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

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?

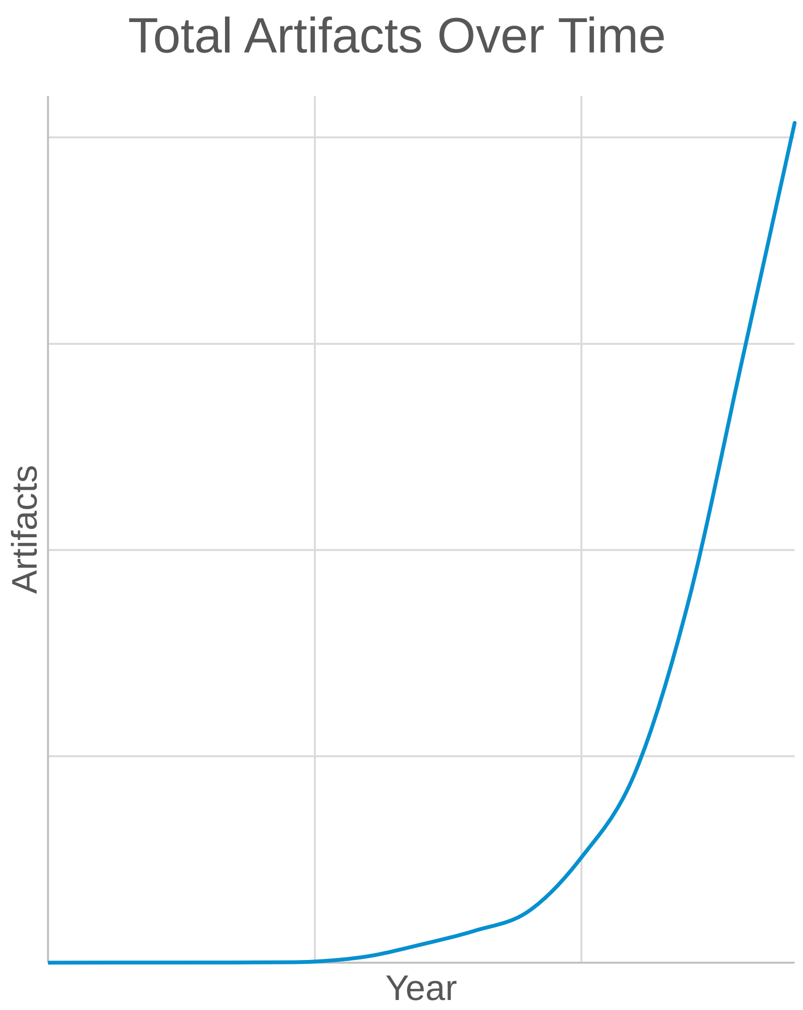
Analysts

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

Machine Learning

- Scalable
- High dimensional
- Precisely specified

A Motivating Problem – Malware Classification



