Deep Reinforcement Learning at Scale

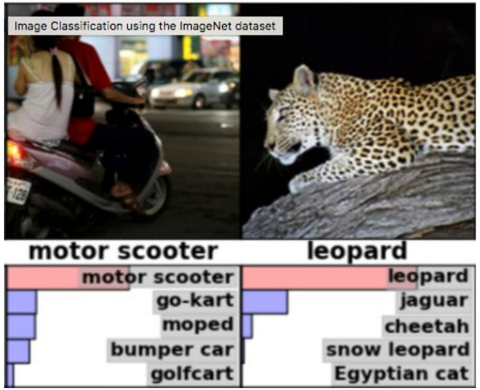
Timothy Lillicrap
Research Scientist, DeepMind & UCL

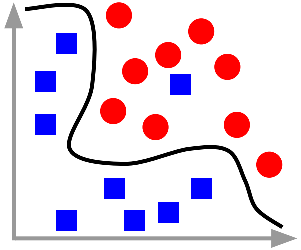
Deep Learning at Supercomputer Scale | NIPS Workshop

DeepMind

What is Reinforcement Learning?

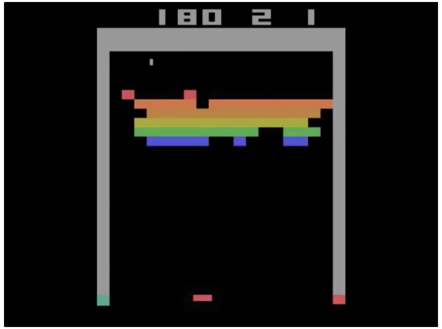
Supervised Learning

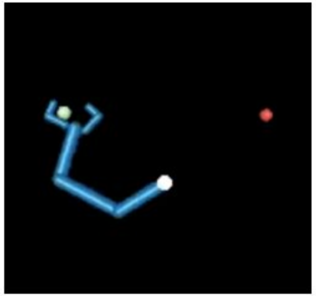




Fixed dataset

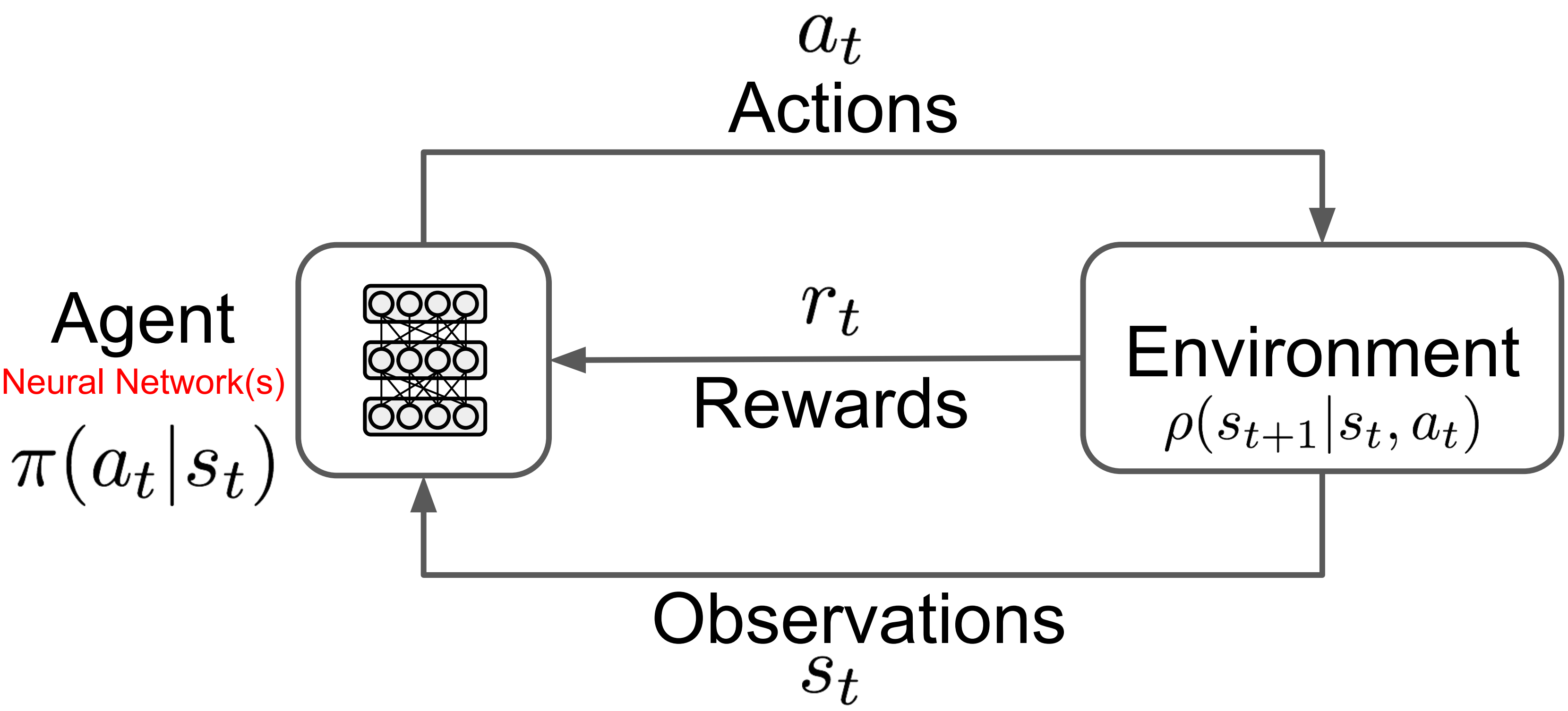
Reinforcement Learning



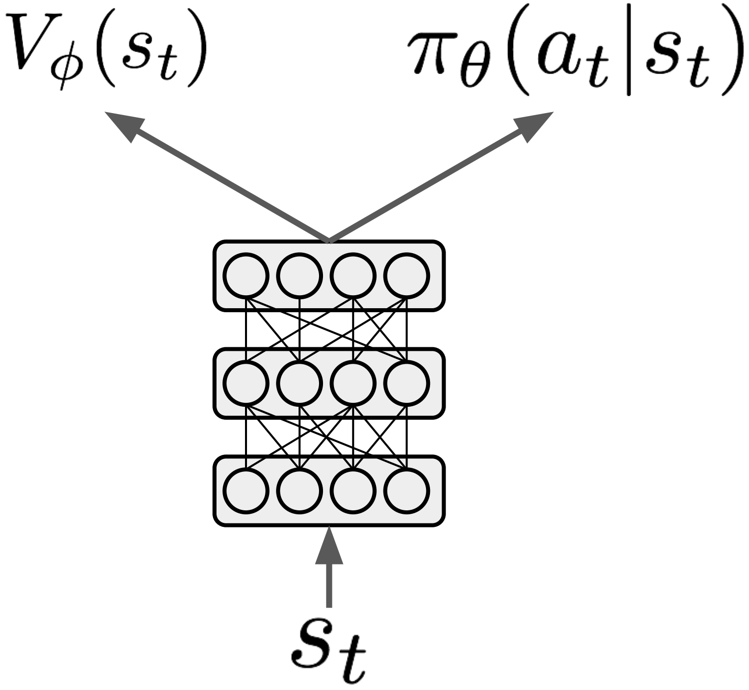


Data depends on actions taken in environment

Formalizing the Agent-Environment Loop



Advantage Actor-Critic (A3C)



$V_{\phi}(s_t)$

$\pi_{\theta}(a_t|s_t)$

s_t

Mnih et al., *ICML* 2016

A Single Trial (with Advantage Actor-Critic)

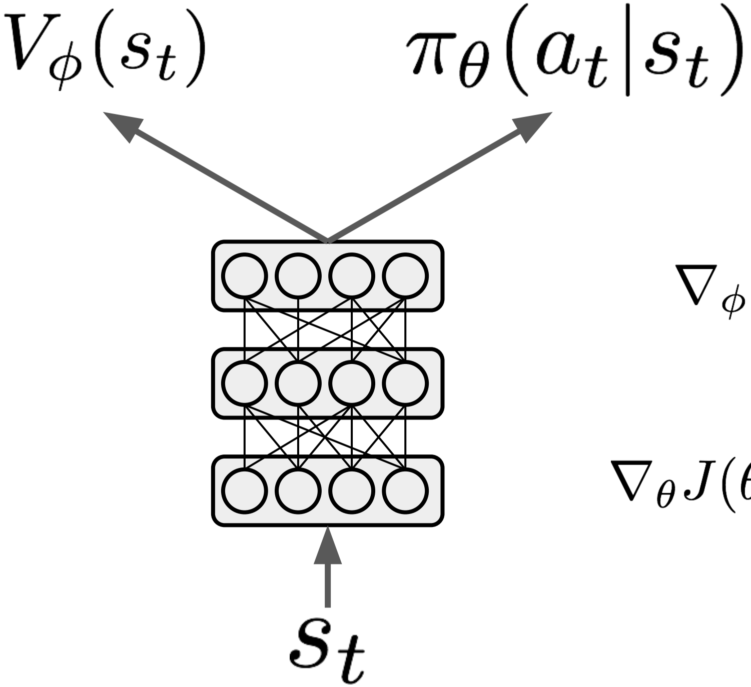
$$\begin{array}{ccccc}
A_0, & A_1, & A_2, & \dots, & A_T \\
V_0, & V_1, & V_2, & \dots, & V_T \\
R_0, & R_1, & R_2, & \dots, & R_T \\
r_0, & r_1, & r_2, & \dots, & r_T \\
a_0, & a_1, & a_2, & \dots, & a_T \\
\pi(a_0|s_0), & \pi(a_1|s_1), & \pi(a_2|s_2), & \dots, & \pi(a_T|s_T) \\
s_0, & s_1, & s_2, & \dots, & s_T \\
 & \rho(s_1|s_0,a_0), & \rho(s_2|s_1,a_1), & \dots, & \rho(s_T|s_{T-1},a_{T-1})
\end{array}$$

