Safe Reinforcement Learning of Real World Processes

Felix Berkenkamp

@Swissmem Workshop

ETH zürich

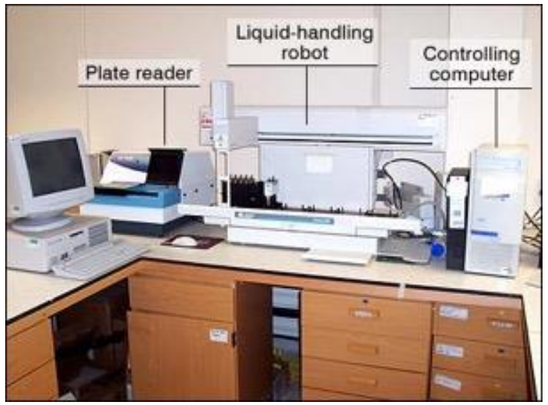
Open Philanthropy Project

LAS

What is artificial intelligence?

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

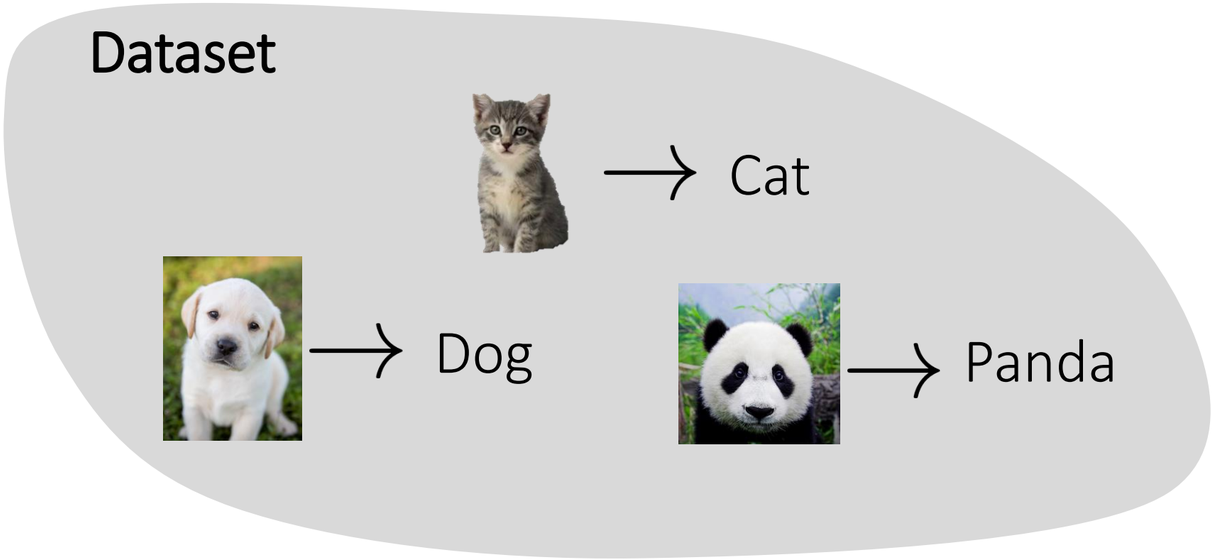




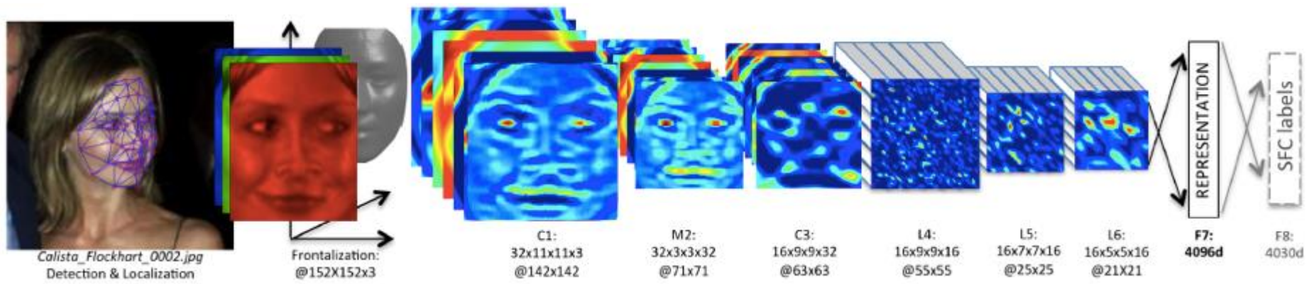
Recently, major breakthroughs through **machine learning**!

