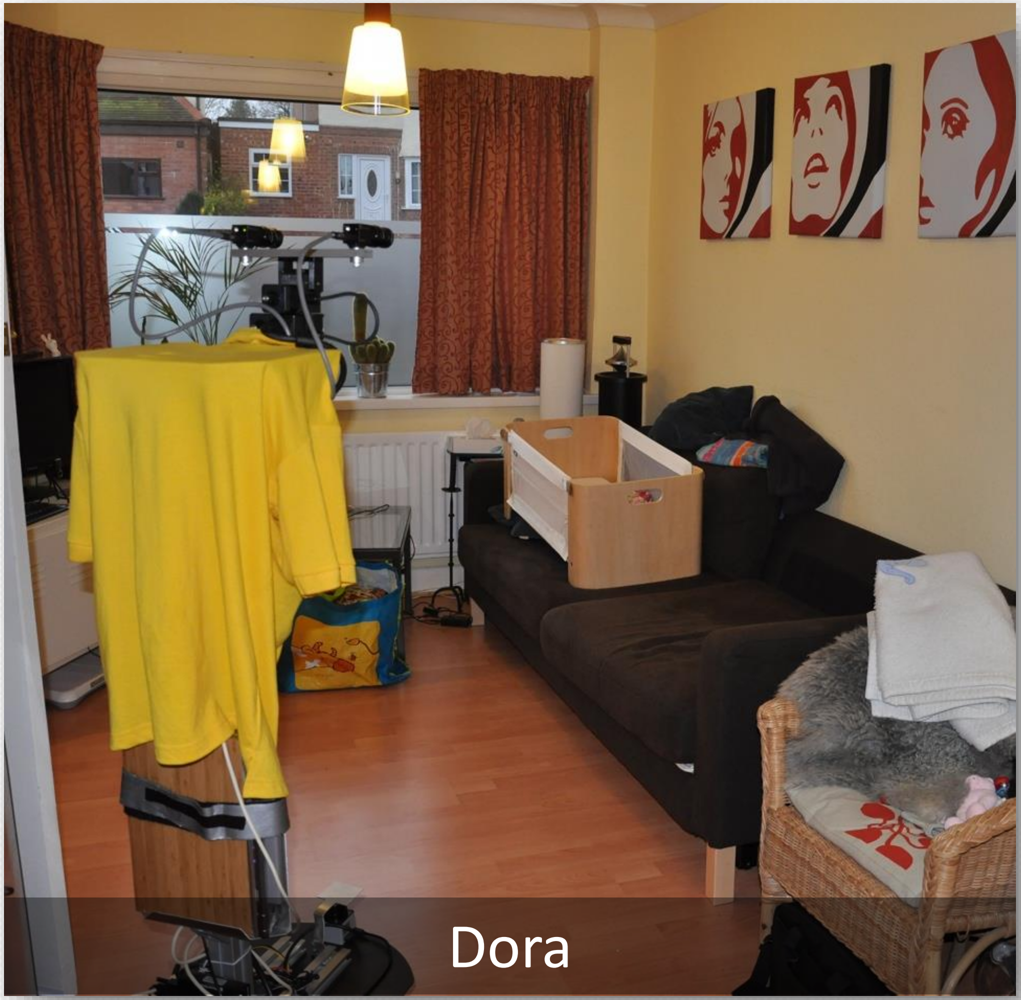
FROM SEMANTIC WORLD UNDERSTANDING TO COLLABORATION WITH DEEP REPRESENTATIONS

Andrzej PRONOBIS

University of Washington
KTH Royal Institute of Technology

www.pronobis.pro

SCENARIOS AND SYSTEMS



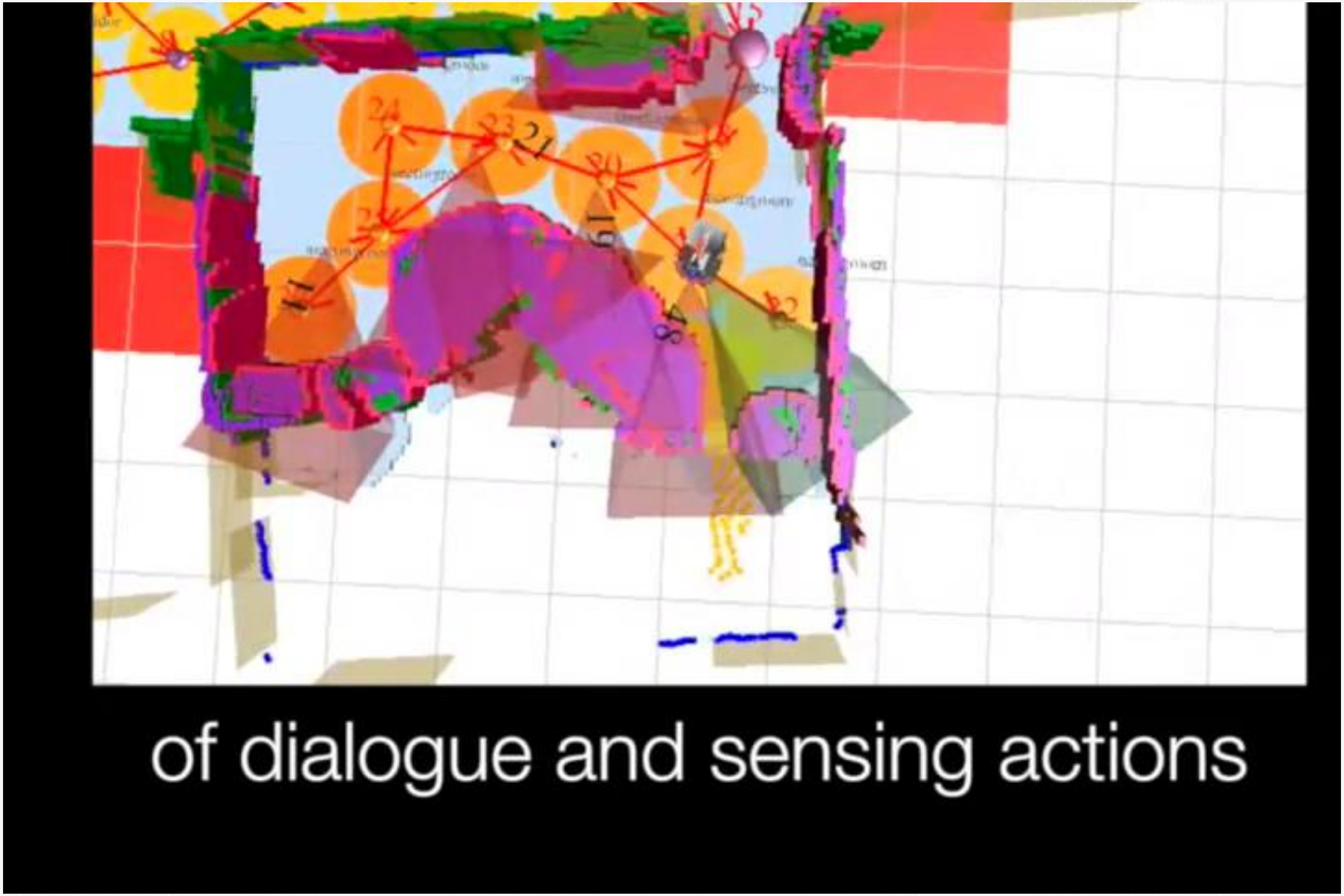
[Hanheide et al., AI'17]

[Chung*, Pronobis* et al., IROS'15, IROS'16]

DORA: ACTIVE PLACE/OBJECT SEARCH [Hanheide et al., AI'17]

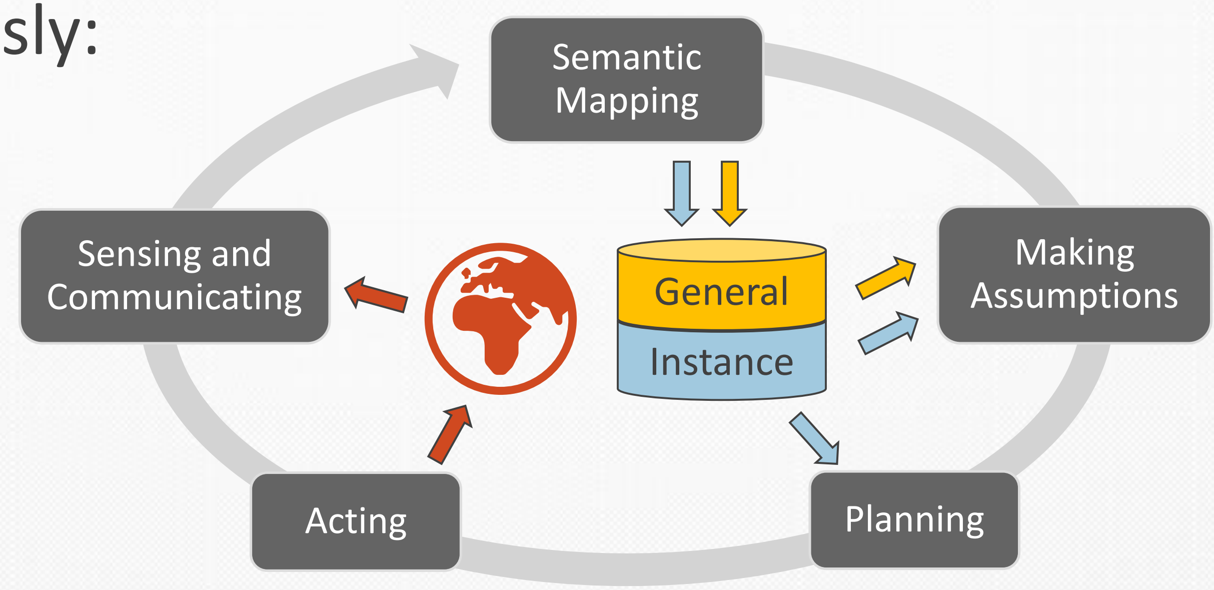
DORA: APPROACH

[Hanheide et al., AI'17]



DORA: CONTINUAL PLANNING WITH SEMANTICS

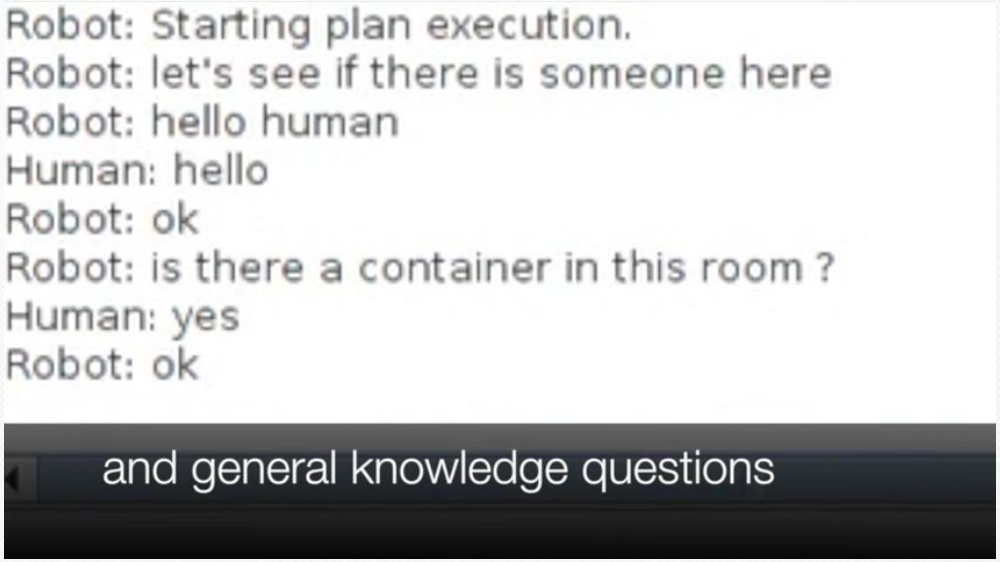
- Continual planning paradigm
- Semantic world knowledge is key
  - Instance – about current environment
  - General/common-sense – about human environments
- Continuously:



- Trades exploration vs. exploitation in a principled way

DORA: LANGUAGE-BASED INTERACTION [Hanheide et al., AI'17]

- Receiving goals in terms of semantic attributes
- Verbalizing explanations of failures
- Gathering unobservable instance knowledge
  - Polar and open questions
- Gathering common-sense knowledge



SCENARIOS AND SYSTEMS

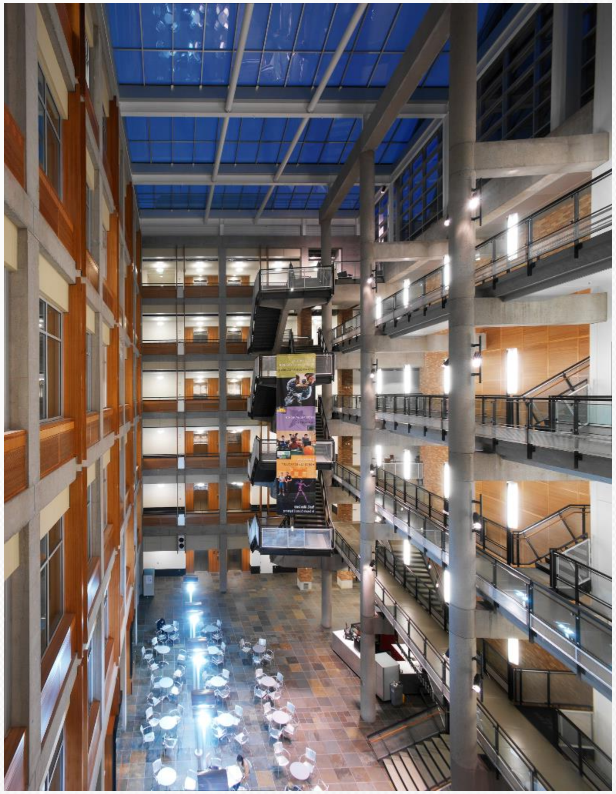
[Hanheide et al., AI'17]

[Chung*, Pronobis* et al., IROS'15, IROS'16]

INFOBOT: CONCEPT

[Chung*, Pronobis* et al., IROS'15, IROS'16]

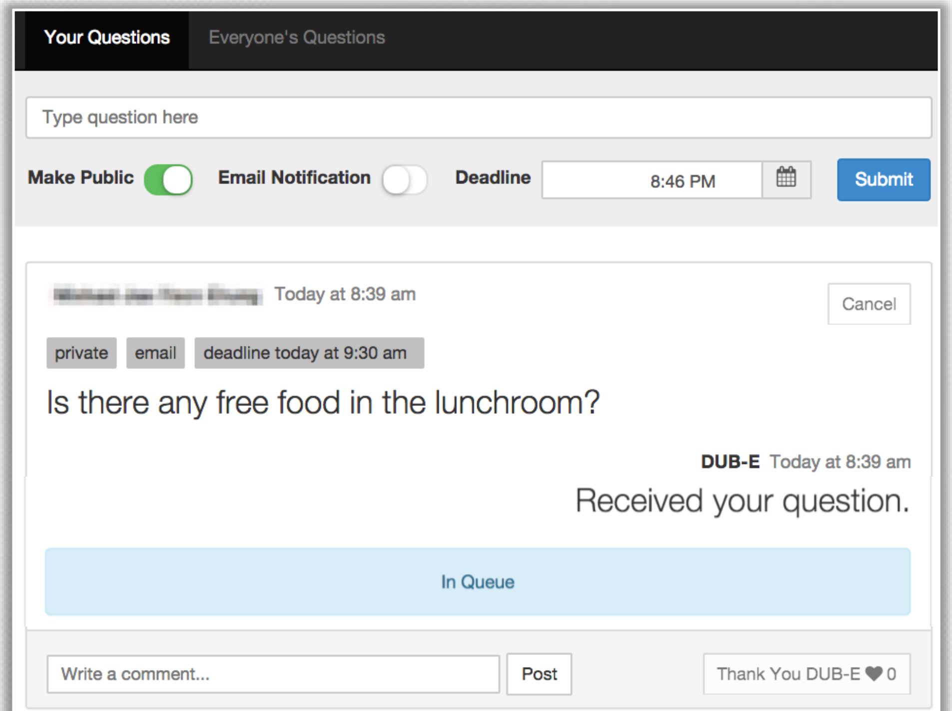
- Real-world environments are dynamic
- Humans in large environments spend time collecting information about current state:
- New type of service robots
  - Assistants answering questions about up-to-date state
  - Assist human co-inhabitants in daily tasks



INFOBOT: IMPLEMENTATION

[Chung*, Pronobis* et al., IROS’15, IROS’16]

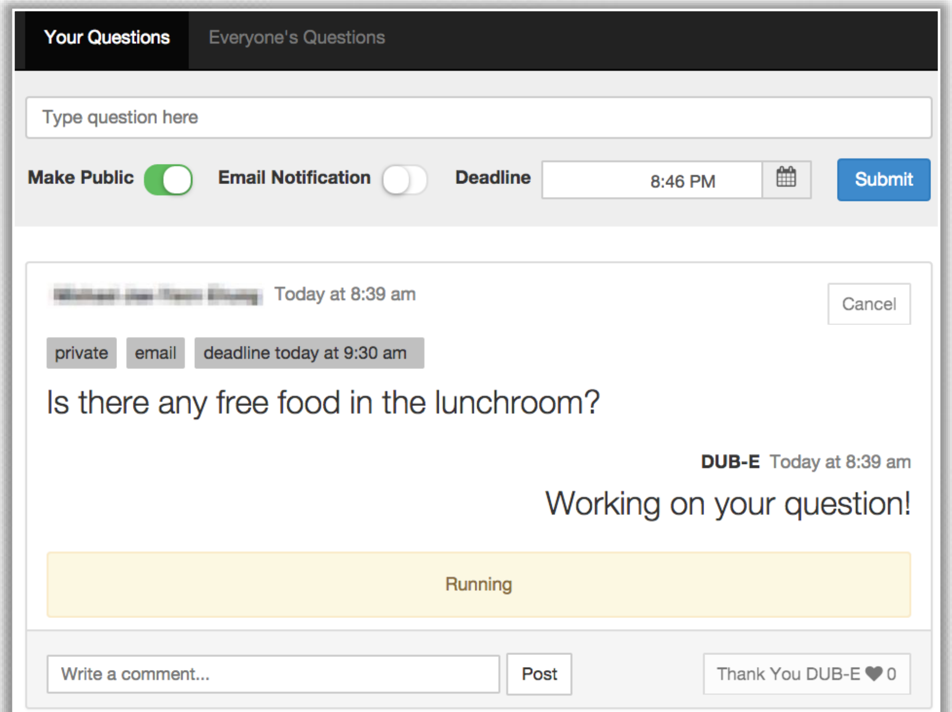
End-to-end system with web interface



INFOBOTS CONCEPT

[Chung*, Pronobis* et al., IROS’15, IROS’16]

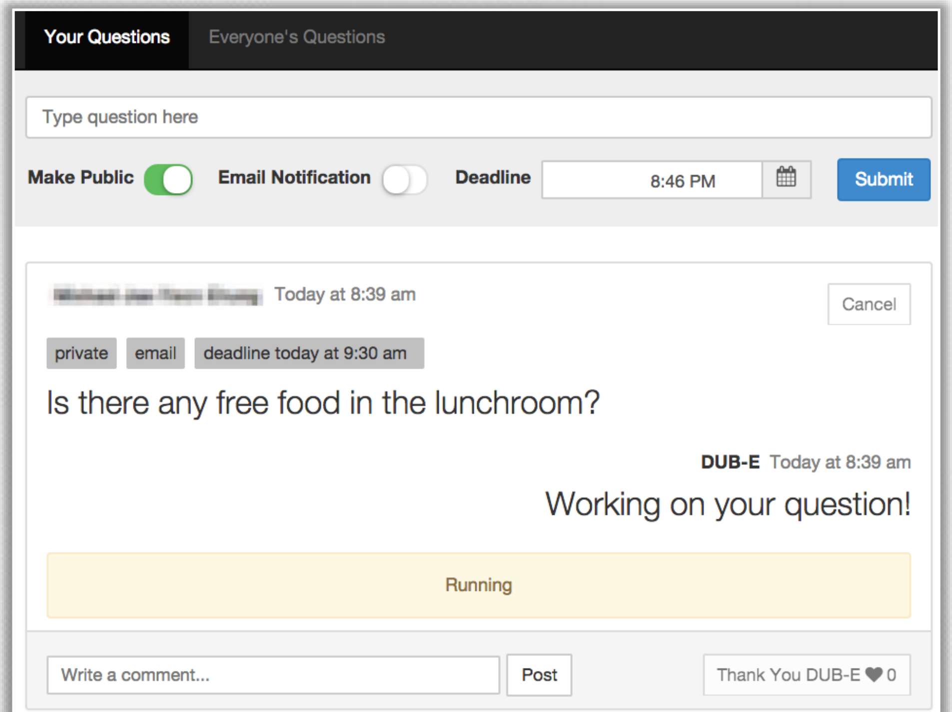
End-to-end system with web interface



INFOBOTS CONCEPT

[Chung*, Pronobis* et al., IROS’15, IROS’16]

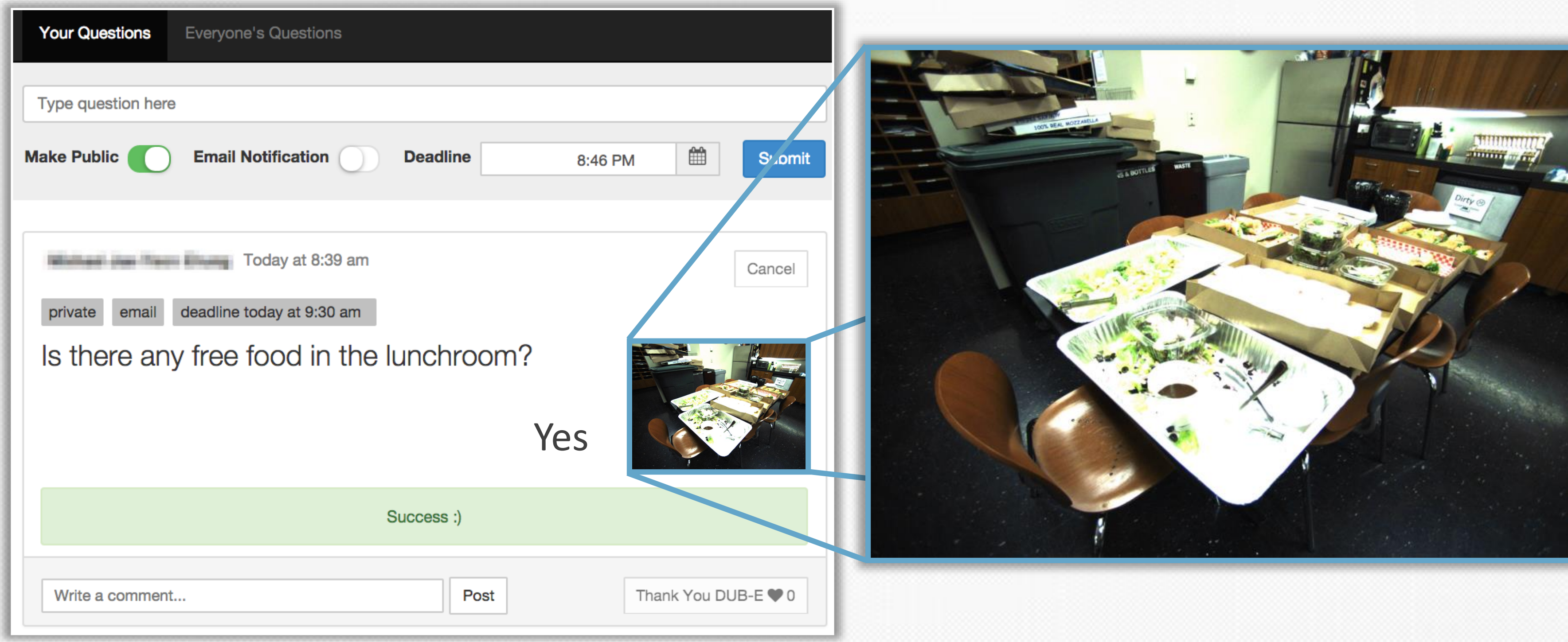
End-to-end system with web interface



INFOBOTS CONCEPT

[Chung*, Pronobis* et al., IROS’15, IROS’16]

End-to-end system with web interface



INFOBOT: HRI STUDY

[Chung et al., IROS'15]

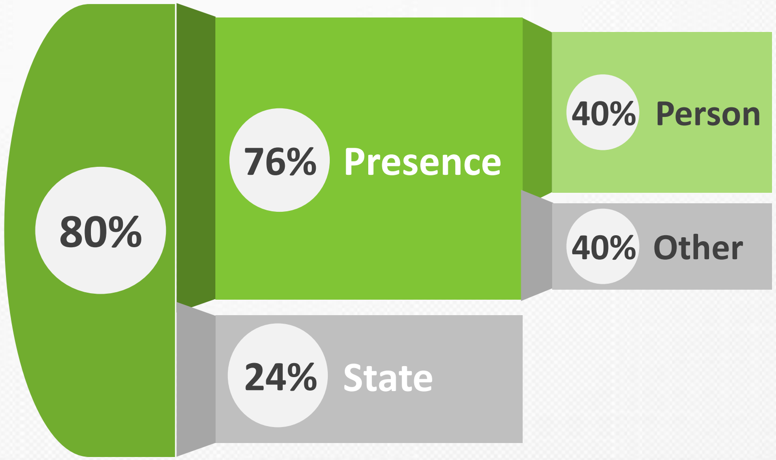
- Goal: practical usage, typical questions
- Comprehensive survey (2 buildings, 111 responses)
- Wizard-of-Oz deployment (4 days, 45 unique users)
  - Users use the web interface
  - Operator tele-operates the robot and posts answers

INFOBOT: HRI STUDY

[Chung et al., IROS'15]

- Goal: practical usage, typical questions
- Comprehensive survey (2 buildings, 111 responses)
- Wizard-of-Oz deployment (4 days, 45 unique users)
  - Users use the web interface
  - Operator tele-operates the robot and posts answers
- Questions:

Is there anyone in {location}?
Is {person} in his/her office?
Is there any food in the downstairs kitchen?
Is there anything in my mailbox?
Is the door to the conference room open?
Is the reception still open?
How noisy is it in the atrium right now?

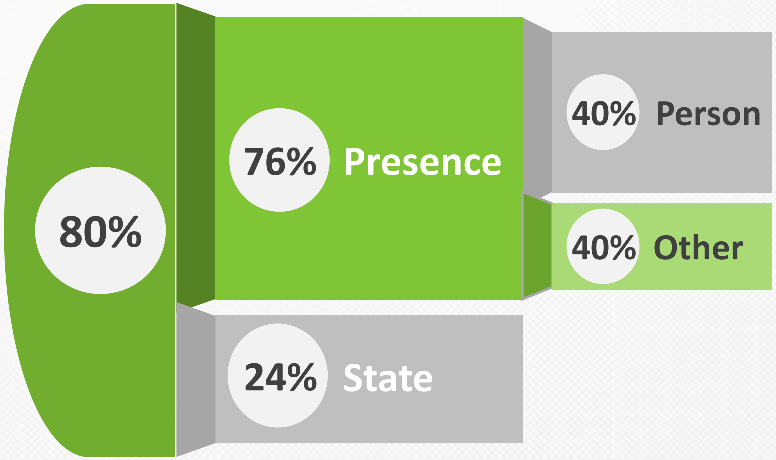


INFOBOT: HRI STUDY

[Chung et al., IROS'15]

- Goal: practical usage, typical questions
- Comprehensive survey (2 buildings, 111 responses)
- Wizard-of-Oz deployment (4 days, 45 unique users)
  - Users use the web interface
  - Operator tele-operates the robot and posts answers
- Questions:

Is there anyone in {location}?
Is {person} in his/her office?
Is there any food in the downstairs kitchen?
Is there anything in my mailbox?
Is the door to the conference room open?
Is the reception still open?
How noisy is it in the atrium right now?

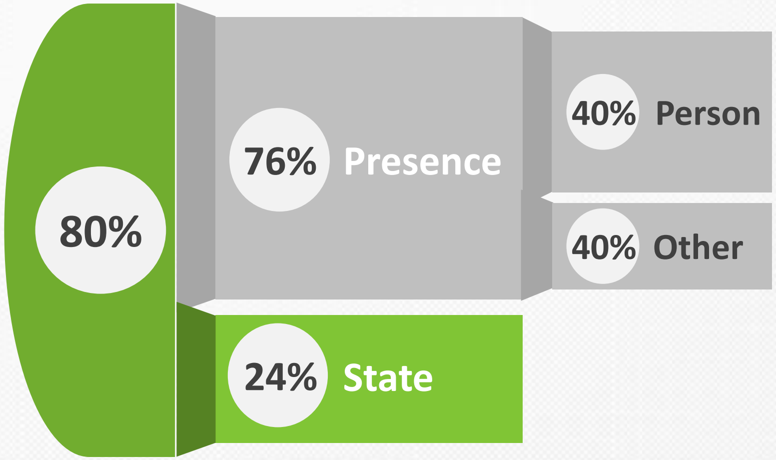


INFOBOT: HRI STUDY

[Chung et al., IROS'15]

- Goal: practical usage, typical questions
- Comprehensive survey (2 buildings, 111 responses)
- Wizard-of-Oz deployment (4 days, 45 unique users)
  - Users use the web interface
  - Operator tele-operates the robot and posts answers
- Questions:

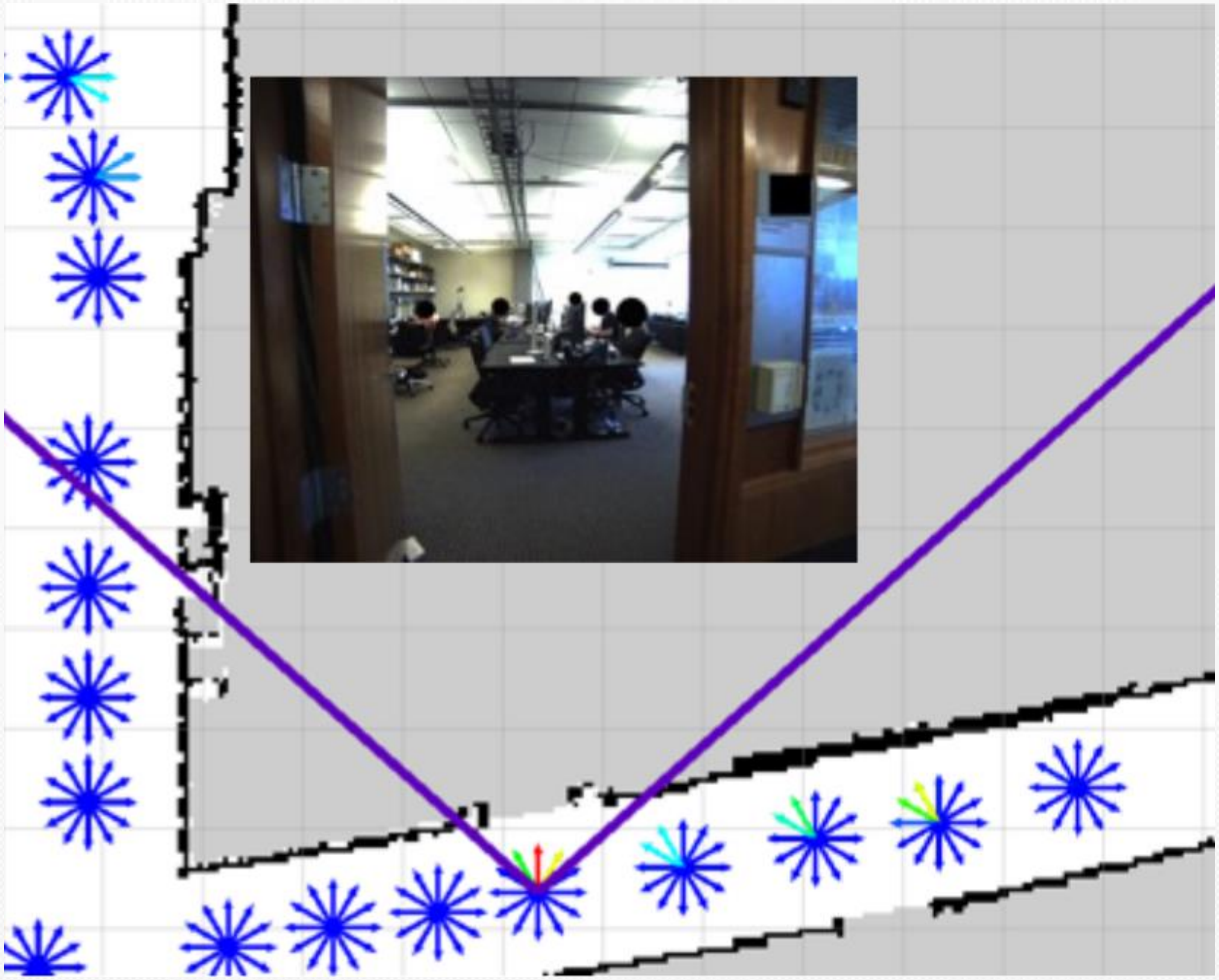
Is there anyone in {location}?
Is {person} in his/her office?
Is there any food in the downstairs kitchen?
Is there anything in my mailbox?
Is the door to the conference room open?
Is the reception still open?
How noisy is it in the atrium right now?



