From Predict to Control
From RL to Offline RL

Zhi-Hong Deng

邓智鸿

2020.11.3

Outline

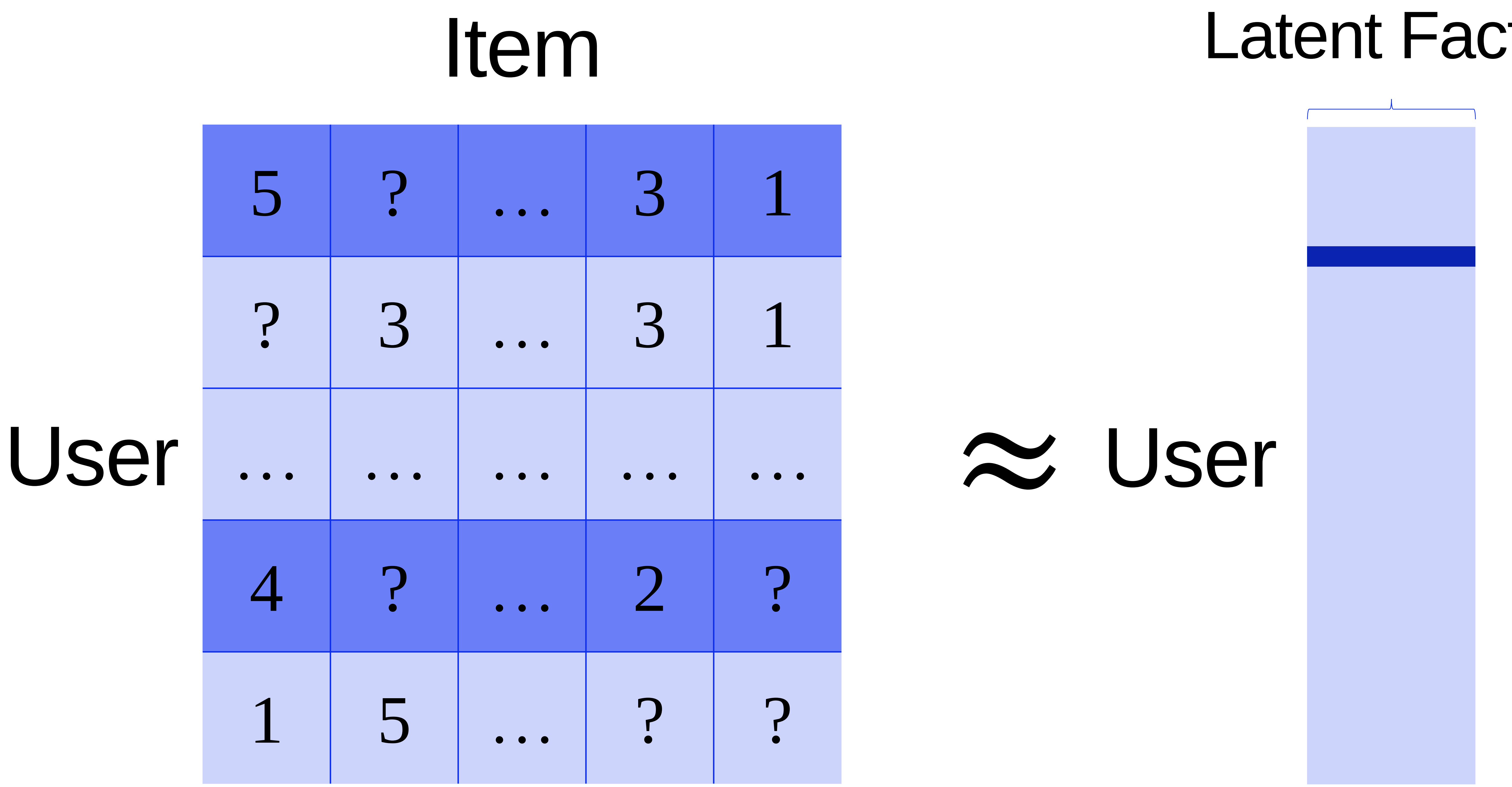
1. Recommender Systems

2. Reinforcement Learning

3. Offline Reinforcement Learning

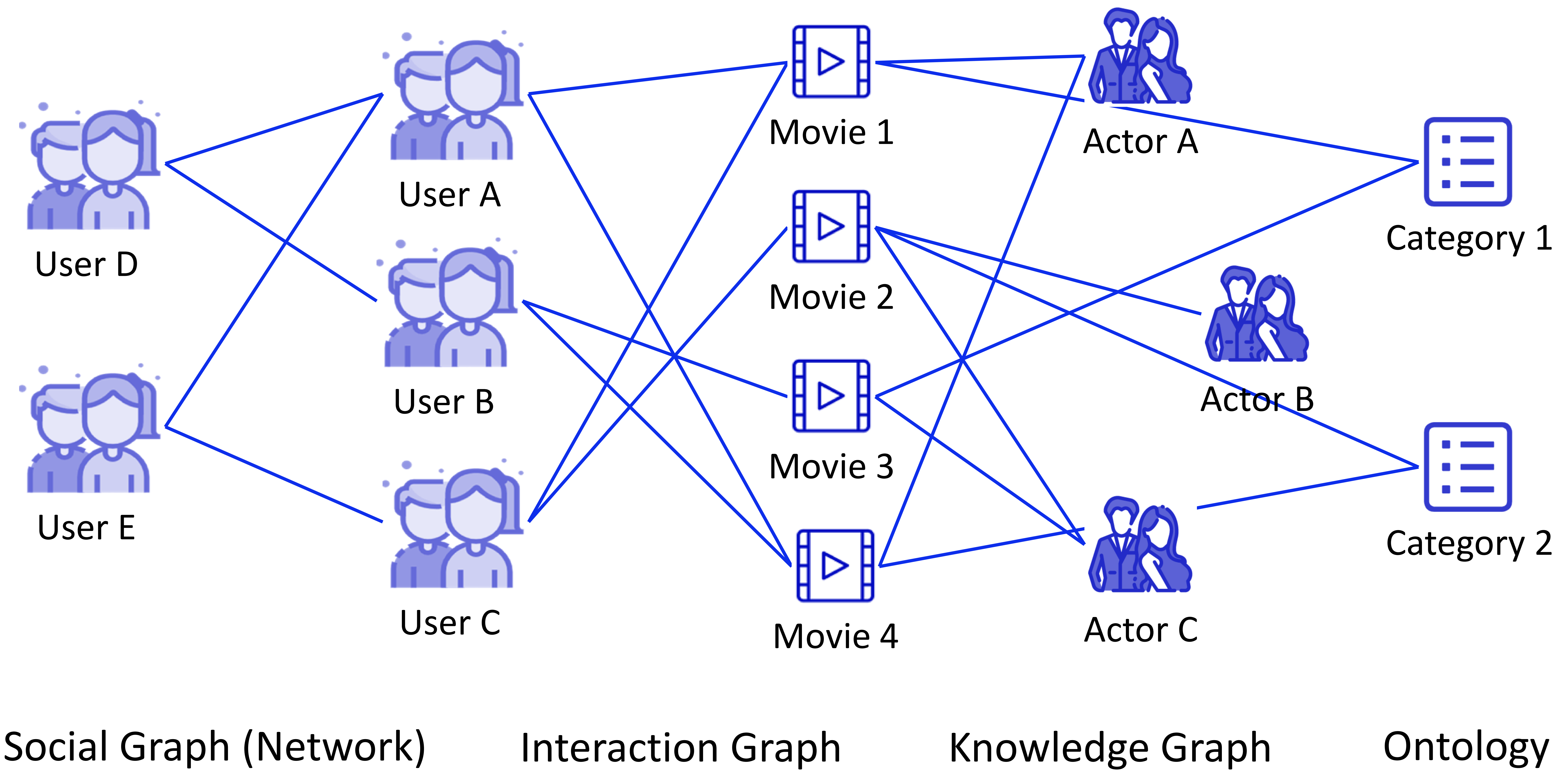
Recommender Systems

The Three Viewpoints of the Recommendation problem —— Matrix

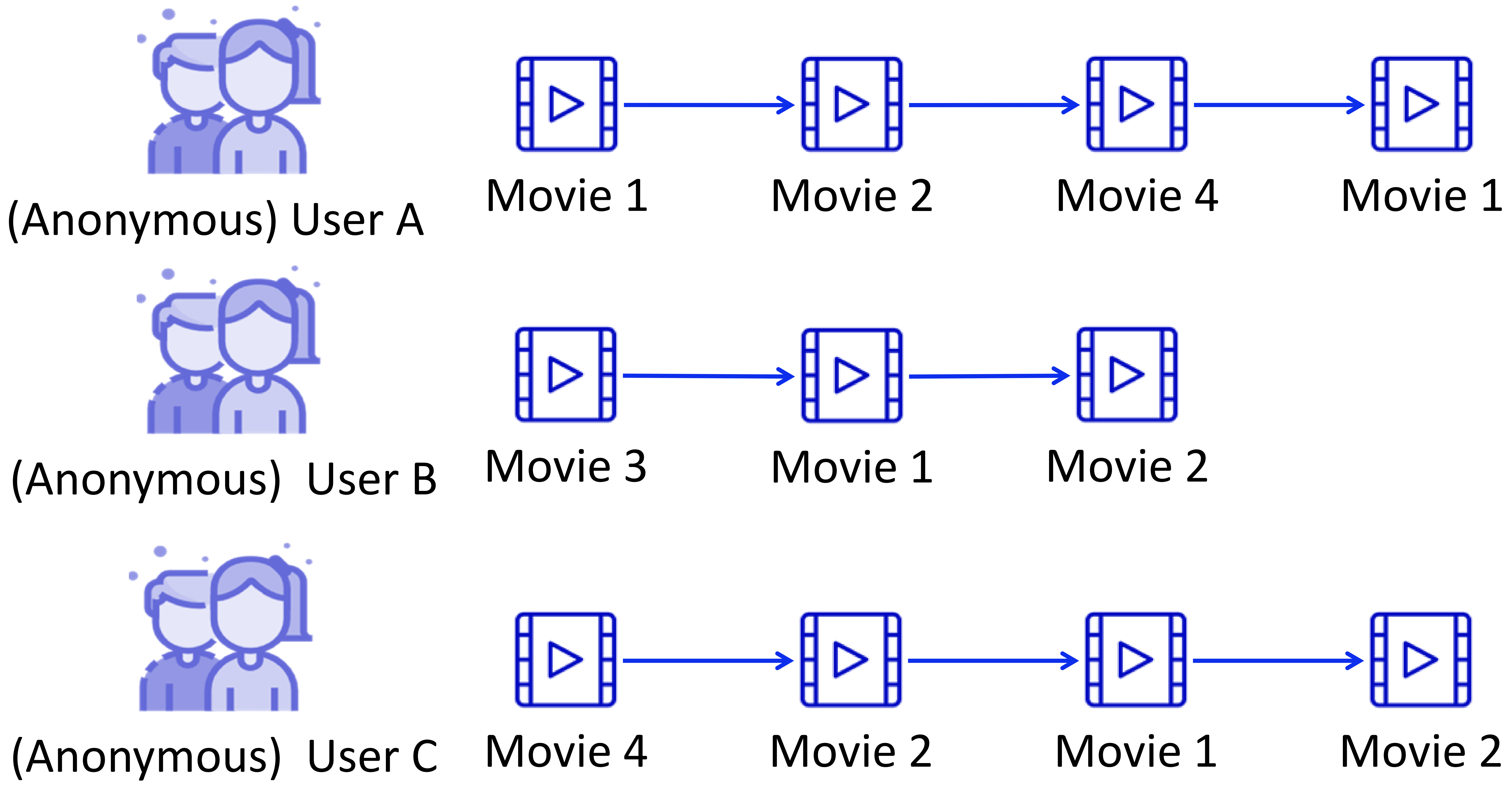


Project users and items into a common latent space, and then use the dot product of latent factors to estimate the rating.

