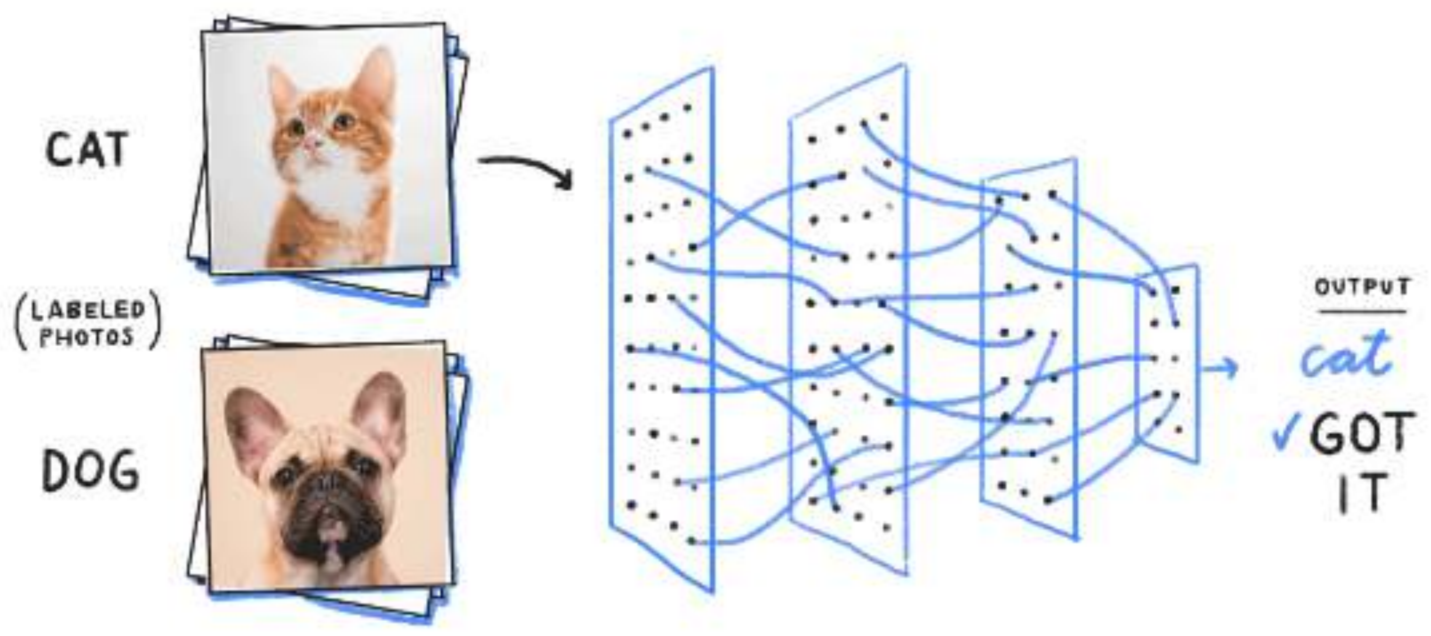
Learning How-To Knowledge from the Web

Yuke Zhu

IROS 2019

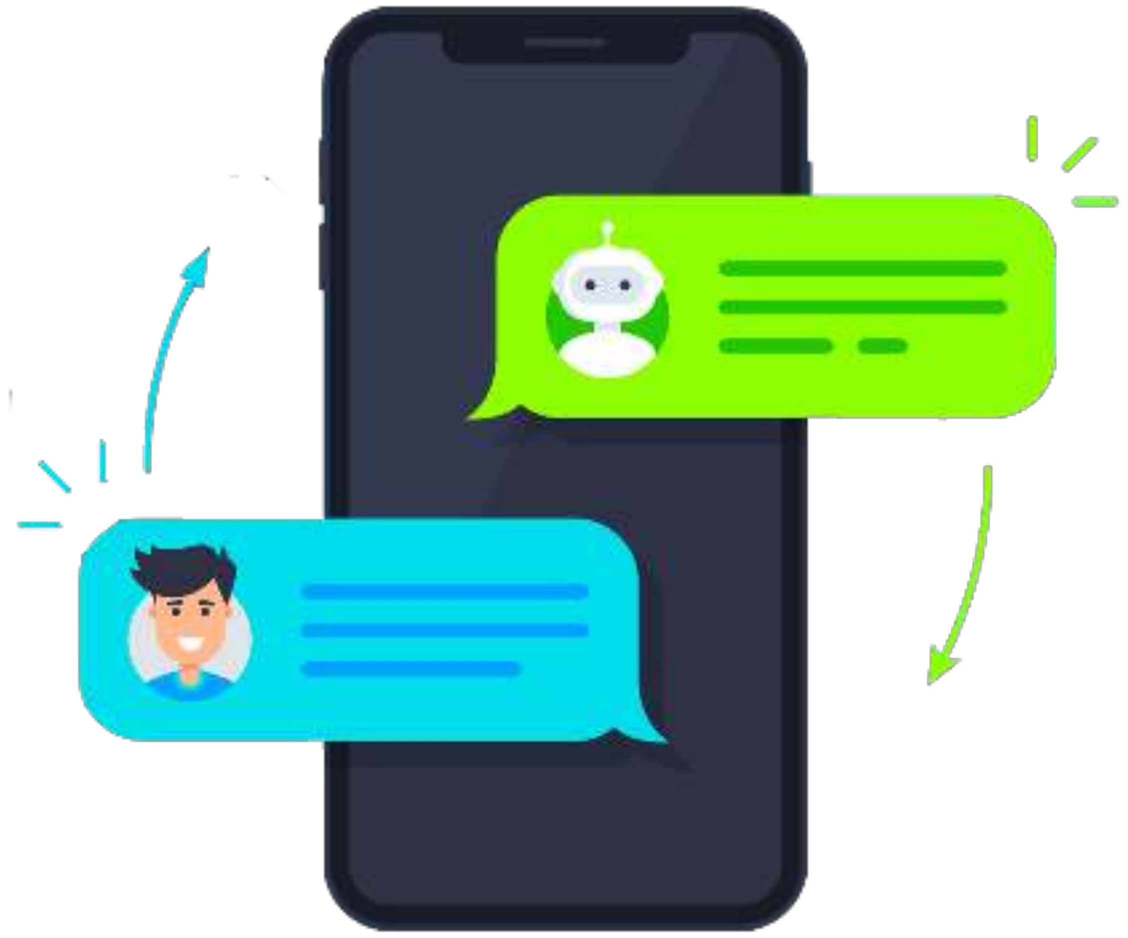
Stanford University
NVIDIA
TEXAS
The University of Texas at Austin

Advances in Artificial Intelligence



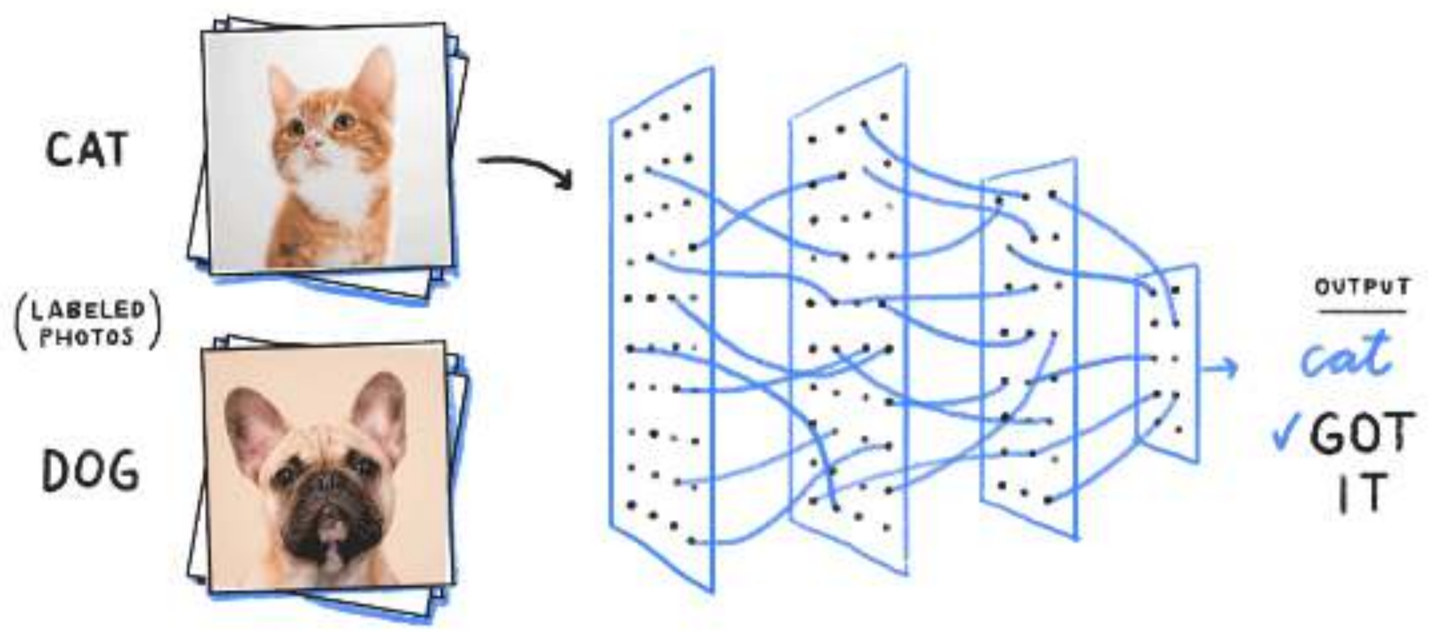
Visual Recognition

Machine Translation



Question Answering

The Unsung Hero: Web Data



Visual Recognition

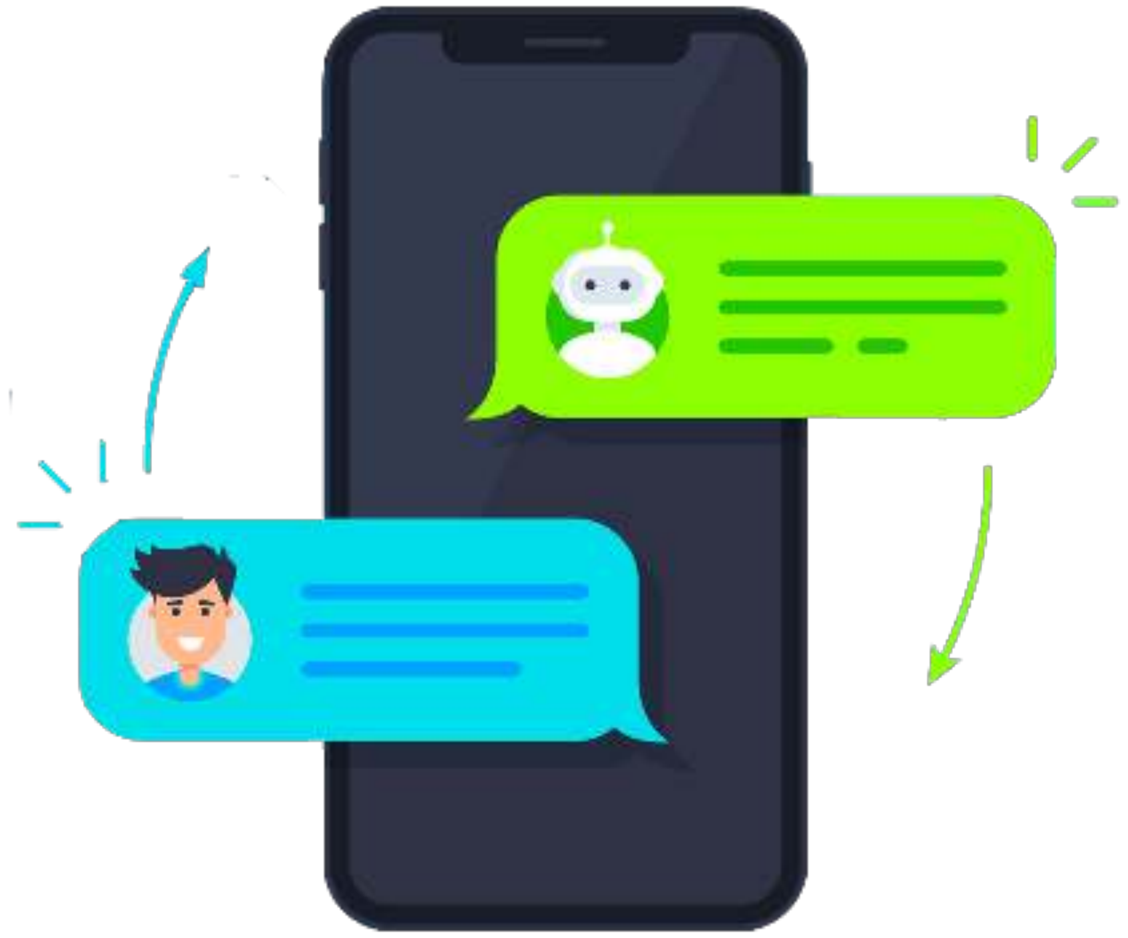
ImageNet
[Deng et al. 2009]

14 million web images
annotated by AMT workers

Machine Translation

Google NMT
[Wu et al. 2016]

WMT En→Fr dataset
with 36M sentence pairs



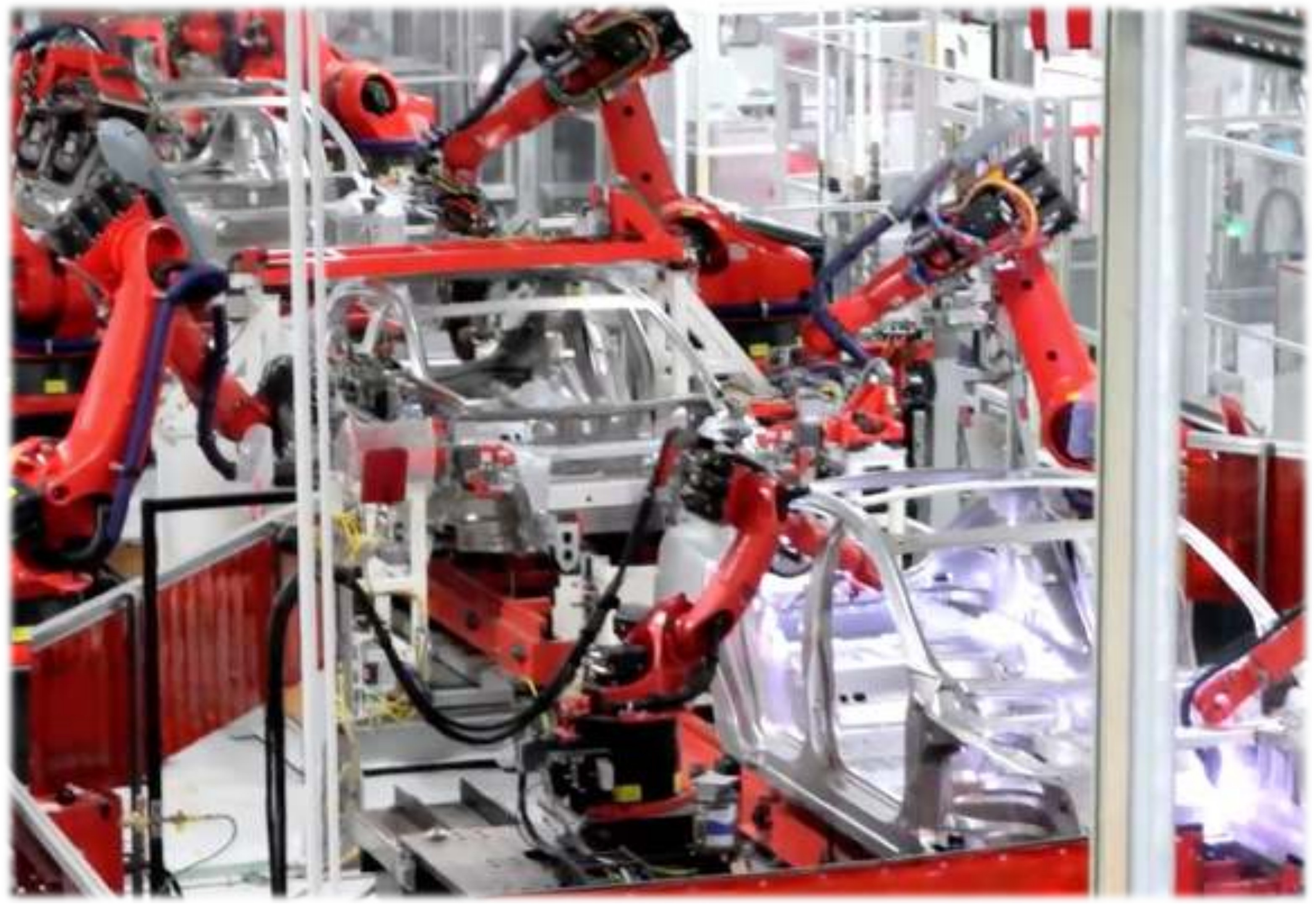
Question Answering

SQuAD QA Dataset
[Rajpurkar et al. 2016]

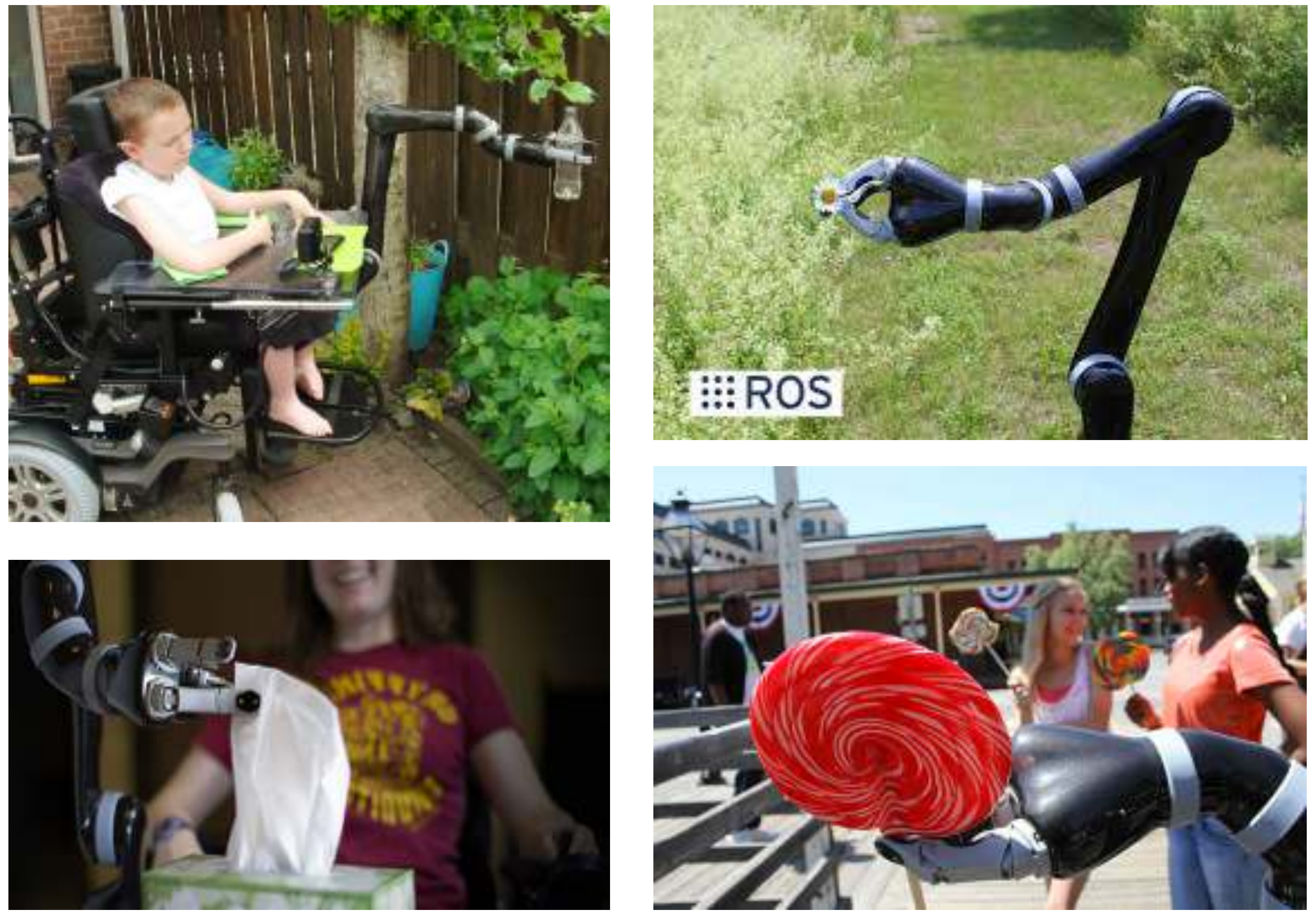
100,000+ questions posed
by crowdworkers on a set
of Wikipedia articles

The Unsung Hero: Web Data

What’s the role of web data in improving robot intelligence?



?



Traditional form of automation

Intelligent robots in real world

What knowledge do we need for robotics?

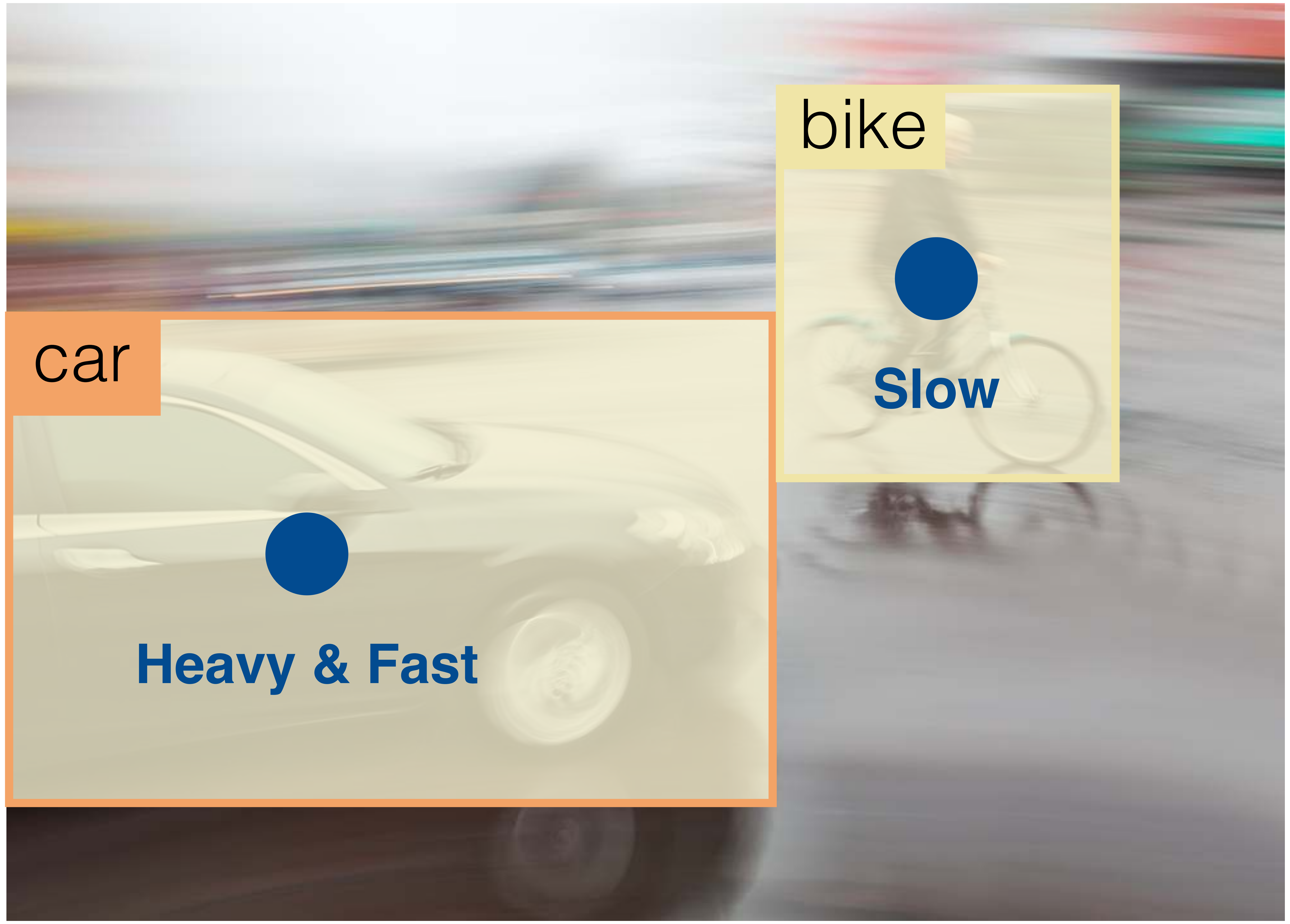
“To accelerate or to brake?”

What knowledge do we need for robotics?

Declarative knowledge

Understanding the world

- ❖ Describes facts of the world
- ❖ Easy to articulate (conscious)



Knowledge of “That-Is”

What knowledge do we need for robotics?

