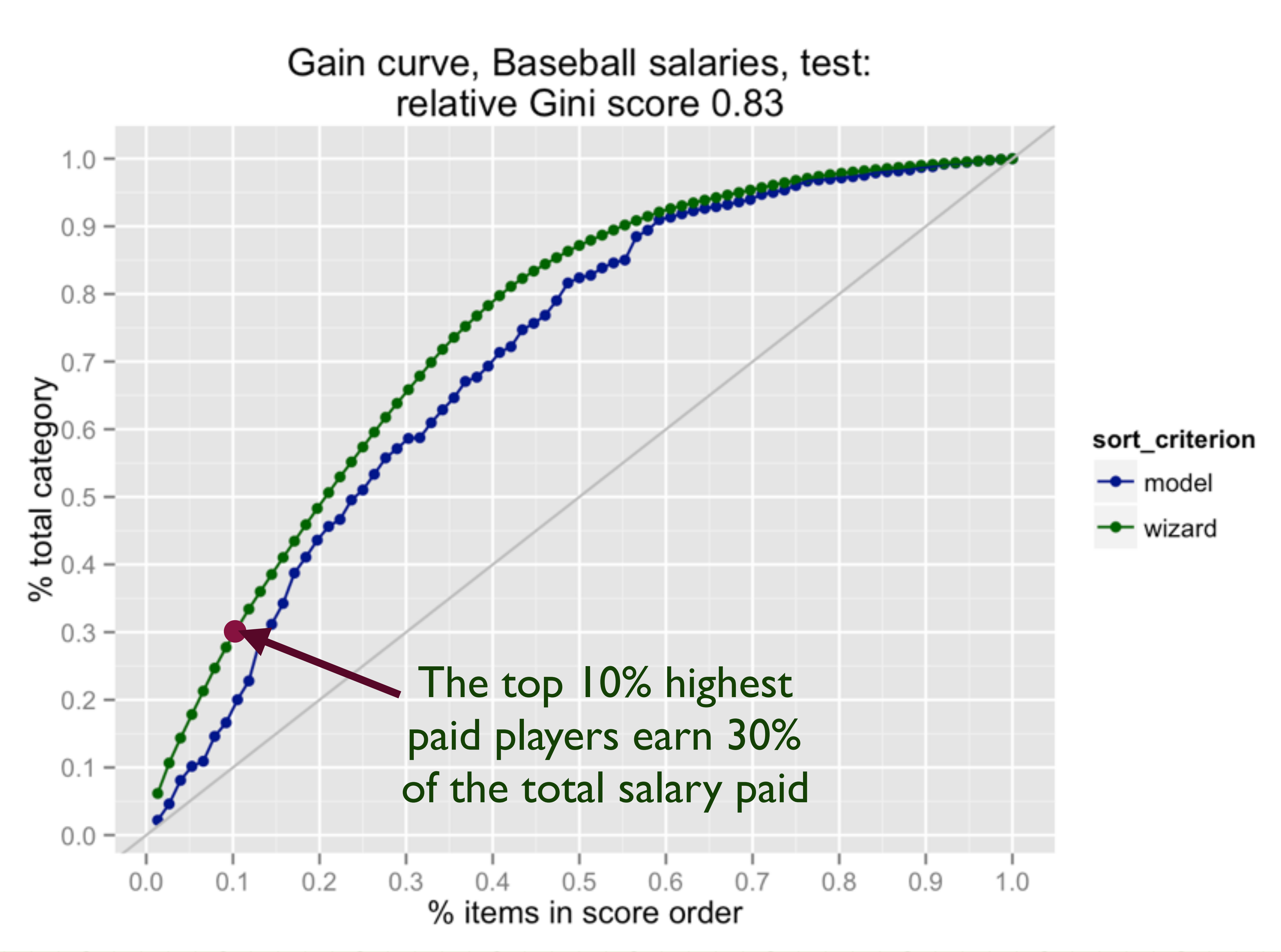
- Estimates the average unsigned error between outcome and prediction
  - Large differences dominate less
  - Arguably what people intuitively consider “average” error
- Optimizing MAE does *not* get expectations and totals correct

Minimizing Relative Error

- For example, by predicting log(y)
- Useful when outcome spans several orders of magnitude
  - $5 error on $1,000 different from $5 error on $10
- Errors on small magnitude outcomes dominate

