Behavior learning research team

Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface

Language used in Robotics Field

Yutaka Nakamura (Team leader)

Behavior learning research team
Guardian Robot Project
Riken

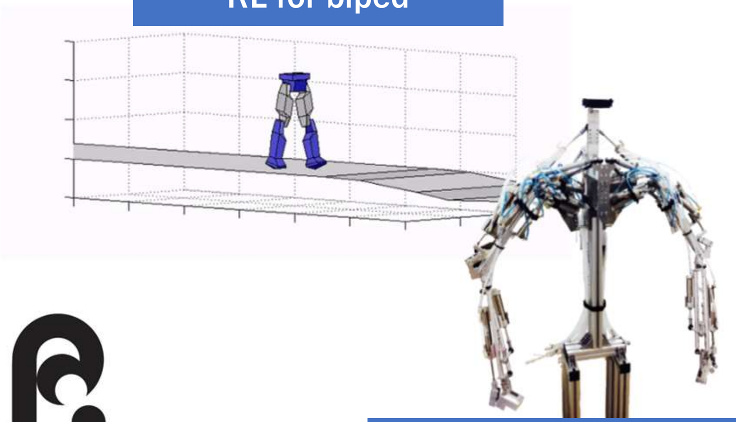
RIKEN ガーディアンロボットプロジェクト Guardian Robot Project

https://grp.riken.jp/labs/behav_learn/

Self introduction

- □Yutaka Nakamura (Team leader,)
  - □ ～2006: Nara institute of science and technology (Ishii lab)
  - □ ～2020: Osaka University (Ishiguro Lab)
  - □ ～Current: Team leader of BLRT, GRP, Riken
- □Research interests
  - □Machine learning, reinforcement learning, human robot interaction (HRI), communication robot, morphological computation, generative model

