Why and How to exploit OOB Validation for Ensemble Size

Philip Kegelmeyer, Sandia National Labs, wpk@sandia.gov

CASIS, November 16, 2007

National Nuclear Security Administration

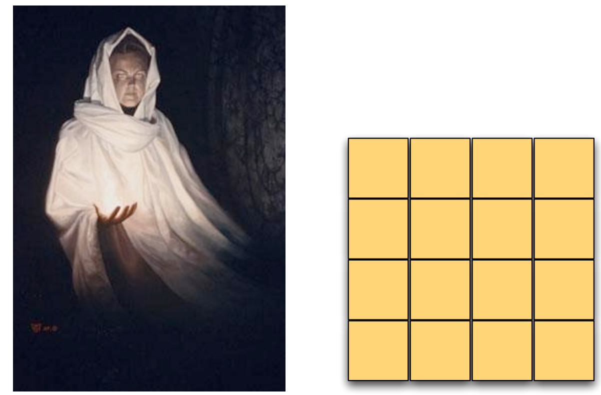
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

SAND Document #2007-6608A, R&A #5256574

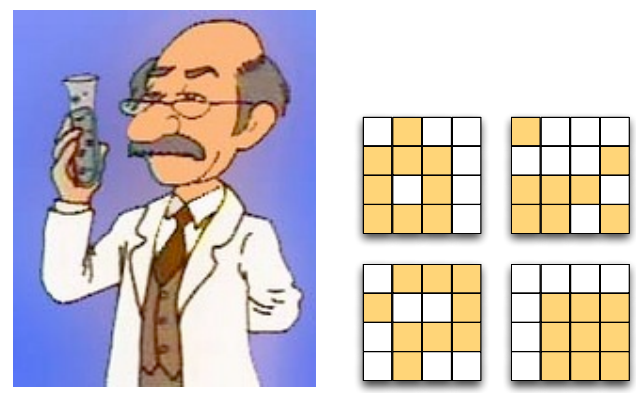
Machine Learning, With Ensembles

Traditional: Use 100% of training data to build a sage.



Sage sees all the data.

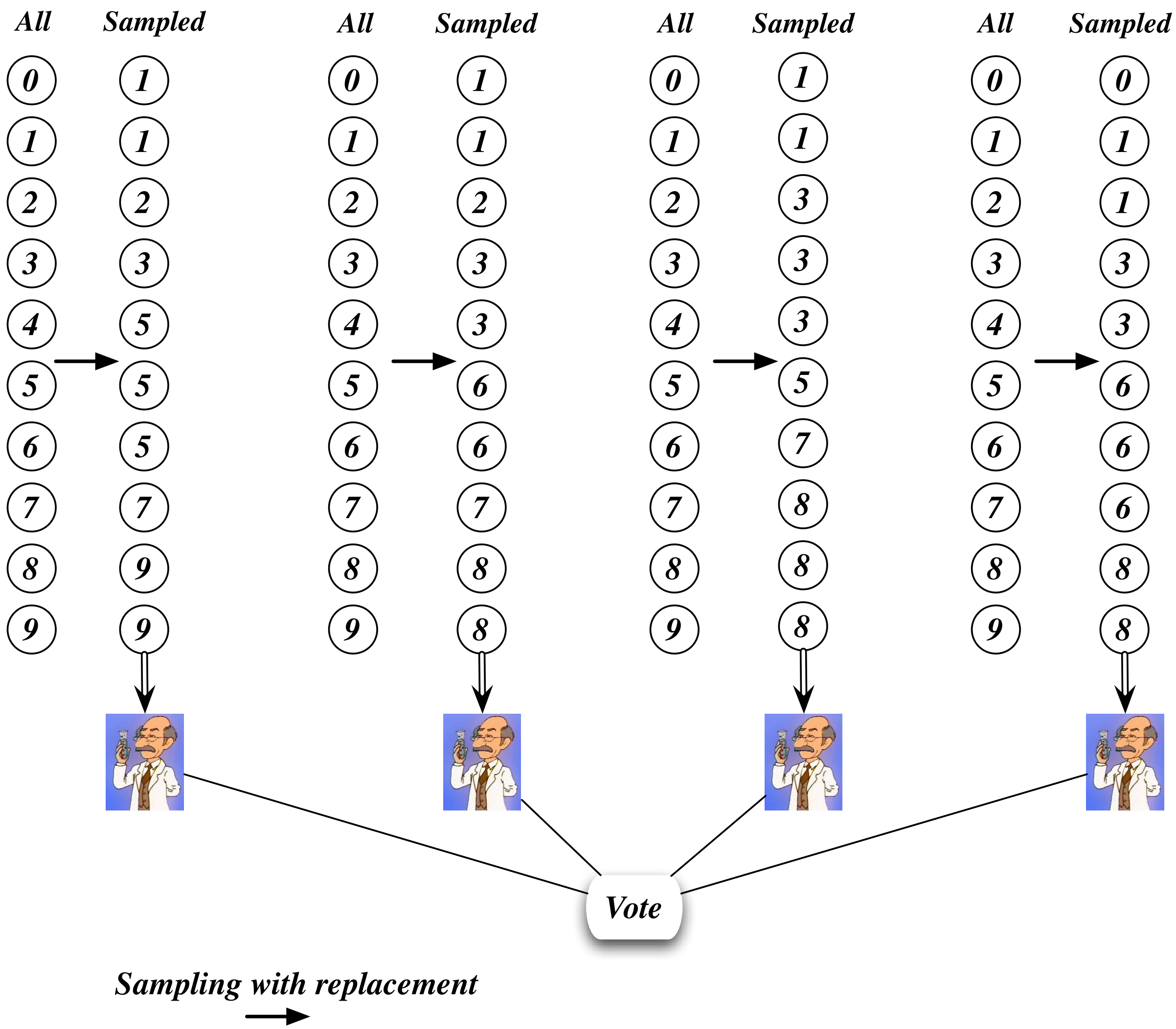
Ensembles: Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.



