Vision and Action:
Visual Approaches to Robotic Control

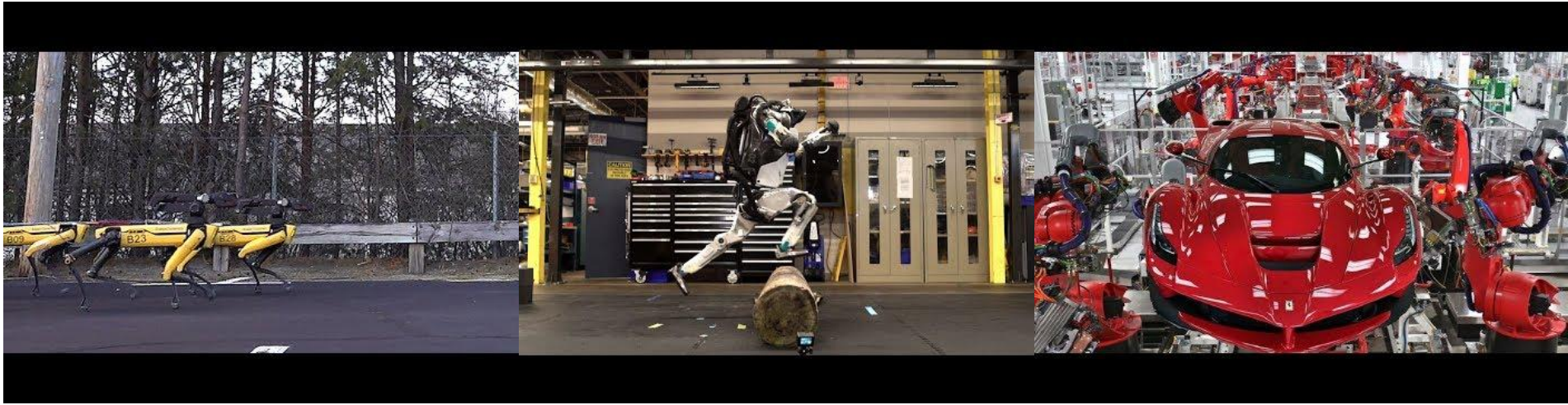
Dinesh Jayaraman

TEXAS
The University of Texas at Austin

Berkeley
UNIVERSITY OF CALIFORNIA

Penn
University of Pennsylvania

Robots without Perception



Robots in the Wild

Robots 10 years ago ...*

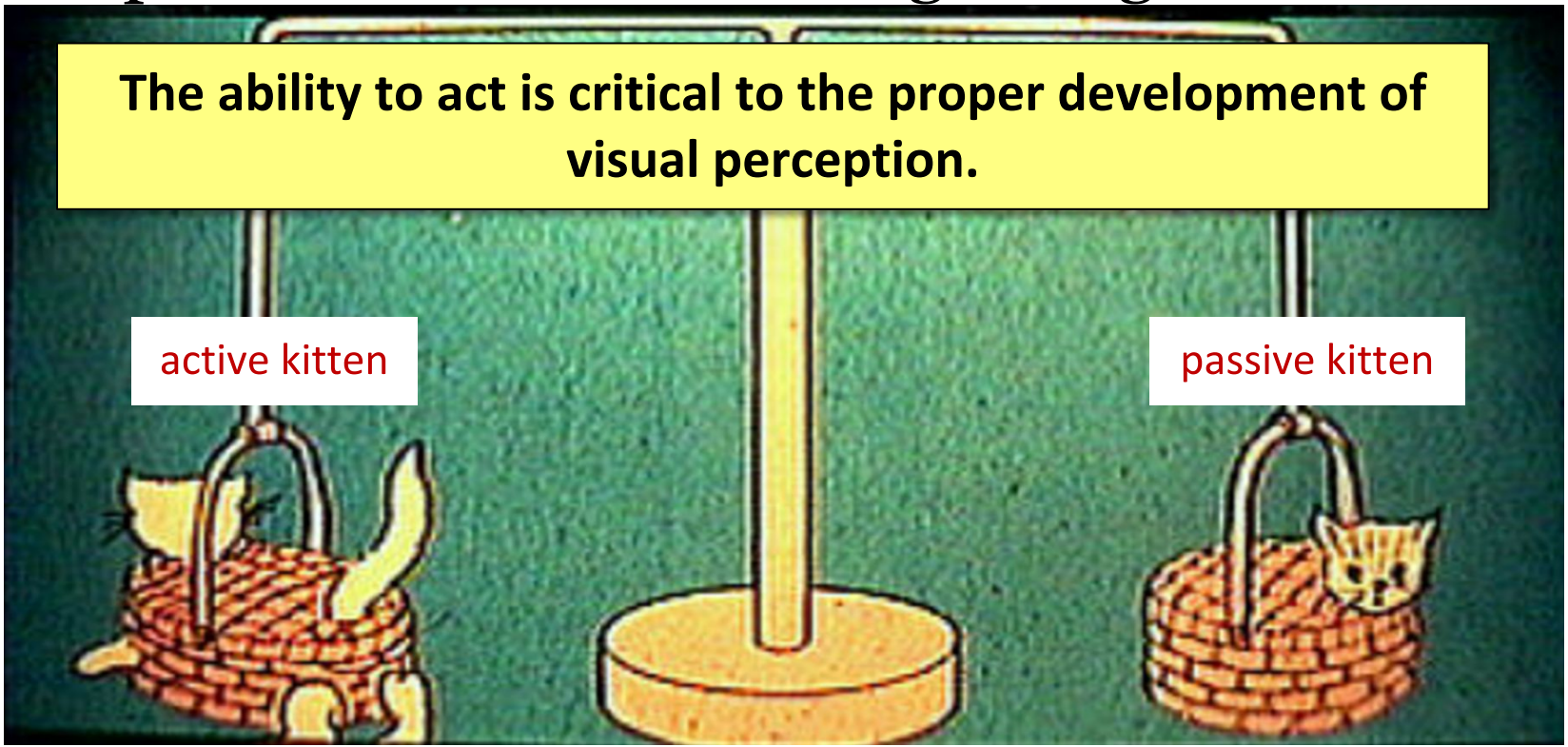
*Caveat:
Human-teleoperated!

Moral: *Control* is the bottleneck.

