Deep Multi-Task and Meta Learning

CS 330

The Plan for Today

1. Course goals & logistics
2. Why study multi-task learning and meta-learning?

Key learning today: what is multi-task learning??

Introductions

Chelsea Finn
Instructor

Amelie Byun
Course Coordinator

Suraj Nair
Head TA

Yoonho Lee
TA

Eric Mitchell
TA

Max Sobol Mark
TA

Fahim Tajwar
TA

Kyle Hsu
TA

Eric Frankel
TA

Welcome!

First question: How are you doing?

(answer by raising hand)

Information & Resources

Course website: http://cs330.stanford.edu/ ← We have put a lot of info here
Please read it. :)

Ed: Connected to Canvas

Staff mailing list: cs330-aut2223-staff@lists.stanford.edu

Office hours: Check course website & Canvas, ***start on Weds***.

OAE letters can be sent to staff mailing list or in private Ed post.

What will you learn in this course?

1. The foundations of modern deep learning methods for learning across tasks
2. How to implement and work with practical multi-task & transfer learning systems (in PyTorch)
3. A glimpse into the scientific and engineering process of building and understanding new algorithms

Topics

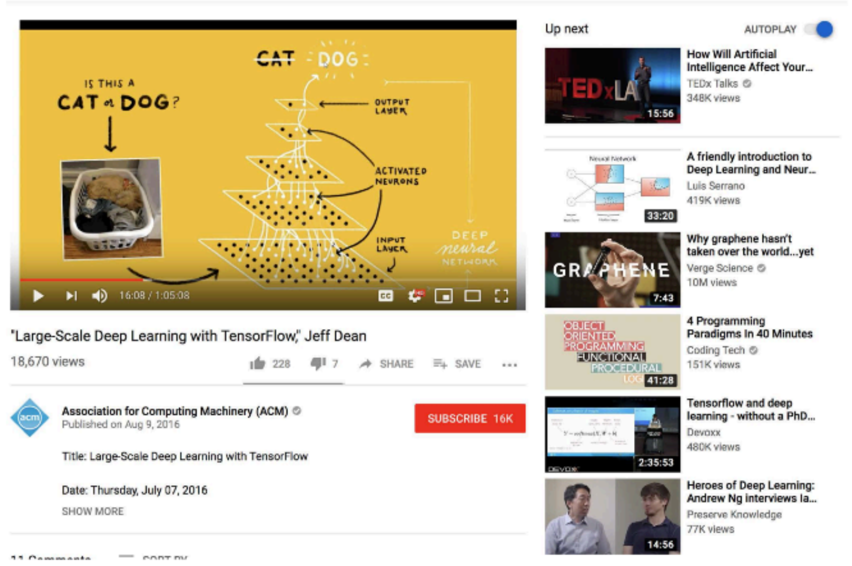
1. Multi-task learning, transfer learning basics
2. Meta-learning algorithms
   (black-box approaches, optimization-based meta-learning, metric learning)
3. Advanced meta-learning topics
   (meta-overfitting, unsupervised meta-learning, Bayesian models)
4. Unsupervised pre-training for few-shot learning
5. Relation to foundation models & in-context learning
6. Domain adaptation & generalization
7. Lifelong learning
8. Open problems

$\left.\begin{array}{l}4.\\5.\\6.\end{array}\right\}$ New!

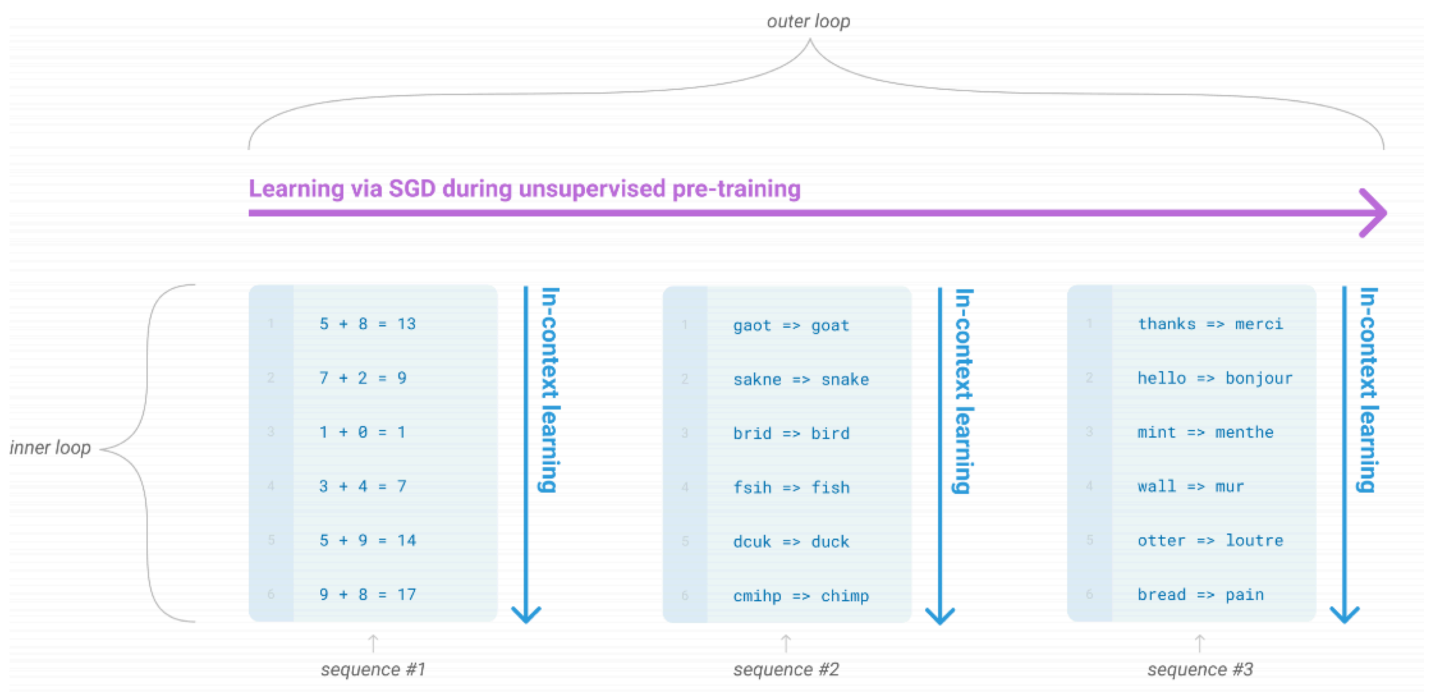
New: No RL lectures or HWs.

Emphasis on **deep learning** techniques.

Case studies of important & timely applications
- Multi-task learning in recommender systems
- Meta-learning for land cover classification, education
- Few-shot learning in large language models



Zhao et al. Recommending What Video to Watch Next. 2019



Brown et al. Language Models are Few-Shot Learners. 2020

Why no RL? What if I want to work on it?

New: No RL lectures or HWs.

- **New course** in Spring quarter (CS224R: Deep Reinforcement Learning)
- Removing RL makes the course **more accessible**.
- You can still explore RL topics in final project.

Lectures & Office Hours

Lectures

- In-person, livestreamed, & recorded
- Two guest lectures (TBD)

Ask questions!

- by raising your hand
- by entering the question in zoom chat

Office hours

- mix of in-person and remote

Pre-Requisites

Pre-requisites: CS229 or equivalent.

PyTorch:

- Assignments will require training networks in **PyTorch**.
- Fahim will hold a PyTorch review session on Thursday, Sep 29, 4:30 pm PT.

Assignments

Homework 0: Multi-task learning basics — 5% of grade

Homework 1: Multi-task data processing, black-box meta-learning

Homework 2: Gradient-based meta-learning & metric learning — 15% of grade each

Homework 3: Fine-tuning pre-trained models

Homework 4 (optional): Bayesian meta-learning & meta-overfitting — (replaces 15% of HW or project)

Grading: 50% homework, 50% project

6 late days total across: homeworks, project-related assignments
maximum of 2 late dates per assignment

Collaboration policy: Please read course website & honor code.
Document collaborators & write up HW solutions on your own.

Final Project

Research-level project of your choice

- in groups of **1-3 students**
- if applicable, encouraged to **use your research**!
- can share with other classes, with slightly higher expectation
- same late day policy as HWs
  (but no late days for poster)

Poster presentation on December 7th.

Initial Steps

1. Homework 0 is out — due Monday at 11:59 pm PT
2. Start forming final project groups if you want to work in a group

The Plan for Today

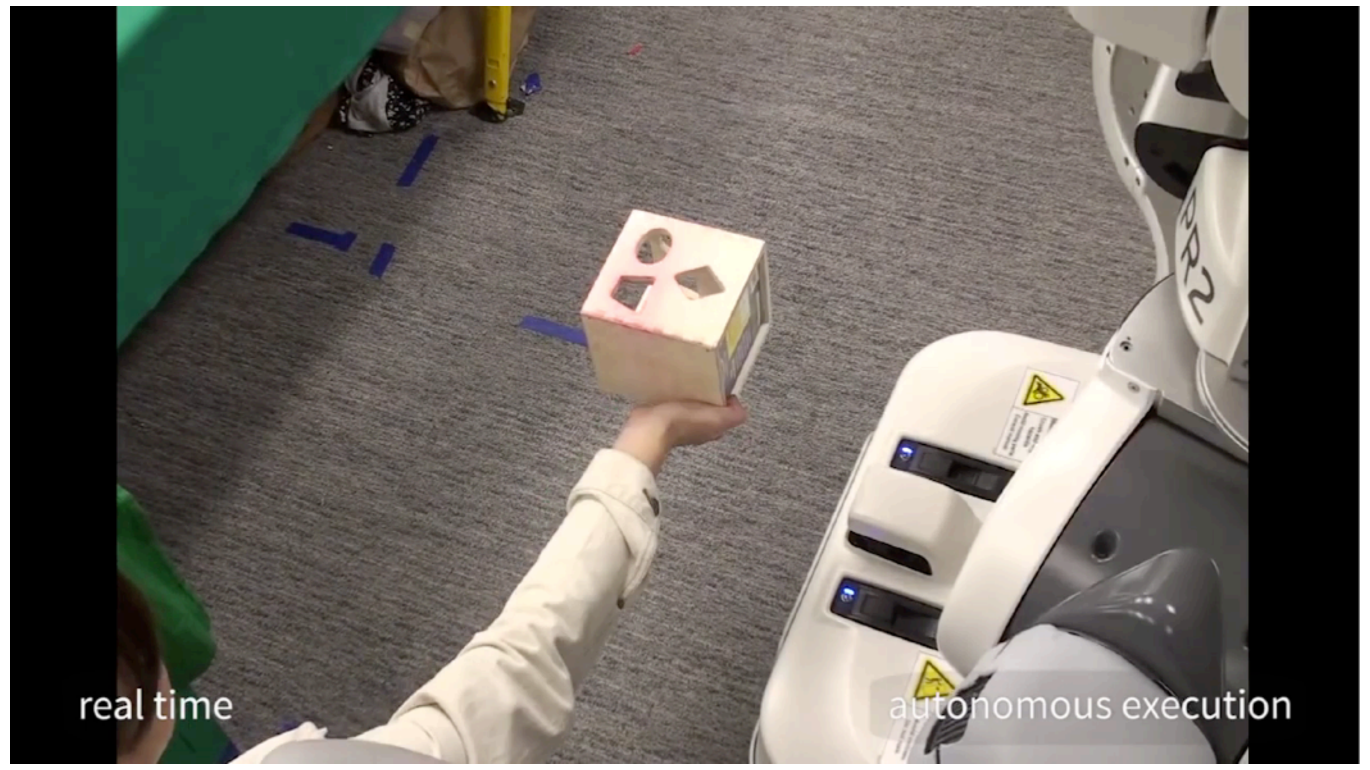
1. Course goals & logistics
2. **Why study multi-task learning and meta-learning?**

