
# Challenges in Developing Learning Algorithms to Personalize Treatment in Real Time

JOOL HEALTH


BariFit

Susan A Murphy



Στατιστική
Statistics


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

The Methodology Center
advancing methods, improving health

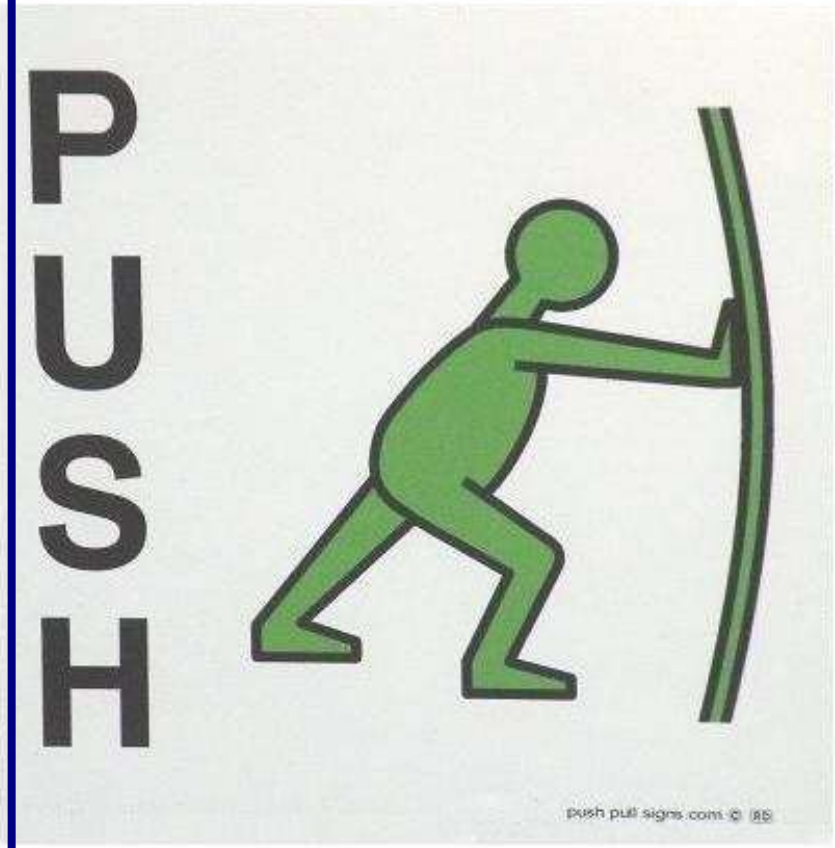
# Outline

- 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

# Mobile Health Treatments



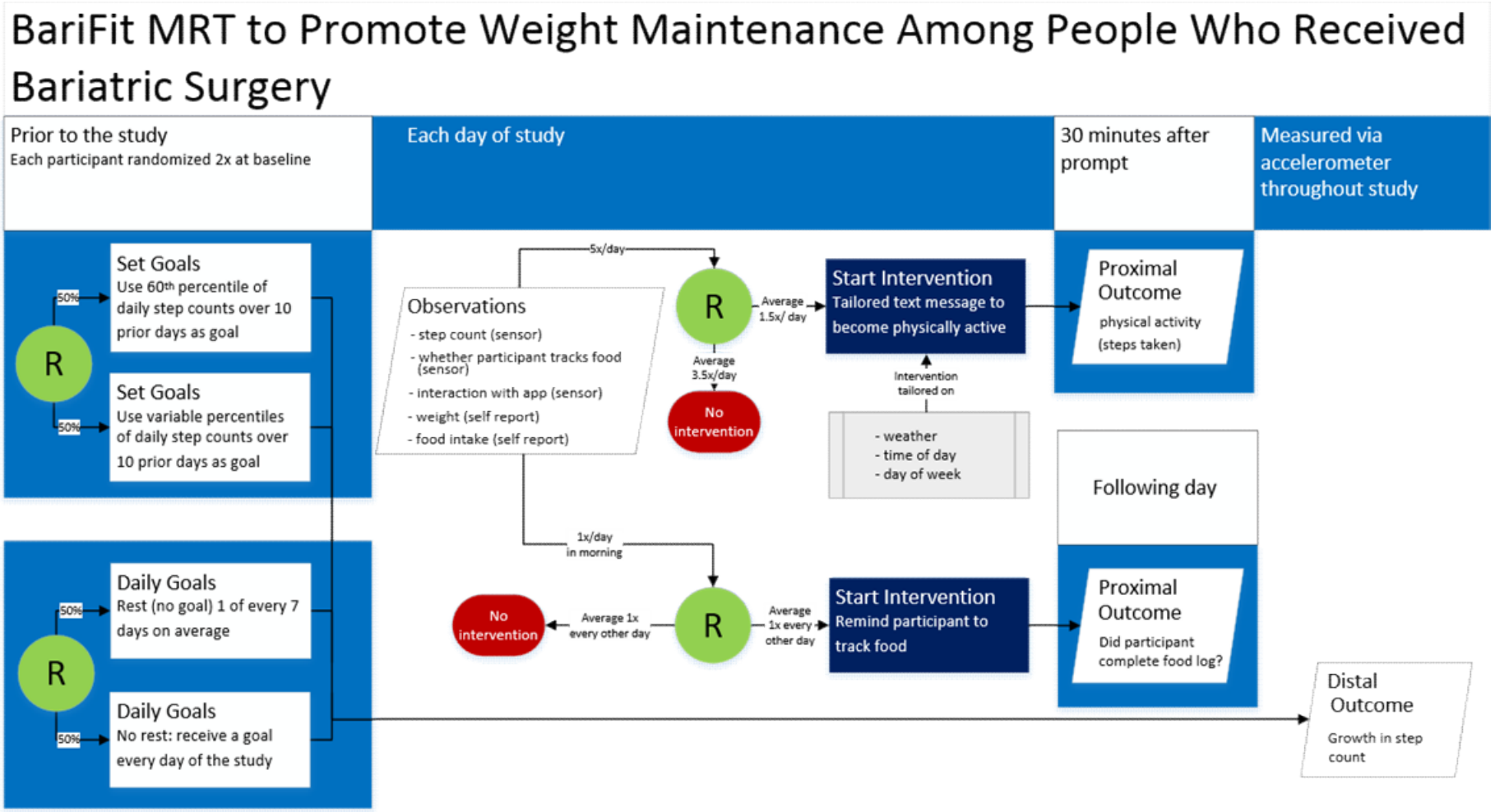

# Outline

- 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 a sequential experimentation strategy
- Sequential experimentation may use online forecasting and reinforcement learning
- Multiple treatment factors occurring at different time scales and which target different outcomes/rewards
- Probabilistic budgets on # of treatment pushes to manage habituation/burden
- Design permits causal inference analyses at study end

