COLLEGE OF ENGINEERING

Toward Visualization Methods for Interactive Improvement of MDP Specifications

Sean McGregor, Thomas G. Dietterich,
Ronald Metoyer

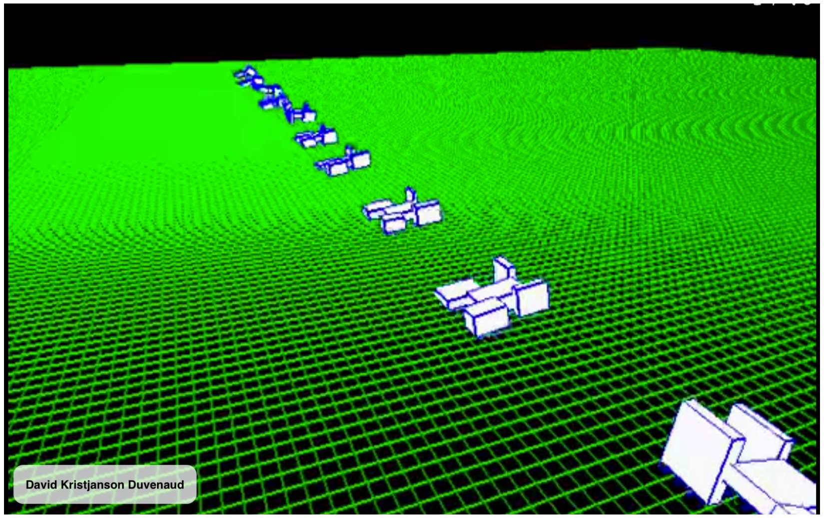
Motivation

Large MDPs → "Complexity"

- Complex software systems are often buggy or misspecified.
  - Does the policy exploit **bugs** in the MDP definition?
  - Does the policy **balance disparate objectives** in an acceptable way?
- Stakeholders lack a means of interrogating the intersection of simulator, values, and policies.
  - How can **stakeholders believe** the policy recommendation?

Examples of "Success"

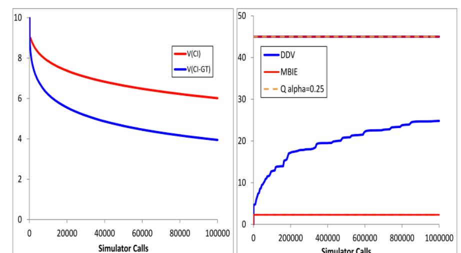
- Debugging
  - Physics Bugs [0]
- Objectives
  - Vibrating Soccer Players [1]
  - Circling Bicycle [1]



[0] https://www.youtube.com/watch?v=STkfUZtR-Vs
[1] Ng, A. Y. (2003). Shaping and policy search in reinforcement learning. University of California, Berkeley.

Specific Motivation of Wildfire

- **Immensely complex models** with numerous potential integration points: vegetative growth, numerous fires spreading spatially, wood products markets, city encroachment, climate change, etc.
- Given a natural wildfire, **SUPPRESS or LET-BURN**



(a) Combination Lock (b) Tamarisk

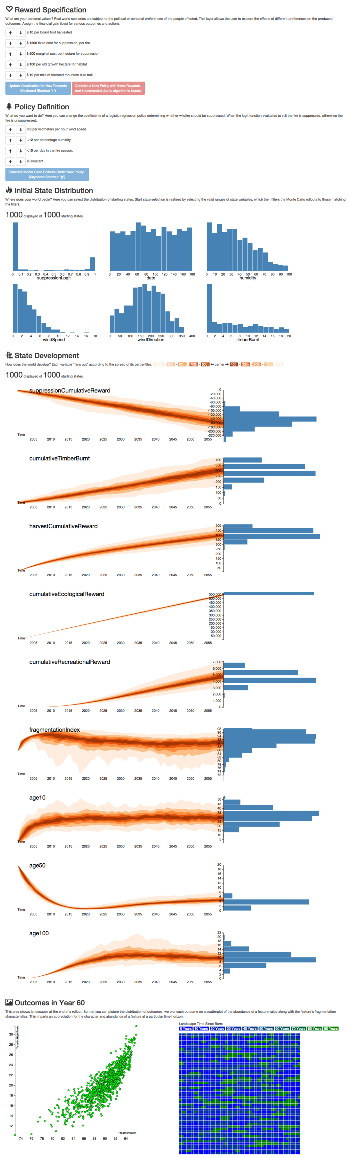
Figure 2: Left: Learning curve for DDV with and without incorporating Good-Turing confidence bounds. Right: Learning curves for MBIE, Q-learning, and DDV on a Tamarisk management MDP.

