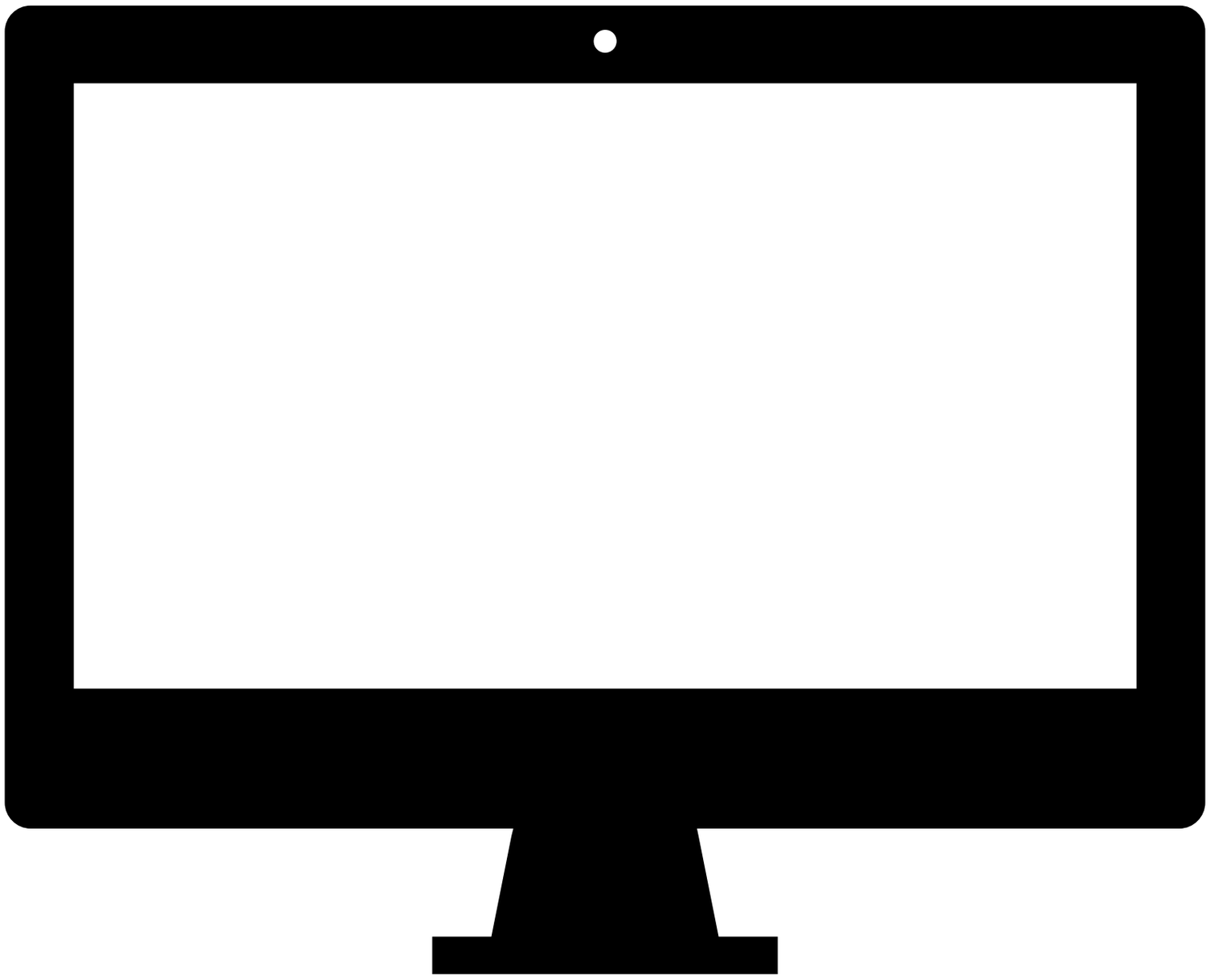
Meta-Learning Contextual Bandit Exploration