- User satisfaction: only 16% not satisfied with answers

INFOBOT: AUTONOMOUS SYSTEM

Is there anyone in the mobile robotics lab?

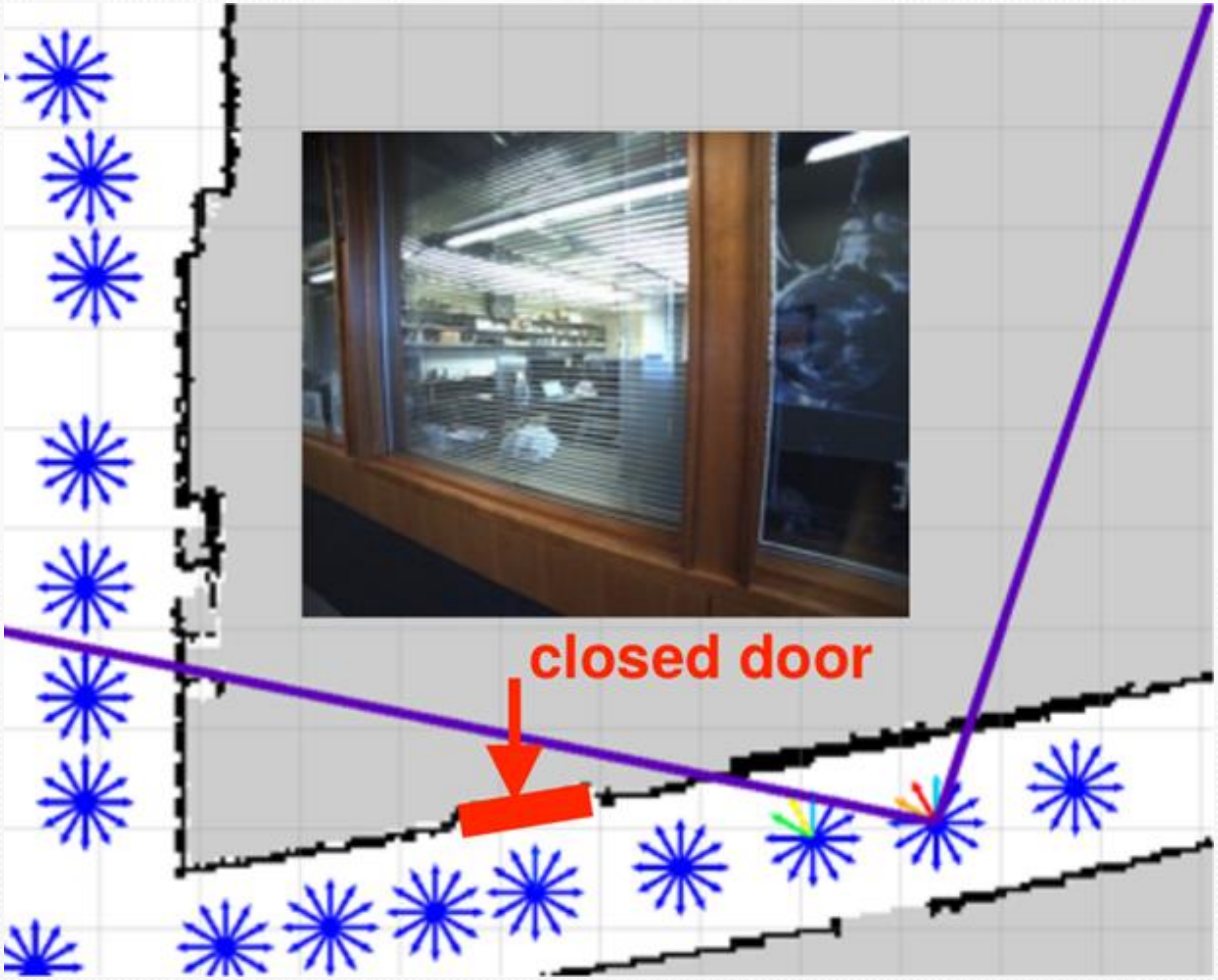


Ground truth: **Yes**

INFOBOT: AUTONOMOUS SYSTEM

[Chung*, Pronobis* et al., IROS'16]

Is there anyone in the mobile robotics lab?



Ground truth: **No**

INFOBOT: AUTONOMOUS SYSTEM

[Chung*, Pronobis* et al., IROS'16]

Is the breakout area occupied?

Ground truth: **No**

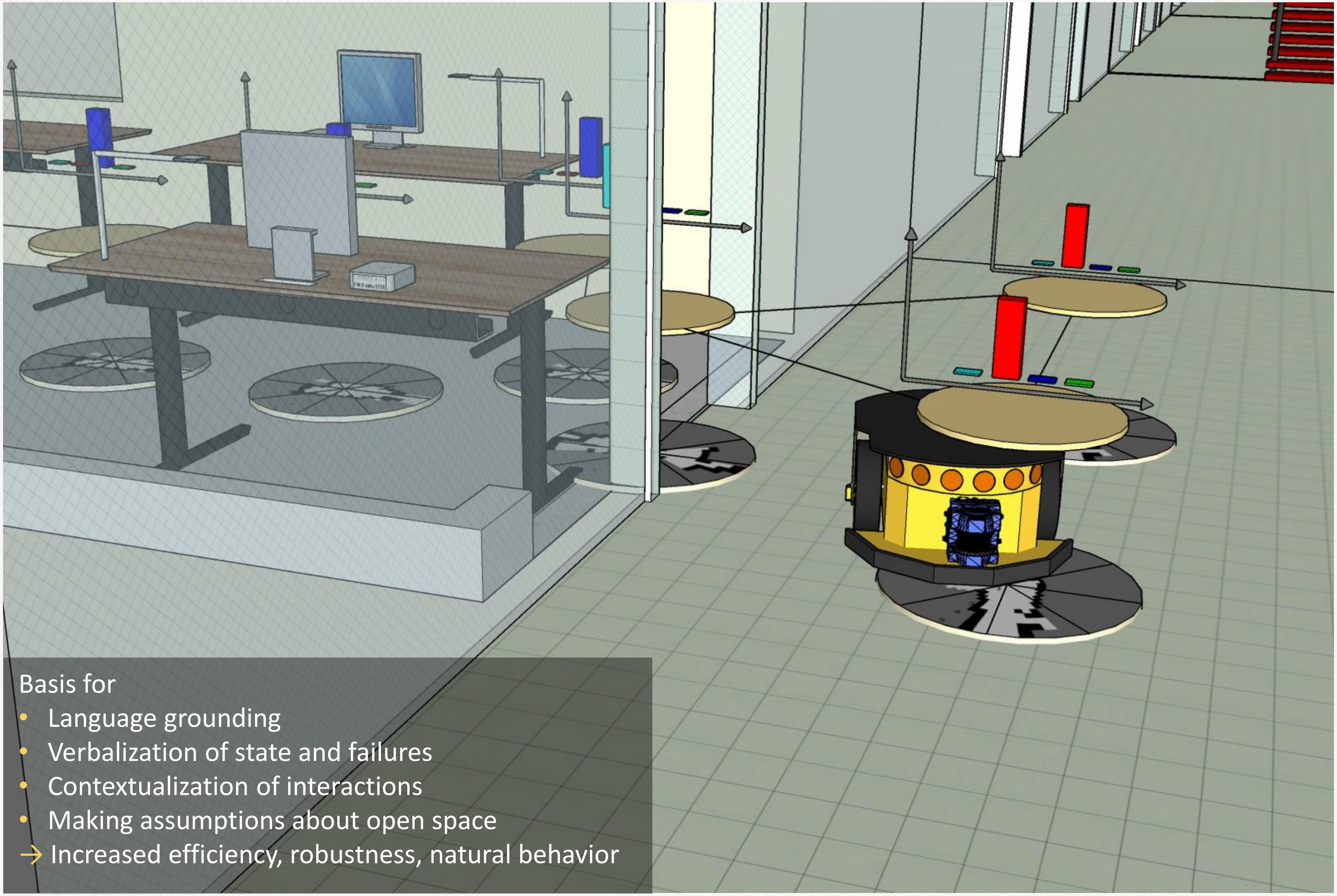
INFOBOT: AUTONOMOUS SYSTEM

[Chung*, Pronobis* et al., IROS'16]

Is the breakout area occupied?

Ground truth: **Yes**

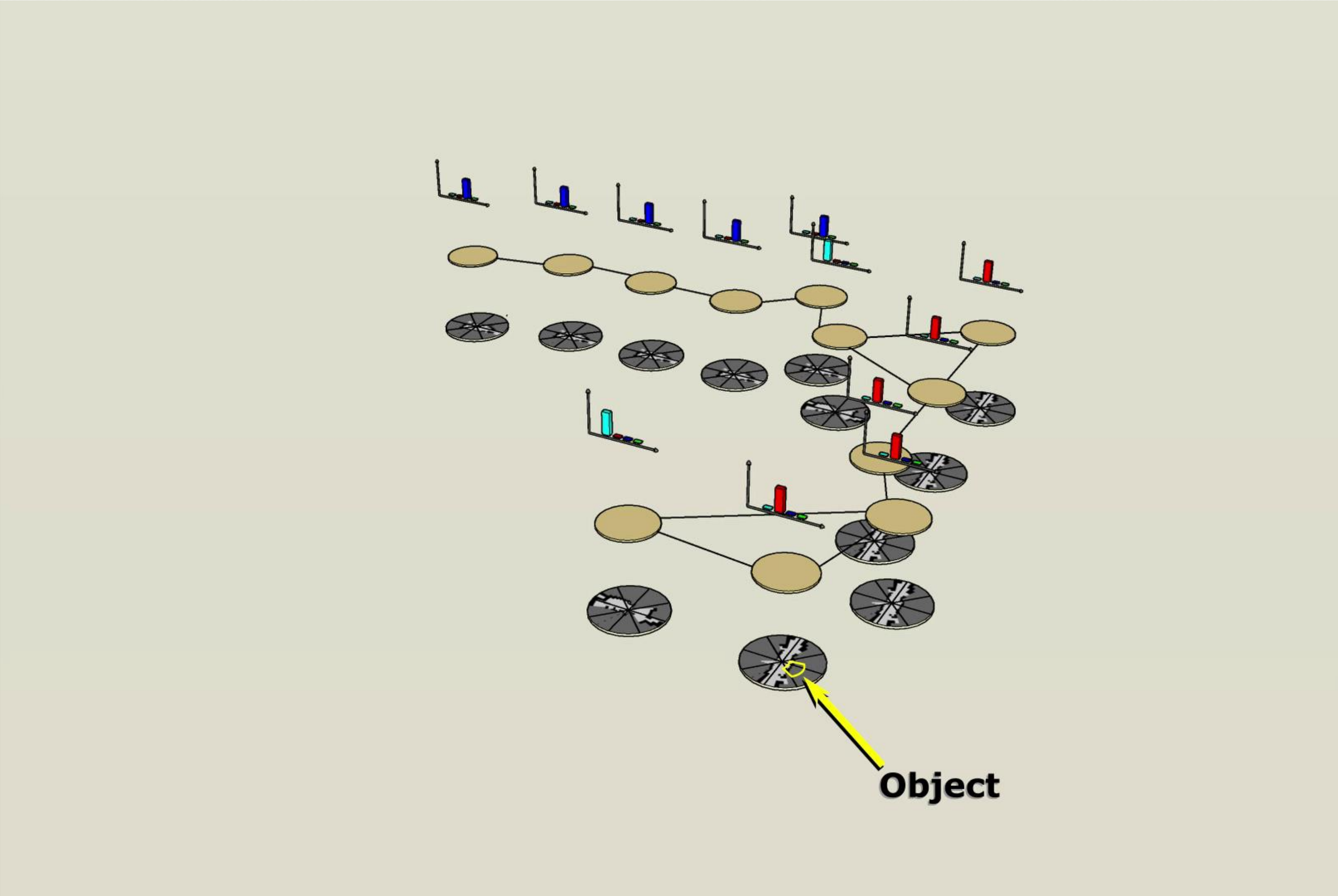
SPATIAL KNOWLEDGE REPRESENTATION



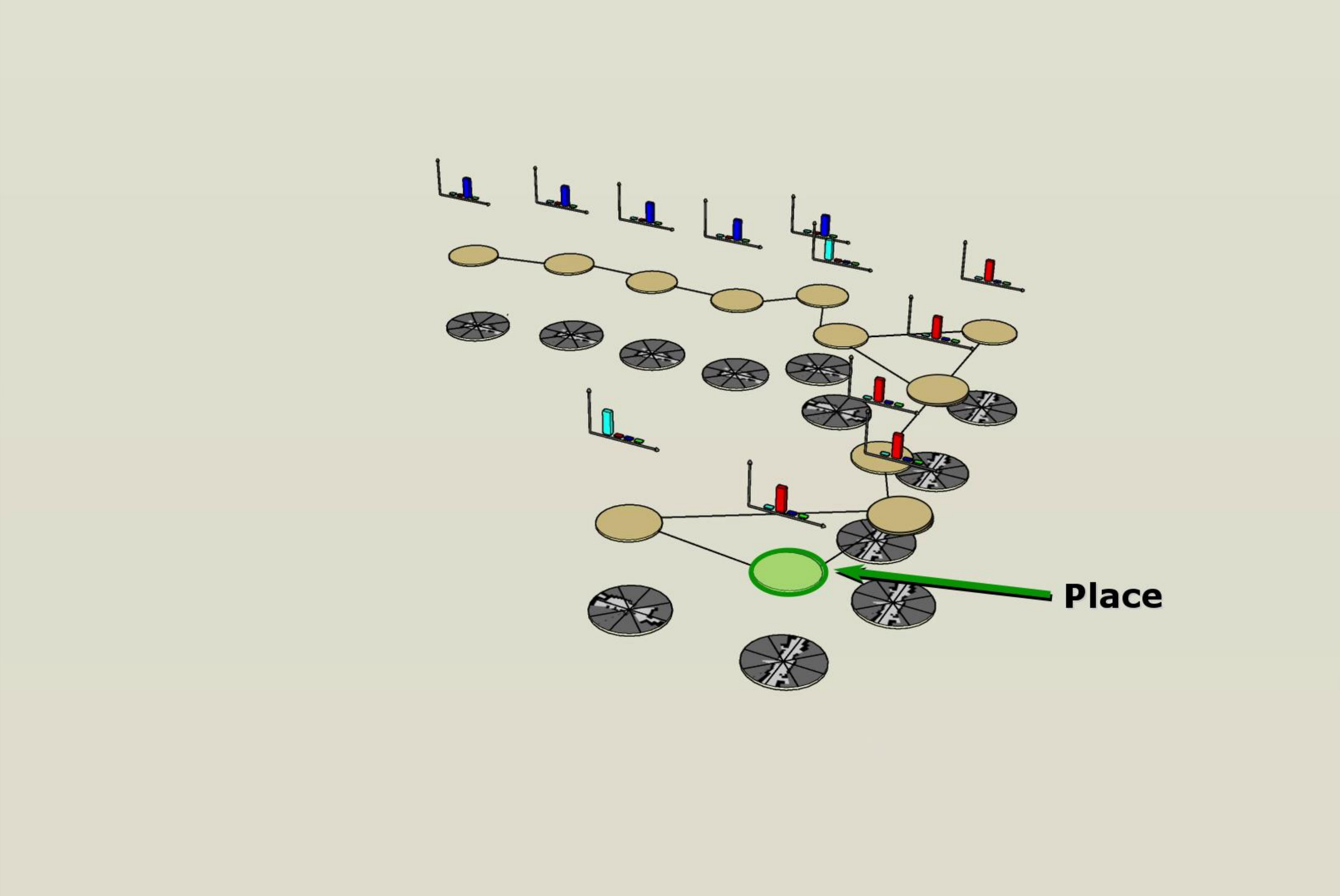
Basis for
- Language grounding
- Verbalization of state and failures
- Contextualization of interactions
- Making assumptions about open space

