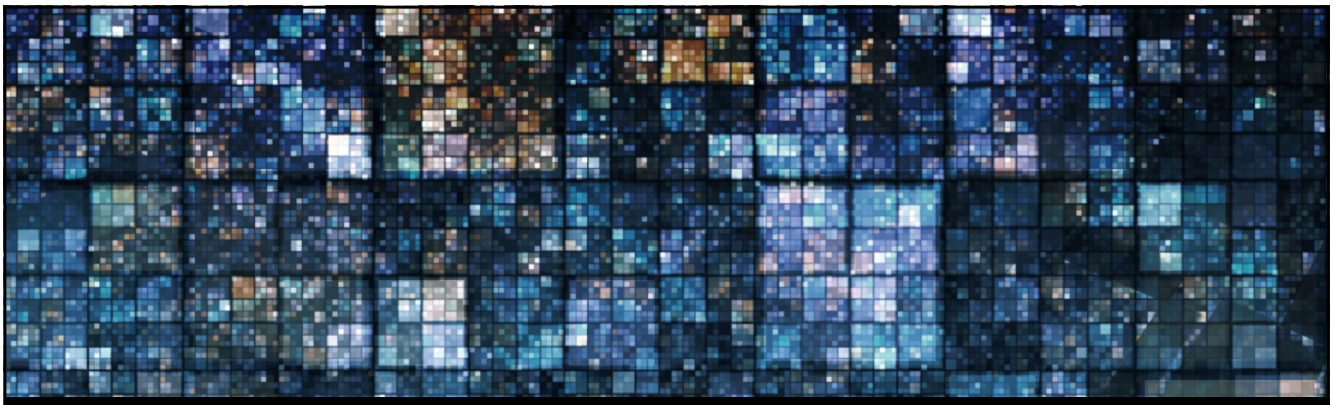
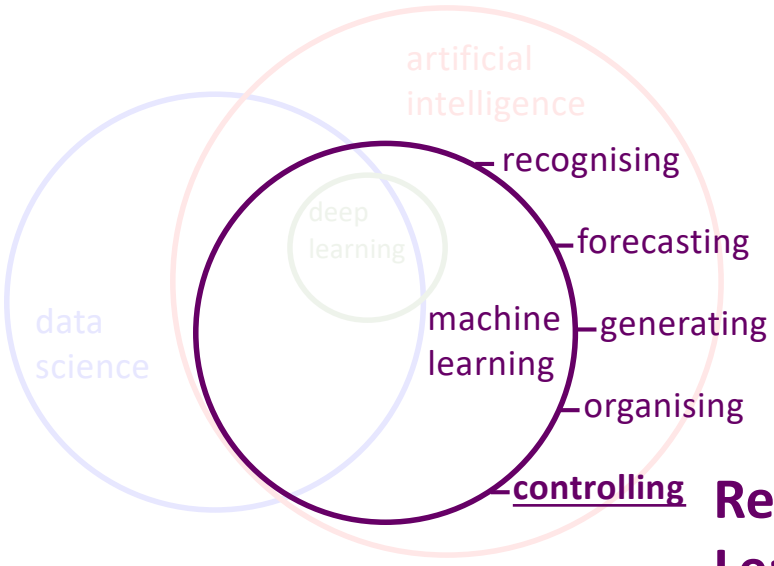


COMP47590
Advanced Machine Learning
Reinforcement Learning - Introduction

Dr. Brian Mac Namee

Information

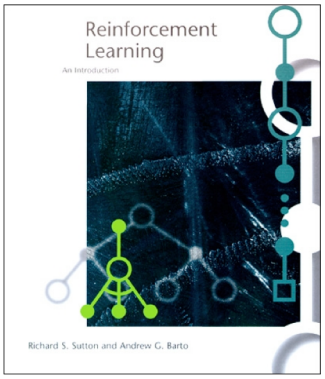
Email:	Brian.MacNamee@ucd.ie
Course Materials:	All material posted on UCD CS moodlehttps://csmoodle.ucd.ie/moodle/course/view.php?id=663 Enrolment key **UCDAdvML2017**



Reinforcement Learning

Reference Book

Much of the content in this section will follow "Reinforcement Learning An Introduction", 2nd Edition, Richard S. Sutton and Andrew G. Barto, MIT Press, 2018 (to appear)



A legitimate online version of this book is available at:

www.incompleteideas.net/book/the-book-2nd.html

INTRODUCING REINFORCEMENT LEARNING

Reinforcement Learning

"**Reinforcement learning** is learning what to do - how to map situations to actions - so as to maximize a numerical reward signal."