Supervised learning with neural networks

$f:$  $\longrightarrow$ Cat



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

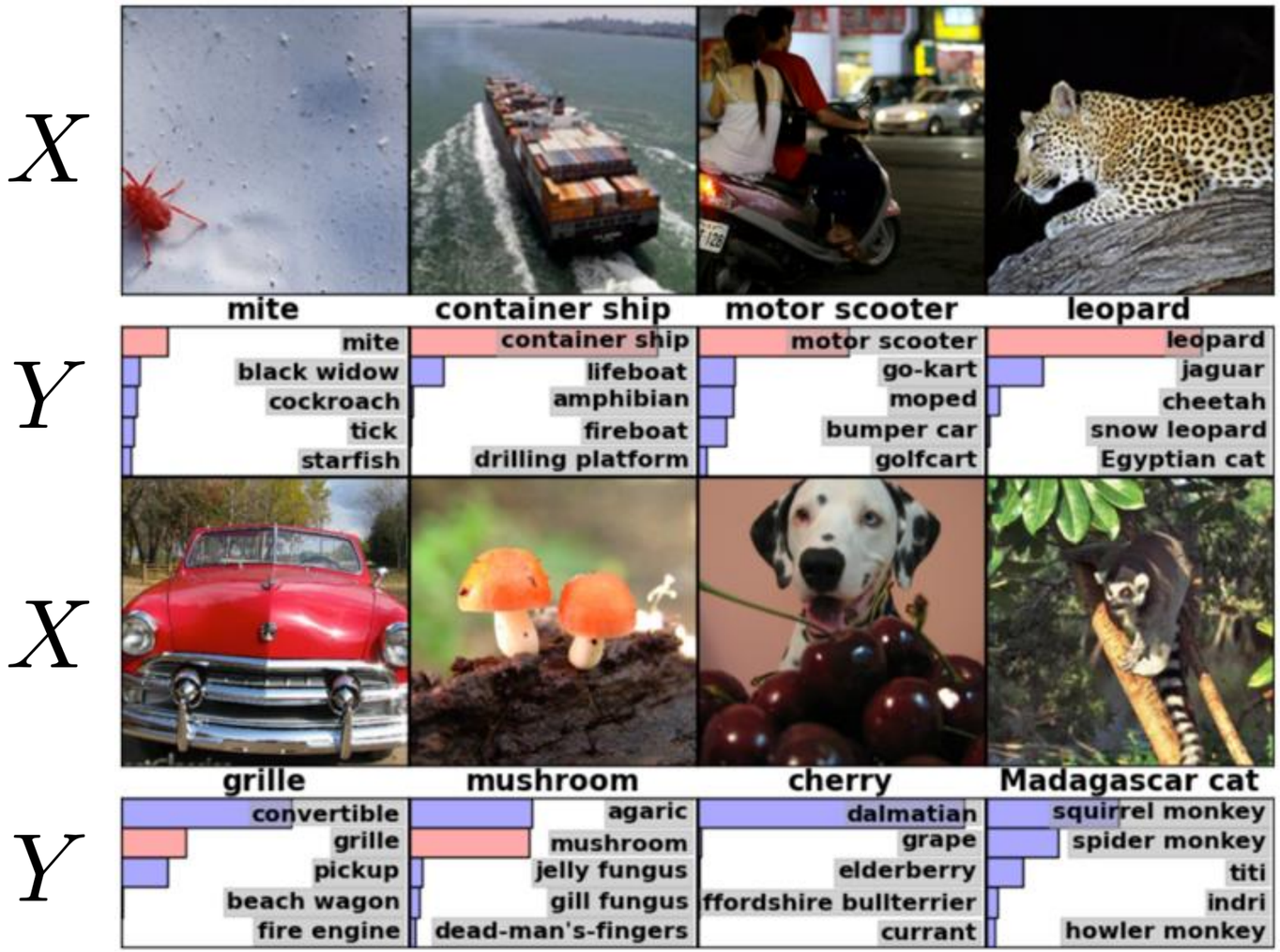


Flexible nonlinear functions with many parameters

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

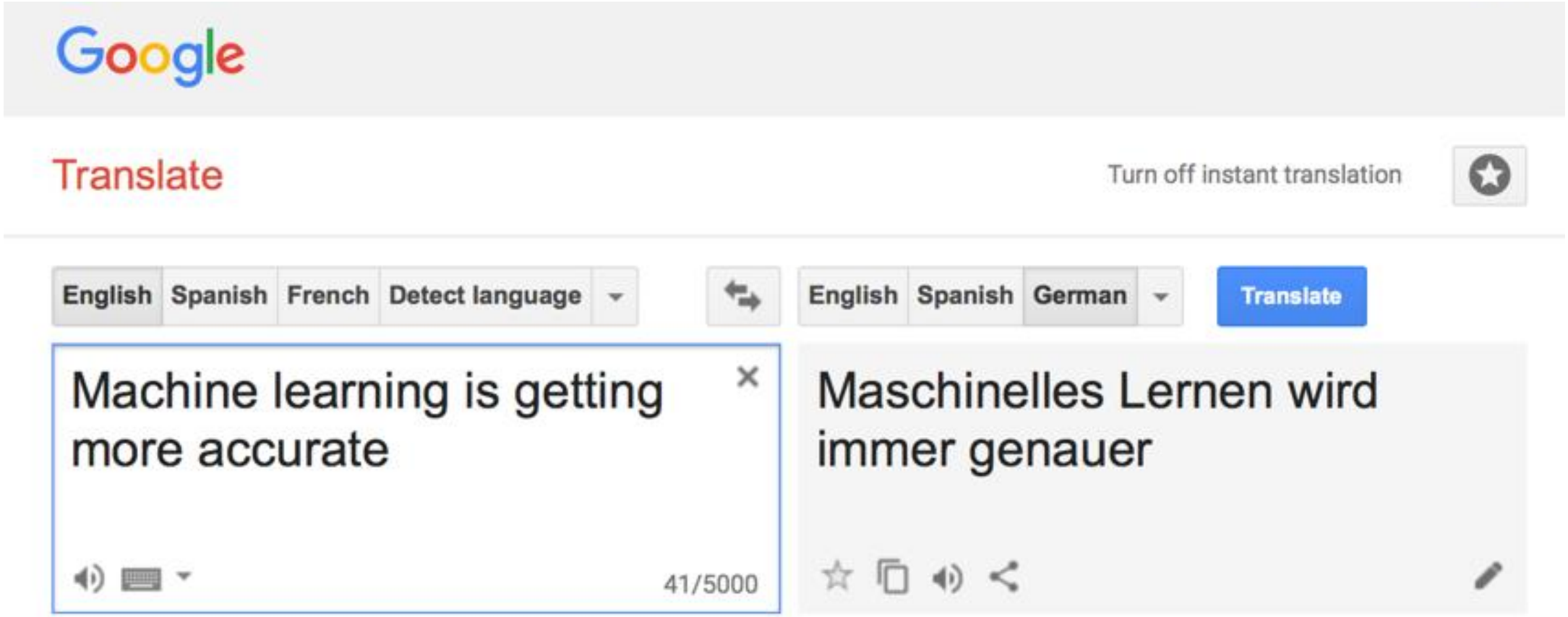
X

Y

X

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

Advances in deep learning



X Y

Challenges in modern machine learning

Data & Computation

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

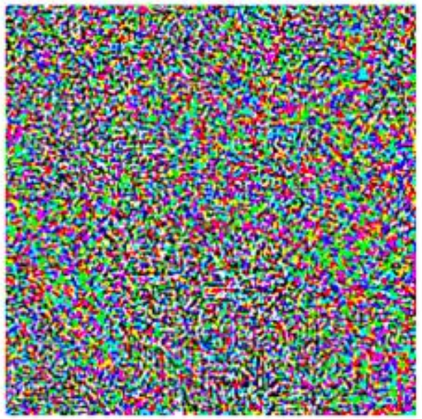
Black box models

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

