Penn Engineering

GRASP LABORATORY

WASHINGTON STATE UNIVERSITY

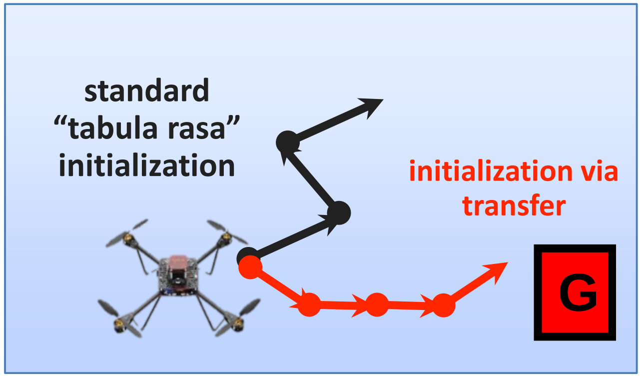
Lifelong Learning for Disturbance Rejection on Mobile Robots

David Isele, José Marcio Luna, Eric Eaton,
Gabriel V. de la Cruz, James Irwin,
Brandon Kallaher, Matthew E. Taylor

Motivation

Problem 1: Without prior knowledge, RL in a new task is slow

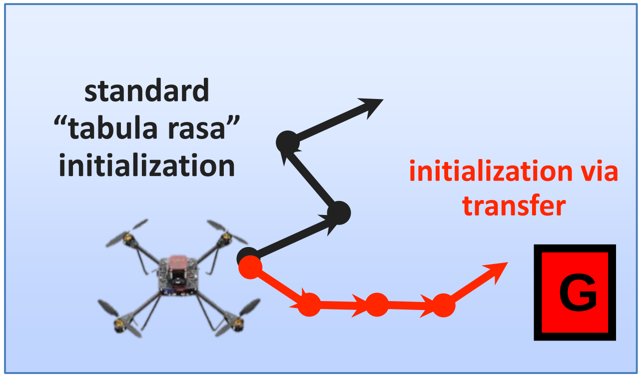
Idea: Reuse knowledge from previously learned tasks

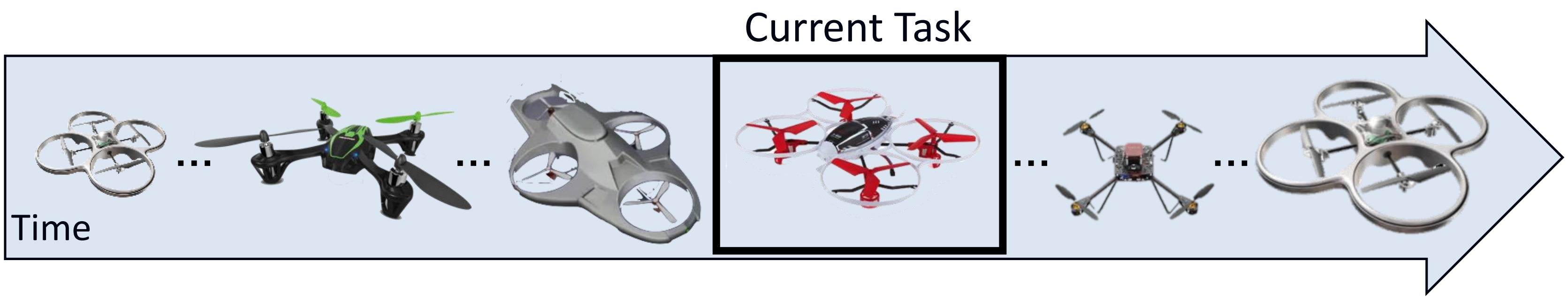


Motivation

Problem 1: Without prior knowledge, RL in a new task is slow

Idea: Reuse knowledge from previously learned tasks



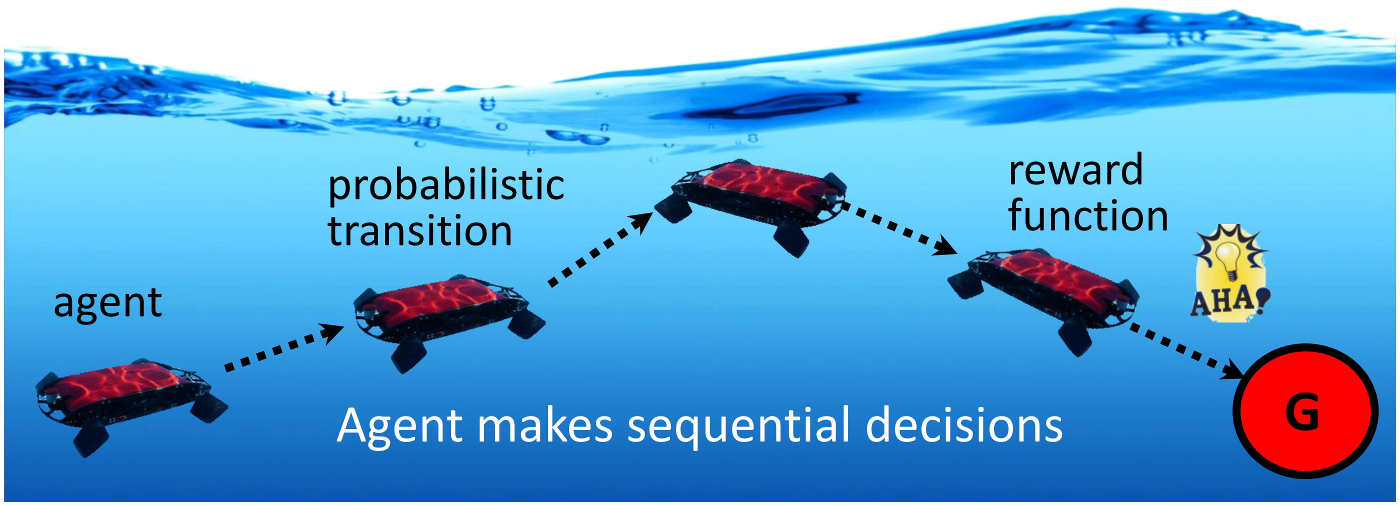


We focus on the **lifelong learning** case:
Agent learns multiple tasks consecutively
Want stability guarantees as the number of tasks grows large

