Intro to Reinforcement Learning Data Science Utrecht

Rob Romijnders

Eindhoven, University of Technology

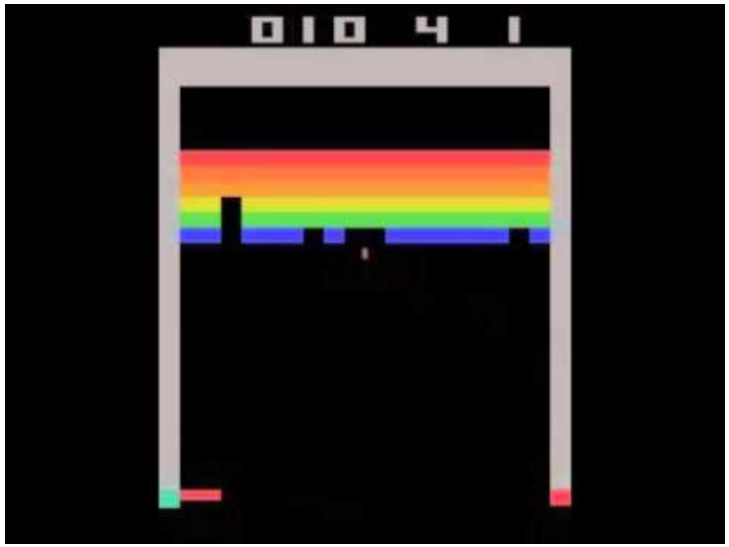
RomijndersRob@gmail.com

March 23, 2017

Overview

1. Why
   - Examples
   - Why not
2. State and policy
3. Policy gradient
   - Gradient
   - Variance reduction
4. Case: Alpha Go
5. Questions

Popular examples



(a) Atari

(b) Alpha Go

Figure: Examples of RL

Robots

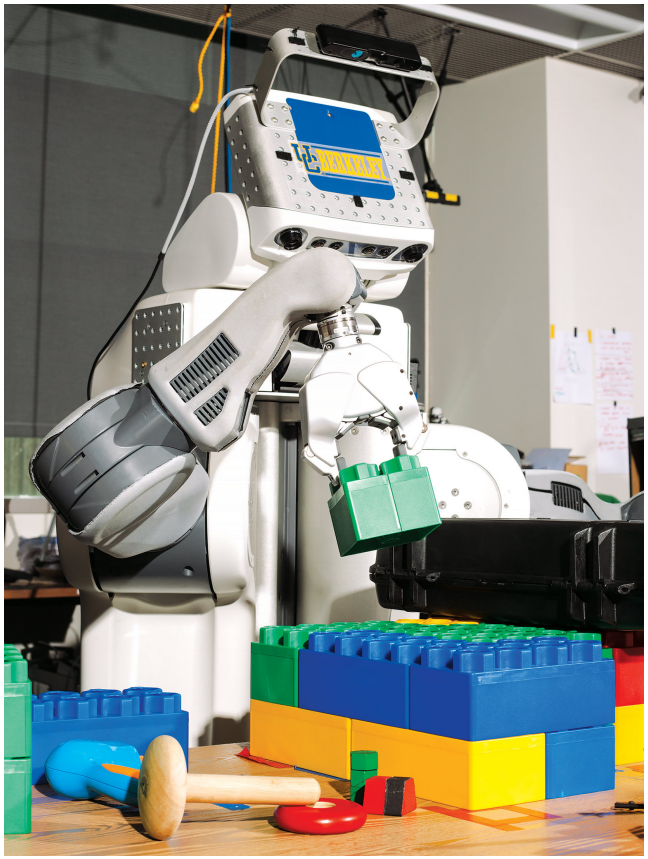
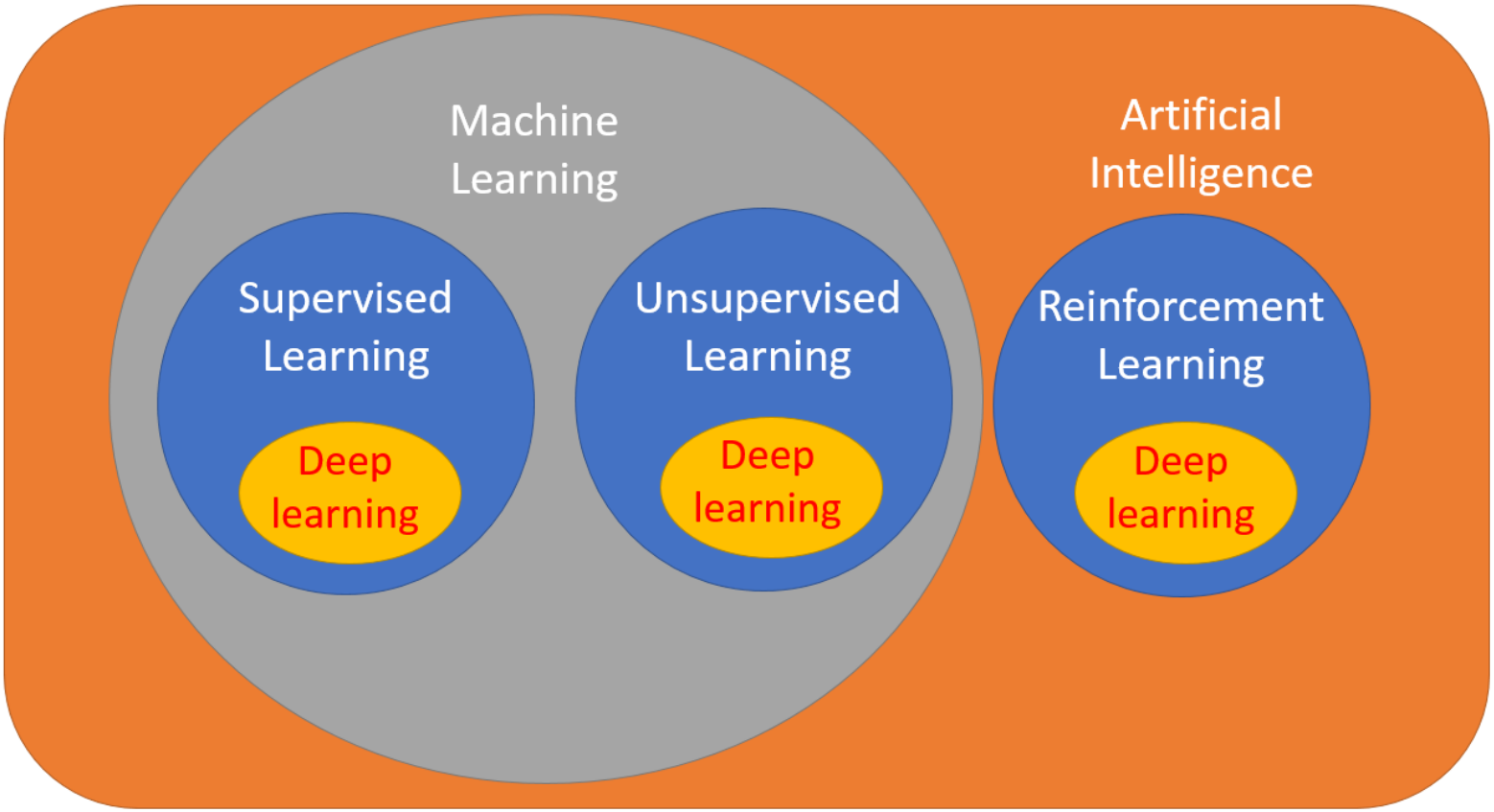


