Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

THE LINUX FOUNDATION OPEN SOURCE SUMMIT NORTH AMERICA

Embedded Linux Conference North America

VIDEO

Q&A WITH SPEAKER SPEAKER BIO

David Aronchick
Head of OSS ML Strategy
Azure/Microsoft

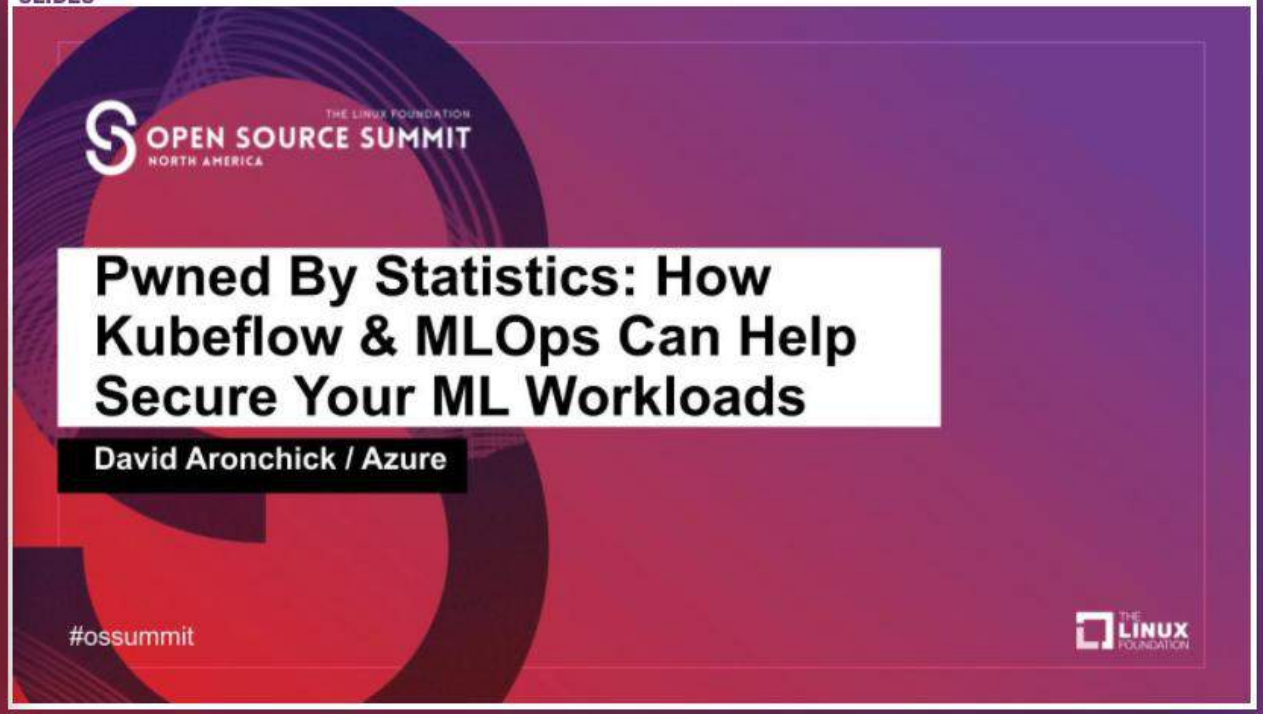
David leads Open Source Machine Learning Strategy at Azure. This means he spends most of his time helping humans to convince machines to be smarter. He is only moderately successful at this. Previously, he led product management for Kubernetes, launched Google Kubernetes Engine and co-founded the Kubeflow project while at Google. David has also worked at Amazon, Chef, and co-founded three startups. David can be found on a mountain (on skis), traveling the world (via restaurants) or participating in kid activities, of which there are a lot more than he remembers than when he was that age.

ABSTRACT

While machine learning is spreading like wildfire, very little attention has been paid to the ways that it can go wrong when moving from development to production. Even when models work perfectly, they can be attacked and/or degrade quickly if the data changes. Having a well understood MLOps process is necessary for ML security!

Using Kubeflow, we will demonstrate how to the common ways machine learning workflows go wrong, and how to mitigate them using MLOps pipelines to provide reproducibility, validation, versioning/tracking, and safe/compliant deployment. We will also talk about the direction for MLOps as an industry, and how we can use it to move faster, with less risk, than ever before.

SLIDES



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

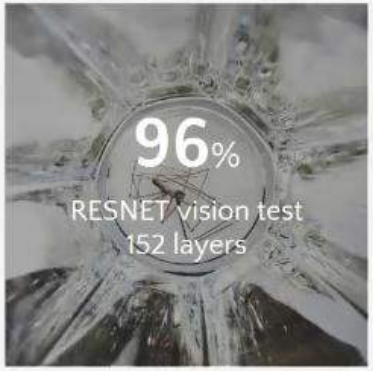
SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0o72a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_4

Microsoft ML breakthroughs

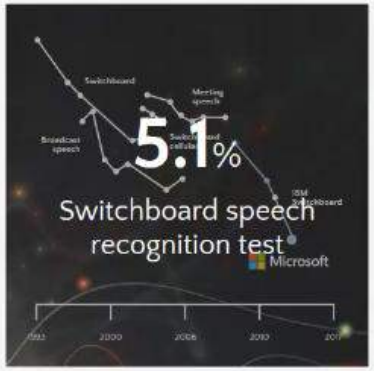
Vision



2016

Object recognition
Human parity

Speech



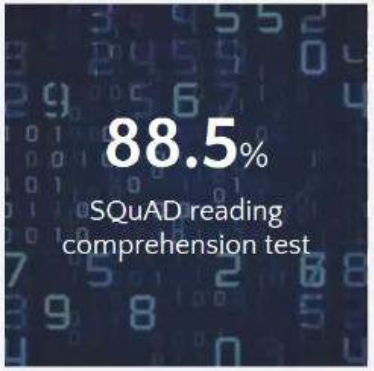
2017

Speech recognition
Human parity

Language

2018

Machine translation
Human parity



2018

Machine reading comprehension
Human parity

Slide 2 | Q&A | Notes | Pointer | Tips

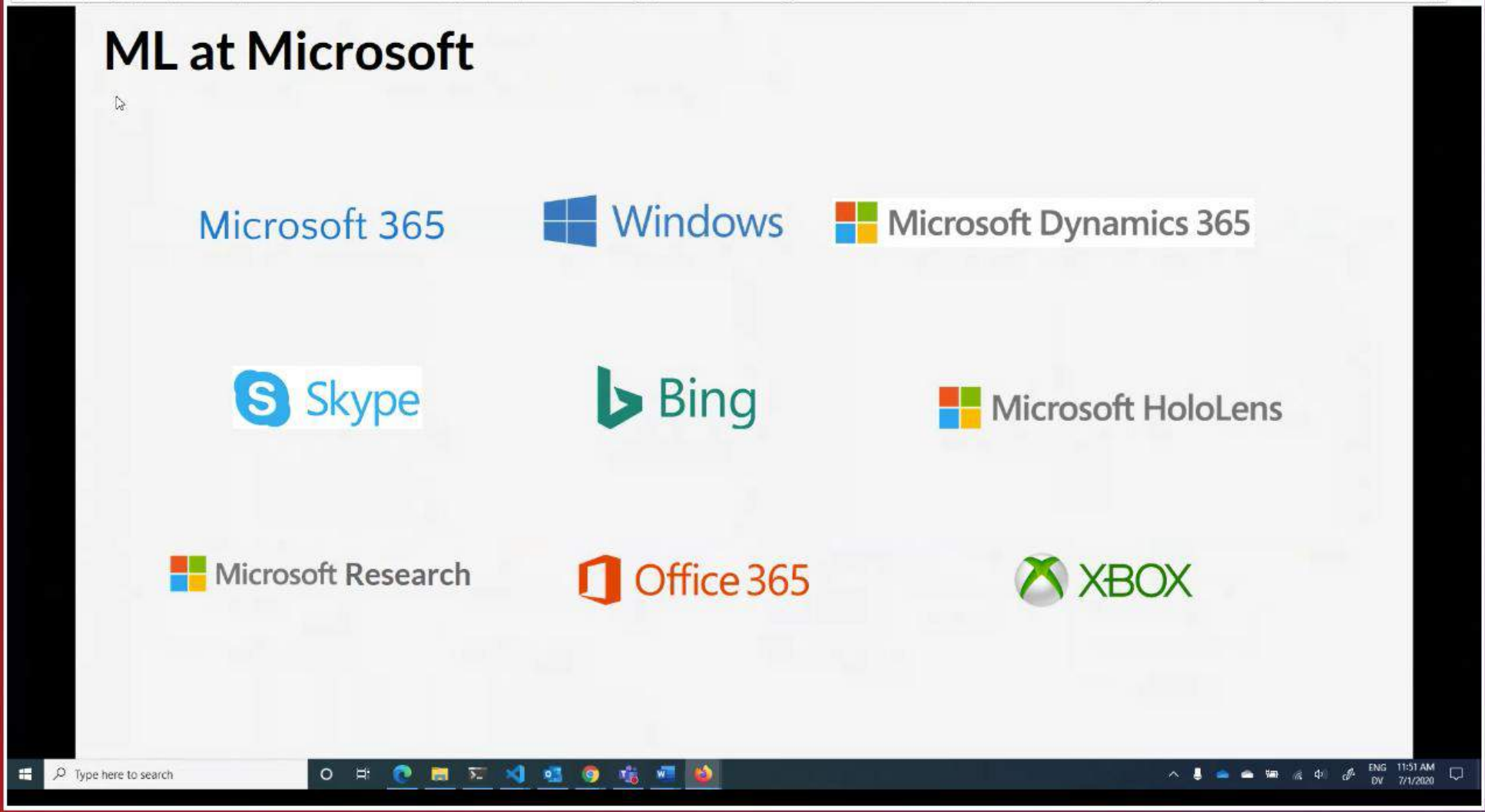
Type here to search

ENG 11:51 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0o72a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_102

ML at Scale

180 million

Monthly active Office 365 users using AI

18 Billion

Questions Asked of Cortana

6.5 Trillion

Number of Signals Analyzed to Block Emerging Threats DAILY

Type here to search

ENG 11:52 AM
DV 7/1/2020

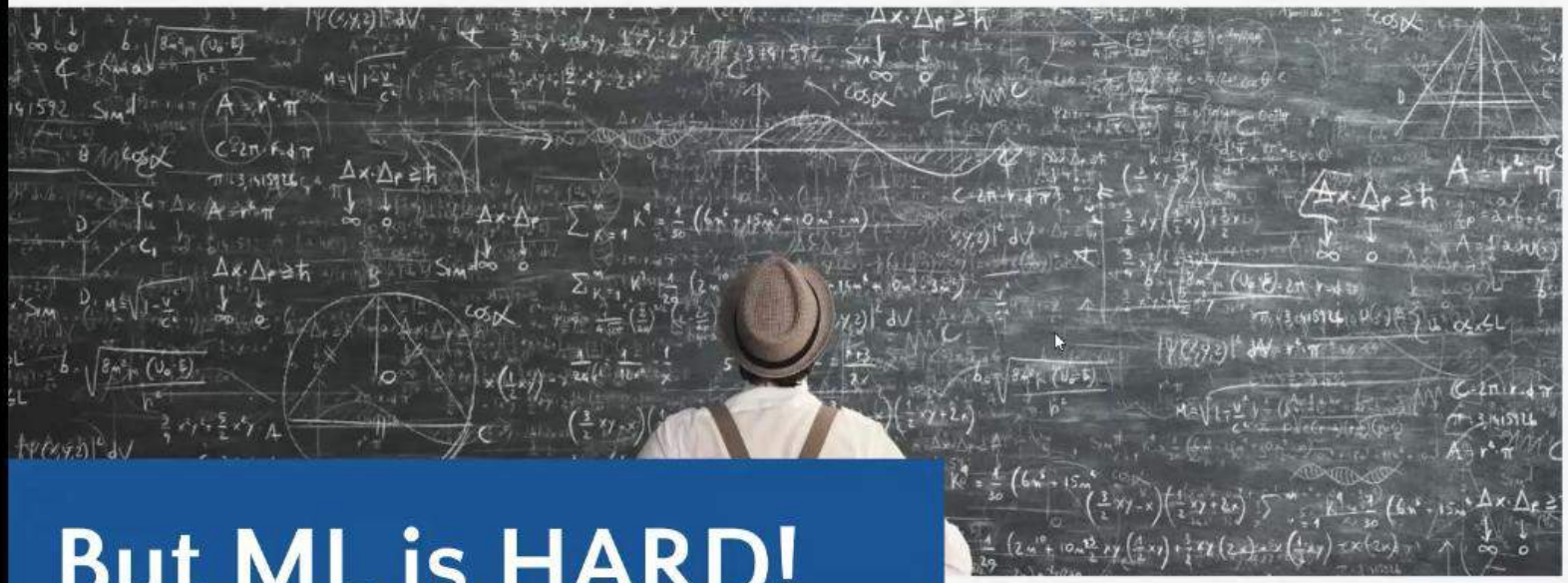
Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUlMbcTnd4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IlTYYz10JTNMfVrg%3A1595628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_112



But ML is HARD!

Slide 5 Q&A Notes Pointer Tips

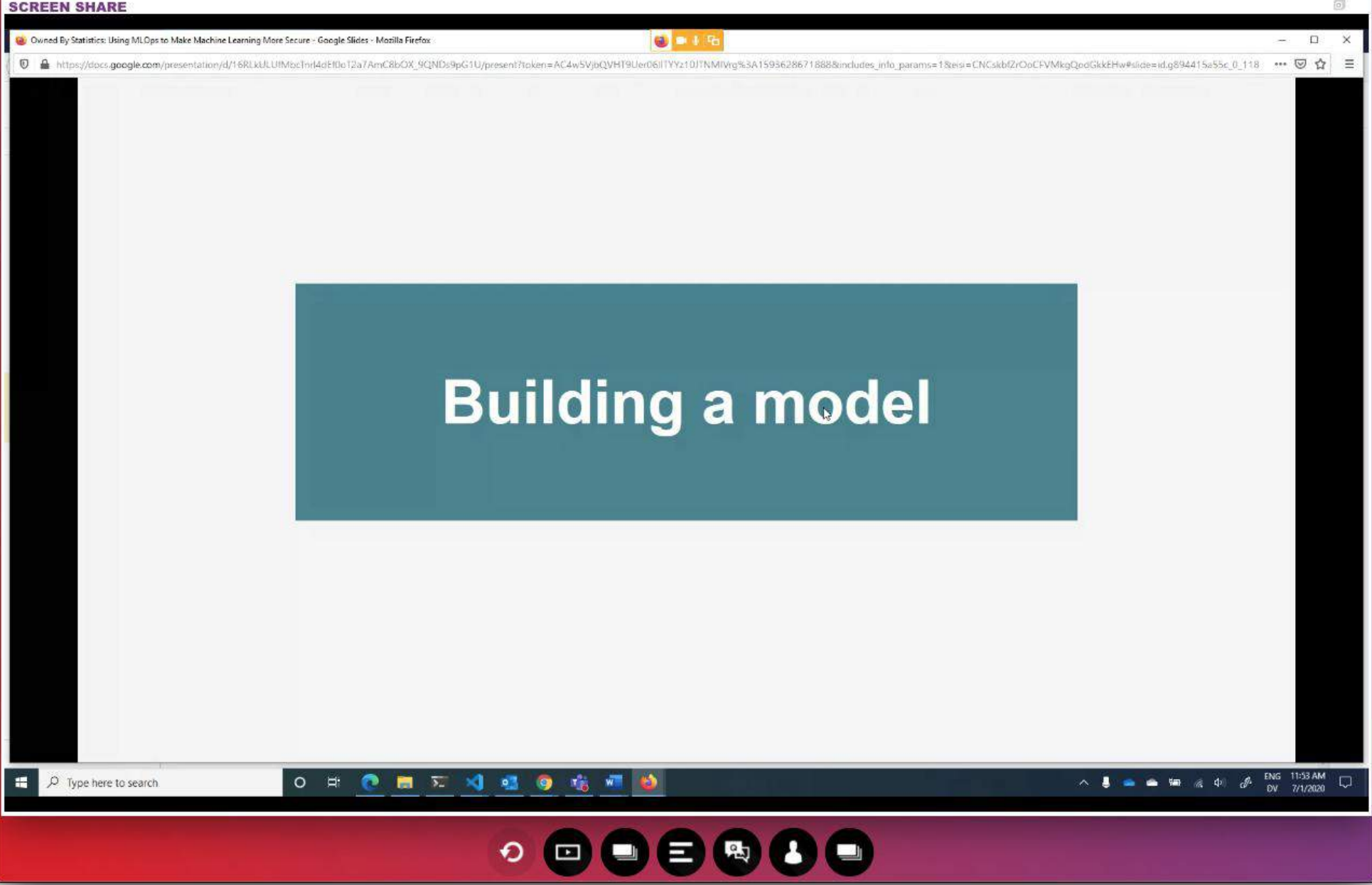
Type here to search

ENG DV 11:53 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

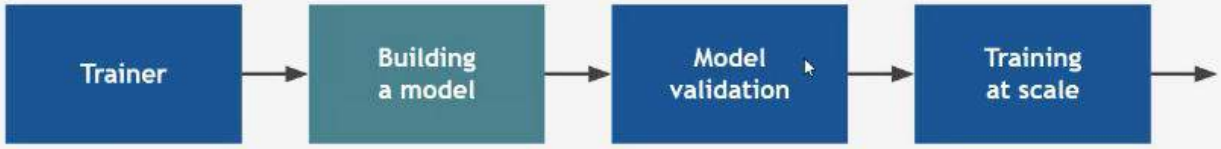


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox



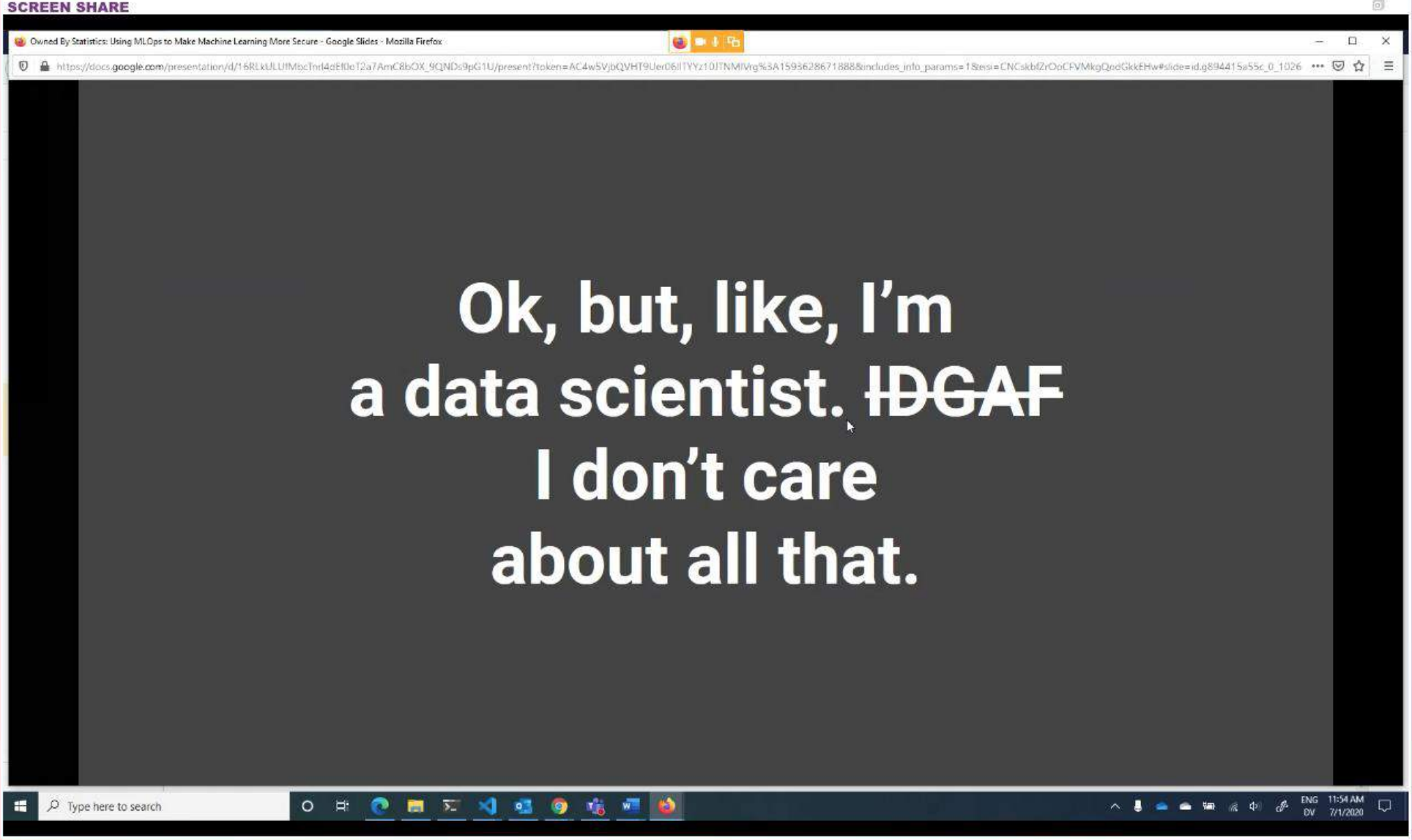
Type here to search

ENG 11:53 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

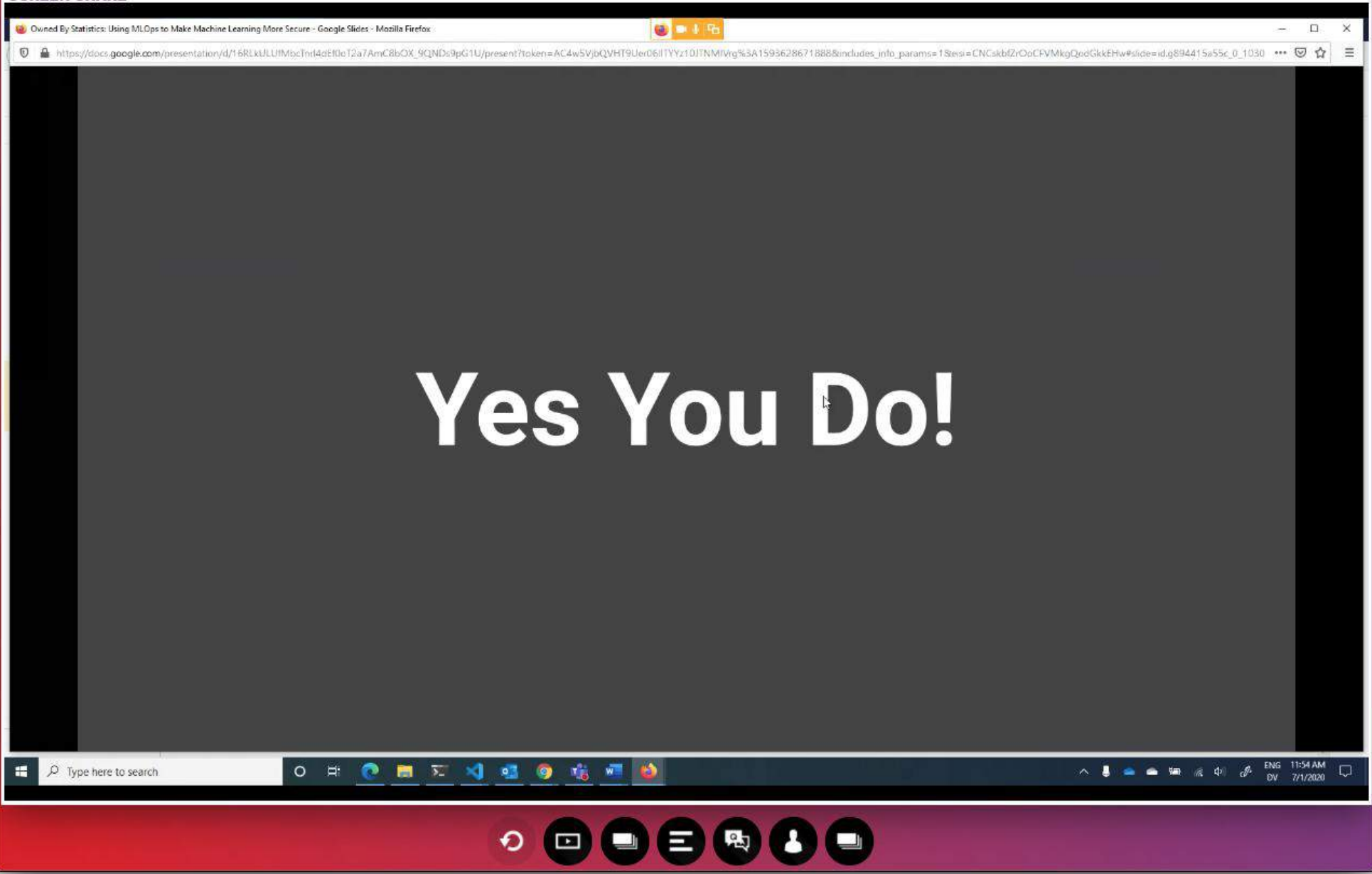
SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkkJLUIMbcTnrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9UerO6llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_1034

ginablaber
@ginablaber

Follow

The story of enterprise Machine Learning: “It took me 3 weeks to develop the model. It’s been >11 months, and it’s still not deployed.” @DineshNirmalIBM #StrataData #strataconf

