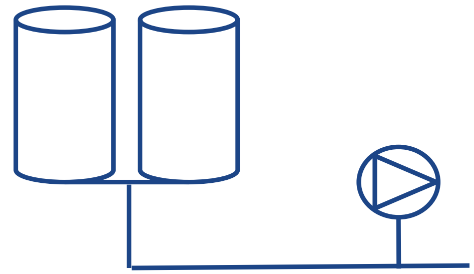
Deep Reinforcement Learning through Imitation Learning and Curriculum Learning: Application to Pump Scheduling in Water Distribution Networks

Henrique Donâncio, Laurent Vercouter

Autonomous Control for Water Distribution Systems

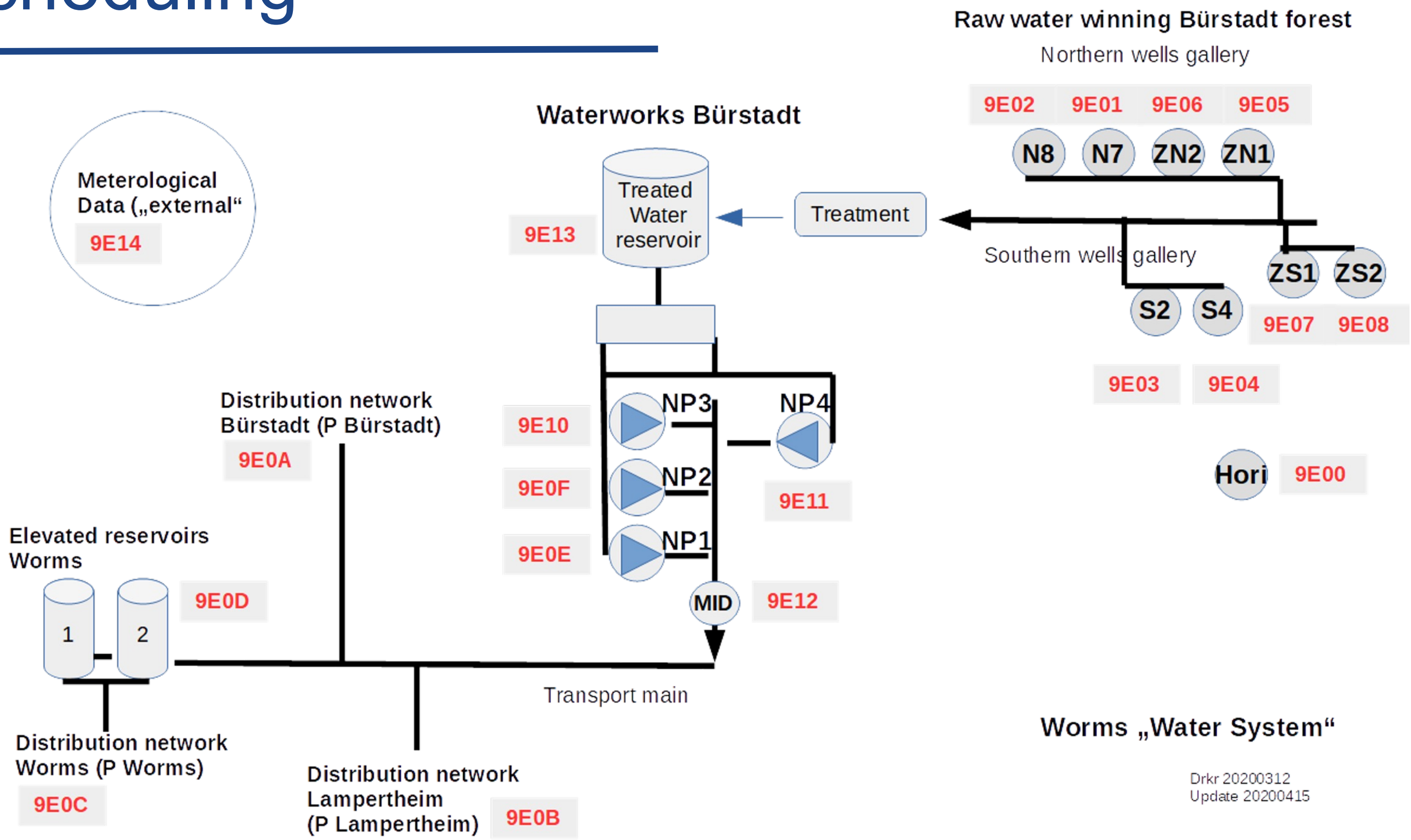
- IoT.H2O Project:
  - TU Kaiserslautern, ULiège, UFMG, Dr. Kraetzig
- Water Distribution System
  - Germany, logged data, and simulator

Water Distribution Systems



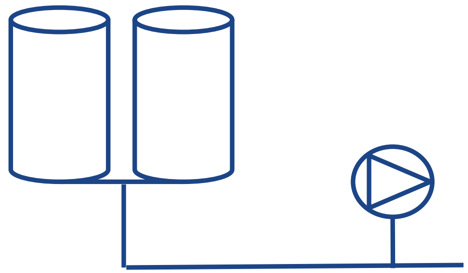
The Pump Scheduling Problem

Pump Scheduling



Raw data format: **9E0C4f77821f0222**
Uinit_ID: 9E0C
Epoch: 4f77821f (decimal: 1333232159)
Pressure: 0222 (decimal: 546 -> 5,46 bar)

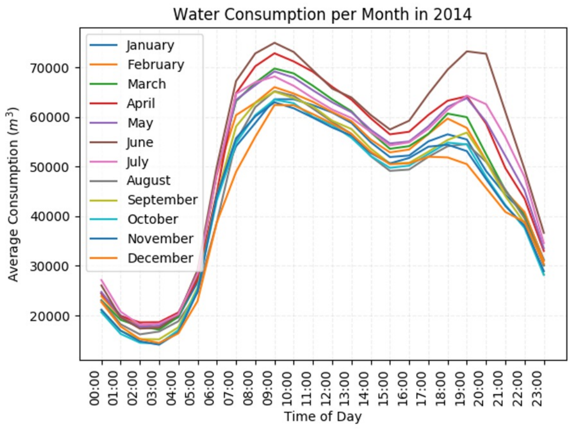
Pump Scheduling



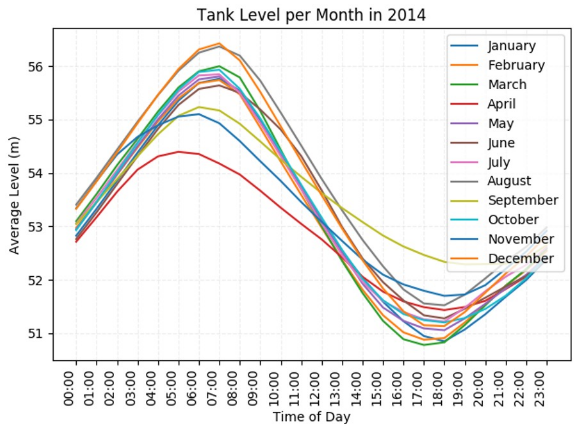
- Which pressure is necessary to deliver a certain amount of water into the system?
  - Water demand has to be delivered
  - Storage tanks must not overflow or run out of water
  - A minimum water reserve has to be in the tanks
  - A minimum pressure must be guaranteed in the pipe network
  - Pumps must be operated efficiently
  - Guarantee water exchange in tanks

Pump Scheduling

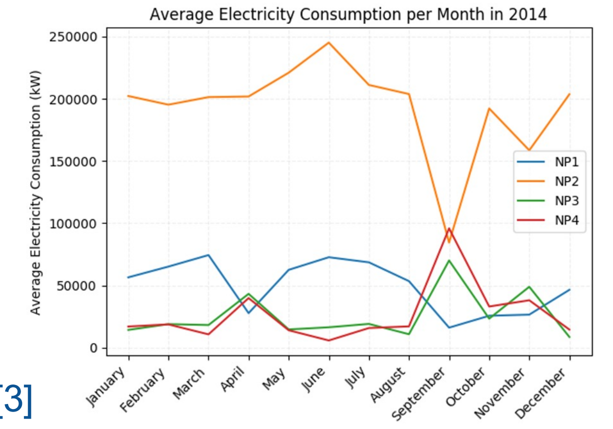
[1]



