# Adaptation, Optimality, and Synthesis

UF UNIVERSITY of FLORIDA

Duke UNIVERSITY

TEXAS The University of Texas at Austin

UC SANTA CRUZ

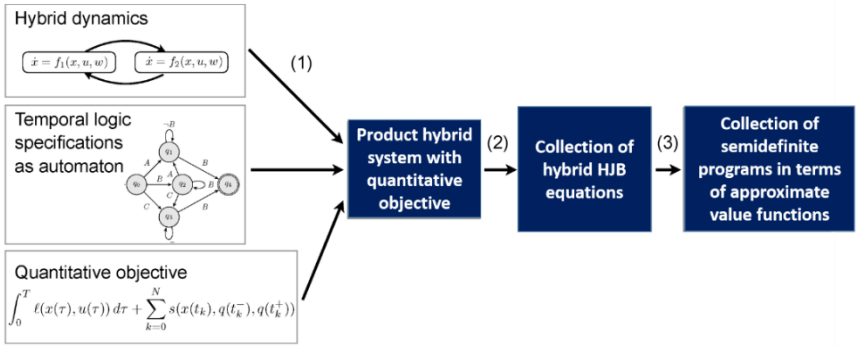
# Adaptation, Optimality, and Synthesis

- Approximately optimal control methods for forward and inverse decision-making problems

- Real-time optimal control methods that can handle uncertainty, complex mission specifications, and rely on sophisticated approximation, learning, and sampling techniques to enhance scalability (avoid explicit discretization of continuous dynamics)

**RT2 will establish new strategies for the development of approximately optimal control methods for continuous and stochastic hybrid systems for forward and inverse decision-making problems under complex mission specifications**

- Tractable optimal control methods under complex mission specifications captured by temporal logic (TL) formulas, and extend them to systems with unknown uncertainties and run-time computational limitations



# Adaptation, Optimality, and Synthesis

- Temporal Logic (TL) Planning and Learning
  - Scalable TL robot planning
  - Abstraction-free TL robot planning
  - Transfer planning for TL tasks
  - Transfer learning with unobserved contextual information
- Approximate Dynamic Programming (ADP) Methods
  - Improved asymptotic performance under ime varying parameters
  - “Safe” (Barrier function) Reinforcement Learning (RL) methods for Approximate Dynamic Programming (ADP)
  - Emerging results on Switched ADP methods
- Eliminate the use of high-accuracy orbit determination to estimate physical parameters of unknown targets.
  - Adaptive control to compensate for unknown physical parameters.
  - Regulation of underactuated system using a single control input.

# Planning and Learning under High-Level Temporal Tasks and Unknown Context

Michael M. Zavlanos
Mechanical Engineering & Materials Science
Electrical & Computer Engineering
Computer Science
Duke University

Assured Autonomy in Contested Environments (AACE)
Spring 2020 Review
April 14, 2020

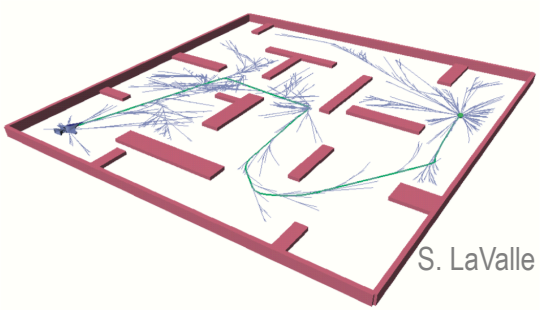
UF UNIVERSITY of FLORIDA
Duke UNIVERSITY
TEXAS The University of Texas at Austin
UC SANTA CRUZ

# Robot Motion Planning

## Point-to-point navigation tasks

- "Starting from point A, reach point B while avoiding obstacles"

L. Kavraki et al (TRA 1996), S. LaValle et al (IJJR 2001),
S. Karaman et al (IJJR 2011), L. Janson (IJRR 2015)

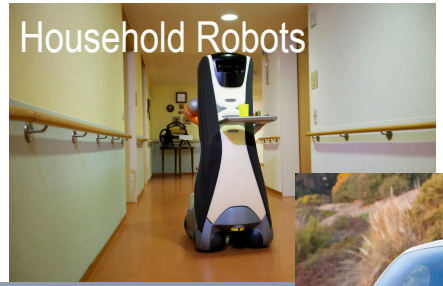


S. LaValle et al (IJJR 2001),

## High-level complex tasks

- "Pick up the mail by visiting houses in a given order"
- "Next visit a delivery site"
- "Never leave the delivery site until a ground robot is present to pick up the mail"
- "Repeat this process every day"

How to express complex tasks in a formal way?
How to synthesize optimal and correct-by-construction controllers?

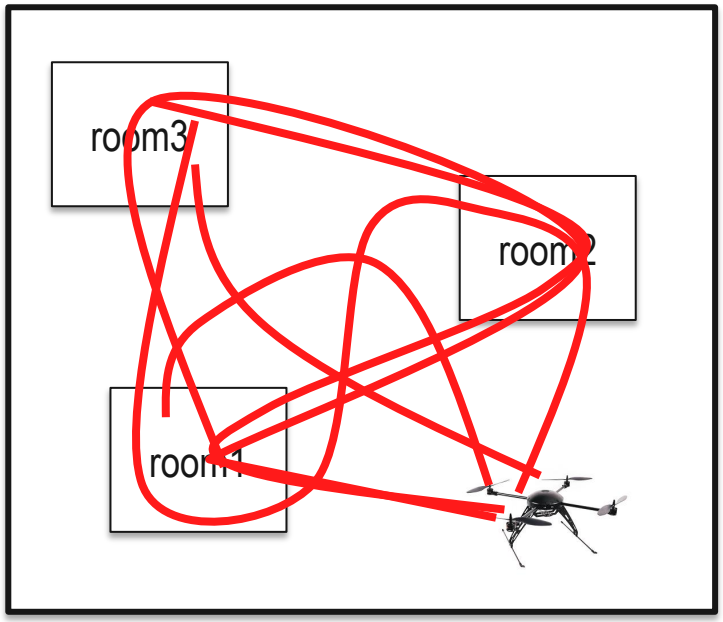





M. Kloetzer et al (TRO 2010), S. Smith et al (IJRR 2011)
A. Ulusoy et al (IJRR 2013), M. Guo et al (IJRR 2015)

Duke UNIVERSITY

# Expressing Complex Tasks using Linear Temporal Logic (LTL)




Reachability task $\Diamond \pi_i^{\text{room1}}$

Reachability with avoidance $\neg(\pi_i^{\text{room1}} \vee \pi_i^{\text{room2}}) \mathcal{U} \pi_i^{\text{room3}}$

Coverage task $\Diamond \pi_i^{\text{room1}} \wedge \Diamond \pi_i^{\text{room2}} \wedge \Diamond \pi_i^{\text{room3}}$

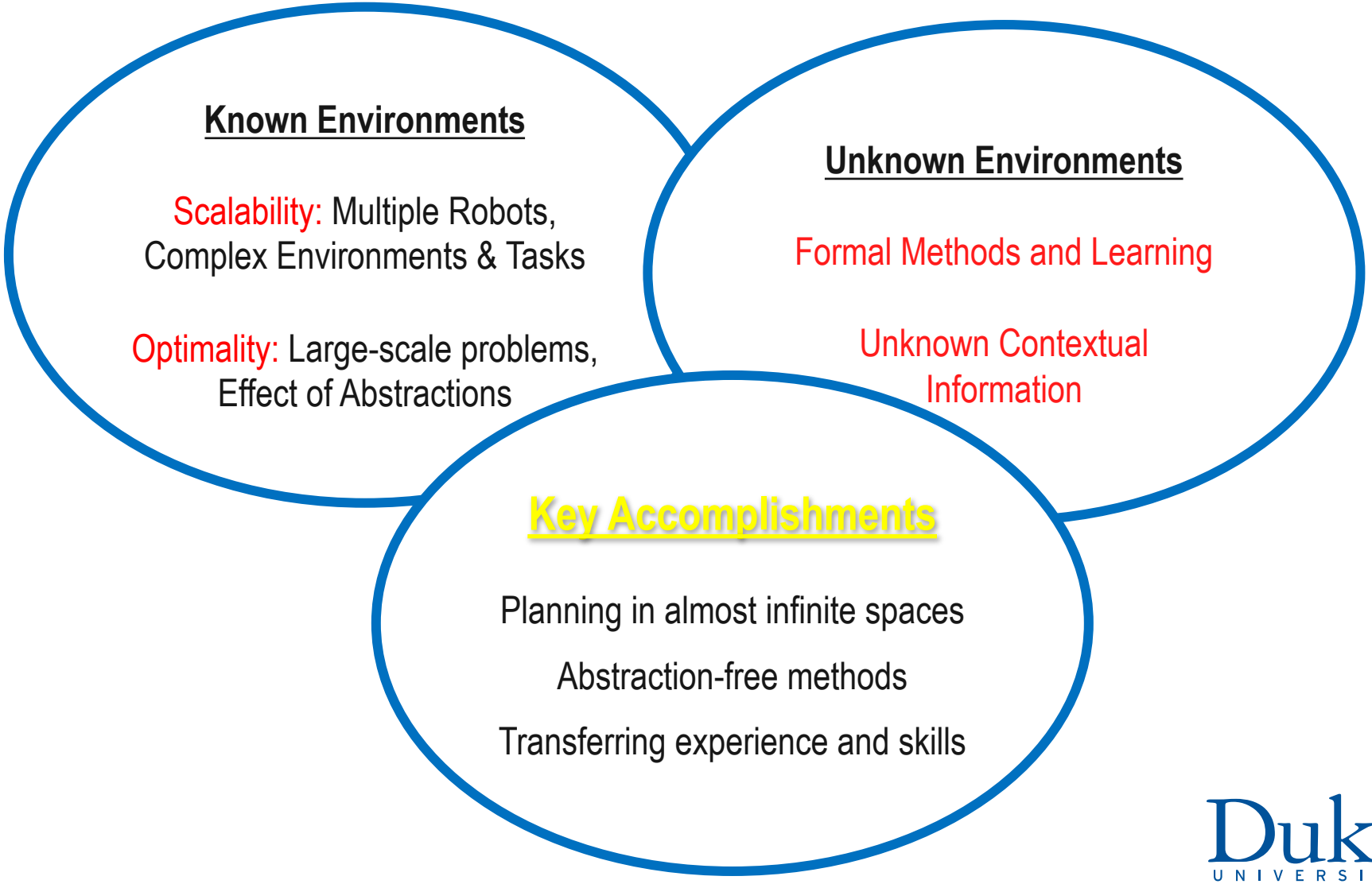
Sequencing $\Diamond(\pi_i^{\text{room1}} \wedge (\Diamond(\pi_i^{\text{room2}} \wedge \Diamond \pi_i^{\text{room3}})))$

Recurrent sequencing $\Box\Diamond(\pi_i^{\text{room1}} \wedge (\Diamond(\pi_i^{\text{room2}} \wedge \Diamond \pi_i^{\text{room3}})))$

Compositional tasks: $\phi = \underbrace{\Box\Diamond(\pi_1^{\text{room1}})}_{\text{Robot 1: visit room1 infinitely often}} \wedge \underbrace{(\neg\pi_1^{\text{room1}} \mathcal{U} \pi_2^{\text{room2}})}_{\text{Robot 1: never visit room1 until robot 2 visits room 2}} \wedge \underbrace{(\Diamond\Box(\pi_2^{\text{room3}}))}_{\text{Robot 2: eventually always visit room 3}}$

