

AIL 8027 / AIL 841

Advanced RL

Sem 2, 2025-26

Lec 1: Course Introduction

06 Jan 2025

Raunak P. Bhattacharyya, ScAI

SL & RL



[Source](#)



[Source](#)

Capability of Autonomous Agents



Source: Scimex



Source: Ghost Robotics



Source: Halo Robotics



Source: Waymo

Exemplar: Self-Driving Cars



[Source](#)

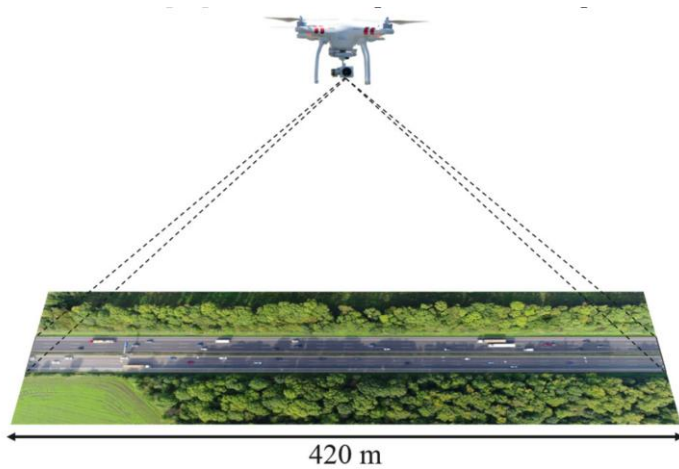
How Does a Waymo Merge into Traffic?



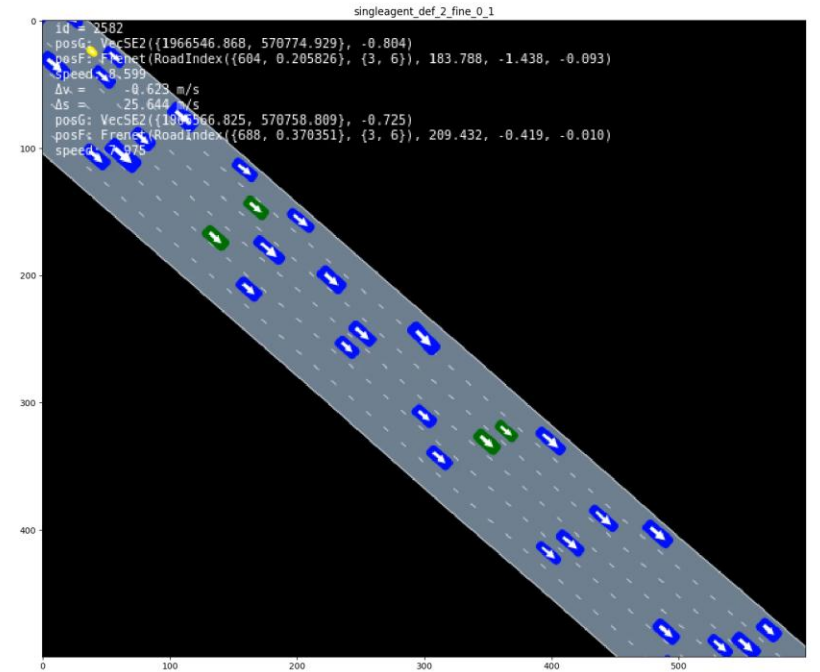
[Source](#)

Can safe autonomous behavior be **learned**?

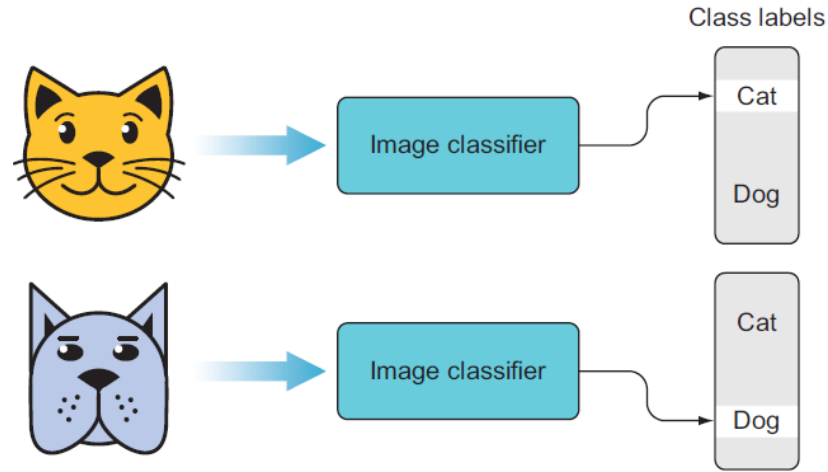
Supervised Learning..., right?