Figure: credits: bloomberg.com/features/2015-preschool-for-robots/

Poker

Figure: credits: cardschat.com Jan, 2017

Where are we?


Artificial Intelligence
Machine Learning
Supervised Learning
Deep learning
Unsupervised Learning
Deep learning
Reinforcement Learning
Deep learning

Why not RL?

1. Know when you are doing supervised learning. And don't force it into RL
2. Consider imitation learning

State Policy

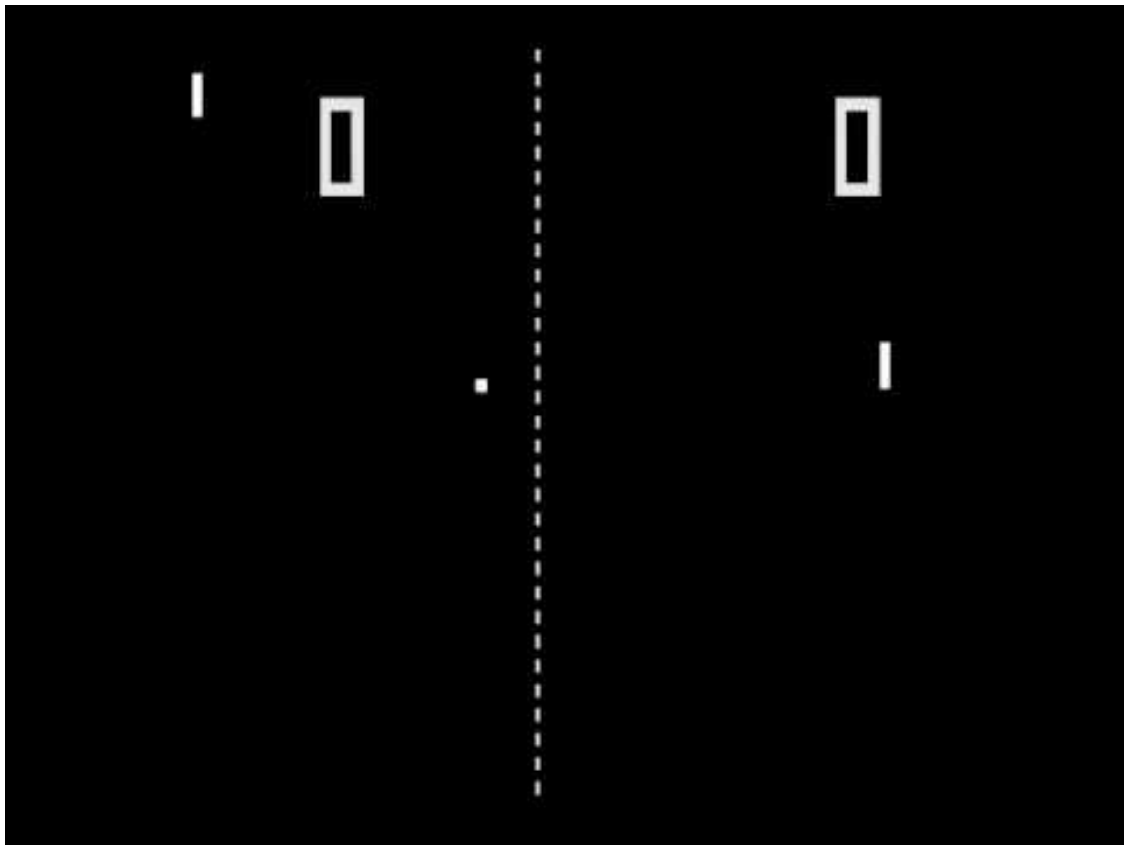
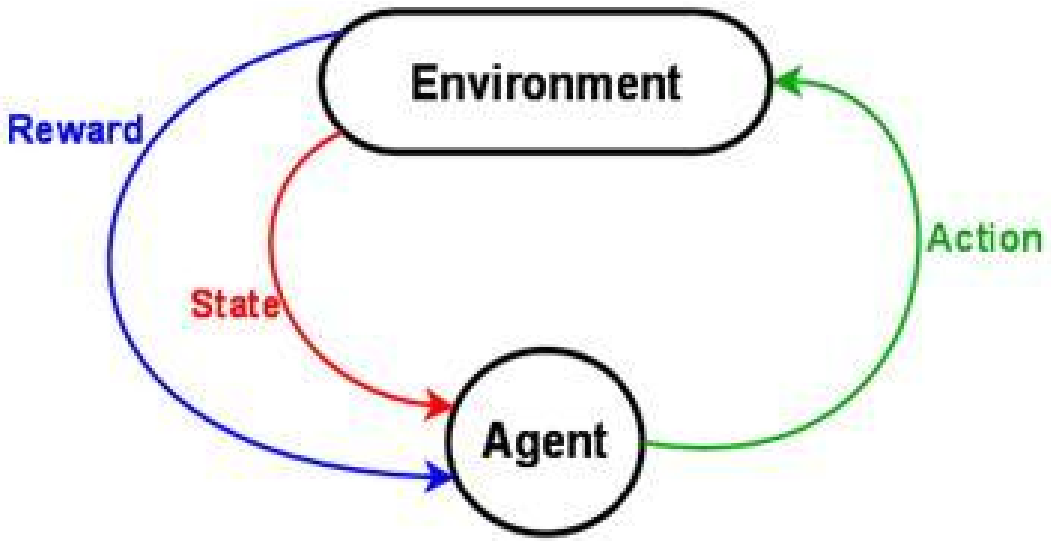
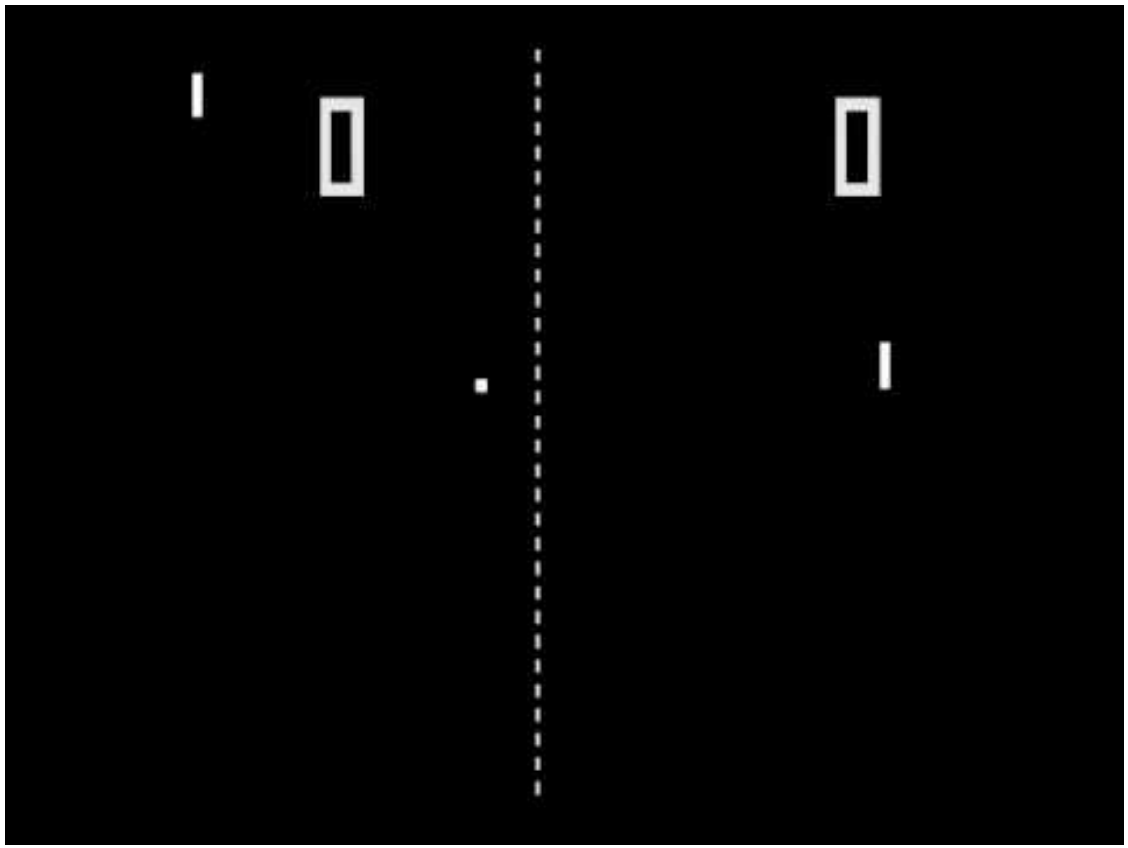


Figure: Playing FPS Games with Deep Reinforcement Learning (Lample, Chaplot. ArXiv 2016)


