

Quantifying Model Errors Using Similarity to Training Data

OR

MAD about MAD...

Rob Brown
Dana Honeycutt
Sarah Aaron

Accelrys Inc

May 2010

- **Problem Statement & Background**
- **Quantifying Model Errors**
 - For Regression Models
 - For Classification Models
- **Summary**

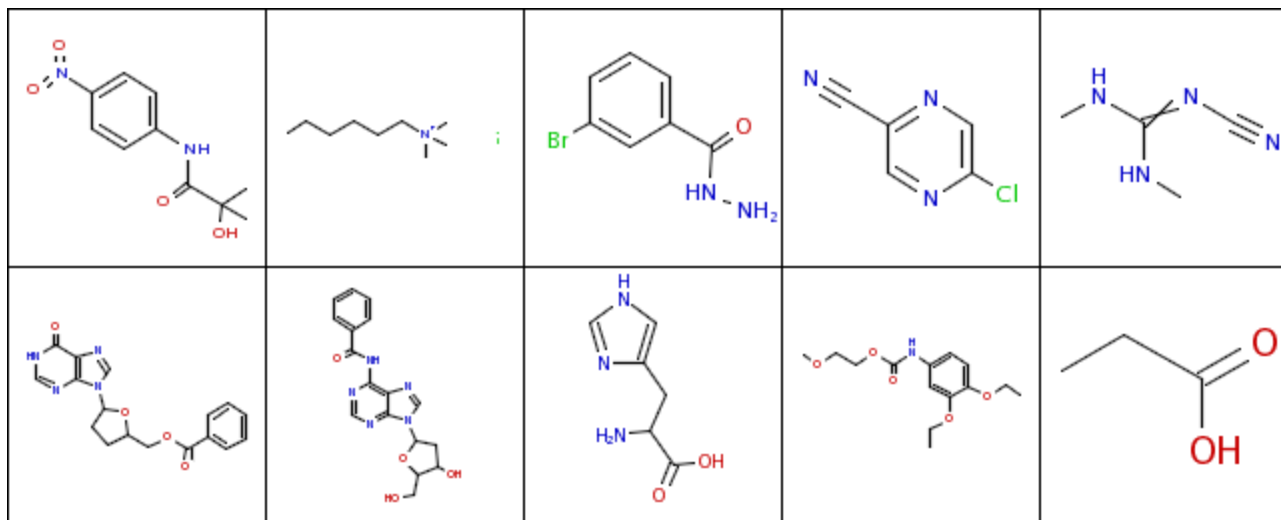
- **Philosophical/scientific grounds**

- A number without a tolerance or uncertainty is meaningless
- “Based on my model I predict the proposed molecule will have an $\log(\text{IC}_{50})$ of -2.3 ”
- “Based on my model I predict the proposed molecule will have an $\log(\text{IC}_{50})$ of -2.3 ± 0.5 ”
- “Based on my model I predict the proposed molecule will have an $\log(\text{IC}_{50})$ of -2.3 ± 5.0 ”

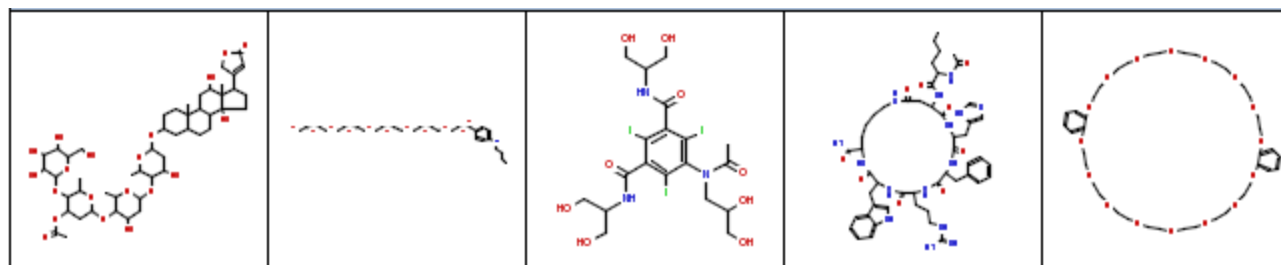
- **Enterprise software grounds**

- User of a model
 - May not be creator of model
 - Won't be familiar with training set
 - Not a data modeler familiar with model pros/cons/caveats.
 - May have either of these incorrect views:
 - “Models are great!” [and I always trust their results]
 - “Models are bogus!” [and I never trust their results]
- Error bars help put predictions in perspective