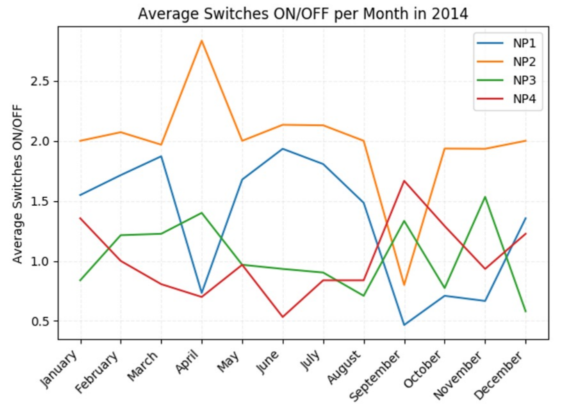
[2]



[3]

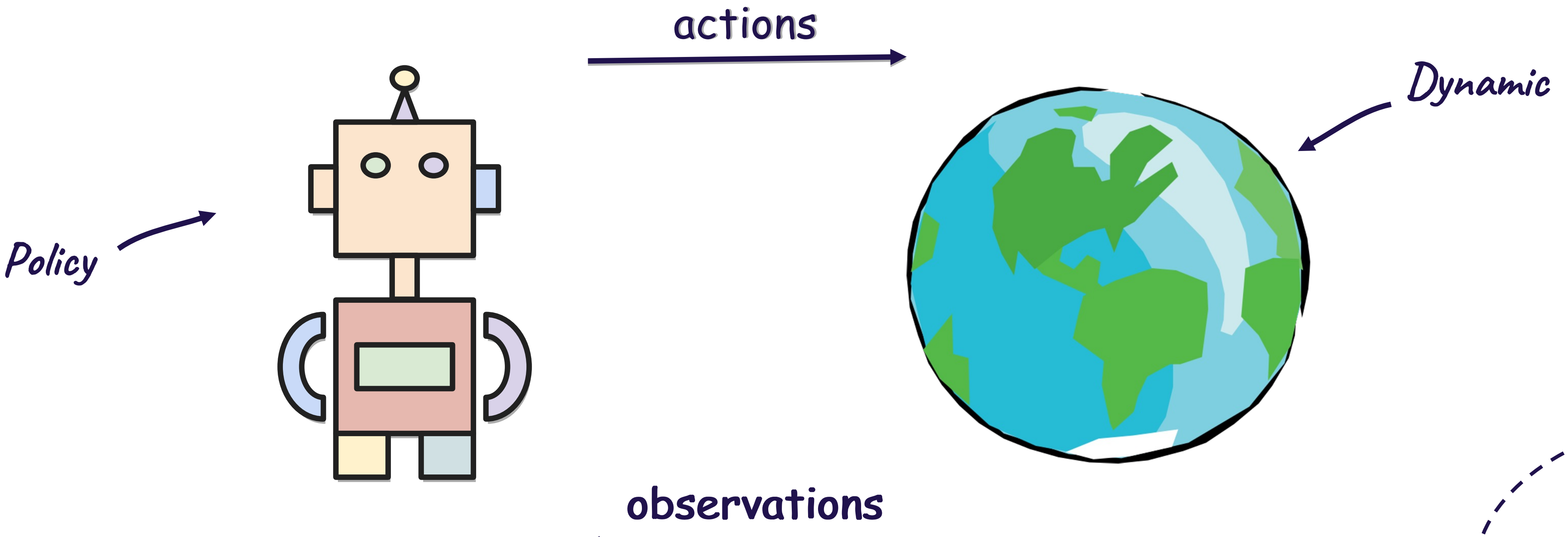


[4]

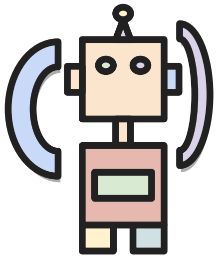


Autonomous control

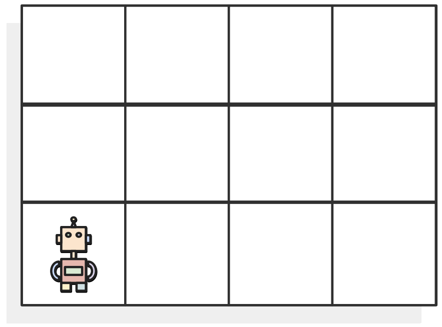
Autonomous Control



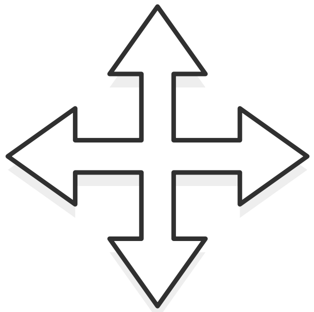
Markov Decision Process (MDP)



