Autonomous exploration through curiosity and social guidance

Manuel Lopes, Pierre-Yves Oudeyer

INRIA, Bordeaux, France

flowers.inria.fr/mlopes

Goals of the talk

1. An overview of active learning and intrinsic motivation on robots
2. Empirical measures of progress
   -> Generalization of Rmax with empirical measures
3. Unified view of several active approaches
   -> strategic student problem

How efficient can learn be?

Requirements

- Good features
- Good generalization capabilities
- Find & Coll
- High-dime nment, non-uniform noise
- Too many things to le

Machine learning

Active Learning

Intrinsic motivation

Development

Active Learning

- The learner selects what to observe next/what to query next
- Advantages:
  - Only informative points are queried
  - Less data/time (for some cases exponential gains can be obtained)
- Disadvantages:
  - Computational cost of making the queries
  - Queries might not be relevant for the task
  - Theoretical analysis is recent

Active Learning for Robots in Real Life

- Find resources (e.g. oil, minerium, …)
  - Each hole costs ~1million$
- Space Exploration
- Environmental Monitoring
- …

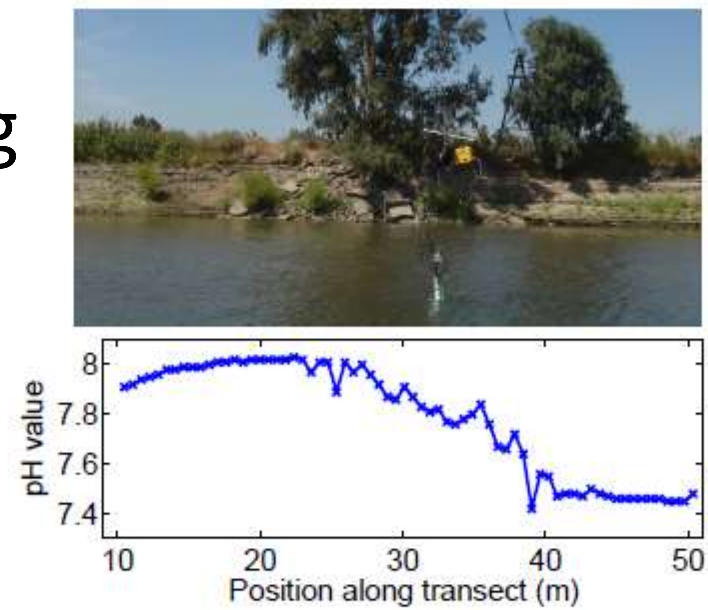


Figure 1. Top: Active sampling using the NIMS sensor (Harmon et al., 2006) deployed at Merced River, CA. The sensor can perform horizontal and vertical traversal. Bottom: Samples of pH acquired along horizontal transect.

Active Learning

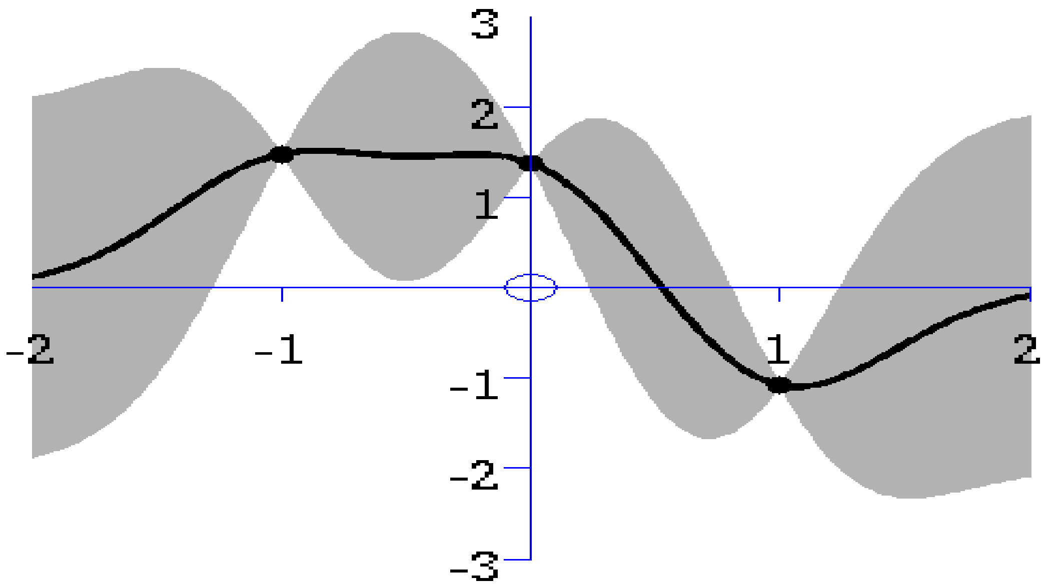
- Learn with reduced time/data
- Fixed tasks
- Learnable everywhere
- Everything can be learned in the limit
- Reduce uncertainty

Intrinsic Motivation

- Learn with reduced time/data
- Tasks change and are selected by the agent
- Parts might not be learnable
- Not everything can be learned during a lifetime
- Improve progress

Gaussian Processes (GP)

- What is the most informative point?

