Applications of Reinforcement Learning in Criminal Justice and Healthcare

Pengyi Shi

Purdue University

About me

Pengyi Shi

Joined Purdue in January 2014

Ph.D. in Industrial Engineering (Georgia Tech)

Research area:

- Healthcare operations, service operations, queueing and stochastic modeling
- Integrate data analytics into decision making (reinforcement learning, online learning)

Agenda

- Reinforcement learning in queueing network control
  - Jail diversion
    - Formulate model to apply RL
  - Hospital unit placement
    - Leveraging queueing structure in RL

Team 1

Academic Partners

Xiaoquan Gao
Purdue -> SMU

Griffin Carter
AAE

Nan Kong
BME

Nicole Adams
Nursing

PURDUE UNIVERSITY

REGENSTRIEF CENTER FOR HEALTHCARE ENGINEERING

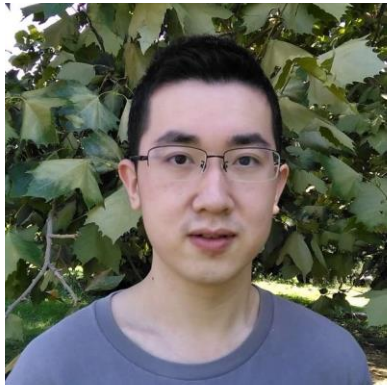
Community Partners

Jason Huber
Director of TCCC

Robert Goldsmith
Sheriff

Team 2

Academic Partners

Amy R. Ward **Zhiqiang Zhang** **Bingxuan Li** **Chuwen Zhang**

Booth School of Business (Purdue -> UCLA) Purdue

Rustandy Center

Community Partners

Government Agency Collaborator in IL

State-wide community-based alternatives to incarceration and improve access to interventions that reduce crime

Data from **Illinois Criminal Justice Information Authority**

ICJIA

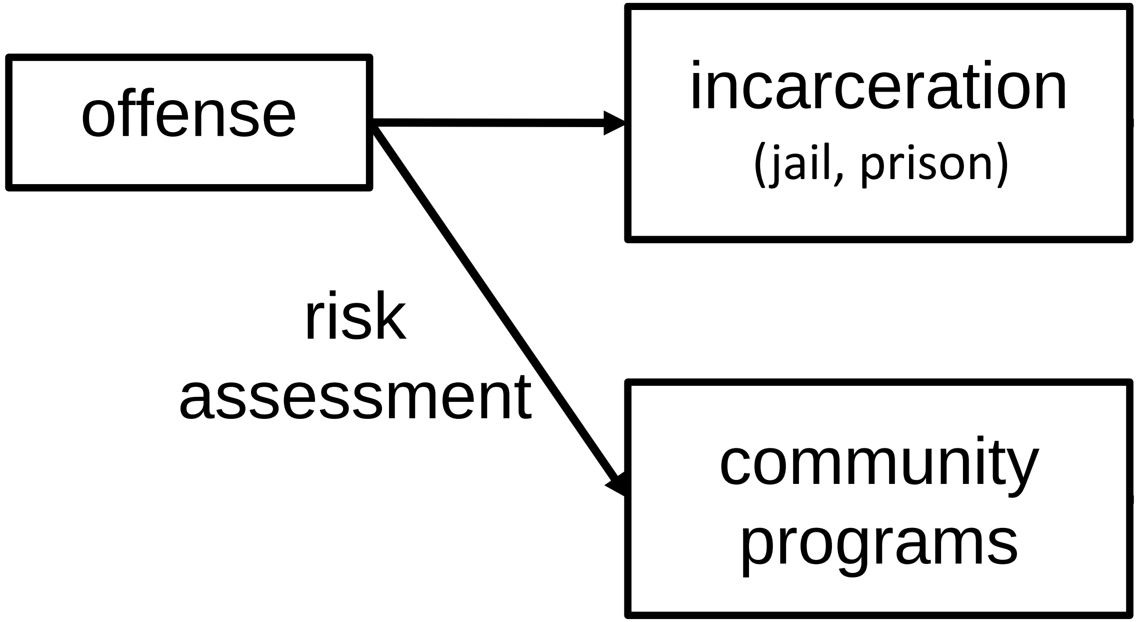
Overcrowding in the Correctional Systems

- Correctional facilities overcrowded
  - 2/3 of jail population have drug-related offences
  - Chronic disease

- Alternatives: Community-based programs
  - Reintegration
  - Treatment – medical and therapy
  - Education, life-skill training

