Guardians of Intelligence

Navigating Supply Chain Security for ML Applications

Dan Fernandez

Illustration by: MidJourney

Agenda

- Intro to ML Applications
- Understanding the Software Supply Chain
- Attack Vectors
- Safeguarding the AI Realm

Illustration by: MidJourney

Dan Fernandez

- Data Nerd
- Product Manager in Cybersecurity
- Instructional Associate @ GaTech

danielfernandez@infosec.exchange

@danielfernandez

/in/dafnz

danielfernandez.medium.com

hey@dafnz.com

Disclaimer

Opinions are my own and not my employers.

Illustration by: MidJourney

Presentation Resources

- The slides will be made available in Github immediately after the presentation.
- Slides contain references to content with additional information.
- 💡Resource: Will look like this

Illustration by: MidJourney

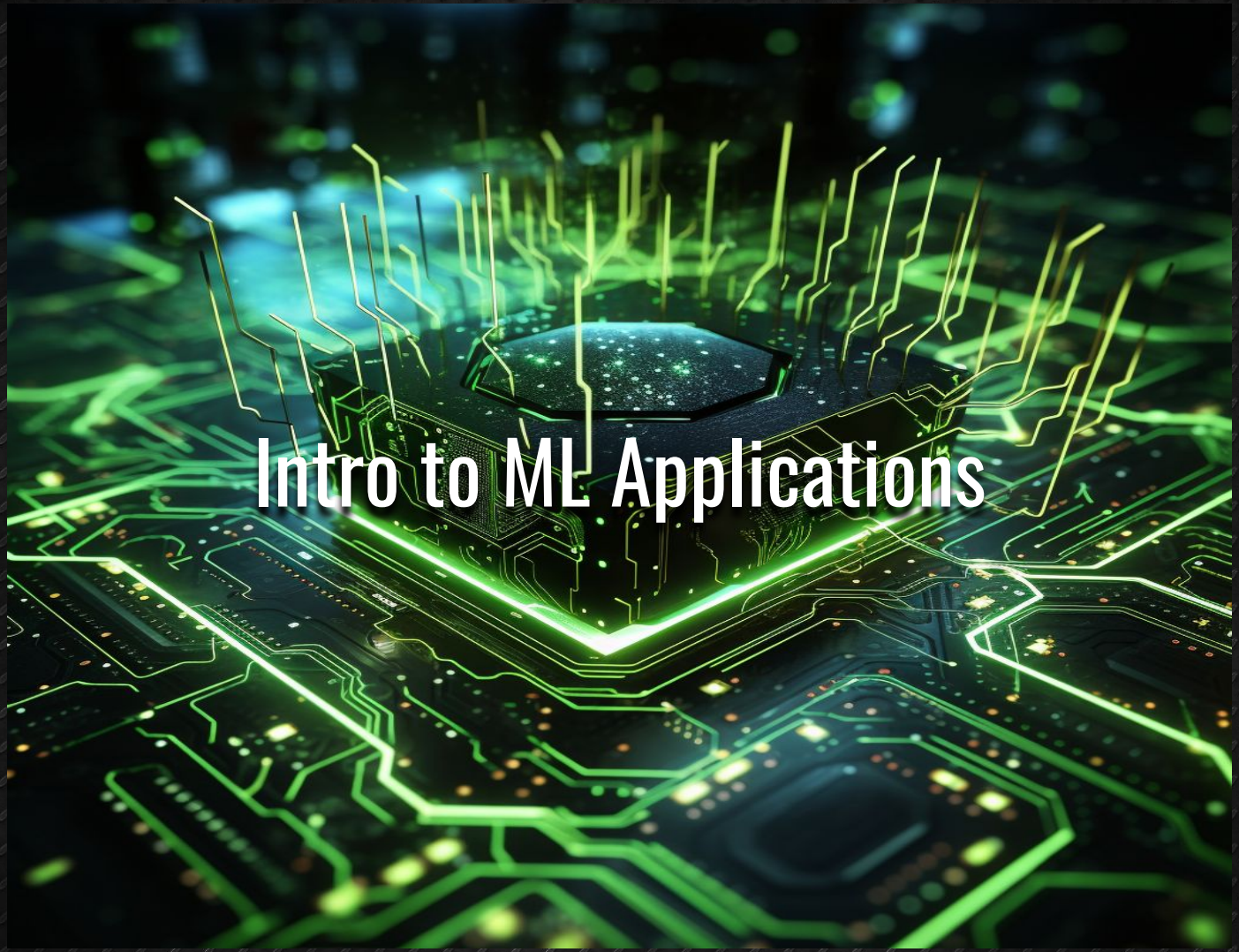


Illustration by: MidJourney

What is ML?

