From Vision to Robotics

Saurabh Gupta

Cambrian Explosion (541 million years ago)

Sea Squirts

We must perceive in order to move, but we must also move in order to perceive.

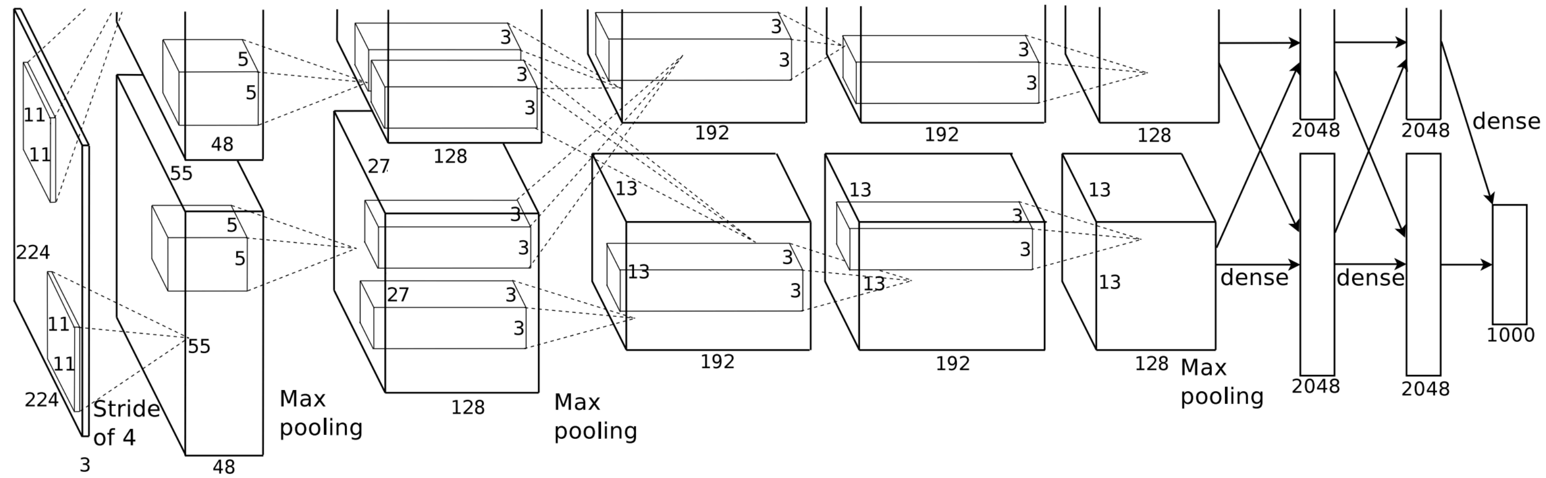
— JJ Gibson

Vision, like other sensory functions, has its evolutionary rationale rooted in improved motor control. Although organisms can of course see when motionless or paralyzed, the visual system of the brain has the organization, computational profile, and architecture it has in order to facilitate the organism's thriving at feeding fleeing, fighting, and reproduction.

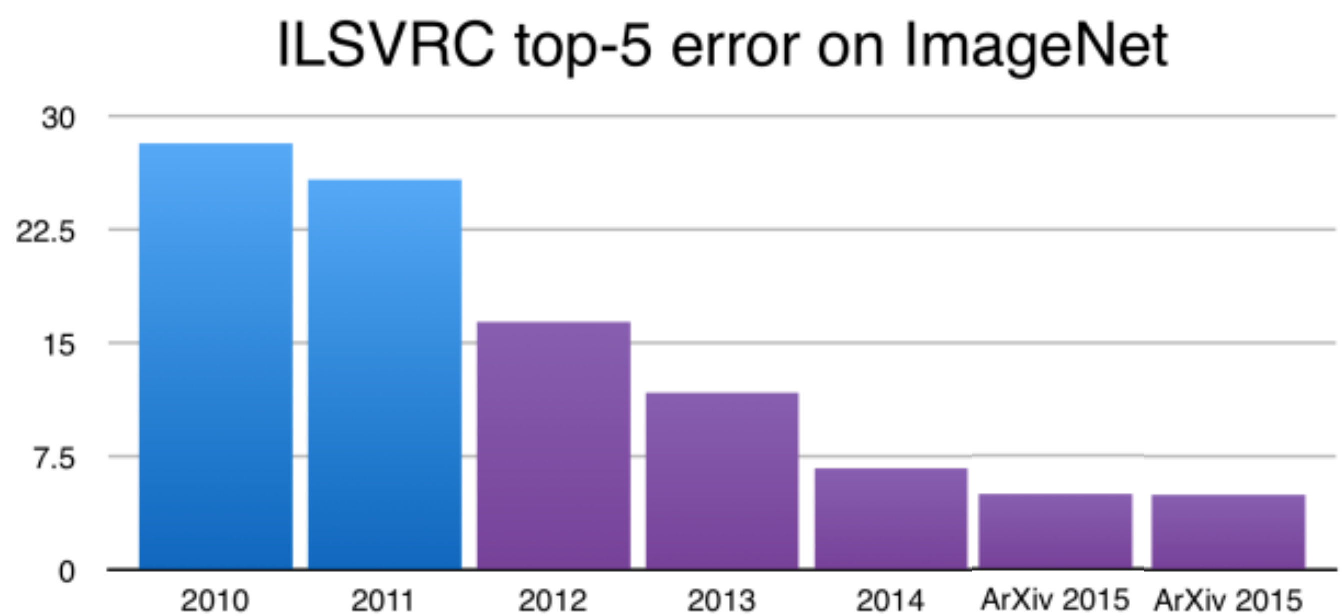
— Churchland, Ramachandran and Sejnowski
A critique of pure vision

Factors Leading to Success in Computer Vision

Large-scale labeled data



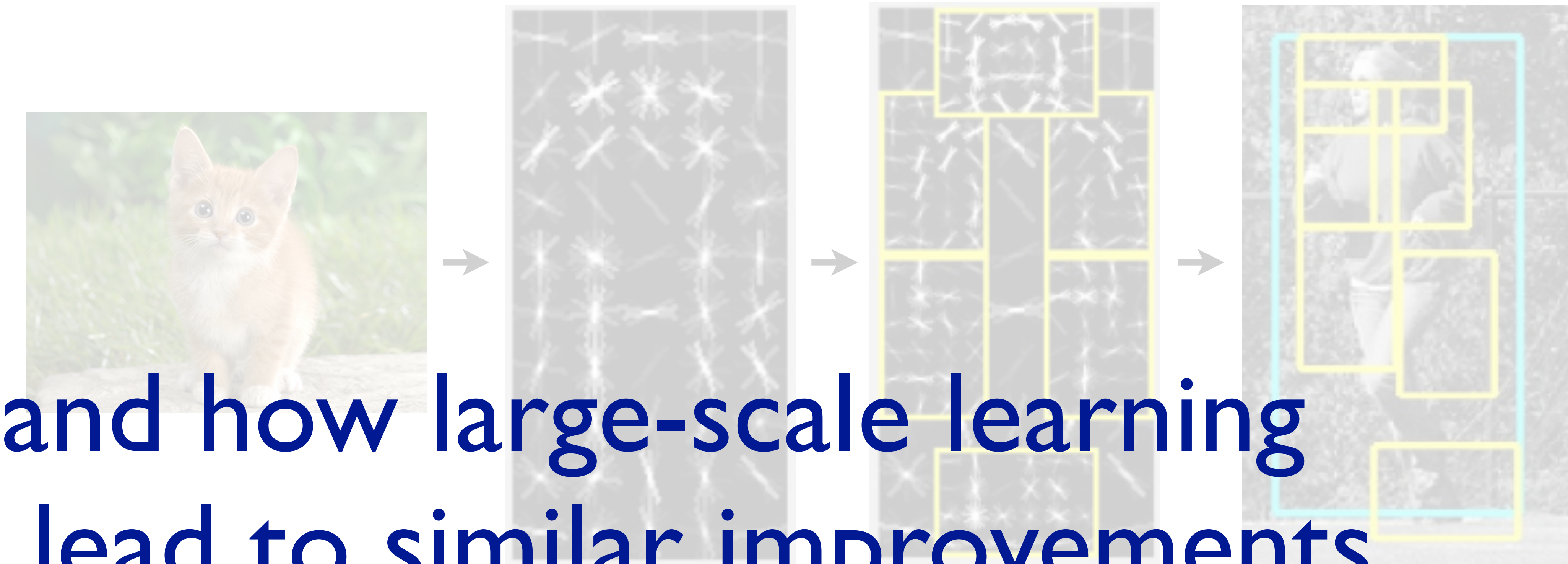
Hand-crafted features
to
End-to-end trained features



A. Krizhevsky et al. **ImageNet Classification with Deep Convolutional Neural Networks.** NIPS 2012
J. Deng et al. **ImageNet: A Large-Scale Hierarchical Image Database.** CVPR 2009

Factors Leading to Success in Computer Vision

Hand-crafted features



features (e.g. HOG)

mid-level features (e.g. DPM)

Felzenszwalb et al.

classifier (e.g. SVM)

If and how large-scale learning can lead to similar improvements in robotics?

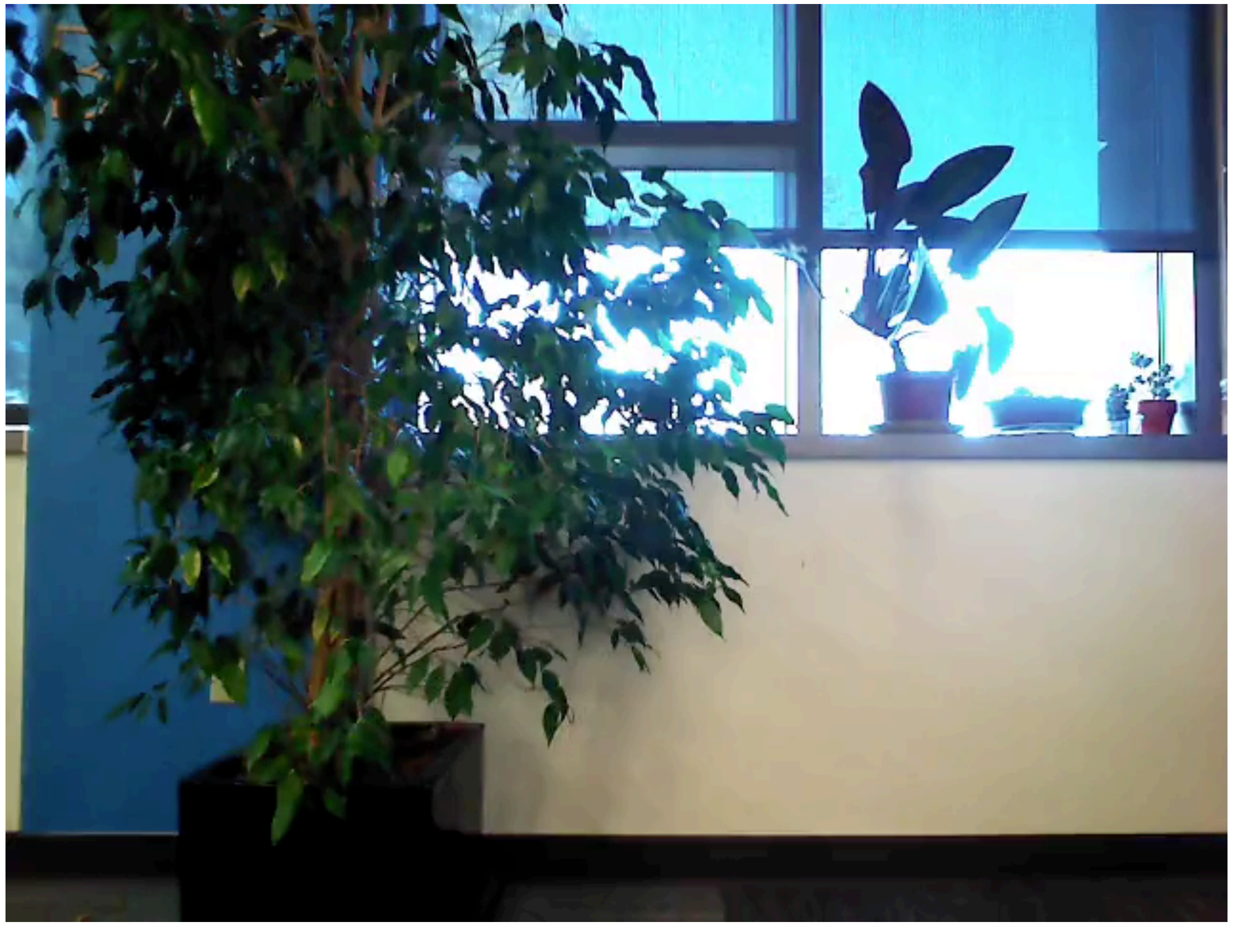
End-to-end trained features

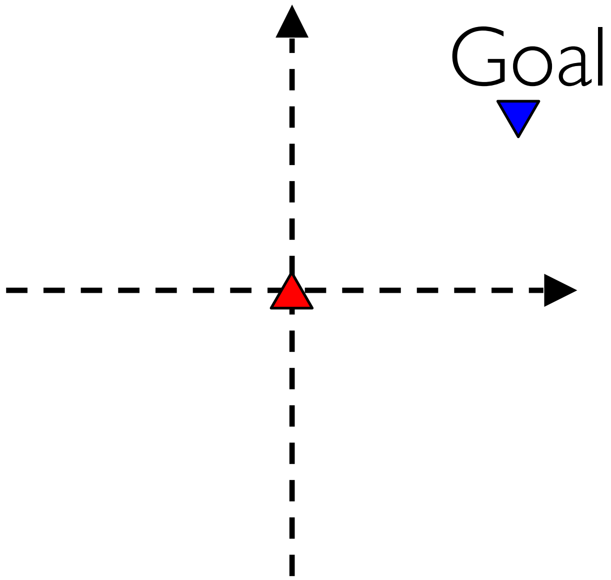
cat

A. Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

Robotic Tasks

Navigation





"Go
300 feet North,
400 feet East"

"Go Find a Chair"

