Model-Based Active Exploration

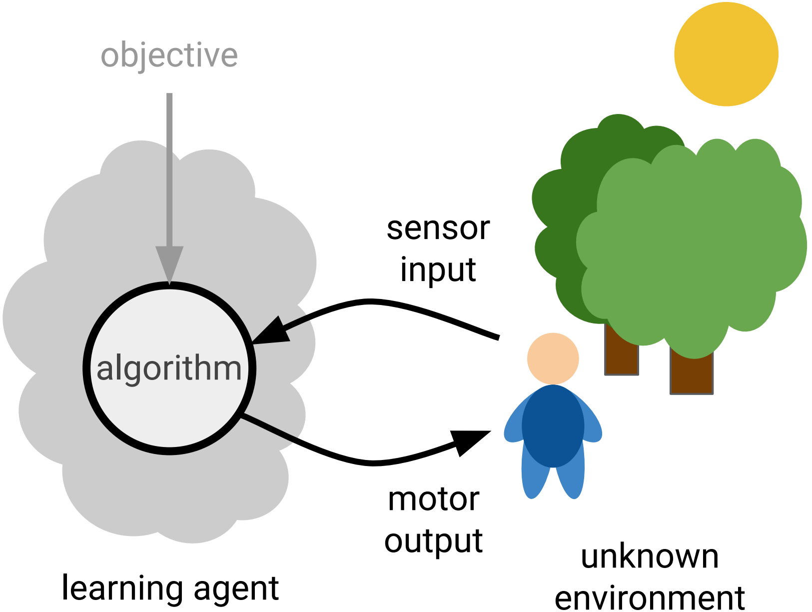
Pranav Shyam, Wojciech Jaskowski, Faustino Gomez

nnaisense

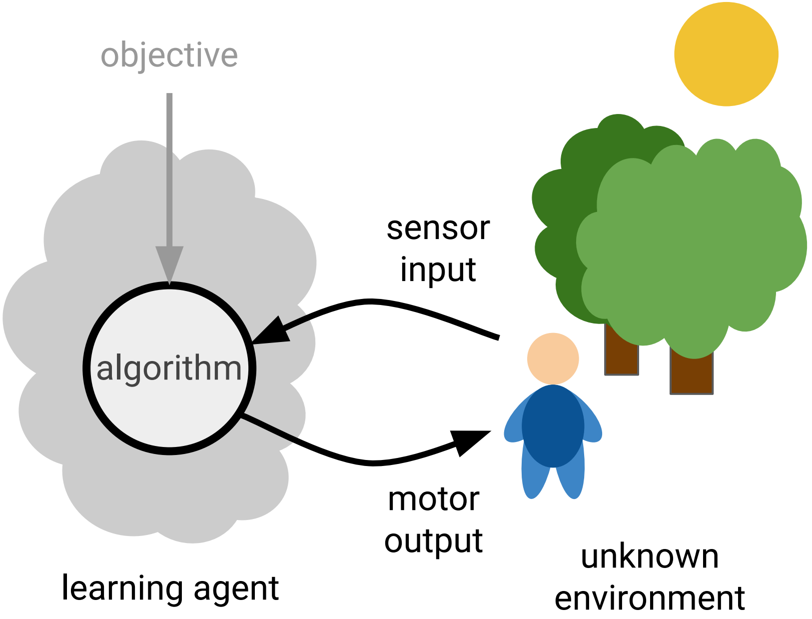
arxiv.org/abs/1810.12162

Presentation by Danijar Hafner

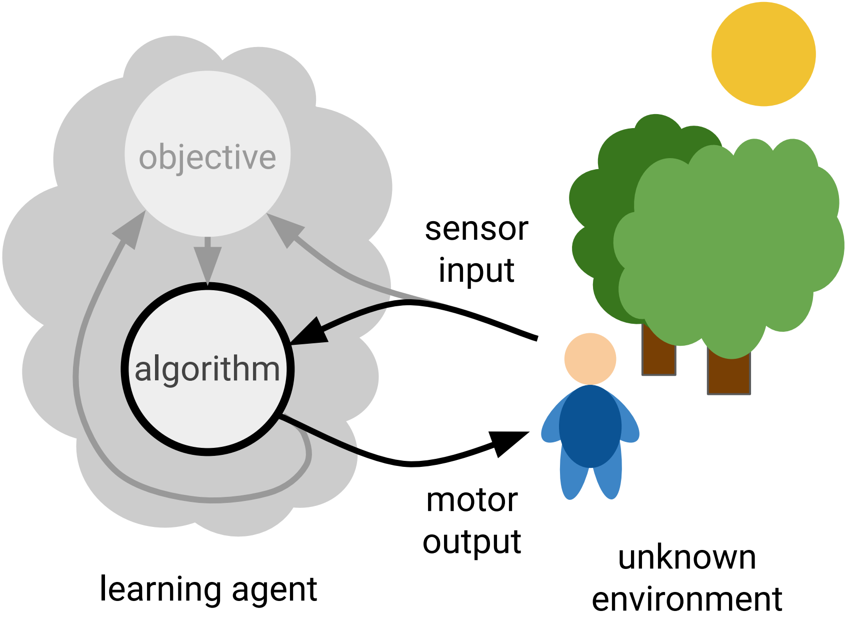
Reinforcement Learning



Reinforcement Learning



Intrinsic Motivation



Many Intrinsic Objectives

Information gain	e.g. Lindley 1956, Sun 2011, Houthooft 2017
Prediction error	e.g. Schmidhuber 1991, Bellemare 2016, Pathak 2017
Empowerment	e.g. Klyubin 2005, Tishby 2011, Gregor 2016
Skill discovery	e.g. Eysenbach 2018, Sharma 2020, Co-Reyes 2018
Surprise minimization	e.g. Schrödinger 1944, Friston 2013, Berseth 2020
Bayes-adaptive RL	e.g. Gittins 1979, Duff 2002, Ross 2007

Information Gain

Without rewards, the agent can only learn about the environment.

Information Gain

Without rewards, the agent can only learn about the environment.

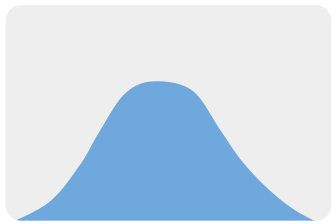
A model W represents our knowledge. E.g.: input density, forward prediction

Information Gain

Without rewards, the agent can only learn about the environment.

A model W represents our knowledge. E.g.: input density, forward prediction

Need to represent uncertainty about W to tell how much we have learned.



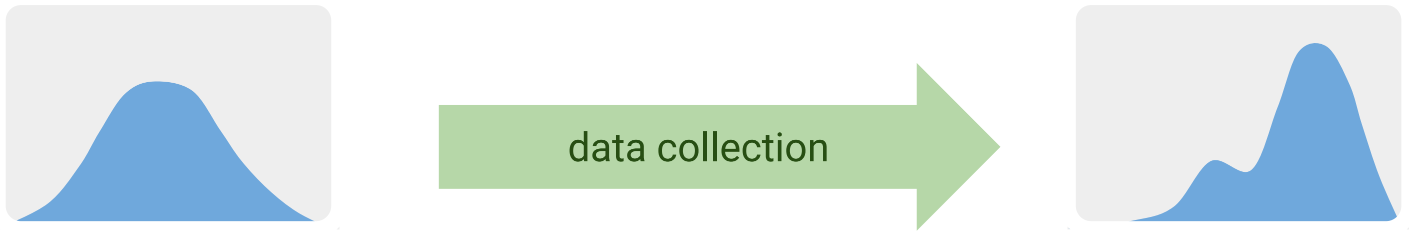
$p(W)$

Information Gain

Without rewards, the agent can only learn about the environment.