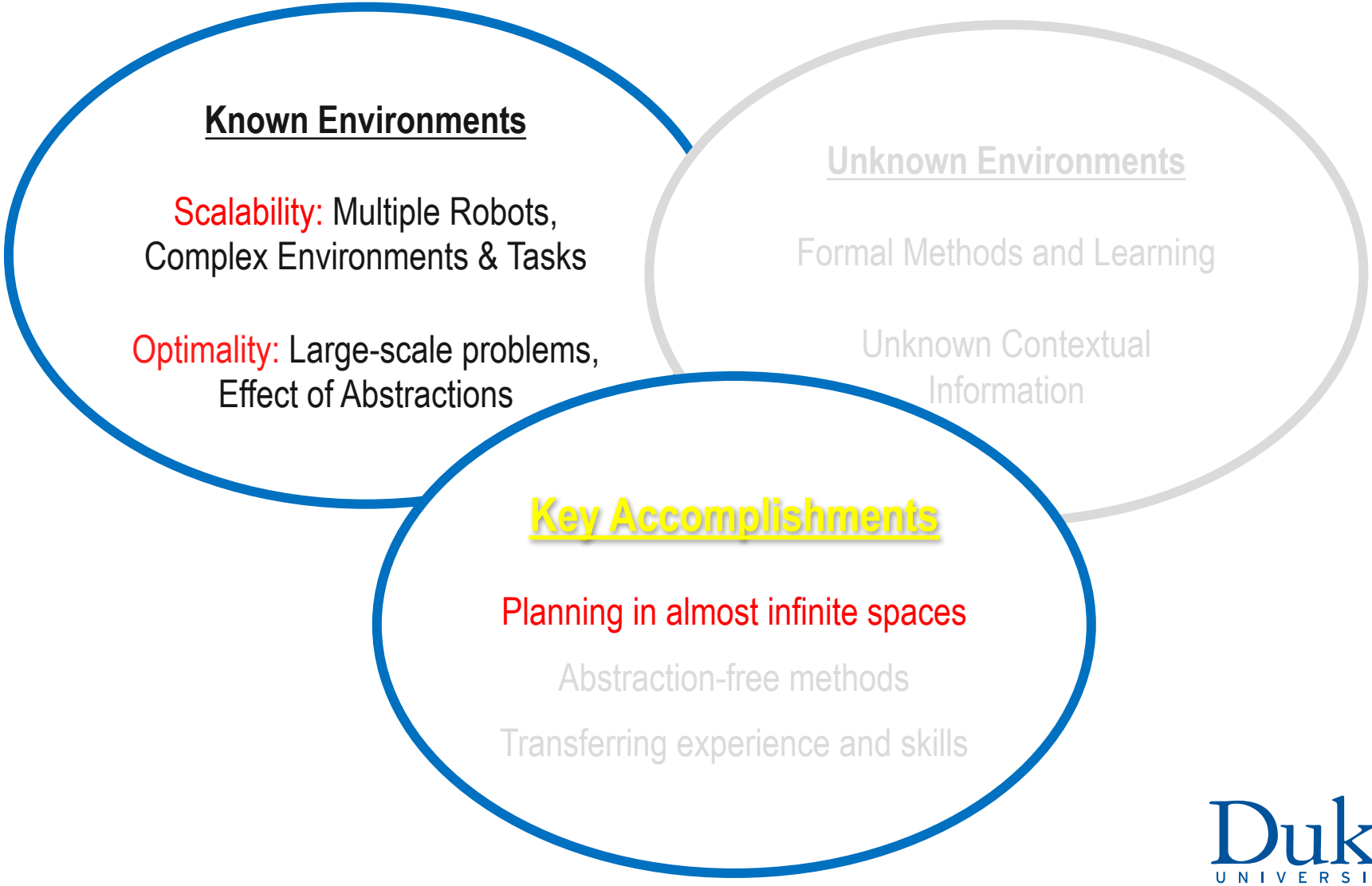
Duke UNIVERSITY

# Challenges & Key Accomplishments




Duke UNIVERSITY

# Challenges & Key Accomplishments

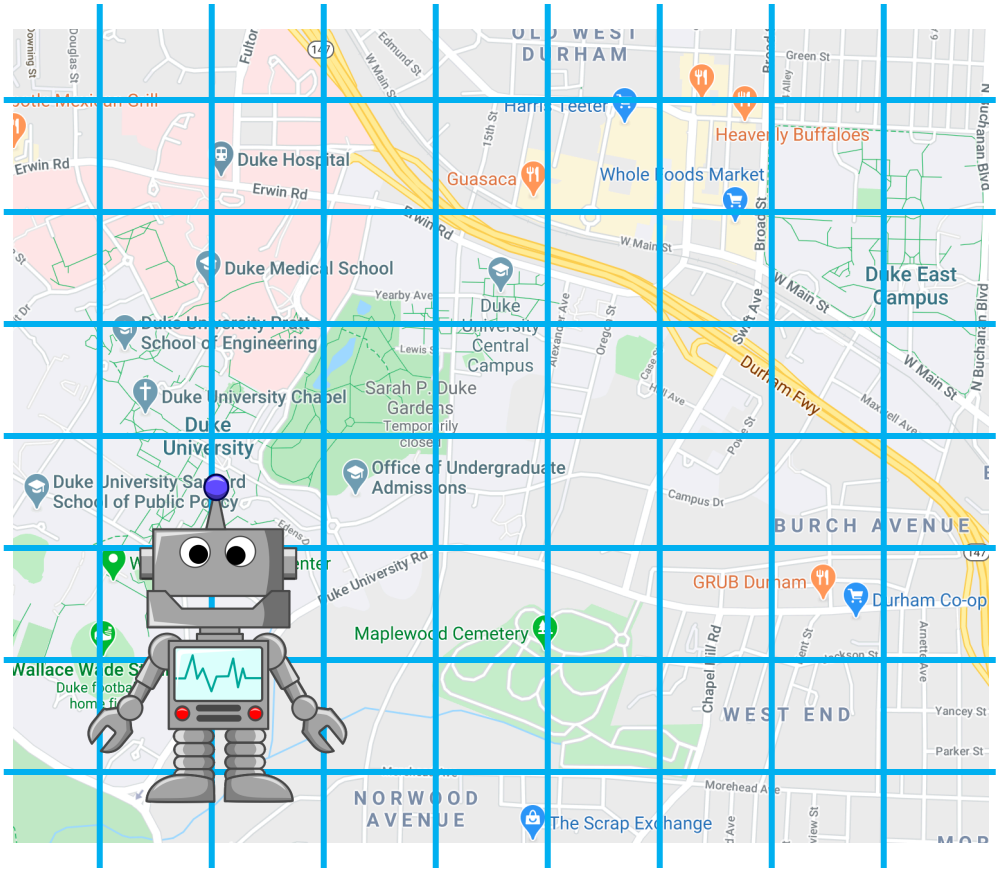


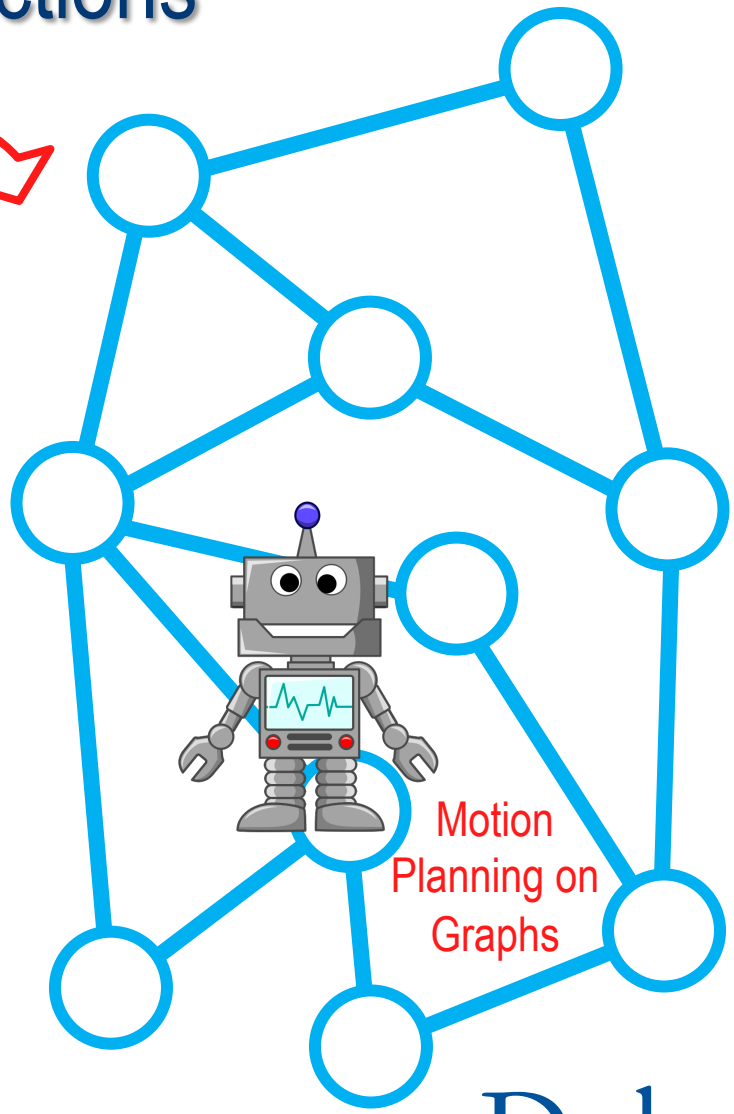

Duke UNIVERSITY

# Discrete Abstractions

Abstraction of Environment and Dynamics

Continuous World

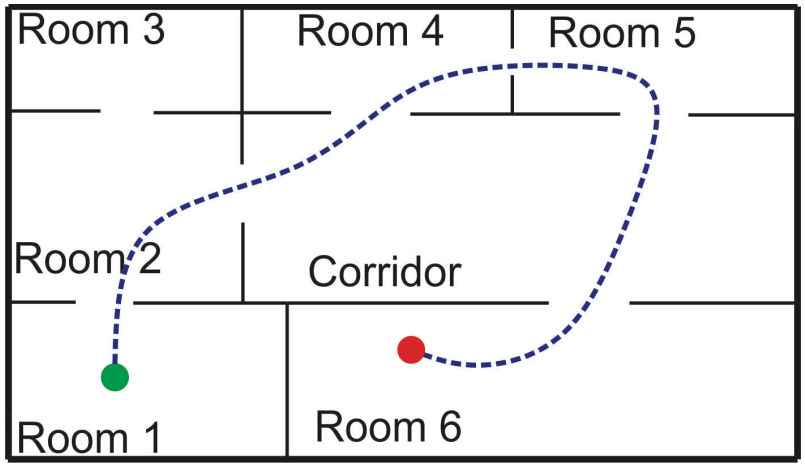



Duke UNIVERSITY

# Optimal Control Synthesis

Given N robots, an abstraction of the environment and robot dynamics, and a collaborative task captured by a global LTL specification $\phi$, synthesize a discrete motion plan $\tau$ such that $\tau \models \phi$ and a user-specified metric $J(\tau)$, such as total traveled distance, is minimized.

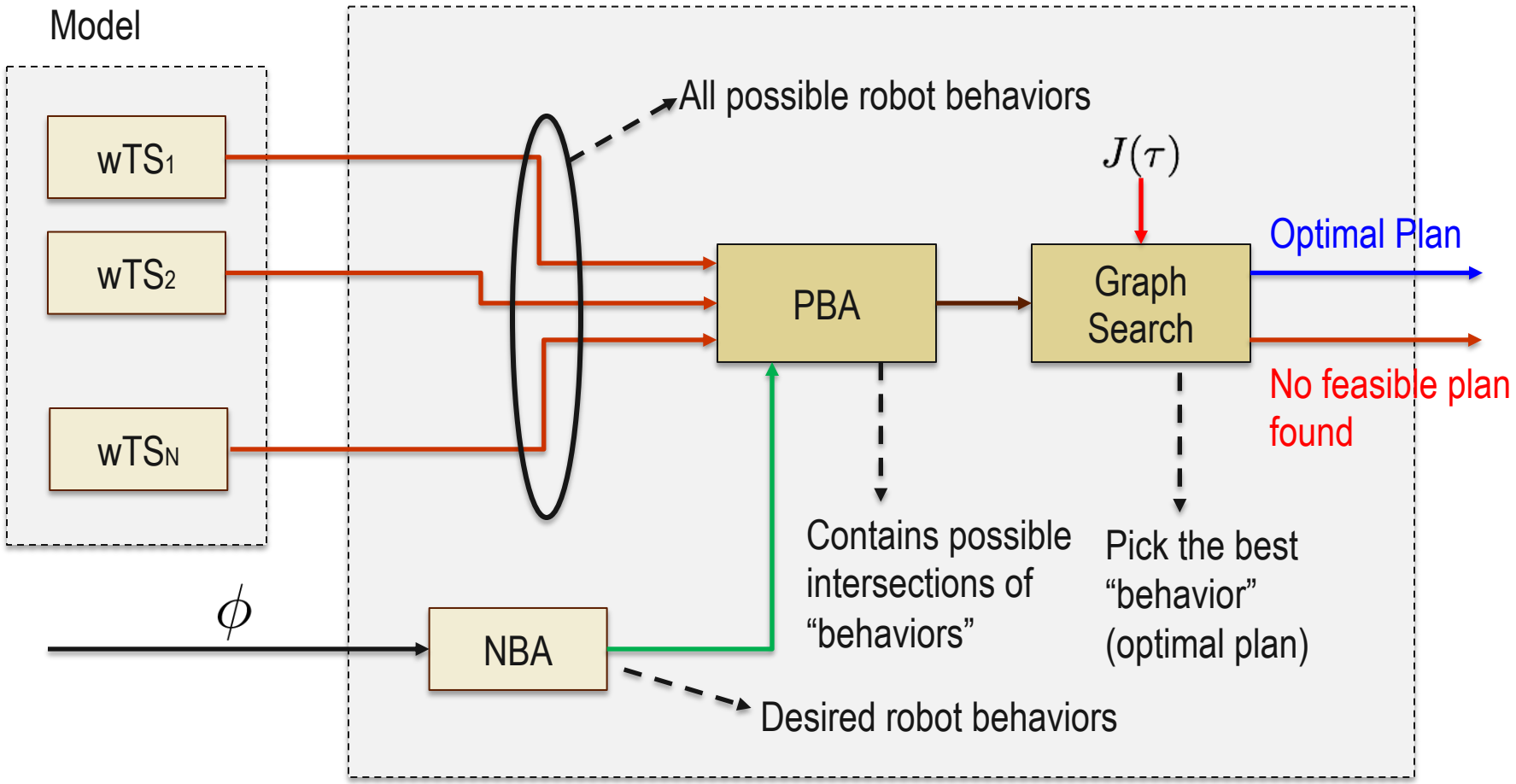



$$\phi = \Diamond(\pi_i^{\text{room2}} \wedge (\Diamond\pi_i^{\text{room4}} \wedge (\Diamond\pi_i^{\text{room5}} \wedge (\Diamond\pi_i^{\text{room6}})))) \wedge (\Diamond\Box\pi_i^{\text{room6}}) \wedge (\Box\neg\pi_i^{\text{room3}})$$

