



KubeCon



CloudNativeCon

Europe 2020

*Virtual*

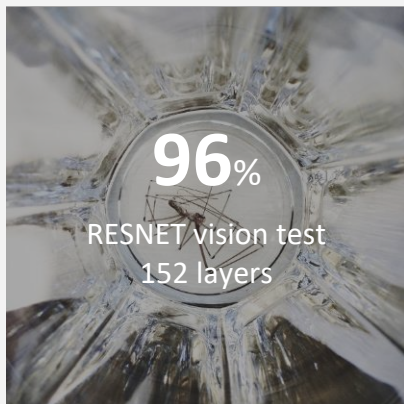
# Owned By Statistics: Attacks on Machine Learning and How to Use MLOps to Defend

**David Aronchick**

**Head of OSS Machine Learning Strategy, Azure Machine Learning  
@aronchick - he/him**

# Microsoft ML breakthroughs

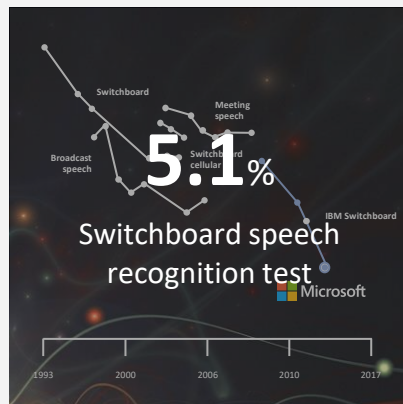
## Vision



2016

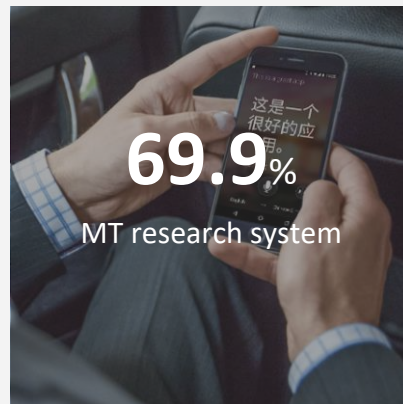
Object recognition  
Human parity

## Speech



2017

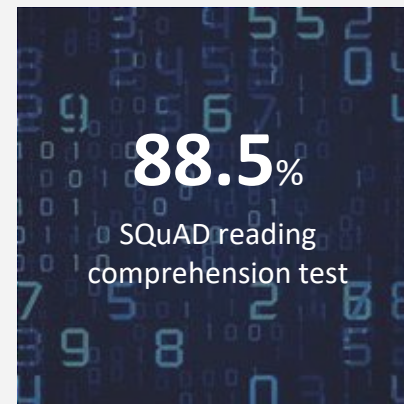
Speech recognition  
Human parity



2018

Machine translation  
Human parity

## Language



2018

Machine reading  
comprehension  
Human parity

# ML at Microsoft

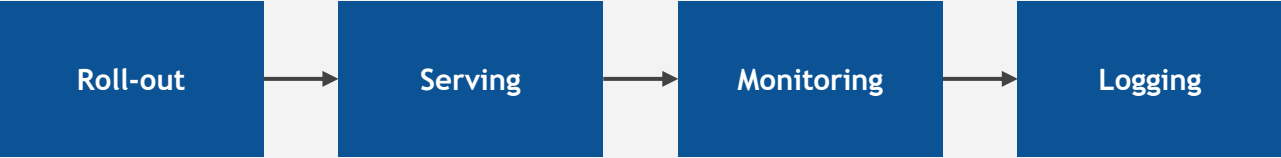
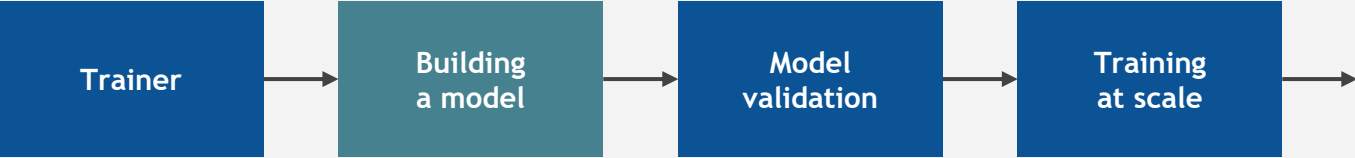
Microsoft 365