Amr Sharaf
University of Maryland
amr@cs.umd.edu

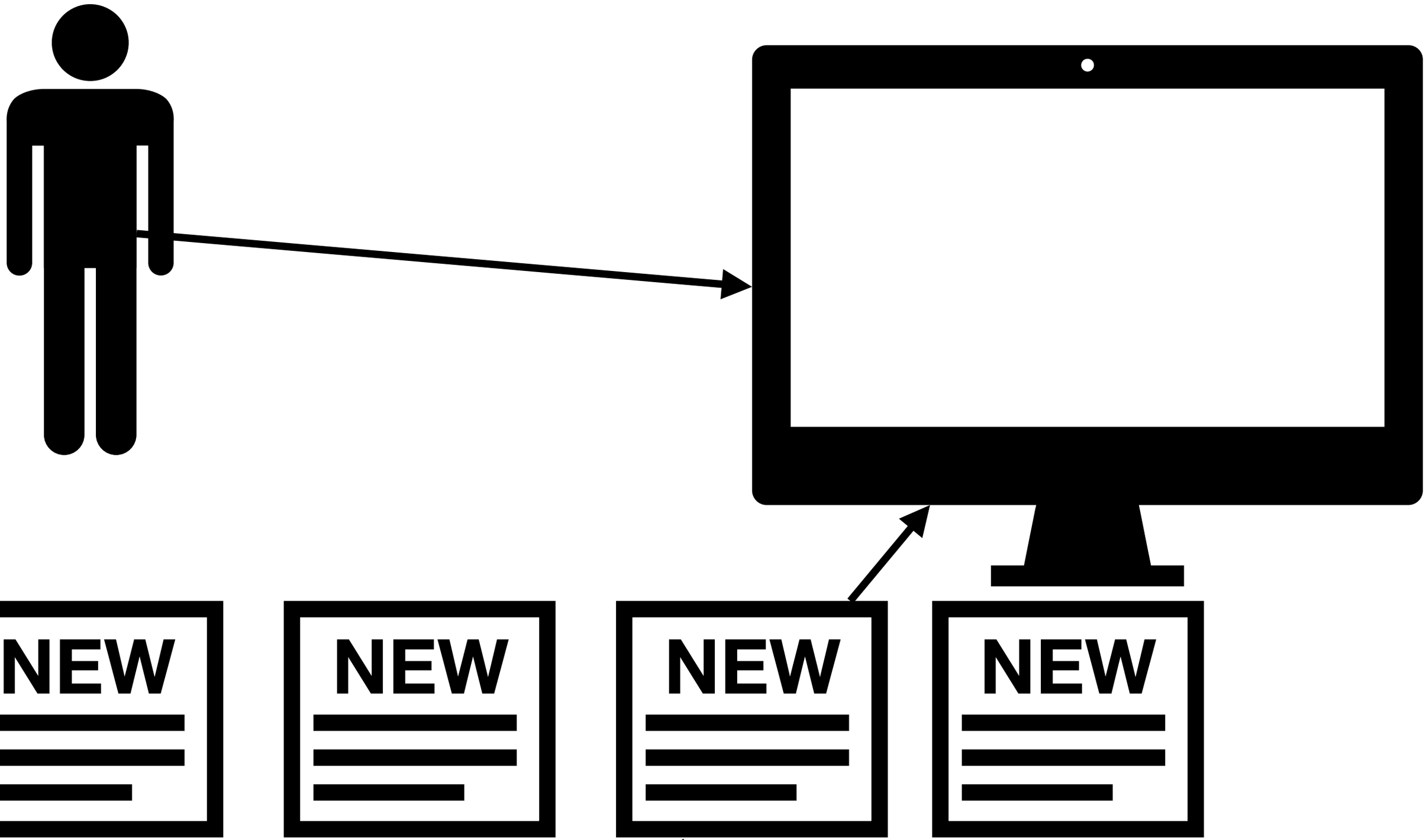
Hal Daumé III
Microsoft Research &
University of Maryland
me@hal3.name

Can we learn to explore in contextual bandits?

