Inferring Knowledge from Behavior in Search-and-rescue Tasks

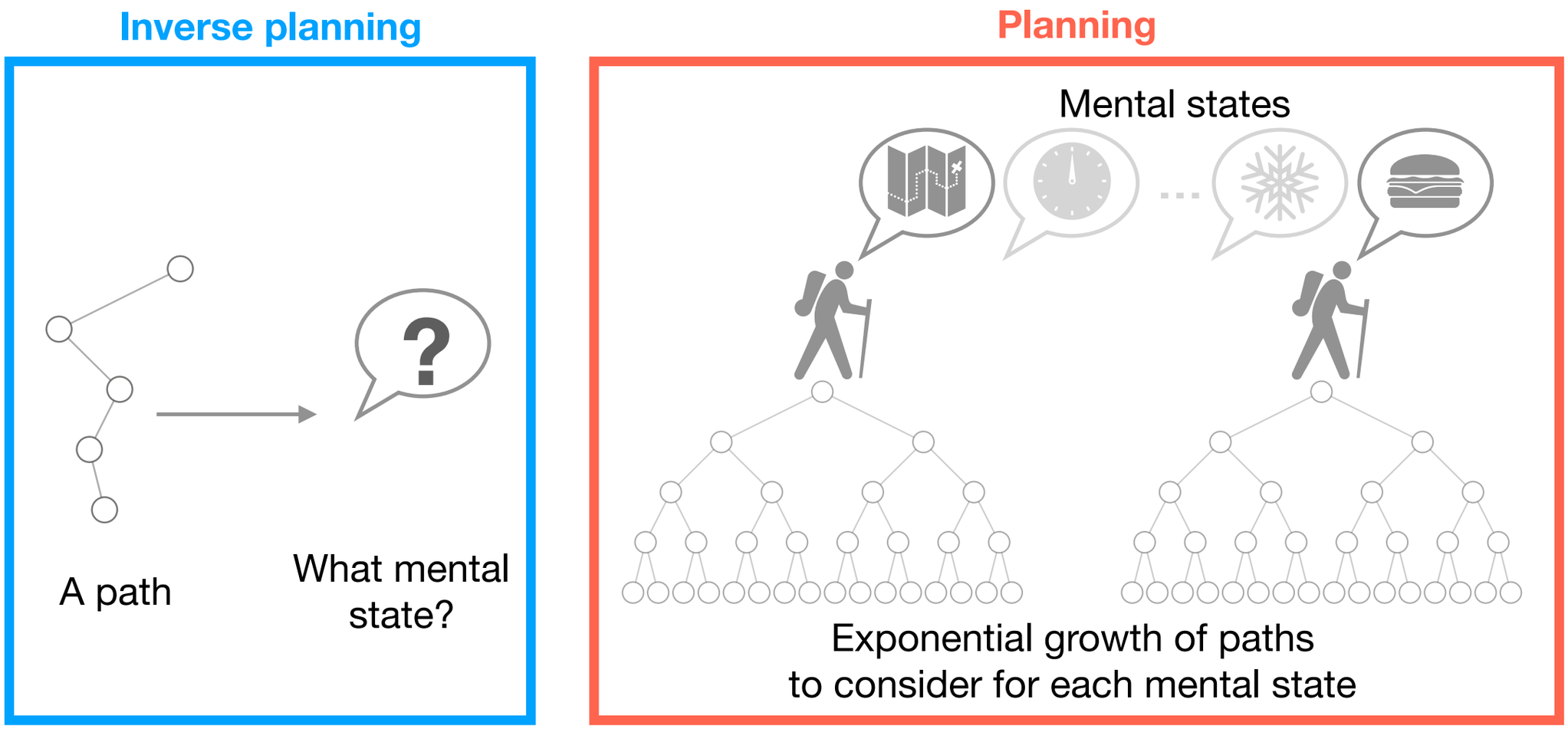
Scott Cheng-Hsin Yang, Sean Anderson, Pei Wang,
Chirag Rank, Tomas Folke, and Patrick Shafto

Department of Math and Computer Science
Rutgers University — Newark

CogSci 2021

Introduction: Theory-of-mind inference as inverse planning

General problem: Theory-of-mind inference is natural for humans but poses significant computational challenges.



