Multi-Objective MDPs for Decision Support

Dan Lizotte† and Eric Laber‡

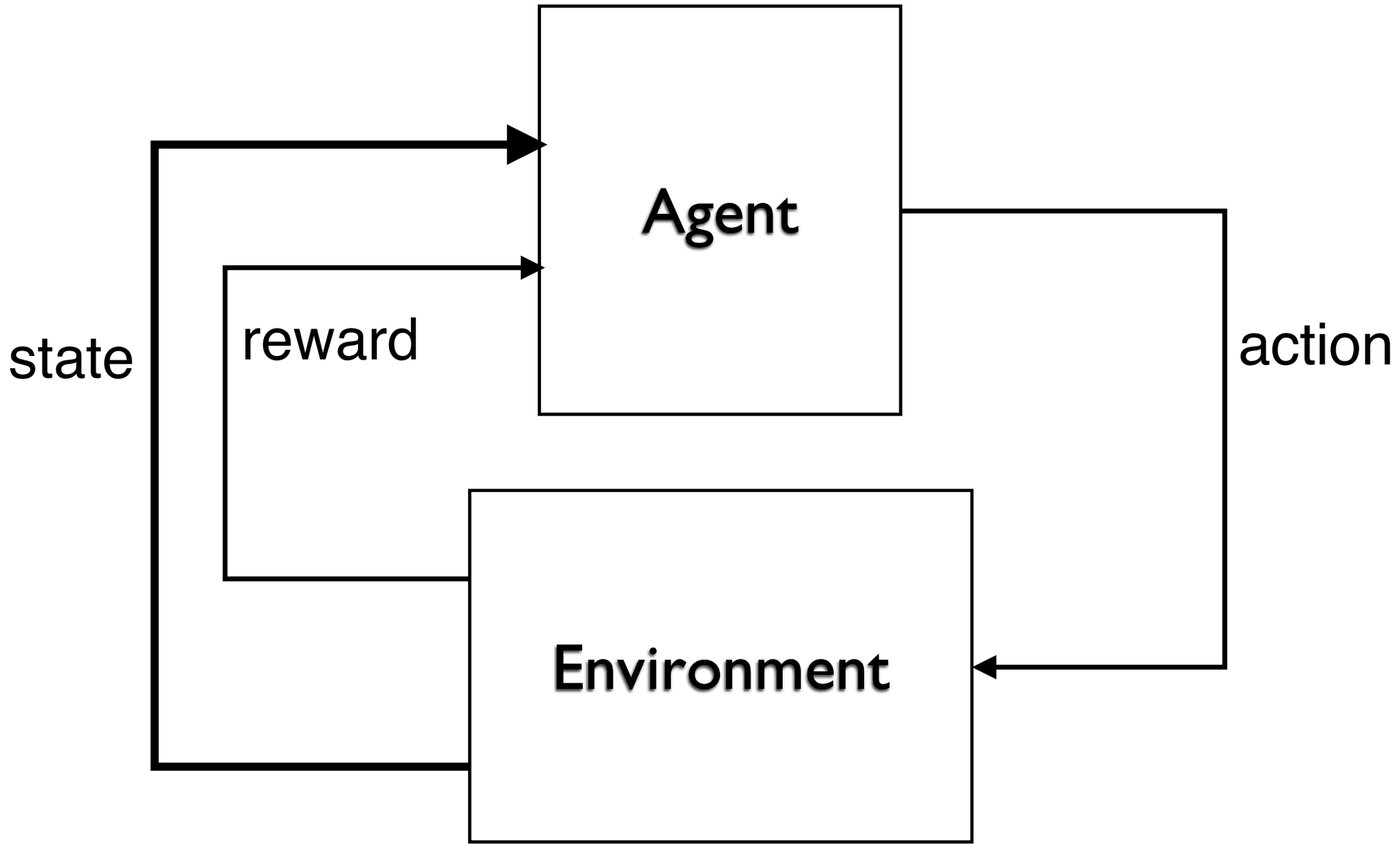
†Computer Science and Epidemiology & Biostatistics
The University of Western Ontario

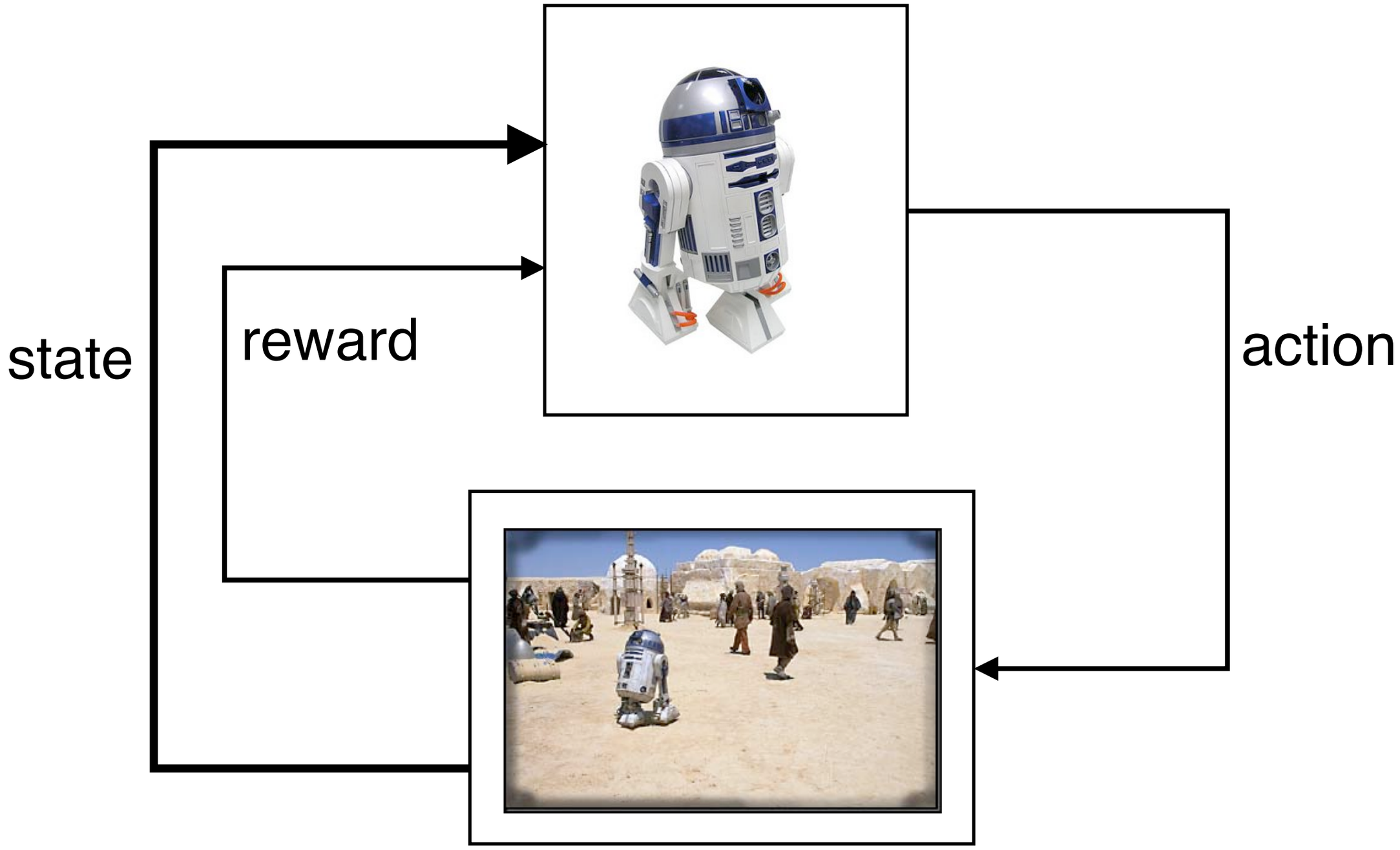
‡Department of Statistics
North Carolina State University