(esp control requiring good perception)

Slide idea: Pieter Abbeel

Perception and Action in Biological Agents



[Held & Hein, 1963]

[Moravec et al, 1984][Wilson et al, 2002] [Smith et al 2005] …

Perception-Action Cycles

Vision aids action

[high-bandwidth information --- useful across tasks and environments]

AND

Action in turn aids vision

[active vision: how can I act to improve my understanding of the world,

state representation: "what is important about my raw visual observations",

…]

The rest of this lecture

- **Intro to Learning for Sequential Decision Making**
  - Overview of the setting
- **Learning from Demonstrations**
  - Behavior Cloning: Imitation as Supervised Learning
  - *Better behavior cloning: DAGGER, examples*
  - Alternatives: Inverse RL, Adversarial Imitation, examples
- **Learning by Trial-and-Error: Model-Based and Model-Free RL**
  - What does RL seek to do? (Notations, Definitions …)
  - *Model-predictive control with visual predictors, examples.*
  - A control systems instantiation of model-based visual control: visual servoing
  - *Q-learning: tabular, and with deep network function approximator, examples*
- **What can Action do for Perception?**
  - Examples: Active Perception, Feature Learning
- **Systematizing the study of robot learning**

Sequential Decision Making

[the framework]

The Sequential Decision Making Problem

Markov Decision Process:

(Sequential decision making formalism)

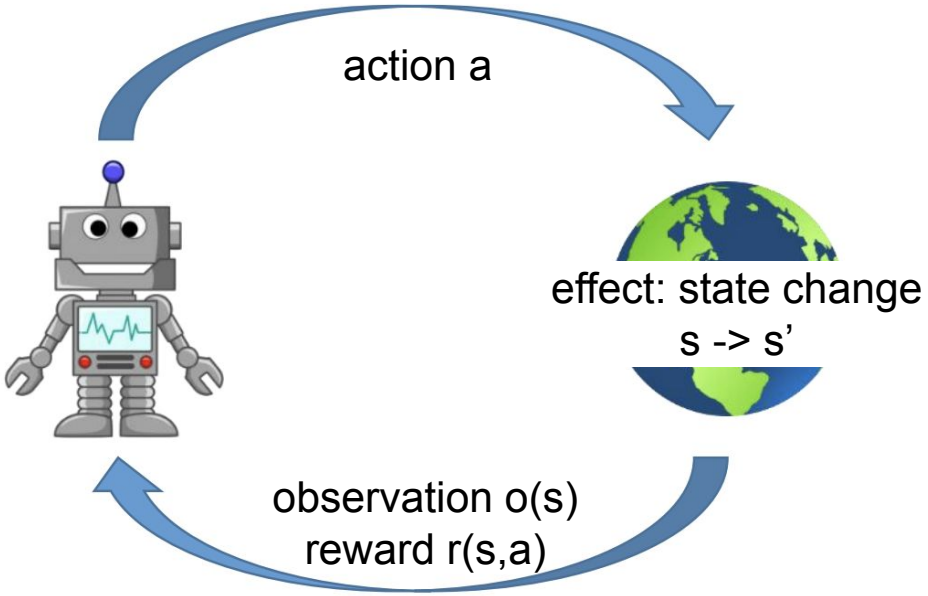
Set of states S, set of actions A

Transition dynamics

$$T(s, a, s') = P(s'|s, a)$$

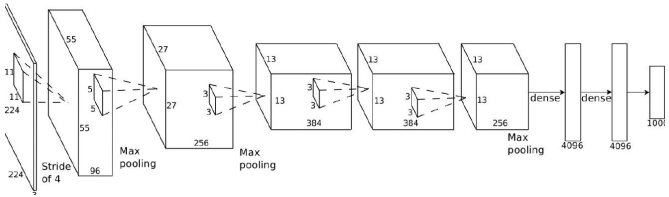
Reward function

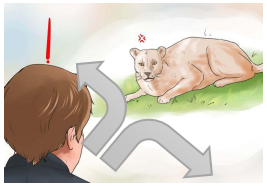
$$R(s, a, s')$$



The Control Problem: Learn a "policy" $\pi(s) : S \rightarrow A$

Terminology & notation

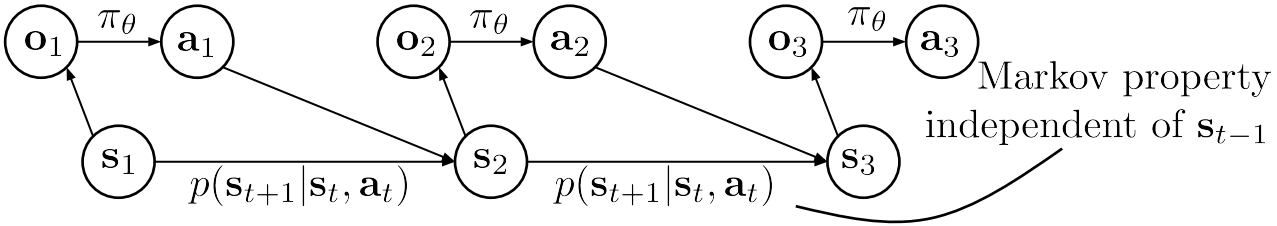




$\mathbf{o}_t$ $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ $\mathbf{a}_t$

$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action

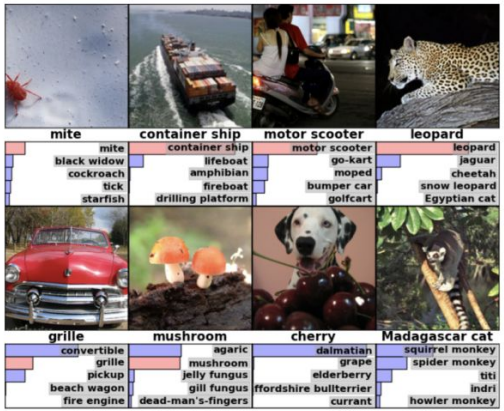
$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)

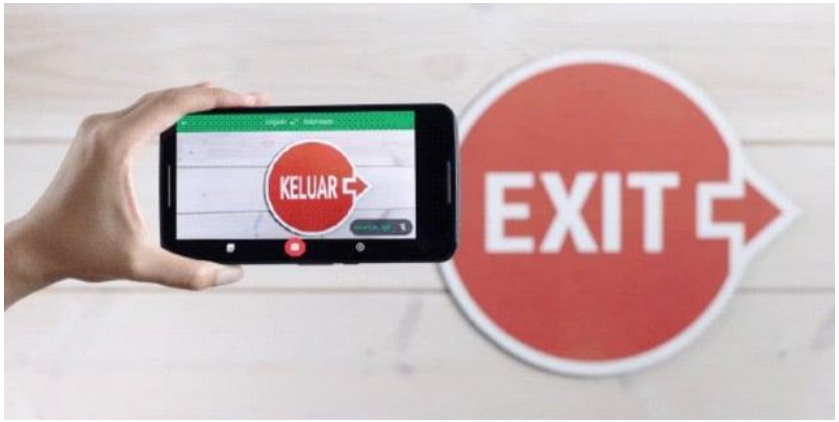


When do we **not** need to worry about sequential decision making?

When your system is making single isolated decision, e.g. classification, regression

When that decision does not affect future decisions





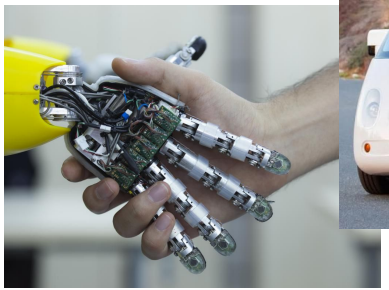
When **should** we worry about sequential decision making?

