# Pattern Recognition for Massive, Messy Data

(Data, data everywhere, and not a thought to think)

**Philip Kegelmeyer**

Michael Goldsby, Tammy Kolda, Sandia National Labs

Larry Hall, Robert Banfield, et al., University of South Florida

Kevin Bowyer, Nitesh Chawla, et al., University of Notre Dame

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

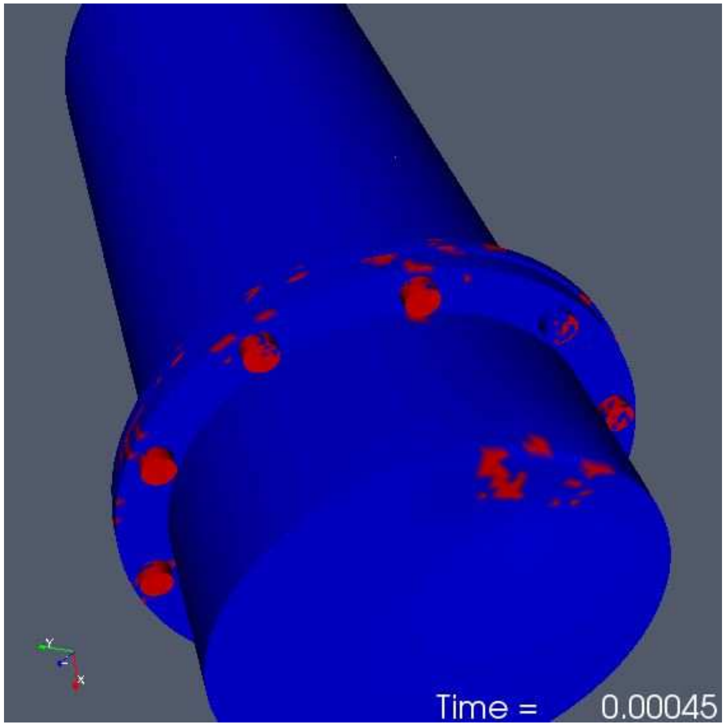
SAND Document #2006-3464, R&A #5243240

# Introduction and Summary

Sandia is developing "commodity" pattern recognition methods which handle data sets that standard methods cannot.

These commodity methods:

- Accept data as is, and in situ.
- Are robust to errors in attributes and labels.
- Scale to terabyte data.
- Are crucial to Stockpile Stewardship post-processing.
- Are broadly applicable, in Sandia and out.

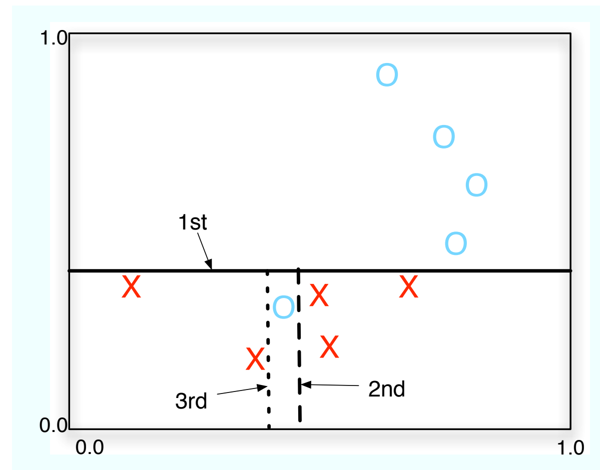



Bolt Failure Detection in ASC Data

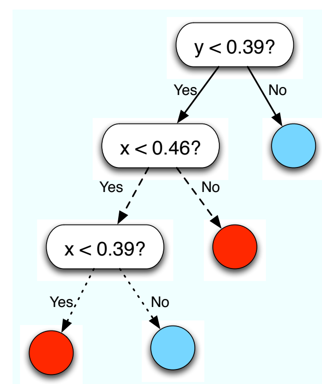
# Pattern Recognition Overview

Also known as: supervised machine learning, statistical inference, data mining.

