“Inverse Preference Elicitation” for Dynamic Treatment Regimes or Considering Multiple Outcomes in Head-to-head Randomized Controlled Trials

Dan Lizotte, Michael Bowling, Susan A. Murphy

University of Michigan, University of Alberta

UNIVERSITY OF MICHIGAN

UNIVERSITY OF ALBERTA

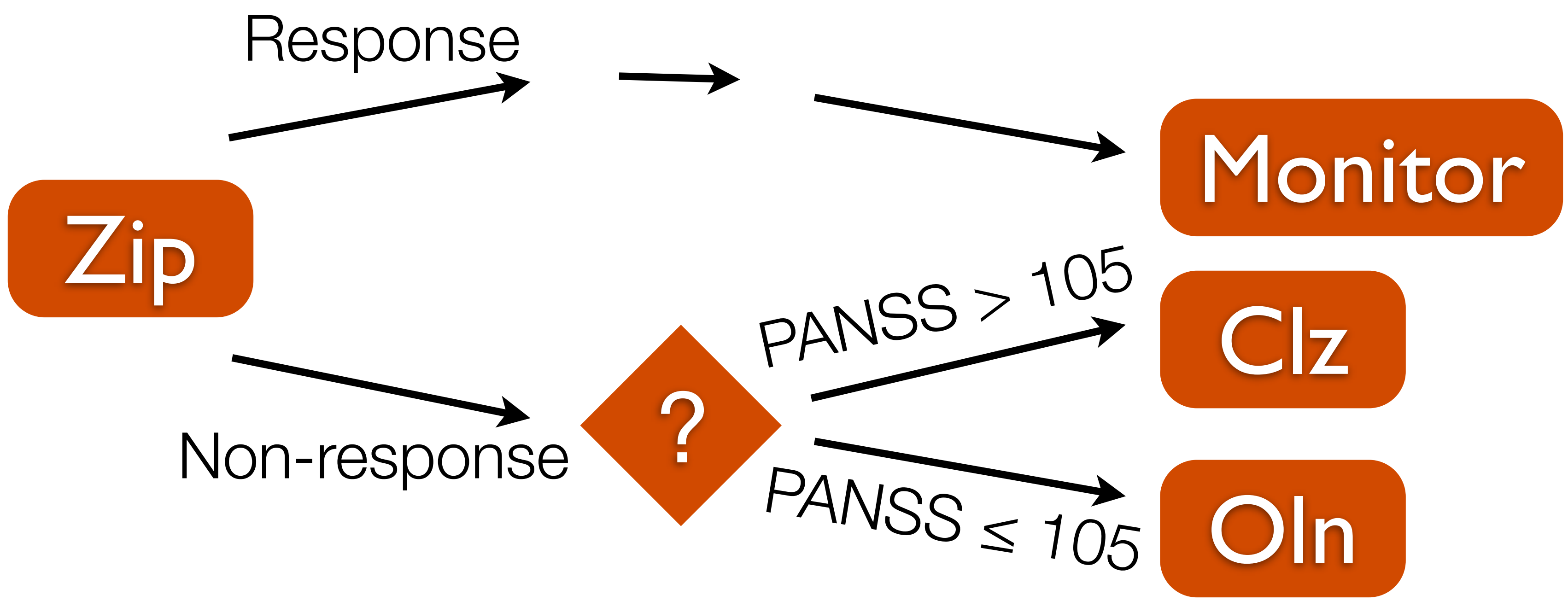
QUAECUMQUE VERA

A Talk in Two Parts

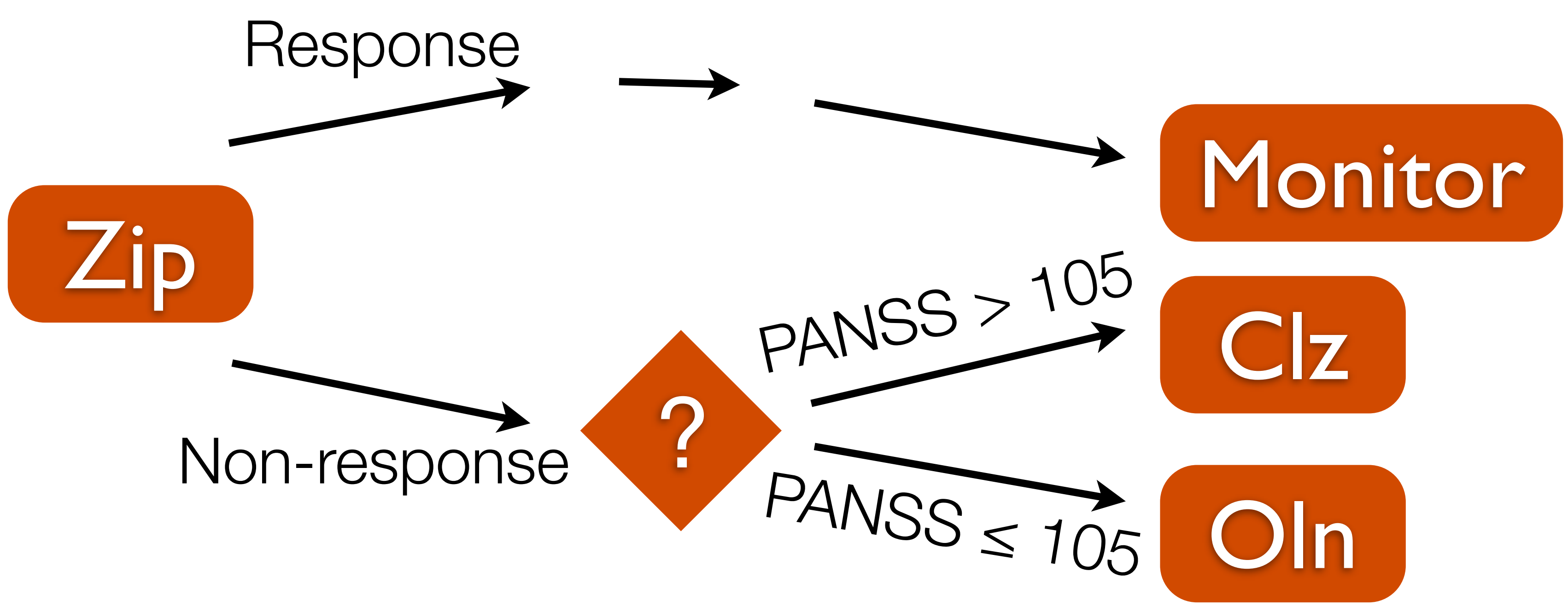
- Part I
  - DTRs, Multiple Outcomes
  - Thoughts about this idea?
- Part II
  - Overview of computational issues

Dynamic Treatment Regimes

- "...are individually tailored sequences of treatments, with treatment type and dosage adapted to the patient."
- "Dynamic"
  - Decisions are **tailored** to individual patients at the time of treatment
- "Regime"
  - A **sequence** of treatment decisions unfolding over time

