

# Counter Adversarial Data Analytics (CADA)



Philip Kegelmeyer, [wpk@sandia.gov](mailto:wpk@sandia.gov)



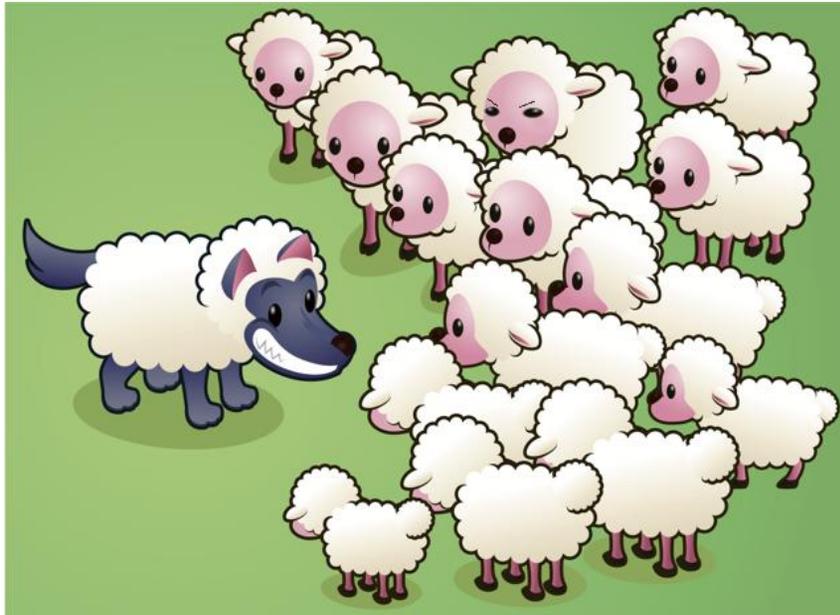
*Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.*



June 2, 2014



# Counter Adversarial Data Analytics (CADA)



IF (white AND fuzzy) Then <Harmless>

- Goals:
  - Discover generalizable, quantifiable counter-adversarial principles.
  - Specifically: investigate a) robust, b) predictive, and c) dynamic defenses.
  - Convert them to relevant, realistic methods with practical implementations.

Sandia makes **critical use of data analytics**, which our adversaries therefore **seek to sap, even suborn**.

Through **understanding our methods**, they seek to produce data which is evolving, incomplete, deceptive, and otherwise **custom-designed to defeat our analysis**.

We **cannot prevent this**: we frequently must depend on data over which our adversaries have extensive influence.

We will thus develop and assess novel data analysis methods to **counter that adversarial influence**.



## Philosophy and Programmatics



“We must learn to love life without ever trusting it.” (G.K. Chesterson)

⇒ “We must learn to love ~~life~~ data without ever trusting it.”

CADA wants to turn this into *quantified*, practical advice.

- 
- A “Data Sciences Research Challenge” incubation project.
  - 1.5 years; April 2013 to September 2014
  - Nascent external work on the effects of data tampering[4, 5].

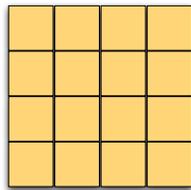


# Review: Ensemble Machine Learning

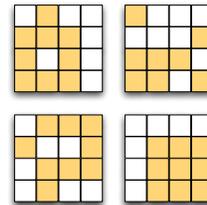


Start with “groundtruth” training data:  
each training sample has attributes and *trusted* labels.

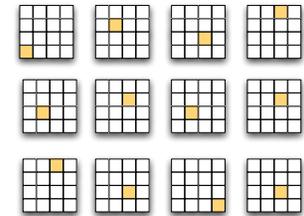
Sage sees all the data.



Experts see diverse subsets.



Each bozo sees a tiny fraction.



The experts beat the sage[1]. The bozos beat the experts[2].

