

Reinforcement Learning @ MSR AI

<https://www.microsoft.com/en-us/research/group/reinforcement-learning-group/>

September 2018

Microsoft Research AI



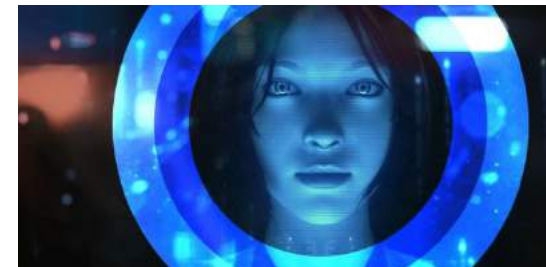
Reinforcement Learning: Stunts & Opportunities



Realistic Non-Player Characters

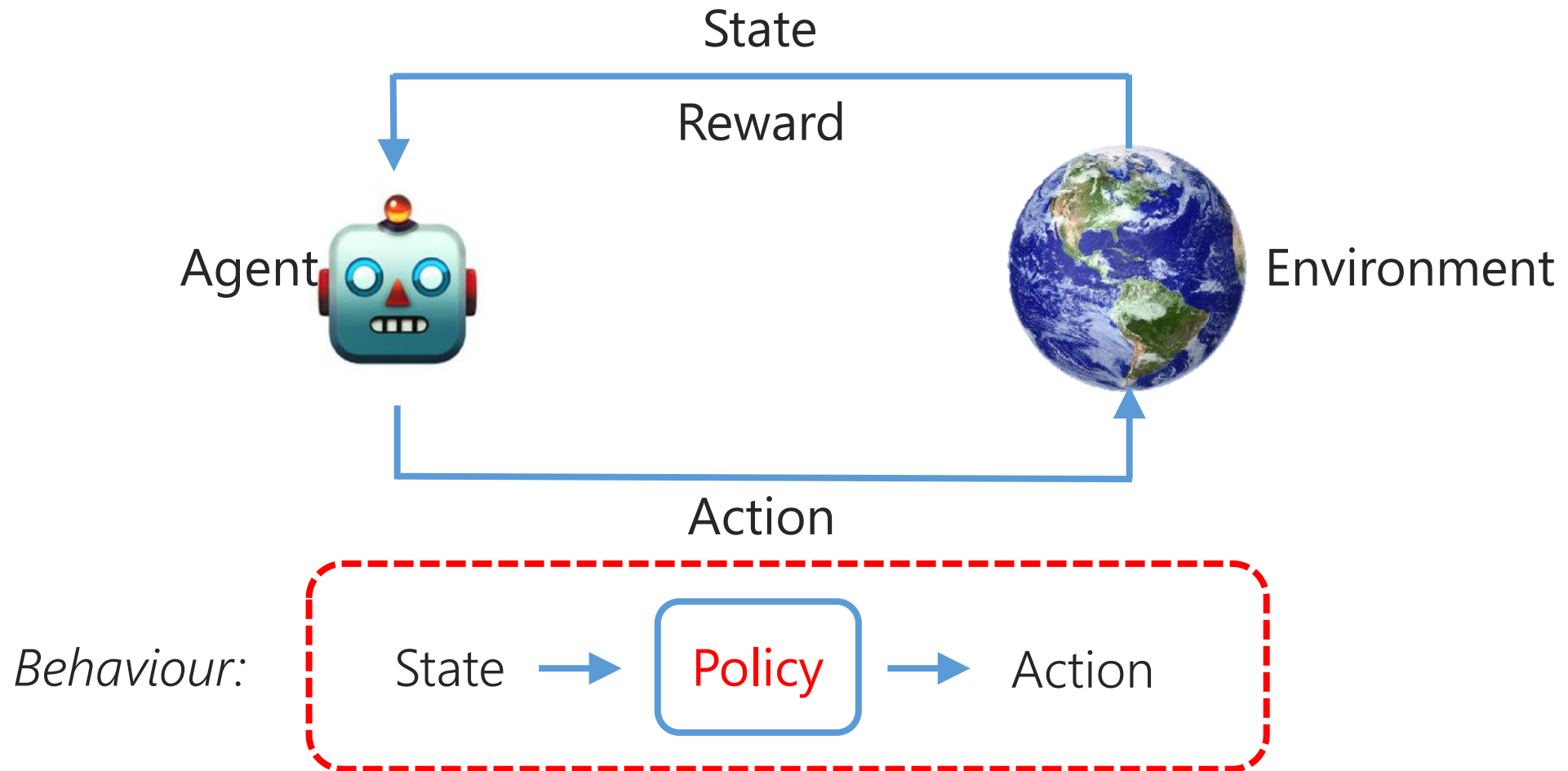


Automated Code Debugging



Advanced, pro-active Cortana

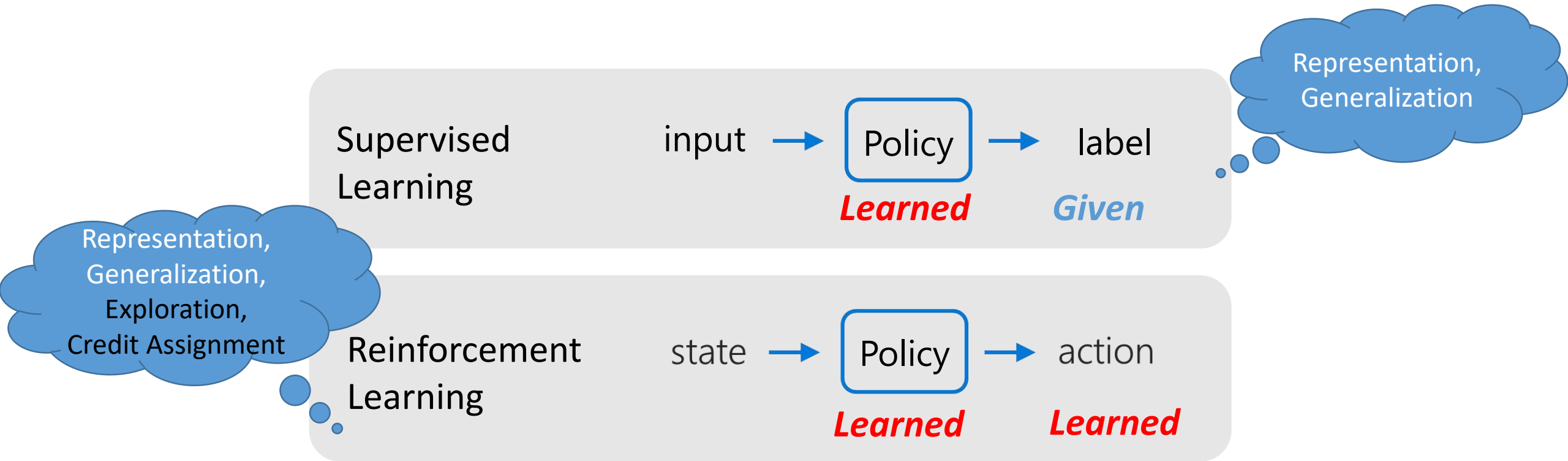
RL is a framework for sequential decision-making under uncertainty



Goal: Find the policy that results in the highest expected sum of rewards.

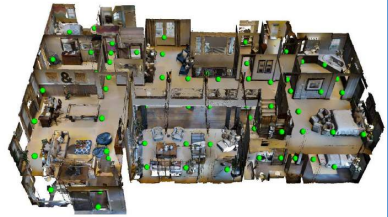
RL differs from Supervised Learning

The agent is not told how it should behave, but what it should achieve.

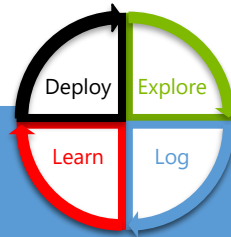


Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL)
to drive real-world AI scenarios.*



Simulation

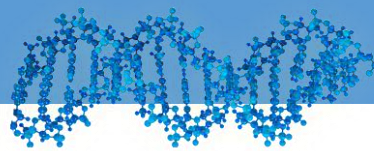


Digital

Physical



Foundations



RL in Simulation



Matthew Hausknecht Adith Swaminathan



1. RL for next-gen videogame AI

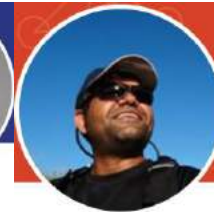
<https://github.com/Microsoft/malmo>

2. AirSim

<https://www.microsoft.com/en-us/research/project/aerial-informatics-robotics-platform/>