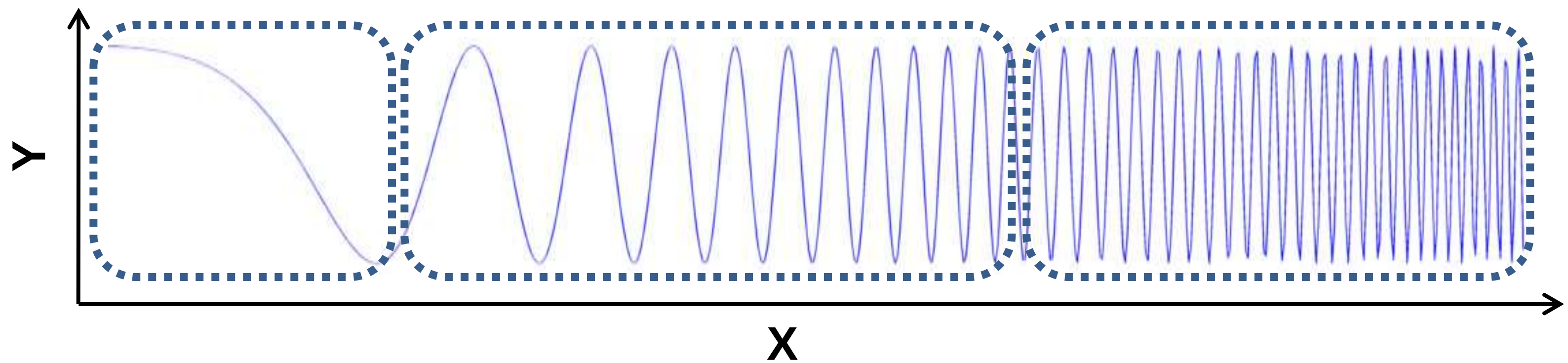


It is the one with *less samples in the neighborhood.*

And this is even **ignoring ALL THE OBSERVATIONS!!!**

Difficulties

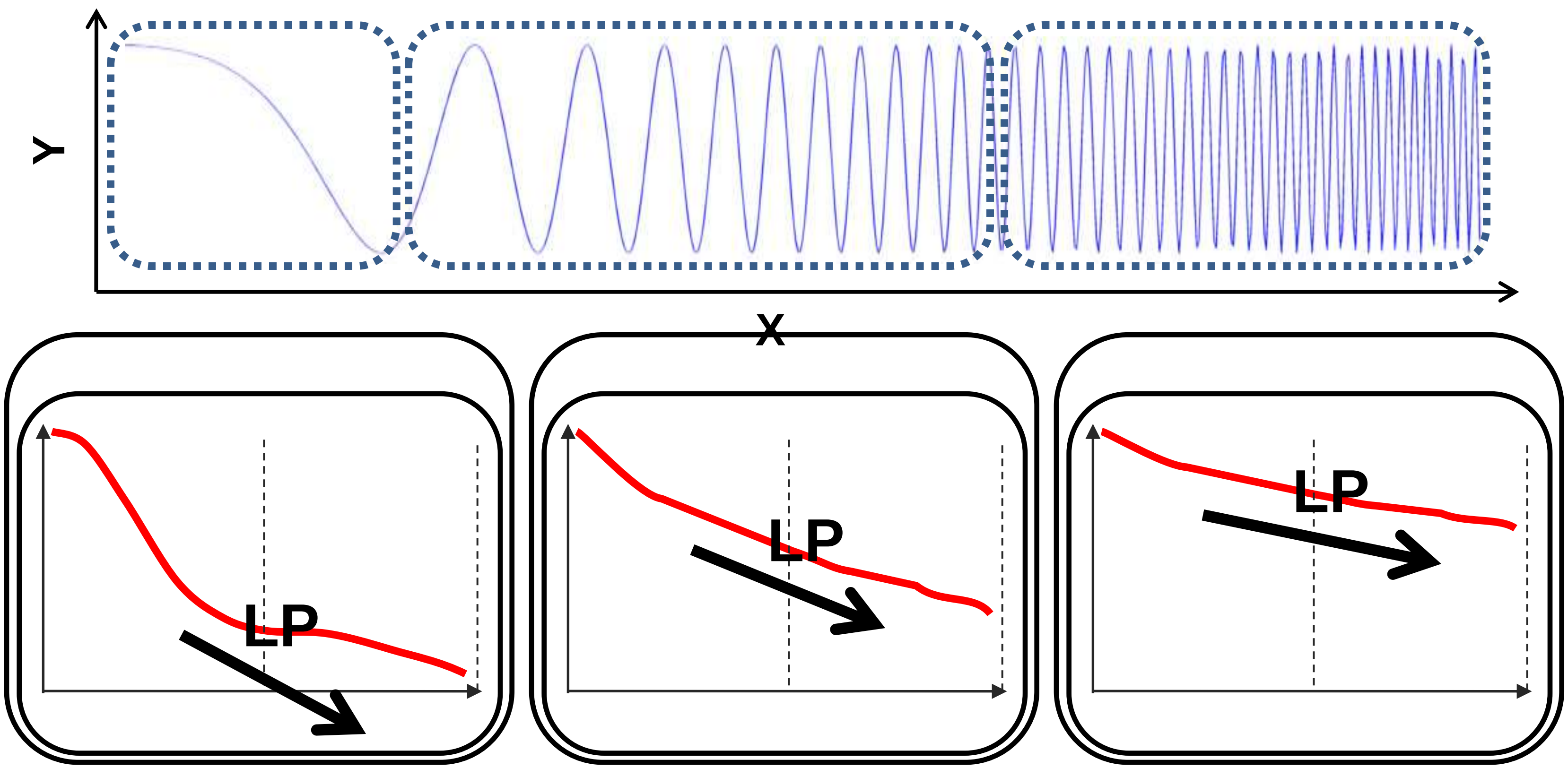
- Non-stationary noise
- Unknown kernels
- The same amount of data would be requested everywhere



Solution:

-> **don’t assume progress, measure it!!!**

Multi-region measure of progress and probabilistic selection of regions



Progress measure = average reduction of the derivative of learning

Can we always choose the region with more progress?

No

a) Measure of progress is noisy b) Progress might not be monotonous

Why Empirical Measures of Success?

"Classical Active Learning"

Given sufficient data:

- Model accurate in whole space
- Time and Space Stationary (recent developments on space)

☺ Easier theoretical study

☺ In the limit guarantees

☹ Model might be too complex

☹

R-max

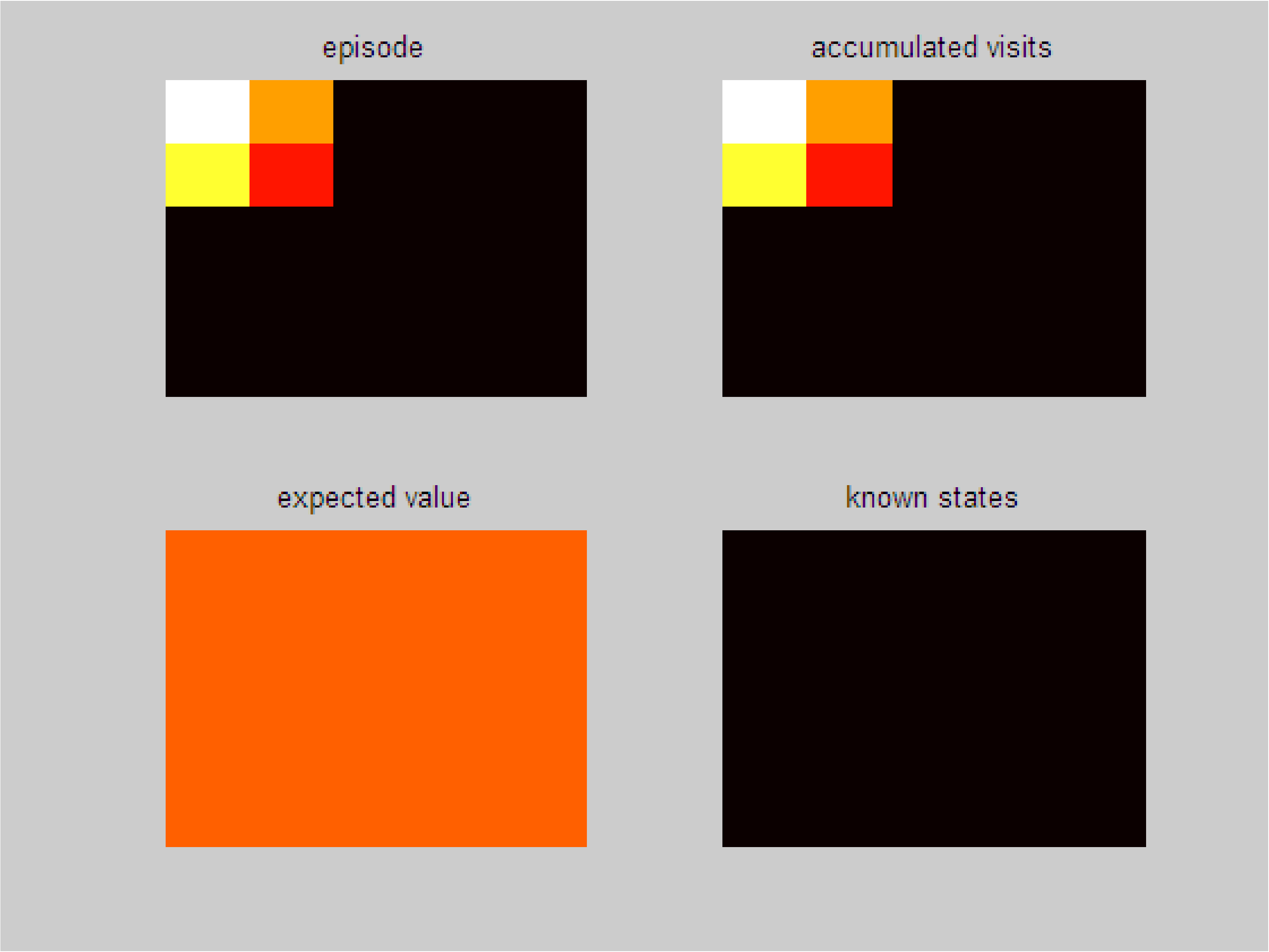
- solves the exploration/exploitation dilemma in model-based RL
- polynomial time approximation of the policy

Algorithm **Rmax** :

1. Divides states in known and unknown
2. Unknown states are optimistically initialized to Rmax
3. At each time step plans in this surrogate model

video

R-max



R-max Limitations

- All unknown states are assumed to provide the same progress
- All states assumed to be similar
  easy to relax but then we need to know exactly how different they are
- Cannot deal with any time of time changes

-> empirical measures of progress

ζ-R-max (zeta-R-max)

Generalization of Rmax with empirical measure of progress

$$\mathcal{R}^{\zeta\text{-R-MAX}}(s,a) = \begin{cases} \mathcal{R}(s,a) & \zeta(s,a) < m \\ R_{max} & \text{else} \end{cases}$$

where ζ is:

$$\zeta(s,a) := \hat{\zeta}(s,a) + \alpha\sqrt{\nu(s,a)}$$

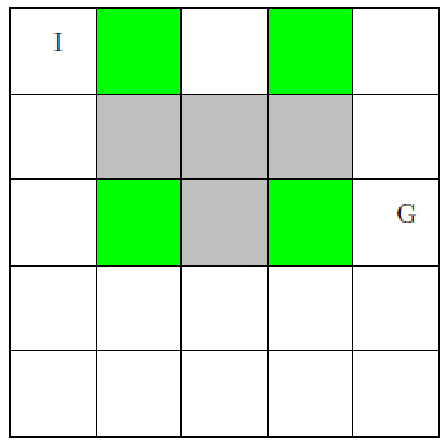
with

$$\hat{\zeta}(s,a) := CV(D_{s,a}^{-k}, s, a) - CV(D_{s,a}, s, a) \approx \mathcal{L}(\hat{\mathcal{T}}^{-k}; D_{s,a}) - \mathcal{L}(\hat{\mathcal{T}}; D_{s,a})$$

