



Acting Engaged: Leveraging Play Persona Archetypes for Semi-Supervised Classification of Engagement

Benjamin D. Nye^{*}, Mark G. Core^{*}, Shikhar Jaiswal^{*},
Avirop Ghosal, Daniel Auerbach

University of Southern California, Institute for Creative Technologies

^{*} Denotes equal contribution.

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.




Engagement

Why? What? How?

Rationale

- What is Engagement?
 - A psychological investment in learning with heightened concentration and interest.
- Why should we study engagement patterns?
 - Lack of engagement leads to lower learning (Baker et al., 2010)
 - Student engagement predicts dropout (Christenson et al., 2012)
 - Engagement can be induced (Lehman, Graesser, et al., 2011)
- Concept: Service for Measuring & Adapting to Real-Time Engagement (SMART-E)

Engagement Time Scales: Current Focus

World (theory)	Time Units		
SOCIAL BAND	Months		Culture of Learning Job/Team Outcomes
	Weeks		
	Days		
RATIONAL BAND	Hours		Continued Learning Course Performance
	10 min		
	Minutes		
COGNITIVE BAND	10 sec		Task Performance Task Interactions/Steps Time on Task
	1 sec		
	100 ms		
BIOLOGICAL BAND	10 ms		Affective Response Neural Responses
	1 ms		
	100 μs		
			

Newell's (1990) Time Scales of Human Action

Engagement Behaviors: Toward Archetypes

- **Diligent (Active Engagement)**: Spends somewhat more time on tasks and shows correspondingly better performance, and more likely to complete optional tasks.
- **Self-Regulated (Active Engagement)**: Seeks out and spends greater time on harder tasks, but may skip or disengage on easier tasks..
- **Cherry Picking (Active Disengagement)**: Seeks out easier tasks or abuses features to make tasks easier (e.g., hint abuse), and avoids harder tasks.
- **Nominal Engagement (Passive Engagement)**: Completes tasks as recommended or assigned, with ordinary time-on-task and performance.
- **Racing/Guessing (Passive Disengagement)**: Rapidly answers (potentially multiple times) despite relatively poor performance.
- **Distracted/Slow (Passive Disengagement)**: Uncommonly delayed or irregular answers, particularly when extra time does not appear to improve performance.
- **Expert/Recall (Passive Engagement)**: Regardless of difficulty level, completes tasks very rapidly and with high performance. Possibly an expert on the content, but might also be shallow recall or lookup.

Concept

Toward generalized engagement services

Why are generalized metrics non-trivial?

Relativity:

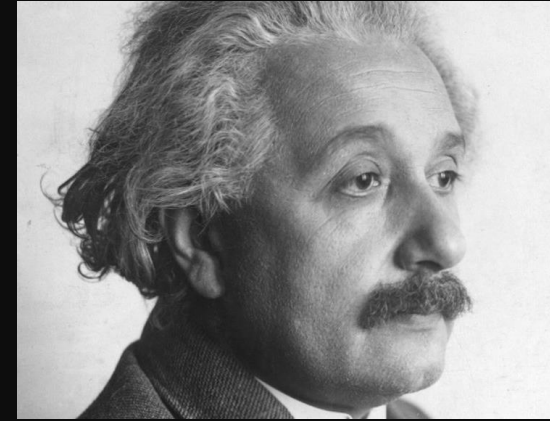
- What actions occur in a system?
- What is a “large amount” of time/clicks/etc...?
- What does “engaged” look like?

Automation:

- How to record data from many systems?
- How to re-use analytics with minimal change?

Usefulness:

- How to communicate metrics to stakeholders?
- What metrics are actionable?



Relatively not engaged.