Learn a mapping between situation (state)
to acceleration and steering (action)



Supervised Learning vs. Reinforcement Learning



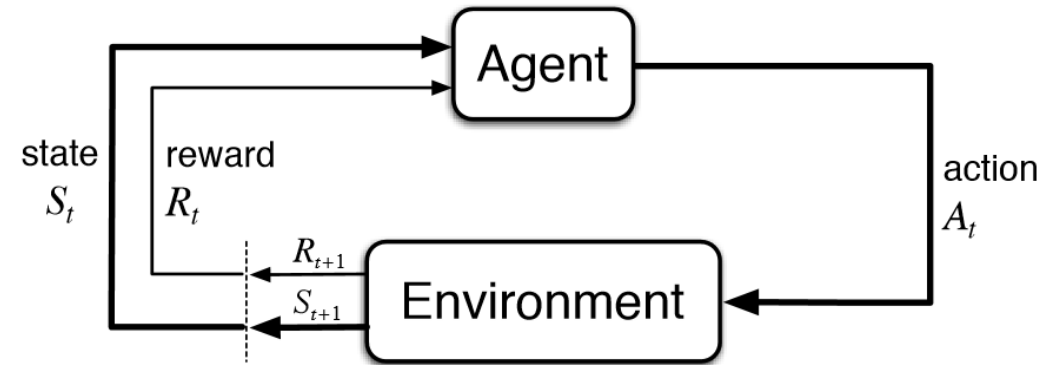
[Source](#)

Input: x

Output: y

Data: $D = \{(x_i, y_i)\}$

Goal: $f_{\theta}(x_i) \approx y_i$



[Source](#)

Input: State s_t at each time step

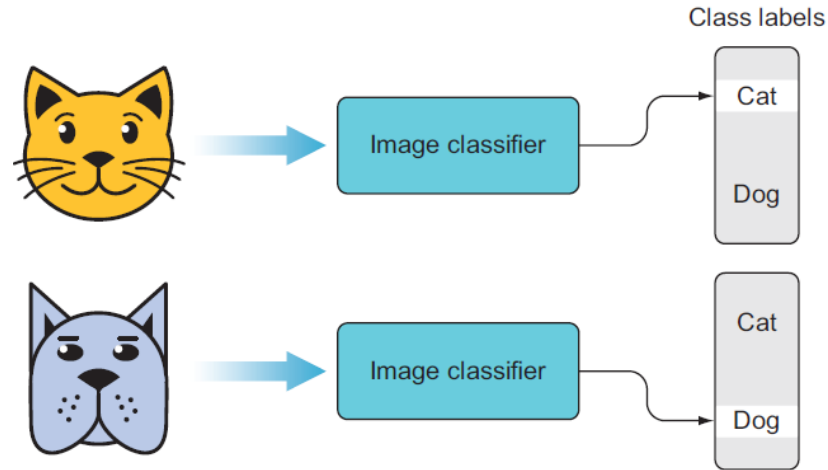
Output: Action a_t at each corresponding time step