Andrey Markov



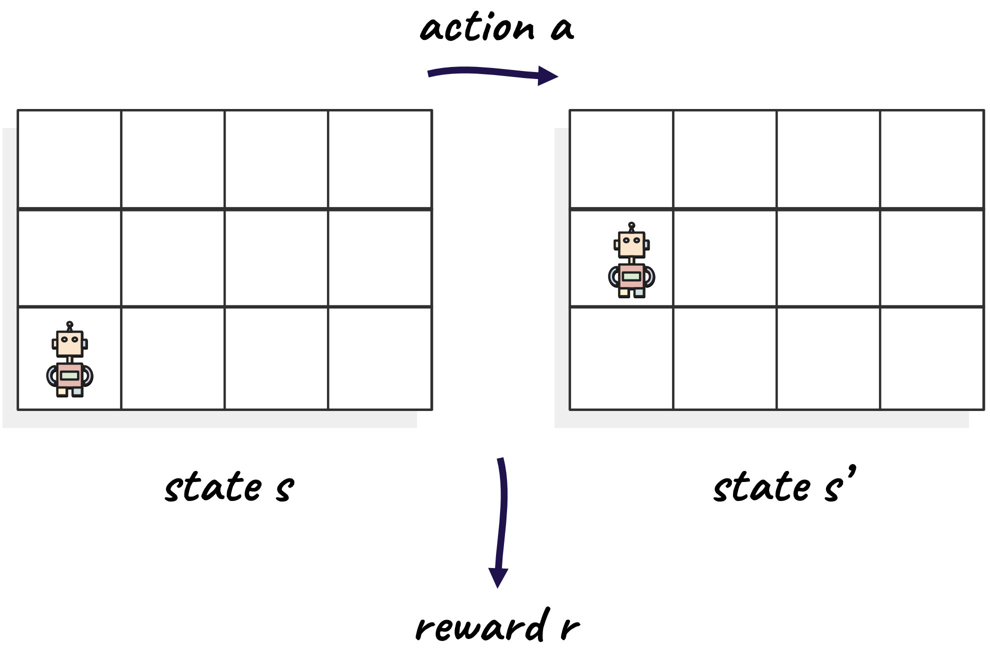
States S
Observations O



Actions A

Reward Function R

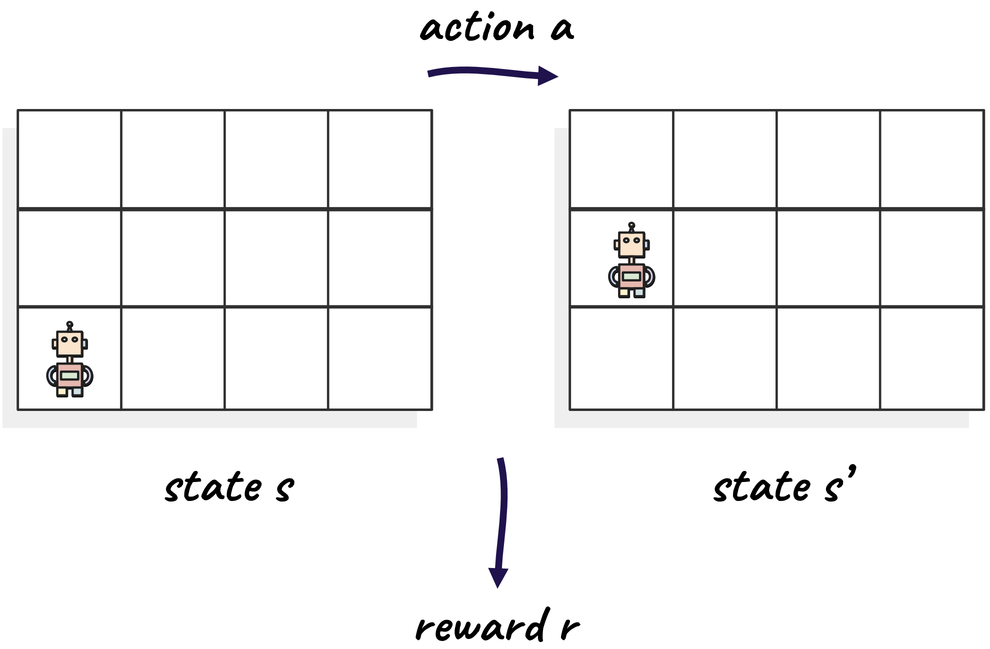
Markov Decision Process (MDP)



Andrey Markov

$$\text{returns} = r(0) + \gamma^1 r(1) + \gamma^2 r(2) \ldots$$

Markov Decision Process (MDP)



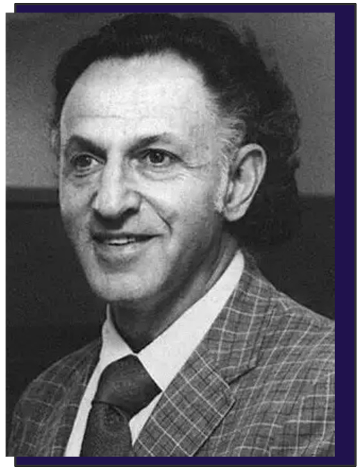
Andrey Markov

The MDP defines an RL problem!

Q-Learning [Watkins & Dayan (1992)]

Q(s,a) How good is being in a state *s* and performing an action *a*?

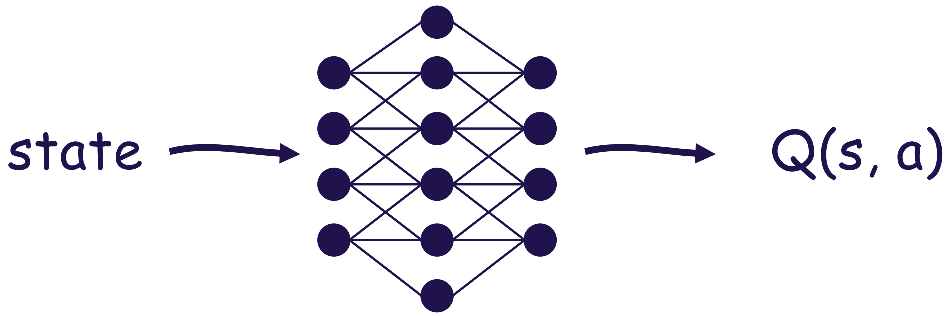
$$Q(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q(s',a') \right]$$



Richard Bellman

Reinforcement Learning (RL) [Sutton & Barto. 2018]

Dealing with high dimensional state spaces!



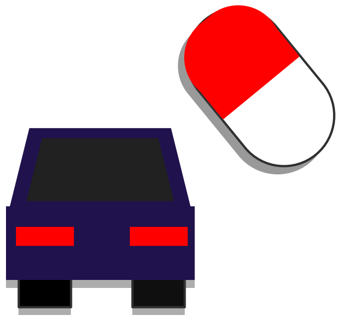
Deep Q-Networks [Mnih et at. 2013]

RL in Real-World [Dulac-Arnold et al. 2017]

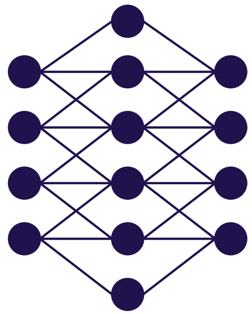
Observations? 

Rewards? 

Define an RL problem

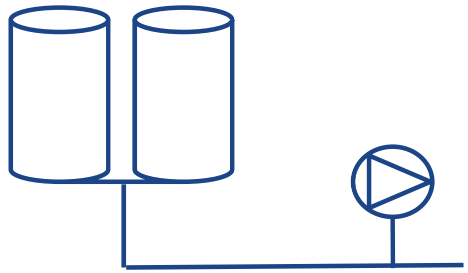


Experiences



Learning

Pump Scheduling: POMDP

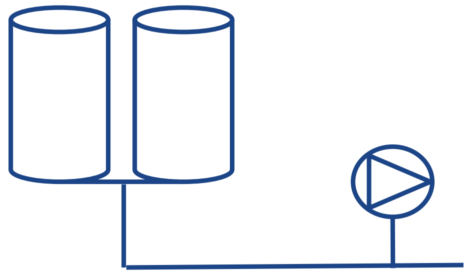


- **State:** < Tank level, Water Consumption, Time of Day, Month, Last Action, Time Pumping, Quality >
- **Actions:** {NP1, NP2, NP3, NP4, NOP}
- **Reward:**