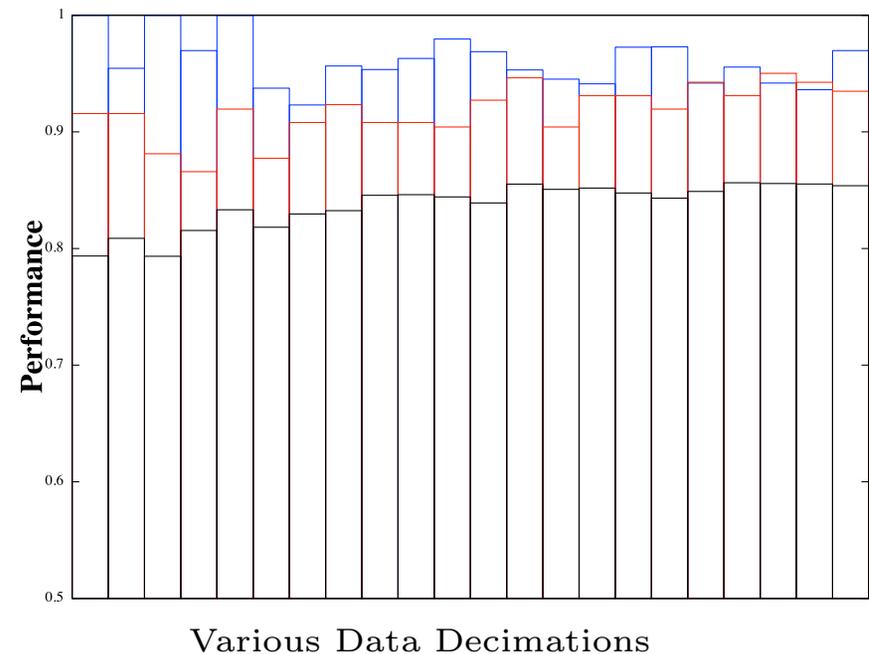
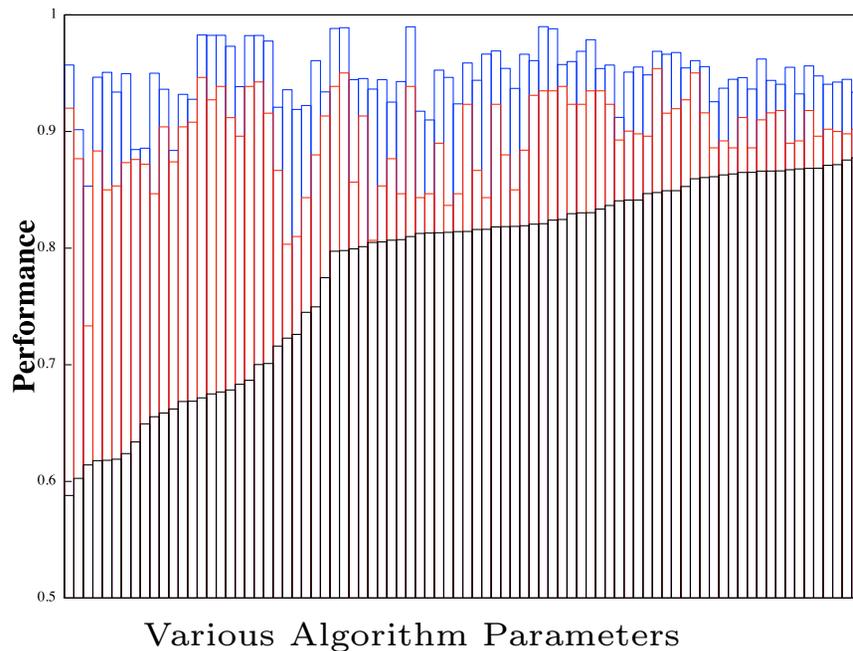


# Review: Performance Assessment Expectations



Typically, one expects:

- **self-assessment on the training data** to be an optimistic estimate ...
- ... of **ensemble performance on test data**, which in turn is better than
- ... **non-ensemble performance on test data**.