**But ML is HARD!**

# Building a model



**Ok, but, like, I'm  
a data scientist. ~~IDGAF~~  
I don't care  
about all that.**

**Yes You Do!**





**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



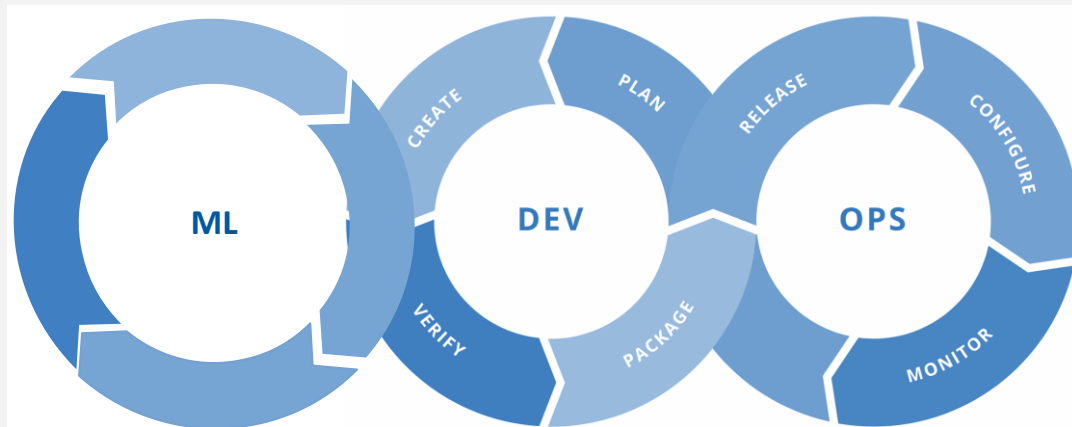
↻ 7



19

**MLOps**

# 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

**Wasn't This Talk  
Supposed to be  
About Security?**

**MLOps is The  
Baseline for  
Security**

**But... it's Just  
Math, How Bad  
Could It Be?**

# **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**

# Three Types of Attacks We'll Talk About Today

1. **Attacker Gets Your ML to Lie To You**

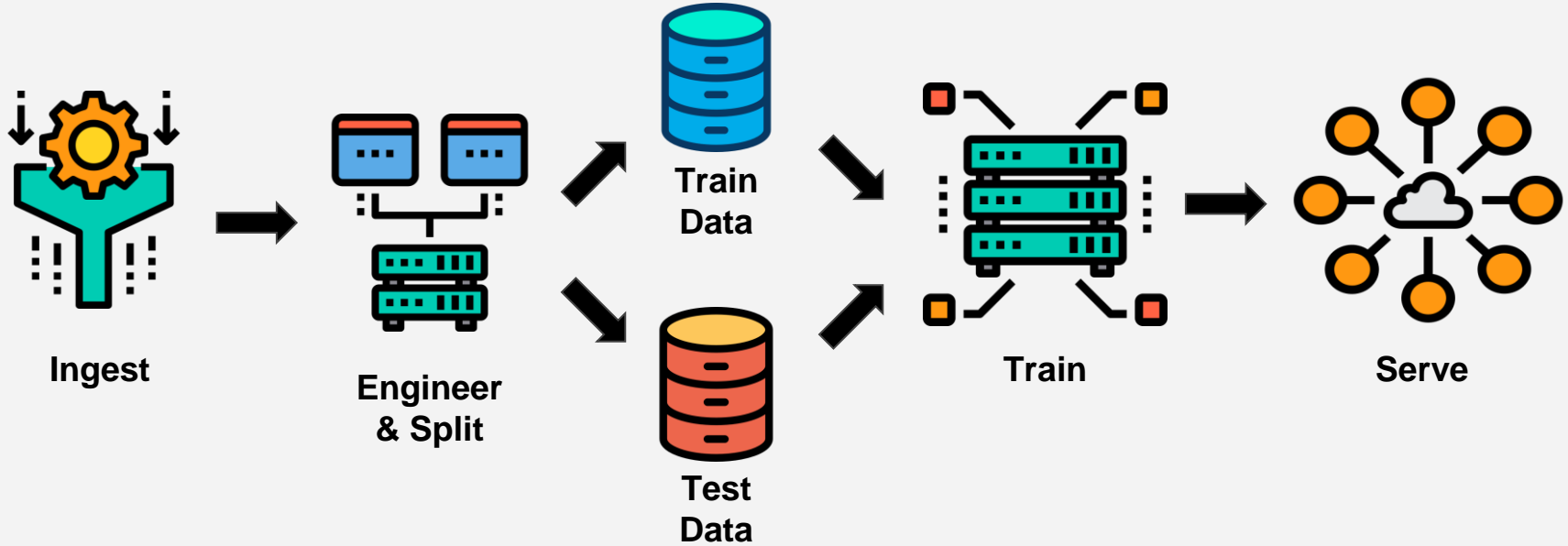
2. **Attacker Takes Your Models**

3. **Attacker Finds Out About Hidden Data**

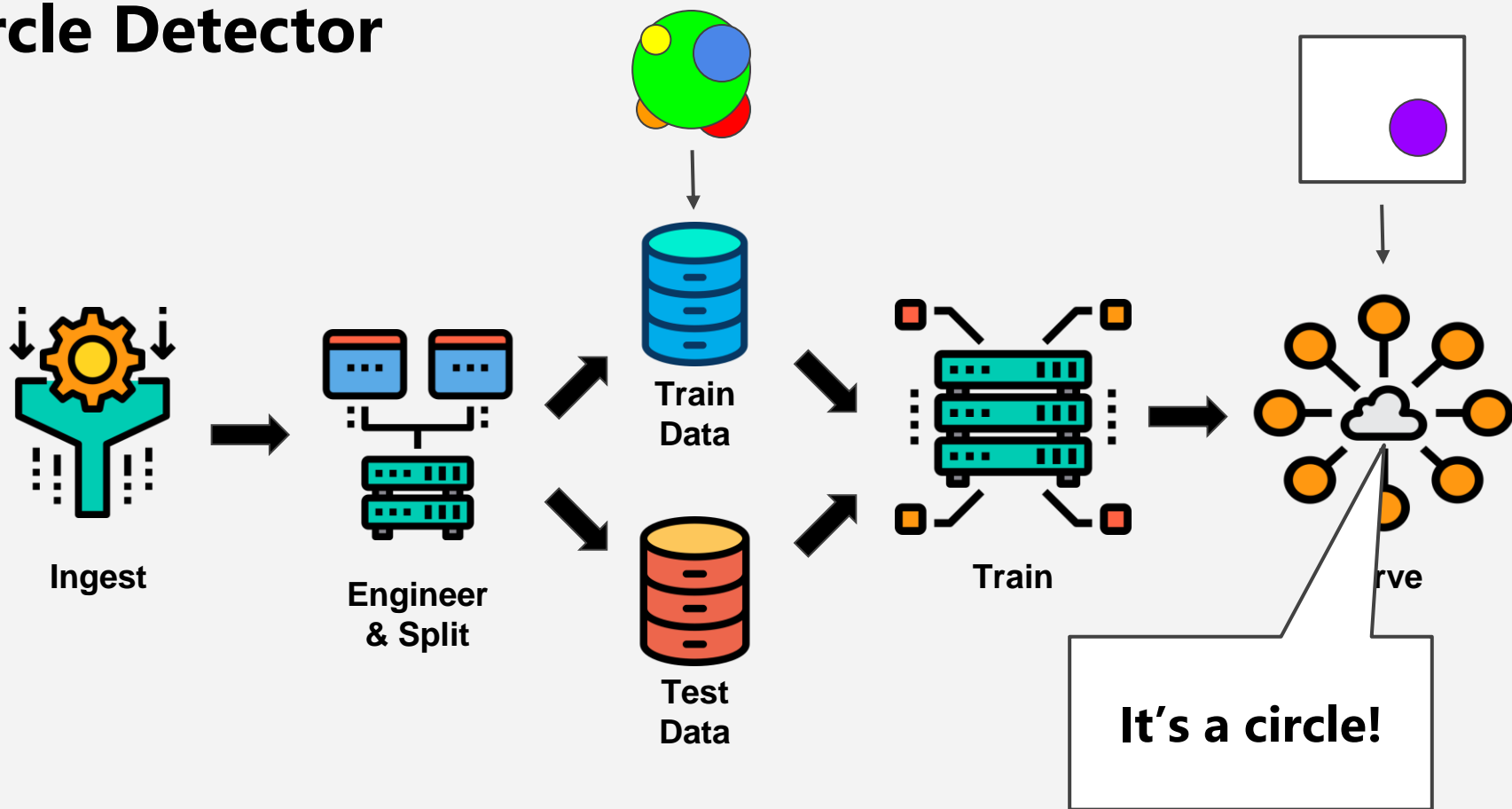


**Attacker Gets  
Your ML to Lie To  
You**

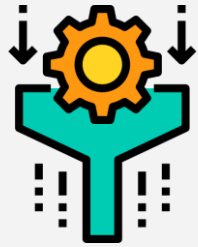
# Circle Detector



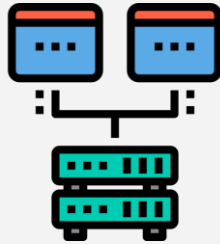
# Circle Detector



# Circle Detector



Ingest



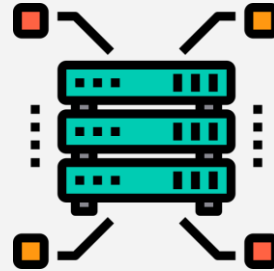
Engineer  
& Split



Train  
Data



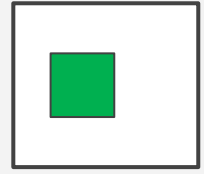
Test  
Data



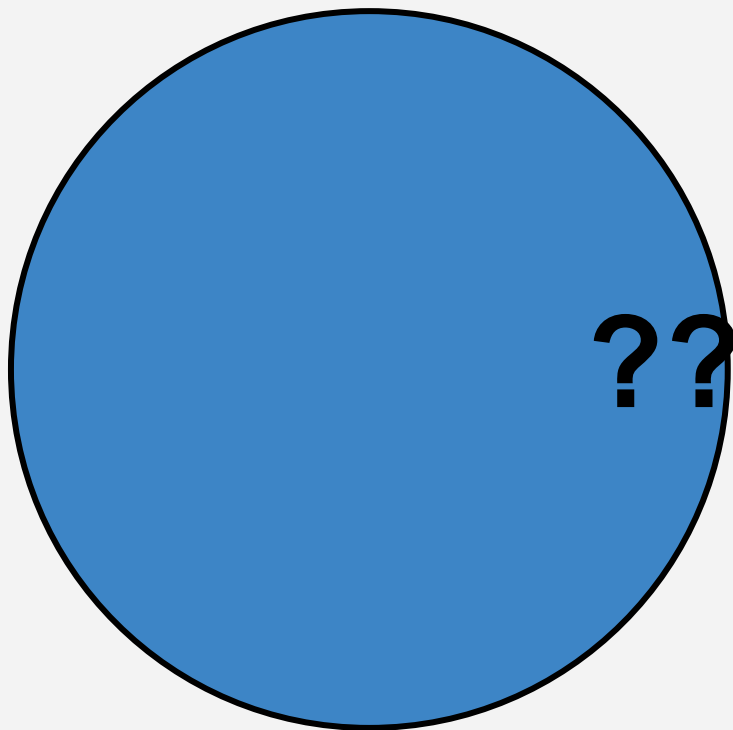
Train



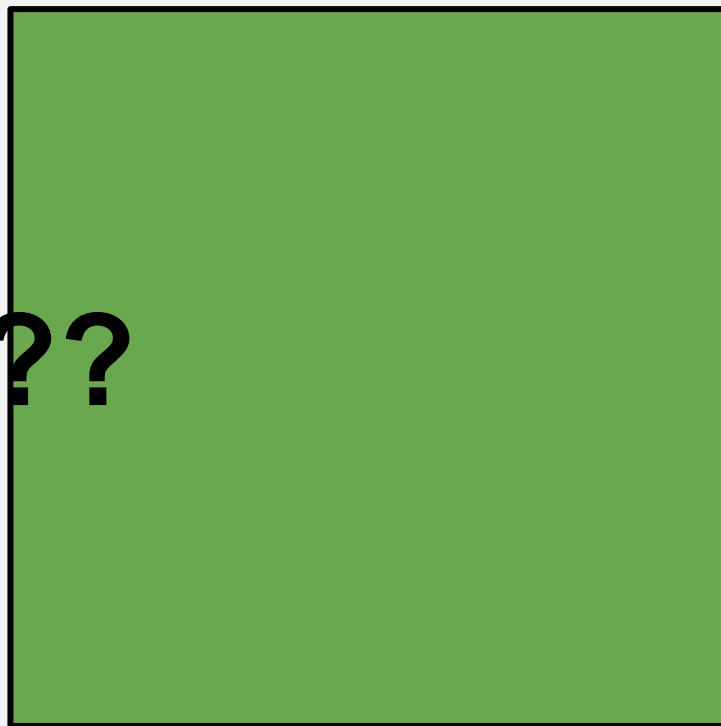
Serve

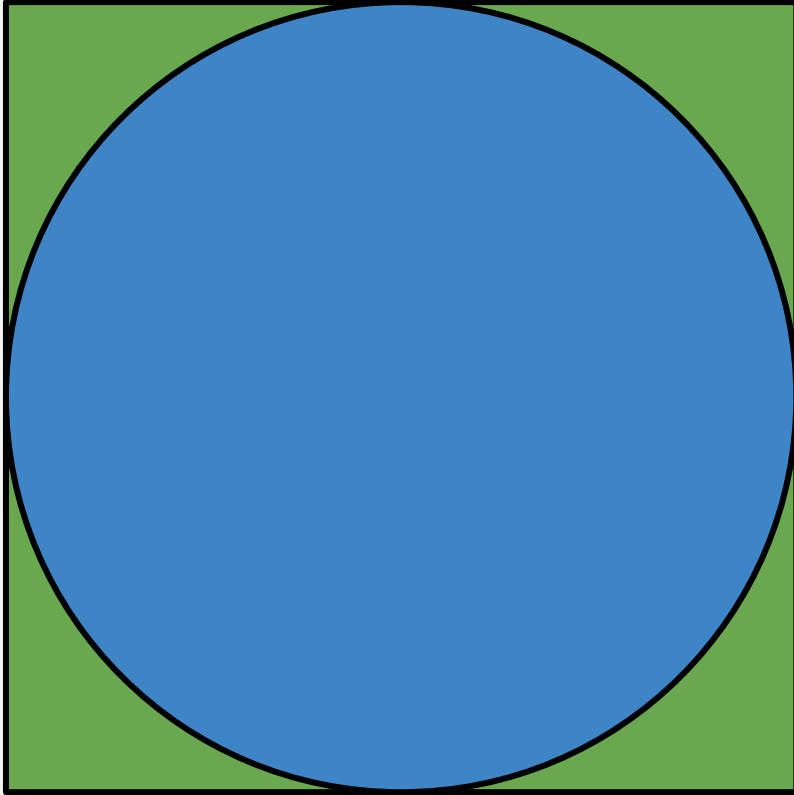


It's a circle!  
???????