RLDM : University of Alberta : 9 June 2015

MOMDPs for Decision Support

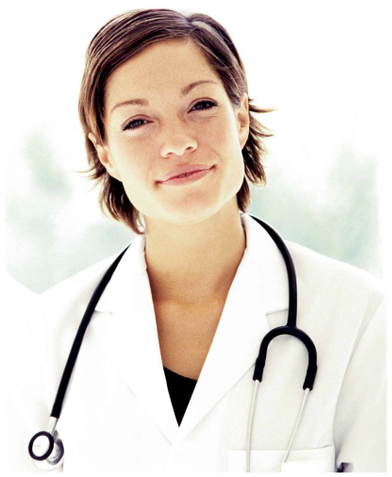
- Main Idea
  - Automating and Supporting Sequential Decision-making are similar but different problems.
- One facet: Accommodating multiple reward signals
  - Part of MOMDP literature (see survey, Roijers et al. 2013 JMLR)
  - "Non-deterministic Fitted-Q for Multiple Objectives"
  - Output is a non-deterministic (set-valued) policy
- Example: CATIE Schizophrenia Trial


Agent
state
reward
action
Environment


state
reward
action

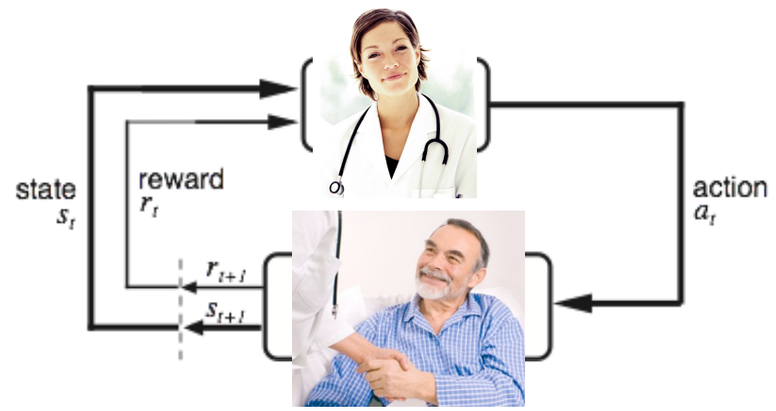
Status

Outcome

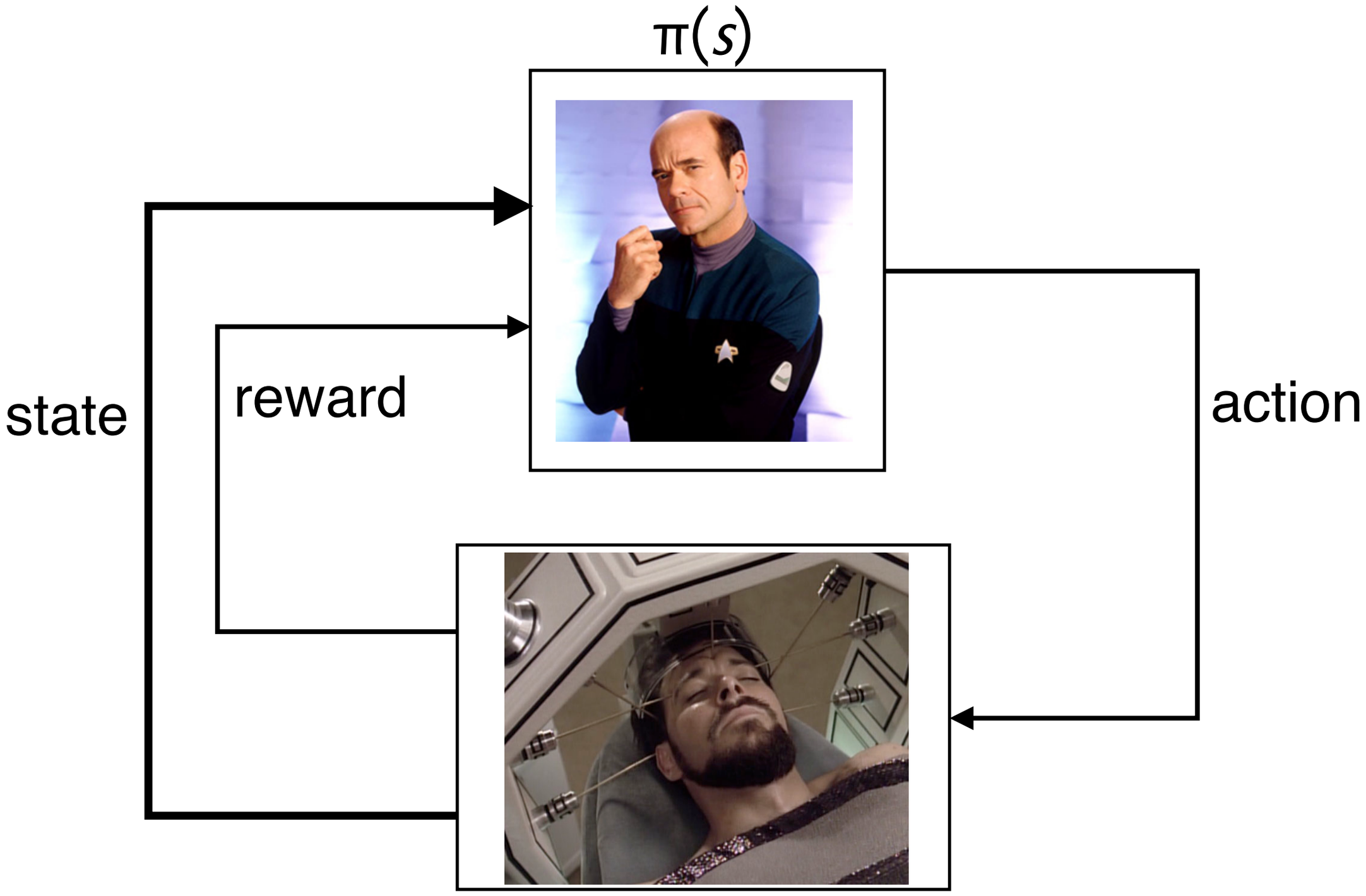


Treatment

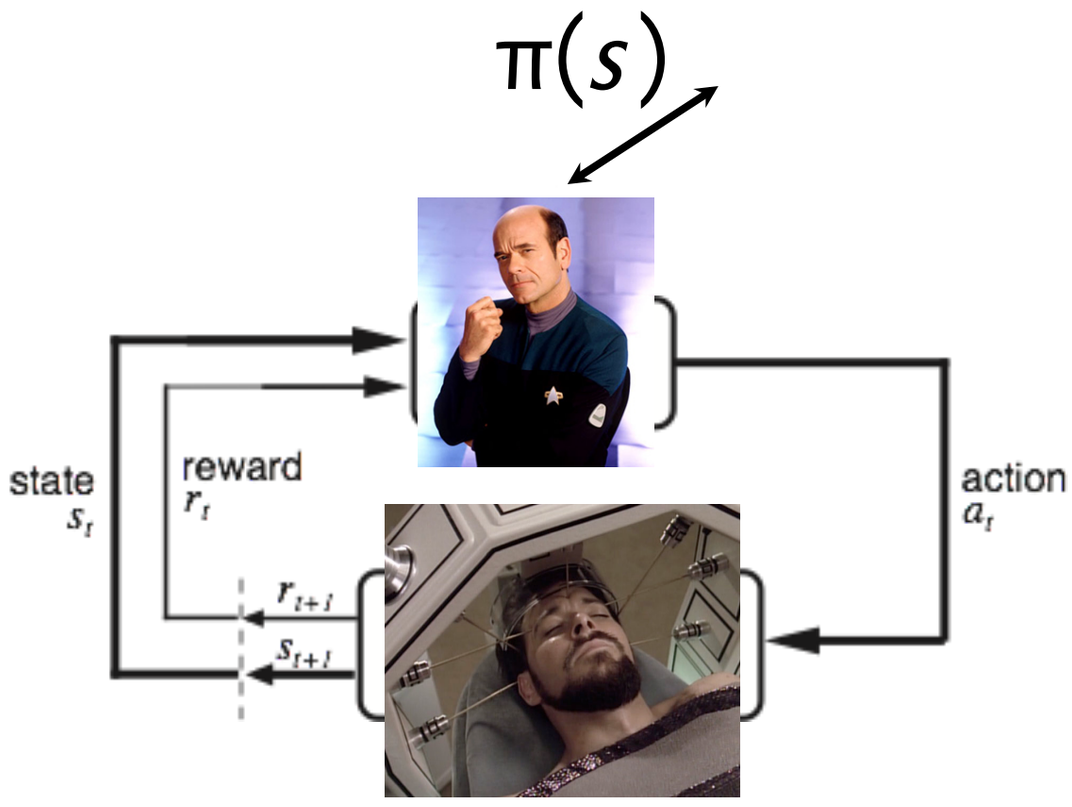
Reinforcement Learning



- Why is are RL methods potentially useful for clinical decision-making?
- Goal is to maximize long term success
- Makes context-dependent, individualized decisions
- Handles uncertain futures naturally
- **Policies can be learned from data**


