# Interpretable Statistical Learning for Real-World Behavioral Data

## Dissertation Defense


**Patrick Emedom-Nnamdi**

**Harvard University**


**Committee:** Dr. Jukka-Pekka Onnela, Dr. Junwei Lu, and Dr. Susan Murphy

May 1, 2023

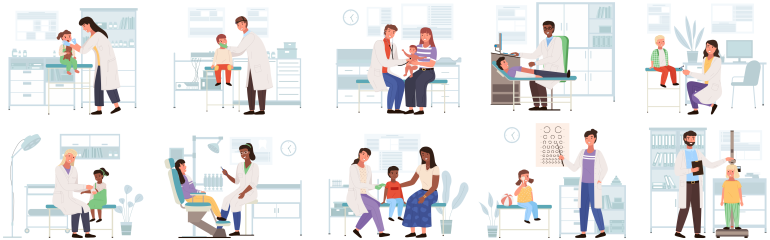
# Outline

1. **Introduction**
2. **Overview of Dissertation Chapters:**
   a. **Chapter 1:** Nonparametric Additive Value Functions: Interpretable Reinforcement Learning with an Application to Surgical Recovery
   b. **Chapter 2:** Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning
   c. **Chapter 3:** Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder
   d. **Chapter 4:** Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – *Piloting the Digital Assessment in Neuro-Oncology*
3. **Conclusion**
4. **Acknowledgements**

# Transition from **Episodic** to Continuous Data Collection

Traditional approaches for evaluating subject behavior in healthcare rely on tools such as **surveys** or **in-clinic interviews**

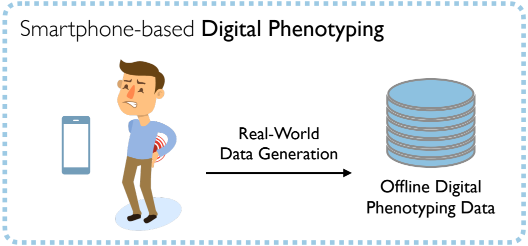
- Administered in fixed periods of time, typically after the event or episode of interest has occurred
  - Resulting in data collection that is **episodic** and **sparse** in nature
- Attempt to build a **chronological ordering of events** based on subject recall
- Build categorizations of patient experience that are heavily **contextualized** to the specific time and location of data collection



# Transition from Episodic to **Continuous** Data Collection

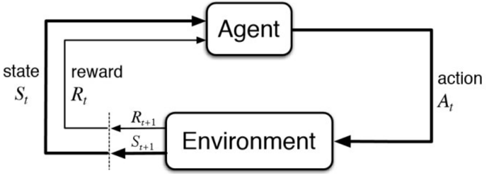
Novel data collection approaches attempt to characterize **real-world** human behavior in an **objective**, **consistent** manner

- Focus on identifying social, behavioral, and cognitive phenotypes that are temporally and contextually dependent
- Aim to improve **reliability** of data collected by minimizing reliance on subject recall
- Leverage devices that have (1) **high adherence** and (2) **use** among a study populations (i.e., smartphones)
- **Digital Phenotyping** is defined as "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular **smartphones**."

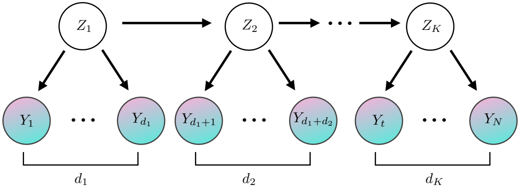


# Research Overview

**Goal of Dissertation:** Develop interpretable statistical machine learning methodology that reveal valuable clinical insight from real-world behavioral data.




(a) **(Offline) Reinforcement Learning**




(b) **Latent State-Space Modeling**

# Smartphone Data Collection and Analysis

**Beiwe Data Collection:** Patients installed the data collection application to their smartphones.

- Beiwe is a high-throughput raw data collection platform in development and use since 2013
- Front-end: native Android and iOS applications for collecting active and passive data
- Back-end: Amazon Web Services (AWS) cloud computing, scalable, globally deployable

**Forest** data analytics library provides a suite of methods for handling data generated from Beiwe

- GPS imputation and summarization, activity recognition (e.g., step counting, walking cadence), survey data aggregation, call/text summarization, etc.
- Implemented in Python and can be run locally or on the AWS backend

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

# Postoperative Spine Disease Recovery

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

**Postoperative recovery** is defined as the period of functional improvement between the end of surgery to the onset of normal functional activity:

- Can vary drastically among patients
- Accompanied by mild to severe complications

**Current Physician-guided Recommendations:**

- Early mobilization activities (e.g., as standing, walking, and getting in and out bed)
- Consistent pain management

**Physicians currently lack:**

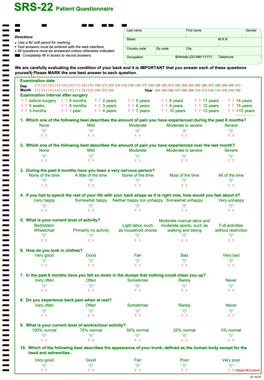
- A quantification of allowable levels of mobilization (e.g., relative step count, walking cadence)
- An understanding of optimal time to initiate mobilization over the course of the recovery period

# Challenges in Current Clinical Practice

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

Current measurements for postoperative assessments rely on **patient-reported outcome measures (PROMs)**

- A set of clinically-validated questionnaires that capture patient outcome
- Administered in person during follow-up visits, or, electronics via surverys
- PROMs are insufficient for most tasks
  - Underutilized by physicians due to administration burden
  - Reliant on patient recall
  - Sparse coverage over the course of patient follow-up
  - Response (shift) bias



**SRS-22** Patient Questionnaire

*Directions*
- *Use a #2 soft pencil for marking.*
- *Text answers must be entered with the web interface.*
- *All questions must be answered unless otherwise indicated.*
- *Completely fill in boxes to record answers.*

Last name | First name | Gender
Street | M.R.N.
Country code | Zip code | City
Occupation | Birthdate (DD.MM.YYYY) | Telephone

**We are carefully evaluating the condition of your back and it is IMPORTANT that you answer each of these questions yourself. Please MARK the one best answer to each question.**

**Examination date**
Day (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31)
Month (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) Year (04) (05) (06) (07) (08) (09) (10) (11) (12) (13) (14) (15) (16)

**Examination interval after surgery**
before surgery, 6 weeks, 3 months, 6 months, 9 months, 1 year, 2 years, 3 years, 4 years, 5 years, 6 years, 7 years, 8 years, 9 years, 10 years, 11 years, 12 years, 13 years, 14 years, 15 years, >15 years

**1. Which one of the following best describes the amount of pain you have experienced during the past 6 months?**
None | Mild | Moderate | Moderate to severe | Severe

**2. Which one of the following best describes the amount of pain you have experienced over the last month?**
None | Mild | Moderate | Moderate to severe | Severe

**3. During the past 6 months have you been a very nervous person?**
None of the time | A little of the time | Some of the time | Most of the time | All of the time

**4. If you had to spend the rest of your life with your back shape as it is right now, how would you feel about it?**
Very happy | Somewhat happy | Neither happy nor unhappy | Somewhat unhappy | Very unhappy

**5. What is your current level of activity?**
Bedridden/ Wheelchair | Primarily no activity | Light labor, such as household chores | Moderate manual labor and moderate sports, such as walking and biking | Full activities without restriction

**6. How do you look in clothes?**
Very good | Good | Fair | Bad | Very bad

**7. In the past 6 months have you felt so down in the dumps that nothing could cheer you up?**
Very often | Often | Sometimes | Rarely | Never

**8. Do you experience back pain when at rest?**
Very often | Often | Sometimes | Rarely | Never

**9. What is your current level of work/school activity?**
100% normal | 75% normal | 50% normal | 25% normal | 0% normal

**10. Which of the following best describes the appearance of your trunk; defined as the human body except for the head and extremities.**
Very good | Good | Fair | Poor | Very poor

*Please fill in back!*

*01.10.07*

# Brigham and Women Hospital's Spine Disease Study Cohort

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

Computational Neuroscience Outcomes Center (CNOC) at Harvard and BWH constructed the first smartphone-based digital phenotyping spine disease study cohort

- **Study Aim:** Quantify real-world patient experience using *reliable* and *objective* measures of quality of life
- $n = 344$ patients with clinically diagnosed spine disease
- Median age of 57 (IQR: 49-68)
- Enrollment period between June 2016 and March 2020, with 6-months of follow-up
- $58.6\%$ of patient received a neurosurgical intervention

# Digital Phenotyping Summary Statistics



**Mobility-based Summary Statistics:**

| | | |
|---|---|---|
| Distance Traveled (km) | Radius of Gyration (km) | Average flight duration (km) |
| Time Spent at Home (hours) | Maximum Diameter (km) | Fraction of the day spent stationary |
| Max. Distance from Home (km) | Num. Significant Places Visited | Time Spent Walking |
| Average flight length (km) | Step Count | Average Cadence |

Table: Subset of GPS and accelerometer-based summary statistics of digital phenotyping.