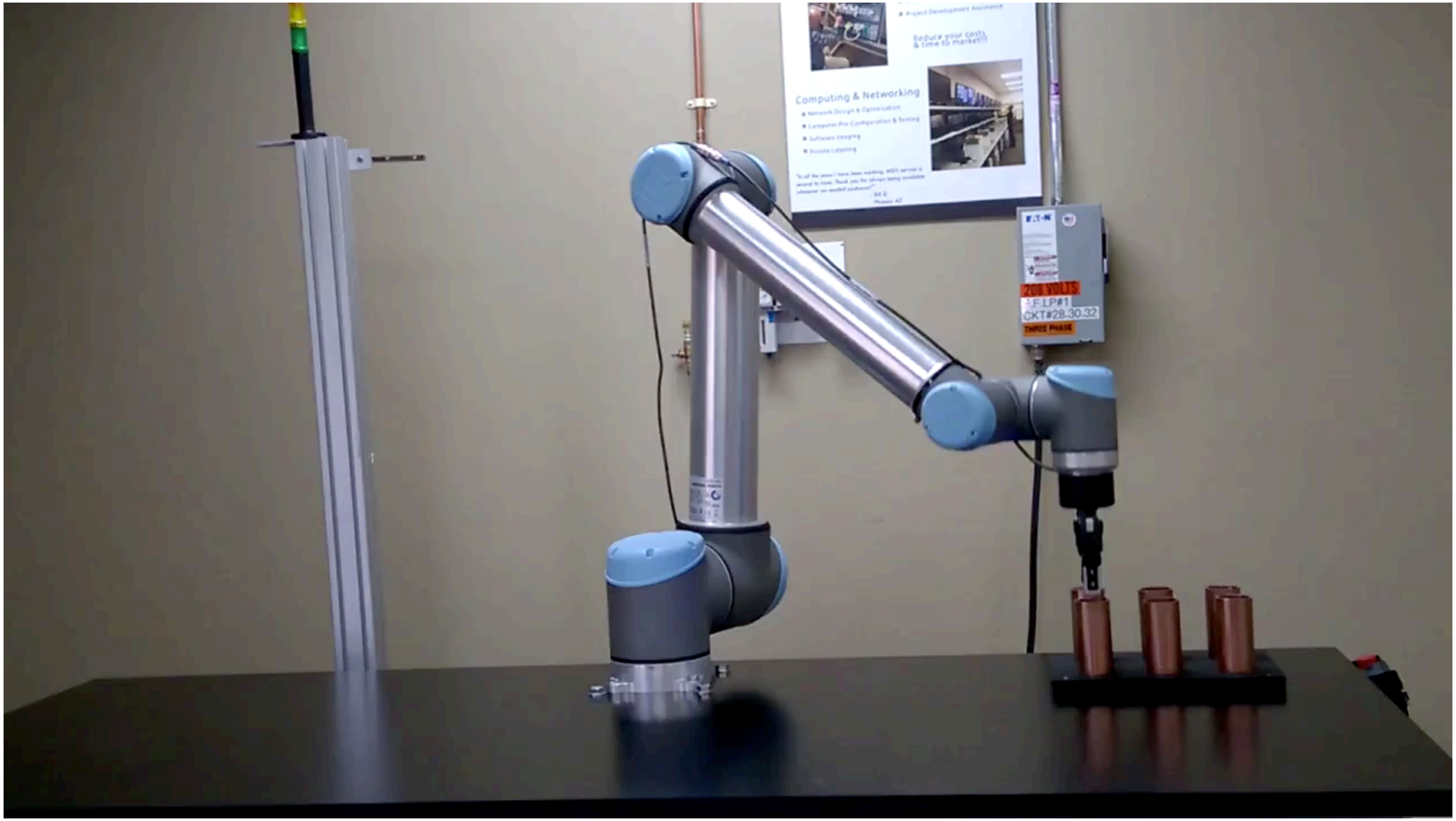
Robot with a first person camera

Dropped into a novel environment

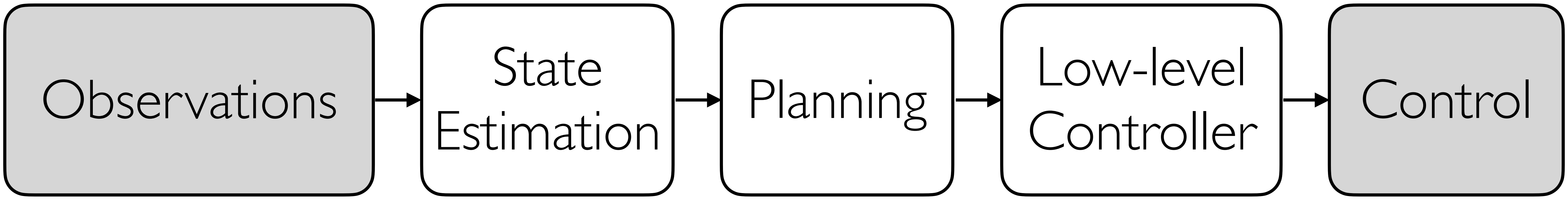
Navigate around

Robotic Tasks

Manipulation



Typical Classical Robotics Pipeline



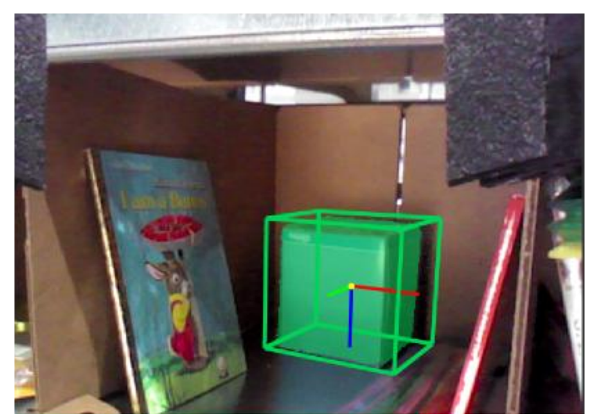
Slide adapted from S. Levine.

Typical Classical Robotics Pipeline

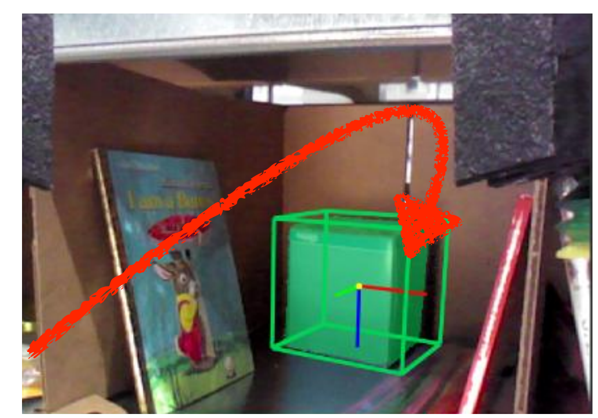
Observations → State Estimation → Planning → Low-level Controller → Control

Manipulation

Observed Images



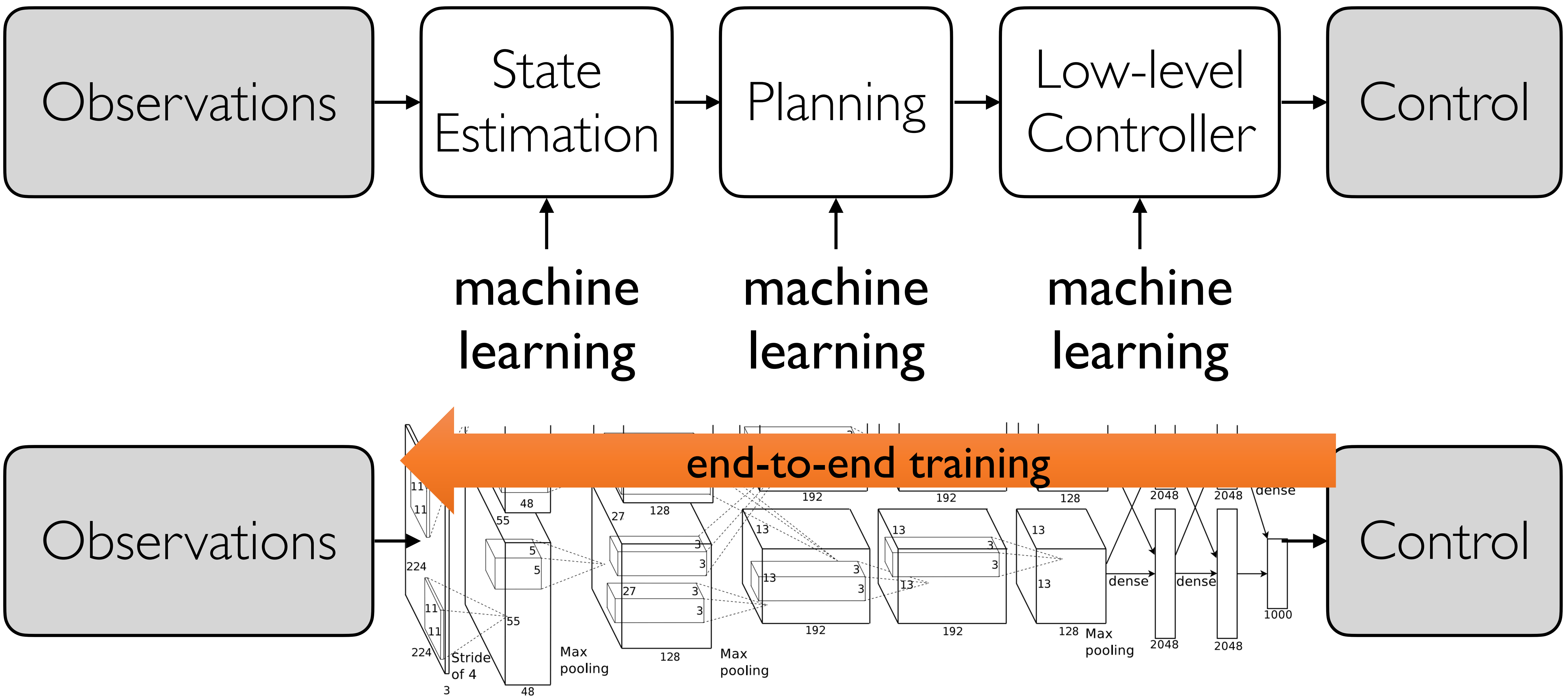
6DOF Pose



Grasp Motion Planning

Slide adapted from S. Levine.

Typical Classical Robotics Pipeline



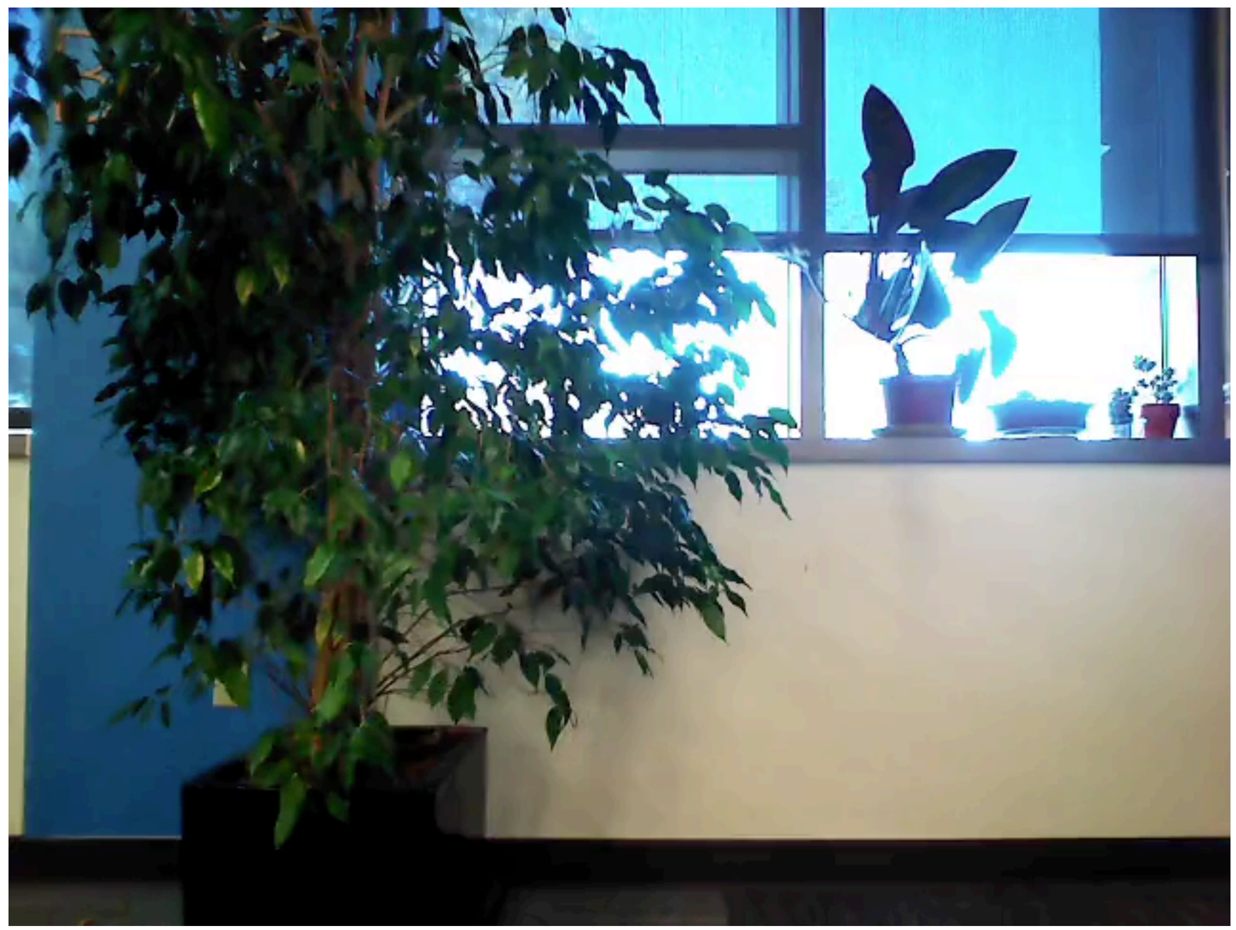
Should it help?

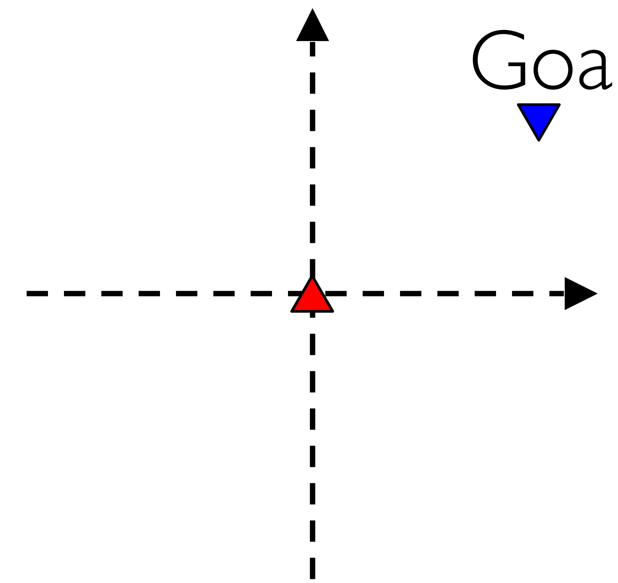
If so, how to do it?

Does it help?

Slide adapted from S. Levine.

Robot Navigation





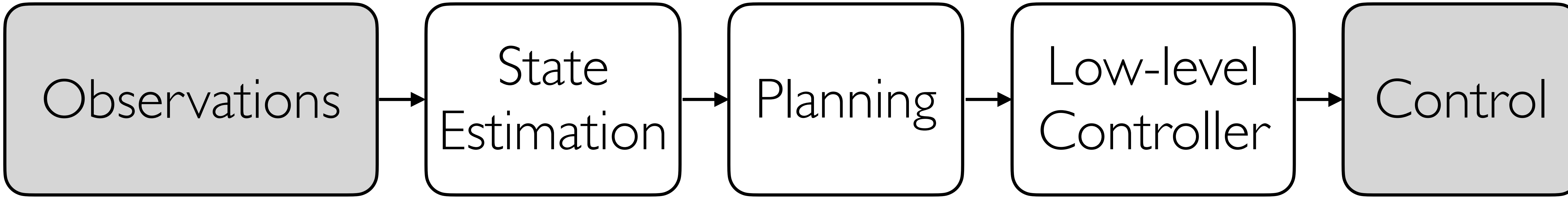
“Go
300 feet North,
400 feet East”

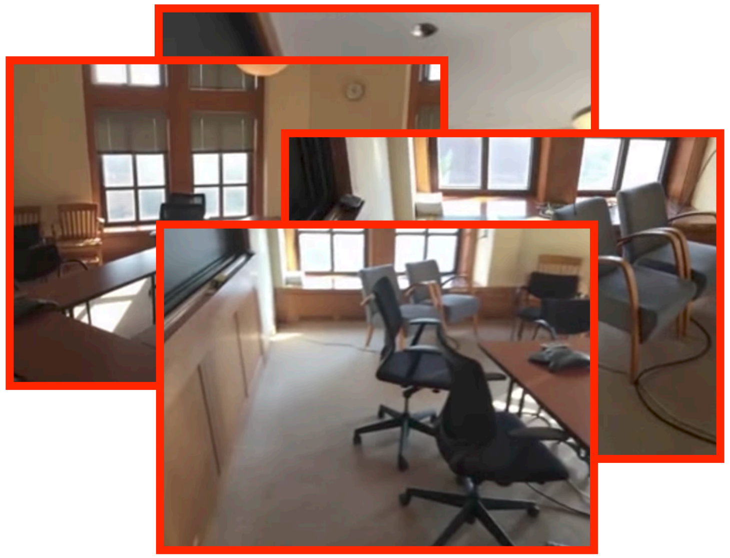
“Go Find a Chair”

