

Human-Computer Decision Systems in Cybersecurity

Brian Lindauer

(with Bronwyn Woods, Shane Moon, Peter
Jansen, and Jaime Carbonell)

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Copyright 2015 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

This material has been approved for public release and unlimited distribution except as restricted below.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM-0002822

Collaboration Between Human Experts and ML

Two typical approaches to classification or categorization:

Human analysts and **machine learning (ML) classifiers**.

Different strengths and weaknesses. Why pick one?

Analysts

- Flexible, adaptable
- Sensitive to context
- Ability to explain

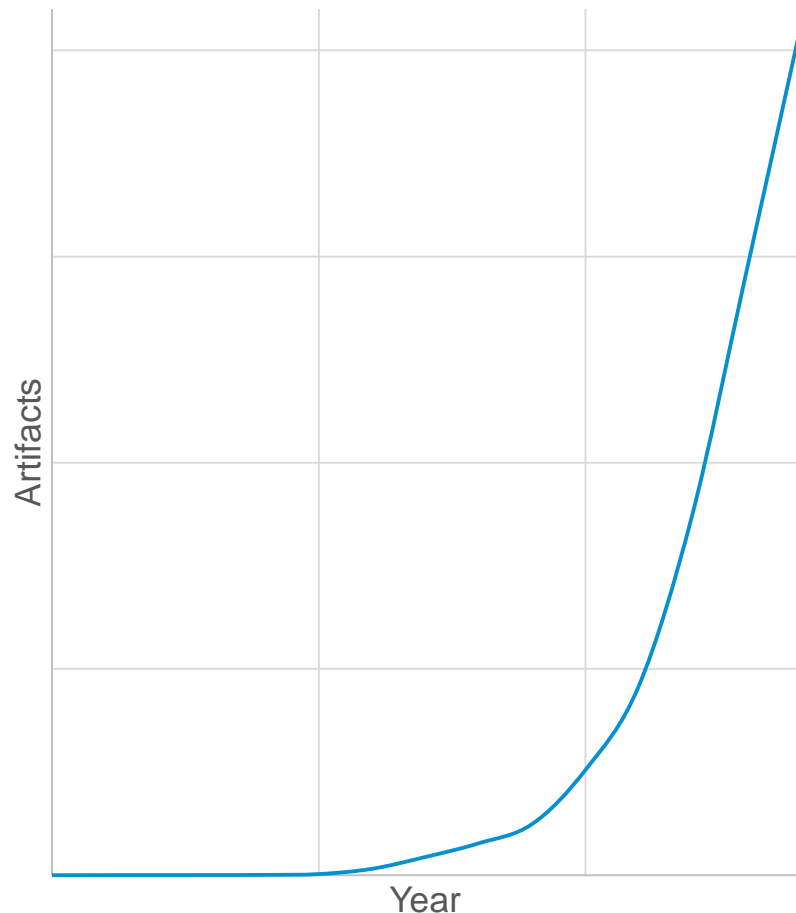
Machine Learning

- Scalable
- High dimensional
- Precisely specified



A Motivating Problem – Malware Classification

Total Artifacts Over Time



- CERT artifact catalog is a valuable resource that depends on expert reverse engineers for labels.
- Sample growth is exponential. Staffing growth is... sub-exponential.
- One-off ML models show promise, but can we do better?
- Other potential domains
 - SOC/CSIRT Triage
 - Insider Threat

