Intrinsic Motivation in Reinforcement Learning

to guide exploration and task-agnostic learning

Georg Martius

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

AUTONOMOUS LEARNING
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Vision: Versatile Learning Robots

Imagine we had robots that can be trained to perform new tasks quickly and that become dexterous…

Valuable assistants for humans in:

- collaborative assembly
- planting trees
- care
- sustainable agriculture
- …

[Kuka]

(frauenhofer ipa)

[bergkvistanna karin@tuvie.com]

Need Learning!

[Polybot.eu]

Developmental Learning

What are the generic driving principles?

Georg Martius <georg.martius@tue.mpg.de>

Intrinsic motivation

gain sensorimotor coordination

information theory and dynamical systems-based intrinsic motivation

gaining understanding

surprise based motivation, predicted information gain in unsupervised reinforcement learning

gaining control of environment and learn skills

competence-based methods in hierarchical reinforcement learning

visit particular states

empowered, causally impactful, or regular situations

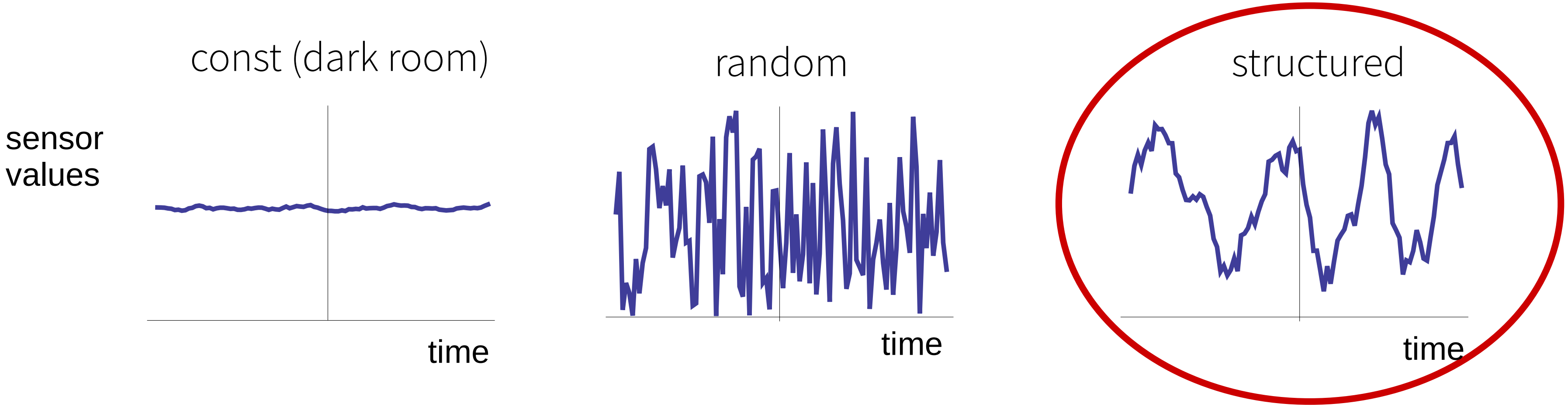
Task-agnostic learning (not comprehensive)

- Sensorimotor coordination – Dynamical balance
  Homeokinesis (Der 2001, Der & Martius, 2011)
  Predictive Information maximization (Martius, Der, Ay, 2013)
  Differential extrinsic plasticity (DEP) (Martius, Der 2015)
  DEP-RL (Schumacher 2023)
- Curiosity, Prediction Error, Surprise (Schmidhuber 1991-, Pathak 2017)
- Free Energy principle (Friston 2006 -)
- Predicted Information Gain (Sommer & Little 2012)
  Reduction of Epistemic Uncertainty (Pathak 2019+, Vlastelica 2021, Sancaktar 2022)
- Learning progress, competence
  (Schmidhuber 1991-, Oudeyer 2005-, Baldassarre 2007-,
  Blaes 2019, Colas 2019-)
- Skill Diversity (Eysenbach 2018, Gumbsch 2018-2023, Vlastelica @EWRL)
- Adversarial selfplay (OpenAI, Plappert et al 2021)
- Empowerment (Polani et al 2005-)
  Causal action influence (Seitzer et al 2021)
- Regularity (Sancaktar @EWRL)

Georg Martius <georg.martius@tue.mpg.de>

Principles of early sensorimotor coordination?

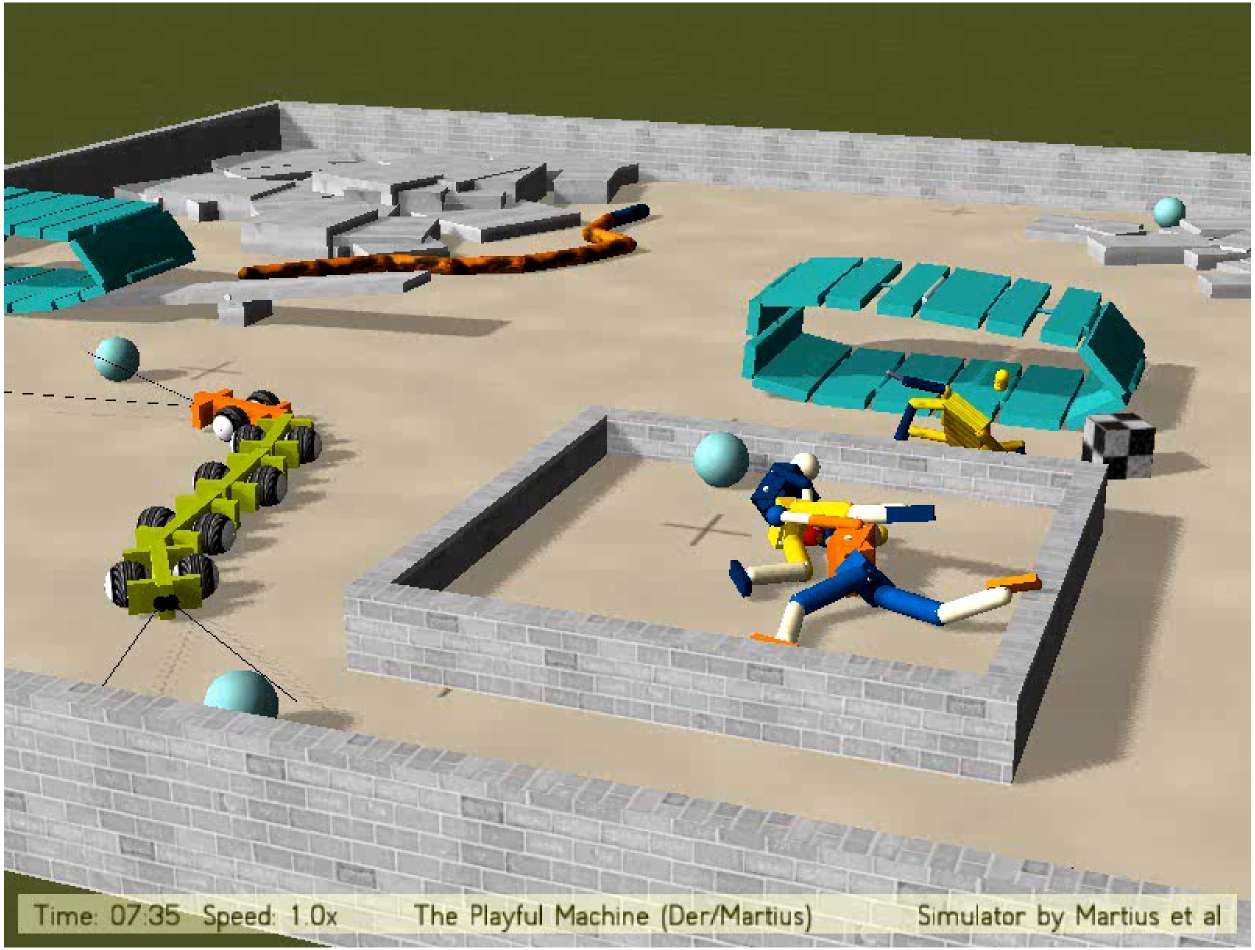
- general principle should avoid trivial solutions



- Dynamical Systems: **no trivial fixed points, balanced dynamics** (critical)
- Information Theory: Predictive information (PI) (Mutual Information between past and future)

Georg Martius <georg.martius@tue.mpg.de>

Self-organizing behavior



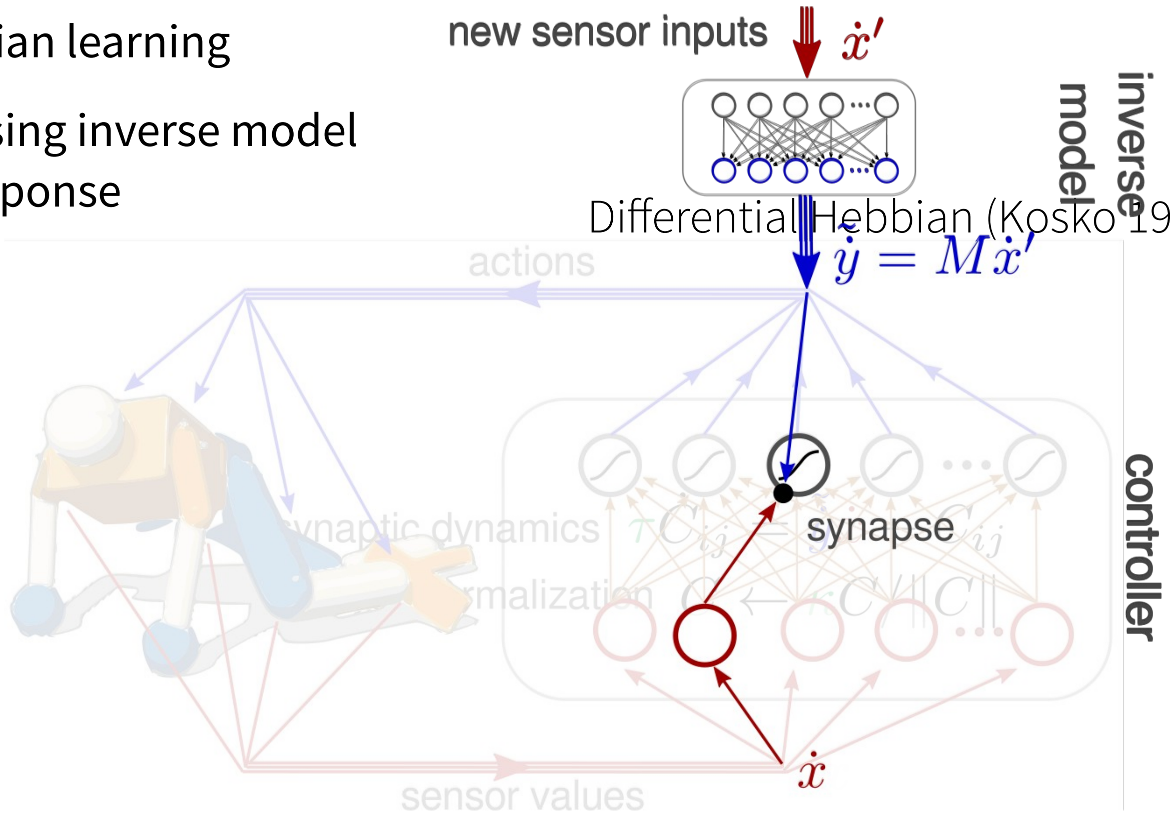
Der, GM. *The Playful Machine*, 2012

Differential Extrinsic Plasticity

intrinsic motivation to create coordinated behavior

Generalization of
differential Hebbian learning

- new term: using inverse model of sensor response



Differential Hebbian (Kosko 1986)

Der, GM, *PNAS* 2015

Georg Martius <georg.martius@tue.mpg.de>

Dynamical self-consistency

Controller: one-layer network

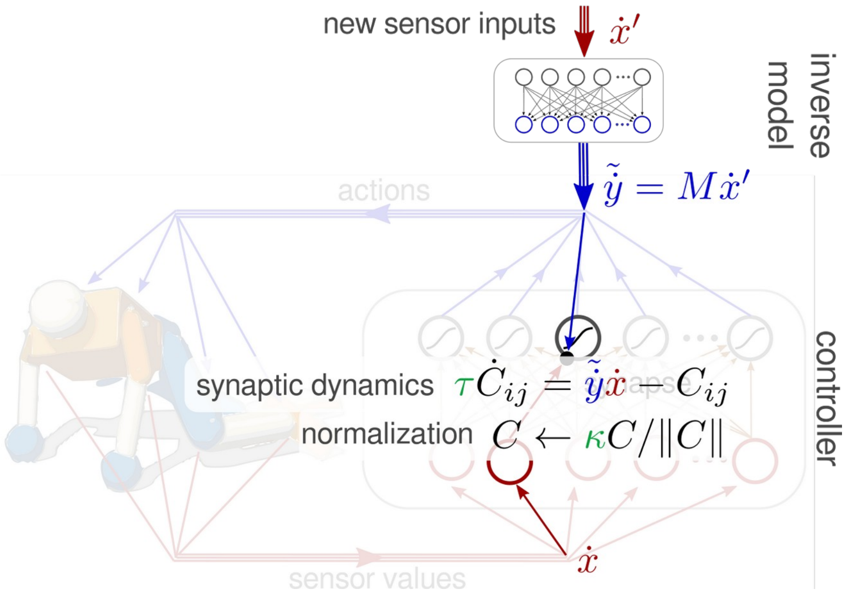
$$y = \tanh(\hat{C}x + h)$$

Weight normalization

$$\hat{C}_{ij} = \kappa \frac{C_{ij}}{\|C_i\|}$$

Inverse Model $(M = \mathbb{I})$

$$F(\dot{x}) = M\dot{x}$$



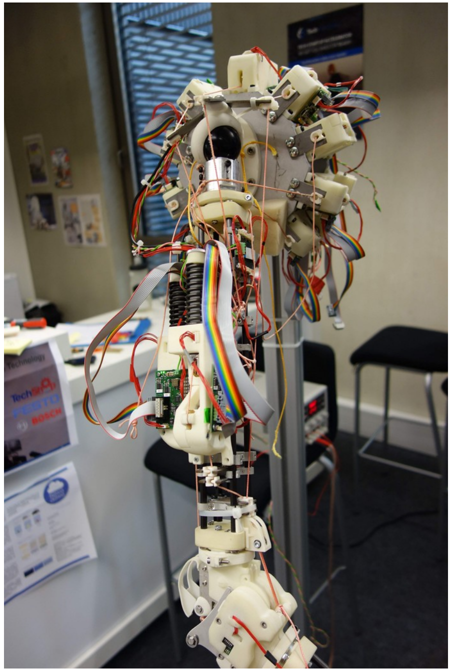
What does it do?