If you build a model from data for compounds that look like this:

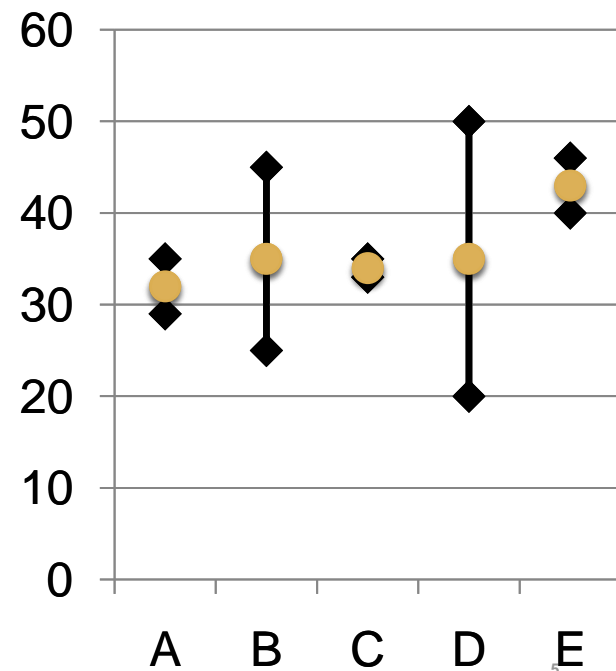


It probably won't predict well for compounds that look like this:



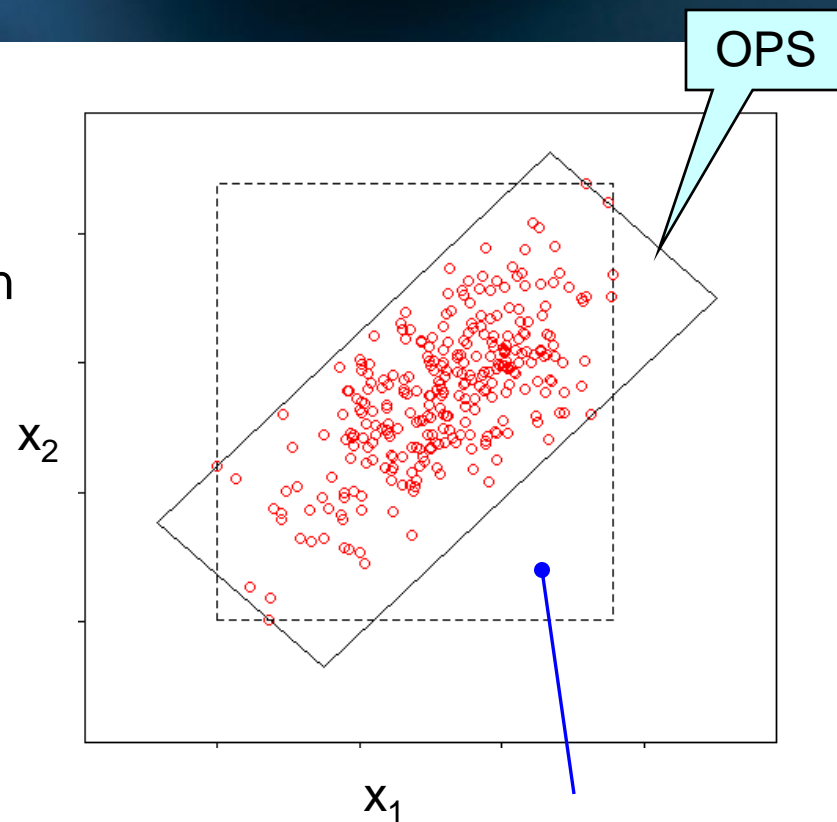
→ The second set of compounds falls outside the model's applicability domain (MAD).