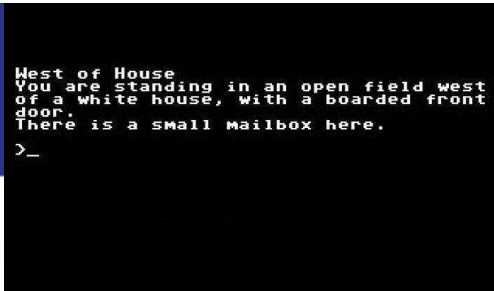
Ashish Kapoor



Shital Shah



Wendy Tay Ricky Loynd



3. Solving Interactive Fiction Games

<https://www.microsoft.com/en-us/research/project/textworld/>

4. Grounded vision-language interaction

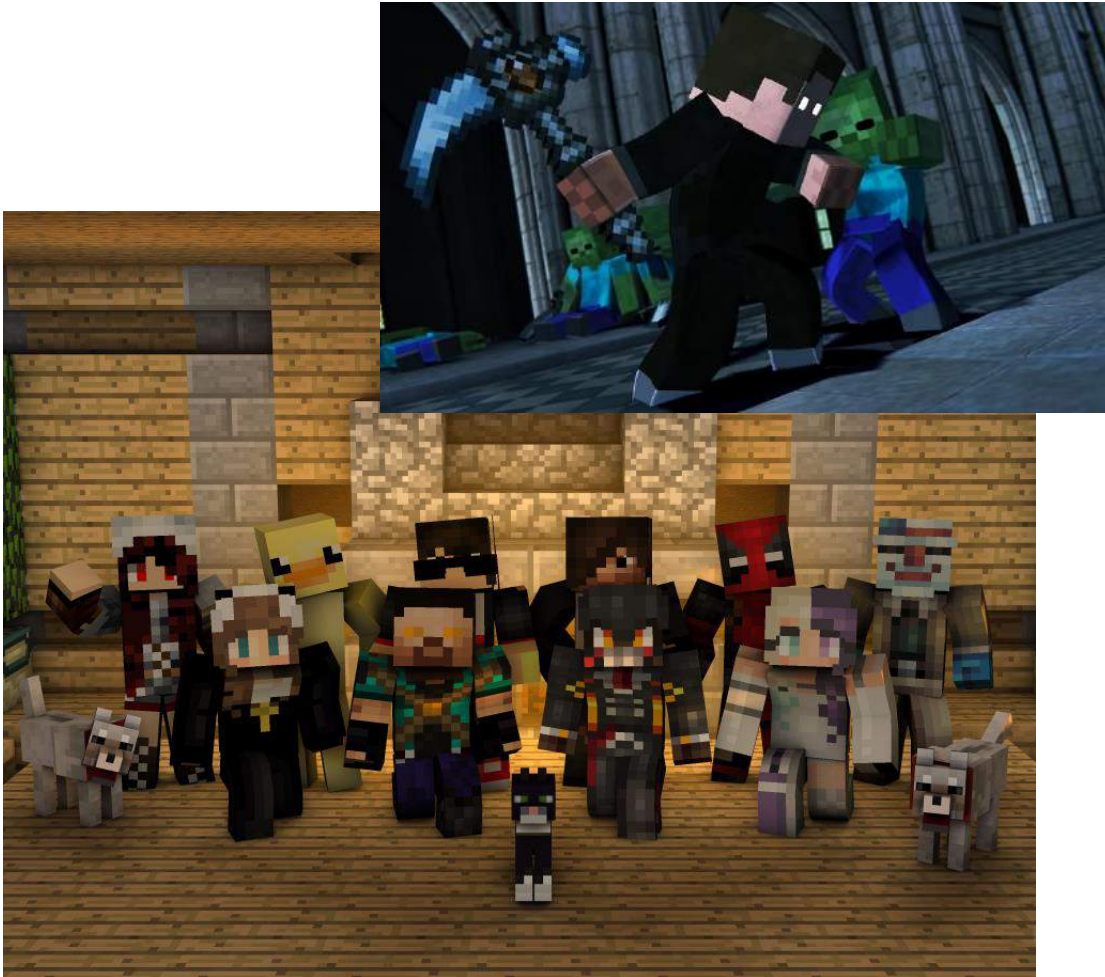


Debadeepta Dey

Bill Dolan



Next-Generation Game AI

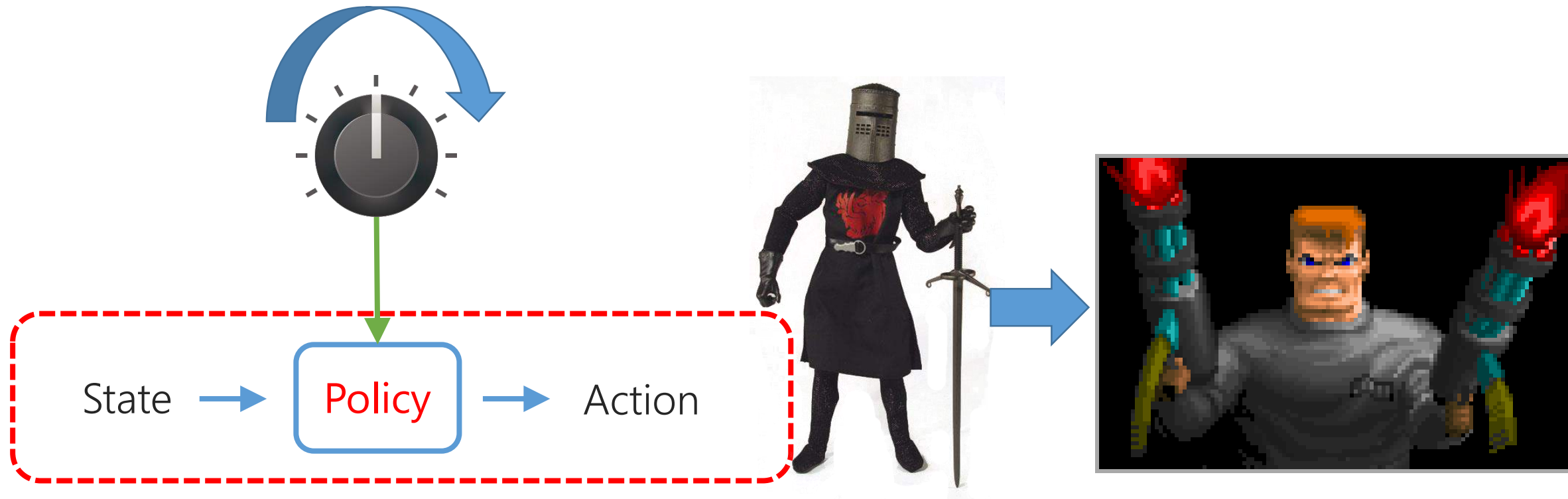


- Create agents that engage and entertain human players, rather than replacing them.
- Build agents capable of learning in open ended worlds like Minecraft.
- Learn policies that we can easily calibrate to specific behaviors/playstyles.



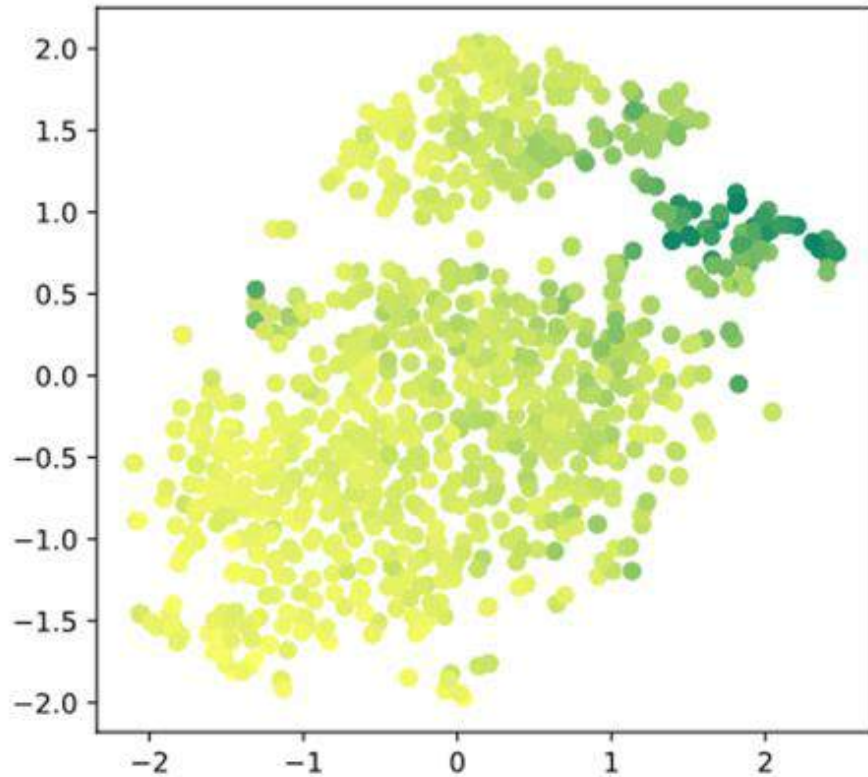
Towards Calibratable Learned Behaviors

Our Goal: Learn policies that we can easily calibrate to specific behaviors/playstyles.

