COMP4620 – Advanced Topics in AI
Partially Observable Markov Decision Processes (POMDP) 3/3

Hanna Kurniawati

http://users.cecs.anu.edu.au/~hannakur/

hanna.kurniawati@anu.edu.au

Australian National University

RESEARCH SCHOOL OF COMPUTER SCIENCE

Topics

- ✓ Lecture 1: What is POMDPs?
- ✓ Lecture 2: How do we solve POMDPs?
- Lecture 3: Applications & What's Next

Applications & What's Next

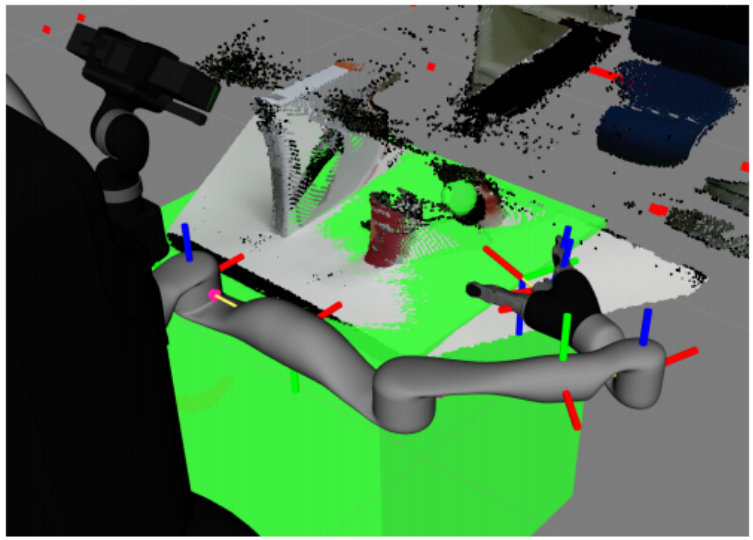
- Different flavours of partial observability in decision making
  - Planning & control problems
  - When influential parameters are not initially known
- POMDP for model generation with limited data
- What's next

A typical example

Somewhat similar problem but Not a typical example

S: 7D continuous (robot's joints) X binary (hand close/open)
A: Increase/decrease of joint angles + hand close/open
O: Position of the cup

Somewhat similar problem but Not a typical example



The world as seen by the robot

Applications & What's Next

- Different flavours of partial observability in decision making
  - ✓Planning & control problems
  - When influential parameters are not initially known
- POMDP for model generation with limited data
- What's next

Making Decisions with Initially Unknown Parameters

- In many problems, there's hidden variables important to make decisions but difficult to assess a priori, e.g.:
  - Robot manipulation: Coefficient of friction
  - Human Robot Interaction/Collaboration: Human intention, human characteristic (e.g., how risk averse)
  - Autonomous pen-testing: How fast the defender patch the n/w?
- In general, this means the effects of actions depend on variables that are not fully known but can be learned from observations

POMDP for Reinforcement Learning (RL)
aka. Bayesian Reinforcement Learning

RL: MDP with missing components

- State space (S_{MDP})
- Action space (A_{MDP})
- Transition function (T_{MDP})
- Reward function (R_{MDP})