Trust

Reliability & Robustness

$+ .007 \times$

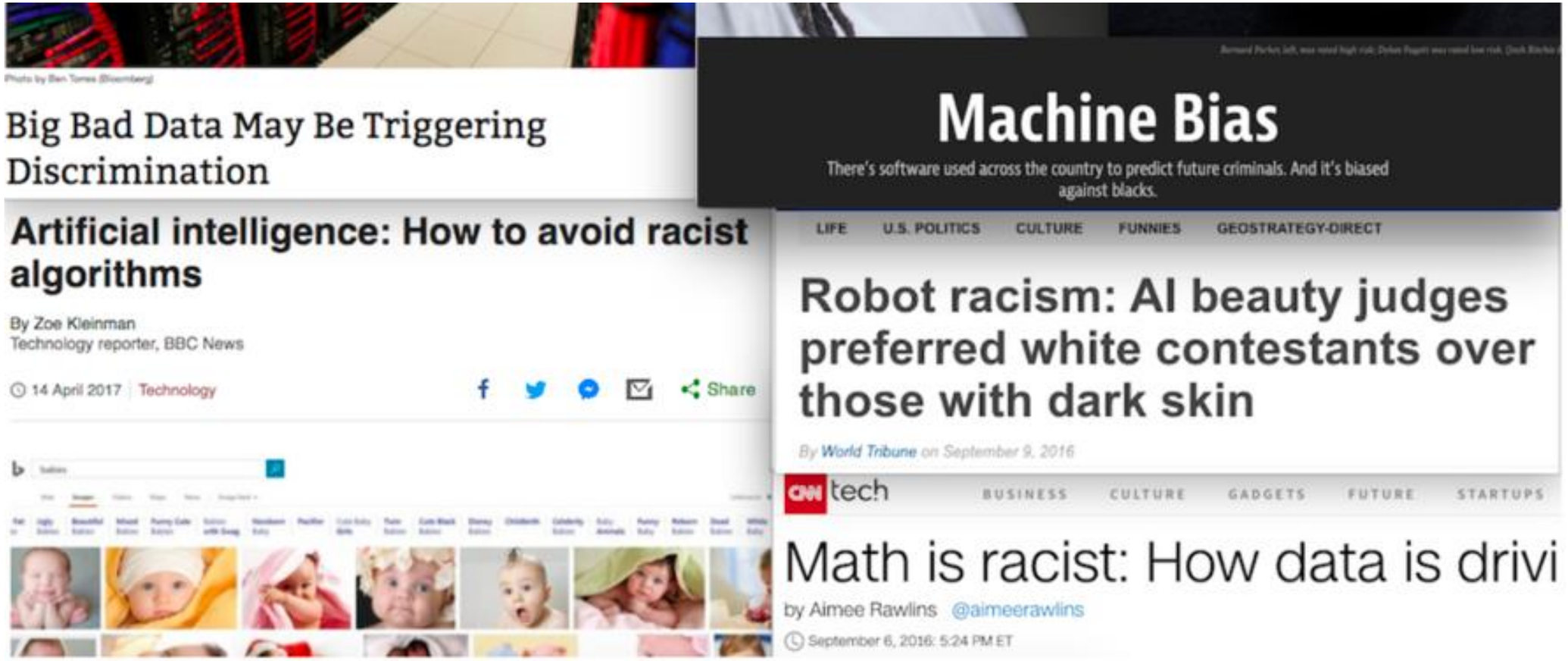


“panda”
57.7% confidence

“gibbon”
99.3% confidence

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

Challenge: Fairness & Bias



[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

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

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

Andreas Krause
ETH Zürich
krausea@ethz.ch

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

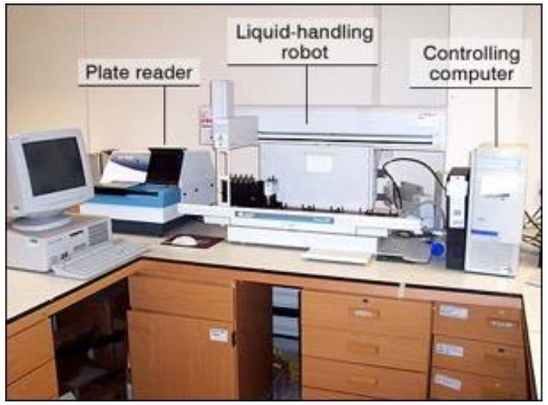
Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Claudio Ferrari
ETH Zürich
ferraric@ethz.ch