Time →

$$R_t = \sum_{k=t}^{T} \gamma^{t-k} r_t \qquad V_t = V_\phi(s_t) \qquad A_t = R_t - V_\phi(s_t)$$

Mnih et al., *ICML* 2016

Combating Variance: Advantage Actor-Critic

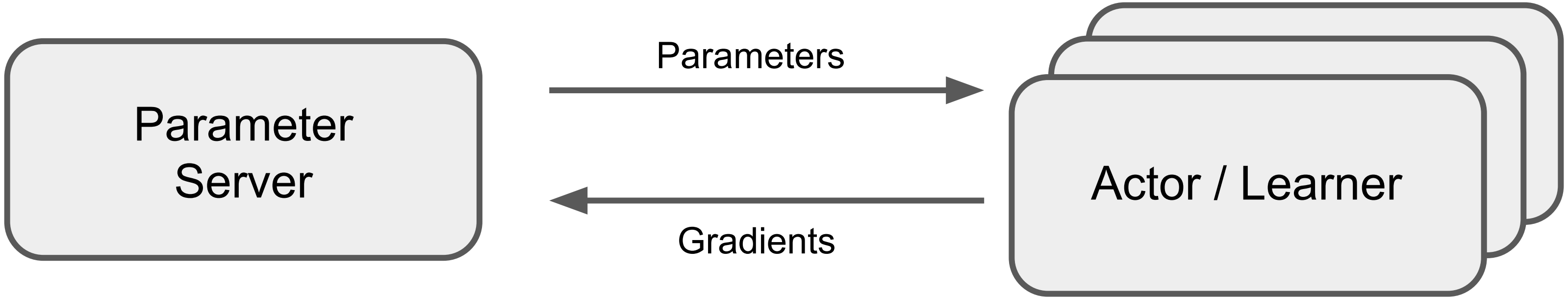


$$\nabla_\phi \mathcal{L} = \sum_{t=0}^{T} \nabla_\phi (\underline{R_t - V_\phi(s_t)})^2$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(\underline{R_t - V_\phi(s_t)})\right]$$

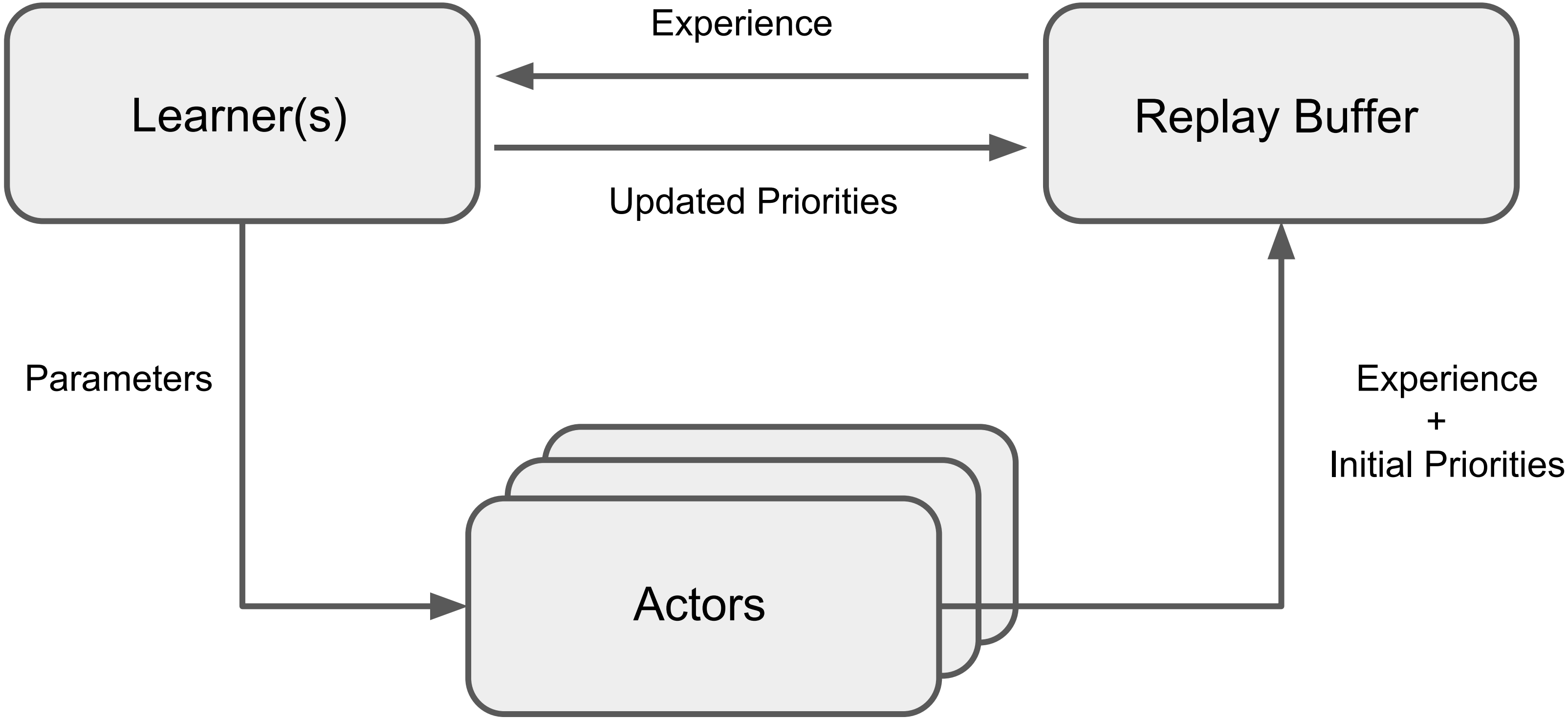
Mnih et al., *ICML* 2016

Scaling Reinforcement Learning (A3C)



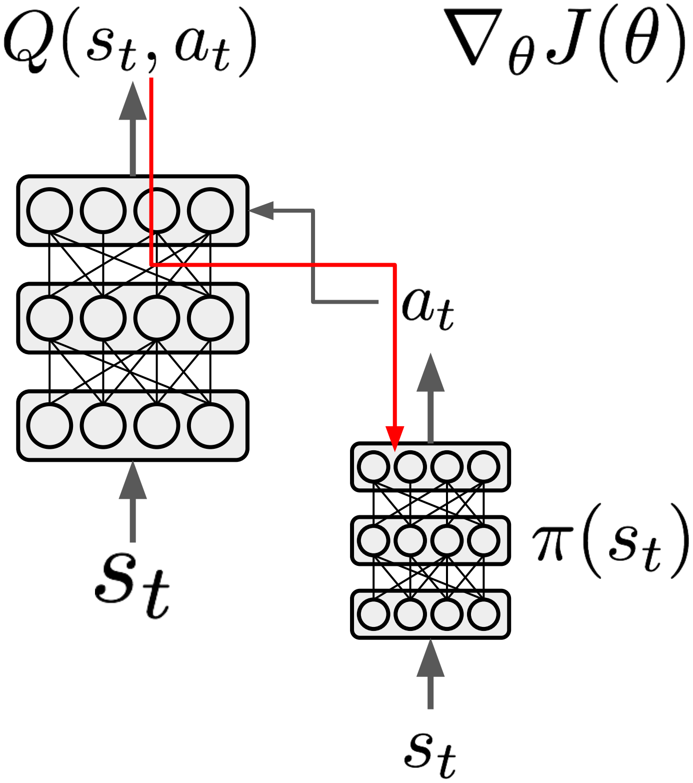
Mnih et al., *ICML* 2016

Scaling Reinforcement Learning



Horgan et al., 2017 & Schaul et al. 2015

Off-policy Actor-Critic for Continuous Actions



$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\mathcal{D}}\left[\nabla_a Q(s,a;\phi)\big|_{a=\pi(s)} \nabla_\theta \pi(s)\right]$$