Each expert sees 2/3rds of the data.

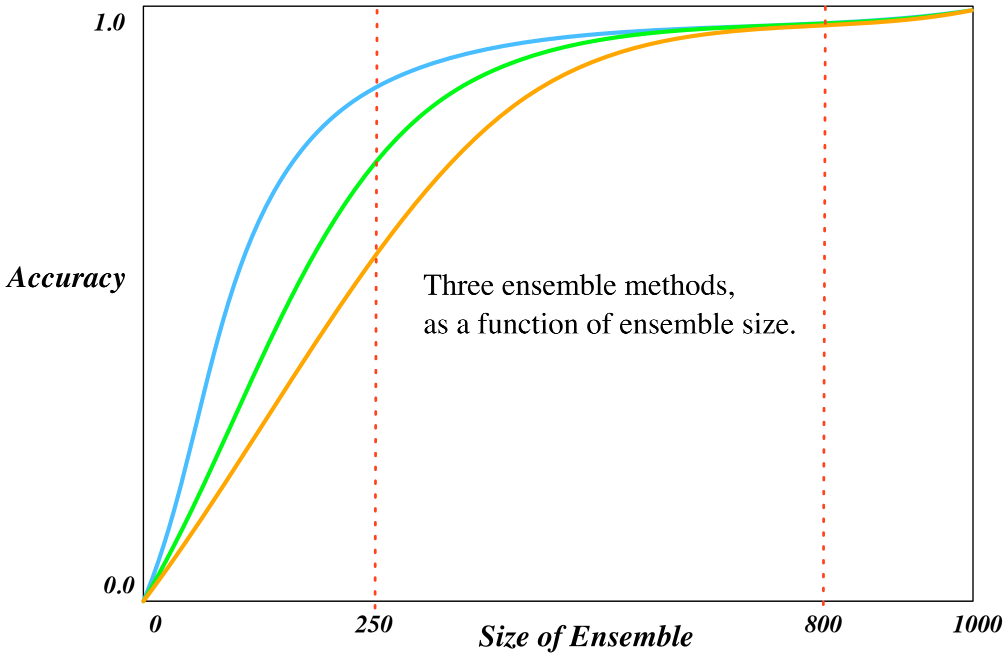
The experts beat the sage[1]!

“Bagging” is the Formal Name for This Method



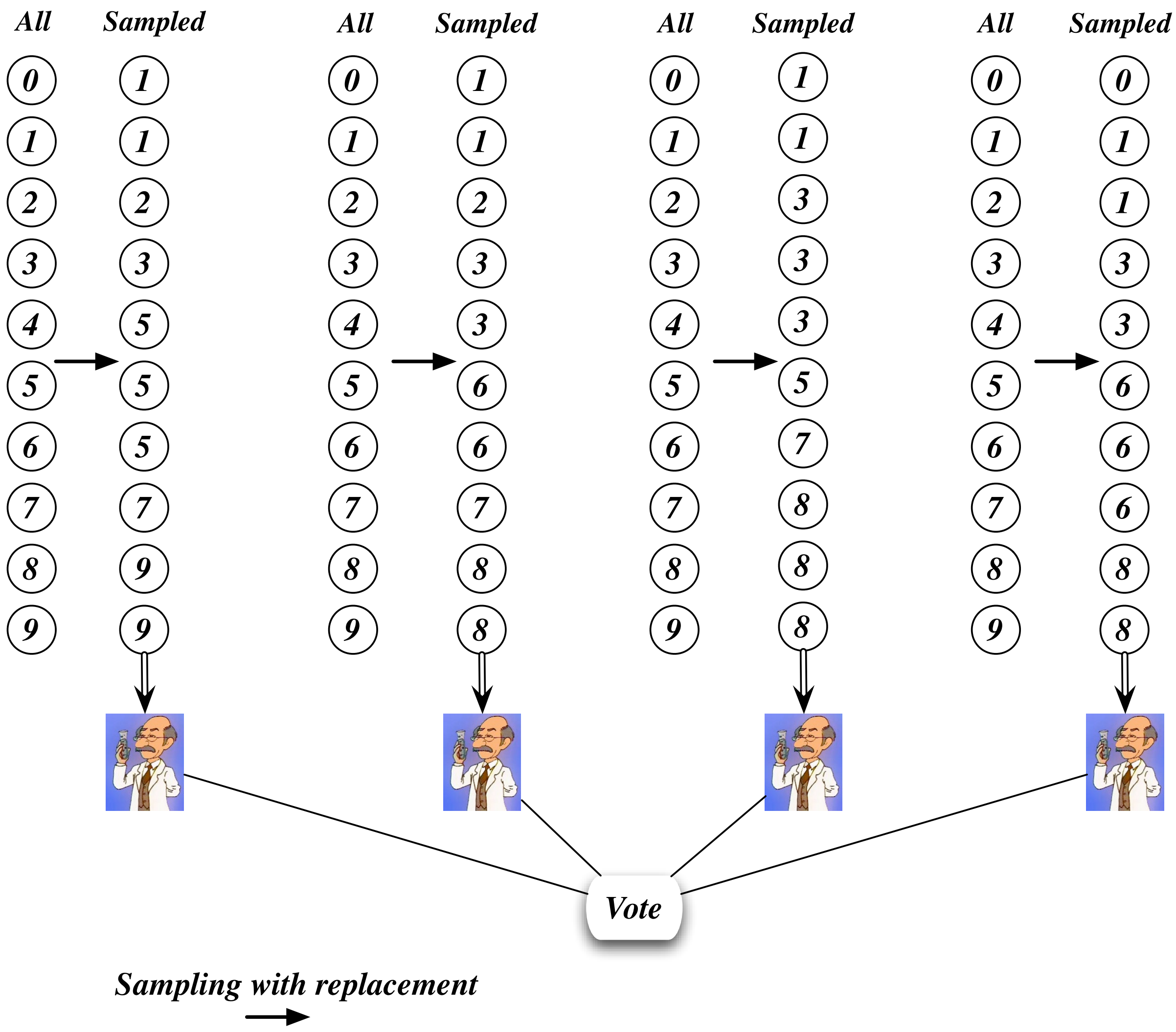
How Big An Ensemble Do You Need?