Limited supervision: you know **what** you want, but not **how** to get it

Actions have consequences

Common Applications

autonomous driving



robotics

language & dialogue (structured

business

finance

Learning from Demonstrations

[a. k. a. Imitation Learning]

Imitation of Televised Models by Infants

Andrew N. Meltzoff
University of Washington

Imitation Approaches

Behavior Cloning

Supervised Learning on expert (s,a) tuples

Inverse Reinforcement Learning

1) Infer reward from human demonstrations
2) Then run reinforcement learning

Others

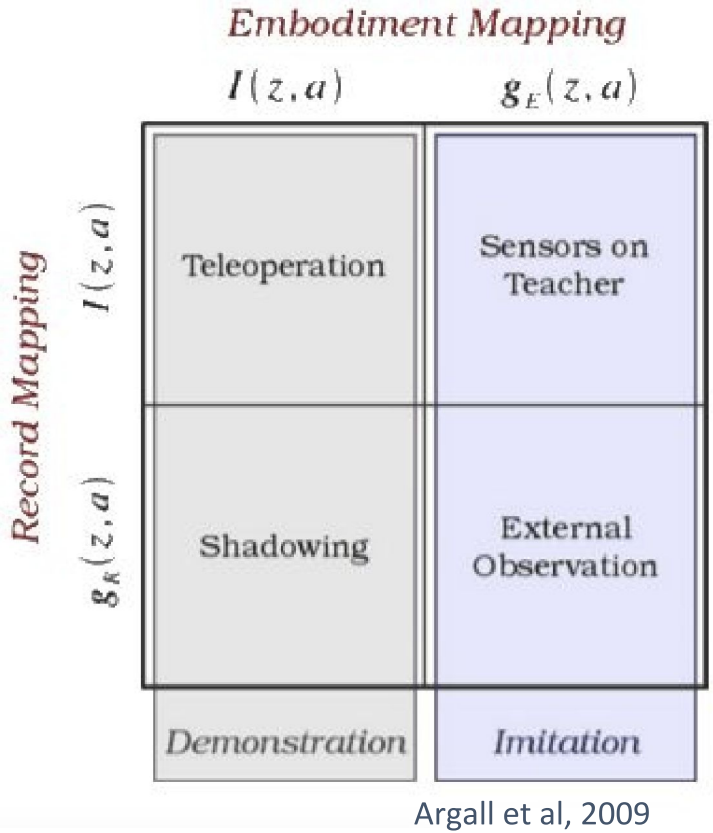
e.g. Adversarial Imitation: learn policies that induce state/state-transition distributions that fool an adversary

E.g. imitation through RL

Demonstration Approaches

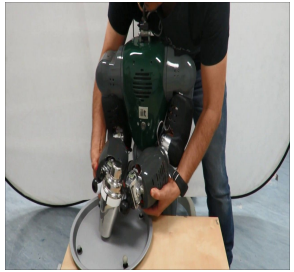
Embodiment mapping: First person (demos) vs third-person (imitation)

Record mapping: are expert demo states directly observed?



Argall et al, 2009

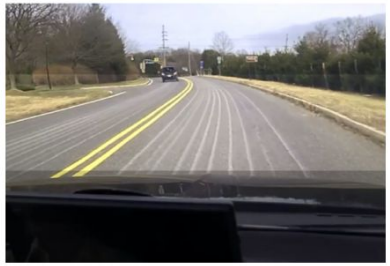
teleoperation



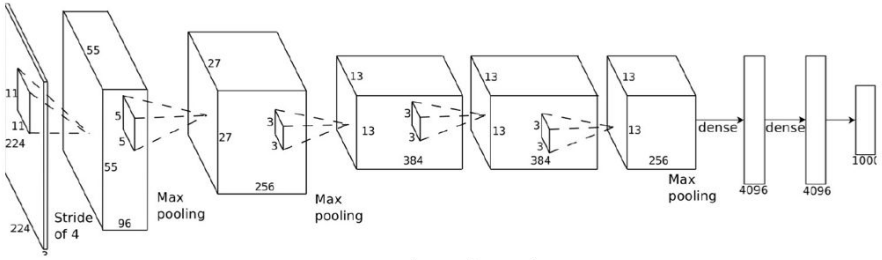
kinesthetic

mocap

web video?



$\mathbf{o}_t$



$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$

$\mathbf{a}_t$

$\mathbf{o}_t$

$\mathbf{a}_t$

training data

supervised learning

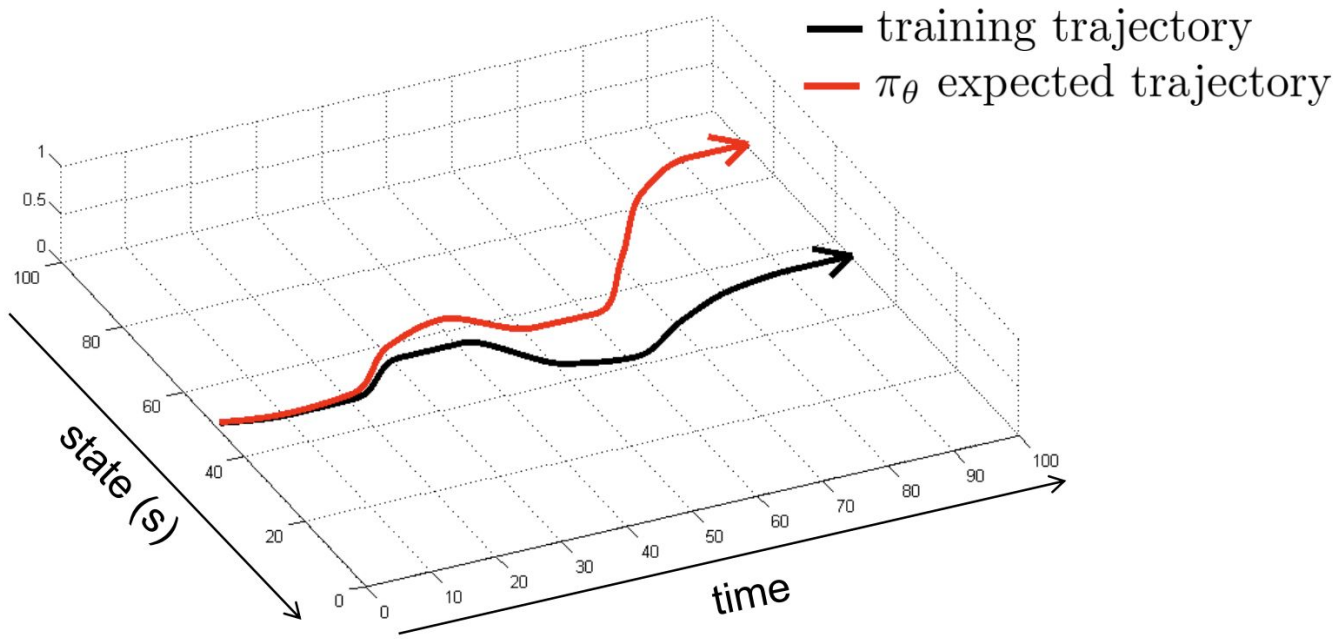
$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$

