Dyna Style Planning with Linear Function Approximation

Alborz Geramifard

April, 2010

Acknowledgments

Richard Sutton

Csaba Szepesvari

Michael Bowling

Cosmin Paduraru

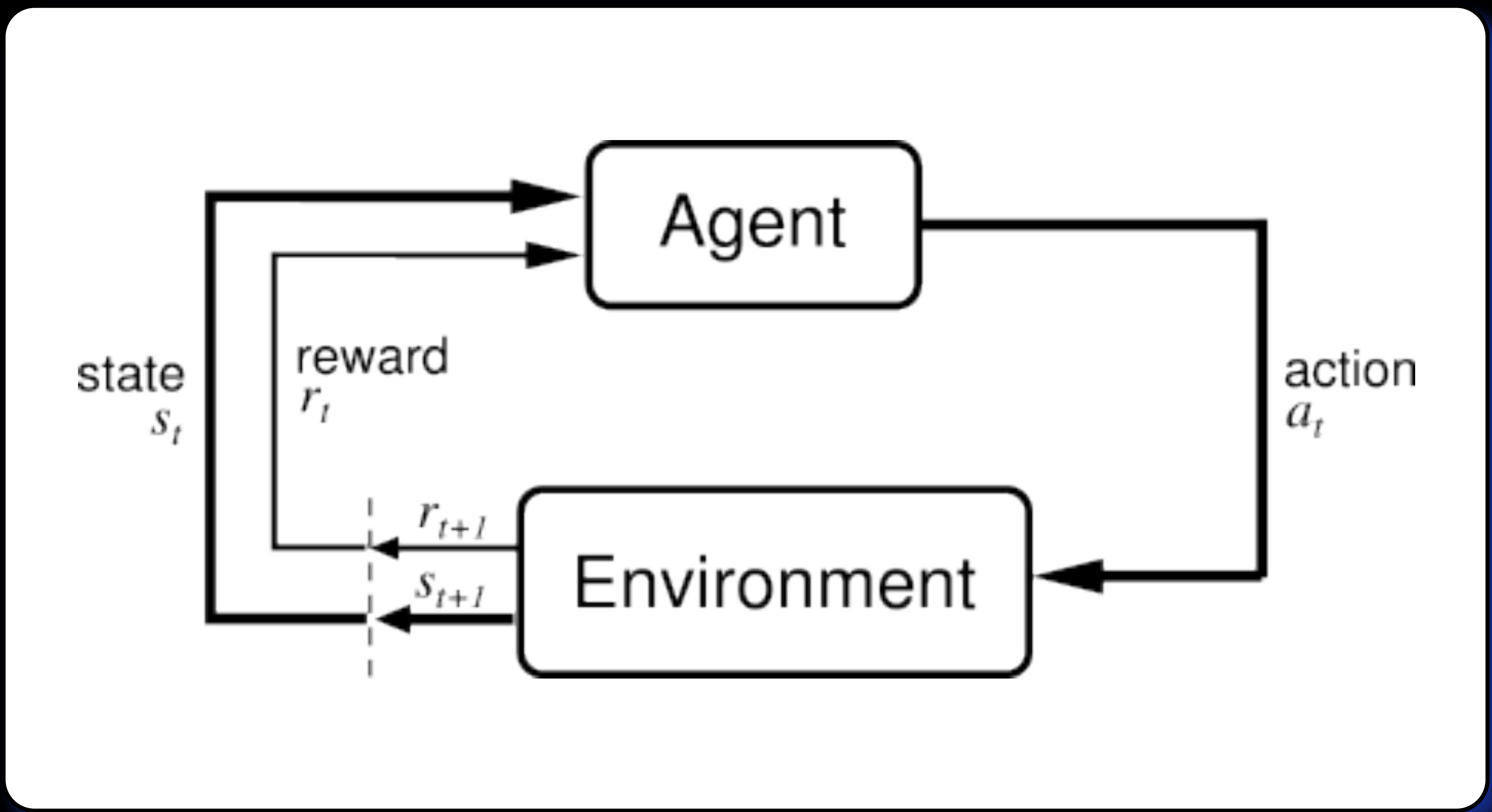
Outline

Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

Background

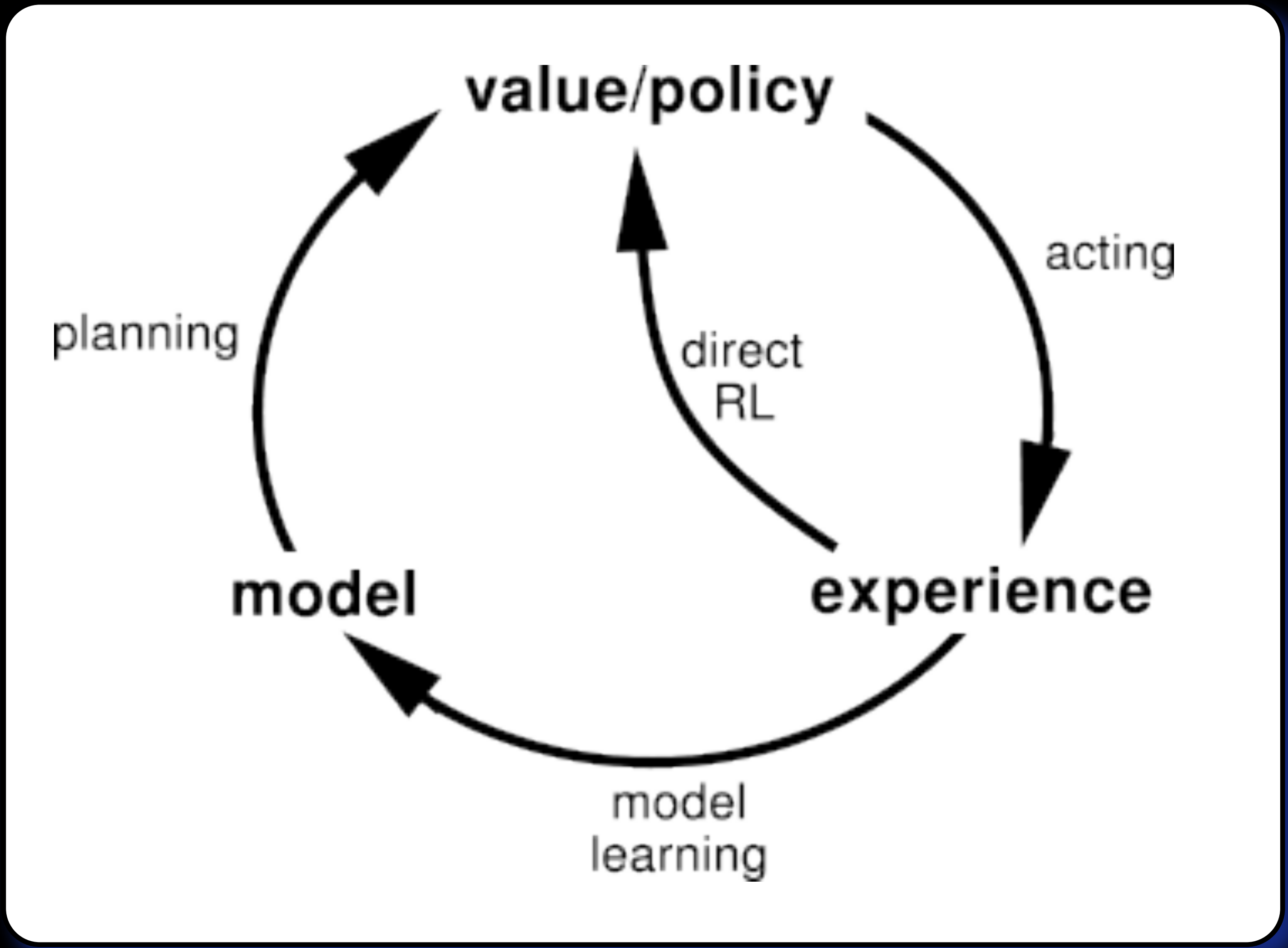
Reinforcement Learning



[Sutton, Barto 1998]

Background

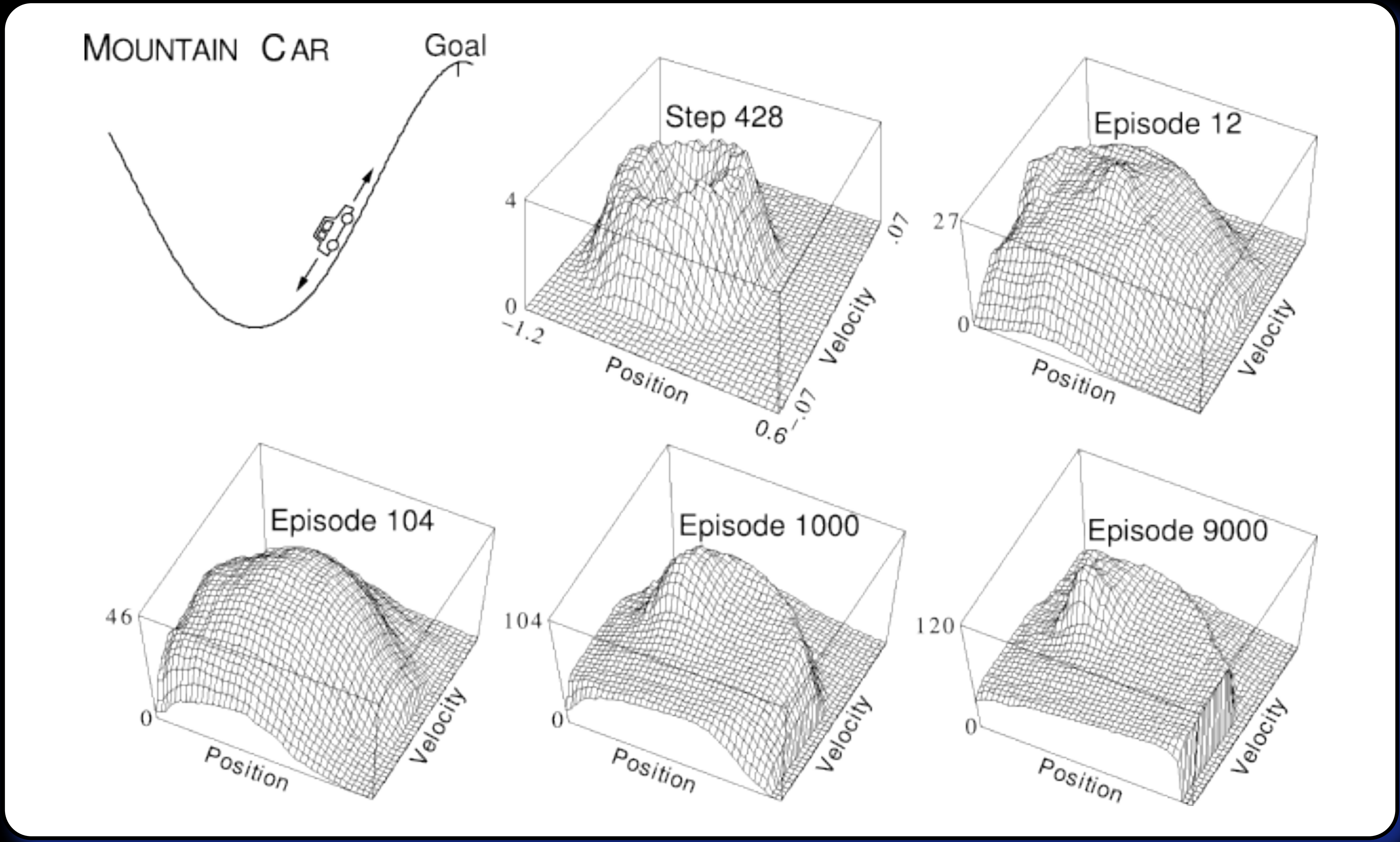
Planning



[Sutton, Barto 1998]

Background

Function Approximation



[Sutton, Barto 1998]

Background

Why planning ?

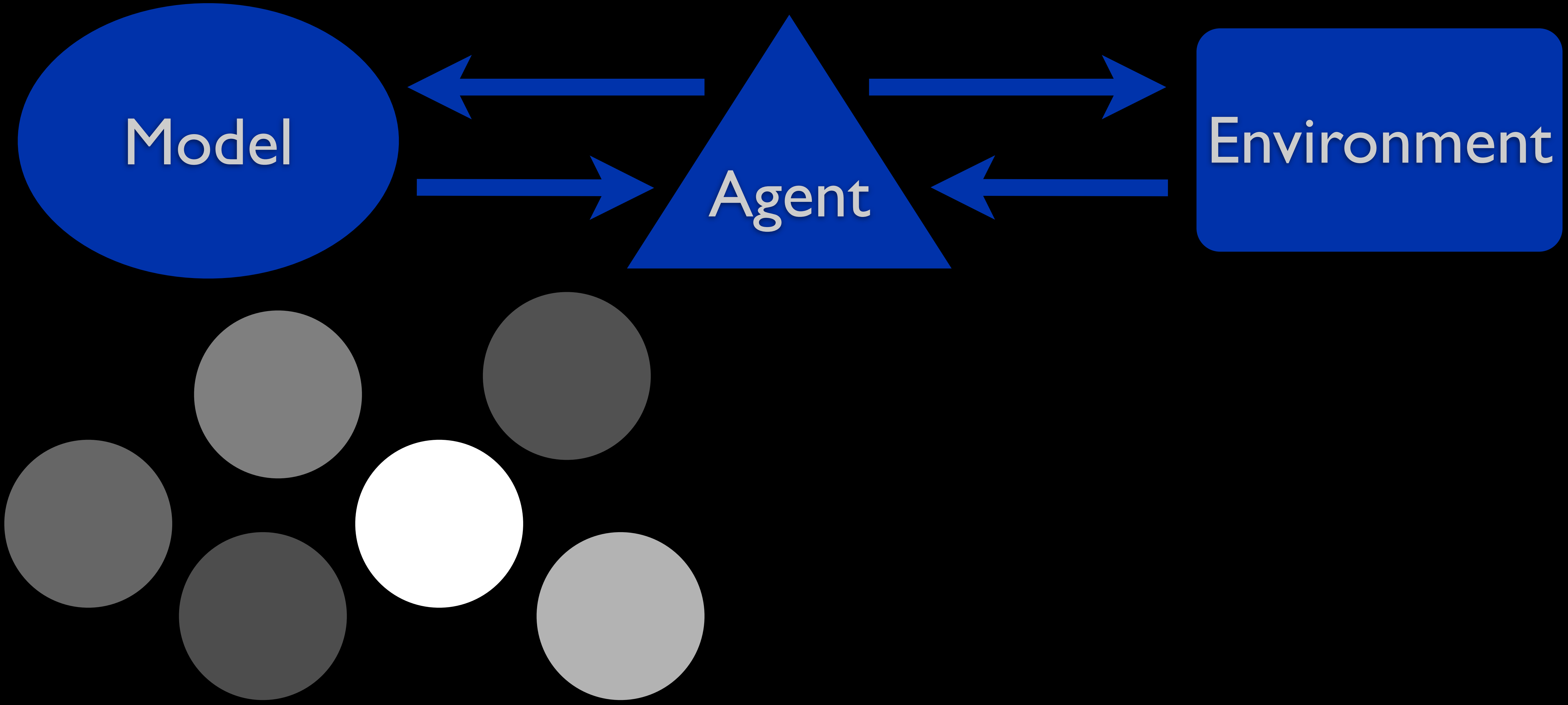
- Expensive data
- Trade off between data and time
- Tracking ...