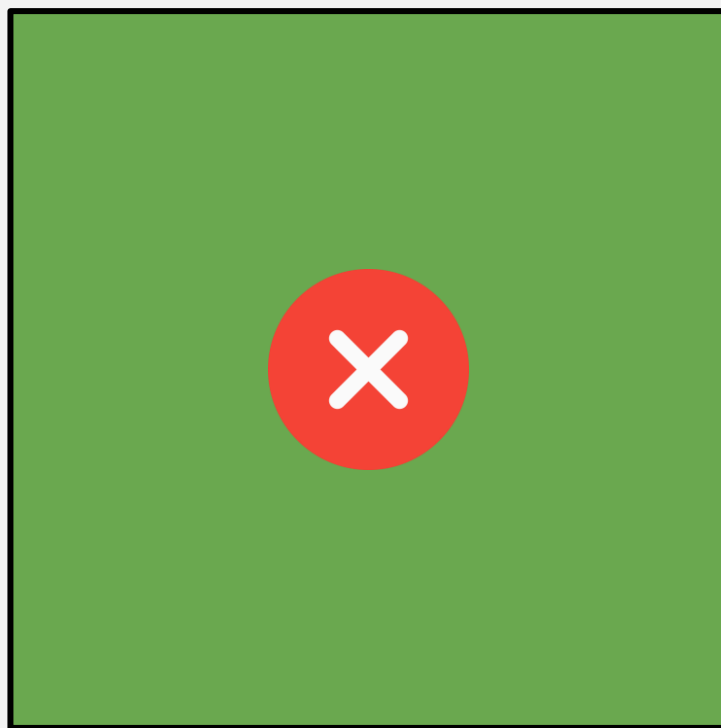
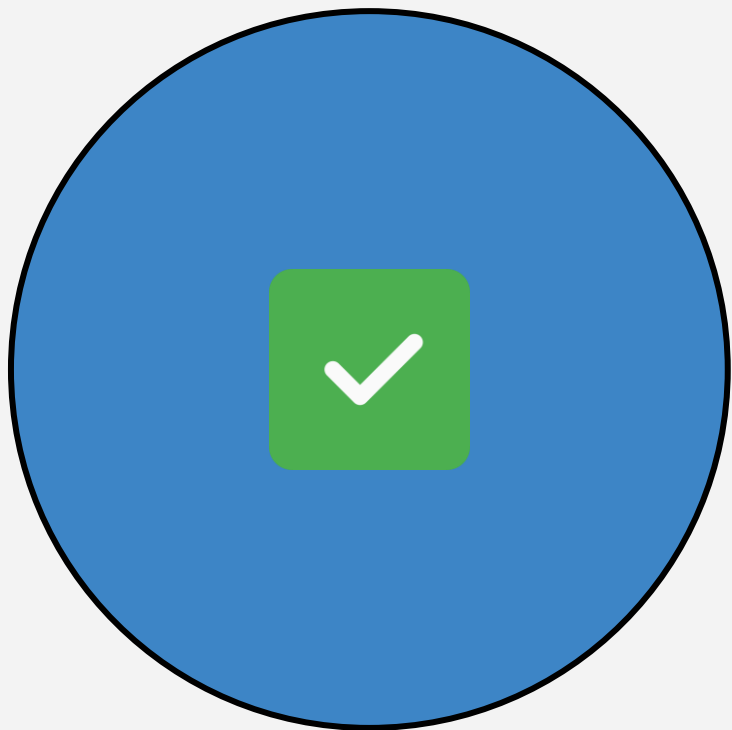
??????





