# Advanced Topics in Deep Reinforcement Learning

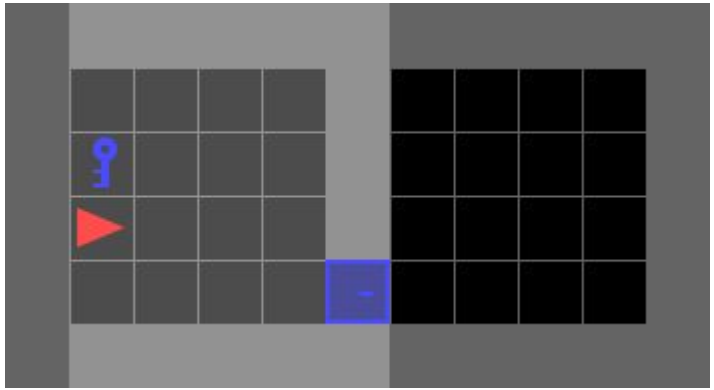
## *Curriculum Learning*

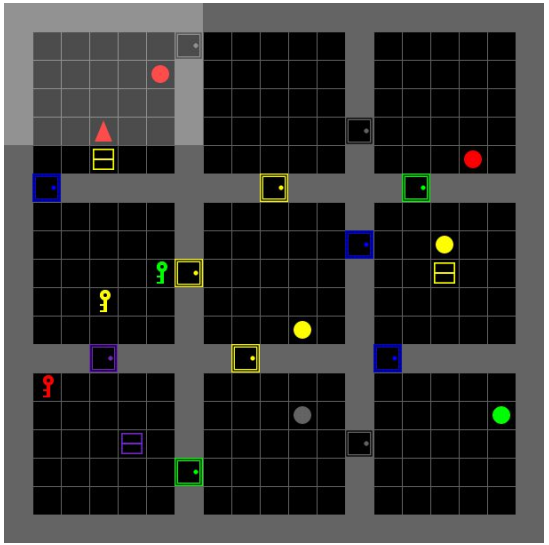
# Why Use Curricula?

- ML systems can benefit from ordering of the training data [Bengio et al. 2019]
- Common image: first learn easy, then hard
- In RL: learn skills on easy task version, the apply in harder ones

# Why Use Curricula?

- ML systems can benefit from ordering of the training data [Bengio et al. 2019]
- Common image: first learn easy, then hard
- In RL: learn skills on easy task version, the apply in harder ones