The Three Viewpoints of the Recommendation problem —— Graph



The Three Viewpoints of the Recommendation problem —— Sequence

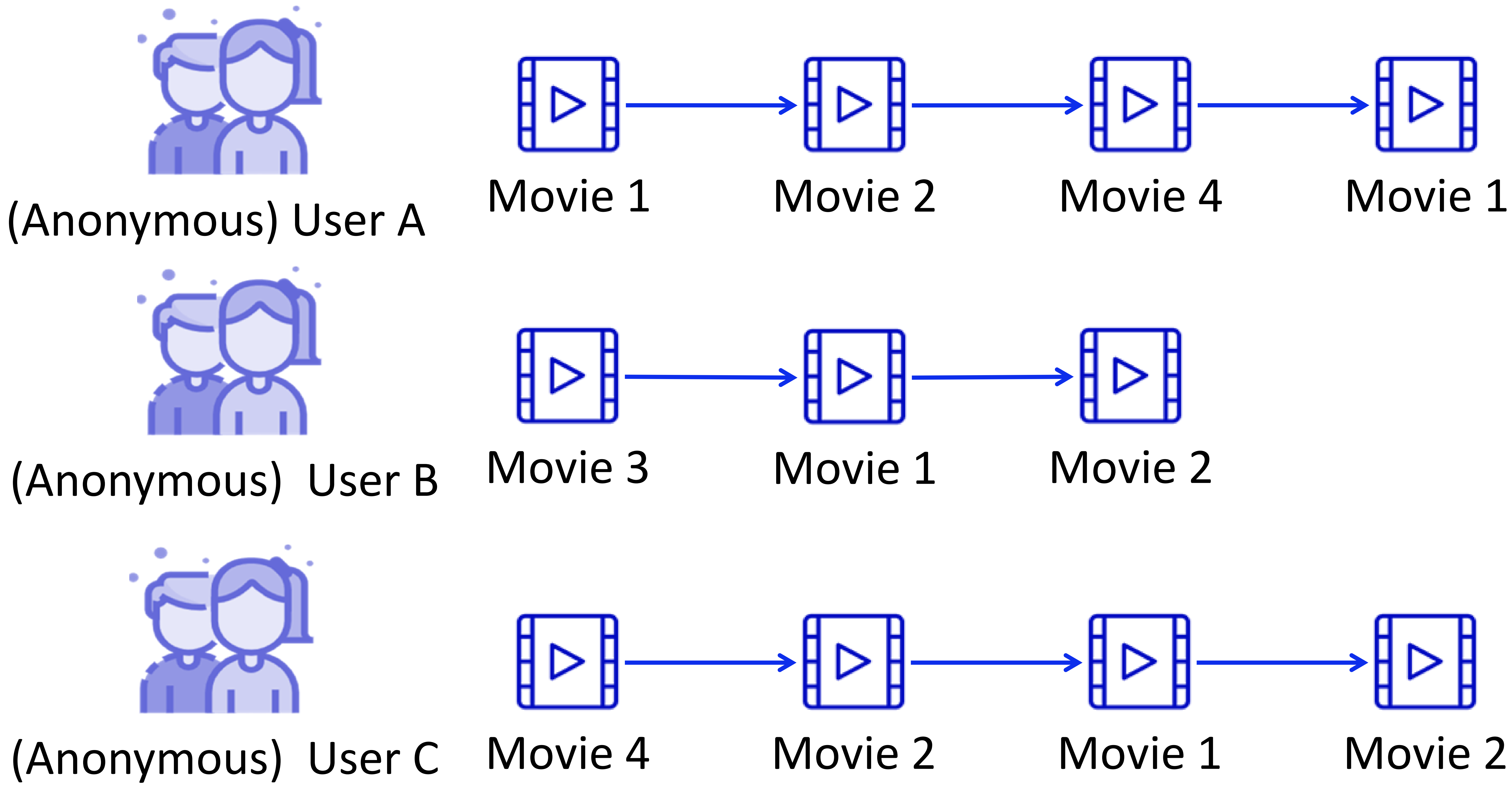


Sequential Recommendation

Session-based Recommendation

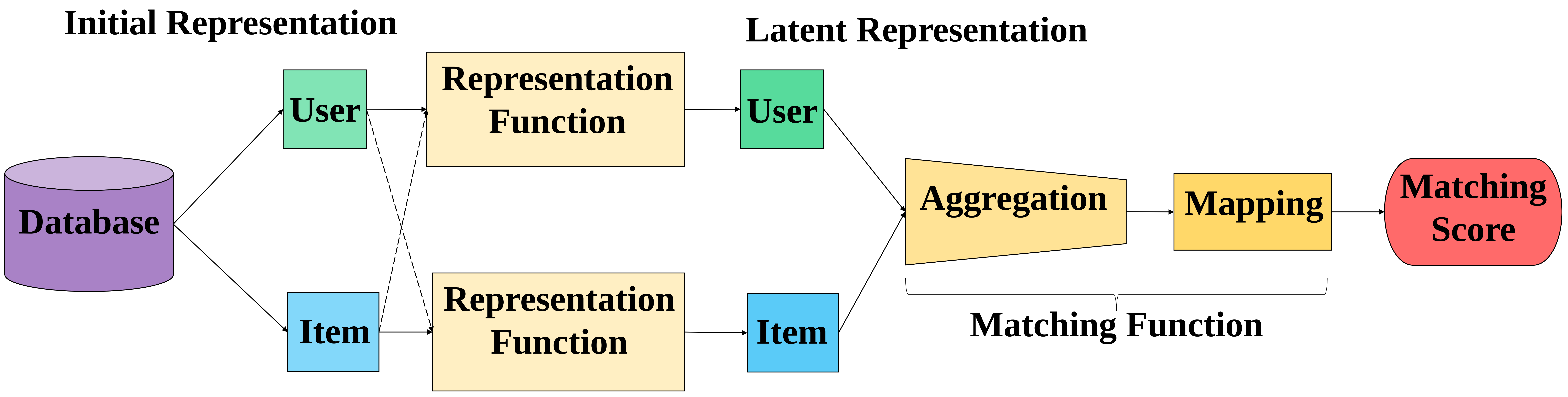
Next-basket Recommendation

The Three Viewpoints of the Recommendation problem —— Sequence



Sequential Recommendation Session-based Recommendation Next-basket Recommendation