Some of Chelsea's Research

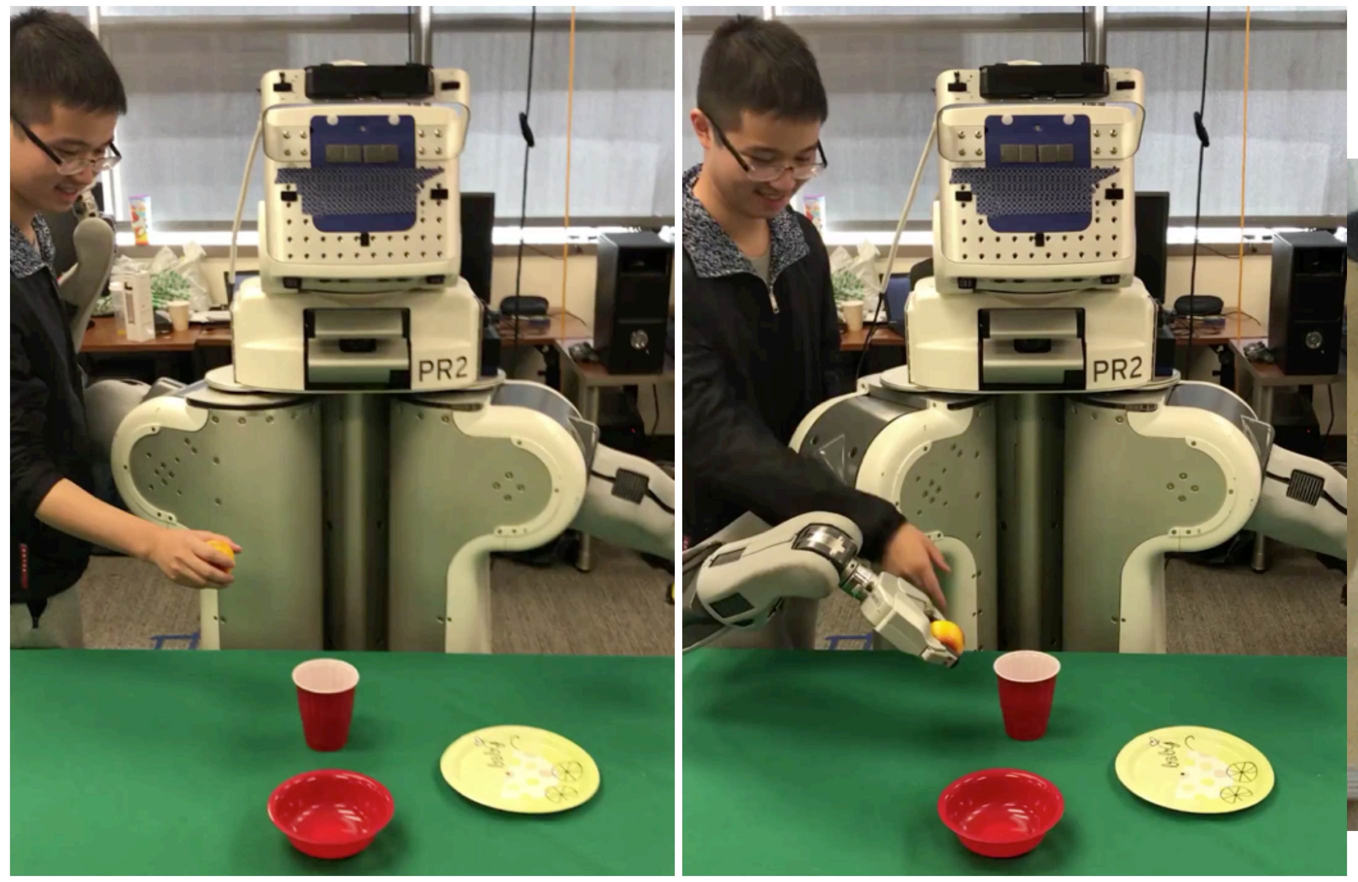
(and why I care about multi-task learning and meta-learning)

How can we enable agents to learn a breadth of skills in the real world?

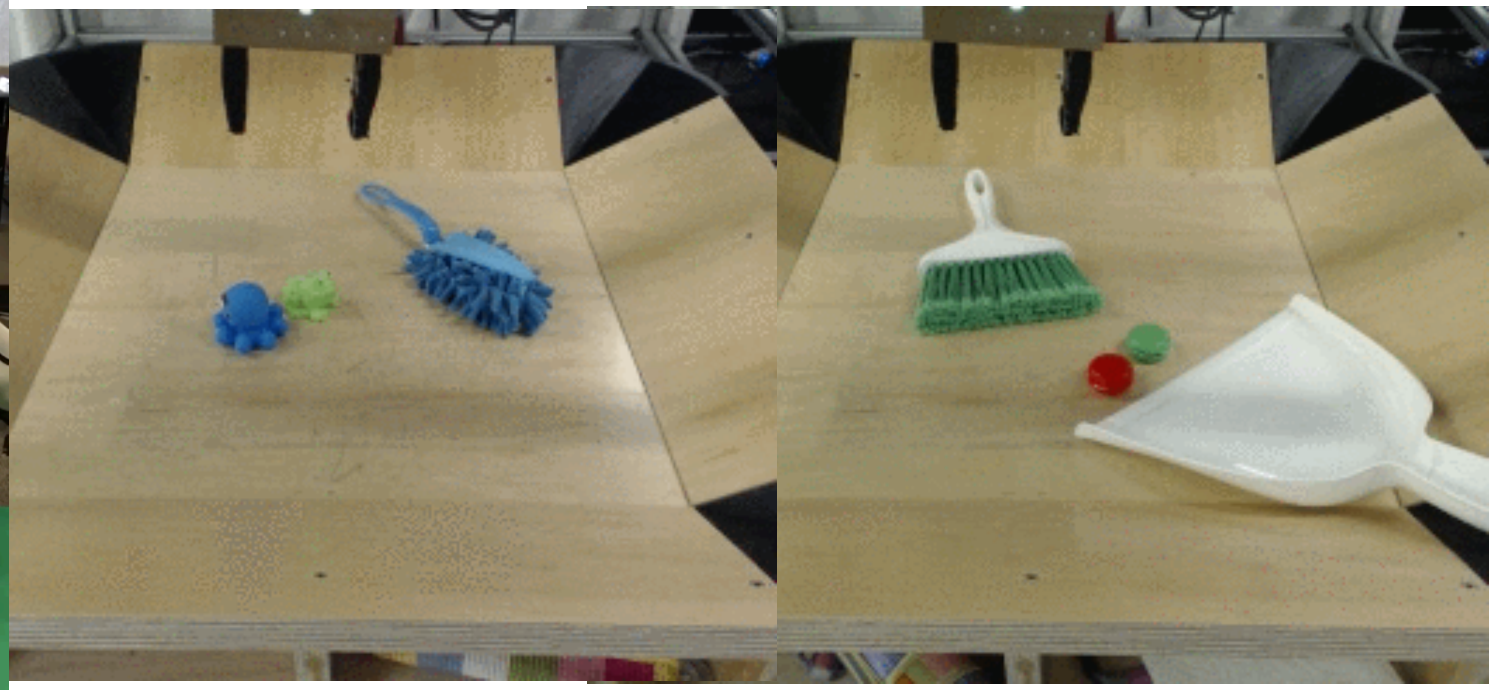
Robots.



Levine*, Finn*, Darrell, Abbeel. JMLR '16



Yu*, Finn*, Xie, Dasari, Zhang, Abbeel, Levine, RSS '18



Xie, Ebert, Levine, Finn, RSS '19

Why robots? Robots can teach us things about intelligence.

