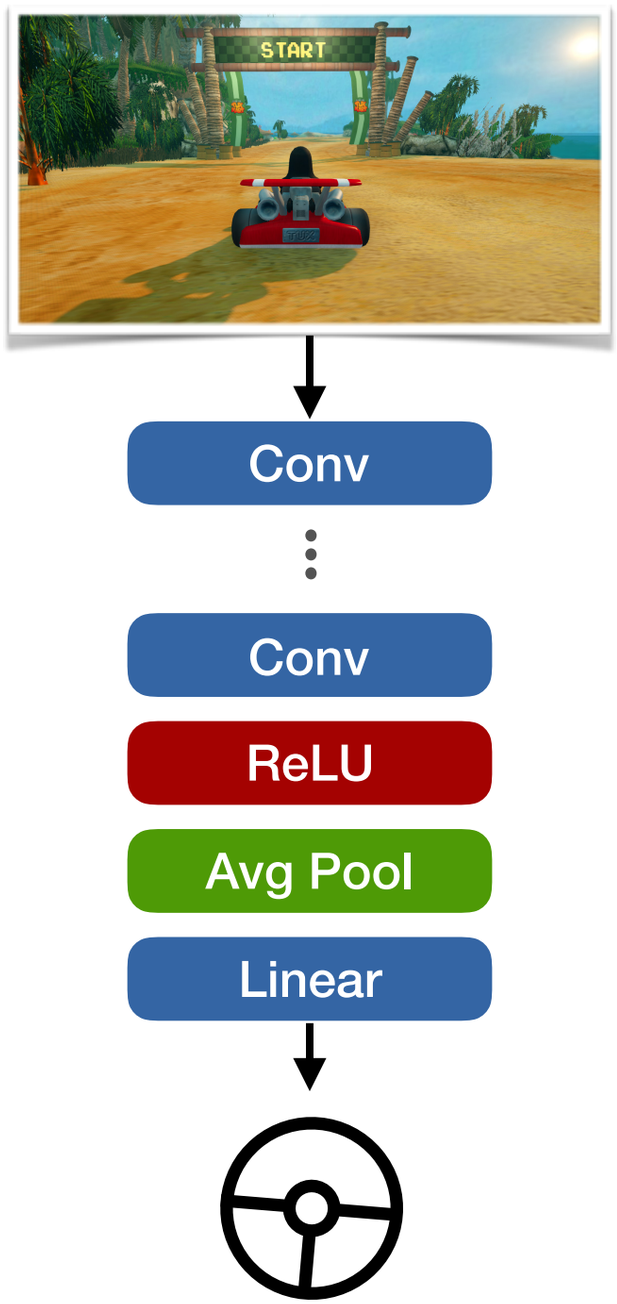
Acting in an environment

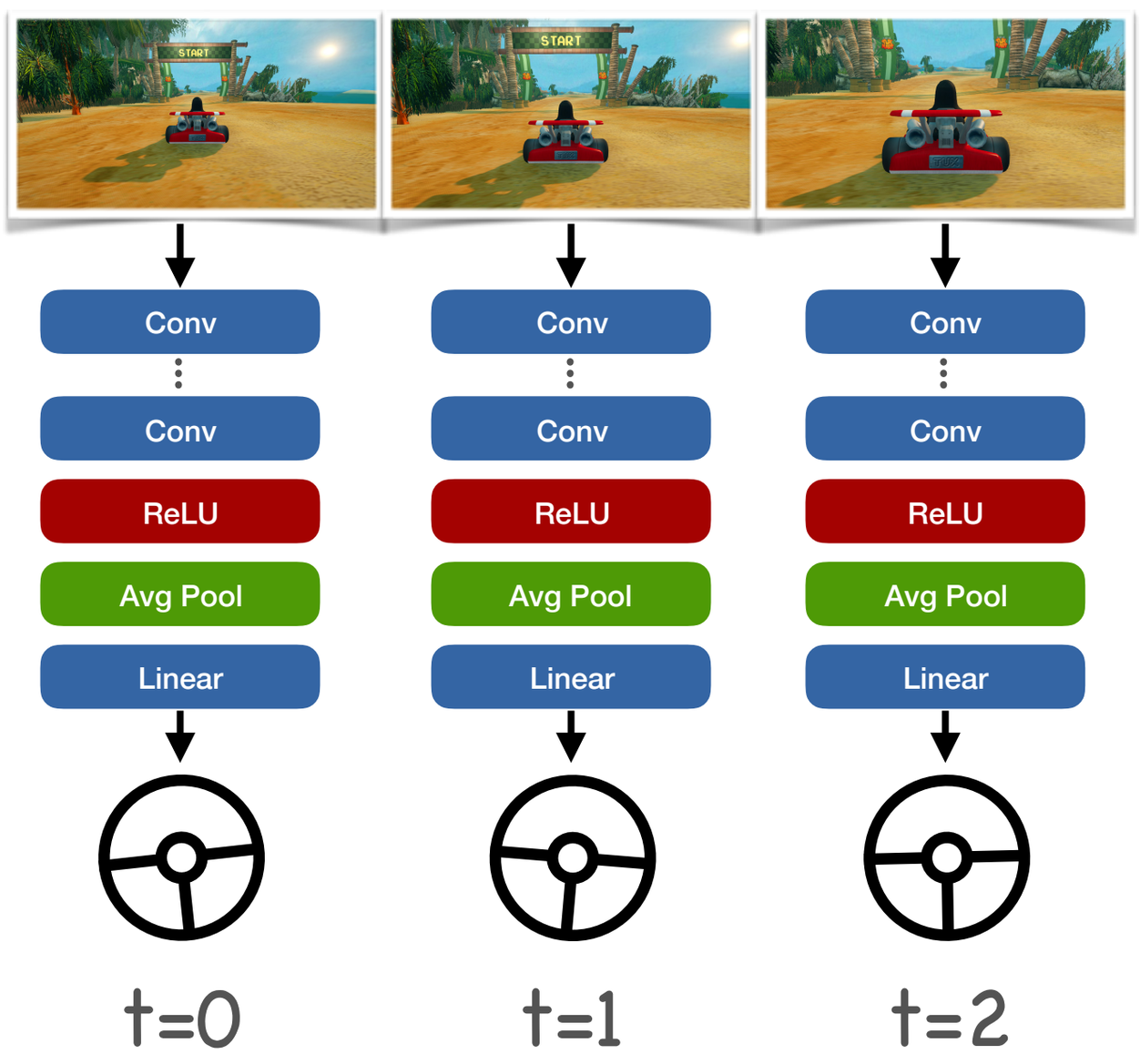
Deep learning for action

- Input
  - Observation
- Output
  - Action



Acting in an environment

- Action changes that state the of the world
- Non-differentiable
- Often non-repeatable
- Long-range dependencies