Background

Background: Policy Gradient Methods for Control

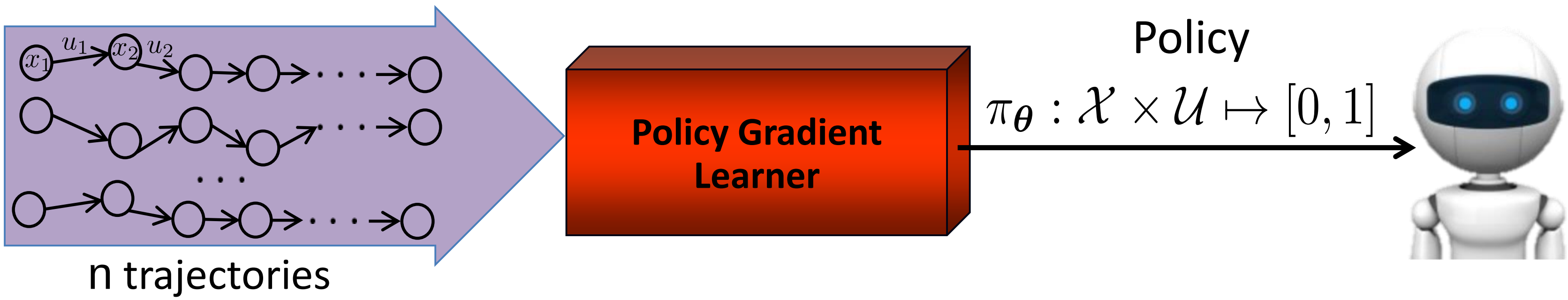
- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control [Kober & Peters 2011; Peters & Schaal 2008; Sutton et al. 2000]



- Formalized as a Markov Decision Process (MDP)

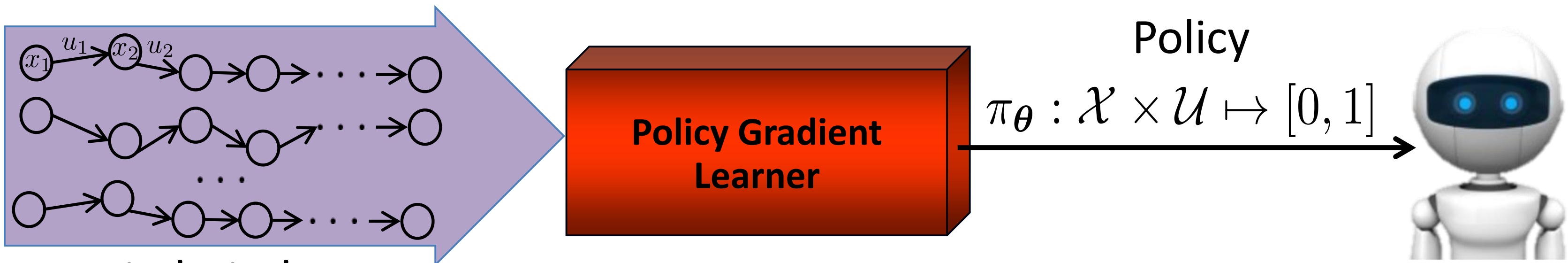
Background: Policy Gradient Methods for Control

- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control
  - [Kober & Peters 2011; Peters & Schaal 2008; Sutton et al. 2000]



Background: Policy Gradient Methods for Control

- Agent interacts with environment, taking consecutive actions
- PG methods support continuous state and action spaces
  - Have shown recent success in applications to robotic control
  - [Kober & Peters 2011; Peters & Schaal 2008; Sutton et al. 2000]



n trajectories

Goal: find policy π_θ that minimizes $\mathcal{J}(\boldsymbol{\theta}) = \int_{\mathbb{T}} p_\theta(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})d\boldsymbol{\tau}$

$$p_\theta(\boldsymbol{\tau}) = p_0(\mathbf{x}_0) \prod_{h=1}^{H} p(\mathbf{x}_{h+1} | \mathbf{x}_h, \mathbf{a}_h)\pi_{\boldsymbol{\theta}}(\mathbf{a}_h | \mathbf{x}_h)$$

probability of trajectory

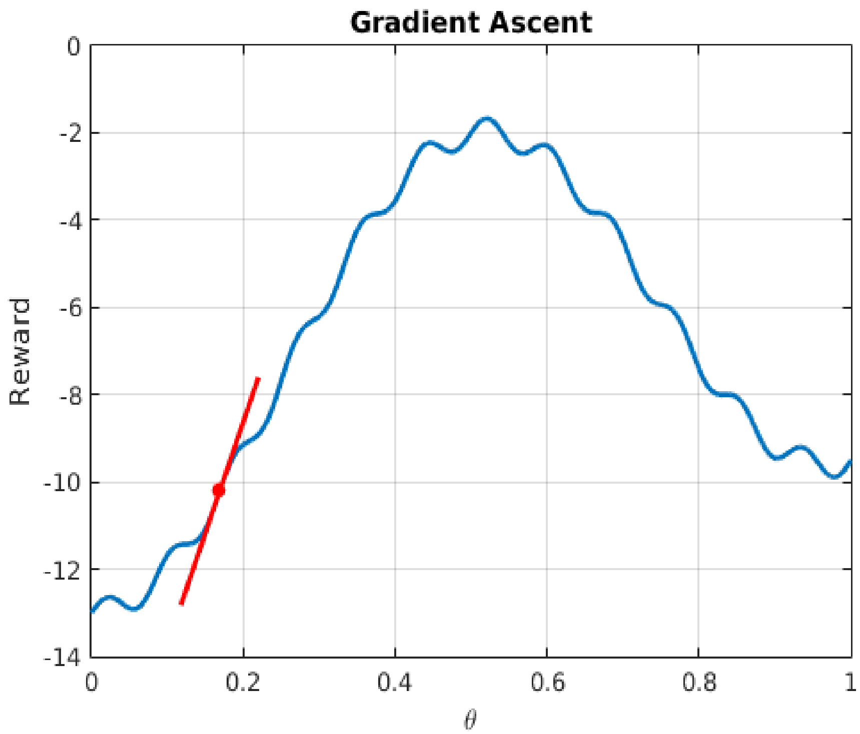
$$\mathcal{R}(\boldsymbol{\tau}) = \frac{1}{H} \sum_{h=0}^{H} r_{h+1}$$

reward function

Background: Finite Difference Policy Gradients

Approximate the change in reward with sampled disturbances

$$\Delta \mathcal{J} \approx \mathcal{J}(\theta + \Delta\theta) - \mathcal{J}(\theta)$$