Environment
Reward
Action
State
Agent



$$state \rightarrow s$$

$$a \sim \pi_\theta(a|s)$$

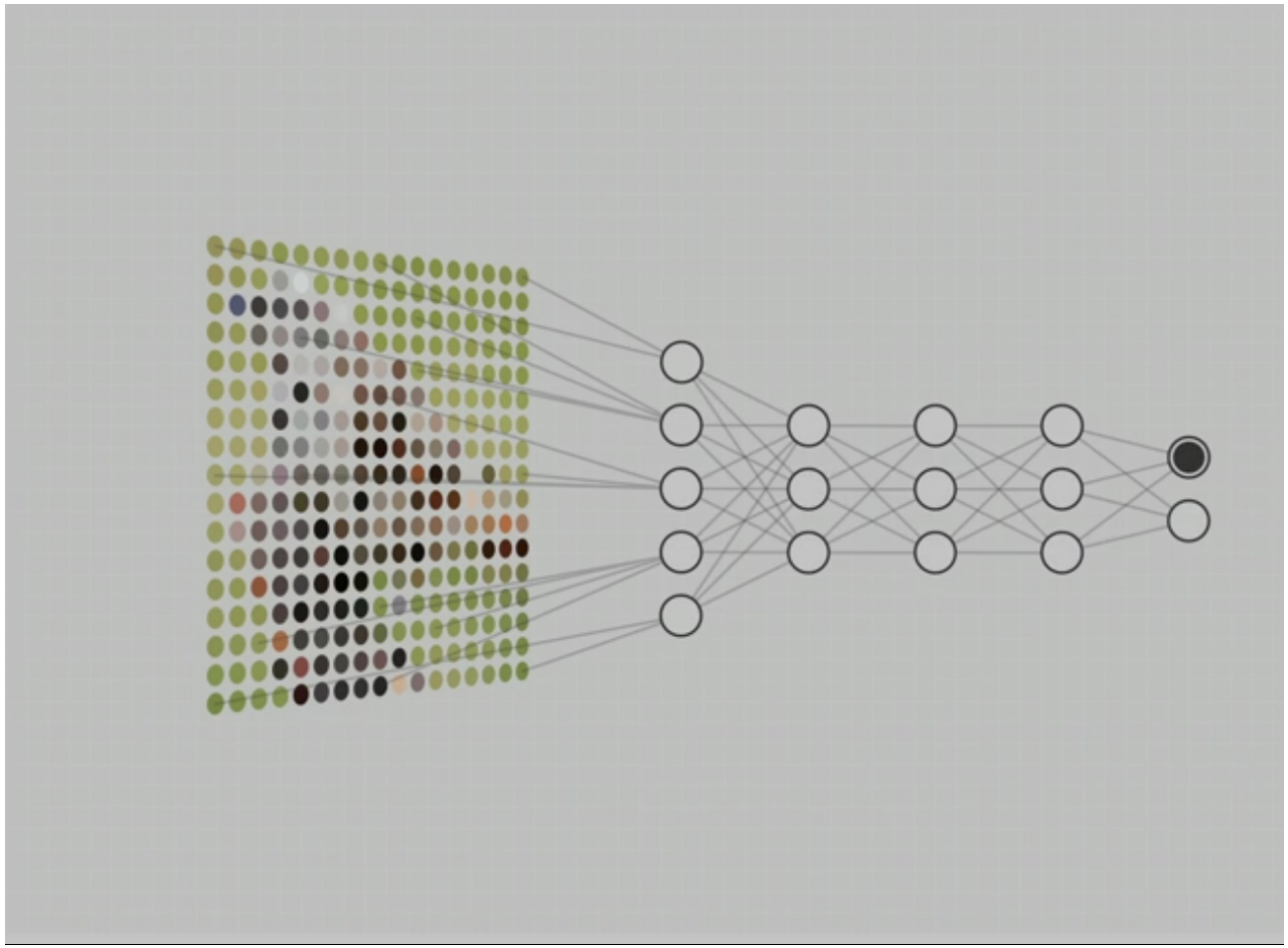


Figure: credits: Blaise Aguera y Arcas

Python code


```
h = np.tanh(np.dot(W1, s) + b1))
logpi = np.dot(W2, h) + b2
p = softmax(logpi)
```

How to train a policy?

Python code


```
h = np.tanh(np.dot(W1, s) + b1))
logpi = np.dot(W2, h) + b2
p = softmax(logpi)
```

