Real World Reinforcement Learning Fundamentals

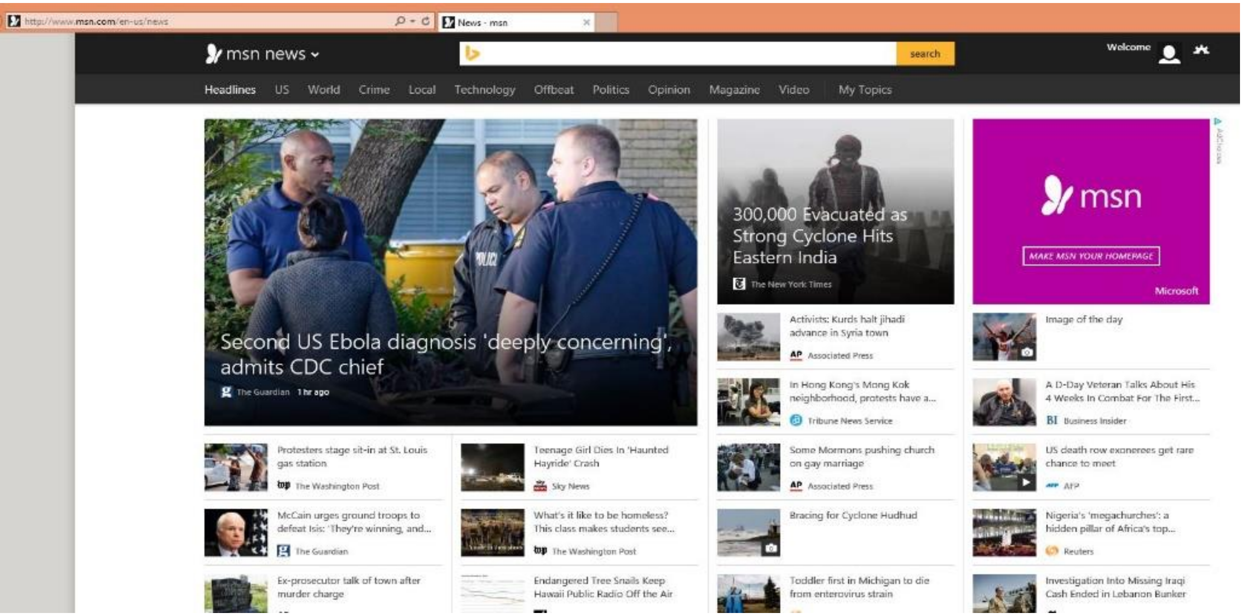
John Langford

NeurIPS 2019, Dec 8

https://vowpalwabbit.org/neurips2019/

Why?

Which News?

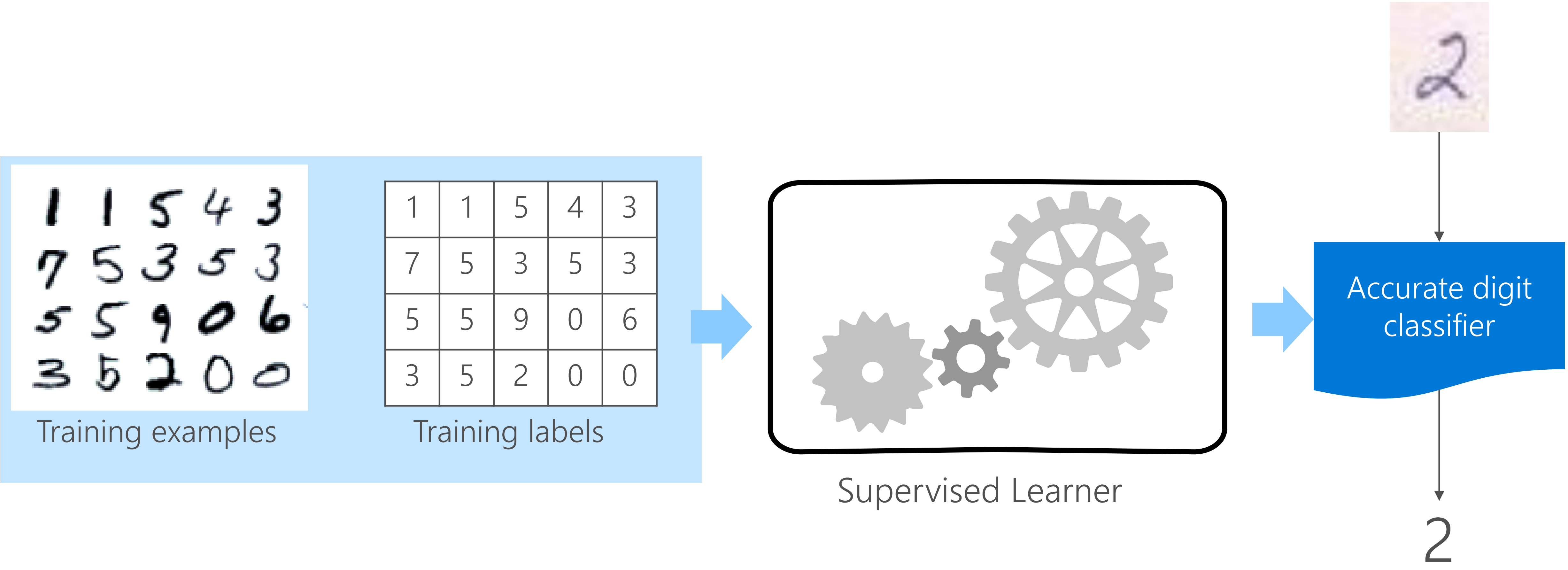


28% lift

Which Game?