Robot with a first person camera

Dropped into a novel environment

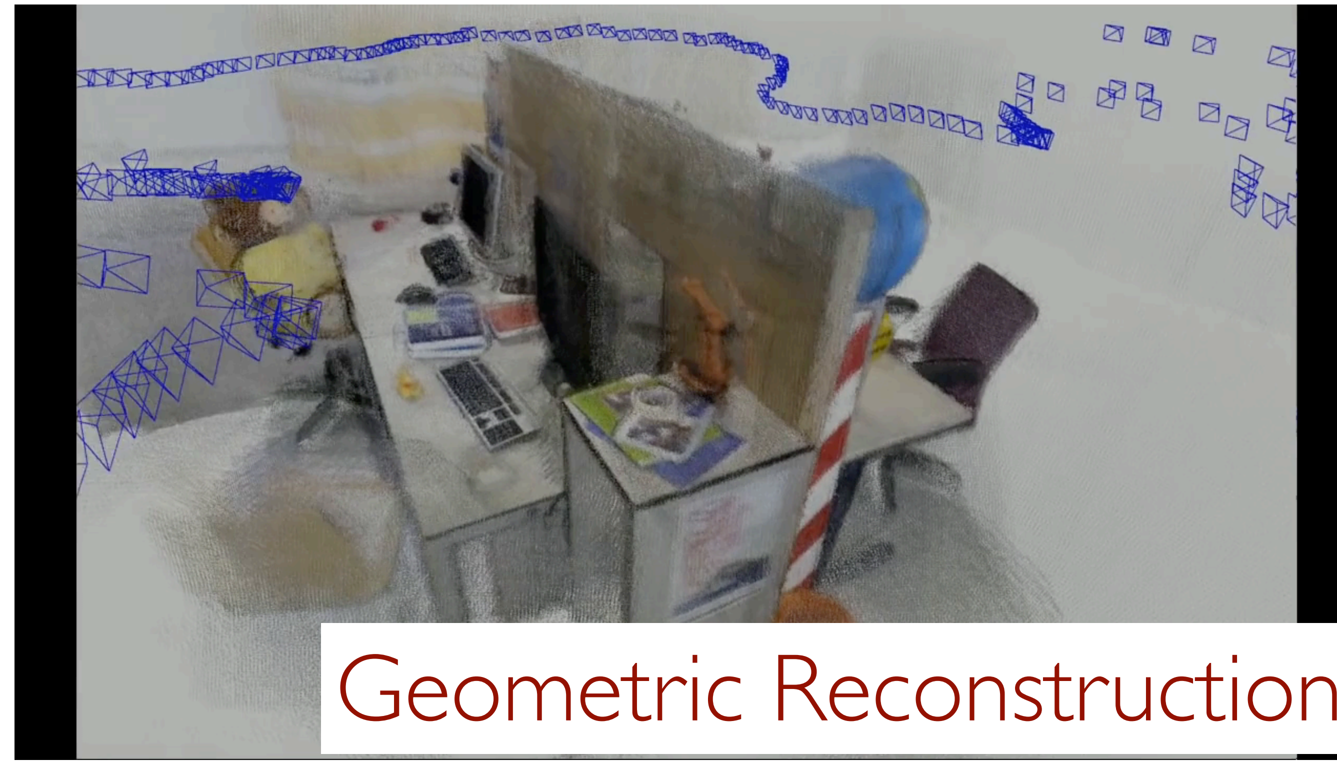
Navigate around



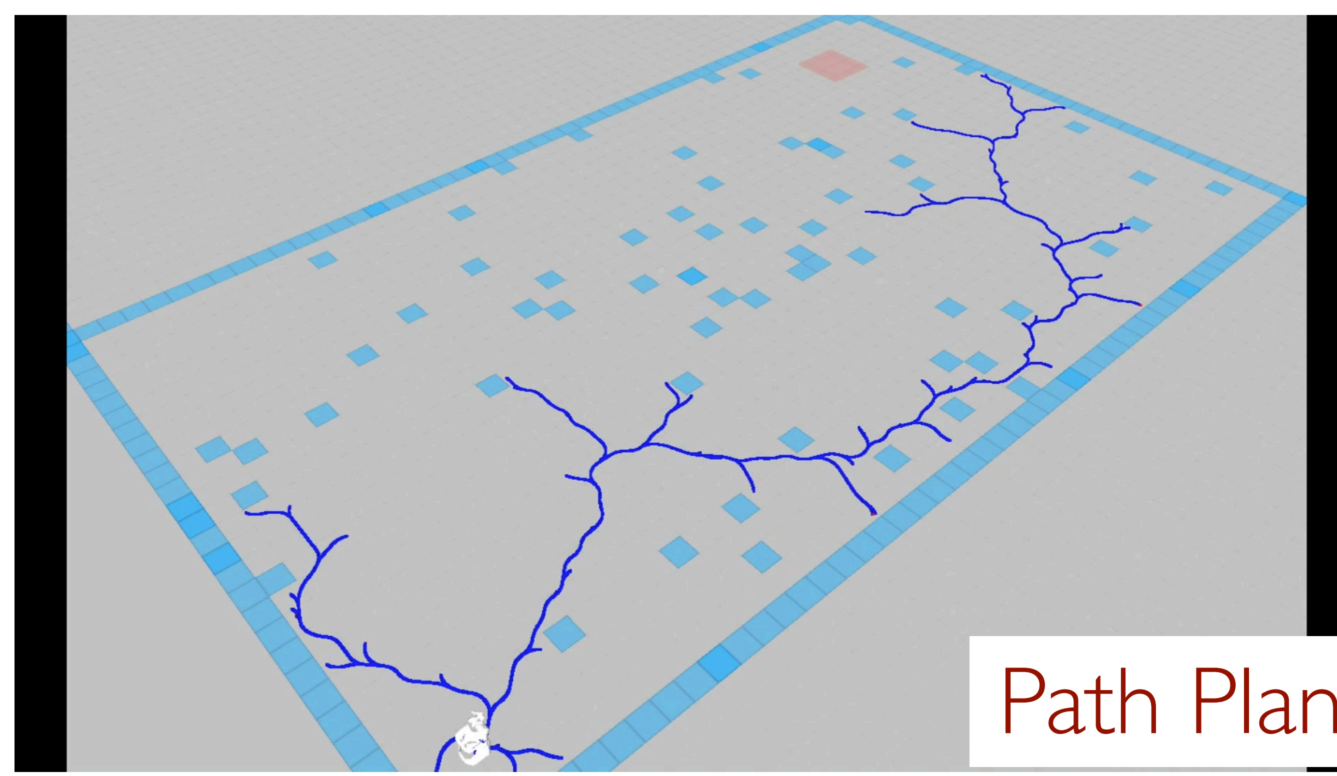


Observed Images

Mapping



Planning



Hartley and Zisserman. 2000. Multiple View Geometry in Computer Vision
Thrun, Burgard, Fox. 2005. Probabilistic Robotics

Canny. 1988. The complexity of robot motion planning.
Kavraki et al. RA1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces.
Lavalle and Kuffner. 2000. Rapidly-exploring random trees: Progress and prospects.

Video Credits: Mur-Artal et al., Palmieri et al.

Geometric 3D Reconstruction of the World

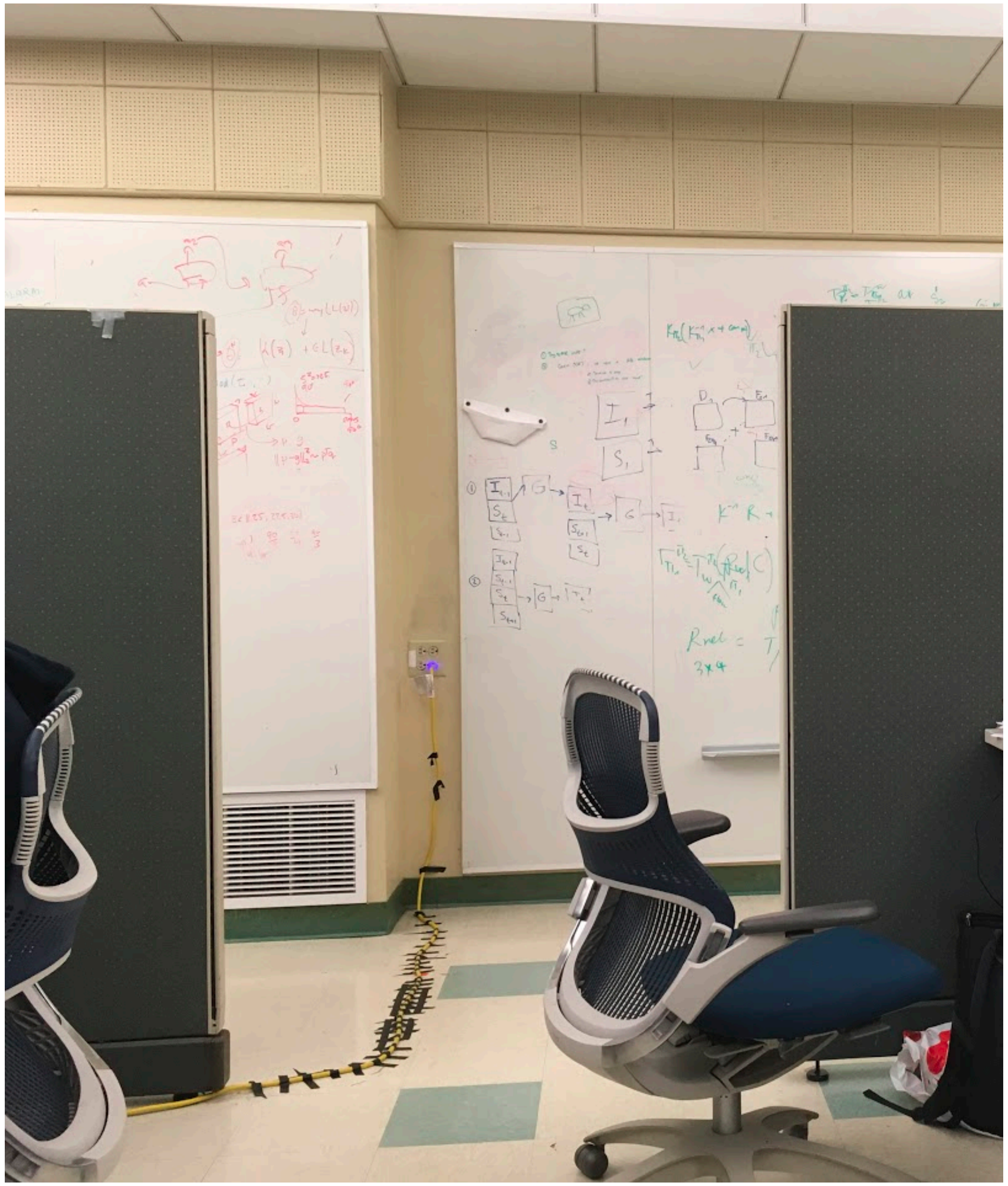
Unnecessary

Do we need to tediously reconstruct everything on this table?

Video Credit: Mur-Artal and Tardos, TRobotics 2016. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras.

Geometric 3D Reconstruction of the World

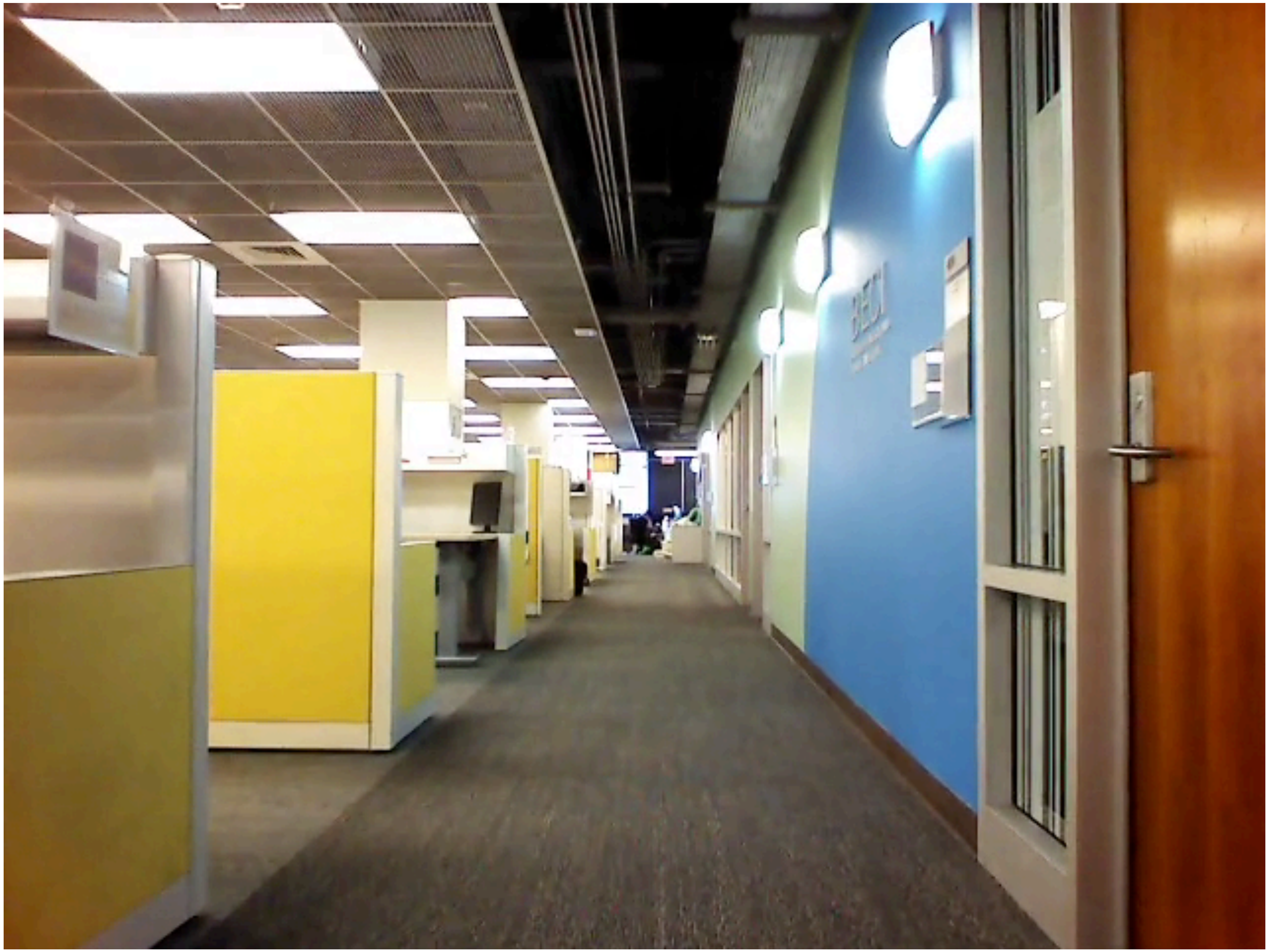
Insufficient



Can’t speculate about space not directly observed.

Geometric 3D Reconstruction of the World

Insufficient



Can't exploit patterns in layout of indoor spaces.

Geometric 3D Reconstruction of the World

Insufficient

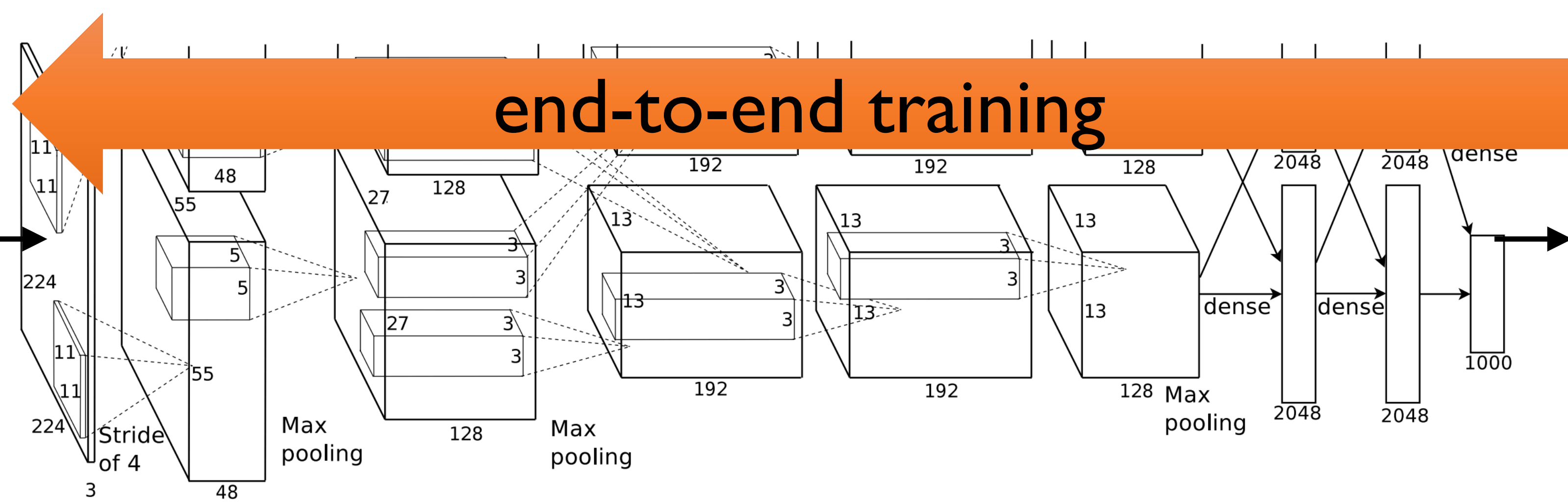
Ignore navigational affordances.