**What We've Got
Here Is Failure To
Communicate**

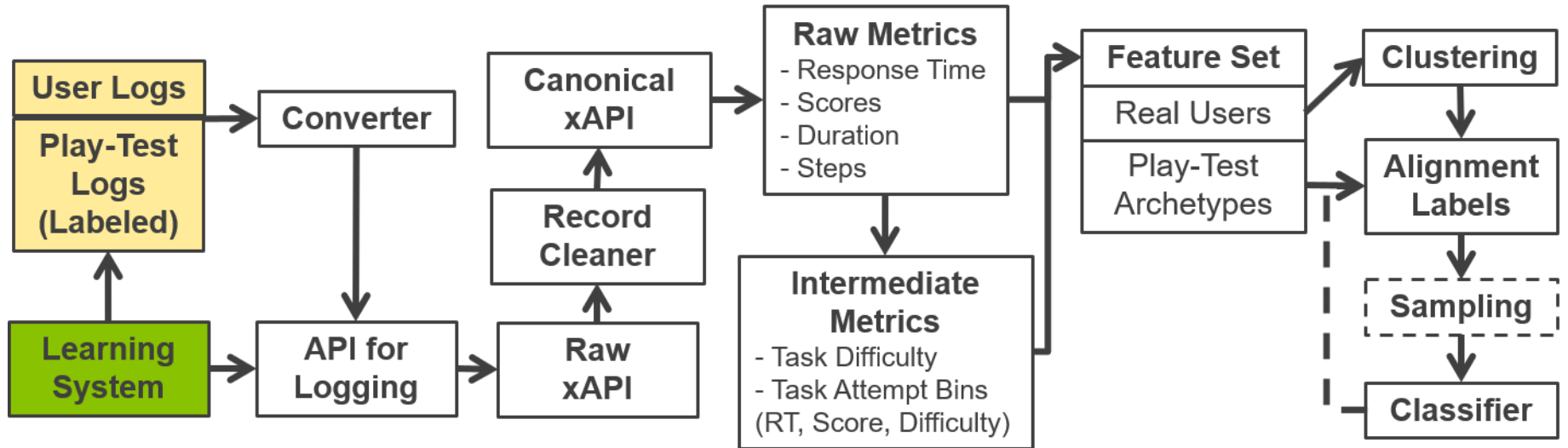
SMART-E Approach: Semi-supervised learning

1. Standards-Based Logs

2. Metrics Engine

3. Learner Feature Vector

4. Classifier



Metrics Pipeline

- Canonical Standards-Based Logs – Levels:
 - Steps: Interactions, inputs
 - Tasks: Activities, usually assessed
 - Lessons: Collections of related activities
 - Session: User actions over a cohesive time period
- Raw Metrics (analyzed at each level):
 - Time-Based (e.g., duration, first step response time)
 - Scores (e.g., score, correctness)
 - Support (e.g., hints used, retry/reattempts)
- Intermediate Metrics:
 - **Average scores, Average task durations, z-scores, etc.**
 - Task difficulty (first-attempt)
 - Interaction Bins: **Difficulty * Duration * Score**

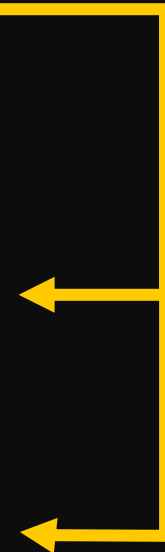
Play-Persona Methodology

- Semi-Supervised:
 - Combine unlabeled real user data with labeled data
 - But where does labeled data come from?
- Engagement Labels:
 - Traditional approaches need training and labor intensive
 - Instead, testers can act out play-personas in the system
 - This approach is often used to play-test games
- Play Personas:
 - Archetypes: Represent different traits or motivations for a session or user
 - Distinct: Exhibit different kinds of behavior, with some differences being diagnostic of the archetype