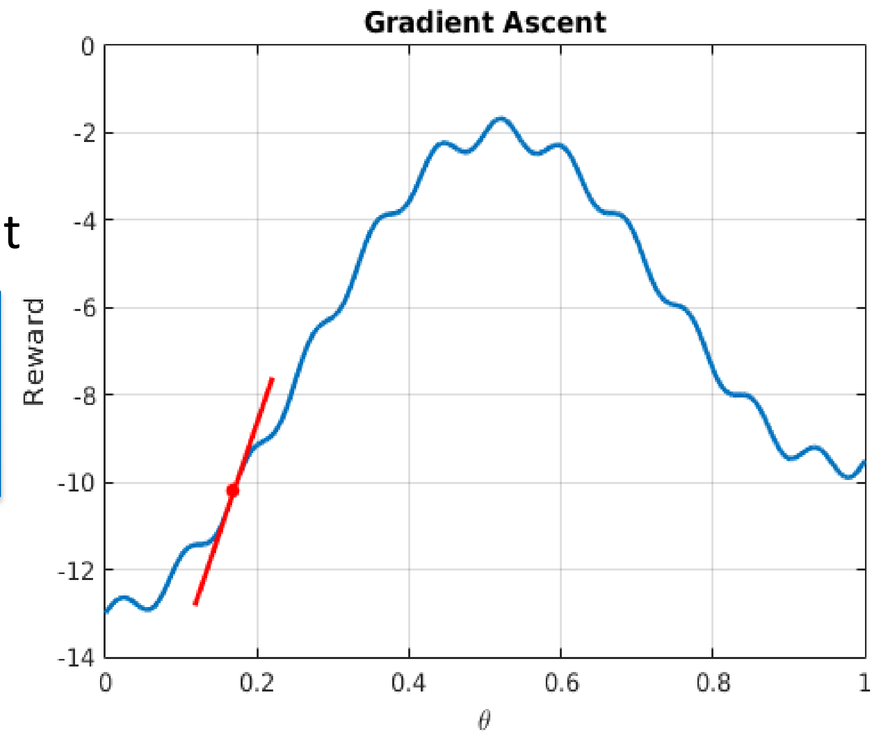
Background: Finite Difference Policy Gradients

Approximate the change in reward with sampled disturbances

$$\Delta\mathcal{J} \approx \mathcal{J}(\theta + \Delta\theta) - \mathcal{J}(\theta)$$

Use the pseudo-inverse to find the gradient

$$\frac{d\mathcal{J}}{d\theta} = (\Delta\theta^\top \Delta\theta)^{-1} \Delta\theta^\top \Delta\mathcal{J}$$



Background: Finite Difference Policy Gradients

Approximate the change in reward with sampled disturbances

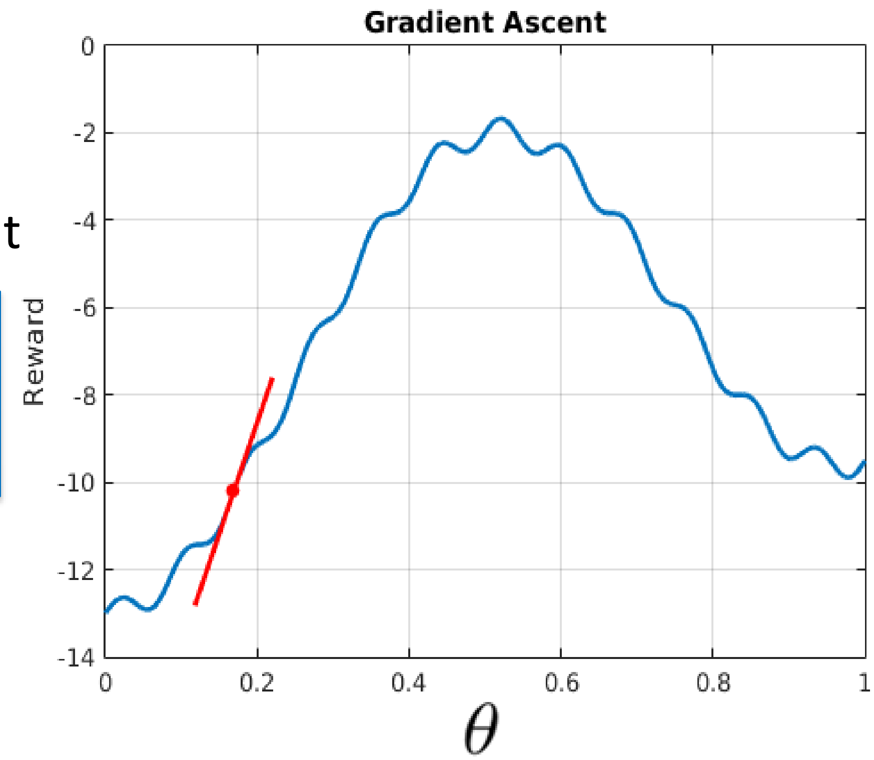
$$\Delta\mathcal{J} \approx \mathcal{J}(\theta + \Delta\theta) - \mathcal{J}(\theta)$$

Use the pseudo-inverse to find the gradient

$$\frac{d\mathcal{J}}{d\theta} = (\Delta\theta^\top \Delta\theta)^{-1} \Delta\theta^\top \Delta\mathcal{J}$$