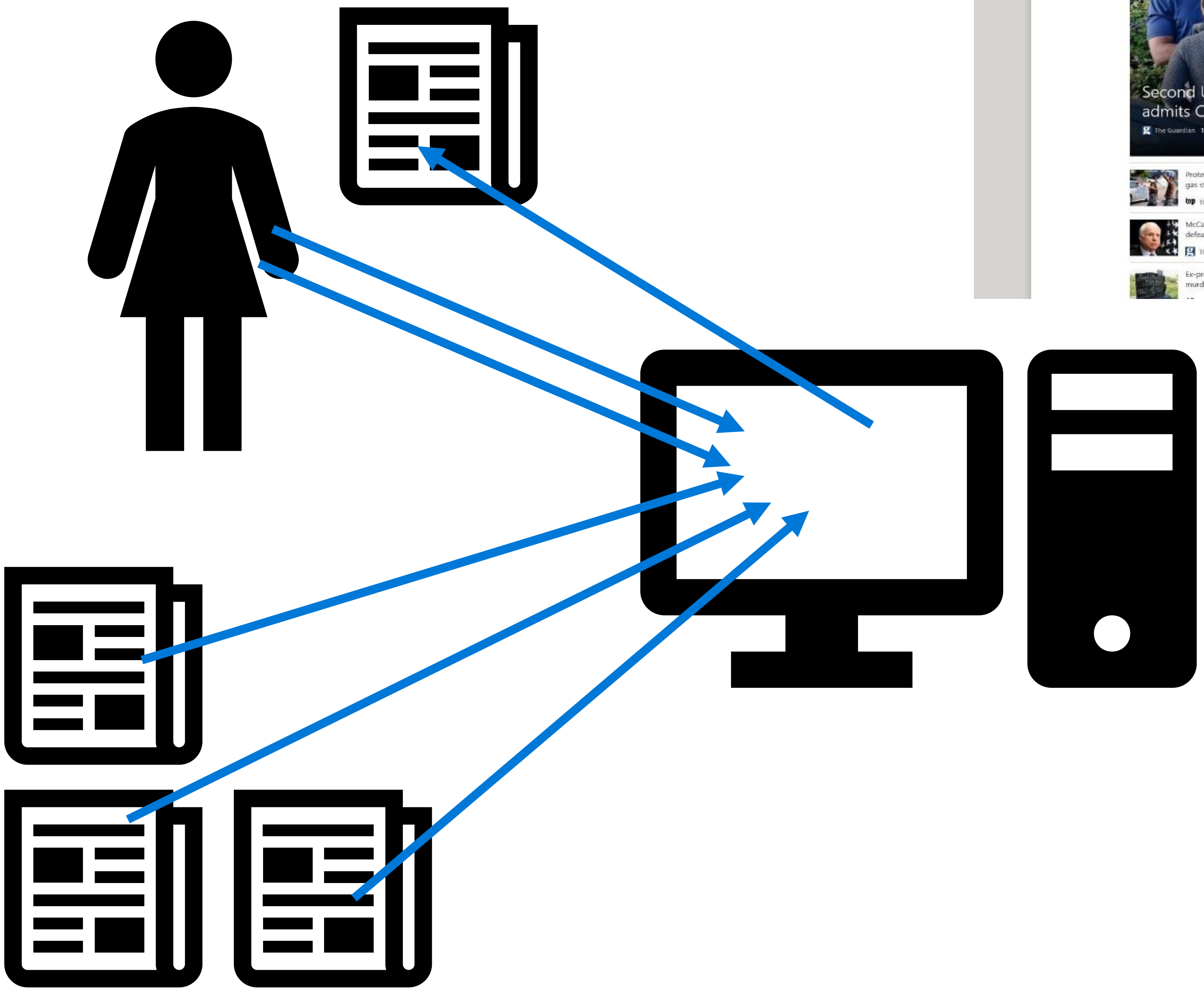
40% lift

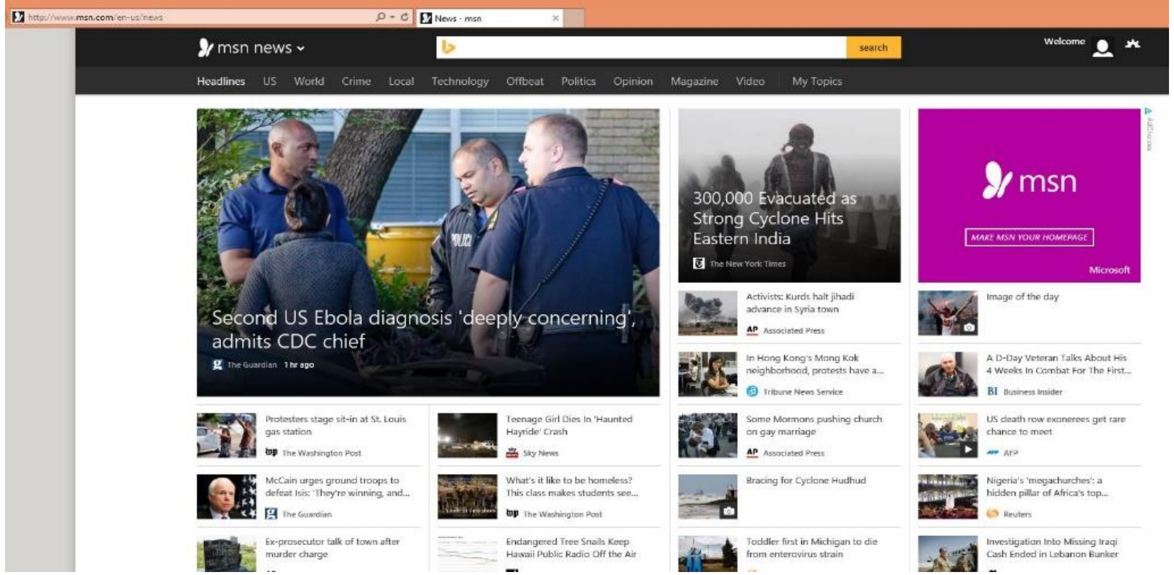
The Baseline: Supervised Learning



Supervised Learning is cool

How about news?





A standard pipeline

1. Collect $(user, article)$ information.
2. Hire editor to judge $relevance(user, article)$
3. Learn $\widehat{relevance}(user, article)$
4. Act with best $article$ from $\widehat{relevance}(user, article)$
