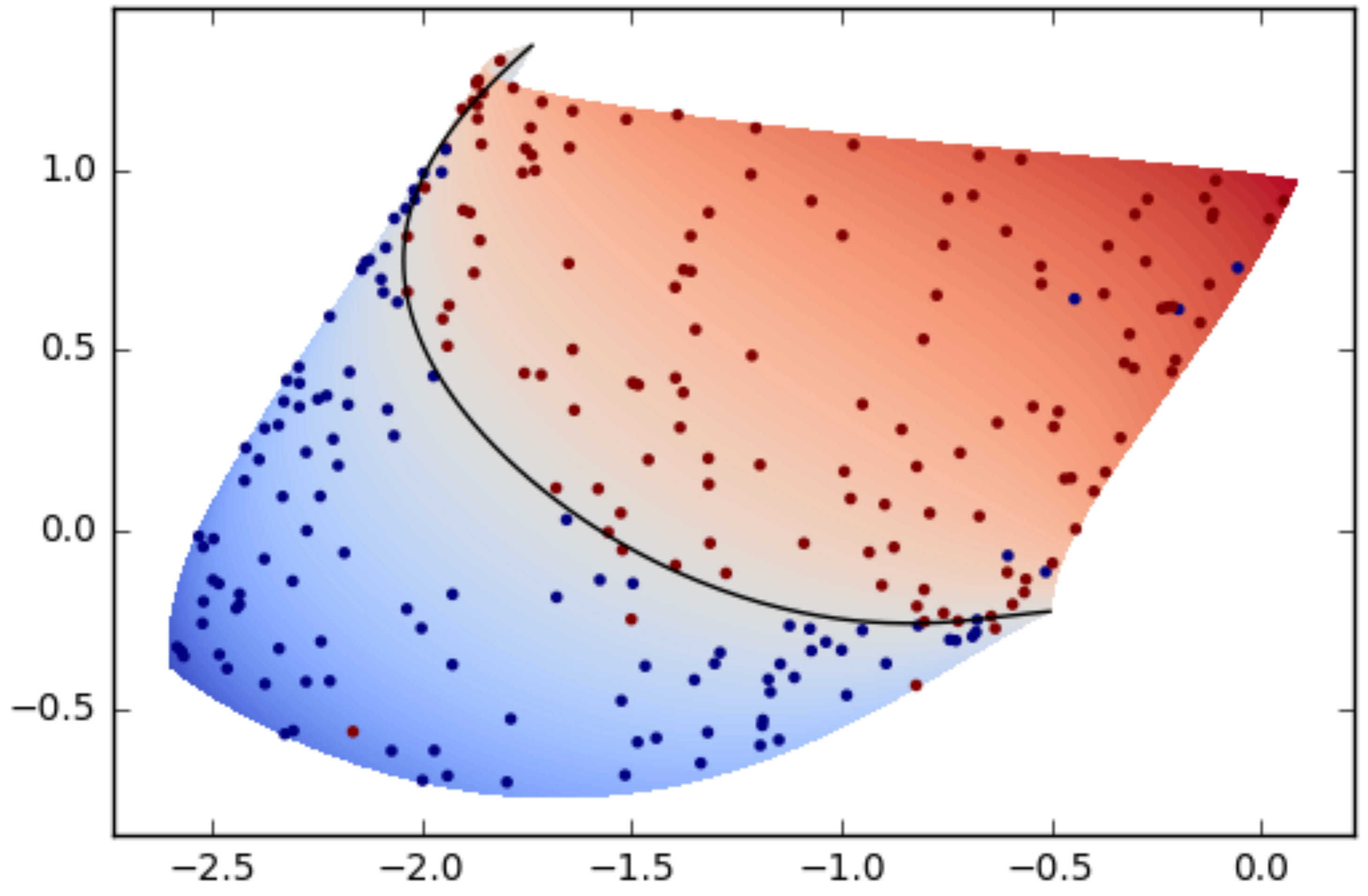
CS 273A: Machine Learning

Winter 2021

Lecture 17: Active and Online Learning

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

All slides in this course adapted from Alex Ihler & Sameer Singh



Logistics

assignments	• Assignment 5 due Thursday
project	• Final report due next Thursday
evaluations	• Evaluations due end of next week
final exam	• Review: next Thursday • Final: Thursday, March 18, 1:30–3:30pm

Today's lecture

Active learning

Online learning

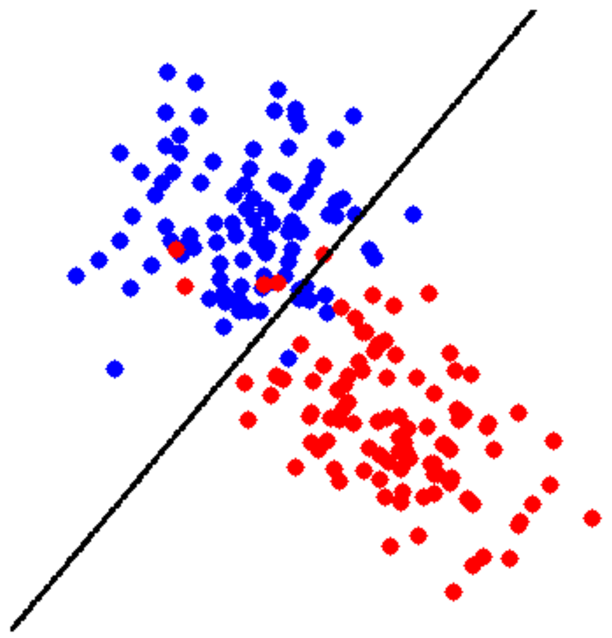
Sequential decision making

Motivation

- Supervised learning: classification
  - Pro: training data $\mathscr{D} = \{(x^{(j)}, y^{(j)})\}$ very informative
  - Con: expert labels $y^{(j)}$ may be expensive to get for big data
- Unsupervised learning: clustering
  - Pro: training data $\mathscr{D} = \{x^{(j)}\}$ may be easier to get
  - Con: discovered clusters may not match intended classes
- Semi-supervised learning: best of both worlds?
  - Few labels $\implies$ class identity; much unlabeled data $\implies$ class borders