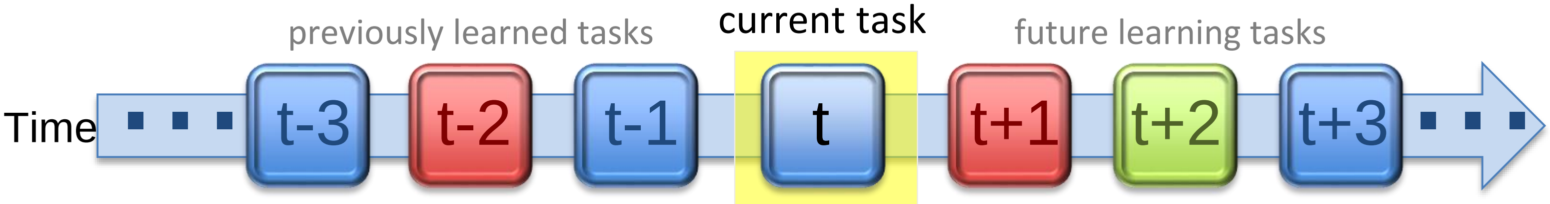
Update the current policy

$$\theta \leftarrow \theta + \alpha \frac{d\mathcal{J}}{d\theta}$$

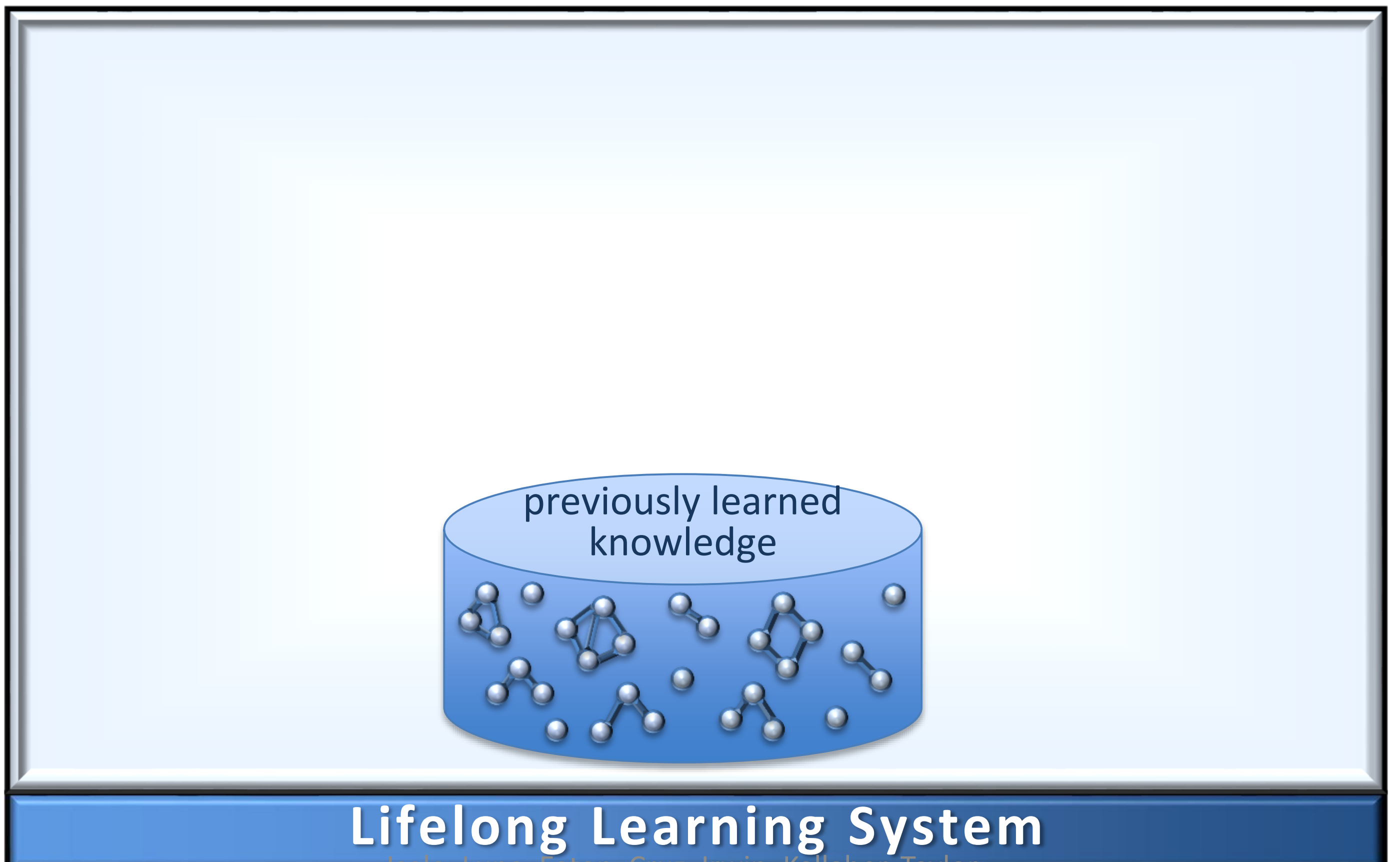


Lifelong PG Learning

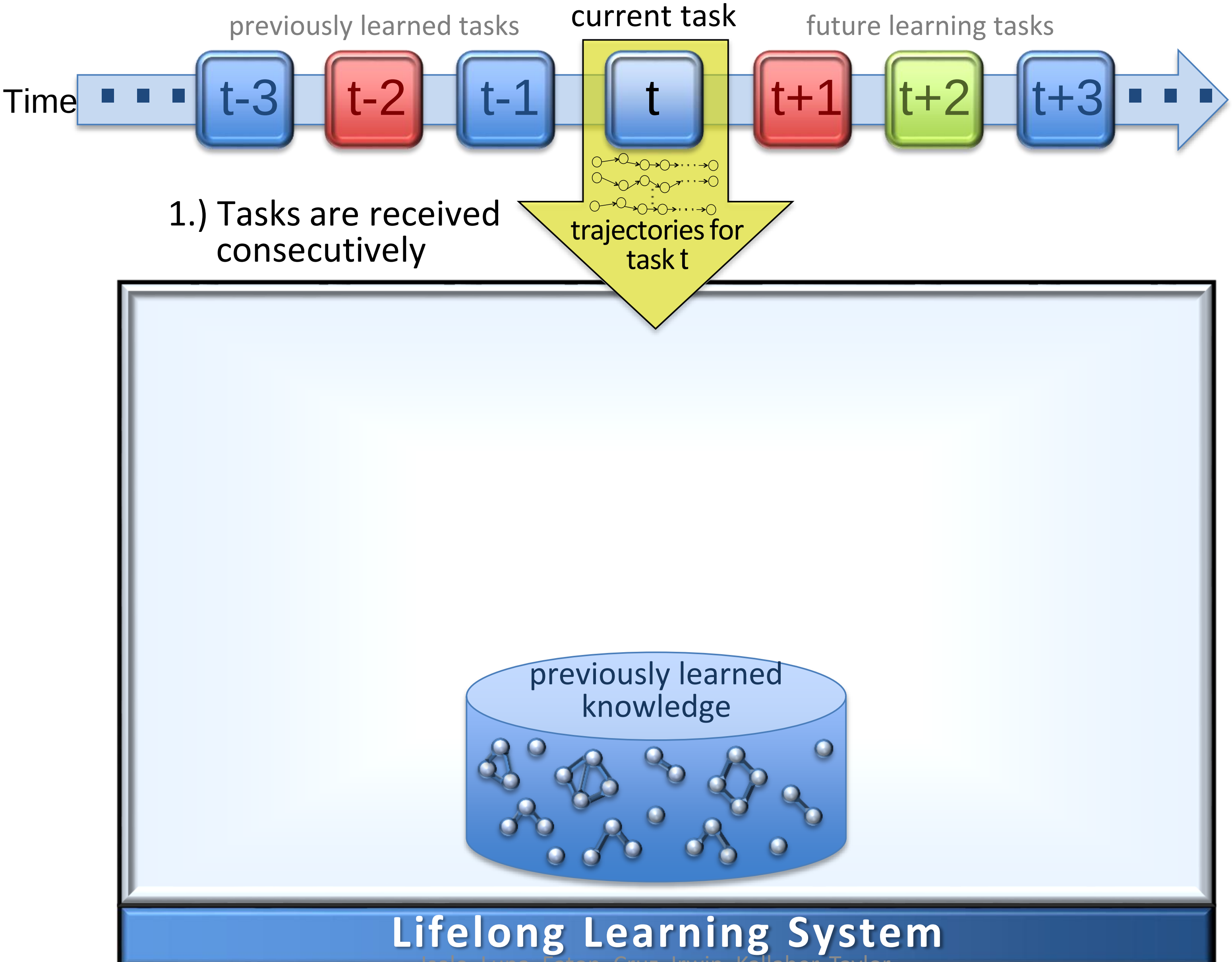
Lifelong Machine Learning



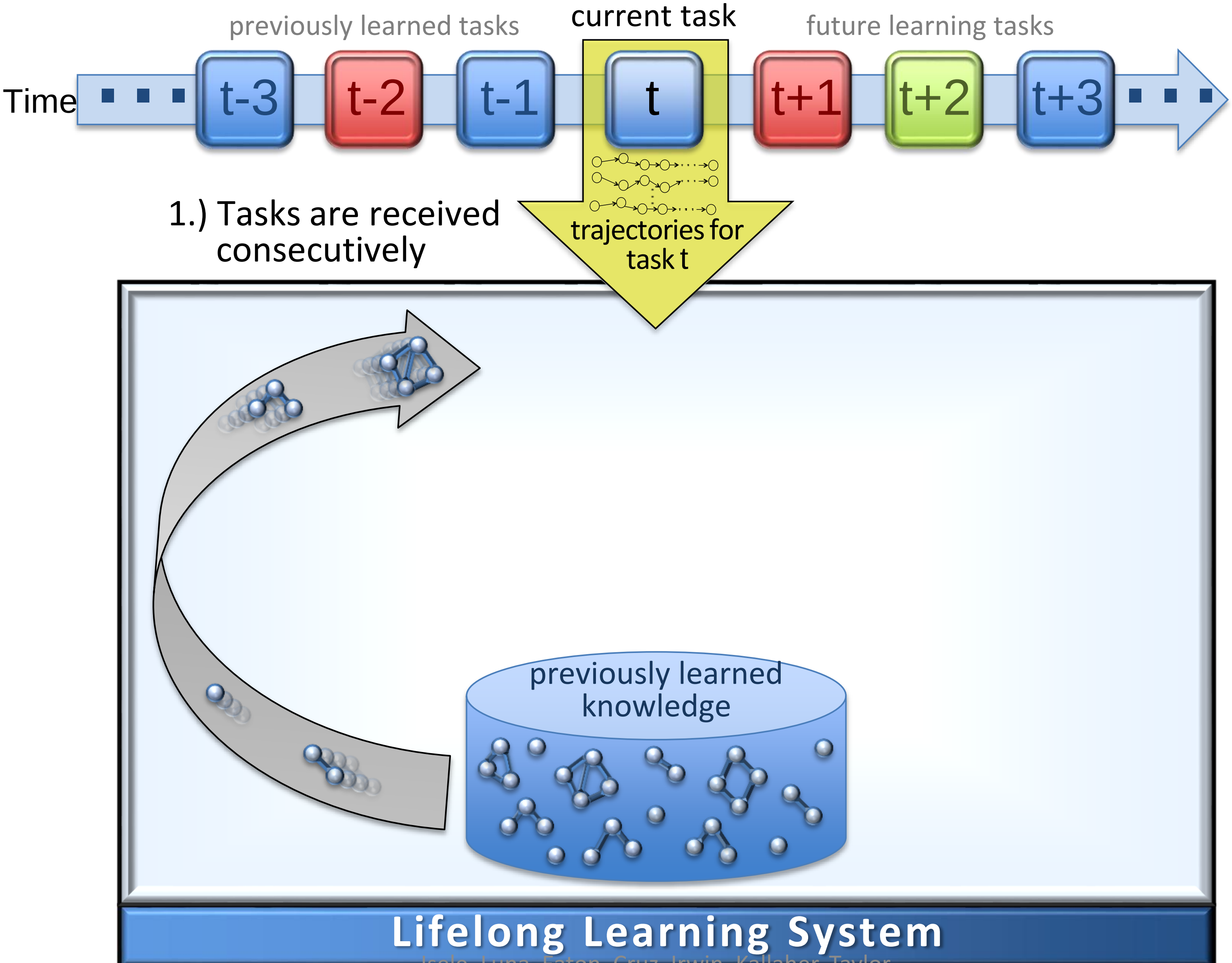
