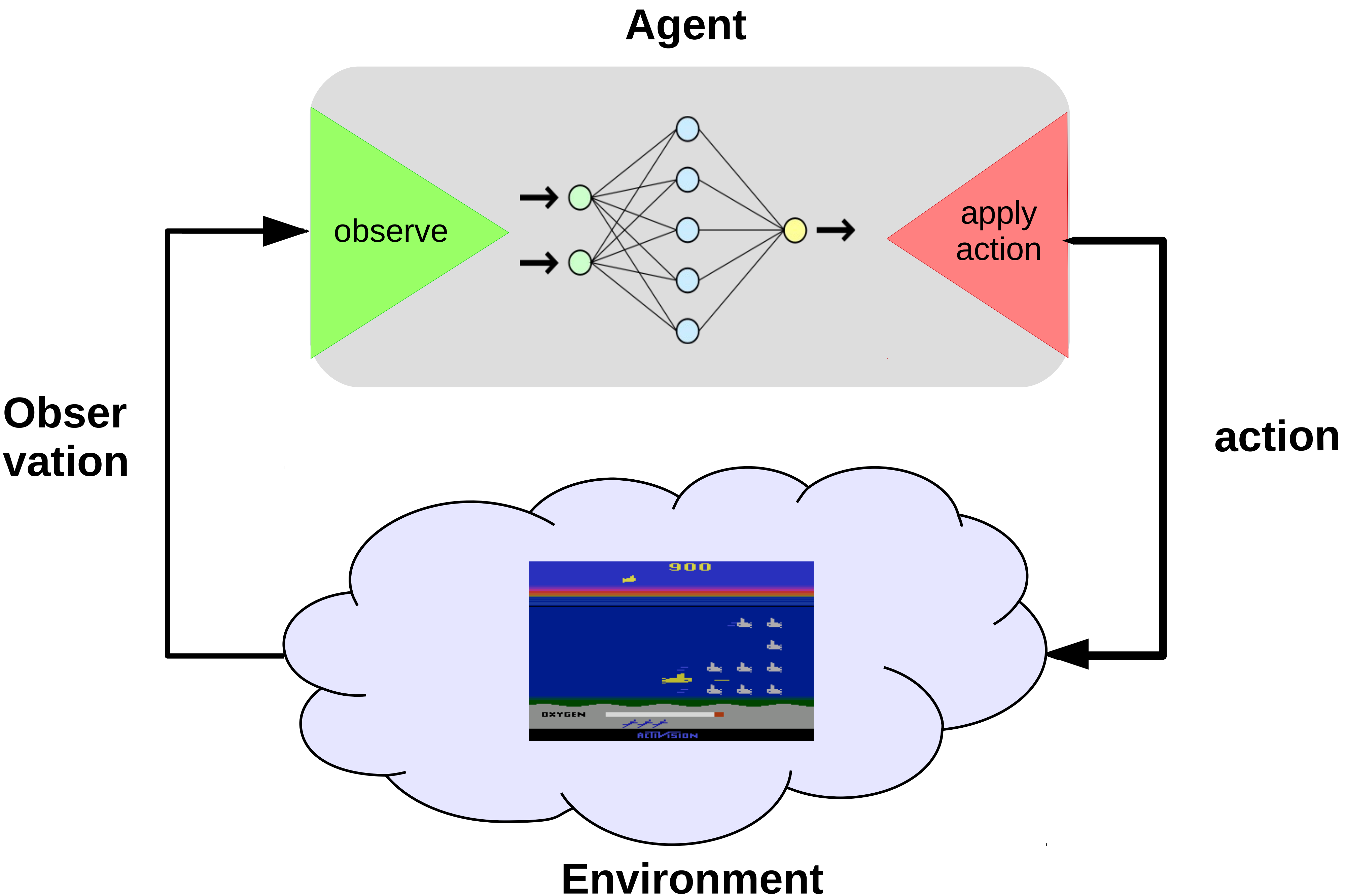
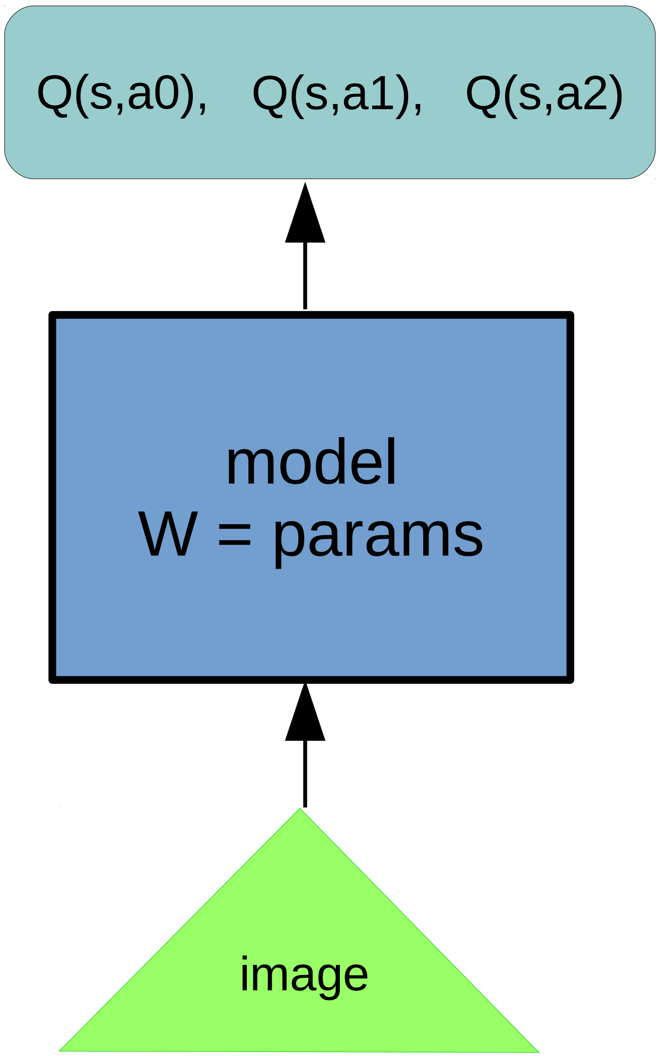
MDP & POMDP