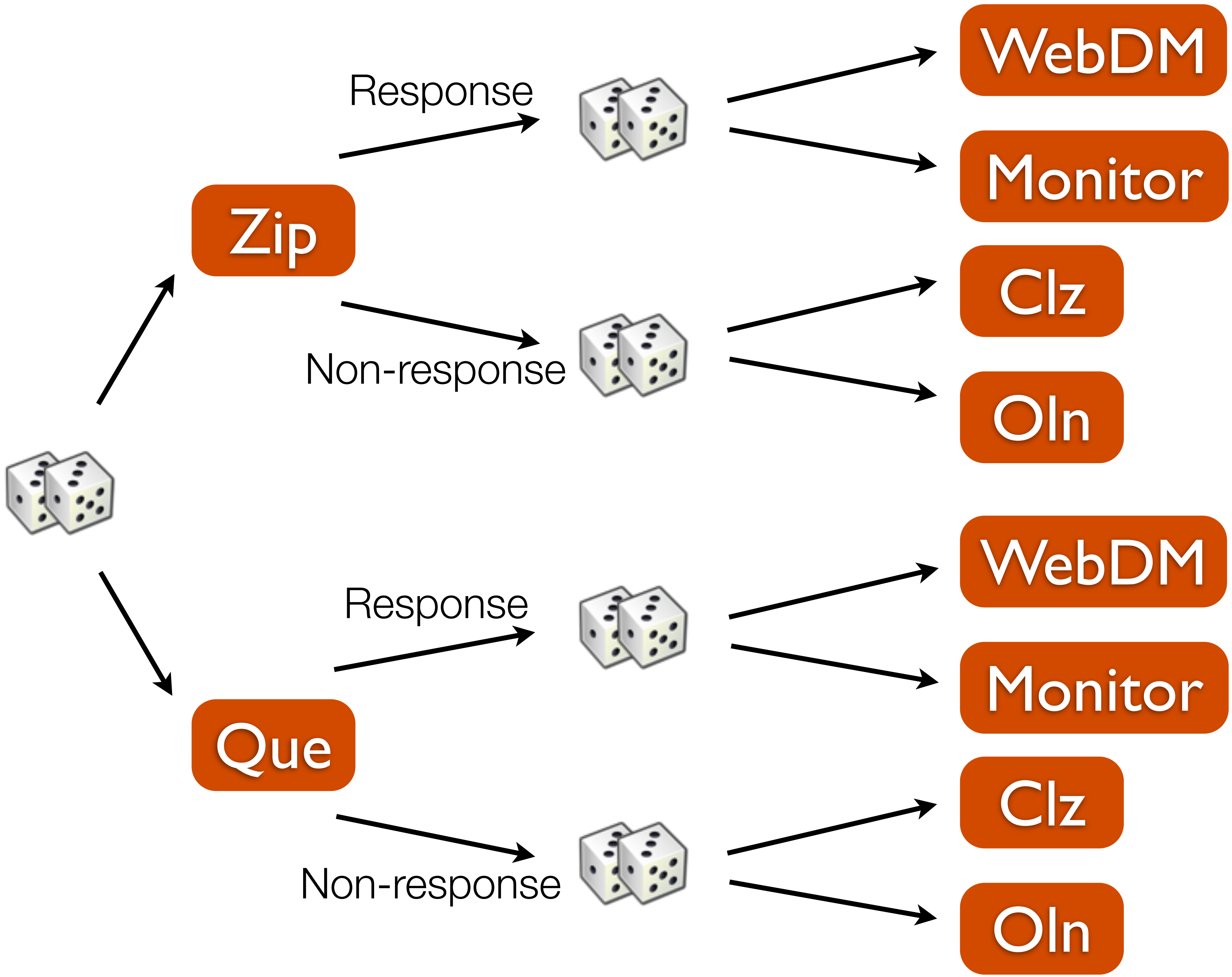


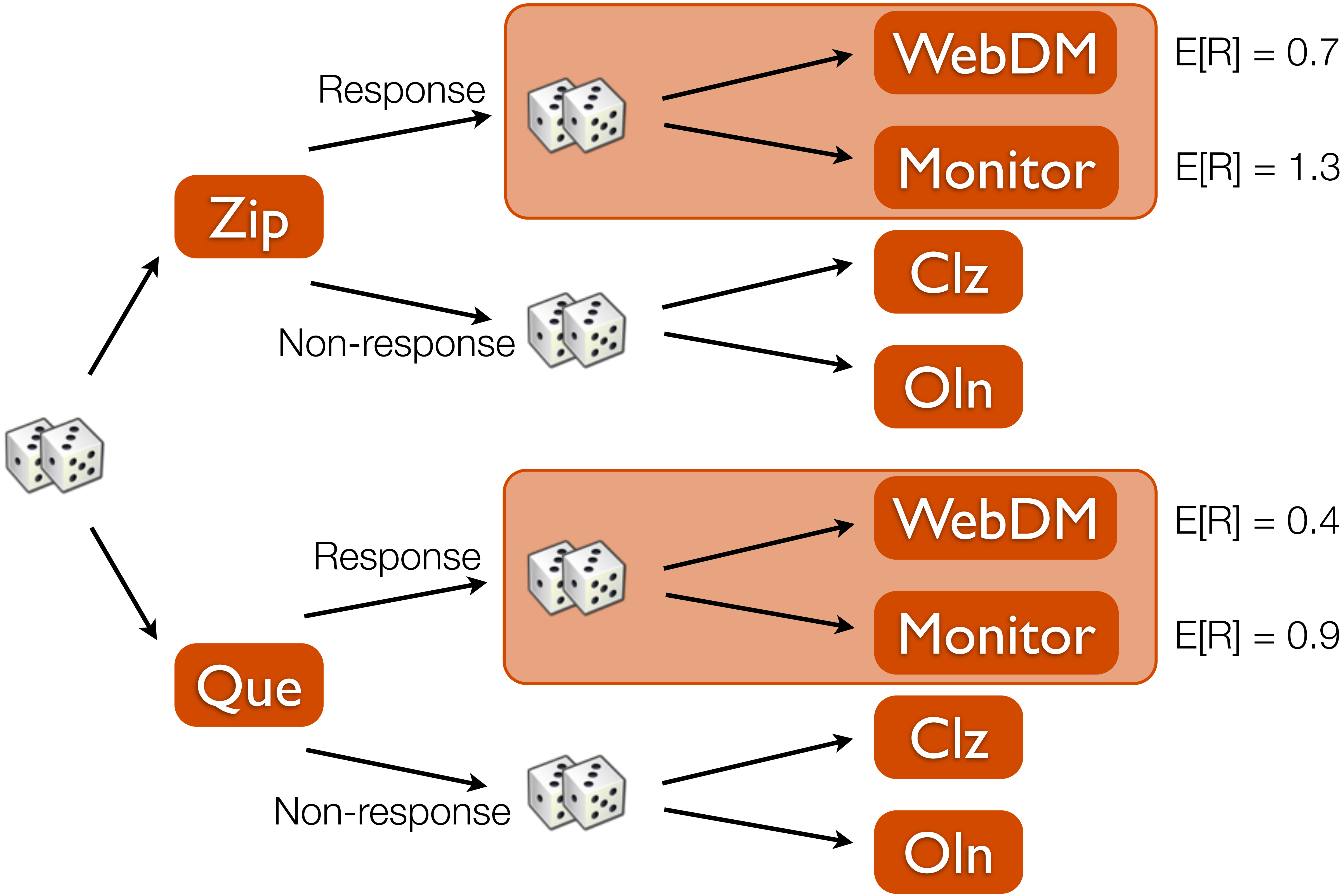
How might we arrive at this strategy?

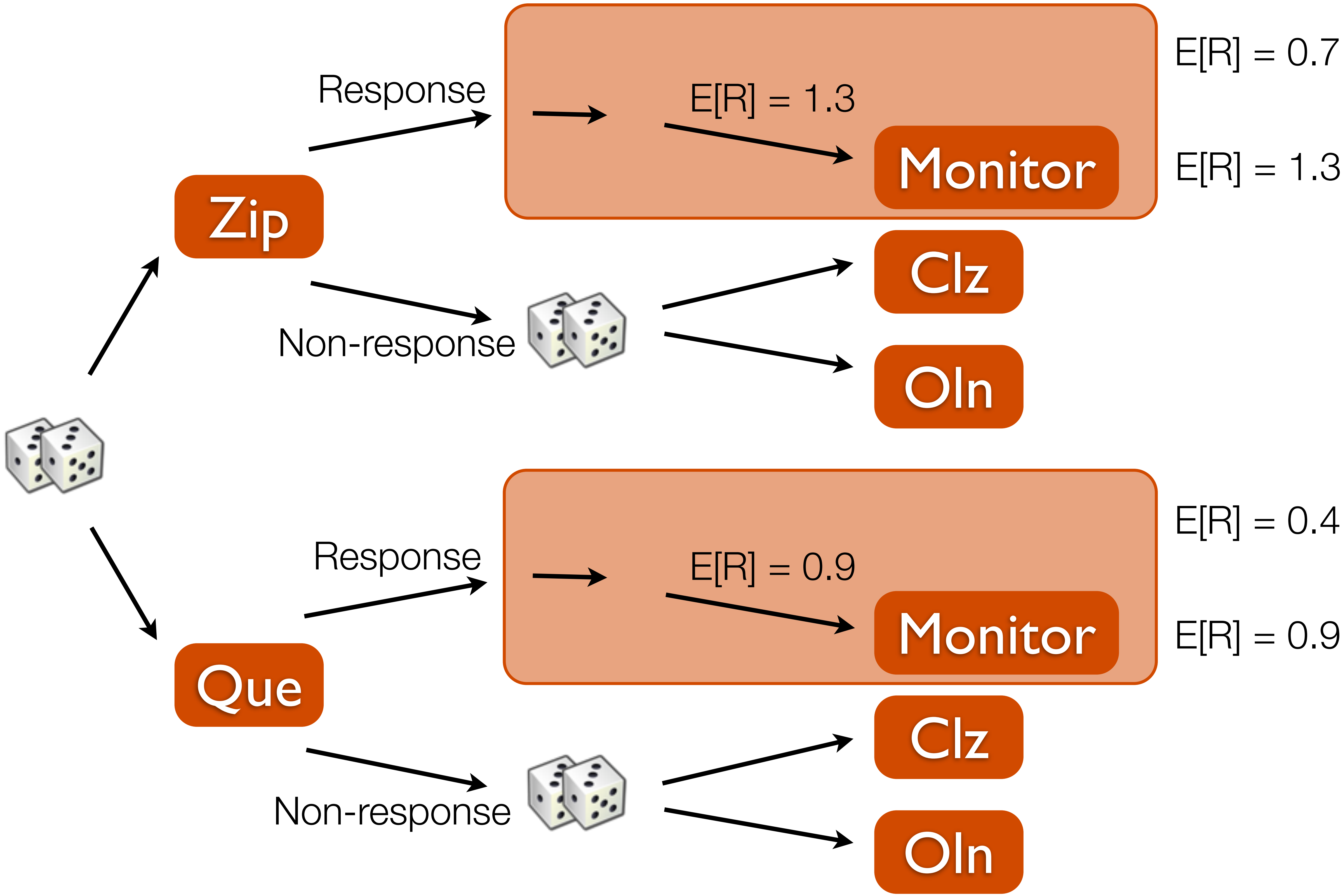


- Run a randomized trial
- Start at the end of the study
- Identify the best final treatment according to some outcome. Call it “Reward” or R. (Larger is better.)

