Toward Visualization Methods for Interactive Improvement of MDP Specifications

Sean McGregor,[1] **Thomas G. Dietterich,**[2] **Ronald Metoyer**[3]
Department of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331
[1]smcgregor@seanbmcgregor.com
[2]tgd@eecs.oregonstate.edu
[3]metoyer@eecs.oregonstate.edu

Abstract

When we define an MDP, specification errors in the reward function can lead to "optimal" policies with unintended behaviors. How can we detect these problems and iteratively correct them? We are developing interactive visualization tools that allow the user to explore Monte Carlo rollouts of one or more proposed policies for an MDP. The tools allow the user to impose constraints, modify the reward function, and obtain approximate feedback on the consequences of these changes. To implement these capabilities, we must solve several algorithmic challenges, which we describe in detail. Our work is inspired by large high-dimensional MDPs whose policies must be validated against a complex set of goals that are not necessarily defined by domain experts.

1 Introduction

As society cedes decision making to computers in tasks as diverse as hiring [8], driving vehicles [7], and wildfire suppression [5], the ability to include humans in the loop of machine optimization gains greater relevance.

Typically a "solution" to a Markov Decision Process (MDP) is a policy maximizing the expected reward, but reward functions may produce tradeoffs unforeseen by the domain experts. Andrew Ng gives examples of domains with unusual policies [9]. In one example, a soccer playing agent is rewarded for touching the ball under the theory that ball possession time is associated with offensive goals. Instead of speeding learning, this shaping reward produced an agent that stands near the ball and "vibrates" to produce the maximum number of ball touches. Such experiences in easily interpretable domains begs the question for large, complex, and simulator-defined MDPs: *does the policy contain the equivalent of a vibrating soccer player and how can we find him?* To bridge machine representations of policies to real-world application, we propose to build from the toolset of information visualization.

Information visualization uses human capacities to examine [14] visual summaries of the nature and relationships contained within data [15]. Traditional machine learning charts can be viewed as rudimentary visualizations, but they do not facilitate exploration.

There are two concomitant challenges in applying information visualization techniques to MDPs. First, a visualization design must be proposed that meaningfully enhances human understanding. Second, the computational issues in supporting real-time interactive visualizations must be solved.

The challenges for machine learning in real-time are accentuated by applications in computational sustainability. This abstract is motivated by the domain of wildfire suppression policies [5], which

attempts to maximize the reward from forested lands while minimizing the management costs of fire suppression and treatments. Wildfire and many other computational sustainability domains are defined by computationally expensive simulators that take minutes to hours to complete a single state transition. Since evaluating a policy typically requires many thousand state transitions, it is infeasible to base policy inspection on being able to generate new transitions in response to user interaction.

In the next section we present work related to visualizing MDPs, including our design that explicitly targets the MDP formulation. Following the visualization we present two challenges for MDP algorithms that must be solved to augment the visualization with additional functionality. While the visualization is useful without solving these algorithm issues, we highlight visualization components in Appendix A that require algorithmic advances before they can function in real time.

2 MDP Visualization

2.1 Related Work

While no MDP information visualizations have been proposed, there are numerous works that could be viewed as visualization for a more restricted class of MDPs.

Several works give systems for exploring decisions at a single time step. Broeksema et al. [3] give a decision analysis tool to examine recommendations made by an expert system. Decisions are plotted as Voronoi diagrams by means of Multiple Correspondence Analysis (MCA), which is a more general version of Multidimensional Scaling (MDS). The Voronoi diagrams don't have a comprehensible coordinate system in two dimensions, but adjacency of attributes plotted over the diagram show how the decision variable changes as other attributes vary.

Migut and Worring [6] compose several information visualizations into a visual analytic dashboard for exploring a dichotomous choice as determined by a machine learning classifier. As is the case for the Broeksema et al. visualization, the system is only applicable to single time steps.

Simulation steering is one branch of visualization that attempts to bring the user into the optimization process by allowing the user to select actions at each state as the system develops. Afzal et al. [1] does simulation steering for epidemic response decisions that shows individual outcomes through time and outcomes on a map. The user can change decisions at various points along a future trajectory to see how the mortality rate responds. This visualization may be viewed as providing user-based optimization for deterministic MDPs. To address the more general case of MDPs with stochastic transitions, it is important to represent a *distribution* of outcomes rather than a single interactive line of actions and responses.