**Well, that green thing  
has all the same pixels  
as a circle... it's a circle!**





**Surely, Advanced  
Models are Better  
... Right?**





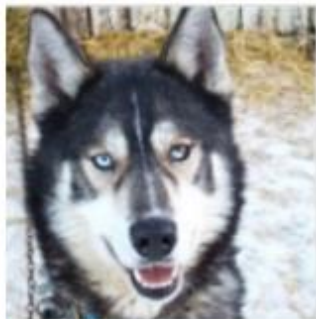
Predicted: **wolf**  
True: **wolf**



Predicted: **husky**  
True: **husky**



Predicted: **wolf**  
True: **wolf**



Predicted: **wolf**  
True: **husky**

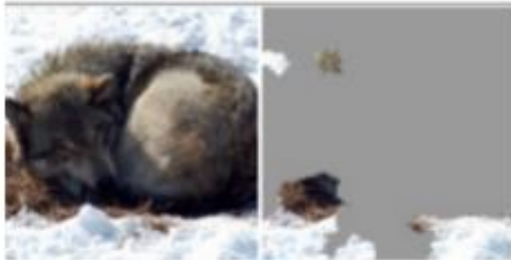


Predicted: **husky**  
True: **husky**

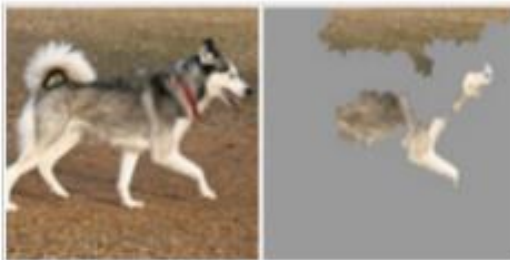


Predicted: **wolf**  
True: **wolf**

# The Explanation Reveals Why



Predicted: **wolf**  
True: **wolf**



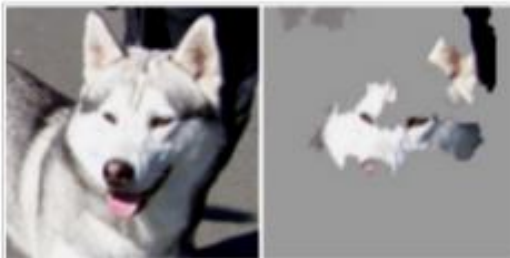
Predicted: **husky**  
True: **husky**



Predicted: **wolf**  
True: **wolf**



Predicted: **wolf**  
True: **husky**

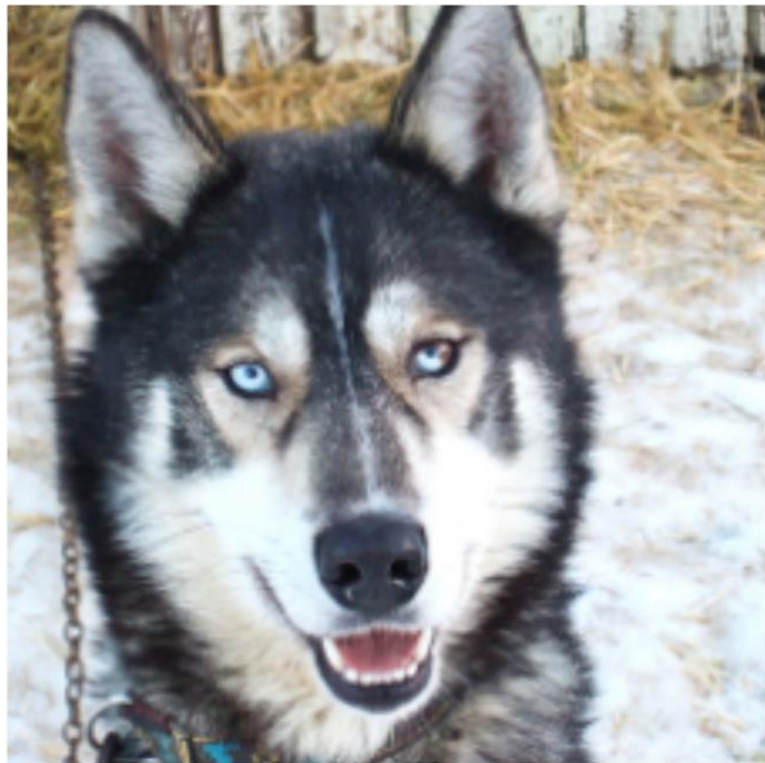


Predicted: **husky**  
True: **husky**

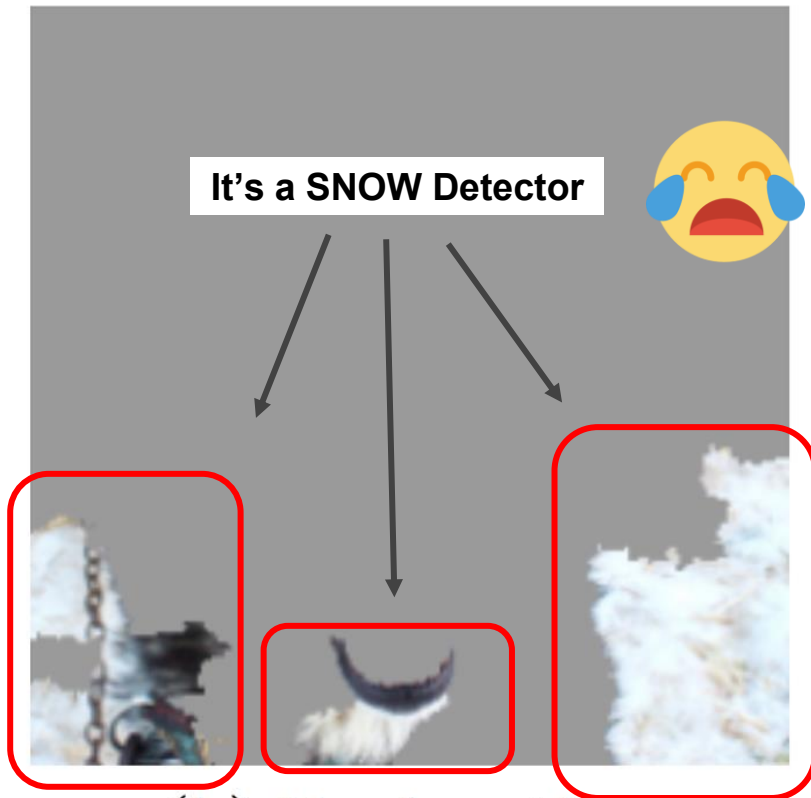


Predicted: **wolf**  
True: **wolf**

# The Explanation Reveals Why



(a) Husky classified as wolf



(b) Explanation

# Adversarial Inference



Speed Limit



Rifle

# Who cares...

GREGORY BARBER

TOM SIMONITE

BUSINESS

05.17.2019 07:00 AM

TOM SIMONITE

BUSINESS

## The Best

US government tes

## Facial is out

Database of  
algorithm  
misidentif

ALLIE FUNK

SECURITY

07.02.2019 09:00 AM

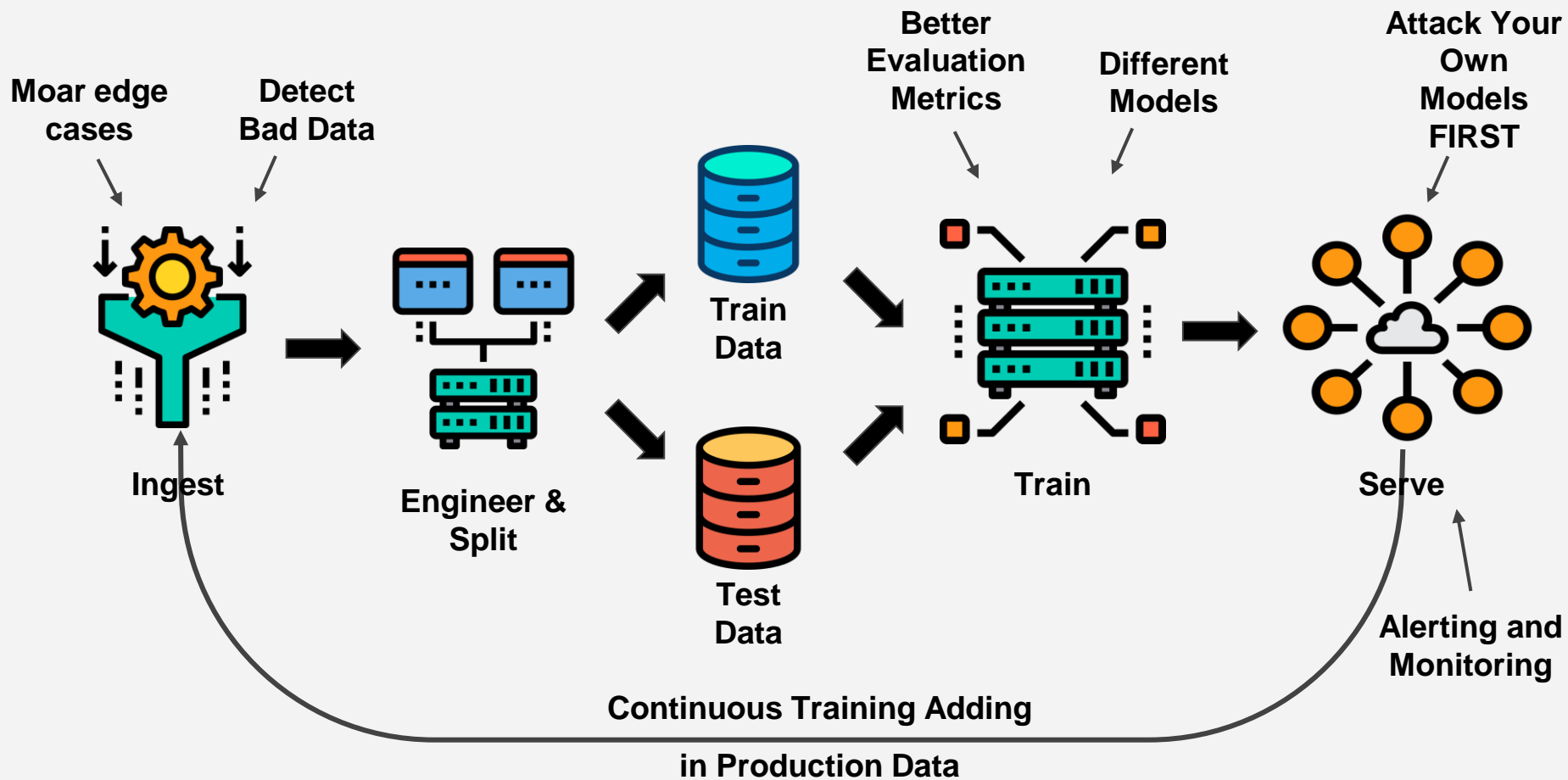
# I Opted Out of Facial Recognition at the Airport — It Wasn't Easy

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



**Terrified yet?**

# Tools to Defend



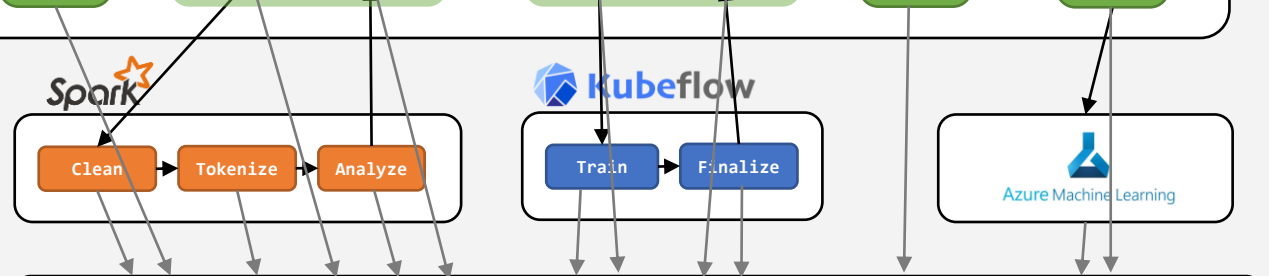
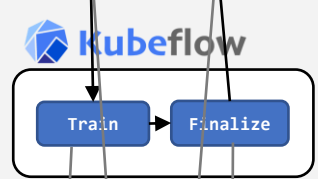
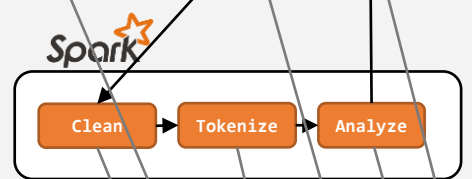
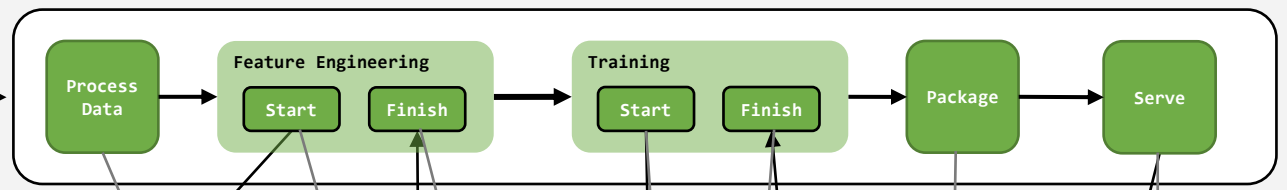
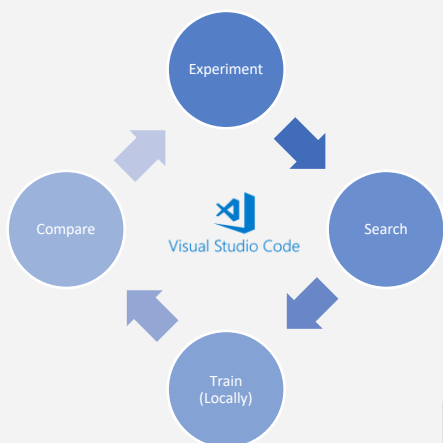
**But How?**

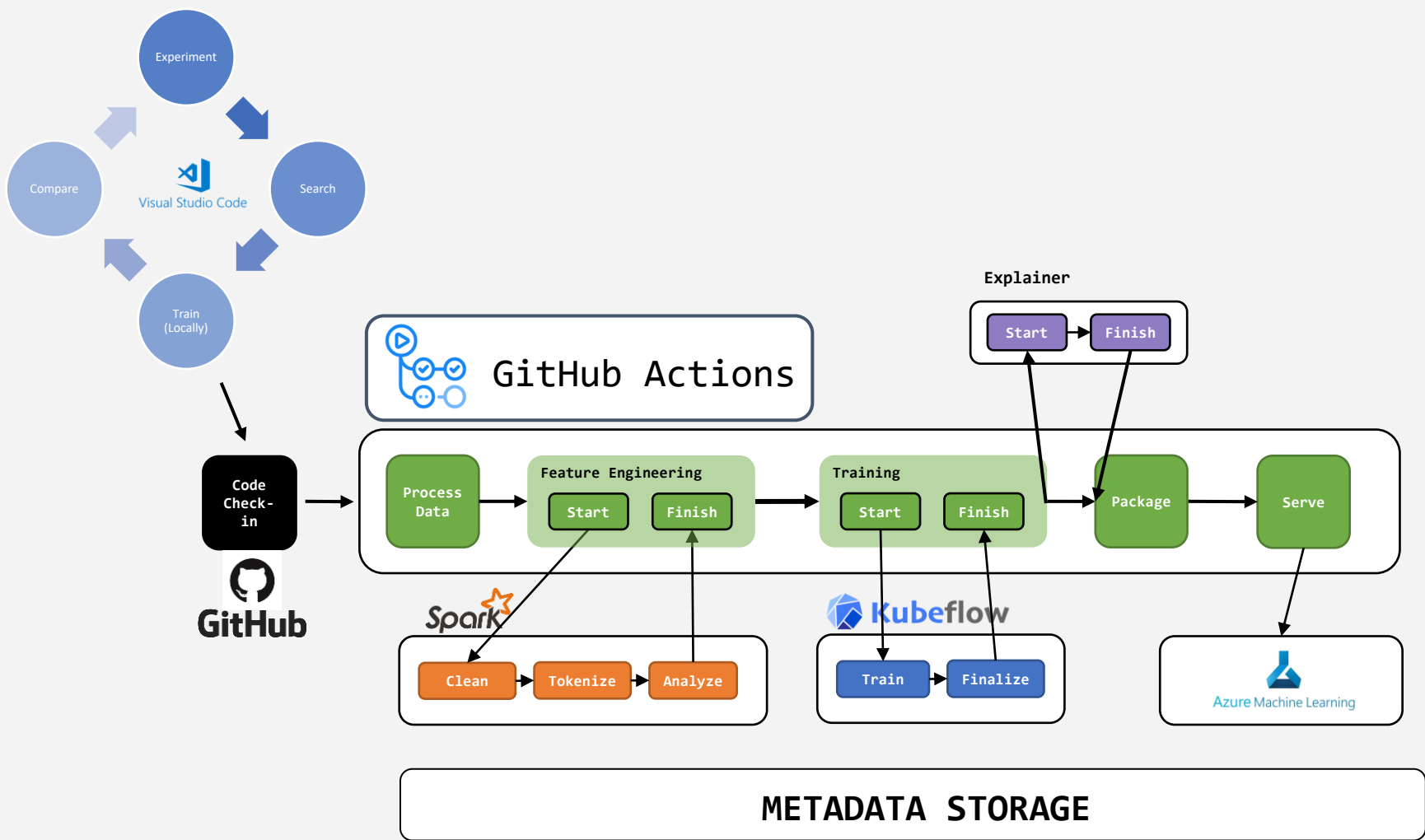


# **MLOps Pipelines!**

# 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





# 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**

# 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**

# **Attacker Takes Your Models**

# 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)

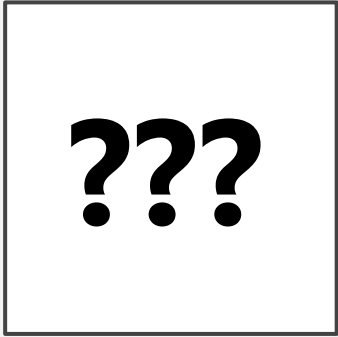
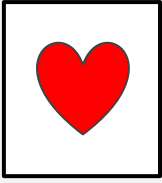
- **Distillation**