Background

Prioritized Sweeping

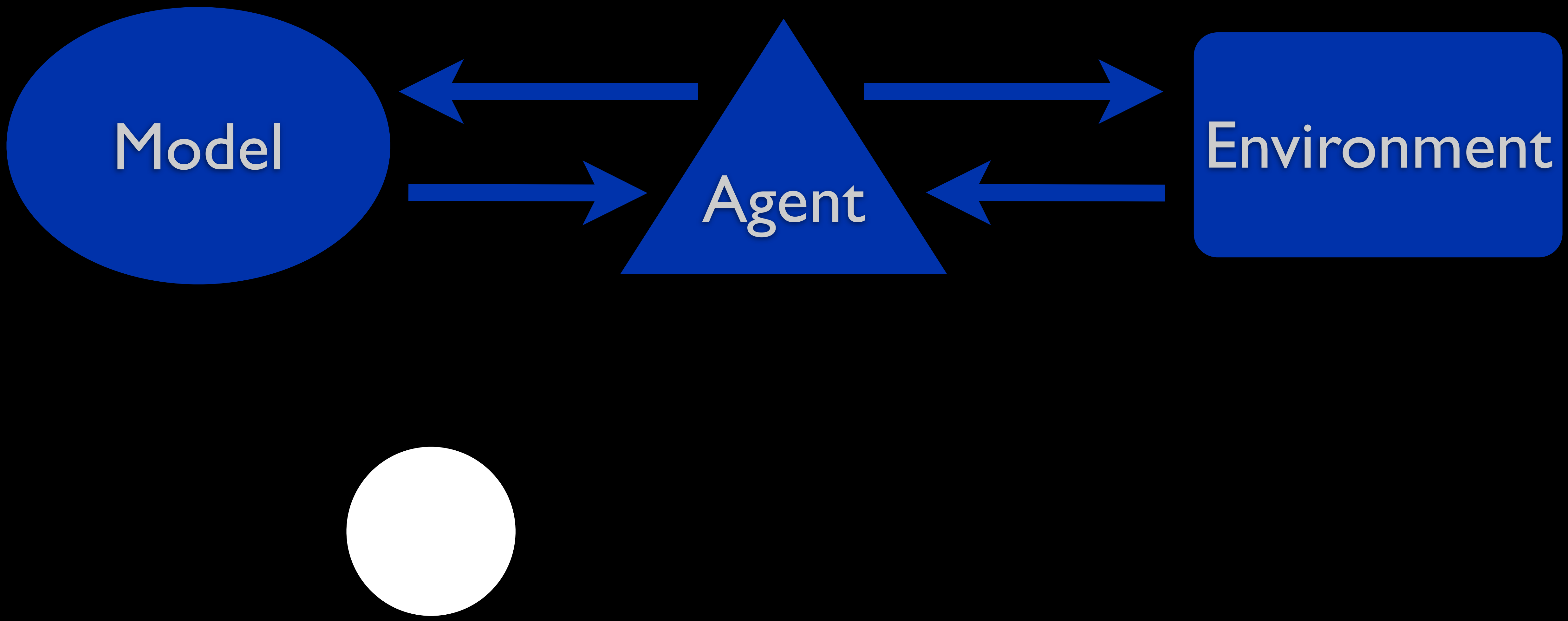
Background

Prioritized Sweeping



Background

Prioritized Sweeping



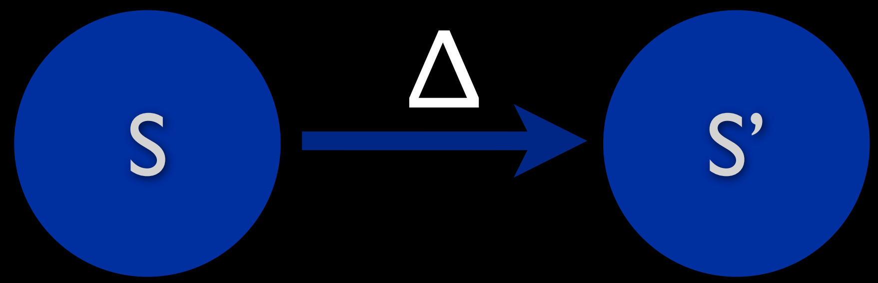
Background

Prioritized Sweeping

[Moore, Atkeson 1993]

Background

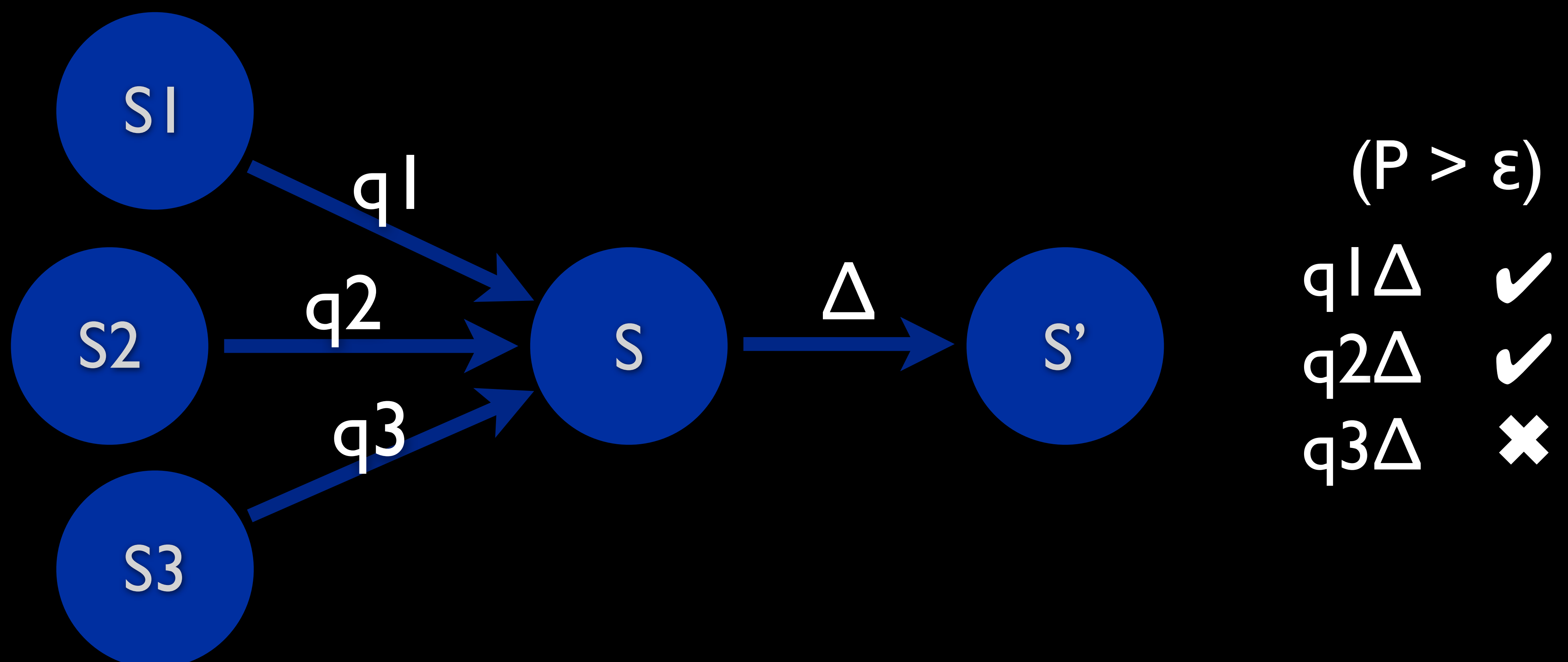
Prioritized Sweeping



[Moore, Atkeson 1993]

Background

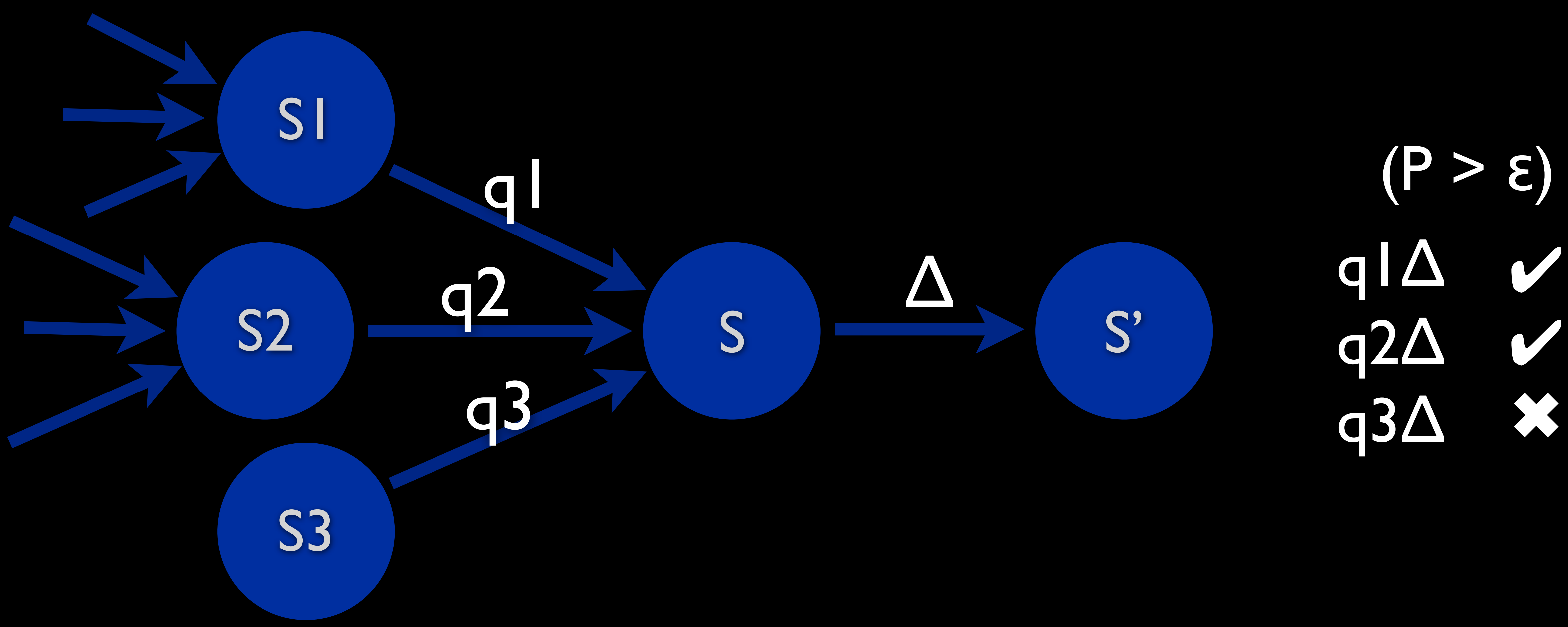
Prioritized Sweeping



[Moore, Atkeson 1993]

Background

Prioritized Sweeping



[Moore, Atkeson 1993]

Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

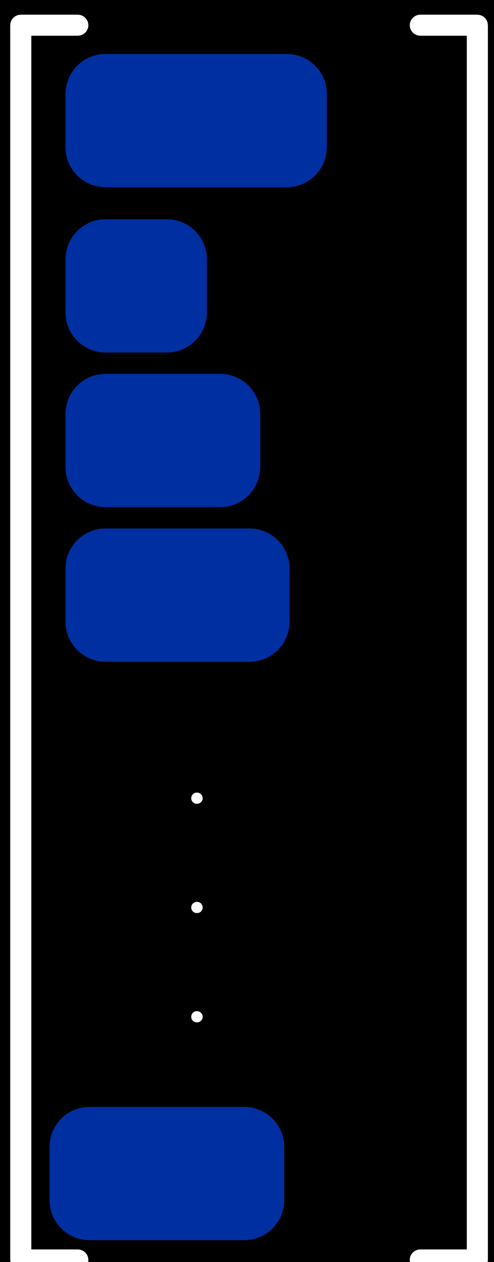
Linear Prioritized Sweeping

- Similar to Prioritized Sweeping
  - State ➠ Features
  - Sparse Features
  - Policy Independent Model