Agent
observe
apply action
Obser vation
action
900
OXYGEN
ACTIVISION
Environment

Approximate Q-learning
nice and simple



Q-values:

$$\hat{Q}(s_t, a_t) = r + \gamma \cdot argmax_{a'} \hat{Q}(s_{t+1}, a')$$

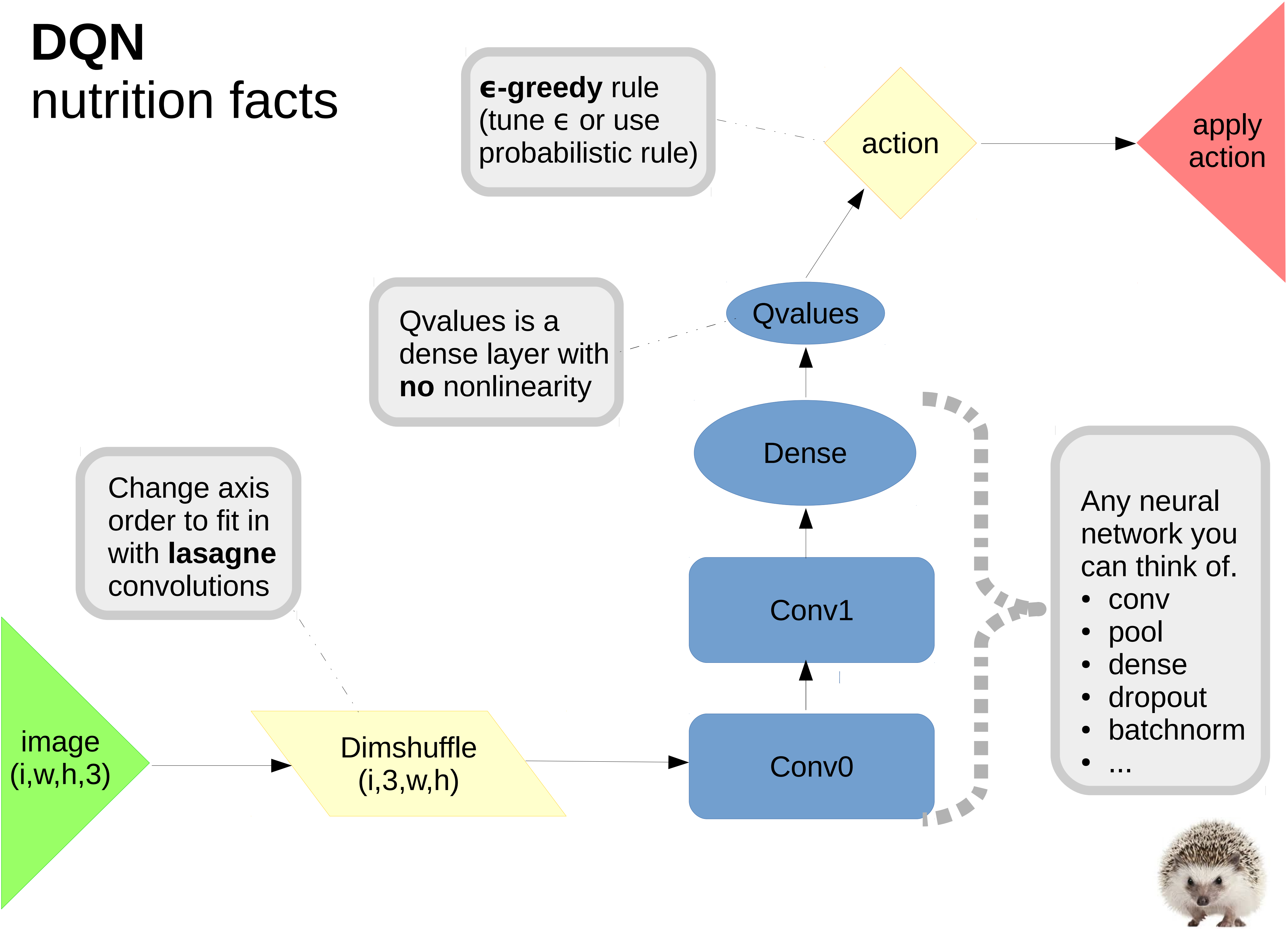