**Active Data Collection (via micro-surveys):**

- Active: 0-10 Self Reported Pain level
  - Prompt: "Please rate your pain over the last 24 hours on a scale from 0 to 10, where 0 is no pain at all and 10 is the worst pain imaginable"

# Evaluating Postoperative Mobility and Recovery

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

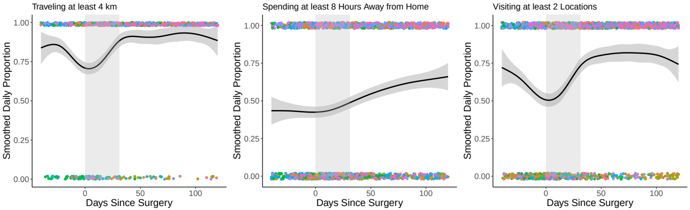



Figure: Smoothed mobility proportions (with standard errors represented in grey) for spine disease cohort centered on day of surgery.

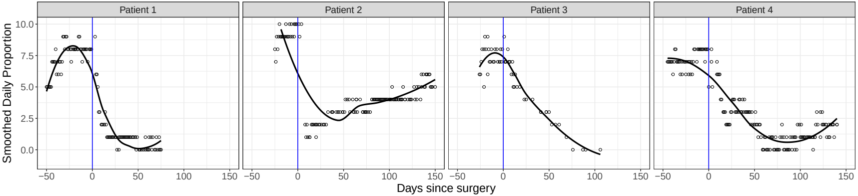



Figure: Pre- and post-operative pain responses with time centered on the day of surgery (i.e., blue line) with a fitted local regression (i.e., black line) for a random selection of patients.

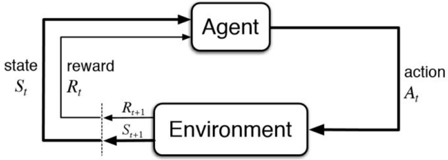
# Research Problem & Objective

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

- **Problem:** How can we improve post-operative rehabilitation for surgical patients?
- **Goal:** Can we learn a *decision-making policy* concerning the *daily post-operative steps* a patient should take in order to improve their recovery process *given their current condition*?
- **Objective:** To estimate **interpretable recovery strategies** for post-operative surgical patients using offline digital phenotyping data

# Re-framing as a Reinforcement Learning Problem

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery




**Policy:** is a any function $\pi : S \to A$ mapping states to actions

**Q-value Function:** provides a mapping from a state-action pair, $(s, a)$, to the *expected total discounted future rewards*

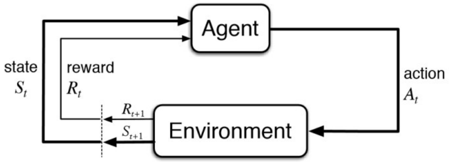
$$Q^{\pi}(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right]$$

when following a given policy $\pi$. Note that $\gamma \in [0, 1)$

Goal is to find **optimal policy**, $\pi^*(s) = \arg\max_{a' \in \mathcal{A}} \underbrace{Q^{\pi^*}(s, a')}_{\text{learn via function approximation}}$

# Need for Interpretable Function Approximators

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery




Current state-of-the-art algorithms approximate $Q^\pi(s, a)$ use **black-box methods**, e.g., neural networks

- Incorporate fitted Q-iteration with modern tools (e.g., replay buffers, target networks)
- Solve complex, high-dimensional decision-making tasks (e.g., Go, chess)

Unfortunately, the success of these modern RL algorithms comes at the **cost of model interpretability**

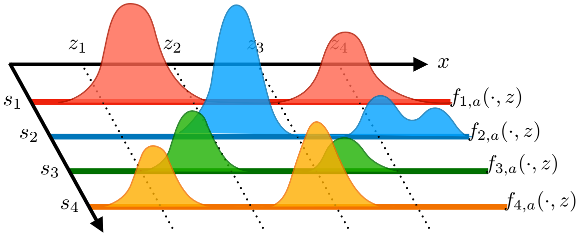
- Unable to examine the contribution each feature makes in producing the model's final suggested action
- Unsuitable in high-risk domains such as health care – hope to anticipate the models performance

# Nonparametric Additive Value Functions

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

We present a **generalized nonparametric, additive** framework for modeling $Q^\pi$:

$$Q^\pi(\mathbf{s}, a, x) = g_a(x) + \sum_{j=1}^{d} f_{j,a}(\mathbf{s}_j, x) + \epsilon$$




- Expand the input space of $Q^\pi$ to include the variable $x \in \mathbb{R}$ (e.g., a candidate state feature, a confounder, or an action)
- $g_a(\cdot)$ represents the **additive marginal effect** of $x$ under action $a$
- $f_{j,a}(\cdot, \cdot)$ represents the **additive joint effect of interactions** between $x$ and state features $\mathbf{s}_j$ under action $a$

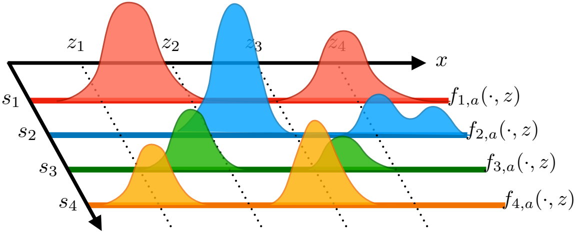
Our model structure allows us to examine **additive nonlinear relationships** that may exist among relevant state features, actions and the variable $x$

# Estimation Strategy for $g_a(x)$ and $f_{j,a}(\mathbf{s}_j, x)$

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

**Step 1 – Basis Expansion:** For an arbitrary fixed constant $z$, we locally express our model using a centered B-spline basis expansion:

$$Q^\pi(\mathbf{s}, a, x = z) \approx \alpha_{a,z} + \sum_{j=1}^{d} \sum_{\ell=1}^{m} \varphi_{j\ell}(\mathbf{s}_j)\beta_{j\ell;a,z}$$




- $\alpha_{a,z}$ is a constant representing the marginal effect $g_a(\cdot)$ at the fixed value $z$
- Each additive component function $f_{j,a}$ is represented as a centered B-Spline basis function

$$f_{j,a}(\mathbf{s}_j, z) = \sum_{\ell=1}^{m} \varphi_{j\ell}(\mathbf{s}_j)\boldsymbol{\beta}_{j\ell;a,z},$$

where $\varphi_{j\ell}(\mathbf{s}) = \psi_\ell(\mathbf{s}) - E[\psi_\ell(\mathbf{s}_j)]$ for the $j$-th component function and the $\ell$-th basis component

# Estimation Strategy for $g_a(x)$ and $f_{j,a}(\mathbf{s}_j, x)$

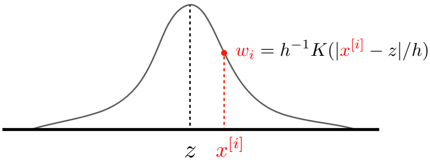


**Step 2 – Kernel (or Locally) Weighted Least Squares Fixed Point Approximation:** Finding a good approximation of $Q^\pi$ equates to forcing the approximate value function to be a fixed point under the Bellman operator $\mathcal{T}_\pi$:

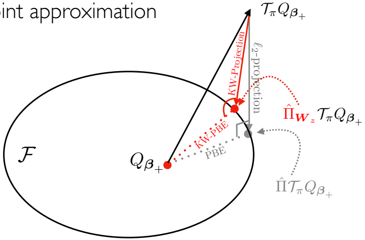
- **Classical Estimation of $Q^\pi$ via LSTDQ (Lagoudakis and Parr 2004)**: $Q_{\boldsymbol{\beta}_+} \approx \Pi \mathcal{T}_\pi Q_{\boldsymbol{\beta}_+}$
- **Kernel Weighted Approach**: $Q_{\boldsymbol{\beta}_+} \approx \Pi_{\boldsymbol{W}_z} \mathcal{T}_\pi Q_{\boldsymbol{\beta}_+}$ (with a group Lasso penalty)

Step a. Construct kernel-weight matrix

$$\boldsymbol{W}_z = \text{diag}(K_h(x^{[1]} - z) \cdots K_h(x^{[N]} - z)) \in \mathbb{R}^{N \times N}$$




Step b. Perform **kernel-weighted** least squares fixed point approximation



# Simulation Study

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

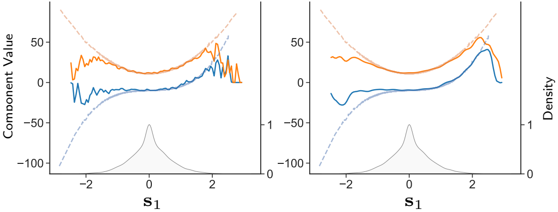
Using the simulated MDP, we examine our algorithm's performance in estimating the marginal nonlinear additive function $g_a(x)$

- We estimate the following nonparametric additive model

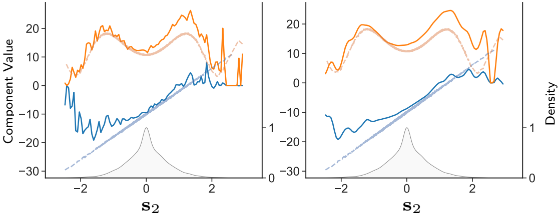
$$Q^{\pi}(\mathbf{s}, a, \mathbf{s}_i) = g_a(\mathbf{s}_i) + \sum_{j \in [d]/i} f_{j,a}(\mathbf{s}_j, \mathbf{s}_i) + \epsilon,$$

  where $x \equiv \mathbf{s}_i$ and $i \in \{1, 2\}$

- We observe that $g_a(\mathbf{s}_i)$ can estimate the true underlying additive component $U_i(\mathbf{s}_i, a)$
  - Performs well primarily for state components that are <u>within the observed distribution</u>




(a) Marginal effect $\hat{g}_a(\mathbf{s}_1)$ vs. Monte-Carlo estimate of $U_1(\mathbf{s}_1, a)$.




(b) Marginal effect $\hat{g}_a(\mathbf{s}_2)$ vs. Monte-Carlo estimate of $U_2(\mathbf{s}_2, a)$.

# Application to Surgical Recovery

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

**Objective:** Estimate a decision-making policy that suggests the **daily number of steps** necessary to reduce long-term ($\gamma = 0.5$) post-operative pain response.

- **Problem Setup:** We construct a simple MDP where each time step $t$ corresponds to a day since surgery
  - a. **State vector** $s^{(t)} \in \mathbb{R}^d$ represents $d = 9$ relevant *digital phenotyping features* and patient-specific *clinical information*
  - b. **Actions** $a^{(t)} \in \{0, 1\}$ are binarized, where $0$ corresponds to moving less than the patient-specific *pre-operative median number of steps* taken per day and $1$ represents moving above this threshold
  - c. **Rewards** $r^{(t)} \equiv$ *negative self-reported pain score*
- We implement a set of nonparametric additive models to estimate action-value functions associated with
  1. A **behavioral policy** $\pi_b$ that aims to mimic decisions commonly taken by subjects, and
  2. An **improved policy** $\pi^*$ retrieved from performing *approximate policy iteration* on the estimated behavioral policy.

# Sample Characteristics

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

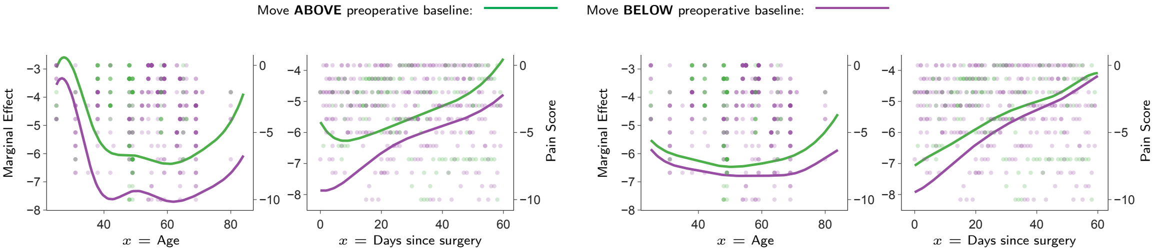
We consider the first 60 days of recovery of $n = 67$ spine patients

- Patients follow-up period of less than 5 days were excluded
- A batch dataset $\mathcal{D} = \{(\mathbf{s}^{(i)}, a^{(i)}, r^{(i)}, \mathbf{s}'^{(i)})\}_{i=1}^{N}$ with $N = 1{,}409$ daily transitions was constructed
- A nonparametric additive model was estimated for each candidate variable of interest:
  - $x \equiv$ age, number of days since surgery

| Variable | $n$ (%) or Median ($25^{th}$–$75^{th}$) |
|---|---|
| **Demographic Data** | |
| Age | 57.0 (48.0–65.5) |
| Female gender | 34 (50.7) |
| **Site of surgery** | |
| Cervical | 19 (28.4) |
| Lumbar | 27 (40.3) |
| Thoracic | 2 (3.0) |
| Multiple | 18 (26.9) |
| **Data Collection** | |
| GPS days of follow-up | 61 (49–61) |
| Accelerometer days of follow-up | 61 (50.5–61) |
| Daily pain survey response rate | 59.4 (42.4–76.9) |
| **Digital Phenotypes** | |
| Number of places visited | 3 (2–5) |
| Time spent at home (hours) | 18.3 (12.9–21.9) |
| Distance traveled (km) | 32.3 (10.8–62.3) |
| Maximum distance from home (km) | 10.6 (4.5–25.5) |
| Radius of gyration (km) | 1.50 (0.18-5.01) |
| Time spent not moving | 21.2 (20.2–22.2) |
| Average cadence | 1.64 (1.55–1.74) |
| Number of steps | 948.6 (356.9–2,005) |

# Marginal Effects $\hat{g}_a(x)$

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery



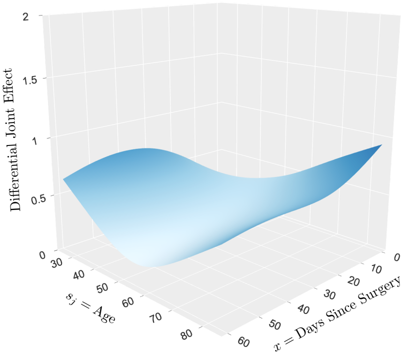

(a) Behavioral policy

(b) Policy Iteration

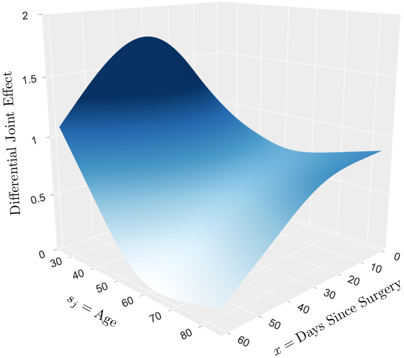
Figure: A comparison of the marginal component function $\hat{g}_a(x)$ of $Q^\pi(\mathrm{s}, a, x)$ estimated under the behavioral policy $\pi = \pi_b$ vs. the improved policy $\pi = \pi^*$.

# Differential Joint Effects $\hat{f}_{j,1}(\mathbf{s}_j, x) - \hat{f}_{j,0}(\mathbf{s}_j, x)$

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery




(a) Behavioral policy




(b) Policy Iteration

Figure: The differential benefit of selecting action $a = 1$ over $a = 0$ with respect to joint effects $\hat{f}_{j,a}(\mathbf{s}_j, x)$ under $Q^{\pi}(\mathbf{s}, a, x)$ estimated for the behavioral policy $\pi = \pi_b$ vs. the improved policy $\pi = \pi^*$.

