Information Theoretic Question Asking to Improve Spatial Semantic Representations

Matthew Walter
Toyota Technological Institute at Chicago

AAAI Fall Symposium
November 13, 2014

Collaborators

Sachi Hemachandra

Seth Teller

Robots as Our Partners

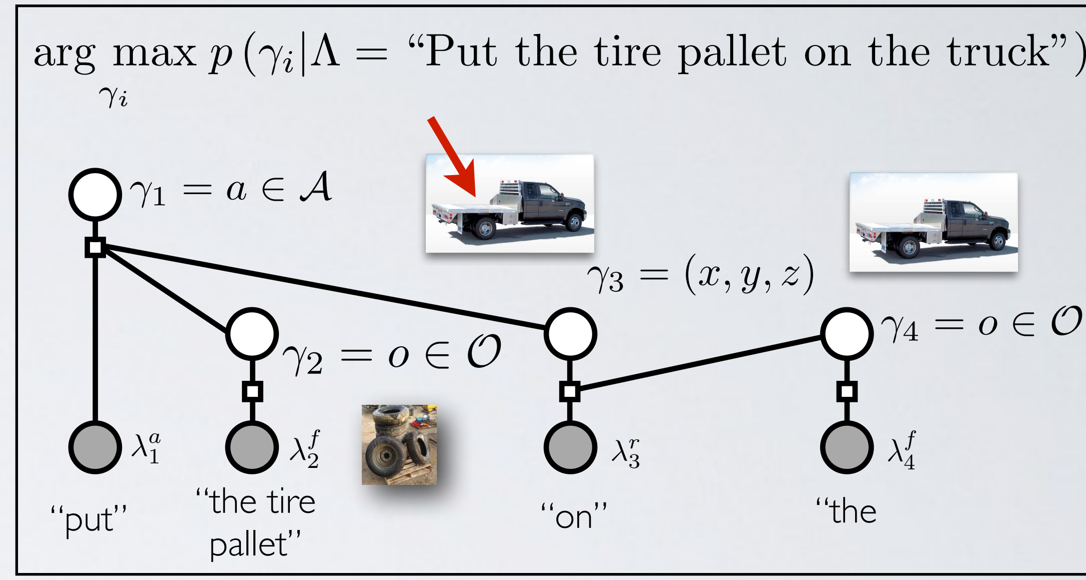
Now: People Accommodate Robots

Courtesy: Kinova Robotics

Where We Need to Be

Natural Language Understanding for Robots

- Knowledge-based map to formal logic [1-4]
  - Exploit structure of language
  - Fixed action space
  - Limited learning



- Statistical-based “Symbol Grounding”
  - Parse language into formal action specifications [6-8]
  - Ground language in physical referents (objects, places, paths, events) [9]
  - Parser and groundings are **learned**

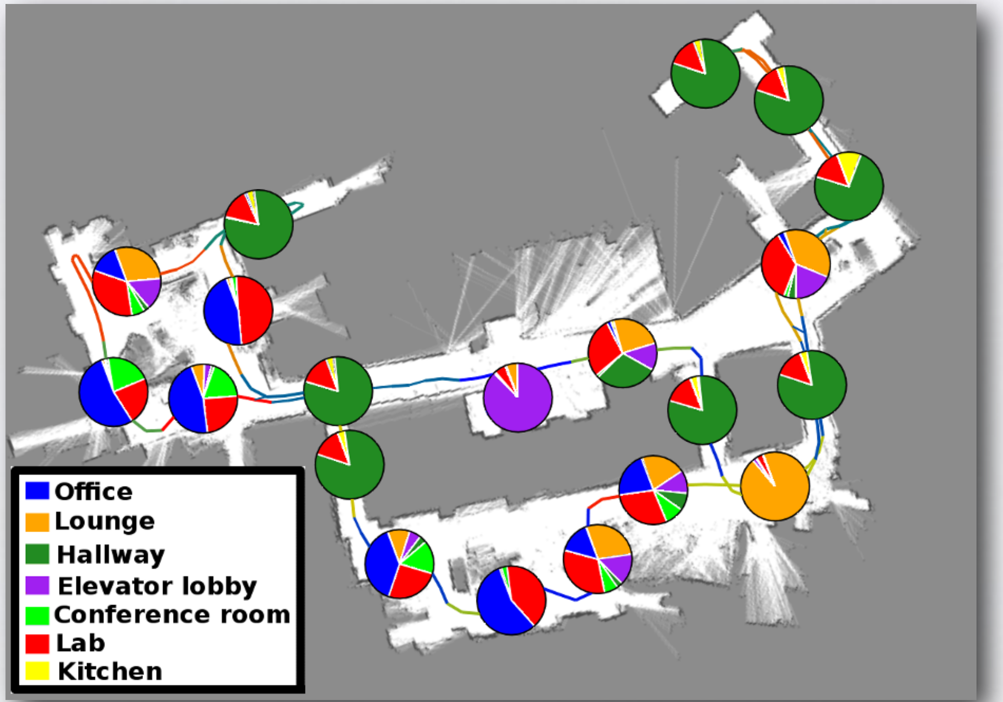
[1] Winograd 1971
[2] MacMahon et al., 2006
[3] Kress-Gazit et al., 2008
[4] Dzifcak et al., 2009
[6] Matuszek et al., 2010
[5] Shimizu & Hass, 2009
[7] Chen et al., 2011
[8] Matuszek et al., 2012
[9] Tellex et al., 2011

NLU as Probabilistic Inference

robot actions locations command

$$\arg\max_{\gamma_i} p(\gamma_o, \gamma_a, \gamma_r, \gamma_p | S, \Lambda)$$

objects relations semantic map



I. Introduction

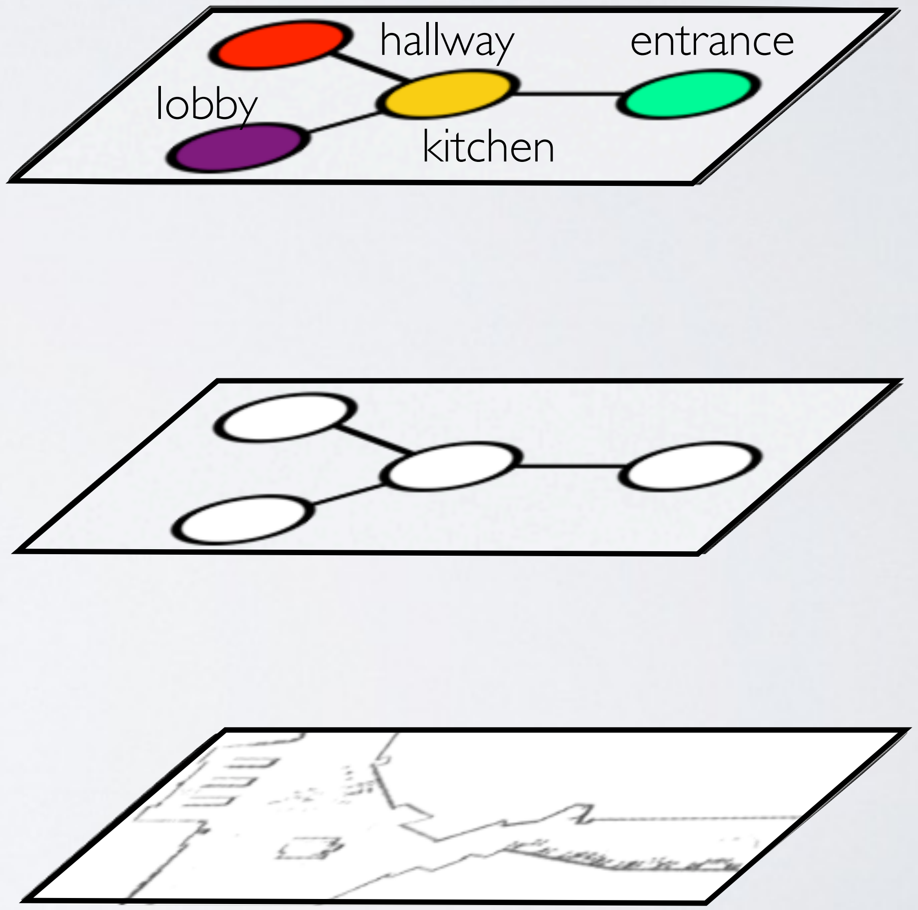
II. Learning Semantic Maps from Natural Language Dialogue

III. Following Directions Without in Unknown Environments

IV. Future Directions & Conclusions

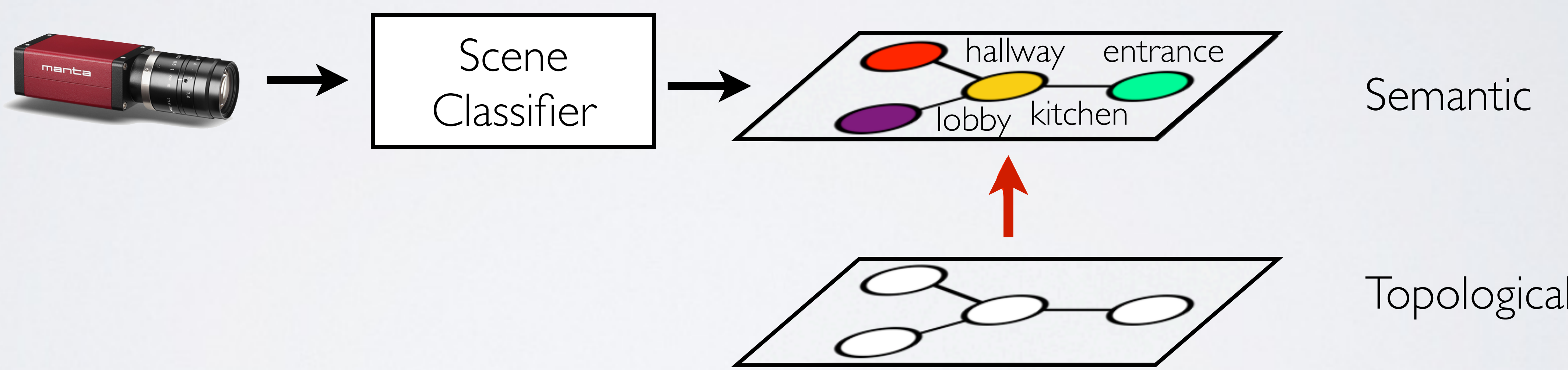
Rich Cognitive Models of Space

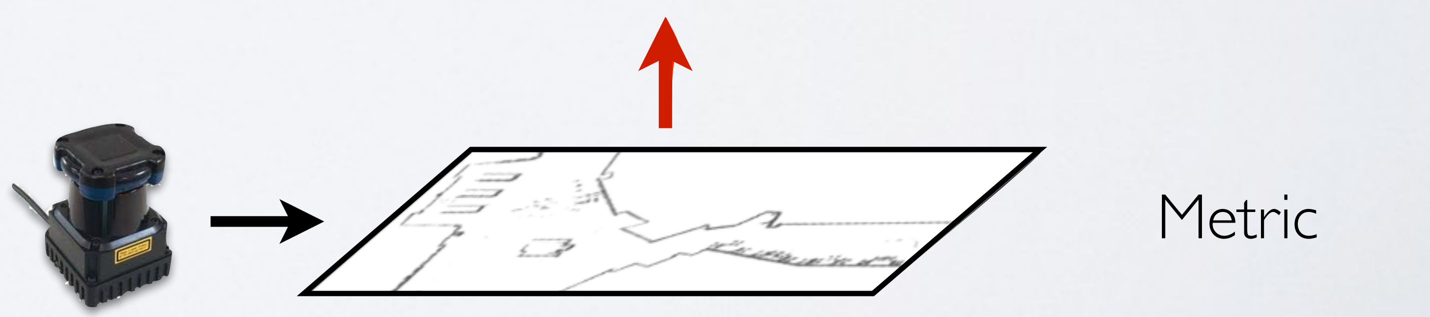
- Formulate human-centric models of the environment
- Models should express:
  - Regional decomposition of space
  - Metric pose (relative or absolute)
  - Connectivity
  - Regions' (room) types
  - Regions' colloquial names
- Models often constructed by hand



State-of-the-Art in Semantic Mapping

- Spatial Semantic Hierarchy [1]
- Augment SLAM map with topological and semantic layers
  - Incorporate scene classification and object detection [2,3]
  - Information flows up from the metric layer, not down





[1] Kuipers, 2000

[2] Zender et al., 2008

[3] Pronobis et al., 2010

Limitations of Semantic Mapping Algorithms

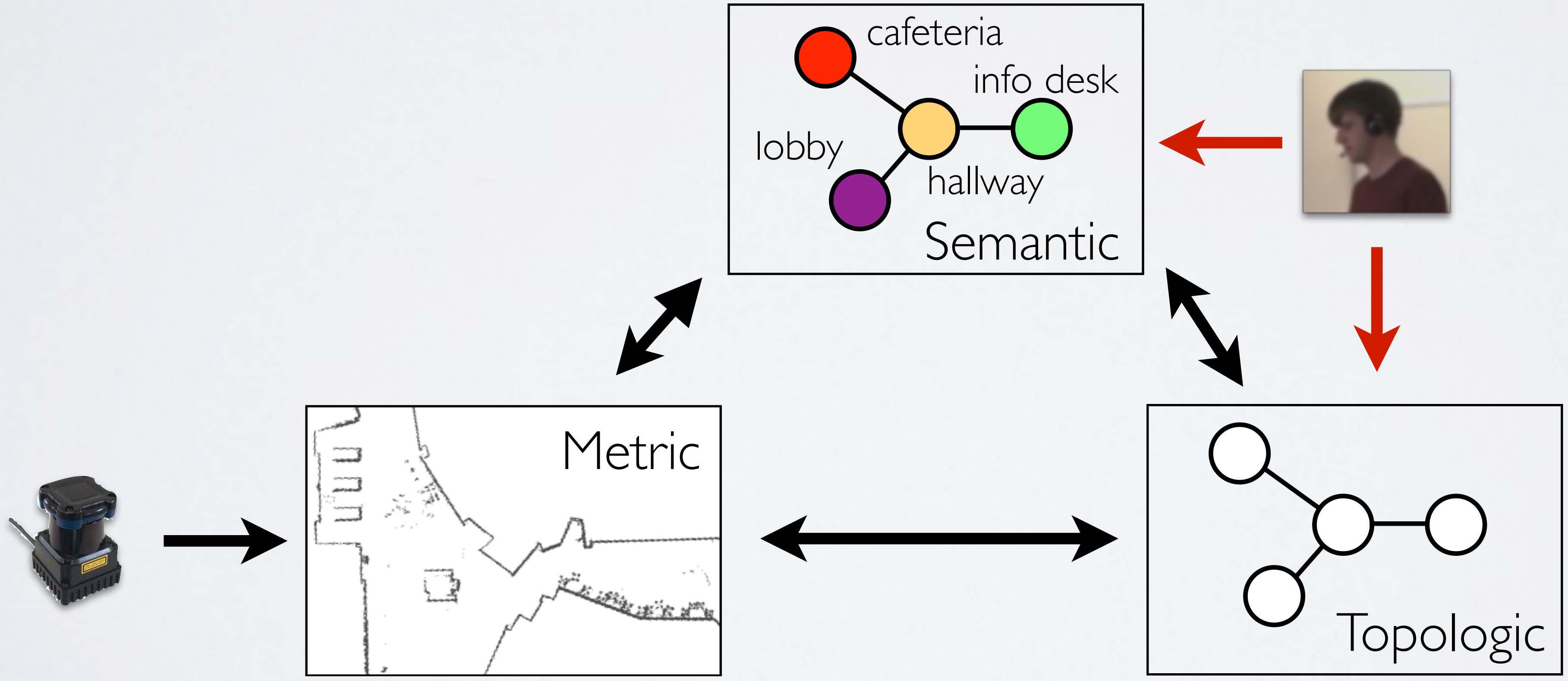
- Rely upon pre-trained classifiers
- Limit generalizability beyond trained envs.
- Restrict to robot's immediate surround
- Require that robot visits each region
- Unable to infer certain properties:
  - Colloquial names
  - Unique objects

Building Semantic Maps with Natural Language

- People can efficiently convey information through speech
- Learn semantic information from natural language descriptions:
  - Colloquial names
  - Room type
  - Spatial relations
- “Observe” beyond robot’s FOV
- Fuse with robot’s sensor stream (i.e., hard & soft information)

Building Semantic Maps with Natural Language

- Learn semantic cues and spatial relations from user's descriptions
- Interpret free-form utterances
- Fully integrate linguistic information



[RSS 2013; ICRA 2014; IJRR 2014 (submitted)]

Challenges to Learning from Natural Language

- Language and sensor streams are uncertain
  - Descriptions are ambiguous
  - Sensor data is noisy
- Language and sensor streams are disparate
  - Language conveys abstract concepts
  - Sensors provide metric observations
- Mapping requires fusing this information

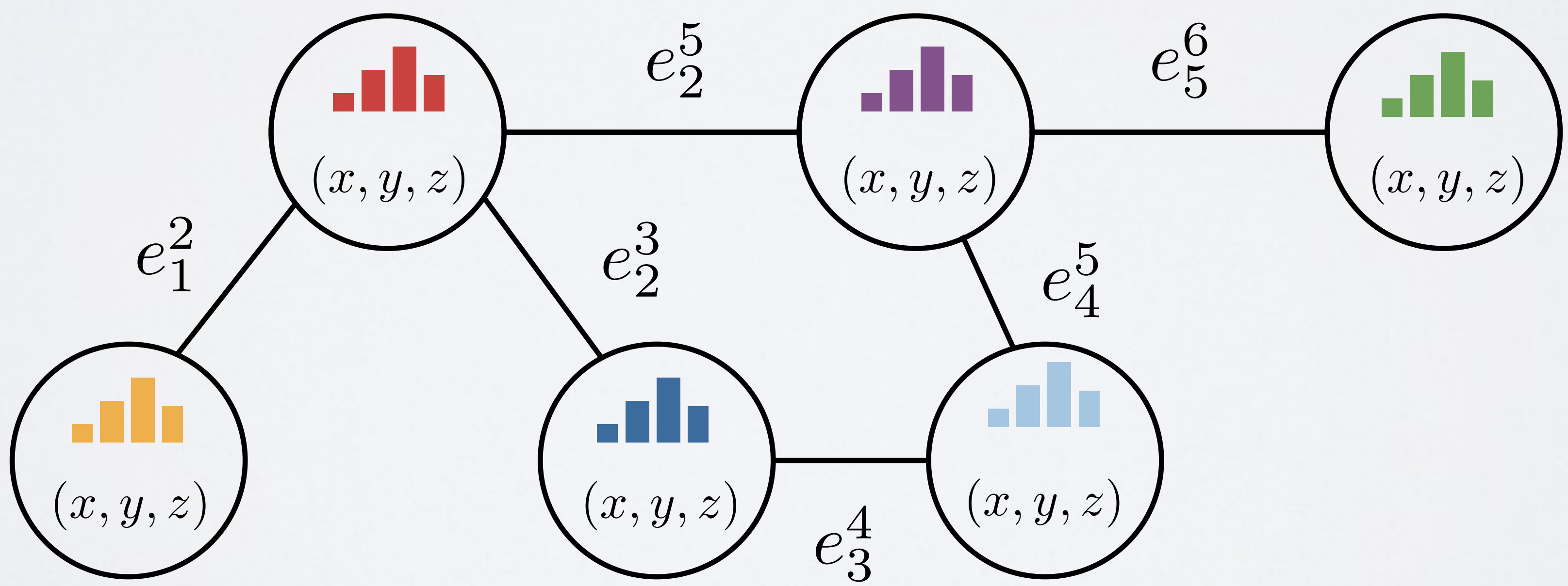
Model: Posterior over Semantic Graphs

Sensor stream Odometry Language

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology $G_t = (V_t, E_t)$ Semantic labels and type

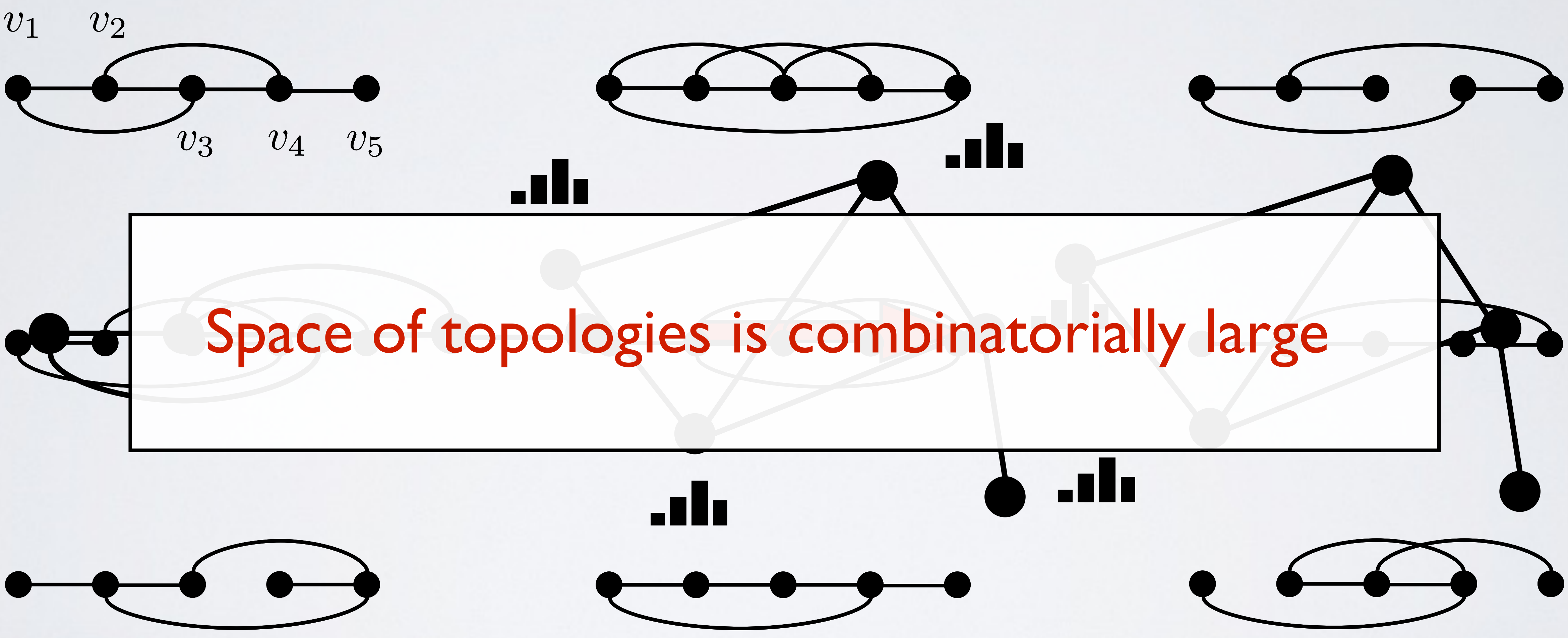
Vertex poses



[RSS 2013; ICRA 2014; IJRR 2014]

Factoring the Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t)\ p(X_t | G_t, z^t, u^t, \lambda^t)\ p(G_t | z^t, u^t, \lambda^t)$$



Space of topologies is combinatorially large

[RSS 2013; ICRA 2014; IJRR 2014]

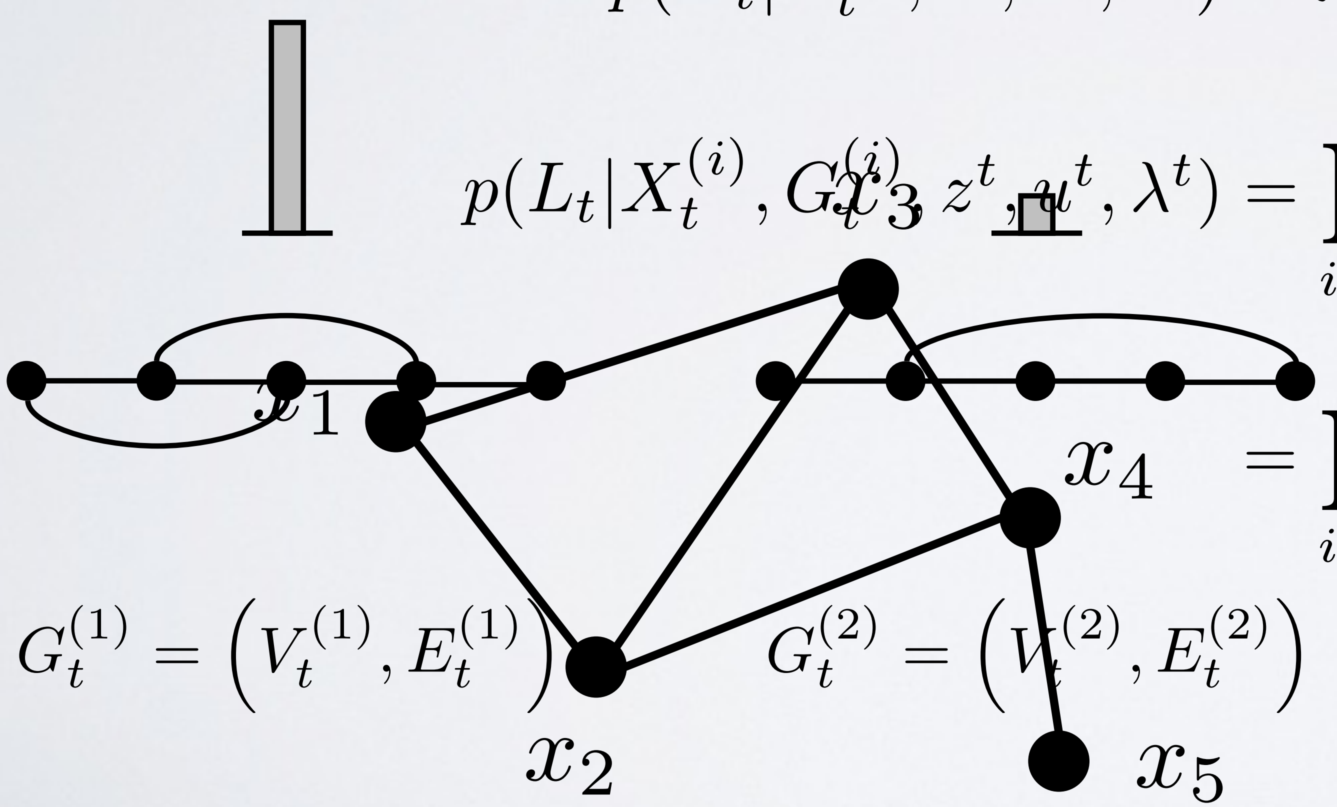
Model: Posterior over Semantic Graphs

$p(G_t, X_t, L_t \mid z^t, u^t, \lambda^t) =$	$p(L_t \mid X_t, G_t, z^t, u^t, \lambda^t)$	$p(X_t \mid G_t, z^t, u^t, \lambda^t)$	$p(G_t \mid z^t, u^t, \lambda^t)$
	Dirichlet	Gaussian (information form)	Sample-based representation