- 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

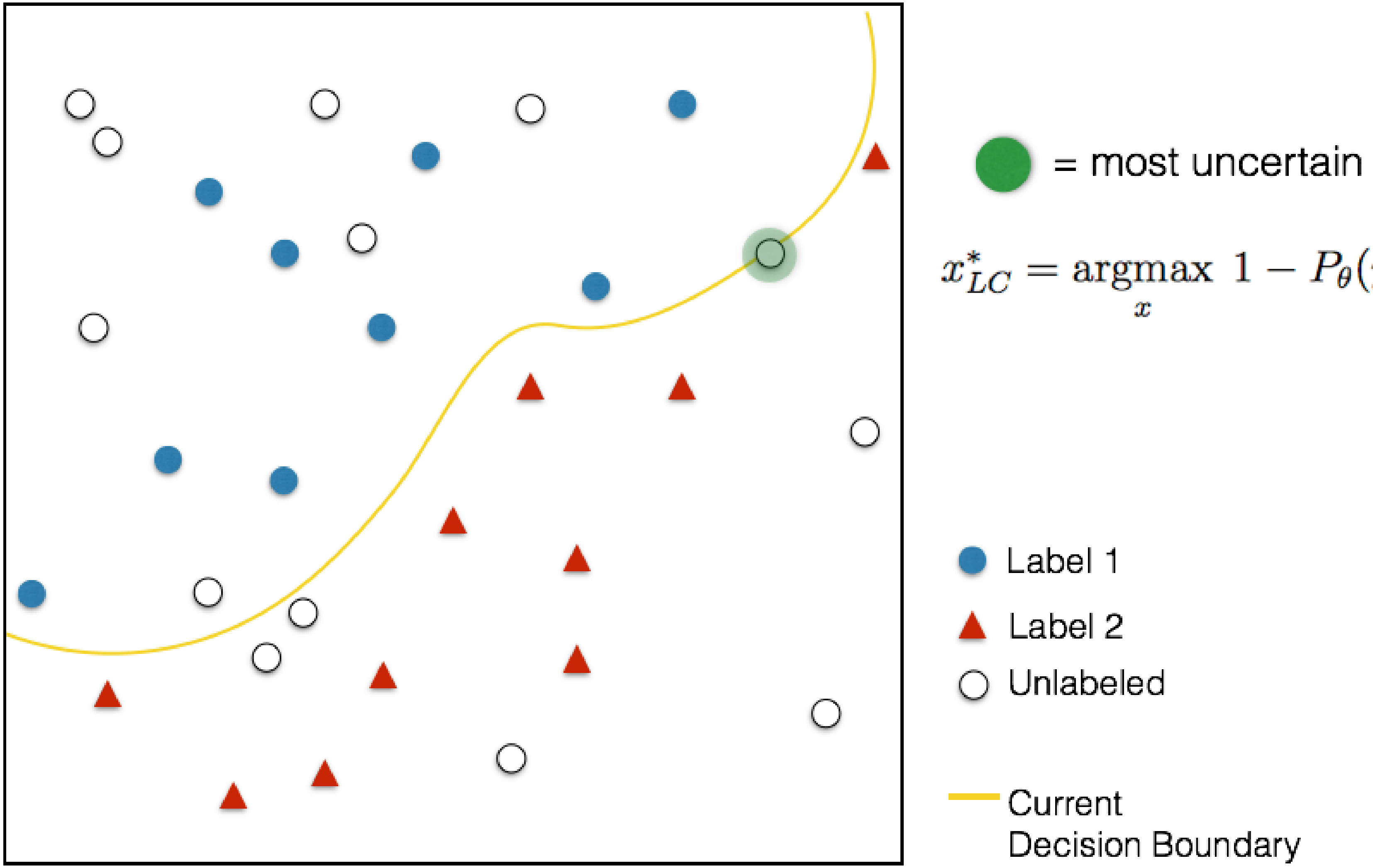
Conclusions and next steps

From Classifiers to Collaboration

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

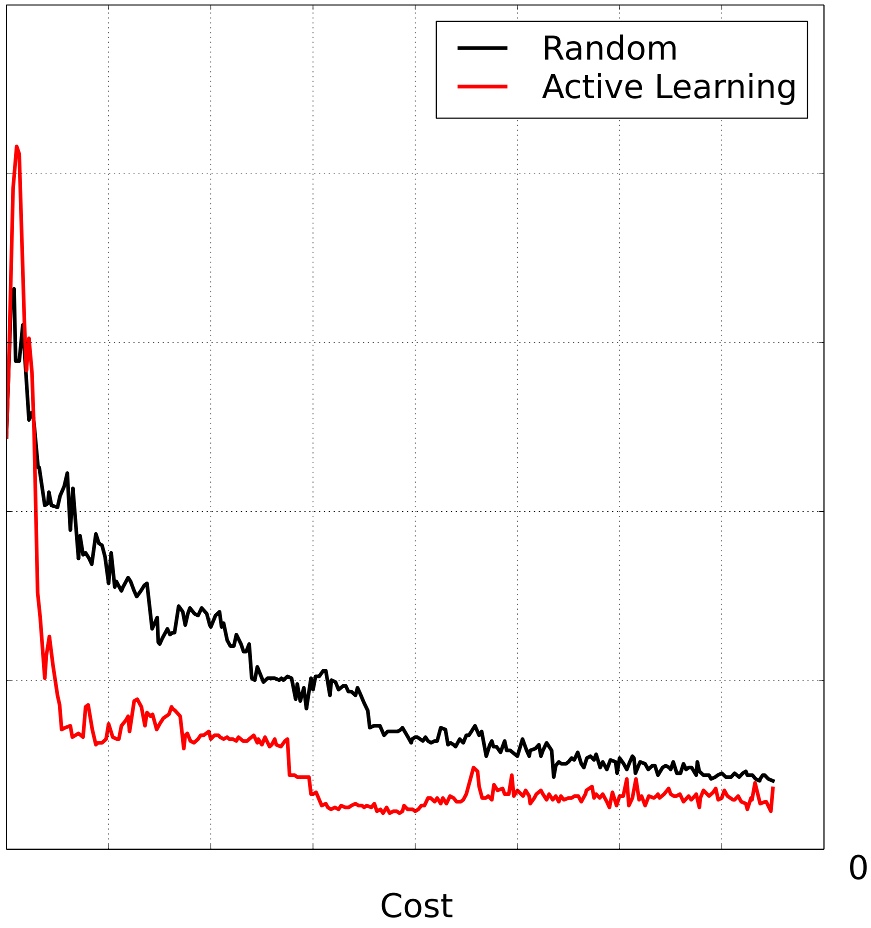
Active learning: the model estimates an ideal sequence of samples and gets labels.