# Conclusion

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

- We introduce a **flexible** and **interpretable representation** for modeling action-value functions
- Our representation allows for the estimation of **non-linear marginal effect** of select variables and **joint effects** between state features
- Our modeling approach can also accommodate **continuous actions** (i.e., $x \equiv a$)
- In our application to surgical recovery, we reveal **recovery strategies** that are in-line with current clinical practice

# **Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

# Motivation

## **Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

In most real-world applications of RL,

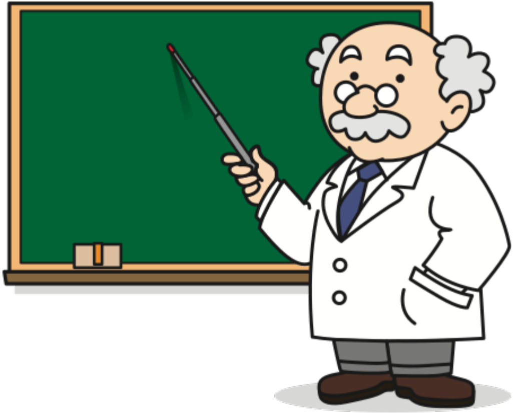
- Continual improvement of decision-making policies during environmental interaction is **infeasible** due to **resource/safety constraints**
- Policies are learned in an offline or a **growing-batch** manner
- Learning optimal policies is difficult due to limited opportunities for **online self-corrections**

# Motivation

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

In many domains, **well-informed/task-specific knowledge** exists and can be queried

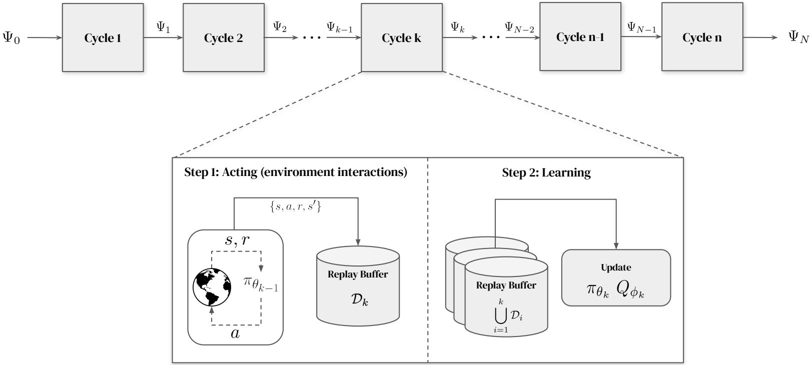
- An external teacher can help alleviate some of these issues, providing **external corrective feedback**
- Teachers can be represented as **human, RL agent, or program**
- Expert knowledge by the teacher can be provided
  1. Up-front via **demonstrations**, or
  2. Throughout the life-cycle of the student agent via **annotations**



**How can we leverage teachers to improve the sample efficiency and performance of value-based RL agents learned in the growing-batch setting?**

# Growing-batch Reinforcement Learning

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning




Policy improvements via offline parameter updates are made only after new batches of experiential data are gathered from the environment.

- **In practice, the number of cycles tend to be small, while the size of each newly gathered batch is large**

# Experimental Setup - Actor Critic Agents



For our experiments, we primarily utilize **D4PG agents** to accommodate continuous actions, where we

1. Learn a critic $Q_\phi(s,a)$ via **distributional TD-Learning**, where $Q_\phi(s,a) = \mathbb{E}Z_\phi(s,a)$ and $Z_\phi(s,a)$ is estimated by minimizing the distributional TD error

$$\mathcal{L}(\phi) = \mathbb{E}_{s\sim\rho^\pi}\left[d\left(\mathcal{T}_{\pi_{\theta'}} Z_{\phi'}(s,a), Z_\phi(s,a)\right)\right]$$

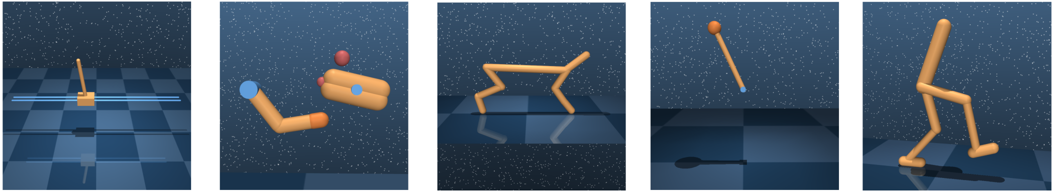
2. Learn a parametric policy $\pi_\theta$ using the **deterministic policy gradient (DPG)**

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{s\sim\rho^\pi}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q_\phi(s,a)|_{a=\pi_\theta(s)}\right]$$

where $\mathcal{J}(\theta) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[Q^{\pi_\theta}(s,a)\right]$ is the expected return.

# Experimental Setup - DeepMind Control Suite





(a) Cart-pole: Balance & Swing-up

(b) Finger-Spin

(c) Cheetah-Run

(d) Pendulum Swing-up

(e) Walker-Run

Figure: The DeepMind Control Suite environments used in our experiments.

**Growing-Batch and Environment Settings:**

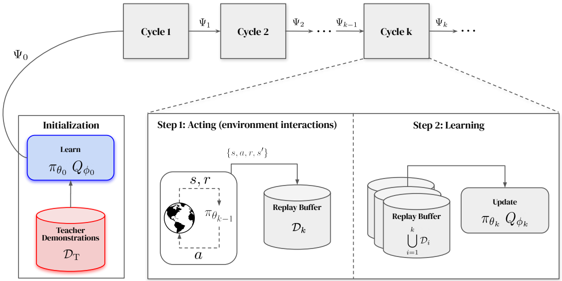
- Evaluated experiments on **six control suite environments**
- Fixed the **total number of actor steps** to 2M, which is the value needed to solve the task by an online agent

# Pre-training Agents using Demonstrations

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

In our first set of experiments, we examine the performance of **naively pre-training a policy** using demonstrations from a teacher performing the task

- Pre-loaded replay buffer with **1M transitions** from 1K episodes
- Pretrained the student's policy $\pi_{\theta_0}$ using **behavioral cloning**
- Using the BC policy, the **critic** is pre-trained using D4PG **policy evaluation**

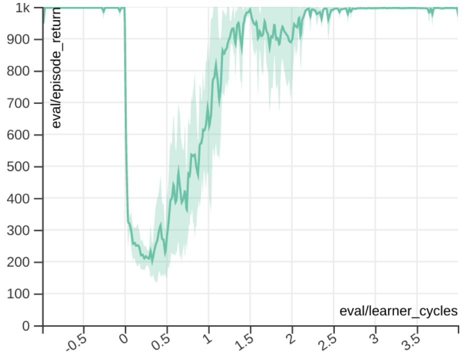


# Pitfalls

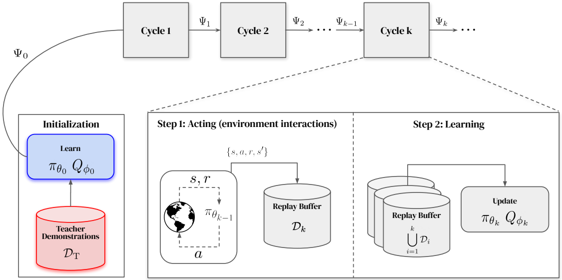
## **Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

Naively pretraining $\pi_0$ using **behavioral cloning performs poorly**

- **Dramatic drop in performance** when training with newly generated data **within the first cycle**




(a) Cartpole Balance, Number of Cycles = 4



# BC-Policy Regularization



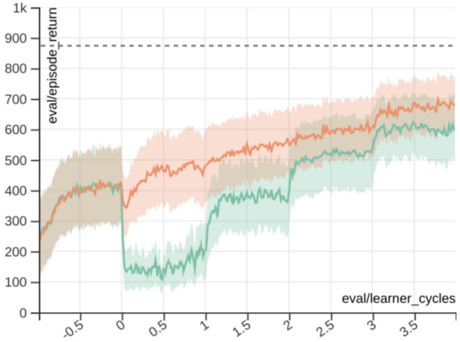
To prevent "forgetting" the initialized policy, we explored incorporating **a regularization term** on the DPG objective function:

$$\mathcal{J}(\theta_k) = \mathcal{J}_{\text{D4PG}}(\theta_k) + \underbrace{\mathbb{E}_{s\sim\rho^\pi}\left[\|\pi_{\text{BC}}(s) - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{BC regularizer}}$$

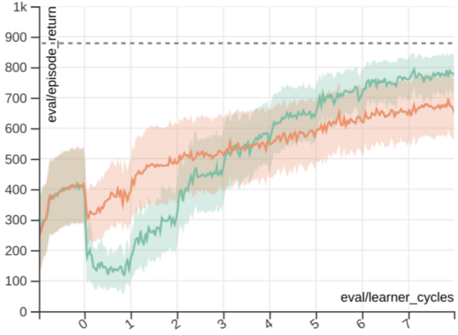
where we explicitly keep our learned policy **close to the BC-initialized policy**.