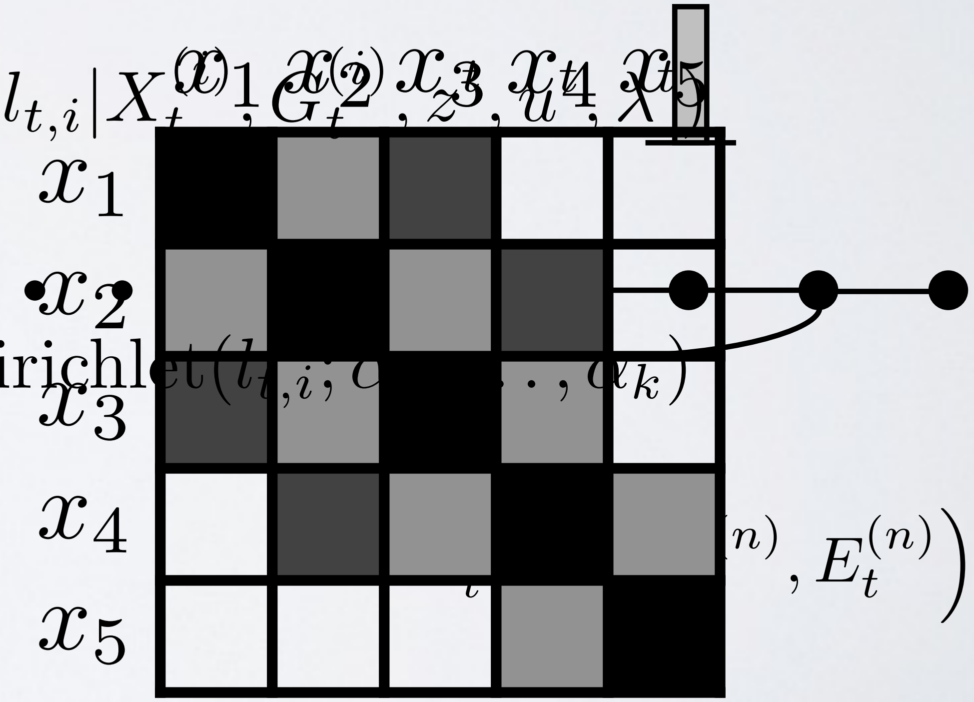
$$p(X_t \mid G_t^{(i)}, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t)$$

$$p(L_t \mid X_t^{(i)}, G_t^{(i)}, z^t, u^t, \lambda^t) = \prod_{i=1}^{t} p(l_{t,i} \mid X_t^{(i)}, G_t^{(i)}, z^t, u^t, \lambda^t)$$

$$= \prod_{i=1}^{t} \mathrm{Dirichlet}(l_{t,i}; \alpha_1, \ldots, \alpha_k)$$



$G_t^{(1)} = \left(V_t^{(1)}, E_t^{(1)}\right)$ $G_t^{(2)} = \left(V_t^{(2)}, E_t^{(2)}\right)$ $G_t^{(n)} = \left(V_t^{(n)}, E_t^{(n)}\right)$



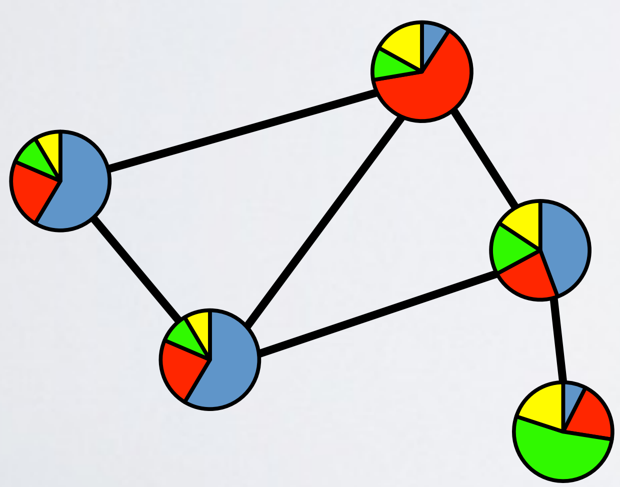
[RSS 2013; ICRA 2014; IJRR 2014]

Model: Posterior over Semantic Graphs

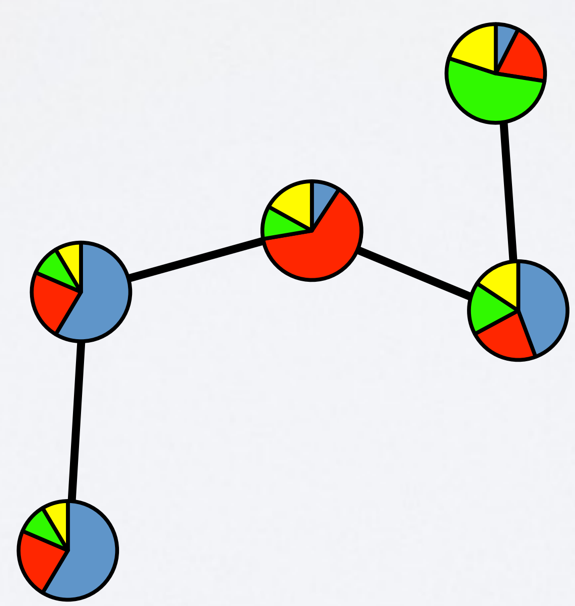
$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t)\ p(X_t | G_t, z^t, u^t, \lambda^t)\ p(G_t | z^t, u^t, \lambda^t)$$

$$\mathcal{P}_t = \left\{ P_t^{(1)}, P_t^{(2)}, \ldots, P_t^{(n)} \right\}$$

$$P_t^{(1)} = \left\{ G_t^{(1)}, X_t^{(1)}, L_t^{(1)}, w_t^{(1)} \right\}$$

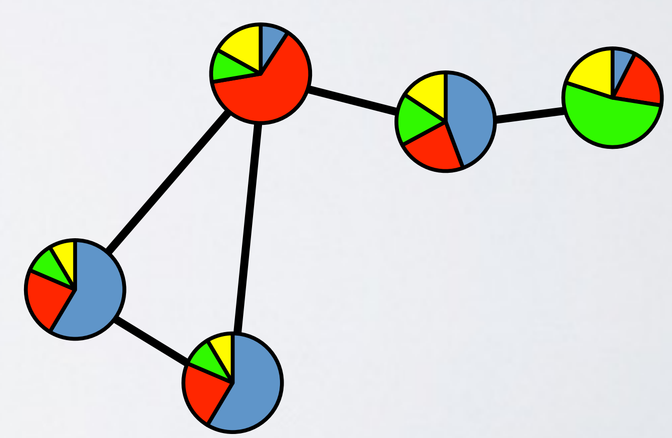


$$P_t^{(2)} = \left\{ G_t^{(2)}, X_t^{(2)}, L_t^{(2)}, w_t^{(2)} \right\}$$



. . .

$$P_t^{(n)} = \left\{ G_t^{(n)}, X_t^{(n)}, L_t^{(n)}, w_t^{(n)} \right\}$$



[RSS 2013; ICRA 2014; IJRR 2014]

Rao-Blackwellized Particle Filter

Input: $\mathcal{P}_{t-1} = \left\{ P_{t-1}^{(1)}, P_{t-1}^{(2)}, \ldots, P_{t-1}^{(n)} \right\}$ where $P_{t-1}^{(i)} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)}, w_{t-1}^{(i)} \right\}$

for each particle i

1) **Proposal**: Modify the topology based on metric and semantic maps

2) **Update**: Perform Bayesian update of Gaussian

3) **Update**: Update Dirichlet over labels based on language

4) **Reweight**: Update weights based on metric observations

Return: $\mathcal{P}_{t} = \left\{ P_{t}^{(1)}, P_{t}^{(2)}, \ldots, P_{t}^{(n)} \right\}$ where $P_{t}^{(i)} = \left\{ G_{t}^{(i)}, X_{t}^{(i)}, L_{t}^{(i)}, w_{t}^{(i)} \right\}$

[RSS 2013; ICRA 2014; IJRR 2014]

Rao-Blackwellized Particle Filter

Input: $\mathcal{P}_{t-1} = \left\{P_{t-1}^{(1)}, P_{t-1}^{(2)}, \ldots, P_{t-1}^{(n)}\right\}$ where $P_{t-1}^{(i)} = \left\{G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$

for each particle i

1) **Proposal**: Modify the topology based on metric and semantic maps

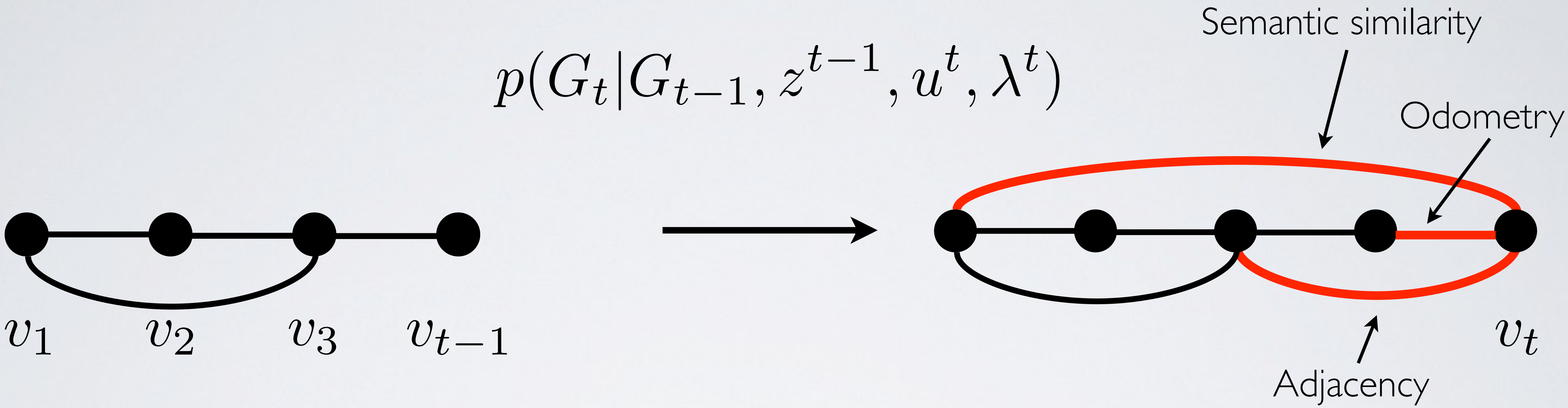
2) **Update**: Perform Bayesian update of Gaussian

3) **Update**: Update Dirichlet over labels based on language

4) **Reweight**: Update weights based on metric observations

Return: $\mathcal{P}_{t} = \left\{P_{t}^{(1)}, P_{t}^{(2)}, \ldots, P_{t}^{(n)}\right\}$ where $P_{t}^{(i)} = \left\{G_{t}^{(i)}, X_{t}^{(i)}, L_{t}^{(i)}, w_{t}^{(i)}\right\}$

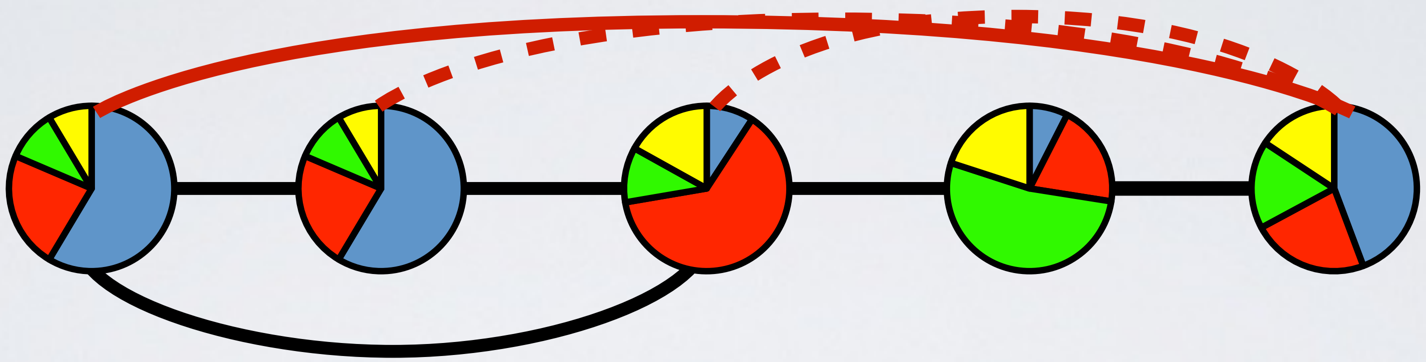
Proposal Distribution: Graph Augmentation



Propose two types of edges expressing collocation:

- Spatial-based edges
- Semantic-based edges

Proposal Distribution: Semantic Map-based Edges



Edges to current node

$$p_s(G_t | G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j: e_{tj} \notin E^-} p(G_t^{tj} | G_t^-, \lambda_t)$$

Assume edges are independent

$$\approx \prod_{j: e_{tj} \notin E^-} \sum_{l_t^-, l_j^-} p(G_t^{tj} | l_t^-, l_j^-, G_t^-) p(l_t^-, l_j^- | G_t^-)$$

Labels for node pair

Cosine similarity

Rao-Blackwellized Particle Filter

Input: $\mathcal{P}_{t-1} = \left\{ P_{t-1}^{(1)}, P_{t-1}^{(2)}, \ldots, P_{t-1}^{(n)} \right\}$ where $P_{t-1}^{(i)} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)}, w_{t-1}^{(i)} \right\}$

for each particle i

1) **Proposal**: Modify the topology based on metric and semantic maps

2) **Update**: Perform Bayesian update of Gaussian

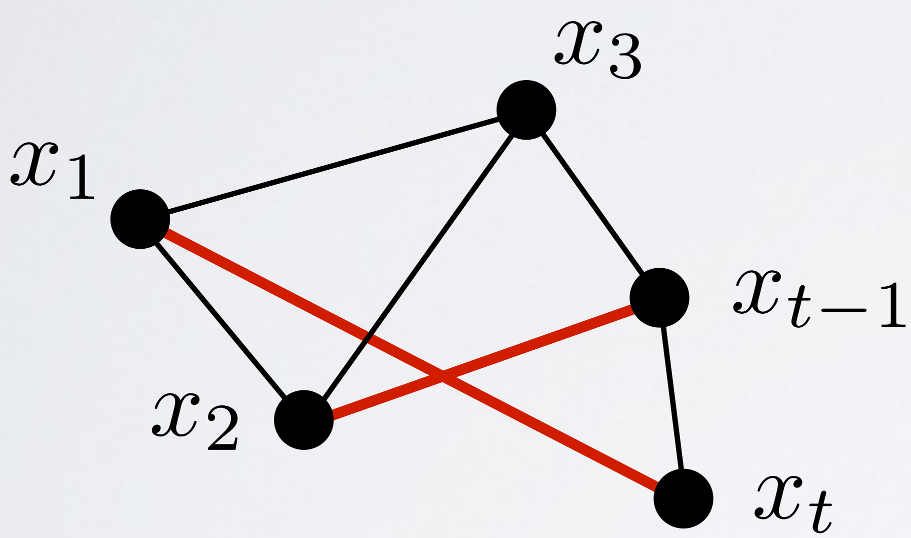
3) **Update**: Update Dirichlet over labels based on language

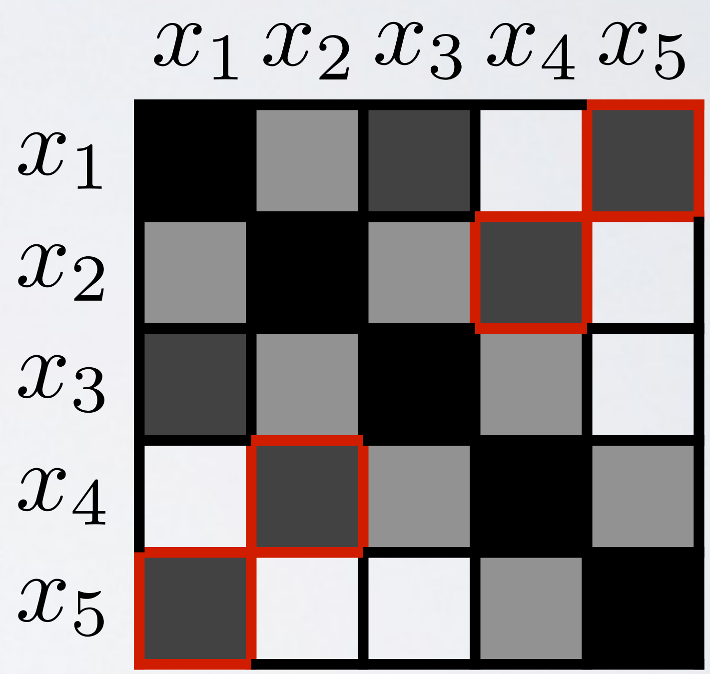
4) **Reweight**: Update weights based on metric observations

Return: $\mathcal{P}_{t} = \left\{ P_{t}^{(1)}, P_{t}^{(2)}, \ldots, P_{t}^{(n)} \right\}$ where $P_{t}^{(i)} = \left\{ G_{t}^{(i)}, X_{t}^{(i)}, L_{t}^{(i)}, w_{t}^{(i)} \right\}$

Gaussian Update

$$p(X_t | G_t, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t)$$





Rao-Blackwellized Particle Filter

Input: $\mathcal{P}_{t-1} = \left\{ P_{t-1}^{(1)}, P_{t-1}^{(2)}, \ldots, P_{t-1}^{(n)} \right\}$ where $P_{t-1}^{(i)} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)}, w_{t-1}^{(i)} \right\}$

for each particle i

1) **Proposal**: Modify the topology based on metric and semantic maps

2) **Update**: Perform Bayesian update of Gaussian

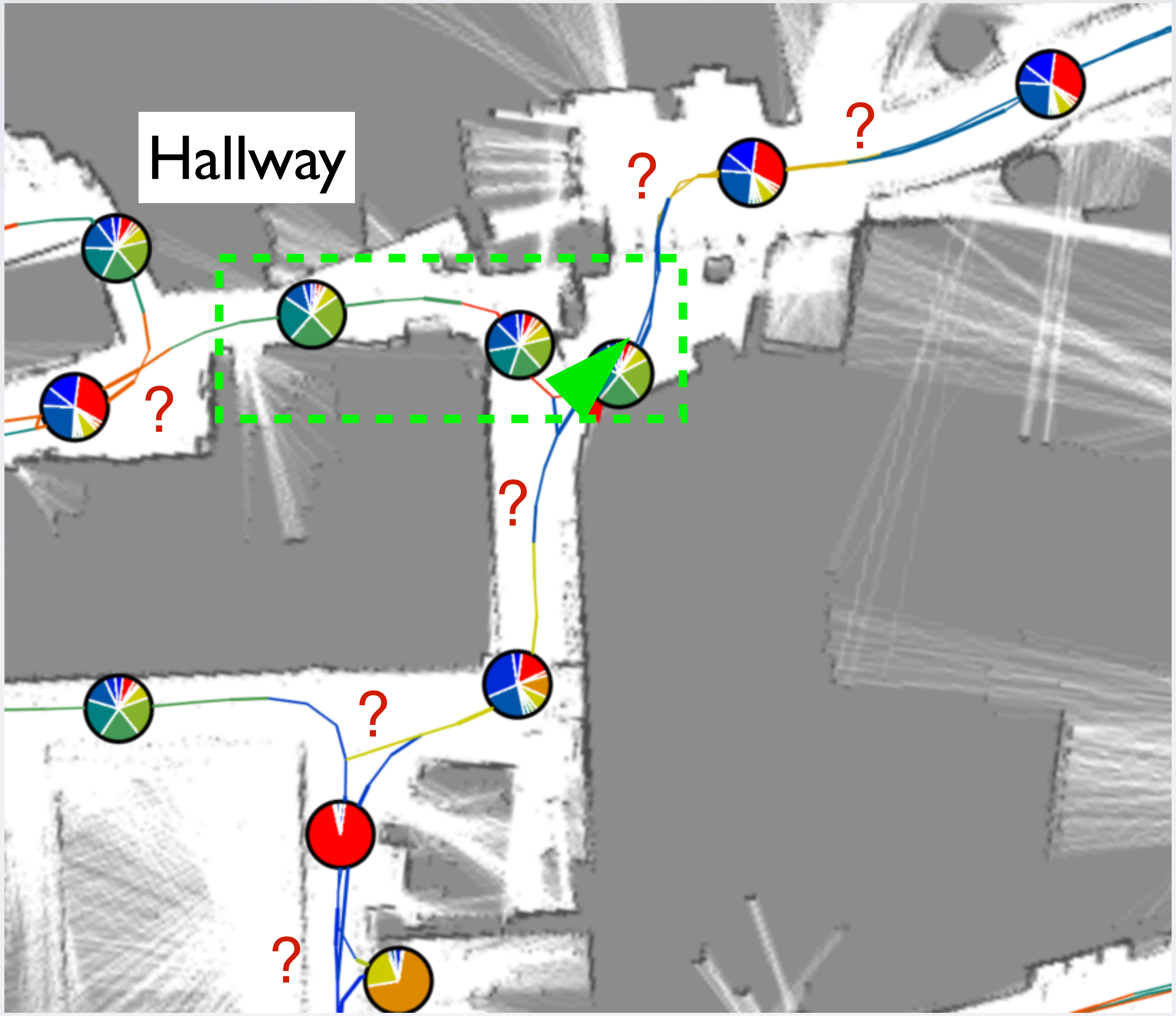
3) **Update**: Update Dirichlet over labels based on language

4) **Reweight**: Update weights based on metric observations

Return: $\mathcal{P}_{t} = \left\{ P_{t}^{(1)}, P_{t}^{(2)}, \ldots, P_{t}^{(n)} \right\}$ where $P_{t}^{(i)} = \left\{ G_{t}^{(i)}, X_{t}^{(i)}, L_{t}^{(i)}, w_{t}^{(i)} \right\}$

Updating the Dirichlet Distribution

"The kitchen is down the hallway"



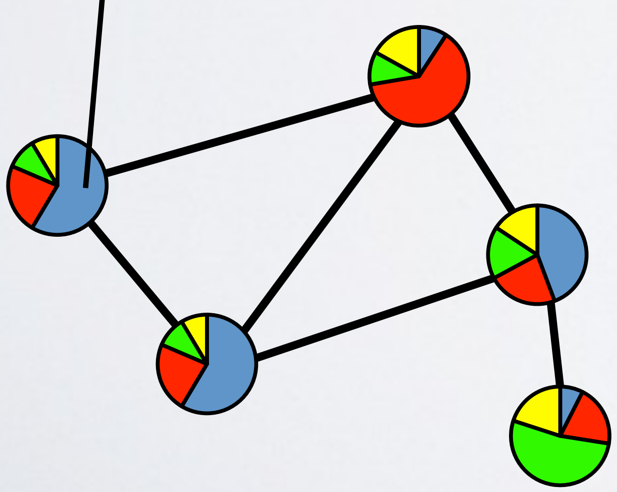
Updating the Dirichlet Distribution

$\lambda_t =$ "The kitchen is down the hallway"

$$p(L_t | X_t^{(i)}, G_t^{(i)}, z^t, u^t, \lambda^t) = \prod_{i=1}^{t} p(l_{t,i} | X_t^{(i)}, G_t^{(i)}, z^t, u^t, \lambda^t)$$

likelihood that language references region i

$$p(l_{t,i} | \lambda_t, l_{t-1,i}) = \frac{\Gamma(\sum_1^K \alpha_i^{t-1} + \Delta\alpha)}{\Gamma(\alpha_1^{t-1}) \times \ldots \times \Gamma(\alpha_k^{t-1} + \Delta\alpha) \times \ldots \times \Gamma(\alpha_K)} \prod_{k=1}^{K} l_{t,i,k}^{\alpha_k - 1}$$



Symbol Grounding Problem

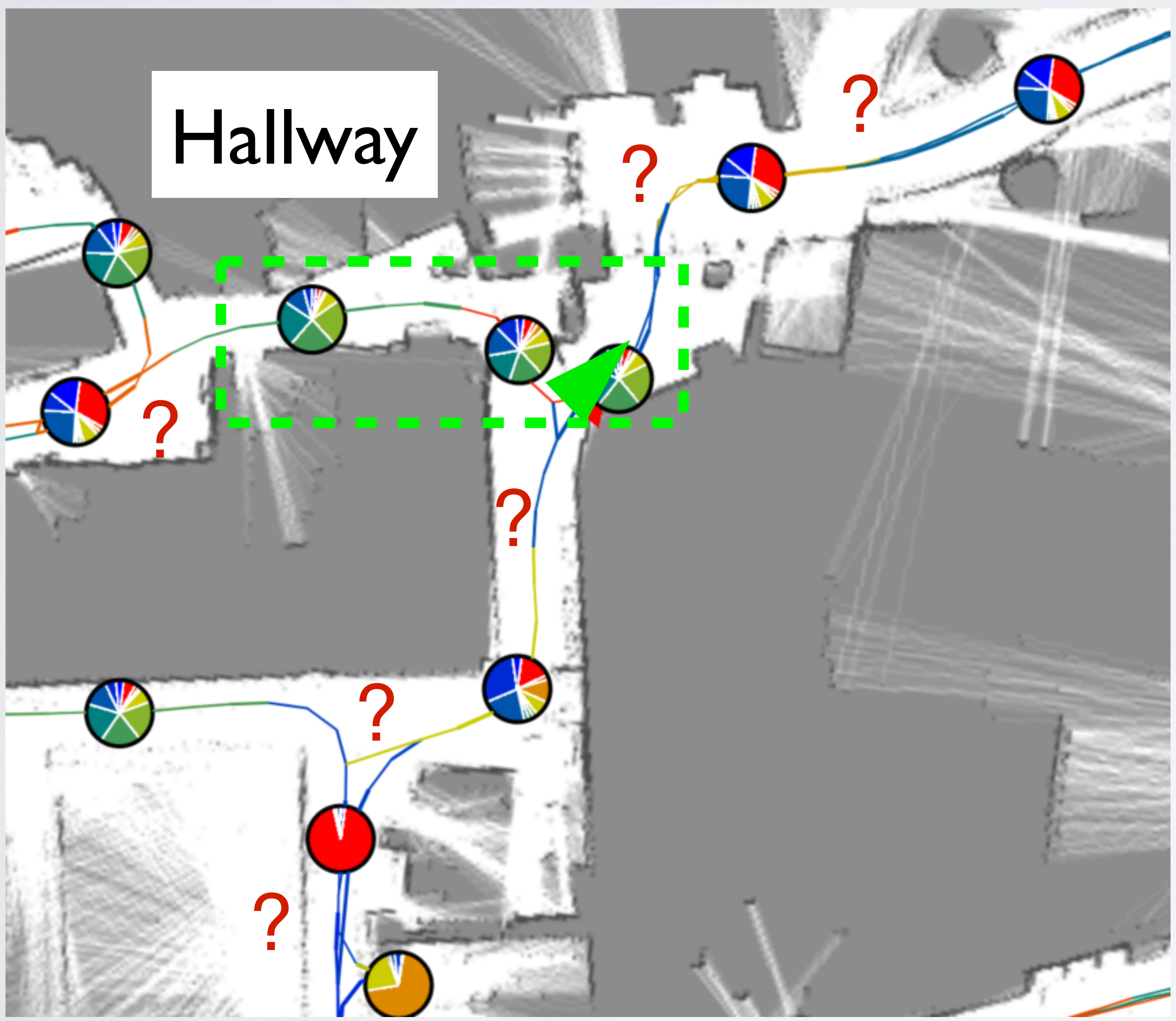
Linguistic elements → Correct referents in the robot's world model

"Grounding"

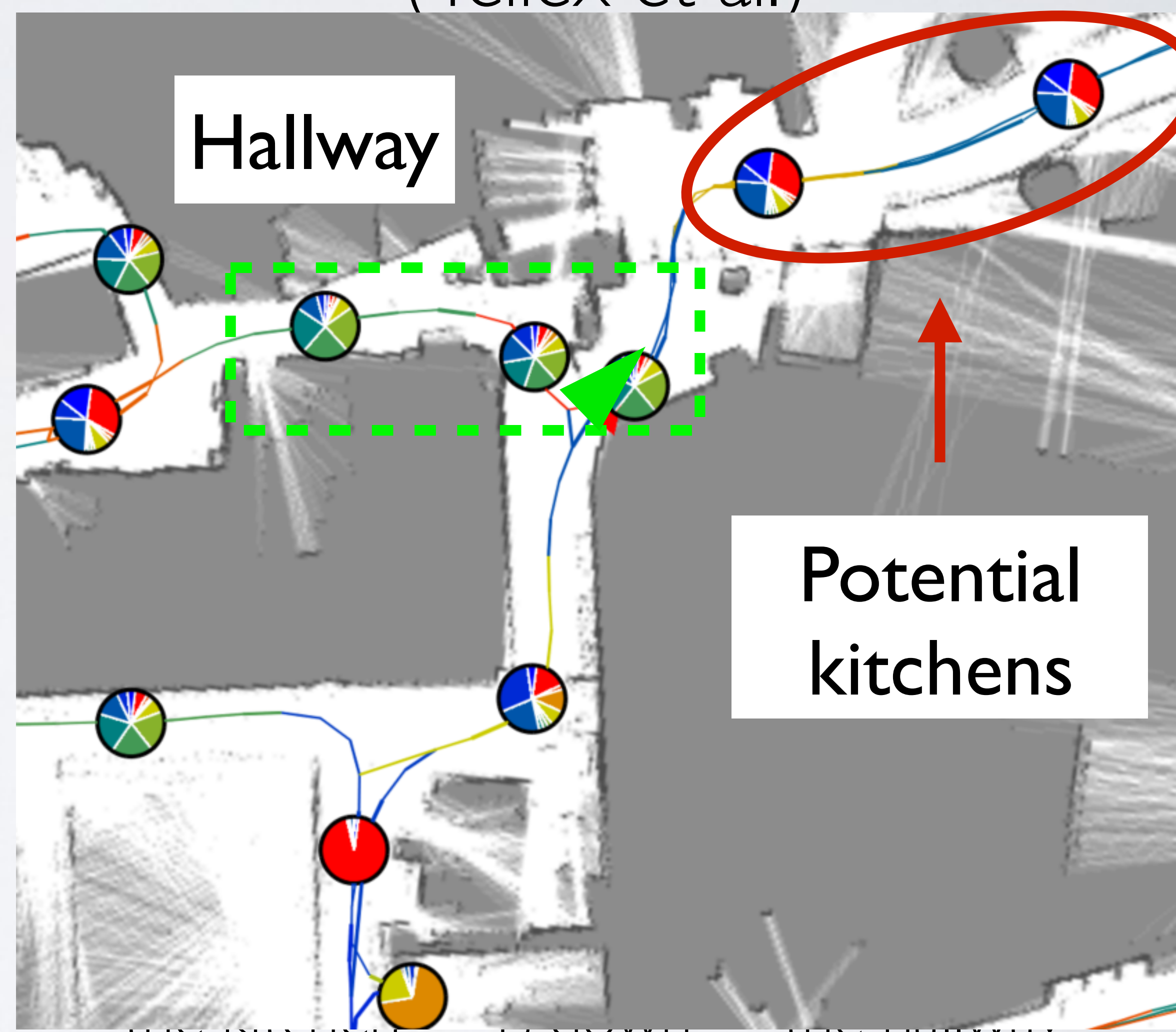
The kitchen is down the corridor.
The kitchen is behind you.
Down the hall, you'll find the kitchen past the exit.
The galley is down the corridor to the left.
The Stata kitchen is on the right, past the tall filing cabinet.
The kitchen is through the double doors at the end of the hall.
The Stata Center's kitchen is behind you, just beyond the doors to the elevator lobby.