Data: $(s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_T, a_T, r_T)$

Goal: Learn policy $\pi_{\theta} : s_t \rightarrow a_t$

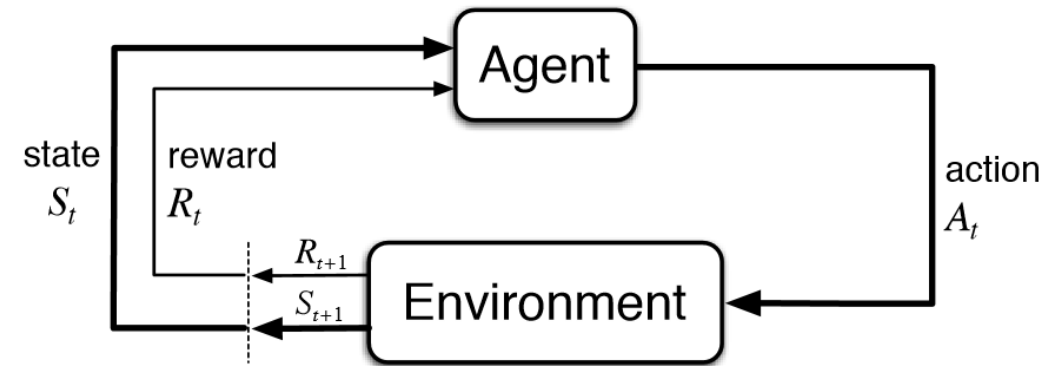
to maximize total reward obtained

SL vs. RL: Inputs and Outputs



[Source](#)

- Someone gives you the labels
 - Known ground truth during training
- i.i.d. assumption

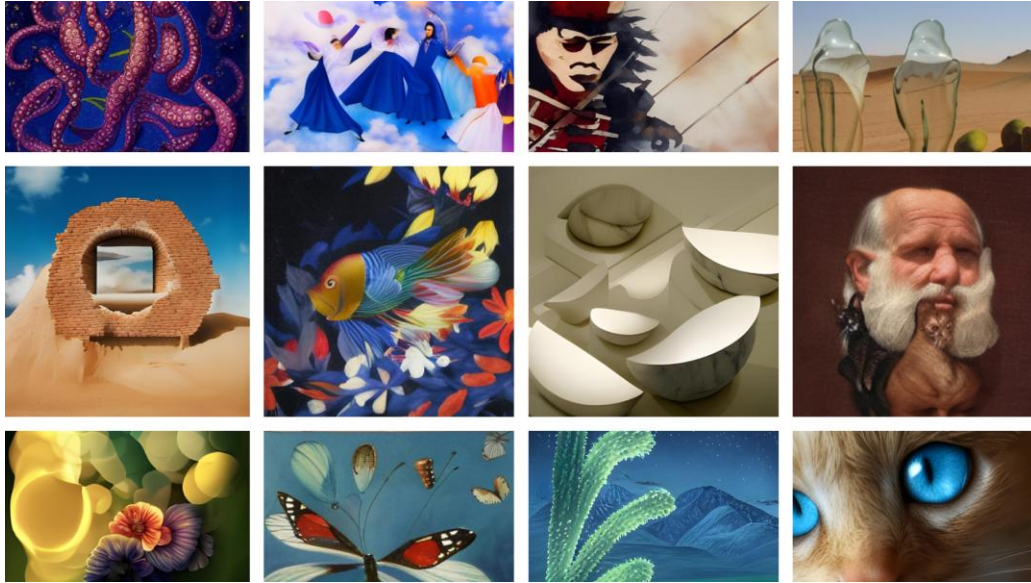


[Source](#)

- No ground truth
 - Instead, we have rewards
- Data is not i.i.d.
 - Prev o/p influence future i/p