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

Andreas Krause
ETH Zürich
krausea@ethz.ch

Artificial intelligence for decision making



Decisions are happening in closed loop

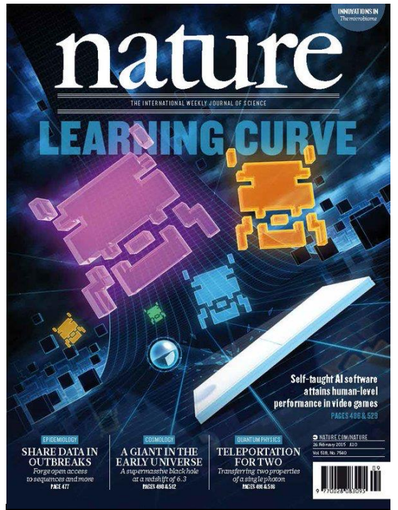
Decisions have long-term consequences

Trust even more important!

Reinforcement Learning

Actively generate dataset

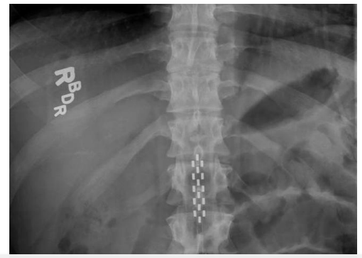
Need to trade off exploration & exploitation



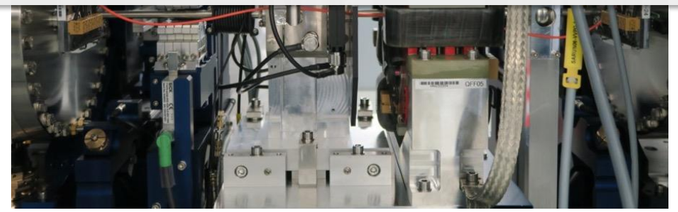
Reinforcement Learning: An Introduction
R. Sutton, A.G. Barto, 1998

Towards safe reinforcement learning





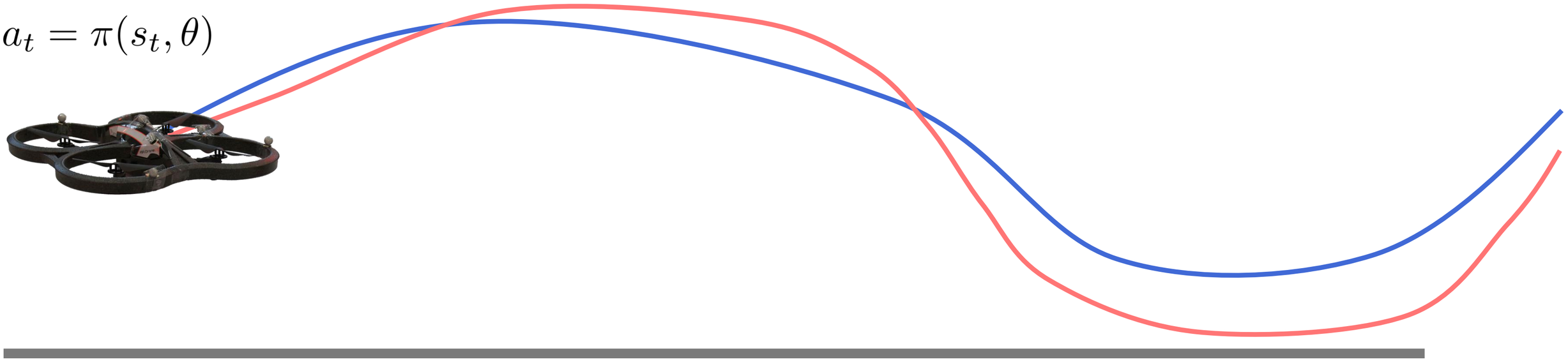
How can we **learn** to act
safely in unknown environments?




...