5. Deploy in A/B test for 2 weeks
6. A/B test fails ☹

Q: What goes wrong?

Is Ukraine 

interesting to John  ?

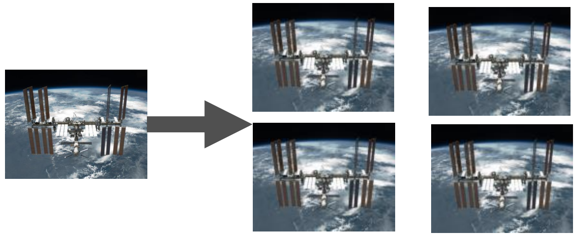
A: Need Right Signal for Right Answer

What goes wrong?

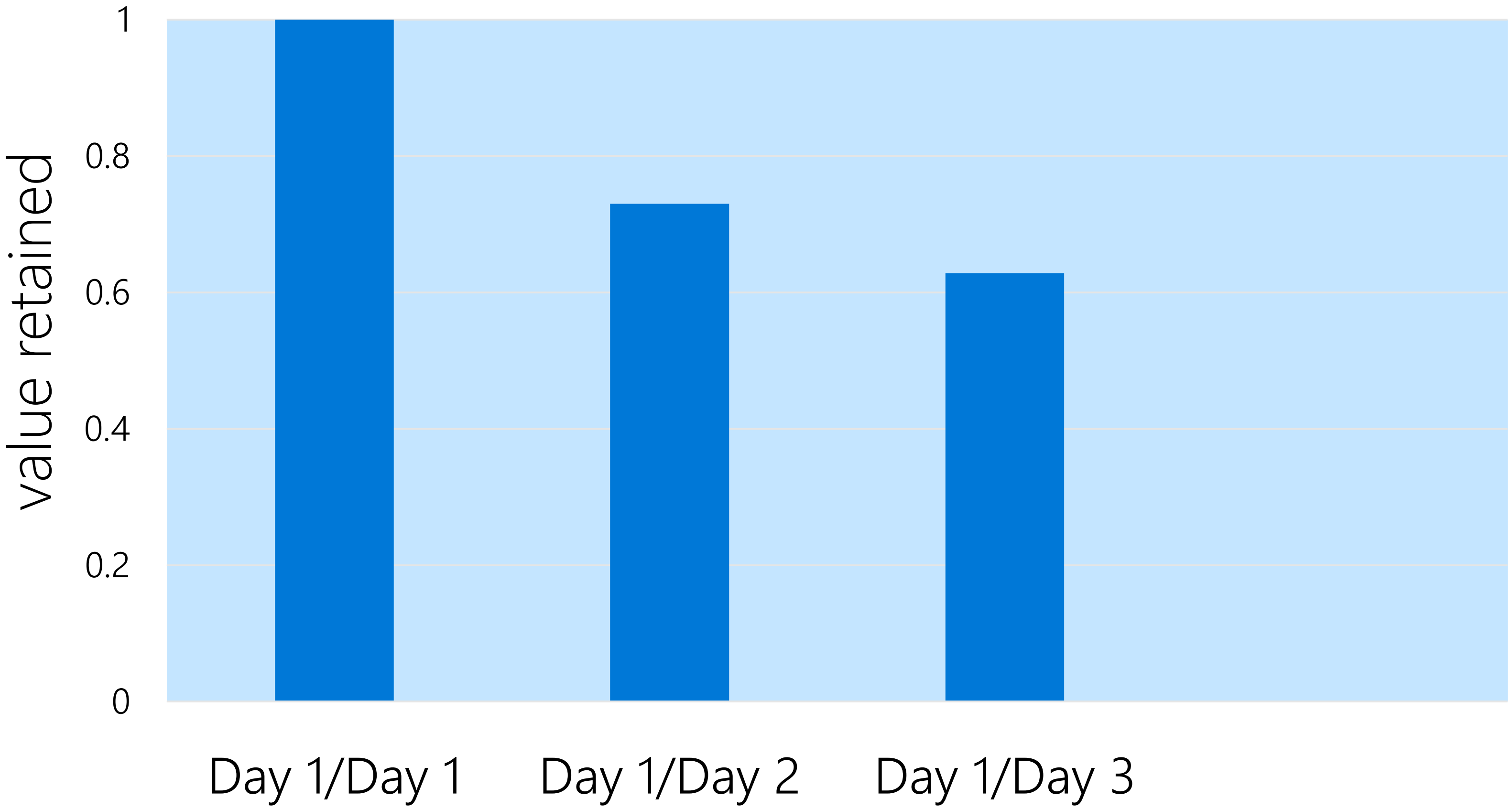
What is the probability of click on a food article

If you only display a space article?