- Input: "ground truth" data.
  - Samples, with attributes, and *labels*.
  - Example ASC context:
    - Samples: nodes, elements.
    - Attributes: variable values.
    - Labels: breach, bolt failure, "interesting".
- Apply suitable method:
  decision trees, neural nets, SVMs.
- Output:
  rules for labeling new, *unlabeled* data.
  Equivalently:
  a partitioning of attribute space.

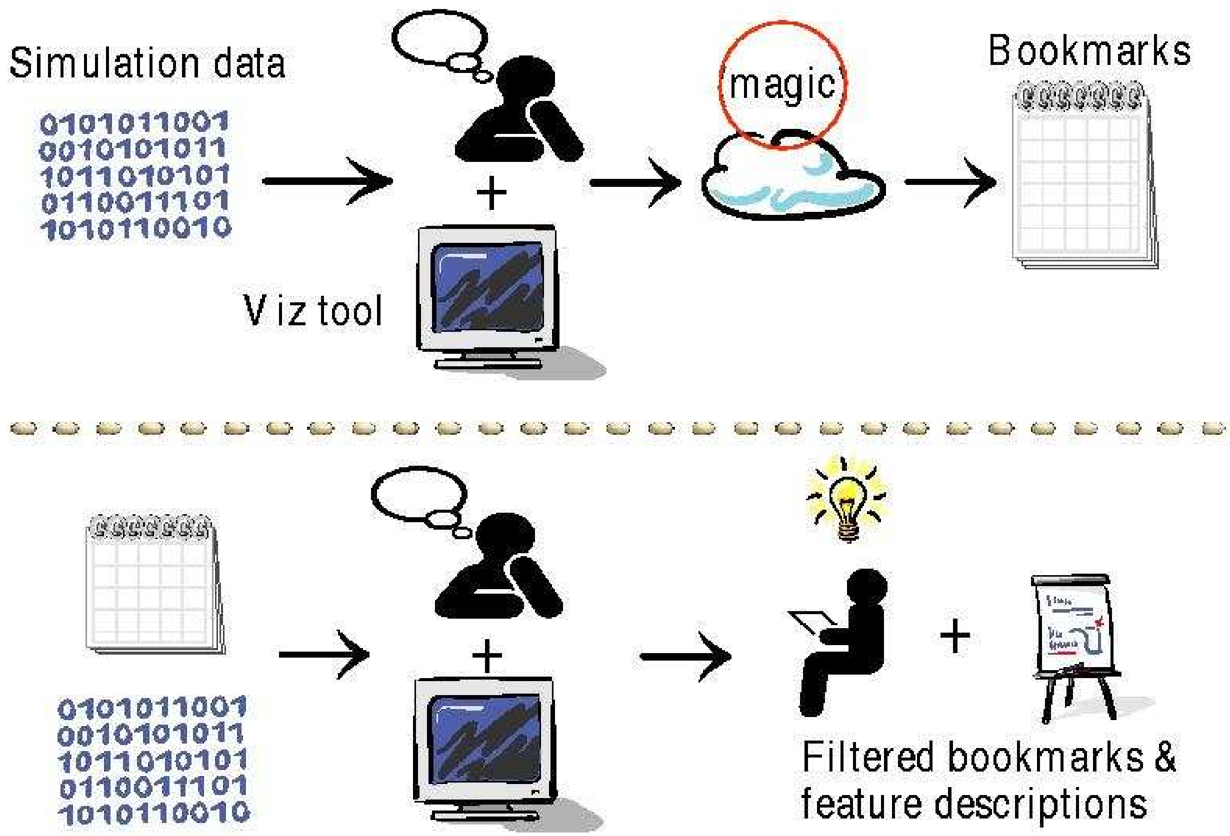



Attribute space partitioned.



