Real World Reinforcement Learning

John Langford

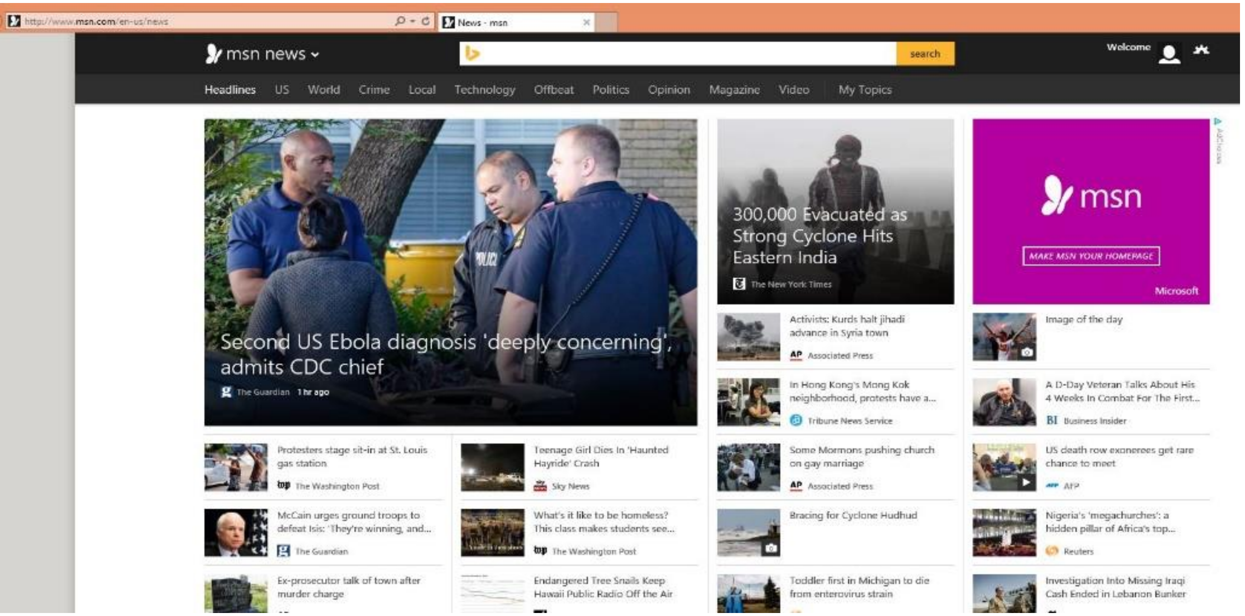
AI Nextcon, NYC, July 23

+Workshop July 25, 9-12

https://vowpalwabbit.github.io/workshop/

Why?

Which News?



Second US Ebola diagnosis 'deeply concerning', admits CDC chief
The Guardian 1 hr ago

300,000 Evacuated as Strong Cyclone Hits Eastern India
The New York Times

Protesters stage sit-in at St. Louis gas station
The Washington Post

Teenage Girl Dies In 'Haunted Hayride' Crash
Sky News

McCain urges ground troops to defeat Isis: 'They're winning, and...
The Guardian

What's it like to be homeless? This class makes students see...
The Washington Post

Ex-prosecutor talk of town after murder charge

Endangered Tree Snails Keep Hawaii Public Radio Off the Air

Activists: Kurds halt jihadi advance in Syria town
AP Associated Press

In Hong Kong's Mong Kok neighborhood, protests have a...
Tribune News Service

Some Mormons pushing church on gay marriage
AP Associated Press

Bracing for Cyclone Hudhud

Toddler first in Michigan to die from enterovirus strain

Image of the day

A D-Day Veteran Talks About His 4 Weeks In Combat For The First...
BI Business Insider

US death row exonerees get rare chance to meet
AFP

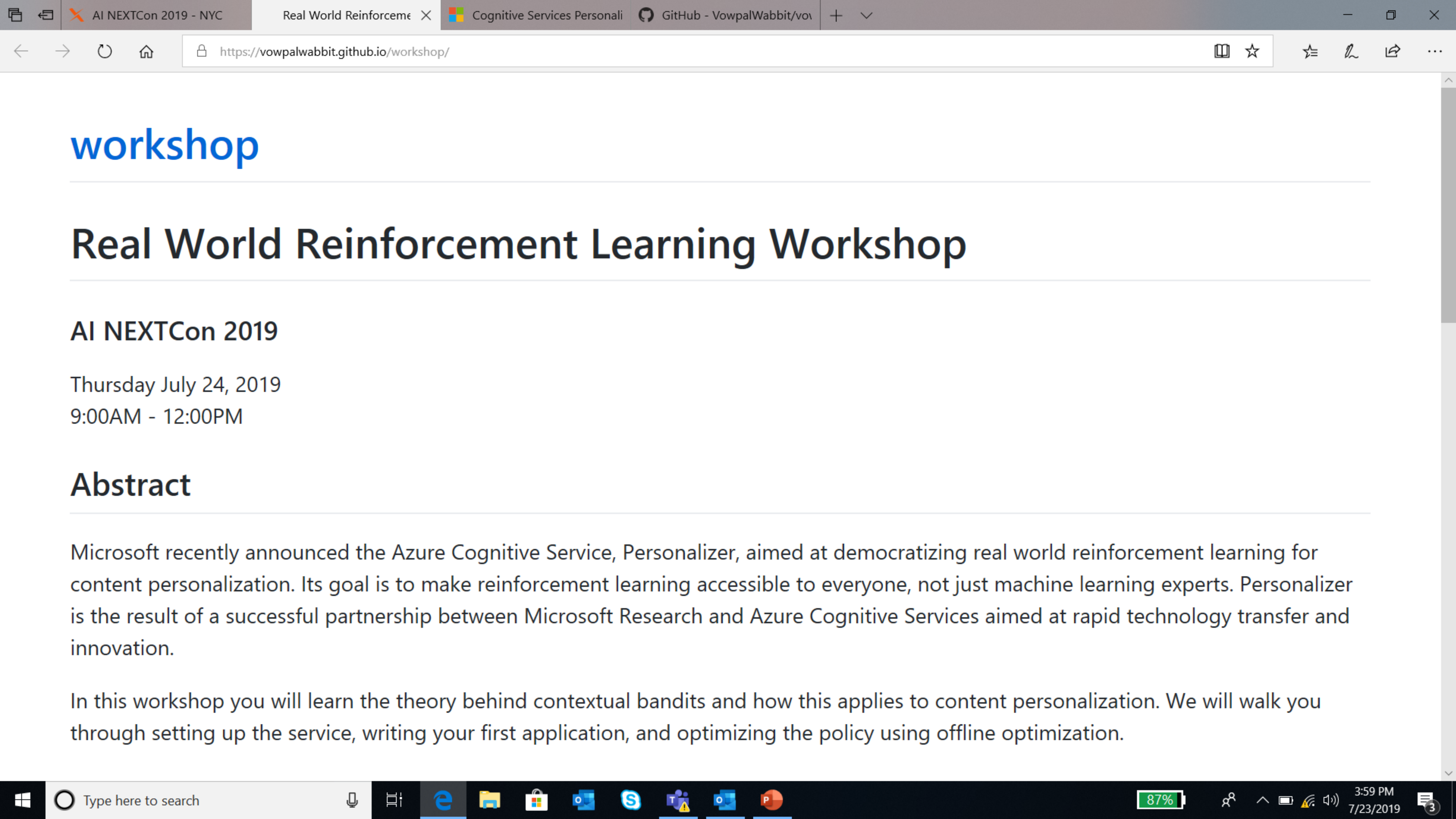
Nigeria's 'megachurches': a hidden pillar of Africa's top...
Reuters

Investigation Into Missing Iraqi Cash Ended in Lebanon Bunker

28% lift

Which Game?