When Order is Important: Gain Curve Revisited

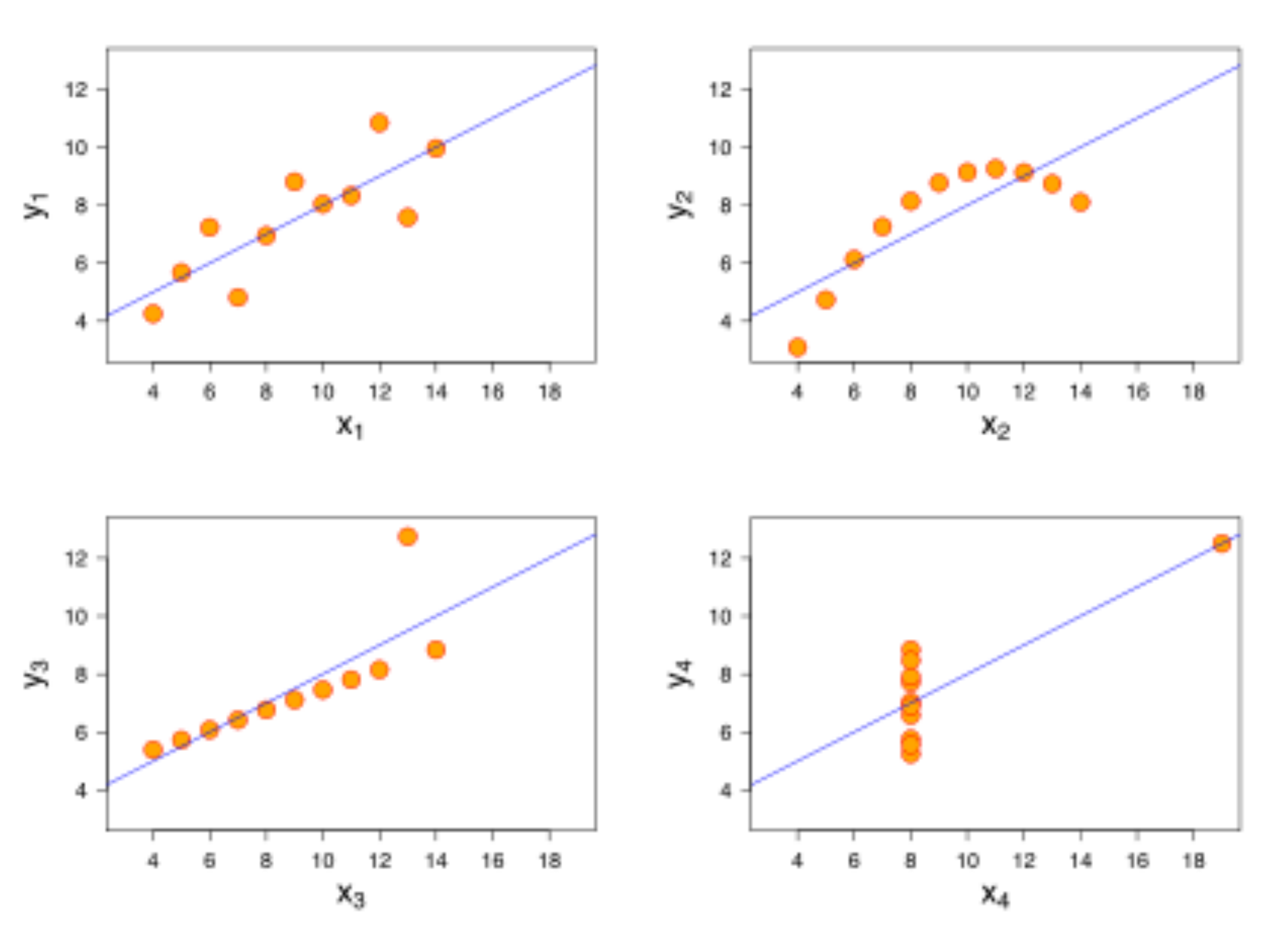
- Sort items by predicted value, descending
- Measure accumulated true value as a fraction of total
- Measures if your predictions are roughly in the right order
    - Do you predict large values as large, and small values as small?
    - Random sort: x=y line



Summaries can be deceptive

- Regression: Anscombe's quartet

- Ranking: distraction of indistinguishable pairs



Win-Vector LLC

labs/Lab03RankingIssues

Which Metrics Are Appropriate?

Question	Metric	Example
Must we predict the value accurately?	RMSE, mean absolute error	Predicting home sale price
Must we predict the value to a good relative tolerance?	Mean relative error	Predicting home sale price to within 10%
Do we mostly need the order to be right?	Relative Gini Score	Predicting lifetime customer value (Identify most valuable customers)

