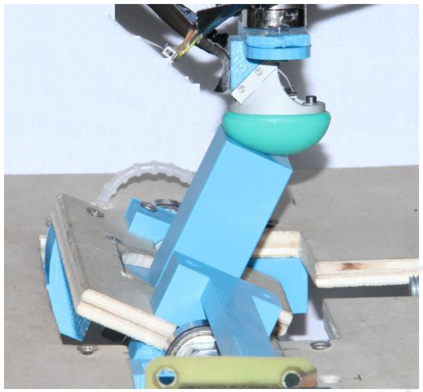
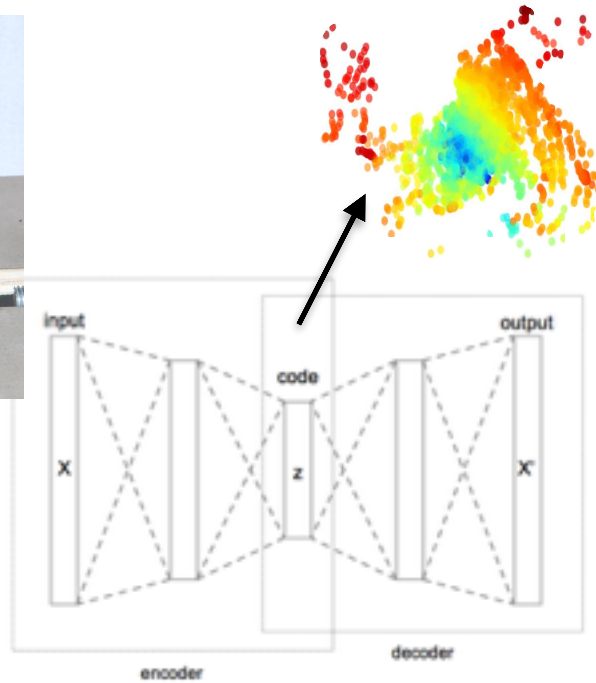
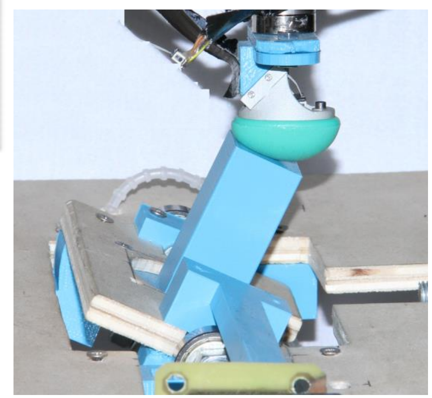
Stable Reinforcement Learning from Sensor Data







Herke van Hoof

McGill

Reinforcement learning from sensor data

Pixabay / Sasint / CC 0

Phasmatinox / Wikipedia / CC BY 3.0

Repetitive tasks
Standardized environment

Different tasks, differences within tasks
Changing environment
Feedback from sensor data

Designing a program for a single task and environment is time consuming
Changes and novelty would require continuous re-programming

Reinforcement Learning

Alternative: autonomous skills acquisition?

Reinforcement learning studies how to optimise behaviour through **trial and error**

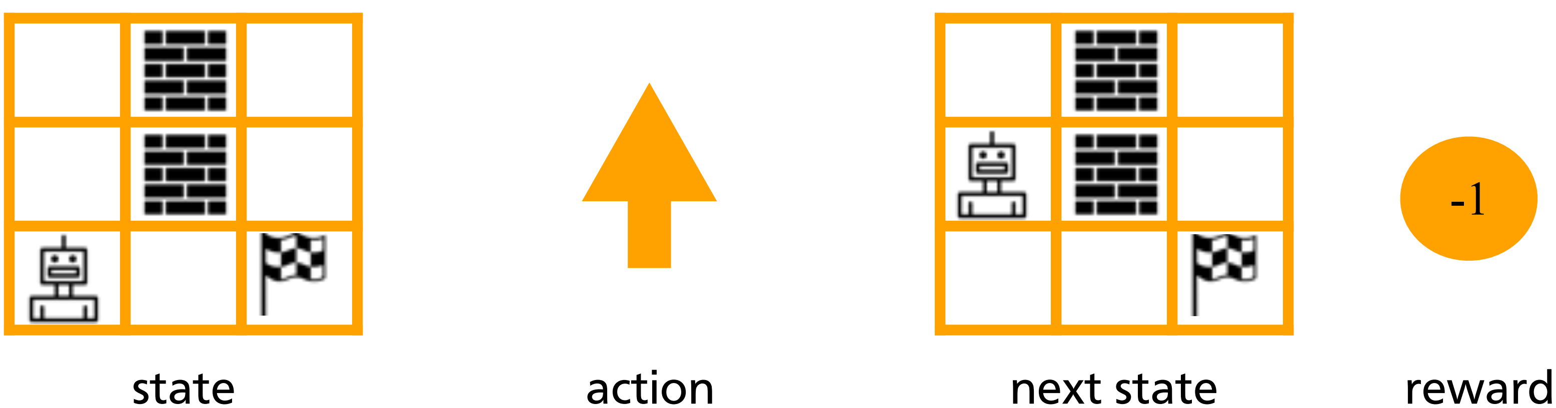
No need for a dataset with demonstrated 'correct actions'