- **QUESTION: Which Model Applicability Domain Measure correlate best with model error?**
 - Work by Sheridan et al., Horvath et al., and others shows that distance from test to training points correlates well with model error
 - Are there other model applicability domain (MAD) measures that correlate better?
 - Of various possible distance measures, which are the best?
- **QUESTION: Can we derive a standard procedure for computing error bars from MAD measures**
 - that gives accurate results?
 - that is simple to apply and understand?
 - that applies to any type of model – regression or classification?
 - that has few or no adjustable parameters?



- **Qualitative: Is the Sample in or out of the MAD?**

- Are the properties of the sample within the same range as the training data properties?
- For molecules
 - Does the sample have any structural features not seen in the training compounds?
 - Does the sample lack any structural features seen in all the training compounds?



- **Quantitative: How close to the training data is the sample?**

- Distance to closest training sample(s)
- Distance to center of data (centroid)
- Which distance measure to use (Euclidean, Tanimoto,...)?

In-range for single variables, but out of OPS

- **“Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR”, R. P. Sheridan, B. P. Feuston, V. N. Maiorov, and S. K. Kearsley, J. Chem. Inf. Comput. Sci., 44, 1912 (2004).**
- **Multiple data sets, multiple learner types**
- **Dice similarity with atom-pair fingerprints as “distance” measure**
- **Strong correlation between prediction error and**
 - similarity to closest training point
 - nearest-neighbor count (based on a similarity cutoff)
- **Windowing of data to get better correlations**

- **“Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models”, D. Horvath, G. Marcou, A. Varnek, J. Chem. Inf. Mod., 49, 1762 (2009).**
- **Notion of MAD as a separately trained meta-model (classifier)**
- **Test set for original model is training set for MAD**
 - Class 1: test samples well-predicted by original model
 - Class 2: test samples poorly-predicted by original model

- **Problem Statement & Background**
- **Quantifying Model Errors**
 - For Regression Models
 - For Classification Models
- **Summary**

- **Data Sets**

- Car prices (205),
- logP (subsets of 500-1000s)
 - chosen to make sure that there are test observations outside the model domain
- Fatheadminnow (100s),
- hERG (100s)

- **Descriptors**

- Continuous and Pipeline Pilot fingerprints (except Car data)

- **Modeling: GFA (type of linear regression)**

- Automated variable selection
- Consensus of multiple models gives good predictions

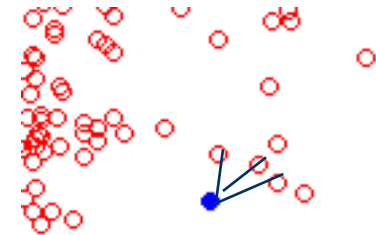
- **For each data set**
 - **Resample** data to generate 100s of training/test sets
 - Split each data set into 50/50 training/test
 - Build GFA model from training
 - Predict test set
 - Calculate residual errors
 - Calculate each MAD measure (next slide)
 - Correlate MAD measure to test set error

- Qualitative Measures

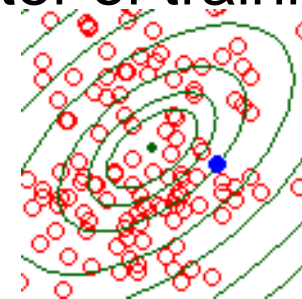
- **Nprop**: number of out-of-range property warnings
- **Nops**: number of out-of-OPS warnings
- **Nfp**: number of missing/unknown fingerprint feature warnings
- **Ntot**: = $Nops + Nprop + Nfp$ (total number of applicability domain warning)
- **NnonFP**: $Nops + Nprop$

- Quantitative Measures

- **MinDist/MaxDist/MeanDist**: min/max/mean distance to 3 (or 5) closest training samples