Speaker Change

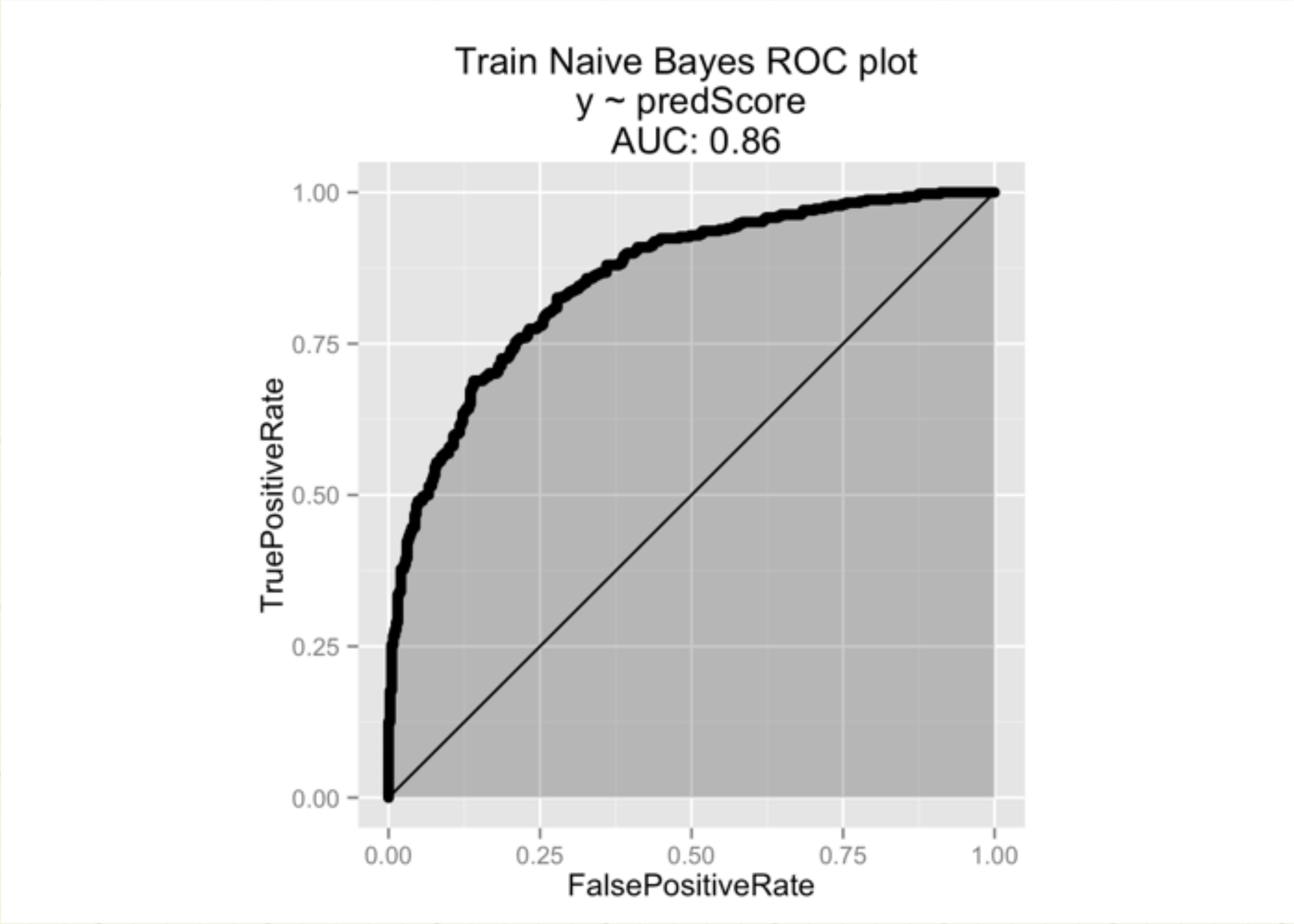
Is it *really* a Good Model? Estimating *out of sample* performance

