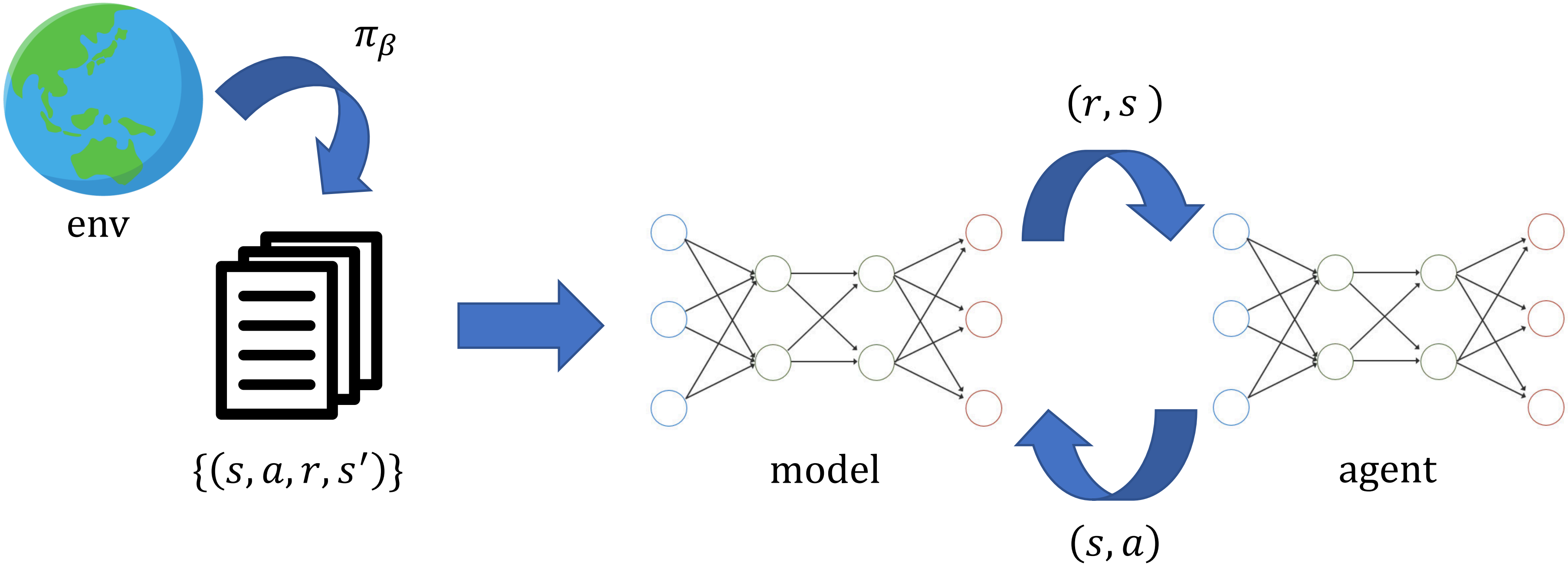
Weekly Report

M2 Ibuki Takeuchi

- Model-based offline RL(MOPO)



π_β

env

$\{(s, a, r, s')\}$

model

$(r, s\ )$

(s, a)

agent

Weekly Report

M2 Ibuki Takeuchi

- Reading a paper
  - S. Levine, A. Kumar, G. Tucker and Justin Fu. “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems ”, *arXiv preprint arXiv: 2005.01643,* 2020.

- Variation of offline RL
  - Policy gradient with importance sampling (difficult, low quality)
  - Approximate dynamic programming

$$Q_\theta^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q_\theta^\pi(s',a')]$$

π_β : data collection policy
π : learning policy

    - If $\underbrace{\Pr\left(a' \sim \pi_\beta(\cdot\,|s')\right)}_{\text{Action distributional shift}} = 0$, $Q_\theta^\pi(s',a')$ might returns high value erroneously

      (State shift can be ignored)
    - There is no method to evaluate unknown state $s_{unknown}$
  - Model-based approach

Weekly Report

M2 Ibuki Takeuchi

- Model-based offline RL (MOPO[1])
  - Estimate transition model $T(s'|s,a)$
  - Both action and state distributional shift should be concerned
  - Utilize uncertainty

- This week
  - Systematically summarize

[1] : T.Yu,G.Thomas,L.Yu,S.Ermon,J.Zou,S.Levine,C.FinnandT.Ma. “MOPO: Model-based Offline Policy Optimization” *arXiv preprint arXiv: 2005.13239,* 2020.