$$\tau = \text{room1,room2,corridor,room4,}\ \text{room5,corridor,room6,}\ [\text{room6}]^{\omega}$$

Duke UNIVERSITY

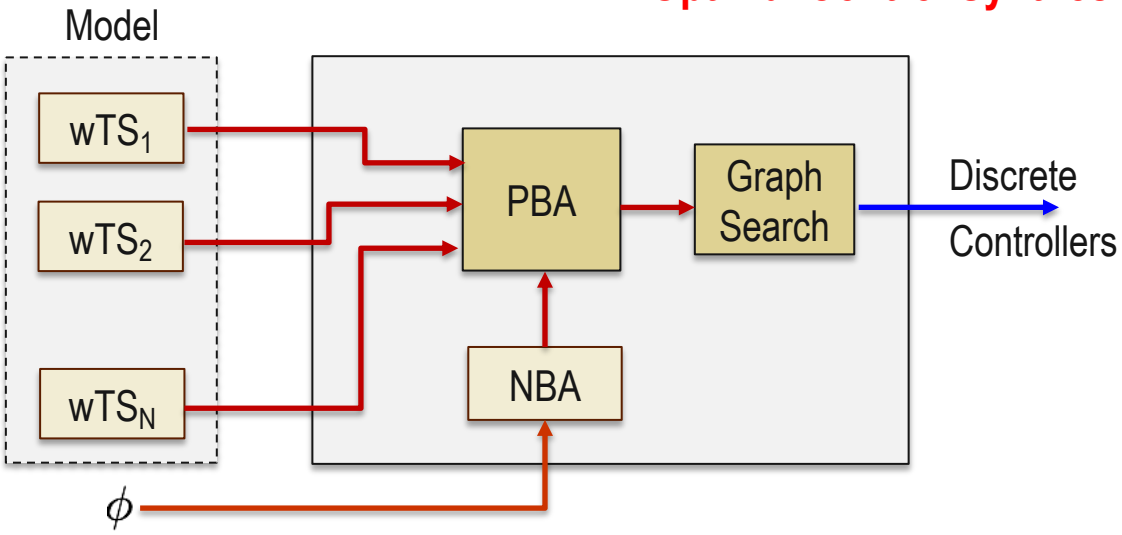
# Key Idea




Duke UNIVERSITY

# Challenges

## Optimal Control Synthesis




M. Kloetzer (TRO 2010)
S. Smith et al (IJRR 2011)
A. Ulusoy et al (IJRR 2013)
M. Guo et al (IJRR 2015)

State explosion, Computationally expensive, Centralized (less than ~$10^7$ states)

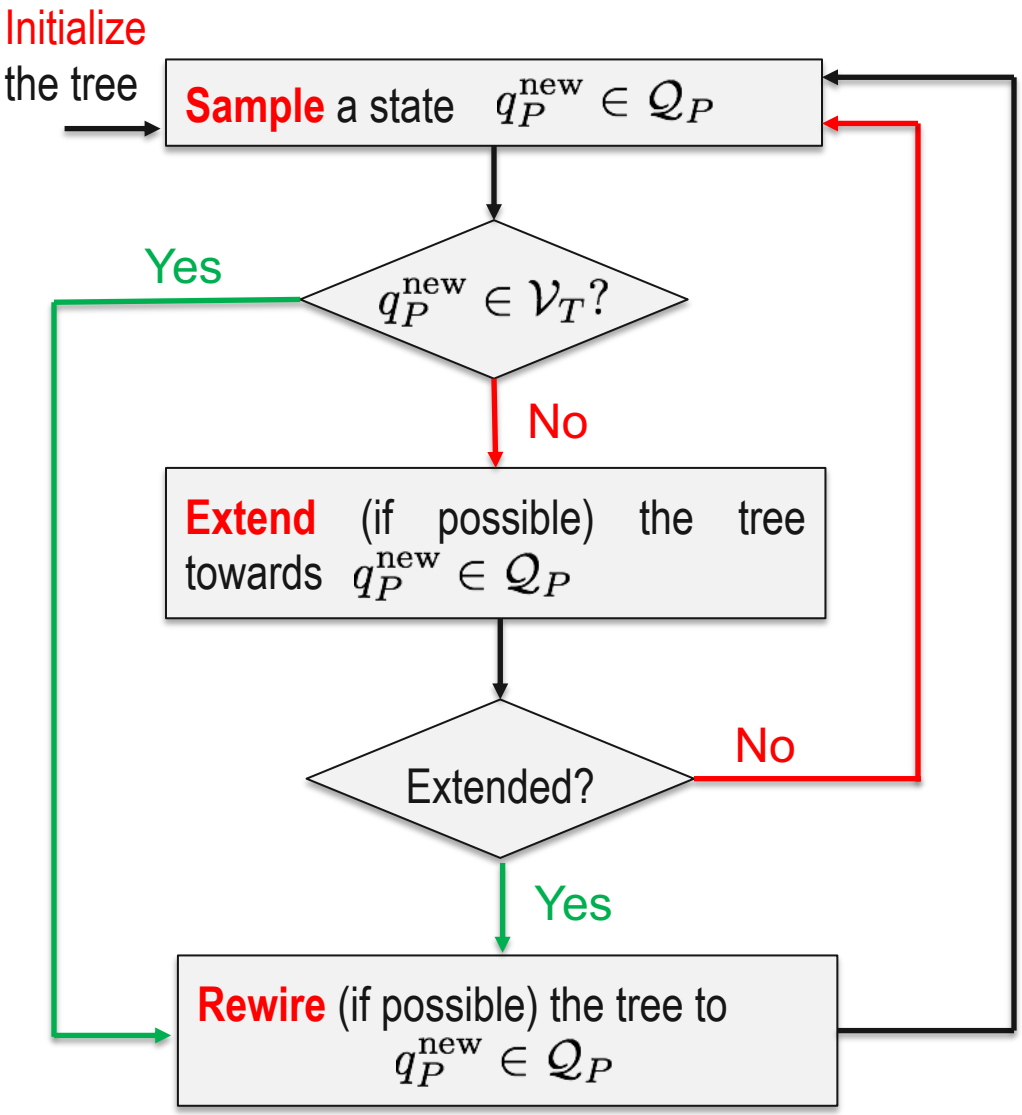
## Model Checking / Verification

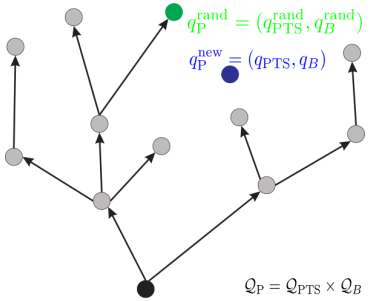
NuSMV 2, nUxmv, SPIN, SPOT

More scalable (~$10^{30}$ states) but no optimality guarantees. Return a feasible, and not the optimal, solution.

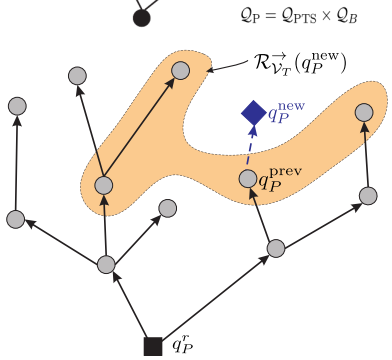
We propose an algorithm that can solve optimally hundreds of orders of magnitude larger planning problems than state-of-the-art methods (~$10^{800}$ states and beyond).

# STyLuS*: Large-Scale Temporal Logic Synthesis

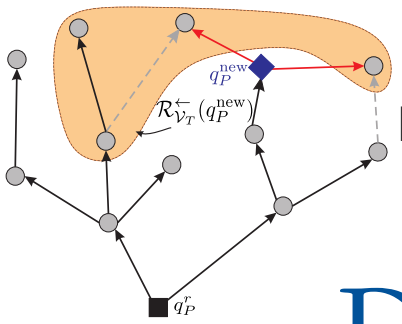



Sample




Extend




Rewire

Duke UNIVERSITY

# Completeness and Optimality

**Theorem:** The proposed sampling-based algorithm is **probabilistically complete.**

**Theorem:** The proposed sampling-based algorithm is **asymptotically optimal**, i.e.,

$$\mathbb{P}\left(\left\{\lim_{n_{\max}^{\text{pre}}\to\infty, n_{\max}^{\text{suf}}\to\infty} J(\tau_{n_{\max}^{\text{pre}}}^{n_{\max}^{\text{suf}}}) = J^*\right\}\right) = 1$$

Duke UNIVERSITY

# Convergence Rate Analysis

**Theorem:** Let p denote a feasible prefix or suffix path

$$\mathrm{p} = q_P^1, q_P^2, ..., q_P^{K-1}, q_P^K$$

Then there exist parameters $\alpha_n(\mathrm{p}) \in (0, 1]$ such that the probability $\Pi_{\mathtt{suc}}(q_P^K)$ of finding the feasible prefix/suffix path p within $n_{\mathtt{max}}$ iterations satisfies

$$1 \geq \Pi_{\mathtt{suc}}(q_P^K) \geq 1 - e^{-\frac{\sum_{n=1}^{n_{\mathtt{max}}} \alpha_n(\mathrm{p})}{2} n_{\mathtt{max}} + K}, \quad \text{if } n_{\mathtt{max}} > K$$

Depend on the selected sampling functions
**NEW BIASED SAMPLING METHOD !!!**

**Theorem:** Let p* denote the optimal prefix or suffix path

$$\mathrm{p}^* = q_P^1, q_P^2, ..., q_P^{K-1}, q_P^K$$

Then there exist parameters $\alpha_n(\mathrm{p}^*) \in (0, 1]$ and $\gamma_n(q_P^k) \in (0, 1]$ and iterations $n_k$ for every state $q_P^k$ in the optimal path such that the probability of finding the optimal path within $n_{\mathtt{max}} > 2K$ iterations satisfies

$$\Pi_{\mathtt{opt}}(\mathrm{p}^*) \geq \left(1 - e^{-\frac{\sum_{n=1}^{\bar{n}} \alpha_n(\mathrm{p}^*)}{2} + K}\right) \prod_{k=1}^{K-1} \left(1 - e^{-\frac{\sum_{n=n_{k-1}}^{n_{\mathtt{max}}} \gamma_n(q_P^k)}{2} + 1}\right)$$

