Unsupervised Reinforcement Learning

in the
Multiverse of Downstream Tasks

DMQA Open Seminar (2023.09.08)

Data Mining & Quality Analytics Lab.

차민성 (Minsung Cha)

발표자 소개

❖ 차민성 (Minsung Cha)

- 고려대학교 일반대학원 산업경영공학과 재학
- Data Mining & Quality Analytics Lab. (지도교수 : 김성범 교수님)
- M.S. Student(2023.03 ~ Present)

❖ Research Interest

- Reinforcement Learning
- Deep Learning Algorithms

❖ Contact

- E-mail : djpanda1217@korea.ac.kr

Table of Contents (1/2)

Table of Contents (2/2)

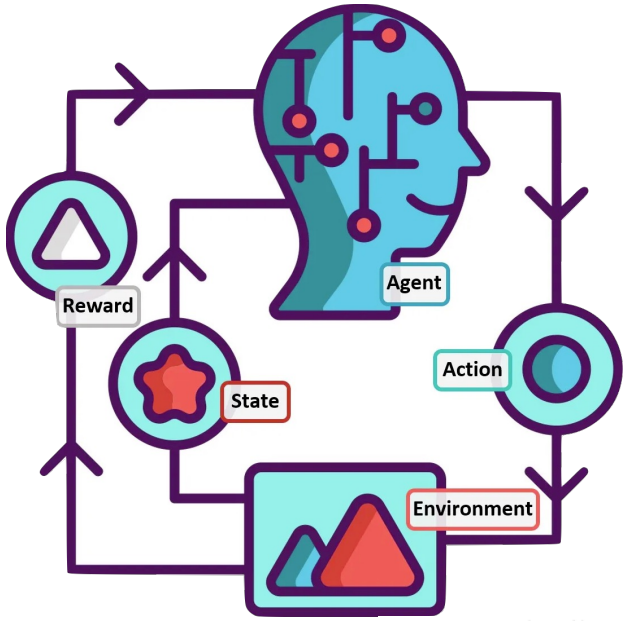
230908 DMQA Open Seminar:
Unsupervised Reinforcement Learning

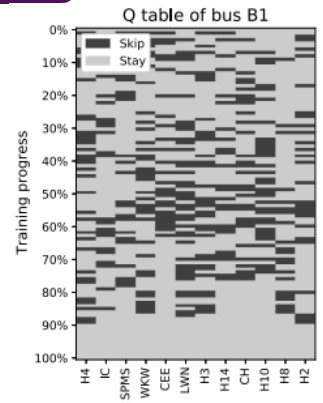
1. URL : Unsupervised Reinforcement Learning

- ✓ 강화학습 (Reinforcement Learning)
- ✓ 기존 강화학습의 한계
- ✓ Introducing URL : Unsupervised Reinforcement Learning
- ✓ Preliminaries
- ✓ Unsupervised Reinforcement Learning의 개념과 평가 방식

강화학습(Reinforcement Learning)

URL : Unsupervised Reinforcement Learning





❖ 강화학습(Reinforcement Learning)이란?

- **Agent가 환경Environment와 상호작용하며 학습**하도록 하는 머신러닝 학습 방법론
  - 어떤 환경의 특정 상태State에서
  - Agent가 행동Action을 취하면
  - 이에 맞추어 환경은 변화하고
  - Agent에게는 환경으로부터 보상Reward이 입력되며
  - Agent는 환경을 관찰하여 다음 상태를 입력받음
- Agent는 **누적 보상의 합Sum of Cumluative Reward이 최대**가 되는 것을 목표로 **최적의 행동을 학습**
- 로봇과 같이 복잡한 Control 과업이나 의사 결정, 할당 문제들을 해결하는 데에 뛰어난 능력을 보이고 있음

L. Vismara, et al. "Optimal assignment of buses to bus stops in a loop by reinforcement learning" Physica A (2021)
Y. Tassa, et al. "dm_control: Software and Tasks for Continuous Control" arXiv (2020)
Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

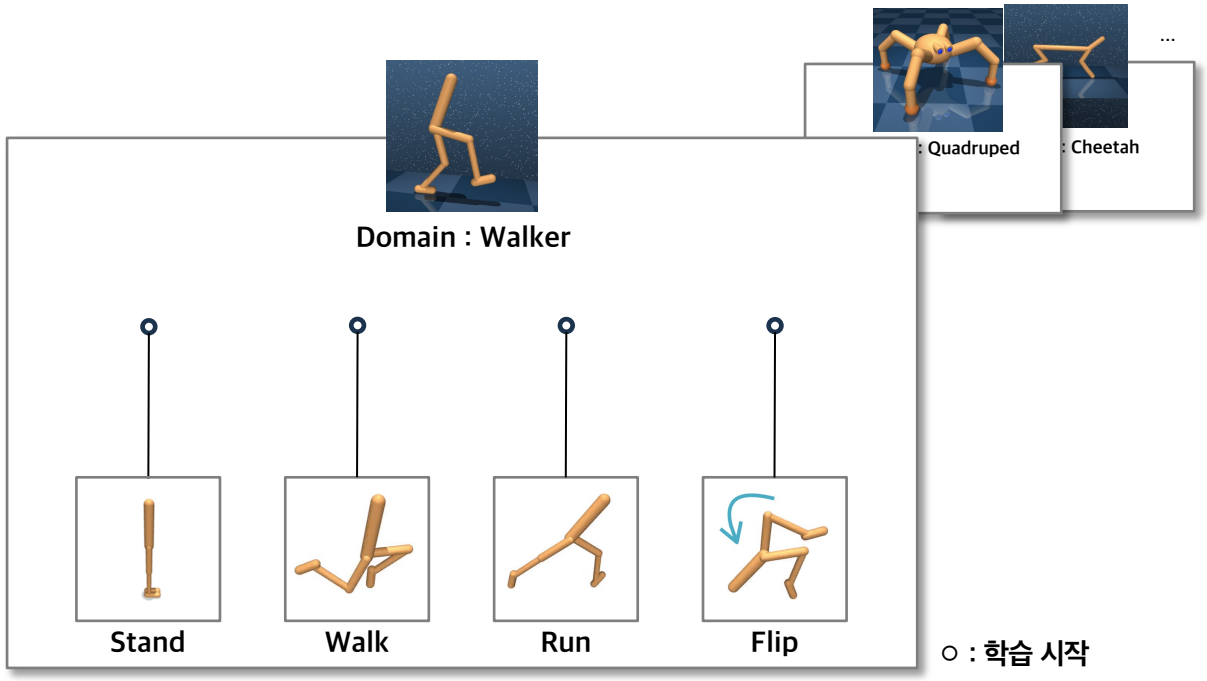
기존 강화학습의 한계

URL : Unsupervised Reinforcement Learning

❖ 일반화의 어려움

- 강화학습은 특정 Task를 풀 수 있는 강력한 능력을 갖출 수 있으나, **환경이 제공하는 보상Reward은 task-specific**
- **학습한 Task만 수행할 수 있는 한계점**을 갖고 있음Powerful but Narrow

  → 같은 Domain이라도 새로운 Task가 등장할 때마다 처음부터 학습시켜야하기 때문에 **비효율적**이고, Domain 내 **일반화가 어려움**



Y. Tassa, et al. “dm_control: Software and Tasks for Continuous Control” arXiv (2020)

기존 강화학습의 한계

URL : Unsupervised Reinforcement Learning

❖ 일반화의 어려움

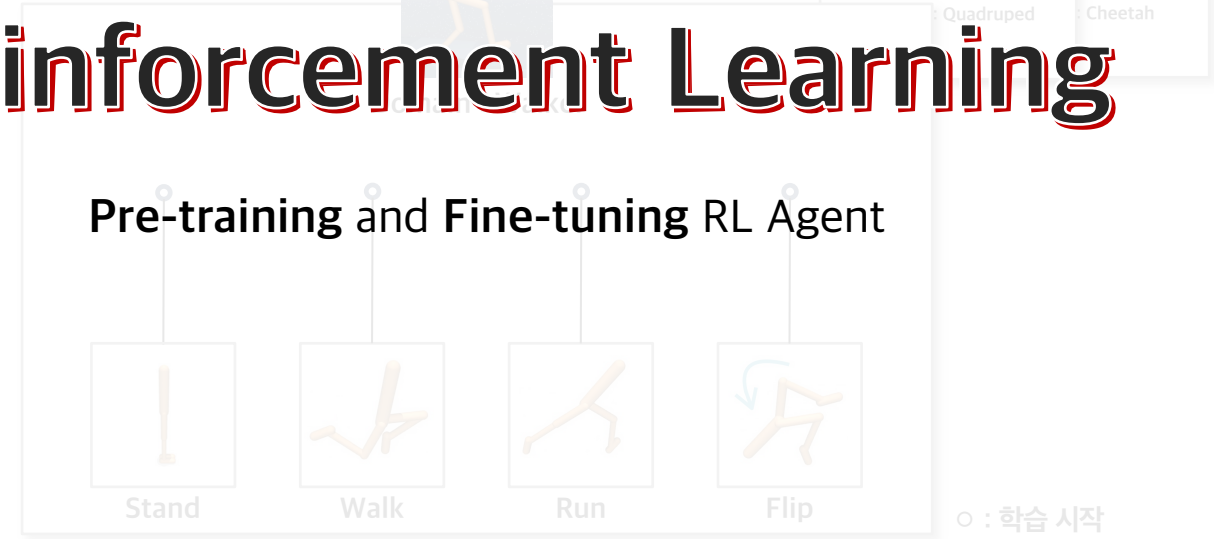
- 강화학습은 특정 Task를 풀 수 있는 강력한 능력을 갖출 수 있으나 **환경이 제공하는 보상**Reward**은 task-specific**
- **학습한 Task만 수행할 수 있는 한계점**을 갖고 있음

→ 같은 Domain이라도 새로운 Task가 등장할 때마다 처음부터 학습시켜야하기 때문에 **비효율적**이고, Domain 내 **일반화가 어려움**

URL

Unsupervised Reinforcement Learning

Pre-training and **Fine-tuning** RL Agent



○ : 학습 시작

Y. Tassa, et al. “dm_control: Software and Tasks for Continuous Control” arXiv (2020)

Introducing URL : Unsupervised Reinforcement Learning

❖ URL : Unsupervised Reinforcement Learning (비지도 강화학습)

- 각 Task 별로 처음부터 학습을 시작하는 기존 강화학습과 달리, URL은 **두 개의 step**으로 이루어짐

Pre-training and **Fine-tuning** RL Agent

Pre-training

- ✓ for **Domain**
- ✓ with **Self-supervised Task**
- ✓ **Intrinsic** Reward
- ✓ Goal : **Generalization**

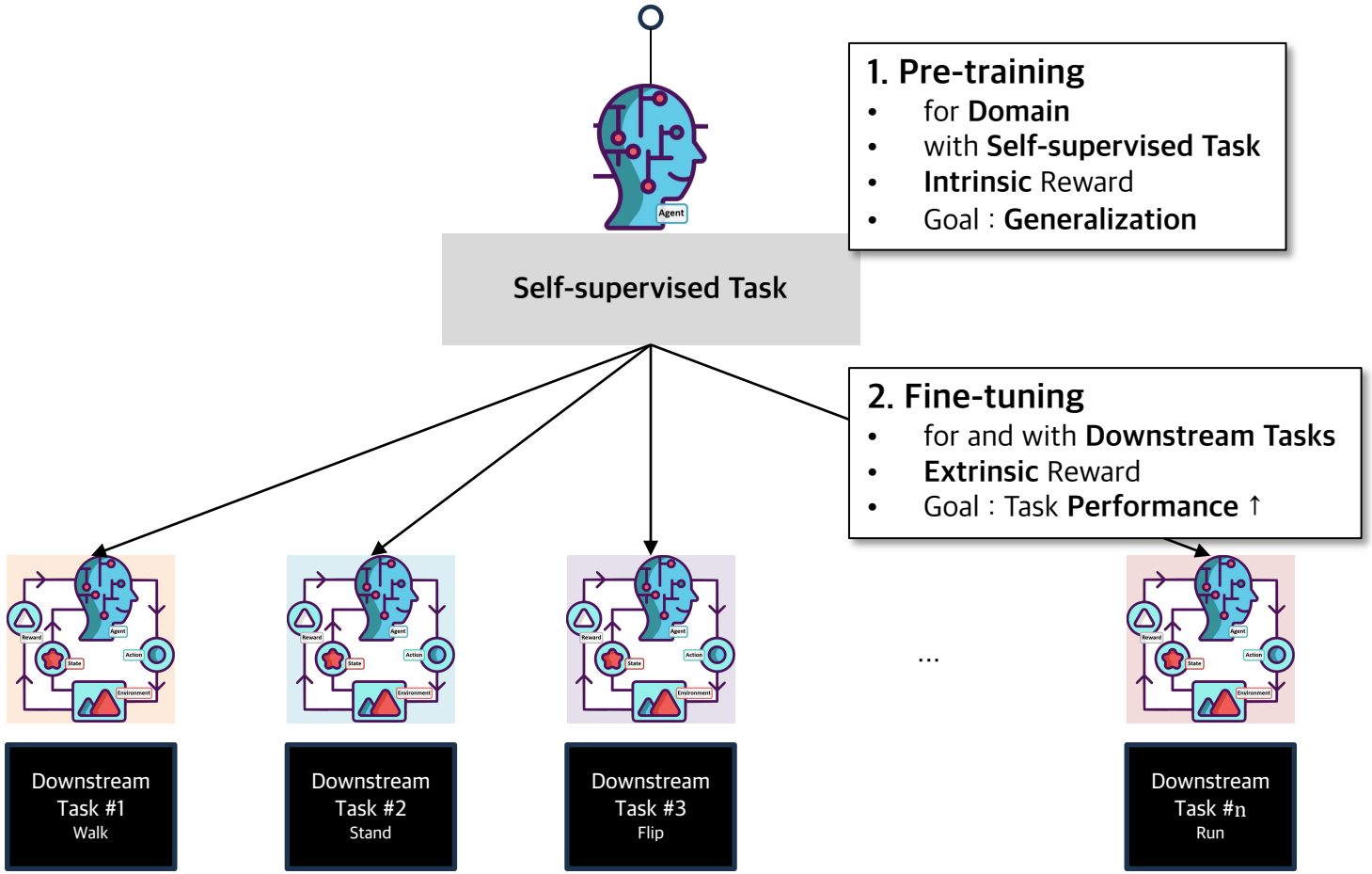
Fine-tuning

- ✓ for and with **Downstream Task**
- ✓ **Extrinsic** Reward
- ✓ Goal : **Task Performance** ↑

Introducing URL : Unsupervised Reinforcement Learning

❖ URL : Unsupervised Reinforcement Learning (비지도 강화학습)

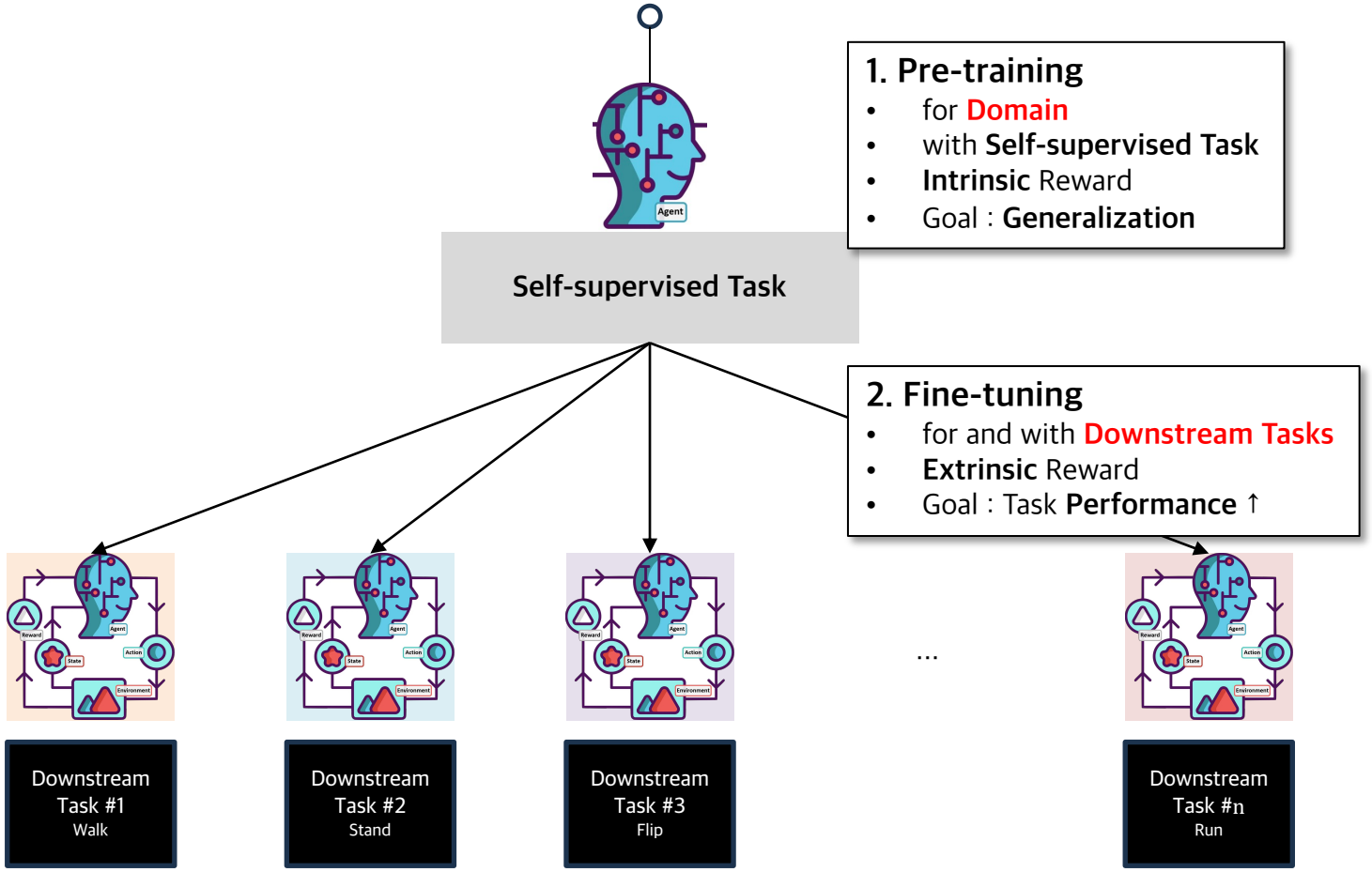
- 각 Task 별로 처음부터 학습을 시작하는 기존 강화학습과 달리, URL은 **두 개의 step**으로 이루어짐



Reinforcement Leaning
(image by Flat-Icons on IconScout under license to Chris Mahoney)

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

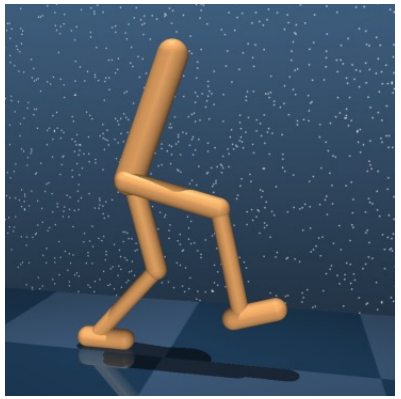


Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

Domain

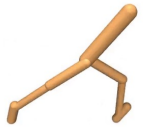


Walker

Downstream Task

Stand

Downstream Task

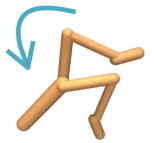


Walk

Downstream Task

Run

Downstream Task

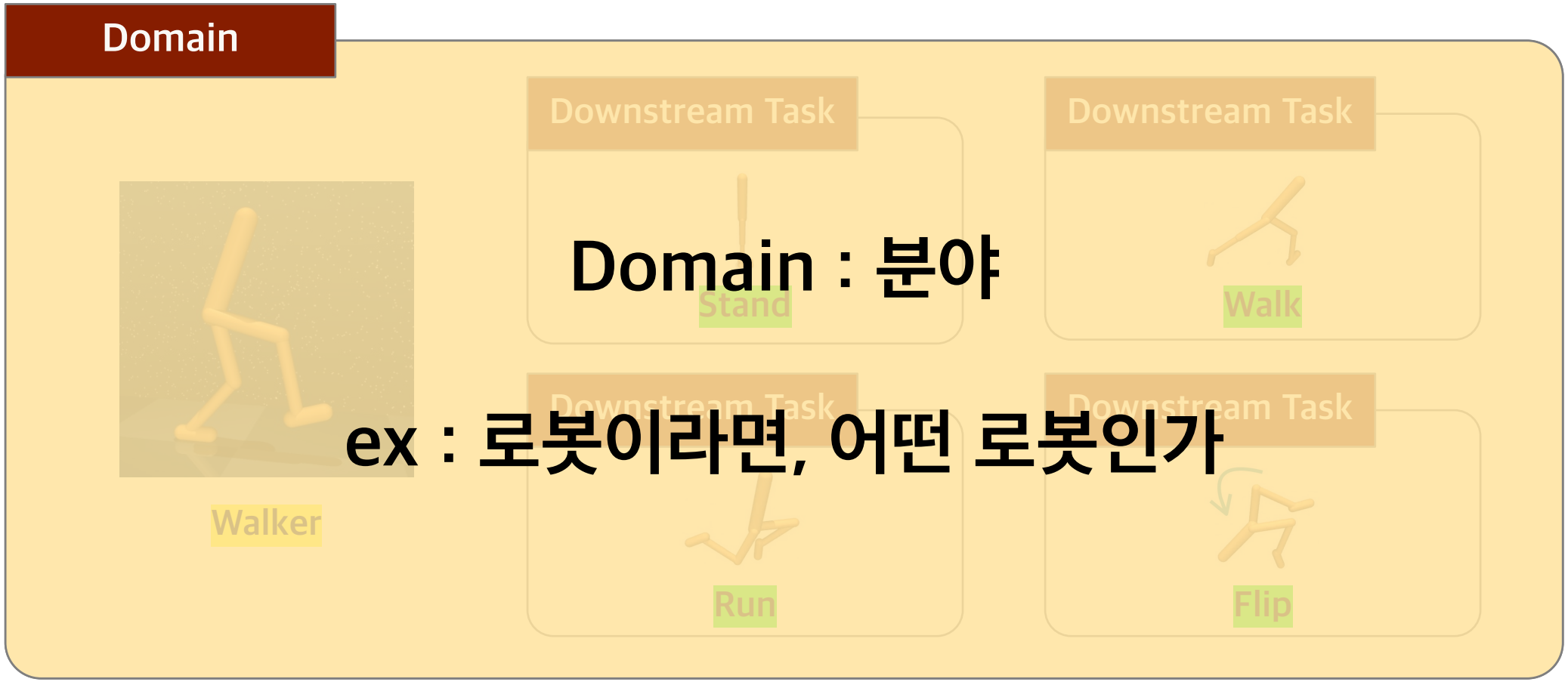


Flip

Y. Tassa, et al. “dm_control: Software and Tasks for Continuous Control” arXiv (2020)
Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

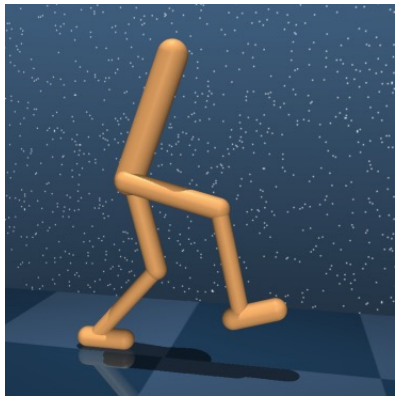


Y. Tassa, et al. “dm_control: Software and Tasks for Continuous Control” arXiv (2020)
Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

Domain



Walker

Downstream Task

Stand

Downstream Task

Walk

Downstream Task : 과업

ex) 그 로봇이 구체적으로 어떤 일을 할 것인가?