- Challenges: Capacity is limited

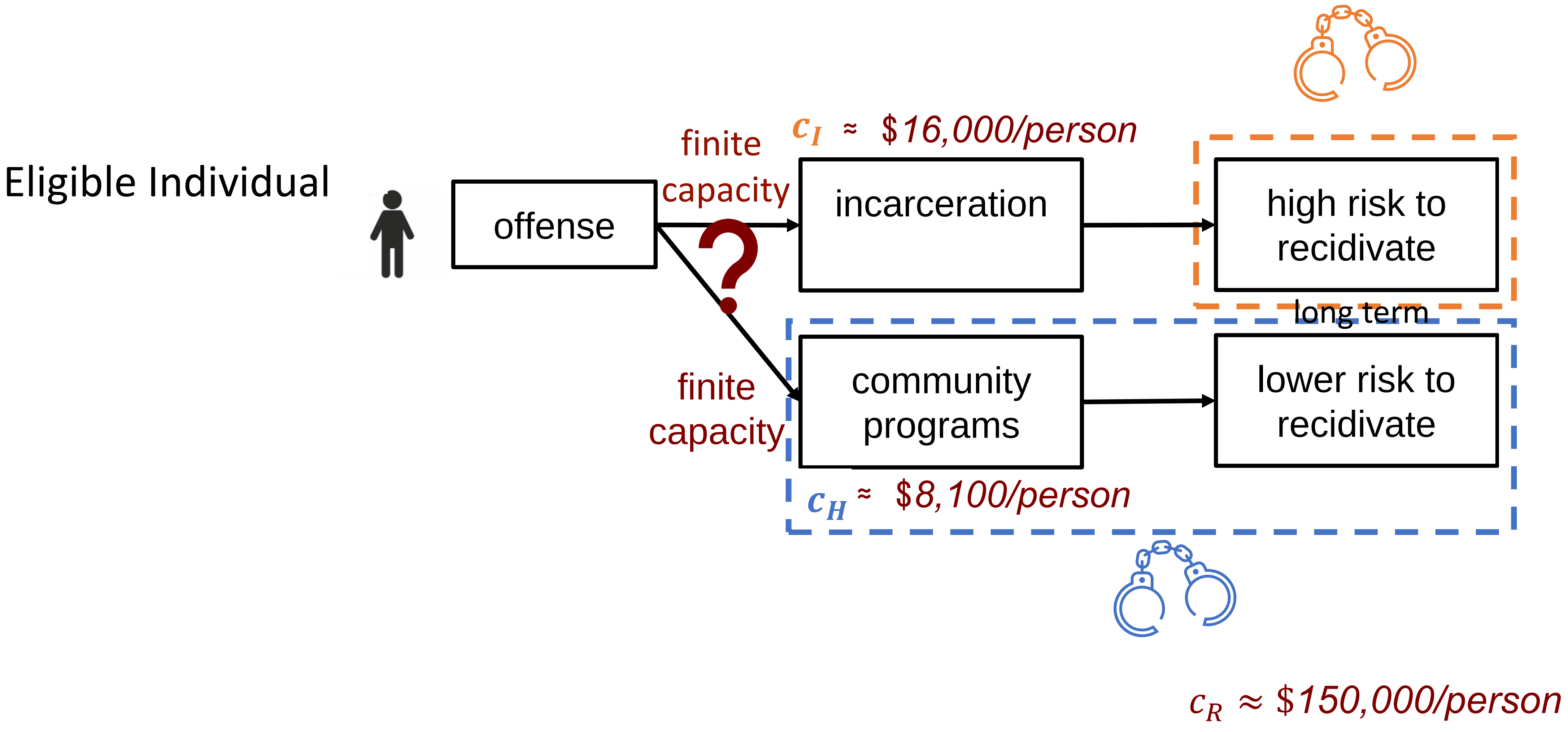
Background: Process Flow



8% and 32% for cognitive behavioral programs
30% for substance abuse treatment programs
20% for education and employment programs

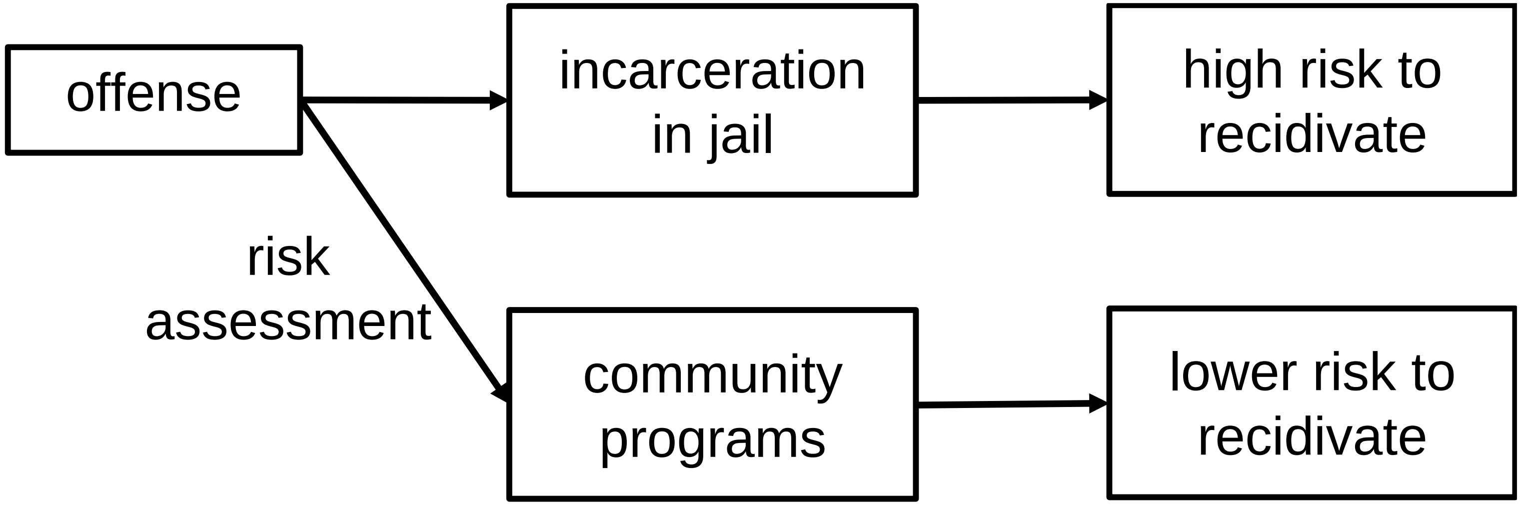
LSI-R: identify an individual's risks and needs with regard to recidivism.
https://cech.uc.edu/about/centers/ucci/products/assessments.html

Tradeoffs



Eligible Individual
offense
finite capacity
$c_I \approx$ $16,000/person
incarceration
high risk to recidivate
long term
finite capacity
community programs
lower risk to recidivate
$c_H \approx$ $8,100/person
$c_R \approx$ $150,000/person

Prescriptive Program Placement:
Problem Formulation – MDP



- State $X_{m,j,d}$ number of clients – [class (risk type), facility (jail/CC), **LOS**]
- Decision: routing $A_{m,j}$ – [class (risk type), facility (jail/CC)]
- Cost function: convex occupancy cost + violation cost + recidivism cost

Prescriptive Program Placement Overview

High-dimensional MDP

Structural management insights

Convexity on occupancy

Gao, Shi, Kong (2023) "Stopping the Revolving Door: MDP-Based Decision Support for Community Corrections Placement." Major Revision in *Operations Research*.

Structural Property

Main Result: Superconvexity

THEOREM 1. *Under Assumption 1 and some mild technical condition, the optimal value function* V^* *satisfies* $SuperC(e_{jail,l}, e_{CC,l})$, *i.e., for all* $s \in \mathcal{S}$ *and* $l = 0, 1, ..., \min\{d_{jail}, d_{CC}\}$,

$$V^*(s+2e_{CC,l}) - V^*(s+e_{CC,l}) \geq V^*(s+e_{jail,l}+e_{CC,l}) - V^*(s+e_{jail,l}), \tag{6}$$

$$V^*(s+2e_{jail,l}) - V^*(s+e_{jail,l}) \geq V^*(s+e_{jail,l}+e_{CC,l}) - V^*(s+e_{CC,l}). \tag{7}$$

Cost decomposition + Policy deviation bounding + Coupling → Value function comparison

Implication 1: one optimum → Policy Gradient Algorithm with Theorem 1 as theoretical support

Implication 2: The optimal policy has a "**switch curve**" structure

Policy Gradient Algorithm

Leverage switch-curve structure to enhance learning

Algorithm 1: Tabular batched actor-critic policy gradient

Input : Step sizes $\alpha_\theta, \alpha_\omega$. Batch size N. Number of iterations T.