$\pi(s)$
state
reward
action

Decision Support Agent

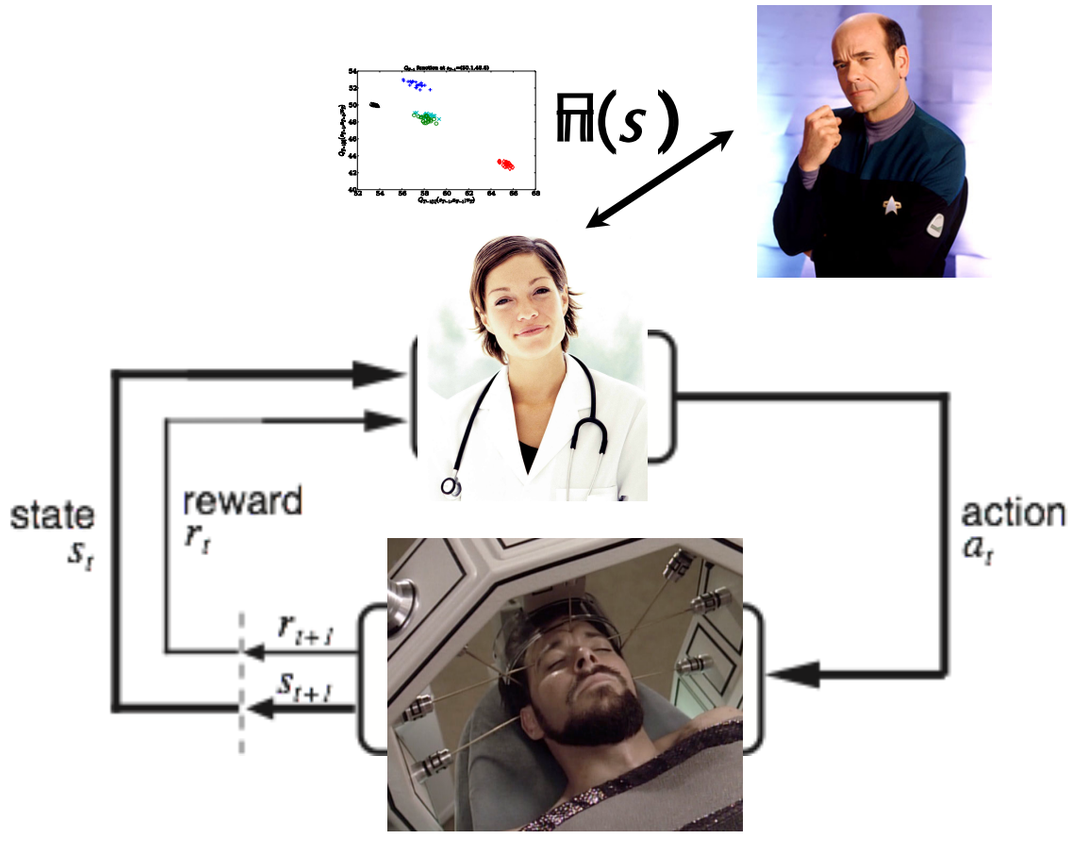


Recommend rather than administer treatments, still leverage autonomous agent idea

Defining the Reward

- **What is the "right" scalar reward?**
  - Symptom relief?
  - Side-effects?
  - Cost?
  - Quality of life?
  - Some combination?
- Potentially different function for every decision-maker, every patient…
- …and therefore no unique optimal action.
- **BUT** we have a human expert in the loop

Decision **Support**



- Provide richer output than just a single action
  - "*Linear Fitted-Q with Multiple Reward Functions*"
    (DL, Bowling, Murphy, *ICML* and *JMLR*)
  - "*Set-Valued DTRs for Competing Outcomes*"
    (Laber, DL, Ferguson, *Biometrics*)
  - "*Multi-Objective Markov Decision Processes for Data-Driven Decision Support*"
    (Lizotte, Laber, in submission)

Non-Deterministic Policies from Multiple Rewards

1. Consider a list of relevant rewards
2. Screen out actions that are "definitely bad"
3. Recommend the set of remaining actions

“Fitted-Q,” Briefly

- Given Data: $(S_1, A_1, S_2, A_2, R)^n$
- $Q_2(s_2, a_2) \approx \mathbb{E}_R[R \mid S_2 = s_2, A_2 = a_2]$
- $\pi_2(s_2) = \text{arg}\max_a Q_2(s_2, a)$
- $Q_1(s_1, a_1) \approx \mathbb{E}_{S_2}[Q_2(S_2, \boldsymbol{\pi_2(S_2)}) \mid S_1 = s_1, A_1 = a_1]$
- $\pi_1(s_1) = \text{arg}\max_a Q_1(s_1, a)$

Non-deterministic (Set-Valued) Policies

Suppose two different rewards are important: *Y*, *Z*

Screen out an action if it is ***dominated*** by another action, i.e., if ***worse according to both rewards*** than some other action.

Recommend remaining **set** of actions.

Output:

Non-deterministic policy that maps states to **sets** of actions.

Stage 2 Decision

- Consider a state s_2, suppose:
  - $(Q^Y_2(s_2,1), Q^Z_2(s_2,1)) = (9,9)$
  - $(Q^Y_2(s_2,-1), Q^Z_2(s_2,-1)) = (2,2)$
- Clearly treatment 1 is preferable to treatment -1. (E.g., point (9,9) Pareto dominates point (2,2).)
- Recommend {1} (and screen out {-1})
- $\Pi_2(s_2) = \{1\}$

Stage 2 Decision

- Consider another state s_2, suppose:
  - $(Q^Y_2(s_2,1), Q^Z_2(s_2,1)) = (8,1)$
  - $(Q^Y_2(s_2,-1), Q^Z_2(s_2,-1)) = (1,8)$
- Which stage 2 treatment should we assume at this s_2?
  Either is plausible depending on preference.
  (Neither dominates the other.)
- Recommend {1,-1} (no screening)
- $\Pi_2(s_2) = \{1,-1\}$

Q-Learning Redux

- Data: $(S_1, A_1, S_2, A_2, Y, Z)^n$

- Estimate
  - $Q^Y{}_2(s_2, a_2) \approx \mathbb{E}[Y | S_2 = s_2, A_2 = a_2]$
  - $Q^Z{}_2(s_2, a_2) \approx \mathbb{E}[Z | S_2 = s_2, A_2 = a_2]$

- Determine **set-valued** policy Π_2

- Estimate
  - $Q^Y{}_1(s_1, a_1) \approx \mathbb{E}[Q^Y{}_2(S_2, ??) | S_1 = s_1, A_1 = a_1]$
  - $Q^Z{}_1(s_1, a_1) \approx \mathbb{E}[Q^Z{}_2(S_2, ??) | S_1 = s_1, A_1 = a_1]$

Compatibility

- We will consider a collection of stage 2 policies π_2 that are "compatible" with the **set-valued policy Π_2** and then estimate
  - $Q^Y_1(s_1,a_1;\pi_2) \approx \mathbb{E}[Q^Y_2(S_2, \pi_2(S_2))|S_1=s_1,A_1=a_1]$
  - $Q^Z_1(s_1,a_1;\pi_2) \approx \mathbb{E}[Q^Z_2(S_2, \pi_2(S_2))|S_1=s_1,A_1=a_1]$