Dietterich, T., Taleghan, M., & Crowley, M. (2013). PAC Optimal Planning for Invasive Species Management: Improved Exploration for Reinforcement Learning from Simulator-Defined MDPs. Twenty-Seventh AAAI Conference on Artificial Intelligence.

Oregon State UNIVERSITY

Solution: MDP Visualization

1. Control the rewards
2. Control the policy
3. Filter Initial State Distribution
4. View State Evolution
5. Filter Final States
6. View Results



Large MDP Visualization Requirements

- Have a basis in the **MDP formulation**
- **Scale well**
- Provide for **real-time interaction + exploration**
- Explore **distribution of outcomes** rather than single realizations
- Interactively **explore the policy space – Challenge**
- For rapid debugging, **generate new policies** based on changing rewards – **Challenge**

MDP Formulation of Visualization

Reward Definition: $R(s,a)$
Policy Definition: $\pi(s,a)$
Initial State Distribution: P_0
State Development Distribution: P
Final State Examination: S

MDP Formulation of Visualization

Reward Definition: R(s,a)

♡ Reward Specification

- ↑ ↓ **$10** per board foot harvested
- ↑ ↓ **$1,500** fixed cost for suppression, per fire
- ↑ ↓ **$500** marginal cost per hectare for suppression
- ↑ ↓ **$100** per old growth hectare for habitat
- ↑ ↓ **$10** per mile of forested mountain bike trail

[Update Visualization for New Rewards (Keyboard Shortcut "r")]

[Optimize a New Policy with these Rewards. (not implemented due to algorithmic issues)]

Oregon State UNIVERSITY

MDP Formulation of Visualization

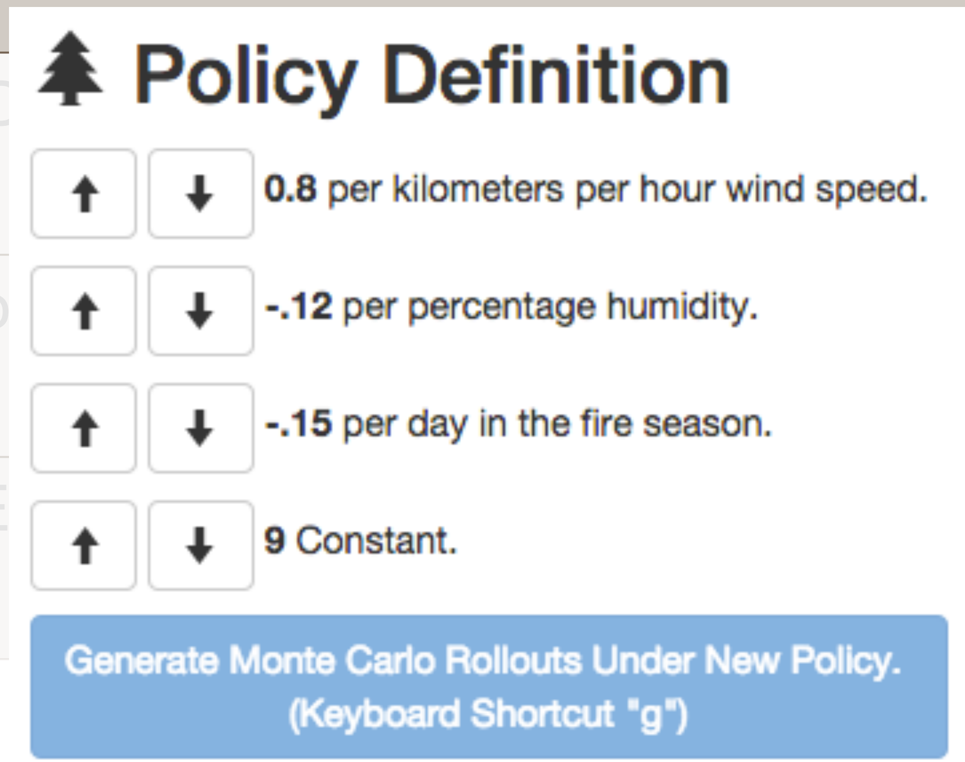
Reward Definition: R(s,a)