Waser et al. [16] give another simulation steering visualization, "World Lines," where the user is invited to control emergency response to a flooding event. This visualization concentrates on generating a small set of alternative futures based on an action in the present. Stochasticity is not an element of this visualization so it also falls well short. Later refinements to Waser et al's approach [12, 13, 17, 18] expands World Lines to support stochasticity through secondary simulation controls (random levee breach locations) on the probabilistic parameters of the model. These works advance representing the full complexity of MDPs, but the World Lines representation fails to show the multitude of outcomes in a compact representation. Further, the visualization is aimed at interactively constructing a user-generated policy, which ignores machine-generated policies.

In contrast to the current approaches in literature that attempt to give the best visualization possible for a particular problem domain, our approach looks to define the visualization in terms of the abstract properties of a class of problems that has heretofore not been explicitly visualized. While possible to produce more useful visualizations for particular problem domains through specialization, we believe abstract formulation of the visualization leads to maximal re-use of visualization effort and serves as a starting point for specializing visualizations to the properties of specific MDPs.

Next we express our novel visualization in terms of the MDP formulation.

2.2 MDPVis Prototype

We employ the standard formulation of an infinite horizon discounted Markov Decision Process (MDP) with a designated start state distribution [2, 11] $\mathcal{M} = \langle S, A, P, R, \gamma, P_0 \rangle$. S is a finite set of states of the world; A is a finite set of possible actions that can be taken in each state; $P : S \times A \times S \mapsto [0, 1]$ is the conditional probability of entering state s' when action a is executed in state s; $R(s, a)$ is the (deterministic) reward received after performing action a in state s; $\gamma \in (0, 1)$ is the discount factor, and P_0 is the distribution over starting states.

For the purposes of presenting both the diversity of outcomes for a given policy and the relative probability of the outcomes, we give a visualization that breaks the MDP into five parts. The parts are displayed in layers such that each layer fixes a component of the MDP for the subsequent layers. Each layer is premised on having access to a database of Monte Carlo rollouts. An example rendering for a wildfire suppression domain is given in Appendix A. Proceeding from the top layer to the bottom, we have:

a. *Rewards Specification:* $R(s, a)$. Here we permit the user to change the reward function. Interaction at this level requires no novel algorithms since we can reevaluate the existing set of Monte Carlo rollouts. How to quickly optimize a new policy based on the new reward function is covered in section 3.

b. *Policy Definition:* $\pi(s) \mapsto a$. In this layer we state the policy by which we generate the Monte Carlo rollouts. Adding real-time interaction to this layer requires novel approaches outlined in section 3.

c. *Initial State Distribution:* P_0. In this layer we determine the state(s) the domain starts in. This is expressed as a set of histograms displaying the starting state variables. By brushing (interactively subsetting) the histograms the user can update P_0 to fix initial state values. The combination of brushing operations across all histograms supports contingency analysis by showing how the Monte Carlo rollouts develop under different starting conditions and the same policy. In the case of our wildfire domain this sets the state variables of the first fire experienced. By examining the worst case and best case initial fires the user can check the robustness of the policy for more extreme conditions.

d. *State Variable Evolution:* P. In this layer we fix the policy and initial state distribution and present the set of outcomes on a per-variable basis. The state development is given as a fan chart of the percentiles of the state variable values (see Appendix A). By giving the percentiles, we are able to show both the diversity of outcome and the probability of particular values. Annotations to the fan chart provide the ability to explore constraints on the state variables, such as the maximal or minimal probability of a variable value at any time step. Exploration of different time horizons is supported by brushing histograms giving the frequency of variable values at the current time horizon. This supports exploration of both the typical and extreme results of the policy by selecting different percentile ranges.

e. *Outcome Exploration:*

After selecting and brushing the other components of the MDP, we can now show the set of outcomes in more detail. The central challenge at this layer is to present outcomes over many state variables for any time frame of interest. Here we either apply a domain-specific visualization (see fire example in Appendix A), or we run Multi-Dimensional Scaling (MDS) or Multiple Correspondence Analysis (MCA), which collapses high dimensional attribute spaces into the two dimensional screen space [3]. The goal of this layer is to present the outcomes in a way that groups similar states near each other on the screen. The user can then explore the induced clusters of states to create a mental model of the diversity of outcomes.