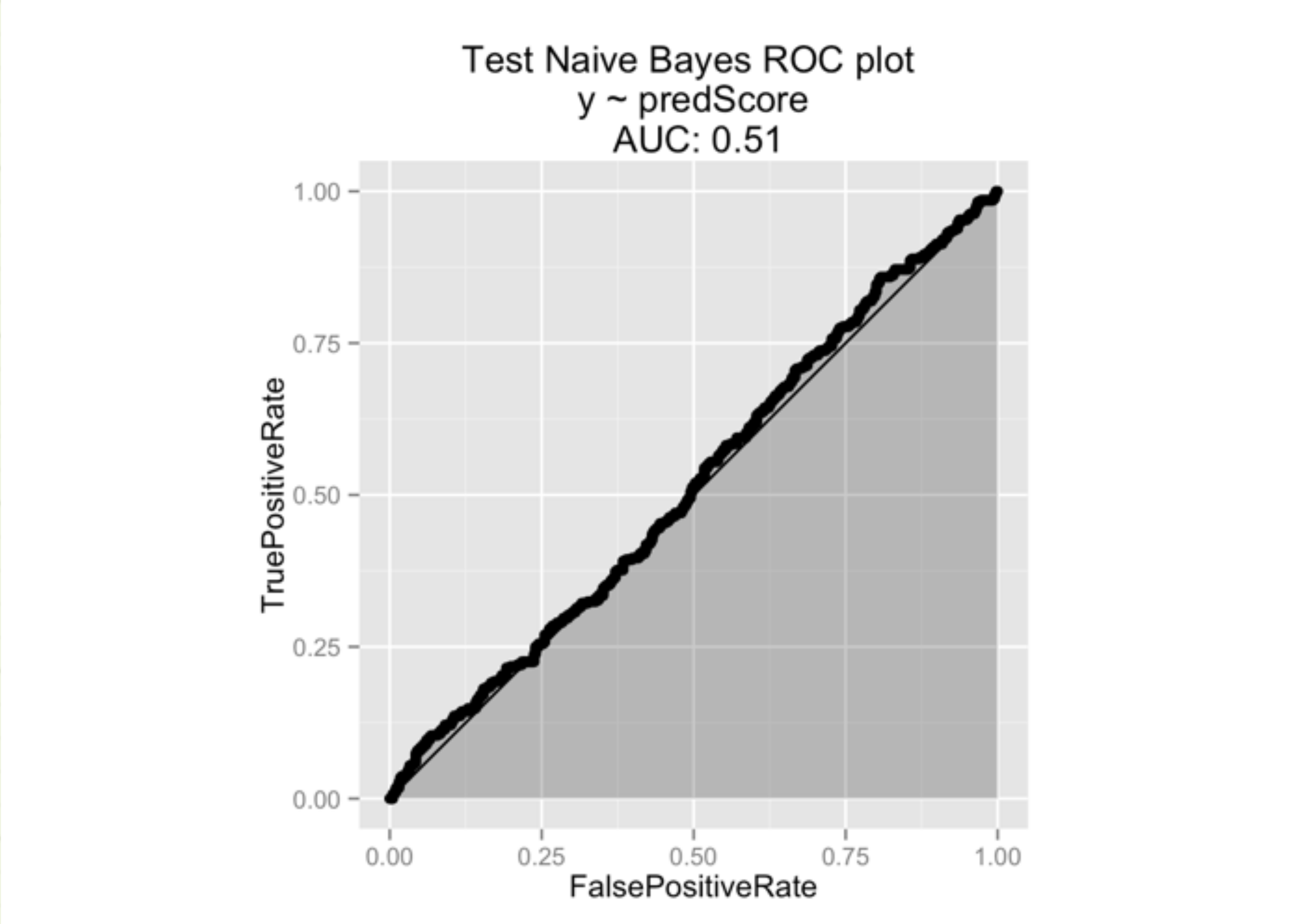
Principles

- Evaluation is comparison: distinguish good models from bad (or less good).
- Throughout: Estimate out-of-sample performance by repartitioning training data.
- Treat variable selection as choosing models by building throw-away single variable models.
- Work in probability units (not in effect sizes) wherever practical.

Are you predicting *anything* at all?

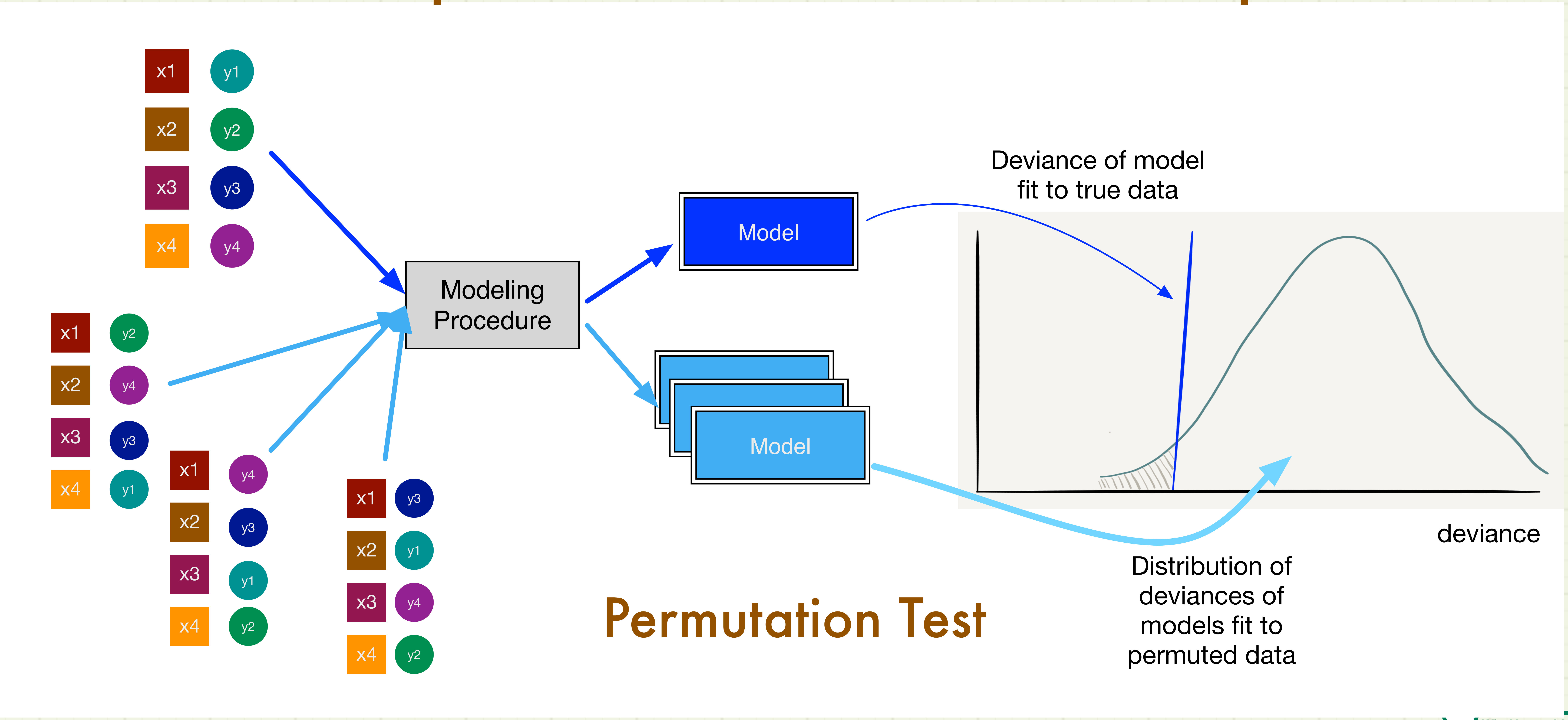
- Or: would a model fit to noise score this well?
- Or: Is the output *really* related to the inputs?