- “High Accuracy and High Fidelity Extraction of Neural Networks” - Jagielski, Carlini, Berthelot, Kurakin, Papernot
- Uses sampled data from same original distribution (usually)

- **Model Extraction**

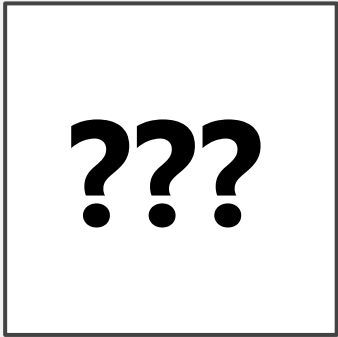
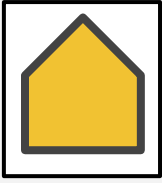
- ‘Thieves on Sesame Street! Model Extraction of BERT-based APIs’ - Krishna, Tomar, Parikh, Papernot and Iyyer
- Targets BERT style transformer models

# Distillation Attack

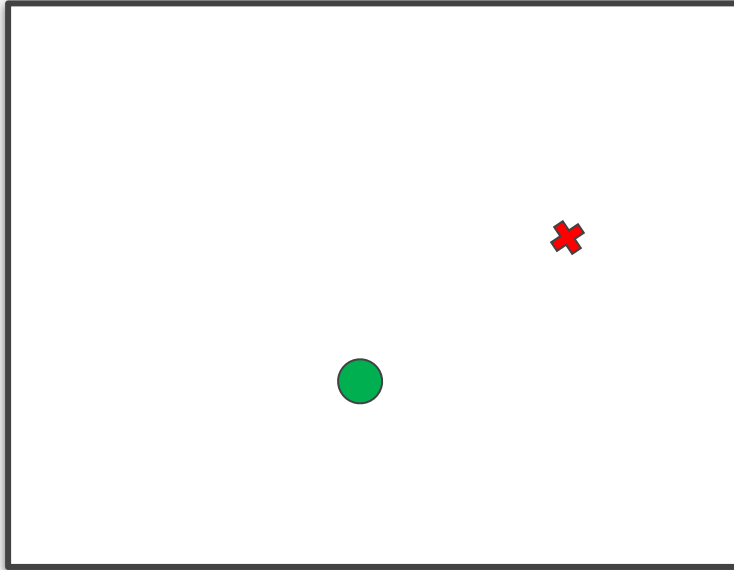
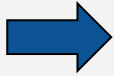


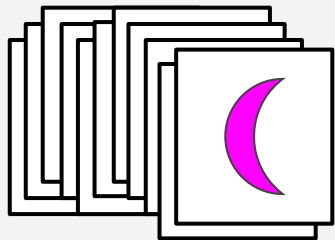
**Model**



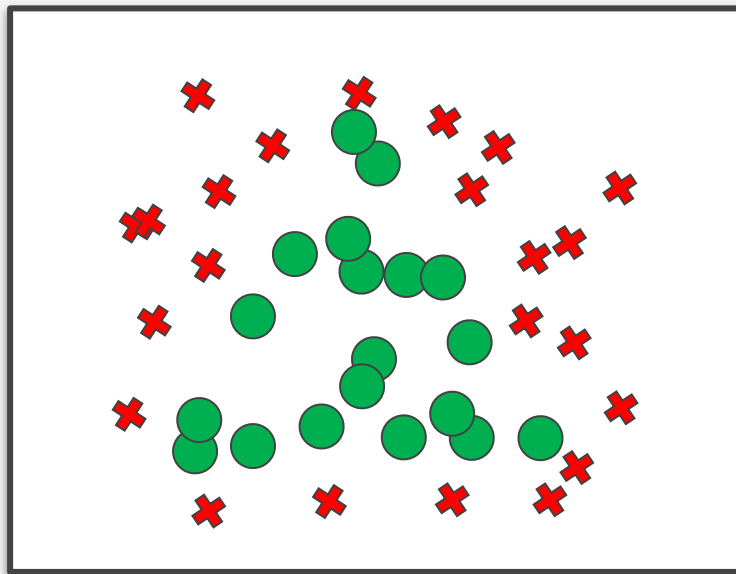
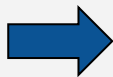


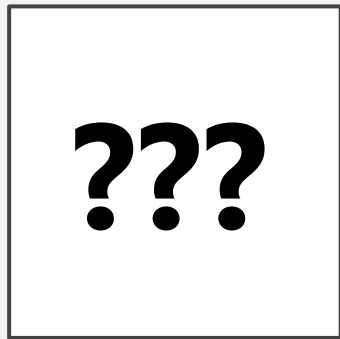
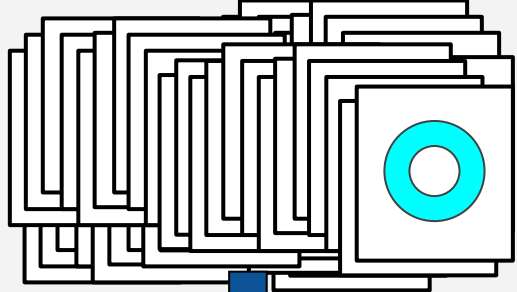
**Model**



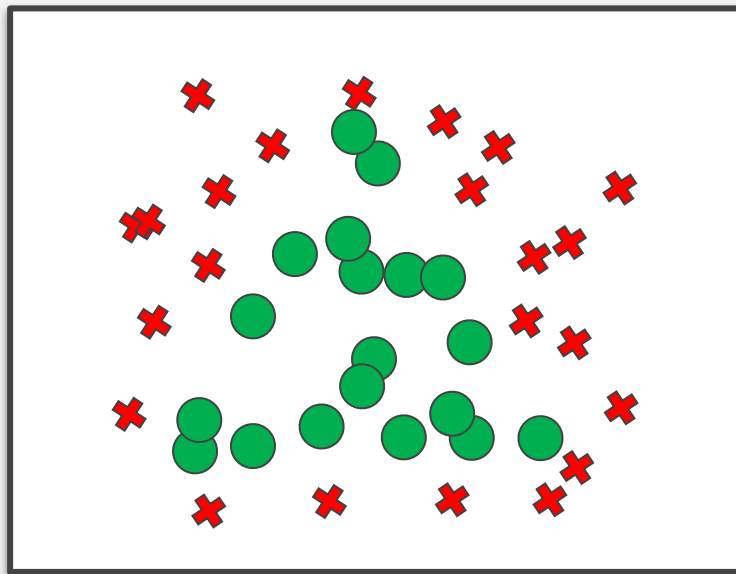


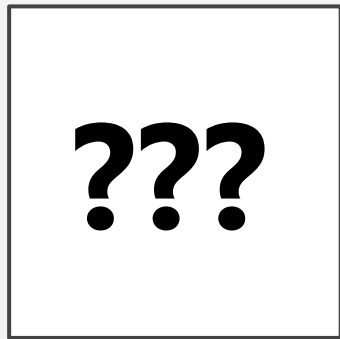
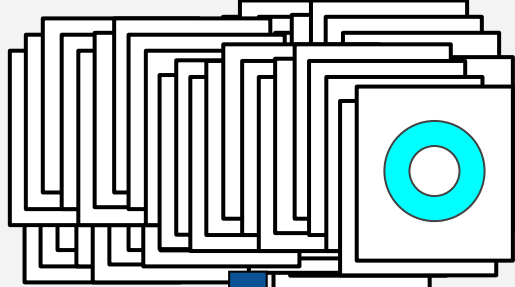
**Model**



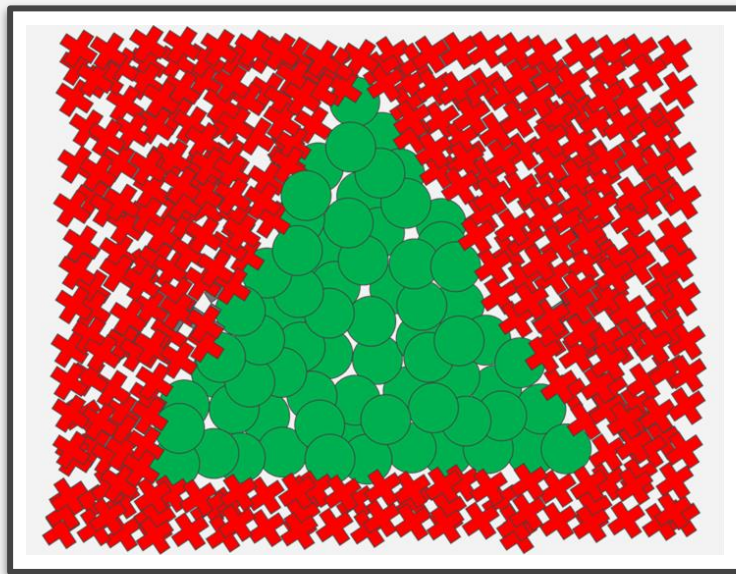
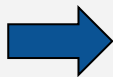