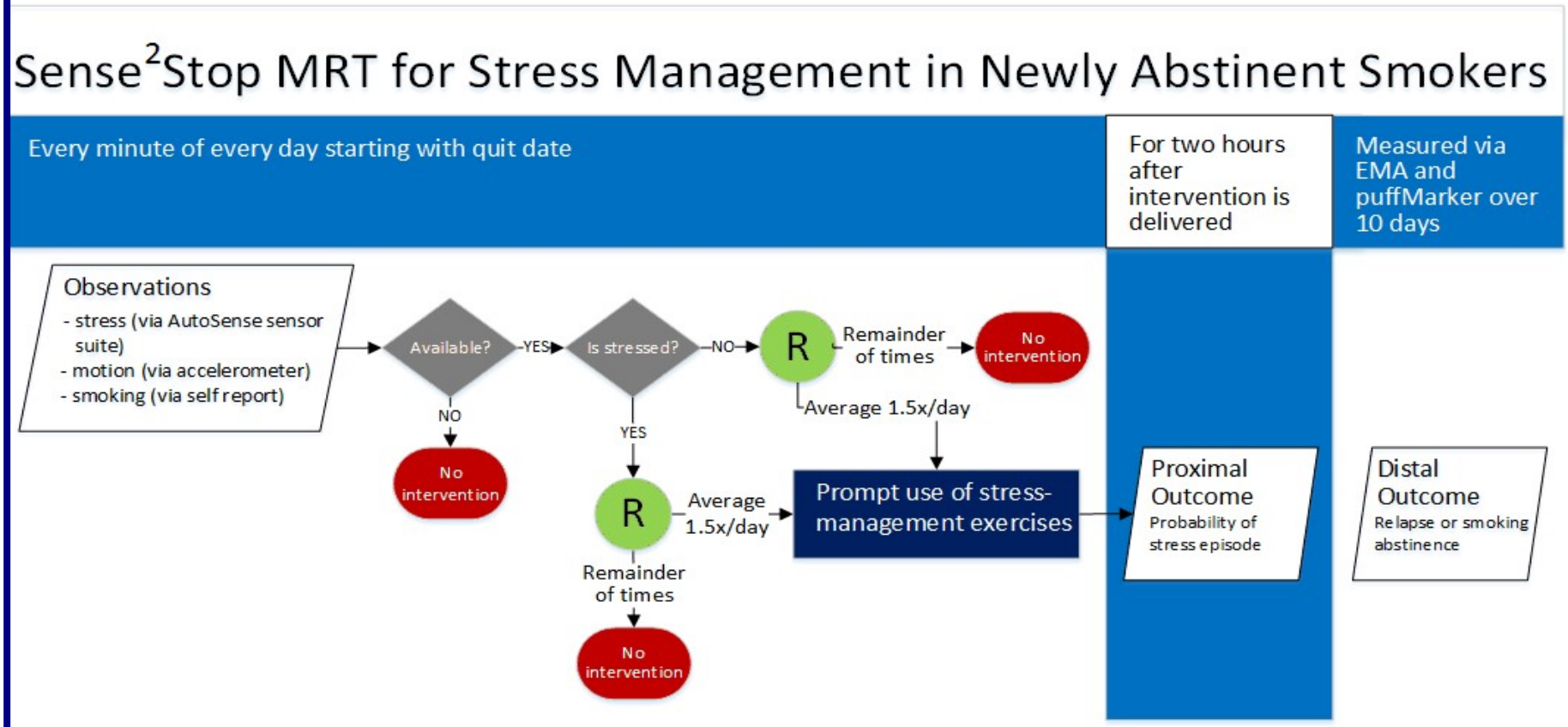
# BariFit




**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

# Data from Micro-Randomized Trial

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

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

# Outline

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

# Reinforcement Learning (RL)

- RL algorithms use sequential experimentation to learn the optimal policy: e.g. how to best select the action $A_t$ after observing context, $S_t$

- The optimal policy maximizes a criterion. This criterion is often a discounted sum of rewards:

$$E\left[\sum_{t\geq 0}\gamma^t R_{t+1}\right] = E\left[\sum_{t\geq 0}\gamma^t r(S_t, A_t)\right]$$

$\gamma$ is discount rate

# Challenges to RL

- State is large yet partially observed: "unknown-unknowns"

  → Non-stationary reward

- Variety of stakeholders with different goals

  → Facilitate end of study causal inference to further develop behavioral science

# Challenges to RL

- High variance within users & complex reward function

  → Slow learning rate

- Treatment actions that tend to have positive effects on immediate rewards but negative impact on future rewards via user habituation/burden.

