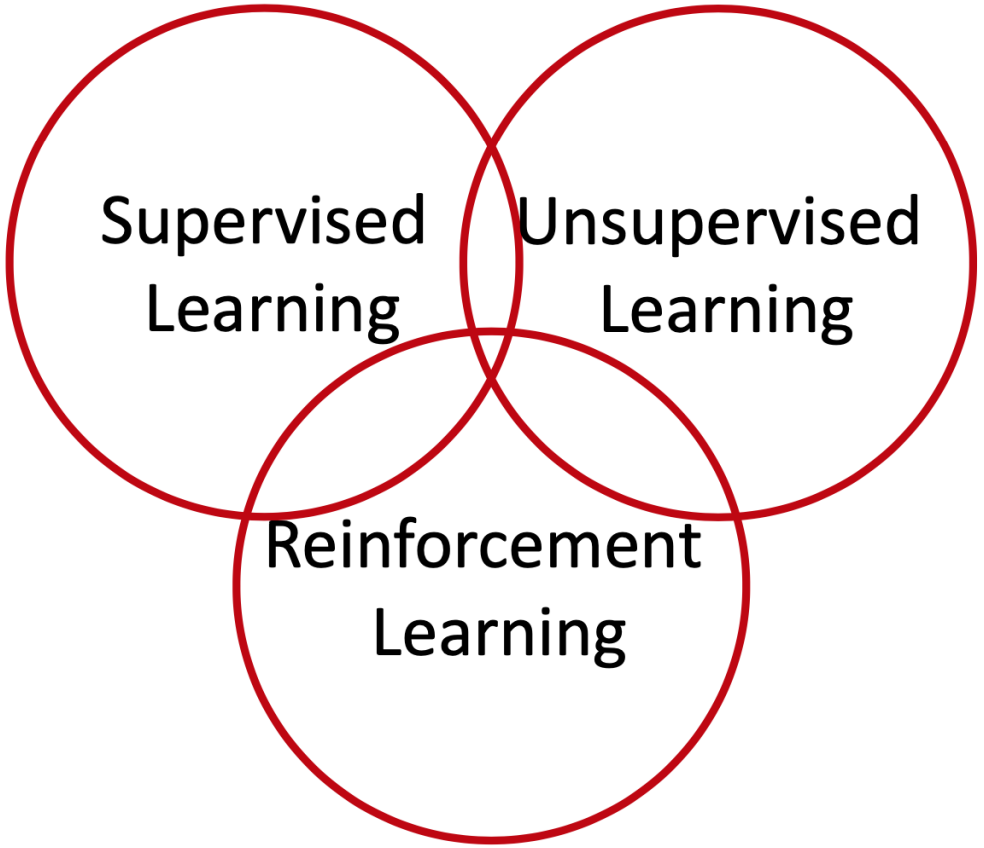
University of Salford MANCHESTER

Artificial Intelligent – Week 10

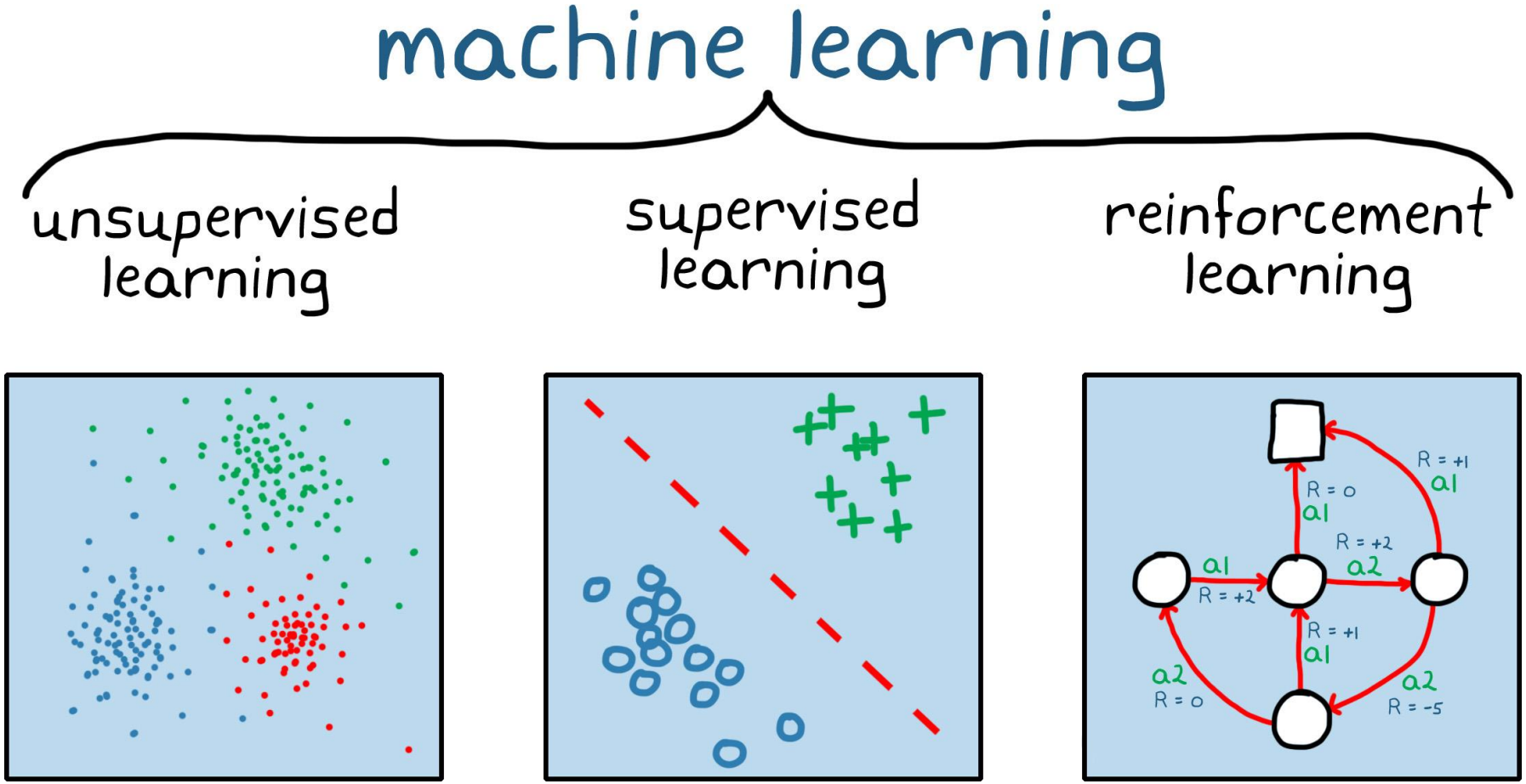
Dr Salem Ameen

S.A.AMEEN1@SALFORD.AC.UK

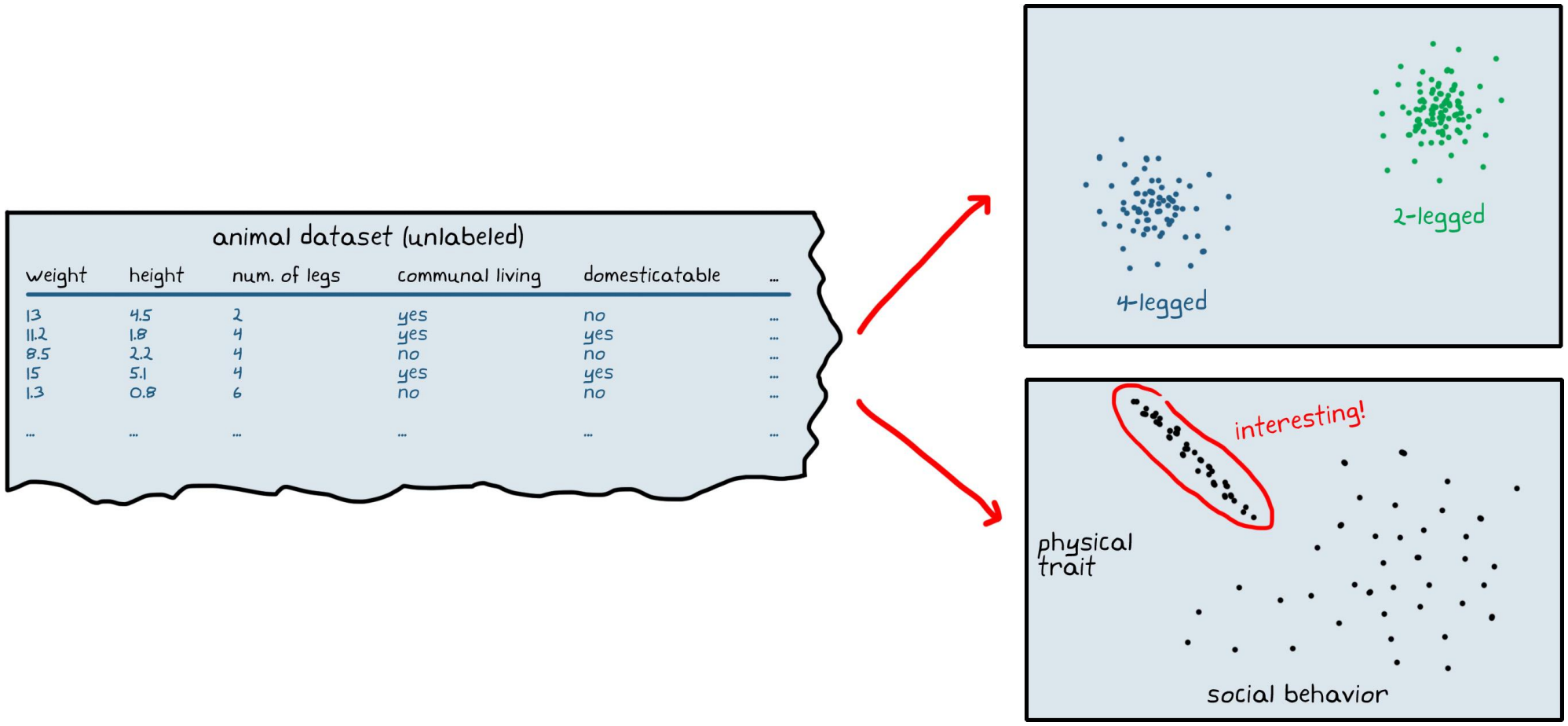
Week 10 - ML


Supervised Learning
Unsupervised Learning
Reinforcement Learning

Week 10 - ML


machine learning
unsupervised learning
supervised learning
reinforcement learning
R = +1
a1
R = 0
a1
R = +2
a2
a1
R = +2
R = +1
a1
a2
R = 0
a2
R = -5

Week 10 – Unsupervised Learning

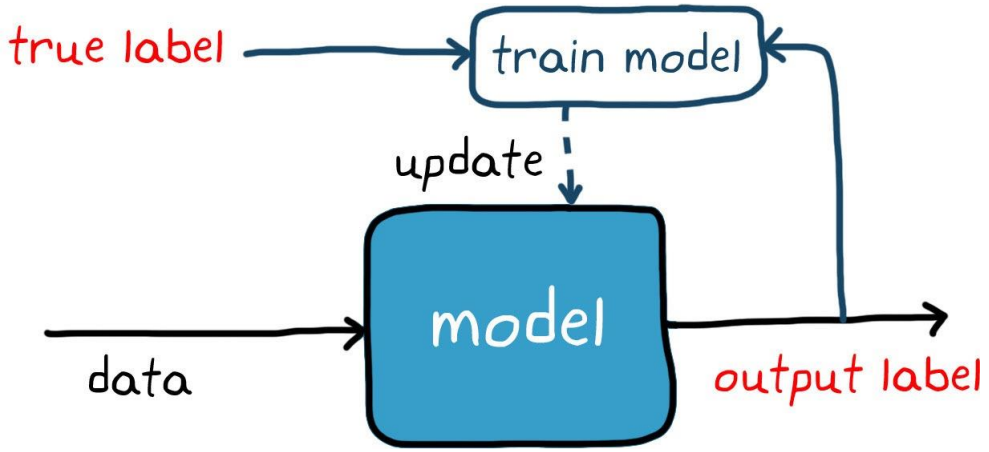


Week 10 – Supervised Learning

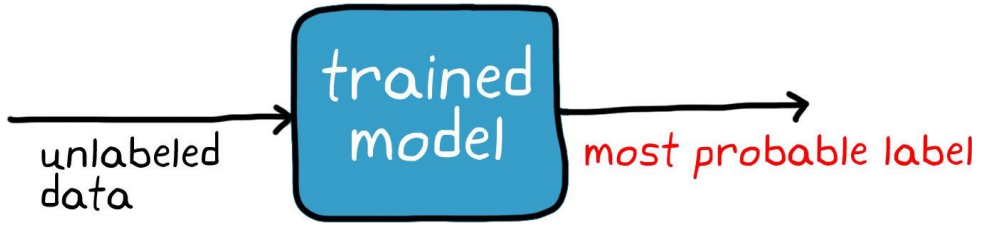
University of Salford MANCHESTER

animal dataset (labeled)

species	weight	height	num. of legs	communal living	domesticatable	...
rat	1.3	1.1	4	yes	yes	...
robin	1.2	0.8	4	no	no	...
elephant	48.5	12.2	4	yes	no	...
rabbit	25	2.1	4	yes	yes	...
spider	0.1	0.2	8	no	no	...
...	...	...	...	...	...	...



Week 10 – Supervised Learning



Week 10

University of Salford MANCHESTER

	Supervised vs Unsupervised	Regression vs Classification	Parametric vs Non-Parametric	Generative vs Discriminative
Linear Regression	Supervised	Regression	Parametric	Discriminative
Logistic Regression	Supervised	Classification	Parametric	Discriminative
k-NN	Supervised	either	Non-Parametric	Discriminative
Decision Tree	Supervised	either	Non-Parametric	Discriminative
PCA	Unsupervised	neither	Non-Parametric	neither
Clustering	Unsupervised	neither	Non-Parametric	Generative

Week 10 – Reinforcement Learning

University of Salford MANCHESTER

Reinforcement learning: given a set of rewards or punishments, learn what actions to take in the future.