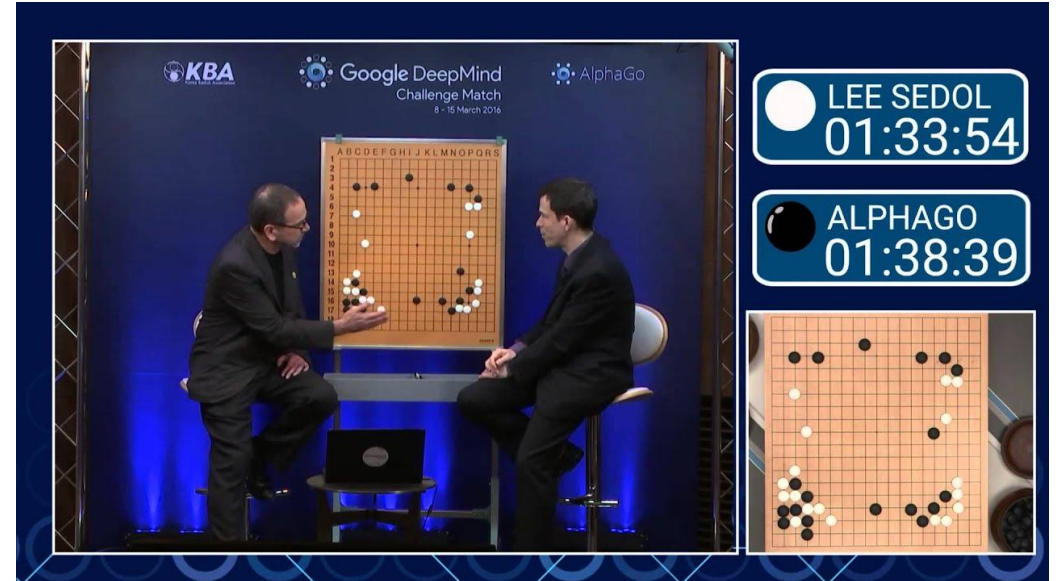
SL and Distribution Shift

SL vs. RL: Causes of Excitement



[Source](#)

Something a person may have drawn



[Source](#)

Unexpected: Better than what a human may have done

Contextualising RL

Retrieval

Reasoning



Reflex-based

State-based

Variable-based

Logic

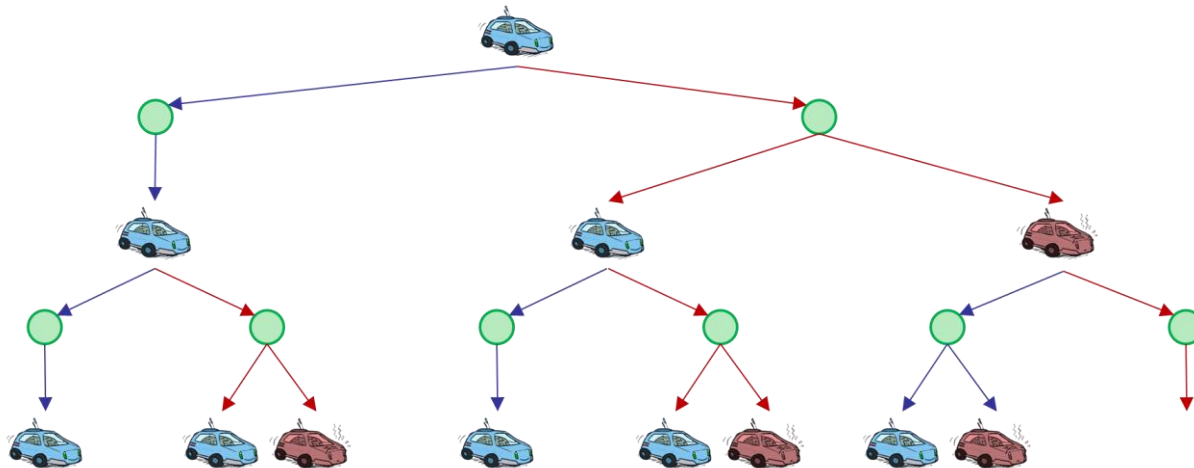
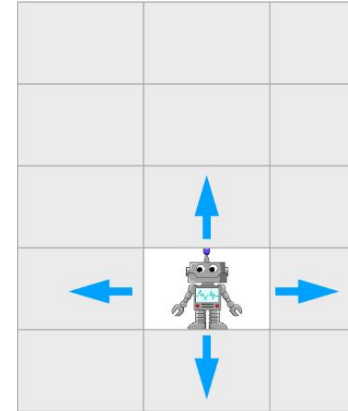
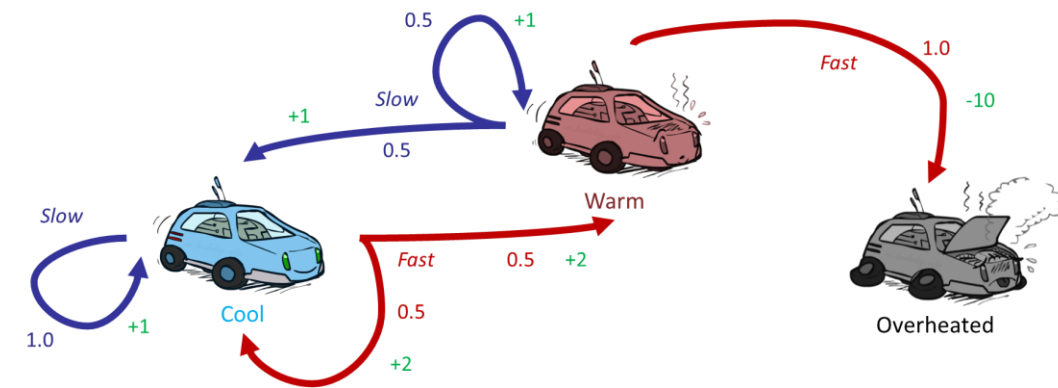
Regression,
Nearest
Neighbours