Sutton & Barto

Reinforcement Learning: An Introduction , Second edition, in progress
Richard S. Sutton and Andrew G. Barto, MIT Press, 2017
www.incompleteideas.net/book/the-book-2nd.html

Reinforcement Learning

The learner is not told which actions to take, but must discover which actions yield the most reward by trying them

- Often actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards

The two most important distinguishing features of reinforcement learning are:

- trial-and-error search
- delayed reward

Donal Michie's MENACE (1960s)

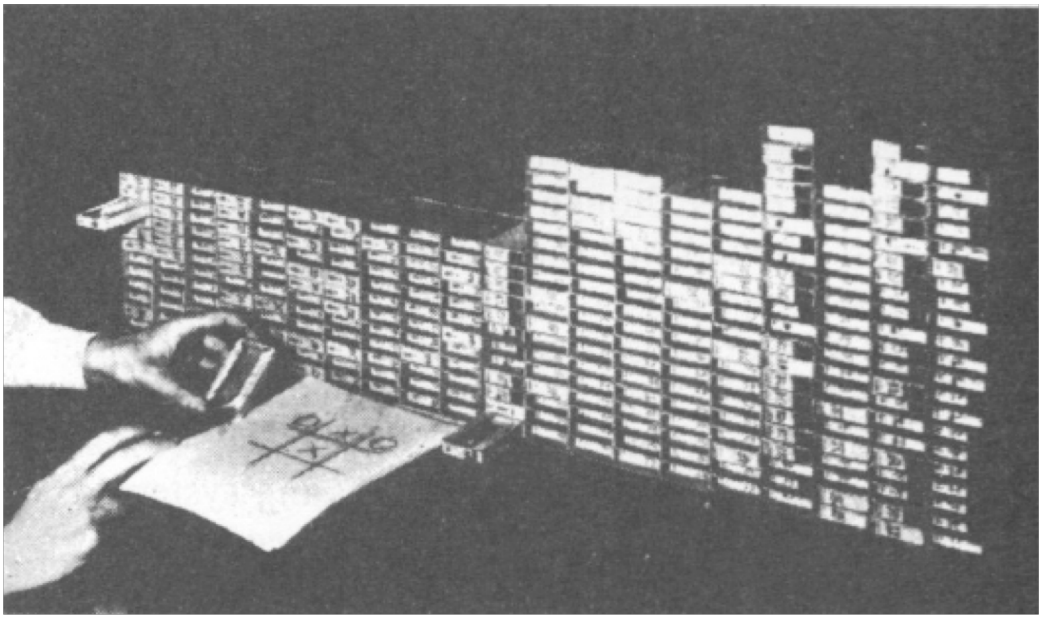
In the 1960s University of Edinburgh academic Donald Michie built MENACE, a *machine* that could *learn* to play the game of Xs & Os (or noughts and crosses, or tic-tac-toe)

Constructed from 304 matchboxes carefully filled with precise numbers of collared beads

With the help of a human operator mindlessly following some simple rules MENACE could play Xs & Os and learn to get better at it!

Future of Robotics and Machine Learning: Machine Learning Explained, Rodney Brooks, 2017
https://rodneybrooks.com/forai-machine-learning-explained/

Donal Michie's MENACE (1960s)

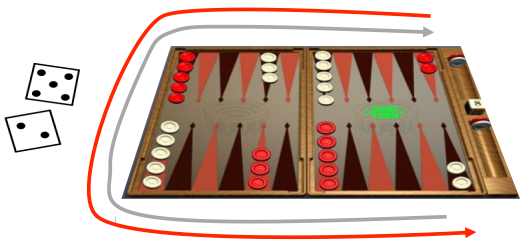


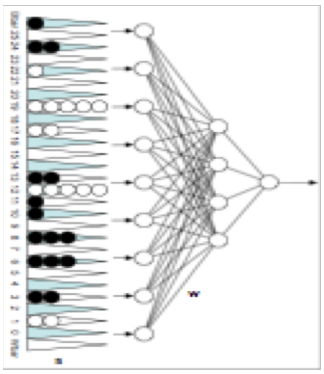
Future of Robotics and Machine Learning: Machine Learning Explained, Rodney Brooks, 2017
https://rodneybrooks.com/forai-machine-learning-explained/

Donal Michie's MENACE (1960s)

A New THEORY of AWESOMENESS and MIRACLES, Being NOTES and SLIDES on a talk given at PLAYFUL 09, concerning CHARLES BABBAGE, HEATH ROBINSON, MENACE and MAGE, 2009, James BRIDLE
http://shorttermmemoryloss.com/menace/

TD-Gammon (1990s)





estimated state value
(≈ prob of winning)

Start with a random network

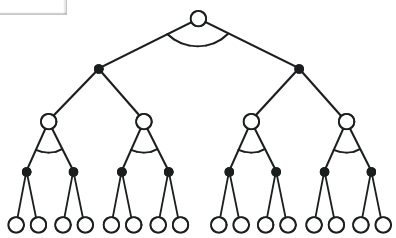
Play millions of games against itself

Learn a value function from this simulated experience

6 weeks later it is the best player of backgammon in the world!