10:19 AM - 7 Mar 2018

7 Retweets **19** Likes

7 19

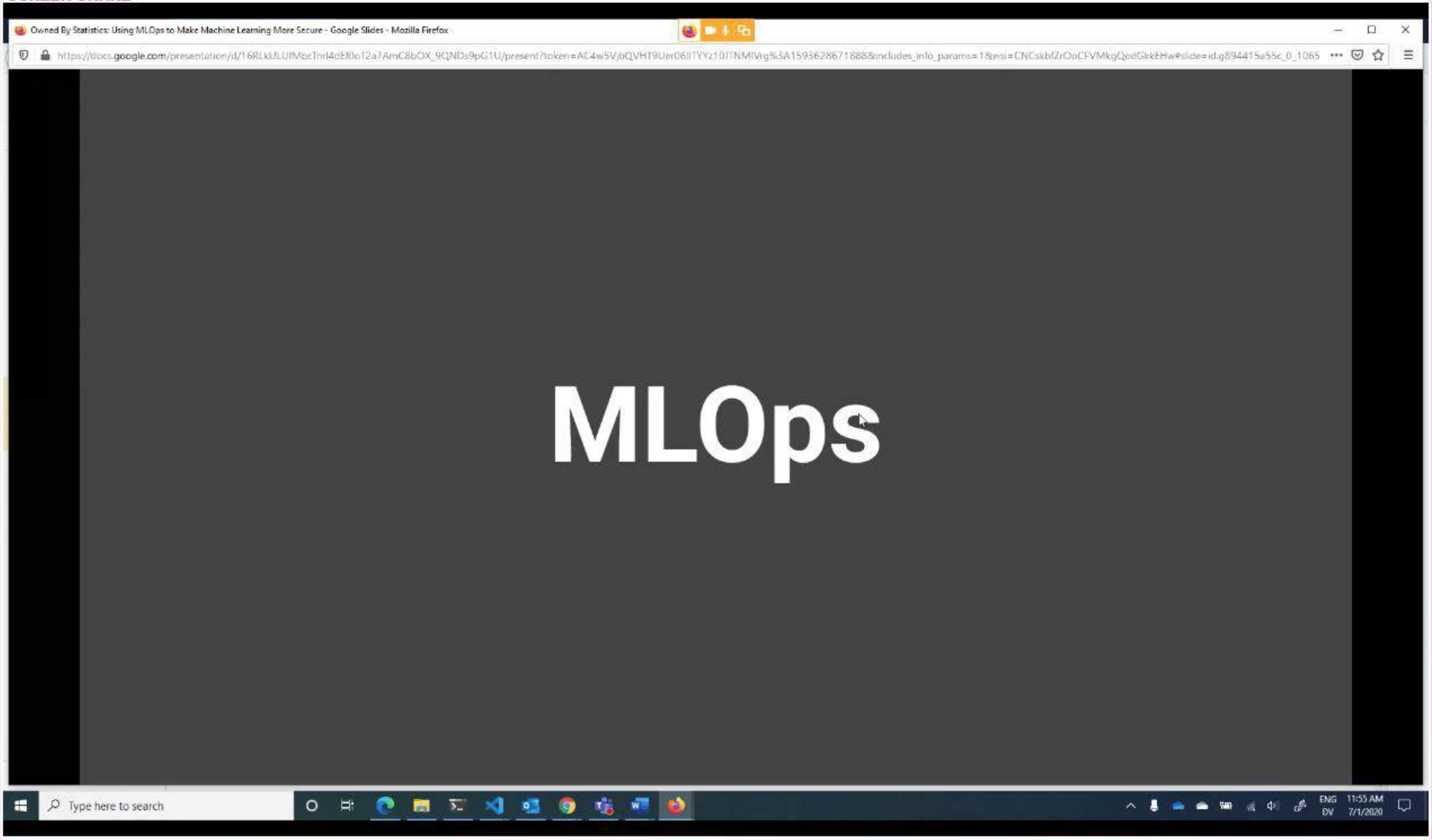
Type here to search

ENG DV 11:54 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

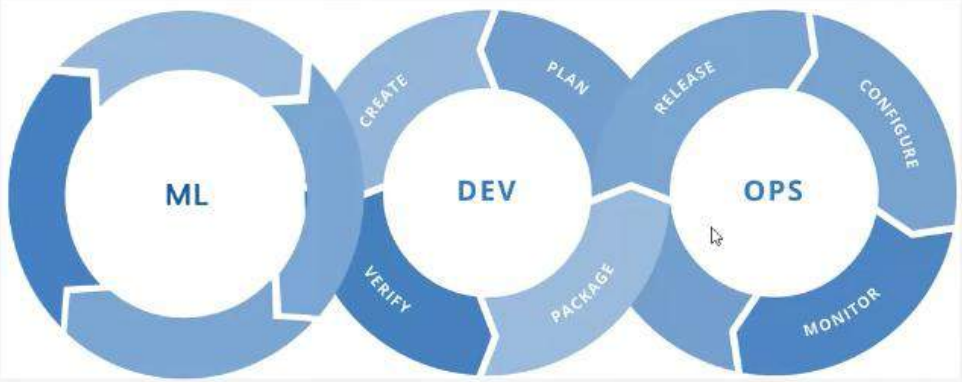
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUlMbcTnrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_1069

MLOps = ML + DEV + OPS



Experiment
Data Acquisition
Business Understanding
Initial Modeling

Develop
Modeling Testing
Continuous Integration
Continuous Deployment

Operate
Continuous Delivery
Data Feedback Loop
System + Model Monitoring

Type here to search

ENG DV 11:55 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnd4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IlTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_1090

MLOps Benefits

Automation / Observability

- Code drives **generation** and **deployments**
- Pipelines are **reproducible** and **verifiable**
- All artifacts can be **tagged** and **audited**

Validation

- SWE best practices for quality control
- Offline comparisons of model **quality**
- Minimize **bias** and enable **explainability**

Reproducibility/ Auditability

- Controlled rollout capabilities
- Live comparison of predicted vs. expected performance
- Results fed back to watch for drift and improve model

== VELOCITY and SECURITY (For ML)

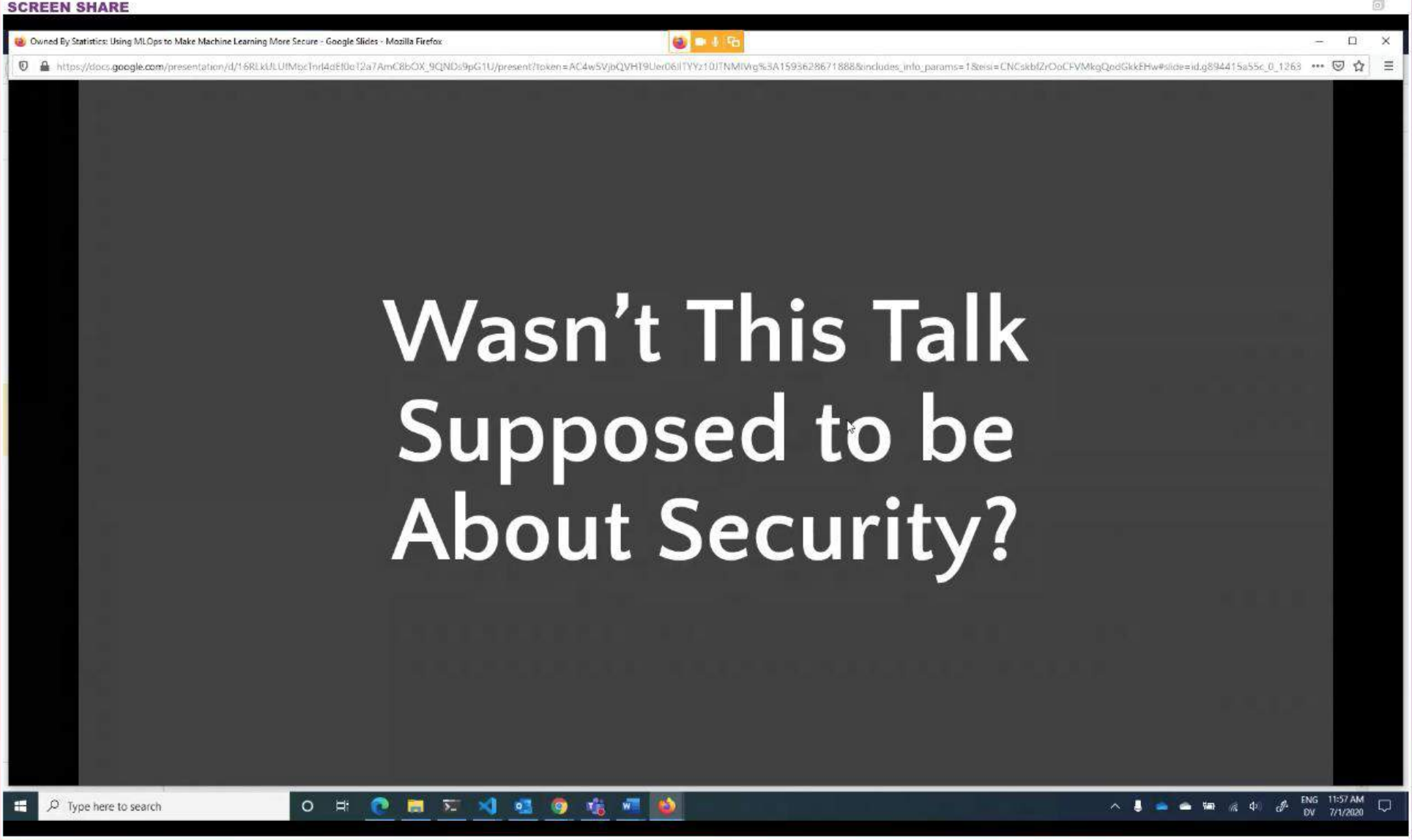
Type here to search

ENG 11:56 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

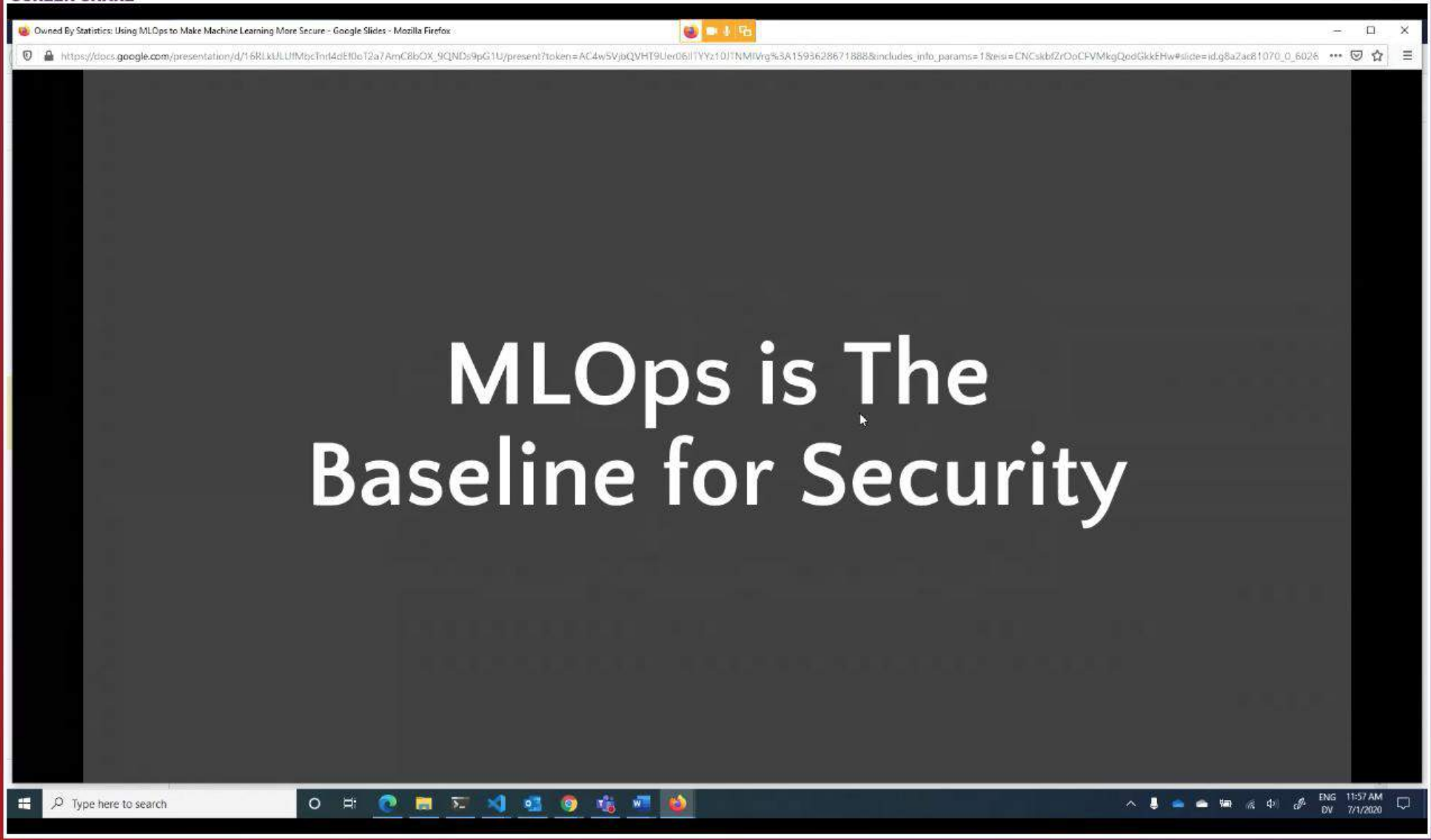
SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

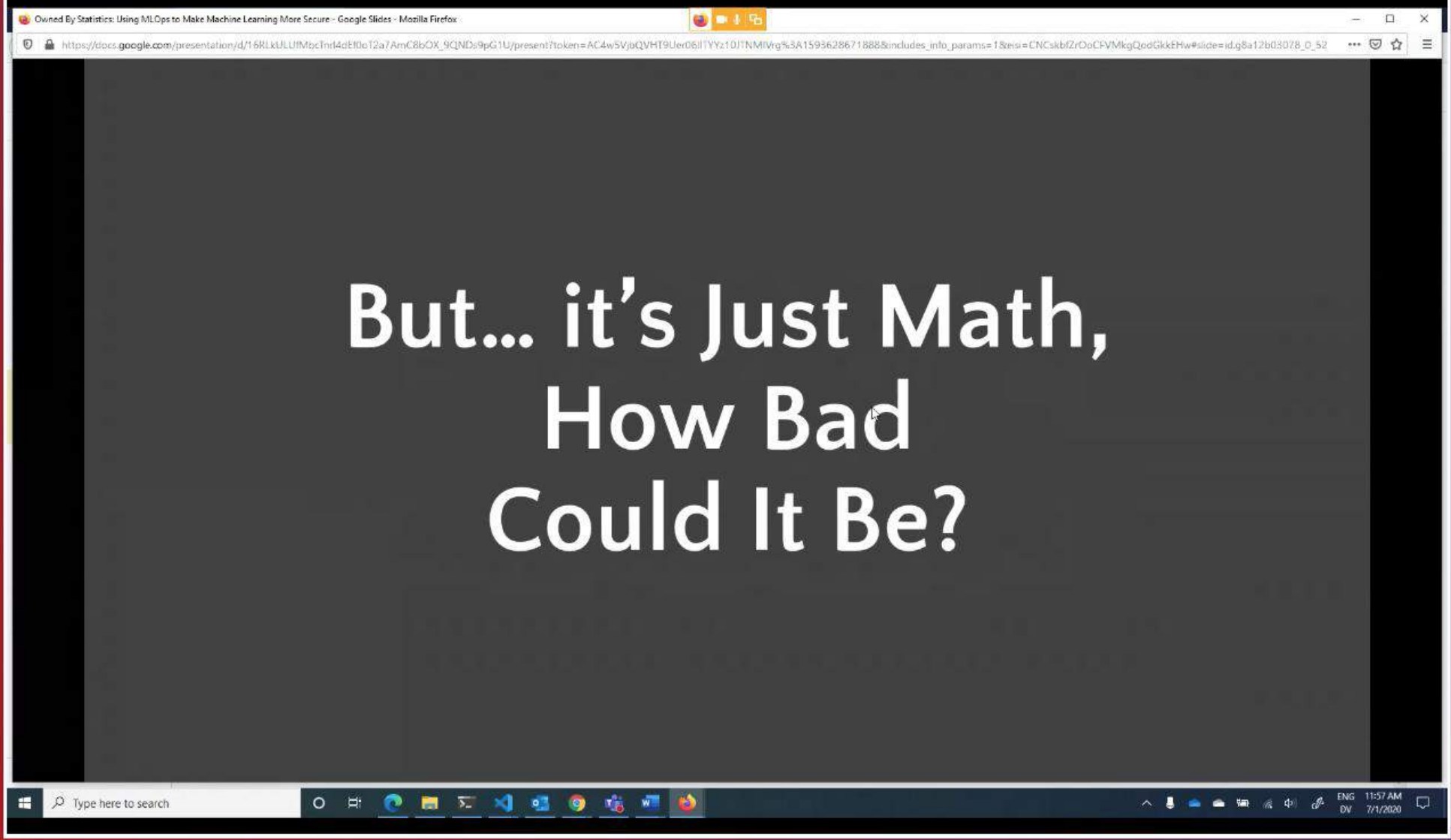
SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMtjcTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10ITNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5865

Three Types of Attacks We'll Talk About Today

1. Attacker Gets Your ML to Lie To You
2. Attacker Takes Your Models
3. Attacker Finds Out About Hidden Data

Type here to search

ENG 11:57 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

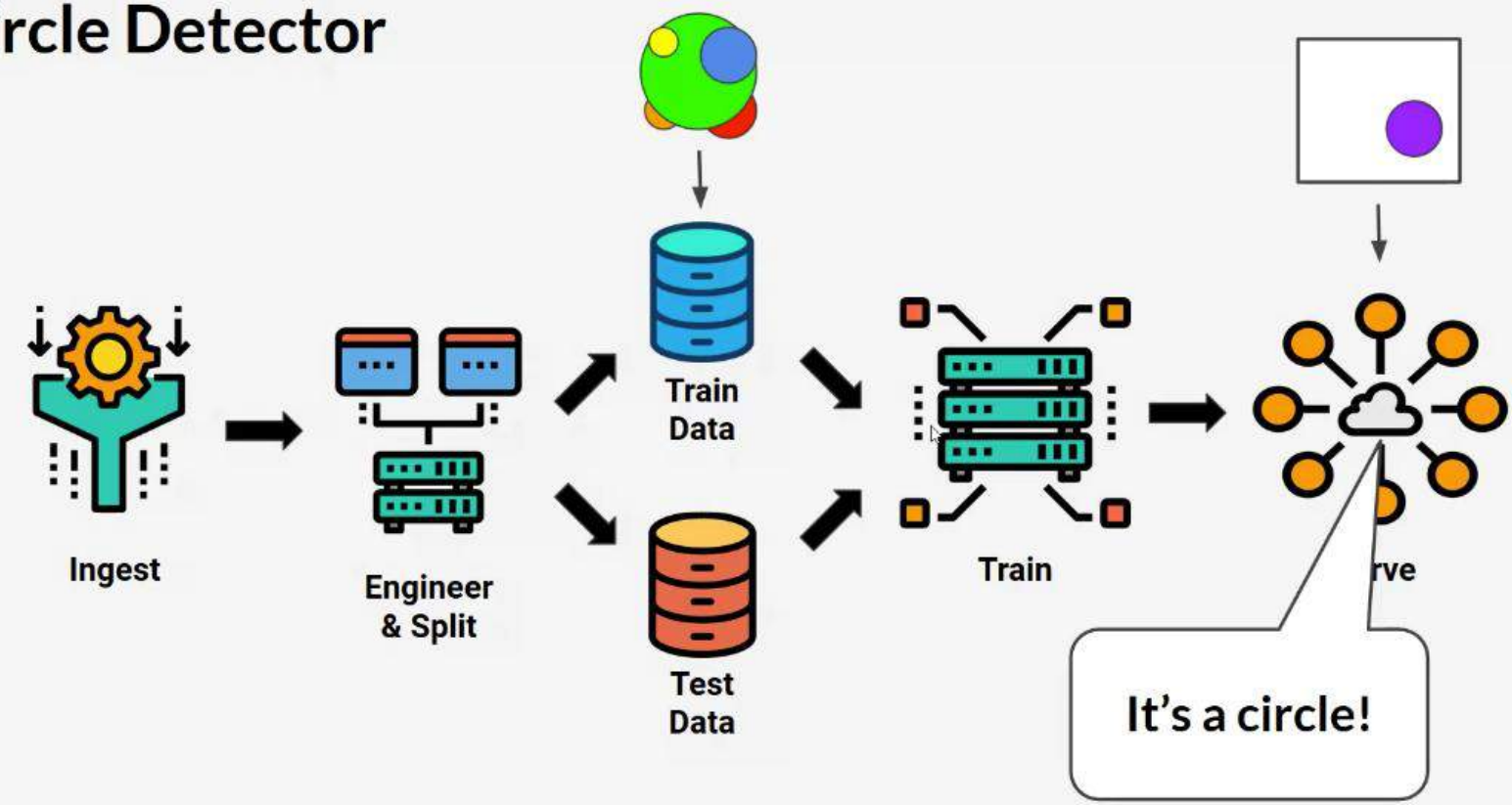
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_7