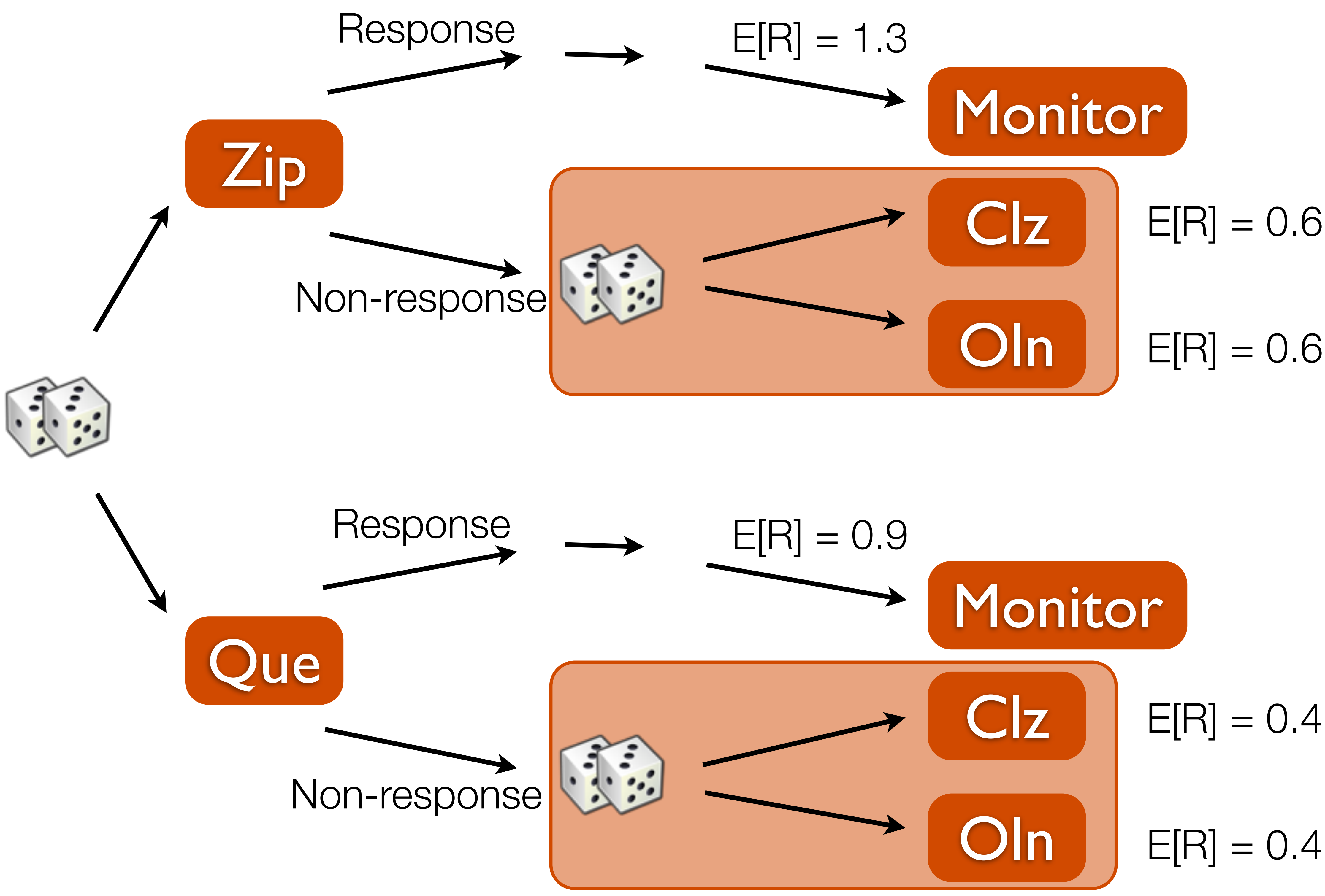
How might we arrive at this strategy?


Zip
Response
Non-response
WebDM
Monitor
Clz
Oln
Que
Response
Non-response
WebDM
Monitor
Clz
Oln

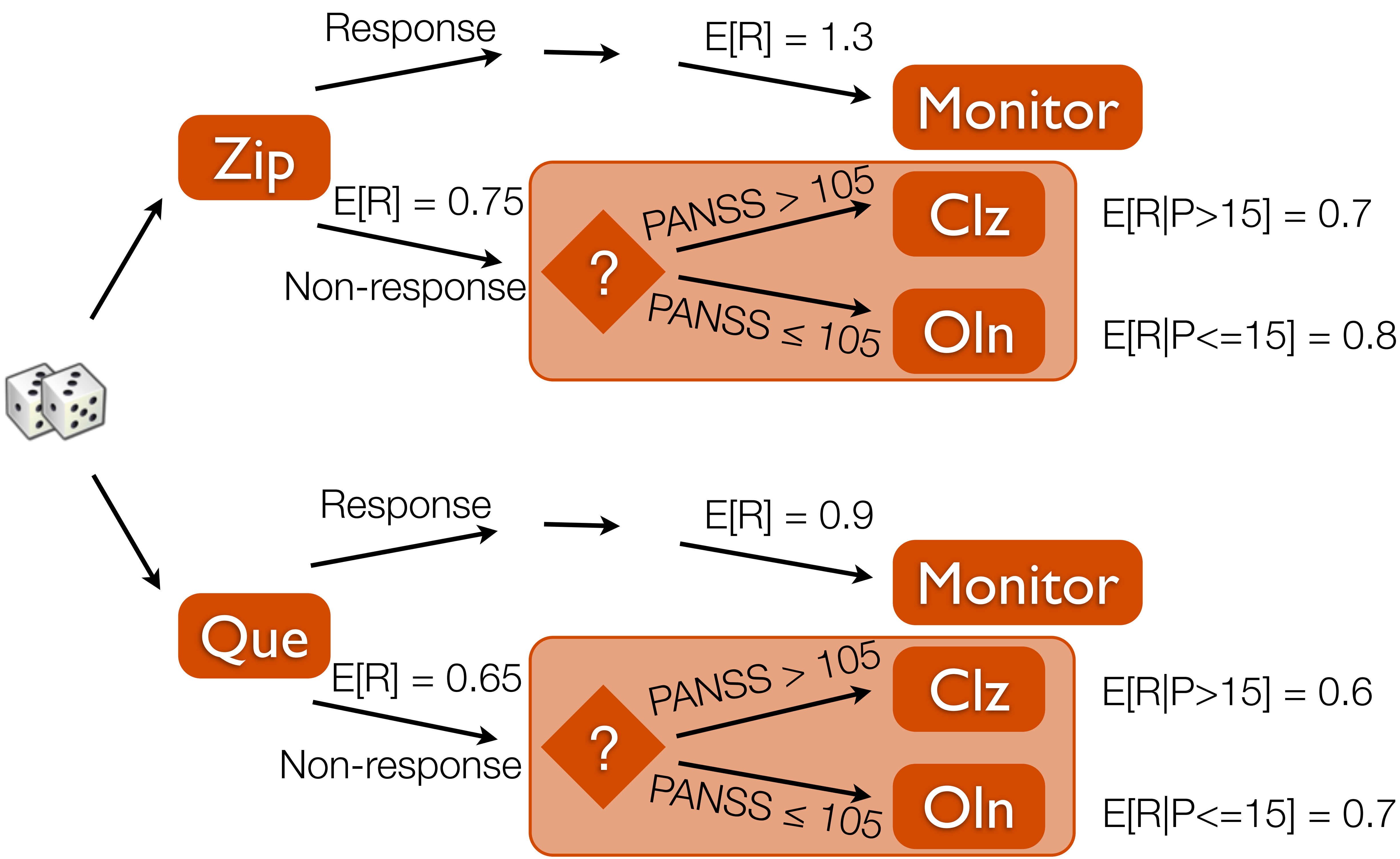

Response
WebDM
Monitor
E[R] = 0.7
E[R] = 1.3
Zip
Non-response
Clz
Oln
Response
WebDM
Monitor
E[R] = 0.4
E[R] = 0.9
Que
Non-response
Clz
Oln