Policy Definition: π(s,a)

Initial State D

State Develo

Final State E



MDP Formulation of Visualization

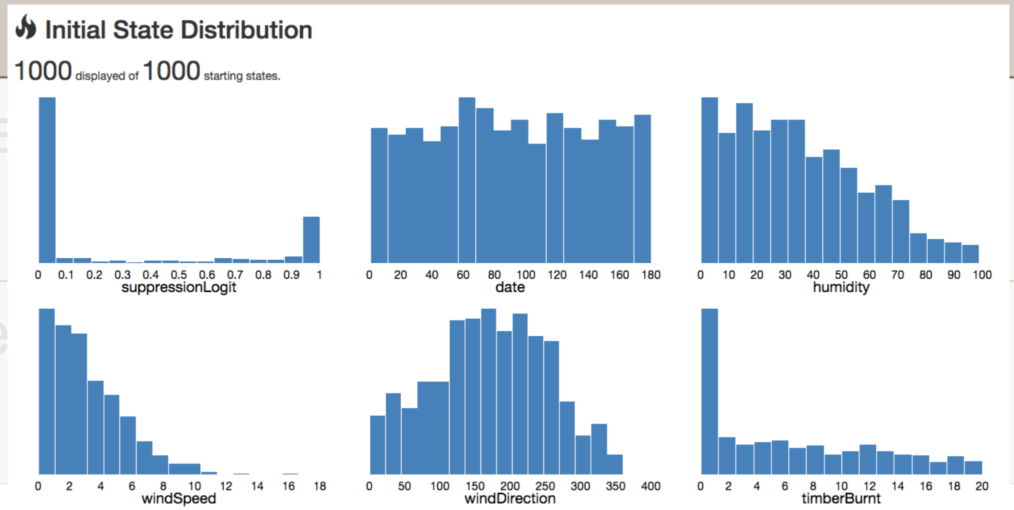
Reward Definition: R(s,a)

Policy Definition: π(s,a)