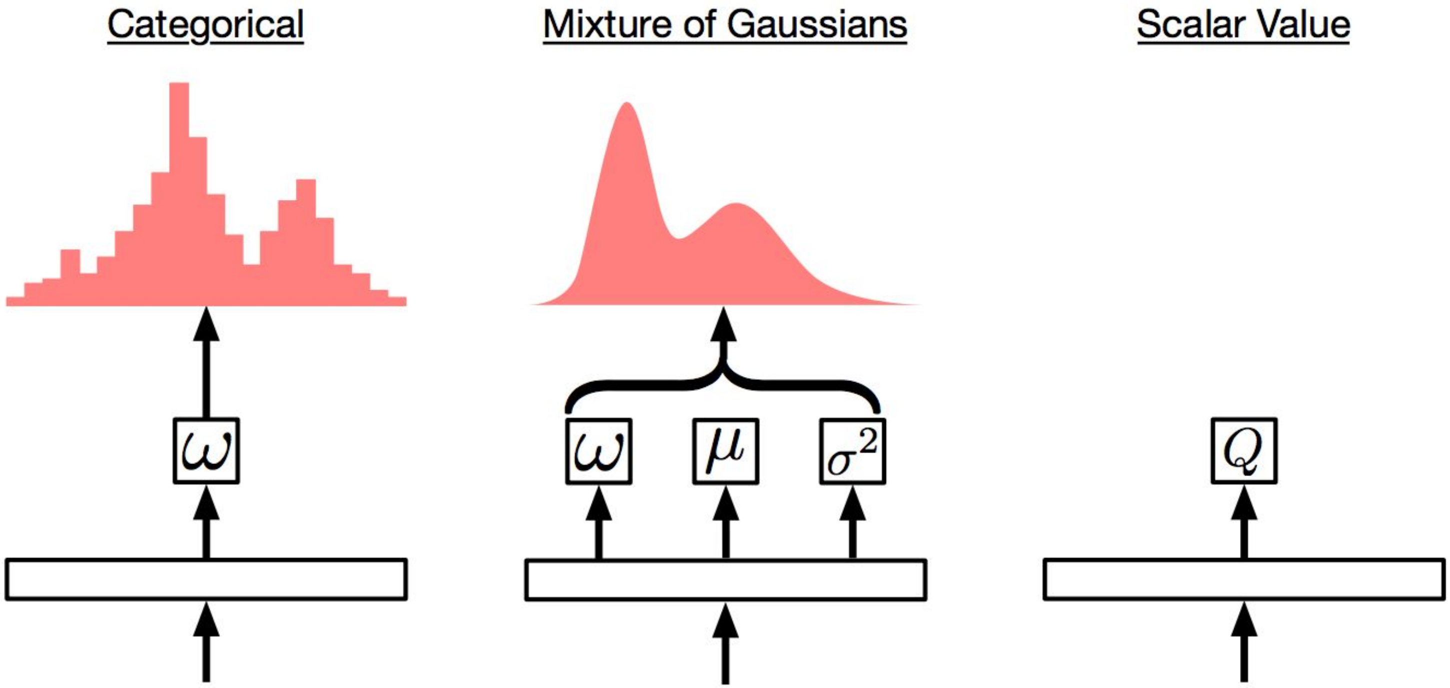
$$(s, a, r, s') \sim U(\mathcal{D})$$

$$y = r + \gamma Q(s', \pi(s'); \phi^{\text{target}})$$

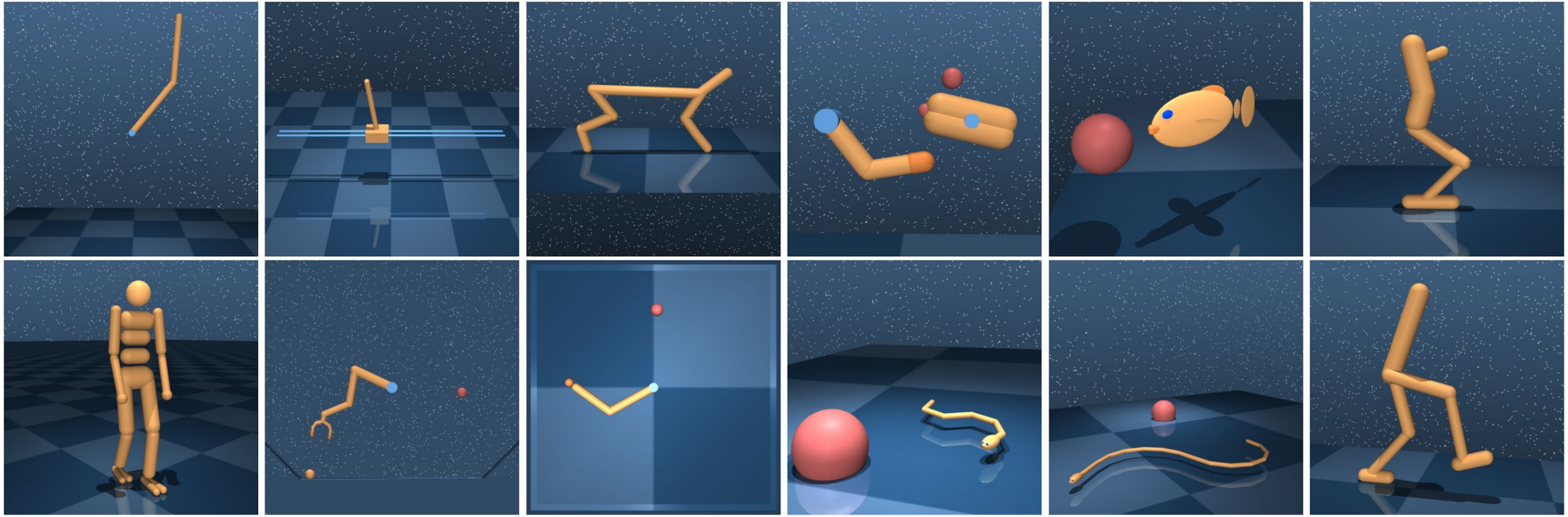
$$L(\phi) = (y - Q(s,a;\phi))^2$$

Lillicrap et al., *ICLR* 2016

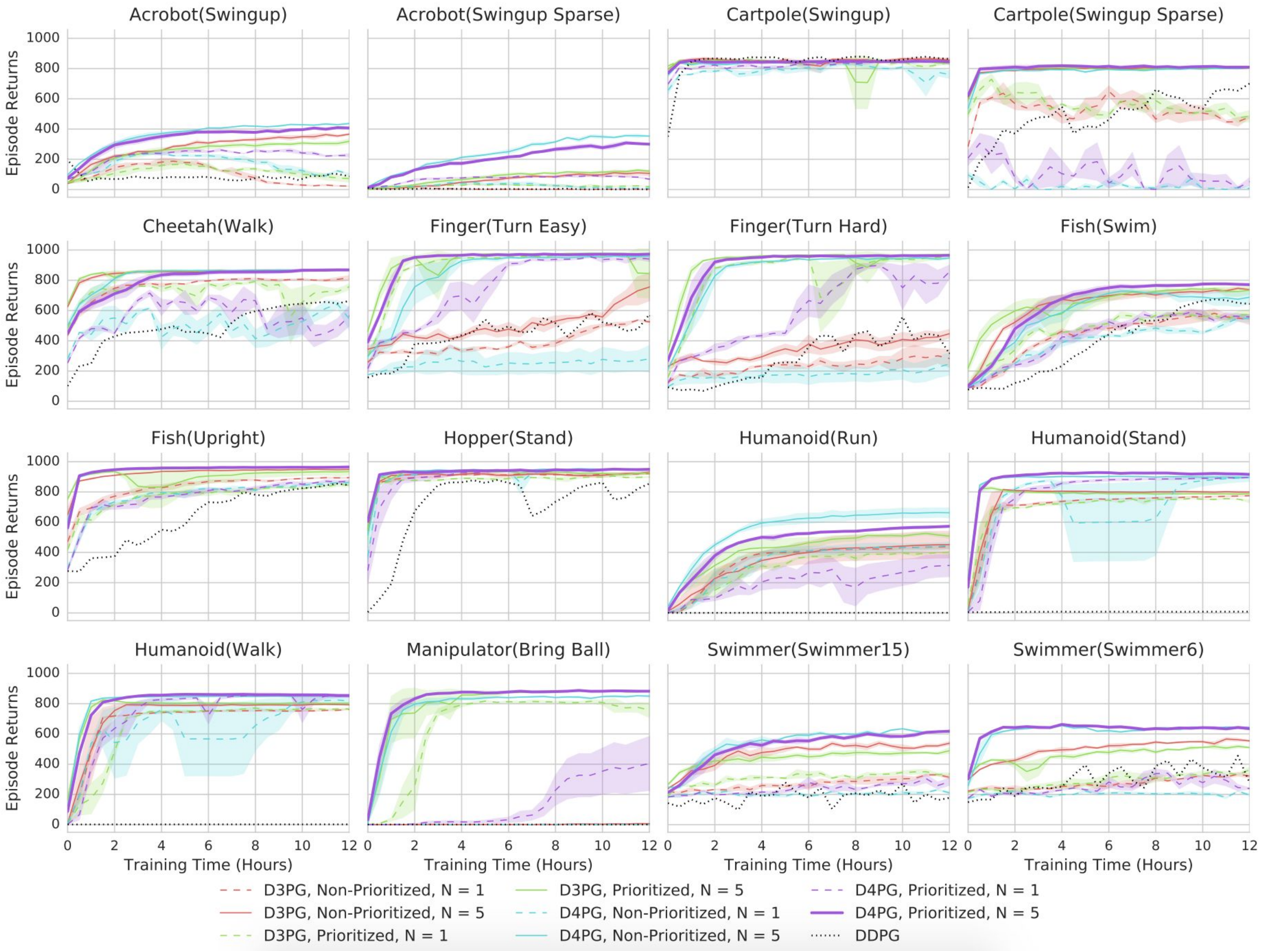
Distributional Distributed DDPG (D4PG)