Approach: abstraction + myopic rational agent analysis

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329– 349.

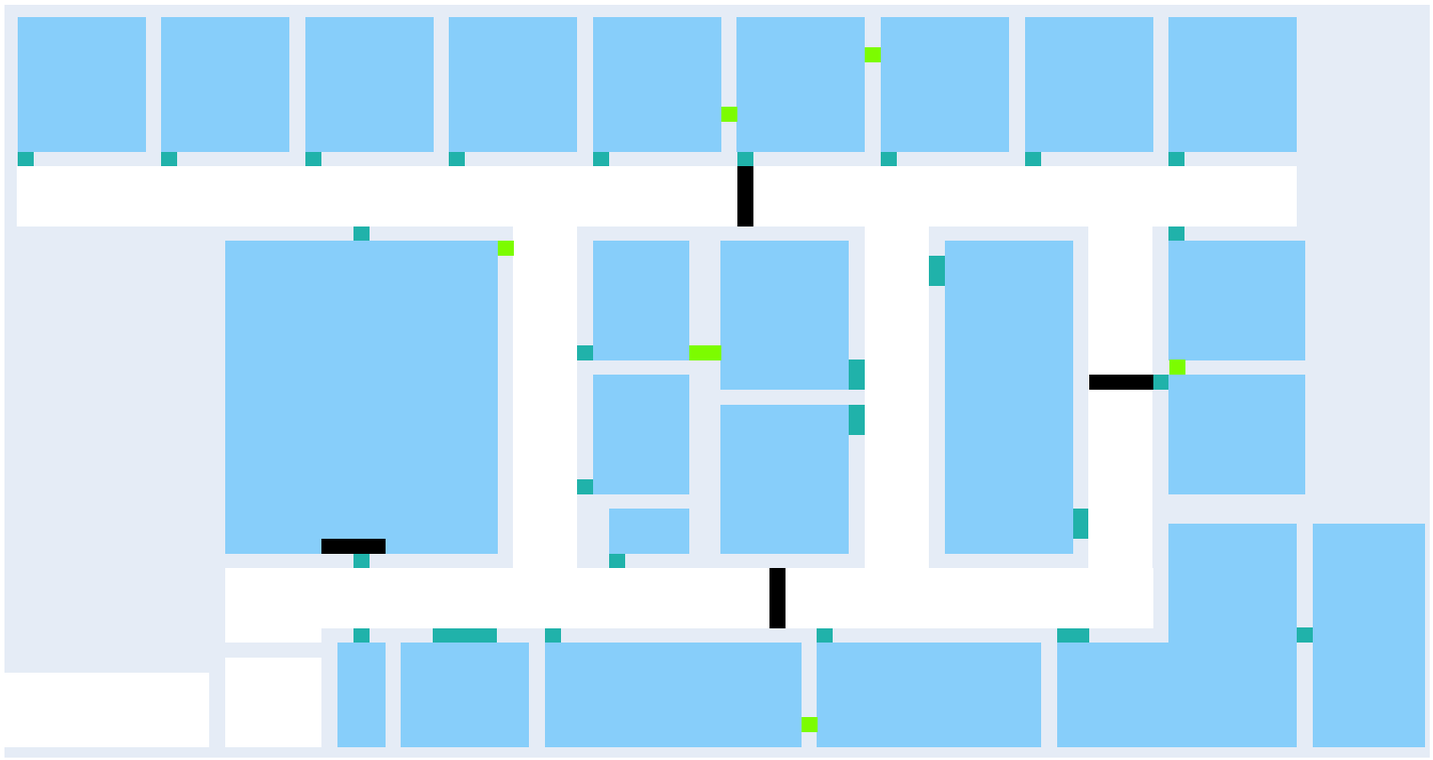
Experiment: Search-and-rescue task in Minecraft

- Each participant engaged in a single-person, first-person-view, search-and-rescue task.

Huang, L., Freeman, J., Cooke, N., et al. (2021). Using humans' theory of mind to study artificial social intelligence in Minecraft search and rescue. Technical report.