Reinforcement Learning

Alternative: autonomous skills acquisition?

Reinforcement learning studies how to optimise behaviour through **trial and error**

No need for a dataset with demonstrated 'correct actions'



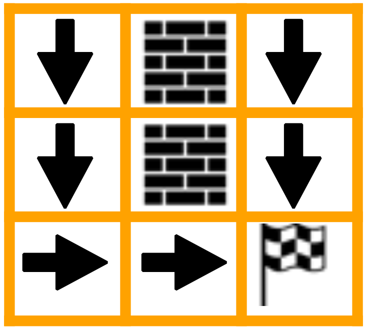
Reinforcement Learning

Alternative: autonomous skills acquisition?

Reinforcement learning studies how to optimise behaviour through **trial and error**

No need for a dataset with demonstrated 'correct actions'

Limitation: Classical algorithms assume states discrete or low-dimensional vector. Can't handle e.g. image input.



Optimal policy

Deep Reinforcement Learning

Deep RL: use deep networks as representation of e.g. policy

Can learn complex task directly from sensor data

- Human-level control in Atari games
  [Mnih et al., 2015]
- Control of complex simulated robots
  [Schulman et al., 2016]
- World-champion in boardgames Go
  [Silver et al., 2016]

However, end-to-end learning tends to be data-hungry

(Bellemare et al., 2013)

(Schulman et al., 2016)

Small batches of complex sensor data

Goal: Learn a policy $\pi(\mathbf{a}|\mathbf{s})$ that maximizes expected reward

- Policy Evaluation: Estimate expected long-term reward (value)
- Policy Improvement: Take actions that lead to best long-term reward

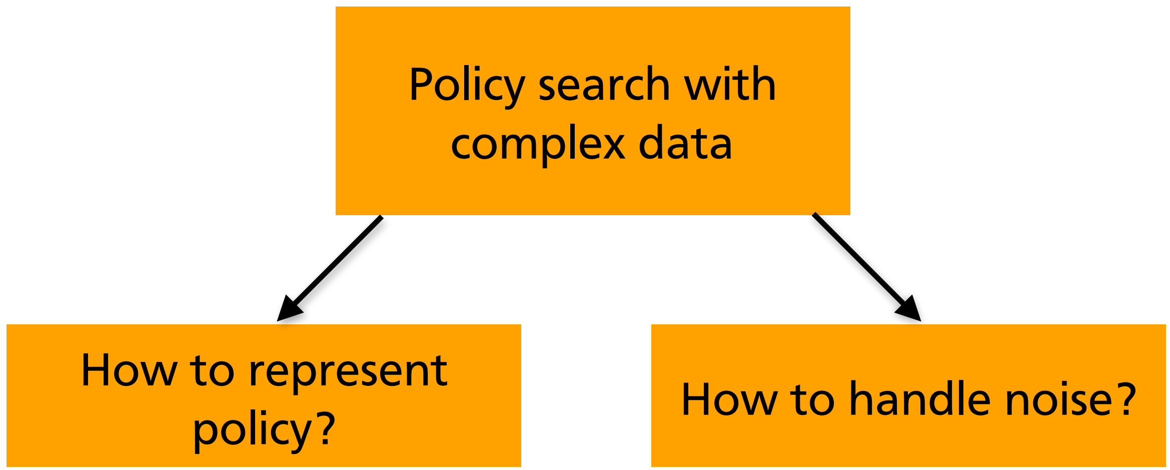
With small batches of data, **overfitting** is always a risk

- Estimation of expected reward can be imprecise
- Greedy maximization risks instability, premature convergence

Policy search methods limit the change to policy per iteration

- Policy gradient methods
- Trust region policy optimization
- **Relative entropy policy search**