Week 10 – Reinforcement Learning

University of Salford MANCHESTER

Reinforcement learning is a different beast altogether. Unlike the other two learning frameworks, which operate using a static dataset, RL works with data from a **dynamic environment**. And the goal is not to **cluster data or label data**, but to find the **best sequence of actions** that will generate the optimal outcome. The way reinforcement learning solves this problem is by allowing a piece of software called an **agent** to explore, interact with, and learn from the environment.

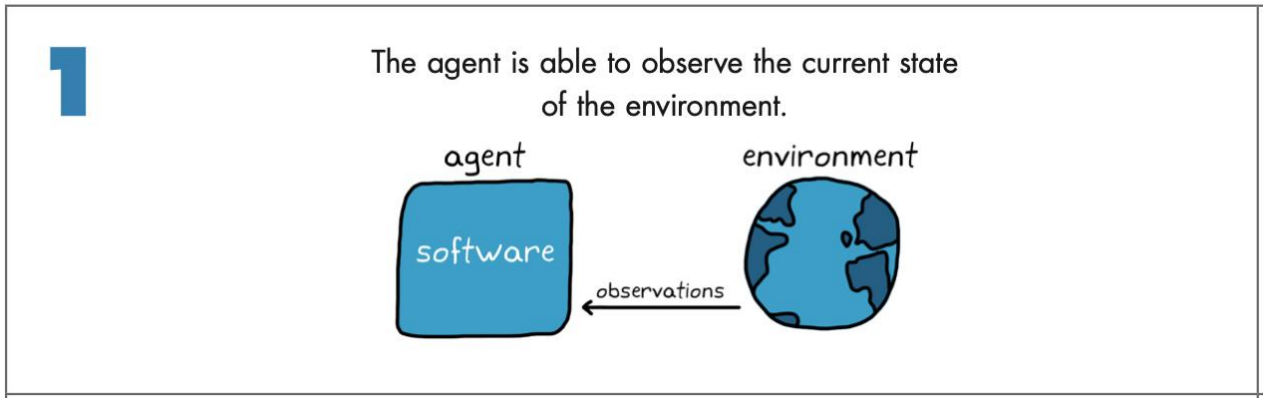
Week 10 – Reinforcement Learning

University of Salford MANCHESTER

Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.

From the book

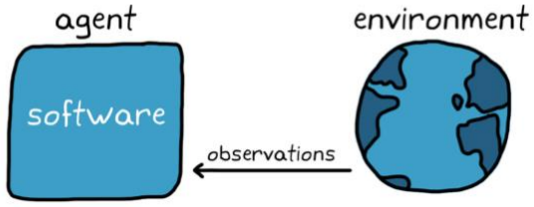
Week 10 – Reinforcement Learning



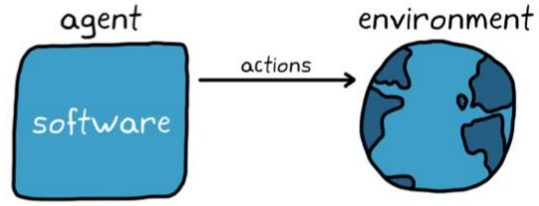
Week 10 – Reinforcement Learning

University of Salford MANCHESTER

1 The agent is able to observe the current state of the environment.



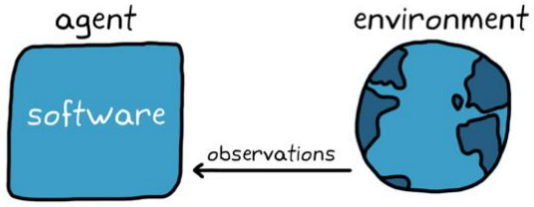
2 From the observed state, it decides which action to take.



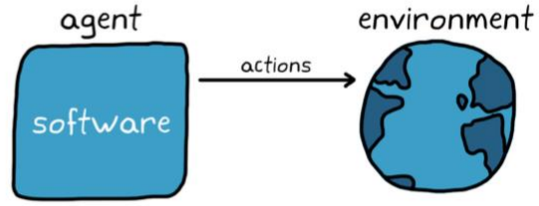
Week 10 – Reinforcement Learning

University of Salford MANCHESTER

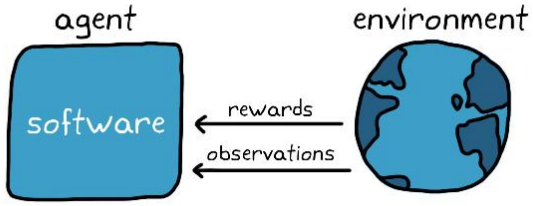
1 The agent is able to observe the current state of the environment.



2 From the observed state, it decides which action to take.



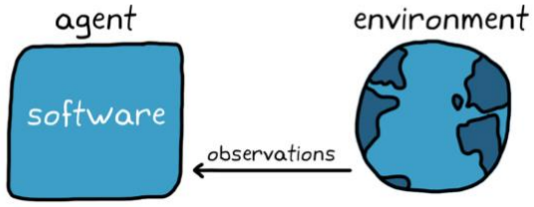
3 The environment changes state and produces a reward for that action. Both of which are received by the agent.



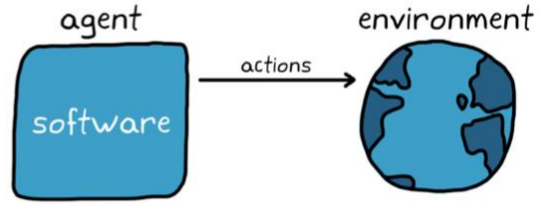
Week 10 – Reinforcement Learning

University of Salford MANCHESTER

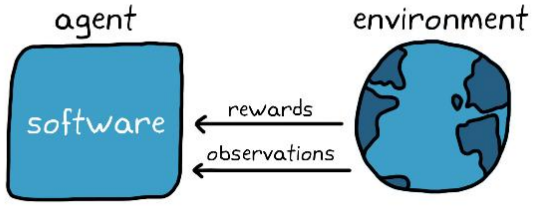
1 The agent is able to observe the current state of the environment.



2 From the observed state, it decides which action to take.

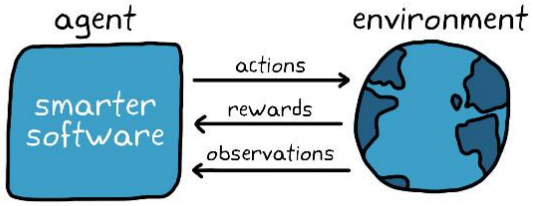


3 The environment changes state and produces a reward for that action. Both of which are received by the agent.

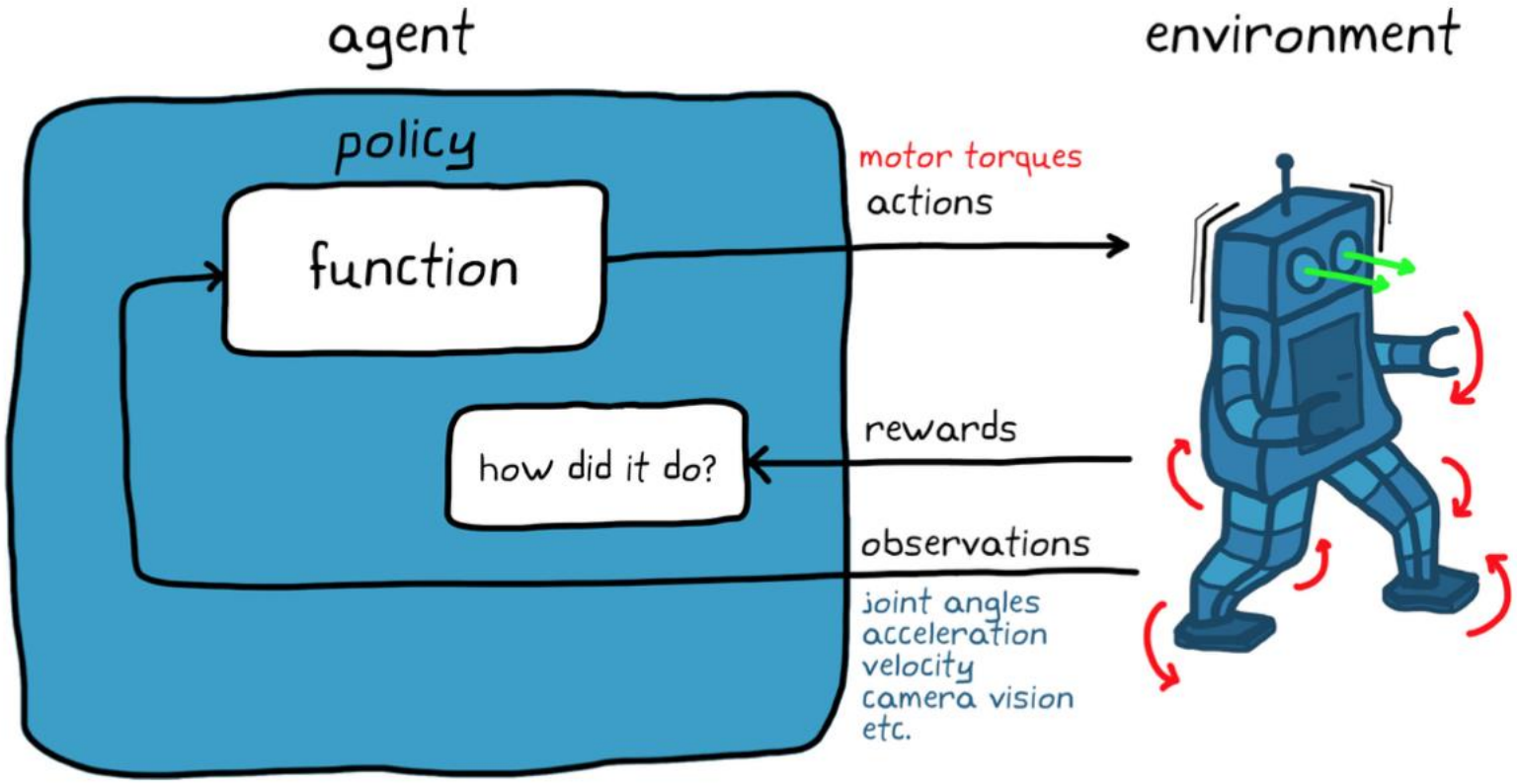


4 Using this new information, the agent can determine whether that action was good and should be repeated, or if it was bad and should be avoided.

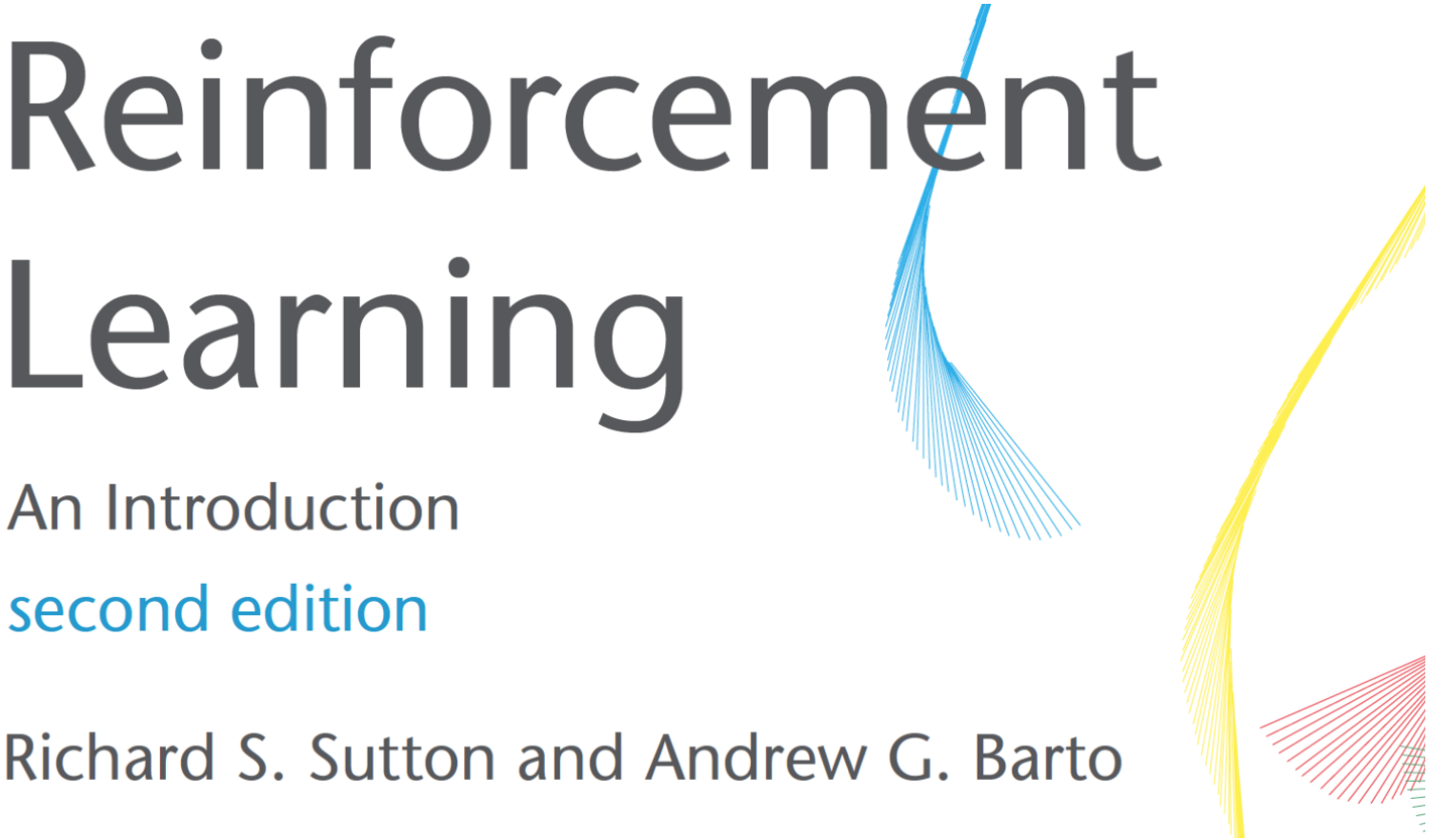
The observation-action-reward cycle continues until learning is complete.