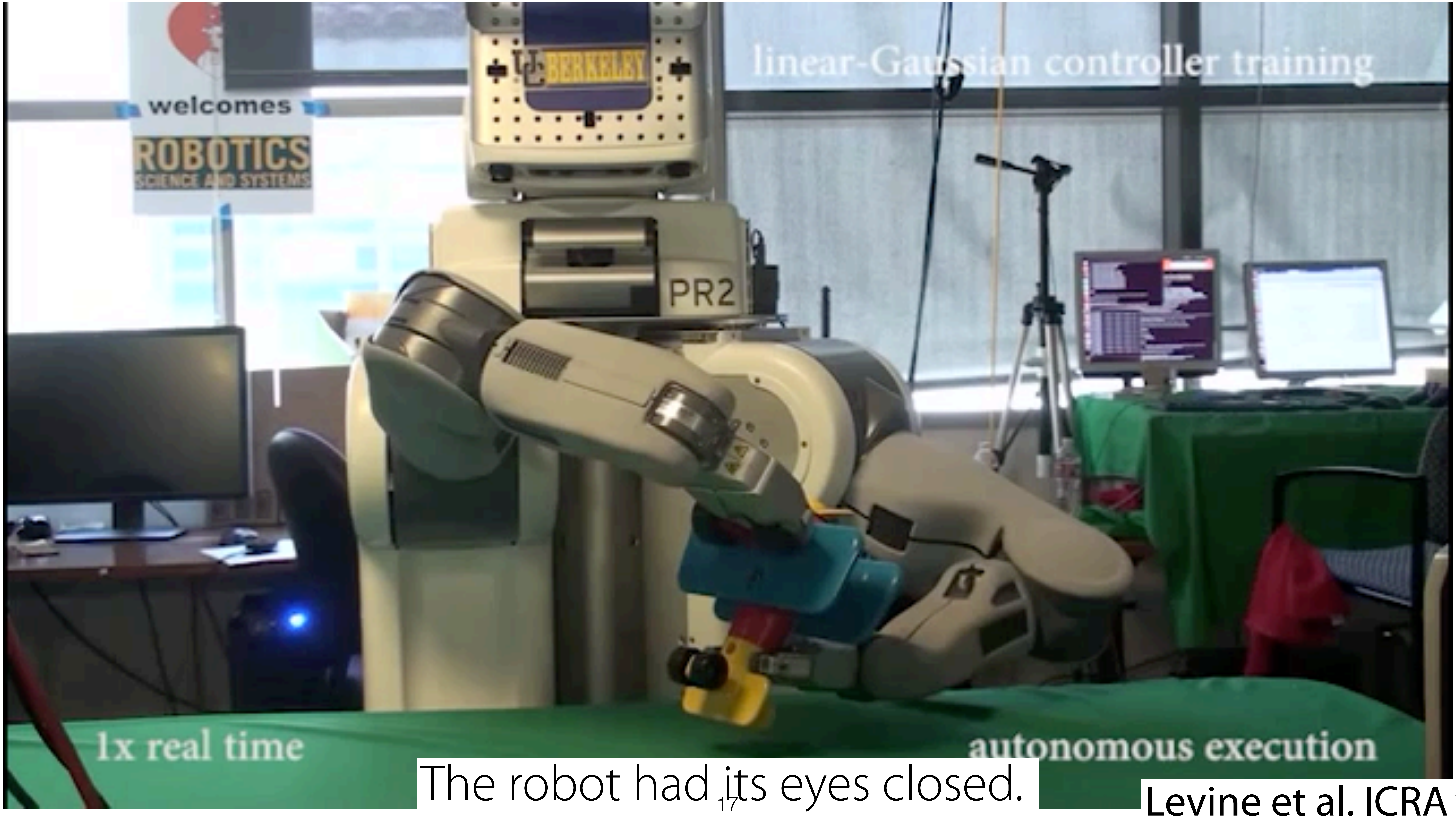
faced with the **real world**

must **generalize** across tasks, objects, environments, etc

need some **common sense understanding** to do well

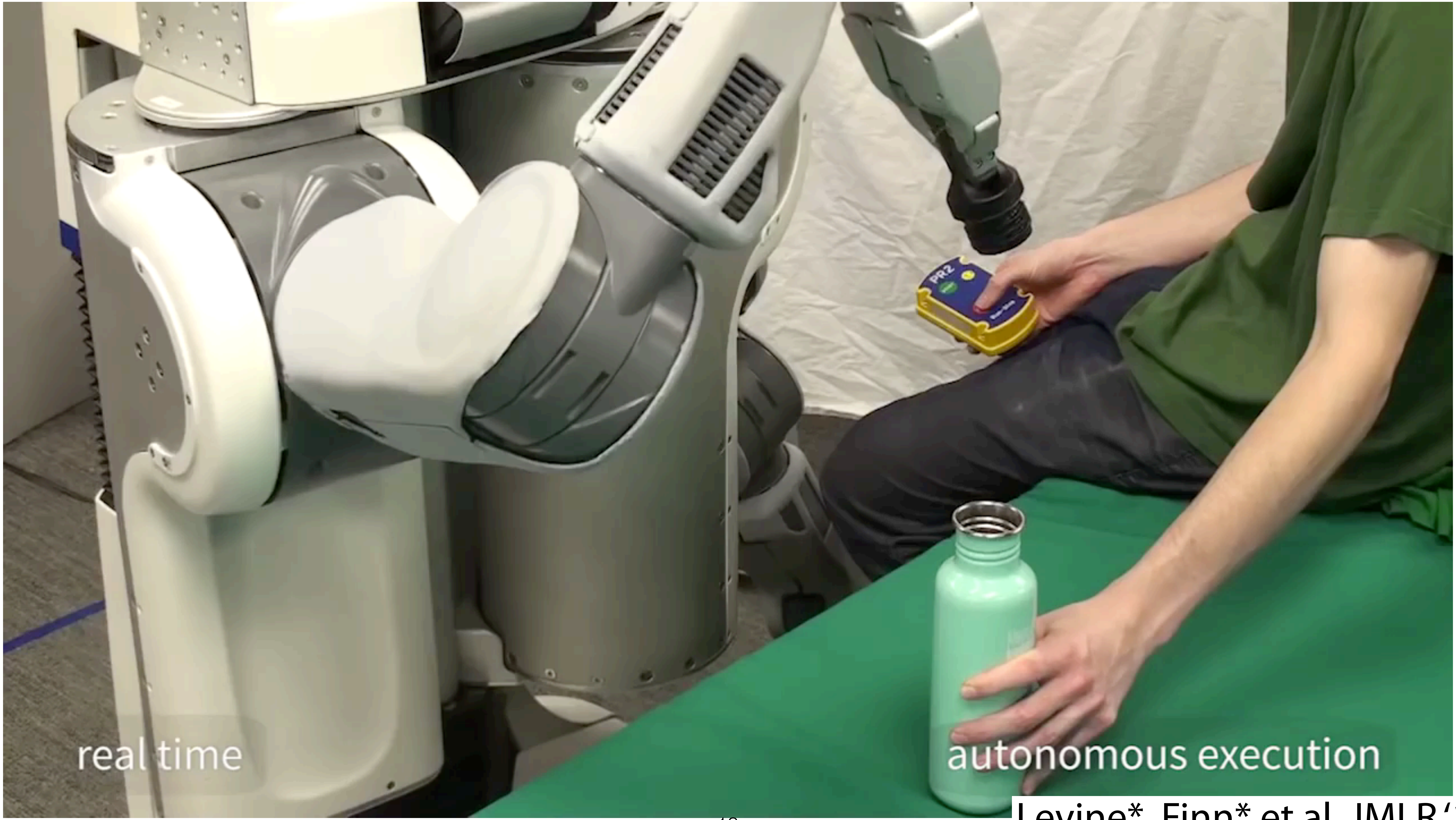
supervision can't be taken for granted

Beginning of my PhD



The robot had its eyes closed.