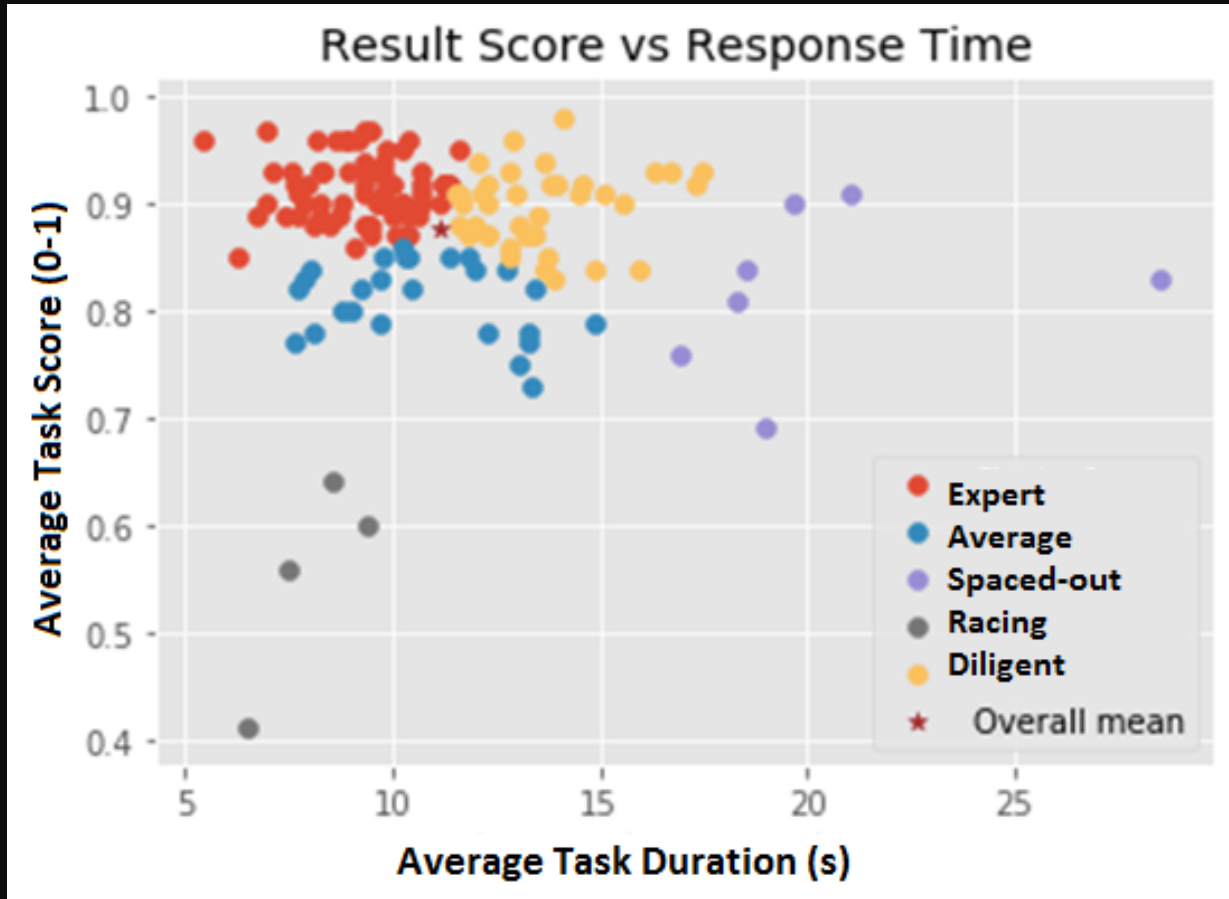
Goal: Generate actionable insights

- Use machine learning to recognize learner "archetype"
- Intervene in cases of disengagement

Archetype	Engaged	Speed	Performance
diligent	yes	slow	high
distracted	no	slow	low
nominal	yes	average	average
expert/recall	yes	fast	high
racing	no	fast	low



Semi-Supervised Learning: Why and How?



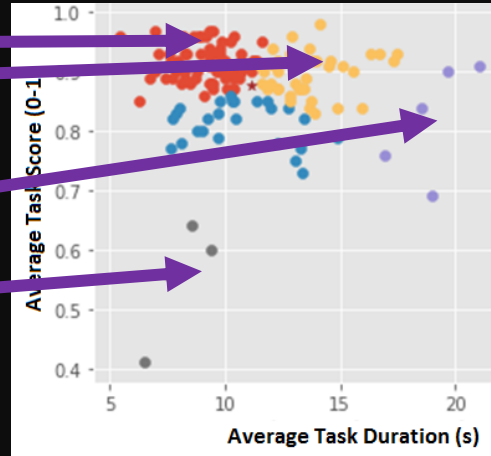
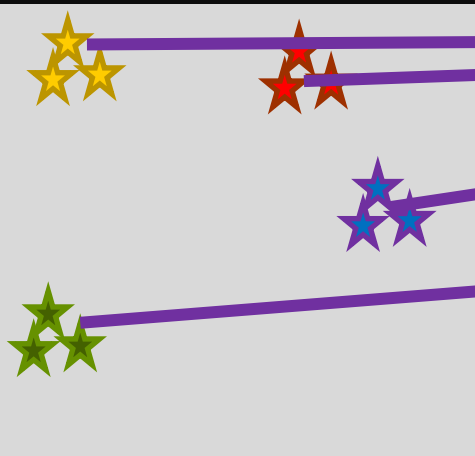
Metrics Engine: Semi-Supervised Classifier