Setting: Search-and-rescue task in Minecraft

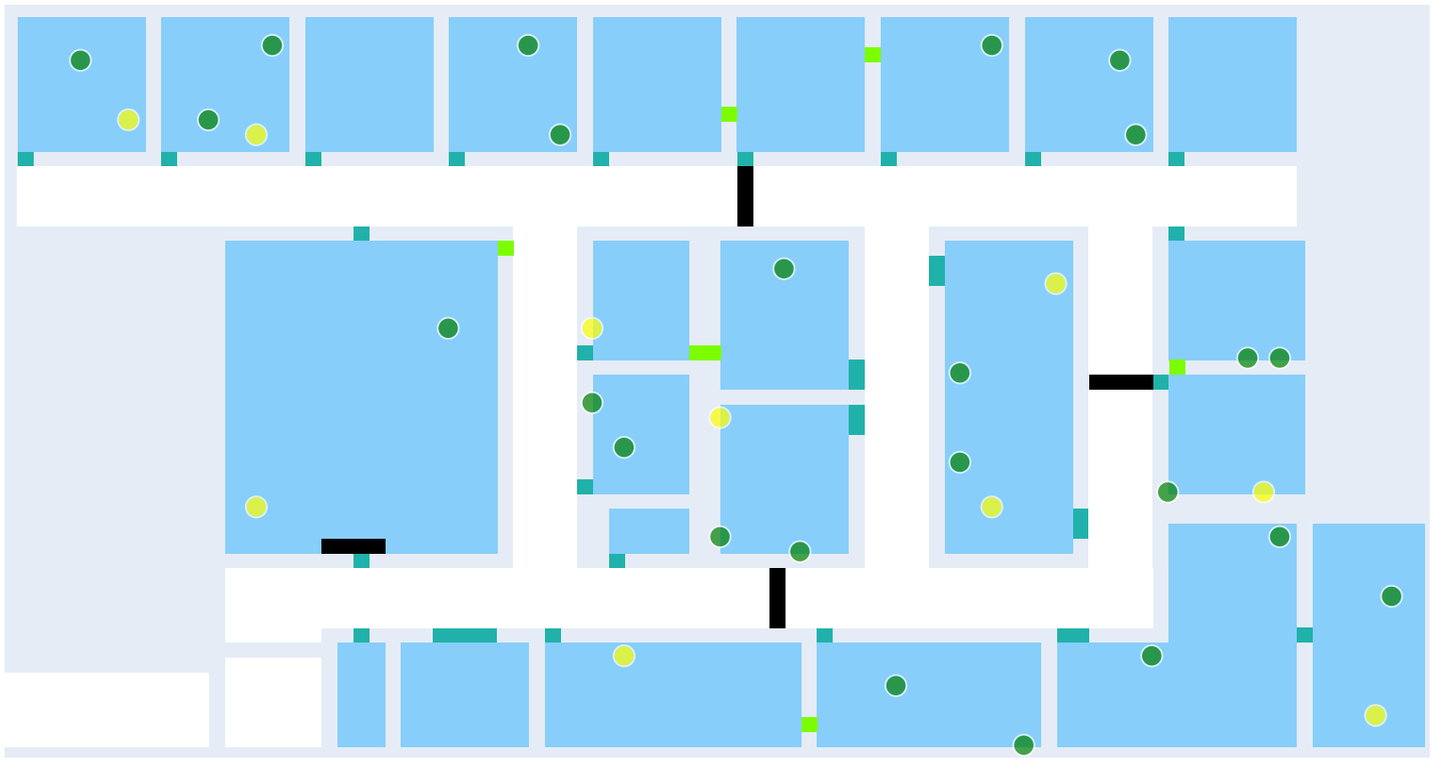
- Each participant engaged in a single-person, first-person-view, search-and-rescue task.
- The environment was a collapsed office building consisting of rooms and corridors.



Huang, L., Freeman, J., Cooke, N., et al. (2021). Using humans' theory of mind to study artificial social intelligence in Minecraft search and rescue. Technical report.