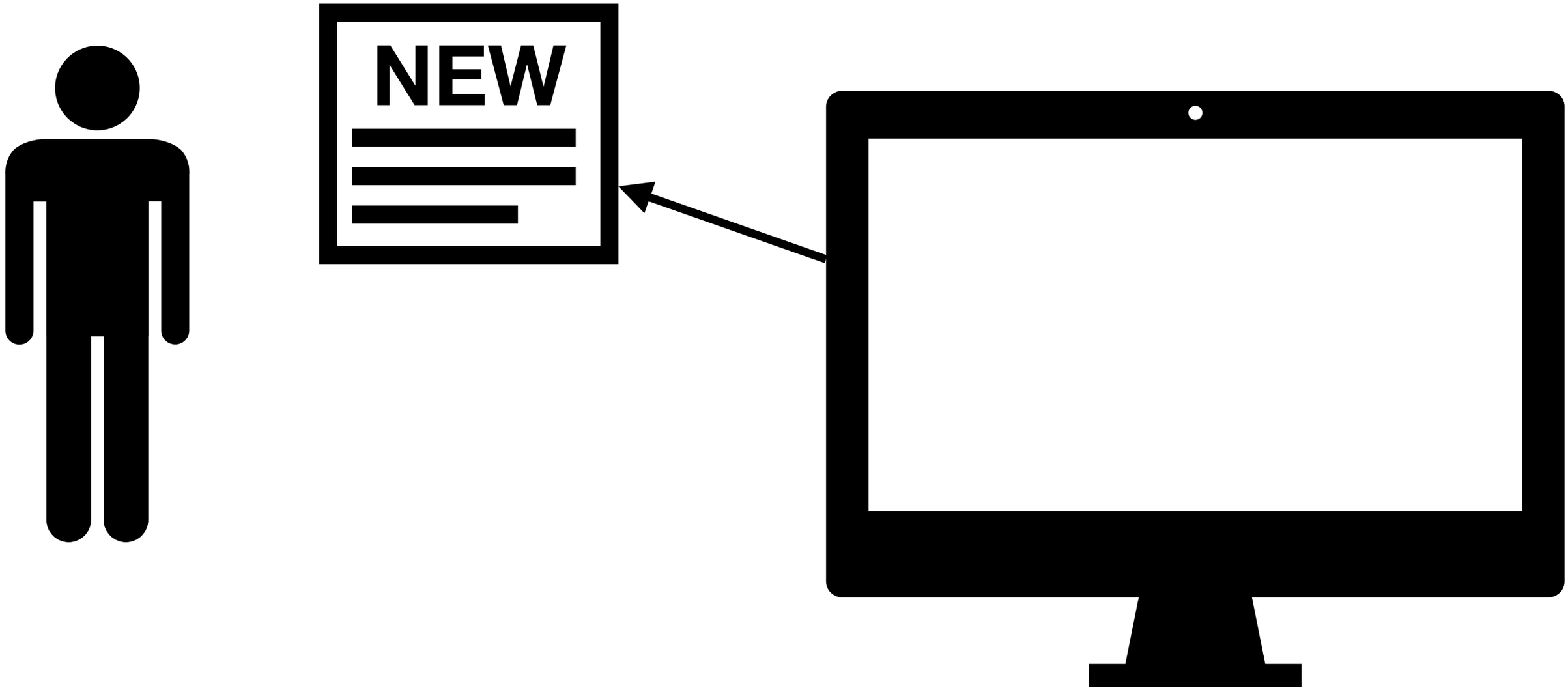
Contextual Bandits: News Display



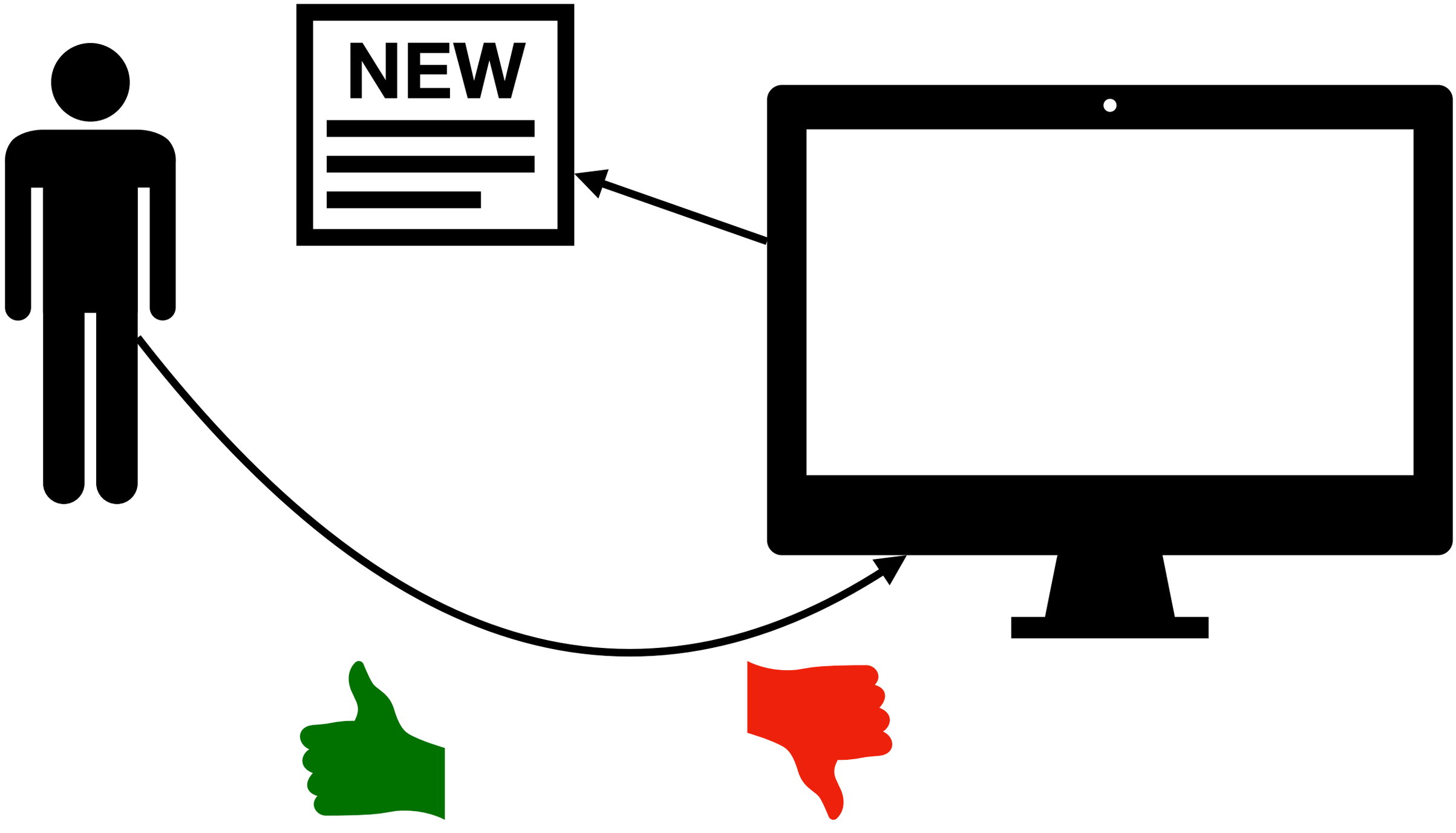
Contextual Bandits: News Display