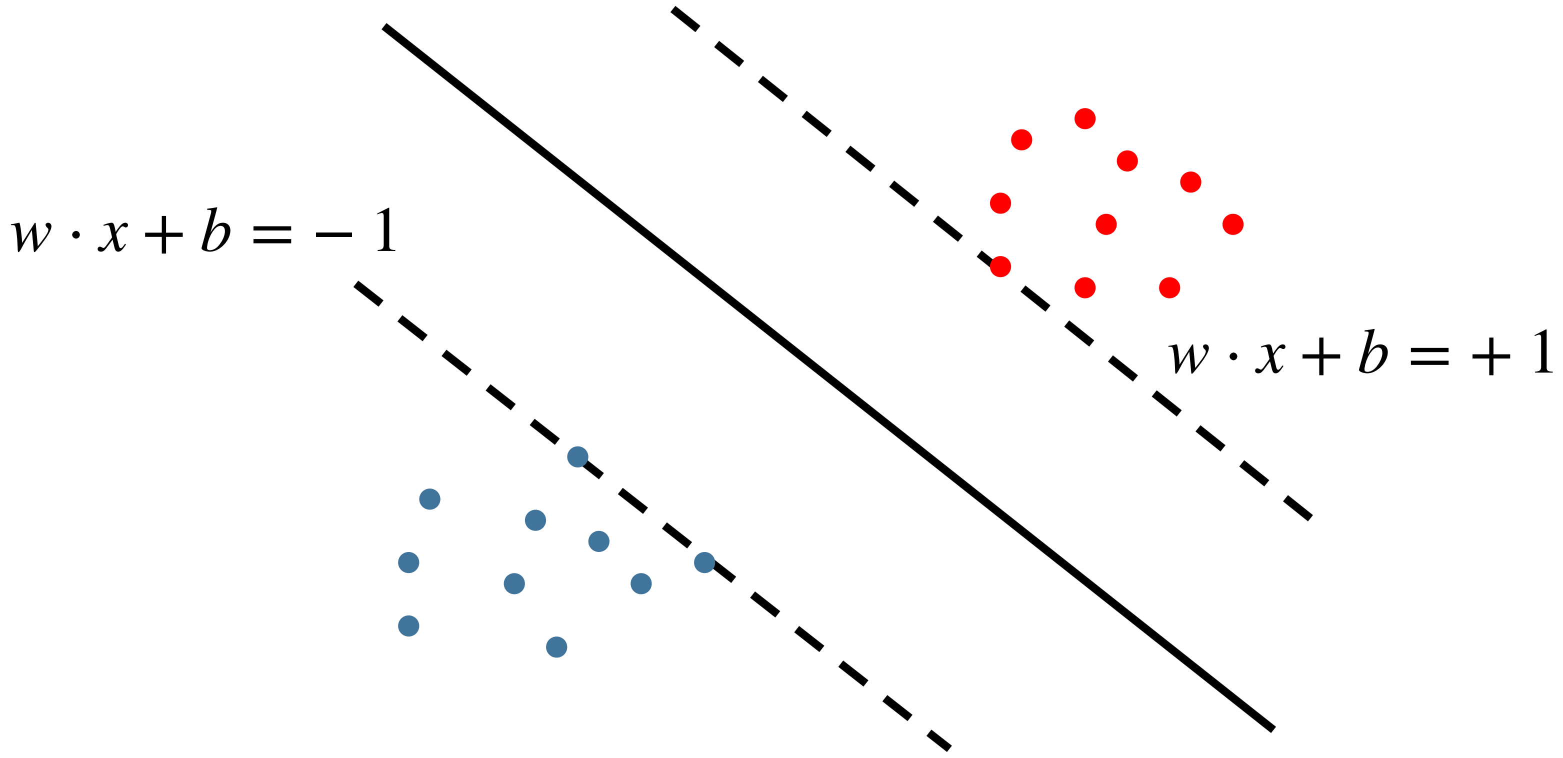
Example: semi-supervised SVM

- Problem: only few instances are labeled
  - Do unlabeled instances violate the margin constraints $y^{(j)}(w \cdot x^{(j)} + b) \geq 1$?
    - We don't know $y^{(j)}$...



- Let's assume labels are correct $\Longrightarrow y^{(j)} = \text{sign}(w \cdot x^{(j)} + b)$
  - Constraint becomes $|w \cdot x^{(j)} + b| \geq 1 \iff x^{(j)}$ outside margin on either side
- Constraints no longer linear
  - Can solve with Integer Programming

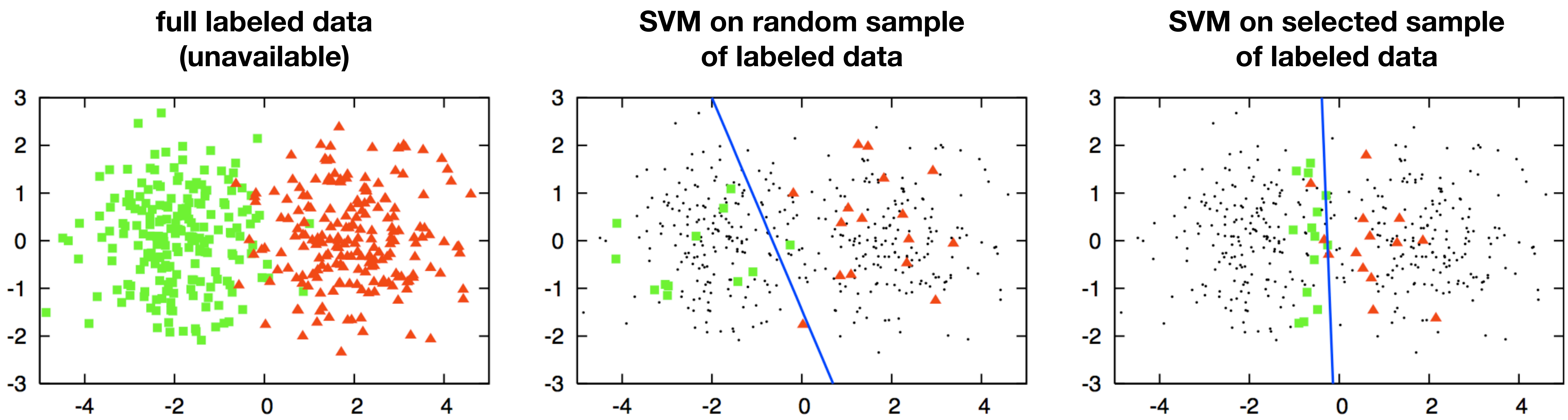
    or other approximation methods



Who selects which instances to label?