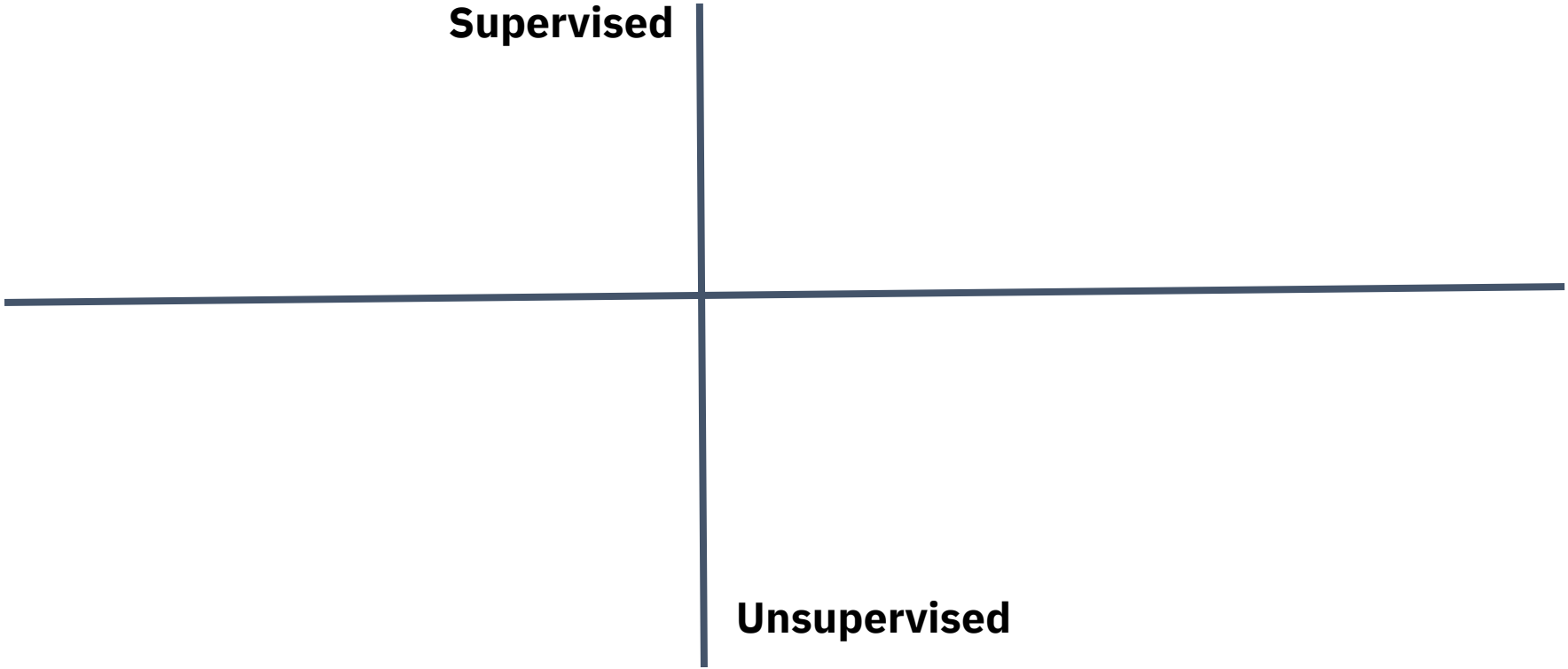




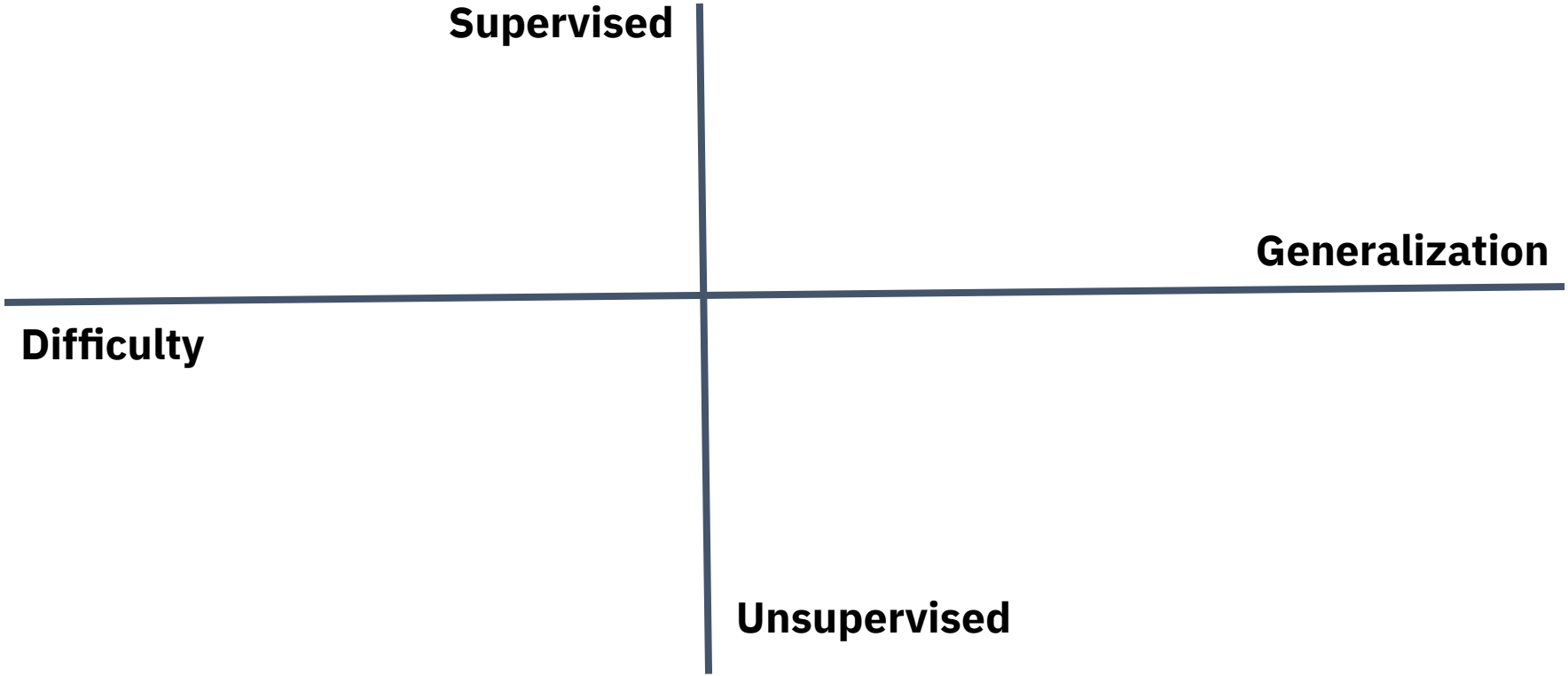
# Reminder: cMDPs

- Context varies reward and transition functions
- Example context: gravity in physics simulation
- Usually defines a generalization task
- Can in CL also be used with the objective to perform well on only the hardest contexts