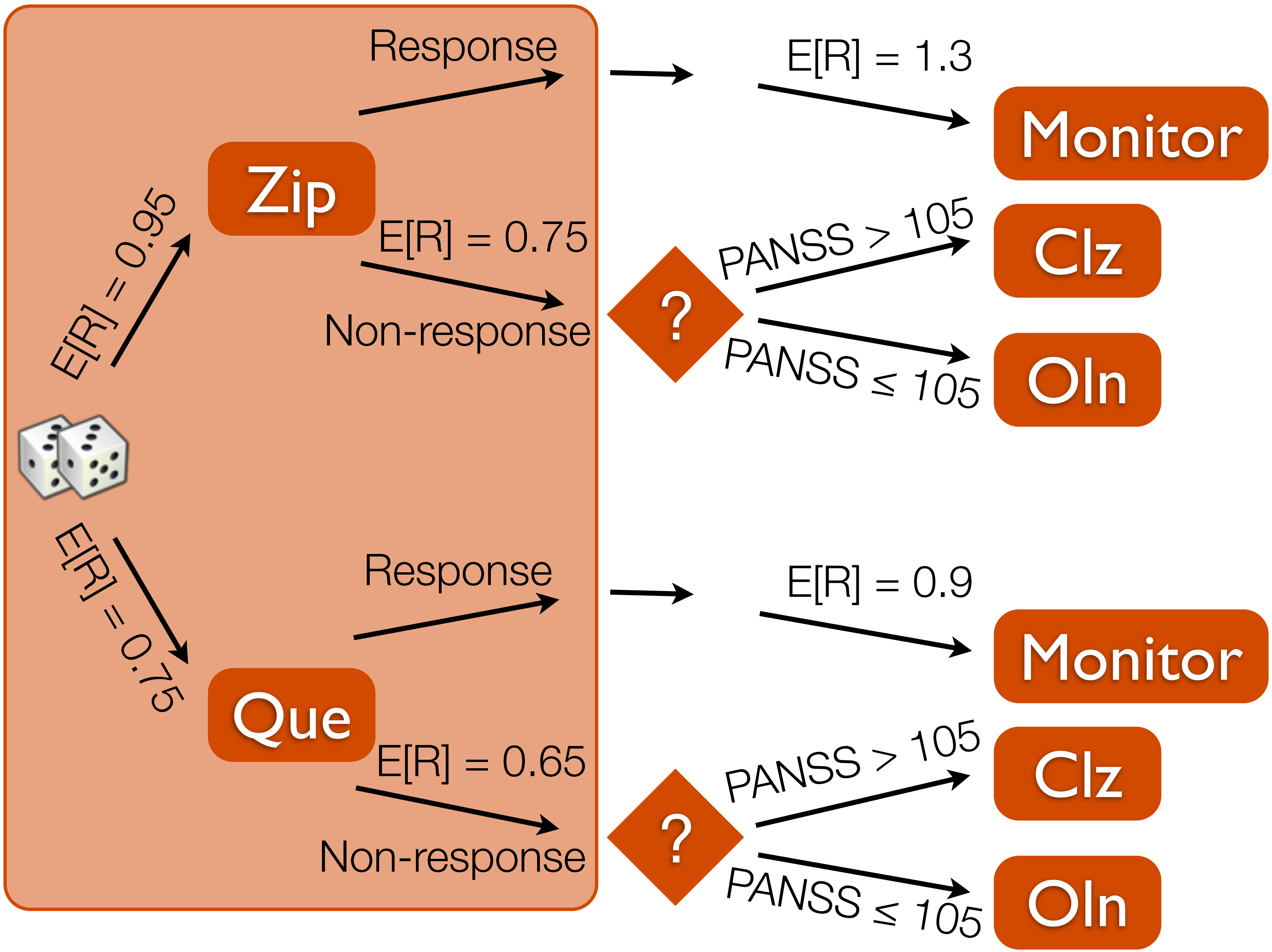


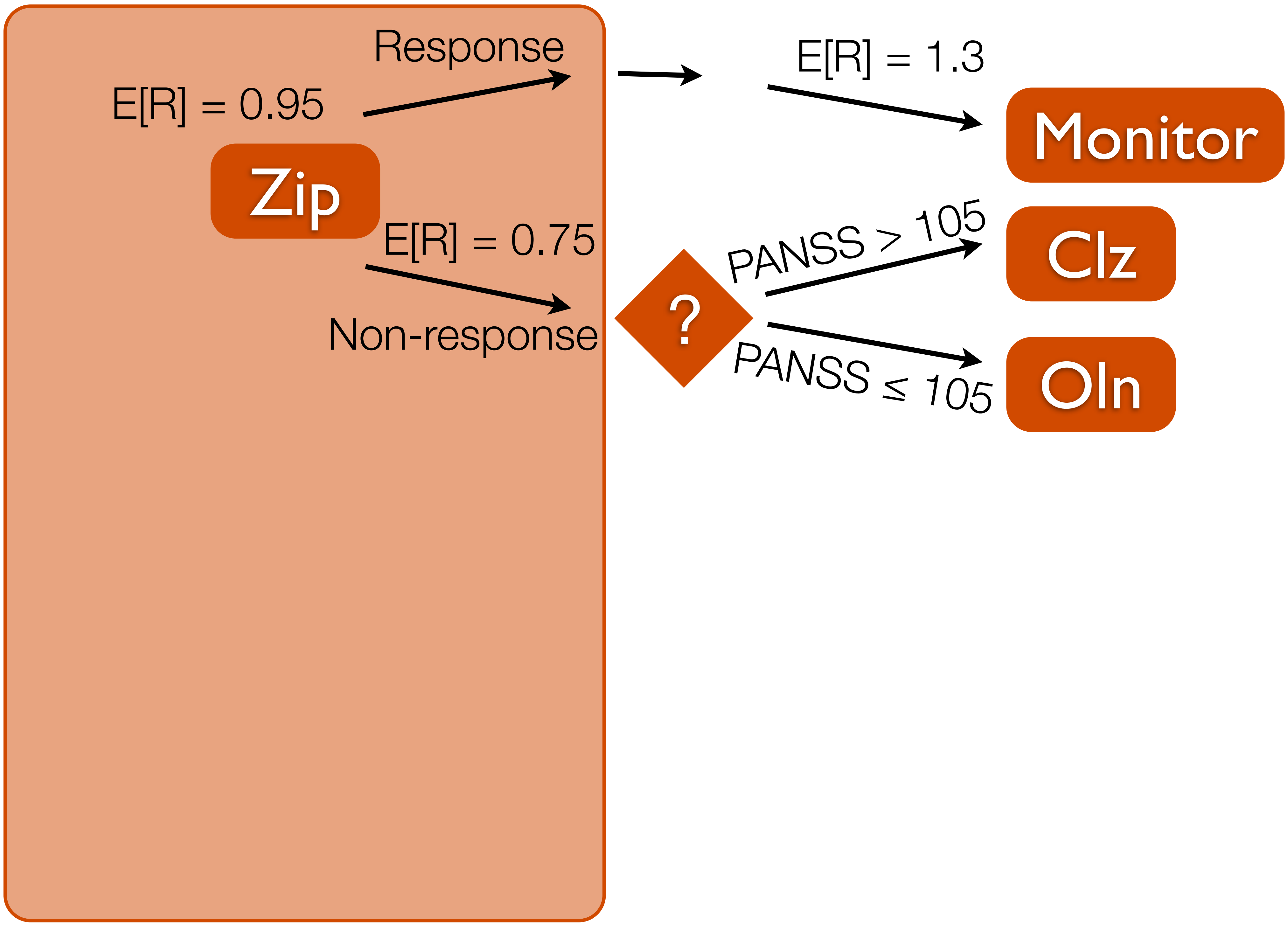
Response
E[R] = 1.3
Monitor
E[R] = 0.7
E[R] = 1.3
Zip
Clz
Non-response
Oln
Response
E[R] = 0.9
Monitor
E[R] = 0.4
E[R] = 0.9
Que
Clz
Non-response
Oln

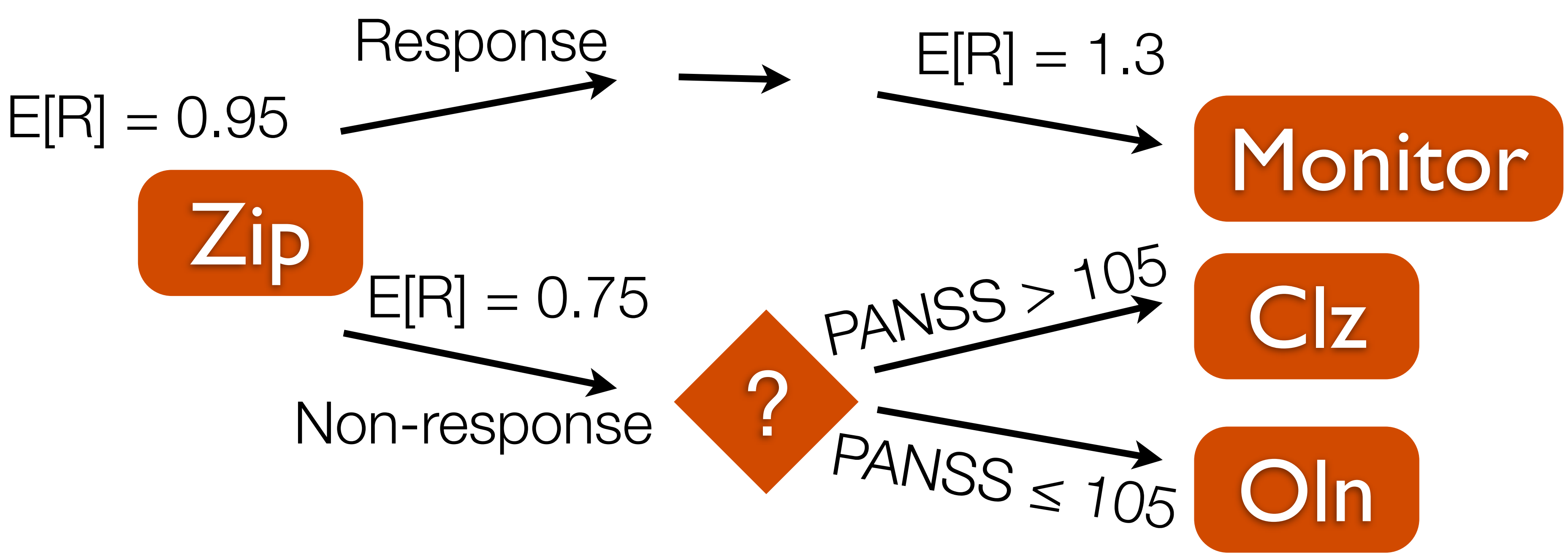


Response
E[R] = 1.3
Monitor
Zip
Clz
E[R] = 0.6
Non-response
Oln
E[R] = 0.6
Response
E[R] = 0.9
Monitor
Que
Clz
E[R] = 0.4
Non-response
Oln
E[R] = 0.4


Response
E[R] = 1.3
Monitor
Zip
E[R] = 0.75
PANSS > 105
Clz
E[R|P>15] = 0.7
?
Non-response
PANSS ≤ 105
Oln
E[R|P<=15] = 0.8
Response
E[R] = 0.9
Monitor
Que
E[R] = 0.65
PANSS > 105
Clz
E[R|P>15] = 0.6
?
Non-response
PANSS ≤ 105
Oln
E[R|P<=15] = 0.7


E[R] = 0.95
E[R] = 0.75
Zip
Que
Response
Non-response
E[R] = 0.75
Response
Non-response
E[R] = 0.65
E[R] = 1.3
E[R] = 0.9
?
?
PANSS > 105
PANSS ≤ 105
PANSS > 105
PANSS ≤ 105
Monitor
Clz
Oln
Monitor
Clz
Oln


Response
E[R] = 0.95
Zip
E[R] = 0.75
Non-response
E[R] = 1.3
Monitor
PANSS > 105
Clz
?
PANSS ≤ 105
Oln



- Run a randomized trial
- Start at the end of the study
- Identify the best final treatment according to some outcome

Goal: Medical Decision Support

- Our goal is data analysis for decision support:
  1. Take comparative effectiveness clinical trial data
  2. Produce a DTR based on patient covariates (state)
  3. Give the DTR to a clinician
- But really, a DTR is too prescriptive.
  - Our data are noisy and incomplete, causing uncertainty in the learned optimal DTR - other projects at Michigan and elsewhere.
  - We do not know what Reward to optimize.

Example: Schizophrenia

- In treatment of schizophrenia, one wants few symptoms but good functionality. This is often unachievable.

    1. This is a chronic disease. Patient state changes over time.
    2. The effect of different treatments varies from patient to patient.
    3. Different people may have very different preferences about which to give up. Each has a different reward function/objective.

- Properties 1. and 2. make the problem amenable to analysis in terms of Dynamic Treatment Regimes. The goal of this work is to deal with 3. by not *a priori* committing to a single outcome.