Don't use fixed size ensembles. They will short-change you and deceive you. Instead, stop when accuracy levels off.



But how to measure accuracy? *Don't* just use the training data.
Use a separate validation set? Sure, but they are rare and costly.
Out-of-bag (OOB) validation is easy and cheap.

Every Classifier Lacks a Fraction of the Samples

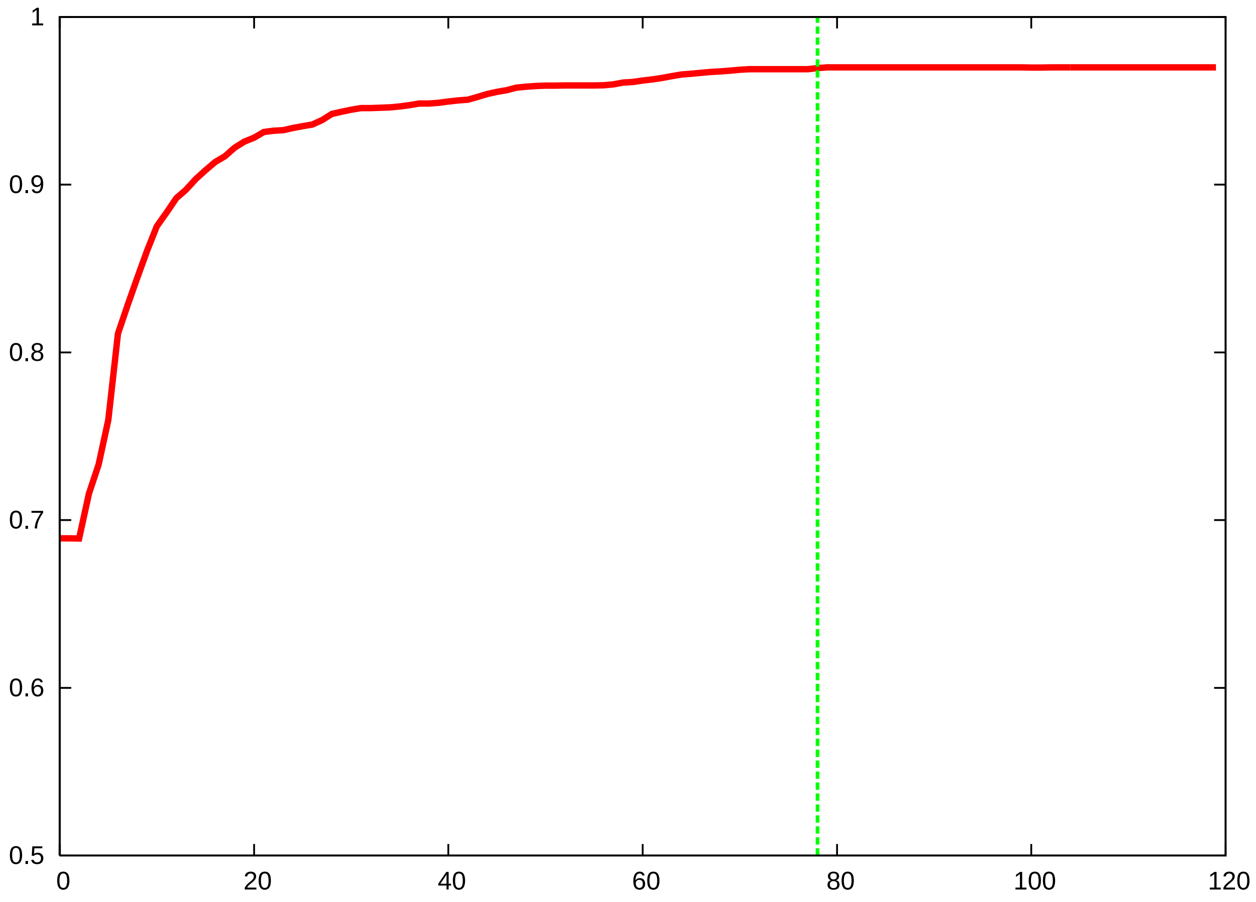


Every Sample Lacks a Fraction of the Classifiers!!