A model W represents our knowledge. E.g.: input density, forward prediction

Need to represent uncertainty about W to tell how much we have learned.



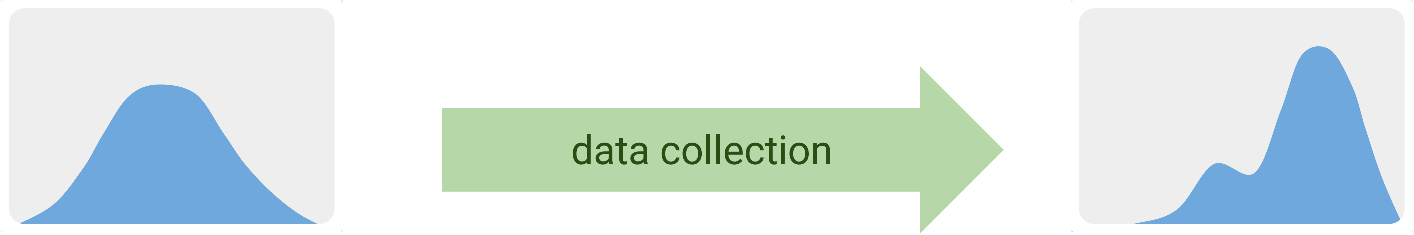
$p(W)$ $p(W \mid X)$

Information Gain

Without rewards, the agent can only learn about the environment.

A model W represents our knowledge. E.g.: input density, forward prediction

Need to represent uncertainty about W to tell how much we have learned.



$\mathrm{p}(W)$ $\qquad\qquad$ $\mathrm{p}(W \mid X)$

To gain the most information, we aim to maximize the mutual information between future sensory inputs X and model parameters W:

$$\max_a \mathrm{I}(X; W \mid A{=}a)$$

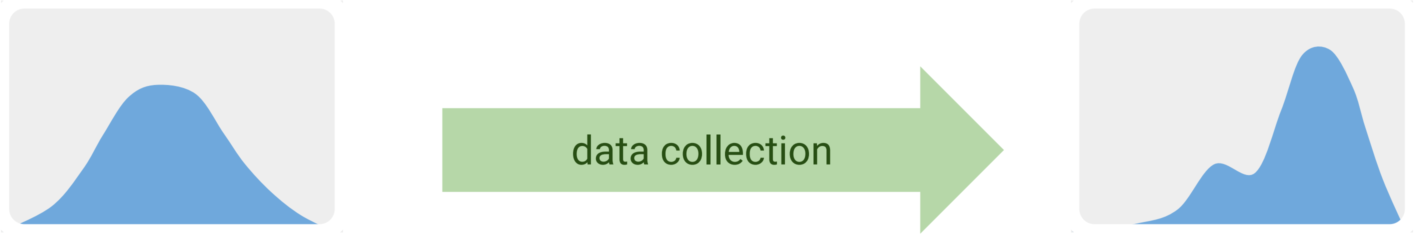
Both W and X are random variables

Information Gain

Without rewards, the agent can only learn about the environment.

A model W represents our knowledge. E.g.: input density, forward prediction

Need to represent uncertainty about W to tell how much we have learned.



$p(W)$ $\qquad\qquad$ $p(W \mid X)$

To gain the most information, we aim to maximize the mutual information between future sensory inputs X and model parameters W:

$$\max_a \mathrm{I}(X; W \mid A{=}a) = ?$$

Both W and X are random variables

Retrospective Infogain

e.g. VIME, ICM, RND

$$\mathrm{KL}[p(W \mid X, A=a) \,||\, p(W \mid A=a)]$$

Collect episodes, train world model, record improvement, reward the controller by this improvement

Infogain depends on agent's knowledge that keeps changing, making it a non-stationary objective

The learned controller will lag behind and go to states that were previously novel but are not anymore

Expected Infogain

e.g. MAX, PETS-ET, LD

$$\mathrm{I}(X; W \mid A=a)$$

Need to search for actions that will lead to high information gain without additional environment interaction

Learn a forward model of the environment to search for actions by planning or learning in imagination

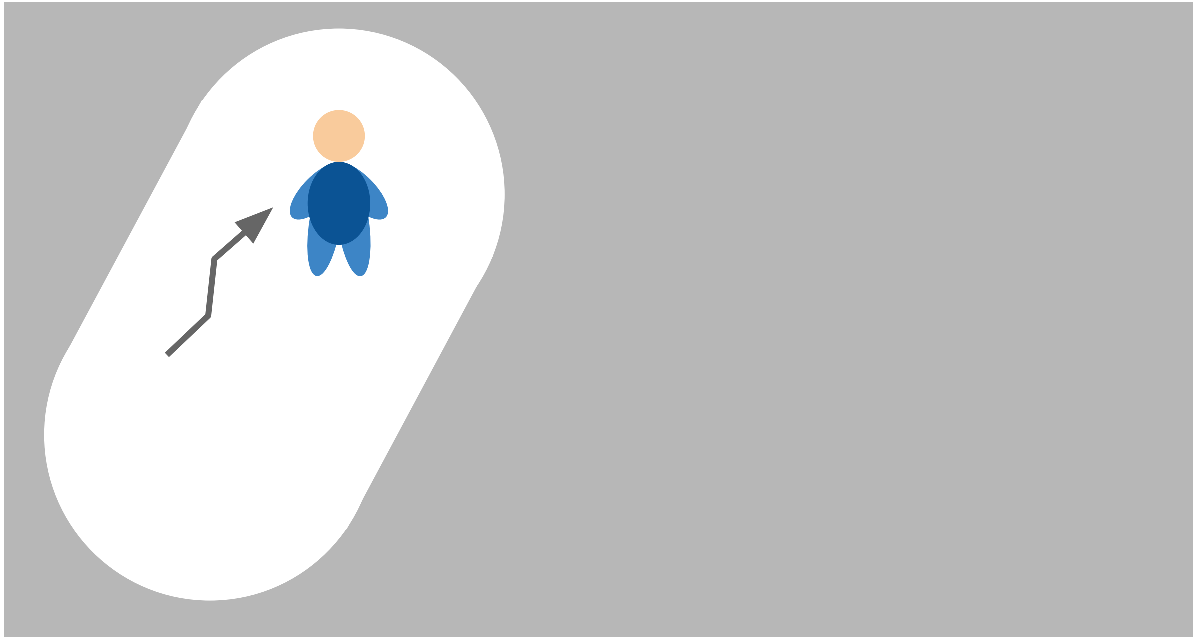
Computing the expected information gain requires computing entropies of a model with uncertainty estimates