Traditional Active Learning (uncertainty-based)



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

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.

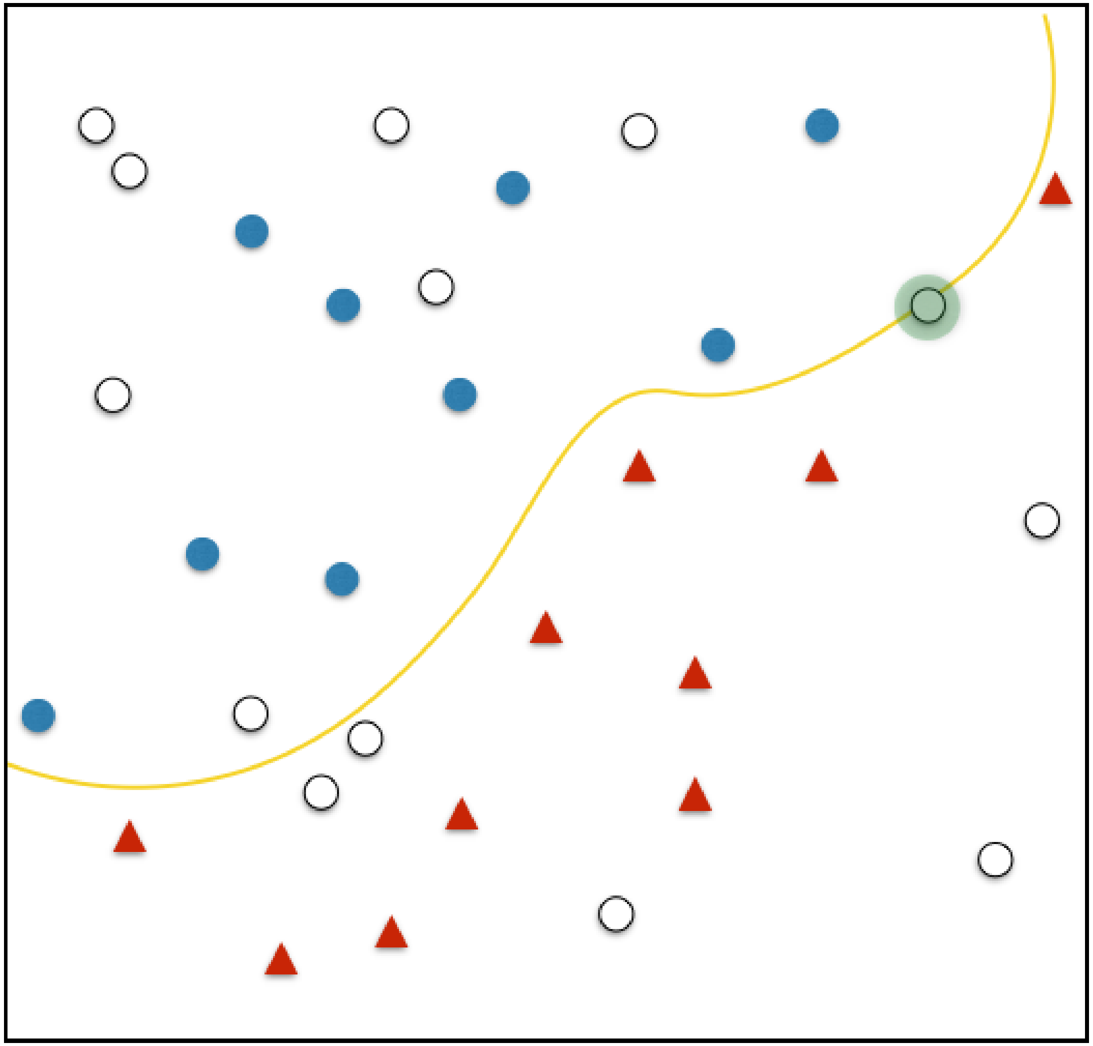
Background and goals

Learning theory progress

Experimental progress

Conclusions and next steps

Traditional Active Learning (uncertainty-based)



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

Proactive Learning With Multiple Domain Experts

Problem Formulation (Objective)