Declarative knowledge

Understanding the world

- ❖ Describes facts of the world
- ❖ Easy to articulate (conscious)

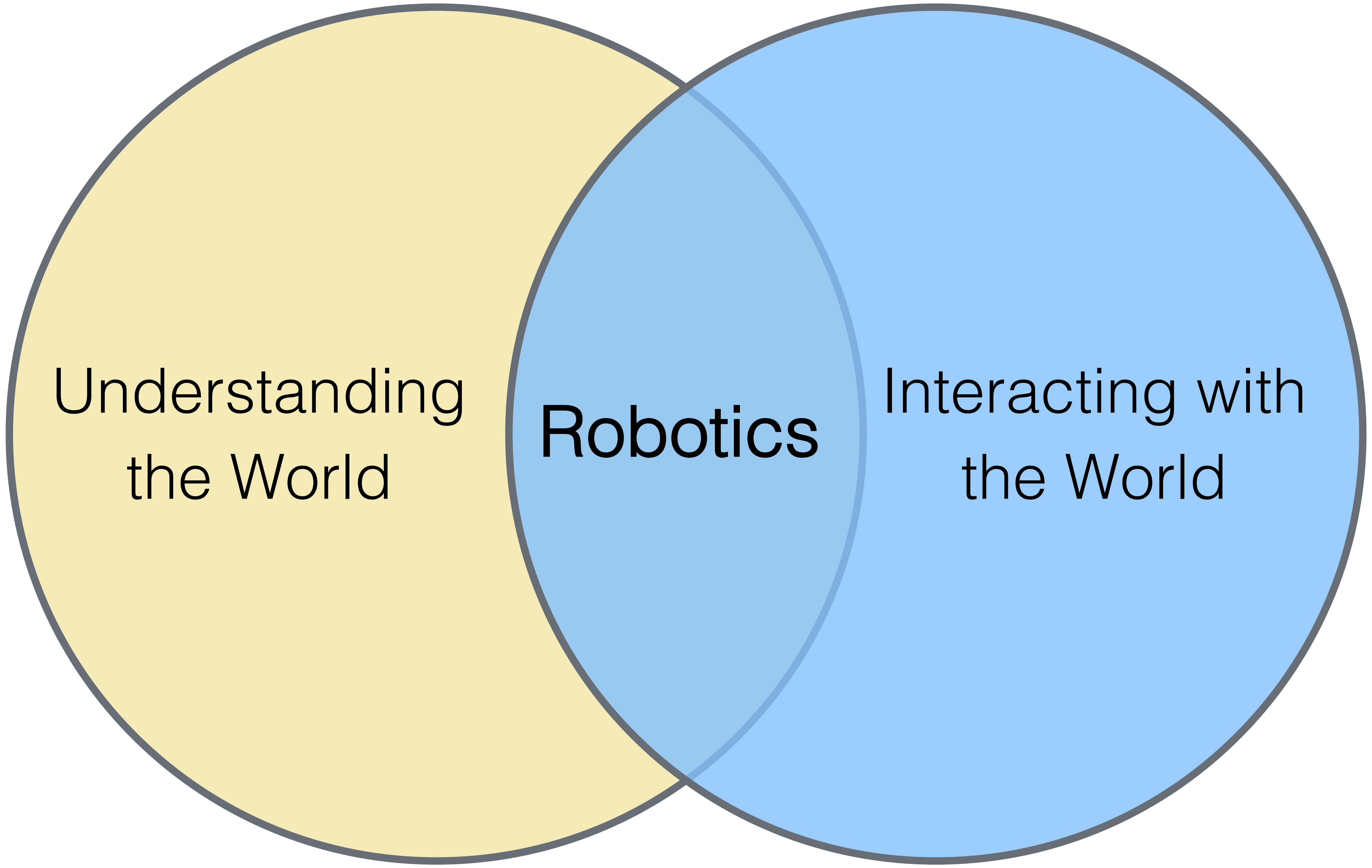
Knowledge of "How-To"

Procedural knowledge

Interacting with the world

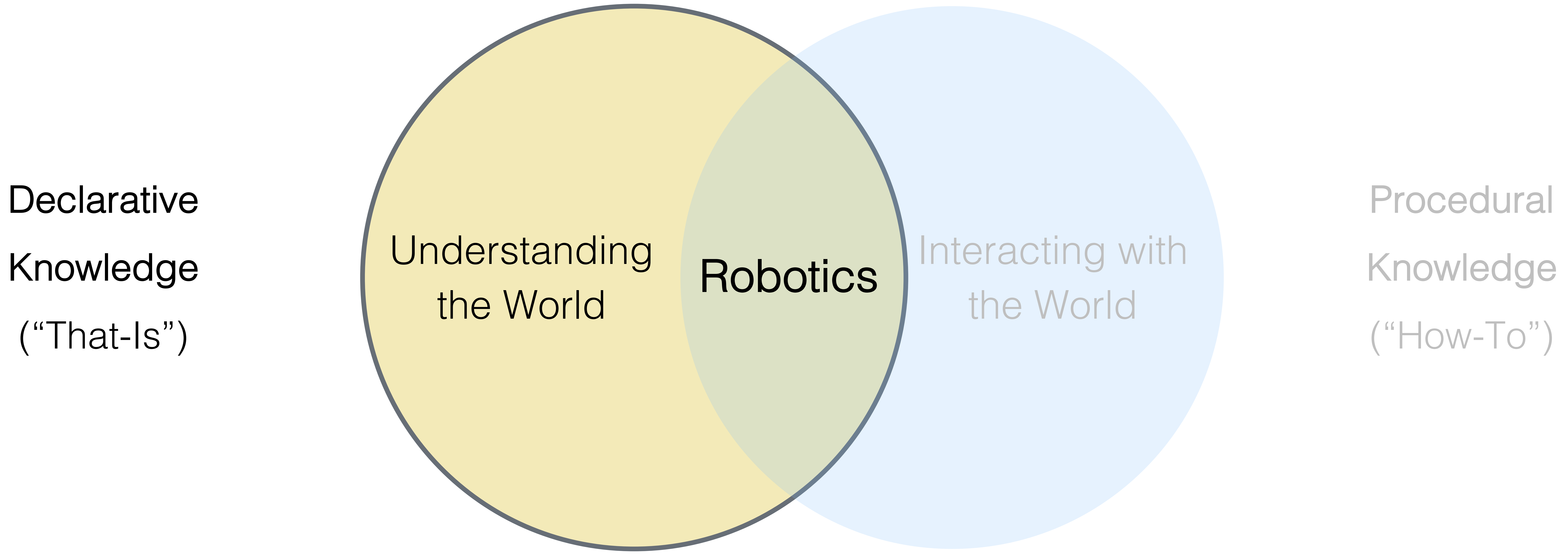
- ❖ Describes how to perform tasks
- ❖ Hard to pinpoint (unconscious)

Declarative Knowledge ("That-Is")



Procedural Knowledge ("How-To")

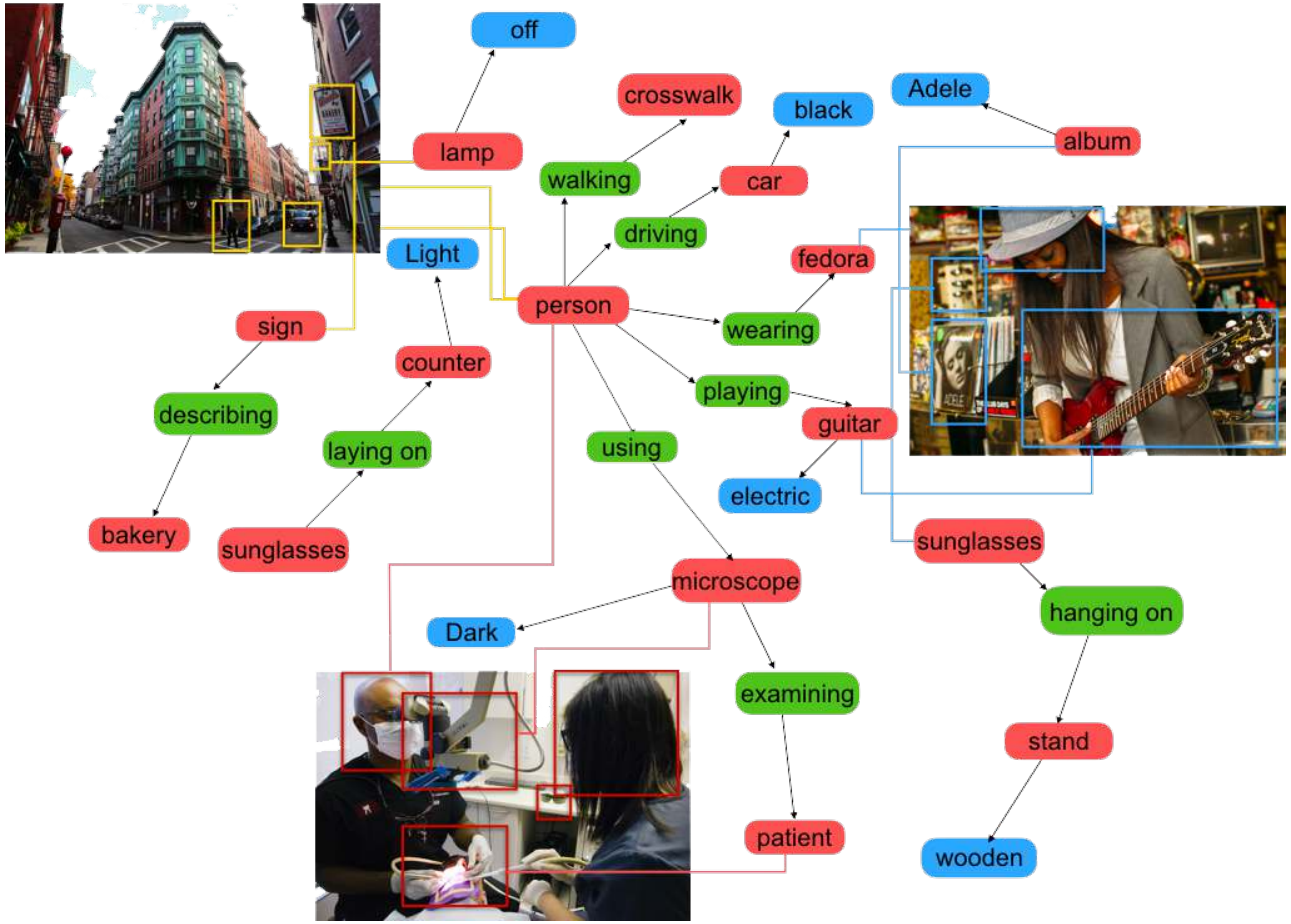
Learning Declarative (“That-Is”) Knowledge from the Web



Understanding the world is the cornerstone of interacting with the world.

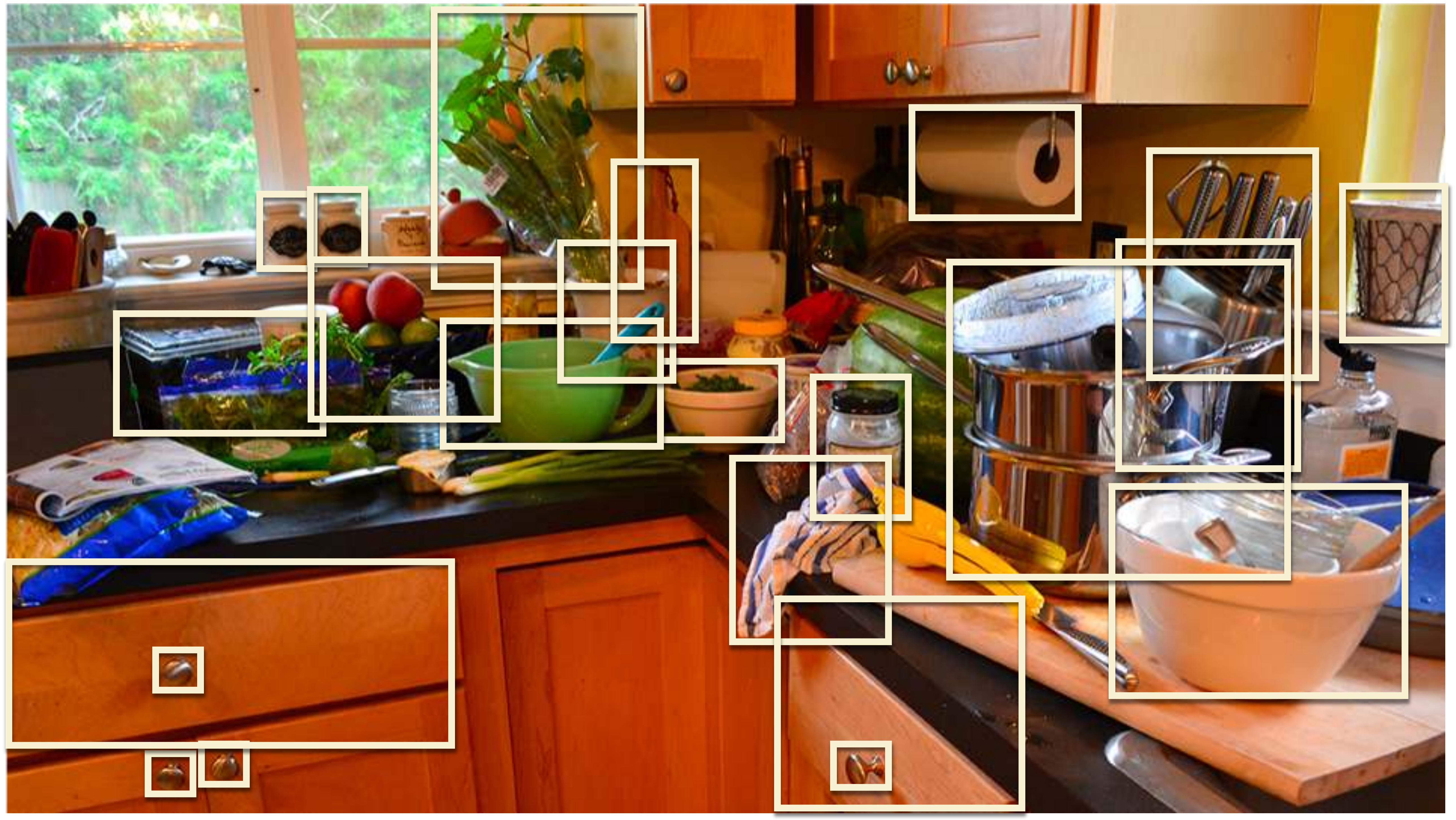
The Visual Genome Project

A large-scale visual knowledge base of structured image concepts

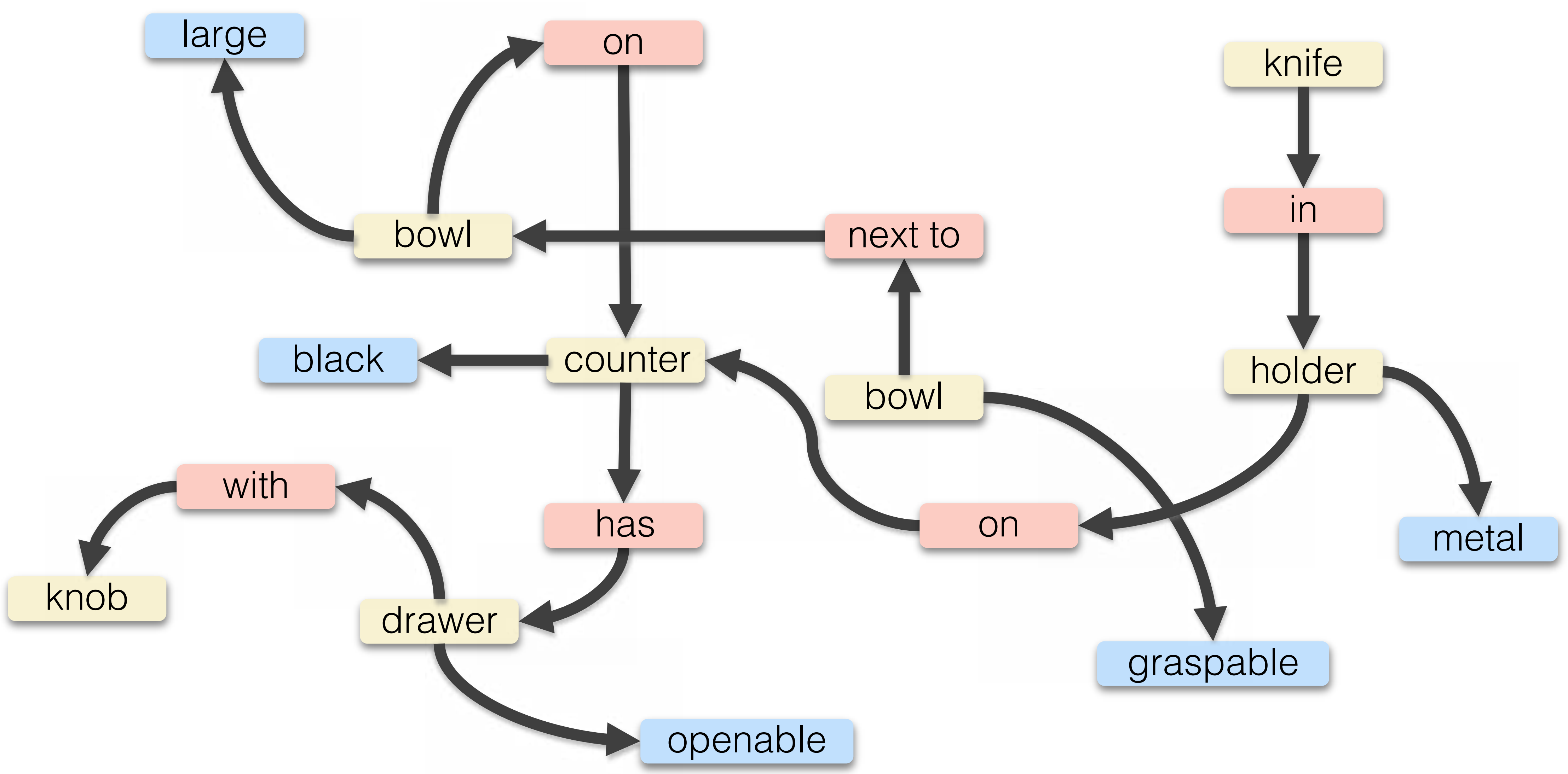


Krishna, **Zhu**, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei, IJCV 2017

Visual Genome



Scene Graph: Objects + Attributes + Relationships



Questions

1. Q: What's the color of the counter? A: Black.
2. Q: How many drawers can you see? A: Two.
3. Q: What's the material of the pots? A: Metal.

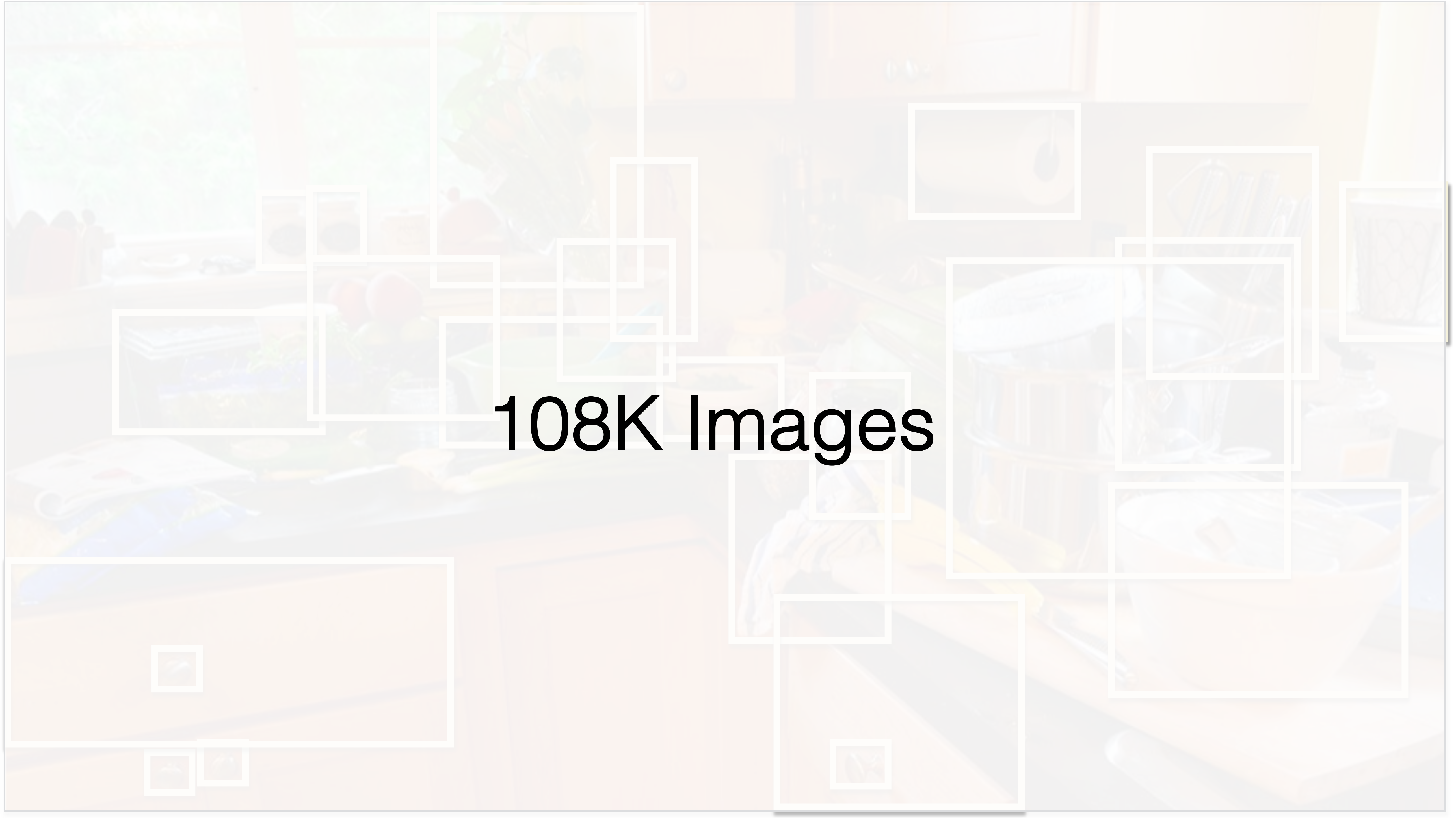
……

Region Descriptions

1. There a green bowl on the black counter.
2. The cabinet door is closed.
3. Six knives are placed in the knife holder.

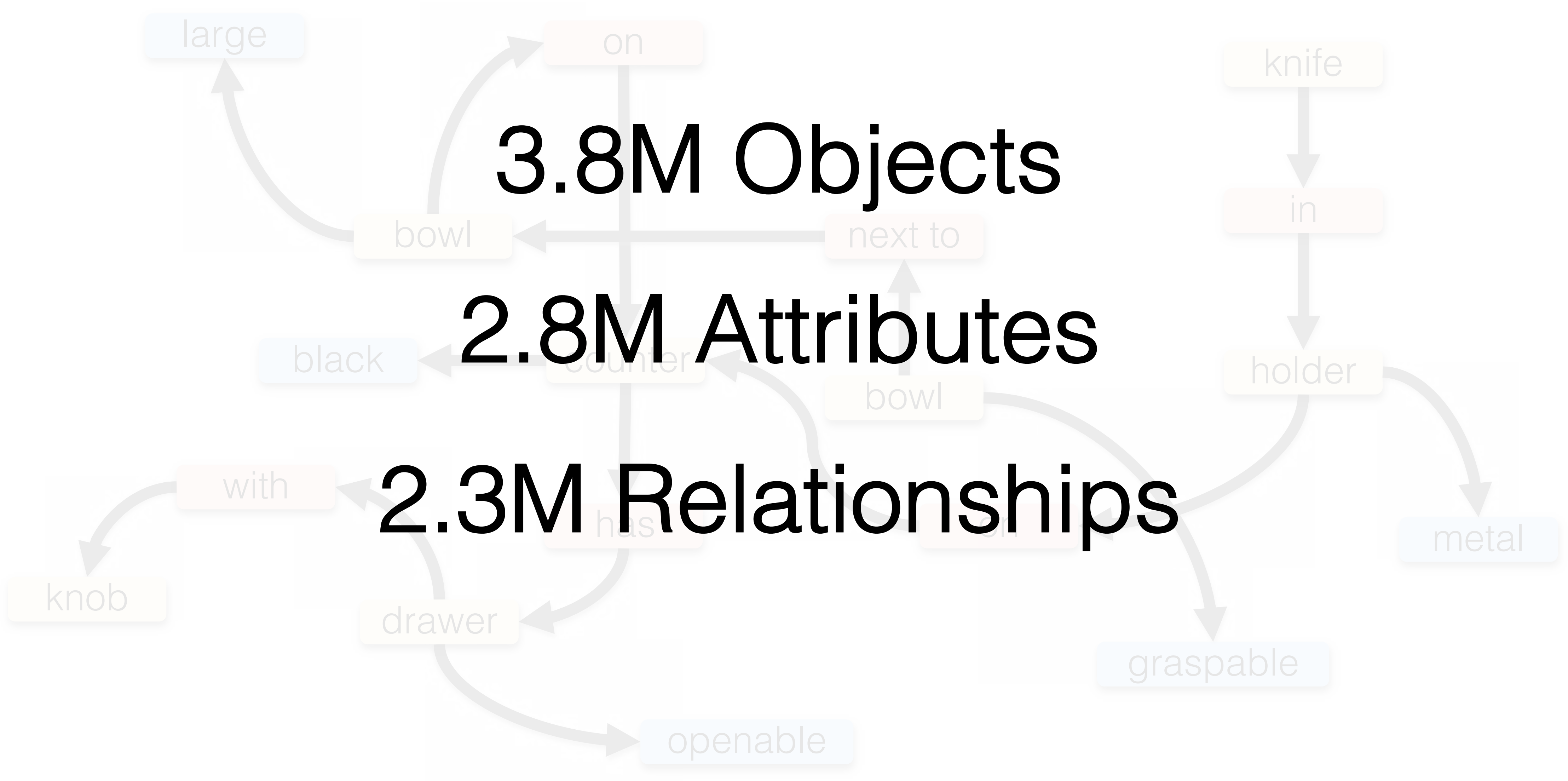
……

Visual Genome



108K Images

Scene Graph: Objects + Attributes + Relationships



3.8M Objects

2.8M Attributes

2.3M Relationships

Questions

1. Q: What's the color of the counter? A: Black.
2. Q: How many drawers can you see? A: Two.
3. Q: What's the material of the pots? A: Metal.