→ Increased efficiency, robustness, natural behavior

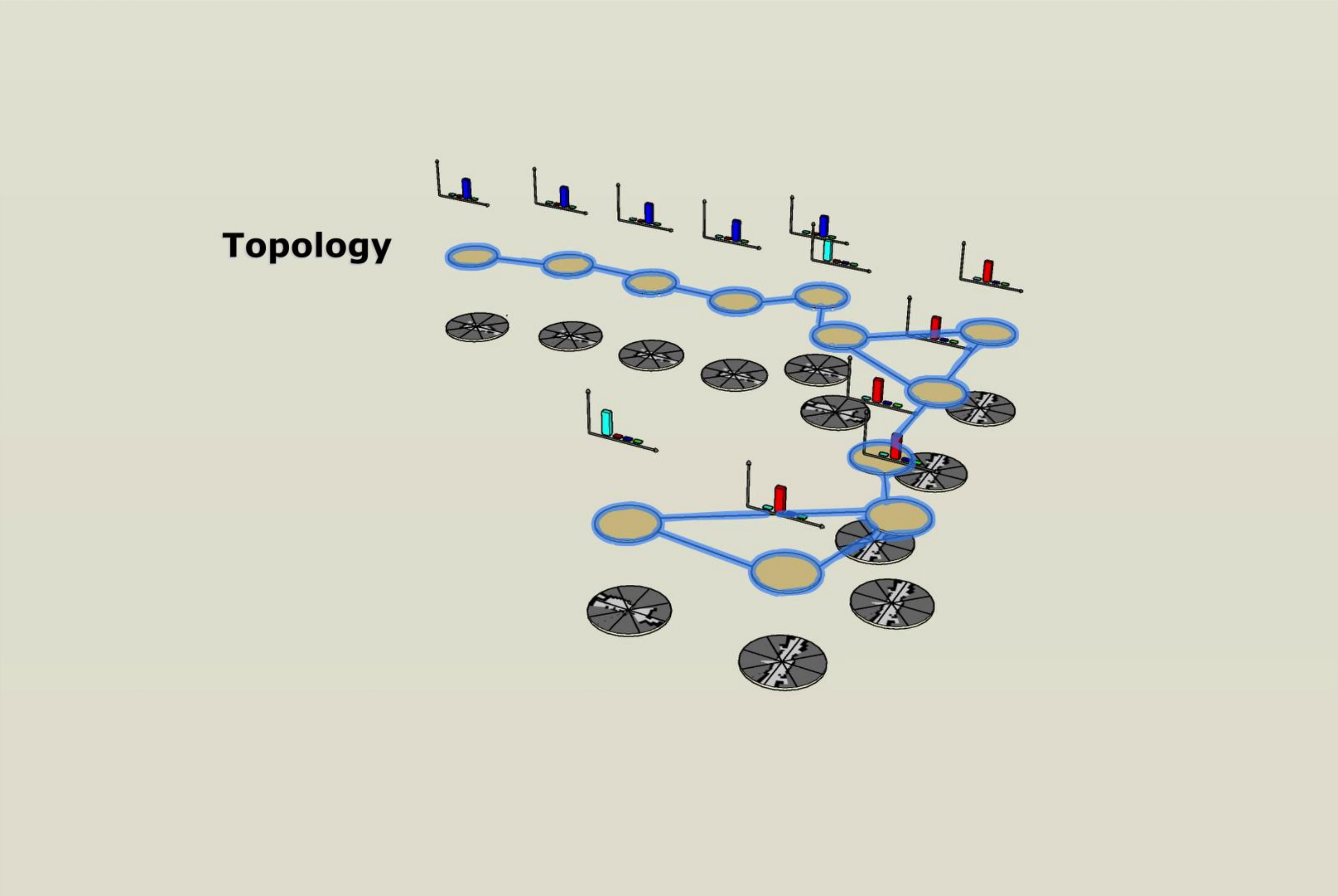
SPATIAL KNOWLEDGE REPRESENTATION



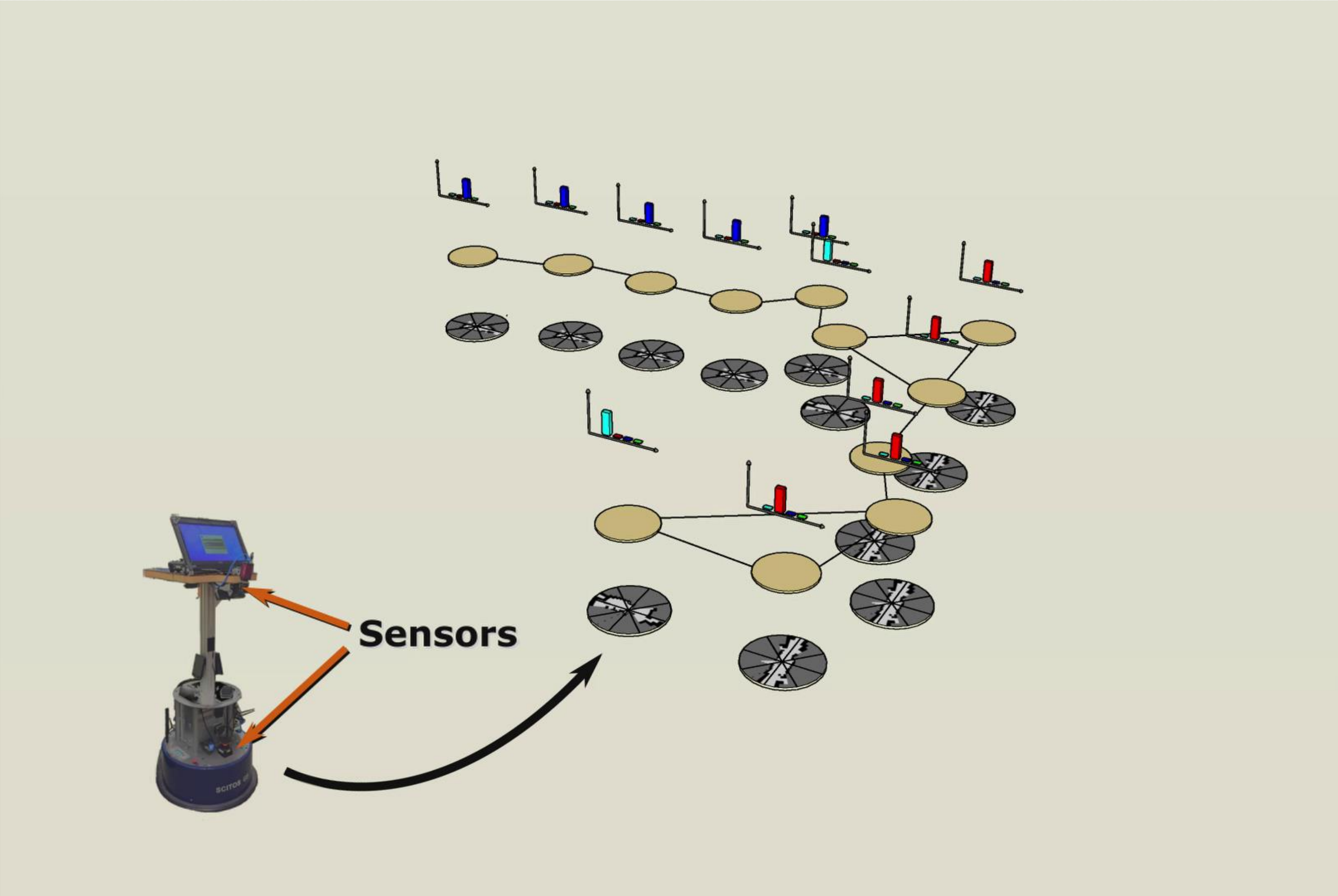
SPATIAL KNOWLEDGE REPRESENTATION



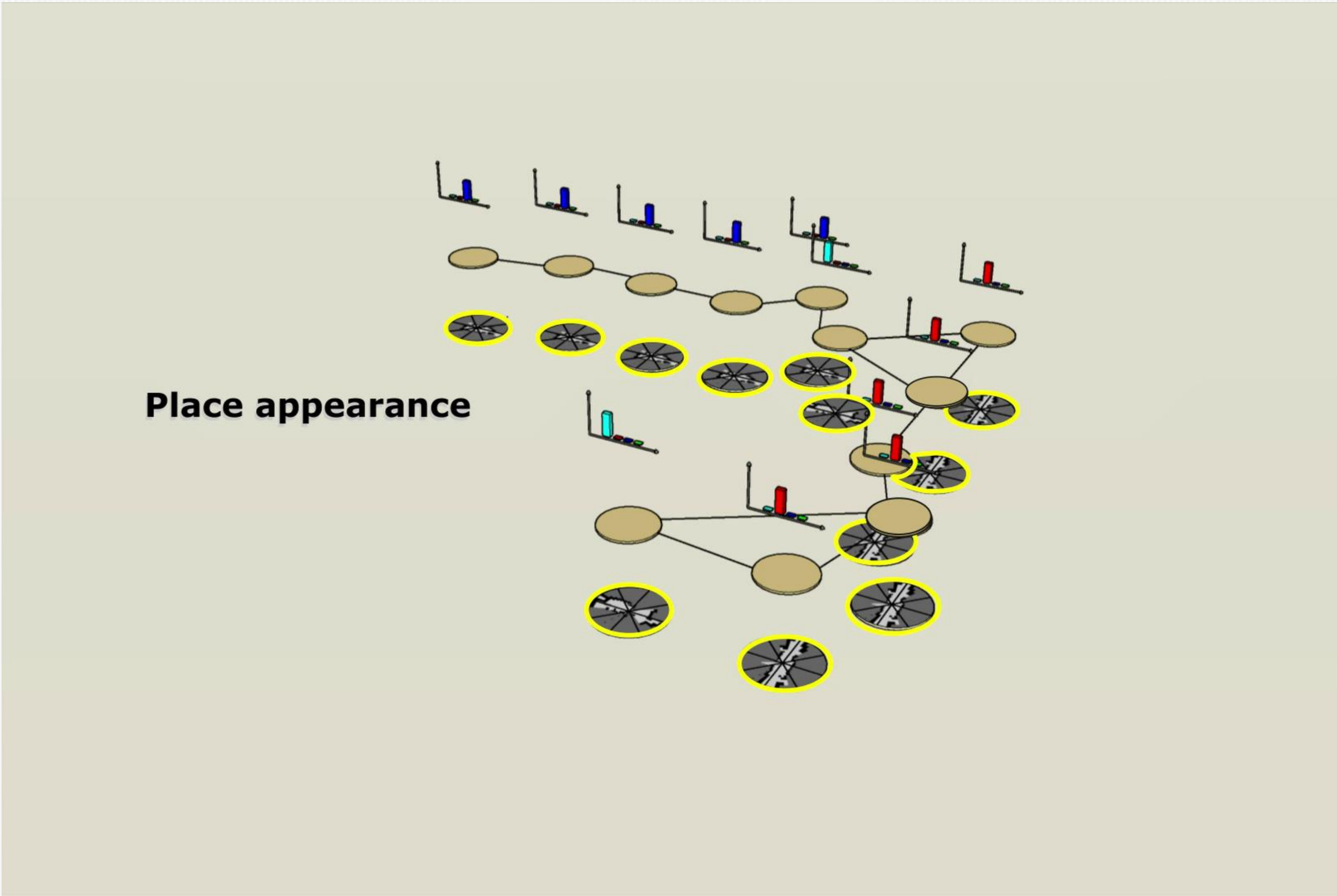
SPATIAL KNOWLEDGE REPRESENTATION



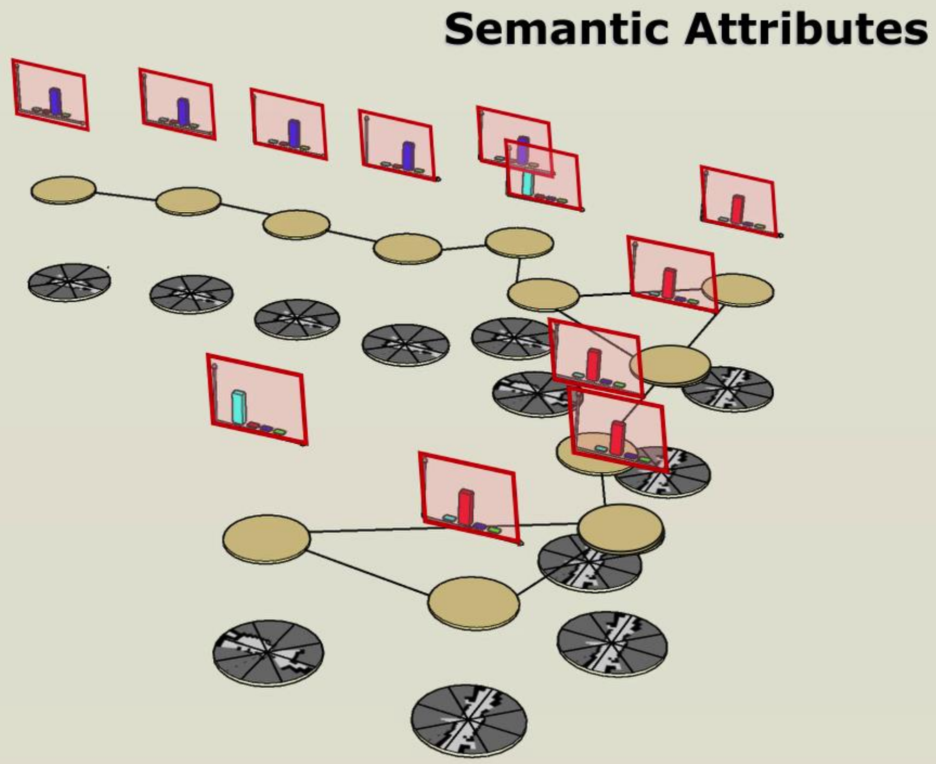
SPATIAL KNOWLEDGE REPRESENTATION



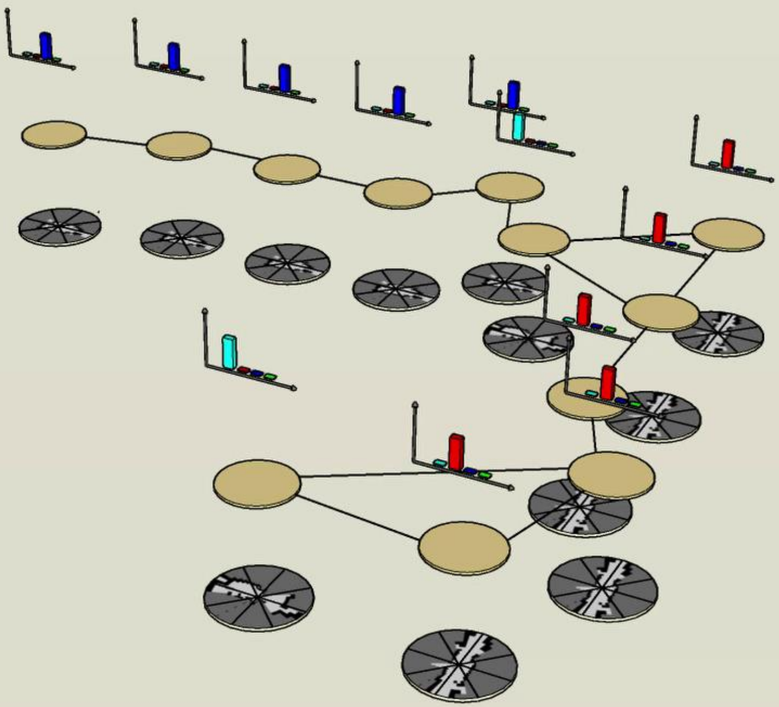
SPATIAL KNOWLEDGE REPRESENTATION



SPATIAL KNOWLEDGE REPRESENTATION

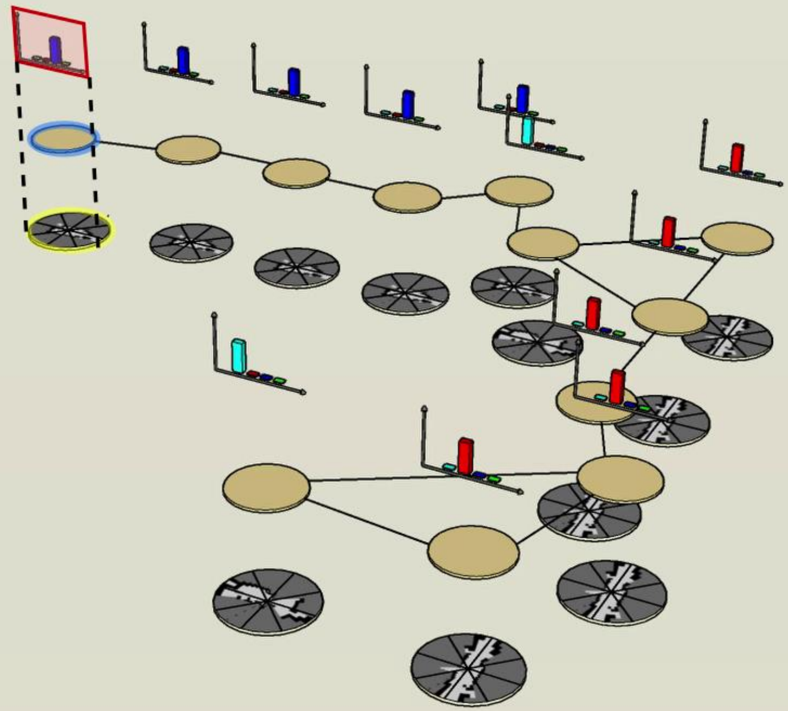


SPATIAL KNOWLEDGE REPRESENTATION

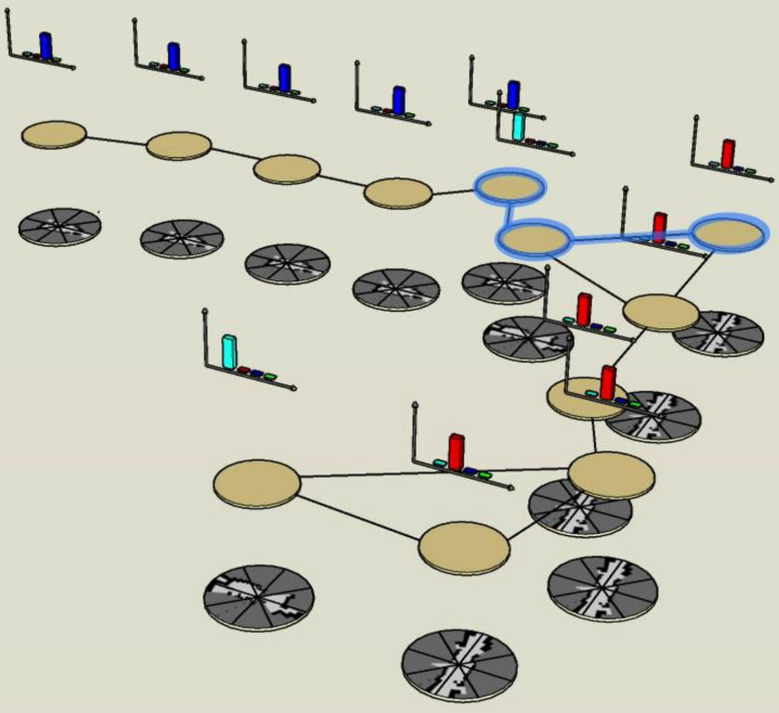


Sensory observations:
local, partial, noisy

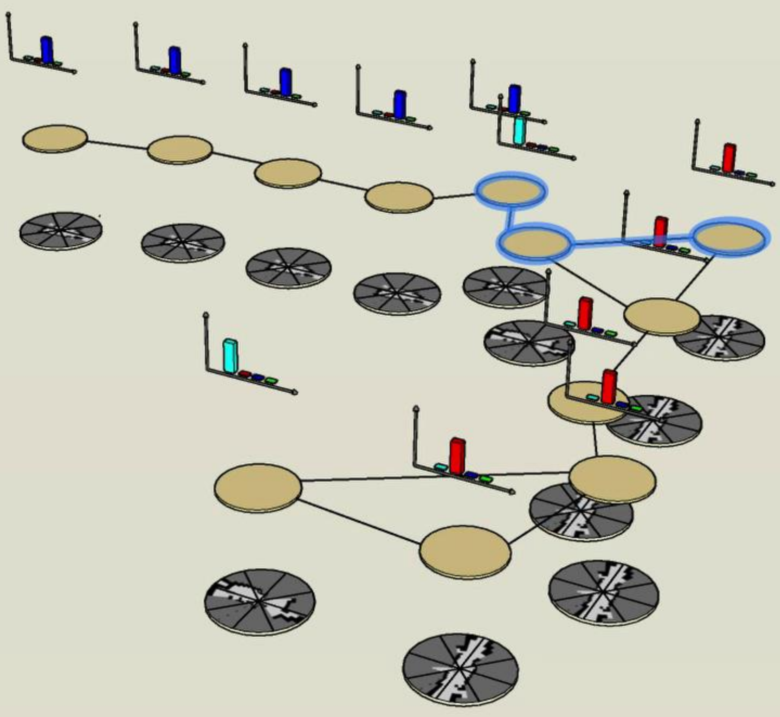
SPATIAL KNOWLEDGE REPRESENTATION



SPATIAL KNOWLEDGE REPRESENTATION



SPATIAL KNOWLEDGE REPRESENTATION

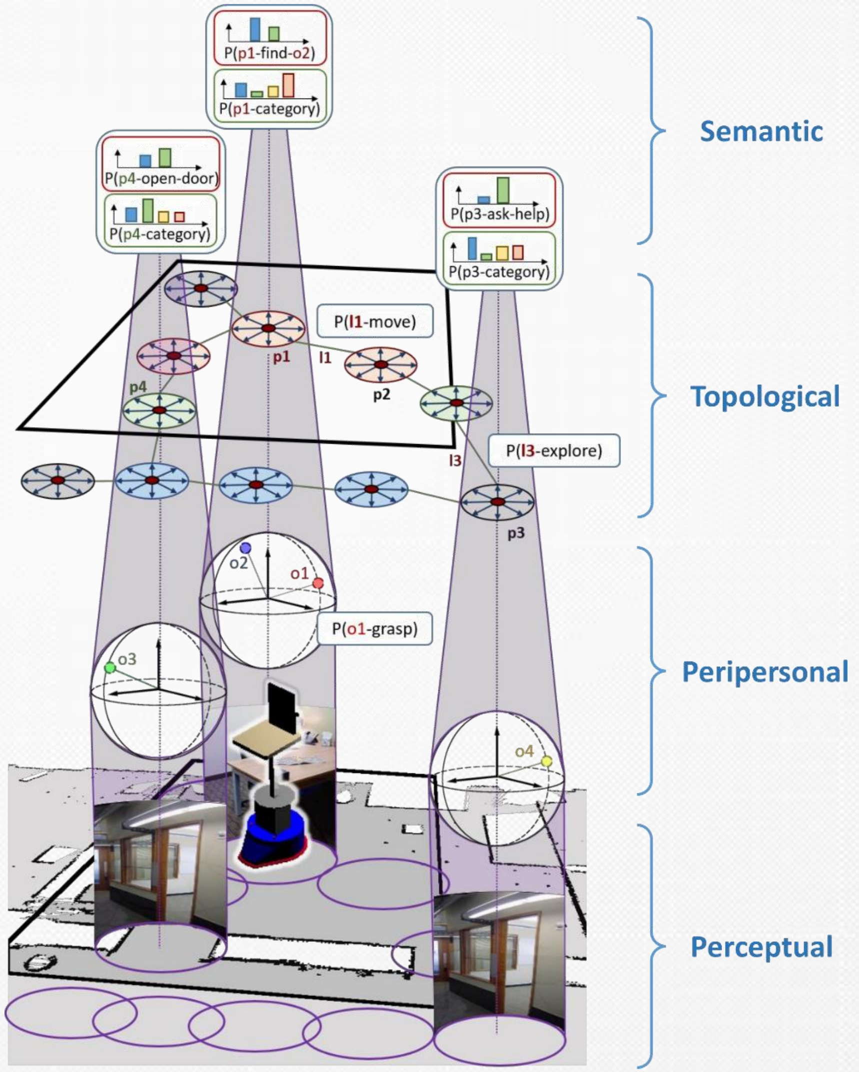


How to structure the body of spatial knowledge for a deliberative agent?

DASH: DEEP SPATIAL AFFORDANCE HIERARCHY

[Pronobis, Riccio, Rao, RSS-SSRR'17]

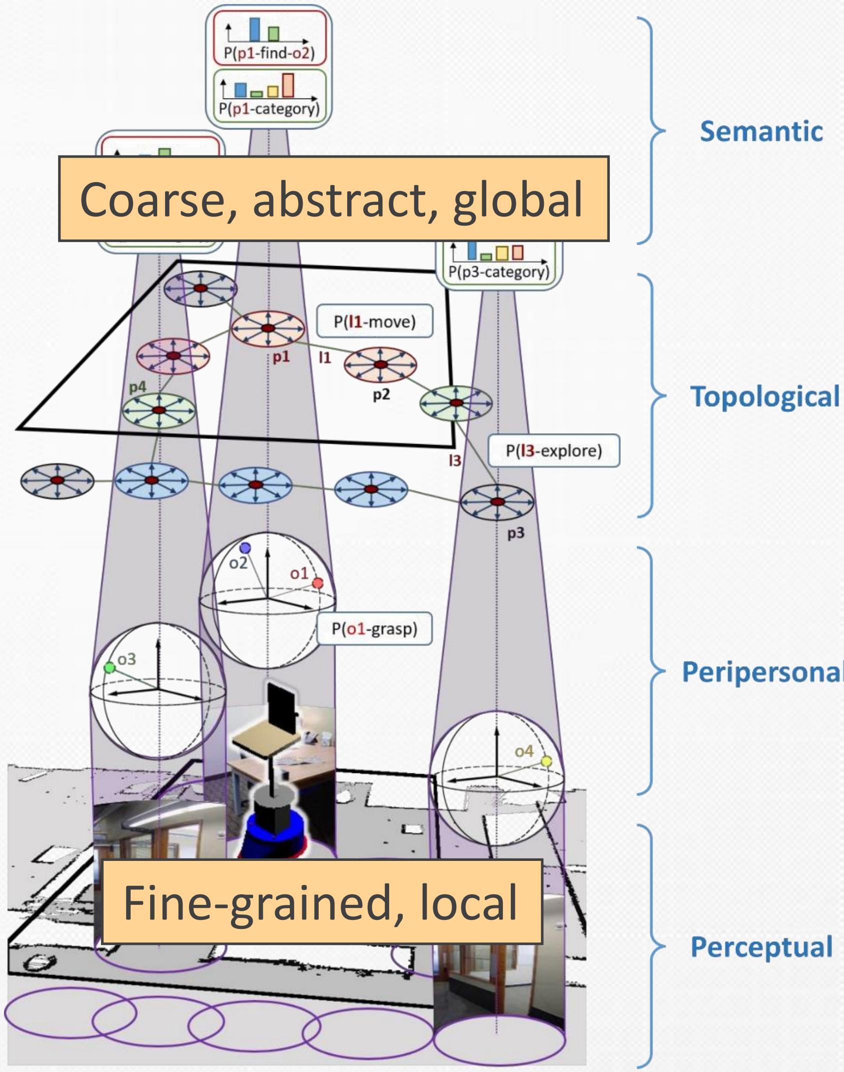
- Hierarchical, layered
- New representation and deep probabilistic model



DASH: DEEP SPATIAL AFFORDANCE HIERARCHY

[Pronobis, Riccio, Rao, RSS-SSRR'17]

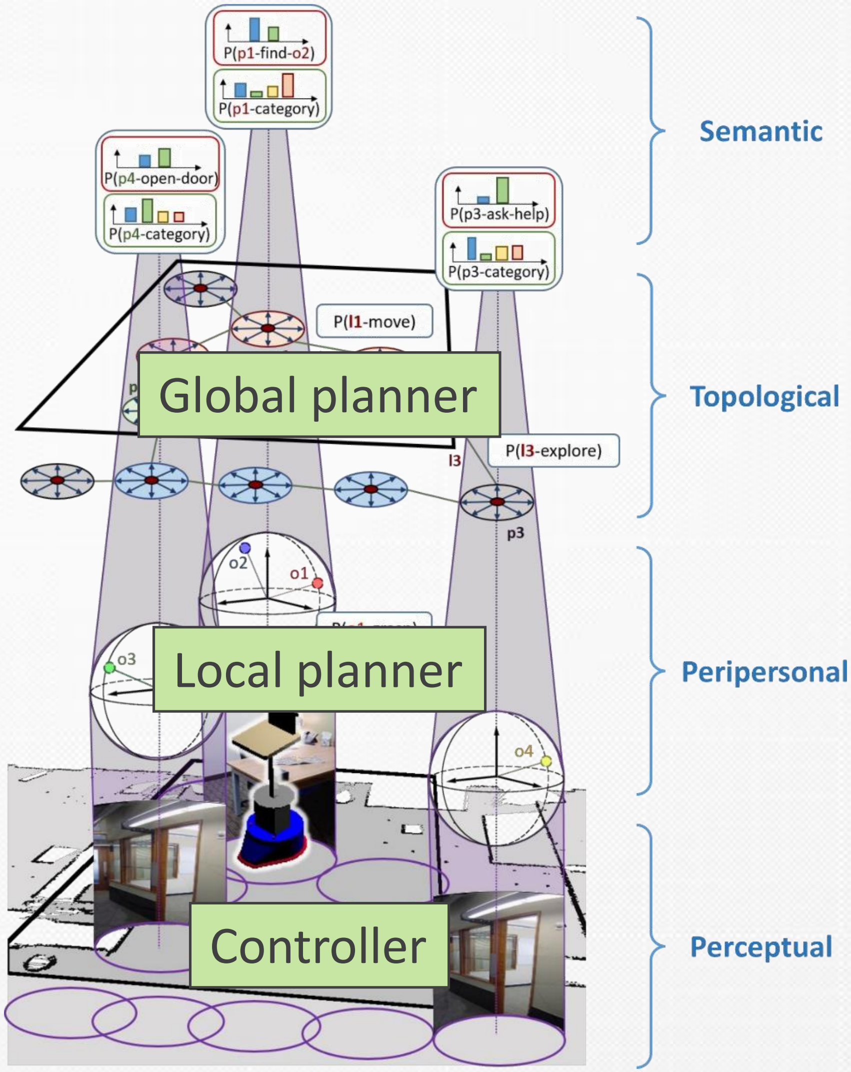
- Hierarchical, layered
- New representation and deep probabilistic model
- 4 layers, different:
  - Aspects of the world
  - Levels of abstraction
  - Spatial scales
  - Frames of reference



DASH: DEEP SPATIAL AFFORDANCE HIERARCHY

[Pronobis, Riccio, Rao, RSS-SSRR'17]

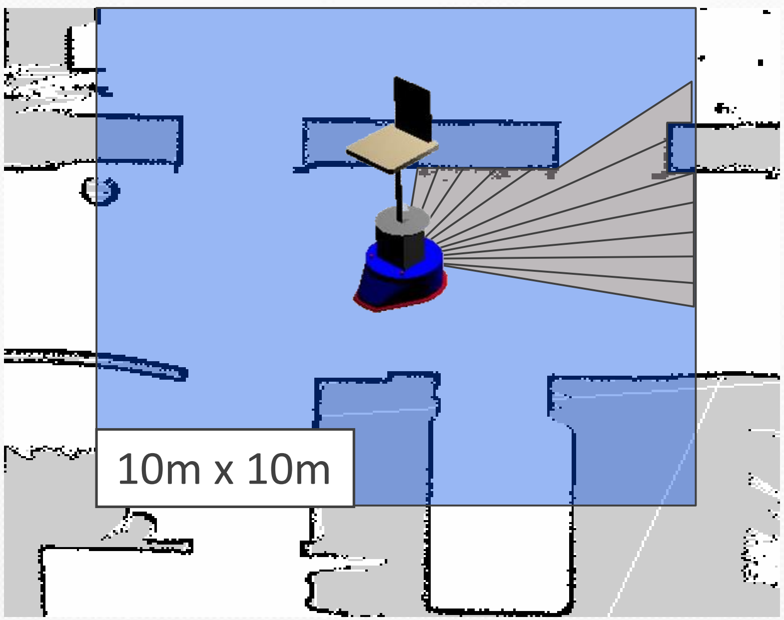
- Hierarchical, layered
- New representation and deep probabilistic model
- 4 layers, different:
  - Aspects of the world
  - Levels of abstraction
  - Spatial scales
  - Frames of reference

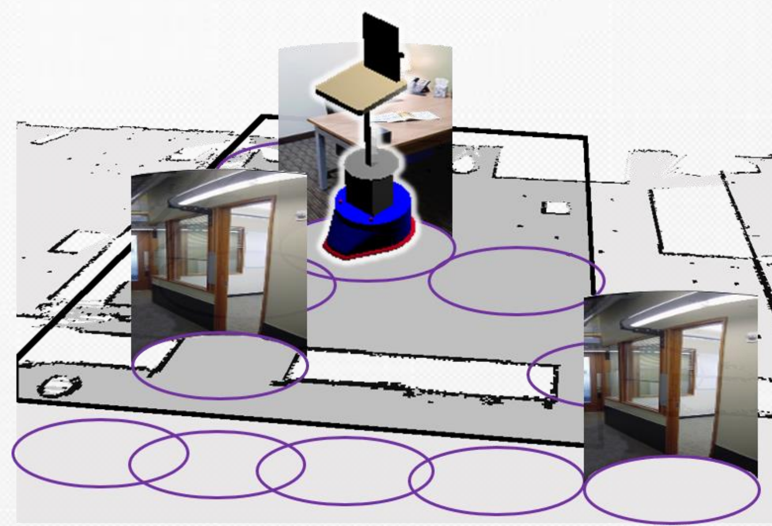


DASH: PERCEPTUAL LAYER

[Pronobis, Riccio, Rao, RSS-SSRR'17]

- Spatio-temporal integration of sensory data
- Accurate geometry and appearance
- Realized as sliding window following the robot

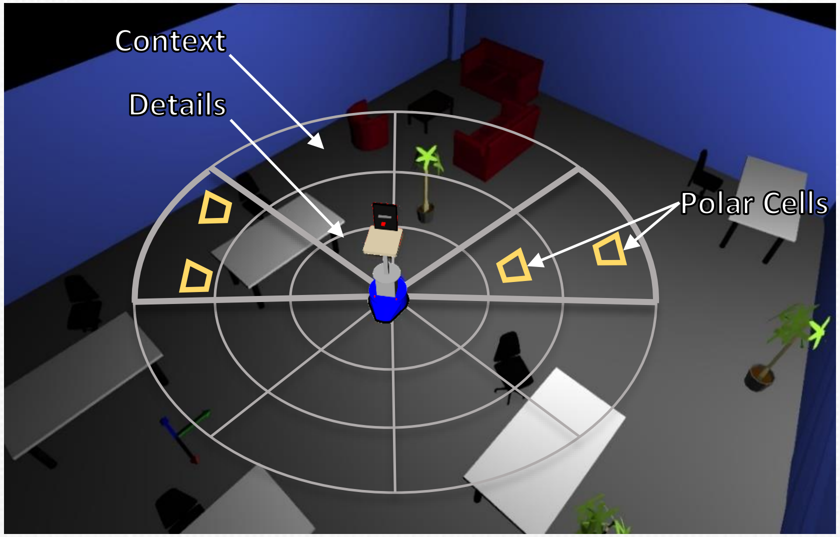


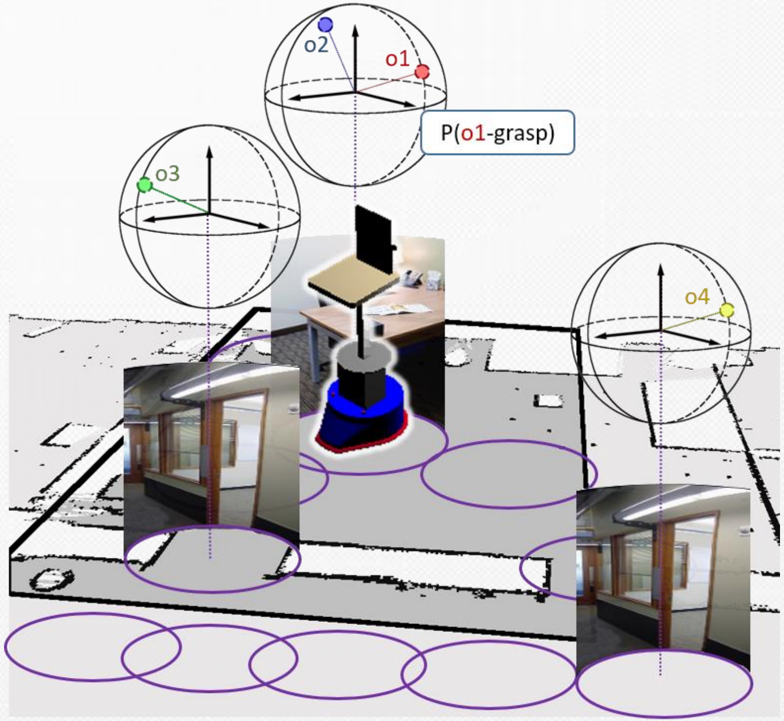


DASH: PERIPERSONAL LAYER

[Pronobis, Riccio, Rao, RSS-SSRR'17]

- Collection of ego-centric representations, each:
  - Models space immediately reachable or observable
  - From perspective of robot at a specific place
  - Updated when robot visits a place
- Realized using collection of polar occupancy grids



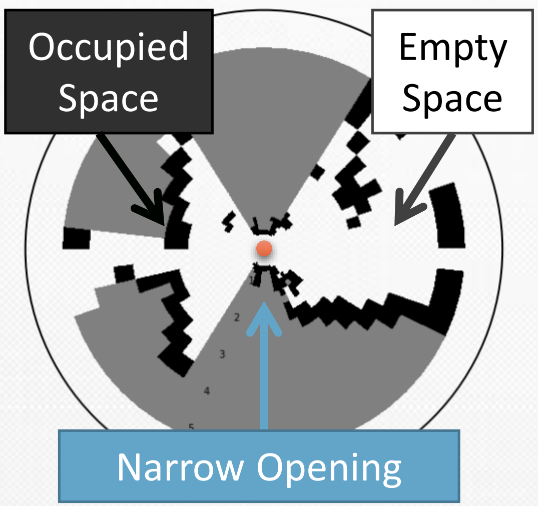


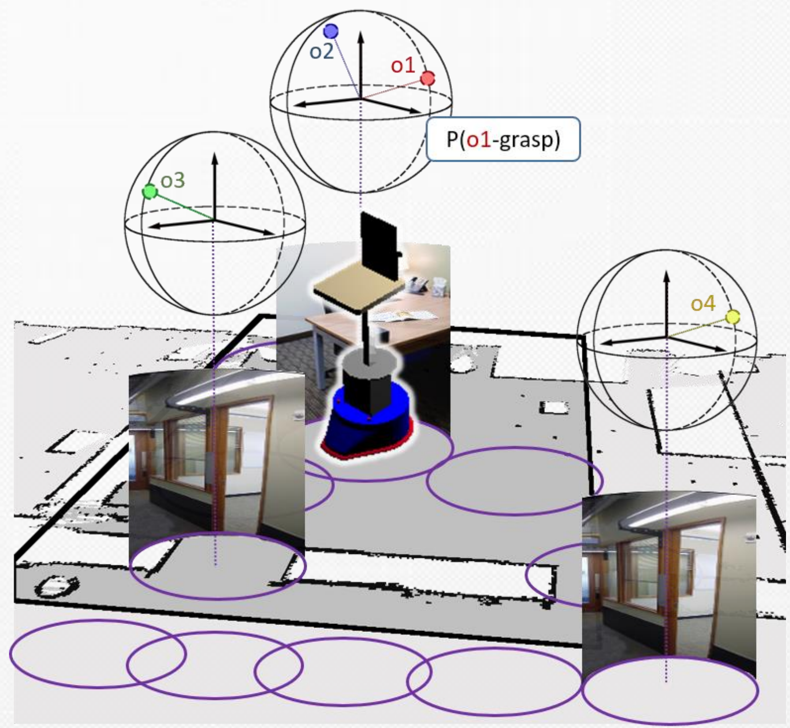
DASH: PERIPERSONAL LAYER

[Pronobis, Riccio, Rao, RSS-SSRR'17]

- Collection of ego-centric representations, each:
  - Models space immediately reachable or observable
  - From perspective of robot at a specific place
  - Updated when robot visits a place
- Realized using collection of polar occupancy grids

For a doorway:

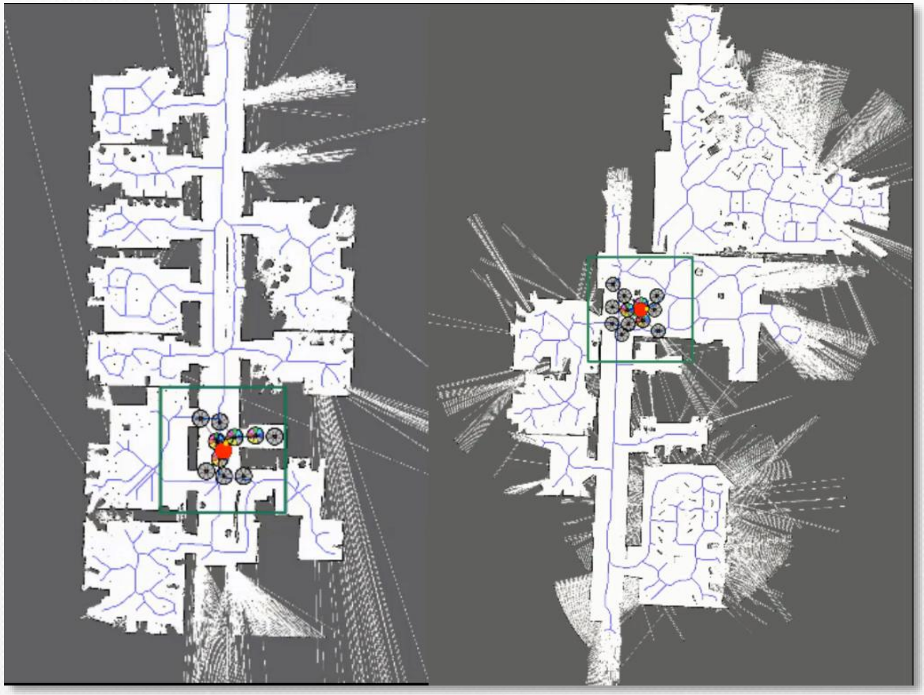


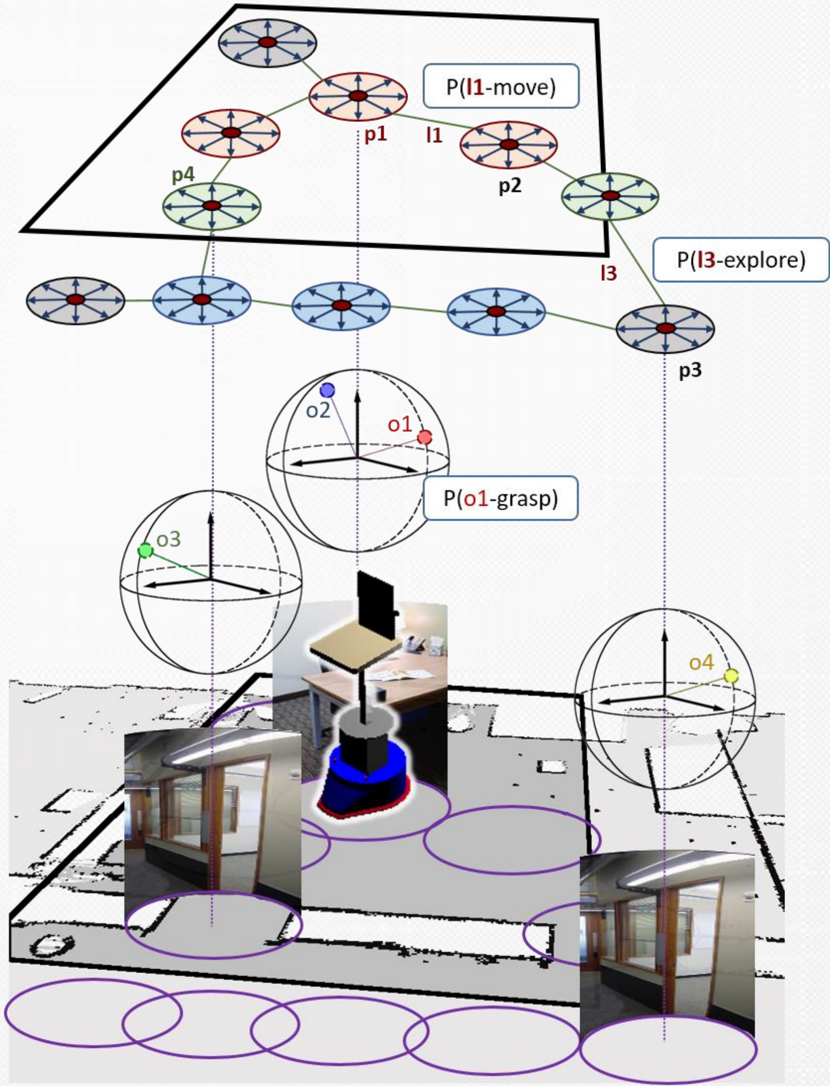


DASH: TOPOLOGICAL LAYER

[Pronobis, Riccio, Rao, RSS-SSRR'17]

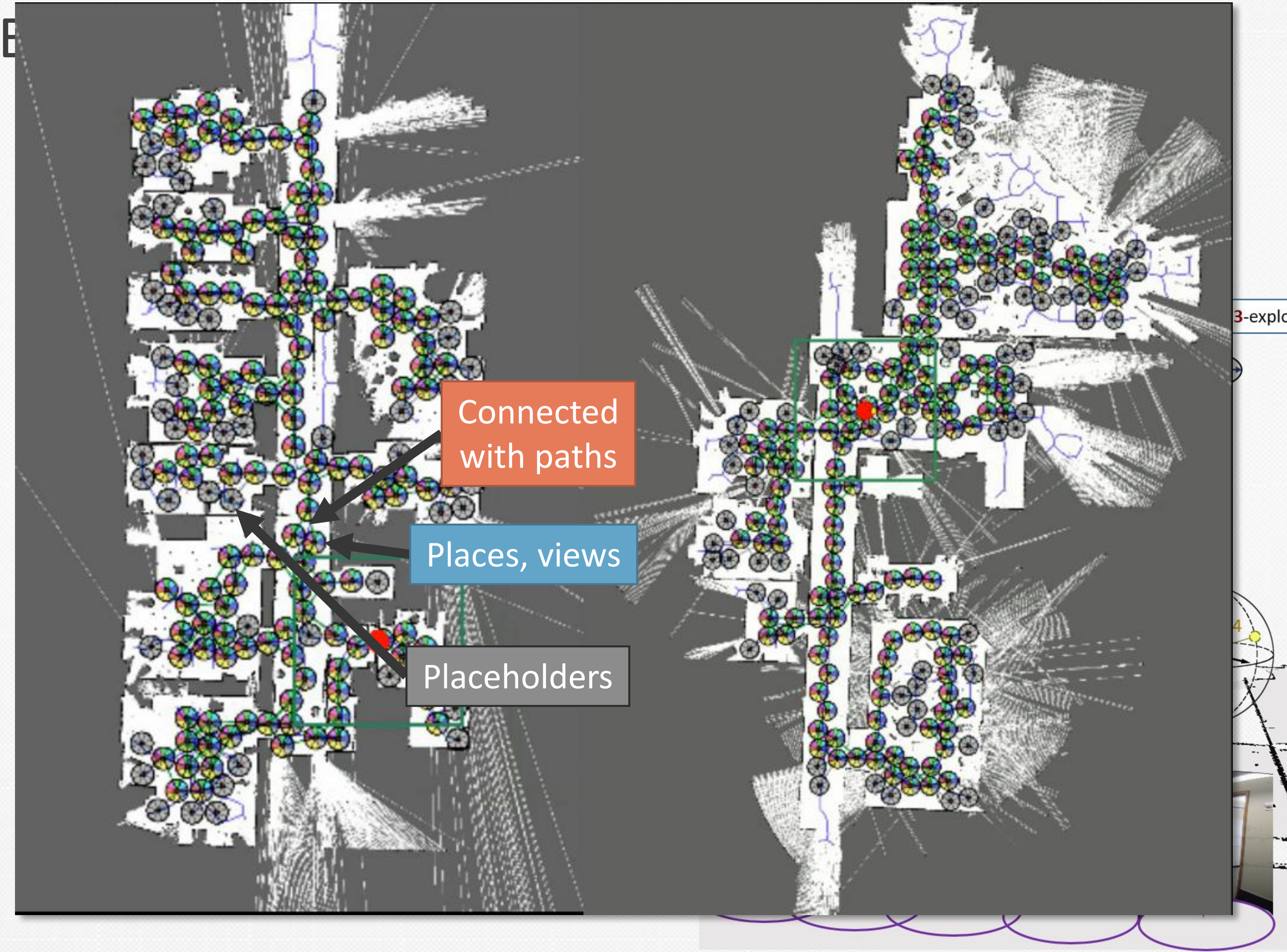
- Efficient representation of large-scale space
  - Coarse global geometry
  - Topology