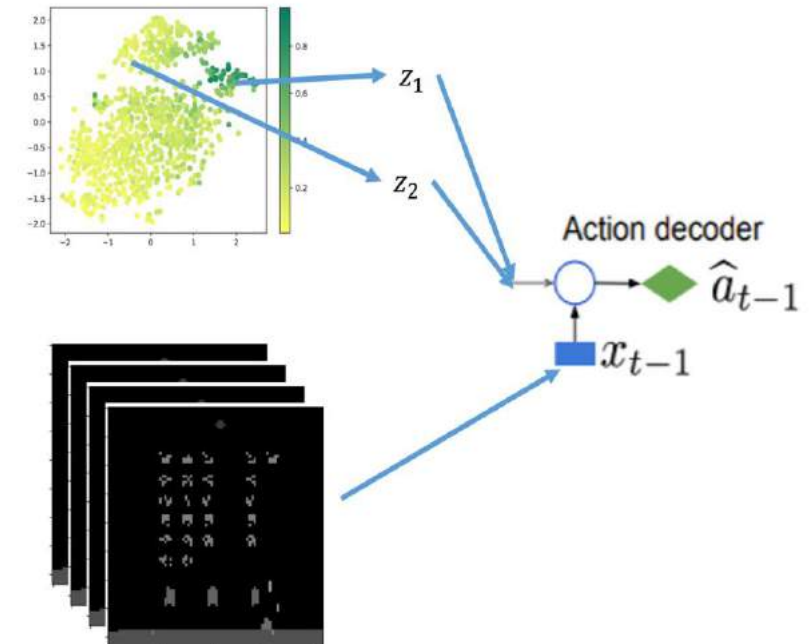
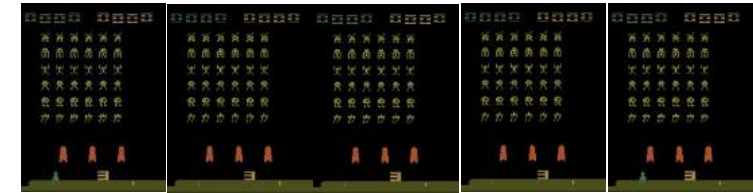


A Solution: Trajectory embedding + Imitation Learning

<http://atarigrandchallenge.com/data>

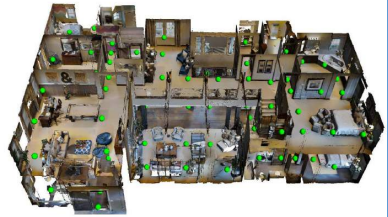


Frequency of firing

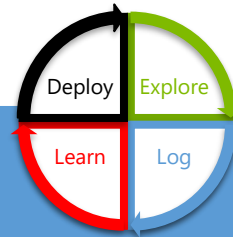


Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL)
to drive real-world AI scenarios.*



Simulation

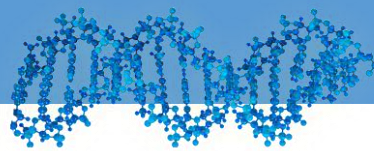


Digital

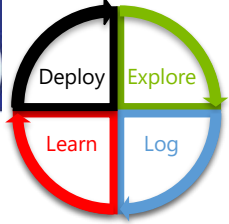
Physical



Foundations



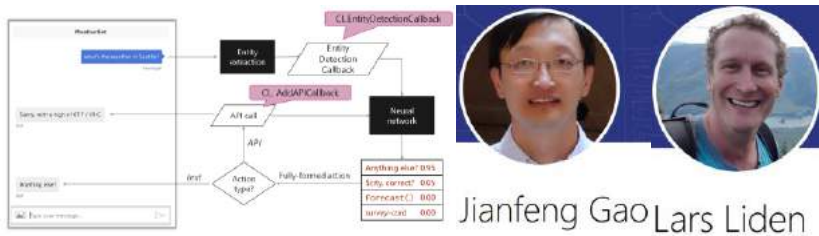
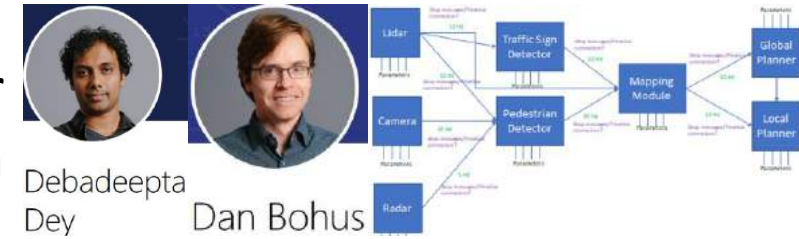
RL in the Digital World



1. Decision Service

<https://ds.microsoft.com>

2. Meta-reasoning for pipeline optimization



Jianfeng Gao Lars Liden

3. RL for Dialogue Systems

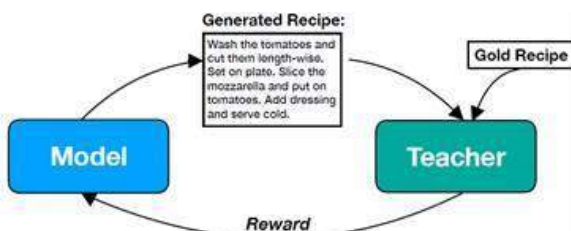
<https://labs.cognitive.microsoft.com/en-us/project-conversation-learner>

4. Next-gen Web Crawler for Bing

<http://www.pnas.org/content/115/32/8099>



Andrey Kolobov Yuval Peres



Asli Celikyilmaz Yizhe Zhang

5. RL for language generation

<https://www.microsoft.com/en-us/research/project/deep-communicating-agents-natural-language-generation/>

Decision Service (<https://ds.microsoft.com>)

End-to-end RL service on Azure for problems with immediate rewards (contextual bandits)