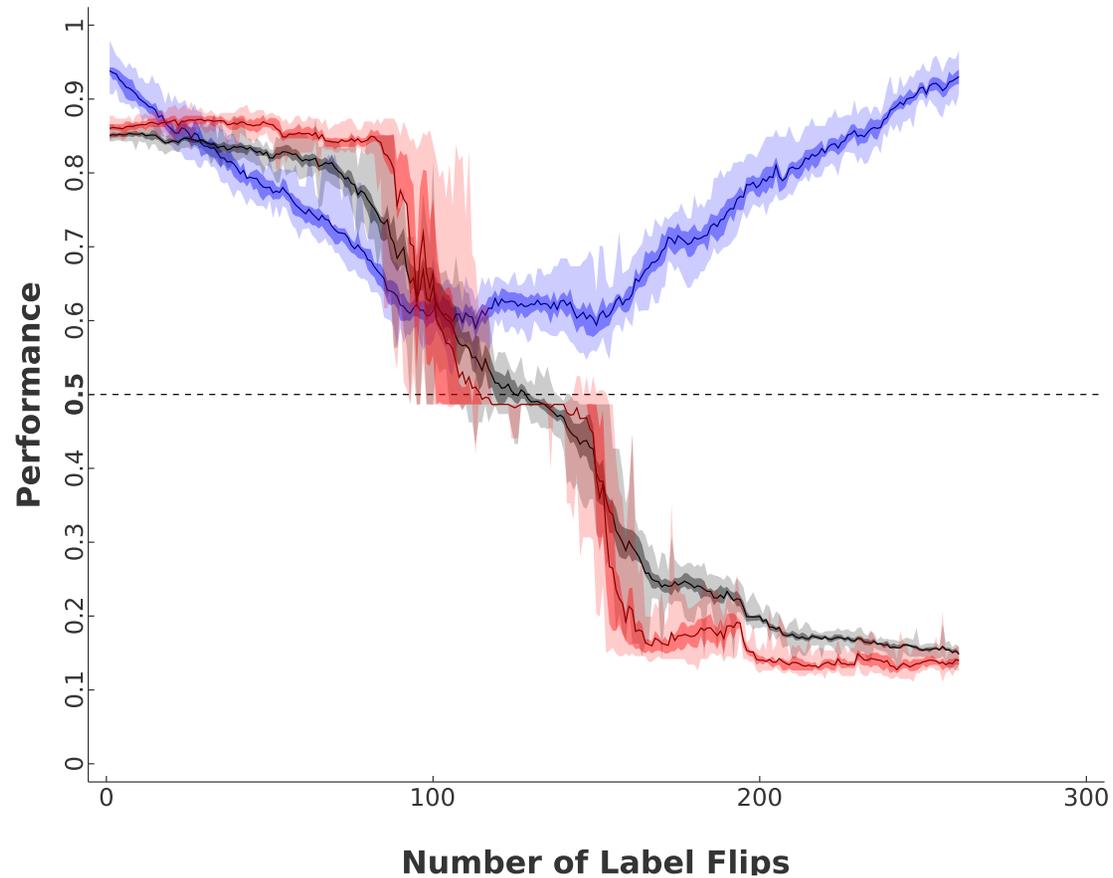




# Random Tampering



Mindless, random flipping of labels is effective enough.