Grounding Natural Language

"The kitchen is down the hallway"



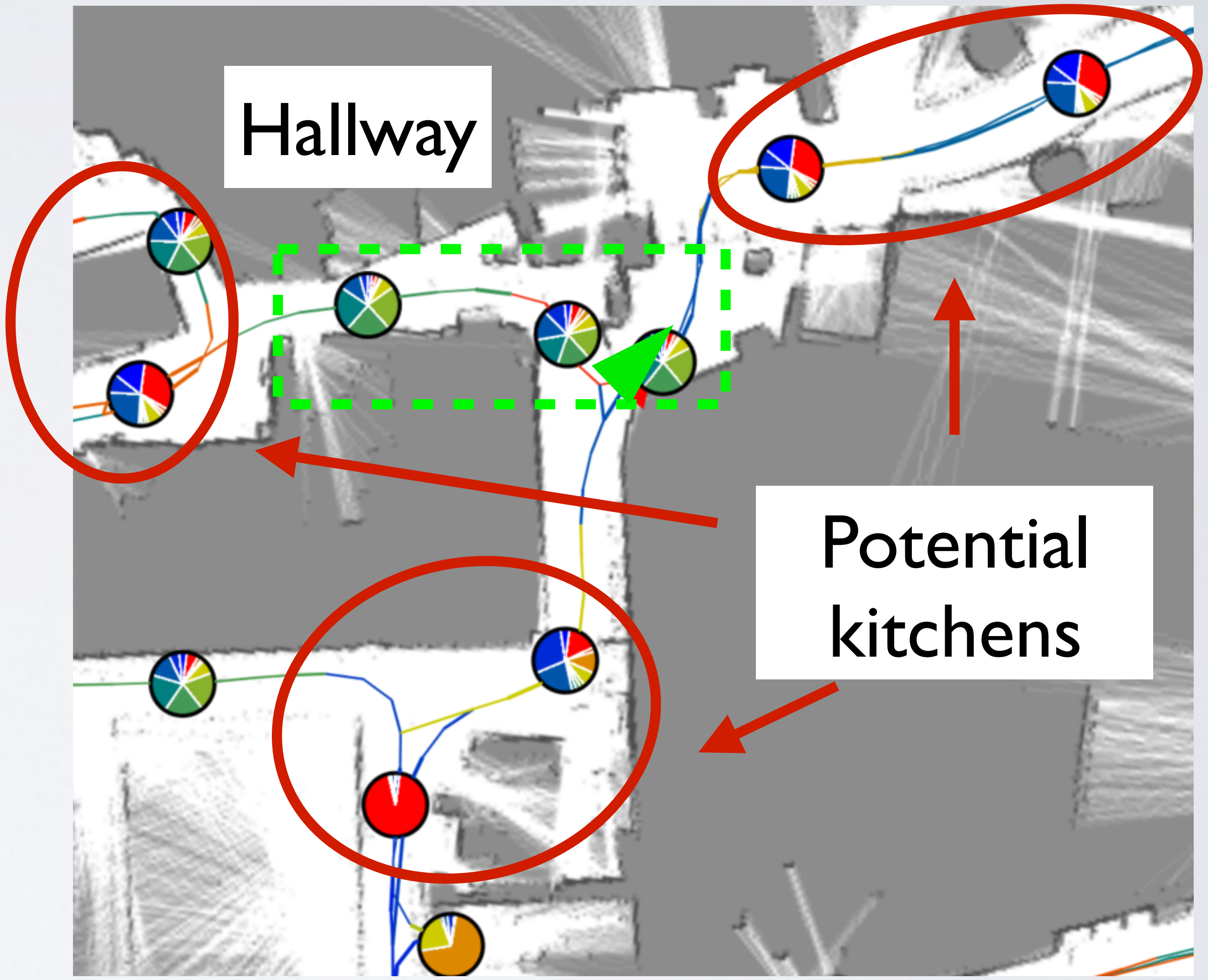
Generalized Grounding Graph
(Tellex et al.)



[AAAI 2011; AI Magazine 2011]

Language Grounding Ambiguity

- Descriptions are often ambiguous
- “The kitchen is down the hallway”
  - Multiple hallways (known & unknown)
  - Multiple regions “down” hallways
- Robot’s role is traditionally passive



Resolving Ambiguity Through Dialogue

- Robot can explore to resolve uncertainty
  - Physical exploration
  - **Dialogue**
- Dialogue: Robot asks questions that disambiguate groundings

Challenges to Dialogue

- Decide whether to ask a question
- Decide which region to ask about
- Deal with partially known environments
- Provide sufficient context to the user
- Model frame-of-reference

Problem Formulation

- Model next state as tuple
  - Previous semantic map
  - Question
  - Answer
- At each time t, robot selects from a set of actions:
  - Follow the user
  - Ask a question
- Define question asking actions a_i for each language utterance
- Answers (states) are uncertain $\longrightarrow$ (Q)MDP

Problem Formulation

- Per-particle state:

semantic map particle — semantic map particle

$$S_{t+1} = \left\{ P_t^{(i)}, a_t, z_t^a \right\}$$

action

- Actions:
  - Follow the user
  - Stay in place
  - Ask a question
- Transition function based on answer likelihood
- Reward reflects information gain and cost

Action Selection

Plan a one-step policy:

particle weight

$$a_t^B = \arg\max_{a_t} \sum_{S_t} p(S_t) Q(S_t, a_t)$$

where

value = function(information gain)

cost = function(burden)

$$Q(S_t, a_t) = \sum_{S_{t+1}} \gamma V(S_{t+1}) \times p(S_{t+1} | S_t, a_t) - \mathcal{C}(a_t)$$

$$= \gamma \mathbb{E}(V(S_{t+1})) - \mathcal{C}(a_t)$$

Action Selection

- Cost of an action:

$$\mathcal{C}(a_t) = \mathcal{F}(f(a_t))$$

- Time since last question
- Time since last asking about grounding
- Number of questions asked

- Value of the next state:

$$V(S_{t+1}) = \mathcal{F}(I(a_t))$$

- Information gain for (question, answer) pair:

NLU figure (region) grounding

$$I(a, z^a) = H(\gamma_f | \Lambda) - H(\gamma_f | \Lambda, a, z^a)$$

Action Selection

Plan a one-step policy:

$$a_t^B = \arg\max_{a_t} \sum_{S_t} p(S_t) Q(S_t, a_t)$$

where

$$Q(S_t, a_t) = \gamma \mathbb{E}(V(S_{t+1})) - \mathcal{C}(a_t)$$

$$\mathbb{E}(V(S_{t+1})) = \sum_{z_j^a} \mathcal{F}(I(a|z_j^a)) \times p(z_j^a|S_t, a)$$

possible answers for question a

Choosing Question Structure

- Consider binary (yes/no) questions: $z_j^a \in \{\text{yes}, \text{no}\}$
- Questions follow structured template:

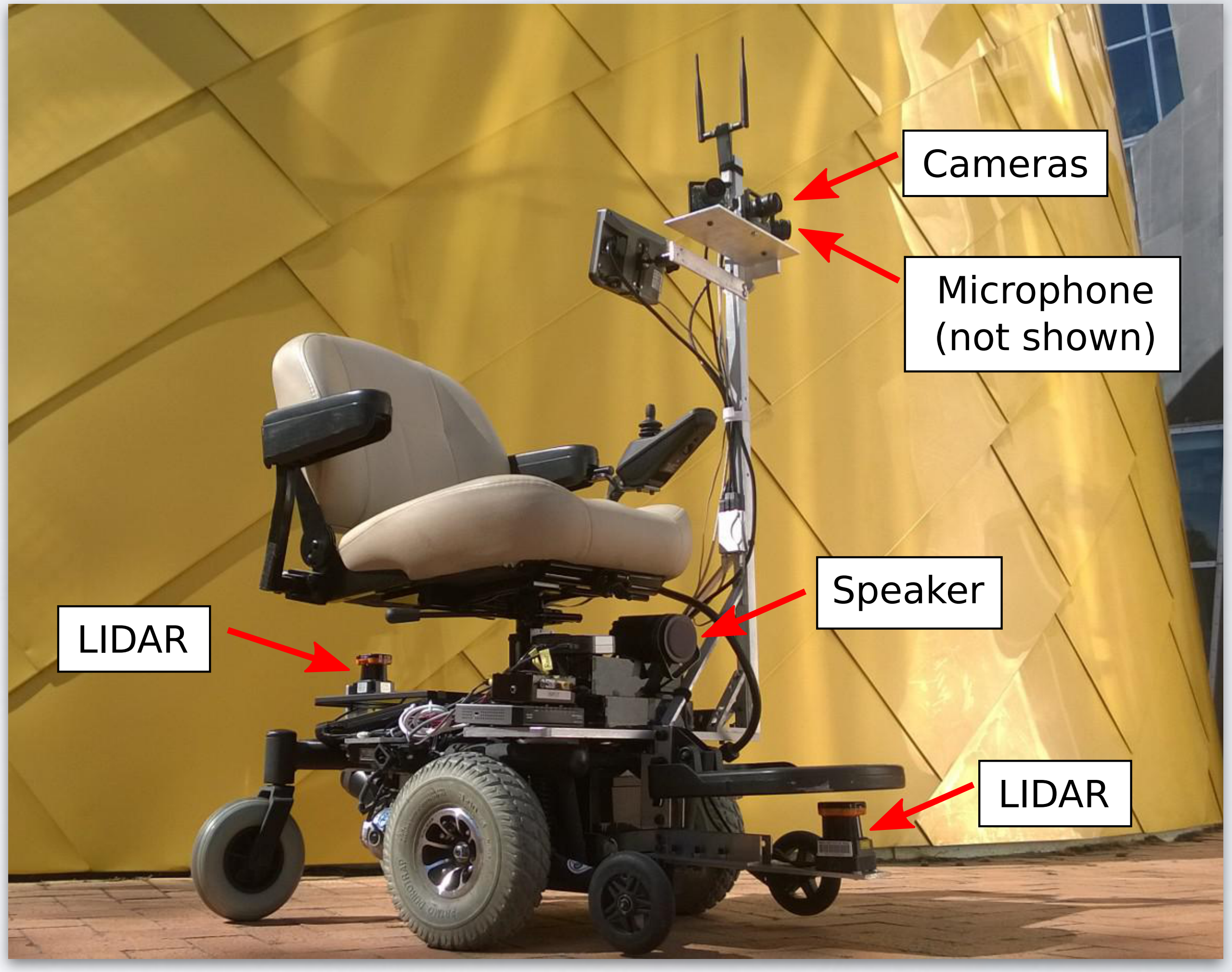
$$<\text{figure}> \ <\text{relation}> \ <\text{landmark}>$$

"Is the kitchen in front of me?"

- Two types of landmarks
  - Robot: "Is the kitchen in front of me?"
  - An environment region: "Is the kitchen across from the cafeteria?"
- Context: Choose landmark (and relation) that provides most information
- Assume robot's frame-of-reference

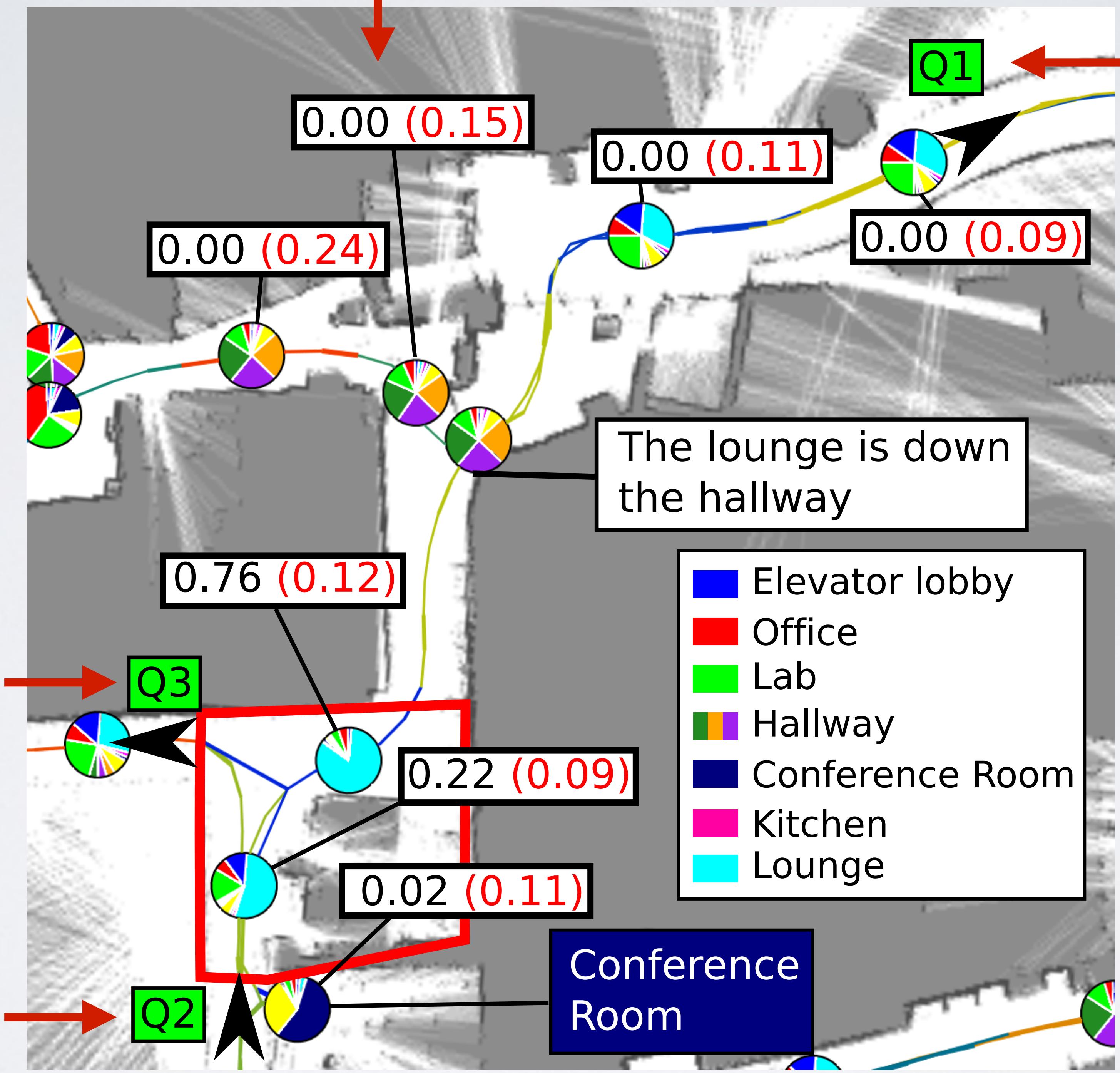
Experiment

- Gave narrated guided-tour of the MIT Stata Center
- Robotic wheelchair equipped with
  - Two LIDARs
  - Three monocular cameras
- User provided 9 descriptions
  - 6 egocentric
  - 3 allocentric
- Robot asked 5 questions



Results

Grounding likelihood with (without) dialogue



Q1: "Is the lounge near the conference room?"

A1: "Yes"

Q3: "Is the lounge behind me?"

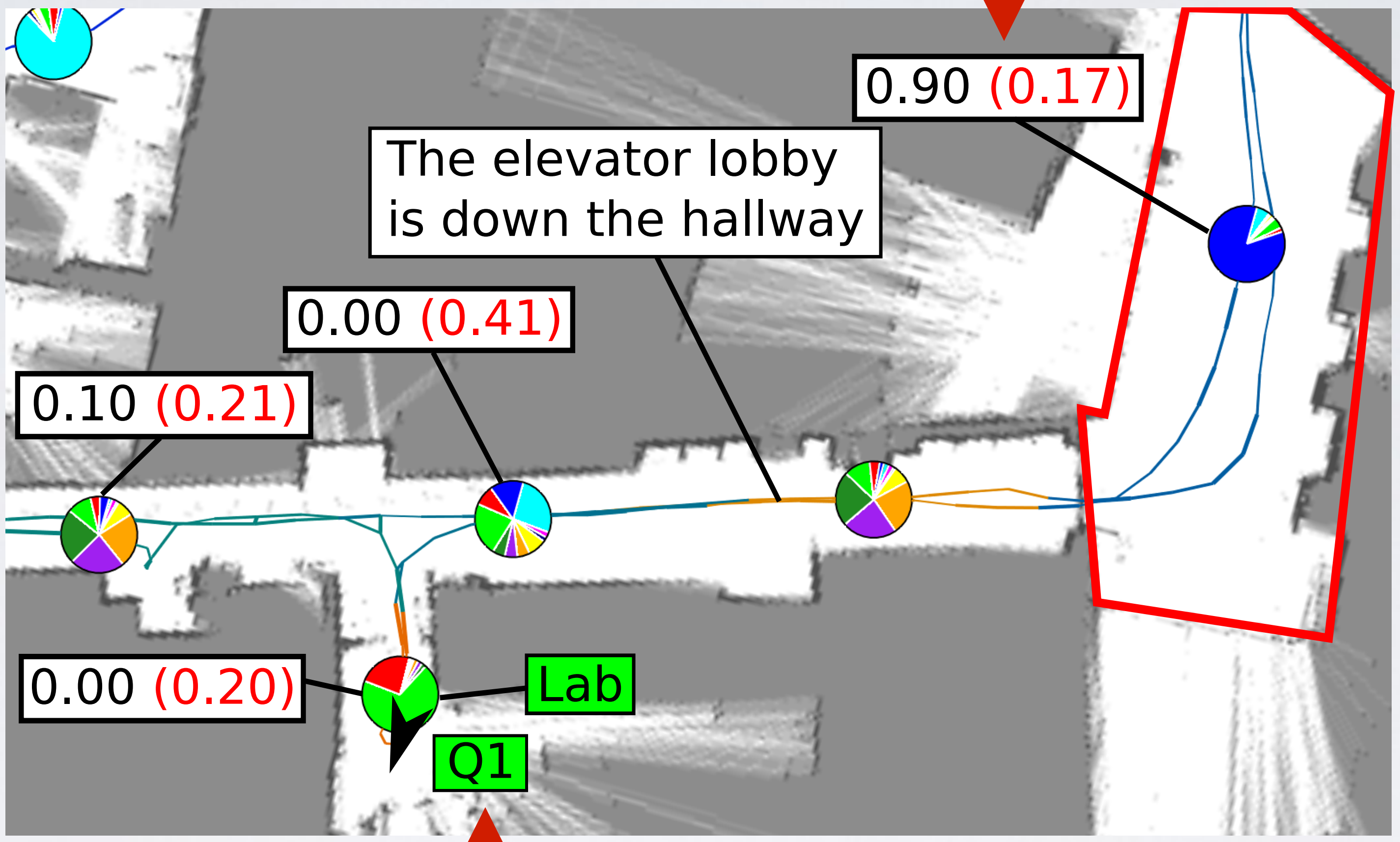
A3: "Yes"

Q2: "Is the lounge on my right?"

A2: "No"

Results

Grounding likelihood with (without) dialogue



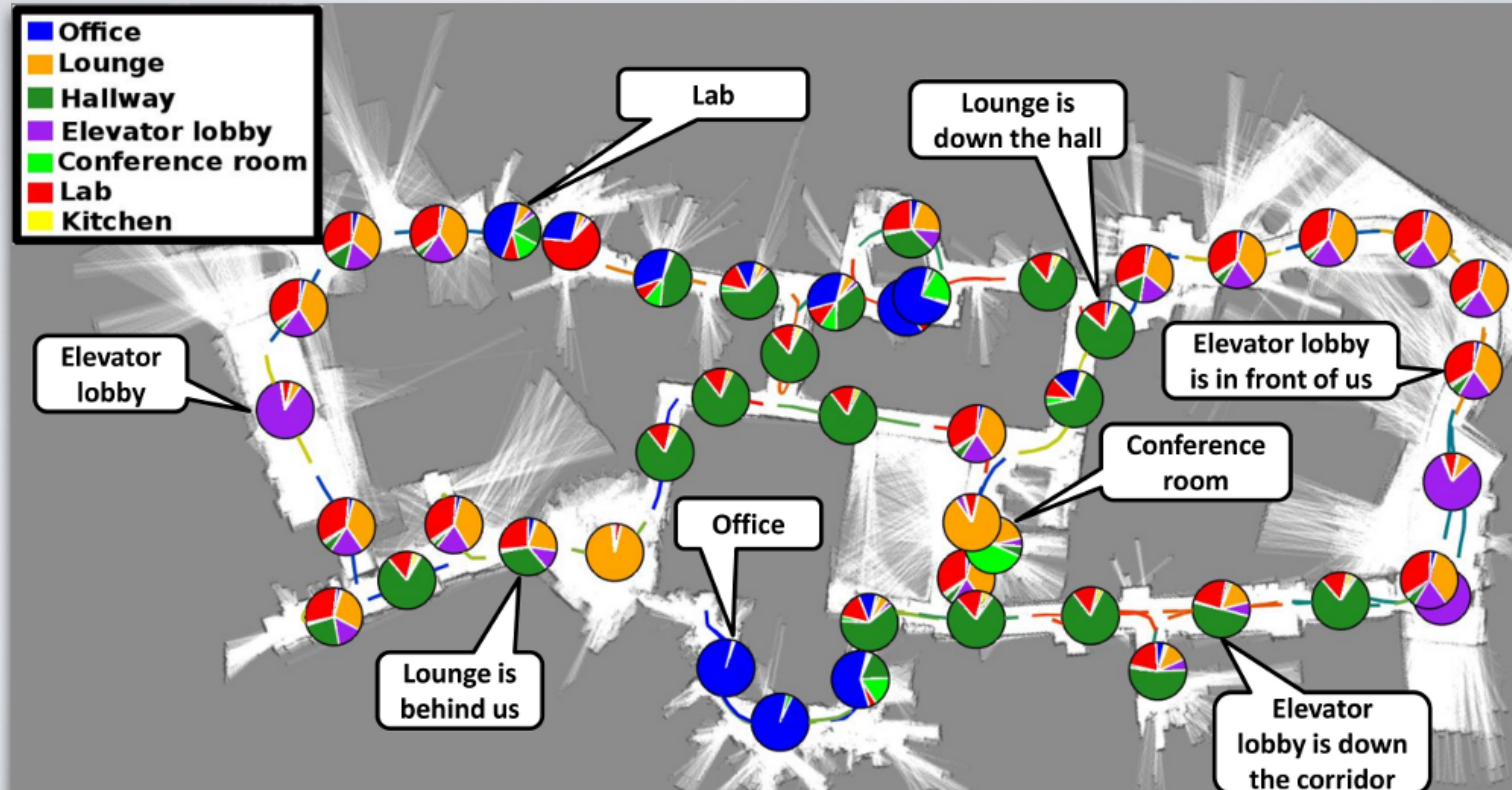
Q1: "Is the elevator lobby near me?"