Significant gains in first and third-party applications



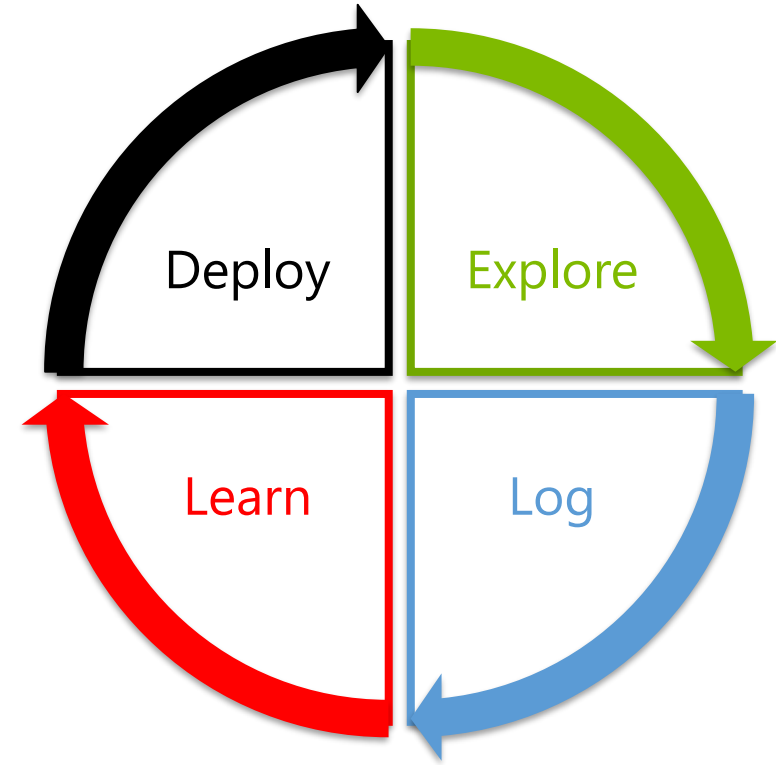
26% Lift vs.
Editorial



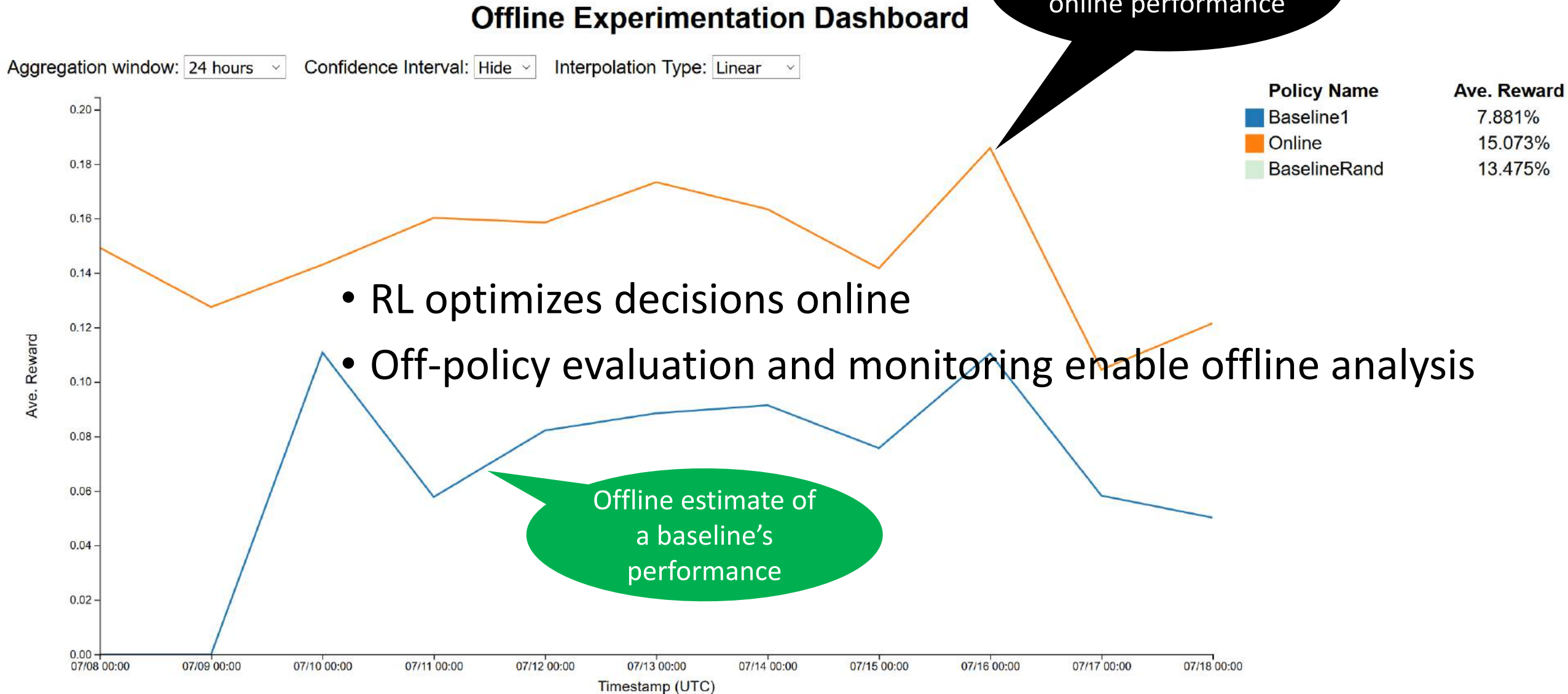
40% Lift vs.
Editorial



30% Revenue-
per-click
Improvement

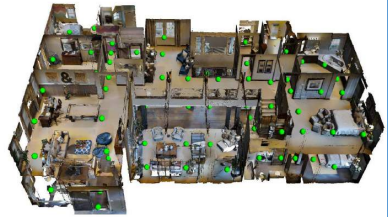


Counterfactual dashboard

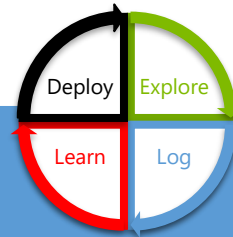


Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL)
to drive real-world AI scenarios.*



Simulation

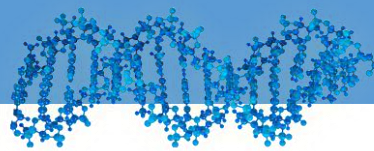


Digital

Physical



Foundations



RL in the Physical World

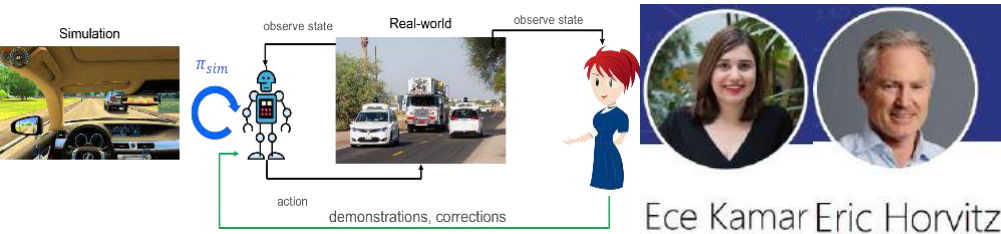


1. AI for Autonomous Soaring

<https://www.microsoft.com/en-us/research/project/project-frigatebird-ai-for-autonomous-soaring/>

2. Optimal Control for Indoor Agriculture

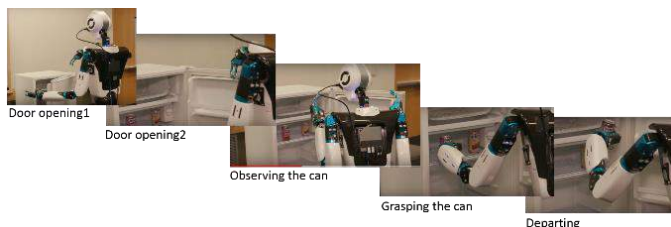
<https://www.microsoft.com/en-us/research/project/deep-reinforcement-learning-for-operational-optimal-control/>



3. Blind Spots in RL

4. Mobile Social Robotics on Ψ ($\backslash\psi$)

<https://www.microsoft.com/en-us/research/project/platform-situated-intelligence/>



5. Programming-by-Demonstration & RL

<https://blogs.microsoft.com/ai/step-inside-the-microsoft-envisioning-center/>

Project Frigatebird: AI for Autonomous Soaring

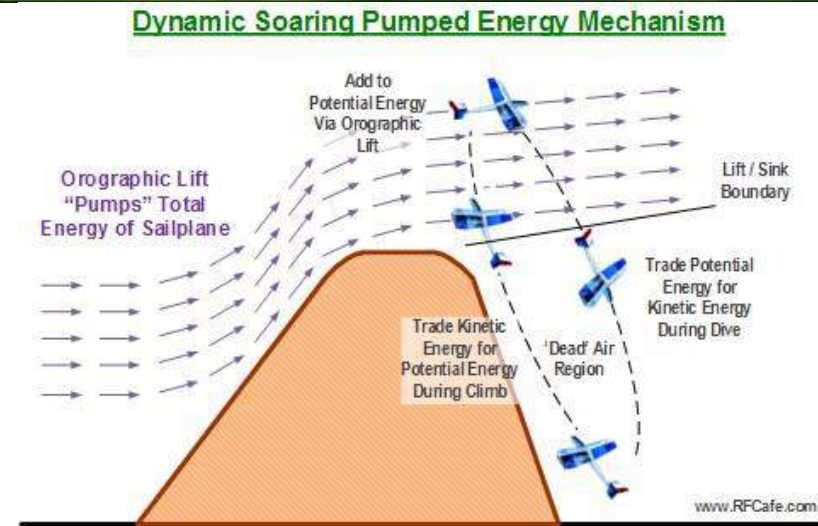
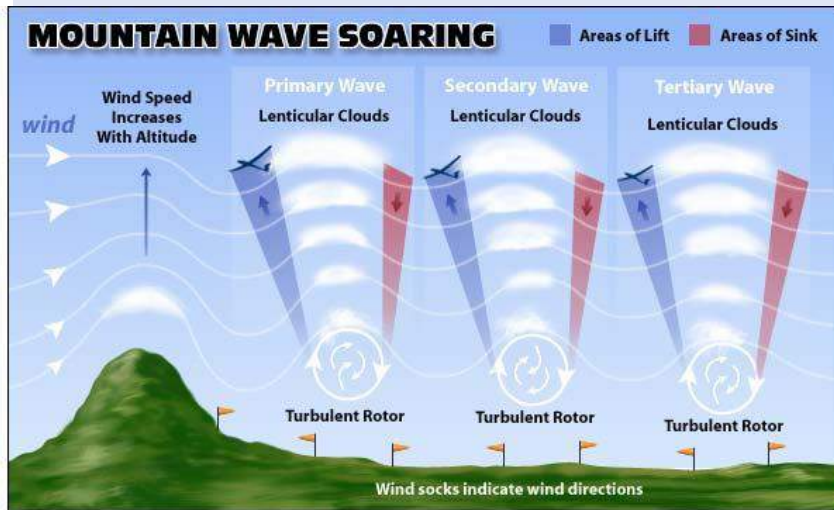
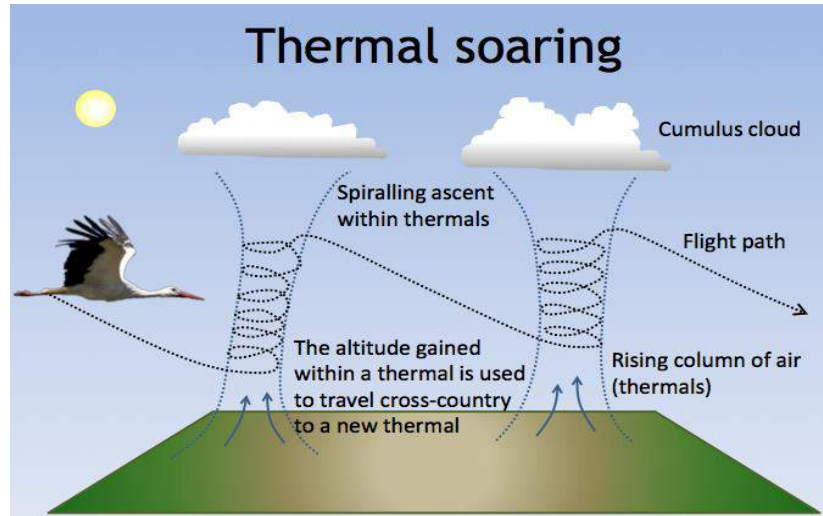


Frigatebirds and some other species can stay aloft for hours with hardly a wing flap, on energy they extract from thin air.