Not known

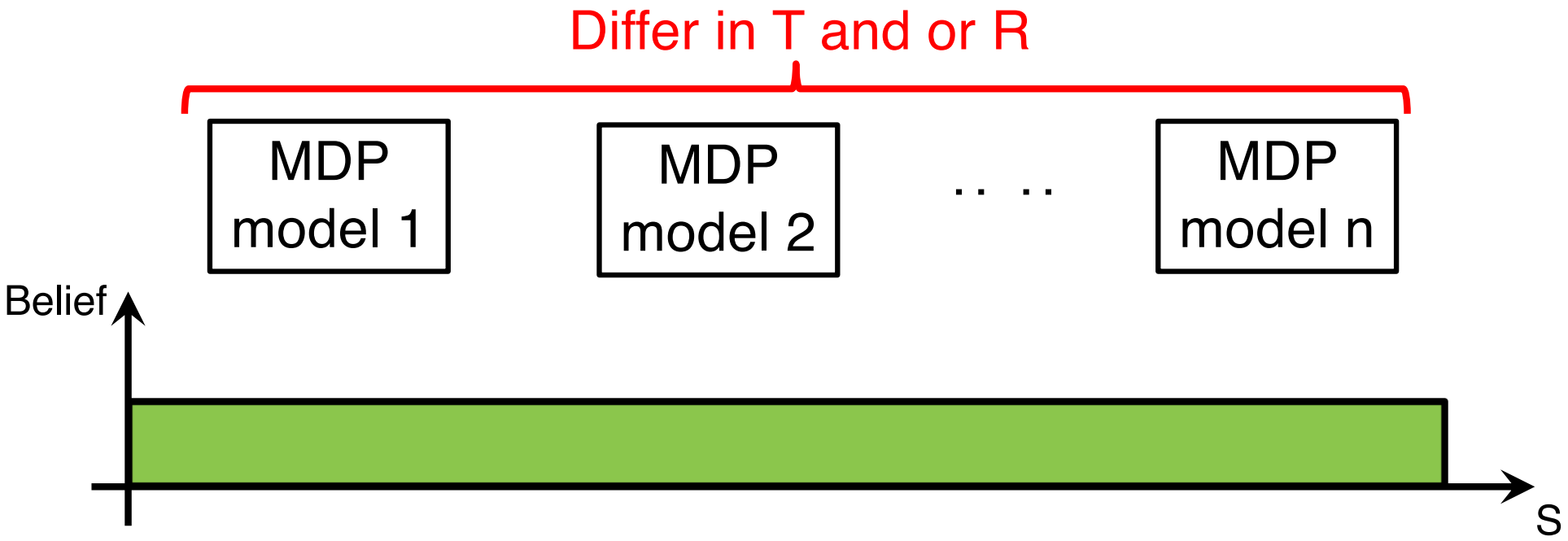
Construct a POMDP

- Where the states are MDP states X parameters of the T_{MDP} & R_{MDP}
- Essentially, partial observability on which MDP model is the right model
- A, T, R follows from the particular MDP model
- O & Z are observations & observation function about which MDP model is correct

Bayesian RL

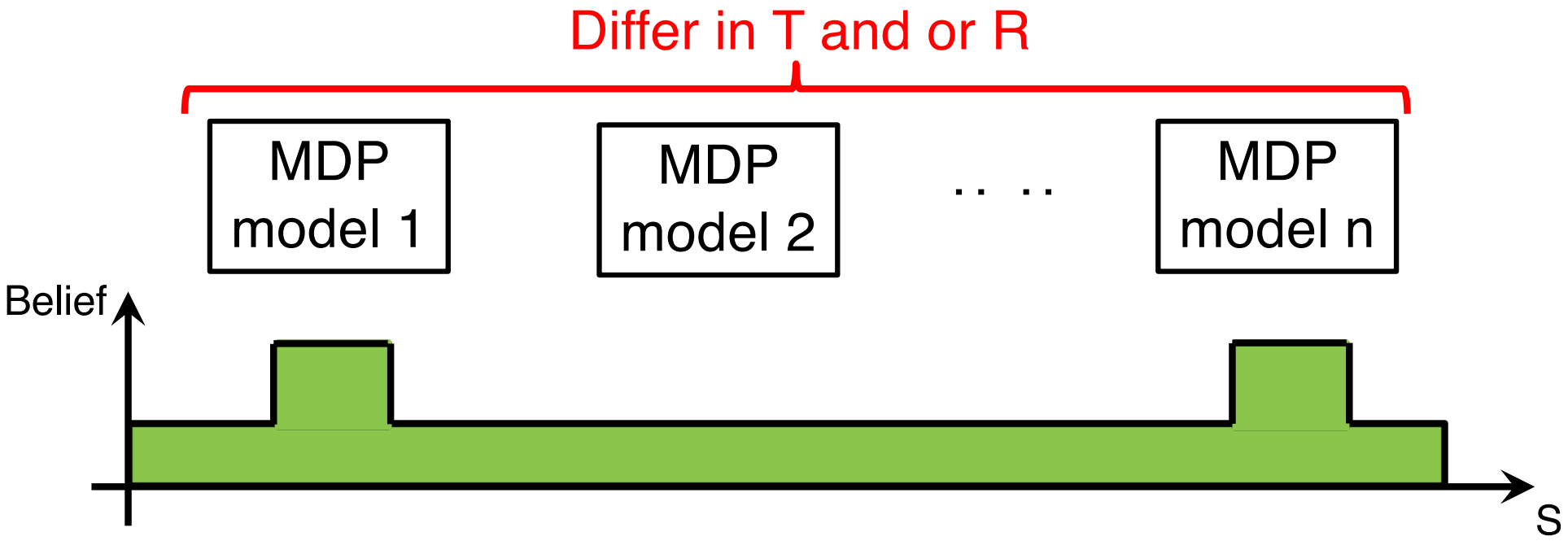
POMDP for Reinforcement Learning (RL)
aka. Bayesian Reinforcement Learning

