# Big Data, Bigger Audience: A Meta-algorithm for Making Machine Learning Actionable for Analysts

Dylan Cashman
Tufts University
dylan.cashman@tufts.edu

Stephen Kelley
MIT Lincoln Laboratory
stephen.kelley@ll.mit.edu

Diane Staheli
MIT Lincoln Laboratory
diane.staheli@ll.mit.edu

Cody Fulcher
MIT Lincoln Laboratory
cody.fulcher@ll.mit.edu

Marianne Procopio
MIT Lincoln Laboratory
procopio@ll.mit.edu

Remco Chang
Tufts University
remco@cs.tufts.edu

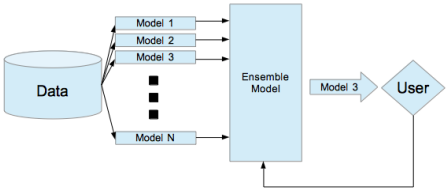
## Abstract

In the current security climate, much more data is being accumulated from sensors than can be realistically viewed by analysts. Statistical methods offer promising solutions to the data filtering problem. However, many statistical methods are a black box, and their opacity makes it difficult for the analyst to infer any actionable solution from the methods' suggestions.

We present a meta-algorithm for building interpretable data analysis tools by borrowing the concept of small multiples from visualization research and applying it to statistical modeling[8]. This technique, rather than producing a single, high-dimensional data model, builds a multitude of small, interpretable models used in an ensemble to analyze data. Aside from visualization, similar concepts have shown theoretical merit[3, 4]. We show a prototype used to analyze high-dimenional network traffic data, and demonstrate its effectiveness on the VAST 2013 MC3 dataset.

**Index Terms:** K.6.5 [Management of Computing and Information Systems]: Security and Protection—; H.5.2 [Information Interfaces and Presentation]: Theory and methods—


(a) A traditional visual analytics pipeline.




(b) Our proposed pipeline. Different subspaces of the data are used to train multiple kinds of models, which are then combined to make an ensemble judgement. The user only sees those small models that contributed most to the judgement.

## 1 Introduction

In 2015, the United States Department of Defense listed the first priority of the organization as training and maintaining a cyber workforce[2] The number of working cybersecurity analysts will likewise grow exponentially. At the same time, their job becomes harder; The adversarial nature of the game between attacker and defender has meant that attack patterns have grown monotonically more sophisticated. One potential solution lies in using newer results from the Machine Learning and Statistics communities that are able to account for unseen patterns. However, many of these methods require a deep understanding of the statistical mechanisms underneath to tune and interpret. In addition, cyber analysts have traditionally had a lack of trust in statistical solutions as they can lead to nonsensical false positives[7]. In this work, we propose a meta-algorithm for integrating modern, robust machine learning techniques into an interpretable system by constraining the statistical models we use. Instead of using a single, complicated model that incorporates all features of the data to answer a single decision problem, we generate many small, interpretable models and then use ensemble techniques to combine them.

We developed a proof-of-concept system to demonstrate the benefits of adaptable ensembles of interpretive models in the cyber security domain. We analyzed three months of network traffic data from a private corporate network of approximately 450 hosts with the intent of finding anomalous activity. To accomplish this, we first generated numerical features from the data. Next, we used Gaussian Processes to find time series anomalies and Gaussian Mixture Models to find cluster anomalies. Then, we used a multiplicative weight update based on implicit user feedback to combine the models into a single anomaly detector.

## 2 Framework

Suppose we have a data analysis task in $n = 10$ dimensions. A standard approach would be to build a model in 10 dimensions to optimize for statistical accuracy. Instead, we generate a large group of simple models that cover orthogonal sections of the data with the assumption that this *model mesh* approximates a more complex model.

To generate this model mesh, we partition the data into 1- and 2-dimension subspaces, and for each subspace, we train several types of machine learning models. These models are chosen with two priority: they should be *orthogonal* in that they should cover different qualities of the data, and they should have some interpretable output. Then, we utilize techniques from ensemble learning to integrate these models into a single model for the data task at hand[1, 5]. When that ensemble model offers a solution, it shows the user the small model that most contributed to the solution. In that way, we are able to cover many qualities of the data in our statistical analysis, but we only expose interpretable models to the analyst. Lastly, we include a feedback mechanism to refine these confidence measures to better align with a user's mental model. In this manner, we find statistically sophisticated solutions to questions in high-dimensional spaces that are interpretable and adaptable to the particular user. Our meta-algorithm is illustrated in Figure 2.

Figure 1: A visualization with idealized data of the framework presented in this paper. Note the 3 main visual components: first, a time series view of features displays information about the raw feature data. On the bottom left, there exists a list of time windows that open an existing interface for inspection of raw data. A series of contextualized anomaly detections stretches in the remaining space along the bottom of the display to collect feedback on cue-relevance for adaptation.