40% lift



workshop

Real World Reinforcement Learning Workshop

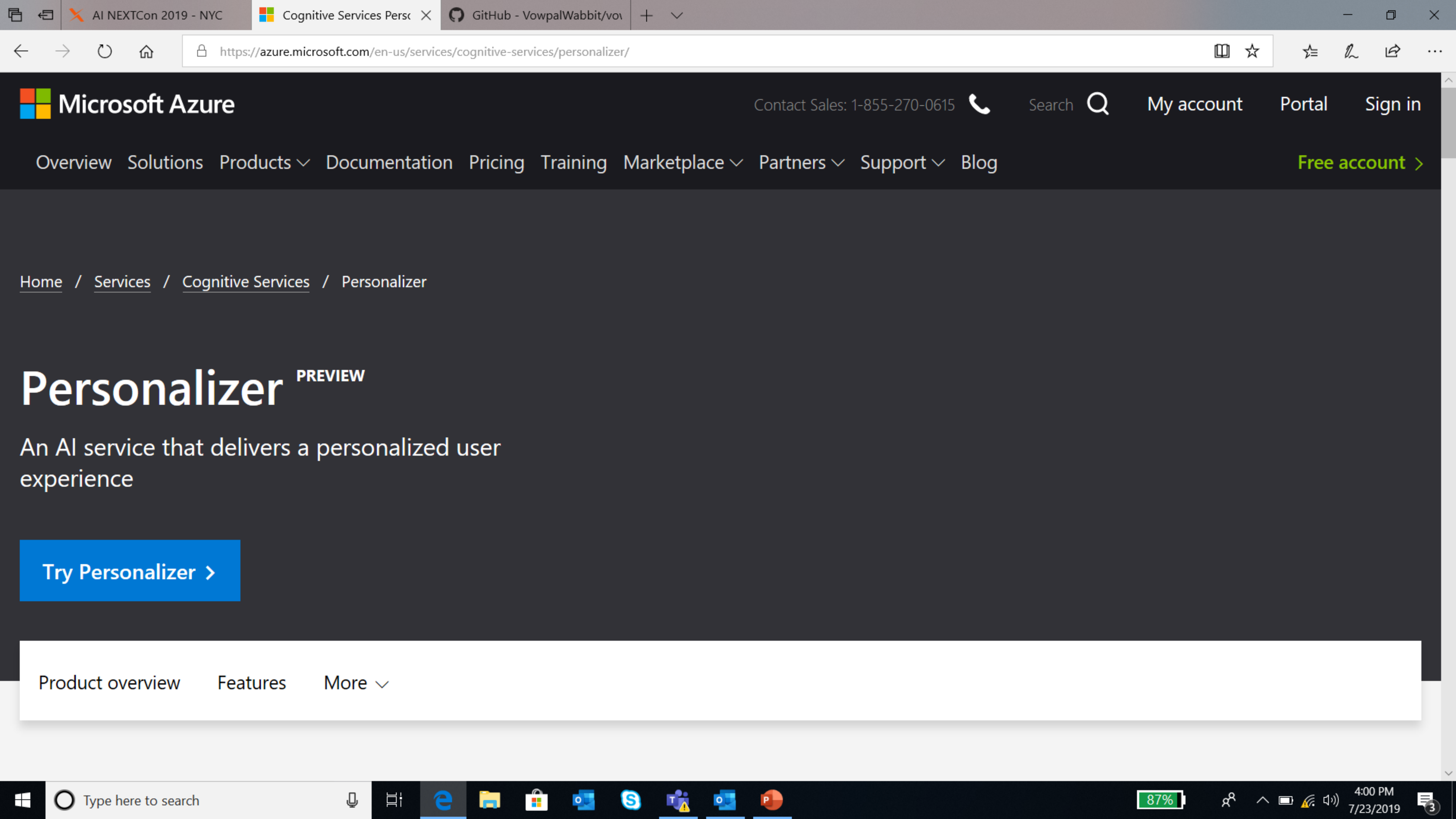
AI NEXTCon 2019

Thursday July 24, 2019
9:00AM - 12:00PM

Abstract

Microsoft recently announced the Azure Cognitive Service, Personalizer, aimed at democratizing real world reinforcement learning for content personalization. Its goal is to make reinforcement learning accessible to everyone, not just machine learning experts. Personalizer is the result of a successful partnership between Microsoft Research and Azure Cognitive Services aimed at rapid technology transfer and innovation.

In this workshop you will learn the theory behind contextual bandits and how this applies to content personalization. We will walk you through setting up the service, writing your first application, and optimizing the policy using offline optimization.



AI NEXTCon 2019 - NYC | Cognitive Services Persc | GitHub - VowpalWabbit/vo

https://azure.microsoft.com/en-us/services/cognitive-services/personalizer/

Microsoft Azure

Contact Sales: 1-855-270-0615 Search My account Portal Sign in

Overview Solutions Products Documentation Pricing Training Marketplace Partners Support Blog Free account