# Comparative Results: Large NBA

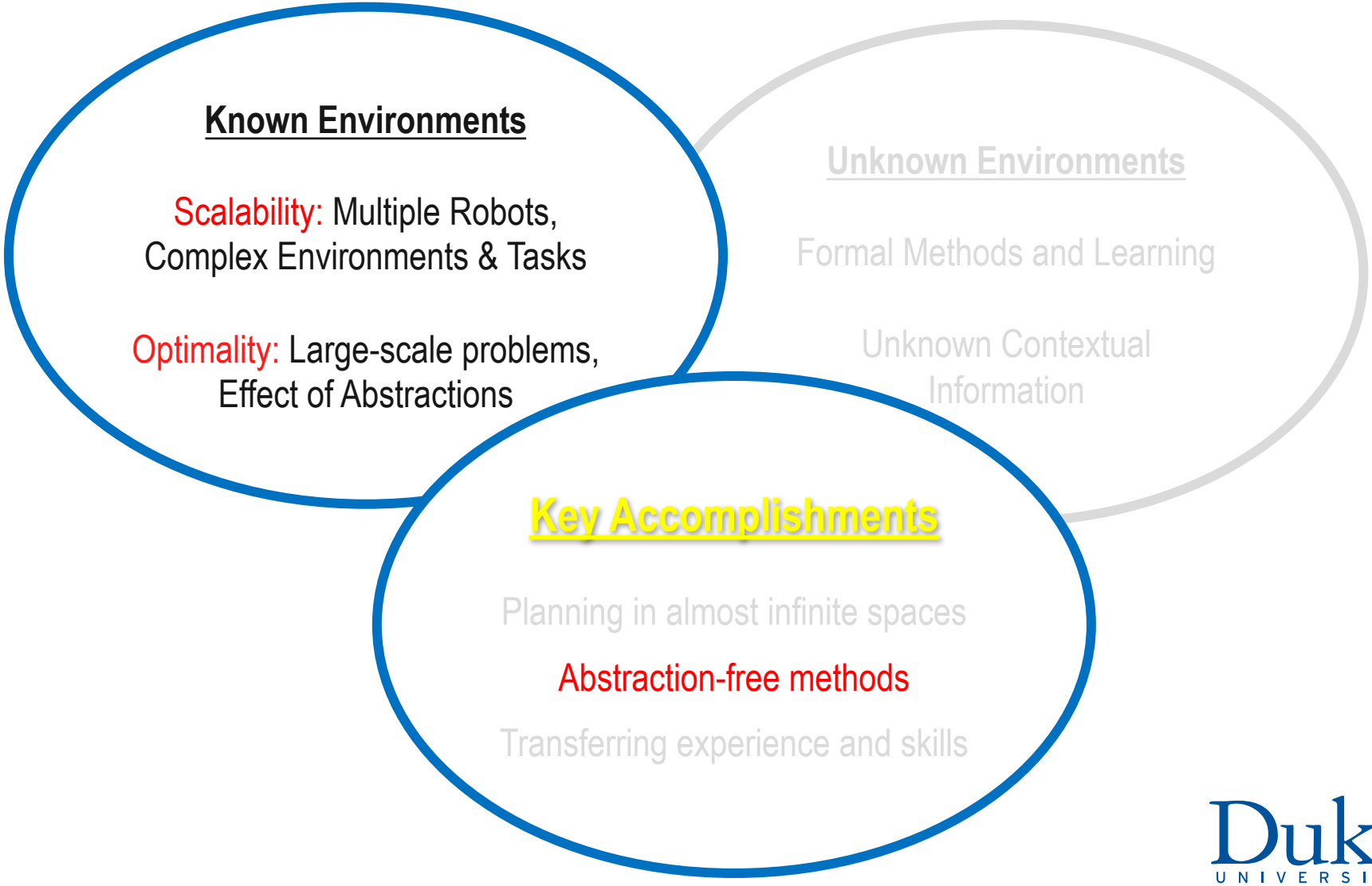
MATLAB runtimes to detect the first feasible plan

TABLE II
FEASIBILITY AND SCALABILITY ANALYSIS: $|\mathcal{Q}_B| = 59$

| $N$ | $\lvert\mathcal{Q}_i\rvert$ | $\lvert\mathcal{Q}_P\rvert$ | $n_{\text{Pre1}} + n_{\text{Suf1}}$ | $\lvert\mathcal{V}_T^{\text{Pre1}}\rvert + \lvert\mathcal{V}_T^{\text{Suf1}}\rvert$ | `Pre1+Suf1` | NuSMV/nuXmv |
|---|---|---|---|---|---|---|
| 1 | 100 | $10^3$ | 54 + 92 | 533 + 274 | 2.18+1.55 (secs) | $<$ 1 sec |
| 1 | 1000 | $10^3$ | 78 +51 | 326 + 252 | 1.84+1.37 (secs) | $<$ 1 sec |
| 1 | 10000 | $10^4$ | 150 + 107 | 769 + 364 | 19.2+11.2 (secs) | M/M |
| 9 | 9 | $10^{10}$ | 93 + 27 | 400 + 168 | 20.7+18.9 (secs) | $<$ 1 sec |
| 10 | 100 | $10^{21}$ | 51+ 39 | 650 + 239 | 2.1+0.74 (secs) | $\approx$ 3/2 secs |
| 10 | 1000 | $10^{31}$ | 36 + 154 | 450 + 404 | 3.9+6.1 (secs) | $\approx$ 80/65 secs |
| 10 | 2500 | $10^{35}$ | 61 + 98 | 710 + 516 | 10.4+11.9 (secs) | M/$\approx$ 32 mins |
| 10 | 10000 | $10^{41}$ | 47 + 164 | 722+604 | 56.6 + 98.1(secs) | M/M |
| 100 | 100 | $10^{200}$ | 21 + 117 | 154 + 1431 | 1.6+18.5 (secs) | F/F |
| 100 | 1000 | $10^{300}$ | 52 + 74 | 401 + 856 | 19.8+53.32 (secs) | M/M |
| 100 | 10000 | $10^{400}$ | 39 + 89 | 398 + 1621 | 5.1+28.3 (mins) | M/M |
| 150 | 10000 | $10^{600}$ | 39 + 112 | 526+1864 | 8.3 + 60.11 (mins) | M/M |
| 200 | 10000 | $10^{800}$ | 48 + 103 | 588 + 1926 | 11.7+65.9 (mins) | M/M |

Duke UNIVERSITY

# Challenges & Key Accomplishments

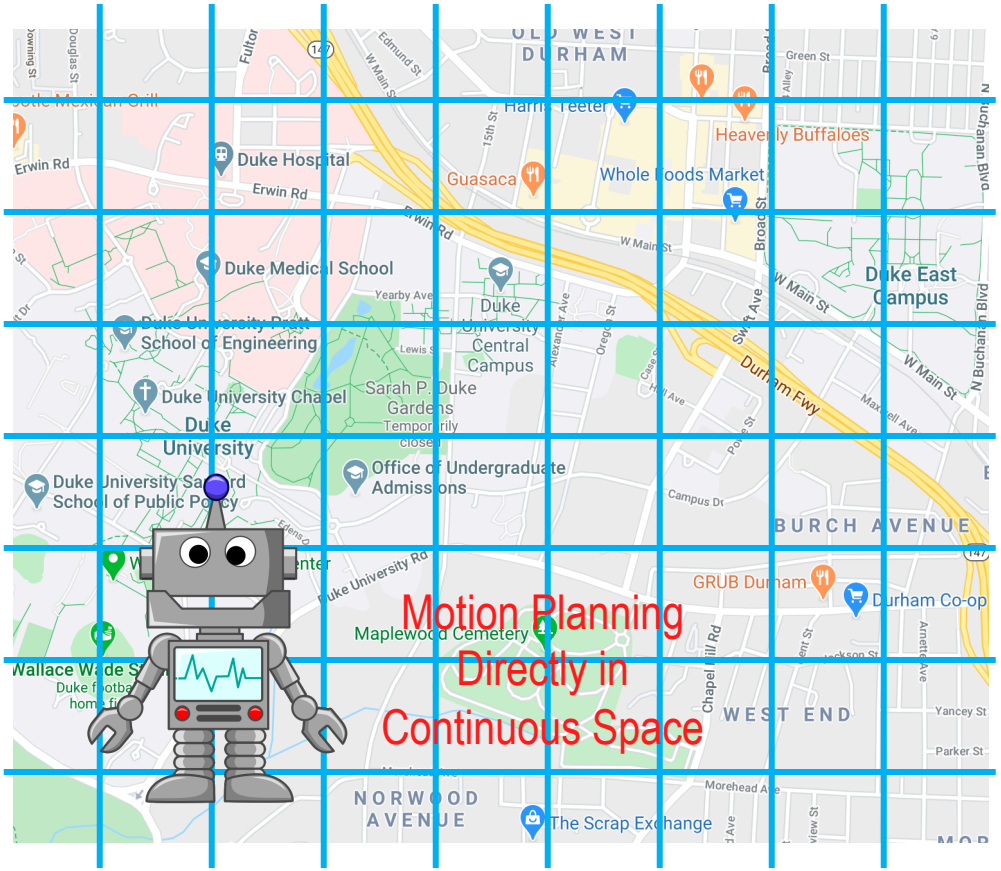


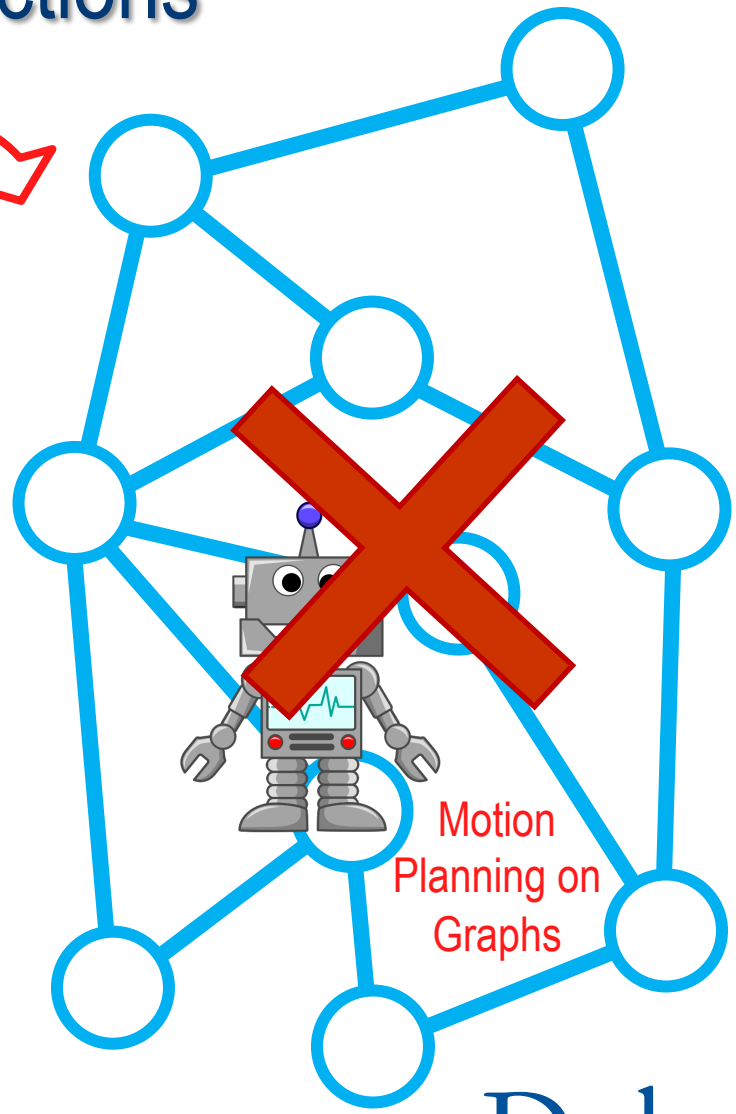

Duke UNIVERSITY

# Discrete Abstractions

Abstraction of Environment and Dynamics

Continuous World



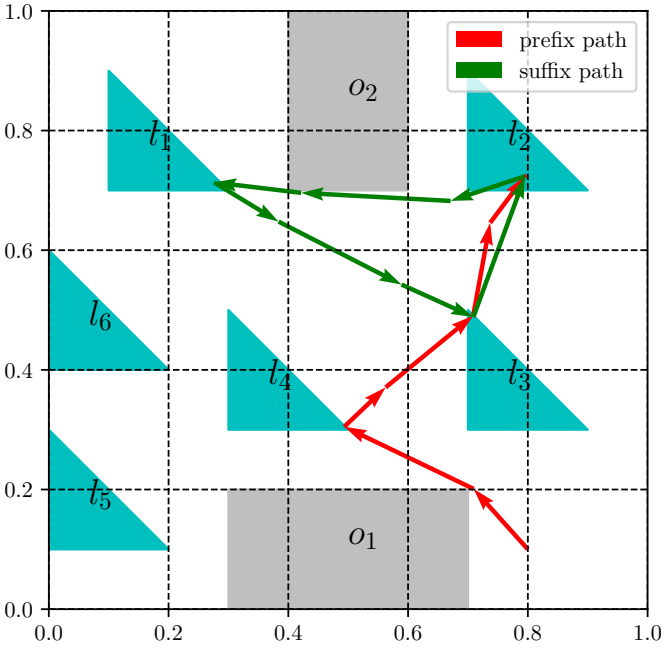




Duke UNIVERSITY

# Abstraction-Free Optimal Control Synthesis

Given N robots, **a continuous environment** and a collaborative task captured by a global LTL specification $\phi$, synthesize a discrete motion plan $\tau$ such that $\tau \models \phi$ and a user-specified metric $J(\tau)$, such as total traveled distance, is minimized.