$$r_t = e^{1/(-Q_t/kW_t)} - B * \psi + \log(1/(P + \omega)) \ (1)$$

$$r_t = -e^{(-1/kW_t)} - B * \psi + \log(1/(P + \omega)) \ (2)$$

Pump Scheduling: POMDP

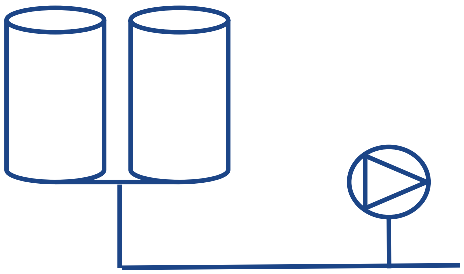


Efficiency

$$r_t = \boxed{e^{1/(-Q_t/kW_t)}} - B * \psi + \log(1/(P+\omega)) \quad (1)$$

$$r_t = -e^{(-1/kW_t)} - B * \psi + \log(1/(P+\omega)) \quad (2)$$

Pump Scheduling: POMDP

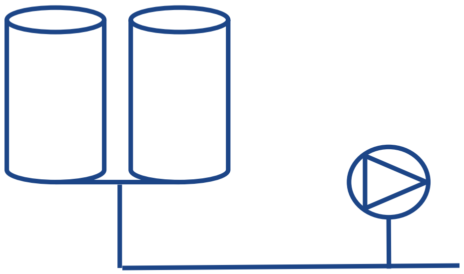


Safety
Constraints

$$r_t = e^{1/(-Q_t/kW_t)} \boxed{- B * \psi} + \log(1/(P + \omega)) \quad (1)$$

$$r_t = -e^{(-1/kW_t)} - B * \psi + \log(1/(P + \omega)) \quad (2)$$

Pump Scheduling: POMDP

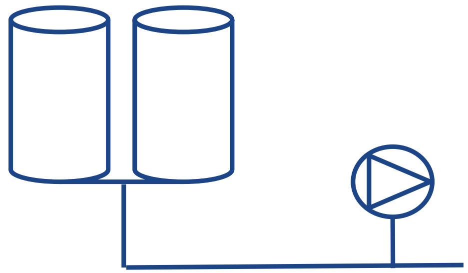


Pump Use/Switch

$$r_t = e^{1/(-Q_t/kW_t)} - B * \psi + \boxed{\log(1/(P + \omega))} \quad (1)$$

$$r_t = -e^{(-1/kW_t)} - B * \psi + \log(1/(P + \omega)) \quad (2)$$

Pump Scheduling: POMDP

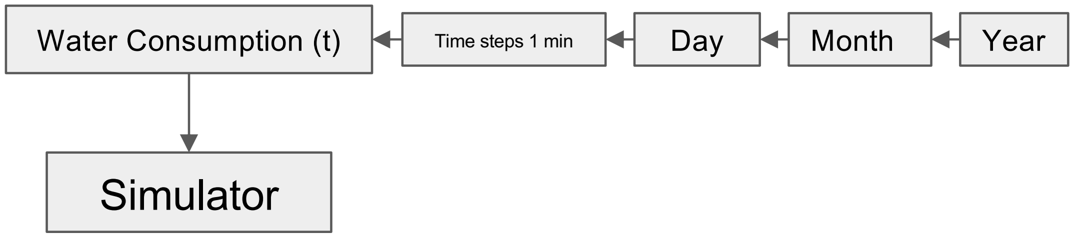


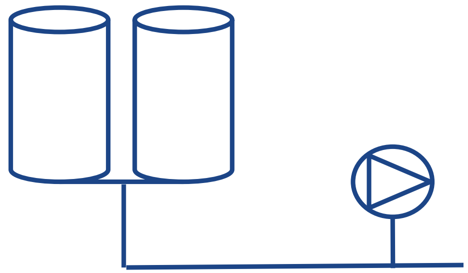
$$r_t = e^{1/(-Q_t/kW_t)} - B * \psi + \log(1/(P + \omega)) \quad (1)$$