Retrospective Novelty

Episode 1

Everything unknown

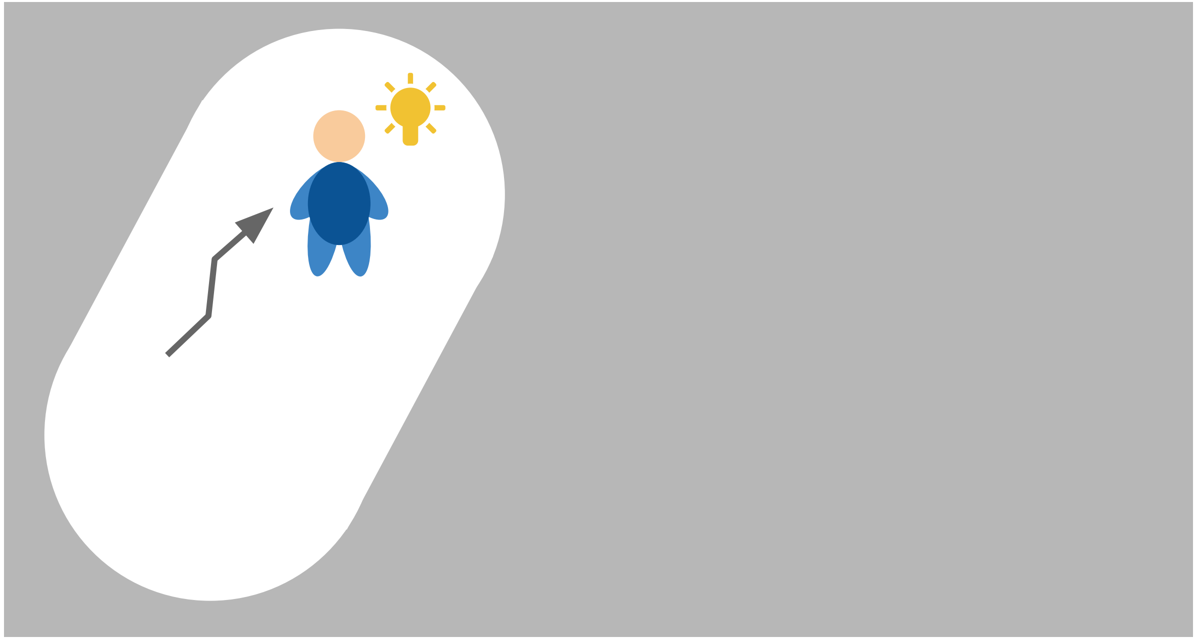
Retrospective Novelty



Episode 1

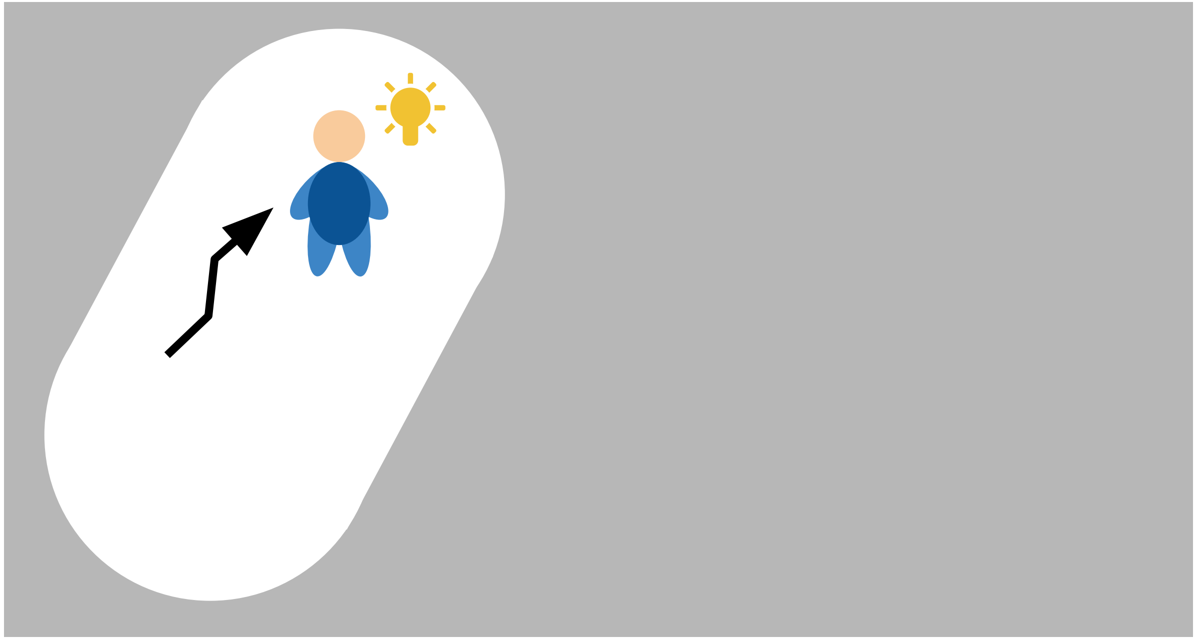
Random behavior

Retrospective Novelty



Episode 1 High novelty

Retrospective Novelty



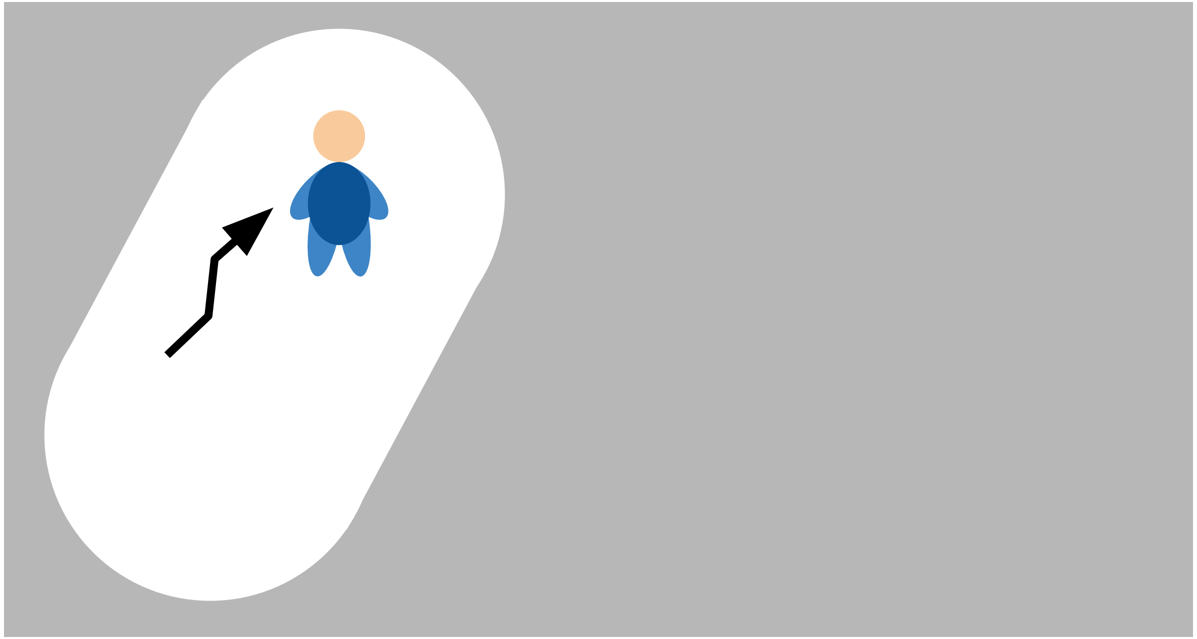
Episode 1

Reinforce behavior

Retrospective Novelty

Episode 2 Repeat behavior

Retrospective Novelty



Retrospective Novelty

Episode 2

Not surprising anymore :(