(Lopes et al, NIPS'12)

R-max vs ζ-R-max

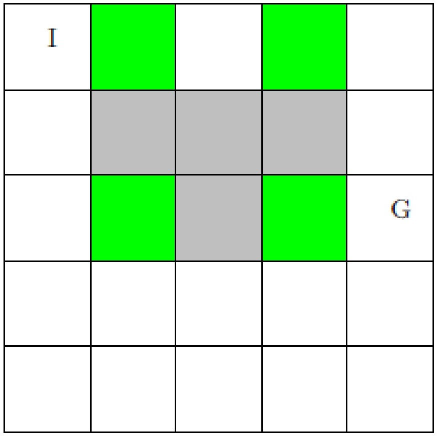
- Goal:
  Learn the dynamical model of a typical maze

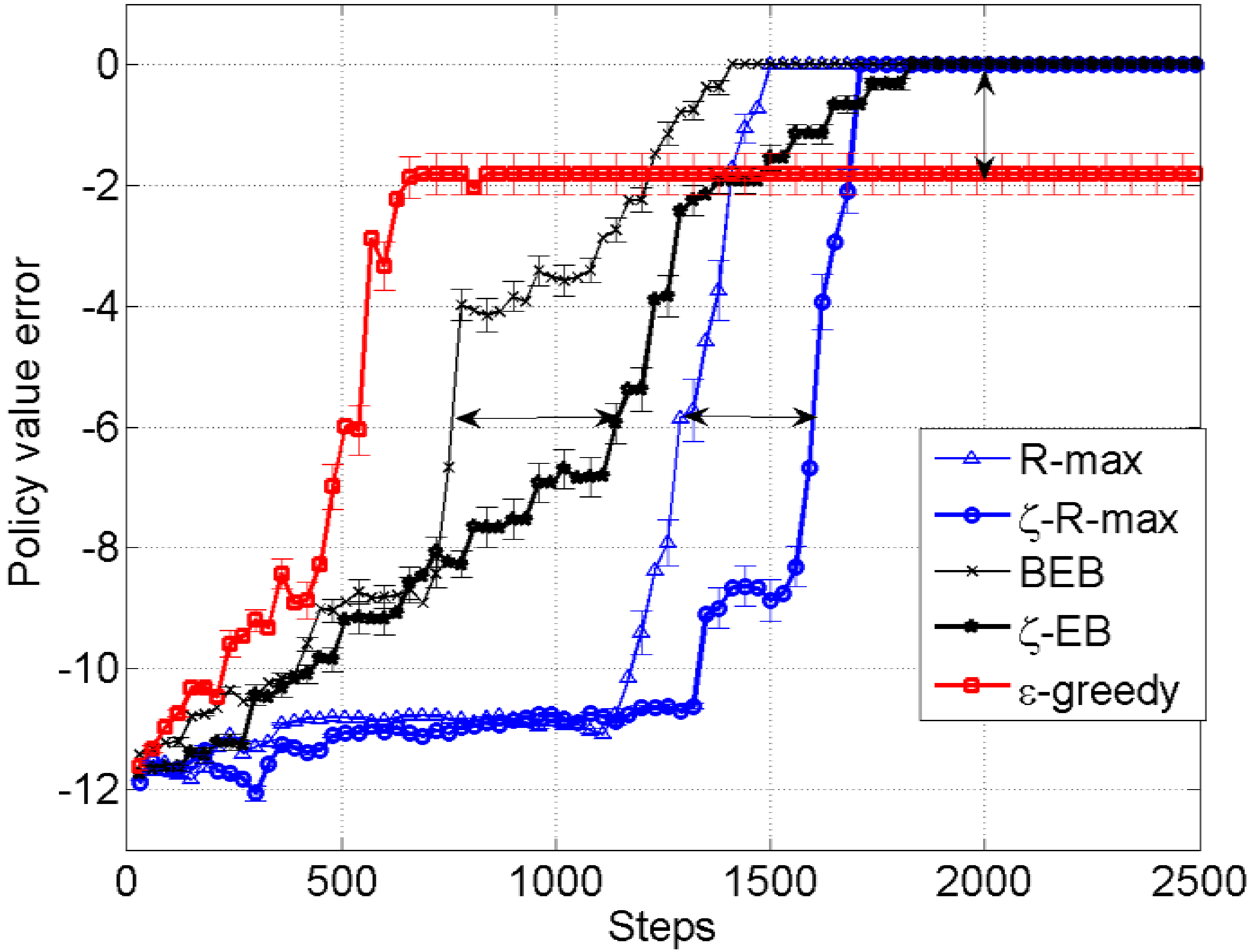


- Grey: Obstacles; Green: stochastic transitions
  I: Initial State; G: Goal State

ζ-R-max with correct assumptions

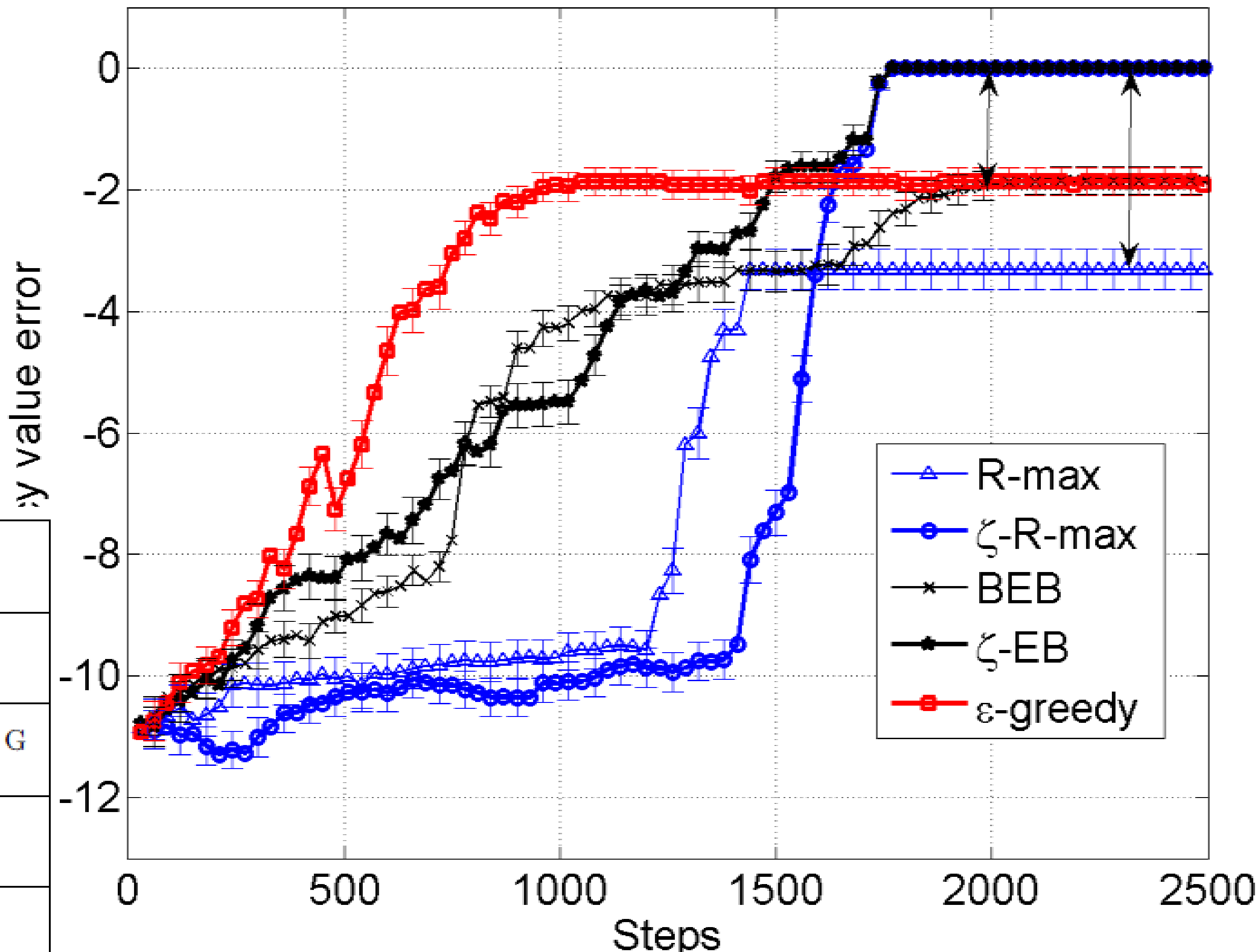
i.e.
The noise levels of white and green states is known