Search,
Markov
Decision
Processes

Constraint
Satisfaction,
Bayes Nets

Temporal
Logic, Model
Checking

Example MDPs



-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.2
-0.1	0	0	0	0	0	0	0	0	-0.1
-0.1	0	0	0	0	0	0	3	0	-0.1
-0.1	0	0	0	0	0	0	0	0	-0.1
-0.1	0	0	-5	0	0	0	0	0	-0.1
-0.1	0	0	0	0	0	0	0	0	-0.1
-0.1	0	0	0	0	0	0	0	0	-0.1
-0.1	0	0	-10	0	0	0	0	10	-0.1
-0.1	0	0	0	0	0	0	0	0	-0.1
-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.2

Example MDP: Queuing Problem



- Line initially empty
- Two modes of service
- Each timestep, new cust with p
- Horizon T
- Waiting cost: $Y \cdot \text{queue length}$

$\mathcal{S} = \{0, 1, 2, \dots\}$: Length of the queue x_t $x_0 = 0$

$\mathcal{U} = \{\text{Fast (F)}, \text{Slow (S)}\}$ Completion probs: $q(F) > q(S)$

$c(x_t, u_t) = \gamma x_t + d(u_t)$ Service costs: $d(F) > d(S)$

If $x = 0$:

$$p(x' = 1 \mid x = 0, u = F/S) = p$$

$$p(x' = 0 \mid x = 0, u = F/S) = 1 - p$$

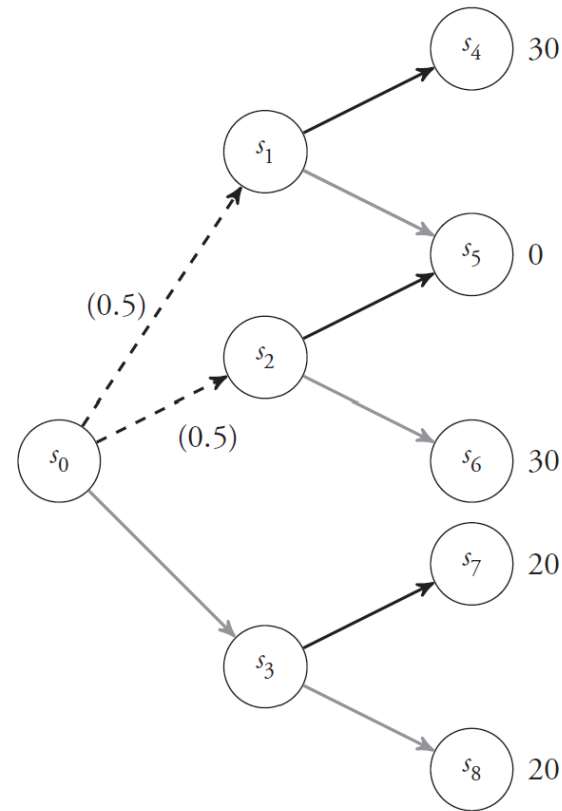
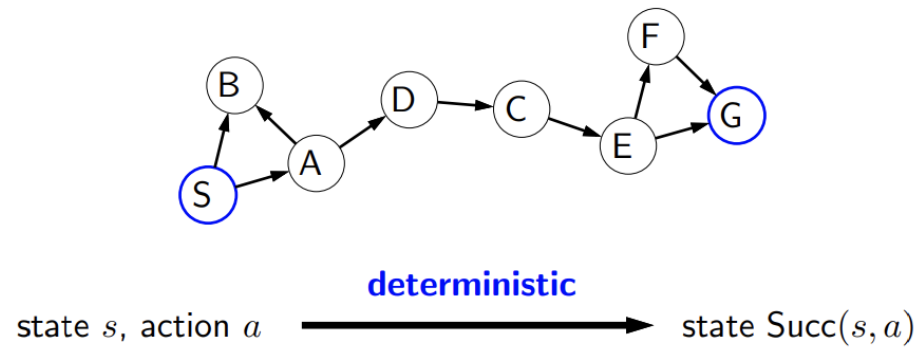
If $x > 0$:

$$p(x' = x + 1 \mid x, u) = p \cdot (1 - q(u))$$

$$p(x' = x \mid x, u) = (1 - p) \cdot (1 - q(u)) + p \cdot q(u)$$

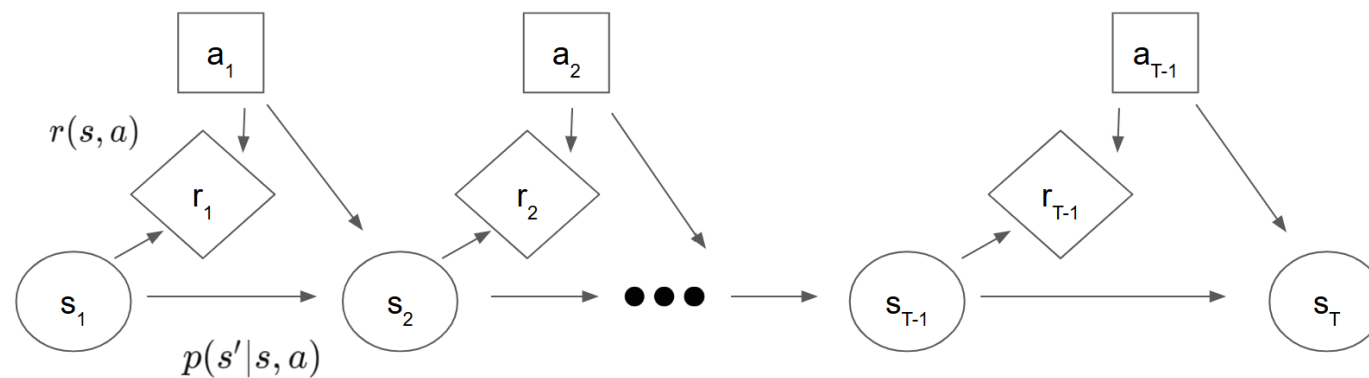
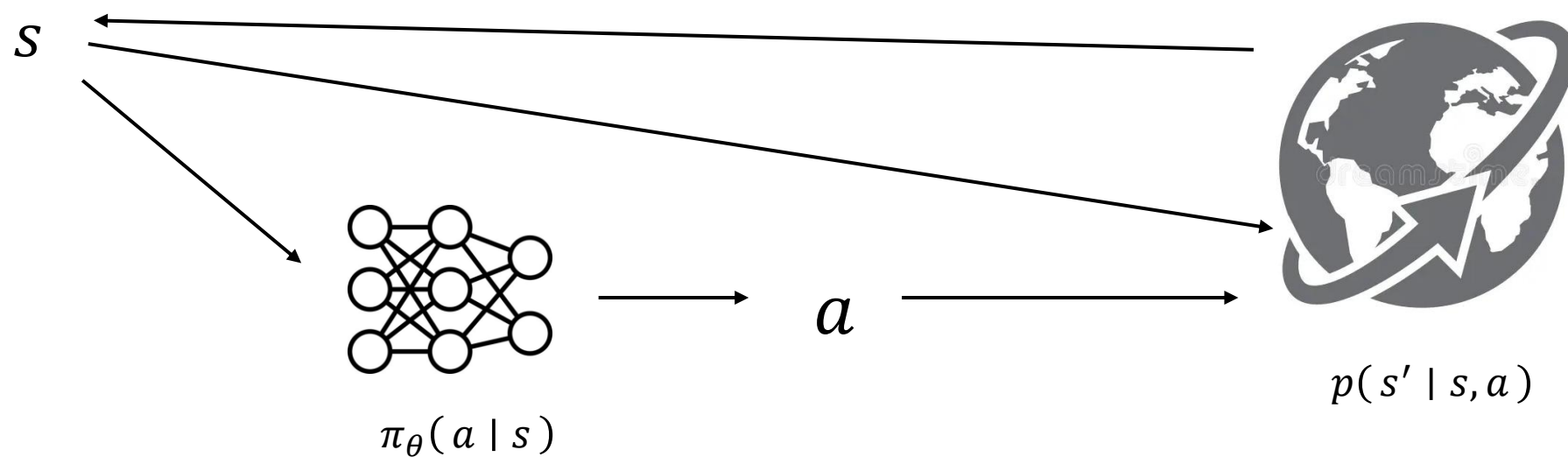
$$p(x' = x - 1 \mid x, u) = q(u) \cdot (1 - p)$$