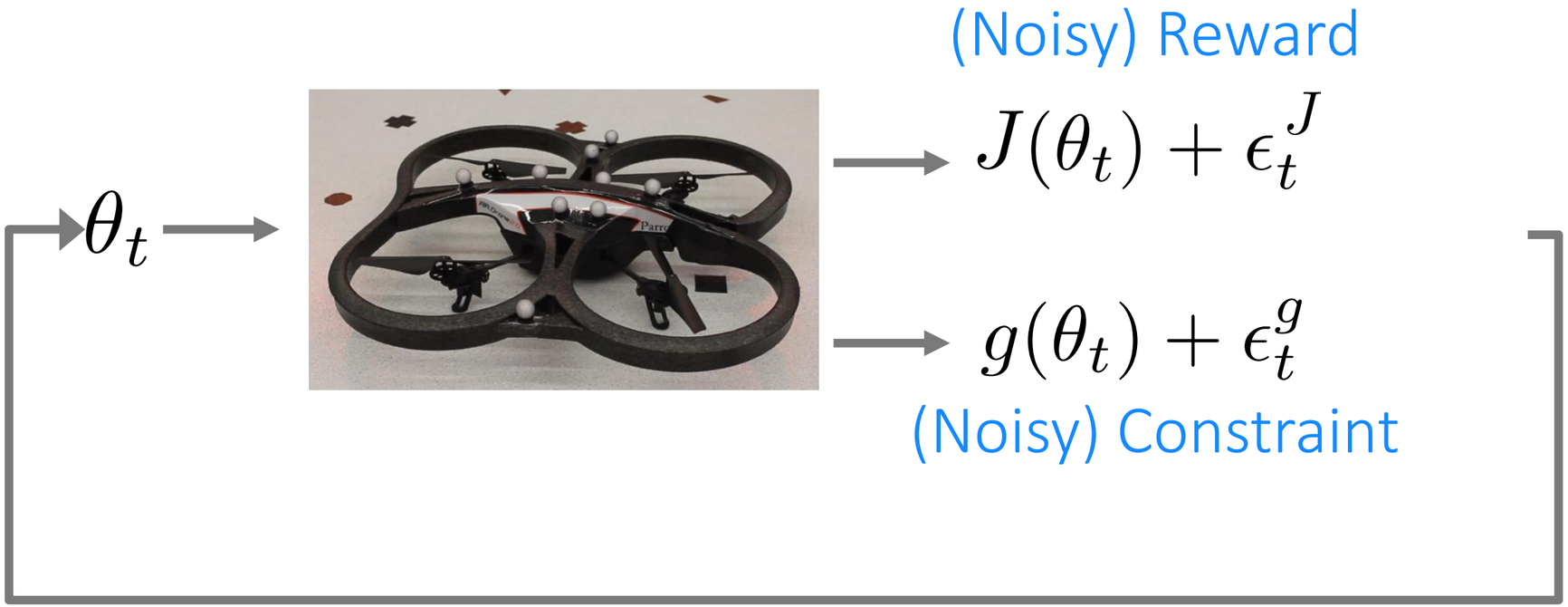
Model-free reinforcement learning

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



Tracking performance	$\max_\theta J(\theta)$	**Few, noisy experiments**
Safety constraint	$g(\theta) \geq 0$	**Safety for all experiments**