  → Delayed effects

# Outline

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

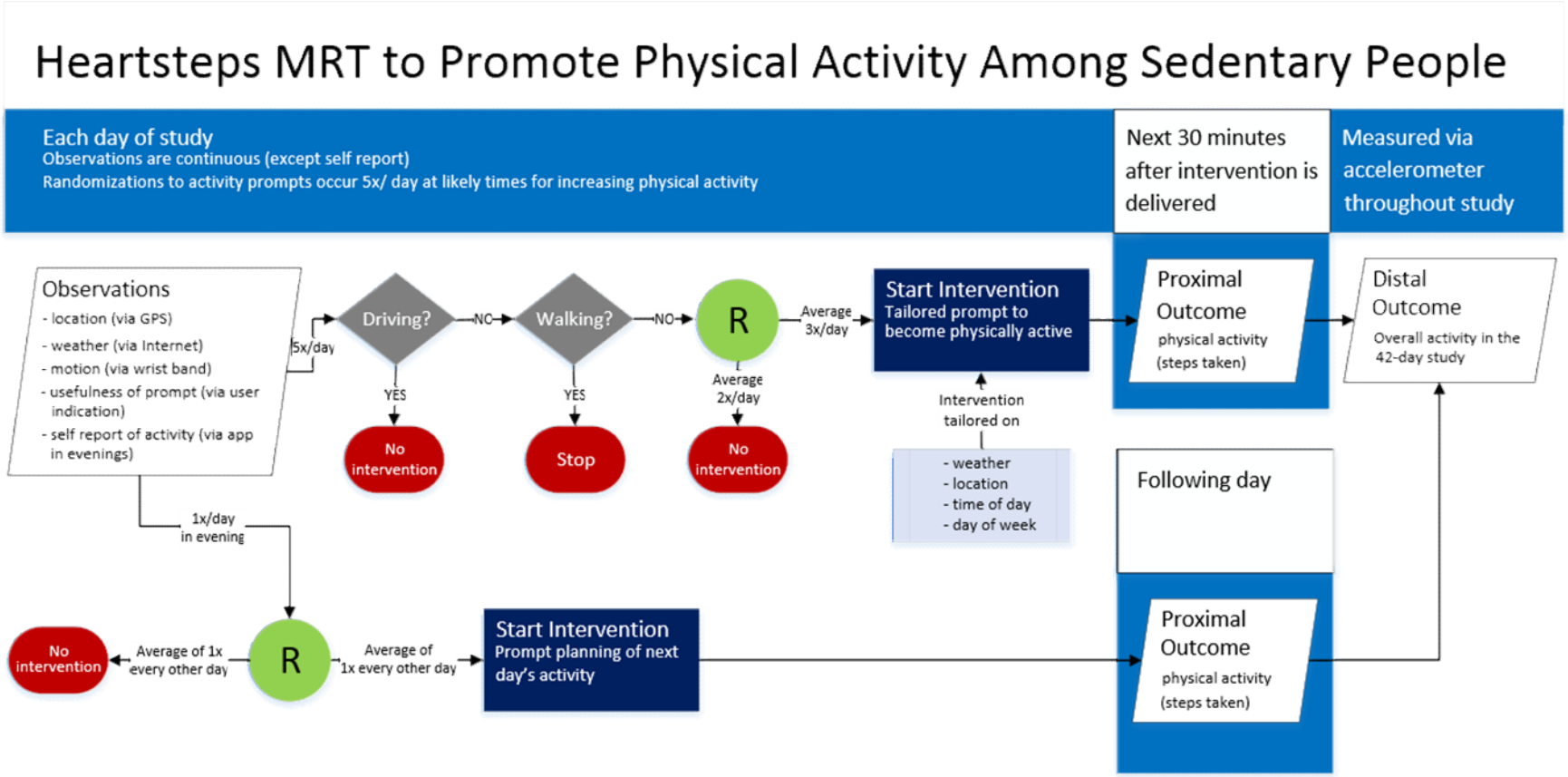
# HeartSteps (PI Klasnja)

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

Three iterative studies:

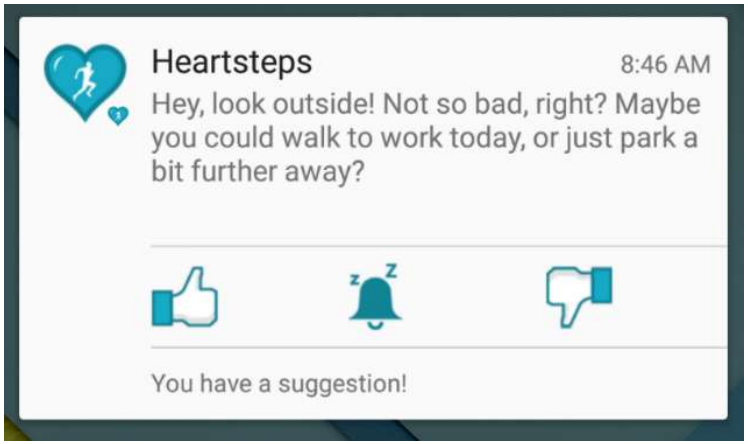
- V1: 42 day micro-randomized pilot study with 37 sedentary individuals,
- V2: 90 day micro-randomized (partially via a bandit) study,
- V3: 365 day personalized study