Circle Detector



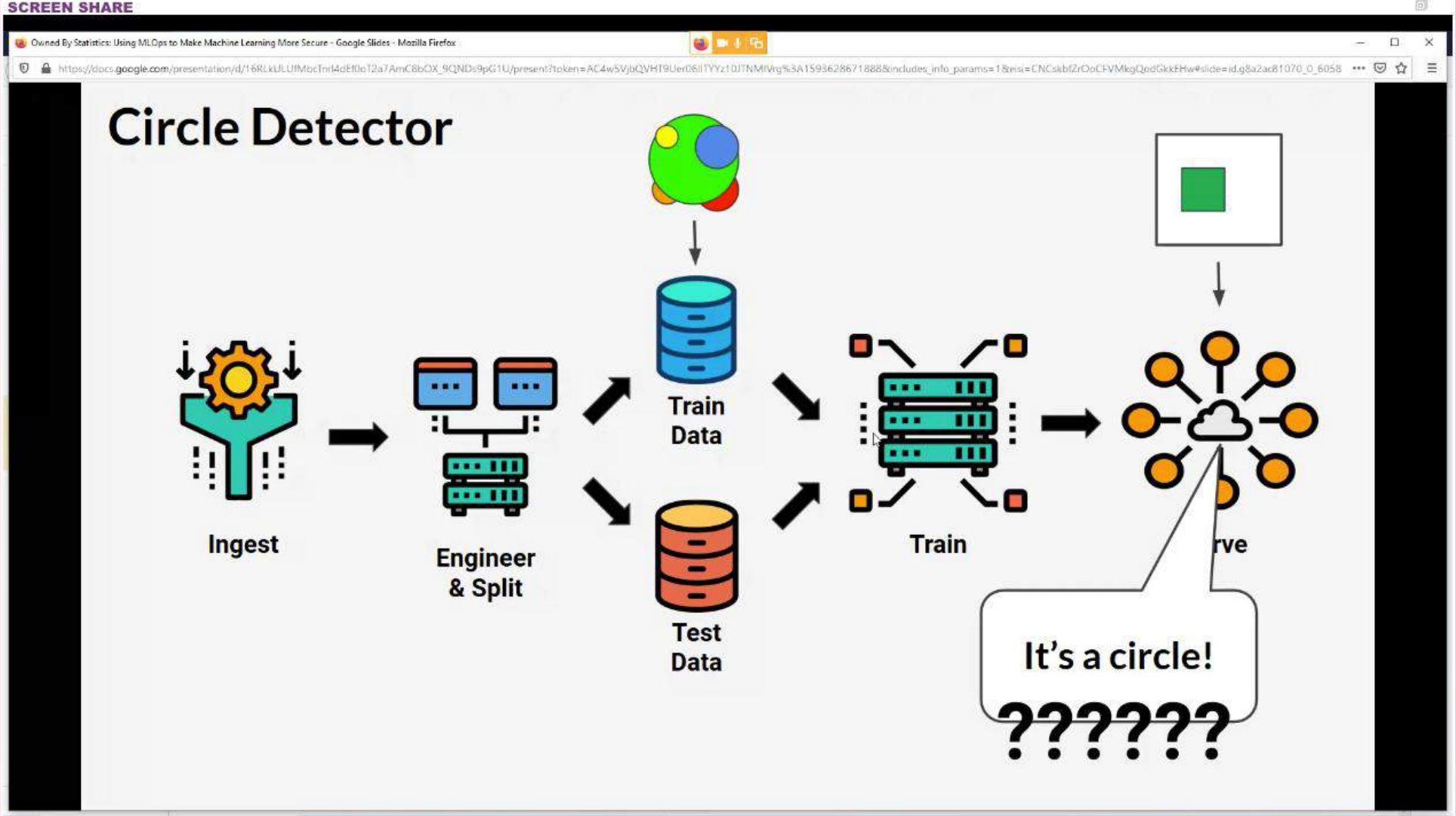
Type here to search

ENG DV 11:59 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



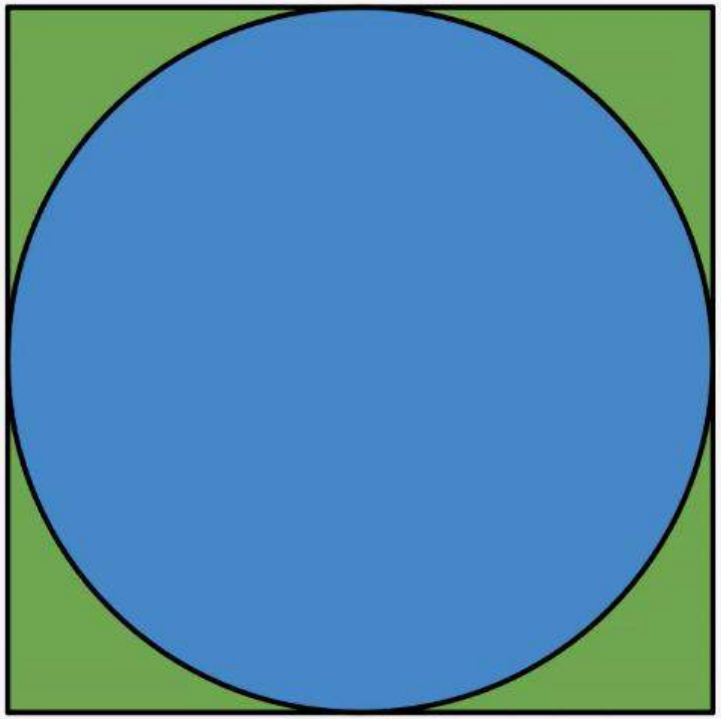
Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

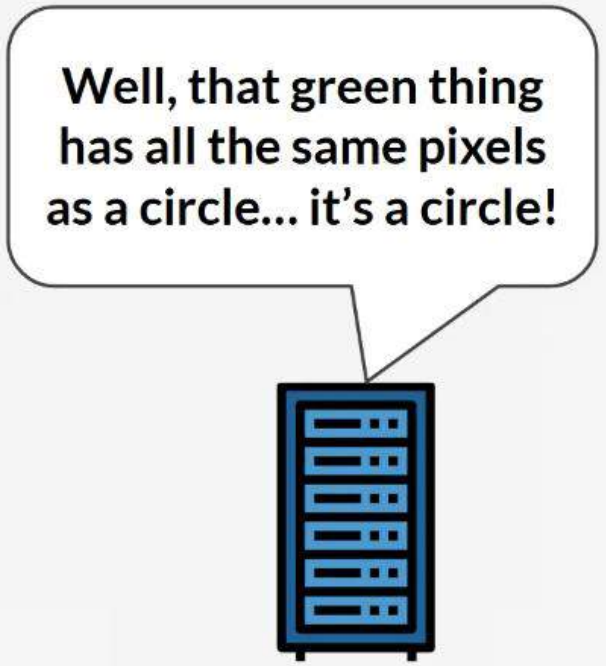
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4cEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_6101





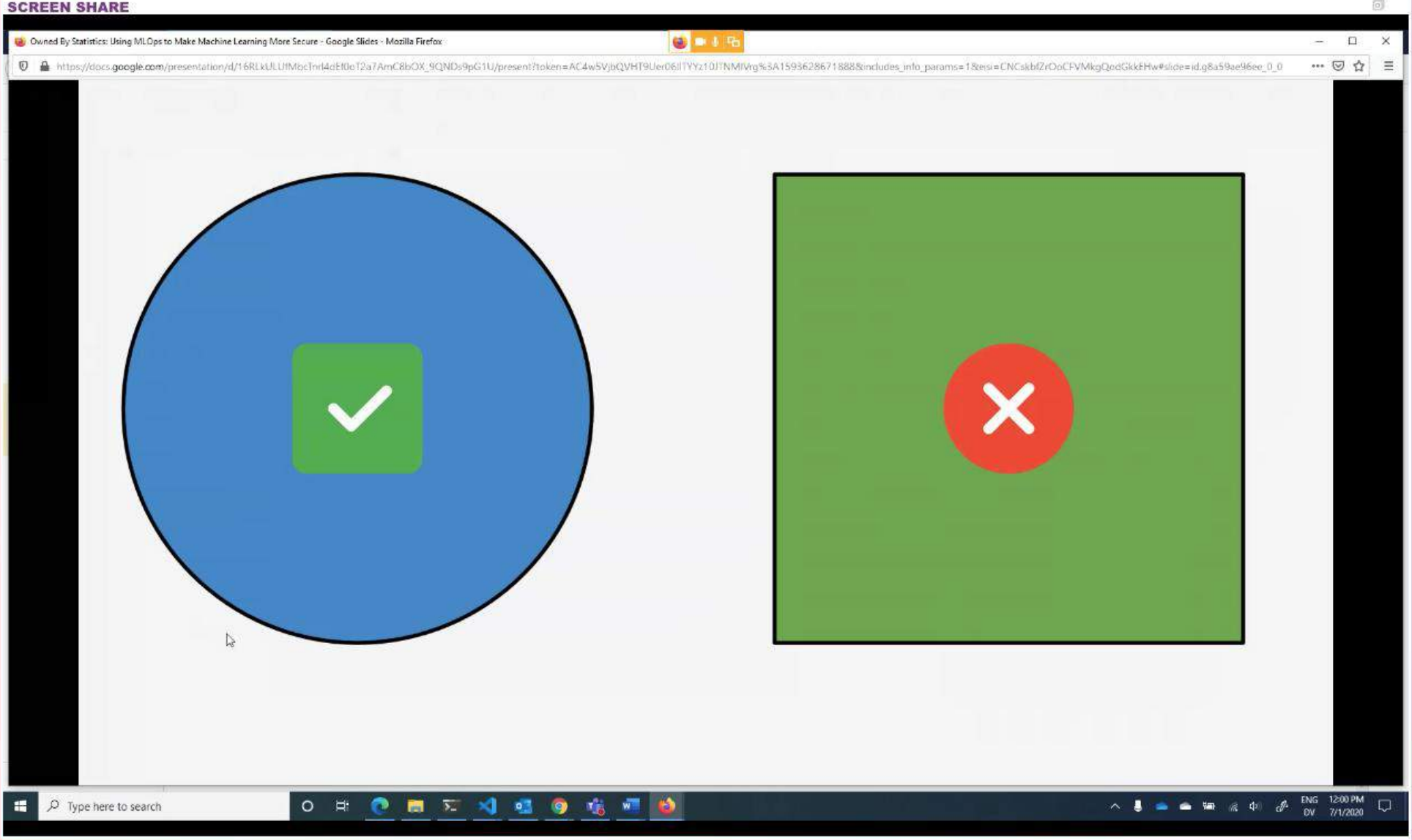
Type here to search

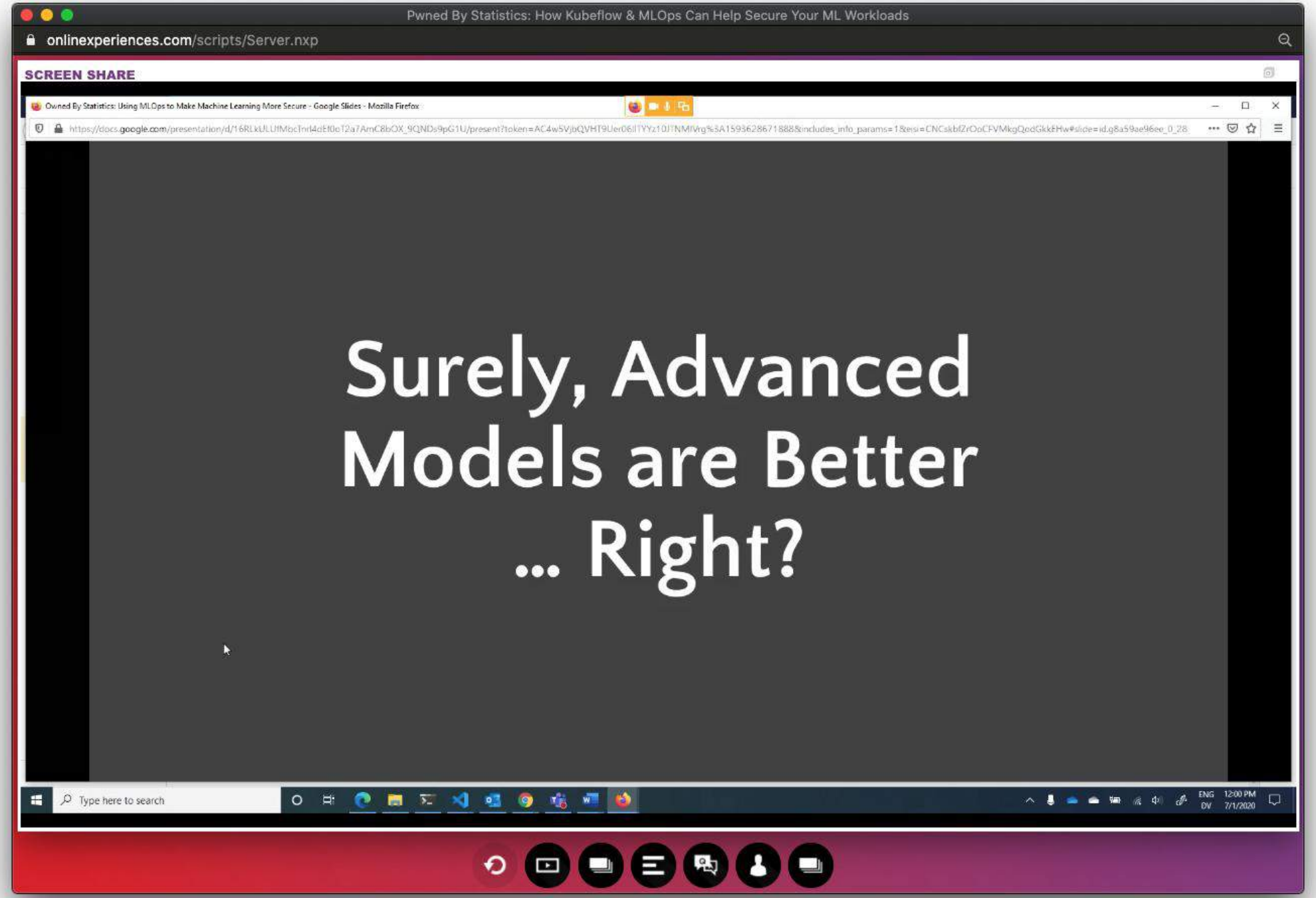
ENG 11:59 AM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE




Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Surely, Advanced Models are Better ... Right?
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcInrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g894415a55c_0_1324

"Why Should I Trust You?" Explaining the Predictions of Any Classifier - Ribeiro, Singh, Guestrin

Type here to search

ENG 12:01 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

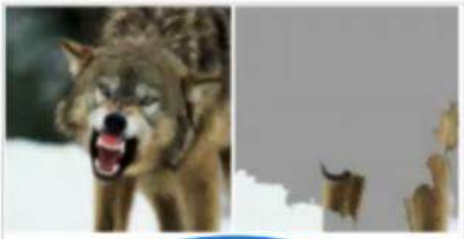
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMpcTnrl4dEf0oT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8b61ac44f6_0_15

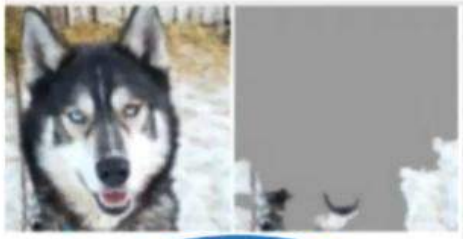
The Explanation Reveals Why

Predicted: wolf
True: wolf

Predicted: husky
True: husky



Predicted: wolf
True: wolf



Predicted: wolf
True: husky

Predicted: husky
True: husky

Predicted: wolf
True: wolf

Type here to search

ENG DV 12:01 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

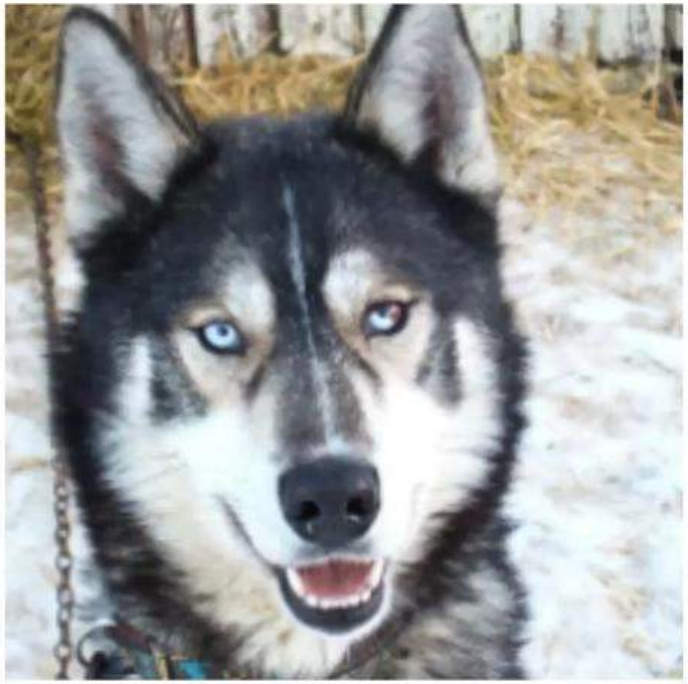
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

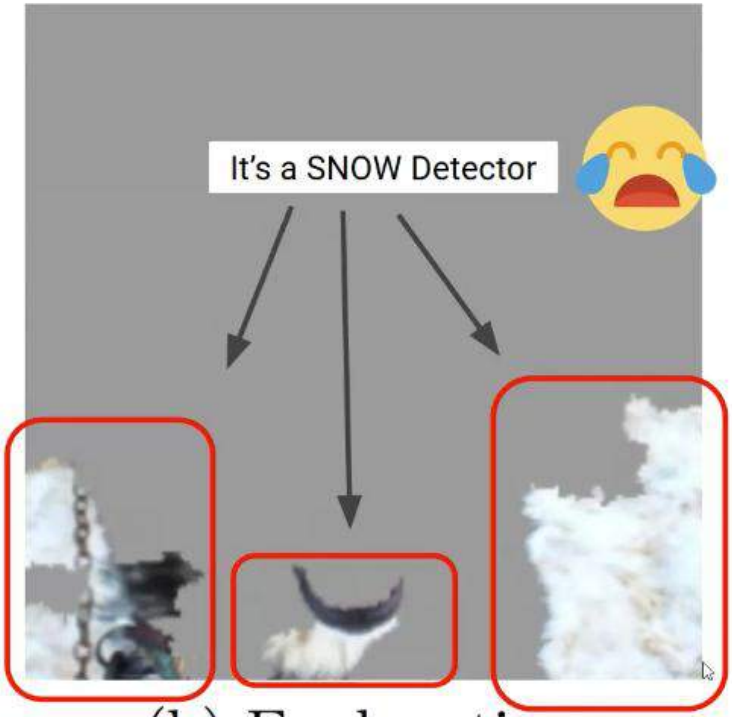
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMpcInrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10ITNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8b61ac44f6_0_4

The Explanation Reveals Why



(a) Husky classified as wolf



(b) Explanation

Type here to search

ENG DV 12:01 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMtcTnrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10ITNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_44

Detecting NSFW Content

VICE Video Queers Built This News Tech Music

MOTHERBOARD
TECH BY VICE

Tumblr's Algorithm Thinks Vomiting Unicorns, Raw Chicken, and Boot Cleaners Are Porn

False Positives

Type here to search

ENG DV 12:02 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMtjcTnrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06iITYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_144

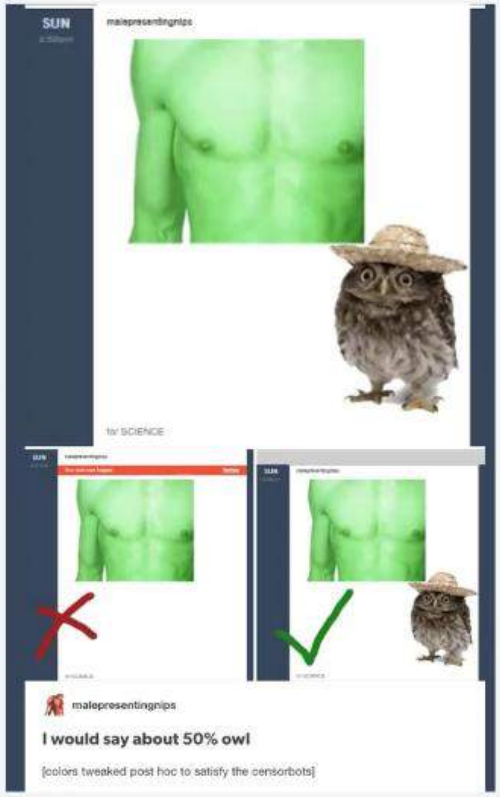
Detecting NSFW Content

VICE Video Queers Built This News Tech Music

MOTHERBOARD
TECH BY VICE

Tumblr's Algorithm Thinks Vomiting Unicorns, Raw Chicken, and Boot Cleaners Are Porn

False Negatives



Type here to search

ENG DV 12:02 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_501

Adversarial Inference

Speed Limit

Rifle

Synthesizing Robust Adversarial Examples - Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok

Type here to search

ENG 12:03 PM
DV 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

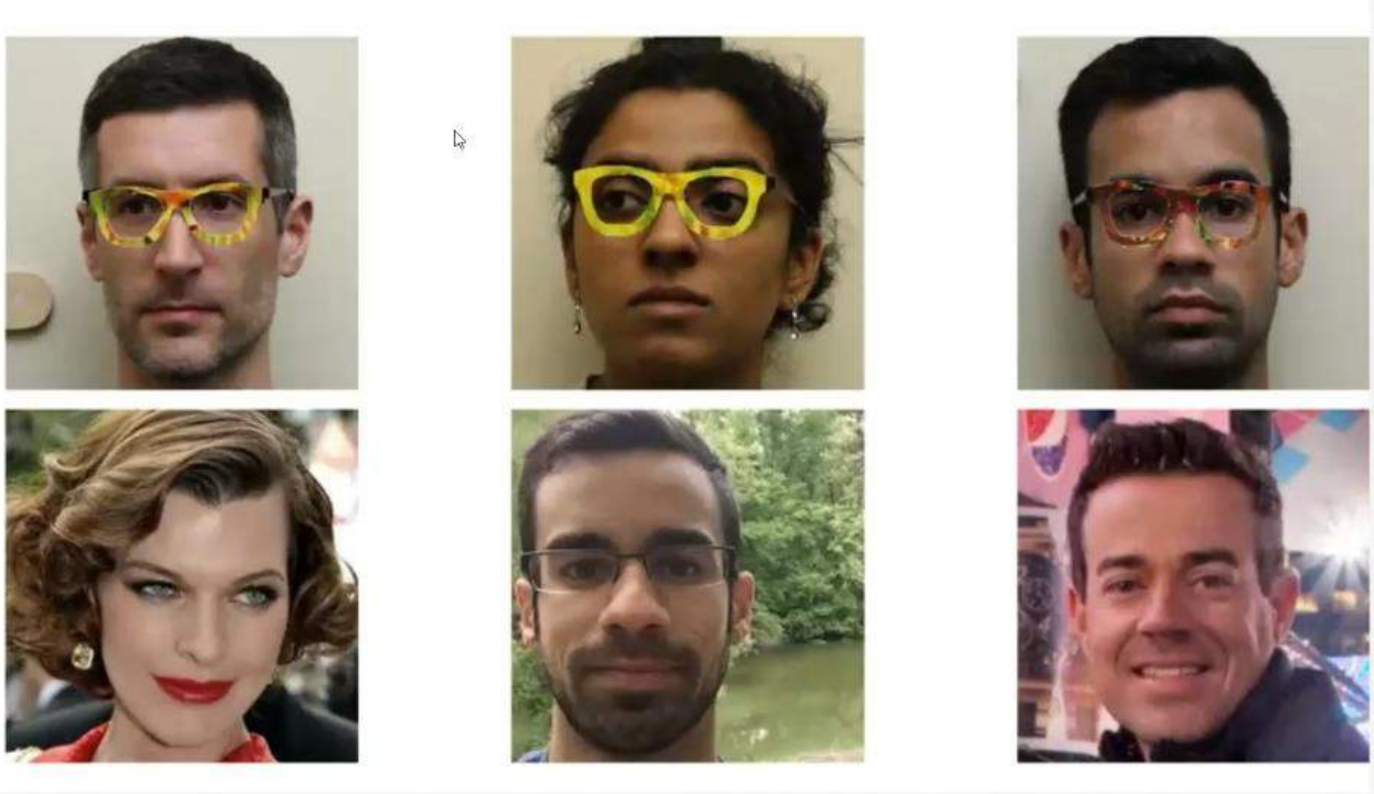
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMocInrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10ITNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_488

Adversarial Inference



Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition - Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter

Type here to search

ENG DV 12:04 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMocTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10ITNMIVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5778

Who cares...

