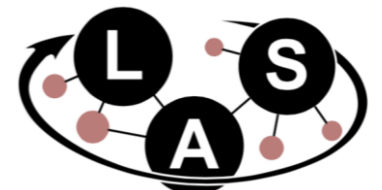


Safe Reinforcement Learning of Real World Processes

Felix Berkenkamp

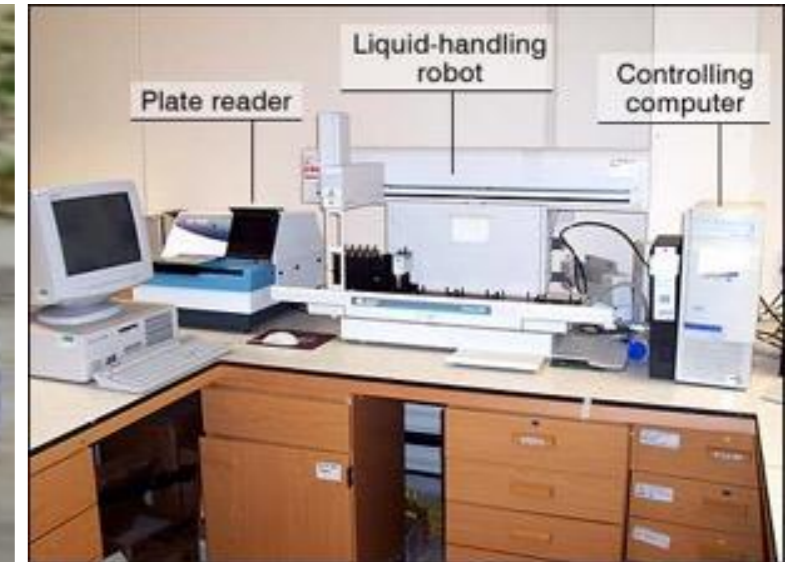
@Swissmem Workshop

ETH zürich



What is artificial intelligence?

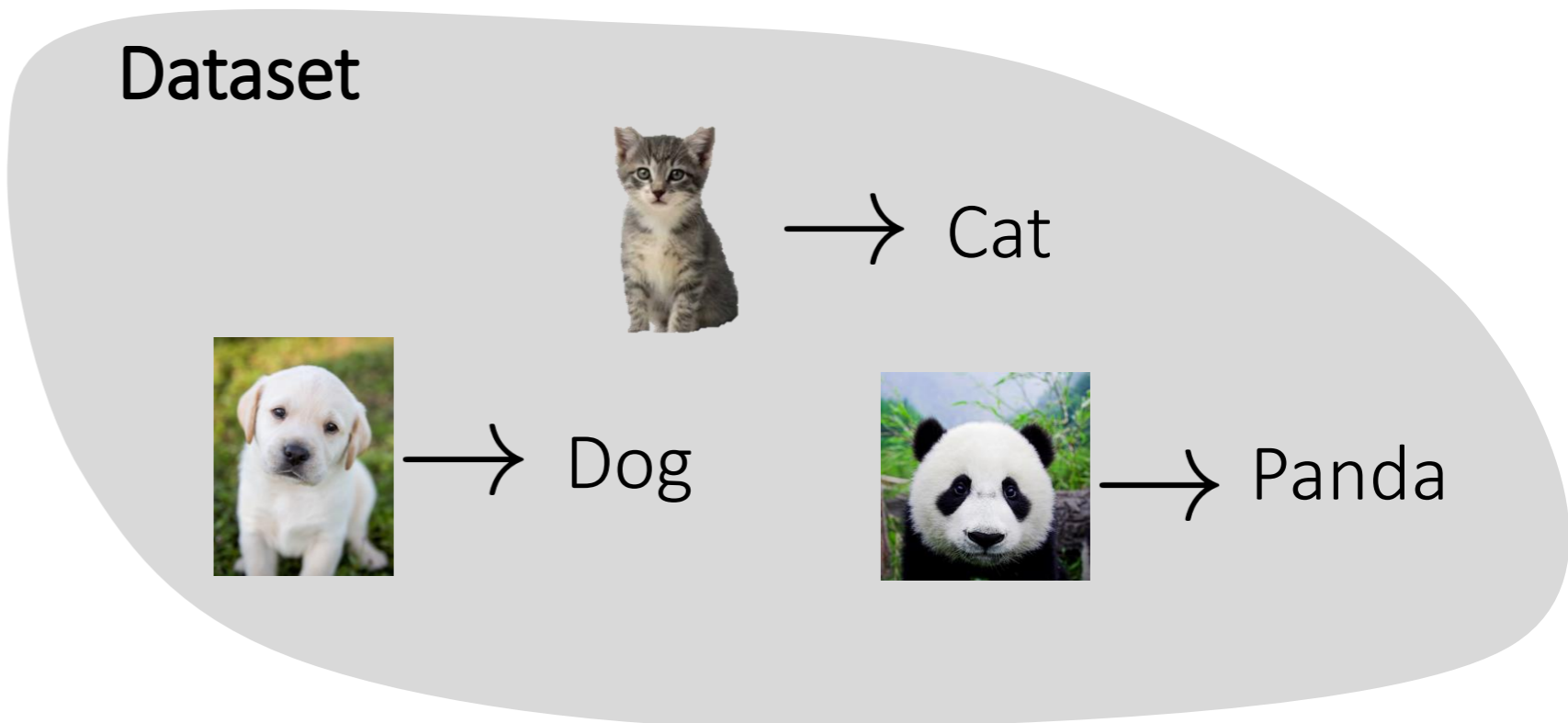
“Science and engineering of how to build **systems that solve tasks** commonly associated with requiring human-level intelligence.”