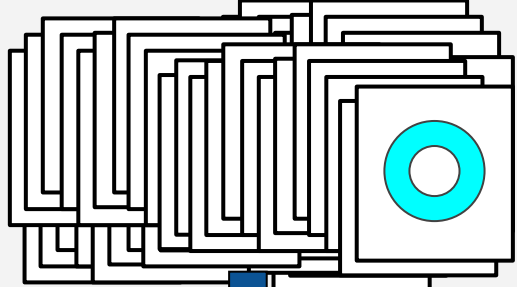
**Model**



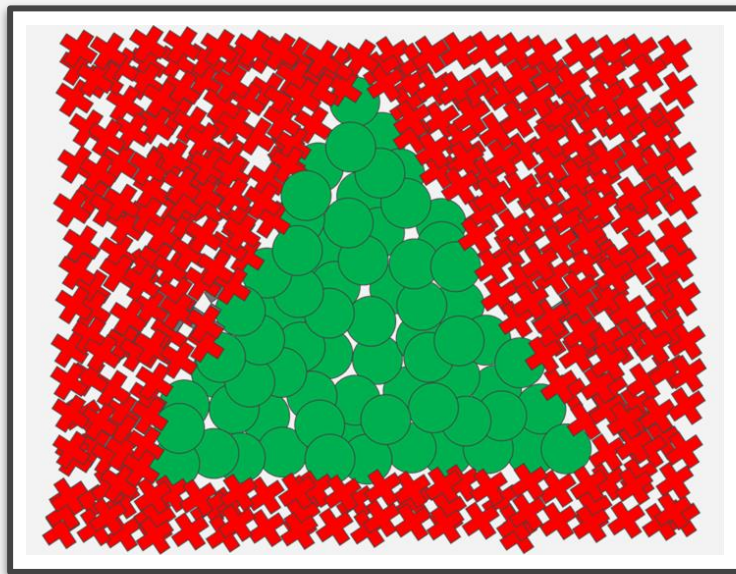
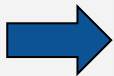


**Model**

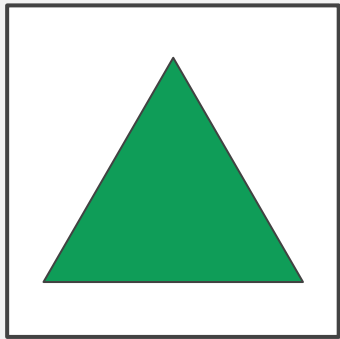
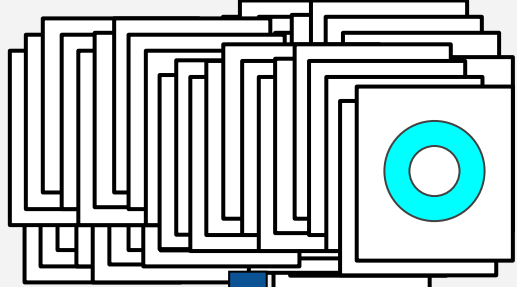




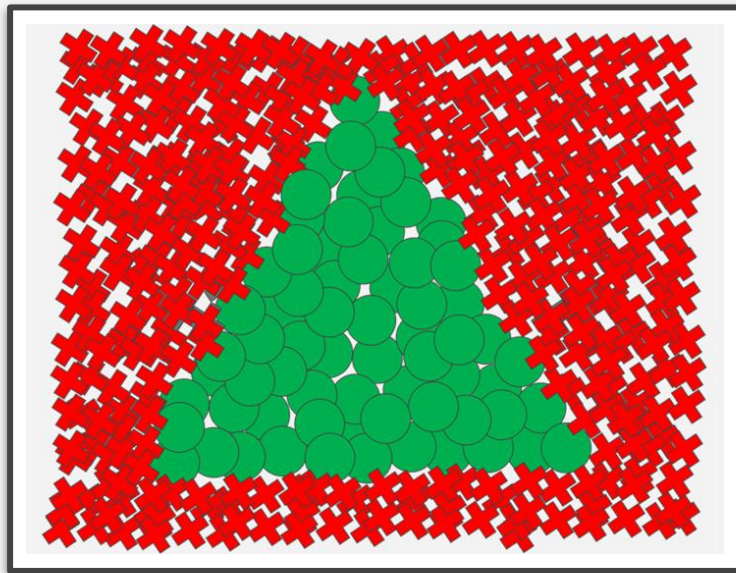
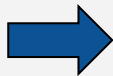
**Model**

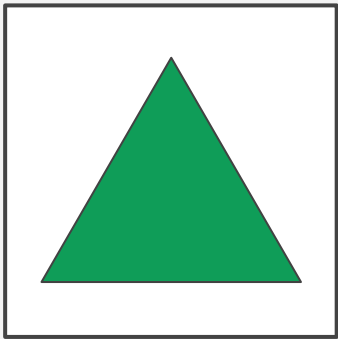




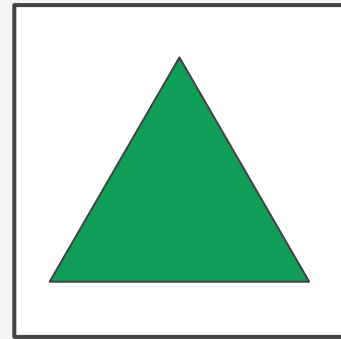
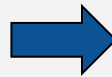
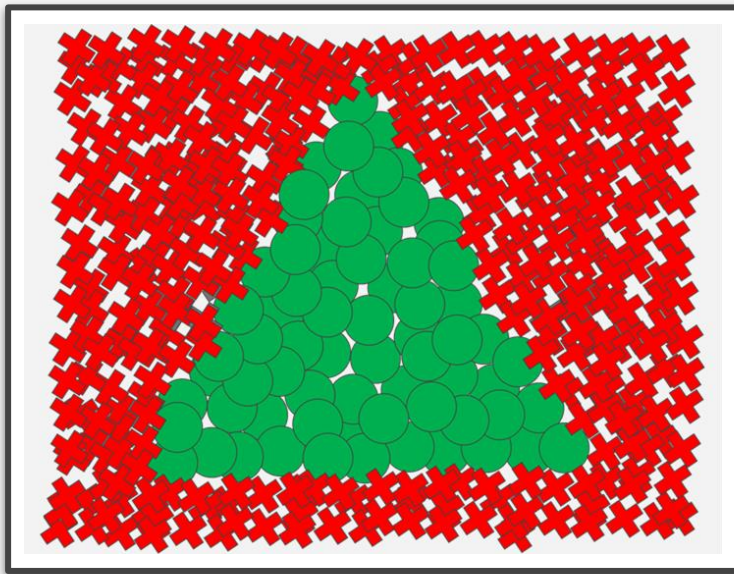
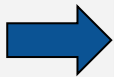


**Model**

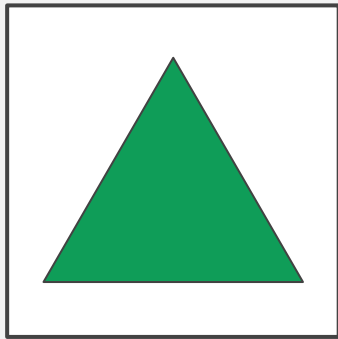




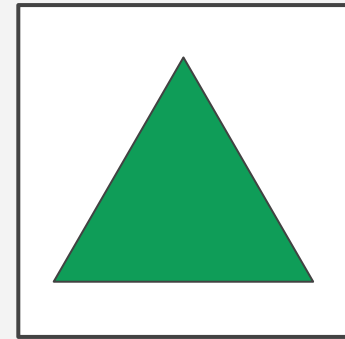
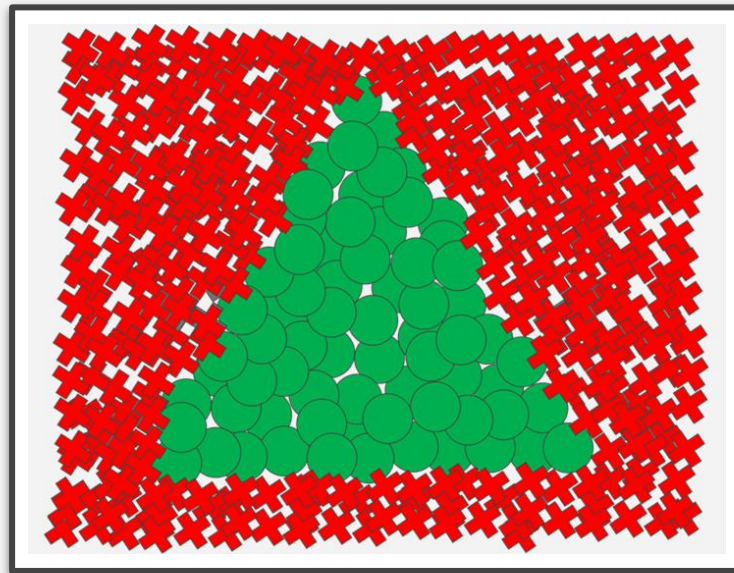
**Model**



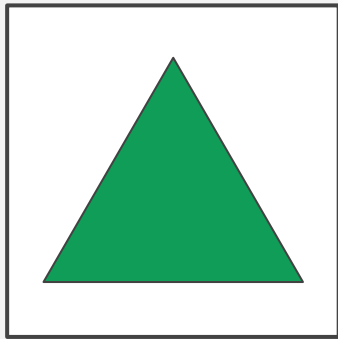
**Duplicated  
Model**



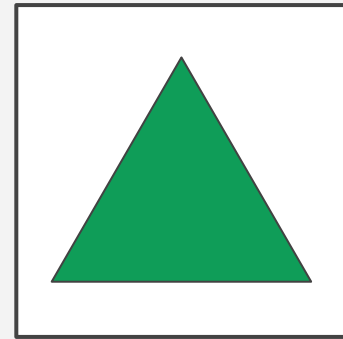
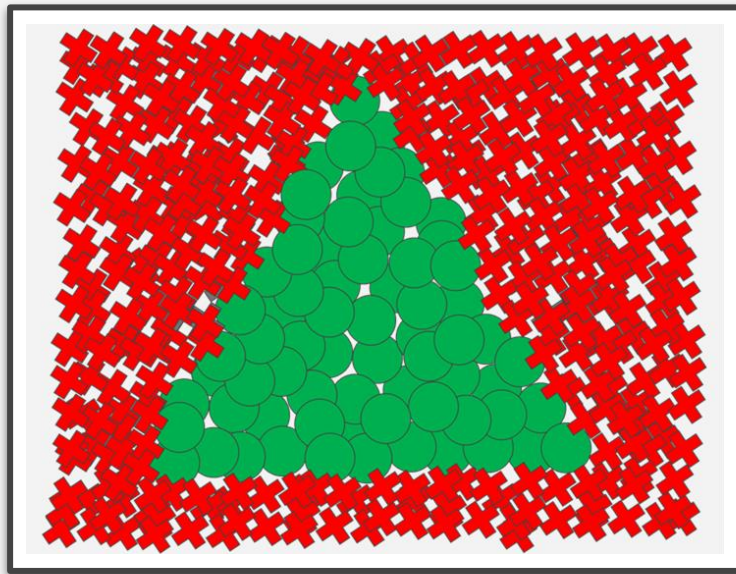
**Model**