Output: Value function $V(\tilde{s}), \tilde{s} \in \tilde{\mathcal{S}}$.

```
Initialize {θ_{j,m}(s̃)}_{j,m}, s̃ ∈ S̃ at random, V(s̃) = 0, s̃ ∈ S̃. Initialize state s̃_1 at random.
for t = 1, 2, ..., T do
    for n = 1, 2, ..., N do
        Set current state s̃_t.
        Sample and store the placement of new arrivals ã_n ~ π_Θ(ã|s̃_t) and the next state s̃'_n.
    end
    Update the policy parameters:
```

$$\boxed{\theta_{j,m} \leftarrow \theta_{j,m} - \alpha_\theta V(\tilde{s}_t) \cdot \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta_{j,m}} \ln \pi_\Theta(\tilde{a}_n | \tilde{s}_t),}$$

Enforce switching curve structure

```
    Update the value function with TD(0):
```

$$V(\tilde{s}_t) \leftarrow V(\tilde{s}_t) + \alpha_\omega \left(\frac{1}{N} \sum_{n=1}^{N} \left(C(\tilde{s}_t) + \gamma \cdot V(\tilde{s}'_n) \right) - V(\tilde{s}_t) \right). \tag{10}$$

```
    Sample the next states s̃_{t+1} ~ P(s̃'|s̃_t, ã_N).
end
```

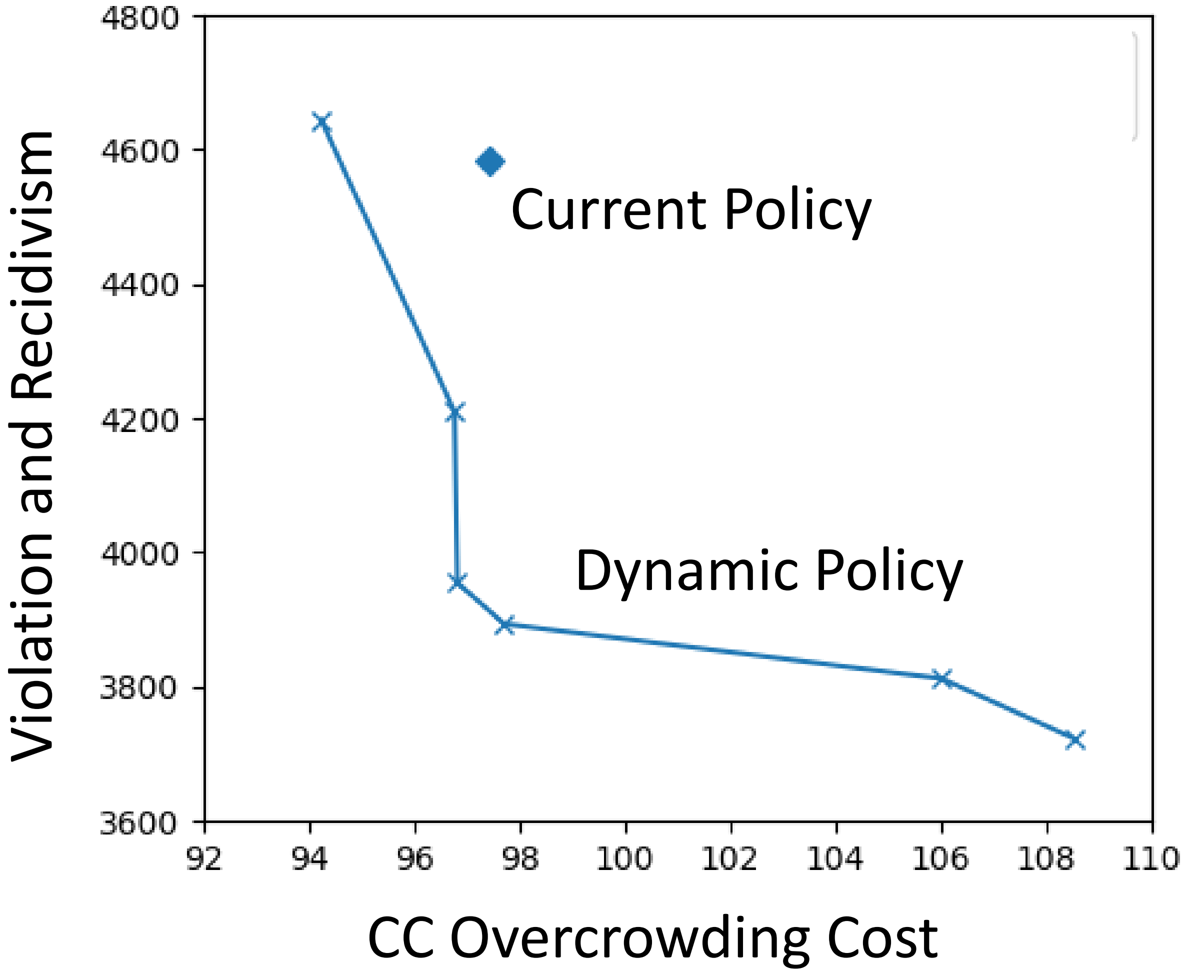
Case Study: Data

- Tippecanoe County Community Corrections data
  - Individual-level data: demographics, criminal history, programs, appointments
  - ~56,000 records in 2010 - 2019

- Jail data
  - Population-level data: demographics, jail admissions, arrests, re-arrests
  - Tippecanoe county, monthly summary 2015 - 19

Test on Historical Data

- Efficiency frontier over different cost parameters



Leadership Buy-in: Capacity Planning

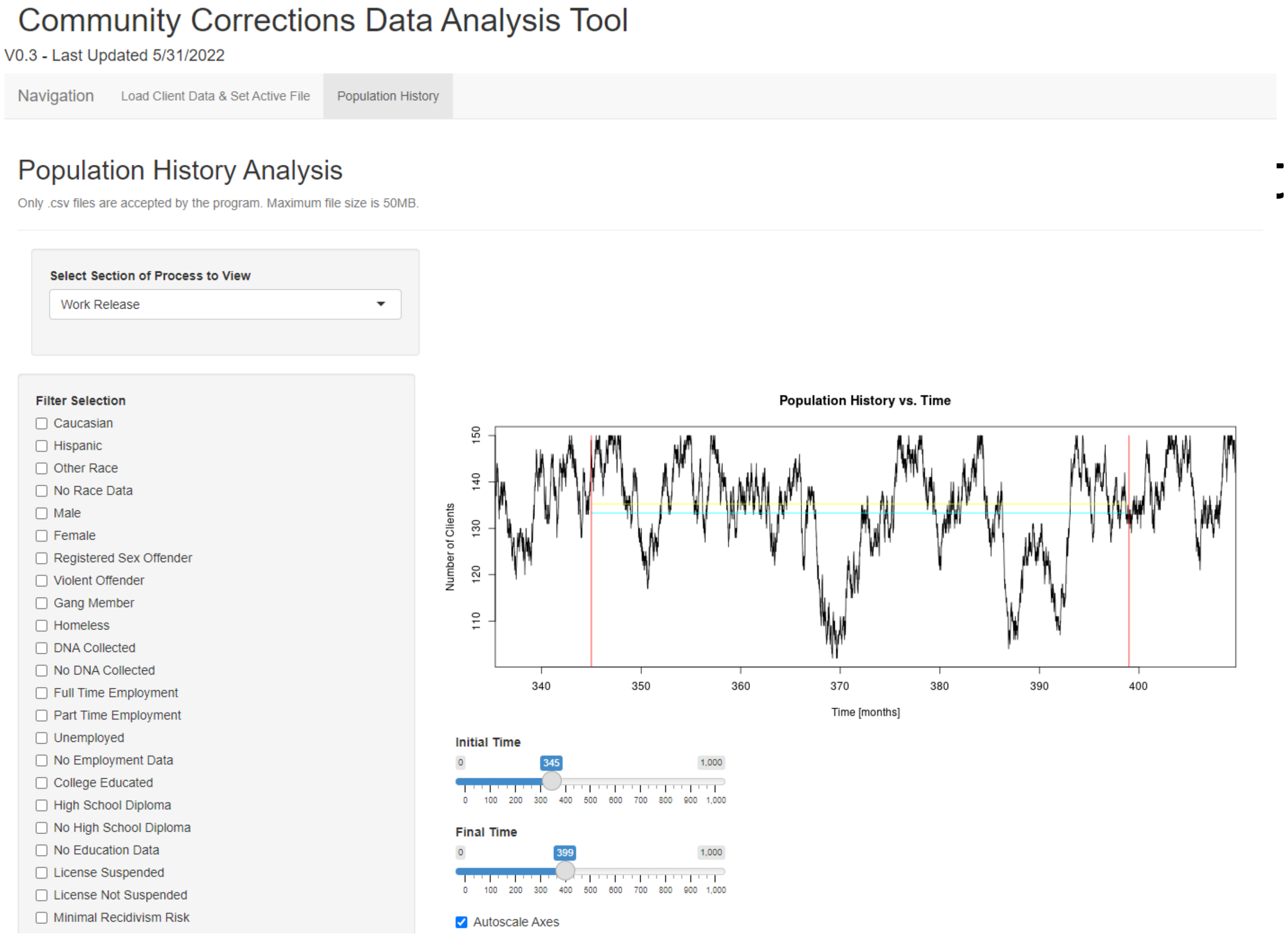
Recruit 4 more case managers: Reduce recidivism and violations in 3-yr window by 15%-38%.