……

1.7M Questions

Region Descriptions

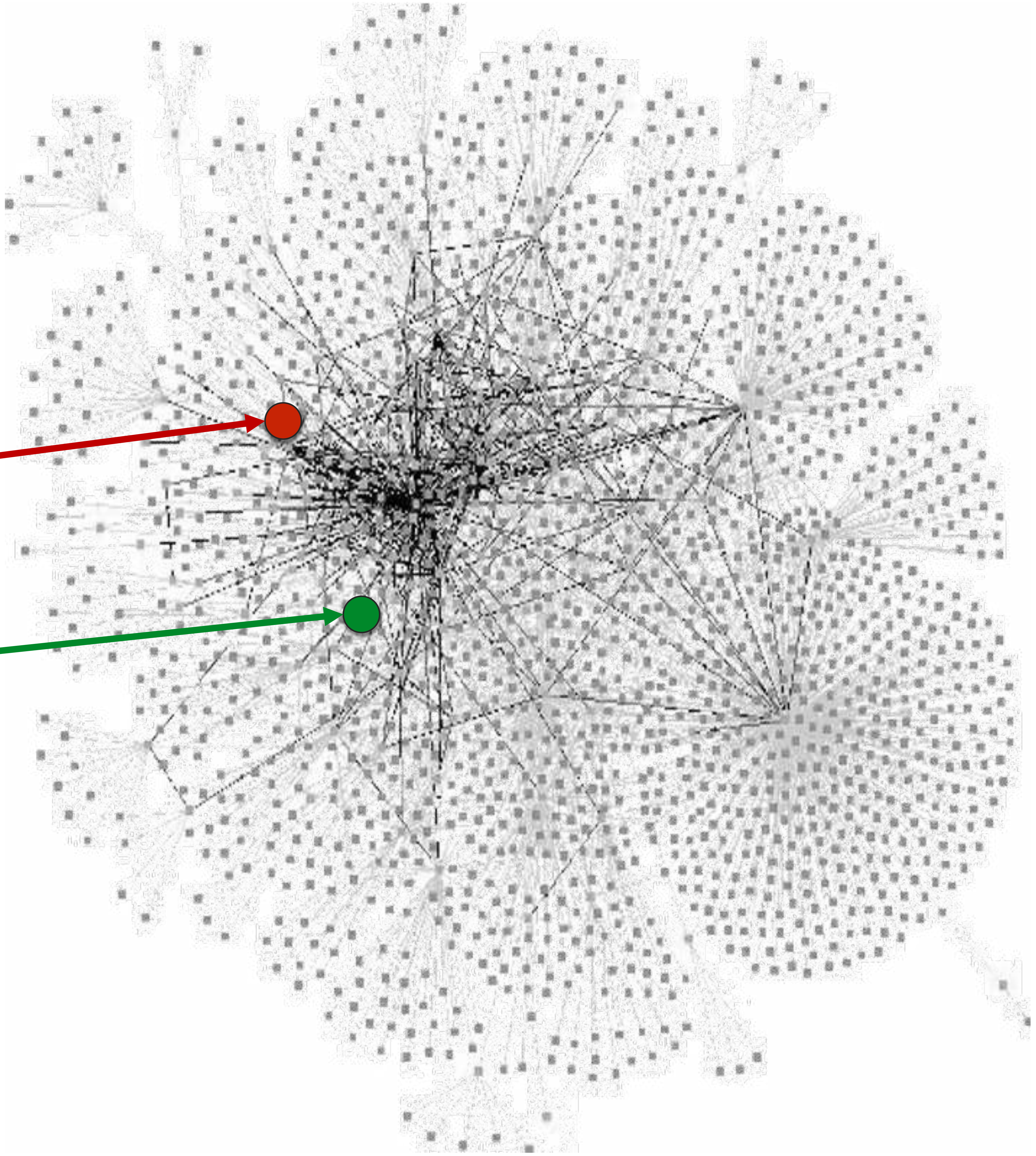
1. There a green bowl on the black counter.
2. The cabinet door is closed.
3. Six knives are placed in the knife holder.

……

5.4M Region Descriptions

Visual Genome

https://visualgenome.org



An ontology of visual concepts

Johnson et al. CVPR'16; Krishna, **Zhu**, et al. IJCV'17

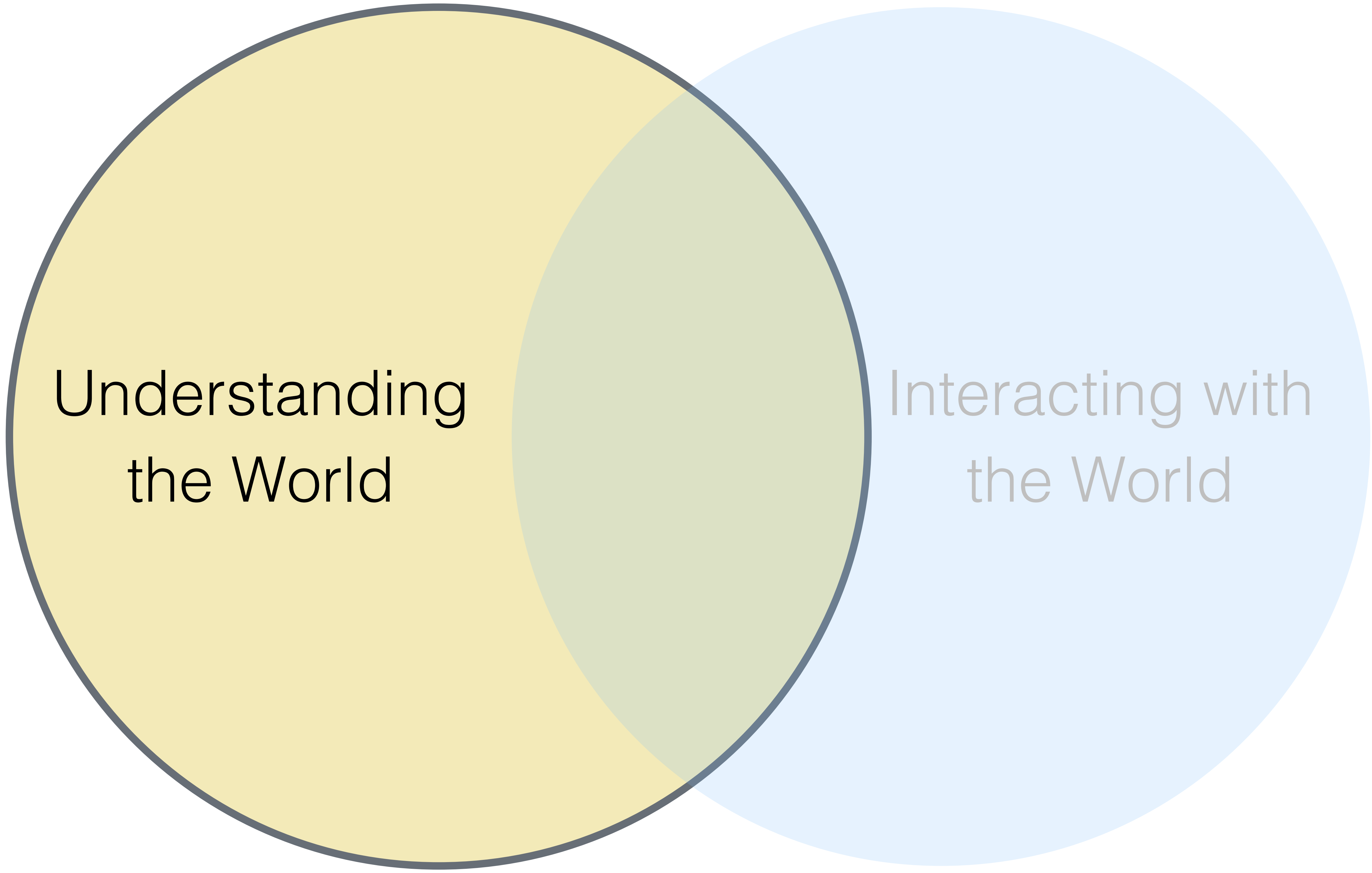
Zhu et al. CVPR’16, **Zhu** et al. CVPR’17

Xu, **Zhu**, Choy, Fei-Fei, CVPR'17

Visual Genome learns Declarative Knowledge from the web.

We built a large-scale visual knowledge base via online crowdsourcing.

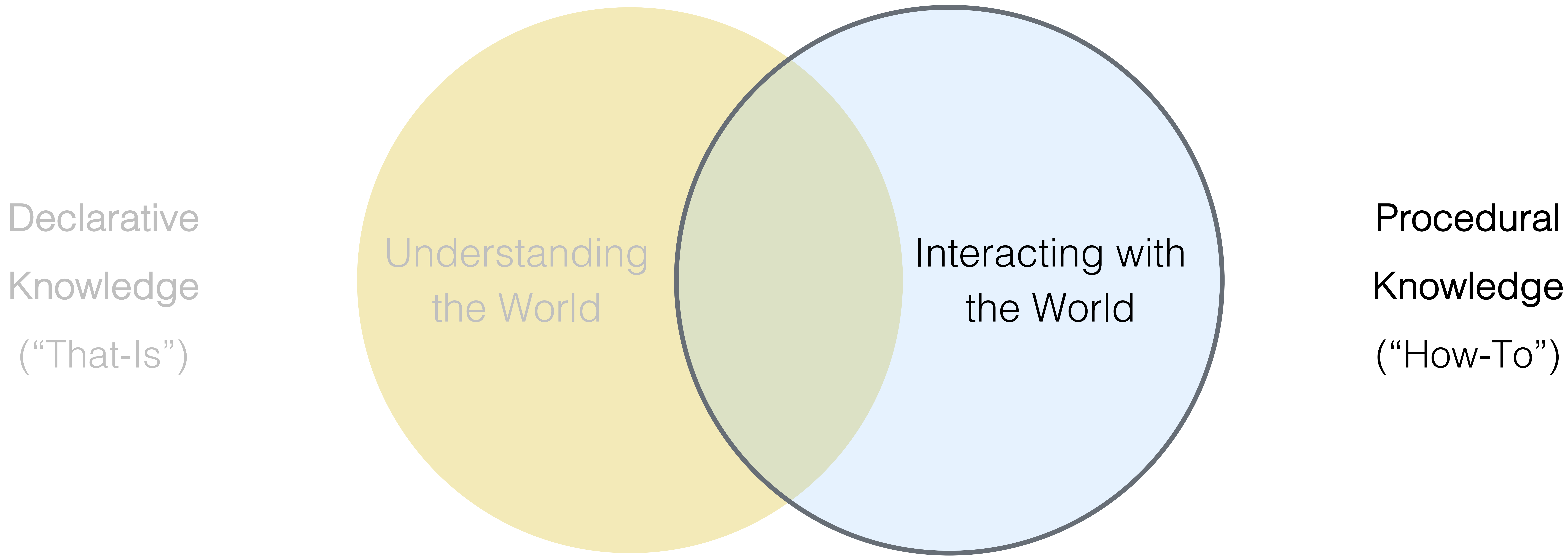
Declarative Knowledge ("That-Is")



Procedural Knowledge ("How-To")

Learning Procedural Knowledge needs new methodology.

It is hard to pinpoint and difficult to verbally described.



Learning Procedural ("How-To") Knowledge from the Web

Three Key Questions

- ❖ What's a good representation of procedural knowledge?
- ❖ How do we learn procedural knowledge from the web?
- ❖ How can robots take advantage of such knowledge?

Part I: Learning from Video Demonstrations

Part II: Learning from Crowd Teleoperation

Part I: Learning from Video Demonstrations

Part II: Learning from Crowd Teleoperation

Web videos supply massive knowledge of how to solve new tasks.

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

TECH \ YOUTUBE \ CULTURE

10

Half of YouTube viewers use it to learn how to do things they've never done

Some of us are on there just to pass the time, though

By Patricia Hernandez | @xpatriciah | Nov 7, 2018, 12:36pm EST

SHARE

how to

how to make slime
how to tie a tie
how to draw
how to basic
how to get boogie down dance
how to cake it
how to train your dragon 3
how to get the galaxy skin in fortnite
how to make slime without glue
how to solve a rubik's cube

Report search predictions

NOVEMBER 7, 2018

Many Turn to YouTube for Children's Content, News, How-To Lessons

An analysis of videos suggested by the site's recommendation engine finds that users are directed toward progressively longer and more popular content

BY AARON SMITH, SKYE TOOR AND PATRICK VAN KESSEL

(MaaHoo Studio/Getty Images)

Source: The Verge, Pew Research Center

Humans learn efficiently from video demonstrations.

Imitation of Televised Models by Infants

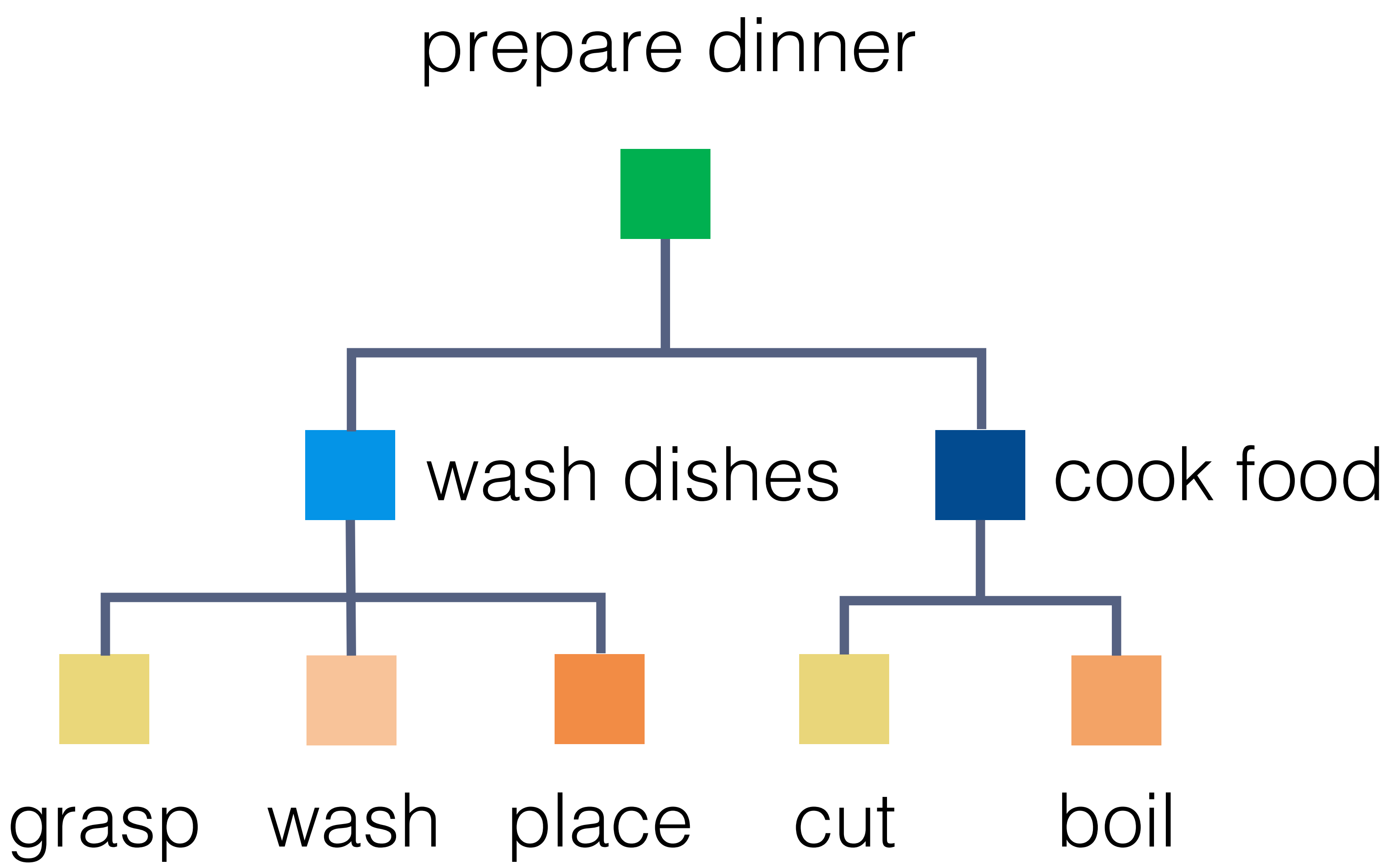
Andrew N. Meltzoff, *Child Development* 1988

Babies (14-24 months) can learn by imitating demonstrations from the TV screen.

Meltzoff & Moore 1977; Meltzoff & Moore 1989, Meltzoff 1988

Our Goal: Learning procedural knowledge as compositional task structures from video demonstrations of a task

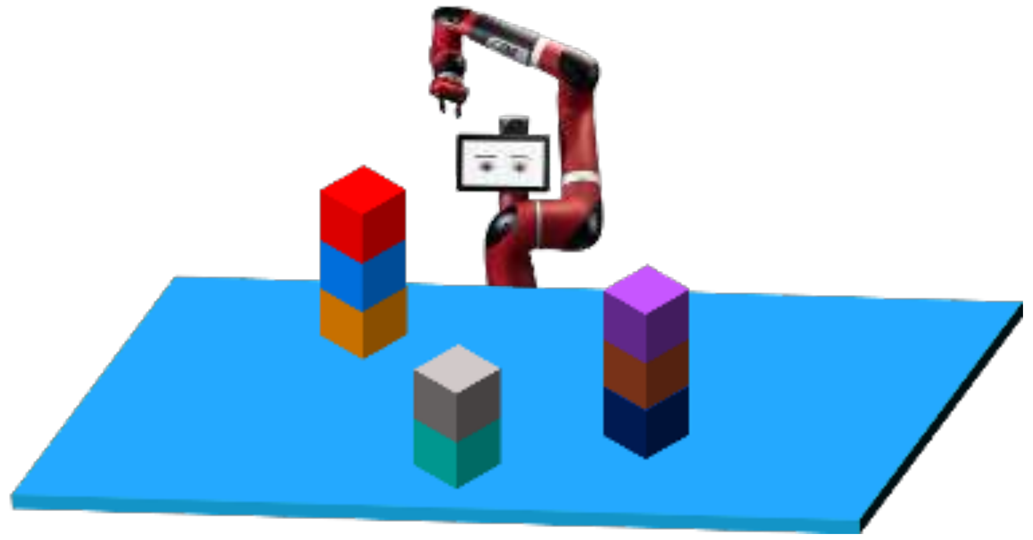
prepare dinner



One-Shot Imitation Learning from Videos

single video
demonstration

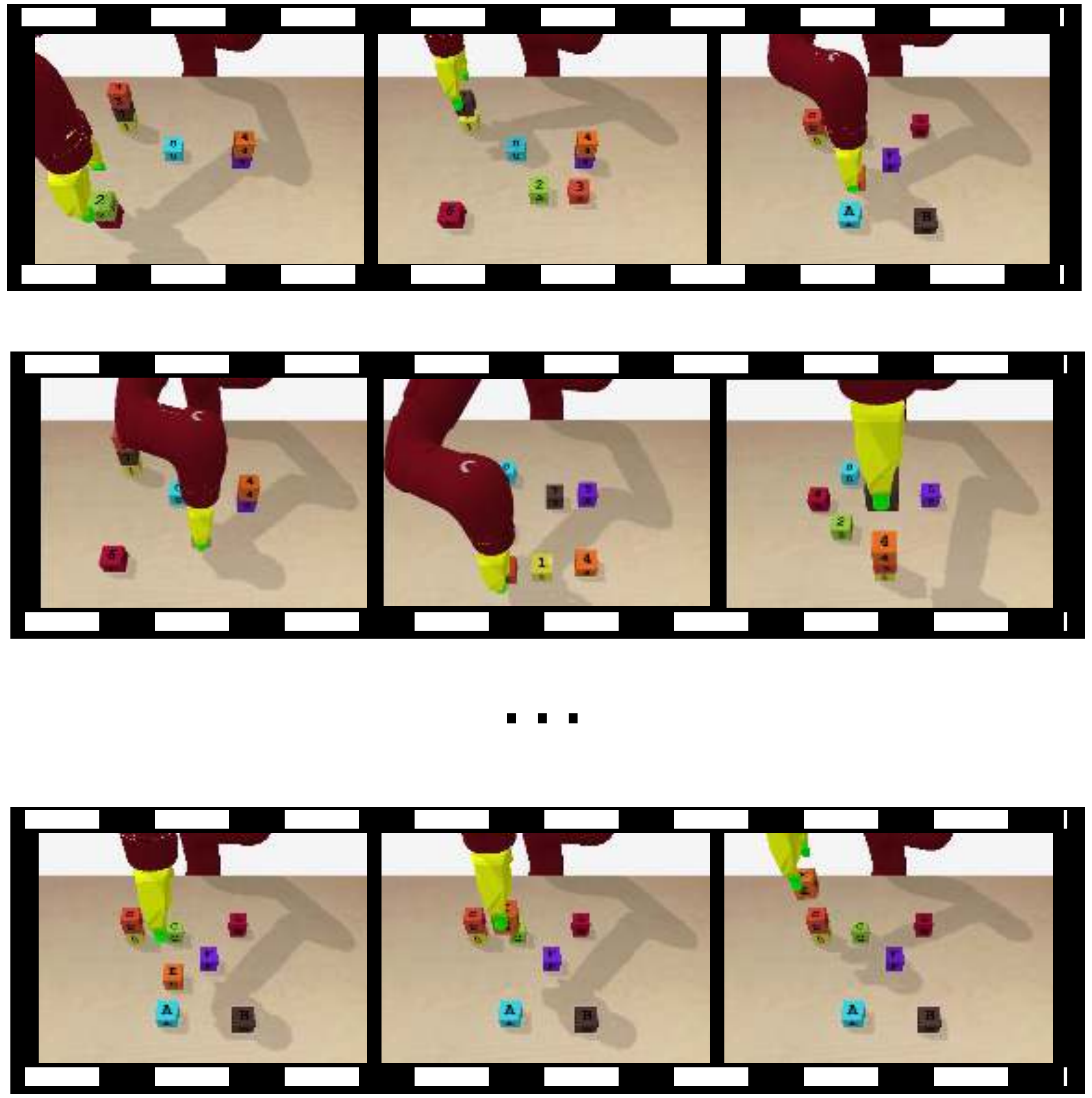
meta-learning
model



policy for the
demonstrated task

Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

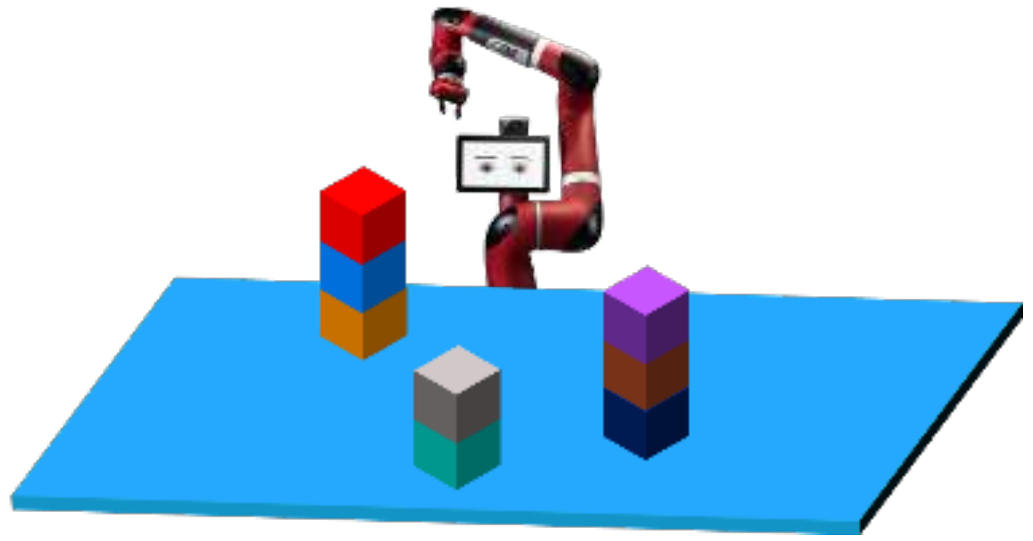
One-Shot Imitation Learning from Videos



a lot of training videos
(seen tasks)

supervision

meta-learning
model



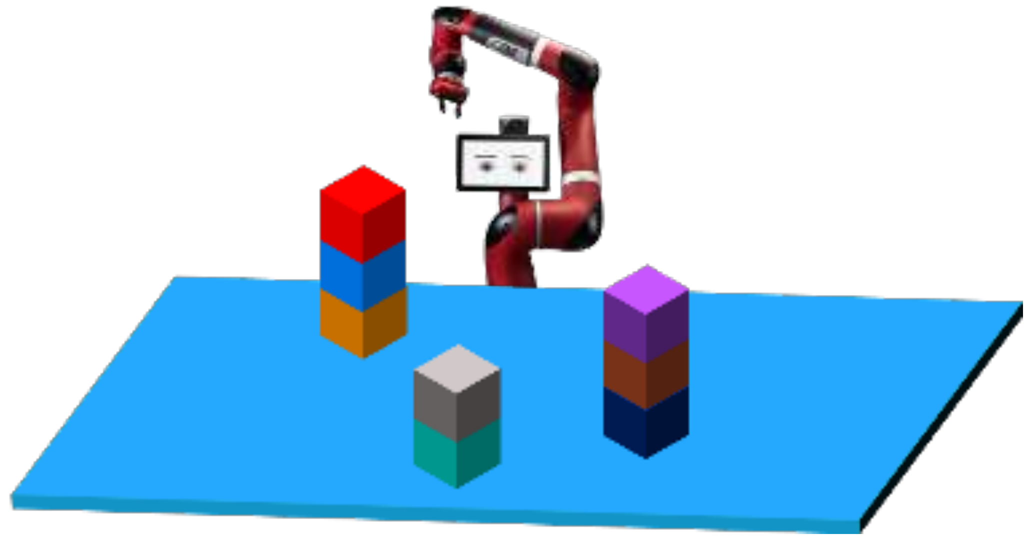
policy for the
demonstrated task

Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

One-Shot Imitation Learning from Videos

single test video
(unseen task)