- **MD**: Mahalanobis distance to center of training data



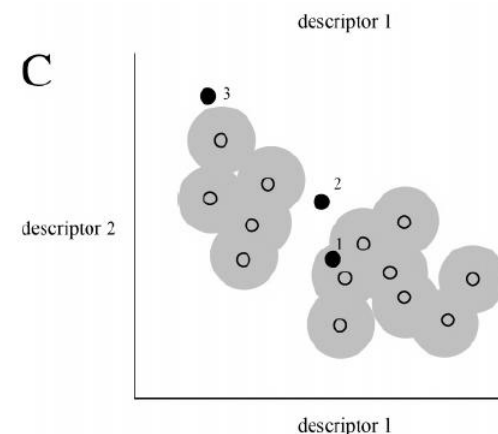
- **Correlate MAD measures to test set error**
 - Correlation to binned mean-squared error (MSE)
 - Divide non-discrete MAD measure into ~3–6 bins
 - Correlate bin location with average error for bin
 - For discrete MAD measures, each discrete value is a bin
- **Binning approaches**
 - Uniform bin width
 - Uniform bin populations – quartiles/percentiles
 - Better statistics than uniform width bins
 - Windowed (overlapping bin) averages to capture functional dependence of error
 - Sort data by increasing distance
 - Compute moving averages over a specified window size

Data Set	Typical r^2 (prediction)	RMSE (in-domain)*	RMSE (out-of-domain)*	Best MSE Indicator	Best MSE Correlation
Auto price	0.85	0.18	0.24	MaxDistBin	0.89
LogP	0.85	0.51	0.68	MDBin	0.92
Fathead tox	0.65	0.61	1.05	MaxDistBin	0.96
hERG	0.45	1.09	1.21	MaxDistBin	0.73

*as determined by presence/absence of warnings in arising from out-of-range property values, out-of-range OPS values, or unknown fingerprint features

- **Distance to 3rd (5th) nearest neighbour or Mahalanobis distance best correlates to error**

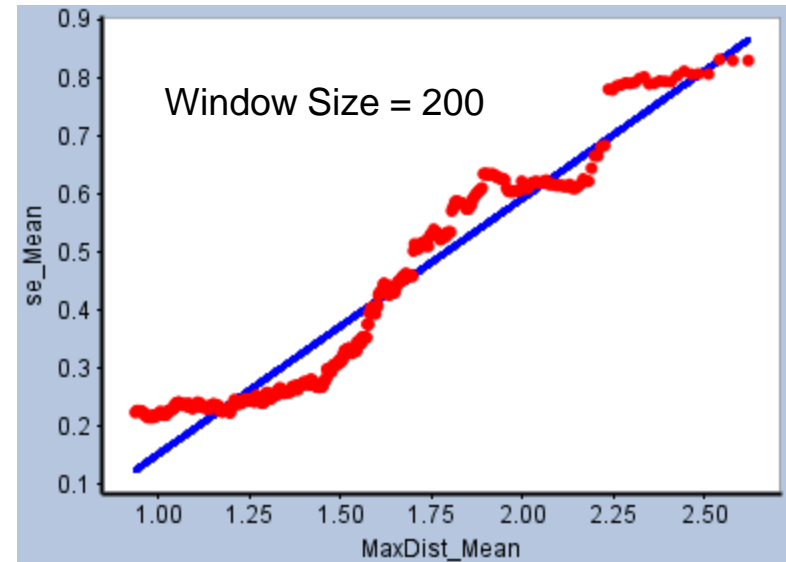
— Consistent with Sheridan *et al*



- **Uniform number of samples per bin**
 - E.g., based on quartiles or percentiles
 - Yields better statistics than uniform bin width
 - Better distance-error correlation for good models; worse for bad models

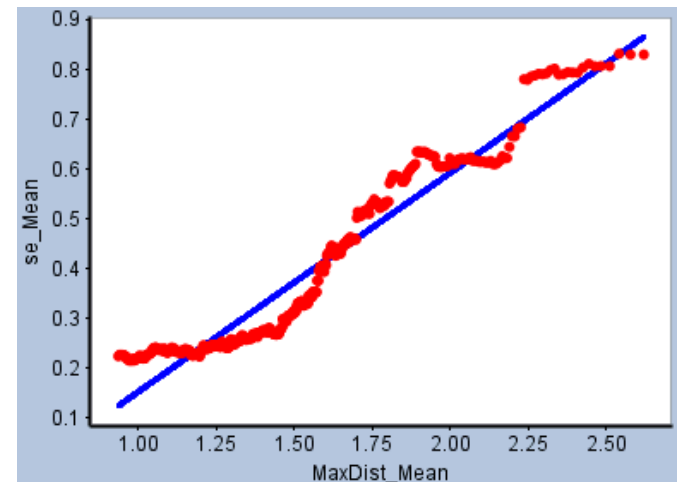
Data Set	Typical r^2 (prediction)	Uniform bin MaxDist-MSE correlation	Quartile-based MaxDist-MSE correlation
Auto price	0.85	0.89	0.91
LogP	0.85	0.84	0.98
Fathead tox	0.65	0.96	0.88
hERG	0.45	0.73	0.56