Background and goals

Learning theory progress

Experimental progress

Conclusions and next steps



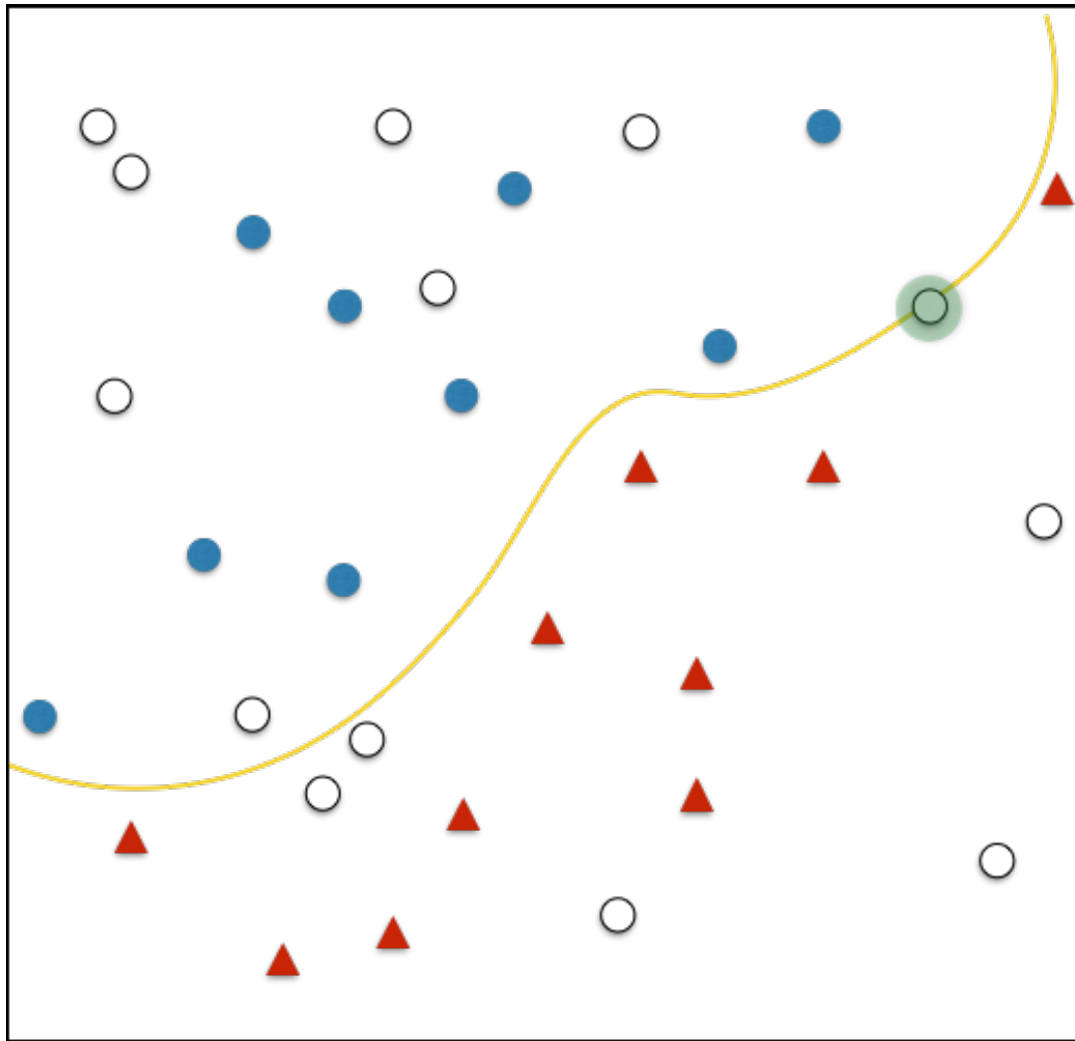
From Classifiers to Collaboration


Traditional machine learning: select a random sample to label for training data.

Active learning: the model estimates an ideal sequence of samples and gets labels.



Traditional Active Learning (uncertainty-based)