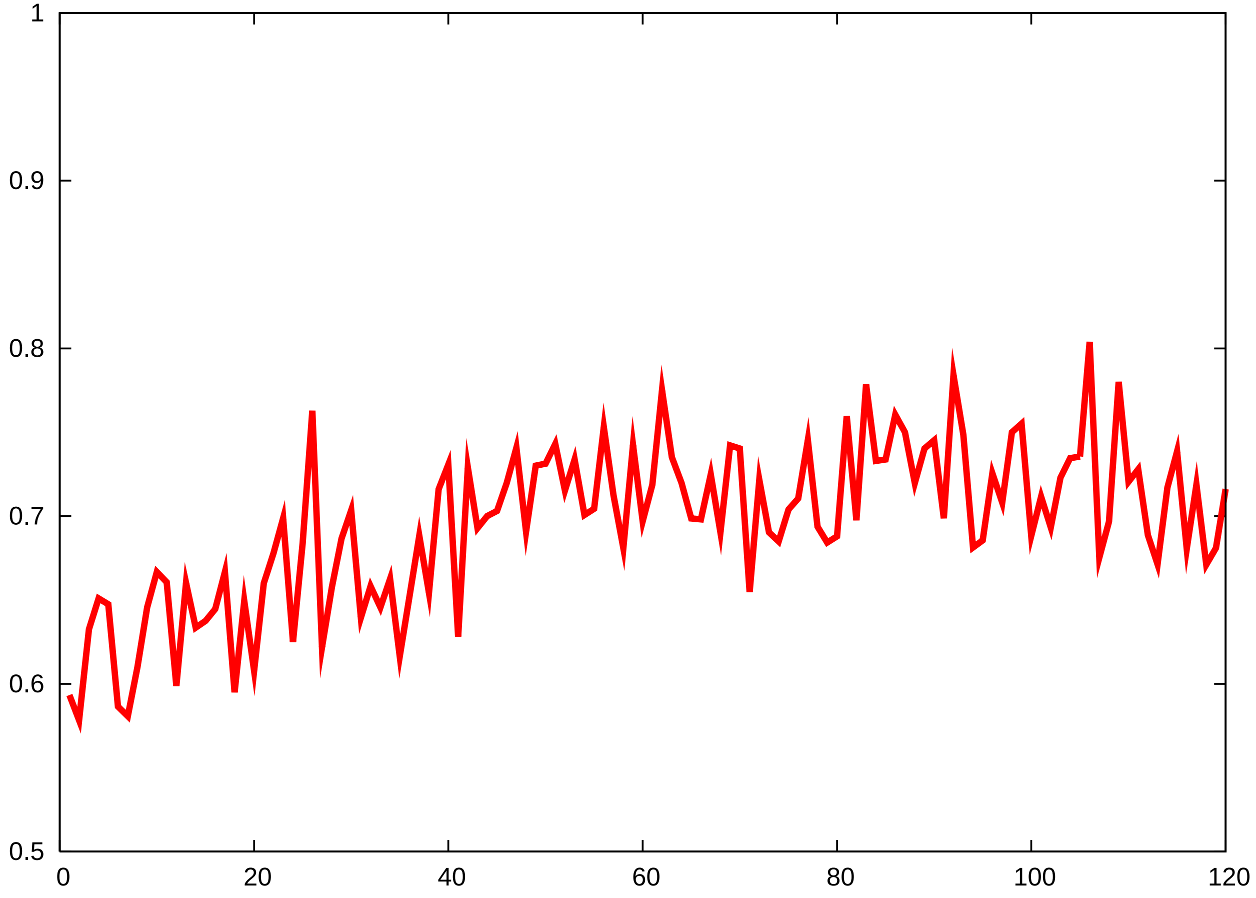
The classifiers that didn't see the sample can be fairly used to test it.

Sample 2 can be tested by E3 and E4; Sample 4 by E1, E2, E3 and E4.
Each sample can be tested by a substantial fraction of the classifiers.
So the over all accuracy is accumulated, one sample at a time.