Multiple Rewards

- Notation: s represents patient covariates ("state"), a represents treatment.
- Consider a pair of important rewards. Suppose $r_t^{(0)}$ reflects level of symptoms and $r_t^{(1)}$ reflects level of functionality.
- Consider the set of convex combinations of these two reward functions, e.g.

$$r_t(s,a,\delta) \equiv (1 - \delta) \cdot r_t^{(0)}(s,a) + \delta \cdot r_t^{(1)}(s,a)$$

Multiple Rewards

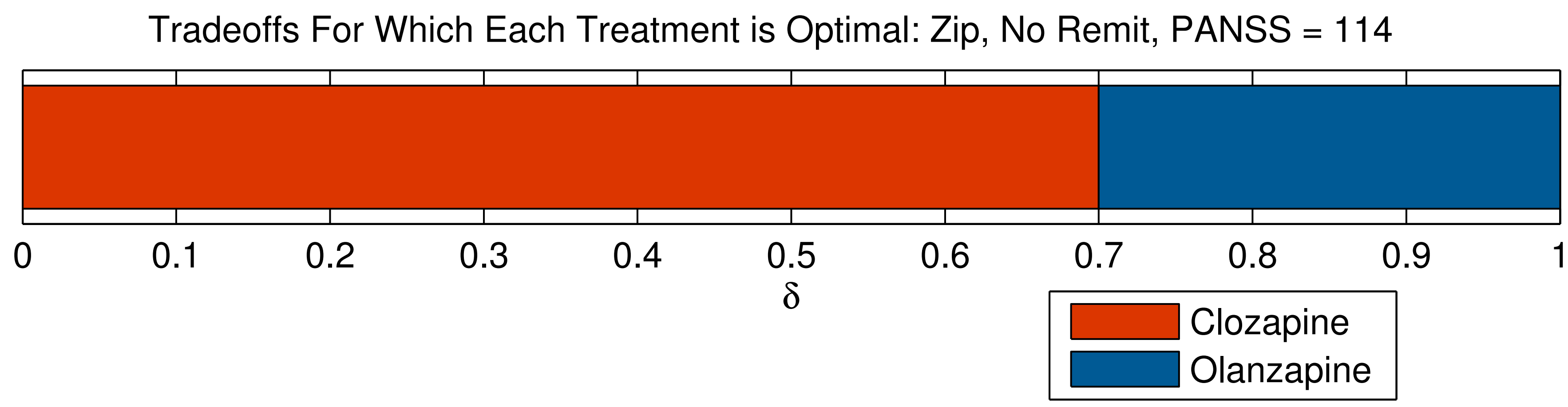
$$r_t(s,a,\delta) \equiv (1 - \delta) \cdot r_t^{(0)}(s,a) + \delta \cdot r_t^{(1)}(s,a)$$

- Each δ identifies a specific reward function, and induces a corresponding estimated optimal DTR. Depending on δ, the optimal DTR "cares more" about $r_t^{(0)}$ or $r_t^{(1)}$.
- δ determines the "exchange rate" between $r_t^{(0)}$ and $r_t^{(1)}$
- Closest "standard" approach: "Preference Elicitation"
  - Try to determine the decision-maker's *true* value of δ via time tradeoff, standard gamble, visual analog scale,...
- Given s and δ for a patient, the resulting DTR selects a treatment.

“Inverse”
Preference Elicitation

- We propose a different approach
- Take $r(s,a,\delta) \equiv (1 - \delta) \cdot r^{(0)}(s,a) + \delta \cdot r^{(1)}(s,a)$
- Run analysis to find optimal actions *given all* δ
- Given a new patient’s state, e.g. “Got Zip, didn’t remit, PANSS = 114” report, for each treatment, the range of δ for which it is optimal.
    - For some treatments this range might be empty.

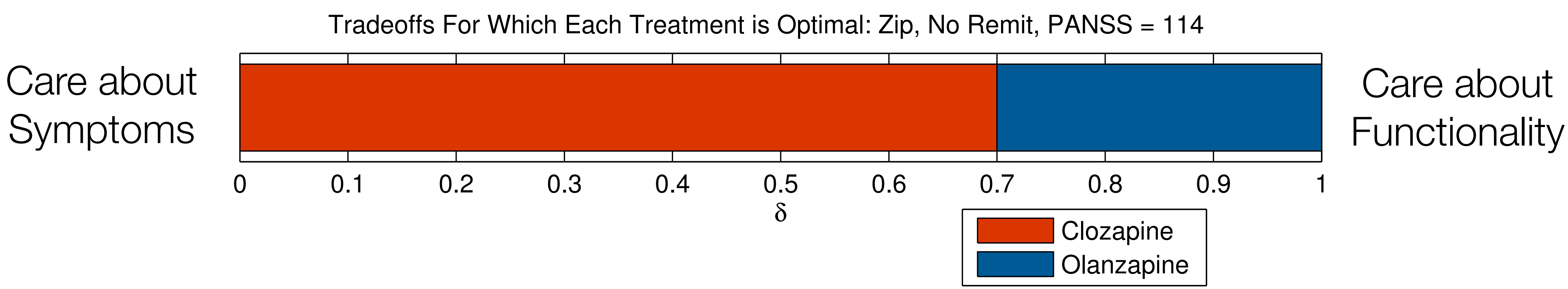
Care about Symptoms



Care about Functionality

“Inverse”
Preference Elicitation

- “Choosing Clozapine indicates willingness to give up **at most** 2.3 units of symptoms in order to gain 1 unit of functionality.”
- “Choosing Olanzapine indicates willingness to give up **at most** .42 units of functionality to gain 1 unit of symptoms.”
- “Choosing Clozapine indicates willingness to give up **at least** .42 units of functionality in order to gain 1 unit of symptoms.”
- “Choosing Olanzapine indicates willingness to give up **at least** 2.3 units of symptoms in order to gain 1 unit of functionality.”



End of Part I

- Part I
  - DTRs, Multiple Outcomes
  - Thoughts about this idea?
- Part II
  - Overview of computational issues

Begin Part II

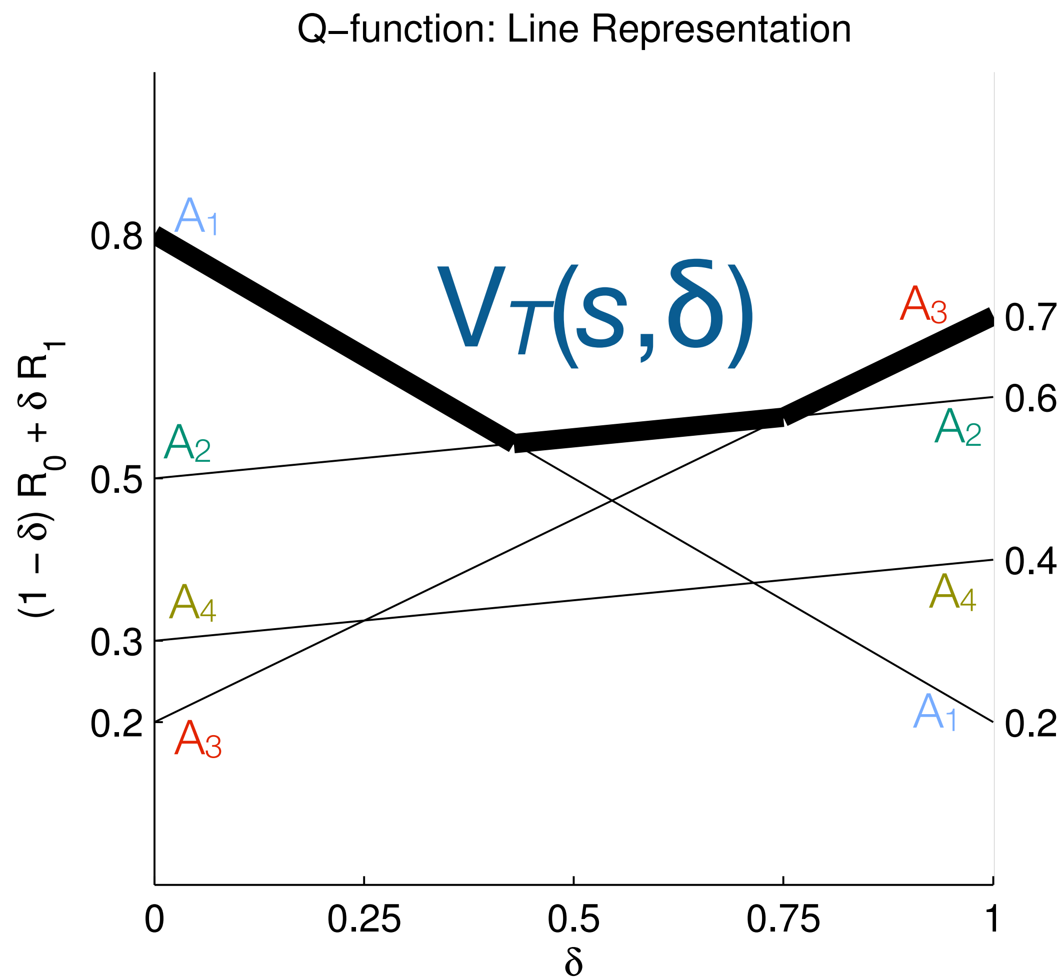
- Part I
  - DTRs, Multiple Outcomes
  - Thoughts about this idea?
- Part II
  - Overview of computational issues