1. Collect Labeled Archetypes

2. Align to Unlabeled Clusters (e.g., GMM)

3. Training Set (Sampled)

4. Train Classifier (e.g., SVM)



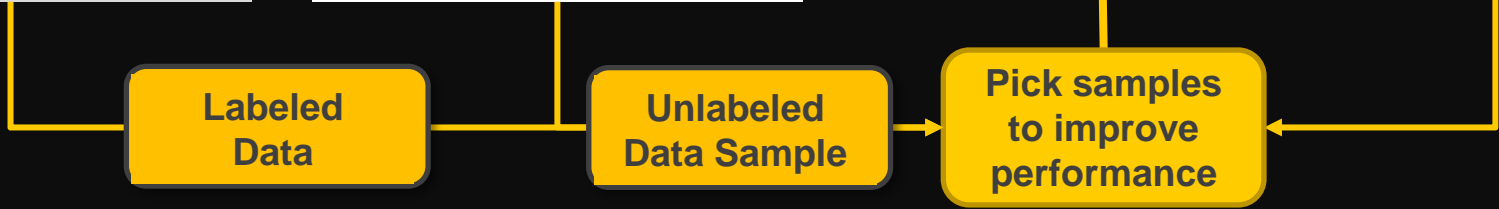
A:Y1	A:Y2	A:Y3
0.0	0.0	0.0
391.0	384.0	543.0
562.0	478.0	584.0
746.0	798.0	715.0
823.0	754.0	669.0
736.0	846.0	742.0
832.0	855.0	799.0
923.0	750.0	816.0
801.0	854.0	826.0
811.0	795.0	864.0
942.0	831.0	938.0

Engagement
Classifier

Labeled
Data

Unlabeled
Data Sample

Pick samples
to improve
performance



Research Questions

- Q1 (Distinctiveness): Are the data patterns for a set of play-tester archetypes distinct (different testers act similarly, given similar instructions)?
- Q2 (Alignment): Will play-test archetypes align with unsupervised clusters producing labeled clusters similar to how experts would label them?
- Q3 (Semi-Supervised Comparison): Will a semi-supervised approach that builds a classifier from play-test and aligned data label individual learners more consistently than relying only on bottom-up clusters?
- Q4 (Basic Features): Will average response time and scores, in simple systems, be sufficient for reasonable engagement labels?
- Q5 (Expanding Features): Will increasing the number of features to include task difficulty and feature interactions lead to greater consistency in fewer samples?

Data Set

ELITE-Lite Interactive Scenarios

Data Set: ELITE-Lite

ELITE Lite Counseling ITS:

- Trains interpersonal skills to help subordinates with personal and performance problems
- Core skills: Active listening, checking for underlying causes of the problem, asking additional questions and verifying information, and responding with a course of action
- Versions in use at multiple military sites:
 - ❖ USMA (ELITE-Lite)
 - ❖ Milgaming (ELITE-Lite)
 - ❖ Officer Training Command Newport (INOTS)



Data Set: ELITE-Lite

Virtual Human

Transcript

The screenshot shows a virtual human interface with a central video window of a woman in a military uniform. To the right is a chat window with a transcript of a conversation. Below the chat is a 'RESPONSE CHOICES' section with three options. A feedback bubble is overlaid on the left, and a 'VIRTUAL COACH' window is at the bottom left.

CHAT

Hey, Sir. Do you have a...

Sure. What's up?

I don't even know if you're the person I should be telling this to, but I had to come to somebody.

This can't be good.

I want to be transferred to a different unit.

RESPONSE CHOICES

Why? What happened?

What? You're not serious, are you? What happened?

I really don't have time for this.

Feedback: You did not check for underlying causes.

VIRTUAL COACH

Feedback: You did not check for underlying causes.

Hint: You should check for underlying personal issues.

Feedback: You did not check for underlying causes.

