mHealth Adventures in Sequential Experimentation & Reinforcement Learning!

JOOL HEALTH

BariFit

Susan A Murphy

07.07.18

Sense²Stop

MD2K Center of Excellence for Mobile Sensor Data-to-Knowledge

Statistics Harvard

The Methodology Center advancing methods, improving health

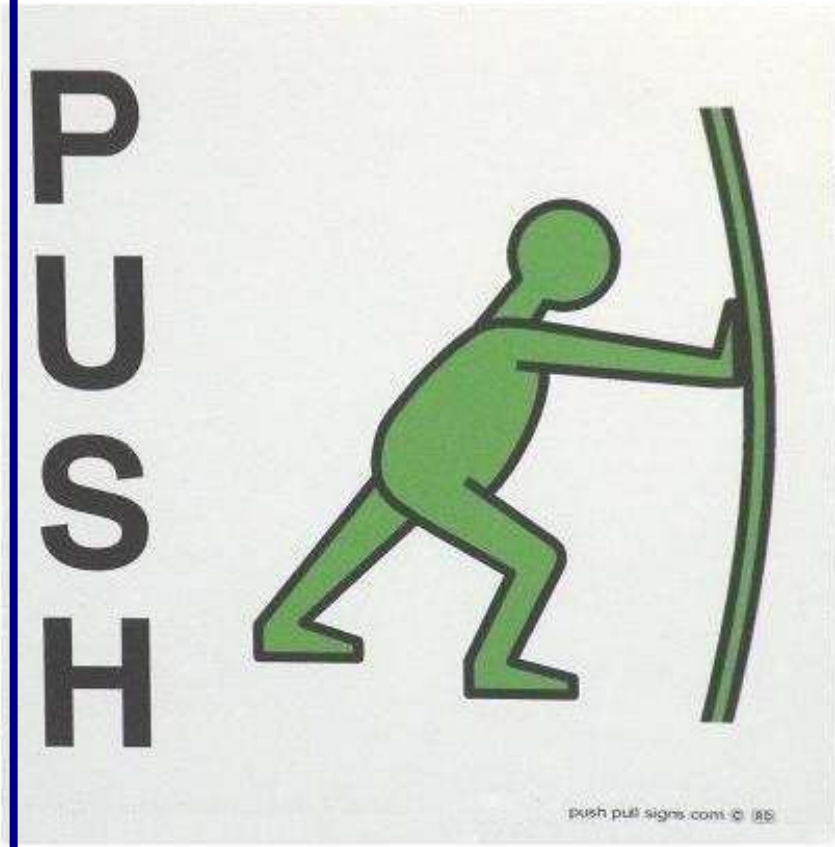
Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

mHealth Science Goals

- Promote behavior change and maintenance of this change
  - Assist user in achieving long term goals
    - Recovery from addictions; avoid heart attacks; maintain independence
  - Manage chronic illness
- Test, evaluate, develop causal behavioral science