Goal: *Build AI to let sailplane (a.k.a. glider) UAVs fly long distances fully autonomously without active propulsion, using only soaring.*

Andrey Kolobov, Iain Guilliard, Rick Rogahn, Chris Lovett, Debadeepta Dey

How do Soaring Birds and Sailplanes Stay Aloft?



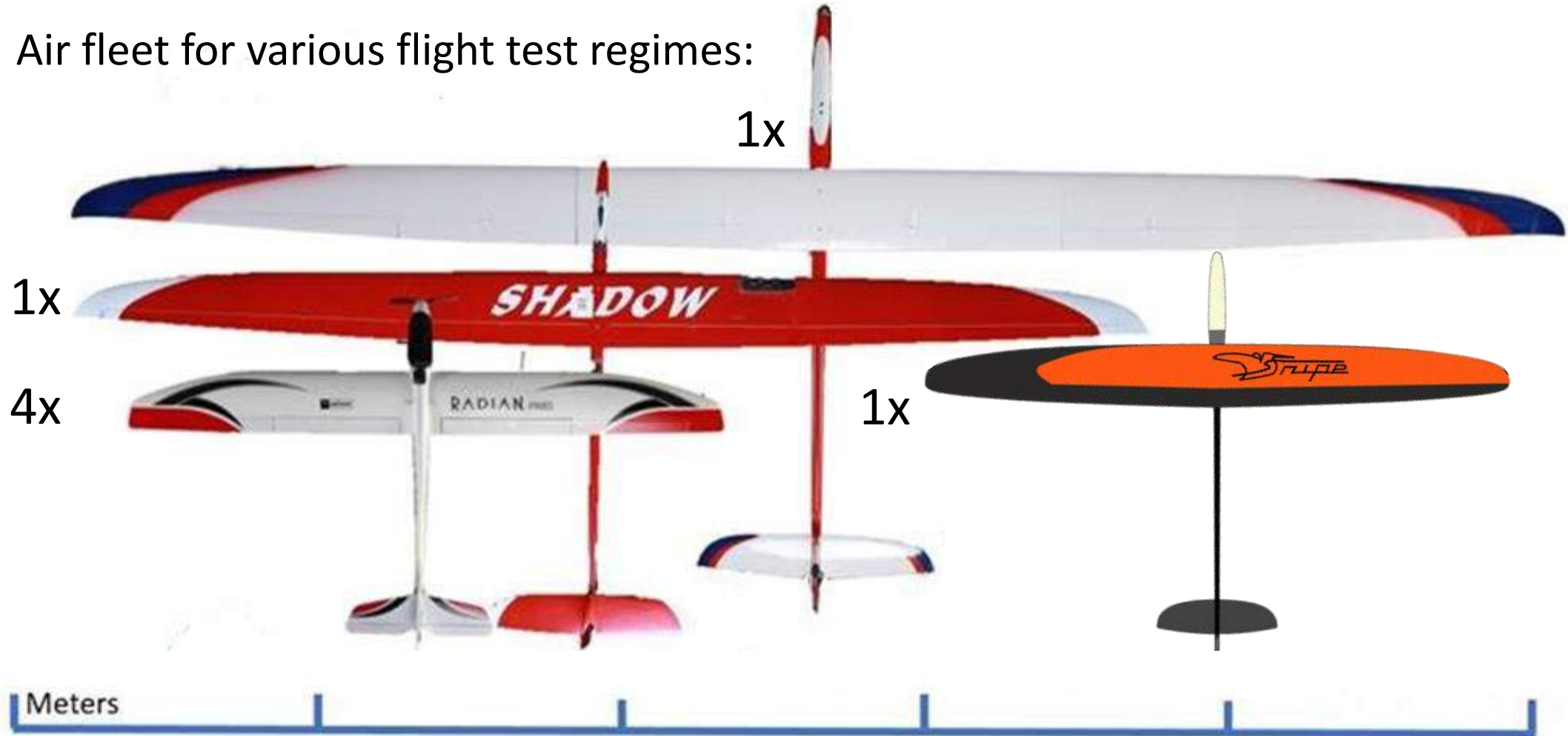
- By exploiting 3D wind patterns
- Wind patterns not directly visible, their locations not known with certainty
- Air movement can be sensed with onboard equipment...
- ...but no 2 windfields are alike – limited generalizability

Research challenges:

Learn to identify, exploit, predict, and plan for highly probabilistic atmospheric phenomena from little data

If It Doesn't Fly (Autonomously), It Doesn't Count!

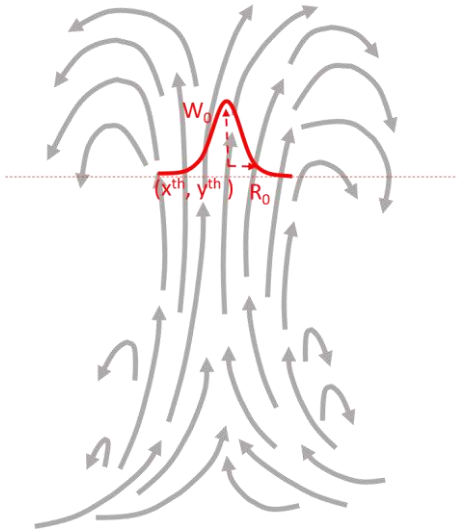
- Air fleet for various flight test regimes:



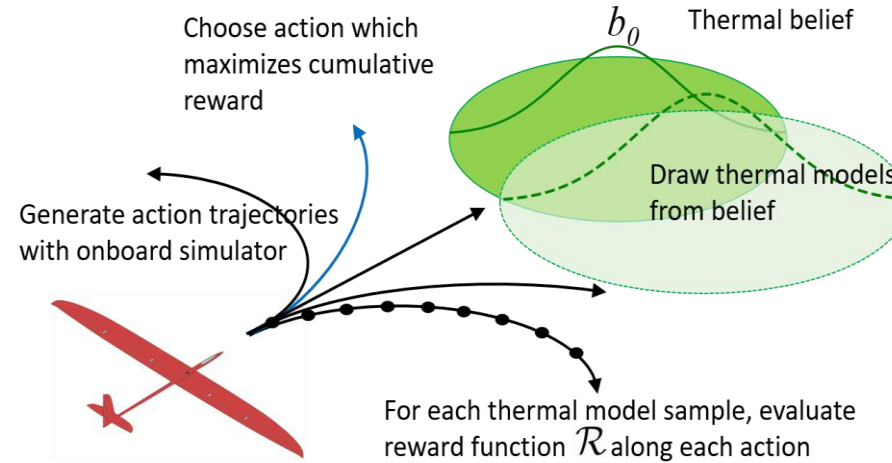
- Each carries GPS, airspeed sensor, etc., & onboard compute for autonomous flight
- Use soaring flight simulators (SilentWings, purpose-built) for sanity checks on the ground

Soaring in Thermals and Beyond

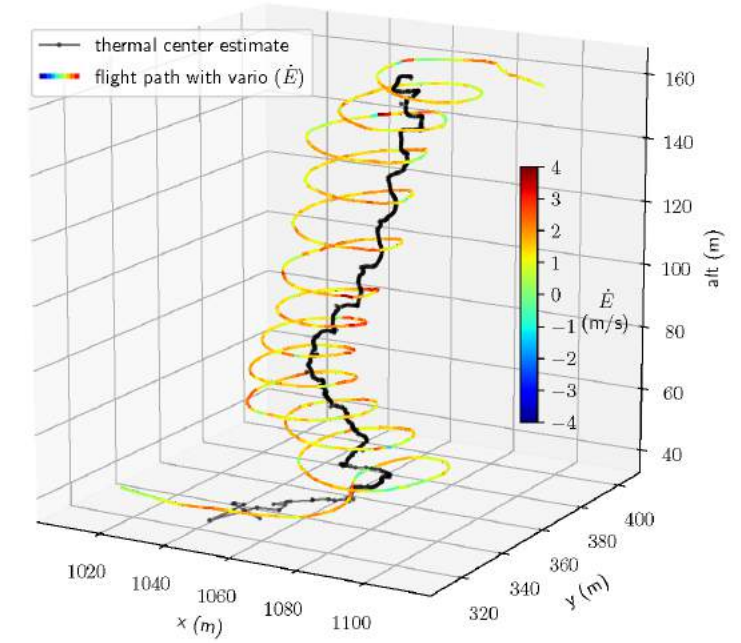
- First step: thermal soaring (RSS-2018, IROS-2018)



Thermal: irregular column of rising air



Approach: Bayesian RL done in real time aboard the sailplane



Results: successful thermal exploitation in real-world flights in adverse conditions

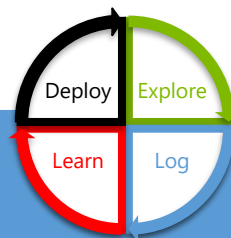
- Current research: vision for long-endurance flight planning in uncertain conditions
- More info, code, and data on Project Frigatebird's webpage. Come talk to the crew!

Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL)
to drive real-world AI scenarios.*



Simulation

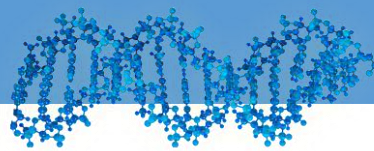


Digital

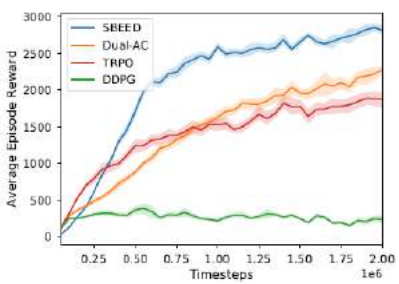
Physical



Foundations



RL Foundations



Lin Xiao

Zeyuan Allen-Zhu

1. Interplay of Optimization-Representation-RL

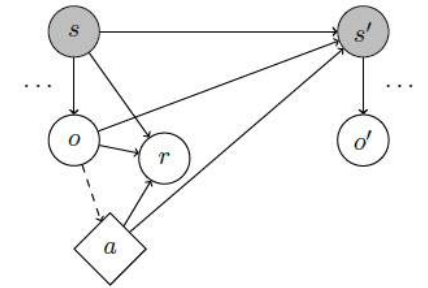
e.g. <https://www.microsoft.com/en-us/research/publication/sbeed-convergent-reinforcement-learning-with-nonlinear-function-approximation/>

2. Exploration

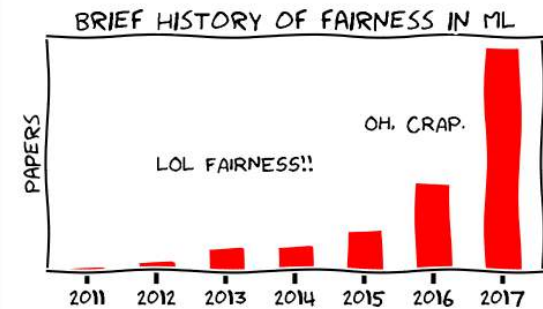
e.g. <https://arxiv.org/abs/1807.03765>,
<https://arxiv.org/abs/1802.03386>,
<https://arxiv.org/abs/1711.01037> etc.



Sébastien Bubeck Alekh Agarwal



(c) A unified model that subsumes (a) and (b) and yields low Bellman rank.



Hanna Wallach

Fernando Diaz

3. Social and Ethical Aspects

e.g. <https://www.microsoft.com/en-us/research/publication/exploring-or-exploiting-social-and-ethical-implications-of-autonomous-experimentation-in-ai/>

4. Imitation Learning

e.g. <https://www.microsoft.com/en-us/research/publication/learning-gather-information-via-imitation/>



Debadeepta Dey Ashish Kapoor



SBEED: Convergent RL w/ Function Approximation

SBEED (out of the Deadly Triad):

Convergent Reinforcement Learning with
Nonlinear Function Approximation

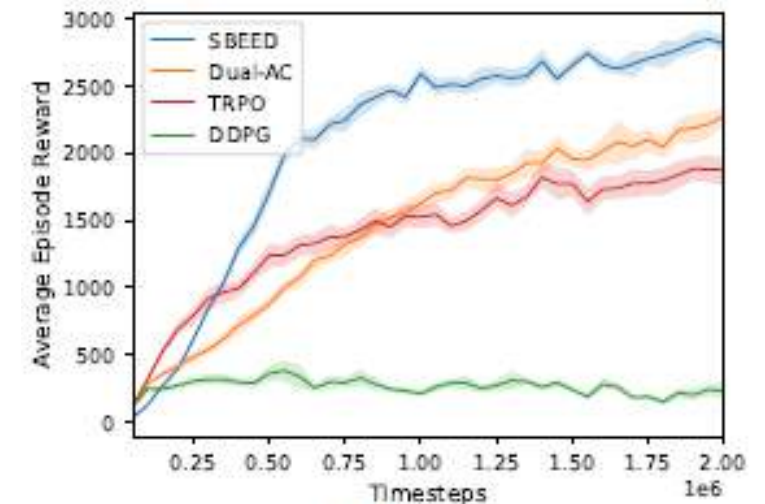
Bo Dai^{1→2}, Albert Shaw¹, Lihong Li², Lin Xiao³, Niao He⁴,
Zhen Liu¹, Jianshu Chen⁵, Le Song¹

¹Gatech, ²Google Brain, ³Microsoft Research, ⁴UIUC, ⁵Tencent AI

(Appeared in ICML 2018, arXiv:1712.10285)

Stability/Convergence of RL algorithms

- Impressive empirical success of DeepRL, but,
- No convergence guarantees, often diverges!
- Limited theory and algorithms (e.g. linear)
- Major Open Problem for decades



(e) Hopper

Smoothed Bellman Error EmbeDing (SBEED)

- First provably convergent ADP/RL algorithm with general nonlinear function approximation
- Tackle RL problems by directly solving the Bellman equation

$$\min_V \mathbb{E}_s \left[\left(V(s) - \max_a (R(s, a) + \gamma \mathbb{E}_{s'|s,a} [V(s')]) \right)^2 \right]$$

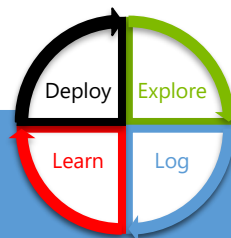
- An obvious attempt, but has two difficulties
 - #1: max operator is nonsmooth (hard for analysis and unstable in practice)
 - ✓ Solution: smoothing using entropy regularization over policy simplex
 - #2: conditional expectation inside square, causes biased stochastic gradient
 - ✓ Solution: primal-dual lifting into minimax problem using Fenchel conjugate

Our Mission

*We conduct ground-breaking research in Reinforcement Learning (RL)
to drive real-world AI scenarios.*