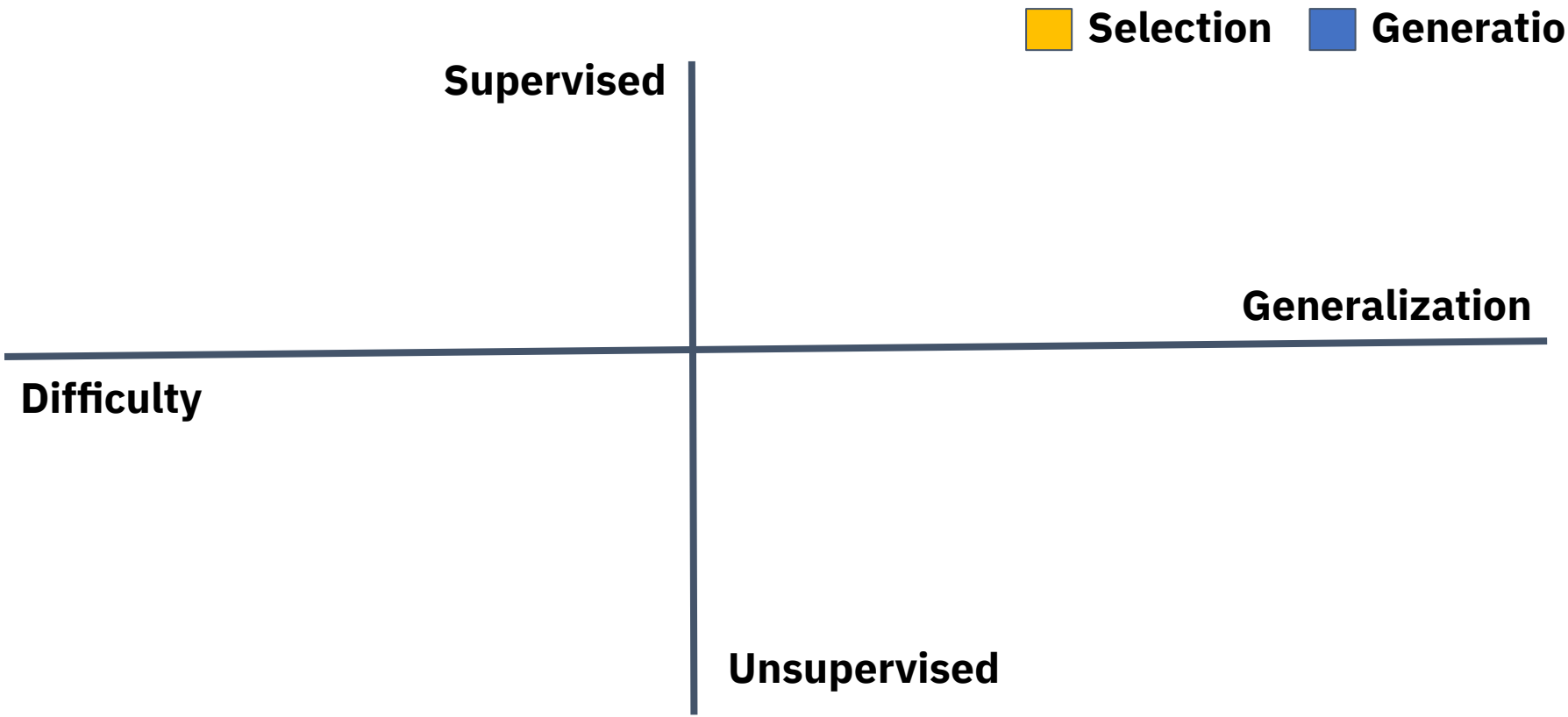
# Curriculum Learning Methods



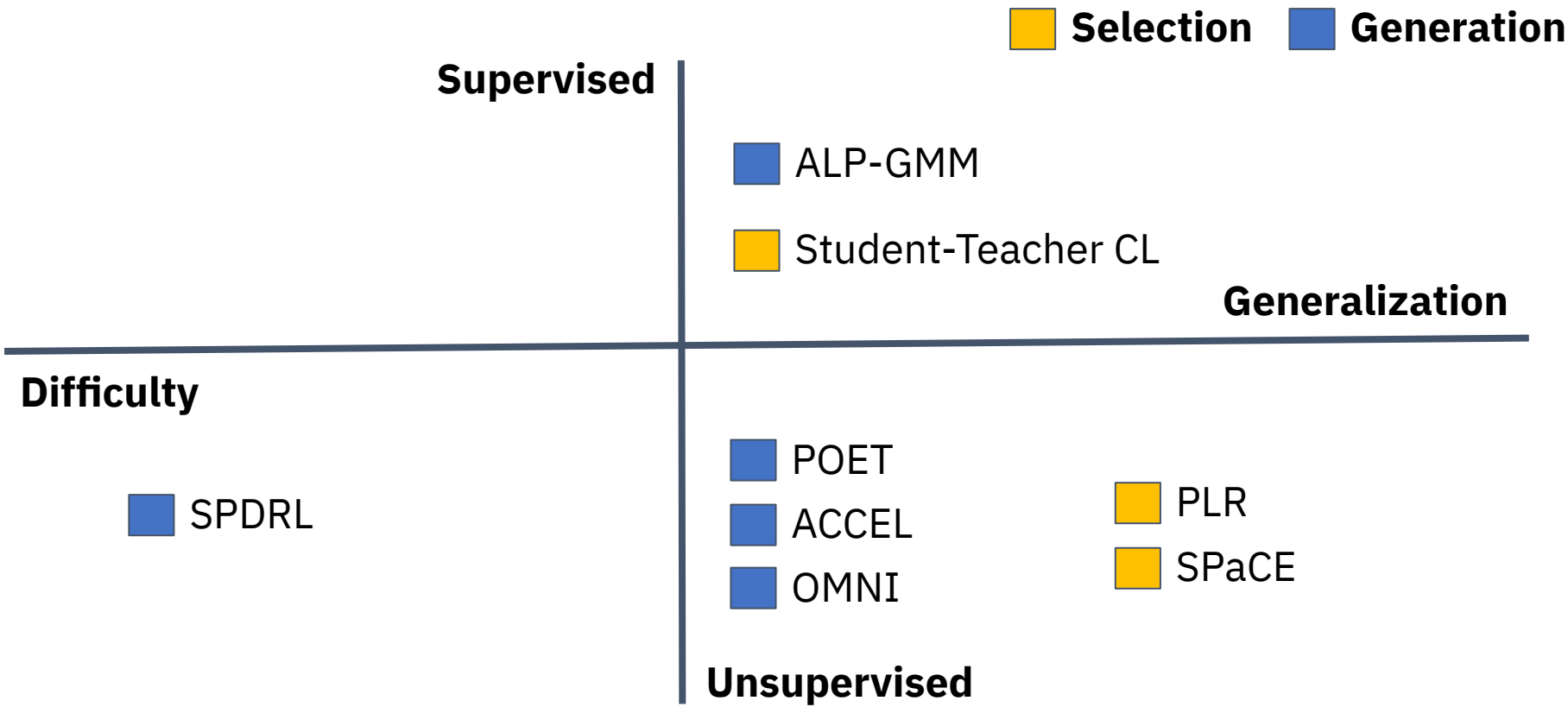
# Curriculum Learning Methods



# Curriculum Learning Methods



# Curriculum Learning Methods

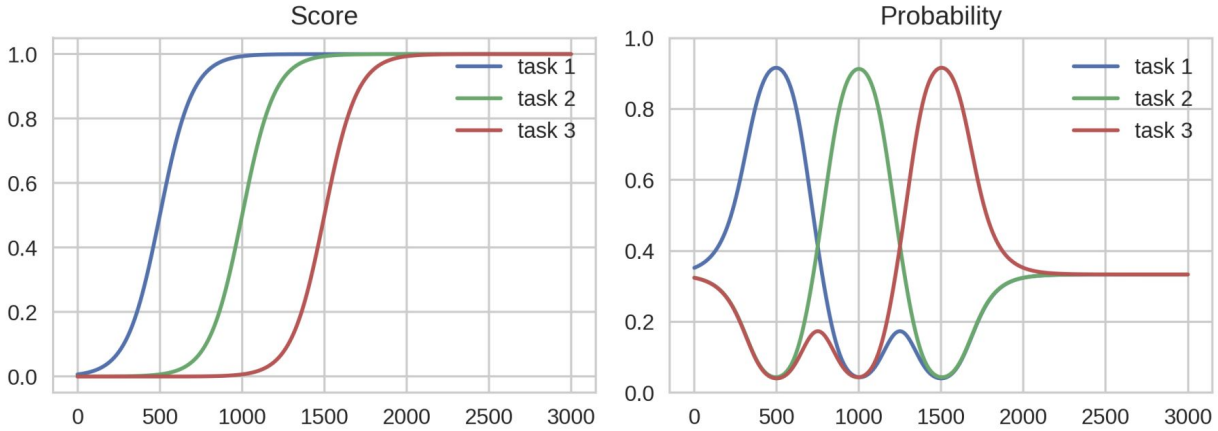


# Supervised Context Selection [Matiisen et al. 2017]

- Idea: Teacher picks next task to train on
- Modelled as a hierarchical RL task
- Teacher reward is student return
- Scales poorly to large amounts of contexts

# Supervised Context Selection [Matiisen et al. 2017]

- Idea: Teacher picks next task to train on
- Modelled as a hierarchical RL task
- Teacher reward is student return
- Scales poorly to large amounts of contexts



# Supervised Context Generation

- Train a generation mechanism to generate new contexts instead of picking from existing ones
- Problem: Shifting target
- Example: ALP-GMM [Portelas et al. 2019]
  - fits a Gaussian Mixture Model to generate contexts
  - to ensure sampling in spaces with a lot of progress, use nearest neighbour for each point to get $alp_{new} = |r_{new} - r_{old}|$