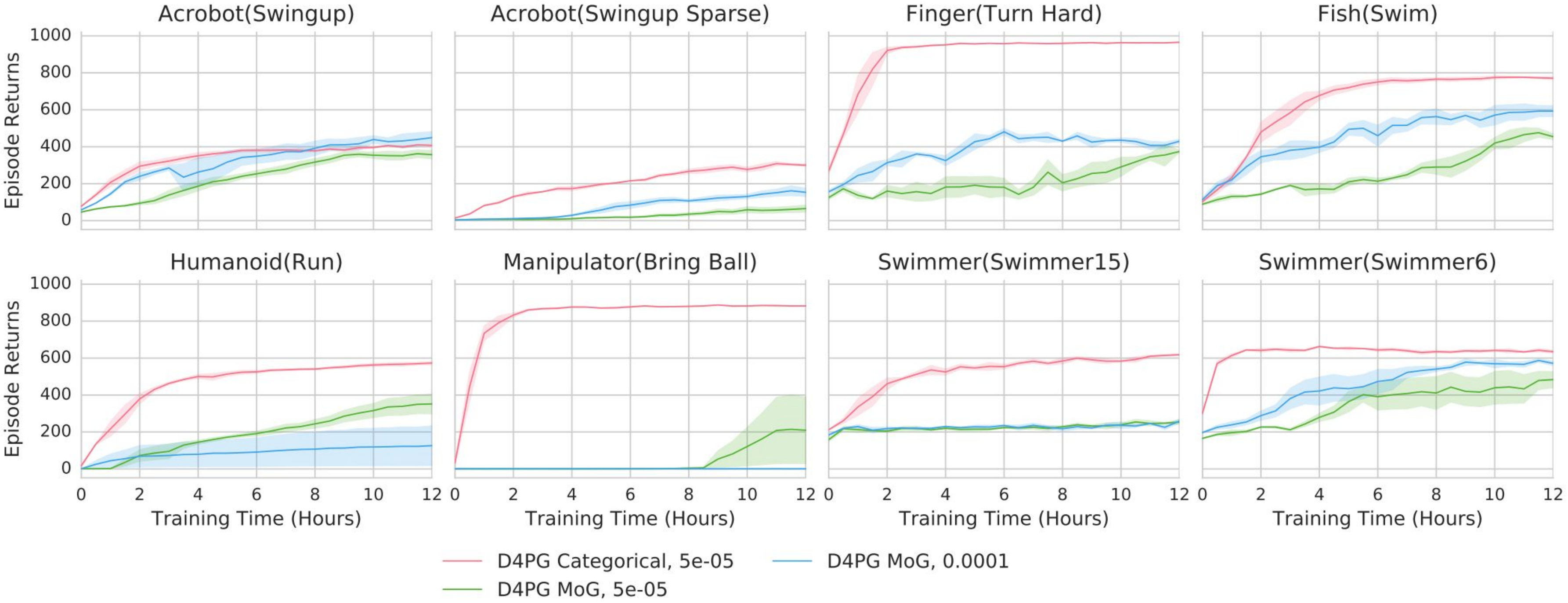
Hoffman, Barth-Maron et al., 2017



Hoffman, Barth-Maron et al., 2017

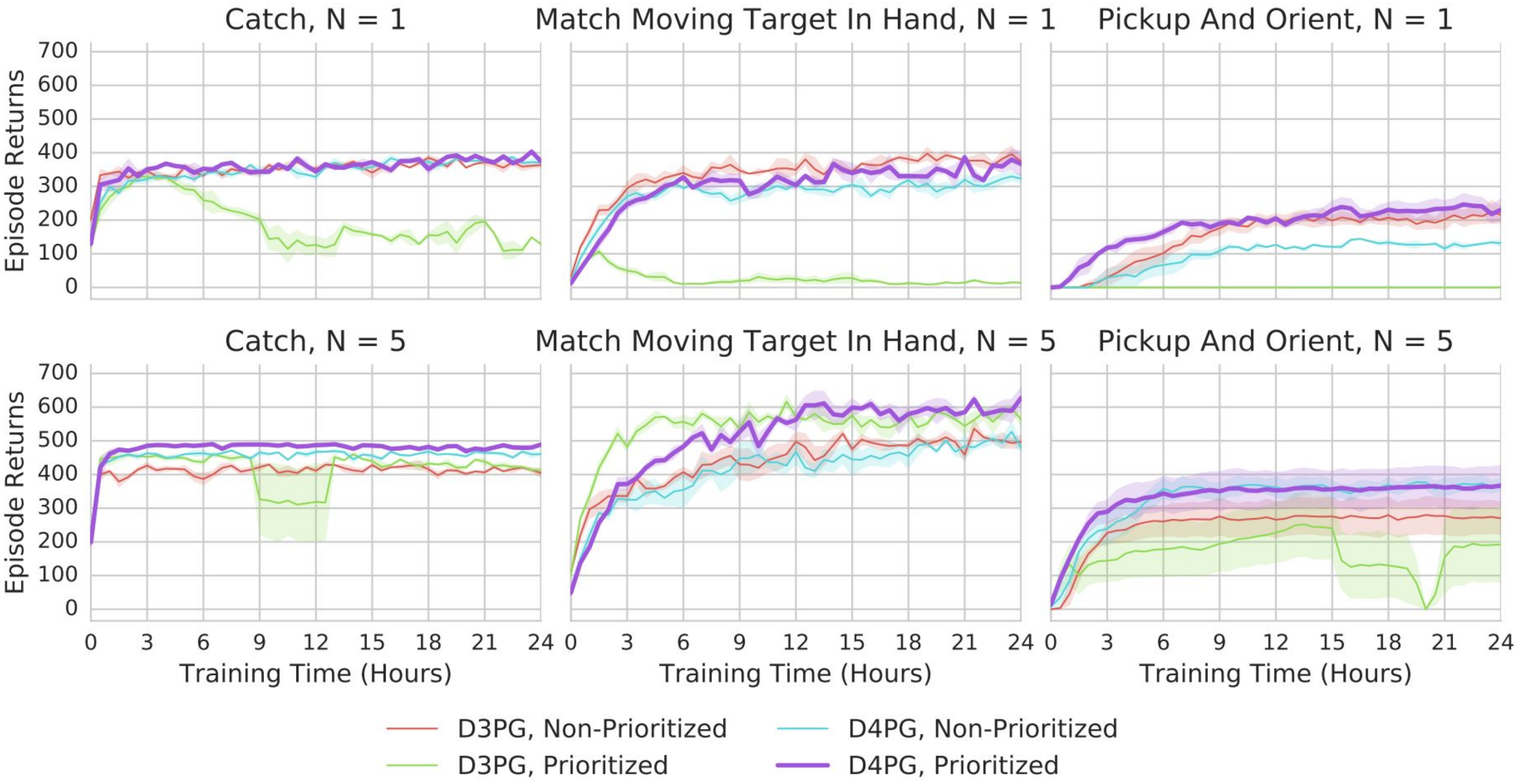


Hoffman, Barth-Maron et al., 2017

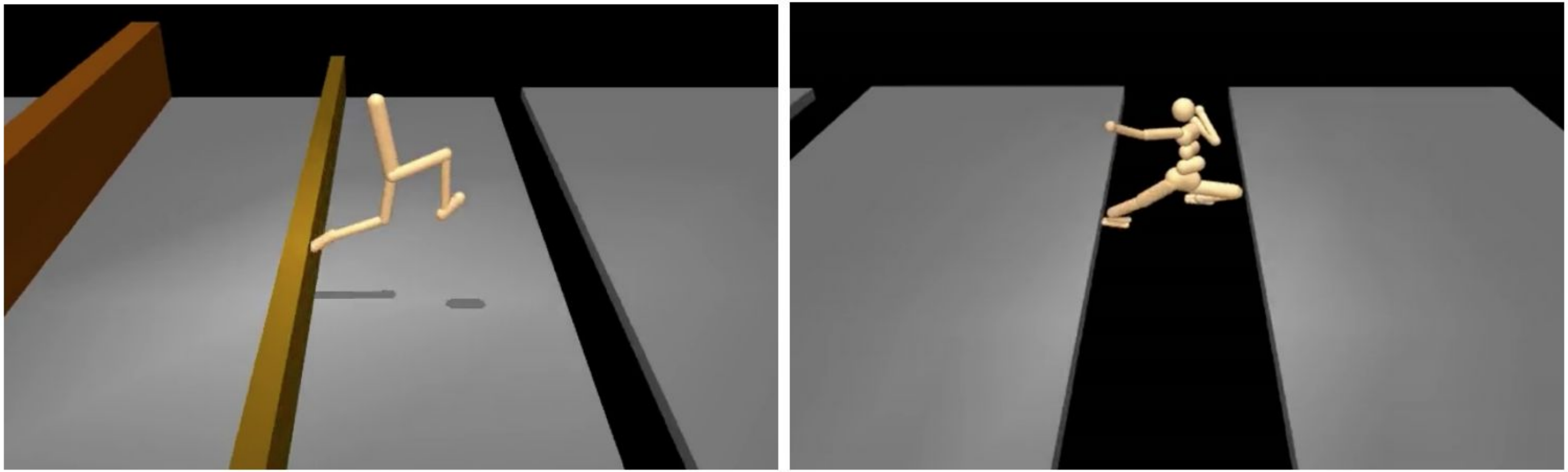


Hoffman, Barth-Maron et al., 2017

Hoffman, Barth-Maron et al., 2017



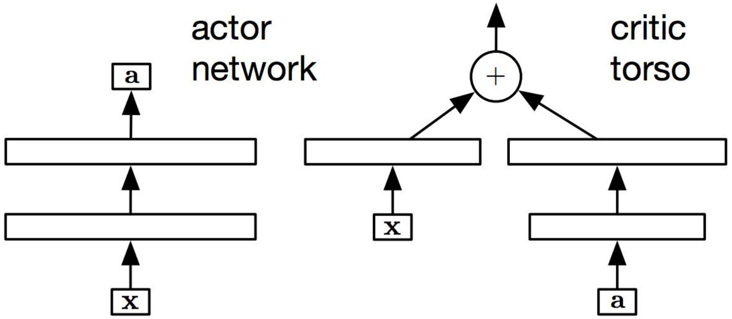
Hoffman, Barth-Maron et al., 2017