Small batches of complex sensor data



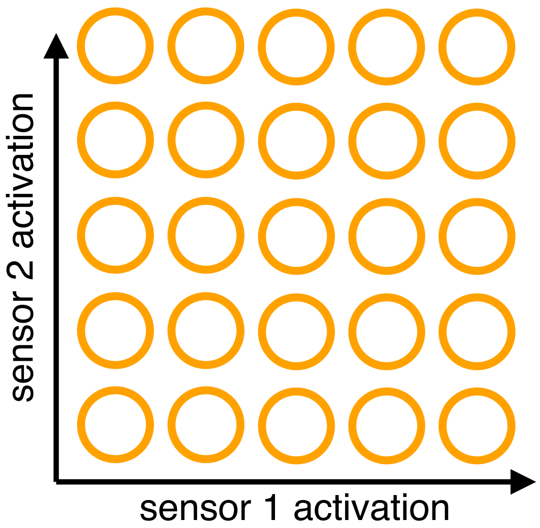
Policy representation

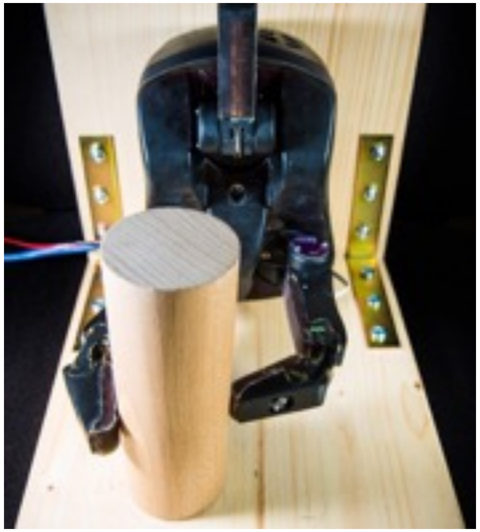
Linear representations require little data $a = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{s}) + \epsilon$

Need to 'design' task specific features?

Local basis functions are highly flexible!

However, might need many to cover the space….





Policy representation

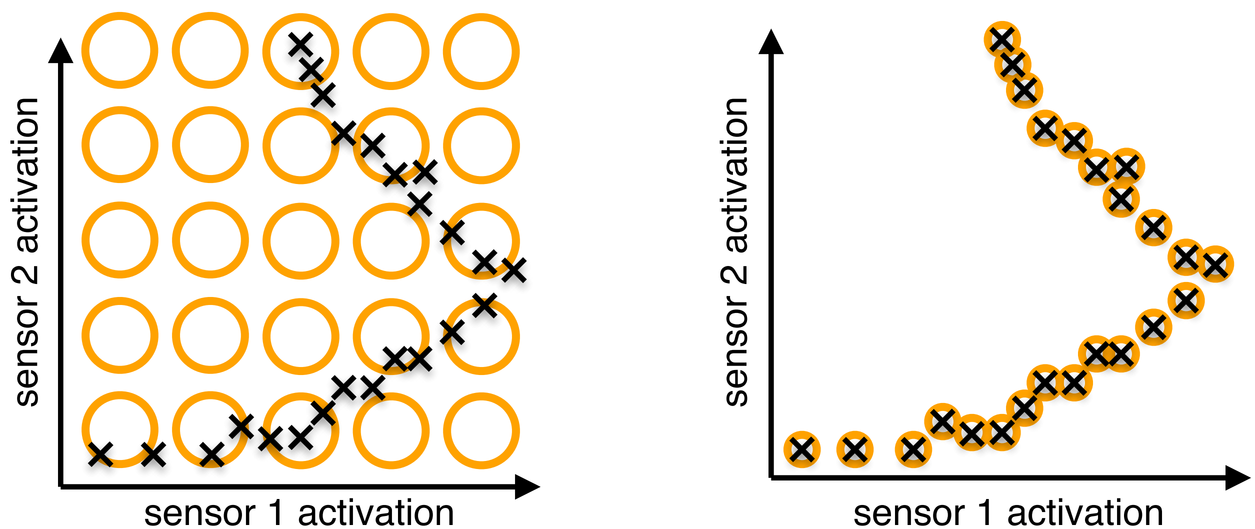
Luckily, dimensions often not independent