Recently, major breakthroughs through **machine learning!**

Supervised learning with neural networks

$$f : \text{Kitten} \rightarrow \text{Cat}$$

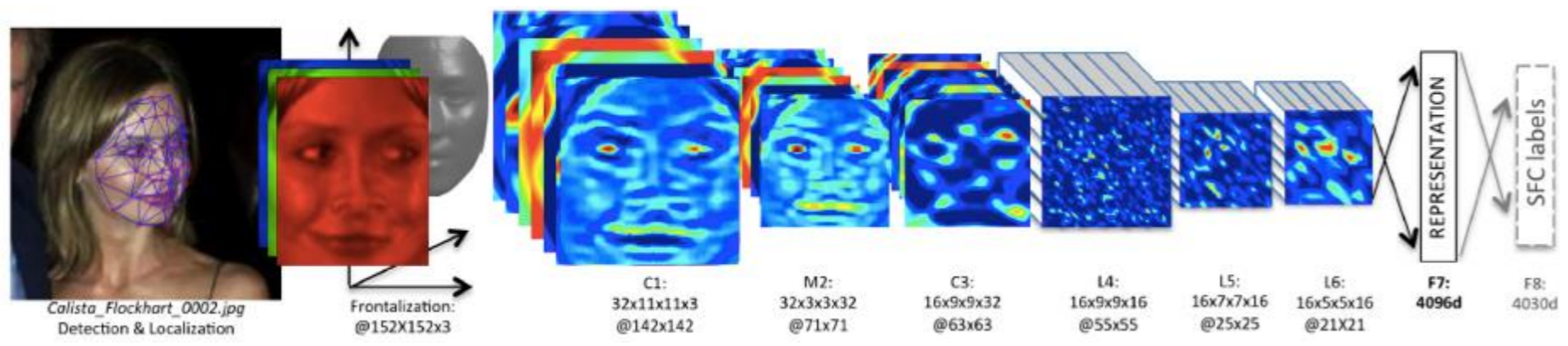


$$f(x; \mathbf{w}) = \phi(\mathbf{W}_1 \phi(\mathbf{W}_2 \phi(\dots \phi(\mathbf{W}_l \mathbf{x}))))$$

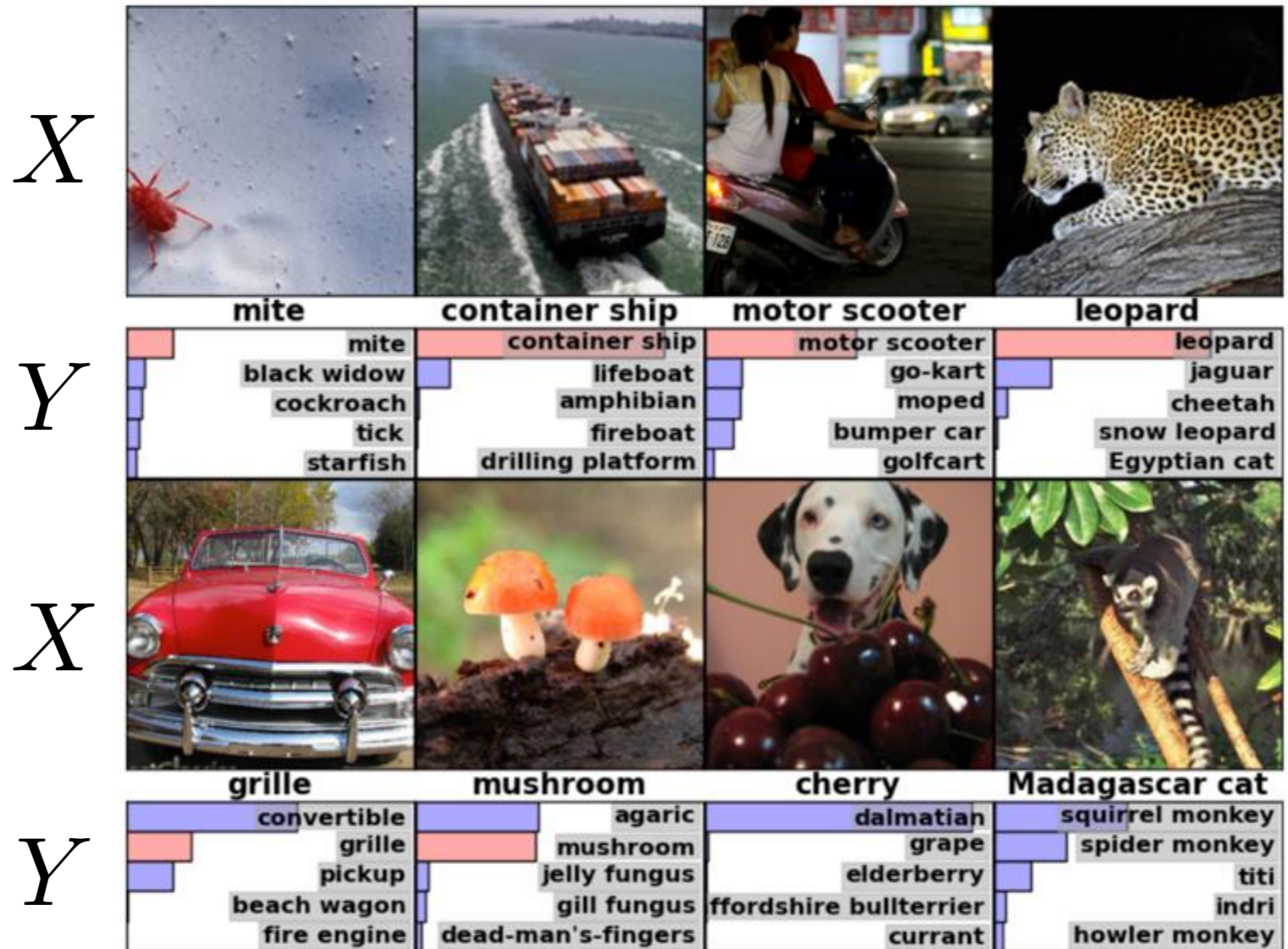
Flexible nonlinear functions with many parameters

Deep = nested in many layers

Loosely inspired by biological neuronal networks



Advances in deep learning



Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks '12

Advances in deep learning

Y

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A skateboarder does a trick on a ramp.