Non-Deterministic Policy

$\Pi_2(s_2^{(1)}) = \{1\}$

$\Pi_2(s_2^{(2)}) = \{1\}$

…

$\Pi_2(s_2^{(n-3)}) = \{1,-1\}$

$\Pi_2(s_2^{(n-2)}) = \{1,-1\}$

$\Pi_2(s_2^{(n-1)}) = \{1,-1\}$

$\Pi_2(s_2^{(n)}) = \{1,-1\}$

Compatible Policy

$\Pi_2(s_2^{(1)}) = \{1\}$ $\quad \pi_2(s_2^{(1)}) = 1$

$\Pi_2(s_2^{(2)}) = \{1\}$ $\quad \pi_2(s_2^{(2)}) = 1$

…

$\Pi_2(s_2^{(n-3)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-3)}) = -1$

$\Pi_2(s_2^{(n-2)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-2)}) = -1$

$\Pi_2(s_2^{(n-1)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-1)}) = -1$

$\Pi_2(s_2^{(n)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n)}) = -1$

Compatible Policy

$\Pi_2(s_2^{(1)}) = \{1\}$ $\quad \pi_2(s_2^{(1)}) = 1$

$\Pi_2(s_2^{(2)}) = \{1\}$ $\quad \pi_2(s_2^{(2)}) = 1$

…

$\Pi_2(s_2^{(n-3)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-3)}) = 1$

$\Pi_2(s_2^{(n-2)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-2)}) = 1$

$\Pi_2(s_2^{(n-1)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-1)}) = -1$

$\Pi_2(s_2^{(n)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n)}) = -1$

Compatible Policy

$\Pi_2(s_2^{(1)}) = \{1\}$ $\quad \pi_2(s_2^{(1)}) = 1$

$\Pi_2(s_2^{(2)}) = \{1\}$ $\quad \pi_2(s_2^{(2)}) = 1$

…

$\Pi_2(s_2^{(n-3)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-3)}) = 1$

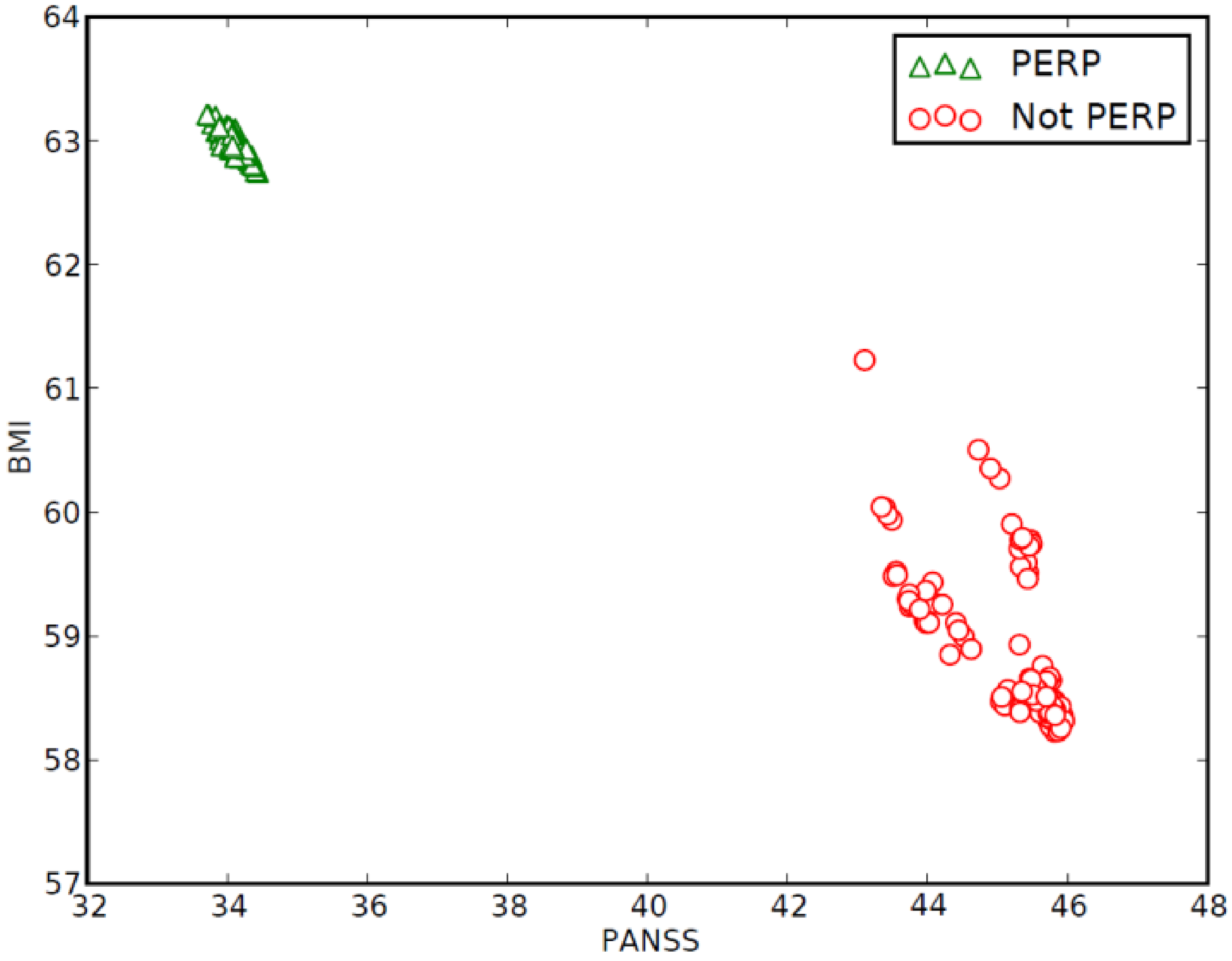
$\Pi_2(s_2^{(n-2)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-2)}) = -1$

$\Pi_2(s_2^{(n-1)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n-1)}) = 1$