Retrospective Novelty

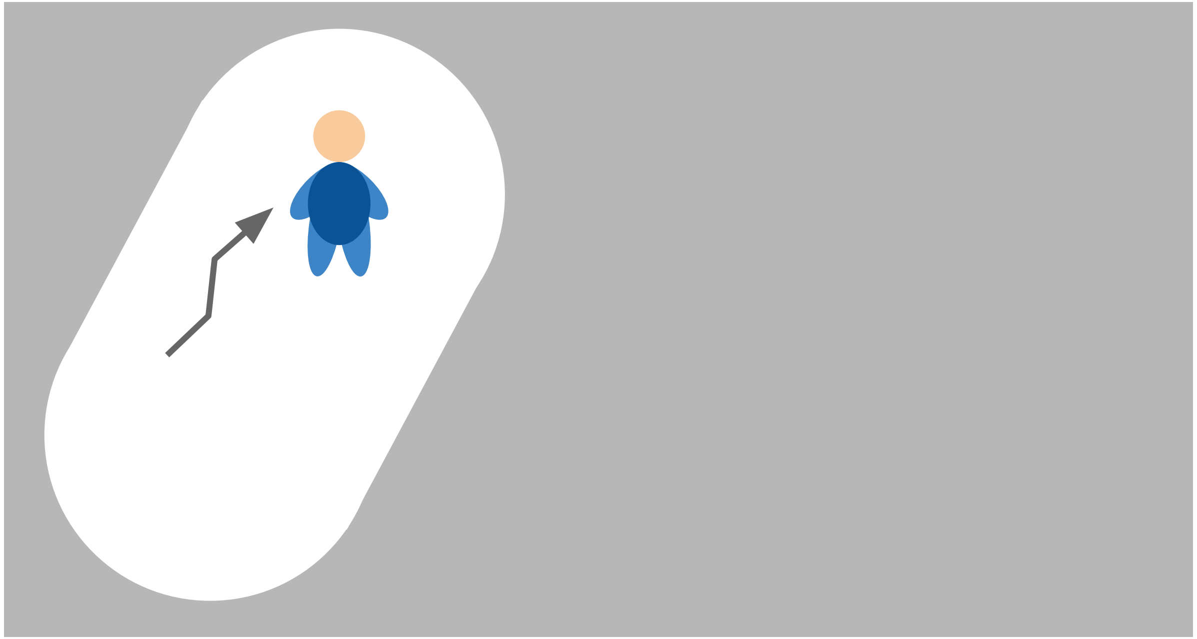
Episode 2

Unlearn behavior

Retrospective Novelty

Episode 3 Repeat behavior

Retrospective Novelty



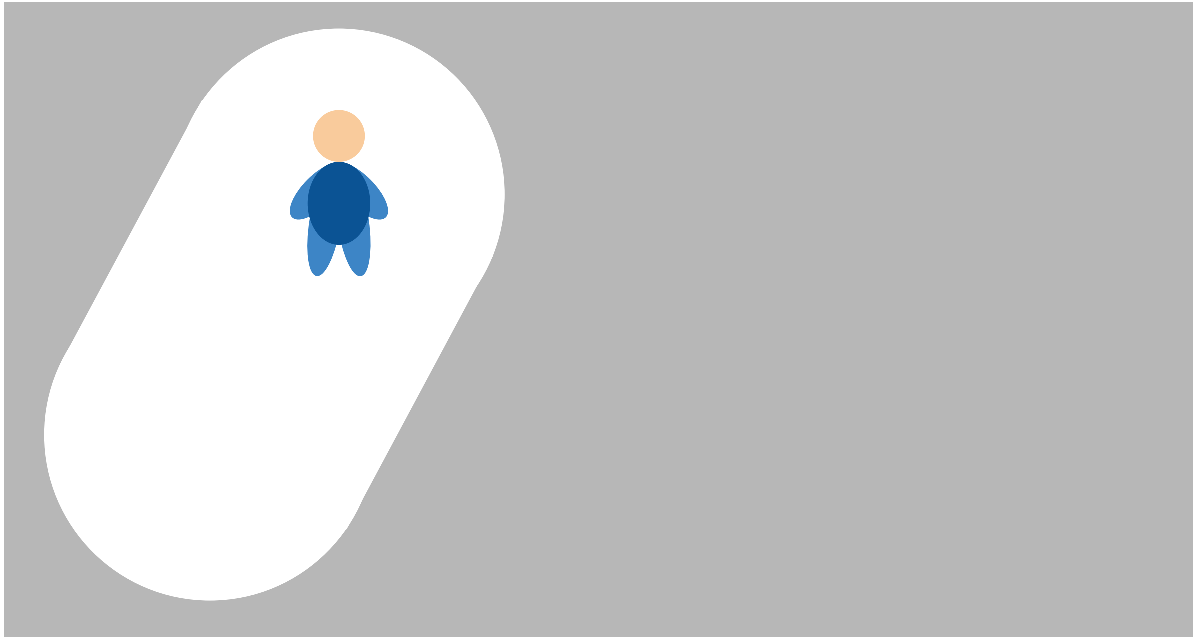
Episode 3

Repeat behavior

Retrospective Novelty

Episode 3 Still not novel

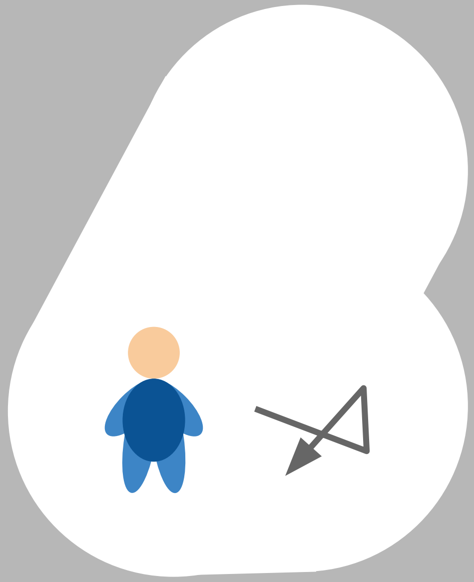
Retrospective Novelty



Episode 3

Unlearn behavior

Retrospective Novelty



The agent builds a map of where it was already and avoids those states.