X

Y

A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



X

Vinyals et al. *Show and Tell: A Neural Image Caption Generator* '14

Advances in deep learning

The screenshot shows the Google Translate web interface. At the top left is the Google logo. Below it, the word "Translate" is written in red. On the right side, there is a link "Turn off instant translation" and a star icon. The main interface has two language selection boxes. The left box is set to "English" and "Detect language". The right box is set to "English", "Spanish", and "German". A blue "Translate" button is positioned between the two boxes. Below the language boxes, there are two text areas. The left text area contains the English text "Machine learning is getting more accurate" and is highlighted with a blue border. Below this text area are icons for a speaker and a keyboard, and the character count "41/5000". The right text area contains the German translation "Maschinelles Lernen wird immer genauer". Below this text area are icons for a star, a copy icon, a speaker, and a share icon, along with a pencil icon for editing.

X

Y

Challenges in modern machine learning

Data & Computation

- Privacy?
- Cost of data generation?
- Energy consumption?

Black box models

- Assurances / reliability?
- Bias / fairness
- Interpretability?



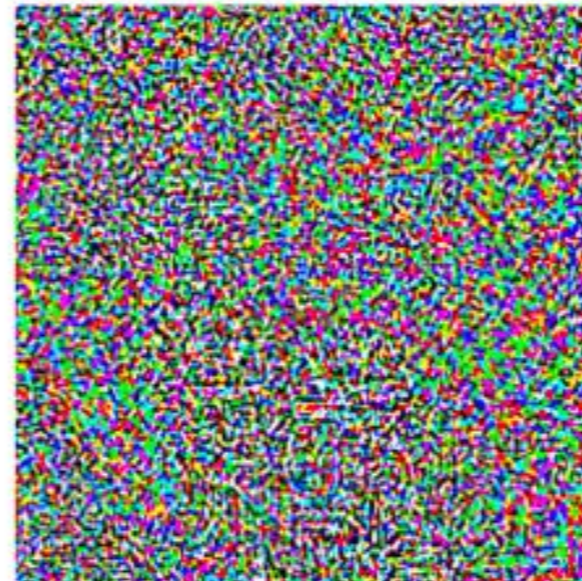
Trust

Reliability & Robustness



“panda”
57.7% confidence

+ .007 ×



“gibbon”
99.3% confidence

[Explaining and Harnessing Adversarial Examples, Goodfellow, Shlens & Szegedy ICLR '15]

Challenge: Fairness & Bias

The image is a collage of several news articles and social media posts related to machine bias and discrimination. At the top left, there's a photo of a DNA helix and a person's hand. Below it, a headline reads "Big Bad Data May Be Triggering Discrimination". To the right, a large black box contains the title "Machine Bias" and a sub-headline: "There's software used across the country to predict future criminals. And it's biased against blacks." Below this, there are three main article snippets: 1) "Artificial intelligence: How to avoid racist algorithms" by Zoe Kleinman, dated 14 April 2017, with social media share icons. 2) "Robot racism: AI beauty judges preferred white contestants over those with dark skin" by World Tribune, dated September 9, 2016, with a "tech" category tag. 3) "Math is racist: How data is driven" by Aimee Rawlins, dated September 6, 2016, with a Twitter handle @aimeerawlins. At the bottom left, there's a grid of baby faces from a dataset.

[Joseph, Kearns, Morgenstern '16] [Dwork, Hardt, Pitassi, Reingold, Zemel '11] [Heidari, Krause '18]

Challenge: Fairness & Bias

Bias in → Bias out

Even ignoring protected attribute, discrimination can occur due to correlation

Key issues

- What is fairness?
- How to enforce that trained ML models are fair?



Hoda Heidari

A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Michele Loi
University of Zürich
michele.loi@uzh.ch

Andreas Krause
ETH Zürich
krausea@ethz.ch

Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Claudio Ferrari
ETH Zürich
ferrari@ethz.ch

Andreas Krause
ETH Zürich
krausea@ethz.ch

Artificial intelligence for decision making



Decisions are happening in closed loop

Decisions have long-term consequences

Trust even more important!