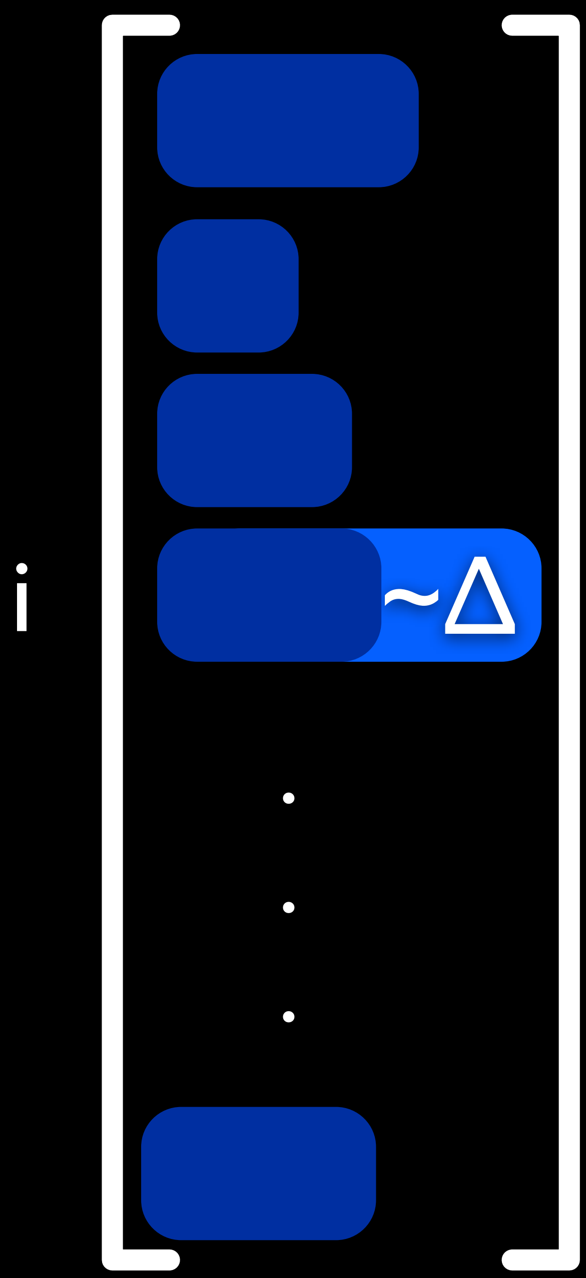
Linear Prioritized Sweeping

- Similar to Prioritized Sweeping
  - State ➠ Features
  - Sparse Features
  - Policy Independent Model

Dyna (PWMA)

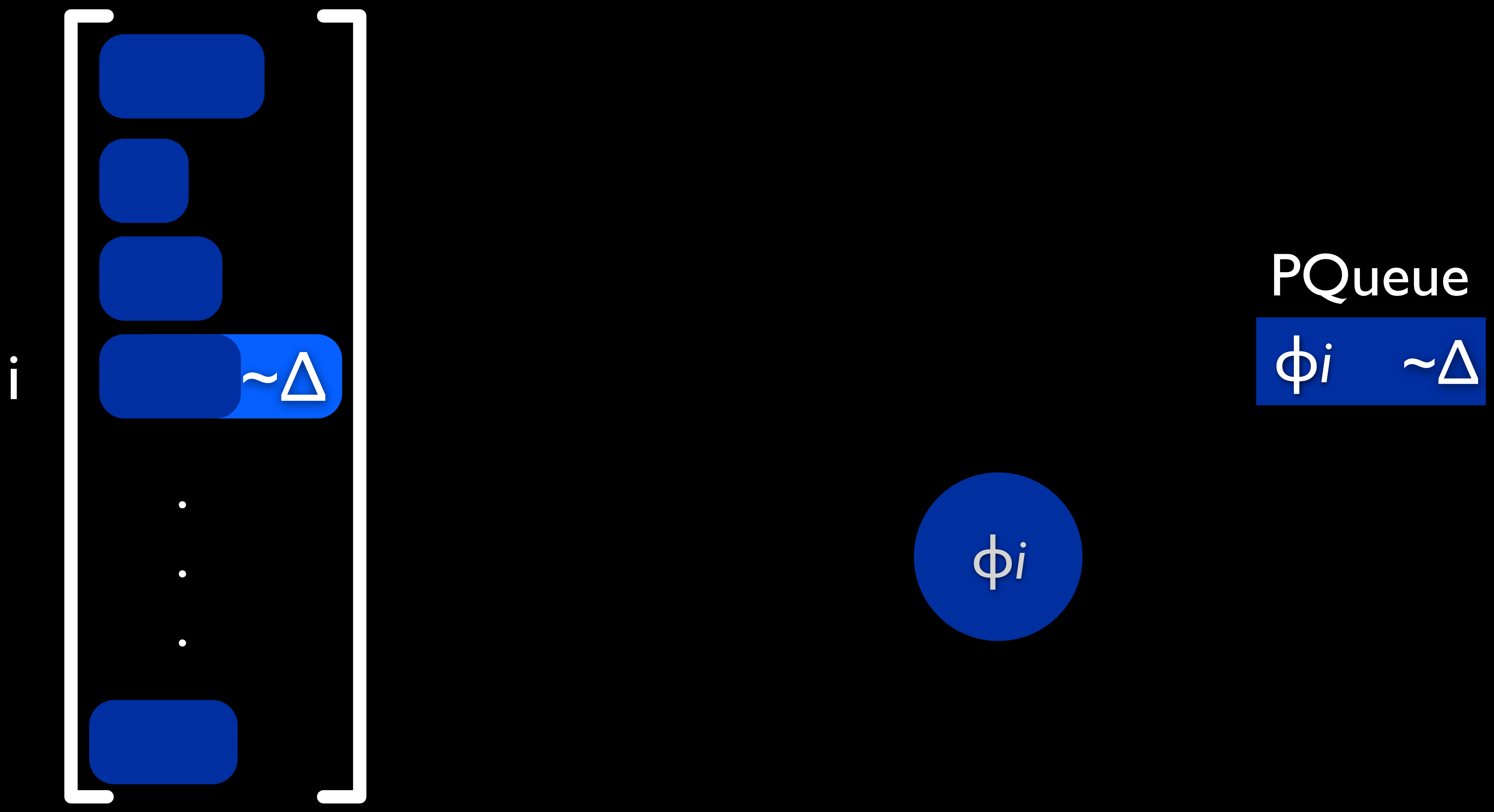


Dyna (PWMA)

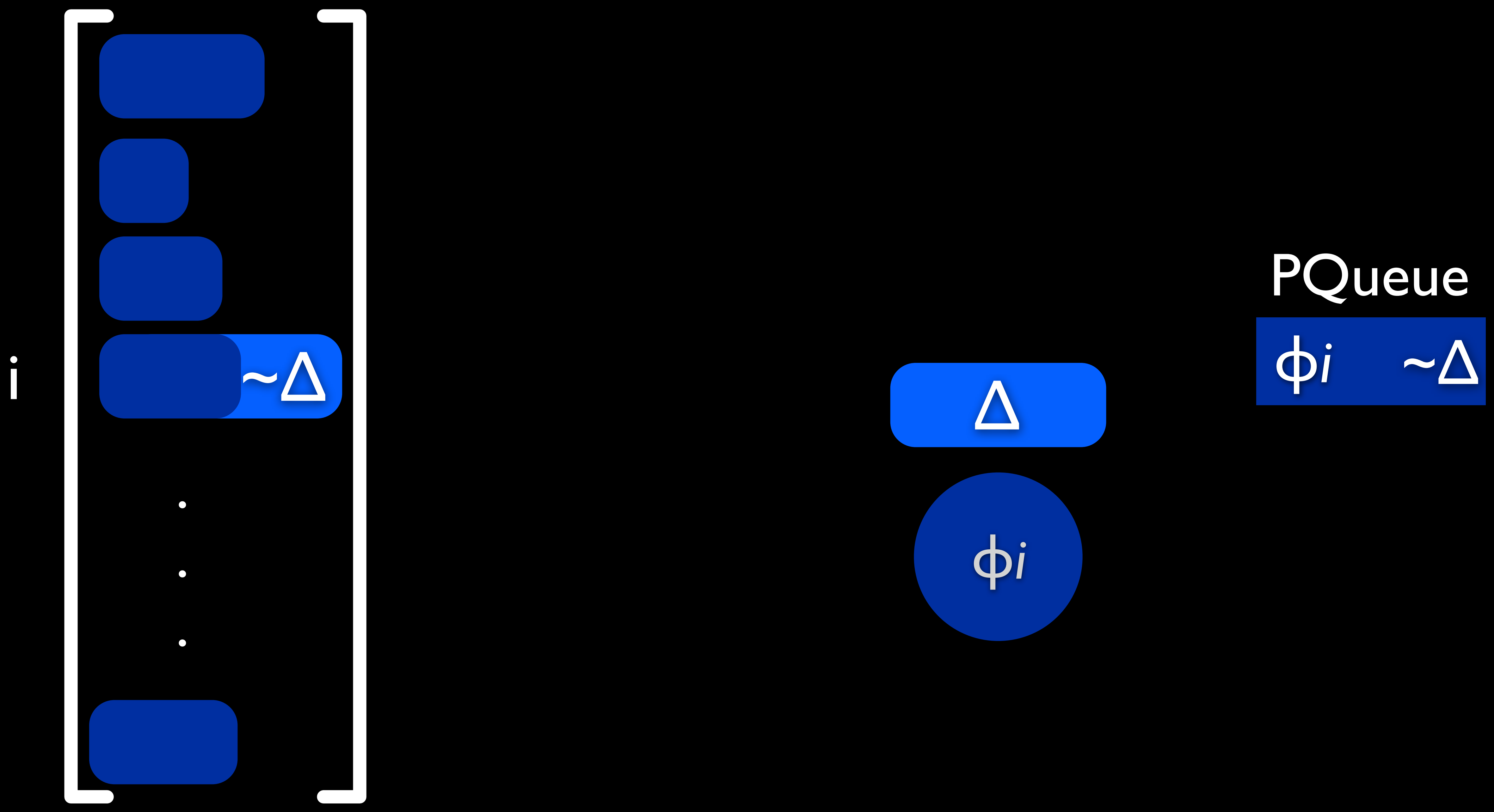


Dyna (PWMA)

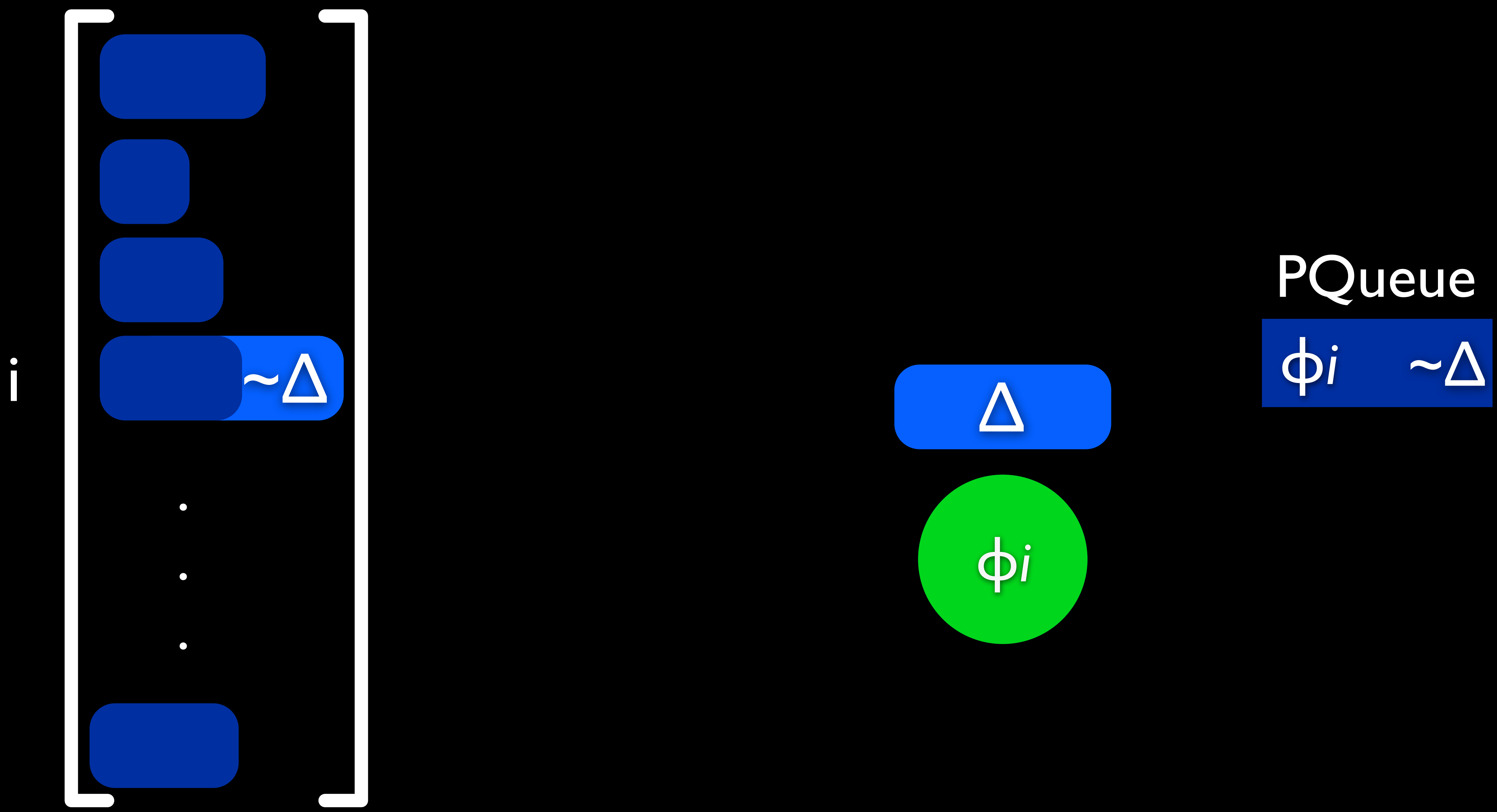
Dyna (PWMA)