A1: "No"

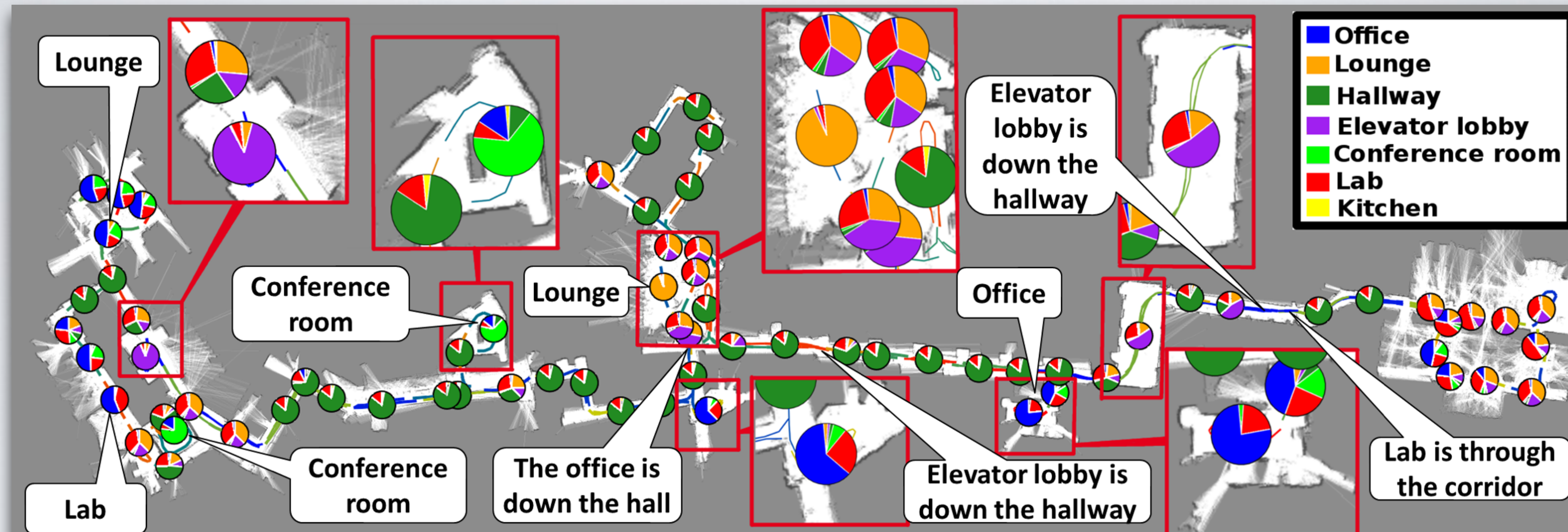
Results

Utterance	Entropy		Accuracy		No. of Questions
	Without Questions	With Questions	Without Questions	With Questions	
"The lounge is down the hallway"	1.911	0.237	17.3%	90.6%	2
"The elevator lobby is down the hallway"	1.574	0.566	35.8%	70.9%	2
"The lounge is behind you"	0.403	0.095	87.2%	98.4%	1
"The lab is down the hall"	2.041	0.310	14.6%	91.6%	3
"The conference room is down the hallway"	2.061	0.664	6.5%	65.5%	8
"The lounge is in front of us"	1.053	0.107	20.6%	43.8%	2

Stata Center Third Floor Semantic Graph



Multi-Building Semantic Graph



I. Introduction

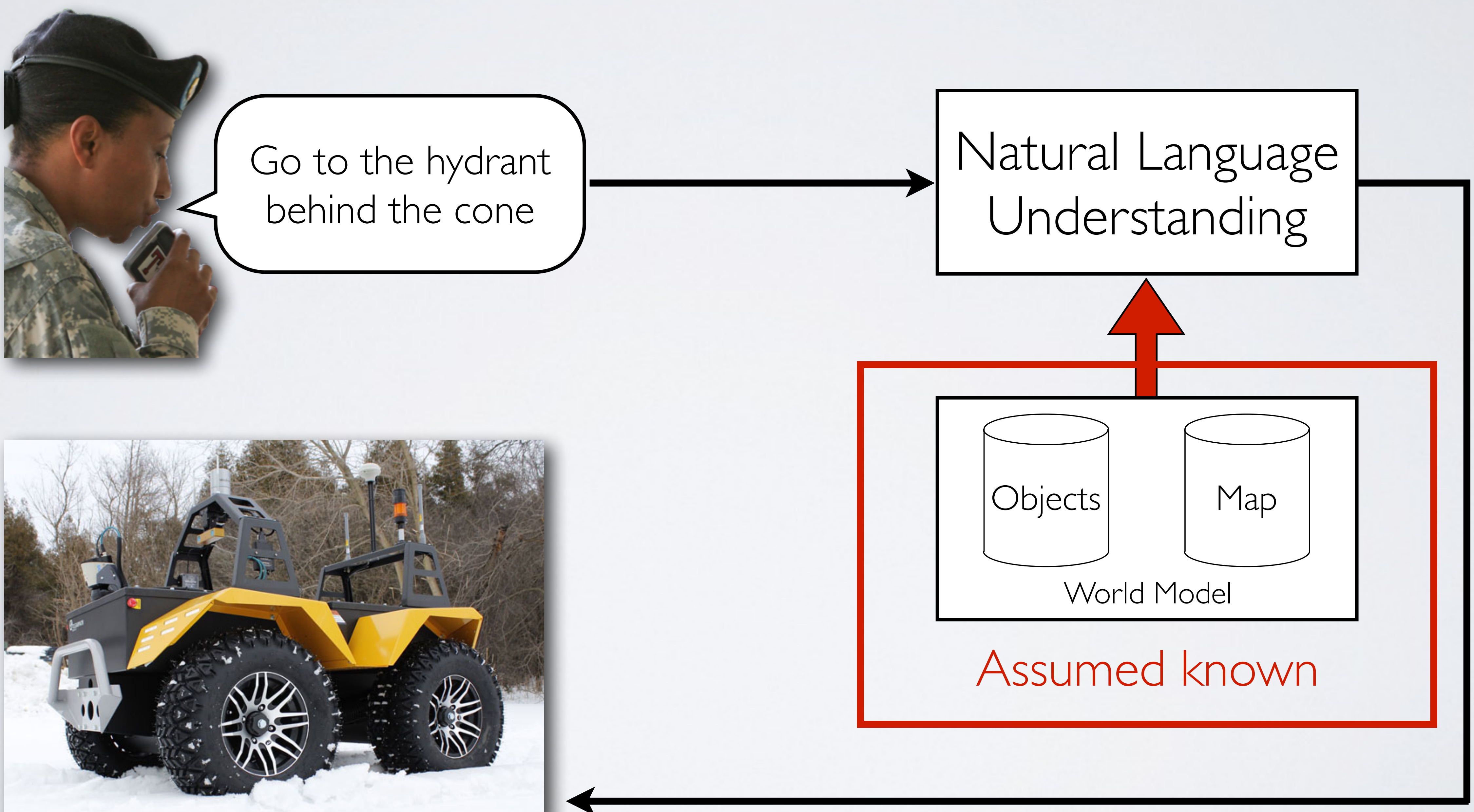
II. Learning Semantic Maps from Natural Language Dialogue

III. Following Directions Without in Unknown Environments

IV. Future Directions & Conclusions

Language Understanding Without a Map

$$\arg\max_{\gamma_i} p(\gamma_o, \gamma_a, \gamma_r, \gamma_p | S, \Lambda)$$



Language Conveys Two Types of Information

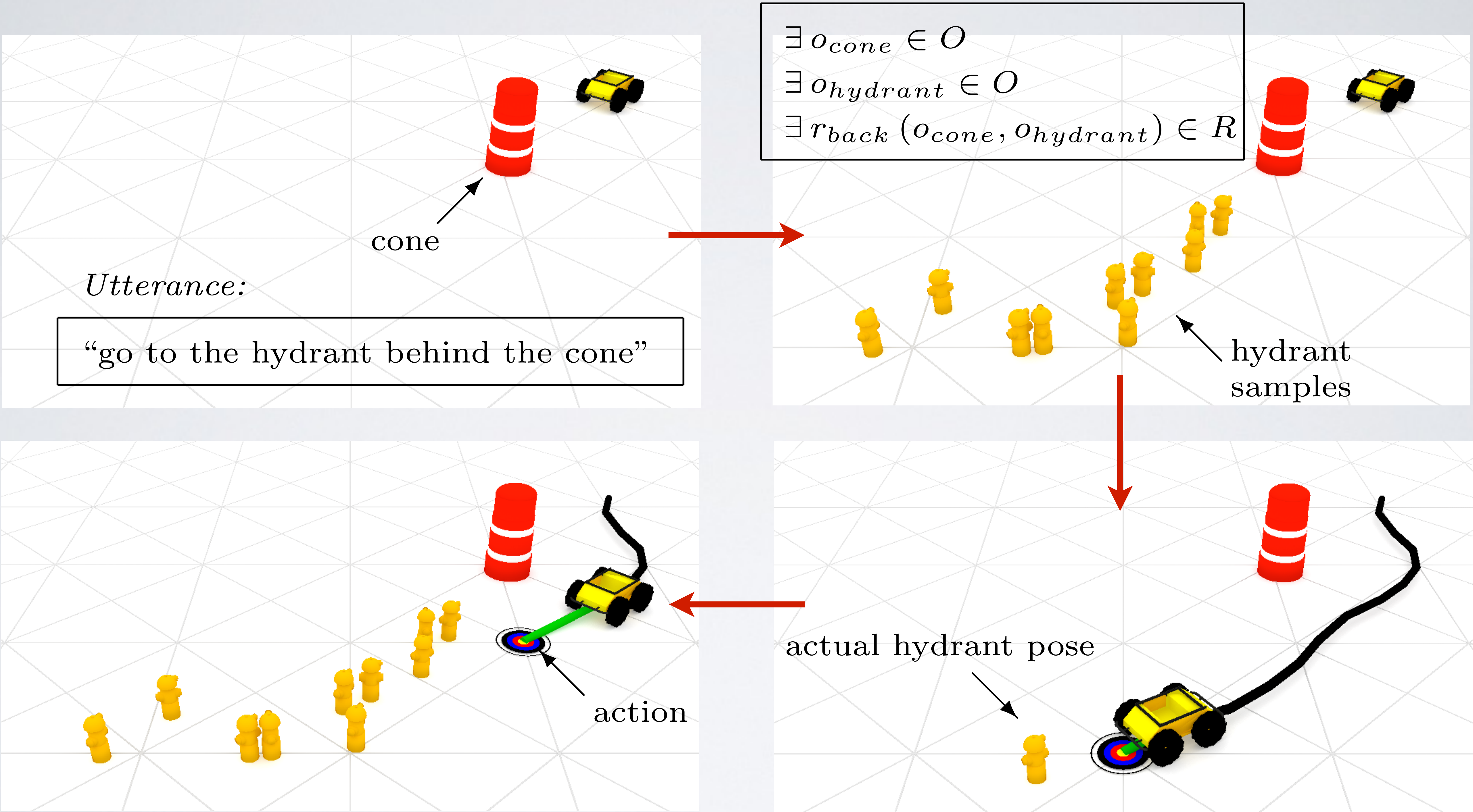
Go to the hydrant behind the cone

Explicit: Task
"Go to the hydrant"

Implicit:
Description of the world

[ISER 2014]

Joint Map & Behavior Inference



[ISER 2014]

Joint Map & Behavior Inference

Sensor stream

Robot trajectory Language Odometry

$$\underset{x(t)\in\mathbb{R}^n}{\arg\max}\; p\left(x(t)|\Lambda^t, z^t, u^t\right)$$

[ISER 2014]

Joint Map & Behavior Inference

Semantic Graph

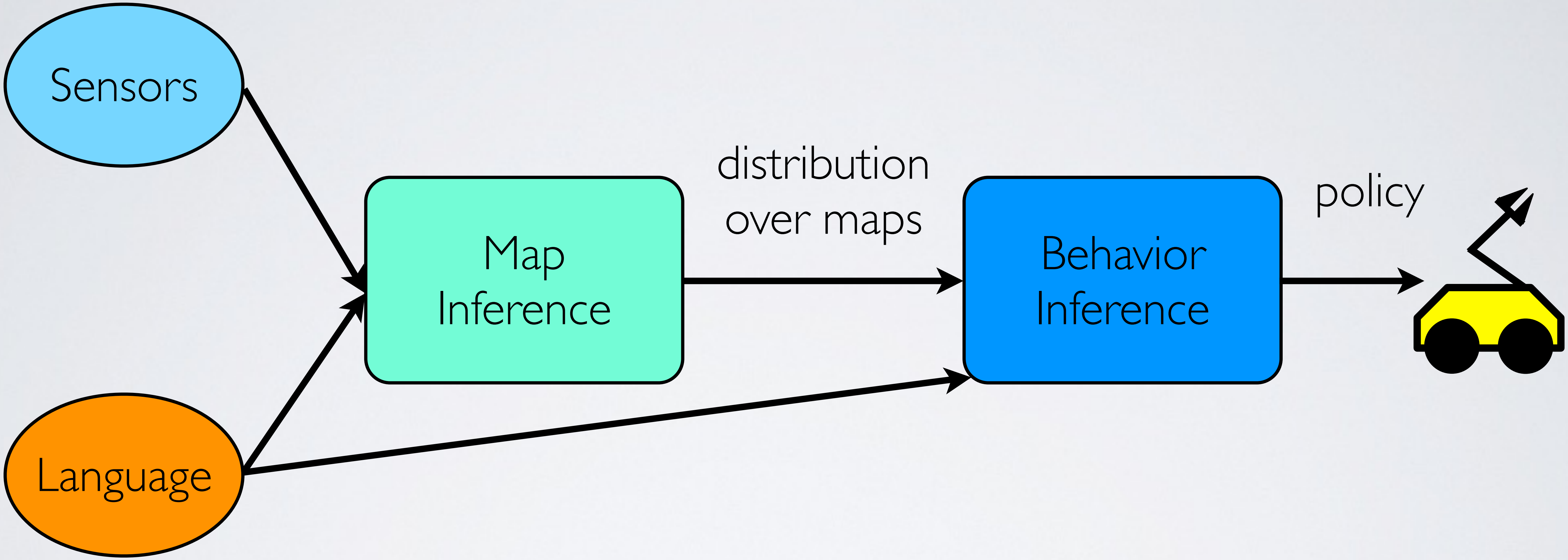
$$\arg\max_{x(t)\in\mathbb{R}^n} \int_{S_t} p\left(x(t)|S_t,\Lambda^t\right) p\left(S_t|\Lambda^t,z^t,u^t\right) dS_t$$

Policy learning

Semantic Mapping

[ISER 2014]

Joint Map & Behavior Inference



[ISER 2014]

Extracting Facts About the World

$$p(S|\Lambda, z, u) \approx p(S|\alpha, z, u)$$

DCG [10] $\longrightarrow$

Annotations
(objects, relations)

$$\exists \; r_{\text{behind}}(o_{\text{cone}}, o_{\text{hydrant}}) \in \mathcal{R}$$

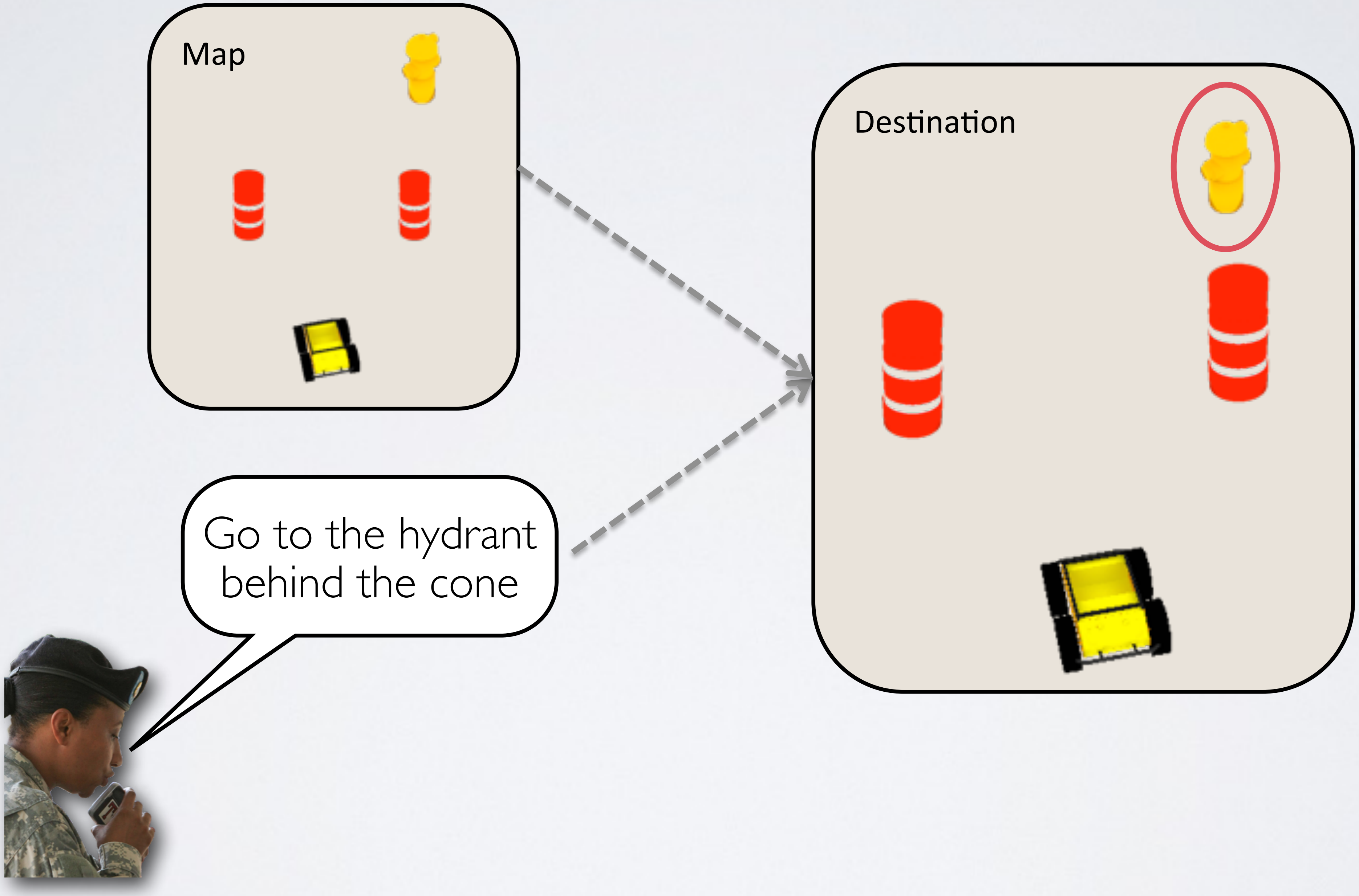
$$\exists \; o_{\text{cone}} \in \mathcal{O}$$

$$\exists \; o_{\text{hydrant}} \in \mathcal{O}$$

[10] Howard et al. 2014

[ISER 2014]

Behavior Inference: Behaviors given Map Distribution

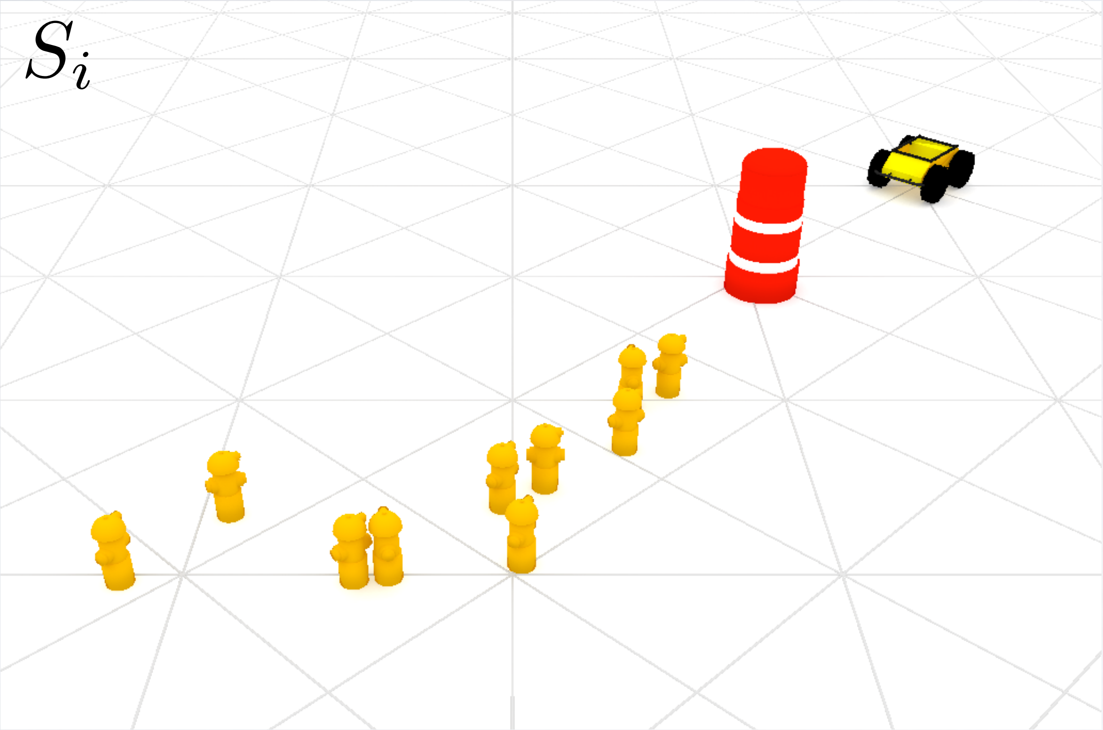


\

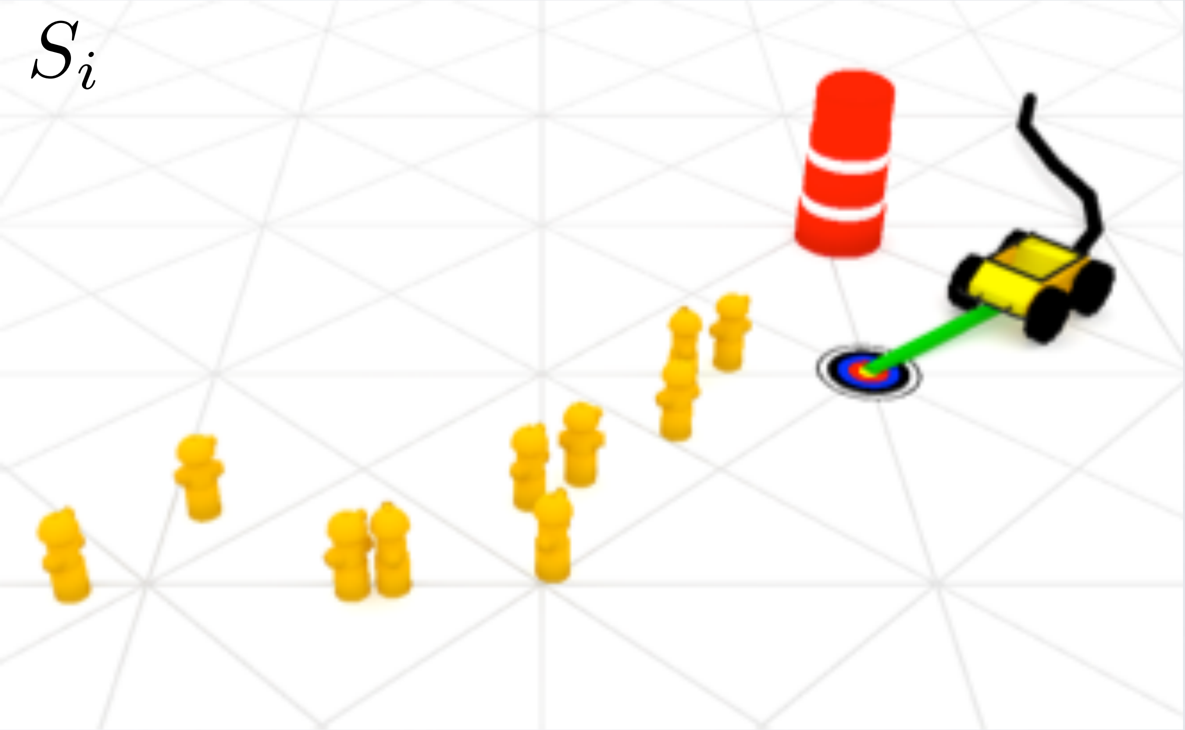
[ISER 2014]

Find Actions Consistent with Inferred Behavior

Action Set: One step destination



What is the next destination?