Reinforcement Learning



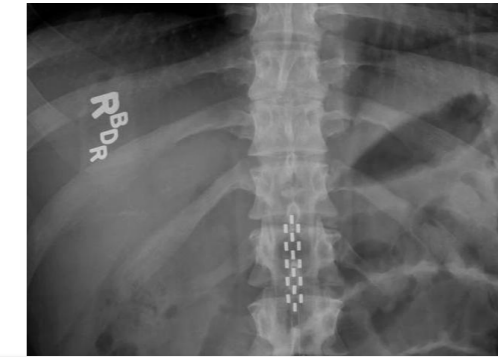
Actively generate dataset

Need to trade off **exploration** & **exploitation**

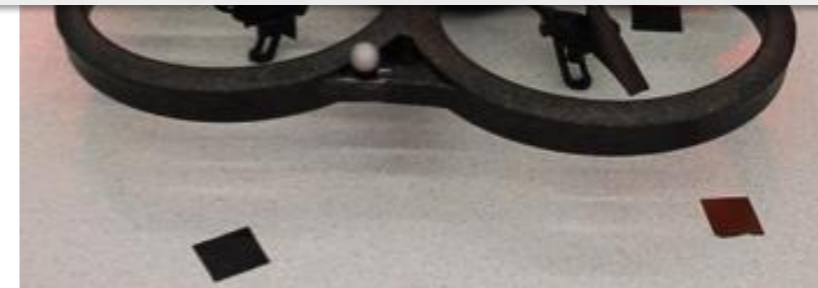
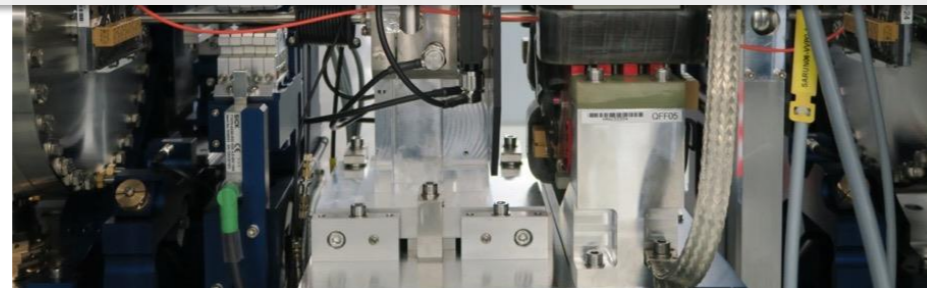
Reinforcement Learning: An Introduction

R. Sutton, A.G. Barto, 1998



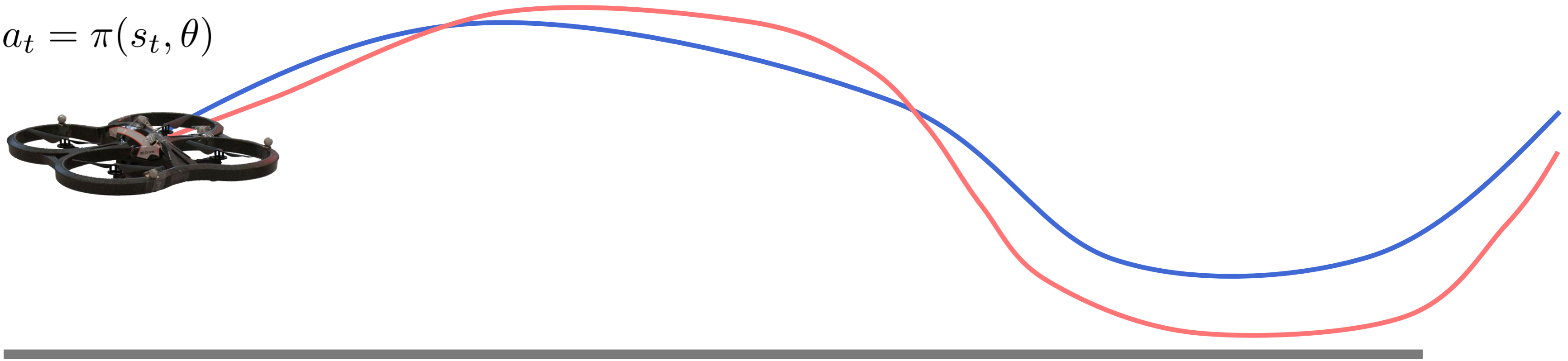


How can we **learn** to act
safely in unknown environments?