Cost: Personnel cost (salary) +Variable cost (From increased HD population) $= \$2{,}373{,}900$

Benefit: Jail congestion mitigation + Recidivism reduction $= \$40{,}657{,}524$

- Results presented by Director of TCCC at townhall meeting for budget
- ***Sustainable*** workforce via better workload plan and reduced burnout

Ongoing Work – Interpretable



ions

Co-PI:
Nicole Adams
Nursing, Purdue

RA:
Griffin Carter
AAE, Purdue

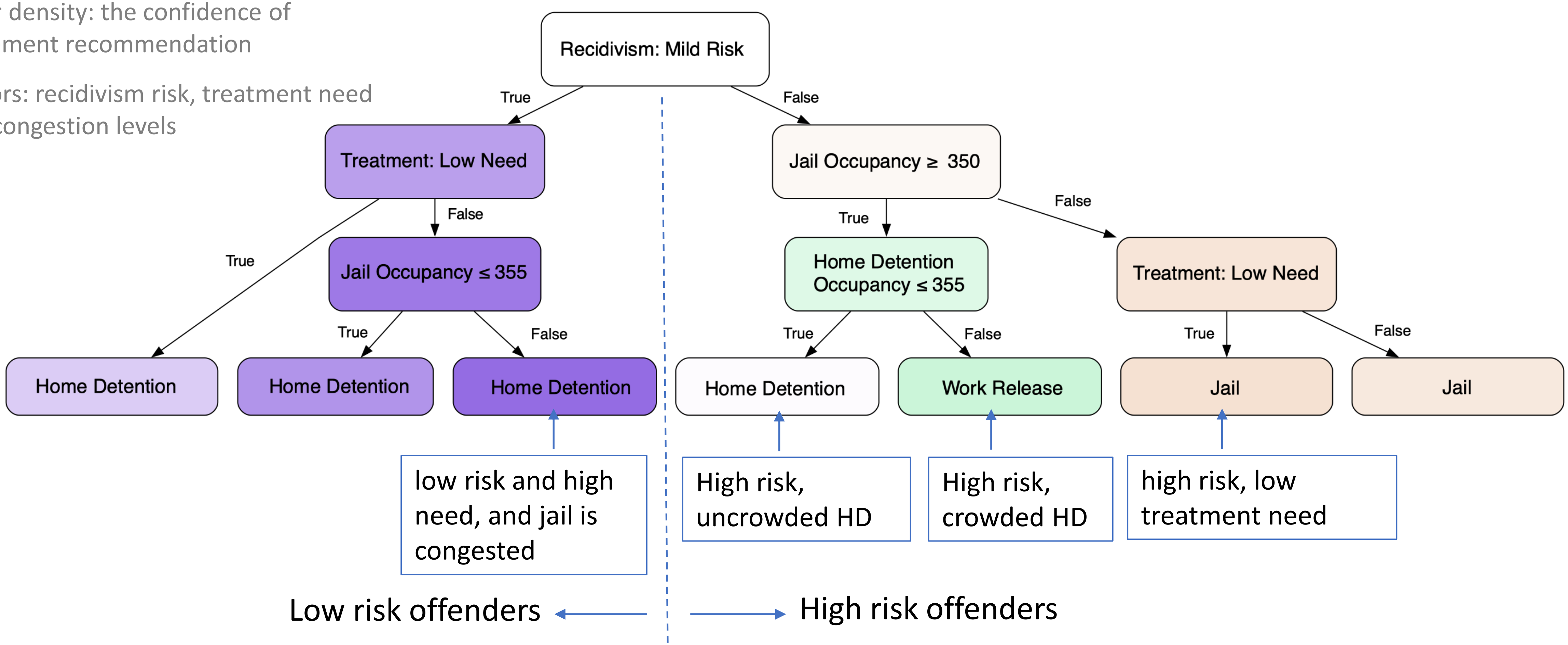
"A Community Approach for Racial Justice with Data-driven Analytics." 2021 Engagement Scholarship Research/Creative Activities Grant (**PI: Shi**)

Interpretable Placement Decision

Decision tree extracted from the RL-based policy to help understand the placement recommendation