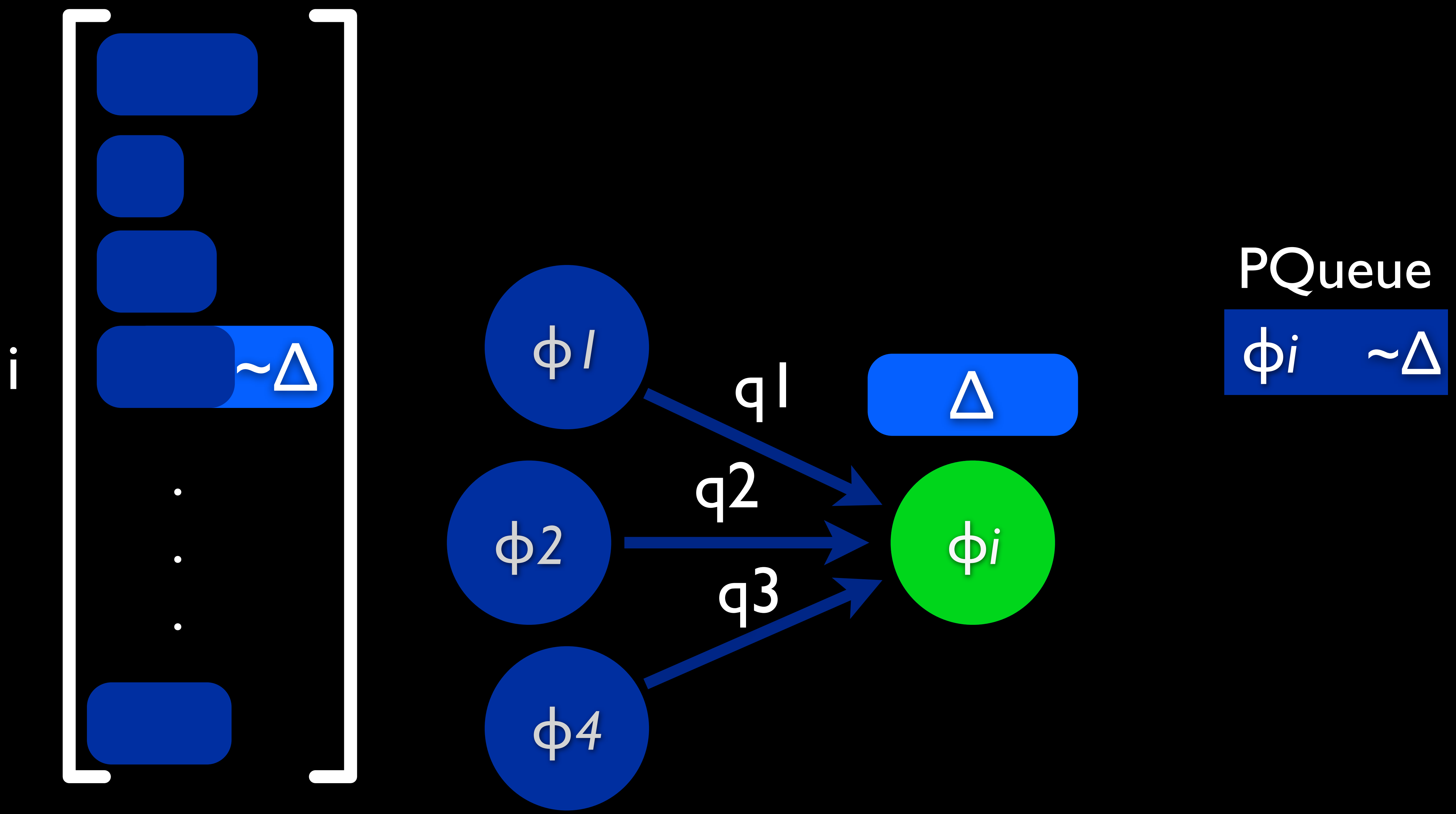
Dyna (PWMA)



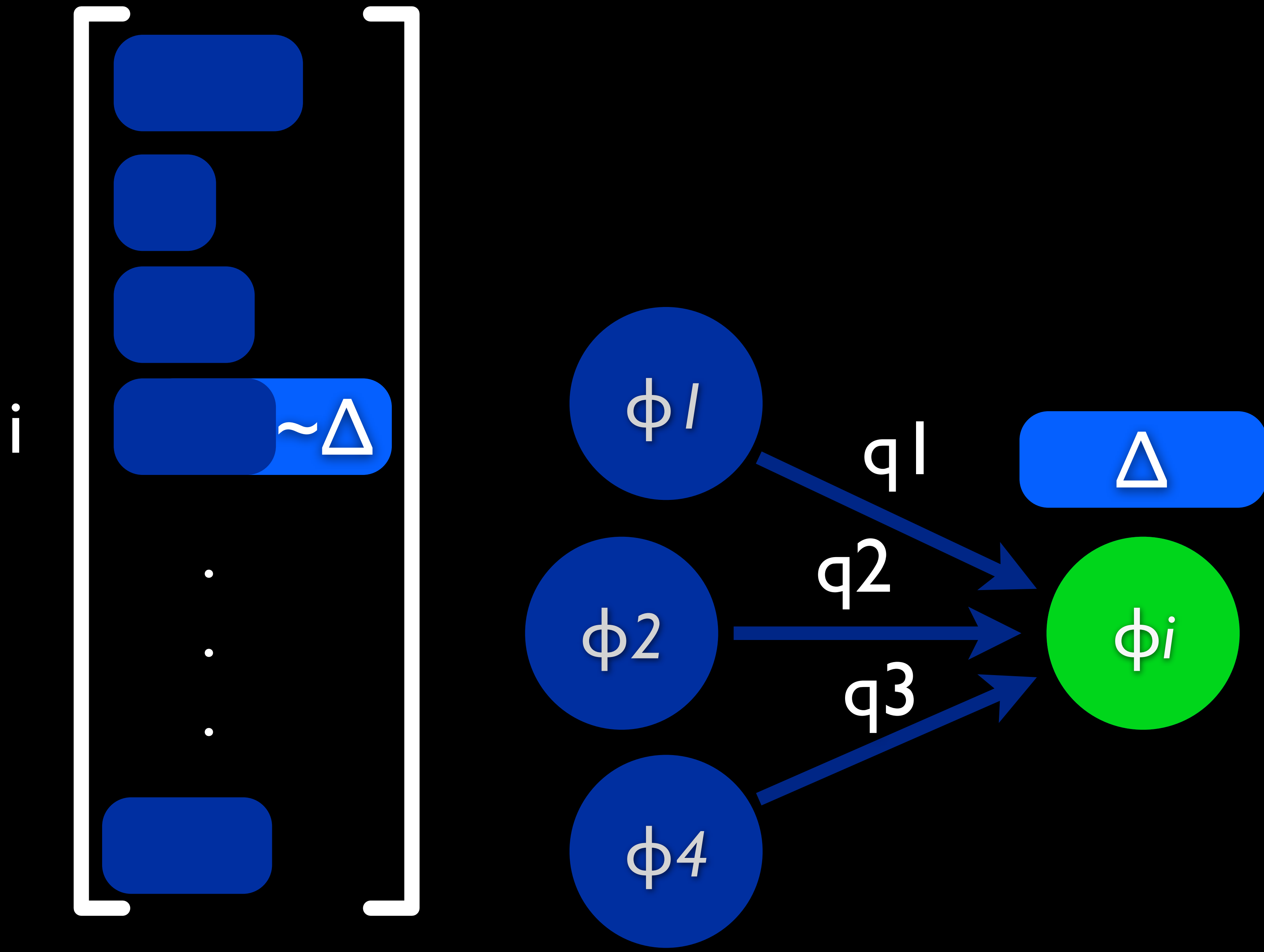
Dyna (PWMA)



Dyna (PWMA)



Dyna (PWMA)



PQueue

ϕ1	q1Δ
ϕ2	q2Δ
ϕ4	q3Δ

Building the Model

Transition Model : F_a

$$F_a \phi = \phi'$$

Reward Model : b_a

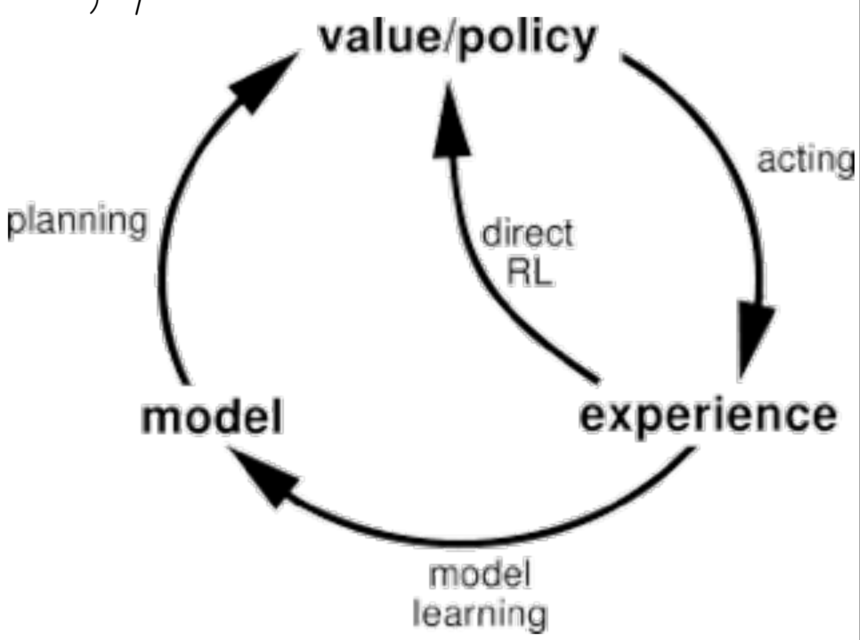
$$b_a^T \phi = r$$

Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

Obtain initial ϕ, θ, F, b

For each time step:

Take action a according to the policy. Receive r, ϕ'



Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

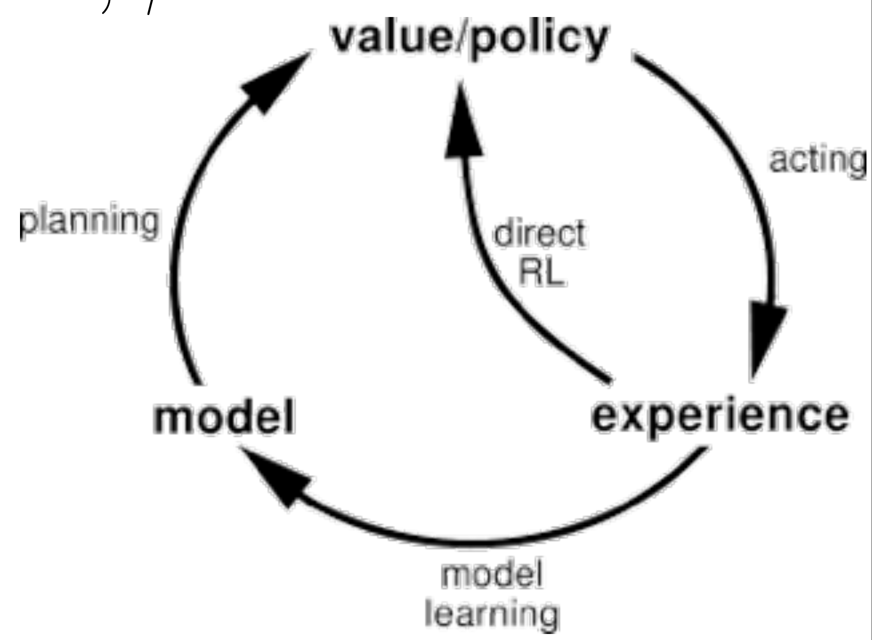
Obtain initial ϕ, θ, F, b

For each time step:

Take action a according to the policy. Receive r, ϕ'

$\delta \leftarrow r + \gamma\theta^\top \phi' - \theta^\top \phi$

$\theta \leftarrow \theta + \alpha\delta\phi$



Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

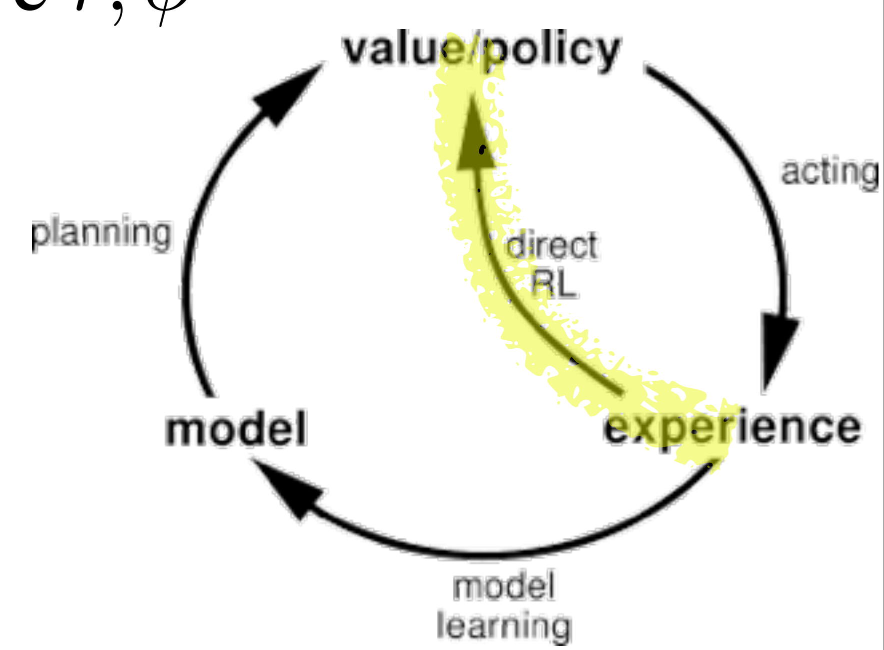
Obtain initial ϕ, θ, F, b

For each time step:

Take action a according to the policy. Receive r, ϕ'