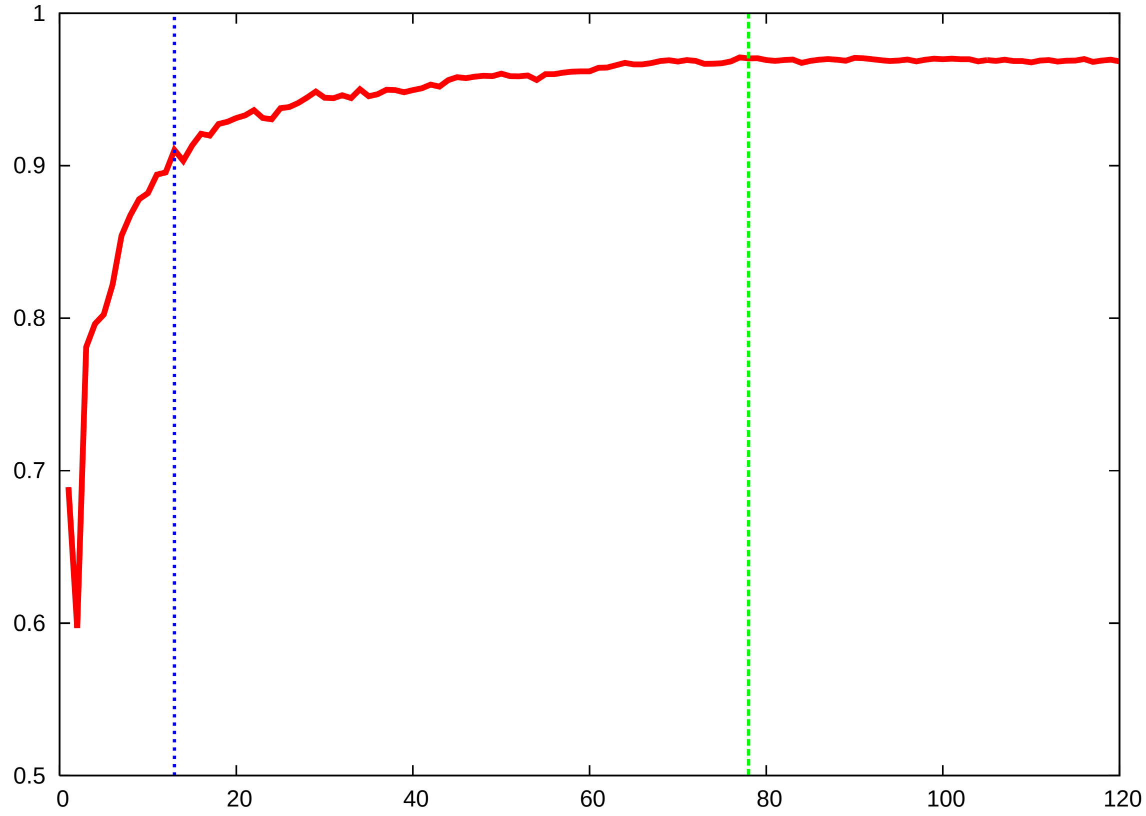
When To Stop? When Accuracy Flattens Out



But: Accuracy Can Increase Erratically!



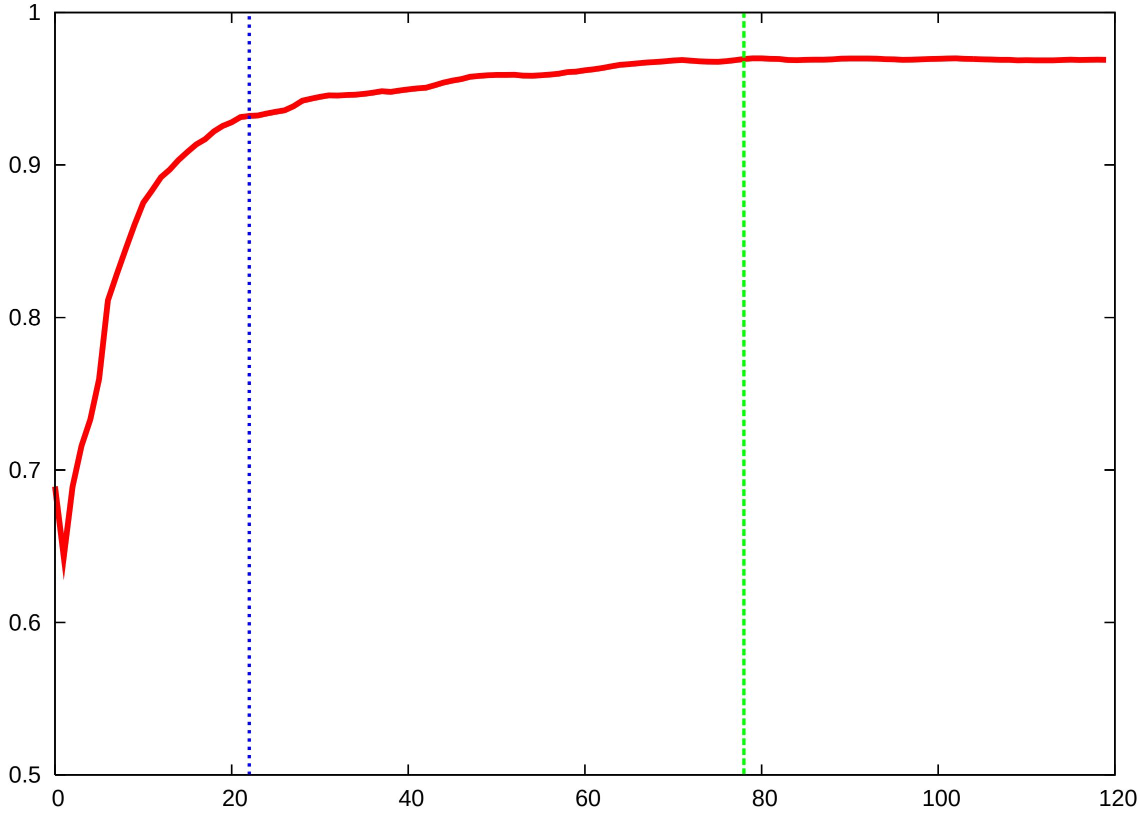
Simple Raw Accuracy Curve (From NIF Data)



Can't stop at first peak or plateau; accuracy curve must be smoothed.