mHealth Treatment Actions



Outline

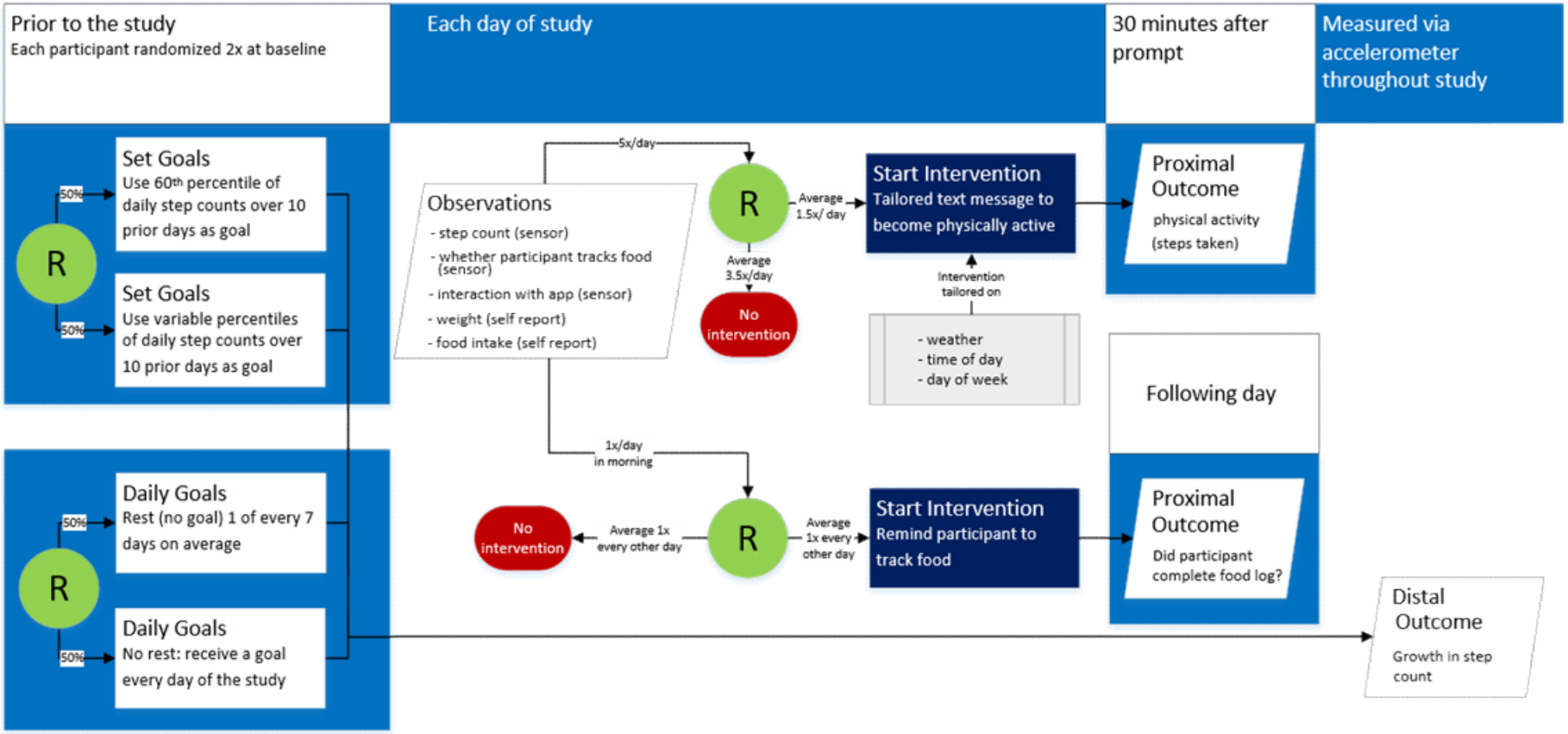
- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

Micro-randomized Trial

- Micro-randomized trial = combination of factorial experimental design with explicitly controlled exploration
- Exploration via use of online forecasting (and RL algorithms)
- Multiple treatment factors occurring at different time scales and which target different rewards
- Probabilistic budgets on # of treatment pushes to manage habituation/burden
- Off-policy, after study is over, causal inference