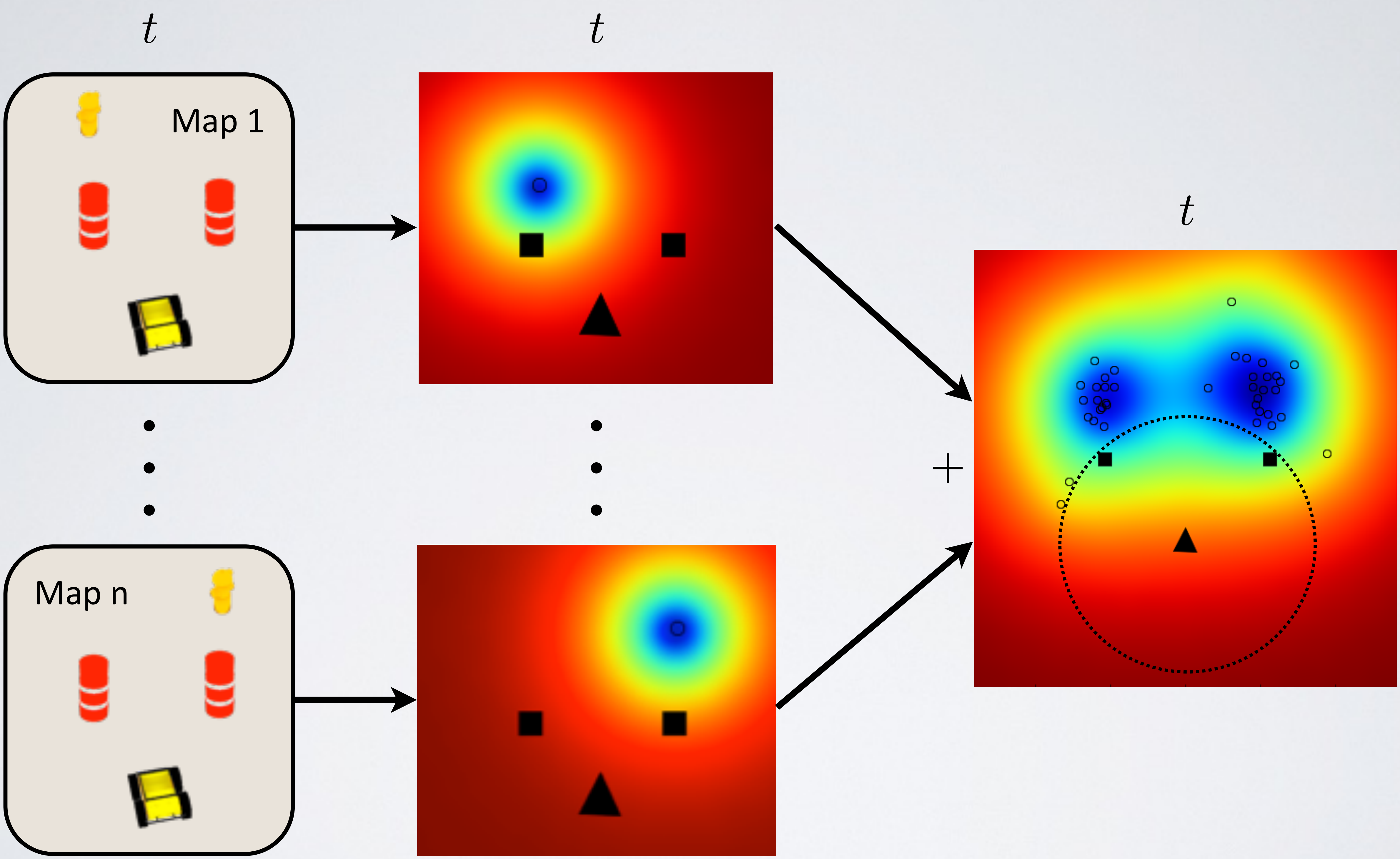


Policy π

\

[ISER 2014]

Find Actions Consistent with Inferred Behavior

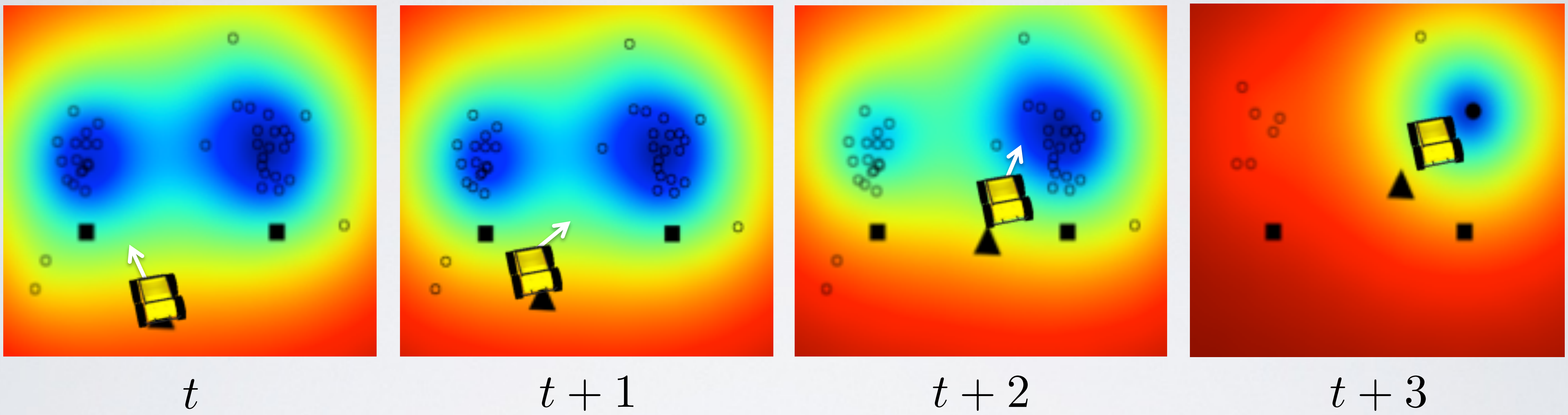


QMDP [Littman et al., 1995]

\

[ISER 2014]

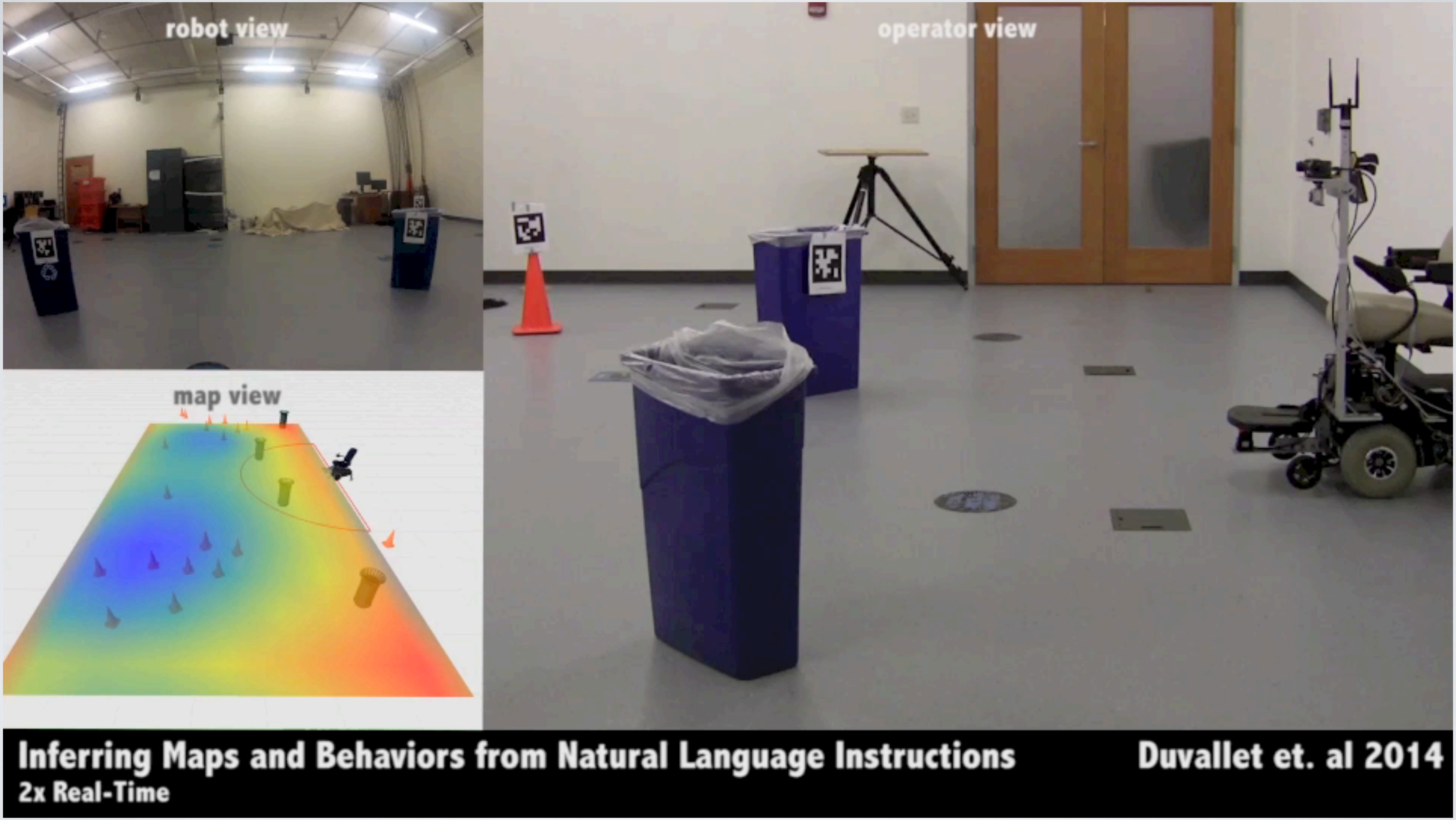
Find Actions Consistent with Inferred Behavior



QMDP [Littman et al., 1995]

\

[ISER 2014]


robot view
operator view
map view
Inferring Maps and Behaviors from Natural Language Instructions
2x Real-Time
Duvallet et. al 2014

Following Route Directions in Unknown Envs.

I. Introduction

II. Learning Semantic Maps from Natural Language Dialogue

III. Following Directions Without in Unknown Environments

IV. Future Directions & Conclusions

Semantic Map Learning

- Learn from additional semantic cues (e.g., objects, text)
- Learn from non-situated descriptions
- Information-gathering actions:
  - Physical exploration
  - Dialogue

Matthew Walter

Future Work

- Extend dialogue beyond user-referenced locations
- Consider less-structured questions
  - Open-ended answers
  - Free-form questions
- Account for figures that refer to unknown regions
- Go beyond a hand-crafted measure of cost (burden)
- Reason over frame-of-reference
- Incorporate physical exploration
- Move towards fully non-situated dialogue

Conclusions

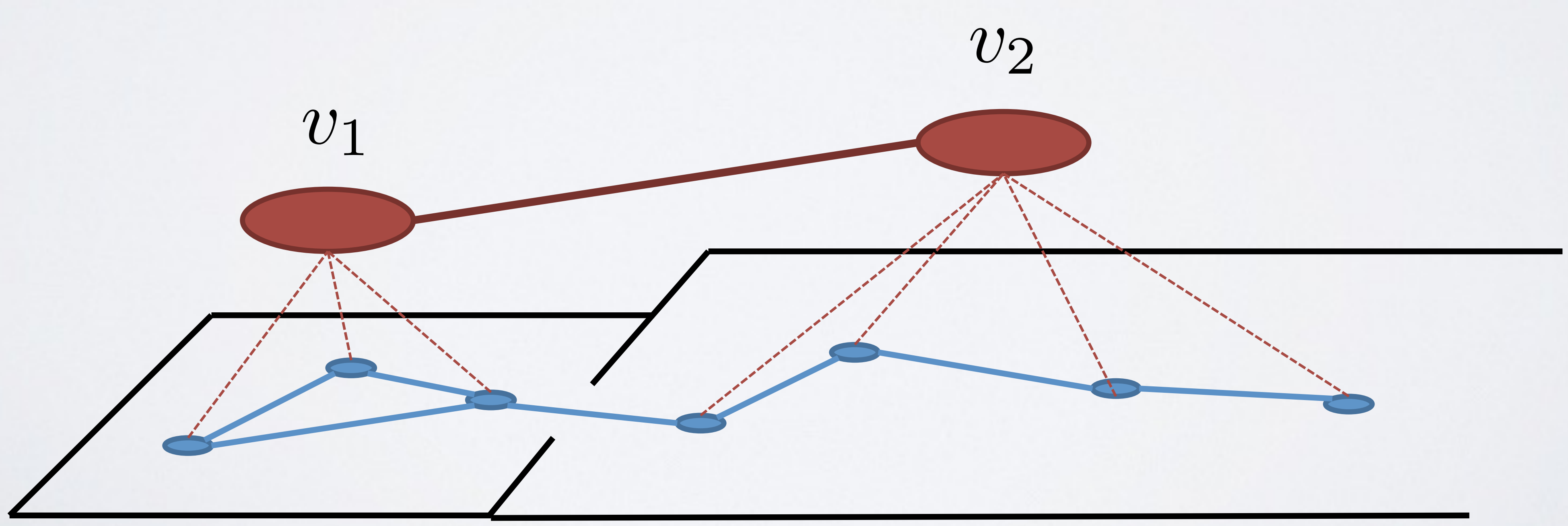
- Natural language understanding for robots requires cognitive models
- Argued that language is an effective means of sharing our cognitive models
- Described an algorithm that learns semantic environment models from natural-language dialogue
- Described an algorithm for joint map and behavior inference
- Outlined ongoing and future work

Questions?

mwalter@ttic.edu

Topological Map Representation

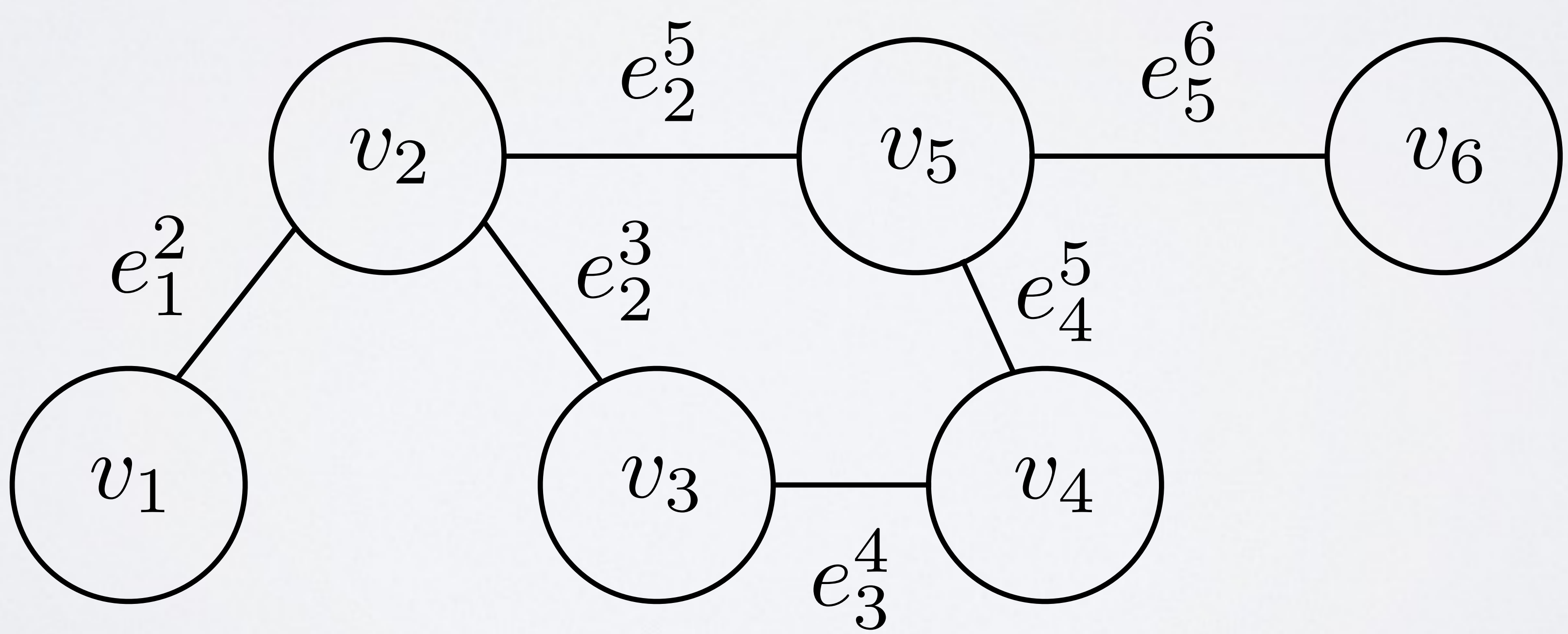
- Node v_i denotes a distinct (semantically meaningful) region
- Edges e_i represent connectivity
  - Observed with robot's sensors (e.g., scan-matching)
  - Traversed (odometry)
  - Inferred from description
- What defines a "region"?
  - Local, spatial consistency ← Spectral clustering of laser scans
  - Semantic attributes (e.g., room type)



Semantic Graph

$$\{G_t, X_t, L_t\}$$

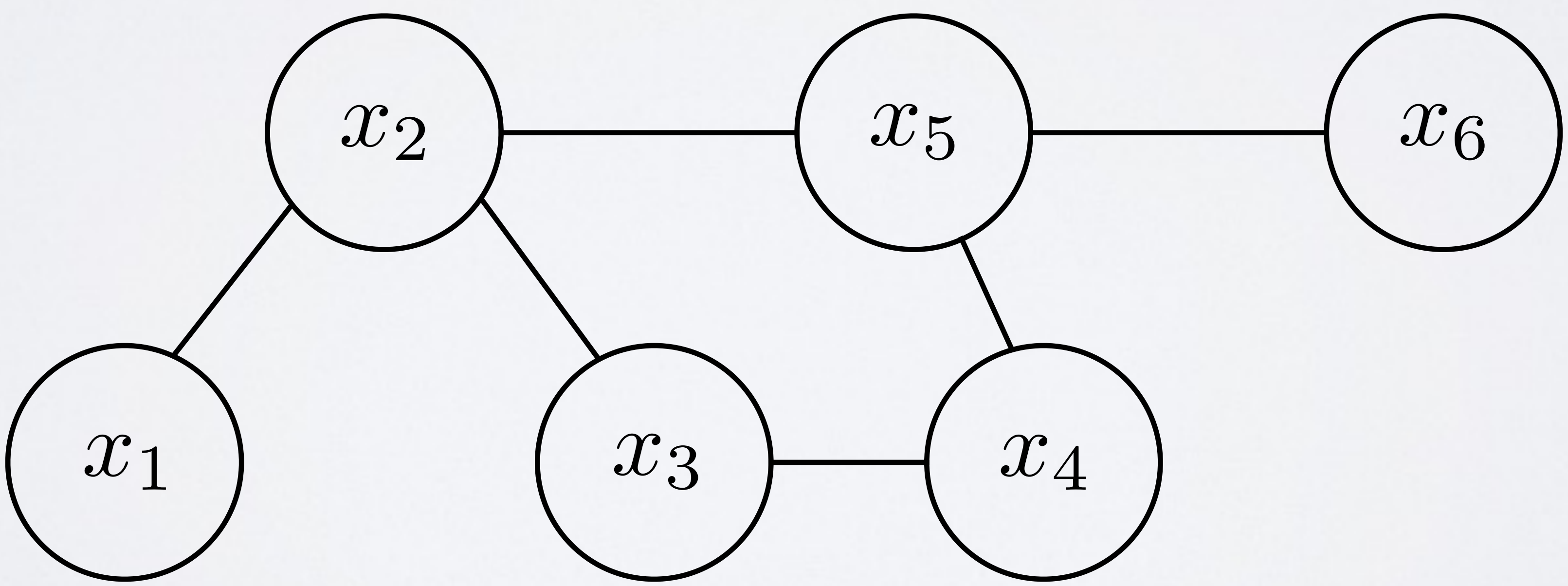
- **Topological map:** $G_t = (V_t, E_t)$
- Metric map: $X_t = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$
- Semantic map: $L_t = \{l_1, l_2, \ldots, l_n\}$



Semantic Graph

$$\{G_t, X_t, L_t\}$$

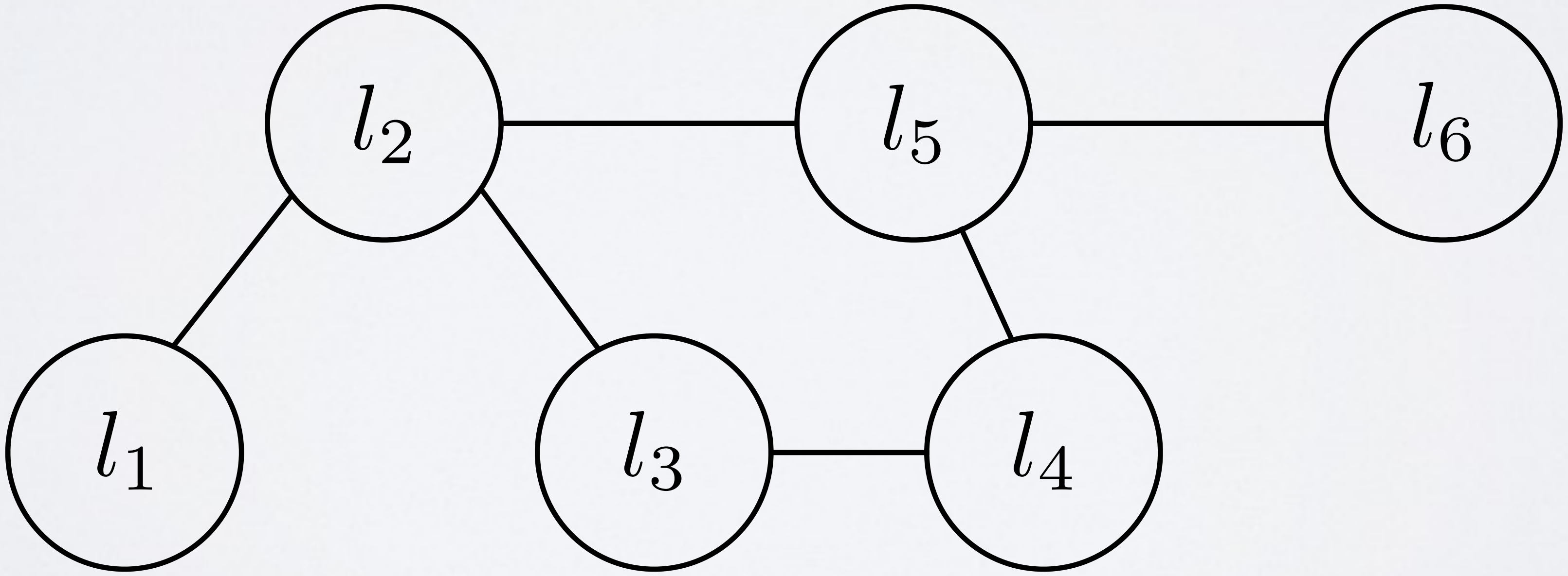
- Topological map: $G_t = (V_t, E_t)$
- **Metric map**: $X_t = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$
- Semantic map: $L_t = \{l_1, l_2, \ldots, l_n\}$



Semantic Graph

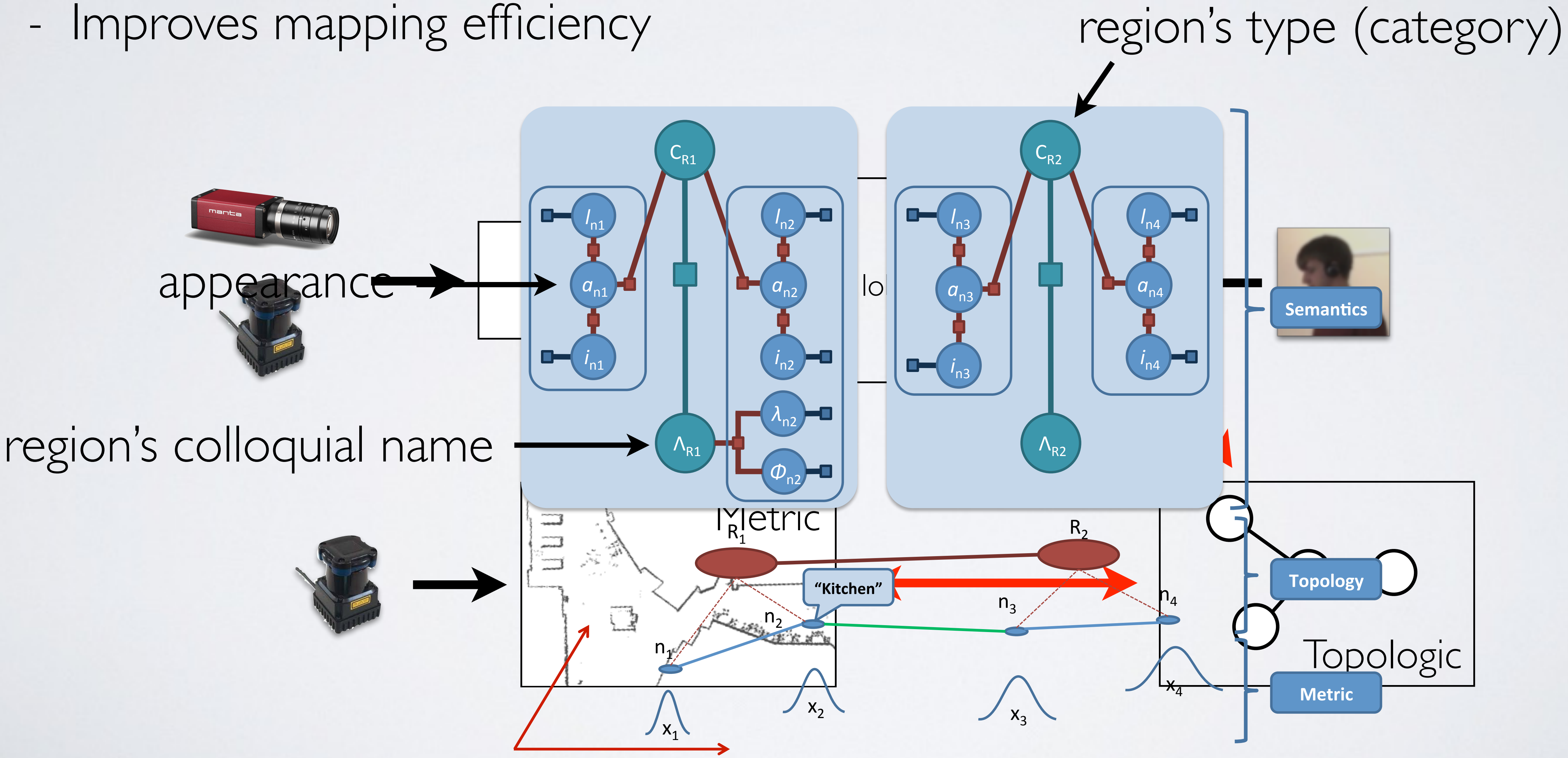
$$\{G_t, X_t, L_t\}$$

- Topological map: $G_t = (V_t, E_t)$
- Metric map: $X_t = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$
- **Semantic map**: $L_t = \{l_1, l_2, \ldots, l_n\} \quad l_i = (\text{colloquial name}, \text{region type})$



Semantic Attributes via Scene Classification

- Jointly reason over semantic hierarchy: region type and colloquial name
- Infer region type from sensor data
  - Improves allocentric language grounding
  - Improves mapping efficiency

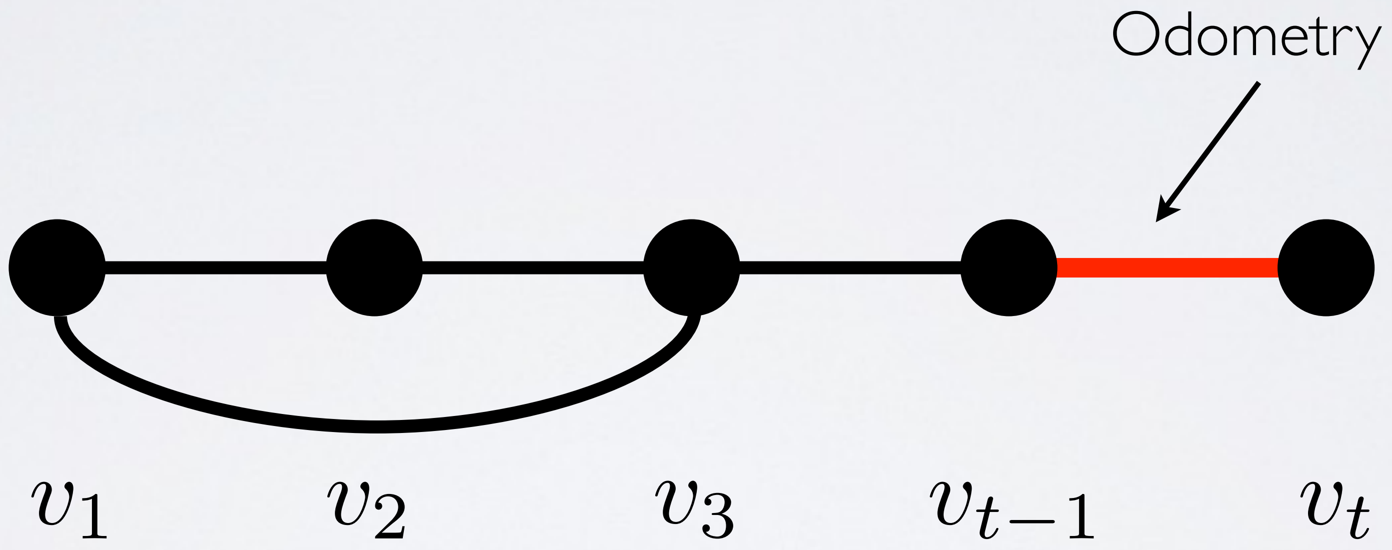


Matthew Walter

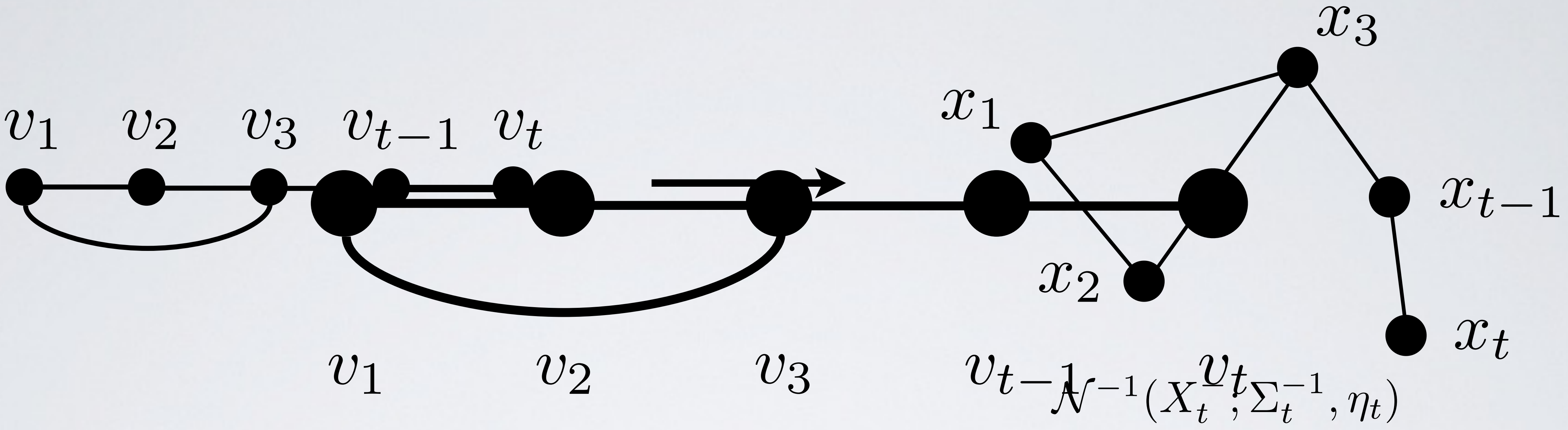
Proposal Distribution: Odometry Annotation

$$p(G_t^-|G_{t-1}, z^{t-1}, u^t, \lambda^t)$$

$G_t^- = \{V_t^-, E_t^-\}$



Proposal Distribution: Metric Map-based Edges



Edges to current node

$$p_a(G_t | G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j:e_{tj} \notin E^-} p(G_t^{tj} | G_t^-)$$