# Results - BC-Policy Regularization

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning




(a) All Environments, Number of Cycles = 4




(b) All Environments, Number of cycles = 8

Figure: Baseline with BC initialization only vs. BC-policy regularizer

# BC-Policy Regularization (Blissful Ignorance)

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

Assuming that the BC-initialized policy is **sub-optimal**, we would like the agent to **surpass it's performance**:

- Incorporated an **exponential decay weight** $\alpha \in (0, 1)$ as a hyper-parameter

$$\mathcal{J}(\theta_k) = (1 - \alpha)\, \mathcal{J}_{\text{D4PG}}(\theta_k) + \alpha \underbrace{\mathbb{E}_{s \sim \rho^\pi}\left[\|\pi_{\text{BC}}(s) - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{BC regularizer}}$$

- By treating $\alpha$ as a **function of total learner steps**,
  1. We first **stay close** to the initialized policy
  2. Then, gradually **prioritize solely learning from the DPG** loss component

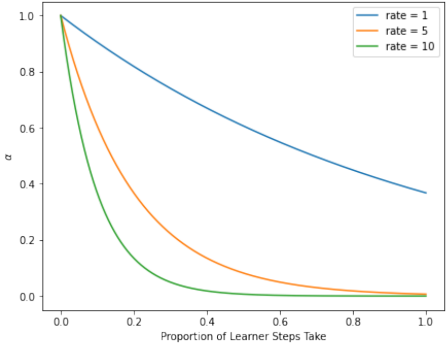



Figure: Exponential decay weight $\alpha$ for various rates.

# BC-Policy Regularization (Blissful Ignorance)

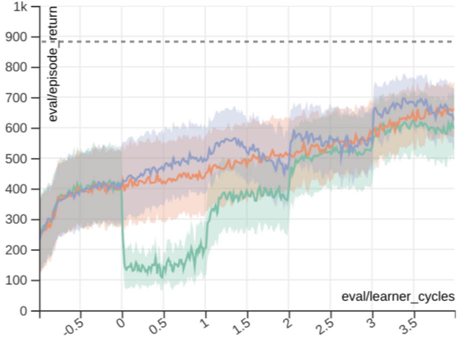


Assuming that the BC-initialized policy is **sub-optimal**, we would like the agent to **surpass it's performance**:

- Incorporated an **exponential decay weight** $\alpha \in (0, 1)$ as a hyper-parameter

$$\mathcal{J}(\theta_k) = (1 - \alpha)\, \mathcal{J}_{\text{D4PG}}(\theta_k) + \alpha\, \underbrace{\mathbb{E}_{s \sim \rho^{\pi}}\left[\|\pi_{\text{BC}}(s) - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{BC regularizer}}$$

- By treating $\alpha$ as a **function of total learner steps**,
  1. We first **stay close** to the initialized policy
  2. Then, gradually **prioritize solely learning from the DPG** loss component




(a) All Environments, Number of cycles = 4

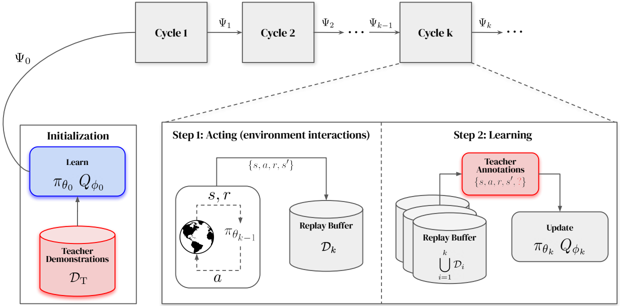
Figure: Baseline with BC initialization only vs. BC-policy regularizer with decay rate $= 1$ and decay rate $= 5$.

# Teacher Guided Annotations

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

While BC provides an initialization to a good policy, our agent still risks learning a sub-optimal policy due to the following issues related to (1) **insufficient state-action coverage** and (2) **overestimation bias**:

- We considered various forms of **annotations** provided by a teacher at training time
- We explore using **teacher-suggested actions** as a mechanism for mimicking the teacher



# Teacher-Suggested Actions - Varying Decay Rates

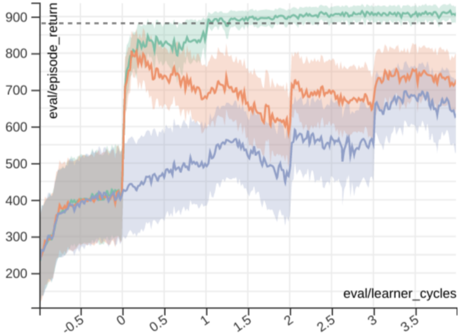


We query the teacher's action $a^* \sim \pi^*(s)$ during training time:

- We incorporate an $\ell_2$ loss component that evaluates the difference between **the student's policy** and **the teacher-suggested action**
- Additionally, we incorporate a **between-cycle regularizer** to promote monotonic improvement between cycles

$$\mathcal{J}(\theta_k) = (1-\alpha)\,\mathcal{J}_{\text{D4PG}}(\theta_k) + \beta_k \underbrace{\mathbb{E}_{s\sim\rho^\pi}\left[\|\pi_{\theta_{k-1}}(s) - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{Between-cycle policy regularizer}} + \alpha \underbrace{\mathbb{E}_{s\sim\rho^\pi}\left[\|a^* - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{Teacher-action Component}}$$

- The weight $\beta_k \in (0,1)$ decays with respect to cycle $k$




(a) All Environments, Number of Cycles = 4

Figure: Baseline using BC-policy regularizer vs. Teacher-action auxiliary loss with decay rate $= 1$, and decay rate $= 5$.

# Filtered Teacher-Suggested Actions

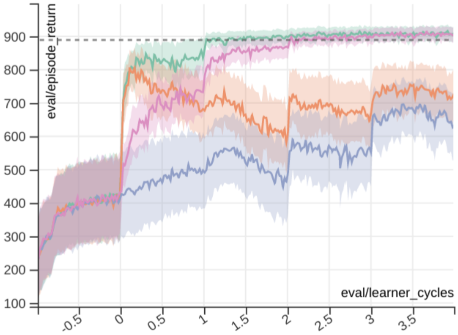


We explore using a Q-filter to **determine which loss component (Teacher-action vs. D4PG) to learn from** for a given transition:

$$\nabla \mathcal{J}(\theta_k) = \mathbb{E}_{s\sim\rho^\pi}\Big[\underbrace{\left[1-\delta(s)\right]\nabla_{\theta_k}\pi_{\theta_k}\nabla_a Q_{\phi_k}(s,a)\big|_{a=\pi_{\theta_k}(s)}}_{\text{D4PG component}}$$
$$+ \underbrace{\delta(s)\,\nabla_{\theta_k}\|a^* - \pi_{\theta_k}(s)\|_2^2}_{\text{Teacher-action component}}\Big],$$

where $\delta(s) = 1\left[Q_{\phi_k}(s, a^*) \geq Q_{\phi_k}(s, \pi_\theta(s))\right]$ is a **Q-filter**.

- Ignore (or de-prioritize) learning from actions that produce lower values than using $\pi_{\theta_k}(s)$



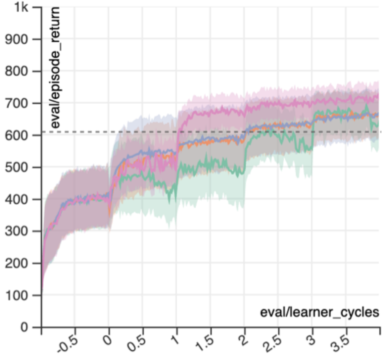

(a) All Environments, Number of Cycles = 4

Figure: Baseline using BC-policy regularizer vs. Filtered Teacher-action vs. Teacher-action auxiliary loss with decay rate $= 1$, and decay rate $= 5$.