- Random = semi-supervised learning
  - Labeled points $\sim p(x, y)$, unlabeled points from marginal distribution $\sim p(x)$
  - Equivalently: select instances $\sim p(x)$, select uniformly which to label $\sim p(y \mid x)$
- Teacher = exact learning, curriculum learning
  - Teacher identifies where learner is wrong, provides corrective labels
  - Some learners benefit from gradual increase in complexity (e.g. boosting)
- Learner = active learning
  - Automate the process of selecting good points to label

Why active learning?



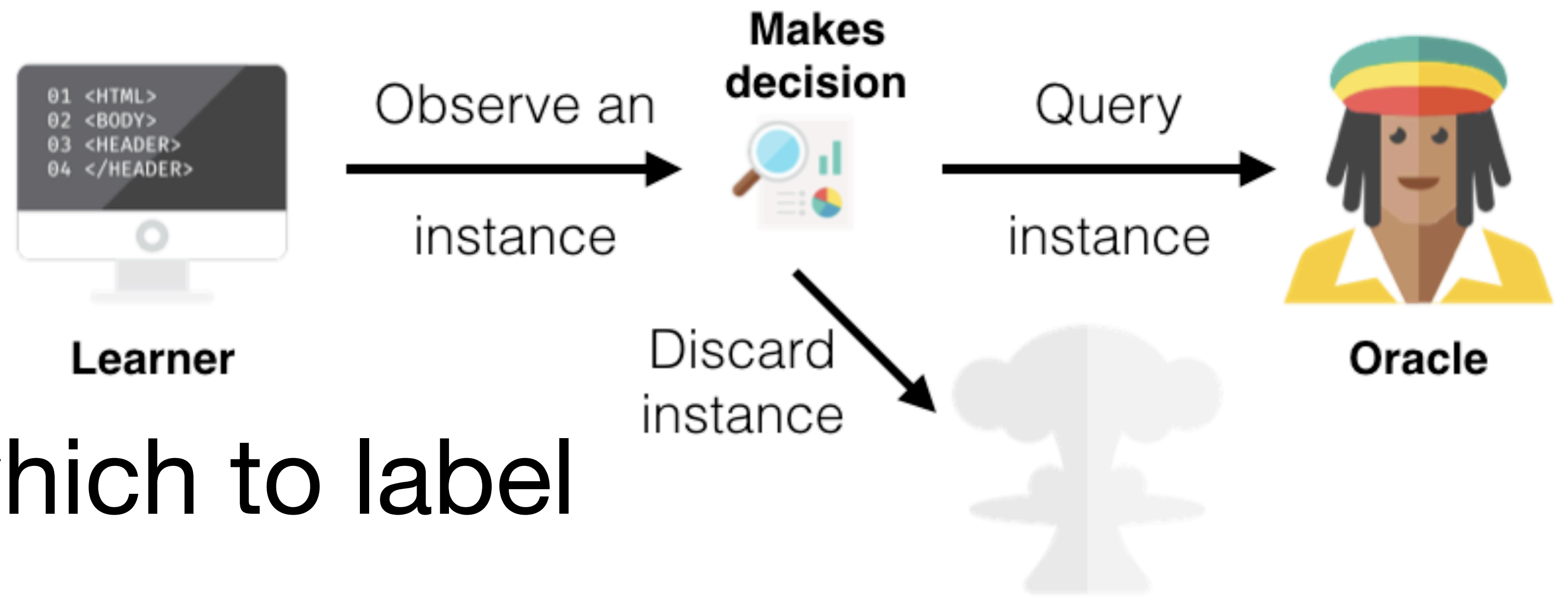
Source: https://www.datacamp.com/community/tutorials/active-learning

- Expensive labels $\Longrightarrow$ prefer to label instances relevant to the decision
- Selecting relevant points may be hard too $\Longrightarrow$ automate with active learning
- Objective: learn good model while minimizing #queries for labels

Active learning settings

- Pool-Based Sampling
  - Learner selects instances in dataset $x \in \mathscr{D}$ to label

- Stream-Based Selective Sampling
  - Learner gets stream of instances $x_1, x_2, \ldots,$ decides which to label

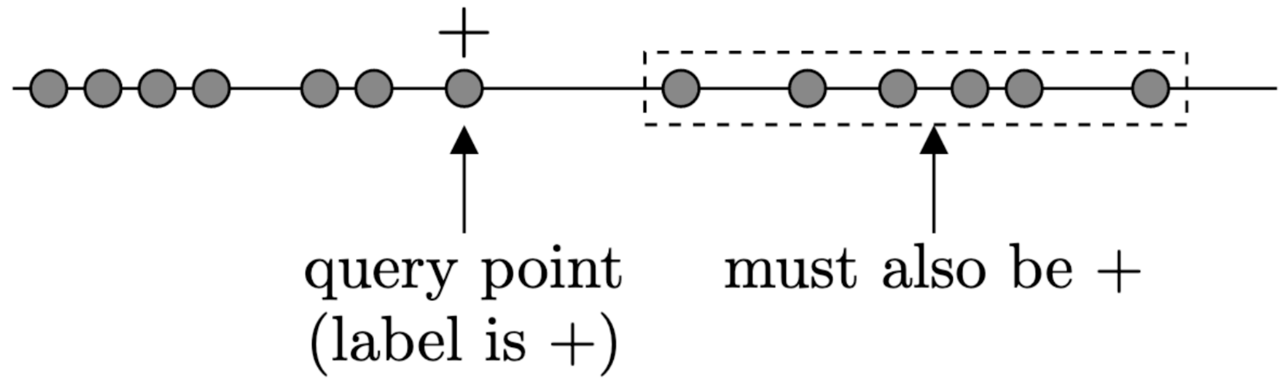


- Membership Query Synthesis
  - Learner generates instance x
  - Doesn't have to occur naturally = $p(x)$ may be low
    - $\implies$ May be harder for teacher to label ("is this synthesized image a dog or a cat?")

Source: https://www.datacamp.com/community/tutorials/active-learning

Simple example: find decision threshold

- When building decision tree on continuous features
  - Where to put the threshold on a given feature?
- If all data points are labeled and sorted $\implies$ binary search
  - Split data in half until you find switch point of $-1 \to +1$
- Active learning = ask for labels



  - Same strategy: query mid point, if -1 / $+1 \implies$ determines left / right half
  - #queries = $\log m$

How to select relevant data points?

