Pretraining for Intelligent Agent

Offline Data and Simulation Perspective

Hojoon Lee, Byungkun Lee

KAIST AI

Short Bio

Hojoon Lee

- Ph.D student at KAIST AI, advised by Jaegul Choo.
- During M.S, his papers have been rejected 7 times in a row.
- **Research Interest**: Representation learning for decision making.

Byungkun Lee

- Ph.D student at KAIST AI, advised by Jaegul Choo.
- During M.S, his papers have been rejected 6 times in a row.
- **Research Interest**: I'm exploring my interest based on eps-greedy.

What is Intelligent Robot?


TSLA LIVE

What is Intelligence?

The ability to perceive, plan, act, and adapt.

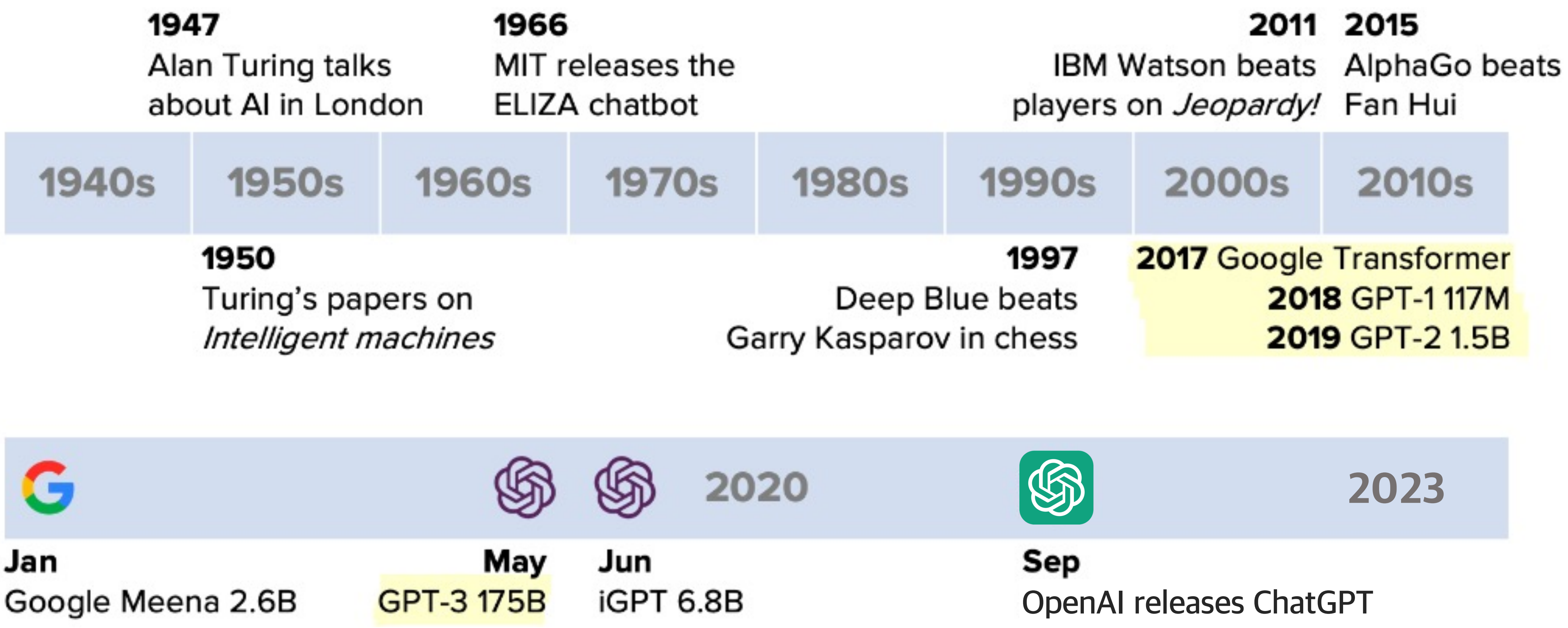
Patrick Winston

What is Artificial Intelligence?

The study of computations that can learn to perceive, plan, act, and adapt.

Patrick Winston

History of Artificial Intelligence 1947-2023



Rule-based system

Input → Hand-designed perceiver → Hand-designed actor → Output

IBM DeepBlue

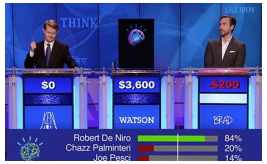
Rule-based system

Input → Hand-designed perceiver → Hand-designed actor → Output

IBM DeepBlue

Classical machine-learning system (act)

Input → Hand-designed perceiver → Learned actor → Output



IBM Watson

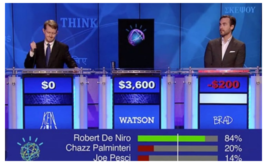
Rule-based system

Input → Hand-designed perceiver → Hand-designed actor → Output

IBM DeepBlue

Classical machine-learning system (act)