Originally used expert, handcrafted features, later repeated with raw board positions

Action selection
by a shallow search



Tesauro, Gerald. "Temporal difference learning and TD-Gammon."
Communications of the ACM 38.3 (1995): 58-68.
http://enzodesiage.com/wp-content/uploads/2017/08/tesauro-tdgammon-1995.pdf

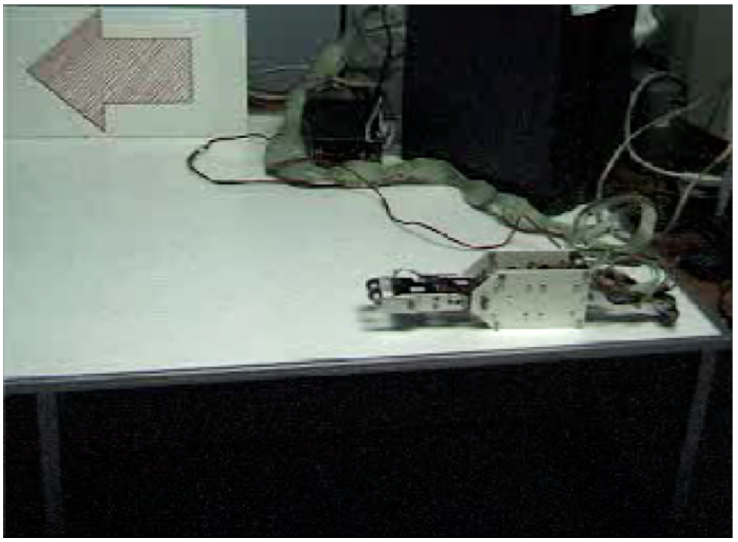
Hajime Kimura's RL Robots (1990s)

Small robots composed of a simple two link arm controlled with two servo motors

A small wheel sensor to record movement is included

The goal is to learn control rules to move to the front or back - a move is taken about every 0.2 sec

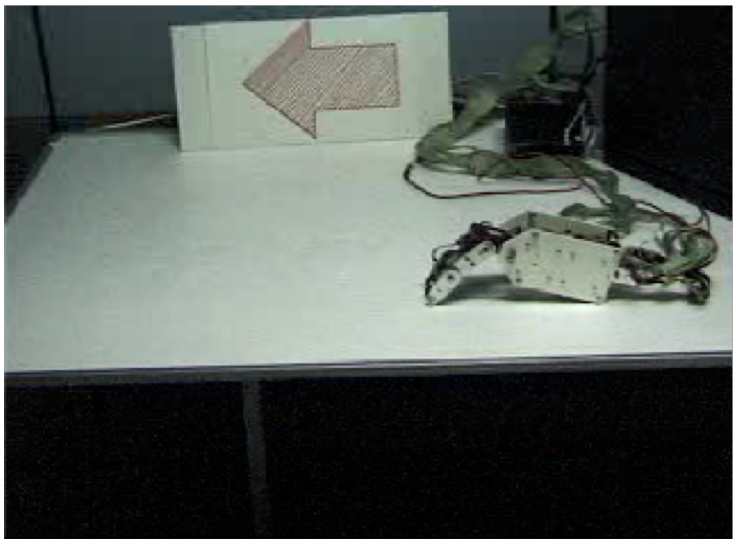
The body's movement for each time step is given to the agent as an instantaneous reward



Reinforcement Learning using Stochastic Gradient Algorithm and its Application to Robots, The Transaction of the Institute of Electrical Engineers of Japan, Vol.119, No.8 (1999) Hajime Kimura, Shigenobu Kobayashi
http://sysplan.nams.kyushu-u.ac.jp/gen/papers/denki99.pdf (in Japanese!)