$$r_t = \boxed{-e^{(-1/kW_t)}} - B * \psi + \log(1/(P + \omega)) \quad (2)$$

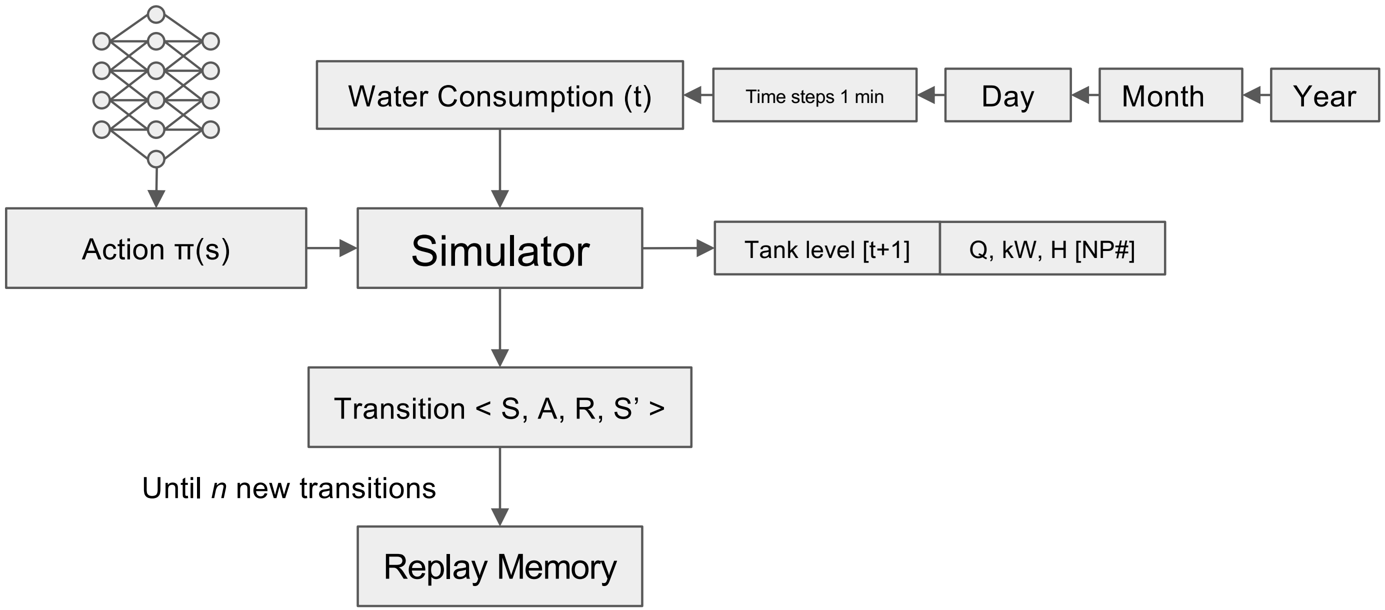
Electricity

Water Distribution Simulator

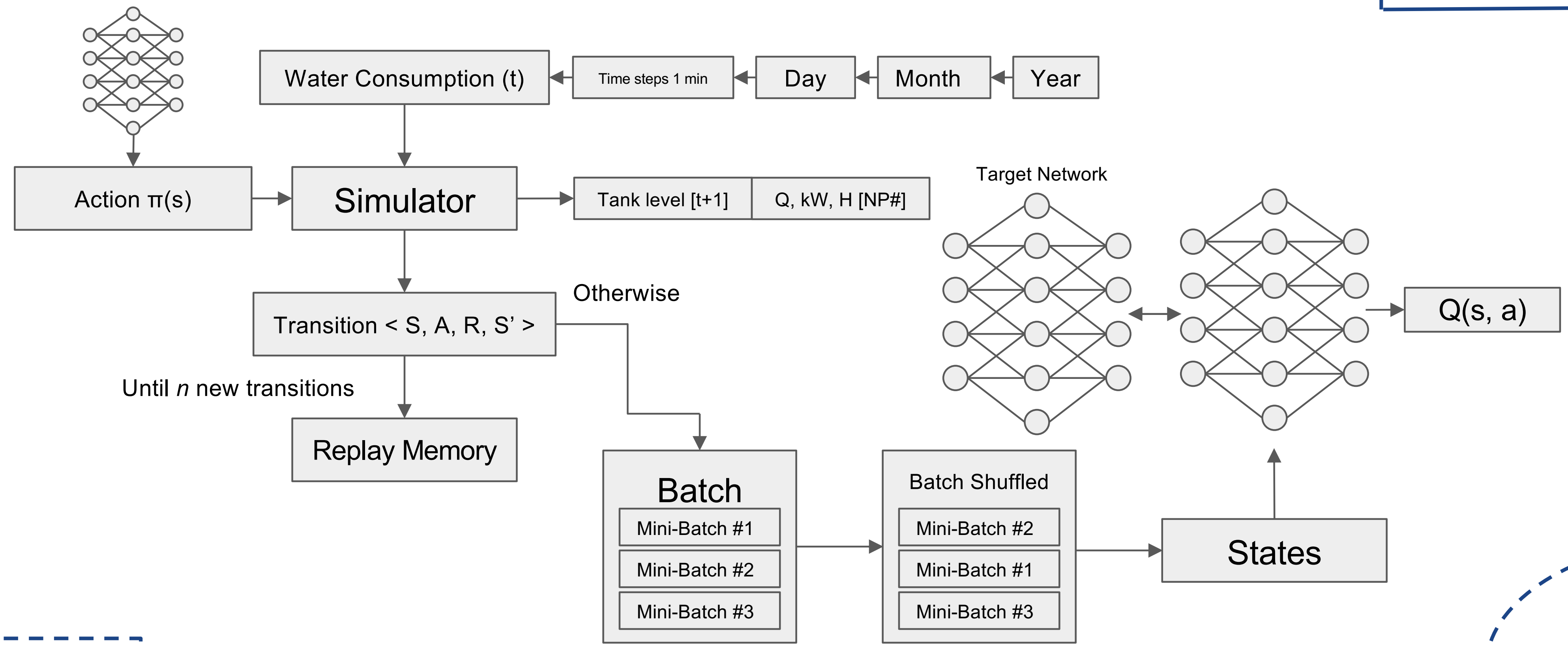




Water Distribution Simulator



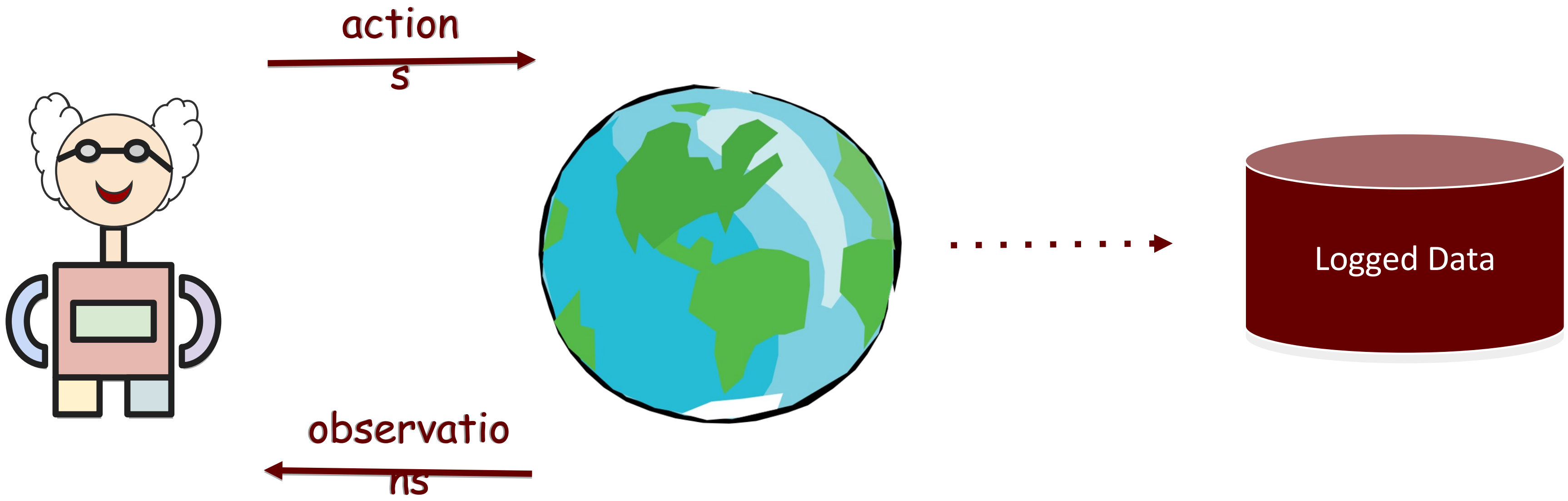
Water Distribution Simulator



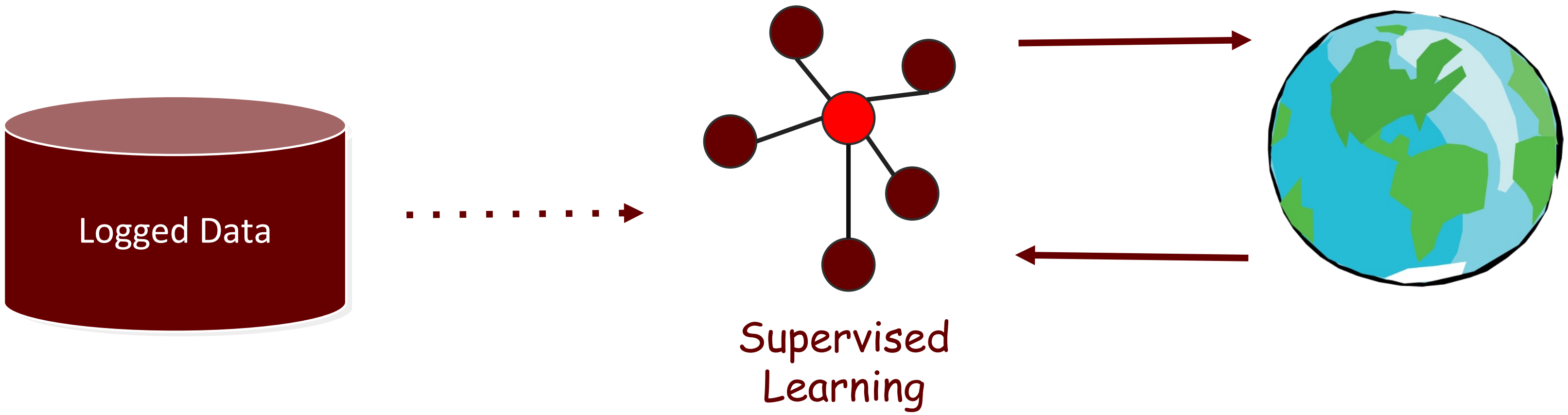
- **[Mnih.2015]** Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness,Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, GeorgOstrovski, et al.2015. Human-level control through deep reinforcement learning. nature 518, 7540 (2015), 529–533.

Safety through Intrinsically Motivated Imitation Learning

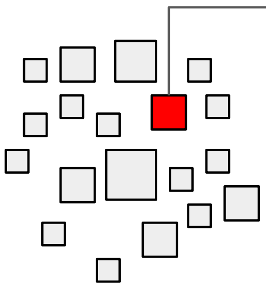
Imitation Learning



Imitation Learning



Safety through Intrinsically Motivated Imitation Learning (SIMIL)



state K-Nearest Neighbors

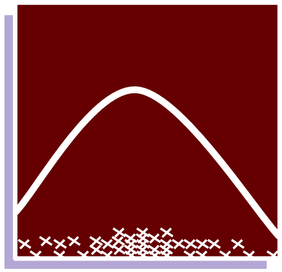
↓

choose the action most often

↓

apply this action

augment the reward according to the state likelihood under the demonstration distribution

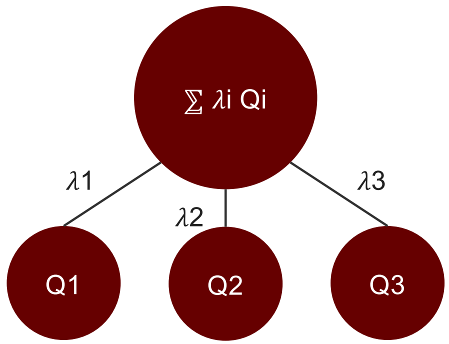


$$r'_t = r_t + \rho\eta(s_t)$$
$$\delta_t = r'_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$$

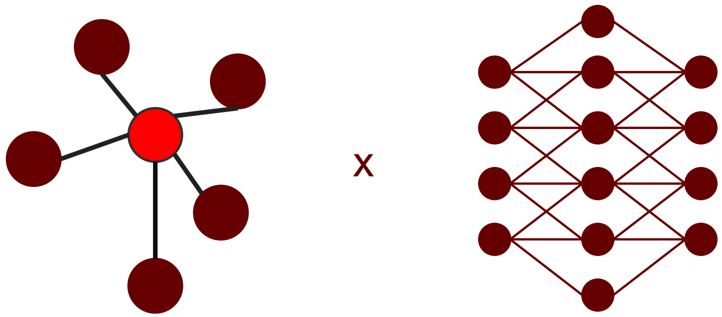
Offline RL Algorithms

- Compare performance with Offline RL Algorithms

    - ❑ Random Ensemble Mixture (REM)



    - ❑ Batch Constrained deep Q-Learning (BCQ)

