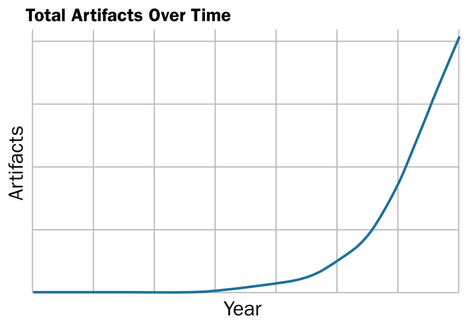
Human-Computer Decision Systems

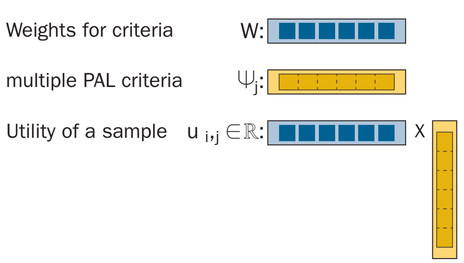
Security decision systems aim to distinguish malicious activity from benign and often use a combination of human expert and automated analysis, including machine learning (ML). Systems using only human experts scale badly; pure ML systems are susceptible to structured attack by adversaries and, in most cases, have unsatisfactory performance on their own.

- Many operational security problems depend on a small number of skilled analysts to process a large and growing firehose of potentially malicious data.
- Traditional active learning tries to address this situation by suggesting allocation of limited analysis resources that optimize the convergence of a machine learning classifier.

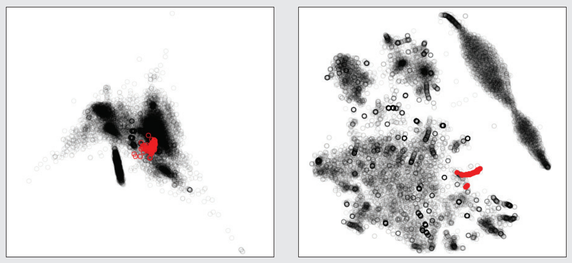
Growth of CERT Artifact Catalog



Dynamic Proactive Learning



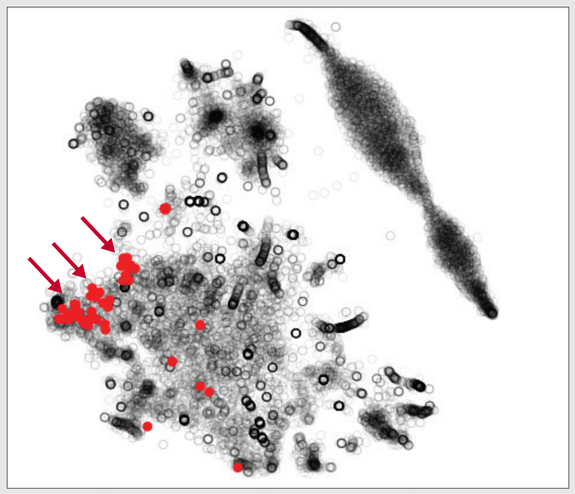
How good is your cheap feature?



PCA (for comparison only) t-SNE

- 20k observations of 545 mnemonic counts reduced to two dimensions.
- Red points are a specific IAT hash of interest.
- This IAT hash (cheap) is well localized in t-SNE space (expensive)
- Knowing this IAT hash is likely good enough to define this family.
- Expert analysis concludes this is a single family.

Cheap can be noisy… a different IAT hash

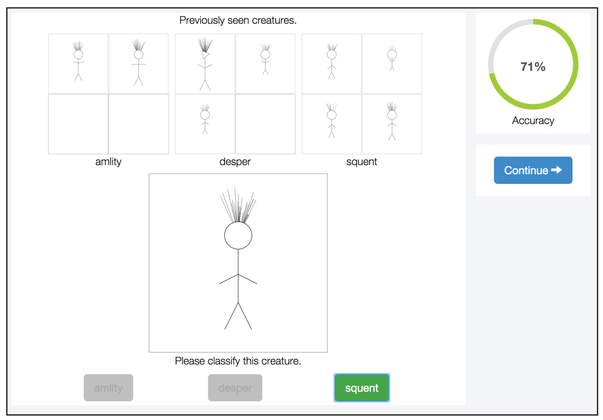


Embedding reduces the number cases to reverse engineer and increases confidence

- Current analyses methods conclude this IAT hash is one family.
- **t-SNE + IAT = family would have cost less.**

Result: t-SNE-based visualizations paired with IAT section hashes greatly reduce the number of manual binary analyses required to understand new groups of binaries

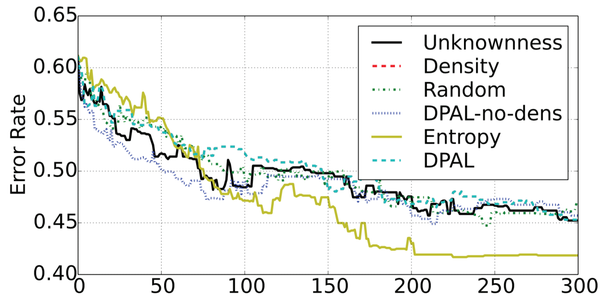
- The human-computer collaboration model will improve upon traditional active learning by optimizing not simply for convergence of the ML component, but also for future performance of the overall system, including mutable human analysts.
- We test the performance of new models not only through simulation, but also through human-subject experiments.
- Because conducting these experiments using real security analysts performing their normal tasks would be prohibitively expensive, we instead developed a proxy problem of identifying fictional creatures and leveraged non-experts on Amazon's Mechanical Turk platform. The process of generating the fictional creatures adheres to the statistical distributions of real malware classes.



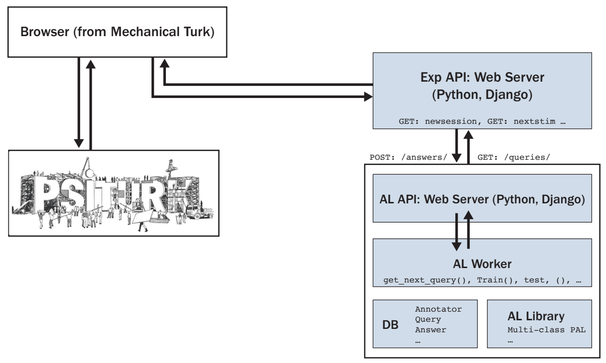
A screenshot of the experimentation system built using Mechanical Turk and Psiturk.

DPAL provides a framework to combine multiple factors in choosing points, including factors related to analyst performance. It shows promise in simulation and will be put to the test in a human-subject experiment.

DPAL with Real Users



- Entropy (very simple!) wins.
- **Runtime features are too discriminative** for DPAL to gain an advantage.



Future work includes joint optimization of classifier and analyst objectives, extension of the experimentation software to support multi-session and team experimental trials, and a test of transferability of the model problem results to the target domain.

To keep pace with adaptive adversaries, our cybersecurity defenses must take advantage of both machine learning and human analyst strengths. Future solutions should optimize for success of the overall system.

Contact: Brian Lindauer | lindauer@sei.cmu.edu