$$\max_{S \subset UL} E[V(S)] - \lambda(\sum_k t_k \cdot C_k)$$

$$\text{s.t.} \sum_k t_k \cdot C_k \leq B, \sum_k t_k = |S|$$

S : the set of instances to be sampled

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

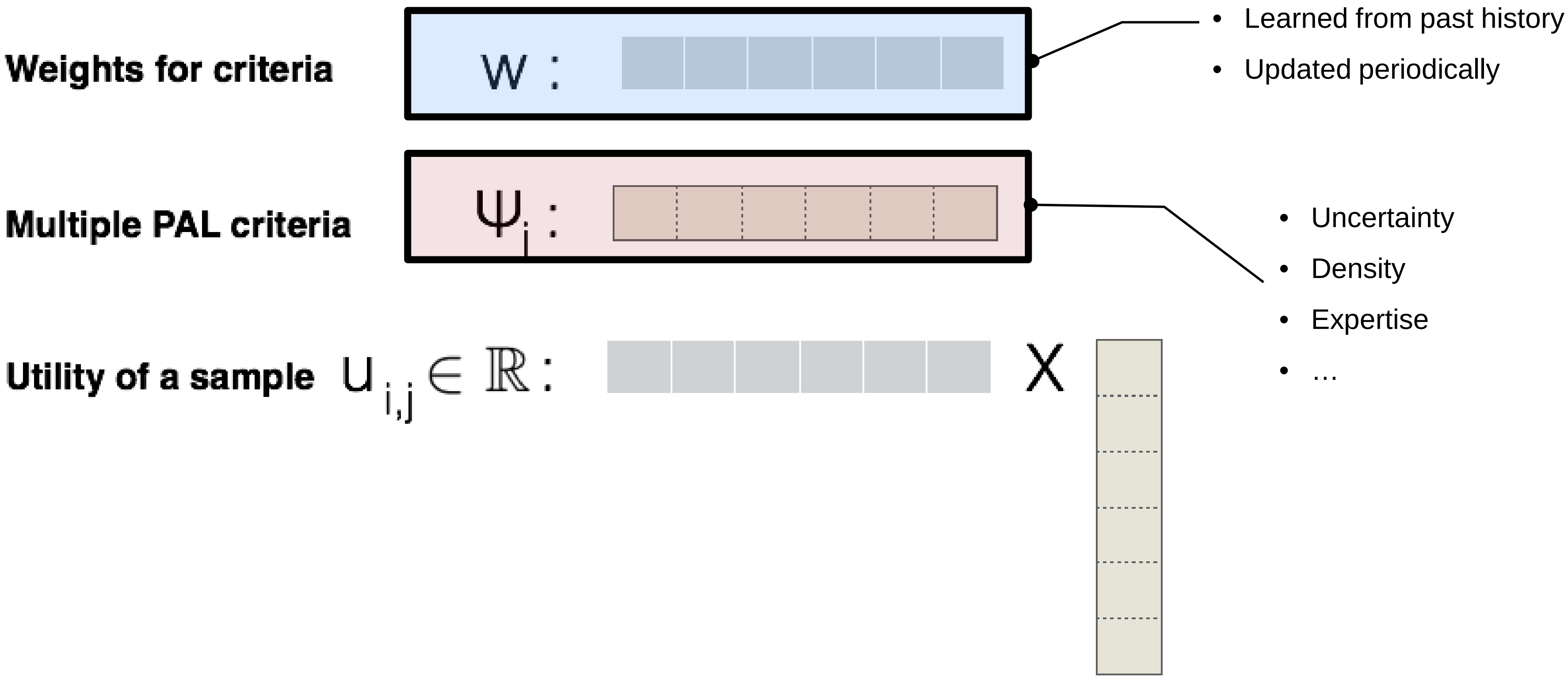
C_k : cost of the chosen expert k

Greedy Approximation

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

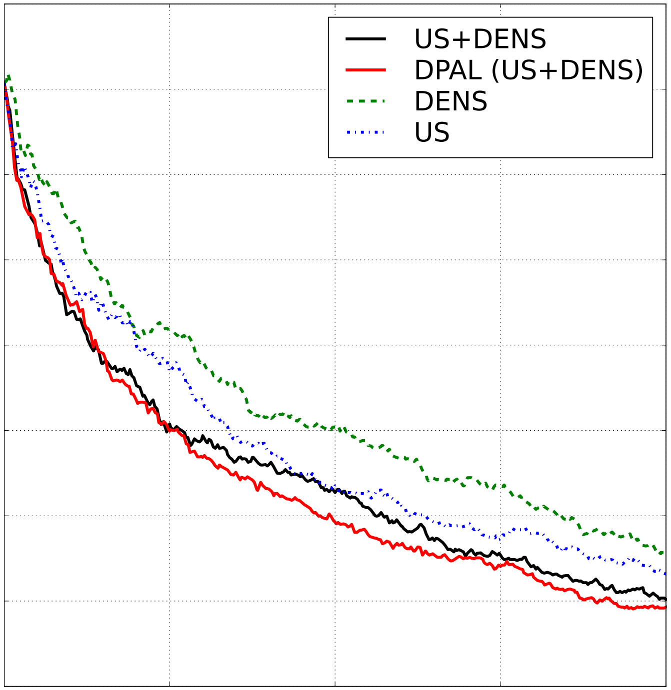
[Moon and Carbonell, 2014]