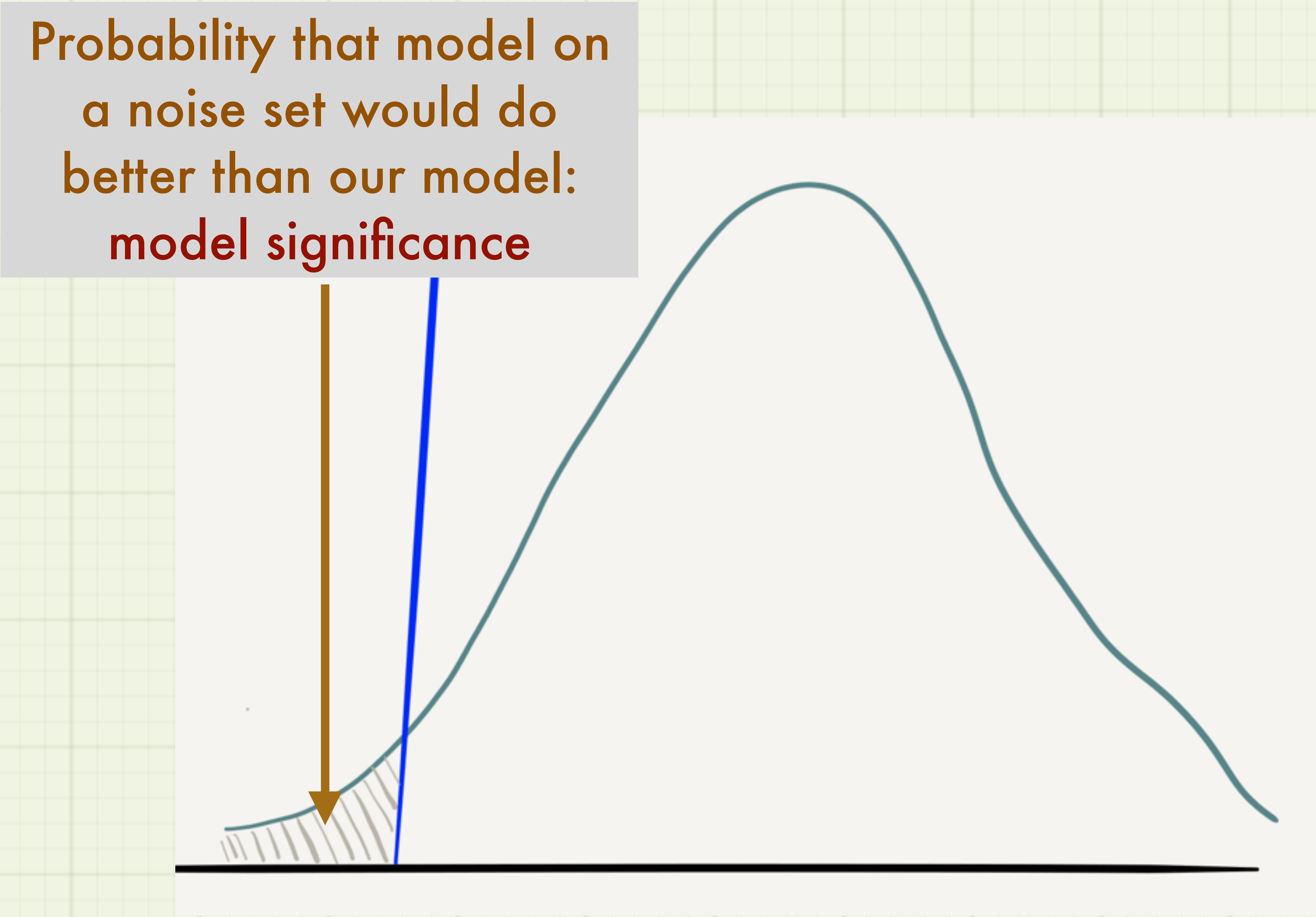
Win-Vector LLC

Thought Experiment:
Is the input related to the output?



Model Significance

- In (data science) practice: hold-out set
- Useful for finding variables with signal
    - Fit a one-variable model
    - linear regression: F test
    - logistic regression: χ^2 test



Model Significance in R

Logistic Regression

```
# get the significance of glm model
get_glm_significance = function(model) {
  delta_deviance = model$null.deviance - model$deviance
  df = model$df.null - model$df.residual
  pchisq(delta_deviance, df, lower.tail=FALSE)
}
```

Linear Regression (from `summary(model)`)

```
# get the significance of lm model
get_lm_significance = function(model) {
  fs = summary(model)$fstatistic
  pf(fs["value"], fs["numdf"], fs["dendf"],lower.tail=FALSE)
}
```

labs/ Lab04PermTestVarSel

“Ok. I know the model is predicting something.
But how well is it doing?”

Can I trust my model evaluation? In-sample vs. Out-of-sample estimates

- Many common model metrics are in-sample
  - R^2, model likelihood
  - RMSE or deviance on training set
  - **Performance estimates are upwardly biased**