Objective:

$$L = \left(Q(s_t, a_t) - r + \gamma \cdot argmax_{a'} Q(s_{t+1}, a')\right)^2$$

Gradient step:

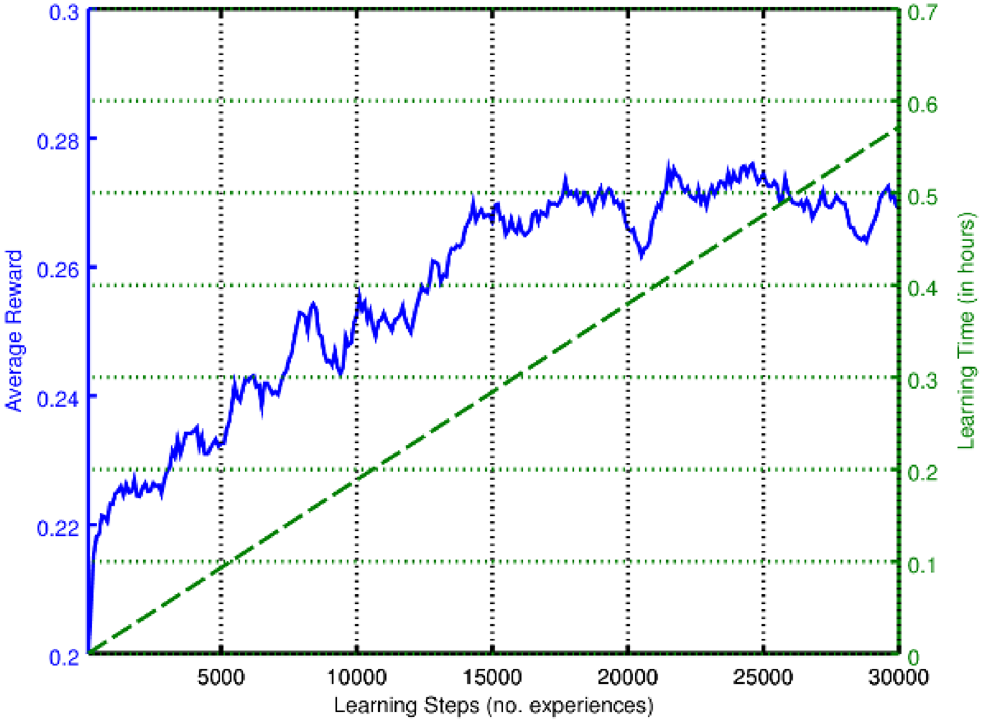
$$w_{t+1} = w_t - \alpha \cdot \frac{\delta L}{\delta w}$$