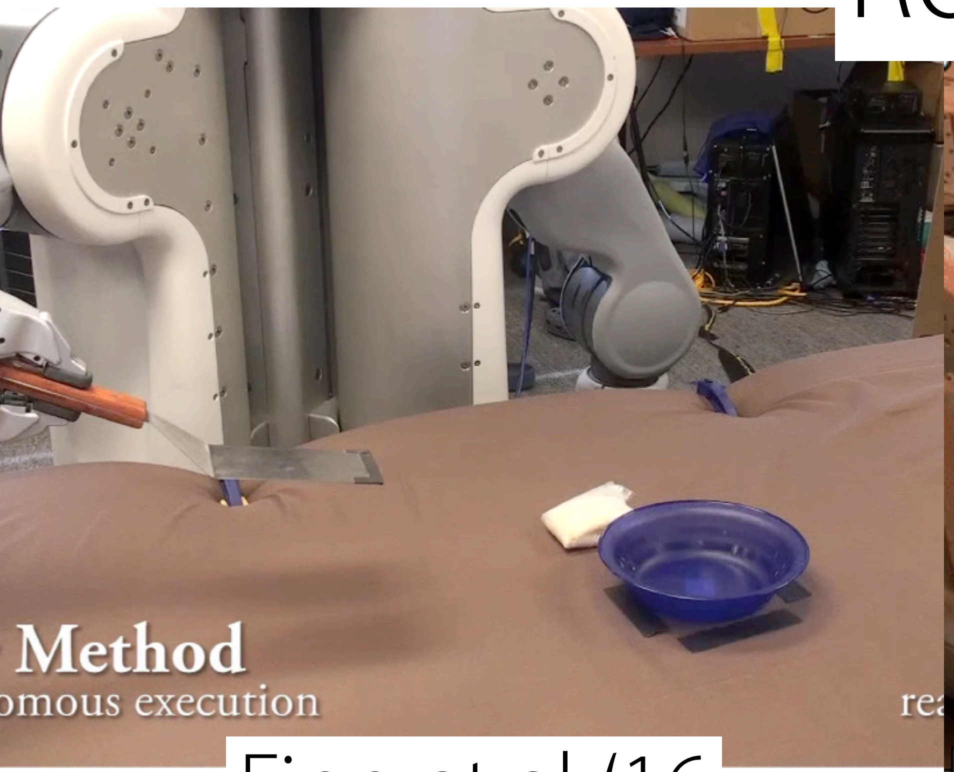
Levine et al. ICRA '15



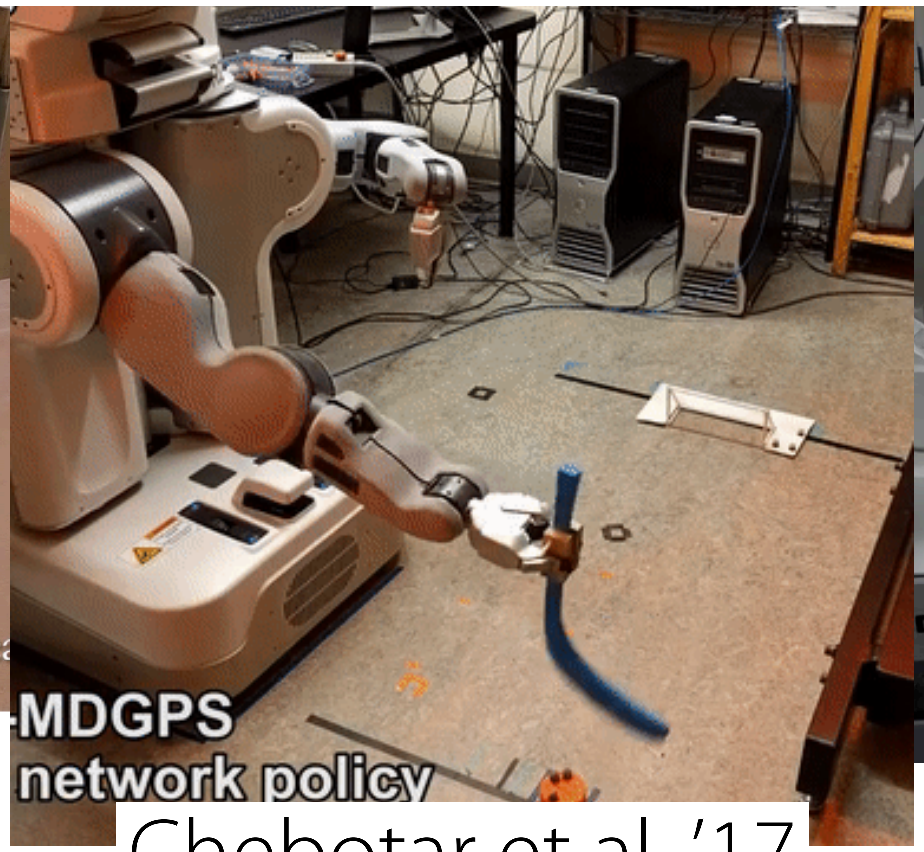
Levine*, Finn* et al. JMLR '16

Finn et al. ICRA '16

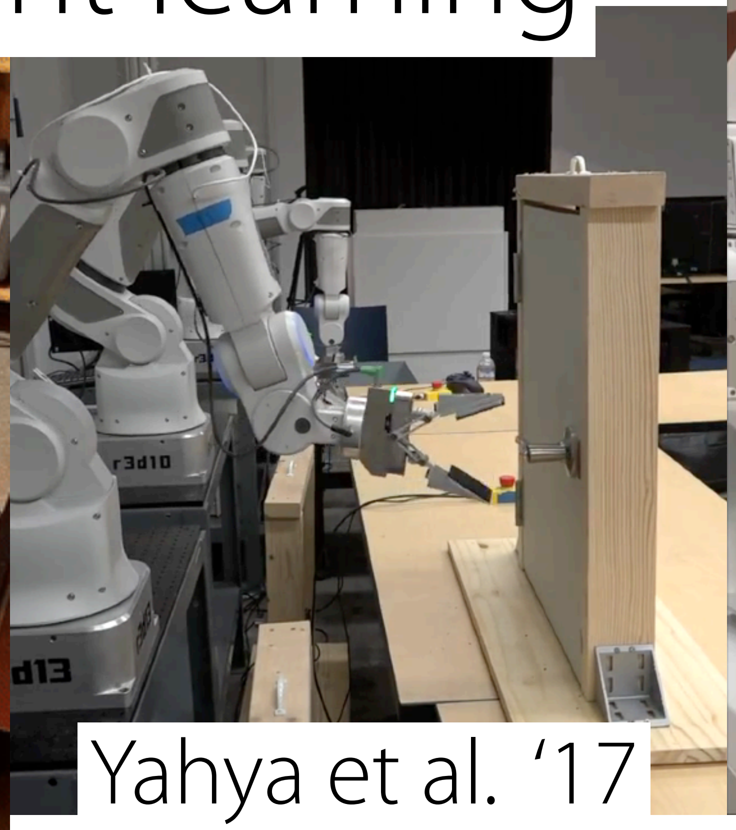
Robot reinforcement learning



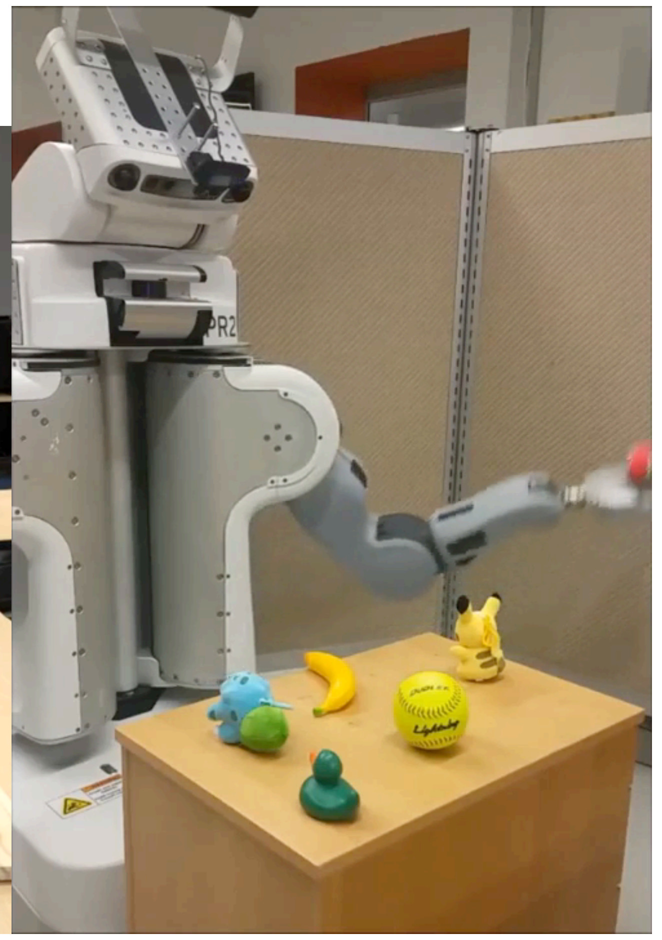
Finn et al.'16



Chebotar et al. '17

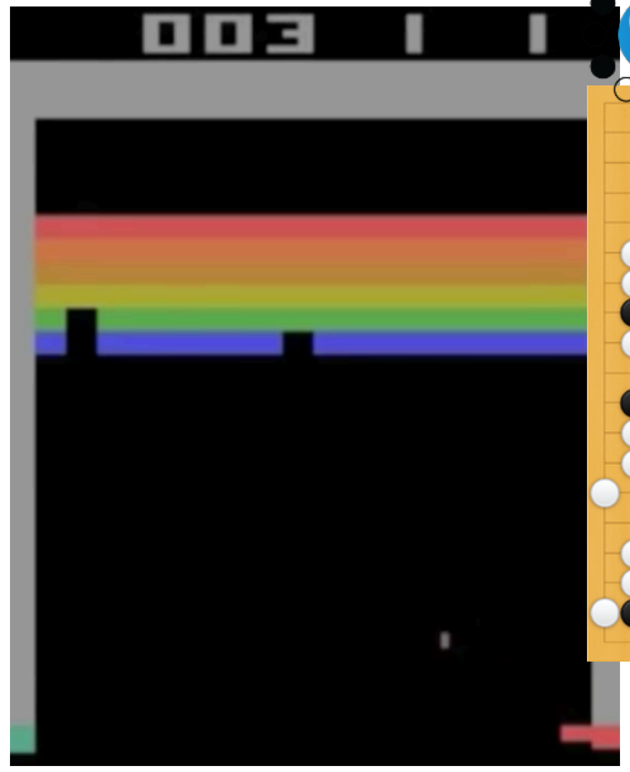


Yahya et al. '17



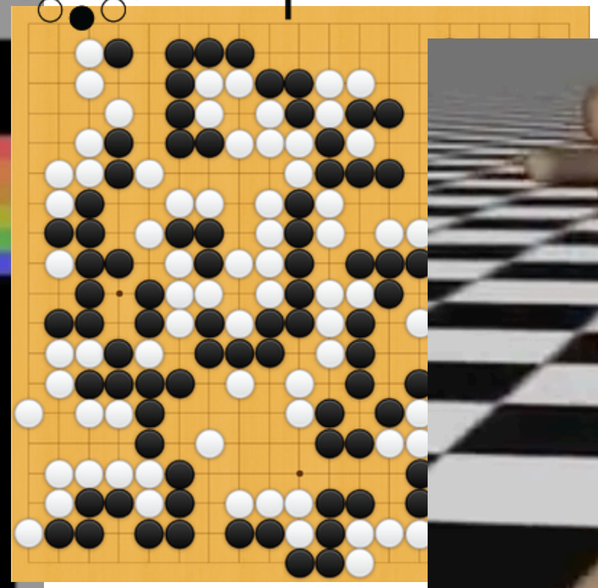
Ghadirzadeh et al. '17

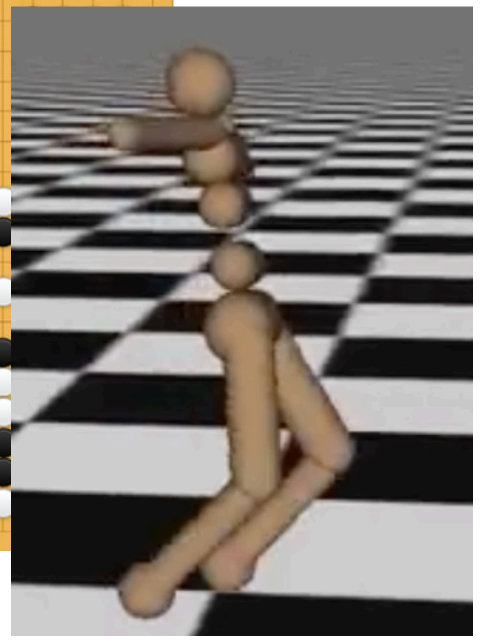
Reinforcement learning



Atari

AlphaGo

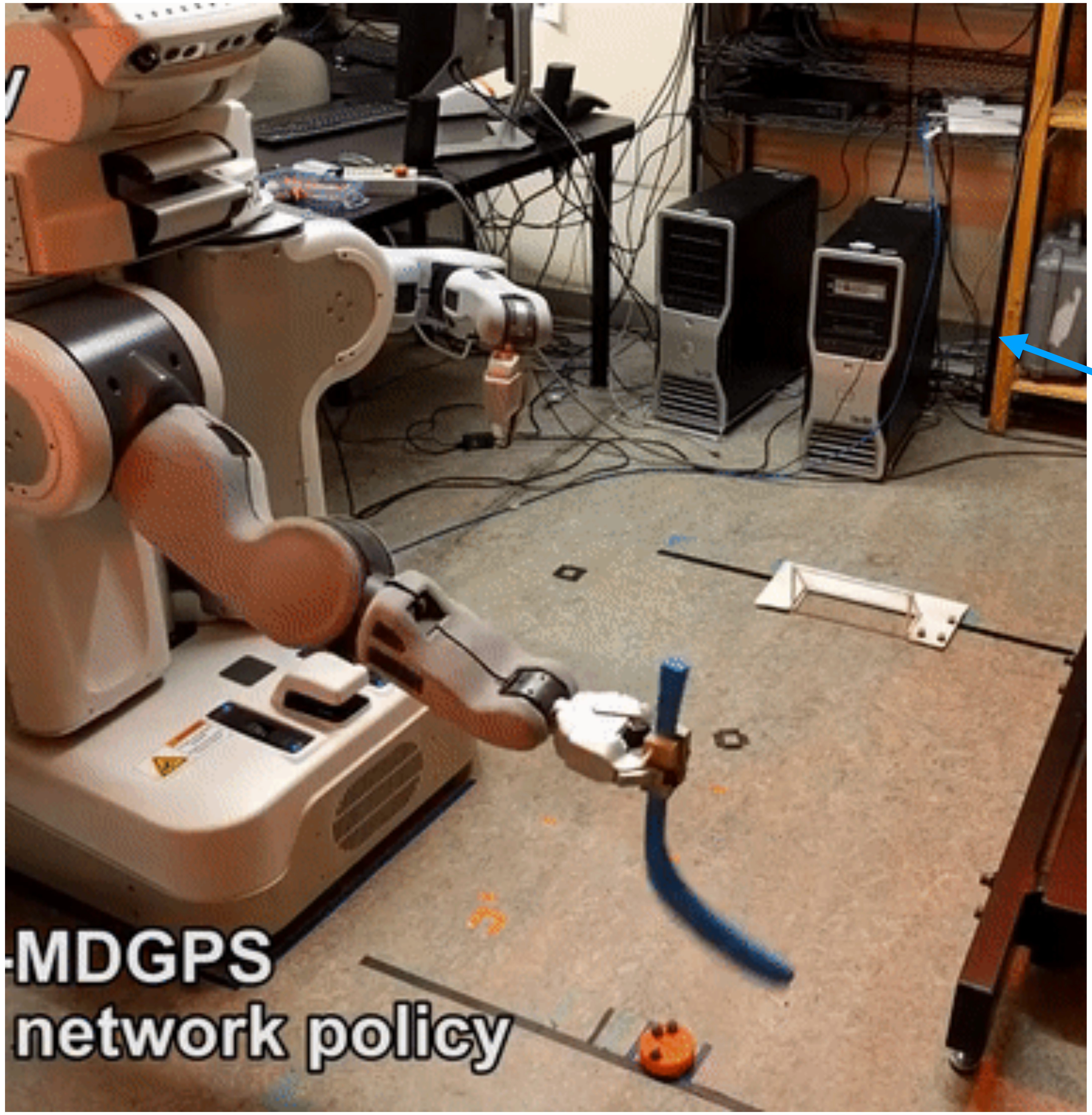




locomotion

Learn **one task** in **one environment, starting from scratch**

Behind the scenes…

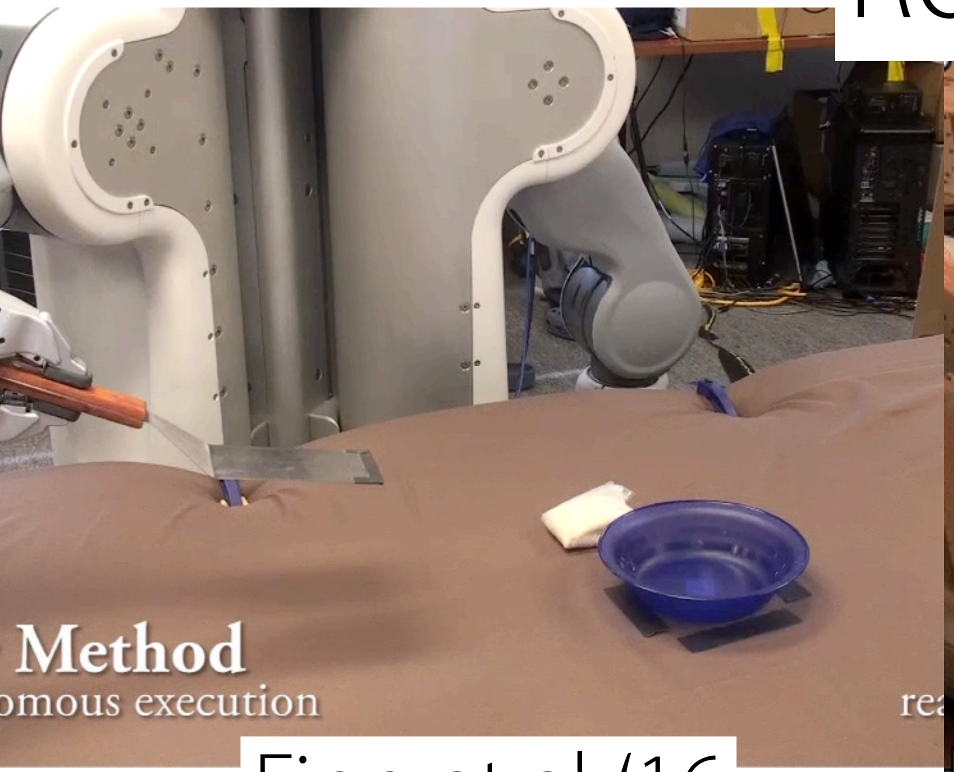


Yevgen

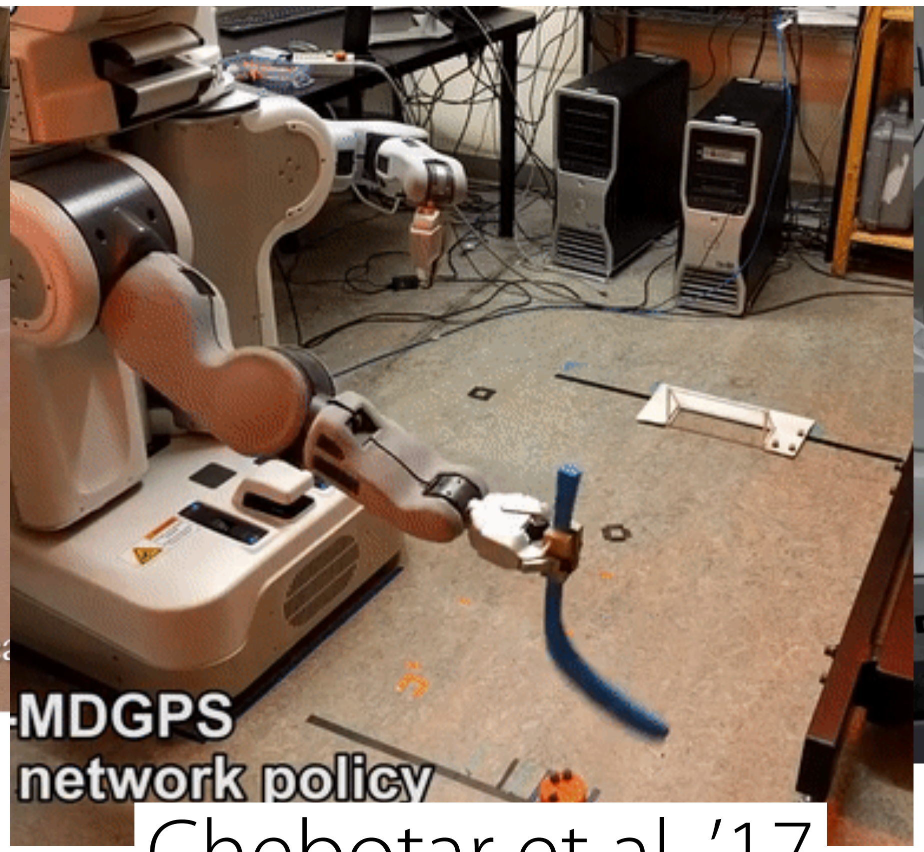
Yevgen is doing more work than the robot!

It's not practical to collect a lot of data this way.