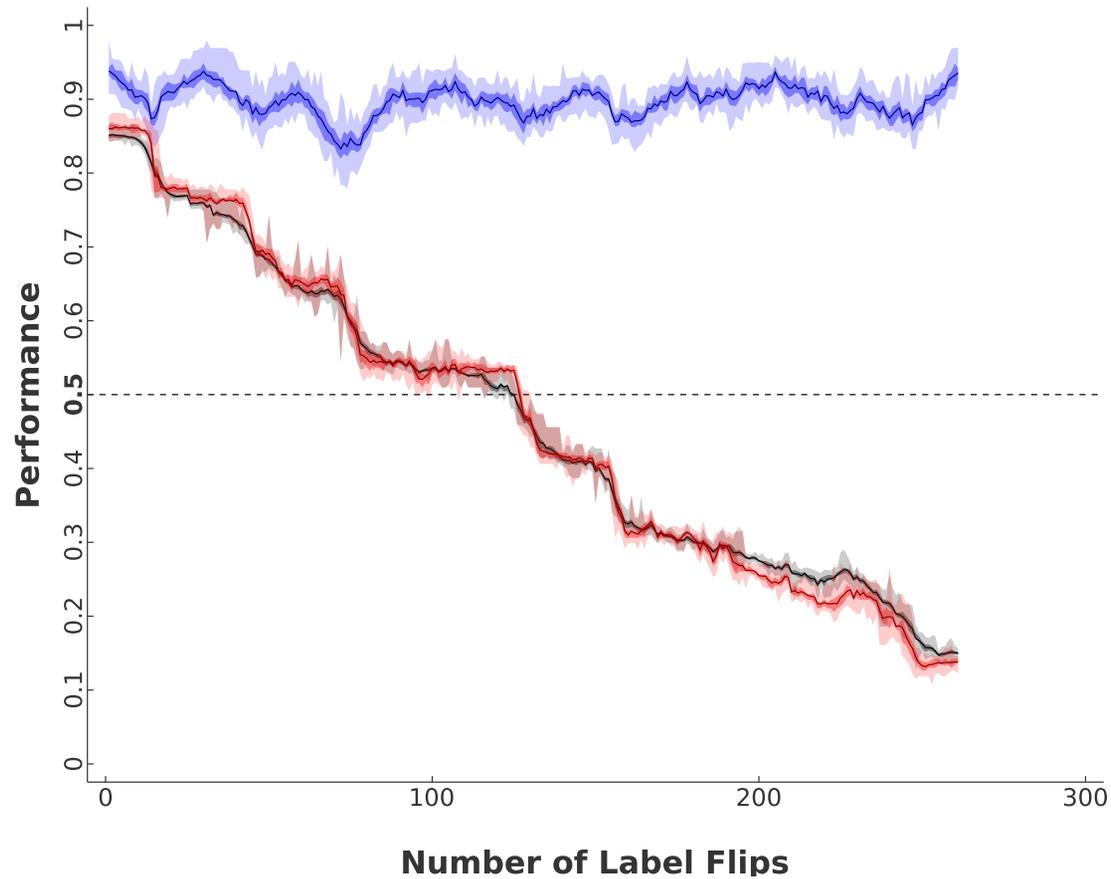
NIF optics inspection (OI) data



# Attack Clusters One at a Time



Smarter attacks suppress the cross-validation “dip” signature.