GREGORY BARBER | TOM SIMONITE | BUSINESS | 05.17.2019 07:00 AM

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28 members of Congress, identifying

The Best AI

US government tests find

Facial r

is out o

Database cont

algorithm is w

misidentify bl

Type here to search

ENG DV 12:04 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMocTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrDoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5778

Who cares...

GREGORY BARBER | TOM SIMONITE | BUSINESS | 05.17.2019 07:00 AM

ALLIE FUNK | SECURITY | 07.02.2019 09:00 AM

I Opted Out of Facial Recognition at the Airport —It Wasn't Easy

Opinion: We've been assured that facial recognition technology is secure, reliable, and accurate. That's far from certain.

Type here to search

ENG DV 12:05 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_371

Poisoned Federated Learning

How To Backdoor Federated Learning

Eugene Bagdasaryan
Cornell Tech, Cornell University
eugene@cs.cornell.edu

Andreas Veit
Cornell Tech, Cornell University
andreas@cs.cornell.edu

Yiqing Hua
Cornell Tech, Cornell University
yiqing@cs.cornell.edu

Deborah Estrin
Cornell Tech, Cornell University
destrin@cs.cornell.edu

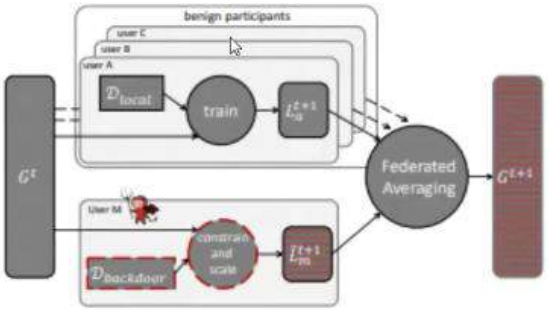
Vitaly Shmatikov
Cornell Tech, Cornell University
shmat@cs.cornell.edu

Abstract

Federated learning enables thousands of participants to construct a deep learning model without sharing their private training data with each other. For example, multiple smartphones can jointly train a next-word predictor for keyboards without revealing what individual users type.

Federated models are created by aggregating model updates submitted by participants. To protect confidentiality of the training data, the aggregator by design has no visibility into how these updates are generated. We show that this makes federated learning vulnerable to a model-poisoning attack that is significantly more powerful than poisoning attacks that target only the training data.

A malicious participant can use *model replacement* to introduce



How To Backdoor Federated Learning - Bagdasaryan, Veit, Hua, Estrin, Shmatikov

Type here to search

ENG DV 12:05 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

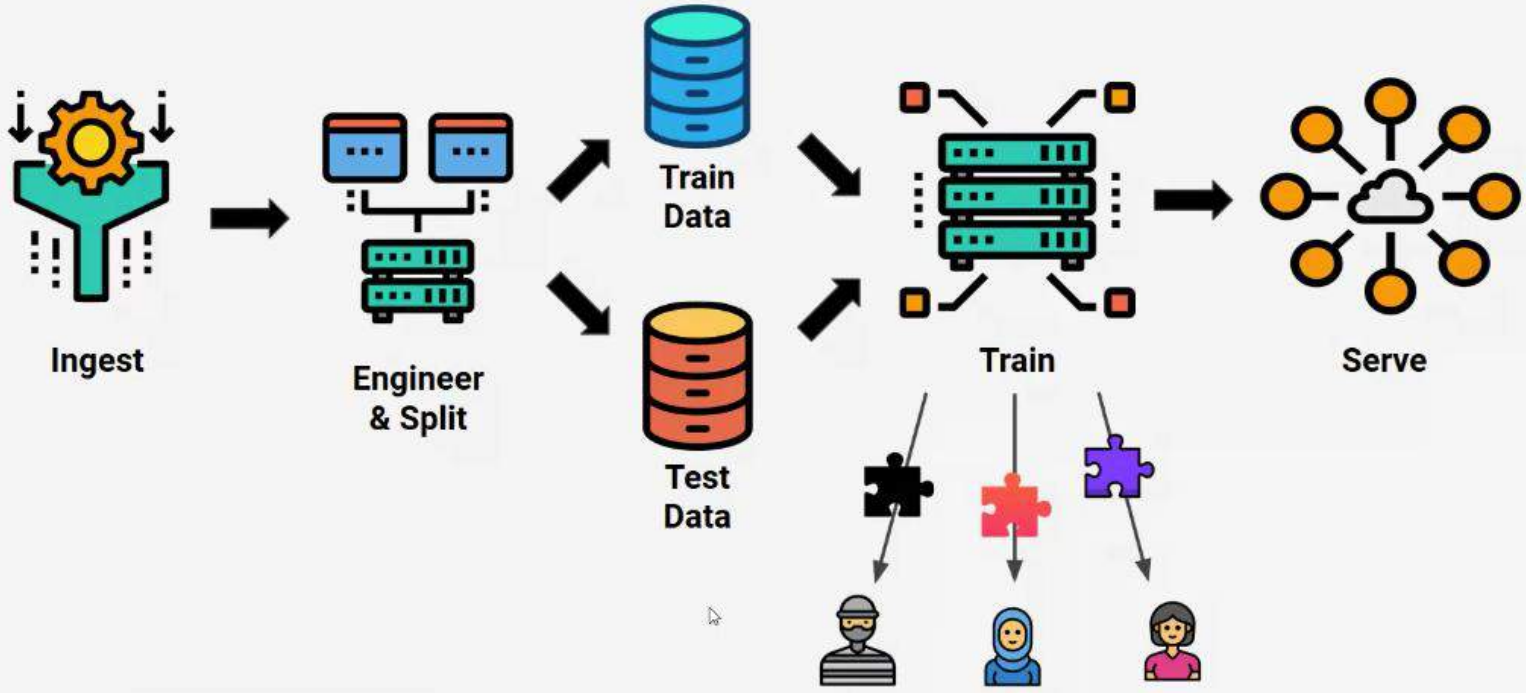
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_567

Federated learning makes things harder!



Type here to search

ENG DV 12:06 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

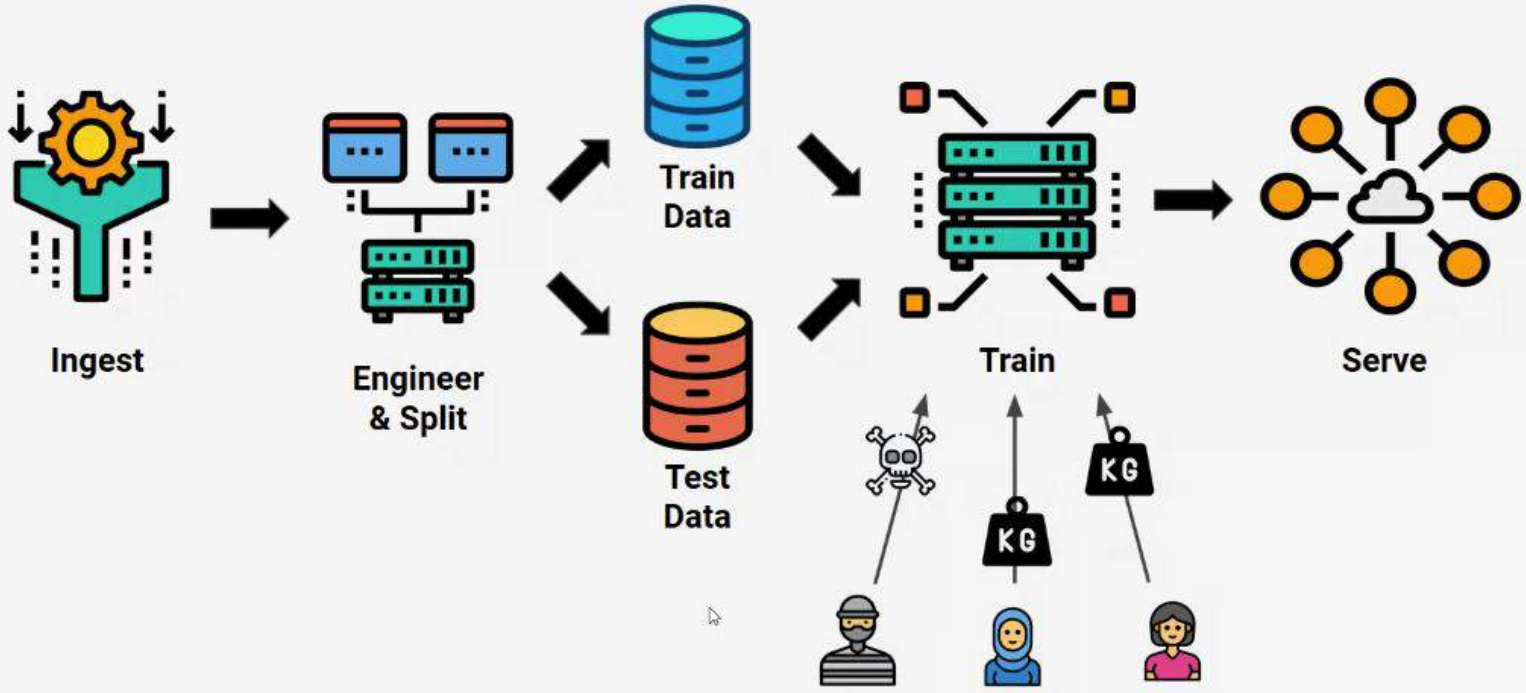
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_614

Federated learning makes things harder!



Type here to search

ENG DV 12:06 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

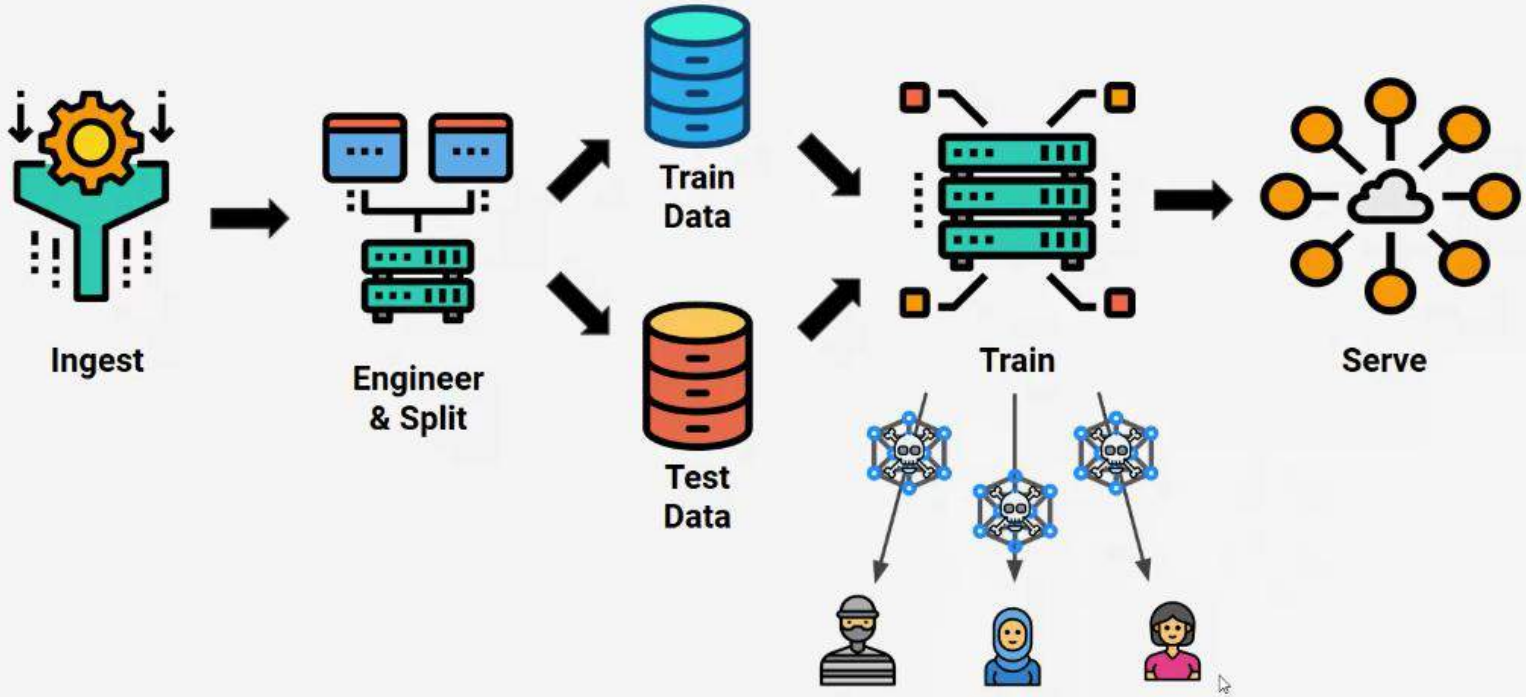
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskblZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_662

Federated learning makes things harder!



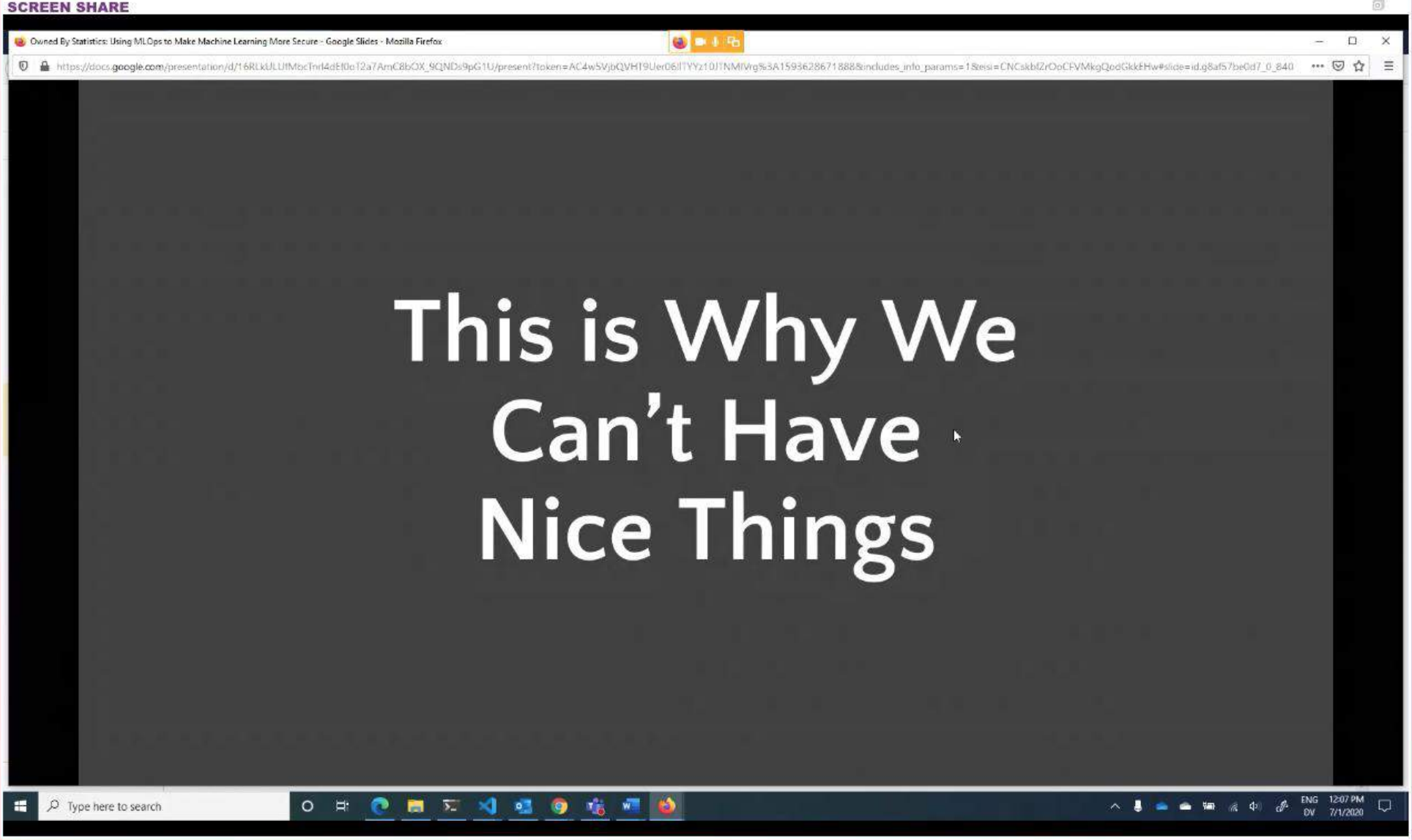
Type here to search

ENG DV 12:07 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMbcTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5879

Poisoned Federated Learning Defense

Learning to Detect Malicious Clients for Robust Federated Learning

Suyi Li[1], Yong Cheng[2], Wei Wang[1]
Yang Liu[2], Tianjian Chen[2]

[1]The Hong Kong University of Science and Technology
[2]AI Department, WeBank

{slida, weiwa}@cse.ust.hk, {petercheng, yangliu, tobychen}@webank.com

Abstract

Federated learning systems are vulnerable to attacks from malicious clients. As the central server in the system cannot govern the behaviors of the clients, a rogue client may initiate an attack by sending malicious model updates to the server, so as to degrade the learning performance or enforce targeted model poisoning attacks (a.k.a. backdoor attacks). Therefore, timely detecting these malicious model updates and the underlying attackers becomes critically important. In this work, we propose a new framework for robust federated learning where the central server learns to *detect and remove* the malicious model updates using a powerful

harmful model updates, initiating *adversarial attacks* on the global model [Kairouz *et al.*, 2019]. In this paper, we consider two types of adversarial attacks, namely the *untargeted* attacks and the *targeted* attacks. The untargeted attacks aim to degrade the overall model performance and can be viewed as Byzantine attacks which result in model performance deterioration or failure of model training [Li *et al.*, 2019a; Wu *et al.*, 2019]. The targeted attacks (a.k.a. backdoor attacks) [Bhagoji *et al.*, 2019; Bagdasaryan *et al.*, 2019; Sun *et al.*, 2019], on the other hand, aim to modify the behaviors of the model on some specific data instances chosen by the attackers (e.g., recognizing the images of cats as dogs), while keeping the model performance on the other data instances unaffected. Both the untargeted and targeted attacks can result in catastrophic consequences. Therefore, attack-

Learning to Detect Malicious Clients for Robust Federated Learning - Li, Cheng, Wang, Liu, Chen

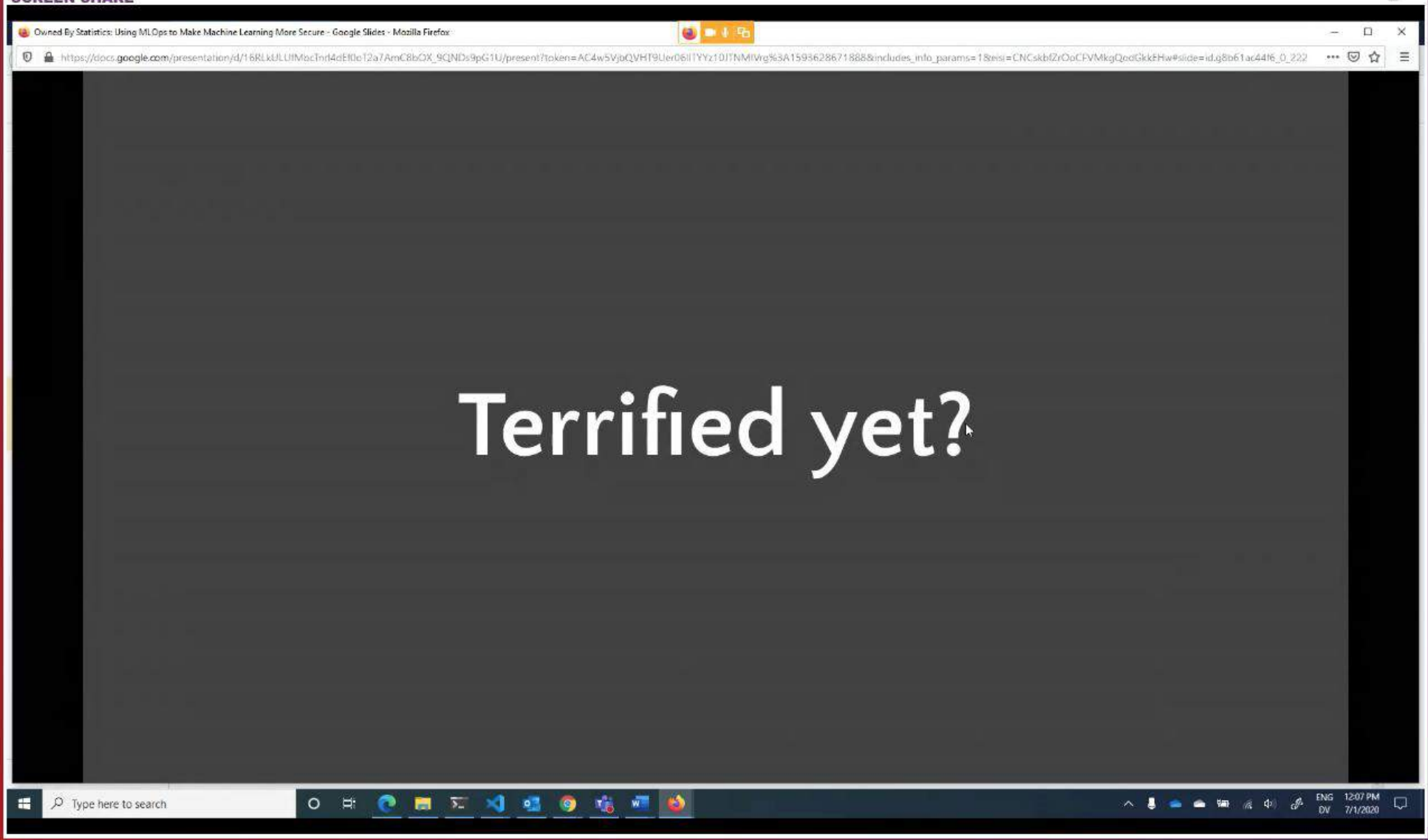
Type here to search

ENG 12:07 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

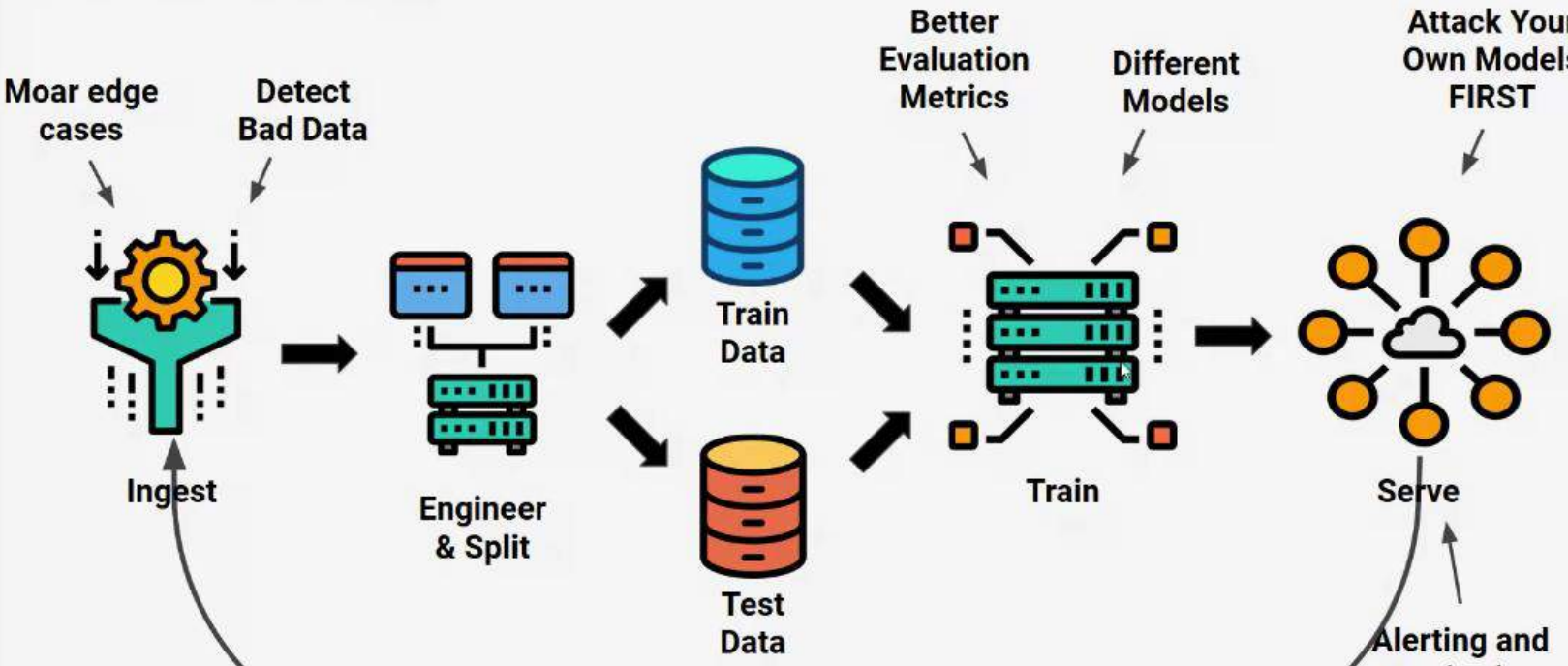
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