All roads lead to “Matching”

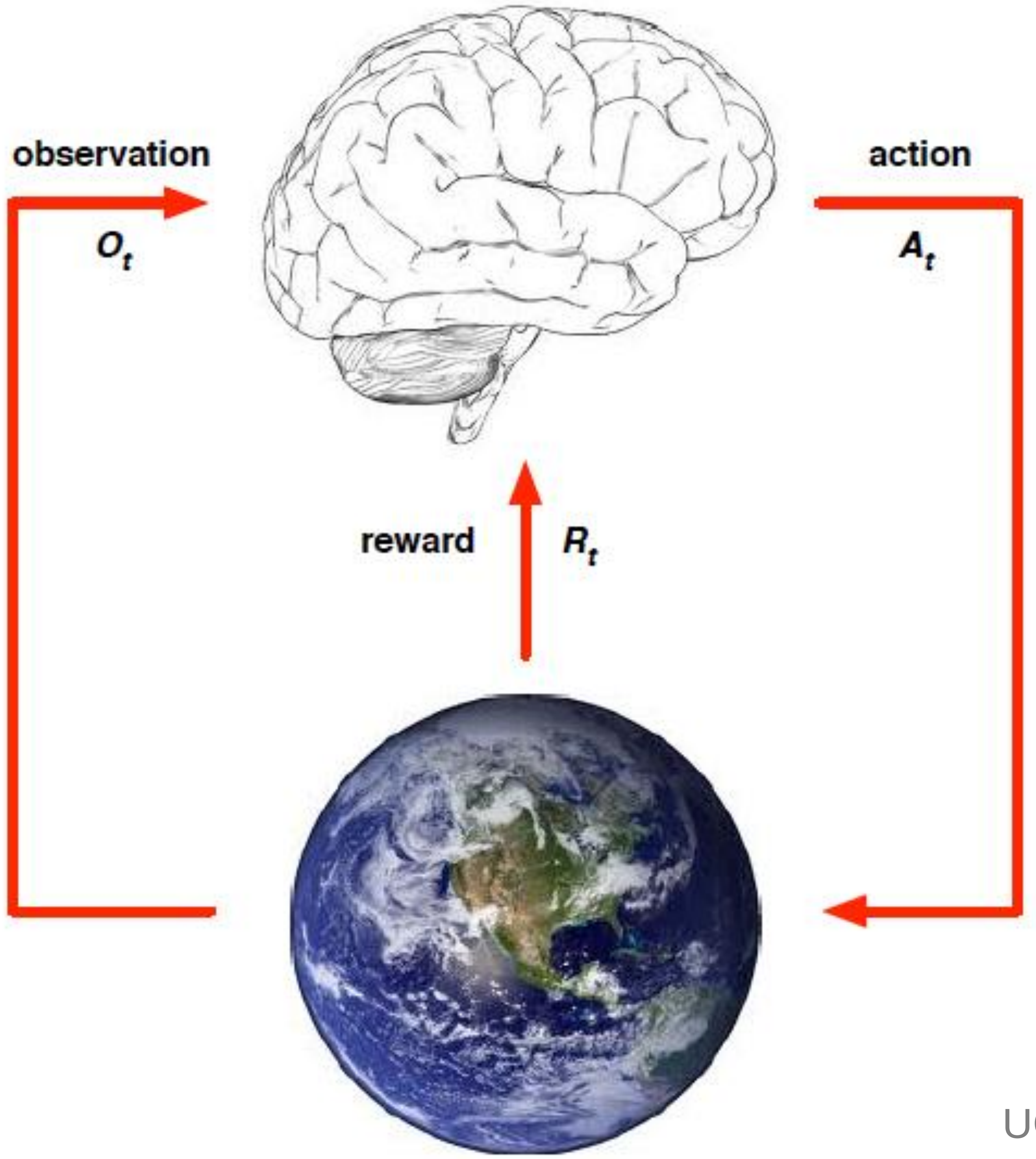


- Matching is a much broader topic in the domain of Information Retrieval.
- Matching can be viewed as a special type of classification problems which aims to predict the most relevant items/documents/answers.

Is real-world recommendation a prediction task?

Reinforcement Learning

Reinforcement Learning



RL is a general-purpose framework for decision-making.

- An agent selects actions
- Its actions influence its future observations
- Success is measured by a scalar reward signal

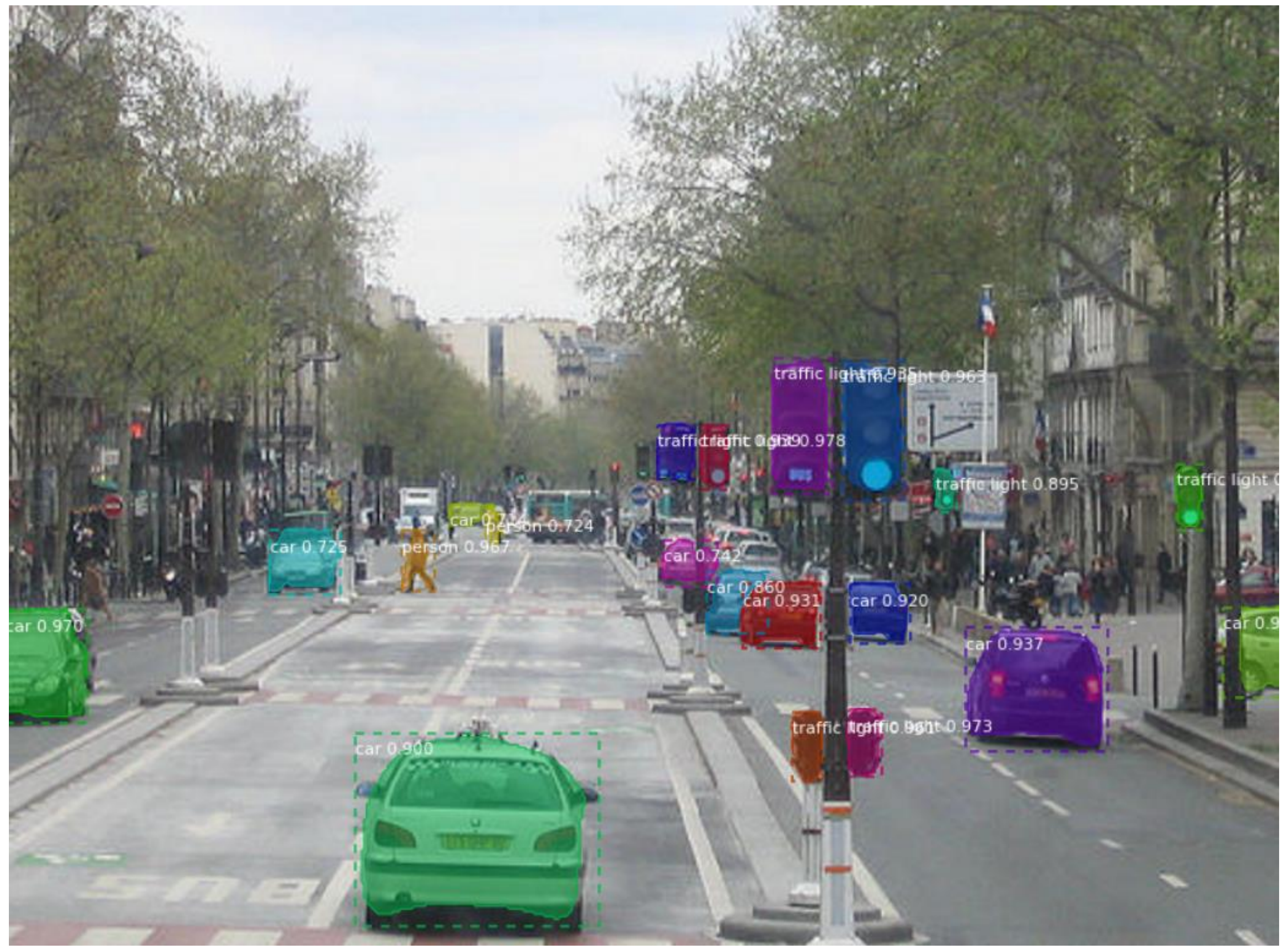
Goal: select actions to maximize future rewards