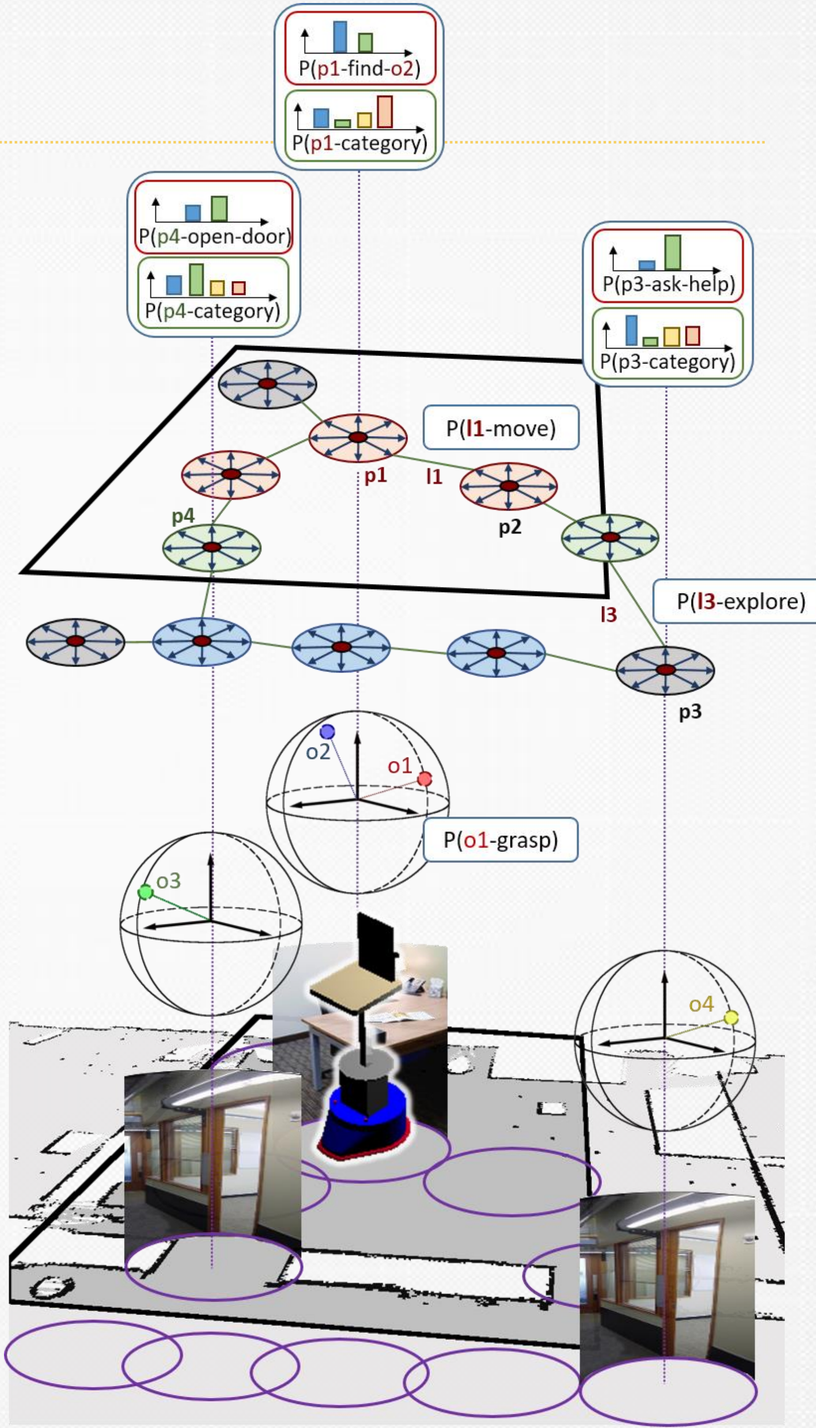
DASH: TOPOLOGICAL LAYER

[Pronobis, Riccio, Rao, RSS-SSRR'17]



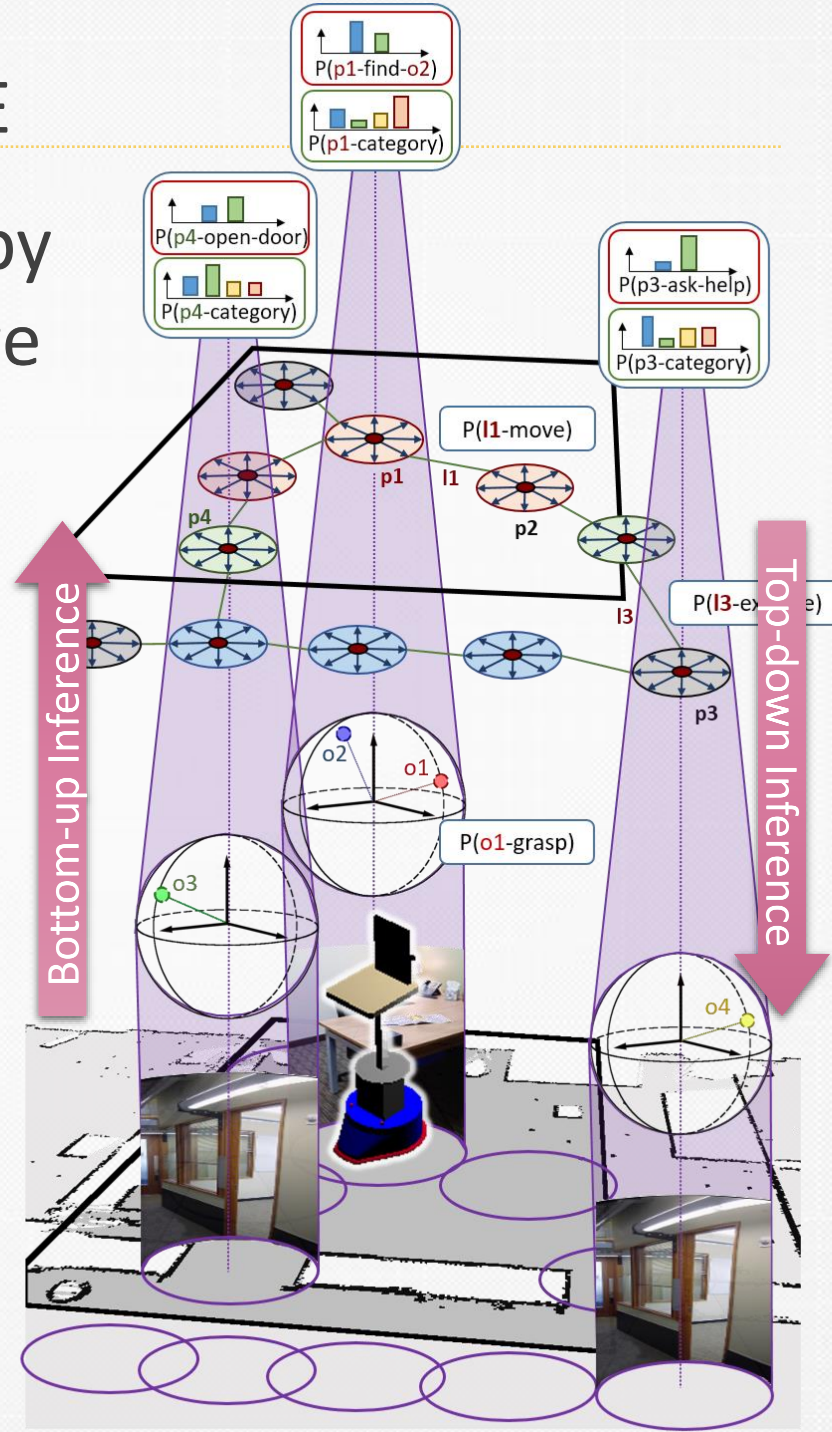
DASH: SEMANTIC LAYER

- Simple probabilistic relational representation
- Relates entities to semantic concepts
  - "place1 is-a kitchen"



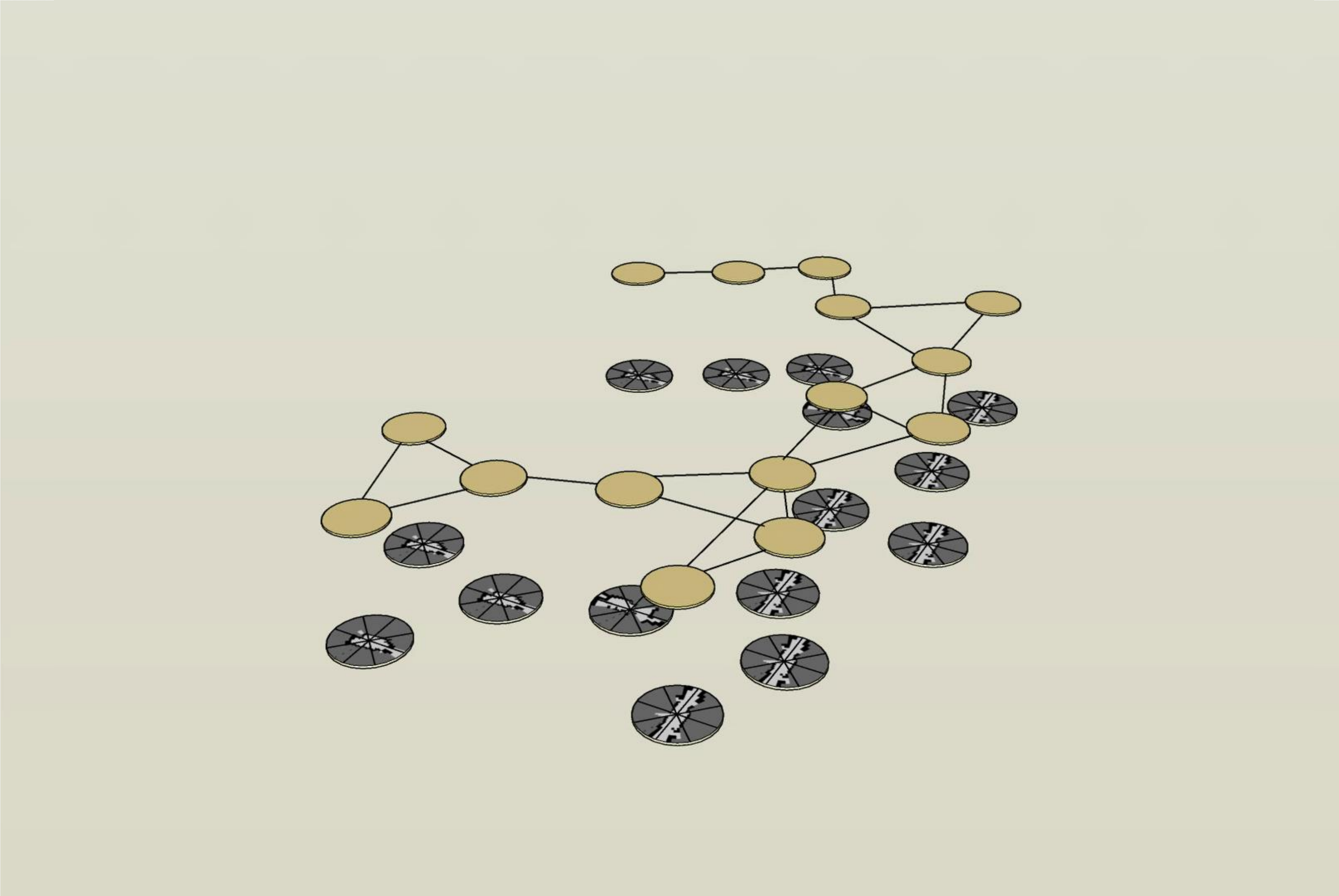
DASH: GENERAL KNOWLEDGE

- **Instance** layers connected by model of **general** knowledge
- Enables top-down and bottom-up inferences
  - Filling missing data **(what's behind robot?)**
  - Inferring latent info **(what room is this?)**
  - Resolving ambiguities

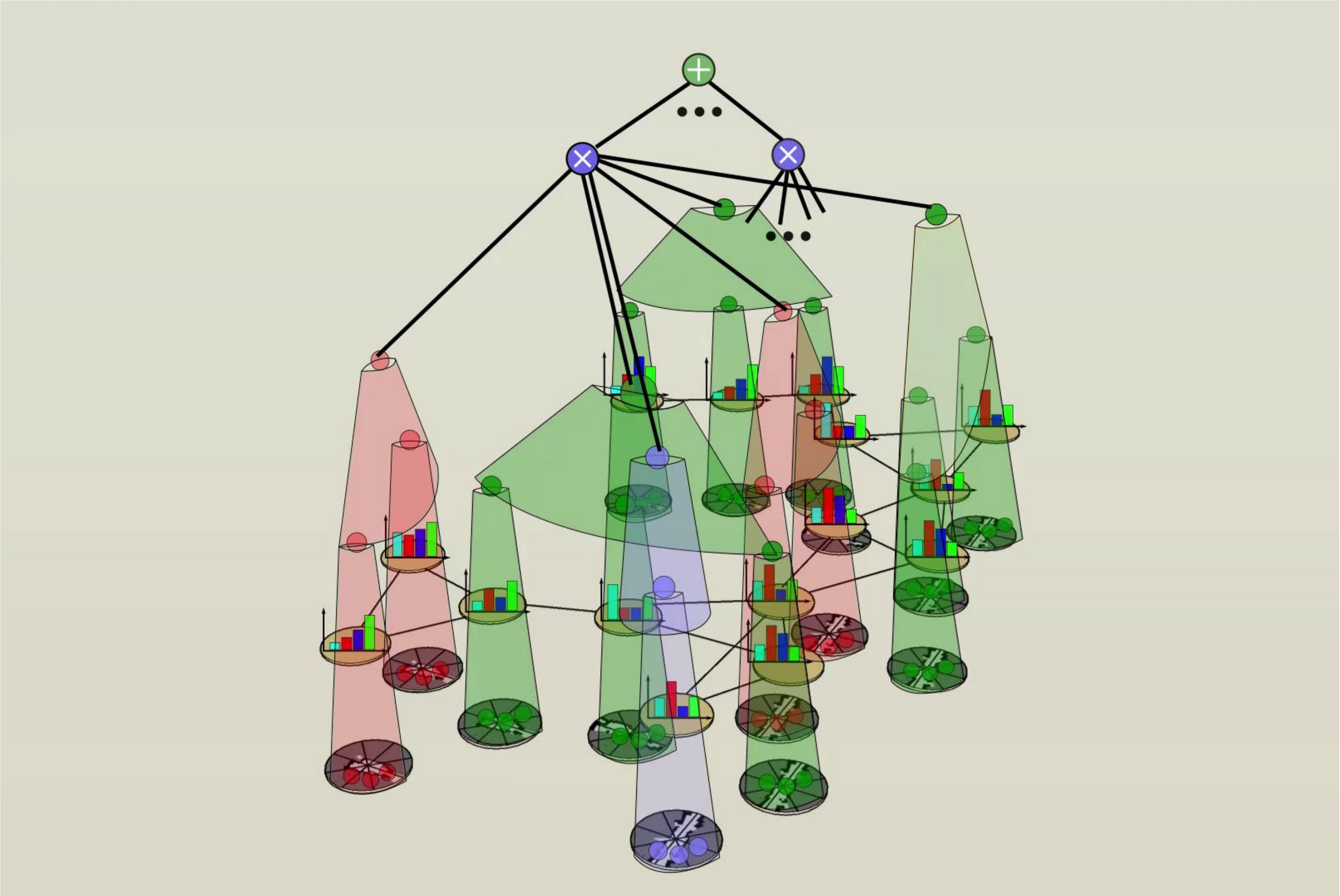


GENERAL KNOWLEDGE: END2END DEEP APPROACH

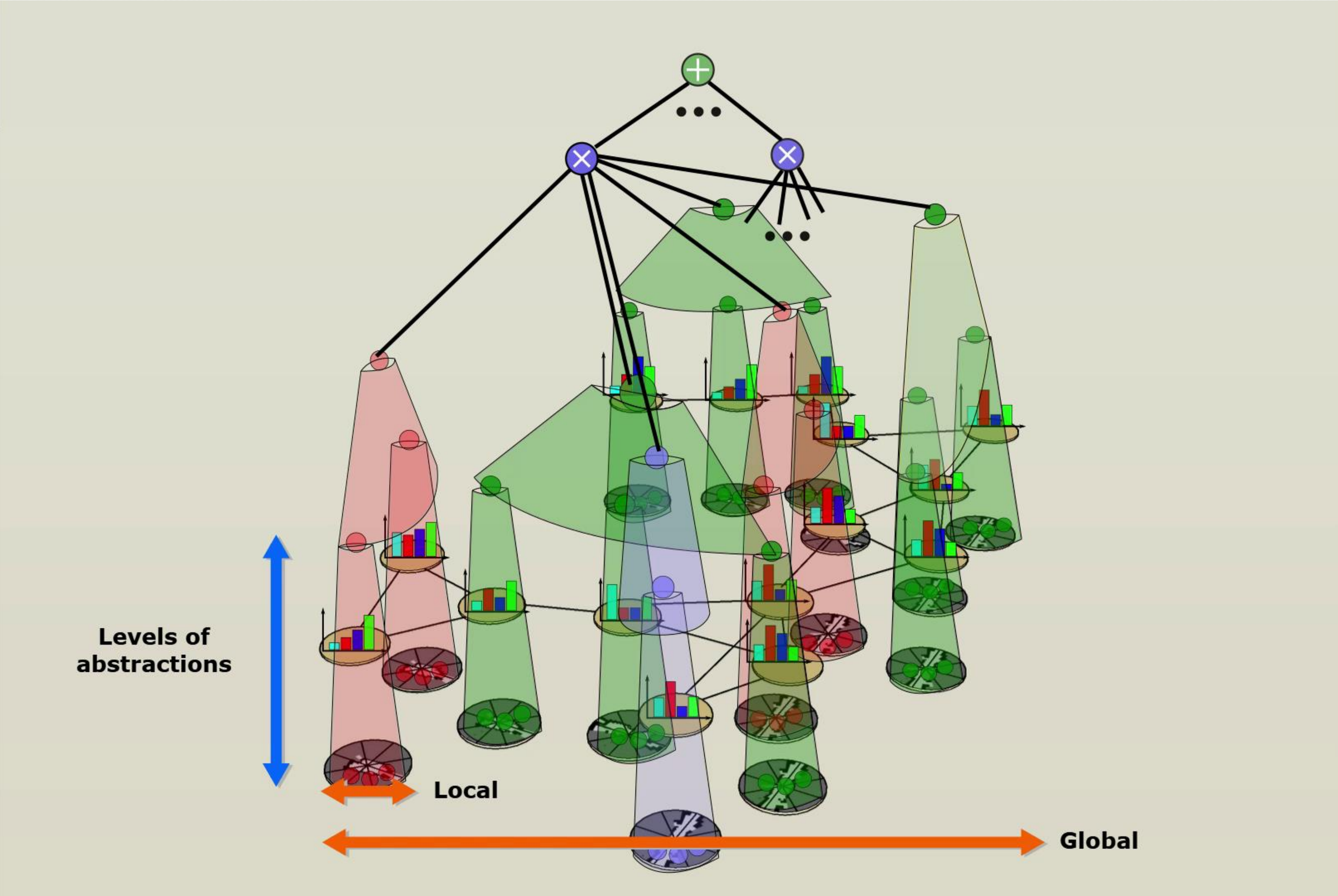
GENERAL KNOWLEDGE: END2END DEEP APPROACH



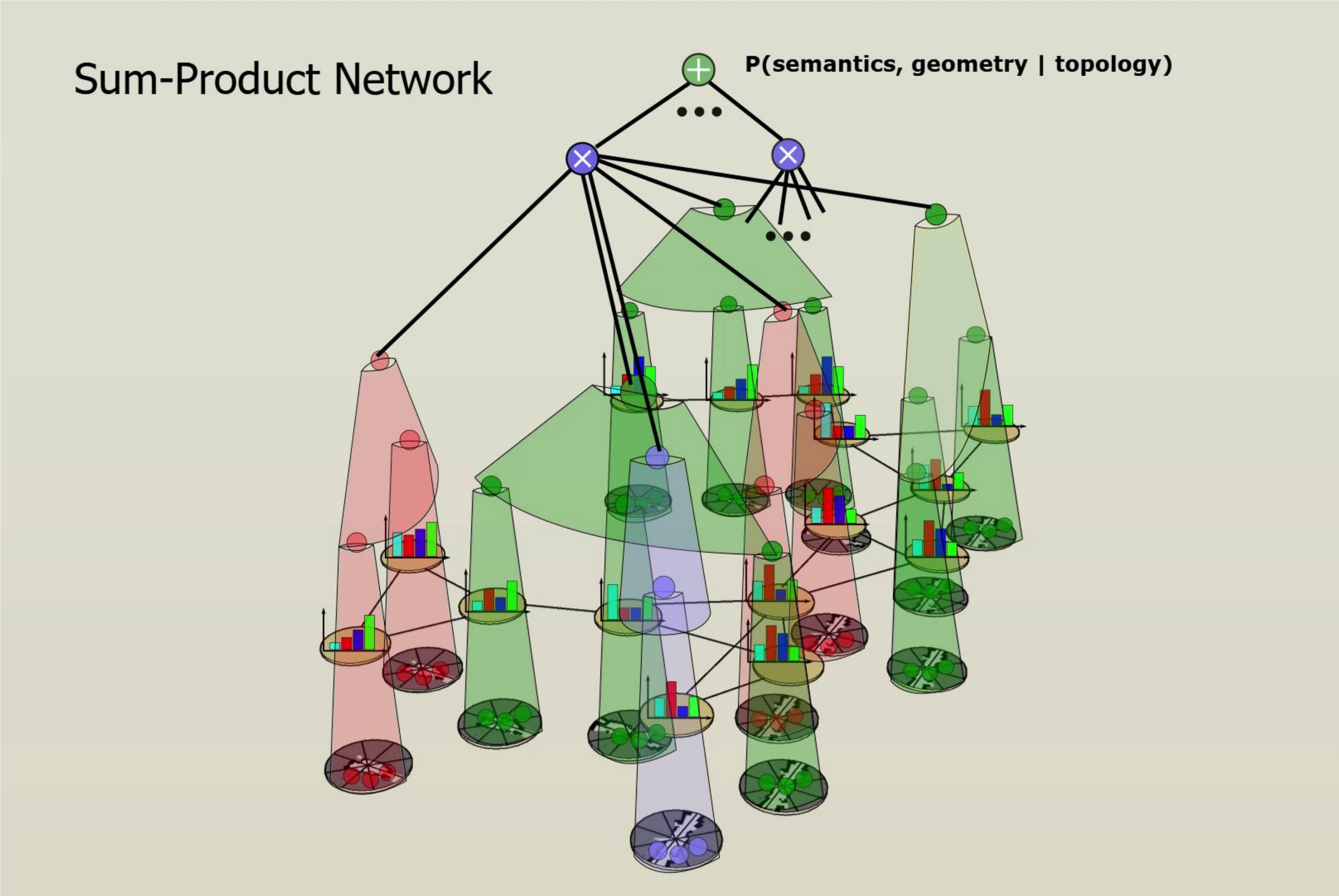
GENERAL KNOWLEDGE: END2END DEEP APPROACH



GENERAL KNOWLEDGE: END2END DEEP APPROACH



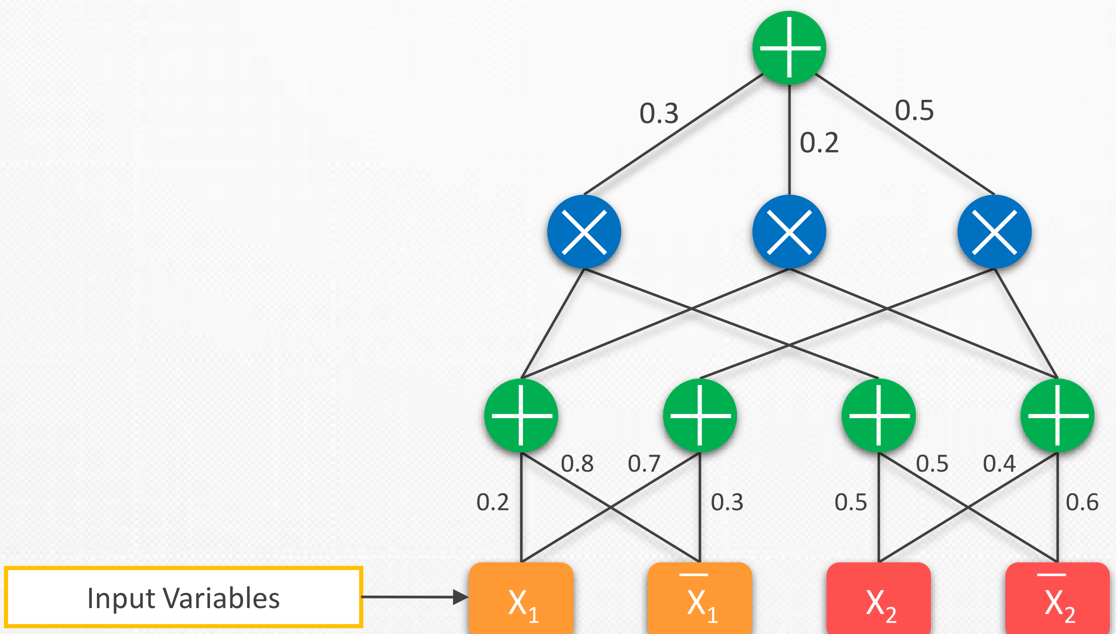
GENERAL KNOWLEDGE: END2END DEEP APPROACH



SUM-PRODUCT NETWORKS

[Poon & Domingos, UAI'11, Friesen & Domingos, ICML'16]

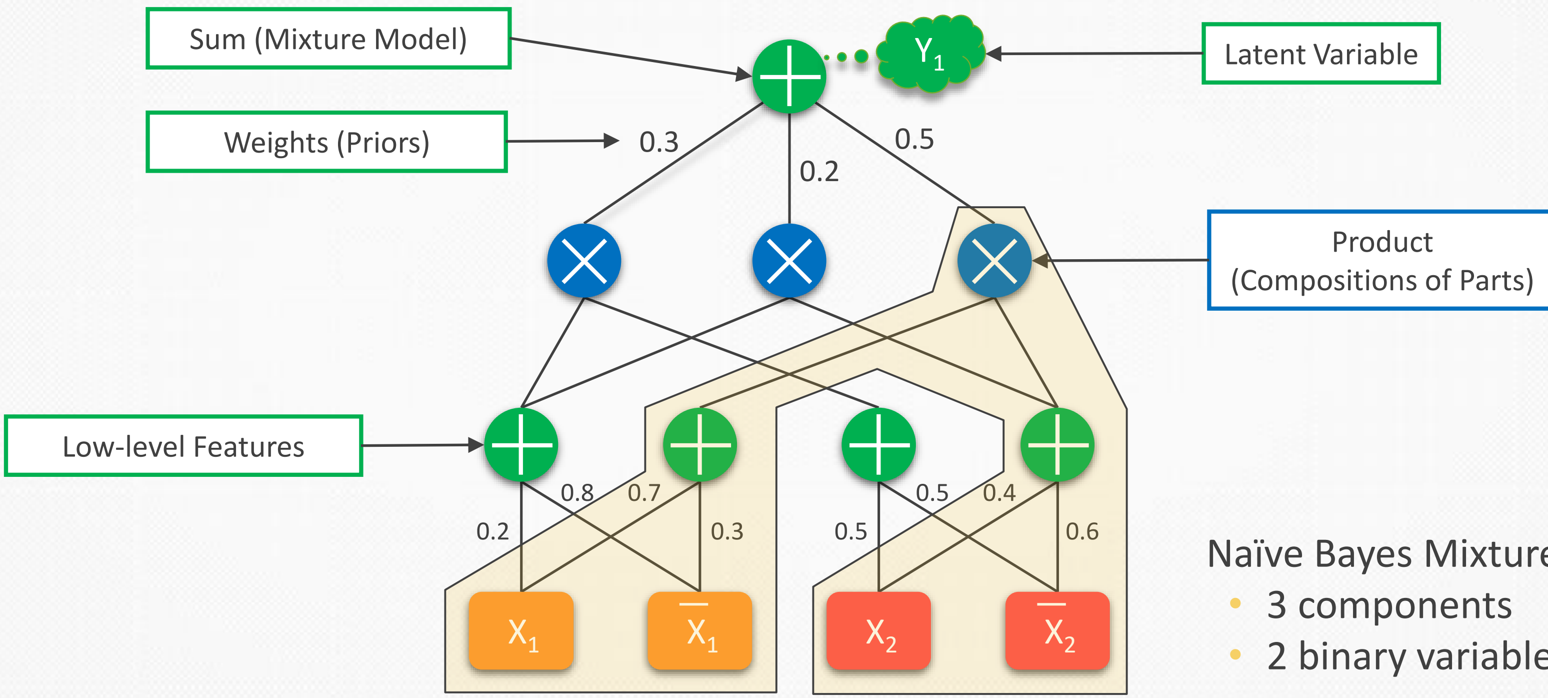
- 2 Views: Deep architecture and Graphical model
- Learn conditional or joint distributions
- Tractable partition function, exact inference
- Structure semantics: hierarchical mixture of parts



Naïve Bayes Mixture Model
- 3 components
- 2 binary variables

SUM-PRODUCT NETWORKS

[Poon & Domingos, UAI'11, Friesen & Domingos, ICML'16]



Naïve Bayes Mixture Model
- 3 components
- 2 binary variables

- Large SPNs can be very deep

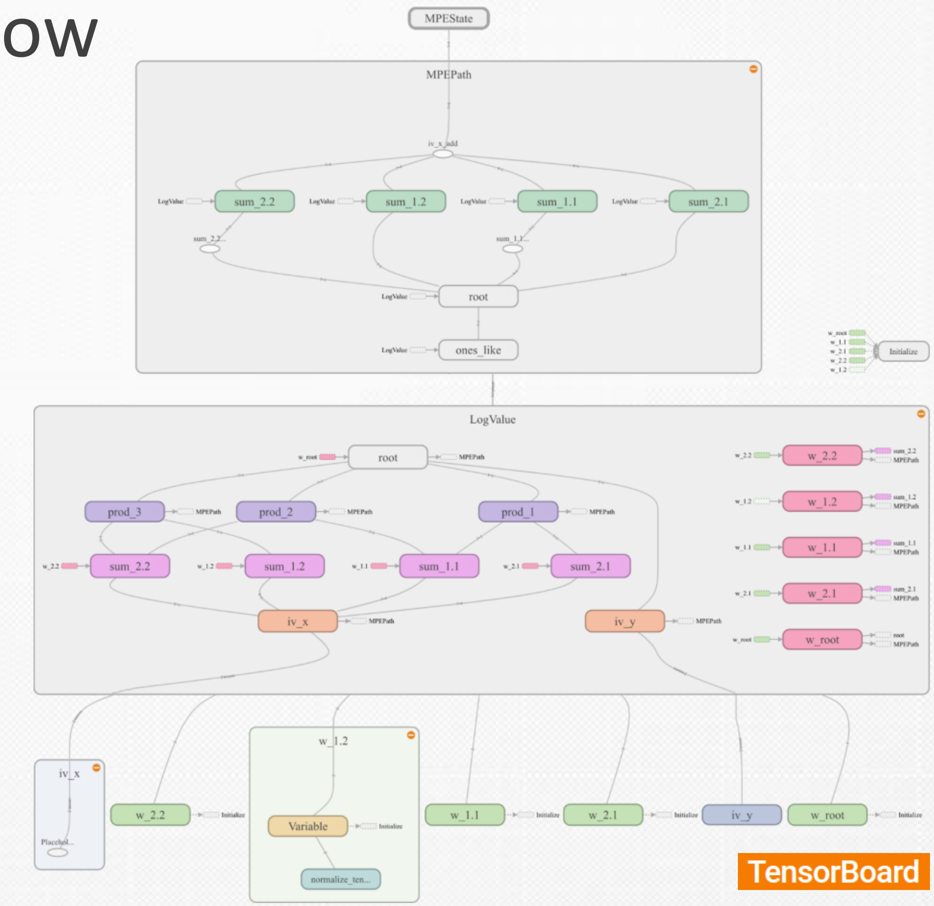
SPNS: LEARNING AND INFERENCE

- Learning
  - Generative (EM/GD) [Poon & Domingos, UAI'11]
  - Discriminative (GD) [Gens & Domingos, NIPS'12]
  - Simultaneous learning of parameters and structure
    [Gens & Domingos, ICML'13, Hsu et al., ICLR'17]
- Inference
  - Single up/down pass through the network
  - Upwards pass:
    - Probability of evidence
  - Downwards pass:
    - Gradients to obtain marginals
    - MPE state of variables