# Unsupervised Context Selection

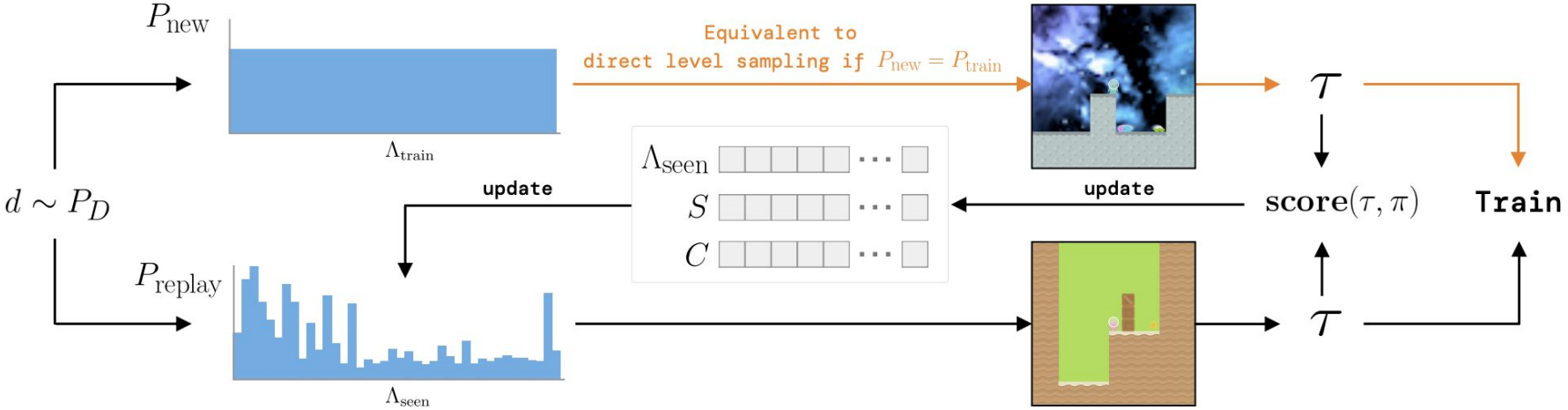
- Popular paradigm with different flavors
- Better scaling than supervised
- Often easy to implement
- Can utilize agent statistics for selection

# Prioritized Level Replay [Jiang et al. 2021]

- Keeps a level buffer similar to a replay buffer
- Training tasks are sampled from that buffer
- Prioritized by last value loss
- Robustified version through diversity [Jiang et al. 2022]
- Likely still state of the art overall

# Prioritized Level Replay [Jiang et al. 2021]

- Keeps a level buffer similar to a replay buffer
- Training tasks are sampled from that buffer
- Prioritized by last value loss
- Robustified version through diversity [Jiang et al. 2022]
- Likely still state of the art overall



# SPaCE [Eimer et al. 2021]

- Similar to PLR, but instead of keeping a buffer of losses, we do a single forward pass over all contexts
- We then train on tasks with largest change
- Similar idea to ALP-GMM but using value predictions
- No buffer needed, so added flexibility
- But: relies on consistency of predictions which might make it less reliable than PLR

# Unsupervised Context Generation

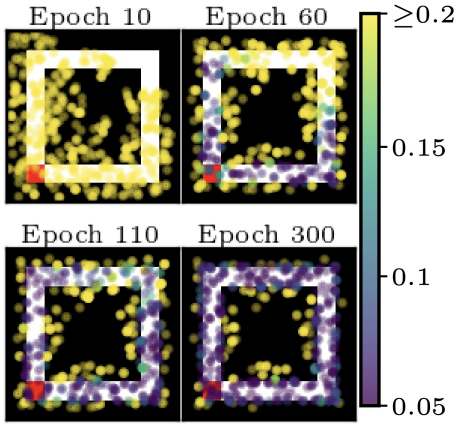
- Also: unsupervised environment design
- Generates context according to heuristic
- Requires context generation ability

# Self-Paced Deep RL [Klink et al. 2022]

- Goal: solve very hard tasks
- Curriculum transitions from hard to easy gradually
- Forgetting early instances and generalization don’t matter
- Originally: CL through shifting of context distribution mean [Klink et al. 2020]
- Updated version uses optimal transport for distribution interpolation