So Smooth . . .

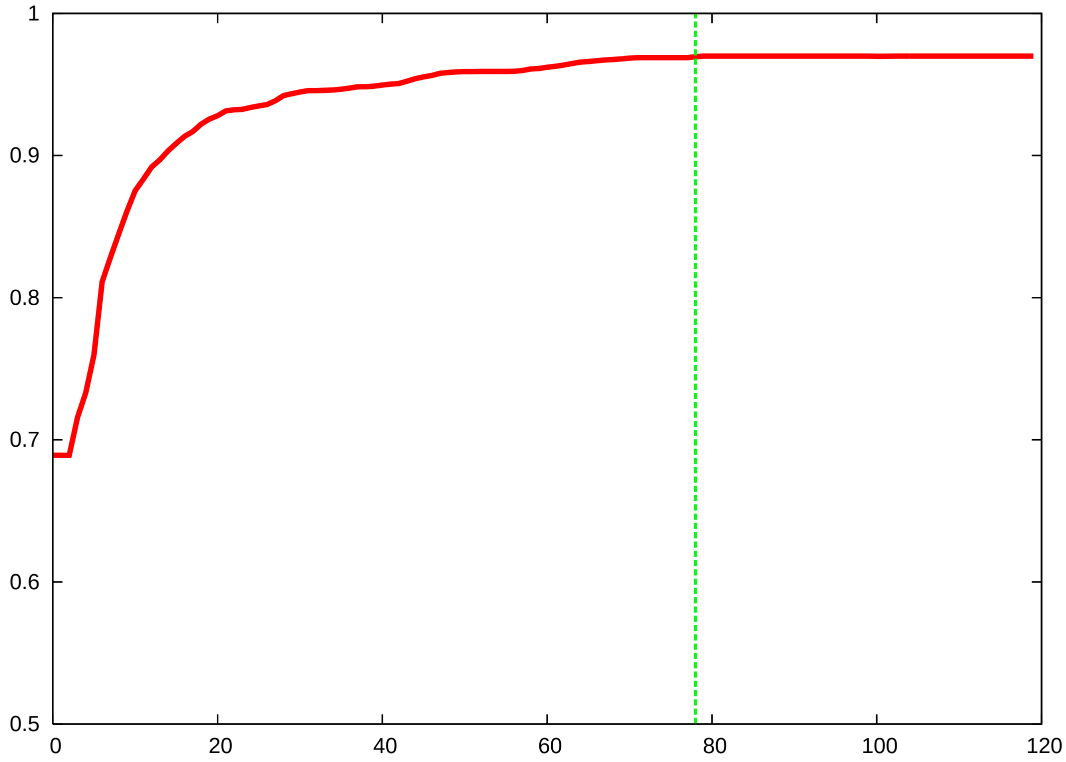
Smooth with a running average over a small window w_{small}.



$$w_{\text{small}} = 5$$

. . . and Check “Flatness” over Broad Window

Apply “set to maximum” filter over a broad window w_{large}, set ensemble size to first point that achieved max accuracy.

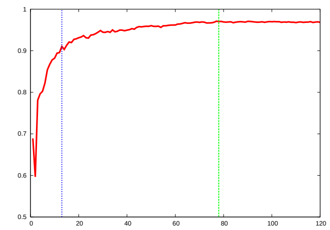


$$w_{\text{large}} = 20$$

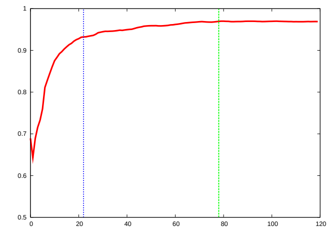
Summary: Stopping Point Selection

Three step algorithm for selecting a stop point[2]:

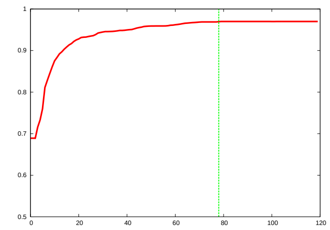
1. Maintain a running average over w_{small} samples, to smooth.
2. Track maximum accuracy over windows of size w_{large} until it doesn't increase.
3. Return size of ensemble that first achieved that accuracy.