Acting in an environment

state /
observation

agent

action

left

right

up

jump

down

shoot

environment

Acting in an environment

Video Game

environment

state / observation

Deep Network

agent

action

left

right

up

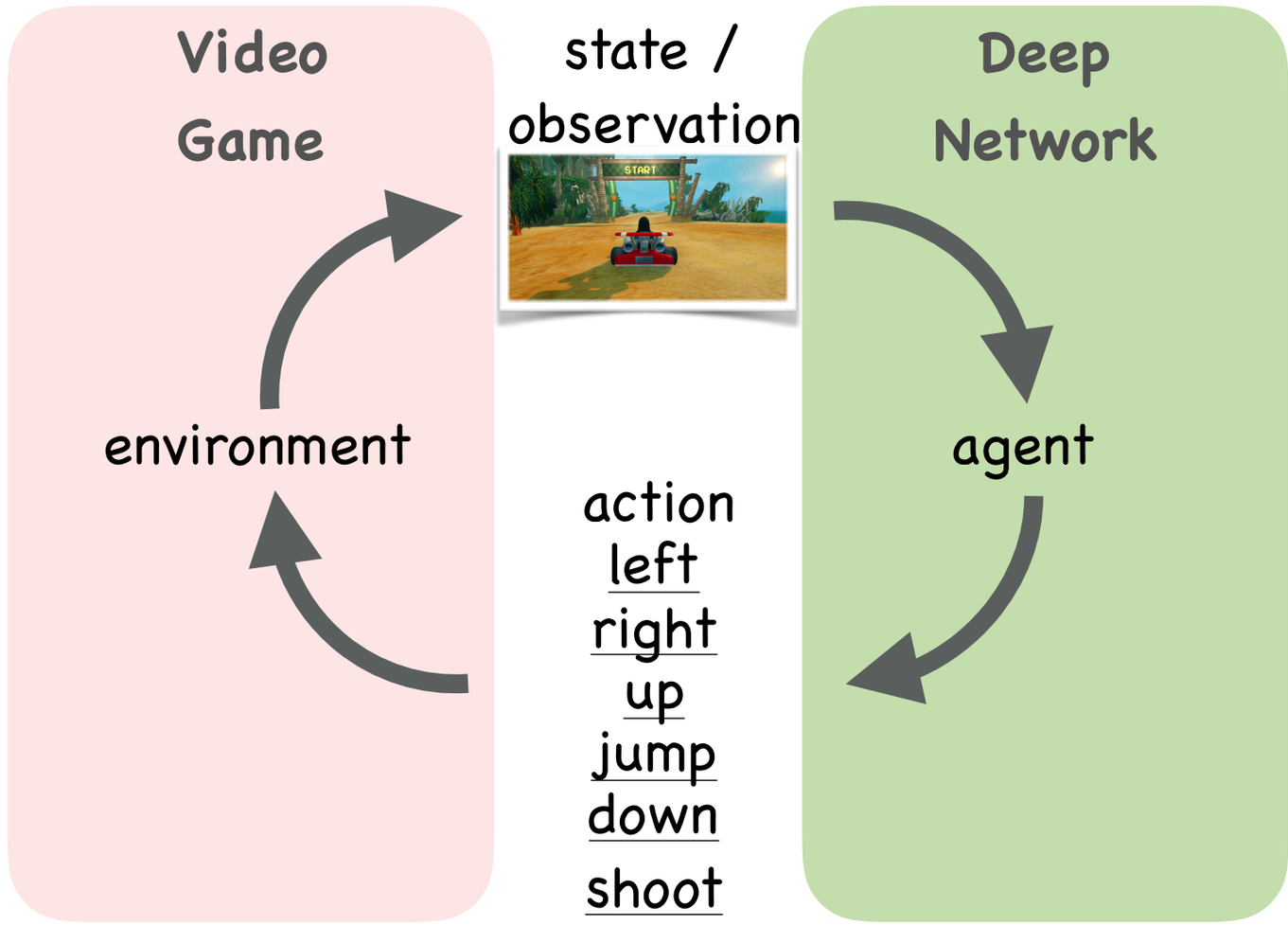
jump

down

shoot

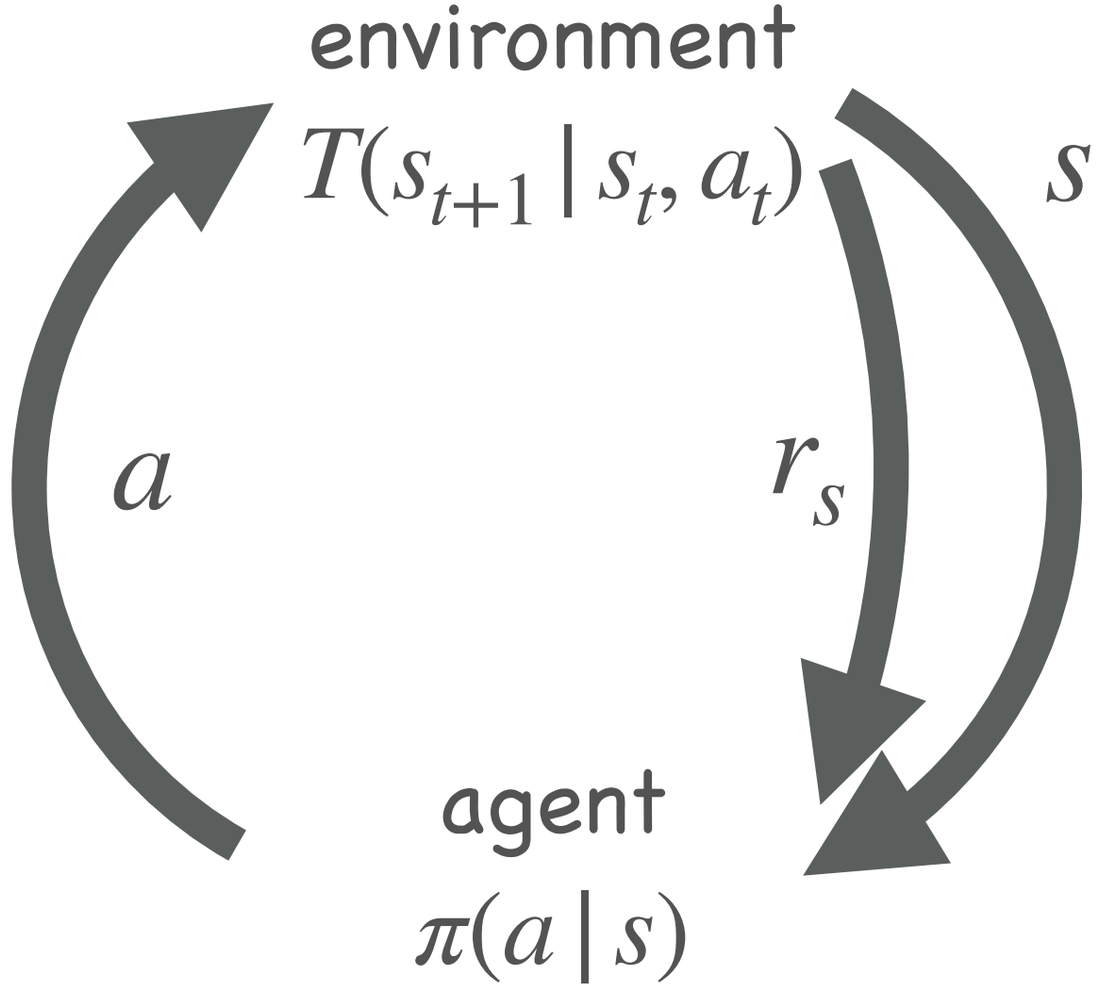
How to train the agent?