# Leveraging Sub-optimal Teachers

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning

Information from sub-optimal teachers can be used to **"jump-start" new policies**:




(a) All Environments, Number of Cycles = 4

Figure: Baseline using BC-policy regularizer vs. Filtered Teacher-Actions vs. Teacher-Gradient auxiliary loss with decay rate $= 1$, and decay rate $= 5$.

At a first glance of the **deterministic policy gradient**, one approach is to replace $\nabla_a Q_\phi(s,a)|_{a=\pi_\theta(s)}$, with the **gradient of the teacher's Q-function** w.r.t to the students policy:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\rho^\pi}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi^{\text{teacher}}}(s,a)\Big|_{a=\pi_\theta(s)}\right]$$

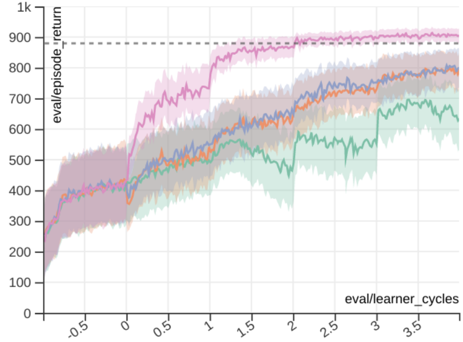
In practice, this can be achieved by modifying the DDPG policy loss as follows:

$$J(\theta_k) = (1-\alpha)\, J_{\text{D4PG}}(\theta_k) + \alpha \underbrace{\|\nabla_a Q^{\pi^{\text{teacher}}}(s,a) + a - \pi_{\theta_k}(s)\|_2}_{\text{Expert Gradient DPG Component}} + \underbrace{\mathbb{E}_{s\sim\rho^\pi}\left[\|\pi_{\text{BC}}(s) - \pi_{\theta_k}(s)\|_2^2\right]}_{\text{BC regularizer}}$$

- **Intuition:** This loss encourages the agent to learn a policy that selects actions in the direction that maximizes the teacher's critic.

# Results - Teacher Gradients

**Chapter 2** - Knowledge Transfer from Teachers to Learners in Growing-Batch Reinforcement Learning




(a) All Environments, Number of Cycles = 4

Figure: Baseline using BC-policy regularizer vs. Filtered DAgger vs. Expert Gradient auxiliary loss with decay rate $= 1$, and decay rate $= 5$.

# Conclusion

1. Initially staying close to the BC policy completely avoids the previous drop in performance, but can dampen overall performance if "forgotten" slowly.
2. Learning directly from the teacher's action provides a substantial performance boost, but is **sensitive to the choice rate parameter** for $\alpha$.
3. While the Q-filter initially under-performs, it **adaptively** reaches superior performance **without** use of hyper-parameter tuning.

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

# Motivation

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

**Bipolar disorder** is associated with significant morbidity and mortality

- 2-3 fold **increased risk** for pre-mature mortality (e.g., cardiovascular disease, diabetes, COPD, unintentional injuries, suicide)
- NIMH estimates BP is prevalent among 2.6% of US adults and 2.9% of adolescents (Ages 13–18)
- Recurrent mood episodes are common (i.e., emotional highs or lows)
  - Risk of **future episodes** increases with each occurrence
  - The number of episodes are **associated with poor outcomes**
  - Symptoms typically emerge prior to the **onset of an episode**
- Meta-analysis (2016) revealed that energy changes, increased activity, sleep problems, and physical agitation are common prior to BP episodes

# Motivation

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

**Study Goal:** Use digital phenotyping to develop features that can help identify the onset of notable mood episodes

**Current Objective:** Decode *resting* and *active* patterns of smartphone use that correlate with sleep

# Study Cohort

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

We study a subset of participants enrolled at NYU Langone Health & Northwell Health

- $n = 43$ adolescents (median age of 16):
  - 20 participants had a **diagnosis of bipolar disorder** and are risk of developing a **new mood episode**
  - 23 participants were selected as **healthy controls**
- Each participant installed the Beiwe research platform on their smartphones
- Examined up to **1-year of data collected** starting from January 1, 2021

| Variable | $n$ (%) or Median ($25^{th}$–$75^{th}$) |
|---|---|
| **Demographic Data** | |
| Age | 16 (15-17) |
| Sex | |
| Female | 23 (53.5%) |
| Male | 18 (41.9%) |
| Other | 2 (4.6%) |
| **Diagnosis** | |
| Bipolar | 22 (51.2%) |
| Typically-developing | 21 (48.8%) |
| **Race** | |
| Asian | 2 (5) |
| Asian, Black or African American | 2 (5%) |
| Asian, Black or African American, White | 1 (2%) |
| American Indian or Alaska Native, Asian | 1 (2%) |
| Black or African American | 3 (7%) |
| Black or African American, White | 2 (5%) |
| South Asian | 3 (7%) |
| White | 29 (67%) |
| **Data Collection** | |
| Phone usage days of follow-up | 223 (116–322) |
| Number of surveys recorded | 98 (53–148) |

# Smartphone Data Collection

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

Data analyzed were primarily collected from smartphone usage logs

- We focused on events that indicate moments of smartphone interactivity:
  - **iOS devices:** the smartphone being "Unlocked"
  - **Android devices:** the smartphone's screen being "turned on"
- Data were recorded as the event occurred in real-time
- Events were binned into **five-minute intervals** starting from the onset of the day of first data collection for patient $i$ (i.e., 00:00:00 - 00:05:00)
- Periods within a day where no usage events were recorded received an entry of 0
- Entire days where no usage events were recorded
- We denote the number of events that occurred with the interval $t$ as $Y_t \in [0, \infty)$
- Across all patients, we collected 2.63M observations

# Hidden Semi-Markov Models

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

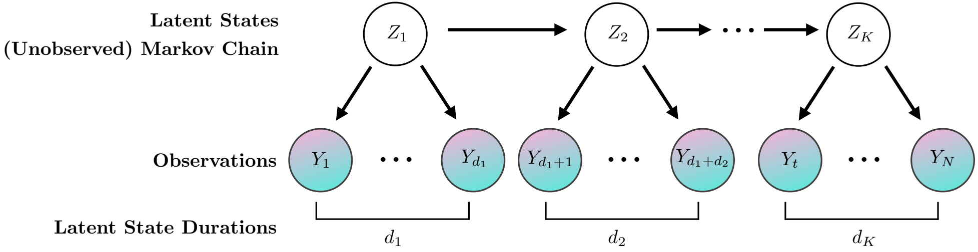
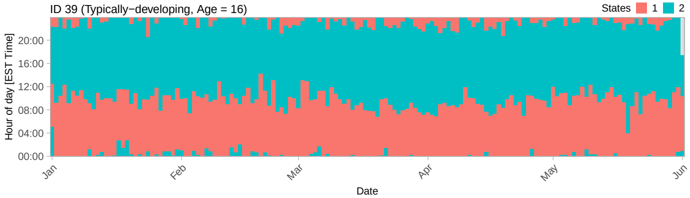


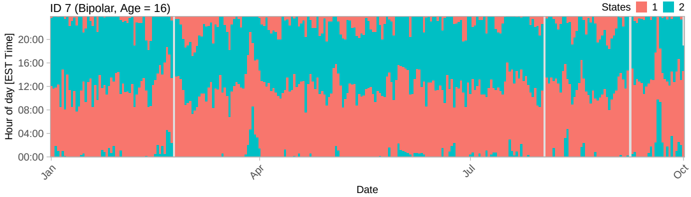

Figure: Representation of a Hidden-Semi Markov Model with state durations $d_j$

# Decoding Viterbi Paths

## **Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder



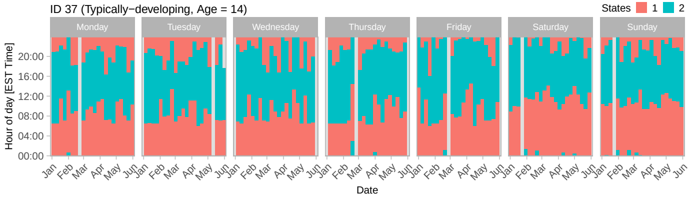

(a) Viterbi path of Participant ID 39 (*typically-developing*, age 16).