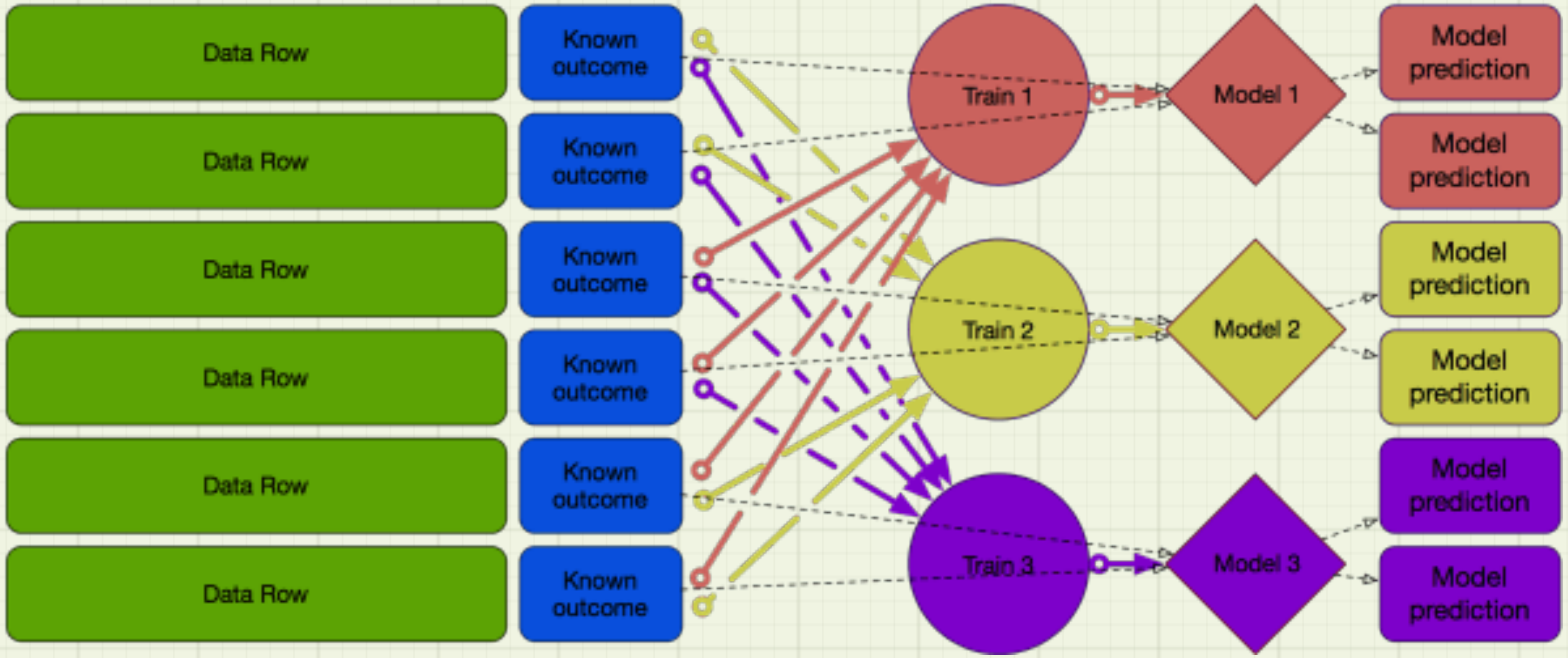
- In-sample metrics that try to compensate for bias
  - Adjusted R^2, AIC
  - *Only* consider bias due to model's degrees of freedom (# of parameters). Miss other multiple comparison issues that we don't tell them about!