Input → Hand-designed perceiver → Learned actor → Output

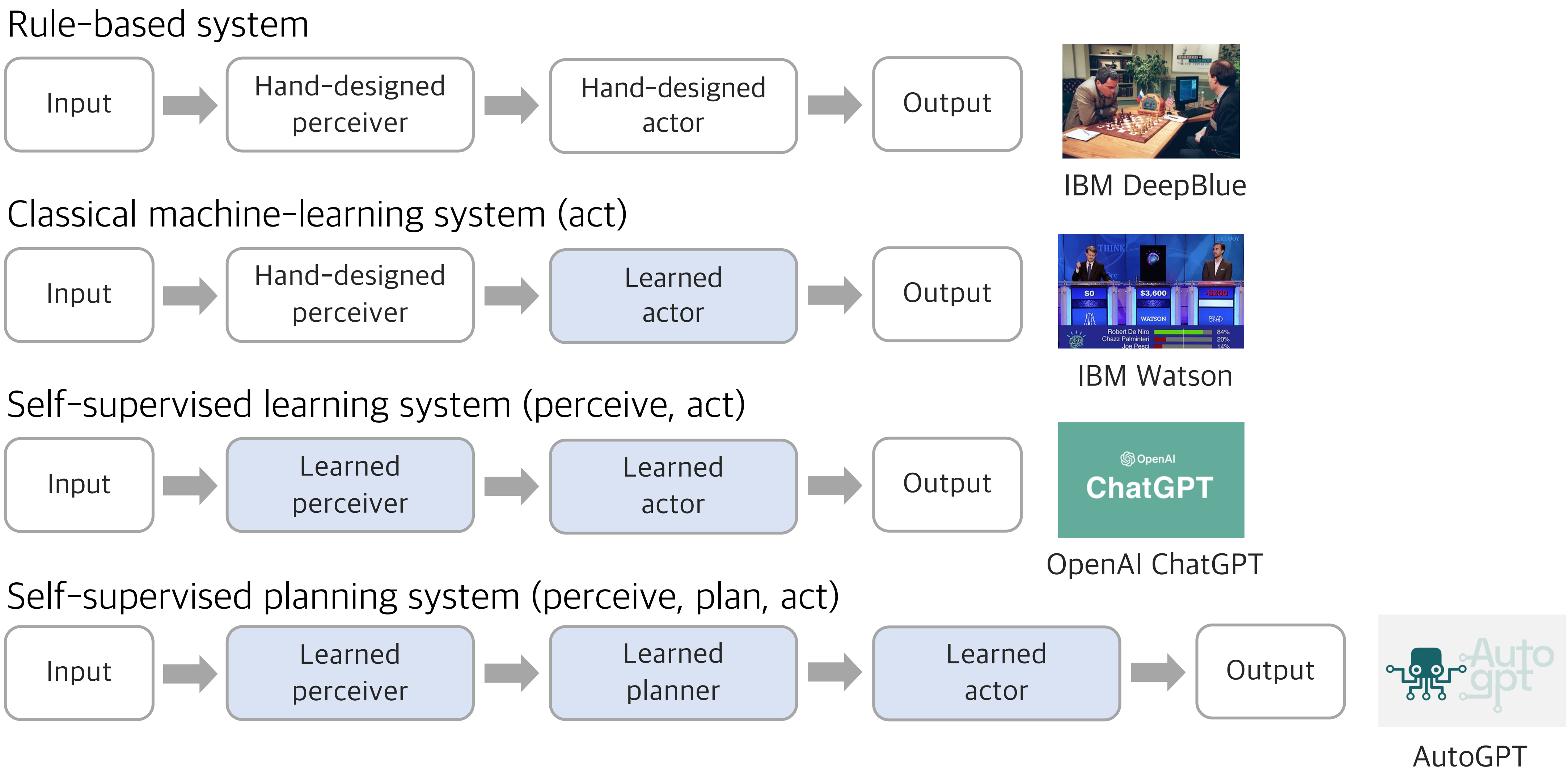


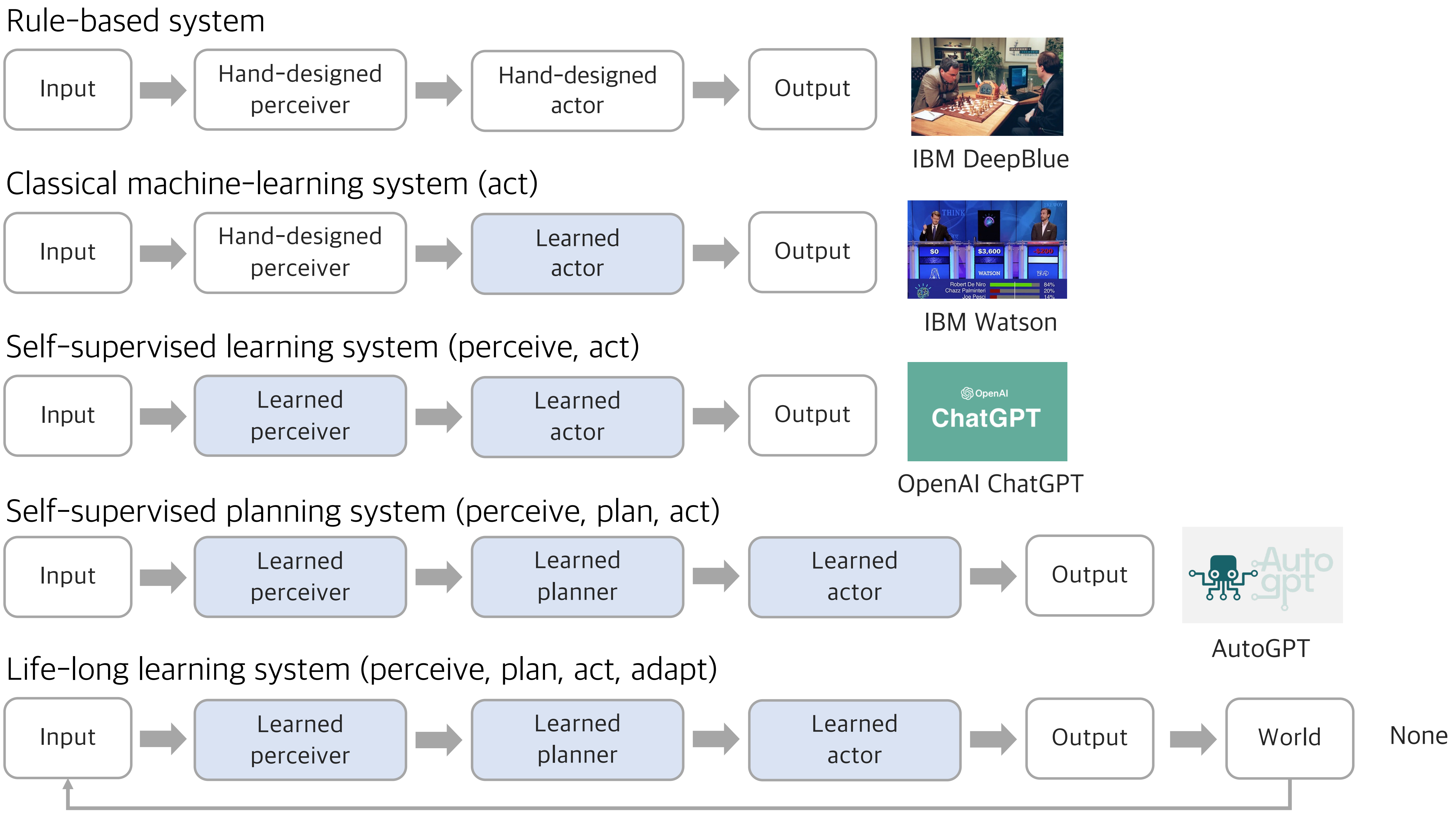
IBM Watson

Self-supervised learning system (perceive, act)

Input → Learned perceiver → Learned actor → Output

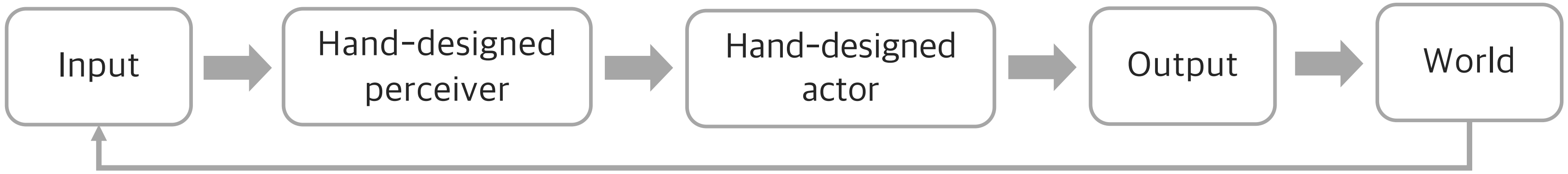
OpenAI ChatGPT

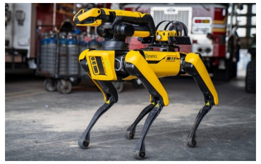

Rule-based system
Input
Hand-designed perceiver
Hand-designed actor
Output
IBM DeepBlue
Classical machine-learning system (act)
Input
Hand-designed perceiver
Learned actor
Output
IBM Watson
Self-supervised learning system (perceive, act)
Input
Learned perceiver
Learned actor
Output
OpenAI
ChatGPT
OpenAI ChatGPT
Self-supervised planning system (perceive, plan, act)
Input
Learned perceiver
Learned planner
Learned actor
Output
Auto gpt
AutoGPT


Rule-based system
Input
Hand-designed perceiver
Hand-designed actor
Output
IBM DeepBlue
Classical machine-learning system (act)
Input
Hand-designed perceiver
Learned actor
Output
IBM Watson
Self-supervised learning system (perceive, act)
Input
Learned perceiver
Learned actor
Output
OpenAI
ChatGPT
OpenAI ChatGPT
Self-supervised planning system (perceive, plan, act)
Input
Learned perceiver
Learned planner
Learned actor
Output
Auto gpt
AutoGPT
Life-long learning system (perceive, plan, act, adapt)
Input
Learned perceiver
Learned planner
Learned actor
Output
World
None

Where are we now for Intelligent Robot?