Hajime Kimura's RL Robots (1990s)

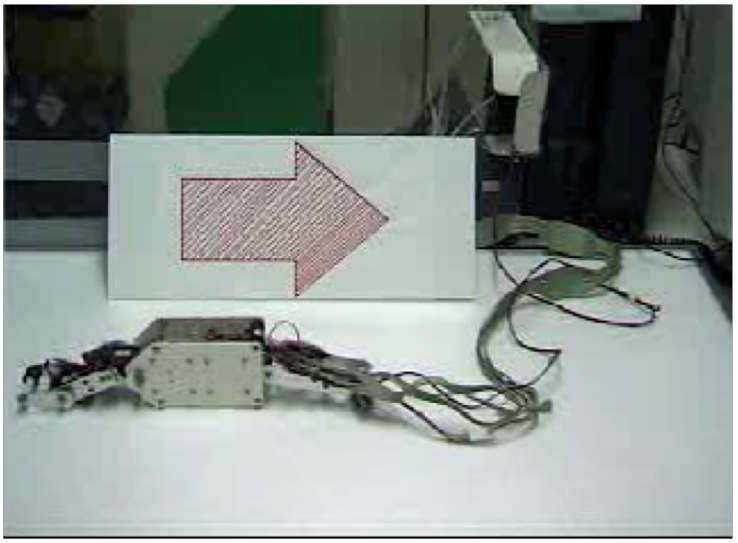
- Small robots composed of a simple two link arm controlled with two servo motors
- A small wheel sensor to record movement is included
- The goal is to learn control rules to move to the front or back - a move is taken about every 0.2 sec
- The body's movement for each time step is given to the agent as an instantaneous reward



Reinforcement Learning using Stochastic Gradient Algorithm and its Application to Robots, The Transaction of the Institute of Electrical Engineers of Japan, Vol.119, No.8 (1999) Hajime Kimura, Shigenobu Kobayashi http://sysplan.nams.kyushu-u.ac.jp/gen/papers/denki99.pdf (in Japanese!)

Hajime Kimura's RL Robots (1990s)

- Small robots composed of a simple two link arm controlled with two servo motors
- A small wheel sensor to record movement is included
- The goal is to learn control rules to move to the front or back - a move is taken about every 0.2 sec
- The body's movement for each time step is given to the agent as an instantaneous reward