Color density: the confidence of placement recommendation

Factors: recidivism risk, treatment need and congestion levels

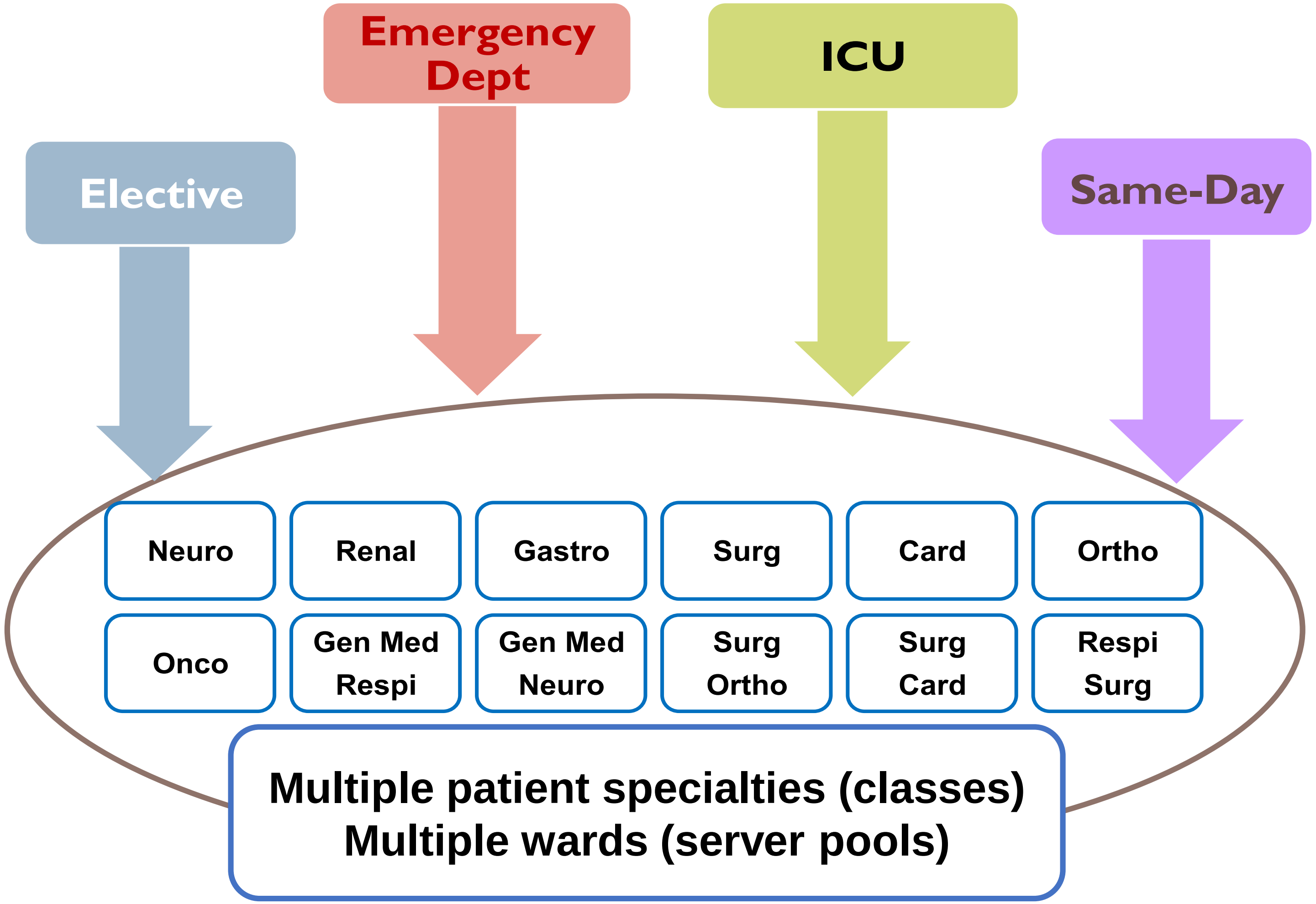


Agenda

- Reinforcement learning in queueing network control
  - Jail diversion
    - Formulate model to apply RL
  - Hospital unit placement
    - Leveraging queueing structure in RL

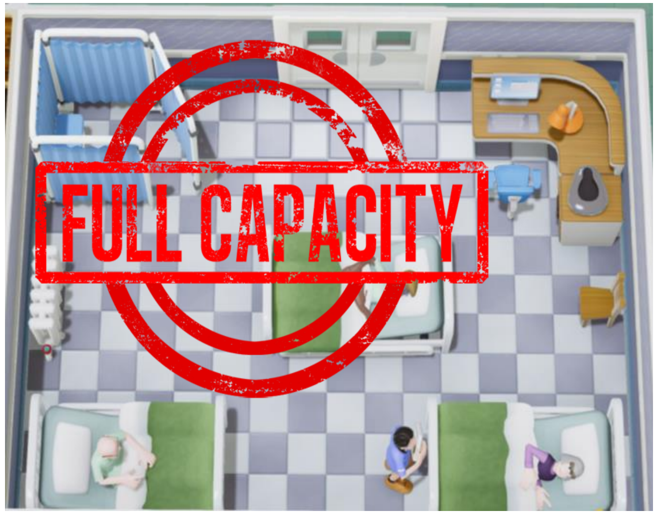
Background: Hospital Inpatient Network

- Different inpatient units
- Different types of patients



Overflow (Off-service Placement)

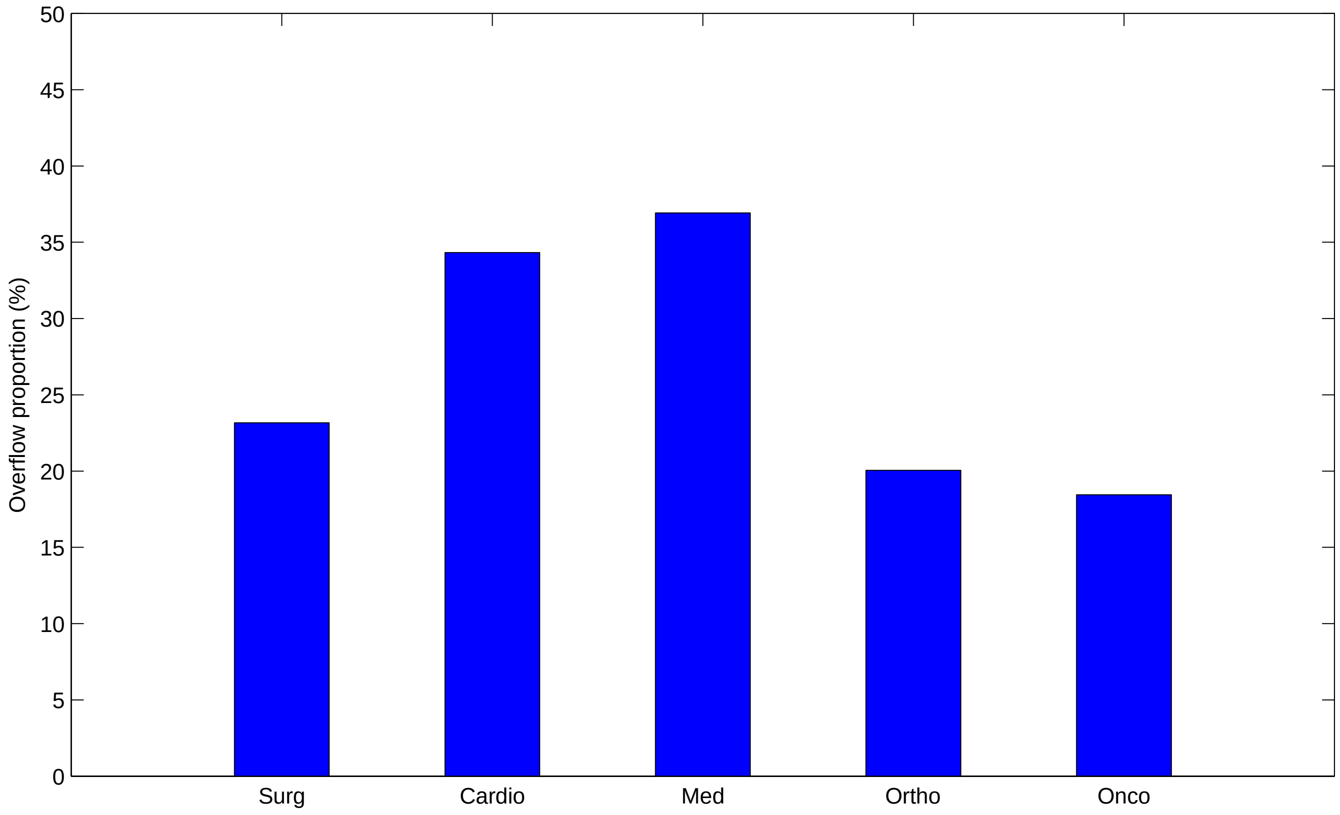
Primary ward



Non-primary ward

Overflow to Reduce ED Crowding

- Overflow 20-30% Emergency Department (ED) patients



[1] Shi et al. (2016); Song et al. (2019)