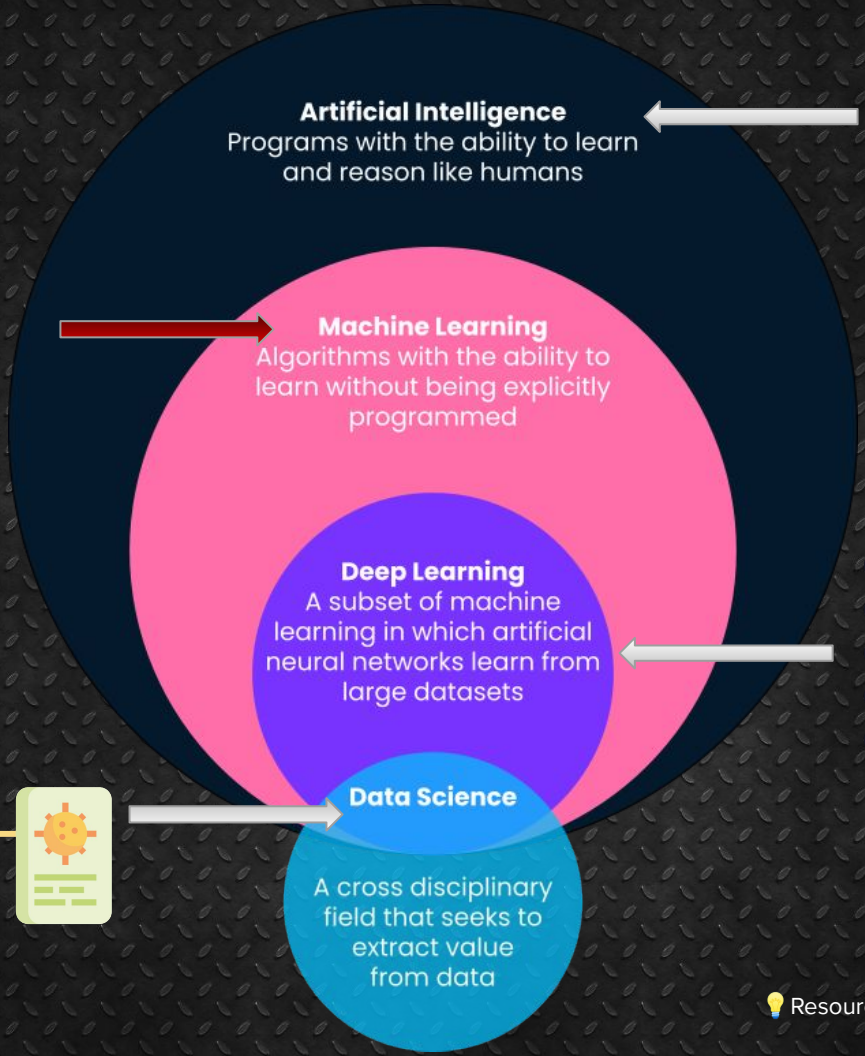


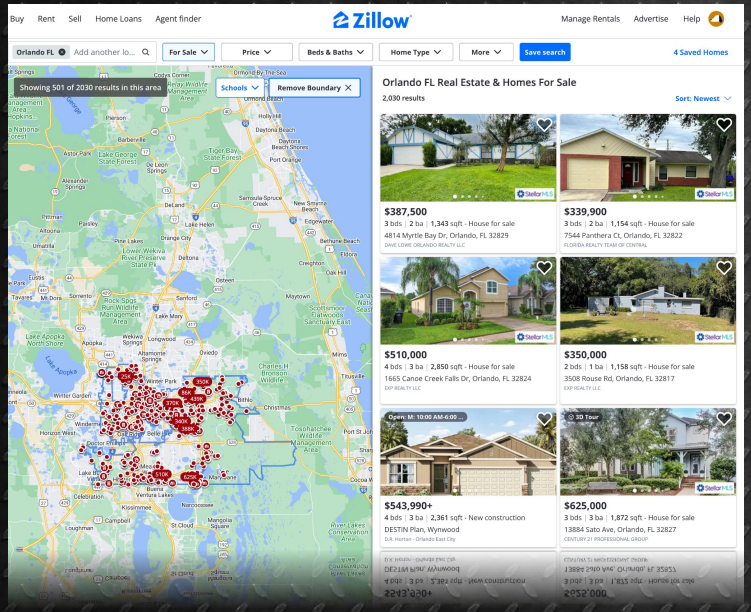
Illustration by: DataCamp

💡 Resource: AI vs ML vs Deep Learning

What are the types of tasks do ML Models Perform?

Supervised Machine Learning

- Regression



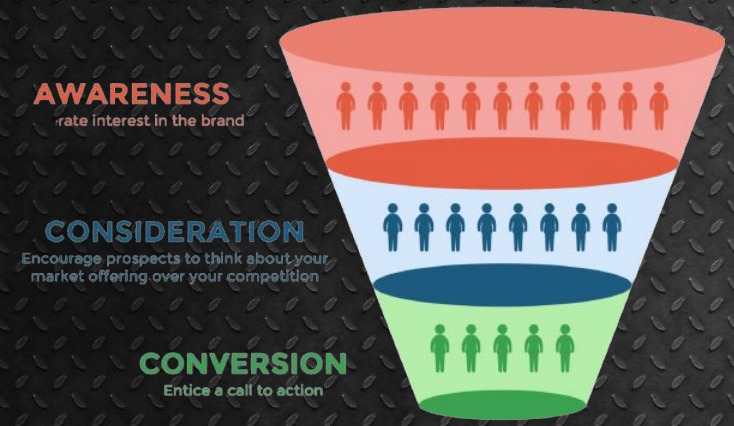
- Classification

💡 Resource: Supervised ML Tutorial

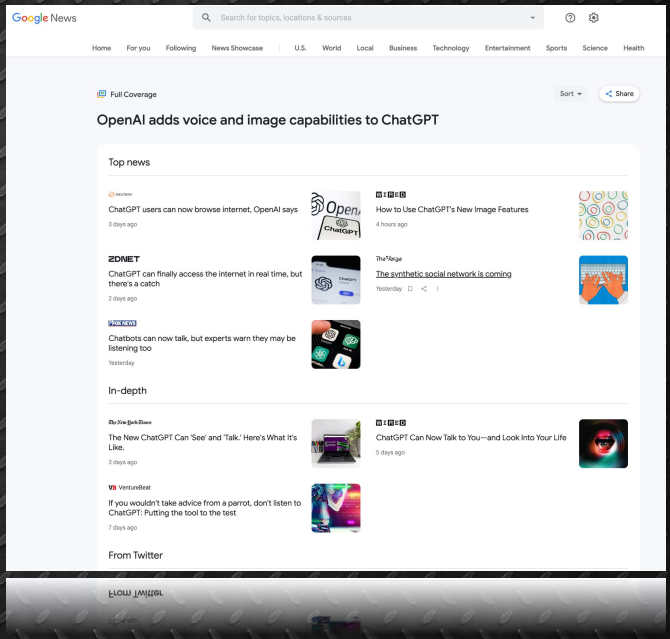
What are the types of tasks do ML Models Perform?

Unsupervised Machine Learning

- Customer Segmentation



- Recommender Systems



💡Resource: Unsupervised ML Tutorial

Why should we secure ML Applications?

Finance

Public Safety

Cybersecurity

Finance: Flash Crash +

REUTERS

World Business Markets Breakingviews Video More

COMMODITIES APRIL 23, 2013 / 5:58 PM / UPDATED 10 YEARS AGO

Analysis: False White House tweet exposes instant trading dangers

By Steven C. Johnson 6 MIN READ

NEW YORK (Reuters) - The upheaval in financial markets caused by a false report of explosions at the White House was brief, but its effect on traders who have come to rely on Twitter may last quite a bit longer.