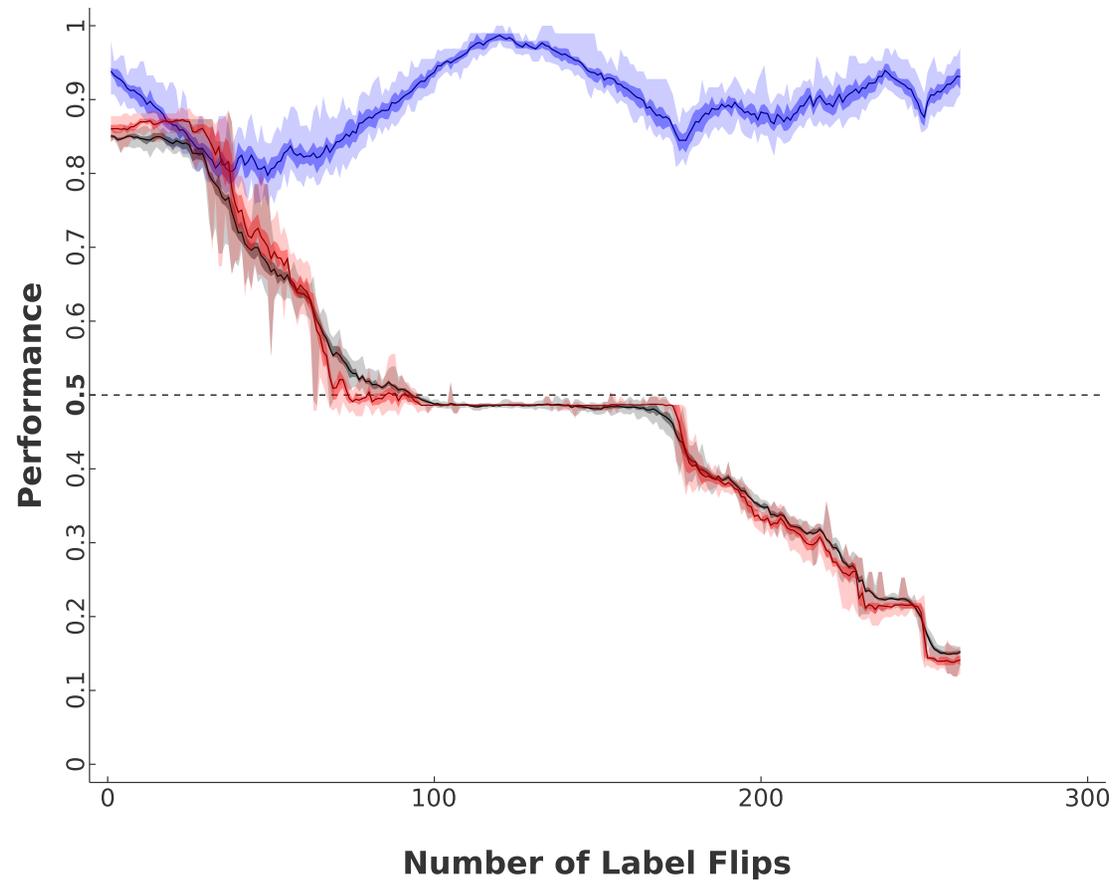
NIF optics inspection (OI) data



# Attack Statistically Significant Samples



Or smart attacks drive down accuracy faster, with less tampering.