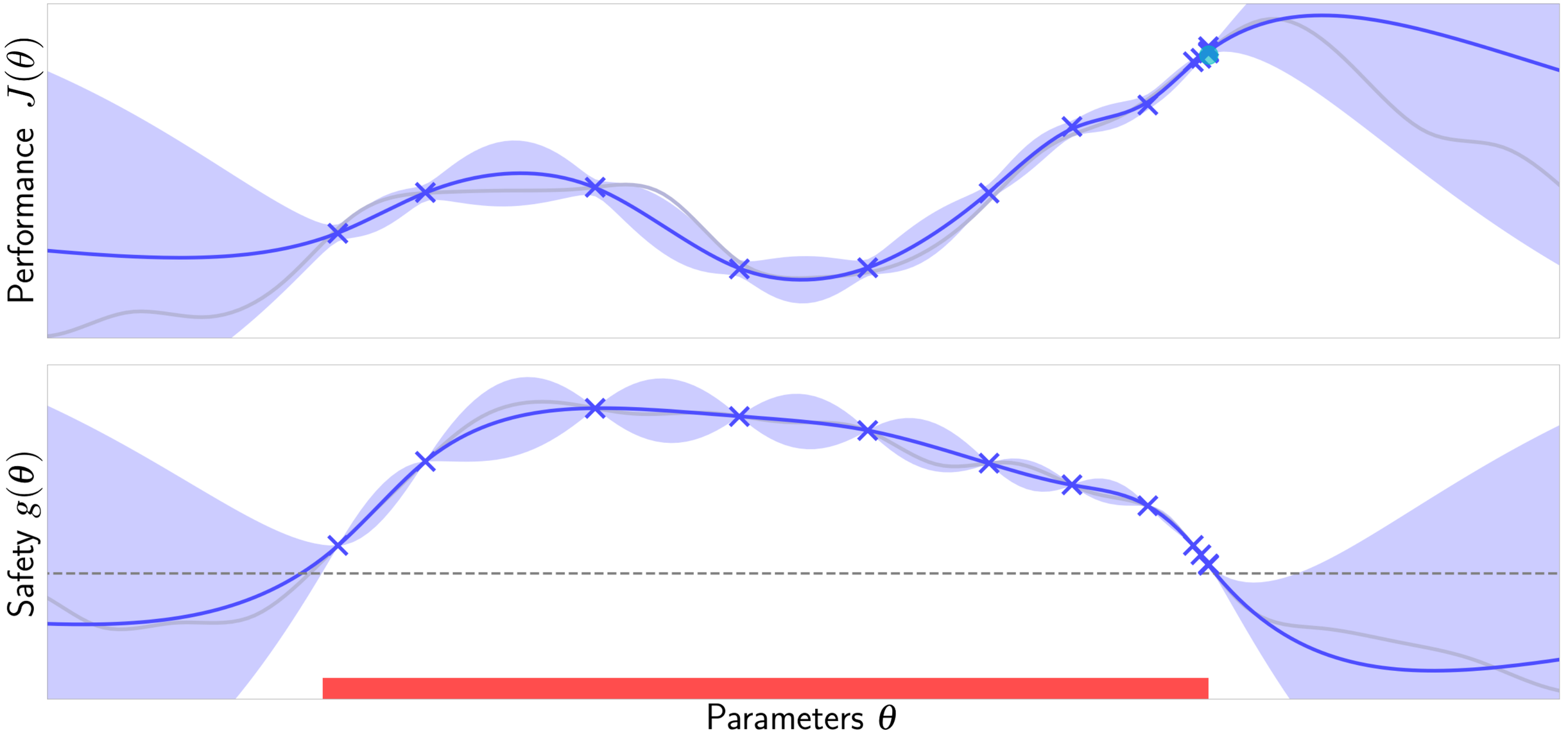
Safe policy optimization



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

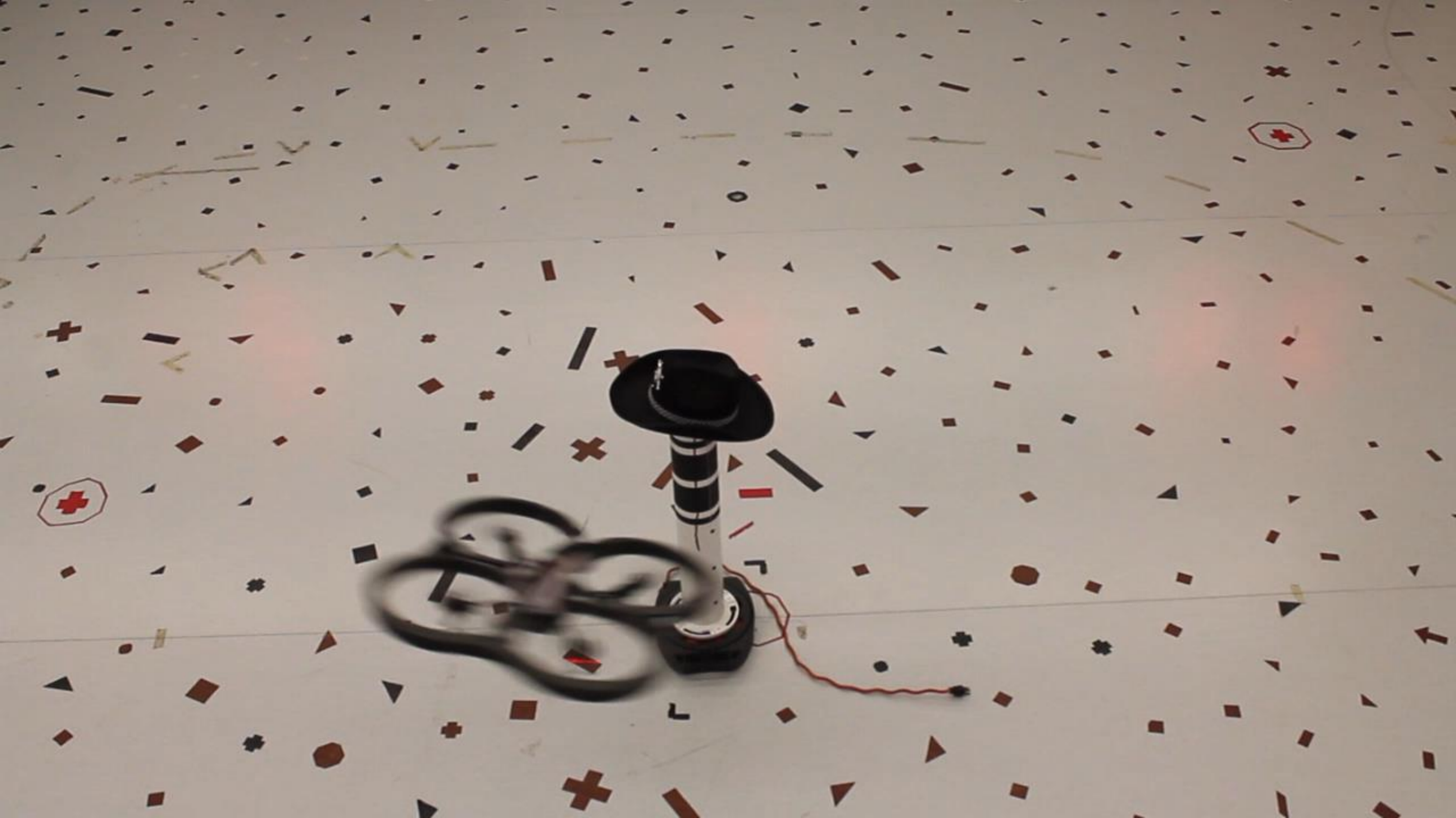
Safety: $g(\theta_t) \geq 0 \text{ for all } t$ with probability ≥ 1-δ