Typical Classical Robotics Pipeline

Observations → State Estimation → Planning → Low-level Controller → Control

machine learning | machine learning | machine learning

Observations →



→ Control

Should it help?

If so, how to do it?

Does it help?

Slide adapted from S. Levine.

Agent Environment Interface

Reinforcement Learning

Image Source: https://www.freecodecamp.org/news/a-brief-introduction-to-reinforcement-learning-7799af5840db/

Markov Decision Process

o_t

Step Back

a_t

o_{t+1}

…

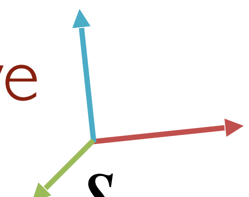
Transition Function

How you move,
how the tiger moves?

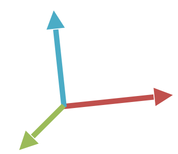
Reward Function

Survived?

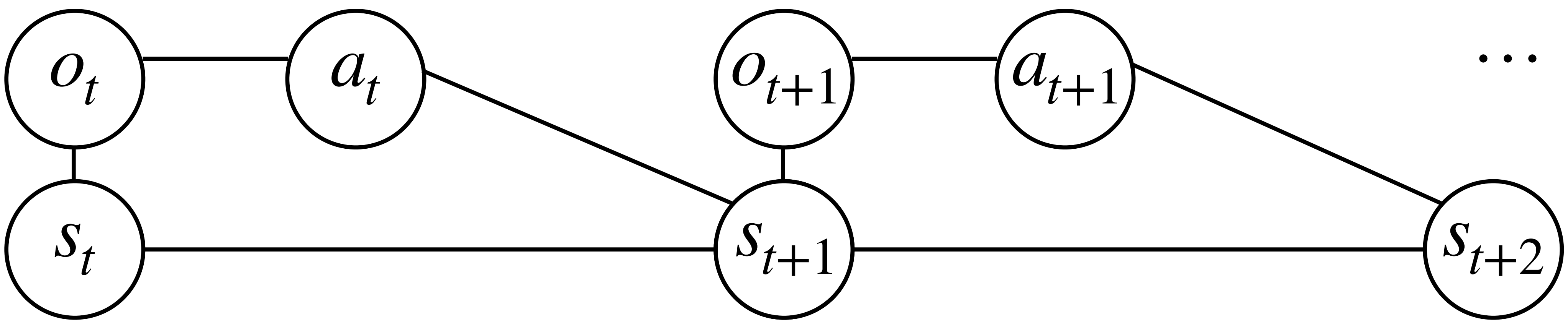
3D Relative Pose



s_t



s_{t+1}



One step dynamics $p(s_{t+1}, r_{t+1} \mid s_t, a_t)$

Transition Function $p(s_{t+1} \mid s_t, a_t)$ $p(s_{t+2} \mid s_{t+1}, a_{t+1})$

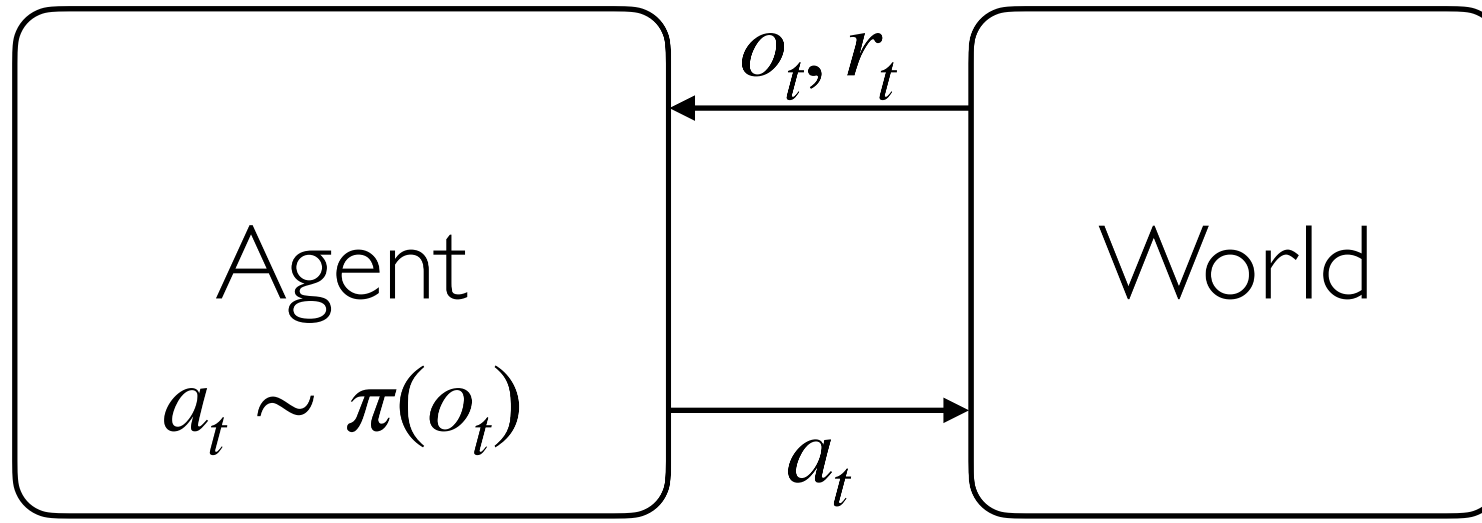
Reward Function $r_{t+1} = R(s_{t+1}, s_t, a_t)$ $r_{t+2} = R(s_{t+2}, s_{t+1}, a_{t+1})$

Goal $\mathrm{argmax}_{a_0,\ldots,a_T} \sum_t \gamma^t r_t$

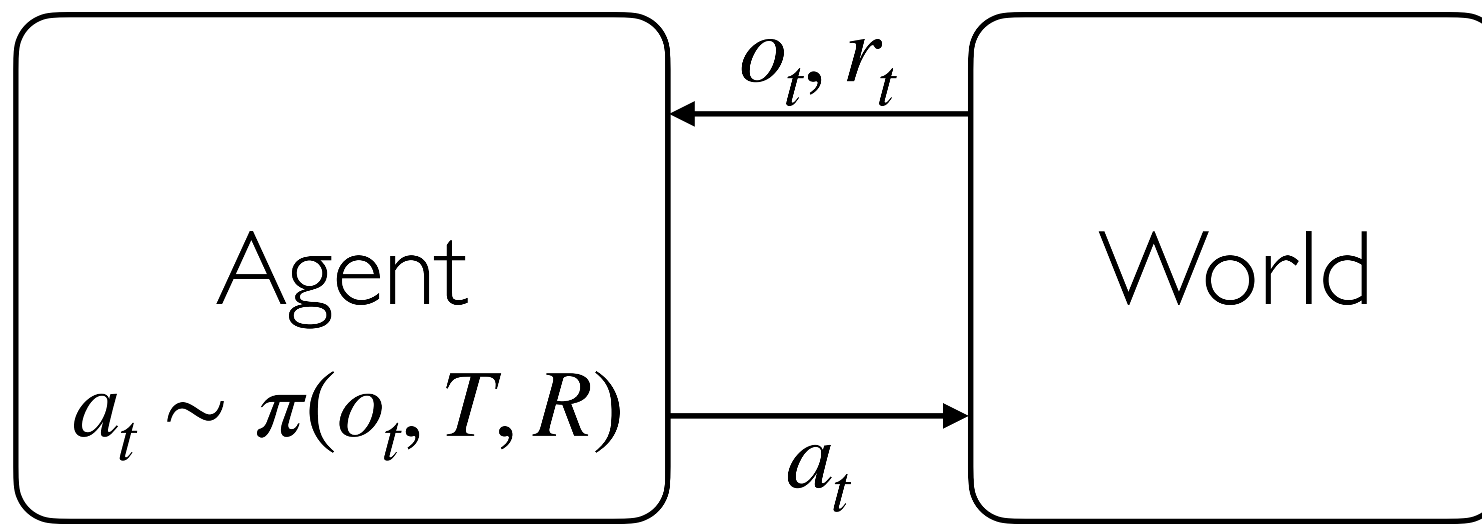
Solving MDPs

Policy: $a_t \sim \pi(o_t)$

Most General Case



More Specific Case



Fully Observed System $o_t = s_t$

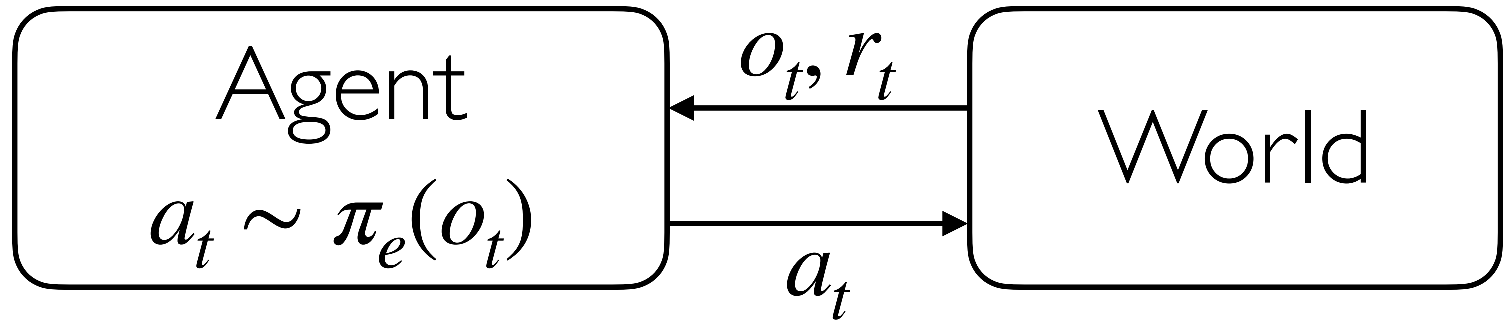
Known Transition Function $s_{t+1} \sim T(s_t, a_t)$

Known Reward Function $R(s_{t+1}, s_t, a_t)$

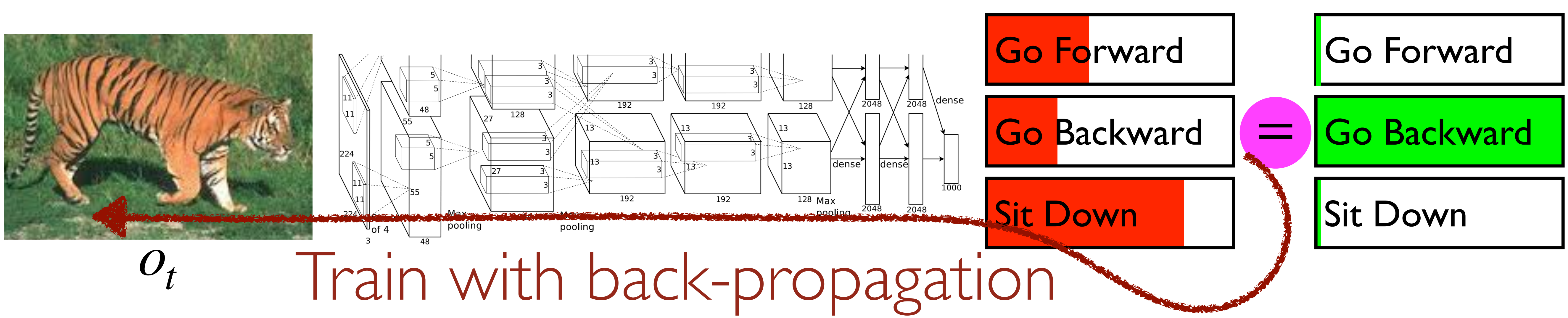
Behavior Cloning

Train Time

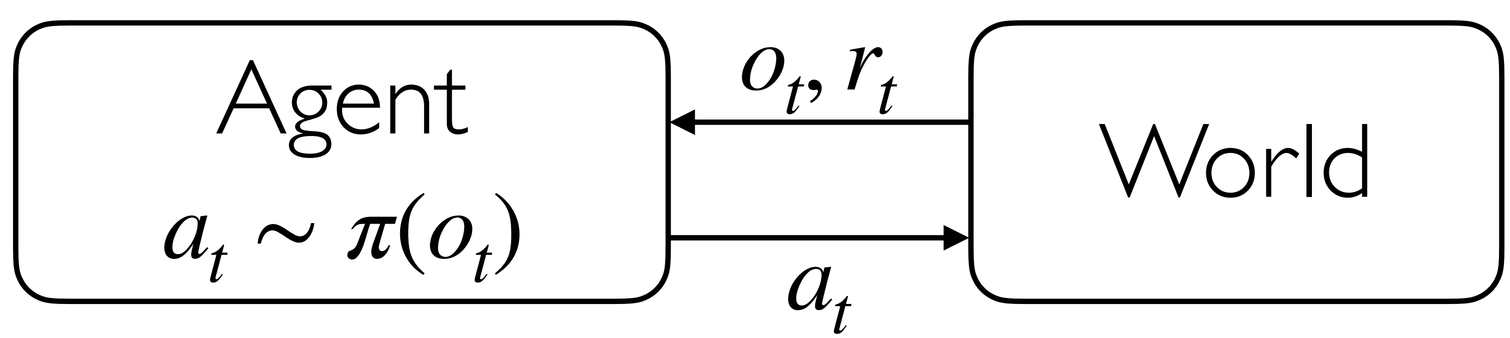
Assume an expert e can solve this MDP.



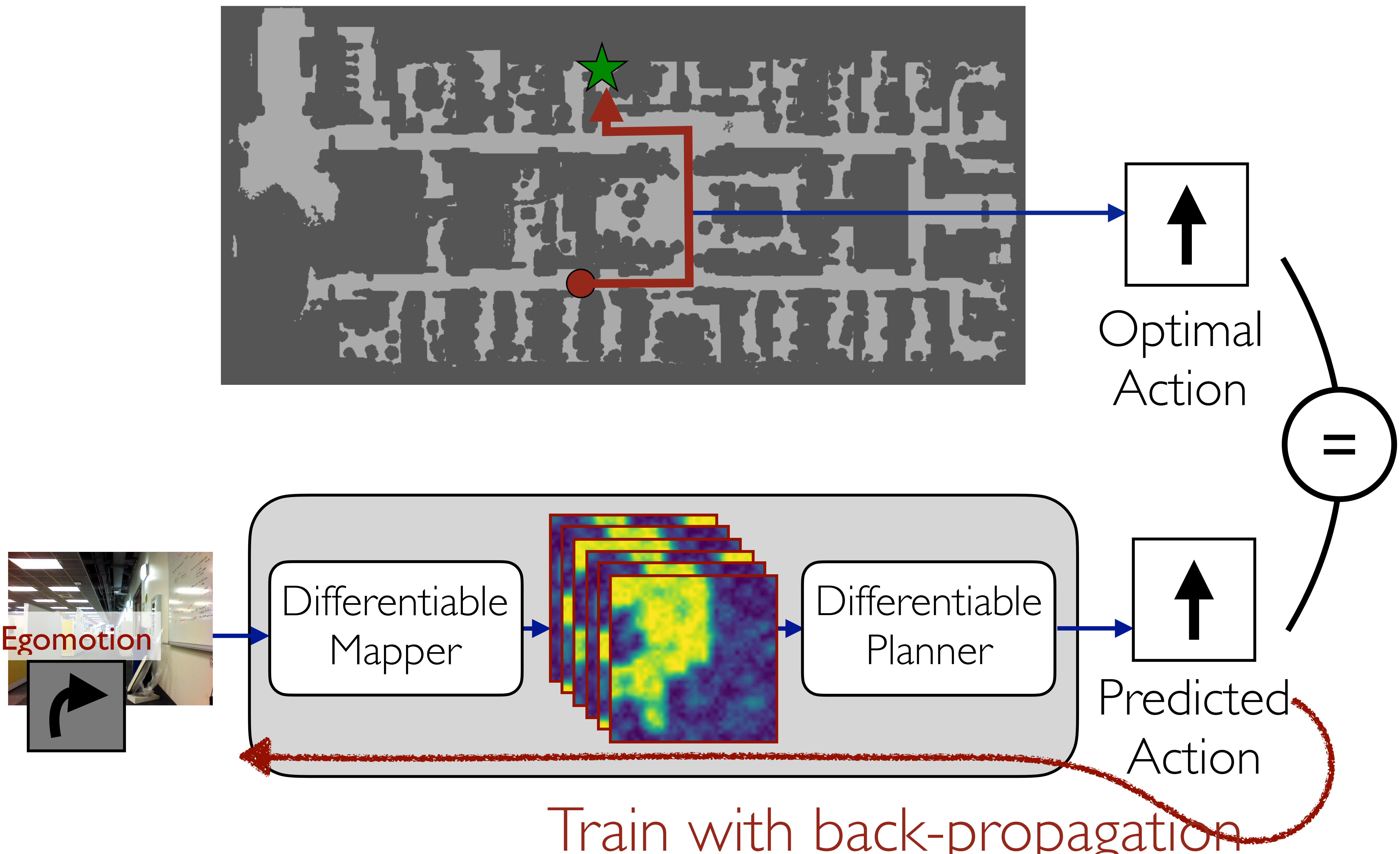
1. *Ask the expert e to solve this MDP.*
2. *Collect labeled dataset D from expert.*
3. *Train a function* $\pi(o_t)$ *that mimics* $\pi_e(o_t)$ *on D.* $\pi(o_t)$ $\pi_e(o_t)$