- Beliefs in POMDP becomes distribution over the possible MDP models



POMDP for Reinforcement Learning (RL)
aka. Bayesian Reinforcement Learning

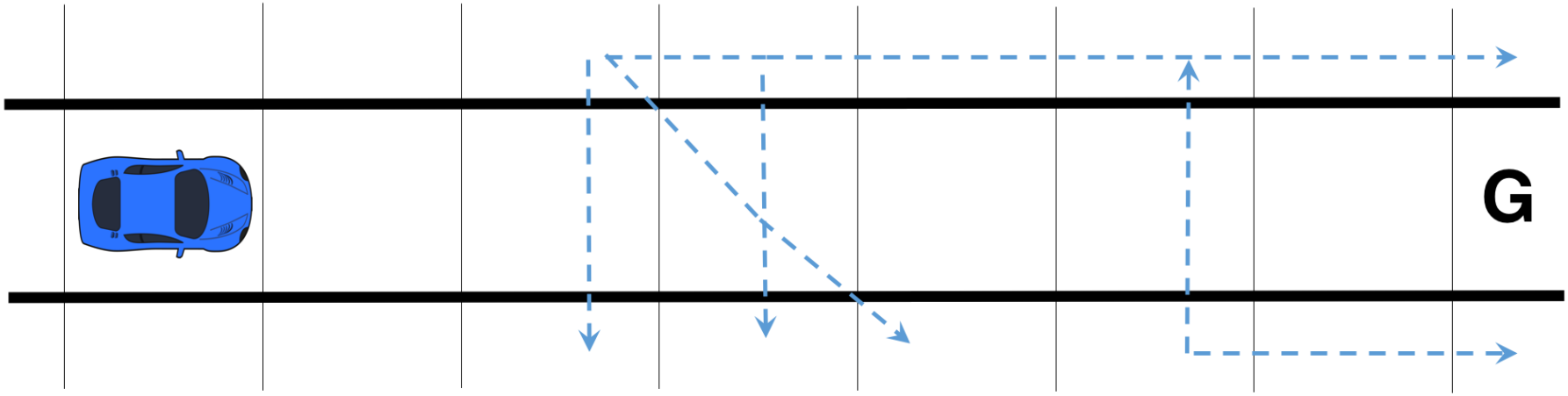
- Belief update means updates in the agent's understanding on which model is correct



POMDP automatically balances the trade-off between generating accurate model and to achieve the goal