Run

Flip

Y. Tassa, et al. “dm_control: Software and Tasks for Continuous Control” arXiv (2020)

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

❖ URLB (Unsupervised Reinforcement Learning) 의 예시

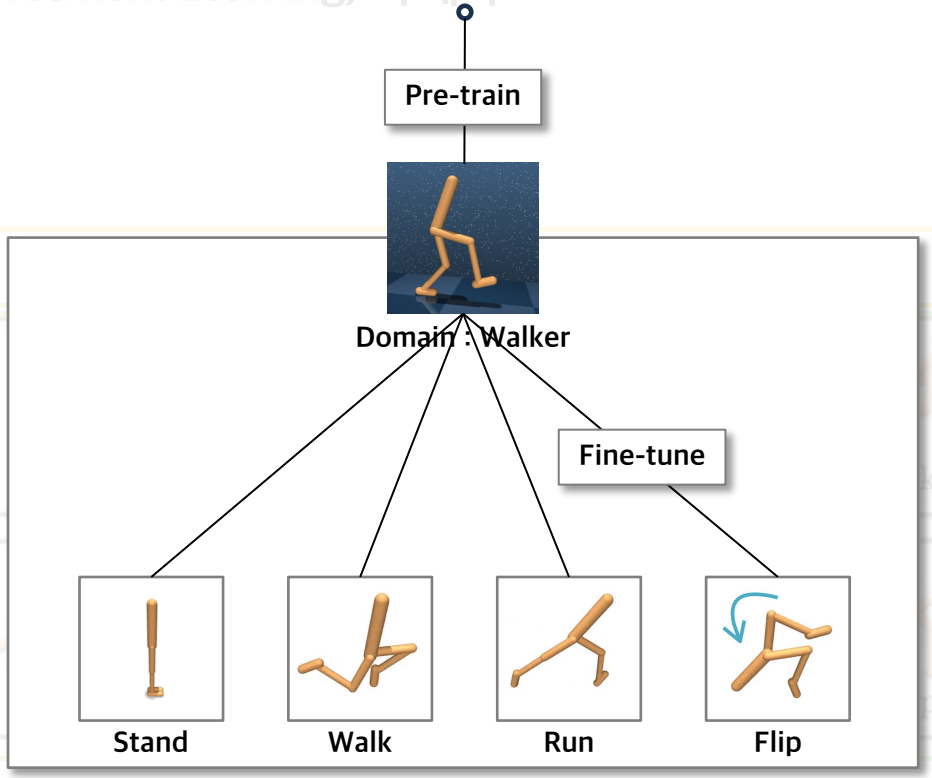
Domain	Walker		Quadruped		Jaco Arm	
Downstream Task	 Stand	 Walk	 Stand	 Walk	 Top left	 Top right
	 Run	 Flip	 Run	 Jump	 Bottom left	 Bottom right

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Preliminaries : Domain, Downstream Task

URL : Unsupervised Reinforcement Learning

❖ URLB (Unsupervised Reinforcement Learning) 의 예시

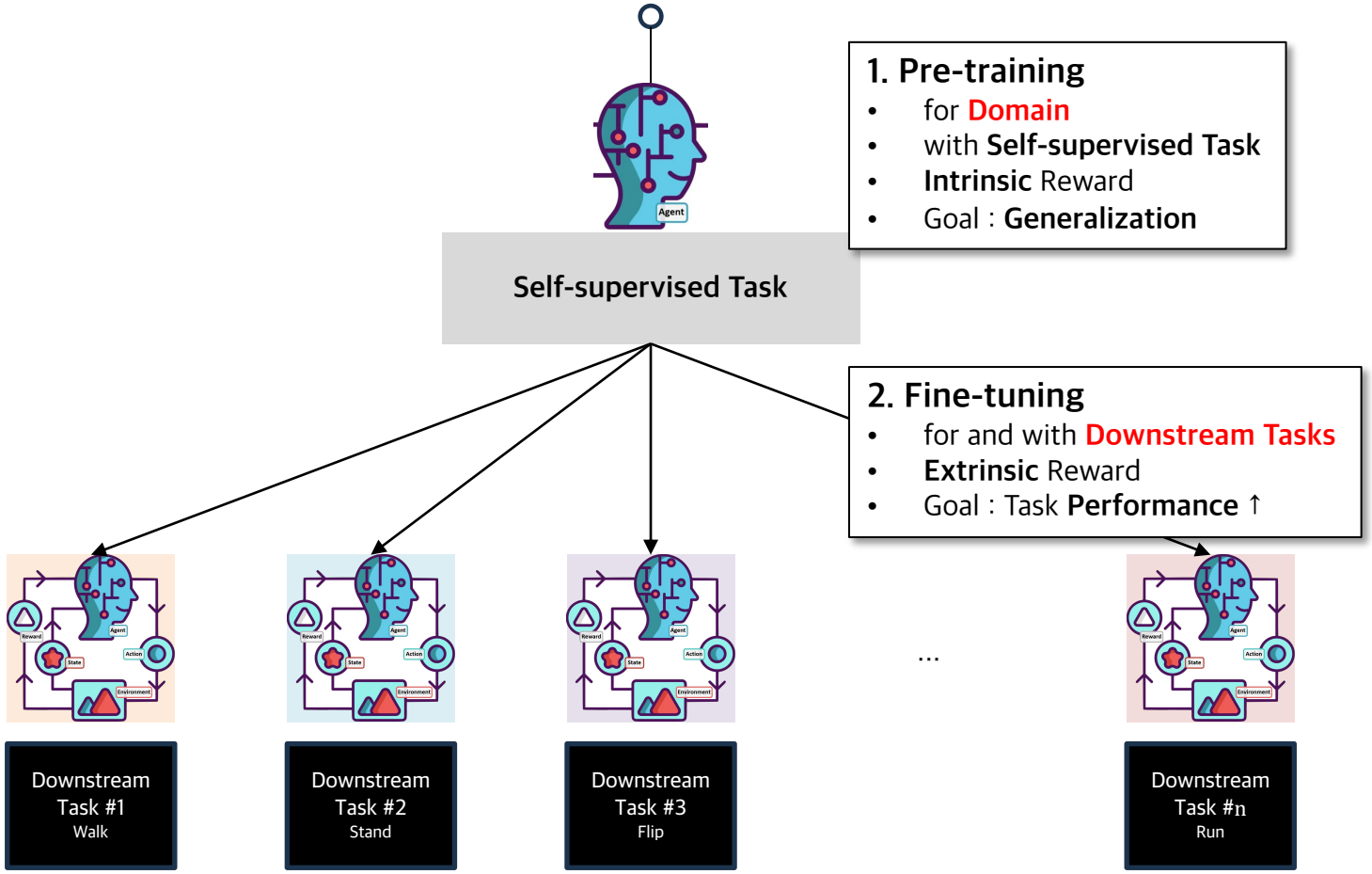


○ : 학습 시작

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Preliminaries : Domain, Downstream Task

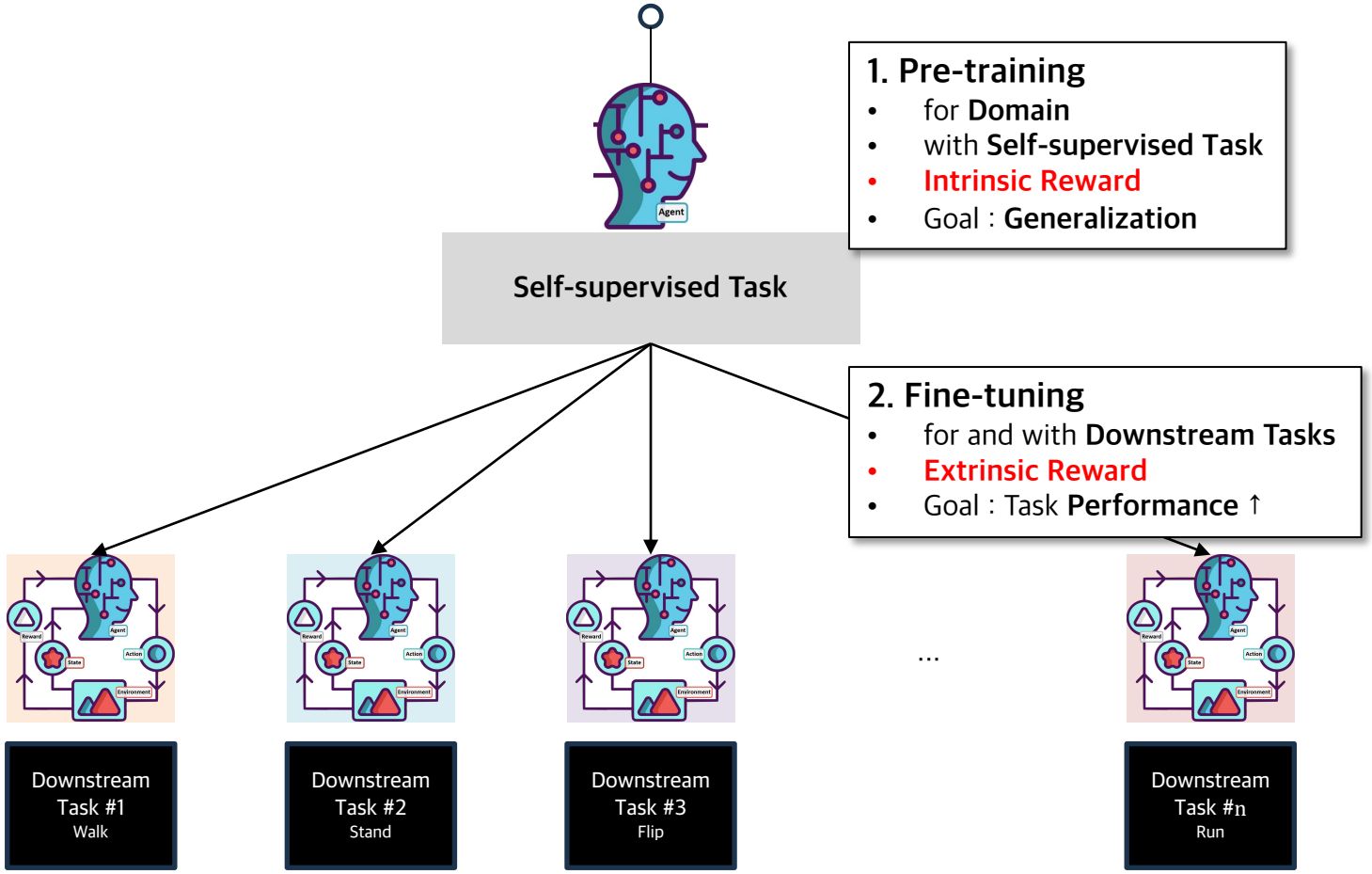
URL : Unsupervised Reinforcement Learning



Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Preliminaries : Extrinsic and Intrinsic Reward

URL : Unsupervised Reinforcement Learning



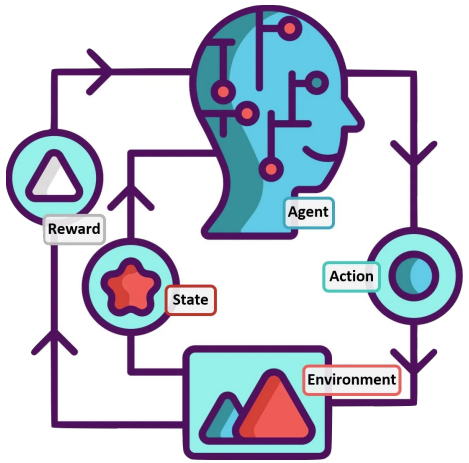
Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Preliminaries : Extrinsic and Intrinsic Reward

URL : Unsupervised Reinforcement Learning

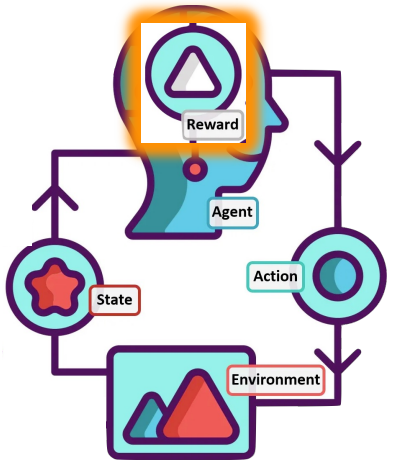
❖ Extrinsic Reward (외부 보상)

- 일반적인 강화학습에서 사용하는 보상Reward과 동일
- Agent가 행동Action을 취하면, **특정 과업Task을 달성하는 데에 어떤 영향을 미쳤는지**를 바탕으로 **환경Environment이 Agent에게 제공**하는 보상Reward
- Agent의 외부인 환경Environment이 제공하는 보상이므로 외부 보상이라고 부름
- 특정 과업Task에 대해서 부여되는 점수이므로 **task-specific**한 보상Reward



❖ Intrinsic Reward (내부 보상)

- **Agent가 행동Action을 취한 뒤, 자기지도Self-supervised방식으로 스스로 계산**하는 보상Reward
- 환경으로부터 제공받는 것이 아닌, Agent 내부에서 스스로 계산하기 때문에 내부 보상이라고 부름
- 특정 task에 대한 것이 아닌, **task-agnostic**한 보상Reward



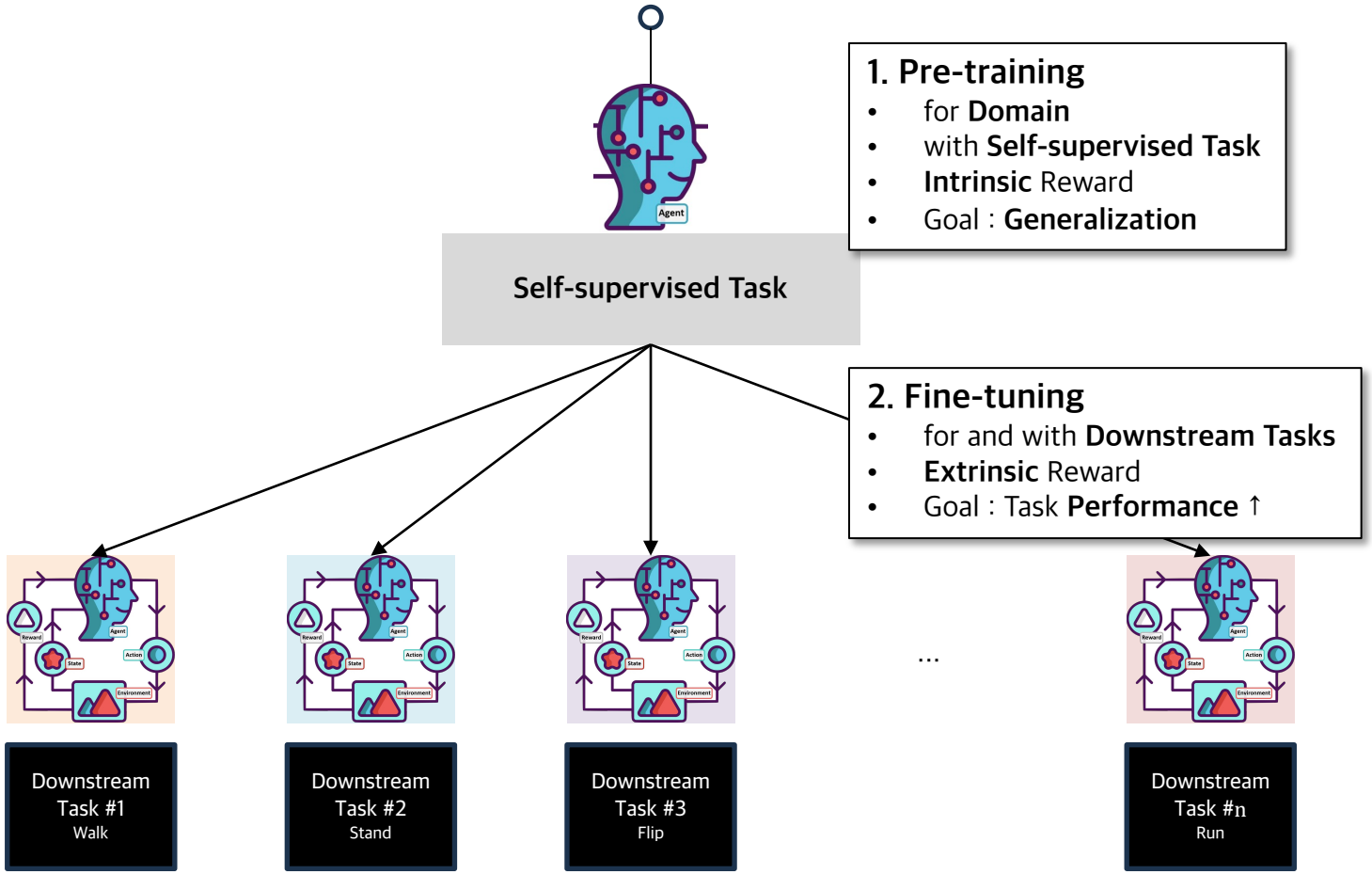
Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Unsupervised Reinforcement Learning의 개념

URL : Unsupervised Reinforcement Learning

❖ URL : Unsupervised Reinforcement Learning (비지도 강화학습)

- 각 Task 별로 처음부터 학습을 시작하는 기존 강화학습과 달리, URL은 **두 개의 step**으로 이루어짐



Reinforcement Leaning
(image by Flat-Icons on IconScout under license to Chris Mahoney)

Unsupervised Reinforcement Learning의 개념

URL : Unsupervised Reinforcement Learning

❖ URL Step 1 : Pre-training

- 구체적인 Task가 무엇인지는 상관 없이,

  Domain 전체에 대해서 진행되는 과정

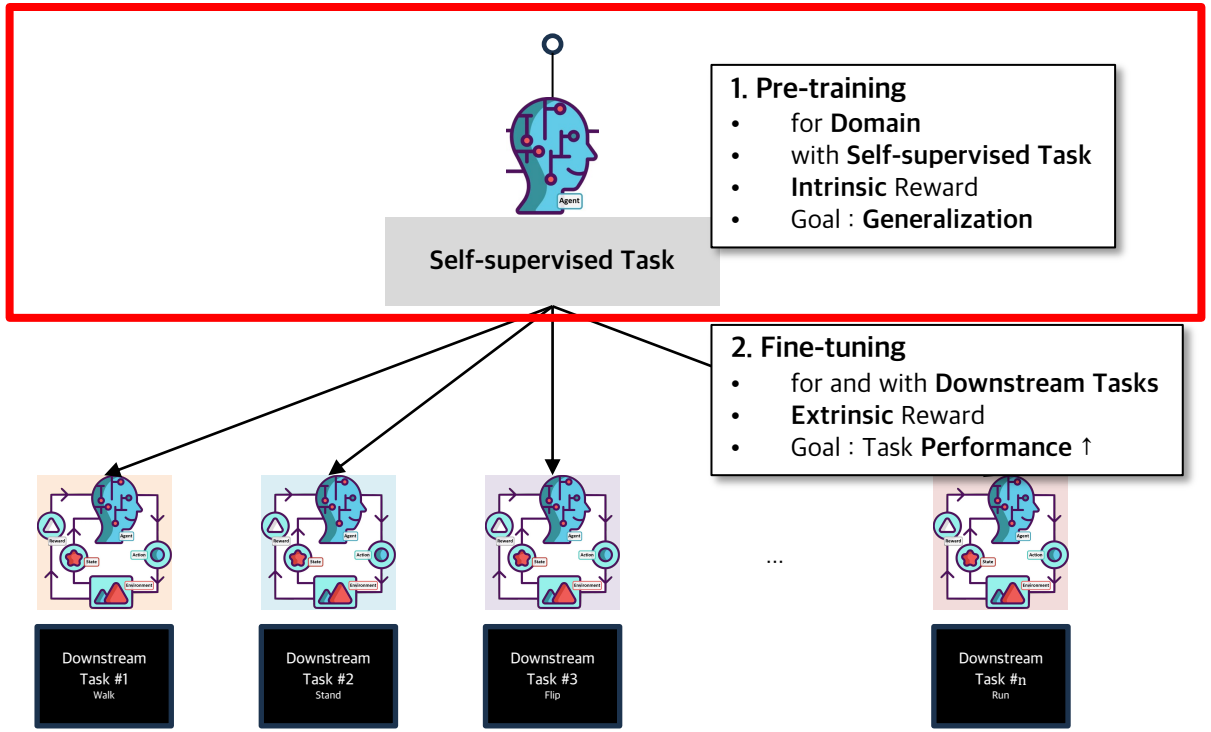
- **(Extrinsic) Reward-free**
  - Task 수행에 대한 결과로 얻는 외부 보상이 아닌,

    Agent가 자기지도Self-supervised방식으로 스스로 계산하는

    task-agnostic한 **내부 보상**Intrinsic Reward, r_t^i 사용
  - 높은 내부 보상Intrinsic Reward, r_t^i 를 얻기 위해

    Self-supervised Task를 학습
- **목적 : Generalization**
  - Domain 내에서 task에 상관 없이 환경Environment을 탐색Exploration

    하거나, 유용한 기술Skill을 습득



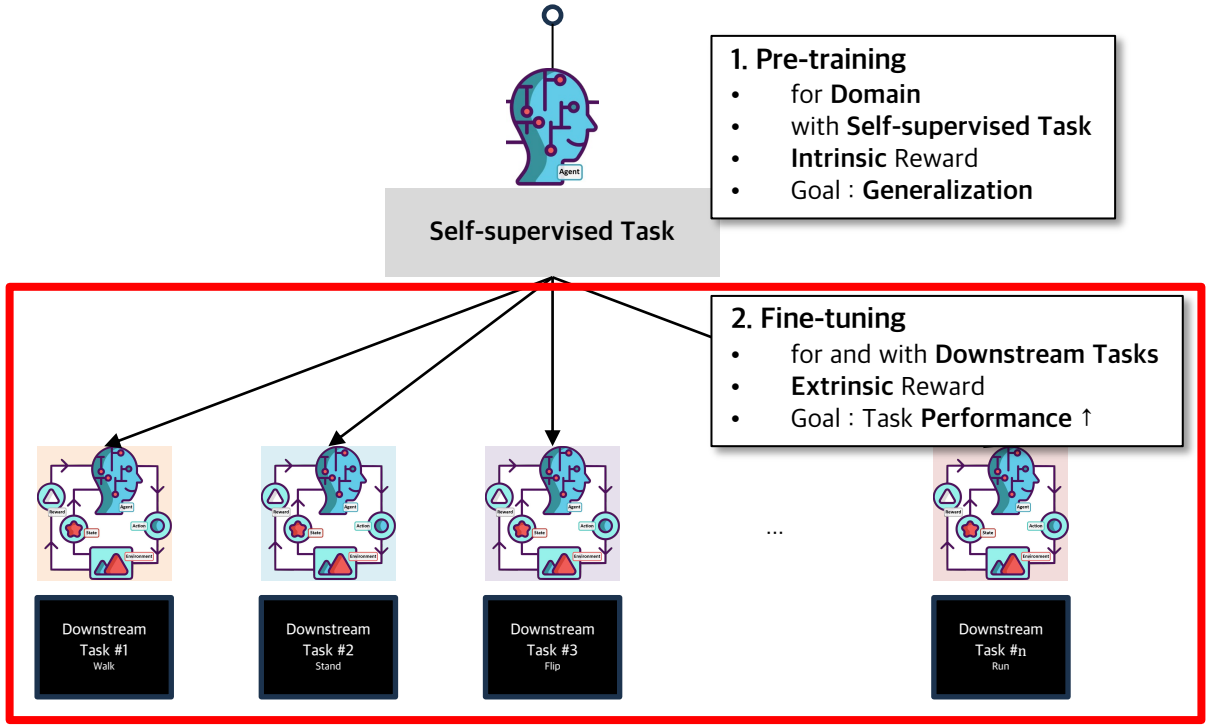
Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Unsupervised Reinforcement Learning의 개념

URL : Unsupervised Reinforcement Learning

❖ URL Step 2 : Fine-tuning

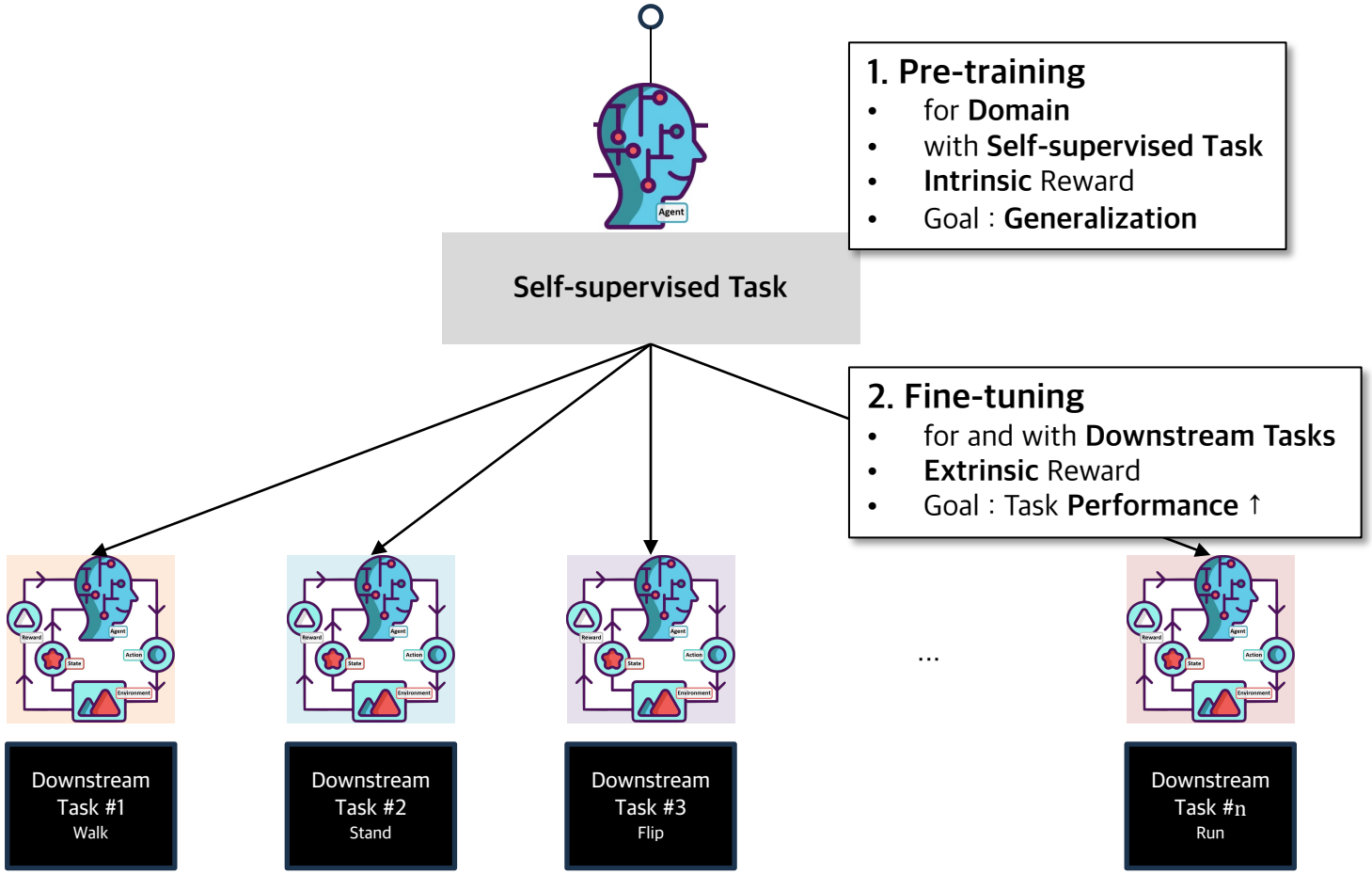
- Domain 내의 특정 **Downstream Task**에 대해 진행되는 과정
- **Reward** :
  - 최종 목표인 Downstream Task를 수행하기 위해 행동Action을 취하고, 그 결과 환경Environment로부터 얻는, task-specific한 **외부 보상**Extrinsic Reward, r_t^e 활용
  - 일반적인 강화학습에서 진행하는 방식과 같음
- **목적 : 특정 Downstream Task 수행 능력 강화**
  - 최종 목적이 되는 특정 Downstream Task에 대한 외부 보상의 합을 최대화하여 해당 Task 수행 능력 강화



Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Unsupervised Reinforcement Learning의 개념

URL : Unsupervised Reinforcement Learning



Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Unsupervised Reinforcement Learning의 개념

URL : Unsupervised Reinforcement Learning

URL

Unsupervised Reinforcement Learning

Pre-training and **Fine-tuning** RL Agent

Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

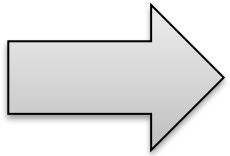
Unsupervised Reinforcement Learning의 평가 방식

URL : Unsupervised Reinforcement Learning

Pre-training and **Fine-tuning** RL Agent

Pre-training

- ✓ for **Domain**
- ✓ with **Self-supervised Task**
- ✓ **Intrinsic** Reward
- ✓ Goal : **Generalization**



Fine-tuning

- ✓ for and with **Downstream Task**
- ✓ **Extrinsic** Reward
- ✓ Goal : **Task Performance** ↑

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

Unsupervised Reinforcement Learning의 평가 방식

URL : Unsupervised Reinforcement Learning

Goal :

Agent에게 Domain 내에서 Pre-training을 잘 시켜서
Domain 내에 각 Task에 대하여
Fine-tuning을 통해서 보다 빠르게 Task 성능을 끌어올릴 수 있게 하자!