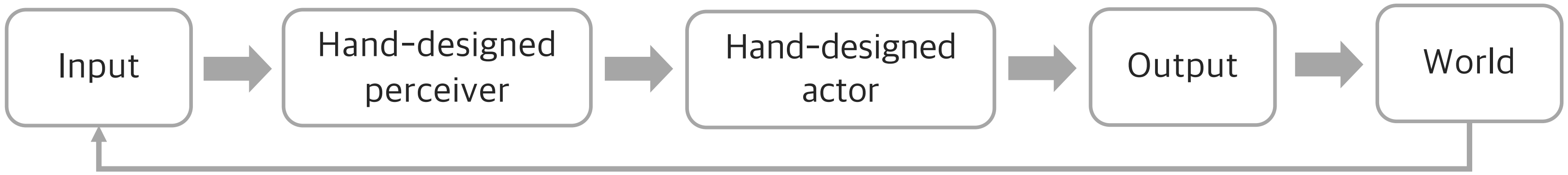
Rule-based system

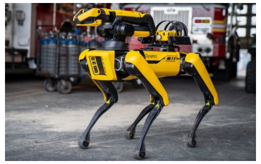




B-Dynamics Spot

Rule-based system





B-Dynamics Spot

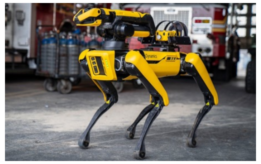
Somewhat mixed system

Tesla Bot

Tesla AutoPilot

Rule-based system

Input → Hand-designed perceiver → Hand-designed actor → Output → World → Input



B-Dynamics Spot

Somewhat mixed system

Input → Mixed perceiver → Mixed actor → Output → World → Input

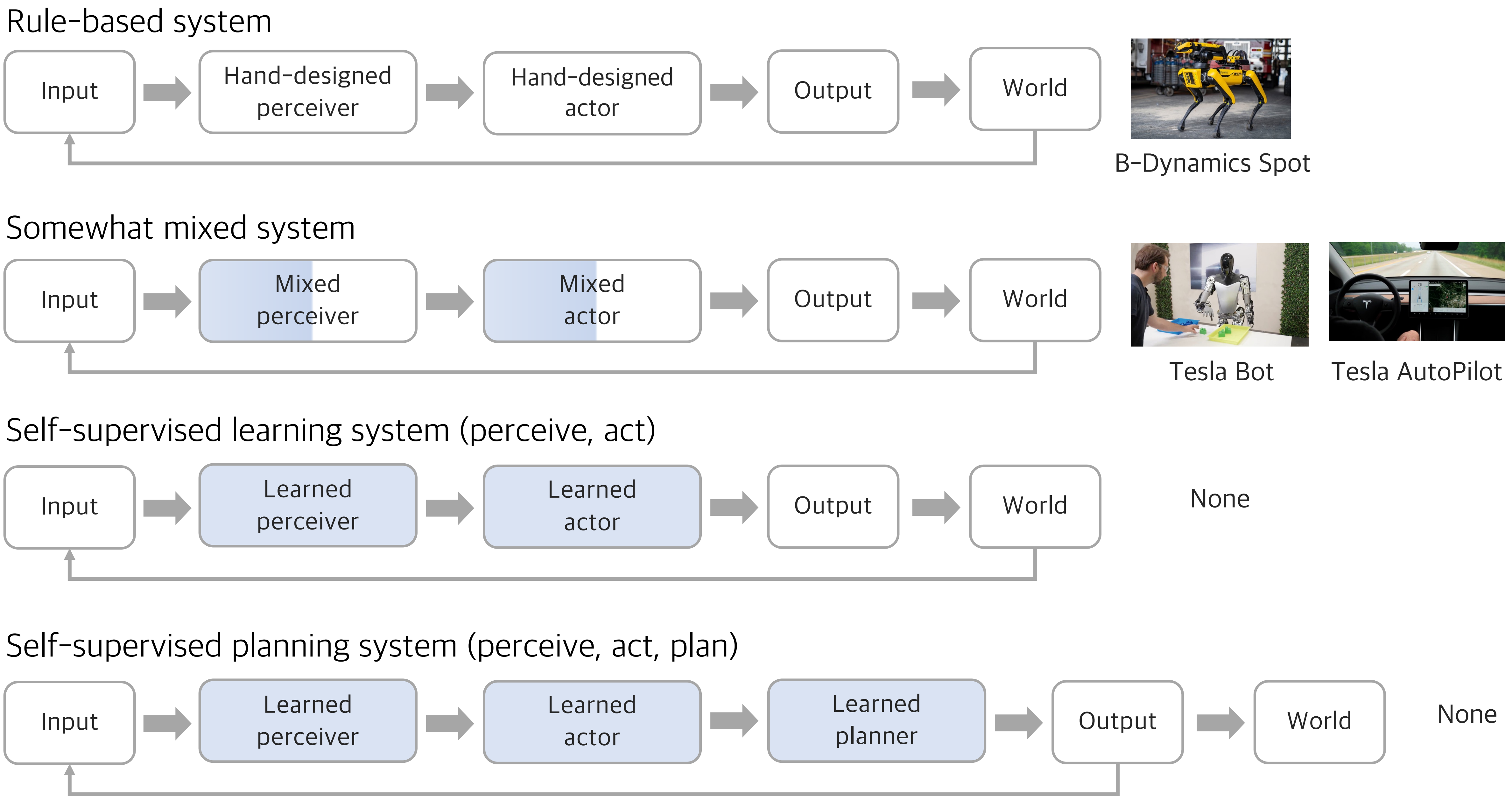
Tesla Bot

Tesla AutoPilot

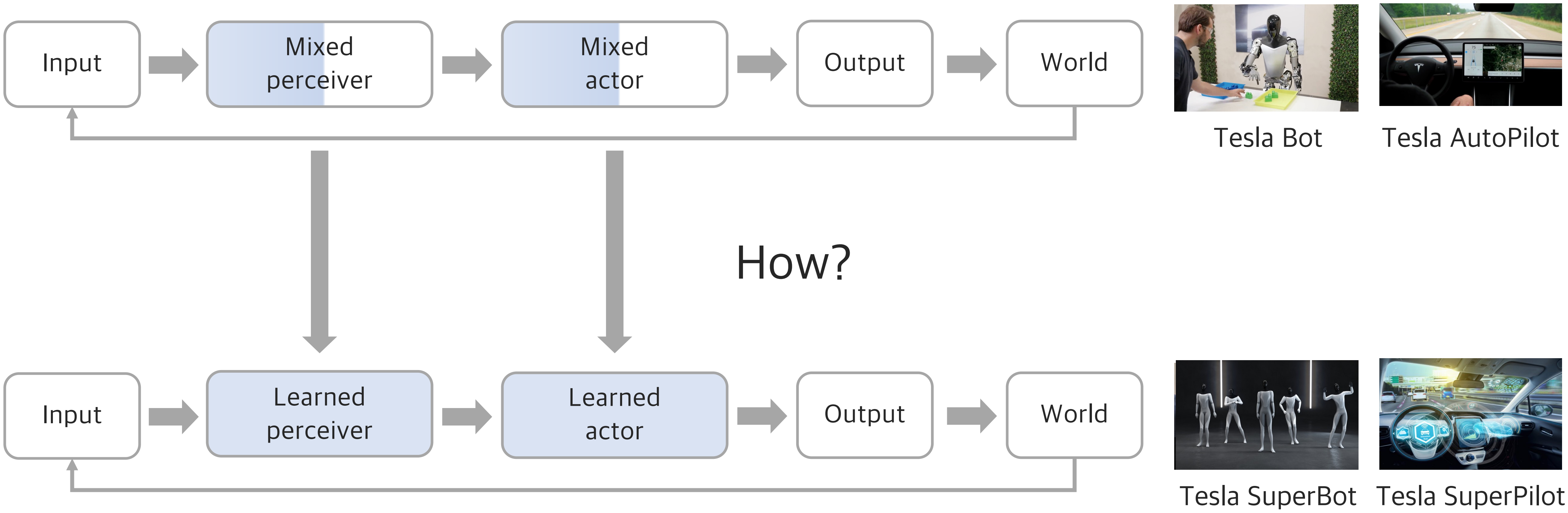
Self-supervised learning system (perceive, act)

Input → Learned perceiver → Learned actor → Output → World → Input

None


Rule-based system
Input
Hand-designed perceiver
Hand-designed actor
Output
World
B-Dynamics Spot
Somewhat mixed system
Input
Mixed perceiver
Mixed actor
Output
World
Tesla Bot
Tesla AutoPilot
Self-supervised learning system (perceive, act)
Input
Learned perceiver
Learned actor
Output
World
None
Self-supervised planning system (perceive, act, plan)
Input
Learned perceiver
Learned actor
Learned planner
Output
World
None

How can we make Intelligent Robot?


Input
Mixed perceiver
Mixed actor
Output
World
Tesla Bot
Tesla AutoPilot
How?
Input
Learned perceiver
Learned actor
Output
World
Tesla SuperBot
Tesla SuperPilot

Source of Training Robot

- Online Dataset
  - (+) explore the uncertain region.
  - (–) time-consuming, expensive to collect.

Source of Training Robot

- Online Dataset
  - (+) explore the uncertain region.
  - (–) time-consuming, expensive to collect.
- Offline Dataset
  - (+) cheaper than online dataset.
  - (–) distribution shift, cannot explore the uncertain region.

RT-X
1M robot trajectory

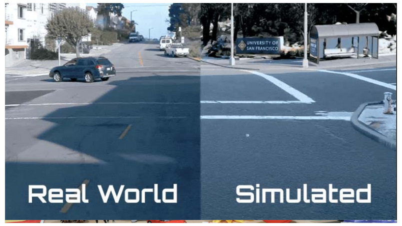
EGO-4D
ego-centric video

Source of Training Robot

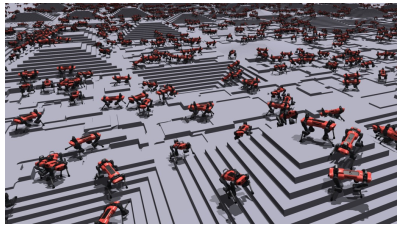
- Online Dataset
  - (+) explore the uncertain region.
  - (-) time-consuming, expensive to collect.
- Offline Dataset
  - (+) cheaper than online dataset.
  - (-) distribution shift, cannot explore the uncertain region.
- Simulator
  - (+) cheaper than online & offline dataset.
  - (-) larger distribution shift, hard to construct.

RT-X
1M robot trajectory

EGO-4D
ego-centric video

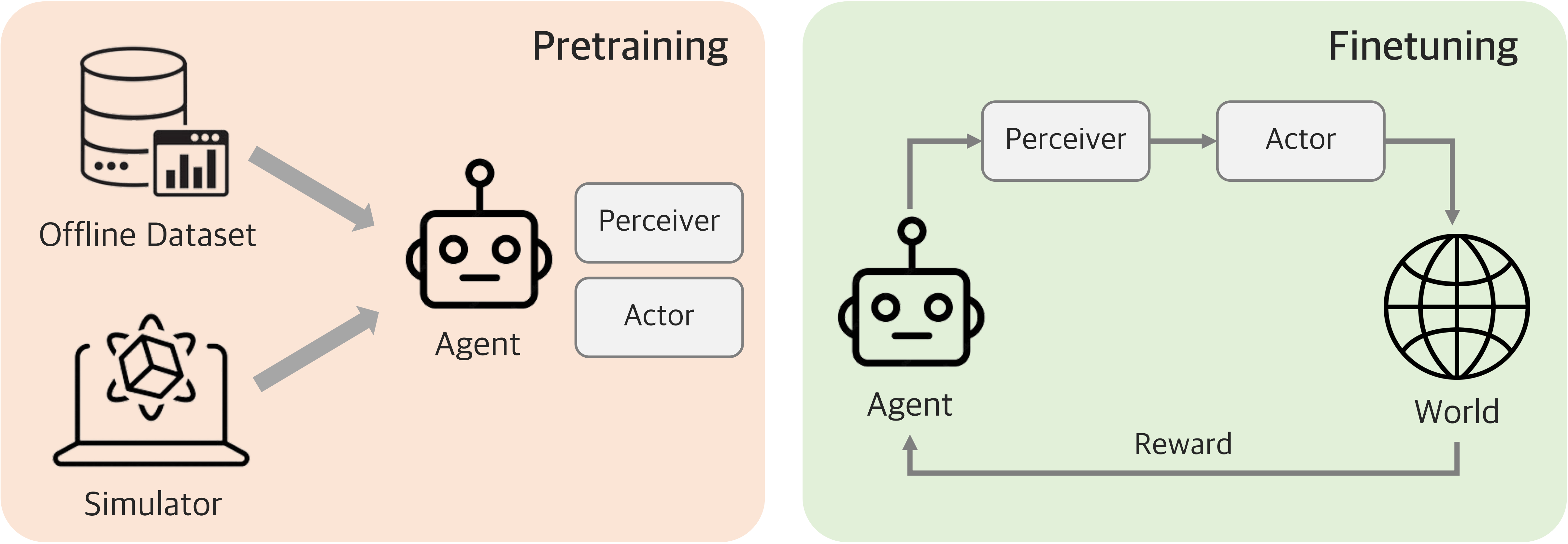


Simulation City
Waymo



Issac Gym
NVIDIA

Pretrain-then-Finetune paradigm



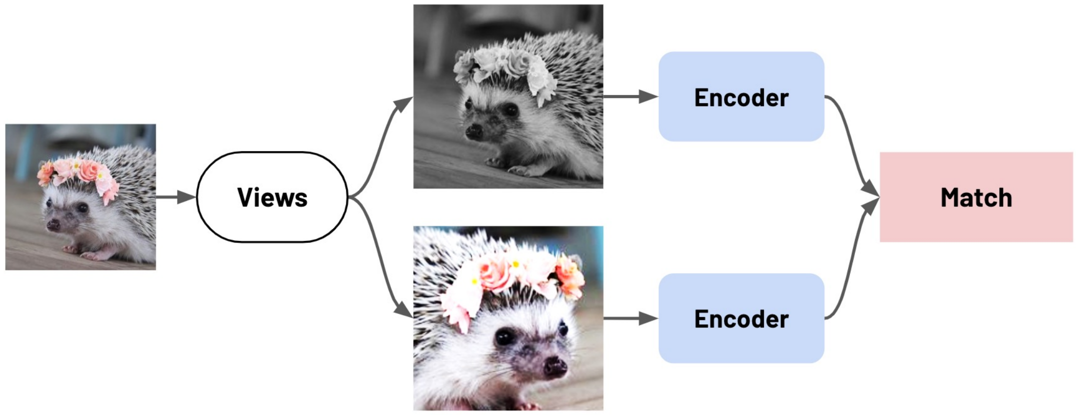
Taxonomy of Pretraining Algorithms

- **Offline Dataset**
  - Perceiver — **Offline Representation Learning**: CURL, ATC, MAE, SiamMAE
  - Perceiver, Actor — **Offline Reinforcement Learning**: CQL, DT
- **Simulator**
  - Perceiver — **Unsupervised Representation Learning**
  - Perceiver, Actor — **Unsupervised Reinforcement Learning**: RND, DIAYN, DISAGREE, P2E, APT