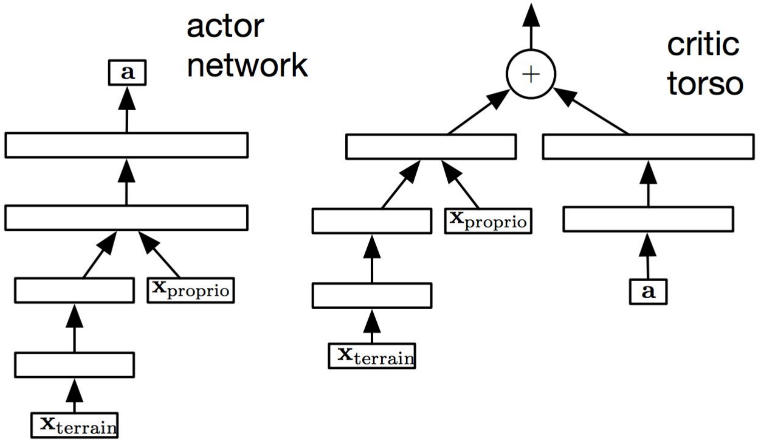
Hoffman, Barth-Maron et al., 2017

Distributional Distributed DDPG (D4PG)

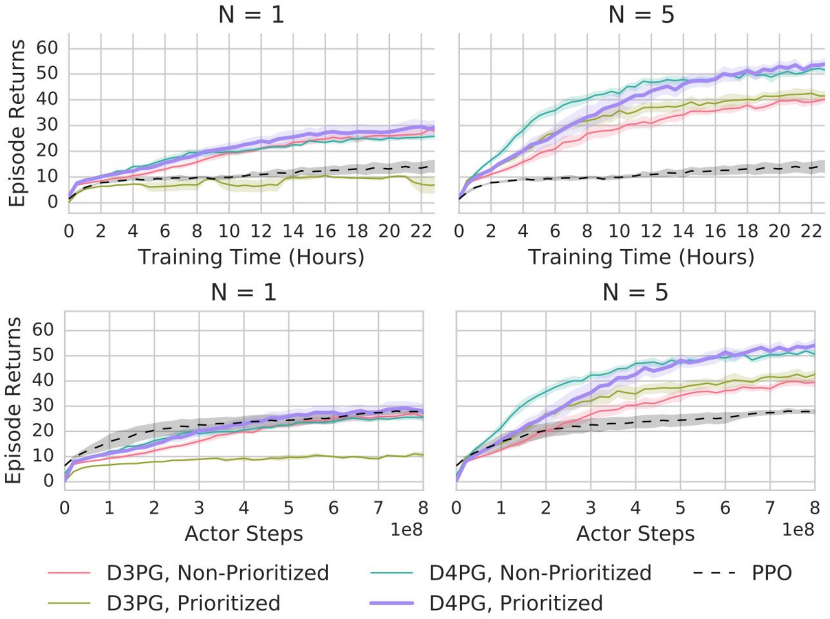
Standard Networks



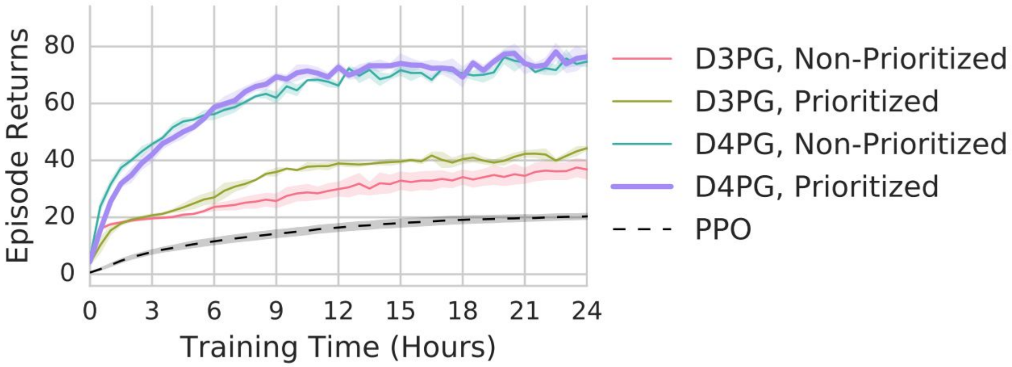
Parkour Networks


actor
network
a
$\mathbf{x}_{\text{proprio}}$
$\mathbf{x}_{\text{terrain}}$
+
critic
torso
$\mathbf{x}_{\text{proprio}}$
$\mathbf{x}_{\text{terrain}}$
a