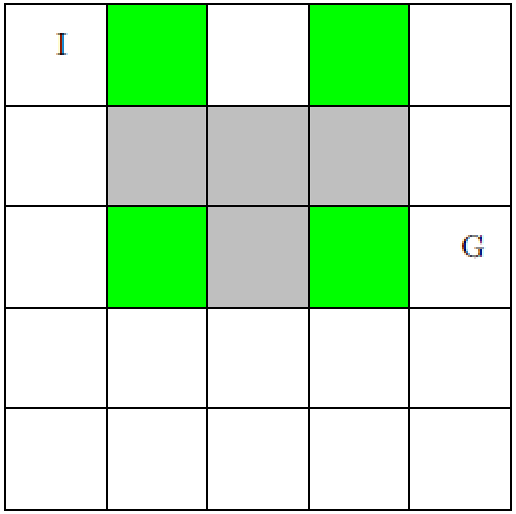




ζ-R-max with violated assumptions

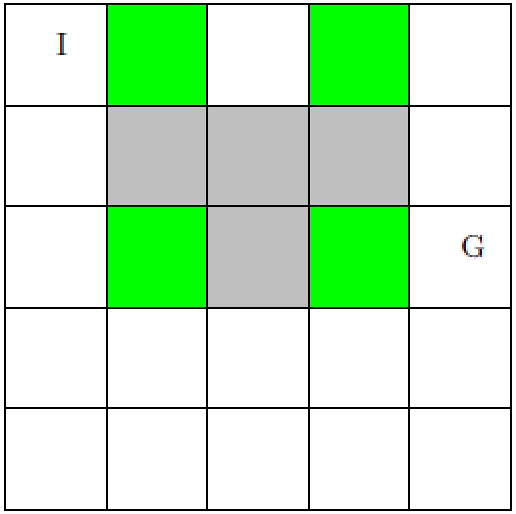
i.e.
The noise levels of white and green states is **not accurately known**

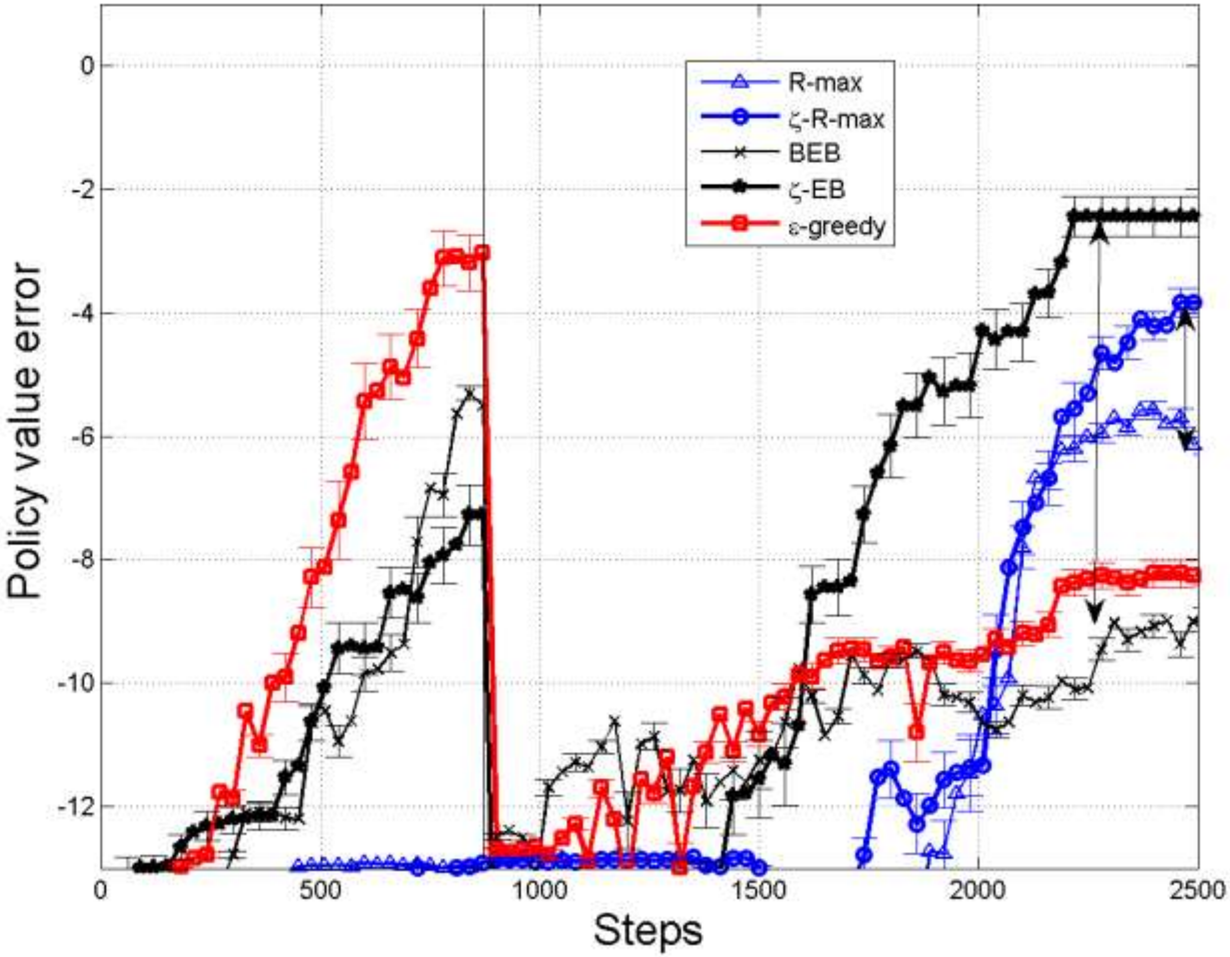




ζ-R-max in time variant domains

A state is the path of the optimal policy **changes at step 900**.





Active Learning in Robots

- Explore / Exploit
  (Rmax, e-greedy, UCB,…)

- Sample informative data
  RIAC,

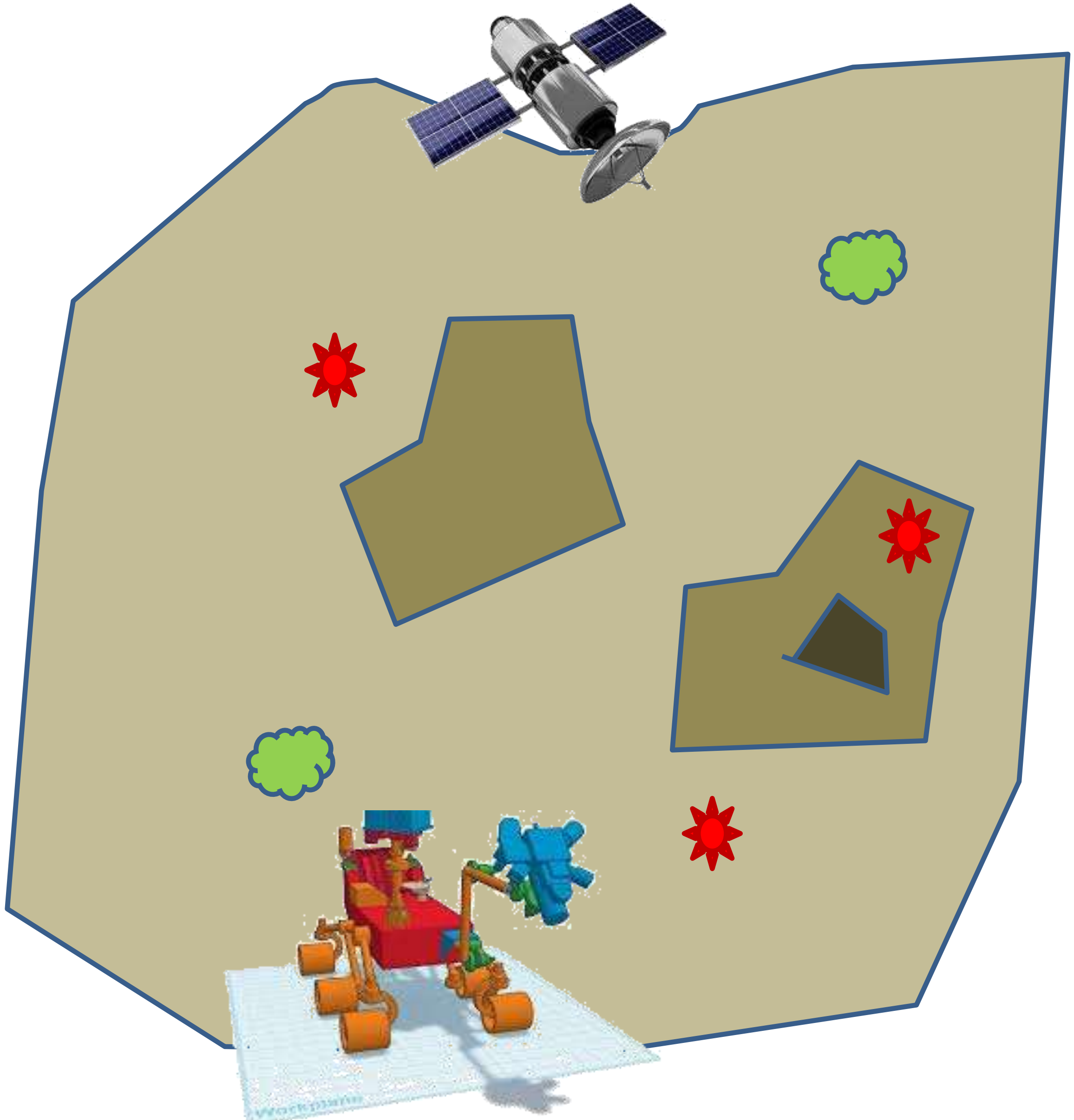
- Select particular points
  actGP, actNN,…