Test Time



Supervision from an Algorithmic Expert

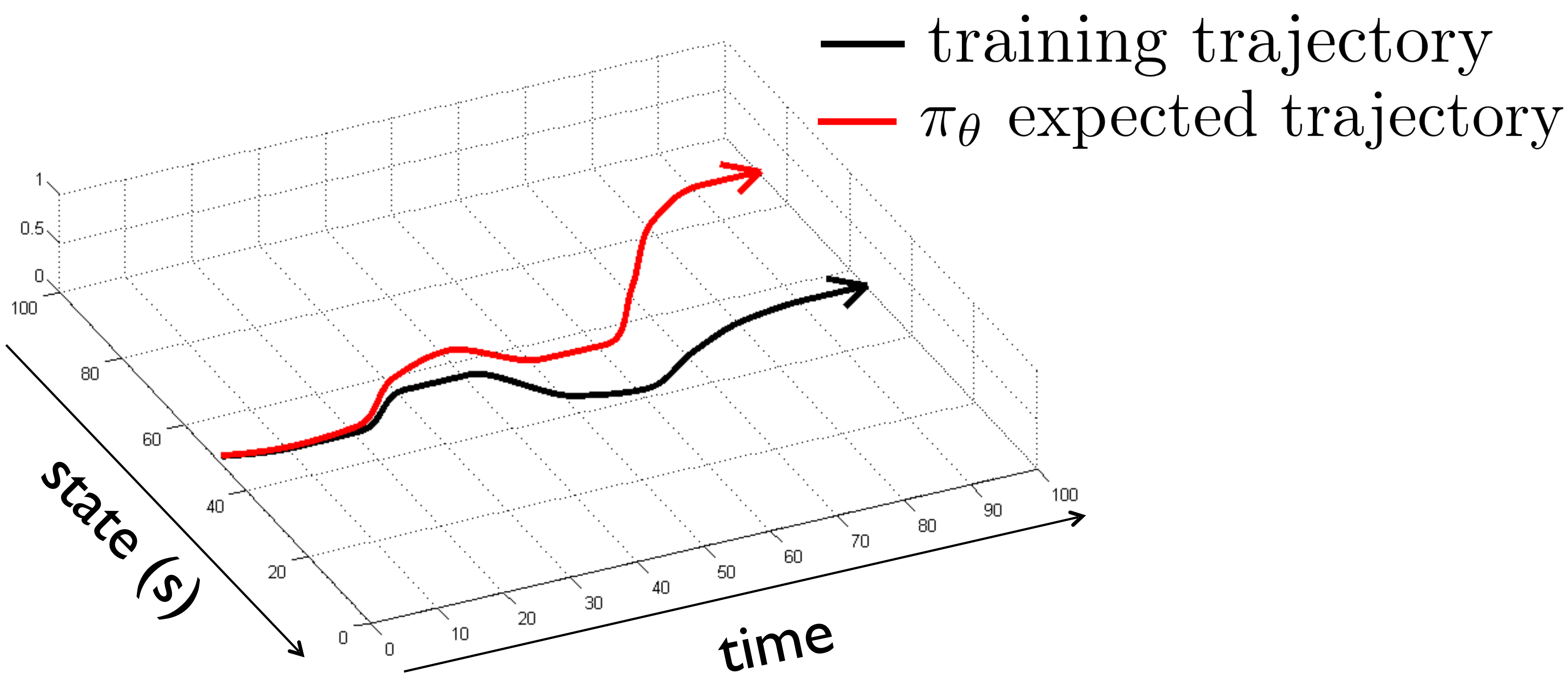


S. Gupta et al. **Cognitive Mapping and Planning for Visual Navigation.** CVPR 2017.

Behavior Cloning

Does it always work?

No, data mis-match problem

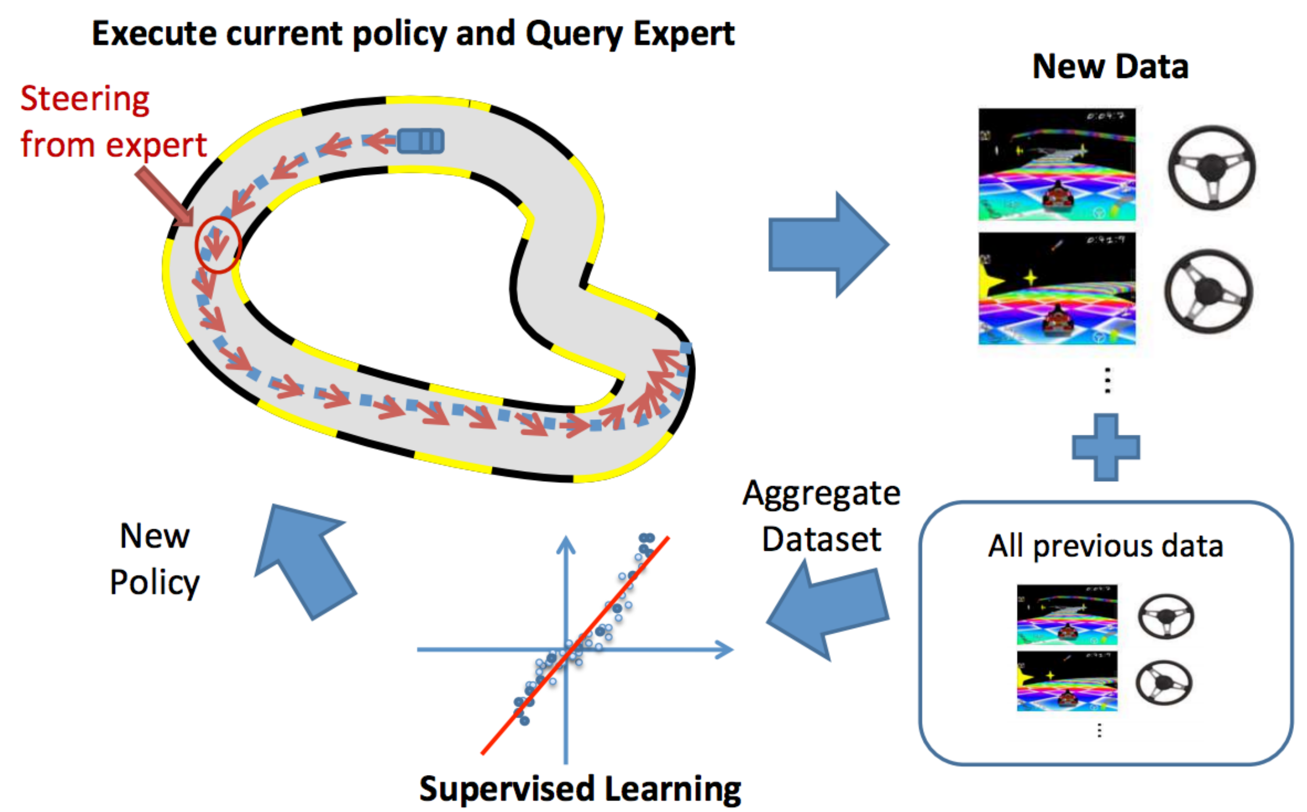


Fix Data Mis-Match Problem

DAgger: Dataset Aggregation

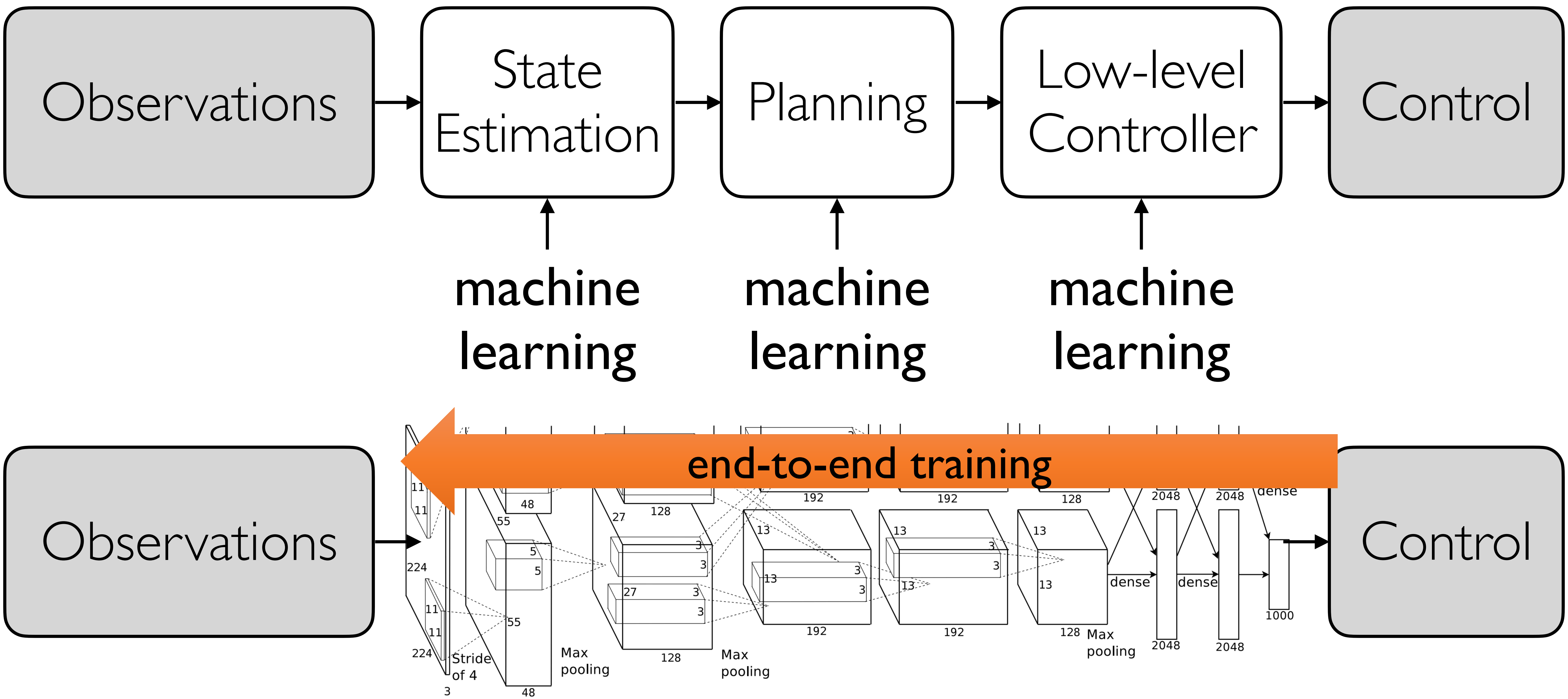
Collect labels on states visited by $\pi(o_t)$ *instead of* $\pi_e(o_t)$.

1. train $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{a}_1, \ldots, \mathbf{o}_N, \mathbf{a}_N\}$
2. run $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $\mathbf{a}_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$



S. Ross et al. **A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.** AISTAT 2011.

Typical Classical Robotics Pipeline



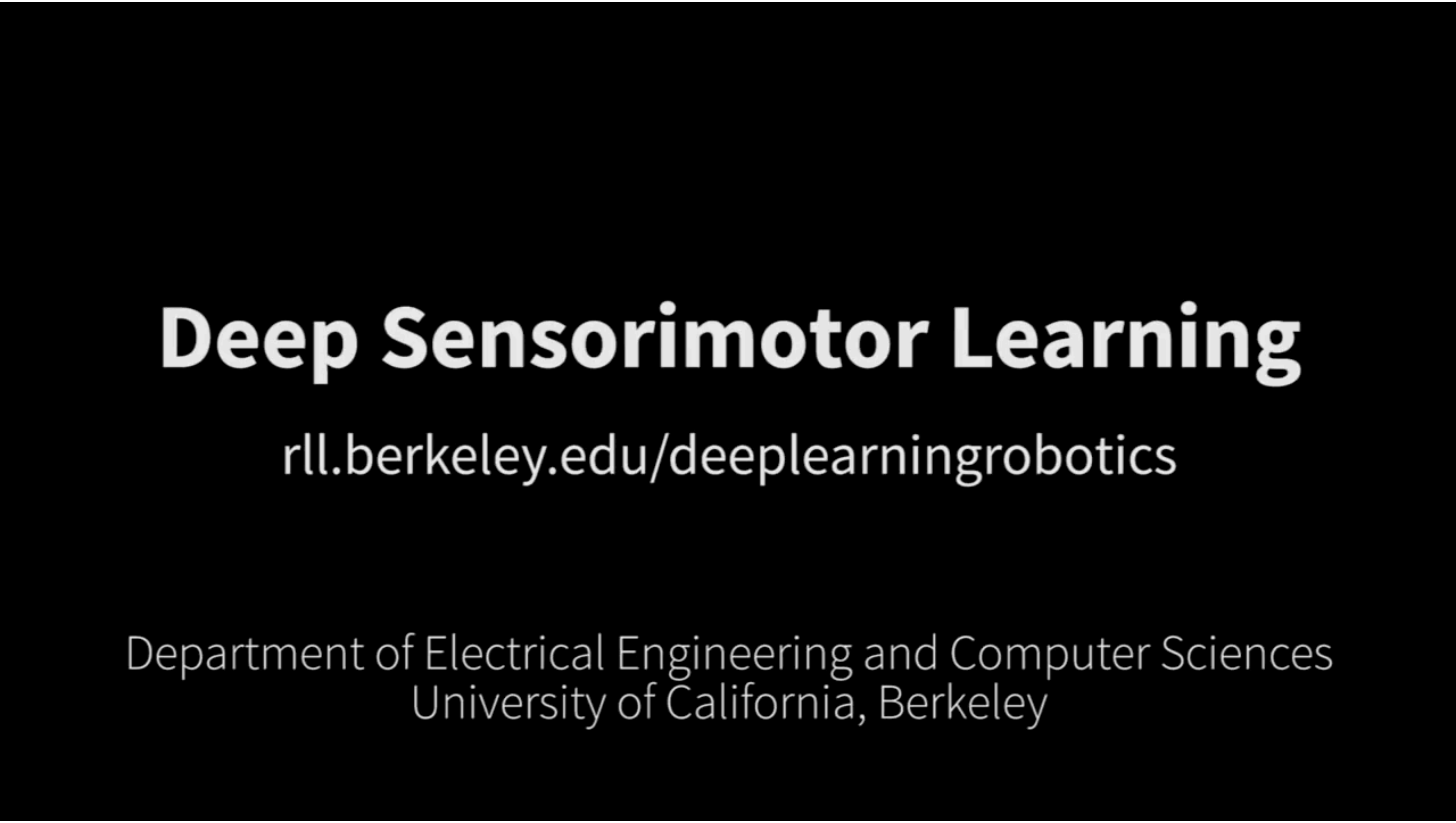
Should it help?

If so, how to do it?

Does it help?

Slide adapted from S. Levine.