- What should the agent learn to do?
  - Minimize loss
    - Reward from environment



Markov decision process (MDP) - Formal definition

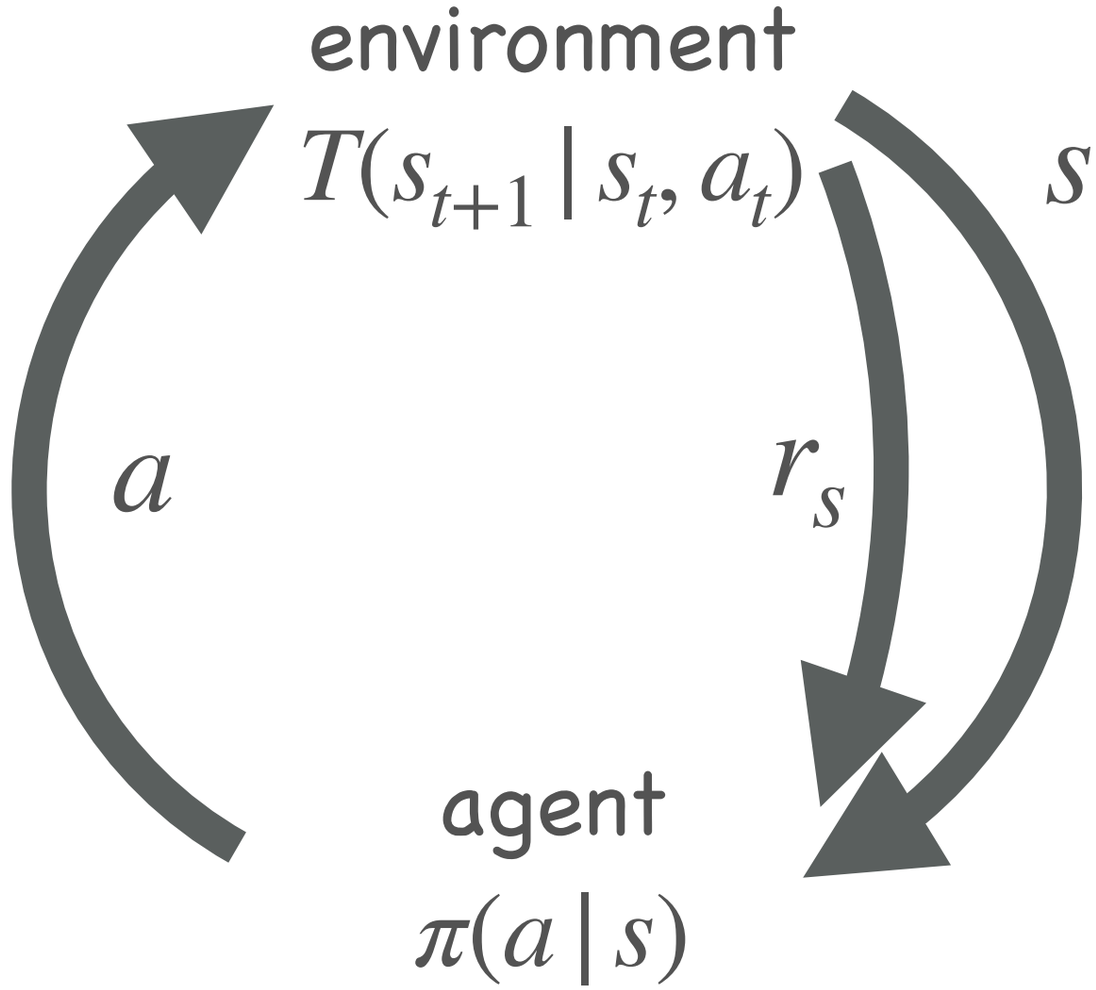
- state $s \in \mathrm{S}$
- action $a \in \mathrm{A}$
- reward $r_s \in \mathbb{R}$
- transition $T(s_{t+1} \mid s_t, a_t)$
- policy $\pi(a \mid s)$