- Pure Learning
  RIAC, actGP, actNN,…

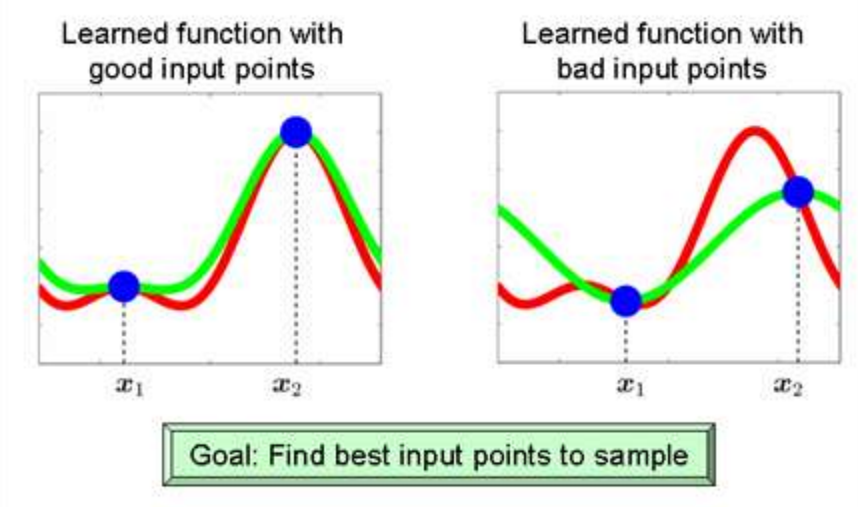
- Plan actions to acquire informative data
  (Rmax, SAGG-RIAC …)

- Select regions/strategies/options
  (IMRL,SAGG-RIAC,SSB,…)

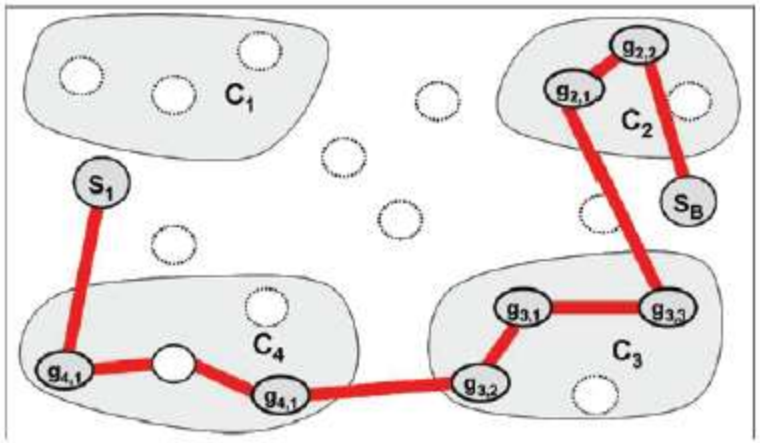
- Goal: Map and locate resources in an environment
- Robots
  - Satellite
    - RGB Camera
    - InfraRed Camera
  - Mobile Robot
    - Camera
    - ChemCam
    - Arm + ChemCam
- Choices:
  - Which Robot to use?
  - Where to sense?
  - Which sensor to use?
- Optimize:
  - Error in localizing resources
  - Quality of map
  - Energy
    (sensor use + motion)
  - Time



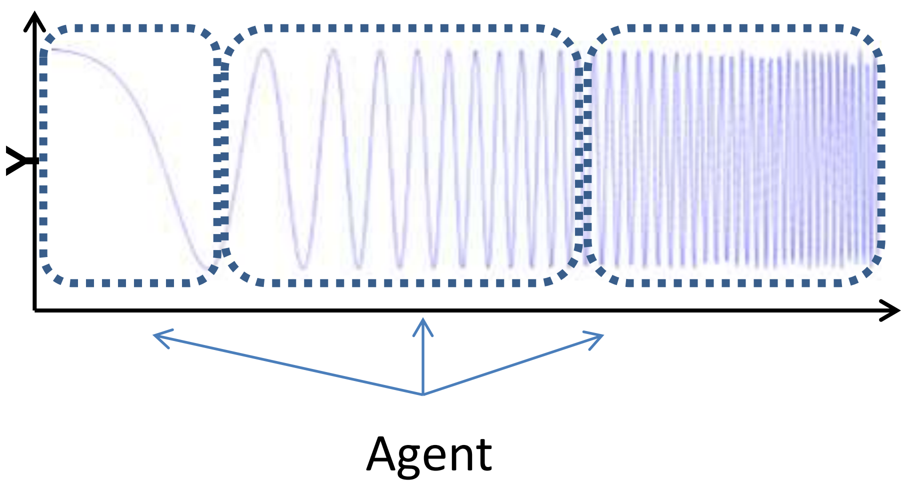
Choosing points



Choosing trajectories



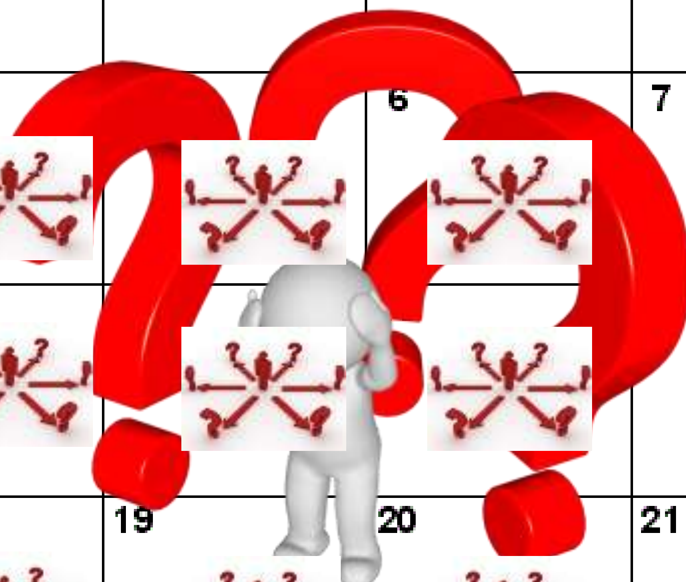
Choosing regions/options



or methods

Strategic Student Problem

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					1	2
3	4		6	7	8	9
10	11				15	16
17	18	19	20	21	22	23
24	25	26	27	28 Exam1	29 Exam2	30 Exam3



