



UNIVERSITY OF  
**PATRAS**  
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

# Imitation Learning in Super Mario Bros: Behavior Cloning and DAgger Using Privileged Information

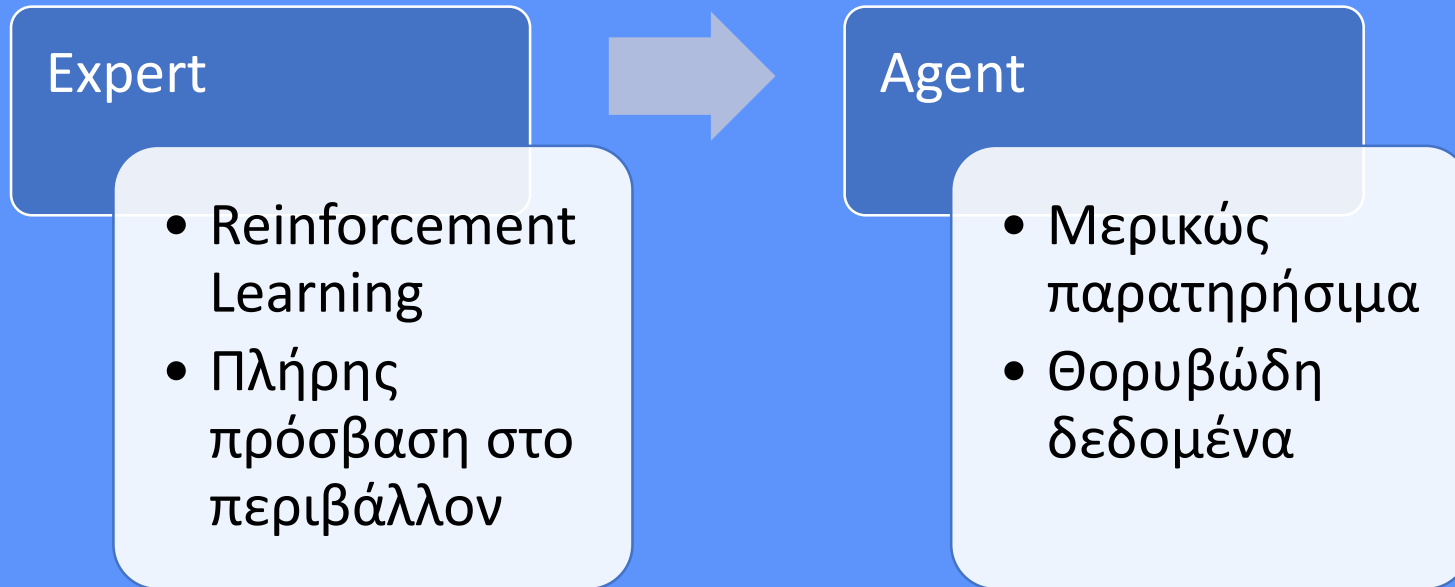
Μαρία - Νίκη Ζωγράφου

[up1096060@ac.upatras.gr](mailto:up1096060@ac.upatras.gr)