BariFit

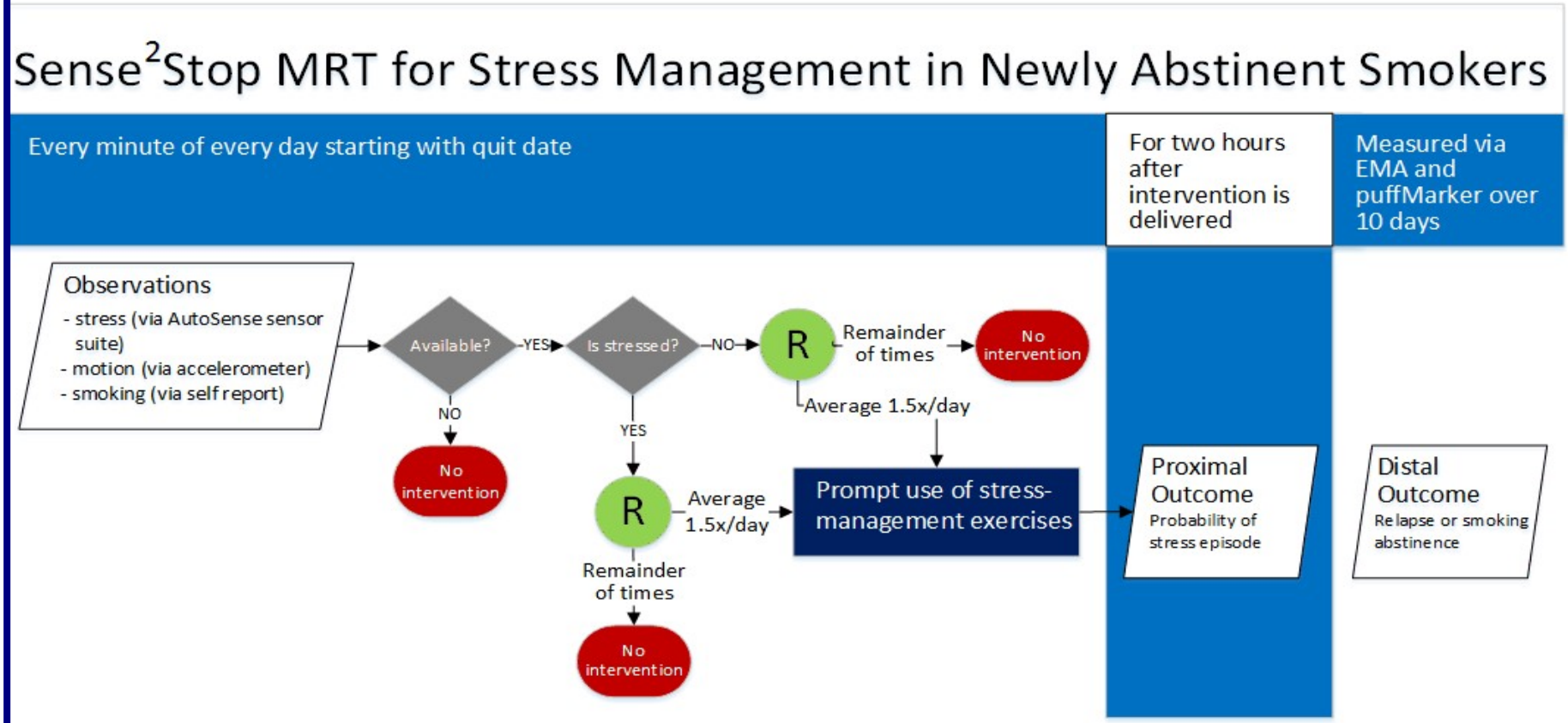
BariFit MRT to Promote Weight Maintenance Among People Who Received Bariatric Surgery



PI: P Klasnja

Location & Funding: Kaiser Permanente

Sense^{2}Stop



PI: S Kumar

Location: Northwestern University, B. Spring, (P.I.)

Funding: NIBIB through funds provided by the trans-NIH Big Data to Knowledge initiative U54EB020404

Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

Challenges to RL

- State is large yet partially observed: "unknown-unknowns"
  - Non-stationary reward
- Treatment actions that tend to have positive effects on immediate rewards but negative impact on future rewards via user habituation/burden.
- High noise within/between user
- Clinical populations (e.g., small numbers of users)
- Off-policy causal inference to further develop behavioral science

Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

HeartSteps (PI Klasnja)

Goal: Develop an mobile activity coach for individuals who are at high risk of adverse cardiac events.

Three iterative studies:

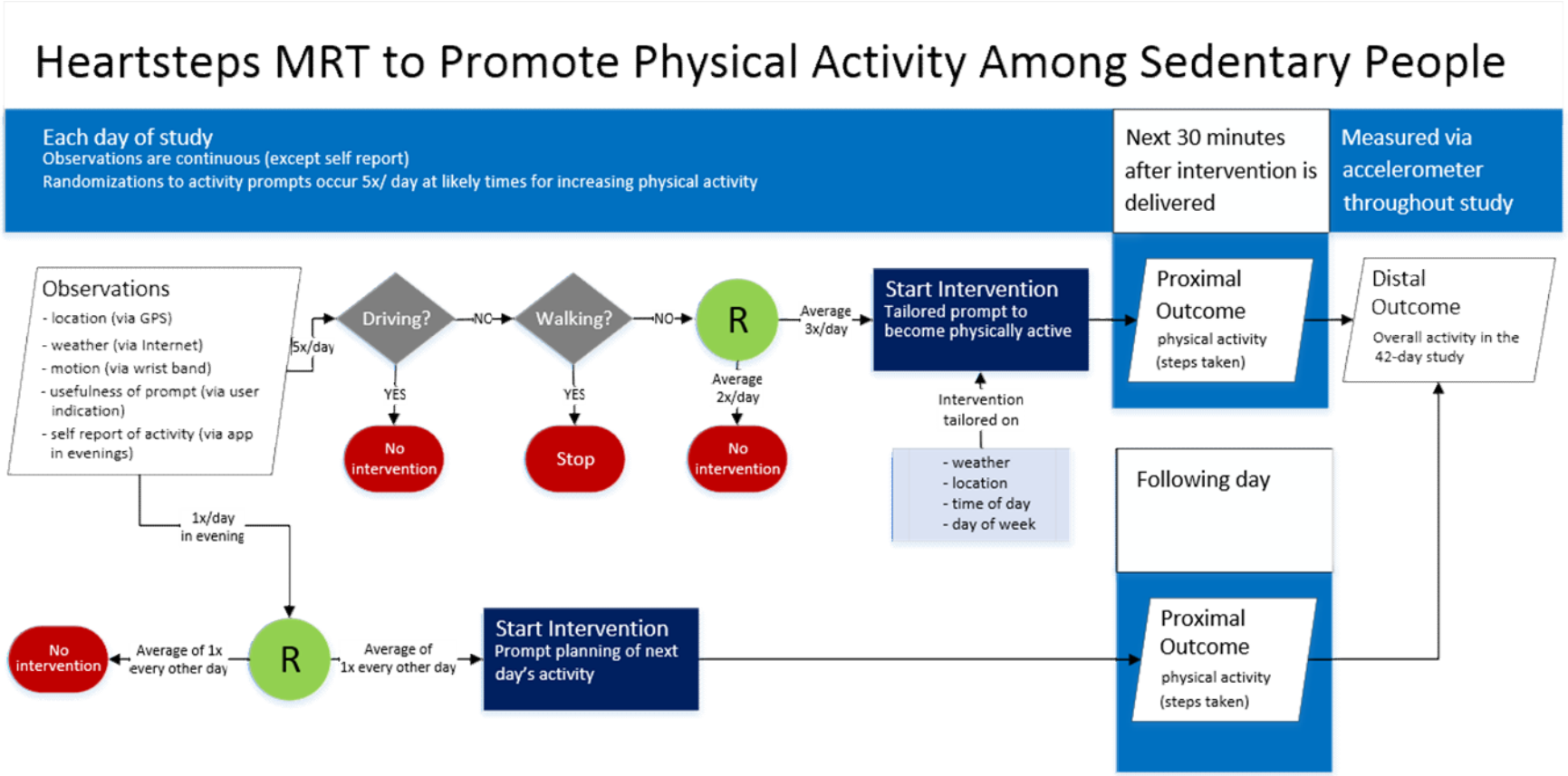
- V1: 42 day micro-randomized pilot study with 37 sedentary individuals,
- V2: 90 day micro-randomized (partially via a bandit) study,
- V3: 365 day personalized study

Data from wearable devices that sense and provide treatments

- On each individual: $S_1\ A_1\ R_2, \ldots, S_t, A_t, R_{t+1}, \ldots$

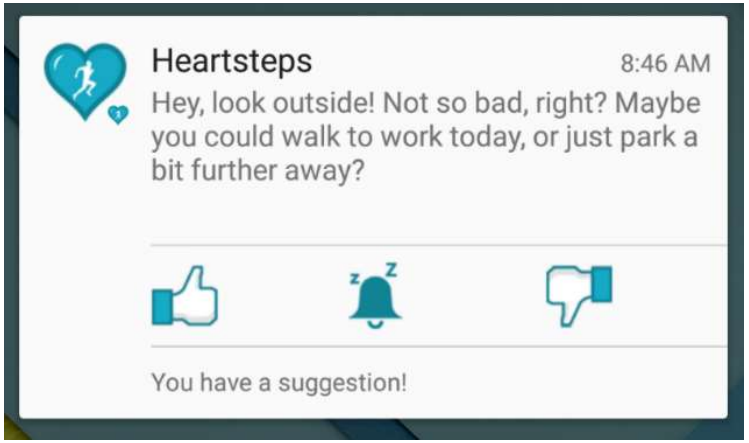
- t: Decision point
- S_t : Context accrued after t-1 and up to/including decision point t (high dimensional)
- A_t : Action at t^{th} decision point (treatment)
- R_{t+1}: Reward (e.g., utility, cost) accrued after time t and prior to time $t+1$

HeartSteps V1



Actions

- Tailored activity suggestion (provide yes/no)



- The set of actions may depend on the context, S_t

Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to predict succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days, prior 30 min step count, total steps on prior day, current temperature

Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to interact with treatment on succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days, prior 30 min step count, total steps on prior day

Goal

During HeartSteps V2 select binary treatment action--whether to provide tailored activity suggestion--online so as to maximize the sum of rewards for each user over the 90 day study (subject to constraints).

1. 5 decision points per day (according to user's work schedule)
2. Reward is the 30 minute stepcount following each decision point *t*.

Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

Linear Thompson Sampling Bandit

1) Linear model for expected reward, e.g. $r(S_t, A_t) = E[R_{t+1}|S_t, A_t] = \eta^T f(S_t, A_t)$
2) Initialize η parameters in expected reward with a prior distribution.
3) Given S_t, A_t, R_{t+1} calculate posterior distribution of η. Mean, covariance matrix of this posterior distribution is η_t, Σ_t.
4) Given S_{t+1}, the probability of selecting treatment, $A_{t+1} = a$, is given by the posterior probability that treatment a has the highest expected reward.

Challenges

1) <u>Noisy data</u>

Our solution: Bandit algorithm

- Bandit algorithms learn faster than full RL algorithms
- The bandit acts as a regularizer (discount rate is 0): trade speed of learning (reduced variance) with bias
- Learn a low dimensional parameterization of the expected reward: linear model in context and treatment. $E[R_{t+1}|S_t, A_t] = \eta^T f(S_t, A_t)$
- Use a Gaussian prior on η with mean, variance based on the data from Heartsteps V1 and a baseline no-treatment week of data from Heartsteps V2

Challenges

2) <u>Nonstationarity: Over longer periods of time, the expected reward function will likely change</u>.

- Due to inability to fully sense, known and unknown, aspects of user's current context.

Our solution:

- Promote continual exploration: Use a Gaussian process prior in the model for the expected reward, e.g. $E[R_{t+1}|S_t, A_t] = f(S_t, A_t)^T \eta_t$ where $\eta_{t+1} = \eta_0 + \gamma(\eta_t - \eta_0) + \varepsilon_t$ $\varepsilon_t \sim N(0, 1-\gamma^2)$

Challenges

3) <u>The immediate effects are primarily positive; the delayed effects are primarily negative.</u> →

- Algorithm may falsely learn that "always treat" is best, yet there are better policies.

Our Solution

- Add a low variance proxy for the "value function" to the current reward

Challenges

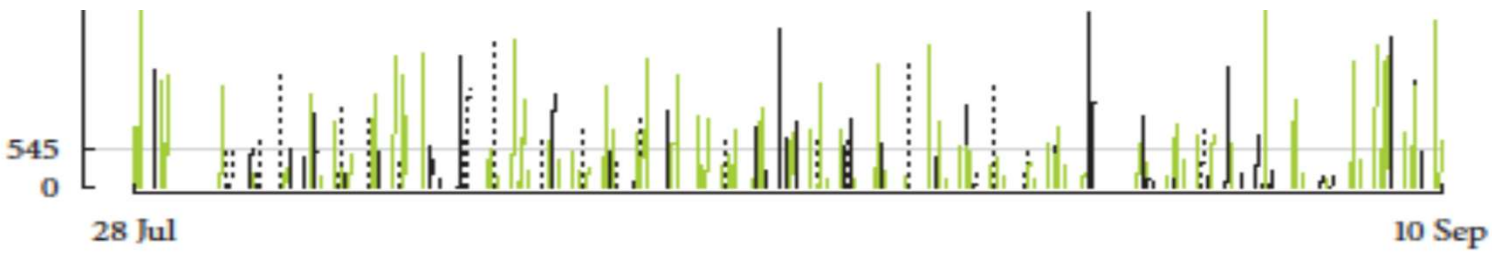
4) <u>Need to ensure ability to conduct "off-policy learning" and causal inference after bandit study</u>

Our solution:

- Use explicit randomization to explore: Thompson Sampling Bandit
- Ensure the no-treatment selection probability lies in an interval bounded away from 0 and 1; here [.2, .9]

Challenges

5) <u>Expected reward, $E[R_{t+1}|S_t, A_t]$ is likely a complex function of context, S_t</u>



Proposed Solution: Center the treatment indicator by binary treatment selection probability, π_t

Action-Centered Bandit

Proposal: For binary A_t:

replace $E[R_{t+1}|S_t = s, A_t = a] = \eta^T f(s, a\ )$

with

$E[R_{t+1}|S_t = s, A_t = a] = b_t(s) + \eta^T f(\mathrm{s})(\mathrm{a} - \pi_t)$