Hoffman, Barth-Maron et al., 2017



Hoffman, Barth-Maron et al., 2017

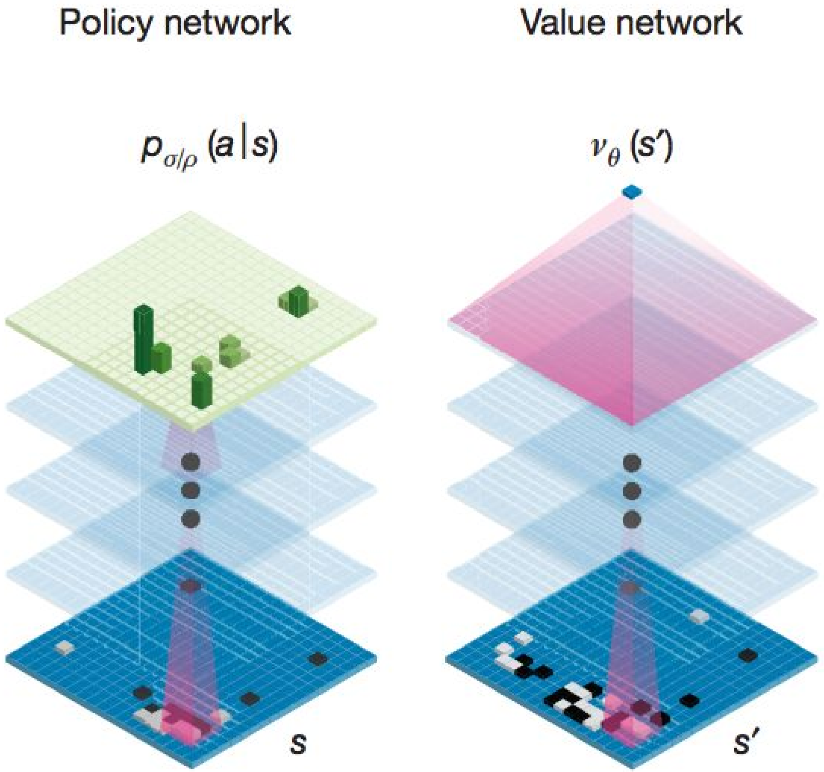


Hoffman, Barth-Maron et al., 2017

Playing Go with Deep Networks and Planning

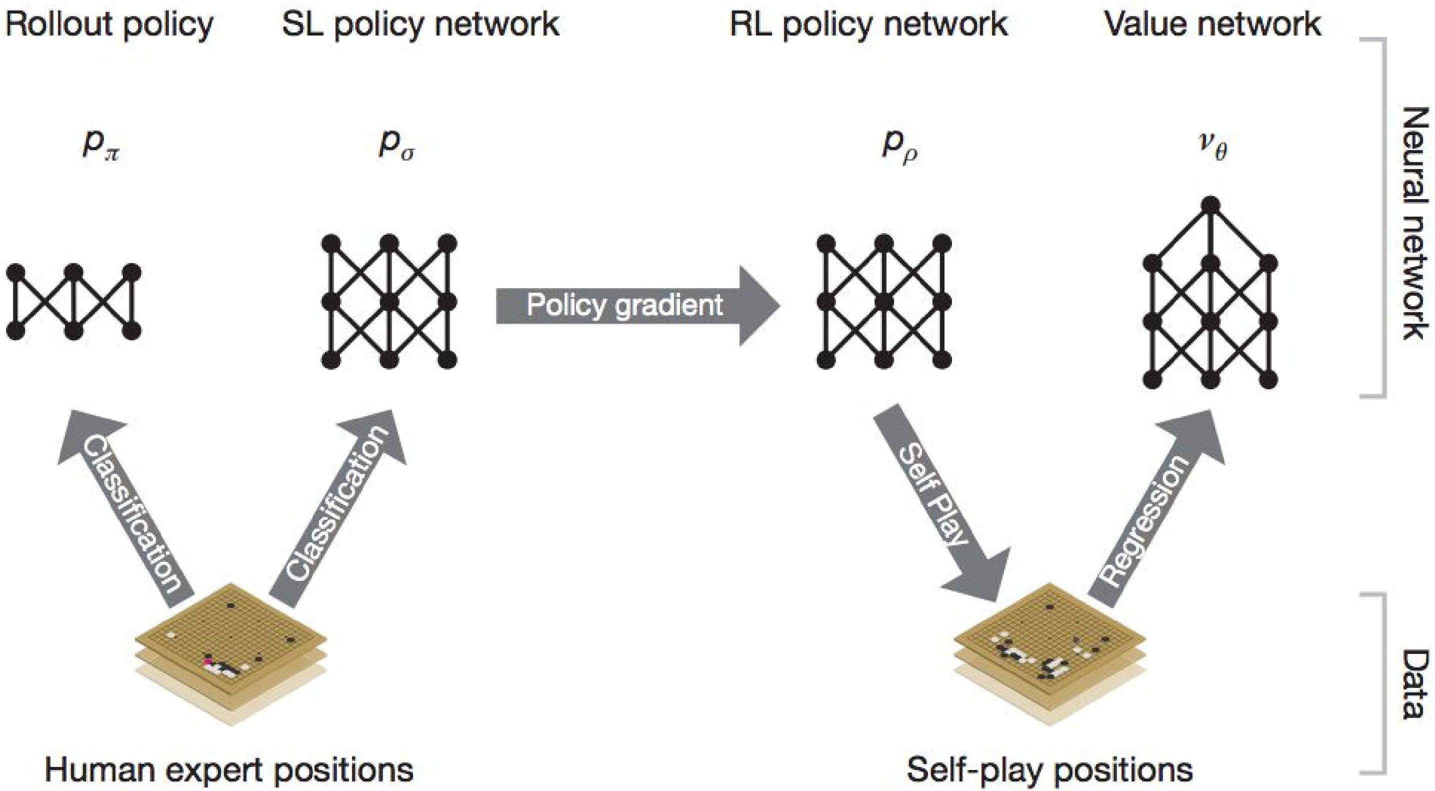
$$\rho\left(s_{t+1} \middle| s_t, a_t\right)$$

Use environment model
in order to plan!



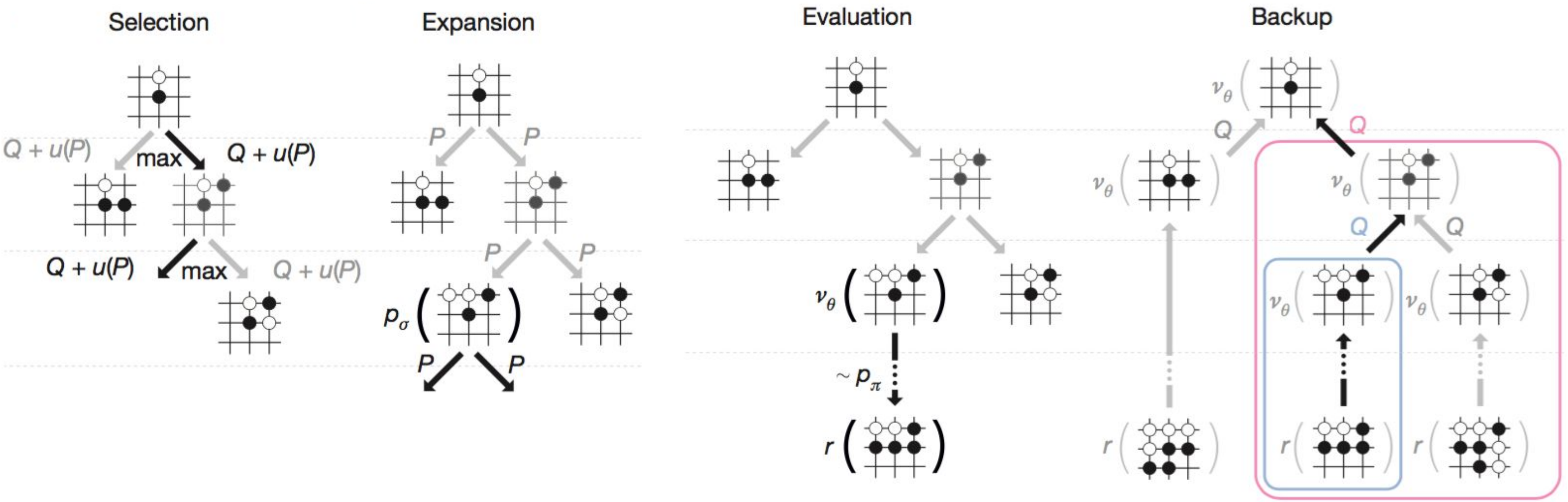
Silver, Huang et al., *Nature*, 2016

Training Policy and Value Networks



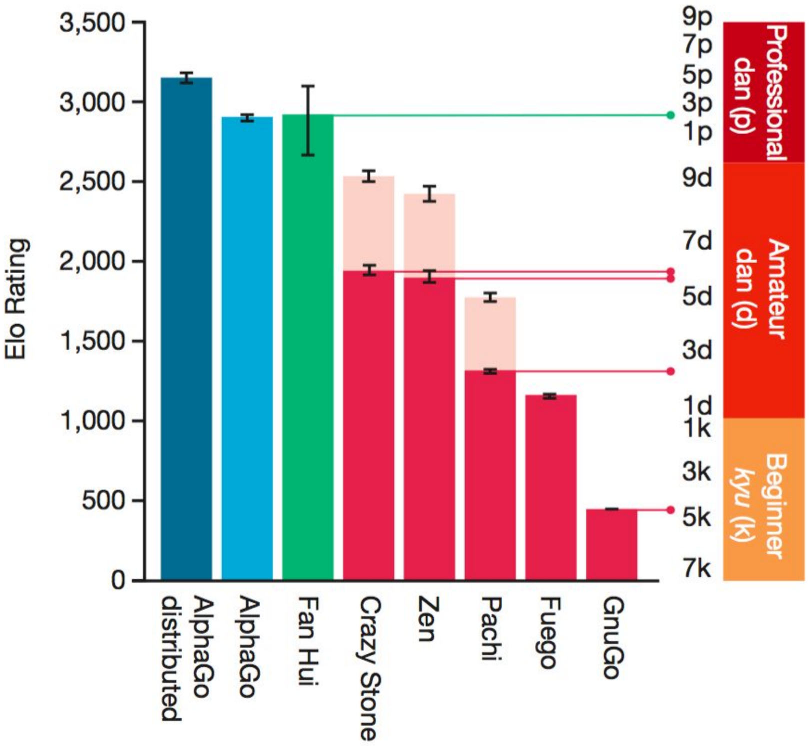
Silver, Huang et al., *Nature*, 2016

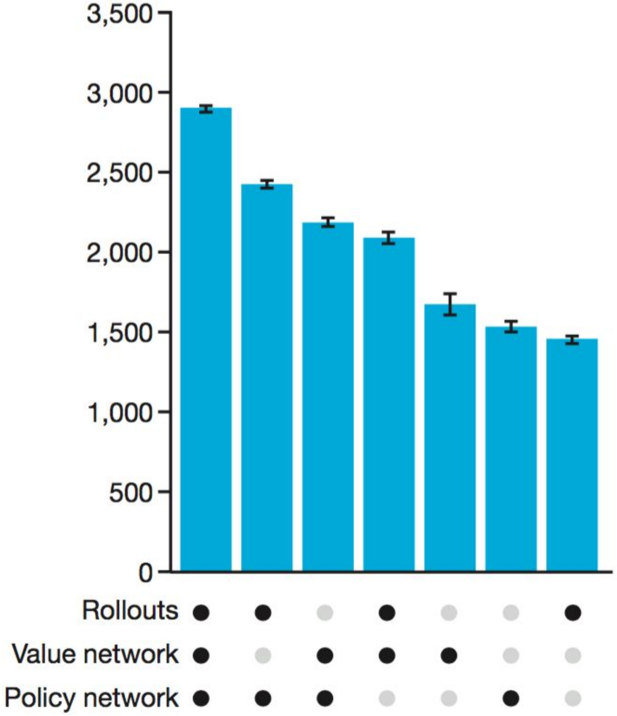
Planning with an Environment Model & MCTS