Setting: Search-and-rescue task in Minecraft

- Each participant engaged in a single-person, first-person-view, search-and-rescue task.
- The environment was a collapsed office building consisting of rooms and corridors.

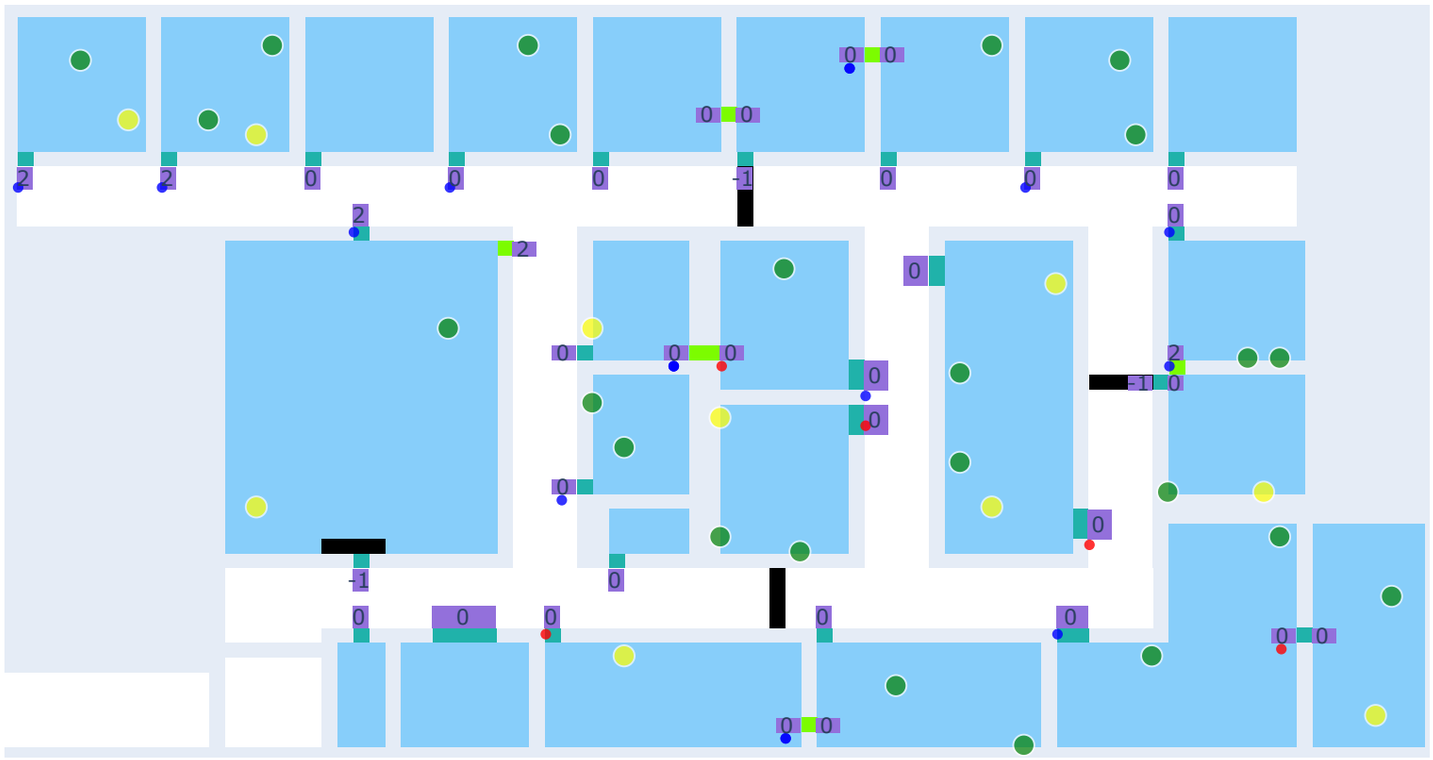


- Victims were inside some of the rooms. The goal was to **maximize points** by triaging victims.

Huang, L., Freeman, J., Cooke, N., et al. (2021). Using humans' theory of mind to study artificial social intelligence in Minecraft search and rescue. Technical report.