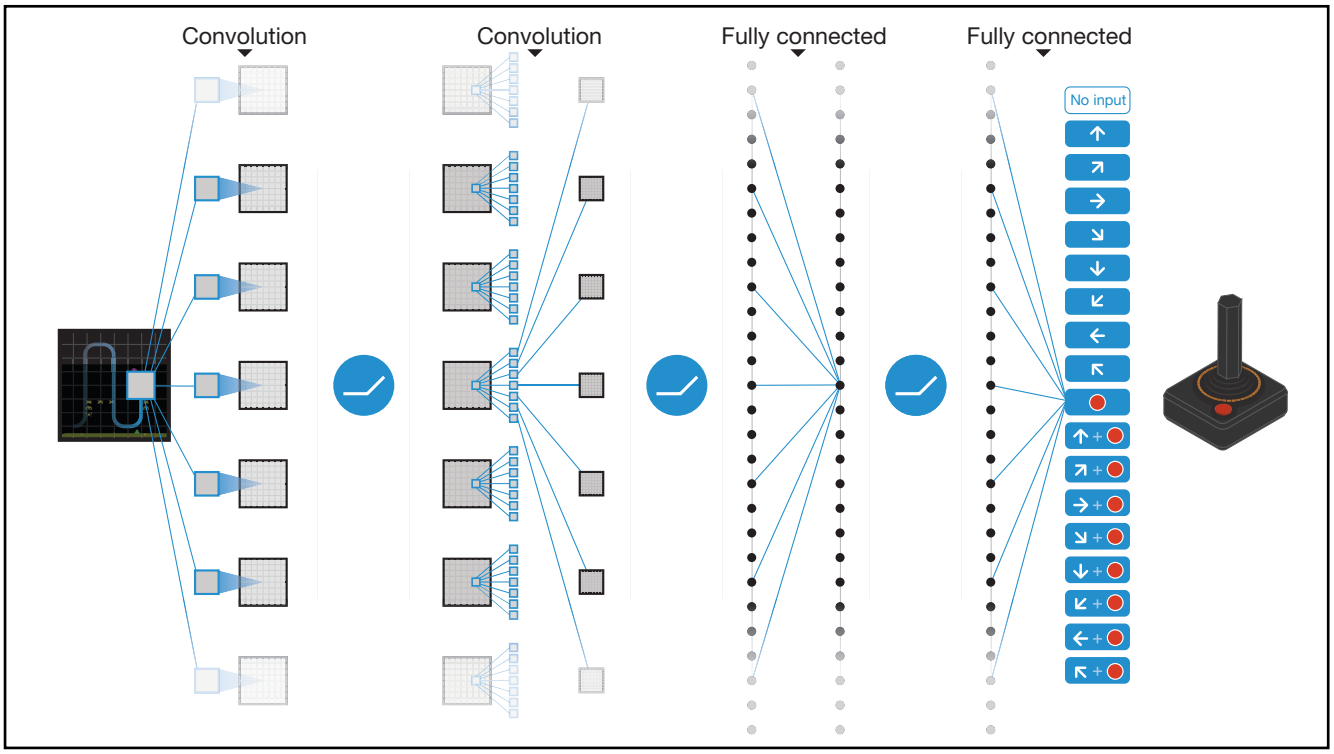


Reinforcement Learning using Stochastic Gradient Algorithm and its Application to Robots, The Transaction of the Institute of Electrical Engineers of Japan, Vol.119, No.8 (1999) Hajime Kimura, Shigenobu Kobayashi http://sysplan.nams.kyushu-u.ac.jp/gen/papers/denki99.pdf (in Japanese!)

Deep Reinforcement Learning Applied to Classic Atari Video Games (2010s)

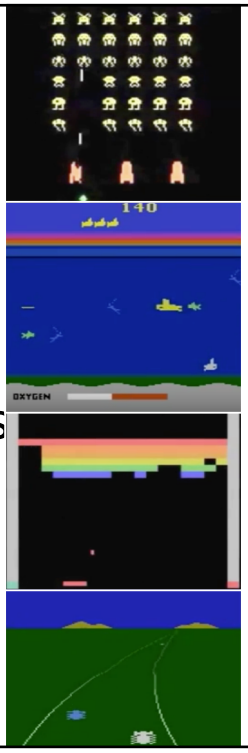
Space Invaders Breakout Enduro

https://www.youtube.com/watch?v=V1eYniJ0Rnk



Deep Reinforcement Learning Applied to Classic Atari Video Games (2010s)

- Learned to play 49 games for the Atari 2600 game console, without labels or human input, from self-play and the score alone
- Learned to play better than all previous algorithms and at human level for more than half the games
- Same learning algorithm applied to all 49 games w/o human tuning!



Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. and Petersen, S., 2015. Human-level control through deep reinforcement learning. *Nature*, *518*(7540), p.529. http://www.davidqiu.com:8888/research/nature14236.pdf

Reinforcement Learning vs Supervised Learning

Supervised learning requires a training set of labelled instances provided by a knowledgeable external supervisor

- Each instance is a description of a situation together with the correct action the system should take
- The object is for the system to generalize its responses so that it acts correctly in situations not present in the training set

In interactive problems it is often impractical to obtain these labelled training sets

- In uncharted territory an agent must be able to learn from its own experience

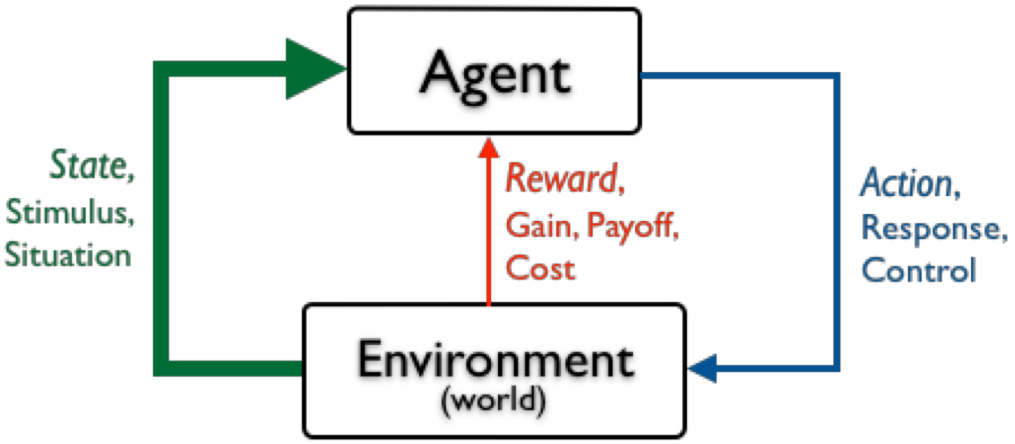
Reinforcement Learning vs Unsupervised Learning

Unsupervised learning is typically about finding structure hidden in collections of unlabelled data