- Least Confidence
  - Query point about which learner is most uncertain of the label
  - Requires learner to know its uncertainty, e.g. a probabilistic model $p_\theta(y \mid x)$
- Margin Sampling
  - Multi-class $\Longrightarrow$ least confident doesn't mean least likely to get confused
    - Example: $p_\theta(y \mid x)$ = [0.3, 0.4, 0.3] vs. [0.45, 0.5, 0.05]
  - Query point about which two classes are most similar (near margin between them)
- Entropy Sampling
  - Query point that has most entropy = maximum information gain by revealing true label

Today's lecture

Active learning

Online learning

Sequential decision making

Online learning

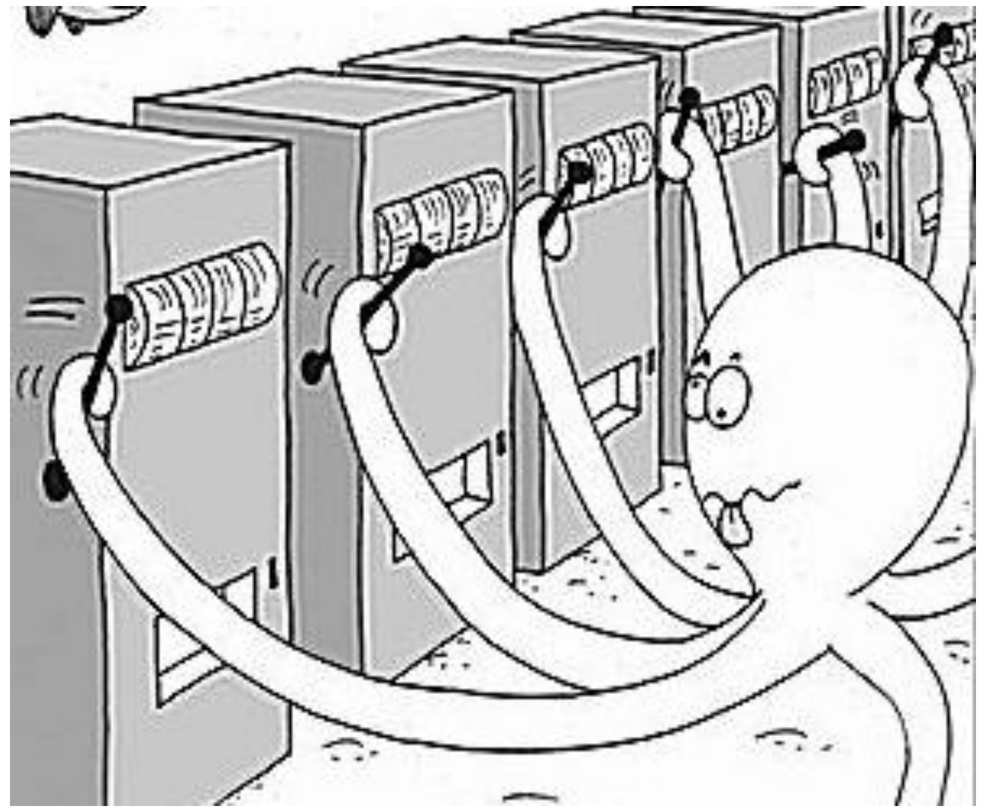
- In multi-class classification, we often assume 0–1 loss $\mathscr{L}(y, \hat{y}) = \delta[y \neq \hat{y}]$
- More generally, we can have different costs $\mathscr{L}(y, \hat{y}) = d(y, \hat{y})$
- Online learning:
  - Stream of instances, need to make predictions / decisions / actions online
  - We don't know the reward = -cost until we actually select $\hat{y}$
  - We'll never know the reward of other actions
- Objective:
  - Make better and better decisions (compared to what? later...)

Multi-Armed Bandits (MABs)

- Basic setting: single instance x, multiple actions $a_1, \ldots, a_k$
  - Each time we take action a_i we see a noisy reward $r_t \sim p_i$
- Can we maximize the expected reward $\max_i \mathbb{E}_{r\sim p_i}[r]$?
  - We can use the mean as an estimate $\mu_i = \mathbb{E}_{r\sim p_i}[r] \approx \frac{1}{m_i} \sum_{t\in T_i} r_t$
- Challenge: is the best mean so far the best action?
  - Or is there another that's better than it appeared so far?

One-armed bandit

Multi-armed bandit



Exploration vs. exploitation

- Exploitation = choose actions that seems good (so far)
- Exploration = see if we're missing out on even better ones
- Naïve solution: learn r by trying every action enough times
    - Suppose we can't wait that long: we care about rewards while we learn
- Regret = how much worse our return is than an optimal action

$$\rho(T) = T\mu_{a*} - \sum_{t=0}^{T-1} r_t$$

- Can we get the regret to grow sub-linearly with T? $\implies$ average goes to 0: $\frac{\rho(T)}{T} \to 0$

Let's play!

- http://iosband.github.io/2015/07/28/Beat-the-bandit.html

Optimism under uncertainty

- Tradeoff: explore less used actions, but don't be late to start exploiting what's known
  - Principle: optimism under uncertainty = explore to the extent you're uncertain, otherwise exploit
- By the central limit theorem, the mean reward of each arm $\hat{\mu}_i$ quickly $\to \mathcal{N}\left(\mu_i, O\left(\frac{1}{m_i}\right)\right)$
- Be optimistic by slowly-growing number of standard deviations: $a = \arg\max_i \hat{\mu}_i + \sqrt{\frac{2 \ln T}{m_i}}$
  - Confidence bound: likely $\mu_i \leq \hat{\mu}_i + c\sigma_i$; unknown constant in the variance $\Longrightarrow$ let c grow
  - But not too fast, or we fail to exploit what we do know
- Regret: $\rho(T) = O(\log T)$, provably optimal

Thompson sampling

- Consider a model of the reward distribution $p_{\theta_i}(r \mid a_i)$
- Suppose we start with some prior $q(\theta)$
  - Taking action a_t, see reward $r_t \Longrightarrow$ update posterior $q(\theta \mid \{(a_{\leq t}, r_{\leq t})\})$
- Thompson sampling:
  - Sample $\theta \sim q$ from the posterior
  - Take the optimal action $a^* = \max_i \mathbb{E}_{r \sim p_{\theta i}}[r]$
  - Update the belief (different methods for doing this)
  - Repeat

Other online learning settings

- What is the reward for action a_i?
  - MAB: random variable with distribution $p_i(r)$
  - Adversarial bandits: adversary selects r_i for every action
    - The adversary knows our algorithm! And past action selection! But not future actions
      - Learner must be stochastic (= unpredictable) in choosing actions
    - Amazingly, there are learners with regret guarantees
- Contextual bandits: we also get instance x, make decision $\pi(a|x)$
  - Can we generalize to unseen instances?

