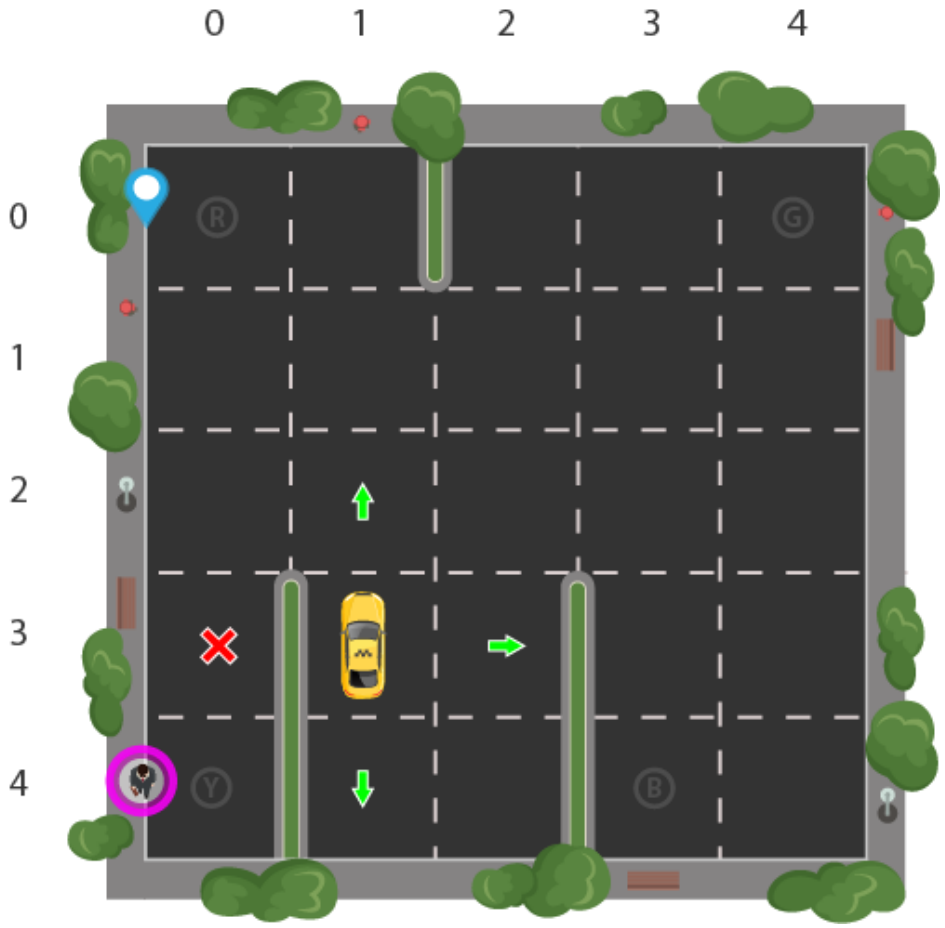
Q-Learning

DATA SCIENCE
ONLINE

Volviendo a nuestro SmartCab…



- Fue capaz de resolver el escenario de la figura…

… en unos discretos cientos de movimientos

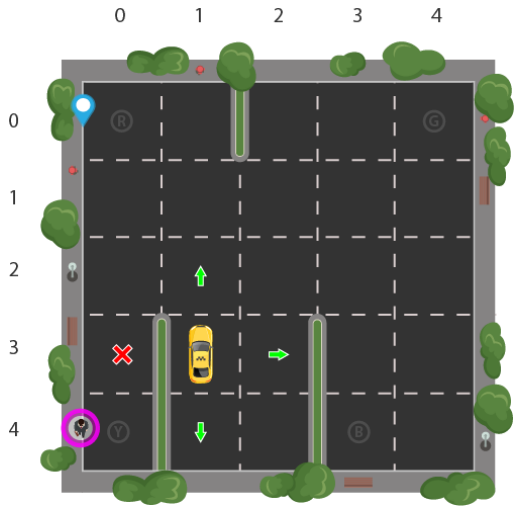
- Lo “aprendido” solo le sirve para ese escenario
- ¿Existe alguna manera de que aprenda realmente?

…imagina que tenemos una tabla…

- Una tabla que tuviera tantas filas como estados y tantas columnas como acciones posibles

Estado	Sur	Norte	Este	Oeste	Recoger	Dejar
0	…	…	…	…	…	…
…	…	…	…	…	…	…
328	-4	-1.8	-2.3	-3.8	-12	-12
…	…	…	…	…	…	…
499	…	…	…	…	…	…

- Cada celda indicara el valor de la mayor recompensa ACUMULABLE que podrías obtener si estando en el estado indicado por la fila ejecutaras la acción indicada por la columna



Estado: 328

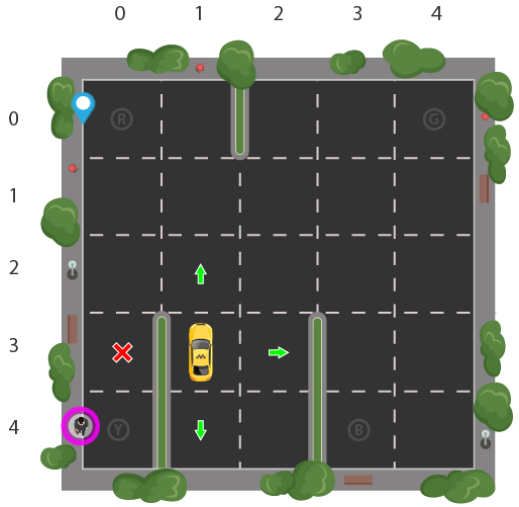
…imagina que tenemos una tabla…

- Una tabla que tuviera tantas filas como estados y tantas columnas como acciones posibles

Estado	Sur	Norte	Este	Oeste	Recoger	Dejar
0	…	…	…	…	…	…
…	…	…	…	…	…	…
328	-4	-1.8	-2.3	-3.8	-12	-12
…	…	…	…	…	…	…
499	…	…	…	…	…	…

- Cada celda indicara el valor de la mayor recompensa ACUMULABLE que podrías obtener si estando en el estado indicado por la fila ejecutaras la acción indicada por la columna

- Q-Learning: Dada esta tabla, el agente escoge siempre la acción con el valor máximo de su celda



Estado: 328

…una Q-Table (o tabla de valores Q)

Estado	Sur	Norte	Este	Oeste	Recoger	Dejar
0	…	…	…	…	…	…
…	…	…	…	…	…	…
328	-4	-1.8	-2.3	-3.8	-12	-12
…	…	…	…	…	…	…
499	…	…	…	…	…	…

- La Q-Table nos da para cada combinación (estado,acción) lo que se denomina su valor Q
- Los valores Q se representan como Q(s,a) donde s es el estado y a la acción
- En el ejemplo el Q(328,”sur”) es -4

¿Cómo se obtiene la tabla-Q?

- El agente explorará de forma más o menos aleatoria el entorno
- Obteniendo valores de recompensa y actualizando una tabla-Q intermedia
- Continúa haciéndolo hasta alcanzar un criterio de parada