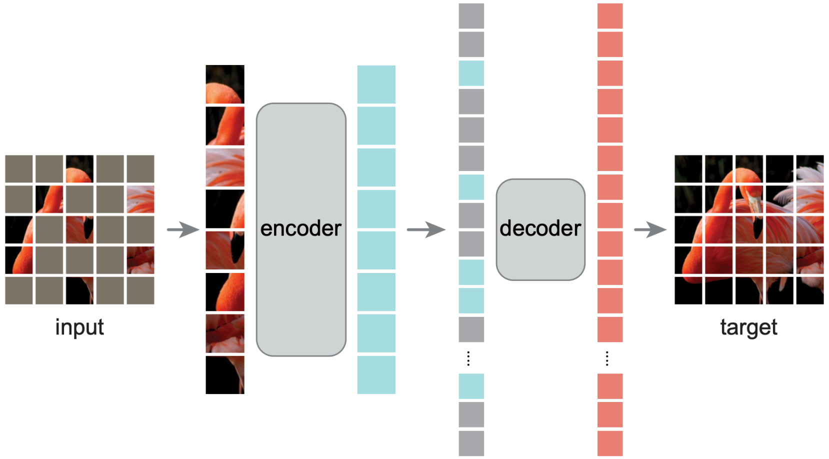
Pretraining from Offline Dataset

Offline Representation Learning

- Train effective Perceiver from offline dataset.
- Invariance vs Generation
  - Let X be data, $Z(X)$ be representation.
  - Let I denotes mutual information of two random variables.
    - **Invariance**: Maximize $I(Z(X_1); Z(X_2))$ where X_1, X_2 are invariant data.
    - **Generation**: Maximize $I(Z(\bar{X}); X)$ where $\bar{X}$ is perturbed image of X.



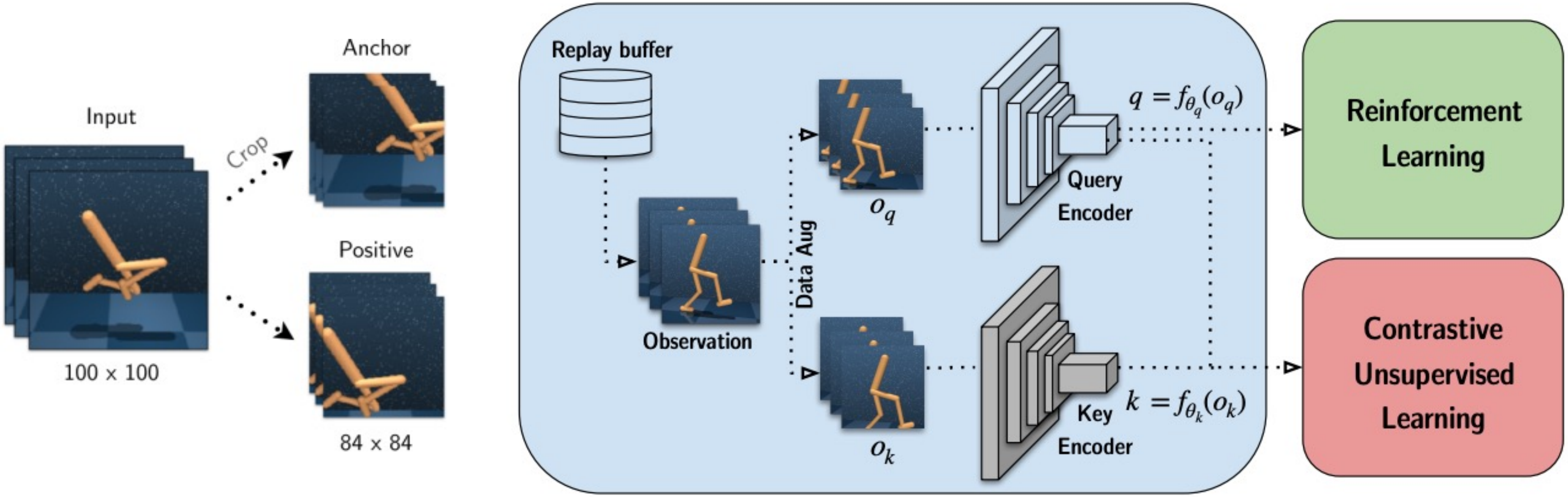
Invariance



Generation

Offline Representation Learning

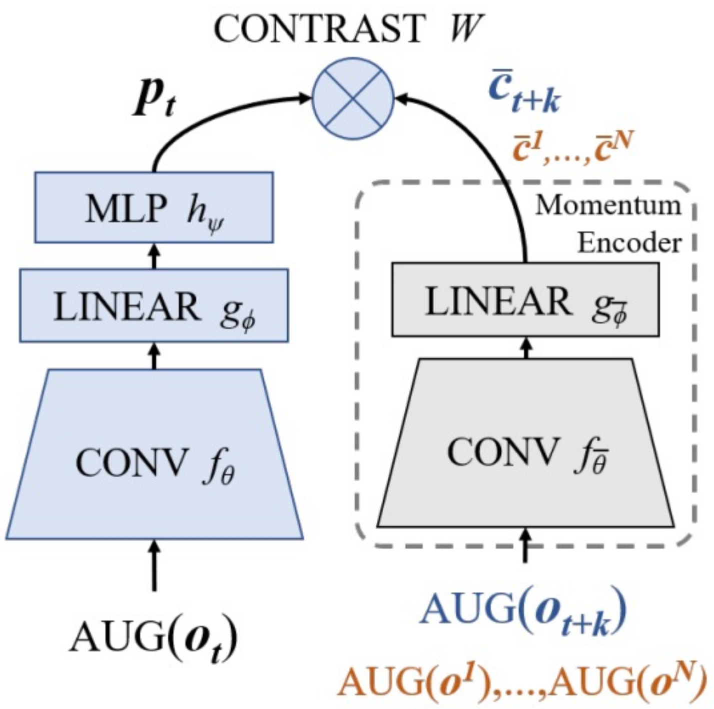
- CURL (Learning Spatial Invariance)
  - Maximize $I(Z(X_1); Z(X_2))$ where X_1, X_2 are same data with different augmentation.

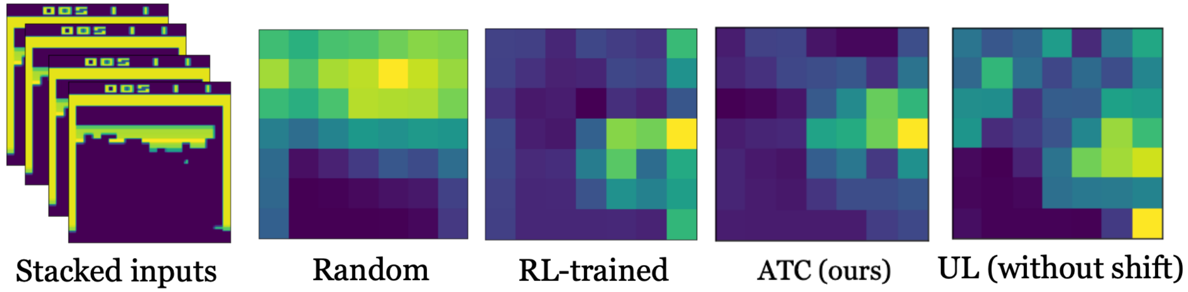


CURL: Contrastive Unsupervised Representations for Reinforcement Learning, Srinivas et al. ICML, 2020.

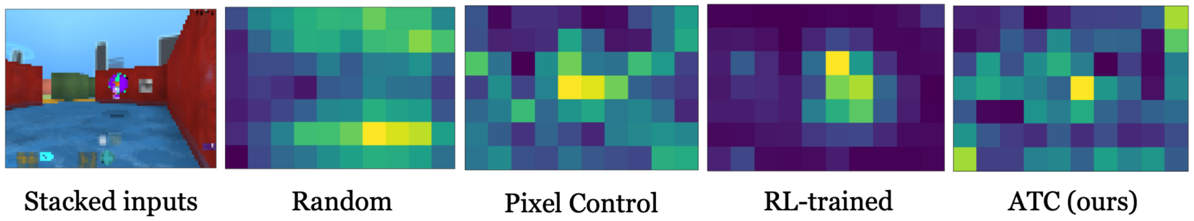
Offline Representation Learning

- ATC (Learning Spatiotemporal Invariance)
  - Maximize $I(Z(X_{t+k}); Z(X_t))$ where X_{t+k}, X_t are data from same trajectory.