Week 10 – Reinforcement Learning - Example



Week 10 – Reinforcement Learning - Book



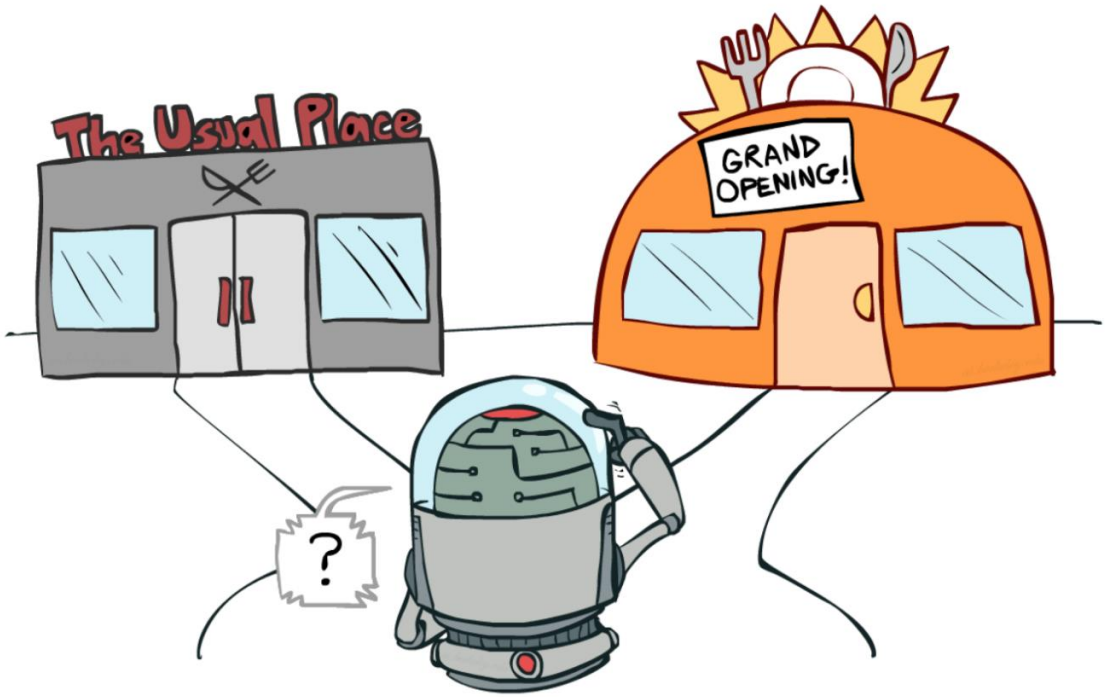
Week 10 – RL
Exploration vs. Exploitation Dilemma

University of Salford MANCHESTER

- Online decision-making involves a fundamental choice:
  - **Exploitation** Make the best decision given current information
  - **Exploration** Gather more information
- The best long-term strategy may involve short-term sacrifices
- Gather enough information to make the best overall decisions

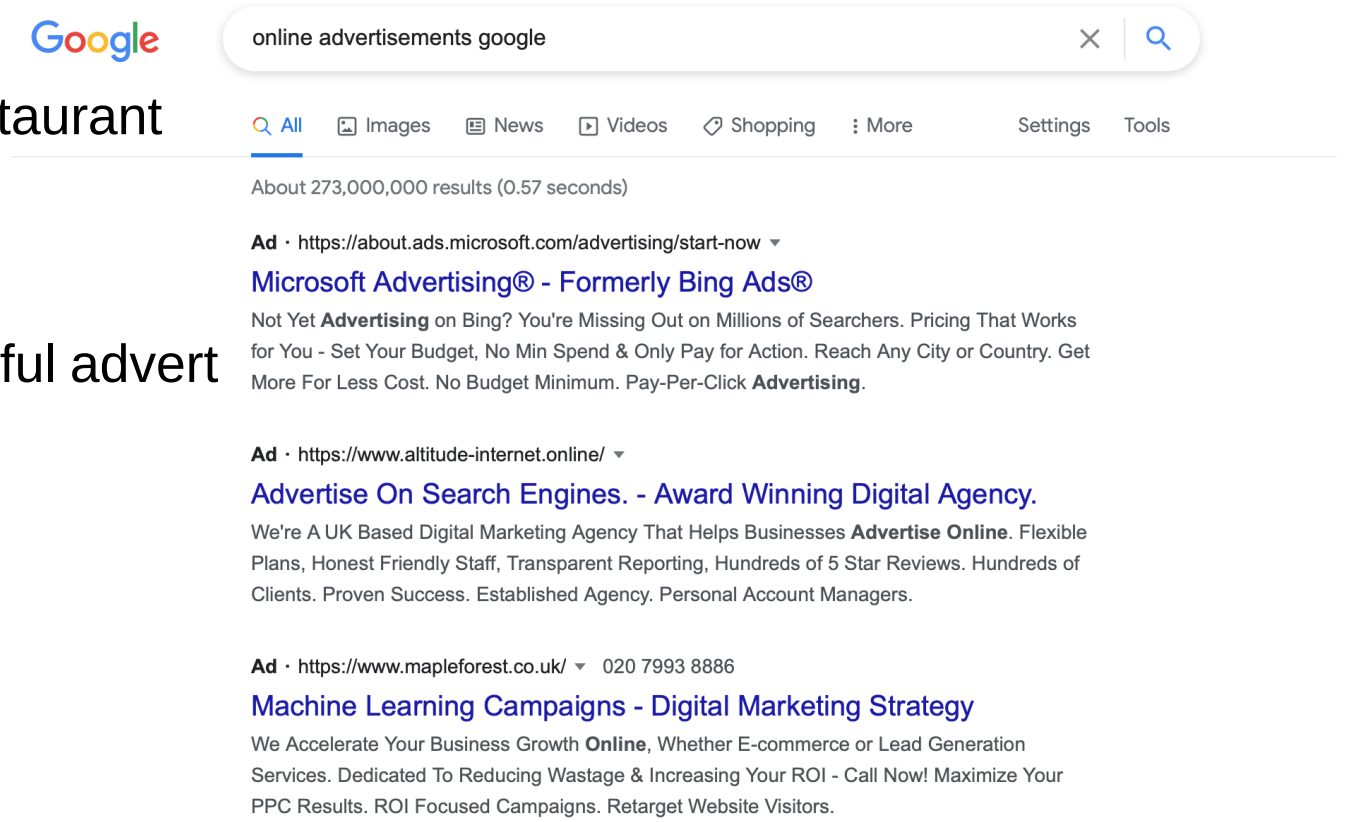
Week 10 – RL
Exploration vs. Exploitation Examples

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant



Week 10 – RL
Exploration vs. Exploitation Examples

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert



Week 10 – RL
Exploration vs. Exploitation Examples

University of Salford MANCHESTER

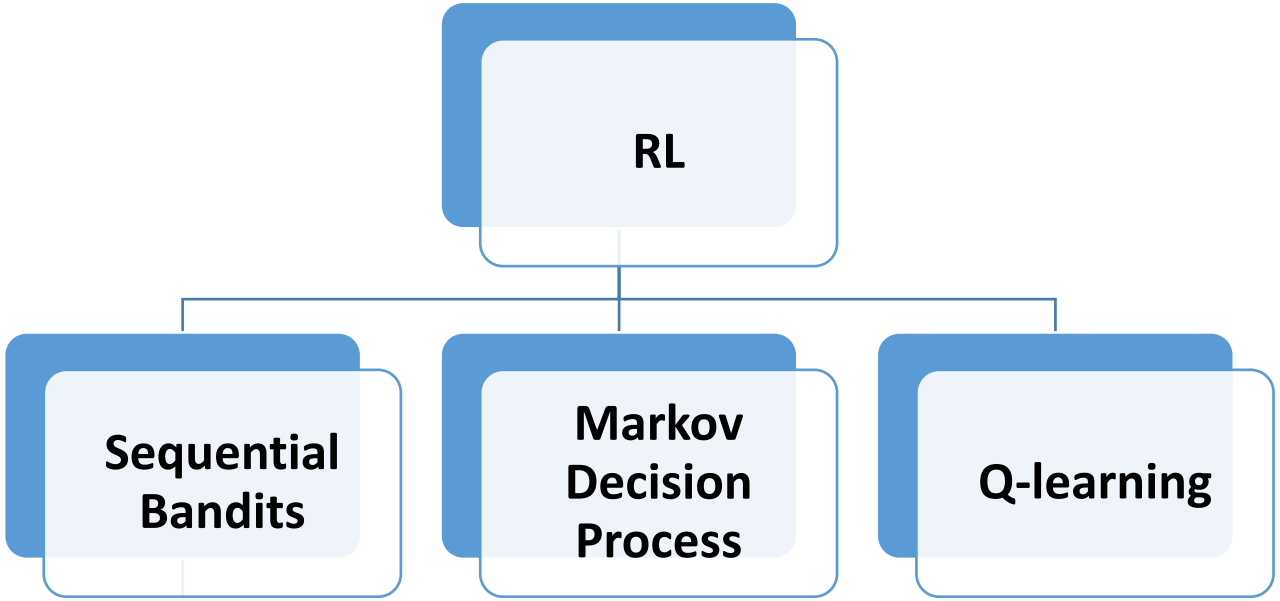
- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert
- Oil Drilling
  - Exploitation Drill at the best known location
  - Exploration Drill at a new location

Week 10 – RL
Exploration vs. Exploitation Examples

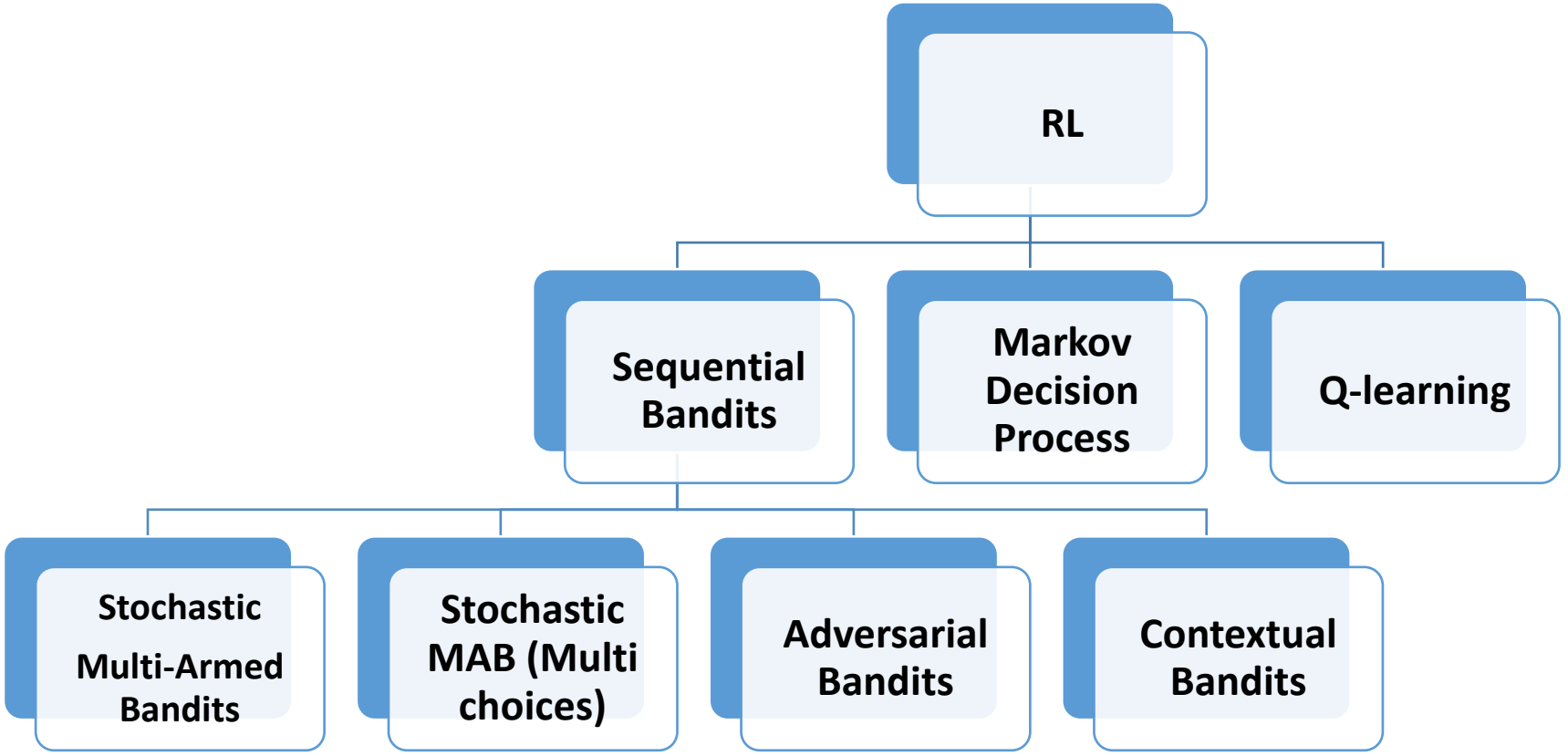
University of Salford MANCHESTER

- Restaurant Selection
  - Exploitation Go to your favourite restaurant
  - Exploration Try a new restaurant
- Online Banner Advertisements
  - Exploitation Show the most successful advert
  - Exploration Show a different advert
- Oil Drilling
  - Exploitation Drill at the best known location
  - Exploration Drill at a new location
- Game Playing
  - Exploitation Play the move you believe is best
  - Exploration Play an experimental move