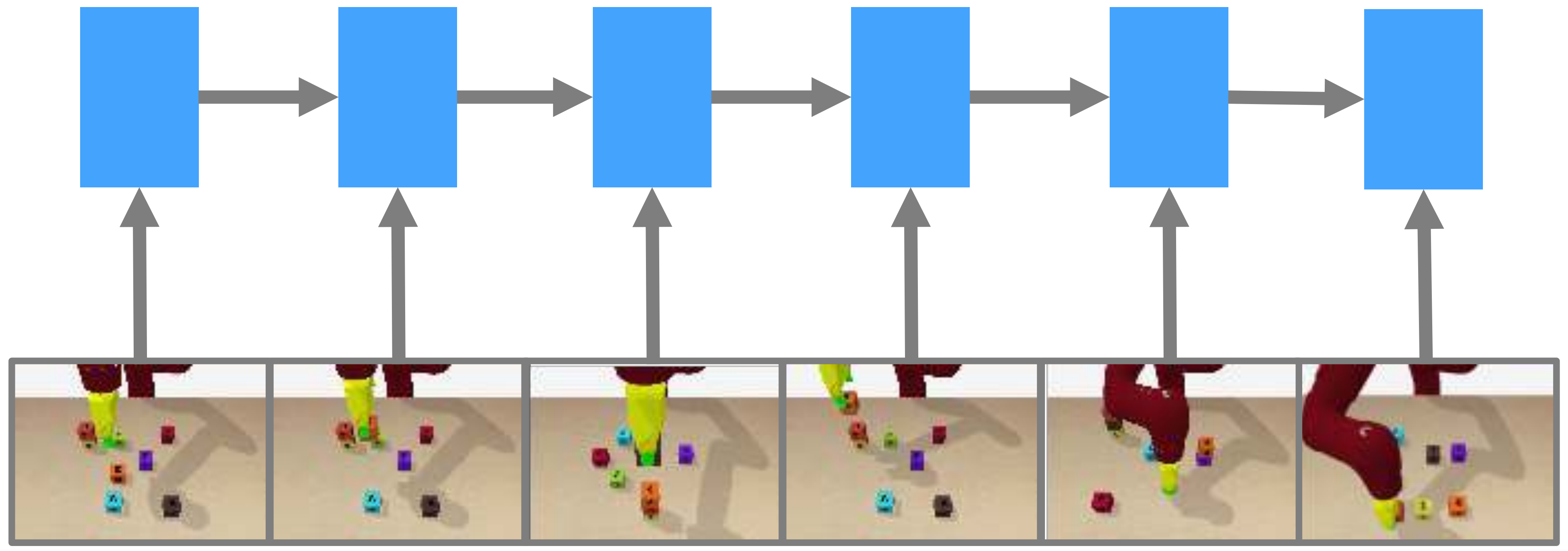
meta-learning
model



policy for the
demonstrated task

Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

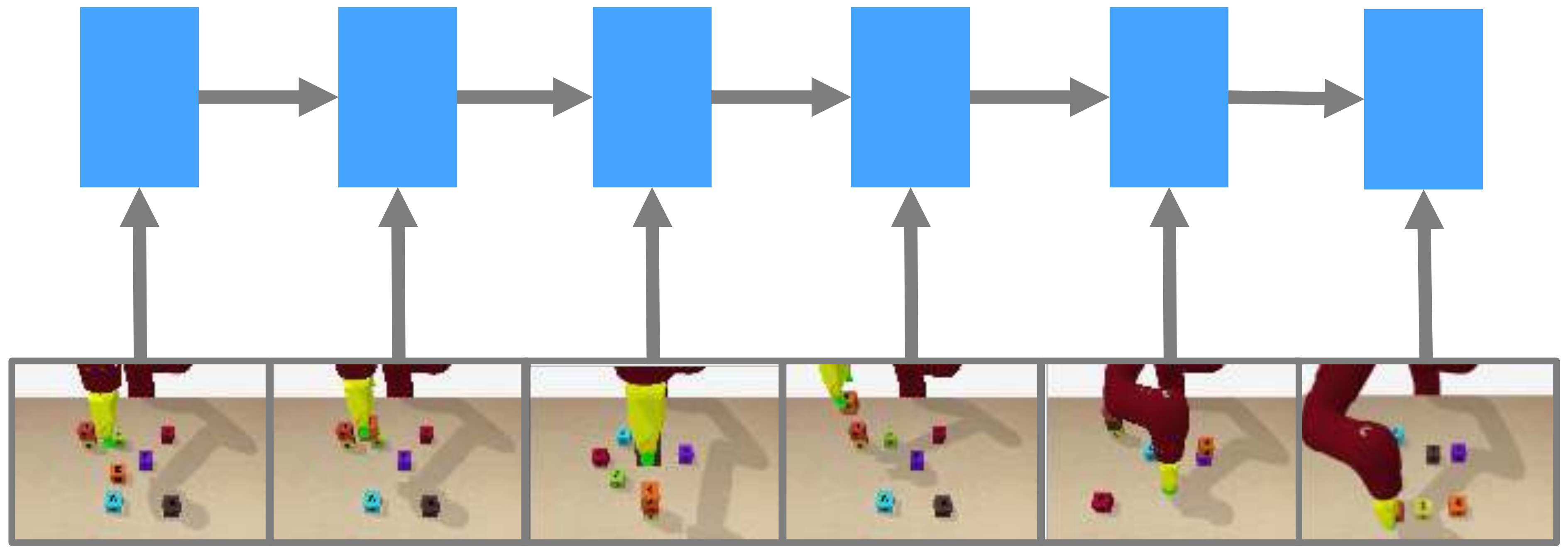
One-Shot Imitation Learning from Videos



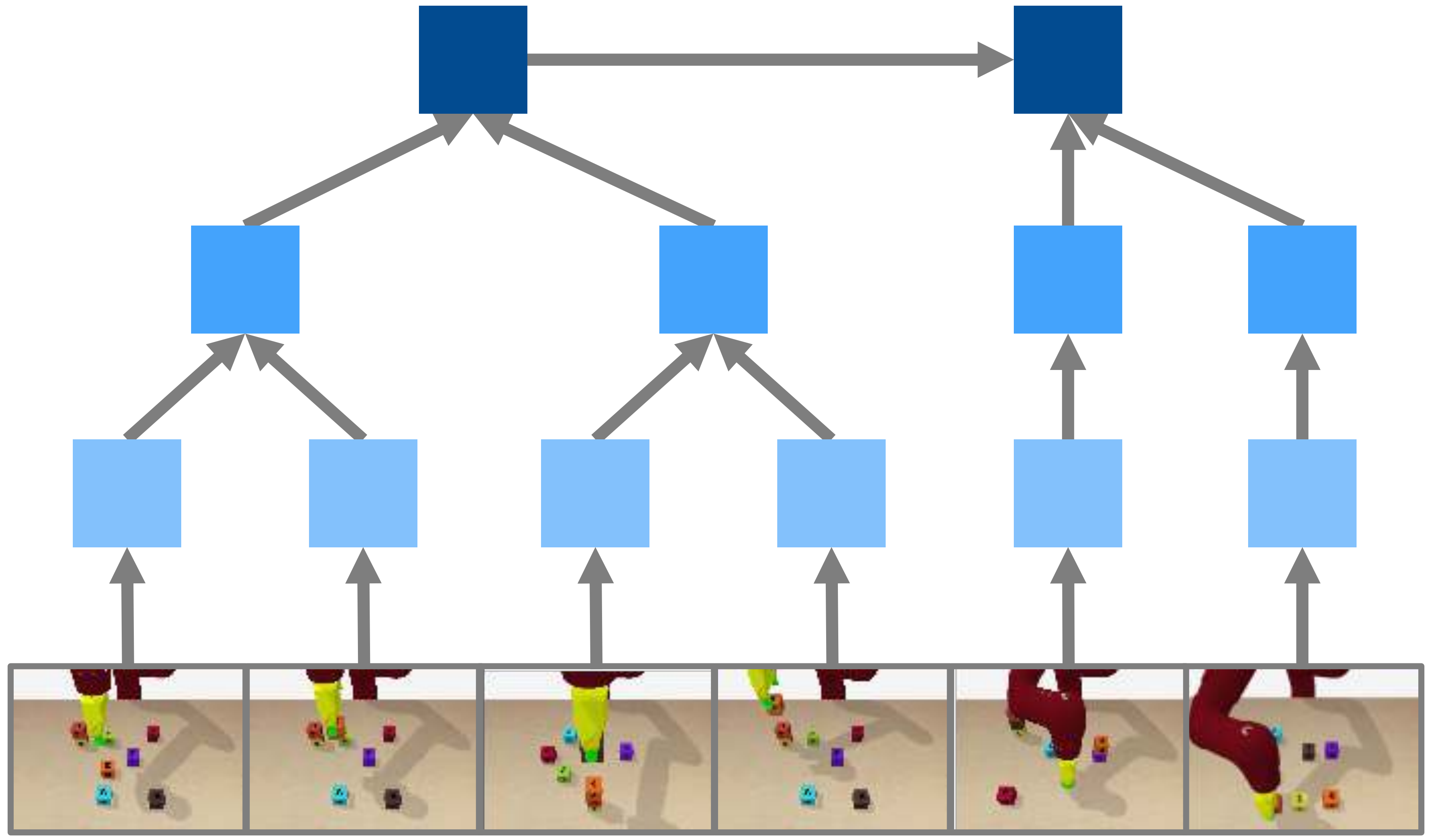
[Duan et al. 17; Finn et al. 2017; Wang et al. 2017; Yu et al. 2018]

modeling demonstration
as a **flat sequence**

One-Shot Imitation Learning from Videos



[Duan et al. 17; Finn et al. 2017; Wang et al. 2017; Yu et al. 2018]

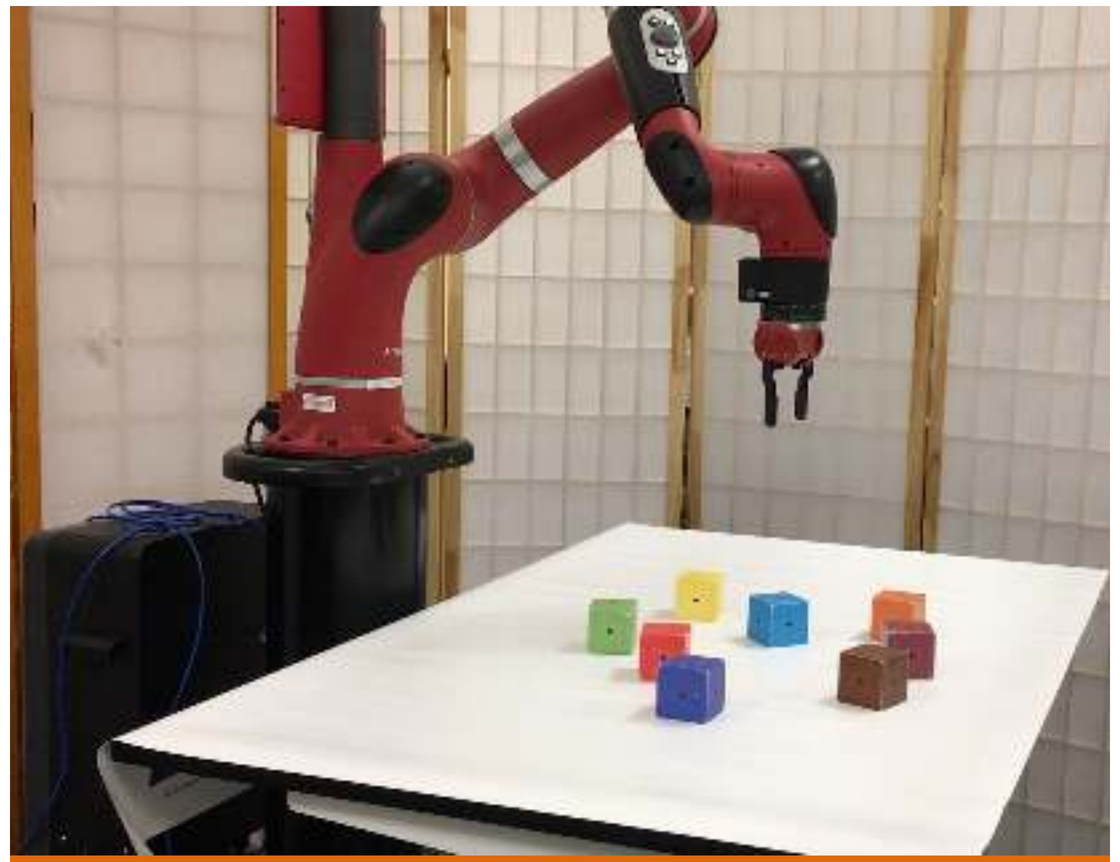


modeling demonstration
as a **flat sequence**

modeling demonstration
as a **compositional structure**

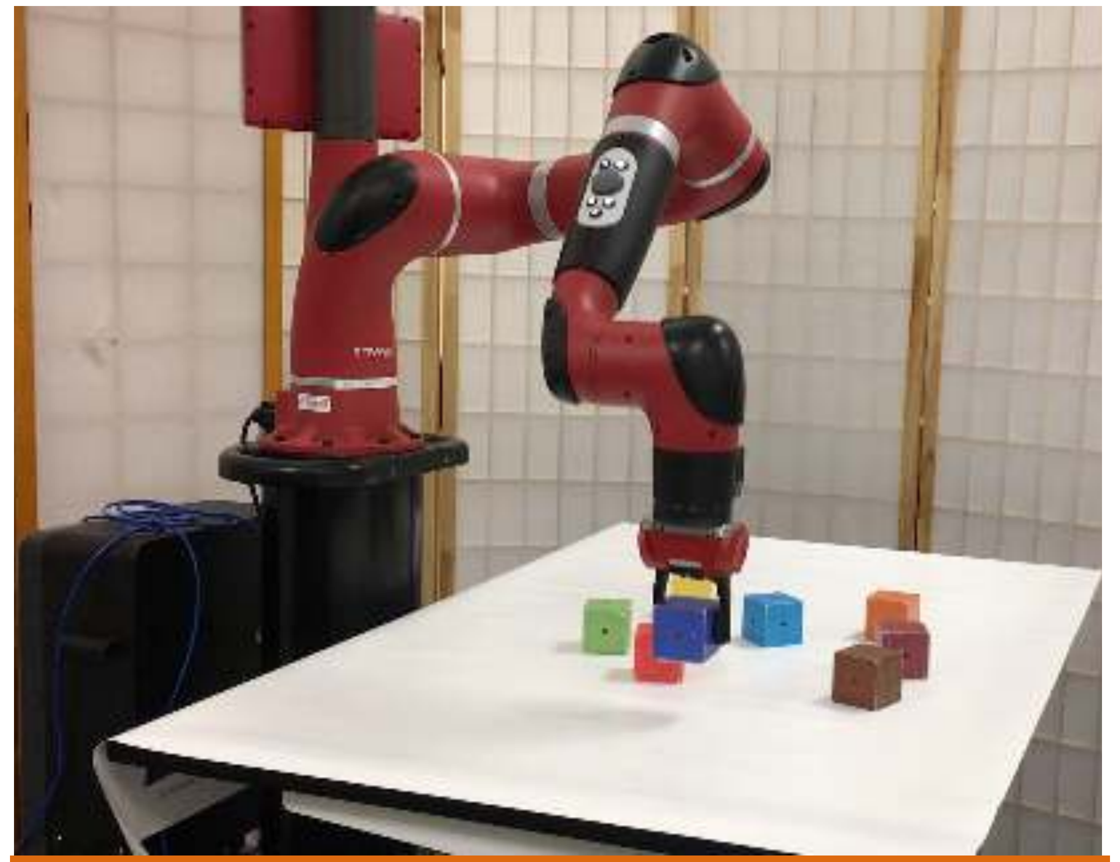
Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

Neural Task Programming (NTP): Hierarchical Policy Learning as Neural Program Induction



Move_to (Blue)

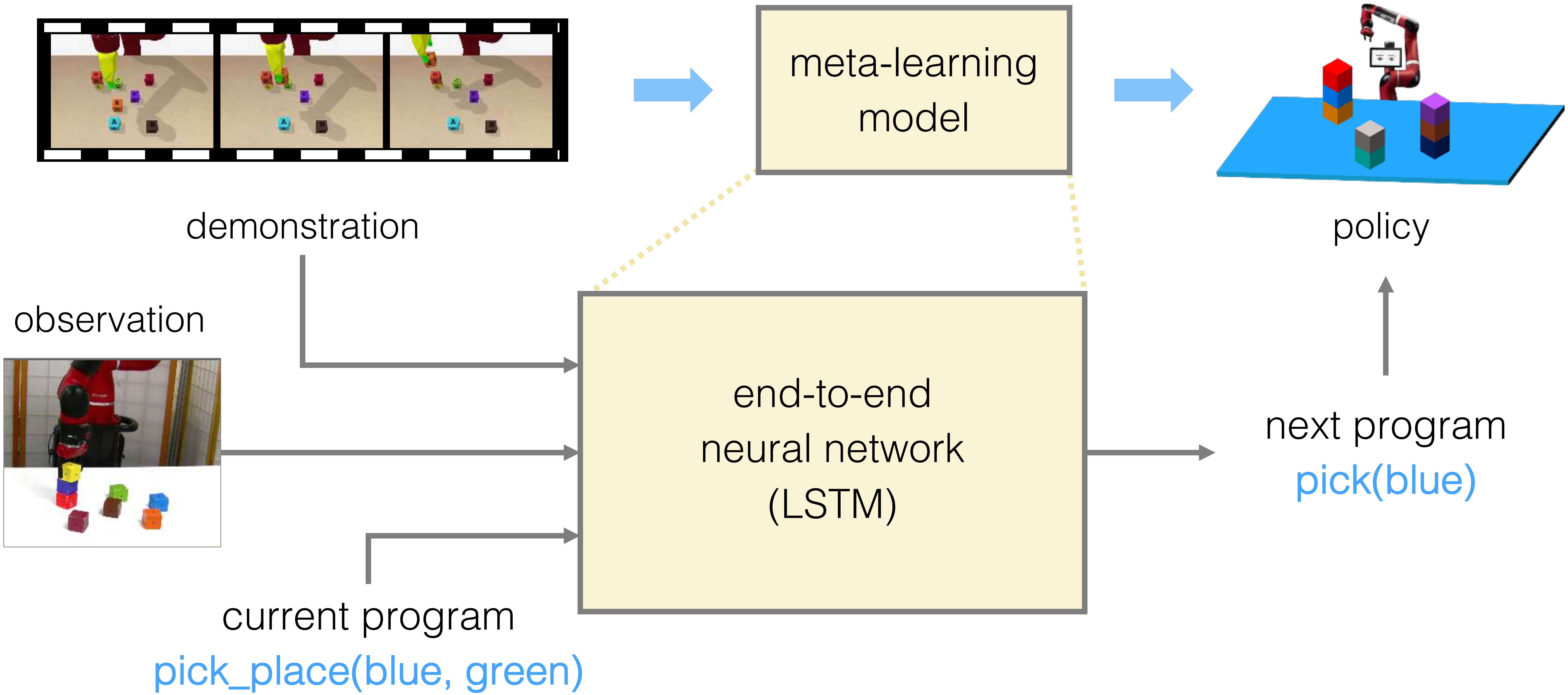
Grip (Blue)



Move_to (Red)

Release( )

One-Shot Imitation Learning from Videos: Neural Task Programming (NTP)



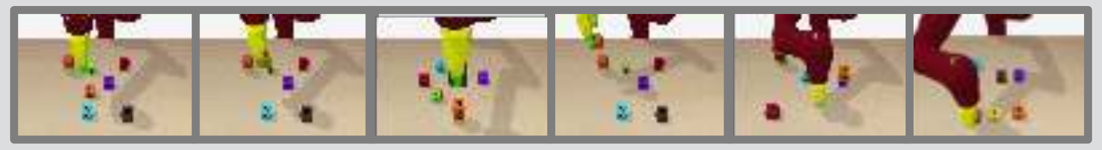
Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

One-Shot Imitation Learning from Videos: Neural Task Programming (NTP)

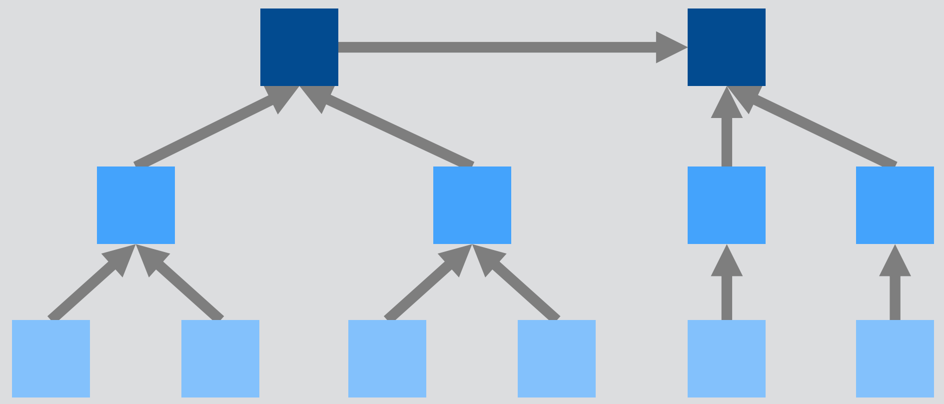
demonstration

Training supervision

{(



,



)}

video demonstration

hierarchical program trace

observation

end-to-end
neural network
(LSTM)

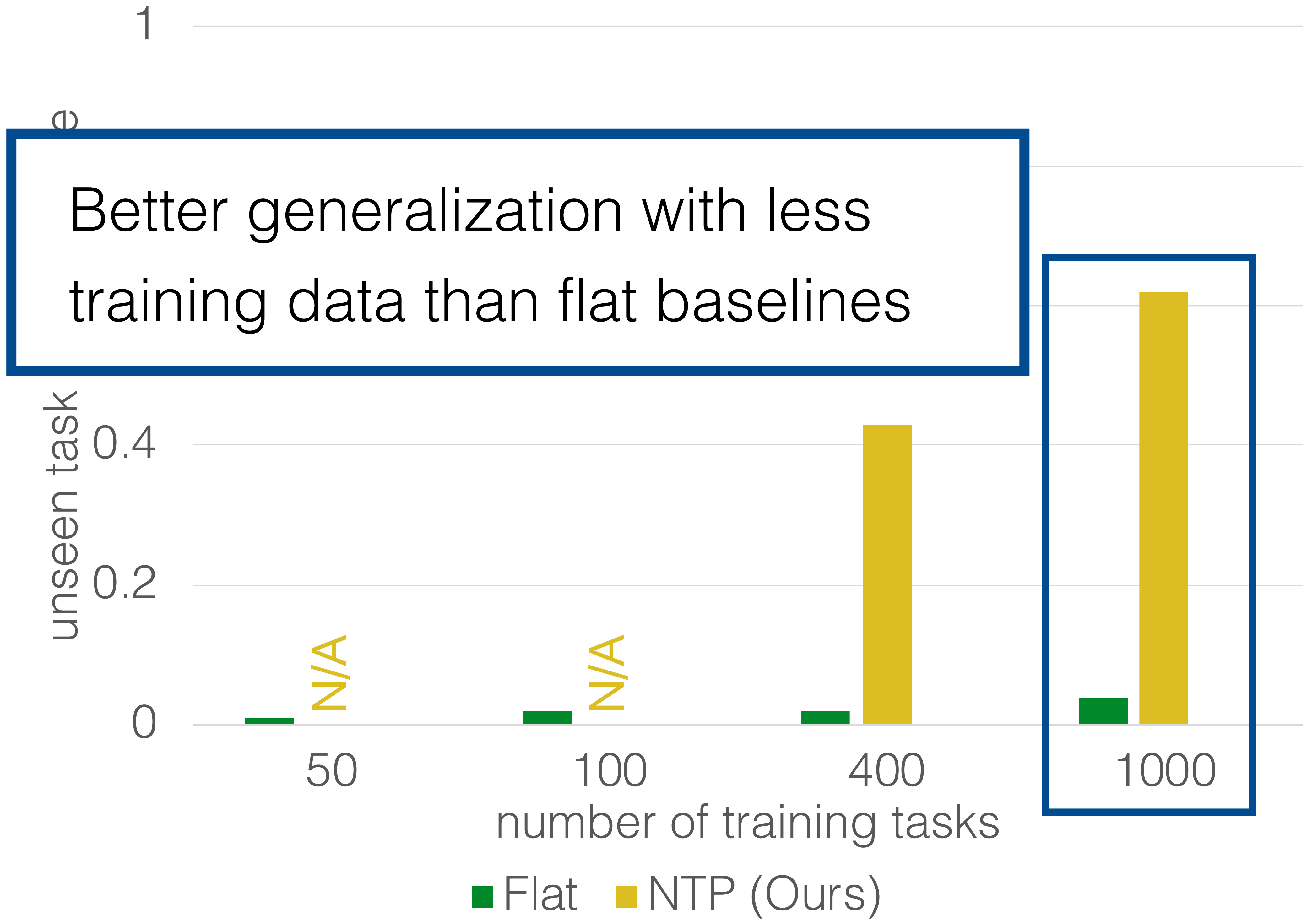
next program
pick(blue)

current program
pick_place(blue, green)

Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

One-Shot Imitation Learning from Videos: Neural Task Programming (NTP)