# HeartSteps V1



# Actions

•Contextually 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

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

3) 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

# Goal

HeartSteps V2: use RL algorithm to select binary treatment action--whether to provide tailored activity suggestion-- 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 (set 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

# A "Bandit" Algorithm

Overview:

1) Initialize parameters in mean reward, $r(S, A)$, given the treatment, $A$ and context feature, $S$.
2) At time point $t$: input current features, $S_t$ and select treatment, $A_t$
3) After the time point: input the reward, $R_{t+1}$.
4) A learning algorithm updates the mean reward, as a function of both the treatment and features.
5) Given the context features at the next time point, $S_{t+1}$ the algorithm uses the updated mean reward to select next treatment, $A_{t+1}$. Go to 3) with $t = t + 1$.

# A "Bandit" Algorithm

Overview:

1) Initialize parameters in mean reward, $r(S, A)$, given the treatment, $A$ and context feature, $S$.
2) At time point $t$: input current features, $S_t$ and select treatment, $A_t$
3) After the time point: input the reward, $R_{t+1}$.
4) A learning algorithm updates the mean reward, as a function of both the treatment and features.
5) Given the context features at the next time point, $S_{t+1}$ the algorithm uses the updated mean reward to select next treatment, $A_{t+1}$. Go to 3) with $t = t + 1$.

# Linear Thompson Sampling Bandit

1) Linear model for mean reward, e.g.
   $E[R_{t+1}|S_t, A_t] = r(S_t, A_t) = \eta^T f(S_t, A_t)$
2) Initialize $\eta$ parameters in mean reward with a prior distribution (here a Gaussian).
3) Given $S_t, A_t, R_{t+1}$ update posterior distribution of $\eta$. Posterior distribution has mean, covariance denoted by $\eta_t, \Sigma_t$.
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 mean reward.

# Challenges in Mobile Health

1) High within user variance

Our solution: Bandit algorithm

- Bandit algorithms learn faster than full RL algorithms
- The bandit acts as a regularizer (discount rate $\gamma$=0): trade speed of learning (reduced variance) with bias
- Use a low dimensional parameterization of the mean 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 $\eta$ with mean, variance based on the data from Heartsteps V1 and a baseline micro-randomized week of data from Heartsteps V2

# Challenges

2) Nonstationarity: Over longer periods of time, the mean reward function will likely change.

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

A solution:

- Promote continual exploration: Use a Gaussian process prior in the model for the mean reward, e.g. $E[R_{t+1}|S_t, A_t] = f(S_t, A_t)^T \eta_t$ where $\eta_{t+1} = \mu_0 + \tau(\eta_t - \mu_0) + \varepsilon_t$

$\varepsilon_t \sim N(0, 1-\tau^2)$

# Challenges

3) The immediate effects are primarily positive; the delayed effects are primarily negative. →

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

Our Solution

- Add to current reward difference between: proxy for the total future rewards if send activity message & proxy for the total future rewards if do not send
- The total future reward is value function for an MDP in which dose evolves deterministically and all other states are iid across time.

# Challenges

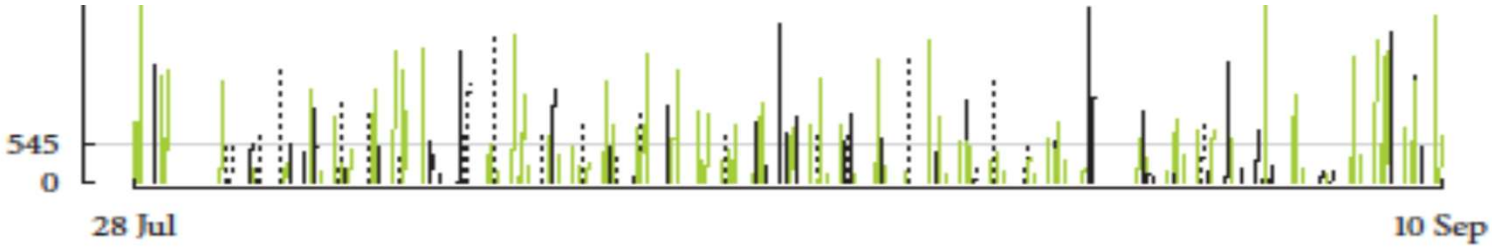
4) Need to ensure ability to conduct "off-policy learning" and causal inference after bandit study