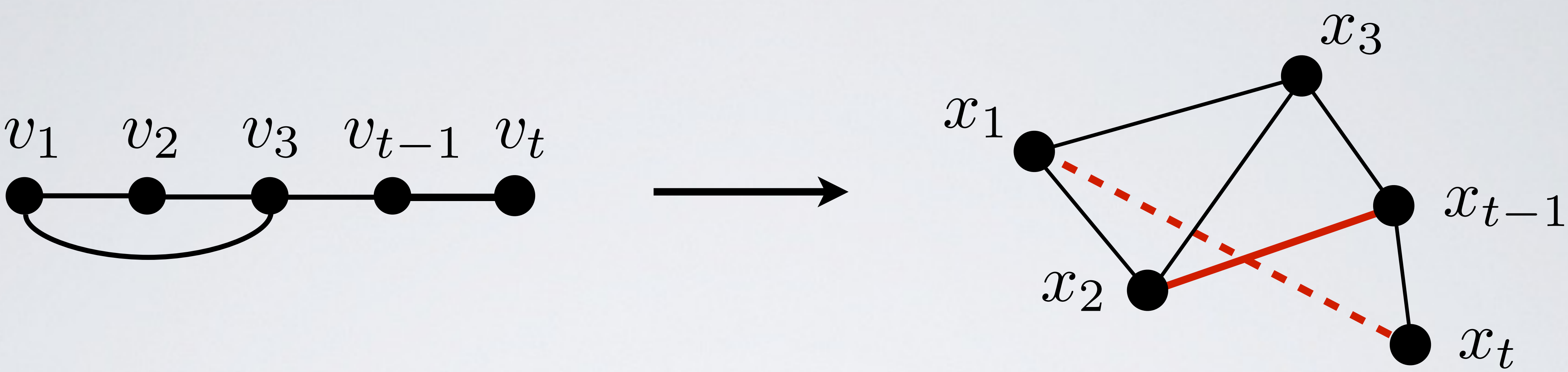
Assume edges are independent

Marginalize over the metric map

$$= \prod_{j:e_{tj} \notin E^-} \int_{X_t^-} p(G_t^{tj} | X_t^-, G_t^-, u_t) p(X_t^- | G_t^-)$$

$$d_{tj} = |x_t - x_j|_2 \qquad \approx \prod_{j:e_{tj} \notin E^-} \int_{d_{tj}} p(G_t^{tj} | d_{tj}, G_t^-) p(d_{tj} | G_t^-)$$

Proposal Distribution: Metric Map-based Edges



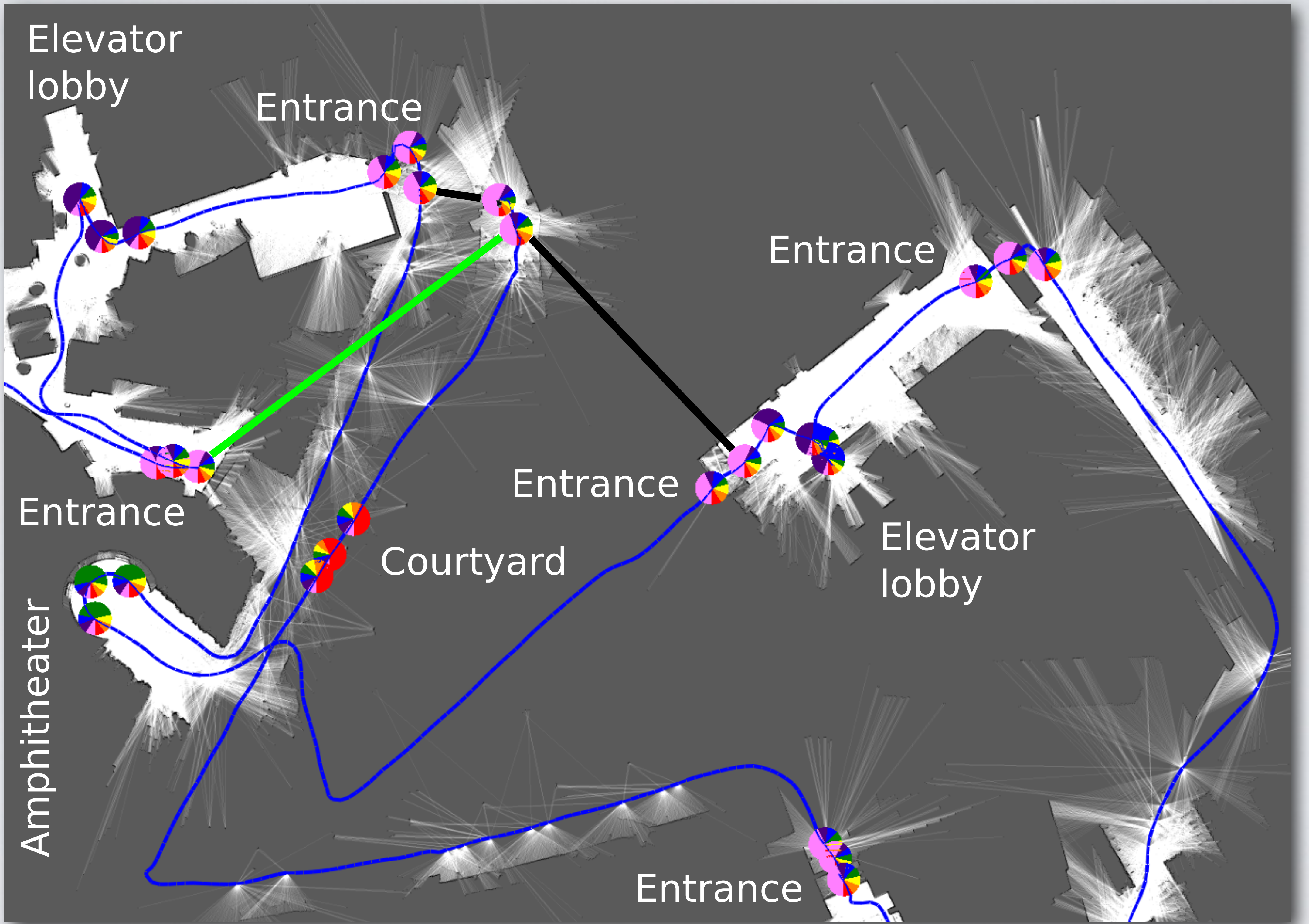
$$\mathcal{N}^{-1}(X_t^-; \Sigma_t^{-1}, \eta_t)$$

$$p_a(G_t | G_t^-, z^{t-1}, u^t, \lambda^t) = \prod_{j : e_{tj} \notin E^-} \int_{d_{tj}} p(G_t^{tj} | d_{tj}, G_t^-) p(d_{tj} | G_t^-)$$

Folded Gaussian

$$p(G_t^{ij} | d_{ij}, G_t^-, z^{t-1}, u^t) \propto \frac{1}{1 + \gamma d_{ij}^2}$$

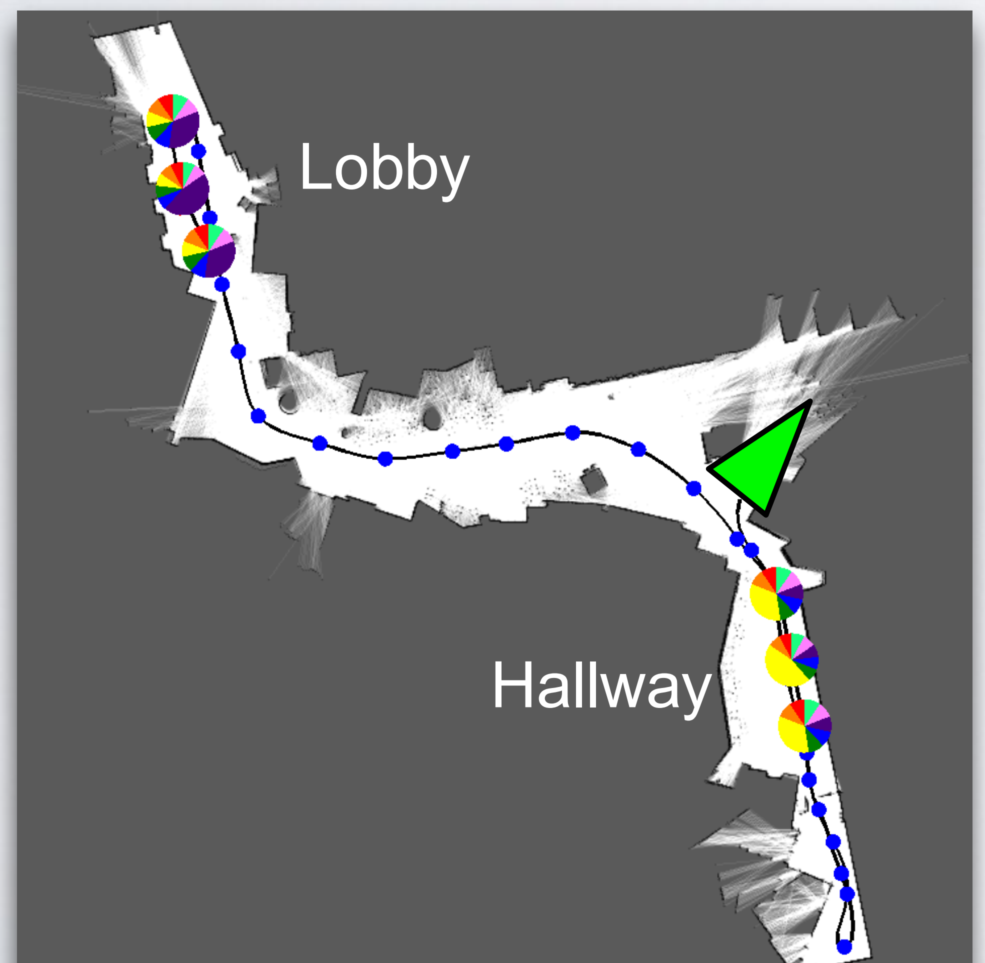
Proposal Distribution: Graph Augmentation



Two Forms of Natural Language Descriptions

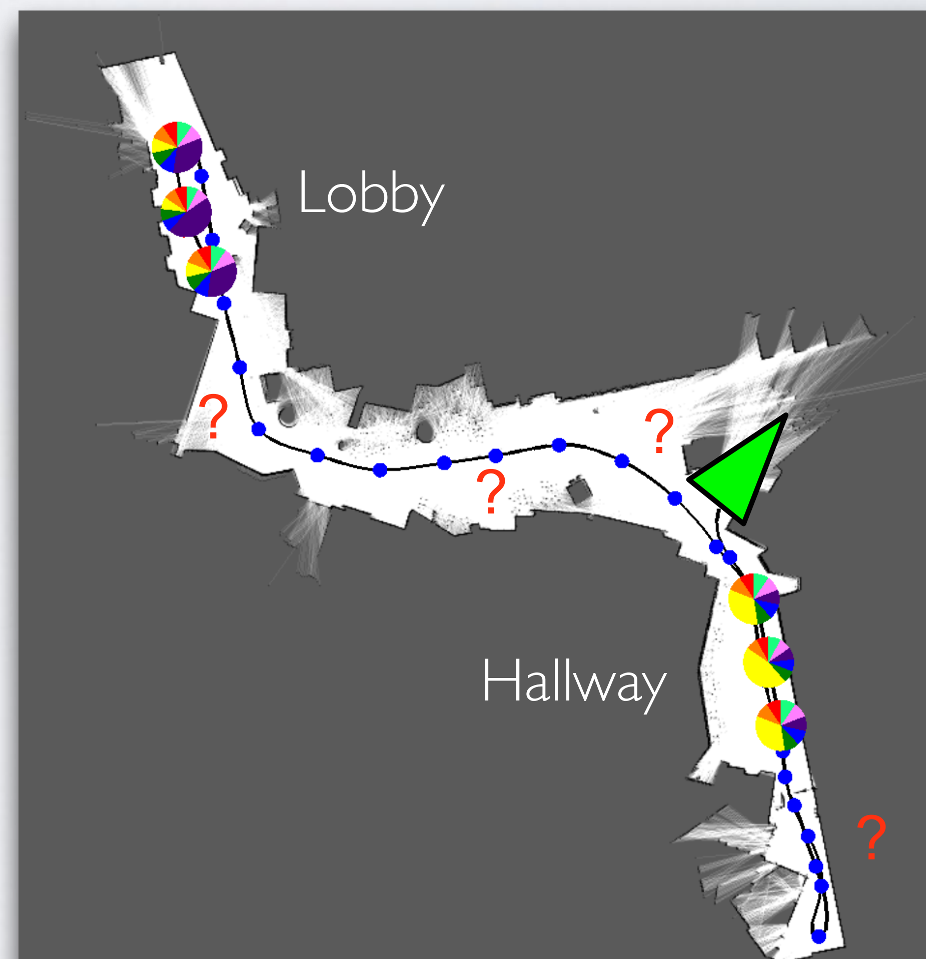
Egocentric

"This is the kitchen"



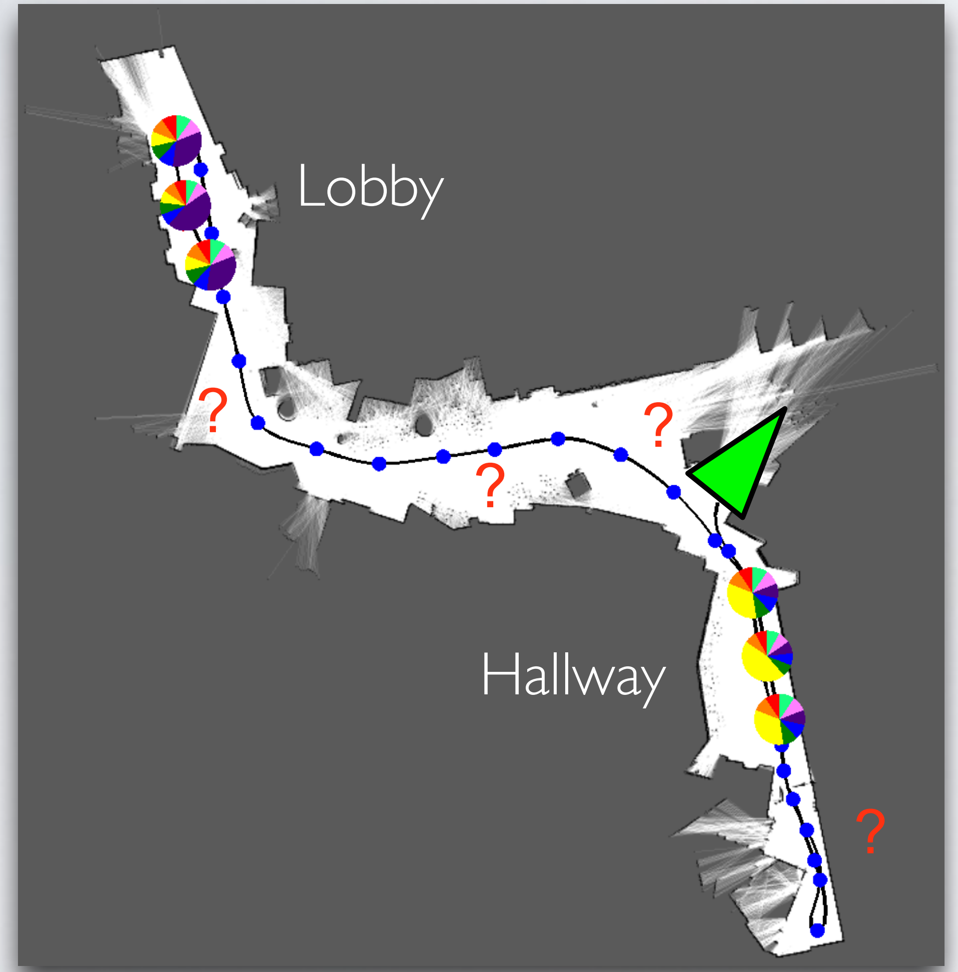
Allocentric

"The kitchen is down the hall"



Grounding Allocentric Language

"The kitchen is down the hall"



$$\text{<figure>}\ \text{<relation>}\ \text{<landmark>}$$

likelihood of the relation

$$p(\phi^f_{R_i} = \mathtt{T}) = \sum_{R_j} p(\phi^f_{R_i} = \mathtt{T} \mid \gamma_l = R_j, SR_k)\, p(\gamma_l = R_j)$$

likelihood that language references region R_i

likelihood that region R_j is the landmark

$$p(\gamma_l = R_j) = \frac{p(\phi^l_{R_j} = \mathtt{T})}{\sum_{R_j} p(\phi^l_{R_j} = \mathtt{T})}$$

Symbol Grounding Problem

Linguistic elements → Correct referents in the robot's world model

"Grounding"

The gym is down the corridor.

The workout center is behind you.

Down the hall, you'll find the gym past the exit sign.

The fitness center is down the corridor to the left.

The Alumni gym is on the right, past the tall filing cabinet.

The weight room is through the double doors at the end of the hall.

The Stata Center's gym is behind you, just beyond the doors to the elevator lobby.

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

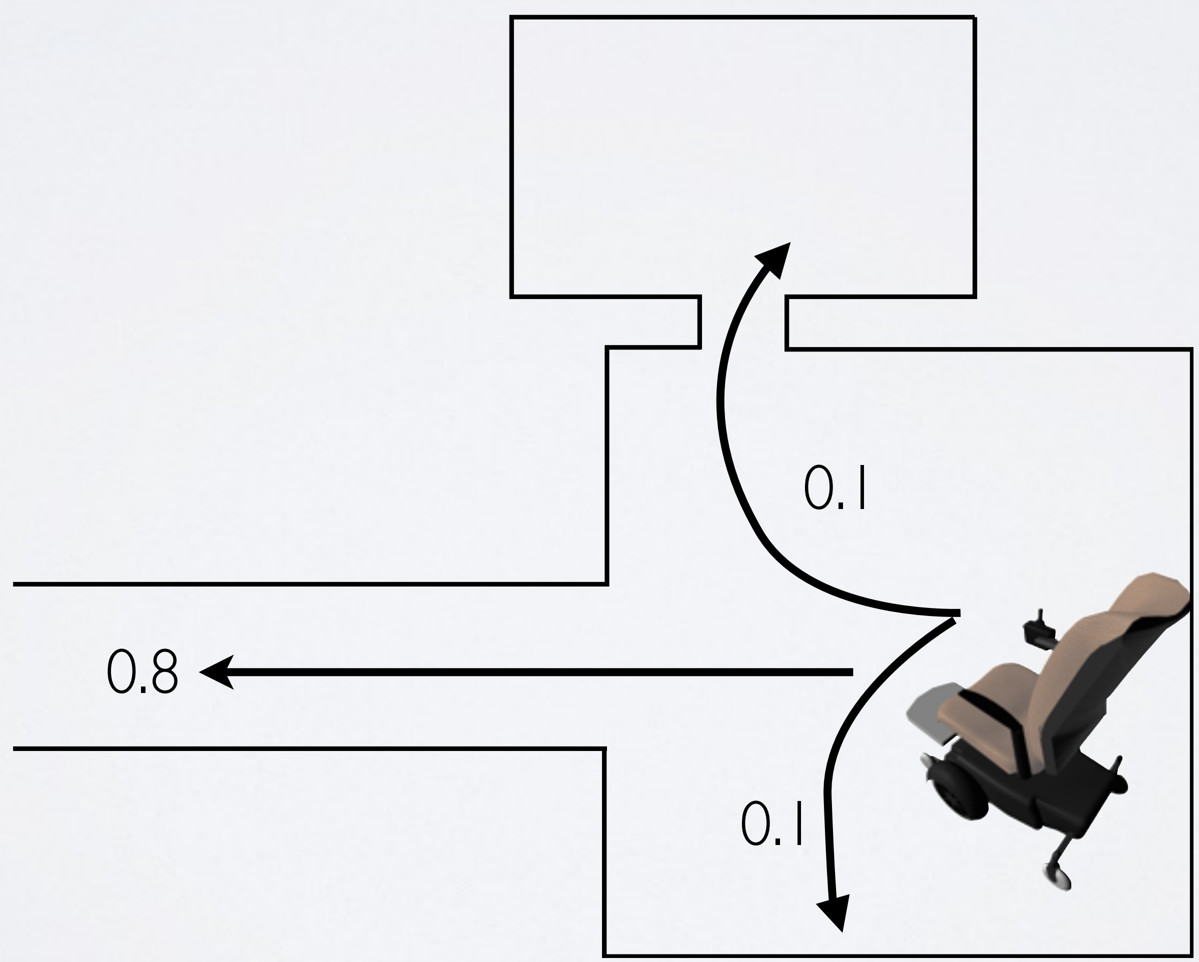
$$\arg\max_{\text{groundings}} p\,(\text{groundings}|\text{language})$$

"Down the hall"

$$\arg\max_{\gamma \in \mathcal{X}} p\,(\gamma|\lambda)$$

objects, actions, **relations**, **places**

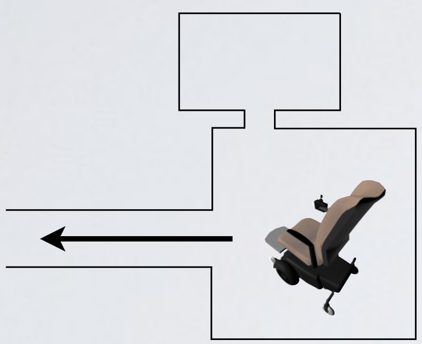
"The



Learning the Grounding Distributions

Training Set

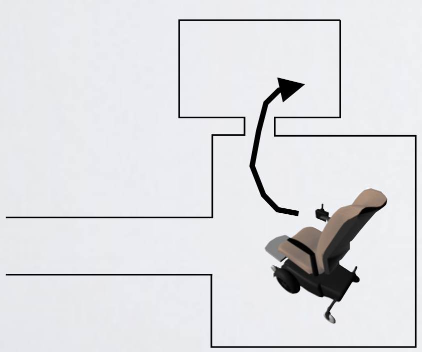
"Down the hall"



$\gamma = (x_1, y_1, z_1)$

$\vdots$

!("Down the hall")



$\gamma = (x_2, y_2, z_2)$

Extract features →

Spatial Features

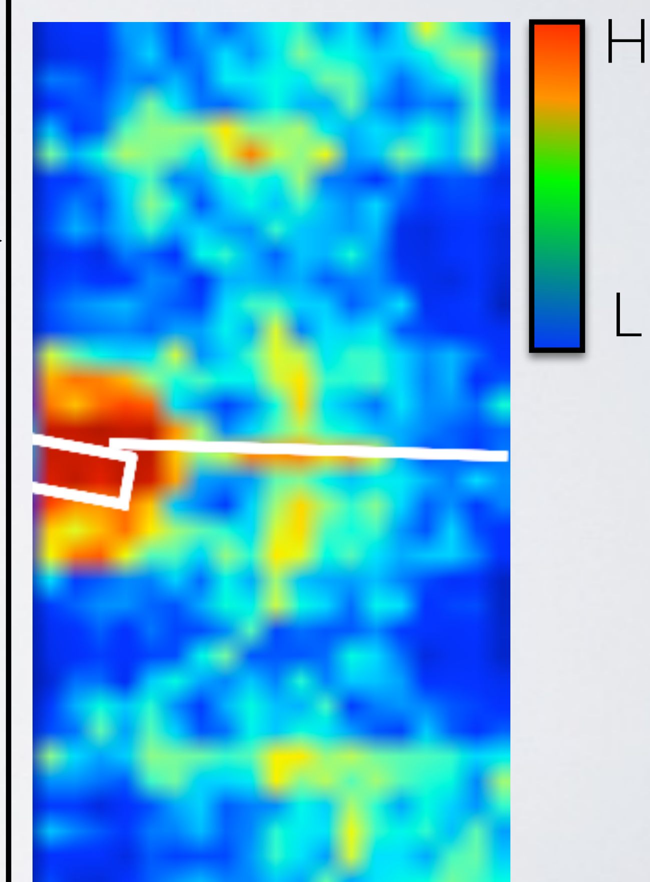
$$f_1 = 4.3$$
$$f_2 = 10.1$$
$$f_3 = 2.3$$
$$\vdots$$

$$f_1 = 2.5$$
$$f_2 = 1.1$$
$$f_3 = 8.3$$
$$\vdots$$

Optimize weights →

Learned Model

$$p(\gamma | \text{Down the hall})$$



Grounding Natural Language Speech

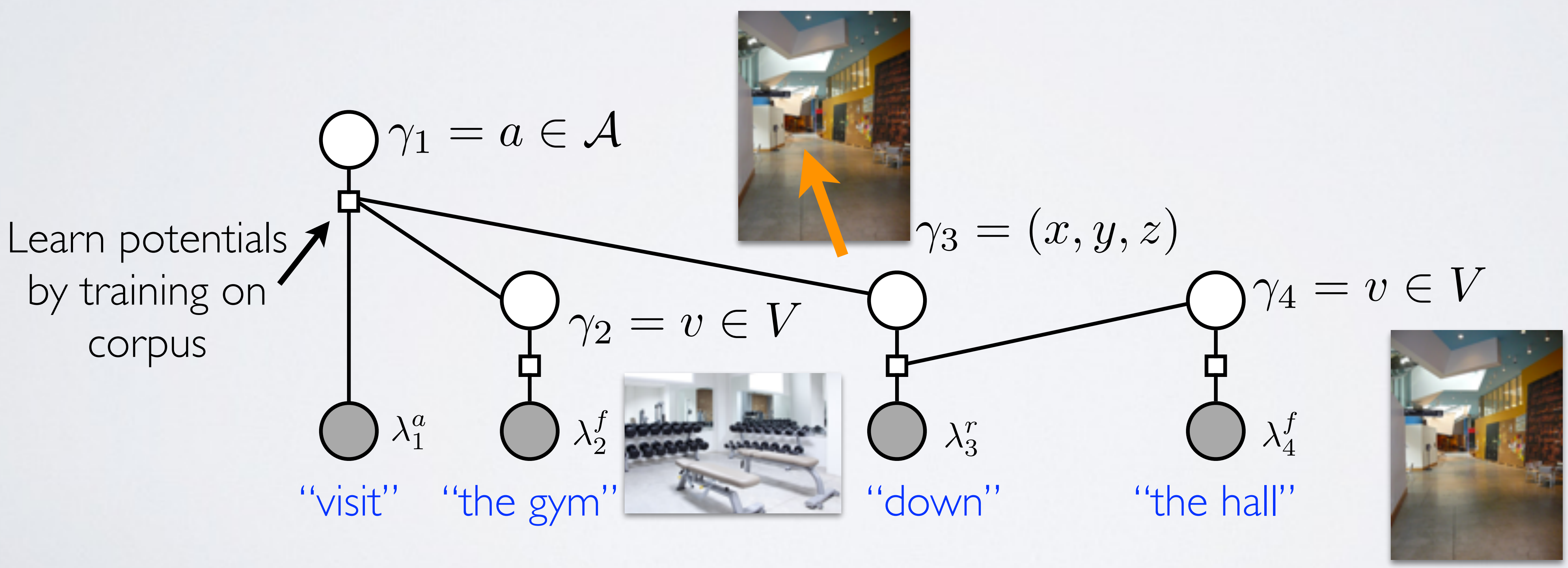
$$\underset{\text{groundings}}{\arg\max}\ p\left(\text{groundings}|\text{language}\right)$$