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

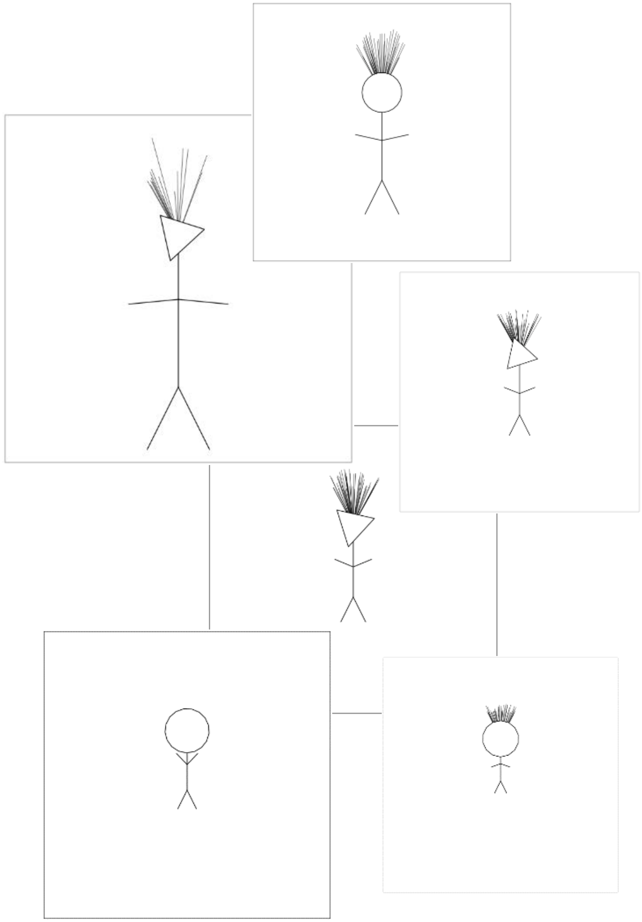
Background and goals

Learning theory progress

Experimental progress