Today's lecture

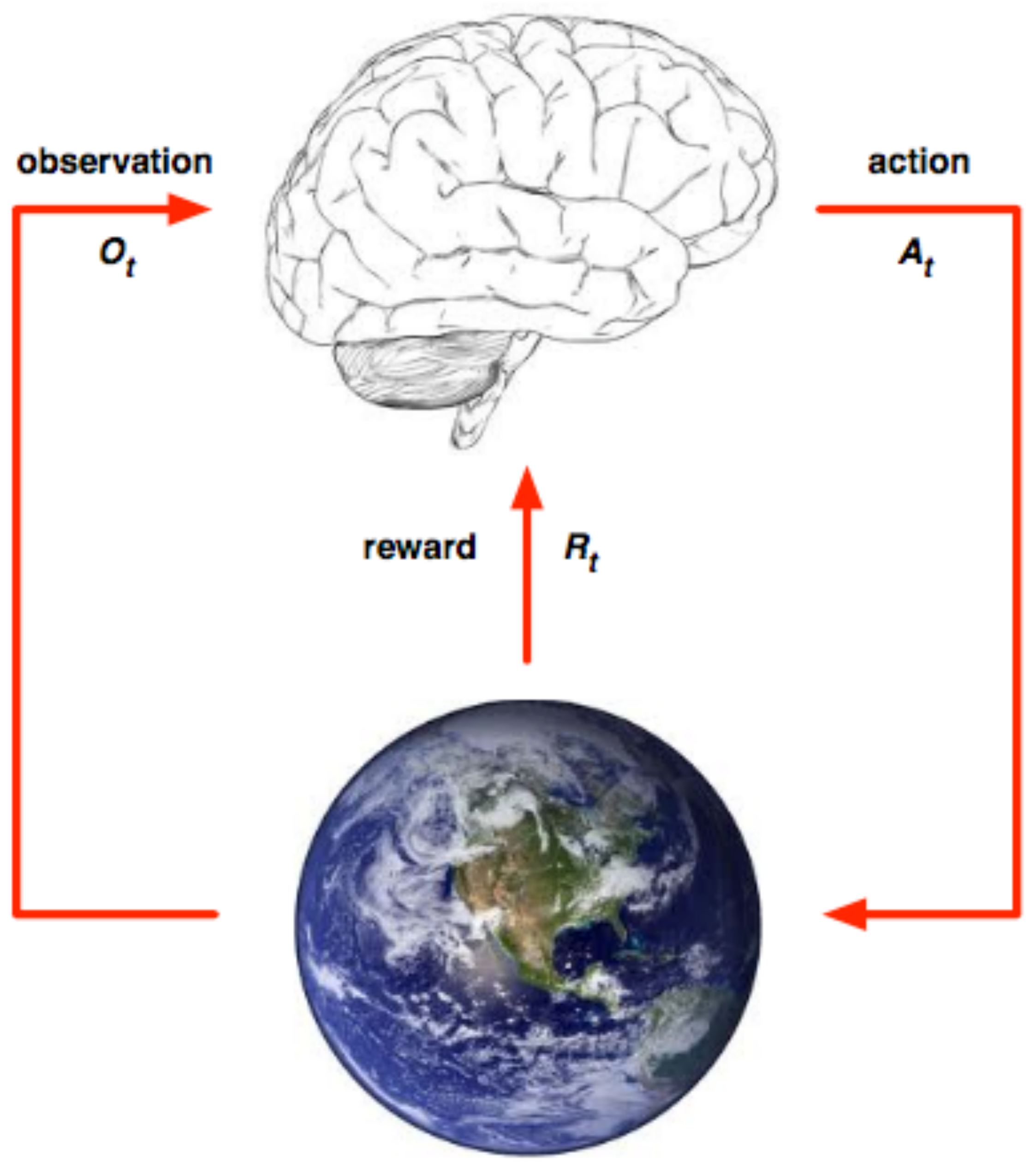
Active learning

Online learning

Sequential decision making

Agent–environment interface

- Agent
  - Decides on next action
  - Receives next reward
  - Receives next observation
- Environment
  - Executes the action → changes its state
  - Generates next observation
  - Supervisor: reveals the reward