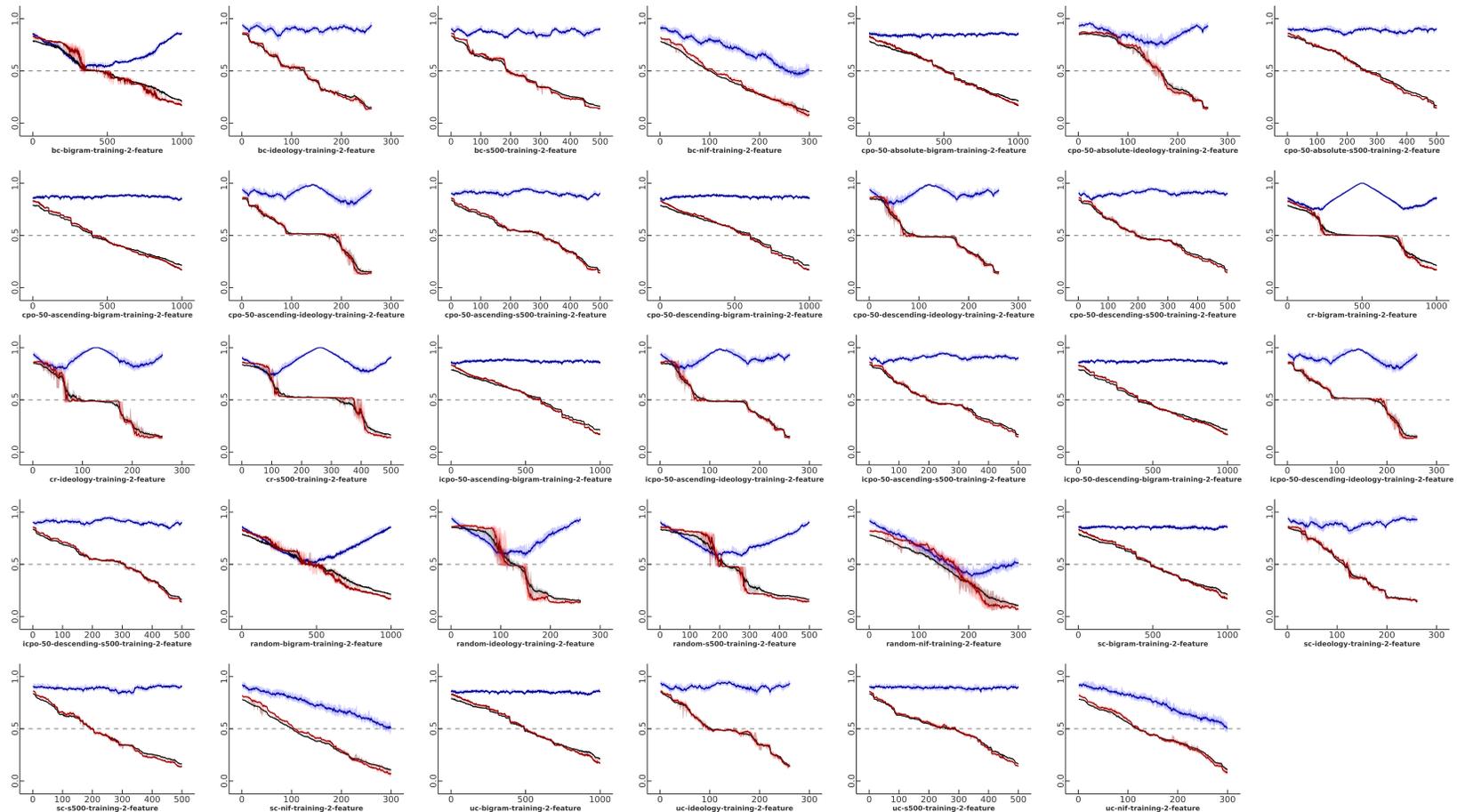
NIF optics inspection (OI) data



# We've Invented Many Attacks; More Coming



Plus new methods, and metrics, for quantification and visualization.