Conclusions and next steps

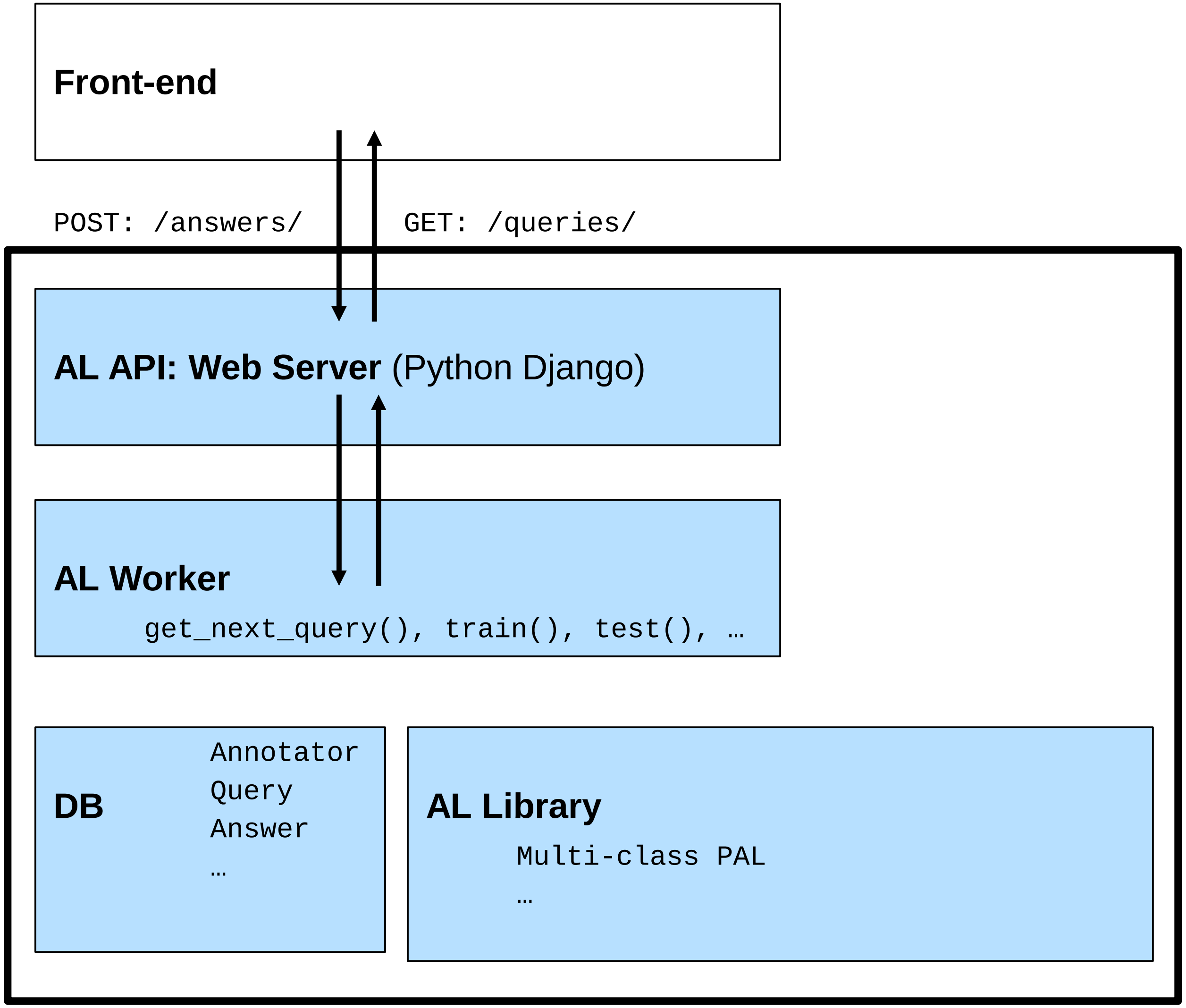
Experiments In the Wild



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

Browser (from Mechanical Turk)