Non-parametric (NP) methods represent relevant area

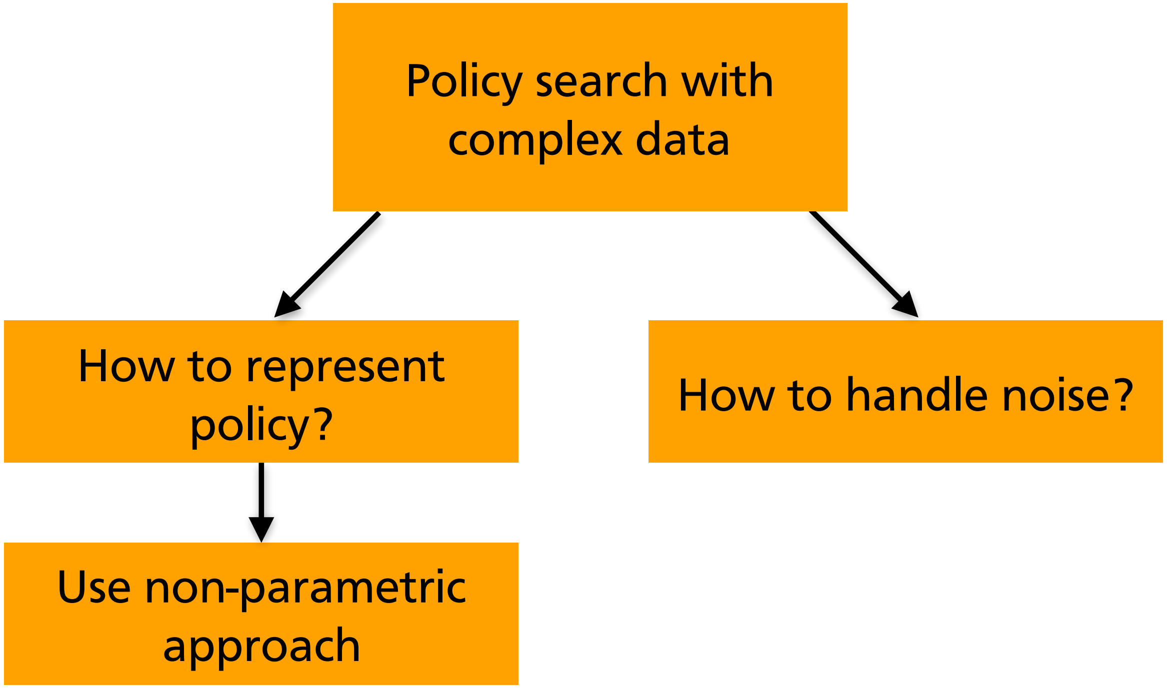
(Grünewälder et al., 2012; Pazis & Parr, 2011; Deisenroth & Rasmussen, 2011; Bagnell & Schneider, 2003; ...)

NP representation of policy and expected long-term reward

(van Hoof, Peters and Neumann, AIstats 2015; JMLR 2017)



Small batches of complex data



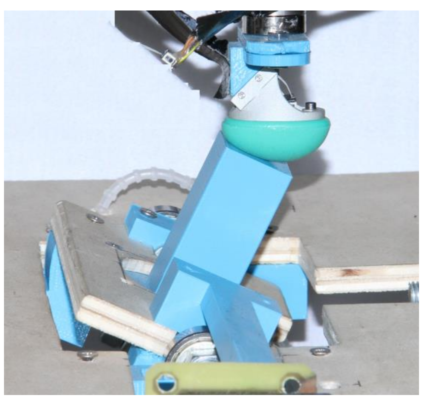
Noise in high-dimensional sensor data

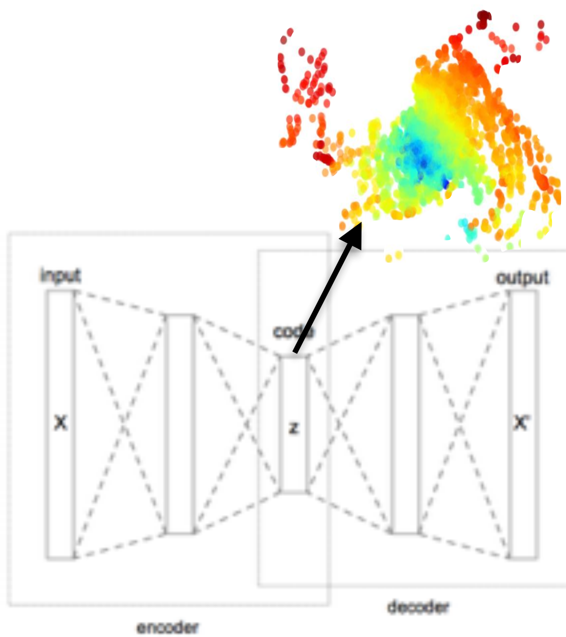
Non-parametric method needs little data, relies on distance

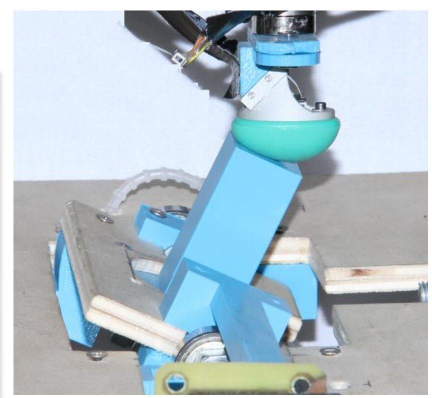
Distances are perturbed by sensor noise, especially in high-d