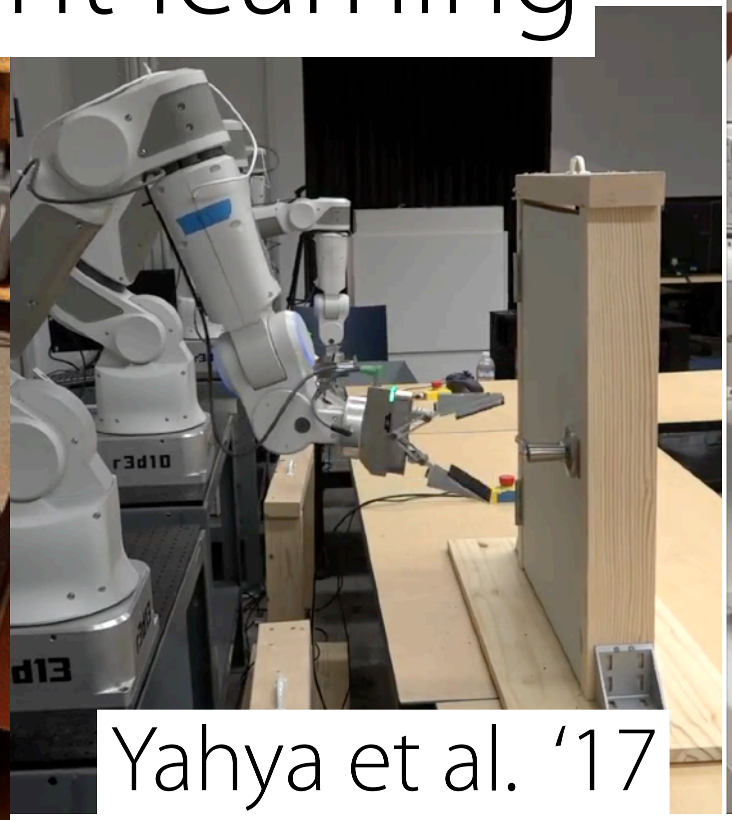
Robot reinforcement learning



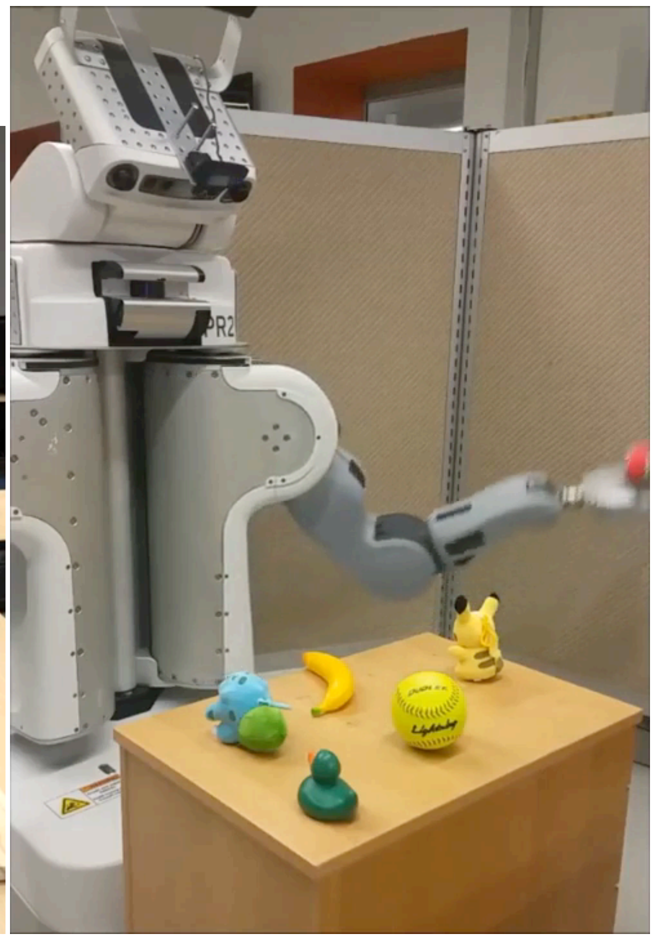
Finn et al.'16



Chebotar et al. '17



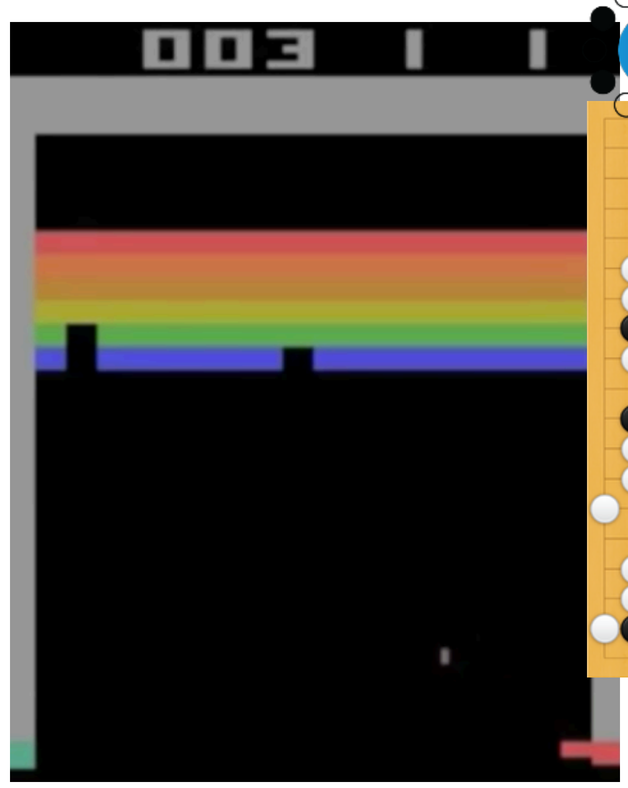
Yahya et al. '17



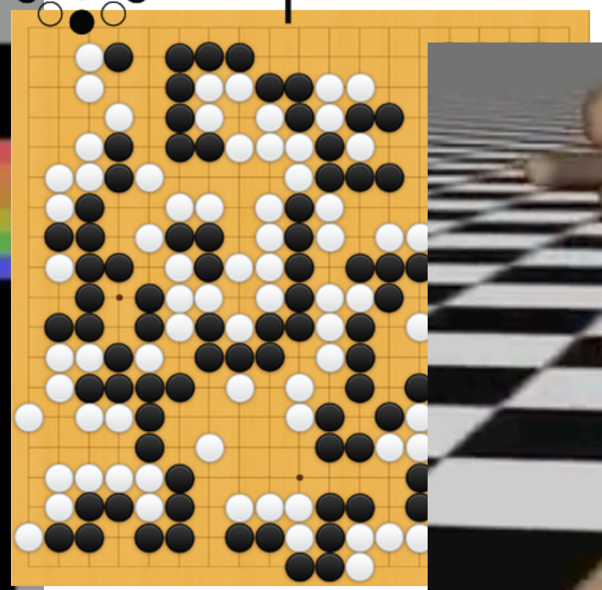
Ghadirzadeh et al. '17

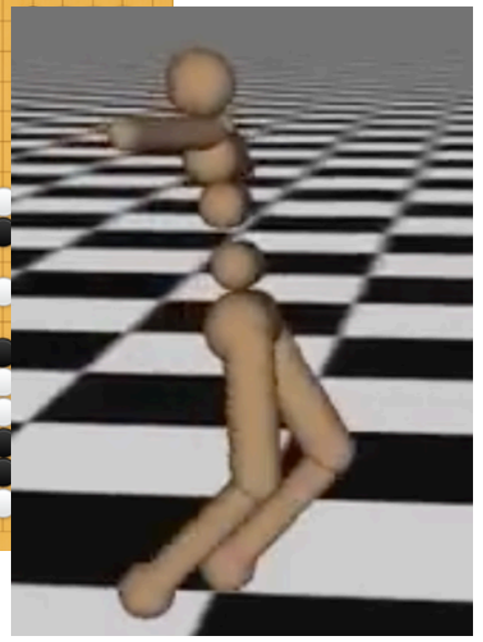
Reinforcement learning

AlphaGo



Atari





locomotion

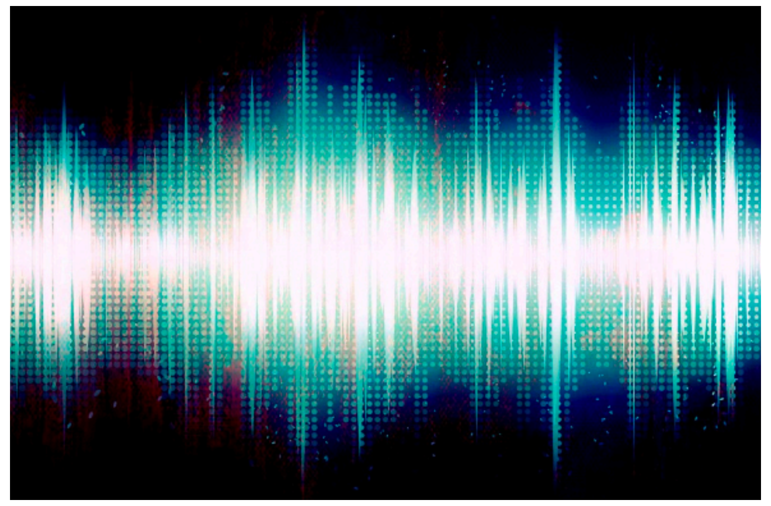
Learn **one task** in **one environment**, **starting from scratch** using **detailed supervision**

Not just a problem with reinforcement learning & robotics.

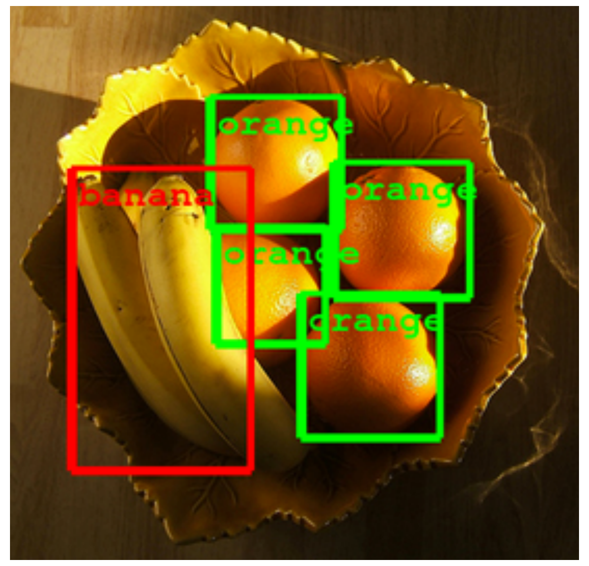
specialists

[single task]

machine translation



speech recognition



object detection

More diverse, yet still **one task**, **from scratch**, with **detailed supervision**

Source: https://youtu.be/8vNxjwt2AqY

vs.

Source: https://i.imgur.com/hJlVfZ5.jpg

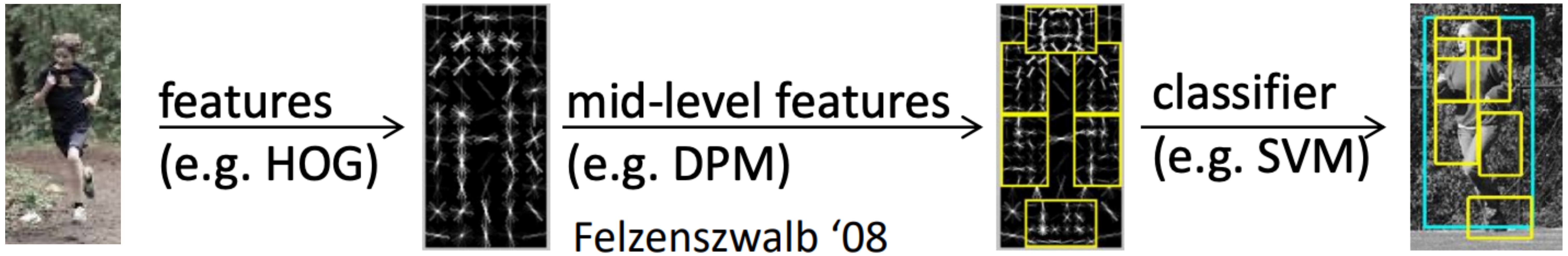
Why should we care about multi-task & meta-learning?

…beyond the robots and general-purpose ML systems

deep
V
Why should we care about multi-task & meta-learning?