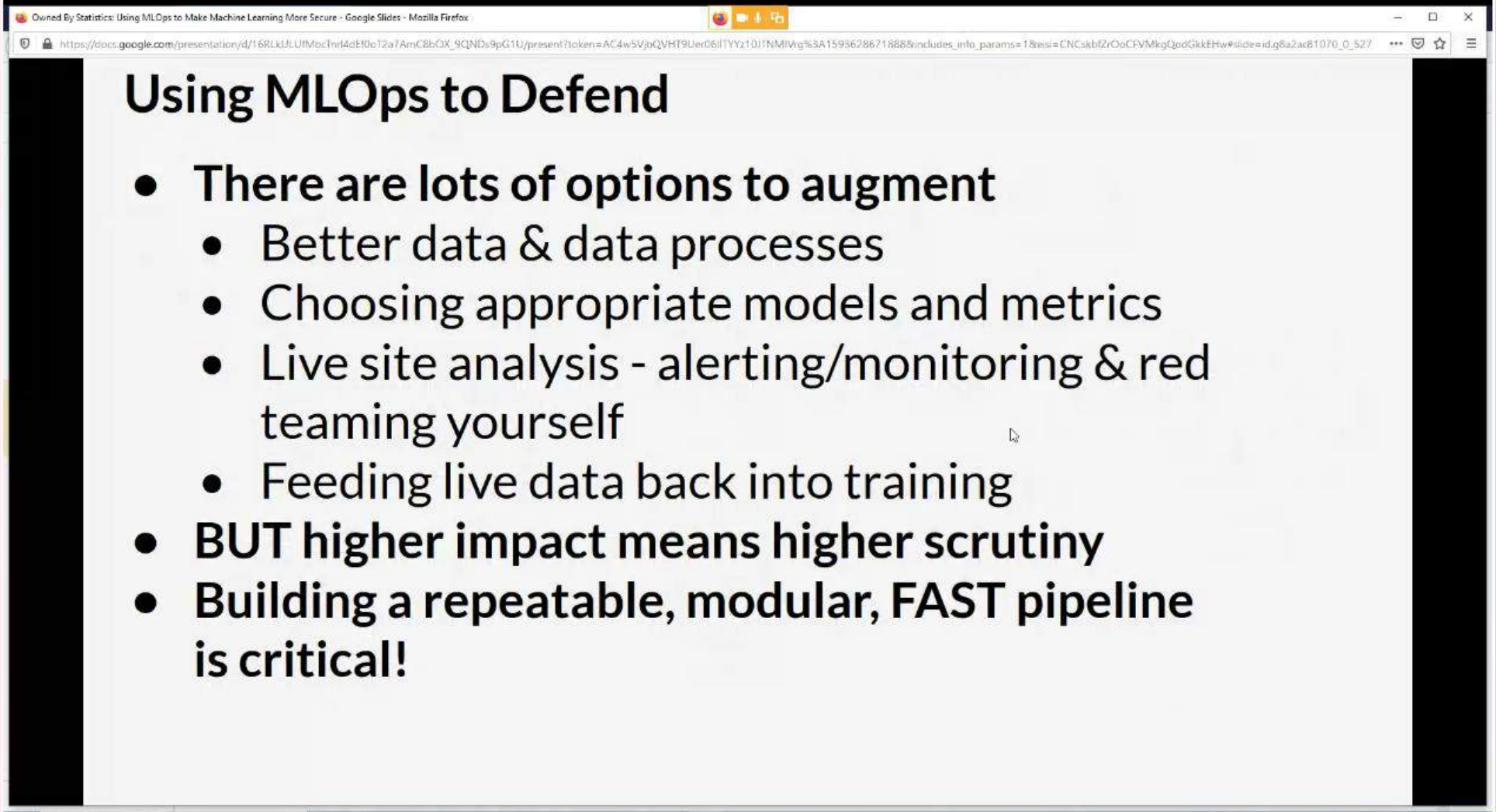
https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1595628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkFHw#slide=id.g8b61ac44f6_0_226

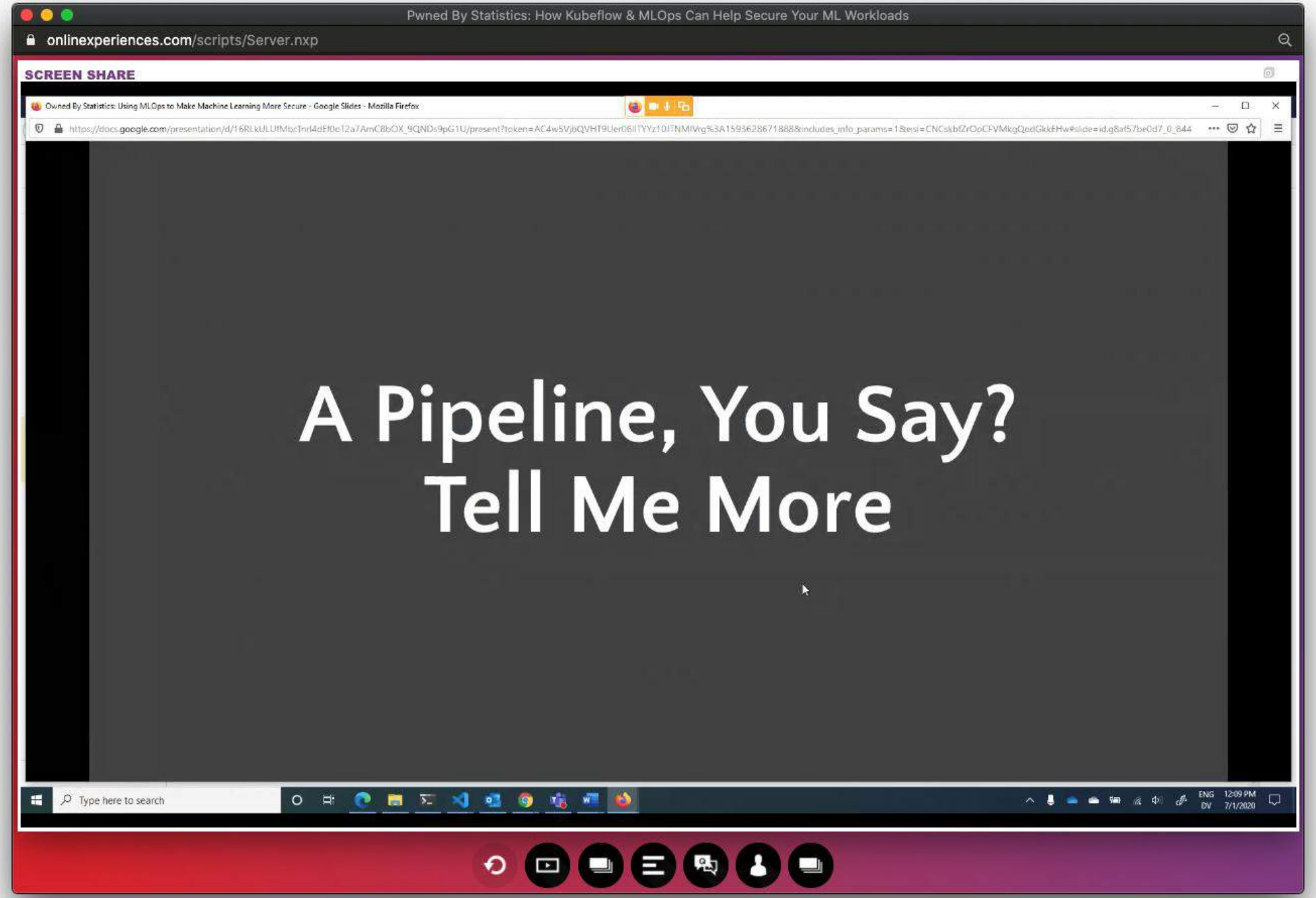
Tools to Defend



Type here to search

ENG DV 12:08 PM 7/1/2020


Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Using MLOps to Defend
There are lots of options to augment
Better data & data processes
Choosing appropriate models and metrics
Live site analysis - alerting/monitoring & red teaming yourself
Feeding live data back into training
BUT higher impact means higher scrutiny
Building a repeatable, modular, FAST pipeline is critical!
Type here to search


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
A Pipeline, You Say?
Tell Me More
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbc1nrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IlTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_0

Building a Pipeline

1. Use a CI/CD Platform - e.g. GitHub Actions, Jenkins
2. Add modular components
   a. Loosely coupled, microservice oriented
   b. Mix and match! (e.g. on-prem, cloud, self-hosted)
   c. Use pre-built solutions - http://mlops-github.com/
3. Measure, measure, measure and **UPDATE**
   a. Models go stale QUICKLY
   b. Don't let adversaries be the ones to alert you that your systems are out of date

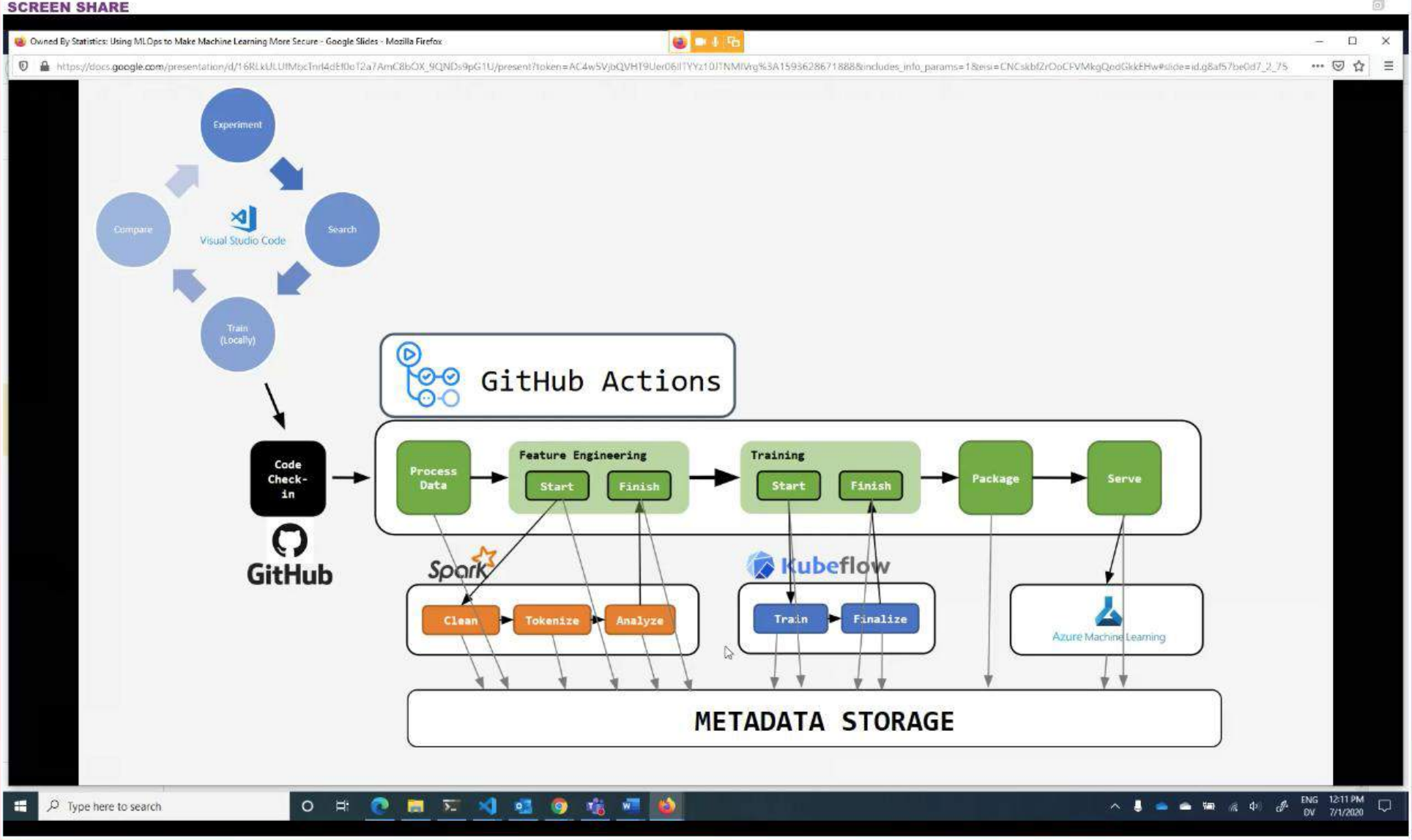
Type here to search

ENG 12:09 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

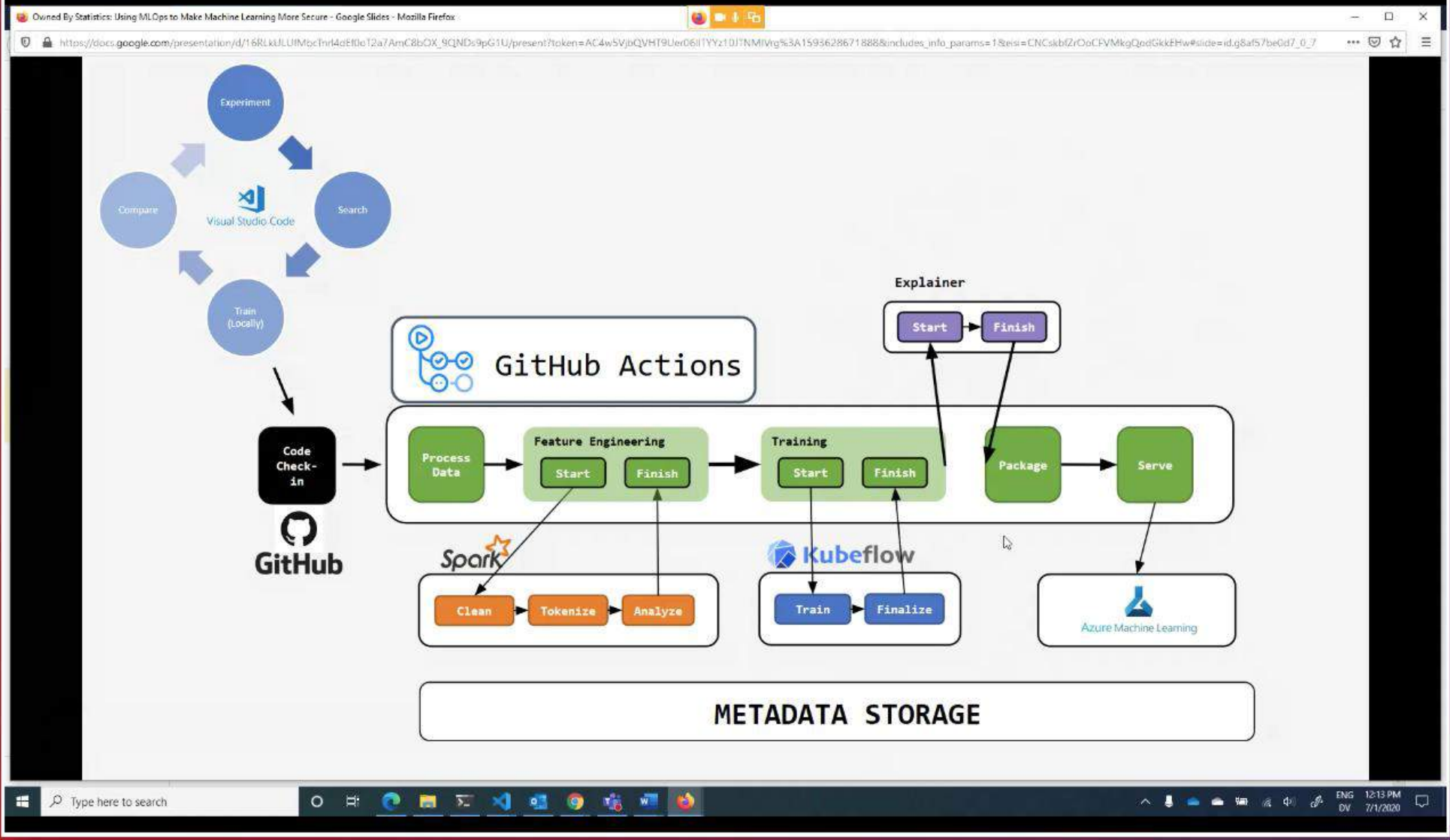
SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkkJLUIMbcTnrl4dEf0oT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVig%3A1593628671888&includes_info_params=1&eisi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5931

Three Types of Attacks We'll Talk About Today

1. Attacker Gets Your ML to Lie To You
2. Attacker Takes Your Models
3. Attacker Finds Out About Hidden Data

Type here to search

ENG DV 12:13 PM 7/1/2020

Motivation

- Malicious user attempts to reproduce the original model
  - Primary goal is just private access
  - *Mostly* correct performance is secondary (but important)
- Gives foothold for further attacks down the line
  - More complete/accurate extraction
  - Extract private information built into the model
  - Construct adversarial examples
- **VERY** hard to defend against completely

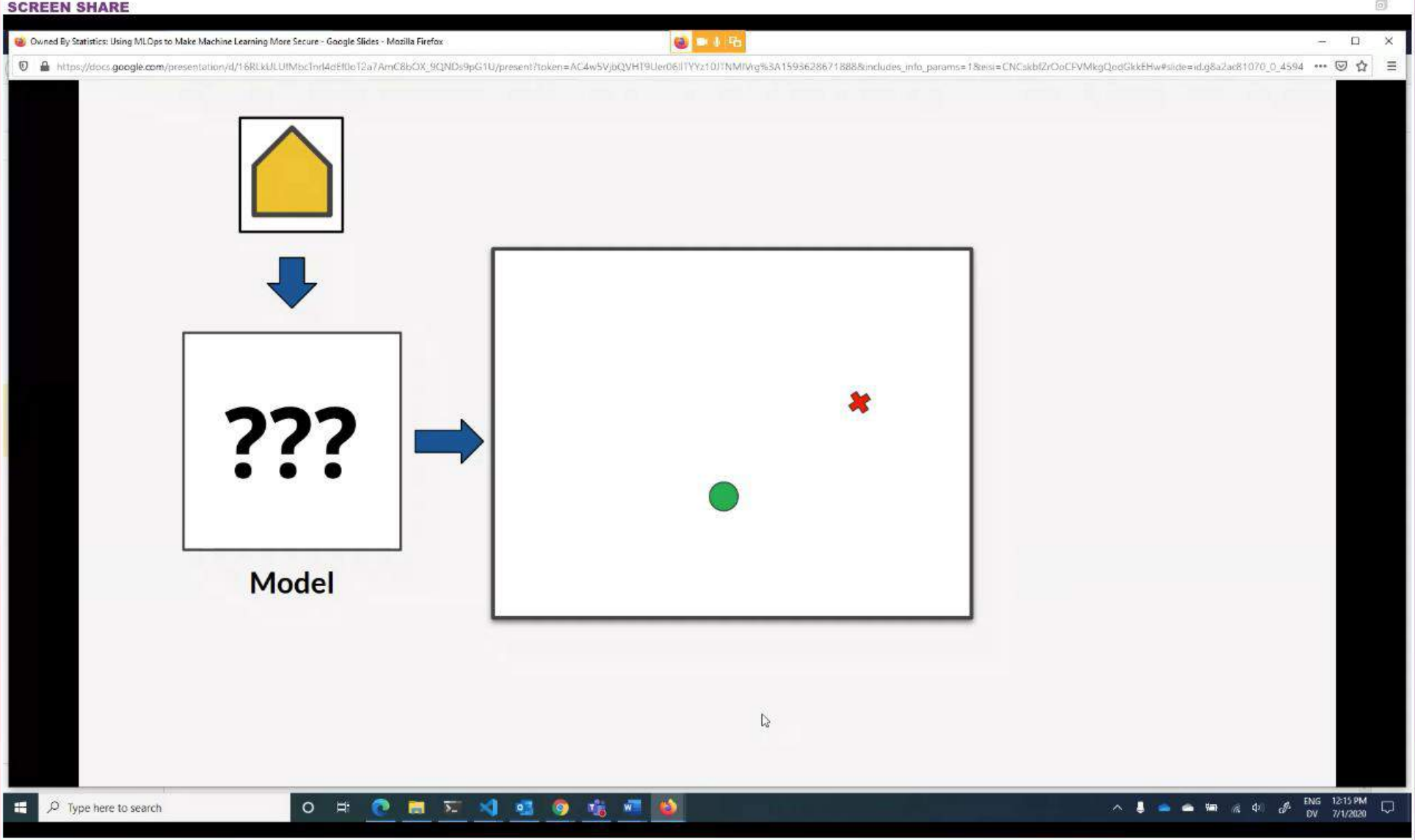
Two Main Avenues (to date)

- **Distillation**
  - "High Accuracy and High Fidelity Extraction of Neural Networks" - Jagielski, Carlini, Berthelot, Kurakin, Papernot
  - Uses sampled data from same original distribution (usually)
- **Model Extraction**
  - 'Thieves on Sesame Street! Model Extraction of BERT-based APIs' - Krishna, Tomar, Parikh, Papernot and Iyyer
  - Targets BERT style transformer models