3 Algorithmic Challenges

The extreme cost of simulators in computational sustainability domains raises the following questions for any visualization that permits the user to modify either the policy or a domain's reward function:

How can we quickly do "what if" analysis for user-defined policies? Typically the policy would be evaluated by examining a set of Monte Carlo rollouts, but generating rollouts is computationally expensive. Methods of avoiding calls to the expensive simulator are necessary.

How can we quickly perform optimizations on a changing reward function as a user "debugs" domain specification? While the MDP visualization given above is functional for exploring a single policy with a pre-generated set of Monte Carlo rollouts, it does not currently optimize new policies. This limits the user's ability to explore alternative domain specifications. Most large MDP algorithms assume they will not be run interactively, which means there is a dearth of approaches that are applicable to visualization.

3.1 Generating Representative Rollouts without Simulation

Fonteneau et al. [4] give a method for synthesizing new artificial rollouts by stitching together state transitions from transitions that have similar ending states to transitions that have similar starting states. This technique, which has been successfully applied to optimizing toy domains, could potentially be applied to a very large database of state transitions as a way of pre-computing the results for many different policies. When a new policy needs Monte Carlo rollouts, the database is queried for transitions rather than calling an expensive simulator.

3.2 Updating the Optimal Policy when the User Changes the Reward Function

Domain specification for complex tasks is an iterative process that can often involve changes to the reward function. In the case of the wildfire suppression problem, such changes could involve changing the relative value of species, recreation, timber, fire suppression expenditures, and scenic views. We wish to quickly give users a strong policy as a starting point for exploration.

While obtaining a good policy in realtime for the user remains challenging, we take inspiration from Ng et al. [10] and limit the policy space to stochastically defined policies. We can then solve a linear program over the database of rollouts that minimizes the likelihood of generating the less desirable rollouts. While it is unlikely that policies generated in this manner will be deemed optimal, it should give a starting point.

4 Conclusion

Here we outlined the benefit of exploring MDP policies via visualization, presented a working visualization, and suggested ways to expand the functionality of the visualization to cover more policies and policy generation.

Acknowledgments

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.

References

[1] Afzal, Shehzad, Maciejewski, Ross, and Ebert, Davis S. Visual Analytics Decision Support Environment for Epidemic Modeling and Response Evaluation. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 191–200. IEEE, 2011. ISBN 9781467300131.

[2] Bellman, Richard. *Dynamic Programming*. Princeton University Press, New Jersey, 1957.

[3] Broeksema, Bertjan, Baudel, Thomas, Telea, Alex, and Crisafulli, Paolo. Decision Exploration Lab: A Visual Analytics Solution for Decision Management. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1972–1981, 2013.

[4] Fonteneau, Raphael, Murphy, Susan a, Wehenkel, Louis, and Ernst, Damien. Batch Mode Reinforcement Learning based on the Synthesis of Artificial Trajectories. *Annals of operations research*, 208(1):383–416, September 2013. ISSN 0254-5330. doi:

10.1007/s10479-012-1248-5. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3773886&tool=pmcentrez&rendertype=abstract.

[5] Houtman, Rachel M., Montgomery, Claire A., Gagnon, Aaron R., Calkin, David E., Dietterich, Thomas G., McGregor, Sean, and Crowley, Mark. Allowing a Wildfire to Burn: Estimating the Effect on Future Fire Suppression Costs. *International Journal of Wildland Fire*, 22(7): 871–882, 2013.

[6] Migut, Malgorzata and Worring, Marcel. Visual Exploration of Classification Models for Risk Assessment. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 11–18, 2010. ISBN 9781424494873.

[7] Miller, Joe. Google's driverless cars designed to exceed speed limit, 2014. URL http://www.bbc.com/news/technology-28851996.

[8] National Public Radio. Episode 509: Will A Computer Decide Whether You Get Your Next Job?, 2014. URL http://www.npr.org/blogs/money/2014/01/15/262789258/episode-509-will-a-computer-decide-whether-you-get-your-next-job.