Raw Accuracy Curve

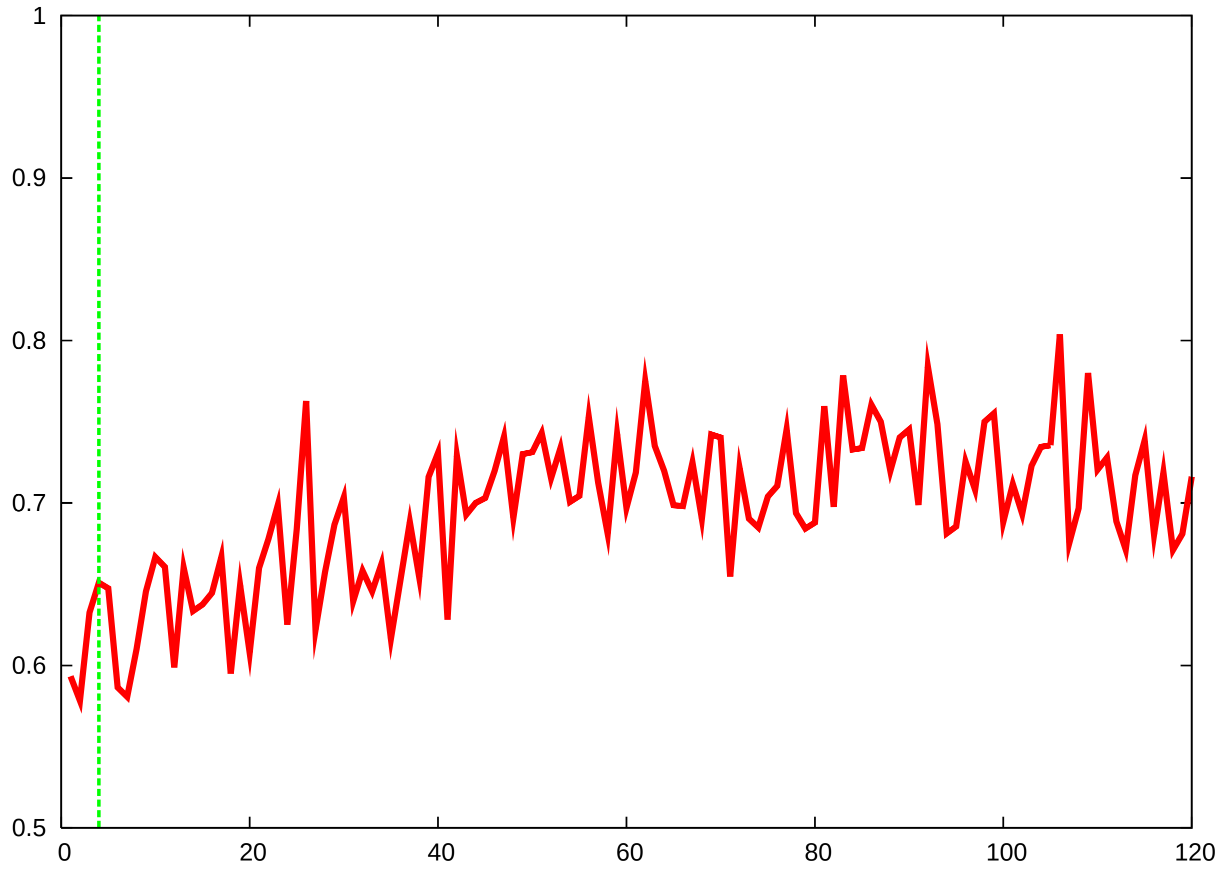


Smoothed Accuracy

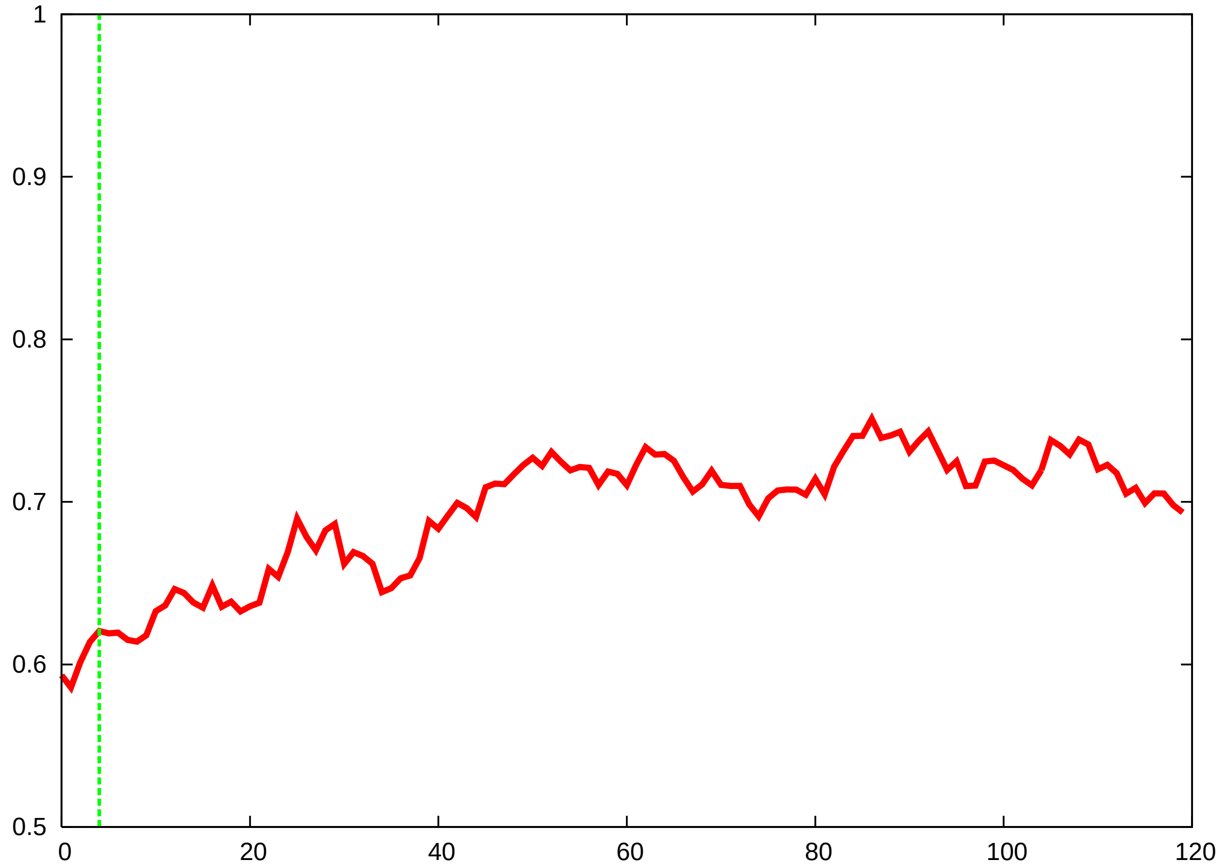


Maximum Filter Accuracy

Trickier Raw Accuracy (Protein Expression Data)