Home / Services / Cognitive Services / Personalizer

Personalizer PREVIEW

An AI service that delivers a personalized user experience

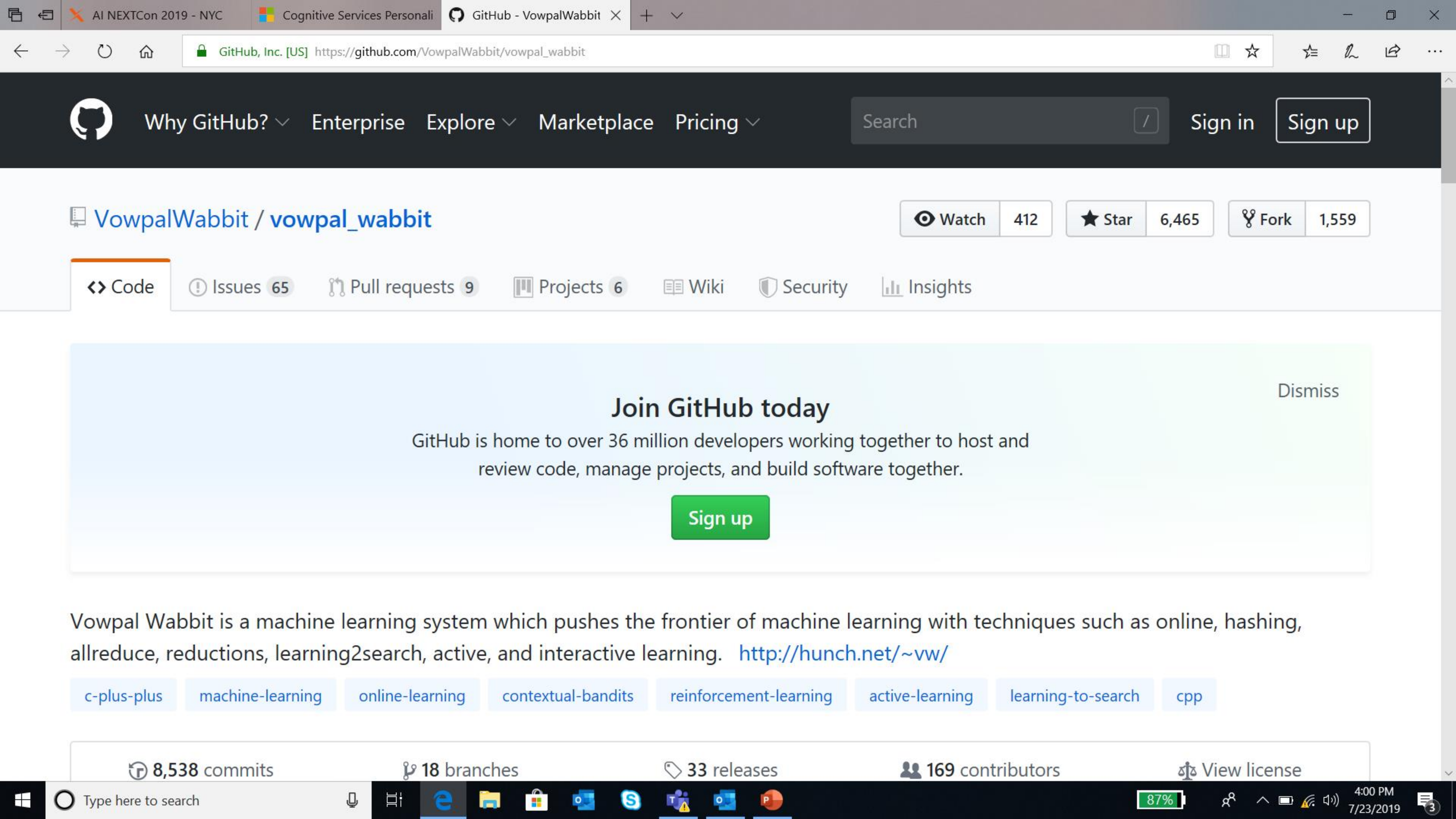
Try Personalizer

Product overview Features More

Type here to search

87%

4:00 PM
7/23/2019

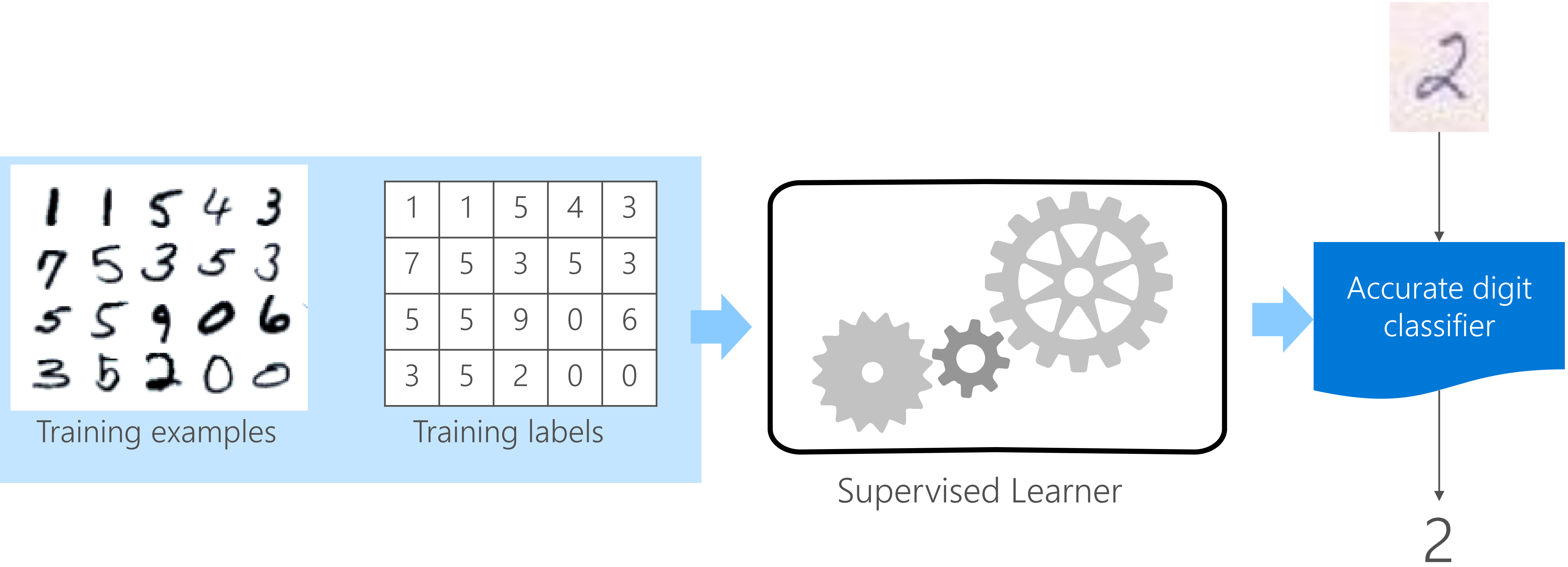


Join GitHub today

GitHub is home to over 36 million developers working together to host and review code, manage projects, and build software together.

Vowpal Wabbit is a machine learning system which pushes the frontier of machine learning with techniques such as online, hashing, allreduce, reductions, learning2search, active, and interactive learning. http://hunch.net/~vw/

The Baseline: Supervised Learning



1	1	5	4	3
7	5	3	5	3
5	5	9	0	6
3	5	2	0	0

Training examples

Training labels

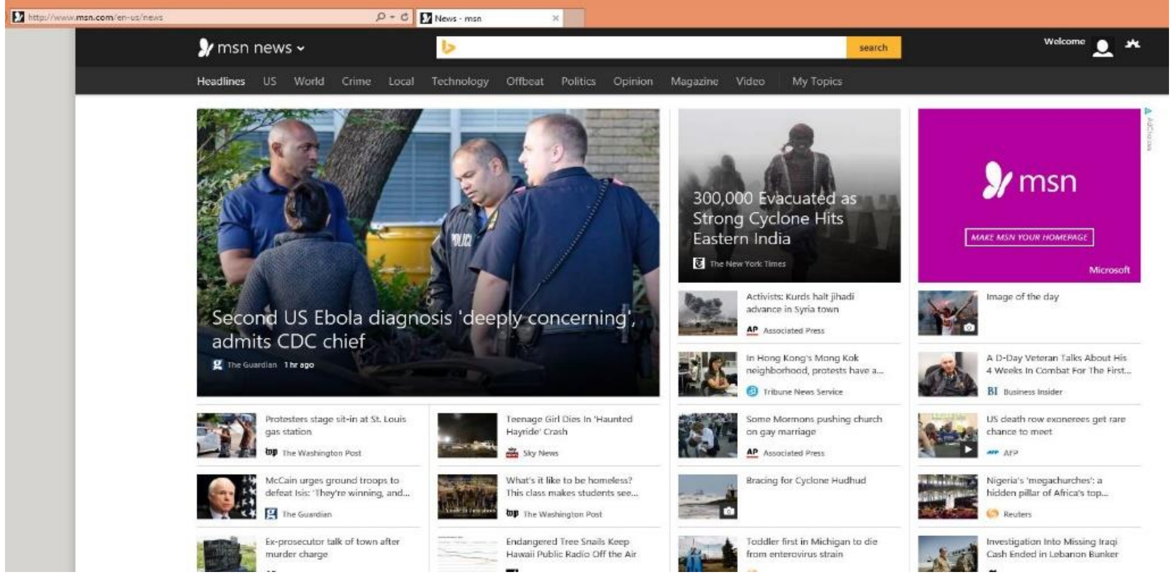
Supervised Learner

Accurate digit classifier

2

Supervised Learning is cool

How about news?



Second US Ebola diagnosis 'deeply concerning', admits CDC chief
The Guardian 1 hr ago

300,000 Evacuated as Strong Cyclone Hits Eastern India
The New York Times

Activists: Kurds halt jihadi advance in Syria town
Associated Press

Image of the day

In Hong Kong's Mong Kok neighborhood, protests have a...
Tribune News Service

A D-Day Veteran Talks About His 4 Weeks In Combat For The First...
Business Insider

Protesters stage sit-in at St. Louis gas station
The Washington Post

Teenage Girl Dies In 'Haunted Hayride' Crash
Sky News

Some Mormons pushing church on gay marriage
Associated Press

US death row exonerees get rare chance to meet
AFP

McCain urges ground troops to defeat Isis: 'They're winning, and...
The Guardian

What's it like to be homeless? This class makes students see...
The Washington Post