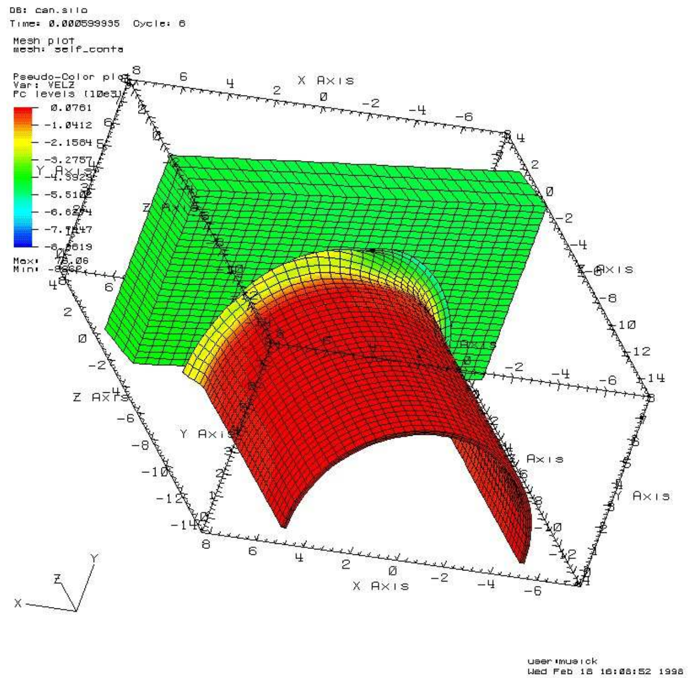

Decision tree representation.

# Pattern Recognition for ASC



# ASC Data is Daunting For Pattern Recognition

- Modern scientific data is:
  deeply skewed, ill-suited, noisy, and wrong.
- ASC data is all that and more:
  - Optimal for simulation, not for feature detection.
  - Highly redundant.
  - Terascale and partitioned.
  - "Interesting" is often the most useful label.
  - Unrelenting.




Simulation variables at every node in the mesh are processed by pattern recognition.

# What to Do?

Give up on the craftsman model of pattern recognition.

Sandia has developed a *commodity* model:

- Accepts data as it is.
- No user tuning required.
- Robust in the face of noise.

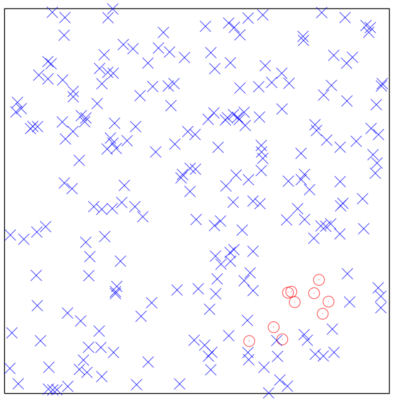
How? Some guiding principles:

1. Use *decision trees* over other methods.
2. Use *ensembles* of decision trees.
3. Embrace *redundancy.*
4. Emphasize *screening.*

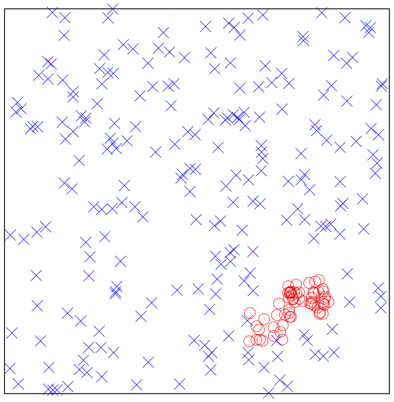
1 was mildly controversial;
2 and 3 *reverse* basic pattern recognition assumptions.

## SMOTE for Skew Populations

- Synthetic Minority Oversampling TEchnique[5].
- Oversample the minority population, but . . .
  . . . simple oversampling induces pathologies.
  So: add *synthetic* samples.
- Method:
  - Pick minority sample.
  - Pick a nearby neighbor.
  - Add new minority sample at a random point between them.
  - Repeat.