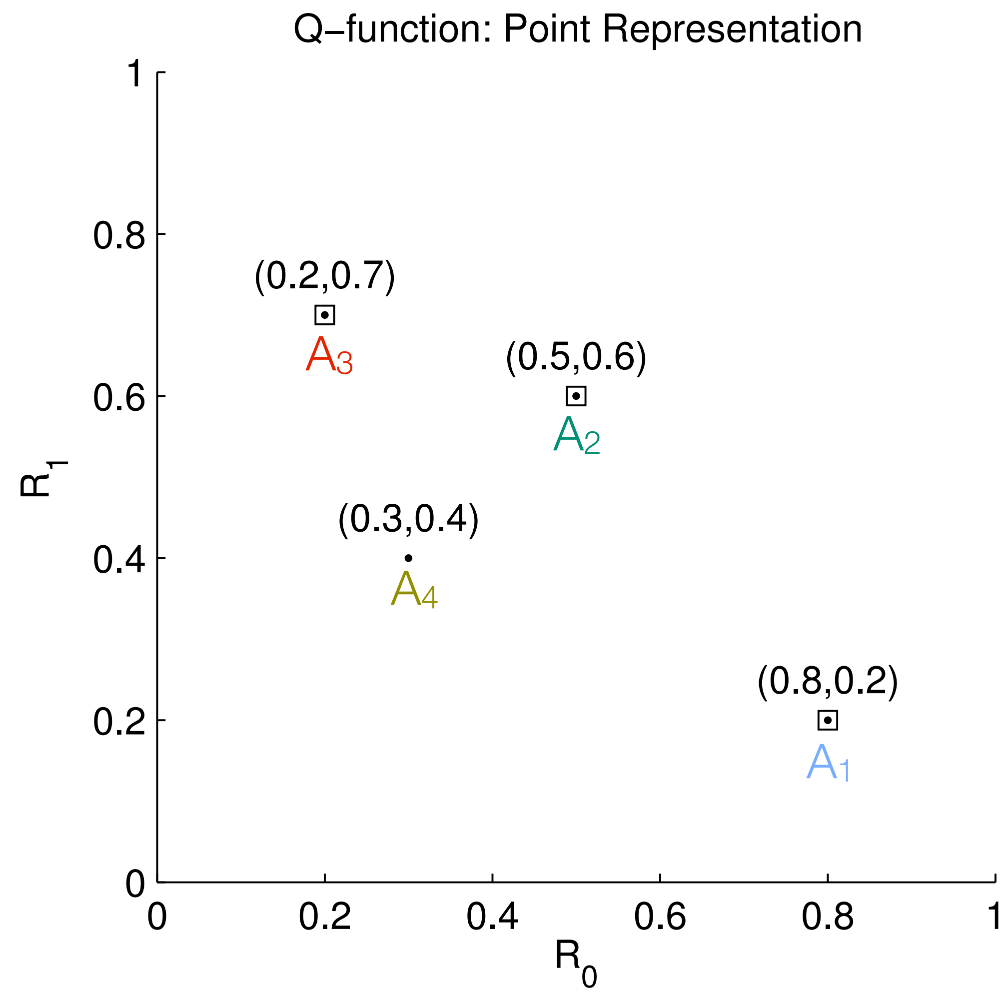
Algorithm for Discrete Covariates

- $r_T(s,a,\delta) \equiv (1 - \delta) \cdot r_T^{(0)}(s,a) + \delta \cdot r_T^{(1)}(s,a)$
- $Q_t(s,a,\delta)$ optimal expected future reward for given s, a, δ
- $V_t(s,\delta) = \max_a Q_t(s,a,\delta)$ optimal expected future reward for s, δ
  - Assumes optimal choice of a now and in future
- Find $Q_{t-1}(s,a,\delta)$ recursively
  - $Q_{t-1}(s,a,\delta) = \mathrm{E}[R_t + V_t(S',\delta)] = \mathrm{E}[R_t + \max_{a'} Q_t(S',a',\delta)]$

Value Backup: $V_t(s,\delta) = \max_a Q_t(s,a,\delta)$

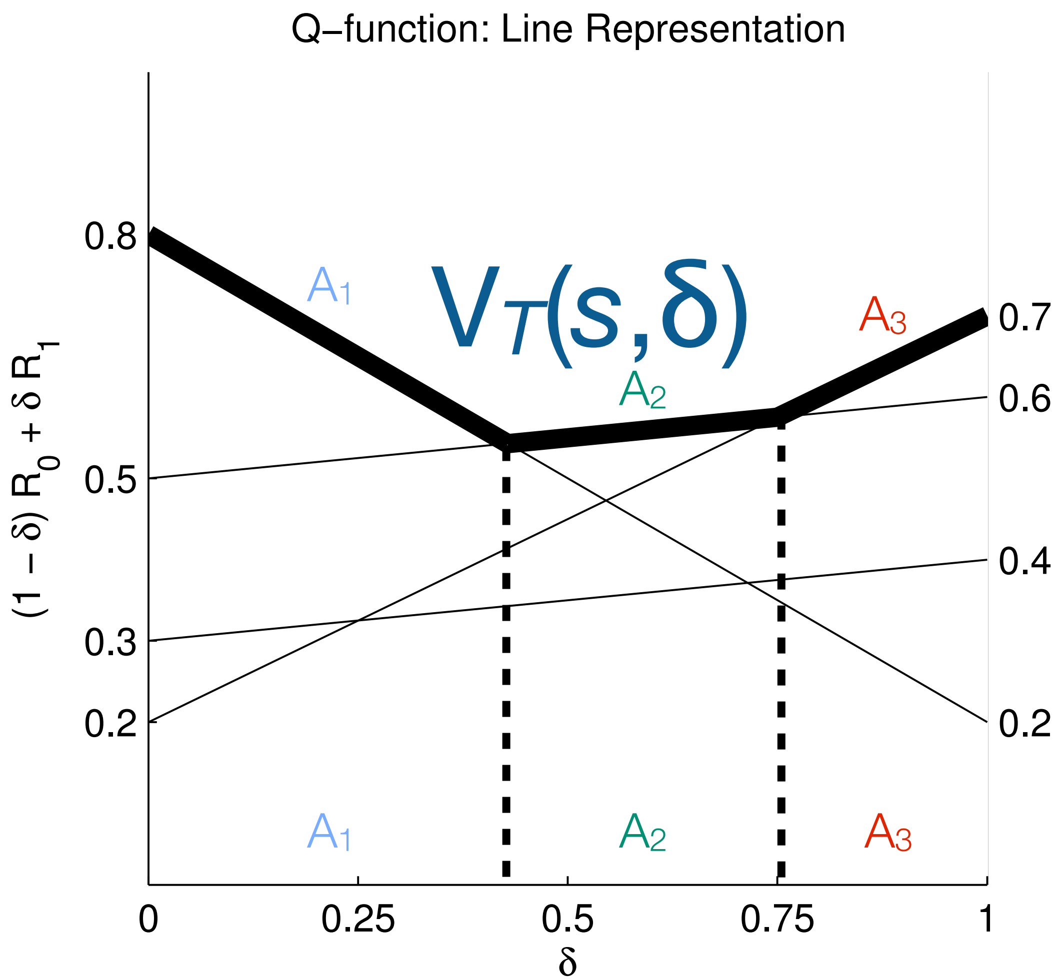
- $Q_T(s,a,0)$, $Q_T(s,a,1)$ are average rewards, $Q_T(s,a,\delta)$ is linear in δ
- $V_T(s,\delta)$ is continuous and piecewise linear in δ
    - Knots introduced by pointwise max over *a* found by convex hull