Gradient Descent

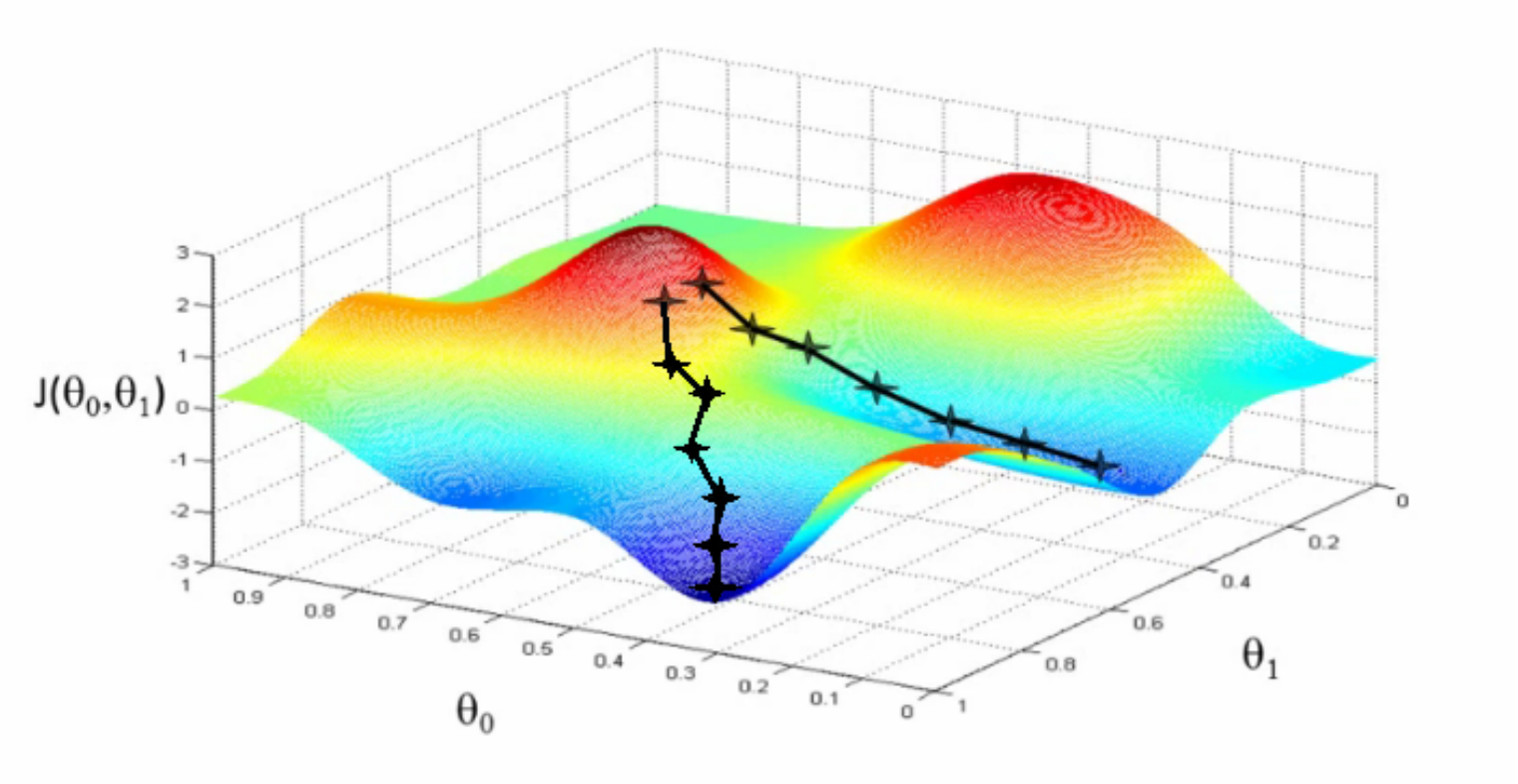


Figure: Credits: ML blog Vasilis Vryniotis

Policy Gradient

1. Reinforcement learning

$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t)R$$

2. Supervised learning

$$grad = \sum_i \nabla_\theta log\ p(y_i|x_i)$$

Roll outs

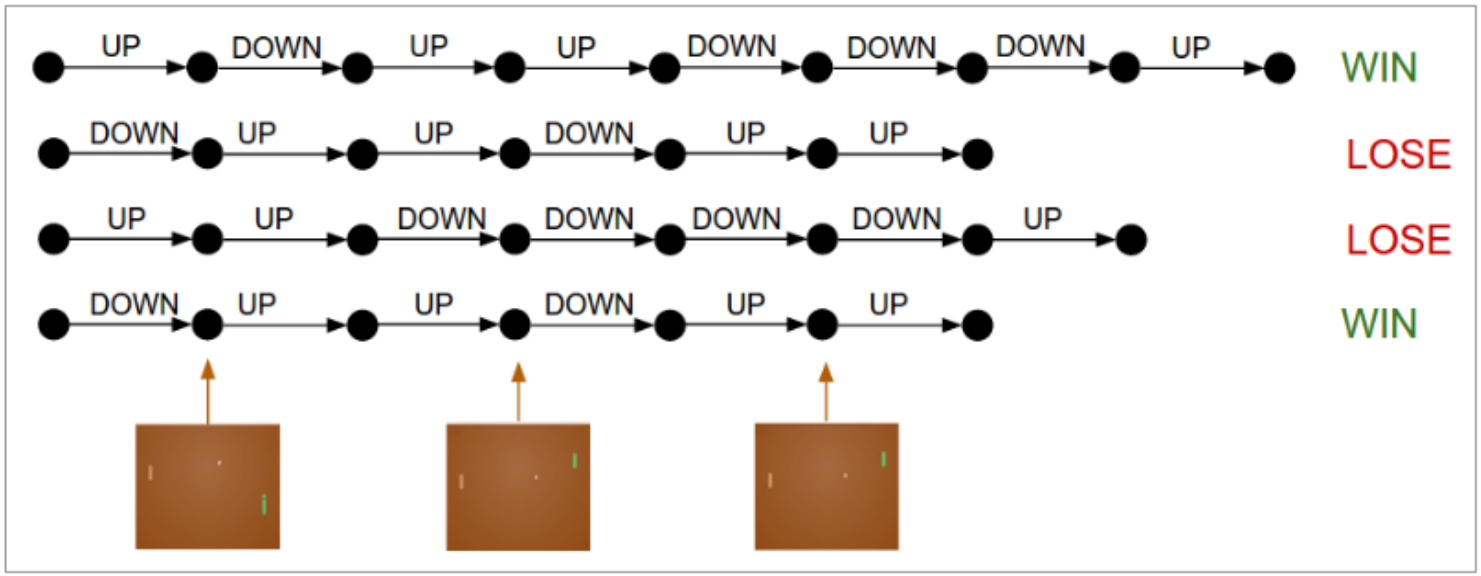


Figure: Credits: karpathy.github.io

Policy Gradient

1. Reinforcement learning

$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t)R$$

2. Supervised learning

$$grad = \sum_i \nabla_\theta log\ p(y_i|x_i)$$

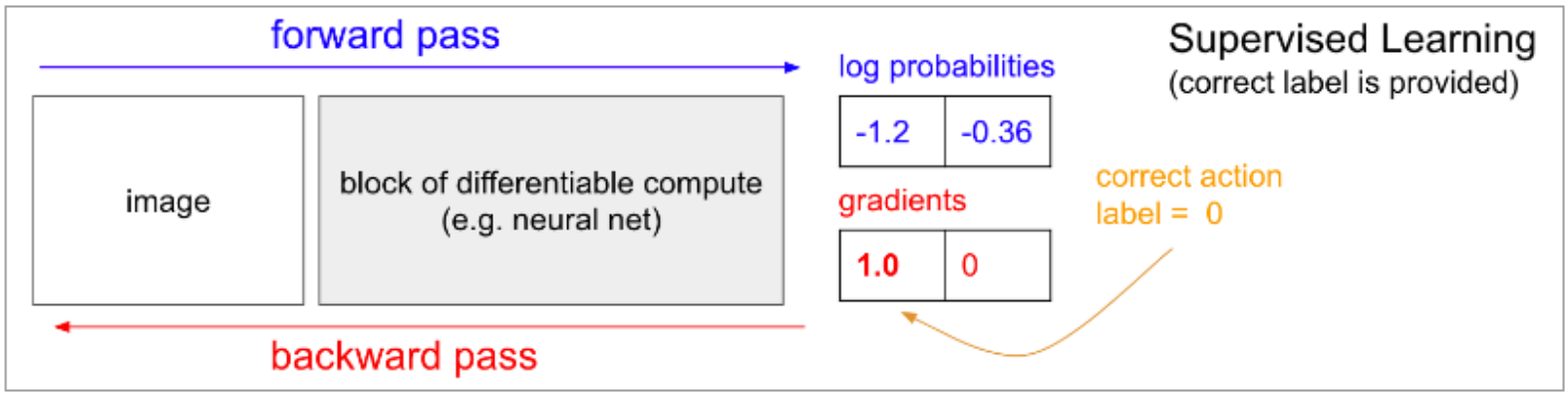


Figure: credits: karpathy.github.io