Qualitative



Quantitative

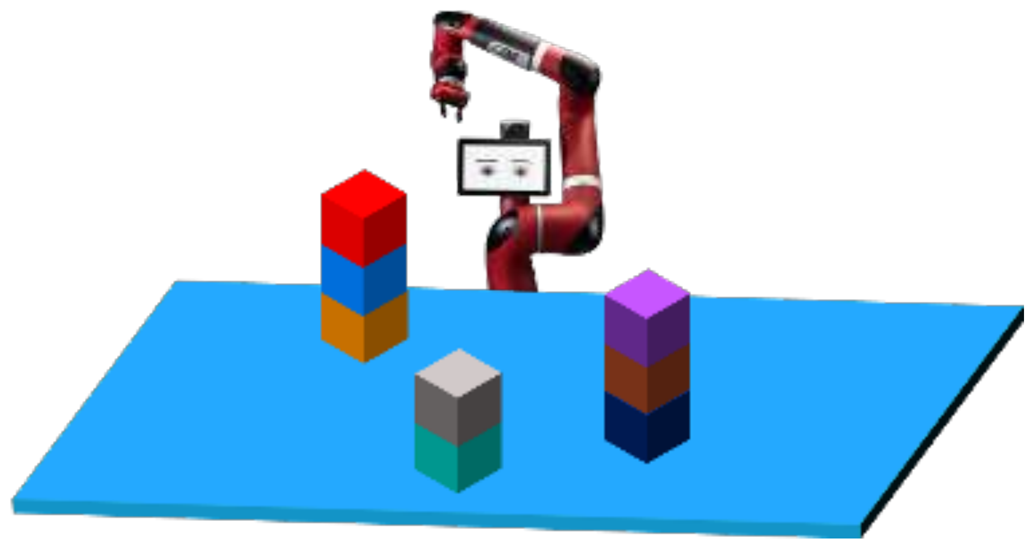
(the higher the better)

Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

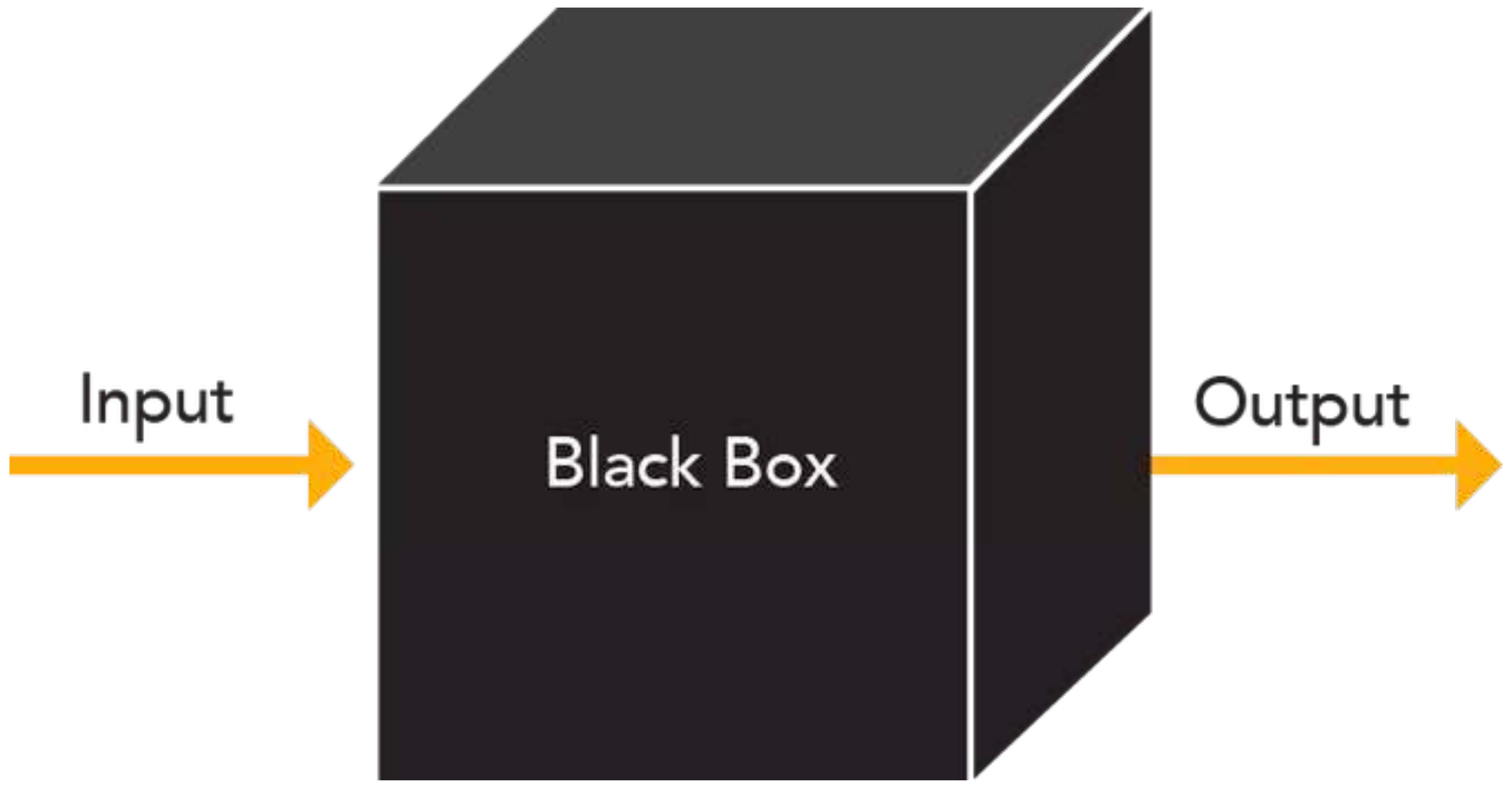
One-Shot Imitation Learning from Videos: Neural Task Programming (NTP)

demonstration

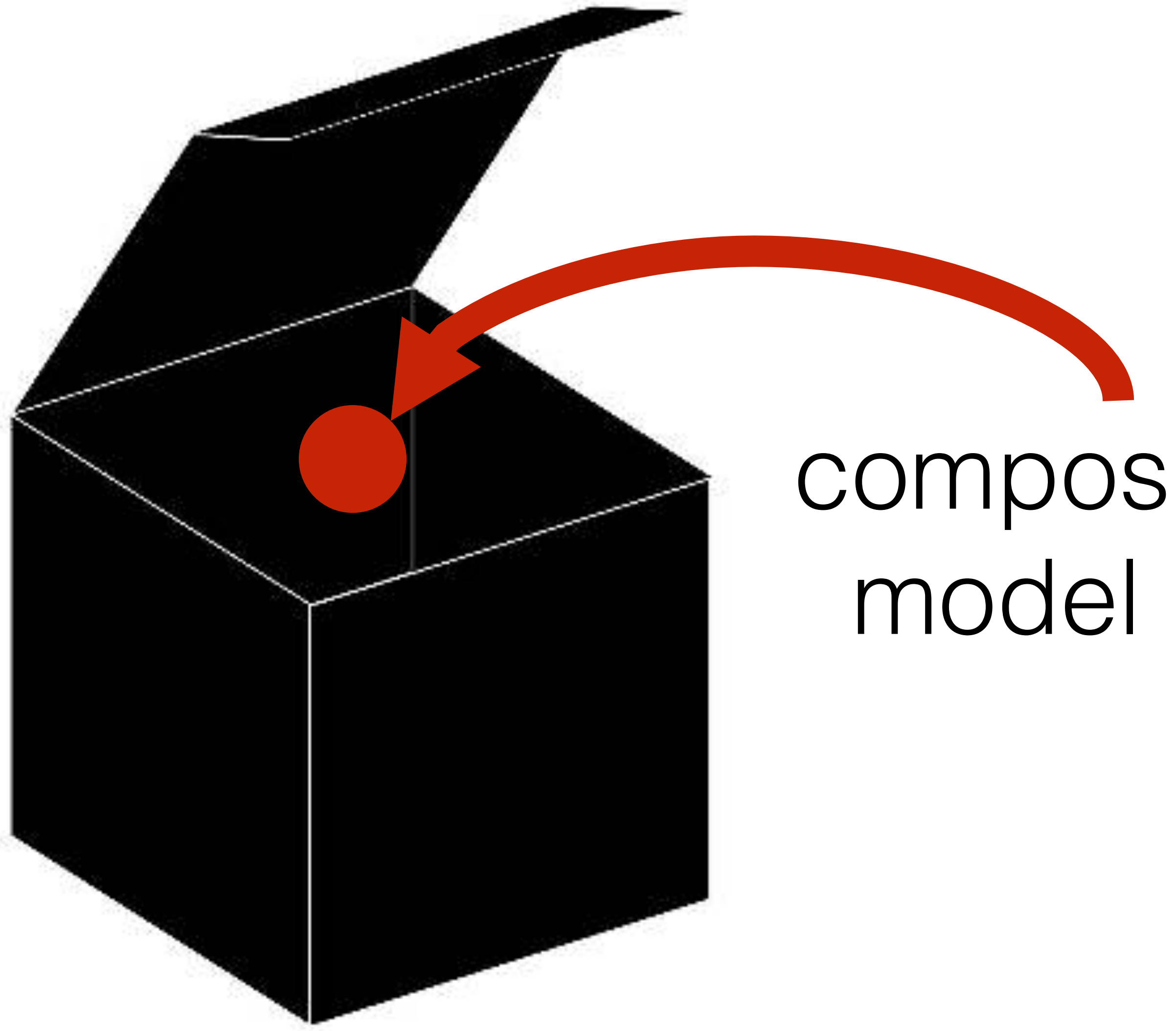
meta-learning model



policy



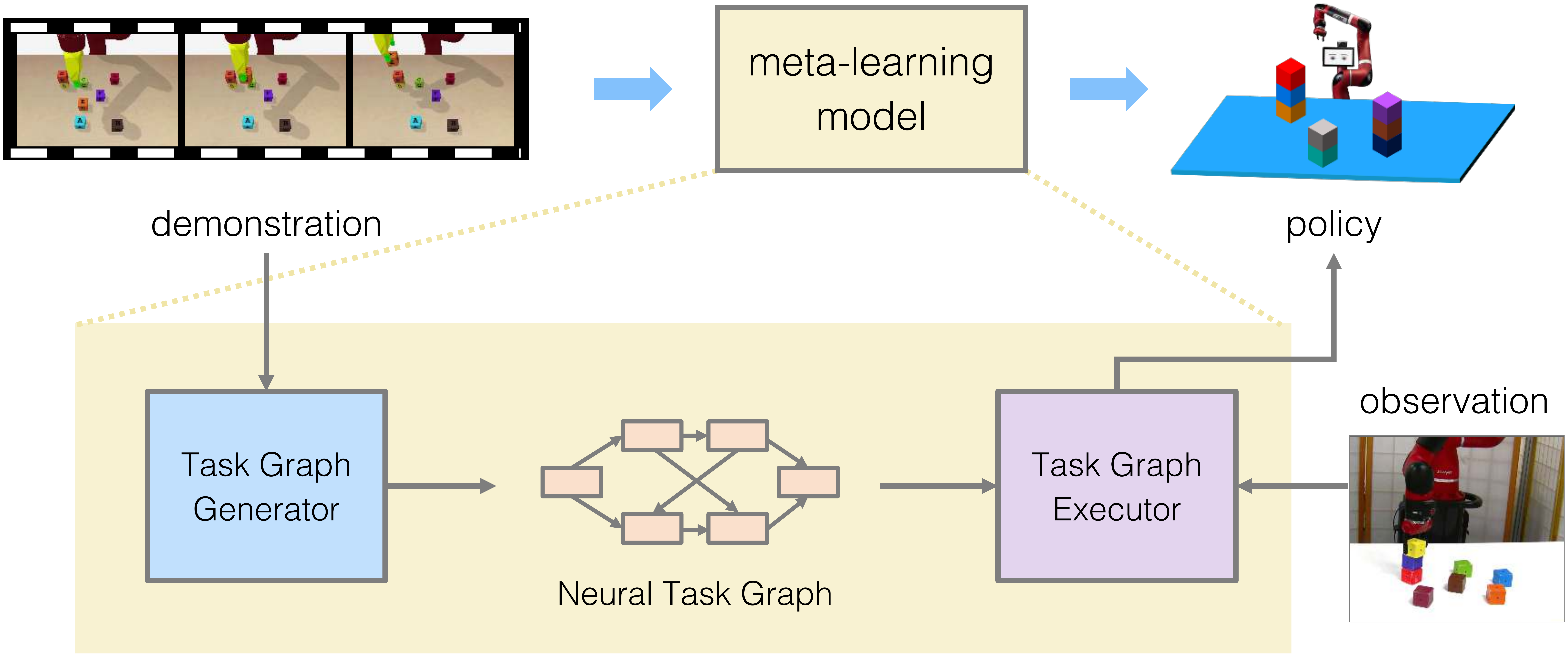
end-to-end
neural network
(LSTM)



compositional
model prior

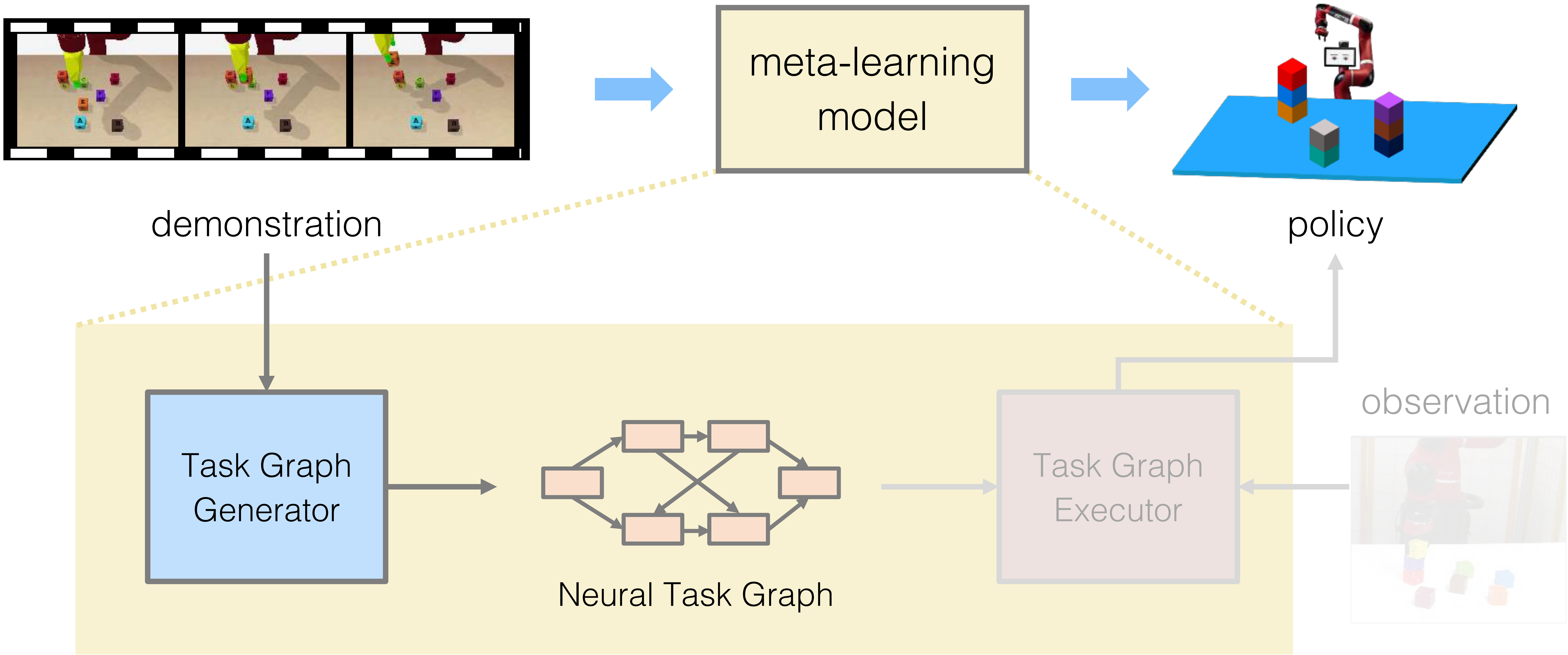
Xu*, Nair*, **Zhu**, Gao, Garg, Fei-Fei, Savarese. *ICRA 2018*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



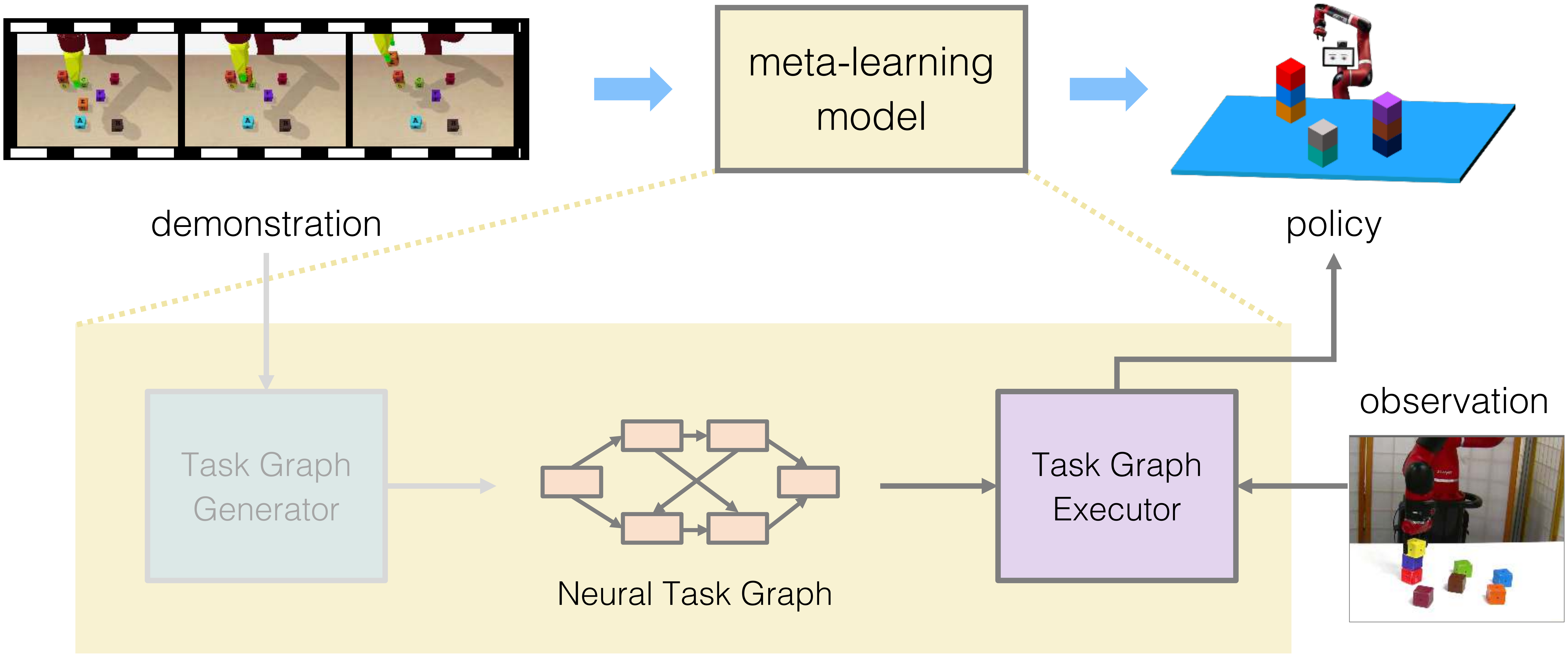
Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

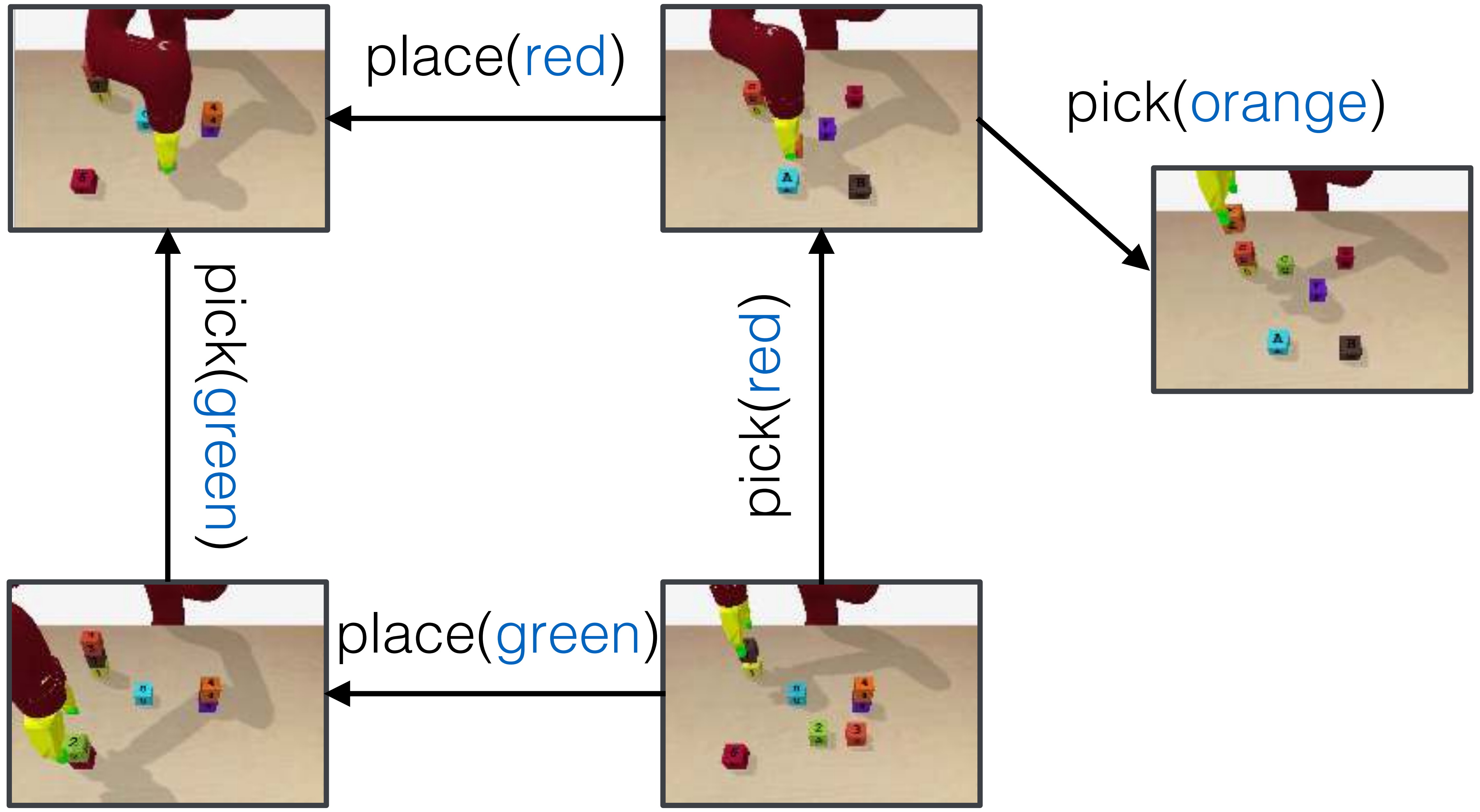
One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

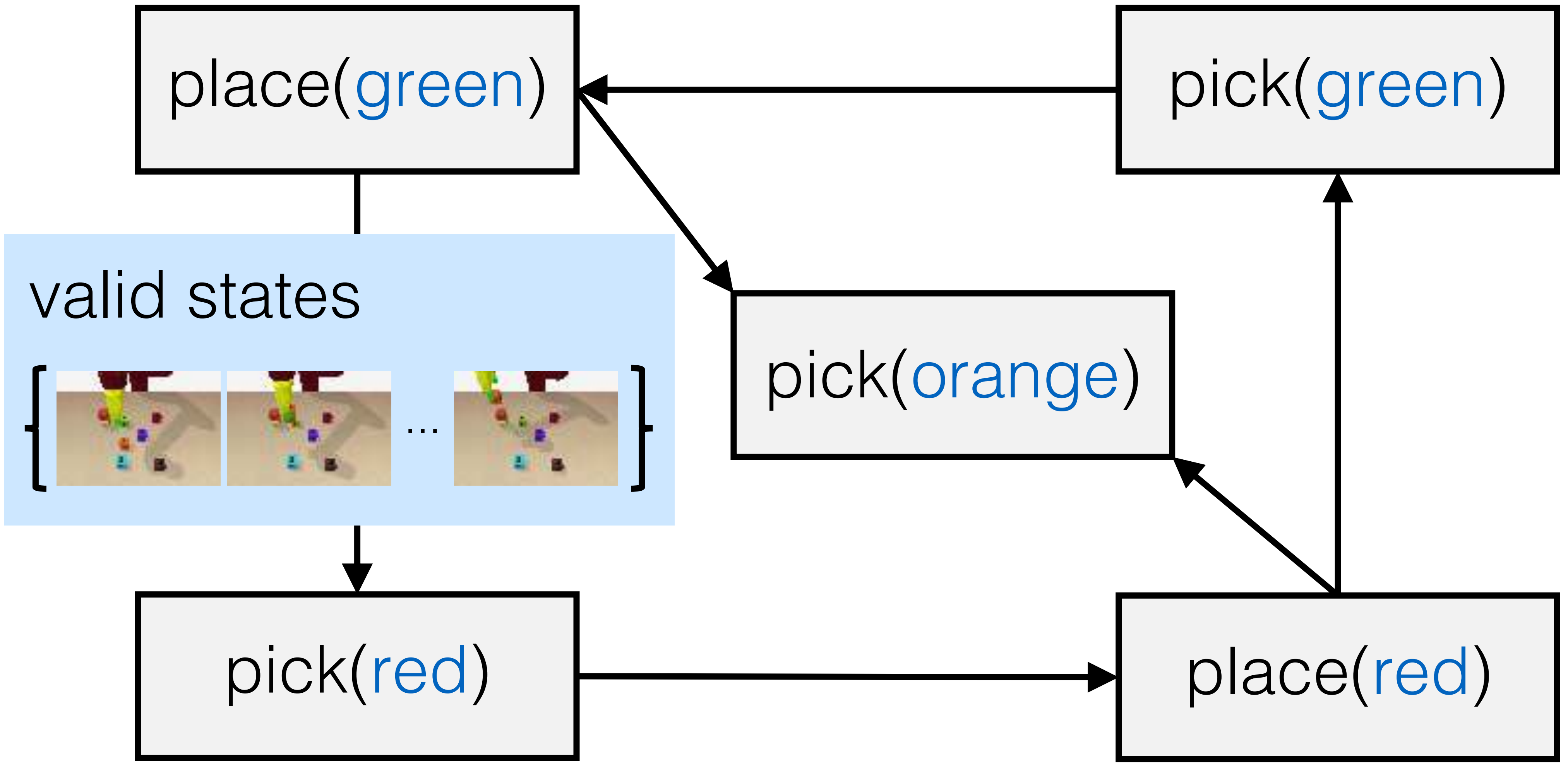
One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)