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) Mean reward, $E[R_{t+1}|S_t, A_t]$ is likely a complex function of context, $S_t$




Proposed Solution: Center the treatment indicator by binary treatment selection probability, $\pi_t$

# 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

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

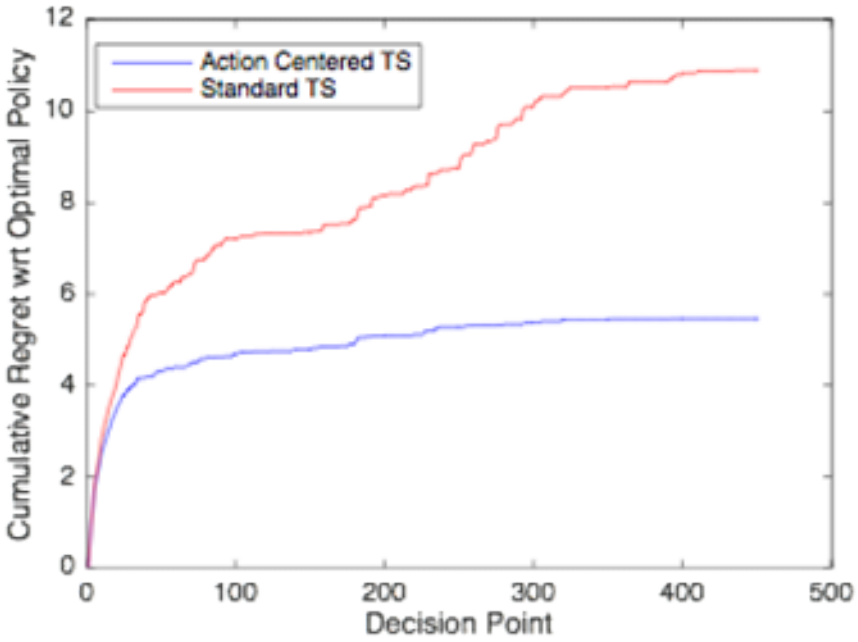
where

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

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

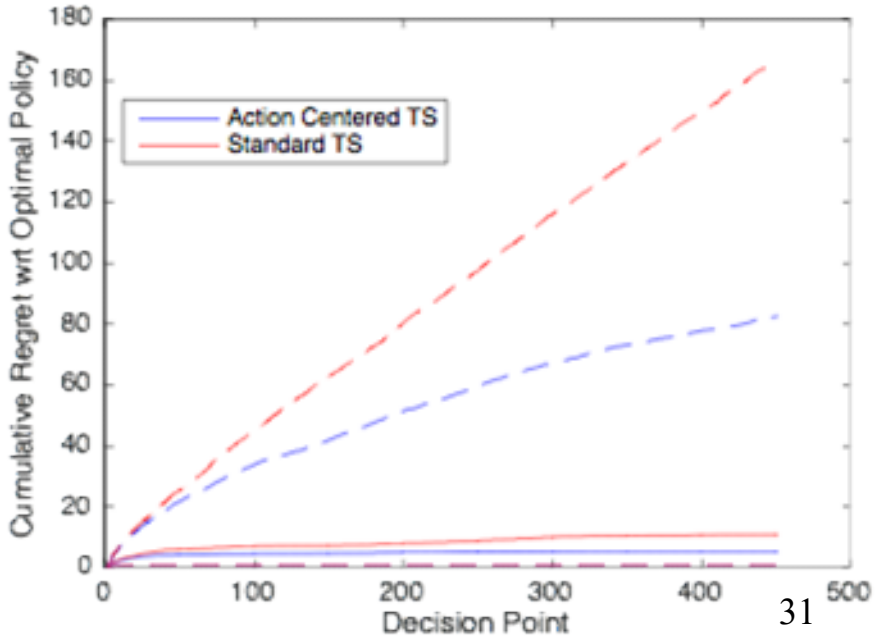
In the Thompson-Sampling update of mean reward use a working (but likely mis-specified) approximation for $b_t(s)$.

## 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 Questions

# Optimality Criterion?

Multiple goals for learning algorithm

- Track best (nonstationary) 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 RL algorithm are false.