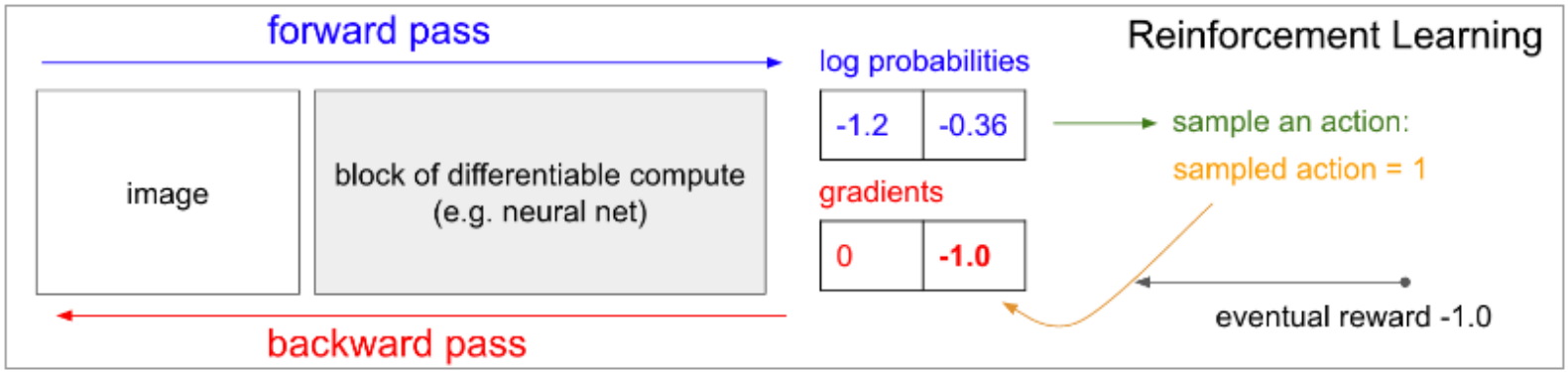


Figure: credits: karpathy.github.io

Policy Gradient

1. Reinforcement learning

$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t)R$$

2. Supervised learning

$$grad = \sum_i \nabla_\theta log\ p(y_i|x_i)$$

Playing Pong with Policy Gradients

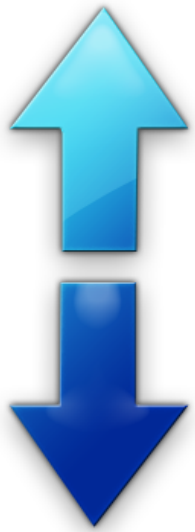
Left	Right	
+250	+180	
+50	-20	+200

Variance reduction

$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t) R$$

$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t) \sum_{t'} r_{t'}$$

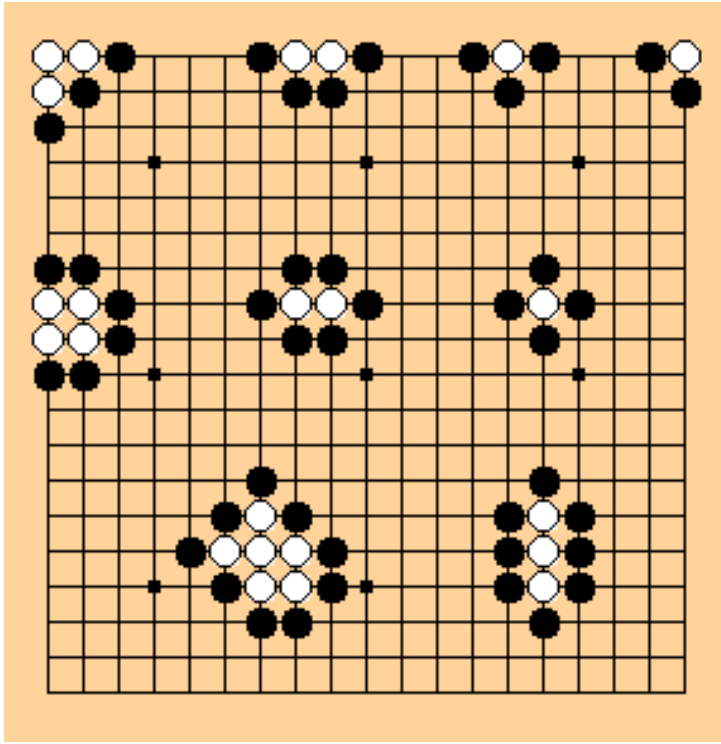
$$grad = \sum_t \nabla_\theta log\ \pi(a_t|s_t) \sum_{t'} (r_{t'} - v(s_t))$$