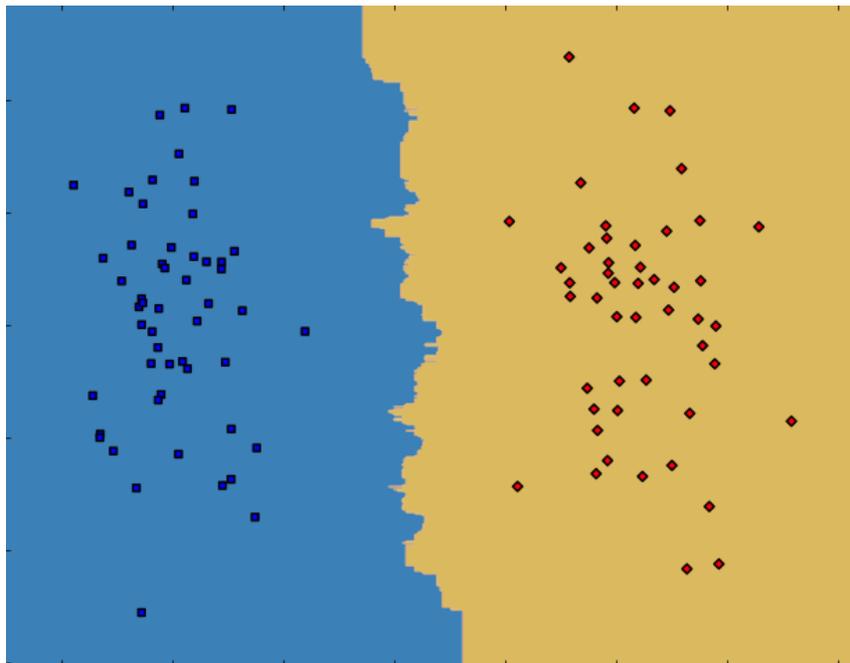




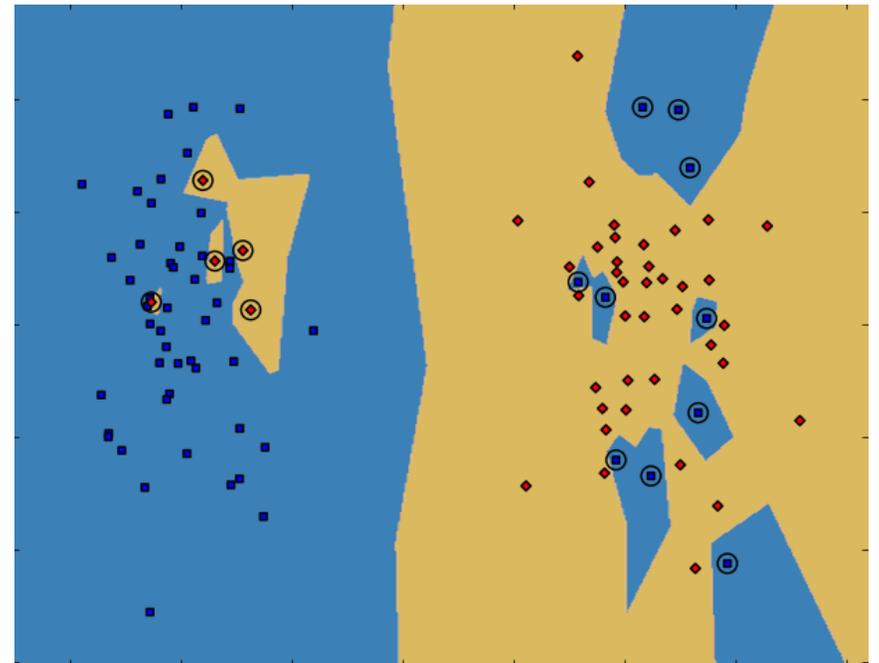
# Is There Any Way to Mitigate the Damage?



## EOM: Ensembles of Outlier Measures



No Label Tampering



15% Label Tampering

Circles indicates points whose label has been tampered with.