ATC focuses on paddle and ball

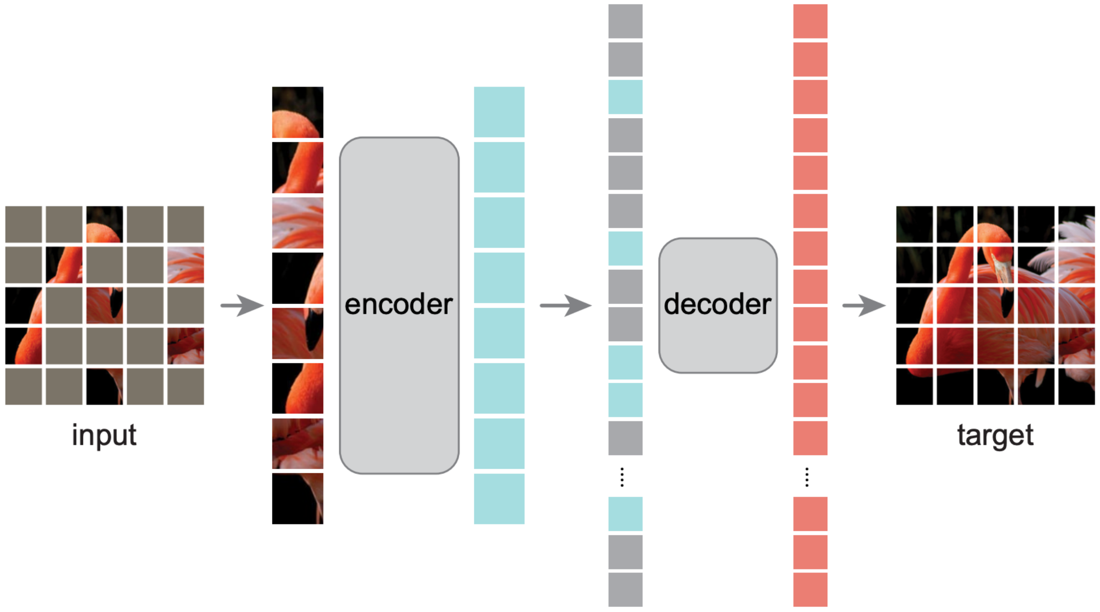


ATC focuses on enemy

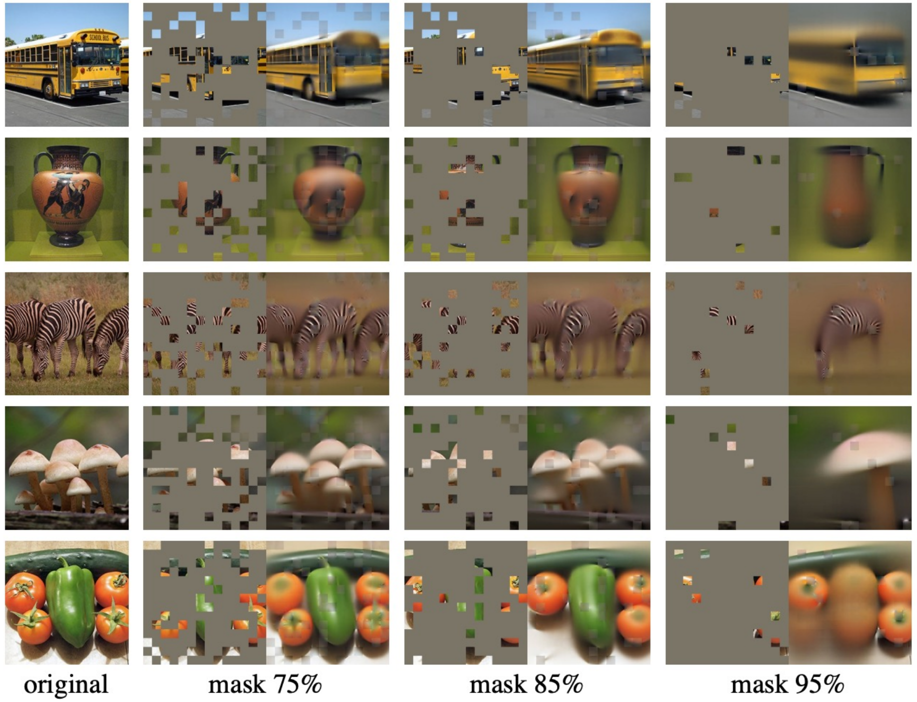
Decoupling Representation Learning from Reinforcement Learning, Stooke et al. ICML, 2021.

Offline Representation Learning

- MAE (Spatial Generation)
  - Maximize $I(Z(\bar{X}); X)$ where $\bar{X}$ is perturbed image of X.



Masked autoencoder (like BERT)

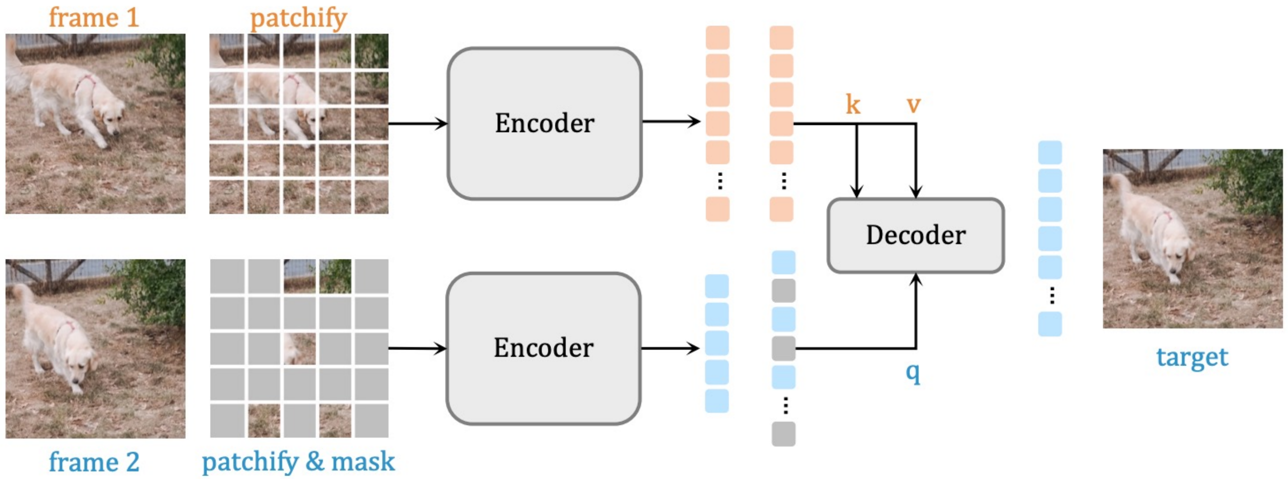


Reconstruction

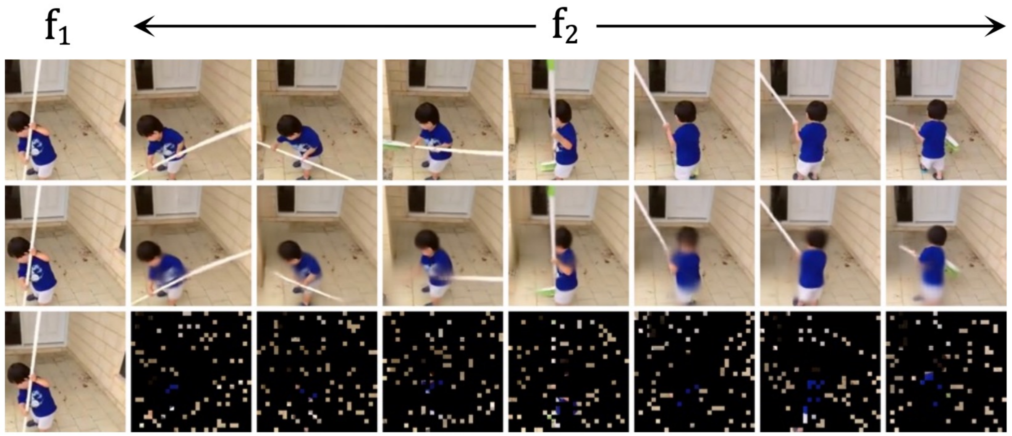
Masked Autoencoders Are Scalable Vision Learners, He et al. CVPR, 2022.

Offline Representation Learning

- SiamMAE (Spatiotemporal Generation)
  - Maximize $I(Z(\bar{X}_{t+k}); X_t)$ where $\bar{X}_{t+k}$ is perturbed image of X_{t+k}.



Masked temporal autoencoder

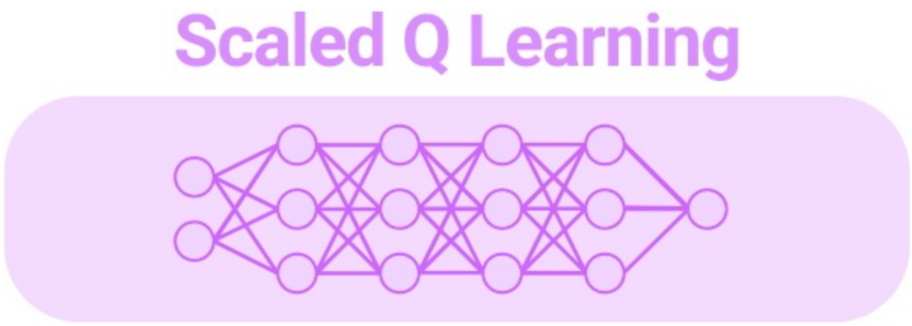


Reconstruction

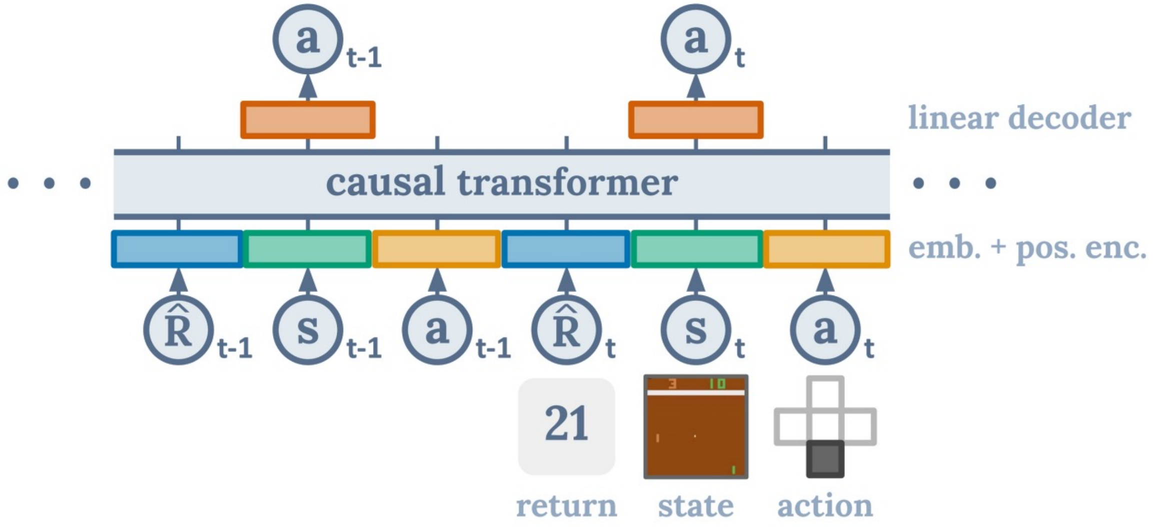
Siamese Masked Autoencoders, Gupta et al. NeurIPS, 2023.