Evaluation :

제한된 Fine-tuning step 안에 얼마나 좋은 성능에 도달할 수 있는가?

URLB

Unsupervised Reinforcement Learning Benchmark

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., ... & Abbeel, P. (2021). URLB: Unsupervised reinforcement learning benchmark. *arXiv preprint arXiv:2110.15191*.

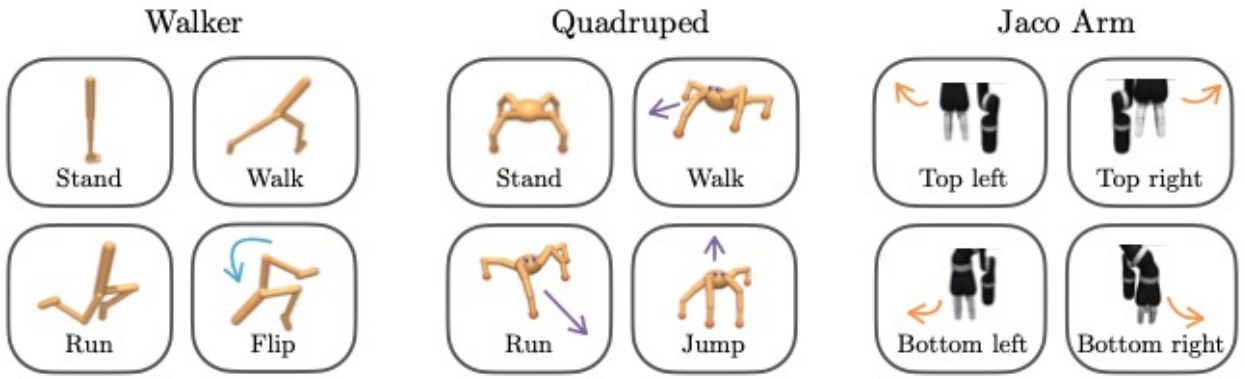
Unsupervised Reinforcement Learning의 평가 방식

URL : Unsupervised Reinforcement Learning

❖ URLB : Unsupervised Reinforcement Learning Benchmark [NeurIPS 2021]

- URL 방법론들에 대해서 통합된 지표로 평가할 수 있는 벤치마크를 제시
  - **평가 방식 : 제한된 fine-tuning step(100k steps)안에 얼마나 높은 점수를 달성할 수 있는가?**
  - DeepMind Control Suite 기반
  - 총 12개의 task:

    세 개의 도메인(Walker, Quadruped, Jaco Arm)에 대하여, 각각 네 개의 task 존재



Laskin, Michael, et al. "URLB: Unsupervised Reinforcement Learning Benchmark." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

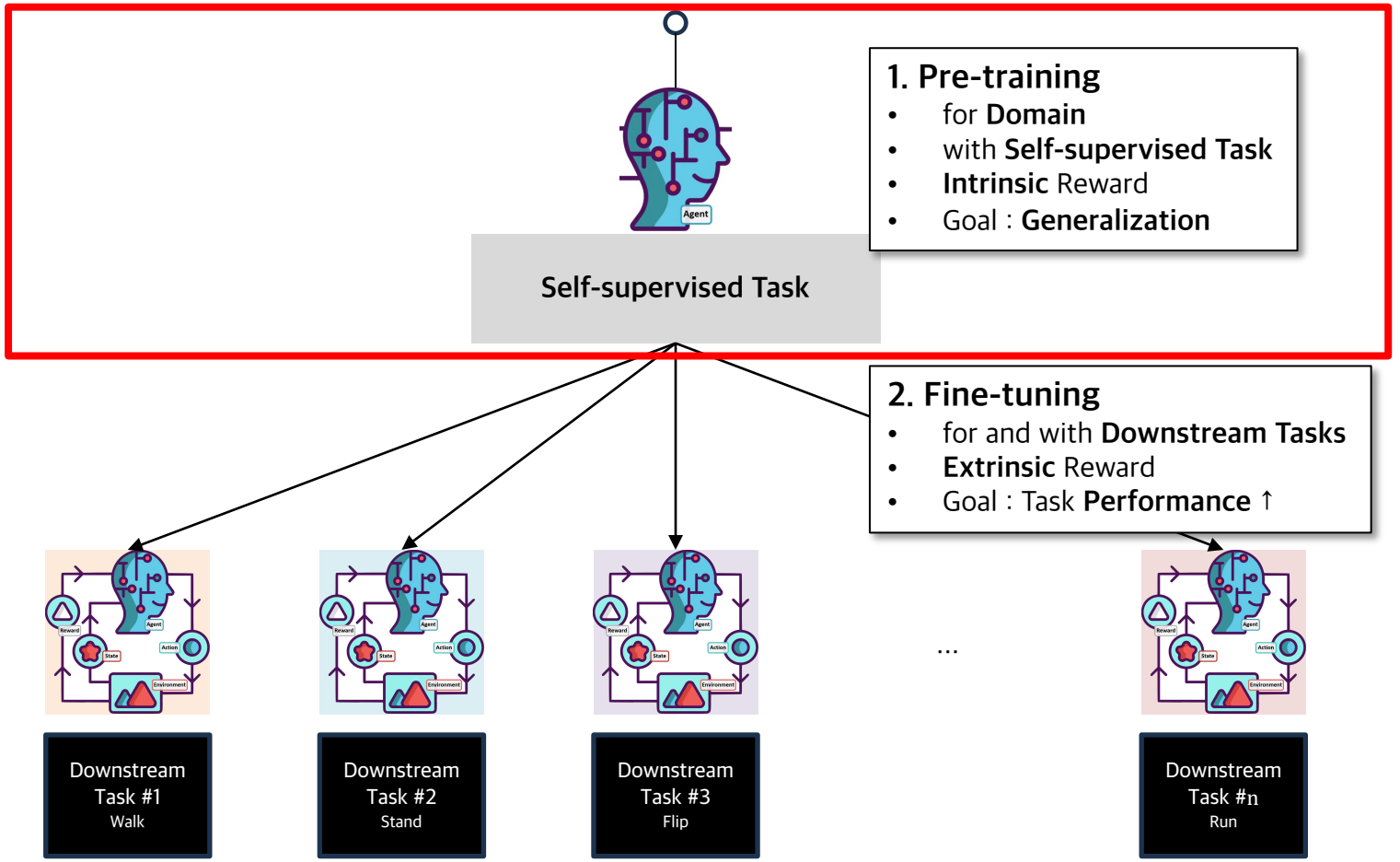
230908 DMQA Open Seminar:
Unsupervised Reinforcement Learning

2. URL의 종류

- ✓ URL의 종류 한 눈에 알아보기
- ✓ Knowledge-based URL : ICM [ICML 2017], Disagreement [ICML 2019]
- ✓ Data-based URL : ProtoRL [ICML 2021]
- ✓ Competence-based URL : CIC [NeurIPS 2022], MOSS [NeurIPS 2022]

URL의 종류 한 눈에 알아보기

URL의 종류



Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

URL의 종류 한 눈에 알아보기

URL의 종류

어떻게 Pre-train 시킬 것인가에 따라
URL은 크게 세 종류로 구분할 수 있음

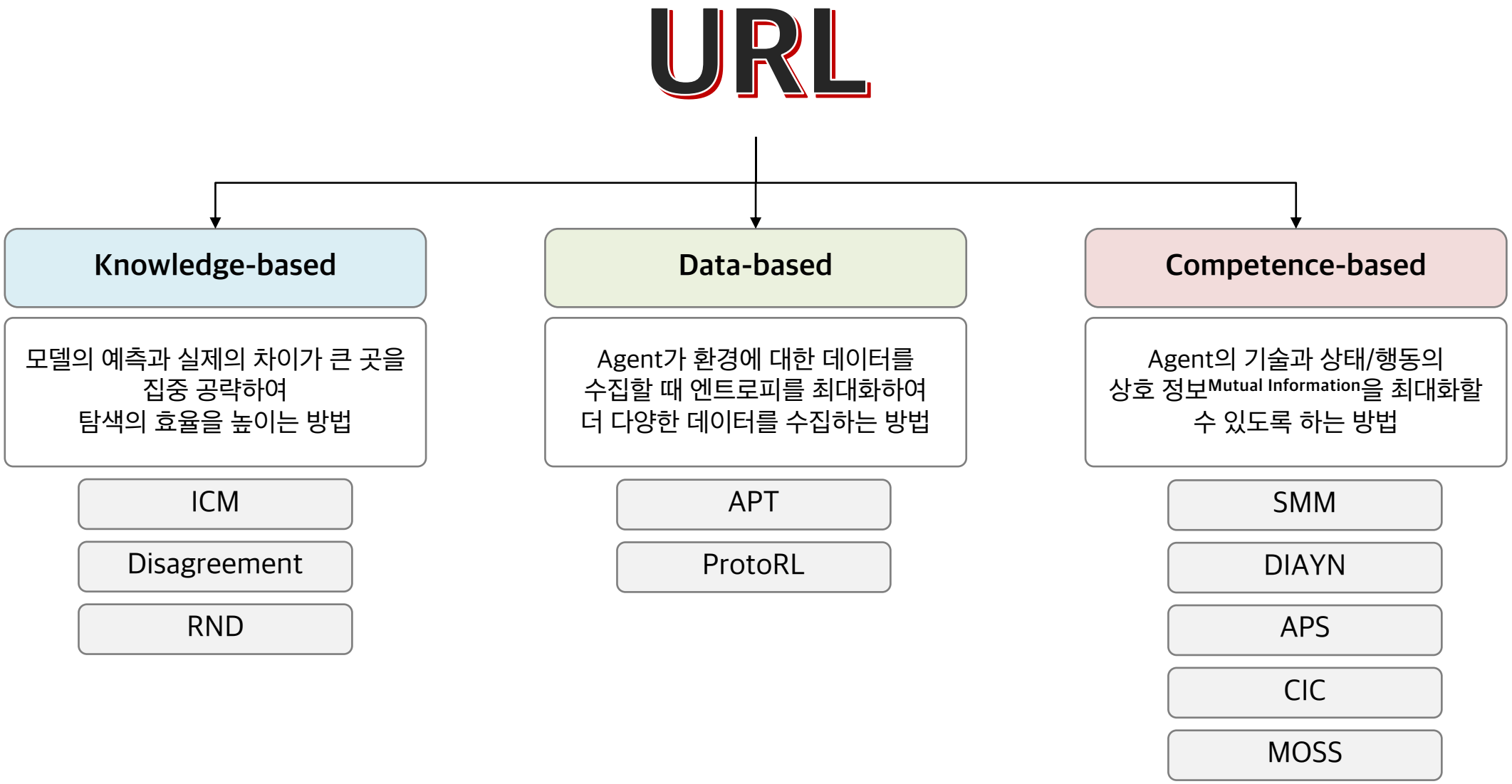
=

What **Self-supervised Task/Intrinsic Reward(r_t^i)** is used?

Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

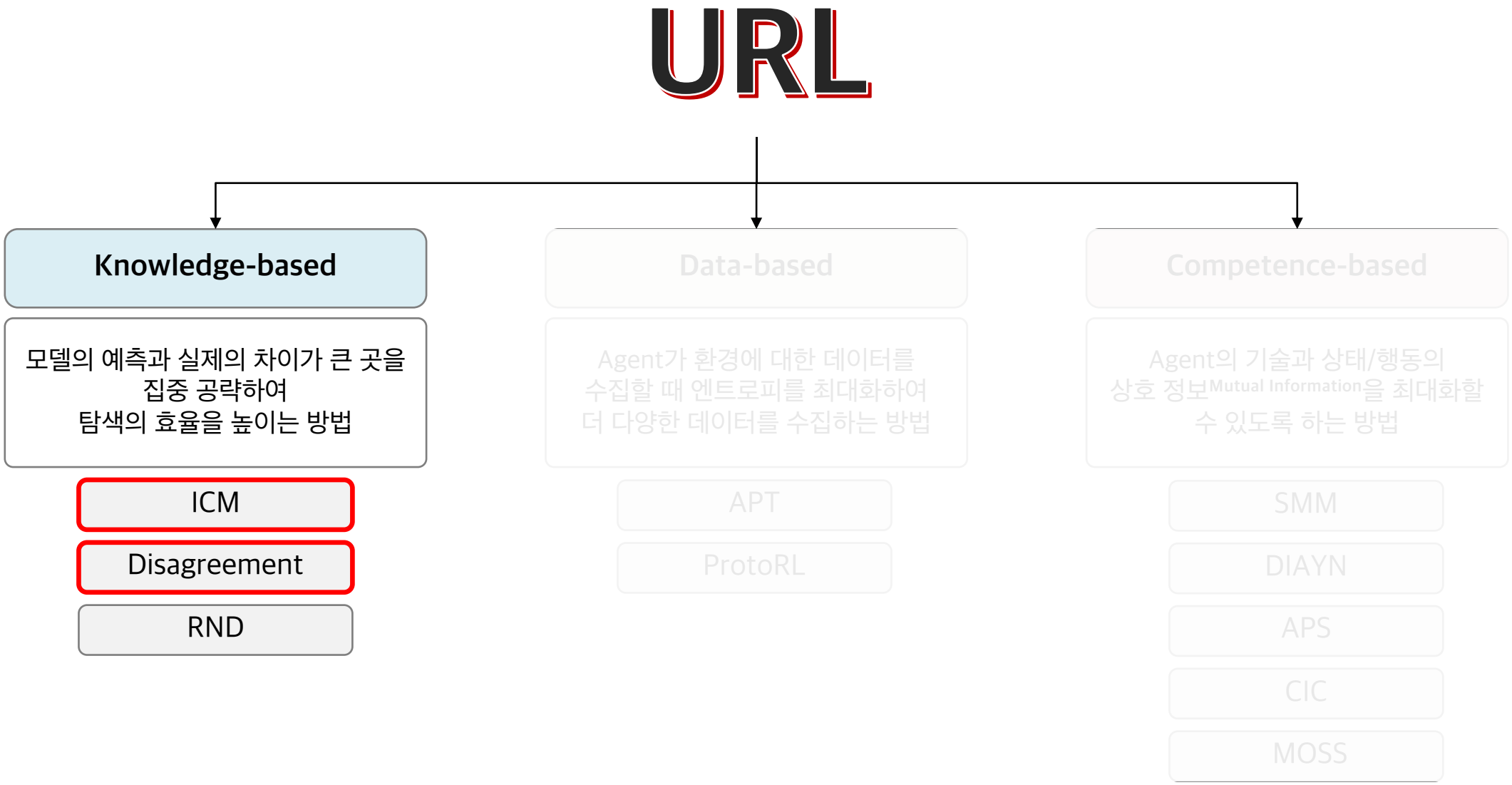
URL의 종류 한 눈에 알아보기

URL의 종류



URL의 종류 한 눈에 알아보기

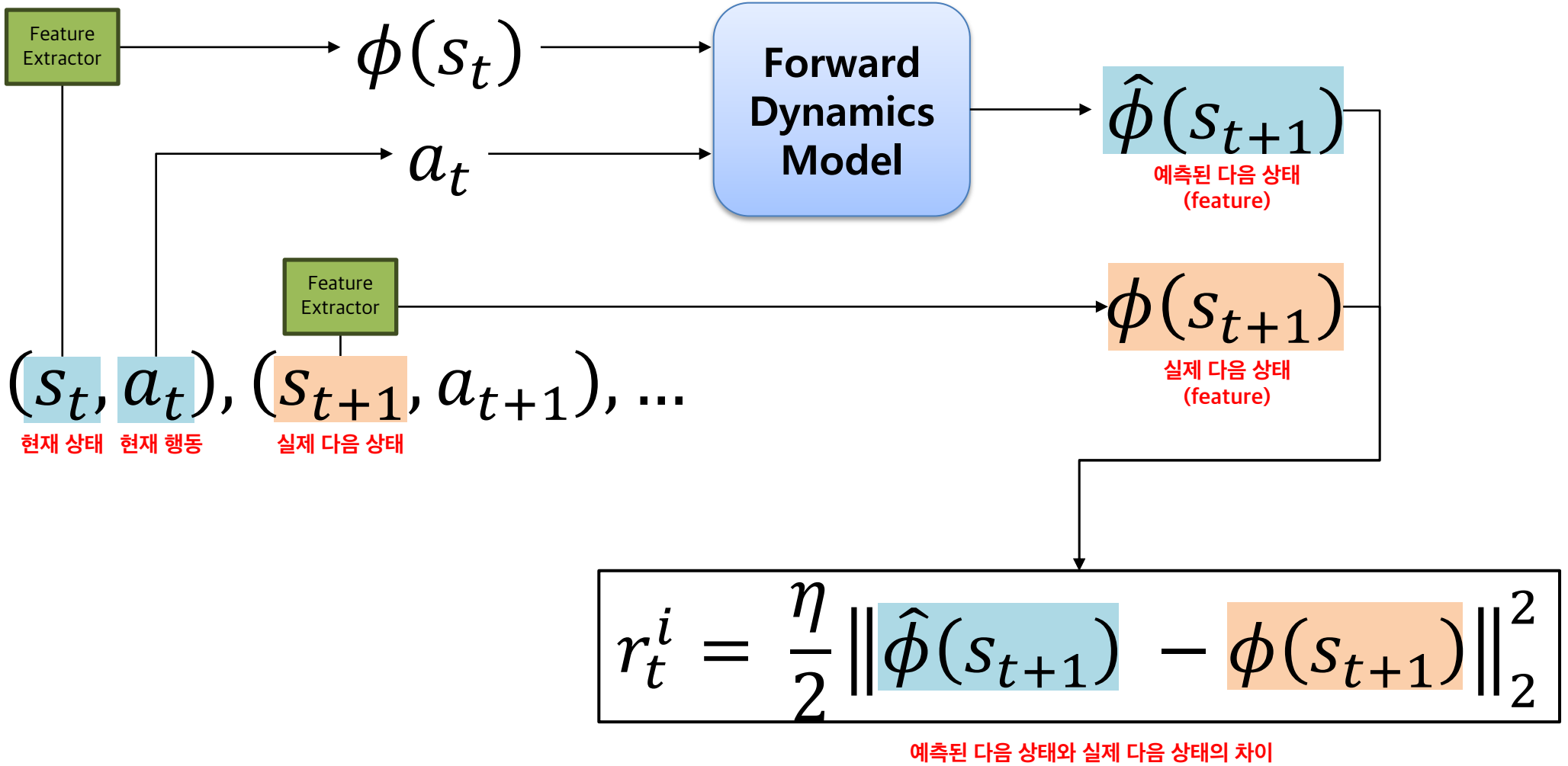
URL의 종류



Knowledge-based URL : ICM [ICML 2017]

URL의 종류

*두 Feature Extractor는 서로 같음

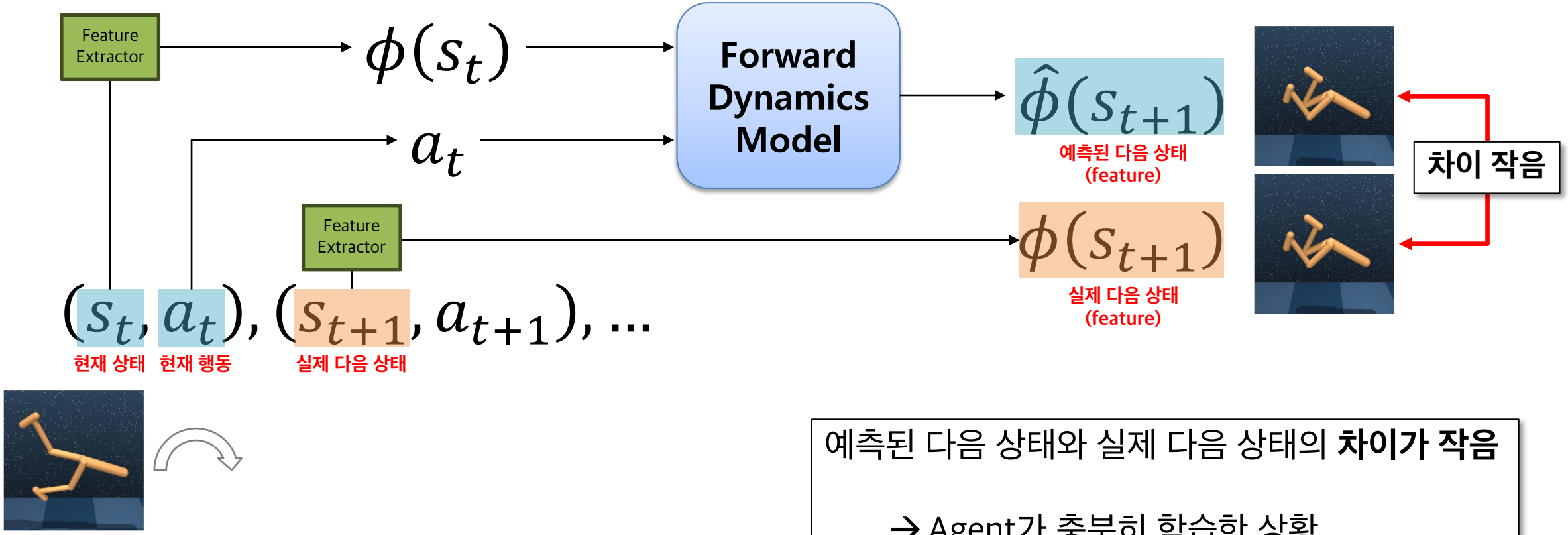


Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Knowledge-based URL : ICM [ICML 2017]

URL의 종류

*두 Feature Extractor는 서로 같음



예측된 다음 상태와 실제 다음 상태의 **차이가 작음**

→ Agent가 충분히 학습한 상황
→ **더이상 탐색하지 않아도 되는 상황**

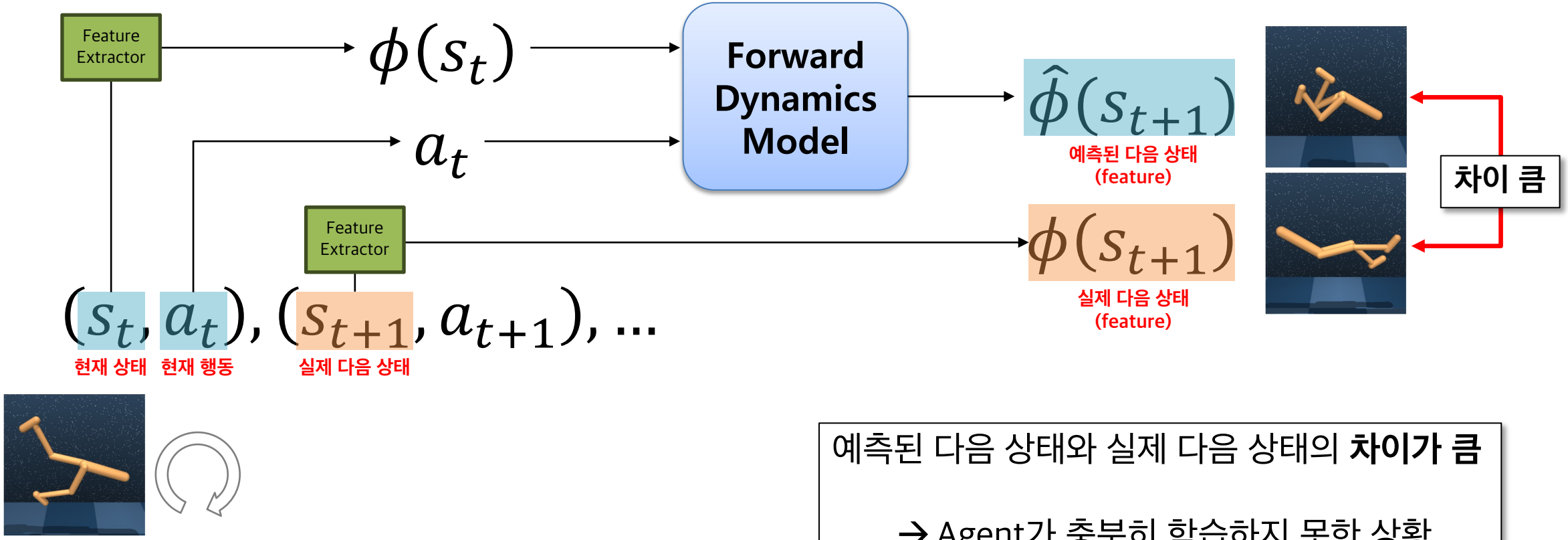
→ **낮은 내부 보상**Intrinsic Reward, r_t^i 부여

Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Knowledge-based URL : ICM [ICML 2017]

URL의 종류

*두 Feature Extractor는 서로 같음



예측된 다음 상태와 실제 다음 상태의 **차이가 큼**

→ Agent가 충분히 학습하지 못한 상황
→ **탐색의 대상이 되는 상황**

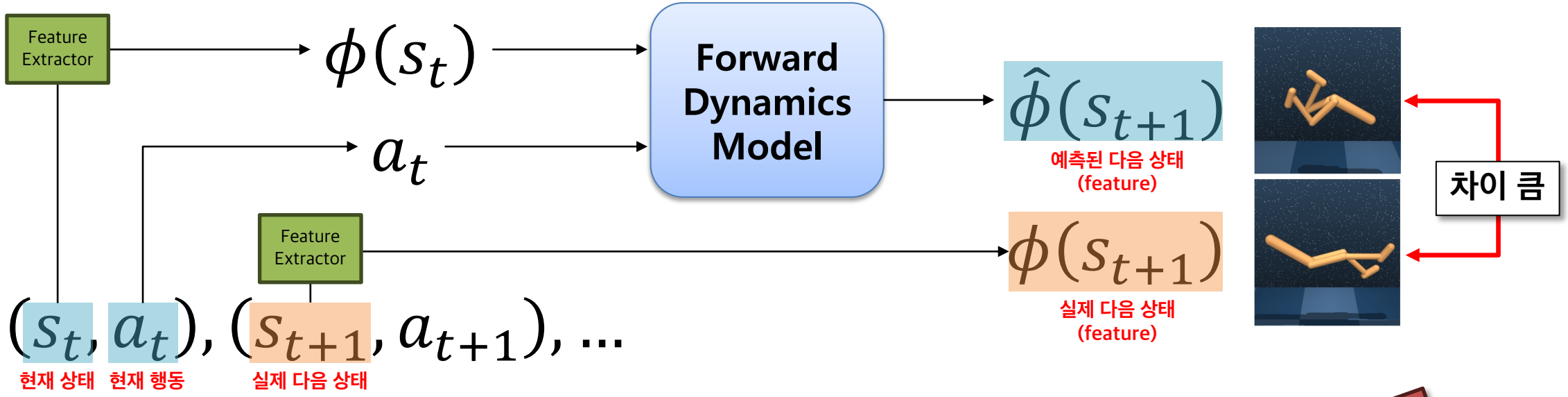
→ **높은 내부 보상**Intrinsic Reward, r_t^i 부여

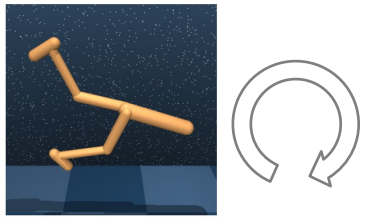
Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Knowledge-based URL : ICM [ICML 2017]

URL의 종류

*두 Feature Extractor는 서로 같음





예측된 다음 상태와 실제 다음 상태의 차이 자체를 내부 보상Intrinsic Reward, r_t^i으로 활용

예측된 다음 상태와 실제 다음 상태의 차이 큼

→ Ag… 상황

→ … 보상Intrinsic Reward, r_t^i 부여

Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Knowledge-based URL : ICM [ICML 2017]

URL의 종류

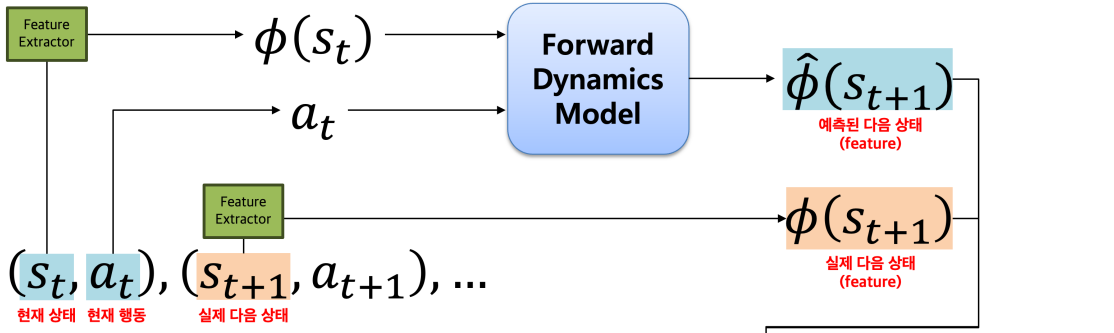
❖ r_t^i : 모델이 예측한 다음 state와 실제 다음 state간의 차이가 얼마나 큰가?

- Agent가 예측한 행동 결과와 실제 행동 결과의 차이가 클수록 높은 보상Reward이 부여
  - Agent의 학습이 충분한 상황, **더이상 탐색하지 않아도 되는 상황이라면** :

    Agent가 다음 state에 대해서 잘 예측할 수 있기 때문에 차이가 크지 않음
  - Agent의 학습이 충분하지 않은, **탐색의 대상이 되어야 하는 상황이라면** :

    Agent가 다음 state에 대해서 잘 예측할 수 없고, 따라서 모델이 예측하는 다음 상태와 실제 다음 상태의 차이가 큼

    → 차이가 큰 쪽을 학습하겠다는 것은 모델이 아직 잘 모르는 상황을 탐색Exploration하겠다는 것과 동일!



*두 Feature Extractor는 서로 같음

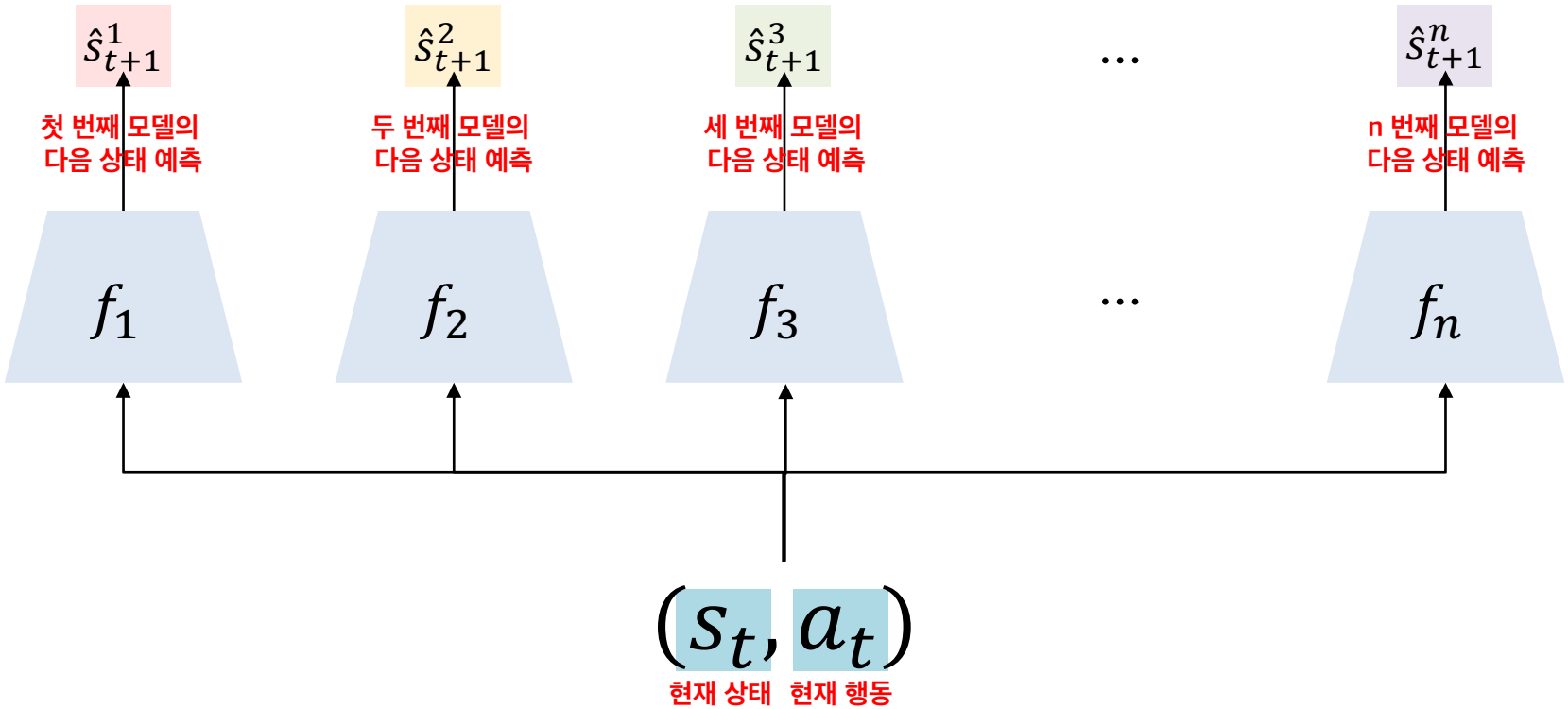
Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Knowledge-based URL : Disagreement [ICML 2019]

URL의 종류

$$r_t^i = Var(\hat{s}_{t+1}^1, \hat{s}_{t+1}^2, \hat{s}_{t+1}^3, \cdots, \hat{s}_{t+1}^n)$$

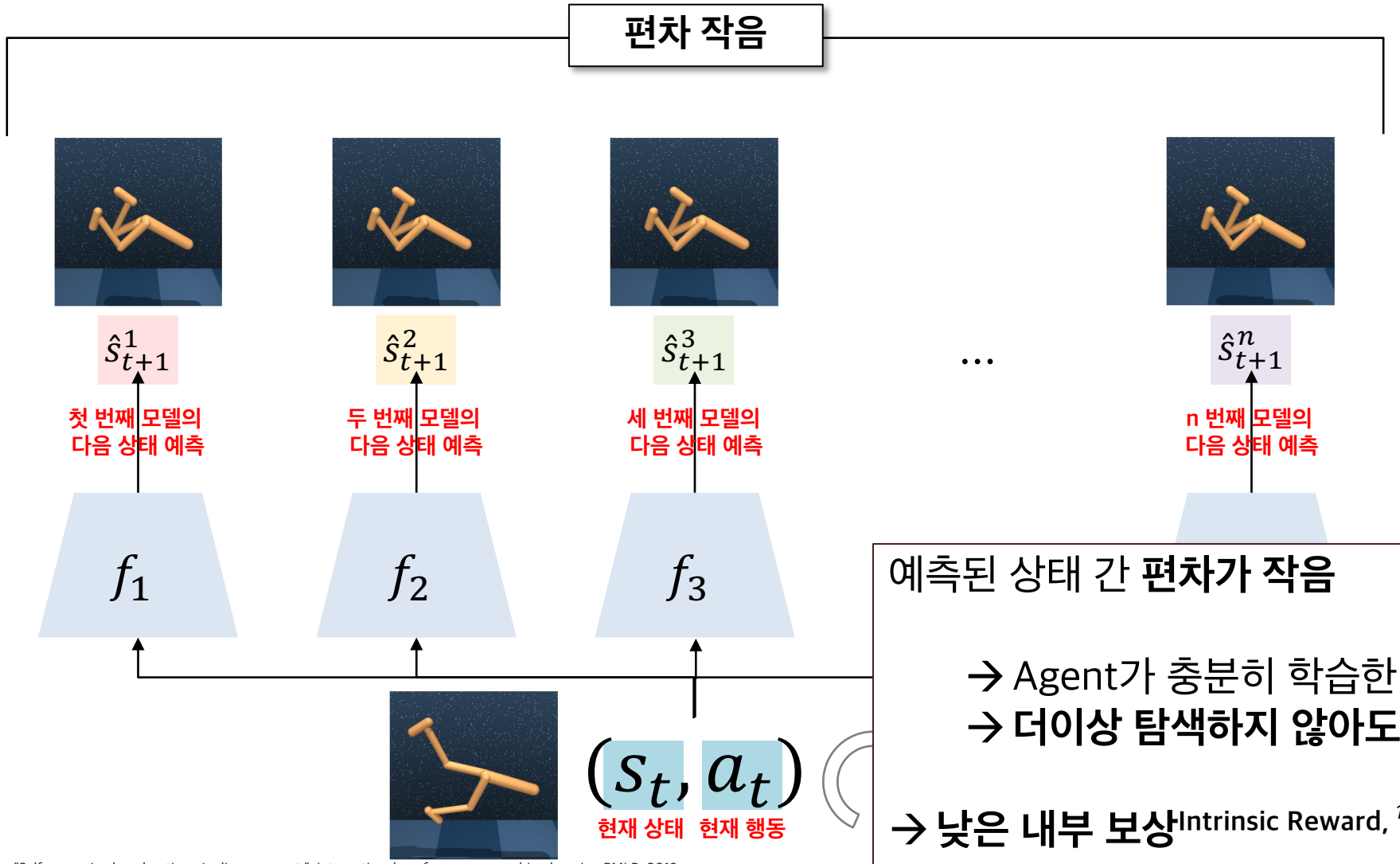
n 개의 모델의 다음 상태 예측에 대한 분산(편차)



Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

Knowledge-based URL : Disagreement [ICML 2019]

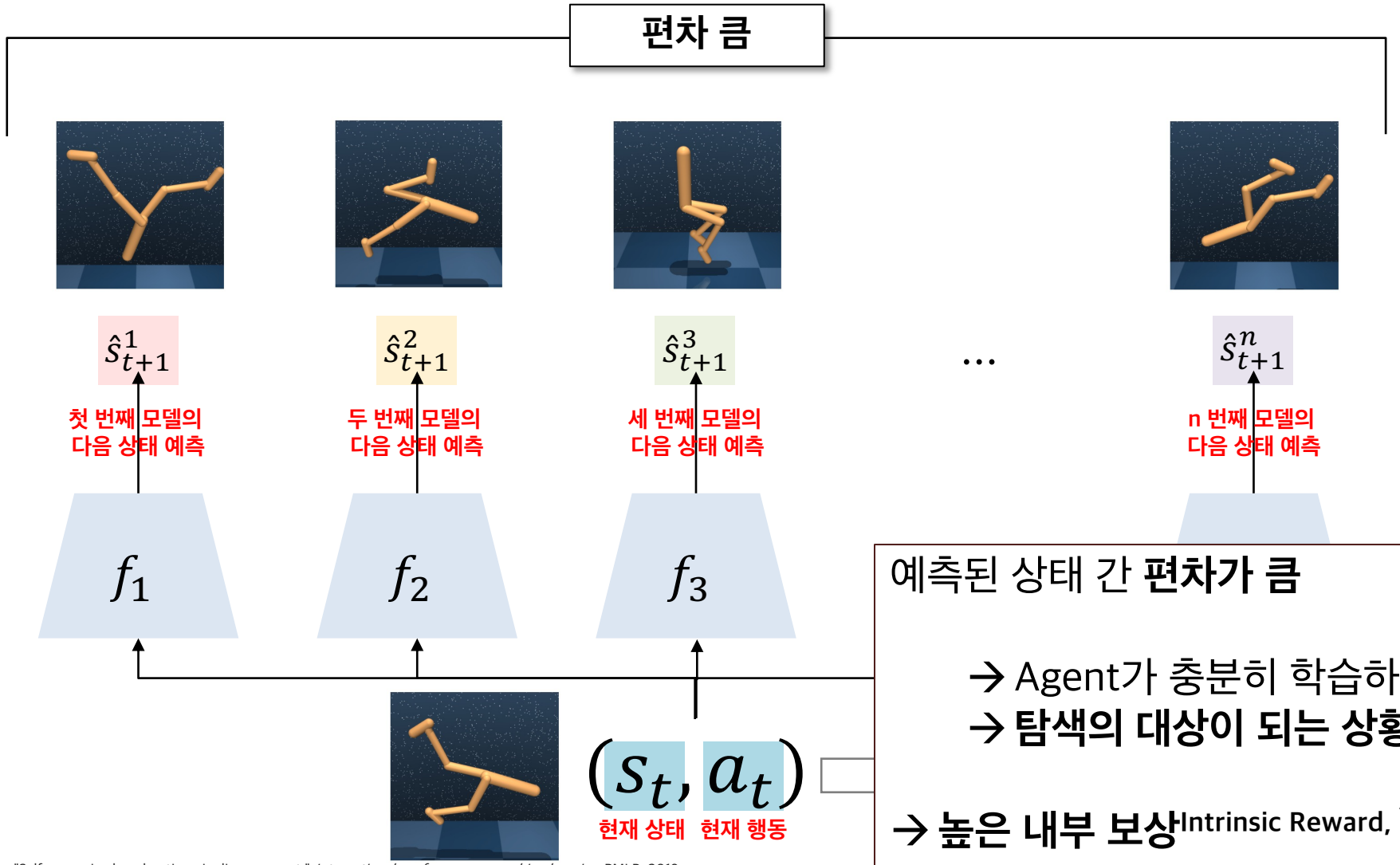
URL의 종류



Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

Knowledge-based URL : Disagreement [ICML 2019]

URL의 종류



예측된 상태 간 **편차가 큼**

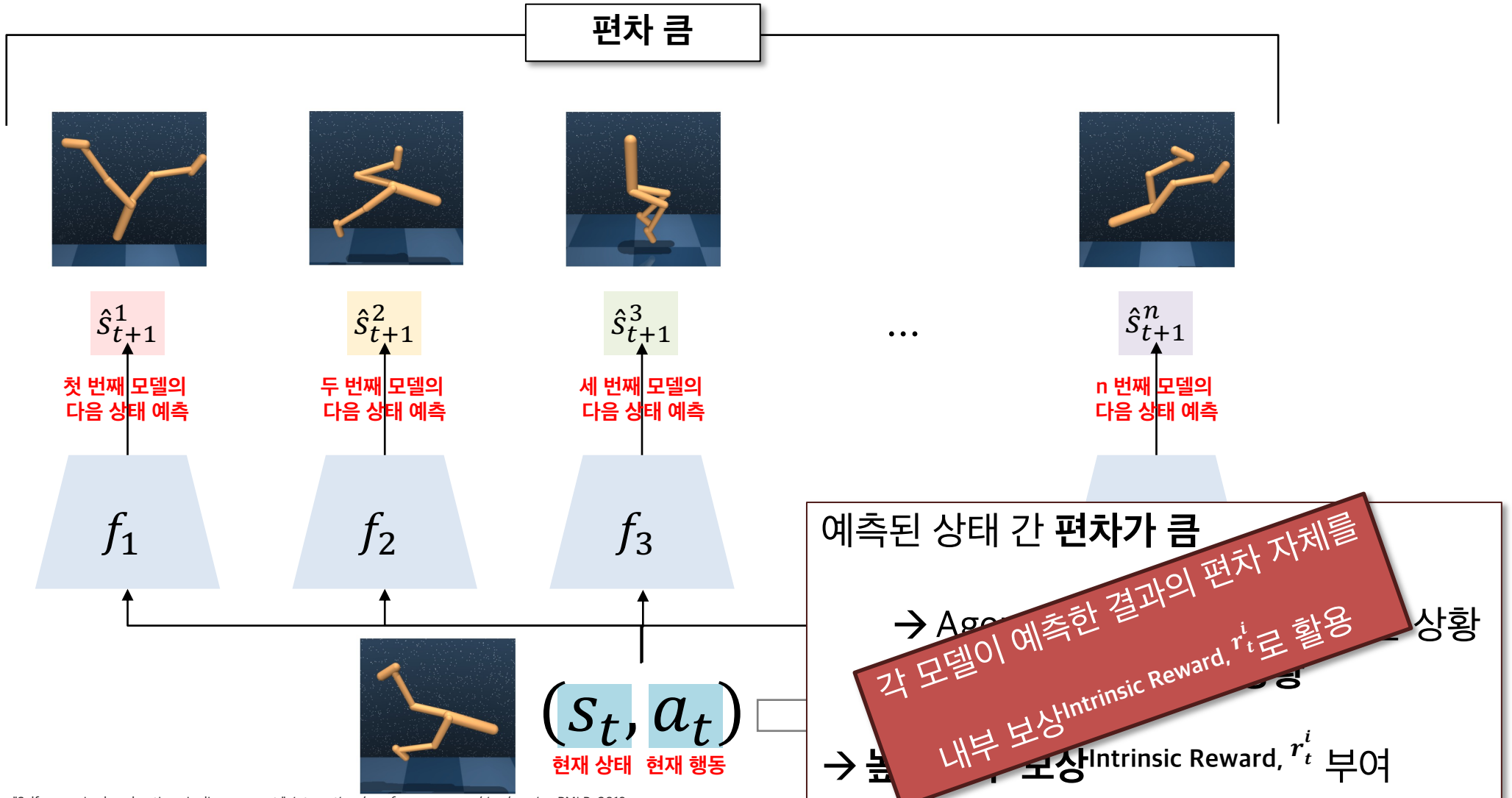
→ Agent가 충분히 학습하지 못한 상황
→ **탐색의 대상이 되는 상황**

→ **높은 내부 보상**Intrinsic Reward, r_t^i 부여

Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

Knowledge-based URL : Disagreement [ICML 2019]

URL의 종류



Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

Knowledge-based URL : Disagreement [ICML 2019]

URL의 종류

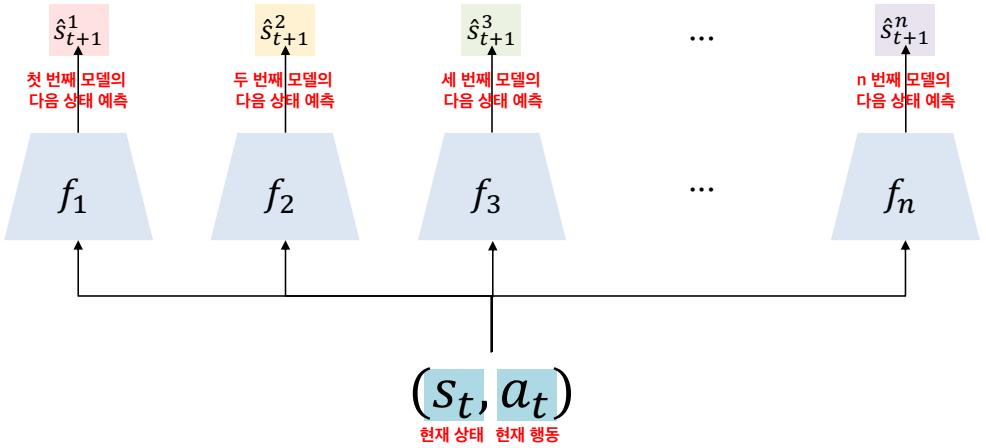
❖ r_t^i : 여러 모델이 예측한 다음 state가 서로 얼마나 불일치하는가?

- 여러 모델을 앙상블하여 사용하며, 모델들의 행동 결과 예측이 서로 불일치할수록 높은 보상Reward이 부여
  - Agent의 학습이 충분한, **더이상 탐색하지 않아도 되는 상황이라면** : 여러 모델이 같은 예측 결과를 보일 것
  - Agent의 학습이 충분하지 않은, **탐색의 대상이 되어야하는 상황이라면** : 여러 모델이 서로 다른 예측 결과를 보일 것

  → 예측 편차가 큰 쪽을 학습하겠다는 것은 Agent가 아직 잘 모르는 상태와 행동을 탐색하겠다는 것과 동일

$$r_t^i = Var(\hat{s}_{t+1}^1, \hat{s}_{t+1}^2, \hat{s}_{t+1}^3, \dots, \hat{s}_{t+1}^n)$$

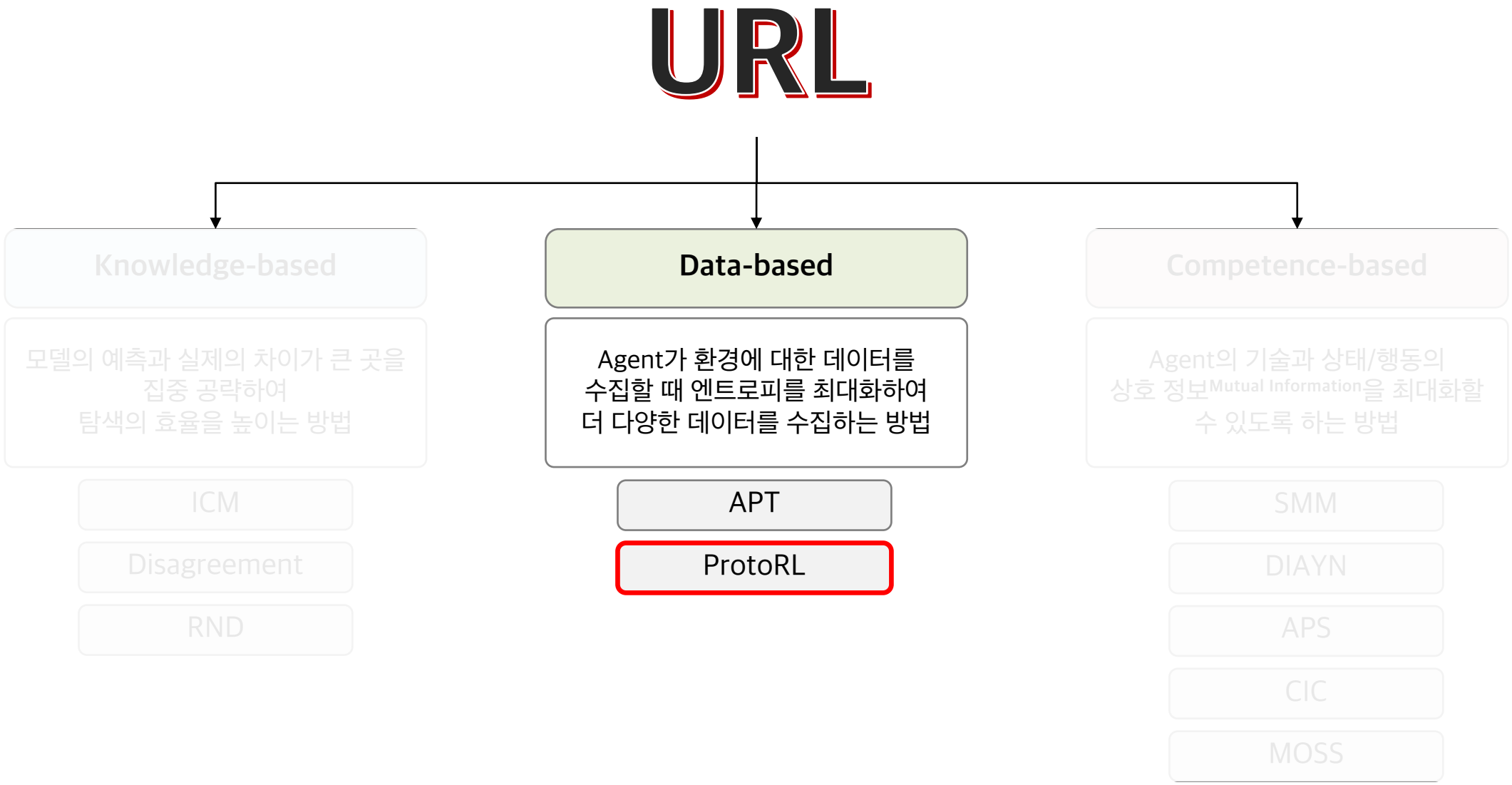
n 개의 모델의 다음 상태 예측에 대한 분산(편차)



Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

URL의 종류 한 눈에 알아보기

URL의 종류



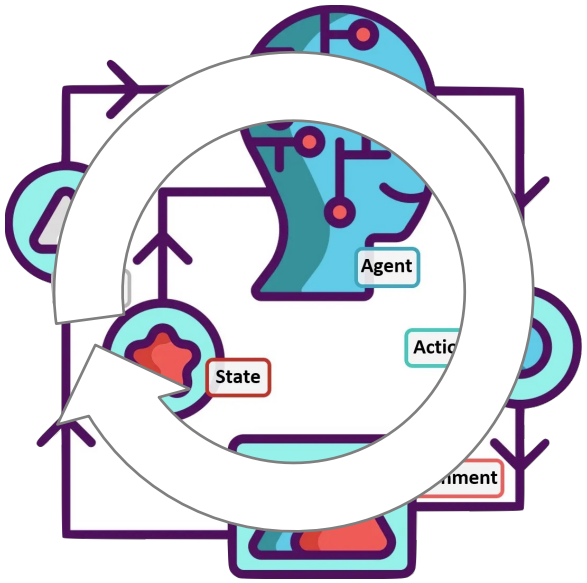
Data-based URL : ProtoRL [ICML 2021]

URL의 종류

❖ What is 'Data' in Data-based URL?