Sequential decision making

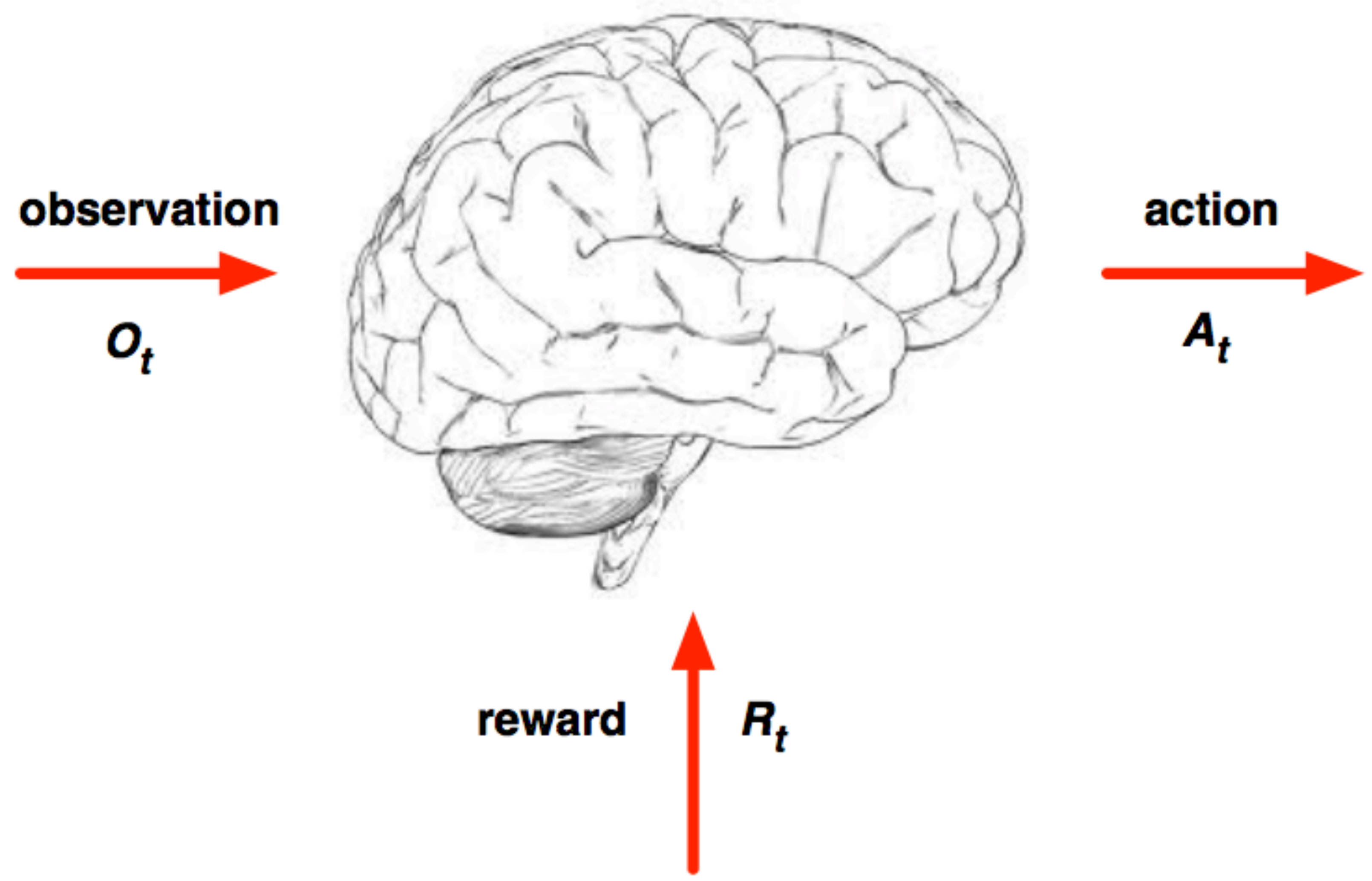
- Reinforcement learning = learning to make sequential decisions
- Challenges:
  - Online learning: reward is only given for actions taken (not for other actions)
  - Active learning: future “instances” determined by what the learner does
  - Sequential decisions: which of the decisions gets credit for a good reward?
- Examples:
  - Fly drone • play Go • trade stocks • control power station • control walking robot
- Rewards: track trajectory • win game • make $ • produce power (safely!)

Long-term planning

- Tradeoff: short-term rewards vs. long-term returns (accumulated rewards)
  - Fly drone: slow down to avoid crash?
  - Games: slowly build strength? block opponent? all out attack?
  - Stock trading: sell now or wait for growth?
  - Infrastructure control: reduce output to prevent blackout?
  - Life: invest in college, obey laws, get started early on course project
- Forward thinking and planning are hallmarks of intelligence

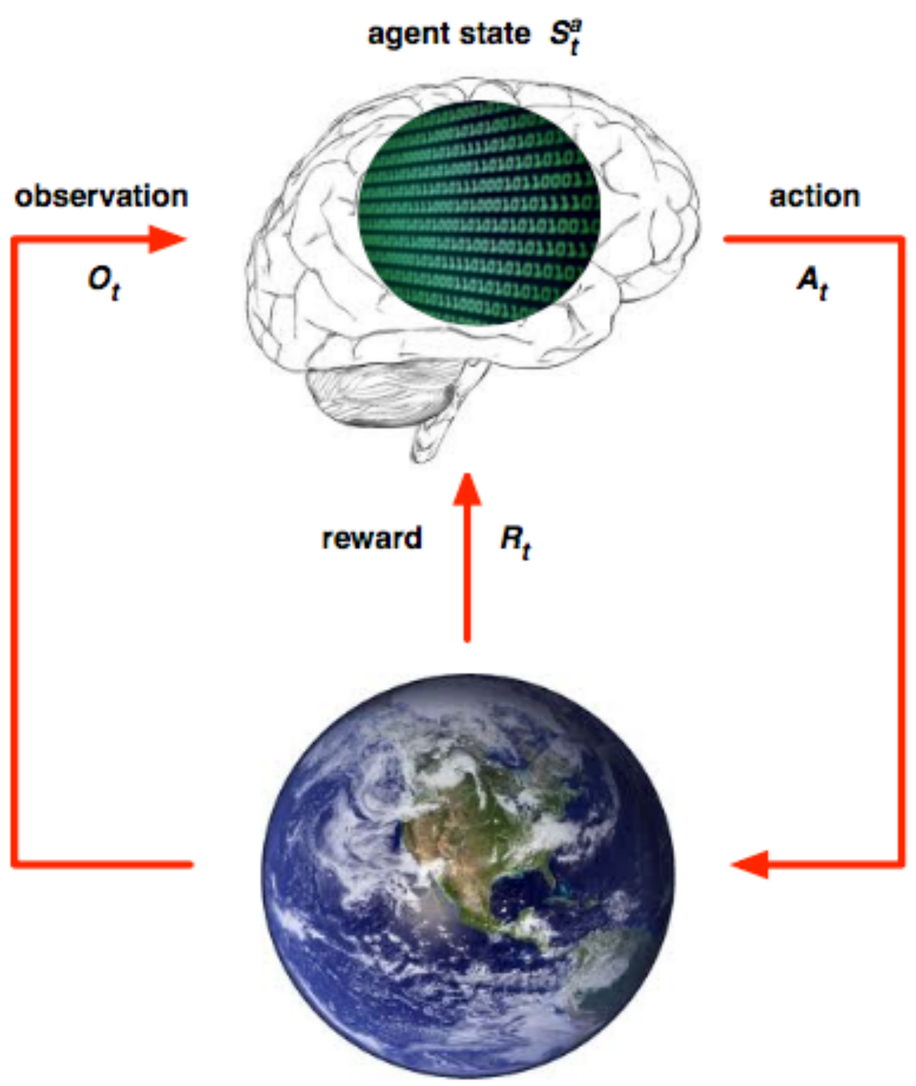
Intelligent agents

- Agent outputs action a_t
  - Function of the context: $a_t = f(x_t)$
    - Perhaps stochastic: $\pi(a_t \mid x_t)$
- What is the context needed for decisions?
  - Ignore all inputs? (open-loop control = sequence of actions)
  - Current observation o_t?
  - Previous action a_{t-1}? reward r_{t-1}?
  - All observations so far $o_{\leq t}$?