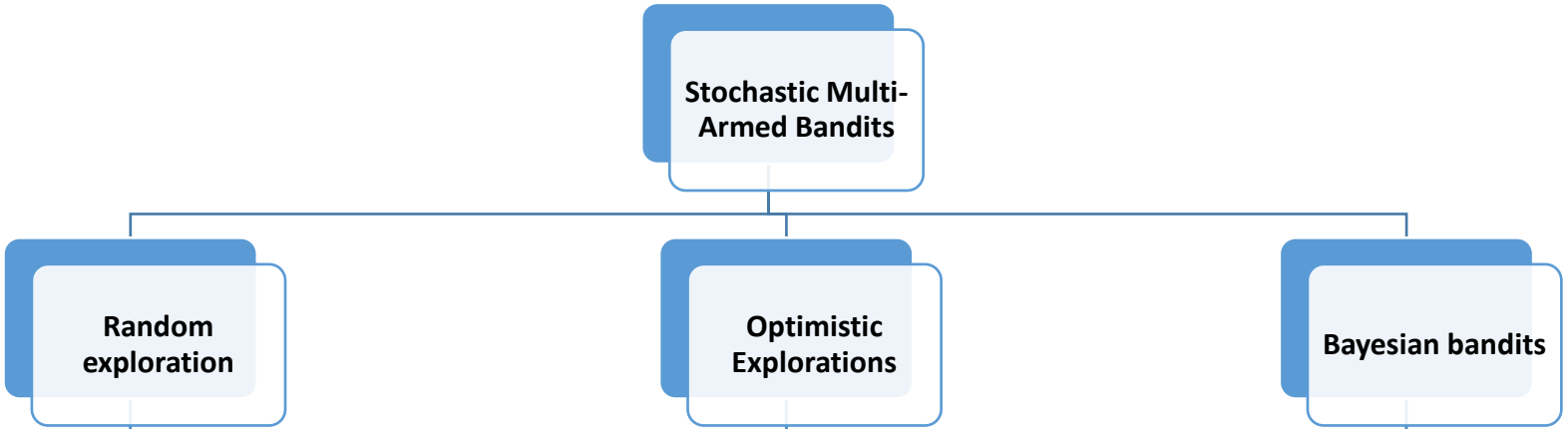
Week 10 – Reinforcement Learning



Week 10 – Reinforcement Learning



Week 10 –
RL - Multi-Armed Bandits



Week 10 –
RL - Multi-Armed Bandits

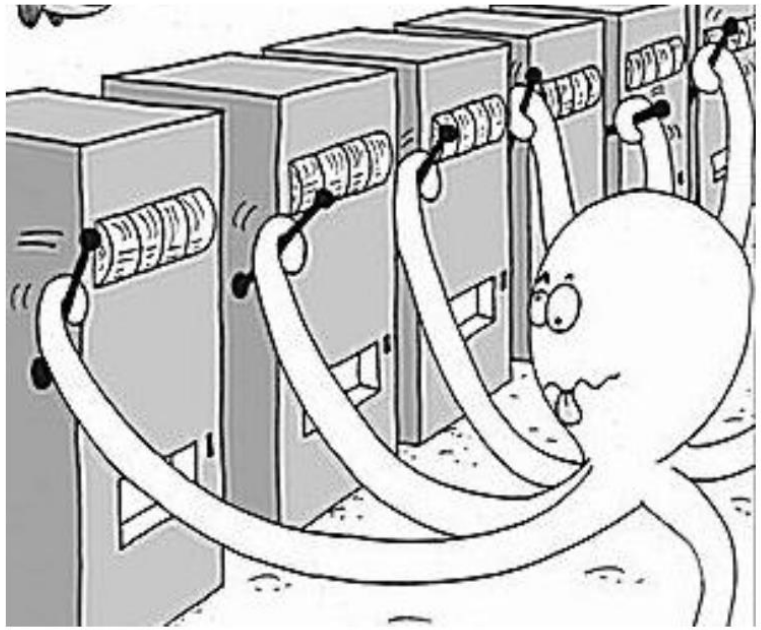
University of Salford MANCHESTER

One-armed bandit= Slot machine (English slang)