 = most uncertain

$$x_{LC}^* = \operatorname{argmax}_x 1 - P_{\theta}(\hat{y}|x)$$

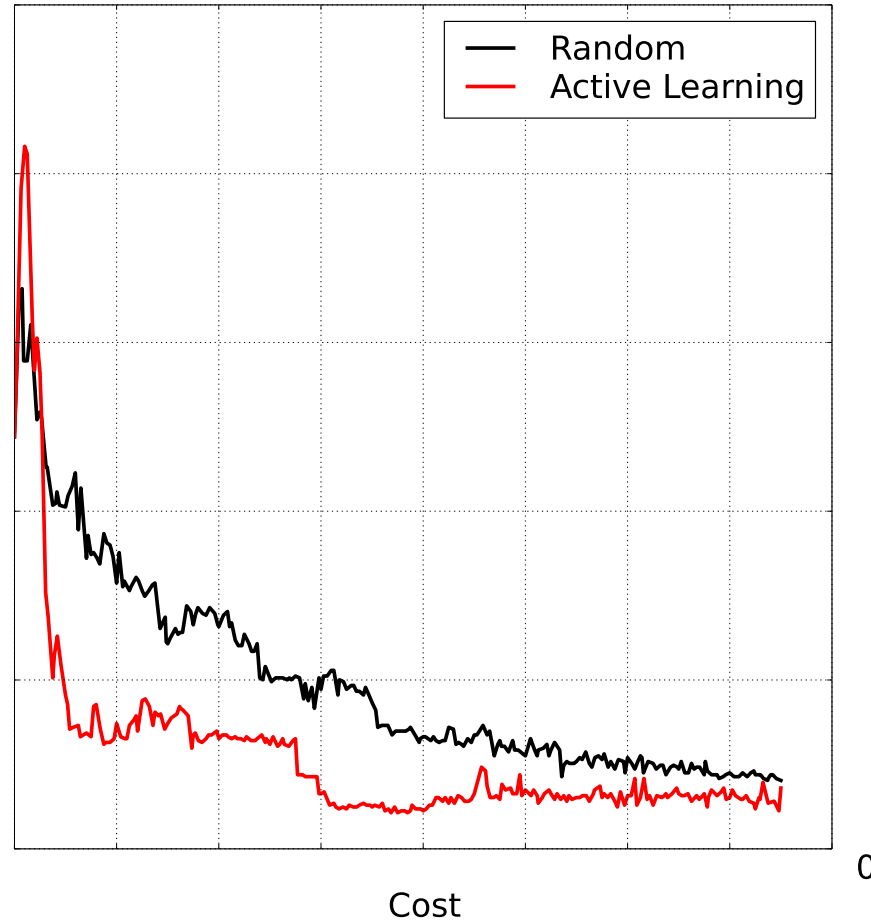
 Label 1

 Label 2

 Unlabeled

 Current
Decision Boundary

Simulated Active Learning



From Classifiers to Collaboration

Traditional machine learning: select a random sample to label for training data.

Active learning: the model estimates an ideal sequence of samples and gets labels.

Proactive learning: active learning, but don't assume the labels are perfect or perfectly reliable since they come from a human, not an oracle.

Human-computer collaboration:

The human experts are a persistent team. The algorithm estimates the best instances to show to each analyst to improve the long-term performance of both the machine and human learners.



Apparatus for HCDS Research

- How is it done today? Simulations, mostly.
- Why isn't that good enough?
 - Proactive learning and human-computer decision systems model and respond to the behavior of humans annotators.
 - Simulated annotators will not have the same behavior (errors and learning patterns) as actual human experts.
- What would we need to know whether a new approach works?
 - Realistic data: Class and feature distributions that relate to a transition domain.
 - Human participants: Actual errors and learning patterns.
 - Ground truth: Because we know labelers are fallible.



What We are Doing

Track 1: Learning theory advances to account for persistent human expert teams.