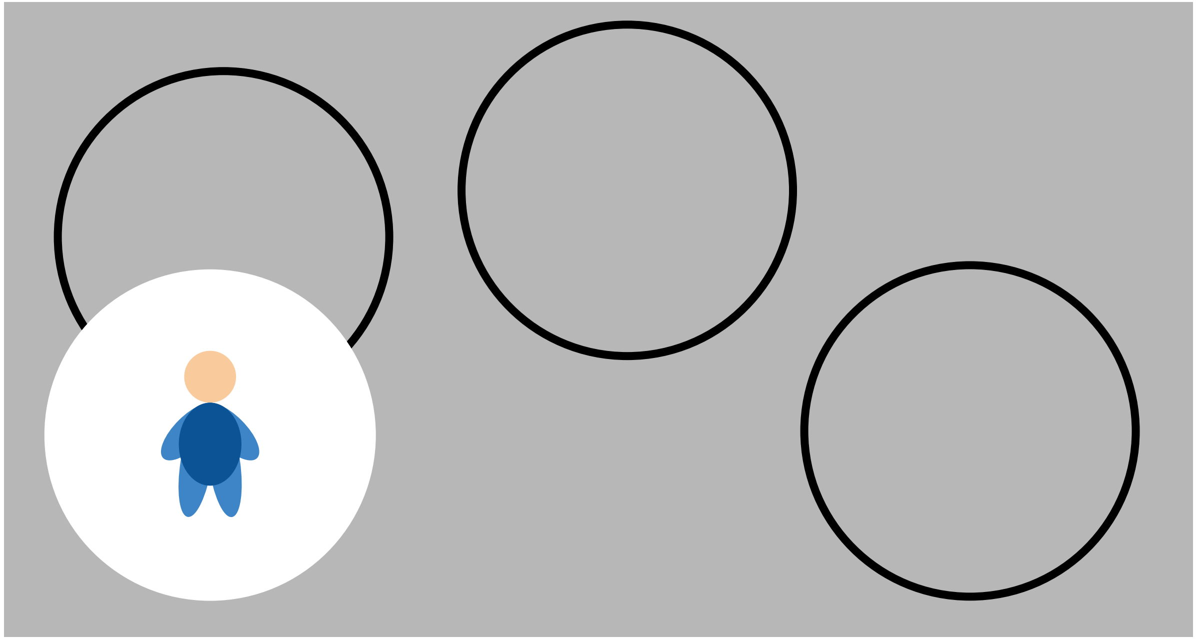
Episode 4

Back to random behavior

Expected Novelty

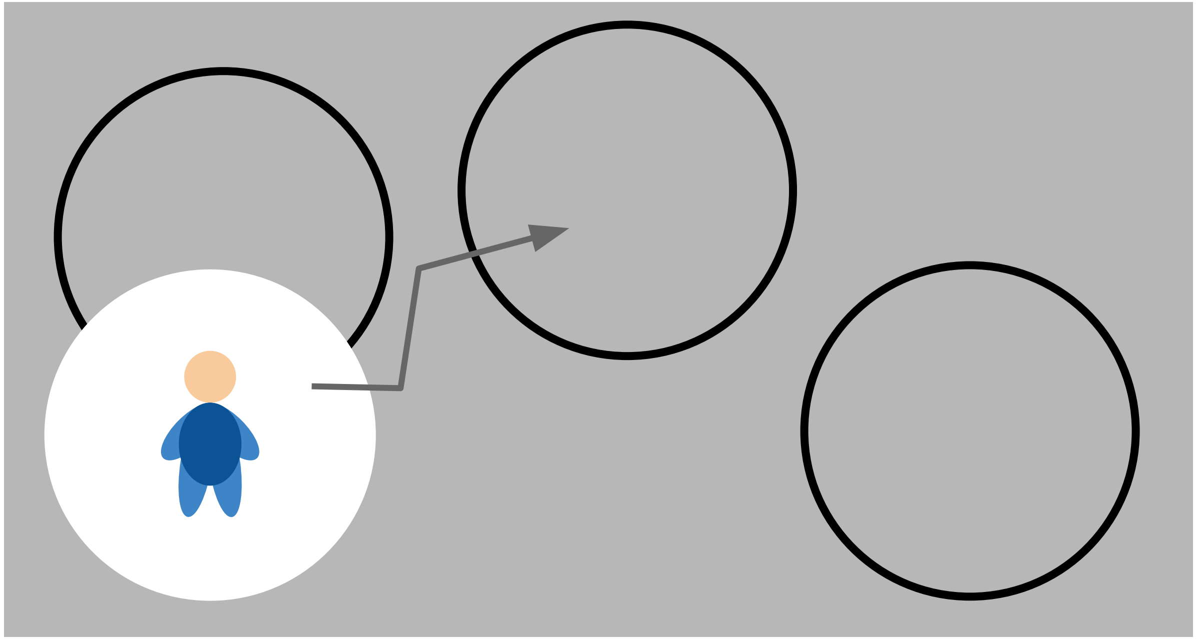
Episode 1 Everything unknown

Expected Novelty



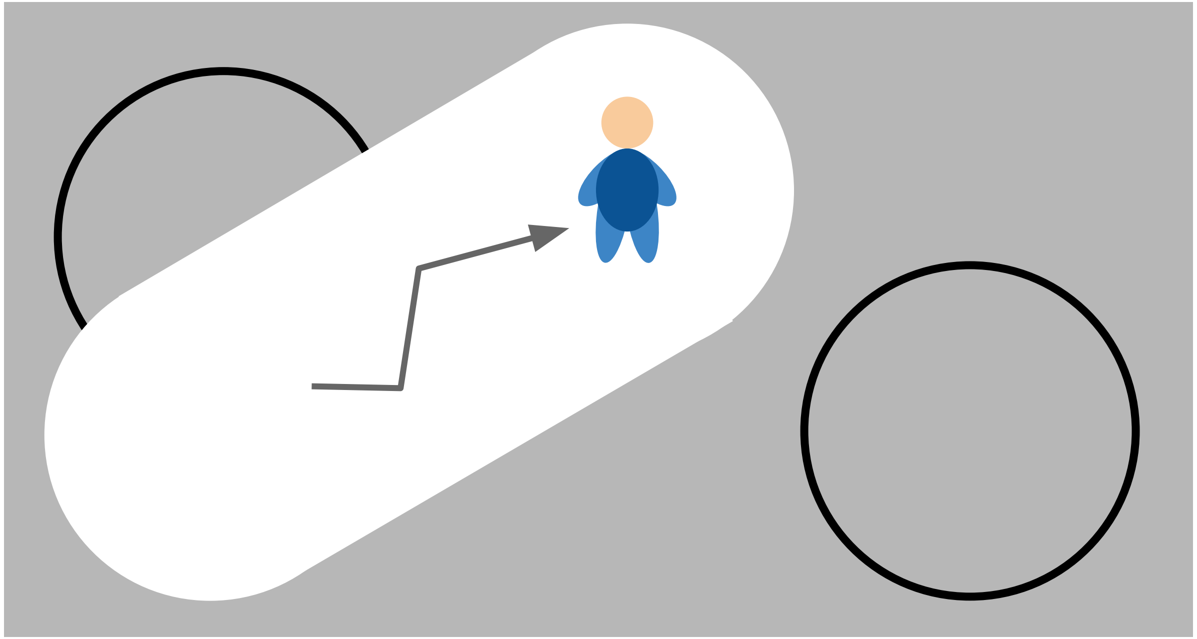
Episode 1 Consider options

Expected Novelty



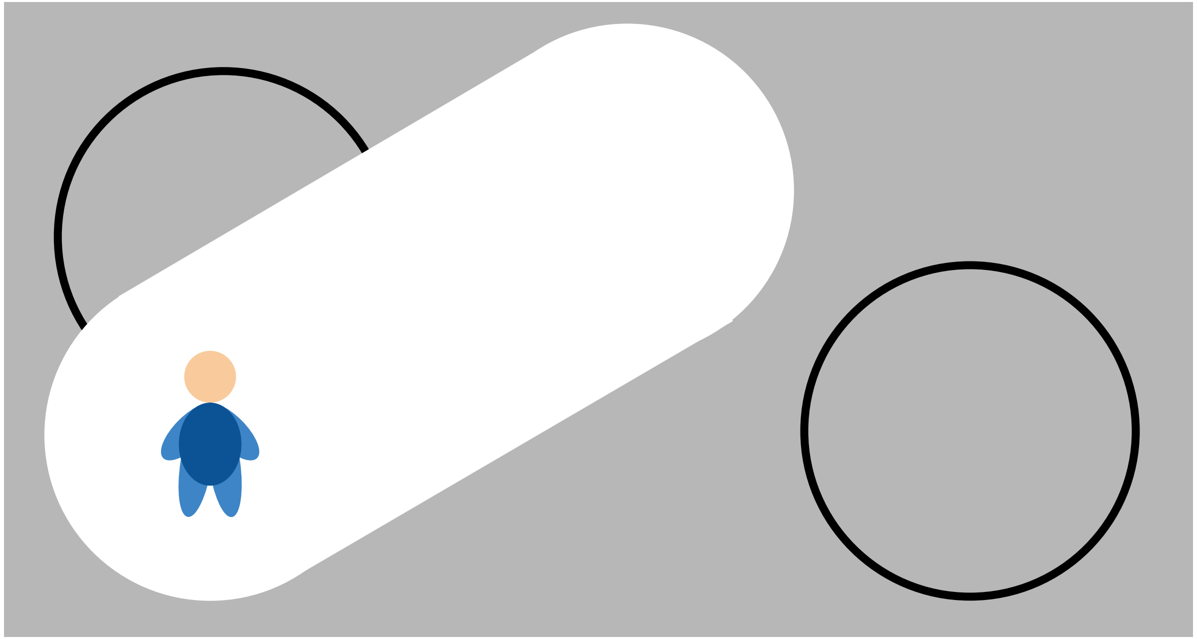
Episode 1 Execute plan

Expected Novelty



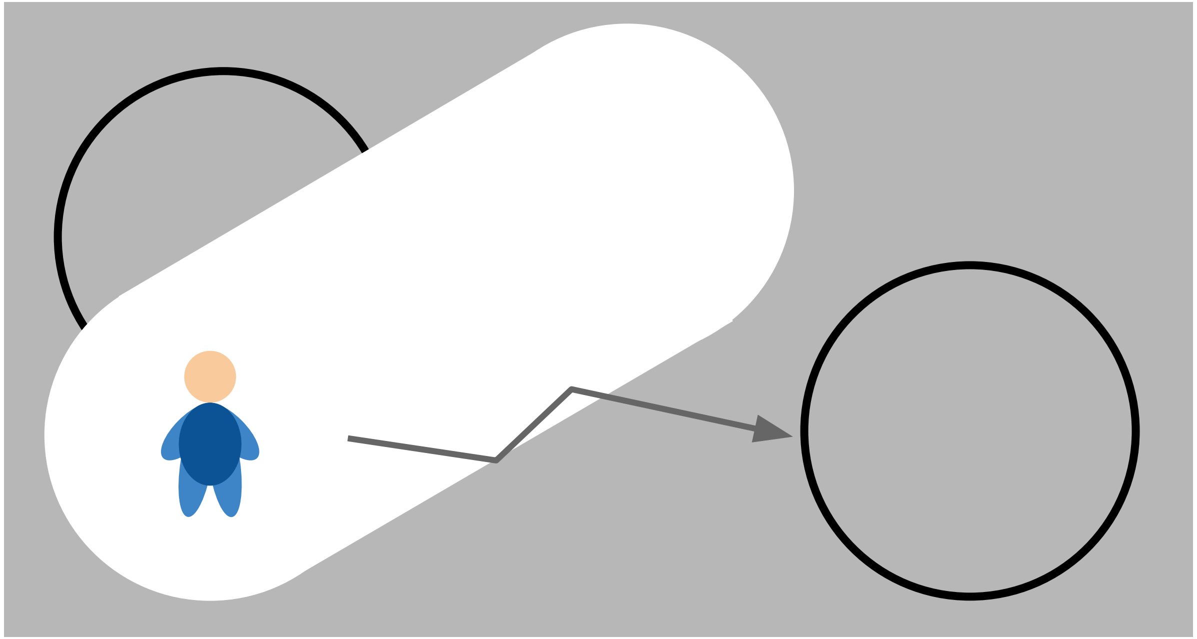
Episode 1 Observe new data

Expected Novelty



Episode 2 Consider options

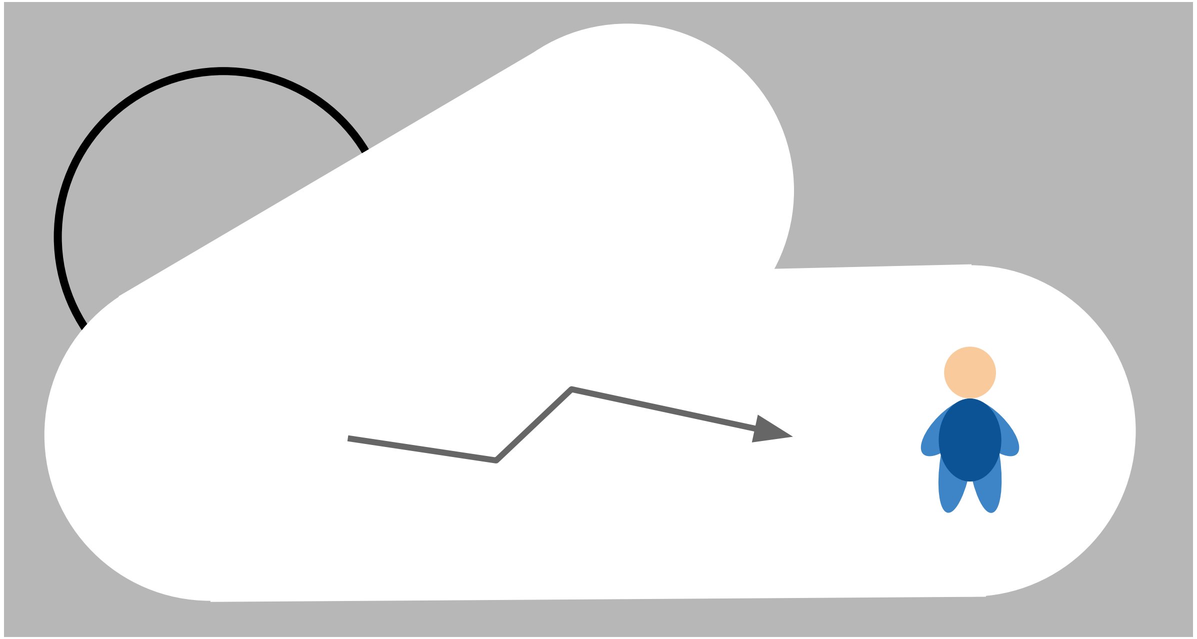
Expected Novelty



Episode 2

Execute plan

Expected Novelty



Episode 2

Observe new data

Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

Capture uncertainty as an ensemble of non-linear Gaussian predictors

Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

Capture uncertainty as an ensemble of non-linear Gaussian predictors

$$\underbrace{\mathrm{I}(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{\mathrm{H}(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{\mathrm{H}(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

Capture uncertainty as an ensemble of non-linear Gaussian predictors