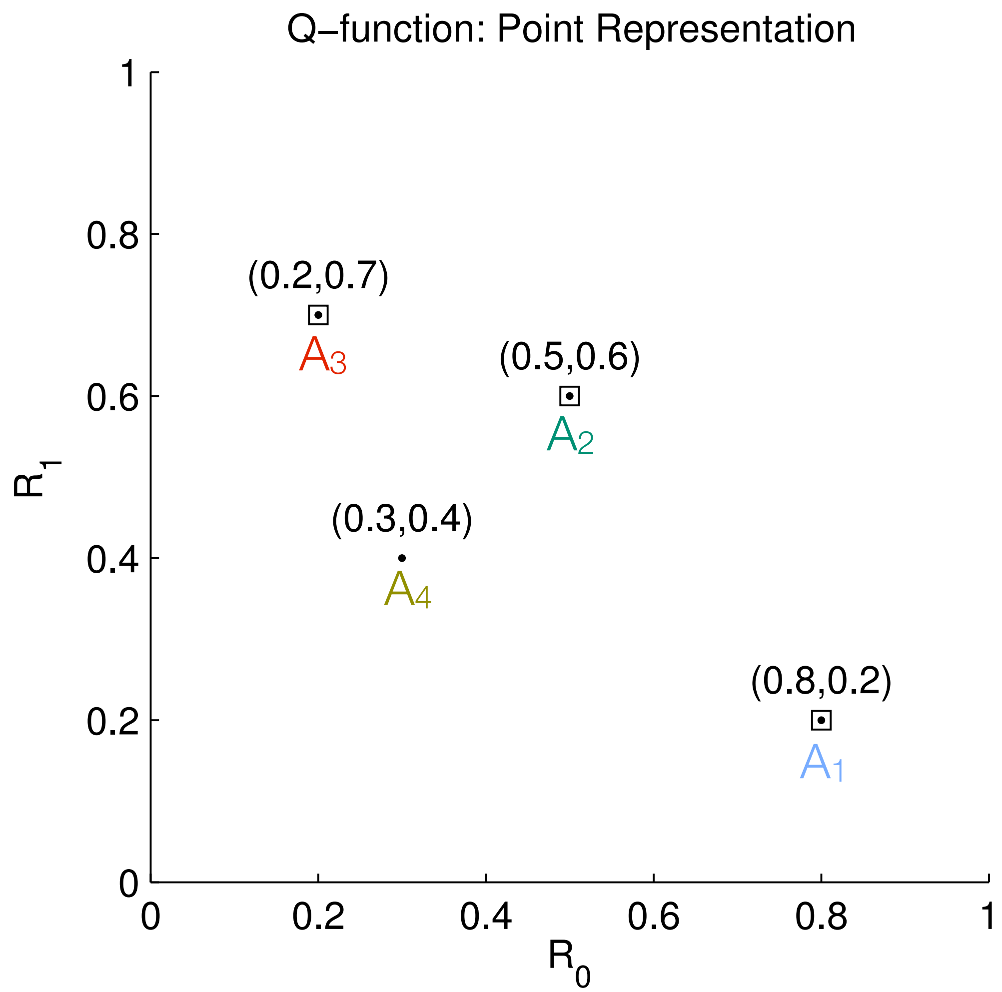




Value Backup: $V_t(s,\delta) = \max_a Q_t(s,a,\delta)$

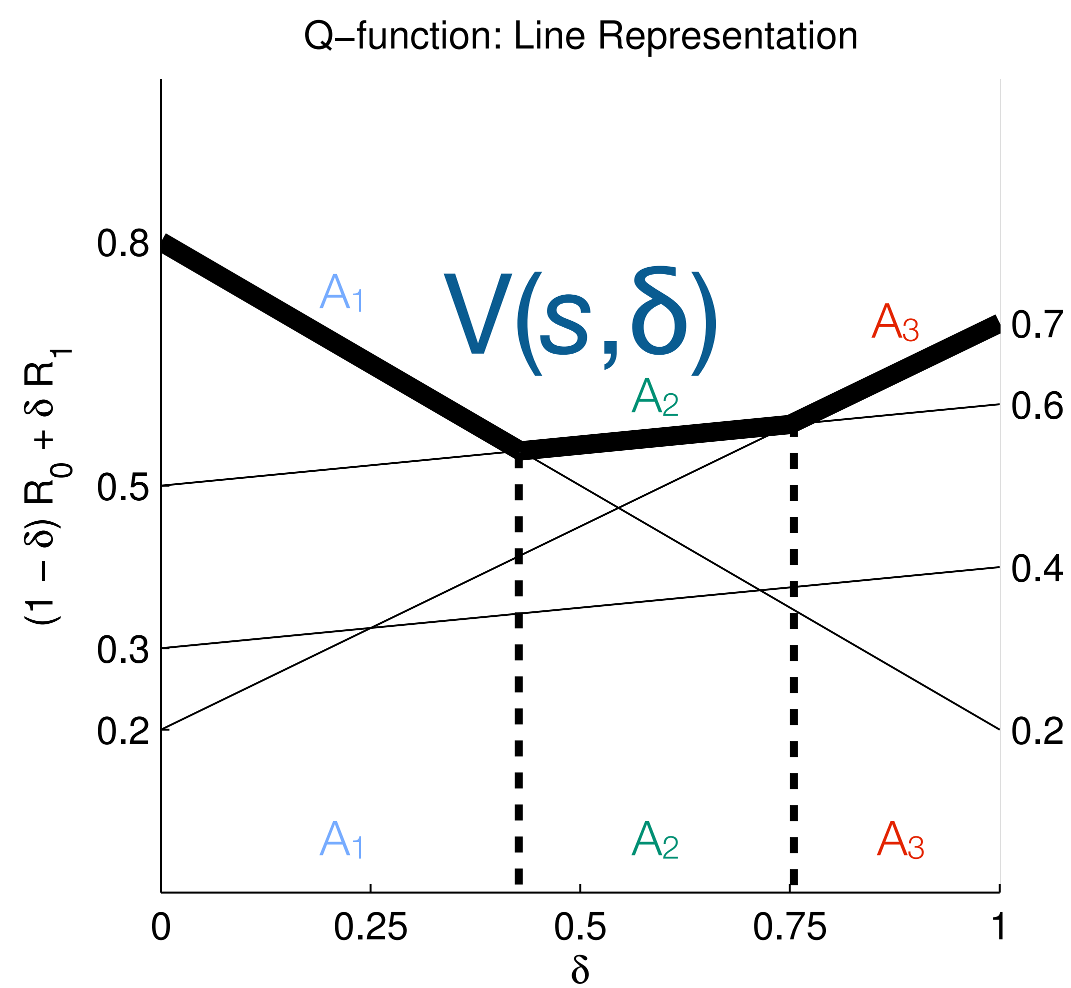
- Our value function representation “remembers” which actions are optimal over which intervals of delta