Week 10 –
RL - Multi-Armed Bandits

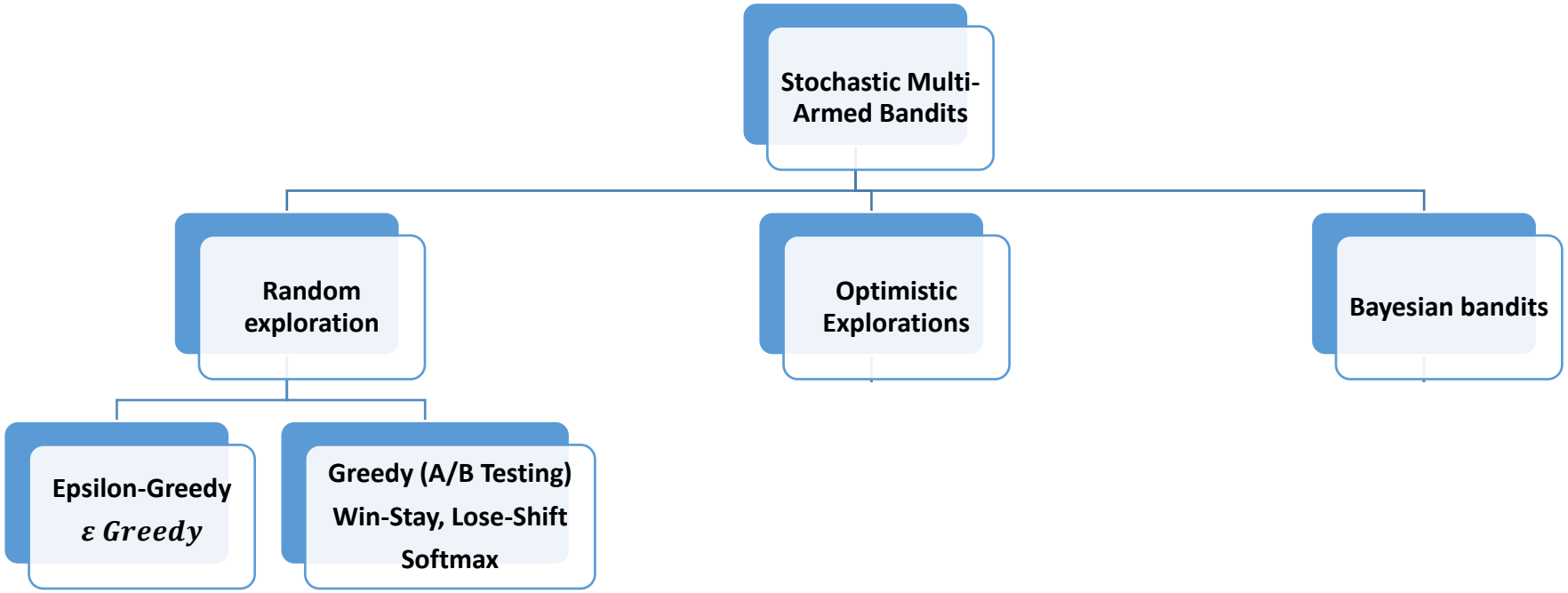
University of Salford MANCHESTER

- Multi-Armed bandit = Multiple Slot Machine

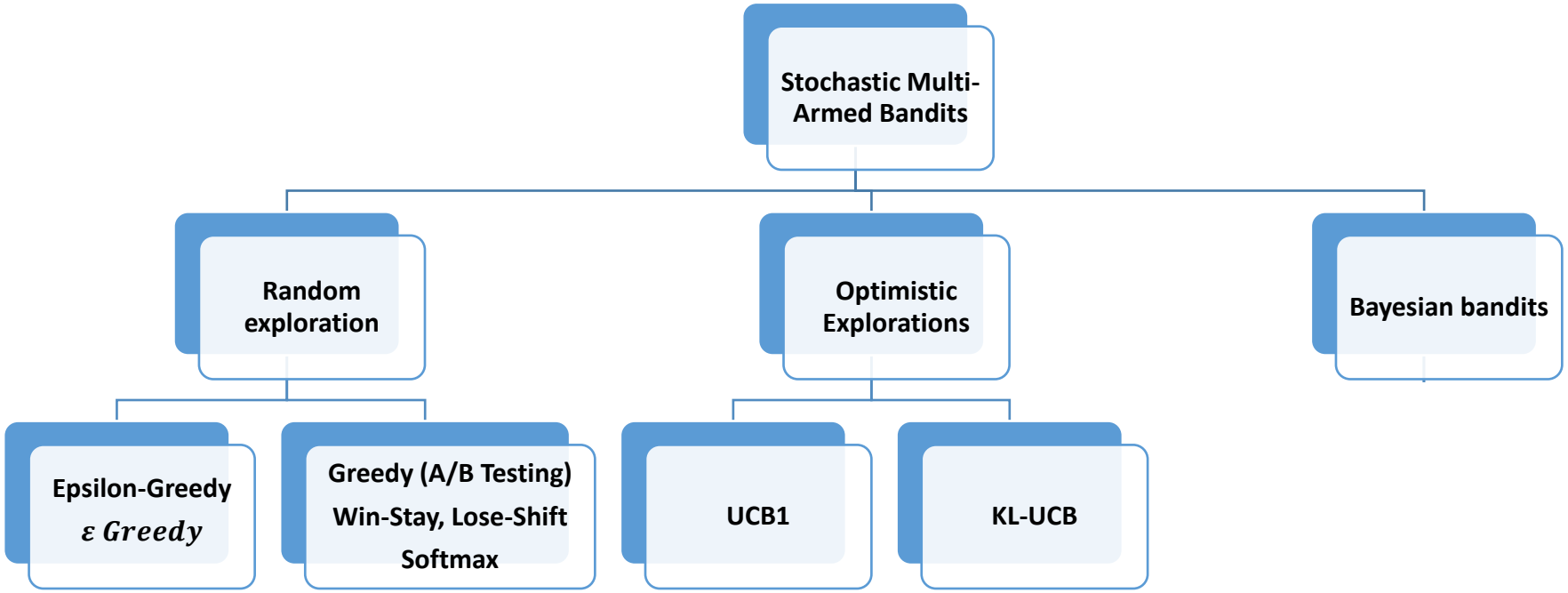


source: Microsoft Research

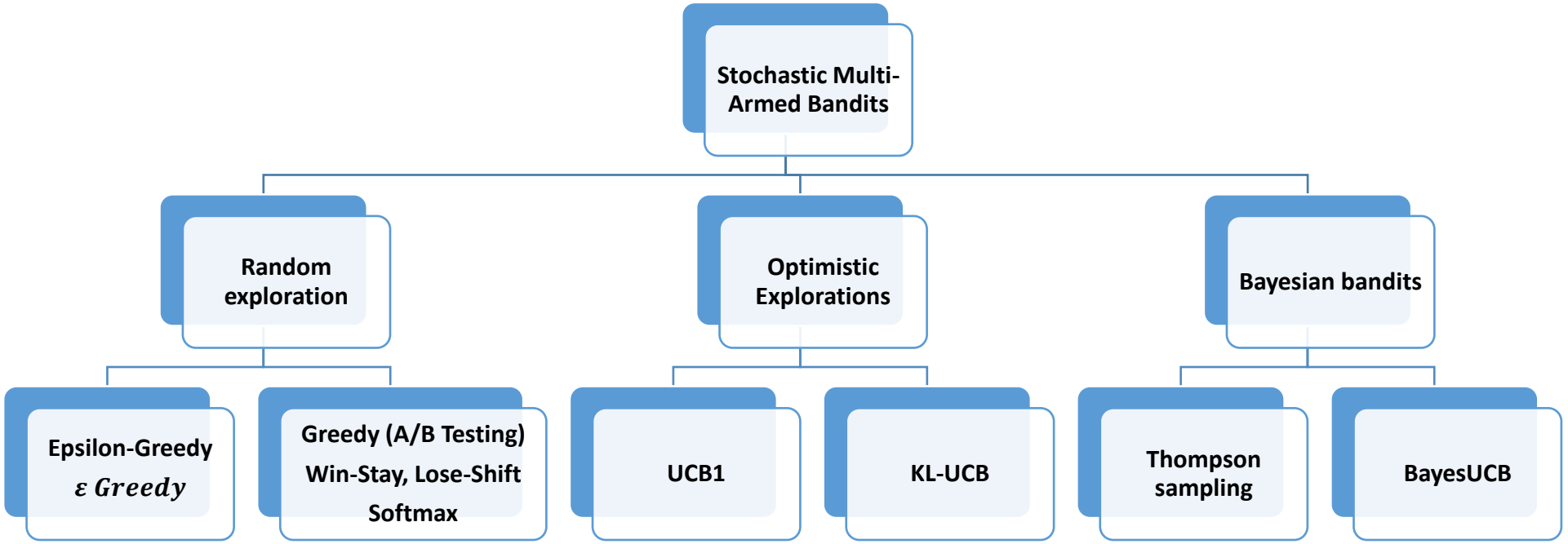
Week 10 –
RL - Multi-Armed Bandits



Week 10 –
RL - Multi-Armed Bandits



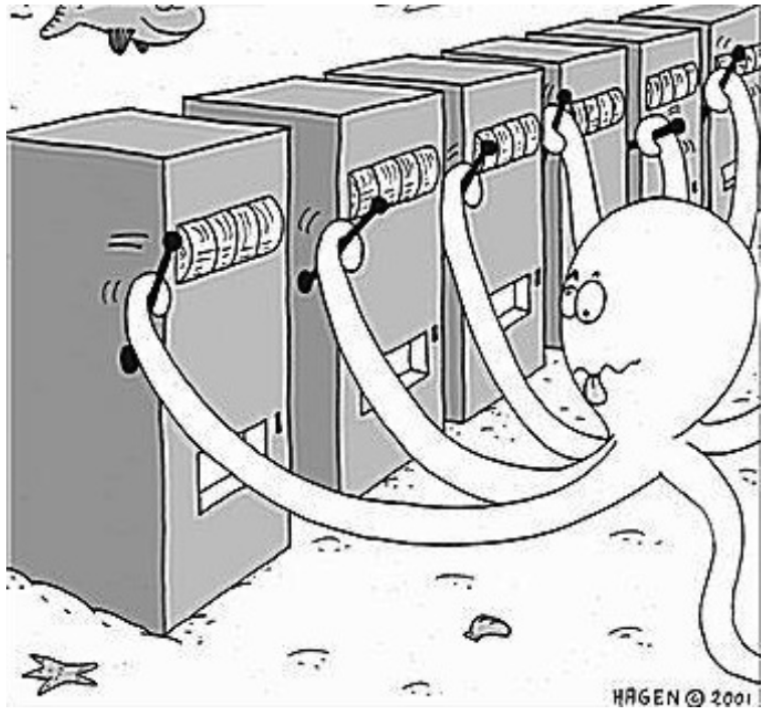
Week 10 –
RL - Multi-Armed Bandits



Week 10 – RL - Multi-Armed Bandits

University of Salford MANCHESTER

- A multi-armed bandit is a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$
- $\mathcal{A}$ is a known set of m actions (or "arms")
- $\mathcal{R}^a(r) = \mathbb{P}[r|a]$ is an unknown probability distribution over rewards
- At each step t the agent selects an action $a_t \in \mathcal{A}$
- The environment generates a reward $r_t \sim \mathcal{R}^{a_t}$
- The goal is to maximise cumulative reward $\sum_{\tau=1}^{t} r_\tau$



Week 10 –
RL - Multi-Armed Bandits

University of Salford MANCHESTER

Regret vs Reward

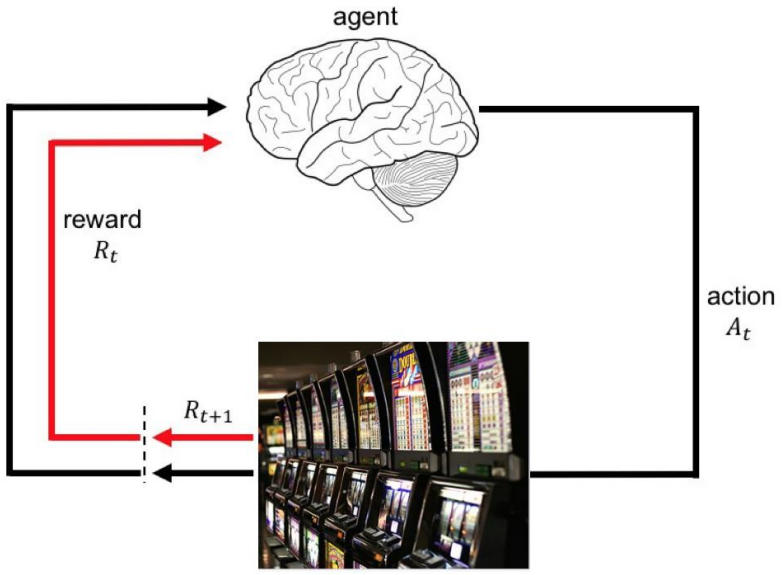
Week 10 –
RL – MAB - ε *Greedy*

University of Salford MANCHESTER

$$a_{t+1} = \begin{cases} argmax\,[\mu_t(1), \ldots \mu_t(k)] & \text{with prob } 1 - \{\epsilon\} \\ select\ randomly\ from \{1 \ldots k\} & \text{with prob } \{\epsilon\}, \end{cases}$$

where $\mu_t(i)$ denotes the average reward for arm i obtained over t rounds.

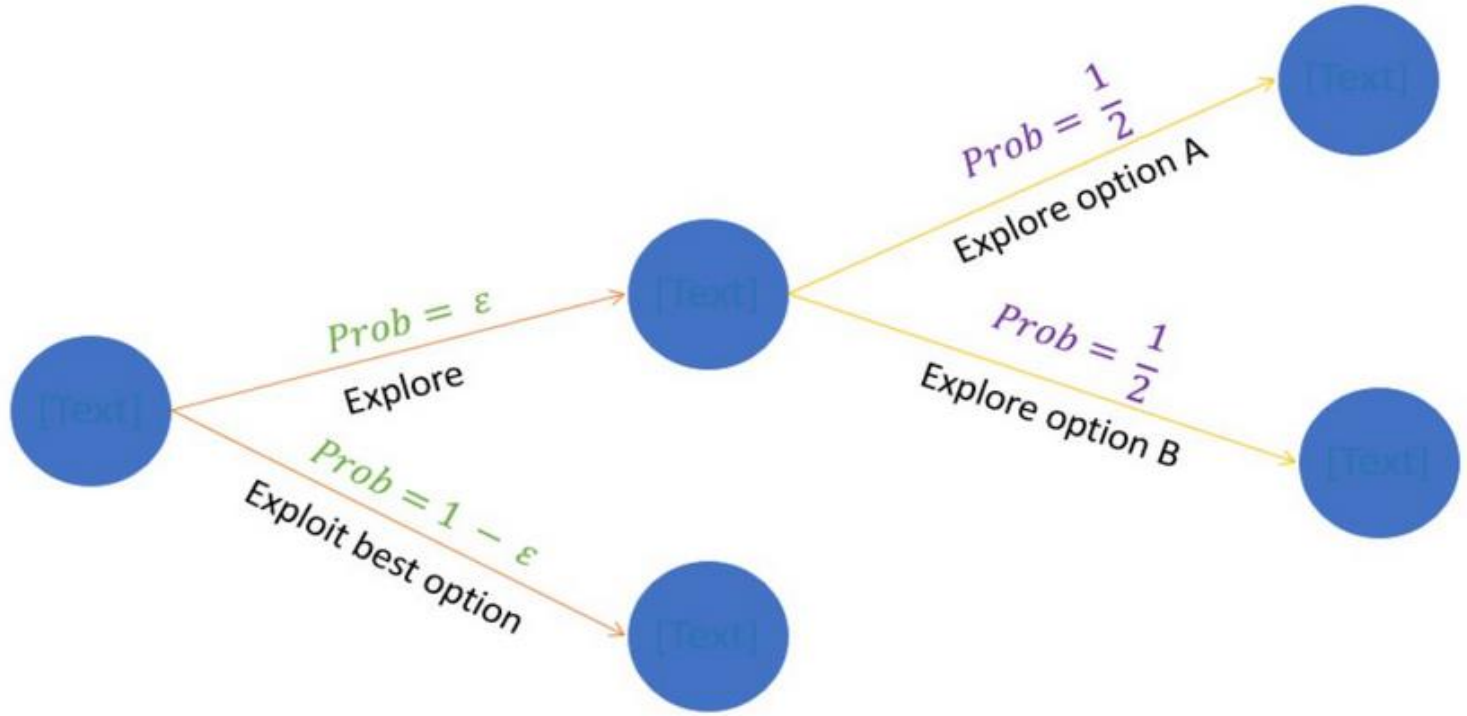
Week 10 – RL – MAB – ε *Greedy*



$$A_t, R_{t+1}, A_{t+1}, R_{t+2}, A_{t+2}, A_{t+3}, R_{t+3}, \ldots$$

The state does not change.

Week 10 – RL – MAB – ε *Greedy*



Week 10 –
RL – MAB - Decay – ε Greedy

$$a_{t+1} = \begin{cases} argmax\,[\mu_t(1), \ldots \mu_t(k)] & \text{with prob 1 - } \{\epsilon\} \\ select\,randomly\,from\{1 \ldots k\} & \text{with prob } \{\epsilon\}, \end{cases}$$

where $\mu_t(i)$ denotes the average reward for arm i obtained over t rounds.

$$\frac{1}{\log(t+\phi)},$$

where ϕ is a small number and t is the number of rounds to date.

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Man City 'make Grealish top target', Kane decides on future and more transfer rumours

Man City have reportedly made Jack Grealish their number one target and more transfer rumours.

SHARE 3 COMMENTS By Steven Railston 19:00, 11 APR 2021

Jack Grealish has been linked with a move to the Etihad.

Advertisement

MOST READ

Manchester Ur
ratings: Paul P
Edinson Cavan
Tottenham

Jack Grealish to Man Utd transfer: Solskjaer's private conversation and the Man City problem

The Aston Villa star continues to be linked with a move to United this summer and there have been a few clues drc recently as to where his future lies

SHARE COMMENTS By Jamie Kemble 06:30, 11 APR 2021

Advertisement

MOST READ

Manchester Unit
ratings: Paul Pog
Edinson Cavani
Tottenham

Jack Grealish revelation gives Man City clear trai advantage over Man United

Aston Villa's talismanic skipper has been the subject of much speculation about his future and has been move to Manchester this summer

SHARE 1 COMMENTS By Leigh Curtis 19:30, 8 APR 2021

Advertisement

MOS

Manc
rating
Edins
Tottenham

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Man City 'make Grealish top target', Kane decides on future and more transfer rumours

Man City have reportedly made Jack Grealish their number one target and more transfer rumours.

SHARE 3 COMMENTS By Steven Railston 19:00, 11 APR 2021

Jack Grealish has been linked with a move to the Etihad.

Advertisement

MOST READ

Manchester Ur
ratings: Paul P
Edinson Cavan
Tottenham

Arm 0

Jack Grealish to Man Utd transfer: Solskjaer's private conversation and the Man City problem

The Aston Villa star continues to be linked with a move to United this summer and there have been a few clues drc recently as to where his future lies

SHARE COMMENTS By Jamie Kemble 06:30, 11 APR 2021

Advertisement

MOST READ

Manchester Unit
ratings: Paul Pog

Arm 1

Jack Grealish revelation gives Man City clear trai advantage over Man United

Aston Villa's talismanic skipper has been the subject of much speculation about his future and has been move to Manchester this summer

SHARE 1 COMMENTS By Leigh Curtis 19:30, 8 APR 2021

Advertisement

MOS

Manc
rating
Edins

Arm 2

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Arm 0

Arm 0
0

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

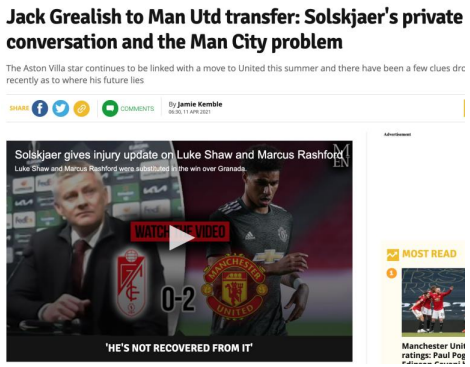
Arm 0

Arm 0
0
1

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Arm 1



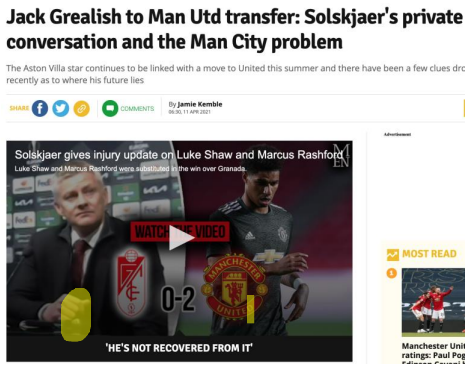
Arm 1
0

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Arm 0

Arm 1



Arm 2

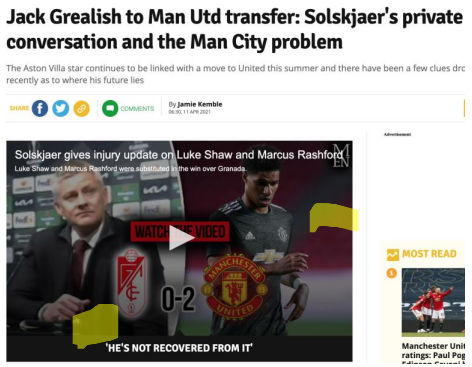
Arm 0	Arm 1	Arm 2
0	0	0
1	0	0
0	0	0
0	0	0
0	1	0

Week 10 –
RL – MAB - ε *Greedy* – *Example*

University of Salford MANCHESTER

Arm 0

Arm 1



Arm 2

Users = 200

	Brute force	Random		ε *Greedy*	
Arm 0	28	0	71	0	161
Arm 1	19	1	71	2	24
Arm 2	13	2	58	1	15

Week 10 – RL – MAB- Optimism in the Face of Uncertainty
UCB1 - Upper Confidence Bounds

University of Salford MANCHESTER

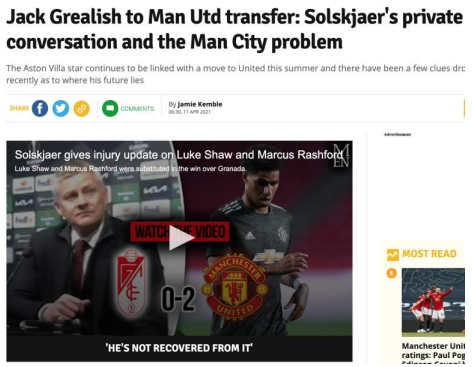
$$a_{t+1} = \underset{i\epsilon\{1\ldots K\}}{\operatorname{argmax}}\left(\mu_i + \sqrt{\frac{2\log t}{n_i}}\right),$$

where n_i is the number of times arm i has been chosen and t is the total number of rounds.