where

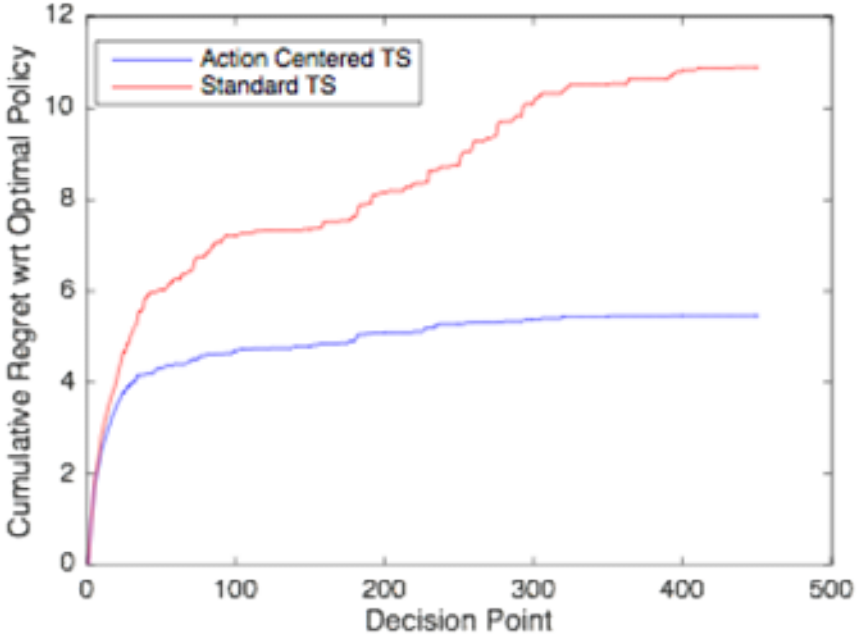
$b_t(s)$ is an unspecified baseline (maybe nonlinear, non-stationary)

$\eta^T f(\mathrm{s})(\mathrm{a} - \pi_t)$ is centered since π_t is the probability of selecting treatment $A_t = a = 1$

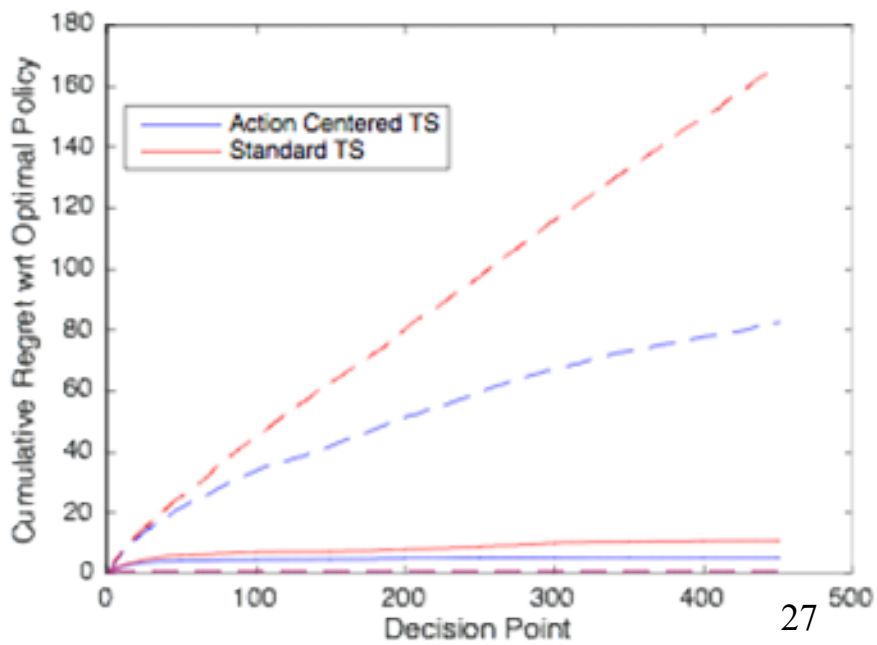
In the Thompson-Sampling update of expected reward use a working (but likely mis-specified) approximation for $b_t(s)$.

Median Regret
500 simulated users

Context, s, is 3 dimensional
True $b_t(s)$ is nonlinear
Linear working model for $b_t(s)$



Quartiles of Regret
500 simulated users



No proxy value
Only Gaussian prior
No Gaussian process prior

Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question

Long Term Goal: Continually learning mHealth App

The learning algorithm is part of the mHealth app

- Incorporate continual learning in the rollout of a mHealth application.
- Learning algorithm makes structural assumptions so as to trade bias and variance in learning

Optimality Criterion?

Multiple goals for learning algorithm

- Track best policy
- Intermittent off-policy inferences
  - causal not just correlational.
  - concern different outcomes than the reward.
  - use different structural assumptions
  - should be valid even if the structural assumptions made by the treatment learning algorithm are false.

Optimality Criterion?