$$\underbrace{\mathrm{I}(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{\mathrm{H}(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{\mathrm{H}(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Information gain targets uncertain trajectories with low expected noise

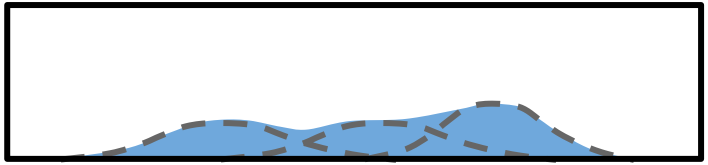
Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

Capture uncertainty as an ensemble of non-linear Gaussian predictors

$$\underbrace{\mathrm{I}(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{\mathrm{H}(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{\mathrm{H}(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Information gain targets uncertain trajectories with low expected noise



Wide predictions mean high expected noise
Overlapping modes means less total uncertainty

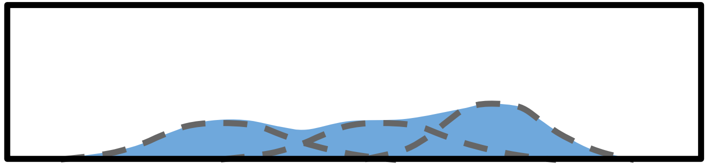
Ensemble of Dynamics Models

Learn dynamics both to represent knowledge and to plan for expected infogain

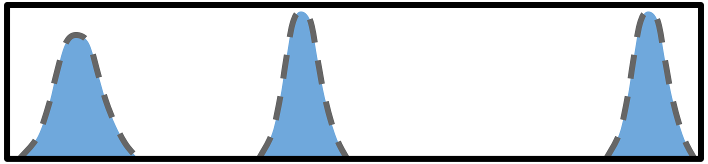
Capture uncertainty as an ensemble of non-linear Gaussian predictors

$$\underbrace{\mathrm{I}(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{\mathrm{H}(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{\mathrm{H}(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Information gain targets uncertain trajectories with low expected noise