[9] Ng, Andrew Y. *Shaping and policy search in reinforcement learning*. Doctor of philosophy, University of California, Berkeley, 2003. URL http://www.cs.ubc.ca/$\sim$nando/550-2006/handouts/andrew-ng.pdf.

[10] Ng, Andrew Y. and Jordan, M. PEGASUS : A policy search method for large MDPs and POMDPs. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000. URL http://dl.acm.org/citation.cfm?id=2073994.

[11] Puterman, Martin. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1st edition, 1994.

[12] Ribicic, Hrvoje, Waser, Jurgen, Fuchs, Raphael, Bloschl, Gunter, and Groller, Eduard. Visual Analysis and Steering of Flooding Simulations. *IEEE transactions on visualization and computer graphics*, 19(6):1062–1075, 2013.

[13] Schindler, Benjamin, Ribicic, Hrvoje, Fuchs, Raphael, and Peikert, Ronald. Multiverse data-flow control. In *IEEE Transactions on Visualization and Computer Graphics*, volume 19, pp. 1005–1019, 2013. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6329370.

[14] Treisman, Anne. Preattentive Processing in Vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177, 1985.

[15] Tufte, Edward R. *The Visual Display of Quantitative Information*, volume 31. Graphics press, Cheshire, CT, 1983.

[16] Waser, Jürgen, Fuchs, Raphael, Ribicic, Hrvoje, Schindler, Benjamin, Bloschl, Gunther, and Groller, M. Eduard. World Lines. *IEEE transactions on visualization and computer graphics*, 16(6):1458–1467, 2010.

[17] Waser, Jürgen, Ribičić, Hrvoje, Fuchs, Raphael, Hirsch, Christian, Schindler, Benjamin, Blöschl, Günther, and Gröller, M Eduard. Nodes on ropes: a comprehensive data and control flow for steering ensemble simulations. *IEEE transactions on visualization and computer graphics*, 17(12):1872–81, December 2011. ISSN 1941-0506. doi: 10.1109/TVCG.2011.225. URL http://www.ncbi.nlm.nih.gov/pubmed/22034304.

[18] Waser, Jürgen, Konev, A, Sadransky, B, Horváth, Z, Ribicic, Hrvoje, Carnecky, R., Kluding, P., and Schindler, B. Many Plans: Multidimensional Ensembles for Visual Decision Support in Flood Management. *Eurographic Conference on Visualization (EuroVis)*, 33(3), 2014.

Appendix A

♡ Reward Specification

More Less **$5** per board foot harvested

More Less **$1,500** fixed cost for suppression

More Less **$500** marginal cost per hectare for suppression

More Less **$10** per old growth tree for habitat

More Less **$100** per mile of forested mountain bike trail

Optimize a new policy with these rewards. (not implemented due to algorithmic issues)

🌲 Policy Definition

Here you can change the coefficients of a logistic regression policy determining whether wildfire should be suppressed.

↑ ↓ **-1.16** for wind (MPH)

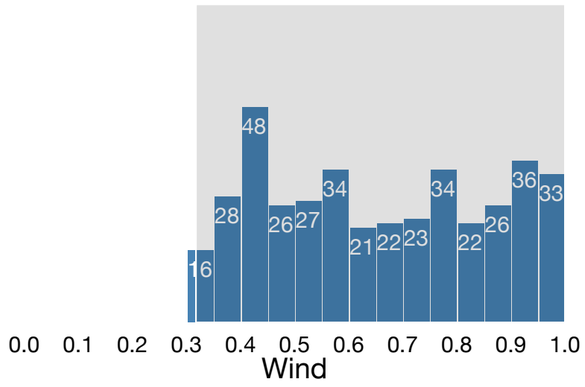
↑ ↓ **0.97** for Energy Release Component