- **Sort by increasing Max Dist**
- **Look as MSE over sliding window**
- **Examine different window sizes**
- **At certain window size → linear correlation between MSE and Mean Max Dist**



The fitted line is used to compute error bars when making predictions.

- **Baseline error bar:**
 - Use test set RMSE as error bar
 - A fit-based error bar must improve on this baseline to be useful
- **Validation Procedure**
 - Split test set into two subsets (1st, 2nd)
 - Use 1st split to fit windowed squared-error (se) vs. $MaxDist$
$$\langle se \rangle = a + b * \langle MaxDist \rangle$$
 - (angle brackets denote average over window)
 - Use a and b to predict se for 2nd split:
$$se_{pred} = a + b * MaxDist$$
 - If (predicted error) < (minimum observed mean error), use min error instead
 - Compare se to se_{pred} to see how much fitted error improves on baseline error
- **Note: A split test set is just one possible approach. Cross-validation errors could just as well be used.**



Results Summary: Window-based errors



Data Set	Typical r^2 (prediction)	Best R Value	Best Window Size	Test set size**	Window / Test Set Size
Auto price	0.85	1.28 (1.37*)	20	50	0.40
LogP	0.85	1.68 (1.65*)	200	400	0.50
LogP	0.85	1.59	125	200	0.63
Fathead tox	0.65	1.42	80	170	0.47
hERG	0.45	1.01	10,15,25	63	0.16,0.24,0.40

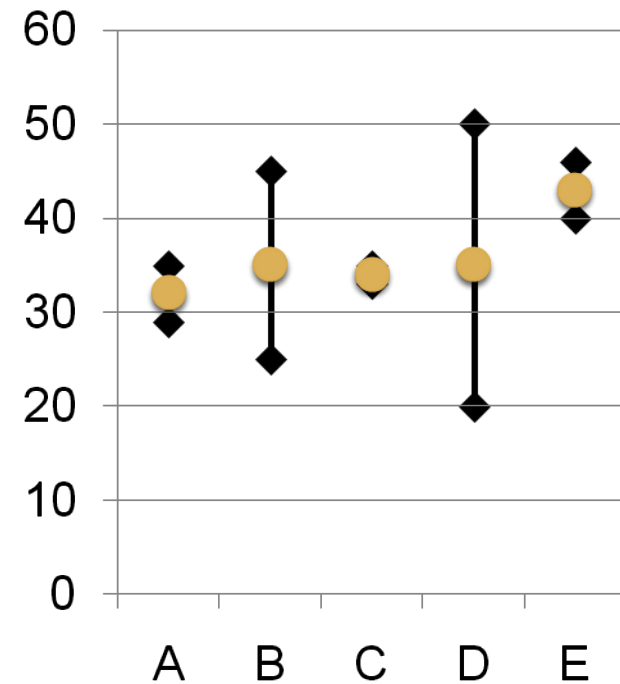
*R values in parentheses use Mahalanobis distance to closest samples (not to centroid) for MaxDist; all others use Euclidean

**Mean size of 1st test set split used to fit error to MaxDist

$$\bullet R = \text{RMS}[\langle se \rangle - \langle se_{\text{mean}} \rangle] / \text{RMS}[\langle se \rangle - \langle se_{\text{pred}} \rangle]$$

• R > 1 implies “good” error bar relative to baseline

- **We can provide estimated error bars for predictions based on the distance of the predicted sample to a number of training set neighbours**
- **Windowed, uniform-bin, and quantile-bin error bars all improve on baseline error...**
 - ...as long as model is adequate (prediction $r^2 > \sim 0.5$)
 - start with a decent model before the question of error bar dependence on MAD even becomes relevant
 - Not yet clear which of 3 methods is best
- **Generality of the approach needs to be validated**
 - with additional data sets
 - with different regression model types



- **Problem Statement & Background**
- **Quantifying Model Errors**
 - For Regression Models
 - For Classification Models
- **Summary**

- **Classifiers**

- A single prediction is either Right or Wrong
- No continuum of error as in regression ($y_i - \hat{y}_i$)
- What does an “error bar” mean in this context?