Control from Imitation: Supervised Learning

Giusti et al, 2016

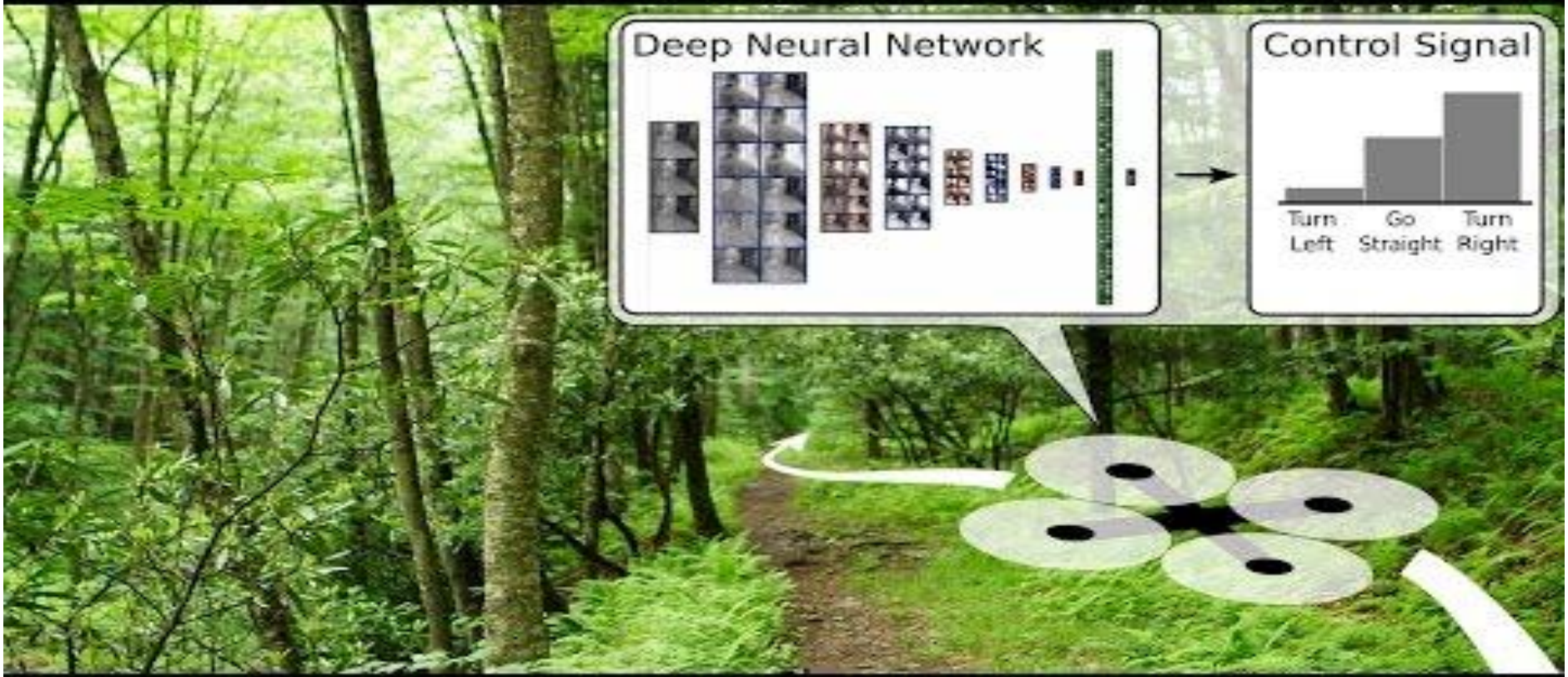
Slide: Sergey Levine

Problem: Distributional Shift in Imitation Learning

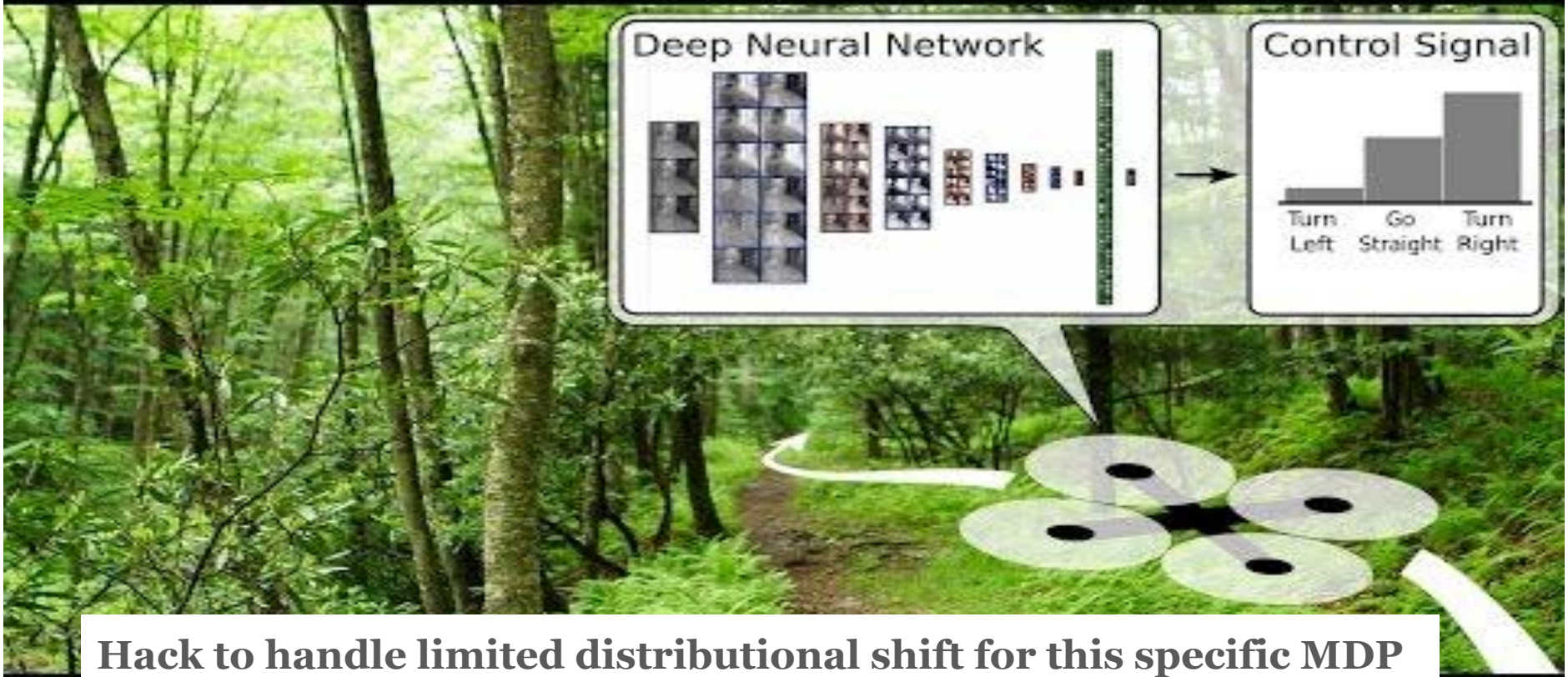


The cloned policy is imperfect; this leads to accumulating errors, and the agent soon encounters out-of-domain states, leading to failure.