Choices

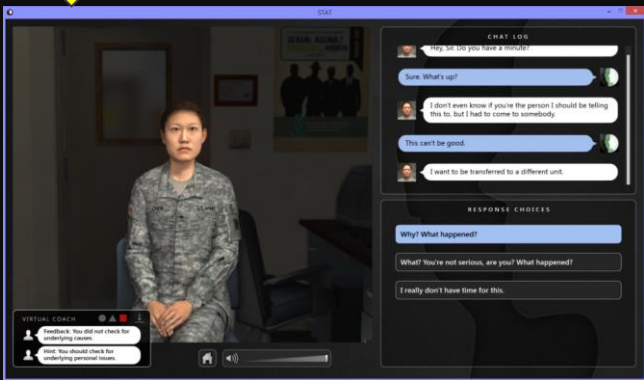
ITS Coach

Data Set: ELITE-Lite

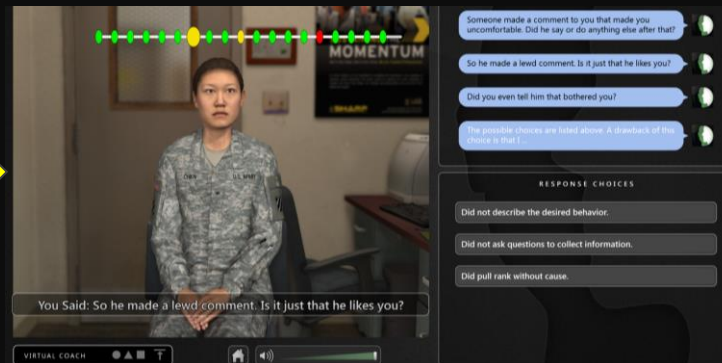
Intro Video



Scenario



After-Action Review



Data Set: Data Set

- **Conditions:**
 - ❖ Baseline data: Using default hint/feedback behavior
 - ❖ Alternate Policies: Different hints/feedback policies (not obvious to one-time user)
- **Two Scenarios:**
 - ❖ “Being Heard” (twice): Requesting transfer & sexual harassment
 - ❖ “Bearing Down”: Conflict between subordinates
- **Pretest/Posttest Design**
- **Interaction Data Logs (xAPI)**
- **Study Population: 145 real user samples from: Georgila et al. (2019)**
- **51 play-test archetype data points**

Results

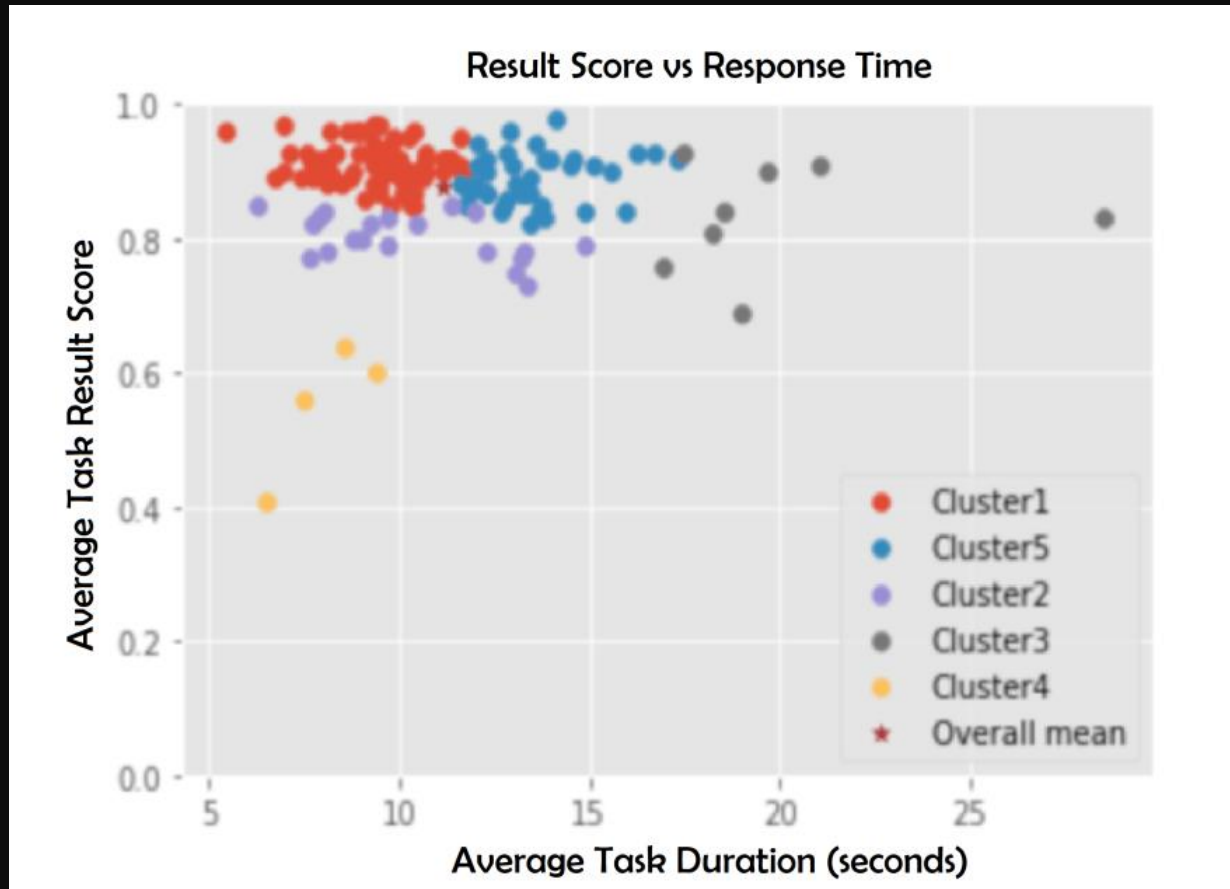
Distinctiveness (Q1)

- Q1: Are the data patterns for a set of play-tester archetypes distinct (different testers act similarly, given similar instructions)?
- Yes. Play-test (Arch) groups are fairly coherent, with limited overlap and reasonable variance.