Strategic Student Problem (SSP) Examples

Choices

- Region to probe

Tasks

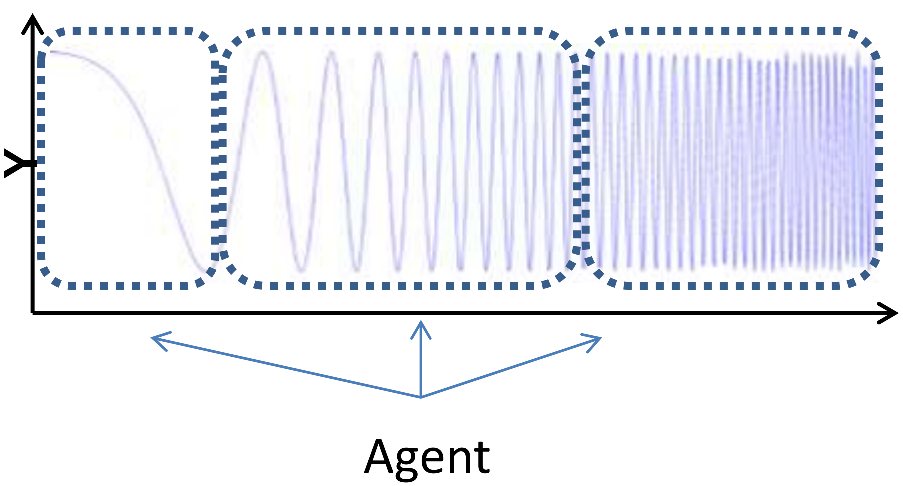
- Learn each region

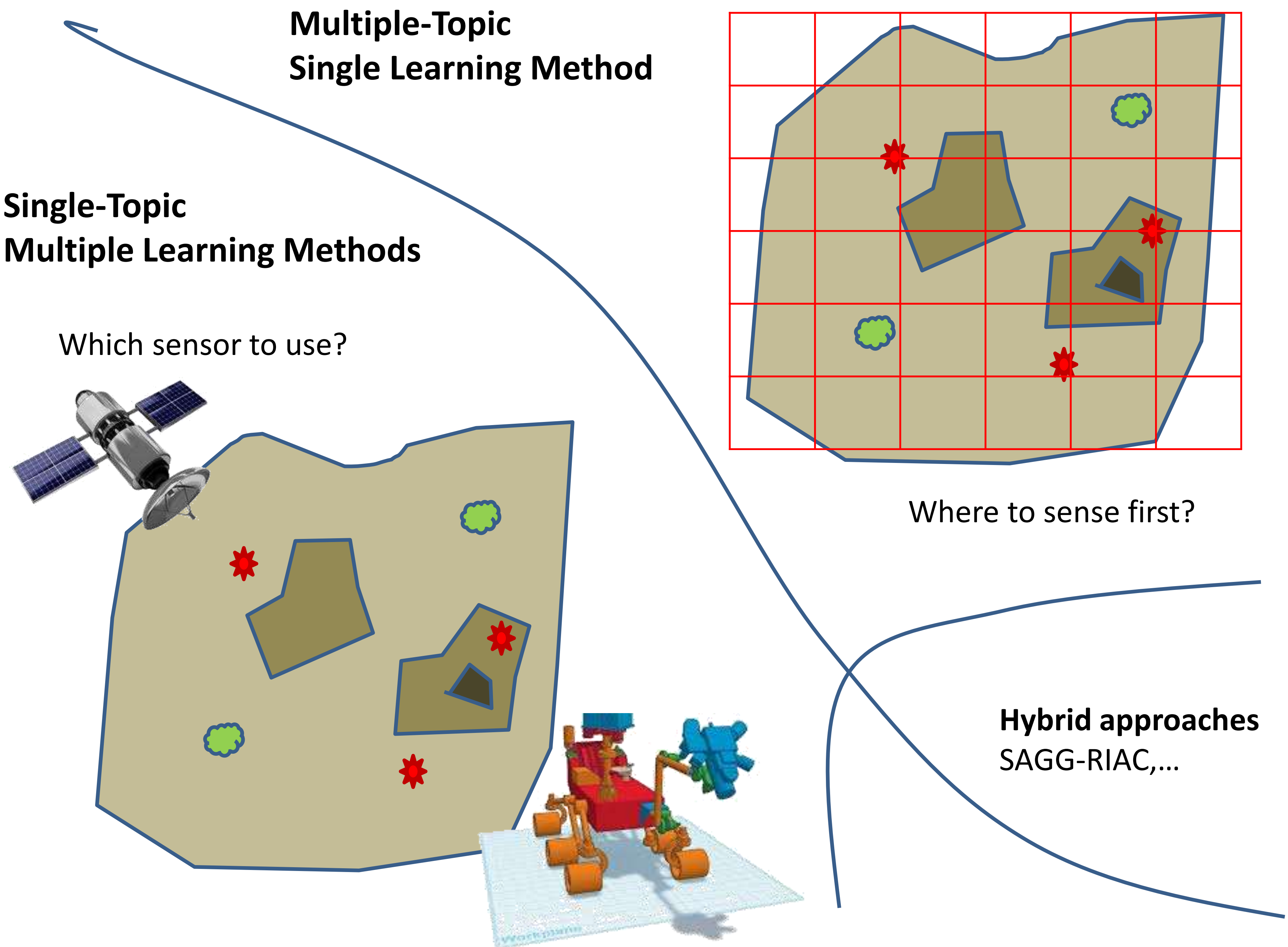
Choices

- Learning method / sensor / action to use

Task

- Learn region




Multiple-Topic
Single Learning Method
Single-Topic
Multiple Learning Methods
Which sensor to use?
Where to sense first?
Hybrid approaches
SAGG-RIAC,...

Strategic Student Problem

At each day study the topic

- randomly

not bad but we might to be able to do better

- with worst expected result

might get stuck on very difficult topics

- with best expected result

improves the best mark but not the average mark

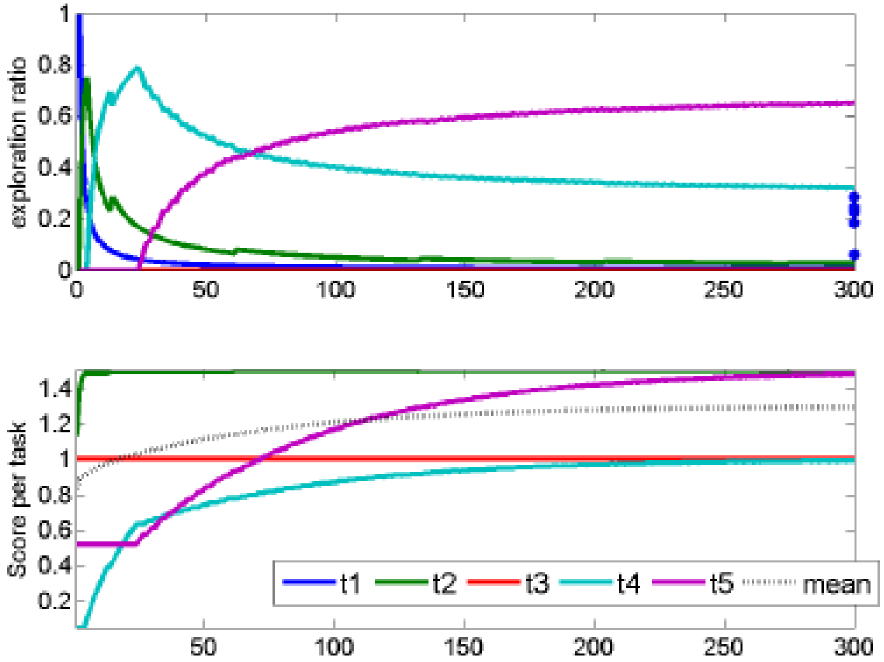
- giving maximum progress on the average mark

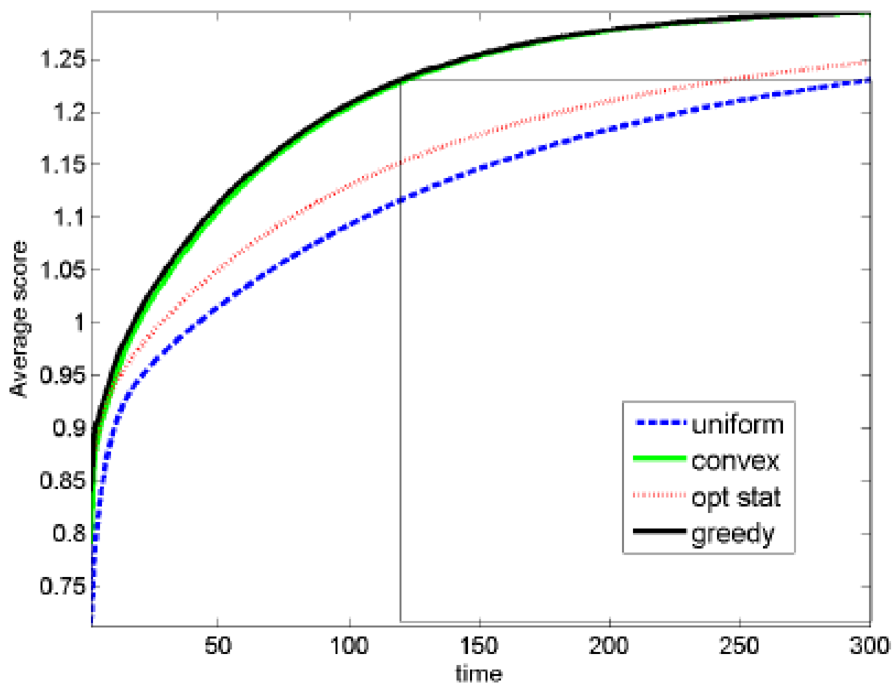
seems a good strategy ☺

SSP – Simple Example

$$\max_{n_i} \sum_i C_i \left(1 - e^{-\frac{n_i}{p_i}}\right) + B_i$$

$$s.t. \sum_i n_i = T \ , \ n_i \geq 0$$





Easier topic are chosen first.
This strategy is optimal

Strategic Student Problem

Consider a function h that gives the 'score' on each topic. G(D) is then the overall score.