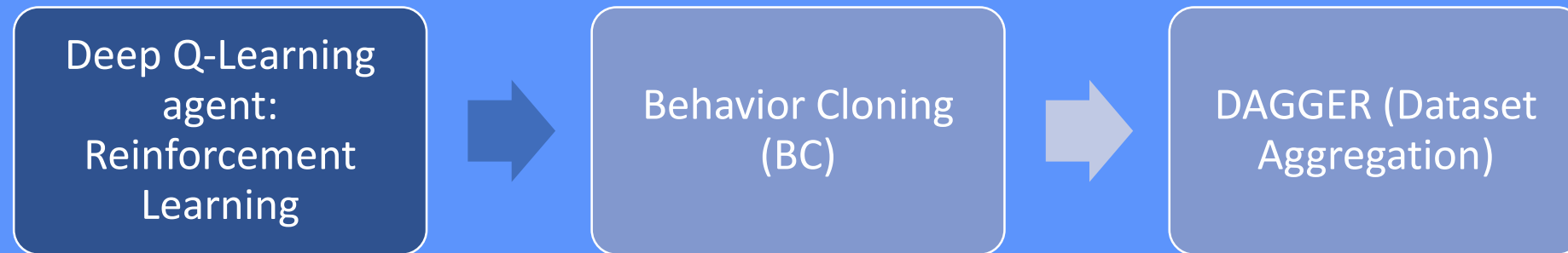
Νικόλαος Γέροντας

[up1092813@ac.upatras.gr](mailto:up1092813@ac.upatras.gr)

# Στόχος Εργασίας



# Εκπαίδευση των Agents



# Expert: Deep Q-Learning



## Q-learning:

$$Q(s, a) = Q(s, a) + \alpha * (r + \gamma \max_z Q(s', a') - Q(s, a))$$

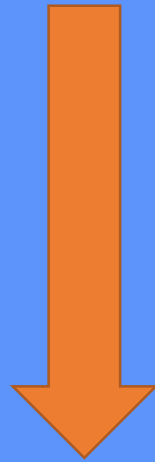
Q: Q-table

s: state

a: action

s': next state

a': action with max future reward



Αντικατάσταση Q-table  
με νευρωνικό δίκτυο

## Deep Q-learning:

Q – value function:  $Q(s, a; \theta)$ , όπου  $\theta$  οι παράμετροι του νευρωνικού