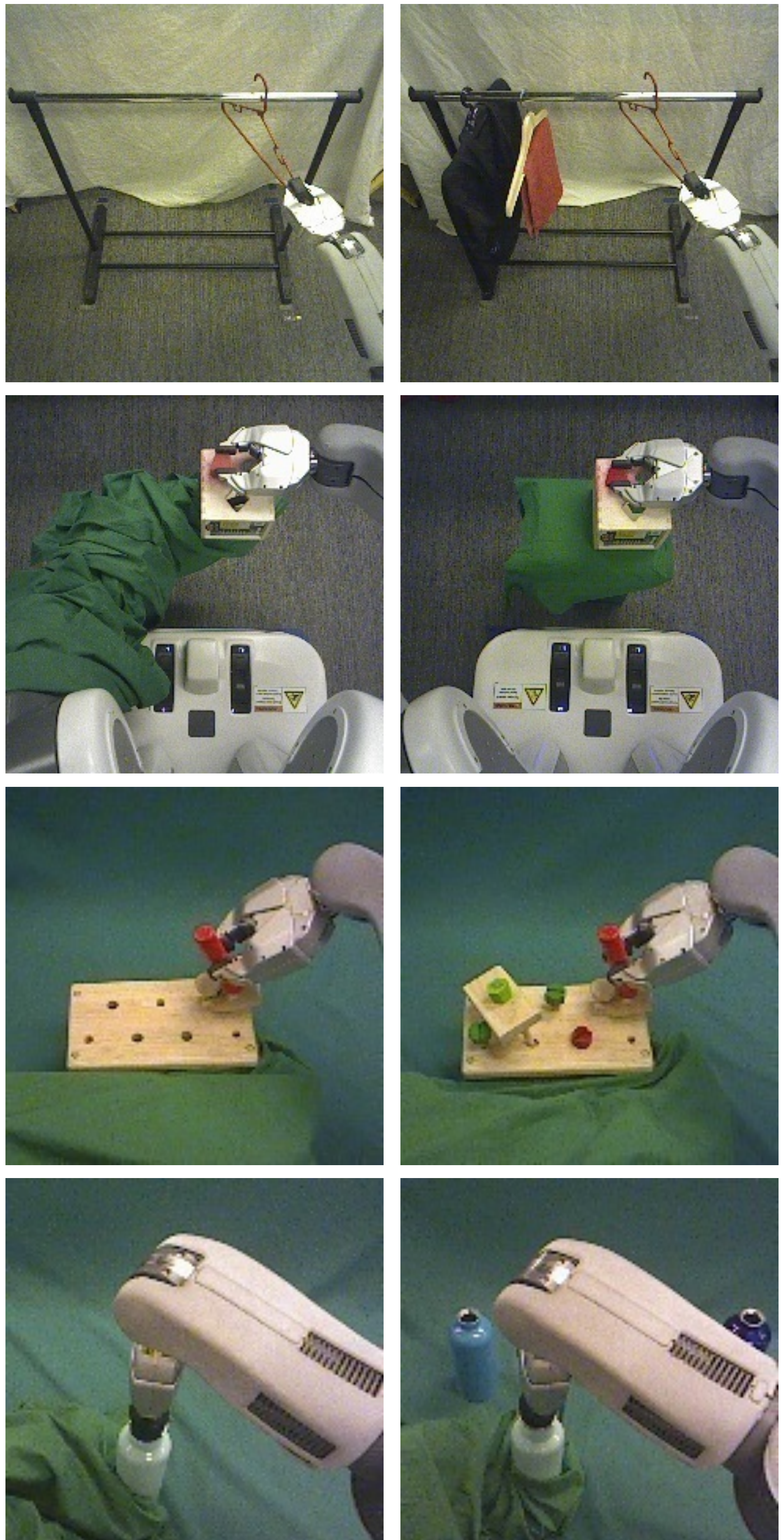
Supervision from an Algorithmic Expert



S. Levine et al. **A End-to-End Training of Deep Visuomotor Policies.** JMLR 2016.

training | visual test

hanger / cube / hammer / bottle

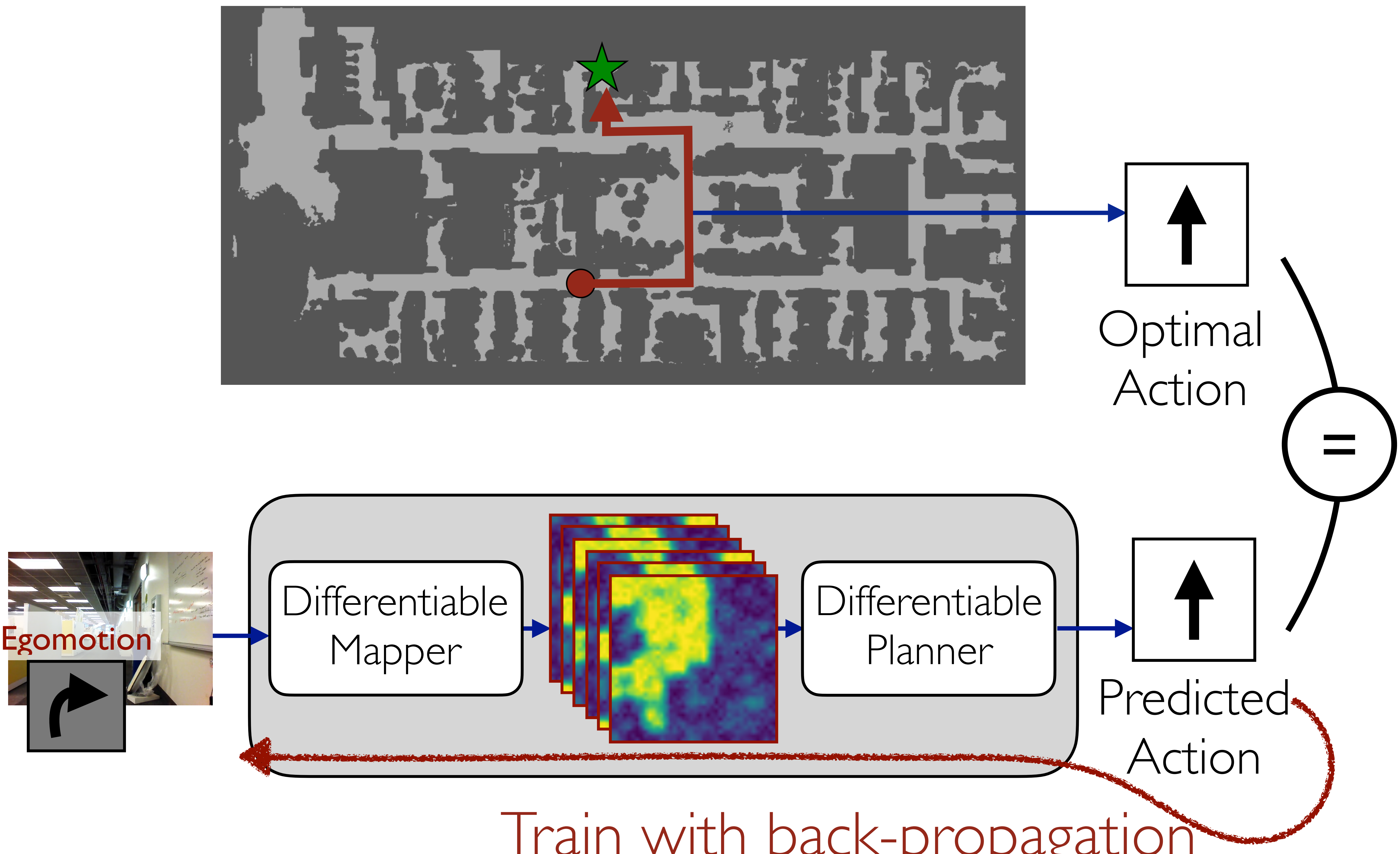


coat hanger	training (18)	spatial test (24)	visual test (18)
end-to-end	**100%**	**100%**	**100%**
pose features	88.9%	87.5%	83.3%
pose prediction	55.6%	58.3%	66.7%
shape cube	training (27)	spatial test (36)	visual test (40)
end-to-end	**96.3%**	**91.7%**	**87.5%**
pose features	70.4%	83.3%	40%
pose prediction	0%	0%	n/a
toy hammer	training (45)	spatial test (60)	visual test (60)
end-to-end	**91.1%**	**86.7%**	**78.3%**
pose features	62.2%	75.0%	53.3%
pose prediction	8.9%	18.3%	n/a
bottle cap	training (27)	spatial test (12)	visual test (40)
end-to-end	**88.9%**	**83.3%**	**62.5%**
pose features	55.6%	58.3%	27.5%

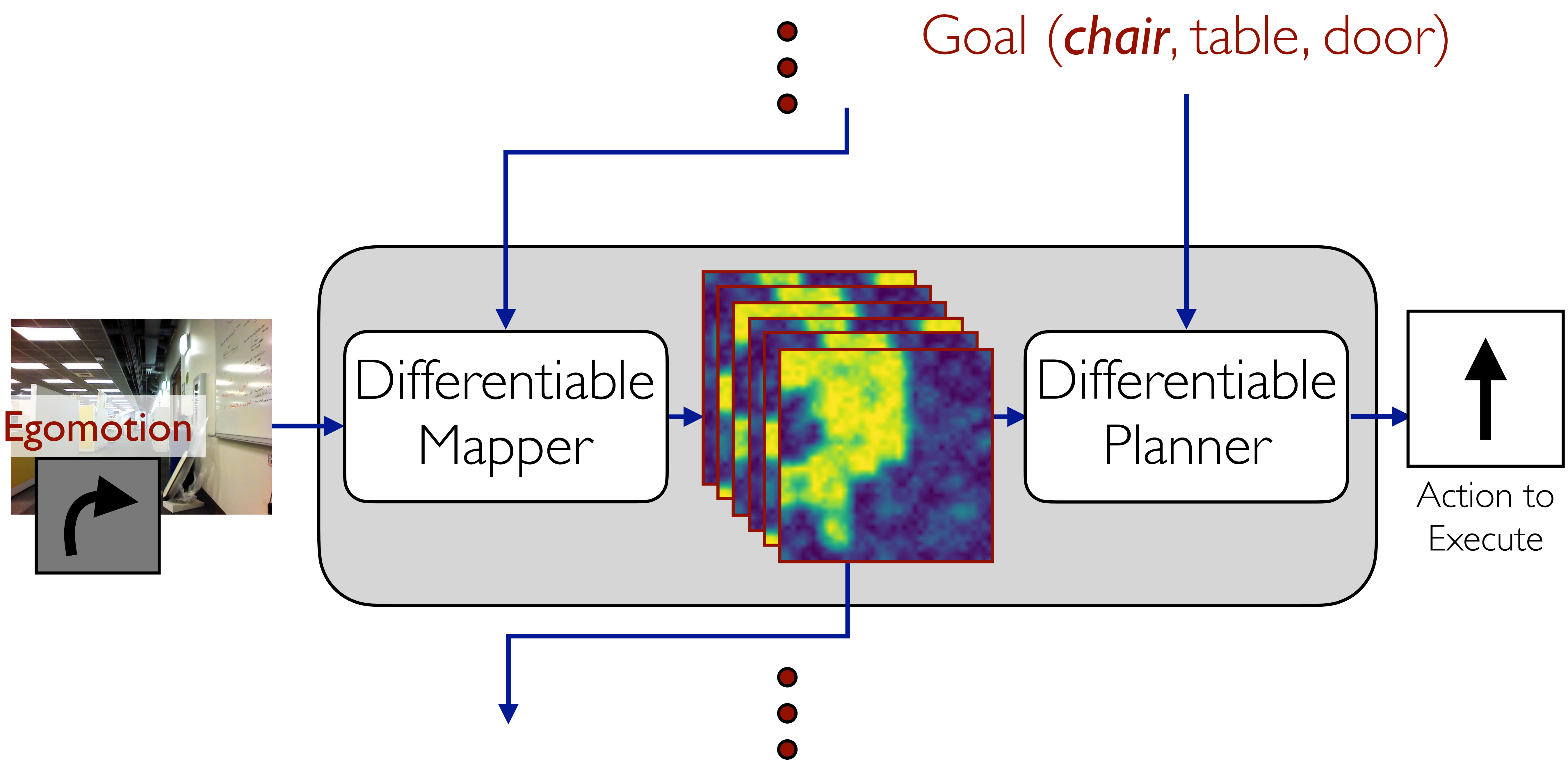
Success rates on training positions, on novel test positions, and in the presence of visual distractors. The number of trials per test is shown in parentheses.