Bracing for Cyclone Hudhud

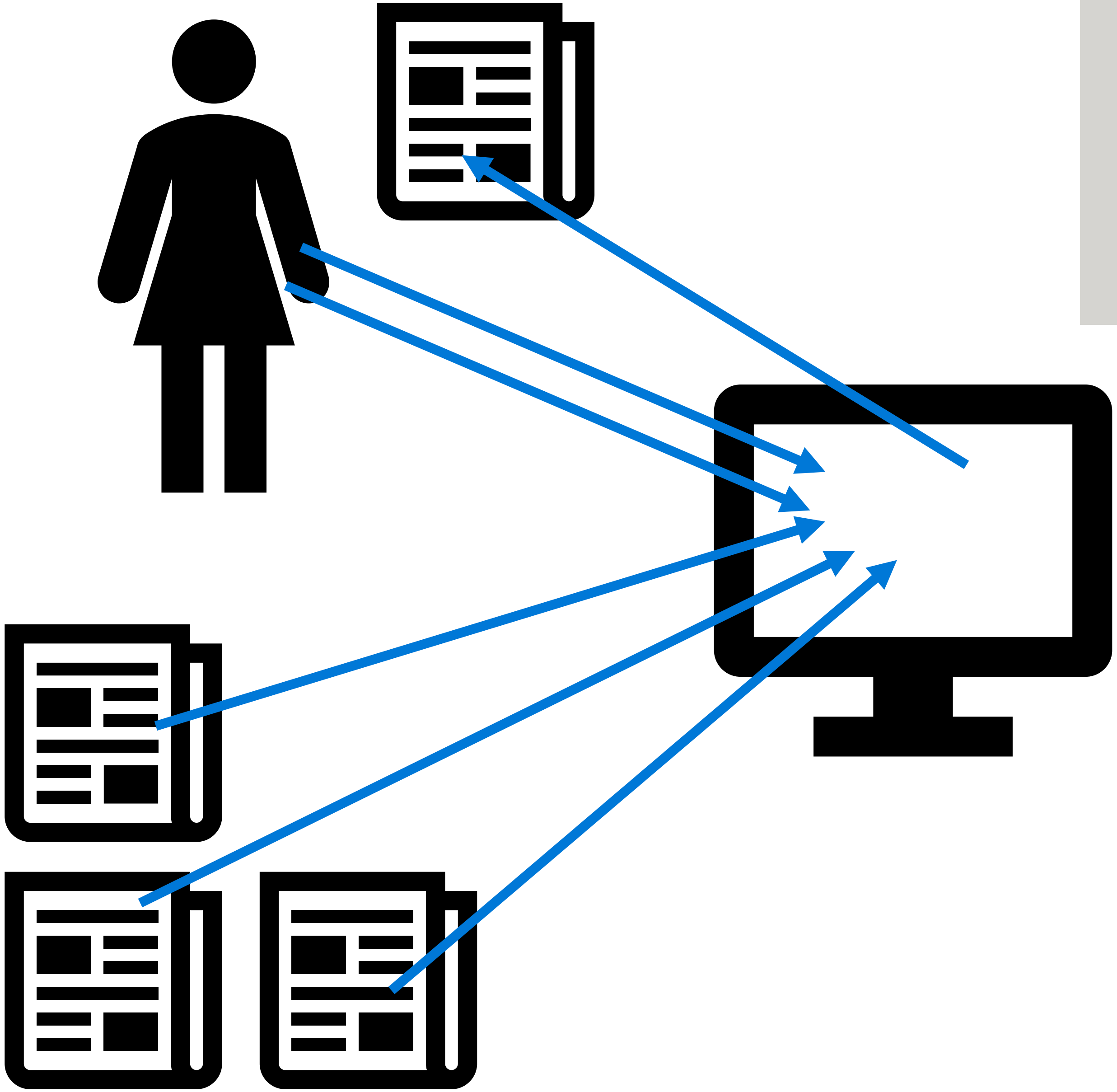
Nigeria's 'megachurches': a hidden pillar of Africa's top...
Reuters

Ex-prosecutor talk of town after murder charge

Endangered Tree Snails Keep Hawaii Public Radio Off the Air

Toddler first in Michigan to die from enterovirus strain

Investigation Into Missing Iraqi Cash Ended in Lebanon Bunker



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 ☹

A standard pipeline

1. Collect $(user, article, click)$ information.
2. Learn $\hat{P}(click|user, article)$
3. Act with best $article$ from $\hat{P}(click|user, article)$
4. Deploy in A/B test for 2 weeks
5. A/B test fails ☹ Why?

Q: What goes wrong?

Is Ukraine

interesting to John ?

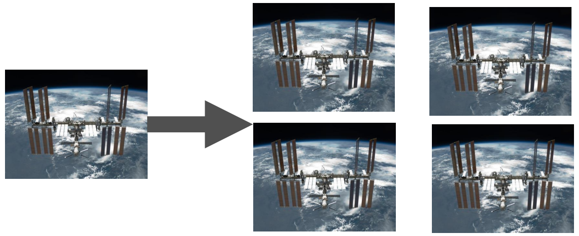
A: Need Right Signal for Right Answer

What goes wrong?

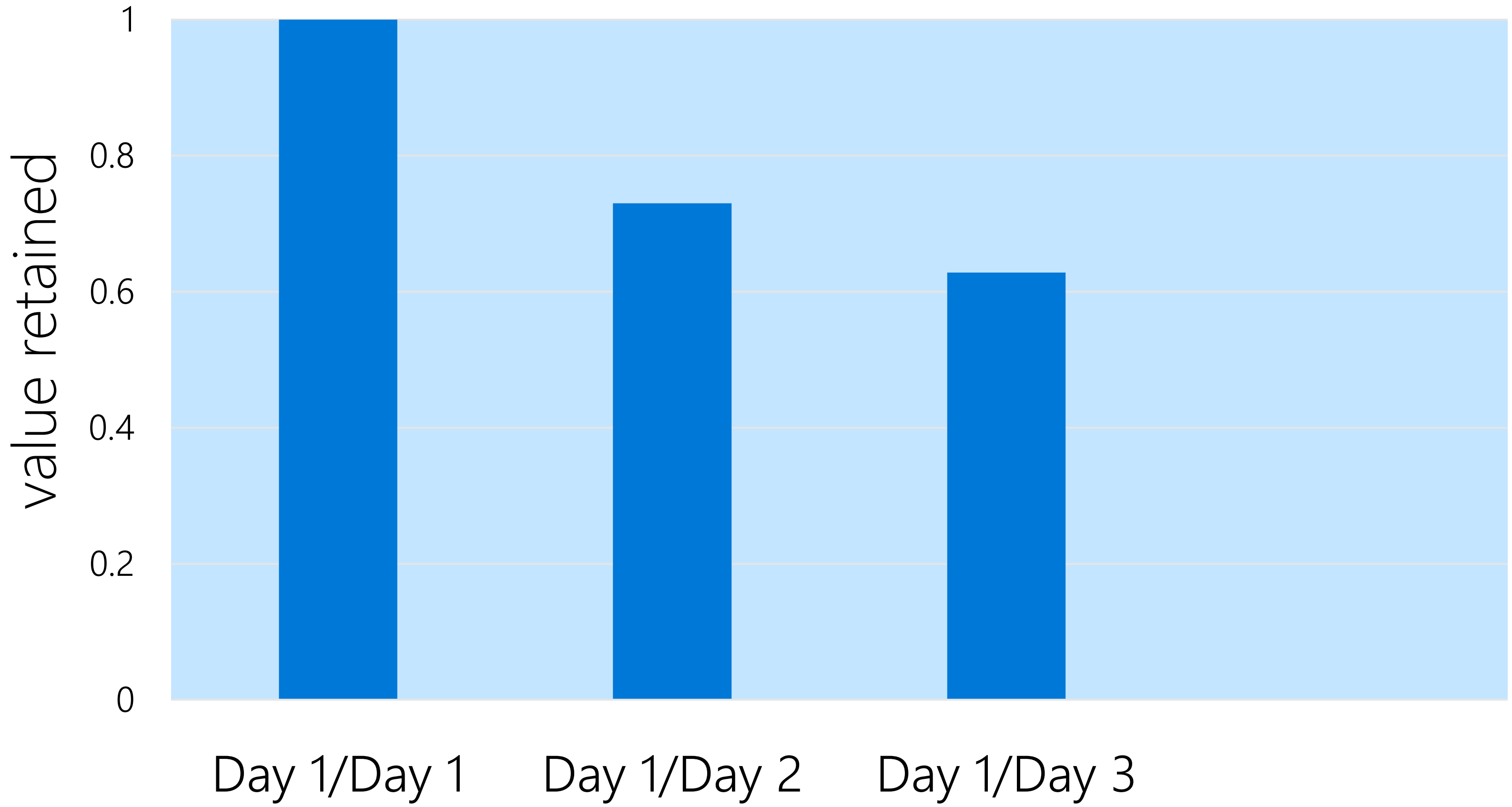
What is the probability of click on a food article

If you only display a space article?