UCL Course on RL by David Silver

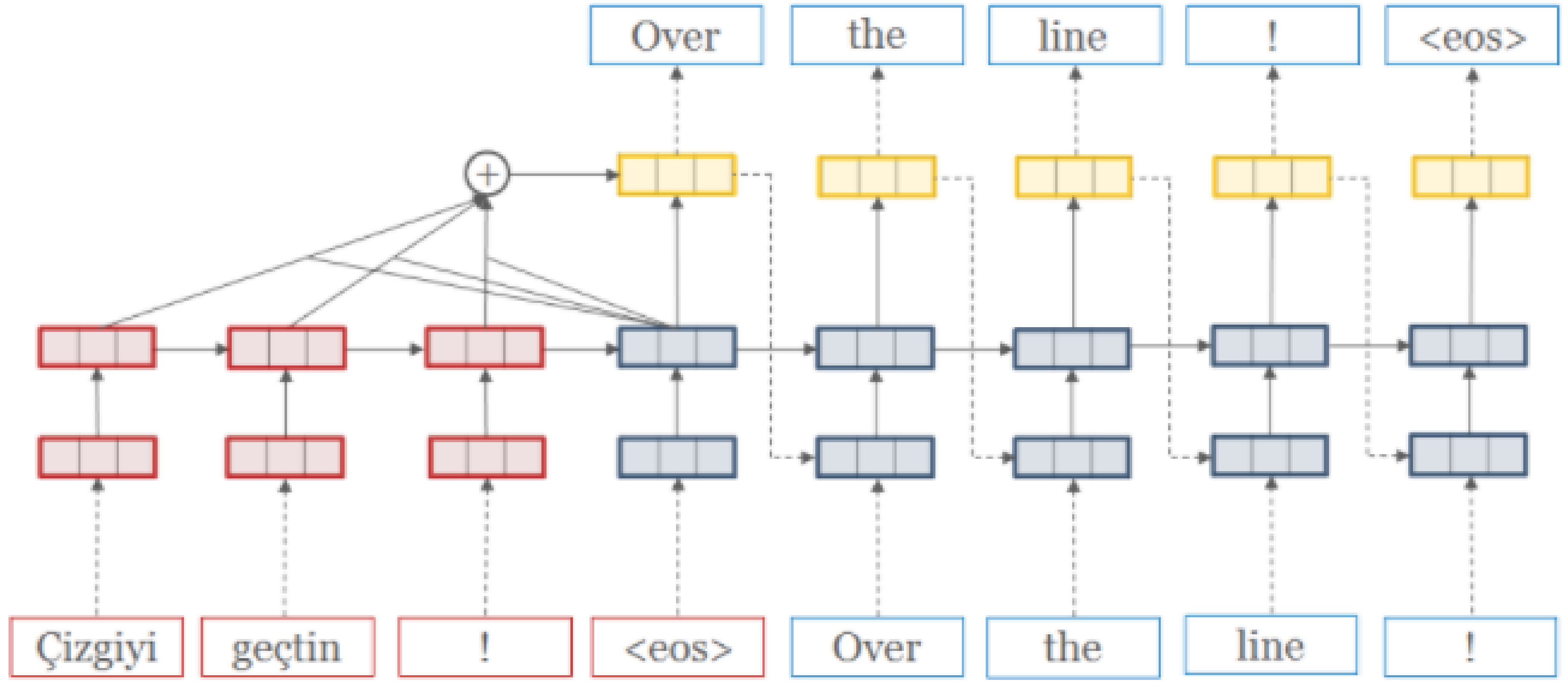
DL + RL = Artificial General Intelligence !

—— David Silver (DeepMind)

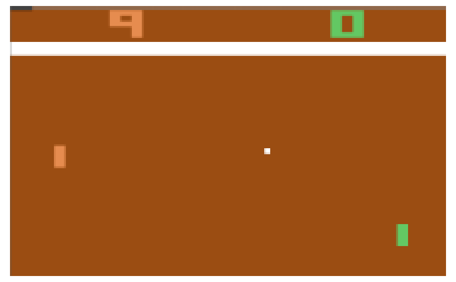
Supervised Learning vs. Reinforcement Learning



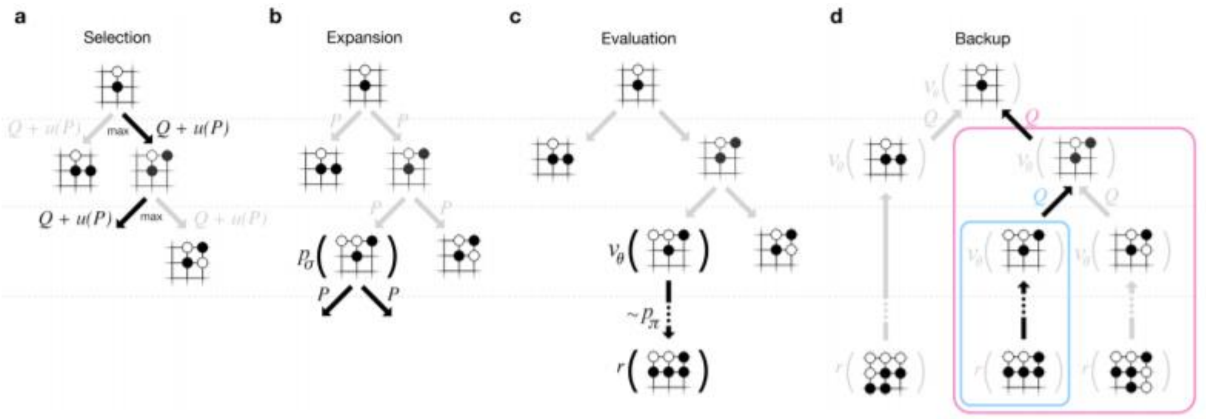
Mask R-CNN



Open NMT

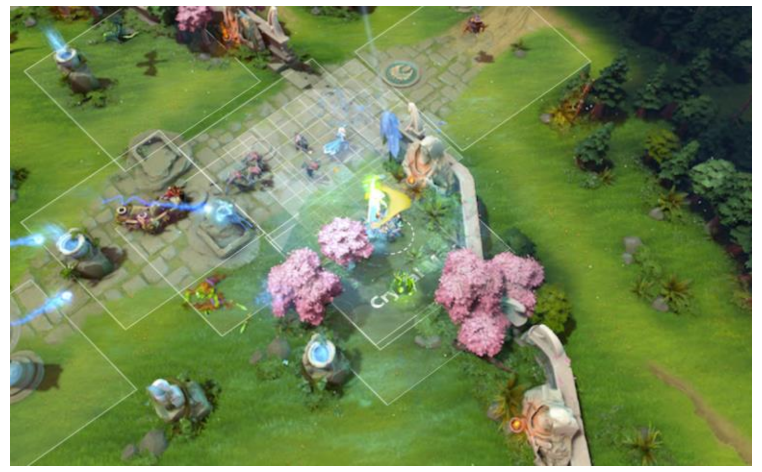


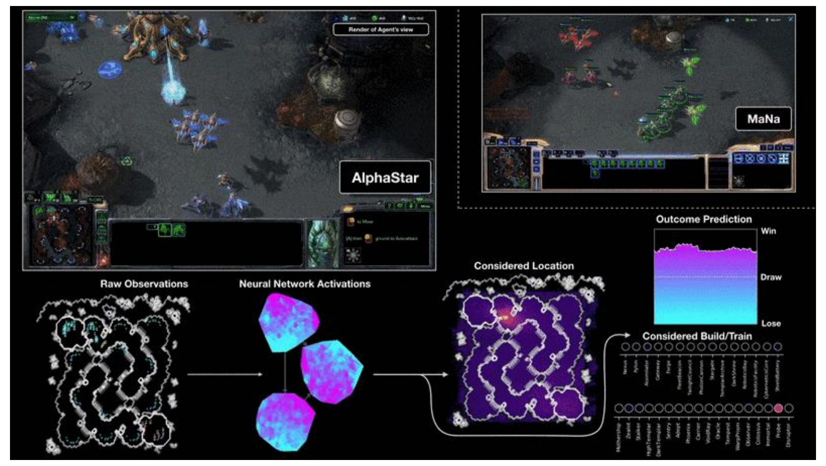
2015 Human-level control through deep reinforcement learning