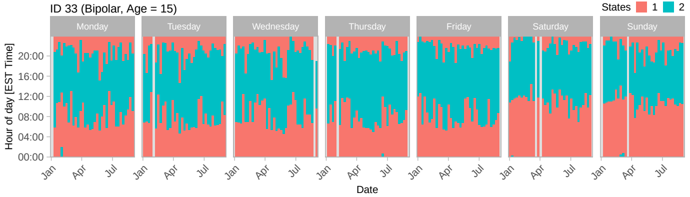

(b) Viterbi path of Participant ID 7 (*bipolar*, age 16).

# Decoding Viterbi Paths – Day of the Week Effects






(a) Viterbi path of Participant ID 37 (*typically-developing*, age 14) stratified by day of the week.




(b) Viterbi path of Participant ID 33 (*bipolar*, age 15) stratified by day of the week.

# Latent Regression Analysis



Using the decoded states, we performed a regression analysis using a linear mixed effects models:

a. **Model A:** To identify temporally-dependent trends between time of day and phone use under the resting and active states
b. **Model B:** To identify a differential state duration between bipolar participants and healthy controls

In both model, we adjust for the same set of potential confounders: age, sex, race, and day of the week.

# Model A – Temporally-dependent Phone Use



**Model A**: To identify temporally-dependent trends between time of day and phone use under the resting and active states

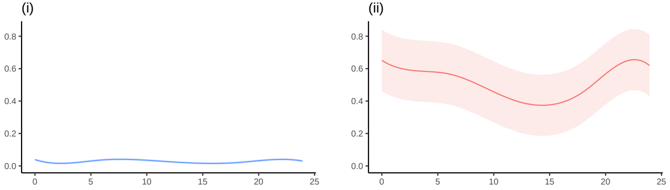
- We fit the following linear mixed effects model:

$$Y_{ti}|Z_{ti} = j \sim \mu + f(X_{ti}) + \mathbf{Z}_{ti}^{\top}\alpha + \gamma_i + \epsilon_{ti}, \quad j \in \{\text{resting}, \text{active}\}$$

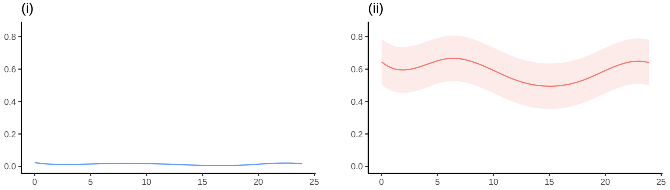
  where $\mu$ is the intercept, $X_{ti}$ is our variable of interest, $\mathbf{Z}_{ti}$ is a set of potential confounders, and $\gamma_i$ is a subject-specific random intercept
- We condition the response variable $Y_{ti}$ on $j$, the value of the decoded latent state
- $X_{ti} \in [0, 24)$ represents time in hours
- $f$ is a B-spline function with knots at 6, 12, and 18

# Model A – Temporally-dependent Phone Use, $\hat{f}(X_{ti})$

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder




(a) Temporally-dependent smartphone use for *typically-developing* adolescents




(b) Temporally-dependent smartphone use for *bipolar* adolescents

# Latent Regression Analysis

## **Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

**Model B**: To identify a differential state duration between bipolar participants and the health controls

- We fit the following linear mixed effects model:

$$\hat{d}_{ti}|Z_{ti} = j \sim \mu + f(X_{ti}) + \mathbf{Z}_{ti}^{\top}\alpha + \gamma_i + \epsilon_{ti}, \quad j \in \{\text{resting}, \text{active}\}$$

  where $\mu$ is the intercept, $X_{ti}$ is our variable of interest, $\mathbf{Z}_{ti}$ is a set of potential confounders, and $\gamma_i$ is a subject-specific random intercept
- We condition the response variable $\hat{d}_{ti}$ on $j$, the value o f the decoded latent state
- $\hat{d}_{ti}$ is the estimated state duration on day $t$ under participant $i$ and state $j$
- $f(X_{ti}) = \beta \cdot X_{ti}$, where $X_{ti}$ is an indicator function for Bipolar status

# Latent Regression Analysis

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

**Model B**: To identify a differential state duration between bipolar participants and the health controls

- We fit the following linear mixed effects model:

$$\hat{d}_{ti}|Z_{ti} = j \sim \mu + f(X_{ti}) + \mathbf{Z}_{ti}^{\top}\alpha + \gamma_i + \epsilon_{ti}, \quad j \in \{\text{resting}, \text{active}\}$$

  where $\mu$ is the intercept, $X_{ti}$ is our variable of interest, $\mathbf{Z}_{ti}$ is a set of potential confounders, and $\gamma_i$ is a subject-specific random intercept

- We condition the response variable $\hat{d}_{ti}$ on $j$, the value o f the decoded latent state
- $\hat{d}_{ti}$ is the estimated state duration on day $t$ under participant $i$ and state $j$
- $f(X_{ti}) = \beta \cdot X_{ti}$, where $X_{ti}$ is an indicator function for Bipolar status

**Results:**

| **Latent State** | $\beta$ | **p-value** | 95%-CI |
|---|---|---|---|
| **1 (Resting)** | 71.03 | 0.002 | $[27.64, 114.43]$ |
| **2 (Active)** | -69.83 | 0.005 | $[-116.34, -23.34]$ |

Table: $\beta$ estimated from LMM controlling for age, sex, race, and day of the week.

# Conclusion

**Chapter 3** - Decoding Resting and Active Patterns of Smartphone Use in Adolescents with Bipolar Disorder

- Decoded states provide a "first-order" approximation of sleep
- Several natural extensions:
  - Incorporate additional modalities (e.g., accelerometer/distance-metrics, screen-time) to model our observations
  - Consider heterogeneous dwell-times using covariates and/or random effects
  - Consider hierarchical hidden semi-Markov models
- **Future:** Build interventions using reinforcement learning
  - Latent states provide us with a direct and interpretable mechanism for representing observations

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

# Background

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

**Glioblastoma (GBM)** is a rare, fast-growing, and aggressive brain tumor

- Survival is poor — 5-year survival rate $< 5\%$
- Clinical trials evaluating new treatments for GBM rely on metrics that ignore the debilitating **functional and symptomatic impact** of GBM and its **associated treatment** on patients
- Data on patient functioning and quality of life in the glioblastoma (GBM) population in free living settings are limited
    - Infrequent and sparse collection of patient-reported outcomes
- Digital phenotyping provides a means to monitor patient treatment response and recovery in an objective, consistent manner, with **minimal patient burden**

**Objective**: Conduct a Digital Assessment in Neuro-Oncology (DANO) pilot by identifying key trends in post-operative recovery among GBM patients

# Study Cohort

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

Our study cohort consisted of 15 GBM patients (mean age $56.5; 46\%$ female)

- 12 were **primary GBM** (no history of previous low-grade tumor)
- 3 were **recurrent/transformed** (i.e., low grade glioma transformed in higher grade)
- Each received **surgical resection** of the tumor
- A **control group** of 30 non-operative spine patients with no history of brain cancer
  - (Mean age $54.2; 33.3\%$ female)

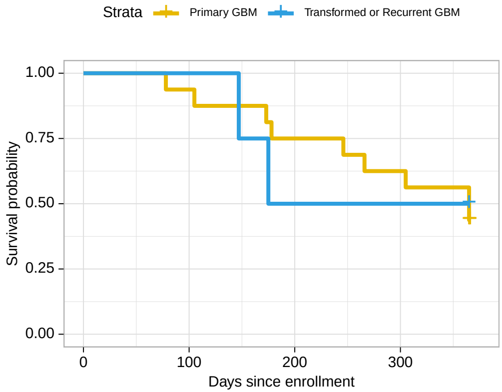



Figure: Kaplan-Meier survival curve from the start of patient enrollment to end of follow-up.

# Smartphone Data Collection

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

**Mobility-based Summary Statistics:**

| | | |
|---|---|---|
| **Distance Traveled** (km) | **Radius of Gyration** (km) | **Num. Significant Places Visited** |
| **Time Spent at Home** (hours) | **Max. Distance from Home** (km) | — |

Table: Subset of GPS-based summary statistics of digital phenotyping.