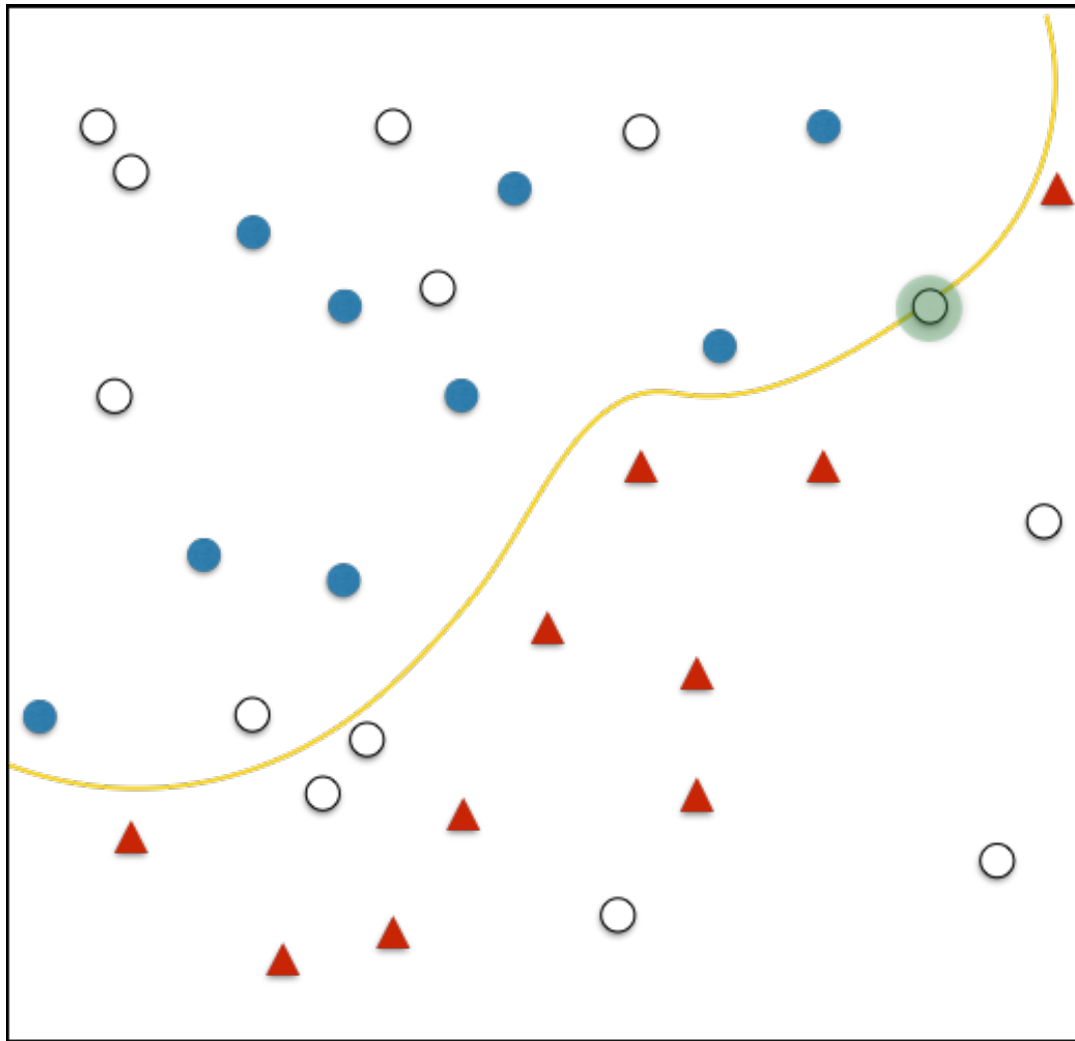
Track 2: Human subjects experiments to validate improvement to system.



Background and goals
Learning theory progress
Experimental progress
Conclusions and next steps



Traditional Active Learning (uncertainty-based)



● = most uncertain

$$x_{LC}^* = \operatorname{argmax}_x 1 - P_{\theta}(\hat{y}|x)$$

● Label 1

▲ Label 2

○ Unlabeled

— Current
Decision Boundary

Proactive Learning With Multiple Domain Experts

Problem Formulation (Objective)

$$\begin{aligned} & \max_{S \subset UL} E[V(S)] - \lambda \left(\sum_k t_k \cdot C_k \right) \\ \text{s.t. } & \sum_k t_k \cdot C_k \leq B, \quad \sum_k t_k = |S| \end{aligned}$$