- Data-based URL에서 '**Data**'란,

  Agent가 환경Environment와 상호작용하며 수집하는 상태State와 행동Action, 그리고 보상Reward에 대한 정보를 통칭



Data

(s_t, a_t, s_{t+1}, r_t)

$(s_{t+1}, a_{t+1}, s_{t+2}, r_{t+1})$

$(s_{t+2}, a_{t+2}, s_{t+3}, r_{t+2})$

$(s_{t+3}, a_{t+3}, s_{t+4}, r_{t+3})$

...

Laskin, Michael, et al. "URLB: Unsupervised Reinforcement Learning Benchmark." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2021]

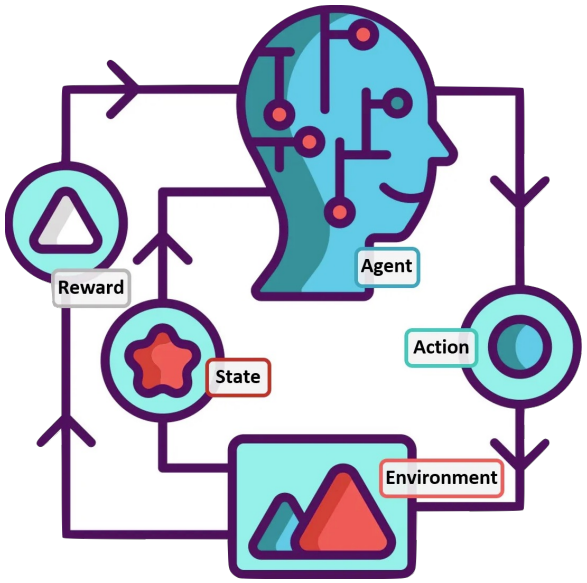
URL의 종류

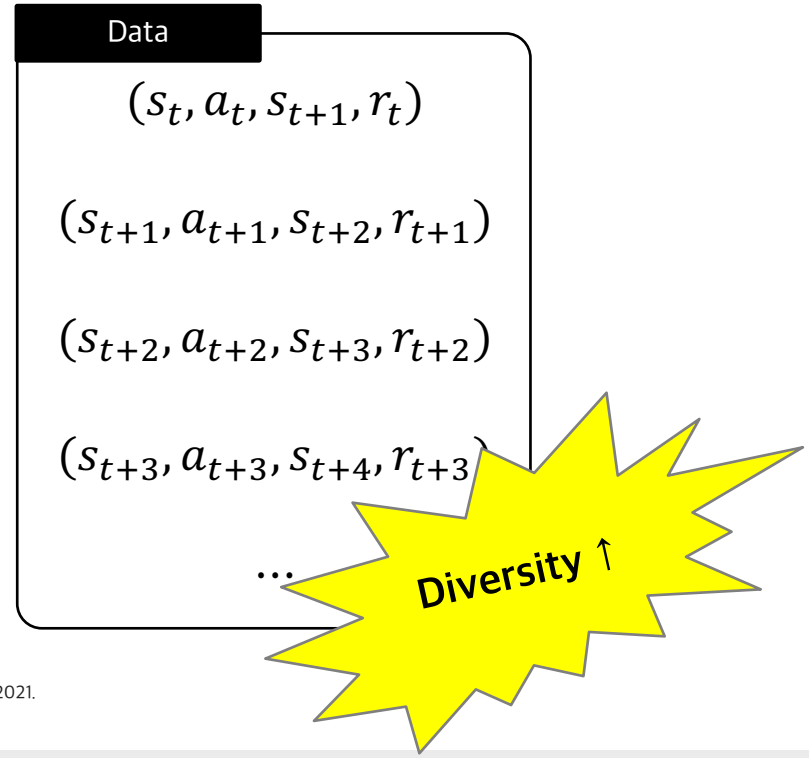
❖ What is 'Data' in Data-based URL?

- Data-based URL에서 '**Data**'란,

  Agent가 환경Environment와 상호작용하며 수집하는 상태State와 행동Action, 그리고 보상Reward에 대한 정보를 통칭

- Data-based URL은 Pre-train 단계에서 **Agent가 수집하는 Data를 최대한 다양하게** 하는 것을 목표로 함





Laskin, Michael, et al. "URLB: Unsupervised Reinforcement Learning Benchmark." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

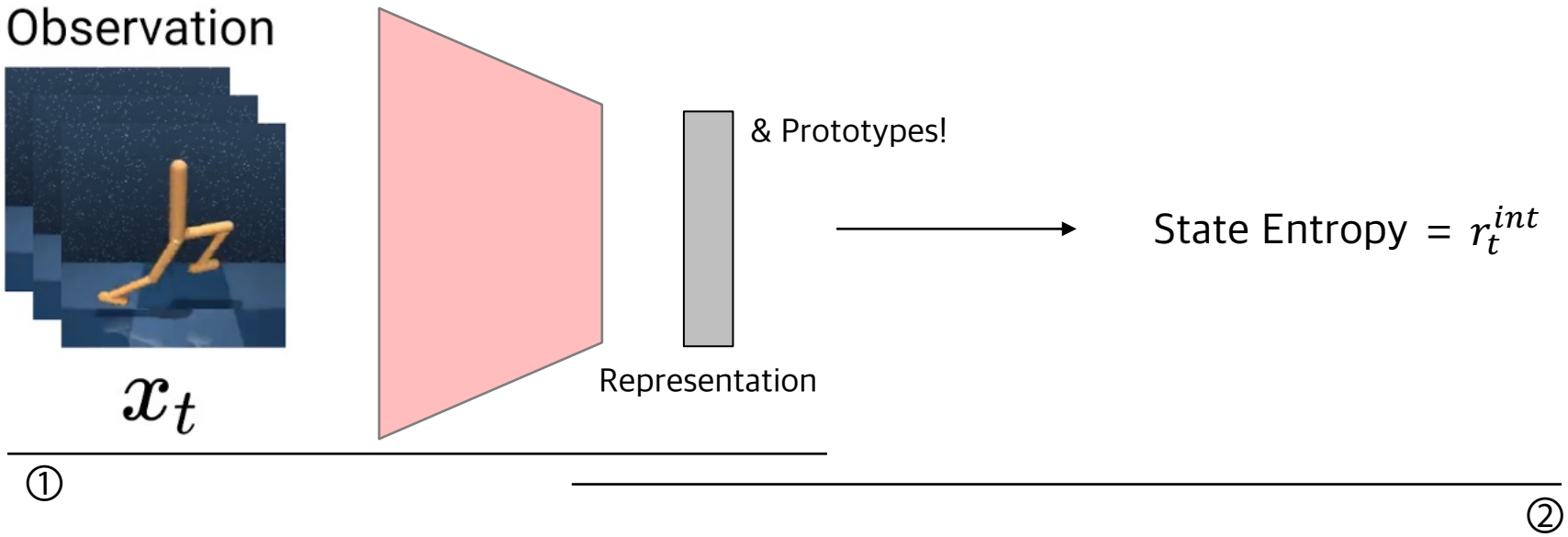
Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ProtoRL의 Pre-training은 두 단계로 구성

- **① Prototypical Representation Learning**
  - Image Observation을 잘 표현할 수 있는 Representation + Latent State Space의 Basis를 잘 표현할 수 있는 Prototype을 얻을 수 있도록 하자!
- **② Maximum Entropy Exploration**
  - 획득한 Representation + Prototype을 활용하여 Agent가 방문하는 State에 대한 Entropy를 최대화하자!



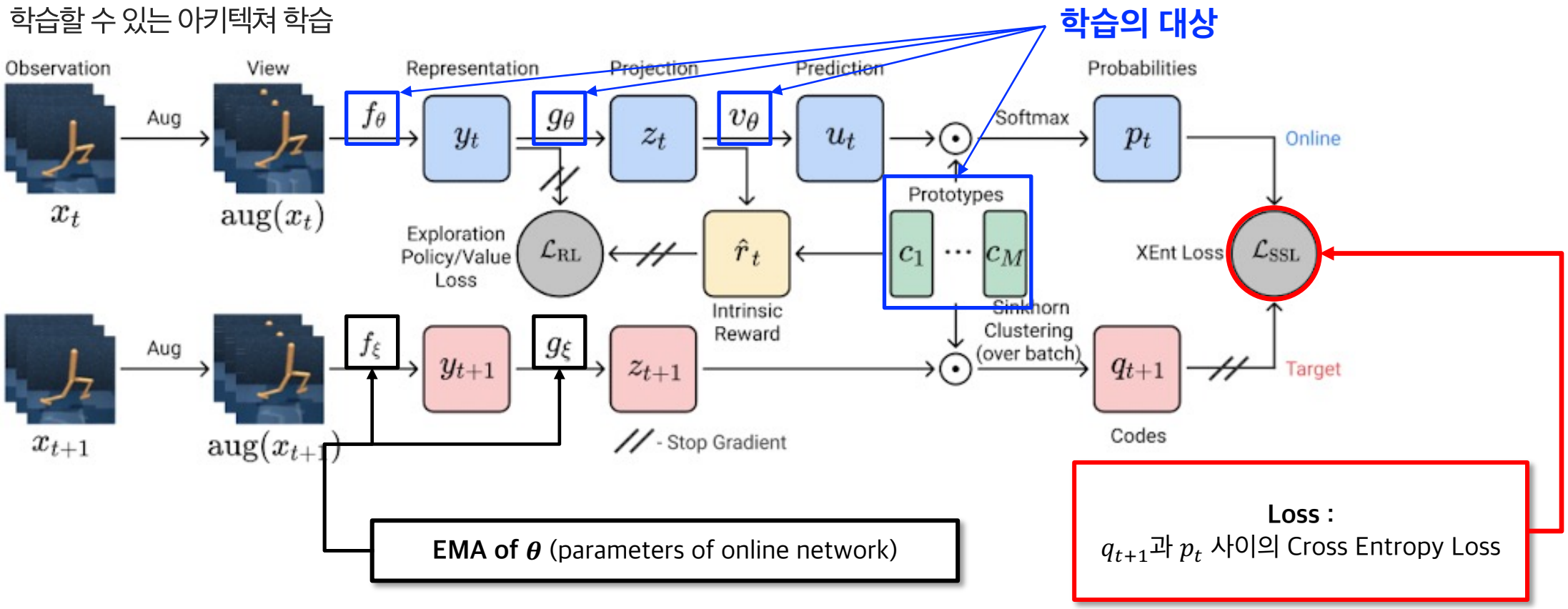
Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ① Prototypical Representation Learning

- SwAV와 유사한 방식
- Image Observation에 대해 Representation을 잘 추출하는 동시에, Prototype이 Latent State Space의 좋은 Basis가 될 수 있도록 잘 학습할 수 있는 아키텍쳐 학습



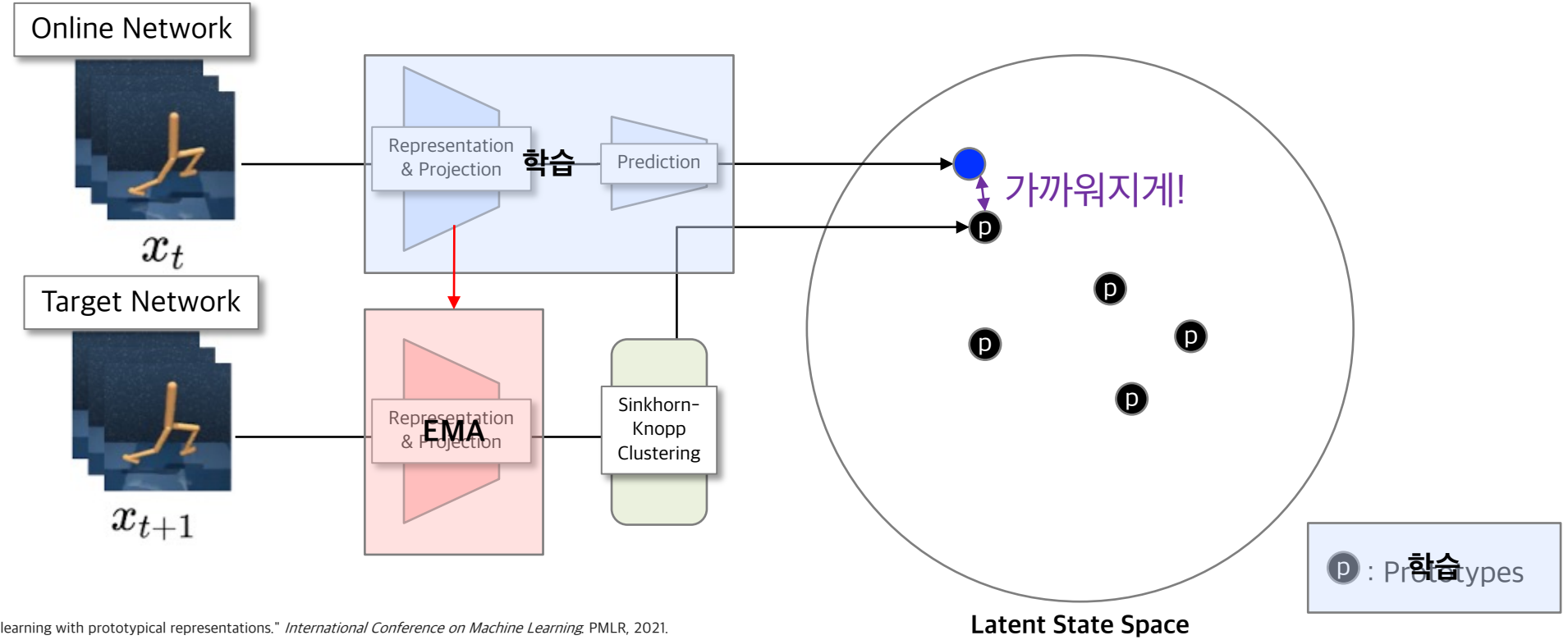
Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ① Prototypical Representation Learning

- SwAV와 유사한 방식
- Image Observation에 대해 Representation을 잘 추출하는 동시에, Prototype이 Latent State Space의 좋은 Basis가 될 수 있도록 잘 학습할 수 있는 아키텍쳐 학습



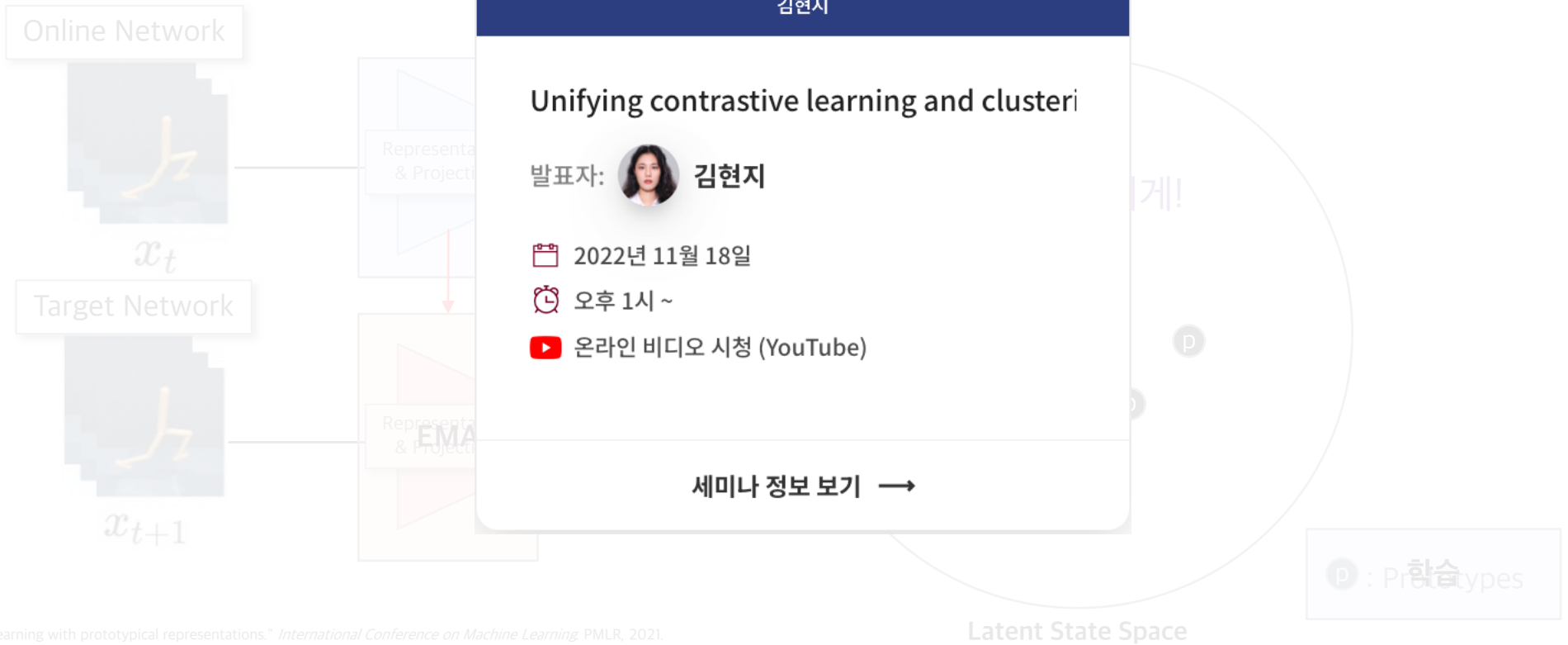
Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ① Prototypical Representation L

- SwAV와 유사한 방식
- Image Observation에 대해 Repres State Space의 좋은 Basis가 될 수 있도록 잘 학습할 수 있는 아키텍쳐 학습



종료

Unifying contrastive learning and clustering

2022.11.18

Data Mining & Quality Analytics Lab.

김현지

Unifying contrastive learning and clusteri

발표자: 김현지

2022년 11월 18일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

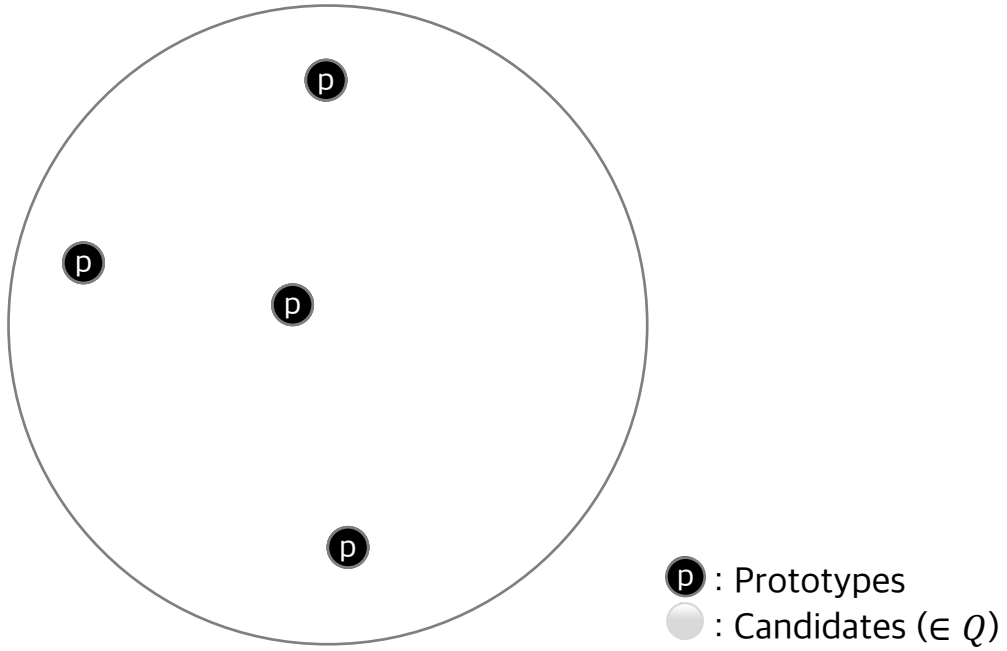
Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ② Maximum Entropy Exploration

- ①에서 학습한 Prototype을 활용해, Candidate Que Q를 Sample
- Q에서 ①에서 학습한 Encoder를 활용해 구한 State Embedding z_i의 k-nearest neighbor를 찾고, z_i와 k-nearest neighbor 사이의 거리를 구해 그 거리의 합을 Intrinsic Reward라고 함



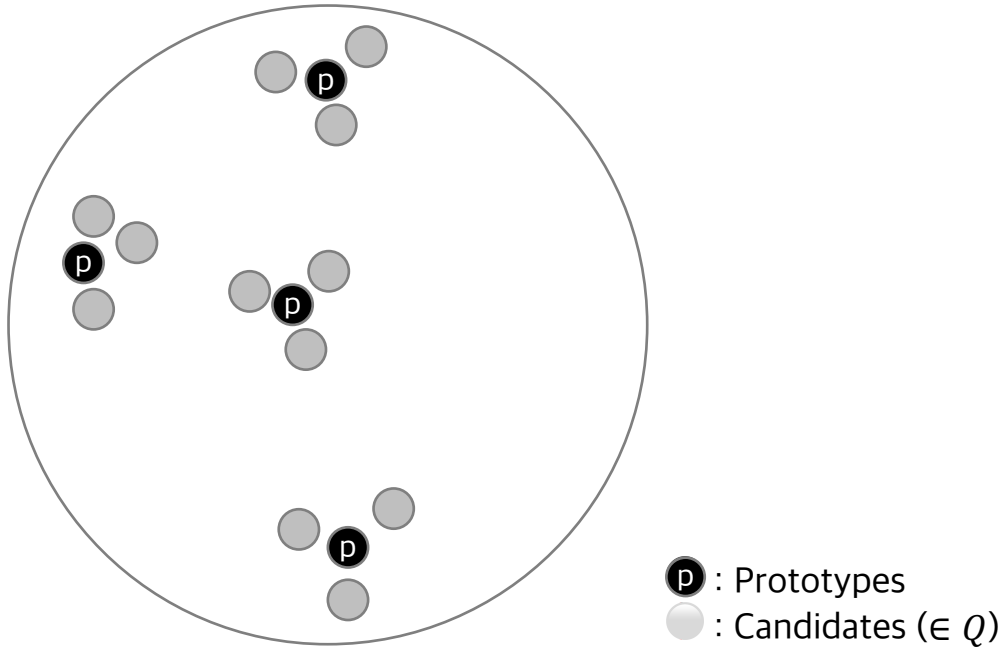
Latent State Space

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

❖ ② Maximum Entropy Exploration

- ①에서 학습한 Prototype을 활용해, Candidate Que Q를 Sample
- Q에서 ①에서 학습한 Encoder를 활용해 구한 State Embedding z_i의 k-nearest neighbor를 찾고, z_i와 k-nearest neighbor 사이의 거리를 구해 그 거리의 합을 Intrinsic Reward라고 함