Setting: Search-and-rescue task in Minecraft

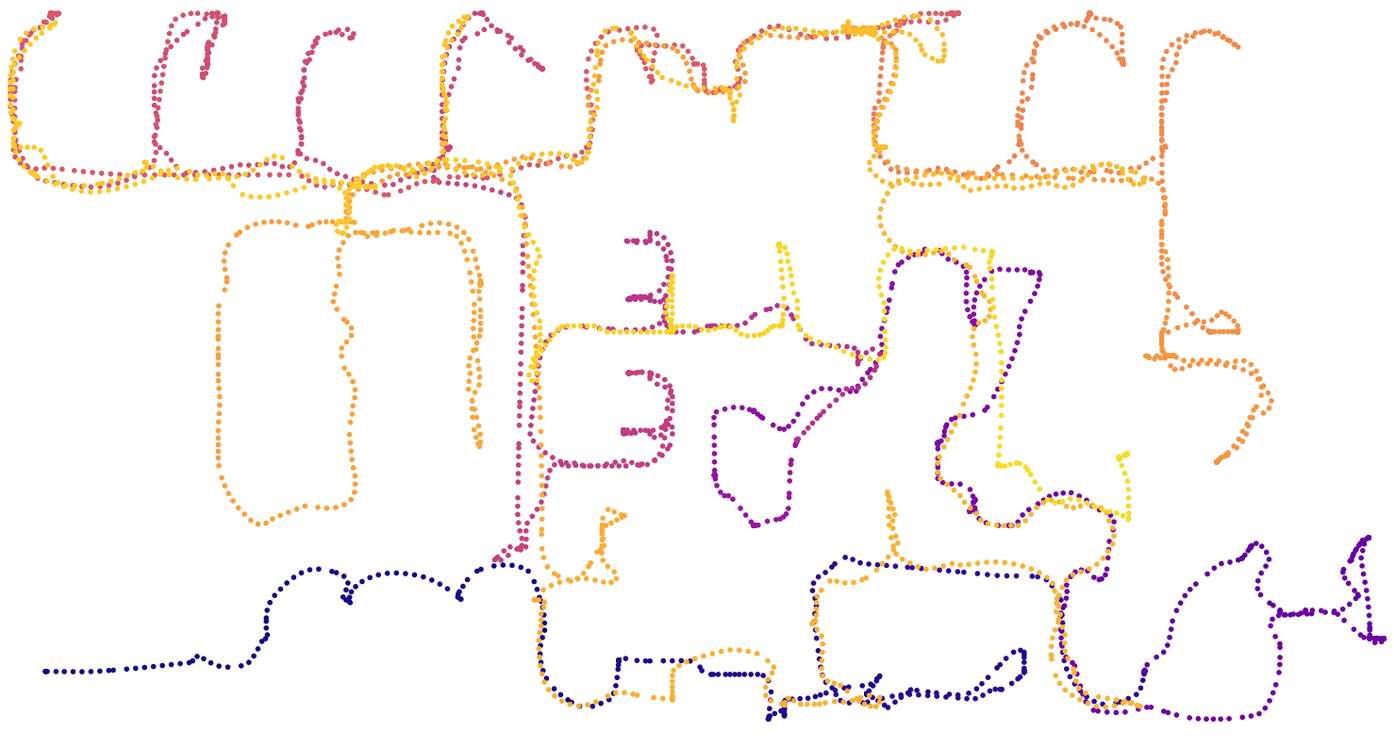
- Each participant engaged in a single-person, first-person-view, search-and-rescue task.
- The environment was a collapsed office building consisting of rooms and corridors.



- Victims were inside some of the rooms. The goal was to **maximize points** by triaging victims.
- Rescue device
  knowledge condition:
  signal (18); **no signal** (36)

Huang, L., Freeman, J., Cooke, N., et al. (2021). Using humans' theory of mind to study artificial social intelligence in Minecraft search and rescue. Technical report.