Policy (vs. Plan)

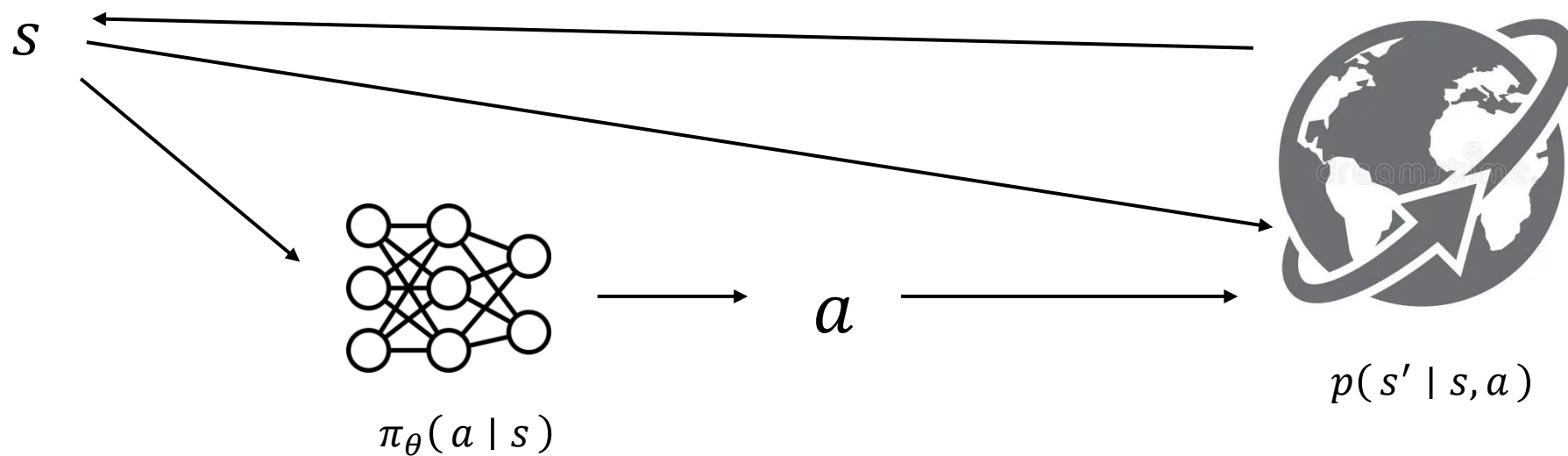


$$\begin{aligned} U(\text{up}, \text{up}) &= 0.5 \times 30 + 0.5 \times 0 = 15 \\ U(\text{up}, \text{down}) &= 0.5 \times 0 + 0.5 \times 30 = 15 \\ U(\text{down}, \text{up}) &= 20 \\ U(\text{down}, \text{down}) &= 20 \end{aligned}$$

The Objective



The Objective



$$\theta^* = \arg \max_{\theta} E_{\tau \sim p_{\theta}(\tau)} \left[\sum_t r(s_t, a_t) \right]$$