RL for biped



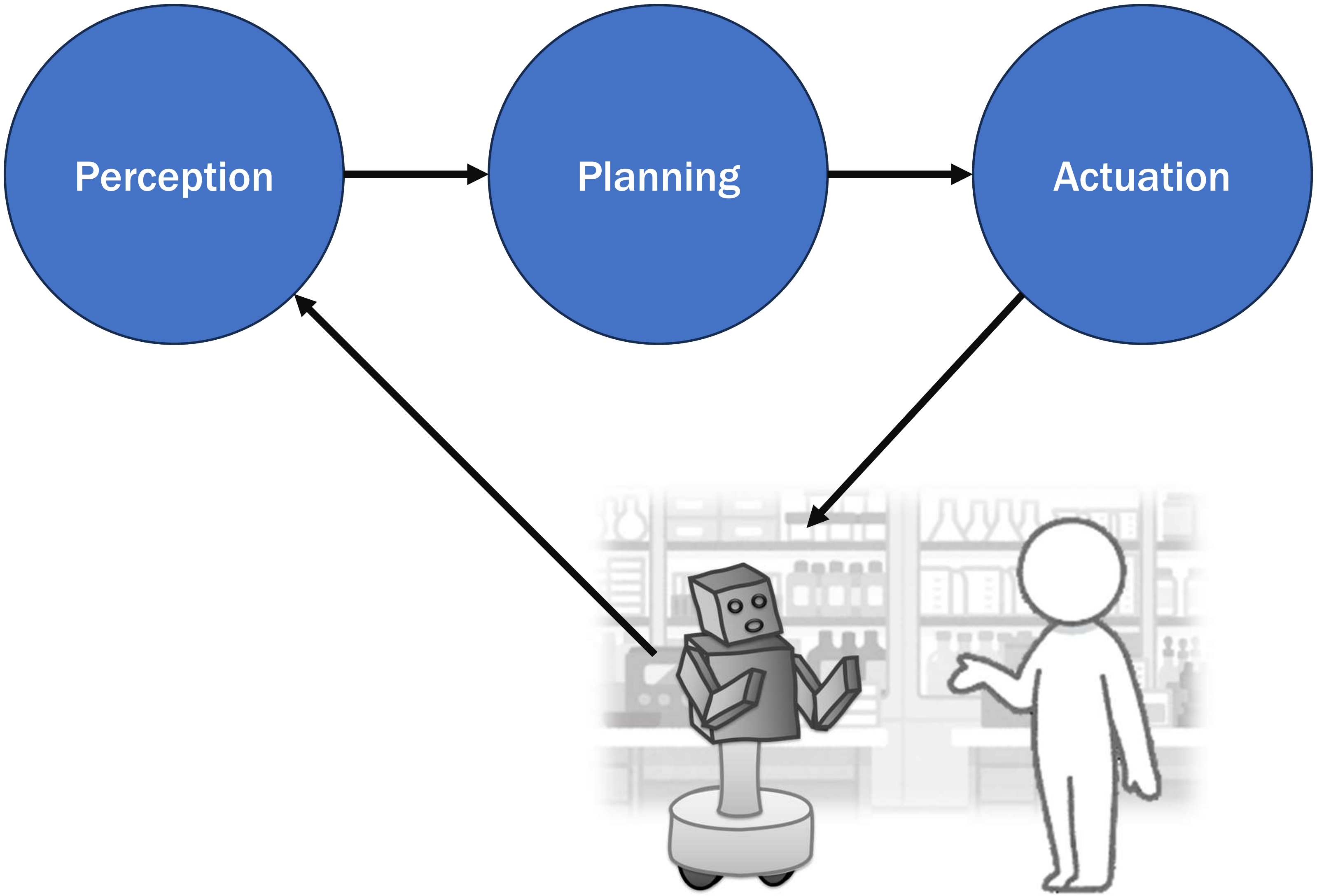
Bio inspired robot

RL for HRI

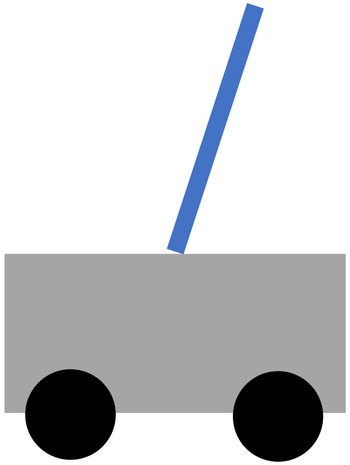
Communication robot

Morphological computation

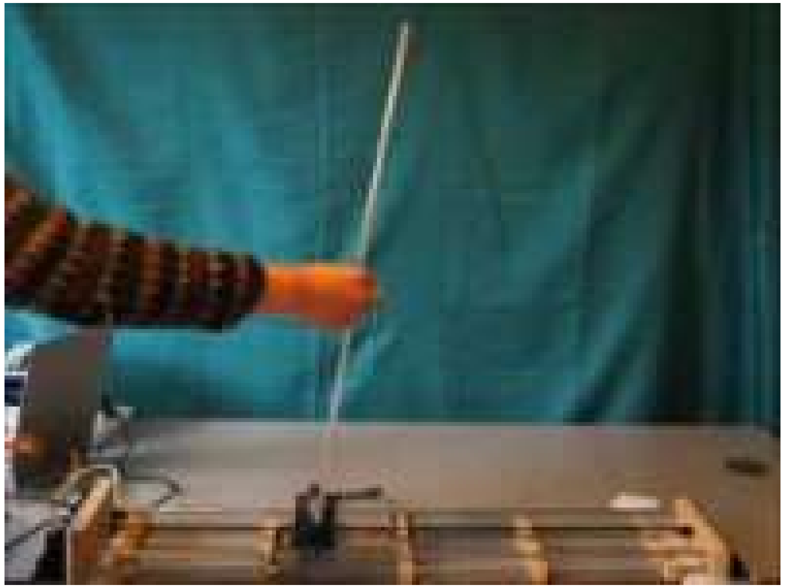
Robotics



Control of a cart pole



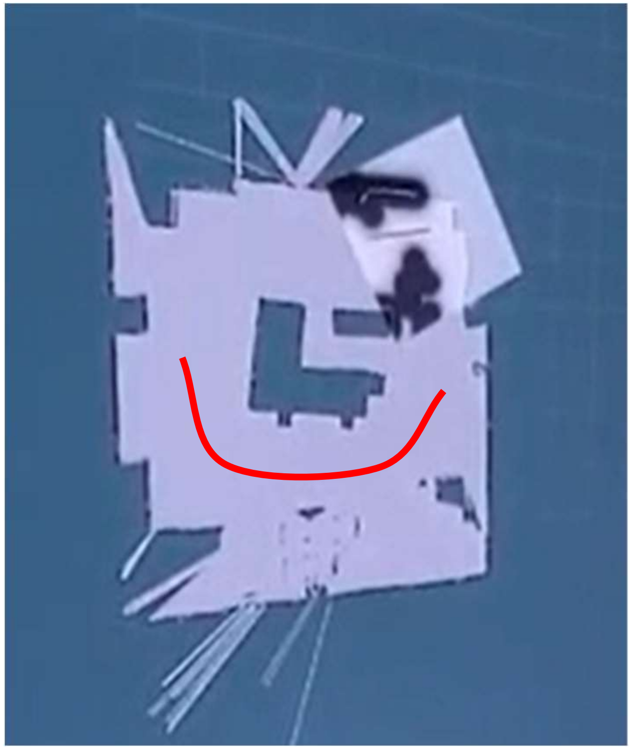
https://www.youtube.com/watch?v=5Q14EjnOJZc



- □ Cartpole task
  - □ Goal: Balancing a pole on a cart
  - □ The pole is connected with an un-actuated joint
  - □ Action: force to move cart
  - □ State: position and velocity of the cart, angle and angular velocity of the pole

Navigation of a rover

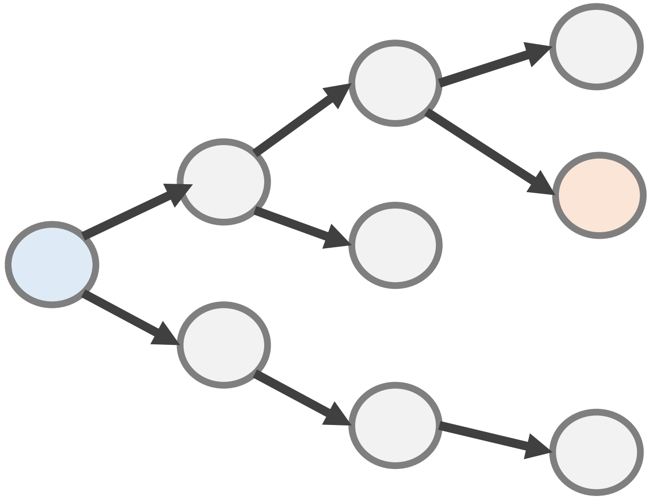
- □ Navigation task
  - □ Goal:
    - □ Move robot to a target location
- □ Technologies used
  - □ SLAM (Simultaneous Localization and Mapping)
    - □ Model of the environment
  - □ Localization
    - □ Estimating the position of the rover
  - □ Planning
    - □ Global planner
      - □ Path within the entire area
    - □ Local planner
      - □ Path planning in the vicinity
      - □ Dynamic change (misalignment, obstacles)