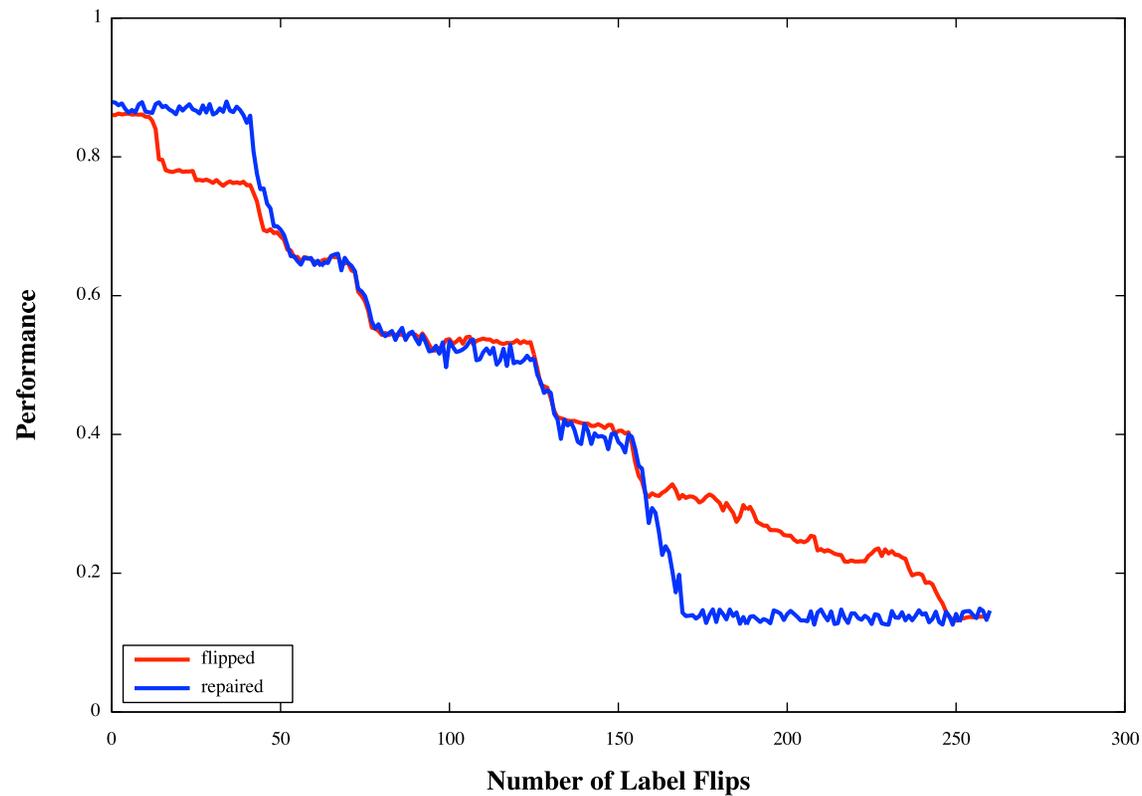


# Detect and Repair Tampering



“Flipped”: The tampered data.

“Repair”: wherever tampered labels are detected, *correct the label*.



NIF optics inspection (OI) data



## Outcomes to Date (Since April 2013)



- Lots and lots of work not discussed today:
  - predictive analytics for *anticipating* adversarial tampering
  - proofs and quantitative analysis for the trade-offs involved in “moving target” defenses
  - a provably pessimal attack against regularized least squares
- Four conference papers already delivered (two invited)[3, 6].
- Two more conference submissions accepted, two in review.
- One journal paper in review, more in preparation.
- Open source release of visualization software (<http://github.com/sandialabs/toyplot>).



## Summary



- We are investigating the effect of well-prepared or insider adversaries on data analytics, starting with machine learning.
- The results are worrisome. Quantifiable, but worrisome.
  - Standard self-assessment methods are quickly led astray.
  - Good attacks not only tank accuracy, they waste time.
- We've discovered some rays of hope.
  - Tentative attacks can actually help the defender. An adversary must invest to be effective.
  - Ensembles can provide cheap protection from highly tailored attacks.
  - EOM (Ensembles of Outlier Methods) can help remediate some attacks.
  - Currently working on a statistical test for distinguishing random vs sentient label noise: quantified paranoia.



## Questions?



- Are such tampering attacks realistic? Is there historical precedence?
- What happens with simpler, less agile classifiers?
- What about “evasion” attacks, such as those of interest to an advanced persistent threat adversary?
- What sort of skills have proven useful in thinking about all this?
- What’s next?



## End Notes



### Collaborators

- Sandians: Tim Shead, Jon Crussell, Dave Zage, Katie Rodhouse, Dave Robinson, Warren Davis, Justin (JD) Doak, Jeremy Wendt, Curtis Johnson
- Ex-Sandians: Rich Colbaugh, Kristin Glass, Brian Jones, Eugene Yevgeniy, Jeff Shelburg

### 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] 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.
- [3] COLBAUGH, R., AND GLASS, K. Moving target defense for adaptive adversaries. In *2013 IEEE International Conference on Intelligence and Security Informatics (ISI)* (2013), IEEE, pp. 50–55.
- [4] DALVI, N., DOMINGOS, P., MAUSAM, SANGHAI, S., AND VERMA, D. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2004), KDD '04, ACM, pp. 99–108.
- [5] LASKOV, P., AND LIPPMANN, R. Machine learning in adversarial environments. *Machine Learning* 81, 2 (2010), 115–119.
- [6] VOROBEYCHIK, Y., AND WALLRABENSTEIN, J. R. Using machine learning for operational decisions in adversarial environments. In *Proceedings of 13th International Conference on Autonomous Agents and Multiagent Systems* (May 2014), Association for Computing Machinery. ISBN 978-1-4503-2738-1.