Silver, Huang et al., *Nature*, 2016

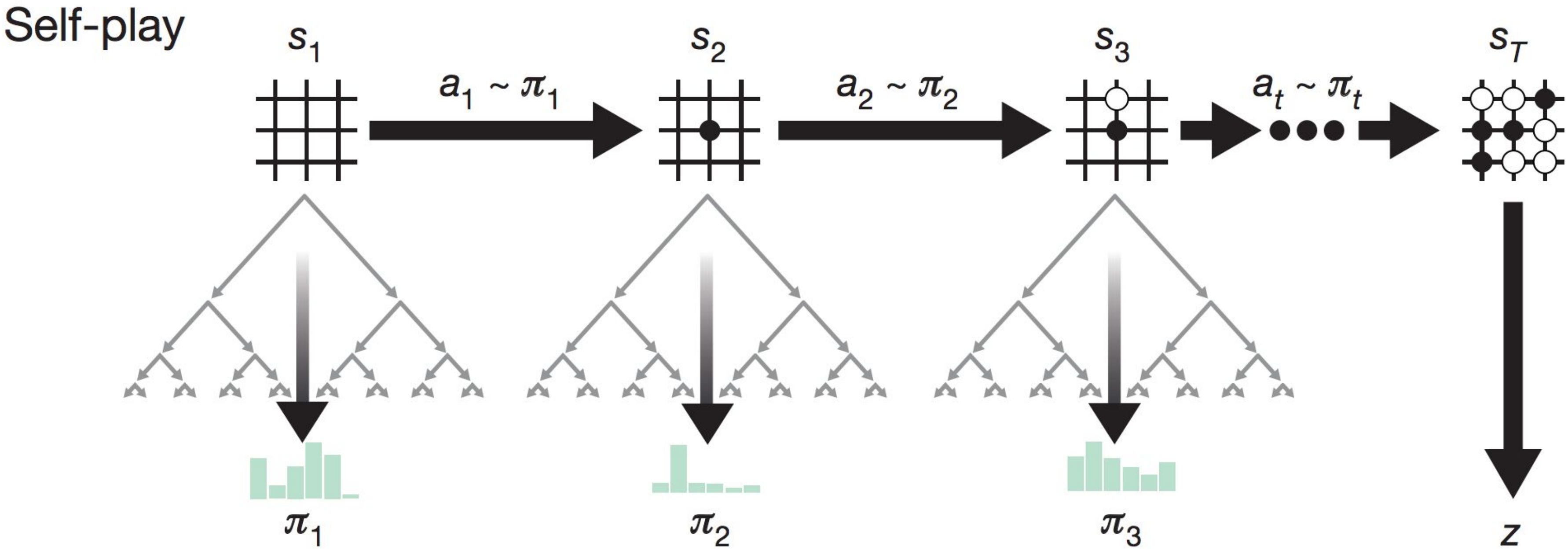
Planning with an Environment Model





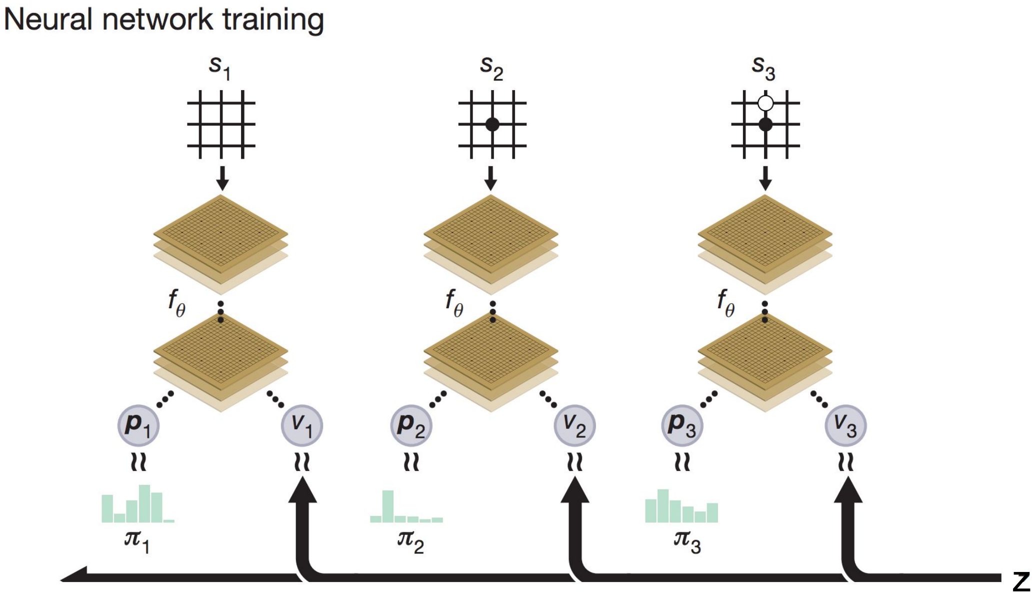
Silver, Huang et al., *Nature*, 2016

Playing Go with Without Human Knowledge



Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

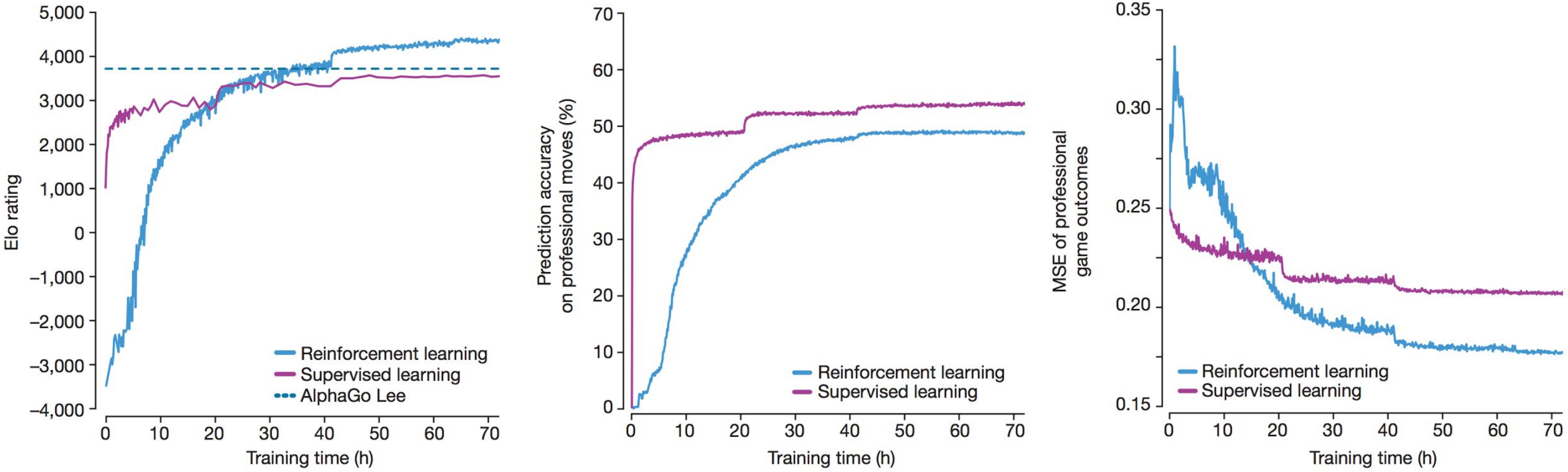
Playing Go with Without Human Knowledge



$$(\boldsymbol{p}, v) = f_\theta(s) \quad \text{and} \quad l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

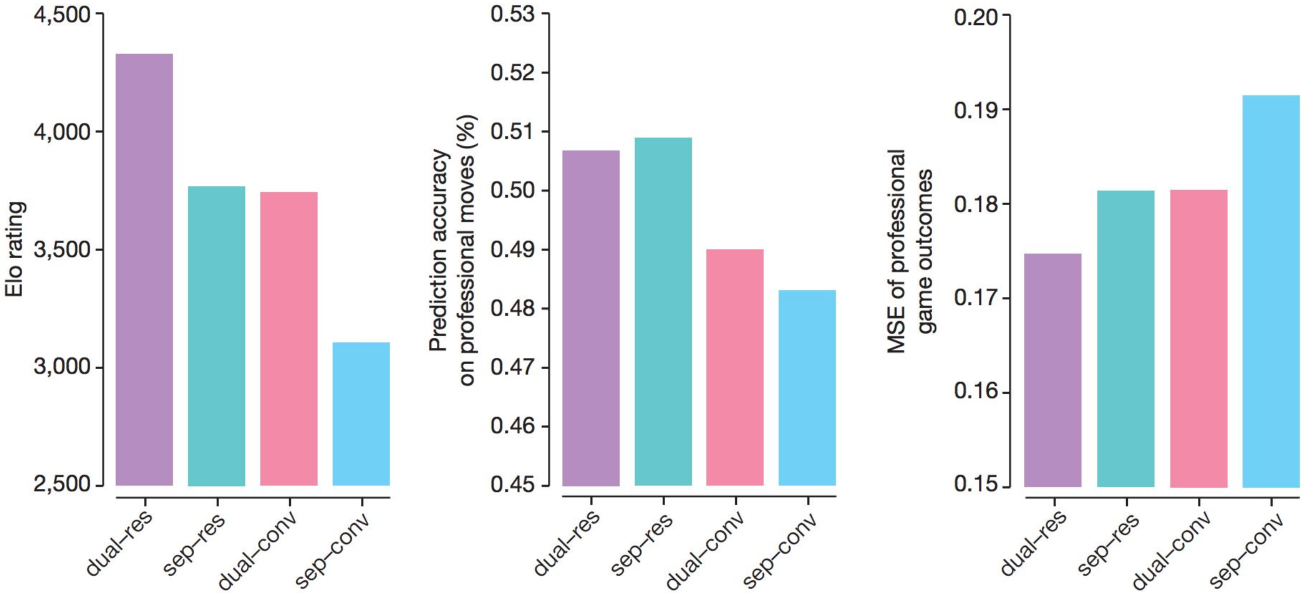
Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