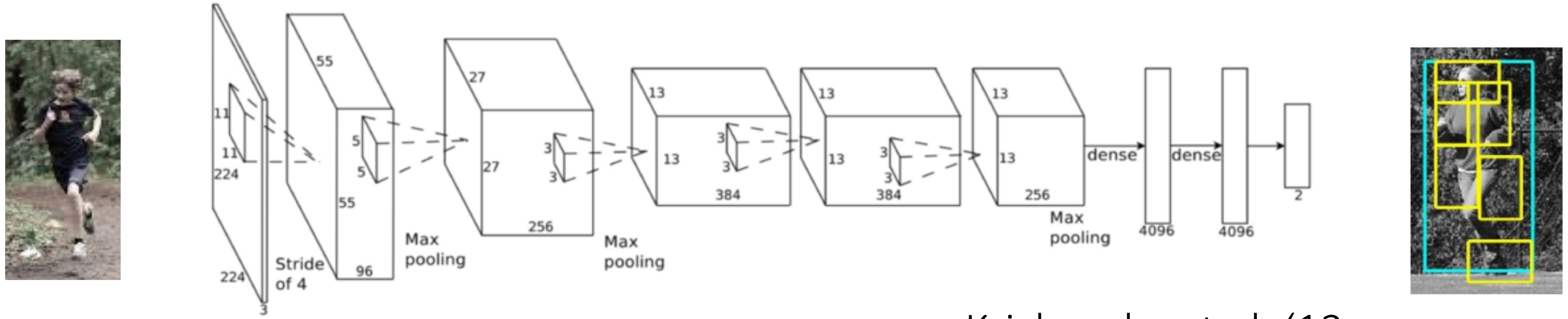
…beyond the robots and general-purpose ML systems

Standard computer vision: hand-designed features



Felzenszwalb '08

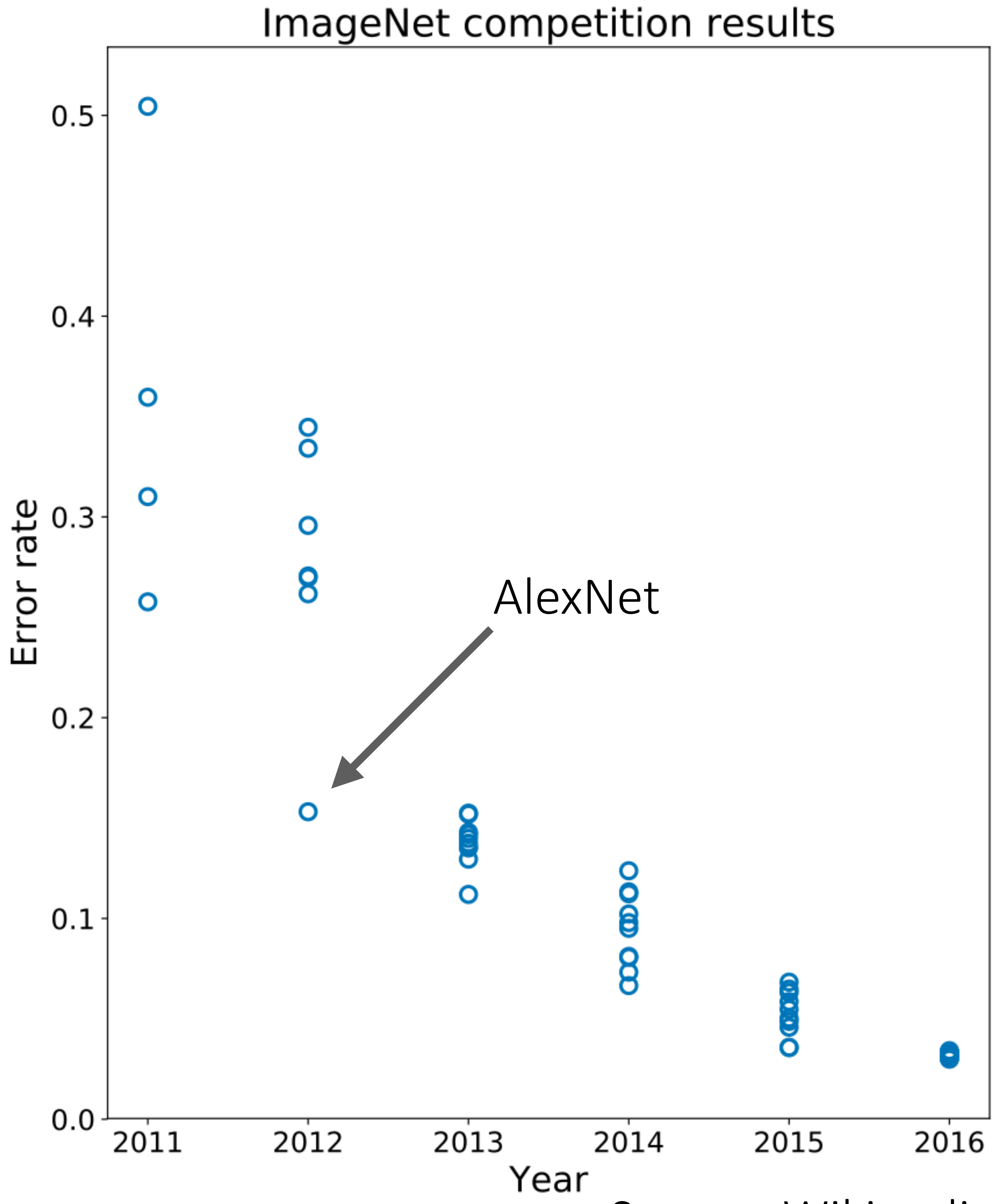
Modern computer vision: end-to-end training



Krizhevsky et al. '12

Deep learning allows us to handle *unstructured inputs* (pixels, language, sensor readings, etc.) without hand-engineering features, with less domain knowledge

Deep learning for object classification



Source: Wikipedia

Deep learning for machine translation

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

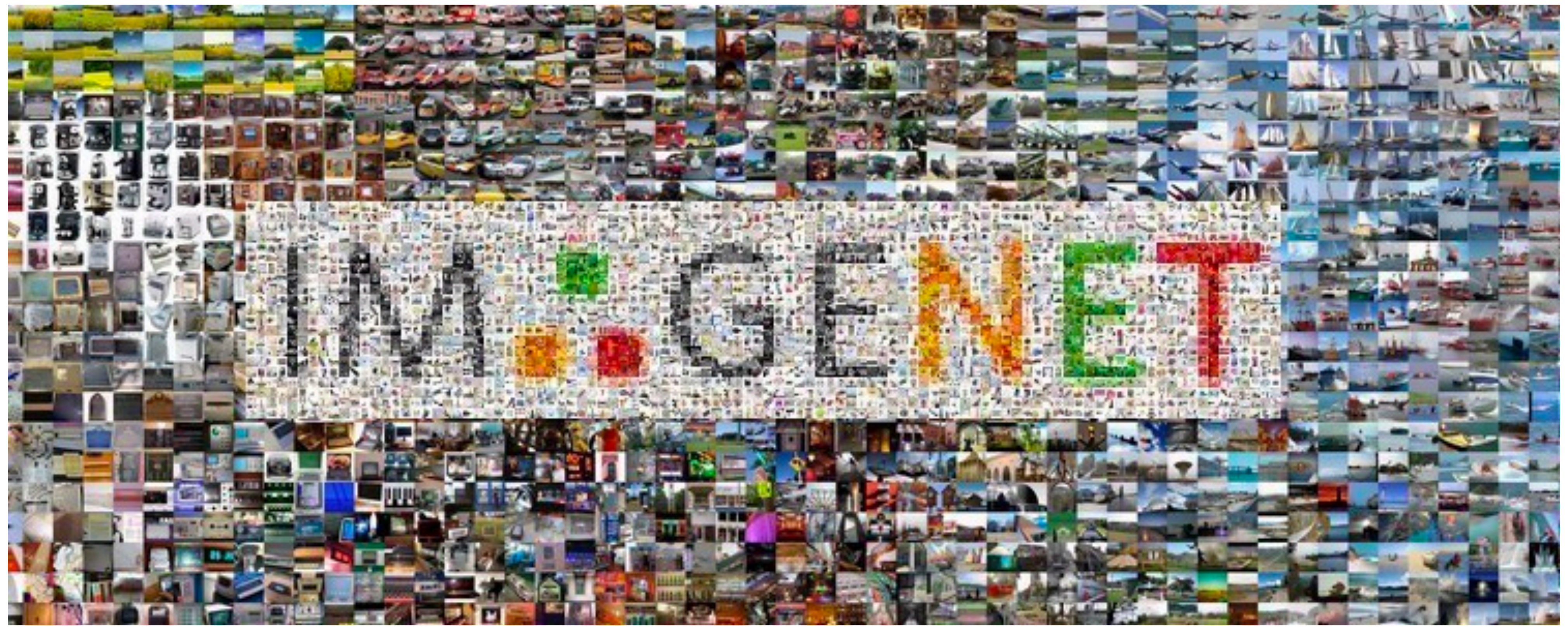
Human evaluation scores on scale of 0 to 6

PBMT: Phrase-based machine translation

GNMT: Google's neural machine translation (in 2016)

Why deep **multi-task** and **meta-learning**?

Large, diverse data (+ large models) $\xrightarrow{\text{deep learning}}$ Broad generalization



Russakovsky et al. '14

Wu et al. '16

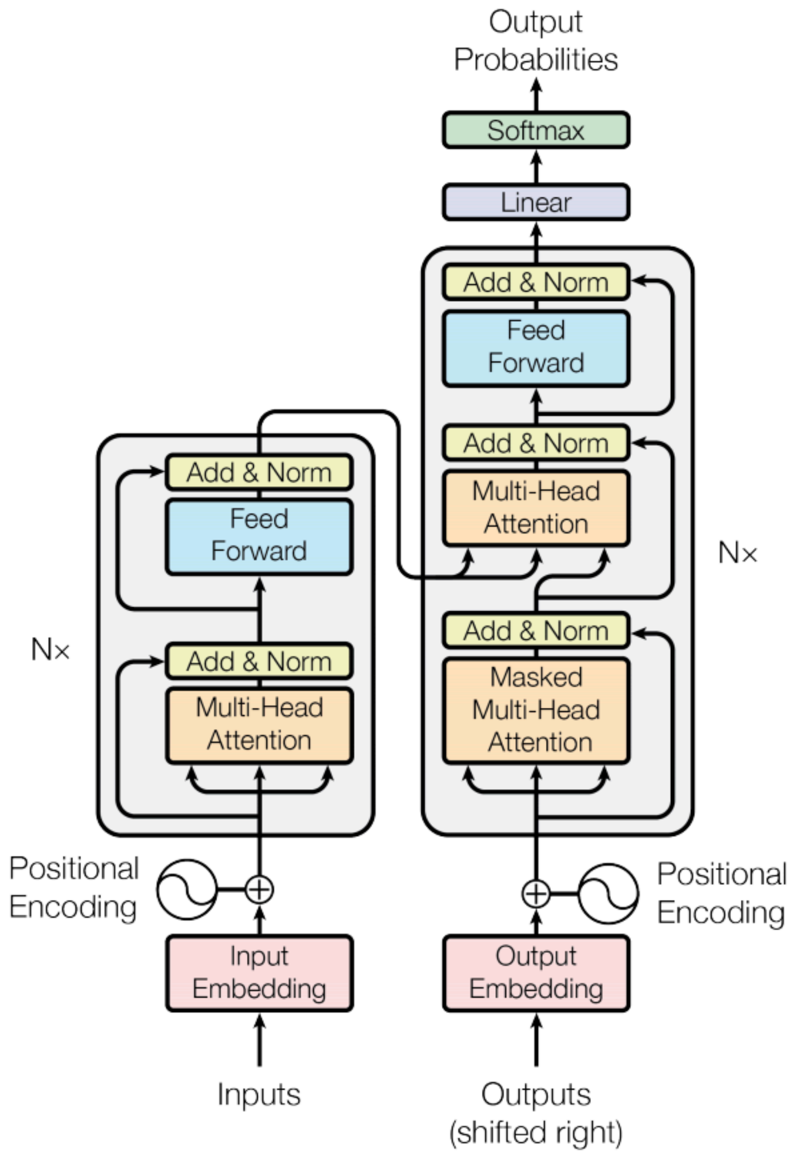


Figure 1: The Transformer - model architecture.

