c)

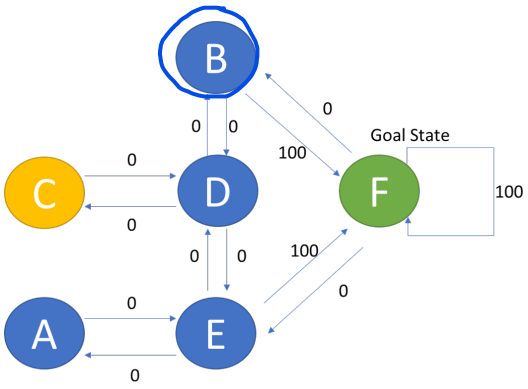
Q-learning is a variant of making an agent experience an environment without knowing the model behind it (learning without knowledge of the MDP). With Q-learning the explicit learning of the *policy* is omitted, instead the *policy* is learned directly.
The following formula applies:

$$Q_{k+1}(s_t, a_t) < -Q_k(s_t, a_t) + \alpha[R(s, a) + \gamma \max(Q_k(s_{t+1}, a)) - Q_k(s_t, a_t))]$$

where α is the *learning rate* and γ is the *discount factor*.
In this example, let the reward matrix R be given as:

state/action	A	B	C	D	E	F
A	-	-	-	-	0	-
B	-	-	-	0	-	100
C	-	-	-	0	-	-
D	-	0	0	-	0	-
E	0	-	-	0	-	100
F	-	0	-	-	0	100

As can be easily seen, there are 6 states $S = A, B, C, D, E, F$ and the actions A that allow an agent to move from a state S_1 to a state S_2 (e.g. from A to E or from D to B,C or E). The example could be a building with rooms, and doors that allow an agent to transition from one room to another.



Apply the Q-learning algorithm step by step. Calculate the following values (k denotes the respective episodes) with $\alpha = 1$ and $\gamma = 0.8$. An episode ends when the goal (=goal state) is reached:

Name: Matrix number:

$Q_{k=1}(B,F) =$ $0 + 1(100 + 0.8(0 - 0))$

$Q_{k=2}(D,B) =$ $100 + (0 + 0.8(100 - 100))$

$Q_{k=3}(C,D) =$ $100 + (0 + 0.8(100 - 0))$

$Q_{k=4}(E,D) =$ $180 + (0 + 0.8(0 - 0))$

What does the $Q - matrix$ look like after the 4 episodes?

$Q =$

Name: Matrix number:

Worksheet