We must avoid "self-fulfilling prophecy"

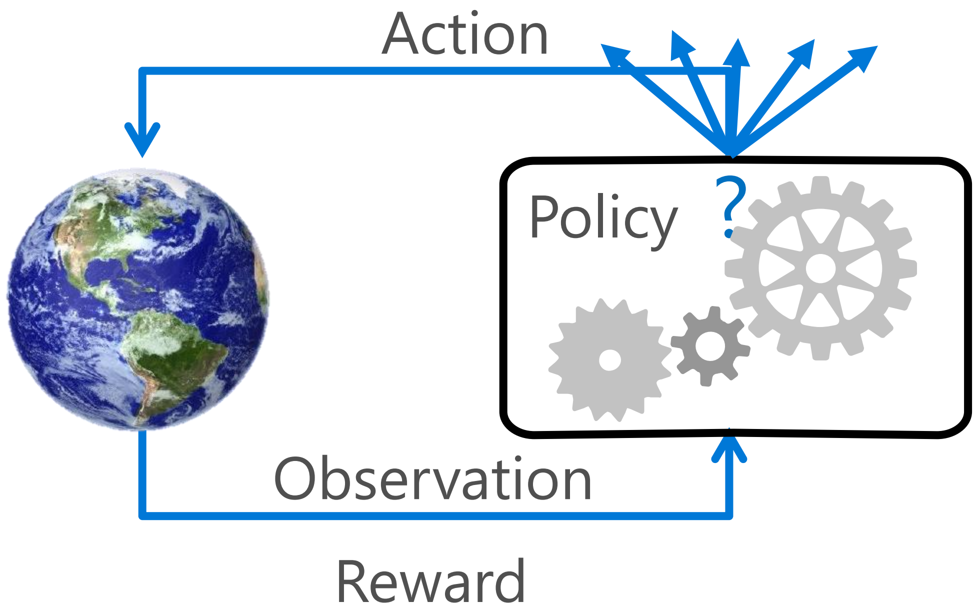


What else goes wrong?



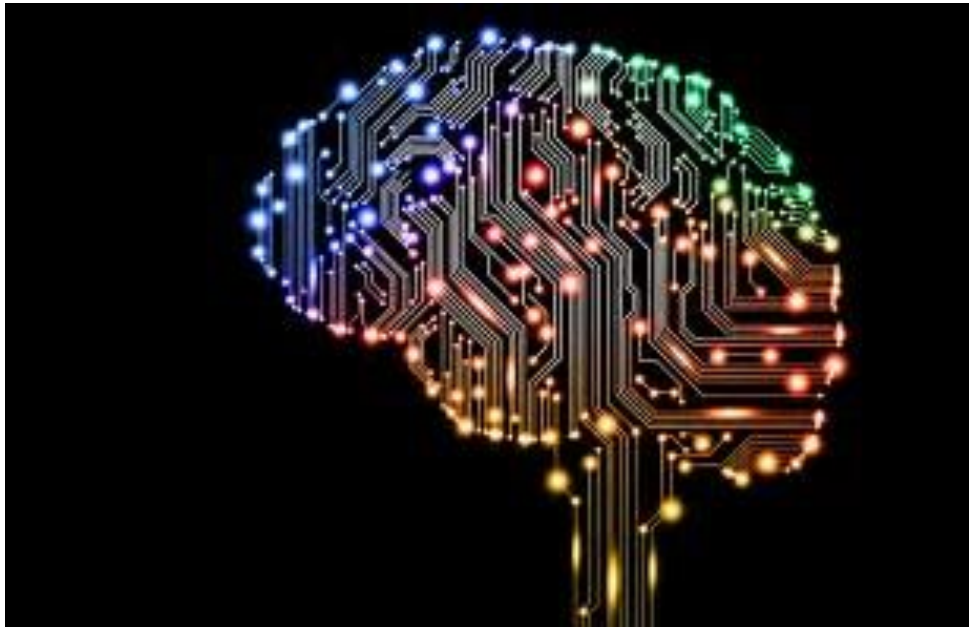
The world changes!

Reinforcement Learning can do this!