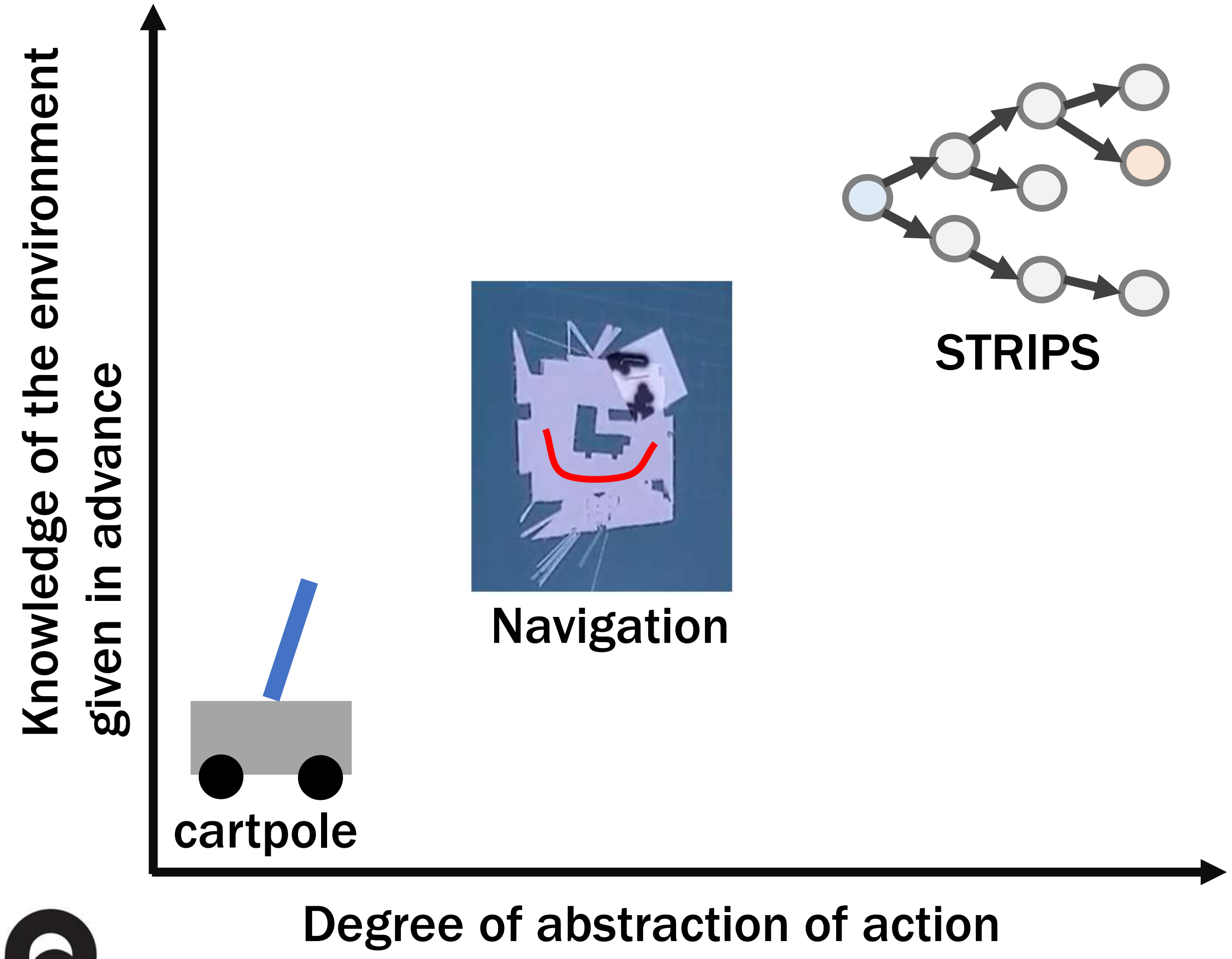
STanford Research Institute Problem Solver

- STRIPS: Automated planner
  - P: set of conditions
  - O: set of operators
  - I: initial state
  - G: goal state

  Task: find a sequence of operators (plan) to get the system from I to G



Degree of abstraction and language


Knowledge of the environment given in advance
STRIPS
Navigation
cartpole
Degree of abstraction of action

Conventional "language" level planning has been used for highly abstract tasks.

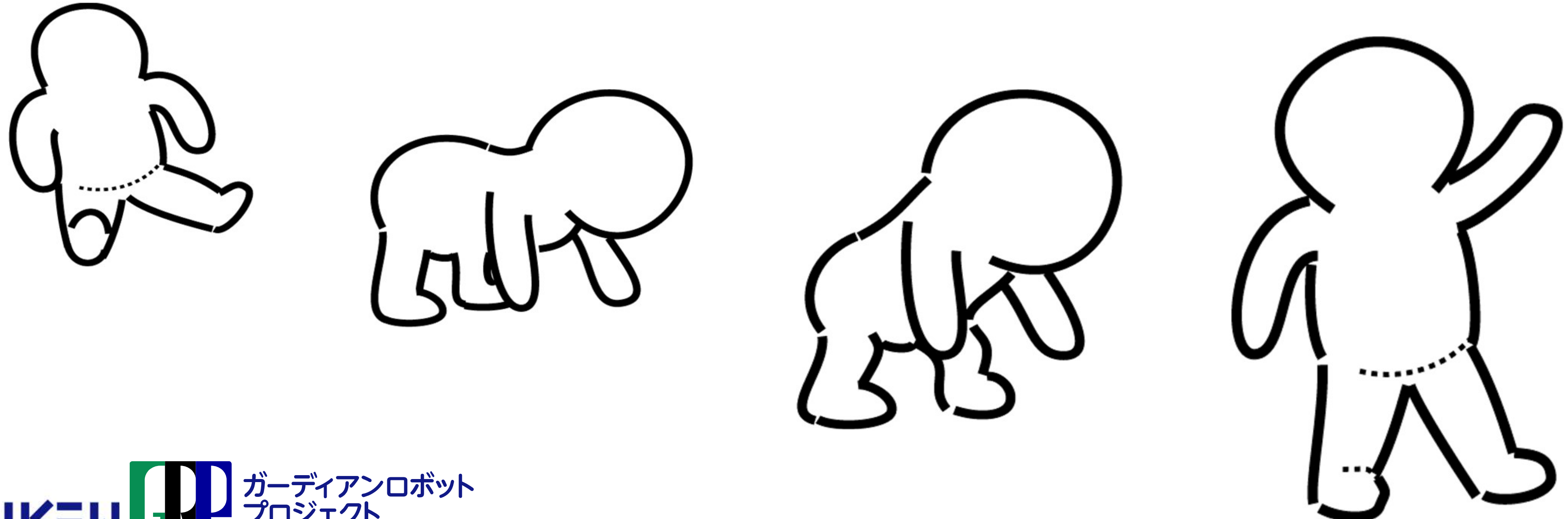
Recently, a number of language-based control frameworks have been created in robotics.