Contextual Bandits: News Display



Contextual Bandits: News Display

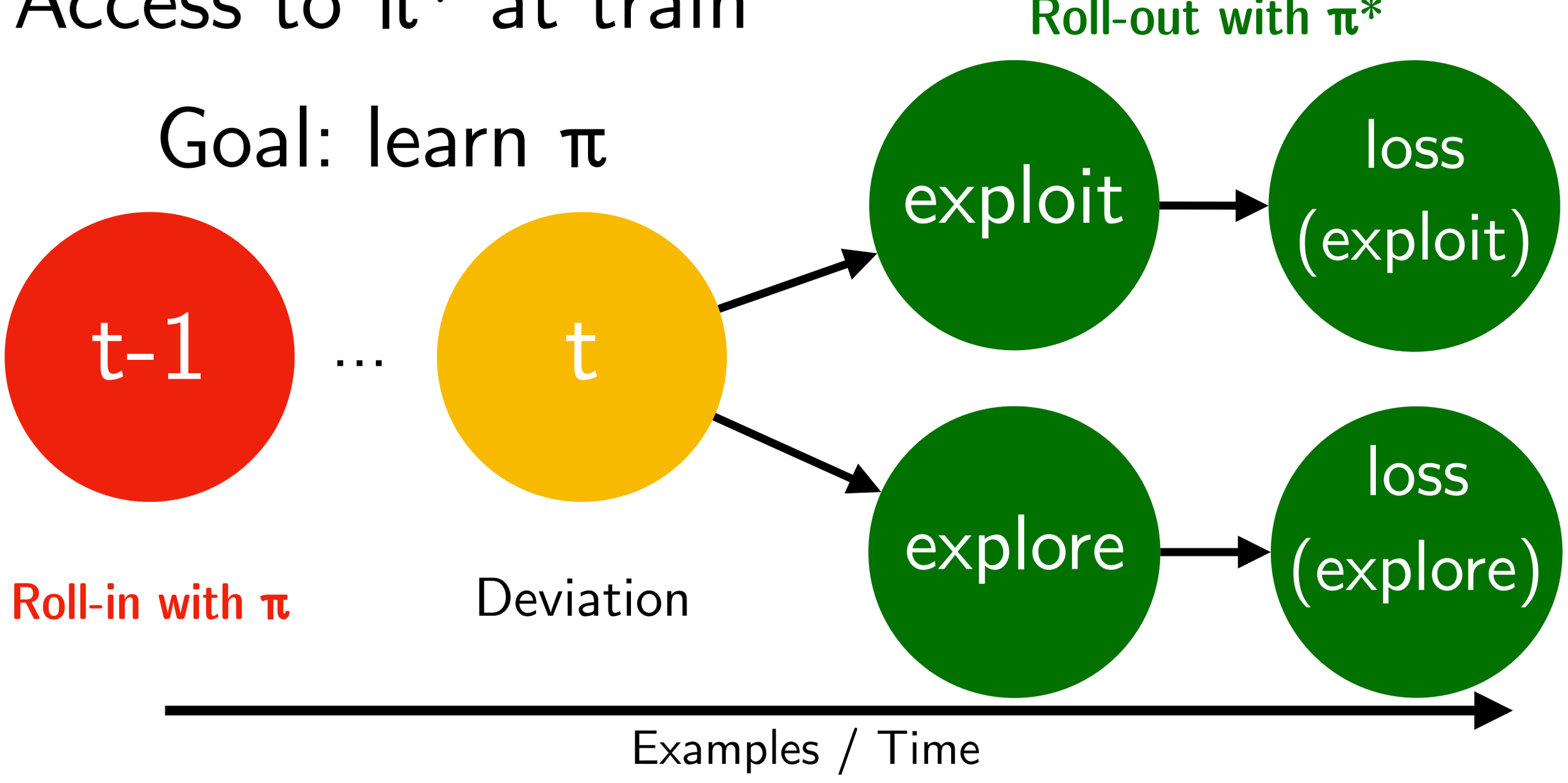


Goal: Maximize Sum of Rewards

Training Mêlée by Imitation

Access to π* at train

Goal: learn π



Generalization: Meta-Features

- No direct dependency on the contexts x.

- Features include:

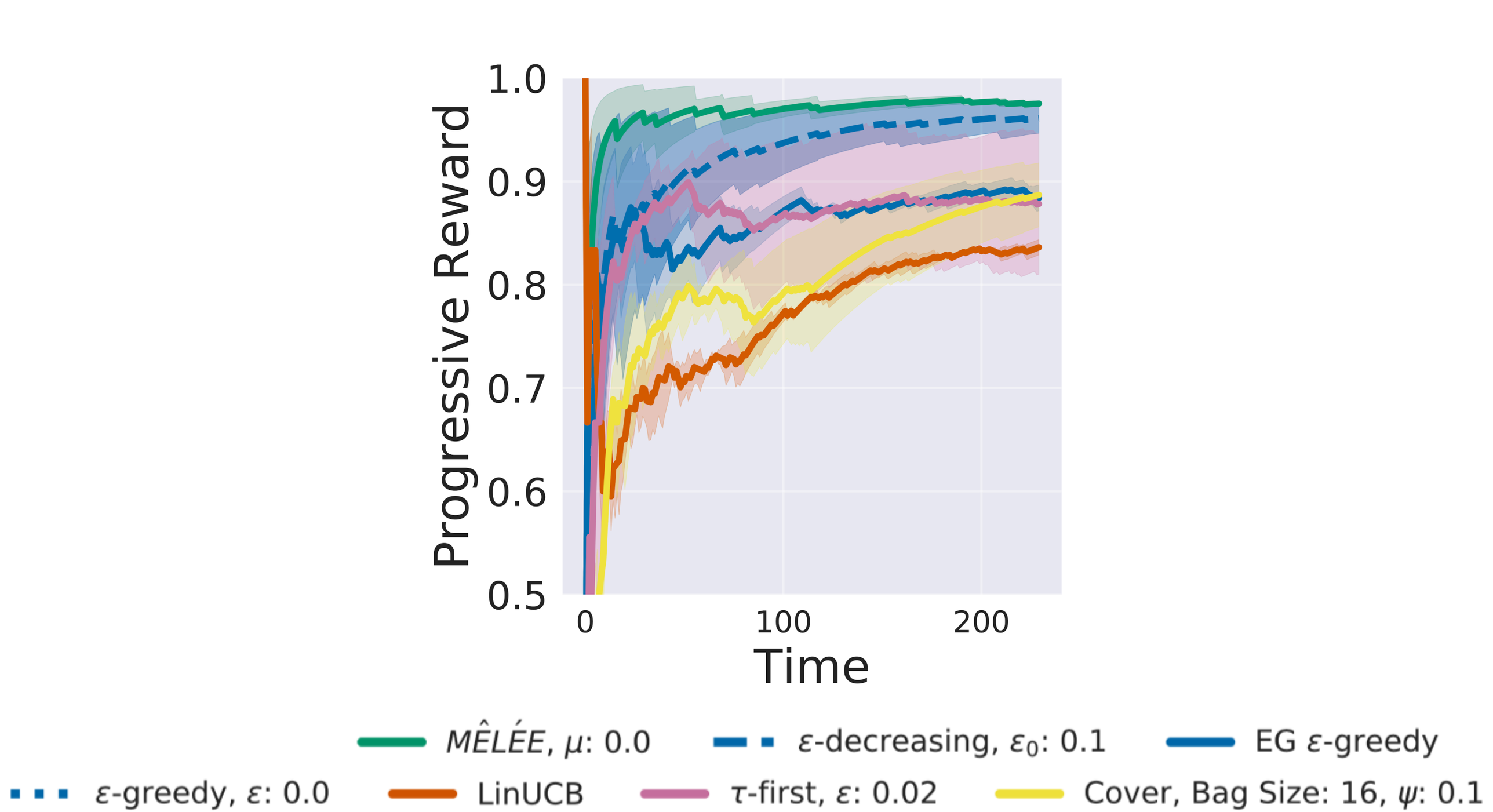
  - Calibrated predicted probability $p(a_t \mid f_t, x_t)$;

  - Entropy of the predicted probability distribution;