Setting: Search-and-rescue task in Minecraft



- Each participant engaged in a single-person, first-person-view, search-and-rescue task.
- The environment was a collapsed office building consisting of rooms and corridors.
- Victims were inside some of the rooms. The goal was to **maximize points** by triaging victims.
- Rescue device
  knowledge condition:
  signal (18); **no signal** (36)
- **Can we infer knowledge condition from the trajectory?**

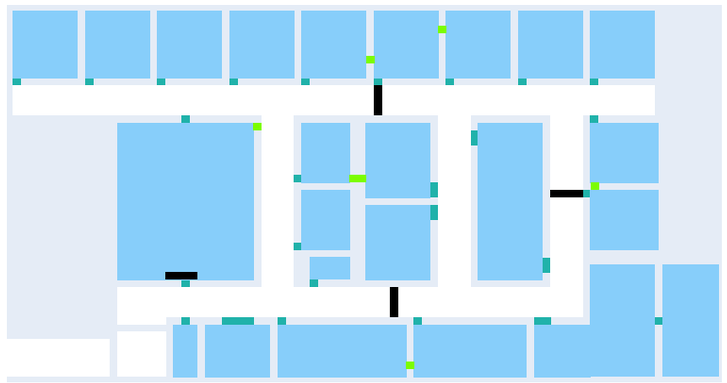
Huang, L., Freeman, J., Cooke, N., et al. (2021). Using humans' theory of mind to study artificial social intelligence in Minecraft search and rescue. Technical report.

Abstraction

Typical

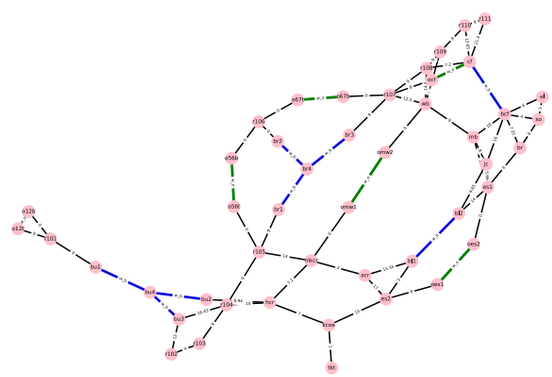
Abstracted

Spatial states



geometry

topology



Reward function

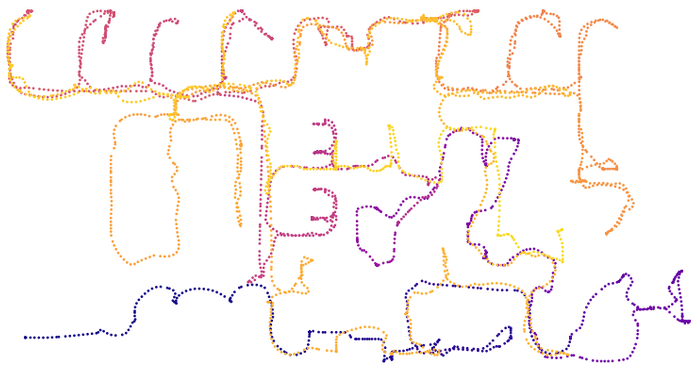
10	-1	-1	-1	x	-1
-1	-1	-1	-1	x	-1
-1	-1	-1	-1	-1	-1
-1	5	-1	-1	x	-1
x	x	x	x	x	-1
-1	-1	-1	-1	-1	-1

value

myopia

Room-based utility ≈
Expected points /
(travel time + total triage time)

Temporal states



continuous

discrete

Decision points: room exists
Decision: next room selected

Rational agent analysis

- **Rational agents:**
  - 2 models: with **signal** knowledge; with **no signal** knowledge
  - Given: a decision point & all information gathered so far
  - Compute: the utility of each room → probabilities on room choices

- **Analysis of human data:**
  - Extract all decision points (room exists) from trajectories
  - Extract the decision (room visited next) at decision points
  - Compare the player's choice with the two rational agents' probabilistic choices

- **Result:** Yes, we can infer knowledge condition from trajectories!

Truth \ Inference	No signal	Signal
No signal (36)	0.611 (22)	0.389 (14)
Signal (18)	0.278 (5)	0.722 (13)

Please see poster 2024 for more modeling detail and results.