2016 Mastering the game of Go with deep neural networks and tree search

2017 Mastering the game of Go without human knowledge



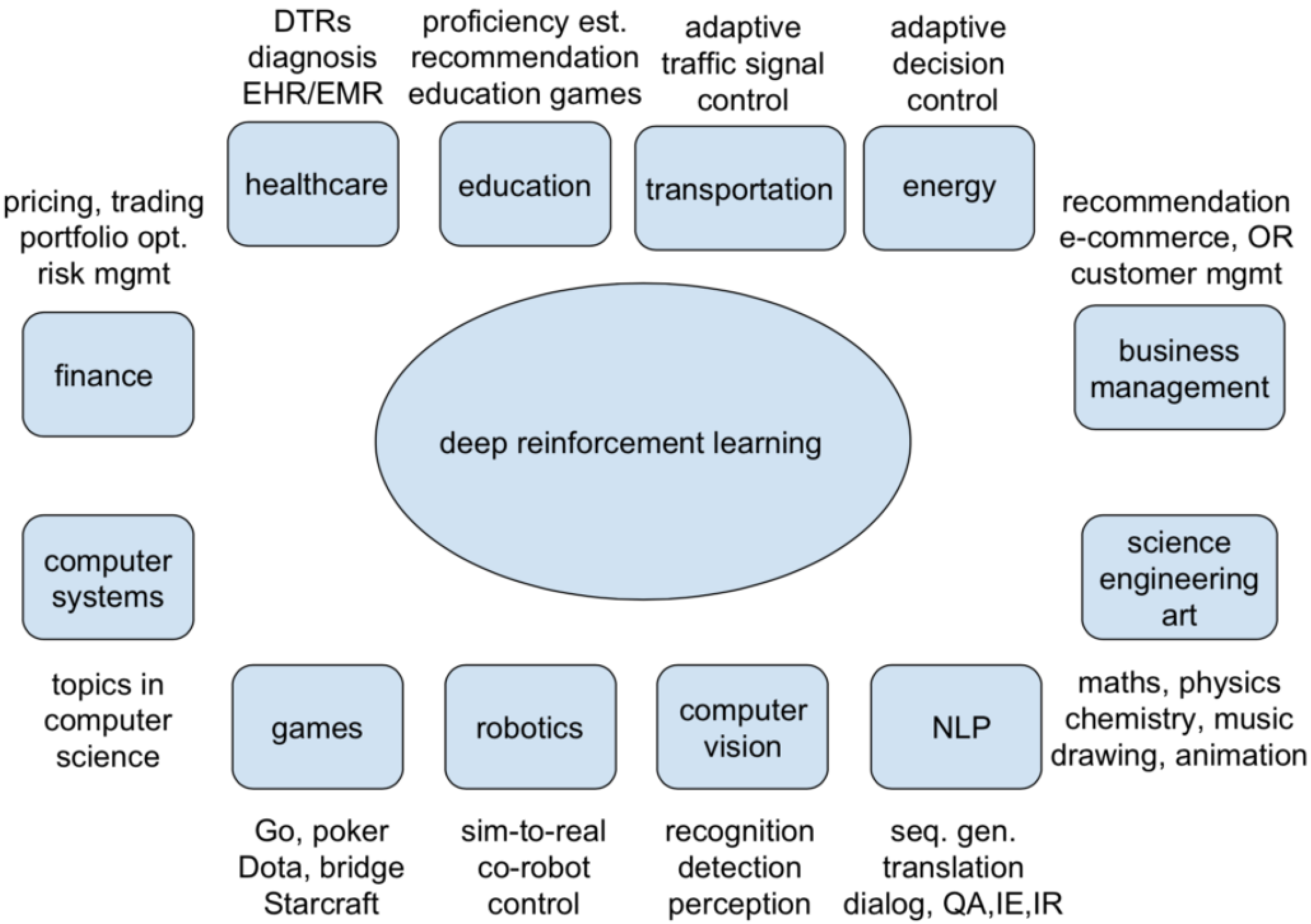


2018 Superhuman AI for heads-up no-limit poker: Libratus beats top professionals

Openai five

2019 Alphastar: Mastering the real-time strategy game starcraft ii

More than Games!



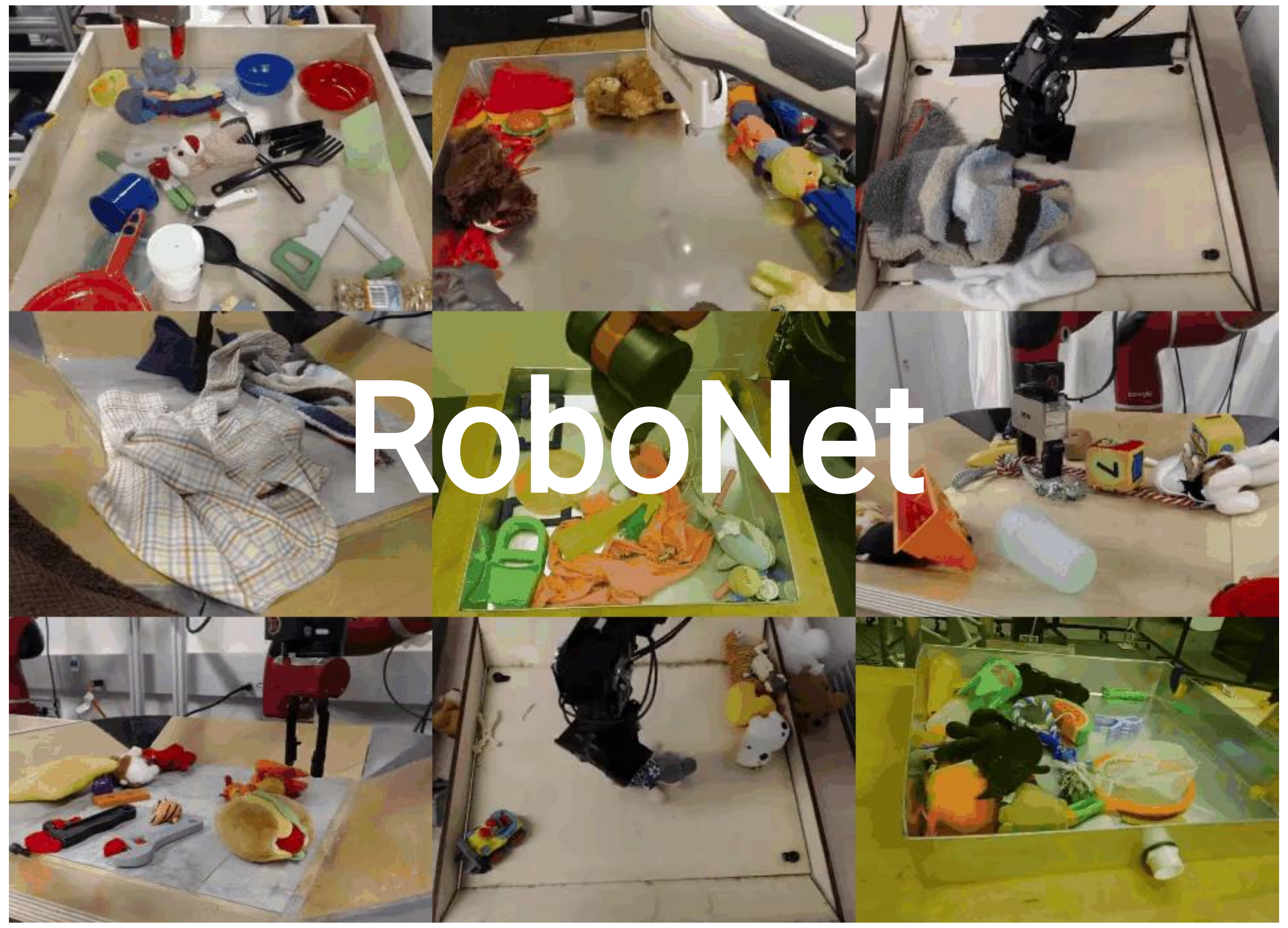
Challenges of Real-World Reinforcement Learning

1. Training off-line from the fixed logs of an external behavior policy.
2. Learning on the real system from limited samples.
3. High-dimensional continuous state and action spaces.
4. Safety constraints that should never or at least rarely be violated.
5. Tasks that may be partially observable, alternatively viewed as non-stationary or stochastic.
6. Reward functions that are unspecified, multi-objective, or risk-sensitive.
7. System operators who desire explainable policies and actions.
8. Inference that must happen in real-time at the control frequency of the system.
9. Large and/or unknown delays in the system actuators, sensors, or rewards.

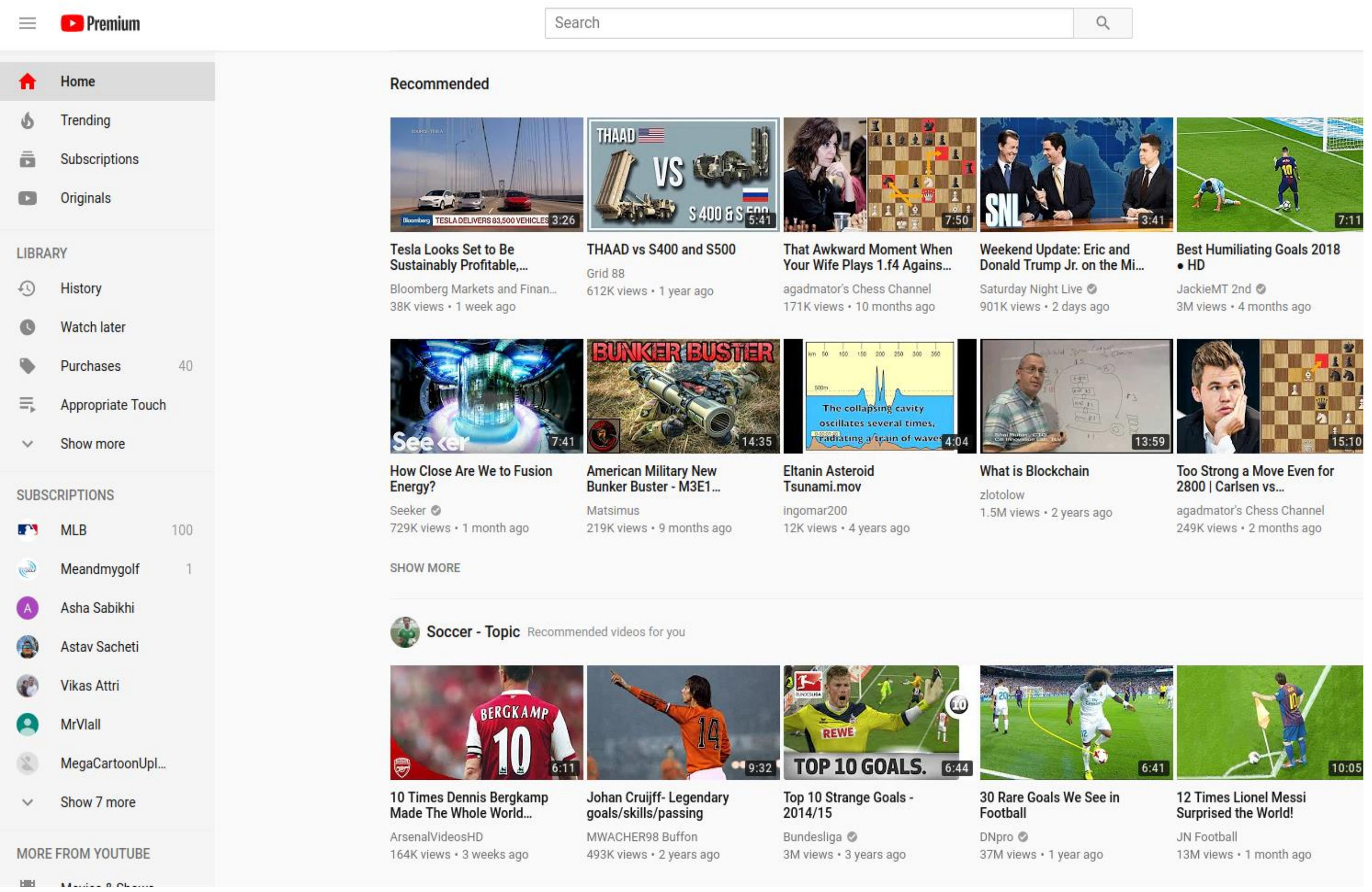
Can We Copy The Success of DL by Offline (Data-driven) RL?