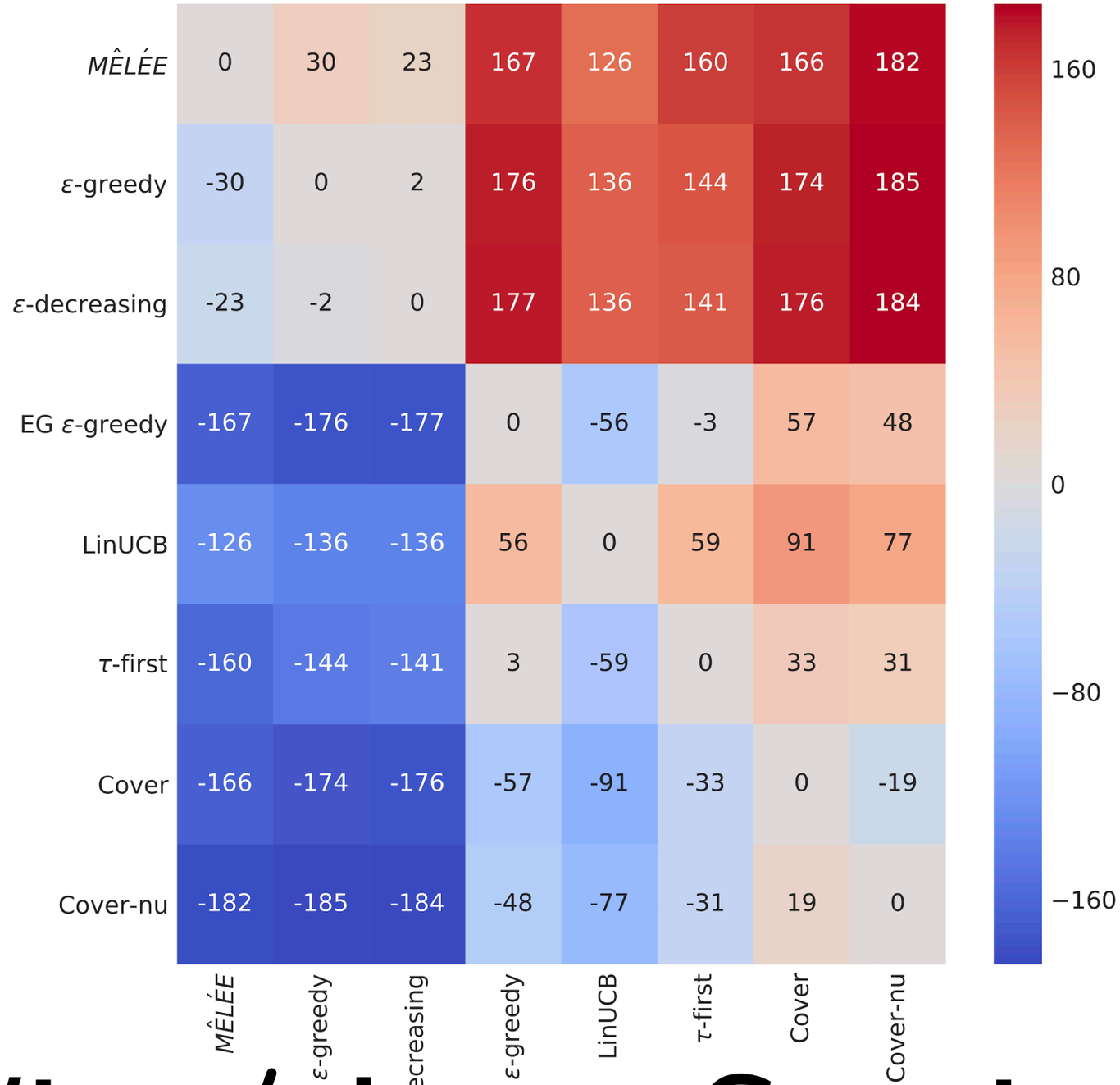
  - A one-hot encoding for the predicted action $ft(x_t)$;

  - Current time step t;

  - Average observed rewards for each action.