$$\phi = \Diamond(\pi_i^{\ell_4}) \wedge \Box\Diamond(\pi_i^{\ell_3} \wedge (\Diamond\pi_i^{\ell_1})) \wedge (\neg\pi_i^{\ell_1} \mathcal{U} \pi_i^{\ell_2}) \wedge \Box(\neg\pi_i^{\ell_5})$$




Duke UNIVERSITY

# TL-RRT*: Temporal Logic RRT*

Initialize the tree

**Sample** a state $q_P^{\text{new}} \in \mathcal{Q}_P$

**Extend** (if possible) the tree towards $q_P^{\text{new}} \in \mathcal{Q}_P$

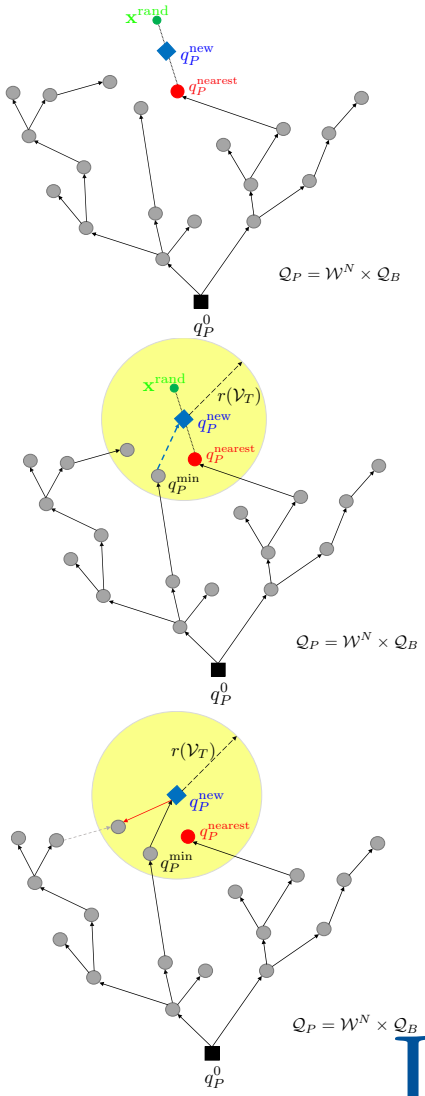
connection radius $r_n(\mathcal{V}_\mathcal{T})$

Extended?

No

Yes

**Rewire** (if possible) the tree to $q_P^{\text{new}} \in \mathcal{Q}_P$




Sample

Extend

Rewire

Duke UNIVERSITY

# Completeness and Optimality

**Theorem:** Let Assumptions 1 and 2 hold and further assume that sampling in the free workspace is unbiased. Then, TL-RRT* is **probabilistically complete**.

**Theorem:** Let Assumptions 1 and 2 hold and further assume that sampling in the free workspace is unbiased. Consider also the connection radius

$$r_n(\mathcal{V}_\mathcal{T}) = \min\left\{\gamma_{\text{TL-RRT}^*}\left(\frac{\log|[\mathcal{V}_\mathcal{T}]_\sim|}{|[\mathcal{V}_\mathcal{T}]_\sim|}\right)^{1/\mathtt{dim}}, \eta\right\},$$

where

$$\gamma_{\text{TL-RRT}^*} > 4\left[\frac{\mu(\mathcal{W}_{\text{free}}^N)}{\zeta_{\mathtt{dim}}}\right]^{1/\mathtt{dim}}.$$

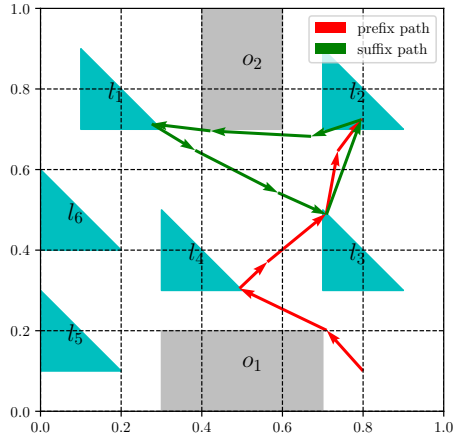
Then, TL-RRT* is **asymptotically optimal**, i.e.,

$$\mathbb{P}\left(\left\{\lim_{n_{\max}^{\text{pre}}\to\infty, n_{\max}^{\text{suf}}\to\infty} J(\tau_{n_{\max}^{\text{pre}}}^{n_{\max}^{\text{suf}}}) = J^*\right\}\right) = 1$$

# Performance for Different Sizes of Regions

TABLE II
COMPARISON OF RUNTIMES AND COST FOR DIFFERENT SIDE LENGTH OF REGIONS

| $s$ | runtime | | | | cost | | | |
|---|---|---|---|---|---|---|---|---|
| | $T^{\text{b}}_{\text{TL-RRT}^*}(s)$ | $T_{\text{SMC}}(s)$ | $T^{\text{u}}_{\text{TL-RRT}^*}(s)$ | $T_{\text{RRG}}(s)$ | $J^{\text{b}}_{\text{TL-RRT}^*}$ | $J_{\text{SMC}}$ | $J^{\text{u}}_{\text{TL-RRT}^*}$ | $J_{\text{RRG}}$ |
| 0.25 | 0.66 | 10.99 | 8.39 | 50.02 | 1.63 | 2.19 | 2.33 | 3.68 |
| 0.20 | 0.61 | 9.70 | 12.43 | 249.67 | 1.67 | 2.68 | 2.75 | 3.67 |
| 0.15 | 2.17 | 10.10 | 20.94 | — | 1.93 | 2.48 | 2.54 | — |
| 0.10 | 3.11 | 10.25 | 106.97 | — | 1.88 | 2.27 | 2.75 | — |
| 0.05 | 8.01 | 14.55 | 444.26 | — | 1.89 | 2.17 | 2.85 | — |





- C. I. Vasile and C. Belta, "Sampling-based temporal logic path planning," IROS 2013.
- Y. Shoukry, P. Nuzzo, A. Balkan, I. Saha, A. L. Sangiovanni-Vincentelli, S. A. Seshia, G. J. Pappas, and P. Tabuada, "Linear temporal logic motion planning for teams of underactuated robots using satisfiability modulo convex programming," CDC 2017.

# Performance for Different Complexity of Tasks

$$\phi = \square\Diamond\xi_1 \wedge \square\Diamond\xi_2 \wedge \square\Diamond\xi_3 \wedge \square\Diamond(\xi_4 \wedge \Diamond(\xi_5 \wedge \Diamond\xi_6)) \wedge \Diamond\xi_7 \wedge \square\Diamond\xi_8 \wedge (!\xi_7 \,\mathcal{U}\, \xi_8). \longrightarrow \xi_e = \wedge_{i=1}^{m} \pi_i^{\ell_j}$$

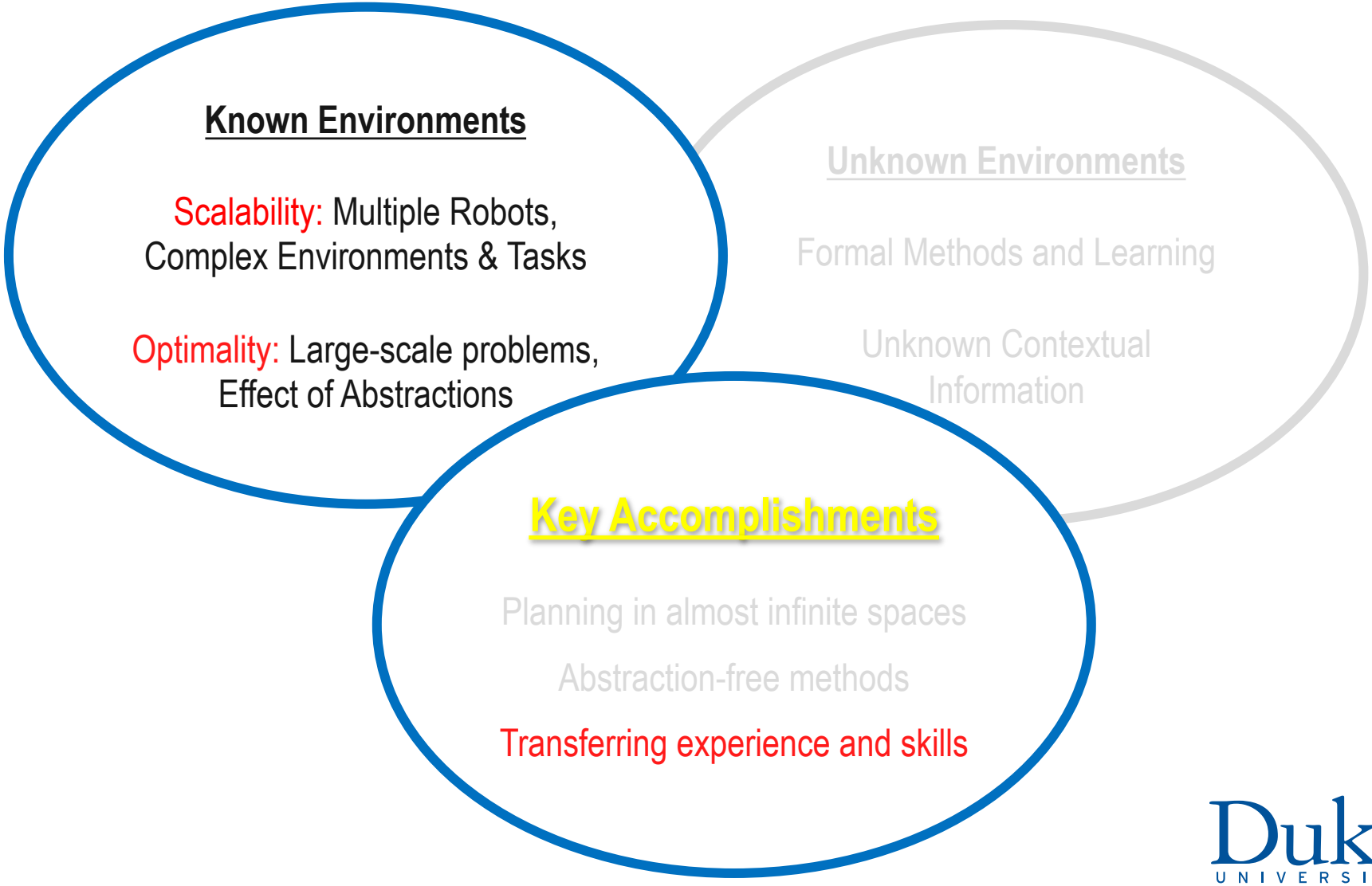
SMC runtimes with "perfect" initial horizons

TABLE IV
COMPARISON OF RUNTIMES AND COST FOR TASKS WITH INCREMENTAL COMPLEXITY

| Task | TL-RRT* | | | | SMC-based | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T_{\text{pre}}(s)$ | $T_{\text{suf}}(s)$ | $T_{\text{total}}(s)$ | $J(\tau)$ | $T_{\text{SAT}}(s)$ | $T_{\text{CPLEX}}(s)$ | $T_{\text{total}}(s)$ | $J(\tau)$ | $T_{\text{total}}^{(1)}(s)$ |
| $\phi_8$ | 0.92 | 0.29 | 1.21 | 2.99 | 8.18 | 0.29 | 8.47 | 3.30 | 12.07 |
| $\phi_{16}$ | 12.05 | 2.88 | 14.93 | 7.73 | 88.34 | 1.47 | 89.81 | 8.27 | 131.66 |
| $\phi_{24}$ | 11.75 | 3.68 | 15.44 | 8.75 | 167.39 | 3.16 | 170.54 | 9.93 | 251.43 |
| $\phi_{32}$ | 34.25 | 39.41 | 73.67 | 13.80 | 314.25 | 7.09 | 321.35 | 11.81 | 470.07 |
| $\phi_{40}$ | 77.94 | 16.77 | 94.71 | 13.45 | 1011.06 | 14.58 | 1025.65 | 14.16 | 1599.50 |
| $\phi_{48}$ | 113.46 | 32.81 | 146.27 | 15.91 | 922.70 | 38.14 | 960.84 | 17.19 | 1380.63 |
| $\phi_{56}$ | 253.13 | 118.70 | 371.84 | 16.69 | 1244.26 | 85.28 | 1329.53 | 17.53 | 1632.21 |

Duke UNIVERSITY

# Challenges & Key Accomplishments




Duke UNIVERSITY

# Transferring Skills in LTL Planning

**A delivery task**

- "Pick up the mail by visiting houses **IN A GIVEN ORDER**"
- "Next visit a delivery site"
- "**NEVER LEAVE THE DELIVERY SITE UNTIL A GROUND ROBOT IS PRESENT TO PICK UP THE MAIL**"
- "Repeat this process every day"

**New Delivery Task**

- "Pick up the mail by visiting houses in **ANY** order"
- "Next visit a delivery site **AND DROP OFF THE MAIL**"
- "Repeat this process every day"

Already know how to visit houses and delivery site.
**Why plan from scratch?**

Duke UNIVERSITY

# Transferring Skills in LTL Planning

Library of Atomic Skills

Sampling-Based Controller Synthesis

Same as Before!