Goal: Find a policy maximizing the sum of rewards

Q: One last Why…

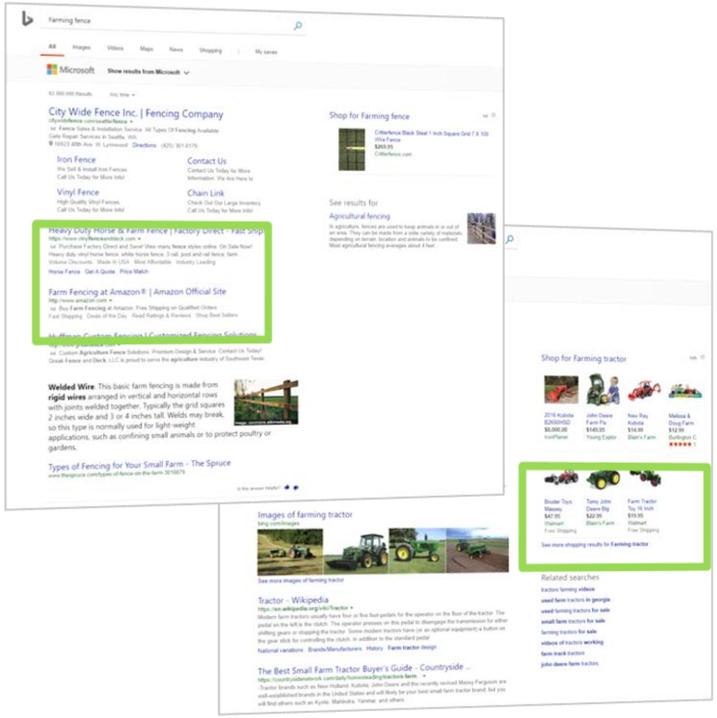


~~AI: A function programmed with data~~

AI: An economically viable digital agent that explores, learns, and acts

Content

Layout

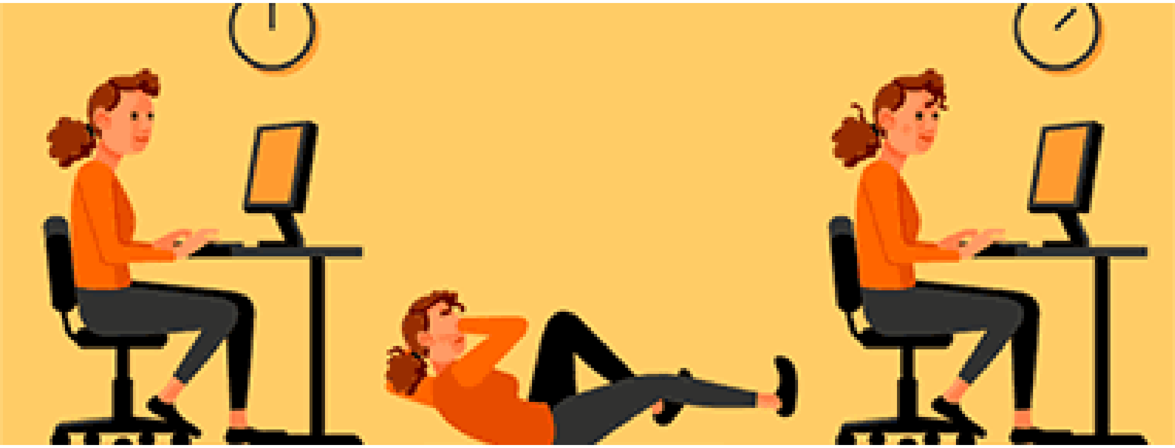


Creative

"Book Your Vacation to Hawaii"