- **When used as rankers**

- Model prediction is likelihood of sample being in one class vs. another
- But still no way to compute error magnitude for a single sample
 - Necessary to use binning/windowing techniques

- **Aim**

- to determine the degree to which prediction trustworthiness changes as we move through and outside the MAD
- NOT to provide a(nother) method for determining the best classifier and/or cut-off

- **Data Sets**

- Ames Mutagenicity (6500/3500)
- Abbott MAO (1250/114)
- NCI AIDS data set (3500/230)

- **Algorithms**

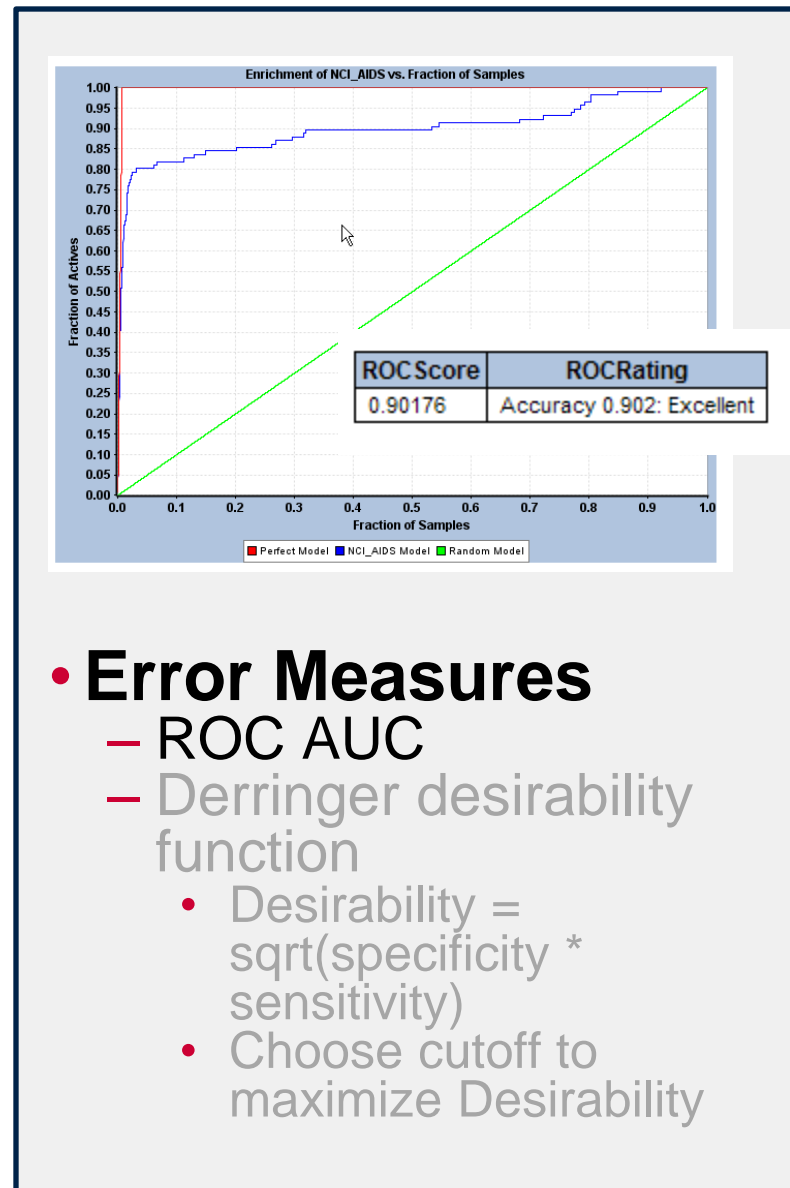
- ECFP4 Fingerprint
- Bayesian Classifier