$\delta \leftarrow r + \gamma\theta^\top\phi' - \theta^\top\phi$

$\theta \leftarrow \theta + \alpha\delta\phi$



Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

Obtain initial ϕ, θ, F, b

For each time step:

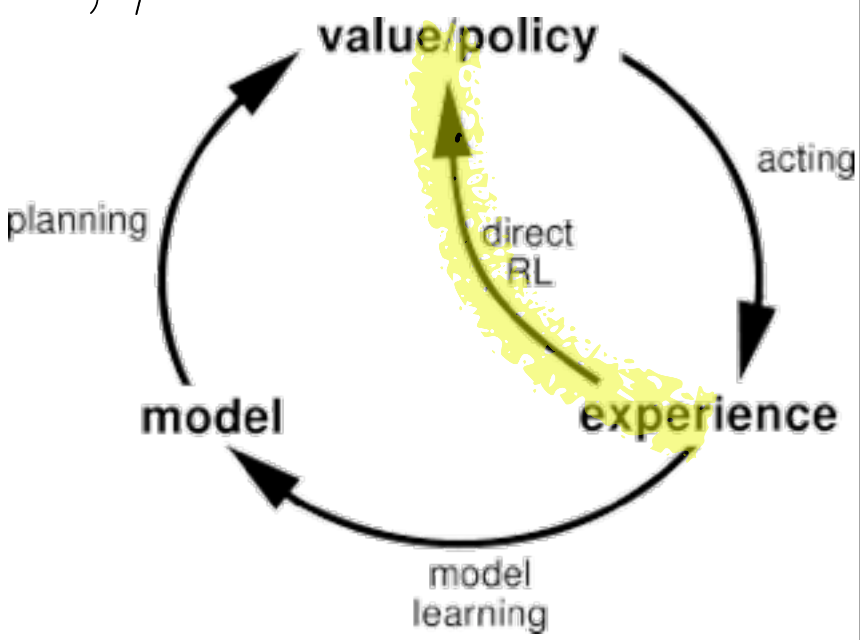
Take action a according to the policy. Receive r, ϕ'

$\delta \leftarrow r + \gamma\theta^\top\phi' - \theta^\top\phi$

$\theta \leftarrow \theta + \alpha\delta\phi$

$F \leftarrow F + \alpha(\phi' - F\phi)\phi^\top$

$b \leftarrow b + \alpha(r - b^\top\phi)\phi$



Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

Obtain initial ϕ, θ, F, b

For each time step:

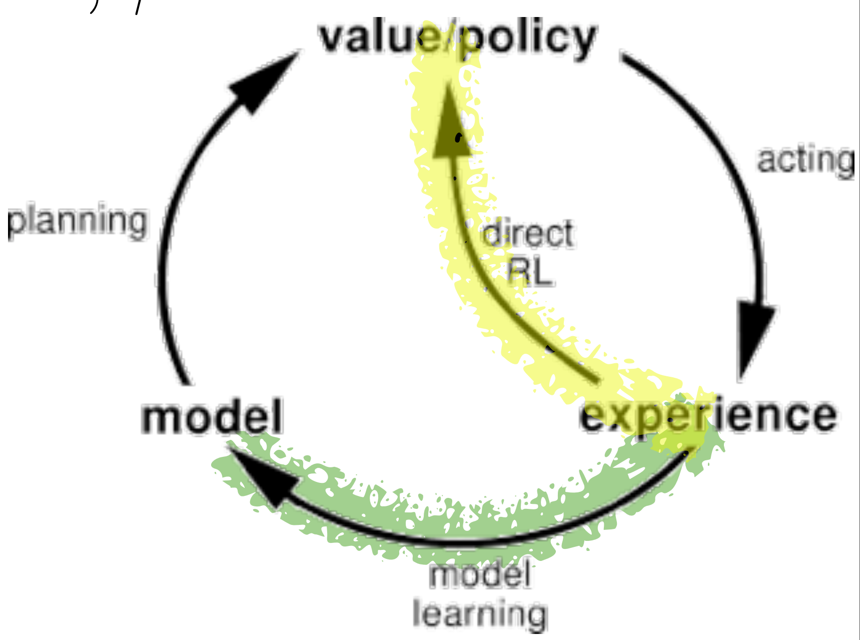
Take action a according to the policy. Receive r, ϕ'

$\delta \leftarrow r + \gamma\theta^\top\phi' - \theta^\top\phi$

$\theta \leftarrow \theta + \alpha\delta\phi$

$F \leftarrow F + \alpha(\phi' - F\phi)\phi^\top$

$b \leftarrow b + \alpha(r - b^\top\phi)\phi$



Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

Obtain initial ϕ, θ, F, b

For each time step:

Take action a according to the policy. Receive r, ϕ'

$\delta \leftarrow r + \gamma\theta^\top\phi' - \theta^\top\phi$

$\theta \leftarrow \theta + \alpha\delta\phi$

$F \leftarrow F + \alpha(\phi' - F\phi)\phi^\top$

$b \leftarrow b + \alpha(r - b^\top\phi)\phi$

For all i such that $\phi(i) \neq 0$:

For all j such that $F^{ij} \neq 0$:

Put j on the PQueue with priority $|F^{ij}\delta\phi(i)|$

Repeat p times while PQueue is not empty:

$i \leftarrow$ pop the PQueue

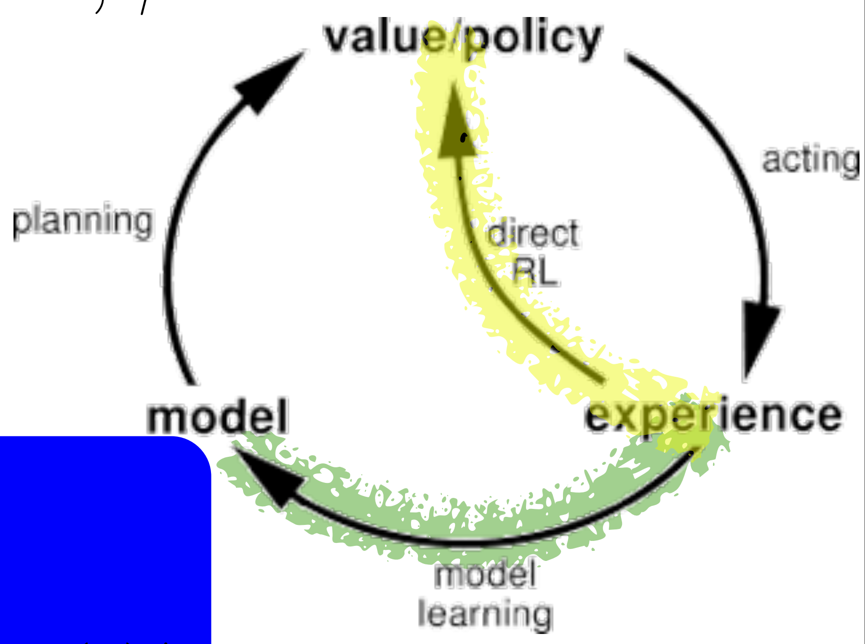
$\delta \leftarrow b(i) + \gamma\theta^\top F e_i - \theta(i)$

$\theta(i) \leftarrow \theta(i) + \alpha\delta$

For all j such that $F^{ij} \neq 0$:

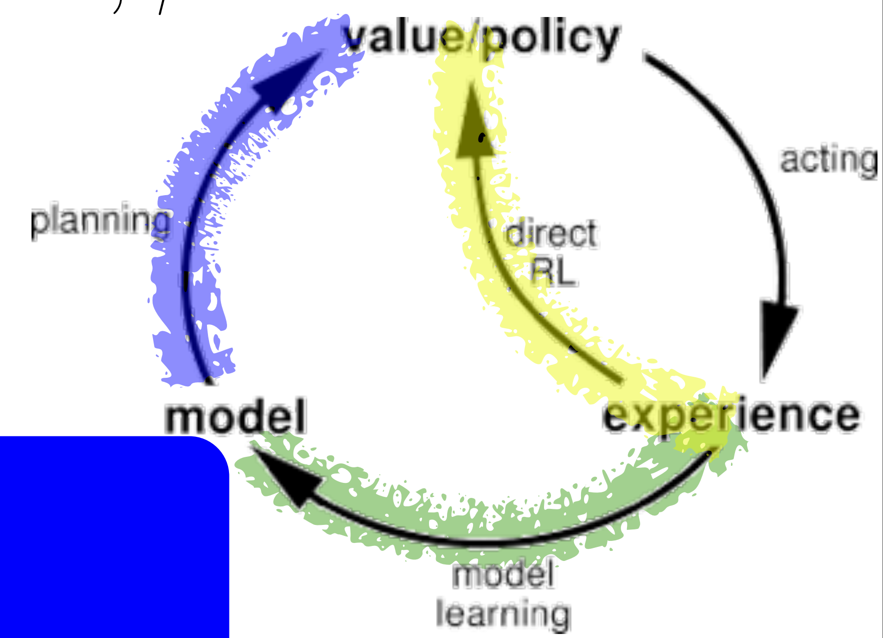
Put j on the queue with priority $|F^{ij}\delta|$

$\phi \leftarrow \phi'$

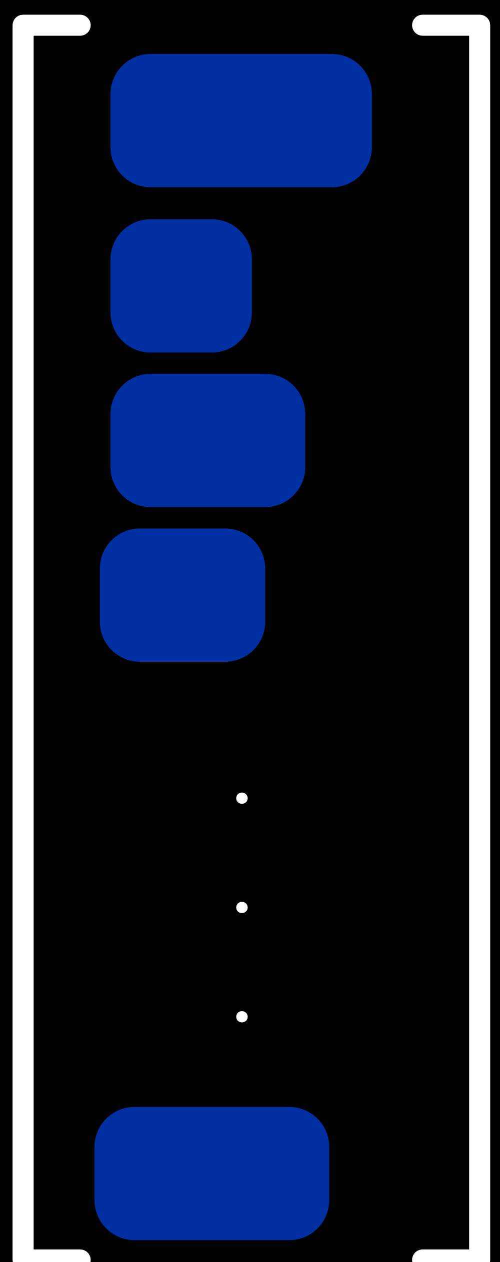


Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)

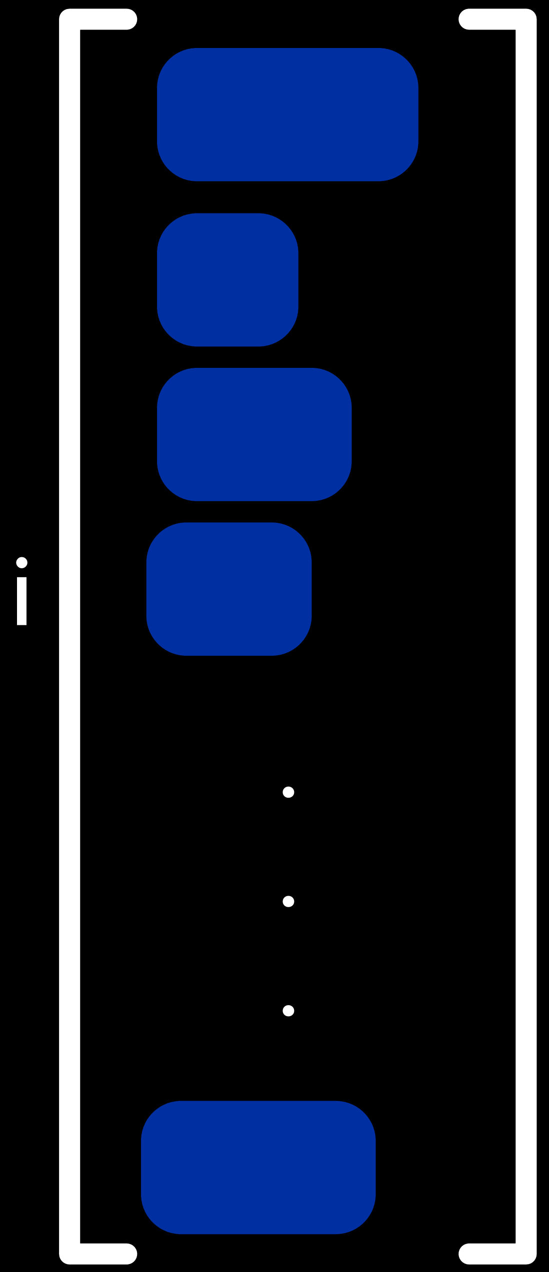
Obtain initial ϕ, θ, F, b
For each time step:
    Take action a according to the policy. Receive r, ϕ'
    $\delta \leftarrow r + \gamma\theta^\top\phi' - \theta^\top\phi$
    $\theta \leftarrow \theta + \alpha\delta\phi$
    $F \leftarrow F + \alpha(\phi' - F\phi)\phi^\top$
    $b \leftarrow b + \alpha(r - b^\top\phi)\phi$
    For all i such that $\phi(i) \neq 0$:
        For all j such that $F^{ij} \neq 0$:
            Put j on the PQueue with priority $|F^{ij}\delta\phi(i)|$
    Repeat p times while PQueue is not empty:
        $i \leftarrow$ pop the PQueue
        $\delta \leftarrow b(i) + \gamma\theta^\top F e_i - \theta(i)$
        $\theta(i) \leftarrow \theta(i) + \alpha\delta$
        For all j such that $F^{ij} \neq 0$:
            Put j on the queue with priority $|F^{ij}\delta|$
    $\phi \leftarrow \phi'$