# Optimality Criterion?

Ideas:

- Maximize finite time T total reward subject to bounds on power to detect a particular causal effect at time T?
- Maximize finite time T total reward subject to bounds on exploration probability?

# Discussion

Challenges:

- Online accommodation/use of missing data
- High between user variance in performance of online algorithms

Randomization assists in post-study causal inferences based on minimal structural assumptions

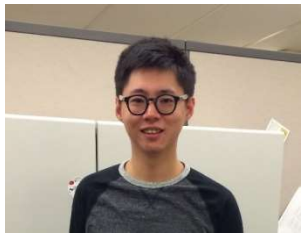
- The bandit algorithm is one way to conduct randomization
- Randomization can also be based more directly on forecasts or predictions

# 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

# Collaborators!




samurphy@fas.harvard.edu

# 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.

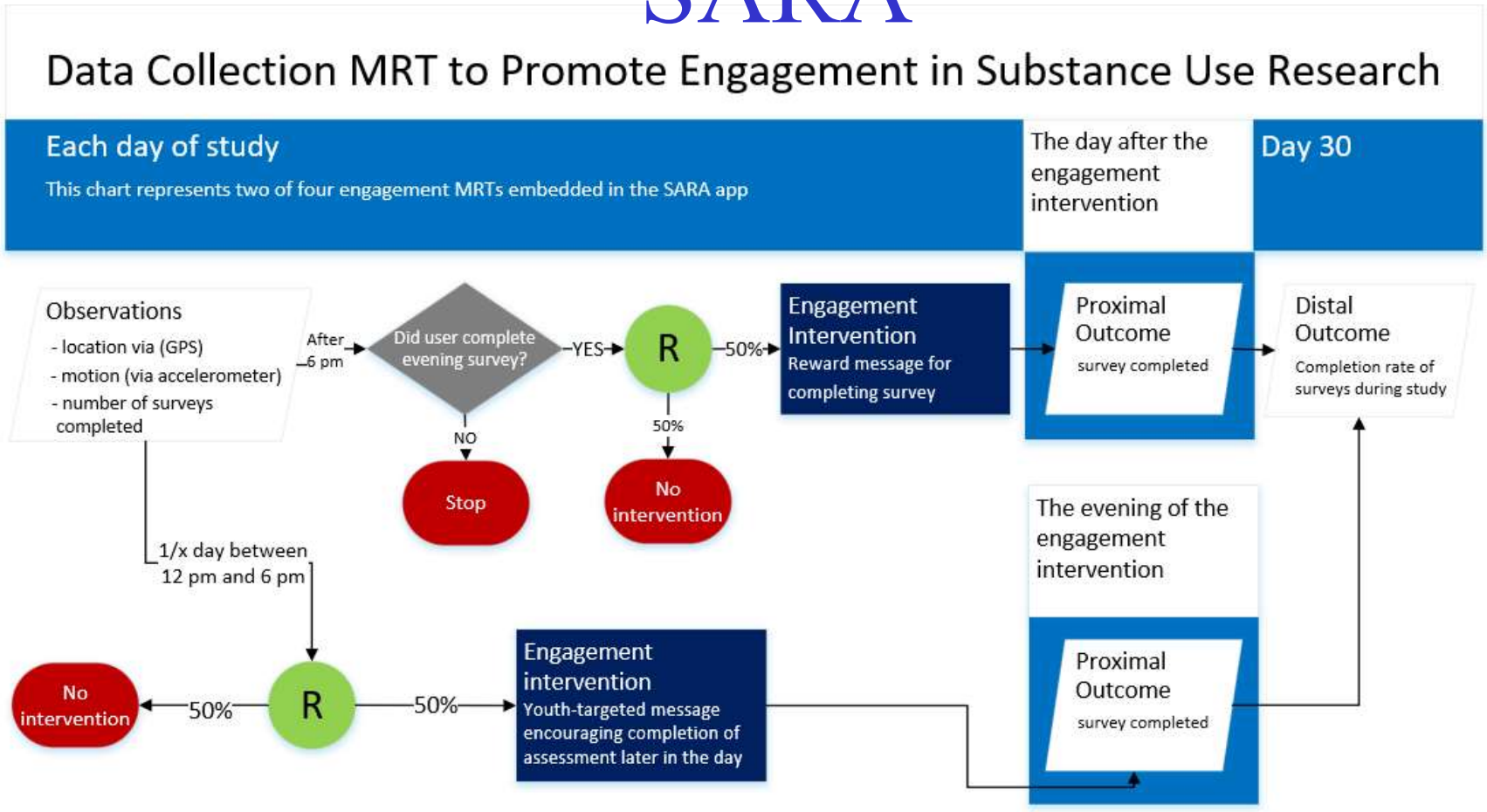
$$r_t = I(a_t = 1)\theta_1^T s_t + I(a_t = 2)\theta_2^T s_t + \eta^T s_t$$