Working quadrotor navigation!



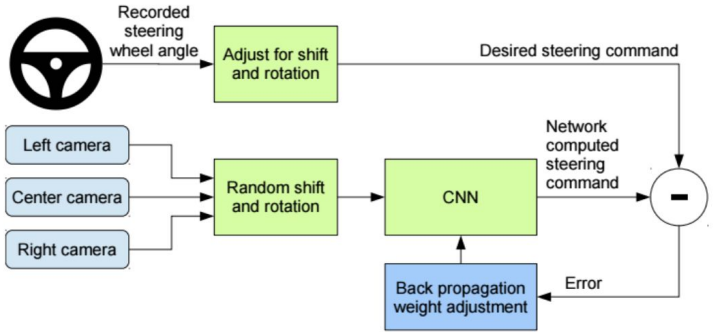
How? Navigation-Specific Data Augmentation Hack

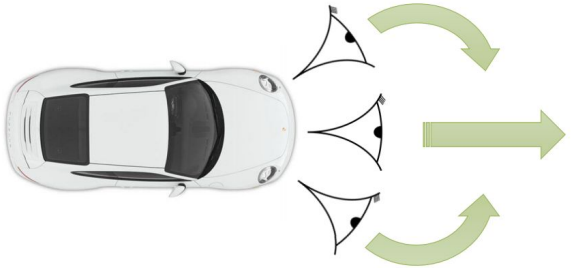


Hack to handle limited distributional shift for this specific MDP

… exploits the geometry of the state and action space.

Car Driving with the Data Augmentation Hack

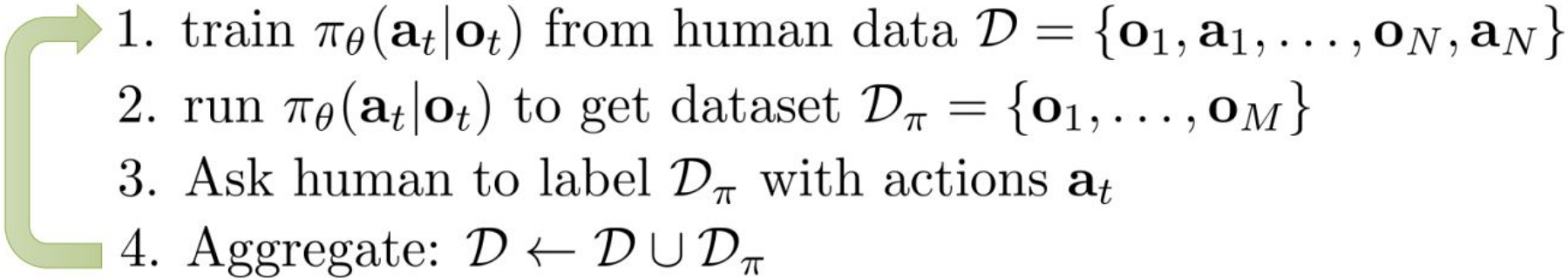




Bojarski et al, 2016

Active Imitation Learning: DAGGER

More general trick for handling distributional shift: requery expert on new states encountered by the initial cloned policy upon execution, then retrain.



Ross et al, DAGGER, 2011

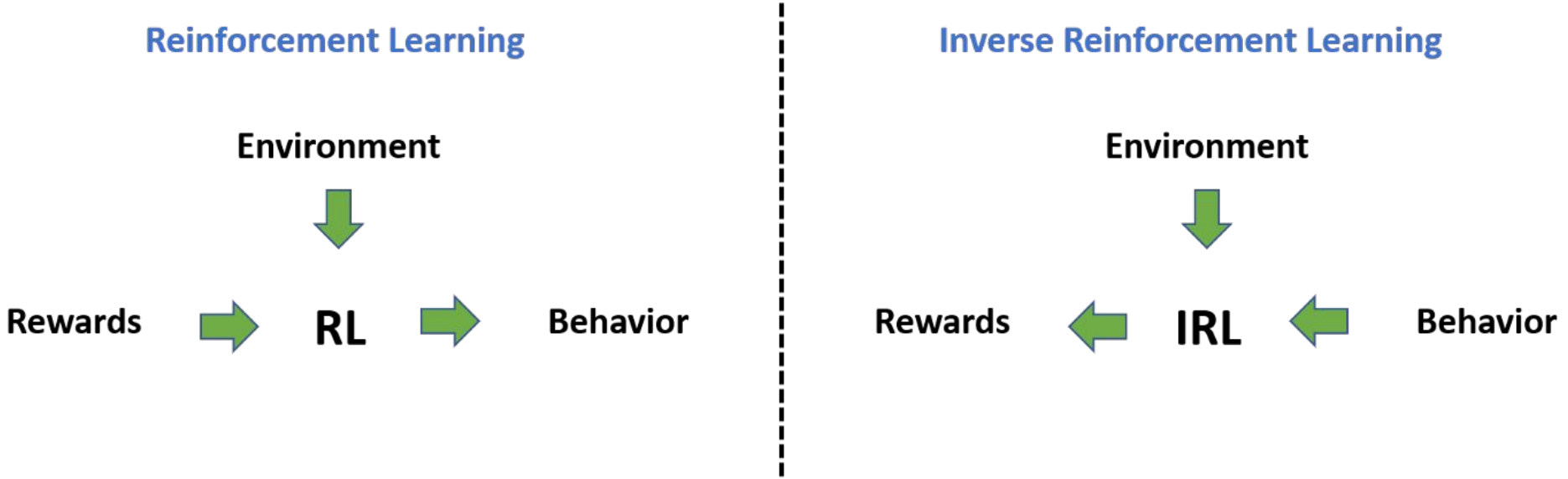
Behavioral Cloning

- Easy to implement, has been shown to work well in real world robots, driving etc.
- Cons:
  a. No guarantee of ``consistent behavior" --- no prior on policies being optimal for some MDP, so you could, e.g., keep switching from A to B and B to A.
  b. Usually requires careful feature engineering to avoid problems with feedback and causality. More on this later!