Although reinforcement learning does not rely on examples of correct behaviour it is not unsupervised

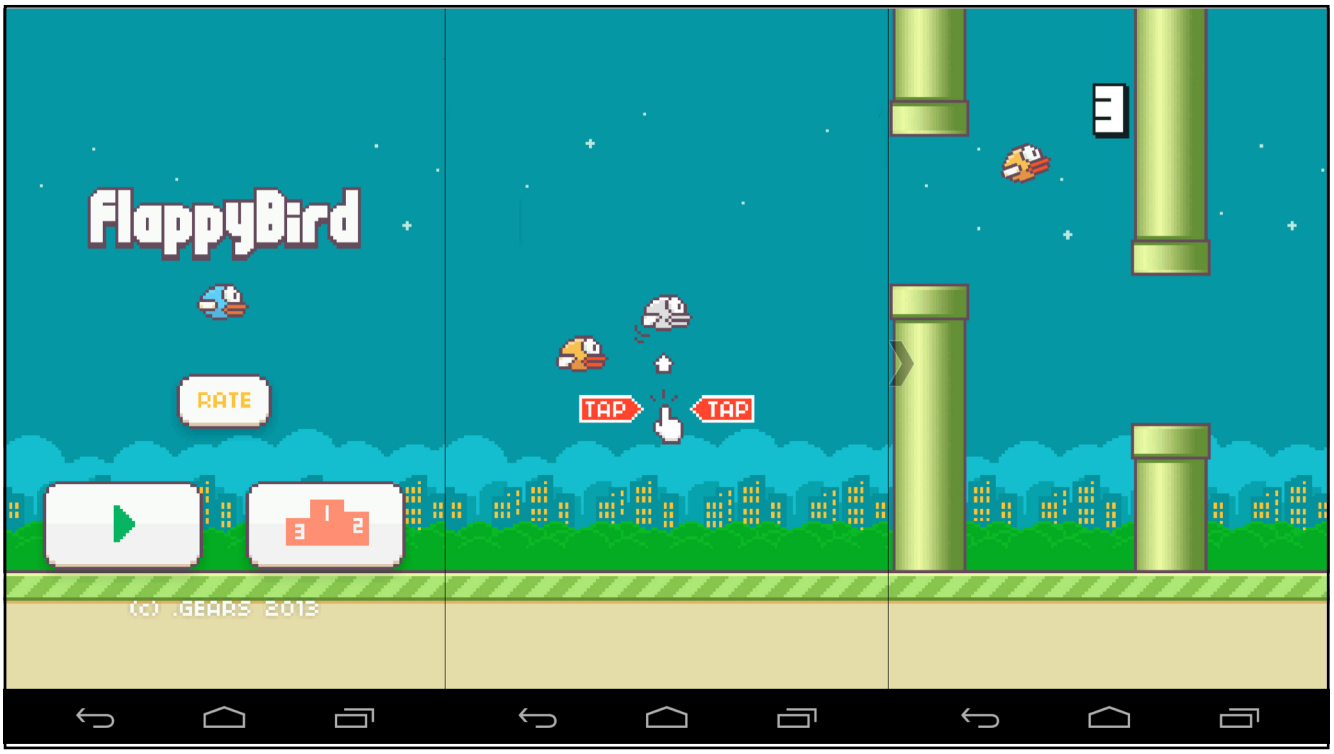
Reinforcement learning maximizes a **reward signal** which is a type of supervision

Reinforcement Learning



Reinforcement Learning: An Introduction , Second edition, in progress
Richard S. Sutton and Andrew G. Barto, MIT Press, 2017
www.incompleteideas.net/book/the-book-2nd.html

SIMPLE RL EXAMPLE



CHARACTERISTICS OF REINFORCEMENT LEARNING

Reinforcement Learning

Reinforcement Learning: An Introduction , Second edition, in progress
Richard S. Sutton and Andrew G. Barto, MIT Press, 2017
www.incompleteideas.net/book/the-book-2nd.html

Key Elements of Reinforcement Learning

Beyond the agent and the environment there are four main sub-elements of a reinforcement learning system:

- a policy
- a reward signal
- a value function
- a model of the environment (optional)

Key Elements of Reinforcement Learning

A **policy** defines the learning agent's way of behaving at a given time.