SafeOPT: Constrained Bayesian optimization

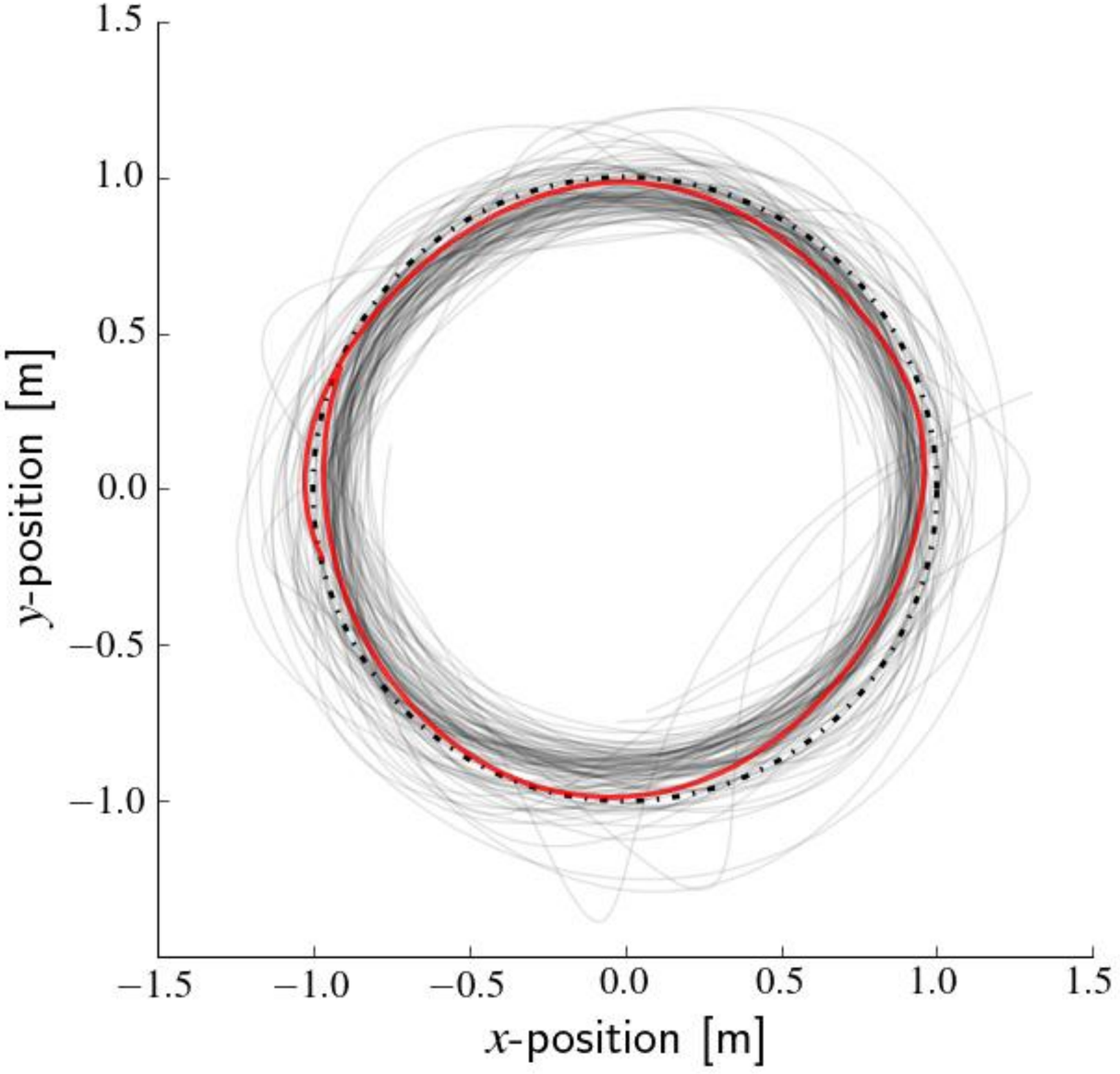


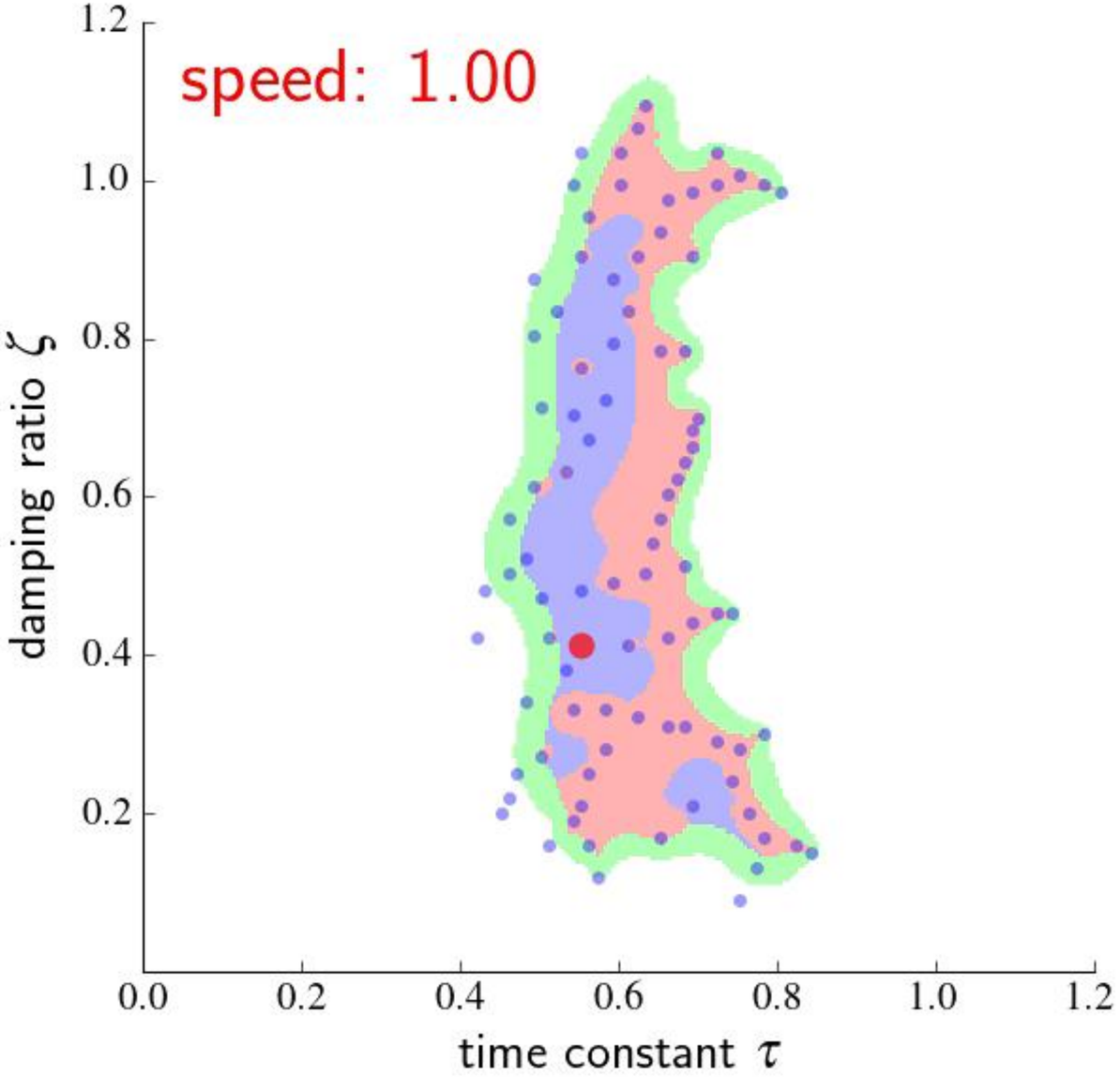
[Sui, Gotovos, Burdick, Krause ICML'15], [Berkenkamp, Schoellig, Krause '16]