(Variational) auto-encoder (VAE) learns low-d representation

(Kingma & Welling, 2013)



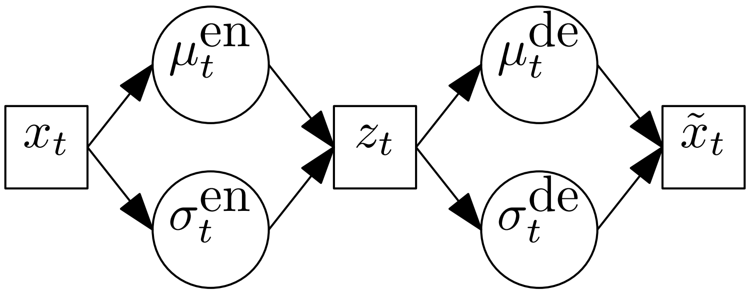




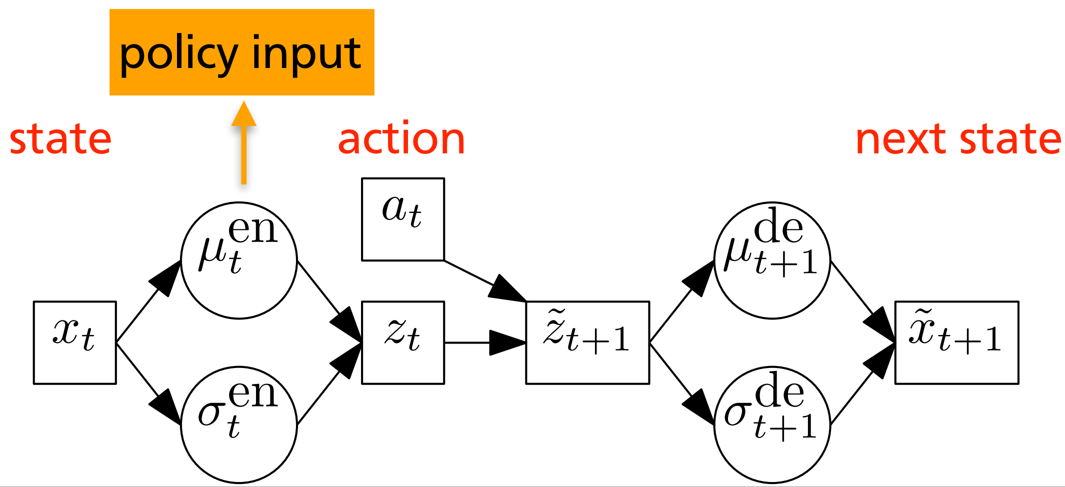
Modified variational auto-encoder

Variational auto-encoder

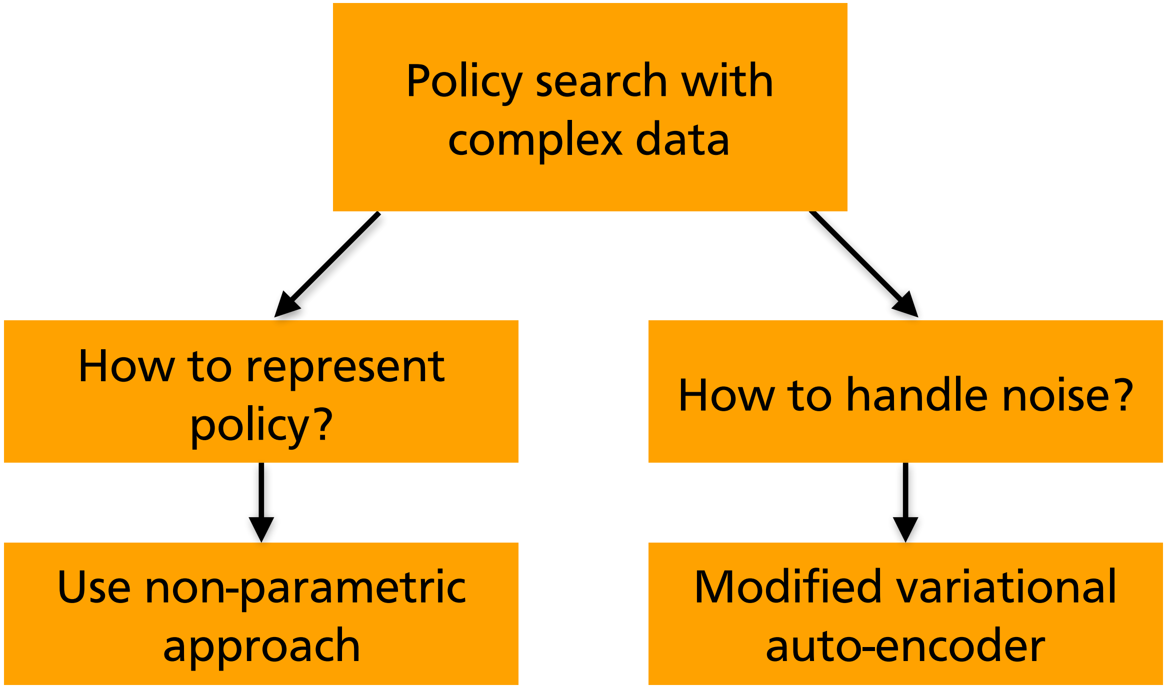
(Kingma & Welling, 2013)