$$G(D) = \int_x h(x; D)dx$$

Our learning task is to probe the system for N examples $D_{1:N}$ in order to maximize G.

Problem 1: The Strategic Student Problem (SSP)

$$\max_D G(D)$$
$$s.t. \#D = N$$

Strategic Bandit

Algorithm 2 Strategic Bandit (SB)

Require: Initialize $D \leftarrow \emptyset$
Require: Set of topics C and choices a

```
Initialize w_g = 1 w_u = 1
Initialize experts: uniform ξ_u = γ/m and greedy: ξ_g = 0
while learning do
    p = w_g ξ_g + w_u ξ_u
    Select choice a proportional to p
    Draw sample x_a using choice a
    Observe output y_a ~ (C_a, x_a) using a and x_a
    D = D ∪ {x_a, y_a}
    r = Ĝ(D) − Ĝ(D \ {x_a, y_a})
    w_i ← w_i exp( γ ξ_i(a) r / (p(a) m) )
    Update greedy expert:
    q_a ← q_a + η (r − q_a)
    ξ_g(a) = e^{β(q_a − min(q))} / Σ_j e^{β(q_j − min(q))}
end while
```

Where:

$$p = w_g\xi_g + w_u\xi_u$$

$$r = \hat{G}(D) - \hat{G}(D \backslash \{x_a, y_a\})$$

$$w_i \leftarrow w_i exp\left(\gamma\xi_i(a)\frac{r}{p(a)m}\right)$$

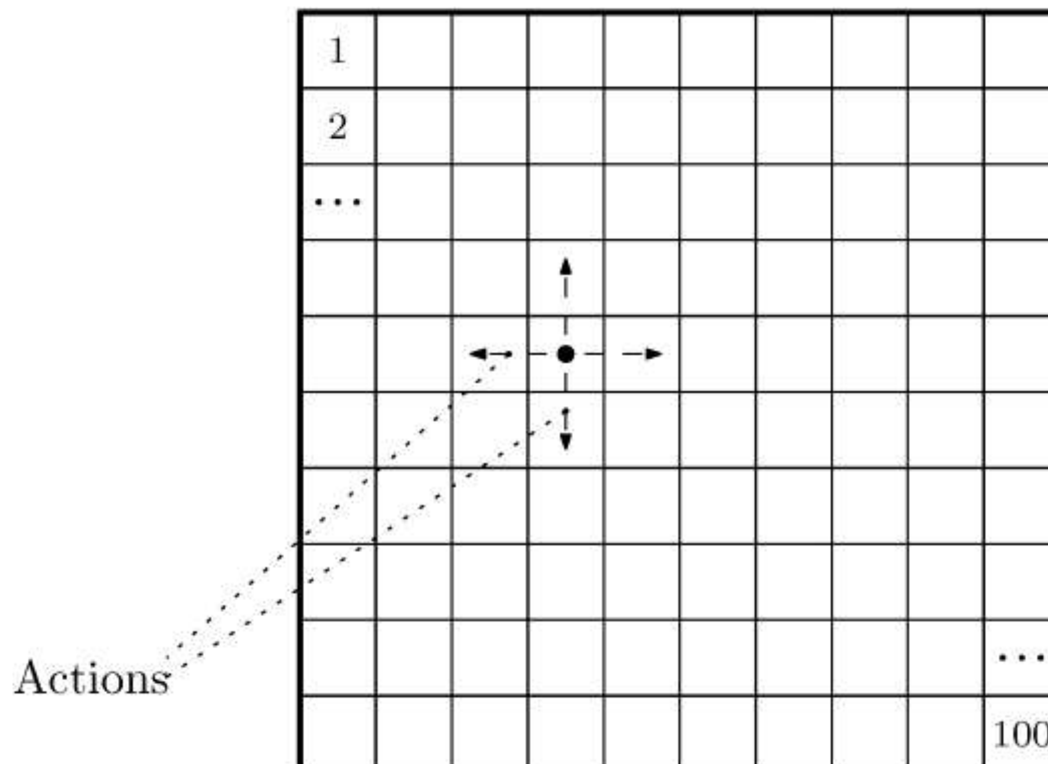
$$q_a \leftarrow q_a + \eta\,(r - q_a)$$

$$\xi_g(a) = \frac{e^{\beta(q_a - \min(q))}}{\sum_j e^{\beta(q_j - \min(q))}}$$

Combination of the ideas of RIAC [Baranes & Oudeyer] algorithm with EXP4 [Auer].