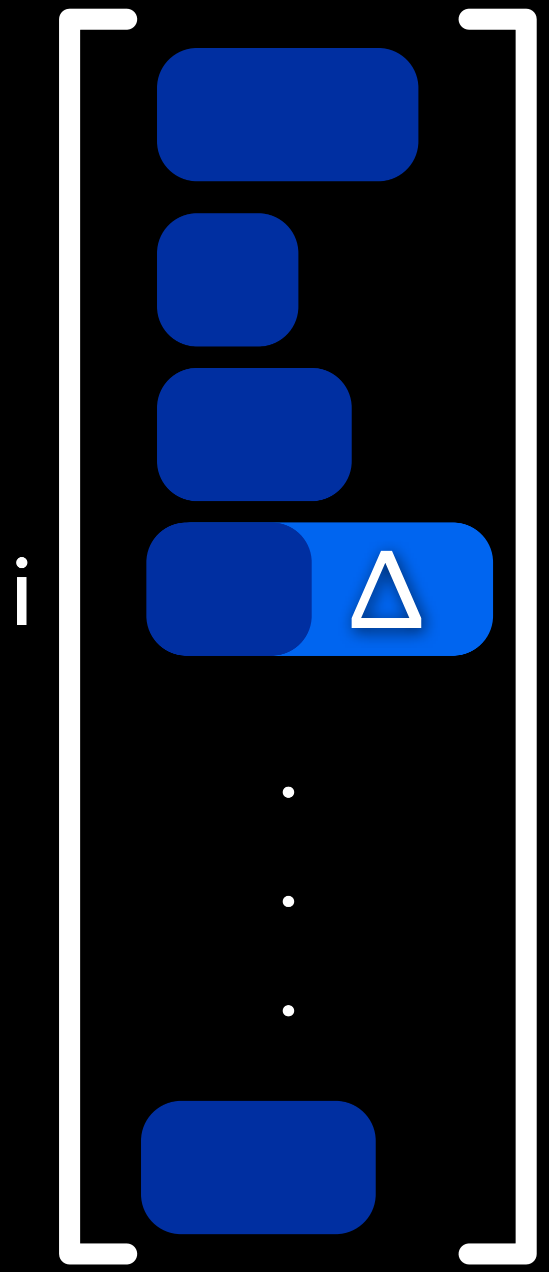
Dyna (MG)



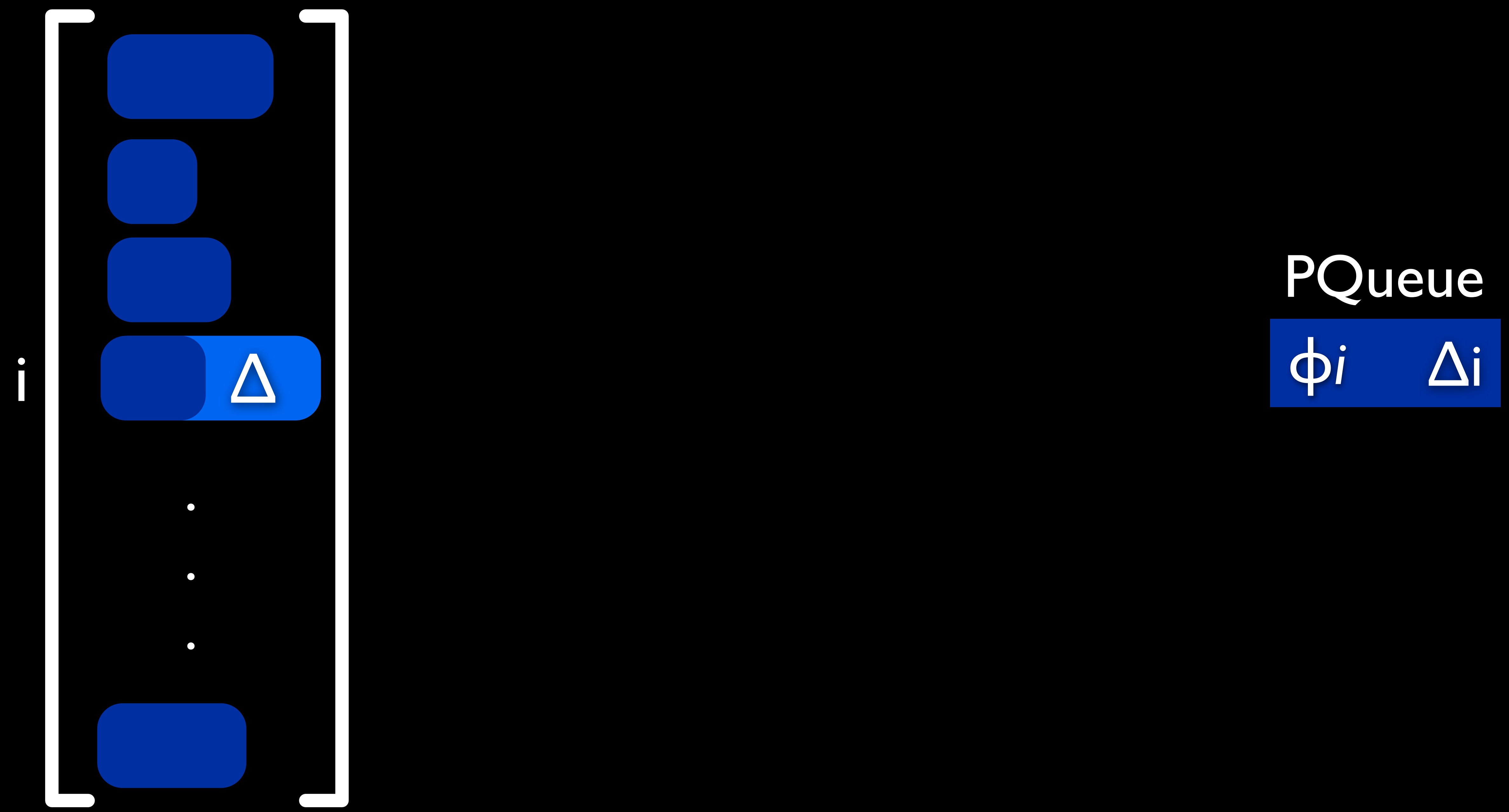
Dyna (MG)



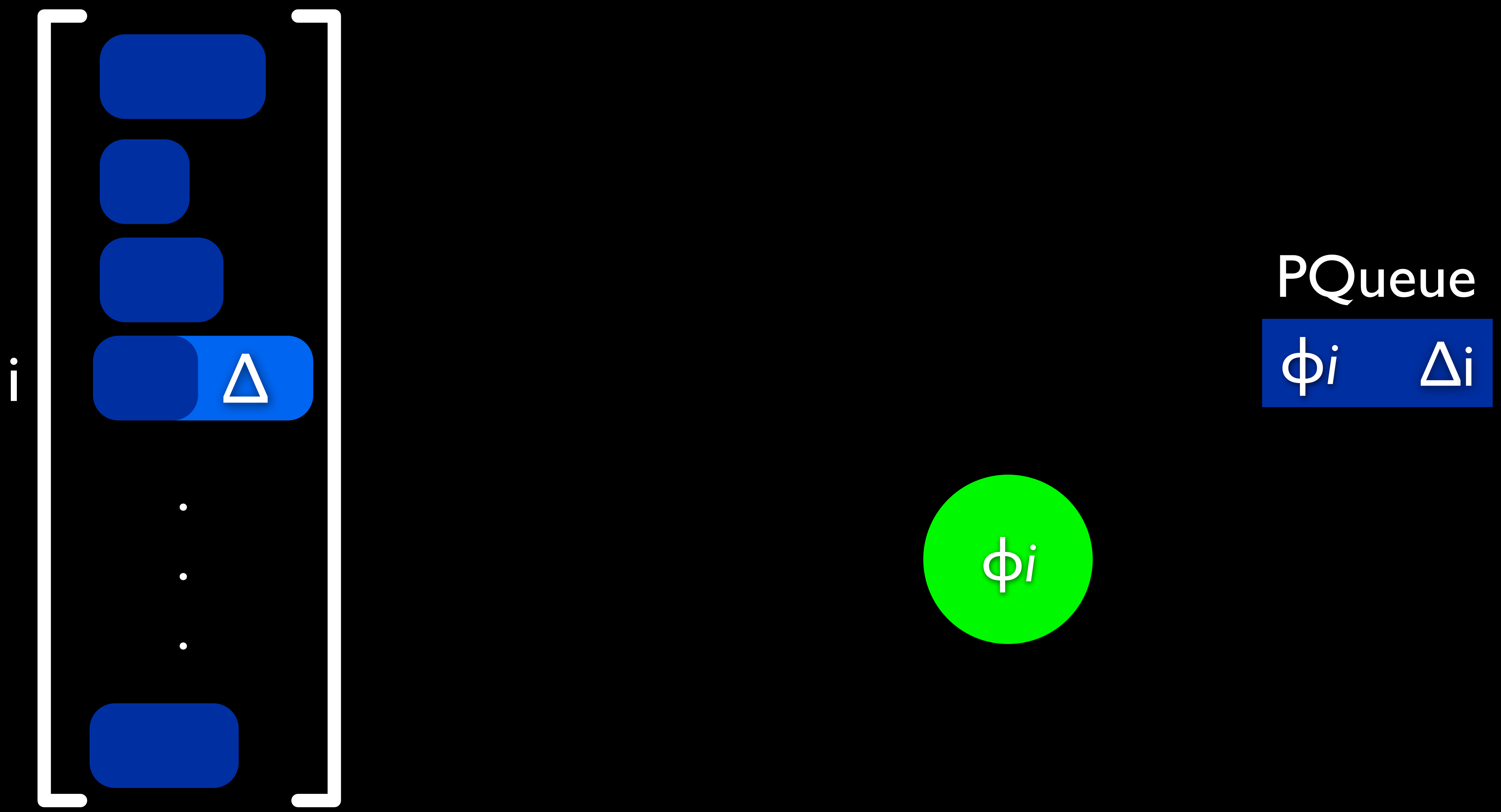
Dyna (MG)



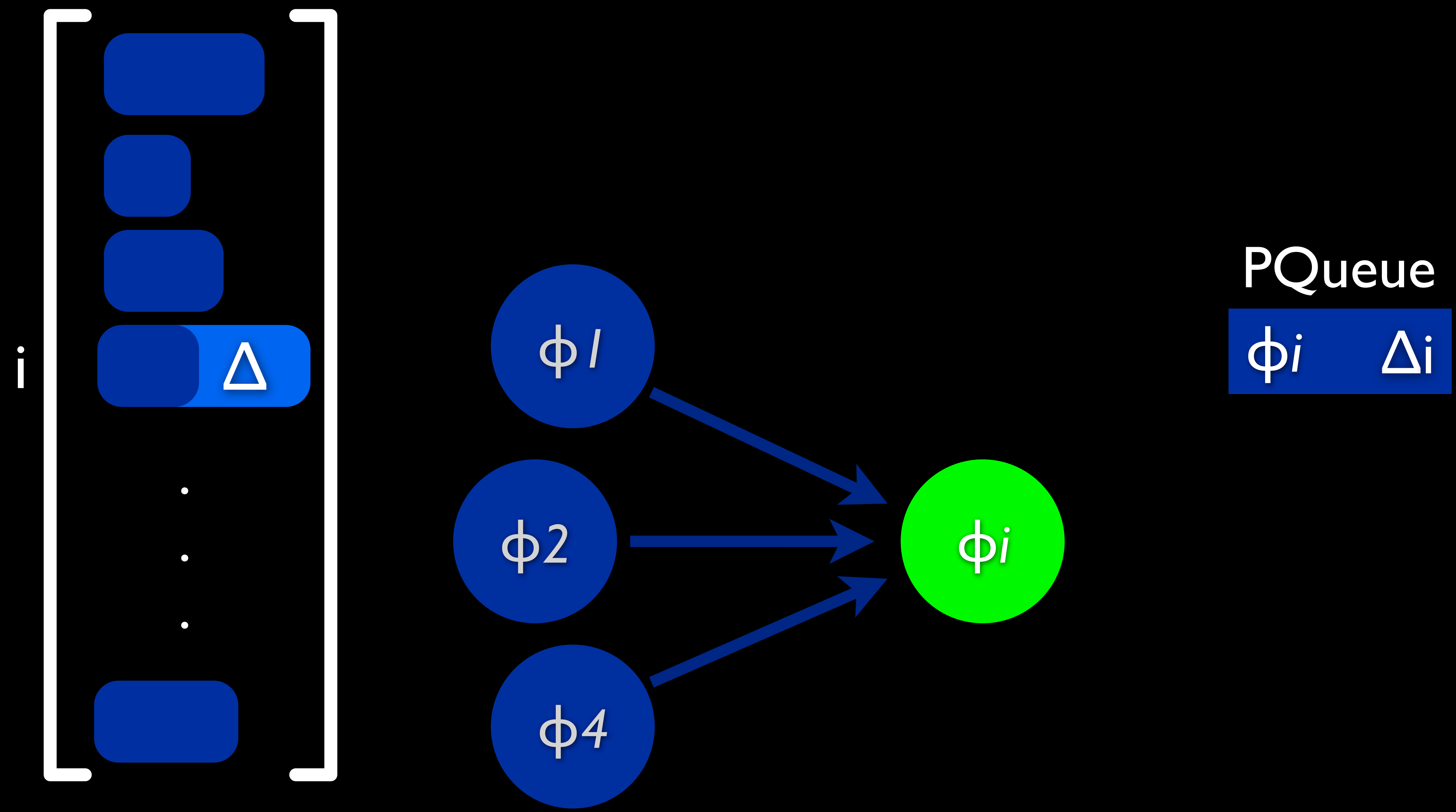
Dyna (MG)