We must avoid “self-fulfilling prophecy”



What else goes wrong?



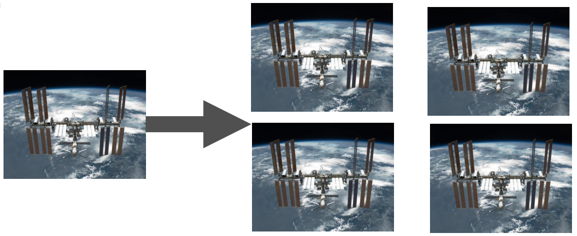
The world changes!

Can we optimize for best outcome?

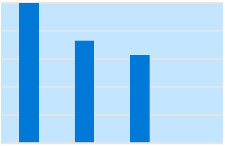
Amongst a given set of choices?

For what matters to individuals?

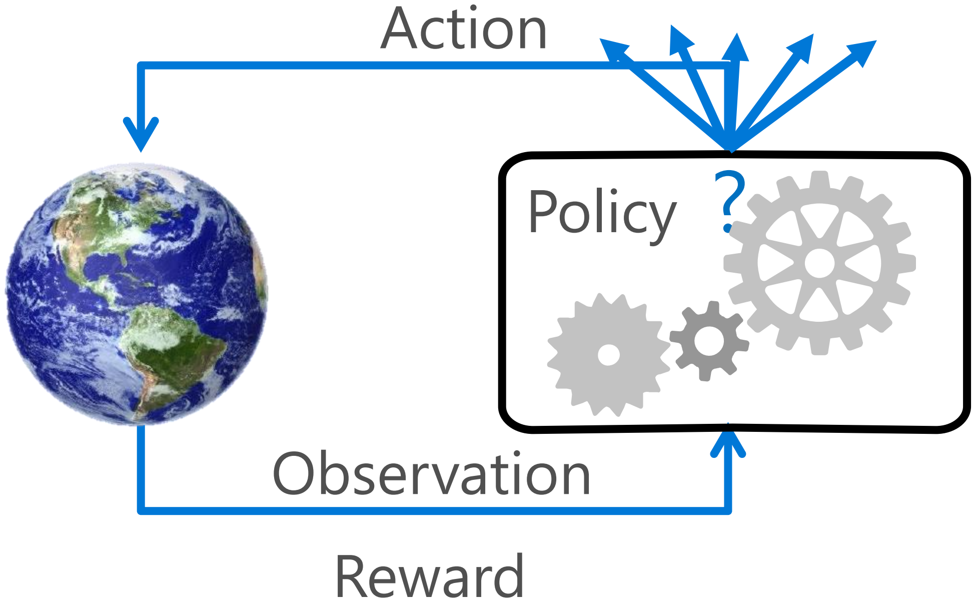
Without self-fulfilling prophecies?



With real-time learning?

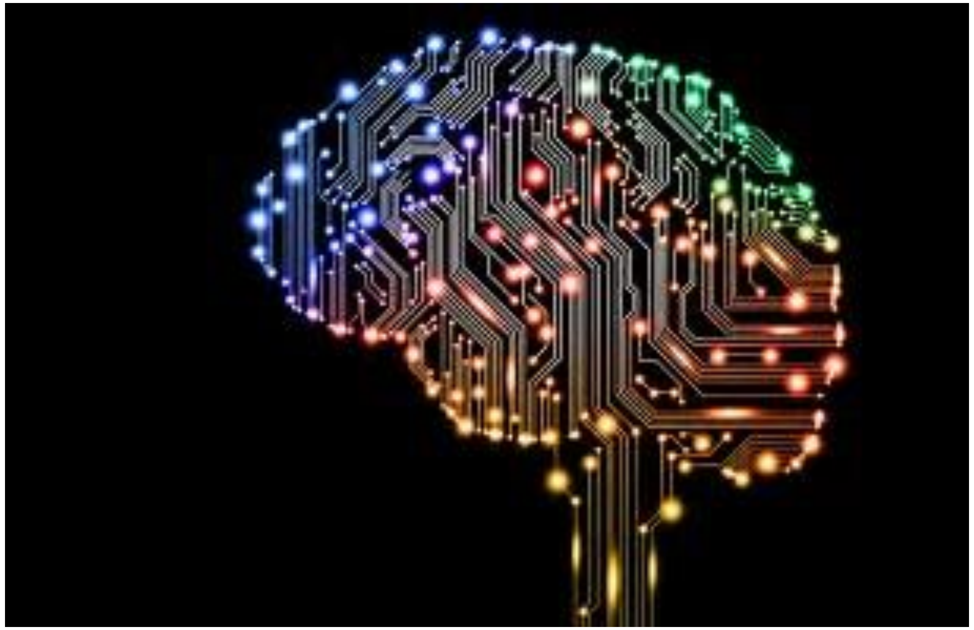


Reinforcement Learning can do this!



Goal: Find a policy maximizing the sum of rewards

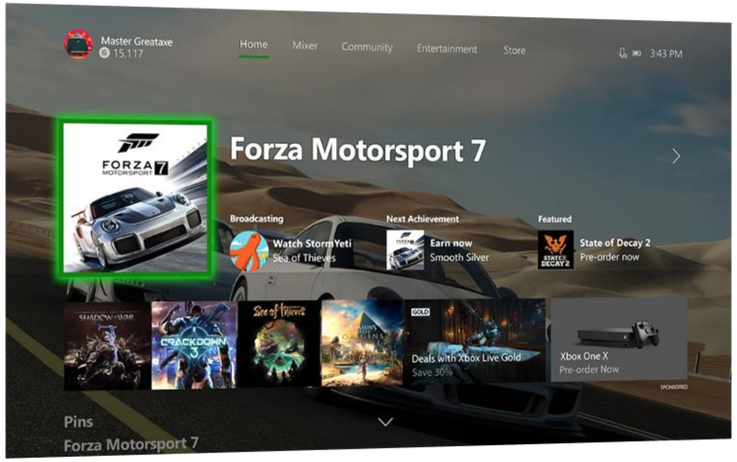
Q: One last Why…



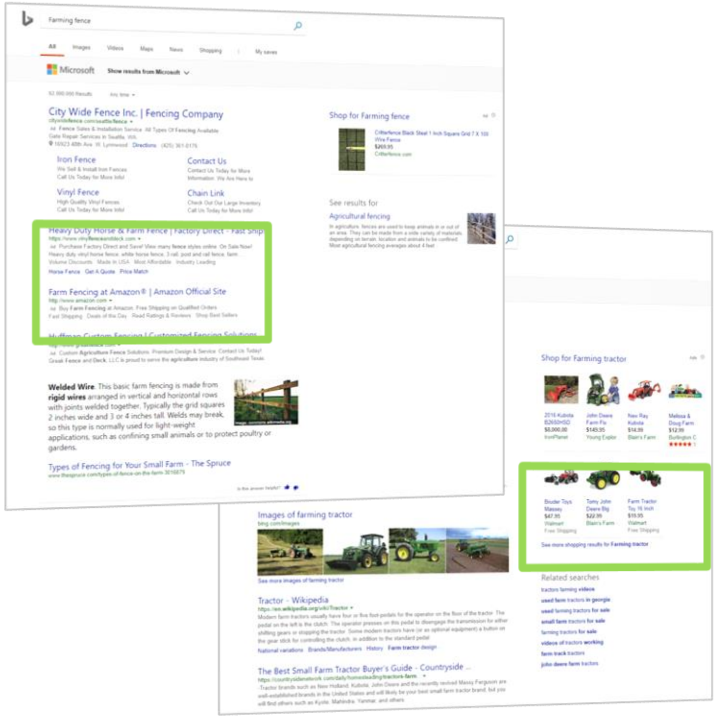
~~AI: A function programmed with data~~

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

Content



Layout


Farming fence
All
Images
Videos
Maps
News
Shopping
My saves
Microsoft
Show results from Microsoft
City Wide Fence Inc. | Fencing Company
Iron Fence
Contact Us
Vinyl Fence
Chain Link
Heavy Duty Horse & Farm Fence | Factory Direct - Fast Ship
Horse Fence
Get A Quote
Price Match
Farm Fencing at Amazon® | Amazon Official Site
Huffman Custom Fencing | Customized Fencing Solutions
Shop for Farming fence
See results for
Agricultural fencing
Types of Fencing for Your Small Farm - The Spruce
Shop for Farming tractor
See more shopping results for Farming tractor
Images of farming tractor
See more images of farming tractor
Tractor - Wikipedia
National variations
Brands/Manufacturers
History
Farm tractor design
The Best Small Farm Tractor Buyer's Guide - Countryside ...
Related searches
tractors farming videos
used farm tractors in georgia
used farming tractors for sale
small farm tractors for sale
farming tractors for sale
videos of tractors working
farm track tractors
john deere farm tractors