Course Contents

- Why do VI and PI work?
 - Proof of Bellman optimality
 - Convergence of Iterative Policy Evaluation
 - Monotonic Improvement in PI
- Deep RL using policies
 - Ideas from Optimisation
 - Approximate PI, Conservative PI, Natural Policy Gradient
 - TRPO, PPO
- Probabilistic Graphical Model view of RL
 - Entropy maximization and soft algorithms: SAC

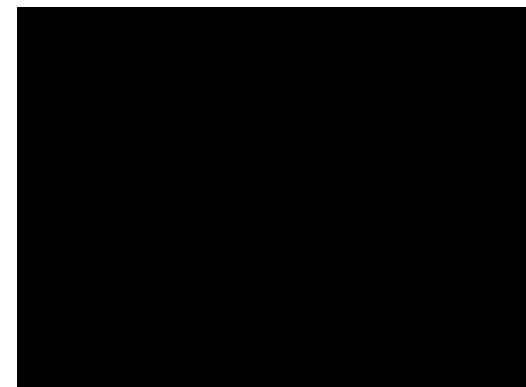
Journey of RL



[Source](#)



[Source](#)



[Source](#)

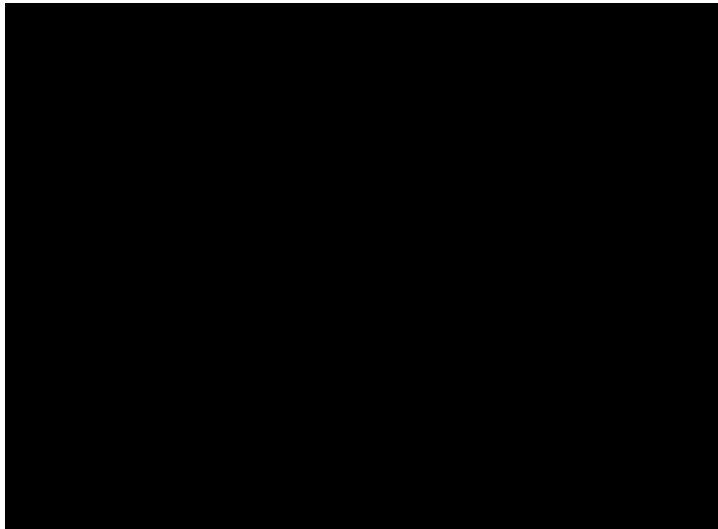


[Source](#)

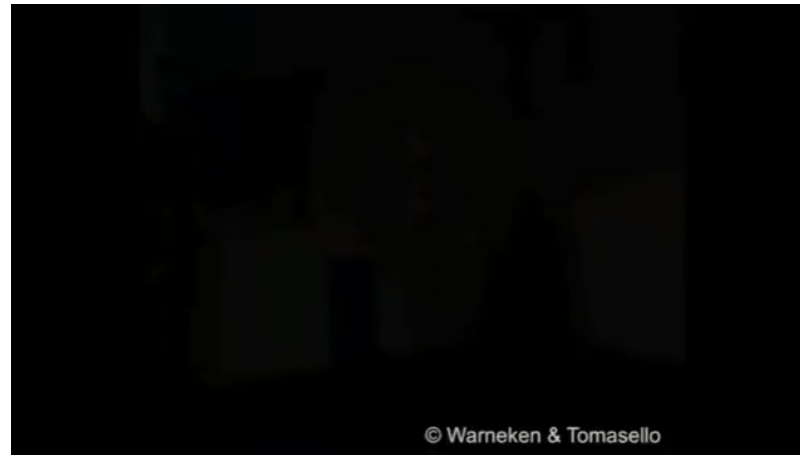


[Source](#)

Learning Without Rewards



[Source](#)



[Source](#)



[Source](#)

Course Contents (contd.)

- Learning without rewards
 - Imitation learning, IRL
- Learning without environment
 - Offline RL
- Safe RL
 - CMDPs, Cost Critics
- Distributional RL
 - CVaR, Quantile Regression
- RL for LLMs
 - RL with sequence models, Learning from preferences