Win-Vector LLC

Estimating Out-of-sample performance with training data

Cross-validation

- No point is evaluated on a model it helped to train



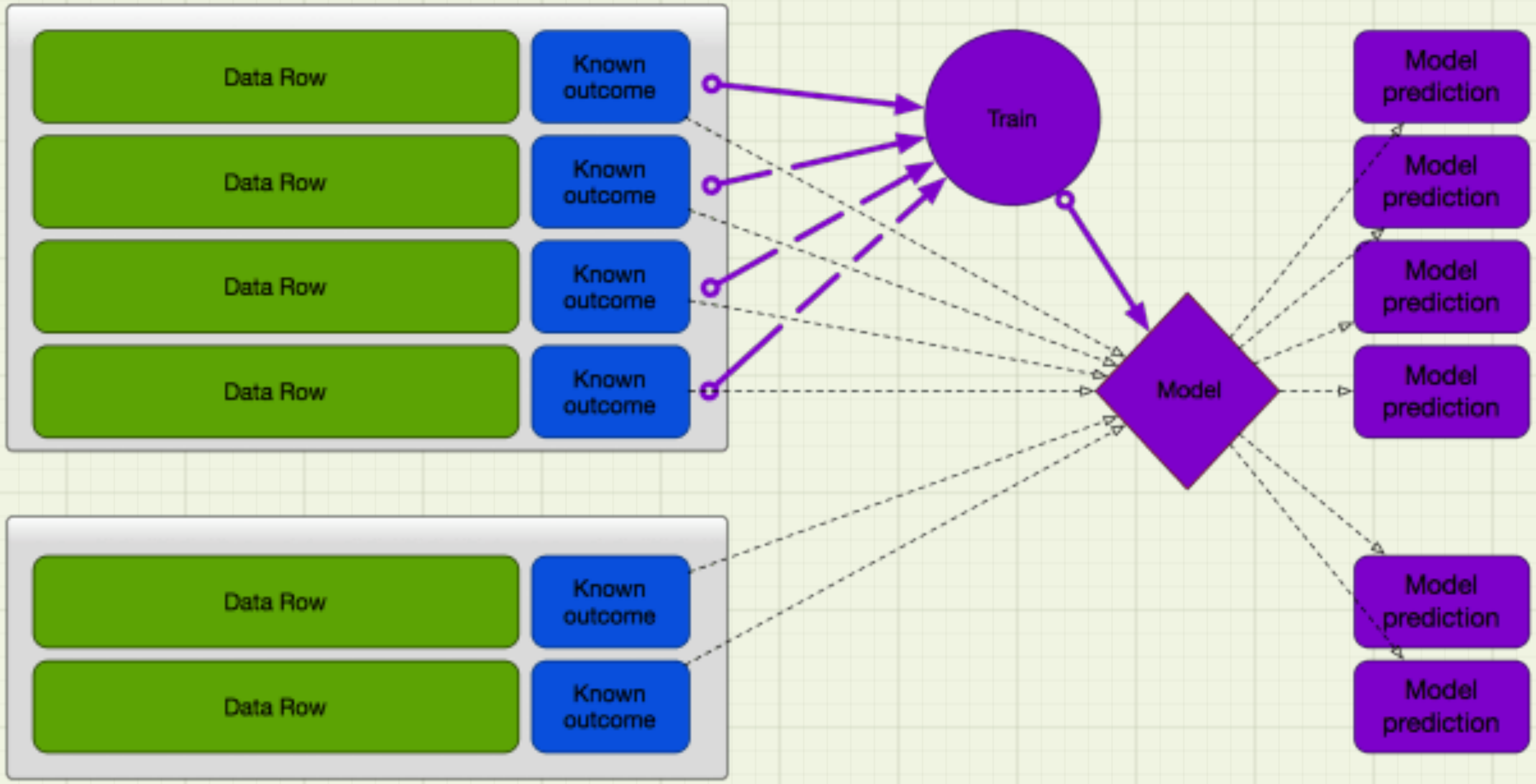
Cross-validation

- Unbiased estimate of out-of-sample error
- Statistically efficient
- Typically gives a point estimate of performance
- Computationally inefficient
    - With some exceptions: PRESS for linear models
- **Evaluates *modeling procedure* — NOT the production model.**

True estimate of out-of-sample performance: holdout data

Test-train split

- Subset of data only used for model evaluation



Win-Vector LLC

Test-train split

- Unbiased estimate of out-of-sample error
  - With no peeking!
- Statistically inefficient
  - Shouldn't split a small data set
- Point estimate only (in standard practice)
- Computationally efficient
- **Evaluates a specific model**

Cross-val vs. Holdout

	Statistically Efficient	Computationally Efficient	Evaluates Model	Evaluates Procedure
Cross-val	✓			✓
Holdout		✓	✓	

- Mnemonic: individual researchers are Bayesians (want to know *their* model works), bosses are frequentists (want to know *the modeling process* works).

Data Science : Data-rich

- Generally, we will prefer test-train split
  - Lots of data to spare for holdout
  - Large data sets make computational efficiency attractive
  - Possible exception: very rare target class

When to Consider Cross-val

- Rare target or rare features of interest
- Setting modeling parameters
  - Especially if data set too small for test-train-calibration split
- Variable selection
- Small data sets

Win-Vector LLC

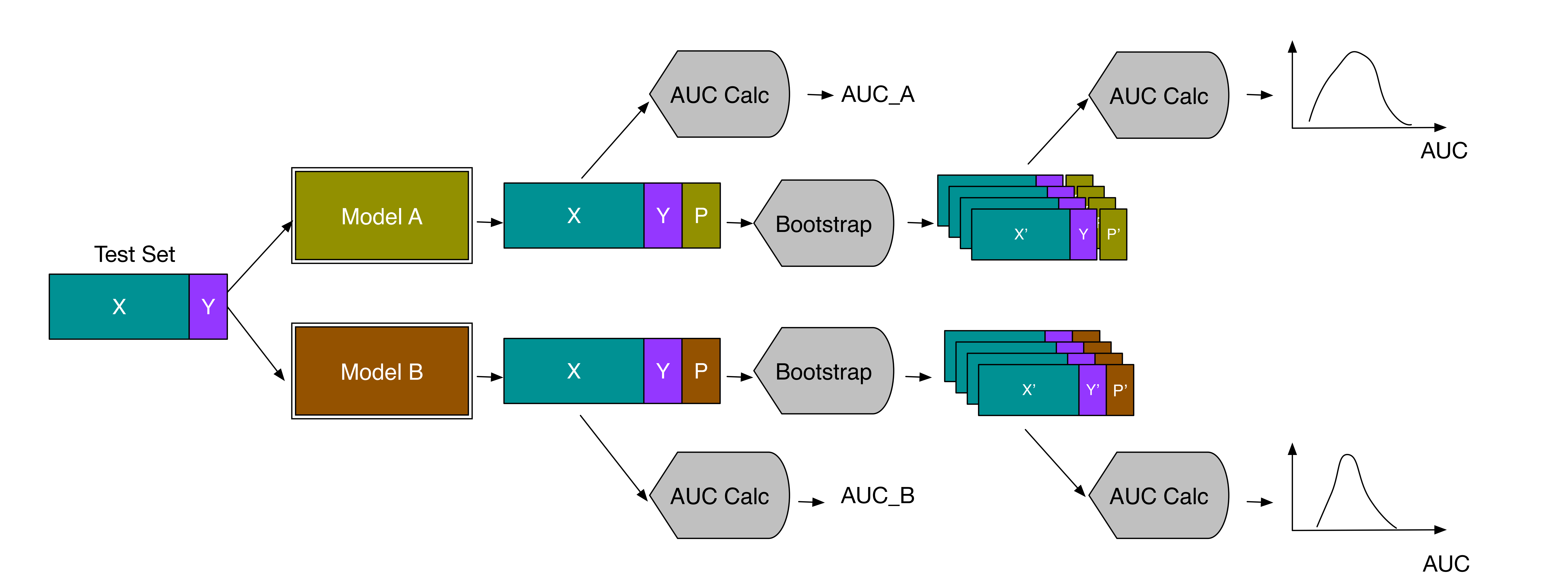
Holdout: No peeking! (Or not too much)

- In practice: fit model->evaluate->tweak model …
  - Too many iterations and performance estimates are upwardly biased again
  - Especially if the holdout set is small

- Recent differential-privacy related results to alleviate this
  - **http://www.win-vector.com/blog/2015/10/a-simpler-explanation-of-differential-privacy/**