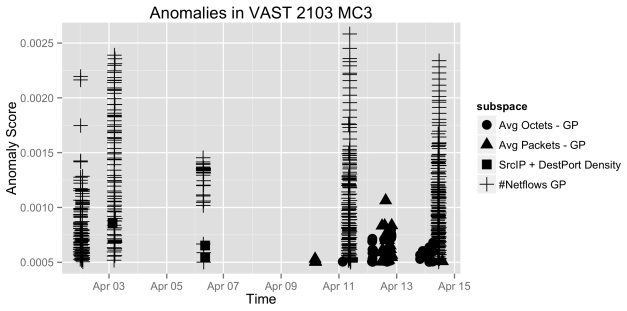



Figure 2: A demonstration of anomalies presented over time within the VAST 2013 MC3 data set. The subspace indicative of each anomaly is indicated by different shapes. This statistical detections found within this framework follow the ground truth story provided with the data closely.

## 3 Applications

### 3.1 Anomaly Detection in Network Traffic Data

To make an example implementation, we tackled the problem of anomaly detection in network traffic data. The dataset was a proprietary collection of network traffic collected at a corporate network of approximately 450 hosts. We started with information on individual connections, and aggregated the data into 30-second intervals. The raw data consisted of IP and port information of the source and destination of each connection, as well as payload size. For each 30-second interval, we generated 9 features: number of unique source/destination IP and port connections, as well as entropy scores[6] for the distribution of source/destination IP, port, and country of origin. Additionally, payload size and number of packets were averaged over connections occurring within each 30 second window. The particular features used in this example were chosen by domain experts as potentially interesting features in the data source.

Given these features, the framework seeks to identify 30 second windows of interest for additional user inspection. Cyber security analysts with expertise in identifying anomalies in raw data formed our user base.

Figure 1 shows a screenshot of the system built for our anomaly detection domain with idealized data. The goal of this system is to help the analyst identify interesting 30 second windows in the data that warrant further investigation. The system can offer suggestions of anomalous data by cueing the user to which model contributed the most to its anomaly score. These anomalies can be contextualized in a variety of ways. We show coarse, natural language contextualization on the top of each column that was generated manually after consultation with analysts on the meaning of each feature. We would also be able to show visualizations of the models, since they are low-dimensional models.

Upon presentation, the user is able to provide feedback on the suggested anomalies. If the user responds that the given model was helpful or unhelpful, an importance weighting of each model's anomaly score is adjusted for future suggestions. We utilize a multiplicative weight update that rewards models that receive positive feedback and penalize those that produce inaccurate results. In this way, the system can adapt weightings on feature subspaces and adapt to produce more accurate suggestions to a user.

### 3.2 Effectiveness on Known Dataset

In order to demonstrate the effectiveness of using multiple small models, we consider the netflow component of the VAST 2013 MC3 data set. For this analysis, we generate features in the same manner as detailed above on the proprietary data set. Here, we view the task as cueing a hypothetical analyst to examine raw data in depth. Figure 2 details all windows that have a combined anomaly score above a threshold selected manually. We can see the types of attacks evolving over time as anomalies in the indicated small models. These types of anomalies correspond to port-scans, denial of service attacks, exfiltrations, and bot-net activity, respectively, which corresponds with the solution story provided by the VAST Mini Challenge.

## 4 Discussion and Conclusion

In this paper we present a framework aimed at constructing and integrating an ensemble of complexity-constrained statistical models. The benefit of building models in this manner are increased interpretability for domain experts to utilize automated models and develop trust in them. Empirical results were given on the representative VAST 2013 MC3 netflow data that show it is effective in highlighting problems found in those slices, and that many anomalies can be found. Work on the proprietary dataset outlined in this text has also produced empirically useful results.

## References


[1] J. B. Cabrera, C. Gutiérrez, and R. K. Mehra. Ensemble methods for anomaly detection and distributed intrusion detection in mobile ad-hoc networks. *Information Fusion*, 9(1):96–119, 2008.

[2] A. Carter. The dod cyber strategy. *Department of Defense: Washington, DC*, 2015.

[3] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *SDM*, volume 73, page 93. SIAM, 2004.

[4] M. Gleicher. Explainers: Expert explorations with crafted projections. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2042–2051, 2013.

[5] P. Kothur, M. Sips, H. Dobslaw, and D. Dransch. Visual analytics for comparison of ocean model output with reference data: Detecting and analyzing geophysical processes using clustering ensembles. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1893–1902, 2014.

[6] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *ACM SIGCOMM Computer Communication Review*, volume 35, pages 217–228. ACM, 2005.

[7] R. Sommer and V. Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 305–316. IEEE, 2010.

[8] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.