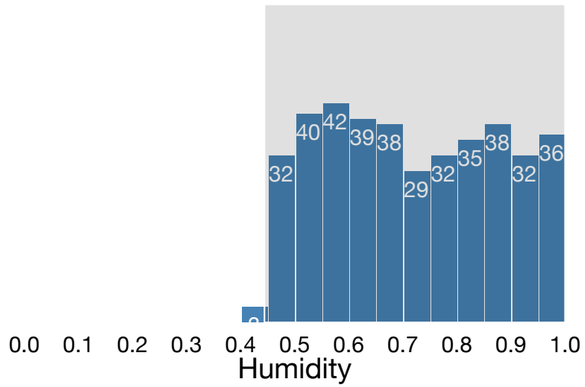
↑ ↓ **-1.05** for Humidity

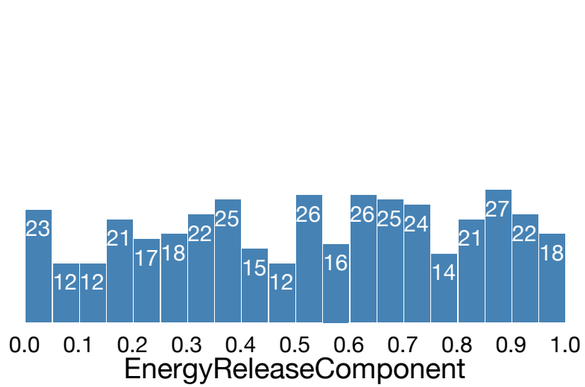
↑ ↓ **0.85** for intercept

Generate Monte Carlo Rollouts. (not implemented due to algorithmic issues)