S : the set of instances to be sampled

$E[V(S)]$: the expected value of information of the sampled data

C_k : cost of the chosen expert k

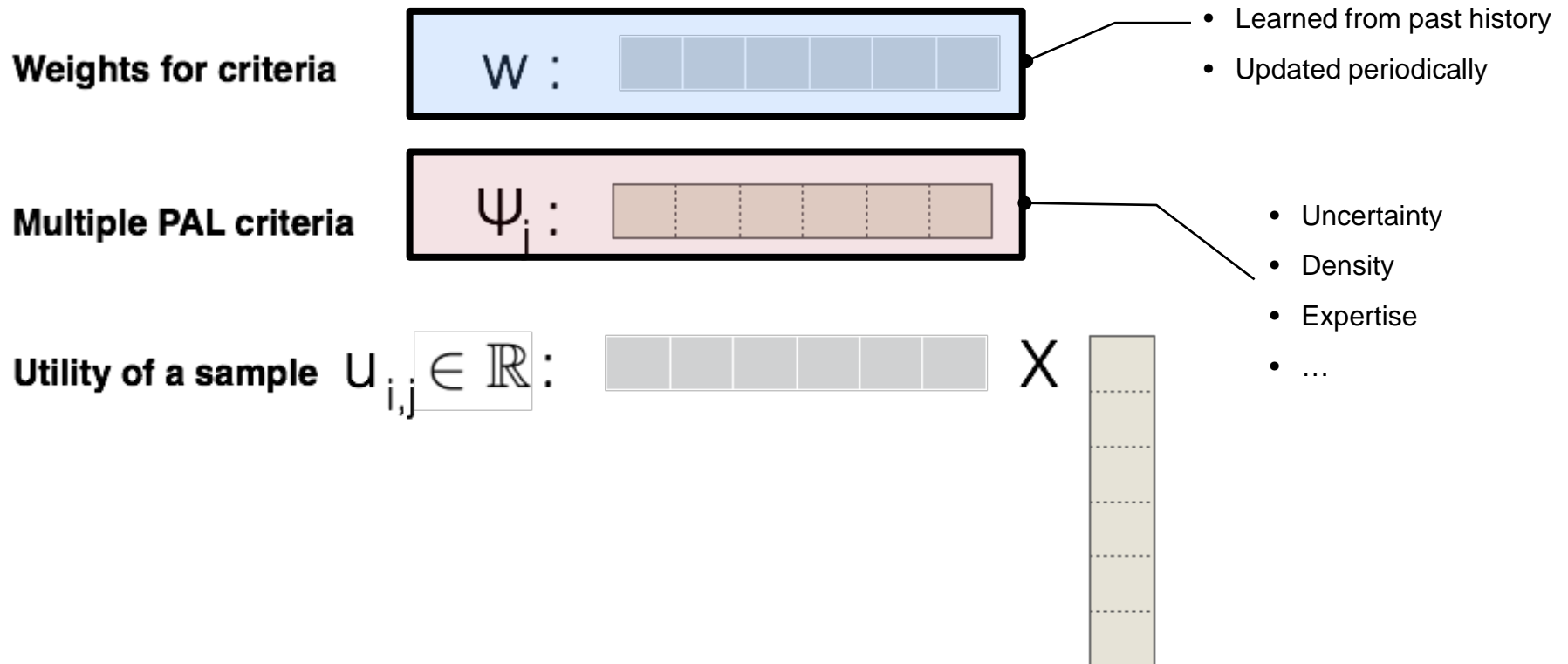
Greedy Approximation

$$(x^*, k^*) = \operatorname{argmax}_{x \in UL, k \in K} U(x, k)$$

[Moon and Carbonell, 2014]

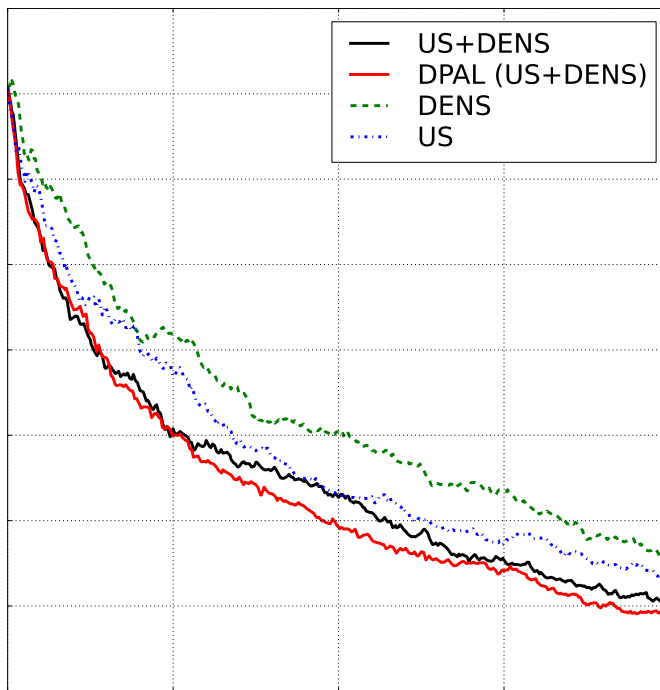


Dynamic Proactive Learning (DPAL)



DPAL is a mathematical framework to support active learning using many simultaneous criteria.