- amplify small movements (κ)
- increase velocity correlations $C \approx \sum_{s=t-d}^{t} \tilde{\dot{y}}_s \dot{x}_s^\top$
- aims for self-consistency

Behavior generated by C reproduces C by the dynamics

Der, GM, *PNAS* 2015

Soft-robot humanoid arm



Robot: Myorobotics arm, TUM

9 muscles for
shoulder and elbow

bottle shaking

GM, Hostettler, Knoll, Der. *IROS* 2017

Exploration is key

Standard noise exploration:

works for torque driven systems

Pierre Schumacher

Isabelle Wochner

Daniel Häufle

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

Exploration is key

Standard noise exploration:

2 DoF
6 muscles

fails for over-actuated systems

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

Exploration is key

Standard noise exploration:

fails for over-actuated systems

2 DoF
6 muscles

Embodied exploration:

DEP: like a Hebbian learning rule:
creates coordinated behavior

[Der, Martius. *PNAS* 2015]

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

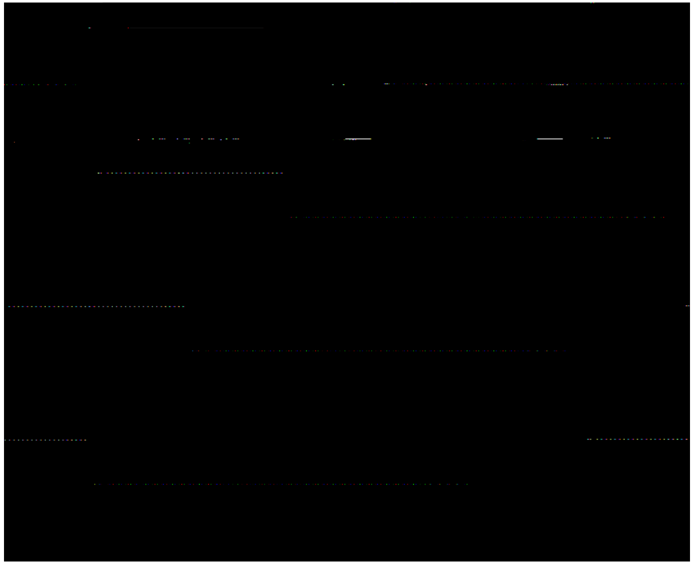
Exploration is key

Standard noise exploration:

fails for over-actuated systems

48 DoF
121 muscles

Embodied exploration:



DEP: like a Hebbian learning rule:
creates coordinated behavior

[Der, Martius. *PNAS* 2015]

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

Georg Martius <georg.martius@tue.mpg.de>

Reinforcement Learning with Embodied Exploration

- DEP-RL: Interleave exploration and policy optimization at random times
- Ostrich

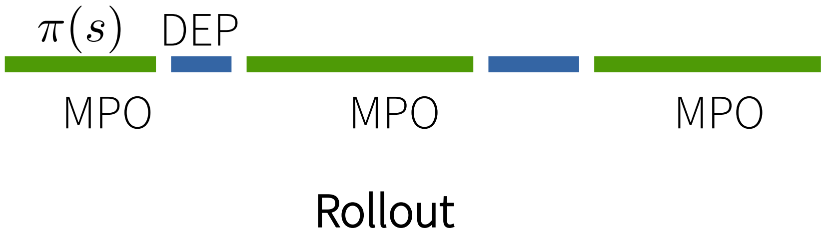
DEP-RL

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

Georg Martius <georg.martius@tue.mpg.de>

Reinforcement Learning with Embodied Exploration

- DEP-RL: Interleave exploration and policy optimization at random times



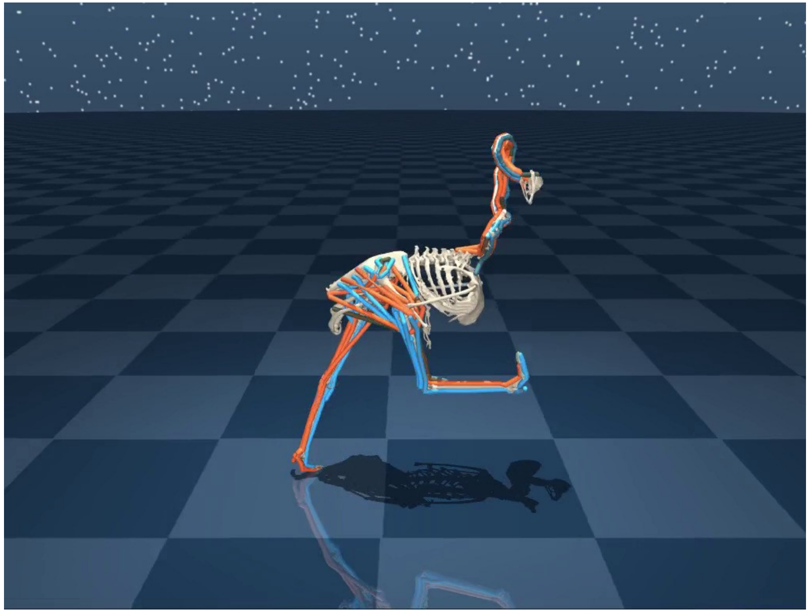
MPO: Maximum a Posteriori Policy Optimisation. Abdolmaleki et al, ICLR 2018

Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

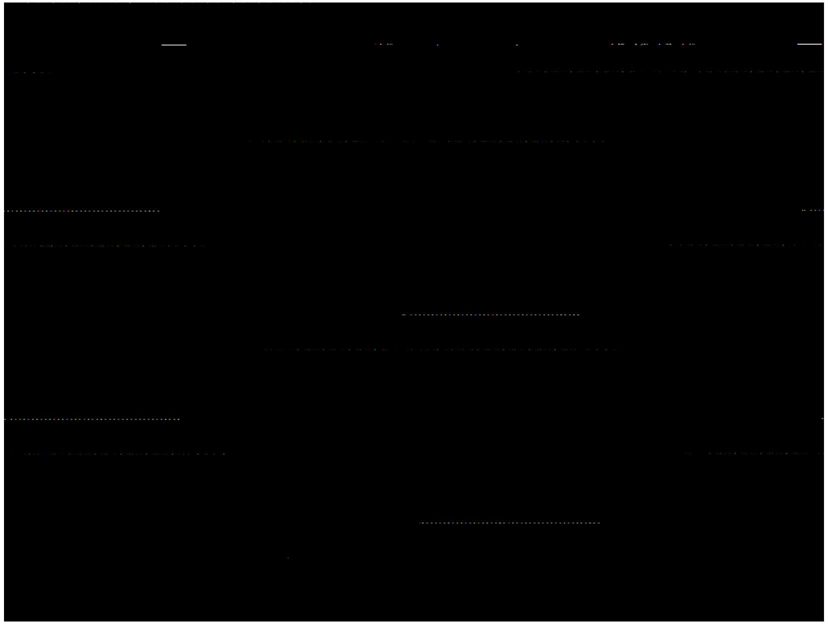
Let it run

- Exploration is key: DEP-RL: Interleave exploration and policy optimization at random times
- Ostrich

DEP-RL



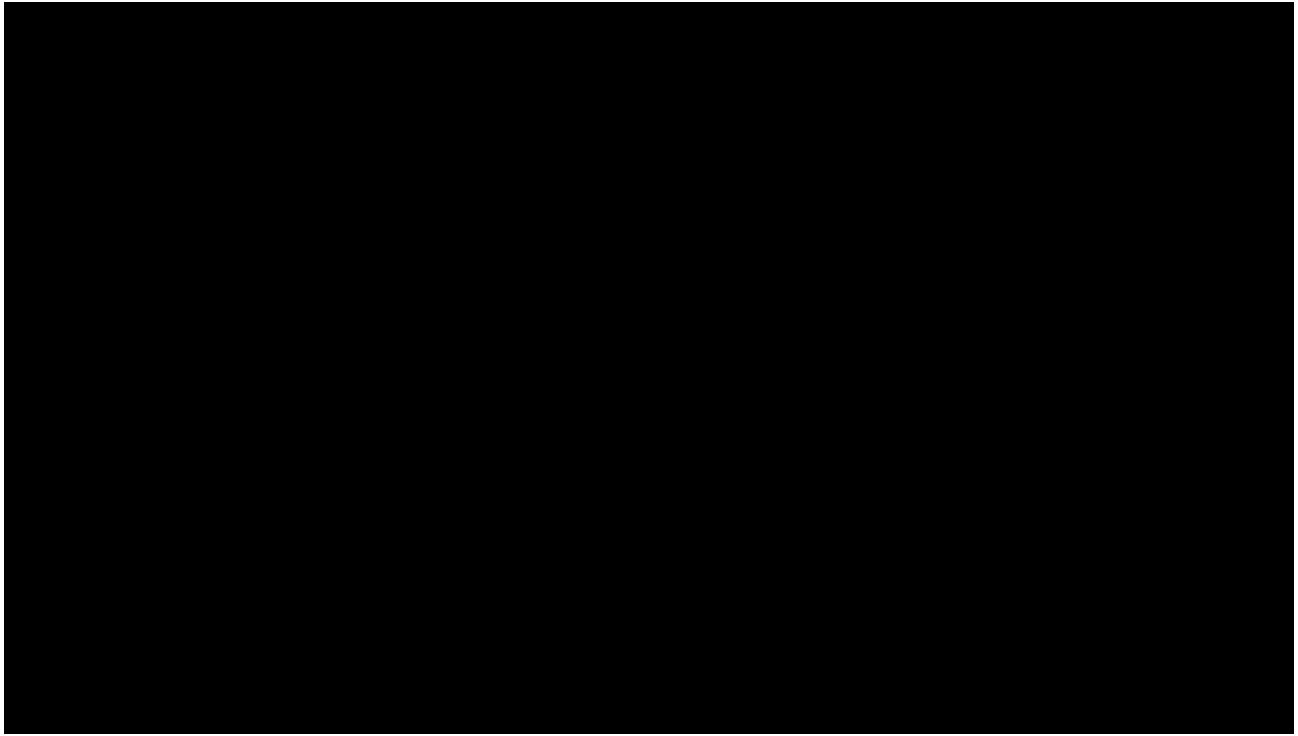
With normal exploration



Schuhmacher, Häufle, Büchler, Schmitt, GM. *ICLR*, 2023

Let it run

- Exploration is key: DEP-RL: Interleave exploration and policy optimization at random times
- What about models of humans?



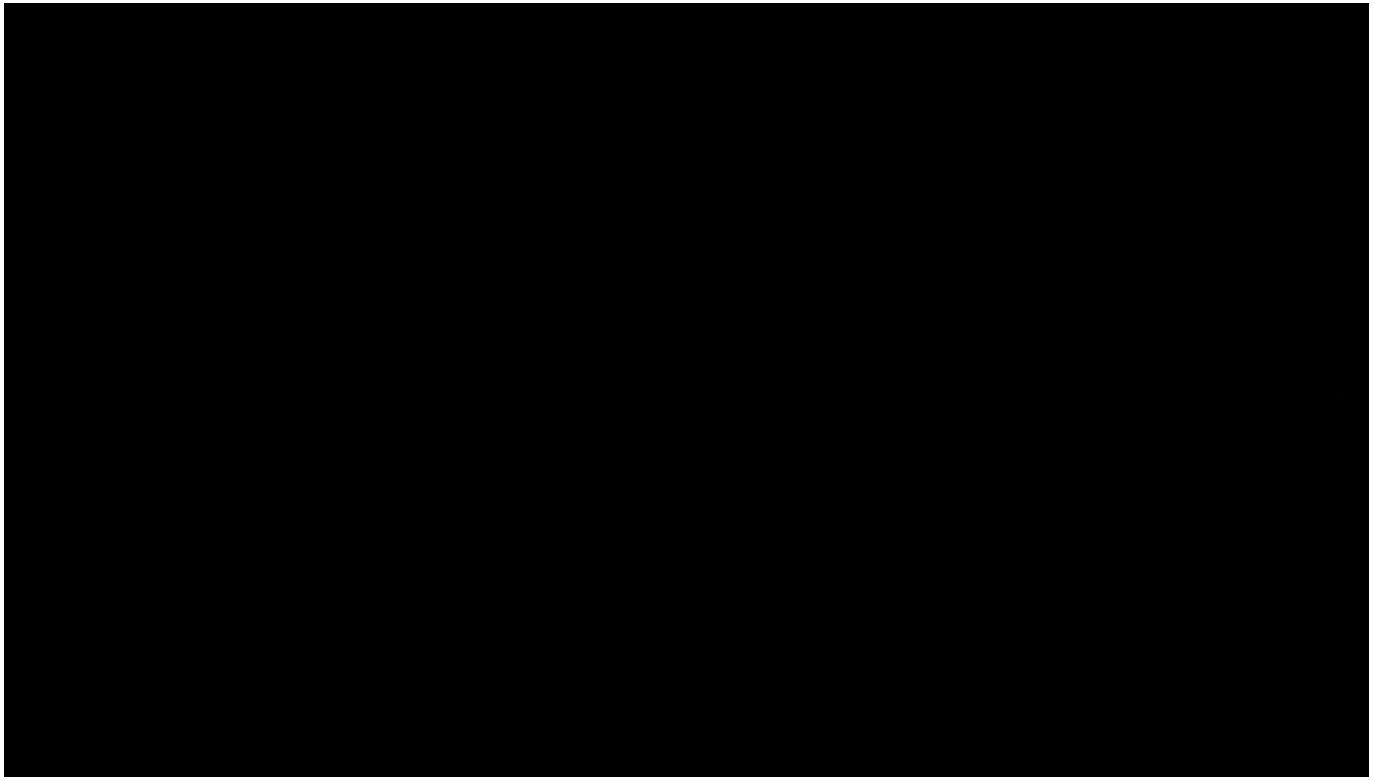
No demonstration
only generic rewards:
+ velocity
- energy
- joint limits

Matches force and angle profiles of humans quite closely

Schuhmacher, …, GM, Häufle, arXiv *2309.02976*

Let it run

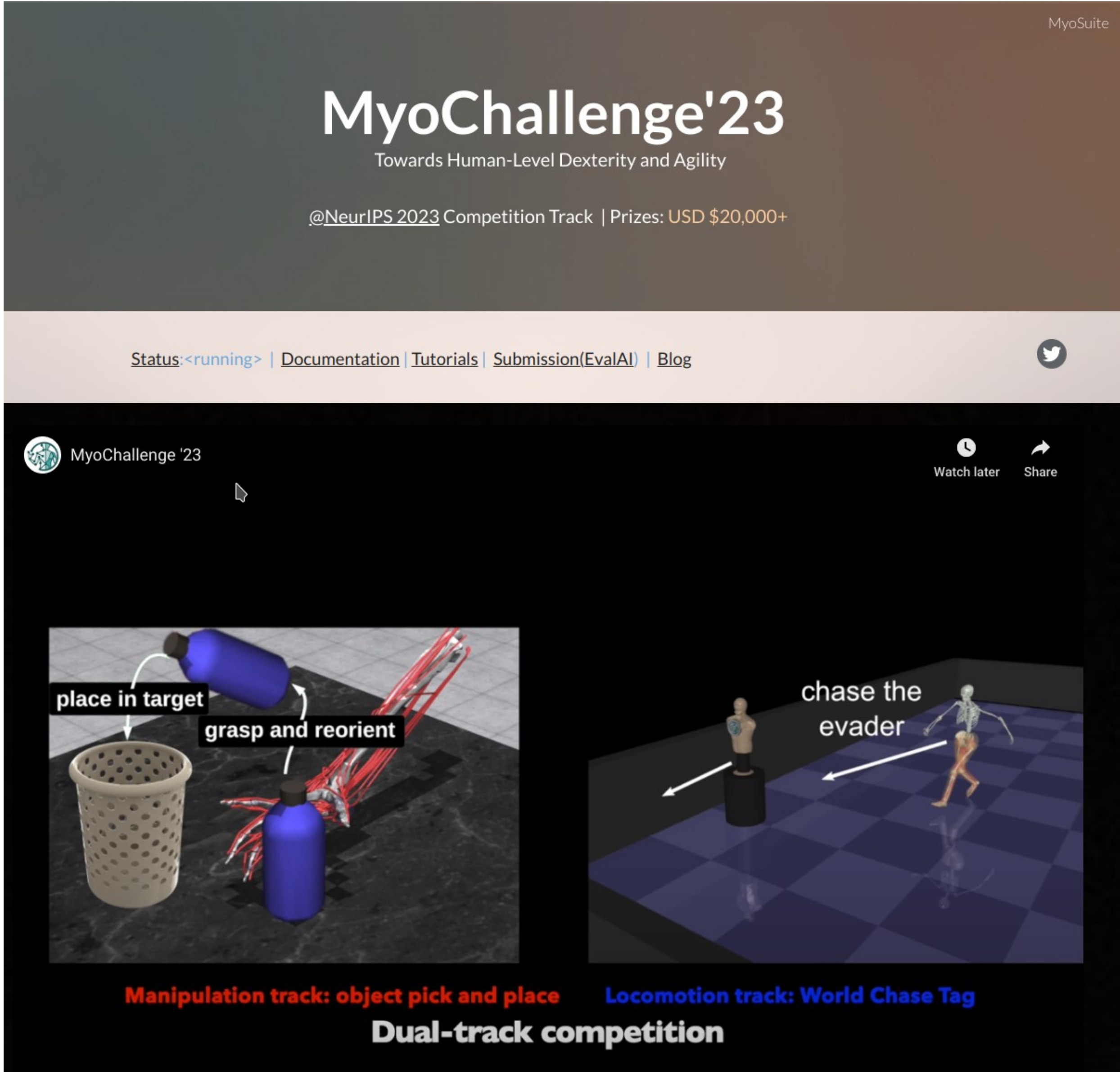
- Exploration is key: DEP-RL: Interleave exploration and policy optimization a random times
- What about models of humans?   Time for the **falling skeleton ;-)**



Schuhmacher, …, GM, Häufle, arXiv *2309.02976*

- NeurIPS 2023 competition
- call to the community to study the control of muscle-skeletal systems.
  - DEP-RL: is a baseline
- manipulation and locomotion

sites.google.com/view/myosuite/myochallenge/myochallenge-2023



Georg Martius <georg.martius@tue.mpg.de>

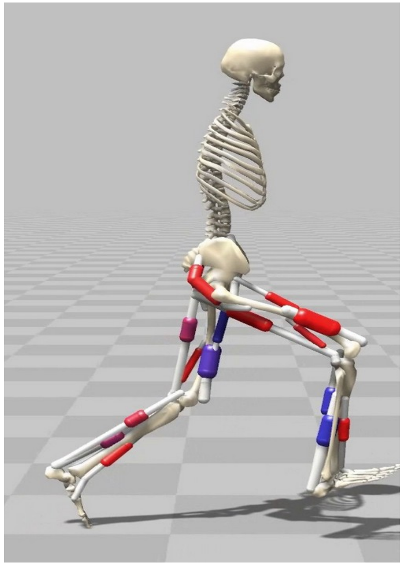
Summary – embodied exploration

✓ over-actuated and/or high-dimensional systems can benefit from embodied exploration:

- take local sensorimotor feedback into account

✓ can learn to control really complicated biophysical models!