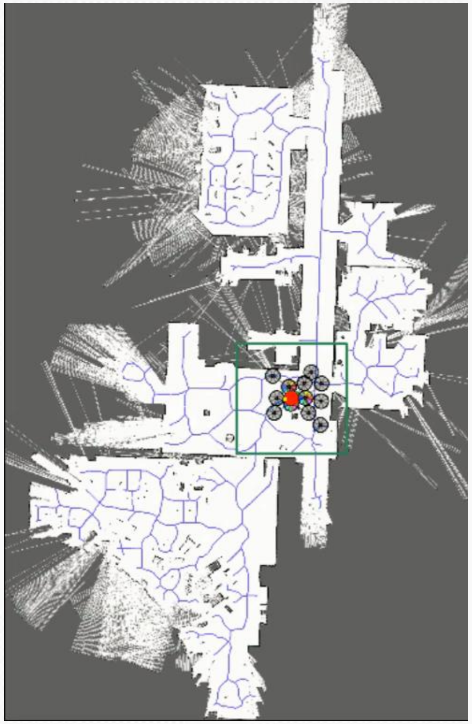
LIBSPN

[Pronobis, Ranganath, Rao, ICML-PADL'17]

- New general-purpose Python library for SPNs
- Learning and inference in large networks
- Integrated with TensorFlow
  - Multi-GPU computations
  - Integration of SPNs with other models
- Open-source soon at: **http://libspn.org**



GRAPHSPNS: STRUCTURED PREDICTION WITH SPNS

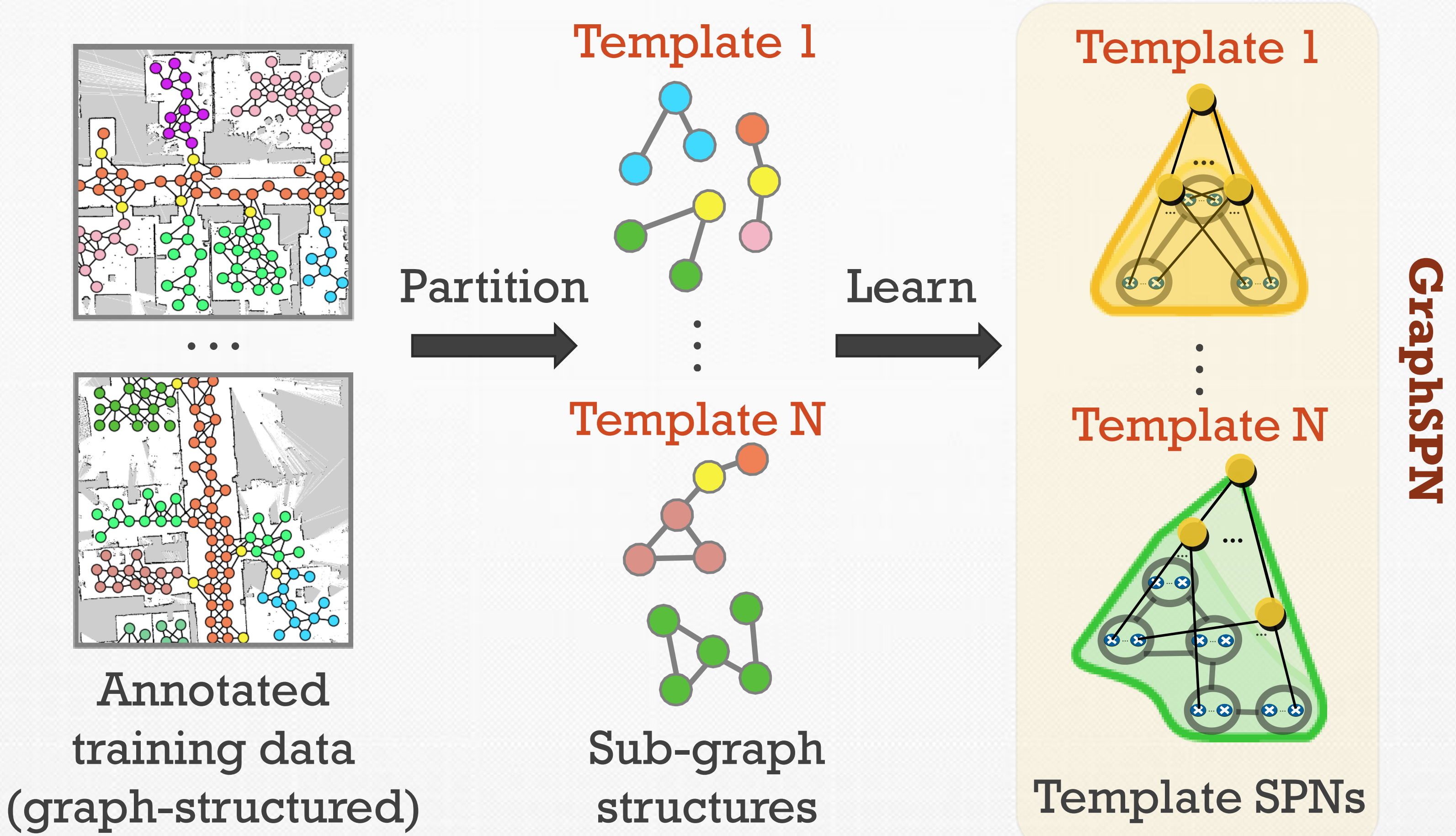


- Large environments pose complex structured prediction problems
  - Structured by graphs of varying size
  - Contaminated with noise
- Many approaches (deep), but:
  - Strict constraints on variable interactions
  - Fixed number of variables
  - Static global structure
- GraphSPNs [Zheng, Pronobis, Rao, AAAI'18]
  - Probabilistic SP approach
  - Dependencies between latent variables expressed in terms of **arbitrary, dynamic graphs**
  - Applicable to many problems (3D scenes, segmentation)

GRAPHSPNS: LEARNING

[Zheng, Pronobis, Rao, AAAI'18]

- Template model defined as a set of template SPNs representing learned, higher-order relations

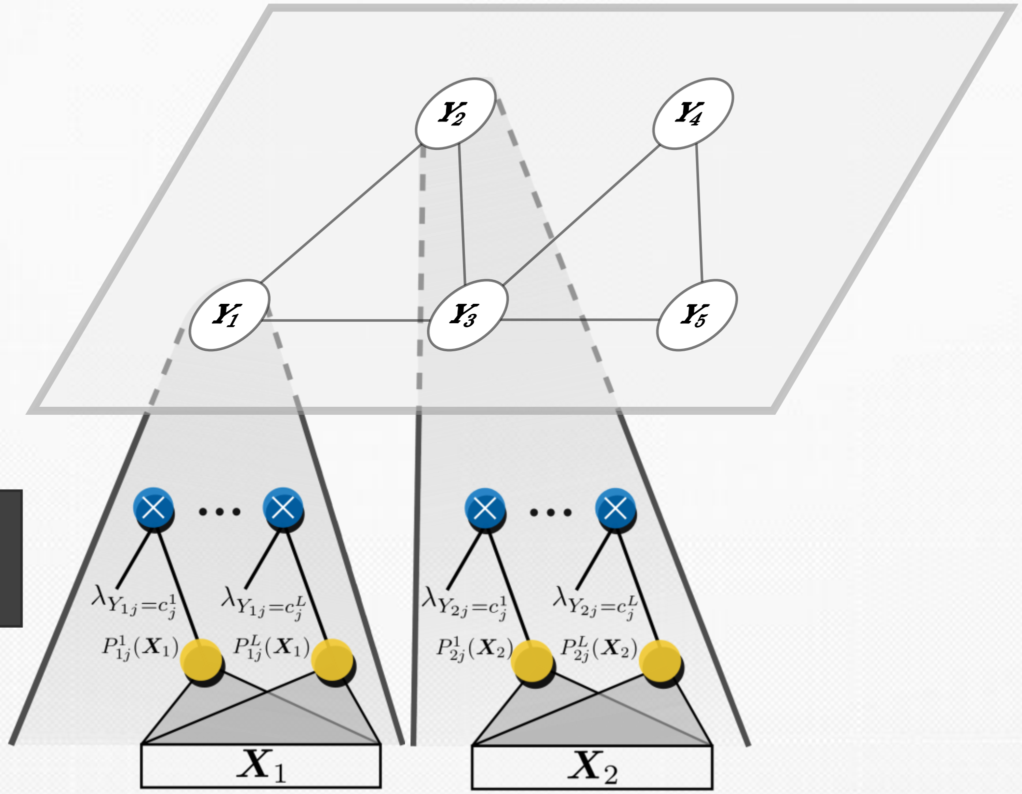


GRAPHSPNS: INFERENCE

[Zheng, Pronobis, Rao, AAAI'18]

- Learned template models form a single distribution $P_D^{\mathcal{G}}(\boldsymbol{X},\boldsymbol{Y})$ for a particular graph-structured problem

Arbitrary local distributions

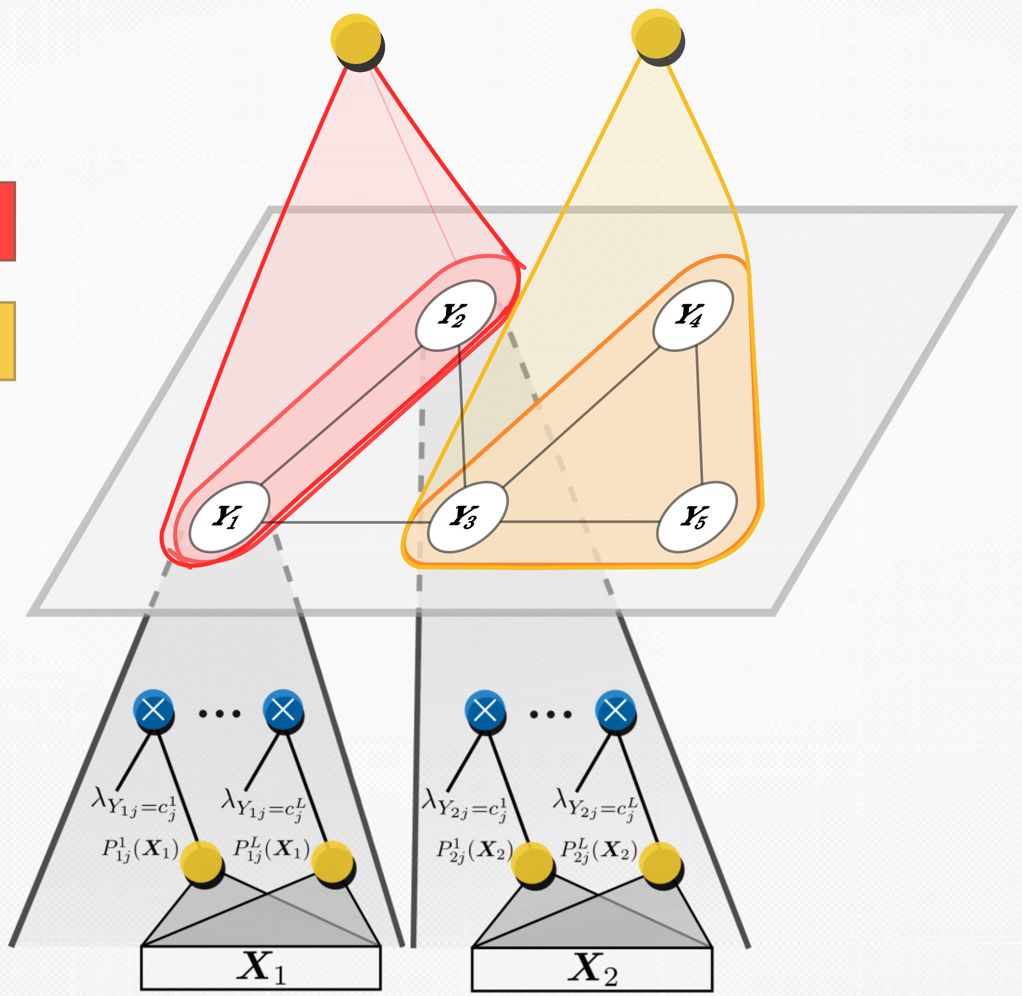


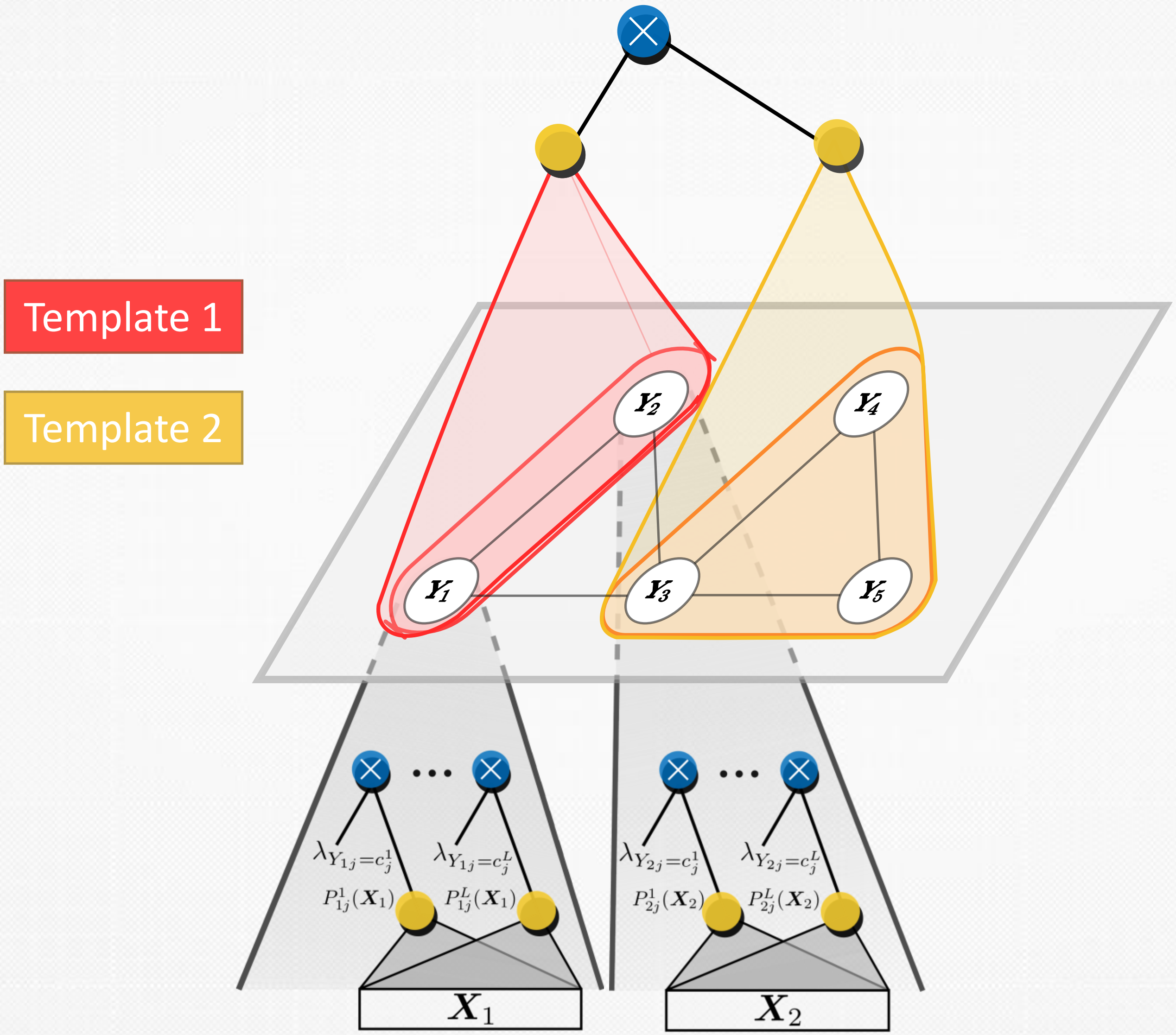
GRAPHSPNS: INFERENCE

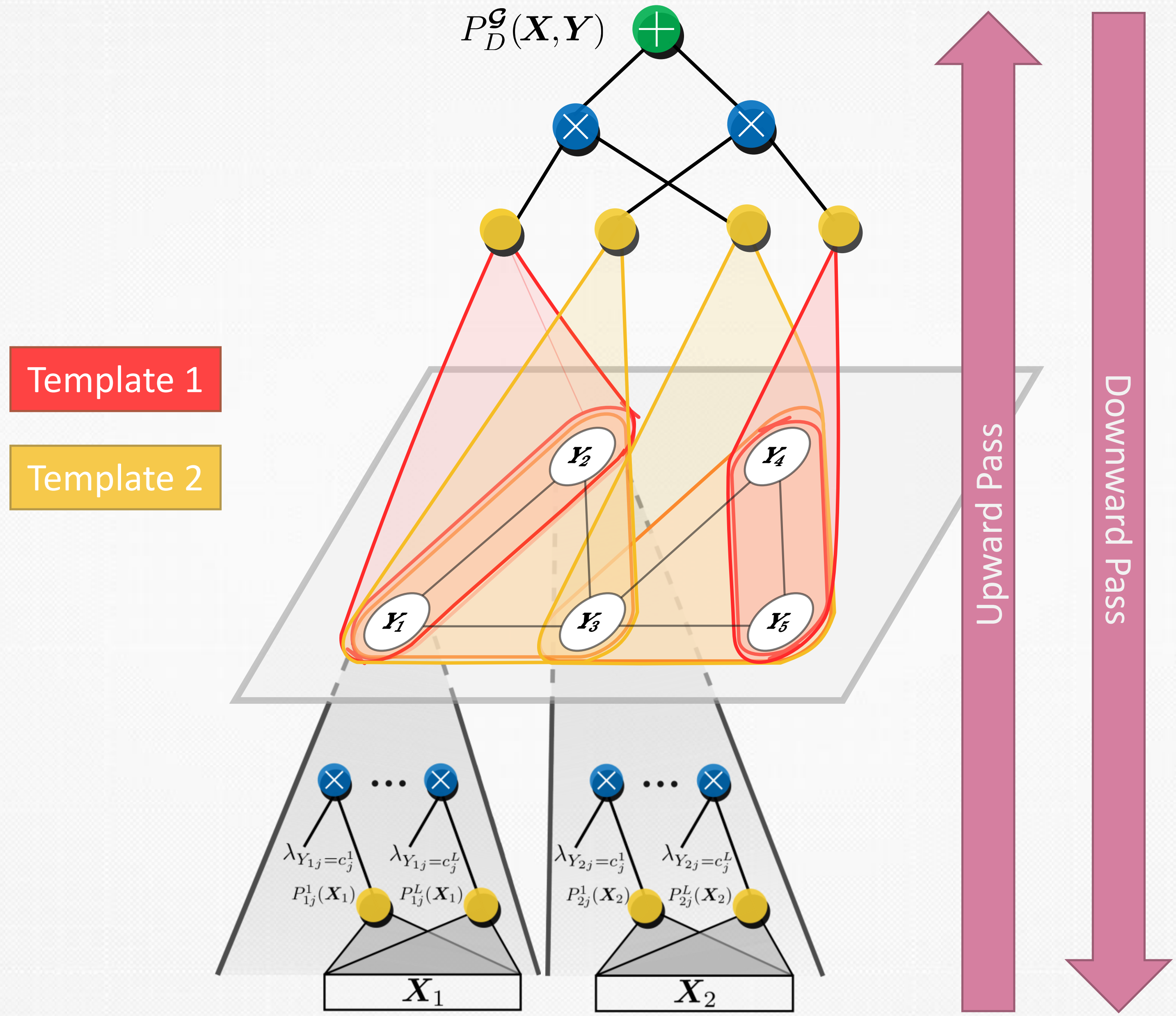
[Zheng, Pronobis, Rao, AAAI’18]

Template 1

Template 2




Template 1
Template 2
Y_1
Y_2
Y_3
Y_4
Y_5
$\lambda_{Y_{1j}=c_j^1}$
$\lambda_{Y_{1j}=c_j^L}$
$\lambda_{Y_{2j}=c_j^1}$
$\lambda_{Y_{2j}=c_j^L}$
$P_{1j}^1(\boldsymbol{X}_1)$
$P_{1j}^L(\boldsymbol{X}_1)$
$P_{2j}^1(\boldsymbol{X}_2)$
$P_{2j}^L(\boldsymbol{X}_2)$
$\boldsymbol{X}_1$
$\boldsymbol{X}_2$

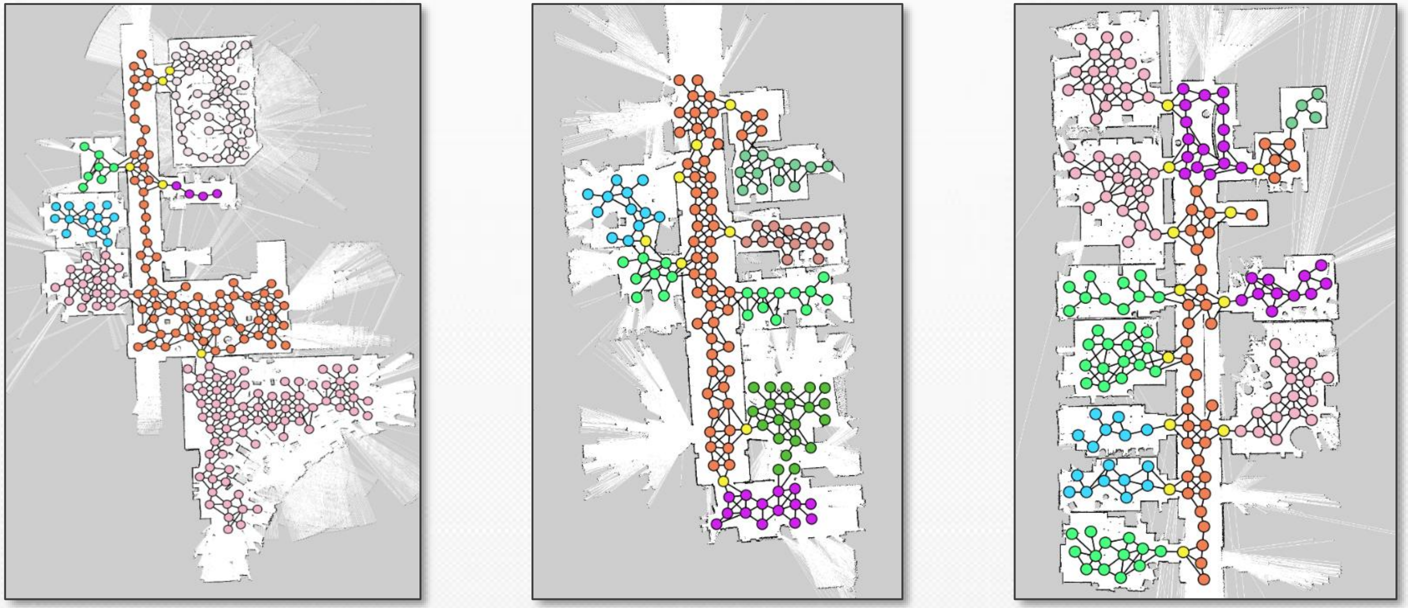


$P_D^{\mathcal{G}}(\boldsymbol{X},\boldsymbol{Y})$
Template 1
Template 2
Y_1
Y_2
Y_3
Y_4
Y_5
$\lambda_{Y_{1j}=c_j^1}$
$\lambda_{Y_{1j}=c_j^L}$
$P_{1j}^1(\boldsymbol{X}_1)$
$P_{1j}^L(\boldsymbol{X}_1)$
$\lambda_{Y_{2j}=c_j^1}$
$\lambda_{Y_{2j}=c_j^L}$
$P_{2j}^1(\boldsymbol{X}_2)$
$P_{2j}^L(\boldsymbol{X}_2)$
$\boldsymbol{X}_1$
$\boldsymbol{X}_2$
Upward Pass
Downward Pass

GRAPHSPNS: EVALUATION

[Zheng, Pronobis, Rao, AAAI'18]

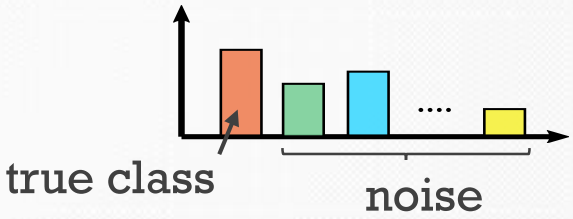
- GraphSPNs vs graphical models (MRFs)
- Data structured by topological graphs
  - 99 graphs: 11 floors in 3 buildings
  - Places labeled with 10 semantic categories
  - Leave-one-building-out procedure



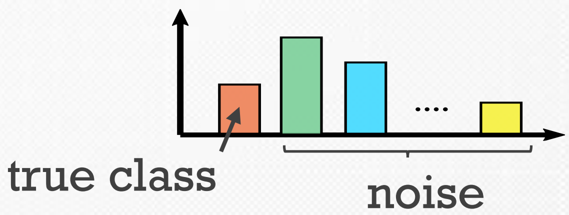
GRAPHSPNS: EVALUATION

[Zheng, Pronobis, Rao, AAAI'18]

- GraphSPNs vs graphical models (MRFs)
- Data structured by topological graphs
  - 99 graphs: 11 floors in 3 buildings
  - Places labeled with 10 semantic categories
  - Leave-one-building-out procedure
- Local evidence:
  - Distributions over semantic categories
  - Noisified ground truth



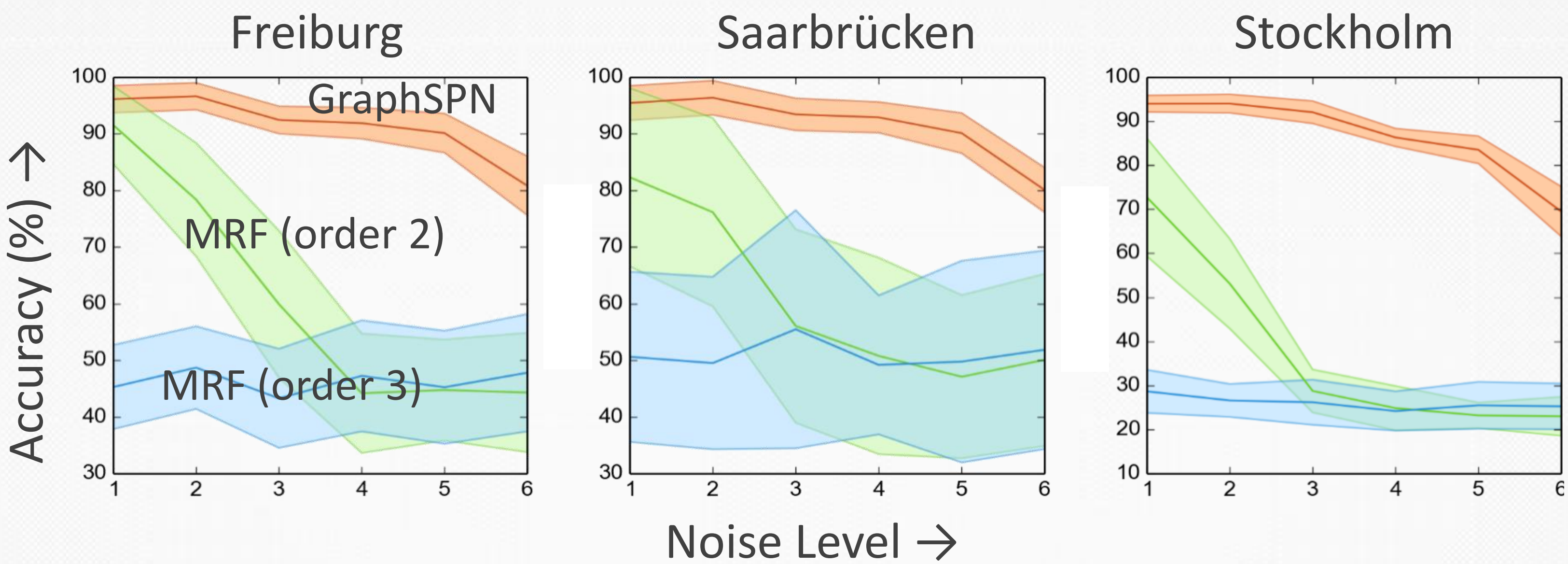
- Local distributions:
  - Unary potentials (MRFs)
  - Corresponding basic distributions (GraphSPN)



- Task: Recover true semantic labels of places

GRAPHSPNS: RESULTS

[Zheng, Pronobis, Rao, AAAI'18]



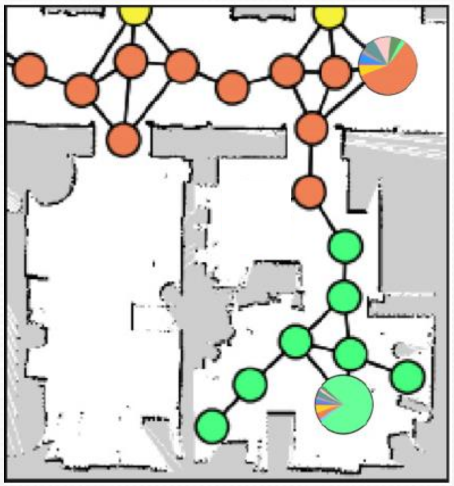
Predictions for nodes without evidence (unexplored placeholders):

GraphSPN

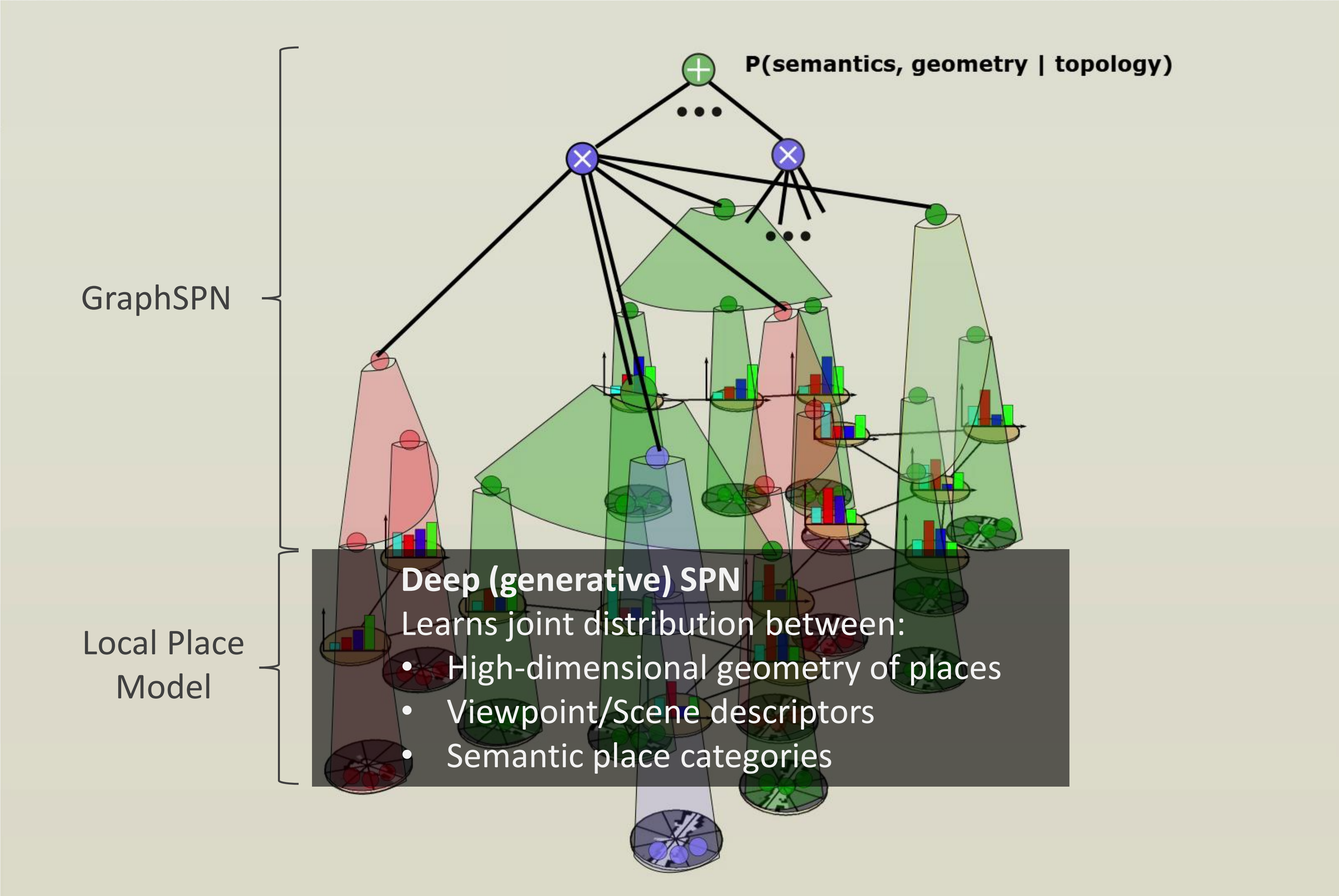
Freiburg	Saarbrücken	Stockholm
67.58%(+/-10.42)	78.15%(+/-9.95)	67.57%(+/-11.11)