Group	N	Avg. RT (s)	Avg. Score
Expert (Arch)	15	8.53 ± 2.43	0.95 ± 0.04
Cluster 1	25	8.10 ± 1.00	0.93 ± 0.03
Diligent (Arch)	14	13.15 ± 3.83	0.89 ± 0.07
Cluster 2	75	11.06 ± 1.61	0.90 ± 0.03
Nominal (Arch)	-	-	-
Cluster 3	13	8.63 ± 1.11	0.82 ± 0.02
Distracted (Arch)	12	22.27 ± 13.80	0.77 ± 0.17
Cluster 4	28	15.81 ± 3.43	0.83 ± 0.07
Racing (Arch)	10	7.18 ± 2.47	0.56 ± 0.17
Cluster 5	4	7.98 ± 1.08	0.55 ± 0.09

Table 1: Cluster vs. Archetype Centers ($\mu \pm \sigma$)

Alignment (Q2)



Alignment (Q2)

- Q2 (Alignment): Will play-test archetypes aligned with unsupervised clusters produce labeled clusters similar to how experts label them?
- Yes. Agreement with archetype-aligned clusters is higher with experts than within experts.
- Experts had high reliability for Experts and Racing.
- Diligent, Nominal (phrased as “Average” in the survey), and Distracted more confused
- Slight wording differences may have impacted experts (e.g., “novice learners” vs. “learners”)

	Expert vs. Expert	Aligned Clusters vs. Expert
Agreement	55%	66%
Fleiss kappa	.44	.57
Krippendorff's alpha	.45	.58

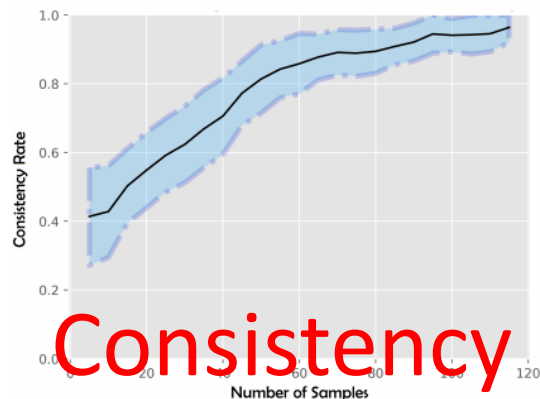
Semi-Supervised Comparison vs. Alignment Only (Q3)

- Q3 (Semi-Supervised Comparison): Will a semi-supervised approach that builds a classifier from play-test and aligned data label individual learners more consistently than relying only on bottom-up clusters?
- How to evaluate?
 - Non-trivial, no single gold standard for behavioral engagement
 - **Cold start Performance:** For labels to be useful, need small amounts of data to fairly quickly agree with larger data set. Test by adding data incrementally to the set, for 20 runs (or 100 for clustering alone)
 - **Consistency:** Agreement of a point's current label vs its final label when all unlabeled data points are observed
 - **Stickiness:** Agreement of a point vs. a prior label when fewer data points were observed

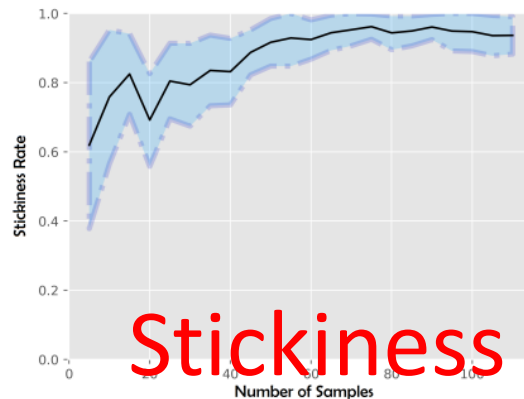
Why not just align clusters?