SayCan, RT-1, RT-2, Palm-e･･･

Reinforcement learning

- □Framework to solve (learn) a sequence of actions
  - □Learning by trial and error
  - □Applicable to model-free cases

    (Model-based case is quite similar to control theory)
  - □In recent years, it has been used in ChatBot and other applications.

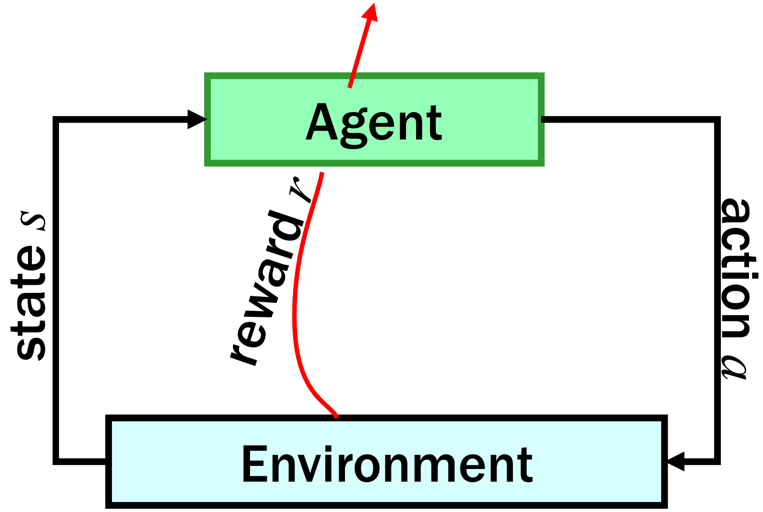


Reinforcement learning

□MDP (Markov Decision Process)

- □ State s, action a
- □ State transition: $s(t+1)$ depends only on $s(t)$ and $a(t)$

  Is defined as state transition probability $P(s'|\, s, a)$
- □ Reward: $r(t)$ depends only on $s(t)$ and $a(t)$



Purpose:

Find the policy function

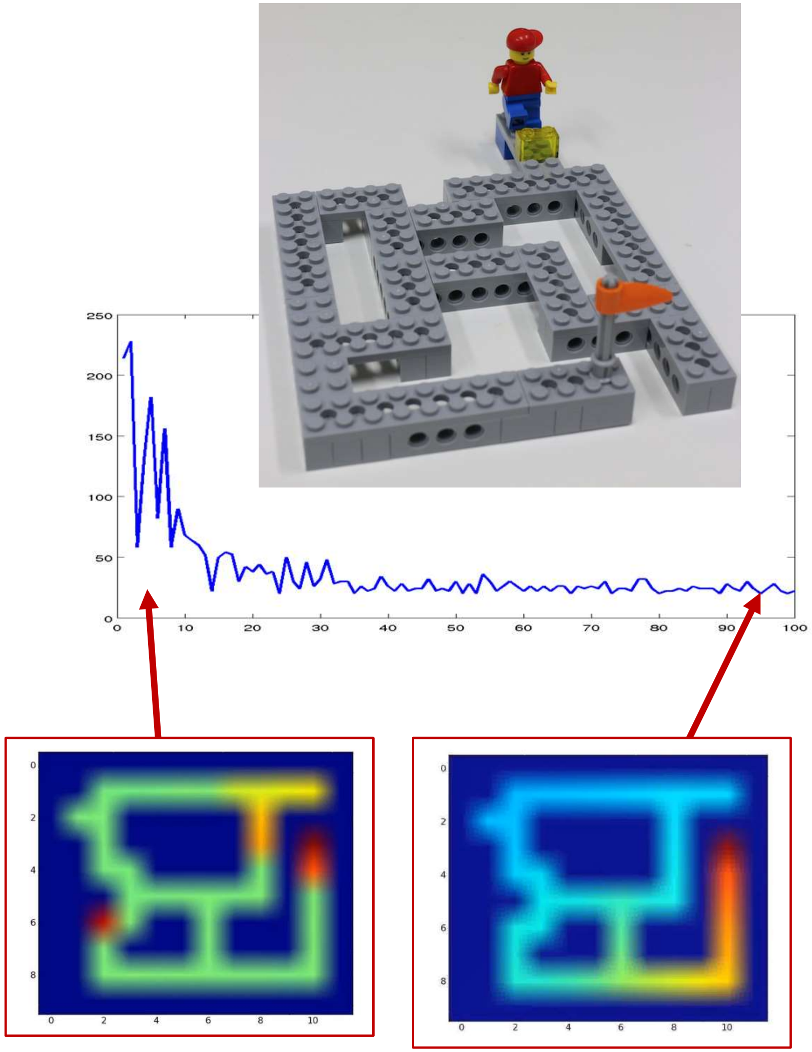
$$a = \pi(s) \sim p(a|s)$$

That maximize the accumulated rewards

$$R(t) = r(t) + r(t+1) + r(t+2) + \cdots + r(t+k) + \cdots$$

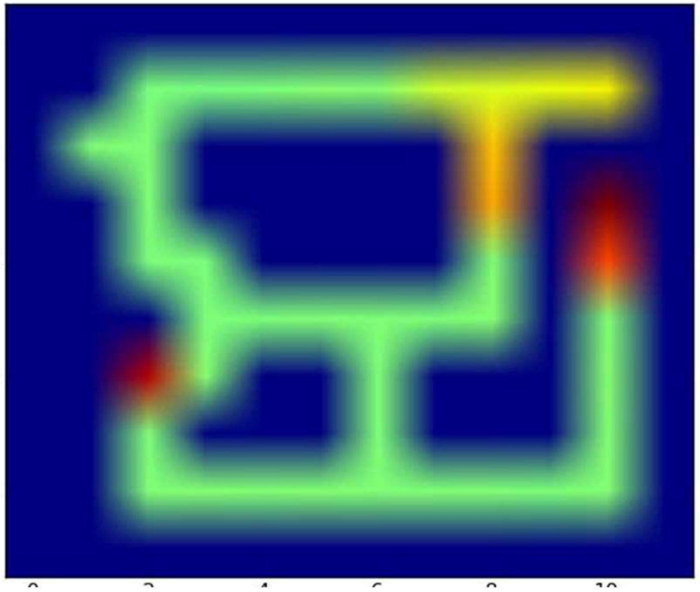
Q learning (shortest path in the maze)

$$Q(s,a) = (1-\eta)Q(s,a) + \eta(r + \gamma \max_{a'} Q(s',a'))$$



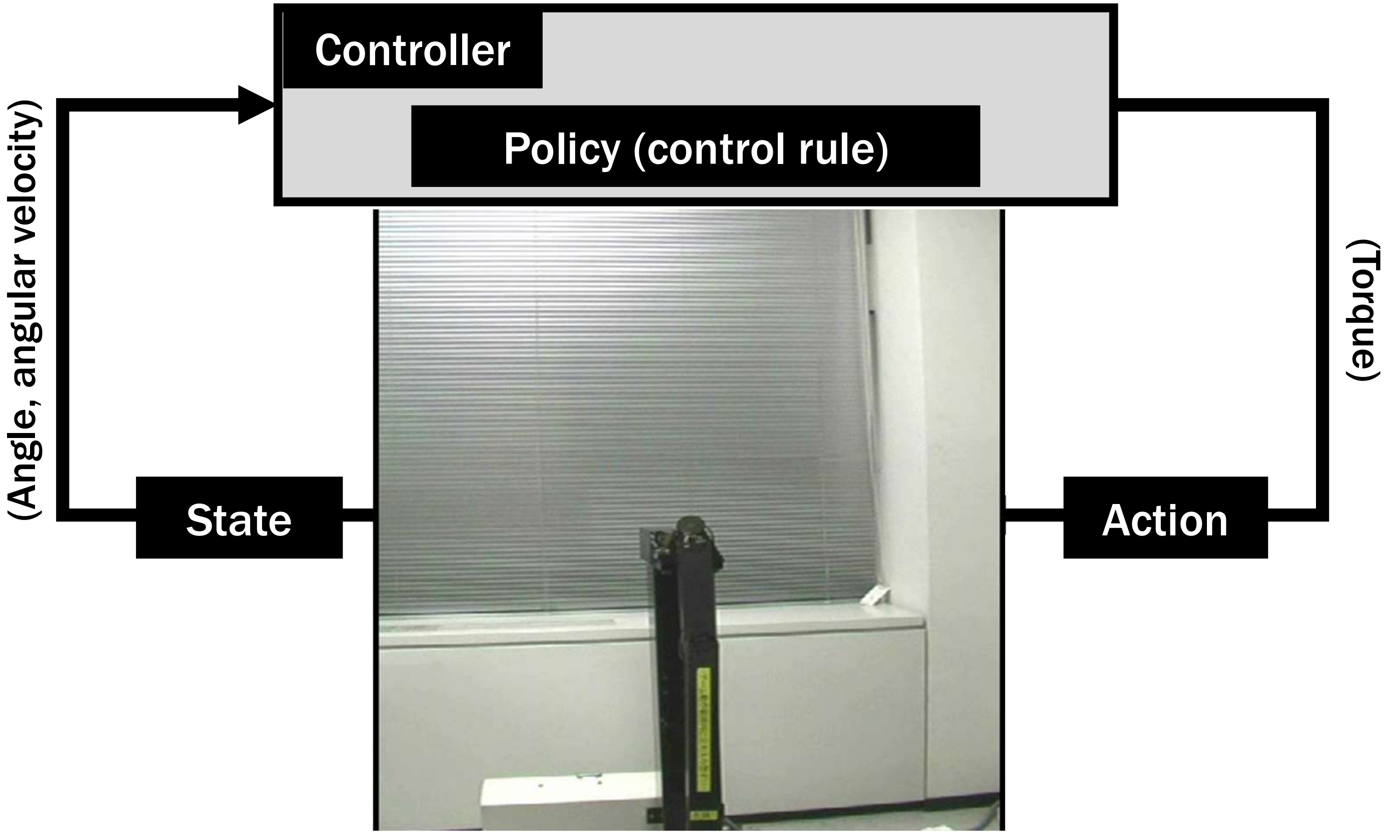
$Q(s,a)$: **Value of choosing action a in state s**

The value function is trained according to the above equation



Control of underactuated robot “Acrobot”

A model of artistic gymnastics (Horizontal bar)



Controller

Policy (control rule)

State

(Angle, angular velocity)

Action

(Torque)

Reward: the height of the 2-nd joint.

