Q-Learning with Stocks: Definitions for our seminar's example	
Immediate view and choices	
Agent	The trader who takes actions.
Long vs. Flat	Business jargon for buying (long) or not trading (flat) an asset. However, we use the term flat here to indicate that we are not invested.
State	All the information the agent *sees right now* when deciding (e.g. last 10 returns and whether he/she was already long or flat). The state should be "enough" to predict what happens next (Markov property).
Action	The choice available in that state – here: *stay/turn flat* (0) or *stay/turn long* (1).
Policy	The agent's rule (possibly random) that maps each state to an action, written $\pi(a \mid s)$.
World the agent interacts with	
Environment	The outside world producing the next state and reward after an action. In this lesson: the SPY price stream plus our trading rules.
Episode	One complete run until the environment signals *done* (here: when we reach the last price bar).
Reward	The immediate payoff after an action (daily return minus transaction cost).
Markov Decision Process (MDP)	The formal "state, action, reward, next state, done" framework underlying the whole problem.
Value estimation	
Q-function (action-value)	$Q^\pi(s,a)$ = expected *total future* reward if you start in state s, take action a once, then follow policy π. It is *computed with* the Bellman equation.
Bellman equation	The recursion tying a Q-value to its own next-step reward: $Q(s,a) = \mathbb{E}[\, r + \gamma \max_{a'} Q(s',a') \,]$. Foundation of Q-learning.
Discount factor γ	Number in $(0,1)$ telling how strongly we value future rewards compared with immediate ones.
Neural networks used	
Online (Q-)network	The neural net we keep updating so its outputs approximate true Q-values; it chooses actions during training.
Target (frozen) network	A periodically copied snapshot of the online network. Its stable Q-values are used as learning targets until the next copy.
Experience storage and training data	
Replay buffer	Rolling memory of past transitions $(s, a, r, s', \text{done})$ so batches are less correlated.
Mini-batch	A random subset of transitions drawn from the replay buffer for one gradient update.
Warm-up memory	Minimum buffer size required before training starts, to ensure batches are diverse.
Exploration mechanisms	
ε-greedy	With probability ε pick a random action (explore); with $1-\varepsilon$ pick the best estimated one (exploit). ε decays over time.
Exploration vs. exploitation	The tension between "try something new to learn" and "use what I already know to earn reward now". ε-greedy is one simple strategy to balance the two.

Table 1: Glossary of key reinforcement-learning terms for the trading Q-learning example.