Minority class overwhelmed.



Minority class filled out by SMOTE.

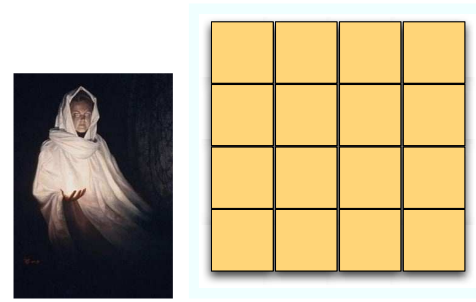
# Ensembles: Democracy Over Meritocracy

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

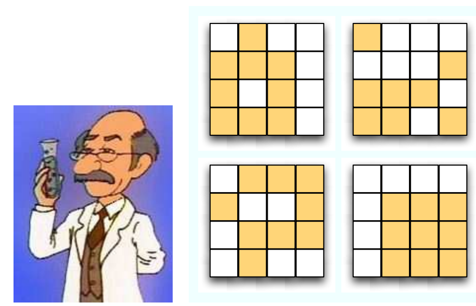
**Ensembles:** Use randomized 100% of
training data to build an expert.
Repeat to build many experts.
Vote them.

**Sandia:** Use a semi-random 1% of the
training data to build a "bozo".
Repeat to build very many bozos.
Vote them.

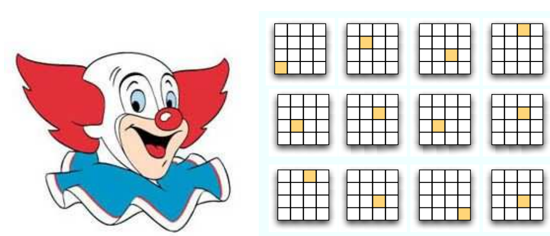
The experts beat the sage[2].
The bozos beat the experts[6].
How?
Averaging reduces measurement error.



Sage sees all the data.



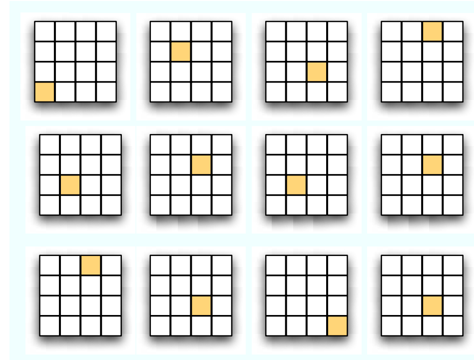
Each expert sees 2/3rds of the data.



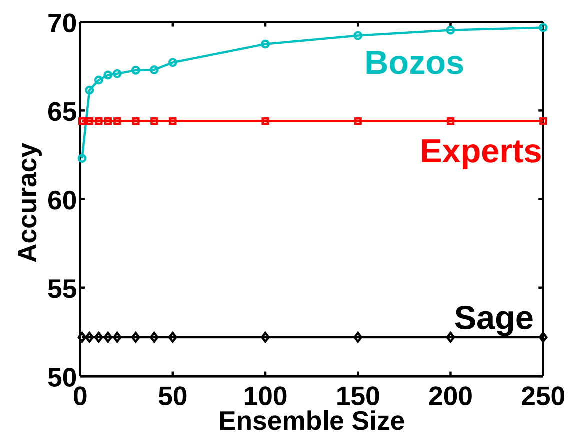
Each bozo sees a tiny fraction.

# Ensembles of Bozos for Distributed Data

- Build separate ensembles on distributed data.
- Use "improvement voting"[6].
  - $e(b)$ is estimate of error rate of $b$ bozos.
  - For (b+1)'st training set:
    - ∗ Accept all misclassified samples.
    - ∗ Accept correct samples with Prob $= e(b)/(1 - e(b))$
- Speed: $O(f \times b \times n \times \log n)$; bozos can be *faster* than sage, as well!