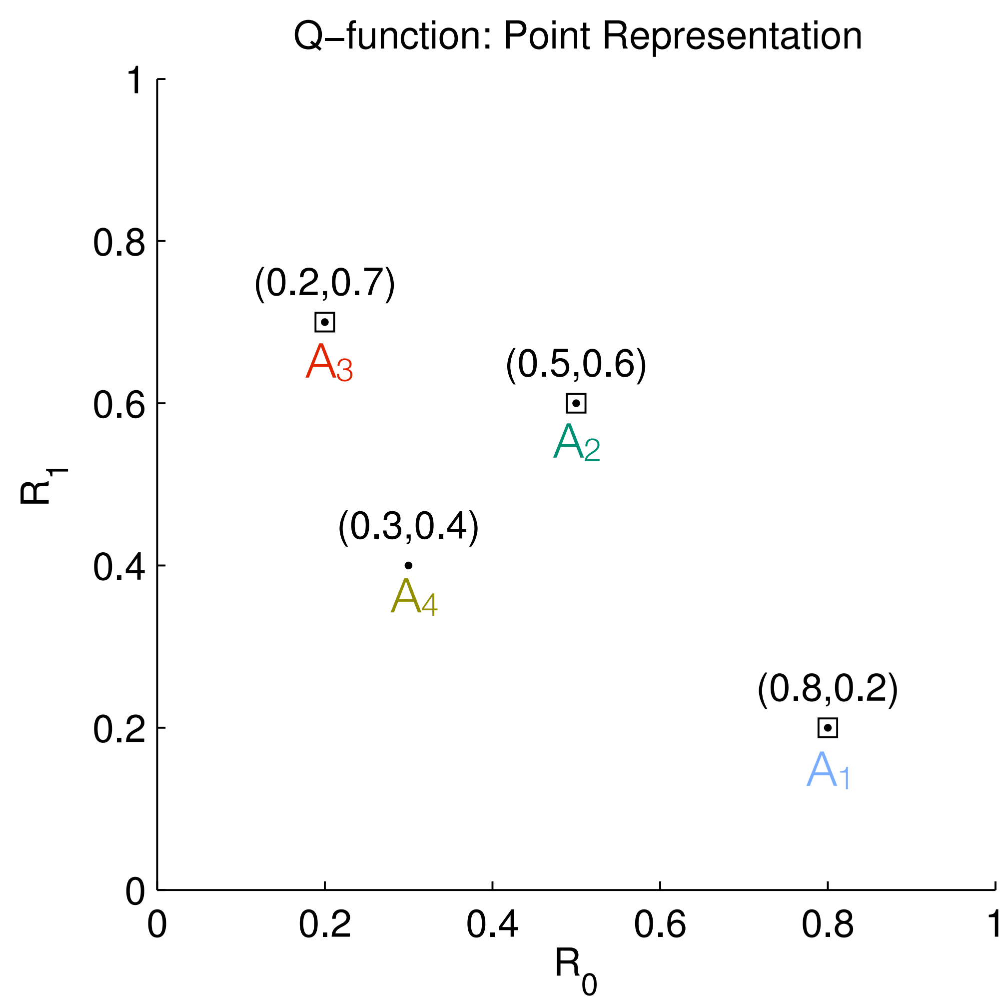




Dominated Actions

- Some actions are not optimal for any δ

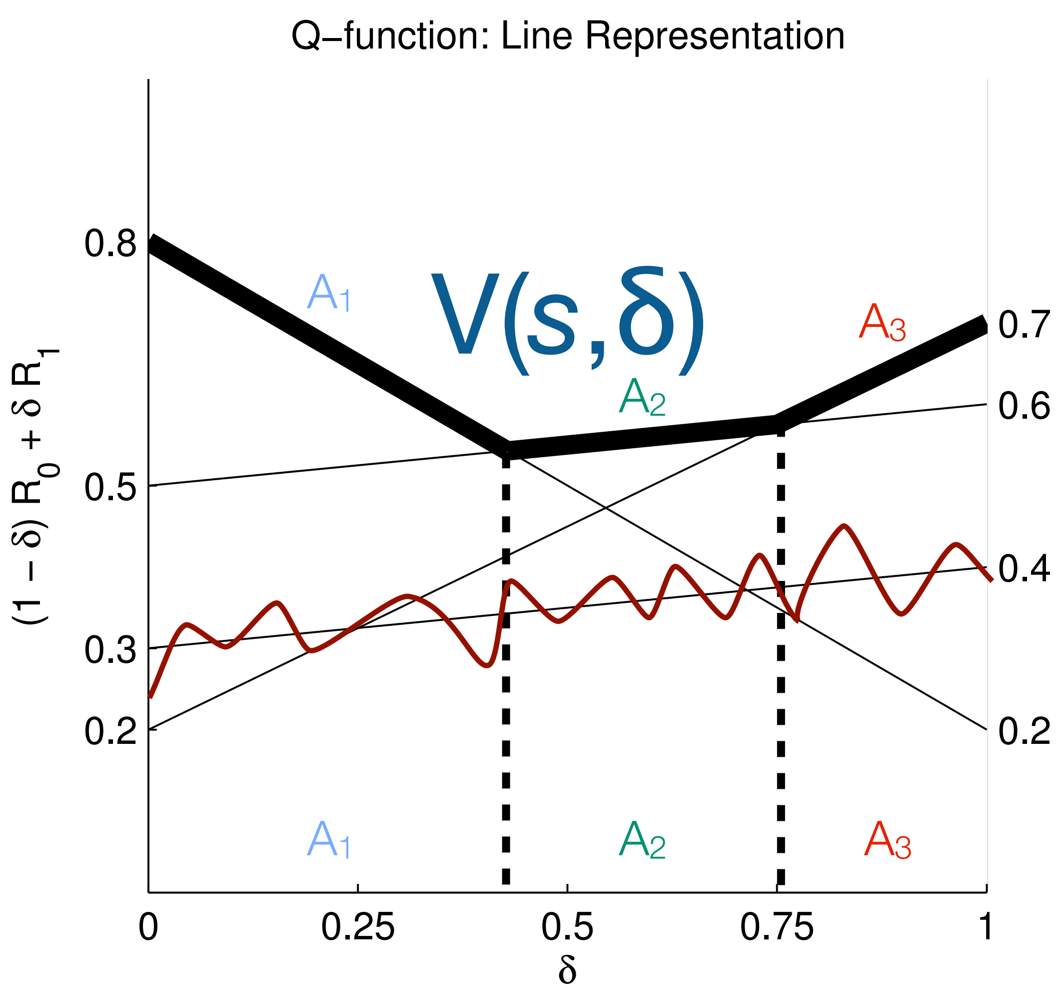


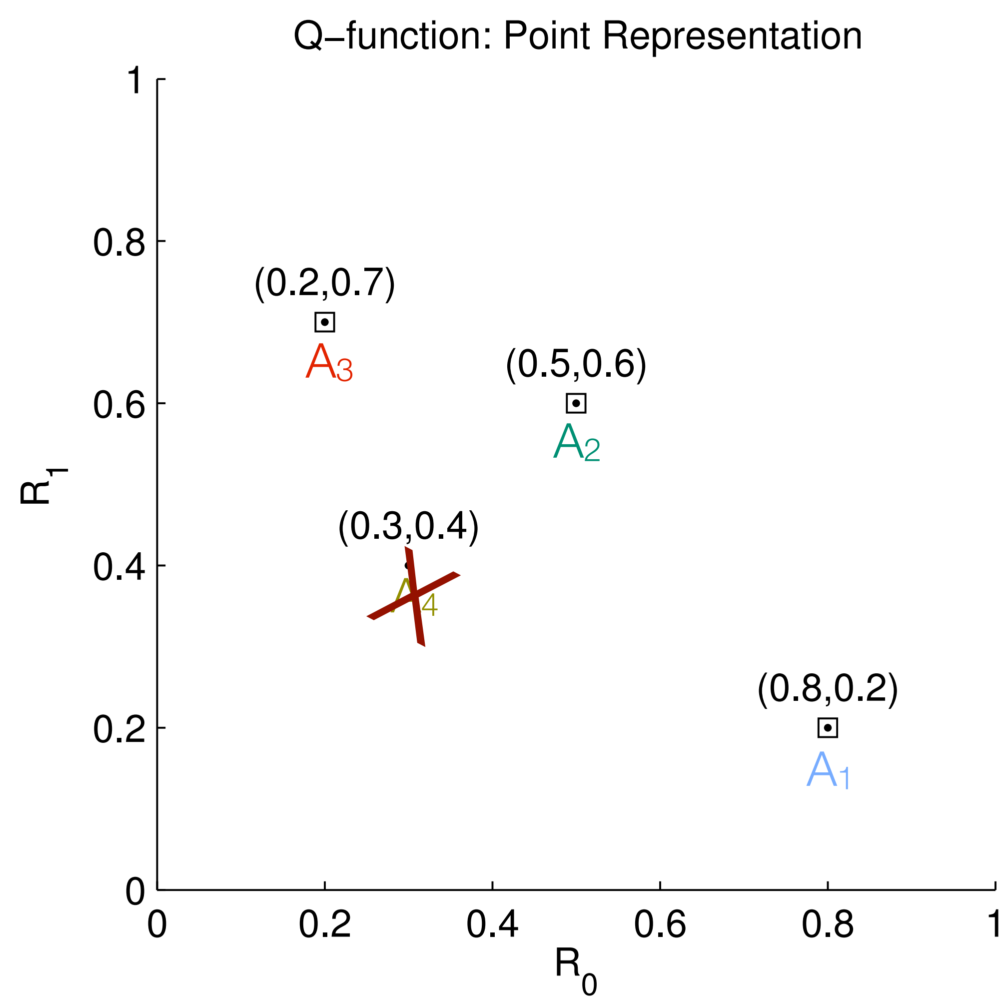


- Some actions are not optimal for any (δ,s)! Can enumerate s to check this.

Dominated Actions

- Some actions are not optimal for any δ

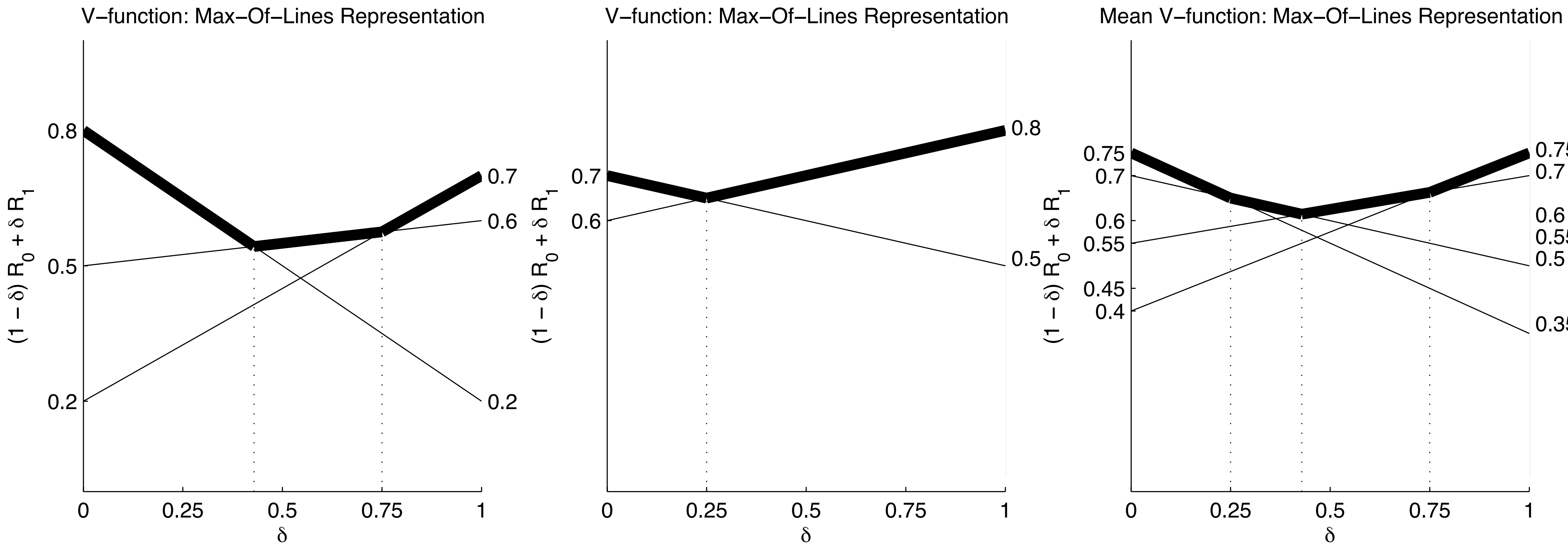




- Some actions are not optimal for any (δ, s)!
  Can enumerate s to check this.

Value Backup: $Q_{t-1}(s,a,\delta) = E[R_t + V_t(s',\delta)]$

- $Q_{T-1}(s,a,\delta)$ is continuous and piecewise linear in δ
  - Pointwise average of $V_T(s',\delta)$



$V_T(s_1,\delta)$ $V_T(s_2,\delta)$ $Q_{T-1}(s,a,\delta) = E_{S'|s,a}[R_{T-1}+V_T(S',\delta)]$

Continuous State Space, Linear Regression

- Recall: $r_T(s,a,\delta) \equiv (1 - \delta) \cdot r_T^{(0)}(s,a) + \delta \cdot r_T^{(1)}(s,a)$
- Construct design matrices S_a ($n_a \times p$), targets $\boldsymbol{r}_a(\delta)$ ($n_a \times 1$) from our data set
- $Q_T(s,a,\delta;\beta) = \beta_a(\delta)^T s$, $\beta_a(\delta) = (S_a^T S_a)^{-1} S_a^T \boldsymbol{r}_a(\delta)$
  - $Q_T(s,a,\delta;\beta)$ linear in β, each element of β is linear in r, and r is linear in δ
- Discrete states:
  - For each s, $Q_{T-1}(s,a,\delta)$ for each s is piecewise linear in δ
- Continuous states:
  - Each regression coefficient of $Q_{T-1}(s,a,\delta;\beta)$ is piecewise linear in δ

Reality Check

- Must compute $\beta_a(\delta)$ at knot δs between linear regions. At time T-t, there could be $O(n^{T-t}|A|^{T-t})$ knots, in the worst case.
- Is this even feasible? Consider 1000 randomly generated datasets, n = 1290, |A| = 3, T = 3, parameters similar to real data
- Maximum time for 1 simulation run is 6.55 seconds on 8 procs.

	Worst-case #knots	Observed Min	Observed Med	Observed Max
t=2	3870	687	790	910
t=1	$1.5 \cdot 10^7$	2814	3160	3916

Future Work - Computing Science, Statistics

- Allow more state variables
  - For backups: Easy! Each element of β is piecewise linear in δ
  - When checking for dominated actions, 2 reward functions plus 2 state covariate is feasible. (Or 3 reward functions + 1 state covariate.)
- Allow more reward functions
  - For backups: 3 reward functions is feasible.
    Representing non-convex continuous piecewise linear functions in high dimensions appears difficult.
- Approximations, now that we know what we are approximating.
- **Measures of uncertainty for preference ranges**

Future Work - Clinical Science

1. Schizophrenia
   - Symptom reduction versus functionality, or weight gain
2. Major Depressive Disorder
   - Symptom reduction versus weight gain, other side-effects
3. Diabetes
   - Disease complications versus drug side-effects

Thanks!

- Supported by National Institute of Health grants R01 MH080015 and P50 DA10075

- Questions?

- Related work:

  Daniel J. Lizotte, Michael Bowling, and Susan A. Murphy. Efficient Reinforcement Learning with Multiple Reward Functions for Randomized Clinical Trial Analysis. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning (ICML)*, 2010.

  Barrett, L. and Narayanan, S. Learning all optimal policies with multiple criteria. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

Preference elicitation for QALYs, [Wikipedia version]

- Time-trade-off (TTO): Respondents are asked to choose between remaining in a state of ill health for a period of time, or being restored to perfect health but having a shorter life expectancy.

- Standard gamble (SG): Respondents are asked to choose between remaining in a state of ill health for a period of time, or choosing a medical intervention which has a chance of either restoring them to perfect health, or killing them.

- Visual analogue scale (VAS): Respondents are asked to rate a state of ill health on a scale from 0 to 100, with 0 representing death and 100 representing perfect health. This method has the advantage of being the easiest to ask, but is the most subjective.