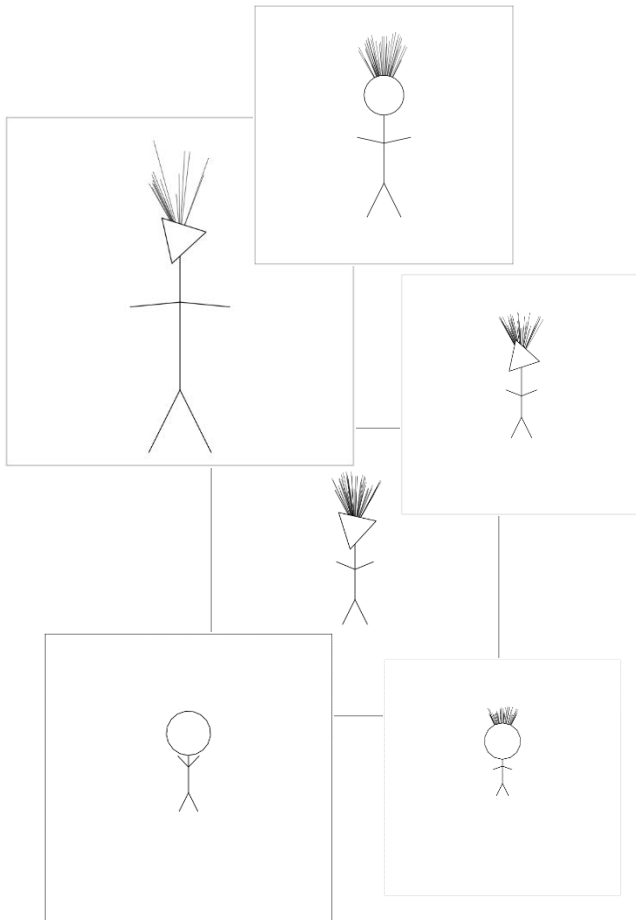
Preliminary DPAL Results



- In simulation, a simple DPAL configuration outperforms other active learning strategies.
- US = Uncertainty Sampling
- DENS = Density Sampling
- US + DENS = static weighting
- DPAL = dynamic weighting

Background and goals
Learning theory progress
Experimental progress
Conclusions and next steps