Inverse Reinforcement Learning

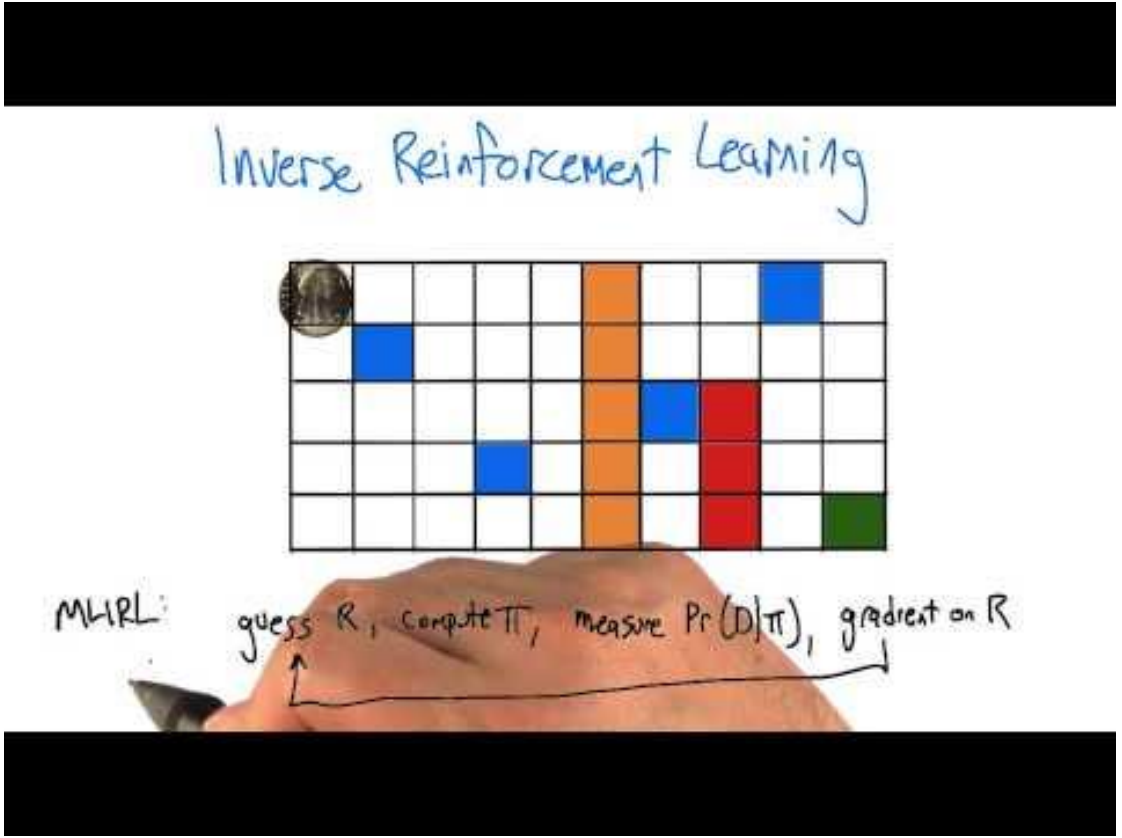
"After all, the entire field of reinforcement learning is founded on the presupposition that the reward function, rather than the policy, is the most succinct, robust, and transferable definition of the task." - Ng and Russell, 2000

Hard to manually specify a reward, so learn it from demos?

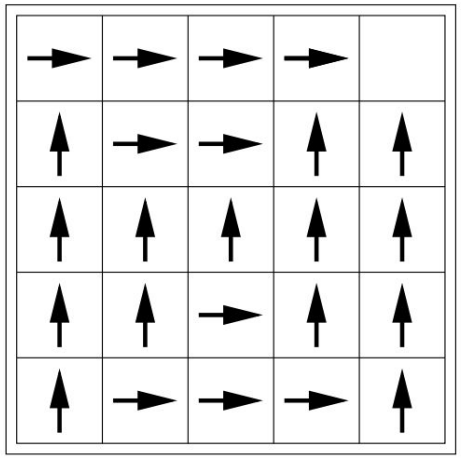


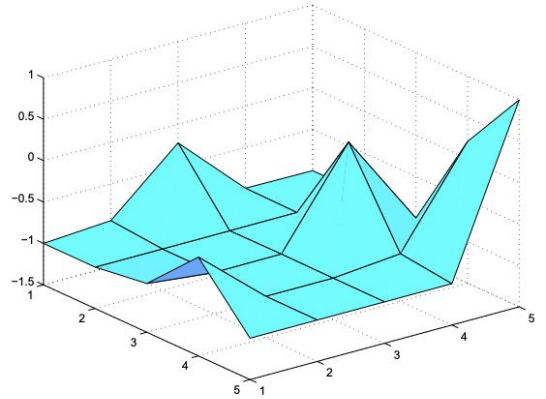
Inverse Reinforcement Learning: Intuition

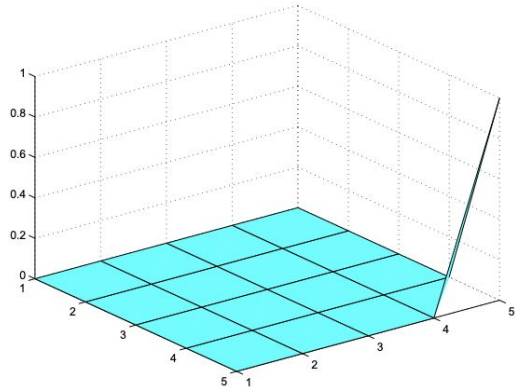
[Udacity video]