S. Levine et al. **A End-to-End Training of Deep Visuomotor Policies.** JMLR 2016.

Can also be applied to navigation

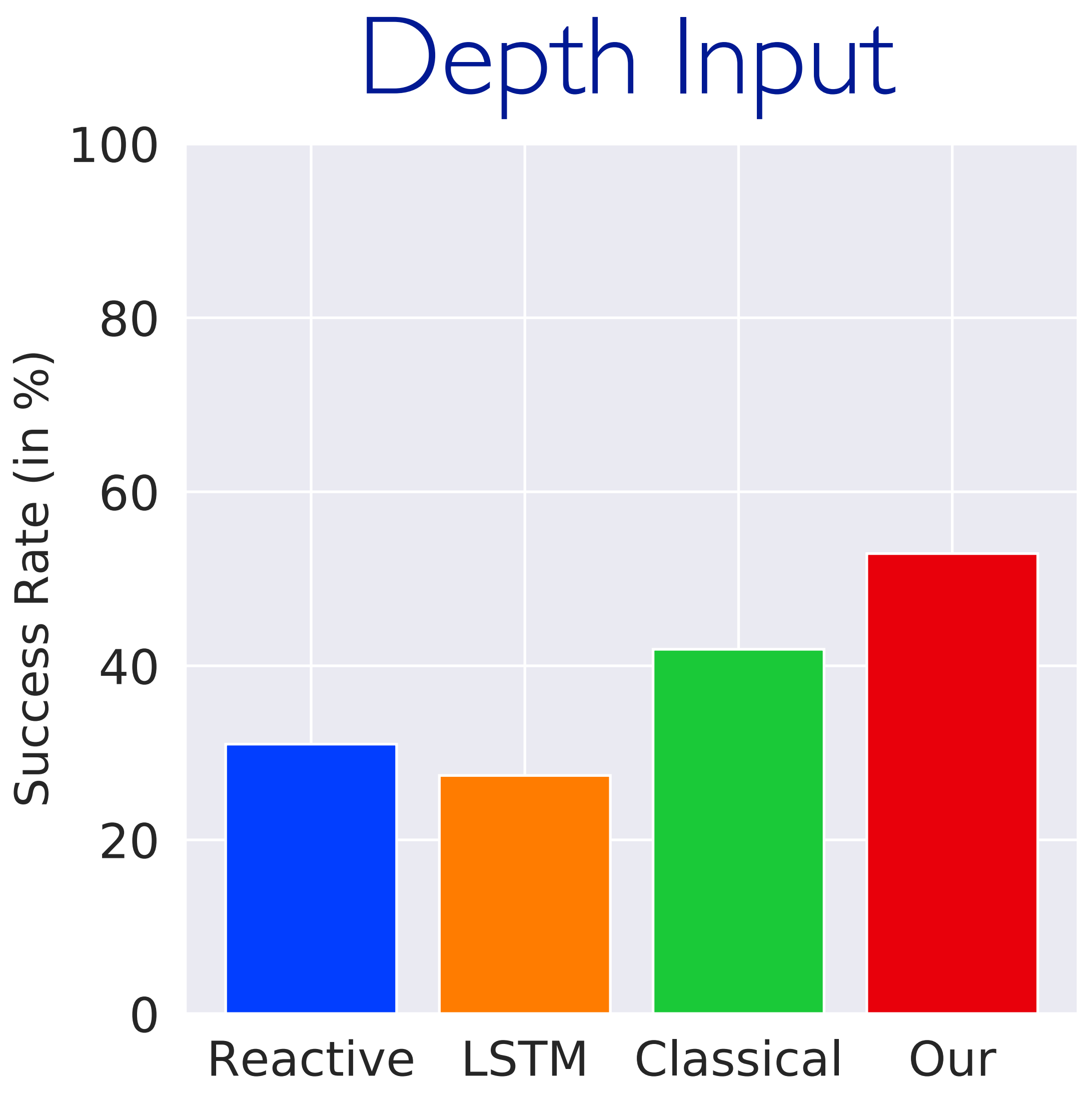


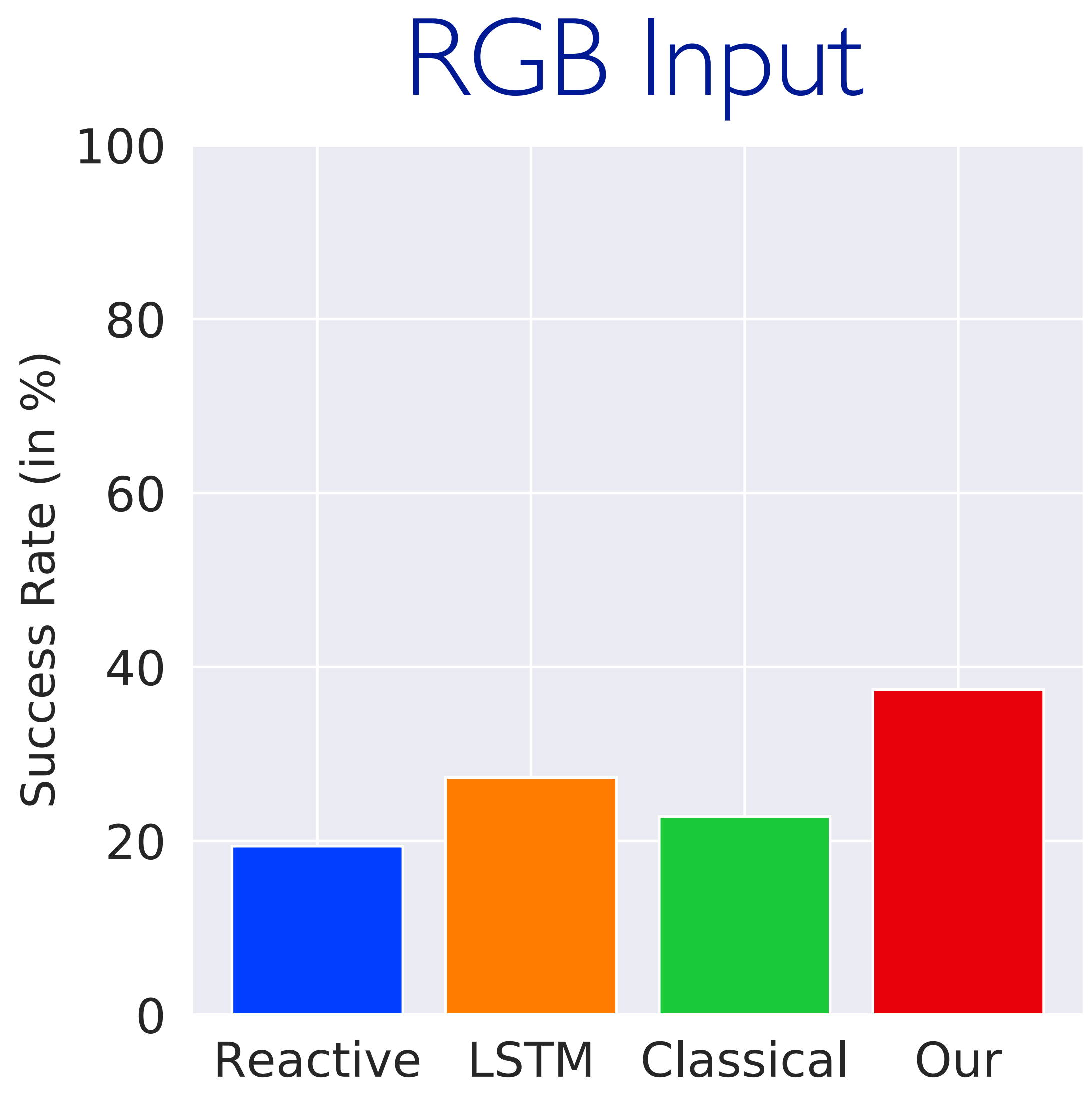
Gupta et al., CVPR 2017. Cognitive Mapping and Planning for Visual Navigation



Gupta et al., CVPR 2017. Cognitive Mapping and Planning for Visual Navigation

Results (Novel Env, Go To Object)

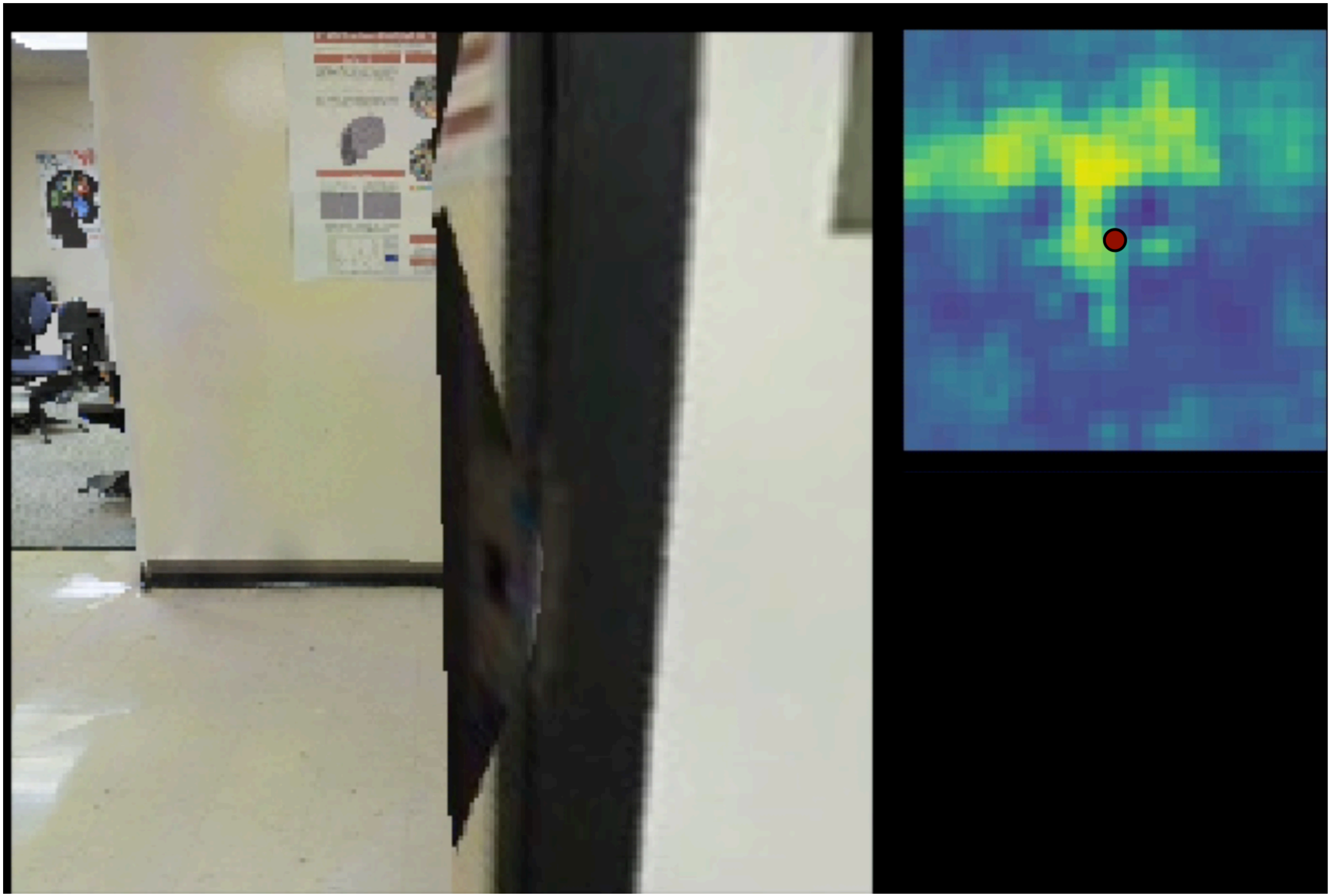




Gupta et al., CVPR 2017. Cognitive Mapping and Planning for Visual Navigation

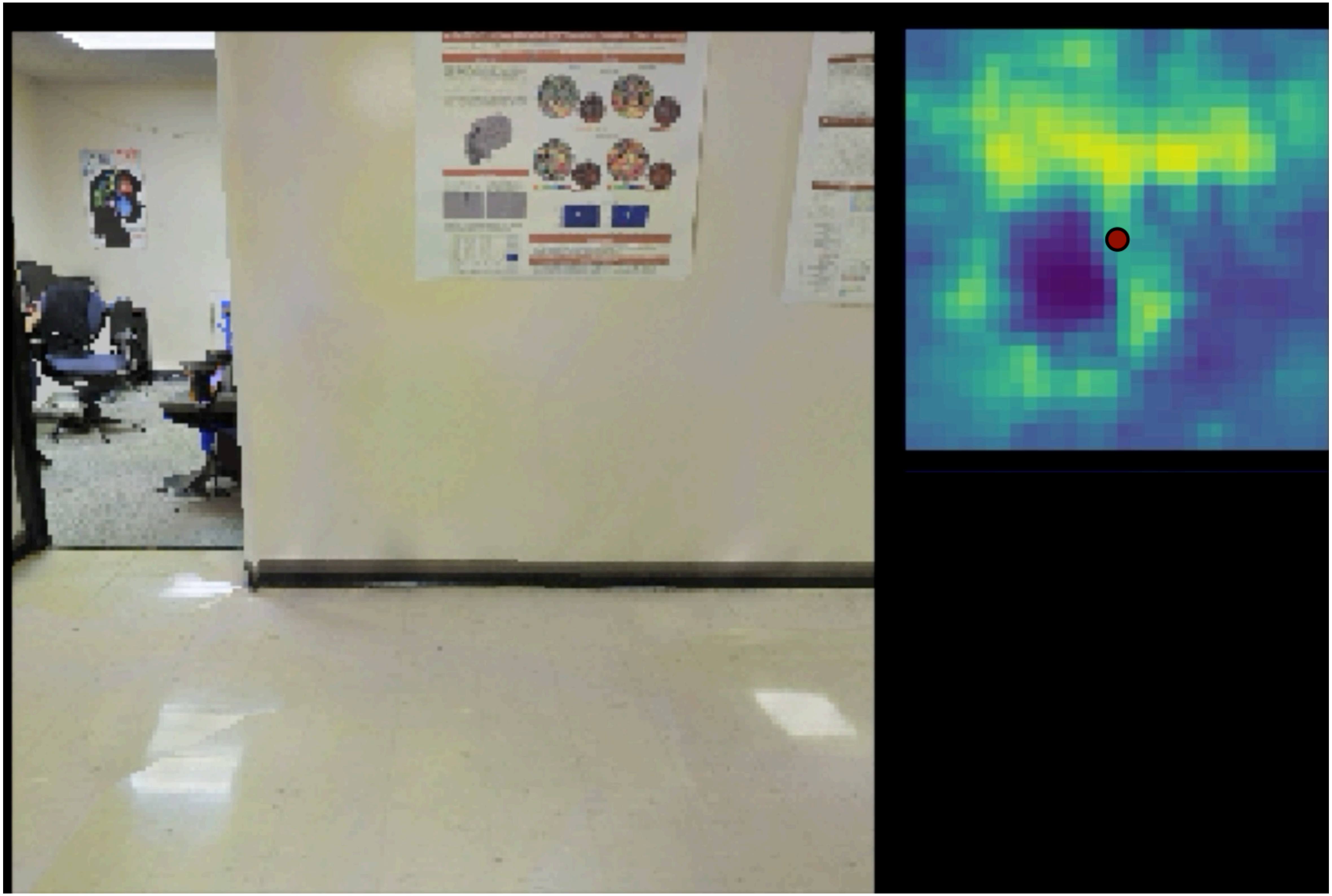
Agent can make predictions about its surroundings