Total params: 12.

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

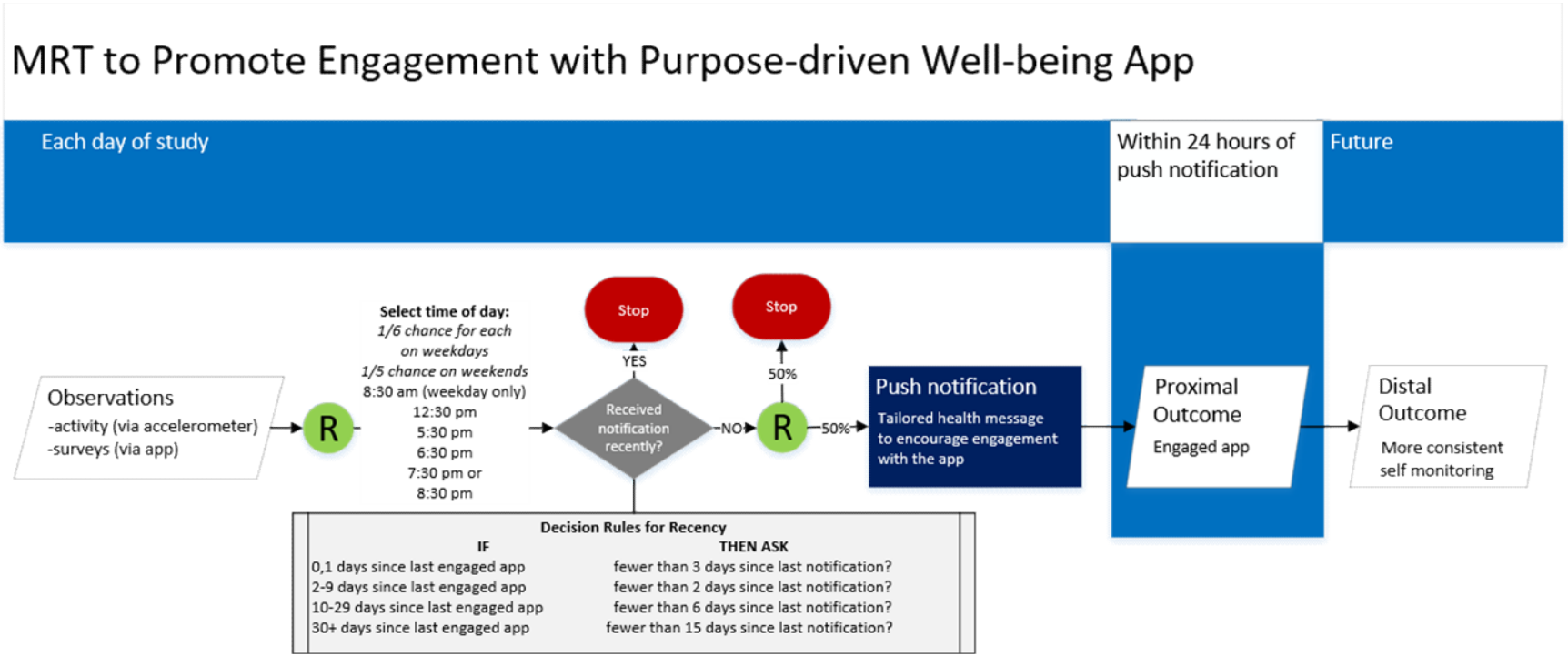
Theta2= [.116, .275, -.233, .0425].

# 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