Task Graph



Nodes	States	infinite
Edges	Actions	

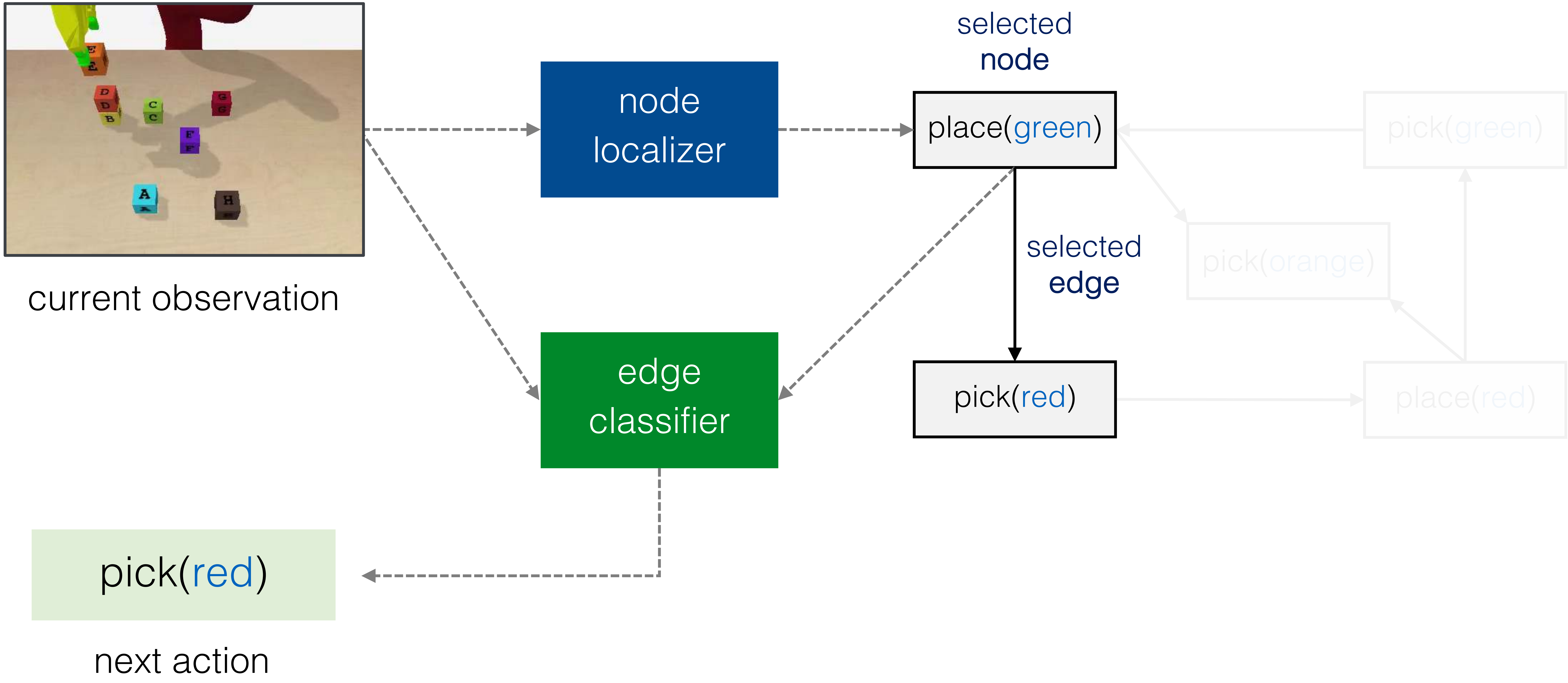
Conjugate Task Graph



Nodes	Actions	finite
Edges	States (Preconditions)	

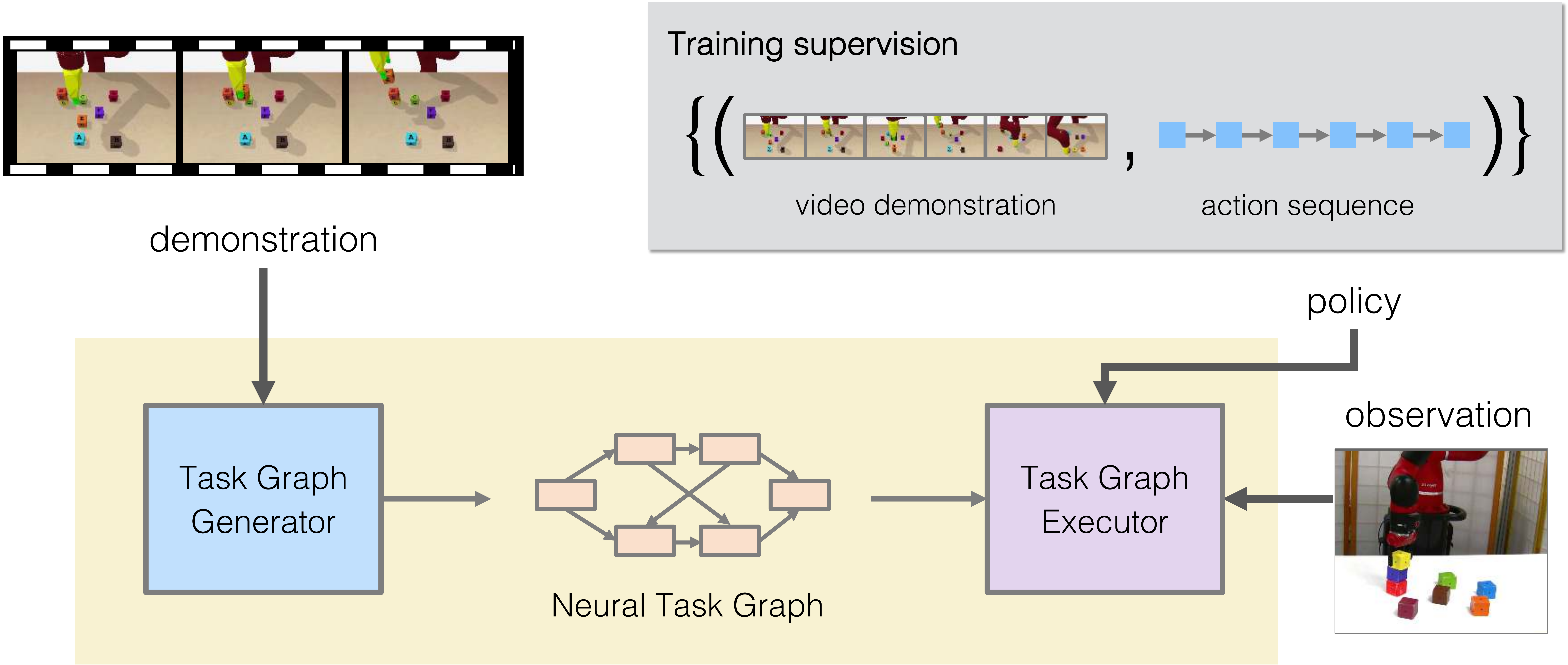
Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



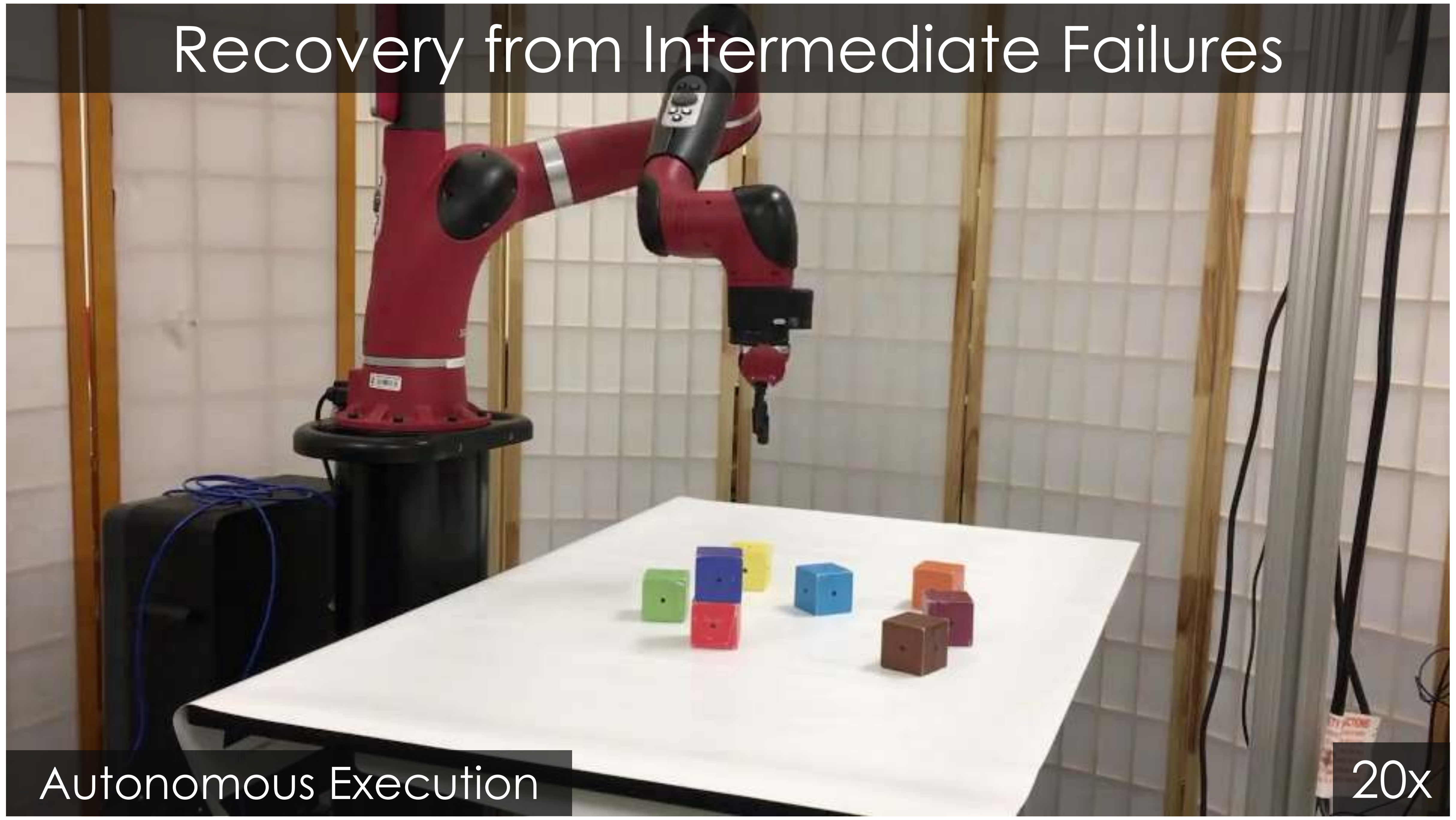
Huang*, Nair*, Xu*, Zhu, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)

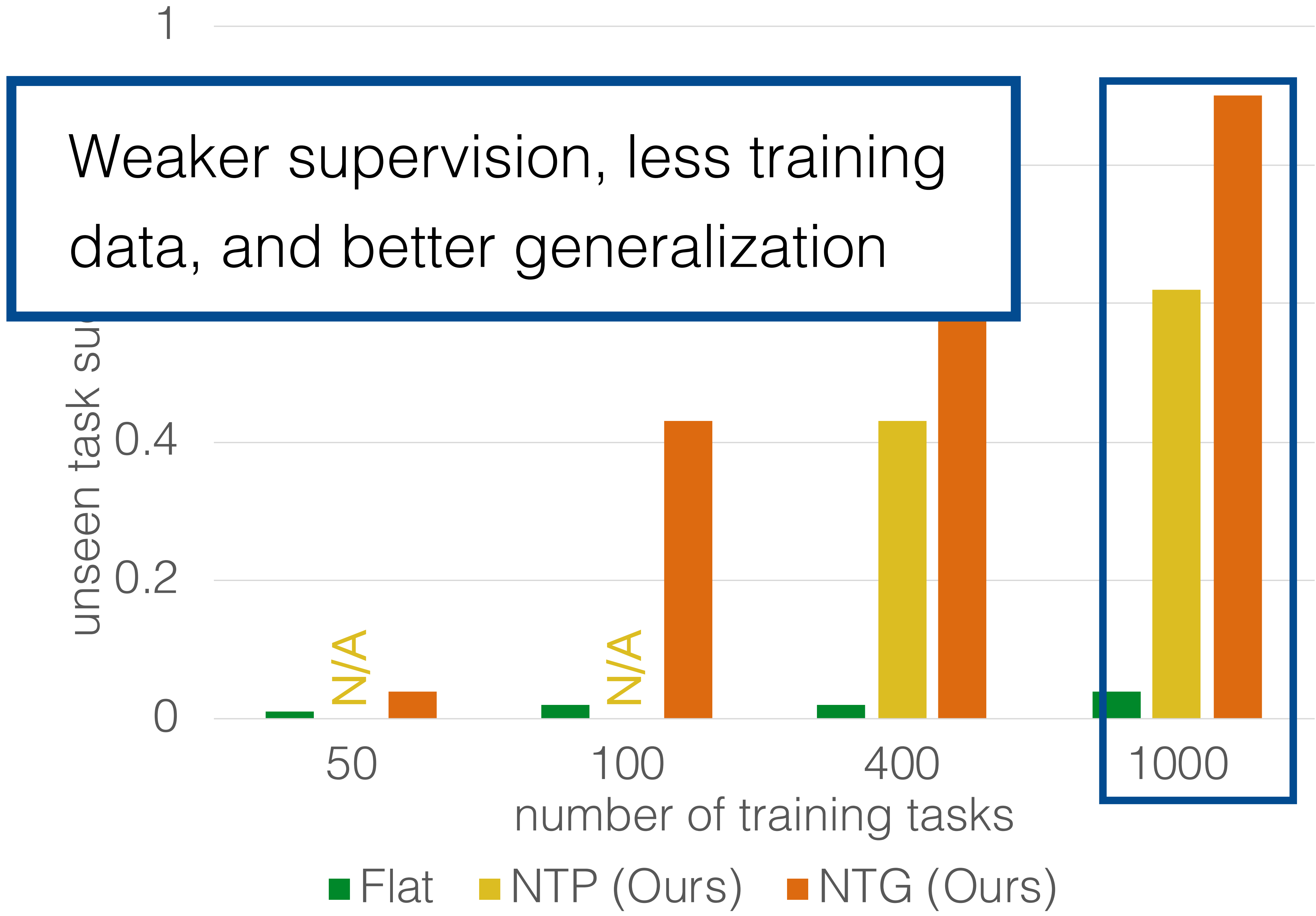


Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



Qualitative

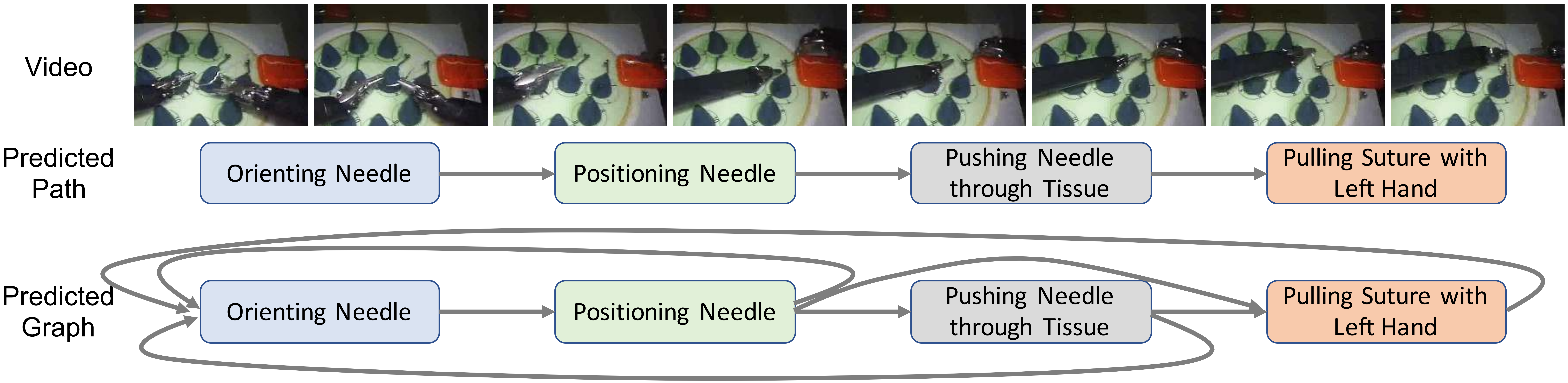


Quantitative

(the higher the better)

Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

One-Shot Imitation Learning from Videos: Neural Task Graphs (NTG)



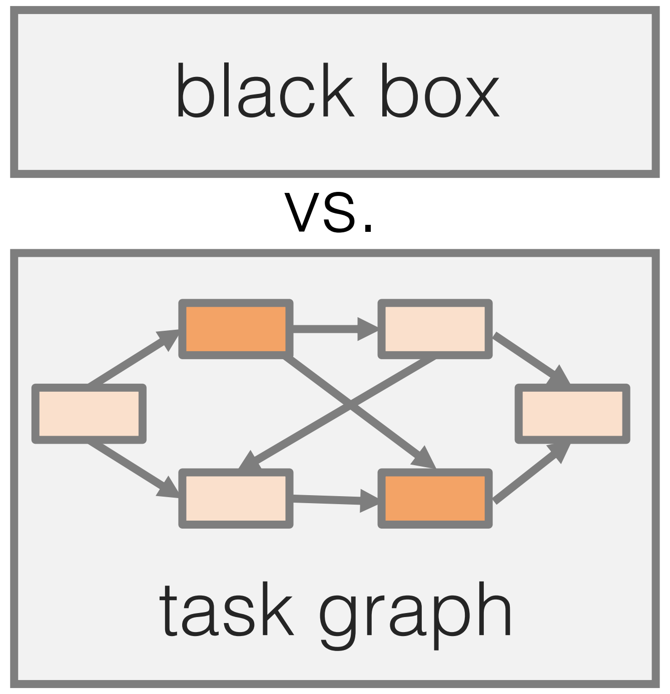
Applying NTG to the real-world surgical video dataset JIGSAWS

Huang*, Nair*, Xu*, **Zhu**, Garg, Fei-Fei, Savarese, Niebles. *CVPR 2019*

Next Goal: Learning task knowledge from web videos

Summary - Part I

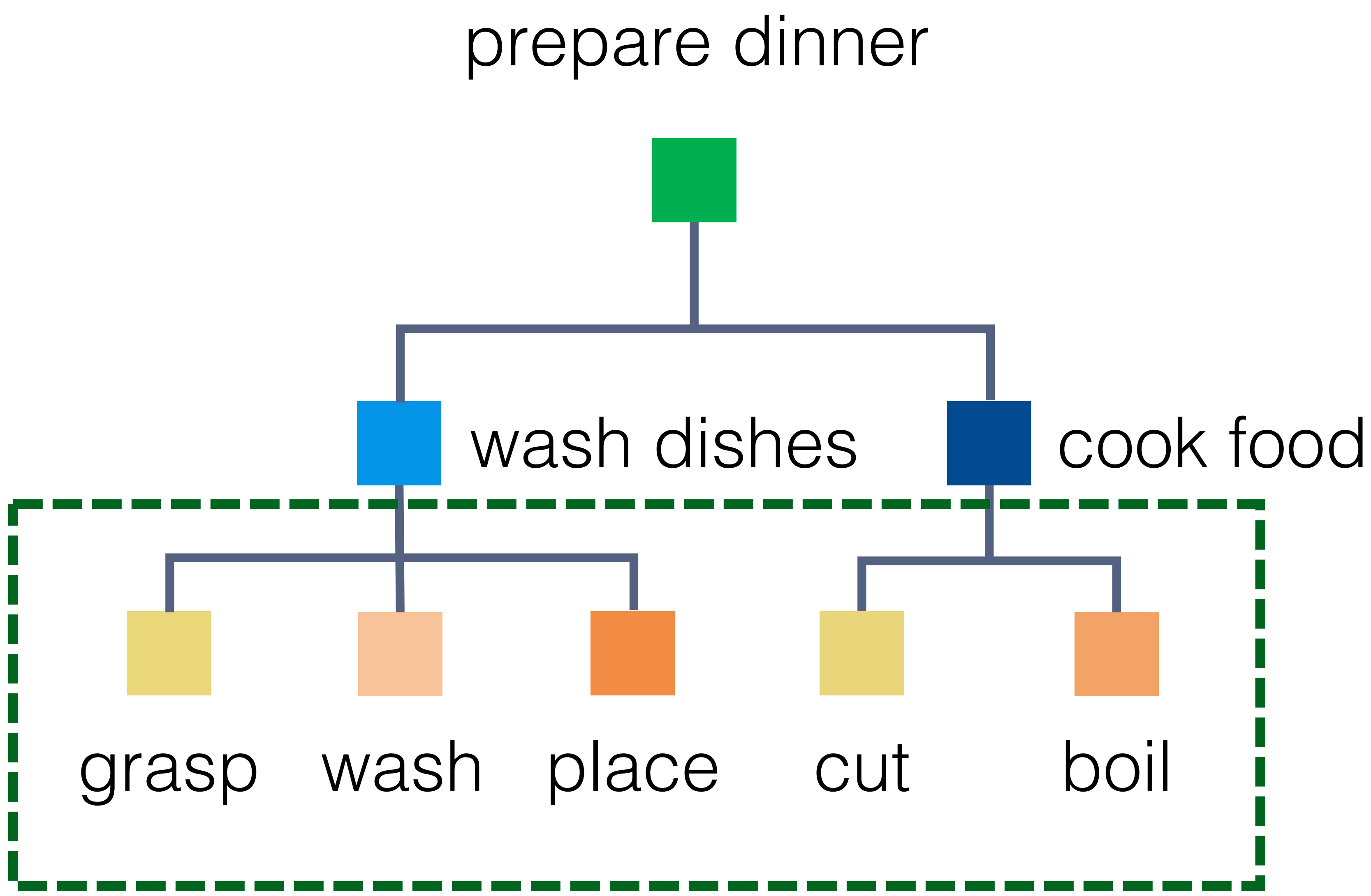
Extracting how-to knowledge about the **compositional task structure** of complex tasks from **video demonstrations**



Meta-learning models with compositional priors generalize better than black-box models

NTP and NTG learn how-to knowledge in the form of compositional task structures while motor skills are abstracted away.

prepare dinner



modeled as pre-defined “API calls”

NTP and NTG learn how-to knowledge in the form of compositional task structures while motor skills are abstracted away.

Manually defining motor skills is intractable.
We need to learn from data.

How can we collect data for learning motor skills from the web?

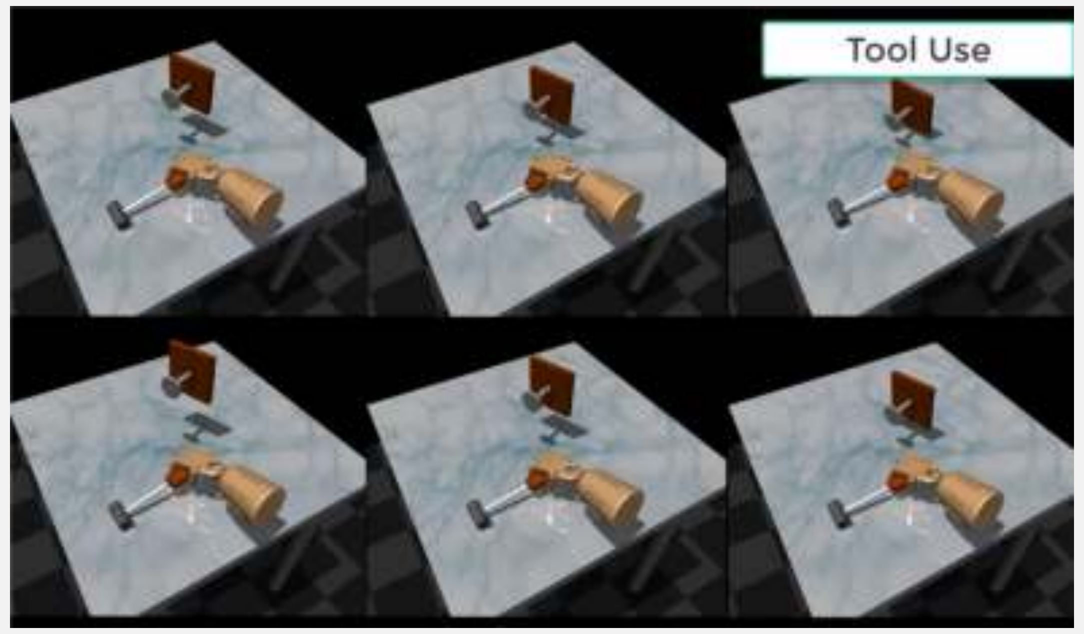
Part I: Learning from Video Demonstrations

Part II: Learning from Crowd Teleoperation

Data is critical for learning robot motor skills.