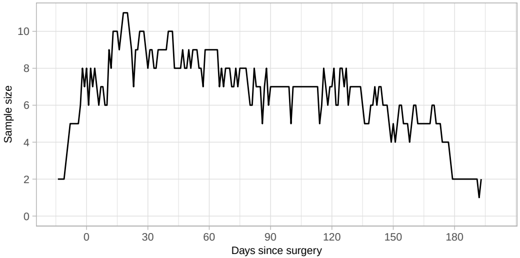
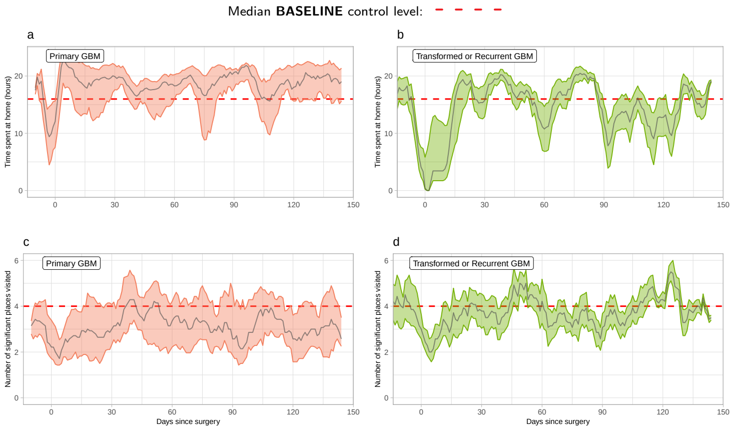



Figure: Sample size as a function of days since surgery. Number of patient-level GPS data available as a function of day since surgery. Average 108.9 days of GPS data per patient (range of 20 to 181).

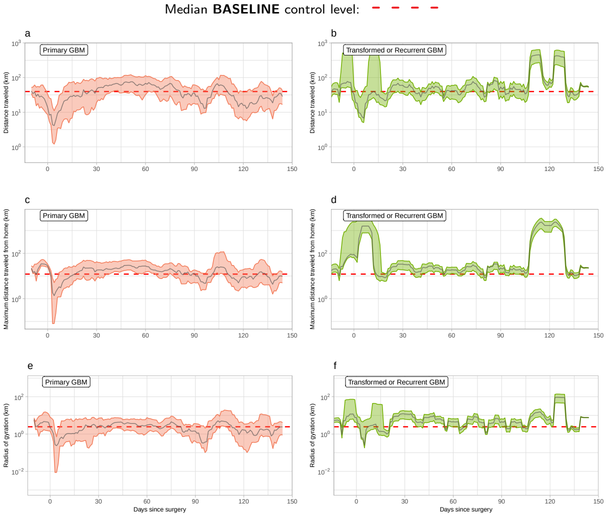
# Evaluating Changes in Mobility – Location-based Summary Statistics

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology



# Evaluating Changes in Mobility – Distance-based Summary Statistics

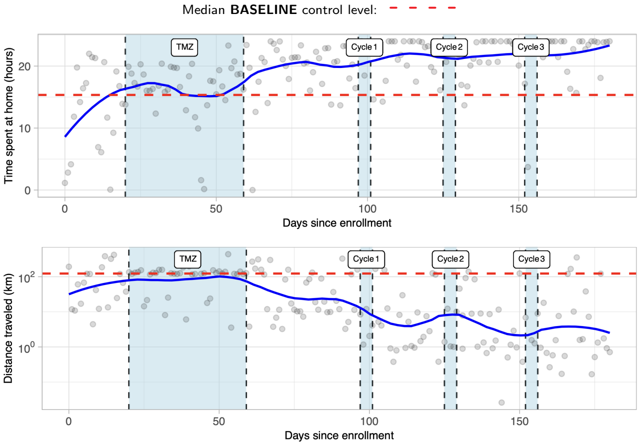
**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology



# Individual-level Changes in Mobility During Active Treatment

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

## TMZ ≡ Temozolomide



# Conclusion

**Chapter 4** - Assessing Mobility in Glioblastoma Patients using Digital Phenotyping – Piloting the Digital Assessment in Neuro-Oncology

- Digital phenotyping has the potential to allow for **quantification of patient behavior and recovery** without the need for active **patient involvement**
- In general, GBM patients appear to be **less mobile** than the control group
  - With largest dip in mobility occurring immediately after surgery
- Individual-level analysis revealed noticeable dips in mobility immediately following treatment
  - Potentially attributable to **accumulating symptom burden**
- **Future** – Determine whether inferences made using mobility-based assessments coincide with **patient-reported outcomes**

# Acknowledgments

# Acknowledgements


**Dissertation Committee:**

- Dr. Jukka-Pekka Onnela
- Dr. Junwei Lu
- Dr. Susan Murphy

**Collaborators:**

- Dr. Timothy R. Smith
- Dr. Anna Van Meter
- Dr. Matt W. Hoffman
- Dr. Nando de Freitas
- Dr. Abram L. Friesen
- Dr. Bobak Shahriari
- Dr. Francesca Siddi
- Dr. Michael Catalino
- John L. Kilgallon
- Noah L.A. Nawabi
- Dr. Jakob V.E. Gerstl
- Jacob G. Ellen
- Krish M. Maniar

**Lab Members:**

- Dr. Marcin Straczkiewicz
- Dr. Debbie Huang
- Dr. Marta Karas
- Dr. Anna Beukenhorst
- Hassan Y. Dawood
- Eli Jones
- Kenzie W. Carlson
- Greyson Liu
- Zach Clement
- Octavious Smiley
- Georgios Efstathiadis
- Dr. Hali Hambridge
- Dr. Thien Le
- Dr. Till Hoffman

**Biostatistics Dept.:**

- Dr. Paige L. Williams
- Jelena Follweiler
- Dr. Marcello Pagano
- Dr. Bethany Hedt-Gauthier
- Dr. John Quackenbush
- Dr. Sebastien Haneuse

# Acknowledgements

# Acknowledgements

# Next Steps

**2023-2024 Harvard Business School Blavatnik Fellow**

Harvard Business School

Blavatnik Fellowship in Life Science Entrepreneurship

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

raia

AI Supportive Care

**Thank you!**

# Appendix: Simulation Study

**Chapter 1:** Nonparametric Additive Value Functions – Interpretable Reinforcement Learning with an Application to Surgical Recovery

We examine our algorithm's performance in estimating the marginal nonlinear additive function $g_a(x)$

- We construct a **simulated MDP:**
  - a. **Initial state vector** $\mathbf{s}^{(0)} \in \mathbb{R}^d$ with each element sampled as $\mathbf{s}_i^{(0)} \in$ Unif$(\frac{1}{2}, \frac{1}{2})$
  - b. **Next state transition** occurs as $\mathbf{s}^{(t)} \sim \mathcal{N}(\mathbf{s}^{(t-1)} + \delta_a, 0.1)$, where $\delta_a = 0.1 \times \mathbb{1}(a^{(t)} = 0) - 0.1 \times \mathbb{1}(a^{(t)} = 1)$
  - c. **Additive reward function** defined as $r(\mathbf{s}, a) = u_1(\mathbf{s}_1, a) + u_2(\mathbf{s}_2, a)$
- Perform Monte-Carlo simulation to estimate the action-value function associated with the random policy $\pi_e(s^{(t)}) \in \{0, 1\}$ with $p = 0.5$ and a discount factor of $\gamma = 0.5$

$$
\begin{aligned}
Q^{\pi_e}(\mathbf{s}, a) &= \mathbb{E}_{\pi_e}\left[\sum_{i=0}^{\infty} \gamma^i r(\mathbf{s}^{(i)}, a^{(i)}) \mid \mathbf{s}^{(0)} = \mathbf{s}, a^{(0)} = a\right] \\
&= \sum_{j=1}^{2} \mathbb{E}_{\pi_e}\left[\sum_{i=0}^{\infty} \gamma^i u_j(\mathbf{s}_j^{(i)}, a^{(i)}) \mid \mathbf{s}^{(0)} = \mathbf{s}, a^{(0)} = a\right] \\
&= U_1(\mathbf{s}_1, a) + U_2(\mathbf{s}_2, a).
\end{aligned}
$$