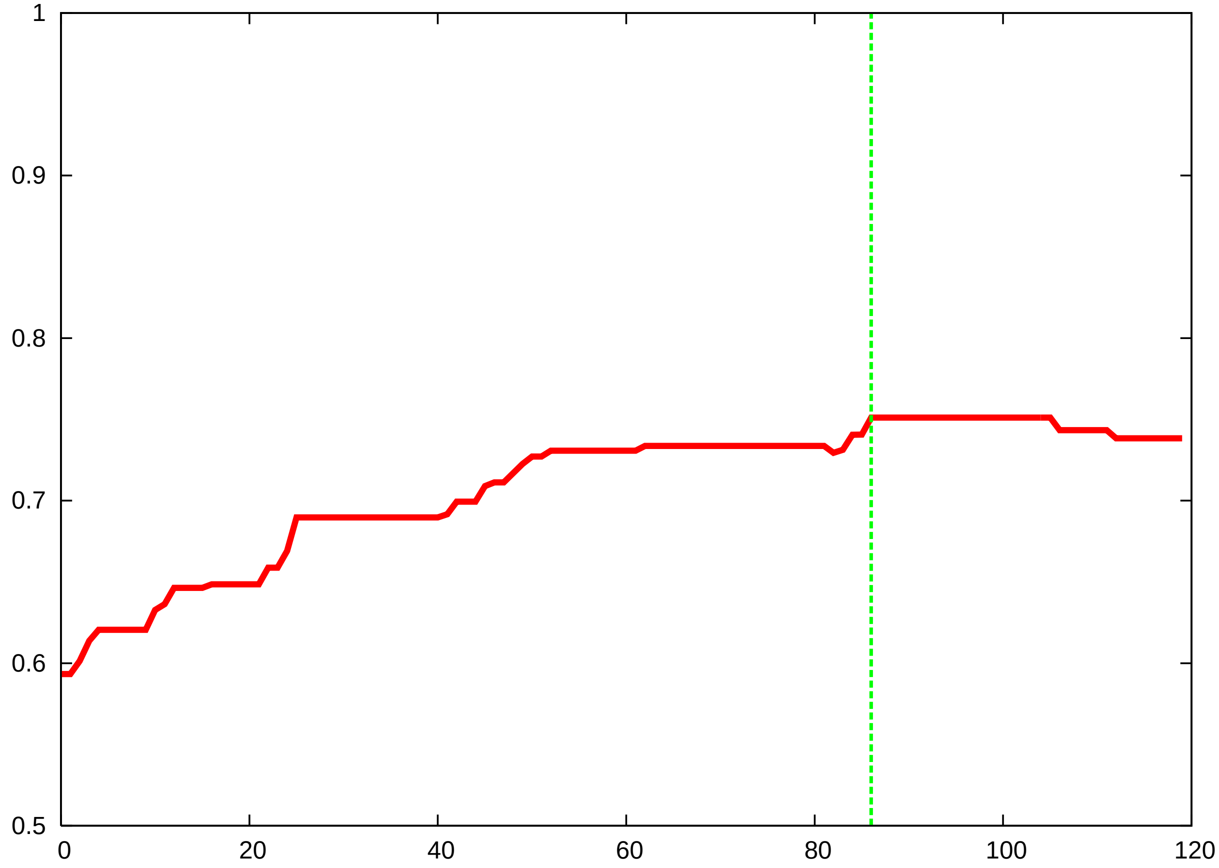


Smooth ...



$w_{\text{small}} = 5$

. . . and Check Flatness over Broad Window



$w_{\text{large}} = 20$

So: Smoothed Maximum Accuracy is Effective . . .

. . . but theoretically unsatisfying.

Next Steps:

- Generate a menagerie of real curves; build intuition.
- Estimate parameters from the curve itself?
  - Extract a non-parametric measure of variability from the raw ensemble data?
  - Explicitly model the "noise", the variation in accuracy?
- Consult with a trained 1D signal processor.

References

[1] BANFIELD, R. E., HALL, L. O., BOWYER, K. W., BHADORIA, D., KEGELMEYER, W. P., AND ESCHRICH, S. A comparison of ensemble creation techniques. In *Proceedings of the Fifth International Conference on Multiple Classifier Systems, MCS2004* (2004), J. K. F. Roli and T. Windeatt, Eds., vol. 3077 of *Lecture Notes in Computer Science*, Springer-Verlag.

[2] BANFIELD, R. E., HALL, L. O., BOWYER, K. W., AND KEGELMEYER, W. P. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 1 (January 2007), 173–180.