MDP - objective

- Trajectory
  $\tau = \{s_0, a_0, s_1, a_1, \dots\}$

- Return
  $R(\tau) = \sum_t \gamma^t r_{s_t}$

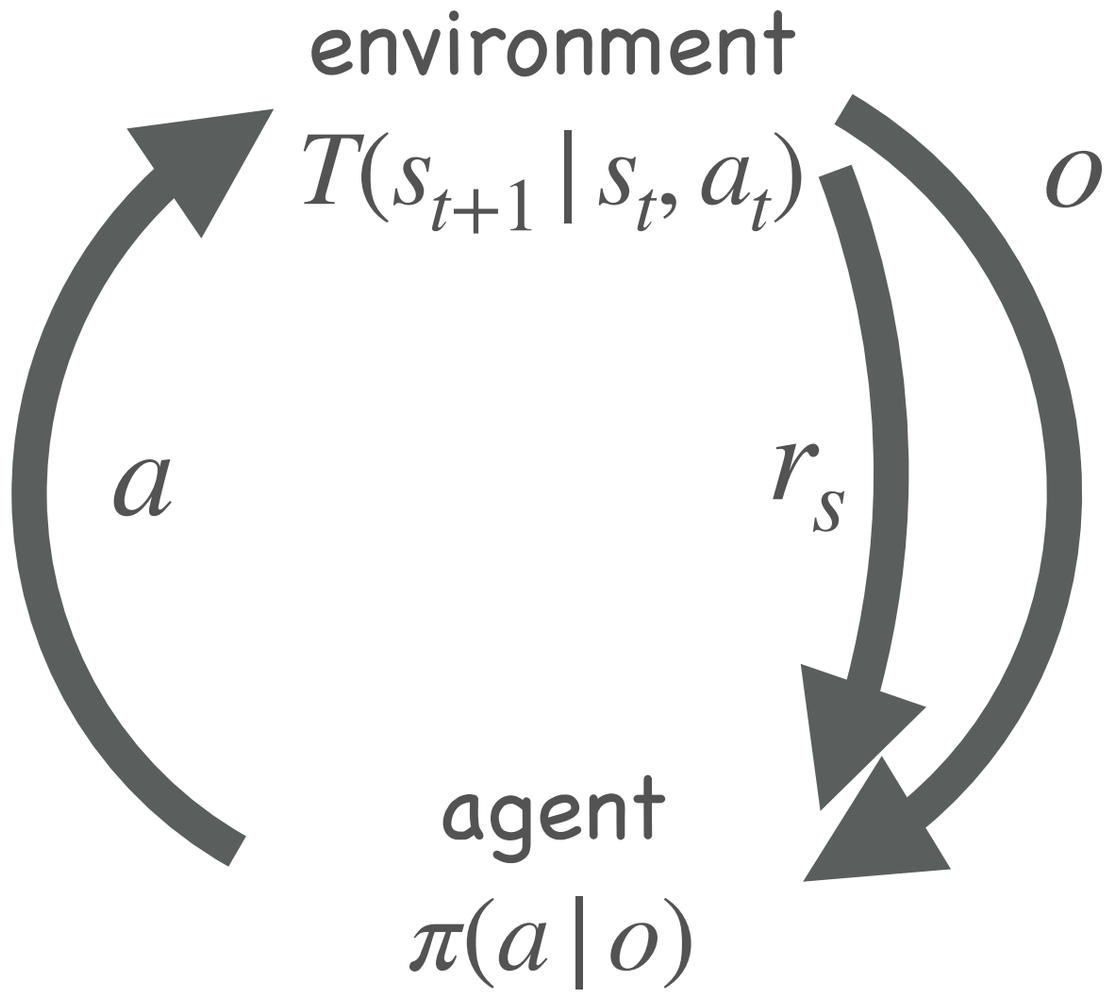


- Objective

  - $\text{maximize}_\pi \; \mathbb{E}_{\tau \sim P_{\pi,T}}[R(\tau)]$

Partially observed Markov decision process (POMDP)

- state $s \in \mathrm{S}$
- action $a \in \mathrm{A}$
- reward $r_s \in \mathbb{R}$
- transition $T(s_{t+1} \mid s_t, a_t)$
- **observation** $o \in \mathrm{O}$
- **observation function** $O(o \mid s)$
- **policy** $\pi(a \mid o)$



Examples - Cart-pole

- MDP
- **objective**: balance a pole on movable cart
- **state**: angle, angular velocity, position, velocity
- **action**: force applied
- **reward**: 1 for each time step pole is upright