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

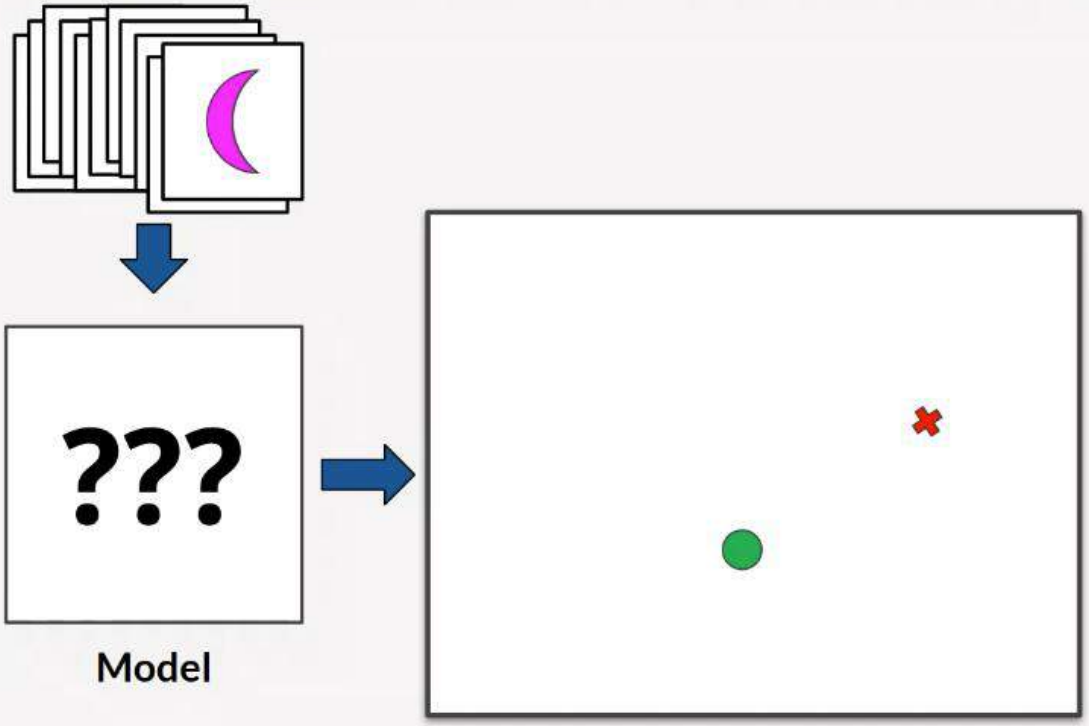


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

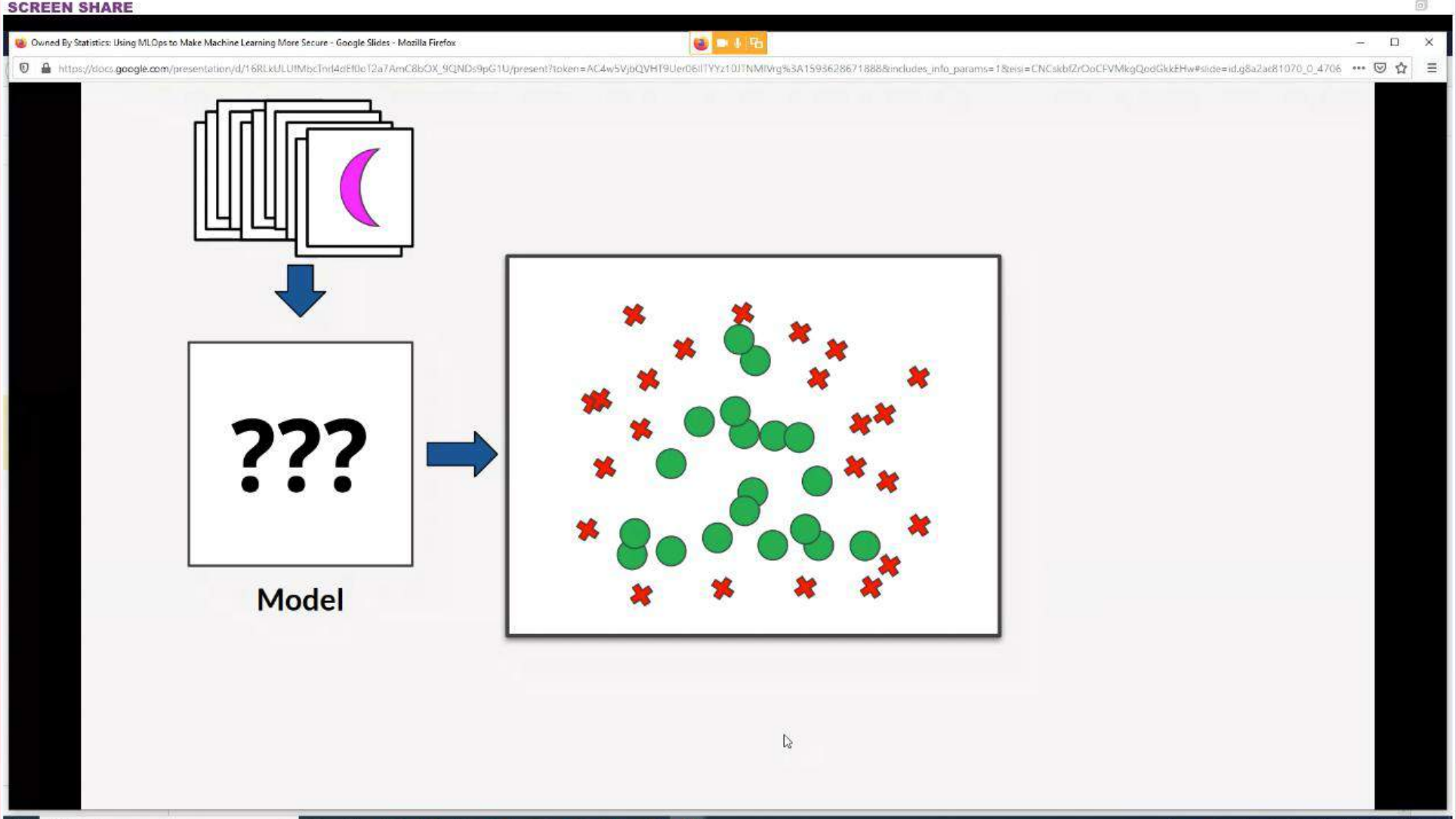


Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

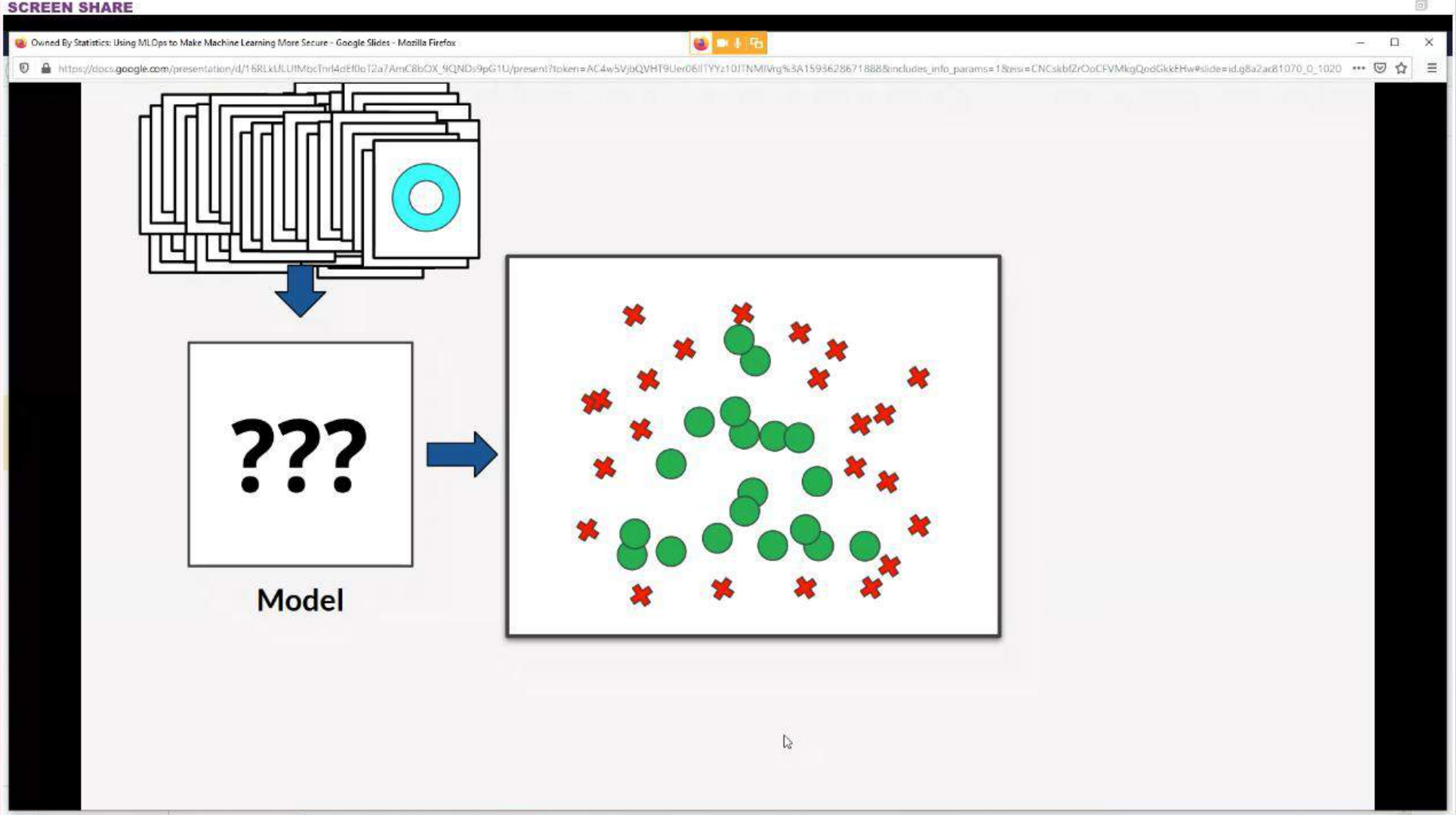
SCREEN SHARE



Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

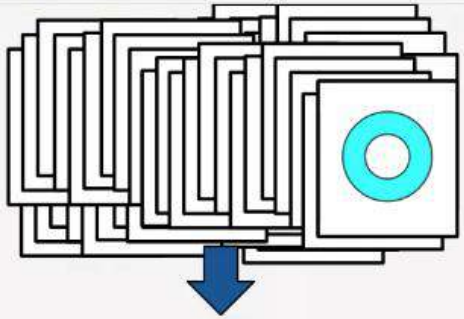


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

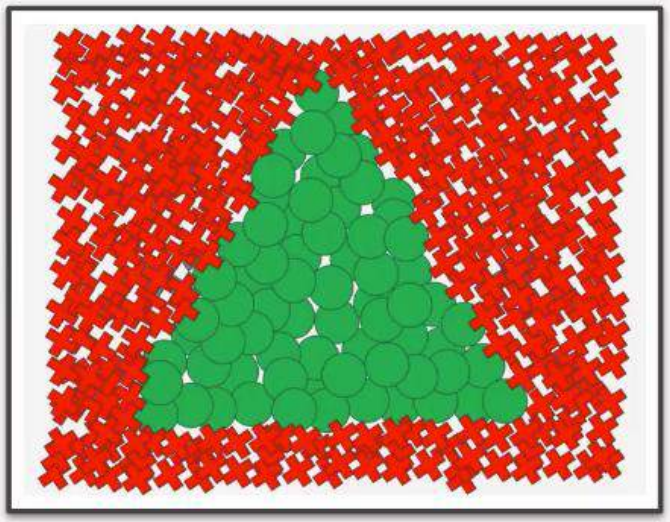
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox



Model



Type here to search

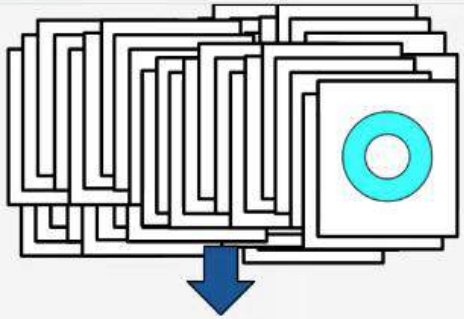
ENG DV 12:15 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

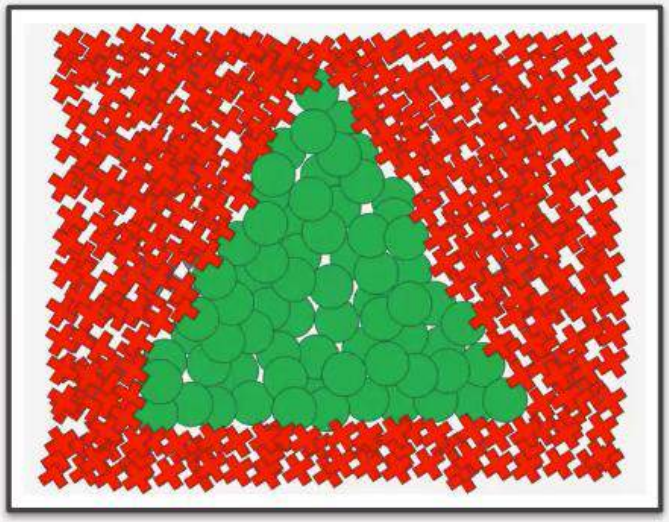
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox



Model



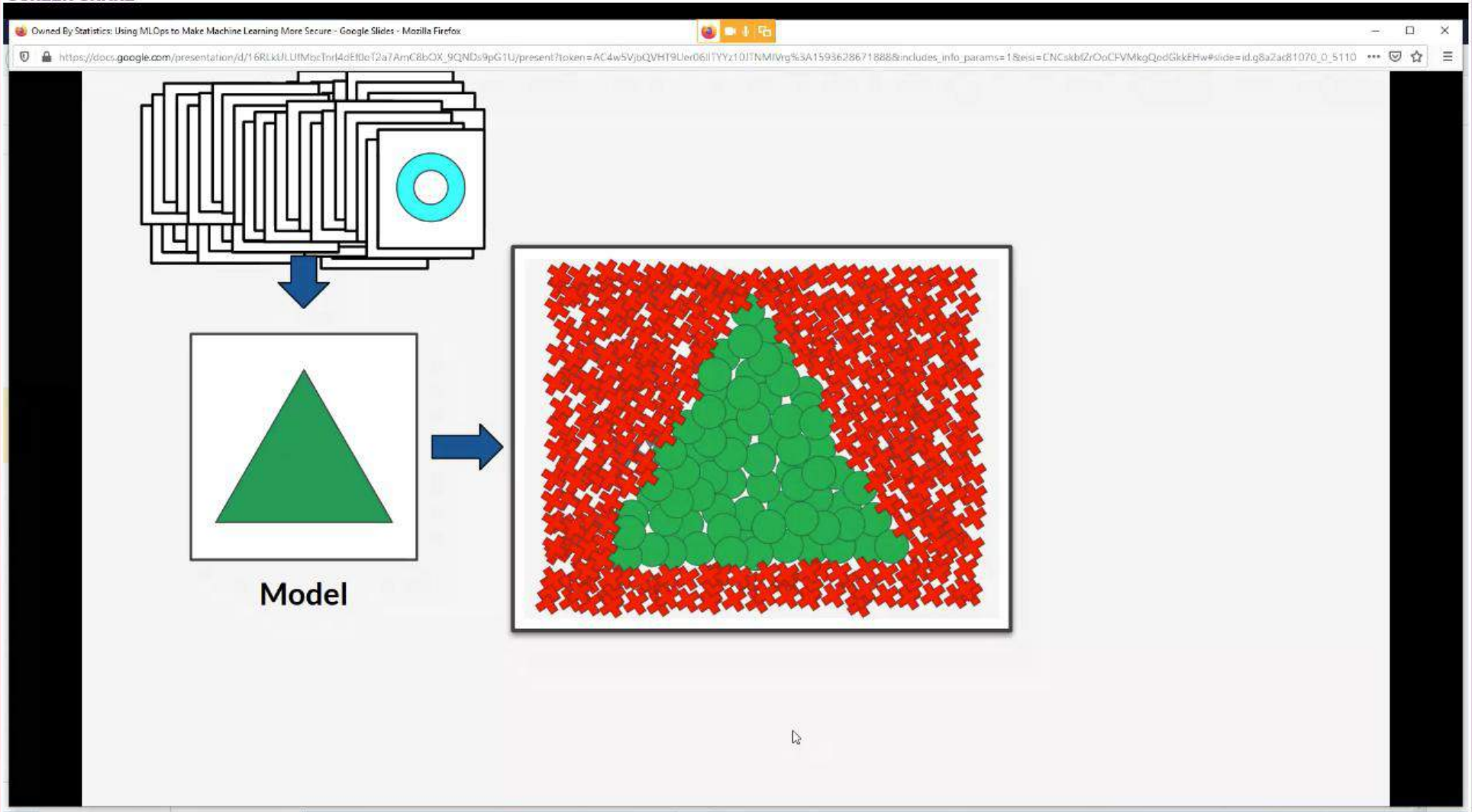
Type here to search

ENG DV 12:15 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

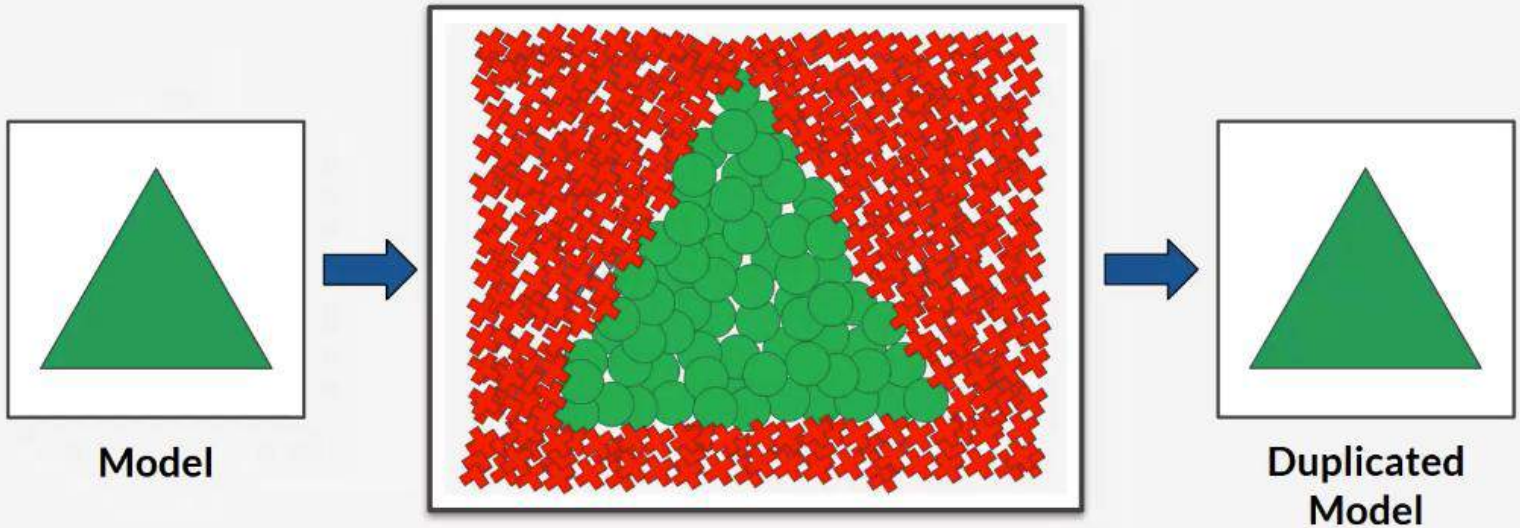


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox



Type here to search

ENG 12:15 PM 7/1/2020

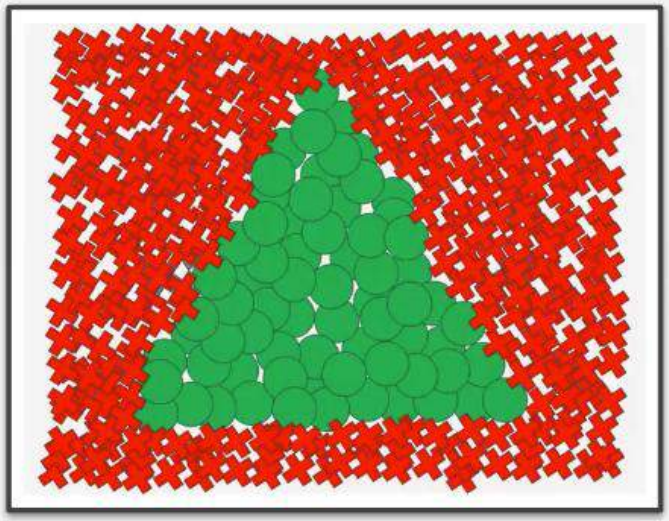
Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

Model



Duplicated Model

Queries to Get To 99% Accuracy = ???????

Type here to search

ENG DV 12:16 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

Service	Model Type	Data set	Queries	Time (s)
Amazon	Lo Regression	Digits	650	70
	Log Regression	Adult	1,485	149
B	Decision Tree	Germa dit	1,150	1
	Decision Tree	Steak S y	4,013	8

Slide 69

Q&A Notes Pointer Tips

Type here to search

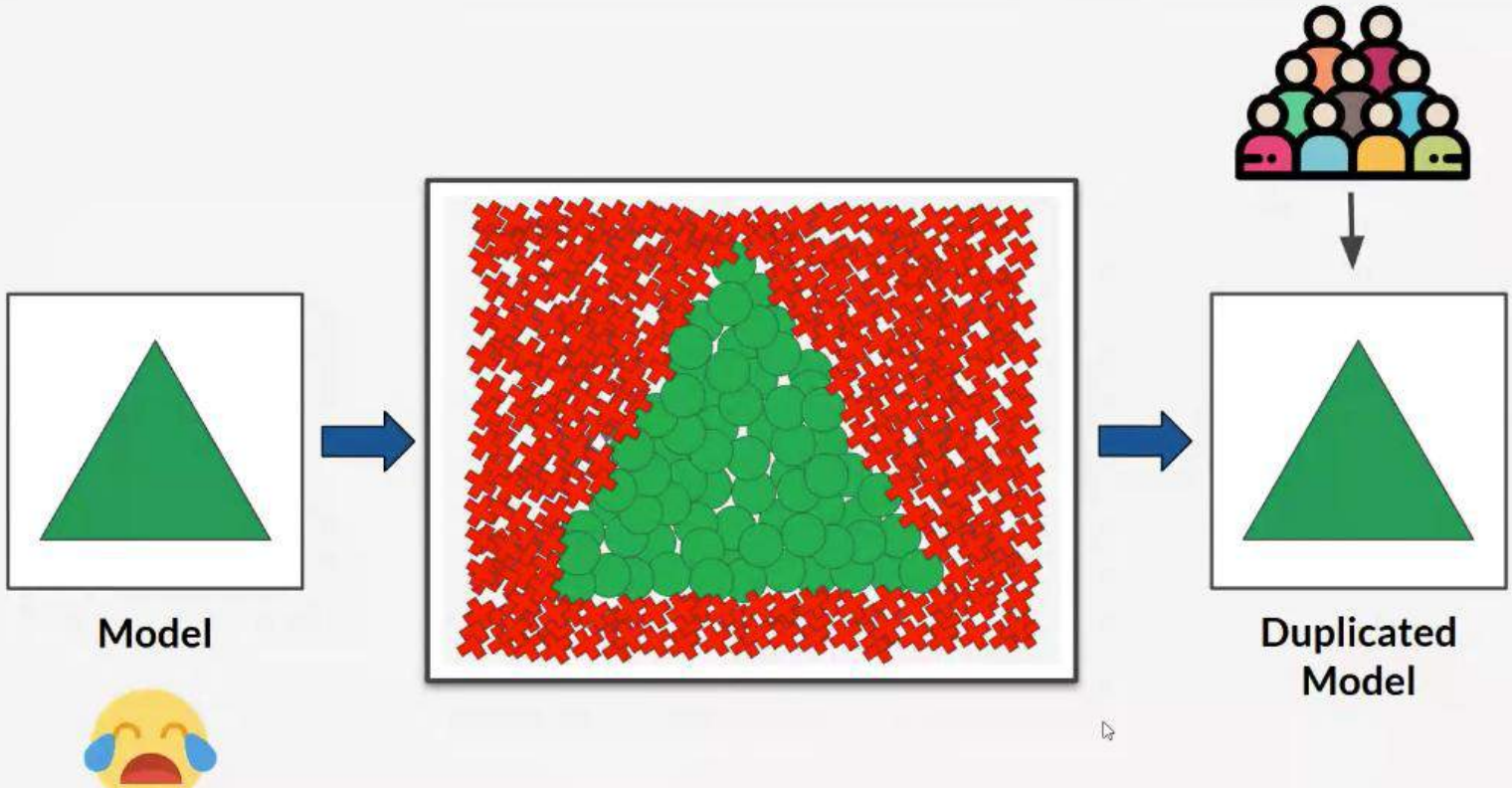
ENG DV 12:16 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