Welded Wire. This basic farm fencing is made from **rigid wires** arranged in vertical and horizontal rows with joints welded together. Typically the grid squares 2 inches wide and 3 or 4 inches tall. Welds may break, so this type is normally used for light-weight applications, such as confining small animals or to protect poultry or gardens.

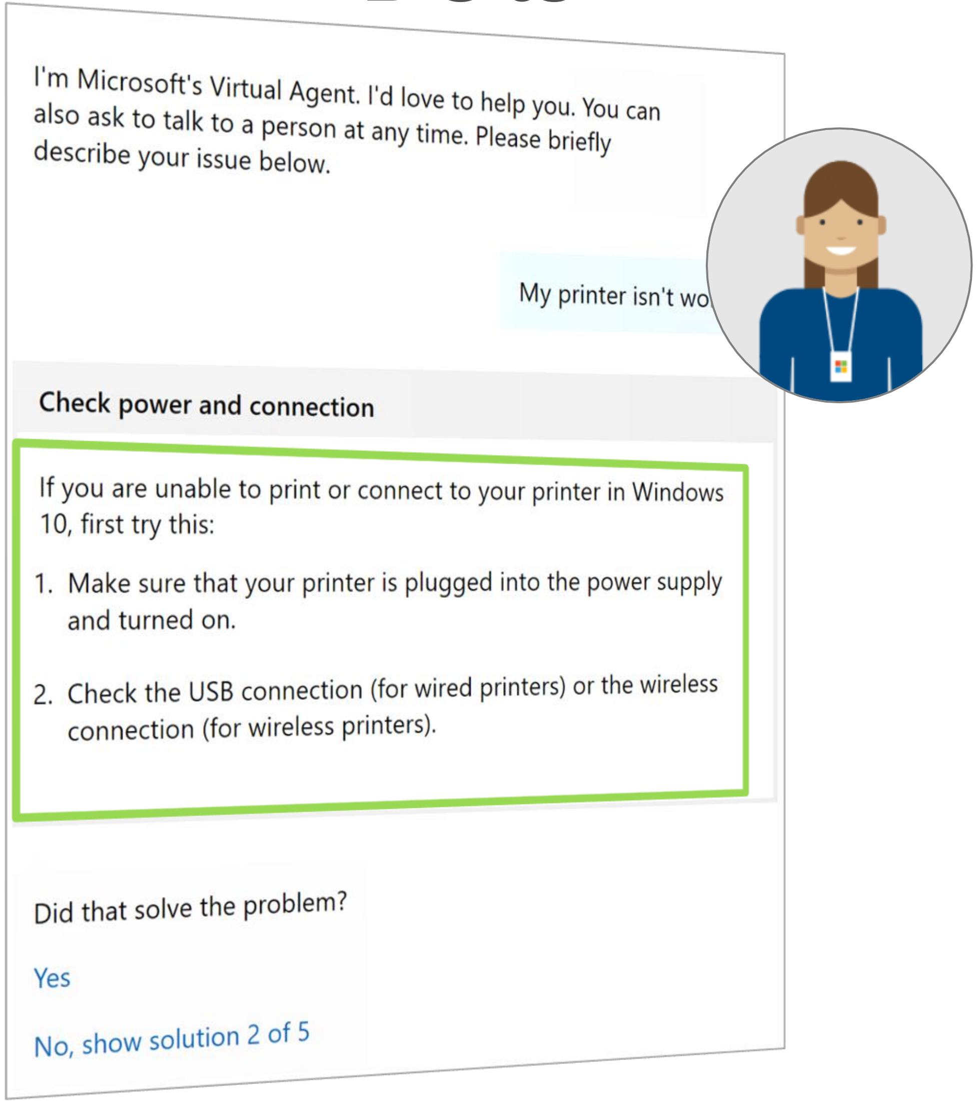
Creative

“Book Your Vacation to Hawaii”

Wellness

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

Bots



I'm Microsoft's Virtual Agent. I'd love to help you. You can also ask to talk to a person at any time. Please briefly describe your issue below.

My printer isn't wo

Check power and connection

If you are unable to print or connect to your printer in Windows 10, first try this:

1. Make sure that your printer is plugged into the power supply and turned on.
2. Check the USB connection (for wired printers) or the wireless connection (for wireless printers).

Did that solve the problem?

Yes

No, show solution 2 of 5

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

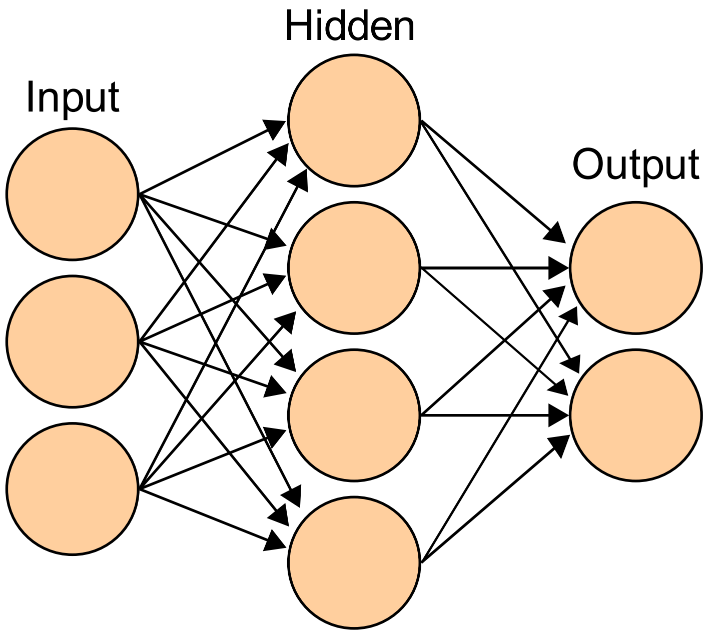
Repeatedly:

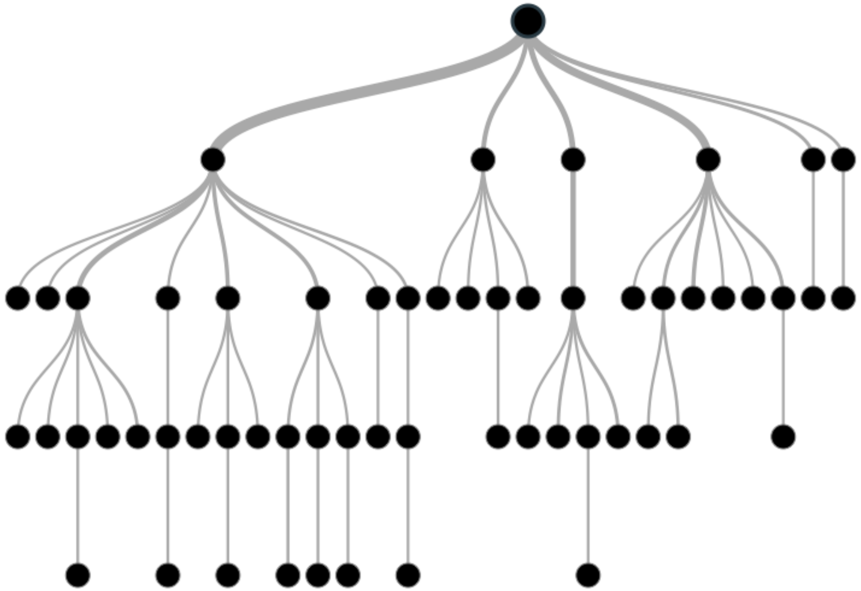
1. Observe features x
2. Choose action $a \in A$
3. Observe reward r