🔥 Initial State Distribution

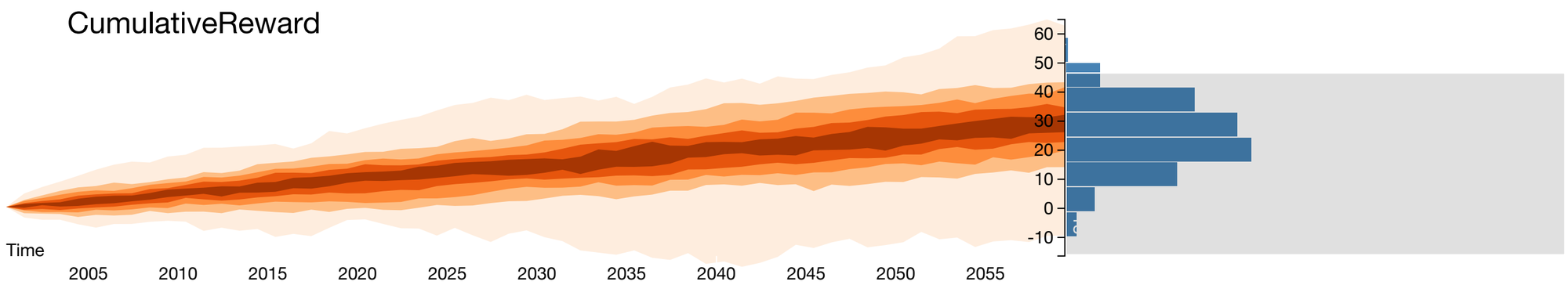


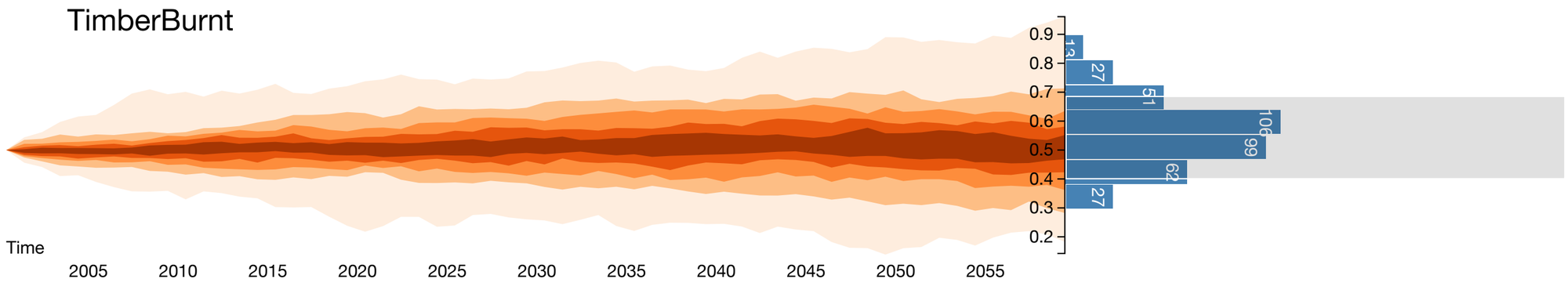


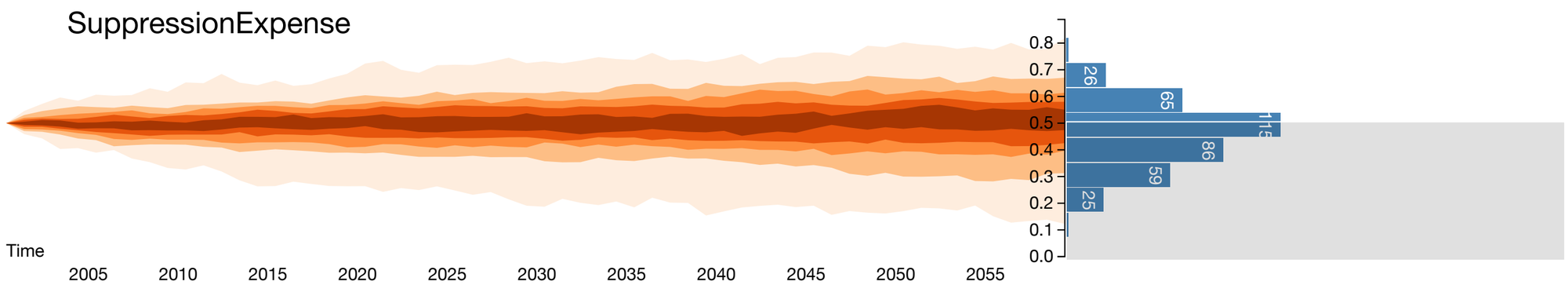


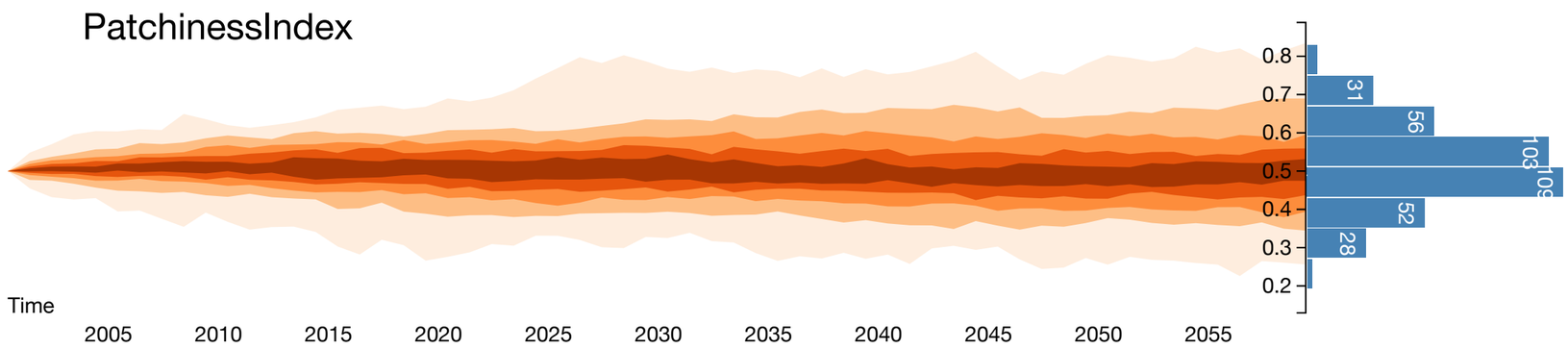
State Development

Each attribute "fans out" according to the spread of its percentiles. The darkest color represents the middle (40th to 60th) 20 percentiles, with each color shift representing a further 10th percentile until the color transitions to blank at the 100th percentile.









Outcomes in Year 60

This area shows landscapes at the end of a rollout. So that you can picture the distribution of outcomes, we plot each outcome on a scatterplot of the abundance of a feature value along with the feature's fragmentation characteristics. This imparts an appreciation for the character and abundance of a feature at a particular time horizon.