Offline Reinforcement Learning

- Train effective Perceiver Actor from offline dataset.
- Q-learning vs Sequence Modeling



Q-Learning (Optimizing Bellman Equation)

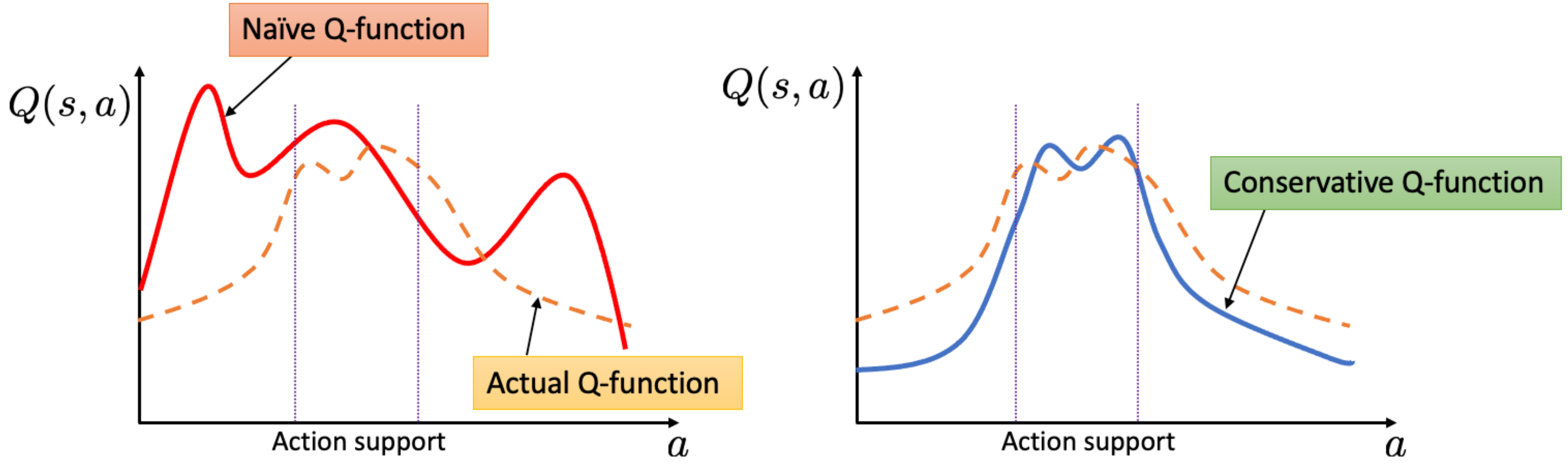


Sequence Modeling

Offline Reinforcement Learning

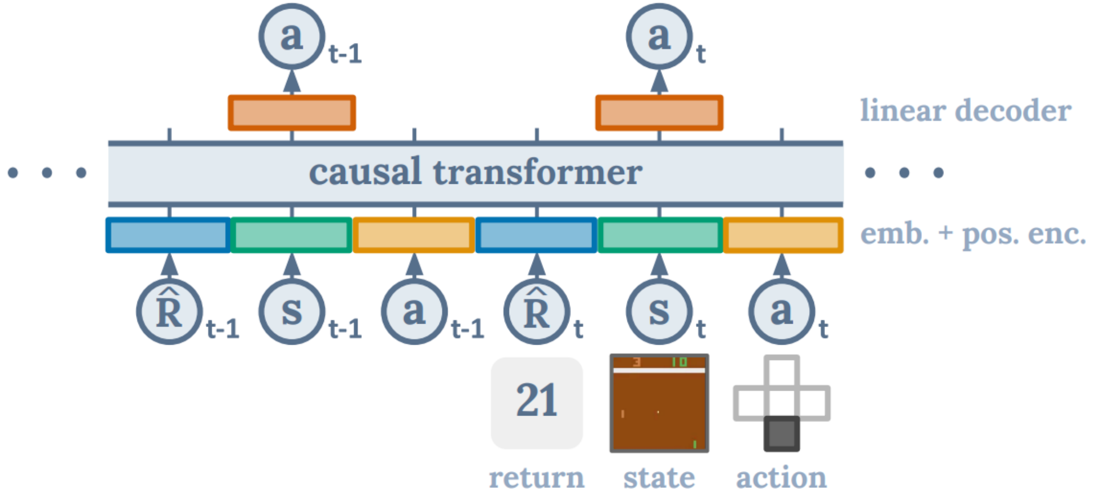
- CQL (Conservative Q-Learning)
  - Applying Q-learning to offline dataset will cause **extrapolation error.**
    - **Extrapolation error** brings **overestimation biases**.
  - Solve it by conservative Q-value estimation to unseen actions.

$$\hat{Q}^{\pi}_{\text{CQL}} := \arg\min_{Q} \alpha \cdot \left(\underbrace{\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]}_{\text{minimize Q-values}} - \underbrace{\mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\hat{\pi}_{\beta}(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})]}_{\text{maximize Q-values under data}} \right) + \underbrace{\frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q - \hat{\boldsymbol{B}}^{\pi}Q\right)^2\right]}_{\text{standard Bellman error}}$$

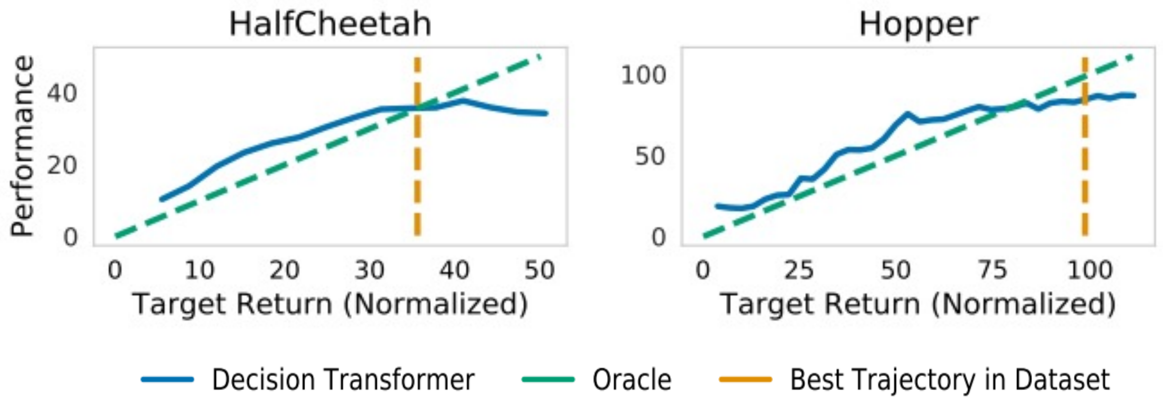


Offline Reinforcement Learning

- **DT** (Decision Transformer)
  - Formulate RL as a big sequence modeling problem.



Causal Transformer as RL agent



Reward-conditioned policy

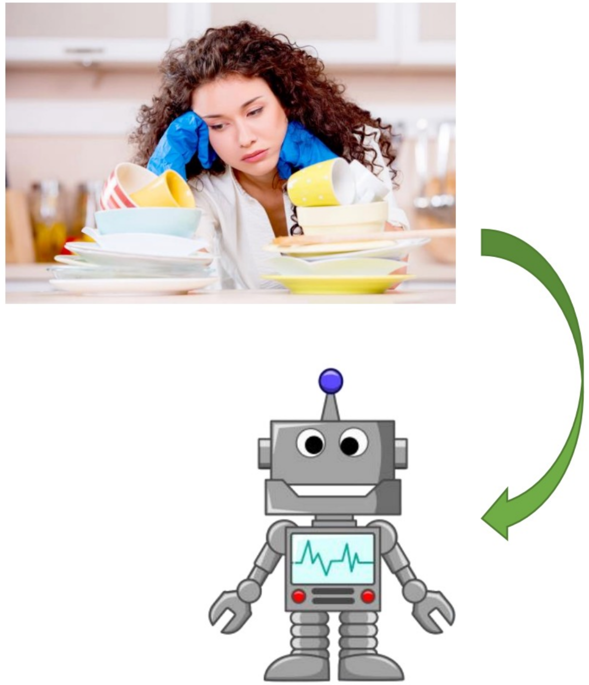
Decision Transformer: Reinforcement Learning via Sequence Modeling., Chen et al., NeurIPS 2021.

Pretraining from Simulator

Unsupervised Reinforcement Learning

- Train effective Perceiver Actor from simulator.
- Assumes we do not have an access to a pre-defined reward function.

training time: unsupervised



Berkely, CS 285: Deep Reinforcement Learning., Sergey Levine.

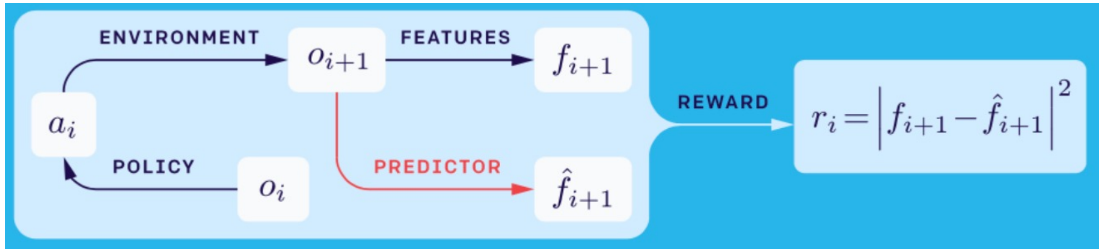
Unsupervised Reinforcement Learning

- Curiosity-driven exploration
    - Explore ‘curious’ states.
- Data coverage maximization
    - Maximize the ‘coverage of data’ collected through pretraining.
- Skill discovery
    - Learn task-agnostic skills.
- World model
    - Learn dynamics of the environment.