MRF-2

Freiburg	Saarbrücken	Stockholm
28.32%(+/-7.53)	39.85%(+/-19.42)	12.44%(+/-3.46)



GENERAL KNOWLEDGE: END2END DEEP APPROACH

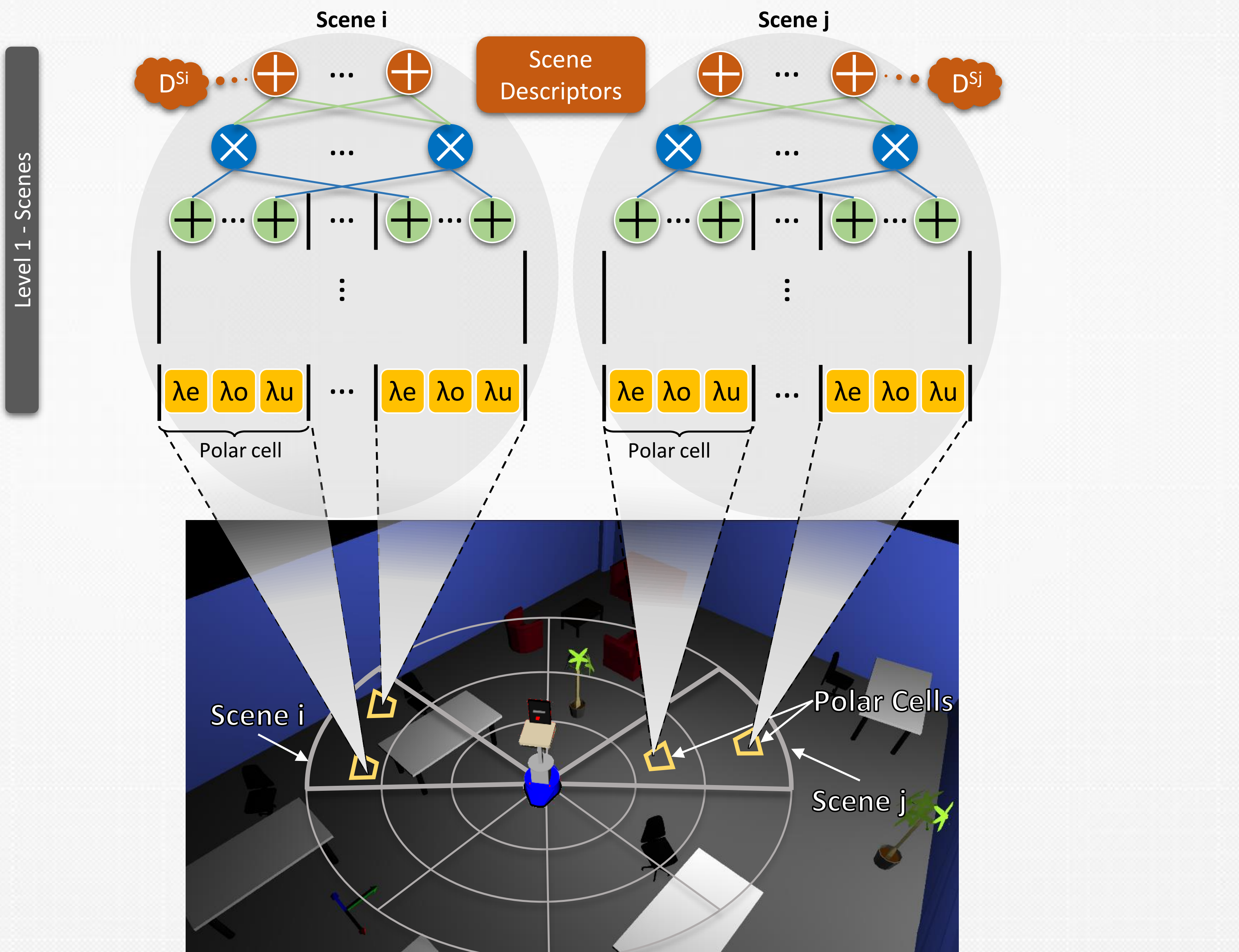


Deep (generative) SPN

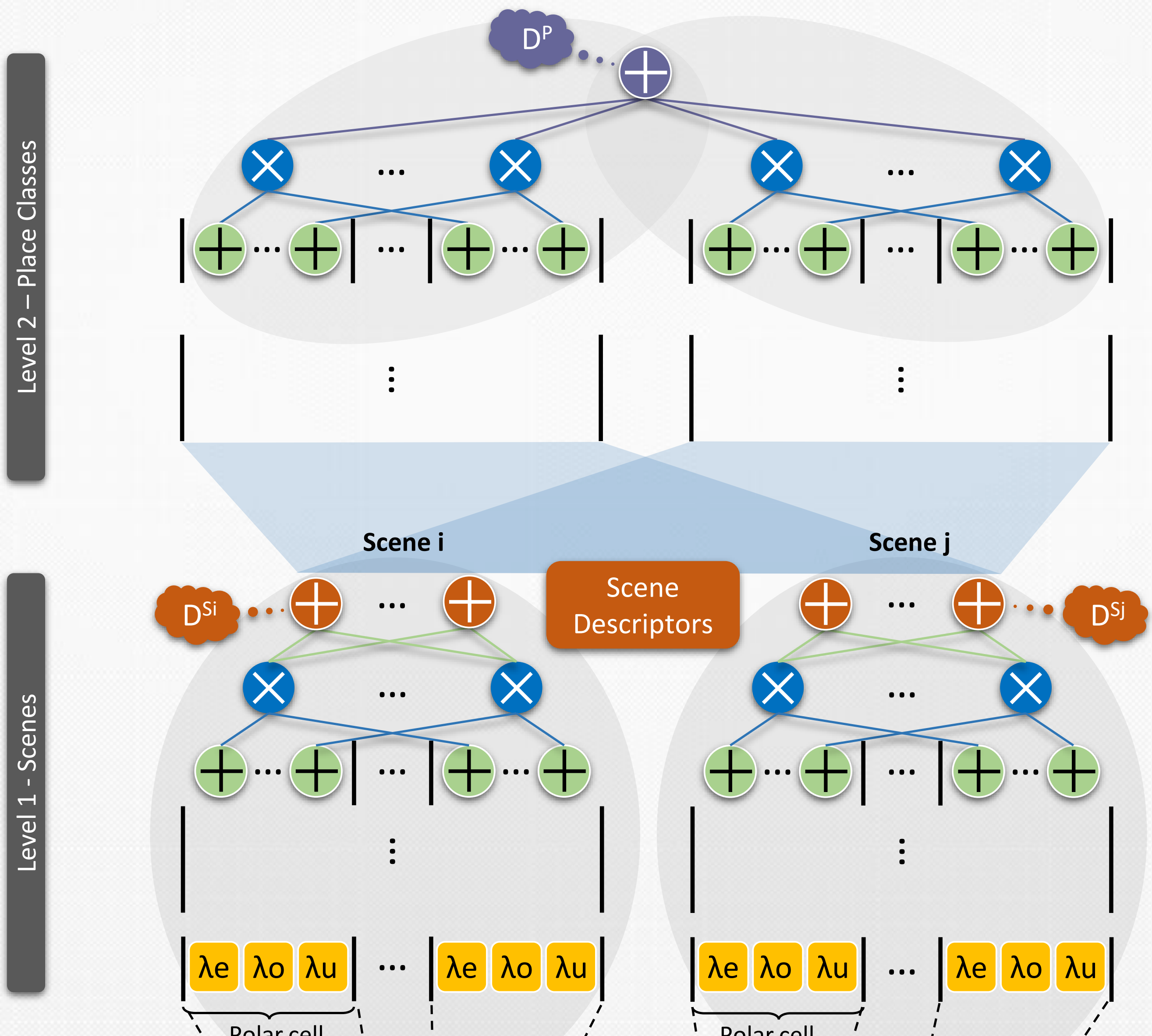
Learns joint distribution between:

- High-dimensional geometry of places
- Viewpoint/Scene descriptors
- Semantic place categories

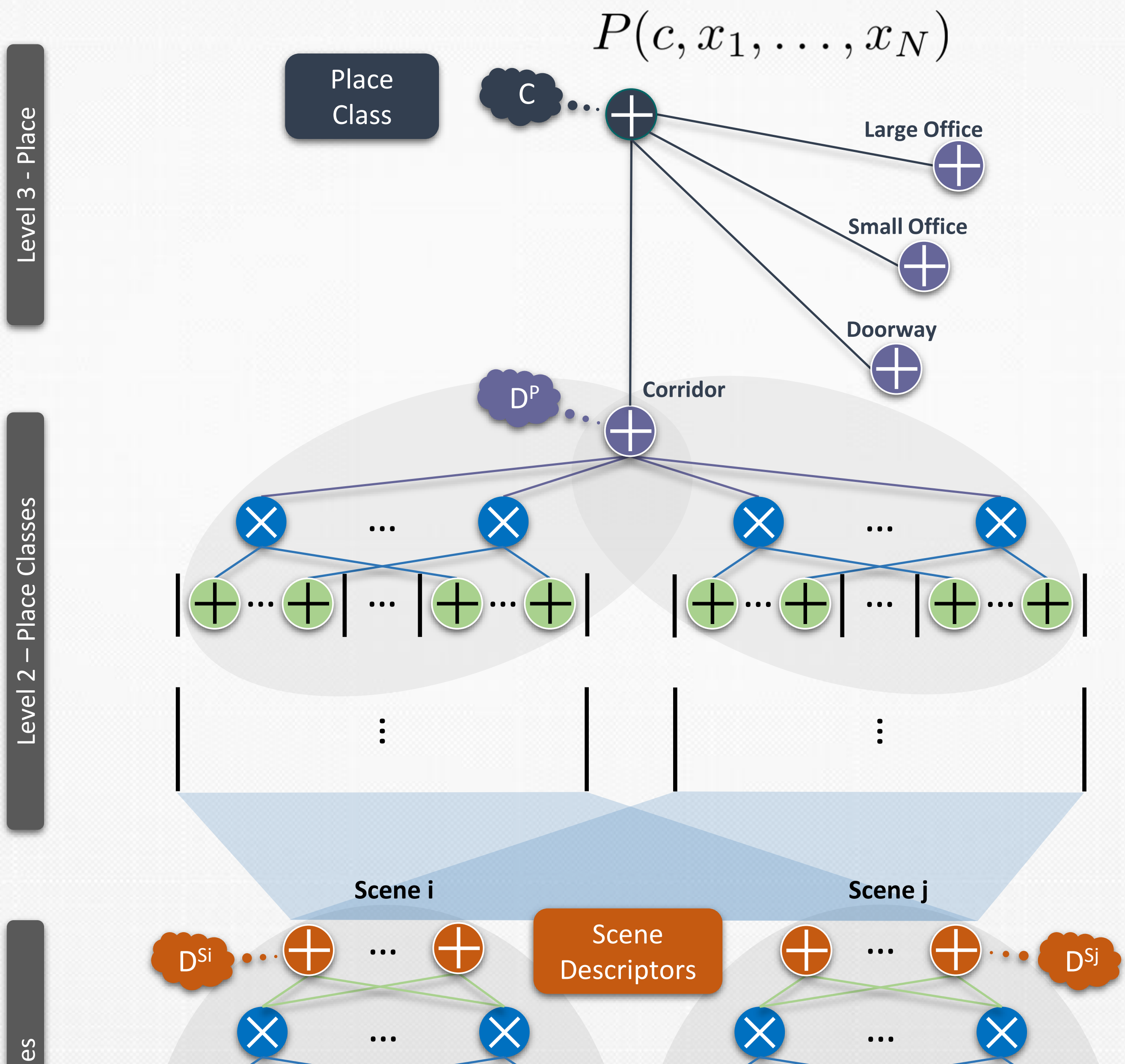
LOCAL PLACE MODEL: ARCHITECTURE [Pronobis, Rao, IROS'17]



LOCAL PLACE MODEL: ARCHITECTURE [Pronobis, Rao, IROS'17]



LOCAL PLACE MODEL: ARCHITECTURE [Pronobis, Rao, IROS'17]



LOCAL PLACE MODEL: EVALUATION

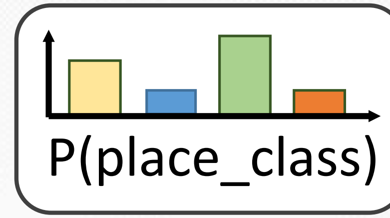
[Pronobis, Rao, IROS'17]

- Semantically annotated sequences of sensory data
  - Robot navigating 4 floors of office building
  - Each floor contains multiple instances of: small office, large office, corridor, doorway **(considered known)**
  - Additional few instances of: kitchen, elevator, lab, living room, meeting room **(considered novel)**
  - Leave-one-floor-out procedure
- Sensor: Laser-range scanner
- Learning: Generative hard EM

BOTTOM-UP INFERENCE: PLACE CATEGORIZATION

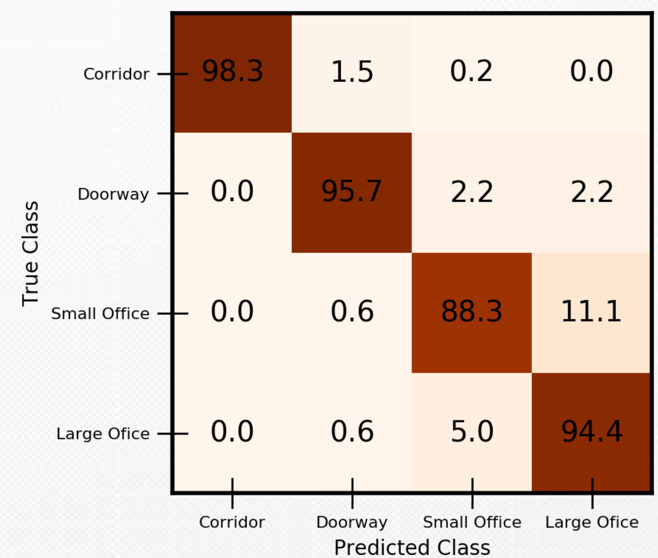
[Pronobis, Rao, IROS'17]

$$c^* = \underset{c}{\operatorname{argmax}}\ P(c, x_1, \ldots, x_N)$$



- Baseline:
  - SVM + RBF Kernel
  - Geometric features [Mozos et al. '05] from high-res 360° virtual scans
- Average CR:
  - **SVM** (discriminative): 85.9%
  - **DGSM (generative): 92.7%**

Confusion Matrix (DGSM)



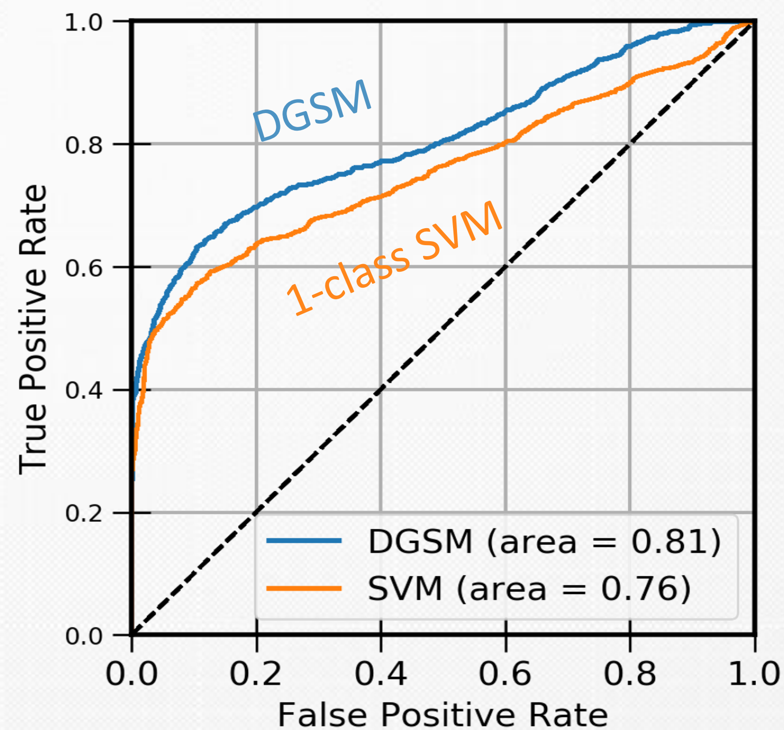
BOTTOM-UP INFERENCE: NOVELTY DETECTION

[Pronobis, Rao, IROS'17]

$$P(x_1, \ldots, x_N) > threshold$$

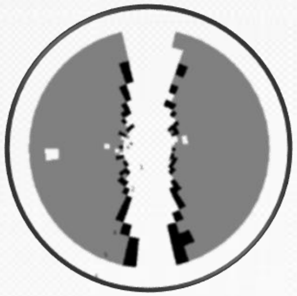
Known / Novel

- Baseline:
  - One-class SVM + RBF Kernel
  - Geometric features



TOP-DOWN INFERENCE: GENERATING PROTOTYPES

[Pronobis, Rao, IROS'17]



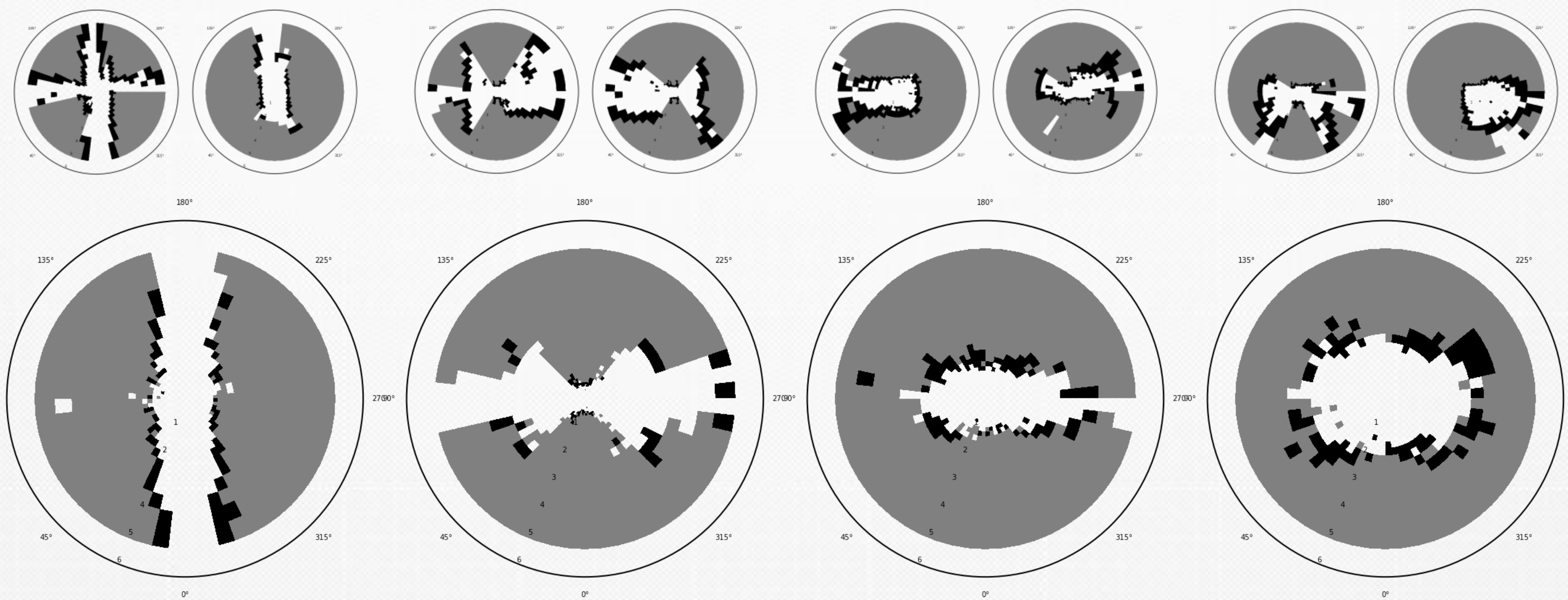
Corridor

$$x_1^*,\dots,x_N^* = \underset{x_1,\dots,x_N}{\operatorname{argmax}}\; P(x_1,\dots,x_N|c)$$

Corridor | Doorway | Small Office | Large Office

True

Generated

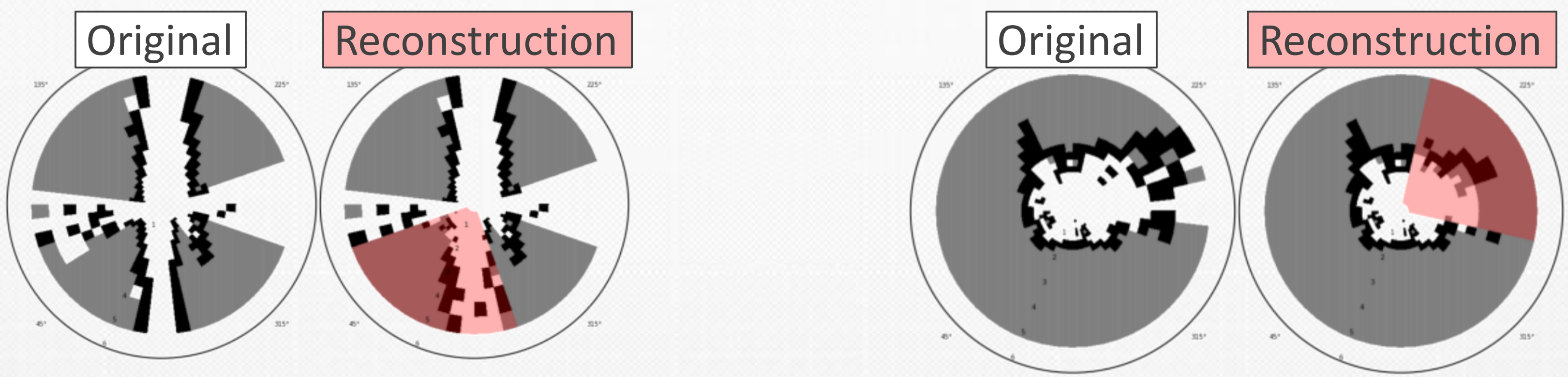


TOP-DOWN INFERENCE: MISSING OBSERVATIONS

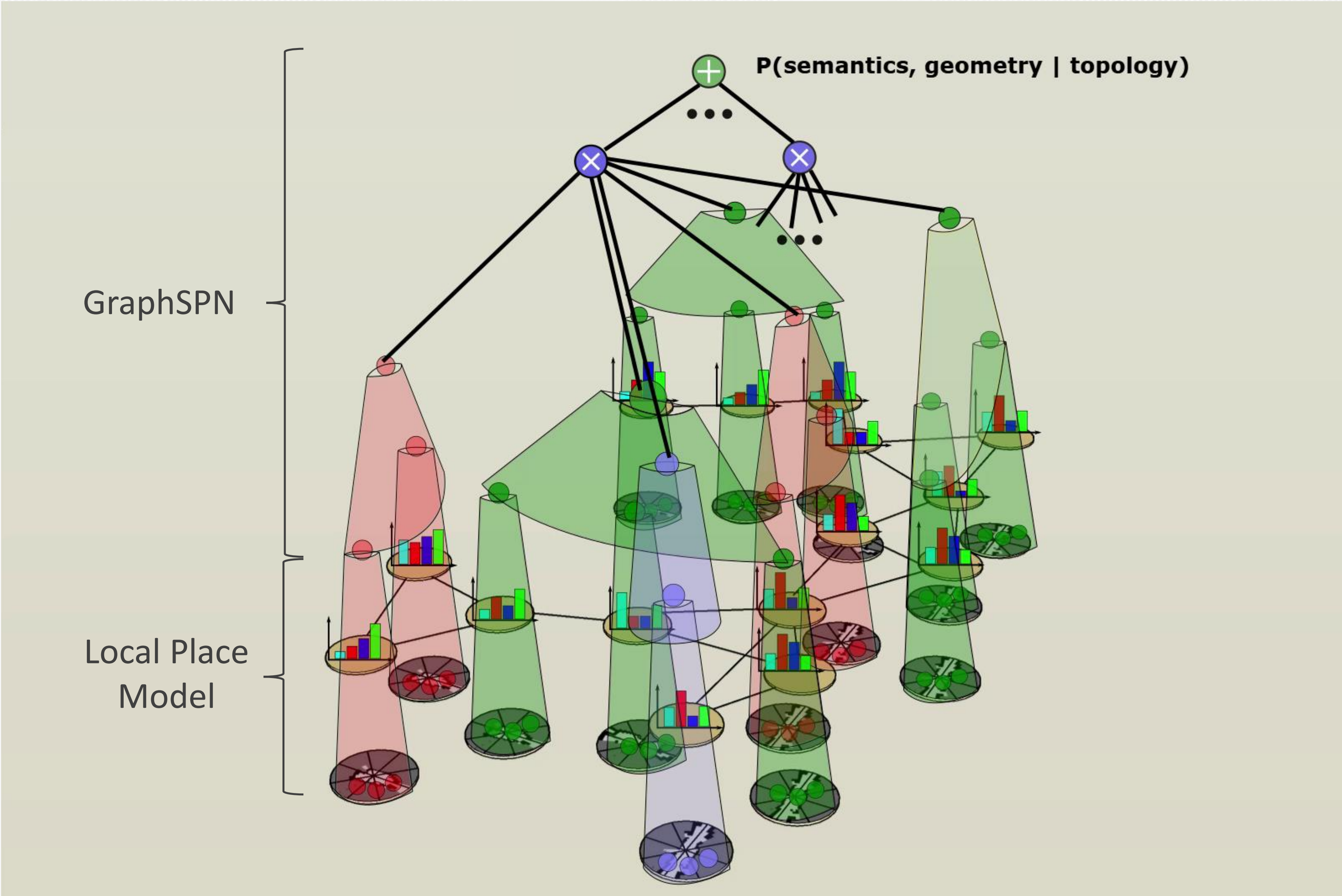
[Pronobis, Rao, IROS'17]

$$x^*_J, \ldots, x^*_N = \operatorname*{argmax}_{x_J, \ldots, x_N} P(x_1, \ldots, x_J, \ldots, x_N)$$

- Baseline: DC-GANs + GD-based inpainting [Yeh et al. '16]
- Correctly predicted: **DGSM: 77.1%** **DC-GAN**: 75.8%

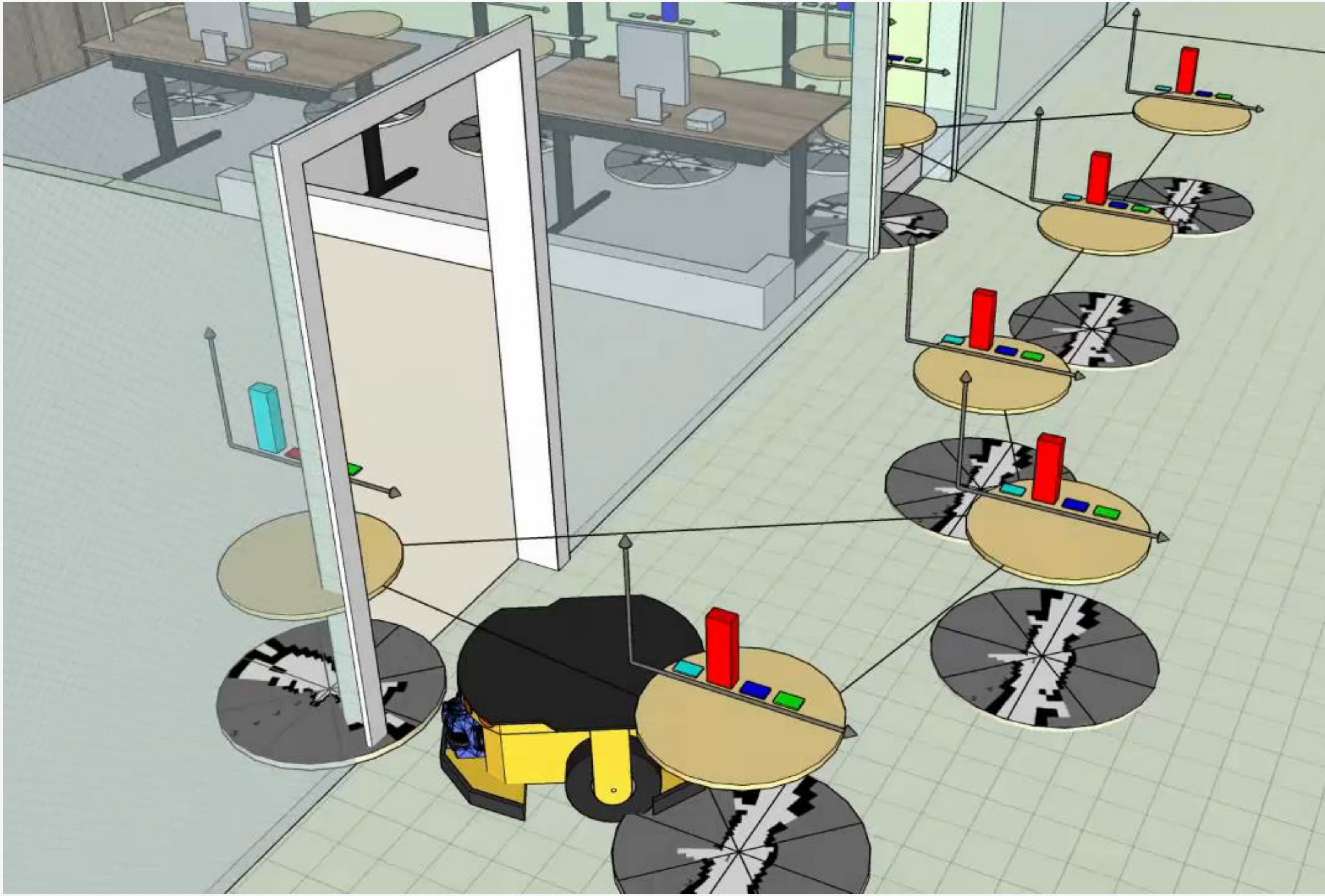


GENERAL KNOWLEDGE: END2END DEEP APPROACH



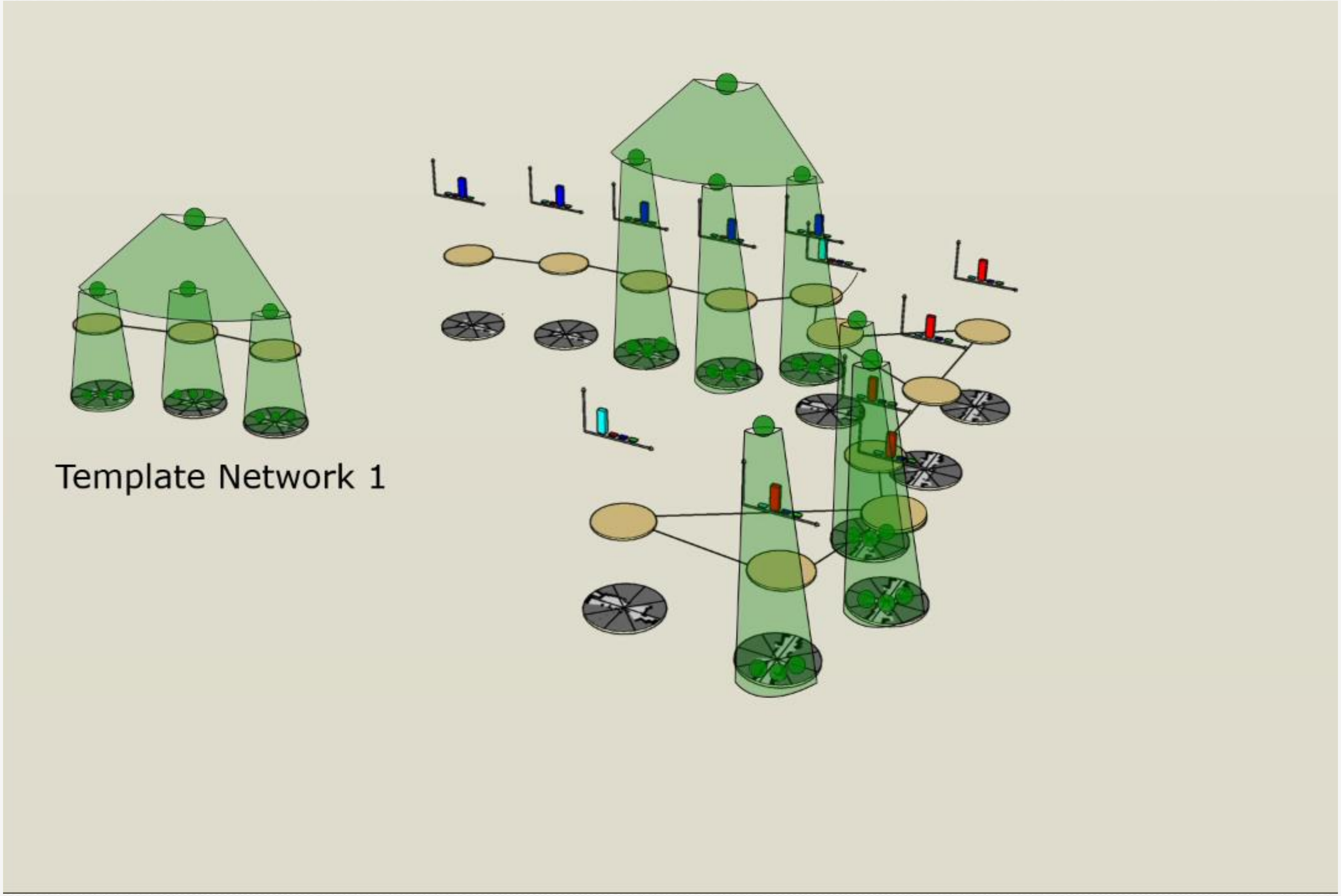
END2END: LEARNING

[Zheng, Pronobis '19]