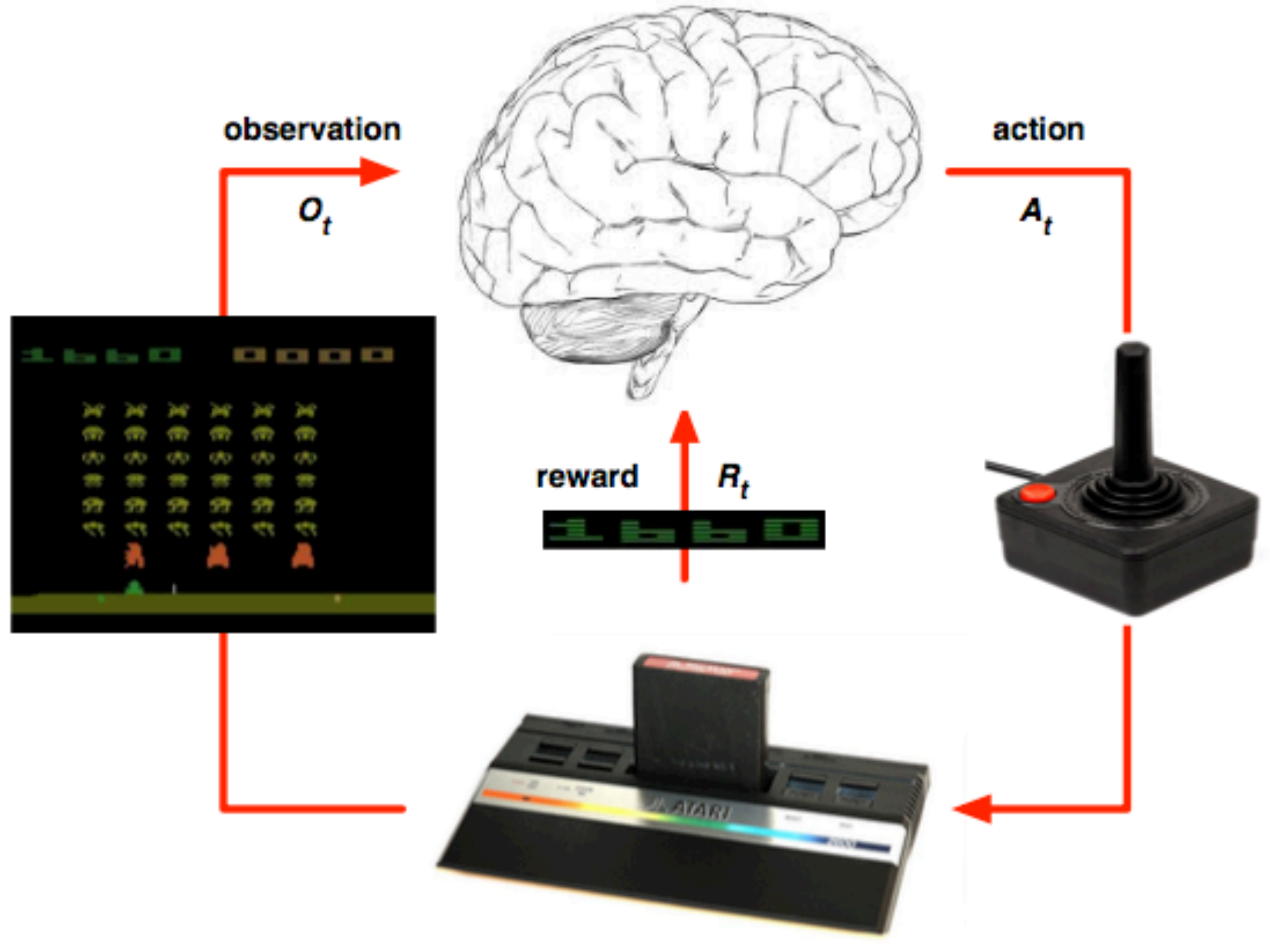


Agent context x_t

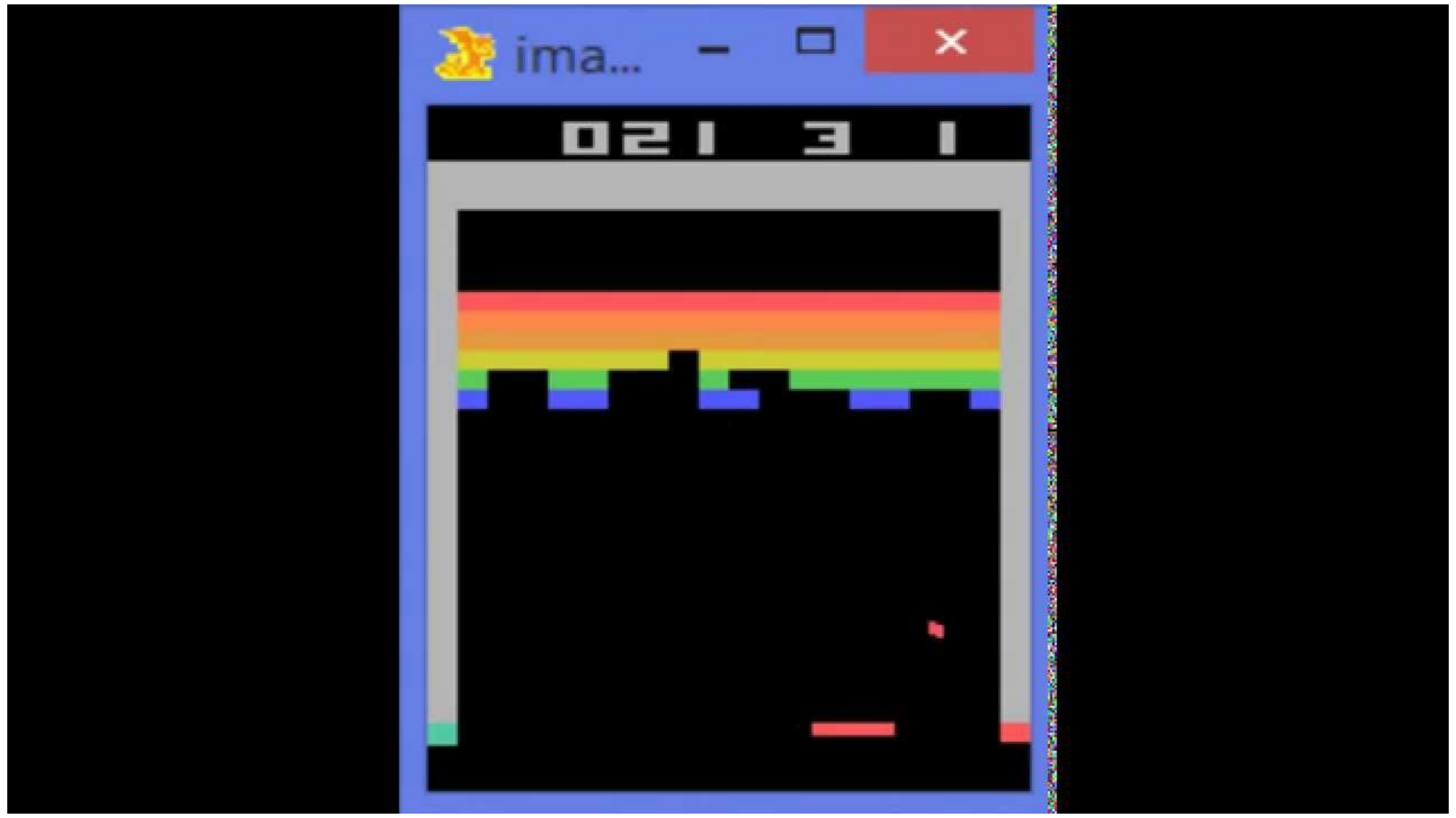
- Observable history: everything the agent saw so far
  - $h_t = (o_1, a_1, r_1, o_2, \ldots, a_{t-1}, r_{t-1}, o_t)$
- The context x_t used for the agent's policy $\pi(a_t | x_t)$ can be:
  - Reactive policy: $x_t = o_t$ (optimal under full observability: $o_t = s_t$)
  - Using previous action: $x_t = (a_{t-1}, o_t) \implies$ can be useful if policy is stochastic
  - Using previous reward: $x_t = (r_{t-1}, o_t) \implies$ extra information about the environment
  - Window of past observations: $x_t = (o_{t-3}, o_{t-2}, o_{t-1}, o_t) \implies$ better see dynamics
  - Generally: any summary (= memory) of observable history $x_t = f(h_t)$