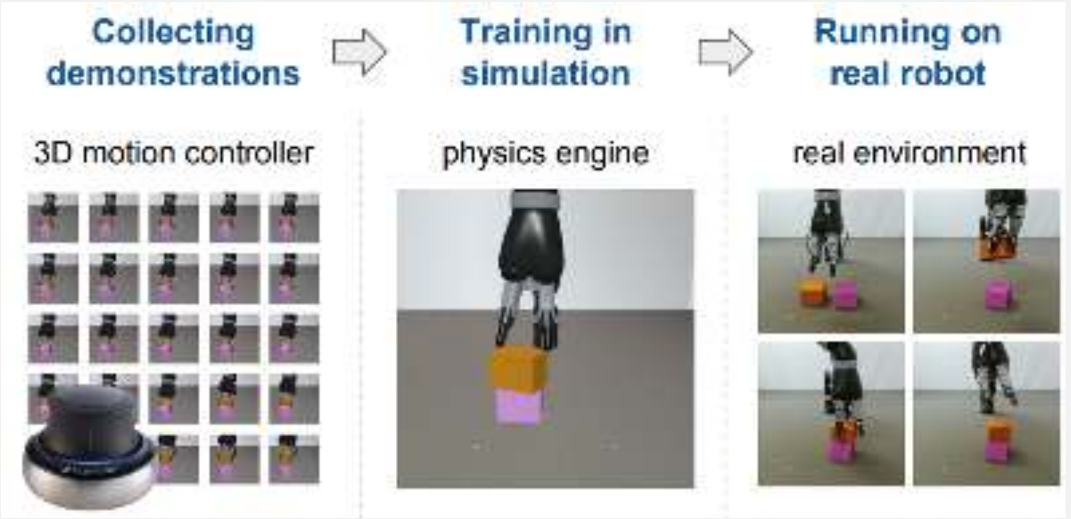
Imitation Learning

Rajeswaran et al. 2018: 25 demos



Finn et al. 2017: 30 demos

Zhu et al. 2018: 30 demos



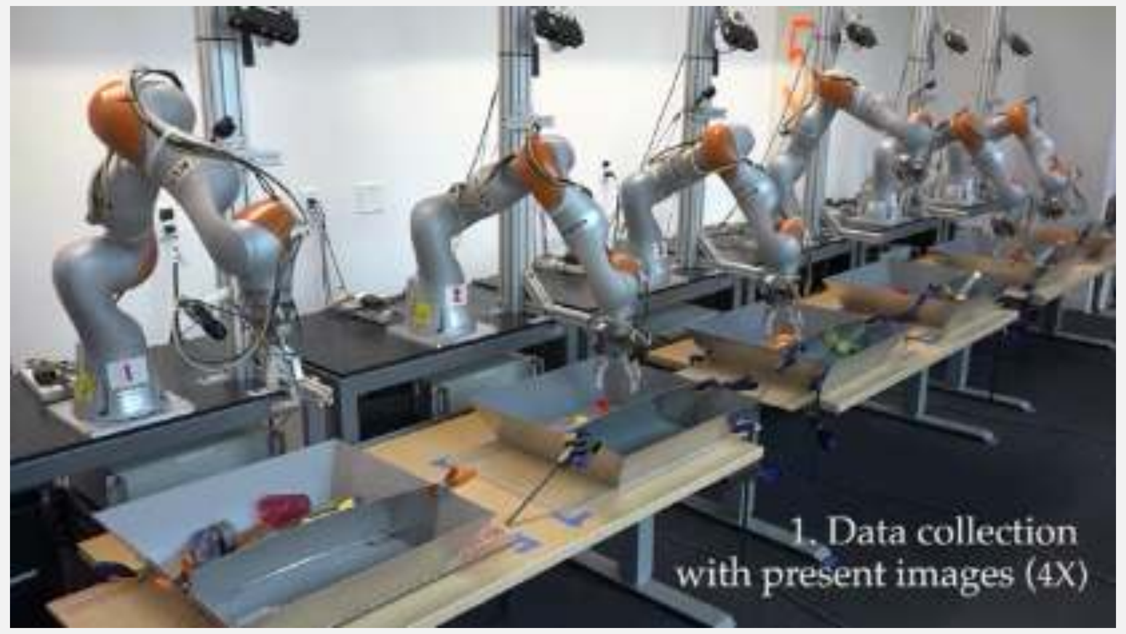
Vecerik et al. 2017: 100 demos

Large demonstration datasets is hard to collect.

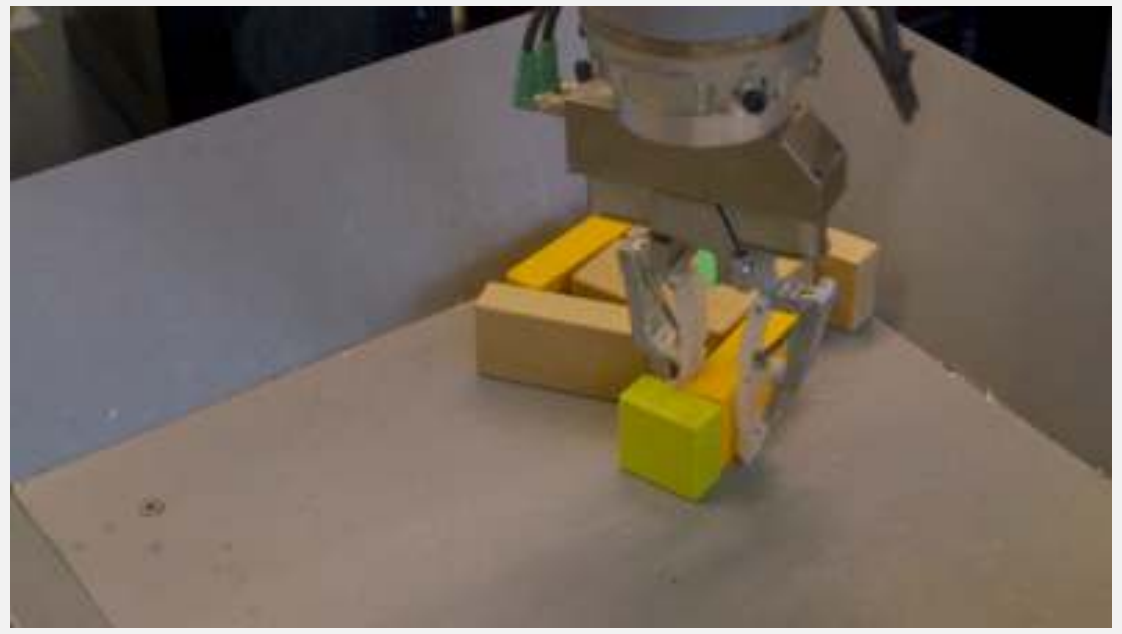
Humans need to demonstrate not label.

Reinforcement & Self-Supervised Learning

Levine et al. 2016



Kalashnikov et al. 2018



Pinto et al. 2016

Fang et al. 2018

Data can be low quality due to lack of expert.

Data is critical for learning robot skills.

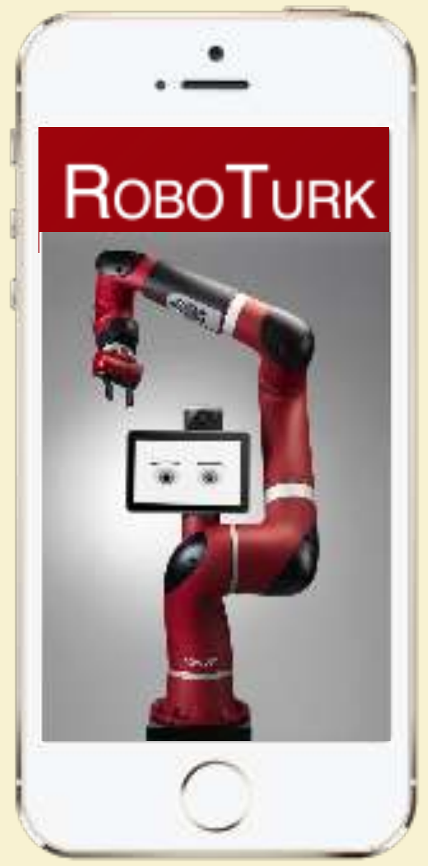
How to scale up high-quality human supervision for robotics?

Provide a natural way for anyone to provide demonstrations

Web-based Crowd Teleoperation with RoboTurk

RoboTurk: Crowdsourcing Platform for Large-Scale Demonstration Collection

RoboTurk in action



6-DoF controller

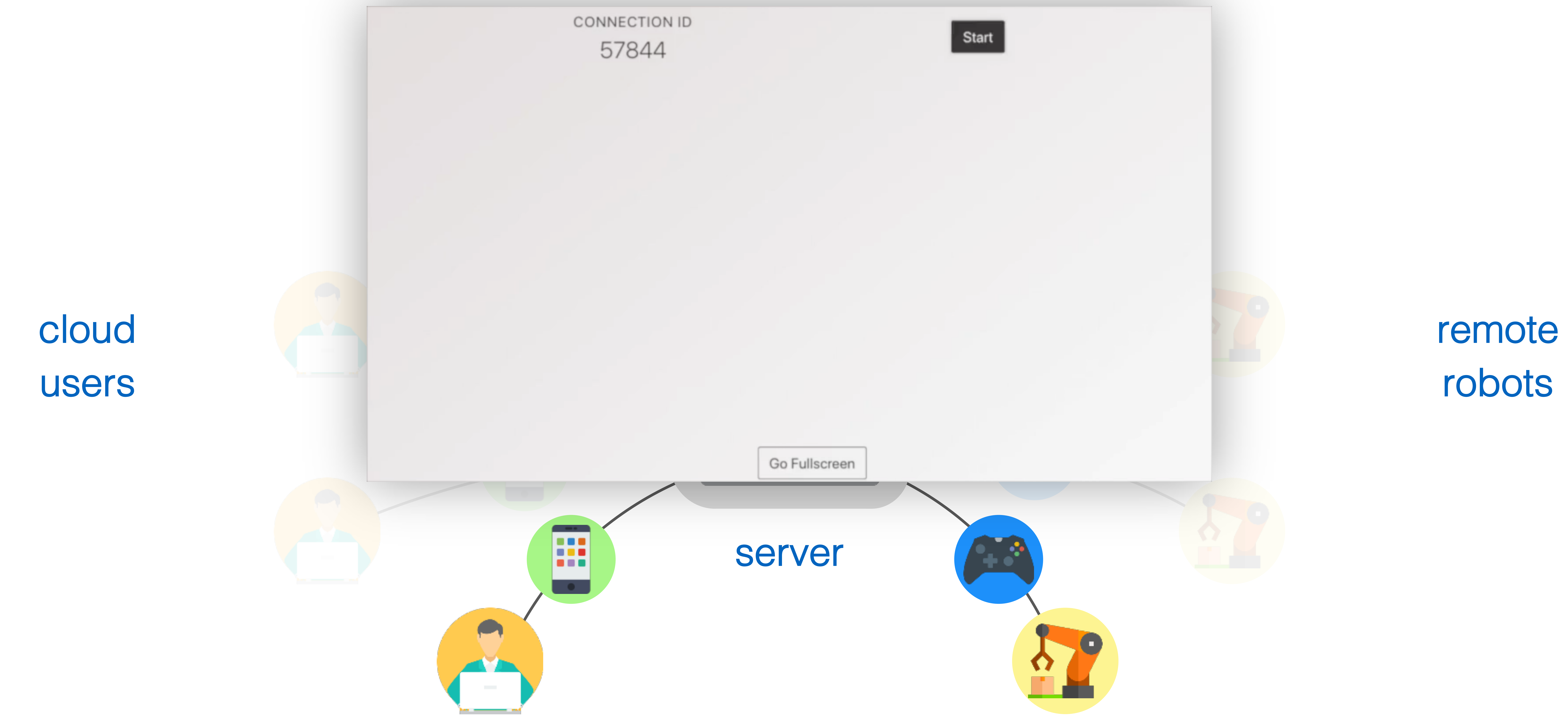
+

real-time streaming from remote robot

roboturk.stanford.edu

Mandlekar, **Zhu**, Garg, Booher, Spero, Tung, Gao, Emmons, Gupta, Orbay, Savarese, Fei-Fei, CoRL 2018

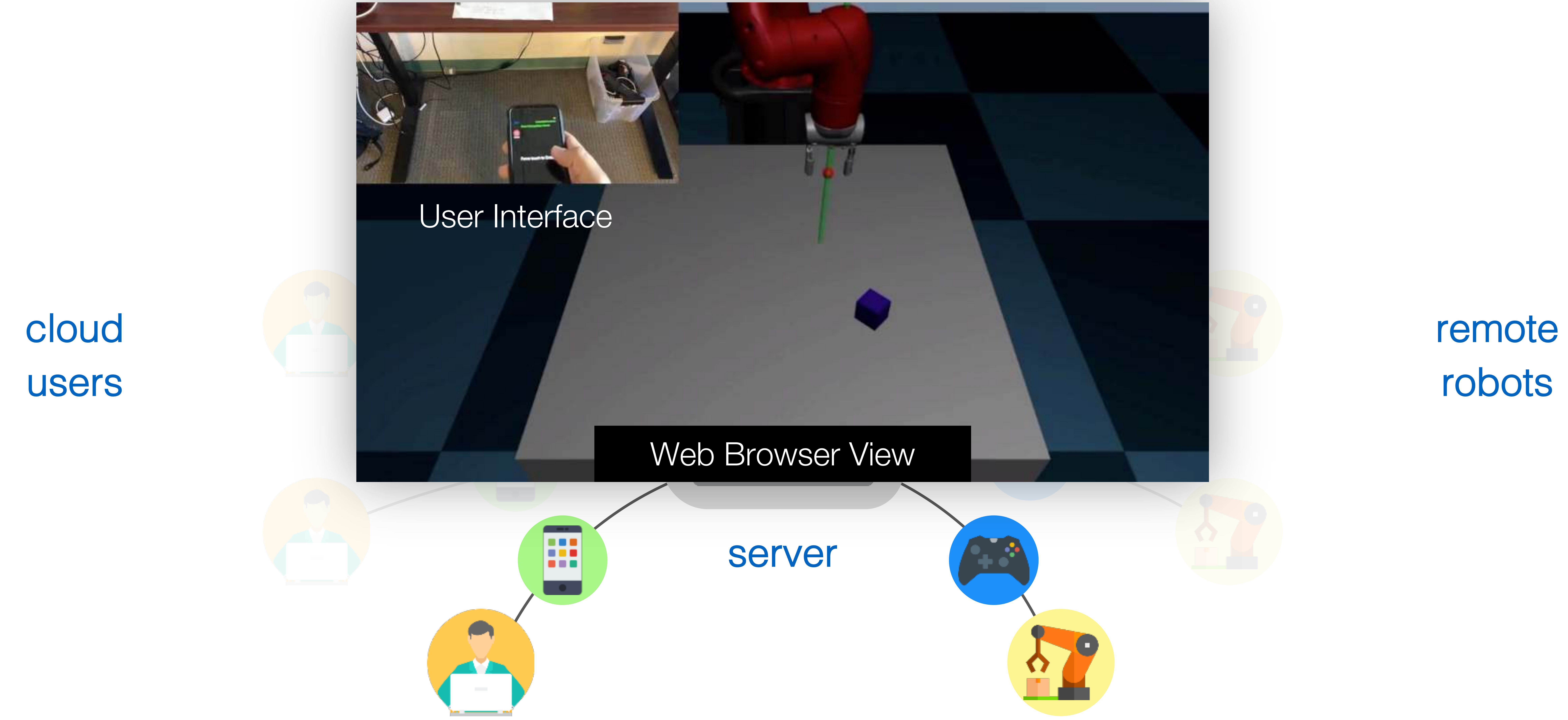
Web-based Crowd Teleoperation with RoboTurk



roboturk.stanford.edu

Mandlekar, **Zhu**, Garg, Booher, Spero, Tung, Gao, Emmons, Gupta, Orbay, Savarese, Fei-Fei, CoRL 2018

Web-based Crowd Teleoperation with RoboTurk



roboturk.stanford.edu

Mandlekar, **Zhu**, Garg, Booher, Spero, Tung, Gao, Emmons, Gupta, Orbay, Savarese, Fei-Fei, CoRL 2018

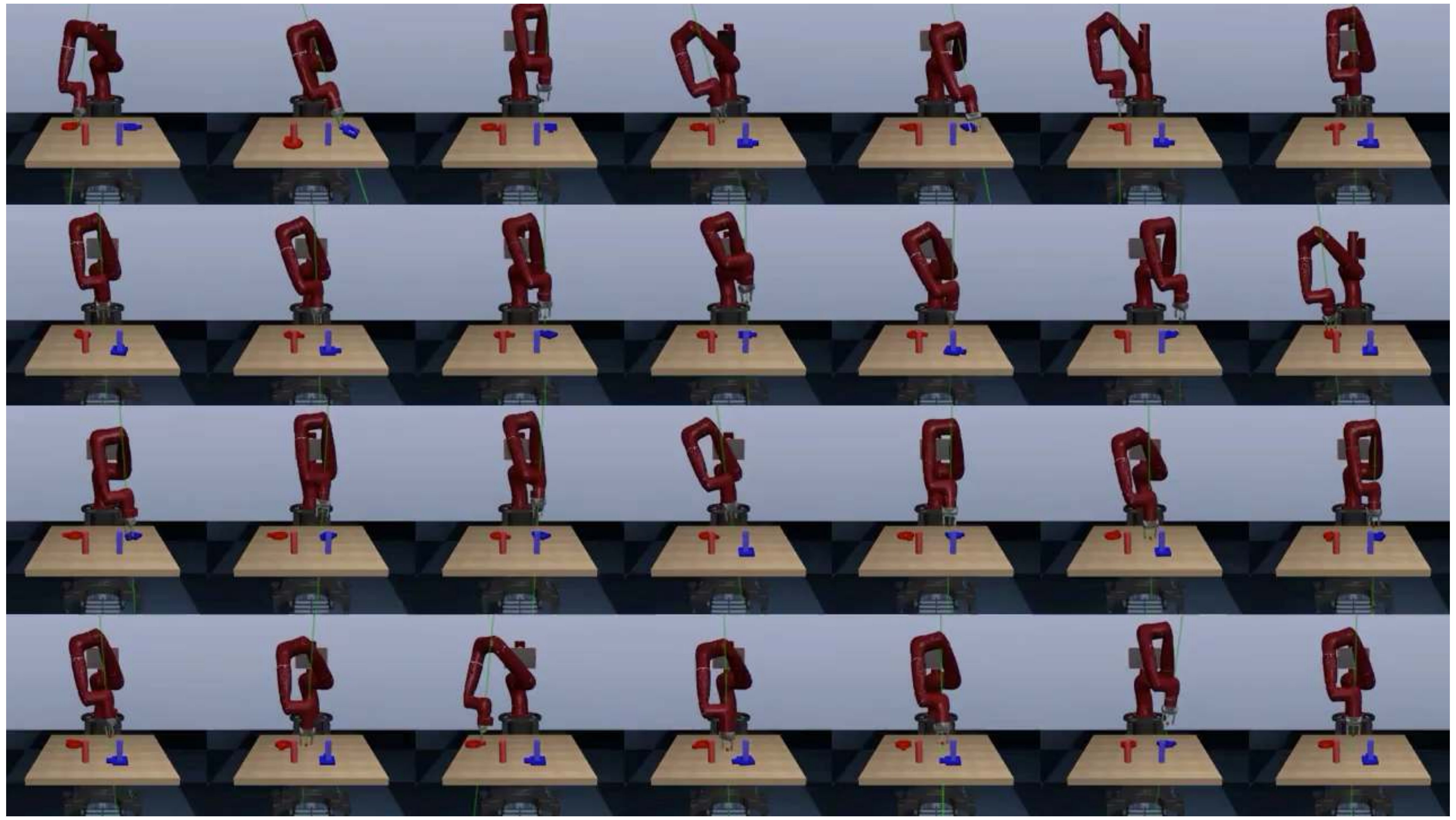
Web-based Crowd Teleoperation with RoboTurk

RoboTurk Pilot Dataset

137.5 hours of demonstrations

22 hours of total platform usage

2218 successful demonstrations



teleoperated demonstrations

surreal.stanford.edu

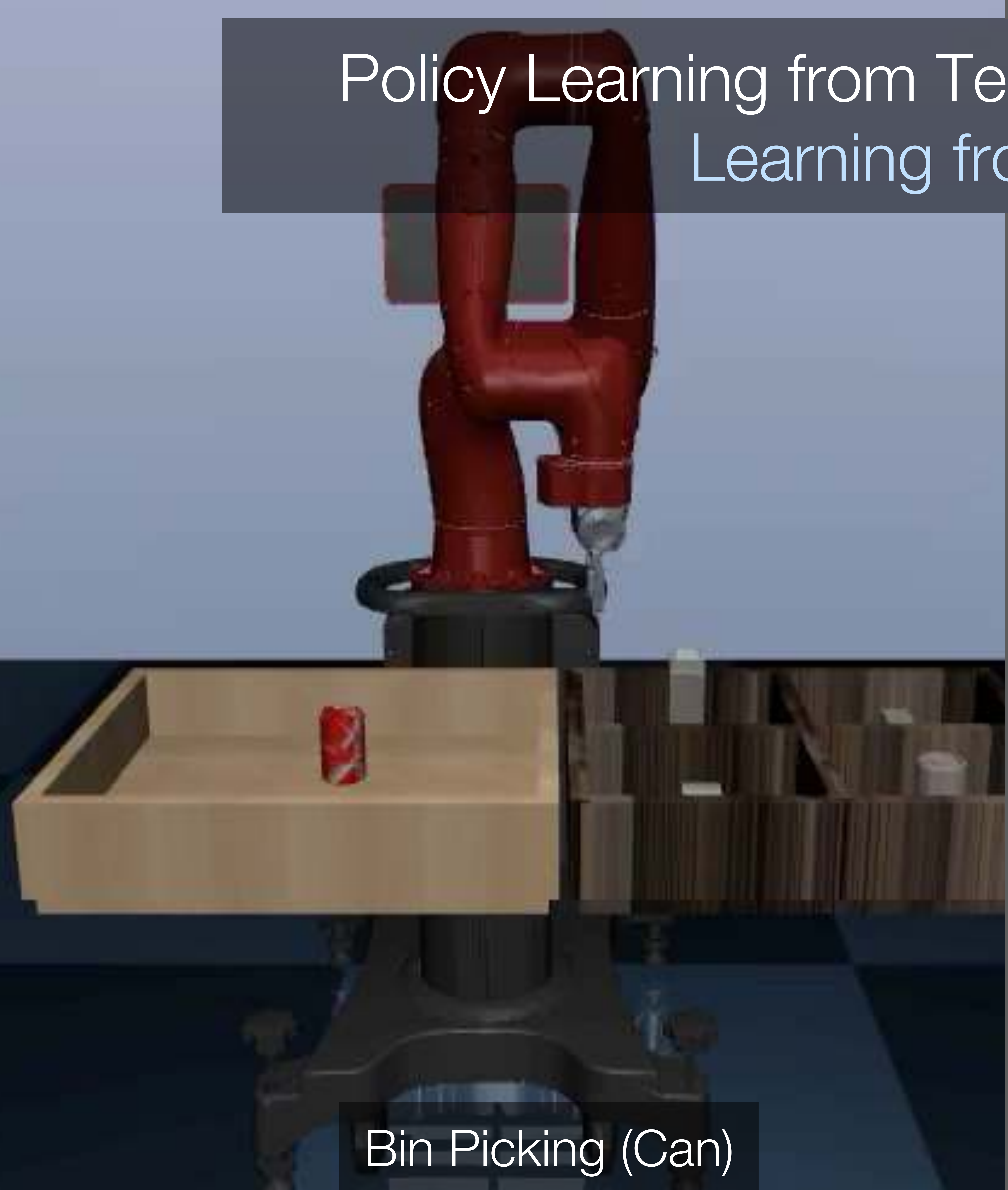
roboturk.stanford.edu

Zhu*, Fan*, Zhu, Liu, Zeng, Gupta, Creus-Costa, Savarese, Fei-Fei, CoRL 2018

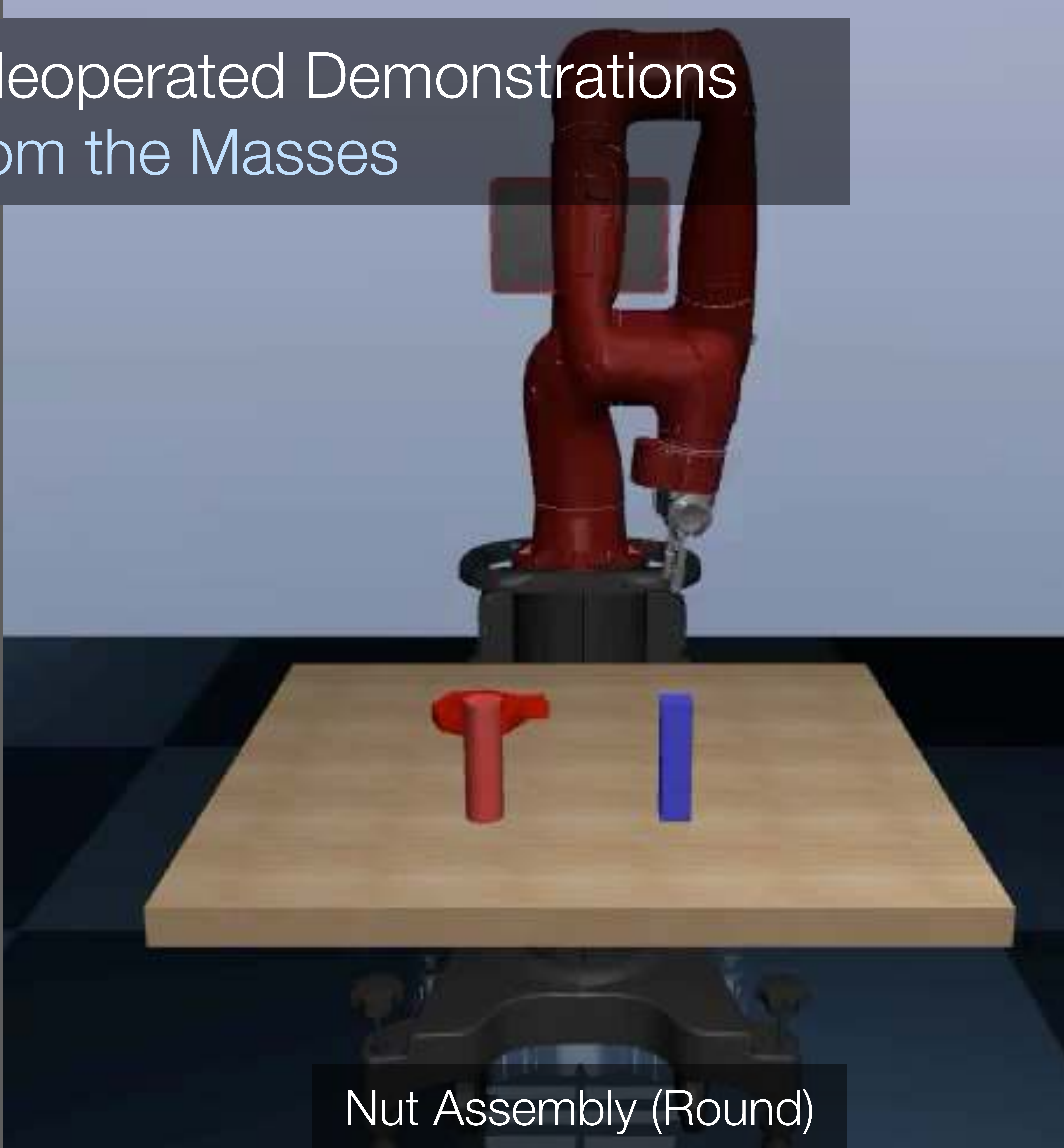
Mandlekar, **Zhu**, Garg, Booher, Spero, Tung, Gao, Emmons, Gupta, Orbay, Savarese, Fei-Fei, CoRL 2018