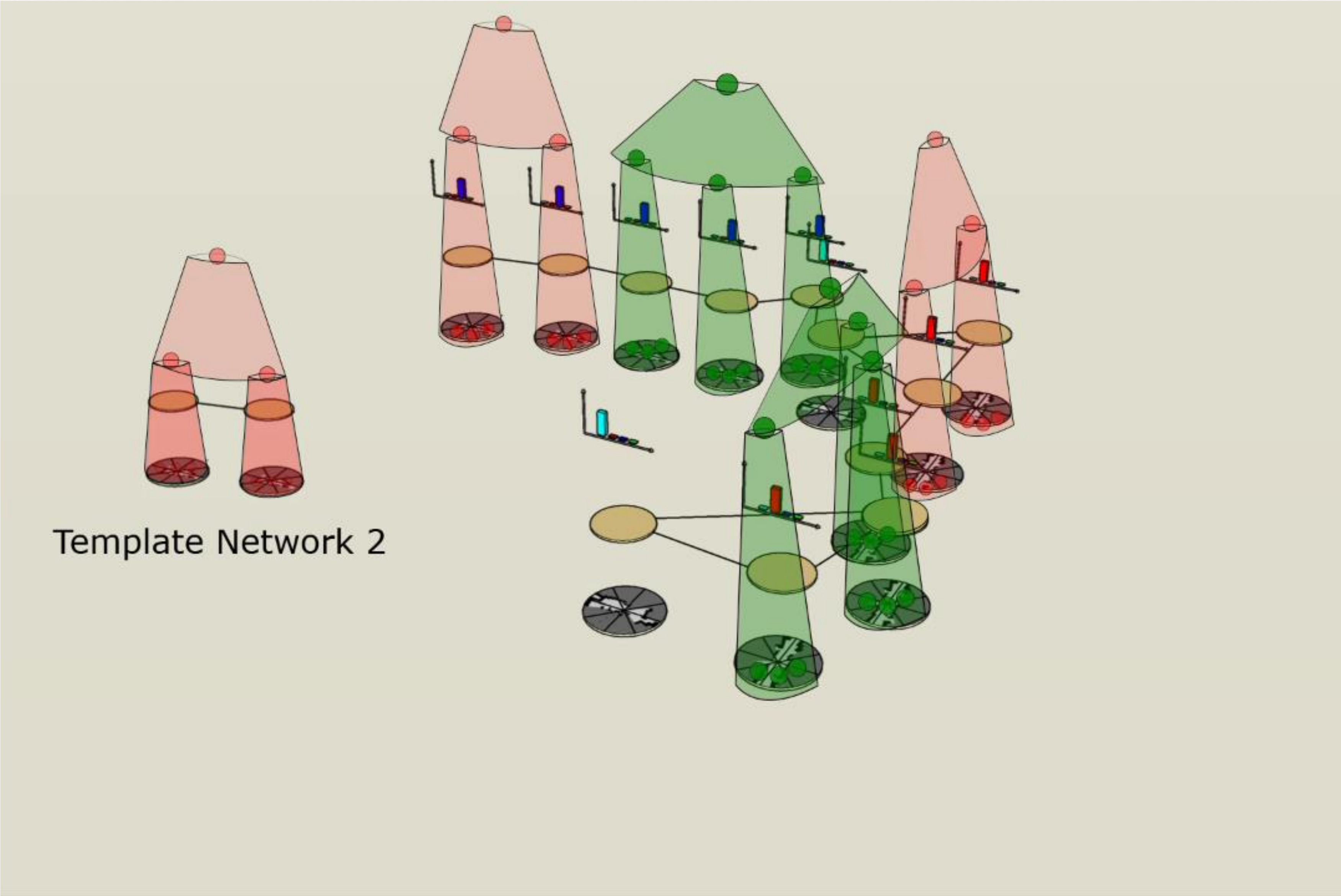
END2END: LEARNING

[Zheng, Pronobis '19]



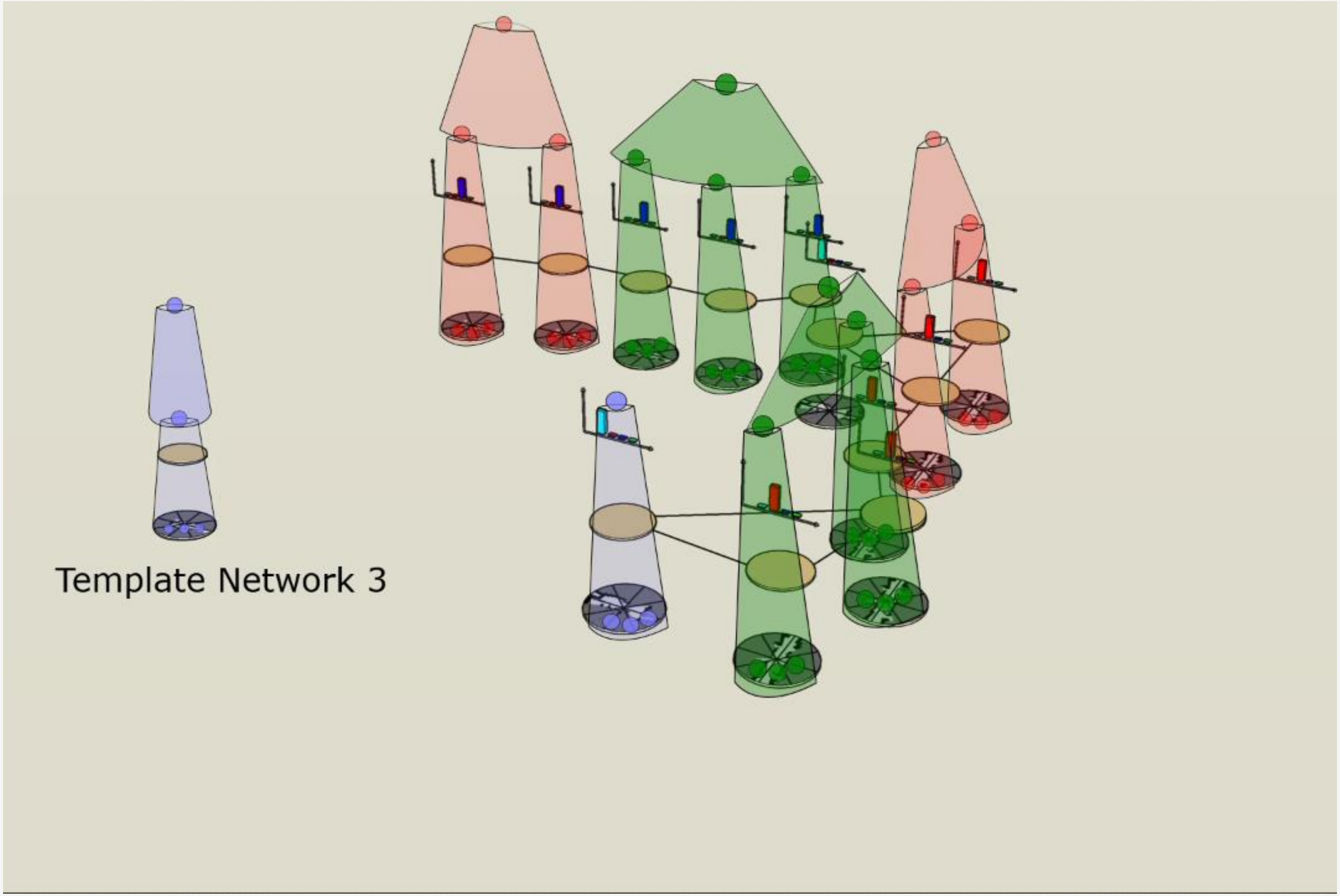
END2END: LEARNING

[Zheng, Pronobis '19]



END2END: LEARNING

[Zheng, Pronobis '19]

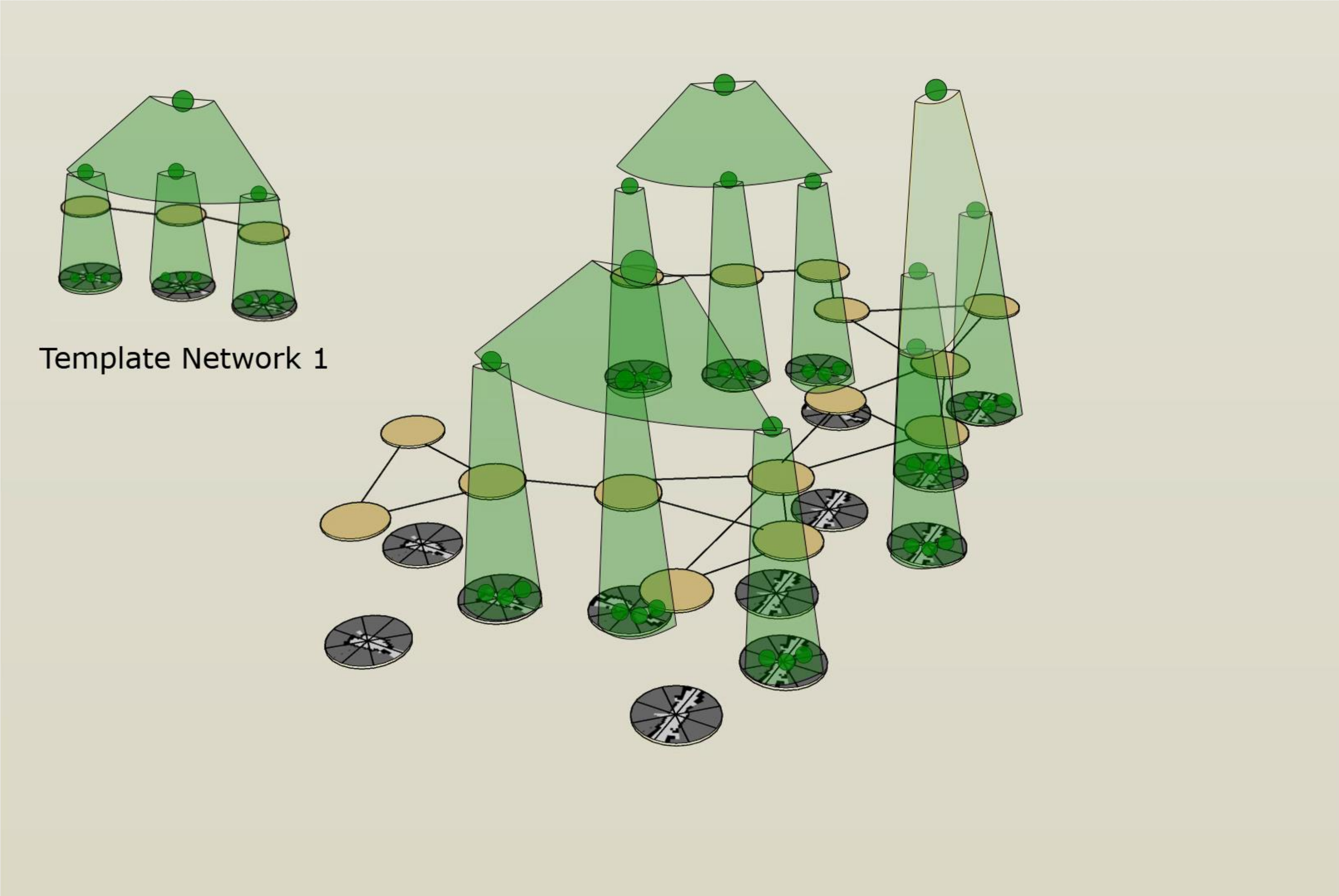


END2END: INFERENCE

[Zheng, Pronobis '19]

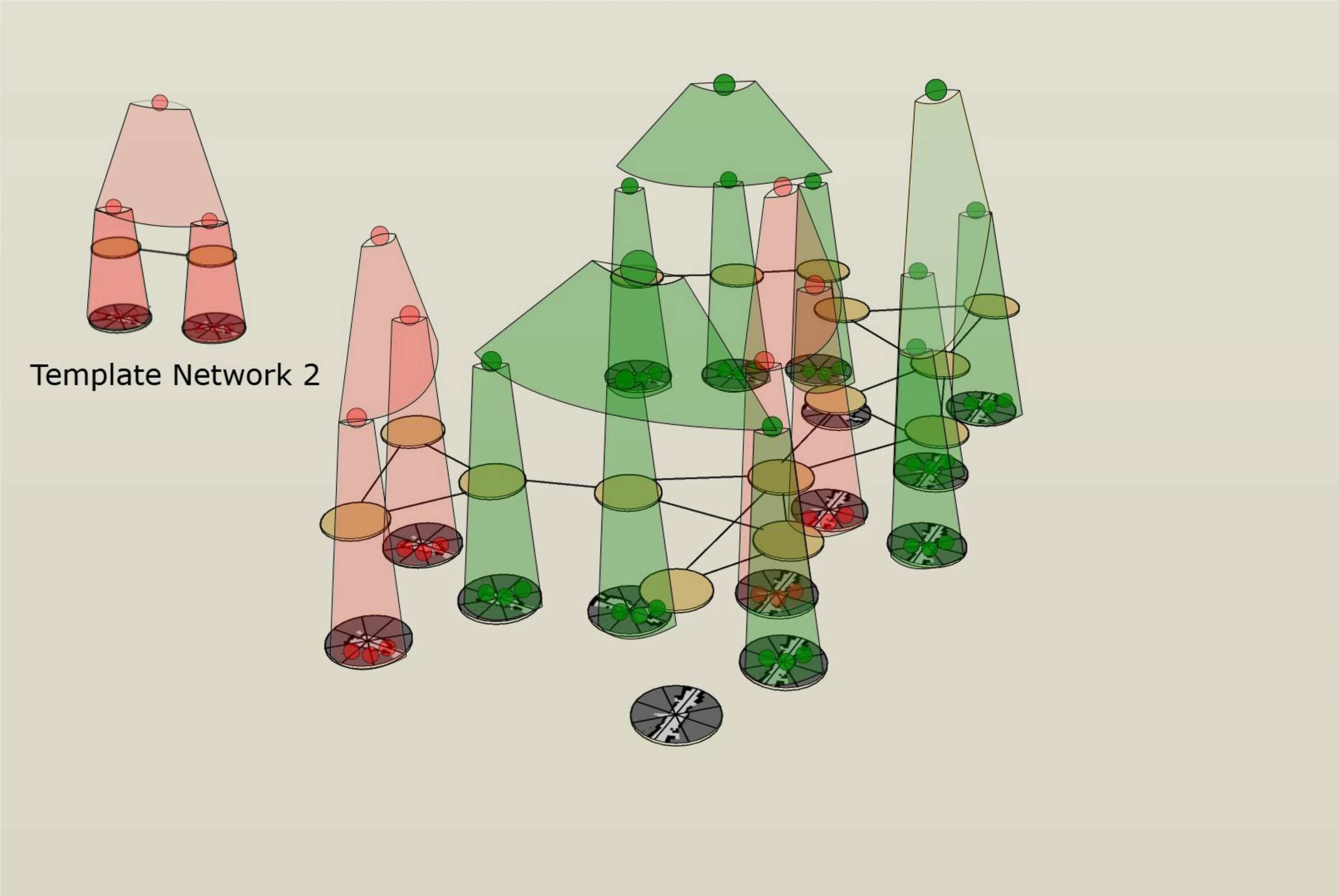
END2END: INFERENCE

[Zheng, Pronobis '19]



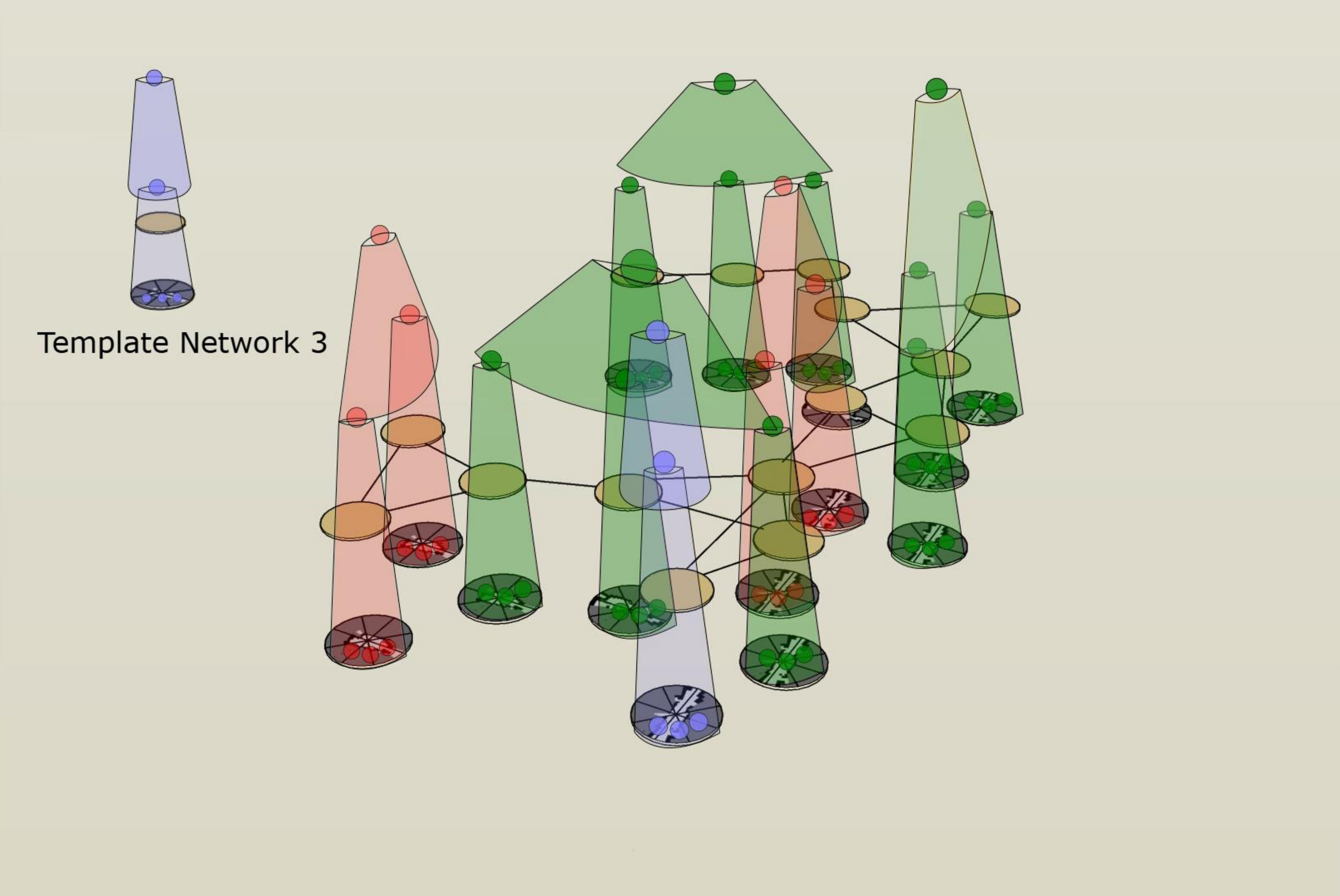
END2END: INFERENCE

[Zheng, Pronobis '19]



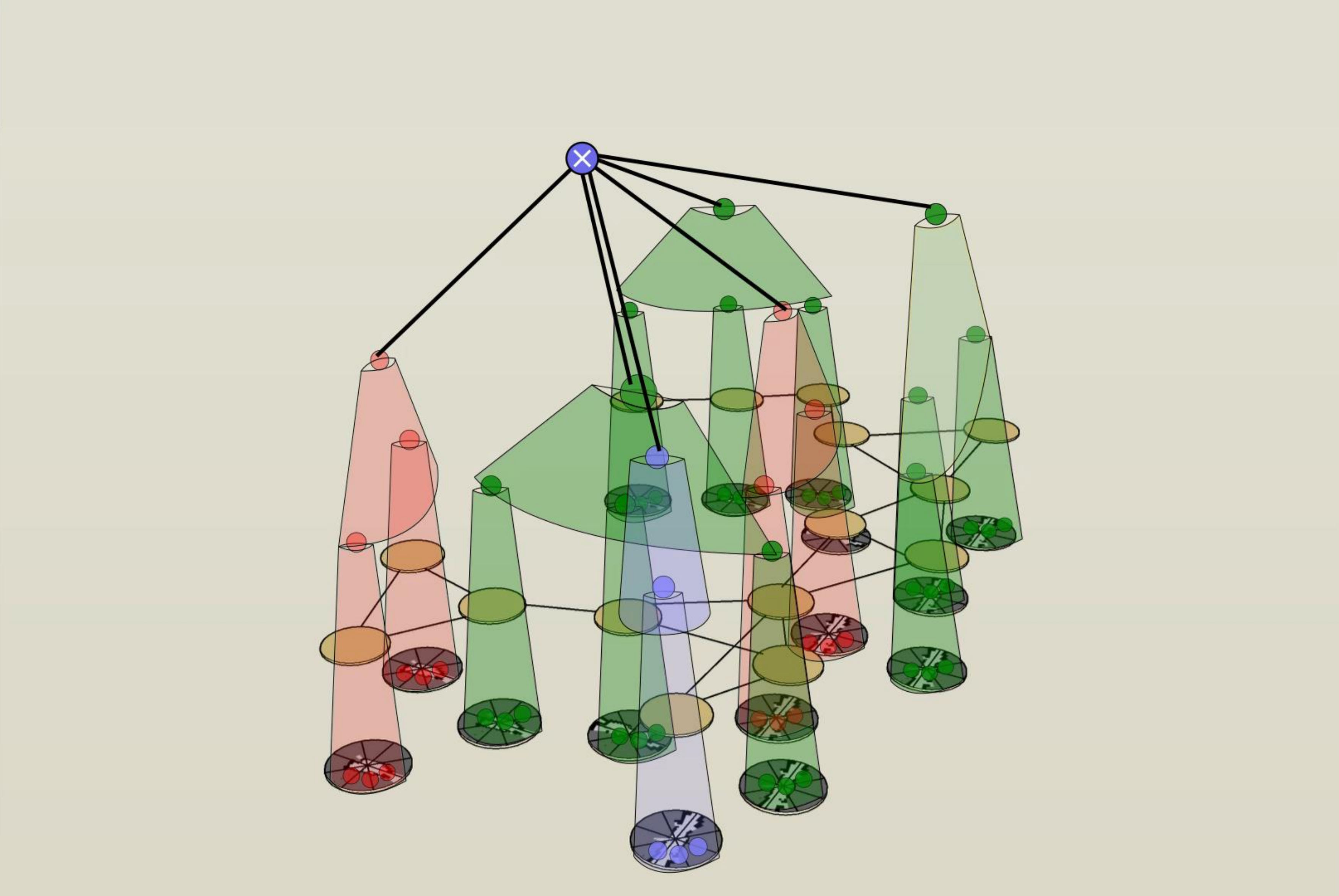
END2END: INFERENCE

[Zheng, Pronobis '19]



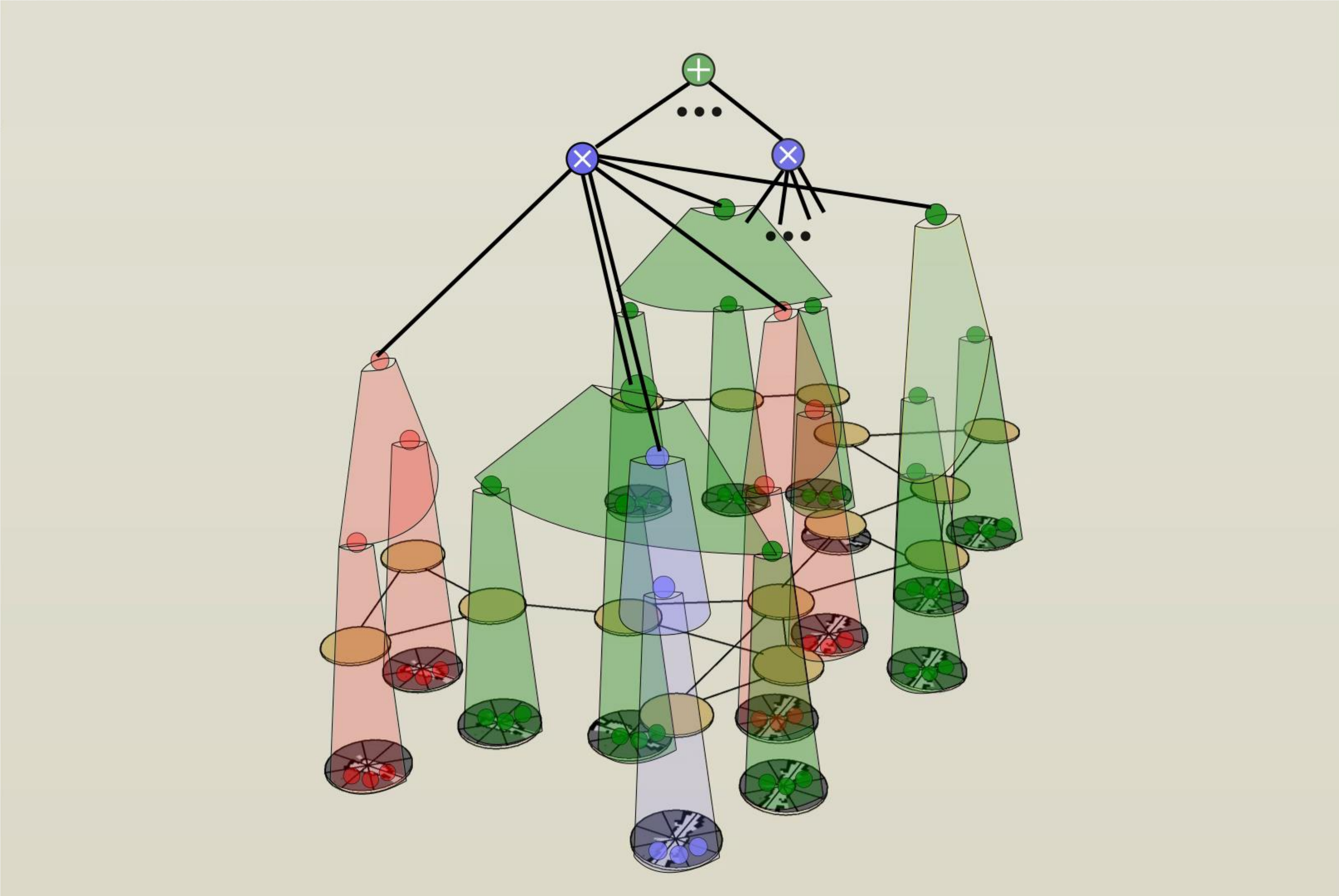
END2END: INFERENCE

[Zheng, Pronobis '19]



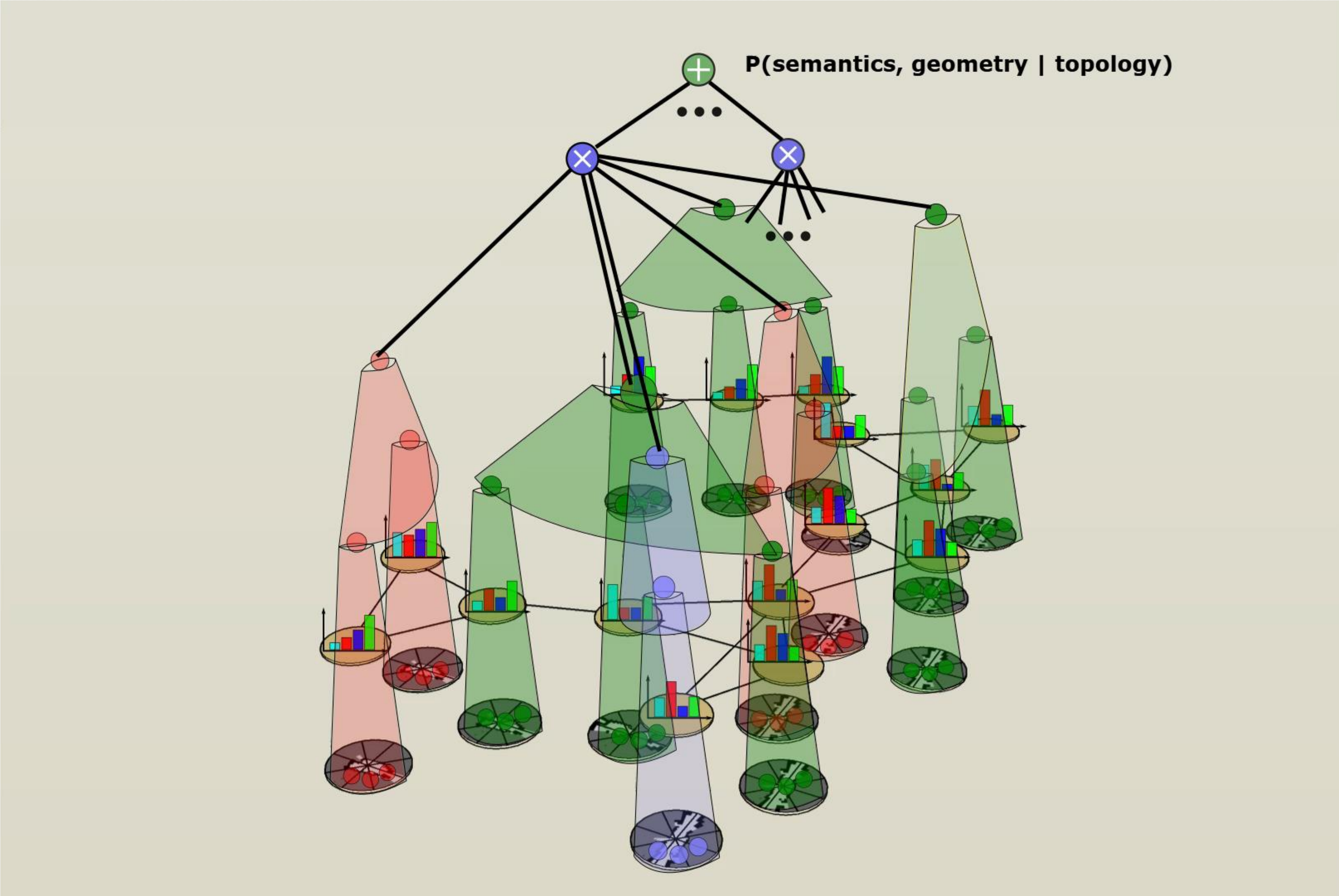
END2END: INFERENCE

[Zheng, Pronobis '19]



END2END: INFERENCE

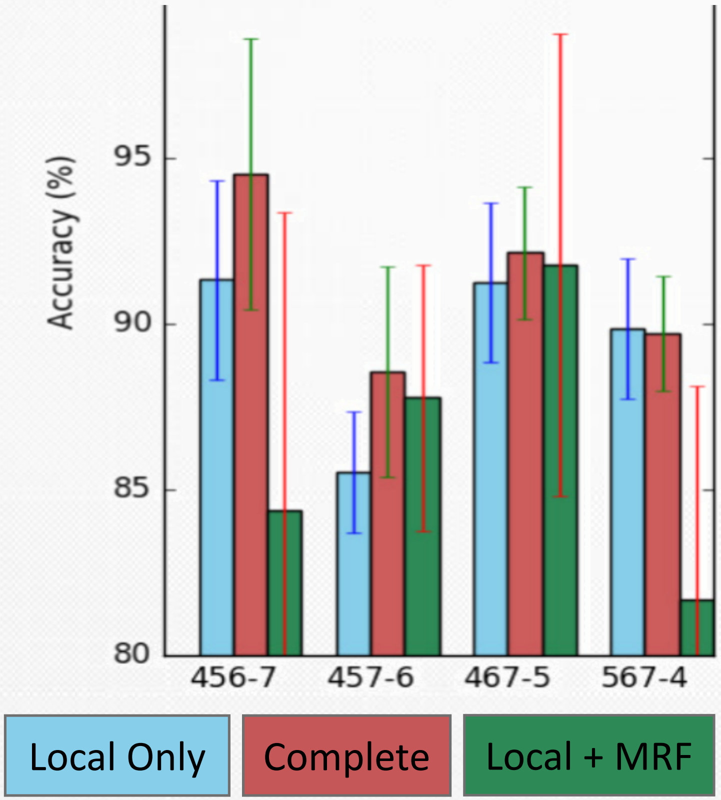
[Zheng, Pronobis '19]