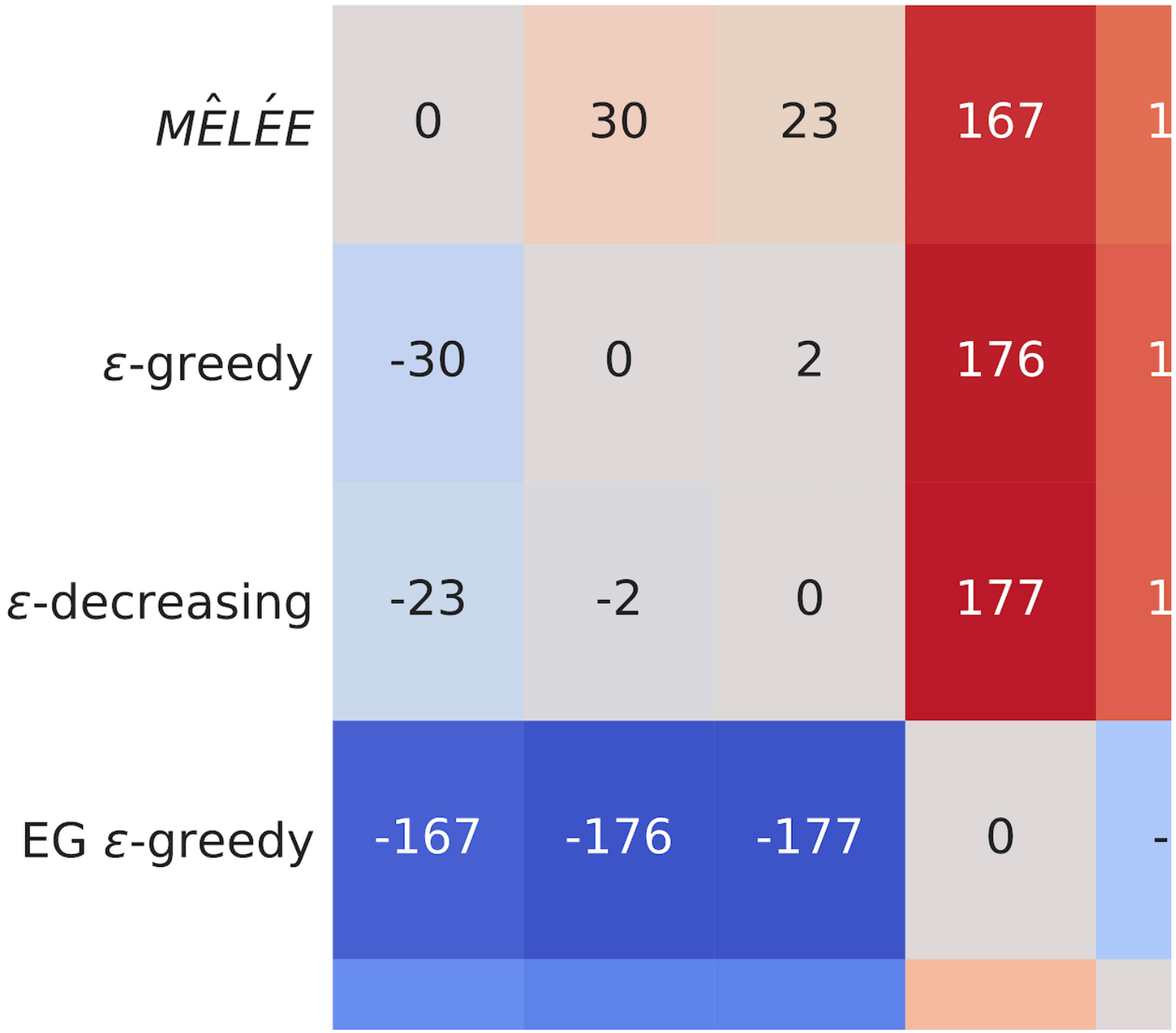


A representative learning curve



Win / Loss Statistics

Win statistics: each (row, column) entry shows the number of times the row algorithm won against the column, minus the number of losses.



Win / Loss Statistics

Win statistics: each (row, column) entry shows the number of times the row algorithm won against the column, minus the number of losses.

Theoretical Guarantees

- The no-regret property of Aggrevate can be leveraged in our meta-learning setting.

- We relate the regret of the learner to the overall regret of π.

- This shows that, if the underlying classifier improves sufficiently quickly, Mêlée will achieve sublinear regret.

Conclusion

- Q: Can we learn to explore in contextual bandits?
- A: Yes, by imitating an expert exploration policy;
- Generalize across bandit problems using meta-features;
- Outperform alternative strategies in most settings;
- We provide theoretical guarantees.