Simulation

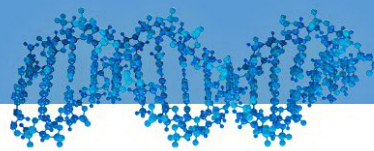


Digital

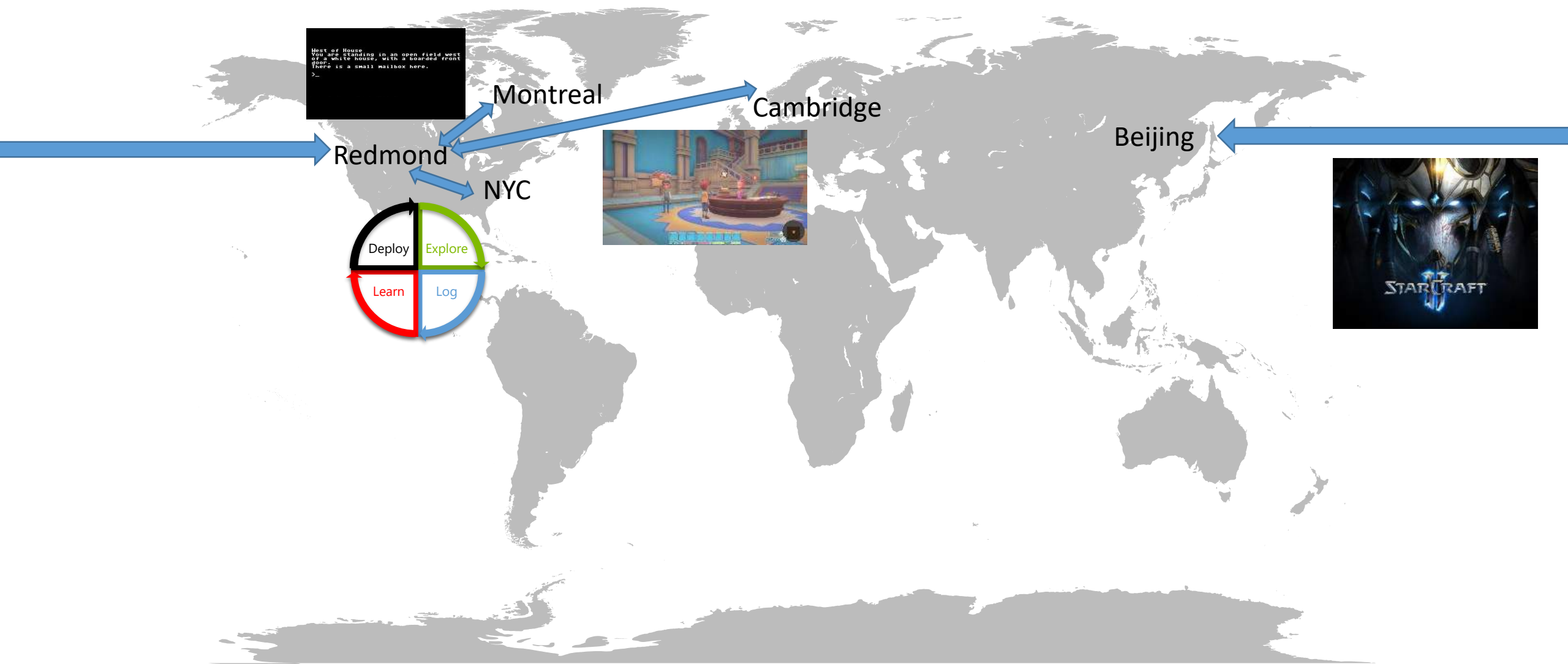
Physical

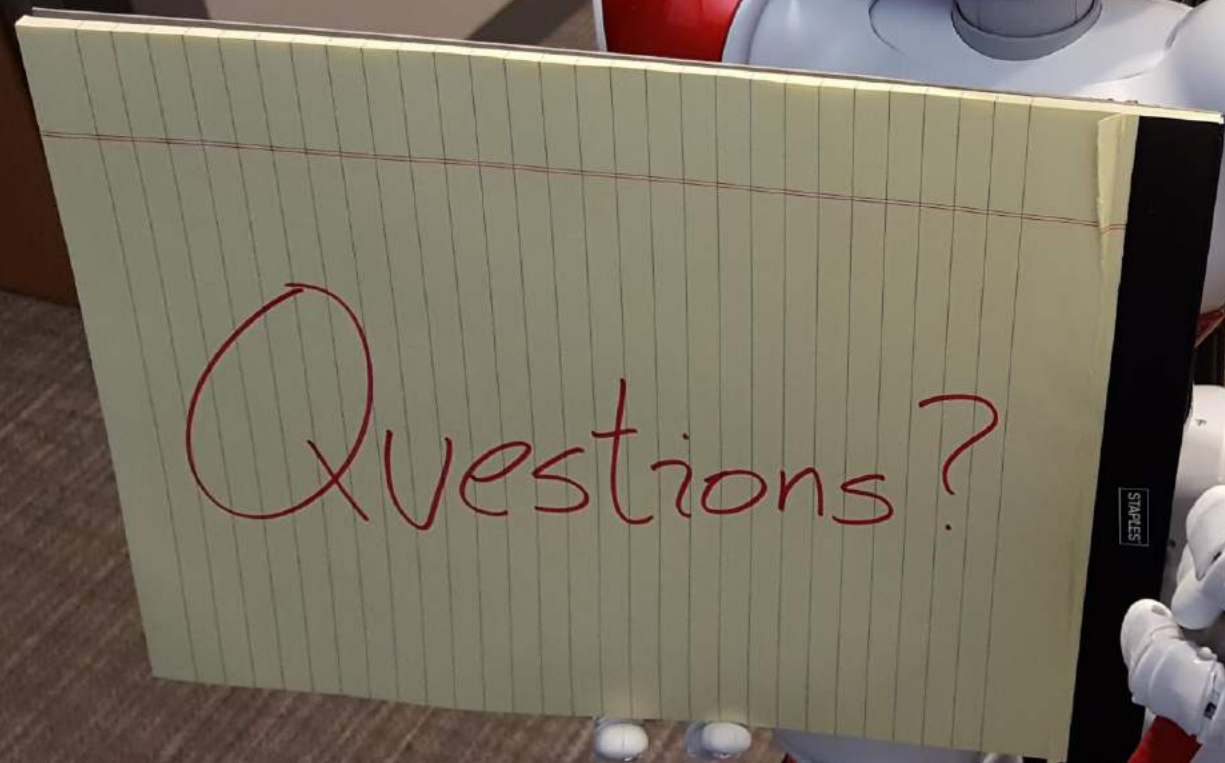


Foundations



RL @ MSR





Questions?

Appendix - Simulation

<https://www.microsoft.com/en-us/research/group/reinforcement-learning-group/>

September 2018

Microsoft Confidential

Microsoft Research AI



Grounded Visual Navigation via Imitation Learning



Find a laptop in one of the bedrooms

Kabi! I am lost! Help me!



Agent **executes** and **decides** when to ask for help

- Long-horizon sequential decision-making.
- Sensing only via vision.
- Photo-realistic real-world indoor datasets (Matterport3D).
- Can setup dialog with human for assistance.
- Requires common-sense reasoning.
- [Test-bed for imitation/reinforcement learning.](#)
- [Sim-to-real transfer to real world robots.](#)

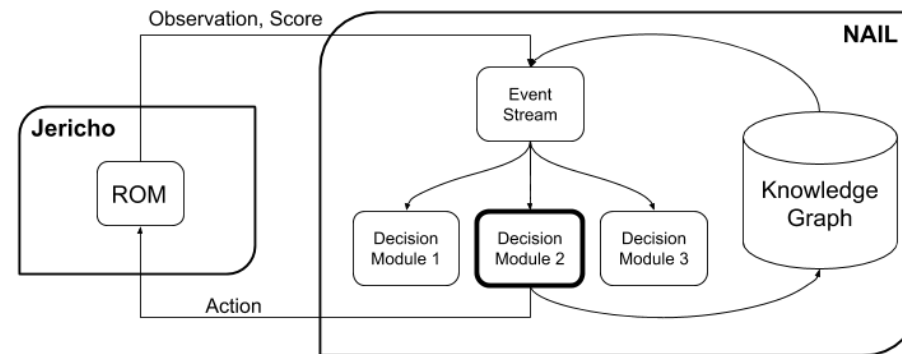


RL in Text-based Adventure Games

- Intersection of RL & NLP
- Agents with language understanding
- Commonsense reasoning
- Map building & Memory



```
West of House
You are standing in an open field west
of a white house, with a boarded front
door.
There is a small mailbox here.
>_
```