Tradeoff between Waiting and Overflow

- Helps reduce waiting time and alleviate congestion temporarily
  - Resource pooling

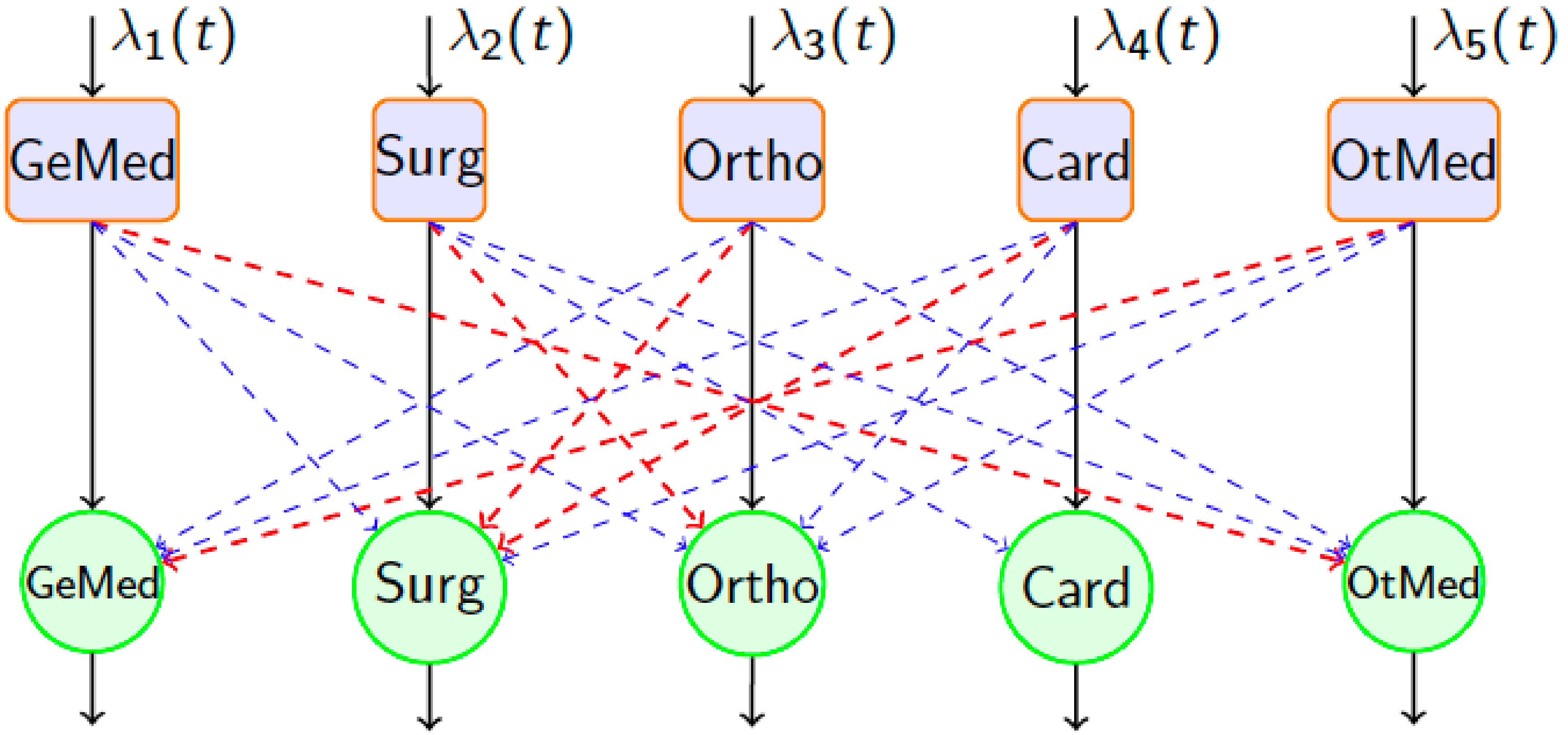
- Not desirable
  - Compromise quality of care (Song et al. 2019)
  - More coordination (Rabin et al. 2012)
  - Overflow patients occupy capacity: “snowball effect” (Dong-Shi-Zheng-Jin 2020)

  https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3306853

- Overflow decisions
  - Overflow the patient now or wait for another hour?
  - Overflow to which wards?
    - Medically closeness, distance to primary wards, ward occupancy, etc

A Five-pool Example

- Challenging to solve with conventional MDP methods
  - Large state space (~10^{14}) and action space
  - Time-dependent arrival and discharge patterns



Modeling overflow decisions

- State at decision epoch t_k

$$S(t_k) = (\underbrace{X_1(t_k), \ldots, X_J(t_k)}, Y_1(t_k), \ldots, Y_J(t_k), \underbrace{h(t_k)})$$

Patient count of each pool j at t_k

To-be-discharge count from pool j

Time-of-day indicator

Modeling Overflow Decisions

- Action

$$f(t_k) = \{f_{ij}(t_k),\ i \neq j,\ i = 1, \ldots, I,\ j = 1, \ldots, J\}$$

- One period cost

$$g\big(S(t_k), f(t_k)\big) = \boxed{\sum_{i=1}^{I} \sum_{j \neq i, j=1}^{J} B_{ij} \cdot f_{ij}(t_k)} + \boxed{\sum_{i=1}^{I} C_i \cdot Q_i(t_k+)}$$

- Minimize long-run average cost = $\boxed{\text{overflow cost}}$ + $\boxed{\text{holding cost}}$

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left(\sum_{k=1}^{n} g\big(S(t_k), f(t_k)\big) \right)$$

Exact Analysis

- Bellman Equation

$$\gamma^* + \nu^*(s) = \min_f \left\{ g(s, f) + \boxed{\sum_{s'} p(s'|s, f)\nu^*(s')} \right\}, \quad s \in \mathcal{S}$$

Cost-to-go

  - *v(s)*: value function for state *s*
  - Large state space $O(X^J Y^J) \sim 10^{14}$ and action space
    - Value iteration or policy iteration becomes infeasible

Tackling the Curse-of-dimensionality

- Large state space
  - Value function approximation + queueing structure

- Large action space
  - Original setup: combinatorial in matching
  - Atomic action: decomposing action into individual level
  - Policy gradient (PPO)
    - Time-varying long-run average setting