Goal: Maximize expected reward

Policies

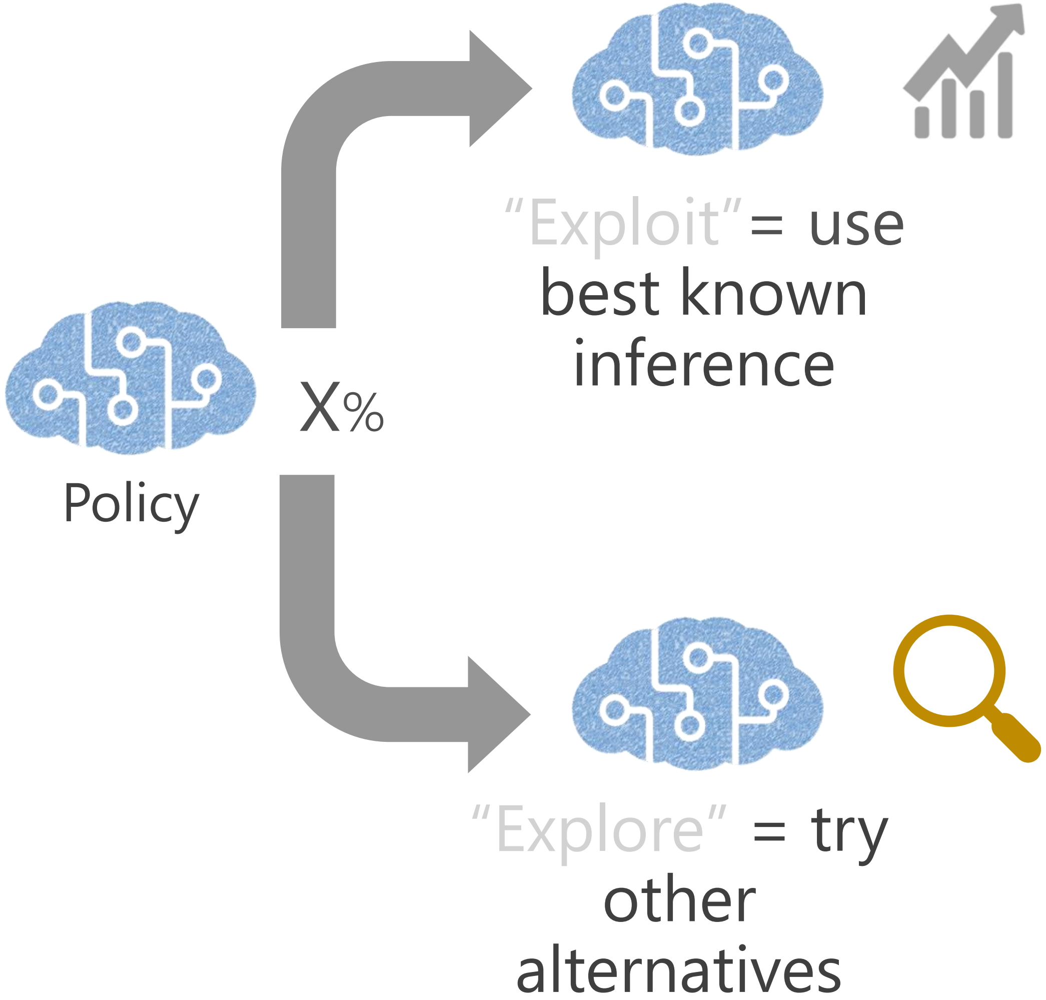
Policy maps features to actions.





Policy = Classifier that *acts*.

Why does it work?



Policy

X%

"Exploit"= use best known inference

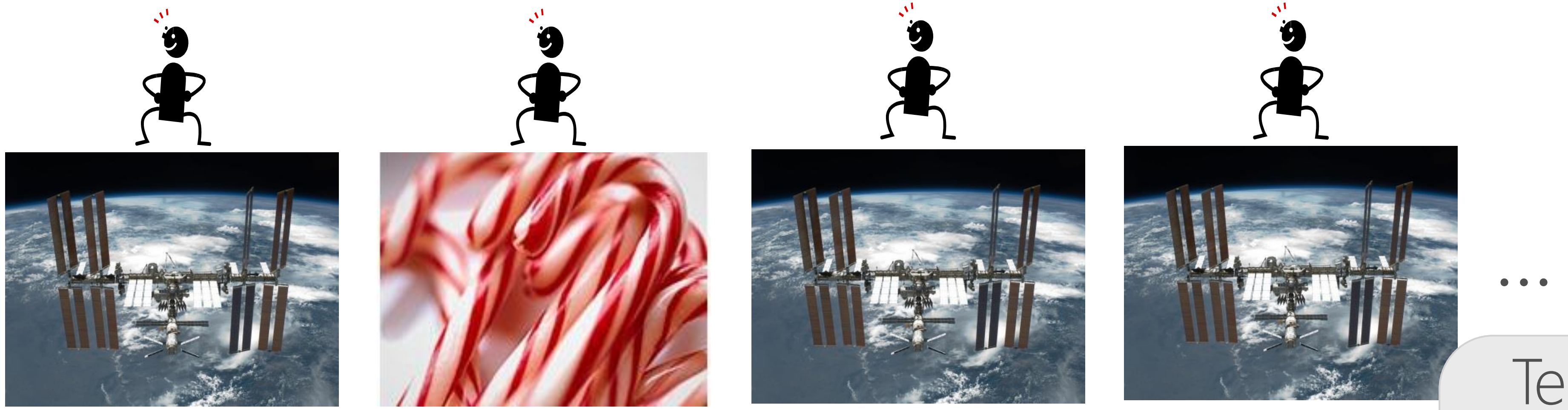
"Explore" = try other alternatives

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 Location rule:

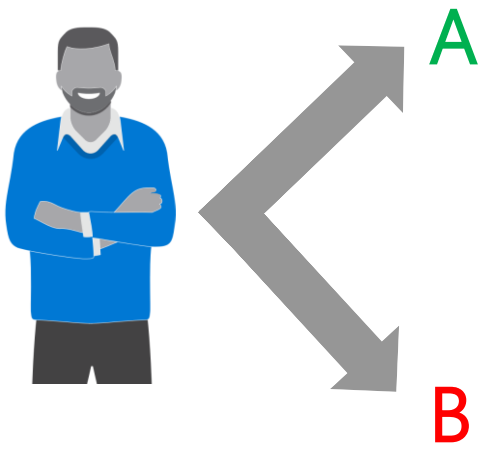
Career rule:

Teacher Engineer Engineer Engineer

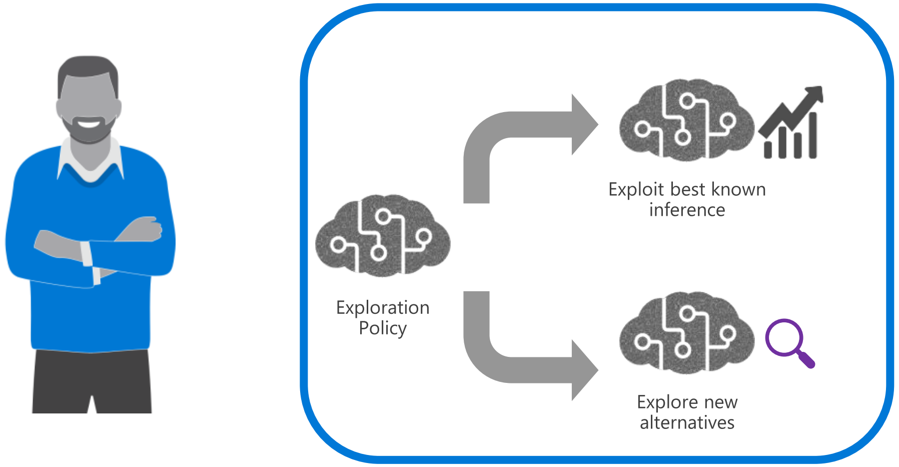
Texas Seattle Seattle Texas

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?