Offline Reinforcement Learning

Reinforcement Learning with Large Real-world Dataset



Robotics

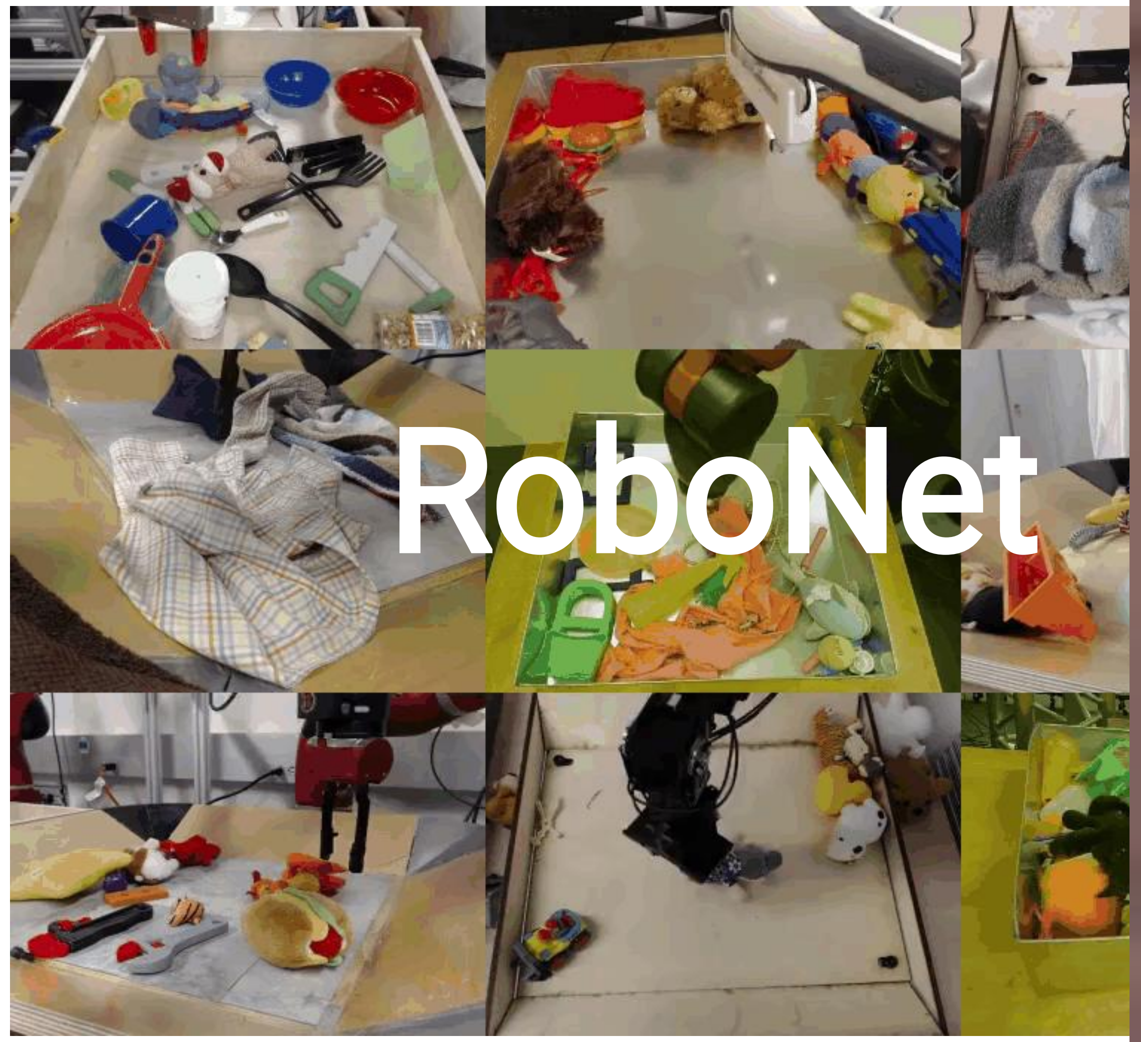


Recommender Systems

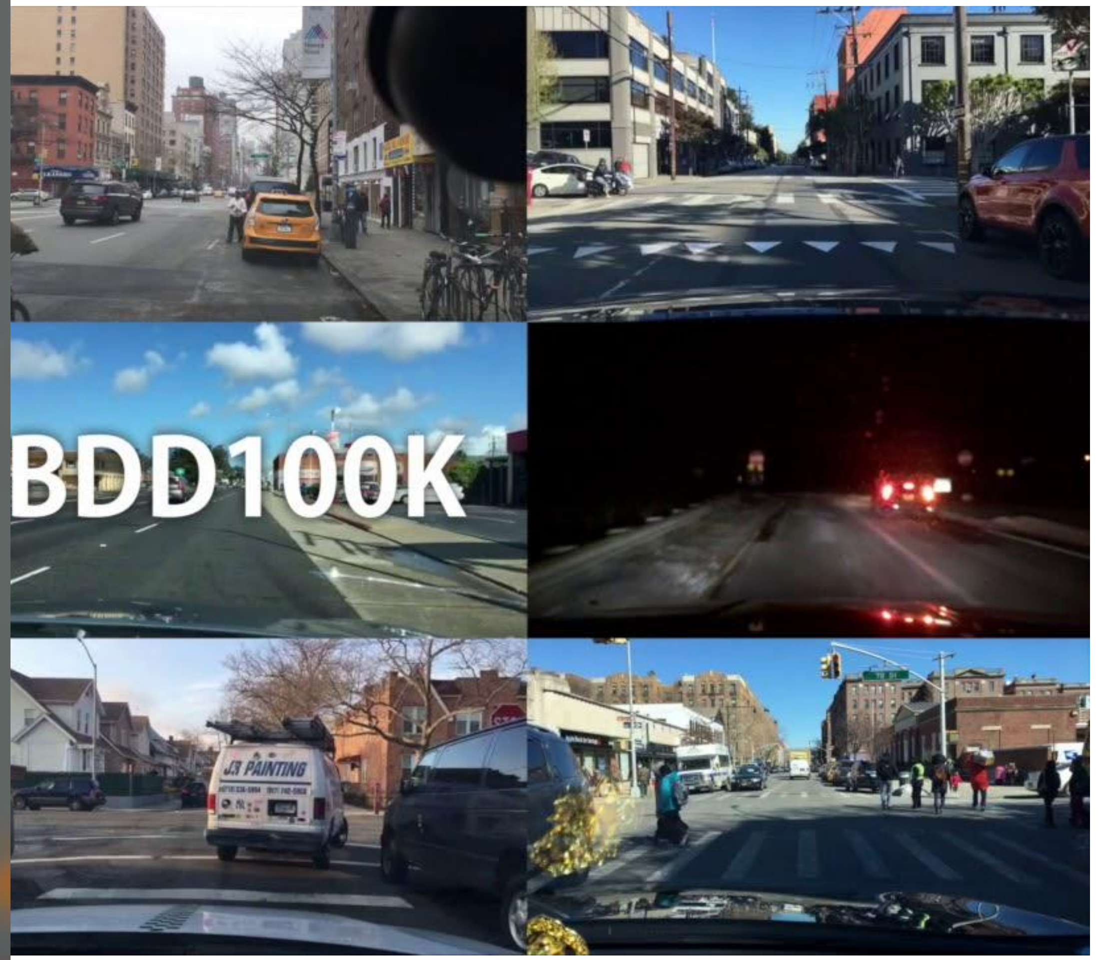
Autonomous Driving

[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.
[2] Yu, Xian, Chen, Liu, Liao, Madhavan, Darrell. BDD100K: A Large-scale Diverse Driving Video Database.