Vaswani et al. '18

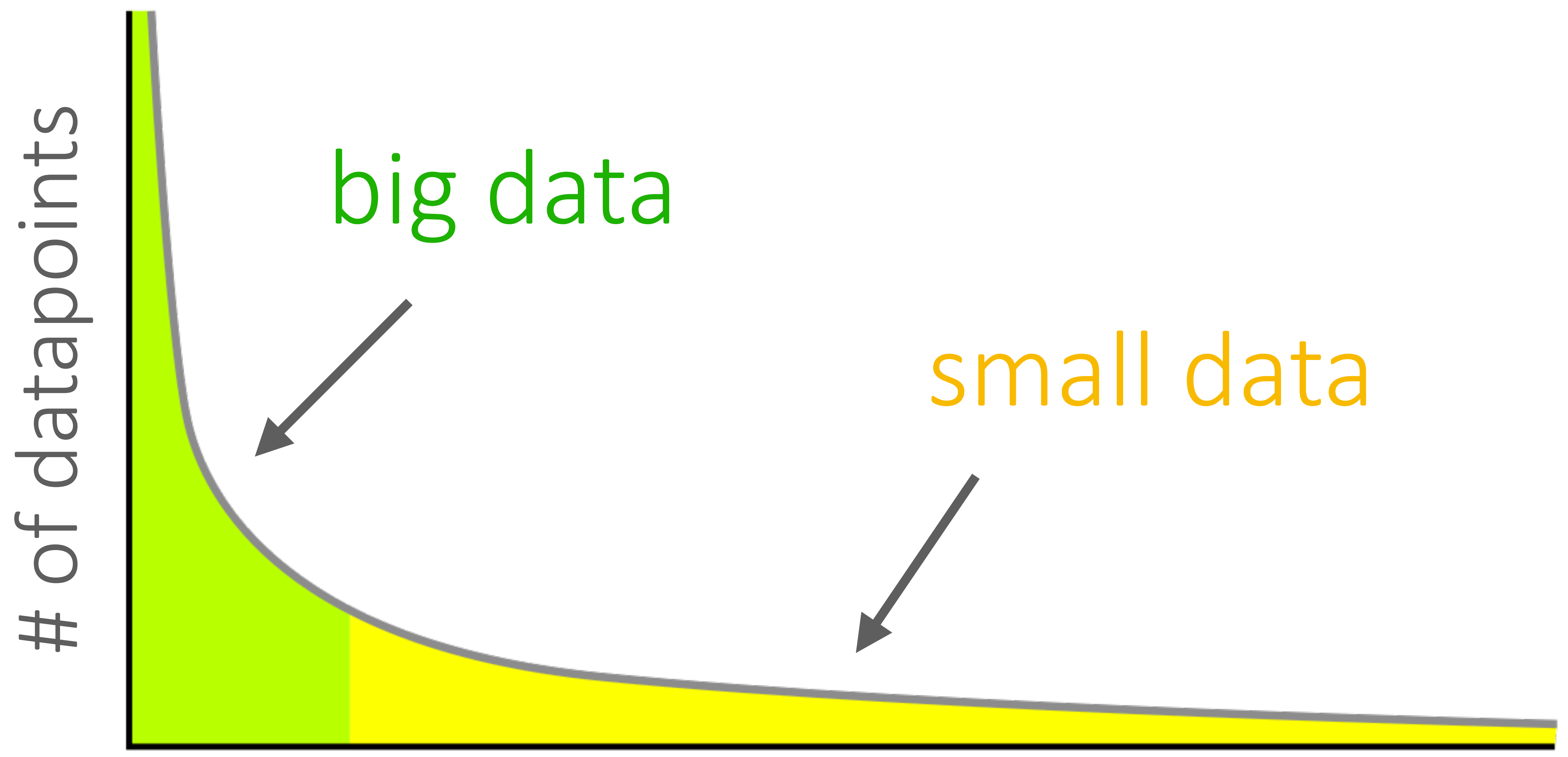
What if you don't have a large dataset?

medical imaging robotics personalized education,

tr ations

Impractical to learn from scratch for each disease, each robot, each person, each language, each **task**

What if your data has a long tail?



objects encountered
interactions with people
words heard
driving scenarios
⋮

This setting breaks standard machine learning paradigms.

What if you need to quickly learn something new?

about a new person, for a new task, about a new environment, etc.

training data

test datapoint

Braque

Cezanne

By Braque or Cezanne?

What if you need to quickly learn something new?

about a new person, for a new task, about a new environment, etc.

“few-shot learning”

How did you accomplish this?

by leveraging prior experience!

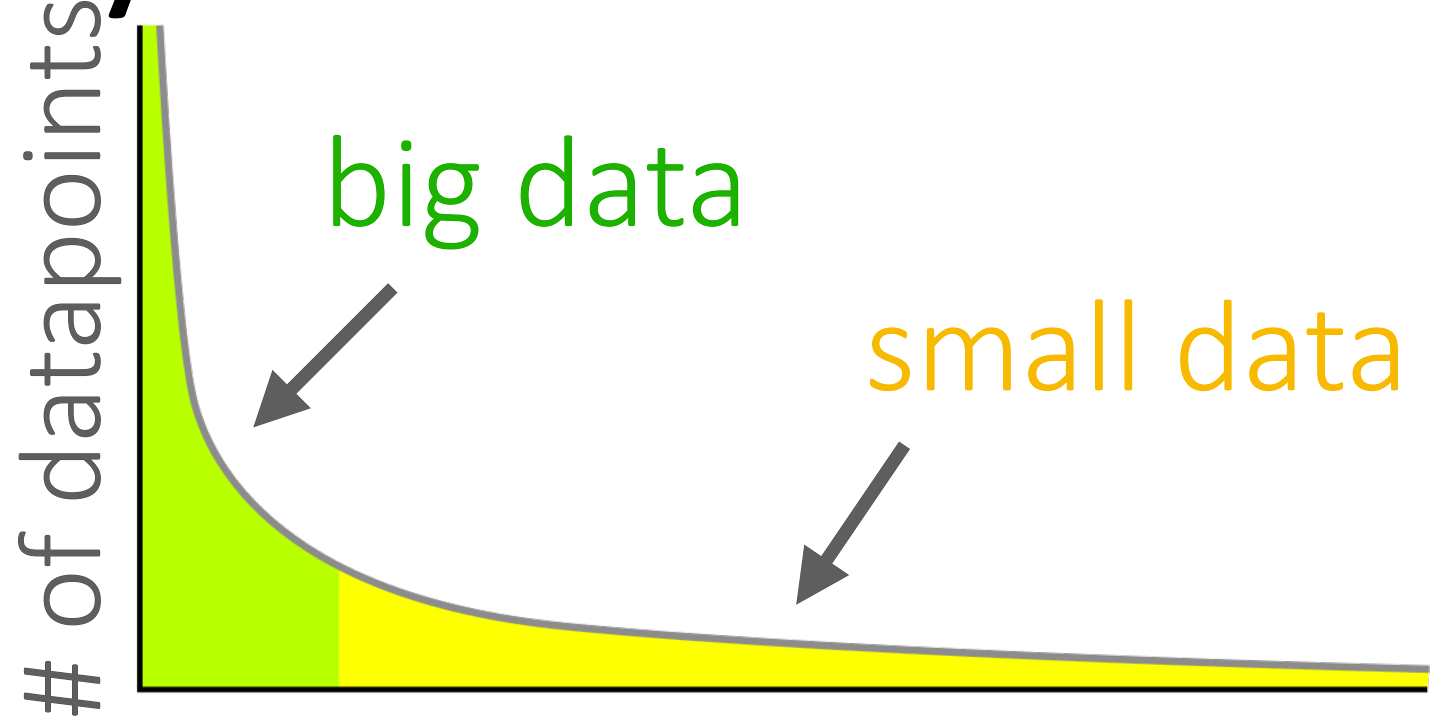
What if you want a more general-purpose AI system?

Learning each task from scratch won't cut it.

What if you don't have a large dataset?

medical imaging robotics personalized education, translation for rare languages medicine, recommendations

What if your data has a long tail?



What if you need to quickly learn something new?

about a new person, for a new task, about a new environment, etc.

This is where elements of multi-task learning can come into play.

Why now?

Why should we study deep multi-task & meta-learning now?

Multitask Learning*

RICH CARUANA

Multitask Learning (MTL) is an inductive transfer mechanism whose principle goal is to improve generalization performance. MTL improves generalization by leveraging the domain-specific information contained in the training signals of *related* tasks. It does this by training tasks in parallel while using a shared representation. In effect, the training signals for the extra tasks serve as an inductive bias. Section 1.2 argues that inductive transfer is important if we wish to scale tabula rasa learning to complex, real-world tasks. Section 1.3 presents the simplest method we know for doing multitask inductive transfer, adding extra tasks (i.e., extra outputs) to a backpropagation net. Because the MTL net uses a shared hidden layer trained in parallel on all the tasks, what is learned for each task can help other tasks be learned better. Section 1.4 argues that it is reasonable to view training signals as an inductive bias when they are used this way.

Caruana, 1997

Is Learning The n-th Thing Any Easier Than Learning The First?

Sebastian Thrun[1]

They are often able to generalize correctly even from a single training example [2, 10]. One of the key aspects of the learning problem faced by humans, which differs from the vast majority of problems studied in the field of neural network learning, is the fact that humans encounter a whole stream of learning problems over their entire lifetime. When faced with a new thing to learn, humans can usually exploit an enormous amount of training data and experiences that stem from other, related learning tasks. For example, when learning to drive a car, years of learning experience with basic motor skills, typical traffic patterns, logical reasoning, language and much more precede and influence this learning task. The transfer of knowledge across learning tasks seems to play an essential role for generalizing accurately, particularly when training data is scarce.

Thrun, 1998

On the Optimization of a Synaptic Learning Rule

Samy Bengio Yoshua Bengio Jocelyn Cloutier Jan Gecsei

Université de Montréal, Département IRO

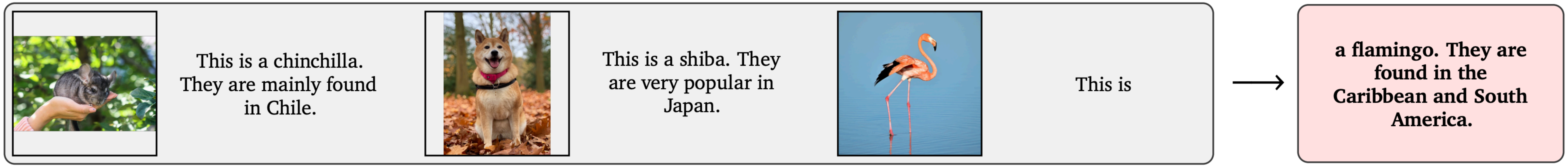
This paper presents a new approach to neural modeling based on the idea of using an automated method to optimize the parameters of a synaptic learning rule. The synaptic modification rule is considered as a parametric function. This function has *local* inputs and is the same in many neurons. We can use standard optimization methods to select appropriate parameters for a given type of task. We also present a theoretical analysis permitting to study the *generalization* property of such parametric learning rules. By generalization, we mean the possibility for the learning rule to learn to solve *new* tasks. Experiments were performed on three types of problems: a

Bengio et al. 1992

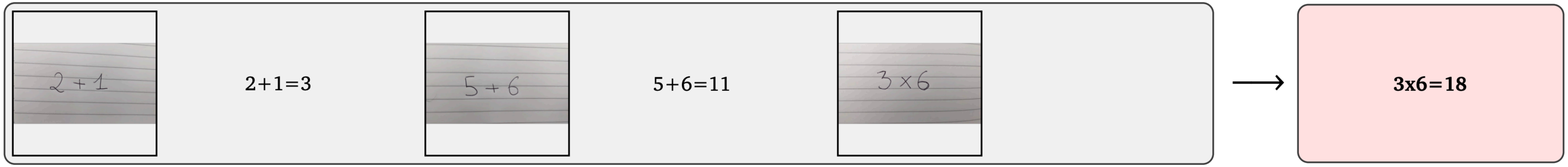
These methods are continuing to play a major role in AI.

Visual language models can learn many distinct tasks.