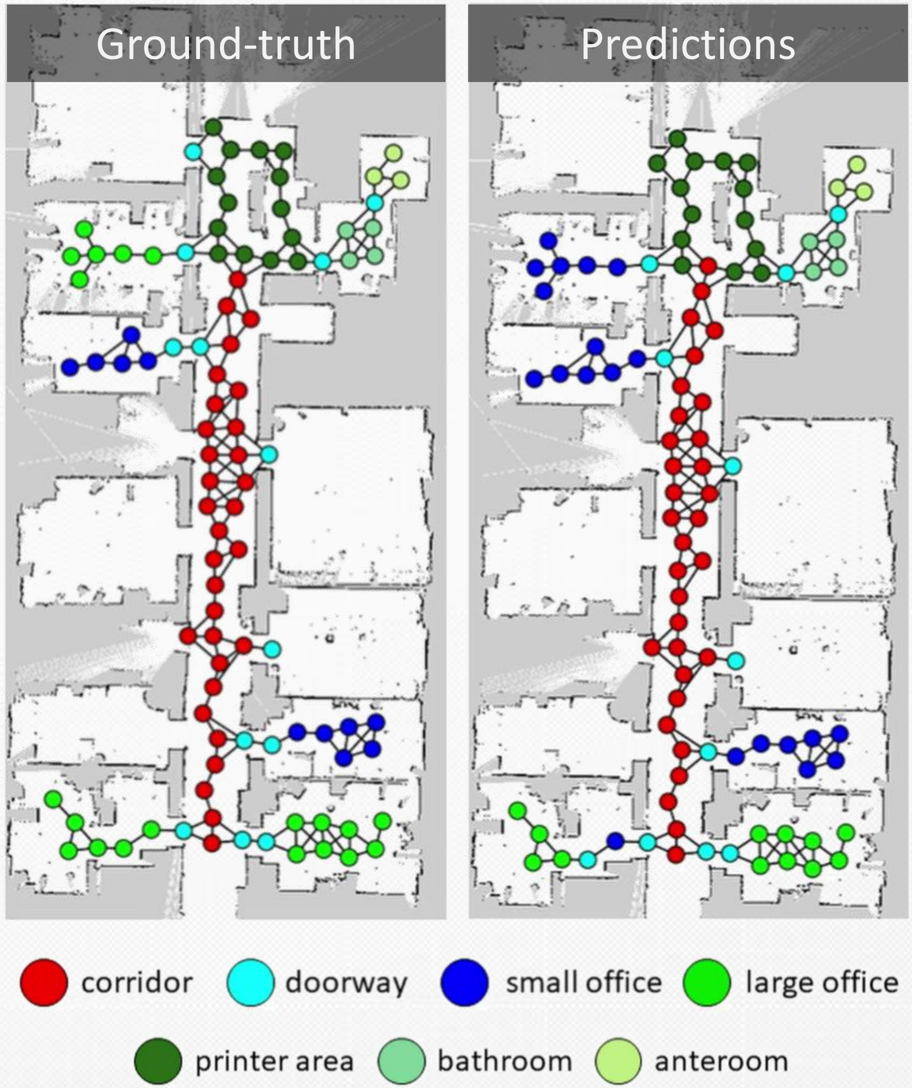


END2END: SEMANTIC MAPPING

[Zheng, Pronobis '19]

- Office building, 4 floors
  - 7 known room classes
- Overall: 93%

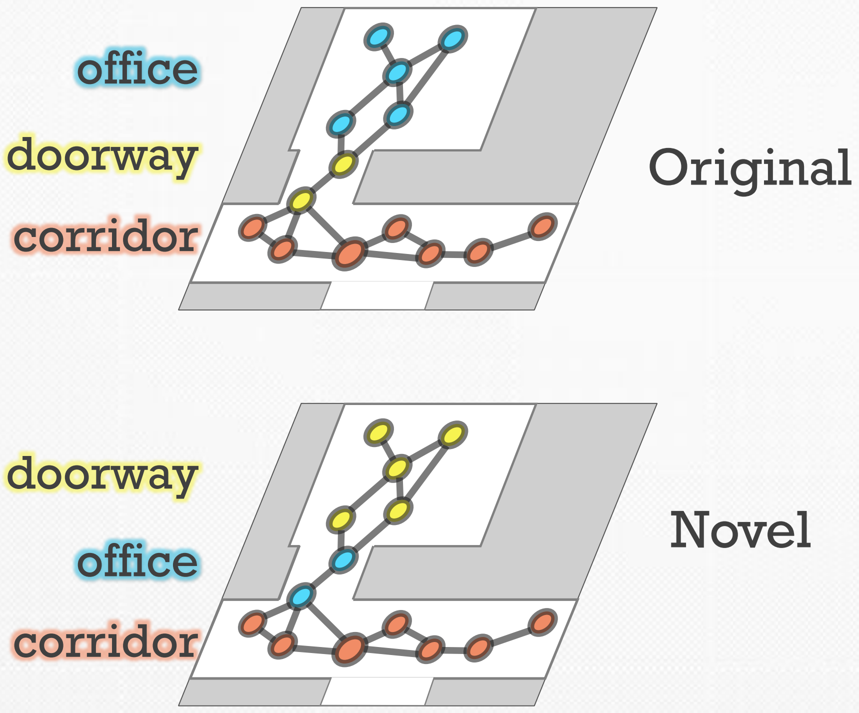


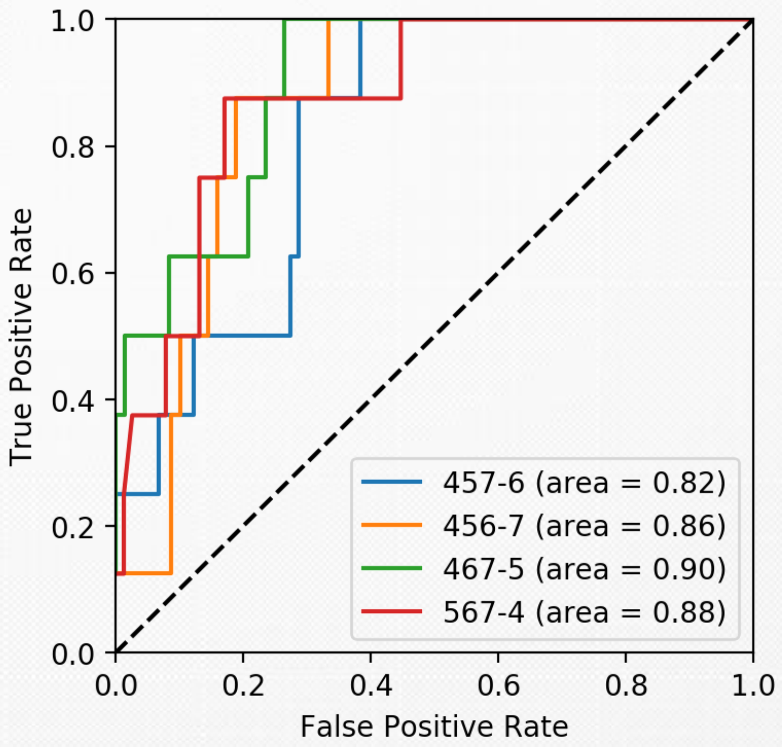


END2END: NOVEL GLOBAL STRUCTURE DETECTION

[Zheng, Pronobis '19]

- *P(semantics, geometry) > threshold* ?
- Novel floor structures by swapping labels of predictions for two random rooms in test map

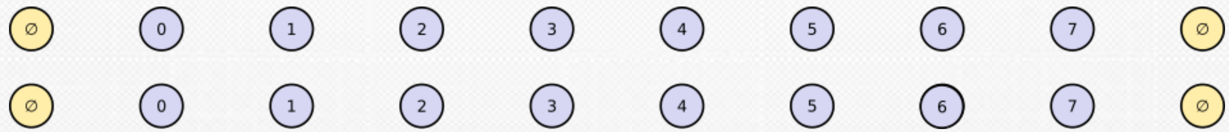




SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

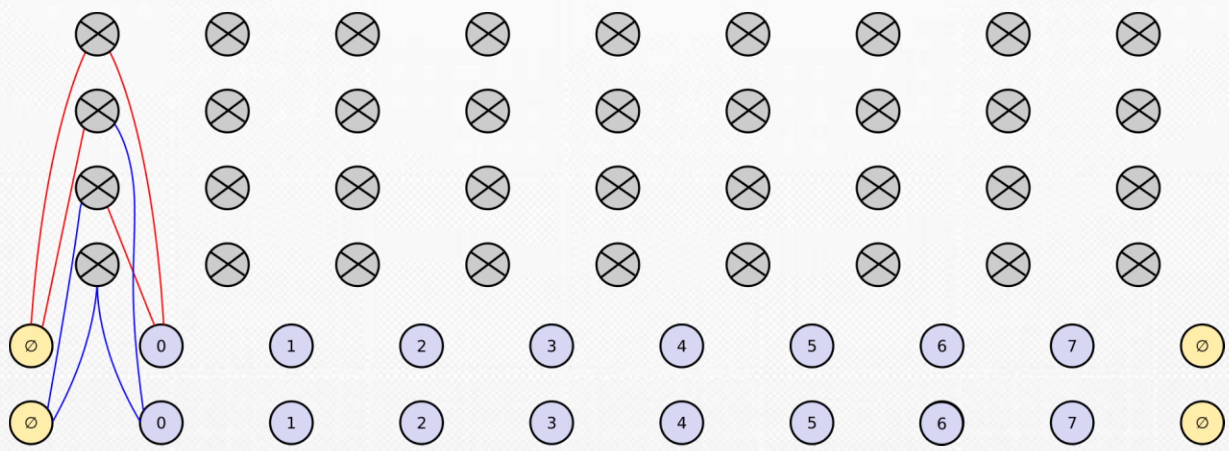
- New SPNs architecture
  - Resembles convolutional neural networks
  - For spatial signals, such as images
- 1D example:
  - Input layer of size 8 with 2 channels/pixel

SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis ‘19]



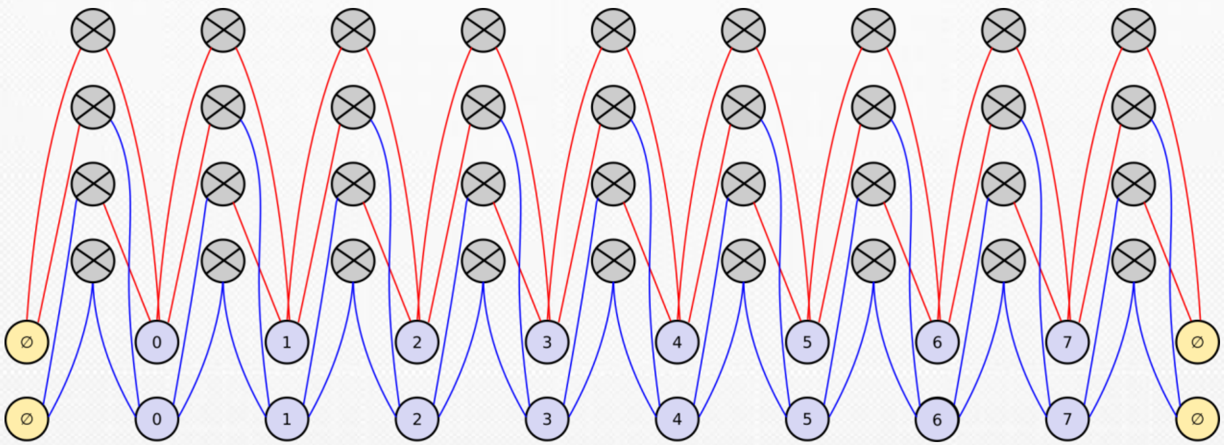
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Spatial products: stride 1, kernel size 2



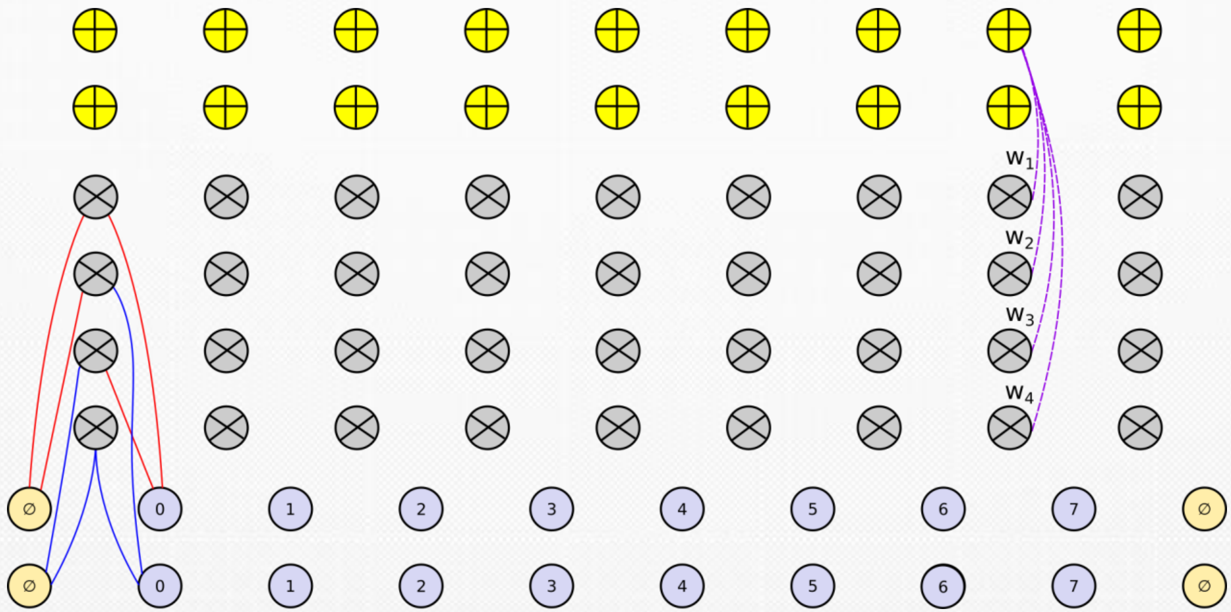
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Spatial products: stride 1, kernel size 2



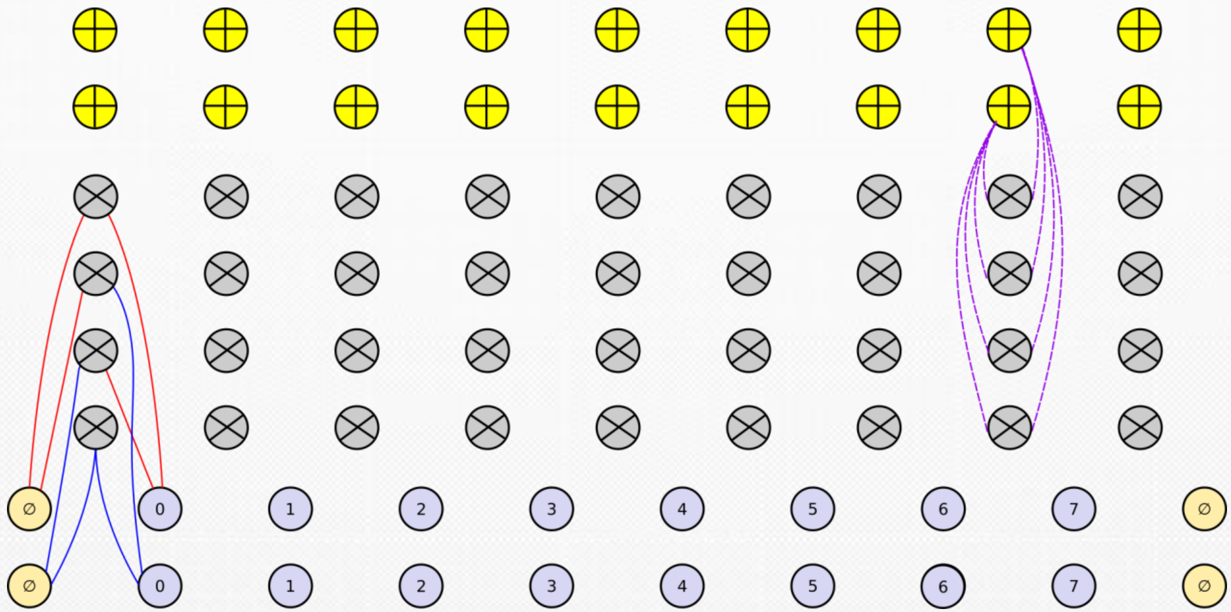
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Sums: 1x1 convolutions



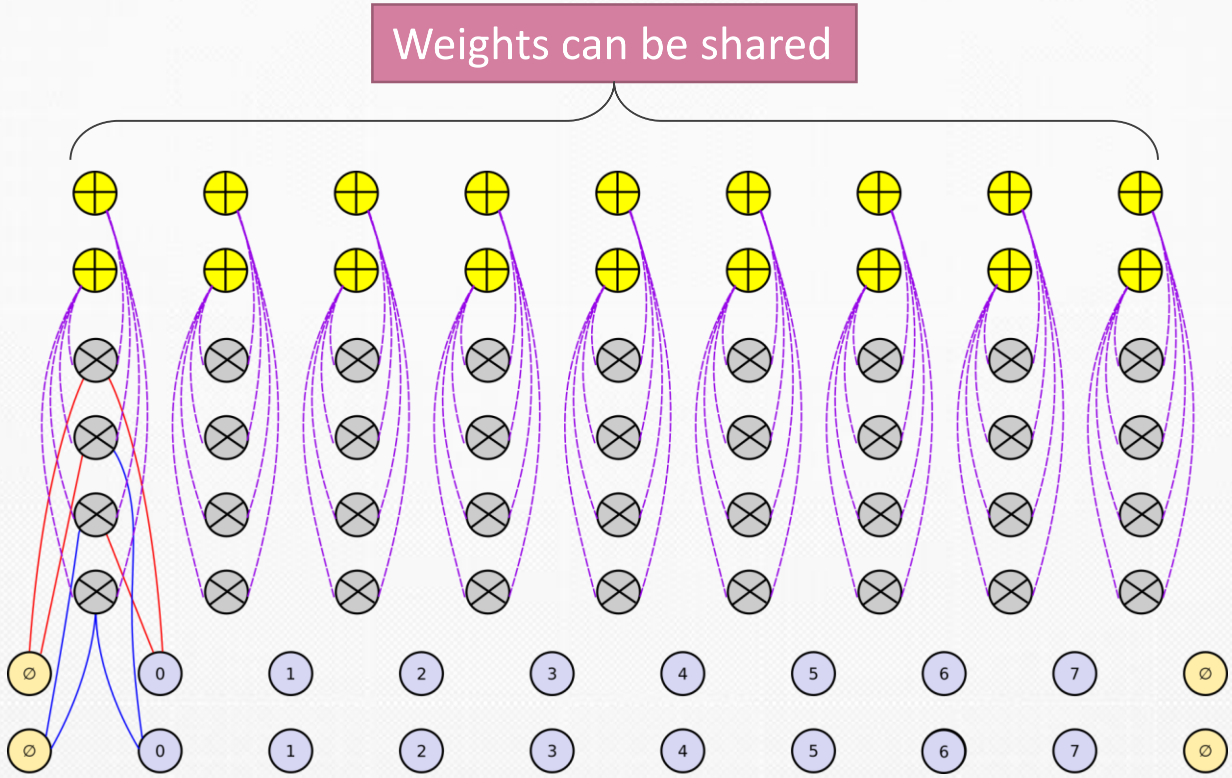
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Sums: 1x1 convolutions



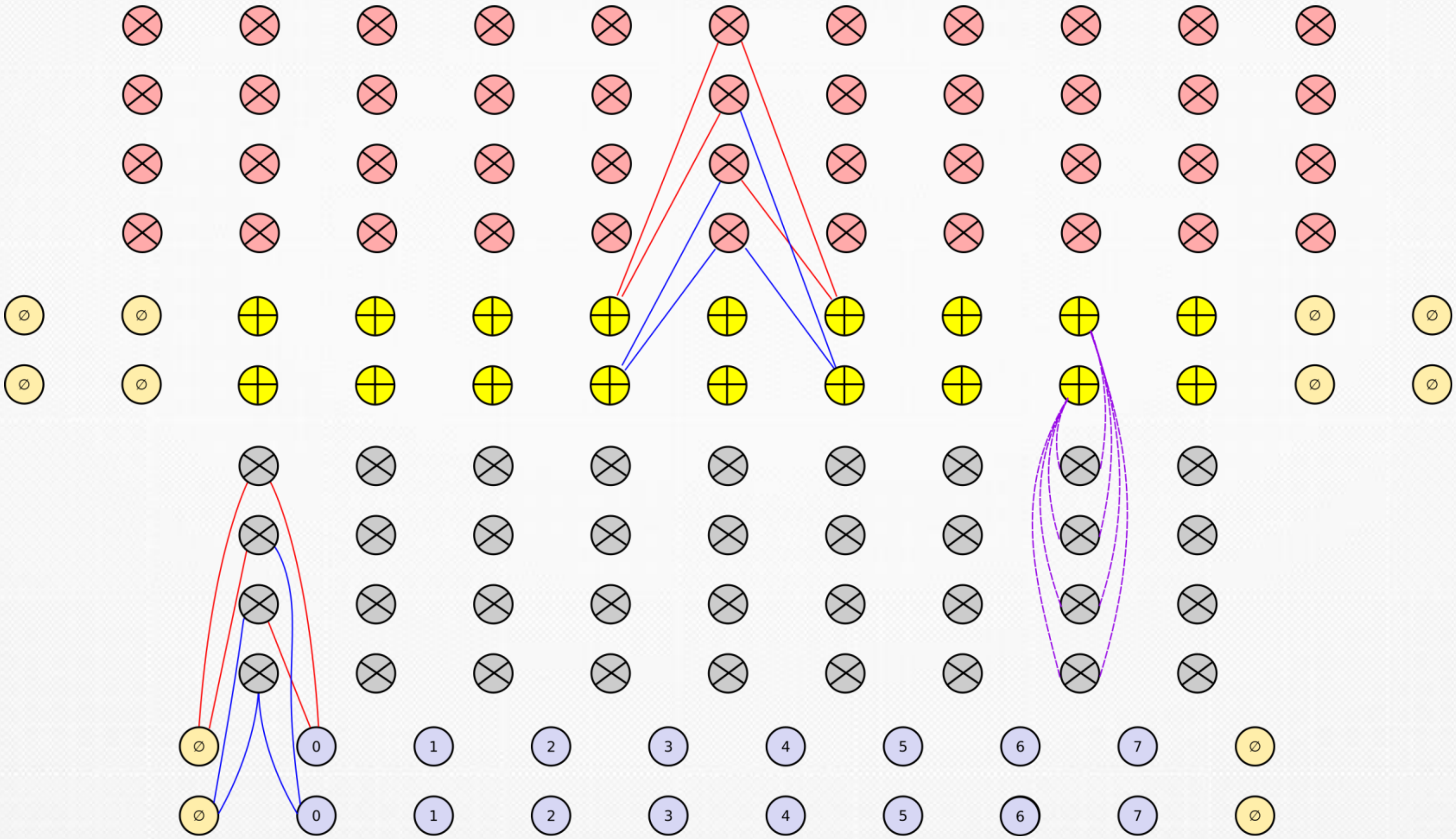
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Sums: 1x1 convolutions



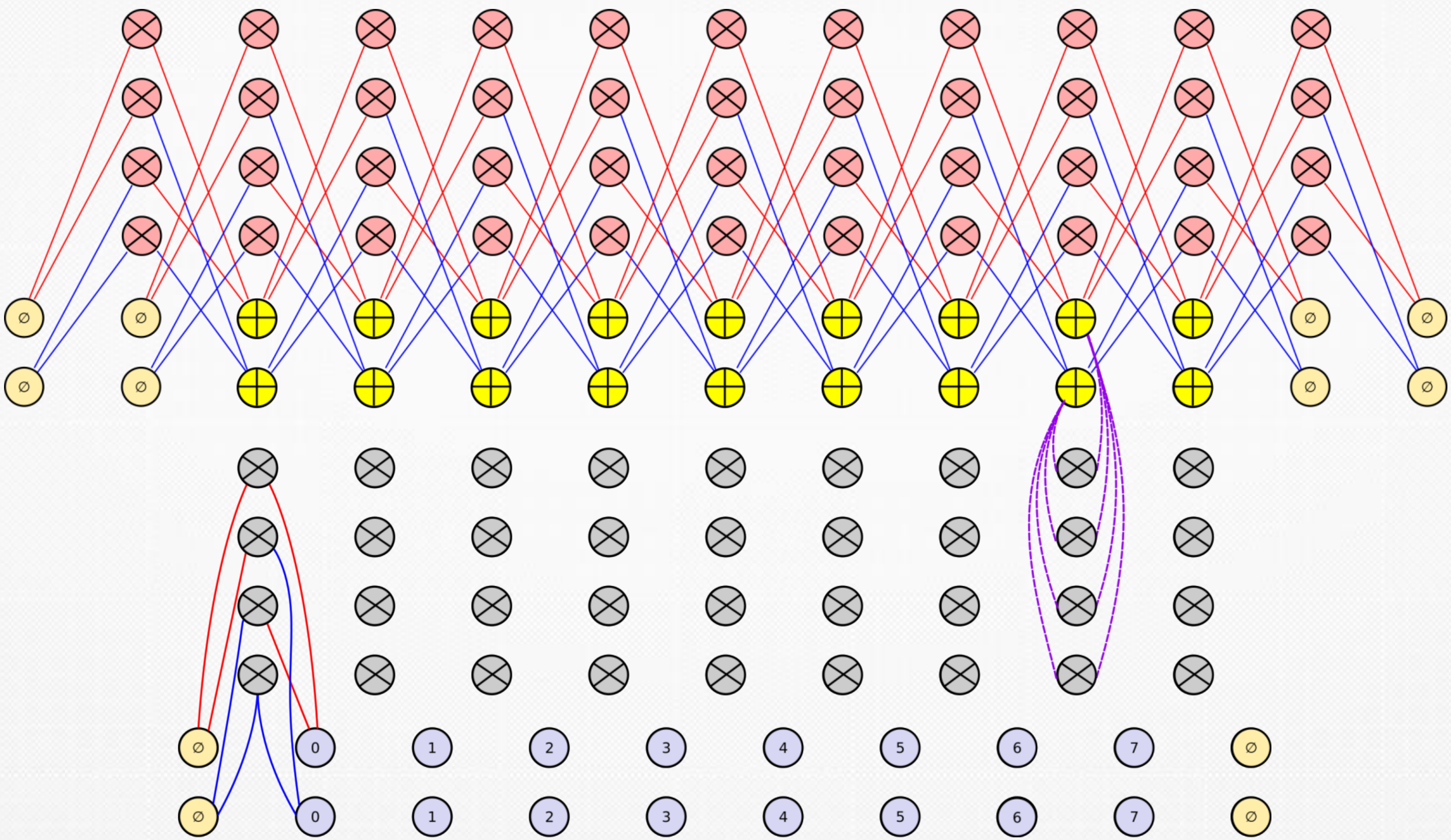
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Spatial products: stride 1, dilation rate **2**



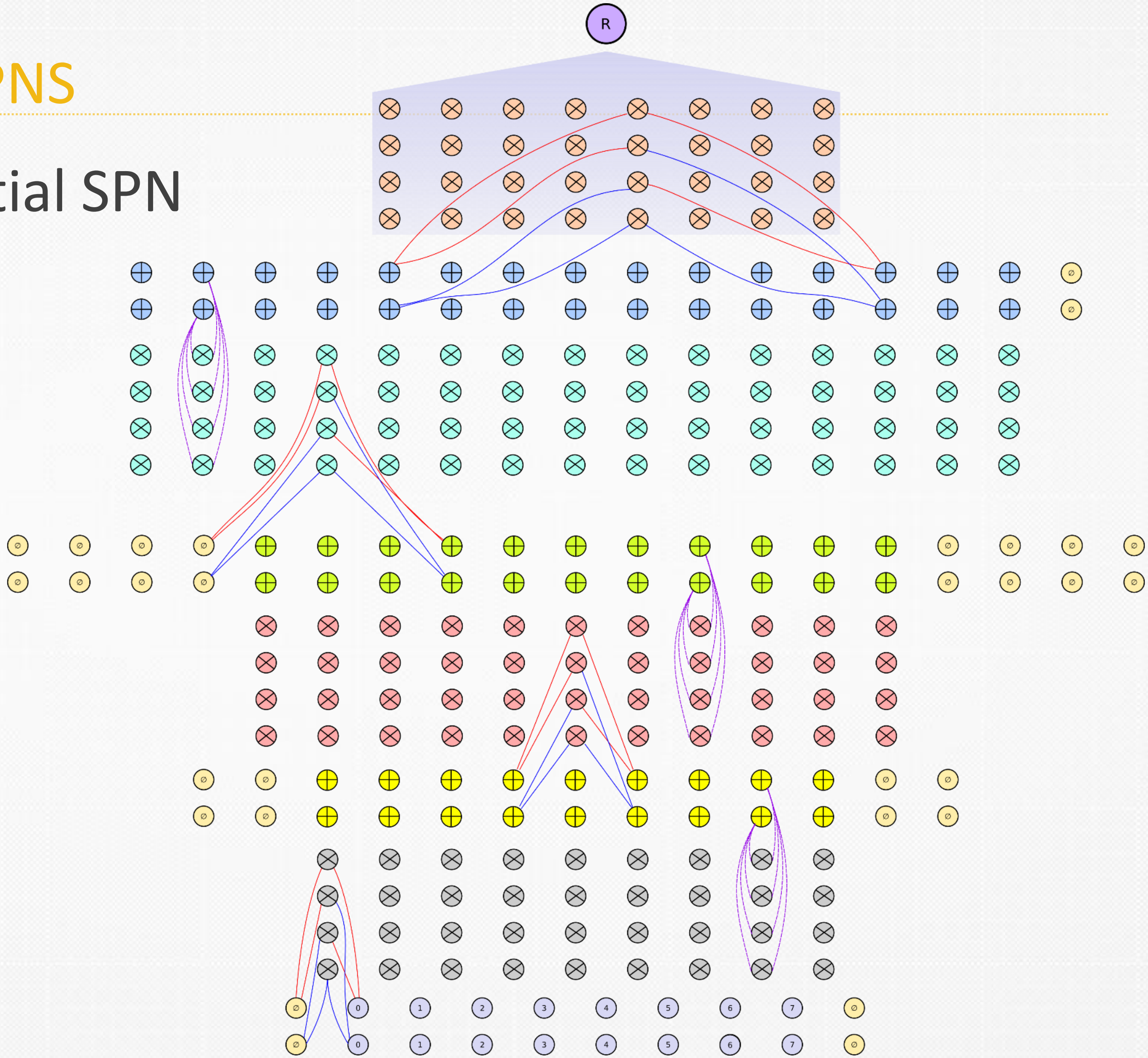
SPATIAL SPNS FOR VISUAL SIGNALS [van de Wolfshaar, Pronobis '19]

- Spatial products: stride 1, dilation rate **2**



SPATIAL SPNS

- Full Spatial SPN



SPATIAL SPNS: INITIAL RESULTS

[van de Wolfshaar, Pronobis '19]

- Dataset: MNIST

Model	Accuracy
RAT-SPN + **discriminative** GD [Peharz et al. 2018]	98.1 %
Vertical/horizontal splits + **discriminative** EBW [Rashwan et al. 2018]	95.0 %
Spatial SPN + **discriminative** GD	$\boxed{98.9\ \%}$
Spatial SPN + **generative** hard EM	$\boxed{96.1\ \%}$

CONCLUSIONS

- Comprehensive spatial representation and its realization using deep probabilistic networks
  - Learns general relationships
    - Between place geometry and semantic descriptions
    - Between pixels of places to topology of buildings
  - Up & down inferences across levels of abstraction
- Pioneers SPNs in the domain of robotics
  - SOTA model for complex structured prediction
  - Even in high-dimensional visual data
- Designed to support planning and communication
- Ongoing work
  - Probabilistic planning in the same architecture
  - Multi-modal representation: visual and depth sensors

THANK YOU

www.pronobis.pro