Bozos extracted in parallel.




Sample bozos, experts, and sage results[6].

## Conclusion: Commodity Fixes for Data Challenges

| Problem | Addressed by |
|---|---|
| Partitioned, terabyte data | ensembles of bozos |
| deeply skewed, | SMOTE |
| ill-suited, | decision trees, screening |
| noisy, | decision trees, ensembles, screening |
| and wrong | ensembles, redundancy, diversity |

- General purpose methods (principles, algorithms, and code) to handle data sets that overwhelm standard methods.
- Broadly applicable; already in use on intelligence applications.
- Shared within Sandia via the AVATAR Tools package, more broadly via the open source OpenDT[1], and through frequent publication[9].

# References


[1] Banfield, R., et al. OpenDT home page. http://opendt.sourceforge.net.

[2] 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.

[3] Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. Decision tree ensemble creation techniques. Accepted for publication on May 12, 2006, by IEEE Transactions on Pattern Analysis and Machine Intelligence, paper # TPAMI-0695-1205.

[4] Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. Ensemble diversity measures and their application to thinning. *Information Fusion Journal 6*, 1 (March 2005), 49–62.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (2002), 321–357.

[6] Chawla, N. V., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research 5* (2004), 421–451.

[7] Condorcet, N. Essai sur l'application de l'analyse à la probabilitè des decisions rendues à la pluralite des voix. Correspondence, 1785. Paris.

[8] Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification.* Wiley-Interscience Publication, 2000.

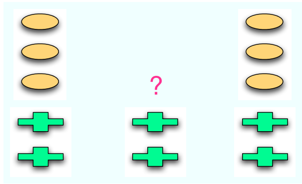
[9] Kegelmeyer, W. P., et al. AVATAR home page. http://morden.csee.usf.edu.

[10] Kolda, T., Dunlavy, D., and Kegelmeyer, W. P. Multilinear algebra for analyzing data with multiple linkages. Submitted to Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, April 2006.

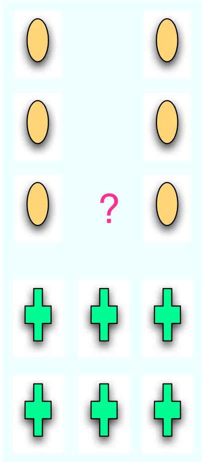
# Background Slides To Follow. . .

# Decision Trees Over Other Methods

- "No Free Lunch"[8] says the method doesn't matter . . .
  but only true for *clean* data!
- Most methods require a attribute distance metric . . .
  so attribute normalization matters.
- Decision trees don't need distance metric.
  - Use ordinal relations only.
  - Attributes need not be normalized.
  - Also, immune to noise attributes.
- With ensembles, no need to prune[6].




Unknown assigned differently . . .