- A policy is a mapping from perceived states of the environment to actions to be taken when in those states
- Corresponds to what psychology would call a set of *stimulus–response rules*
- Policy implementations can range from a simple lookup table to extensive computation such as a search process or a deep learning model
- The policy is the core of a reinforcement learning agent

Key Elements of Reinforcement Learning

A **reward signal** defines the goal in a reinforcement learning problem

- On each time step, the environment sends the agent a single number called the reward
  - Perhaps analogous to the experiences of pleasure or pain
- The agent's sole objective is to maximize the total reward it receives over the long run
- The reward signal is the primary basis for altering the policy
  - if an action selected by the policy is followed by low reward, then the policy may be changed to select some other action in that situation in the future

Key Elements of Reinforcement Learning

Whereas the reward signal indicates what is good in an immediate sense, a **value function** specifies what is good in the long run

- The value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state
- A state might always yield a low immediate reward but still have a high value because it is regularly followed by other states that yield high rewards (or the reverse could be true)
- To make a human analogy eating a cake might have high reward (because it tastes nice) but low value (because it will make me fat)

Key Elements of Reinforcement Learning

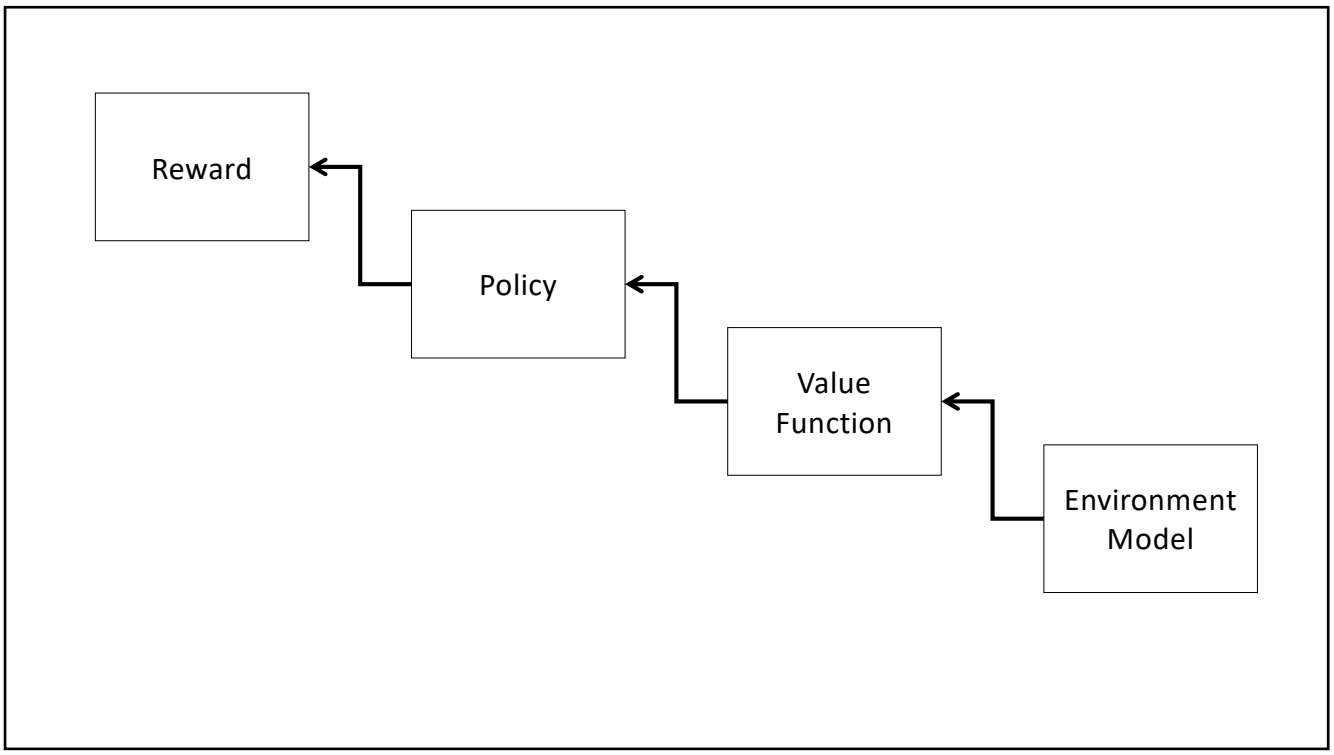
Rewards are primary, while values are secondary

- Without rewards there could be no values, and the purpose of estimating values is to achieve more reward
- However, it is values with which we are most concerned when making and evaluating decisions
- We seek actions that bring about states of highest value, not highest reward, because these actions obtain the greatest amount of reward for us over the long run
- Unfortunately, it is much harder to determine values than it is to determine rewards
  - Rewards are basically given directly by the environment, but values must be estimated and re-estimated from the sequences of observations an agent makes over its entire lifetime.
- In fact, the most important component of almost all reinforcement learning algorithms we consider is a method for efficiently estimating values