Robot ball in a cup

https://youtu.be/qtqubguikMk

Difficulty in RL for robotics

- □ Low sample efficiency
  - □ Large search space
- □ High operating costs
  - □ The robot breaks down.
  - □ Need to restore the initial state.

- □ Human demonstration and imitation learning
  1. Limit the search space by imitation
  2. Optimizing the policy through trial and error (RL)

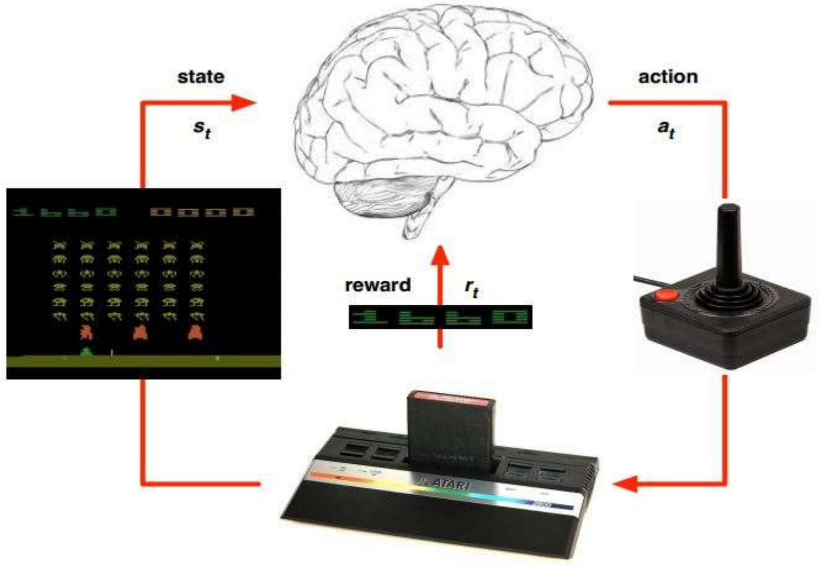
Imitation and Reinforcement Learning for Motor Primitives with Perceptual Coupling, 2010

Deep Q Network (DQN)

□An implementation of Q learning with deep learning

□Human-level control

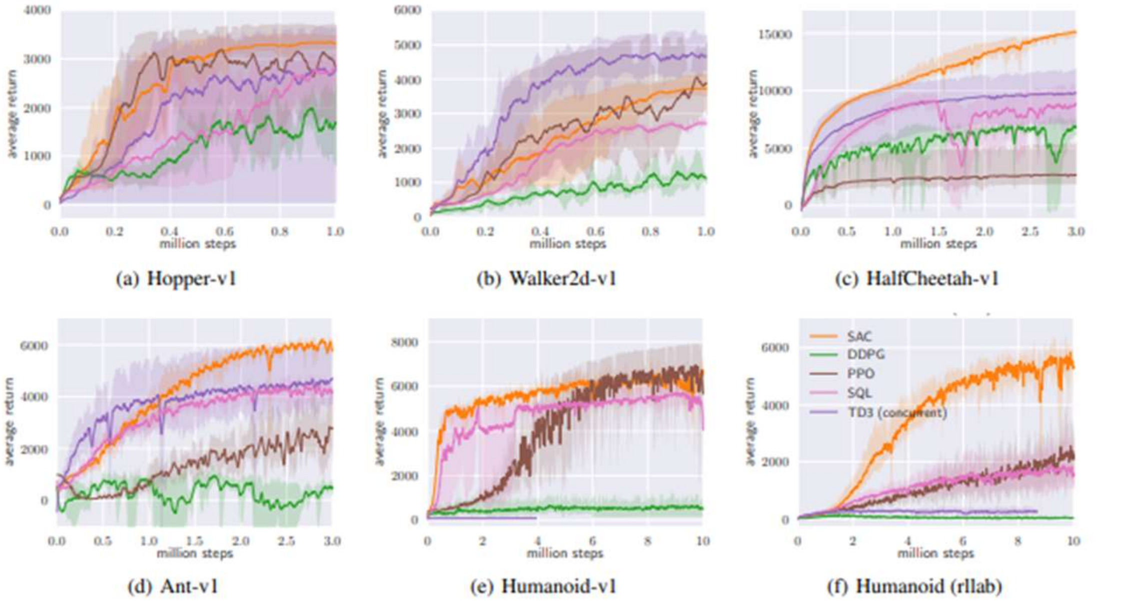
- □Observation: screen
- □Action: Buttons and joysticks for arcade games



Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529-533.