Is Model A decisively better than Model B? Distribution Estimates

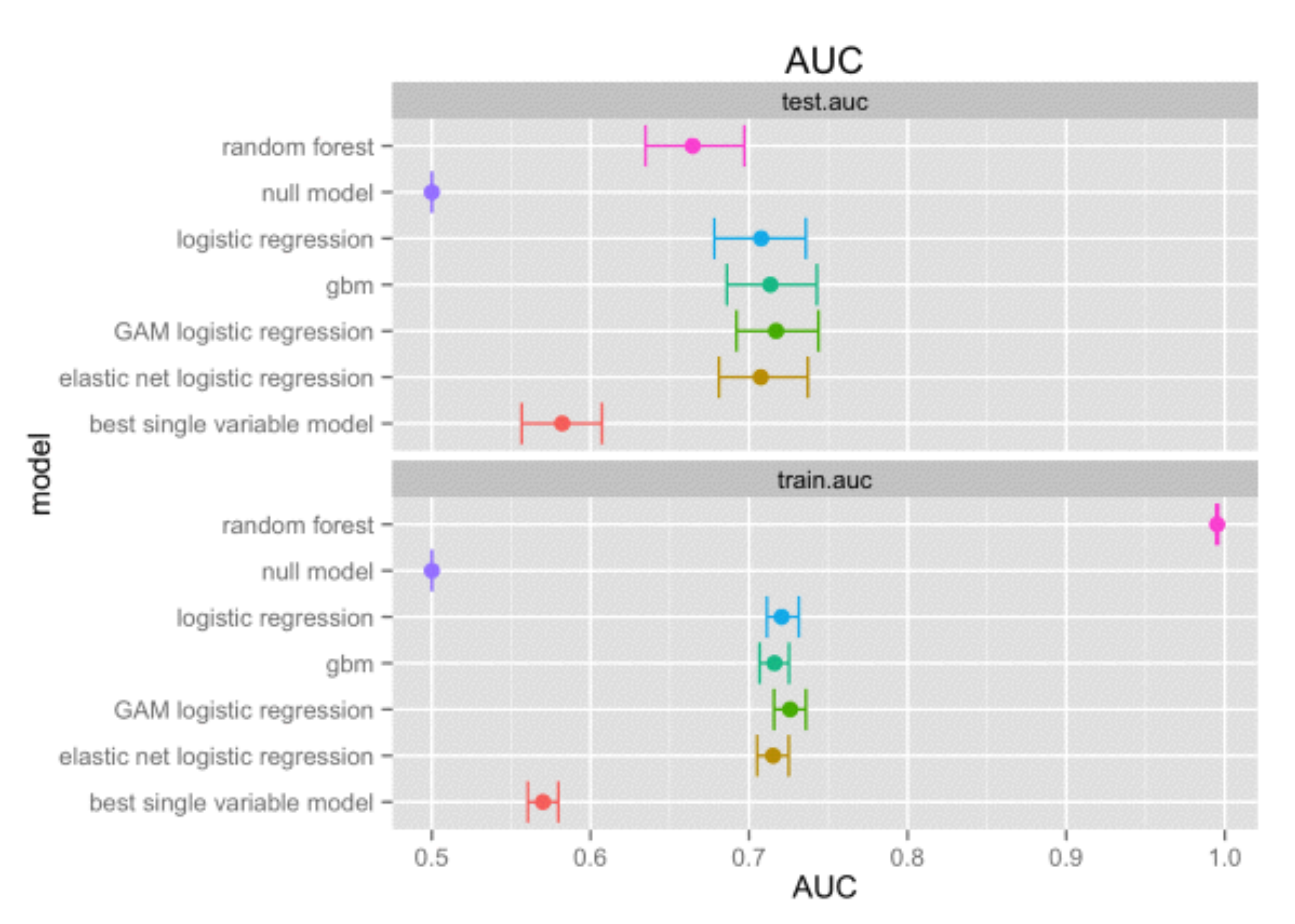
Bootstrapped Scoring



- One test-train split
- Build **one** model
- Bootstrap score the test (and training) sets

Bootstrapped Scoring

- Confidence Interval or Std. Dev. as distance unit
- Detect high variance
    - Situations the model gets really wrong (or really right)
- Does NOT measure stability of training procedure



labs/Lab05BootstrapTest

Conclusions / Take aways

- Model evaluation needs to be integrated into the entire data science process, including project proposal and result delivery.

- Metrics split between “technical” (for the data scientist, more appropriate early in a project) and “business” (adapted to intended use).

- R gives a flexible highly visual framework that implements classical statistical tests, important empirical tests, and any necessary ad-hoc procedures.

More Reading

- “How do you know if your model is going to work?”
  - http://www.win-vector.com/blog/2015/09/isyourmodelgoingtowork/
- “How Do You Know if Your Data Has Signal?”
  - http://www.win-vector.com/blog/2015/08/how-do-you-know-if-your-data-has-signal/
- “Statistics to English Translation”
  - Part 1: Accuracy Measures
  - Part 2a: ’Significant’ Doesn’t Always Mean ’Important’
  - Part 2b: Calculating Significance
- Win-Vector LLC ODSC 2015 “Preparing Data workshop”
  - https://github.com/WinVector/PreparingDataWorkshop/
- White paper and video: Data Preparation in R (require registration)
  - White Paper
  - Video (Pre-recorded webinar)

Thank You