✘ still takes millions of steps



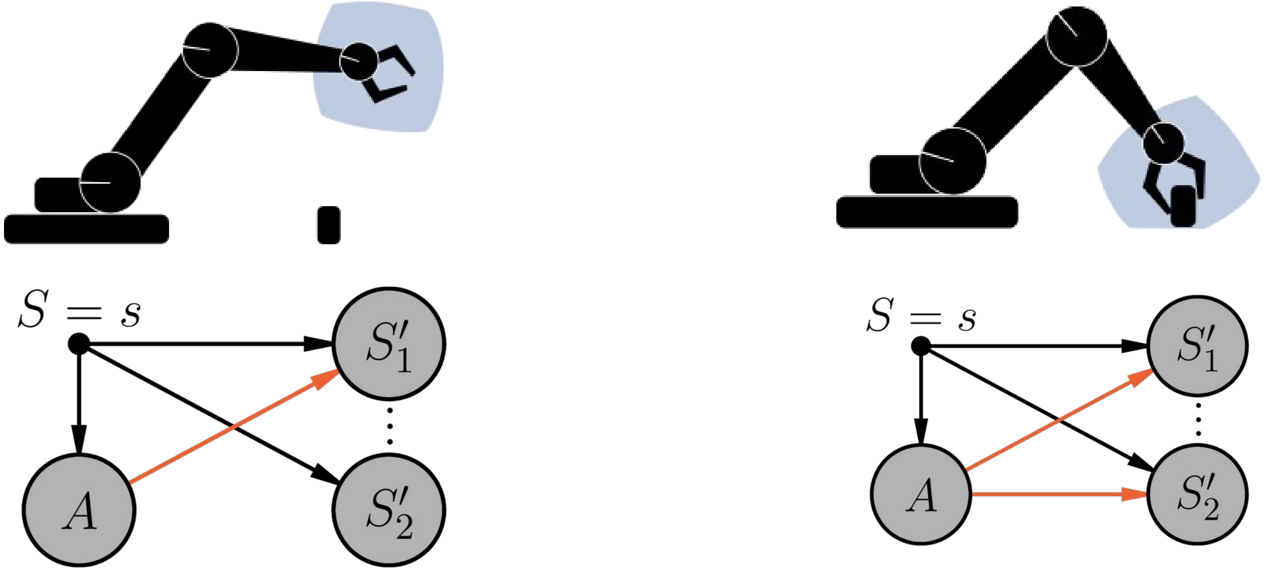
Georg Martius <georg.martius@tue.mpg.de>

Causal Action Influence

Define when actions have causal affect on environment:

- dynamics of object is independent of action

local causal models



MI(Object; Action | Situation)

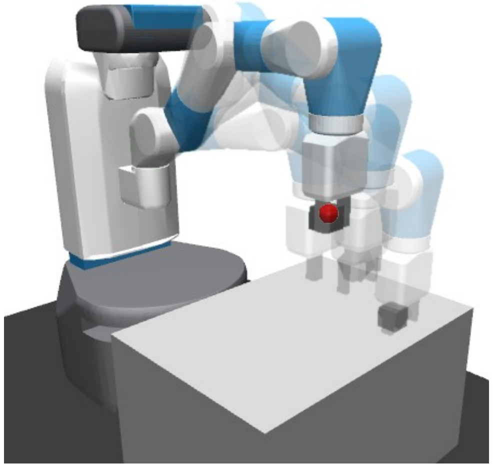
Seitzer, Schölkopf, GM. *NeurIPS* 2021

Georg Martius <georg.martius@tue.mpg.de>

Causal Action Influence

Define when actions have causal affect on environment:

- dynamics of object is not independent of action



CAI: causal action influence

$$C^j(s) := I(S'_j; A \mid S = s) = \mathbb{E}_{a\sim\pi}\left[D_{\mathrm{KL}}\left(P_{S'_j|s,a} \,\middle\|\, P_{S'_j|s}\right)\right]$$

S_j object of interest

probabilistic deep network (gaussian NN)

marginalized (sampling-based)

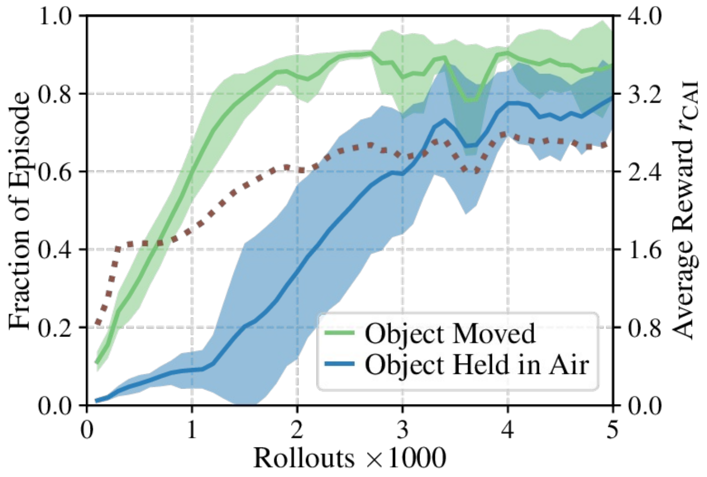
Georg Martius <georg.martius@tue.mpg.de>

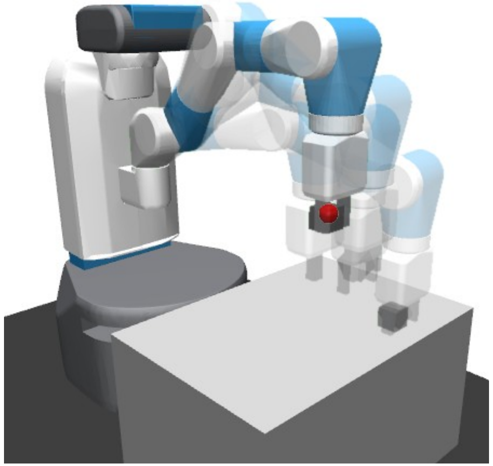
Seitzer, Schölkopf, GM. *NeurIPS* 2021

Causal Action Influence

What can we do with this measure?

➢ use as intrinsic motivation



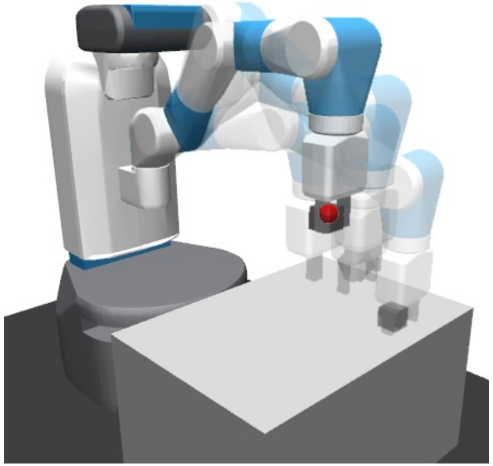


Seitzer, Schölkopf, GM. *NeurIPS* 2021

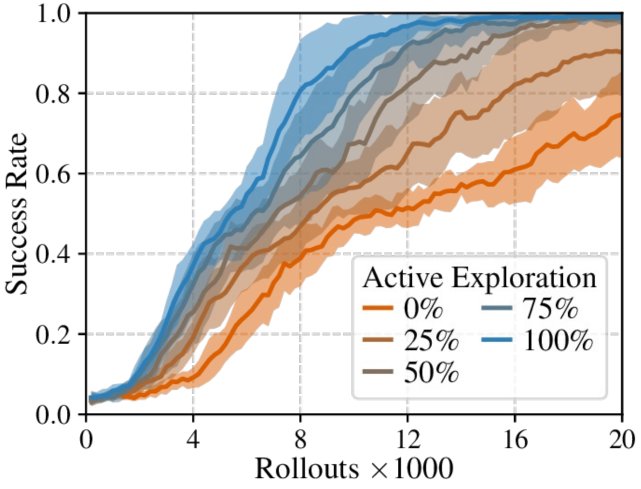
Causal Action Influence

What can we do with this measure?

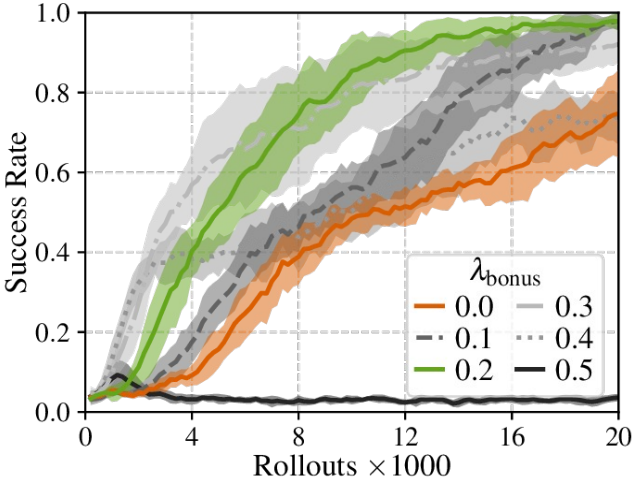
- use as intrinsic motivation
- use for active exploration while aiming for task



active exploration



exploration bonus



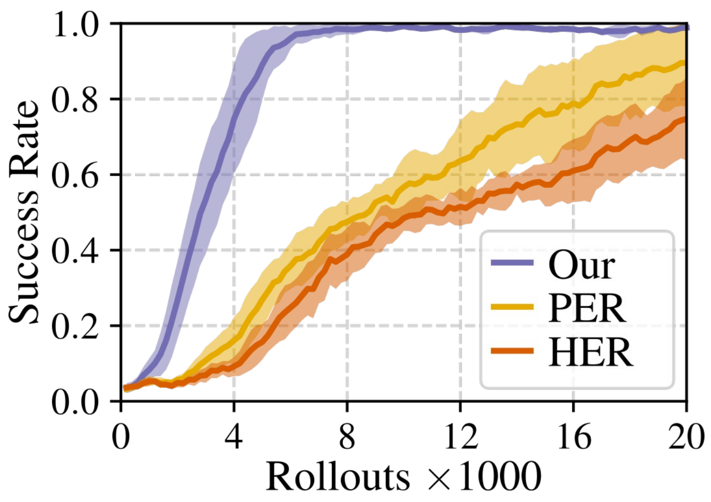
Seitzer, Schölkopf, GM. *NeurIPS* 2021

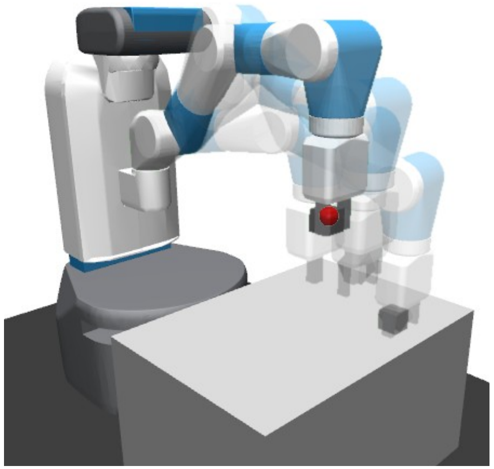
Georg Martius <georg.martius@tue.mpg.de>

Causal Action Influence

What can we do with this measure?

- use as intrinsic motivation
- use for active exploration
- to speed up learning (prioritized replay)





Seitzer, Schölkopf, GM. *NeurIPS* 2021

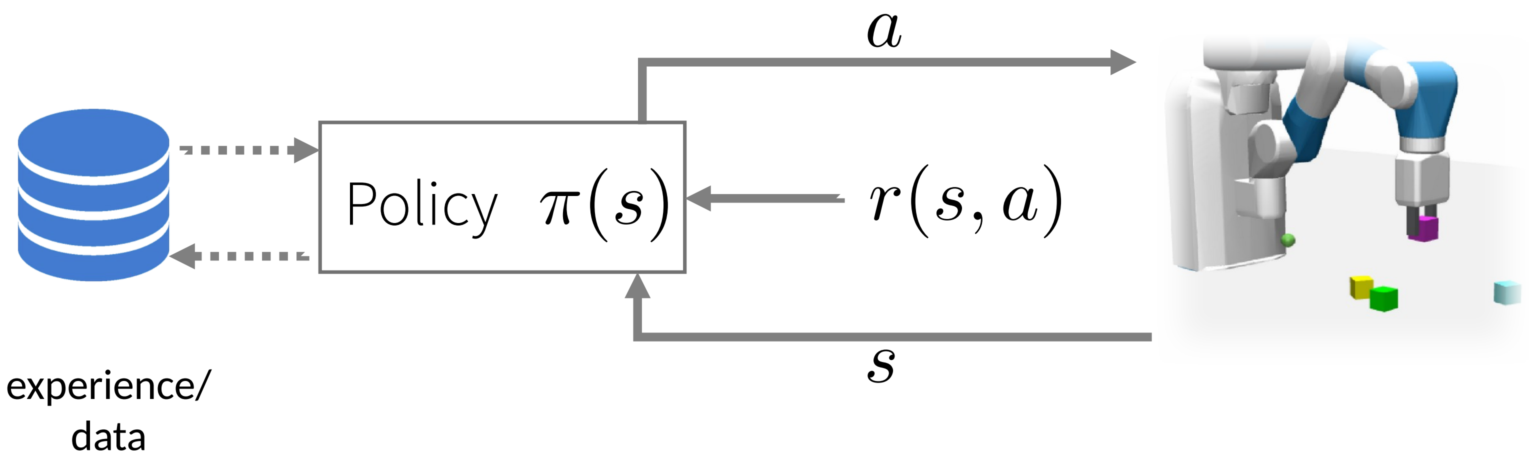
What about autonomous learning?