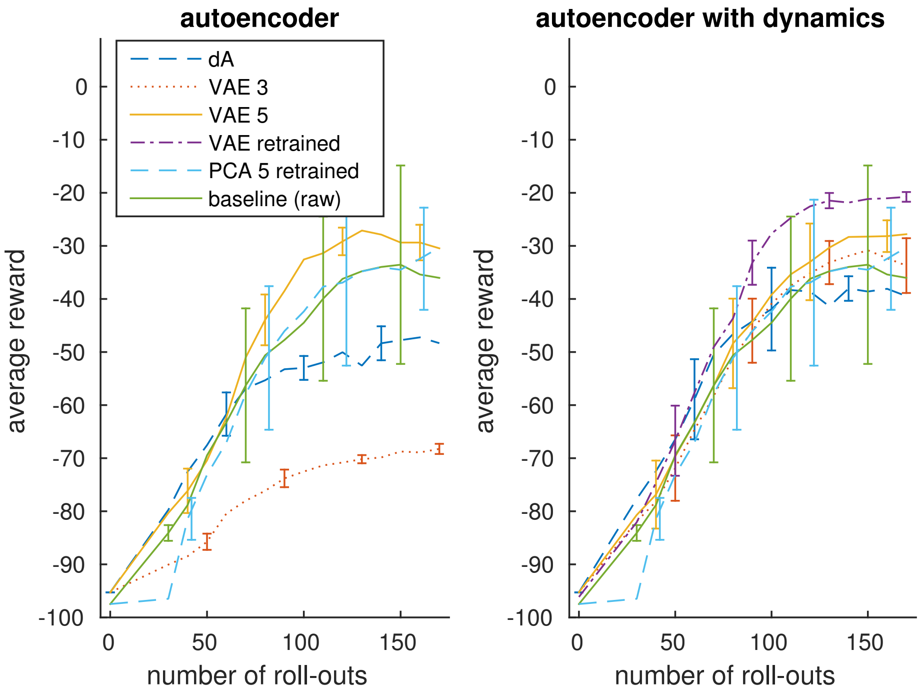
Add problem structure to learn with ‘small data’

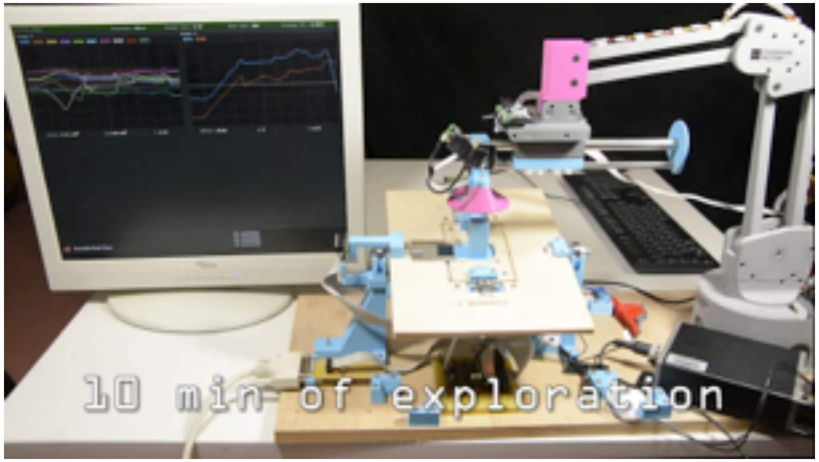


Small batches of complex data



Experiments with sensory data





Pendulum swing-up with visual input

Stabilization with tactile input

Van Hoof, Chen, Karl, Van der Smagt and Peters, Stable Reinforcement Learning with Autoencoders for Tactile and Visual Data, IROS 2016.