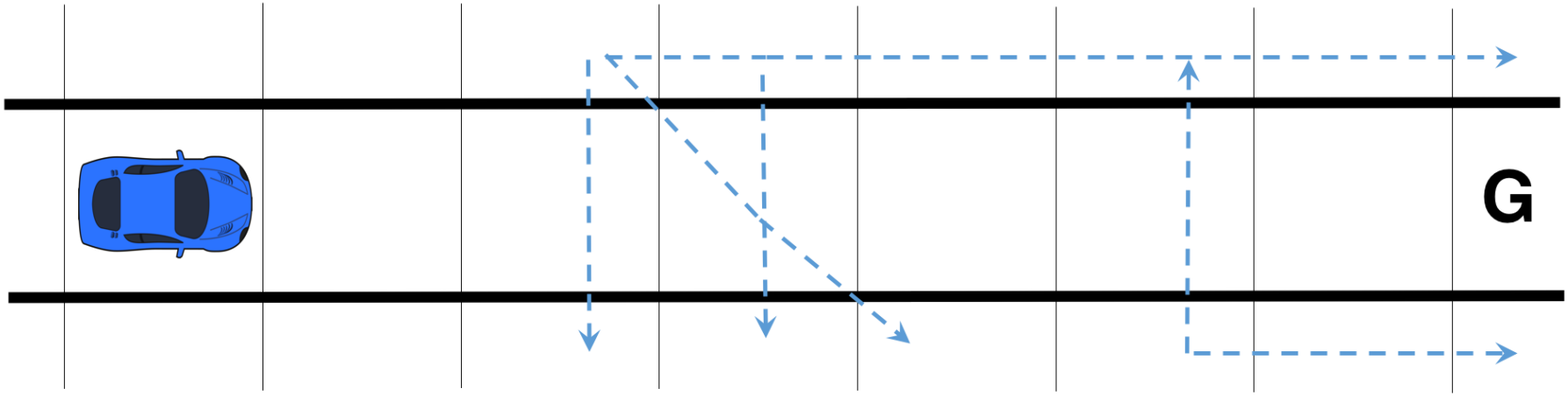
In many cases, can achieve the goal without knowing the exact model

Example: Pedestrian Avoidance



- Suppose a self-driving car has good enough sensors to identify its own position and the pedestrians' positions
- However, pedestrians' intentions are not known before hand
- How should the car behave to reach its goal without colliding with the pedestrians?

Example: Pedestrian Avoidance



Safety?

- If we see RL as learning by doing, then RL for pedestrian avoidance seems like a dangerous idea
- But, if we see RL from its more formal definition, i.e., MDP with missing transition and/or reward functions, then applying RL for pedestrian avoidance might still be okay
  - Bayes RL/POMDP allows us to explore carefully by setting conservative reward function and initial beliefs on the intention of the pedestrian
  - The lookahead component of POMDP helps a lot in ensuring safety

Notes

- Depending on the POMDP solver used, Bayesian RL can provide some guarantee on performance
- But, when many of the parameters are unknown, we need to solve huge POMDP problems
  - Too huge for even current state of the art, despite advances in POMDP over the past decade
  - Sometimes, we can get away by better problem modelling to reduce the number of parameters to learn
    - Example: In tutorial on Autonomous pen-testing

Applications & What's Next

- ✓ Different flavours of partial observability in decision making
  - ✓ Planning & control problems
  - ✓ When influential parameters are not initially known
- POMDP for model generation with limited data
- What's next

POMDP for Model Building

- POMDP can be used to compensate the lack of observational data with imperfect domain knowledge
- The above will then help reduce data requirement in generating predictive models
  - Can help a lot in when data is expensive / difficult to get