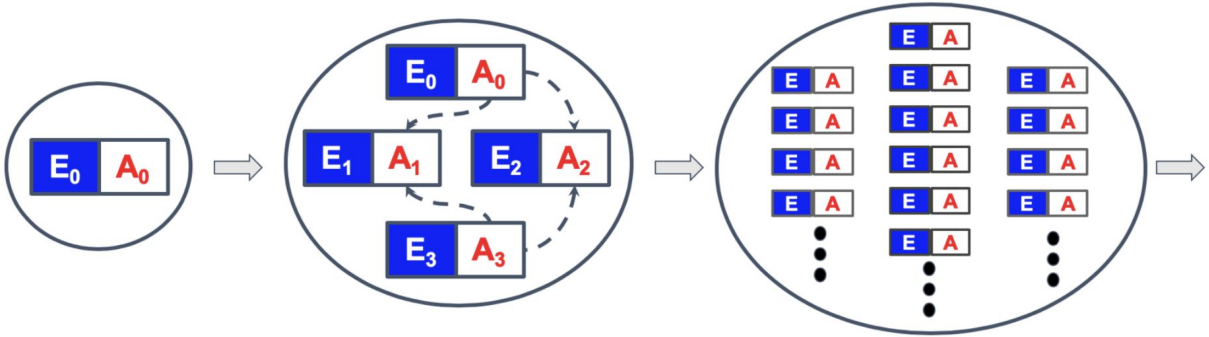
# Self-Paced Deep RL [Klink et al. 2022]

- Goal: solve very hard tasks
- Curriculum transitions from hard to easy gradually
- Forgetting early instances and generalization don’t matter
- Originally: CL through shifting of context distribution mean [Klink et al. 2020]
- Updated version uses optimal transport for distribution interpolation



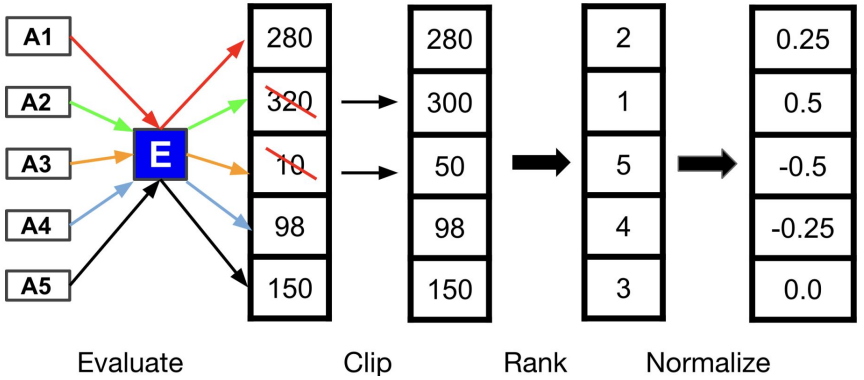
# POET [Wang et al. 2020]

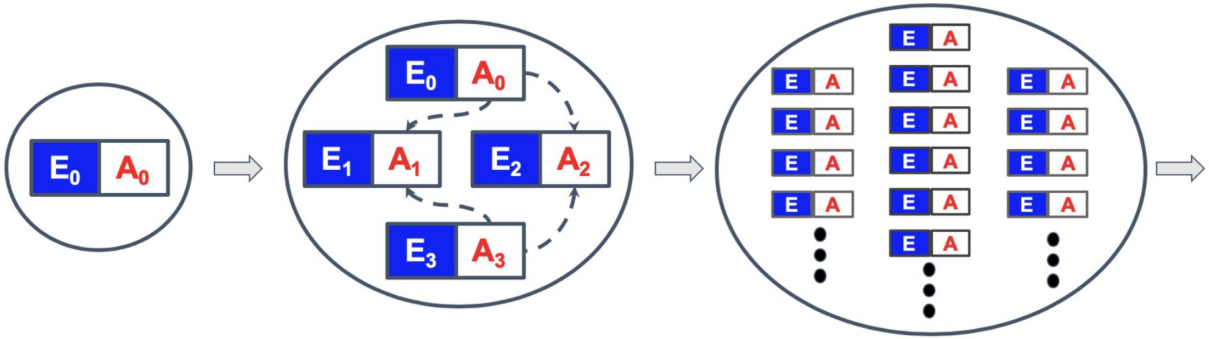
- Co-evolves agents and context
- Prioritizes novel contexts
- Open-ended training