Image-based pendulum swing-up

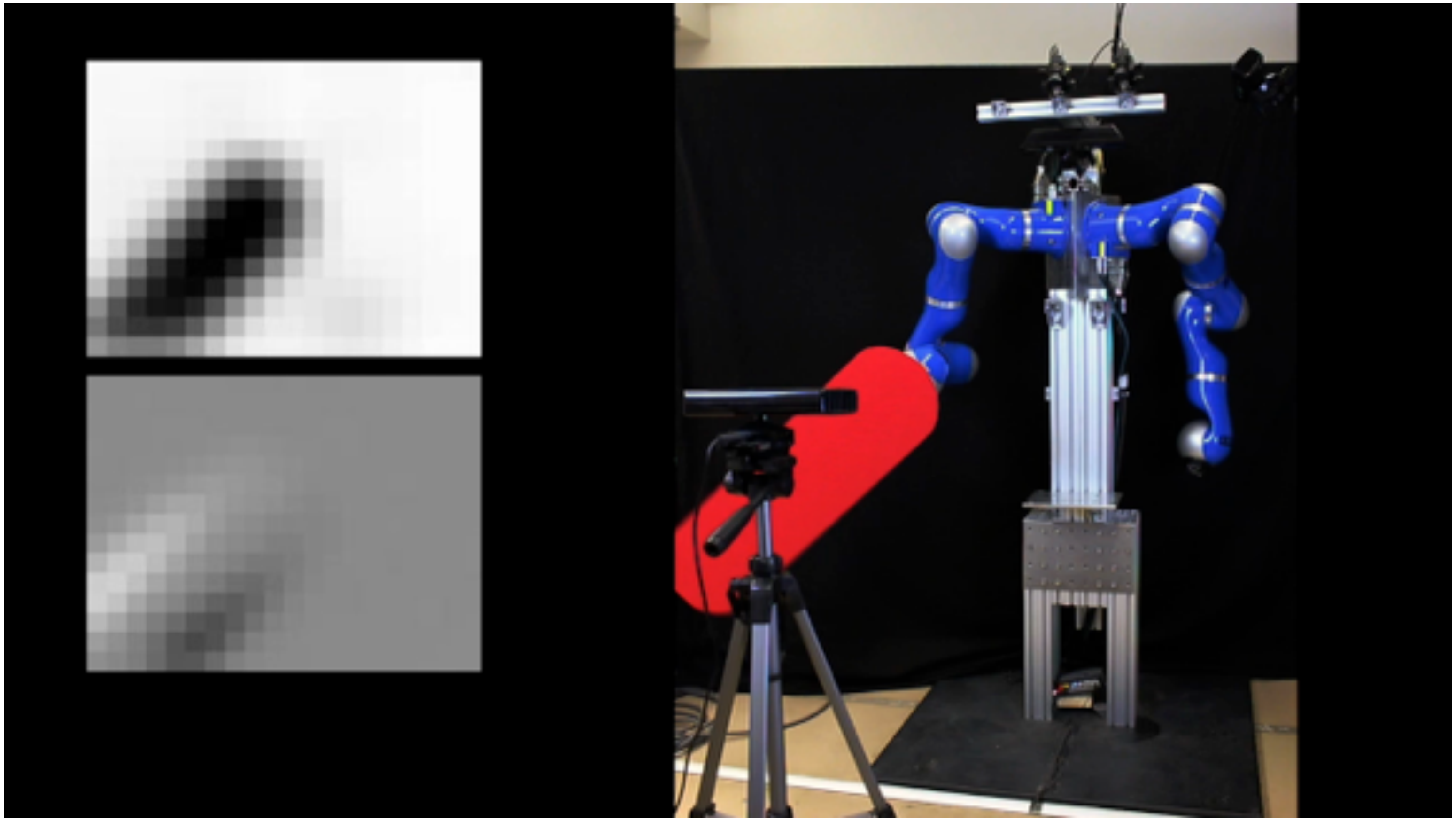
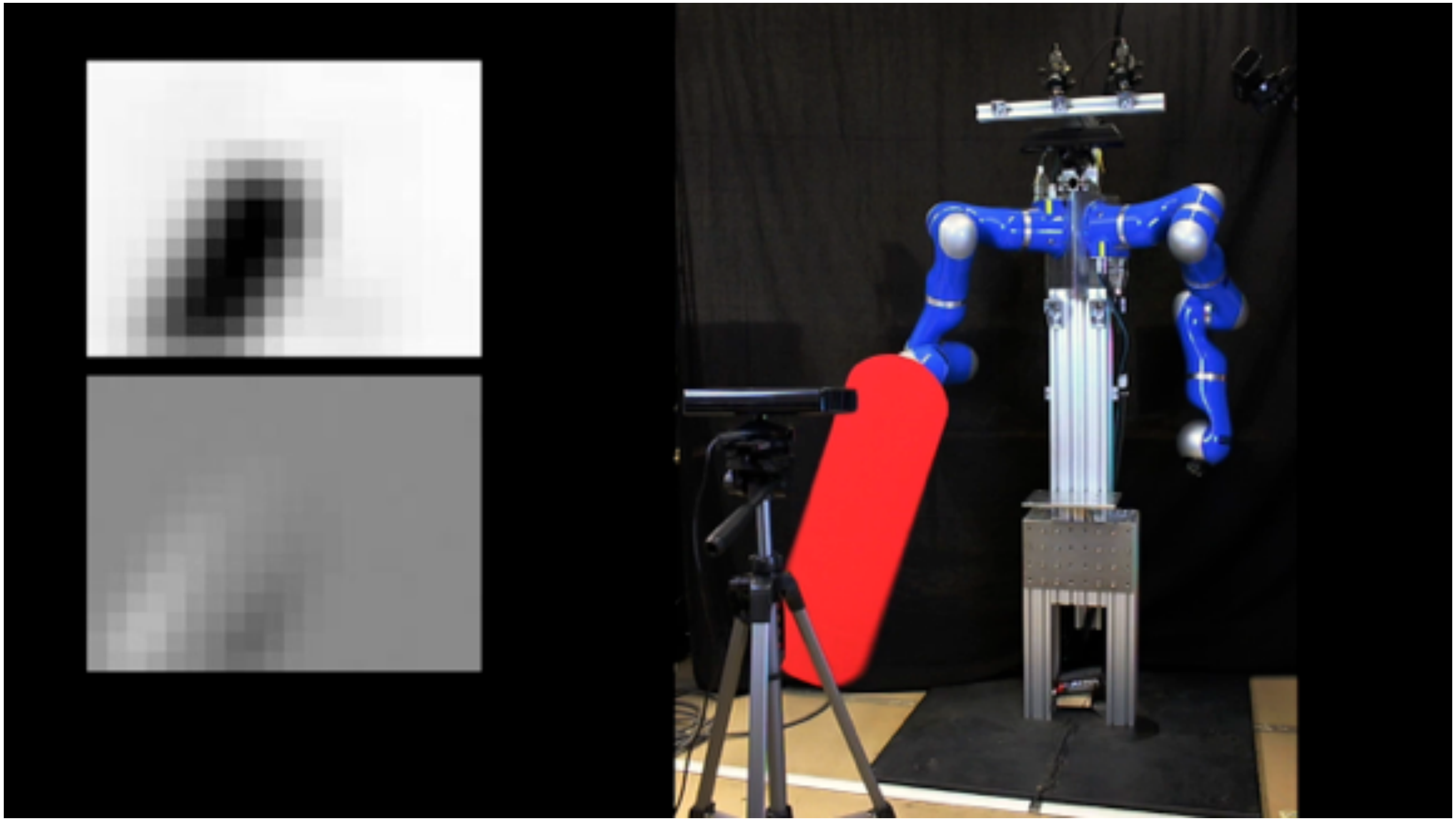