$\Pi_2(s_2^{(n)}) = \{1,-1\}$ $\quad \pi_2(s_2^{(n)}) = -1$

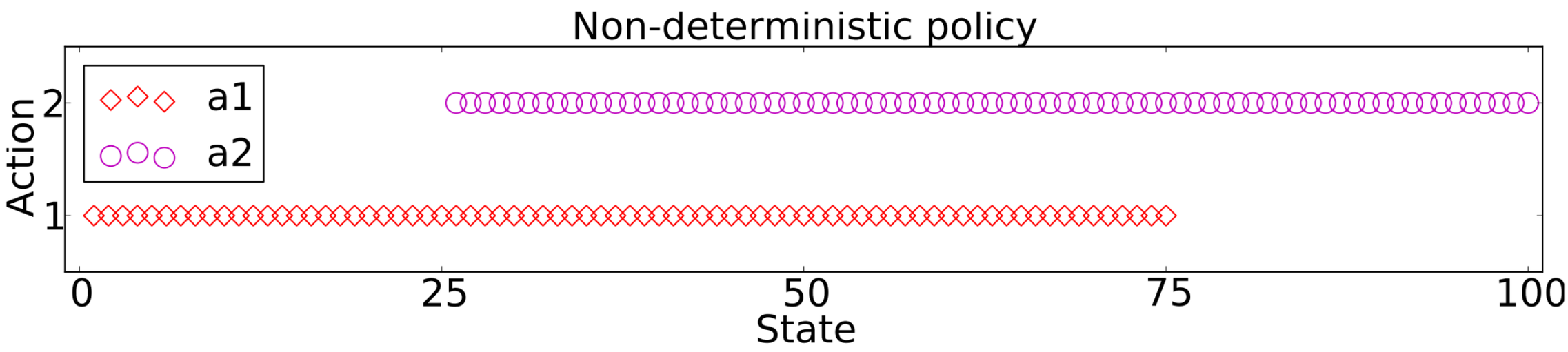
Q-Learning Redux: Learning Q_1

- Learn, for a set of compatible π_2
  - $Q^Y{}_1(s_1,a_1;\pi_2) \approx \mathbb{E}[Q^Y{}_2(S_2, \pi_2(S_2))|S_1=s_1,A_1=a_1]$
  - $Q^Z{}_1(s_1,a_1;\pi_2) \approx \mathbb{E}[Q^Z{}_2(S_2, \pi_2(S_2))|S_1=s_1,A_1=a_1]$
- For any state and action, we get a **set** of possible expected outcomes.


PERP
Not PERP
BMI
PANSS
64
63
62
61
60
59
58
57
32
34
36
38
40
42
44
46
48

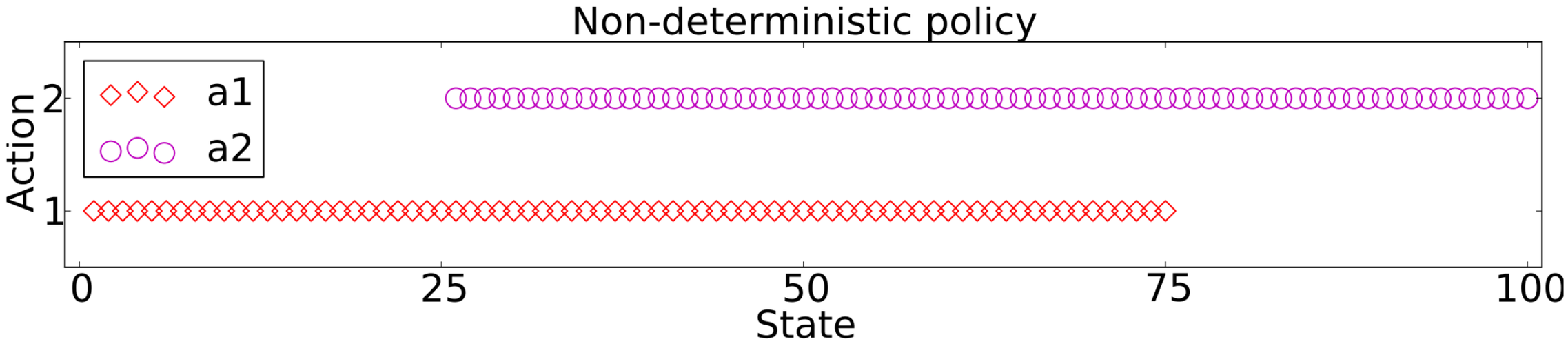
Practical Issues

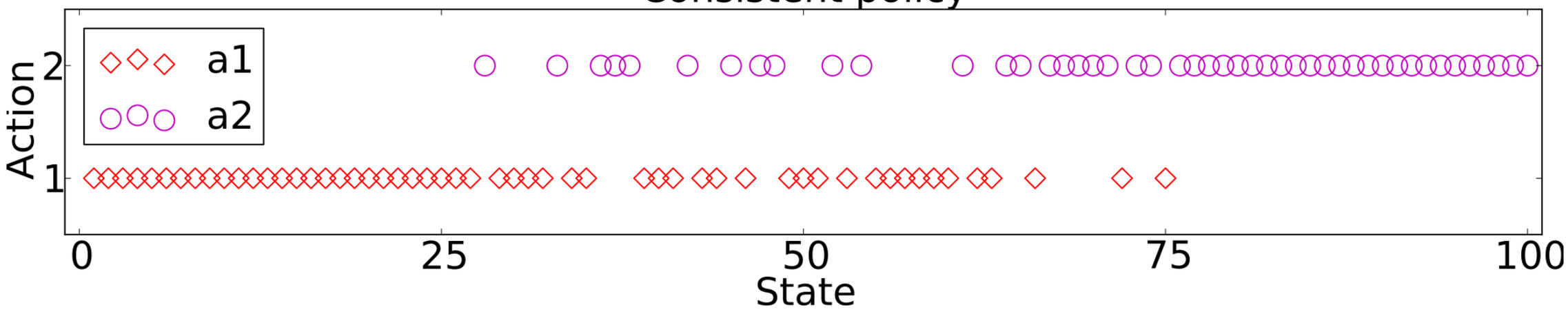
- Worst case, there are $|A|^n$ compatible π_2
- But, many are highly complex and, we argue, inadmissible
- Suppose a linear approximation space for Q. Then there are many compatible policies for which there is **no scalar reward** that would cause us to learn those policies.