# Expert: Deep Q-Learning

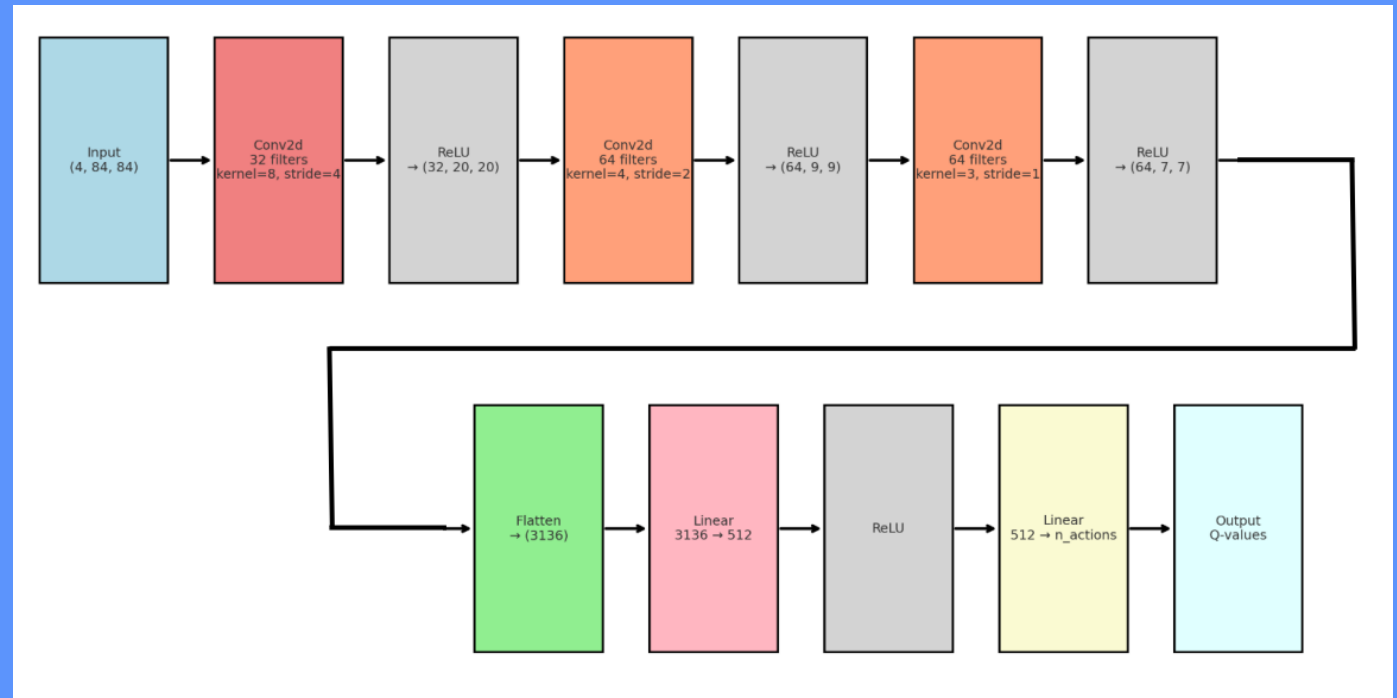
## Input

- 4 τελευταία frames
- μετατροπή σε grayscale
- 84x84 pixels
- frame skipping

## Reward Shaping

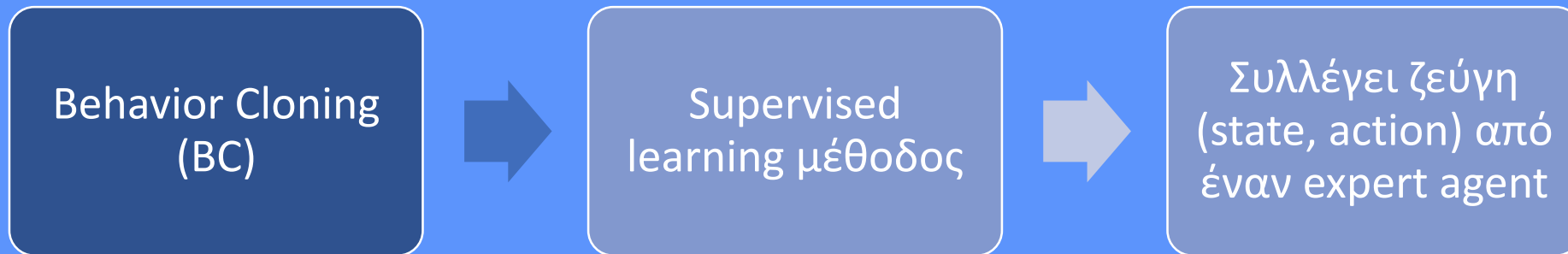
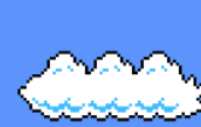
- Πρόσδος προς τα δεξιά +0.1
- Χρονικό penalty -0.1
- Ποινή για θάνατο -10
- Τερματισμός +100

Loss function: **MSE** Loss



Αρχιτεκτονική νευρωνικού

# Imitation Learning



Σύνηθες πρόβλημα:  
Φτάνει σε καταστάσεις  
που δεν έχει δει ο expert

# Dagger (Dataset Aggregation)



Ο DAGGER υπερβαίνει τους περιορισμούς του Behavior Cloning επειδή διορθώνει το distributional shift!

# Dagger με Behaviour Cloning Warm up

Dagger +  
Σύντομο στάδιο  
behavior cloning



Expert  
agreement



Επιτάχυνση  
σύγκλισης του  
Dagger!

Σταθεροποίηση εκμάθησης της πολιτικής  
με ~60 επαναλήψεις, σε αντίθεση με τις  
~1400 του «απλού» Dagger.



# Στατιστική Αξιολόγηση (final loss)

Plain DAGGER

- Median: 0.41
- 25%-75%: 0.34 - 0.42
- 10%-90%: 0.24 - 0.45
- Min-Max: 0.04 - 0.48



DAGGER + BC Warmup

**Improvement: -55.5%**

- Median: 0.18
- 25%-75%: 0.07 - 0.33
- 10%-90%: 0.05 - 0.38
- Min-Max: 0.04 - 0.42

STAGE COMPLETION:

-----  
Plain DAGGER: 70.0% (20 runs)

BC Warmup DAGGER: 100.0% (20 runs)

Improvement: +30.0%



Σας ευχαριστούμε για τον χρόνο σας!