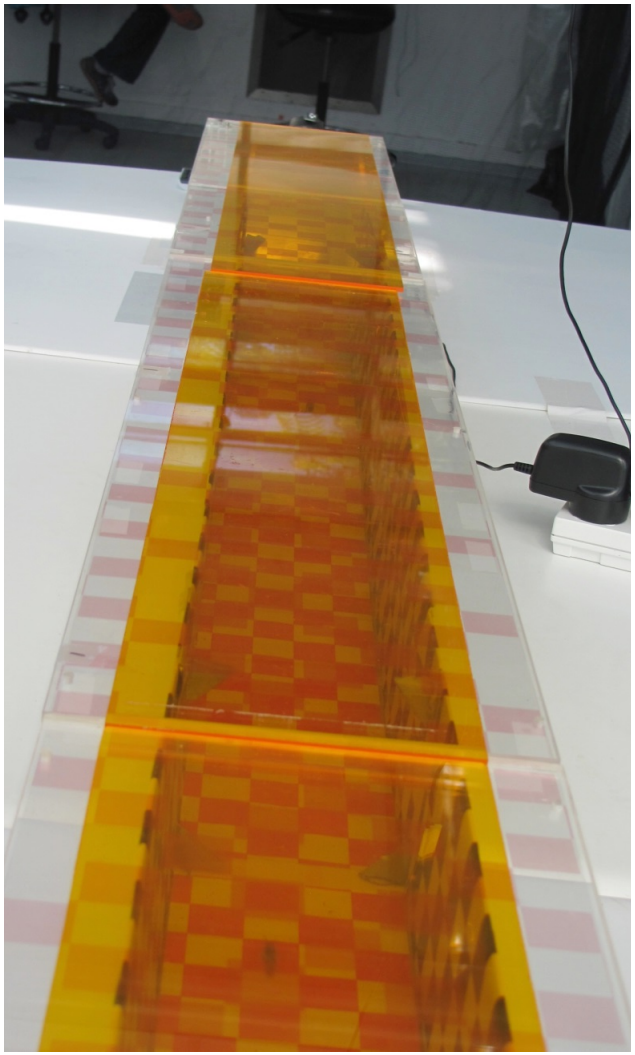
Example: Collision Avoidance Strategies of Honeybees

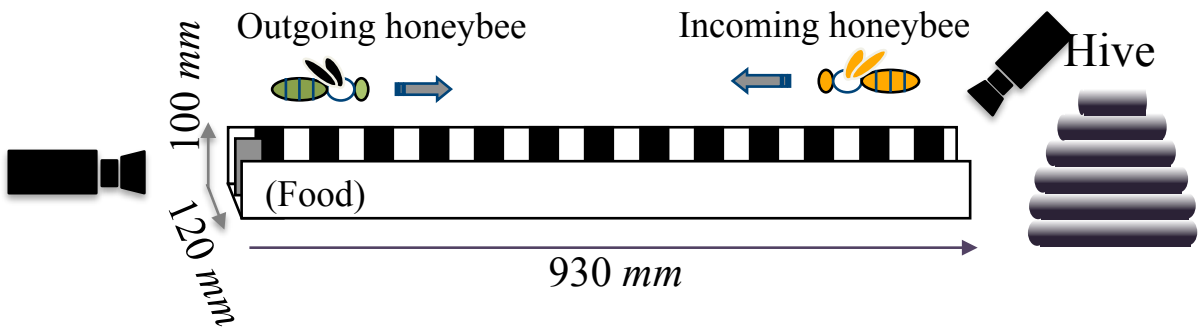
How do they avoid mid-air collision?

How do bees avoid collision?

- Current view: [M.D. Breed and J. Moore. 2012. Animal Behavior.]
  - Animal behavior optimizes certain criteria
  - The question is what criteria is being optimized

How to get the data?