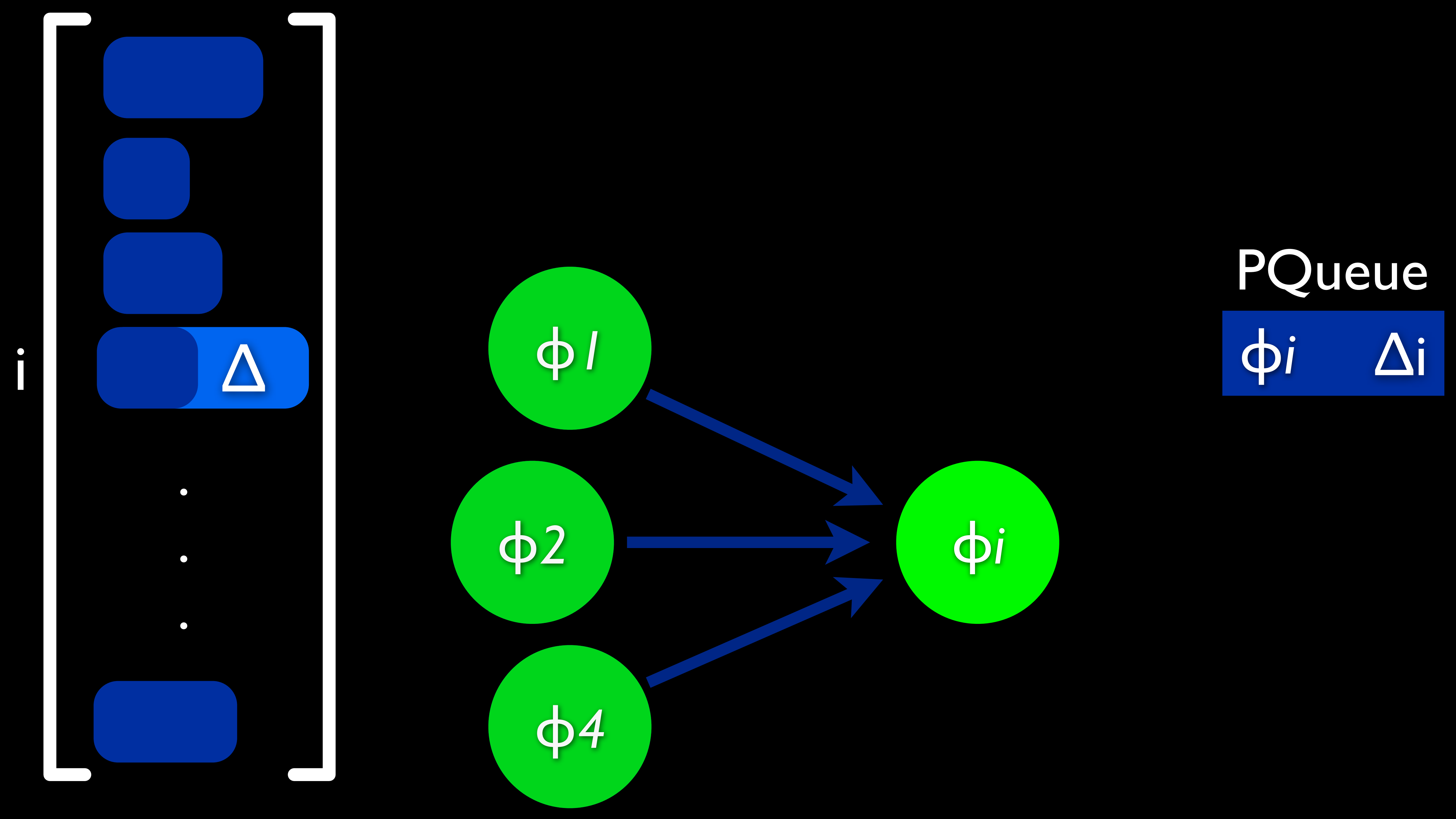
Dyna (MG)



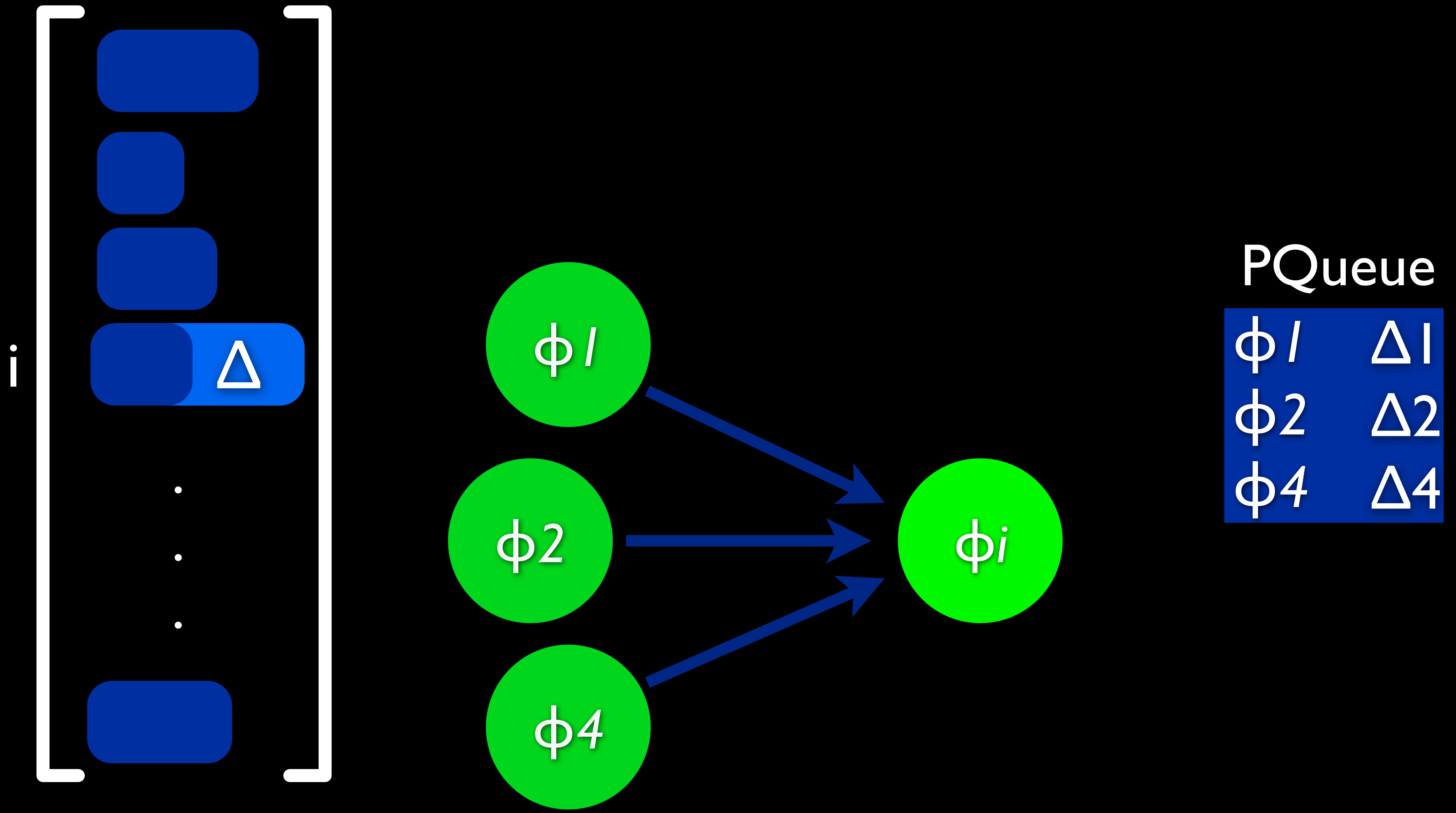
Dyna (MG)



Dyna (MG)



Dyna (MG)



Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

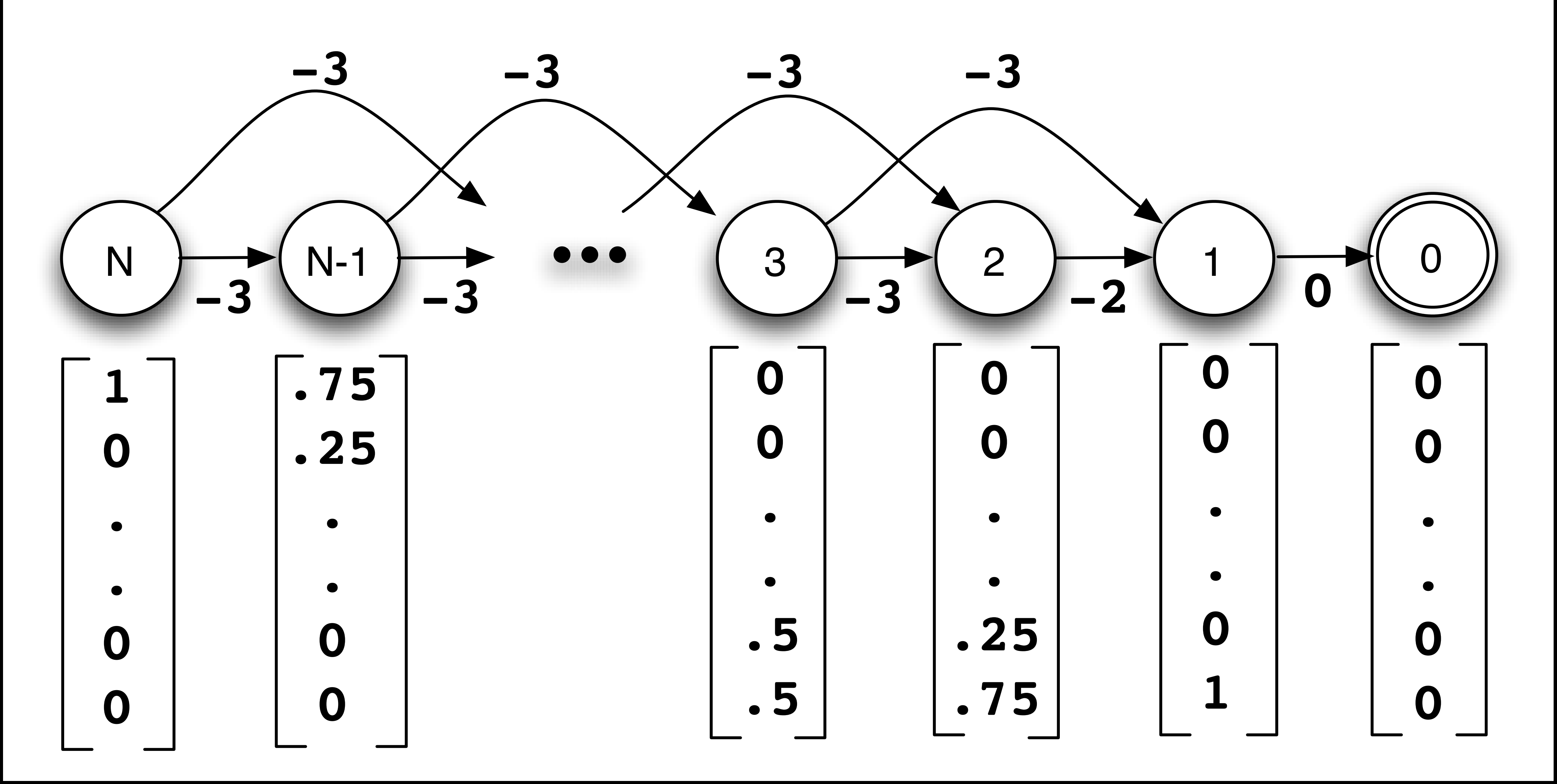
Empirical Results

Settings

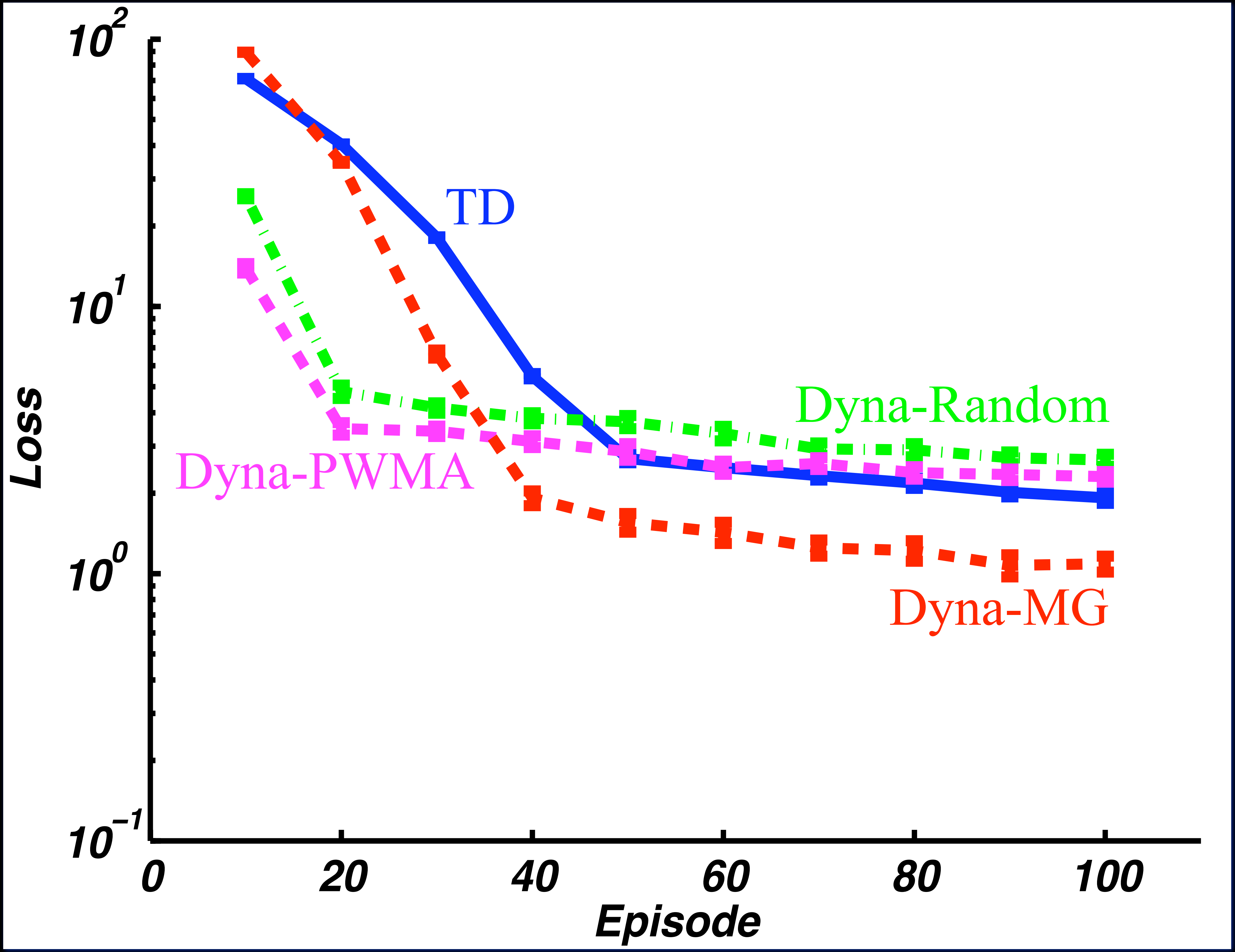
- 30 runs, same set of trajectories
- Best decay parameters in the set
- Results are shifted a bit