Sample

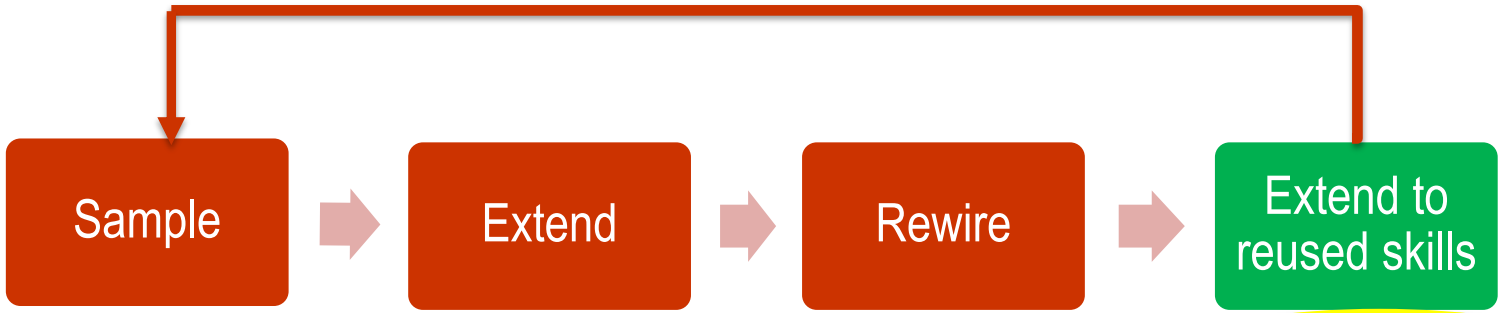
Extend

Rewire

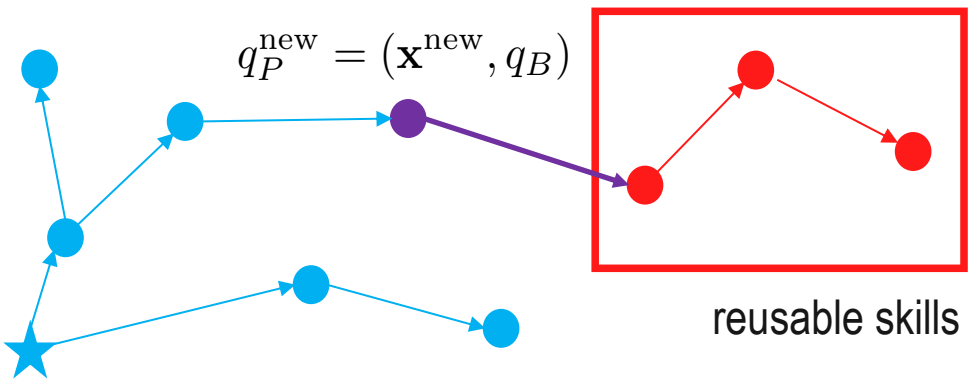
Extend to other subtasks

# Control Synthesis for New LTL Tasks

**Step 1:** Decompose the new LTL into subtasks and **match** with skills in the library.
**Step 2:** Grow a tree by **sampling** and **reusing skills** from the library.




**Theorem:** Probabilistically Complete and Asymptotically Optimal




Duke UNIVERSITY

# Transfer Planning for LTL Different Tasks

Library : $\Diamond\pi^{r}$ , $\Diamond\pi^{r_6}$

Tasks :

$\phi_1 =$

$\phi_2 =$

$\phi_3 =$

$\phi_4 =$

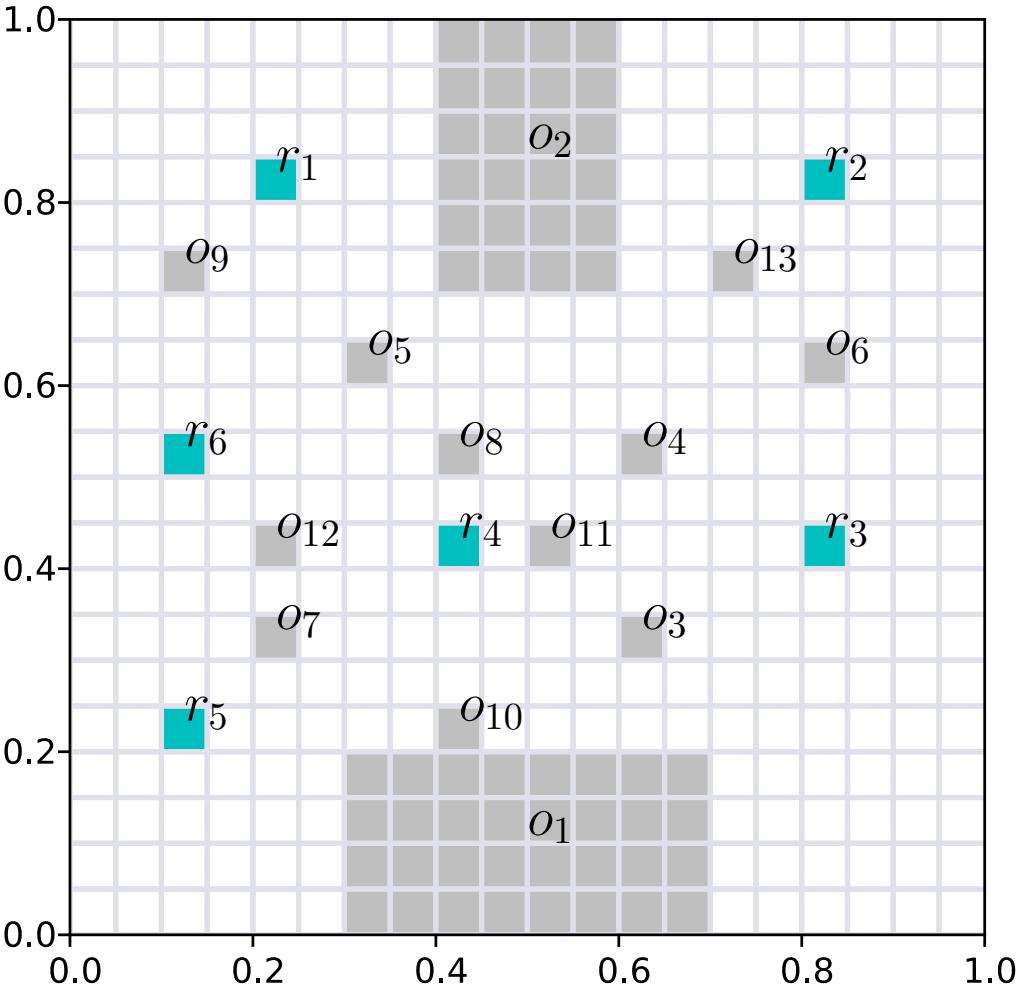
$\phi_5 =$

$\wedge$

$\phi_6 =$

$\wedge$

$\wedge$




$\Diamond r_6$

$r_4))$

$\wedge \Diamond r_6)$

$r_4))$

Duke UNIVERSITY

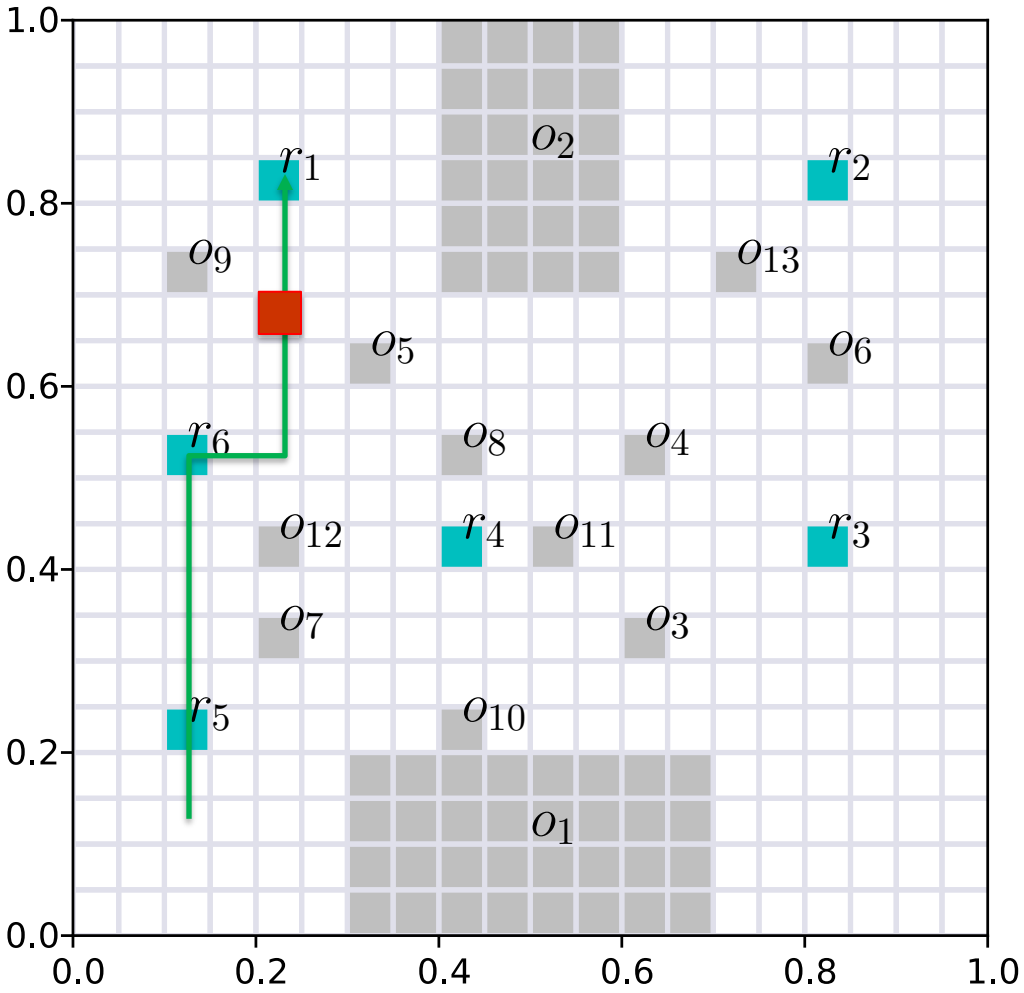
# Transfer Planning for LTL Different Tasks

Table 1: Runtimes and costs for different LTL tasks

| tasks | $t$ | $t$[18] | $J$ | $J$[18] |
|---|---|---|---|---|
| $\phi_1$ | 0.02 | 0.22 | 1.85 | 1.86 |
| $\phi_2$ | 0.01 | 0.20 | 1.60 | 1.47 |
| $\phi_3$ | 0.01 | 0.70 | 3.40 | 3.45 |
| $\phi_4$ | 0.07 | 1.16 | 3.25 | 3.15 |
| $\phi_5$ | 0.20 | 3.40 | 5.10 | 2.83 |
| $\phi_6$ | 0.98 | 9.85 | 4.50 | 3.67 |

[18] Kantaros et.al , "Temporal logic optimal control for large-scale multi-robot systems: 10^400 states and beyond," in 2018 IEEE Conference on Decision and Control (CDC)