Ideas:

- Minimize finite time T regret subject to bounds on power to detect a particular causal effect at time T?
- Minimize finite time T regret subject to bounds on exploration probability?

Collaborators!


















Notes

Context $s_t=[1,s_{t2},s_{t3},s_{t4}]^T$

nonlinear generative model

$$r_t = I(a_t=1)\theta_1^T s_t + I(a_t=2)\theta_2^T s_t + [1.32, 1.26, .58][1, [s_t]_3, [s_t]_4]^T + 2I(|[s_t]_2| \leq 0.8)$$

Analysis model used for action centering.

$$r_t = (I(a_t>0)-\pi_t)\left[I(\bar{a}_t=1)\theta_1^T s_t + I(\bar{a}_t=2)\theta_2^T s_t\right] + \pi_t\left[I(\bar{a}_t=1)\zeta_1^T s_t + I(\bar{a}_t=2)\zeta_2^T s_t\right] + \eta^T s_t$$

Total params: 20.

Analysis model for std Thompson.

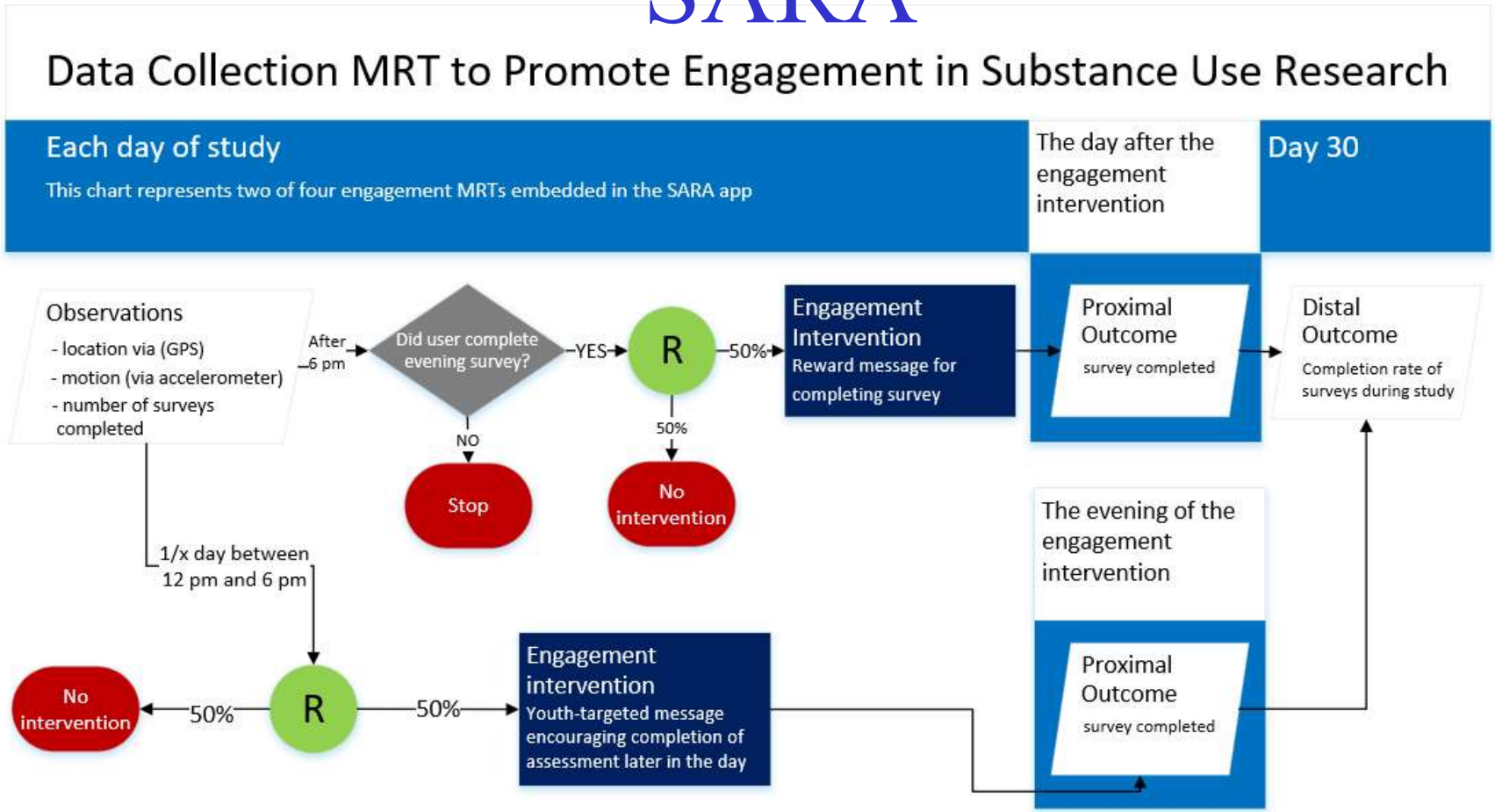
$$r_t = I(a_t=1)\theta_1^T s_t + I(a_t=2)\theta_2^T s_t + \eta^T s_t$$

Total params: 12.