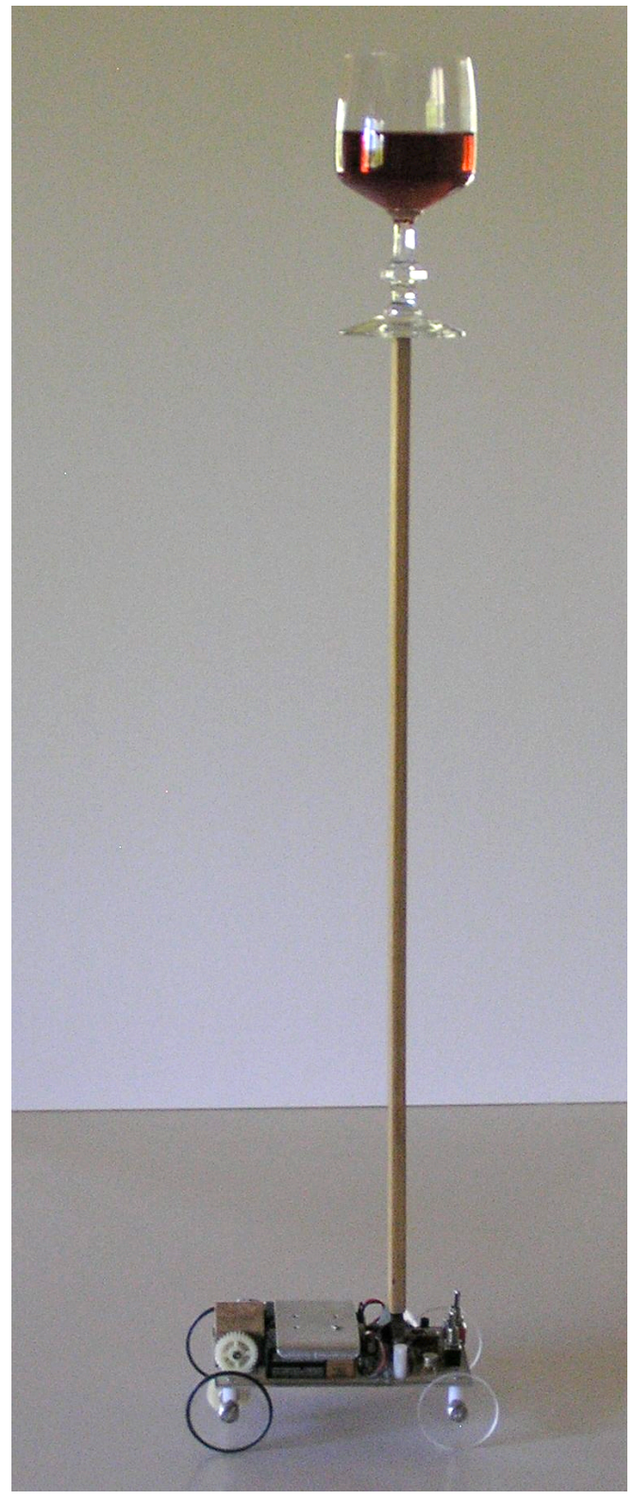
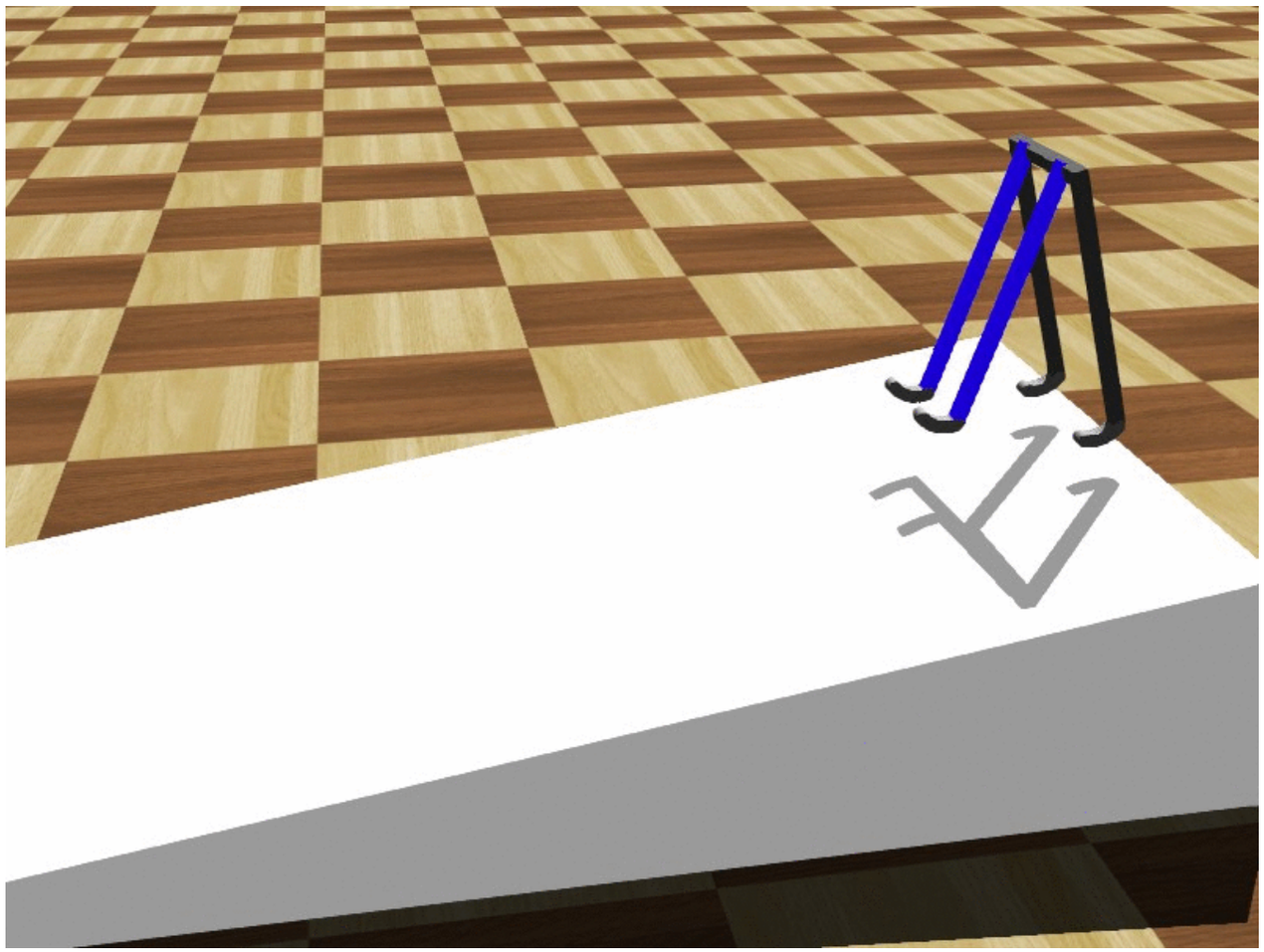


Image source: https://commons.wikimedia.org/wiki/File:Balancer_with_wine_3.JPG

Examples - Robot locomotation

- MDP
- **objective**: make the robot move
- **state**: joint angle and position
- **action**: torques applied to joints
- **reward**: 1 for each time upright + moving



Video source: https://commons.wikimedia.org/wiki/File:Passive_dynamic_walker.gif

Examples - Games

- POMDP
- **objective**: beat the game
- **state**: position, location, state of all objects, agents and world
- **action**: game controls
- **reward**: score increase/ decrease, complete level, die