Week 10 – RL – MAB-UCB1

Arm 0

Arm 1



Arm 2

Users = 200

	Brute force	Random	ε *Greedy*	$UCB1$
Arm 0	28	71	161	169
Arm 1	19	71	24	19
Arm 2	13	58	15	12

Week 10 – RL – MAB- Bayesian Bandits -Thompson Sampling

University of Salford MANCHESTER

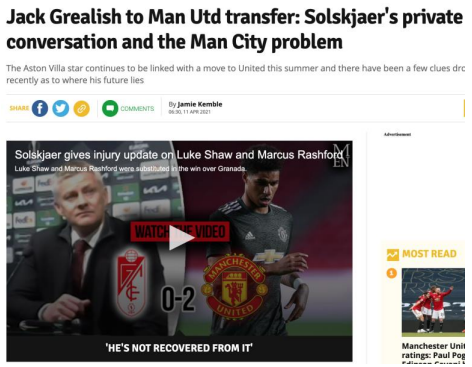
$$P(x) = \frac{(1-x)^{\beta-1} x^{\alpha-1}}{B(\alpha, \beta)},$$

where α is set to $s_a + 1$, β is set to $f_a + 1$ and $B(\alpha, \beta)$ is a normalizing constant.

Week 10 – RL – MAB- Thompson Sampling

Arm 0

Arm 1



Arm 2

Users = 200

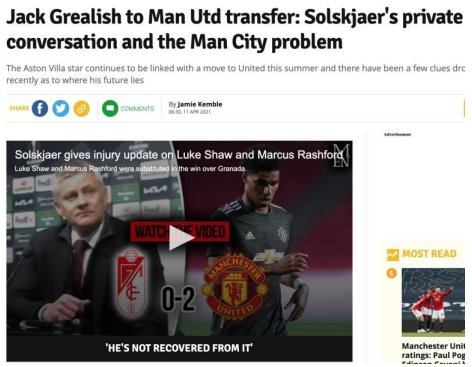
Brute force		Random		ε *Greedy*		$UCB1$		*Thompson Sampling*	
Arm 0	28	0	71	0	161	0	169	0	166
Arm 1	19	1	71	2	24	1	19	2	21
Arm 2	13	2	58	1	15	2	12	1	13

Week 10 – RL – MAB- Thompson Sampling

University of Salford MANCHESTER

Arm 0

Arm 1



Arm 2

Users = 200

Brute force

Arm 0	28
Arm 1	19
Arm 2	13

Random

0	71
1	71
2	58

ε Greedy

0	161
2	24
1	15

$UCB1$

0	169
1	19
2	12

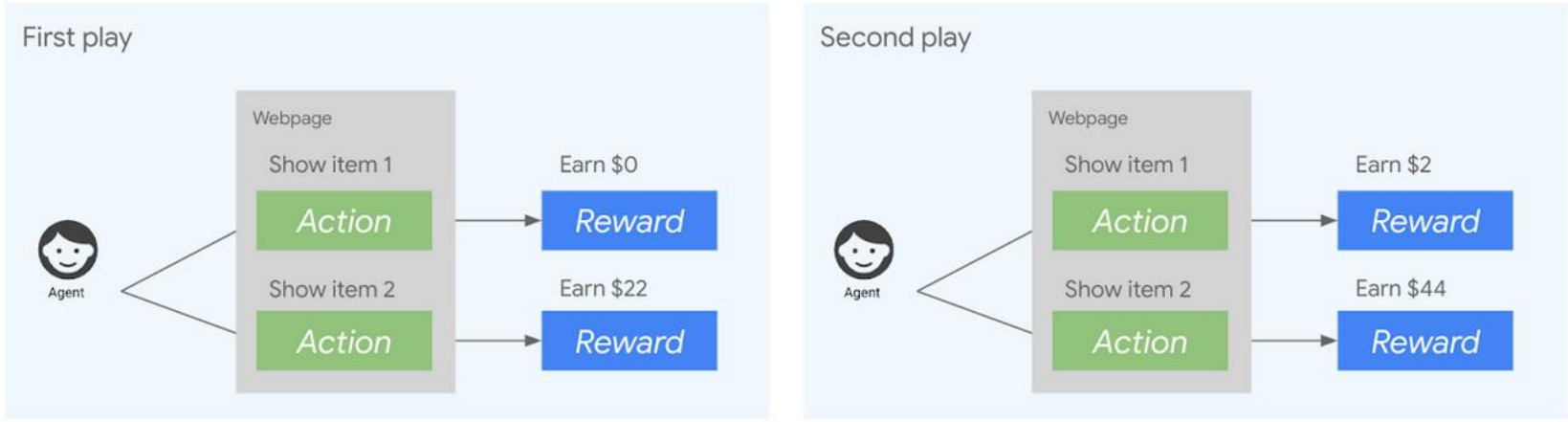
Thompson Sampling

0	166
2	21
1	13

Users = 9619

Arm 0	1406
Arm 1	1000
Arm 2	541

Week 10 – RL – MAB- *Multiple Arms*



No state: Every play (or episode) is independent of each other and rewards received are only related to the action executed, so the agent learns the action that most often yields the best reward.

From Google

Week 10 – RL – MAB - *Multiple Arms*

University of Salford MANCHESTER

Arm 0

Arm 2

Arm 0	Arm 2
0	0

Week 10 – RL – MAB - *Multiple Arms*

University of Salford MANCHESTER

Arm 0

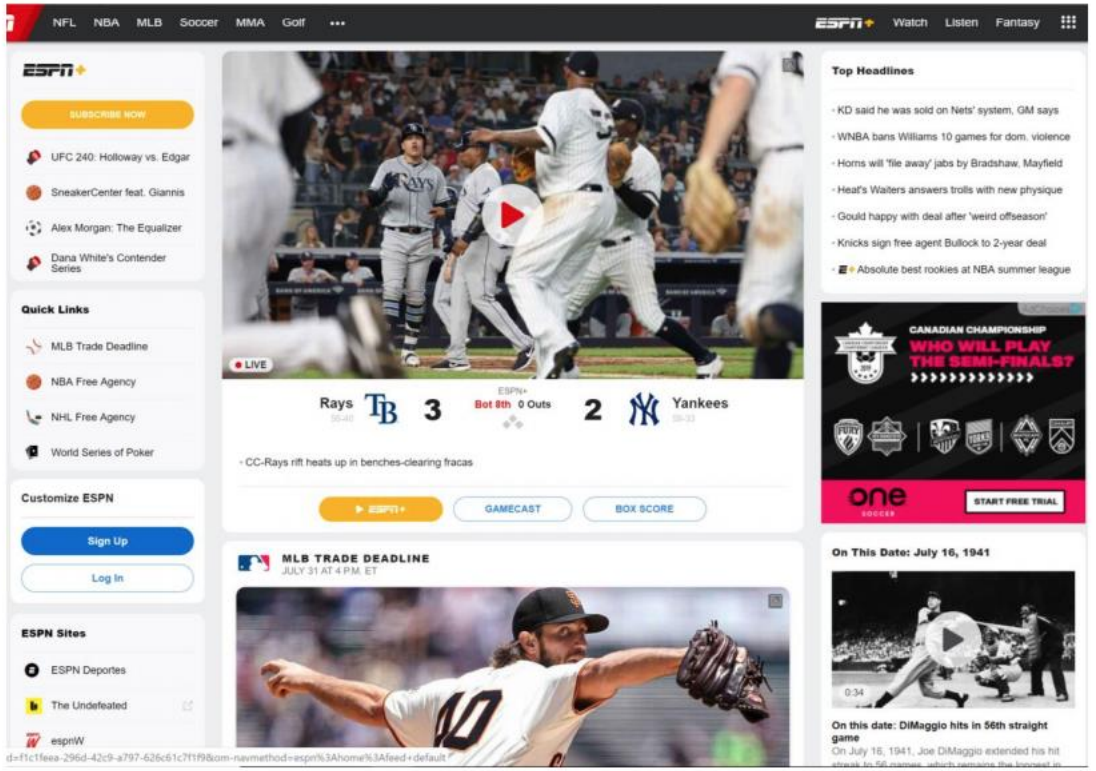
Arm 2

Arm 0	Arm 2
1	0

Week 10 – RL – MAB- Contextual Bandits (Associative Search)

University of Salford MANCHESTER

- Usually, there is some context that can help you make a decision.
- For example:
  - Patient data for clinical trials.
  - Consumer data for news/movie recommendation.
- Today, we will see how we can use this “context” to guide our decision making process.



News that generates most clicks should be front and center. However, the “click through rate” depends on the user!

Week 10 – RL – MAB- Contextual Bandits

University of Salford MANCHESTER

Consider the following example:

- We are running a sports news website. Today, there are K big sports related news stories.
- Every time a user visits our set, we must decide then and there which headlines to display to her on the front page.
- The goal is to maximize the number of clicks.

Bob just visited our website!

1. Bob loves the Toronto Raptors.
2. Hates the Los Angeles Clippers.

Which headline should we show to Bob?

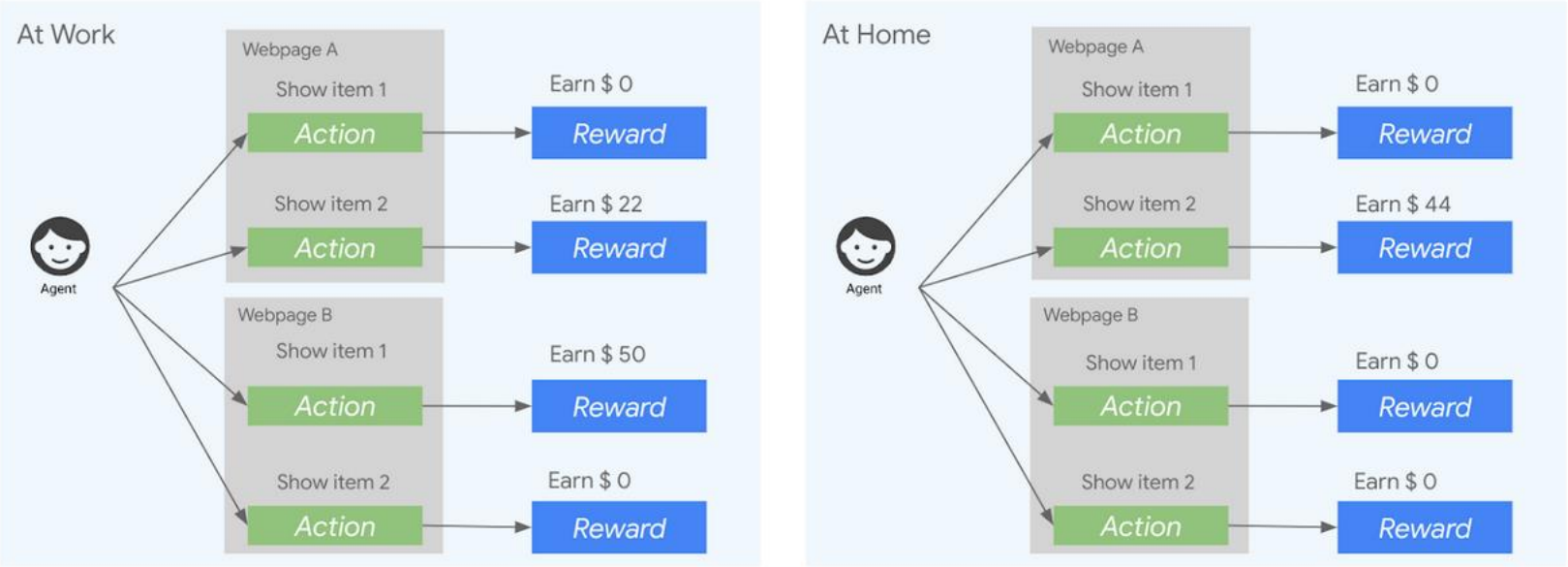
Context

Arms

1.

2.