Theta1 = [.116, -.275, -.233, .0425]

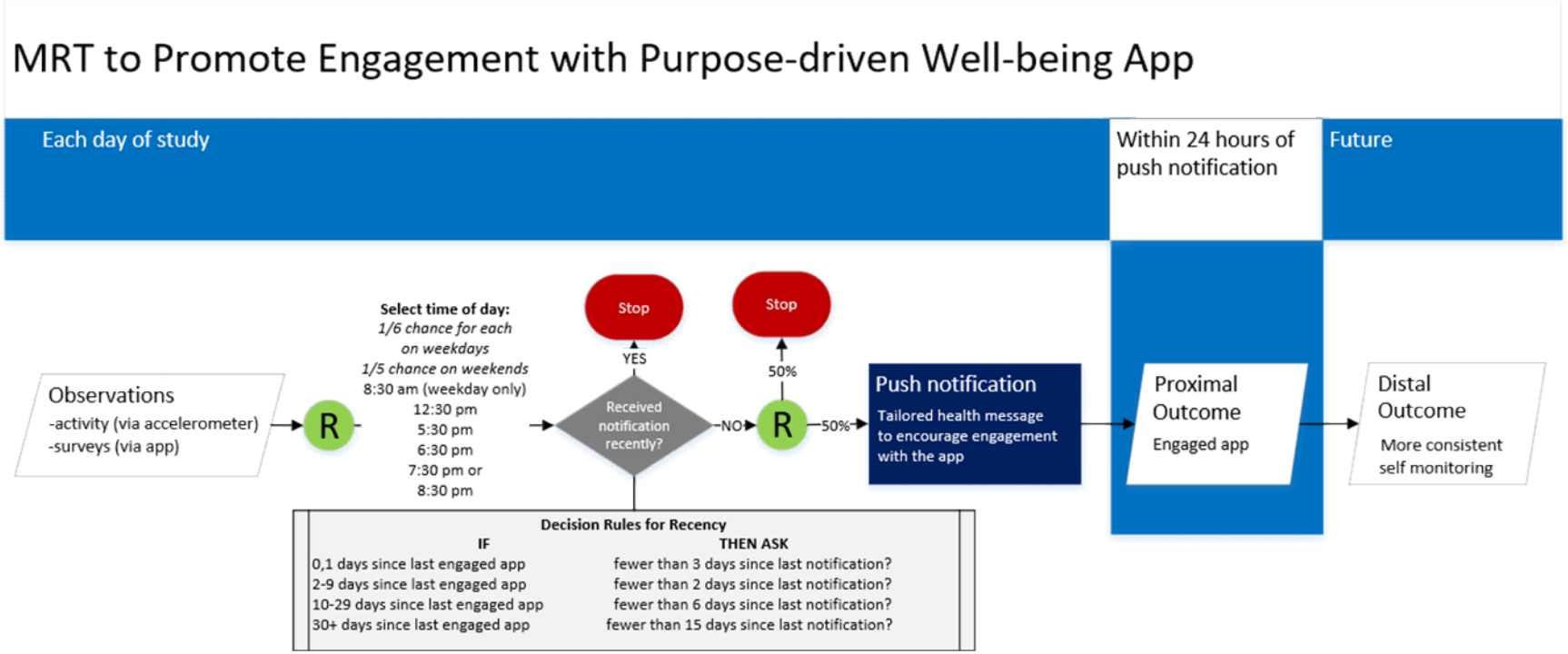
Theta2= [.116, .275, -.233, .0425].

SARA



PIs: M Walton, S Murphy, and M Rabbi Shuvo
Location: University of Michigan
Funding: Michigan Institute for Data Science (PI S. Murphy), University of Michigan Injury Center (PI M. Walton)

Engagement with JOOL



PI: Victor Strecher, PhD, MPH, CEO of JOOL Health
Location & Funding: Ann Arbor, MI
URL: https://www.joolhealth.com