Model-free reinforcement learning

$$a_t = \pi(s_t, \theta)$$



Tracking performance

$$\max_{\theta} J(\theta)$$

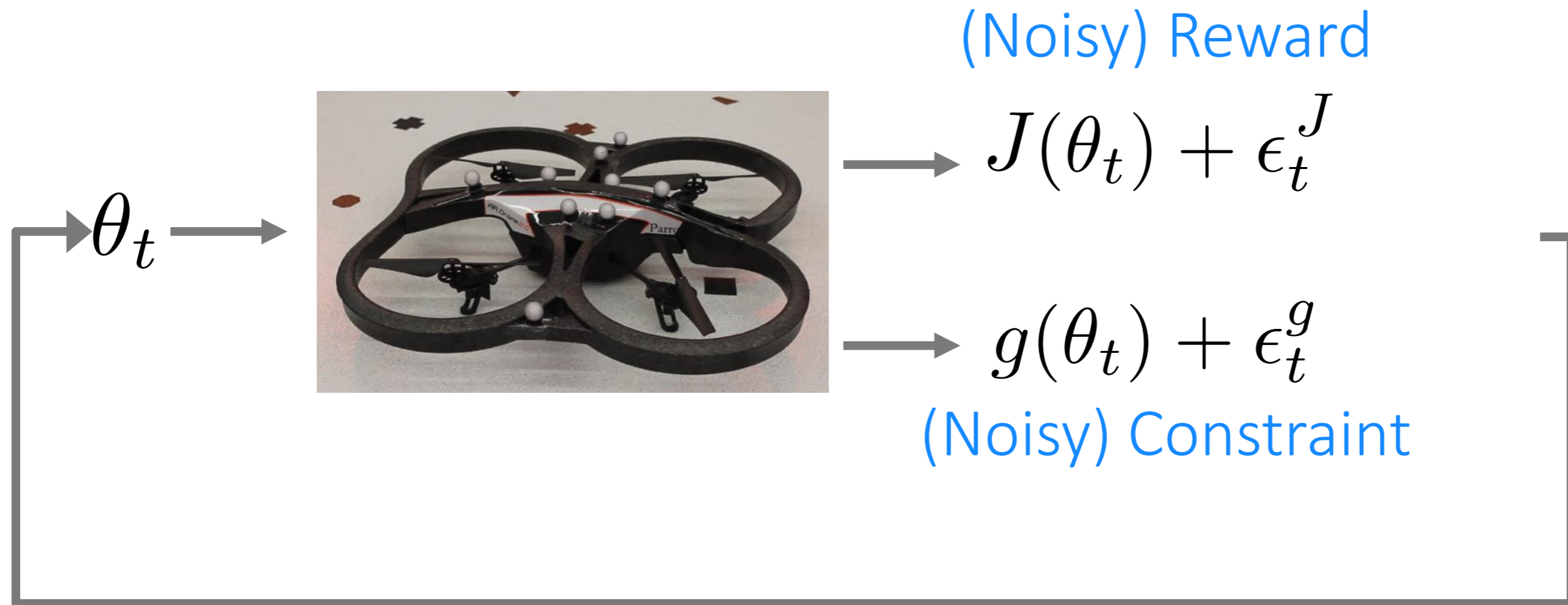
Few, noisy experiments

Safety constraint

$$g(\theta) \geq 0$$

Safety for all experiments