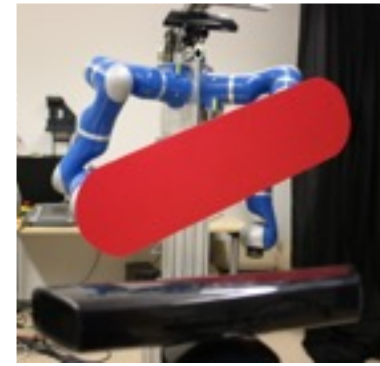
Image-based pendulum swing-up

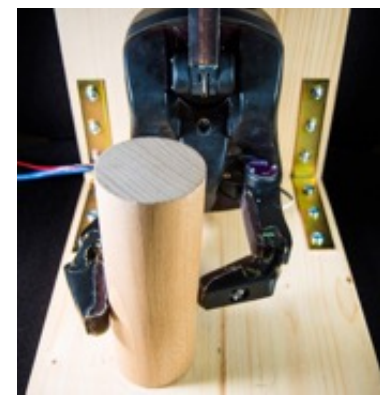


Discussion

Insights

- Reinforcement learning on physical systems poses specific challenges
- With the right tools, reinforcement learning is possible with small sets of complex data
- Variational auto-encoding effective at dealing with noise in high-dimensional data
- Including problem structure helps, especially when dealing with small data sets





Current and planned work

- Integrate deep networks more directly in policy
- Improve performance with irrelevant dimensions
- Use current insights to handle small data sets

Thanks for your attention!

Questions?