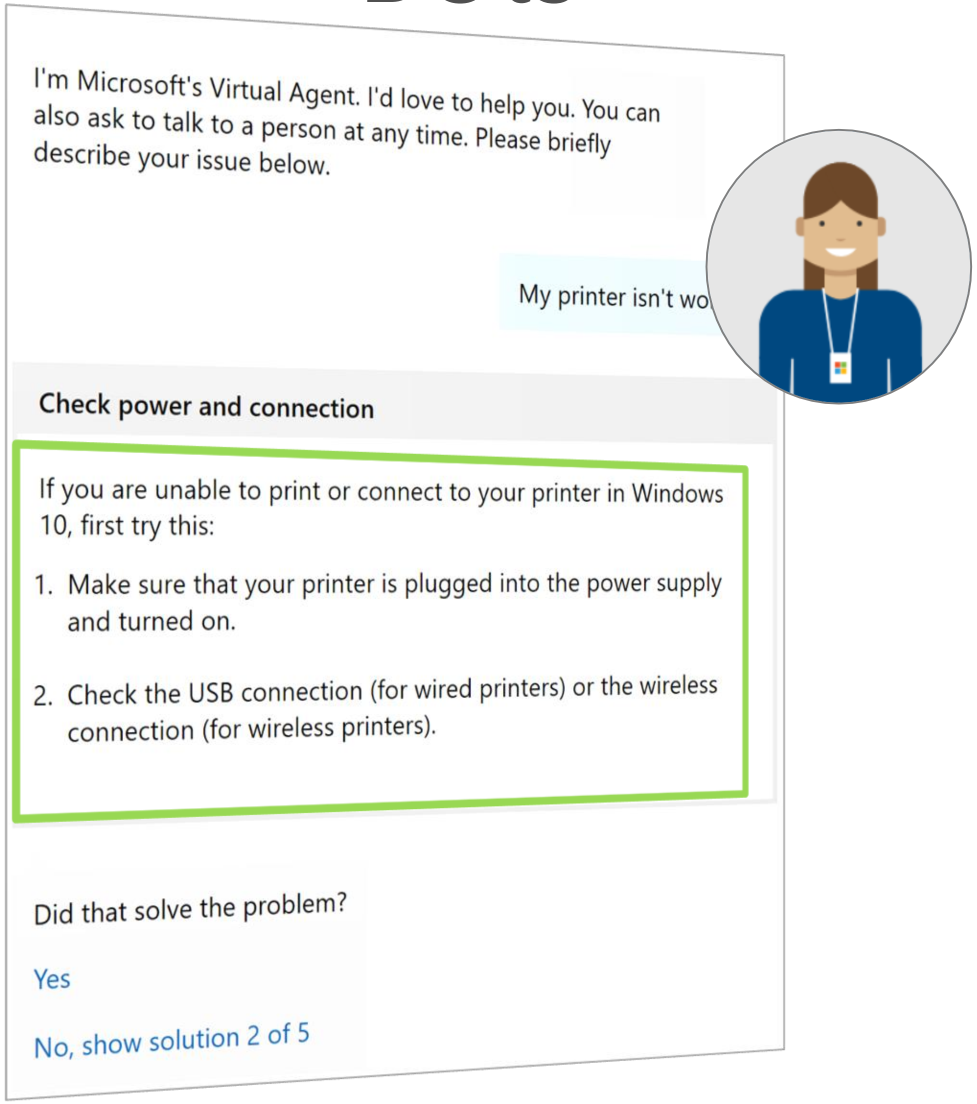



Wellness



[ZKZ ’09, SLLSPM ’11, NSTWCSM ’14, PGCRRH ’14, NHS ’15, KHSBATM ’15, HFKMTY ’16]

Bots



Other Real-world Applications

Ad Choice: [BPQCCPRSS '12]

Ad Format: [TRSA '13]

Education: [MLLBP '14]

Music Rec: [WWHW '14]

Robotics: [PG '16]

Formalism: Contextual Bandits

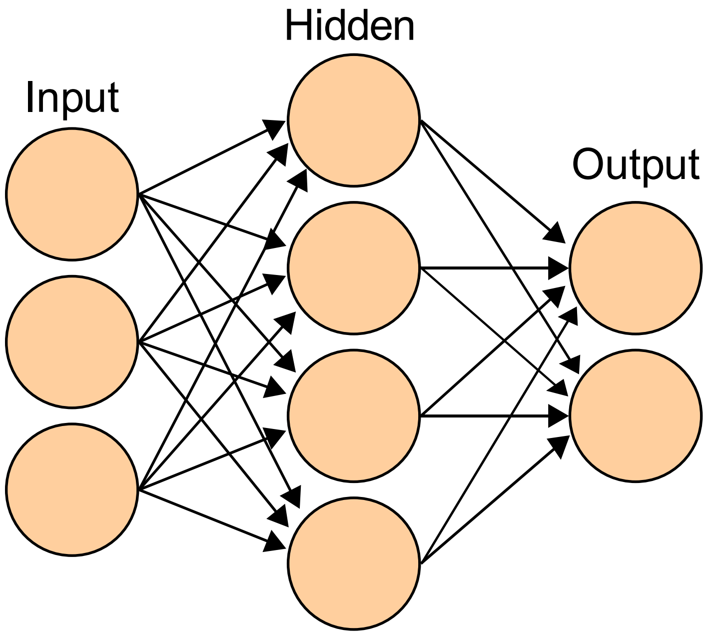
Repeatedly:

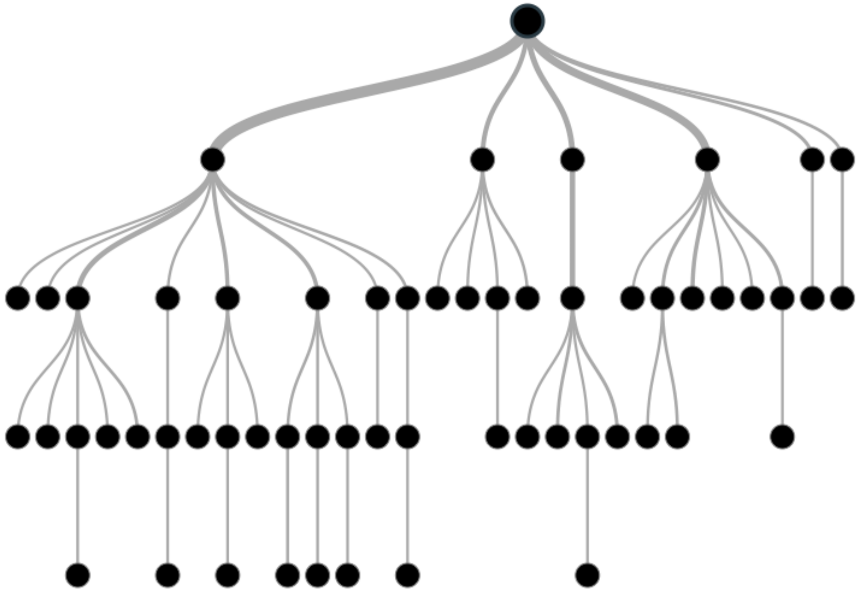
1. Observe features x
2. Choose action $a \in A$
3. Observe reward r