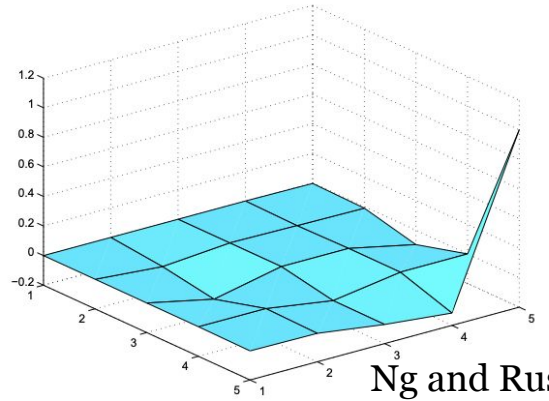


Example: Inverse Reinforcement Learning





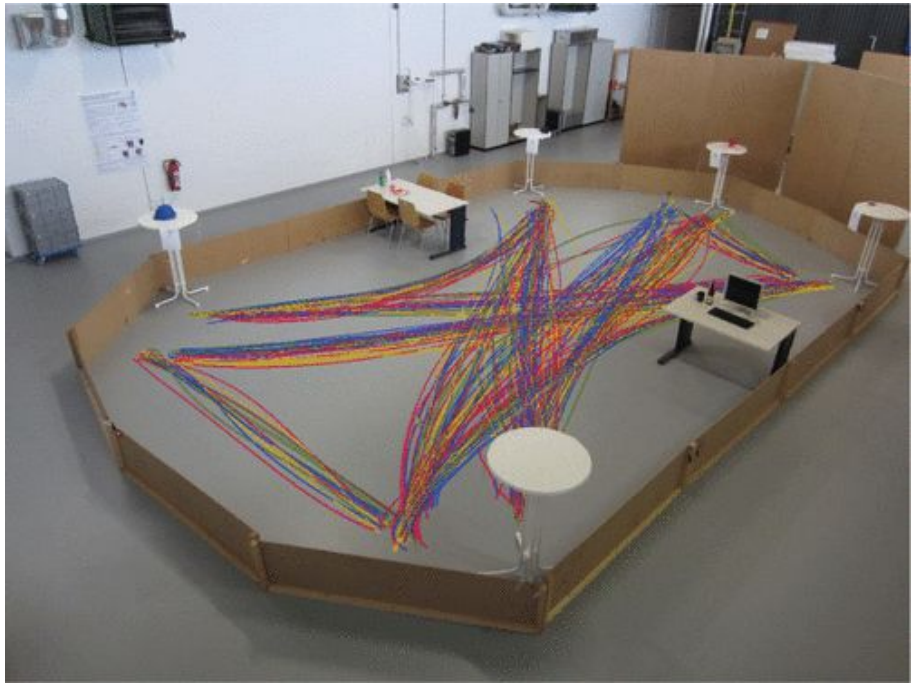


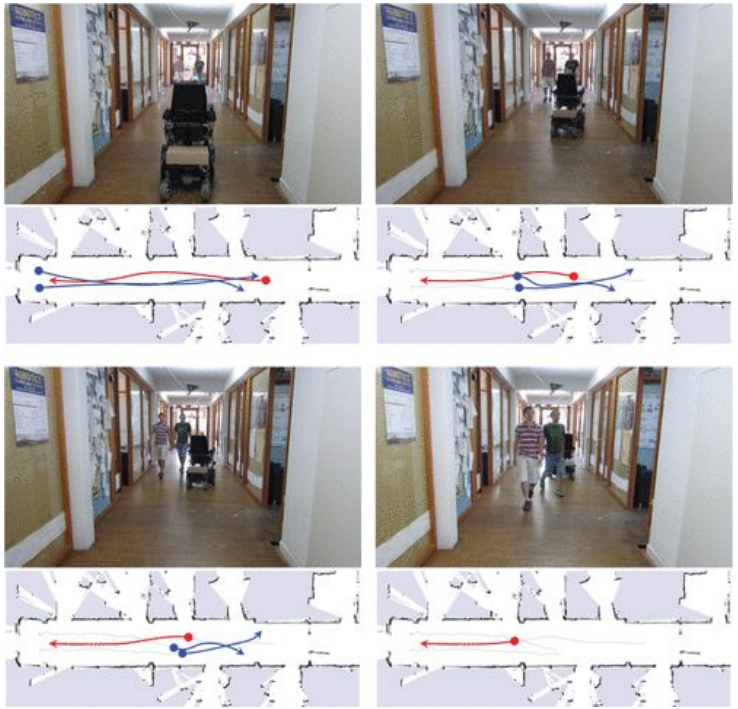


Ng and Russell, 2000

Example: Inverse Reinforcement Learning

H Kretzschmar **et al, IJRR 2016 [Socially compliant mobile robot navigation via inverse reinforcement learning]**