Playing Go with Without Human Knowledge



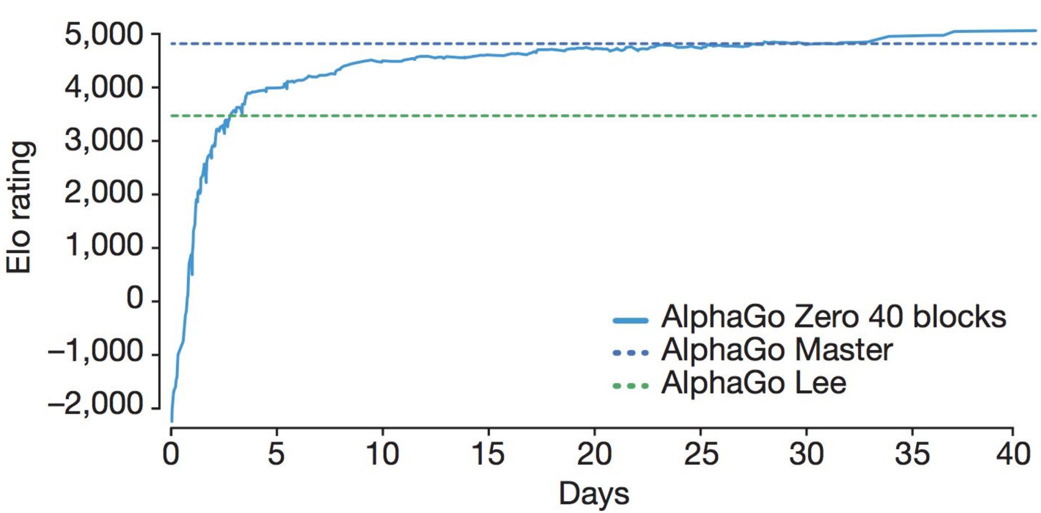
Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

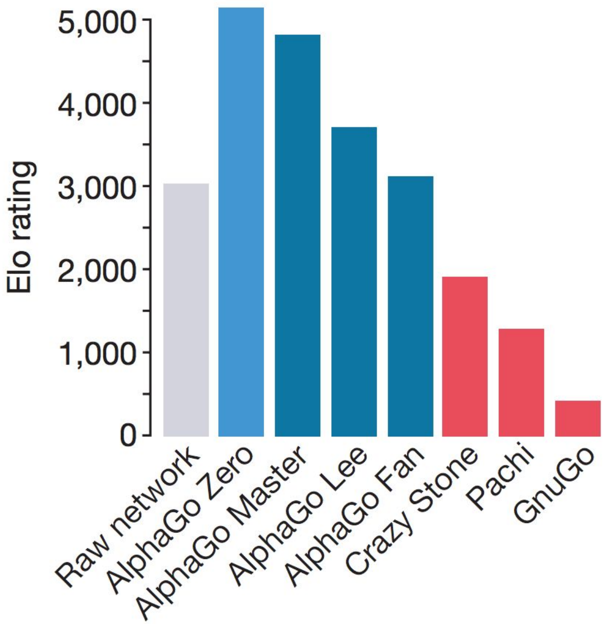
Playing Go with Without Human Knowledge



Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

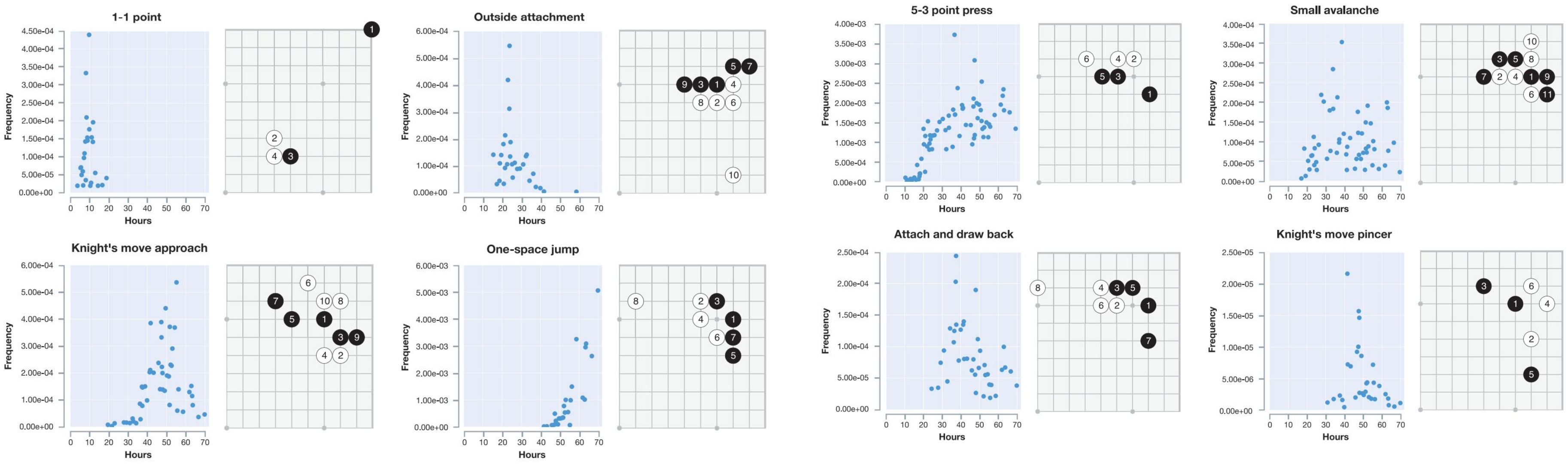
Playing Go with Without Human Knowledge





Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

Playing Go with Without Human Knowledge



1-1 point

Outside attachment

5-3 point press

Small avalanche

Knight's move approach

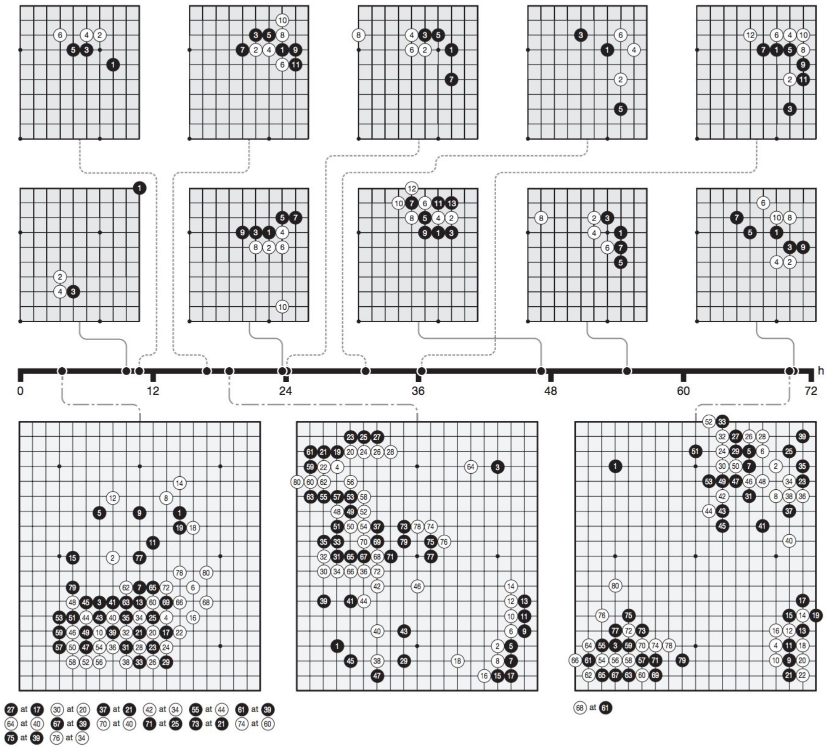
One-space jump

Attach and draw back

Knight's move pincer

Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

Playing Go with Without Human Knowledge



Silver, Schrittwieser, Simonyan, et al. *Nature*, 2017

Questions?