- Want to leave the robot alone: task-agnostic phase / free play
- Later: come and ask it to perform a task

- Ideally sample efficient enough for a real robot

Georg Martius <georg.martius@tue.mpg.de>

Reinforcement learning



Aim: Find policy π that maximizes future reward: $\mathbb{E}_{s_t \sim \pi} \sum_t \gamma^t r(s_t)$

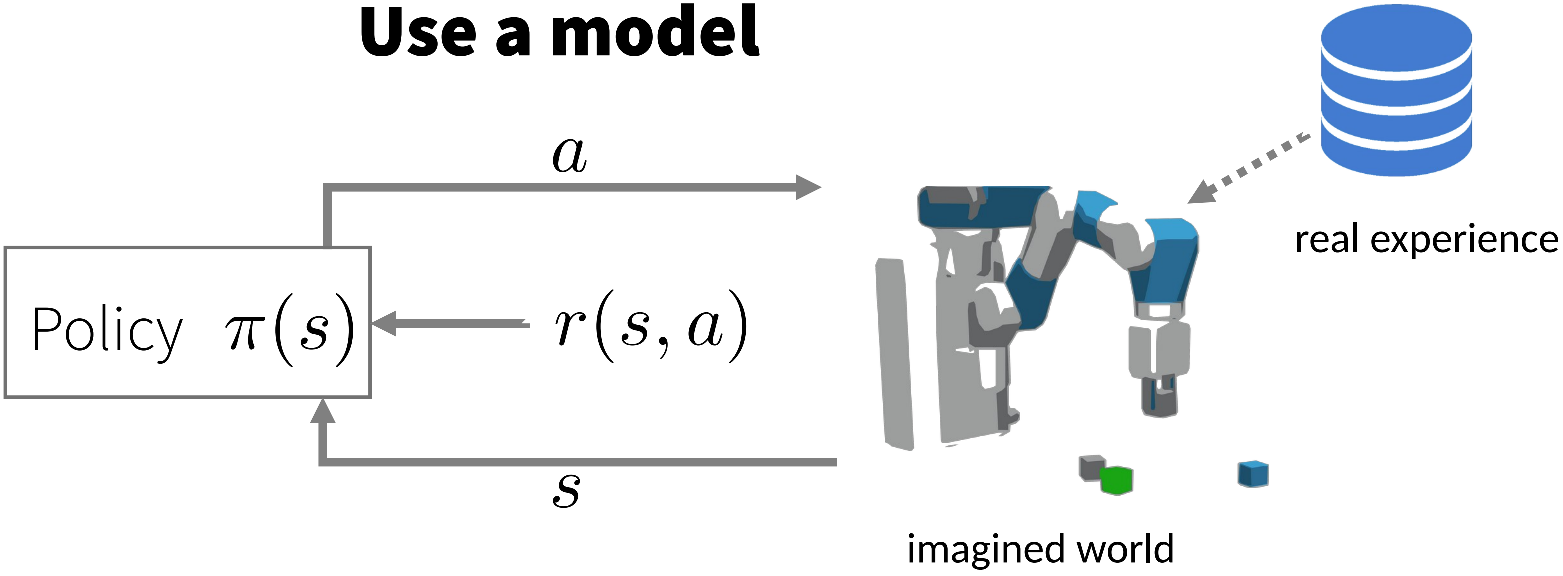
➢ Approach: learn from experience by trial an error

needs a prohibitive amount of interactions

for real-world systems

[ionos.com]

Use a model



Interact with a **model** of the world:

→ can do trial and error learning using the model (mental simulation)

Enables to compute **reward in imagination**

Georg Martius <georg.martius@tue.mpg.de>

Properties of Intrinsic Motivations Signals

In RL: intrinsic motivation is typically an additional reward

- Curiosity, Learning progress, Competence
- Prediction Error (Intrinsic Curiosity Module)
- Novelty search
- Adversarial selfplay

Retrospective
- hard to predict

- Predicted information gain, Reduction of epistemic uncertainty
- Empowerment, Causal action influence
- Skill diversity
- Regularity

Predictable

Why does it matter?

Predictable IM signals can be used in model-based optimization!

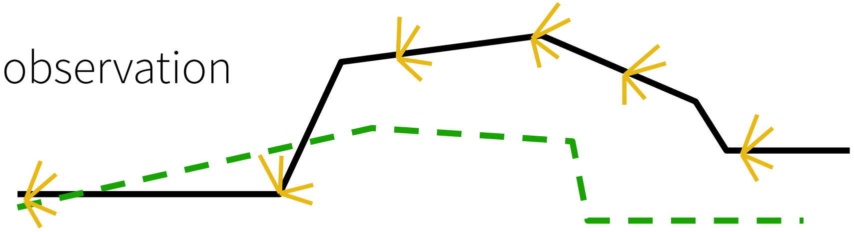
Georg Martius <georg.martius@tue.mpg.de>

Model-based Reinforcement Learning

Two instantiations

Planning at learning time

- ➢ use model to collect data nearby real observation
- ➢ learn to solve a **specific** task
- ➢ **global optimization**

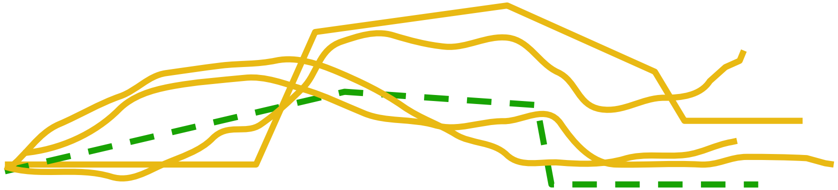


MBPO
Dreamer

MuZero

Planning at run-time

- ➢ use model for planning
- ➢ perform **new task on the fly**
- ➢ **optimize finite horizon** problem



PETS
Plan2Explore

Need:
Fast optimizer
Good model

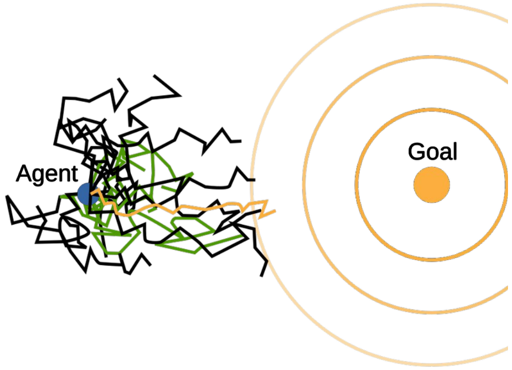
Georg Martius <georg.martius@tue.mpg.de>

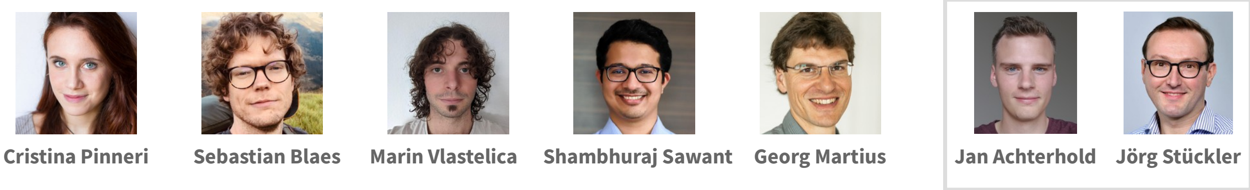
Model-based Planning

Cross Entropy Method (CEM)

- ➢ Sampling based optimization

$$a_{t,\dots,t+H} \sim \mathcal{N}(\mu_i, \sigma_i^2)$$



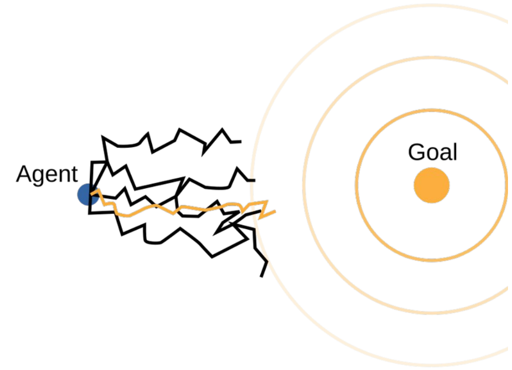


Planning with Temporal Correlation

Cross Entropy Method (CEM)

- Sampling based optimization

$$a_{t,\dots,t+H} \sim \mathcal{N}(\mu_i, \sigma_i^2)$$



improved Cross Entropy Method

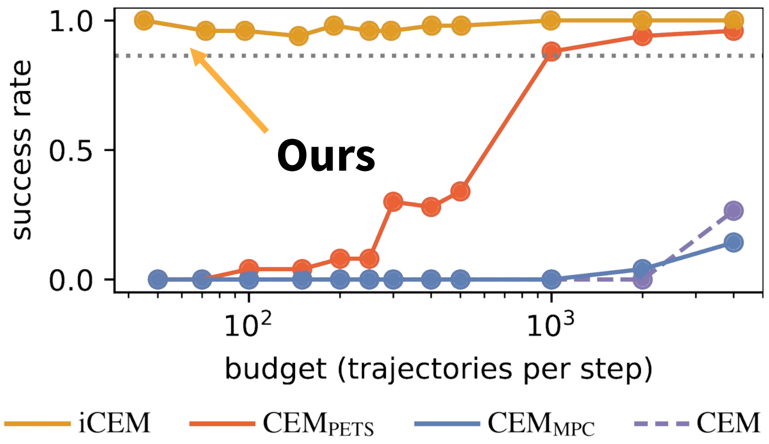
- ✚ Memory
- ✚ Colored noise: temporal correlation

Power Spectral Density $\propto \dfrac{1}{f^\beta}$

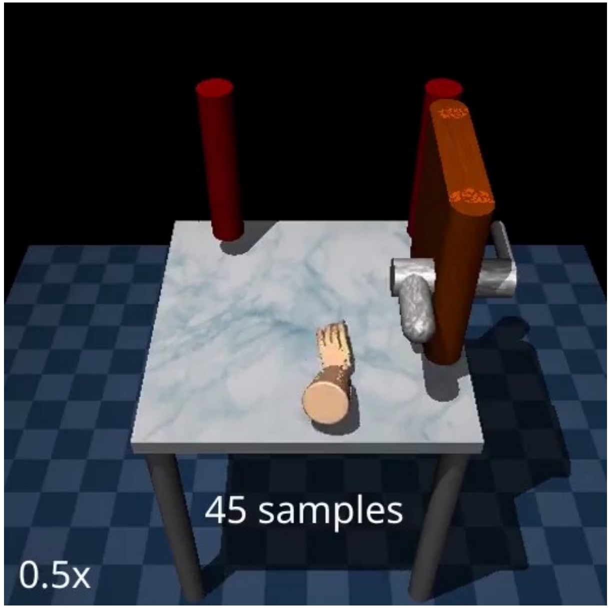
Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Model-based Planning

Door (sparse reward)



ground truth models
(simulator)

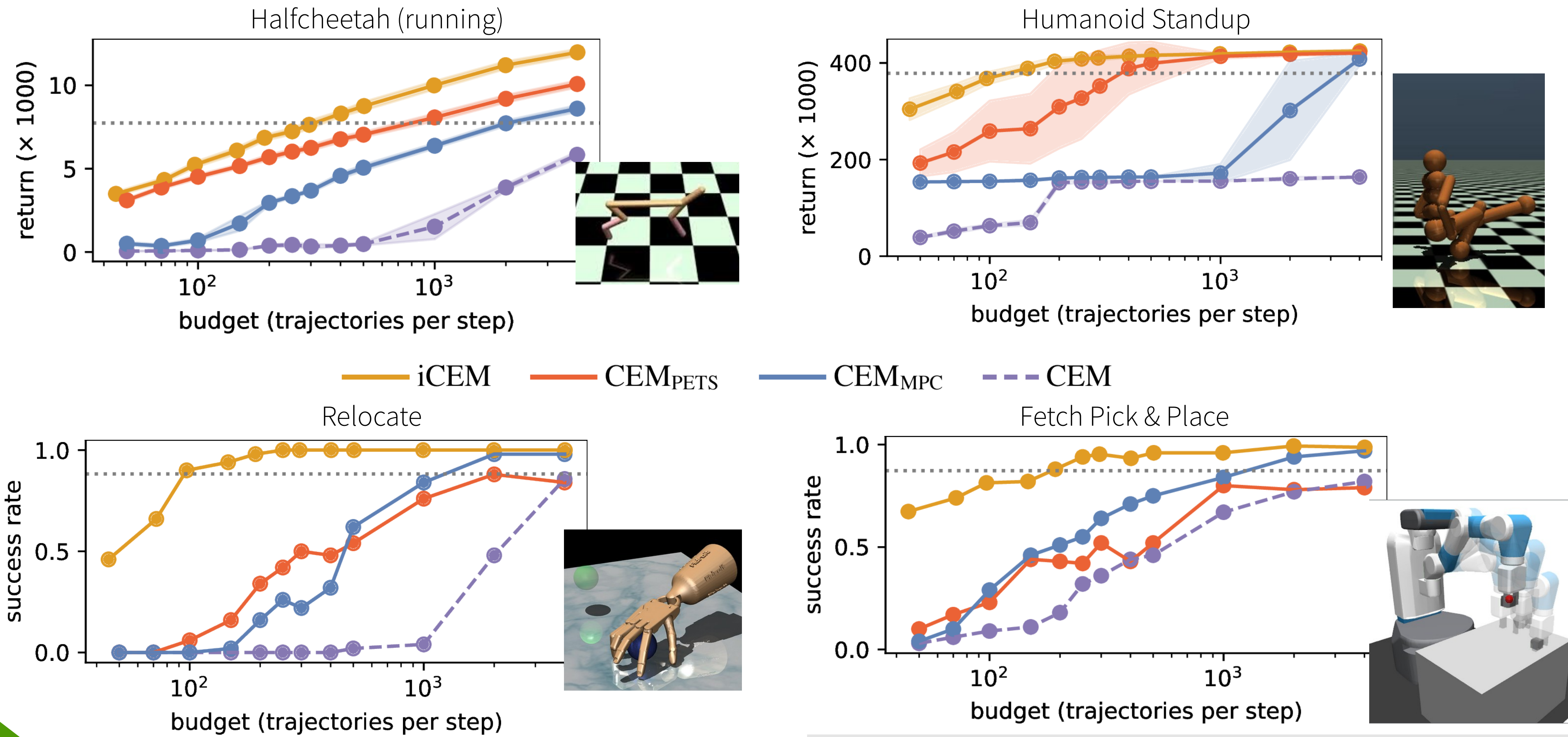


(environment from DAPG project)

Georg Martius <georg.martius@tue.mpg.de>

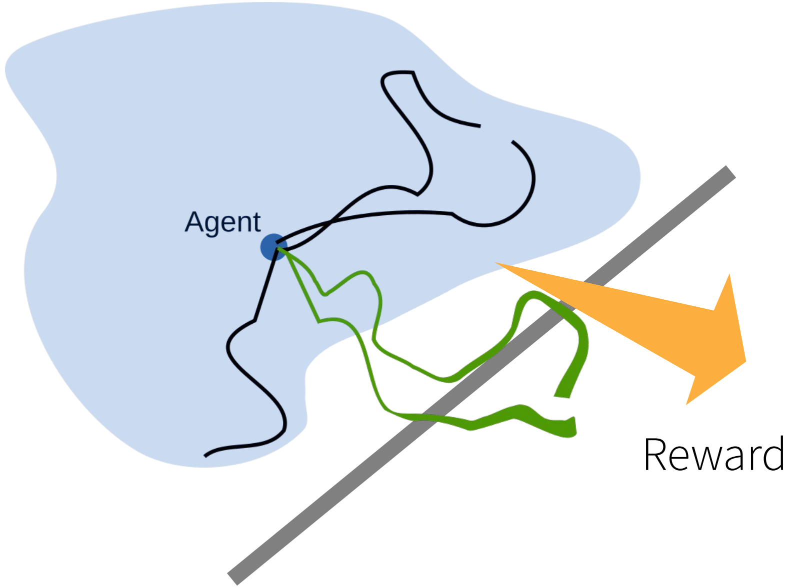
Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Model-based Planning



Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Use learned models... what can go wrong?