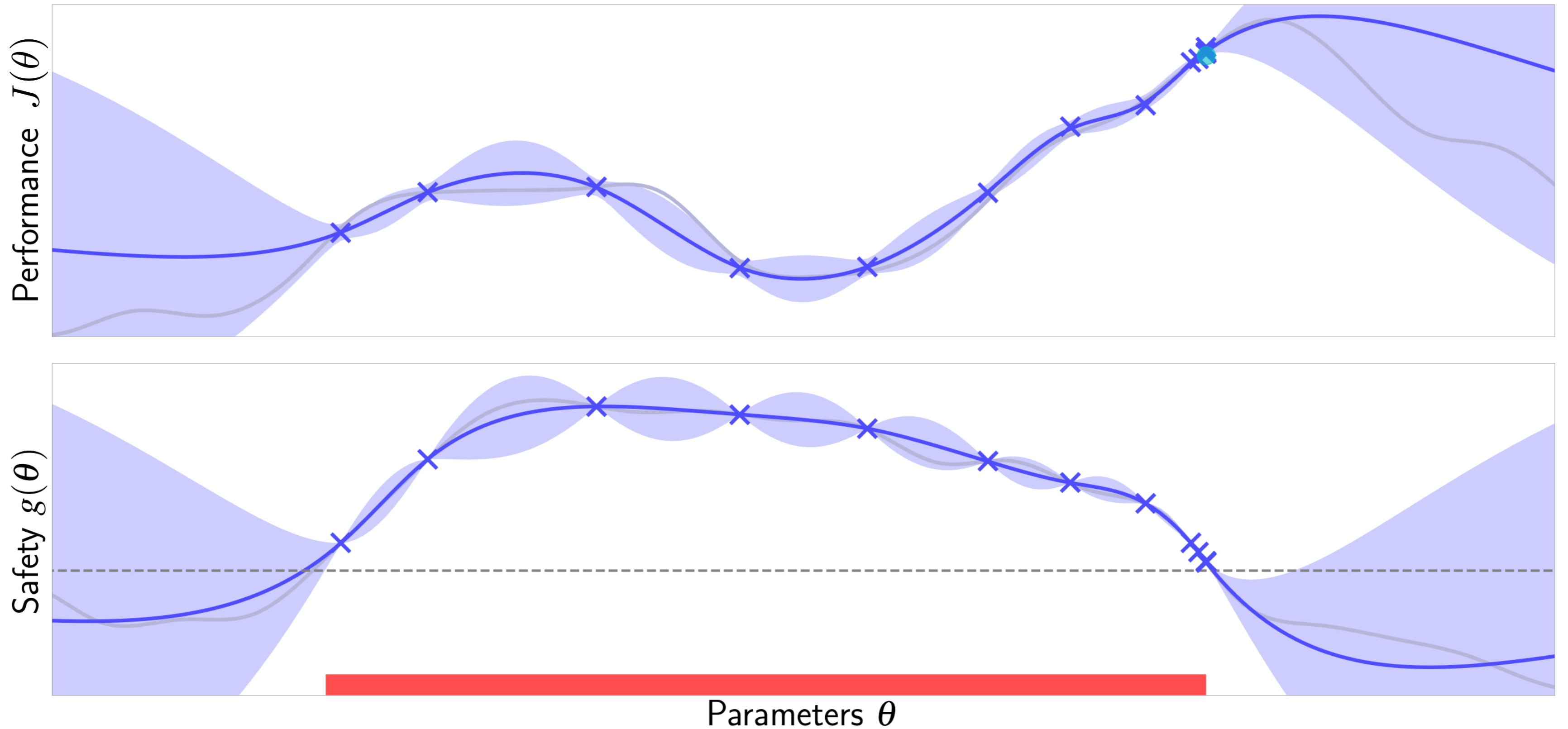
Safe policy optimization



Goal: $\max_{\theta} J(\theta) \text{ s.t. } g(\theta) \geq 0$

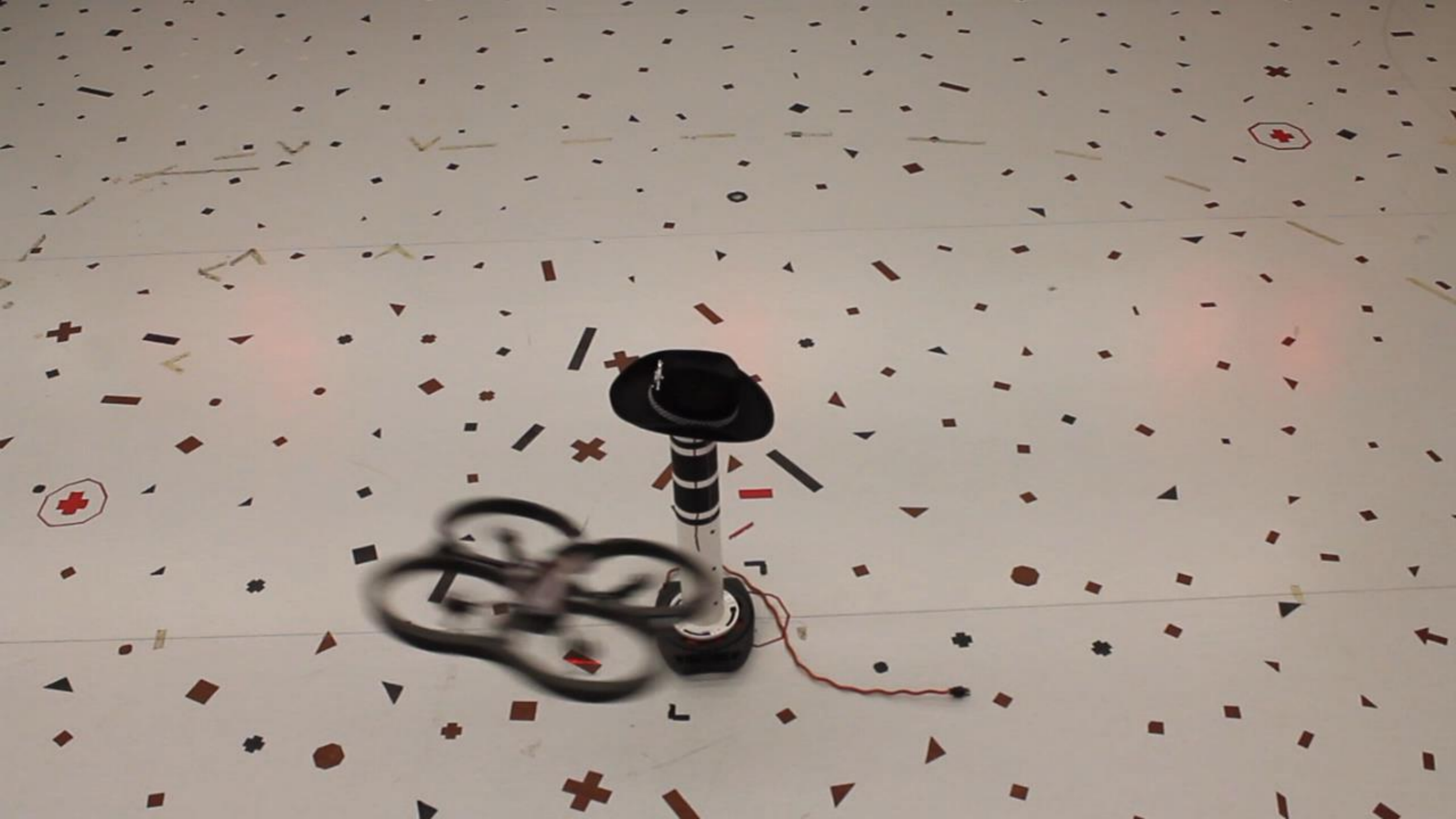
Safety: $g(\theta_t) \geq 0$ for all t with probability $\geq 1-\delta$