Exp API: Web Server (Python Django)
`GET: newsession, GET: nextstim, …`

`POST: /answers/`
`GET: /queries/`

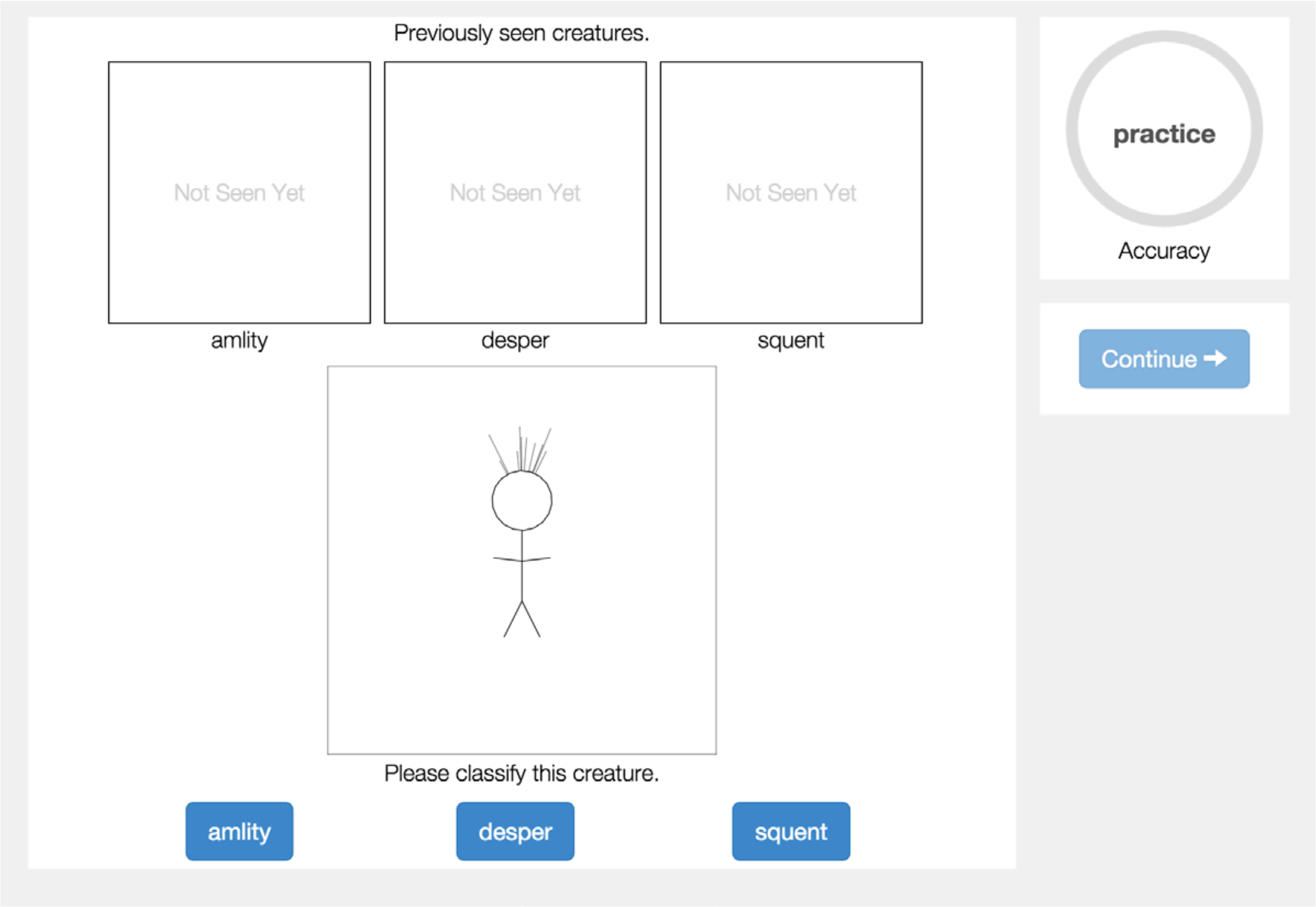
AL API: Web Server (Python Django)