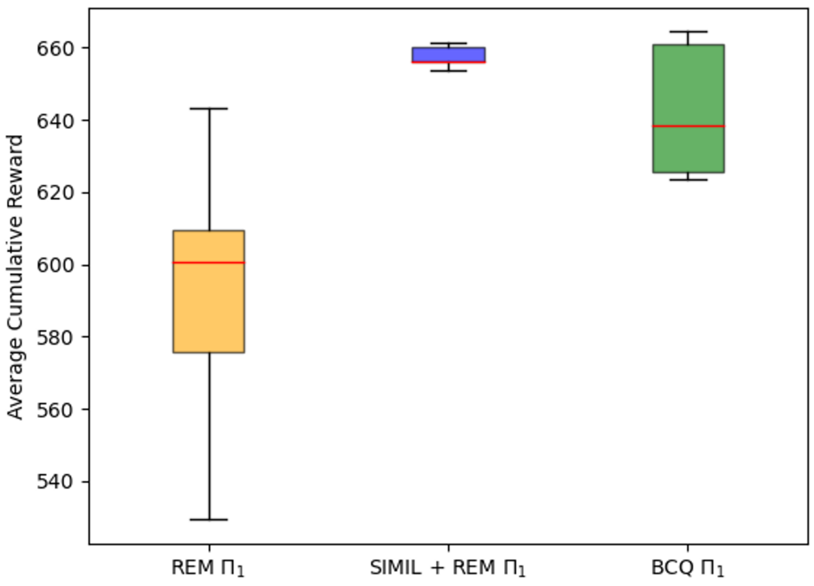


Offline RL Algorithms

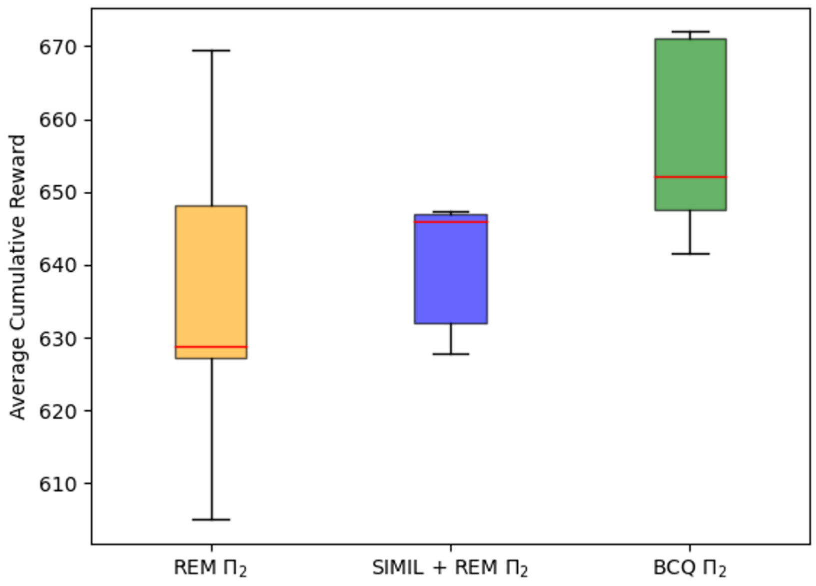
- Compare performance with Offline RL Algorithms
- Generate the same amount of data
- SIMIL + REM
- Evaluate the policies using the water distribution simulator
  - 1 year for learning, 1 year for evaluation
- We average the mean cumulative return of 5 policies

Results

$$r_t = e^{1/(-Q_t/kW_t)} - B * \psi + \log(1/(P+\omega))$$



$$r_t = -e^{(-1/kW_t)} - B * \psi + \log(1/(P+\omega))$$



Comparison Real-World

Policy	Electricity Consumption (kW) (%)
REM Π_1	-1.11 ± 9.78
SIMIL + REM Π_1	**-4.05 ± 1.97**
BCQ Π_1	-3.54 ± 2.71
REM Π_2	4.08 ± 7.93
SIMIL + REM Π_2	**-3.33 ± 5.77**
BCQ Π_2	-1.40 ± 3.33

Knowledge Transfer for Compositional Representations through Curriculum Learning

Curriculum Learning [Bengio et al. 2009]

$1 + 1 + 1 = 3$	$3 \times 1 = 3$	$3 \times (1 + 3) = 12$
$5 - 1 - 2 = 2$	$5 \times 1 - 3 = 2$	$7 \div 2 = 3.5$
$7 - 3 + 4 = 8$	$8 \div 2 \times 2 = 8$	$x \div 2 = 8 + 4$

Task Complexity

Knowledge Transfer for Compositional Representations

State Space S → State Space S'

Action Space A → Action Space A'

Reward Function R → Reward Function R'

Degrees of Freedom!

Knowledge Transfer for Compositional Representations

State Space S → State Space S'

Action Space A → Action Space A'

Reward Function R → Reward Function R'

Policy π_s ← Source Task

Exploration!

Knowledge Transfer for Compositional Representations

State Space S | Action Space A | Reward Function R

State Space S' | Action Space A' | Reward Function R'

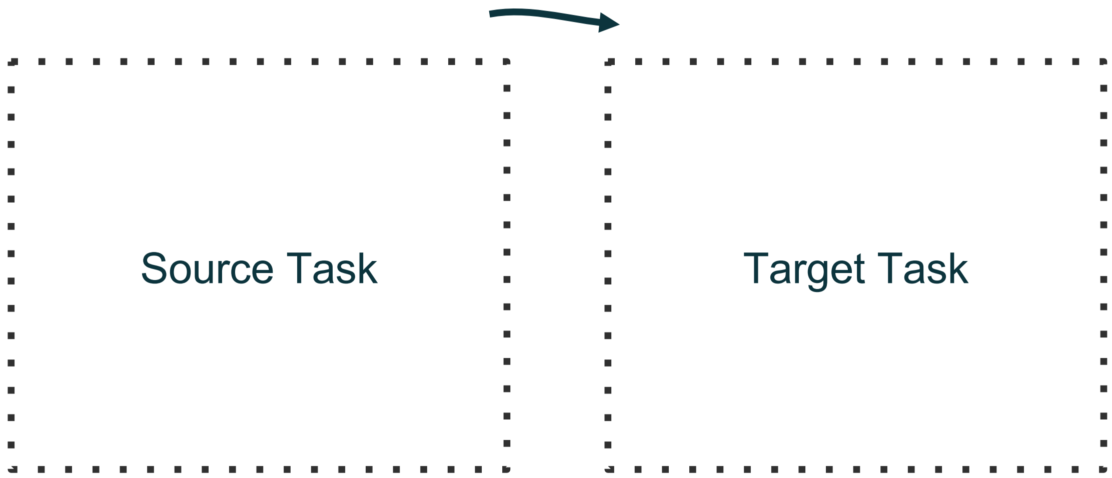
Transfer Learning!

Policy π_s ← Source Task

softmax(max(Q_s), max(Q_t)) → Target Task → Policy π_t

Knowledge Transfer for Compositional Representations

softmax(max(Q_s), max(Q_t))