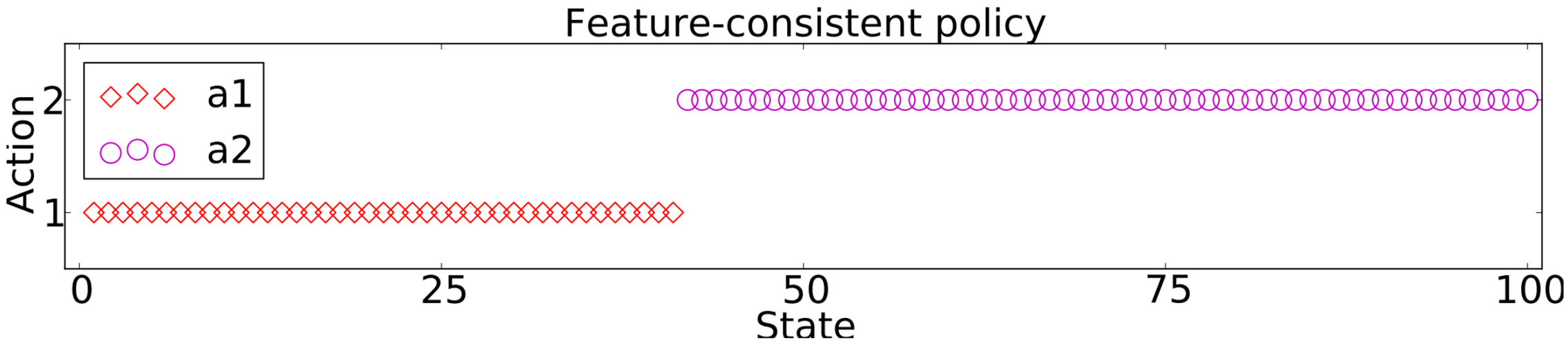

Non-deterministic policy
Action
2
1
a1
a2
0
25
50
75
100
State

"Feature Consistency"

- A policy π is feature-consistent with a set-valued policy Π if:
  - For all *s*, π(*s*) is in Π(*s*) (it's compatible)
  - π is "of the form" $\text{argmax}_a Q(s,a)$
  - These are linearly separable in the state space.
- **There are only $O(n^{\text{Ndim}(Q)})$ of these, where Ndim is the *Natarajan dimension***
- **For GLM Q functions, we enumerate with MIP expressing linear separations of the data. (e.g. CPLEX)**


