# MACHINE TEACHING OF ACTIVE SEQUENTIAL LEARNERS

**Tomi Peltola, Mustafa Mert Çelikok, Pedram Daee, Samuel Kaski**
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland
`firstname.lastname@aalto.fi`

NEURAL INFORMATION PROCESSING SYSTEMS

Aalto University

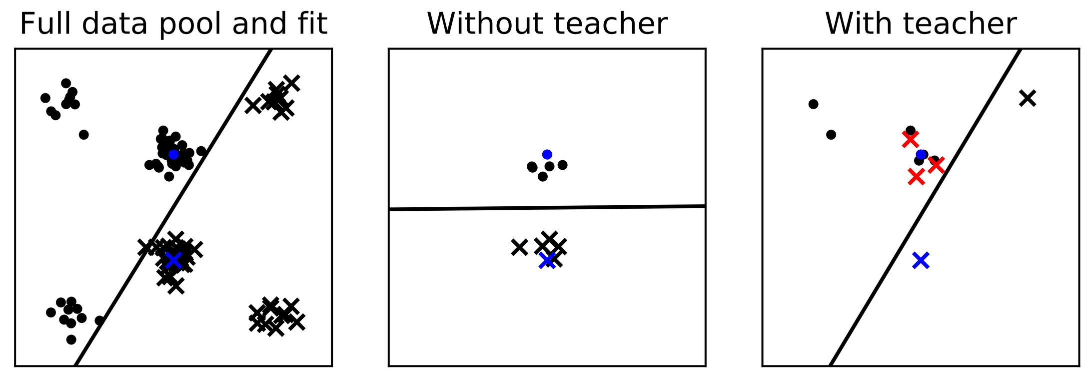
**TL;DR:**
- **PROBLEM: HOW TO STEER AN ACTIVE MACHINE LEARNER THAT QUERIES LABELS SEQUENTIALLY?**
- **SOLUTION: FORMULATE THE TEACHING PROBLEM AS AN MDP, WITH LABEL CHOICE AS ACTION.**
- **RESULT: A TEACHER TEACHING WITH INCONSISTENT LABELS CAN BEAT CONSISTENT LABELS.**
- **FURTHER: ENDOW THE LEARNER WITH A MODEL OF THE TEACHER.**
- **APPLICATION: MODELLING STRATEGIC USER BEHAVIOUR IN INTERACTIVE INTELLIGENT SYSTEMS.**

## INTRODUCTION

**Machine teaching:** Find the best training data that can guide a learning algorithm to a target model with minimal effort.
- Traditionally, the teacher provides data by sampling labels from the true data distribution (consistent teacher).
- Providing true labels can be sub-optimal in finite-horizon tasks for sequential learners that actively choose their queries.

**Contributions**
- We formulate this sequential teaching problem, as an MDP, and allow the teacher to provide data inconsistent with the true distribution ("*With teacher*" panel on the right).
- We address the complementary problem of teaching-aware learning by endowing the learner with a model of the teacher. The final inference problem reduces to inverse reinforcement learning.
- We evaluate the formulation with multi-armed bandit learners in simulated experiments and a user study.

The approach gives tools to taking into account strategic (planning) behaviour of the users in interactive intelligent systems, such as recommendation engines.

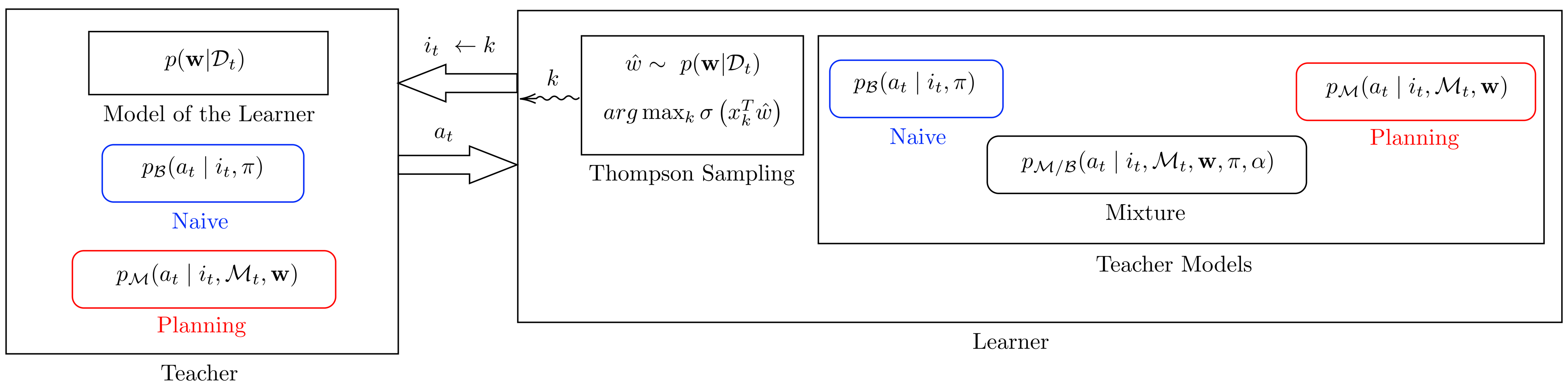



Example of teaching effect on pool-based logistic regression active learner. Starting from blue data,
- the learner without teacher, fails to sample useful points from the pool.
- planning teacher helps the learner by switching some labels (red points).