“Visit the gym down the hall”

$$\arg\max\ (\gamma_1, \gamma_2, \gamma_3, \gamma_4|\lambda)$$

objects, actions, relations, places

“Go to the gym down the hall”

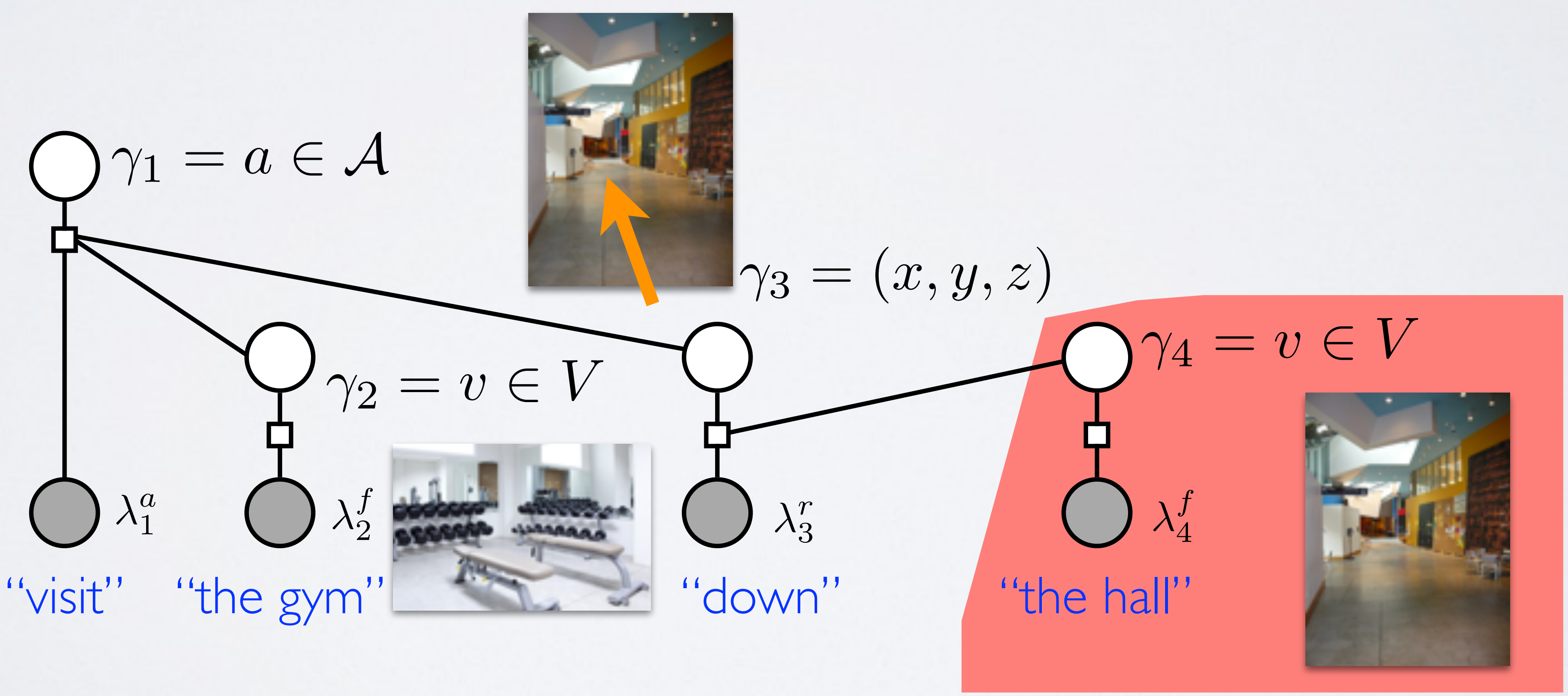


[AAAI 2011; AI Magazine 2011]

Grounding Natural Language Speech

"Visit to the gym down the hall"

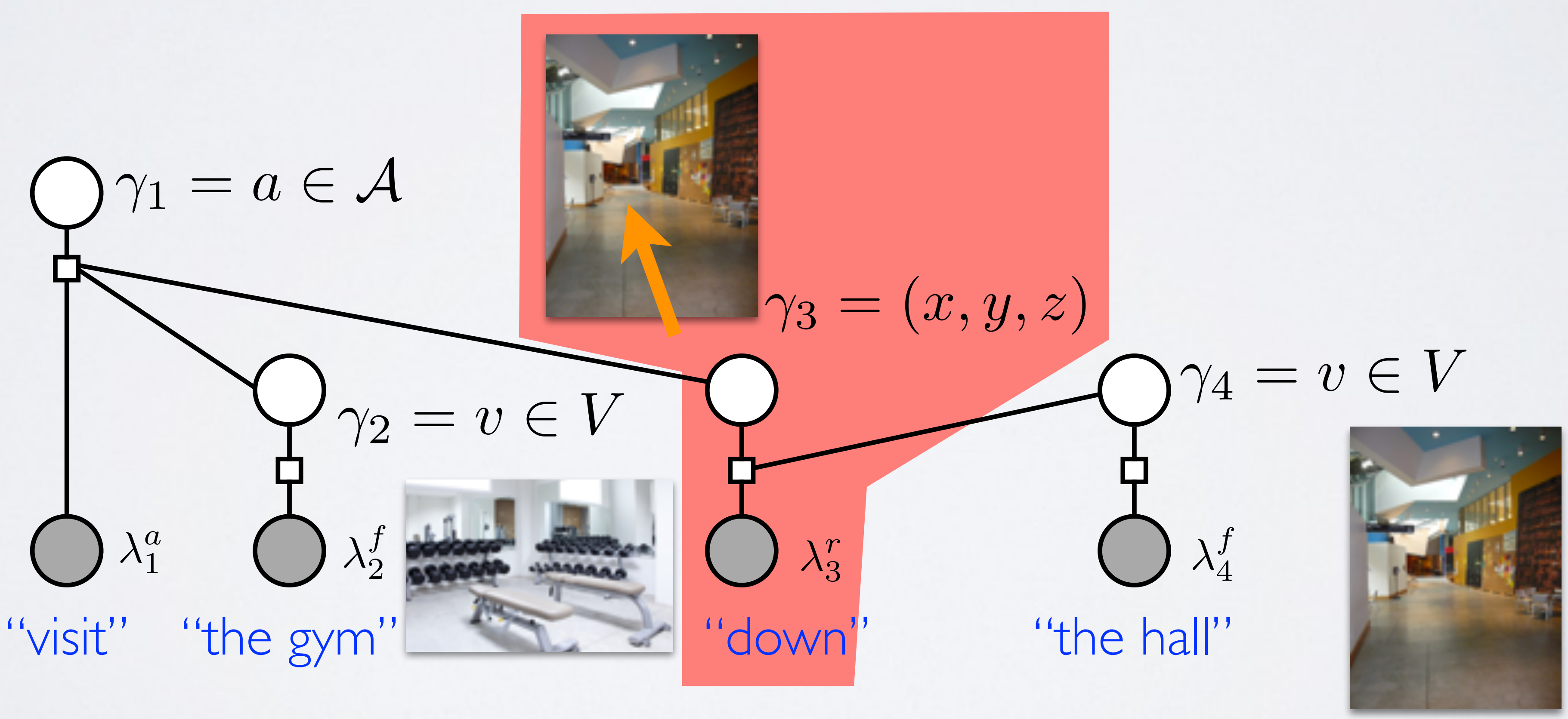
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Visit to the gym **down** the hall"

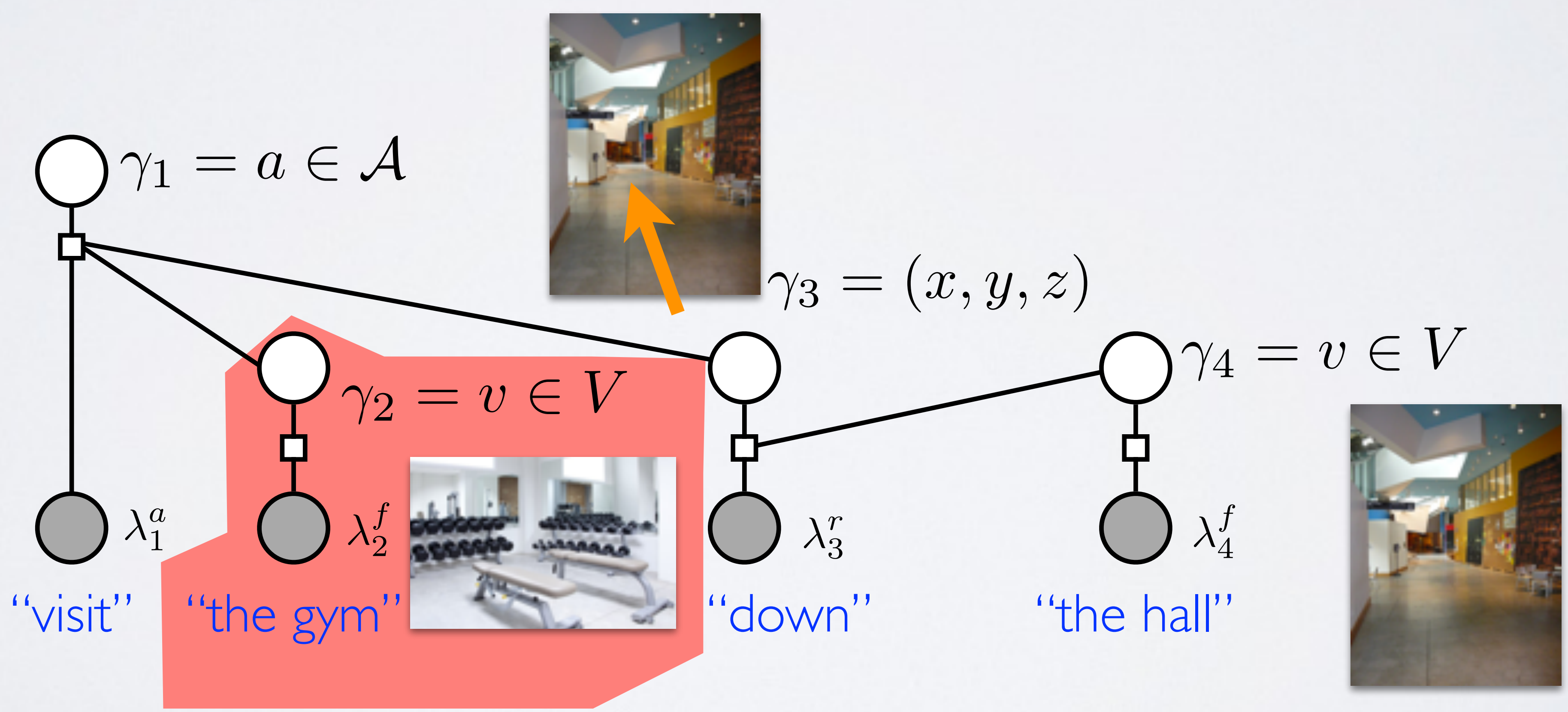
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Visit to **the gym** down the hall"

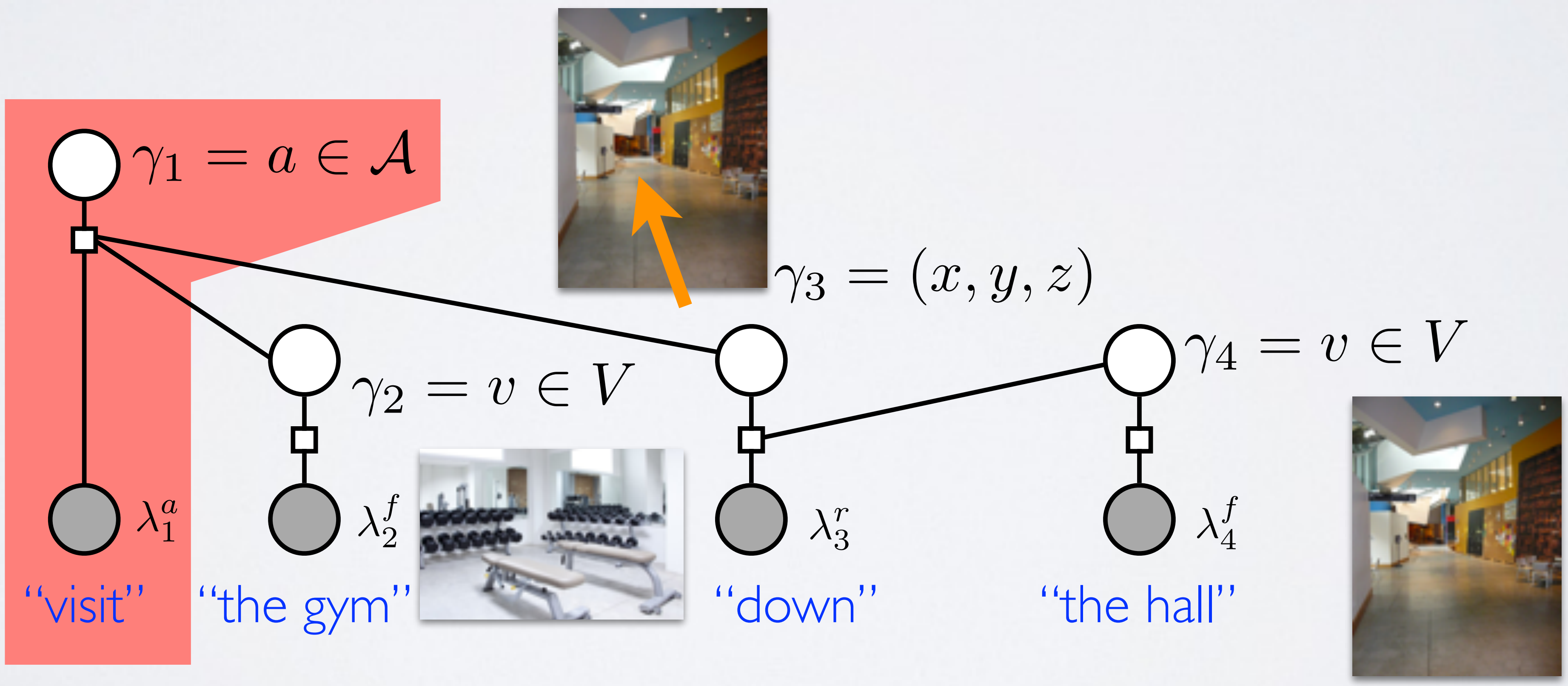
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Visit to the gym down the hall"

$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Rao-Blackwellized Particle Filter

Input: $\mathcal{P}_{t-1} = \left\{P_{t-1}^{(1)}, P_{t-1}^{(2)}, \ldots, P_{t-1}^{(n)}\right\}$ where $P_{t-1}^{(i)} = \left\{G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}$

for each particle i

1) **Proposal**: Modify the topology based on metric and semantic maps

2) **Update**: Perform Bayesian update of Gaussian

3) **Update**: Update Dirichlet over labels based on language

4) **Reweight**: Update weights based on metric observations

Return: $\mathcal{P}_{t} = \left\{P_{t}^{(1)}, P_{t}^{(2)}, \ldots, P_{t}^{(n)}\right\}$ where $P_{t}^{(i)} = \left\{G_{t}^{(i)}, X_{t}^{(i)}, L_{t}^{(i)}, w_{t}^{(i)}\right\}$

Matthew Walter

Updating Particle Weights with Sensor Data

$$w_t^{(i)} = \frac{\text{Target distribution}}{\text{Proposal distribution}} = \frac{p(G_t^{(i)} | z^t, u^t, \lambda^t)}{p(G_t^{(i)} | G_{t-1}^{(i)}, z^{t-1}, u^t, \lambda^t)} w_{t-1}^{(i)}$$

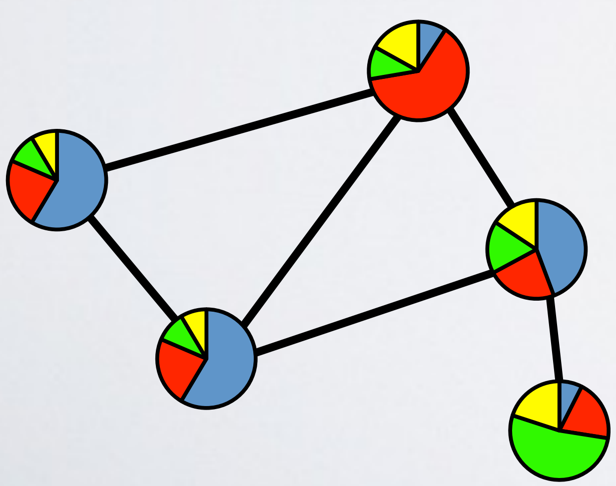
$$\tilde{w}_t^{(i)} = p(z_t | G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \cdot w_{t-1}^{(i)}$$

$$\uparrow$$

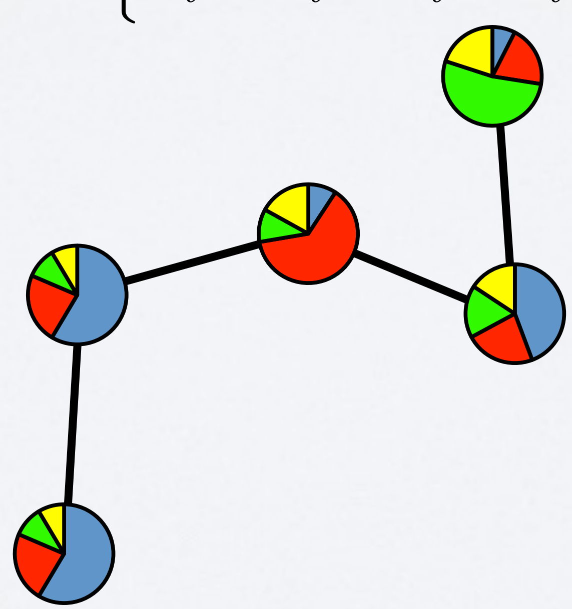
Consistency of LIDAR scans

$$p(z_t | G_t^{(i)}, z^{t-1}, u^t, \lambda^t) = \int_{X_t} p(z_t | X_t^{(i)}, G_t^{(i)}, z^{t-1}, u^t, \lambda^t) p(X_t^{(i)} | G_t^{(i)}, z^{t-1}, u^t, \lambda^t) dX_t$$

$P_t^{(1)} = \left\{ G_t^{(1)}, X_t^{(1)}, L_t^{(1)}, w_t^{(1)} \right\}$

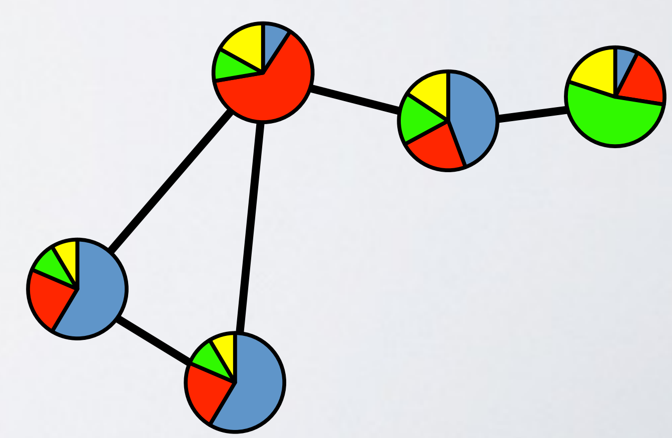


$P_t^{(2)} = \left\{ G_t^{(2)}, X_t^{(2)}, L_t^{(2)}, w_t^{(2)} \right\}$



$\cdots$

$P_t^{(n)} = \left\{ G_t^{(n)}, X_t^{(n)}, L_t^{(n)}, w_t^{(n)} \right\}$



Transition Function

$$\mathbb{E}(V(S_{t+1})) = \sum_{S_{t+1}} V(S_{t+1}) \times p(S_{t+1} | S_t, a_t)$$

$$= \sum_{z_j^a} V(z_j^a) \times p(z_j^a | S_t, a_t)$$

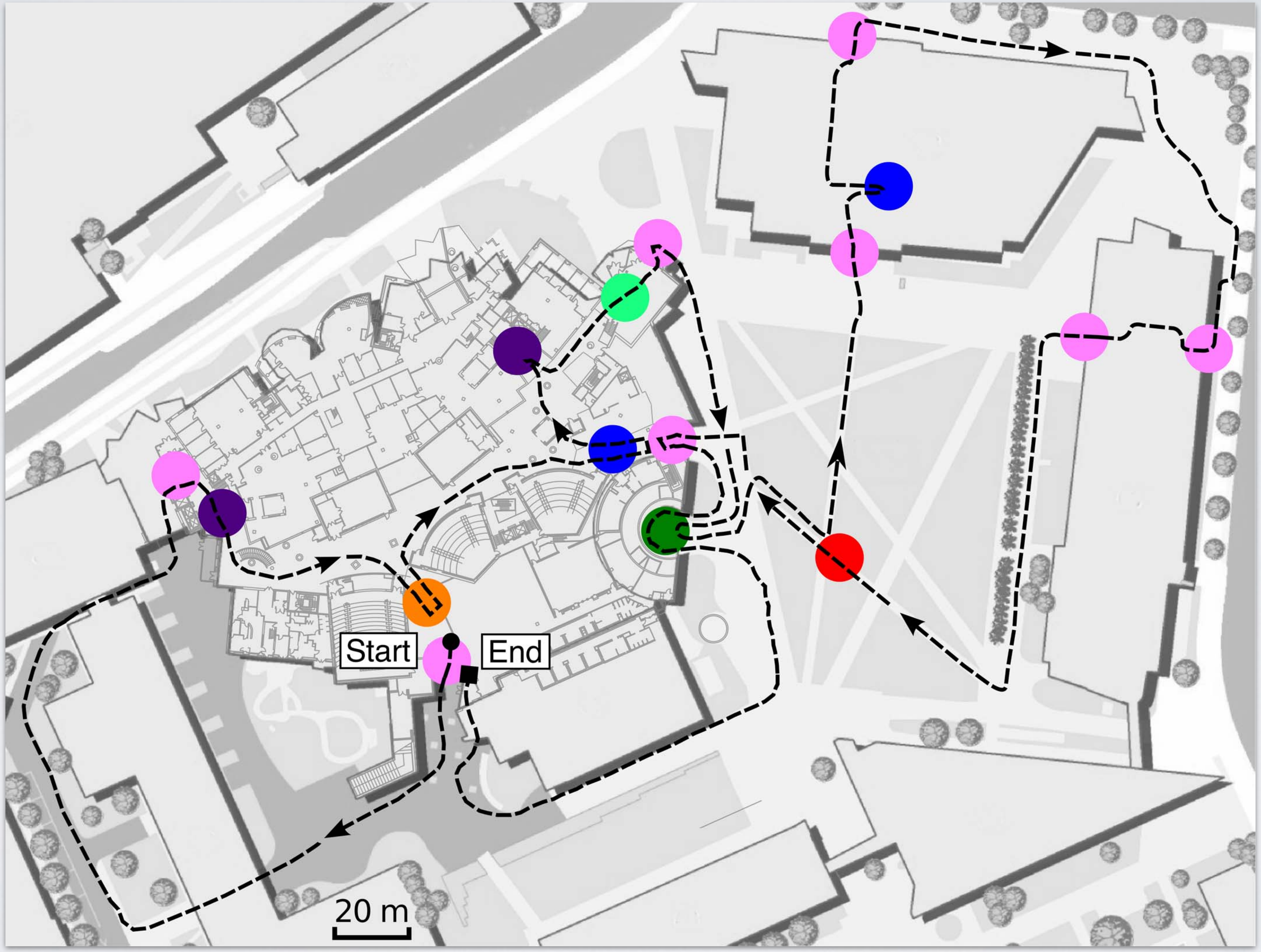
$$\downarrow$$

$$p(z_j^a | S_t, a_t) = \sum_{R_i} p(z_j^a | S_t, R_i, a_t) \times p(R_i | \Lambda_k)$$

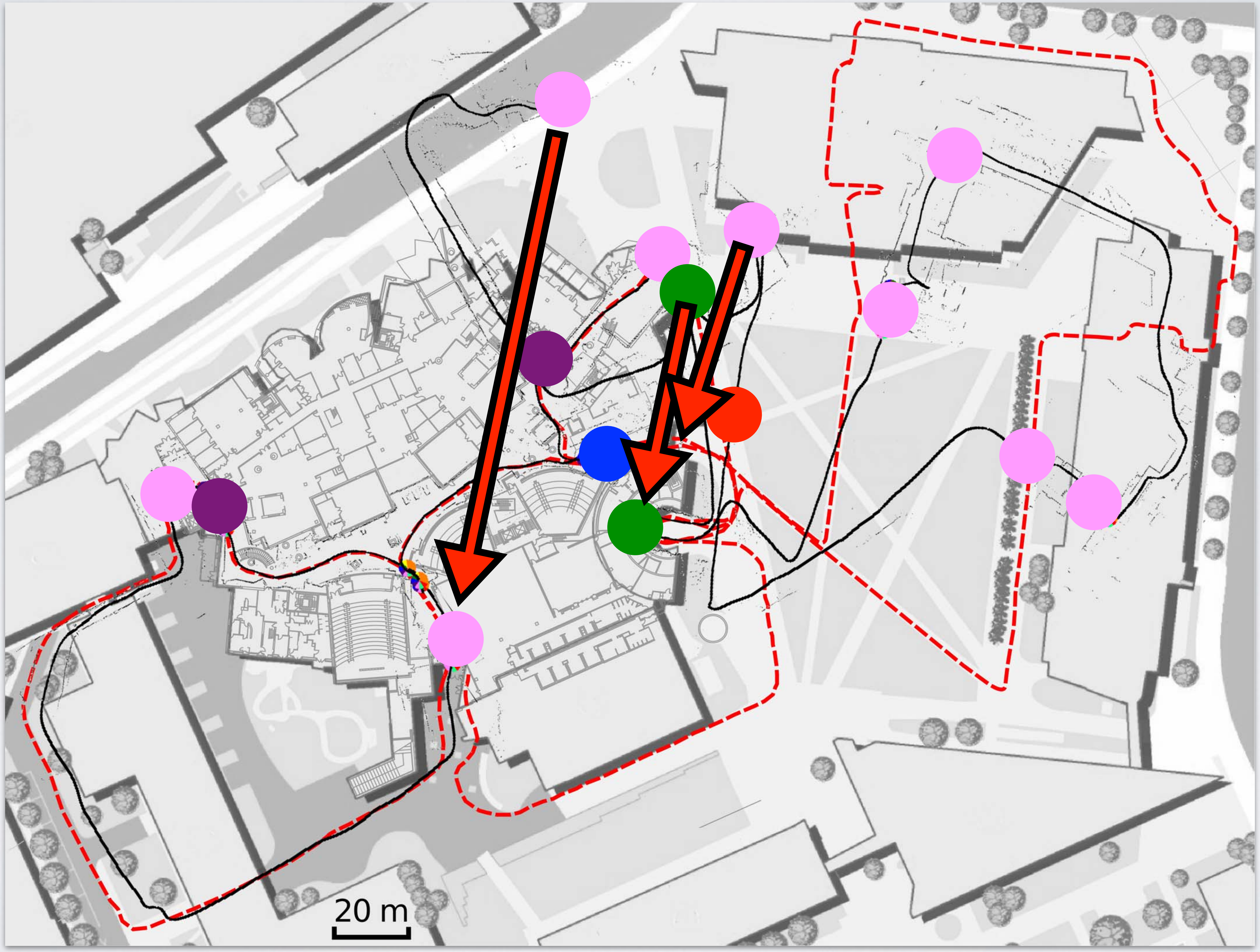
$$p(z_j^a | S_t, R_i, a_t) = \sum_{\phi \in \{F, T\}} p(z_j^a | S_t, R_i, a_t, \phi) \times p(\phi | S_t, R_i, a_t)$$