Non-deterministic policy
Action
2
1
a1
a2
0
25
50
75
100
State

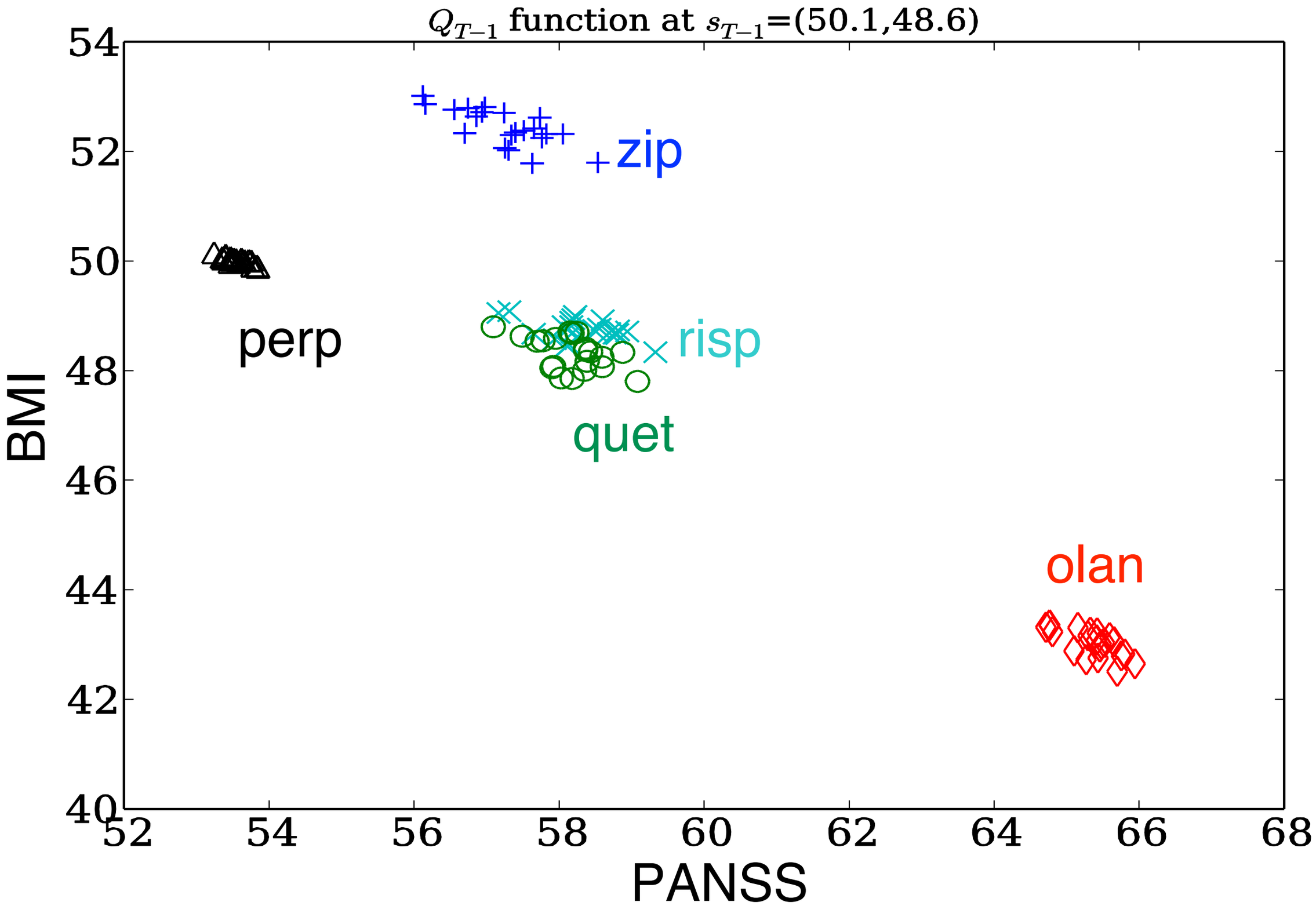

Consistent policy
Action
2
1
a1
a2
0
25
50
75
100
State


Feature-consistent policy
Action
2
1
a1
a2
0
25
50
75
100
State

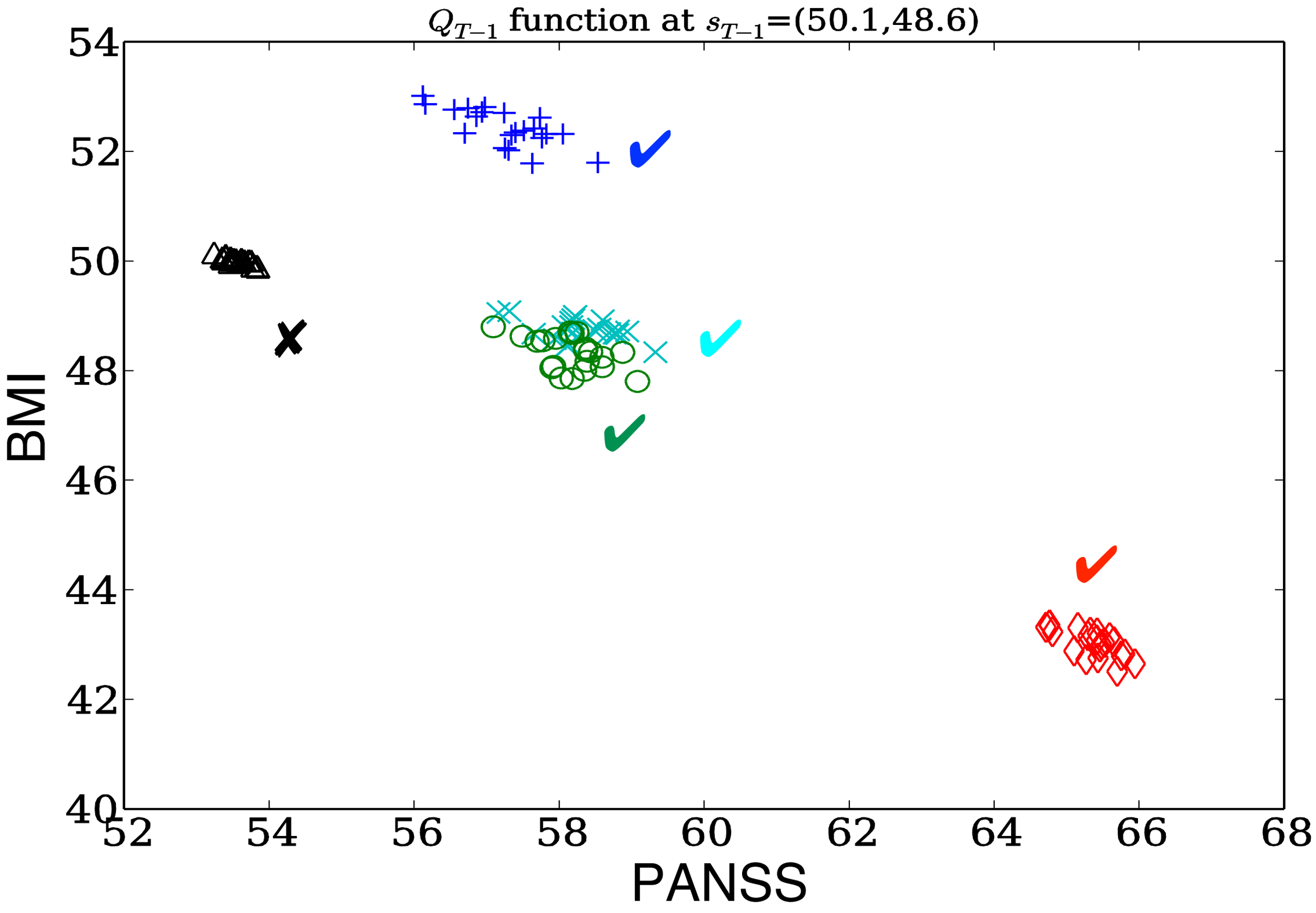
CATIE-Sz

- Clinical Antipsychotic Trials of Intervention Effectiveness
- Phase 1: Patients randomized at intake to {perphenazine, olanzapine, risperidone, quetiapine, ziprasidone}
- Phase 2:
  - If lack of efficacy, re-randomized to {{clozapine}, {olanzapine, risperidone, quetiapine}}
  - If lack of tolerability re-randomized to {olanzapine, risperidone, quetiapine, ziprasidone}
- n = 1460

Set-Valued Policies



Set-Valued Policies



MOMDPs for Decision Support (T22)

- Autonomous Agent model is for sequential decision making but we want decision support.
- Avoid assuming a pre-defined outcome by using **Set-Valued Policies**
  - Consider many possible future policies, identify non-dominated actions
- **My Goal** (what I think we need): A flexible but useful model that includes the **agent**, **environment**, and **decision maker**.

Thank you!

- Dan Lizotte dlizotte@uwo.ca
- Eric Laber laber@stat.ncsu.edu

Acknowledgements:

Natural Sciences and Engineering Research Council of Canada (NSERC)
Data were obtained from the limited access datasets distributed from the NIH-supported "Clinical Antipsychotic Trials of Intervention Effectiveness in Schizophrenia" (CATIE-Sz). The study was supported by NIMH Contract N01MH90001 to the University of North Carolina at Chapel Hill. The ClinicalTrials.gov identifier is NCT00014001. This talk reflects the view of the author and may not reflect the opinions or views of the CATIE-Sz Study Investigators or the NIH