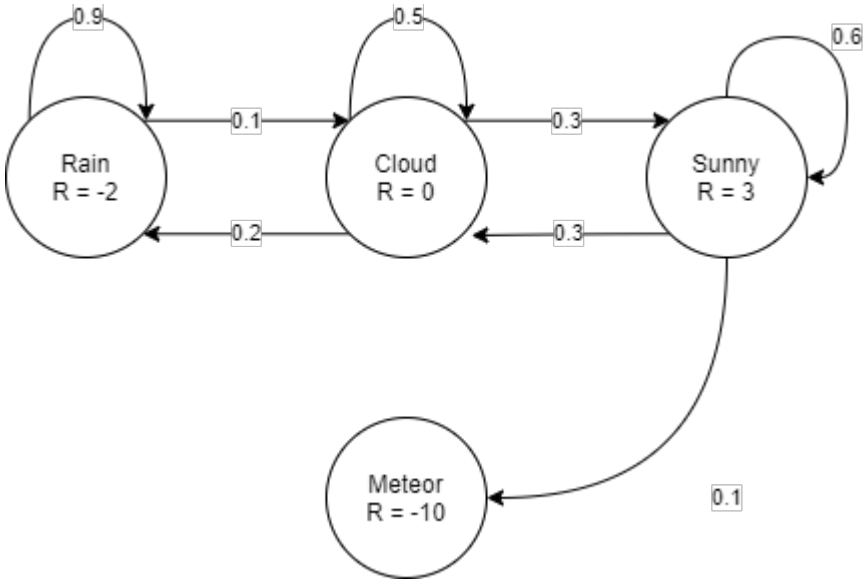
1.1 Markov Chain & 1.2 Markov Reward Process



1.3 Sampling. Een voorbereiding voor Monte-Carlo Policy Evaluation

Markov Chain

	rain	cloudy	sunny	meteor
rain	0,9	0,1	0	0
cloudy	0,2	0,5	0,3	0
sunny	0	0,3	0,6	0,1
meteor	0	0	0	0

Markov Reward

rain	cloudy	sunny	meteor
-2	0	3	-10

rain	cloudy	sunny	meteor		reward
-2	0	3	-10	=	-9

cloudy	cloudy	rain	rain	cloudy	sunny	sunny	sunny	meteor		reward
0	0	-2	-2	0	3	3	3	-10	=	-5

1.4 De value-function bepalen

Markov Chain

	rain	cloudy	sunny	meteor
rain	0,9	0,1	0	0
cloudy	0,2	0,5	0,3	0
sunny	0	0,3	0,6	0,1
meteor	0	0	0	0

Markov Reward

rain	cloudy	sunny	meteor
-2	0	3	-10

iter	rain	cloudy	sunny	meteor
0	0	0	0	0
1	-1,8	0,5	0,8	0
2	-3,37	0,63	1,43	0
3	-4,77	0,57	1,847	0
4	-6,036	0,3851	2,0792	0
5	-7,19389	0,10911	2,16305	0

Zie excel voor de formule toepassing*

1.5 Zelf-onderzoek

Bij een discount factor van 1, zullen de waardens nooit tot stilstand komen en altijd zichzelf aanpassen bij het bepalen van de value.

In het begin van Q-learning zijn die resulaten minder belangrijk dan later in het leer proces. Je ziet vaker dat in verloop van de tijd de discount factor omhoog gaat zodat nieuw geleerde dingen meer mee tellen.