Boyan Chain (PE)

N = 98 states

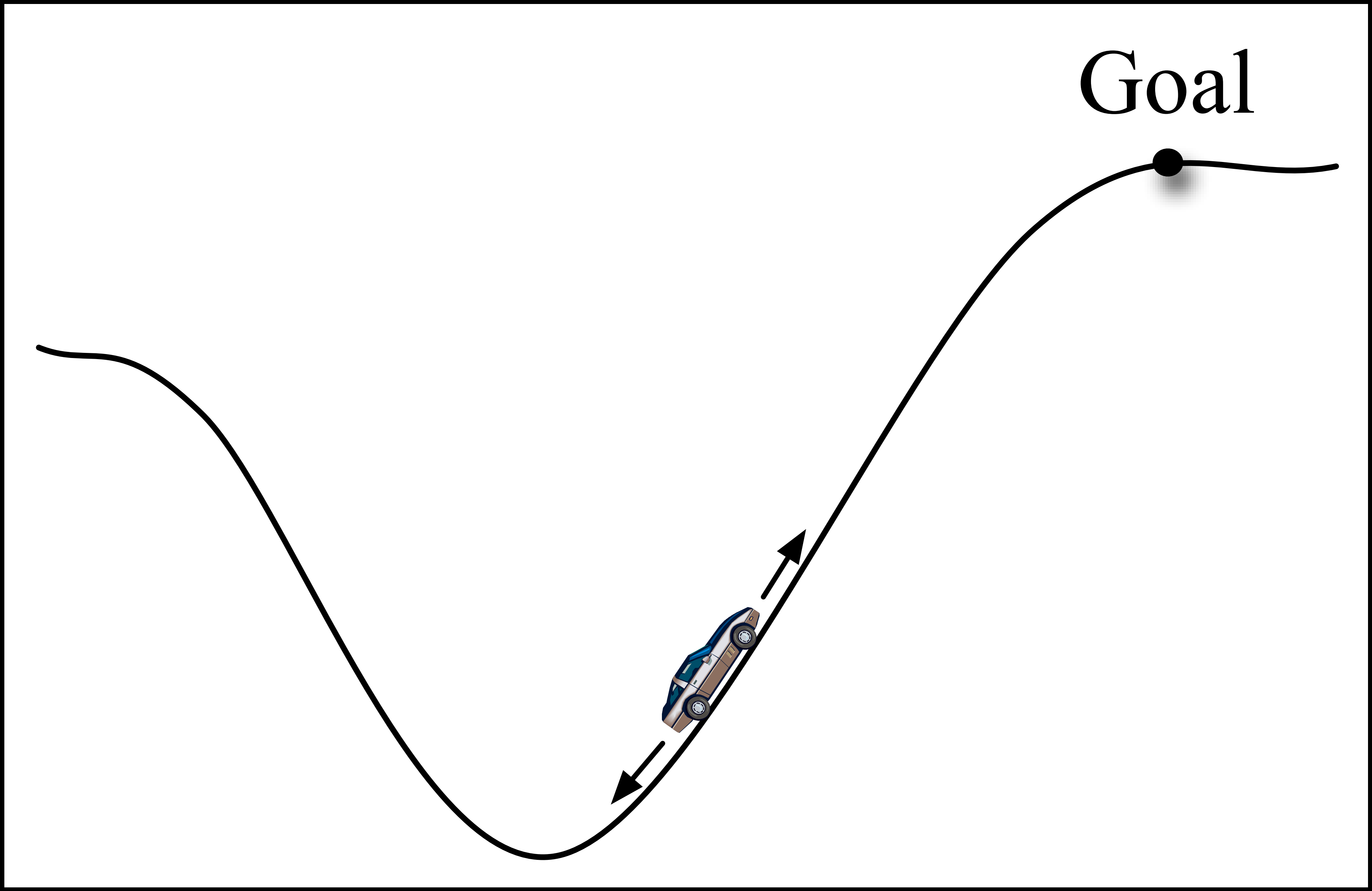


Boyan Chain (No control)

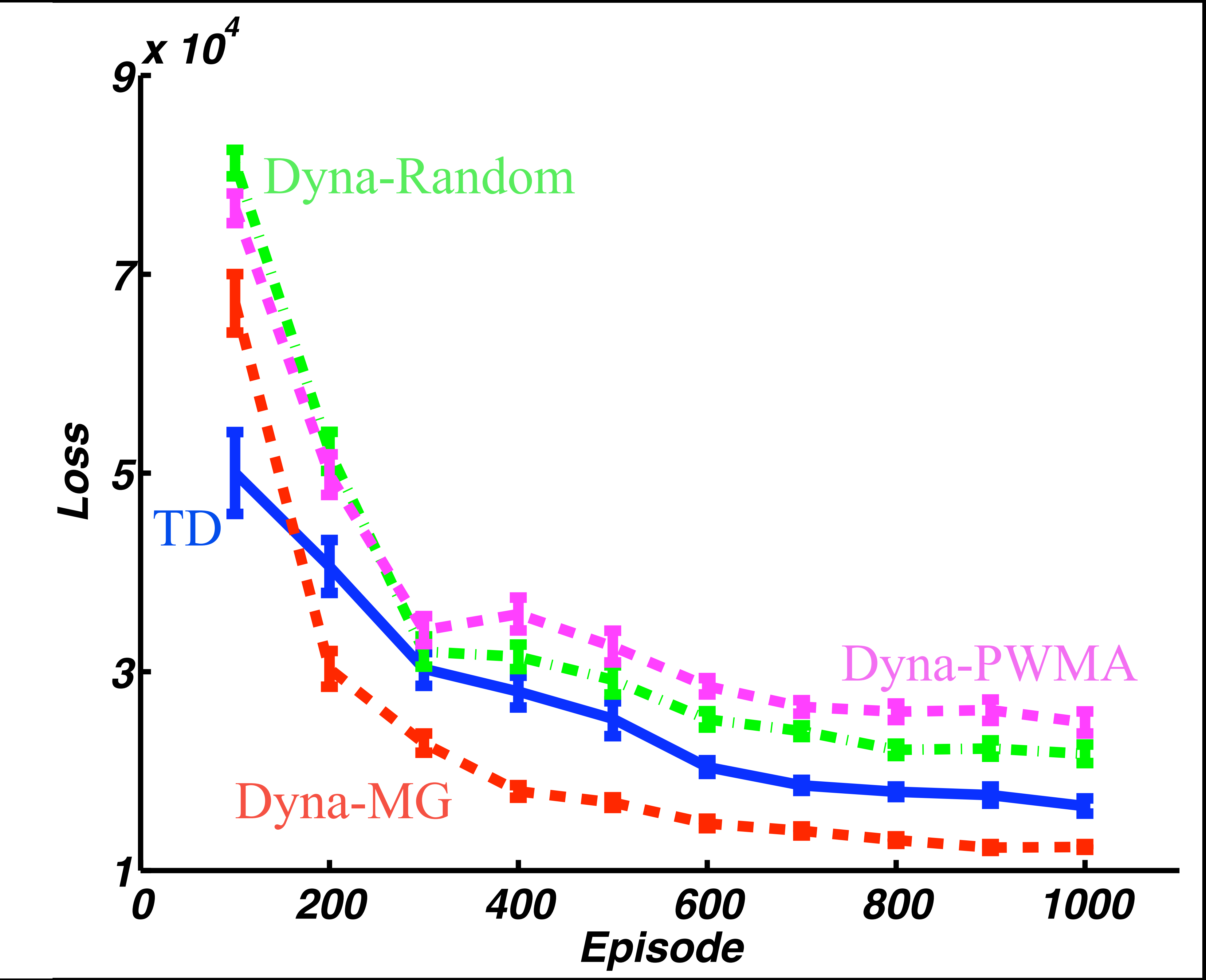


Mountain Car

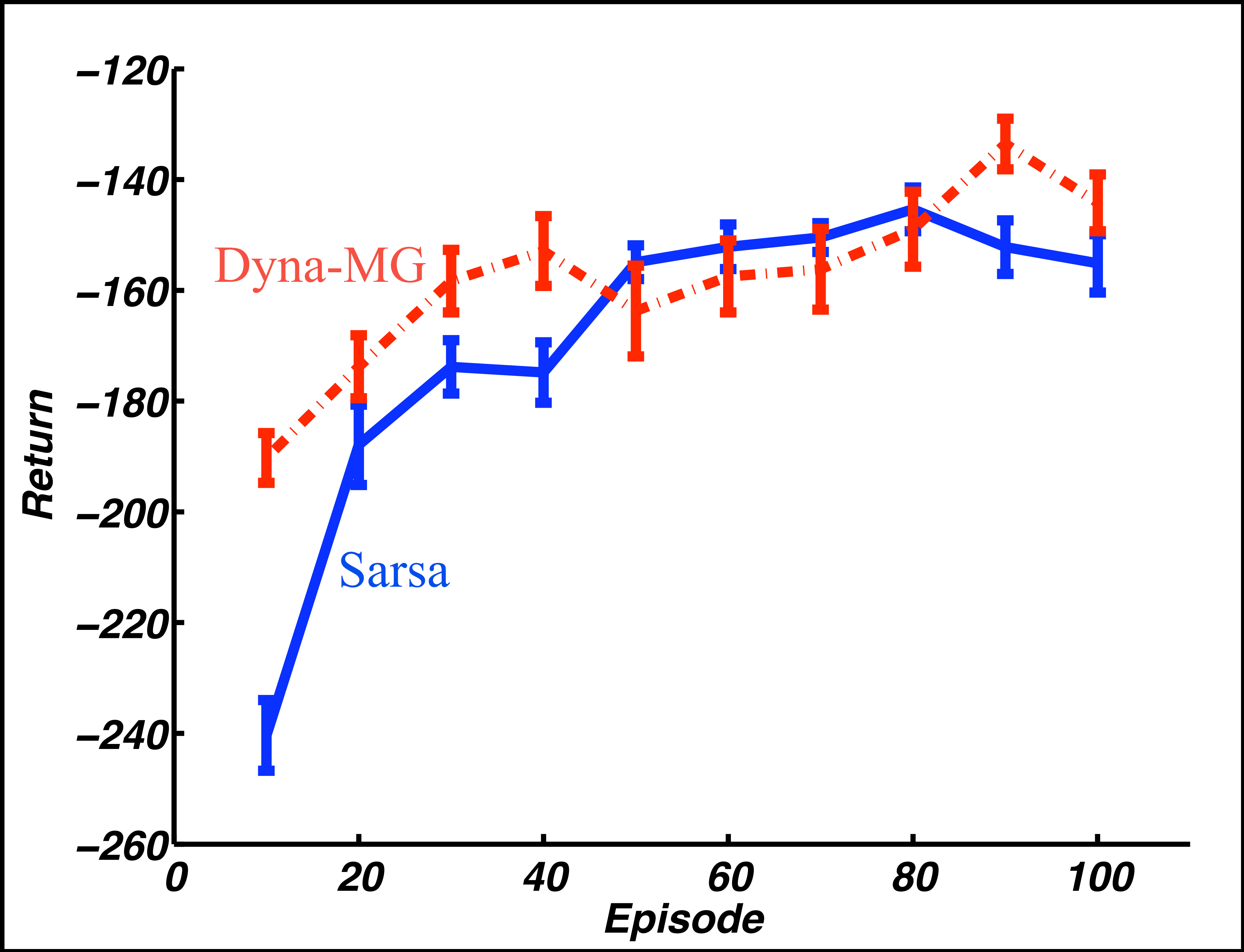
Tile coding (10000 tiles, 10 tillings)



Mountain Car (PE)



Mountain Car (Control)



Outline

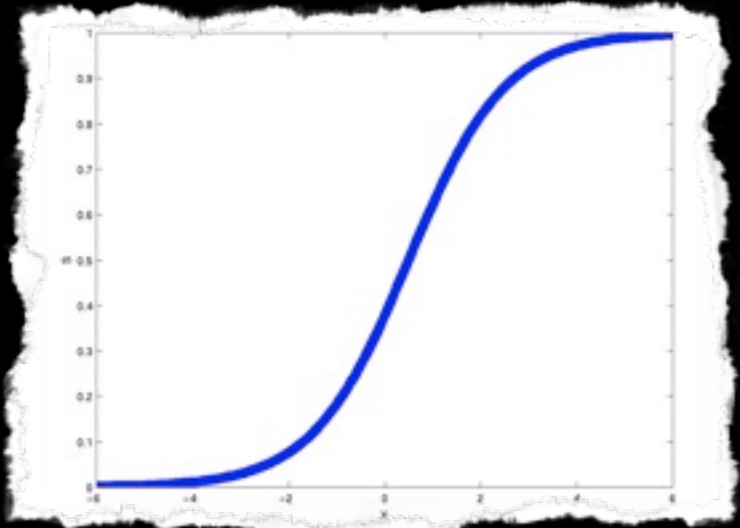
- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

Outline

- Background
- Linear Prioritized Sweeping
- Empirical Results
- Discussion

Discussion

- The whole story is nice, but does it really work? Theory?
- Geared towards sparse features
- Tracking ...
- Possibility: $\phi' \leftarrow \sigma(F\phi)$



Future Work

- More planning on larger problems
- Using Sigmoid function
- Correlation with SPPI and LSPI
- Convergence Proof

Questions?