Key Elements of Reinforcement Learning

Some reinforcement learning systems include a **model of the environment**

- The model should mimic the behaviour of the environment to allow an agent to predict what is likely to happen when they take an action.
- We can distinguish RL techniques as being *model-based* or *model-free* depending on whether or not they include this component



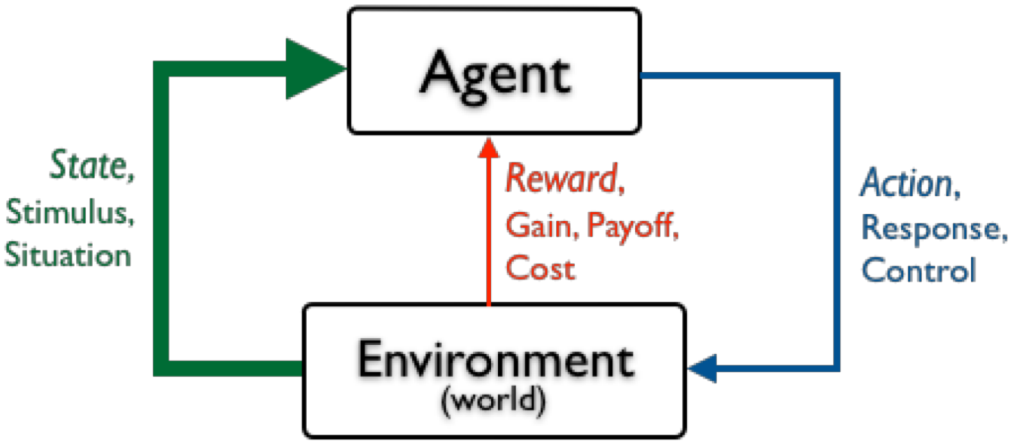
SUMMARY

Summary

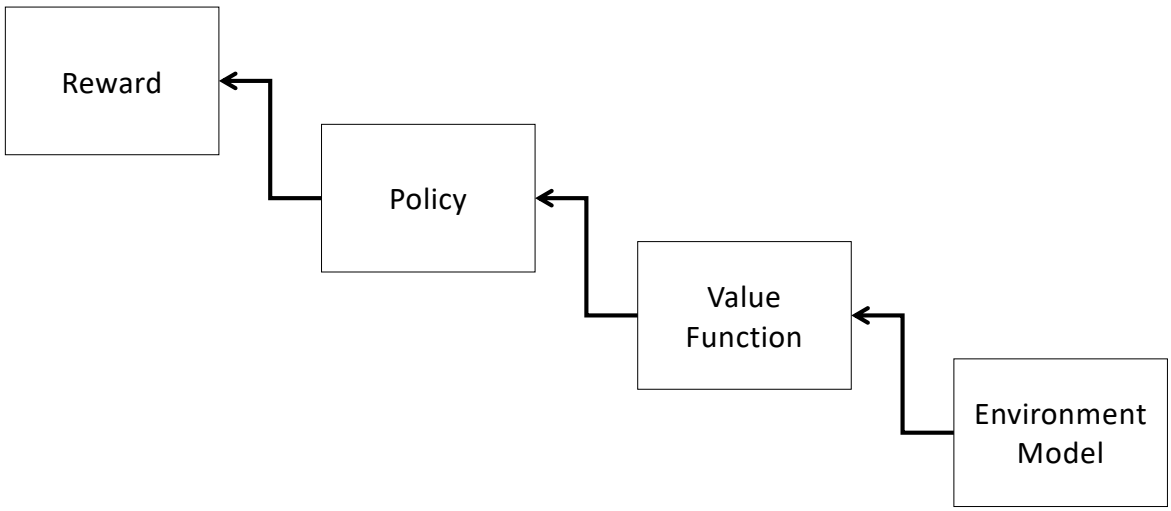
We have introduced the key ideas in reinforcement learning

- an agent
- an environment
- a policy
- a reward signal
- a value function
- a model of the environment (optional)

Reinforcement Learning



Reinforcement Learning: An Introduction , Second edition, in progress
Richard S. Sutton and Andrew G. Barto, MIT Press, 2017
www.incompleteideas.net/book/the-book-2nd.html



Summary

Over the coming lectures we will expand our discussion to cover key ideas in reinforcement learning:

- *k*-armed bandit problems
- Markov decision processes
- Temporal difference learning
- SARSA
- Q learning
- Deep Q learning

Questions