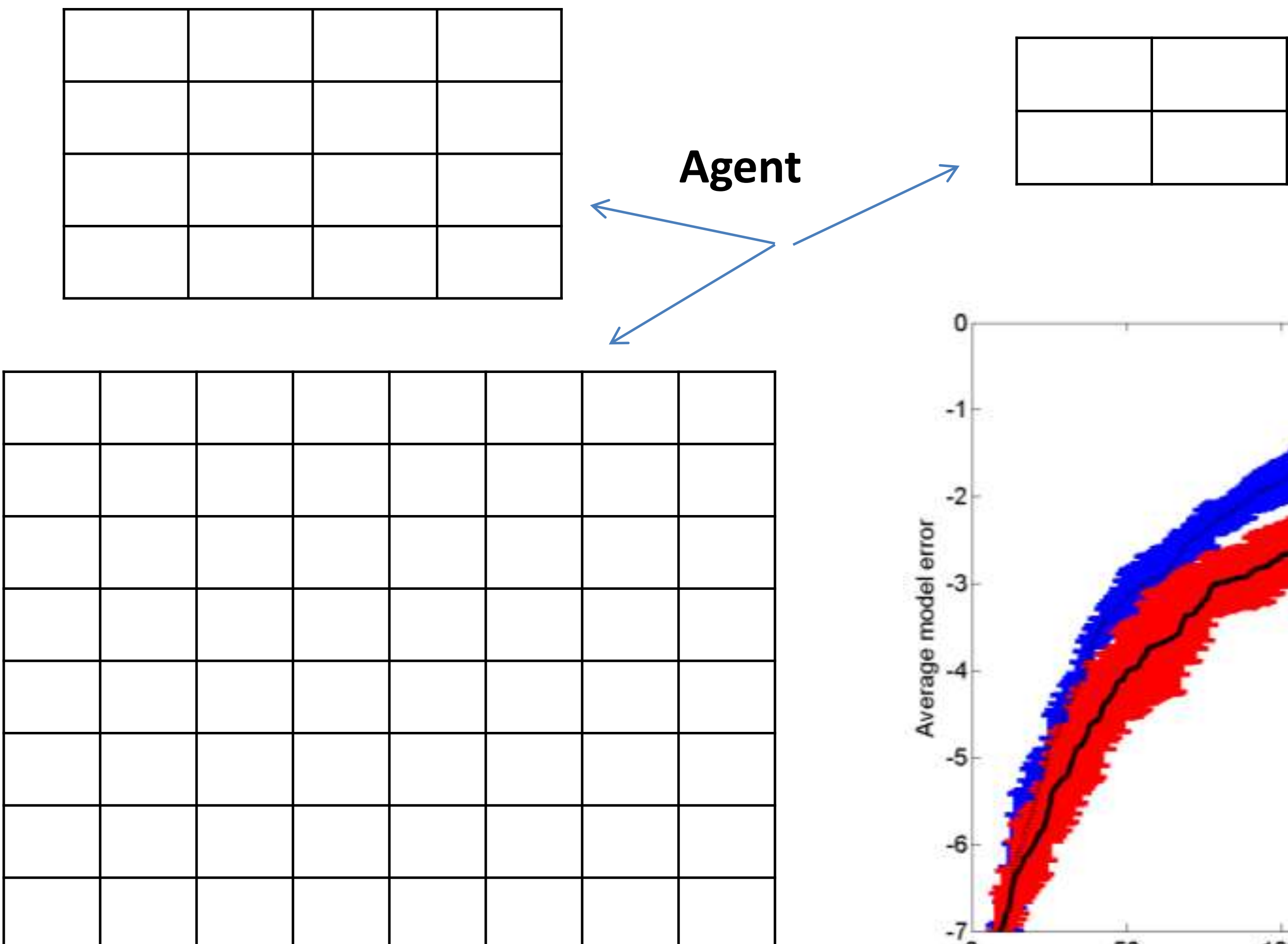
Examples

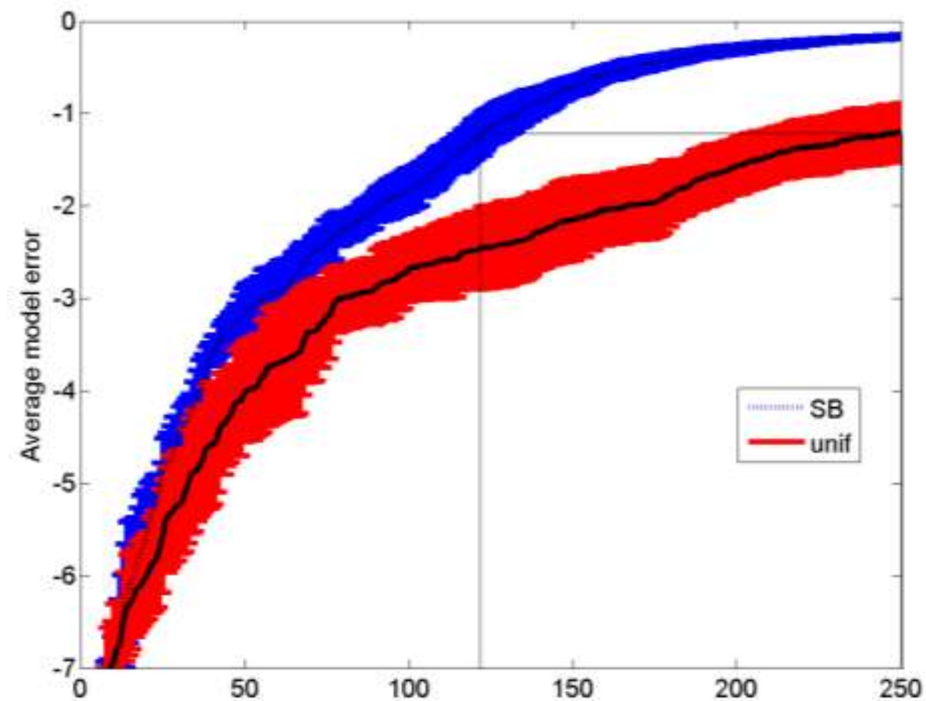
- Four actions available (N, S, E, W)
- Explore the environment to learn the model of the transitions



Examples

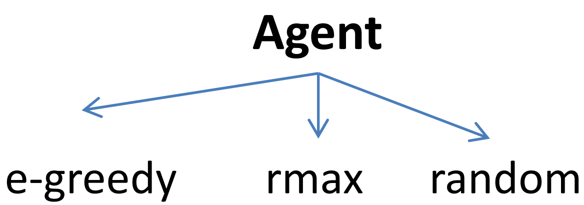
- Learn the model of three environments/options
- At each episode choose which one to explore

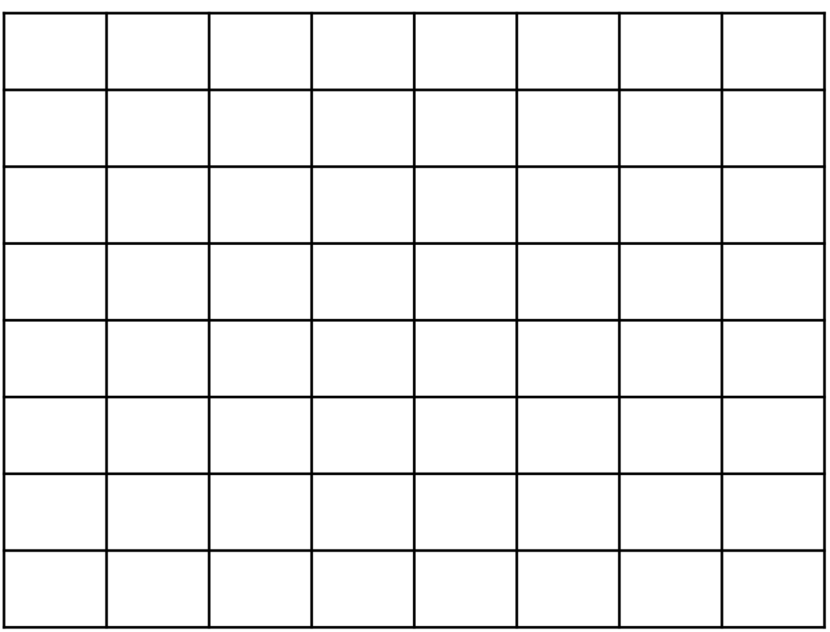


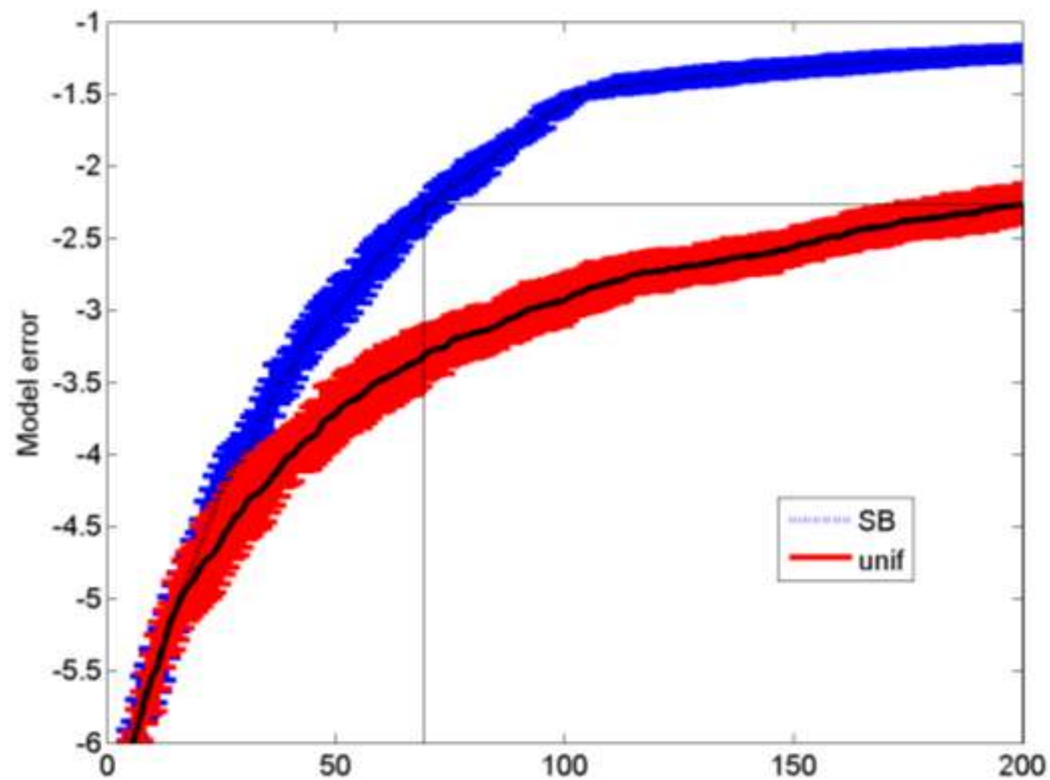


Examples

- Learn the model of an environment using different exploration methods: e-greedy, rmax, random.
- At each episode choose which method to use.







Conclusions

- Active learning can reduce the learning time in many situations
- For robotics active learning can be applied in different problems
- Empirical progress is more robust than simple measures based on uncertainty
- Stochastic approaches are required due to noise and not so well behaved learning functions

- Uncertainty based queries/demos reduce the length of the training sessions and provide measures of quality