# Transfer Planning in Different Environments




Duke UNIVERSITY

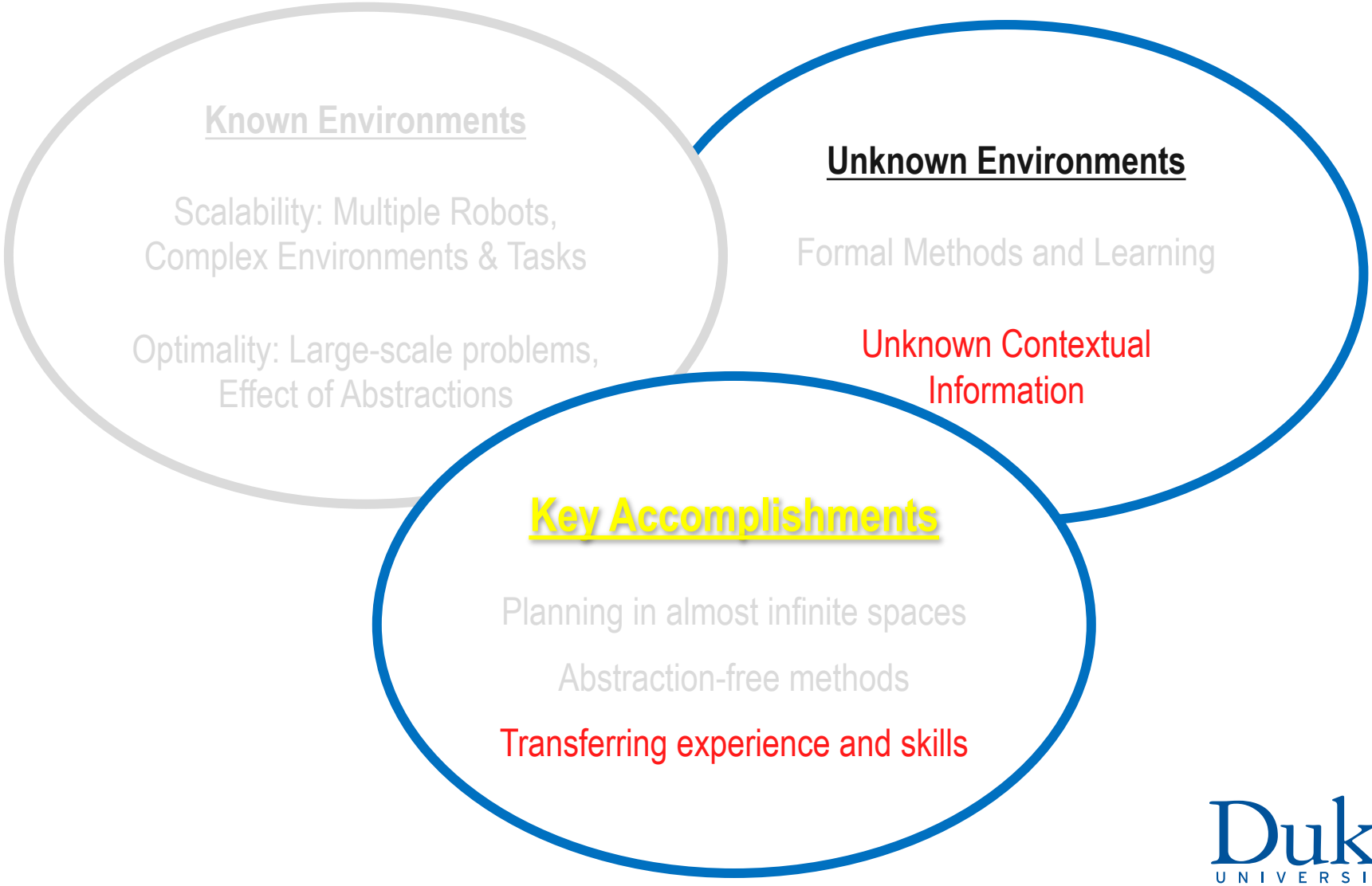
# Transfer Planning in Different Environments

Table 1: Runtimes in the slightly changed environment

| tasks | $m=1$ | | $m=2$ | | $m=3$ | |
|---|---|---|---|---|---|---|
| | $t$ | $t$[18] | $t$ | $t$[18] | $t$ | $t$[18] |
| $\phi_1$ | 0.31 | 0.33 | 0.35 | 0.31 | 0.40 | 0.31 |
| $\phi_2$ | 0.30 | 0.22 | 0.34 | 0.18 | 0.35 | 0.22 |
| $\phi_3$ | 0.35 | 0.76 | 0.39 | 0.89 | 0.43 | 0.84 |
| $\phi_4$ | 0.38 | 1.10 | 0.34 | 1.18 | 0.42 | 0.98 |
| $\phi_5$ | 0.31 | 3.32 | 0.37 | 4.17 | 0.41 | 3.77 |
| $\phi_6$ | 0.34 | 9.49 | 0.38 | 10.17 | 0.44 | 13.38 |

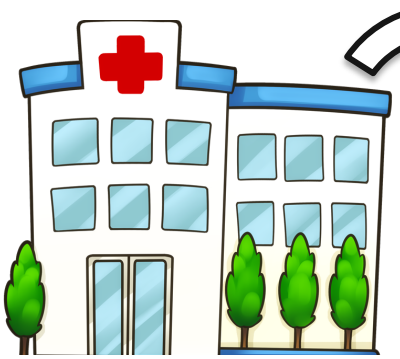
[18] Kantaros et.al , "Temporal logic optimal control for large-scale multi-robot systems: 10^400 states and beyond," in 2018 IEEE Conference on Decision and Control (CDC)

# Challenges & Key Accomplishments

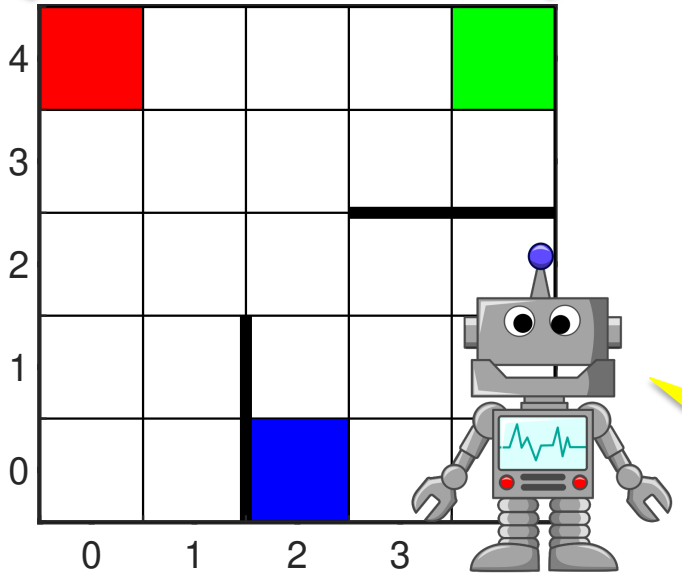


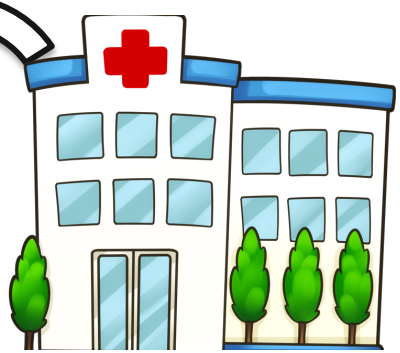

Duke UNIVERSITY

# Contextual Motion Planning



Hospital A specializes in treating disease 1



Hospital B specializes in treating disease 2

TRANSFER PATIENT TO HOSPITAL

Hospital A has larger capacity and, therefore, lower waiting time

Diagnosis

Disease 1

Disease 2

Reward

| u | Red(+10) | Green(+5) |
|---|---|---|
| 0 | 0.6 | 0.3 |
| 1 | 0.1 | 0.8 |

Context

Probabilities

More likely patient with disease 1 will be cured in Hospital A

More likely patient with disease 2 will be cured in Hospital B

Duke UNIVERSITY

# Transfer Learning in Contextual MDPs

**Contextual MDP:** $(s_t, a_t, P^u(s_{t+1}|s_t, a_t), R^u(s_t, a_t, s_{t+1}), \rho(u))$

Transition and reward functions are parameterized by the contextual variable u

Contextual variable u subject to a stationary distribution

In traditional transfer or imitation learning, the demonstrator and learner make decisions based on **the same** information

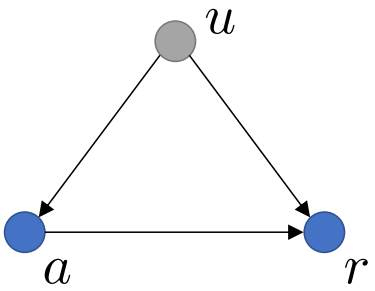
**Transfer Learning under Unobserved Contextual Information**

Given the observed data distribution $P(s_t, a_t, s_{t+1}, r_t)$ collected by a demonstrator agent who makes decisions based on the contextual information, design learning algorithms for a context-unaware learner agent to use these data to find the optimal policy with fewer new data samples.

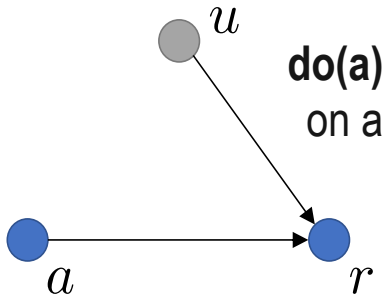
A contextual optimal policy $\pi^*(a_t|s_t, u_t)$ (or an optimal policy $\pi^*(a_t|s_t)$ when the contextual information is unobservable) is the one that maximizes the accumulated reward.

Duke UNIVERSITY

# Causal Inference




**Demonstrator knows context u**




**Learner does NOT know context u**

**do(a) :** intervention on a regardless of the context

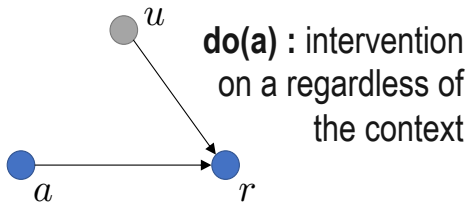
**Causal Inference:** Given the observed data distribution P(r,a,u), or P(r,a), induced from the demonstrator's causal graph, estimate the learner's probability P(r | do(a)) of the outcome r when intervening on the variable a.

The causal effect P(r | do(a)) **cannot be estimated without bias** when there is an unobserved confounder u in the observation data.

Duke UNIVERSITY

# Estimation Bias

The **Learner** cares to estimate: $\mathbb{E}[r|do(a)] = \sum_r r P(r|do(a))$




**do(a) :** intervention on a regardless of the context

SAME DIFFERENT

where: $P(r|do(a)) = \sum_u P(r|a,u)P(u)$ → Learner does NOT KNOW the context $P(u)$

Instead **Learner** can estimate: $\mathbb{E}[r|a] = \sum_r r P(r|a) = \sum_r r \frac{P(r,a)}{P(a)}$ → **Demonstrator's** observational data by executing policy $\pi^\star(a|s,u)$

**Compare** $\mathbb{E}[r|do(a)]$ **and** $\mathbb{E}[r|a]$ **:**

SAME DIFFERENT

$$P(r|a) = \frac{P(r,a)}{P(a)} = \frac{\sum_u P(r|a,u)P(a|u)P(u)}{\sum_u P(a|u)P(u)} = \sum_u P(r|a,u)\frac{P(a|u)P(u)}{\sum_u P(a|u)P(u)}$$

$P(u)$ scaled by demonstrator's data $\pi^\star(a|s,u)$

Since $P(r|do(a)) \neq P(r|a)$ , we have that $\mathbb{E}[r|do(a)] \neq \mathbb{E}[r|a]$

**Estimation Bias**

# Example

**Action space:** UP (1), RIGHT (2), DOWN (3), LEFT (4)

**Context u:** $P(u=0)=0.2,\ P(u=1)=0.8$ (Bernoulli)

**Reward**

| u | Red(+10) | Green(+5) |
|---|---|---|
| 0 | 0.6 | 0.3 |
| 1 | 0.1 | 0.8 |

**Demonstrator's policy:** $P(a=4|u=0)=0.7, P(a=4|u=1)=0.1$

Learner's expected reward **based on demonstrator's observational data:**

$$P(r|a)=\sum_u P(r|a,u)\frac{P(a|u)P(u)}{\sum_u P(a|u)P(u)}=0.6\frac{0.7\cdot 0.2}{0.7\cdot 0.2+0.1\cdot 0.8}+0.1\frac{0.1\cdot 0.8}{0.7\cdot 0.2+0.1\cdot 0.8}=0.418$$

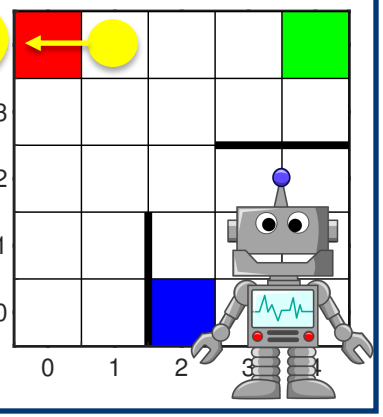
$$\mathbb{E}[r|a]=\sum_r rP(r|a)=10\cdot 0.418+(-1)\cdot(1-0.418)\approx 3.6$$

Learner overestimates reward of moving to Red

Learner's **true expected reward:**

$$P(r|do(a))=\sum_u P(r|a,u)P(u)=0.6\cdot 0.2+0.1\cdot 0.8=0.2$$

$$\mathbb{E}[r|do(a)]=\sum_r rP(r|do(a))=10\cdot 0.2+(-1)\cdot(1-0.2)=1.2$$



# This Bias is why…

… students should not blindly trust their advisors, BUT they should also read and explore their own ideas.