Matthew Walter

Narrated-tour Results

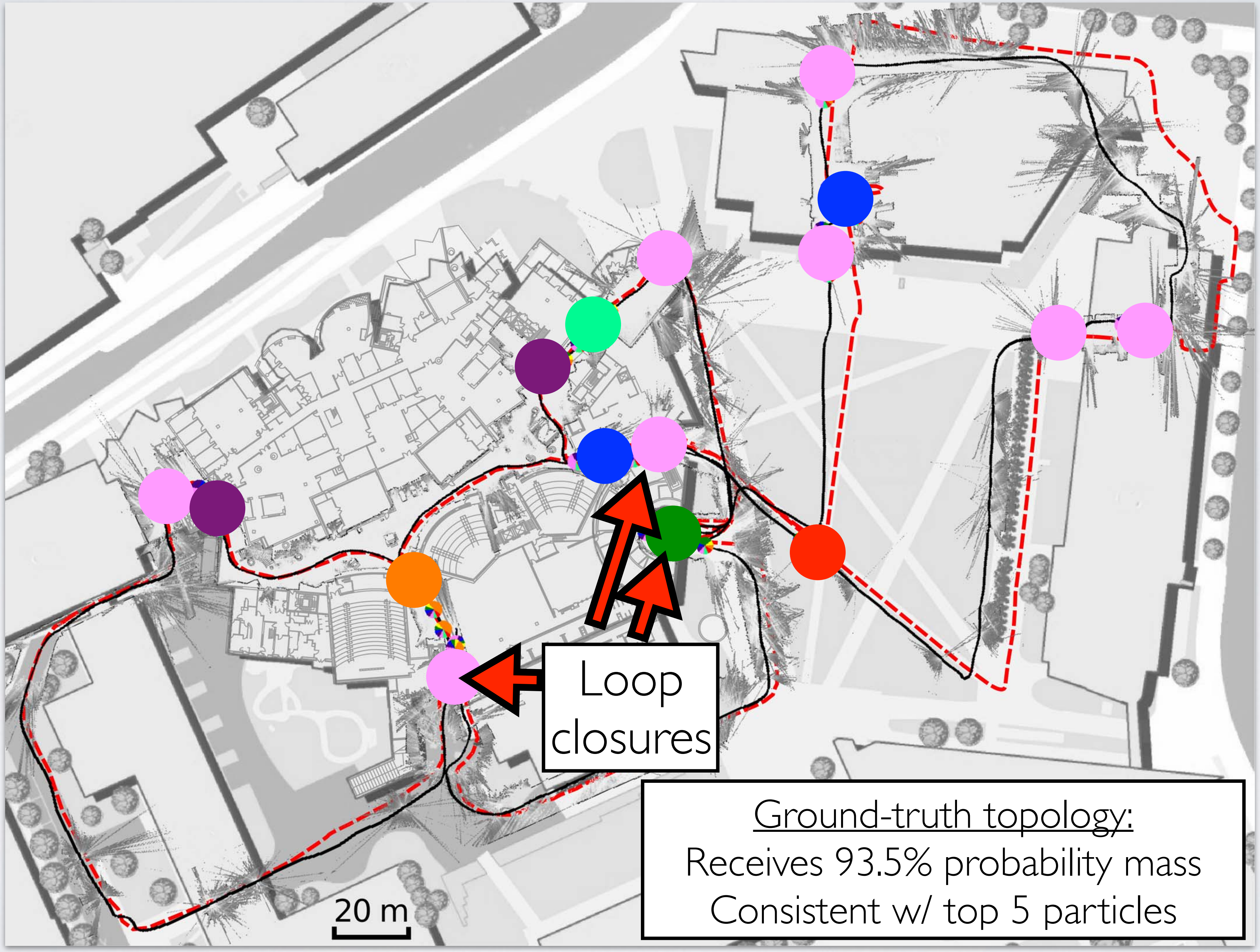


Narrated-tour Results: Baseline



Matthew Walter

Narrated-tour Results: Semantic Graph

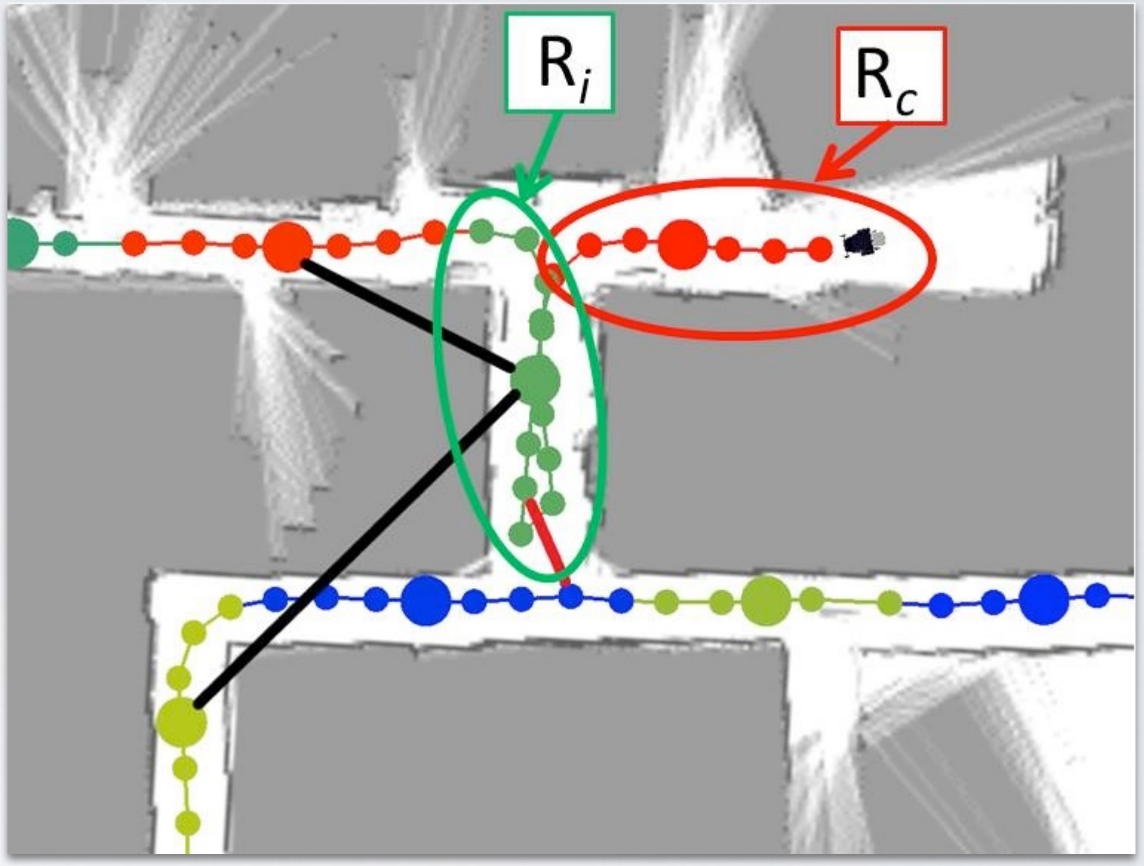


Topological Accuracy

Environment	Accuracy
Stata Floor 3	97.2%
Stata Floor 4	96.3%
Multi-building	96.2%

Region Segmentation Accuracy

Region Type	Stata Floor 3	Multi-building
Conference room	80%	81.7%
Elevator lobby	59.7%	72.8%
Hallway	49.4%	55.7%
Lab	52.8%	30.1%
Lounge	42.9%	39.4%
Office	62.5%	76.1%

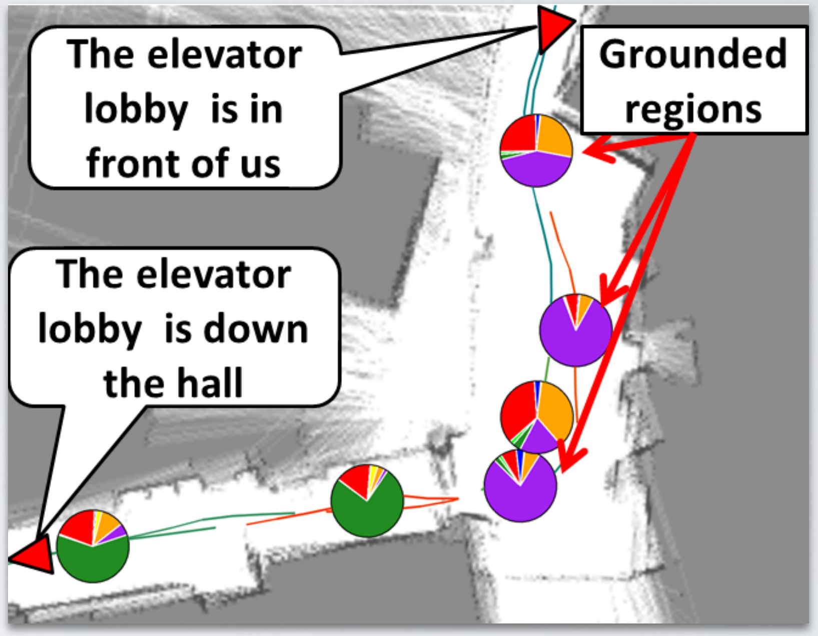


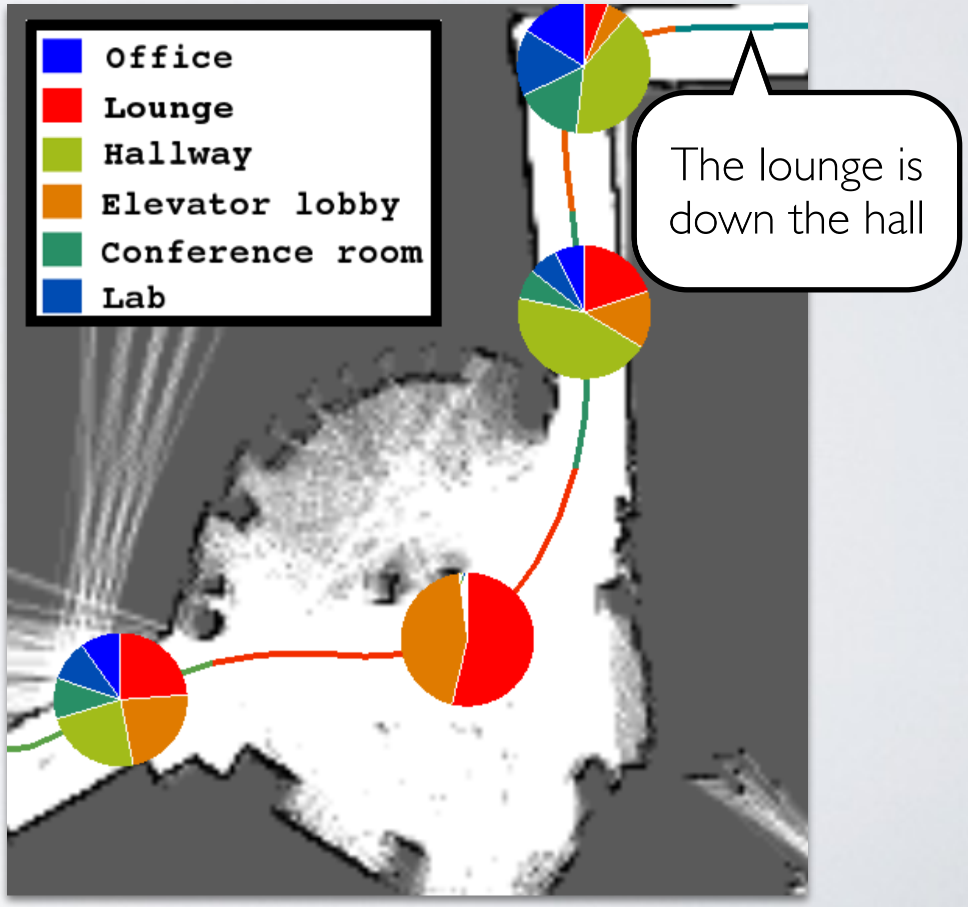
Jaccard similarity $\frac{|V_{R_i} \cap V_{R_{\text{truth}}}|}{|V_{R_i} \cup V_{R_{\text{truth}}}|}$

Cluttered regions prone to over-segmentation

Region Semantic Accuracy

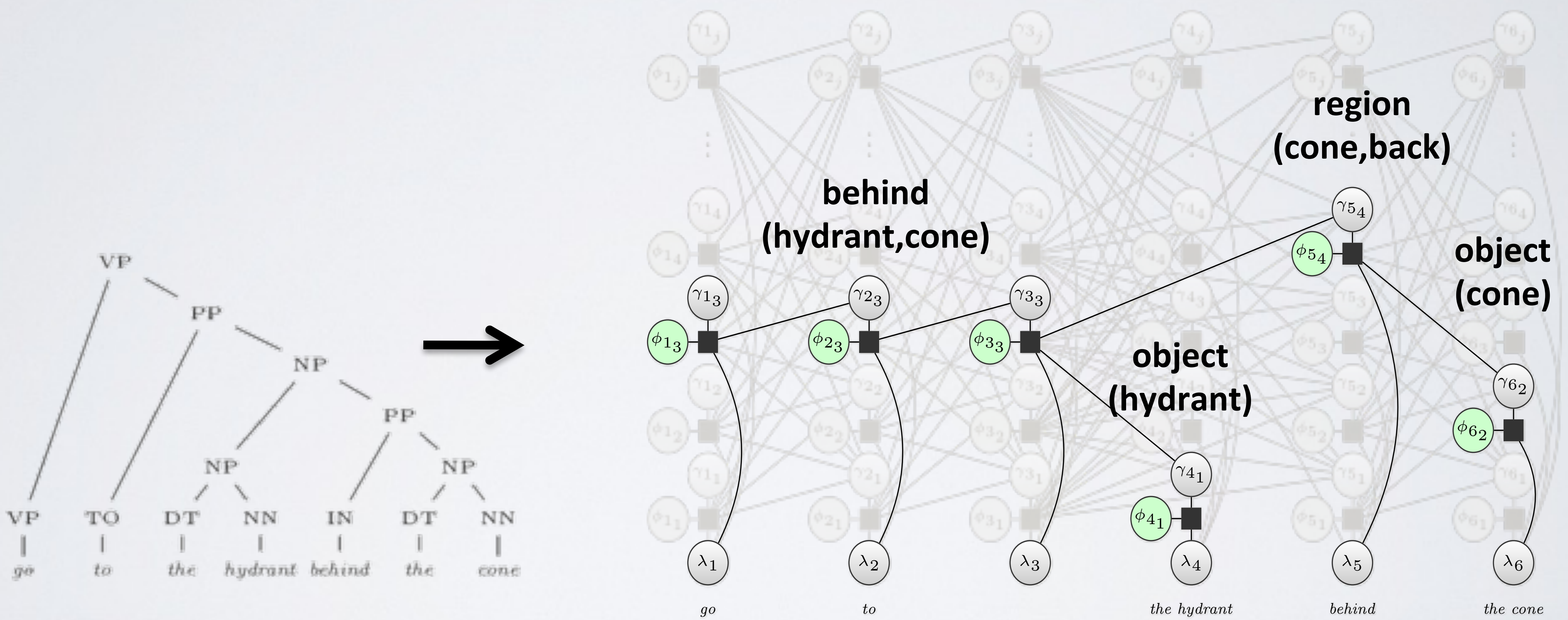
Region Type	Stata Floor 3	Multi-building
Conference room	48.5%	58.7%
Elevator lobby	64.1%	46.4%
Hallway	44.4%	58%
Lab	14.2%	30.6%
Lounge	62%	40.5%
Office	98.6%	60.2%





Annotation Inference

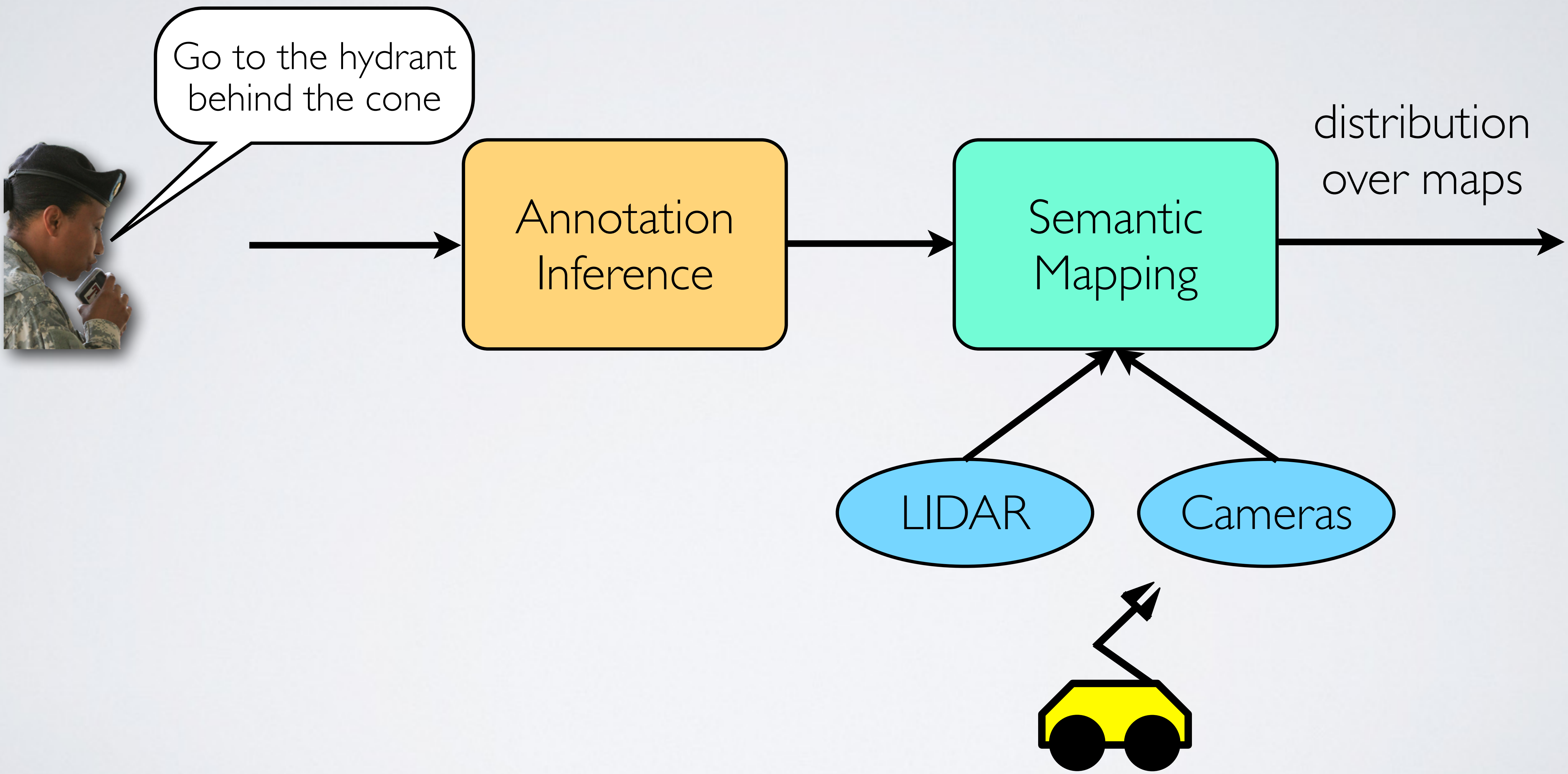
Distributed Correspondence Graph[1]:
Infer objects, locations, and relations from language



[1] Howard et al. 2014

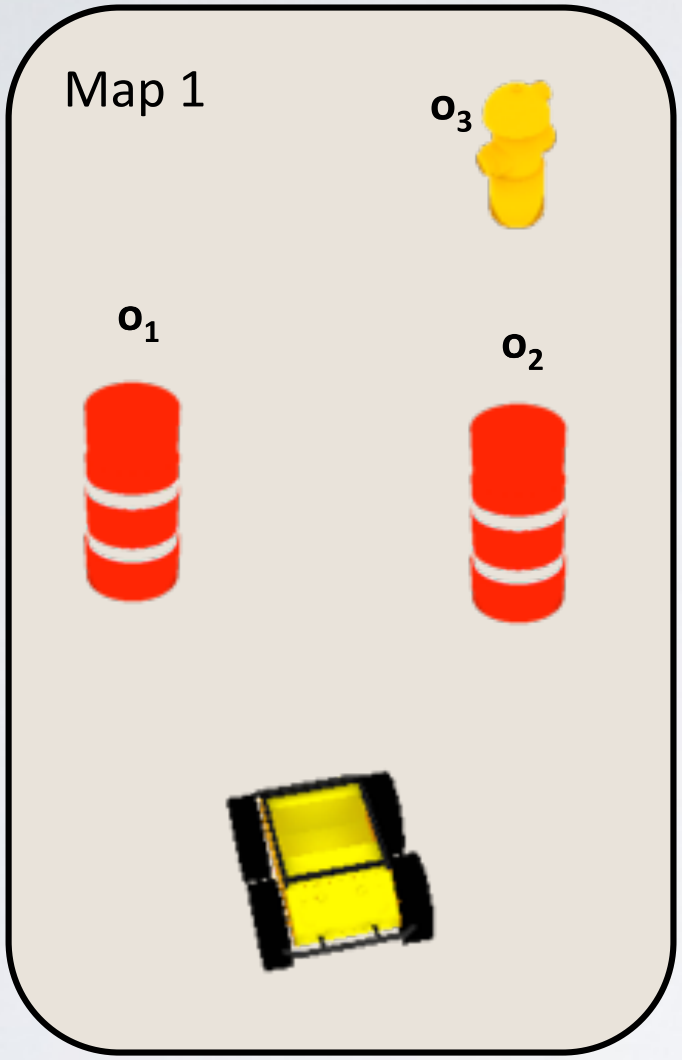
[ISER 2014]

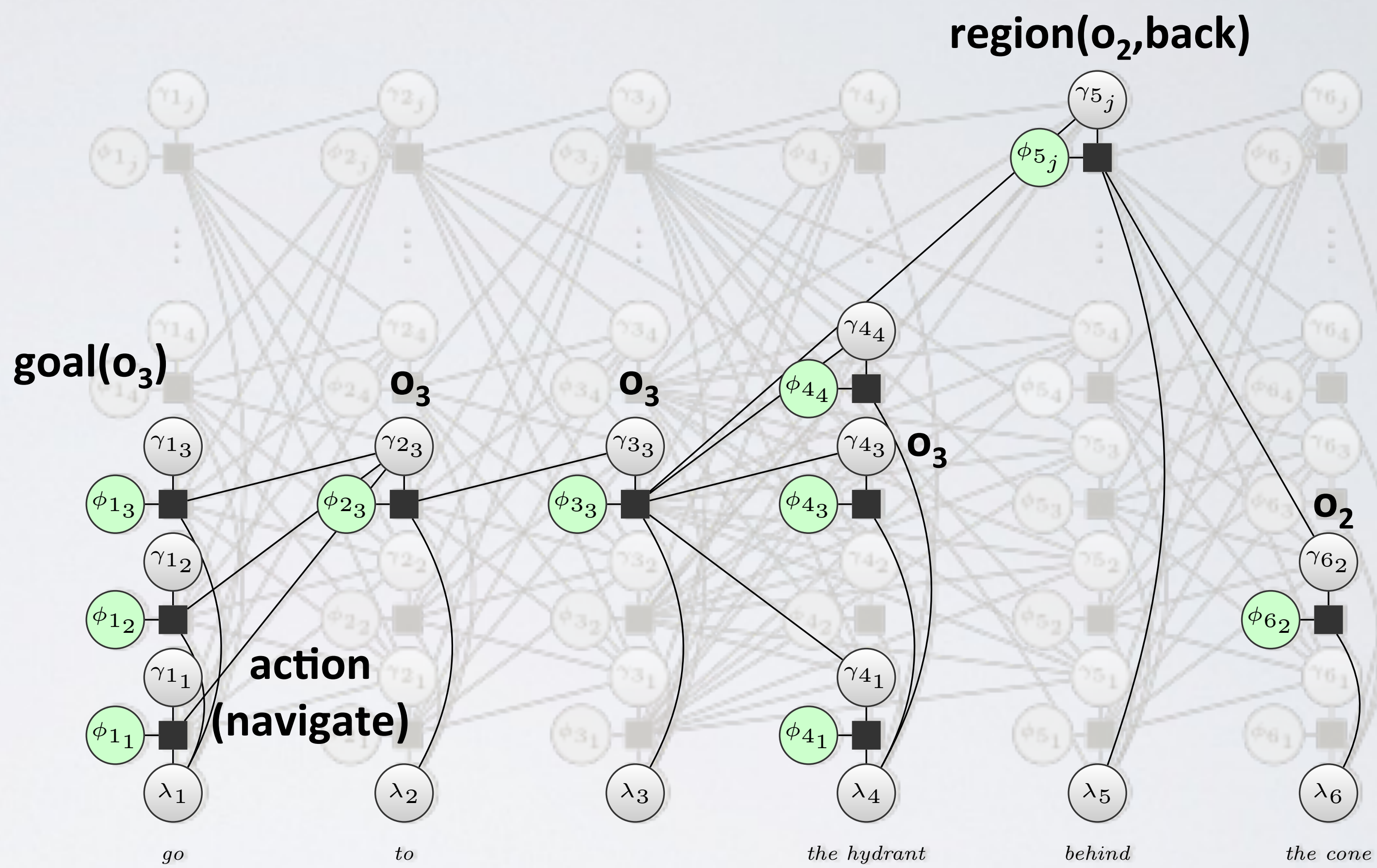
Map Inference



[ISER 2014]

Behavior Inference: Behaviors given Map Distribution





Go to the hydrant behind the cone

DCG [10]

Action: Navigate
Goal: O_3

[10] Howard et al. 2014

[ISER 2014]