# POET [Wang et al. 2020]

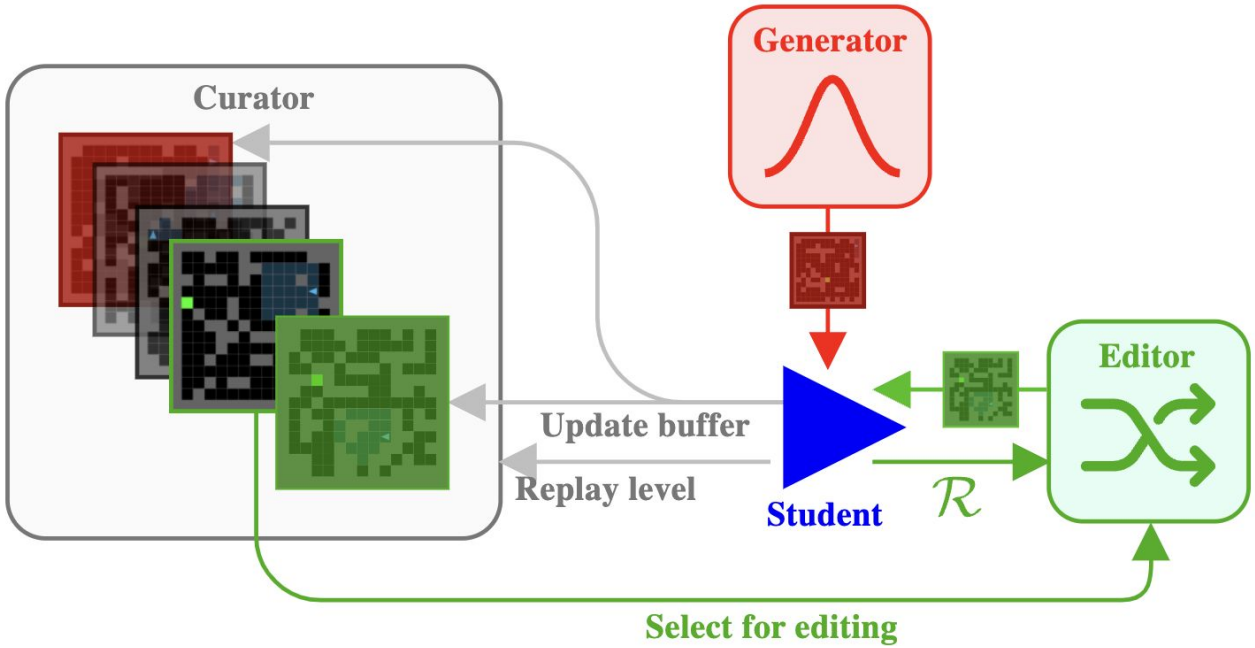
- Co-evolves agents and context
- Prioritizes novel contexts
- Open-ended training

# ACCEL [Parker-Holder et al. 2022]

- Mixture of POET and PLR: keep buffer of context which are evolved over time
- Has been used in algorithm discovery [Jackson et al. 2023]



# OMNI [Zhang et al. 2024]

- Context sampling via language
- Uses environment and task descriptions to measure “interestingness”
- Interestingness score weights the evaluation performance
- Goal: move towards tasks interesting to humans
- Challenge: context sampling becomes more and more involved

# Curricula In Other RL Paradigms

- CL is very online model-free centric
- There's multi-agent version of PLR
- In principle possible for model-based and offline learning as well

# Limitations & Caveats

- Completely on-policy focused since it’s not clear how replay buffers interact with task ordering
- Some findings indicate little benefit from elaborate curricula (see SPaCE)
- In supervised learning: sometimes curricula simply compensate for bad hyperparameter configurations [Weber et al. 2023]
- We have seen similar results in a thesis before for RL (not well explored)

# My Understanding of Curriculum Learning

- I understand why CL can be useful in RL
- I can name one curriculum learning method
- I know the categorizations for CL methods in RL
- Given a CL method, I could characterize it
- I can compare a CL method from each category we discussed
- I know at least one CL method well enough to be able to implement it