## MODELLING

**Common goal of the learner and teacher:** Learn (teach) the best possible model of the true data distribution.
**Learner model:**
- Bayesian Bernoulli bandit with linearly dependent arms. Reward probabilities are modelled as $\pi_i = \sigma(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})$, where $\boldsymbol{w}$ is the linear model parameter. Thompson sampling for exploration–exploitation trade-off.

**Simulated Teacher and Teacher models:**
- Teacher models (naive, planning, mixture) interpret the teacher's actions (likelihood for $\boldsymbol{w}$). Planning teacher thinks the learner is using the naive likelihood. Learner thinks the teacher is: naive, planning, or mixture.




### Teacher models

**Naive:**

$$p_{\mathcal{B}}(a_t \mid i_t, \boldsymbol{\pi}) = \text{Bernoulli}(a_t \mid \pi_{i_t})$$

**Planning:**

$$p_{\mathcal{M}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}) = \frac{\exp\left(\beta Q^*_{\mathcal{M}_t}(s'_0, a'_0; \boldsymbol{w})\right)}{\sum_{a'} \exp\left(\beta Q^*_{\mathcal{M}_t}(s'_0, a'; \boldsymbol{w})\right)}$$

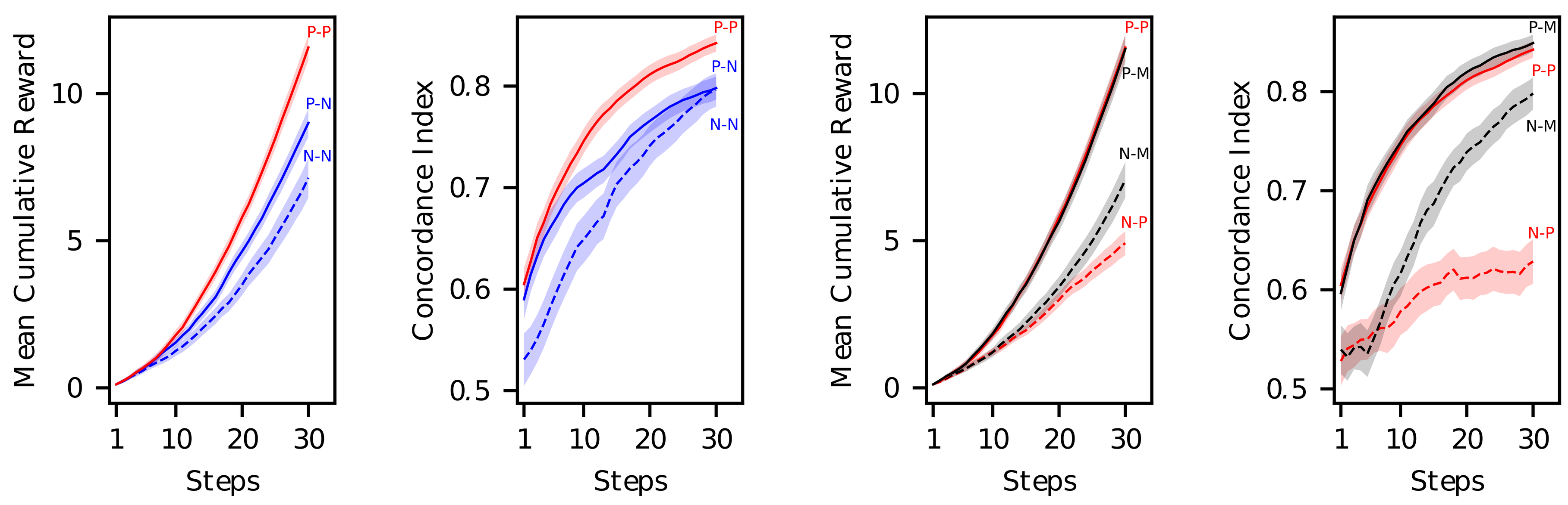
**Mixture:**

$$p_{\mathcal{M}/\mathcal{B}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}, \boldsymbol{\pi}, \alpha) = \alpha p_{\mathcal{M}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}) + (1-\alpha) p_{\mathcal{B}}(a_t \mid \pi_{i_t})$$