Knowledge Transfer for Compositional Representations

$$\mathbf{softmax}(\max(Q_s), \max(Q_t))$$

Task

Target Task

↓

Source Task

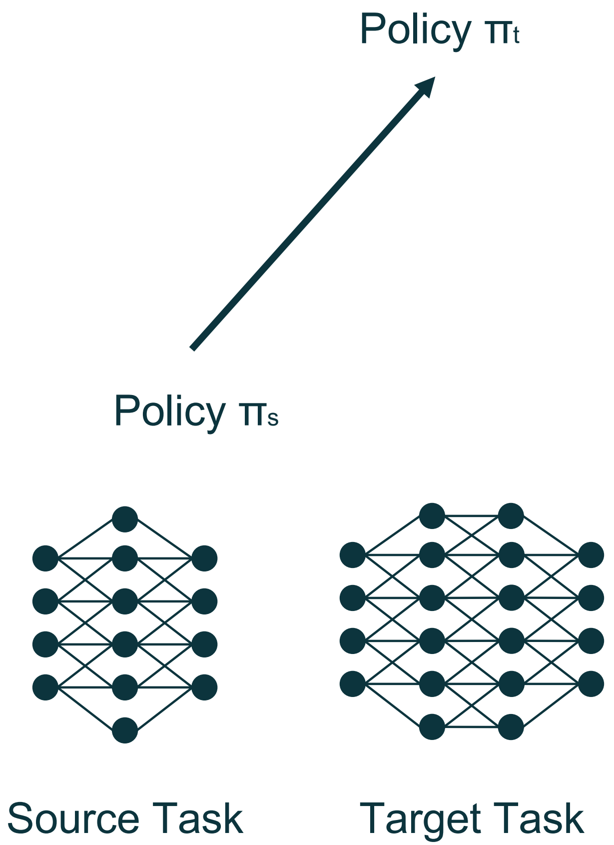
Target Task

…

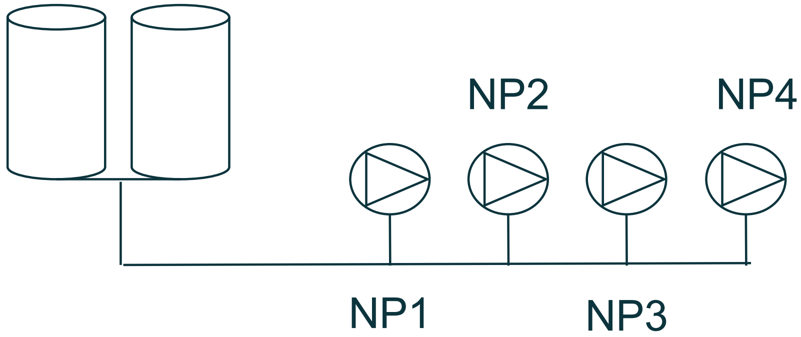
Knowledge Transfer for Compositional Representations

$$softmax(\{max(Q_s), max(Q_t)\})$$

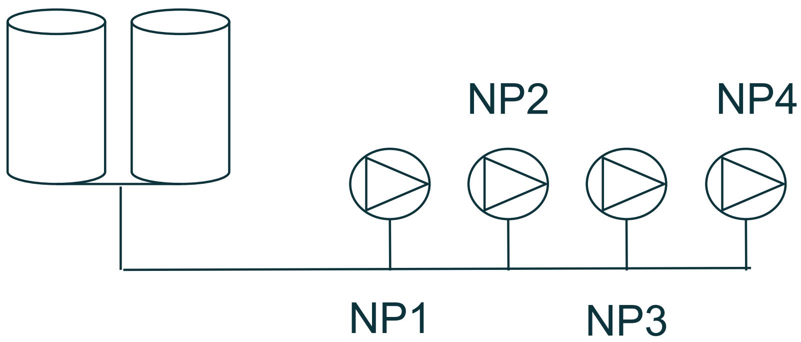
$$a = \begin{cases} argmax(Q_s), softmax_{Q_s} \\ argmax(Q_t), softmax_{Q_t} \end{cases}$$