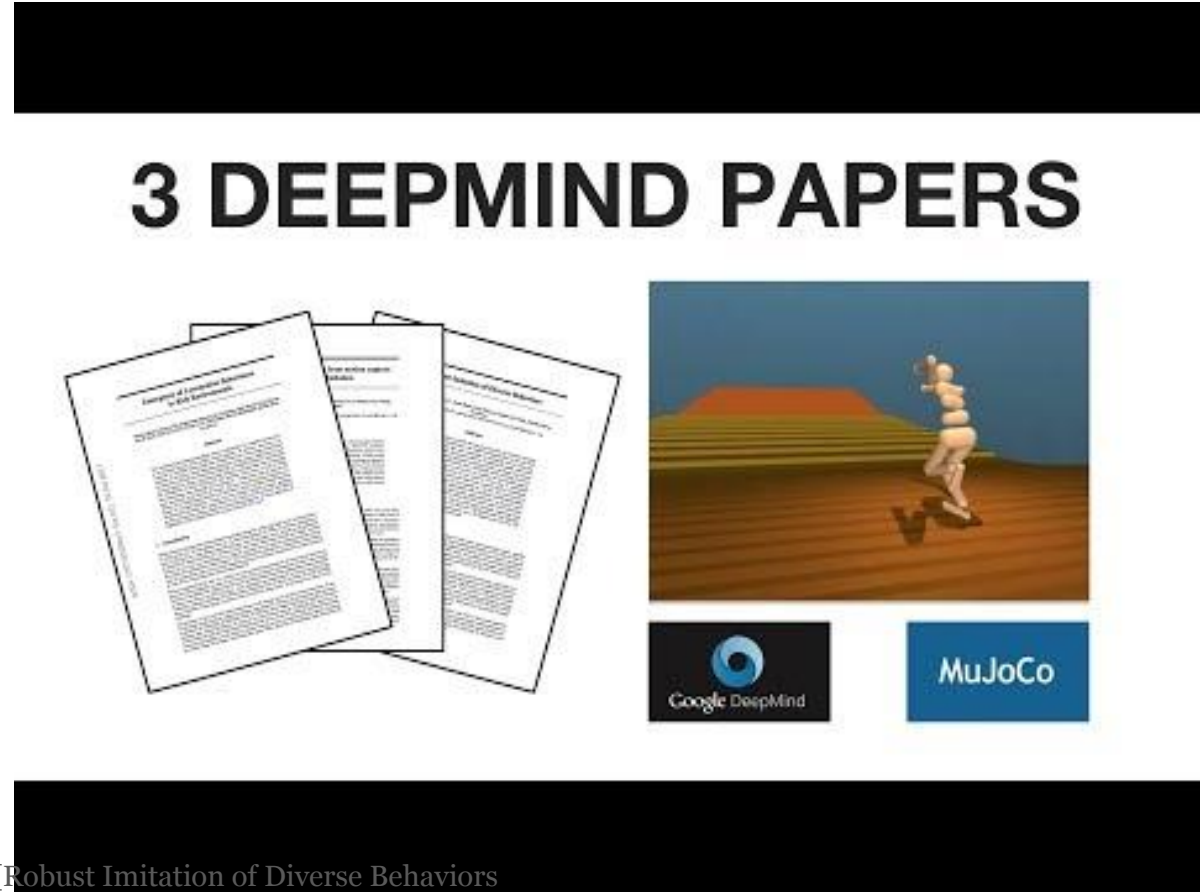




Problem with IRL

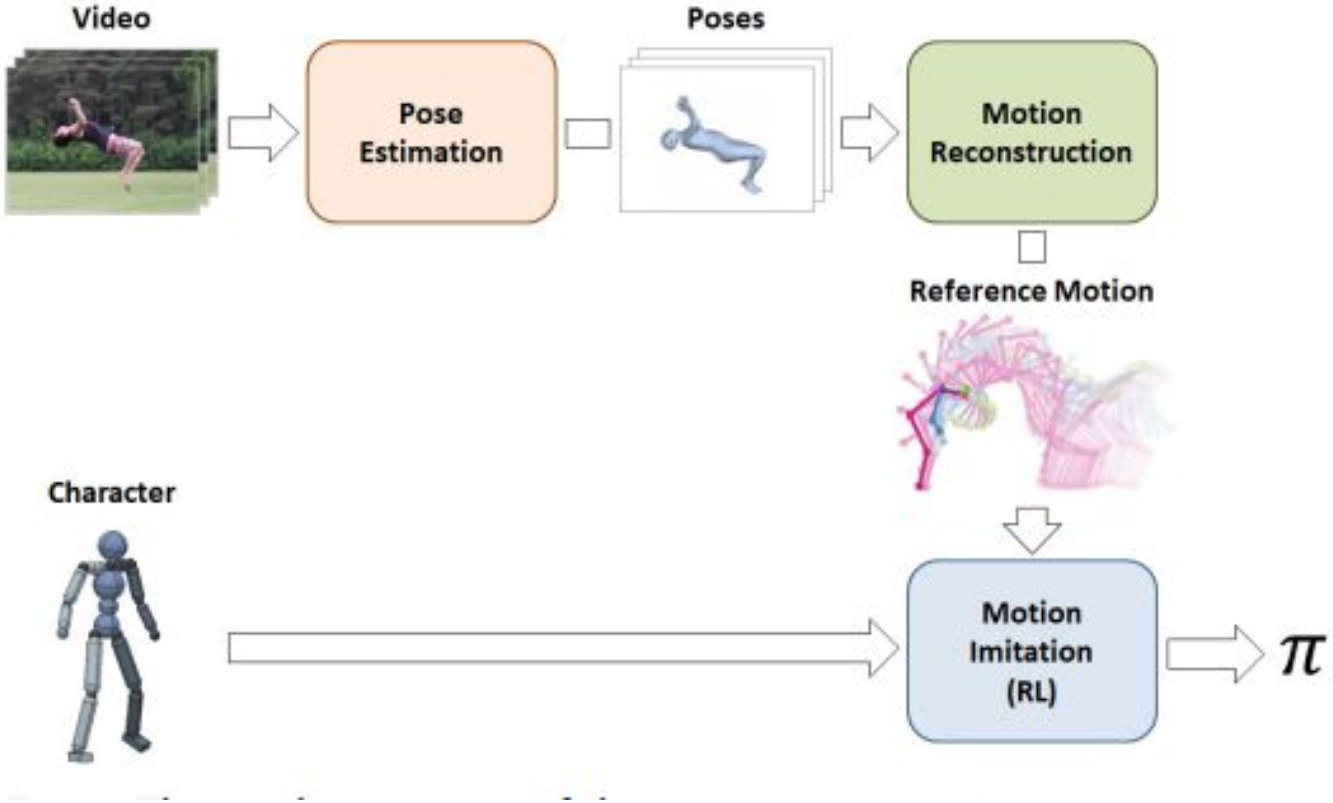
- For most observations of behavior there are many fitting reward functions. The set of solutions often contains many degenerate solutions, i.e. assigning zero reward to all states.
- The IRL algorithms assume that the observed behavior is optimal. This is a strong assumption, arguably too strong when we talk about human demonstrations.

Example: Adversarial Imitation (GAIL)



Wang et al, NIPS 2017 [Robust Imitation of Diverse Behaviors

Imitation with RL (rewards engineered from demos)



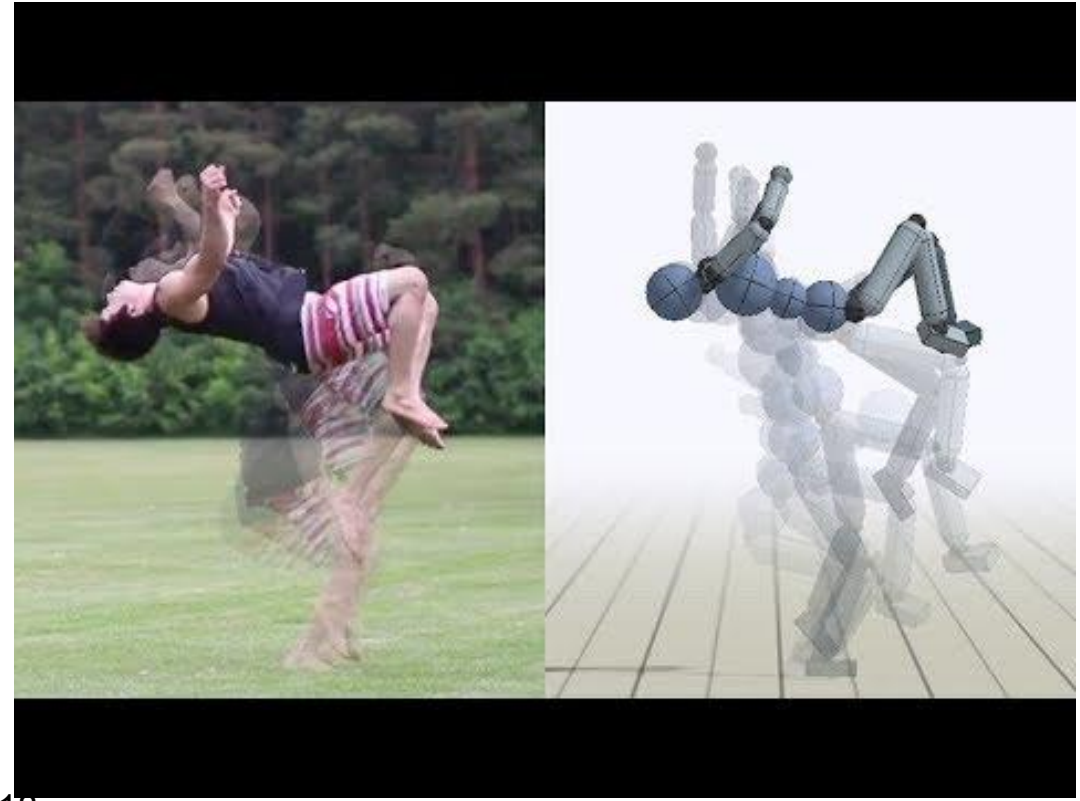
Reward terms:

- Pose matching
- Velocity matching
- End-effector matching
- Center-of-mass matching

Trained with PPO [Schulman et al 2015]

Peng et al, Siggraph ‘18
SFV: Reinforcement Learning of Physical Skills from Videos

Imitation with RL (rewards engineered from demos)



Peng et al, Siggraph ‘18
SFV: Reinforcement Learning of Physical Skills from Videos