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

Propensity Score

What do we know about IPS?

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

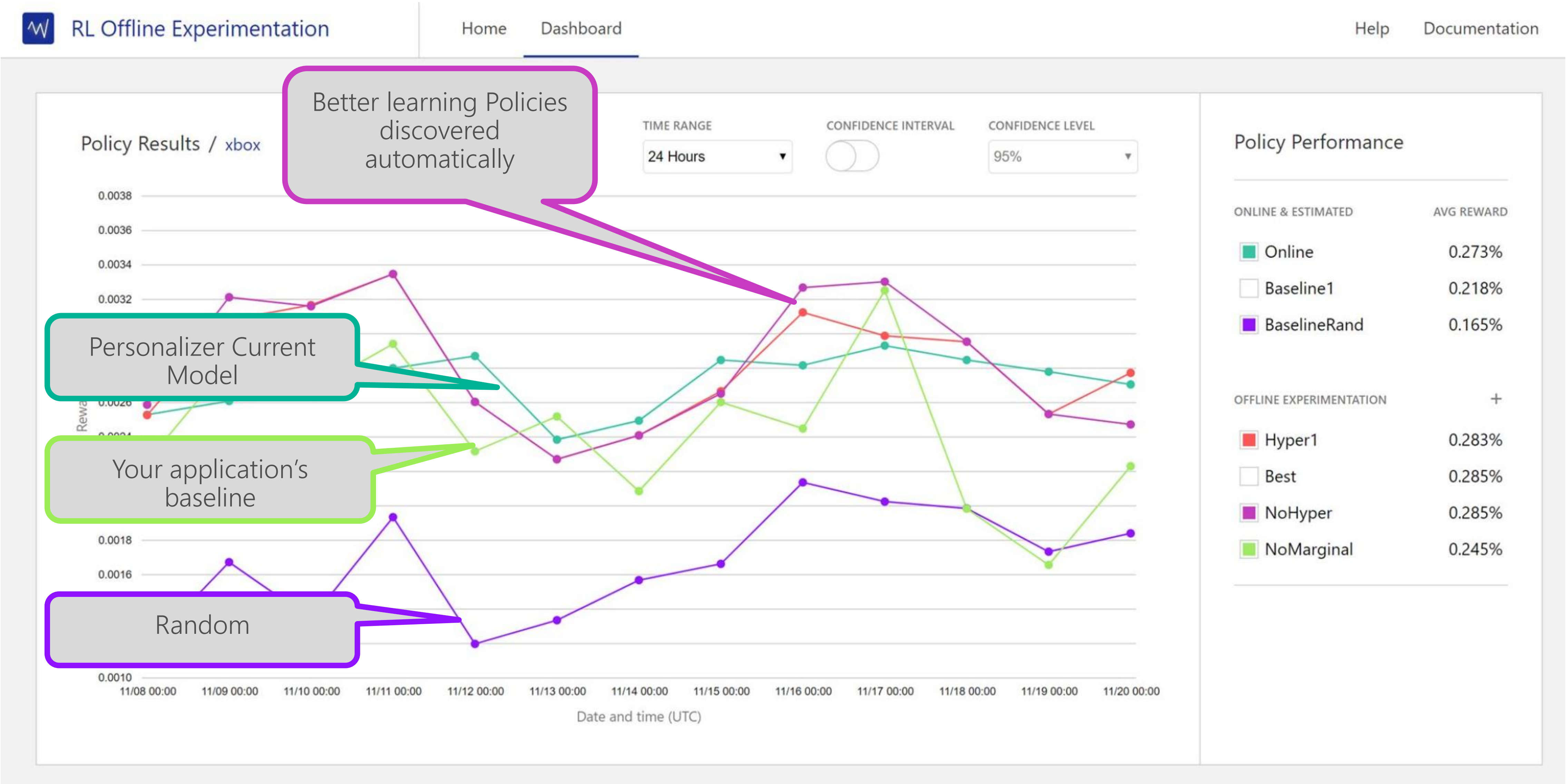
$$\mathrm{E}\left[r_{\pi(x)}\right] = \mathrm{E}[V_{\mathrm{IPS}}(\pi)] = \mathrm{E}\left[\frac{1}{n}\sum_{(x,a,p,r)}\frac{rI(\pi(x)=a)}{p}\right]$$

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

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

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

Why Explore? You can do learning



RL Offline Experimentation

Home | Dashboard | Help | Documentation

Policy Results / xbox

TIME RANGE: 24 Hours ▾

CONFIDENCE INTERVAL: ○

CONFIDENCE LEVEL: 95% ▾

Better learning Policies discovered automatically

Personalizer Current Model

Your application's baseline

Random

Policy Performance

ONLINE & ESTIMATED	AVG REWARD
🟩 Online	0.273%
□ Baseline1	0.218%
🟪 BaselineRand	0.165%

OFFLINE EXPERIMENTATION	+
🟥 Hyper1	0.283%
□ Best	0.285%
🟪 NoHyper	0.285%
🟩 NoMarginal	0.245%

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 $\mathrm{E}[r_{\pi(x)}]$?

Maximize $\mathrm{E}\left[V_{\mathrm{IPS}}(\pi)\right]$ 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]

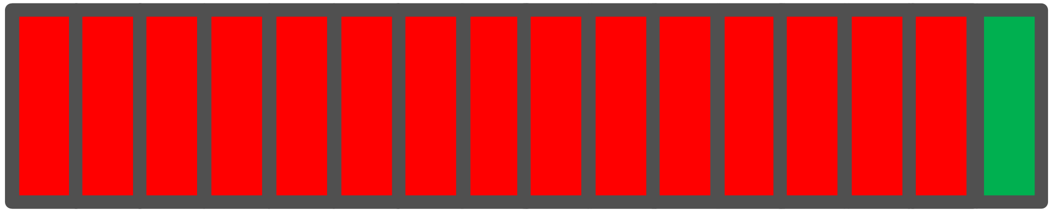
Multitask Regression: [BAL ‘18]

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]

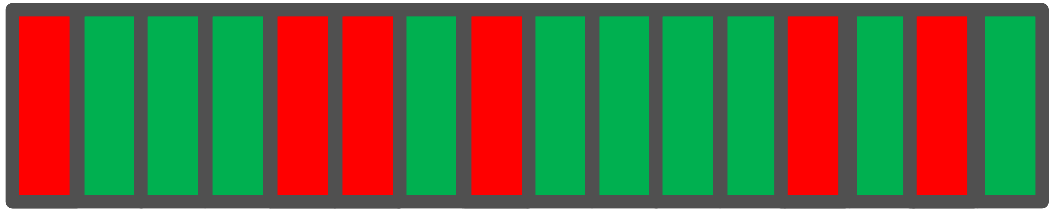
Cover&Bag: [AHKLLS ‘14]

Bootstrap: [EK ‘14]

Evaluating Exploration Algorithms

Problem: How do you take the choice of examples acquired by an exploration algorithm into account?

Answer: Rejection Sample from history. [DELL '12]



Theorem: Realized history is unbiased up to length observed.

Better versions: [DELL '14]

More Details!

AI Nextcon Workshop, July 25 9-12
https://vowpalwabbit.github.io/workshop/

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

Vowpal Wabbit: http://vowpalwabbit.org

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