AR.Drone 2.0

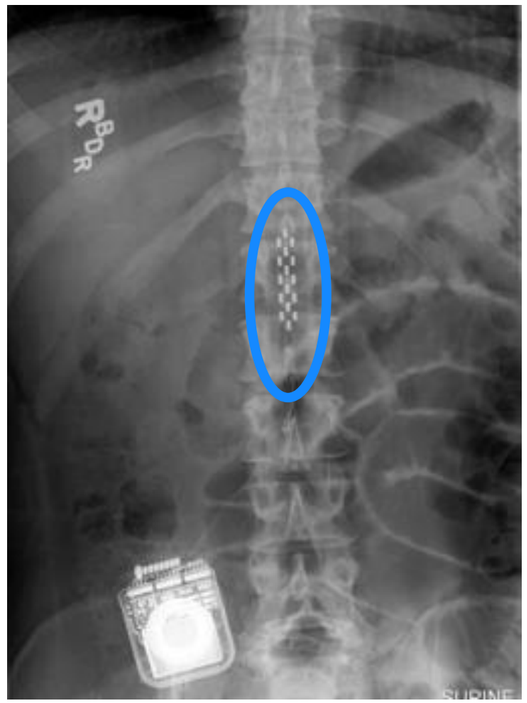


TRANSFER TO HIGHER SPEED (CONTEXT)





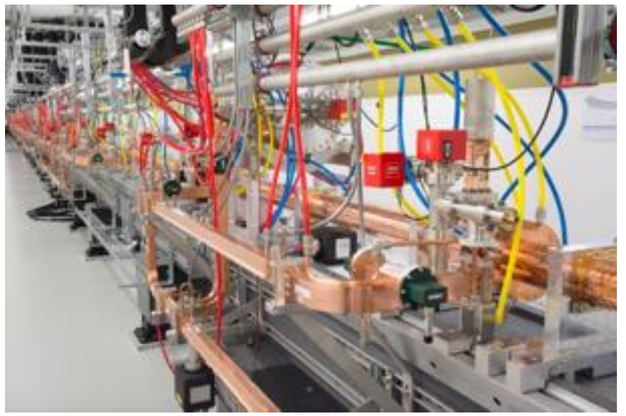
Other applications



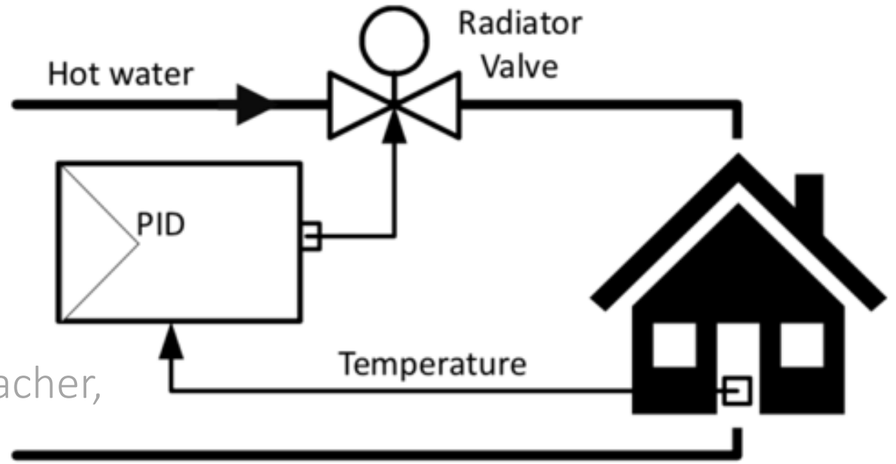
S. Harkema, The Lancet, Elsevier

[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.