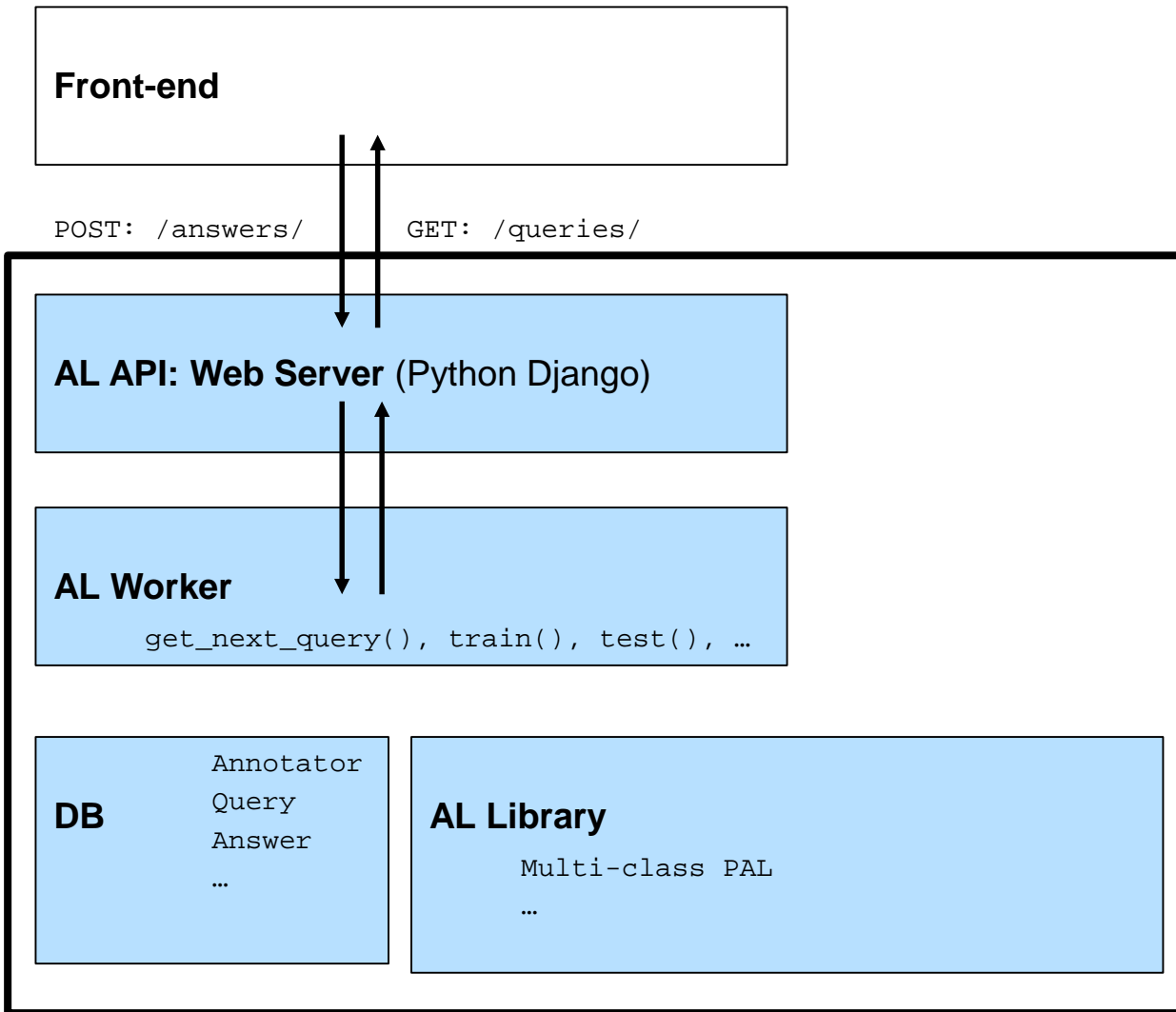
Experiments In the Wild



For research development and validation, we need a **learnable task** and **ground truth**.

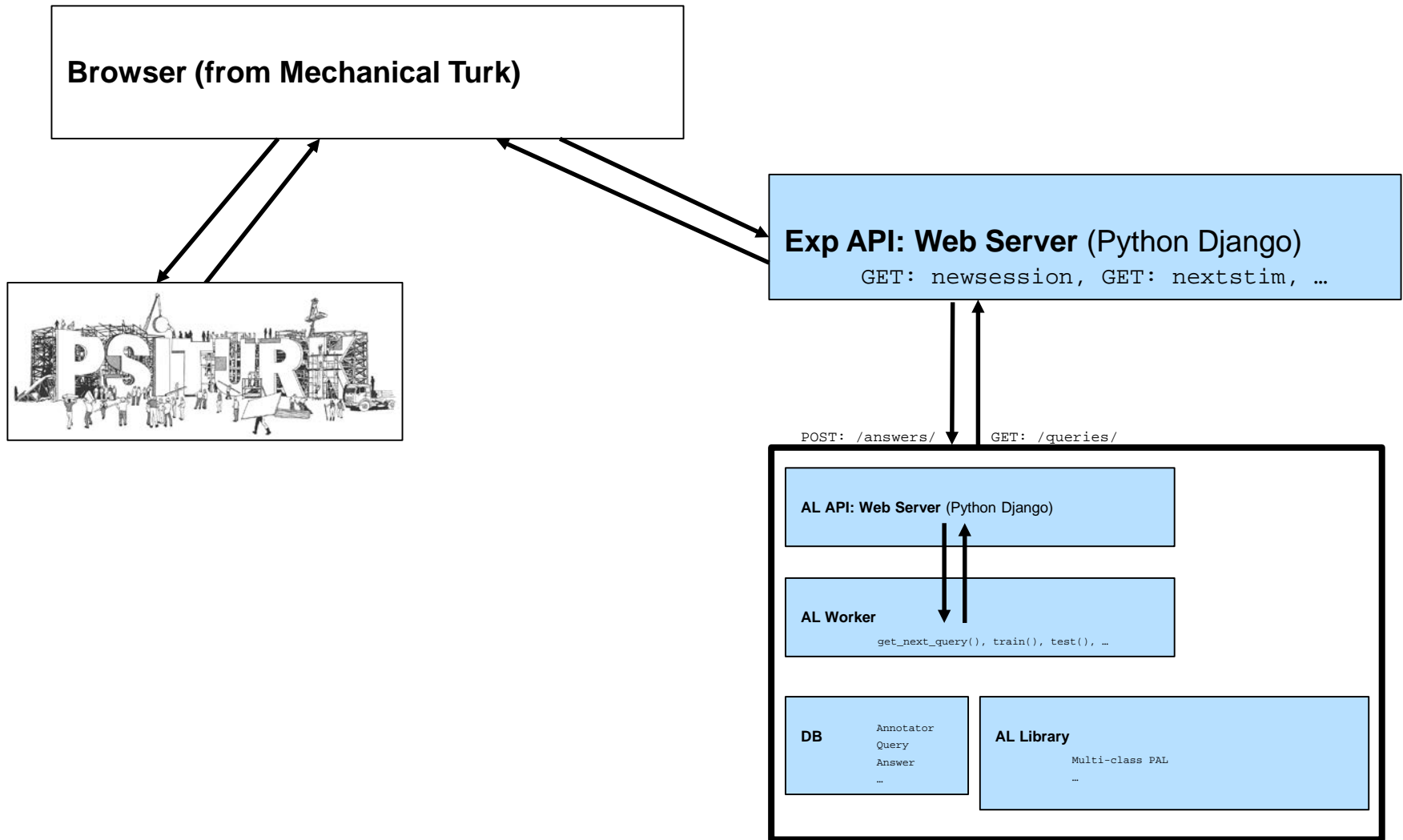
To stay close to the real data, we projected the samples into a three dimensional PCA space, and mapped those dimensions onto stick figures to classify.