SafeOPT: Constrained Bayesian optimization

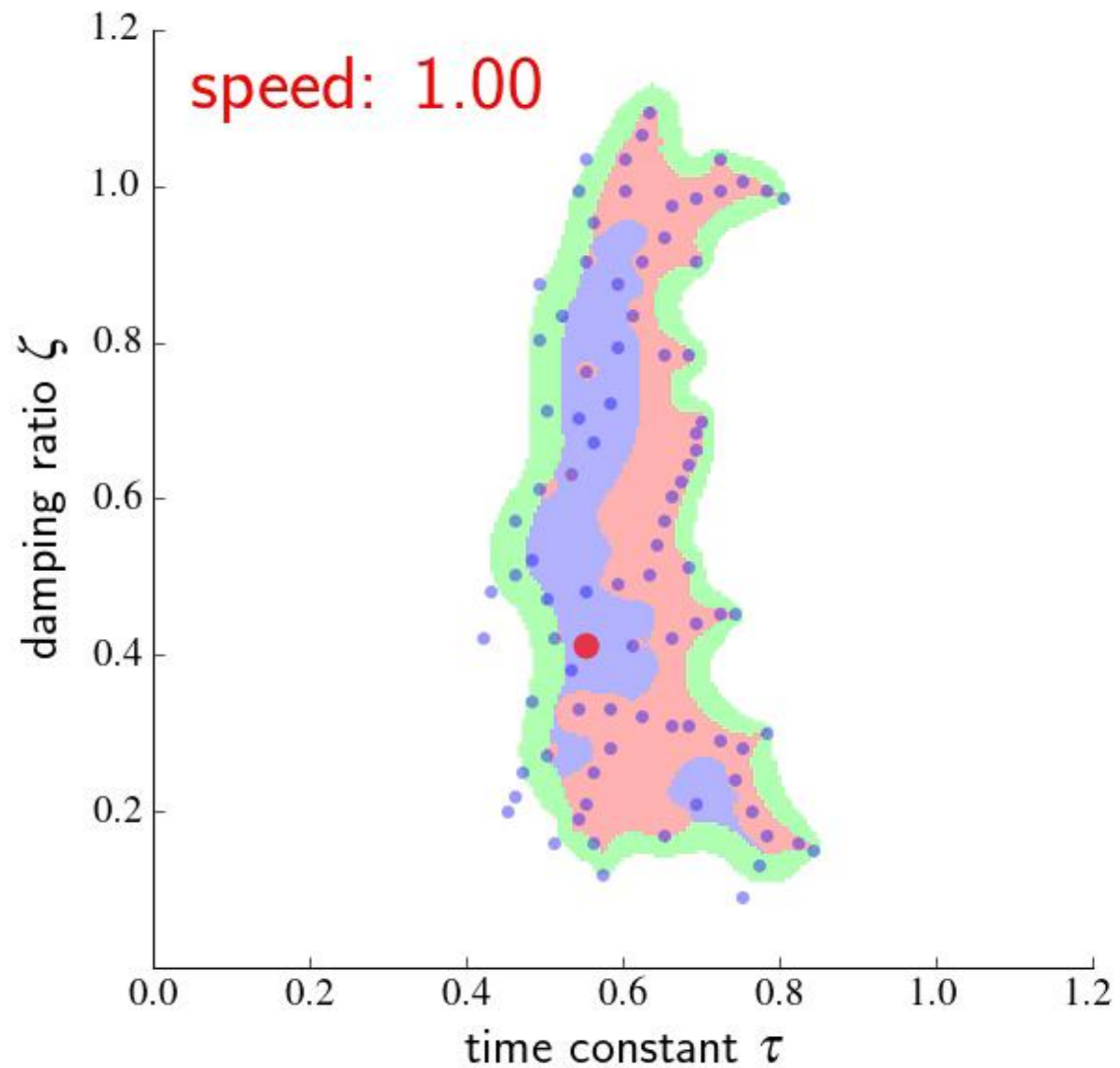
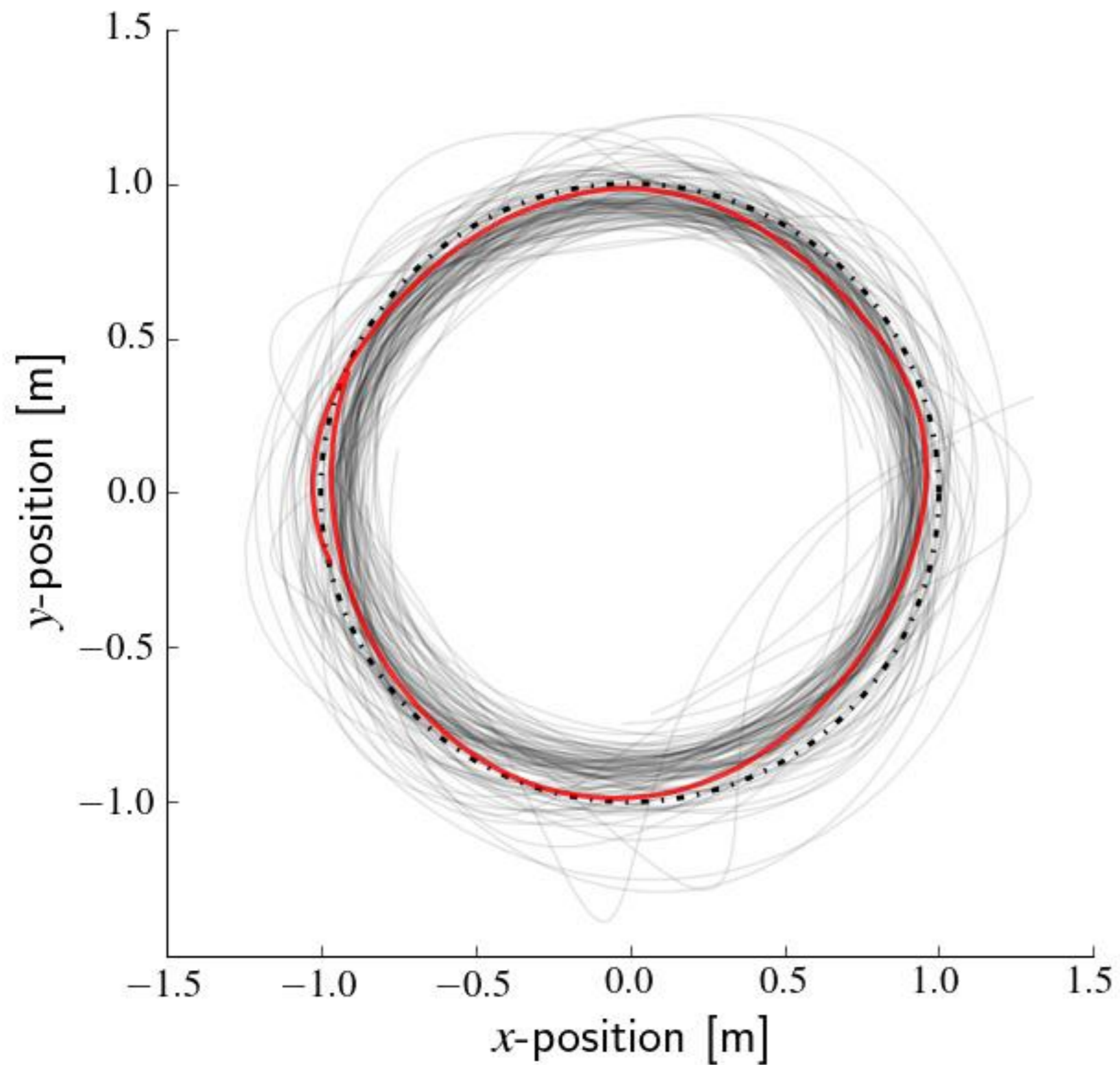


[Sui, Gotovos, Burdick, Krause ICML'15], [Berkenkamp, Schoellig, Krause '16]