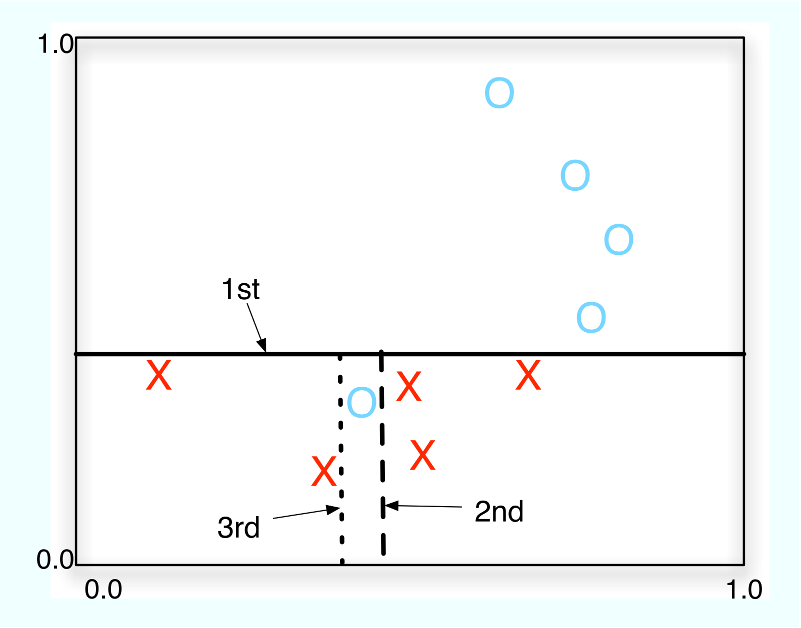



. . . depending on scaling

# Decision Trees and Distance Metrics

- How to partition attribute space?
- For the current population:
  - Consider each attribute separately.
  - Consider each threshold for that attribute.
  - Pick attribute and threshold which "best decreases impurity".
  - Use them to partition the data into two child data sets.

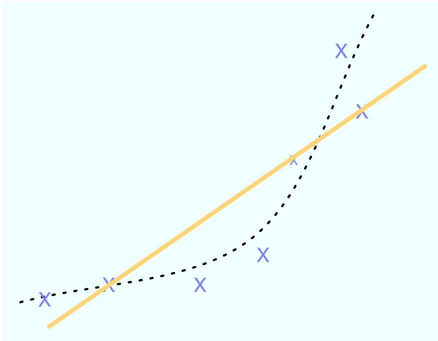
  Repeat with each child.
- Best attribute and threshold is *independent* of scaling.
- Irrelevant attributes ignored in the presence of relevant attributes.



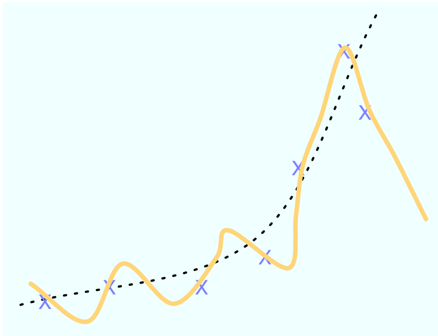

Attribute space partitioned.

# Why Do Ensembles Work? (A)

- A statistical model is a *noisy* model of reality.
- Bias error:
  Model too simple, underfits.
- Variance error:
  Model too complex, overfits.
- Bias/variance is a trade-off.
- Ensembles:
  - Use methods with low bias...
    but high variance ...
    and average to reduce variance!
- Out-of-bag validation picks ensemble size[3].
- Result:
  low bias error *and* low variance error.
  No hand tuning needed.



Too simple a model underfits the data.

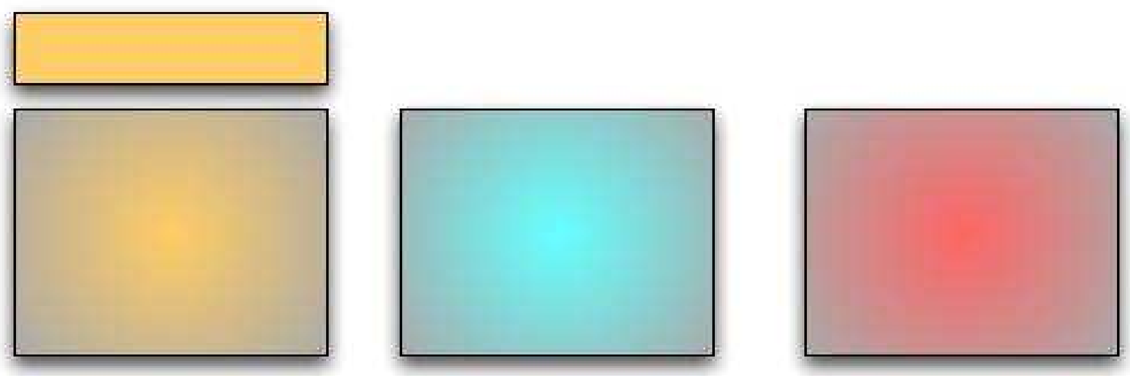


Too complex a model overfits the data.