Object recognition:



Reading & arithmetic:



Counting:

Alayrac*, Donahue*, Luc*, Miech* et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022

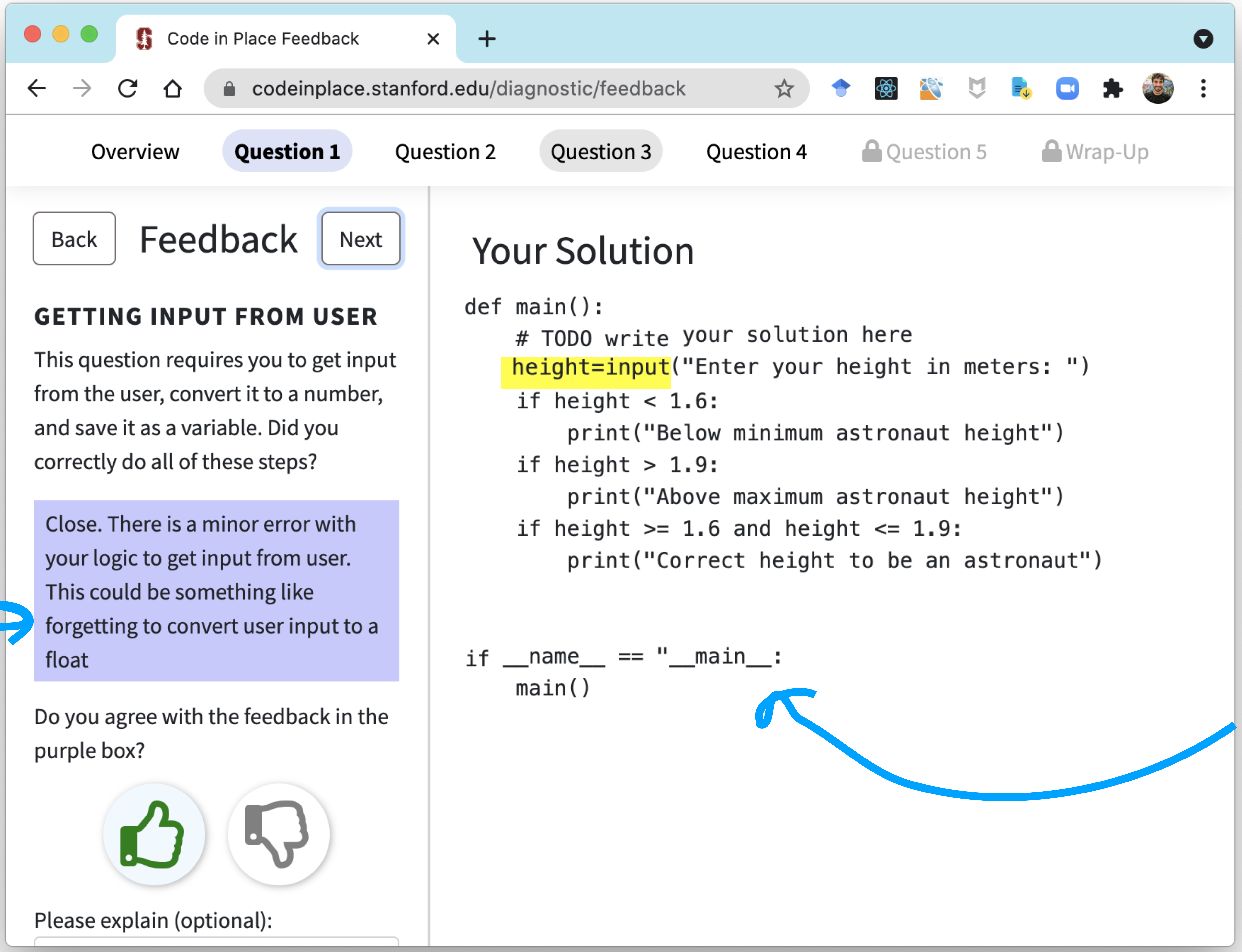
These methods are continuing to play a major role in AI.

Meta-learning enables automatic feedback on student work on new problems.

Code-in-Place 2021

Humans gave feedback on ~1k student programs.
Meta-learning system gave feedback on the remaining ~15k.

Generated feedback



Syntax error here would prevent unit tests from being useful

Wu, Goodman, Piech, Finn. ProtoTransformer: A Meta-Learning Approach to Providing Student Feedback. 2021

These methods are continuing to play a major role in AI.

Multilingual machine translation

Massively Multilingual Neural Machine Translation

Roee Aharoni* Bar Ilan University Ramat-Gan Israel roee.aharoni@gmail.com

Melvin Johnson and **Orhan Firat** Google AI Mountain View California melvinp,orhanf@google.com

while supporting up to 59 languages. Our experiments on a large-scale dataset with 102 languages to and from English and up to one million examples per direction also show promising results, surpassing strong bilingual baselines and encouraging future work on massively multilingual NMT.

NAACL, 2019

YouTube recommendations

Recommending What Video to Watch Next: A Multitask Ranking System

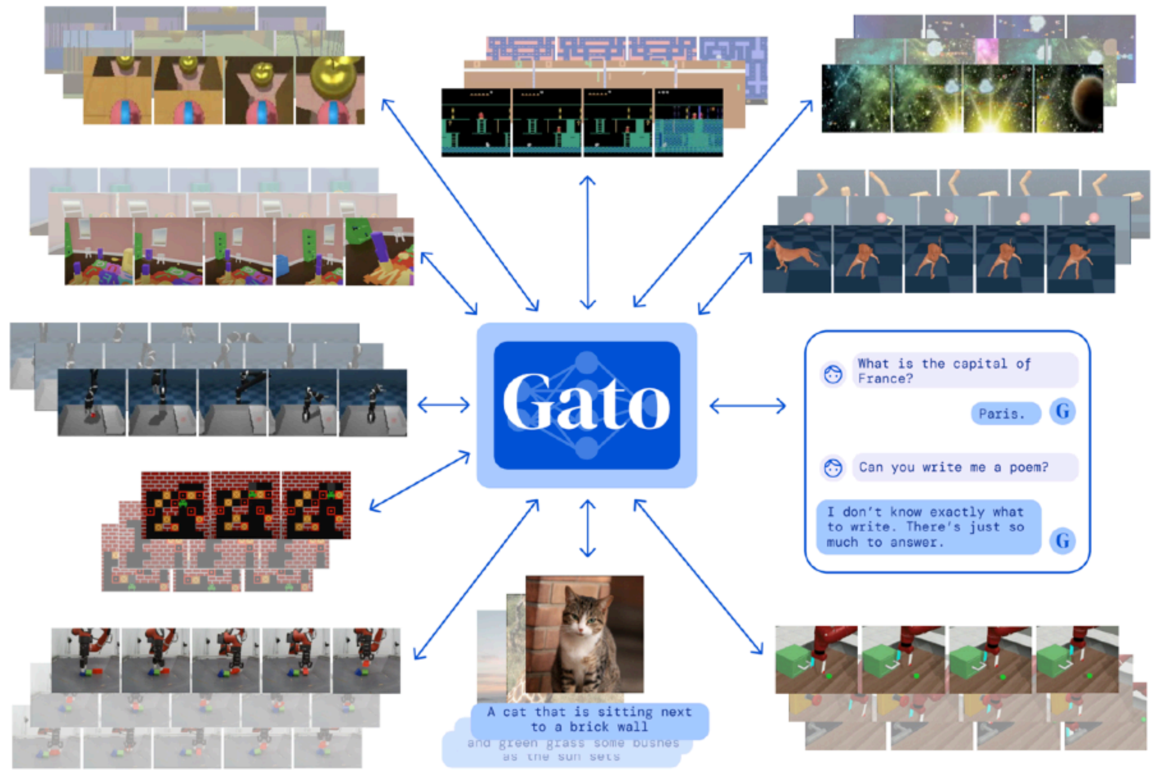
Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi
Google, Inc.
{zhezhao,lichan,liwei,jilinc,aniruddhnath,shawnandrews,aditeek,nlogn,xinyang,edchi}@google.com

In this paper, we introduce a large scale multi-objective ranking system for recommending what video to watch next on an industrial video sharing platform. The system faces many real-world challenges, including the presence of multiple competing ranking objectives, as well as implicit selection biases in user feedback. To

RecSys 2019

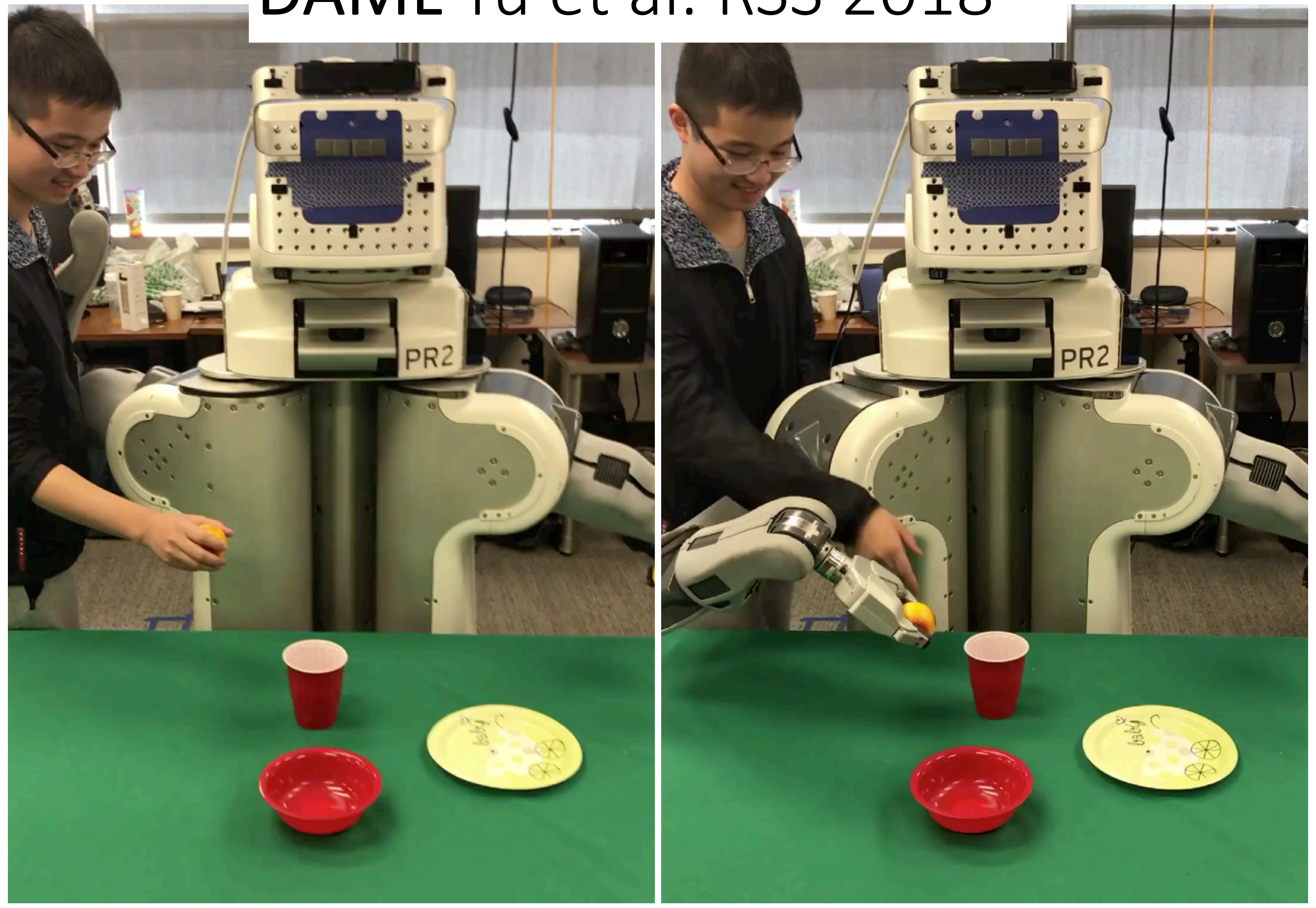
A Generalist Agent

Reed et al. 2022



One-shot imitation from humans

DAML Yu et al. RSS 2018



Its success is important for the **democratization** of deep learning.

ImageNet	1.2 million images and labels
WMT '14 English - French	40.8 million paired sentences
Switchboard Speech Dataset	300 hours of labeled data
Kaggle's Diabetic Retinopathy Detection dataset	35K labeled images
Adaptive epilepsy treatment with RL Guez et al. '08	< 1 hour of data
Learning for robotic manipulation Finn et al. '16	< 15 min of data

But, we also still have many open questions and challenges!

What is a task?

What is a task?

Informally: dataset $\mathcal{D}$ loss function $\mathcal{L}$ $\longrightarrow$ model f_θ

Different tasks can vary based on:

- different objects
- different people
- different objectives
- different lighting conditions
- different words
- different languages
- …

Not *just* different "tasks"

Critical Assumption

The bad news: Different tasks need to share some structure.

If this doesn't hold, you are better off using single-task learning.

The good news: There are many tasks with shared structure!



Even if the tasks are seemingly unrelated:

- The **laws of physics** underly real data.
- People are **all organisms with intentions**.
- The **rules of English** underly English language data.
- Languages all develop for **similar purposes**.

This leads to far greater structure than random tasks.

Informal Problem Definitions

We'll define these more formally later.

The multi-task learning problem: Learn a set of tasks more quickly or more proficiently than learning them independently.

The transfer learning problem: Given data on previous task(s), learn a new task more quickly and/or more proficiently.

This course: anything that solves these problem statements.

Doesn't multi-task learning reduce to single-task learning?

$$\mathcal{D} = \bigcup \mathcal{D}_i \qquad \mathcal{L} = \sum \mathcal{L}_i$$

Are we done with the course?

Doesn't multi-task learning reduce to single-task learning?

Yes, it can!

Aggregating the data across tasks & learning a single model is one approach to multi-task learning.

But, what if you want to learn *new* tasks?

And, how do we tell the model what task to do?

And, what if aggregating doesn't work?

Reminders

1. Homework 0 is out
2. Start forming final project groups if you want to work in a group

Next time (Weds): Multi-Task Learning Basics