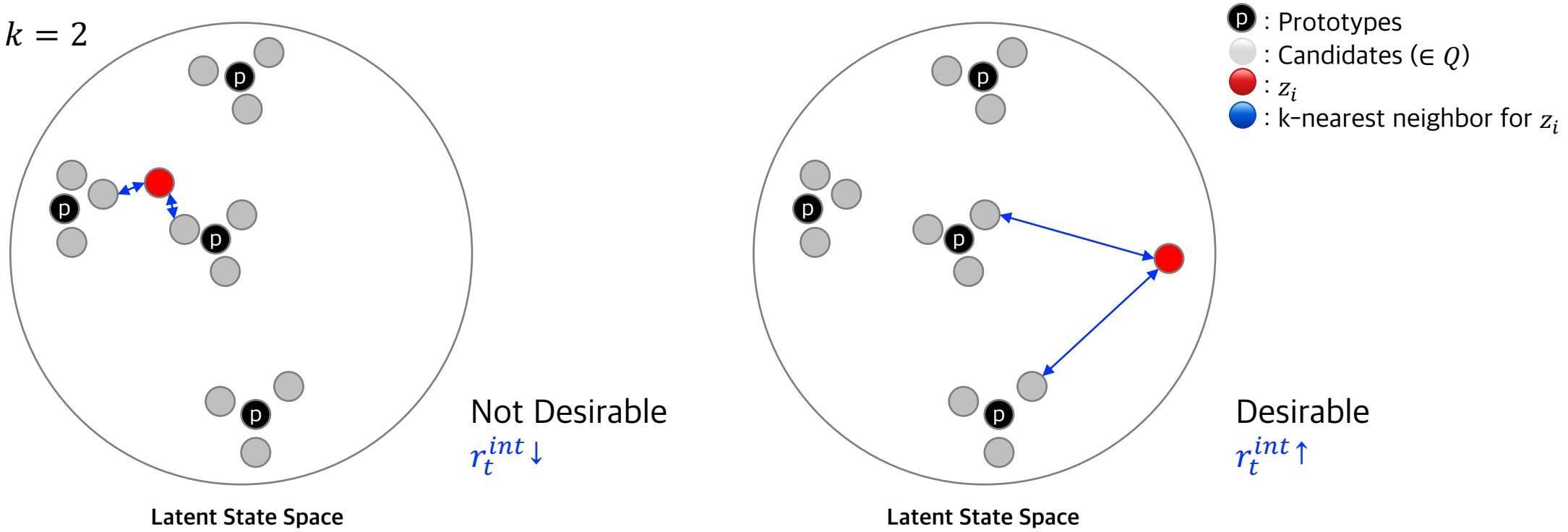
Latent State Space

Data-based URL : ProtoRL [ICML 2022]

URL의 종류

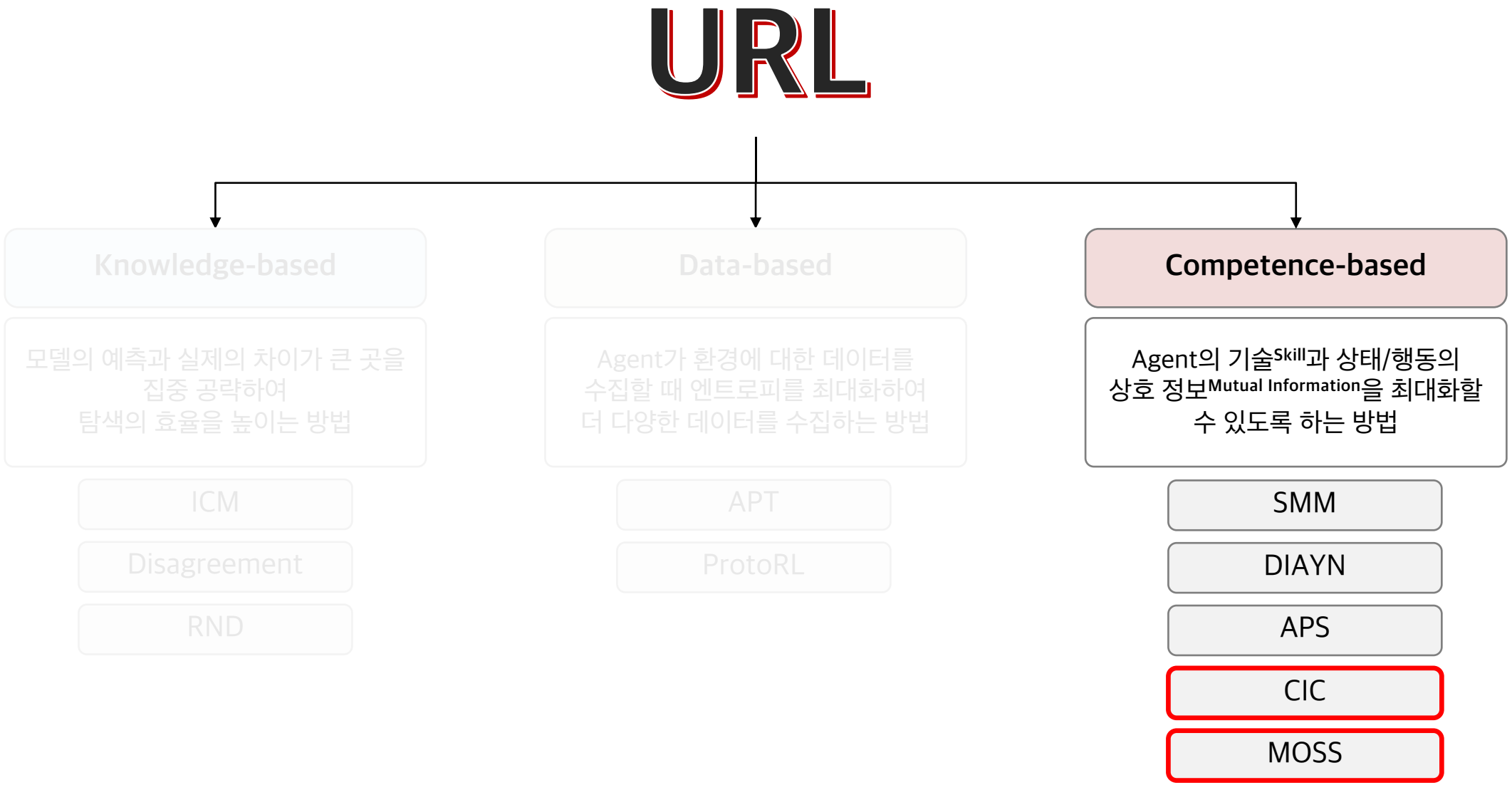
❖ ② Maximum Entropy Exploration

- ①에서 학습한 Prototype을 활용해, Candidate Que Q를 Sample
- Q에서 ①에서 학습한 Encoder를 활용해 구한 State Embedding z_i의 k-nearest neighbor를 찾고, z_i와 k-nearest neighbor 사이의 거리를 구해 그 거리의 합을 Intrinsic Reward라고 함



URL의 종류 한 눈에 알아보기

URL의 종류

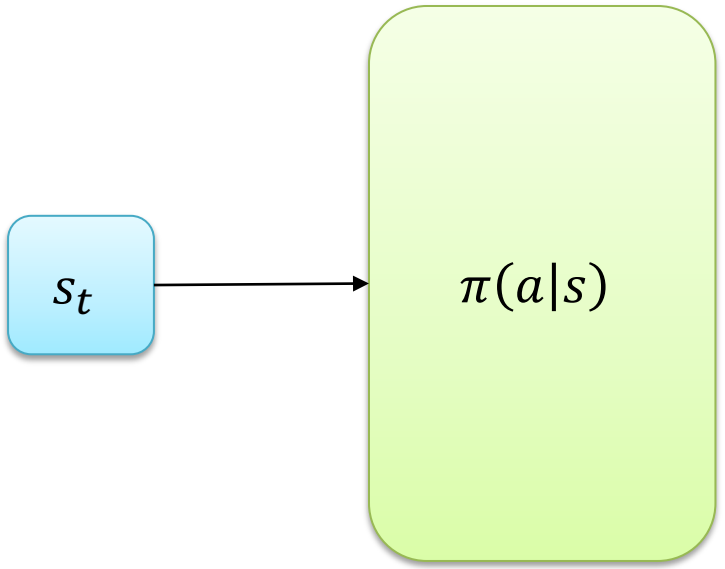


Preliminaries for Competence-based URL

URL의 종류

❖ 기존 강화학습의 Action 학습 및 선택 전략

- 기존 강화학습에서, 현재의 상태(State, s_t)가 주어졌을 때 현재 상태에서 어떤 행동(Action, a_t) 을 선택할지에 대한 확률분포인 정책(Policy, $\pi(a|s)$)이 있으며, 누적 보상의 합이 최대가 되는 방향으로 Action이 선택될 수 있도록 정책 $\pi(a|s)$를 학습하는 것이 강화학습의 목표

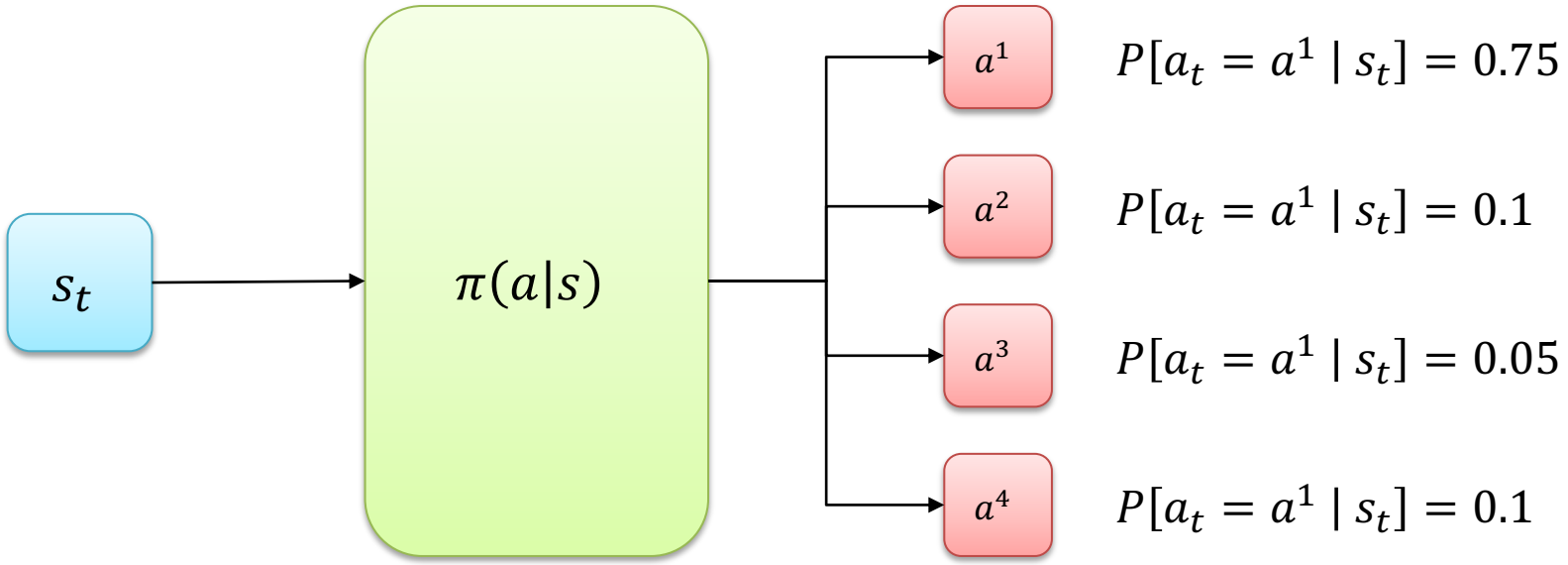


Preliminaries for Competence-based URL

URL의 종류

❖ 기존 강화학습의 Action 학습 및 선택 전략

- 기존 강화학습에서, 현재의 상태(State, s_t)가 주어졌을 때 현재 상태에서 어떤 행동(Action, a_t) 을 선택할지에 대한 확률분포인 정책(Policy, $\pi(a|s)$)이 있으며, 누적 보상의 합이 최대가 되는 방향으로 Action이 선택될 수 있도록 정책 $\pi(a|s)$를 학습하는 것이 강화학습의 목표

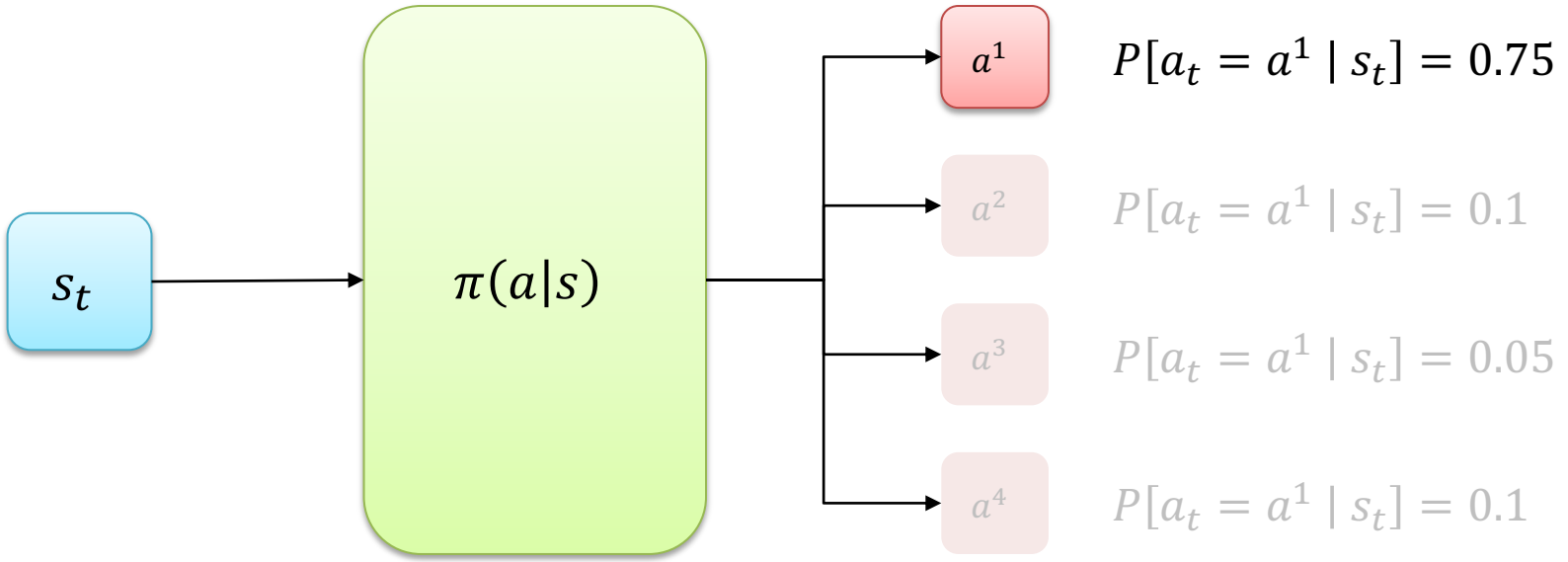


Preliminaries for Competence-based URL

URL의 종류

❖ 기존 강화학습의 Action 학습 및 선택 전략

- 기존 강화학습에서, 현재의 상태(State, s_t)가 주어졌을 때 현재 상태에서 어떤 행동(Action, a_t) 을 선택할지에 대한 확률분포인 정책(Policy, $\pi(a|s)$)이 있으며, 누적 보상의 합이 최대가 되는 방향으로 Action이 선택될 수 있도록 정책 $\pi(a|s)$를 학습하는 것이 강화학습의 목표



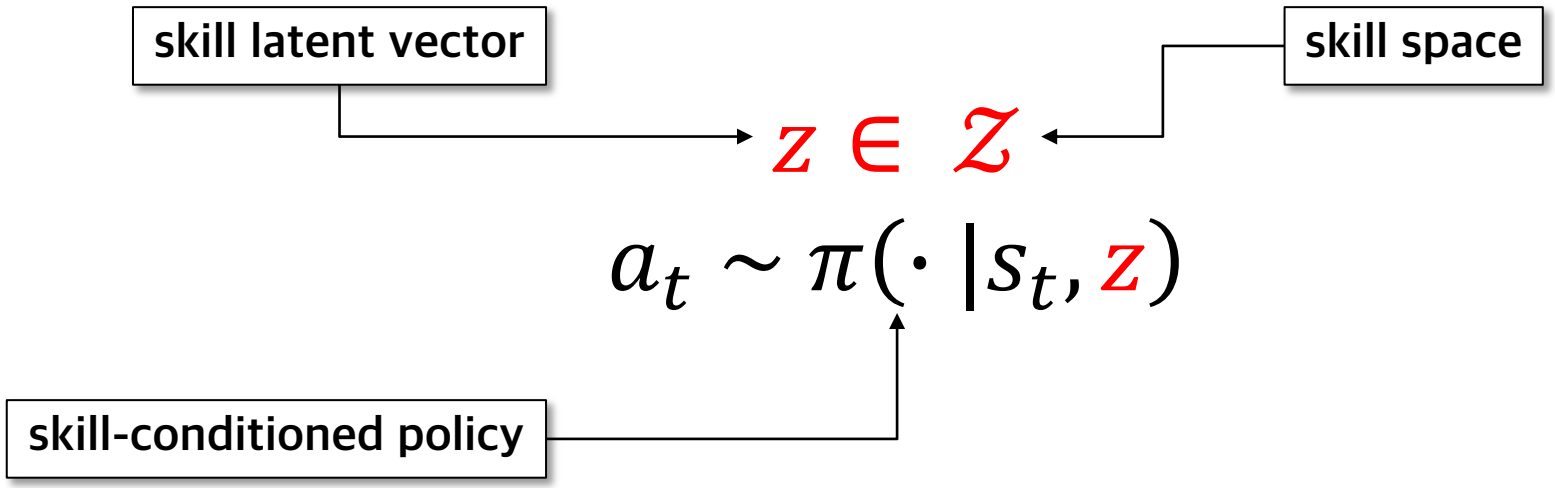
Preliminaries for Competence-based URL

URL의 종류

$$a_t \sim \pi(\cdot \mid s_t)$$

Policy of Competence-based URL

URL의 종류



Preliminaries for Competence-based URL

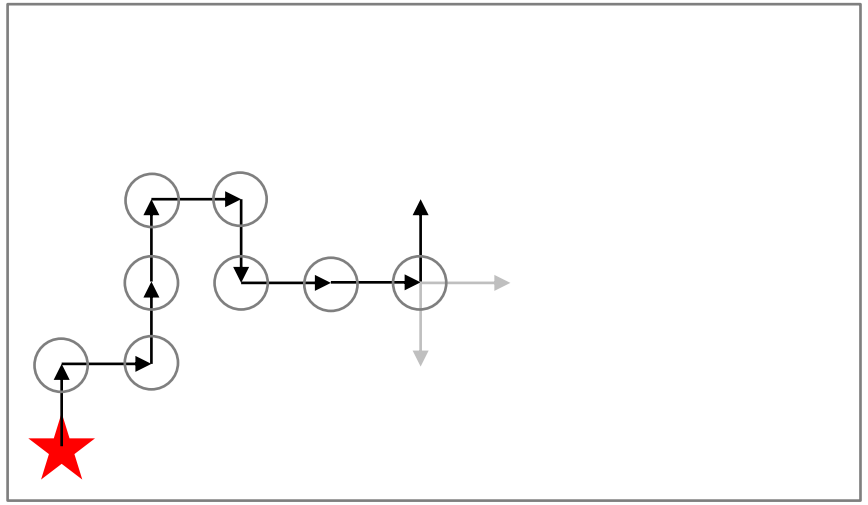
URL의 종류

❖ Competence-based URL의 핵심, Skill

- 정책 $\pi(\cdot)$에, Latent Skill Vector z를 활용한 추가적인 조건을 부여하여 Agent가 일관되게 행동할 수 있도록 한 것

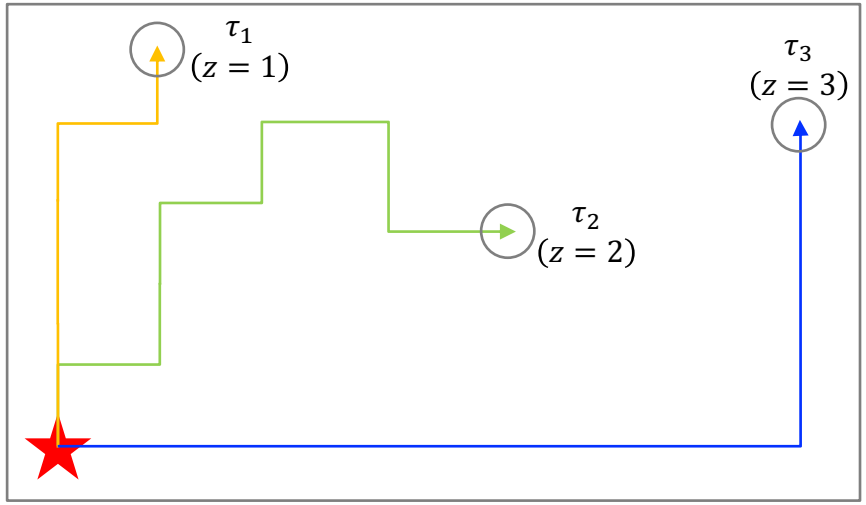
❖ Competence-based URL

- Pre-training 과정에서 Agent가 최대한 다양한 Skill을 습득하여, 이를 Downstream Task의 해결에 활용하도록 하려는 URL 방법론



without skills

$a_t \sim \pi(\cdot\,|s_t)$



with skills

$a_t \sim \pi(\cdot\,|s_t, z)$

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

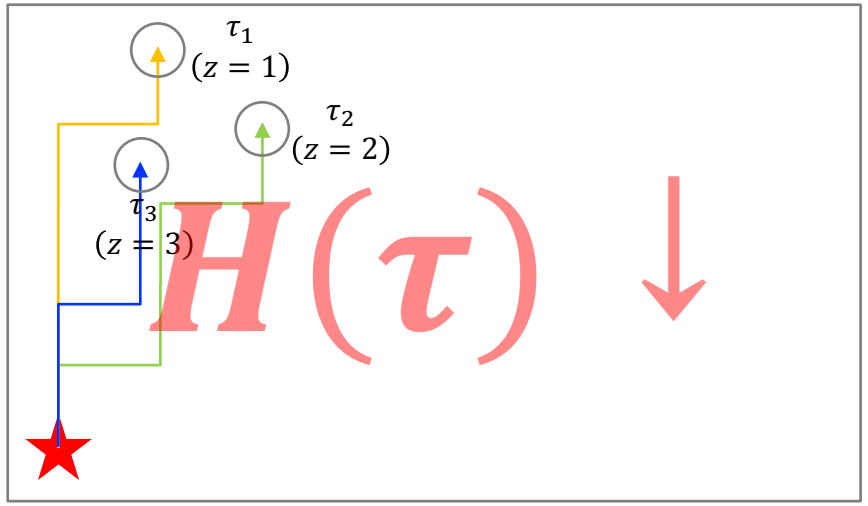
Preliminaries for Competence-based URL

URL의 종류

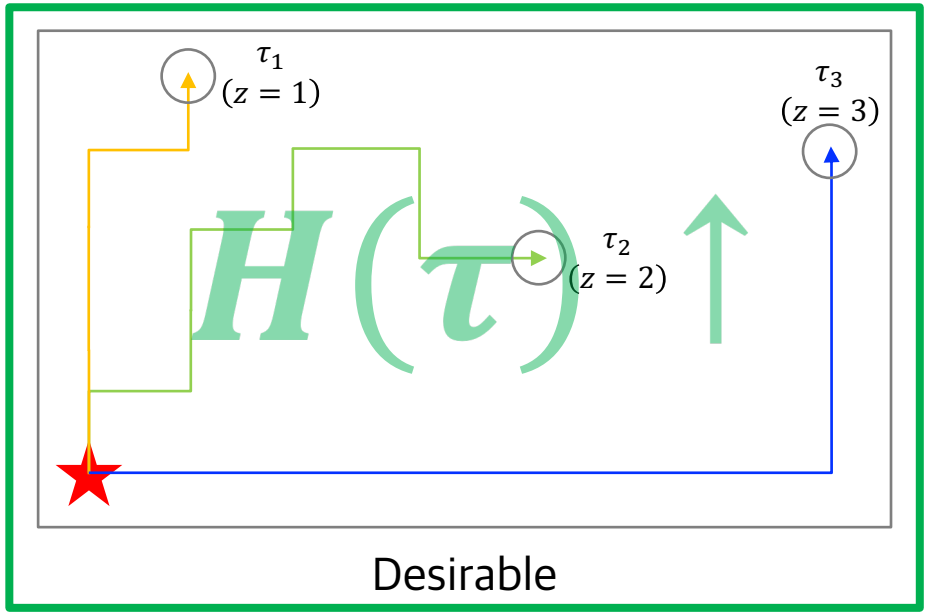
❖ 성공적인 Competence-based URL의 학습을 위한 조건

- Skill은 최대한 다양하게 학습되어야 함
- Skill은 Agent가 방문하는 State(들)를 통해 구분 가능해야함

→ 즉, **최대한 다양한 State를 방문할 수 있는 Skill들**이 학습되어야 함



Not Desirable



Desirable

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Preliminaries for Competence-based URL

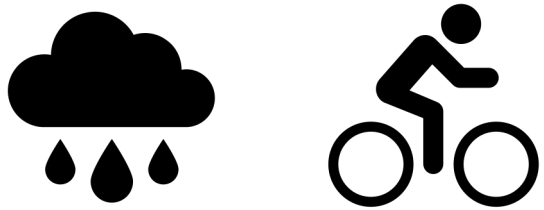
URL의 종류

❖ Mutual Information (상호 정보)

- **두 확률 변수 간의 상호 의존도**를 측정할 수 있는 통계량으로, Entropy를 활용하여 두 확률 변수 X, Y 간의 Mutual Information을 정의하면 다음과 같음

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- 두 확률 변수 간 Mutual Information이 **높다** = 한 변수의 값이 주어졌을 때 나머지 변수의 값을 쉽게 예측할 수 있다
- 두 확률 변수 간 Mutual Information이 **낮다** = 한 변수의 값이 주어진다고 하더라고 나머지 변수의 값을 쉽게 예측할 수 없다