Jingjing Sun
CUHK-Shenzhen

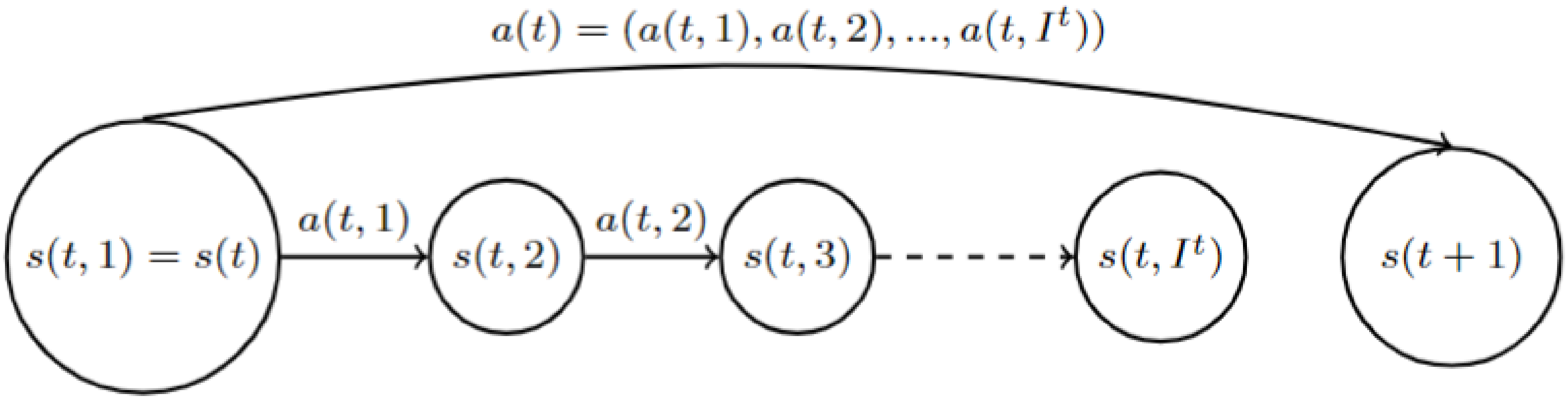
Jim Dai
ORIE, Cornell

Sun, Dai, Shi (2024) "Inpatient Overflow Management with Proximal Policy Optimization."
https://arxiv.org/abs/2410.13767

Atomic Action

- Proximal Policy Optimization (PPO) Method
- Randomized policy: $\pi(a)$

- Macro- and micro-decision process



Algorithm

- In iteration i:
  - Policy evaluation: value function approximation* $v_{\pi_\eta}(t,s)$
  - Minimize loss function w.r.t. θ

$$\sum_{k=0}^{N-1} \max \left[\left(\boxed{\prod_{i=0}^{I_{X(t_k)}-1} \frac{\pi_\theta(a(t_k,i)|t_k,s(t_k,i))}{\pi_\eta(a(t_k,i)|t_k,s(t_k,i))}} \right) \boxed{\hat{A}_\eta(t_k,s(t_k),f(t_k)),} \right.$$
$$\left. \left(\prod_{i=0}^{I_{X(t_k)}-1} clip\left(\frac{\pi_\theta(a(t_k,i)|t_k,s(t_k,i))}{\pi_\eta(a(t_k,i)|t_k,s(t_k,i))}, 1-\epsilon, 1+\epsilon \right) \right) \hat{A}_\eta(t_k,s(t_k),f(t_k)) \right]$$

With

$$A_\pi(t,s,f) = c(s,f) - \bar{c}_\pi + \sum_{y \in S} P(t,s,f,y) h_\pi(t+1,y) - v_\pi(t,s)$$

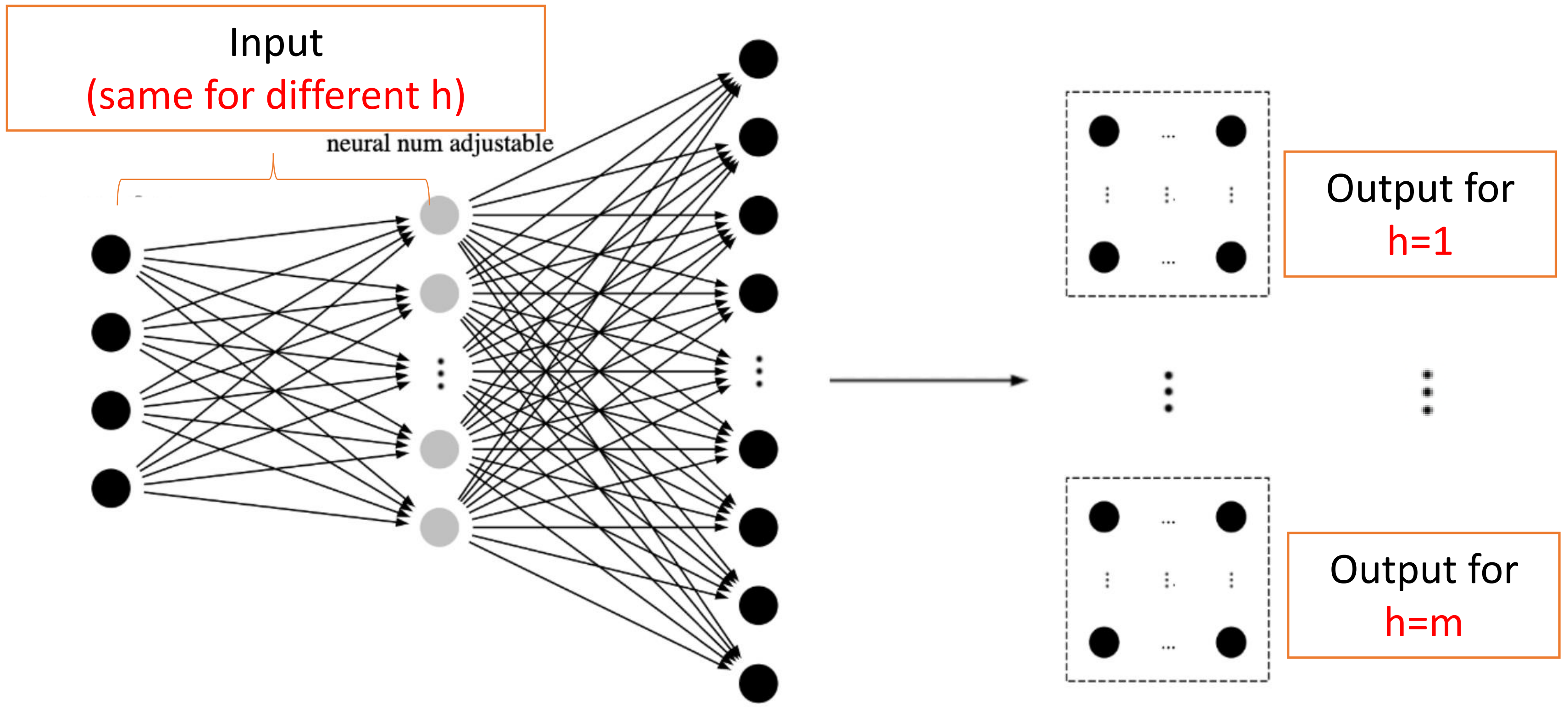
  - Update policy (parameterized NN)

Policy ratio

*Value function approximation using queueing structure, based on pool-wise decomposition