Free Space

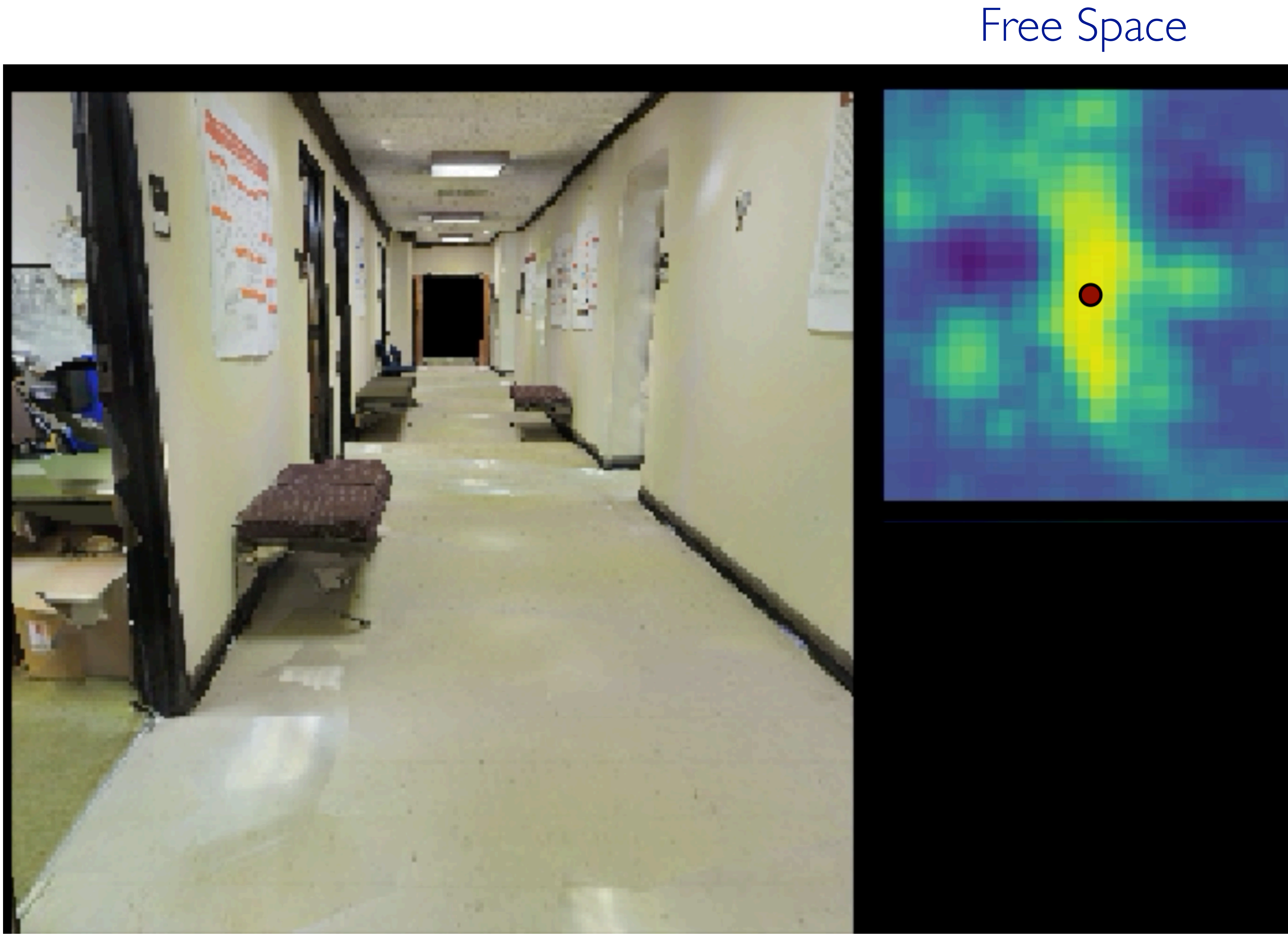


Agent can make predictions about its surroundings

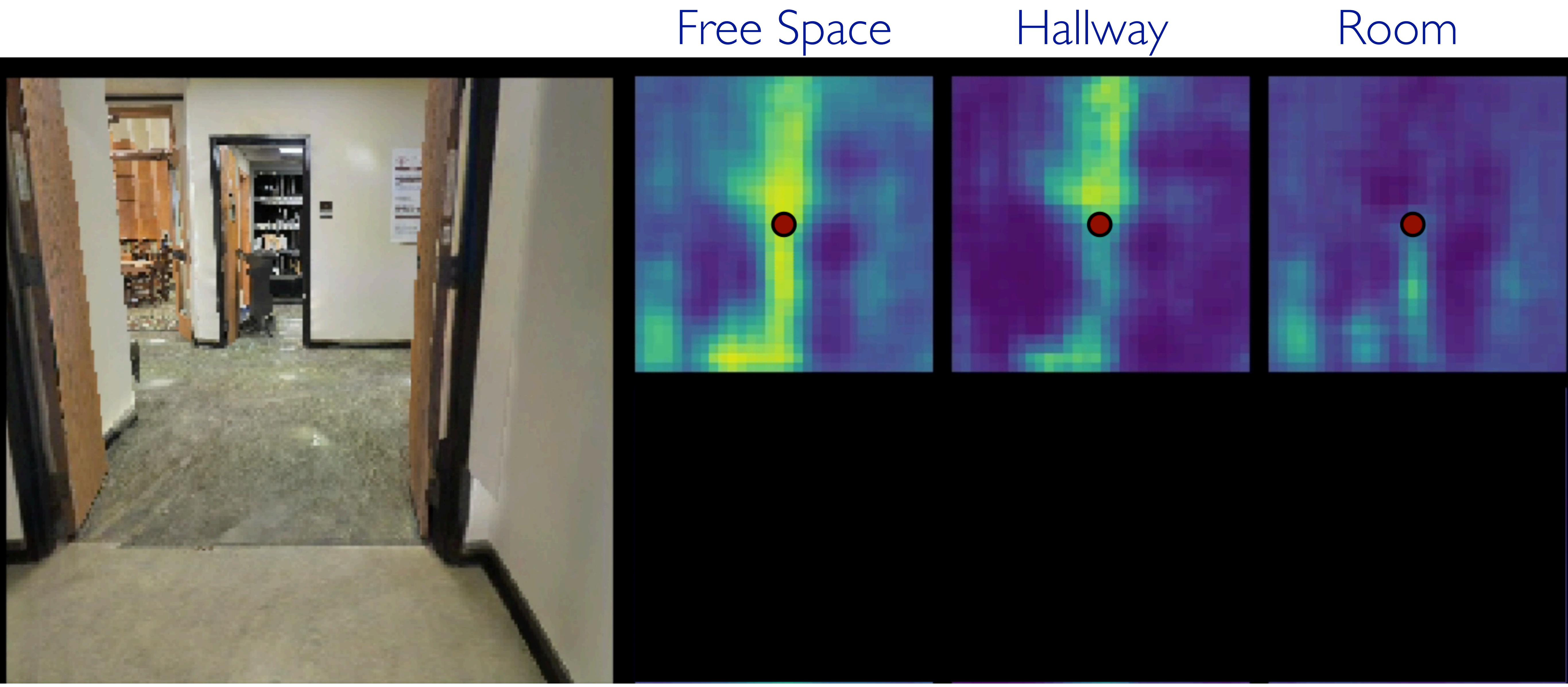
Free Space



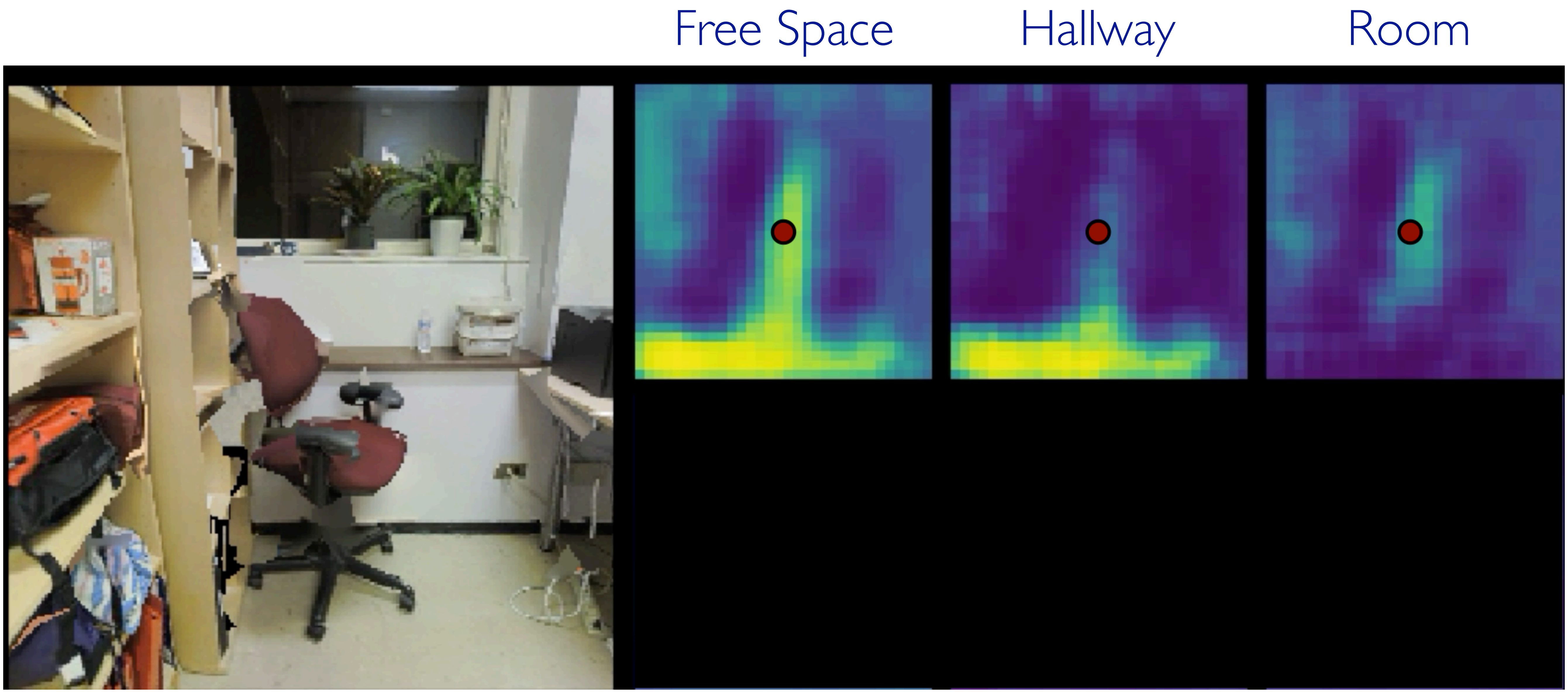
Agent can make predictions about its surroundings



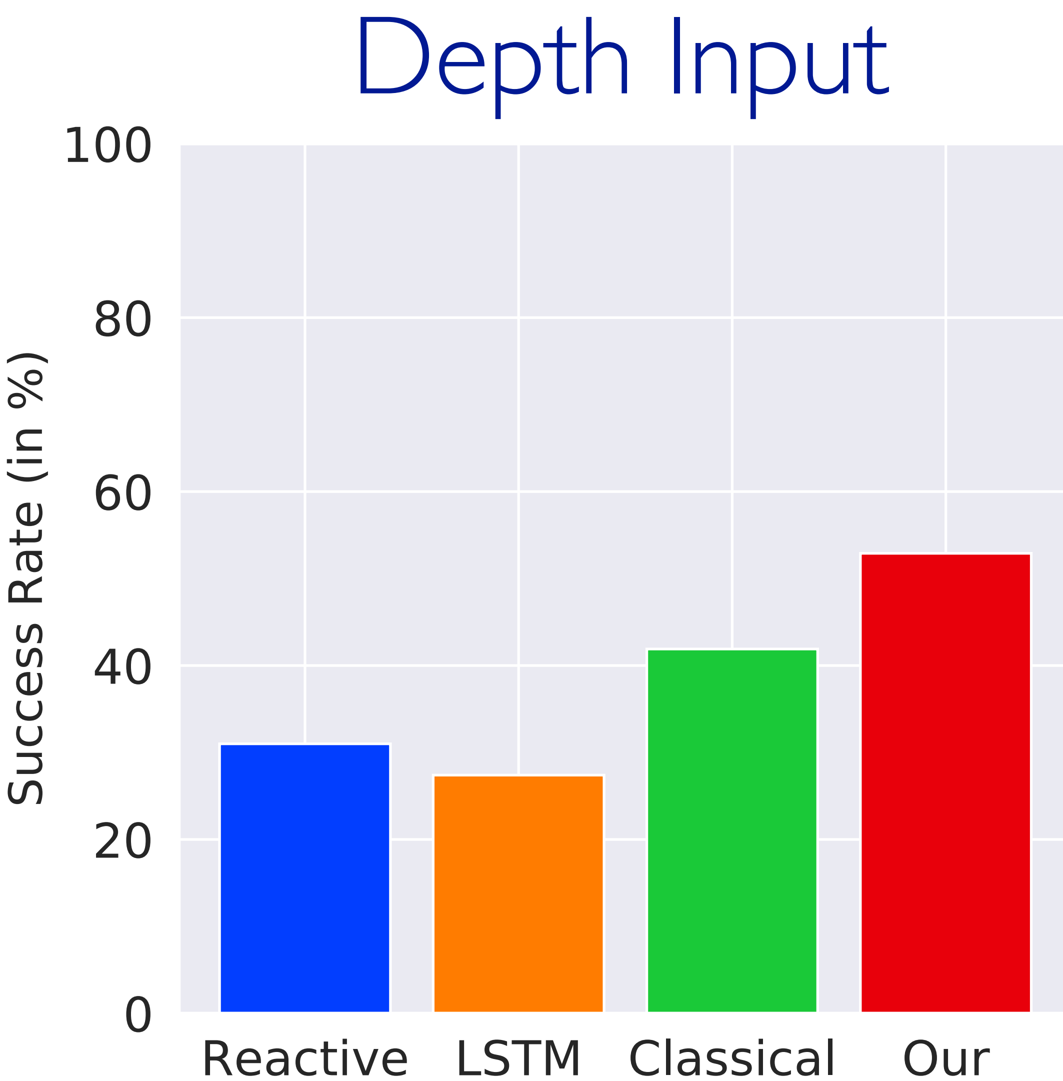
Agent can make predictions about its surroundings

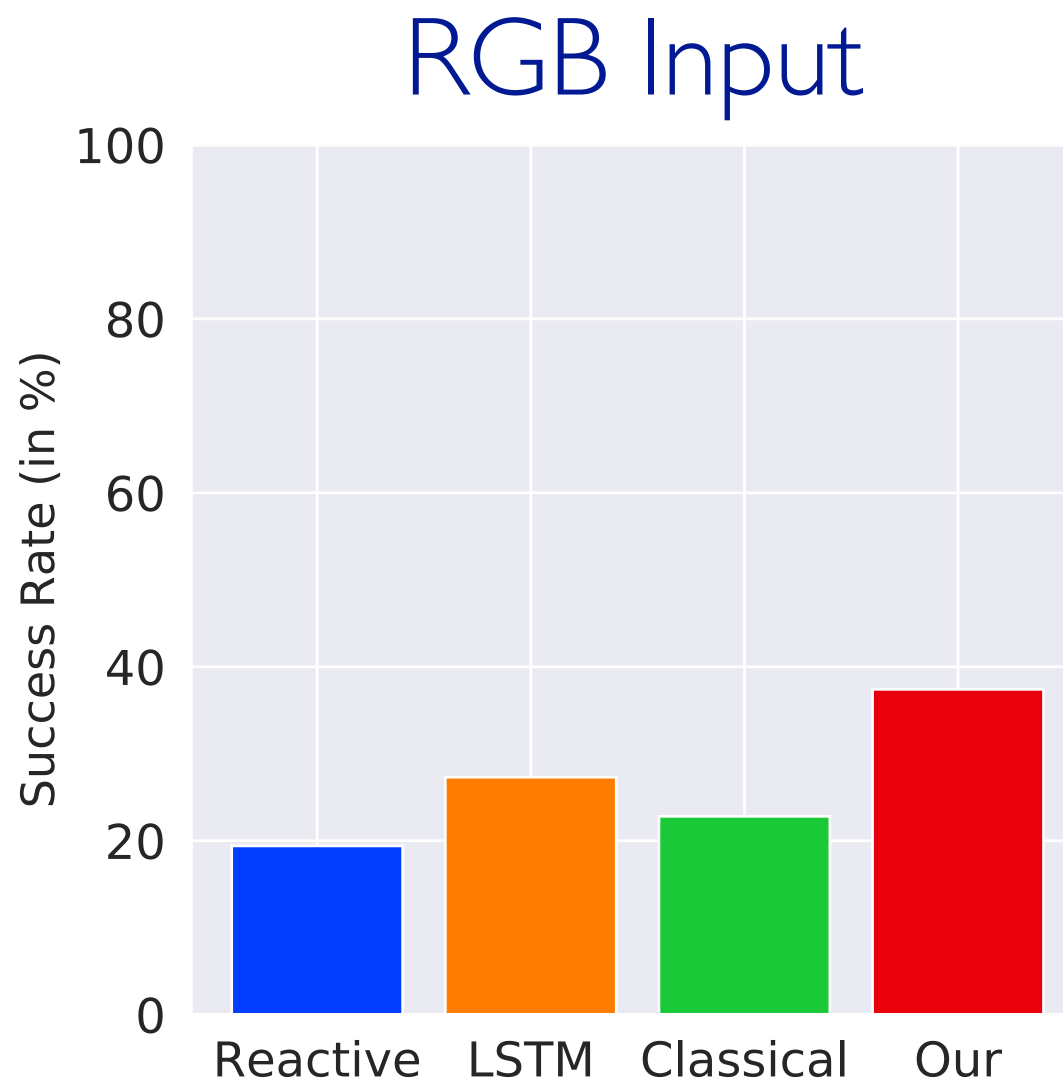


Agent can make predictions about its surroundings



But representations are still important!



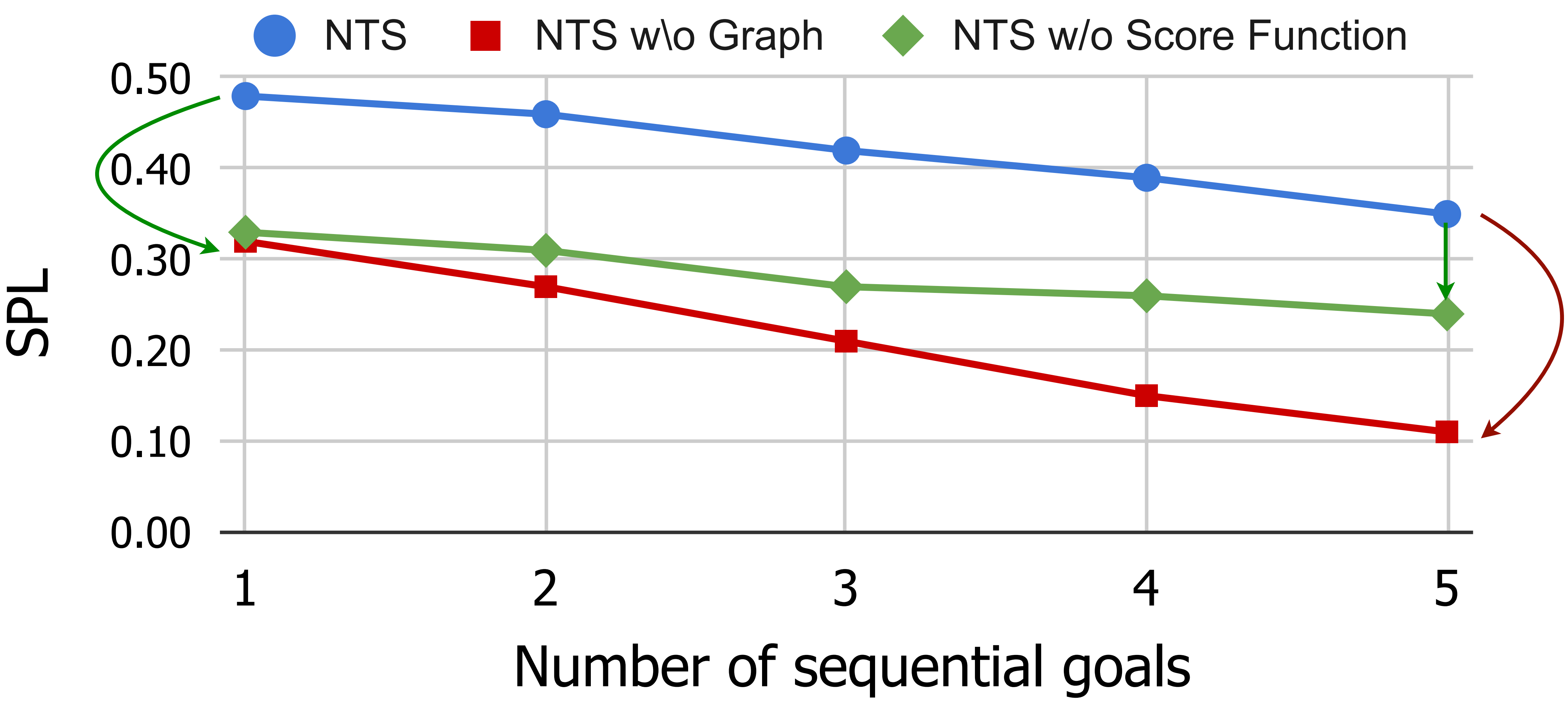


Gupta et al., CVPR 2017. Cognitive Mapping and Planning for Visual Navigation

Visual Semantic Navigation

Semantic score function improves efficiency when no prior experience with environment is available.

As experience in environment increases, utility of semantic function decreases



But, at the same time, importance of the topological representation increases