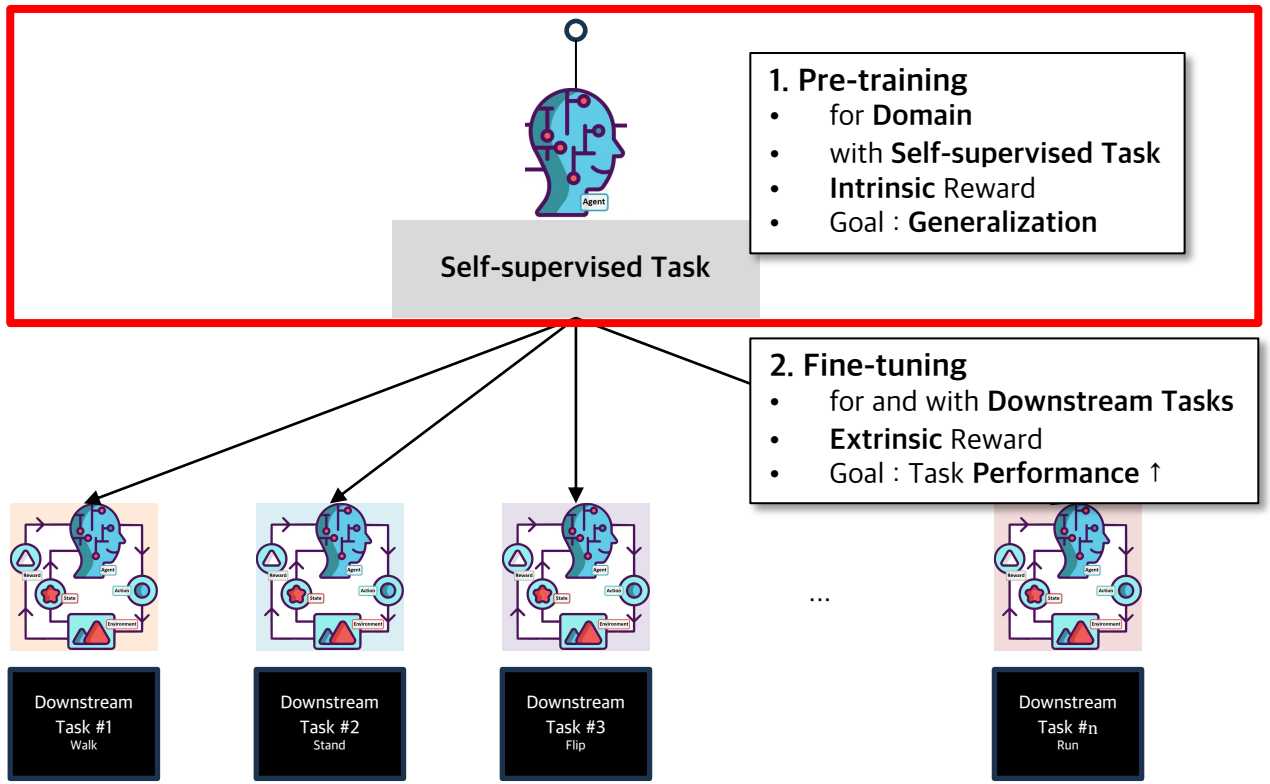
X Y

강수 여부(X)와 자전거를 탈지 말지 결정하는 일(Y)은 상호 정보가 높다

https://en.wikipedia.org/wiki/Mutual_information
https://jrc-park.tistory.com/311

Pre-train : Unsupervised Skill Discovery

❖ Pre-train의 목적은, Downstream Task에서 활용할 Skill들을 학습하는 것



1. **When** :
   Pre-training

2. **What** :
   Downstream Task에서 활용할 다양한 skill 습득

3. **How** :
   Agent가 방문하는 상태와 습득하는 skill 사이의 Mutual Information 최대화

Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

❖ Contrastive Intrinsic Control (CIC) 방법론

- URL의 Pre-train 단계에서 사용
- 방법 : Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)와 Skill(z)의 Mutual Information을 최대화
- **목표** :

1. **Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)와 Skill(z)을 잘 매칭시키자**

   ∵ Skill은 Agent가 방문하는 상태(들)를 통해서 구분할 수 있어야 한다

2. **Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)가 최대한 다양해지도록 하자**

   ∵ Skill은 최대한 다양하게 학습되어야 한다

→ **Agent가 최대한 다양한 Skill(z)을 습득할 수 있도록 하자**

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

Maximize

Lower Bound → Maximize

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

❖ Contrastive Intrinsic Control (CIC) 방법론 제안

- URL의 Pre-train 단계에서 사용
- 방법 : Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)와 Skill(z)의 Mutual Information을 최대화
- **목표** :

1. **Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)와 Skill(z)을 잘 매칭시키자 [①]**

   ∵ Skill은 Agent가 방문하는 상태(들)를 통해서 구분할 수 있어야 한다

2. **Agent가 방문하는 상태의 변화($\tau = (s_t, s_{t+1})$)가 최대한 다양해지도록 하자 [②]**

   ∵ Skill은 최대한 다양하게 학습되어야 한다

→ **Agent가 최대한 다양한 Skill(z)을 습득할 수 있도록 하자 [① + ②]**

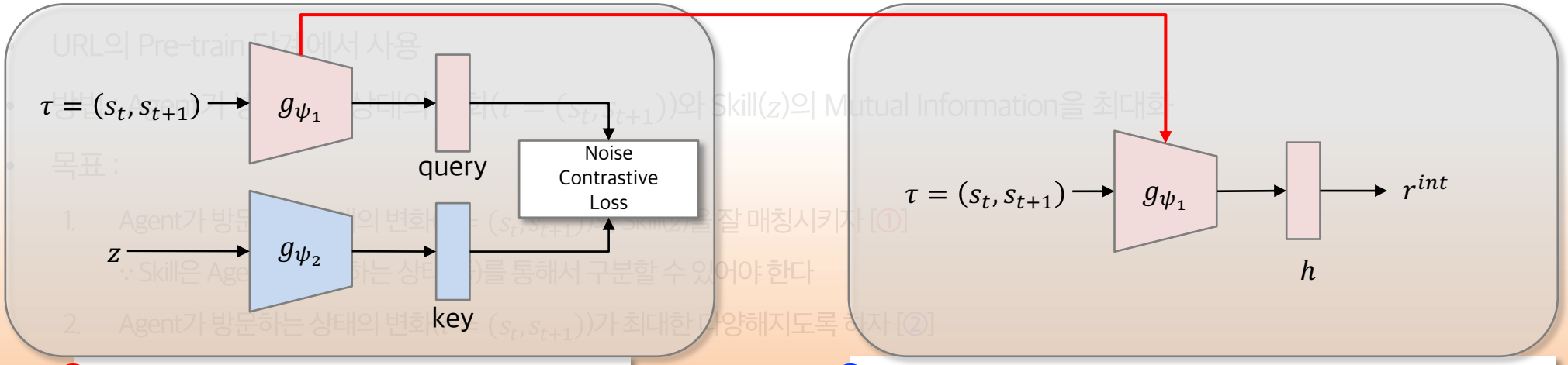
$$I(\tau; z) = H(\tau) - H(\tau|z) \geq \overset{②}{H(\tau)} + \overset{①}{\mathbb{E}_{\tau,z}[\log q(\tau|z)]}$$

Maximize (for $I(\tau; z)$) · Maximize (for $H(\tau)$) · Maximize (for $\mathbb{E}_{\tau,z}[\log q(\tau|z)]$)

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

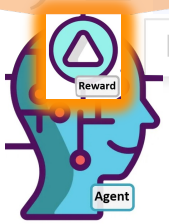


① 방문하는 상태의 변화 τ와 Skill Vector z가 서로 잘 매칭될 수 있도록 Encoder를 학습하고

② 1단계에서 학습한 Encoder를 활용해 Embedding을 구해서 Intrinsic Reward 계산에 활용하자

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

Maximize Maximize Maximize



CIC Agent

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{①}\mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

Maximize (좌) Maximize (우)

❖ ① Discriminator Parametrized with Contrastive Density Estimator

- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

$$\textbf{Maximize} \quad F_{CIC}(\tau) = \log \frac{\exp\left(\frac{g_{\psi_1}(\tau_i)^{\mathsf{T}} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^{\mathsf{T}} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$$

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

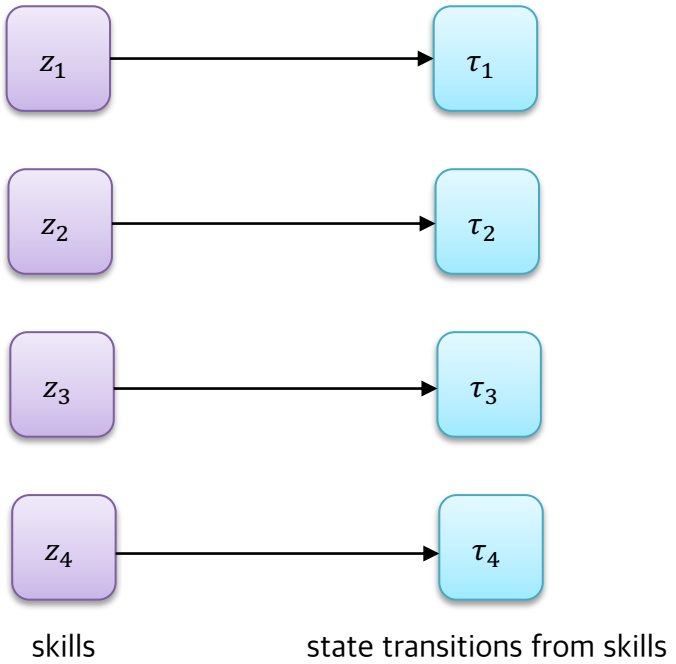
$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{①}\mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

Maximize (좌변) / Maximize (우변)

❖ ① Discriminator Parametrized with Contrastive Density Estimator

- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

Maximize $$F_{CIC}(\tau) = \log \frac{\exp\left(\frac{g_{\psi_1}(\tau_i)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$$



Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{①}\mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

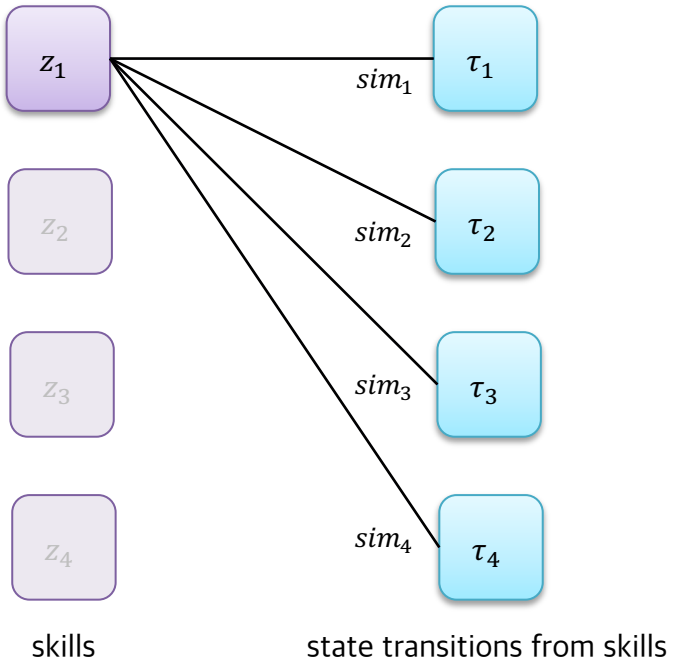
Maximize (좌변) / Maximize (우변)

❖ ① Discriminator Parametrized with Contrastive Density Estimator

- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

Maximize $$F_{CIC}(\tau) = \log \frac{\exp\left(\frac{g_{\psi_1}(\tau_i)^{\top} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^{\top} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$$

특정 skill vector에 대해서,
모든 state transition들과의 Similarity 계산



Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{①}\mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

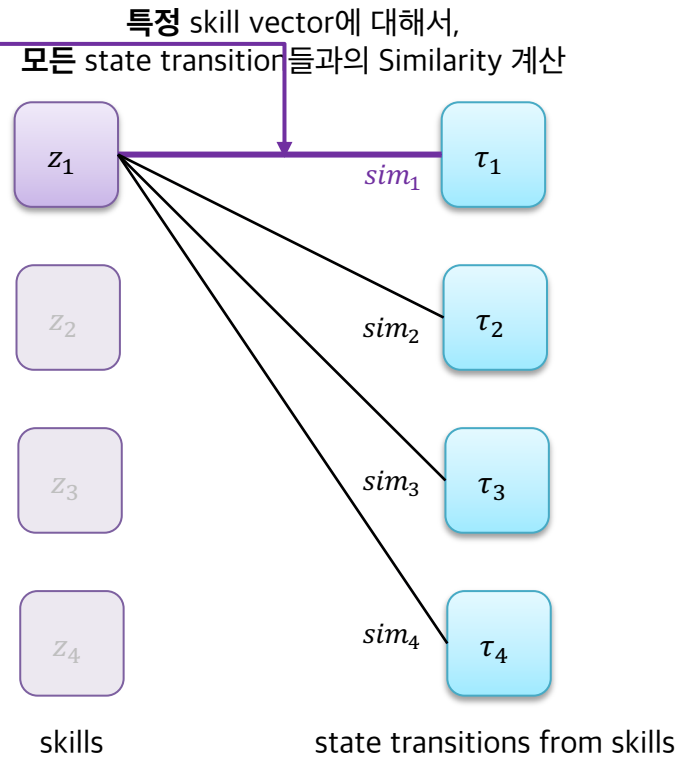
Maximize (좌) / Maximize (우)

❖ ① Discriminator Parametrized with Contrastive Density Estimator

- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

Maximize $F_{CIC}(\tau) = \log \frac{\exp\left(\frac{g_{\psi_1}(\tau_i)^\top g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^\top g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$ ⓟ

이 Similarity는
다른 Similarity들보다 높아야 한다!
(skill과 state transition이 원래의 짝으로 매칭되는 경우)
(=Positive Pair)



Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{①}\mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

(Maximize) (Maximize)

❖ ① Discriminator Parametrized with Contrastive Density Estimator

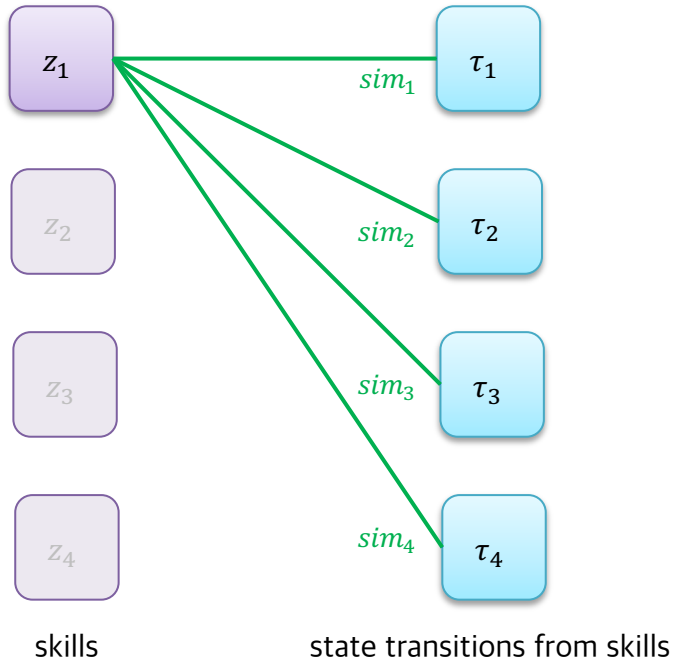
- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

Maximize $F_{CIC}(\tau) = \log \frac{\exp\left(\frac{g_{\psi_1}(\tau_i)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$ ⓝ

모든 Similarity의 평균

(1 Positive Pair, $N-1$ Negative Pairs)

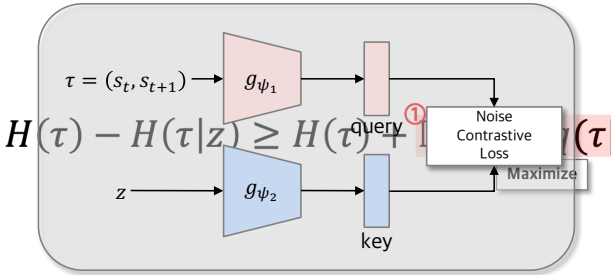
특정 skill vector에 대해서,
모든 state transition들과의 Similarity 계산



Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류



$$I(\tau; z) = H(\tau) - H(\tau|z) \geq H(\tau) + \text{[illegible]}(\tau|z)]$$

Maximize

❖ ① Discriminator Parametrized with Contrastive Density Estimator

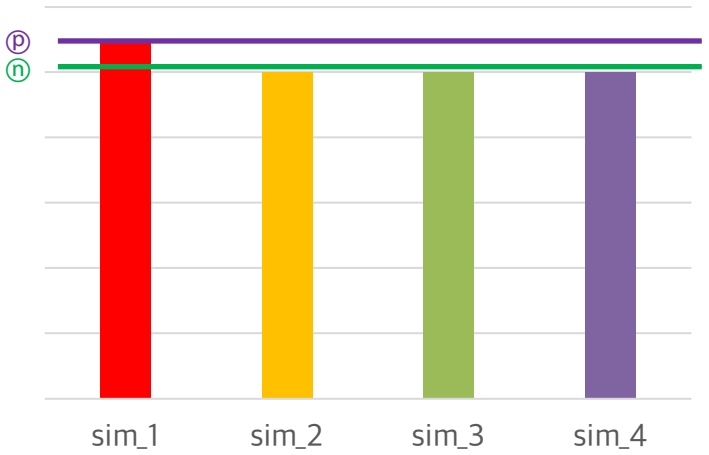
- Discriminator는 특정 State Transition(τ)이 특정 Skill(z)로부터 비롯되었는지 구분하는 역할
- Contrastive Density Estimator를 활용하여 Discriminator를 매개변수화(Parametrize)

Maximize $F_{CIC}(\tau) = \log \dfrac{\exp\left(\frac{g_{\psi_1}(\tau_i)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_i)\| \|g_{\psi_2}(z_i)\| T}\right)}{\frac{1}{N}\sum_{j=1}^{N} \exp\left(\frac{g_{\psi_1}(\tau_j)^{\intercal} g_{\psi_2}(z_i)}{\|g_{\psi_1}(\tau_j)\| \|g_{\psi_2}(z_i)\| T}\right)}$ (ⓟ / ⓝ)

→ **Maximize** $\log \frac{ⓟ}{ⓝ}$

Positive Pair의 Similarity(ⓟ)는
전체 Similarity의 평균(ⓝ)보다 커야 한다

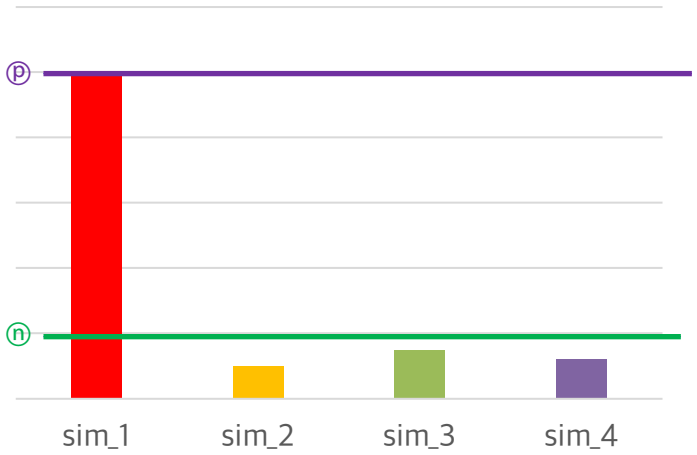
Not Desirable



$ⓟ \approx ⓝ$

$\log \frac{ⓟ}{ⓝ} \rightarrow 0$

Desirable



$ⓟ \gg ⓝ$

$\log \frac{ⓟ}{ⓝ} \uparrow$

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

$$I(\tau; z) = H(\tau) - H(\tau|z) \geq \text{②}\, H(\tau) + \mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

Maximize (for $I(\tau; z)$) Maximize (for ② $H(\tau)$)

❖ ② Particle Entropy Estimate for Exploration [최종 목표]

- Agent는 탐색 과정에서 가능한 한 다양한 State Transition을 방문해야 함
- 현재의 State Transition의 Embedding과 기존에 방문했던 State Transition의 Embedding들 사이의 거리가 최대한 멀어지도록 함

Maximize $H_{particle}(\tau) \propto \frac{1}{N_k} \sum_{h_i^* \in N_k}^{N_k} \log\|h_i - h_i^*\|$

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

Competence-based URL : CIC [NeurIPS 2022]

URL의 종류

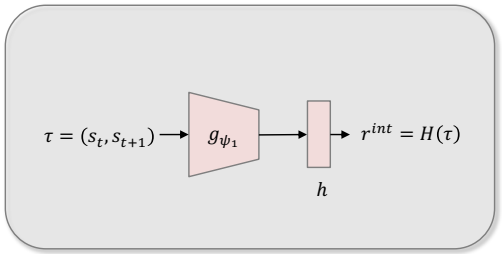
$$I(\tau; z) = H(\tau) - H(\tau|z) \geq \text{②}\,H(\tau) + \mathbb{E}_{\tau,z}[\log q(\tau|z)]$$

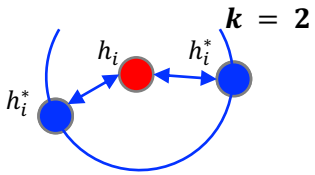
Maximize (첫 번째 항) Maximize (② $H(\tau)$)

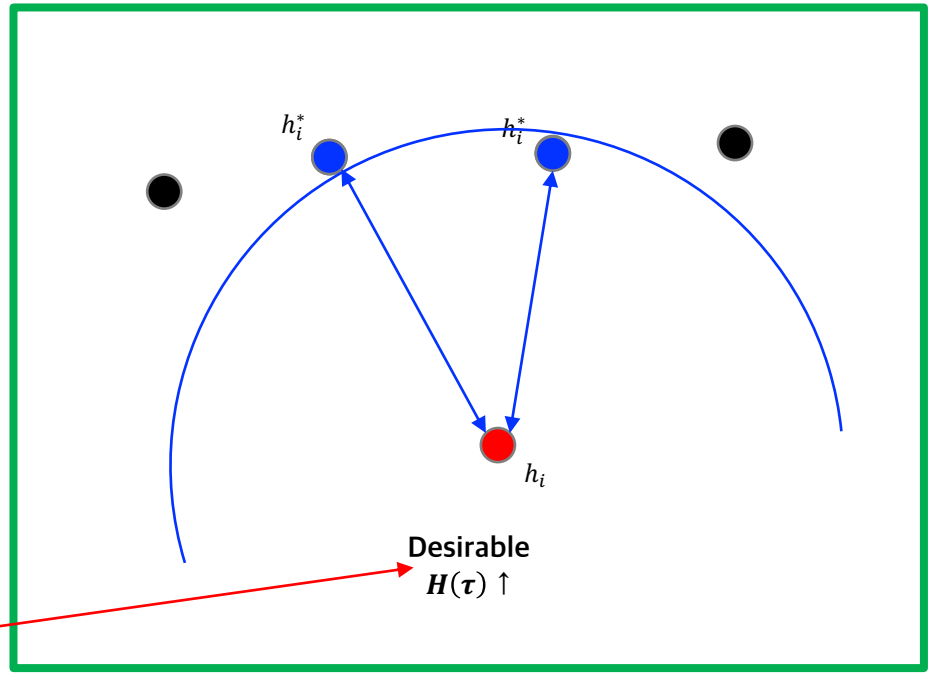
❖ ② Particle Entropy Estimate for Exploration [최종 목표]

- Agent는 탐색 과정에서 가능한 한 다양한 State Transition을 방문해야 함
- 현재의 State Transition의 Embedding과 기존에 방문했던 State Transition의 Embedding들 사이의 거리가 최대한 멀어지도록 함

Maximize $H_{particle}(\tau) \propto \frac{1}{N_k} \sum_{h_i^* \in N_k} \log\|h_i - h_i^*\|$







Not Desirable
$H(\tau) \downarrow$

r^{int}

Desirable
$H(\tau) \uparrow$

■ : 기존에 방문했던 상태의 embedding
■ : 현재 상태의 embedding (h_i)
■ : 기존에 방문했던 상태의 embedding 중 h_i의 k-nearest neighbor (h_i^*)

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

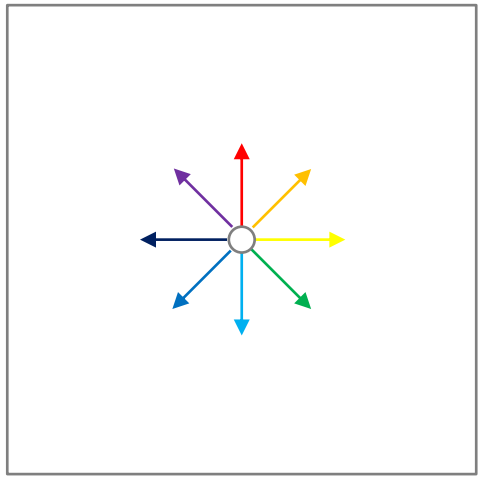
Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

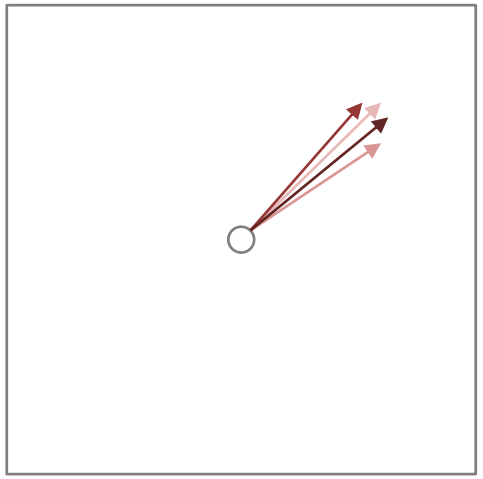
❖ Exploration vs. Exploitation

- Agent가 환경을 이해하기 위해서는 두 가지 행동 전략이 모두 필요
- **Exploration** : Agent가 아직 관찰하지 못한 상태를 찾아 돌아다니는 **탐색**
- **Exploitation** : Agent가 이미 관찰한 상태를 집중 관찰하며 이해도 **향상**
- CIC를 비롯한 기존 URL 알고리즘에서는 Pre-training 단계에서 Exploration에 치중하지만,