- **MAD Measure**

- Mean max_dist

- **Split data into**

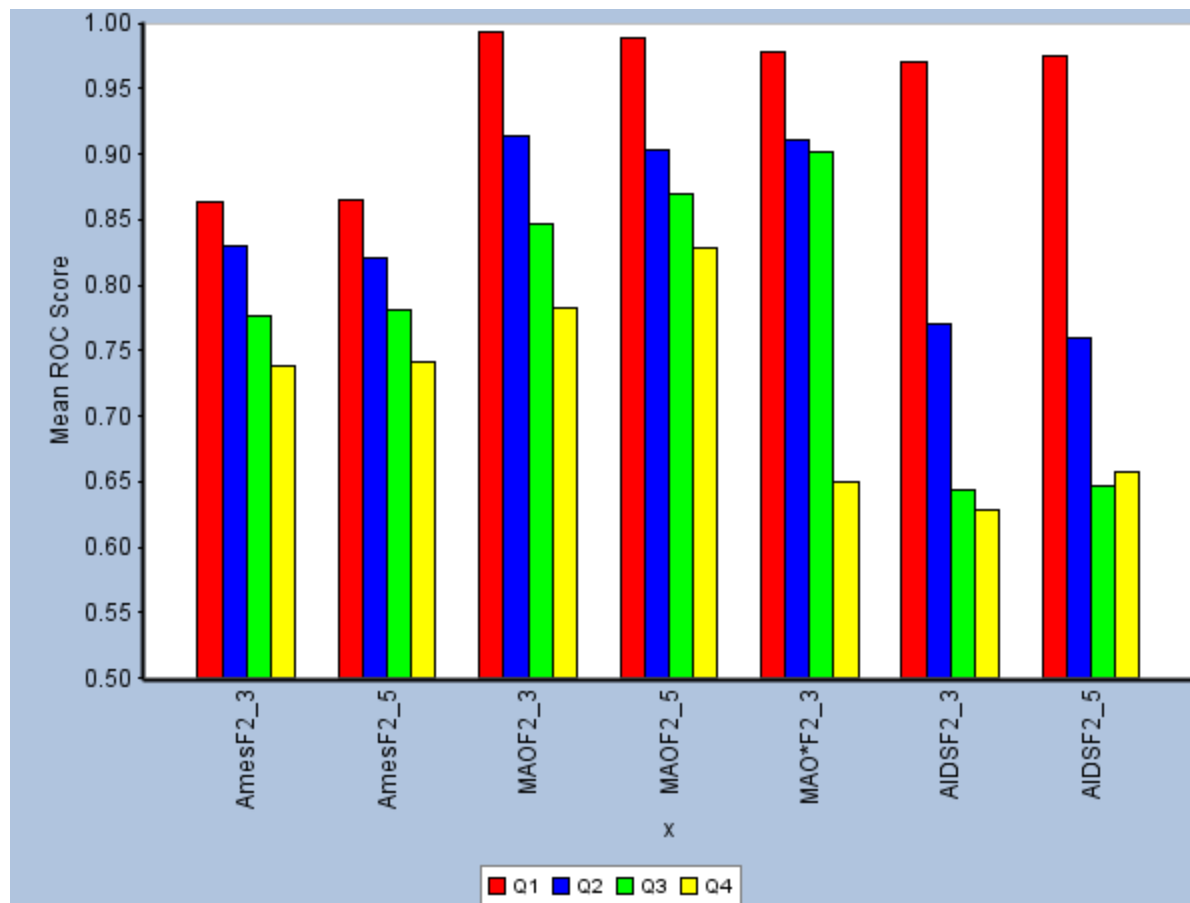
- Training – Build Model
- Test 1 – Predict Errors
 - split the data into quartiles based on max dist (3 or 5)
- Test 2 – Test Error Predictions
 - Apply quartile boundaries from above to divide data