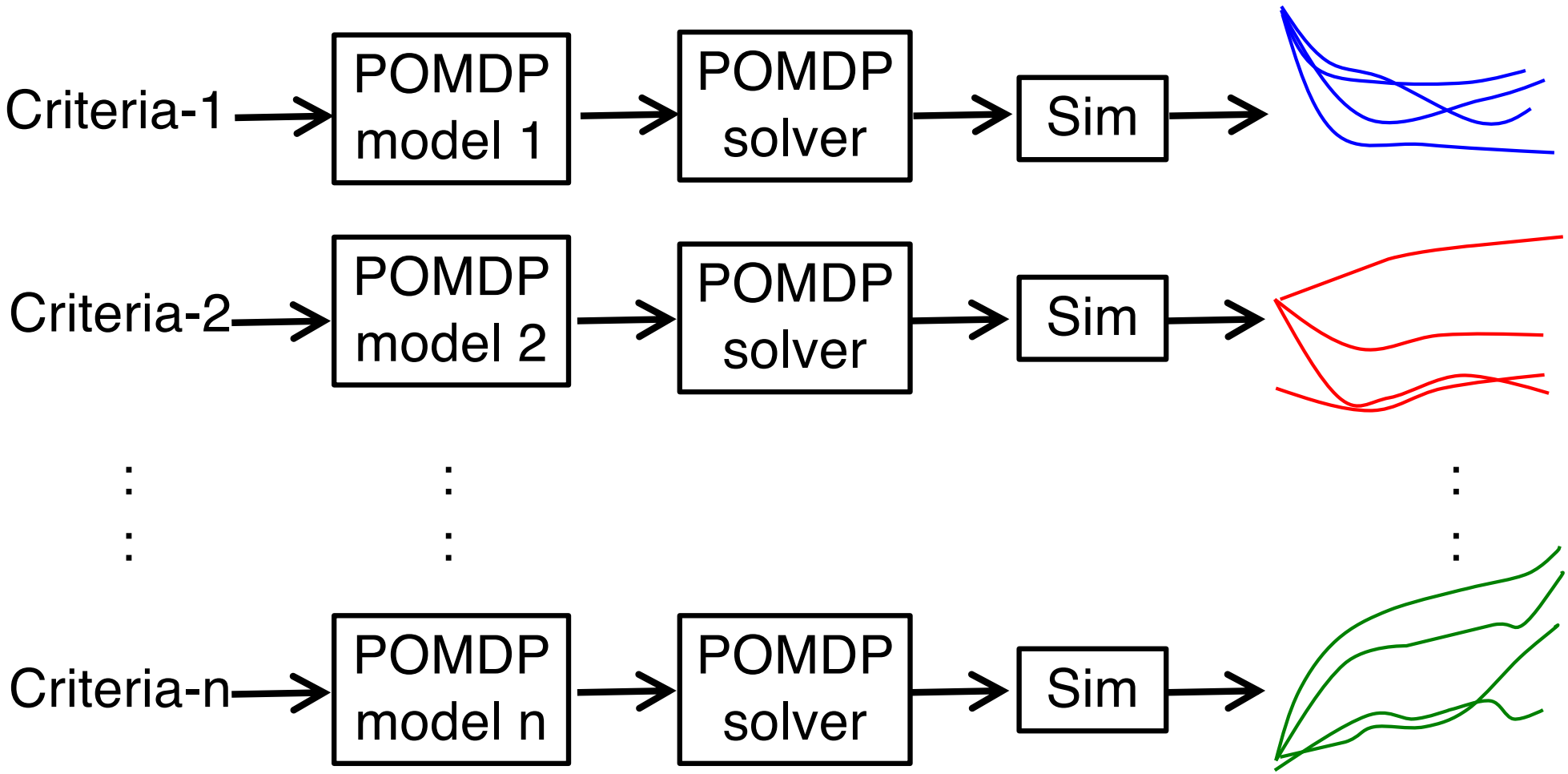


Not really easy to get millions of data

POMDP can be used to rank which criteria are likely to generate the available trajectory data, without knowing the exact properties of the bees (e.g., their mass, wing-span, etc.).