Appendix - Digital

<https://www.microsoft.com/en-us/research/group/reinforcement-learning-group/>

September 2018

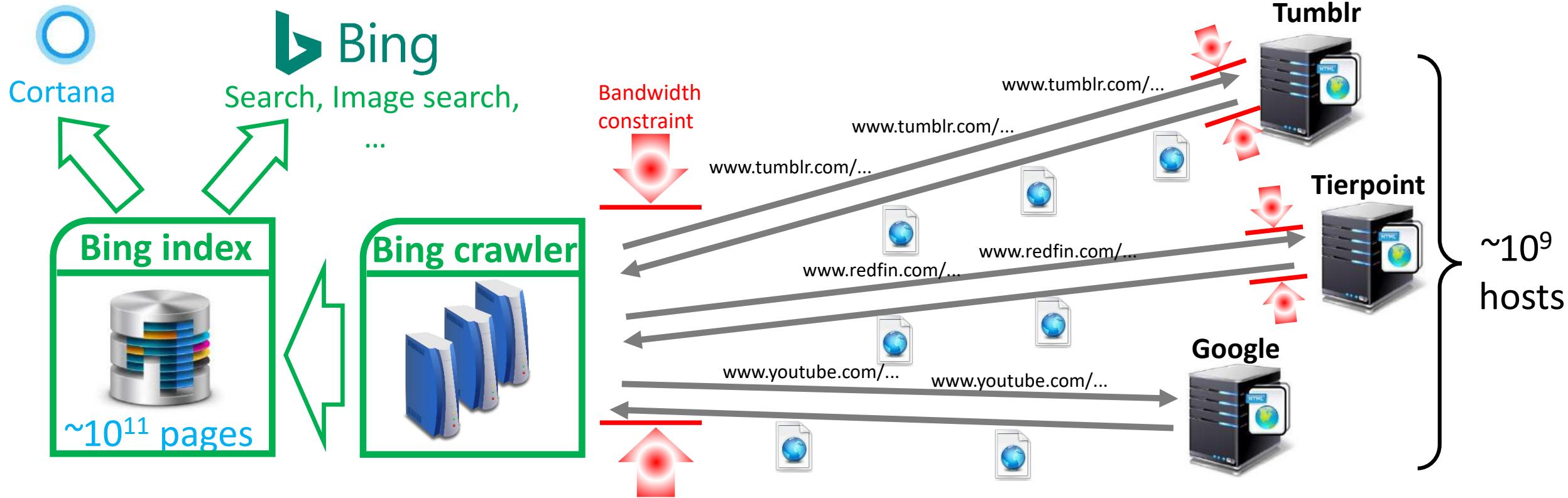
Microsoft Confidential

Microsoft Research AI



Scheduling for Bing's Next-Generation Web Crawler

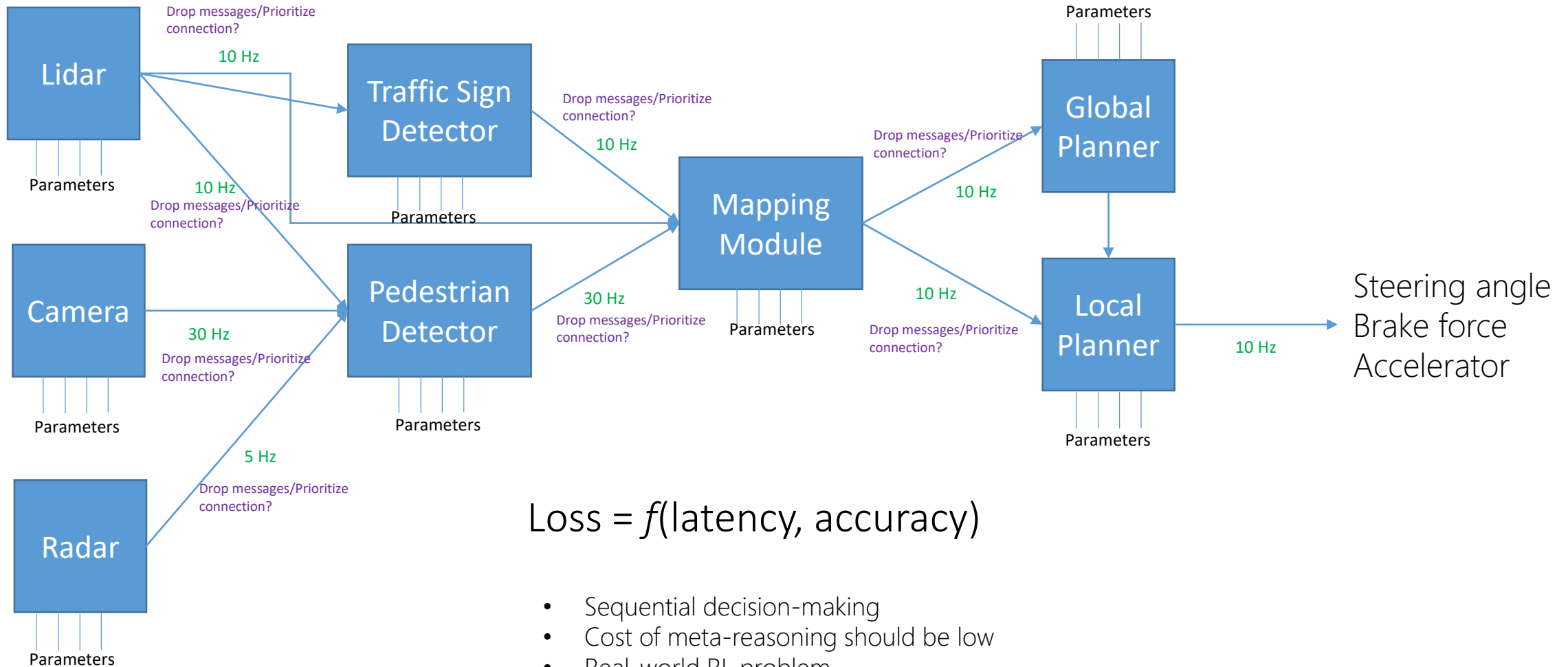
Andrey Kolobov, Yuval Peres, Eric Horvitz, Bing IndexGen team



- Bing's index is a storage of Web content. Web pages change & need to be *recrawled* to keep Bing's index *fresh*.
- How do we compute in *near-linear* time a *crawl scheduling policy* that *maximize Bing index's content freshness*...
 - ... while observing web hosts' and Bing's own constraints on crawl bandwidth ...
 - ... and learning to predict Web page changes ...
 - ... over billions of hosts and 100s of billions of pages?

Meta-Reasoning for Pipeline Optimization

Timing & quality tradeoffs, uncertainties with modular pipelines



Appendix - Physical

<https://www.microsoft.com/en-us/research/group/reinforcement-learning-group/>

September 2018

Microsoft Confidential

Microsoft Research AI



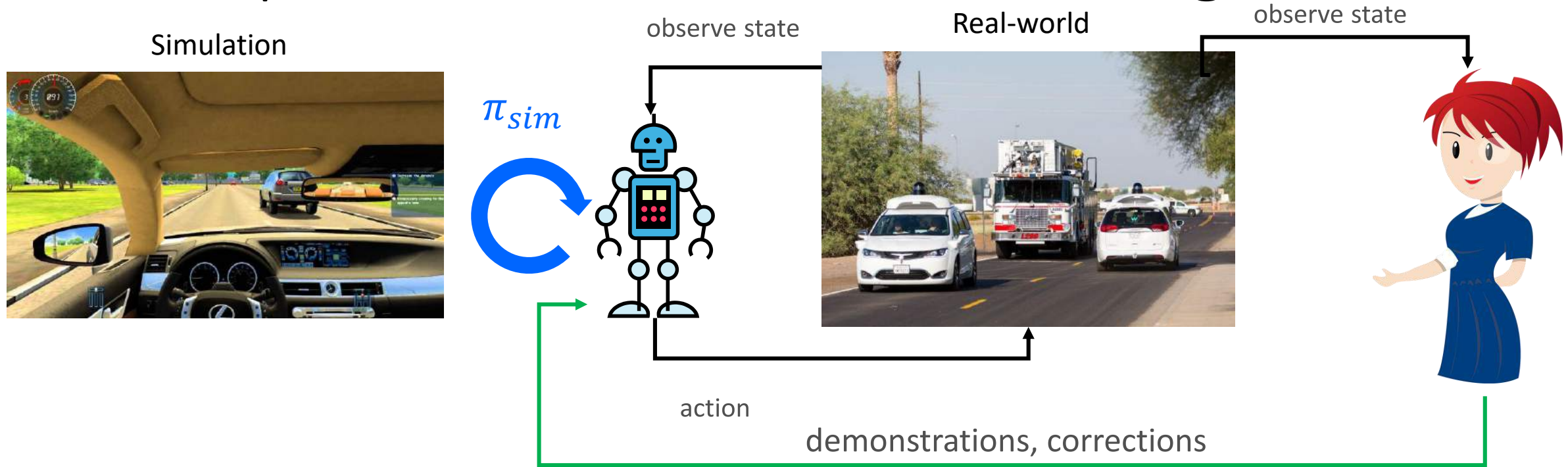
Project Sonoma: Optimal Control for Indoor Farms



- Learn a policy that can optimally control plant growth in indoor farms
- Real-world application that requires advances in model-based RL, transfer RL, POMDP solvers

Kenneth Tran, Ranveer Chandra, Chetan Bansal + external collaborators

Blind Spots in Reinforcement Learning



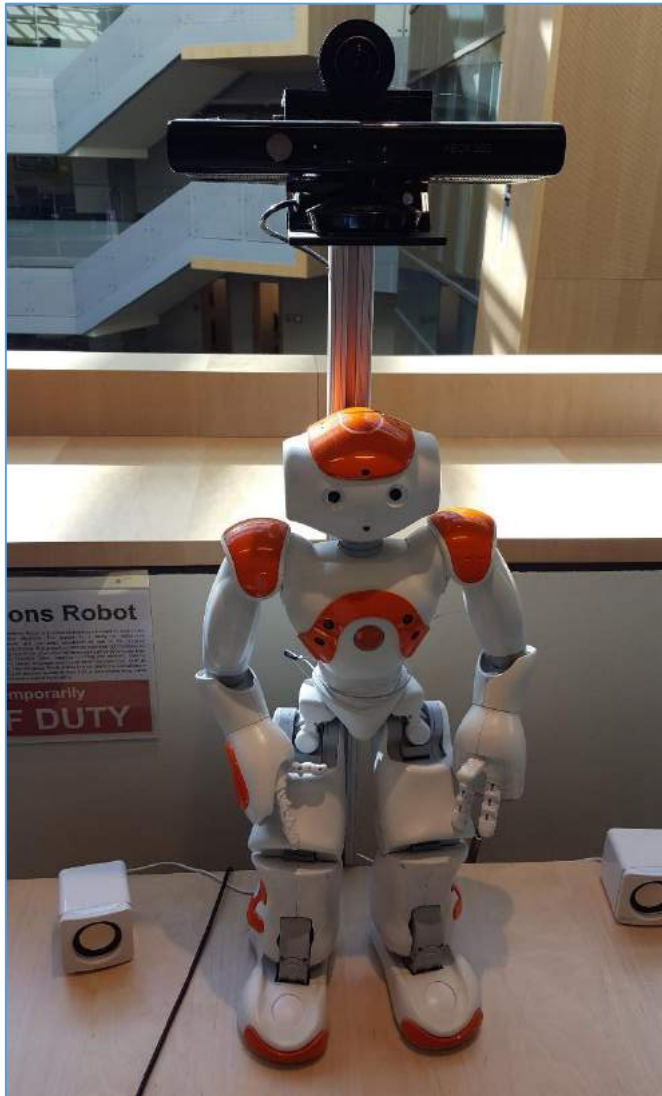
Goal: Create model of blind spots

Blind spot: Systematic input regions with divergence from optimal policy

Complicated by incomplete state representations

Ramya Ramakrishnan, Ece Kamar, Besmira Nushi, Debadeepta Dey and Eric Horvitz

ψ : Assistive, Mobile, Social Robotics with \psi



+



Sean Andrist, Dan Bohus, Ashley Feniello, Eric Horvitz

Programming-by-demonstration and RL

- PbD provides the initial solution
- RL refines the solution