- **Error Measures**

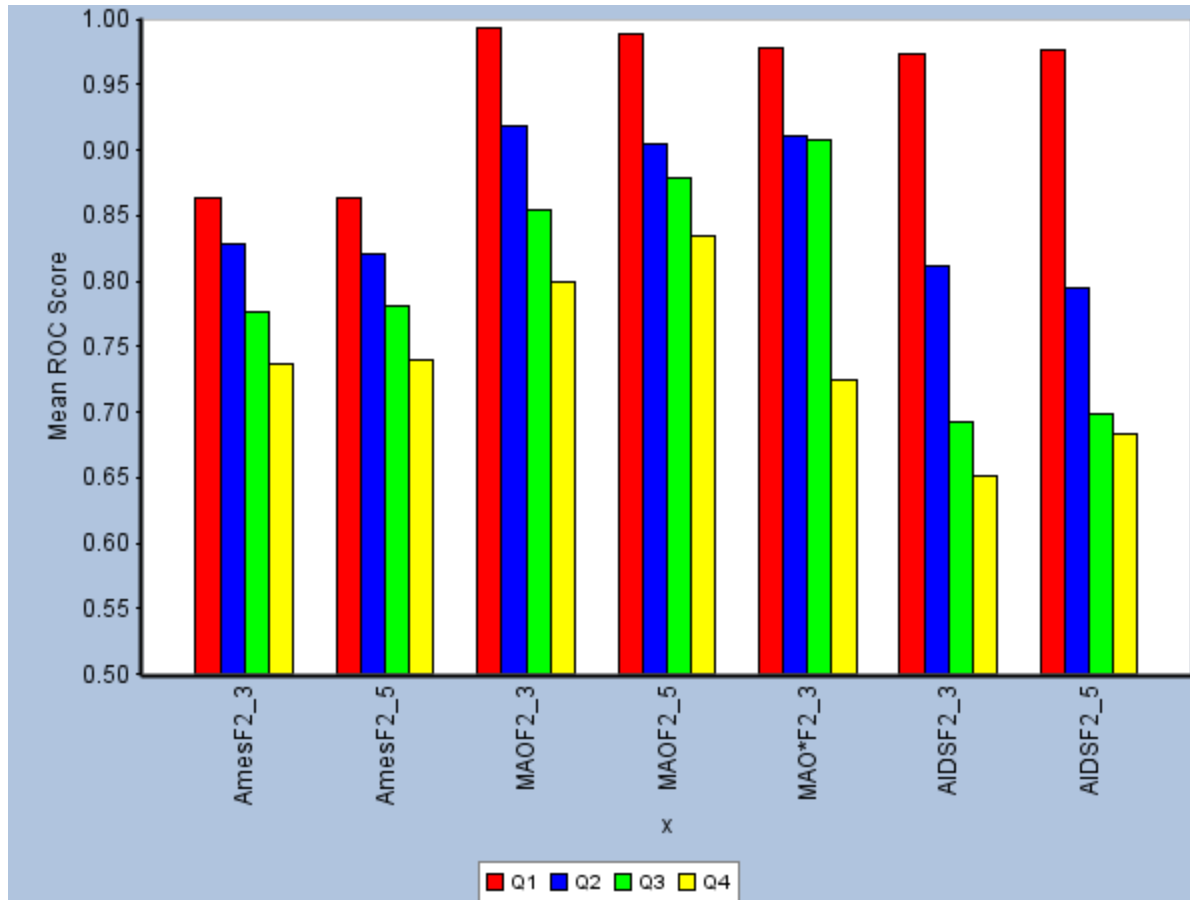
- ROC AUC
- Derringer desirability function
 - Desirability = $\sqrt{\text{specificity} * \text{sensitivity}}$
 - Choose cutoff to maximize Desirability

ROC AUC for Quartiles (all test data)



ROC score goes down as distance goes up – tells us that max dist remains a good measure

ROC AUC for Quartiles (split test data)



Reasonable prediction of ROC score on Test 2 taking the Test 1 to define the quartiles

- **Distance to training set neighbours provides a good measure of model applicability**
- **Quartile-dependent cutoffs give better predictions than a single fixed cutoff**
 - Thus distance-to-training-data information lets us both **assess** and **reduce** prediction error
- **Investigations are ongoing**

- **QUESTION: Can we derive a standard procedure for computing error bars from MAD measures**
 - that gives accurate results?
 - that is simple to apply and understand?
 - that applies to any type of model – regression or classification?
 - that has few or no adjustable parameters?
- **ANSWER: Appears to be Yes**
- **With predicted error bars it becomes safer to put models into the hands of a wider audience**