Reinforcement Learning with Large Real-world Dataset



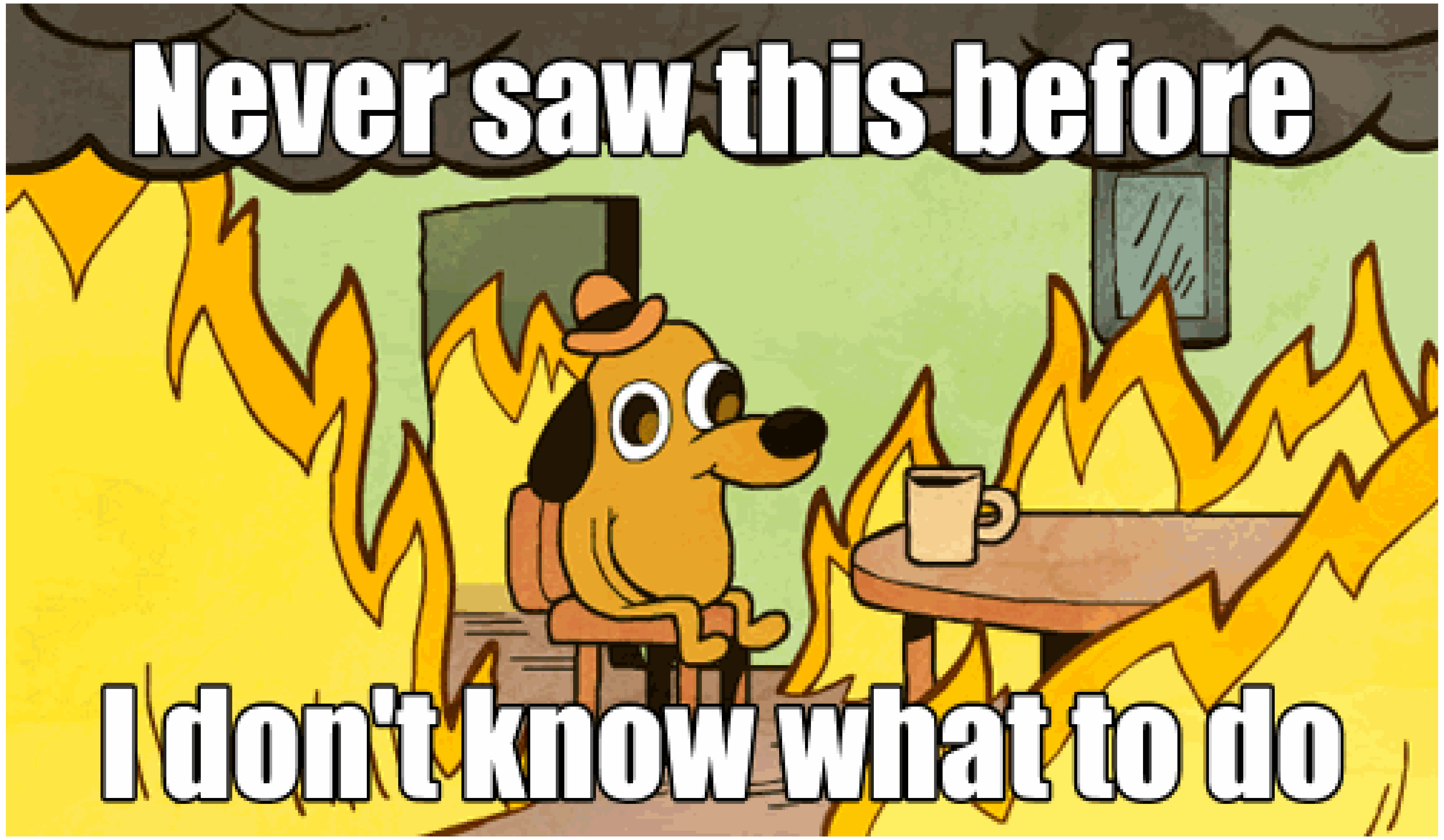
Robotics



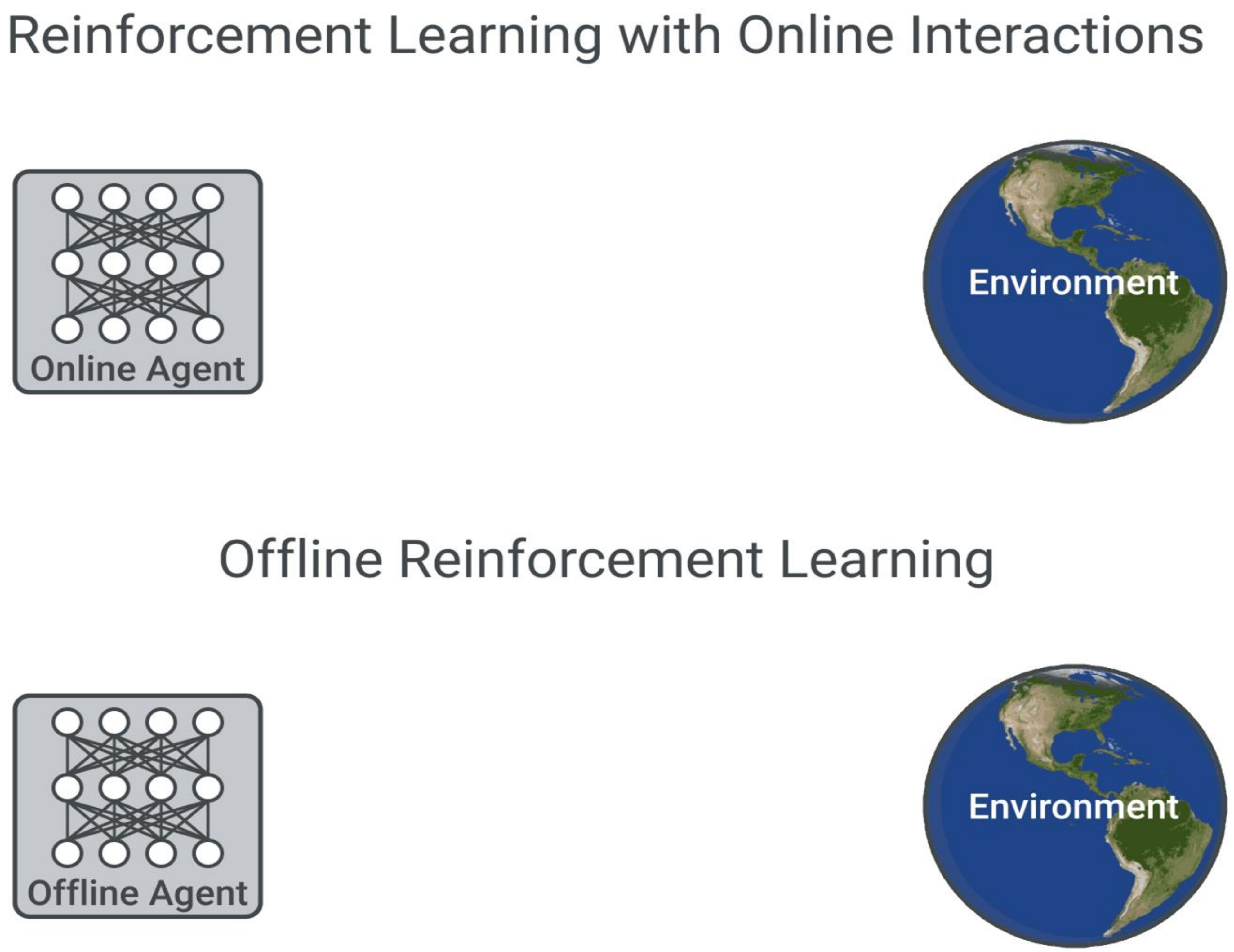
omous Driving

[1] Dasari, Ebert, Tian, Nair, Bucher, Schmeckpeper, .. Finn. RoboNet: Large-Scale Multi-Robot Learning.
[2] Yu, Xian, Chen, Liu, Liao, Madhavan, Darrell. BDD100K: A Large-scale Diverse Driving Video Database.

But .. Offline RL is Challenging!



Distribution mismatch



Online vs. Offline

What Makes Offline Reinforcement Learning Difficult?

Distributional shift:
while our function approximator (policy, value function, or model) might be trained under one distribution, it will be evaluated on a different distribution, due both to the change in visited states for the new policy and, more subtly, by the act of maximizing the expected return.