Goal: Maximize expected reward

Policies

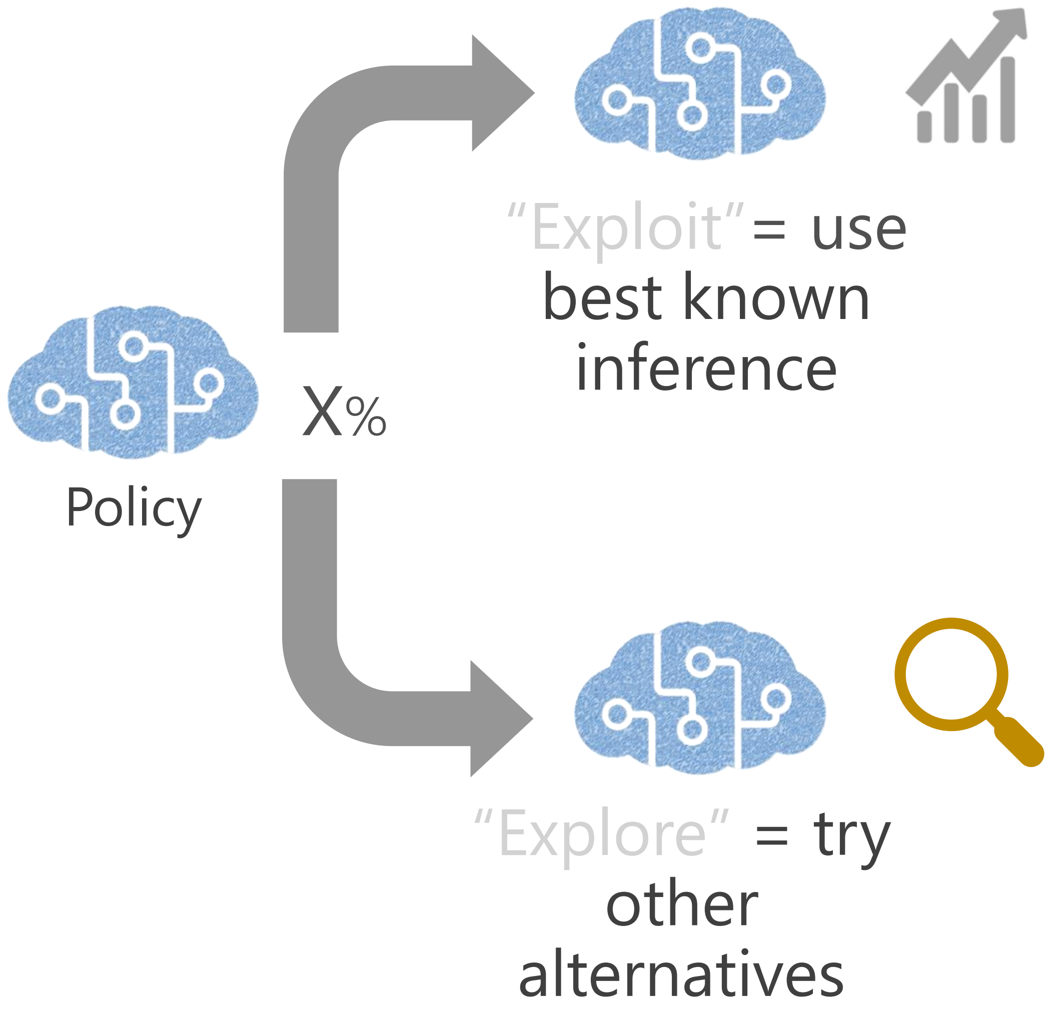
Policy maps features to actions.





Policy = Classifier that ***acts***.

Why does it work?

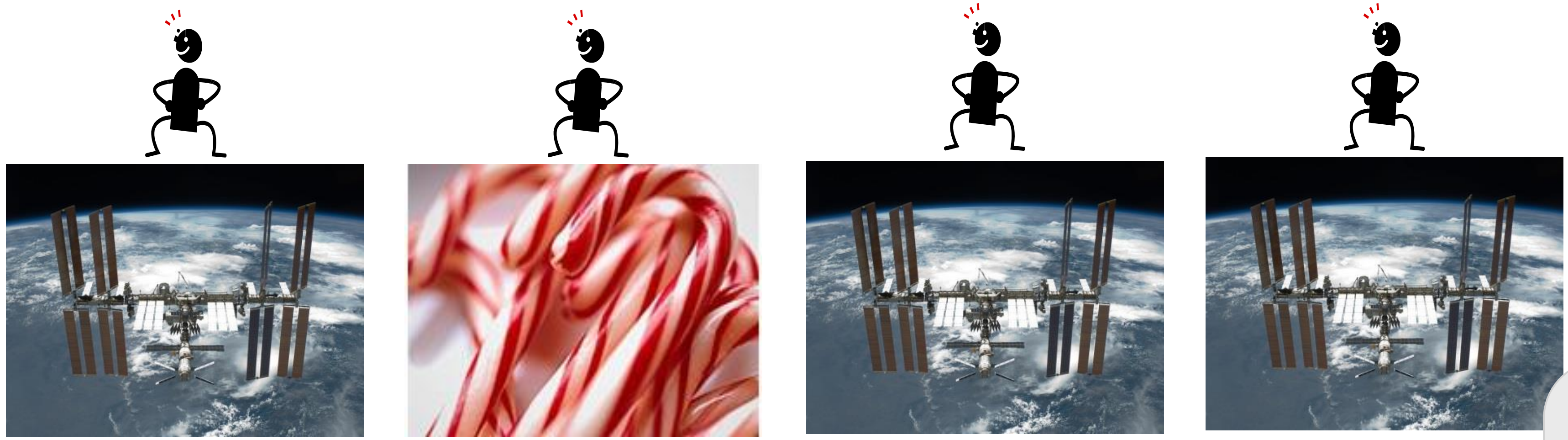


Exploit for performance

"How much should I explore to discover how to best perform?"