Actor Critic at work

Go

(a) Go



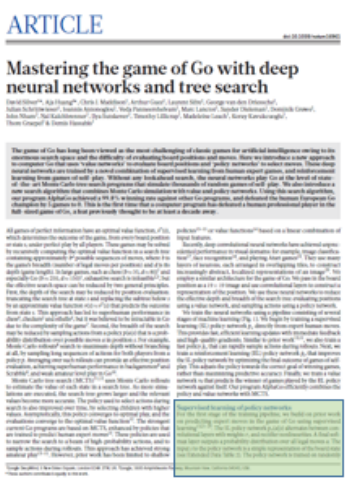
(b) Capture territory

Figure: Go

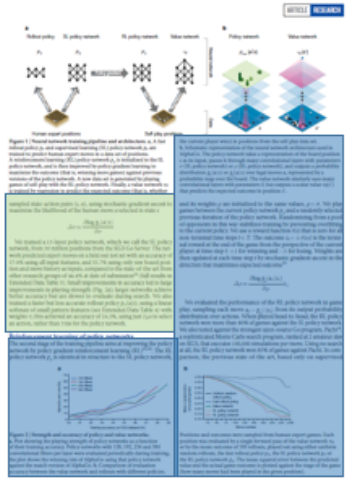
Alpha Go paper in Nature

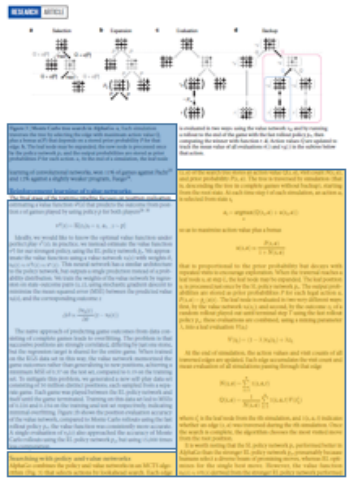
March 23, 2017

Rob Romijnders


ARTICLE

Mastering the game of Go with deep neural networks and tree search

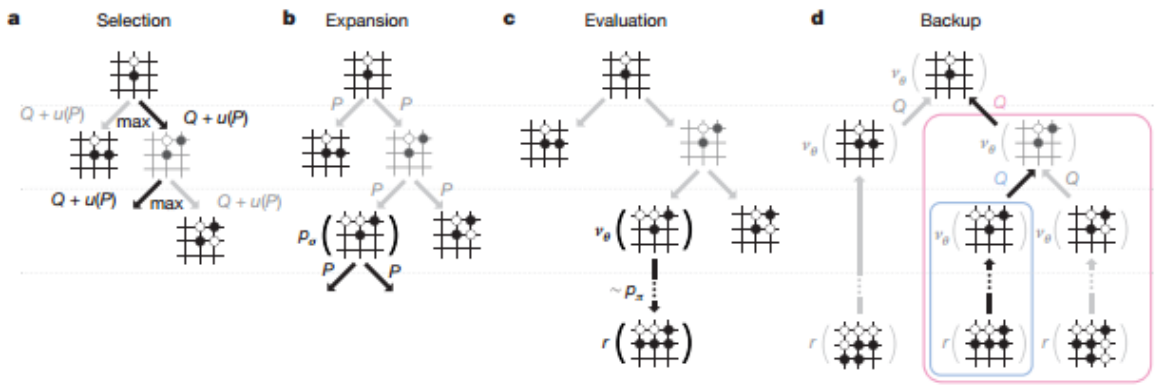




MCTS

March 23, 2017

Rob Romijnders

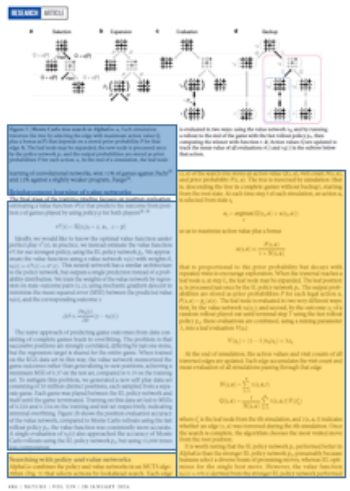


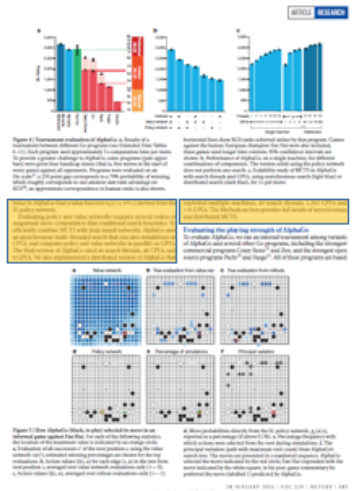
March 23, 2017

Rob Romijnders


ARTICLE

Mastering the game of Go with deep neural networks and tree search





Conclusion

Questions

References on next slide

Further reading

1. Reinforcement learning
   1. David Silver's course at UCL
   2. Book: Reinforcement Learning by Sutton and Barto
2. Deep Reinforcement learning
   1. Deep Reinforcement Learning: Pong from Pixels, Andrej Karpathy
   2. Deep Reinforcement Learning, CS 294 at Berkeley
3. Play yourself
   1. Pol. Grad. in 130 lines using only NumPy
   2. Open AI gym