  MOSS의 저자들은 Pre-training 단계에서 두 전략을 모두 사용해야 함을 강조



Exploration
탐색



Exploitation
향상

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ **Pre-training 단계에서 두 학습 전략을 모두 사용할 수 있도록 하자!**

- Pre-training 단계에서 두 학습 전략을 모두 활용하는 아키텍쳐 구현을 위해 CIC 구조에 **스위치 파라미터(M)**를 도입

$$M \in \{0, 1\}$$

$M = 0$

Exploration (탐색)

Intrinsic Reward : r_{max}^{int}
(State Entropy $H(\tau)$를 **최대화**)

Skills : $Z_{max} \sim P(Z|M = 0)$

$M = 1$

Exploitation (향상)

Intrinsic Reward : $r_{min}^{int}\ (= -r_{max}^{int})$
(State Entropy $H(\tau)$를 **최소화**)

Skills : $Z_{min} \sim P(Z|M = 1)$

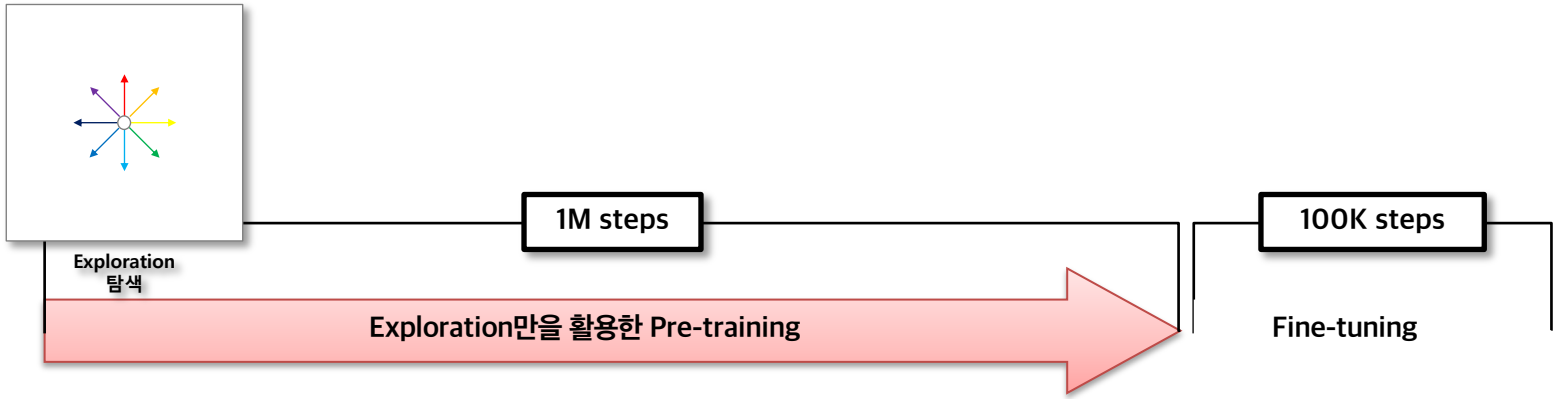
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems, 35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ 기존 CIC의 Pre-training

- CIC는 Pre-training 단계에서 Exploration을 위한 전략만을 사용
- Pre-training은 고정된 길이의 여러 Episode로 이루어짐



Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ 기존 CIC의 Pre-training

- CIC는 Pre-training 단계에서 Exploration을 위한 전략만을 사용
- Pre-training은 고정된 길이의 여러 Episode로 이루어짐

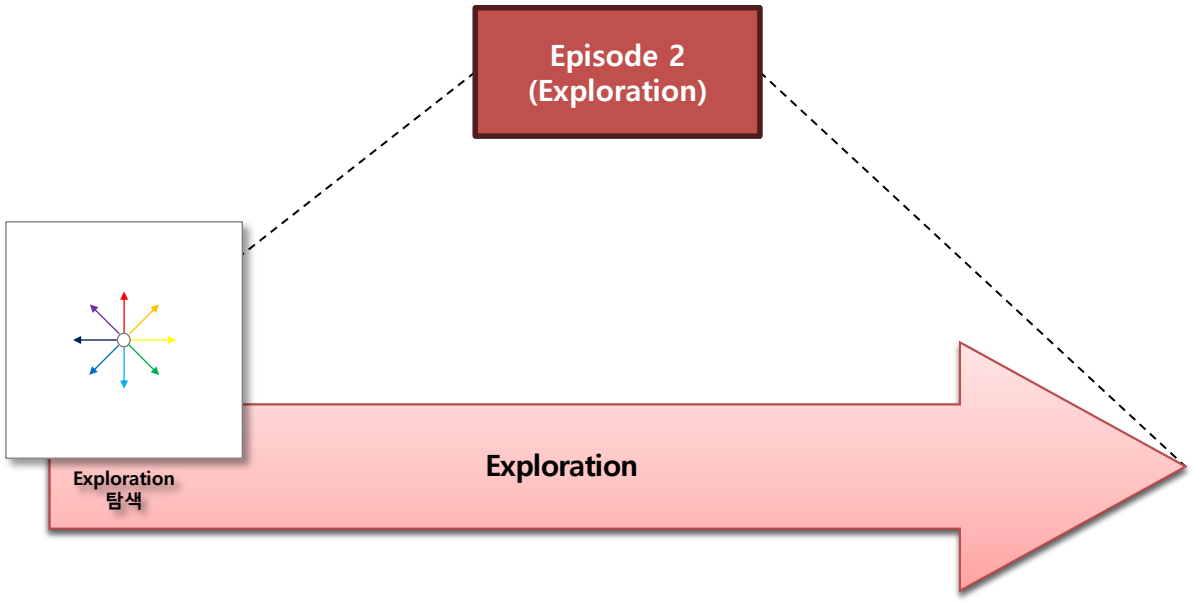
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ 기존 CIC의 Pre-training

- CIC는 Pre-training 단계에서 Exploration을 위한 전략만을 사용
- Pre-training은 고정된 길이의 여러 Episode로 이루어짐



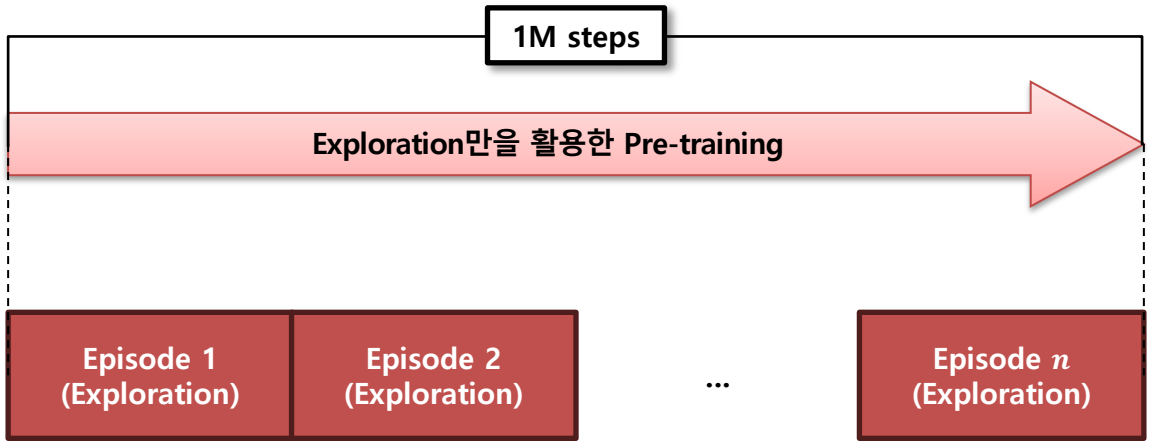
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ 기존 CIC의 Pre-training

- CIC는 Pre-training 단계에서 Exploration을 위한 전략만을 사용
- Pre-training은 고정된 길이의 여러 Episode로 이루어짐



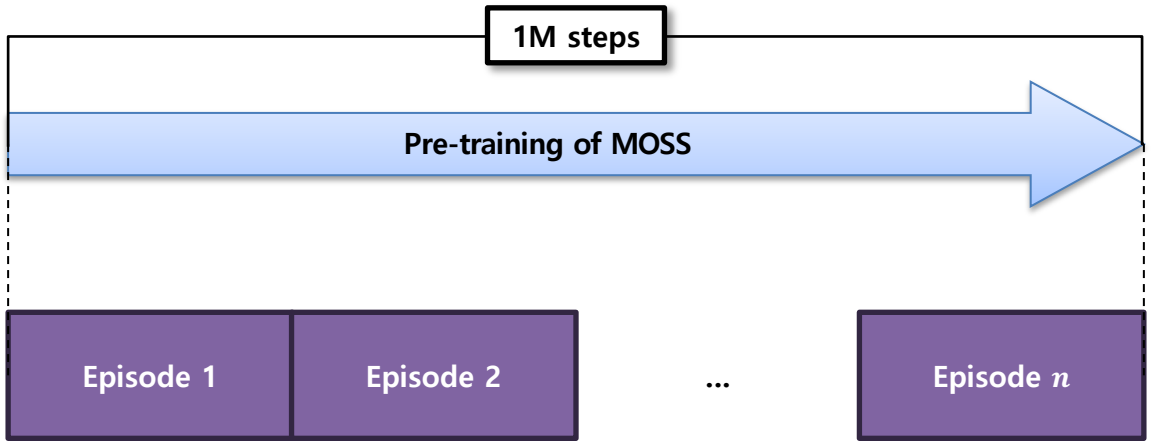
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ **MOSS의 저자들은 Pre-training 단계에서 Exploration과 Exploitation을 모두 활용해야함을 강조**

- 각 Episode를 절반으로 나누어, 탐색과 향상을 순차적으로 진행
- 기존 SOTA였던 CIC 이상의 성능을 달성하며, 두 전략 모두 사용하는 것의 유효성 입증



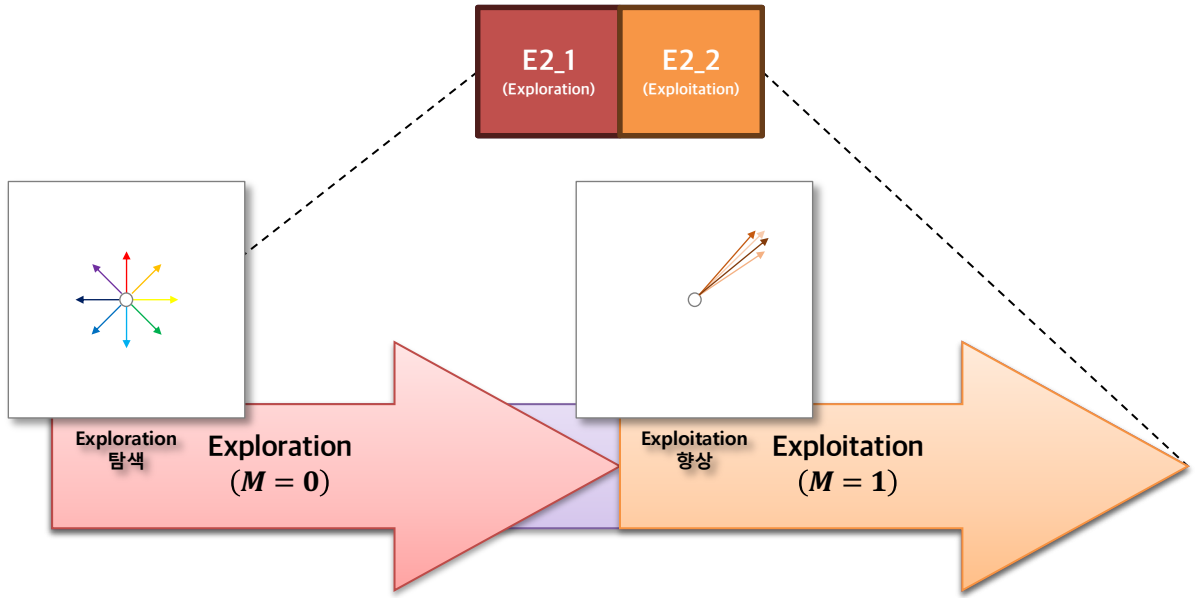
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ **MOSS의 저자들은 Pre-training 단계에서 Exploration과 Exploitation을 모두 활용해야함을 강조**

- 각 Episode를 절반으로 나누어, 탐색과 향상을 순차적으로 진행
- 기존 SOTA였던 CIC 이상의 성능을 달성하며, 두 전략 모두 사용하는 것의 유효성 입증



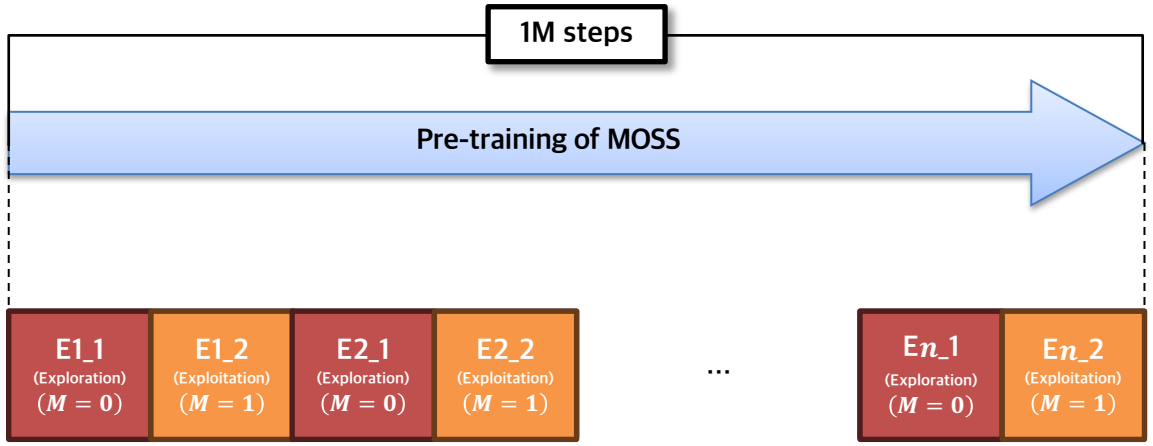
Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems, 35*, 26078-26090.

Competence-based URL : MOSS [NeurIPS 2022]

URL의 종류

❖ **MOSS의 저자들은 Pre-training 단계에서 Exploration과 Exploitation을 모두 활용해야함을 강조**

- 각 Episode를 절반으로 나누어, 탐색과 향상을 순차적으로 진행
- 기존 SOTA였던 CIC 이상의 성능을 달성하며, 두 전략 모두 사용하는 것의 유효성 입증



Zhao, A., Lin, M., Li, Y., Liu, Y. J., & Huang, G. (2022). A mixture of surprises for unsupervised reinforcement learning. *Advances in Neural Information Processing Systems, 35*, 26078-26090.

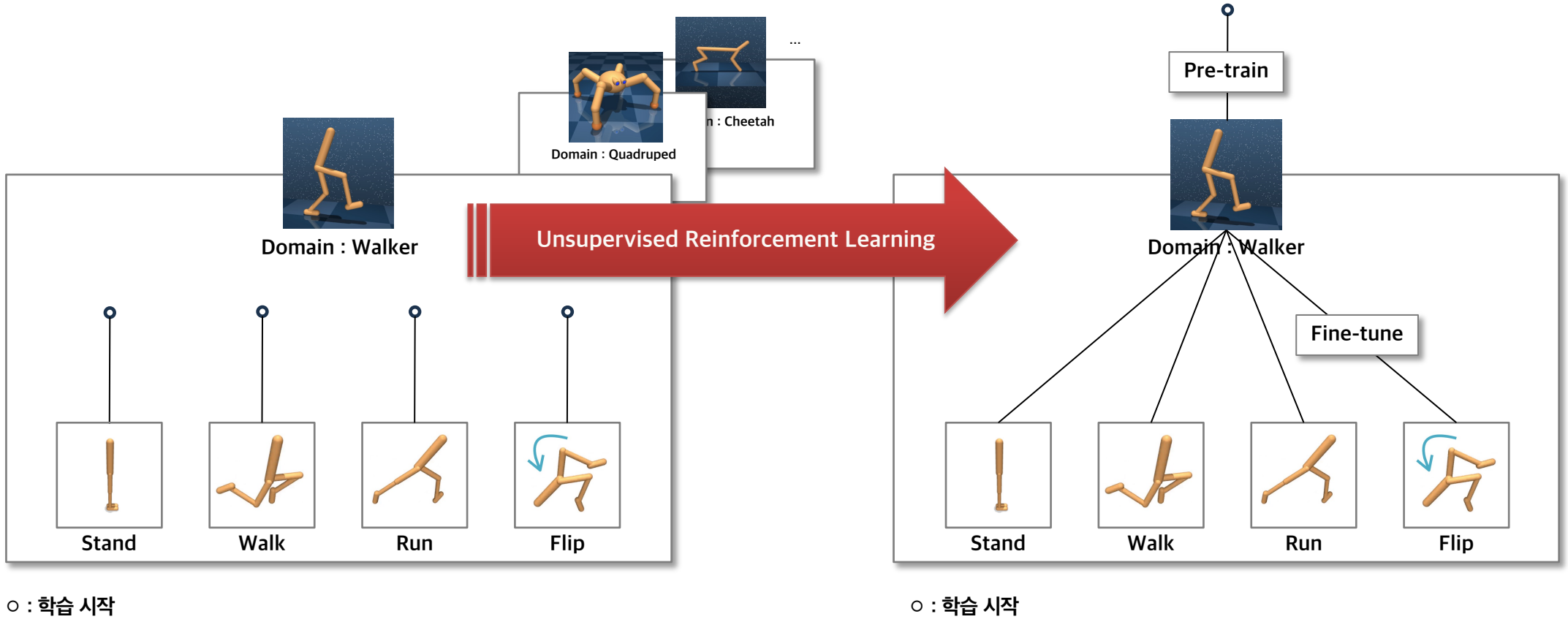
230908 DMQA Open Seminar:
Unsupervised Reinforcement Learning

3. Summary

Summary

❖ Unsupervised Reinforcement Learning

- 일반화가 어렵고, 같은 Domain 내라고 할지라도 Task별로 처음부터 학습을 다시 시켜야 하는 기존 강화학습의 한계를 극복하기 위한 연구 분야 중 하나인 Unsupervised Reinforcement Learning

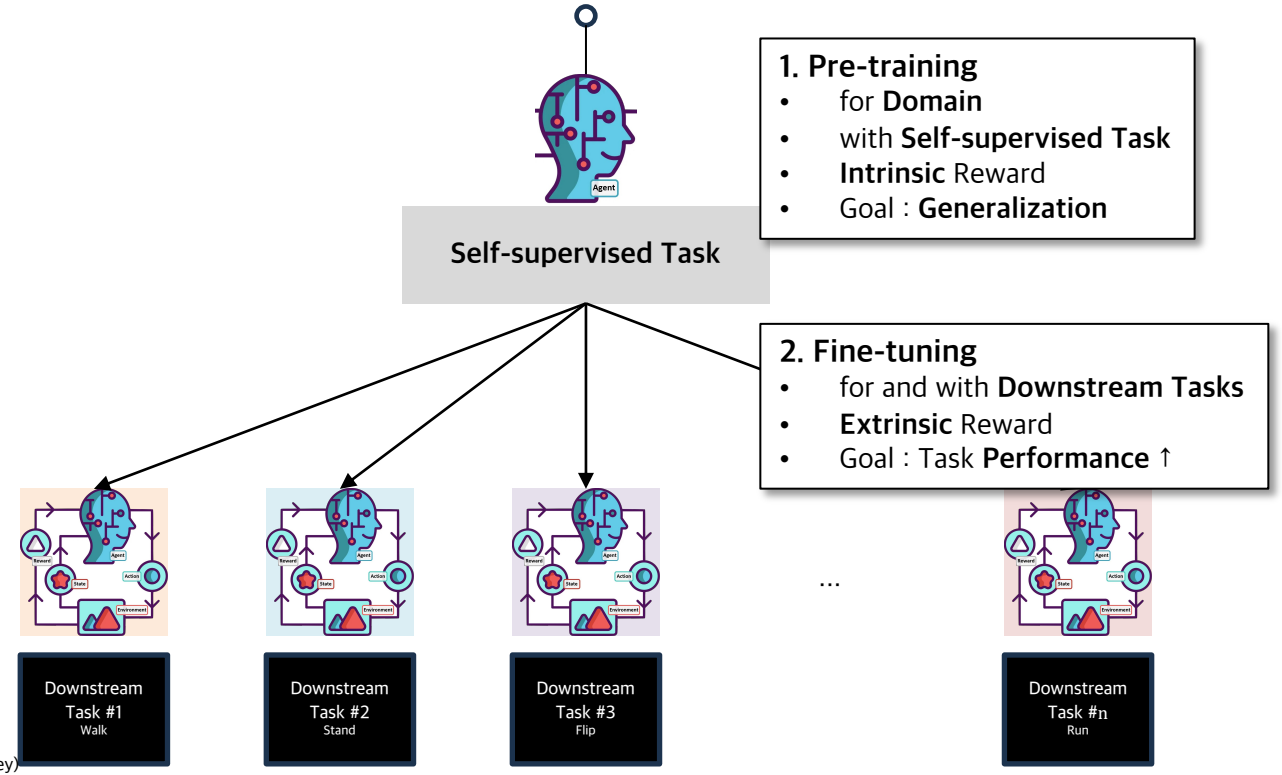


Y. Tassa, et al. "dm_control: Software and Tasks for Continuous Control" arXiv (2020)

Summary

❖ Unsupervised Reinforcement Learning의 의의

- 성공적인 Unsupervised Reinforcement Learning 알고리즘의 개발로,
- Pre-training 과정을 통해 **Domain에 대한 일반화된 학습**을 진행하여
- 해당 지식을 바탕으로 Domain 내 **구체적인 각 Task의 성능을 끌어올리는 과정을 Fine-tuning을 통해 빠르게** 진행할 수 있음

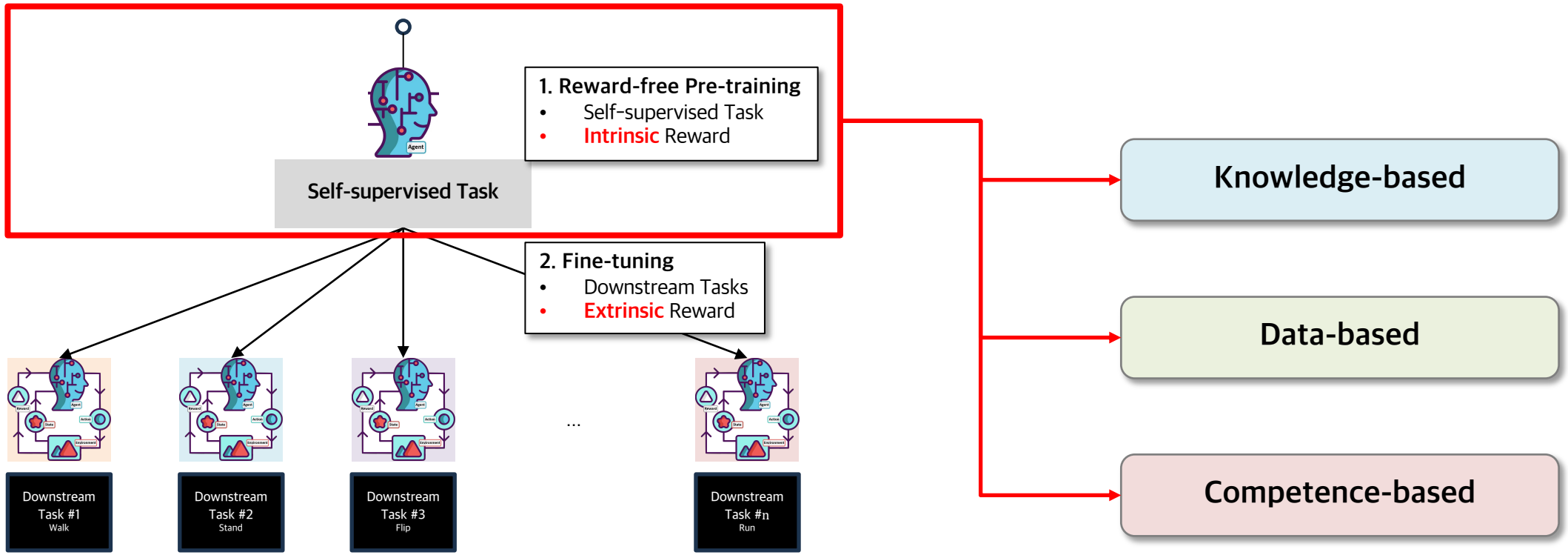


Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Summary

❖ Unsupervised Reinforcement Learning은 Pre-training 과정에서

Self-supervised Task, 또는 내부 보상Intrinsic Reward를 어떻게 정의하느냐에 따라 크게 세 종류로 나뉨



Reinforcement Leaning (image by Flat-Icons on IconScout under license to Chris Mahoney)

Thank You

References

[1] Laskin, Michael, et al. "URLB: Unsupervised Reinforcement Learning Benchmark." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[2] Vismara, Luca, Lock Yue Chew, and Vee-Liem Saw. "Optimal assignment of buses to bus stops in a loop by reinforcement learning." *Physica A: Statistical Mechanics and its Applications* 583 (2021): 126268.

[3] Tunyasuvunakool, Saran, et al. "dm_control: Software and tasks for continuous control." *Software Impacts* 6 (2020): 100022.

[4] Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

[5] Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

[6] Yarats, Denis, et al. "Reinforcement learning with prototypical representations." *International Conference on Machine Learning*. PMLR, 2021.

[7] Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

[8] Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., & Abbeel, P. (2022). Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, *35*, 34478-34491.

[9] Zhao, Andrew, et al. "A mixture of surprises for unsupervised reinforcement learning." *Advances in Neural Information Processing Systems* 35 (2022): 26078-26090.