Proactive Learning API



→ allows for real experiments using the most advanced active learning techniques

Complete Experimentation System

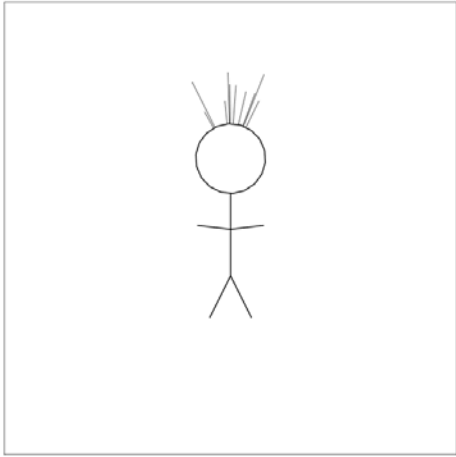


Creature Classification on AMT

Previously seen creatures.

Not Seen Yet	Not Seen Yet	Not Seen Yet
amlity	desper	squent

Please classify this creature.



amlity desper squent










practice

Accuracy

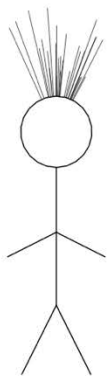
Continue →

Creature Classification on AMT

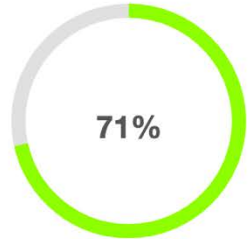
Previously seen creatures.

amilty desper squent



Please classify this creature.



71%

Accuracy

Background and goals
Learning theory progress
Experimental progress
Conclusions and next steps



Next Steps

- Complete pilot: Is the task learnable?
- DPAL vs. baseline
- Joint optimization of analyst and classifier objectives.
- Extension of experimentation software to allow multi-session experiments and team experiments.
- ...
- Test transferability of results to a target task (i.e., malware reverse engineering).



Conclusions

- Including humans makes the system more resilient against adversaries.
- When thinking of machine learning for cybersecurity problems, we should be optimizing for what we really care about – the performance the complete human-computer system.
- Experimentation *with* humans is essential in understanding the true impact of active learning advancements.
- “The ability to accurately represent fully reactionary complex human and group activity in experiments will be instrumental in creating laboratory environments that realistically represent real-world cyber operations.” – Cybersecurity Experimentation of the Future Report