Venn diagram

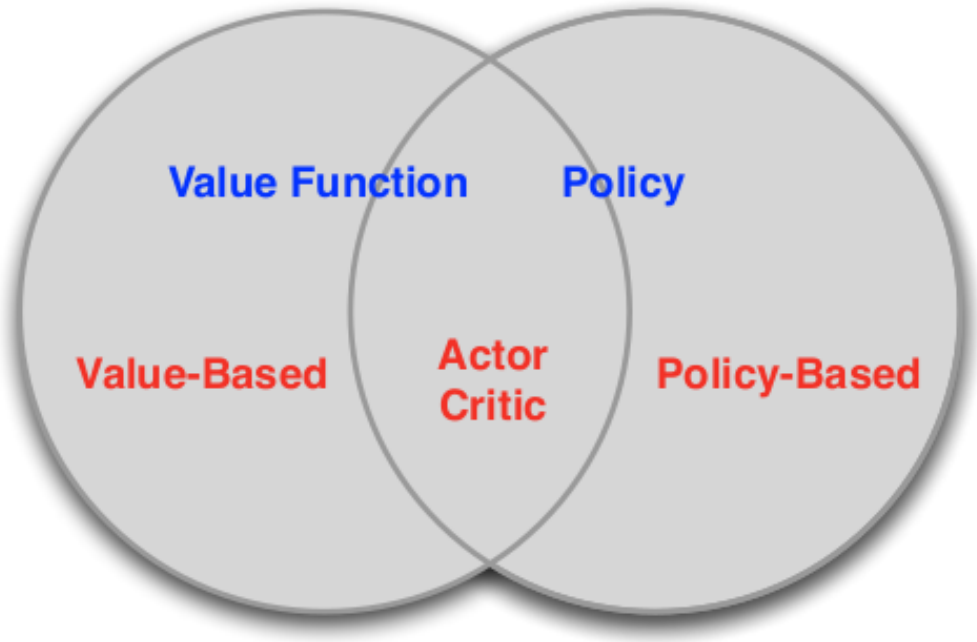


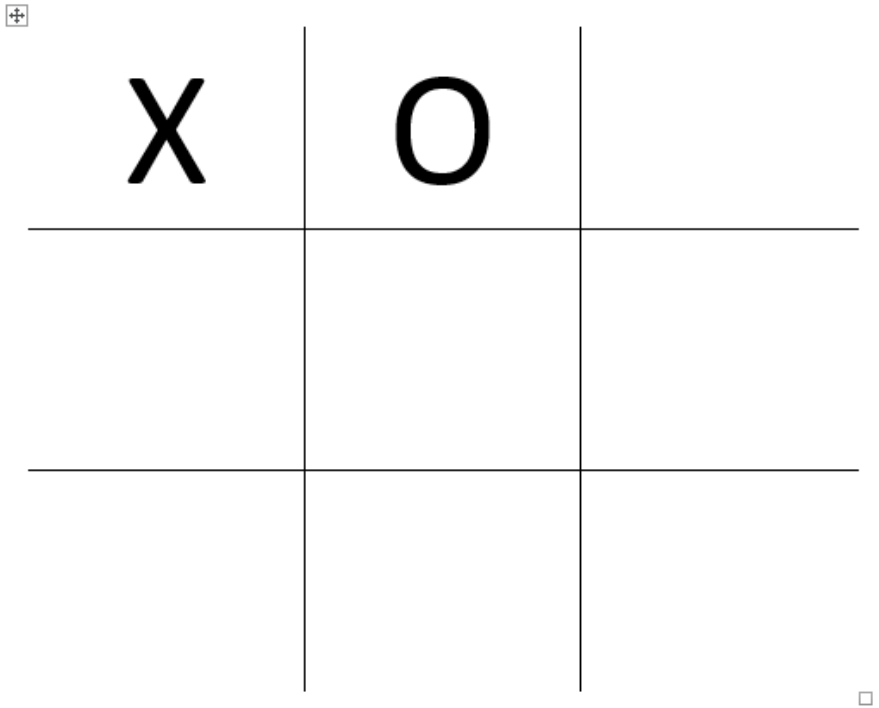
Figure: Credits: David Silver's slides at UCL

Quiz

1. Cluster customers based on their features
2. Play chess with an algorithm
3. Detect the outlier in a stream of reports
4. Predict tomorrow's weather
5. Indicate faces in an image
6. Have the robot find the exit of a maze

State

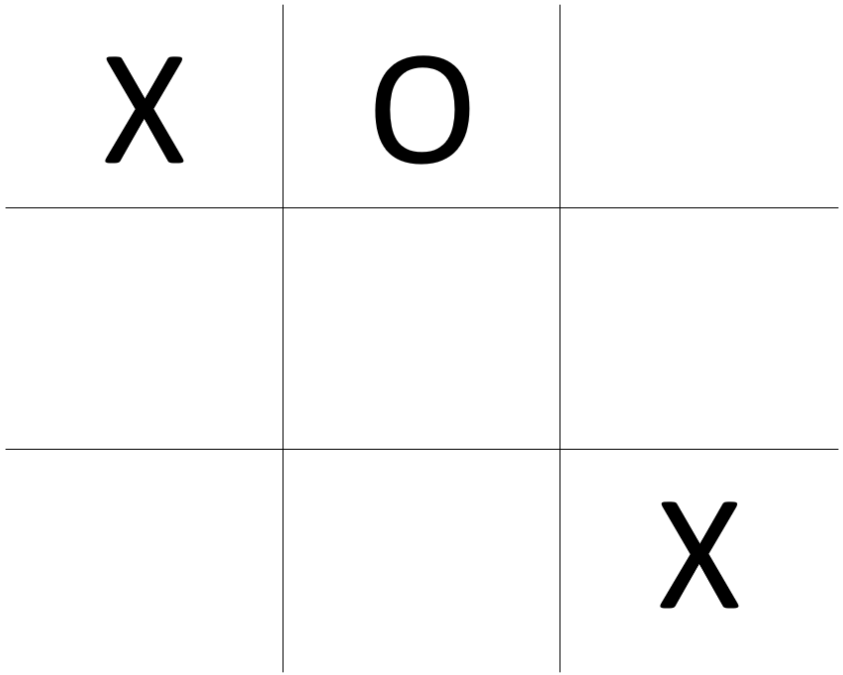
Figure: State of a chess board. Credits: chess.com


X
O

0:X	1:O	2:_
3:_	4:_	5:_
6:_	7:_	8:_

0	X
1	O
2	_
3	_
4	_
5	_
6	_
7	_
8	_

$$s = \begin{pmatrix} 1 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$


X
O
X

0:X	1:O	2:_
3:_	4:_	5:_
6:_	7:_	8:X

0	X
1	O
2	_
3	_
4	_
5	_
6	_
7	_
8	X

$$s = \begin{pmatrix} 1 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$