Wide predictions mean high expected noise
Overlapping modes means less total uncertainty



Narrow predictions mean low expected noise
Distant modes means large total uncertainty

Expected Infogain Approximation

$$\underbrace{\mathrm{I}(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{\mathrm{H}(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{\mathrm{H}(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Expected Infogain Approximation

$$\mathrm{I}(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty total uncertainty aleatoric uncertainty

Ensemble members: $\mathrm{p}(\mathrm{X} \mid \mathrm{W}{=}\mathrm{w}_k, \mathrm{A}{=}\mathrm{a})$

Expected Infogain Approximation

$$\mathrm{I}(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty — total uncertainty — aleatoric uncertainty

Ensemble members: $\mathrm{p}(\mathrm{X} \mid \mathrm{W{=}w_k, A{=}a})$

Aggregate prediction: $\mathrm{p}(\mathrm{X} \mid \mathrm{A{=}a}) = 1/\mathrm{K} \, \Sigma \, \mathrm{p}(\mathrm{X} \mid \mathrm{W{=}w_k, A{=}a})$

Expected Infogain Approximation

$$I(X; W \mid A{=}a) = H(X \mid A{=}a) - H(X \mid W, A{=}a)$$

epistemic uncertainty   total uncertainty   aleatoric uncertainty

Ensemble members: $p(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $p(X \mid A{=}a) = 1/K \, \Sigma \, p(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty:

Expected Infogain Approximation

$$\mathrm{I}(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty total uncertainty aleatoric uncertainty

Ensemble members: $p(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $p(X \mid A{=}a) = 1/K \, \Sigma \, p(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty: ?

Expected Infogain Approximation

$$\mathrm{I}(X;\, W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty   total uncertainty   aleatoric uncertainty

Ensemble members: $\mathrm{p}(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $\mathrm{p}(X \mid A{=}a) = 1/K\, \Sigma\, \mathrm{p}(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty: $1/K\, \Sigma\, \mathrm{H}(\mathrm{p}(X \mid W{=}w_k, A{=}a))$

Expected Infogain Approximation

$$\underbrace{I(X; W \mid A{=}a)}_{\text{epistemic uncertainty}} = \underbrace{H(X \mid A{=}a)}_{\text{total uncertainty}} - \underbrace{H(X \mid W, A{=}a)}_{\text{aleatoric uncertainty}}$$

Ensemble members: $p(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $p(X \mid A{=}a) = 1/K \, \Sigma \, p(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty: $1/K \, \Sigma \, H(p(X \mid W{=}w_k, A{=}a))$

Total uncertainty:

Expected Infogain Approximation

$$\mathrm{I}(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty   total uncertainty   aleatoric uncertainty

Ensemble members: $\mathrm{p}(\mathrm{X} \mid \mathrm{W}{=}\mathrm{w}_k, \mathrm{A}{=}\mathrm{a})$

Aggregate prediction: $\mathrm{p}(\mathrm{X} \mid \mathrm{A}{=}\mathrm{a}) = 1/\mathrm{K}\, \Sigma\, \mathrm{p}(\mathrm{X} \mid \mathrm{W}{=}\mathrm{w}_k, \mathrm{A}{=}\mathrm{a})$

Aleatoric uncertainty: $1/\mathrm{K}\, \Sigma\, \mathrm{H}(\mathrm{p}(\mathrm{X} \mid \mathrm{W}{=}\mathrm{w}_k, \mathrm{A}{=}\mathrm{a}))$

Total uncertainty: ?

Expected Infogain Approximation

$$I(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty — total uncertainty — aleatoric uncertainty

Ensemble members: $p(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $p(X \mid A{=}a) = 1/K\, \Sigma\, p(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty: $1/K\, \Sigma\, H(p(X \mid W{=}w_k, A{=}a))$

Total uncertainty: $H(1/K\, \Sigma\, p(X \mid W{=}w_k, A{=}a))$

Expected Infogain Approximation

$$\mathrm{I}(X; W \mid A{=}a) = \mathrm{H}(X \mid A{=}a) - \mathrm{H}(X \mid W, A{=}a)$$

epistemic uncertainty total uncertainty aleatoric uncertainty

Ensemble members: $p(X \mid W{=}w_k, A{=}a)$

Aggregate prediction: $p(X \mid A{=}a) = 1/K \, \Sigma \, p(X \mid W{=}w_k, A{=}a)$

Aleatoric uncertainty: $1/K \, \Sigma \, H(p(X \mid W{=}w_k, A{=}a))$

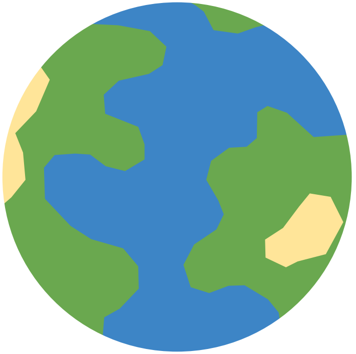
Total uncertainty: $H(1/K \, \Sigma \, p(X \mid W{=}w_k, A{=}a))$

Gaussian entropy has a closed form, so we can compute the aleatoric uncertainty.
GMM entropy does not, sample it or switch to Renyi entropy that has a closed form.

Algorithm 1 MODEL-BASED ACTIVE EXPLORATION

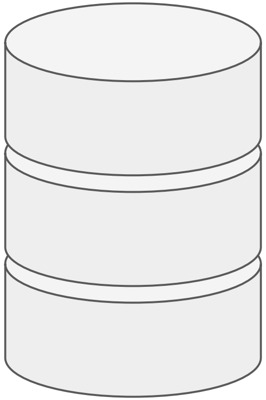
Initialize: Transitions dataset D, with random policy
Initialize: Model ensemble, $\tilde{T} = \{t_1, t_2, \cdots, t_N\}$
repeat
  while episode not complete **do**
    ExplorationMDP $\leftarrow (\mathcal{S}, \mathcal{A}, \text{Uniform}\{\tilde{T}\}, u, \delta(s_\tau))$
    $\pi \leftarrow$ SOLVE(ExplorationMDP)
    $a_\tau \sim \pi(s_\tau)$
    act in environment: $s_{\tau+1} \sim \mathcal{P}(\mathcal{S}|s_\tau, a_\tau, t^*)$
    $D \leftarrow D \cup \{(s_\tau, a_\tau, s_{\tau+1})\}$
    Train t_i on D for each t_i **in** $\tilde{T}$
  end while
until computation budget exhausted

Compared Algorithms



Learning from imagined trajectories (Expected)

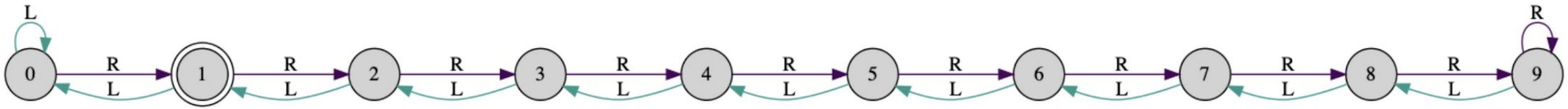
MAX: JSD infogain
TVAX: State variance



Learning from experience replay (Retrospective)