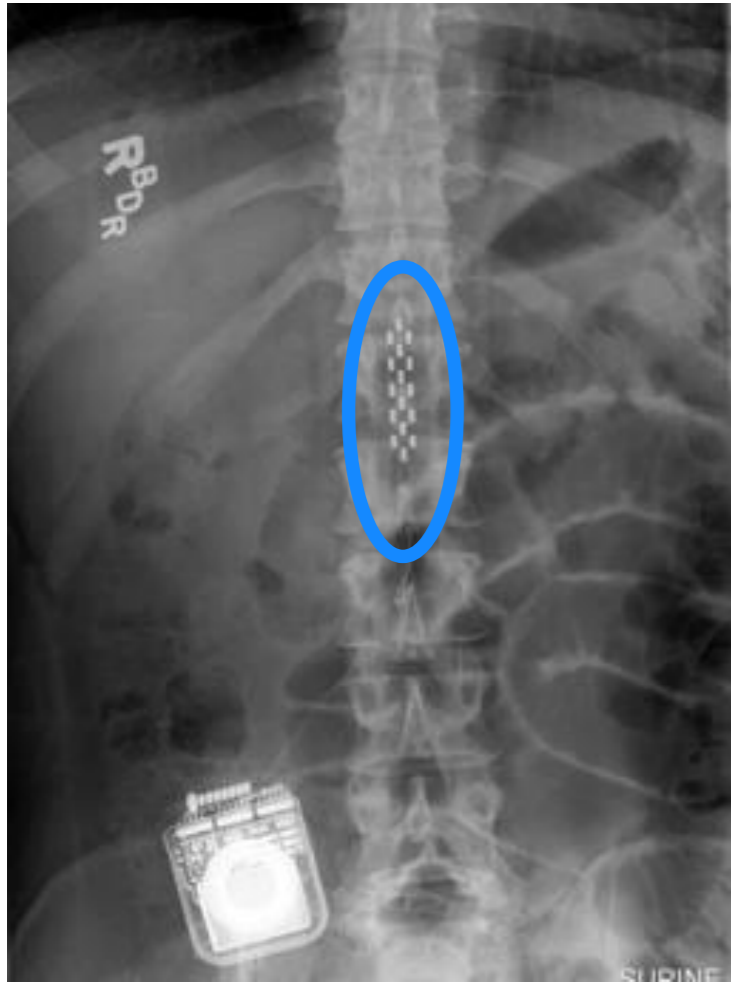




TRANSFER TO HIGHER SPEED (CONTEXT)



Other applications



S. Harkema, The Lancet, Elsevier

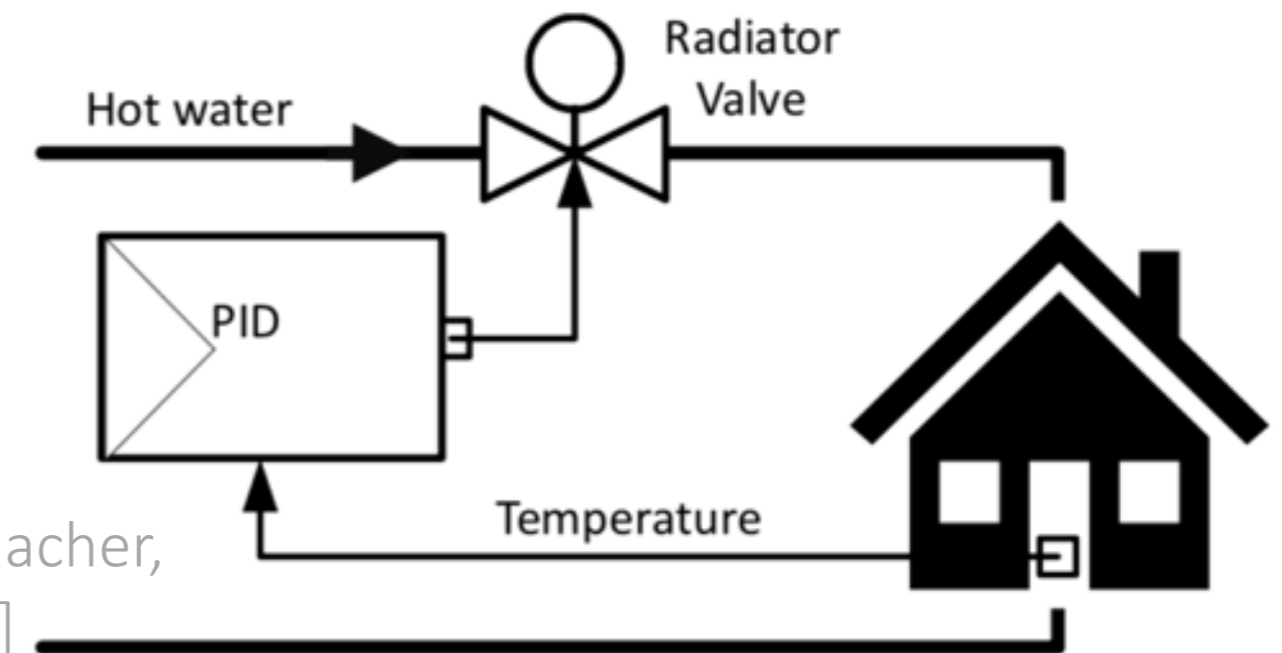
[Sui, Zhuang, Burdick, Yue, ICML 2018]



[ECC'16 Best App. Paper]



[ICML 2019]



[Fiducioso, Curi, Schumacher, Gwerder, K, IJCAI 2019]

Summary

Rapid progress in the field of machine learning

Huge opportunities across many domains

Major impact for science, industry and society

Very far from “General AI”

Key Challenge:

Building learning **systems that one can trust**

Reliability, Robustness, Interpretability, Fairness, etc.