Semi-
Supervised

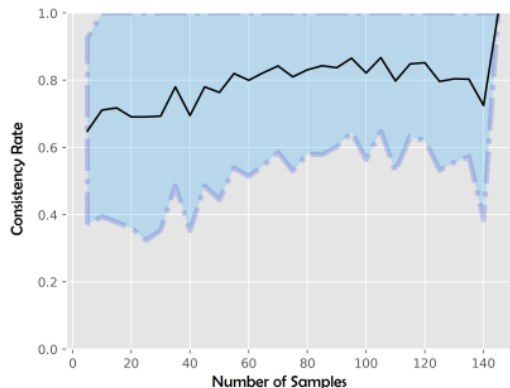
Clustering
Alone



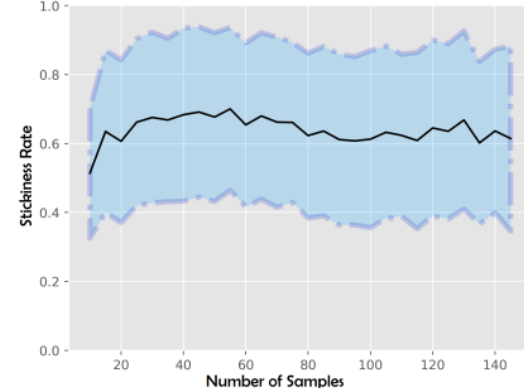
(a) Consistency: Semi-Supervised SVM



(b) Label Stickiness: Semi-Supervised SVM

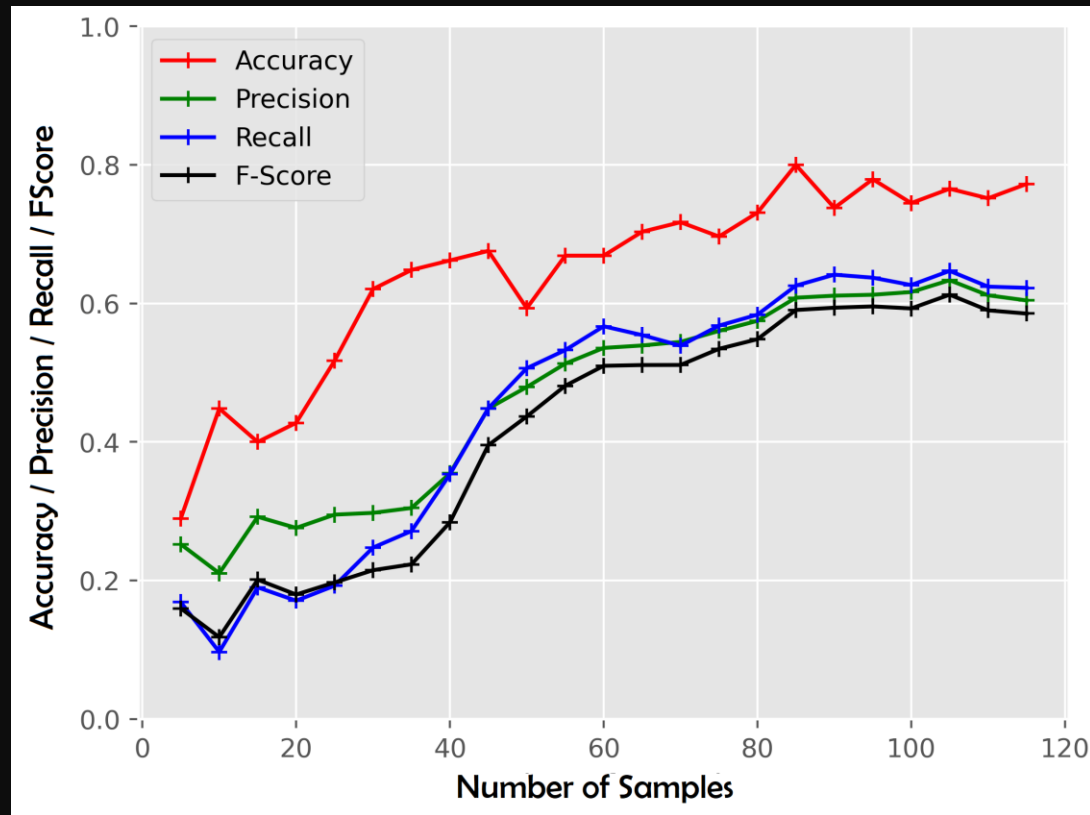


(c) Consistency: Clustering Alone

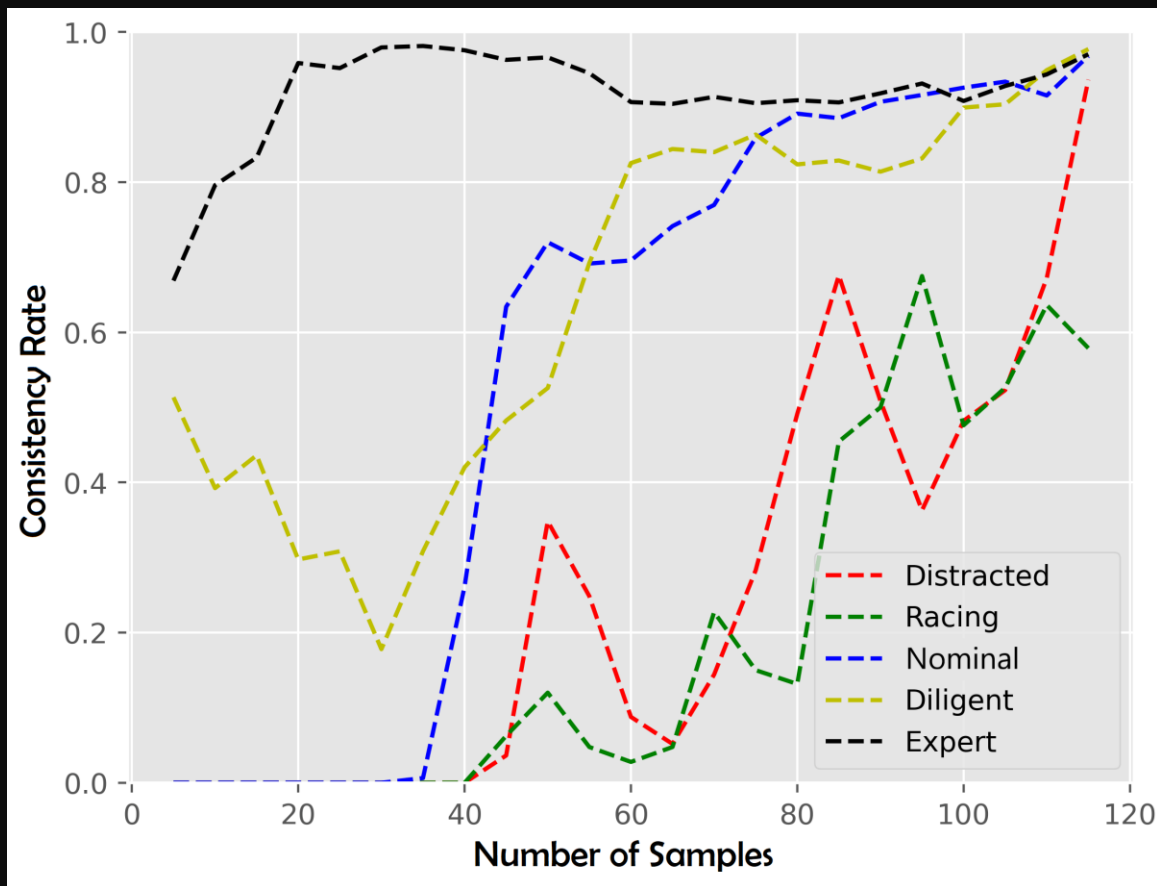


(d) Label Stickiness: Clustering Alone

Semi-Supervised: Class Alignment Labels



Semi-Supervised: Class-Wise Point Consistency



Research Questions

- Q3 (Semi-Supervised Comparison): Will a semi-supervised approach that builds a classifier from play-test and aligned data label individual learners more consistently than relying only on bottom-up clusters?
 - Yes. More consistent and sticky labels than clustering alone.
- Q4 (Basic Features): Will average response time and scores, in simple systems, be sufficient for reasonable engagement labels?
 - Yes. Basic features were sufficient.
- Q5 (Expanding Features): Will increasing the number of features to include task difficulty and feature interactions lead to greater consistency in fewer samples?
 - No / indeterminate. This data set did not show better cold-start behavior for a richer feature set. The results were similar.

Conclusions

- ✓ Classification of Engagement Categories
 - ✓ 5 Categories Labeled: Expert, Diligent, Average, Distracted, Racing
 - ✓ Gives insights equivalent to over 100 samples fairly quickly
 - ✓ 85% consistency by about 52 play-test and 51 real samples
- ✓ Generalized Approach
 - ✓ General xAPI logger for data to process
 - ✓ Any archetypes could be used, if a tester can act them out
 - ✓ Initial results from processing other systems are promising (TALK-ON tank simulation, GIFT Cybersecurity class)

Questions?

Email:
nye@ict.usc.edu