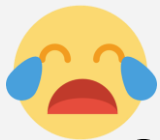
**Duplicated  
Model**



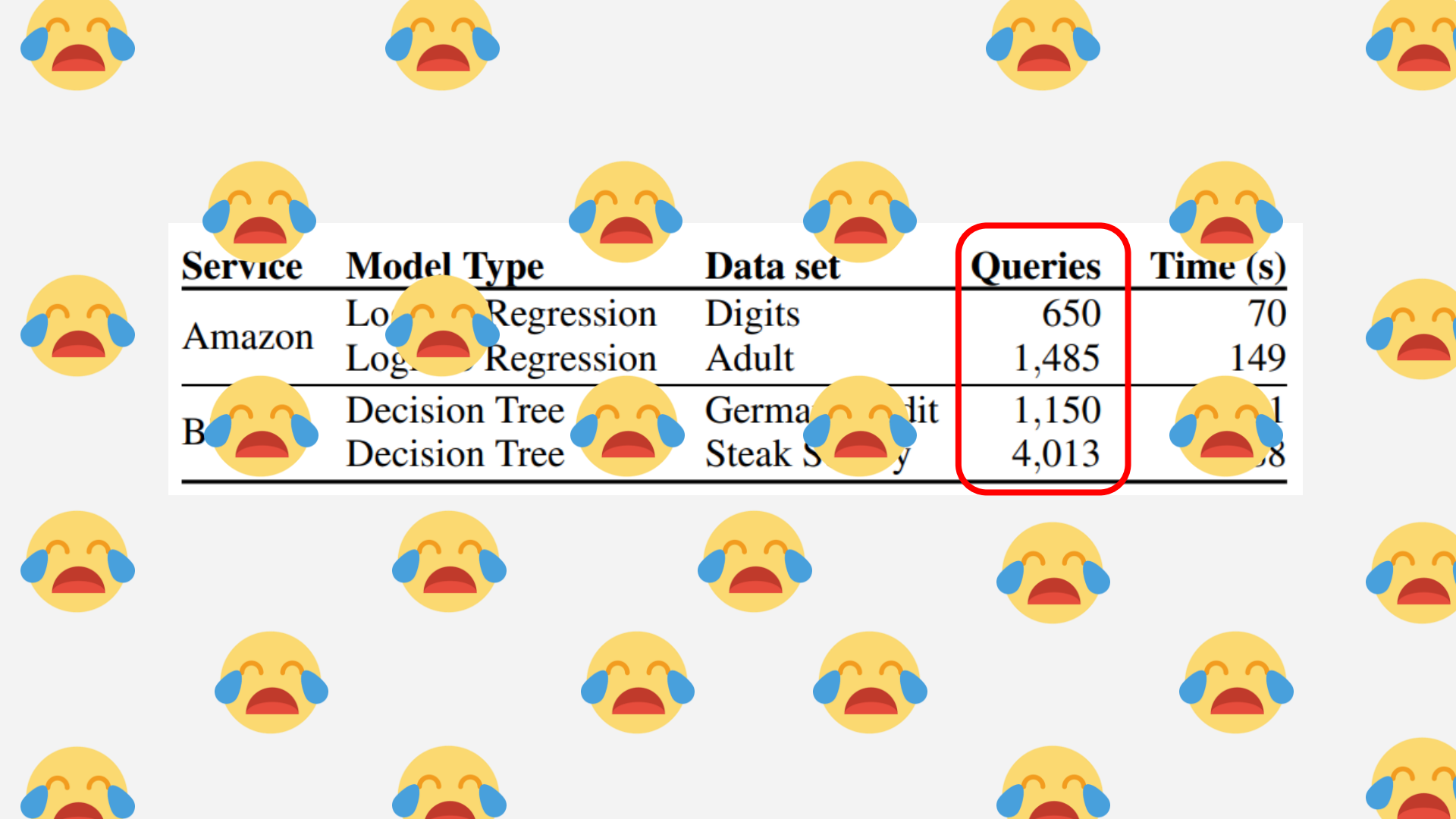
Model



Duplicated Model



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



Service	Model Type	Data set	Queries	Time (s)
Amazon	Logistic Regression	Digits	650	70
	Logistic Regression	Adult	1,485	149
B...	Decision Tree	German Credit	1,150	1...
	Decision Tree	Steak S...	4,013	...

# Model Extraction Attack

# 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

# 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.



# SQuAD Tests

- SQuAD = **S**tanford **Q**uestion **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.

# 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<sup>†,\*</sup>, Nicholas Carlini<sup>\*</sup>, David Berthelot<sup>\*</sup>, Alex Kurakin<sup>\*</sup>, and Nicolas Papernot<sup>\*</sup>

<sup>†</sup>Northeastern University

<sup>\*</sup>Google Research

# 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.

WIKI

**Q:** keyboards a as Nik baritone Cohn Prince on shifts multi-instrumentalist. virtuoso, is 5

**A:** On his first 5 albums.

# The Attack

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

# Using MLOps to Defend

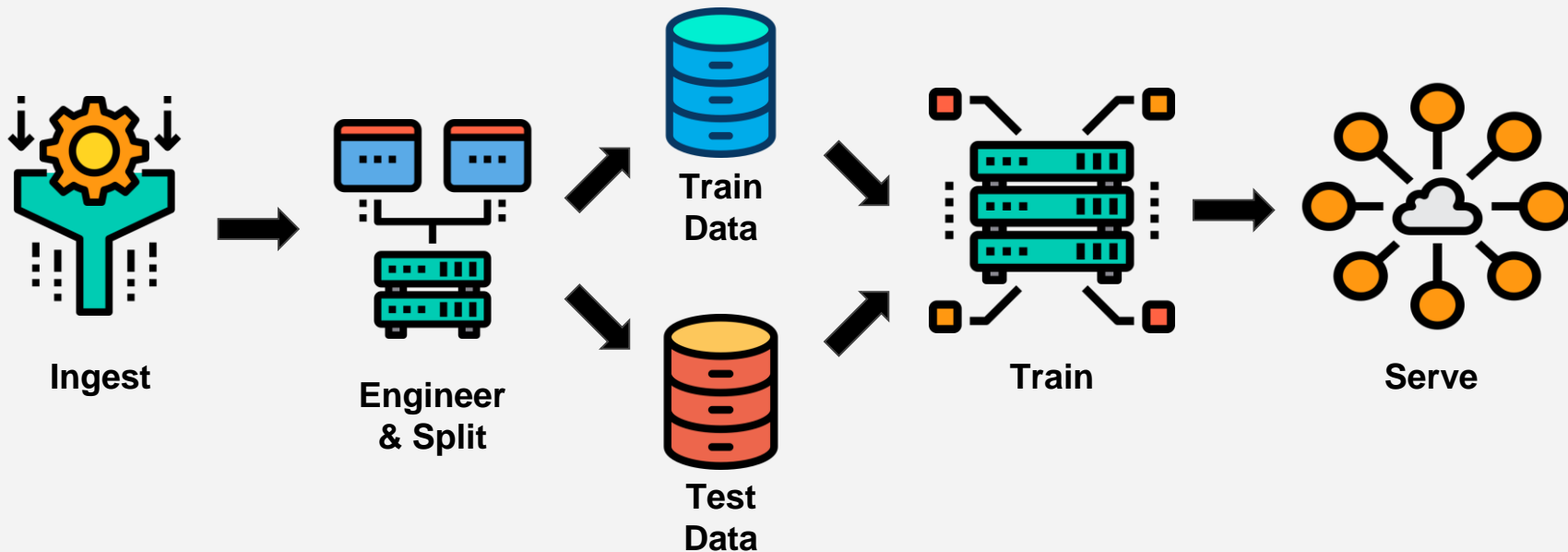
- **Possible defenses**
  - Detecting queries that could be part of an attack
  - Watermarking predictions made by the API
- **REPEAT: The *pipeline* is the value not the *model***
  - Improving domain specificity
  - Continuous retraining for accuracy
  - Faster throughput & SLA
- **If you REALLY need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

# Tools to Defend

Spend the Majority of Engineering Here

Not Here



# 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**

# 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**



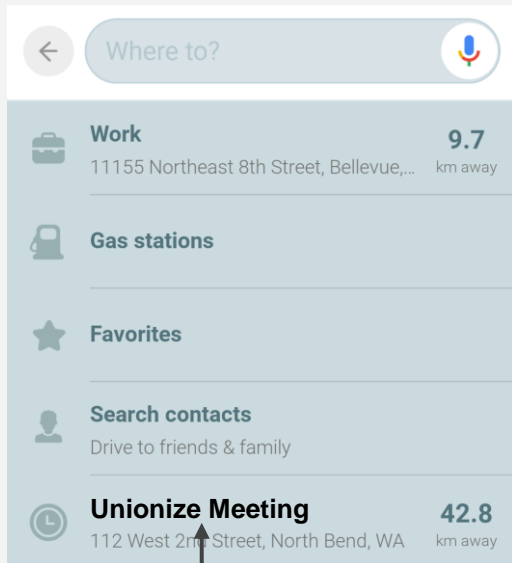
# **Attacker Finds Out About Hidden Data**