1.) Tasks are received consecutively



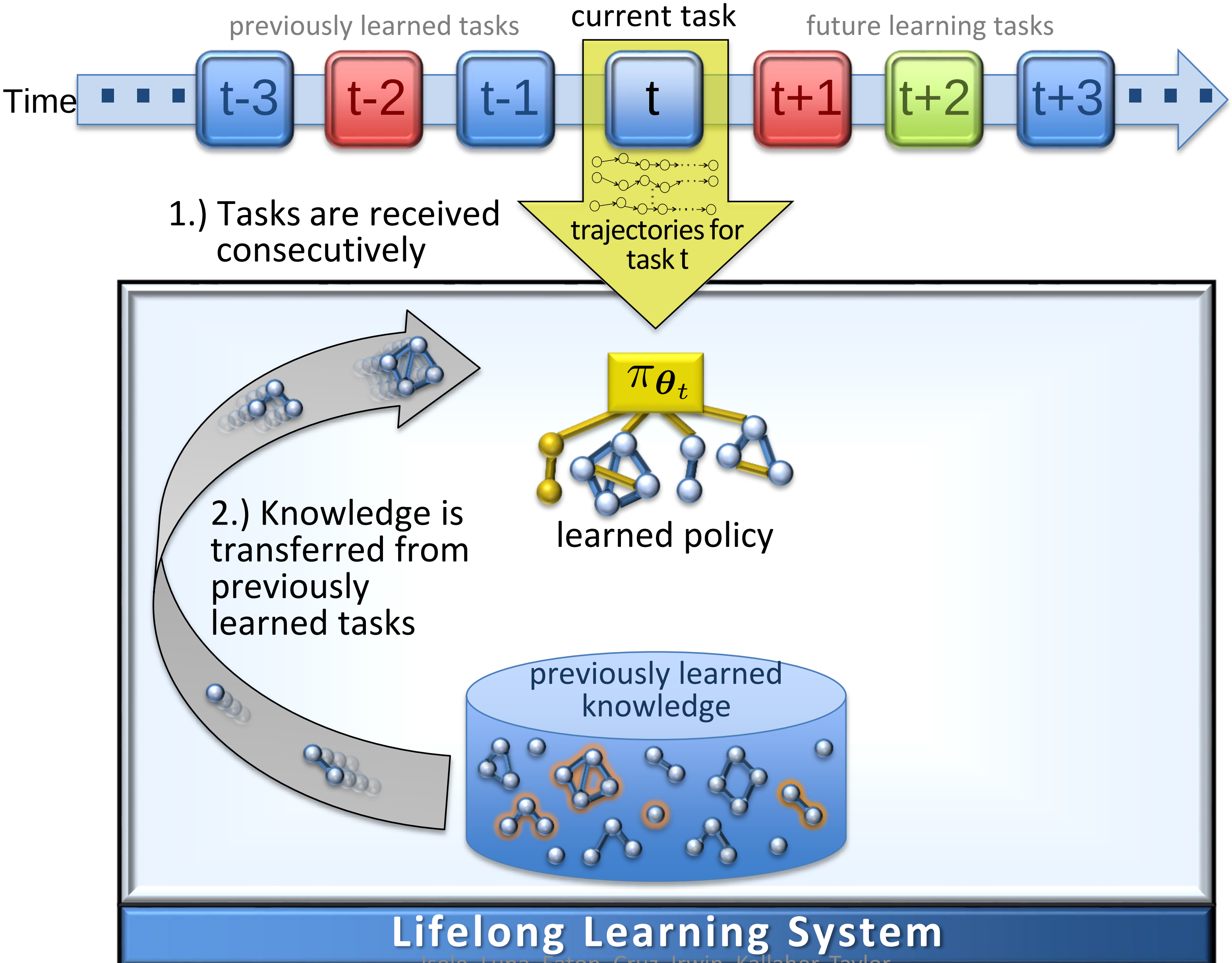
Lifelong Machine Learning



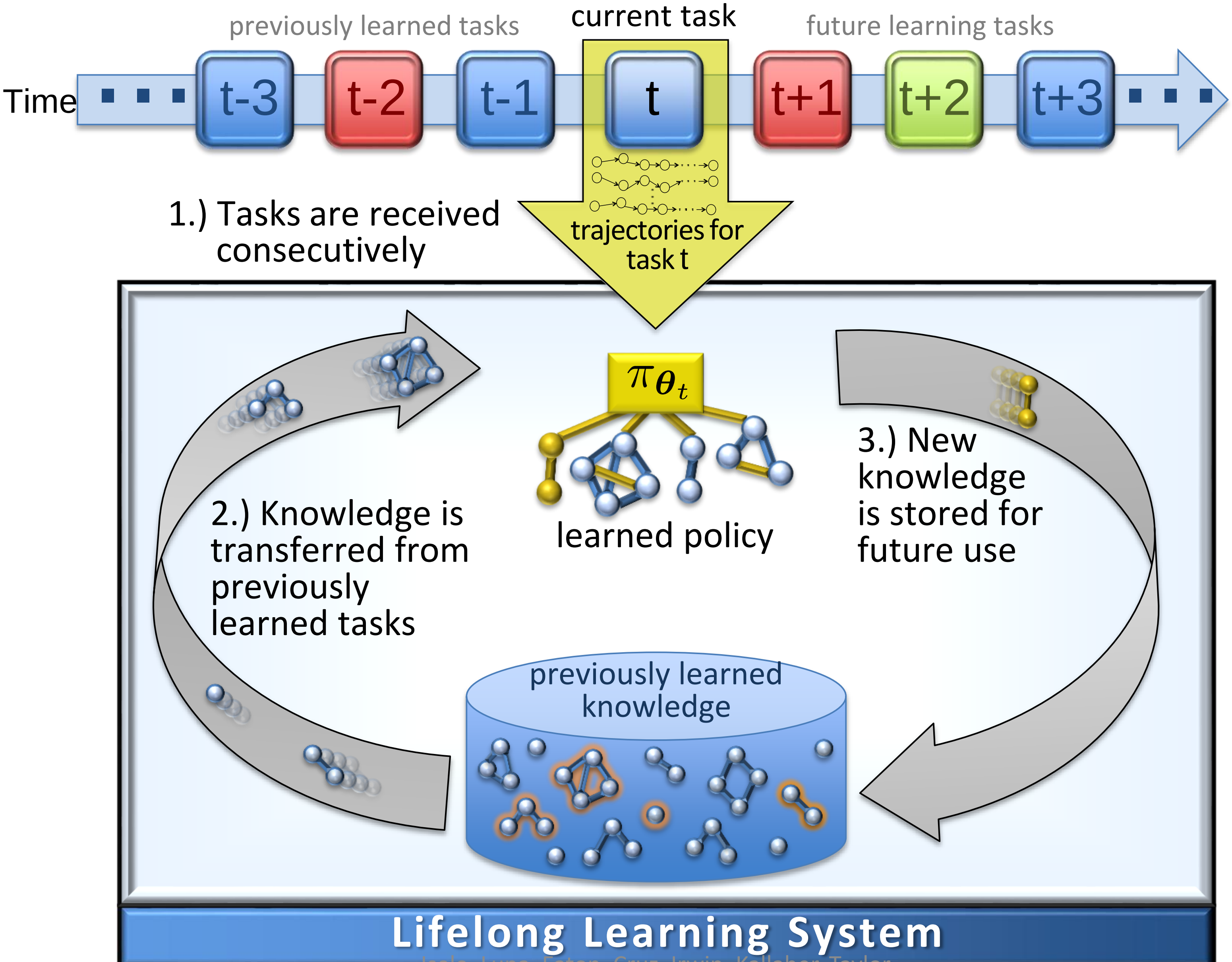
Lifelong Machine Learning



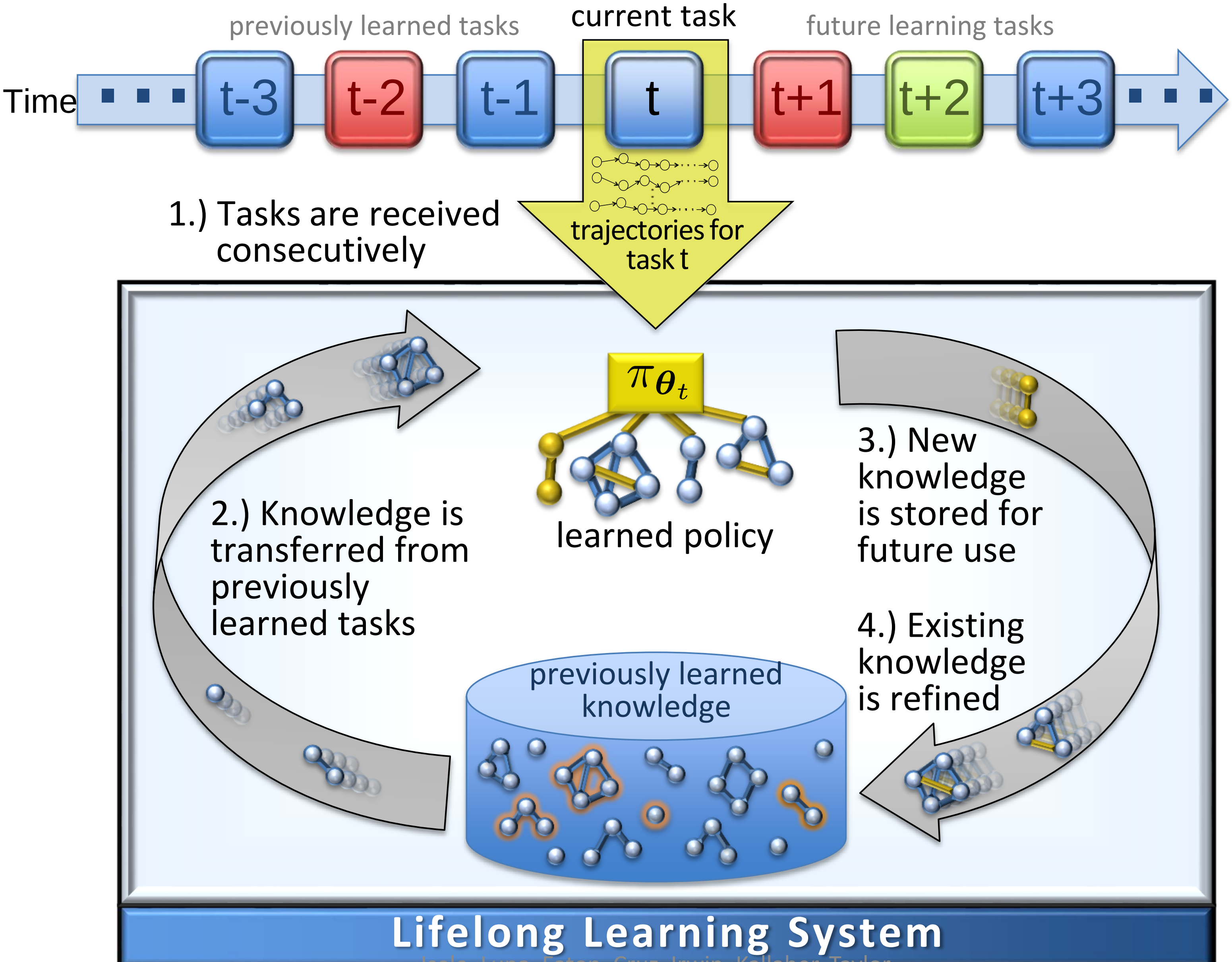
Lifelong Machine Learning



Lifelong Machine Learning



Lifelong Machine Learning



Isele, Luna, Eaton, Cruz, Irwin, Kallaher, Taylor

PG-ELLA Objective

Issue: the objective is dependent on <u>all</u> trajectories

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \left[-\mathcal{J}\left(\boldsymbol{\theta}^{(t)}\right) + \mu \left|\left|\mathbf{s}^{(t)}\right|\right|_1 \right] + \lambda ||\mathbf{L}||_{\mathsf{F}}^2$$

PG-ELLA Objective

Issue: the objective is dependent on <u>all</u> trajectories

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \left[-\mathcal{J}\left(\boldsymbol{\theta}^{(t)}\right) + \mu \left|\left|\mathbf{s}^{(t)}\right|\right|_1 \right] + \lambda ||\mathbf{L}||_{\mathsf{F}}^2$$