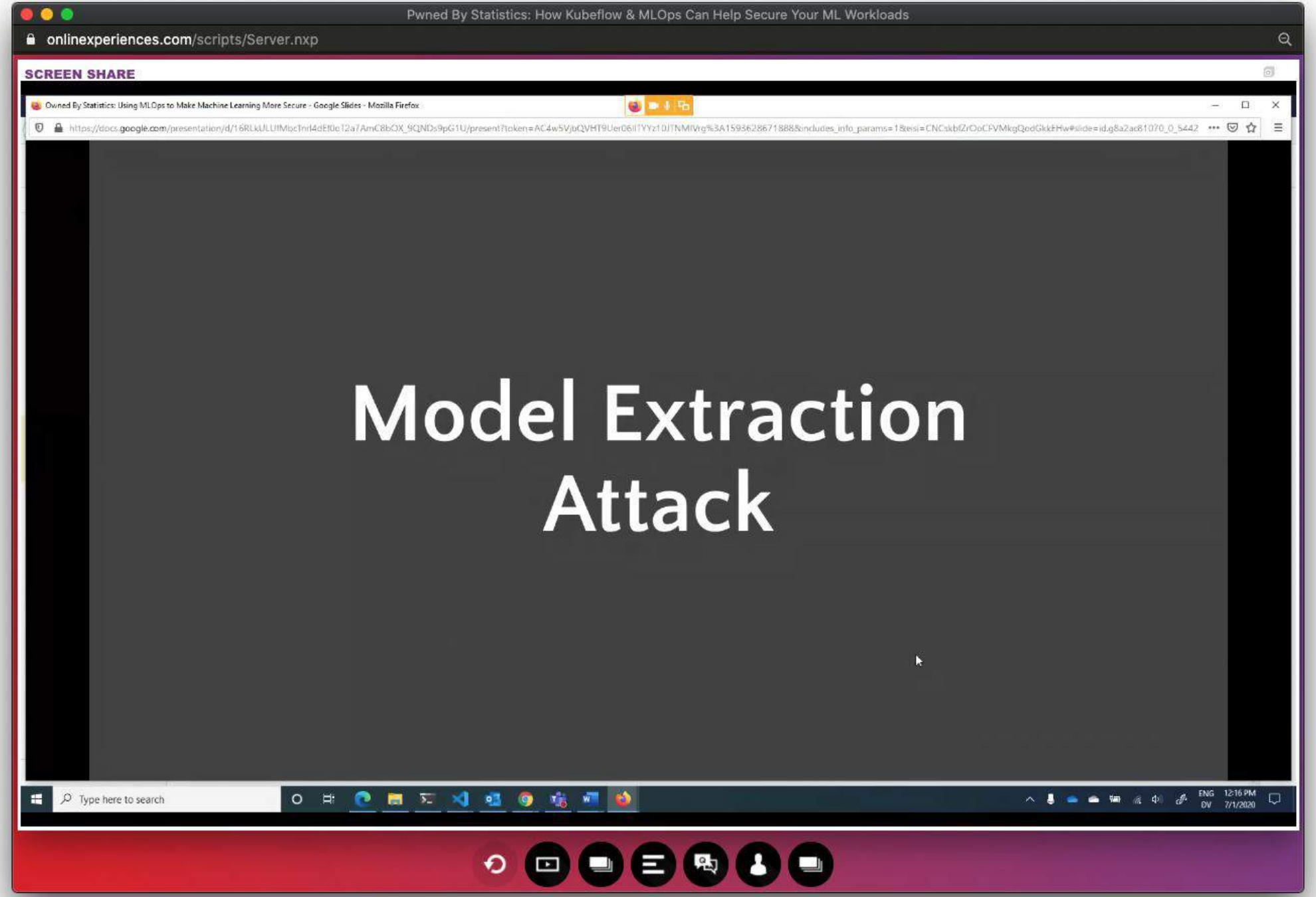
SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox



Queries to Get To 99% Accuracy = < 5000

Type here to search


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Model Extraction Attack
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkkJLUIMocInrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06lITYYz10JTNMfVig%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5459

Language Models

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Devlin, Chang, Lee, Toutanova

Type here to search

ENG DV 12:16 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMscTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9UerO6ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5698

Language Models

Azure Cognitive Services

- Decision
- Language
- Speech
- Vision
- Web search

Extract meaning from unstructured text

Immersive Reader (preview)

Help readers of all abilities comprehend text using audio and visual cues.

Language Understanding

Build natural language understanding into apps, bots, and IoT devices.

QnA Maker

Create a conversational question and answer layer over your data.

Text Analytics

Detect sentiment, key phrases, and named entities.

Translator

Detect and translate more than 60 supported languages.

Type here to search

ENG DV 12:17 PM 7/1/2020

SQuaD Tests

- SQuAD = **S**tanford **Qu**estion **A**nswering **D**ataset
  - Reading comprehension dataset
  - Questions posed against Wikipedia articles
  - Answer is question is a segment of text, or span (or unanswerable).

Wikipedia

Journalist Nik Cohn described him as "rock's greatest ever natural talent". His singing abilities encompassed a wide range from falsetto to baritone and rapid, seemingly effortless shifts of register. Prince was renowned as a multi-instrumentalist. He is considered a guitar virtuoso, a master of drums, percussion, bass, keyboards, and synthesizer. On his first 5 albums, he played nearly all the instruments, including 27 instruments on his debut album, among them various types of bass, keyboards and synthesizers.

Q: How many instruments did Prince play?
A: 27.

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMpcTnrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06iITYYz1UITNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5676

The Attack

THIEVES ON SESAME STREET! MODEL EXTRACTION OF BERT-BASED APIS

Kalpesh Krishna*
CICS, UMass Amherst
kalpesh@cs.umass.edu

Gaurav Singh Tomar
Google Research
gtomar@google.com

Ankur P. Parikh
Google Research
aparikh@google.com

Nicolas Papernot
Google Research
papernot@google.com

Mohit Iyyer
CICS, UMass Amherst
miyyer@cs.umass.edu

High Accuracy and High Fidelity Extraction of Neural Networks

Matthew Jagielski[†,*], Nicholas Carlini[*], David Berthelot[*], Alex Kurakin[*], and Nicolas Papernot[*]

[†]Northeastern University
[*]Google Research

Type here to search

ENG 12:18 PM
DV 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0oT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5658

The Attack

Wikipedia

Journalist Nik Cohn described him as "rock's greatest ever natural talent". His singing abilities encompassed a wide range from falsetto to baritone and rapid, seemingly effortless shifts of register. Prince was renowned as a multi-instrumentalist. He is considered a guitar virtuoso, a master of drums, percussion, bass, keyboards, and synthesizer. On his first 5 albums, he played nearly all the instruments, including 27 instruments on his debut album, among them various types of bass, keyboards and synthesizers.

RANDOM

Q: How workforce. Stop who new of Jordan et Wood, displayed the?

A: His singing abilities encompassed a wide range.

Type here to search

ENG DV 12:19 PM 7/1/2020

The Attack

Task	Model	0.1x	0.2x	0.5x	1x	2x	5x	10x
SST2	VICTIM	90.4	92.1	92.5	93.1	-	-	-
	RANDOM	75.9	87.5	89.0	90.1	90.5	90.4	90.1
(1x = 67349)	WIKI	89.6	90.6	91.7	91.4	91.6	91.2	91.4
MNLI	VICTIM	81.9	83.1	85.1	85.8	-	-	-
	RANDOM	59.1	70.6	75.7	76.3	77.5	78.5	77.6
(1x = 392702)	WIKI	68.0	71.6	75.9	77.8	78.9	79.7	79.3
SQuAD 1.1	VICTIM	84.1	86.6	89.0	90.6	-	-	-
	RANDOM	60.6	68.5	75.8	79.1	81.9	84.8	85.8
(1x = 87599)	WIKI	72.4	79.6	83.8	86.1	87.4	88.4	89.4
BoolQ	VICTIM	63.3	64.6	69.9	76.1	-	-	-
(1x = 9427)	WIKI	62.1	63.1	64.7	66.8	67.6	69.8	70.3

The Cost

- **$62.35** => sentiments on 67,000 sentences
- **$430.56** => speech recognition dataset of 300 hours
- **$2,000** => 1M translation queries (each with 100 characters)

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkkULUfMpcTnrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5362

Using MLOps to Defend

- **Possible defenses**
  - Detecting queries that could be part of an attack
  - Watermarking predictions made by the API
- **REPEAT: The *pipeline* is the value not the *model***
  - Improving domain specificity
  - Continuous retraining for accuracy
  - Faster throughput & SLA
- **If you REALLY need model security, treat accessing your model like accessing source code**

Realistically, if you allow arbitrary access to a model endpoint, it WILL be stolen

(if it's worth it)

Type here to search

ENG 12:21 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

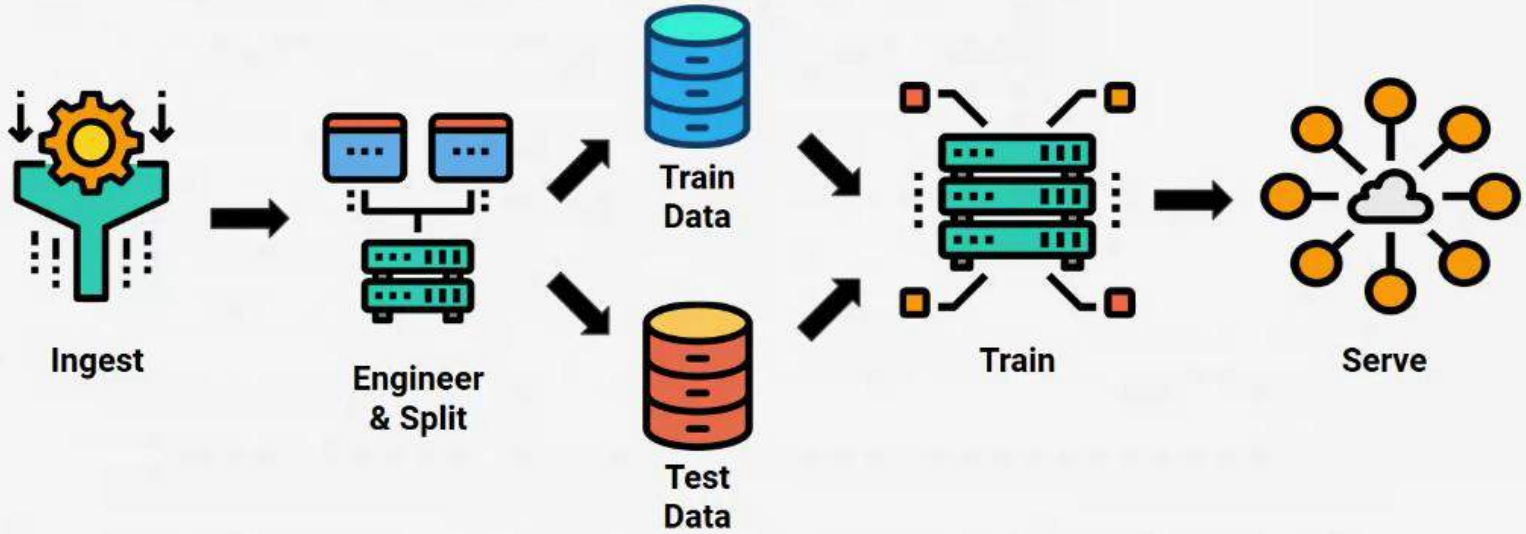
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkFHw#slide=id.g8af57be0d7_0_152

Tools to Defend

Spend the Majority of Engineering Here

Not Here



Type here to search

ENG 12:21 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

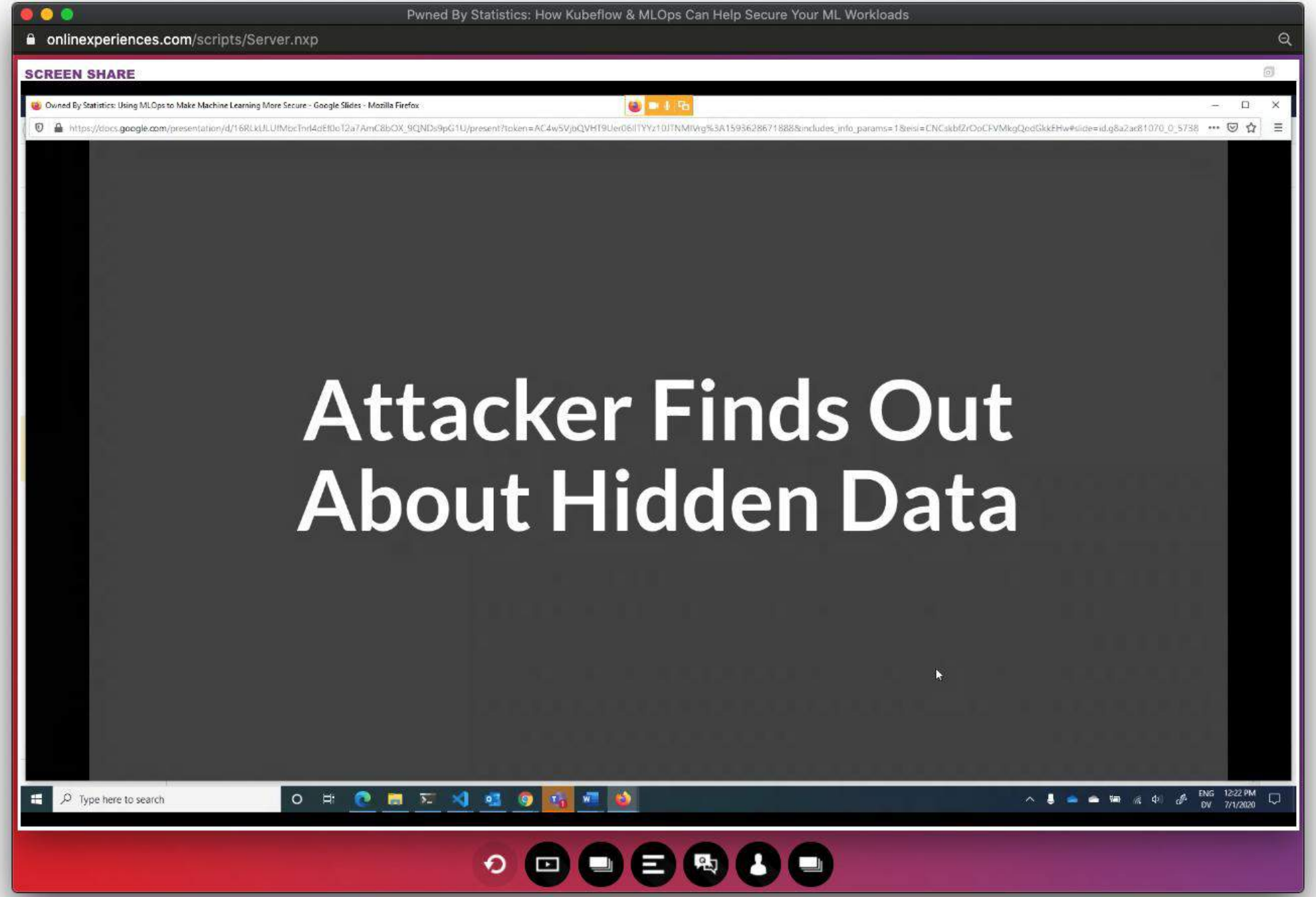
https://docs.google.com/presentation/d/16RLkkJLUIMbcTnrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMfVtg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5733

Three Types of Attacks We'll Talk About Today

1. Attacker Gets Your ML to Lie To You
2. Attacker Takes Your Models
3. Attacker Finds Out About Hidden Data

Type here to search

ENG 12:22 PM 7/1/2020


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Attacker Finds Out
About Hidden Data
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMbcTnrl4dEf0oT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&esi=CNCskbiZrOoCFVMkgQodGkkEHw#slide=id.g8a2ac81070_0_5768

Motivation

- Malicious user wants to find out hidden data
  - Can be for system or users information
  - Probes model using public endpoints
  - Does NOT have to be logged in (but it helps)
- Tough to defend against - looks VERY similar to user behavior
- **You were probably already having this problem, it just became obfuscated (more) by ML**

Type here to search

ENG DV 12:22 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

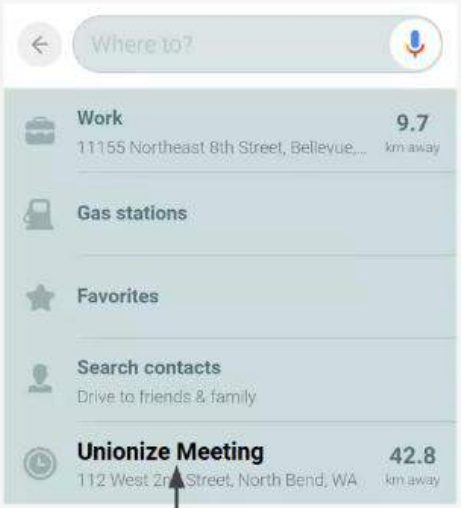
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

Hidden Data Leakage Examples

Recommendations



My Historical Events

Network Graph

My Friend's Graphs

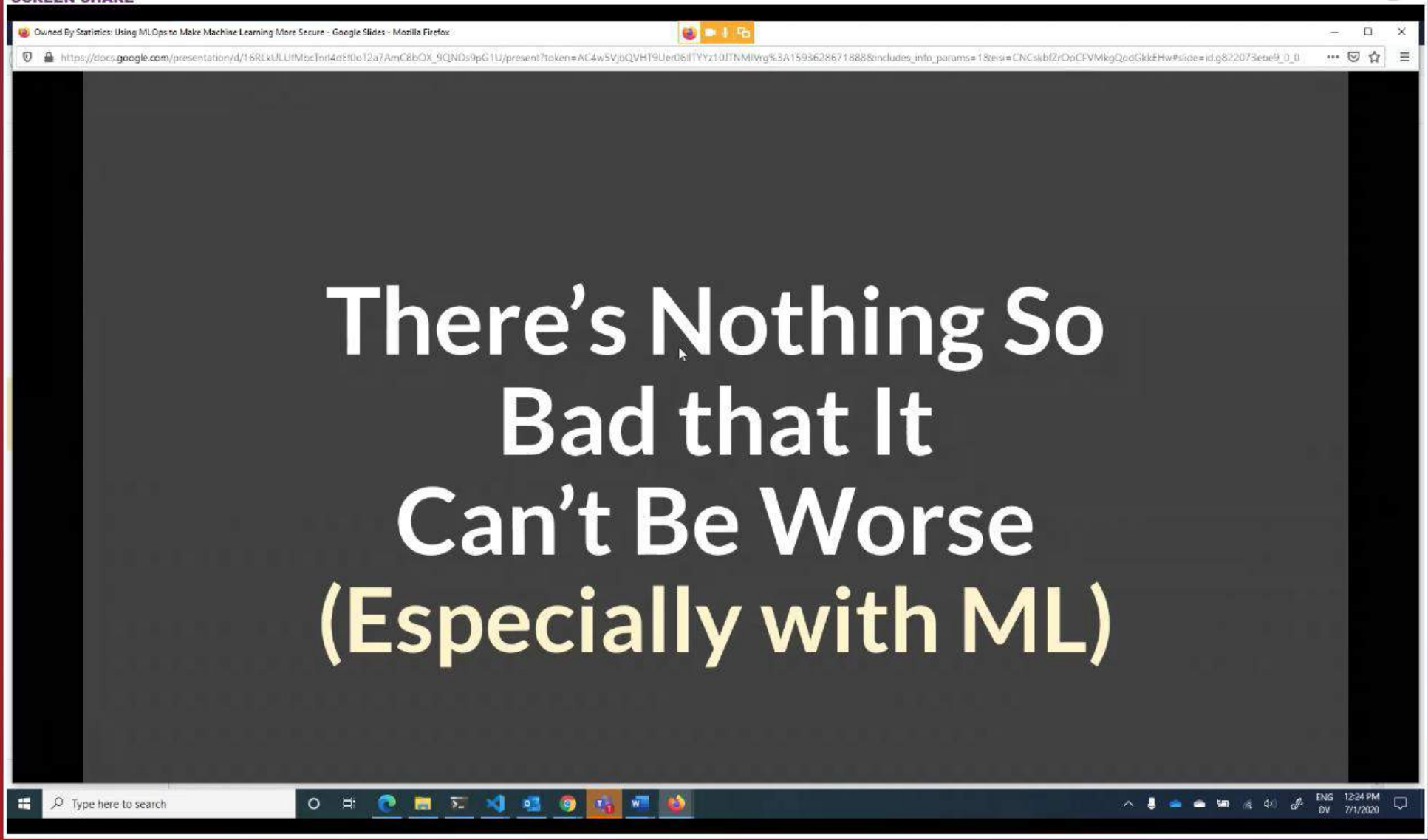
Maps

The Community's Locations

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE



Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

Secret Memorization

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Secret Memorization

- Address: My shipping address is 1101 NE 25th St, #168, Seattle WA 98004
- Phone number: Can you call me at 212 555-1212
- Relationship info: We are planning to visit next week. My partner, Ashley, and I are...
- Credit card: Please put it on my Visa, 4128 1234 5678 9012
- SSN: My social security number is... 262-97-7277

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUfMocTnrl4dEf0cT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskblZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_218

Secret Memorization

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Nicholas Carlini[1,2] Chang Liu[2] Úlfar Erlingsson[1] Jernej Kos[3] Dawn Song[2]

[1]*Google Brain* [2]*University of California, Berkeley* [3]*National University of Singapore*

Abstract

This paper describes a testing methodology for quantitatively assessing the risk that rare or unique training-data sequences are *unintentionally memorized* by generative sequence models—a common type of machine-learning model. Because such models are sometimes trained on sensitive data (e.g., the text of users' private messages), this methodology can benefit privacy by allowing deep-learning practitioners to select means of training that minimize such memorization.

In experiments, we show that unintended memorization is a persistent, hard-to-avoid issue that can have serious consequences. Specifically, for models trained without consideration of memorization, we describe new, efficient procedures that can extract unique, secret sequences, such as credit card numbers. We show that our testing strategy is a practical and easy-to-use first line of defense, e.g., by describing its application to quantitatively limit data exposure in Google's Smart Compose, a commercial text-completion neural network trained on millions of users' email messages.

For example, users may find that the input "my social-security number is. . ." gets auto-completed to an obvious secret (such as a valid-looking SSN not their own), or find that other inputs are auto-completed to text with oddly-specific details. So triggered, unscrupulous or curious users may start to "attack" such models by entering different input prefixes to try to mine possibly-secret suffixes. Therefore, for generative text models, assessing and reducing the chances that secrets may be disclosed in this manner is a key practical concern.

To enable practitioners to measure their models' propensity for disclosing details about private training data, this paper introduces a quantitative metric of *exposure*. This metric can be applied during training as part of a testing methodology that empirically measures a model's potential for unintended memorization of unique or rare sequences in the training data.