AL Worker
`get_next_query(), train(), test(), …`

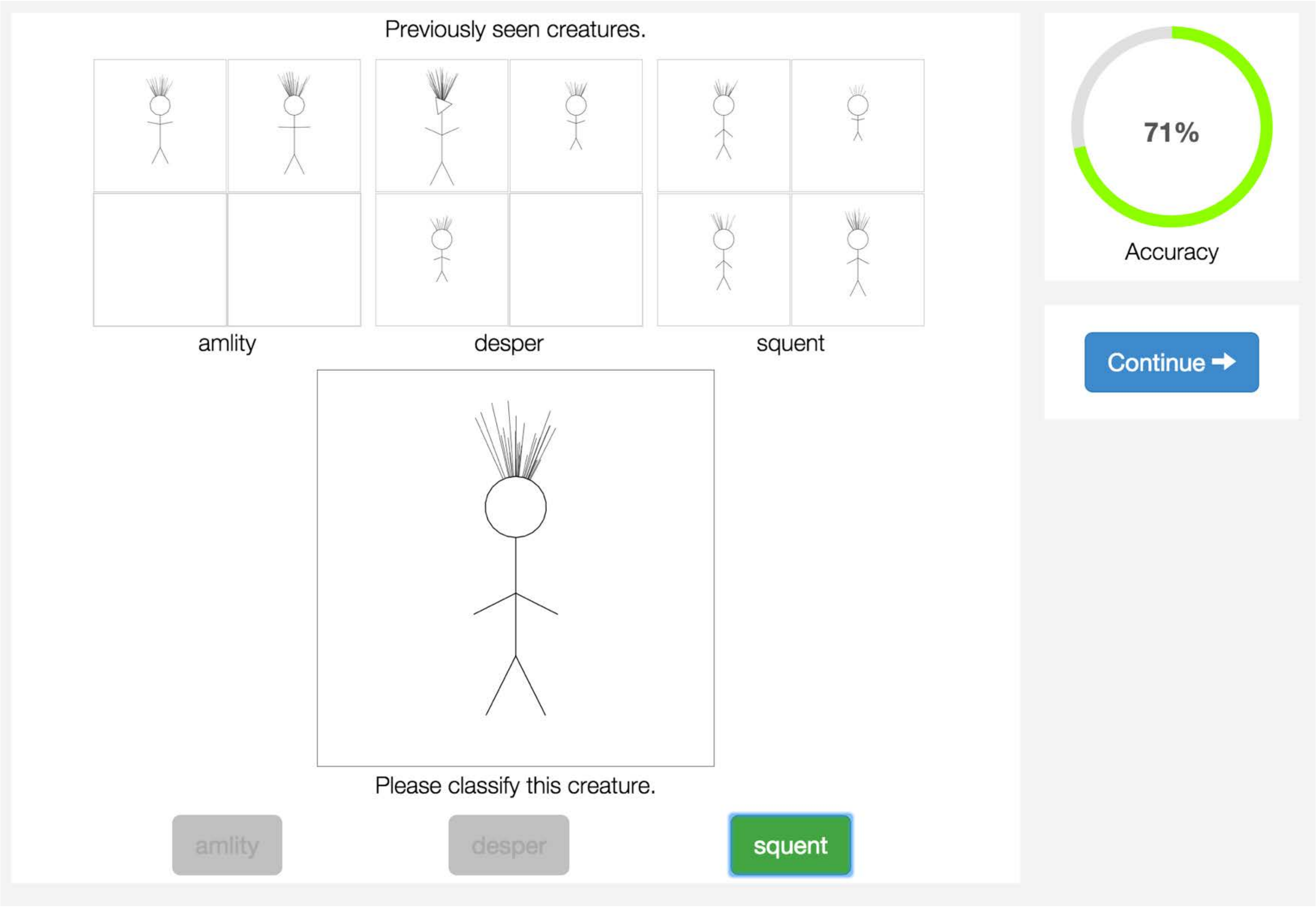
DB
`Annotator`
`Query`
`Answer`
`…`

AL Library
`Multi-class PAL`
`…`

Creature Classification on AMT



Creature Classification on AMT



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