The planner will **exploit model errors**

- ➢ Non-sense behavior is executed
- ➢ Need to **know** what the **model does not know**

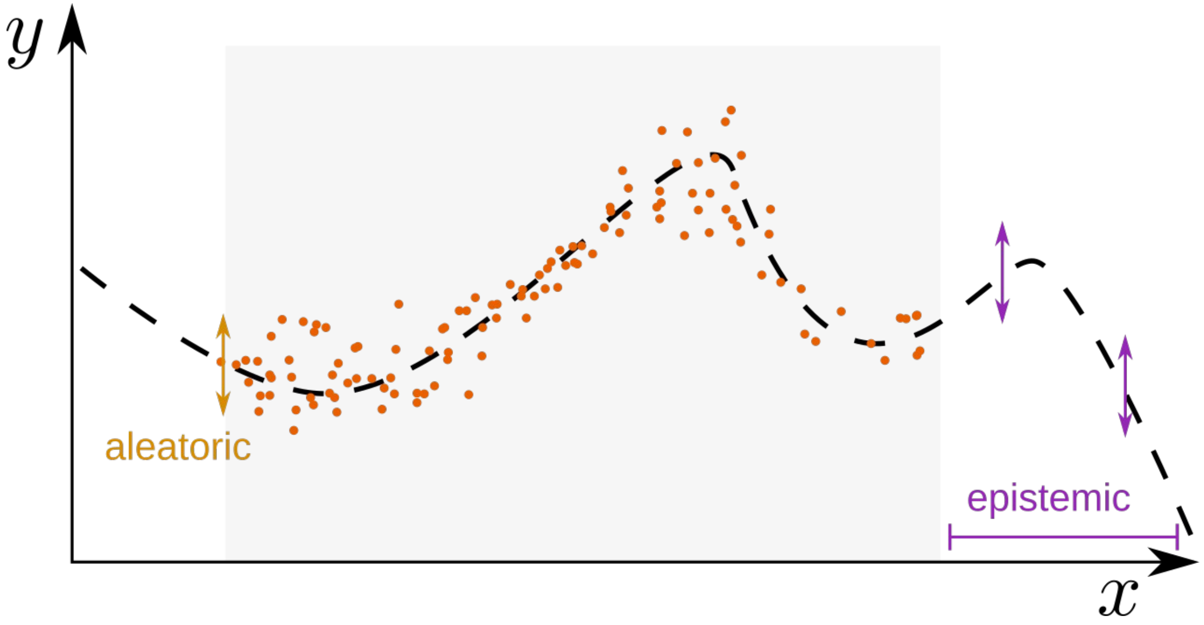
Georg Martius <georg.martius@tue.mpg.de>

Dynamics Models with Uncertainty

- ➢ separation of *aleatoric* and *epistemic* uncertainty

Why?

- ➢ aleatoric: avoid
- ➢ epistemic:
  - seek to reduce during exploration
  - avoid during exploitation



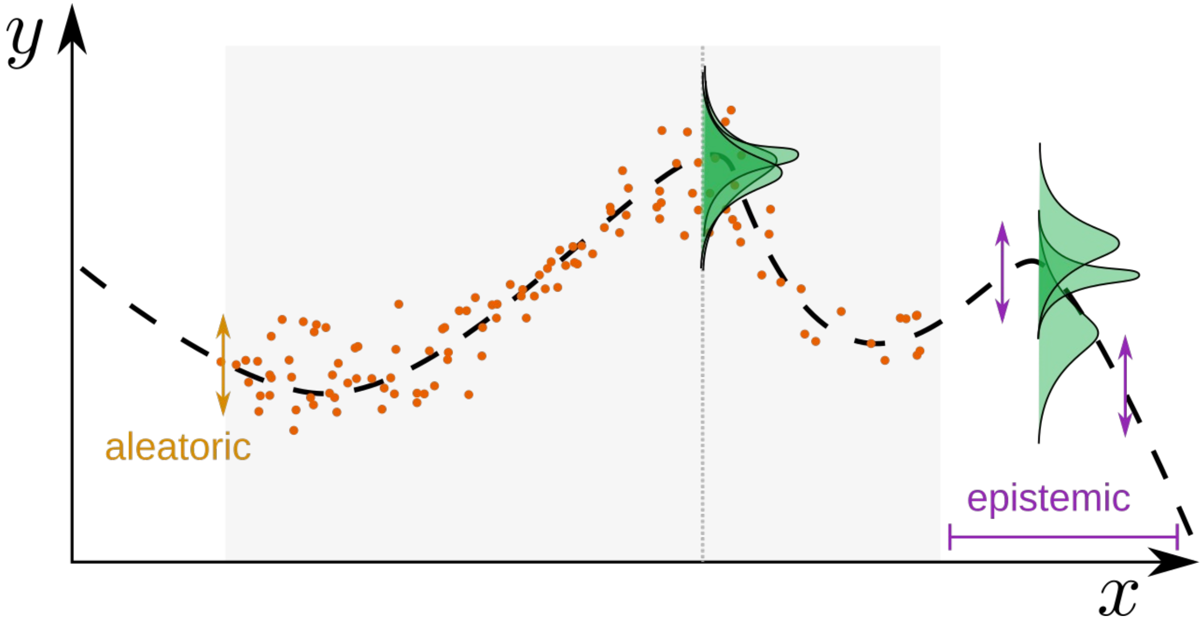
Georg Martius <georg.martius@tue.mpg.de>

Dynamics Models with Uncertainty

- separation of *aleatoric* and *epistemic* uncertainty

Why?

- aleatoric: avoid
- epistemic:
  - seek to reduce during exploration
  - avoid during exploitation



Ensemble of probabilistic Deep Nets

- good estimates of separation both types of uncertainty

Georg Martius <georg.martius@tue.mpg.de>

Dynamics Models with n-step Uncertainty

- ➢ separation of *aleatoric* and *epistemic* uncertainty

Why?

- ➢ aleatoric: avoid
- ➢ epistemic: seek to reduce / avoid during exploitation

What about compounding uncertainties (n-step)

- ➢ Non-trivial, but can be solved practically:
- ➢ PETS: [Chua et al 2018] Probabilistic Ensemble models with Trajectory Sampling
- ➢ RAZER: [Vlastelica, Blaes, Pinneri, GM. CORL 2021]: Disentangle epistemic and aleatoric for n-steps
- ➢ Beta-NLL [Seitzer, Tavakoli, GM. ICLR 2022]: make training of prob. NN models work

Georg Martius <georg.martius@tue.mpg.de>

Properties of Intrinsic Motivations Signals

In RL: intrinsic motivation is typically an additional reward

- Curiosity, Learning progress, Competence
- Prediction Error (Intrinsic Curiosity Module)
- Novelty search
- Adversarial selfplay

Retrospective
- hard to predict

- **Predicted information gain, Reduction of epistemic uncertainty**
- Empowerment, Causal action influence
- Skill diversity
- **Regularity**

Predictable

Why does it matter?

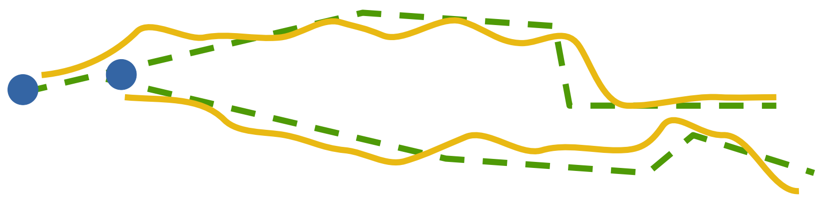
Predictable IM signals can be used in model-based optimization!

Georg Martius <georg.martius@tue.mpg.de>

Planning for Intrinsic Motivation

Planning on the fly for an IM signal

- intrinsic motivation signals are **non-stationary** by design
- can plan for **n-step IM**



In model-free RL:

- need to first find the (intrinsically) rewarding regions (value function and policy)
- then unlearn as new things become more rewarding etc

→ slow

Georg Martius <georg.martius@tue.mpg.de>

Plan for Predicted Information Gain

Learn **autonomously** to prepare for **future tasks**

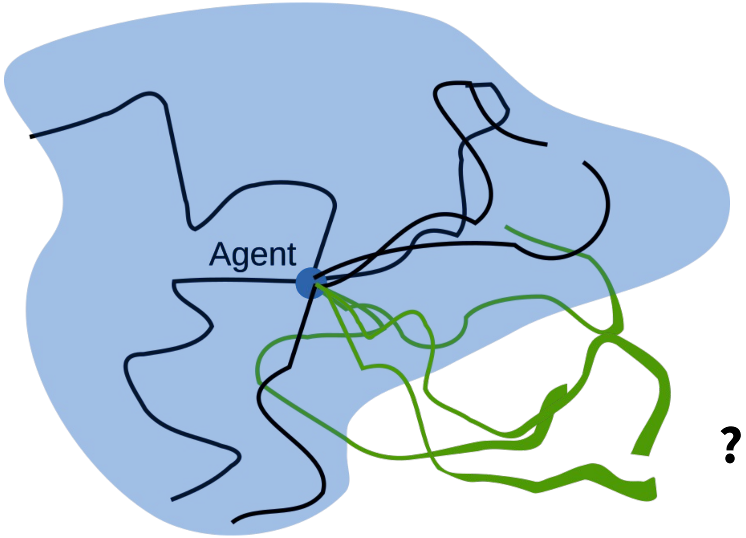
➢ plan for **predicted information gain**

Cansu Sancaktar

Sebastian Blaes

Cristina Pinneri

Marin Vlastelica



Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

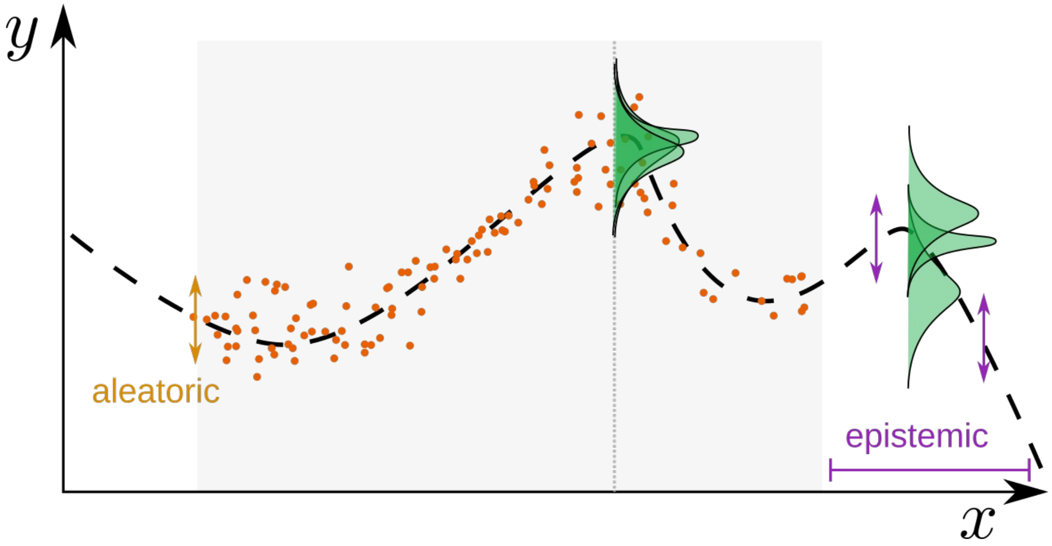
Sancaktar, Blaes, GM. NeurIPS 2022

Georg Martius <georg.martius@tue.mpg.de>

How to measure/predict information gain?

epistemic uncertainty = proxy for information gain

“expect to gain information where uncertain because of lacking data”



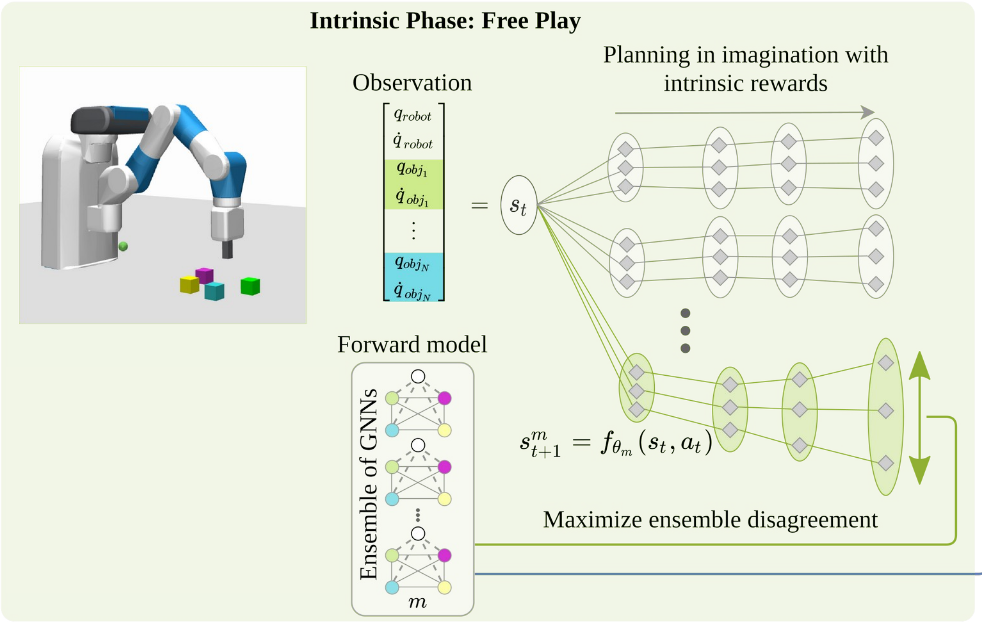
- ➢ Bayesian Neural Nets
- ➢ Ensembles ← are most practical at the moment

Georg Martius <georg.martius@tue.mpg.de>

Plan for Predicted Information Gain

Seeking information:

- Learn a ***structured mental* model** of the world (graph net)
- Plan behavior where the outcome is uncertain / expect to learn something



$$r(s) = \sum_{k=1}^{K} (\mu_k(s) - \bar{\mu}(s))^2$$
$$= \mathrm{Var}(\text{Ensemble predictions})$$

Same objective as in "Plan To Explore"

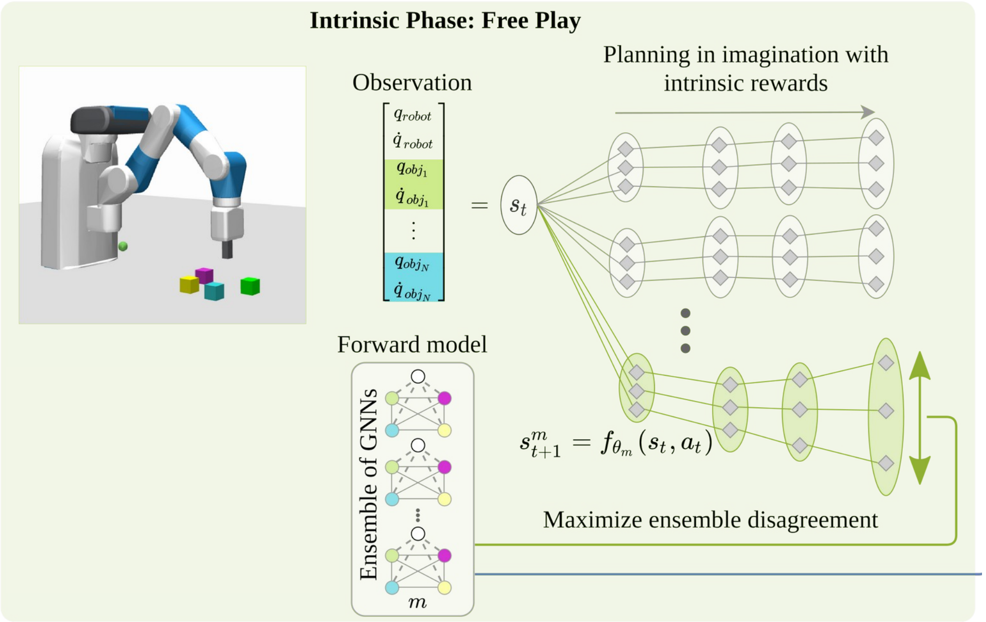
Sancaktar, Blaes, GM. NeurIPS 2022

Georg Martius <georg.martius@tue.mpg.de>

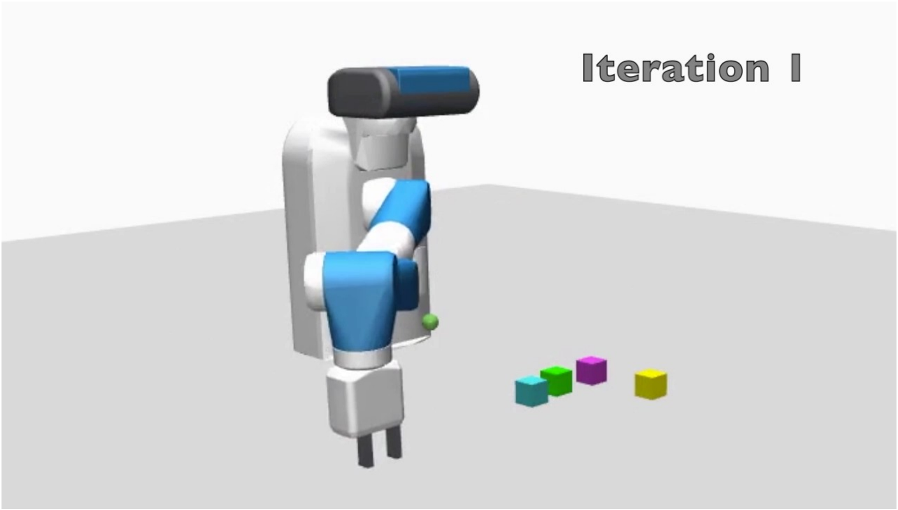
Intrinsically Motivated Learning

Seeking information:

- Learn a *structured mental* model of the world (graph net)
- Plan behavior where the outcome is uncertain / expect to learn something



Free Play

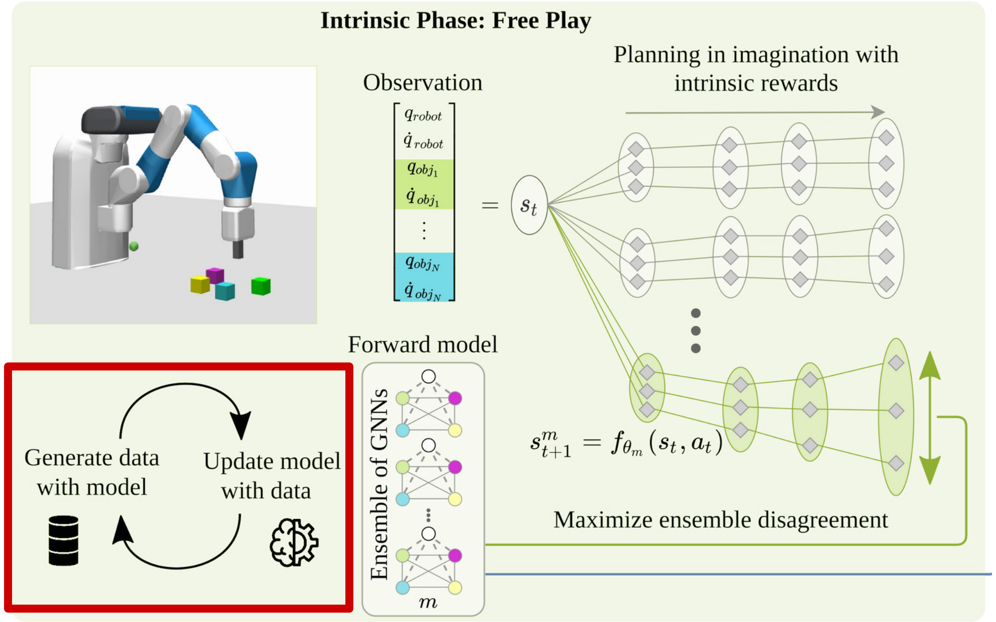


Sancaktar, Blaes, GM. NeurIPS 2022

Intrinsically Motivated Learning

Seeking information:

- Learn a *structured mental* model of the world (graph net)
- Plan behavior where the outcome is uncertain / expect to learn something

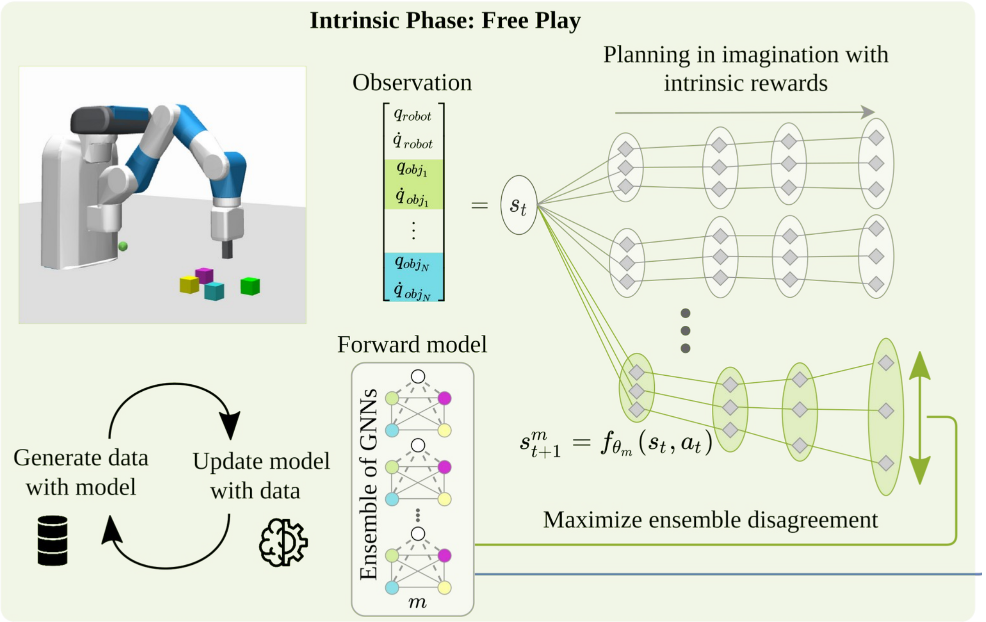


Sancaktar, Blaes, GM. NeurIPS 2022

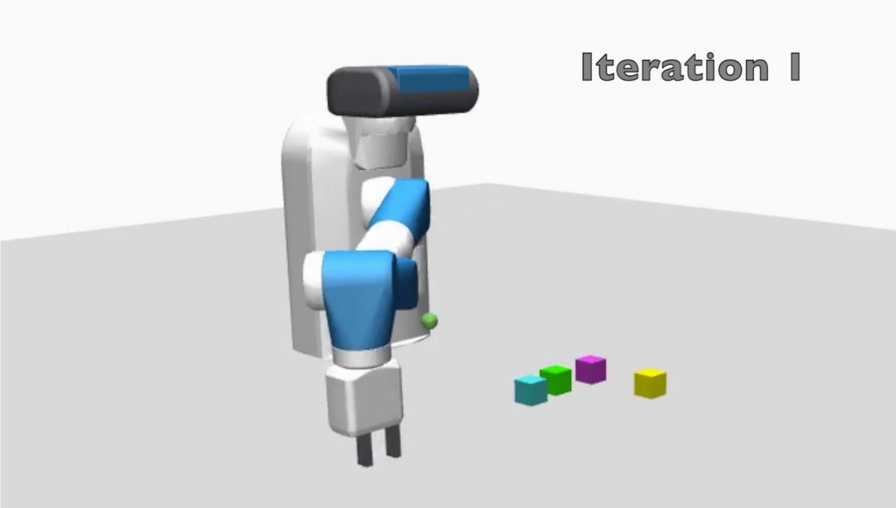
Intrinsically Motivated Learning

Seeking information:

- Learn a *structured mental* model of the world (graph net)
- Plan behavior where the outcome is uncertain / expect to learn something

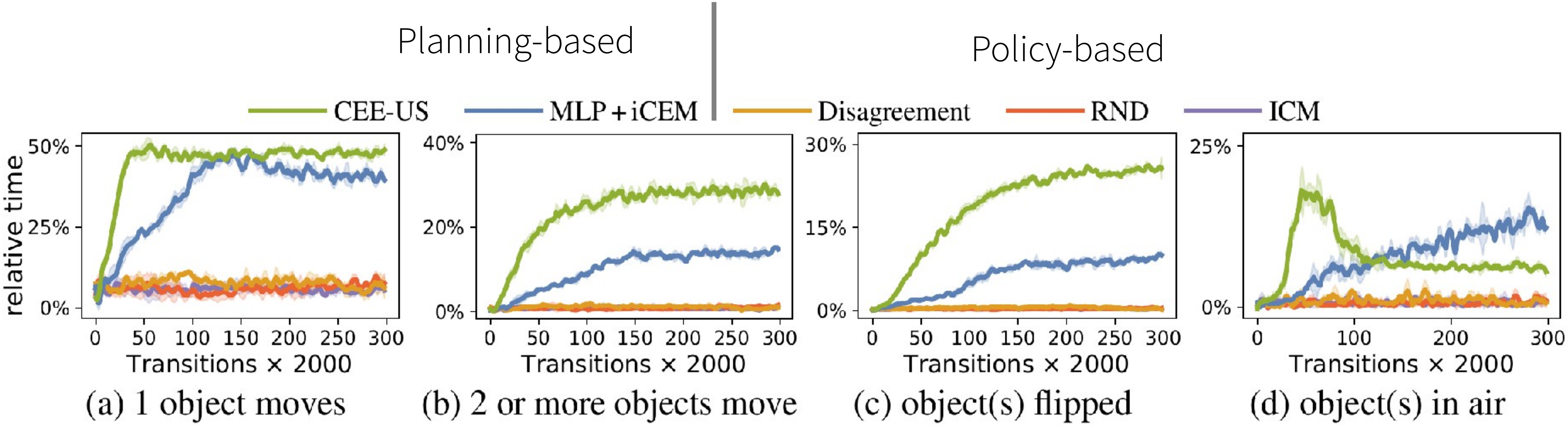


Free Play



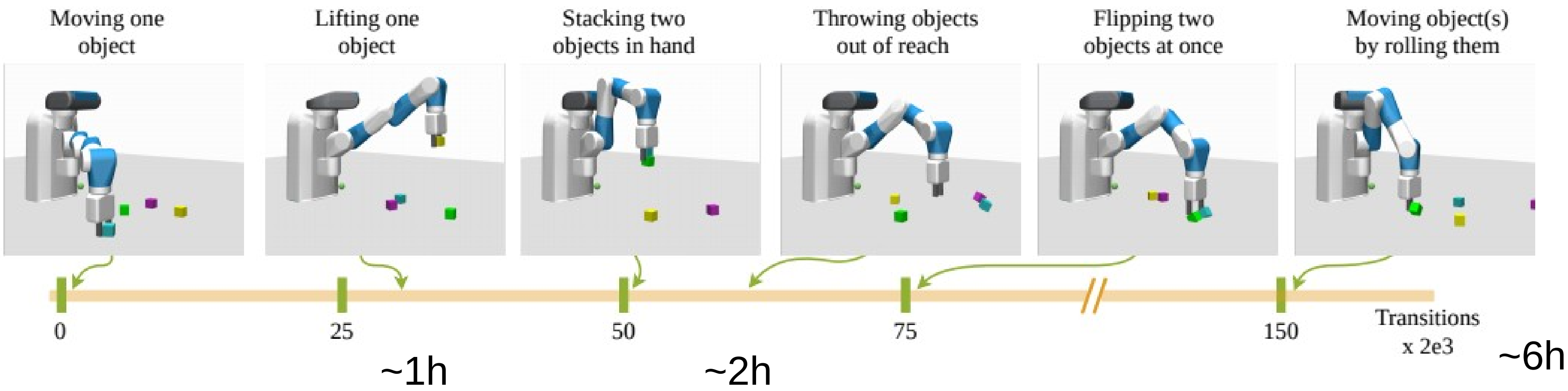
Sancaktar, Blaes, GM. NeurIPS 2022

Interaction Statistics



- Planning (for N-steps) matters
- Structured model (GNN) increases performance

Emergent Behavior



Loose comparison for lifting: **(different environment, ...)**

SELMO: **10M transitions** Groth et al: "Is Curiosity All You Need? On the Utility of Emergent Behaviours from Curious..."

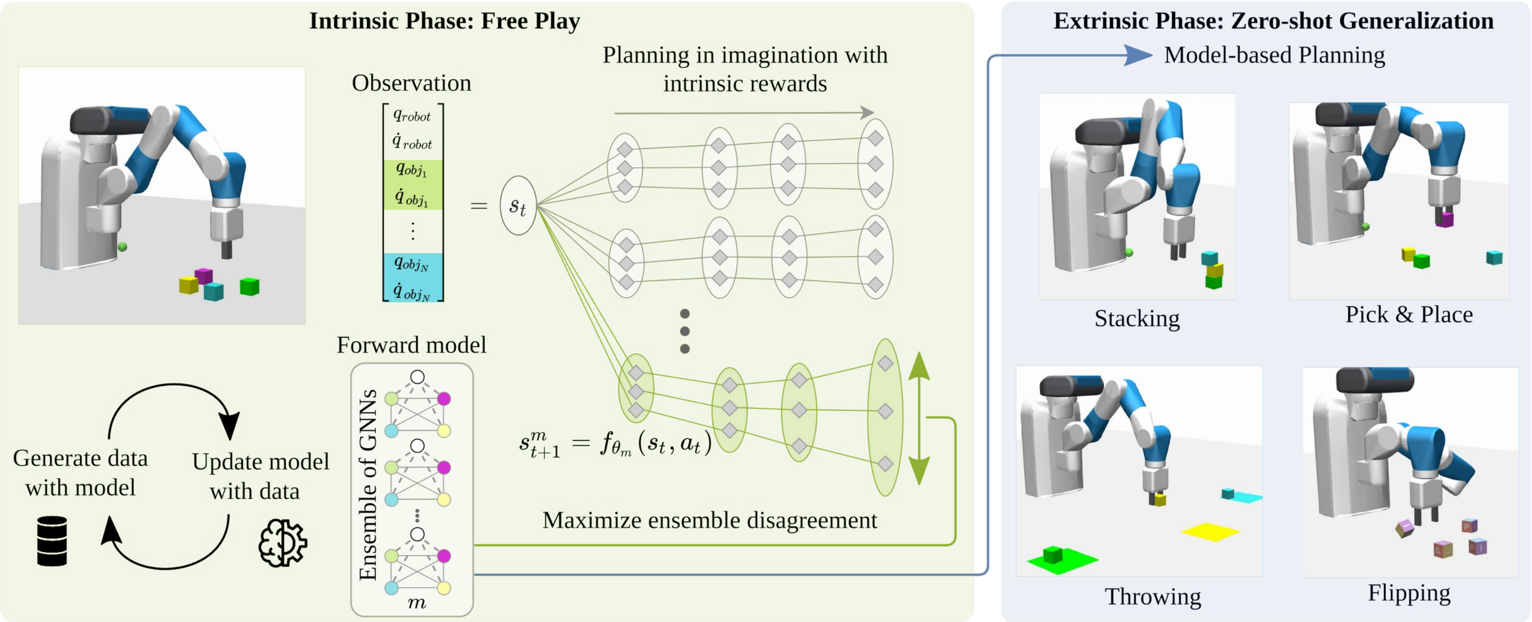
CEE-US: **60K transitions (ours)**

Sancaktar, Blaes, GM. NeurIPS 2022

Perform a task

"Think" and plan to perform a given task:

- use mental model of the world to plan for a given task

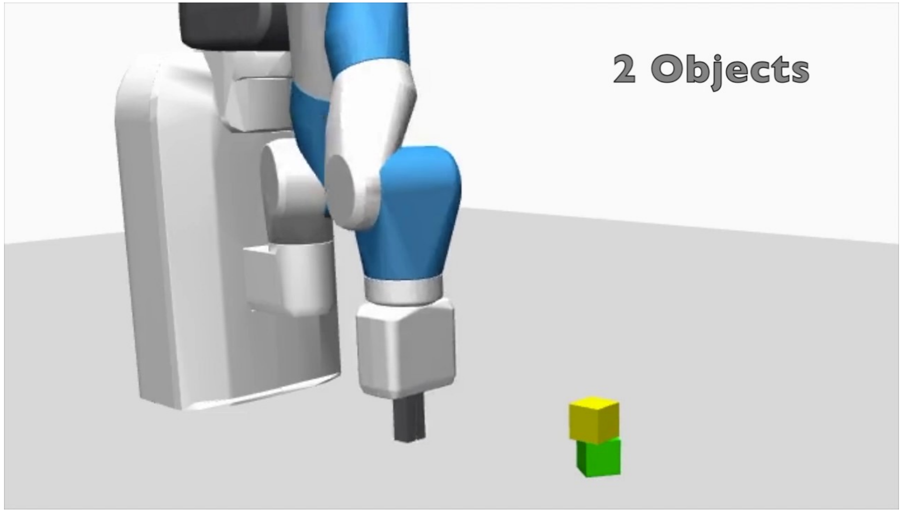


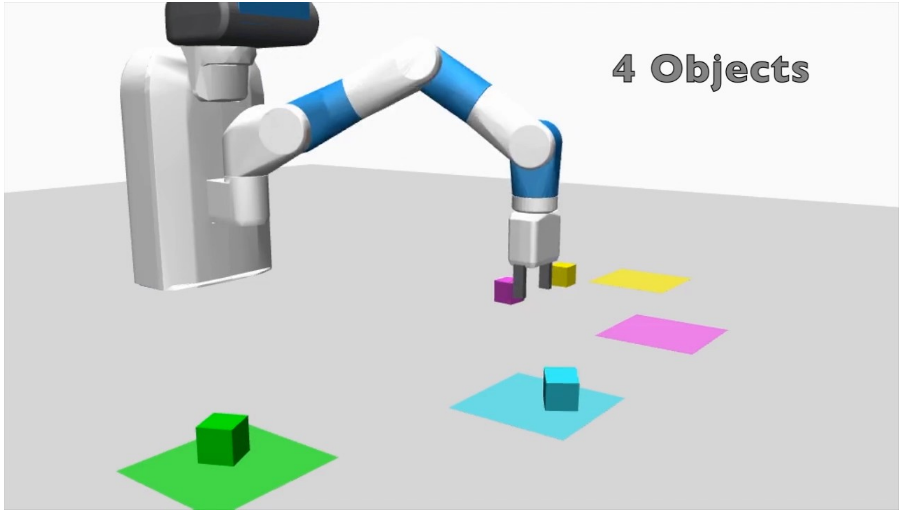
Sancaktar, Blaes, GM. NeurIPS 2022

Perform a task – zero shot generalization

"Think" and plan:

- use mental model of the world to plan for a given task





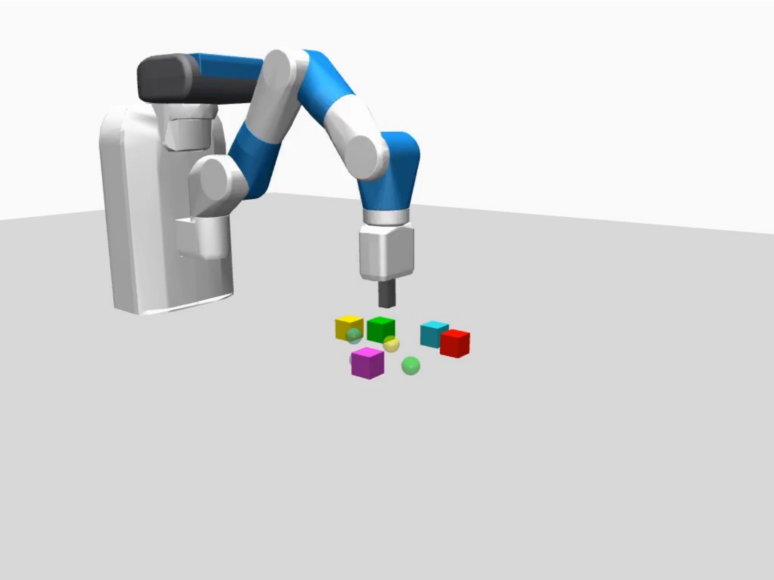
https://cee-us.github.io/

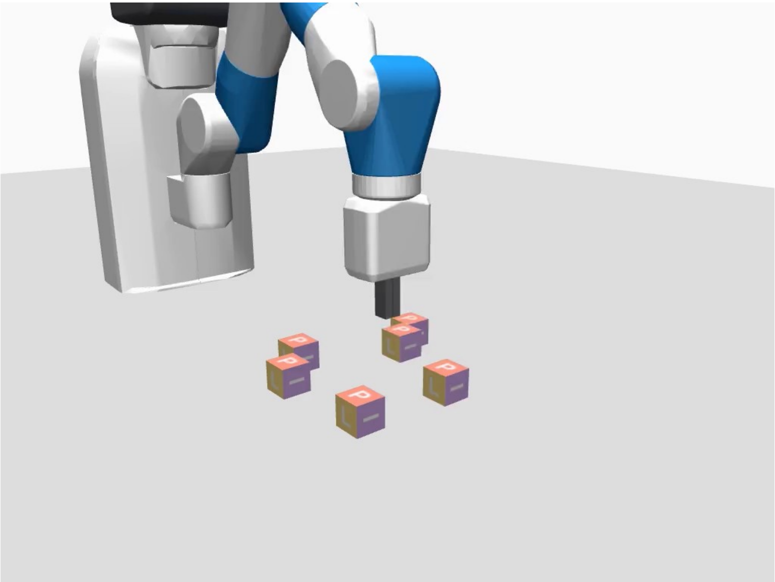
Sancaktar, Blaes, GM. NeurIPS 2022

Perform a task – zero shot generalization

"Think" and plan:

- use mental model of the world to plan for a given task

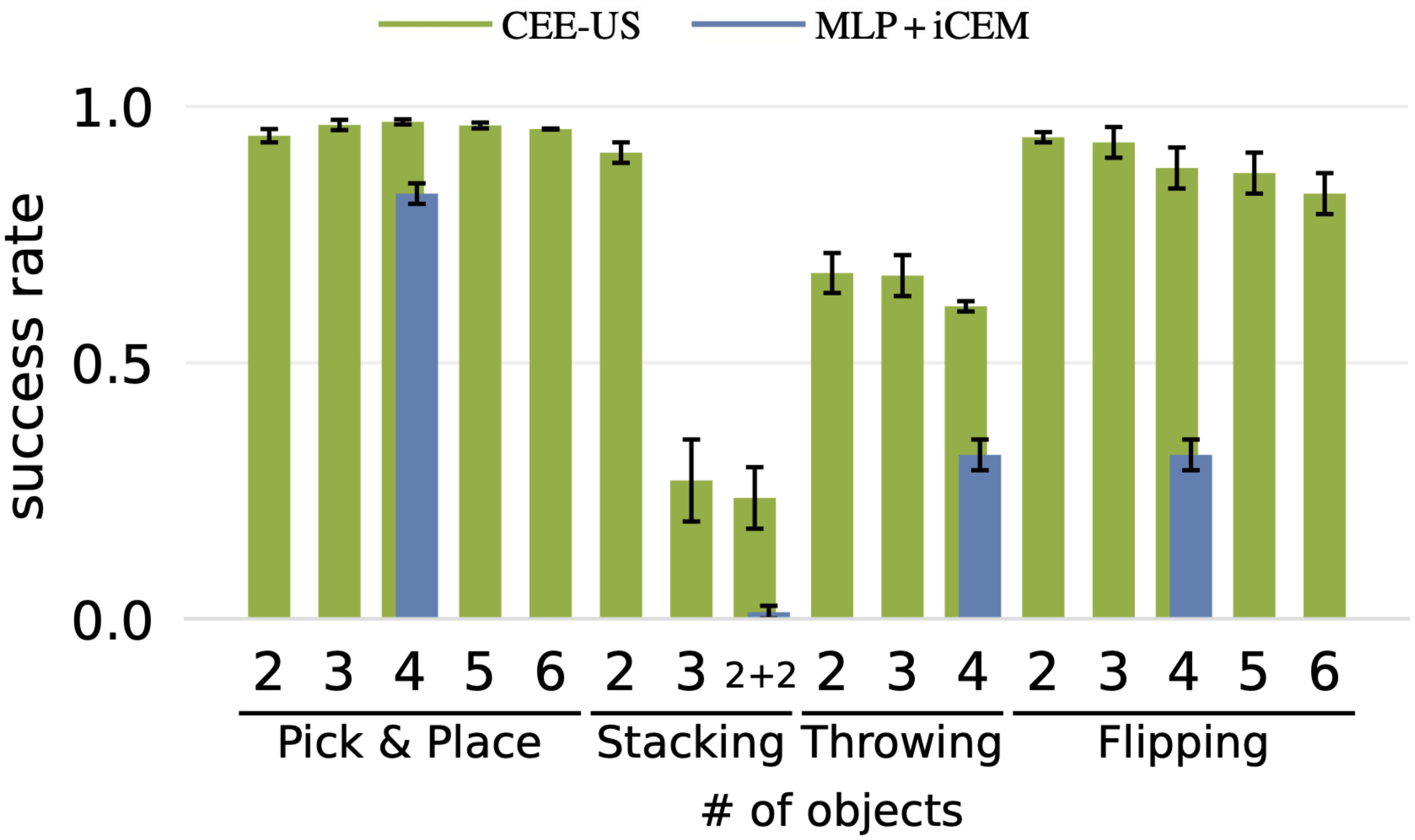




https://cee-us.github.io/

Sancaktar, Blaes, GM. NeurIPS 2022

Perform a task – zero shot generalization



Georg Martius <georg.martius@tue.mpg.de>

Could we also use Offline RL?

Perform offline RL to extract task-policy

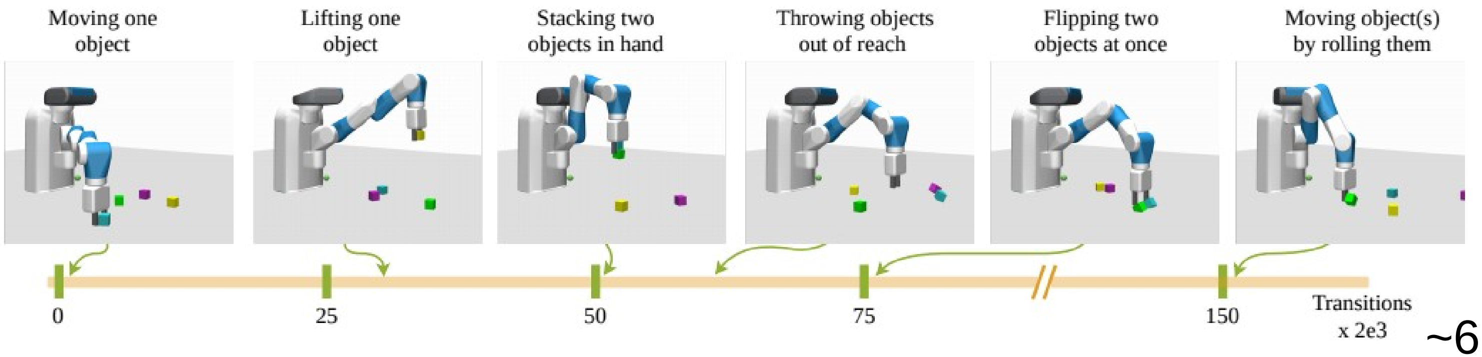
Domain	Task	Disagreement	RND	ICM	MLP + iCEM	CEE-US
CONSTRUCTION	Reach	0.09 ± 0.01	0.19 ± 0.05	0.2 ± 0.03	0.65 ± 0.09	0.94 ± 0.04
600k datapoints	Pick & Place 1 obj.	0.07 ± 0.0	0.07 ± 0.0	0.07 ± 0.01	0.18 ± 0.06	0.43 ± 0.07

- More difficult tasks did not work!
- Lot do to for offline-RL

- Bagatella et al @ EWRL: Goal-conditioned Offline Planning from Curious Exploration
  - Offline RL often suffer from estimation artifacts: can be circumvented with model-based corrections

Intermediate Summary

- **Model-based planning** works with good planners and ensemble network networks
- **Uncertainties** become instrumental: as **intrinsic reward** + to make models **robust**
- **Predictable Intrinsic Motivation** signals + model-based planning → **great sample efficiency**
- First demonstration of: task-agnostic free-play → zero-shot task performance in a difficult setting
- Still lots of limitations (e.g. not full RL setup)



~6h

Georg Martius <georg.martius@tue.mpg.de>

Put more Structure into Play?

Novel ≠ Useful


Cansu Sancaktar


Justus Piater

Put more Structure into Play?

Novel ≠ Useful

What is a generic bias for constructing things?

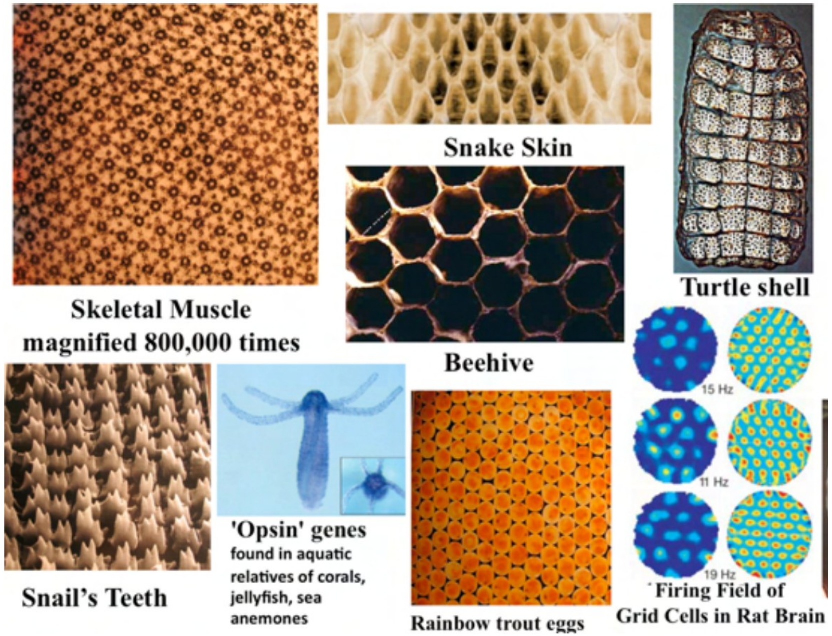
Cansu Sancaktar

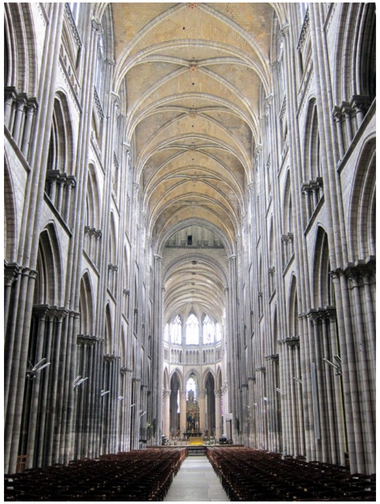
Justus Piater

Put more Structure into Play?

Regularity and symmetries are everywhere.

➢ Regularity as Intrinsic Reward (RaIR)





Rouen Cathedral

Neue Aula, Uni Tübingen

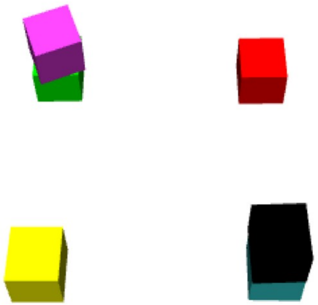
Put more Structure into Play?