Video source: SuperTuxKart 1.0 Official Trailer, https://www.youtube.com/watch?v=Lm1TFDBillg

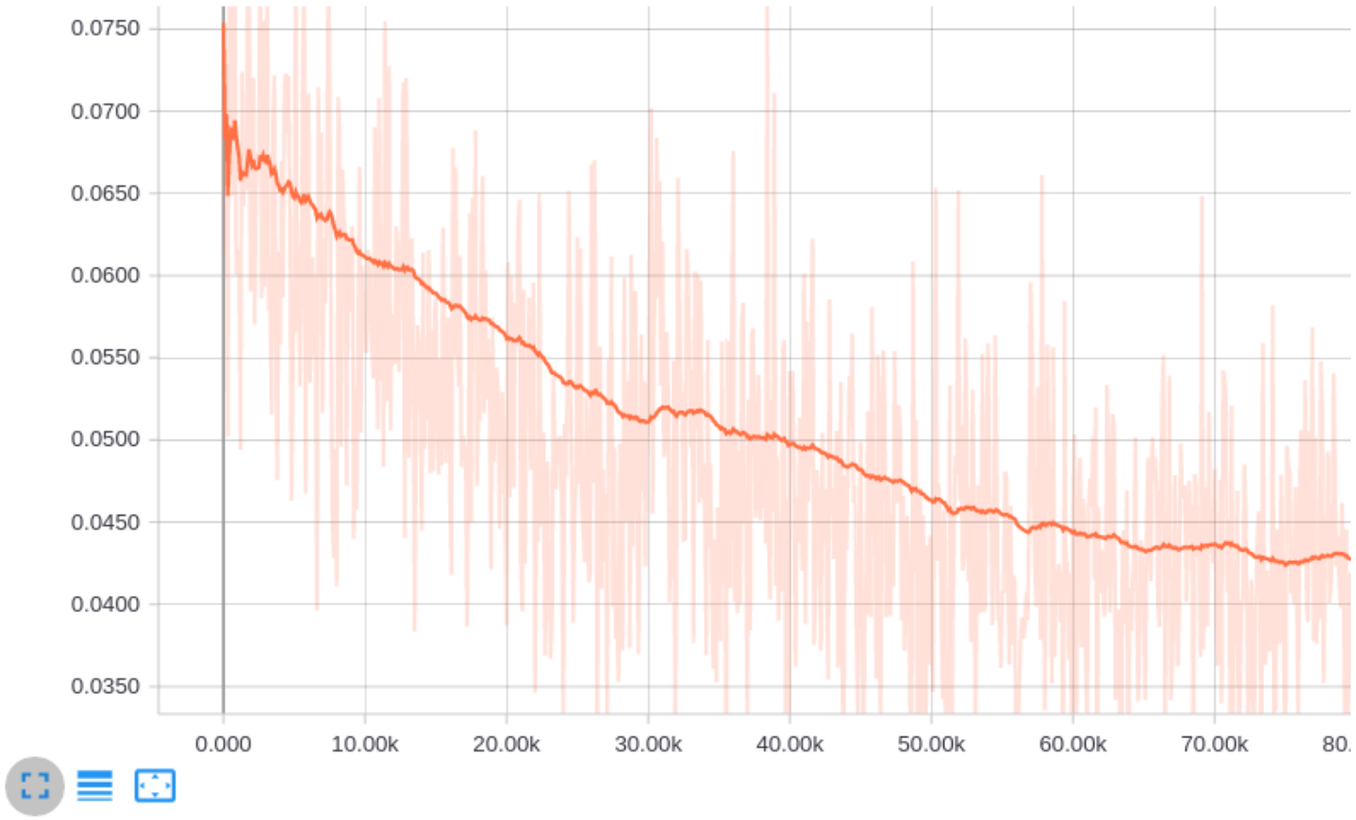
Examples - GO

- MDP
- **objective**: win the game
- **state**: position of pieces
- **action**: next piece
- **reward**: 0 lose, 1 win

Image source: https://en.wikipedia.org/wiki/Go_(game)#/media/File:FloorGoban.JPG

Examples - supervised learning

- MDP
- **objective**: Minimize the training (or validation) loss
- **state**: weights and hyper-parameters
- **action**: gradient update
- **reward**: change in loss



Everything is a (PO)MDP

- Very general concept
  - NP-hard
- Specialized algorithms still work well