Week 10 – RL – MAB- Contextual Bandits



Week 10 – RL – MAB- Contextual Bandits – Algorithms

- Online linear bandits
  - LinUCB *(Upper Confidence Bound)* algorithm
  - LinRel (Linear Associative Reinforcement Learning) algorithm
  - HLINUCBC (Historic LINUCB with clusters)
- Online non-linear bandits
  - UCBogram algorithm
  - Generalized linear algorithms
  - NeuralBandit algorithm
  - KernelUCB algorithm
  - Bandit Forest algorithm
  - Oracle-based algorithm
- Constrained contextual bandit
  - Context Attentive Bandits or Contextual Bandit with Restricted
  - UCB-ALP algorithm
- The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits
- https://arxiv.org/pdf/1802.04064.pdf

Week 10 – RL – MAB- Contextual Bandits – In practical

University of Salford MANCHESTER

- Microsoft
  - Contextual Bandits

    Reinforcement Learning with

    Vowpal Wabbit (Free)
    - https://vowpalwabbit.org/tutorials/contextual_bandits.html#algorithms-and-format
  - Personalizer (Not Free)
    - https://azure.microsoft.com/en-us/services/cognitive-
- Google
  - AutoML for Contextual Bandits (Not Free)
    - https://research.google/pubs/pub48534/
- Amazon
  - Amazon Personalize Create real-time personalized user experiences faster at scale (Not Free)
    - https://aws.amazon.com/personalize/

Week 10 – RL – MAB- Contextual Bandits – In practical

University of Salford MANCHESTER

- Microsoft
  - Contextual Bandits Reinforcement Learning with Vowpal

2019 Inaugural ACM SIGAI Industry Award for Excellence in Artificial Intelligence

acm SIGAI

AWARD

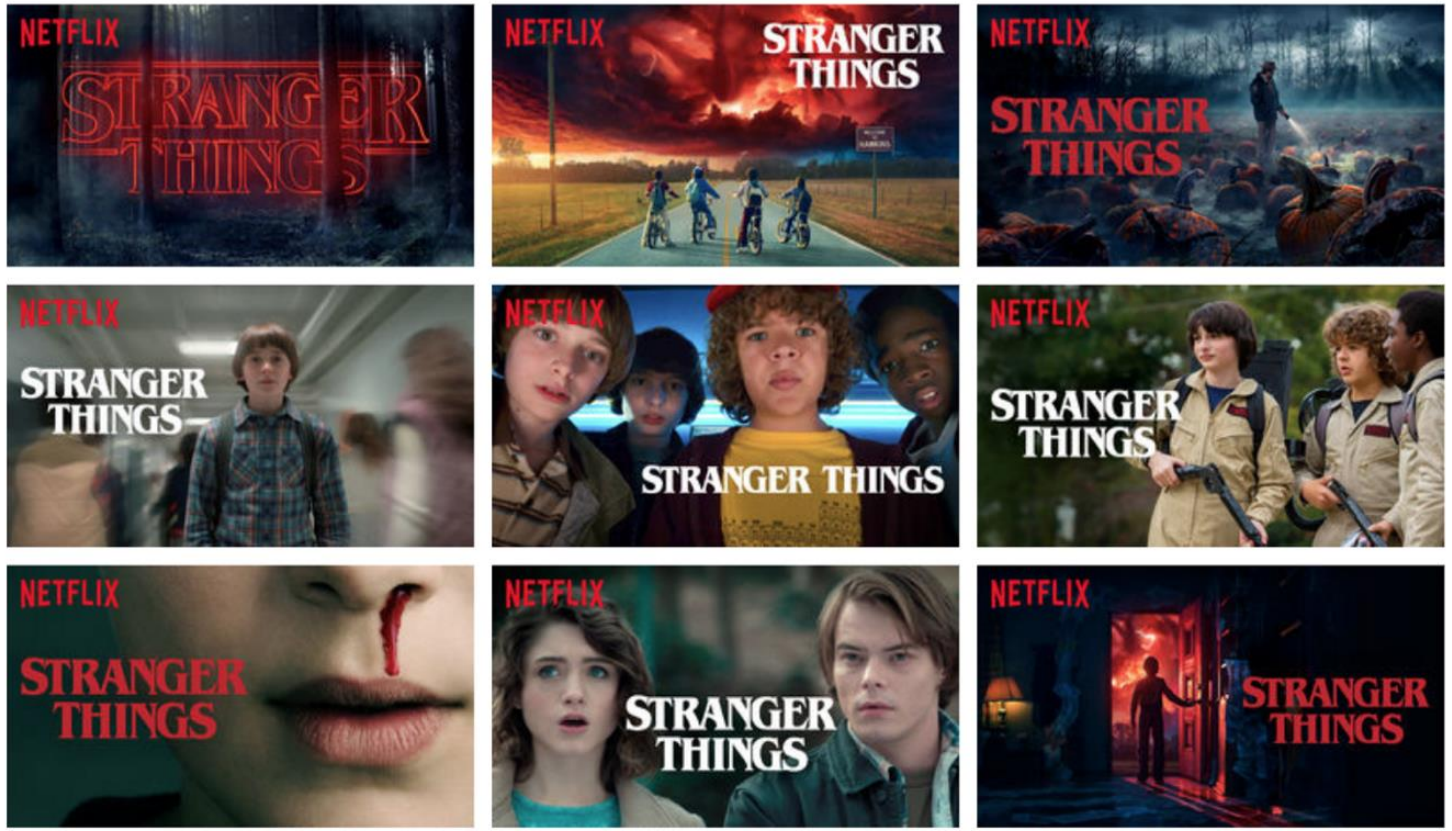
2019 Inaugural ACM SIGAI Industry Award for Excellence in Artificial Intelligence

Winners of the Strata Data Awards 2019

AWARD

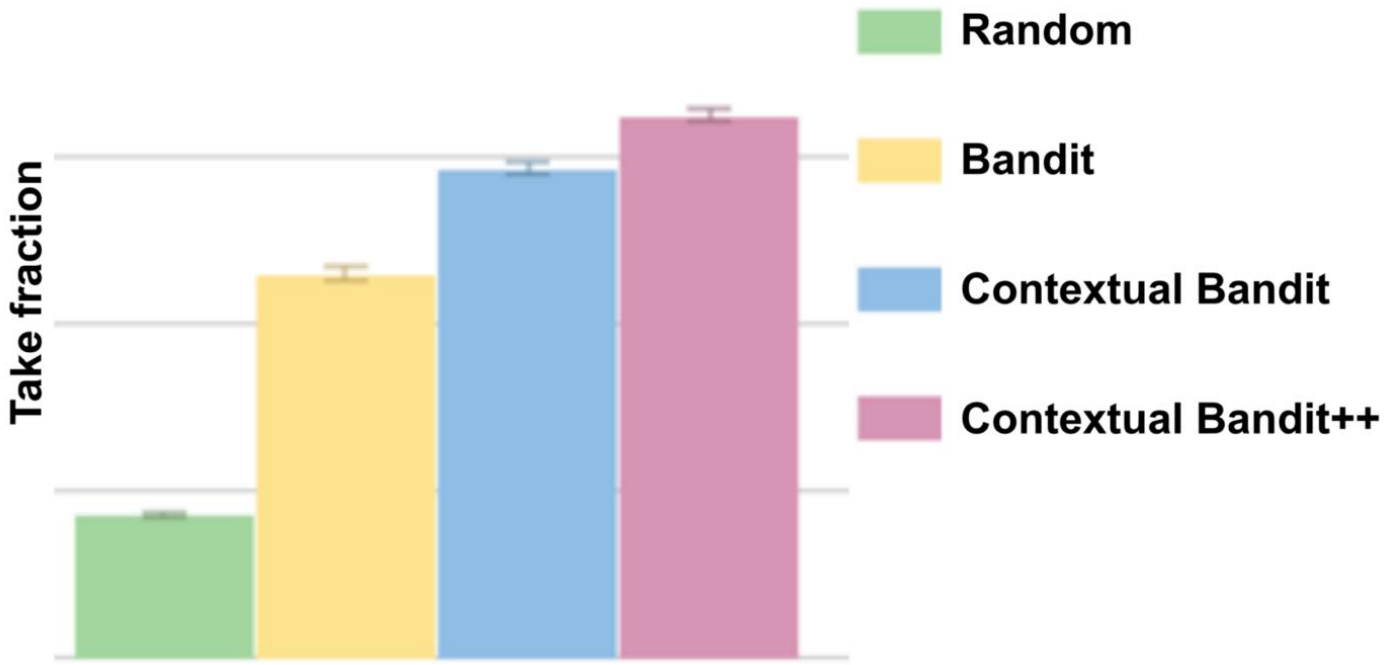
2019 O'Reilly Strata Data Conference Award for Most Innovative Product

Week 10 – RL – MAB- Contextual Bandits – Real World



https://netflixtechblog.com/artwork-personalization-c589f074ad76

Week 10 – RL – MAB- Contextual Bandits – Real World



https://netflixtechblog.com/artwork-personalization-c589f074ad76

Week 10 – RL – MAB- Contextual Bandits – Coding -

- vowpalwabbit

Algorithms and format

There are four main components to a contextual bandit problem:

- **Context (x)**: the additional information which helps in choosing action.
- **Action (a)**: the action chosen from a set of possible actions **A**.
- **Probability (p)**: the probability of choosing **a** from **A**.
- **Cost/Reward (r)**: the reward received for action **a**.

Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

For example:

APP news website:

- **Decision to optimize**: articles to display to user.
- **Context**: user data (browsing history, location, device, time of day)
- **Actions**: available news articles
- **Reward**: user engagement (click or no click)

APP cloud controller:

- **Decision to optimize**: the wait time before reboot of unresponsive machine.
- **Context**: the machine hardware specs (SKU, OS, failure history, location, load).
- **Actions**: time in minutes - {1 ,2 , ...N}
- **Reward**: negative of the total downtime

Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

action	cost	probability	Country	feature2	Gender
1	2	0.4	UK	c	
3	1	0.2	USA	d	
4	1	0.5	UK	b	
2	2	0.3	UK	b	Male
3	1	0.7	UK	d	

Week 10 – RL – MAB- Contextual Bandits – Coding -

- vowpalwabbit

action	cost	probability	Country	feature2	Gender
1	2	0.4	UK	c	
3	1	0.2	USA	d	
4	1	0.5	UK	b	
2	2	0.3	UK	b	Male
3	1	0.7	UK	d	

Arm 0

Week 10 – RL – MAB- Contextual Bandits – Coding -

- vowpalwabbit

action	cost	probability	Country	feature2	Gender
1	2	0.4	UK	c	
3	1	0.2	USA	d	
4	1	0.5	UK	b	
2	2	0.3	UK	b	Male
3	1	0.7	UK	d	

Action 1

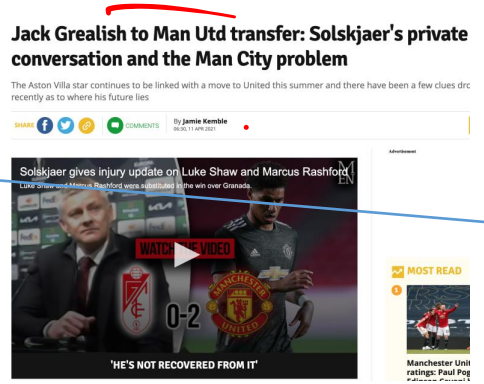
Week 10 – RL – MAB- Contextual Bandits – Coding -

- vowpalwabbit

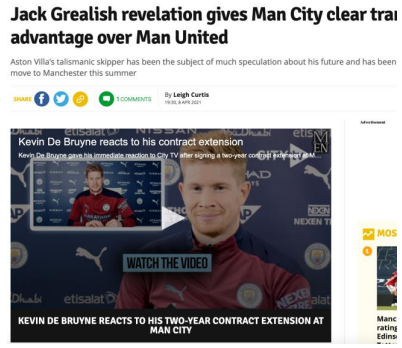
action	cost	probability	Country	feature2	Gender
1	2	0.4	UK	c	
3	1	0.2	USA	d	
4	1	0.5	UK	b	
2	2	0.3	UK	b	Male
3	1	0.7	UK	d	

Action 1

Action 2



Action 3



Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

Each example is represented as a separate line in your data file and must follow the following format:

```
action:cost:probability | features
```

Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

Each example is represented as a separate line in your data file and must follow the following format:

```
action:cost:probability | features
```

action	cost	probability	Country	feature2	Gender
1	2	0.4	UK	c	
3	1	0.2	USA	d	
4	1	0.5	UK	b	
2	2	0.3	UK	b	Male
3	1	0.7	UK	d	

Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

Each example is represented as a separate line in your data file and must follow the following format:

```
action:cost:probability | features
```

```
# Define the parameters

for i in train_df.index:
  action = train_df.loc[i, "action"]
  cost = train_df.loc[i, "cost"]
  probability = train_df.loc[i, "probability"]
  Country = train_df.loc[i, "Country"]
  feature2 = train_df.loc[i, "feature2"]
  Gender = train_df.loc[i, "Gender"]

  # Construct the example in the required vw format.
  learn_example = str(action) + ":" + str(cost) + ":" + str(probability) + " | " + str(feature1) + " " + str(feature2) + " " + str(feature3)

  # Here we do the actual learning.
  vw.learn(learn_example)
```

Week 10 – RL – MAB- Contextual Bandits – Coding -

University of Salford MANCHESTER

- vowpalwabbit

```
from vowpalwabbit import pyvw
#  --cb 4  ==> Number of action
vw = pyvw.vw("--cb 4")
```

```
# Define the parameters

for i in train_df.index:
  action = train_df.loc[i, "action"]
  cost = train_df.loc[i, "cost"]
  probability = train_df.loc[i, "probability"]
  Country = train_df.loc[i, "Country"]
  feature2 = train_df.loc[i, "feature2"]
  Gender = train_df.loc[i, "Gender"]

  # Construct the example in the required vw format.
  learn_example = str(action) + ":" + str(cost) + ":" + str(probability) + " | " + str(feature1) + " " + str(feature2) + " " + str(feature3)

  # Here we do the actual learning.
  vw.learn(learn_example)
```

Week 10 – RL – MAB- Contextual Bandits – Coding - test

- vowpalwabbit

Country	feature2	Gender
UK	c	
UK		female
USA	b	
UK		female

Week 10 – RL – MAB- Contextual Bandits – Coding - test

University of Salford MANCHESTER

- vowpalwabbit

```
print("index Country feature2 Gender ==> predict the action")
for j in test_df.index:
  Country = test_df.loc[j, "Country"]
  feature2 = test_df.loc[j, "feature2"]
  Gender = test_df.loc[j, "Gender"]

  test_example = "| " + str(feature1) + " " + str(feature2) + " " + str(feature3)

  choice = vw.predict(test_example)
  print(j,"\t",feature1,"\t",feature2,"\t",feature3,"\t\t", choice)
```

Week 10 – RL – MAB- Contextual Bandits – Coding - test

University of Salford MANCHESTER

- vowpalwabbit

```
print("index Country feature2 Gender ==> predict the action")
for j in test_df.index:
  Country = test_df.loc[j, "Country"]
  feature2 = test_df.loc[j, "feature2"]
  Gender = test_df.loc[j, "Gender"]

  test_example = "| " + str(feature1) + " " + str(feature2) + " " + str(feature3)

  choice = vw.predict(test_example)
  print(j,"\t",feature1,"\t",feature2,"\t",feature3,"\t\t", choice)
```

```
Country feature2 Gender ==> predict the action
   a       c       b               4
   a               b               4
   a       b       b               3
   a               b               4
```