## Why Do Ensembles Work? (B)

One key is *diversity* [7].

Imagine: three classes, each bozo only 10% accurate, and when wrong, chooses at random among the three classes.
Then the horde of bozos is perfectly, 100% accurate!



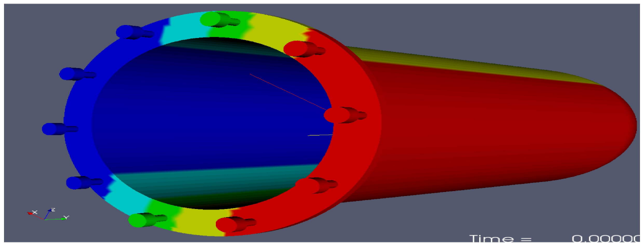
One group of unconfused bozos amid the foggy error.

Note: diverse, *random* error is difficult to achieve[4].

## Next: Inconsistent Class Statistics

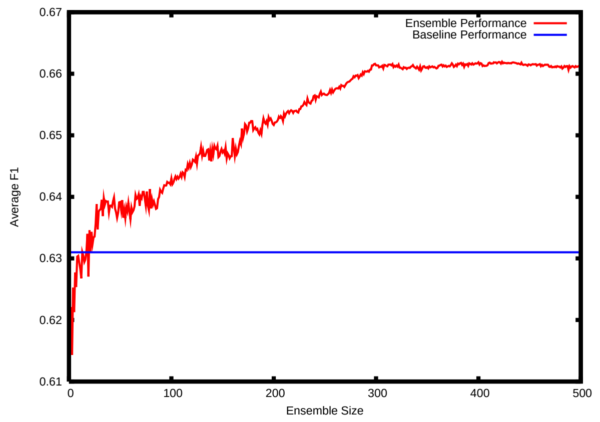
- ASC data is partitioned *and* varies in class statistics.
  - Grow ensembles of bozos on each partition.
  - *Each* ensemble generates a vote.
  - Each vote is weighted by priors:

$$
\begin{aligned}
p(w_i|x) &= \text{percentage of ensembles that vote for } w_i \text{ given } x. \\
P(w_i) &= \text{percentage of ensembles which have seen class } w_i. \\
\text{Classify as } w_m &: argmax_n\left(\frac{p(w_i|x)}{P(w_i)}\right)
\end{aligned}
$$

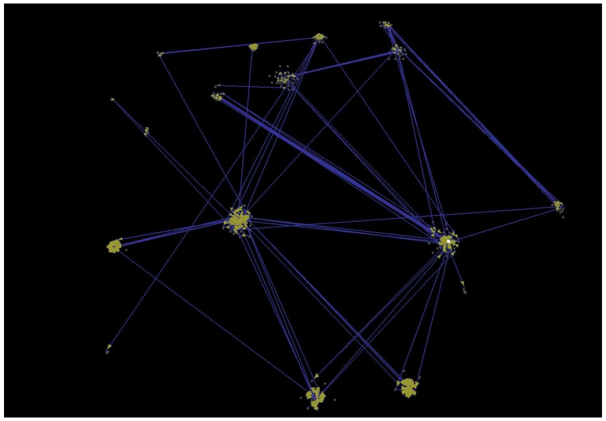


# Impact: Text, Graphs, and Intelligence Analysis

- Intelligence data is often relationship data, and graphs encode relationships.
- Text pattern recognition:
  - Why? To auto-populate graphs.
  - "NER" is phrase classification.
  - Significant improvement on contest data.
- Graph pattern recognition:
  - Classify nodes, edges.
  - Find missing links, subgraphs.
  - Tensors for multilink analysis[10].
- Also, ensembles ease data sharing.




NER improves with ensemble size.



Example multilink graph.