Explore to discover new things

Counterfactual Evaluation



Read | Read | Ignored | Read | …

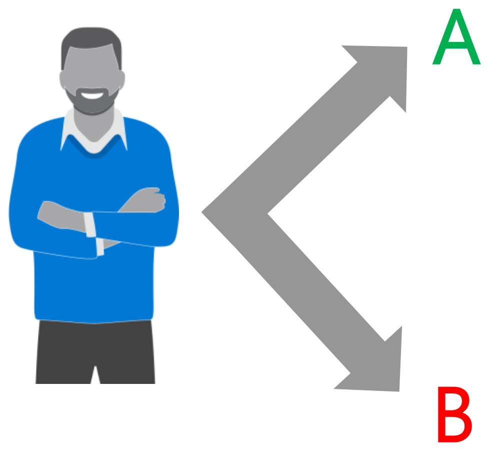
Later evaluate Career rule:

Later evaluate Location rule:

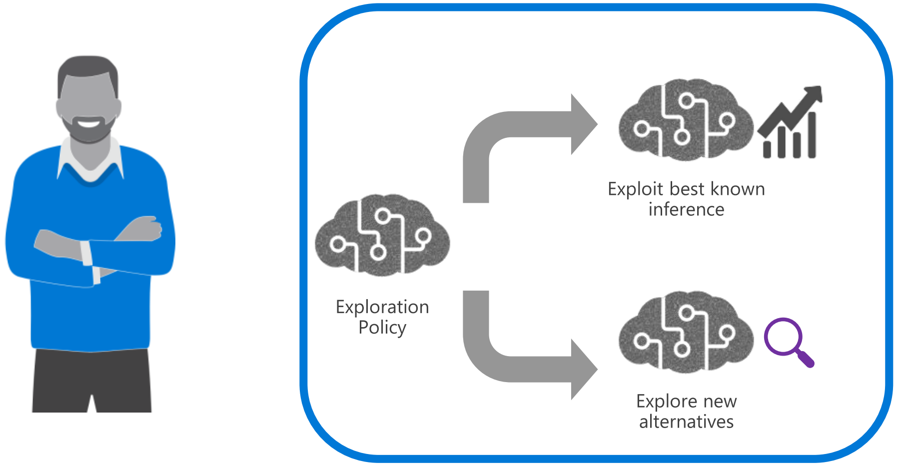
Teacher	Engineer	Engineer	Engineer
Texas	Seattle	Seattle	Texas

Tests can use the same events!

A/B vs. Counterfactual



x1



x100,000

A/B Test:

1. Design the Right Experiment,
2. Test online once
3. Start over

Offline Experiment:

1. Use models that exploit and explore
2. Record User Interaction
3. Find the policy and model that fits reality

Inverse Propensity Score(IPS)

Given experience $\{(x, a, p, r)\}$ and a policy $\pi: x \rightarrow a$, how good is π?

$$V_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{(x,a,p,r)} \frac{r I(\pi(x) = a)}{p}$$

Propensity Score

What do we know about IPS?

Theorem: For all π, for all $D(x,\vec{r})$

$$\mathrm{E}\left[r_{\pi(x)}\right] = \mathrm{E}[V_{\mathrm{IPS}}(\pi)] = \mathrm{E}\left[\frac{1}{n}\sum_{(x,a,p,r)}\frac{rI(\pi(x)=a)}{p}\right]$$

Proof: For all $(x,\vec{r})$, $E_{a\sim\vec{p}}\left[\frac{r_a I(\pi(x)=a)}{p_a}\right]$

$$= \sum_a p_a \frac{r_a I(\pi(x)=a)}{p_a}$$

$$= r_{\pi(x)}$$

Better Evaluation Techniques

Double Robust: [DLL ’11]

Weighted IPS: [K ’92, SJ ’15]

Clipping: [BL ’08]

Empirical Likelihood: [MKL ’19]

Learning from Exploration

Given Data $\{(x, a, p, r)\}$ how to maximize $\mathbf{E}[r_{\pi(x)}]$?

Maximize $\mathbf{E}\left[V_{\text{IPS}}(\pi)\right]$ instead!

$$r_a = \begin{cases} r/p & \text{if } \pi(x) = a \\ 0 & \text{otherwise} \end{cases}$$

Equivalent to:

$$r'_a = \begin{cases} 1 & \text{if } \pi(x) = a \\ 0 & \text{otherwise} \end{cases}$$

with importance weight $\frac{\boldsymbol{r}}{\boldsymbol{p}}$

Importance weighted multiclass classification!

Better Learning from Exploration

Policy Gradient: [W ‘92]

Offset Tree: [BL ’09]

Double Robust for learning: [DLL ‘11]

Multitask Regression: [BAL ‘18]

Weighted IPS for learning: [SJ ‘15]

Evaluating Online Learning

Problem: How do you evaluate an online learning algorithm Offline?

Answer: Use Progressive Validation [BKL '99, CCG '04]

Theorem:

1) Expected PV value = Uniform expected policy value.
2) Trust like a test set error.

How do you do Exploration?

Simplest Algorithm: ϵ-greedy.

With probability ϵ act uniform random

With probability $1 - \epsilon$ act greedily

Better Exploration Algorithms

Better algorithms maintain ensemble and explore amongst actions of this ensemble.

Thompson Sampling: [T '33]

EXP4: [ACFS '02]

Epoch Greedy: [LZ '07]

Polytime: [DHKKLRZ '11]

Cover&Bag: [AHKLLS '14]

Bootstrap: [EK '14]

More Details!

Personalizer Service: http://aka.ms/personalizer

Vowpal Wabbit: http://vowpalwabbit.org

ICML tutorial: http://hunch.net/~rwil

We are hiring: http://aka.ms/rl_hiring