Week 10 – RL – MAB- Contextual Bandits – Coding - Save

University of Salford MANCHESTER

- vowpalwabbit

```
vw.save('cb.model')
```

Week 10 – RL – MAB- Contextual Bandits – Coding - Save

University of Salford MANCHESTER

- vowpalwabbit

```
vw.save('cb.model')
del vw
```

Week 10 – RL – MAB- Contextual Bandits – Coding - Load

University of Salford MANCHESTER

- vowpalwabbit

```
vw.save('cb.model')
del vw

vw = pyvw.vw("--cb 4 -i cb.model")
print("Show action number ", vw.predict('| UK b'))
```

```
action number  4
```

Week 10 – RL – MAB- Contextual Bandits – Coding - Predict

University of Salford MANCHESTER

- vowpalwabbit

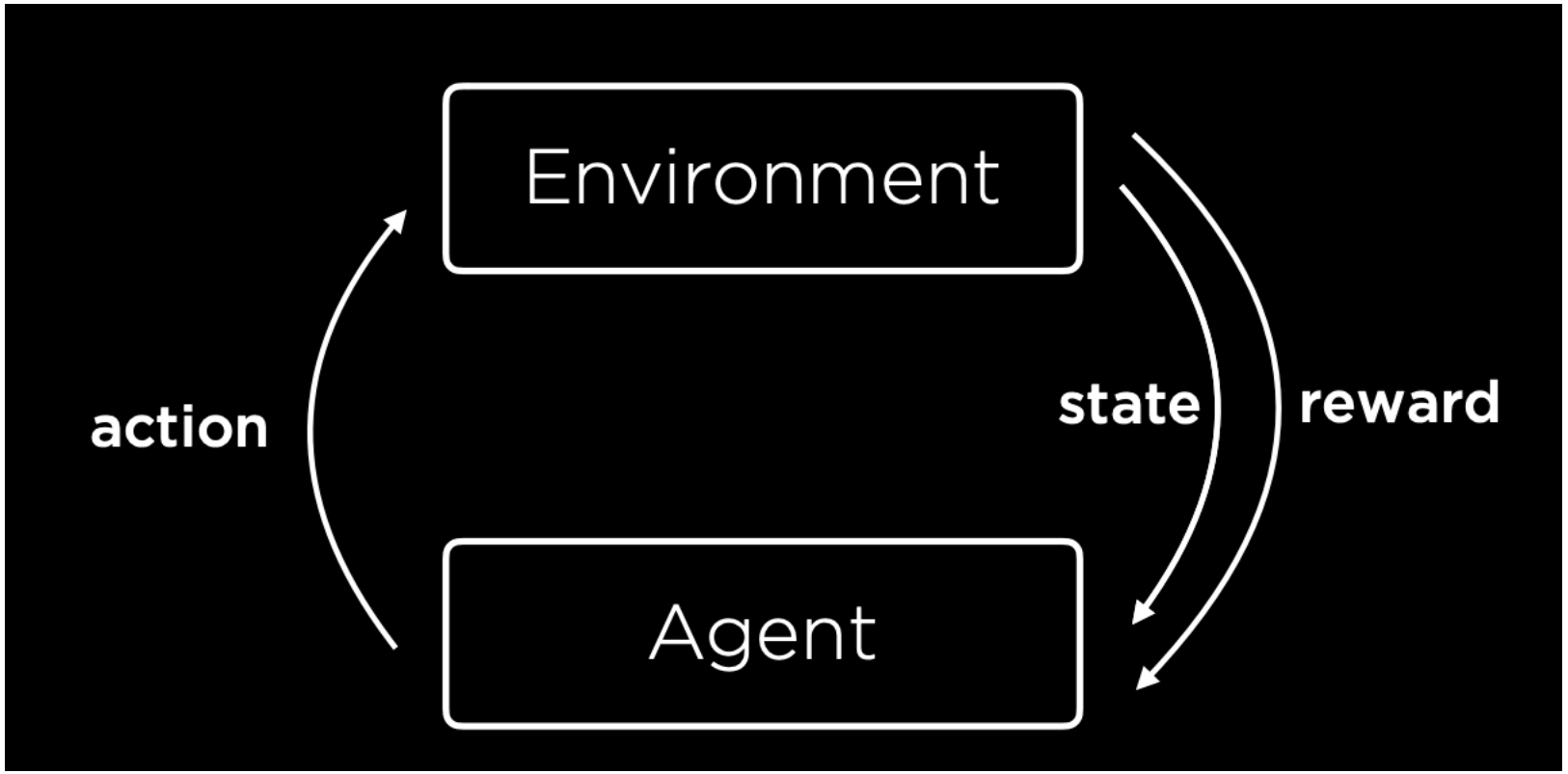
```
vw.save('cb.model')
del vw

vw = pyvw.vw("--cb 4 -i cb.model")
print("Show action number ", vw.predict('| UK b'))
```

```
action number  4
```

```
print("Show action number ",vw.predict('| '))
```

```
action number  4
```

Week 10 – RL



Week 10 – RL
Markov Decision Process

University of Salford MANCHESTER

- model for decision-making,

  representing states, actions, and

  their rewards

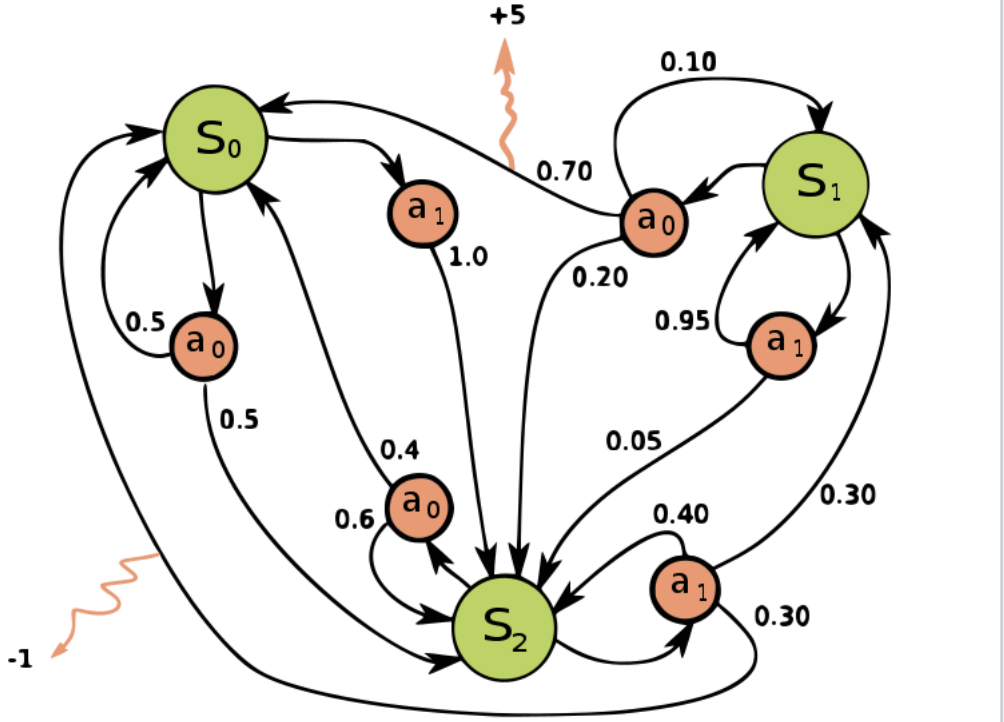
Week 10 – RL
Markov Decision Process

University of Salford MANCHESTER

- model for decision-making, representing states, actions, and their rewards

- Set of states S
- Set of actions ACTIONS(s)
- Transition model P(s' | s, a)
- Reward function R(s, a, s')

Week 10 – RL
Markov Decision Process

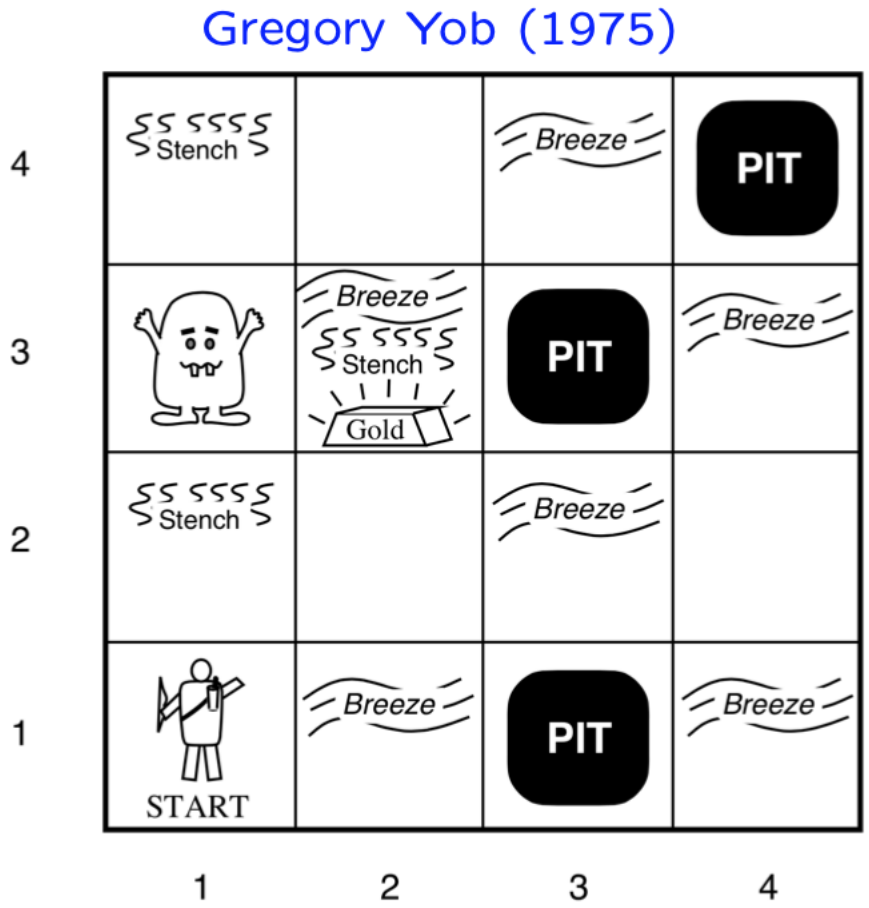


Example of a simple MDP with three states (green circles) and two actions (orange circles), with two rewards (orange arrows).

- Set of states S
- Set of actions ACTIONS(s)
- Transition model P(s' | s, a)
- Reward function R(s, a, s')

https://en.wikipedia.org/wiki/Markov_decision_process

Week 10 – RL
Markov Decision Process - Example



The Wumpus World

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- method for learning a function Q(s, a), estimate of the value of performing action a in state s

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- method for learning a function Q(s, a), estimate of the value of performing action a in state s

- Start with Q(s, a) = 0 for all s, a
- When we taken an action and receive a reward:
  - Estimate the value of Q(s, a) based on current reward and expected future rewards
  - Update Q(s, a) to take into account old estimate as well as our new estimate

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- Start with Q(s, a) = 0 for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(\text{new value estimate} - \text{old value estimate})$$

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- Start with Q(s, a) = 0 for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(\text{new value estimate} - Q(s, a))$$

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- Start with Q(s, a) = 0 for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha((r + \text{future reward estimate}) - Q(s, a))$$

Week 10 – RL
Q-learning

University of Salford MANCHESTER

- Start with Q(s, a) = 0 for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha((r + \max_{a'} Q(s', a'))) - Q(s, a))$$

Week 10 – RL
Q-learning

- Start with $Q(s, a) = 0$ for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha((r + \gamma \max_{a'} Q(s', a'))) - Q(s, a))$$

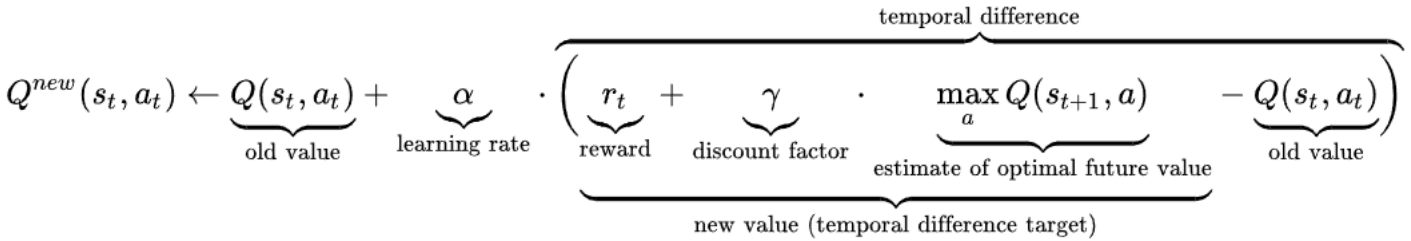
Week 10 – RL
Q-learning – Greedy VS Epsilon Greedy

University of Salford MANCHESTER

- Start with Q(s, a) = 0 for all s, a
- Every time we take an action a in state s and observe a reward r, we update:

$$Q(s, a) \leftarrow Q(s, a) + \alpha((r + \gamma \max_{a'} Q(s', a'))) - Q(s, a))$$

Week 10 – RL
Q-learning – Greedy VS Epsilon Greedy



https://en.wikipedia.org/wiki/Q-learning

Week 10 – RL
Q-learning - Example

University of Salford MANCHESTER

https://www.youtube.com/watch?v=M-QUkgk3HyE

Week 10 – RL
Q-learning – function approximation

University of Salford MANCHESTER

approximating Q(s, a), often by a function combining various features, rather than storing one value for every state-action pair

https://deepmind.com/blog/article/alphago-zero-starting-scratch

Week 10 Summary

University of Salford MANCHESTER

MAB
Contextual Bandits
RL

Week 10

University of Salford MANCHESTER

Recommendation System

PCA

Optimizing Deep learning

Something else

University of Salford MANCHESTER

Thanks for the Attention! ☺