# 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**

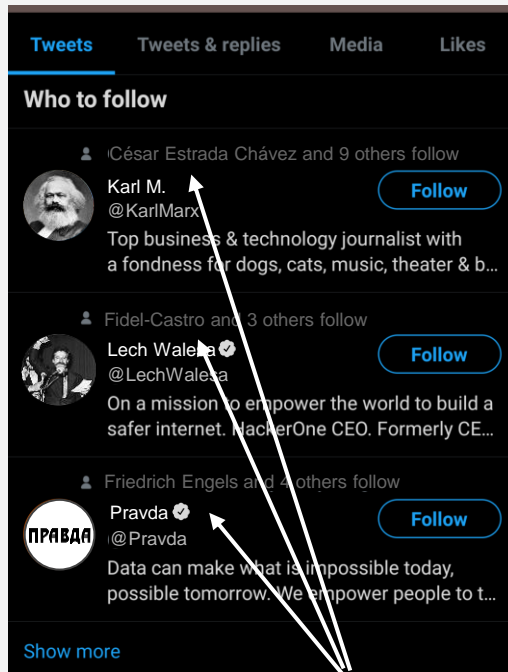
# Hidden Data Leakage Examples

## Recommendations



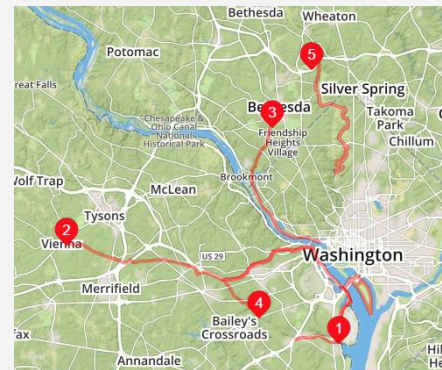
My Historical Events

## Network Graph



My Friend's Graphs

## Maps

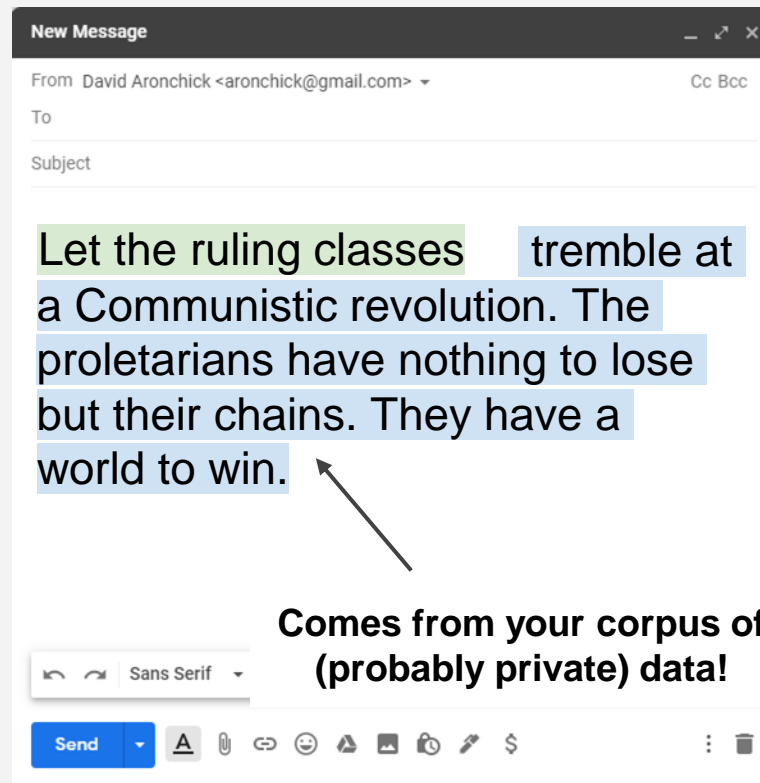


The Community's Locations

**There's Nothing So  
Bad that It  
Can't Be Worse**

**There's Nothing So  
Bad that It  
Can't Be Worse  
(Especially with ML)**

# Secret Memorization



# 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

# Secret Memorization

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

Nicholas Carlini<sup>1,2</sup> Chang Liu<sup>2</sup> Úlfar Erlingsson<sup>1</sup> Jernej Kos<sup>3</sup> Dawn Song<sup>2</sup>

<sup>1</sup>Google Brain <sup>2</sup>University of California, Berkeley <sup>3</sup>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.

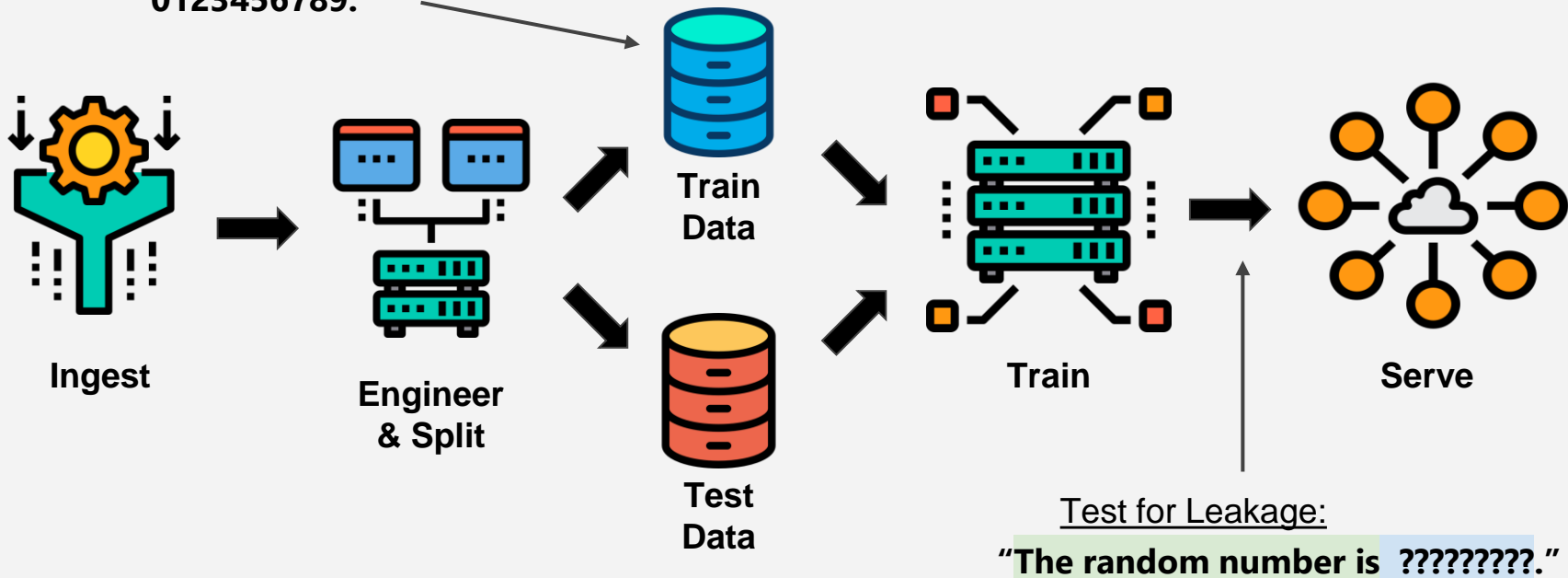
Our exposure metric conservatively characterizes knowledgeable attackers that target secrets unlikely to be discovered by accident (or by a most-likely beam search). As validation of this, we describe an algorithm guided by the exposure met-



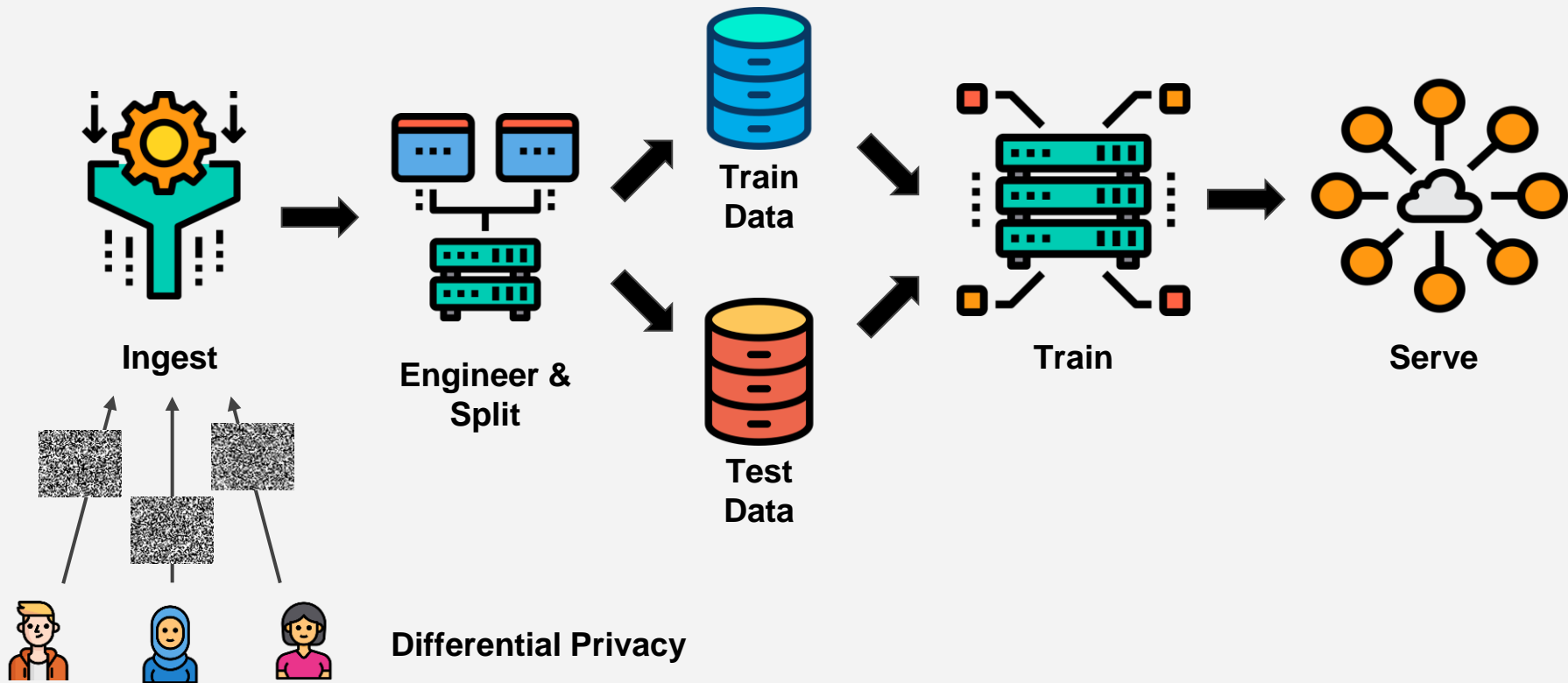
# Secret Memorization

Inject Canary:

**"The random number is  
0123456789."**



# Secret Memorization



# 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

# Summary

# 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**

# 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 everyone take advantage of this transformation?



# Truths You Cannot Avoid

- You WILL be attacked
- Your pipeline WILL have issues
- The game is all about mitigation of harms (and quick recovery)

**me:** David Aronchick

**twitter:** [@aronchick](https://twitter.com/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 Iyer
- 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!**