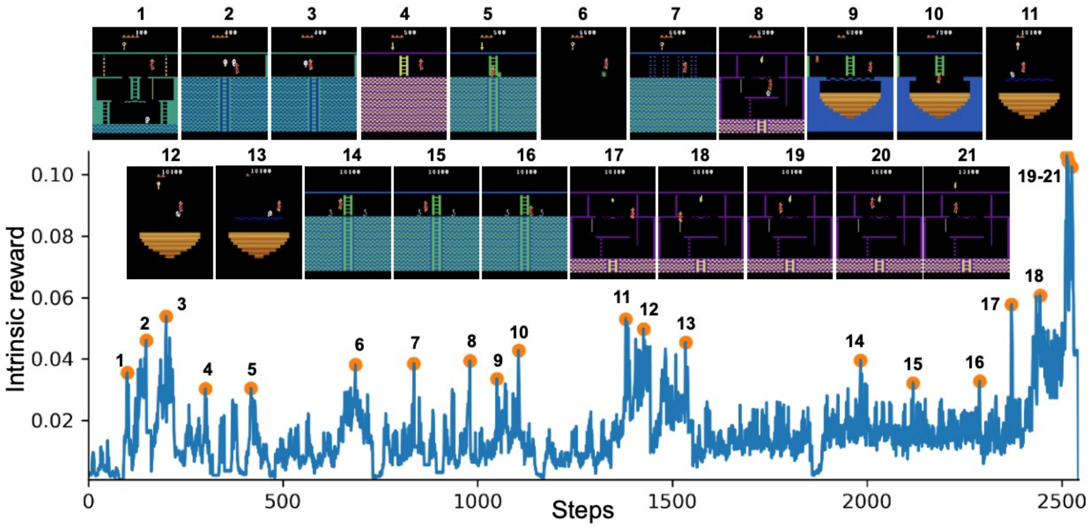
Pretraining in Deep Reinforcement Learning: A Survey, Xie. Arxiv preprint 2022.

Unsupervised Reinforcement Learning

- RND (Curiosity Driven Exploration)
  - Low prediction error: high reward



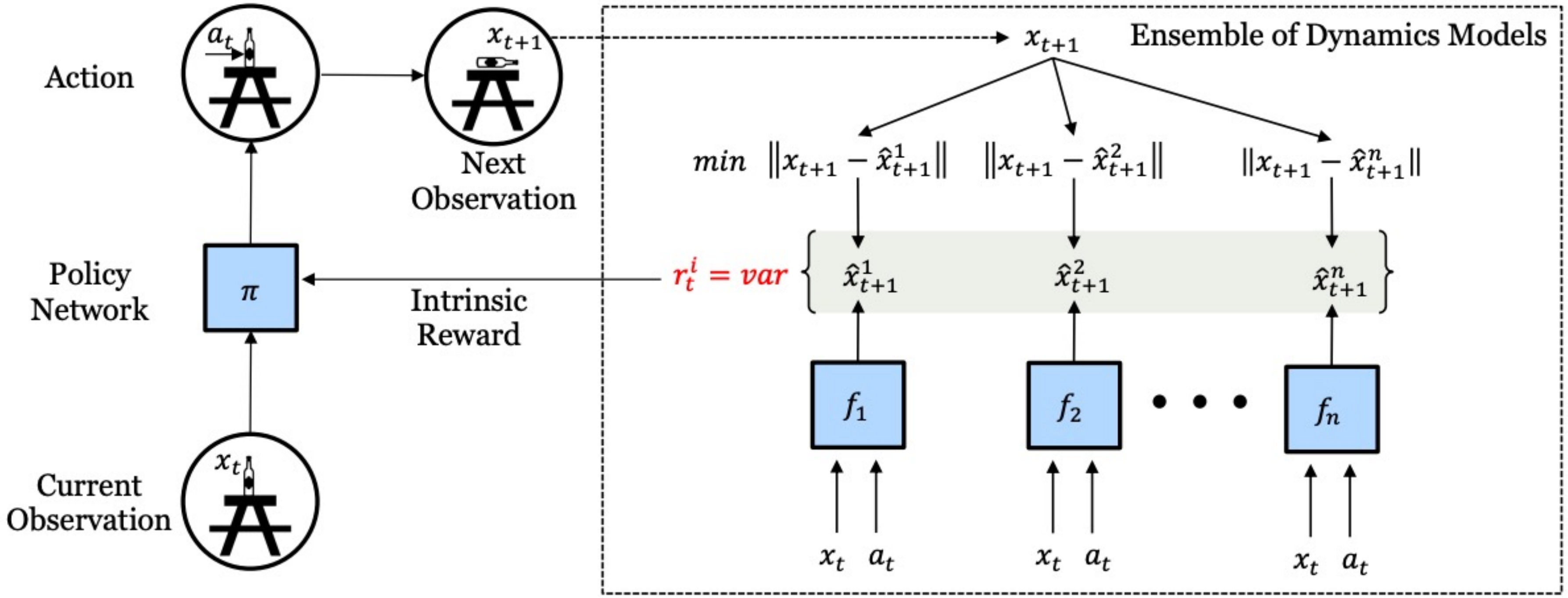
Prediction error as reward.



Spikes in reward correspond to meaningful events.

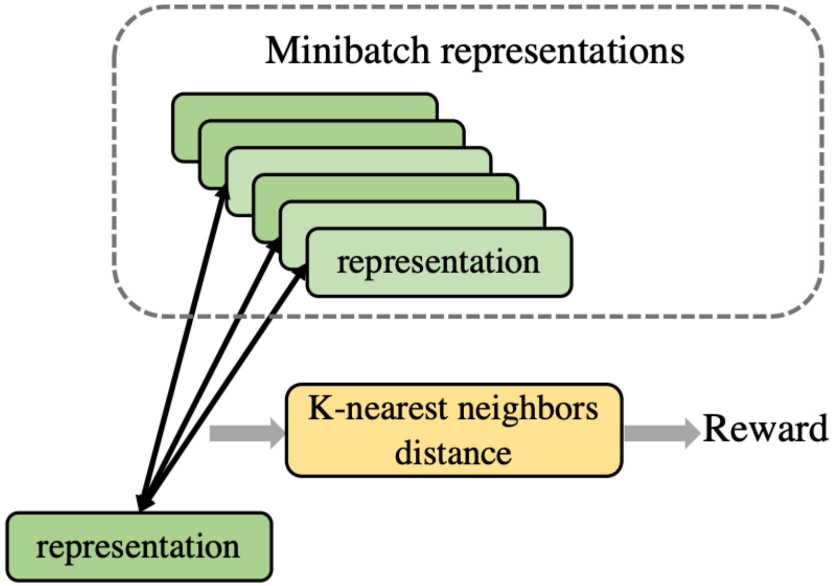
Unsupervised Reinforcement Learning

- DISAGREE (Curiosity Driven Exploration)
  - Disagreement among ensembles: high reward



Unsupervised Reinforcement Learning

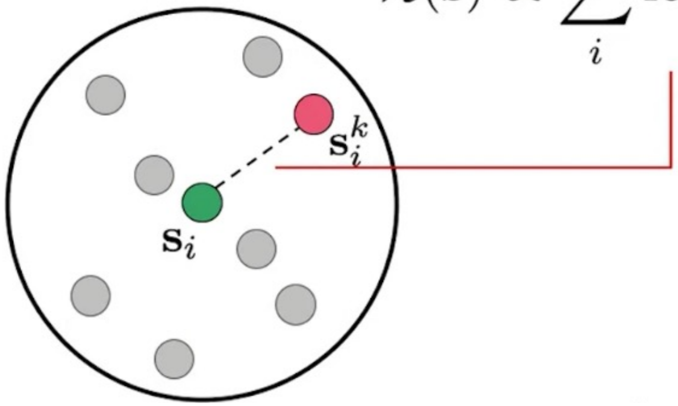
- APT (Data-coverage maximization)
  - Maximize the state entropy ($H(d_\pi)$) in replay buffer



State entropy as reward

$$\mathcal{H}(s) = -\mathbb{E}_{s\sim p(s)}\left[\log p(s)\right]$$

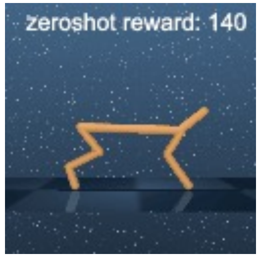
$$\widehat{\mathcal{H}}(\mathbf{s}) \propto \sum_i \log(||\mathbf{s}_i - \mathbf{s}_i^k||)$$



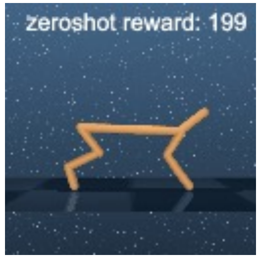
Behavior From the Void: Unsupervised Active Pre-Training, Liu et al. NeurIPS 2021.

Unsupervised Reinforcement Learning

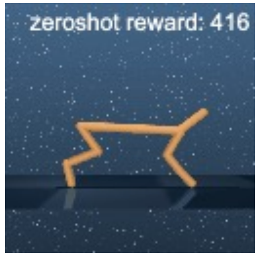
- DIAYN (Skill-Discovery)
  - Maximize the mutual information between state and skill $I(s; z)$



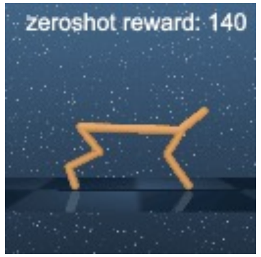
Move forward



Move backward

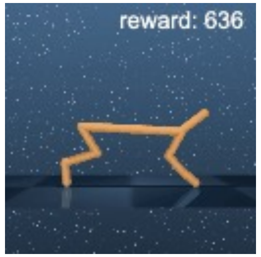


(ready for) flip



Move forward

finetune →



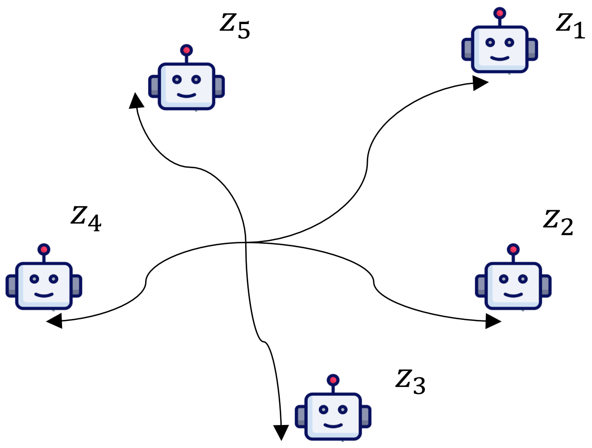
Run forward

Diversity is all you need: learning skills without a reward function, Eysenbach et al. ICLR 2019.