Applications of state-of-the art RL algorithms

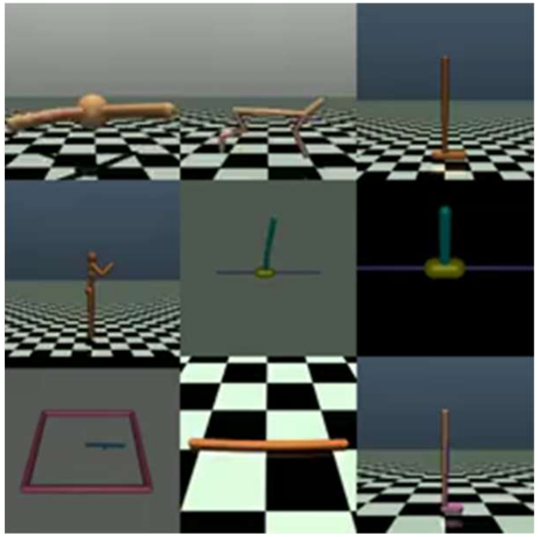
- SAC (soft actor critic), DDPG (deep deterministic policy gradient), PPO (proximal policy optimization) and other algorithms have been developed and are in competition with each other.



(a) Hopper-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) Humanoid-v1 (f) Humanoid (rllab)

Figure 1. Training curves on continuous control benchmarks. Soft actor-critic (yellow) performs consistently across all tasks and outperforming both on-policy and off-policy methods in the most challenging tasks.

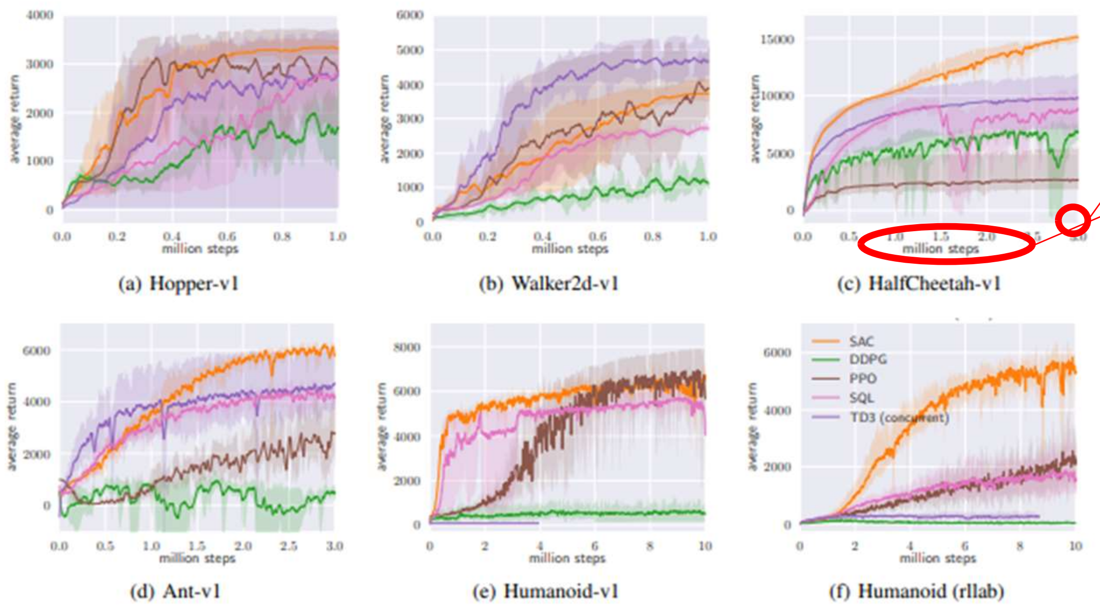


https://www.samplefactory.dev/09-environment-integrations/mujoco/

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, 2018

Applications of state-of-the art RL algorithms

- SAC (soft actor critic), DDPG (deep deterministic policy gradient), PPO (proximal policy optimization) and other algorithms have been developed and are in competition with each other.



(a) Hopper-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) Humanoid-v1 (f) Humanoid (rllab)

Fi

ou

10

Million steps

278 hours at 10fps

Hyperparameter Search

Learning a single behavior in a simplified environment

Real-world applications are not likely to be easy.

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, 2018

Massively parallel deep RL

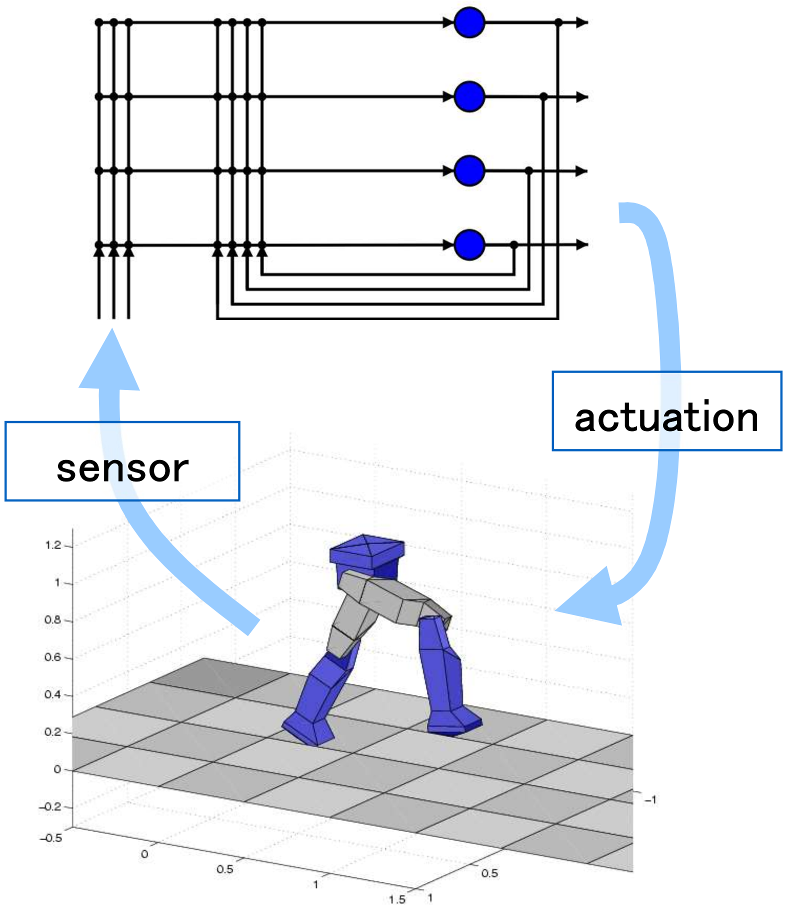
□ One way is to gather experience on a variety of situations by running a large number of simulations.