Example: Atari

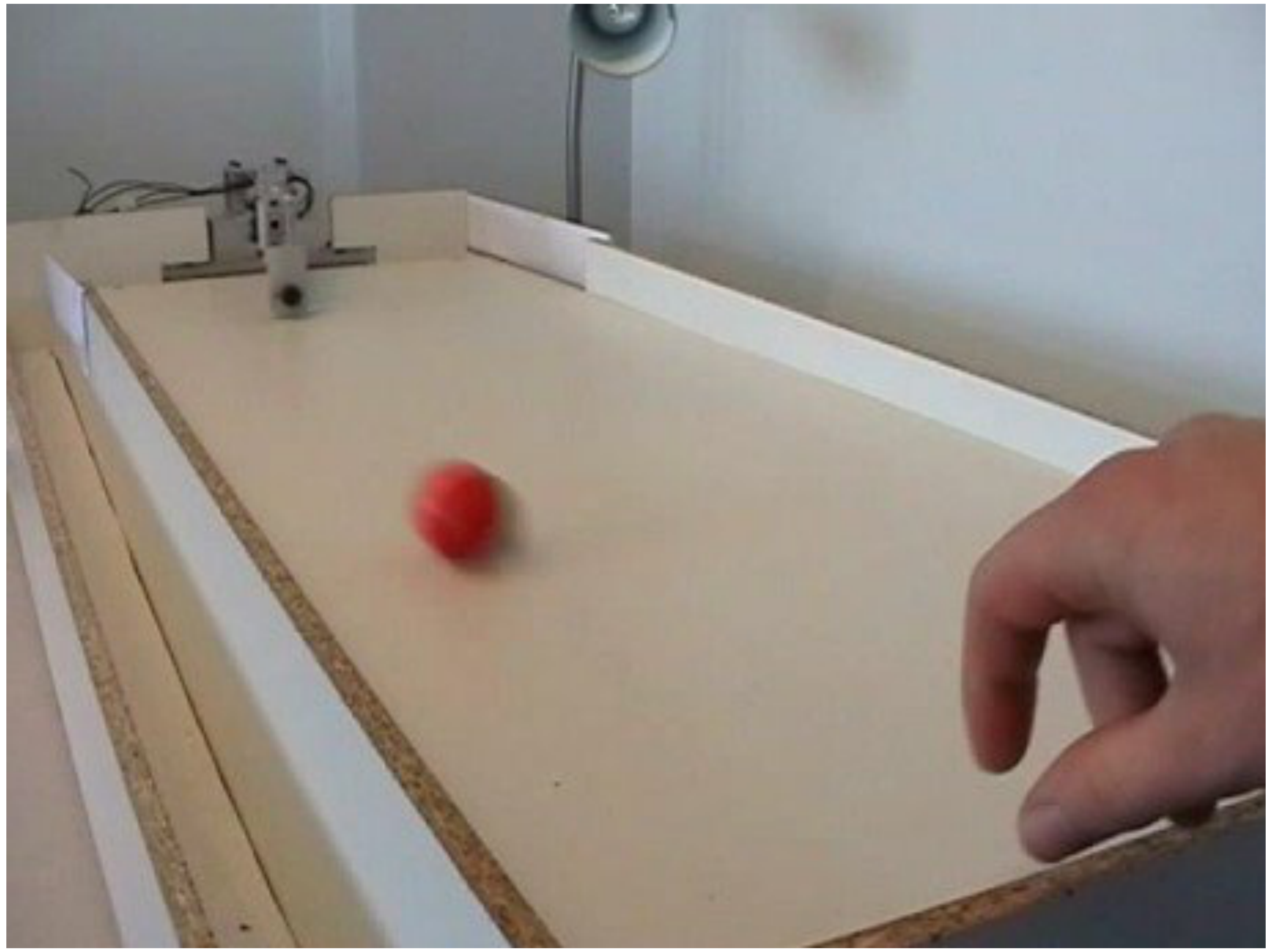


- Rules are unknown
  - What makes the score increase?
- Dynamics are unknown
  - How do actions change pixels?



https://www.youtube.com/watch?v=V1eYniJ0Rnk

Example: Table Soccer



https://www.youtube.com/watch?v=CIF2SBVY-J0

Logistics

assignments	• Assignment 5 due Thursday
project	• Final report due next Thursday
evaluations	• Evaluations due end of next week
final exam	• Review: next Thursday • Final: Thursday, March 18, 1:30–3:30pm