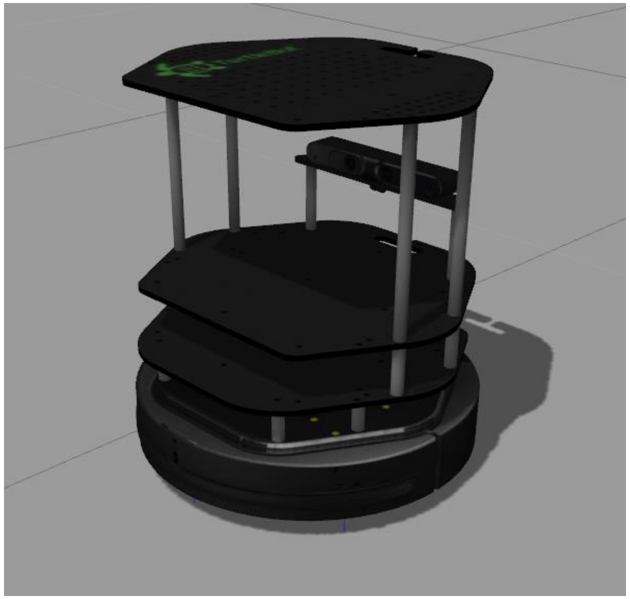
$$\hat{e}_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \left[\left|\left|\boldsymbol{\alpha}^{(t)} - \mathbf{L}\mathbf{s}^{(t)}\right|\right|_{\underset{\text{Hessian}}{\uparrow}\boldsymbol{\Gamma}^{(t)}}^2 + \mu \left|\left|\mathbf{s}^{(t)}\right|\right|_1 \right] + \lambda ||\mathbf{L}||_{\mathsf{F}}^2$$

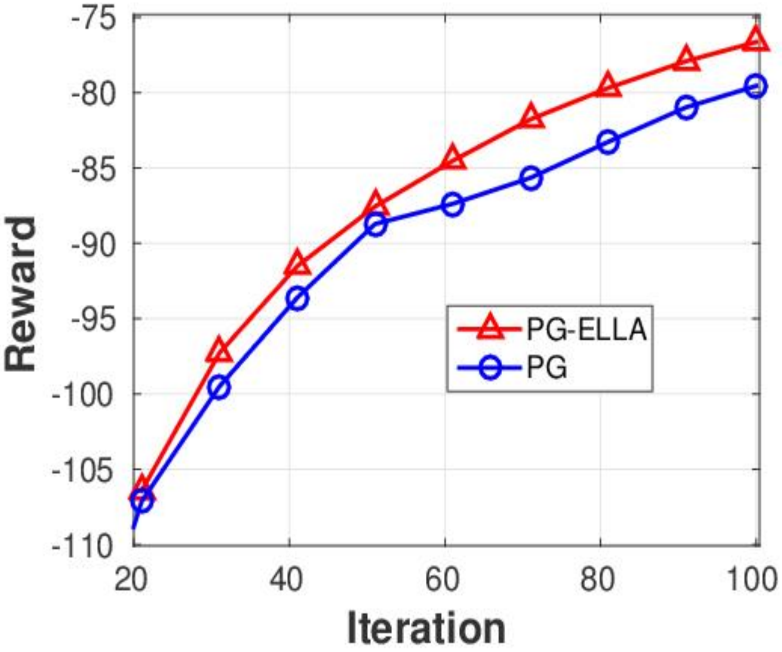
Experiments

Verification on Robots

Results for Robot Go-to-Goal Task

- Run RL on a new robot (goal and disturbance) for a small number of iterations
- Use PG-ELLA to adjust policy according to known solutions
- Continue training

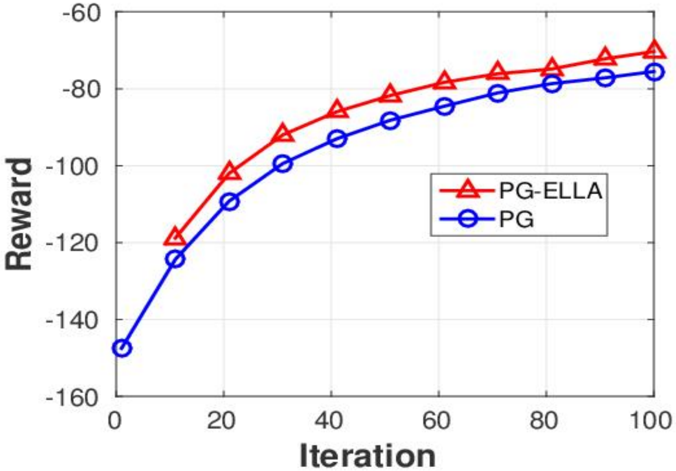


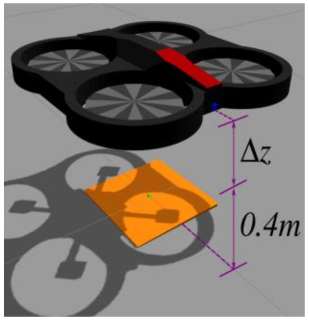


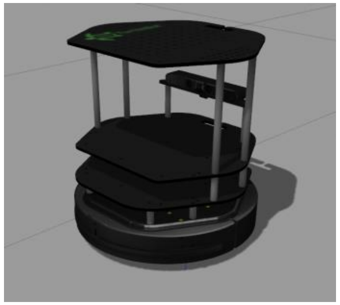
PG-ELLA improves Learning

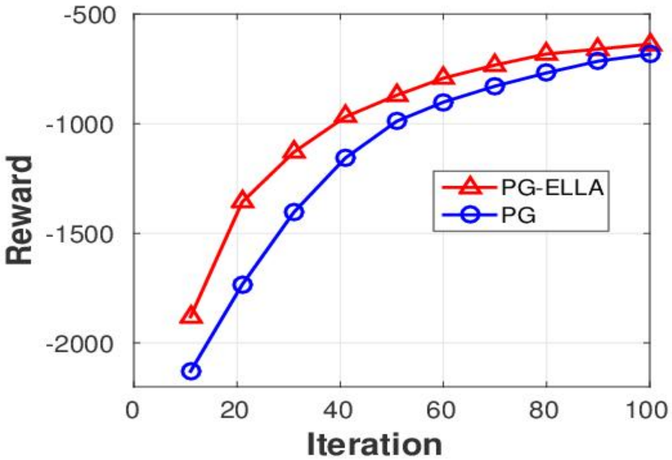
Better Results Incorporating Prior

- Initialization with average policy of other robots improves benefit









PG-ELLA improves Learning

Lifelong Learning for Disturbance Rejection on Mobile Robots

David Isele, José Marcio Luna, Eric Eaton,
Gabriel V. de la Cruz, James Irwin, Brandon Kallaher, Matthew E. Taylor

Thank you!

Questions?

This research was supported by ONR N00014-11-1-0139, AFRL FA8750-14-1-0069, AFRL FA8750-14-1-0070, NSF IIS-1149917, NSF IIS-1319412, USDA 2014-67021-22174, and a Google Research Award.