Unsupervised Reinforcement Learning

- DIAYN (Skill-Discovery)
  - Maximize the mutual information between state and skill $I(s; z)$

    1. Sample skill $z \sim p(z)$
    2. Rollout trajectory $\tau \sim \pi(a|s, z)$
    3. Maximize mutual information between skill z and state s

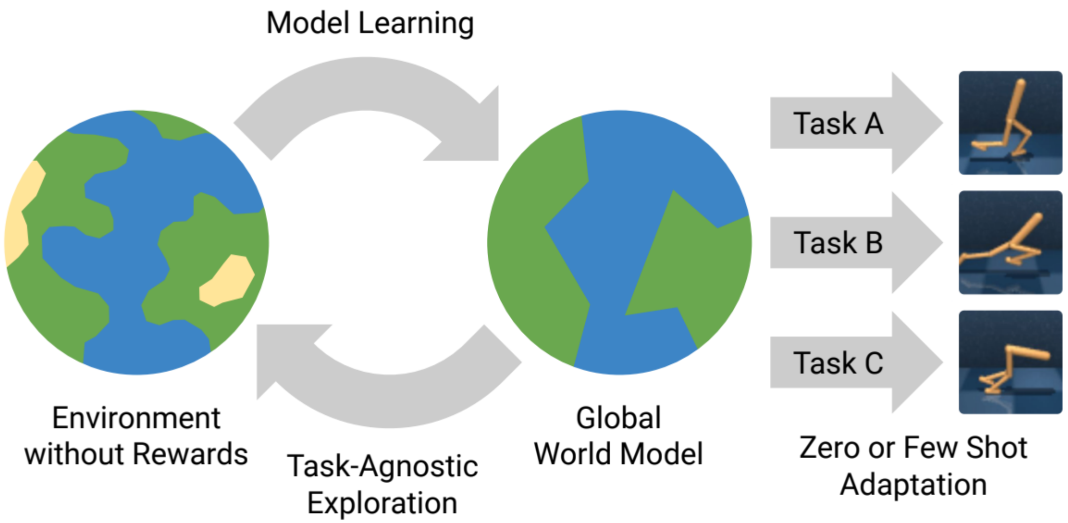


Mutual information as reward

Diversity is all you need: learning skills without a reward function, Eysenbach et al. ICLR 2019.

Unsupervised Reinforcement Learning

- **Plan2Explore** (World-Model)
  - Learn various components to represent the environment.
    1. Dynamics model $f(s_{t+1}|s_t, a_t)$
    2. (optional) reward predictor, image encoder/decoder, ⋯



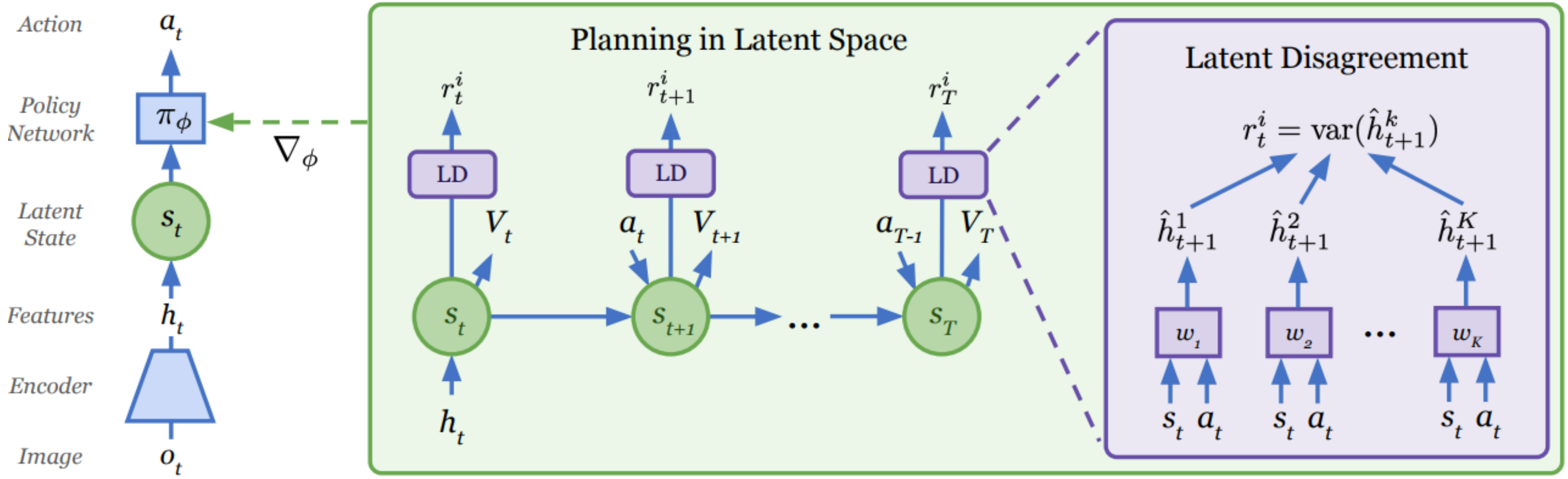
Pretraining: learn world model

Downstream task: now we can **plan** with world model

Planning to explore via self-supervised world models, Sekar et al. ICML 2020.

Unsupervised Reinforcement Learning

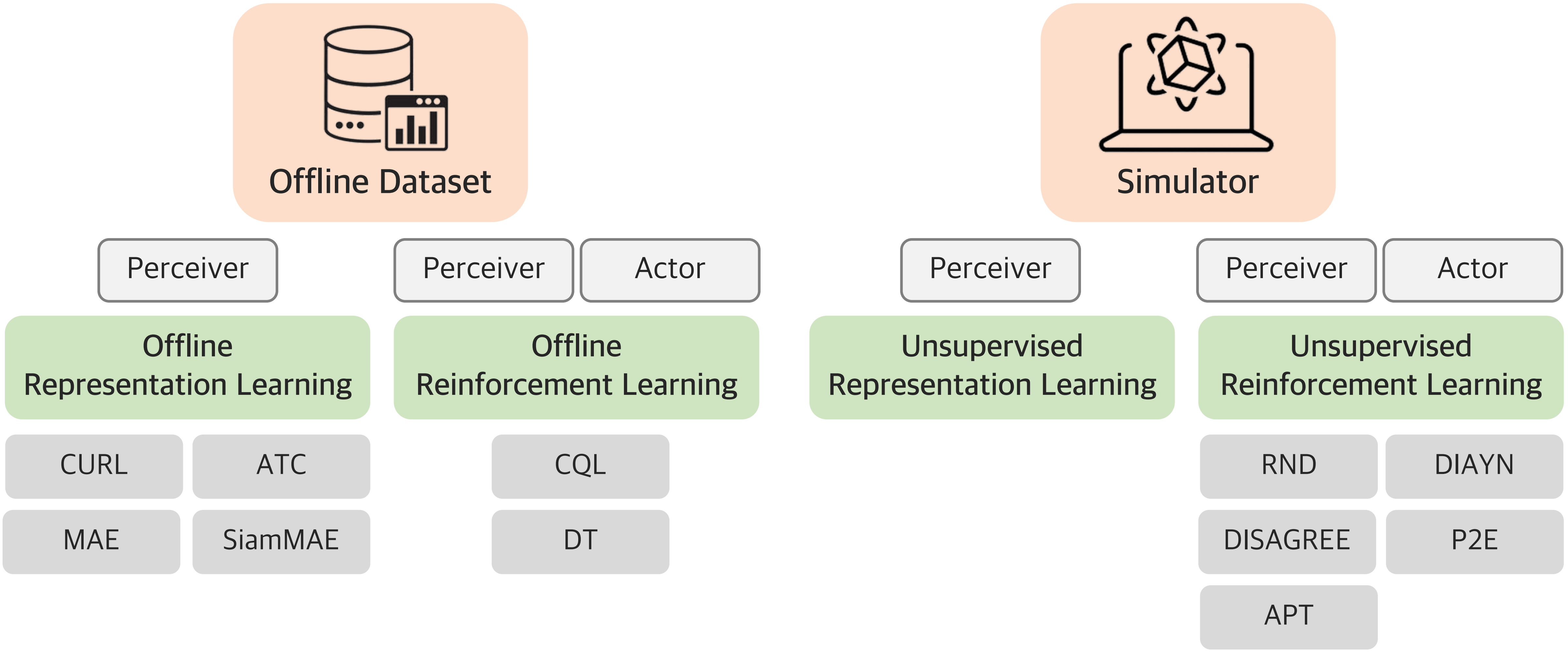
- **Plan2Explore** (World-Model)
  - Learn world model with disagreement



Planning to explore via self-supervised world models, Sekar et al. ICML 2020.

Summary: Let’s Train Robot

Summary


Offline Dataset
Simulator
Perceiver
Perceiver
Actor
Perceiver
Perceiver
Actor
Offline Representation Learning
Offline Reinforcement Learning
Unsupervised Representation Learning
Unsupervised Reinforcement Learning
CURL
ATC
MAE
SiamMAE
CQL
DT
RND
DIAYN
DISAGREE
P2E
APT

Q & A