DQN
nutrition facts



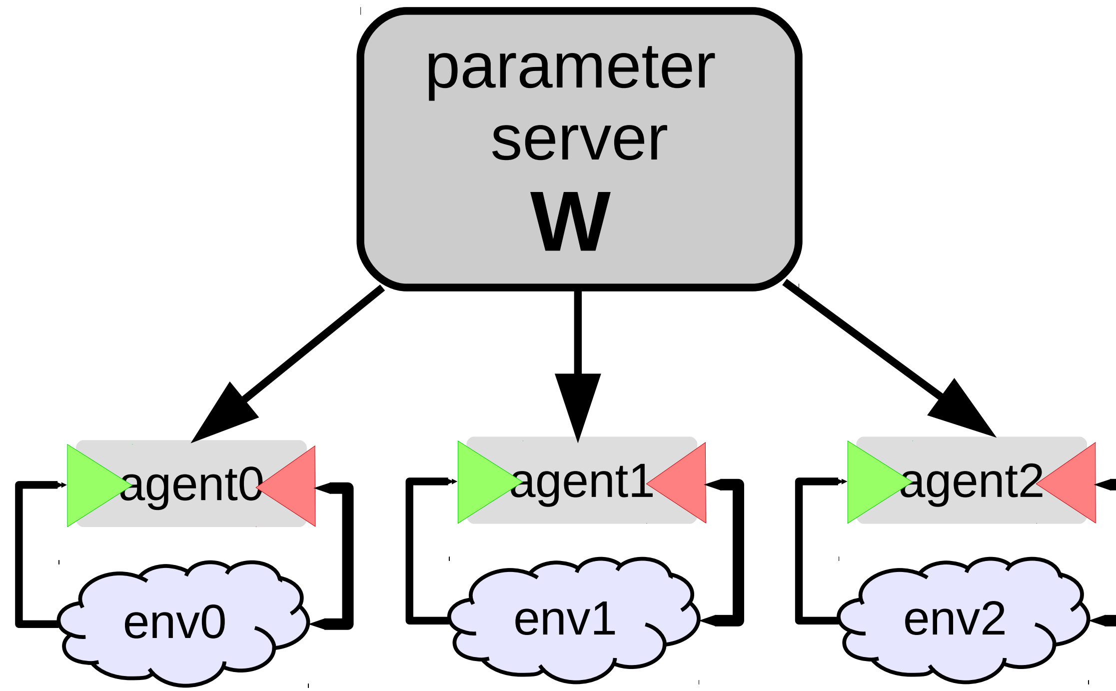
Approximate Q-learning problems

- Training samples are **not “i.i.d”**,
- Model forgets parts of environment it haven't visited for some time,
- Fallbacks on the learning curve
- **Any ideas?**



Deep Q-learning
Multiple agent trick

Idea: Throw in several agents with shared **W**.



- Chances are, they will be exploring different parts of the environment,
- More stable training,
- Requires a lot of interaction,
- Alternative to experience replay.