Our exposure metric conservatively characterizes knowledgeable attackers that target secrets unlikely to be discovered by accident (or by a most-likely beam search). As validation of this, we describe an algorithm guided by the exposure met-

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks Carlini, Liu, Erlingsson, Kos, Song

Type here to search

ENG DV 12:26 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

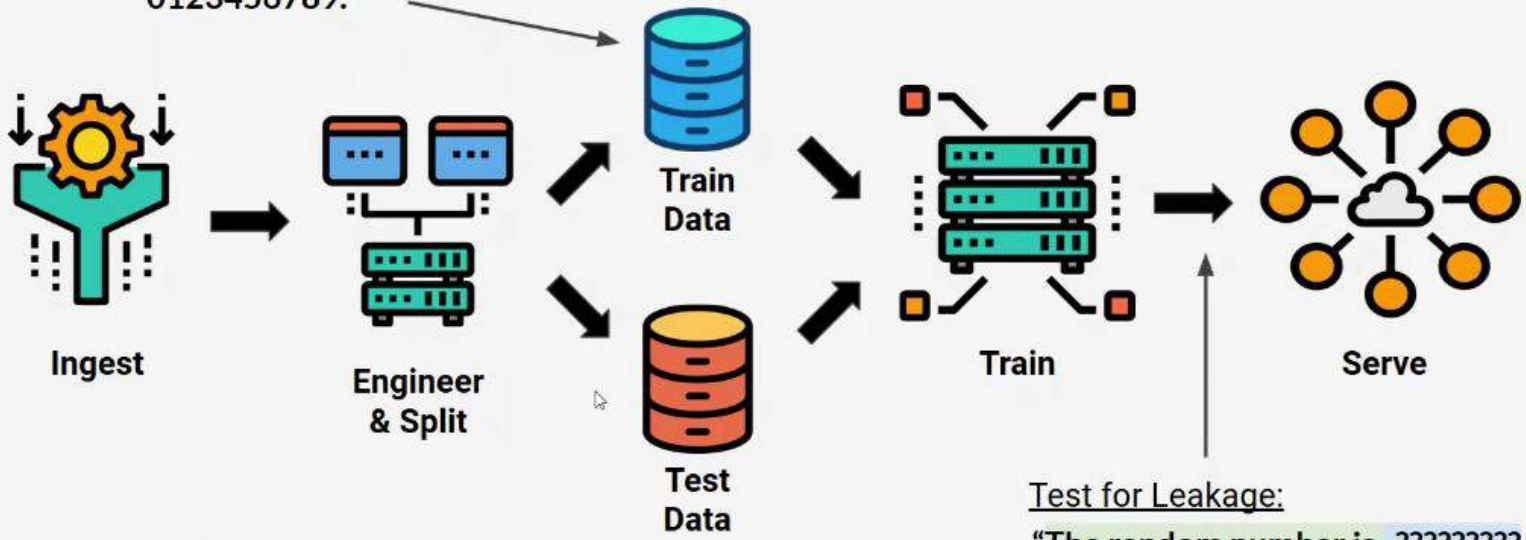
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUIMbcInrl4dEf0o12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06IITYYz10JTNMIVrg%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8a12b03078_0_26

Secret Memorization

Inject Canary:

"The random number is 0123456789."



Test for Leakage:

"The random number is ?????????."

Type here to search

ENG DV 12:27 PM 7/1/2020

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

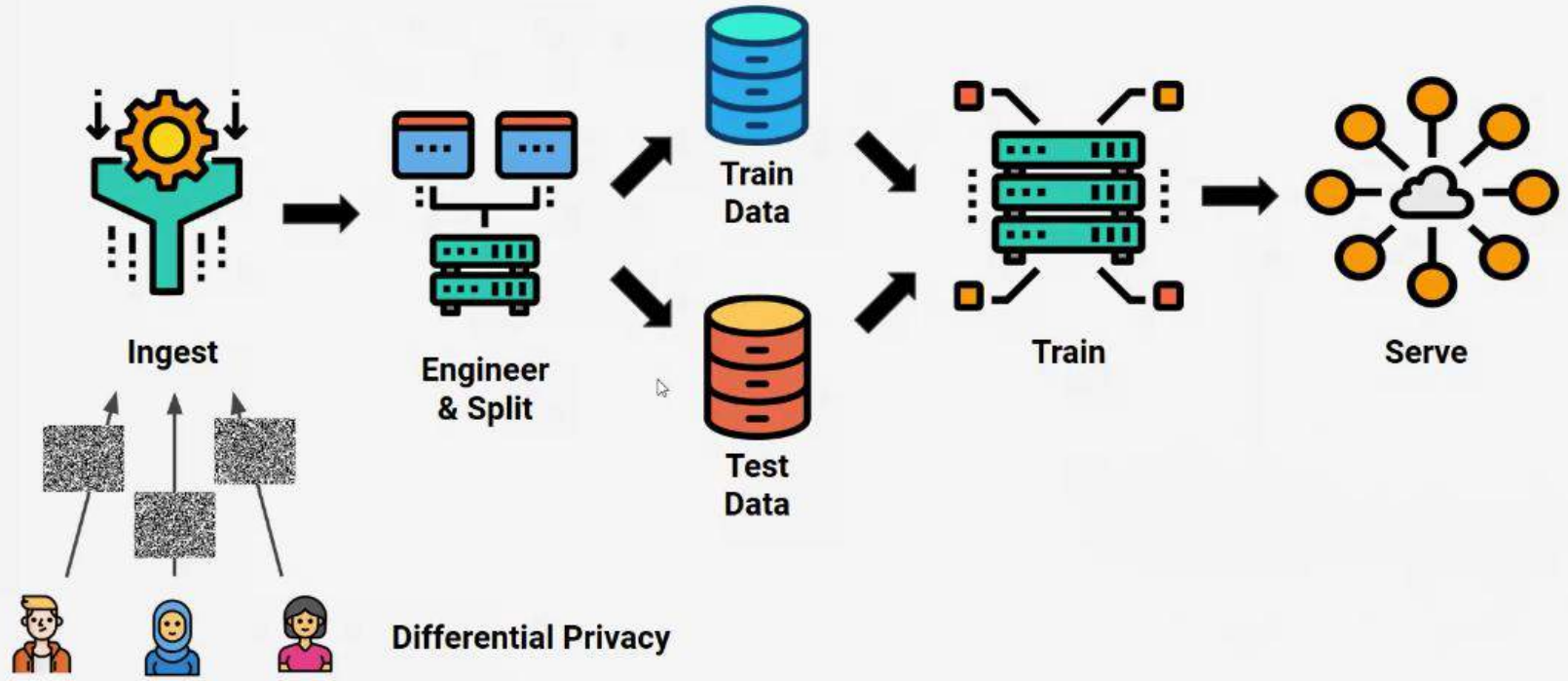
onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

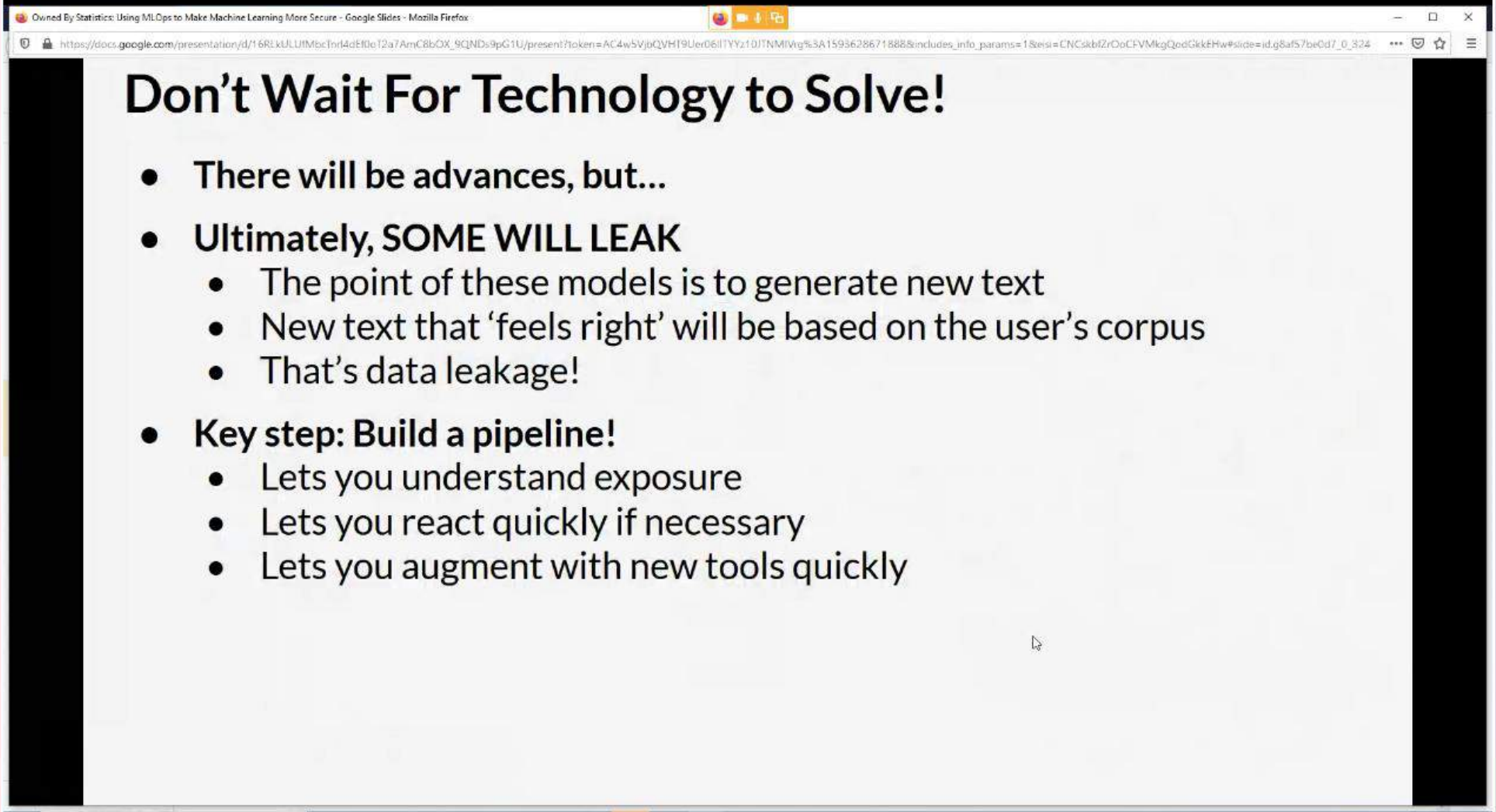
https://docs.google.com/presentation/d/16RLkULUIMbcInrl4dEf0e12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06llTYYz10JTNMIVig%3A1593628671888&includes_info_params=1&esi=CNCskbIZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_276

Secret Memorization



Type here to search

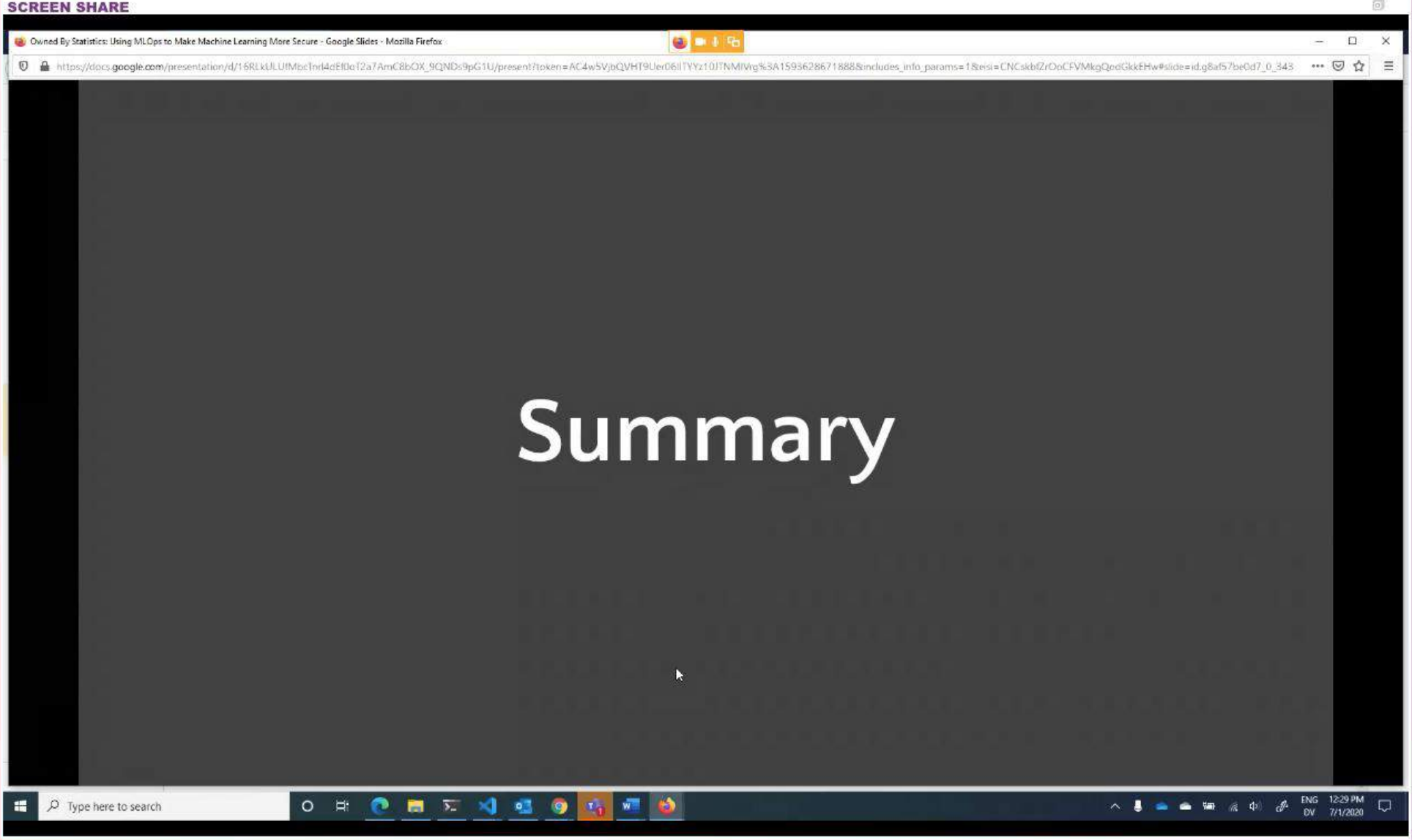
ENG DV 12:27 PM 7/1/2020

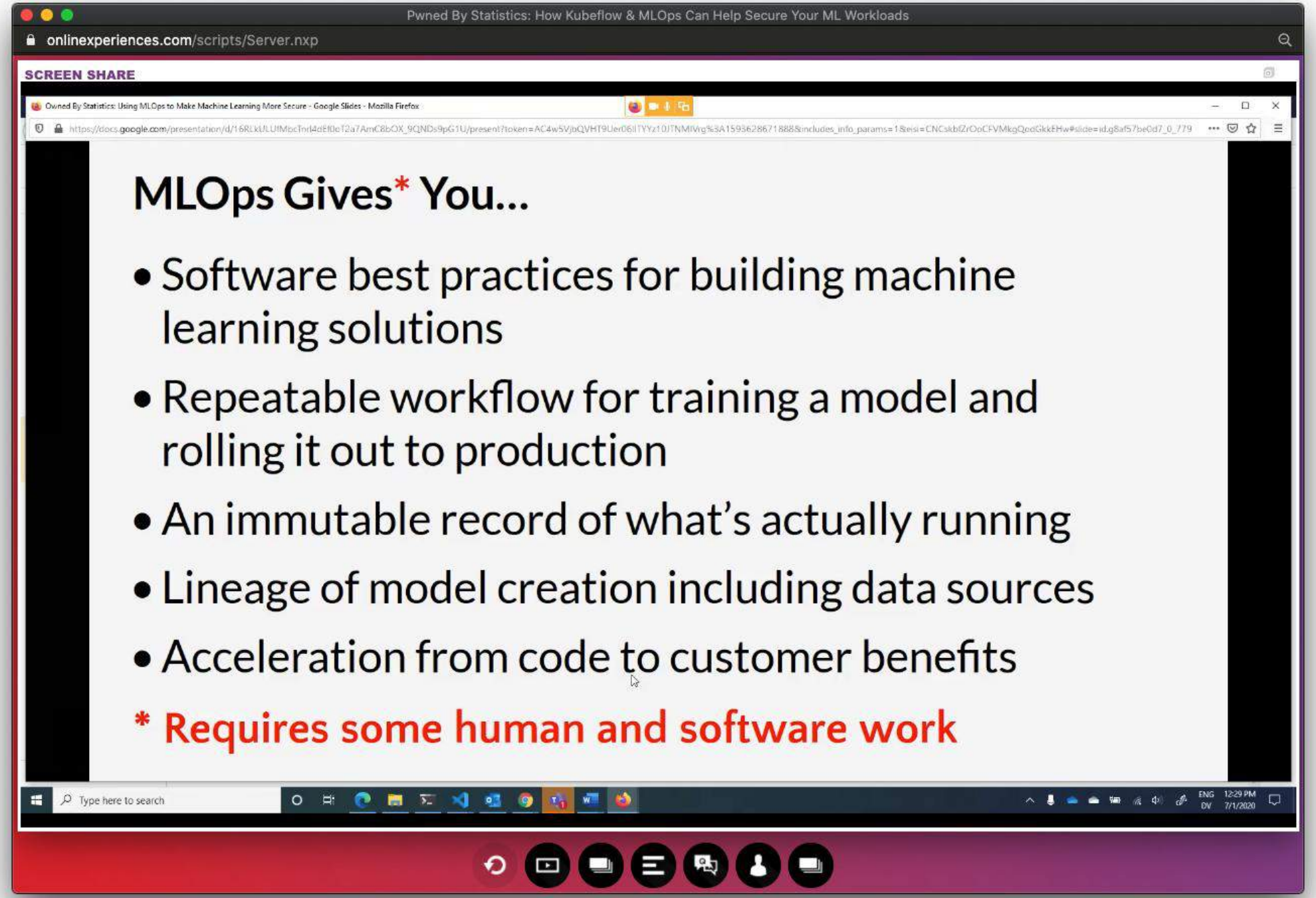

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Don't Wait For Technology to Solve!
There will be advances, but…
Ultimately, SOME WILL LEAK
The point of these models is to generate new text
New text that 'feels right' will be based on the user's corpus
That's data leakage!
Key step: Build a pipeline!
Lets you understand exposure
Lets you react quickly if necessary
Lets you augment with new tools quickly
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE




Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
MLOps Gives* You...
• Software best practices for building machine learning solutions
• Repeatable workflow for training a model and rolling it out to production
• An immutable record of what's actually running
• Lineage of model creation including data sources
• Acceleration from code to customer benefits
* Requires some human and software work
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

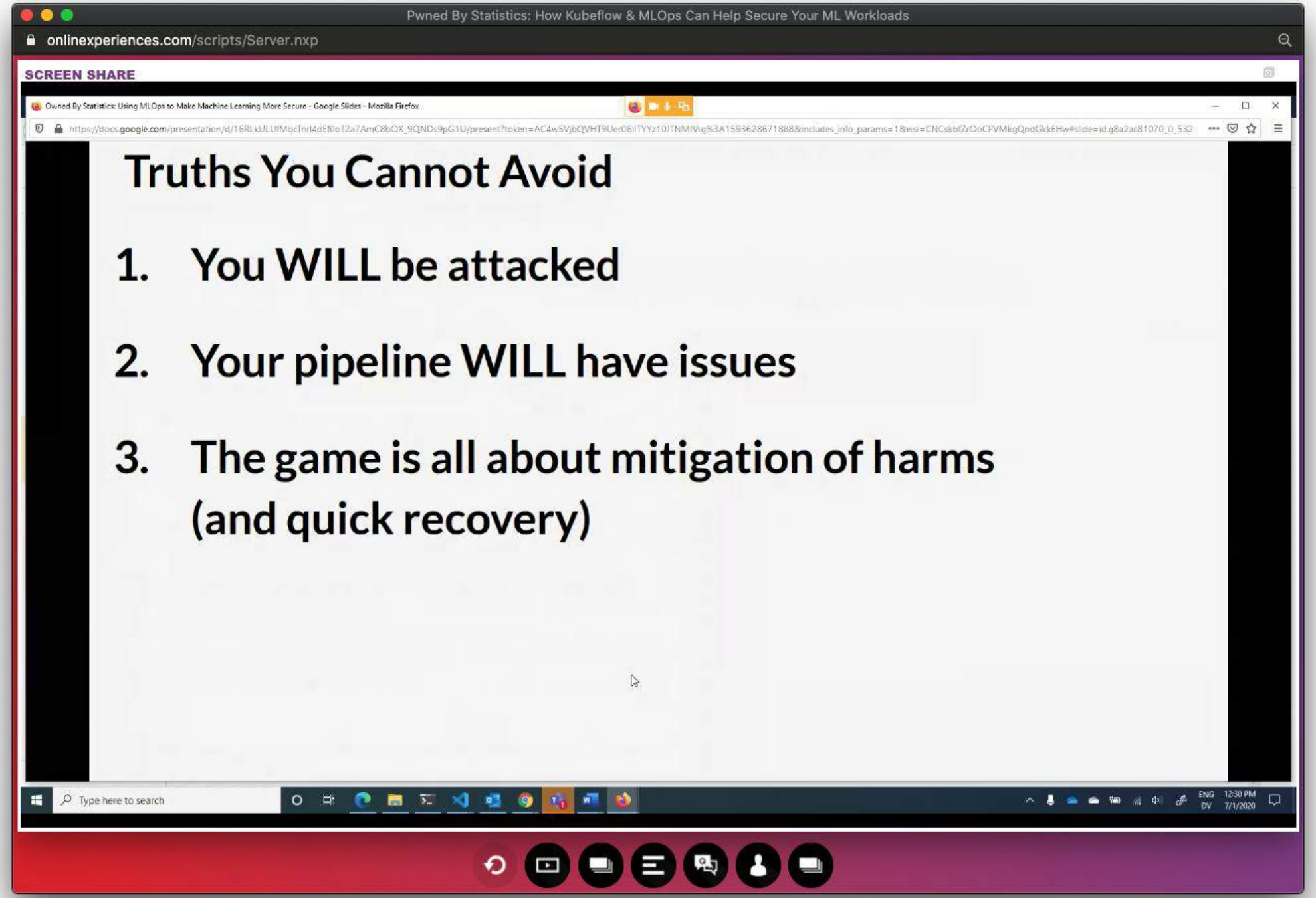
https://docs.google.com/presentation/d/16RLkULUfMscTnrl4dEf0eT2a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06lITYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_791

It's a whole new world

- Data science will touch EVERY industry.
- We can't ask people to become a PhD in statistics though.
- How do WE help <u>everyone</u> take advantage of this transformation?

Type here to search

ENG DV 12:29 PM 7/1/2020


Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads
onlinexperiences.com/scripts/Server.nxp
SCREEN SHARE
Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox
Truths You Cannot Avoid
1. You WILL be attacked
2. Your pipeline WILL have issues
3. The game is all about mitigation of harms (and quick recovery)
Type here to search

Pwned By Statistics: How Kubeflow & MLOps Can Help Secure Your ML Workloads

onlinexperiences.com/scripts/Server.nxp

SCREEN SHARE

Owned By Statistics: Using MLOps to Make Machine Learning More Secure - Google Slides - Mozilla Firefox

https://docs.google.com/presentation/d/16RLkULUlMbcTnrl4dEf0c12a7AmC8bOX_9QNDs9pG1U/present?token=AC4w5VjbQVHT9Uer06ilTYYz10JTNMfVrg%3A1593628671888&includes_info_params=1&esi=CNCskbfZrOoCFVMkgQodGkkEHw#slide=id.g8af57be0d7_0_797

me: David Aronchick (david.aronchick@microsoft.com)

twitter: @aronchick

apps: http://mlops-github.com/

- "Why Should I Trust You?" Explaining the Predictions of Any Classifier - Ribeiro, Singh, Guestrin
- Synthesizing Robust Adversarial Examples - Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok
- Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition - Sharif, Bhagavatula, Bauer, Reiter
- How To Backdoor Federated Learning - Bagdasaryan, Veit, Hua, Estrin, Shmatikov
- Learning to Detect Malicious Clients for Robust Federated Learning - Li, Cheng, Wang, Liu, Chen
- "High Accuracy and High Fidelity Extraction of Neural Networks" - Jagielski, Carlini, Berthelot, Kurakin, Papernot
- 'Thieves on Sesame Street! Model Extraction of BERT-based APIs' - Krishna, Tomar, Parikh, Papernot and Iyyer
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Devlin, Chang, Lee, Toutanova
- The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks Carlini, Liu, Erlingsson, Kos, Song

THANK YOU!

Type here to search

ENG 12:30 PM 7/1/2020