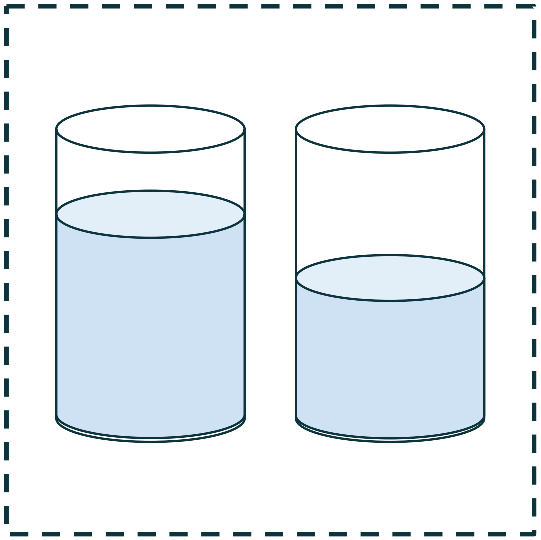


Pump Scheduling: State S and Action A



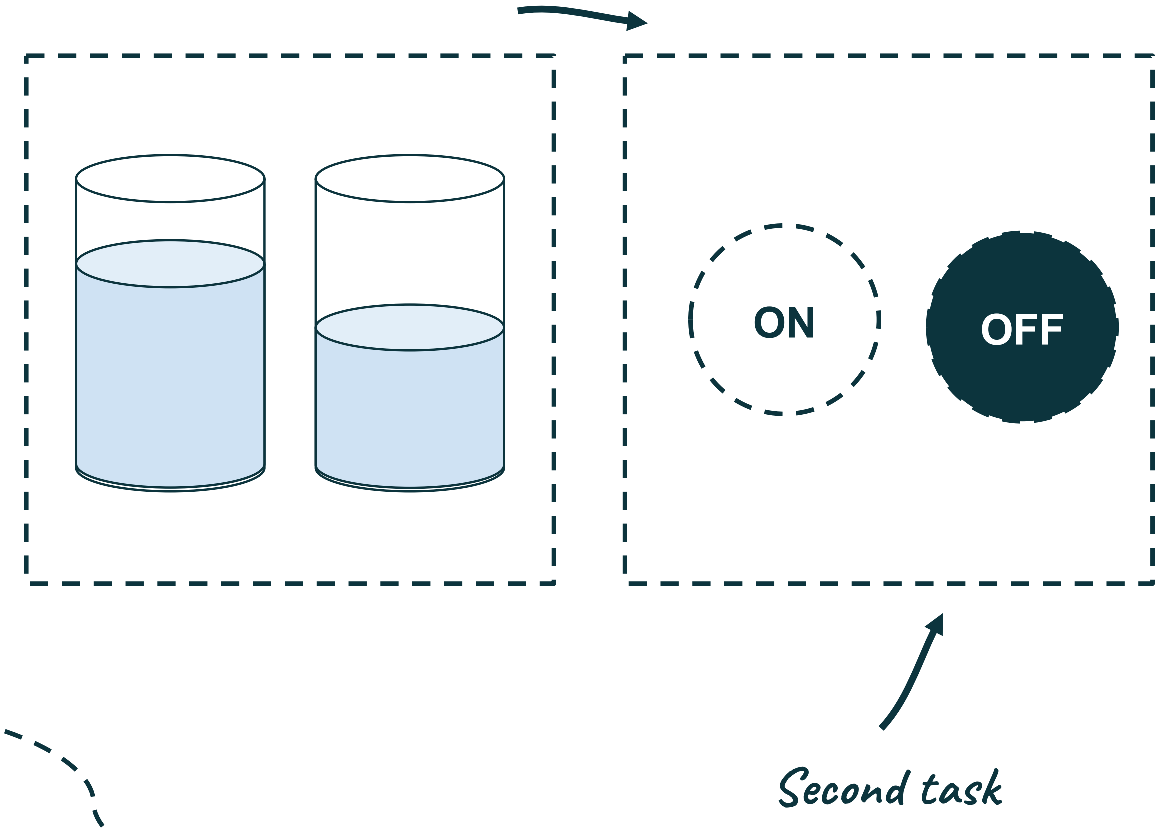
Pump Scheduling: 3 steps curriculum



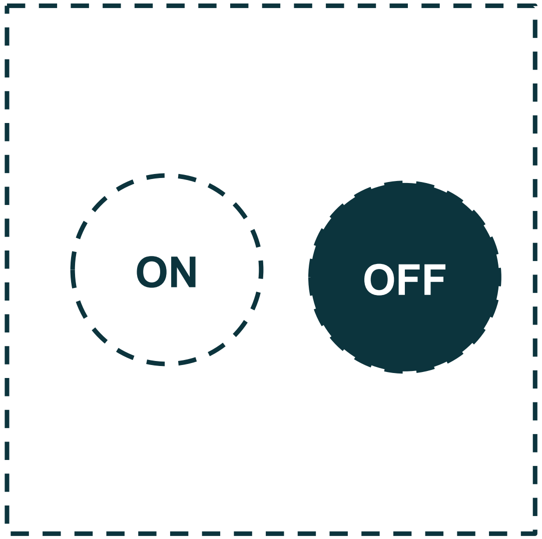


First task

Pump Scheduling: 3 steps curriculum



Pump Scheduling: 3 steps curriculum

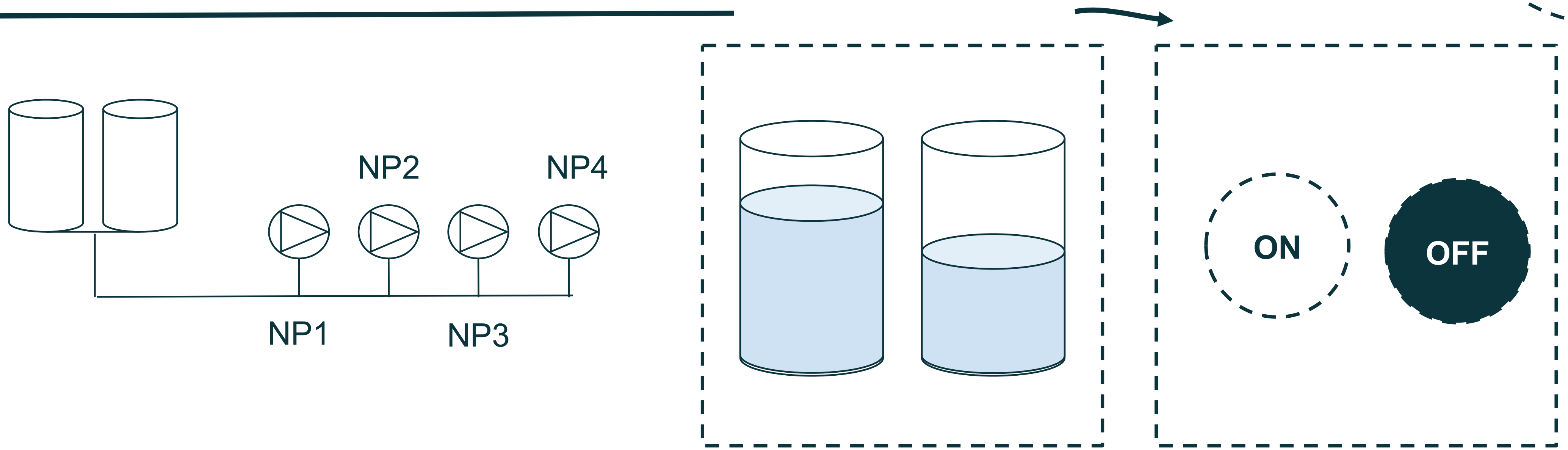


$$Q_s = \{avg(NOP), NP1, NP2, NP3, NP4\}$$

$$softmax(max(Q_s), max(Q_t))$$

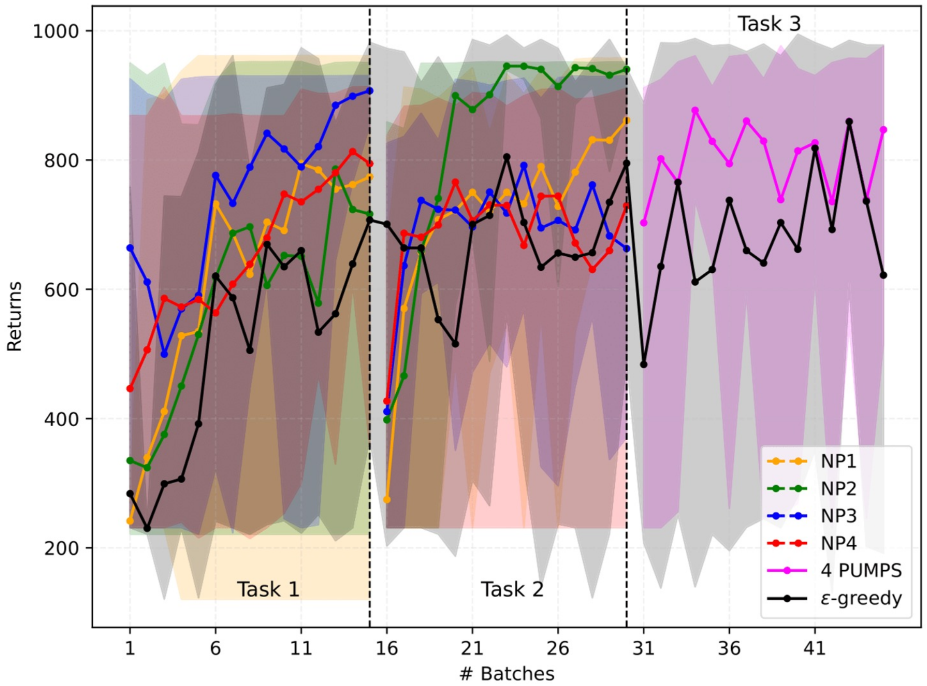
Final task

Pump Scheduling



$$Q_s = \{avg(NOP), NP1, NP2, NP3, NP4\}$$
$$softmax(max(Q_s), max(Q_t))$$

Results: Pump Scheduling



Summary

- POMDP for Pump Scheduling
  - Lead to electricity savings while meeting constraints
- SIMIL
  - Improve policy’s performance over baseline learning algorithm
- Curriculum Learning
  - Can lead to better asymptotic performance compared to standard exploration

Acknowledgments

Harald Roclawski (TU Kaiserslautern) ✧ Aloysio P. M. Saliba (UFMG) ✧ Benjamin Dewals (ULiège) ✧ Anika Theis (TU Kaiserslautern) ✧ Thomas Pirard (ULiège) ✧ Laura Sterle (TU Kaiserslautern) ✧ Thomas Krätzig (Dr. Kraetzig)

AGENCE NATIONALE DE LA RECHERCHE ANR · INSA ROUEN NORMANDIE · Water JPI · 2017 JOINT CALL

The authors acknowledge FAPEMIG, Federal Ministry of Education and Research of Germany, Agence Nationale de la Recherce de France and Fonds de la Recherce Scientifique Belge, for funding this research by the project IoT.H2O (ANR-18-IC4W-0003) on the IC4Water JPI call.

References

Jaques, Natasha, et al. "Social influence as intrinsic motivation for multi-agent deep reinforcement learning." International conference on machine learning. PMLR, 2019.

Lair, Nicolas, et al. "Language grounding through social interactions and curiosity-driven multi-goal learning." arXiv preprint arXiv:1911.03219 (2019).

Bellemare, Marc, et al. "Unifying count-based exploration and intrinsic motivation." Advances in neural information processing systems 29 (2016).

Tang, Haoran, et al. "#Exploration: A study of count-based exploration for deep reinforcement learning." Advances in neural information processing systems 30 (2017).

Gregor, Karol, Danilo Jimenez Rezende, and Daan Wierstra. "Variational intrinsic control." arXiv preprint arXiv:1611.07507 (2016).

Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. 2009.

Nakamoto, Mitsuhiko, et al. "Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning." arXiv preprint arXiv:2303.05479 (2023).

Andrychowicz, Marcin, et al. "Hindsight experience replay." Advances in neural information processing systems 30 (2017).

Schaul, Tom, et al. "Prioritized experience replay." arXiv preprint arXiv:1511.05952 (2015).

Dai, Siyu, Andreas Hofmann, and Brian Williams. "Automatic curricula via expert demonstrations." arXiv preprint arXiv:2106.09159 (2021).

Campero, Andres, et al. "Learning with amigo: Adversarially motivated intrinsic goals." arXiv preprint arXiv:2006.12122 (2020).

Costa, L. H. M., H. M. Ramos, and M. A. H. De Castro. "Hybrid genetic algorithm in the optimization of energy costs in water supply networks." Water Science and Technology: Water Supply 10.3 (2010): 315-326.

Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8 (1992): 279-292.