Policy Learning from Teleoperated Demonstrations

Learning from the Masses



Bin Picking (Can)



Nut Assembly (Round)

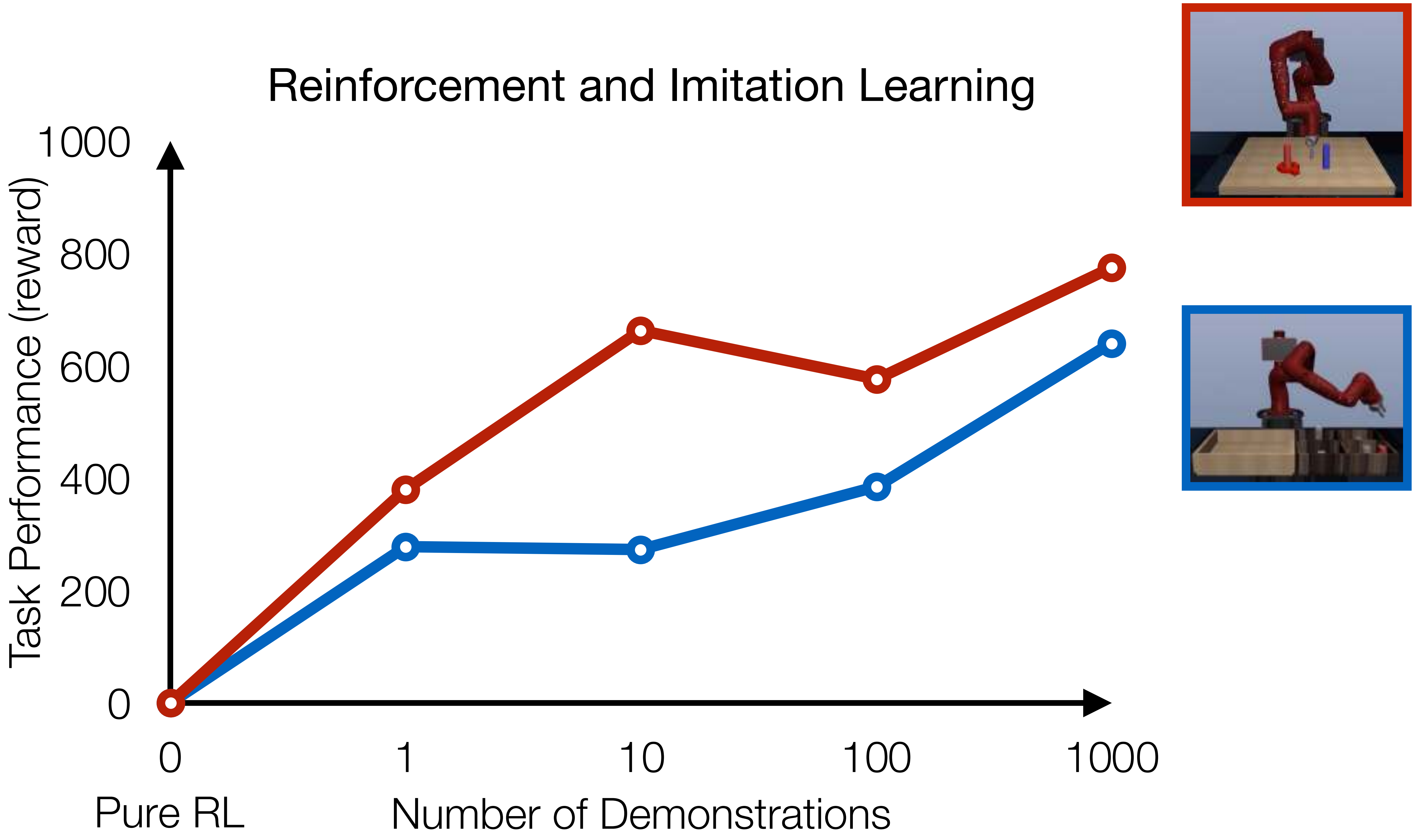
Reinforcement and Imitation Learning: Data

RoboTurk Pilot Dataset

137.5 hours of demonstrations

22 hours of total platform usage

2218 successful demonstrations



surreal.stanford.edu

roboturk.stanford.edu

Zhu*, Fan*, Zhu, Liu, Zeng, Gupta, Creus-Costa, Savarese, Fei-Fei, CoRL 2018

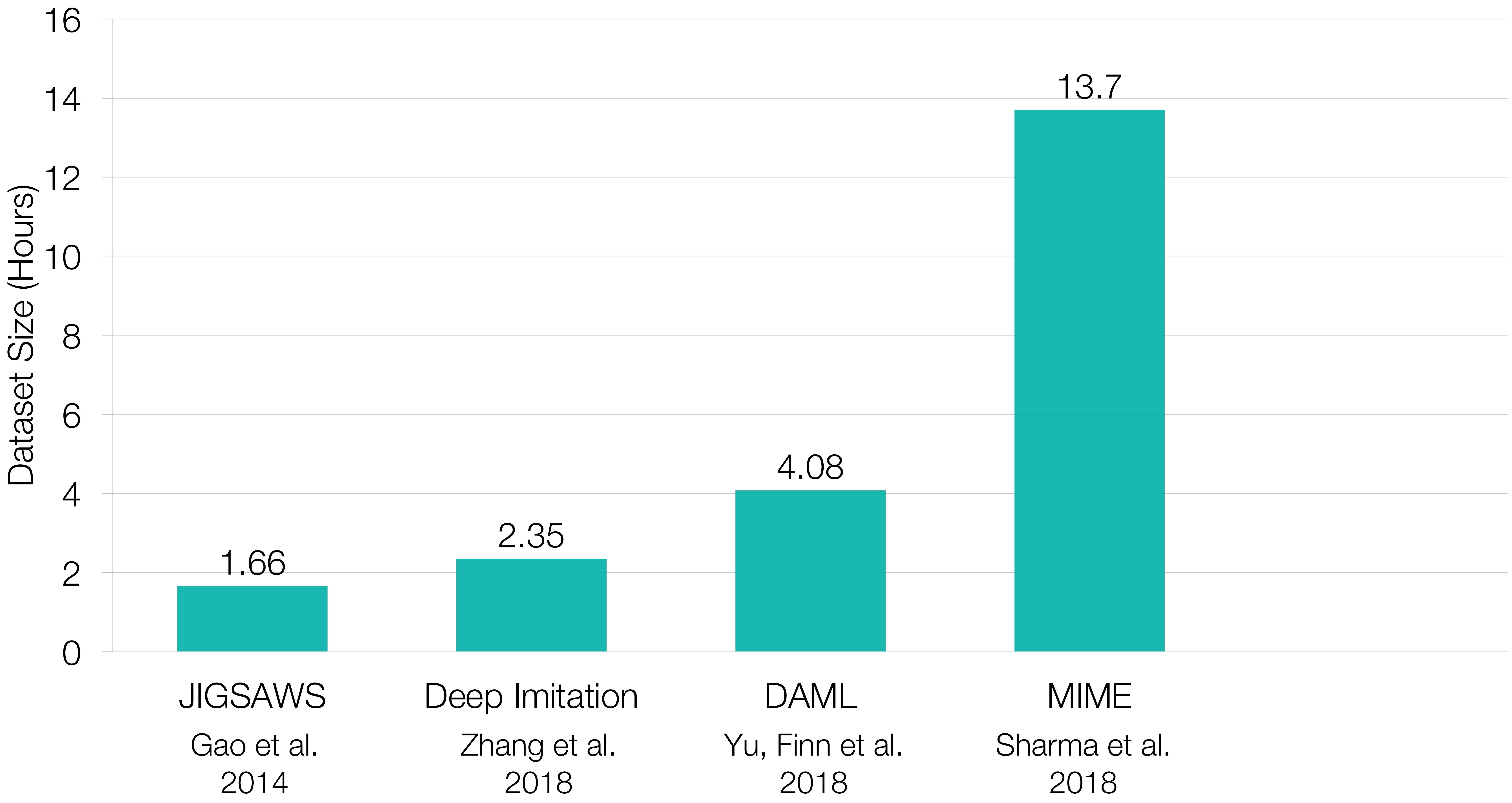
Mandlekar, **Zhu**, Garg, Booher, Spero, Tung, Gao, Emmons, Gupta, Orbay, Savarese, Fei-Fei, CoRL 2018

RoboTurk on Physical Robots

Real-time

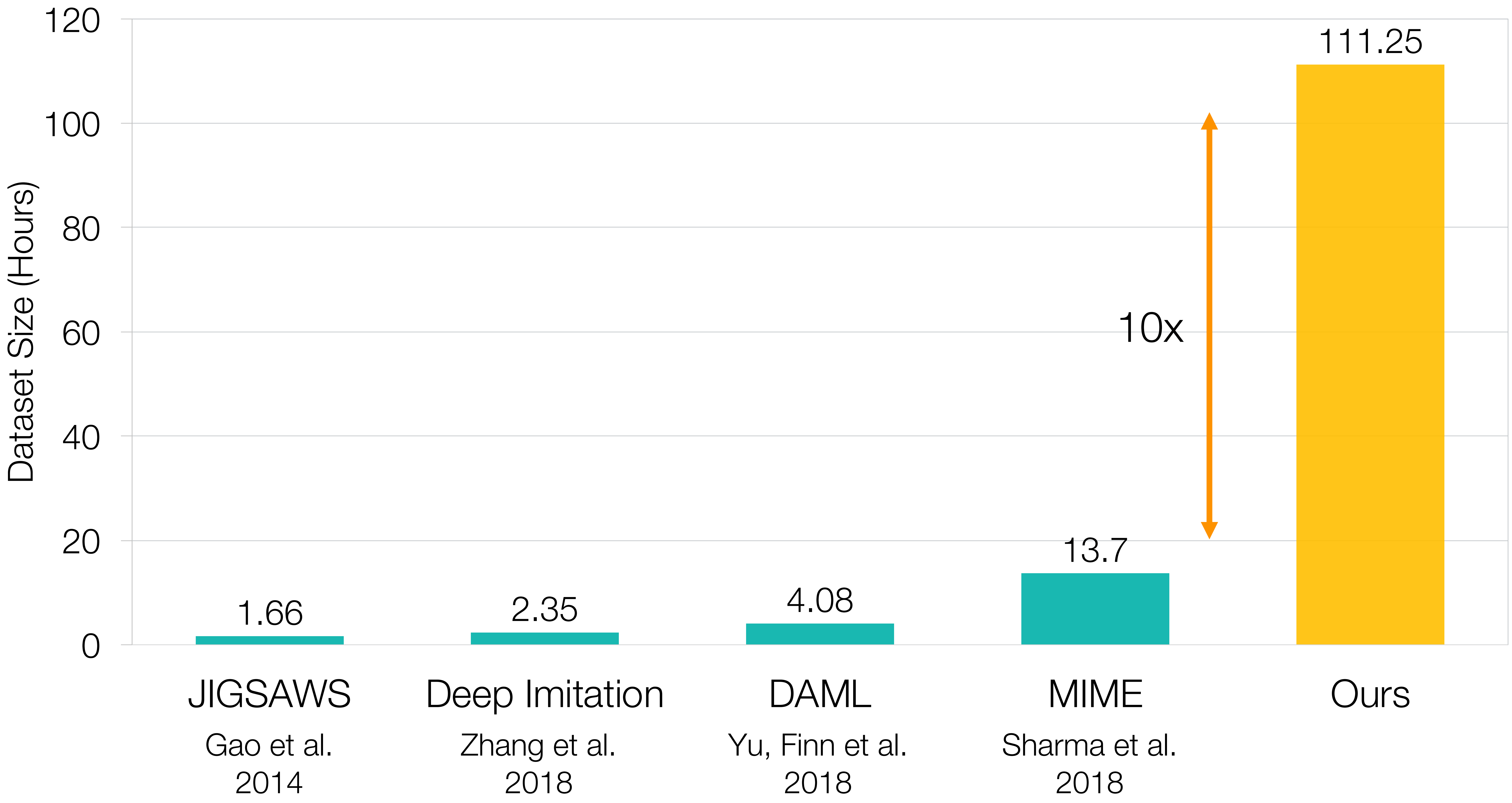
Scalable Data Collection

Dataset Size Comparison



Mandlekar, Booher, Spero, Tung, Gupta, **Zhu**, Garg, Savarese, Fei-Fei, IROS 2019

Dataset Size Comparison

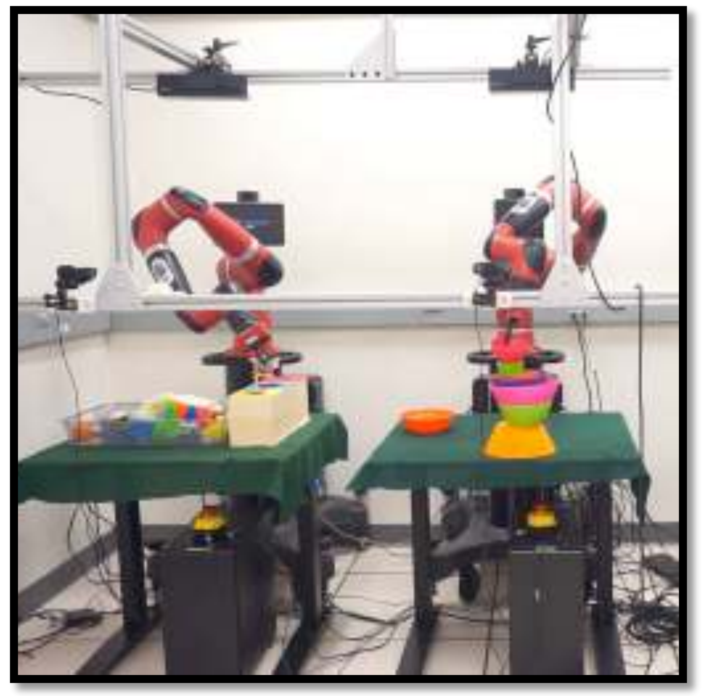


Mandlekar, Booher, Spero, Tung, Gupta, **Zhu**, Garg, Savarese, Fei-Fei, IROS 2019

RoboTurk for

everyone, everywhere

Summary - Part II



RoboTurk scales up demonstration collection with **teleoperated crowdsourcing** from web users

Large-scale **crowdsourced data** enables us to train more effective **motor skill learning** algorithms.

Learn More about RoboTurk?

Come to our IROS Presentation

RoboTurk: Human Reasoning and Dexterity for Large-Scale Dataset Creation

Tuesday 15:45-16:00, Award Session II: Paper TuBT4.5

Part I: Learning from Web Videos

Extracting compositional task structures from video data

Part II: Learning from Crowd Teleoperation

Crowdsourcing teleoperated demonstrations for skill learning

Conclusions

❖ What's a good representation of procedural knowledge?

High-level task structures & low-level motor skills

❖ How do we learn procedural knowledge from the web?

Large-scale web videos & crowd teleoperation from online users

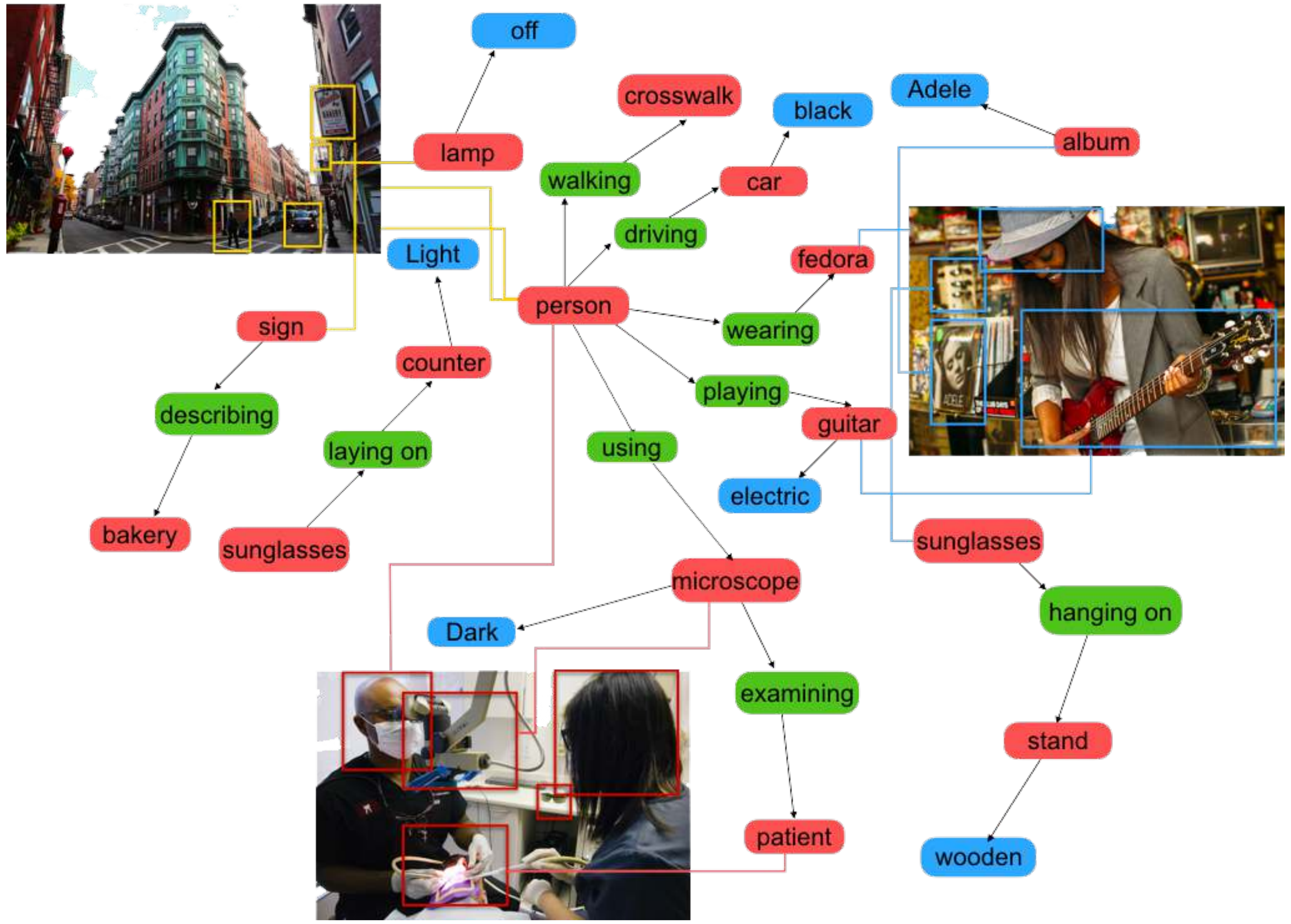
❖ How can robots take advantage of such knowledge?

Machine learning algorithms, e.g., meta-learning & imitation learning

Conclusions

Open Question:

How to integrate procedural knowledge and declarative knowledge into a unified knowledge ontology for building intelligent algorithms in robotics?



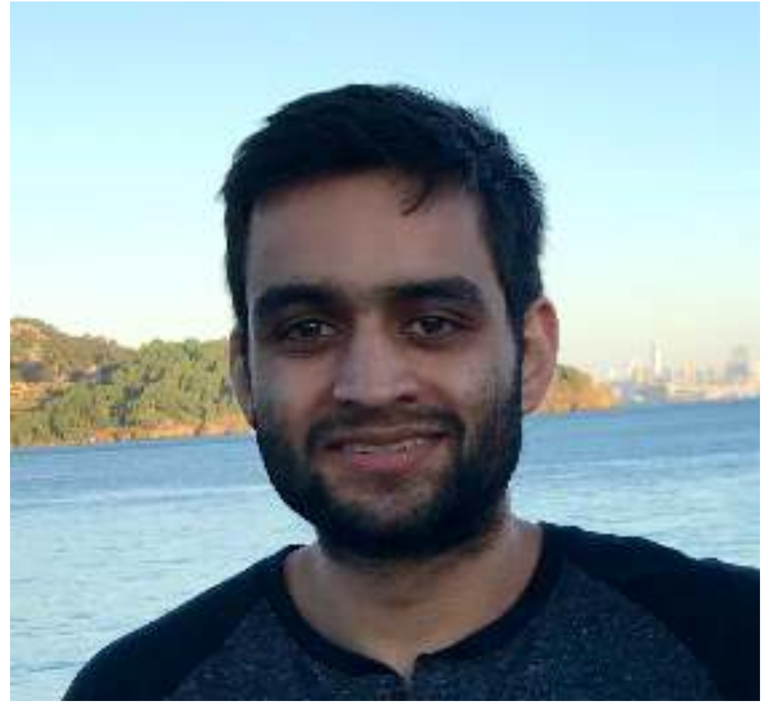
Acknowledgements

Fei-Fei Li

Silvio Savarese

Animesh Garg

Danfei Xu



Ajay Mandlekar

De-An Huang

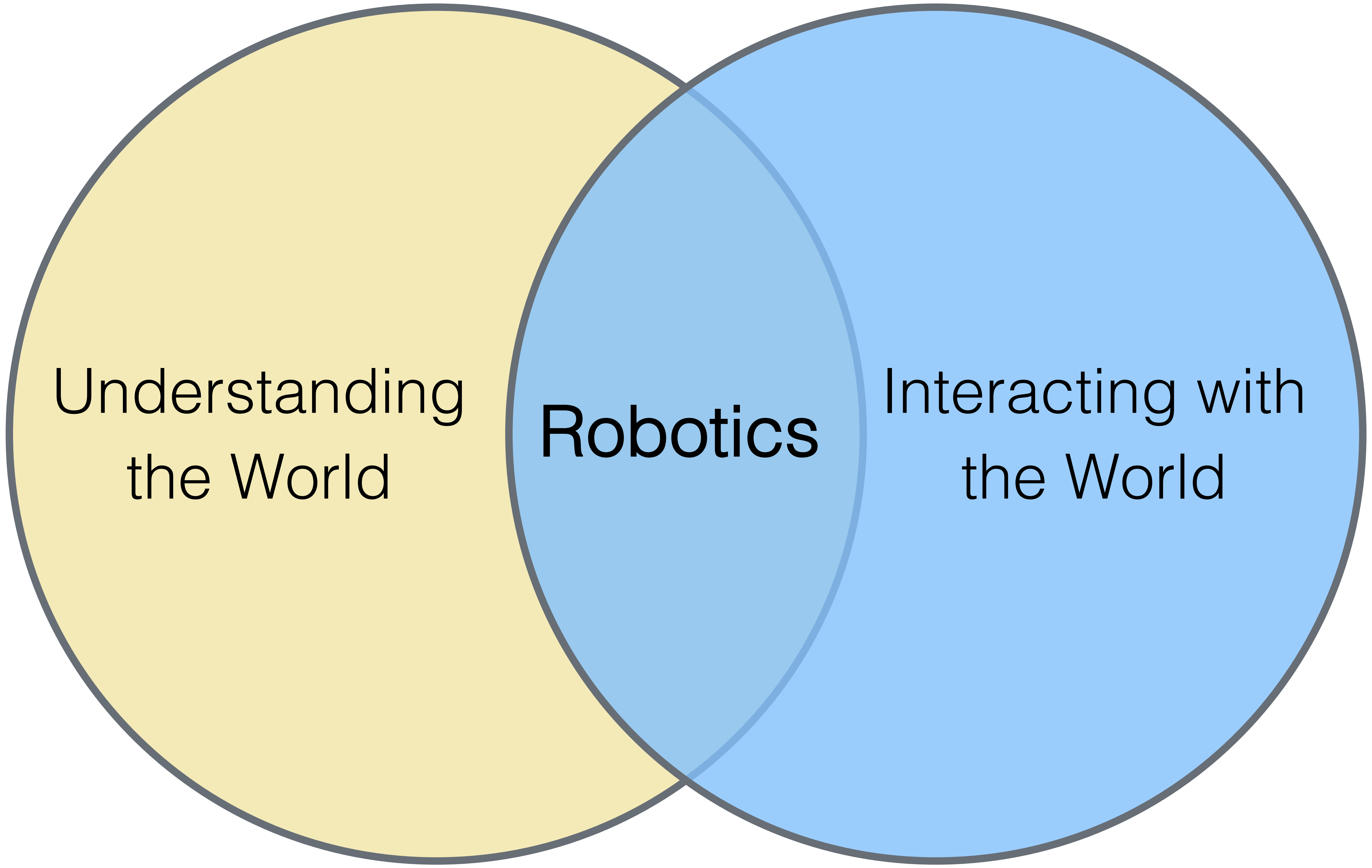
TOYOTA RESEARCH INSTITUTE

Google

JD.COM

DEPARTMENT OF THE NAVY ONR Science & Technology

Declarative Knowledge ("That-Is")



Procedural Knowledge ("How-To")

http://ai.stanford.edu/~yukez/

yukez@cs.stanford.edu

Conclusions

Open Question:

How to integrate procedural knowledge and declarative knowledge into a unified knowledge ontology for building intelligent algorithms in robotics?