H. Wang, H. Kurniawati, S.P.N. Singh, and M.V. Srinivasan. In-silico Behavior Discovery System: An Application of Planning in Ethology. ICAPS 2015. (Outstanding Student Paper Award)

A Hypothesis Ranking System



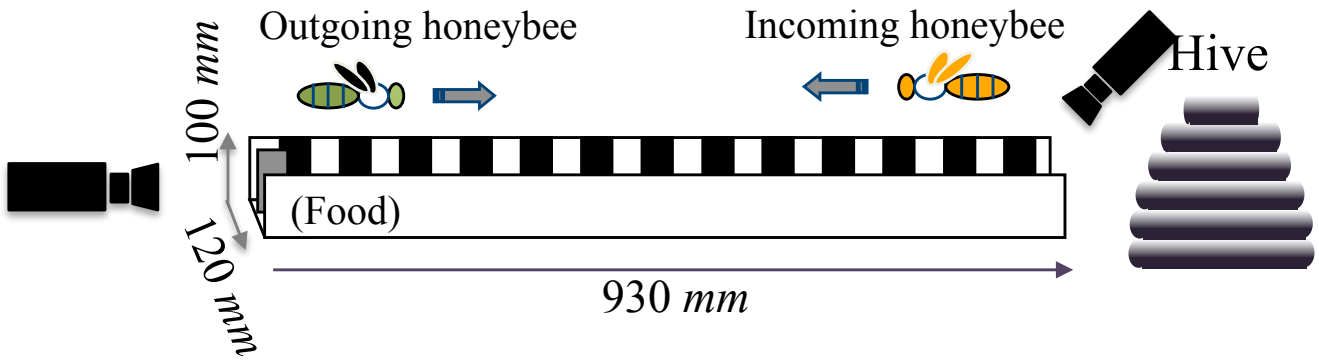
Rank the criteria based on how similar the simulated trajectory is to the (limited) experimental data

How to generate the hypothesis reward functions?

- Domain knowledge helps identify the variables, though not the values
- Optimization criteria that biologists believe to be TRUE
  1. Bees fly toward a destination (not random)
  2. Bees prefer to fly as close as possible to the middle of the tunnel horizontally (because they use optical flow for navigation)

Components of reward function

1. Cost: Collision cost + Movement cost
2. Reaching destination reward
3. Horizontal centering: Penalty proportional to the horizontal deviation from the central area
4. Vertical centering: Penalty proportional to the vertical deviation from the central area



Testing

Hypothesis	Rank
Cost	6
Cost + Destination reward	2
Cost + Horizontal centering	3
Cost + Vertical centering	5
Cost + Destination reward + Horizontal centering	**1**
Cost + Horizontal centering + Vertical centering	4

Correctly rank phototaxis behavior (no vertical centering) + horizontal centering at the top of the bees' behaviour

Applications & What's Next

- ✓ Different flavours of partial observability in decision making
  - ✓ Planning & control problems
  - ✓ When influential parameters are not initially known
- ✓ POMDP for model generation with limited data
- What's next

What's Next?

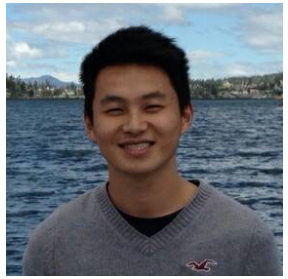
- Extend to multi-agent where different agents may have different objectives
  - Partially Observable Stochastic Games
  - Applications: Assurance of autonomous systems
- What to do when transition, observation, and reward functions are not known in advance?
  - POMDP structure helps learning methods to be more data efficient and generalize better
  - Applications: Robust manipulation
- Available projects at all level
  - E-mail me, projects may not be advertised

Robust Decision Making & Learning Lab

@CSIT bld, level 3, room N323






Applications & What's Next

- ✓ Different flavours of partial observability in decision making
  - ✓ Planning & control problems
  - ✓ When influential parameters are not initially known
- ✓ POMDP for model generation with limited data
- ✓ What's next

Topics

- ✓ Lecture 1: What is POMDPs?
- ✓ Lecture 2: How do we solve POMDPs?
- ✓ Lecture 3: Applications & What's Next

- POMDPs:
  - Not that complicated
  - Are now relatively practical for robust decision making in many realistic problems
  - Has a lot of applications