□ Combined with techniques such as Sim2real, which transfer the policy learning by simulation to a real environment, improved sample efficiency in the real environment is expected.

Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning, 2021

CPG-RL

- Reinforcement learning for controllers with reference to biological control mechanisms
  - CPG: central pattern generator is a nervous system controlling periodic movements



- Deep RL for CPG controller realizes a Quadruped Robot Robust to Environmental Changes and Fast Learning

CPG-RL: Learning Central Pattern Generators for Quadruped Locomotion

[illegible]

CPG-RL: Learning Central Pattern Generators for Quadruped Locomotion, 2022

Curiosity: intrinsic motivation

- □ Curiosity
  - □ Actively explore inexperienced behaviors and state transitions
  - □ Exploitation–exploration balance
    - □ Choose the best behavior experienced
    - □ Search for good behavior never experienced before
- □ Practical advantages
  - □ Addressing *sparse rewards problem*
    - □ When rewards are defined only for the goal, no learning occurs until the goal is reached
  - □ Reward shaping
    - □ It is not guaranteed that the reward design (by human) is good

Curiosity-driven Exploration by Self-supervised Prediction,2017

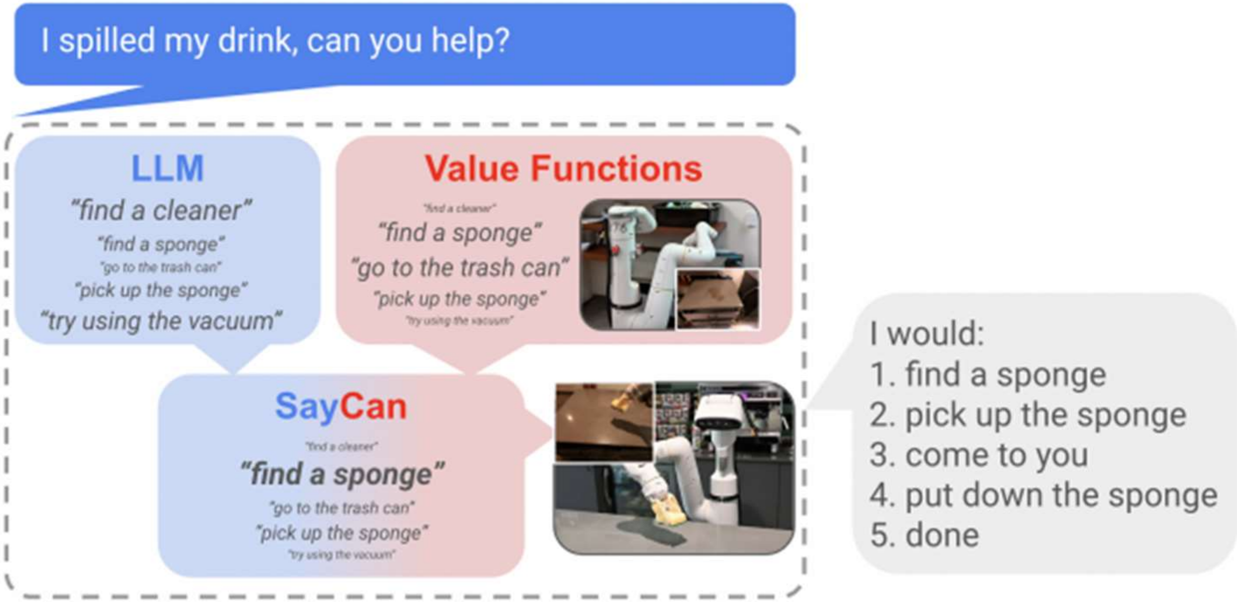
Instruction by natural language

- □Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, 2018
  - □The robot can be operated by unconstrained spoken language instruction
- □RT-1: robotics transformer for real-world control at scale 2022
  - □Mapping from instruction and image to action
  - □Trained by 130k demonstrations
- □MDM: Human motion diffusion model, ICLR2023
  - □Generative model for human behavior
  - □Generate movement according to instructions

Temporal extended instruction

- LLM: Interpret the instruction
  - Calculate the likelihood of each skill
- Affordance function (value function)
  - Estimate the success rate of each skill

→ Planning



Do As I Can and Not As I Say: Grounding Language in Robotic Affordances, 2022

AlphaGo and AlphaZero

- □AlphaGo
  - □In the early stages of learning, action selection rules are learned using game records (policy network)
  - □Evaluates each move using reinforcement learning that pits agents against each other (value network)
- □AlphaZero
  - □Optimize strategies in a battle between agents without using human knowledge

AlphaGo < AlphaZero

In some cases, human knowledge worsens performance

Summary

- □Motion learning at the signal level has been well studied.
  - □Sample efficiency is important
  - □Introducing human knowledge can reduce the number of iterations.
    - □Model based RL (Model of the system is given in advance)
    - □Imitation learning
    - □Refer to a system that is working well (e.g. GPG)
- □Development of systems that provide instructions in natural language
  - □Mapping from Instruction to action using DNN
  - □Integration with large-scale language models enables planning
- □Methods using language are emerging
  - □Methods with high generalization ability.
  - □Effective learning with language