Regularity = Redundancy in scene description

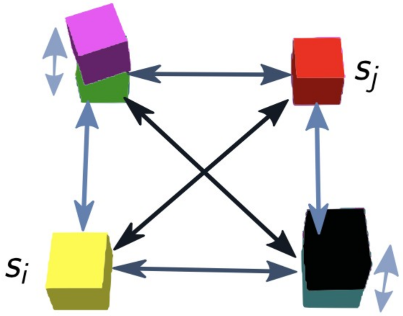
➢ Measured by Entropy of some representation

➢ Example:

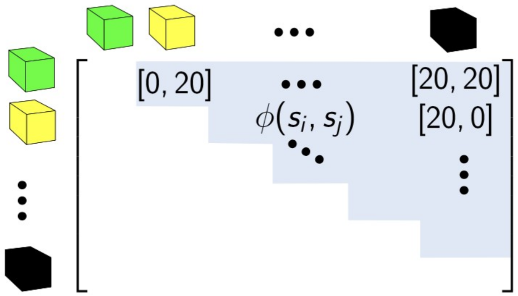
Scene



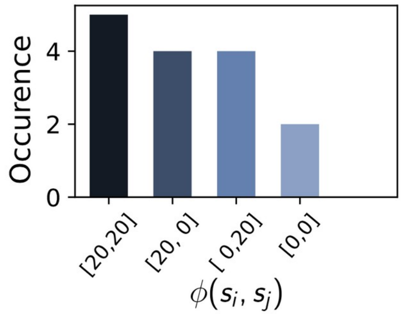
Representation



Relation Matrix



Histogram



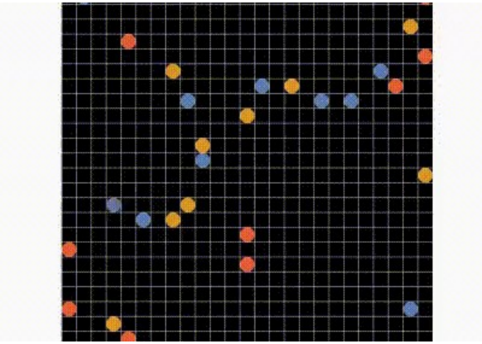
Color is not considered here

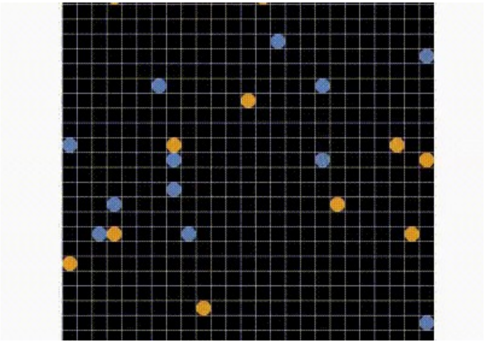
$$r_{\mathrm{RaIR}}(s) := -\mathcal{H}(\Phi(s)) = \sum_{x \in X} p(x) \log p(x)$$

Sancaktar, Piater, GM. @EWRL

Regularity as Intrinsic Reward

What does it do with
a perfect model?



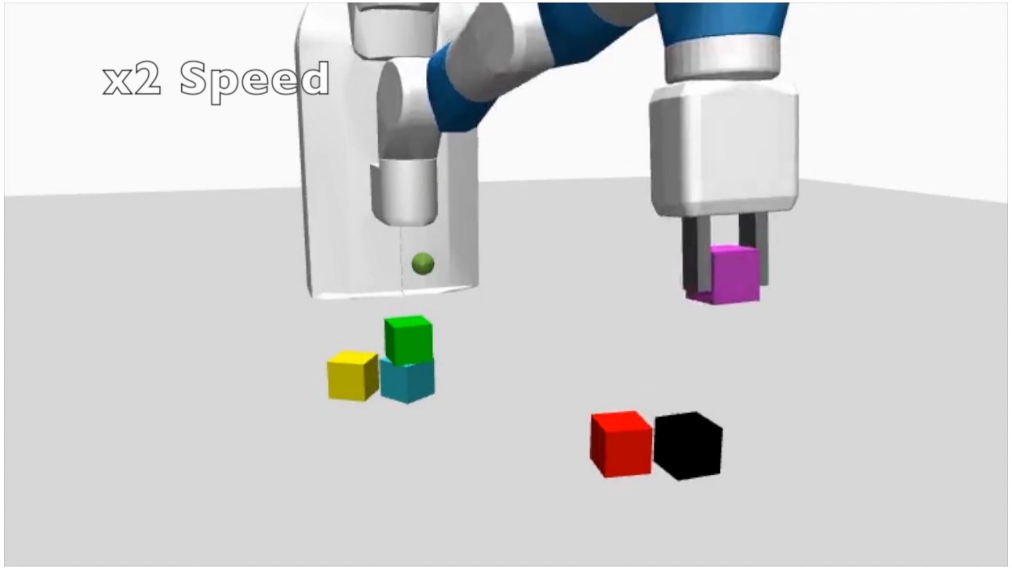


Regularity in relative position and color
Every blob is controlled one after the other.

Georg Martius <georg.martius@tue.mpg.de>

Regularity as Intrinsic Reward

What does it do with
a perfect model?

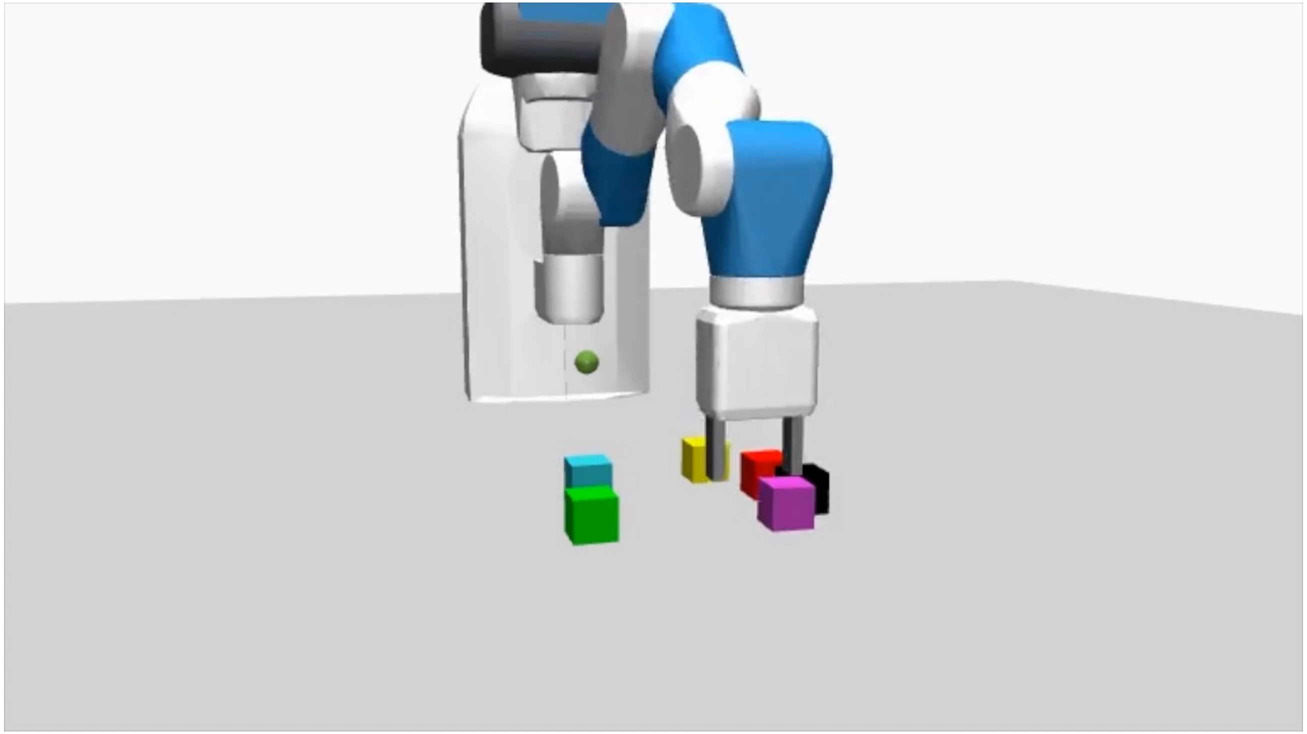


Sancaktar, Piater, GM. under review

Regularity as Intrinsic Reward

Free-play

RaIR + Info-gain

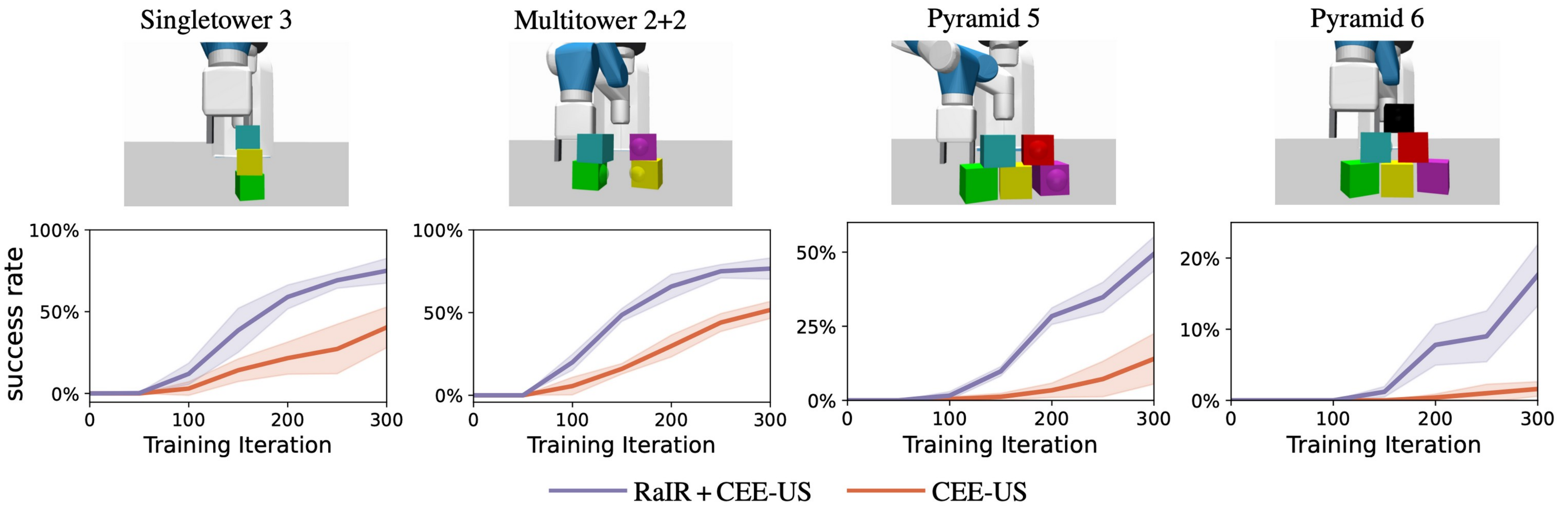


Sancaktar, Piater, GM. under review

Regularity as Intrinsic Reward

Does it help?

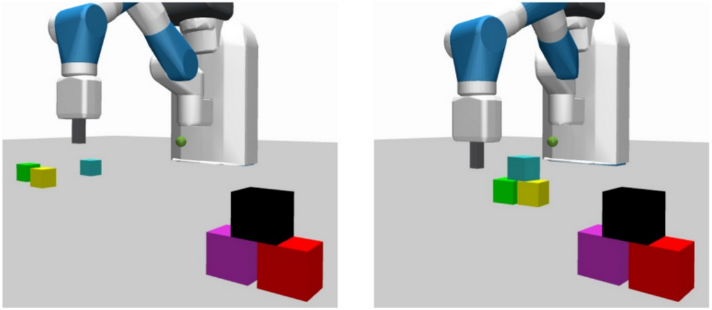
Zero-shot performance:



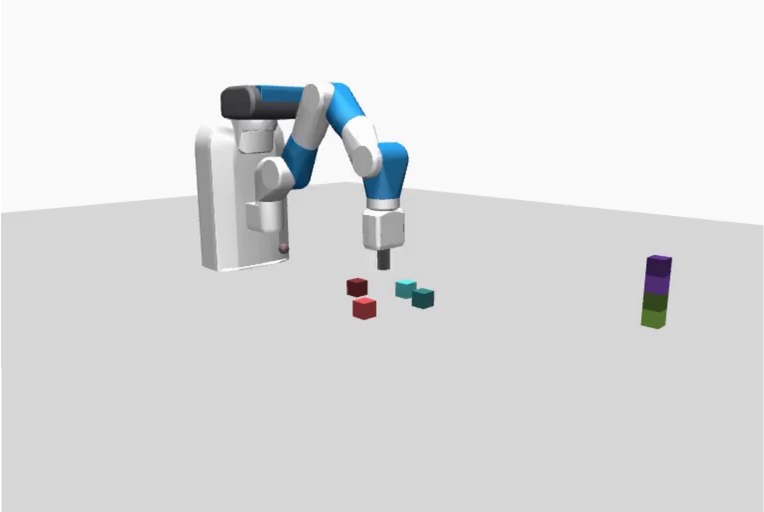
Sancaktar, Piater, GM. under review

Recreate Existing Regularities

- Initialize a regular structure outside of the robot's reach
- Just optimize for RaIR → Repeating existing regularity is an optimum



t=0 t=200



https://sites.google.com/view/rair-project

Sancaktar, Piater, GM. under review

What about Hierarchical Planning?

Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics

Christian Gumbsch*[1,2], Noor Sajid[3], Georg Martius[2], and Martin V. Butz[1]

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

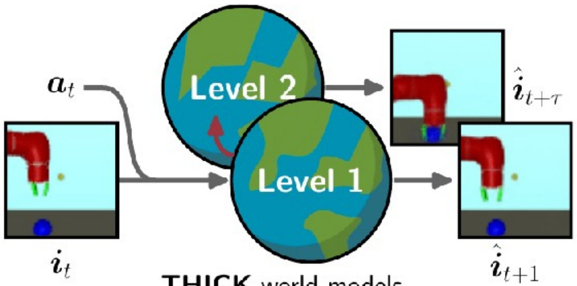
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

At a glance

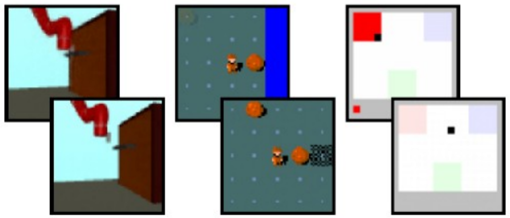
Motivation: How can we learn hierarchical world models for planning across multiple time scales?

Our contributions:

- **C-RSSM**: Context-encoding RSSM [3] extension
- **THICK**: Algorithm to learn world model hierarchy with an adaptive high-level time scale
- **Improved long-horizon planning** when using THICK world models for MBRL and MPC agents



THICK world models



exemplary hierarchical predictions

website

Poster today

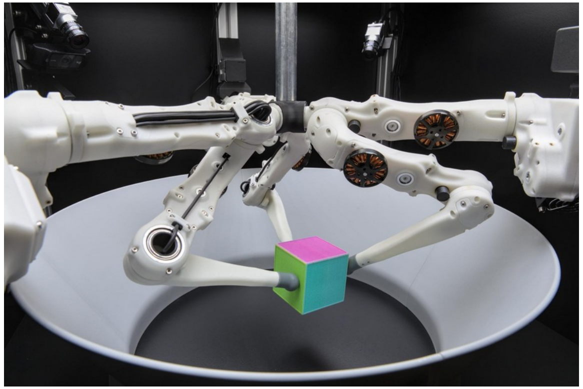
Georg Martius <georg.martius@tue.mpg.de>

Summary

- ➢ Intrinsic motivations help us to formalize exploration strategies

  inductive bias to specify downstream task families

- ➢ Model-based planning + predictive intrinsic motivation is promising
- ➢ Regularity as an addition to the intrinsic motivation zoo ;-)
- ➢ We are close to have playing robots that become useful?!









Thank you!

Tübingen AI Center

imprs-is

CyberValley

MAX PLANCK GESELLSCHAFT

VolkswagenStiftung

machine learning new perspectives for science

Robust Learning

erc