Advisors often guide students (the policy), WITHOUT explaining their thought process (the context).

Still, advisors can help students learn faster and avoid major mistakes…


Duke
UNIVERSITY

# Causal Bound Constrained Q-Learning

While the causal effect $\mathbb{E}[r|do(a)]$ **is unidentifiable** when there is an unobserved confounder u in the observational data, we can compute **causal bounds** on $\mathbb{E}[r|do(a)]$ (and the Q-function) given the demonstrator's observational data.

**Linear Programming!**

**Causal Bound Constrained Q learning**

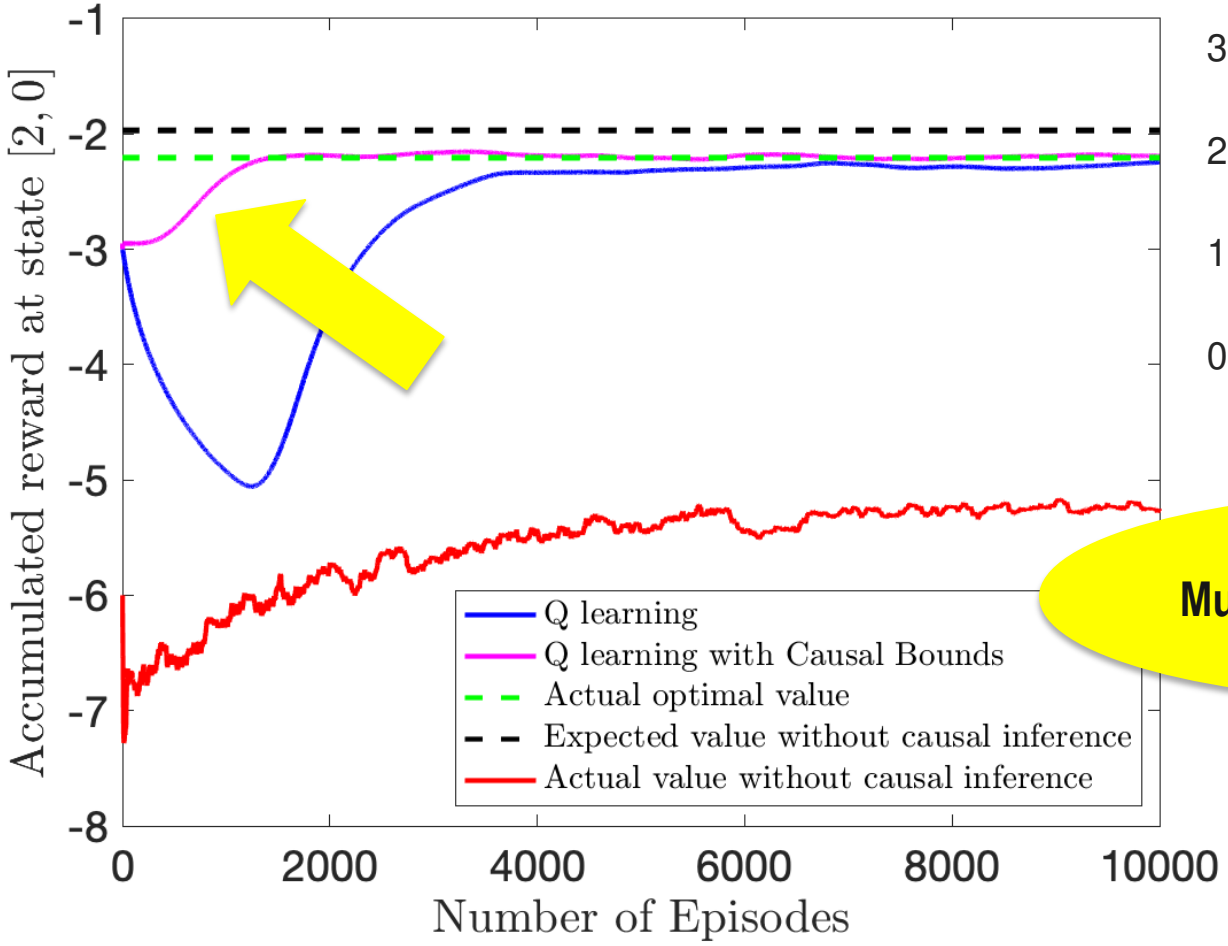
$$a_t \leftarrow \epsilon\text{-Greedy}\big(Q(s_t, a)\big)$$

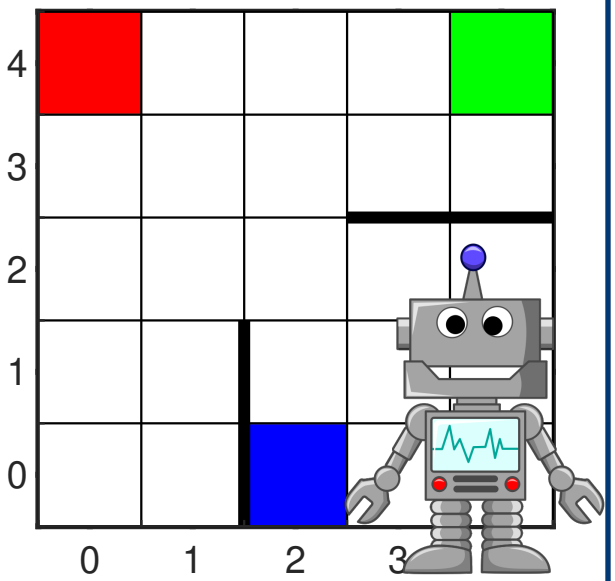
$$Q(s_t, a_t) \leftarrow \Pi_{[\underline{Q}(s_t,a_t),\overline{Q}(s_t,a_t)]}\Big((1-\alpha_t)Q(s_t, a_t) + \alpha_t\big(r(s_t, a_t) + \gamma \max_a Q(s_{t+1}, a)\big)\Big).$$

Projection on causal bounds avoids exploration of the state space in directions that decrease the Q function

Duke UNIVERSITY

# Numerical Experiments

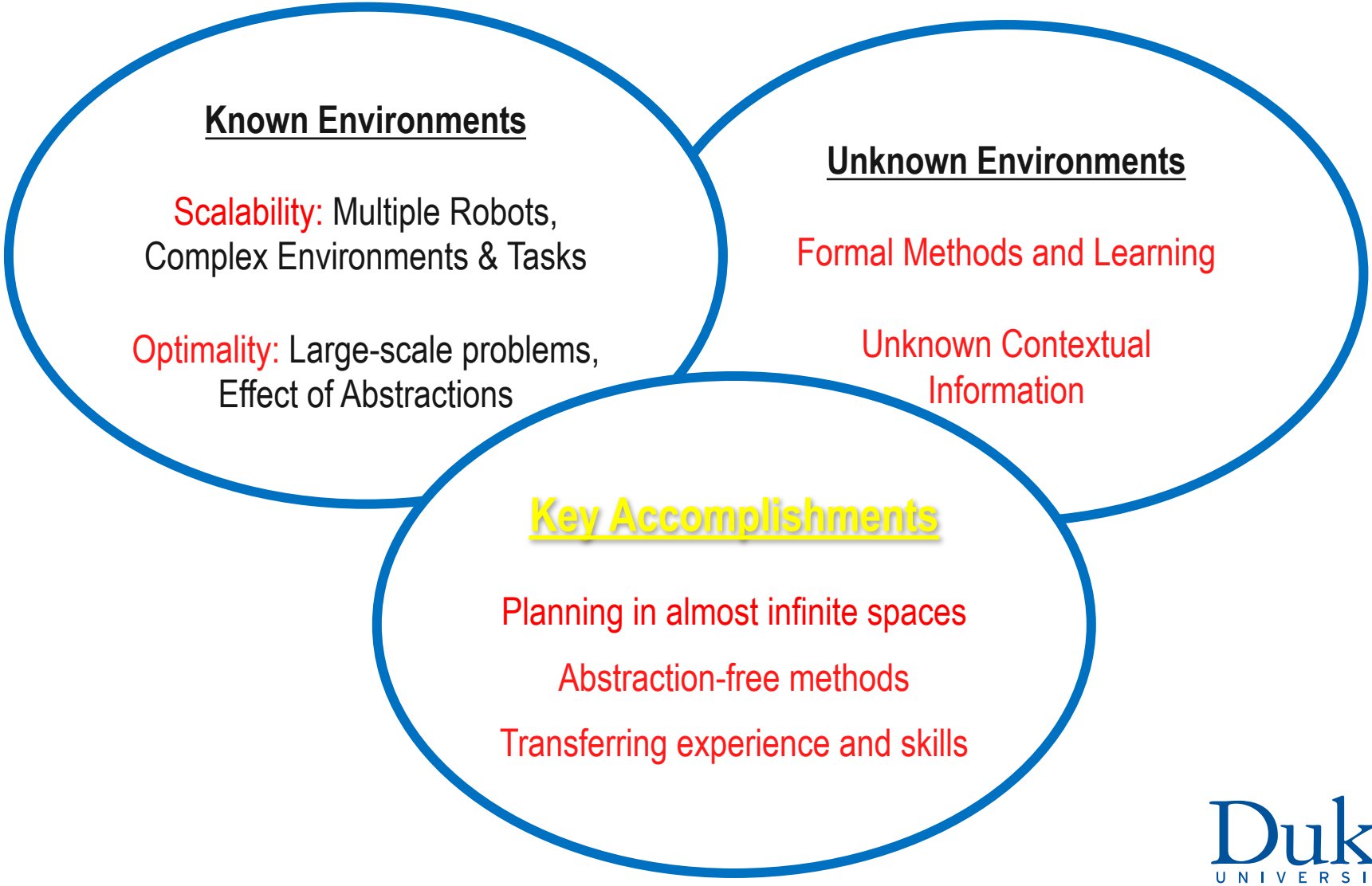



**Much faster convergence!**

Duke UNIVERSITY

# Summary




Duke UNIVERSITY

# Thank You

## STyLuS*: Large-Scale Temporal Logic optimal Synthesis

- Y. Kantaros and M. M. Zavlanos, “Sampling-Based Optimal Control Synthesis for Multi-Robot Systems under Global Temporal Tasks,” IEEE Transactions on Automatic Control, 2019.
- Y. Kantaros and M. M. Zavlanos, “STyLuS*: A Temporal Logic Optimal Control Synthesis Algorithm for Large-Scale Multi-Robot Systems,’” International Journal of Robotics Research, accepted.
- Y. Kantaros and M. M. Zavlanos, “Temporal Logic Optimal Control for Large-Scale Multi-Robot Systems: $10^{400}$ States and Beyond,” 57th IEEE Conference on Decision and Control, 2018.

## TL-RRT*: Temporal Logic RRT*

- X. Luo, Y. Kantaros, and M. M. Zavlanos, “An Abstraction-Free Method for Multi-Robot Temporal Logic Optimal Control Synthesis,” IEEE Transactions on Robotics, under review.

## Transfer Planning and Learning

- X. Luo and M. M. Zavlanos. Transfer planning for temporal logic tasks. Proc. 58th IEEE Conference on Decision and Control (CDC), December 2019.
- Y. Zhang and M. M. Zavlanos. Transfer Reinforcement Learning under Unobserved Contextual Information. 11th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), April 2020.




Duke
UNIVERSITY