Policy Representation

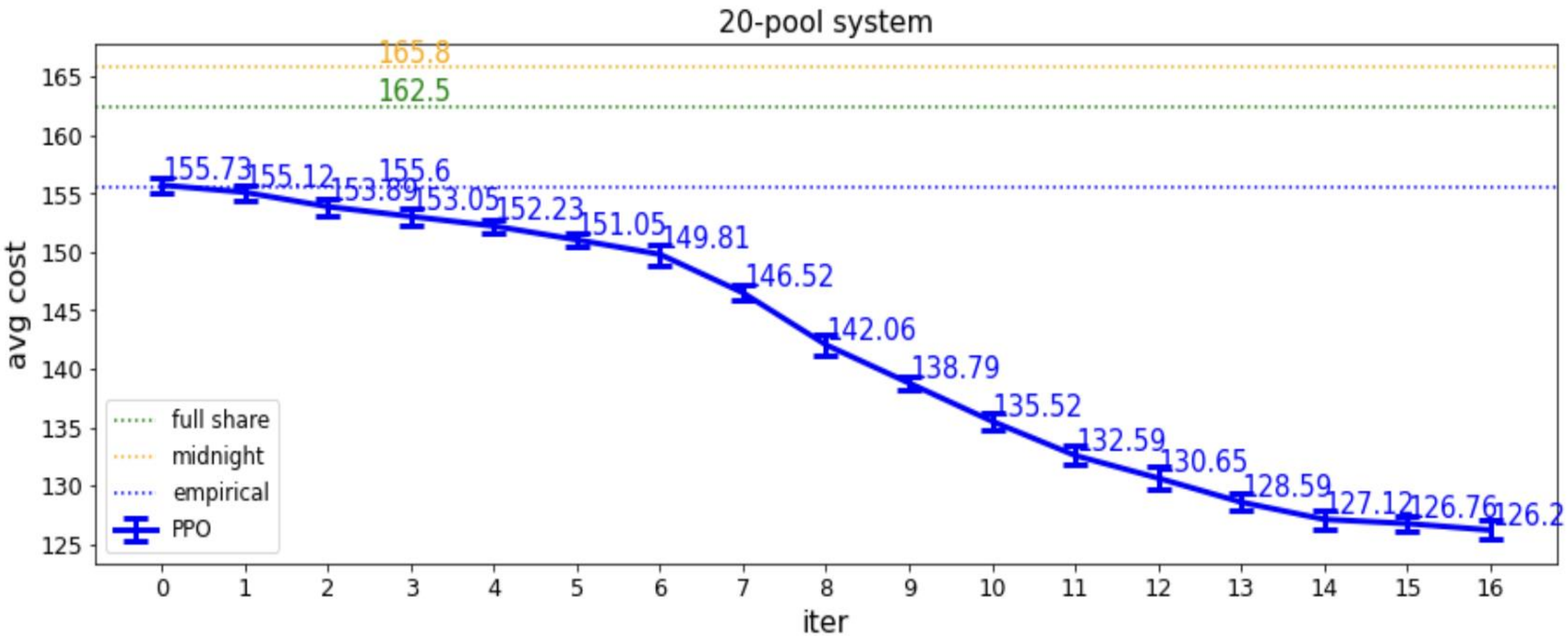
Partially-connected structure



- Combine the strengths of two intuitive designs (fully-separated or fully-connected)

Empirical Success

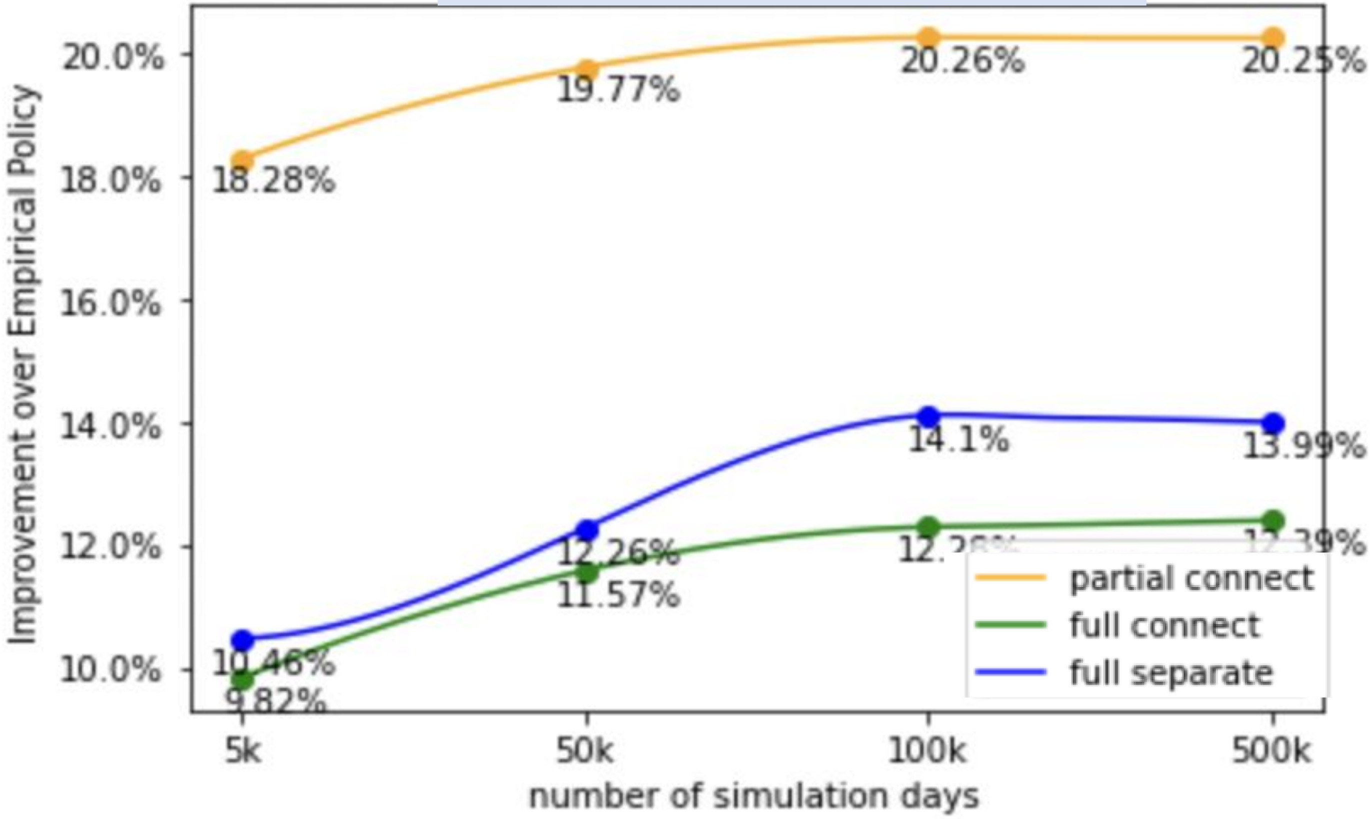
- Scalable algorithm: 20-pool network



Importance of Policy NN Design

- Smaller sample can perform well under right design

Simulation Data Amount



Main Takeaway and Future Research

- Complicated tradeoff
  - Proper modeling and domain knowledge
  - Exciting area

- Future directions
  - Short-term vs long-term fairness (with Chuwen Zhang, Amy Ward)
  - Safe online learning?

Questions?

Email: shi178@purdue.edu