JDRX: JSD infogain
PERX: Prediction error

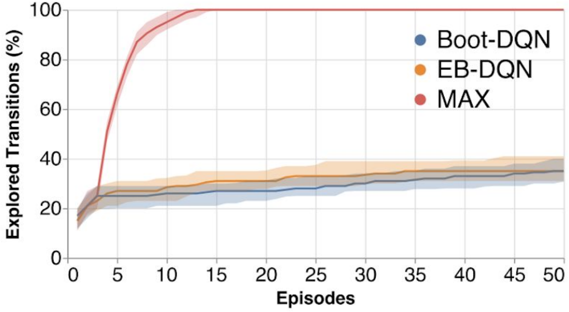
Exploration Chain Domain



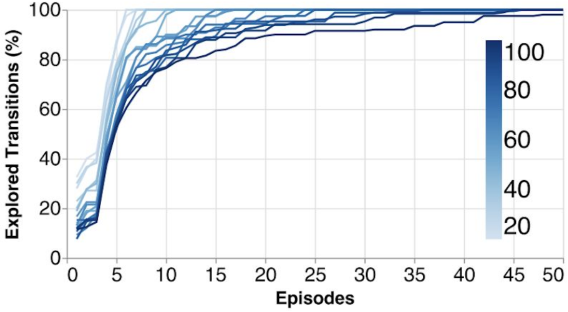
+0.001

(a) Chain Environment of length 10.

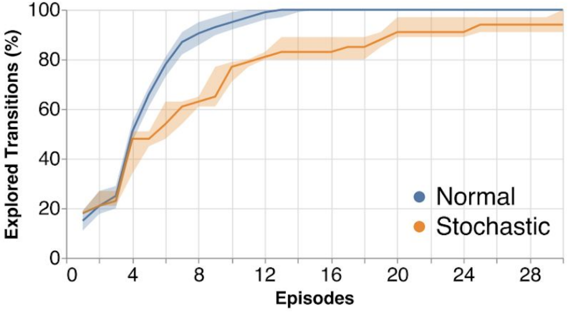
+1



(b) 50-state chain.

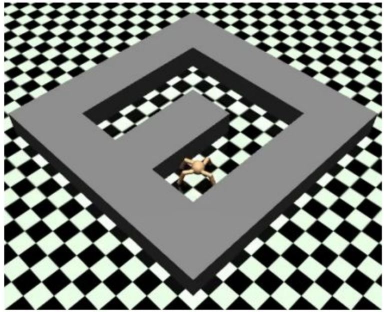


(c) Chain lengths

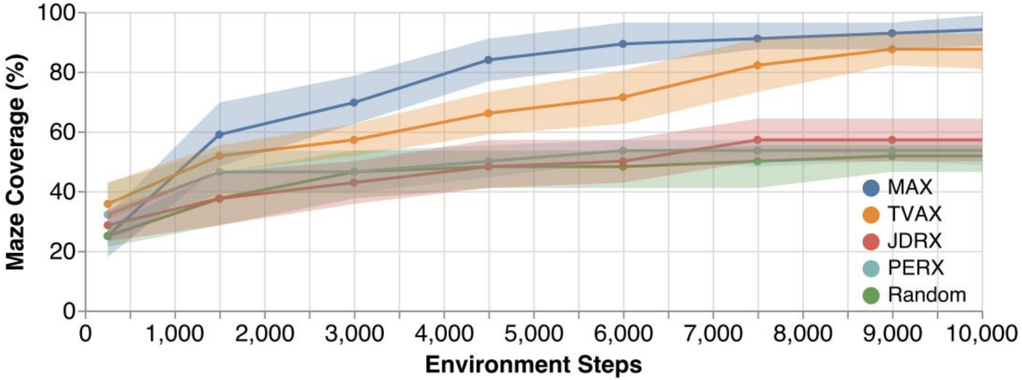


(d) Stochastic trap

State coverage of Ant Maze



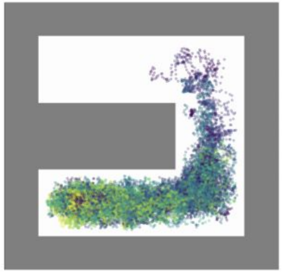
(a) Ant Maze Environment



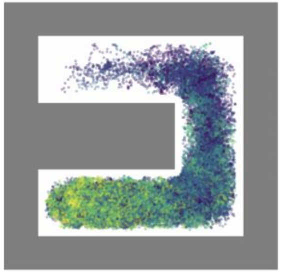
(b) Maze Exploration Performance

(c) 300 steps

(d) 600 steps



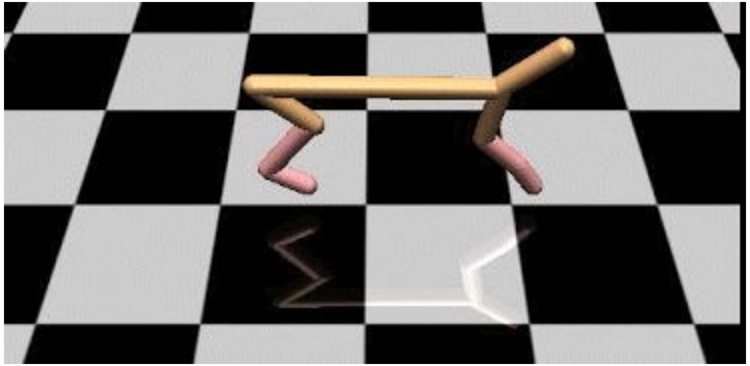
(e) 3600 steps



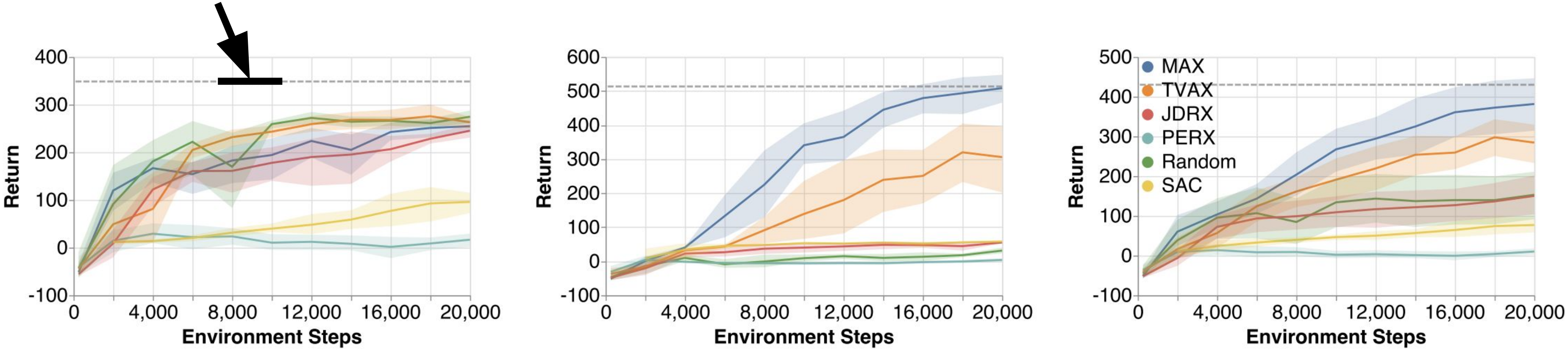
(f) 12000 steps

Zero-Shot Adaptation

Learn evaluation policy inside of learned model given a known reward function



model-free with 10x data



(a) Running task performance

(b) Flipping task performance

(c) Average performance

no exploration needed

exploration needed

Conclusions

Information gain is a principled task-agnostic objective

As a non-stationary objective, it should be optimized in expectation

This requires a dynamics model for planning to explore

Ensemble of Gaussian dynamics is a practical way to represent uncertainty