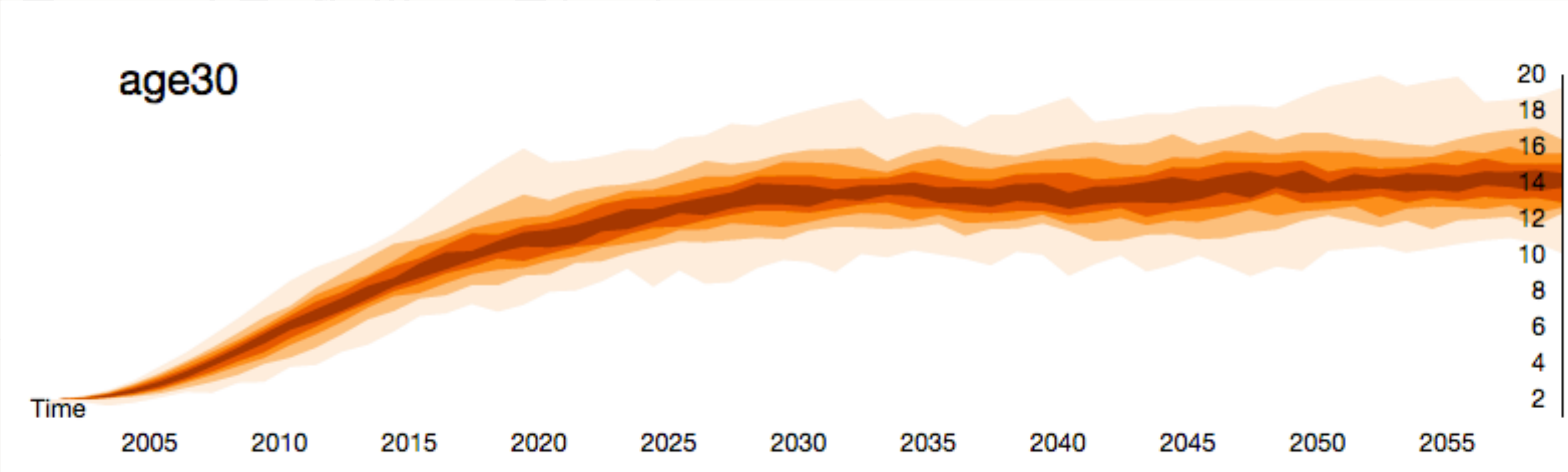
Initial State Distribution: P_0

State Deve

Final State



MDP Formulation of Visualization



State Development Distribution: P

Final State Examination: S

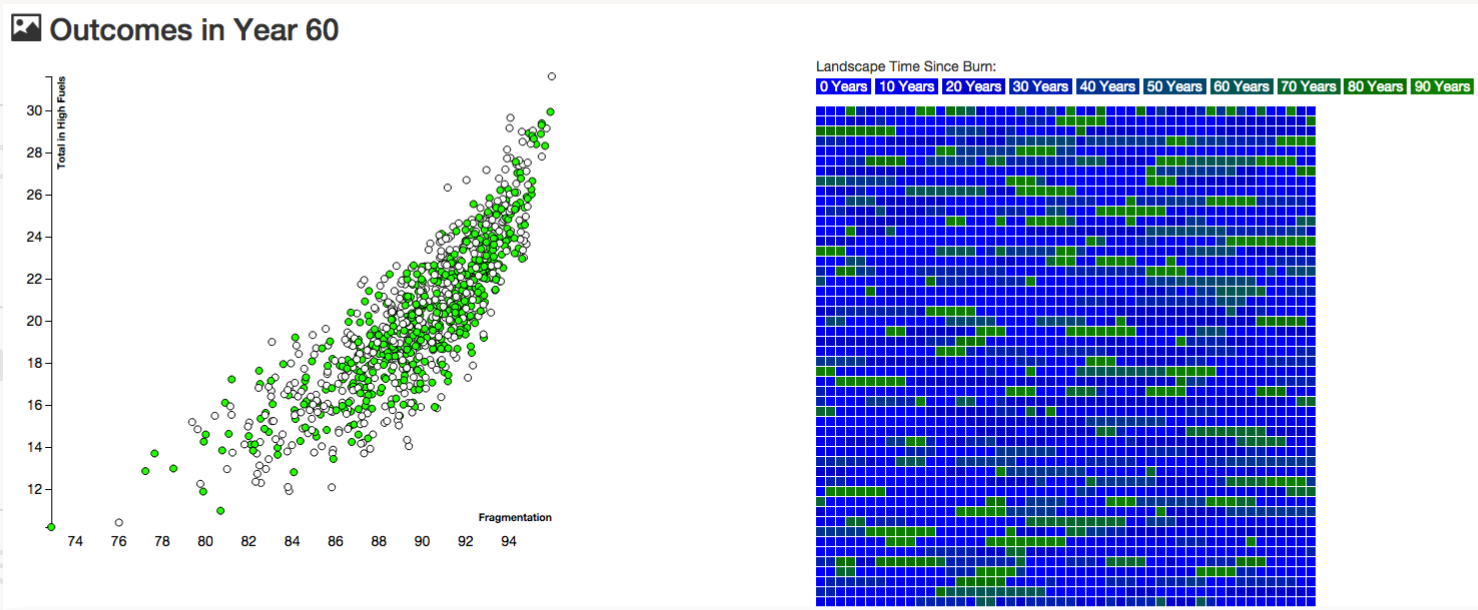
MDP Formulation of Visualization

Reward Definition: R(s,a)

Polic

Initia

State



Final State Examination: S

Reward Specification

What are your personal values? Real world scenarios are subject to the political or personal preferences of the people affected. This layer allows the user to explore the effects of different preferences on the produced outcomes. Assign the financial gain (loss) for various outcomes and actions.

- $ **10** per board foot harvested
- $ **1500** fixed cost for suppression, per fire
- $ **500** marginal cost per hectare for suppression
- $ **100** per old growth hectare for habitat
- $ **10** per mile of forested mountain bike trail

Update Visualization for New Rewards (Keyboard Shortcut "r")

Optimize a New Policy with these Rewards. (not implemented due to algorithmic issues)

Policy Definition

What do you want to do? Here you can change the coefficients of a logistic regression policy determining whether wildfire should be suppressed. When the logit function evaluates to >.5 the fire is suppressed, otherwise the fire is unsuppressed.