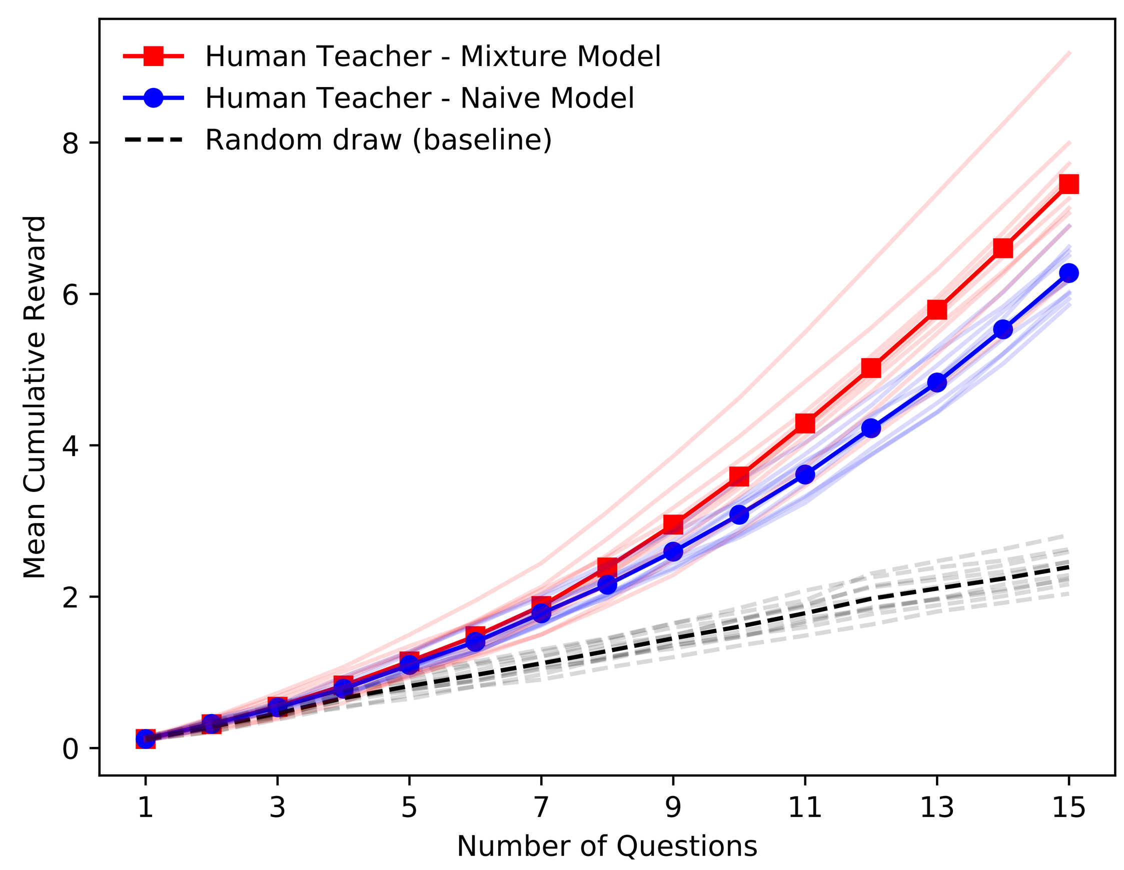
## EXPERIMENTS

**Setup:**
- Word search study: the teacher selects a target word and the learner tries to guess the word by asking sequential questions.
- Learner: "Is this word relevant to the target?", Teacher: Yes/No

**Results with Simulated Teachers:**
- The planning teacher can steer a teacher-unaware learner to achieve a marked increase in performance compared to a naive teacher (P-N vs N-N; left-side panels)
- The performance increases markedly when the learner models the planning teacher (P-P; left-side panels)




**Results with Human Teachers:**
- Participants ($n = 10$) achieved noticeably higher rewards while interacting with a learner having the mixture teacher model (red), compared to the naive teacher model (blue).




## CONCLUSION

- We have introduced a new sequential machine teaching problem, where the learner actively chooses queries (e.g., in active learners and multi-armed bandits) and the teacher provides responses. The new teaching problem is formulated as a Markov decision process, where the solution provides the optimal teaching policy. Using the MDP formulation, teacher-aware learning from the teacher's responses is formulated as probabilistic inverse reinforcement learning.
- The proposed teaching framework holds promise for a feasible and natural computational approach in modelling active user behaviour in interactive intelligent systems.

**See the paper website for more info and the code:** `https://git.io/JeSaU`.