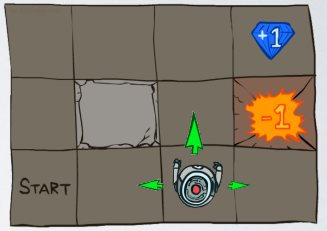
MODEL-BASED AND MODEL-FREE LEARNING

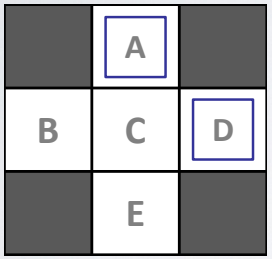
CSE 511A: Introduction to Artificial Intelligence

MODEL-BASED LEARNING

- Model-based learning:
  - Learn and approximate parameters (i.e., transition and reward functions) of the MDP model
  - Plan based on that approximated model just like in offline learning
- Step 1: Learn empirical MDP model
  - Count outcomes s' for each s and a
  - Normalize to give an estimate of $\hat{T}(s,a,s')$
  - Discover each $\hat{R}(s,a,s')$ when we experience (s,a,s')
- Step 2: Solve the learned MDP
  - For example, use value iteration as before



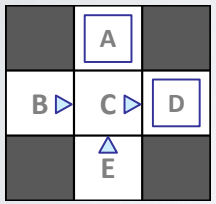
MODEL-BASED LEARNING



MODEL-BASED LEARNING

Input policy π

Observed Episodes (Training)



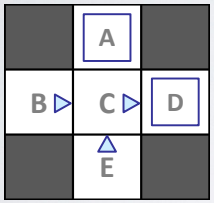
Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Assume $\gamma = 1$

MODEL-BASED LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1

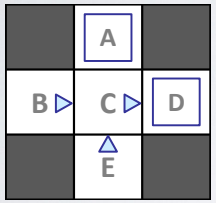
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

MODEL-BASED LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

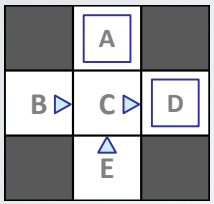
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

MODEL-BASED LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3

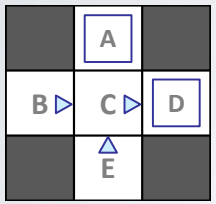
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -10

MODEL-BASED LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -10

Learned Model

$\hat{T}(s,a,s')$

T(B, east, C) = 1.00
T(C, east, D) = 0.75
T(C, east, A) = 0.25
…

$\hat{R}(s,a,s')$

R(B, east, C) = -1
R(C, east, D) = -1
R(D, exit, x) = +10
…

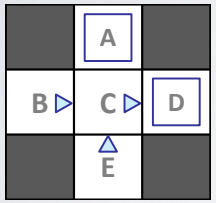
MODEL-FREE LEARNING

- Model-free learning:
  - Ignore MDP model completely
  - Learn and approximate the value functions V*(s) or Q*(s,*a*) directly
  - Extract policy based on approximated values

9

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

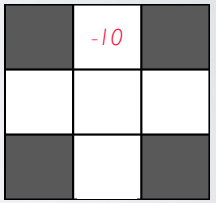
Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

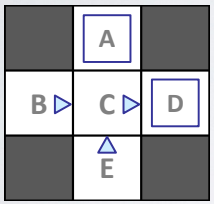
Learned Values



10

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

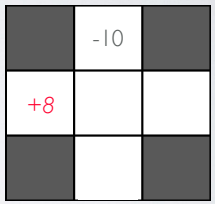
Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

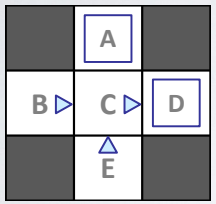
Learned Values



11

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

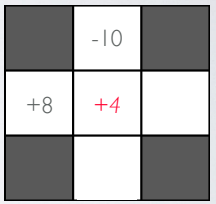
Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

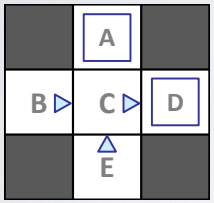
Learned Values



12

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

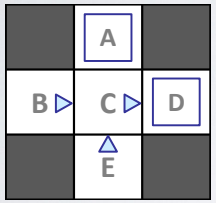
Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

Learned Values

	-10	
+8	+4	+10

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

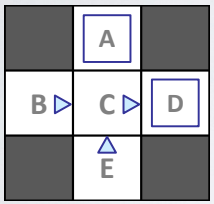
Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

Learned Values

	-10	
+8	+4	+10
	-2	

MODEL-FREE LEARNING

Input policy π



Assume $\gamma = 1$

Observed Episodes (Training)

Episode 1
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

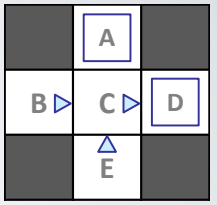
Learned Values

	-10	
+8	+4	+10
	-2	

Any problem with these numbers?

MODEL-FREE LEARNING

- What's good about this approach?
  - It's easy to understand
  - It doesn't require any knowledge of T or R
  - It eventually computes the correct average values, using just sample transitions
- What's bad about this approach?
  - It wastes information about state connections
  - Each state must be learned separately
    - e.g., if both B and E always go to C successfully under this policy, why are their values different?
  - So, it takes a long time to learn



	-10	
+8	+4	+10
	-2	