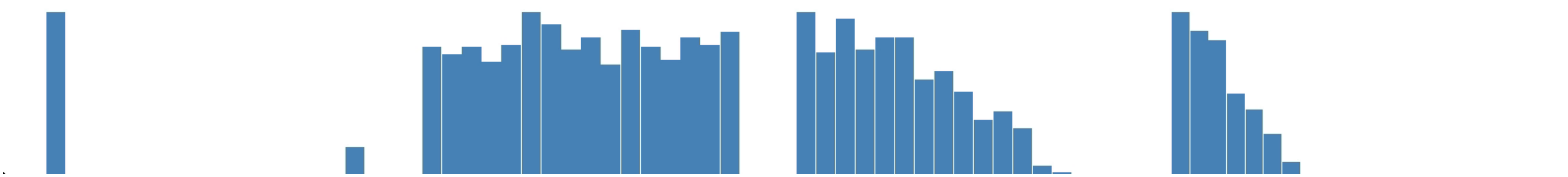
- **0.8** per kilometers per hour wind speed.
- **-.12** per percentage humidity.
- **-.15** per day in the fire season.
- **9** Constant.

Generate Monte Carlo Rollouts Under New Policy. (Keyboard Shortcut "g")

Initial State Distribution

Where does your world begin? Here you can select the distribution of starting states. Start state selection is realized by selecting the valid ranges of state variables, which then filters the Monte Carlo rollouts to those matching the filters.

1000 displayed of 1000 starting states.



Summary

- As **RL matures** it needs **new tools**.
- Powerful tools require **solving algorithmic challenges.**

Algorithmic Challenges

Intelligent Caching of High Dimensional State Transitions → Generate Monte Carlo Rollouts Under New Policy. (Keyboard Shortcut "g")

Quickly optimizing new policies → Optimize a New Policy with these Rewards. (not implemented due to algorithmic issues)

Interactive Demo

AtlasOfLife.com/mdp

Thanks

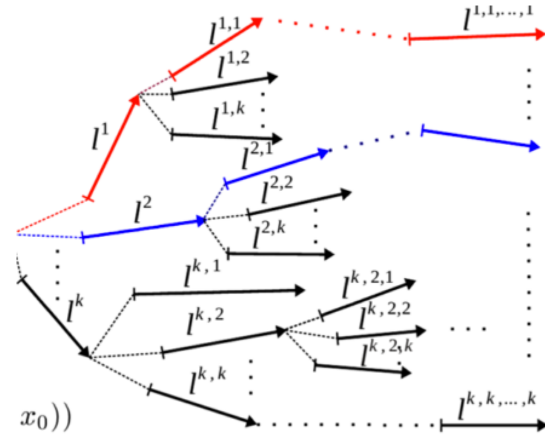
- Workshop organizers
- Thomas Dietterich, Ronald Metoyer
- Claire Montgomery, Rachel Houtman, Mark Crowley, Hailey Buckingham
- National Science Foundation

This material is based upon work supported by the National Science Foundation under Grant No. 1331932.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Generating New Monte Carlo Rollouts

- Problem: Slow simulator prevents changing policies.
- Proposed Solution: pre-compute a database of transitions.



Database "Hit" on (s_0,a_0)

$(s_0,a_0) \longrightarrow (s_{48312},\cdot)$

Database "Miss" on (s_{48312}, a_1)

$(s_{48312},a_1) \cdots\!\!> (s_{9248},a_1) \longrightarrow (s_{42999},\cdot)$

Stitch "Nearest"

Hit

Fonteneau, R., Murphy, S. a, Wehenkel, L., & Ernst, D. (2013). Batch Mode Reinforcement Learning based on the Synthesis of Artificial Trajectories. Annals of Operations Research, 208(1), 383–416.

Updating Policy for New Rewards

- Simulator is still slow
- Optimizing in large MDPs is slow.

"It is important that the physical simulation be reasonably accurate… errors, will inevitably be discovered and exploited... Although this can be a lazy and often amusing approach for debugging a physical modeling system, it is **not necessarily the most practical**." – Karl Sims

Make it practical in non-physical systems!

Sims, K. (1994). Evolving virtual creatures. Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '94, 15–22. doi:10.1145/192161.192167