The expectation of the number of mistakes

$$\ell(\pi) = \mathbb{E}_{p_\pi(\tau)}\left[\sum_{t=0}^{H} \delta(\mathbf{a}_t \neq \mathbf{a}_t^\star)\right].$$

Theorem 2.1 (Behavioral cloning error bound). *If $\pi(\mathbf{a}|\mathbf{s})$ is trained via empirical risk minimization on $\mathbf{s} \sim d^{\pi_\beta}(\mathbf{s})$ and optimal labels $\mathbf{a}^\star$, and attains generalization error ϵ on $\mathbf{s} \sim d^{\pi_\beta}(\mathbf{s})$, then $\ell(\pi) \leq C + H^2\epsilon$ is the best possible bound on the expected error of the learned policy.* offline

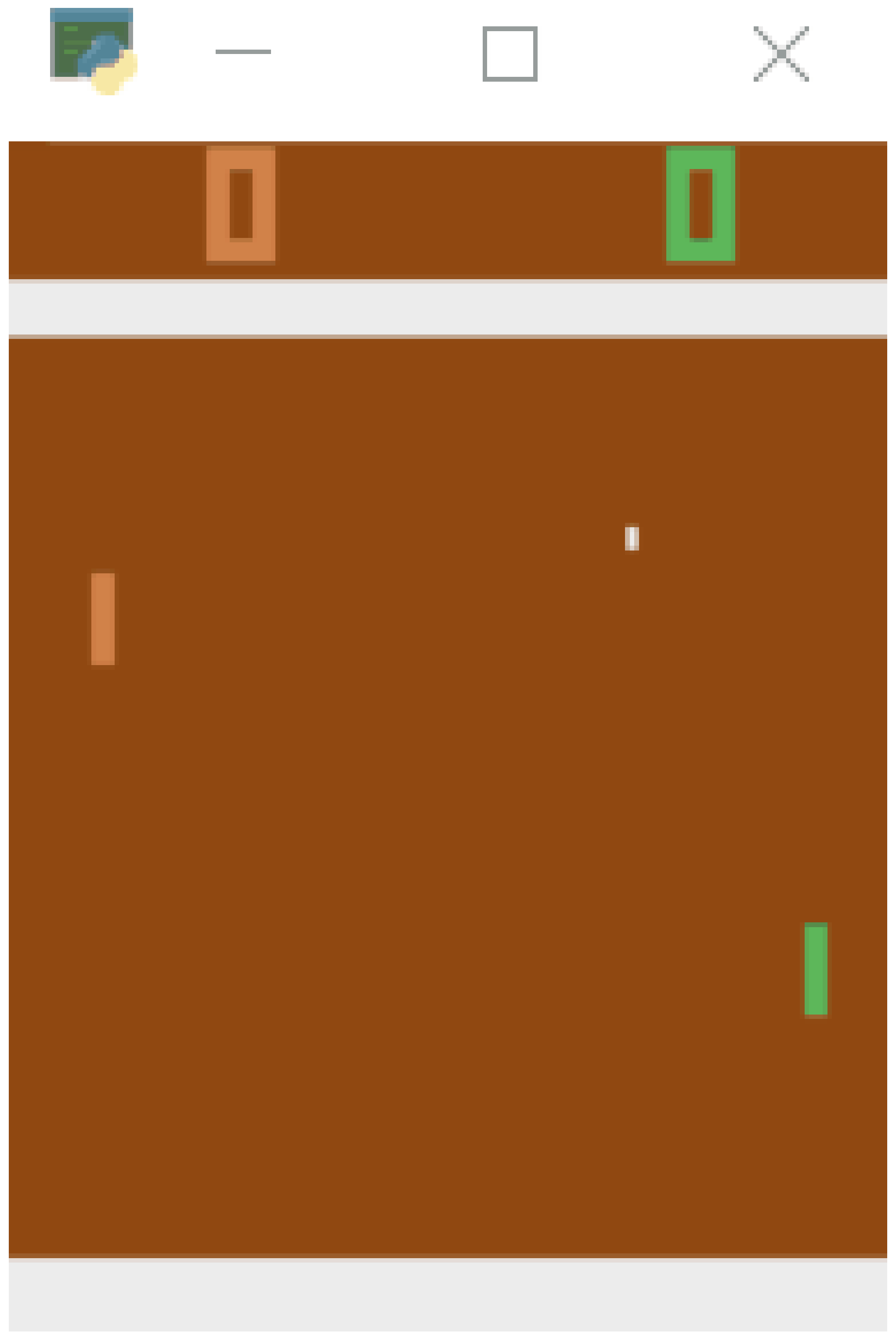
Theorem 2.2 (DAgger error bound). *If $\pi(\mathbf{a}|\mathbf{s})$ is trained via empirical risk minimization on $\mathbf{s} \sim d^{\pi}(\mathbf{s})$ and optimal labels $\mathbf{a}^\star$, and attains generalization error ϵ on $\mathbf{s} \sim d^{\pi}(\mathbf{s})$, then $\ell(\pi) \leq C + H\epsilon$ is the best possible bound on the expected error of the learned policy.* online

a short theoretical illustration of how harmful distributional shift can be on the performance of policies

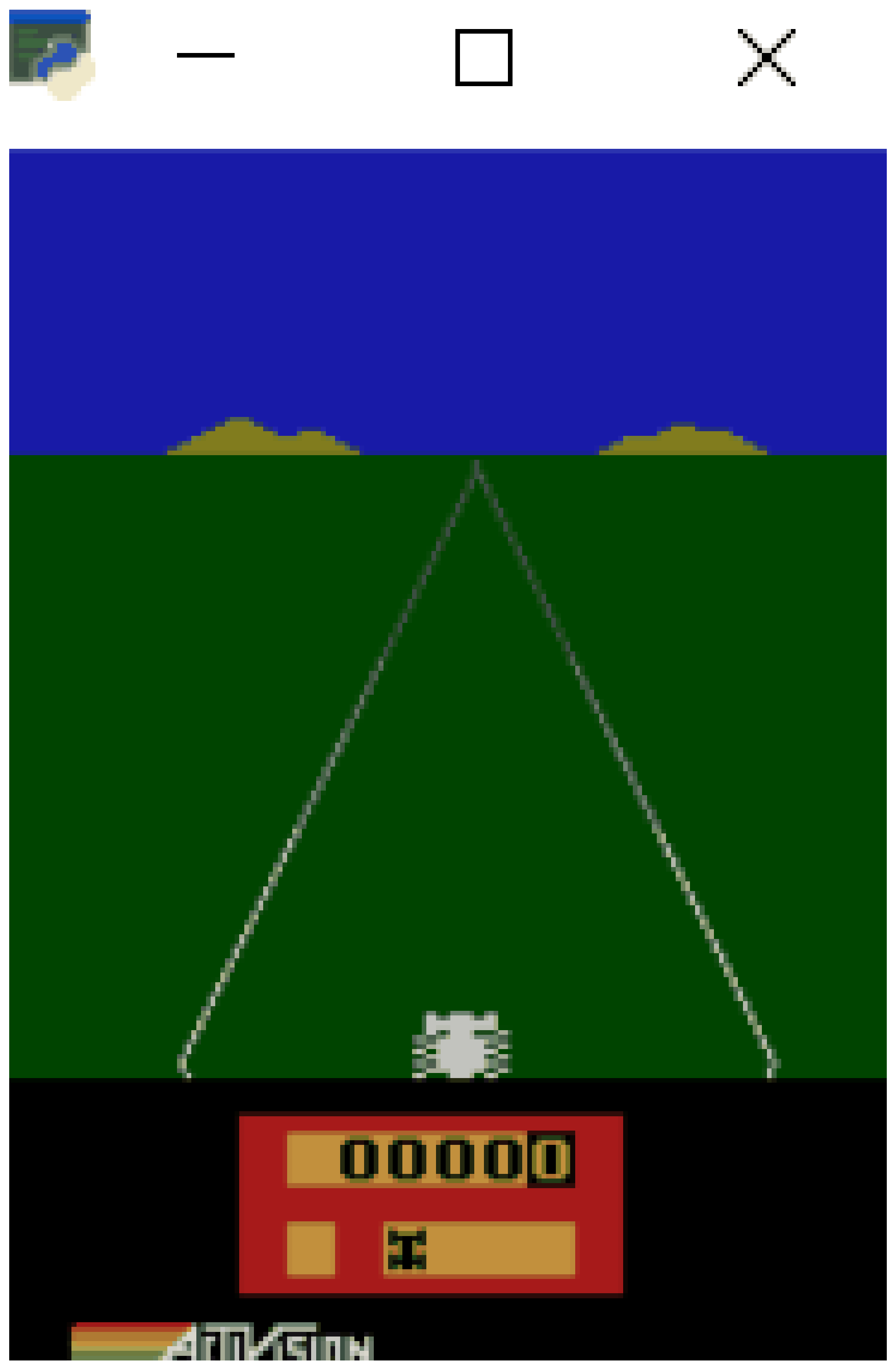
But .. Offline RL is Challenging!

No New Corrective Feedback

Offline Reinforcement Learning has a great potential but we should be careful when we deploy it in real-world production systems.



Pong



Enduro

Breakout

All trained by discrete BCQ, an offline RL algorithm.

Thank you