The U.S. Capitol Building is pictured in Washington, February 27, 2013. REUTERS/Jason Reed

Volatile price moves ricocheted through markets for stocks, bonds, currencies and commodities around 1 p.m. EDT on Tuesday after a tweet purporting to be from the Associated Press said there had been two explosions at the White House and that President Barack Obama had been injured.

Entrepreneur

Sign In Subscribe

Business News

Eli Lilly Stock Plummets After Parody Twitter Account Says Insulin is Now Free

The pharma company felt the affects of a dangerous imitation account on Thursday.

BY EMILY RELLA • NOV 11, 2022 Share

Getty Images

Elon Musk's disastrous back-and-forth rollout of Twitter Blue has caused a multitude of communication problems between users and trolls, with fake news and information inadvertently going viral after users quickly see blue checkmarks and assume what they are reading is from a legitimate source.

Though much of the banter has been in good jest, one company is feeling the very real implications of the viral spreading of misinformation.

Eli Lilly and Company's stock plummeted on Thursday after an account posing to be the pharmaceutical company claimed that insulin would now be free in the United States.

💡Resource: Flash Crash Simulator Whitepaper

Disabling Physical Security

Vox EXPLAINERS CROSSWORD VIDEO PODCASTS POLITICS POLICY CULTURE SCIENCE MORE Give

We have a request

We need 292 more contributions today to hit our goal of adding 2,500 reader gifts in September. Contributions help keep Vox's policy coverage and beyond free. Support quality journalism that's not behind a paywall, Make a contribution today.

Yes, I'll Give

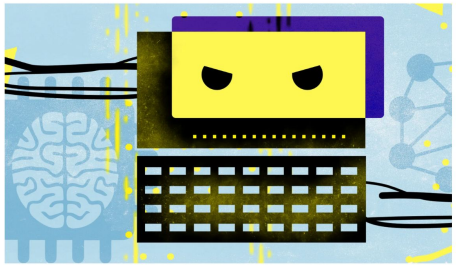
FUTURE PERFECT TECHNOLOGY

It's disturbingly easy to trick AI into doing something deadly

How "adversarial attacks" can mess with self-driving cars, medicine, and the military.

By Sigal Samuel | Apr 8, 2019, 9:10am EDT

SHARE



Javier Zarracina/Vox

Most Read

1. What Dianne Feinstein's death means for California's Senate elections
2. Scientists will unleash an army of crabs to help save Florida's dying reef
3. 1 winner and 3 losers from Fox's dud of a second GOP debate
4. More young, healthy people should be getting Paxlovid when they get Covid
5. Twitter's CEO had a wild, combative appearance at the Code conference

Sigal Samuel is a senior reporter for Vox's Future Perfect and co-host of the Future Perfect podcast. She writes primarily about the future of consciousness, tracking advances in artificial intelligence and neuroscience and their staggering ethical implications. Before joining Vox, Sigal was the religion editor at the Atlantic.

Future Perfect

Each week, we explore unique

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

› Defense Advanced Research Projects Agency › Our Research › Guaranteeing AI Robustness Against Deception

Guaranteeing AI Robustness Against Deception (GARD)

Dr. Alvaro Velasquez

The growing sophistication and ubiquity of machine learning (ML) components in advanced systems dramatically expands capabilities, but also increases the potential for new vulnerabilities. Current research on adversarial AI focuses on approaches where imperceptible perturbations to ML inputs could deceive an ML classifier, altering its response. Such results have initiated a rapidly proliferating field of research characterized by ever more complex attacks that require progressively less knowledge about the ML system being attacked, while proving increasingly strong against defensive countermeasures. Although the field of adversarial AI is relatively young, dozens of attacks and defenses have already been proposed, and at present a comprehensive theoretical understanding of ML vulnerabilities is lacking.

GARD seeks to establish theoretical ML system foundations to identify system vulnerabilities, characterize properties that will enhance system robustness, and encourage the creation of effective defenses. Currently, ML defenses tend to be highly specific and are effective only against particular attacks. GARD seeks to develop defenses capable of defending against broad categories of attacks. Furthermore, current evaluation paradigms of AI robustness often focus on simplistic measures that may not be relevant to security. To verify relevance to security and wide applicability, defenses generated under GARD will be measured in a novel testbed employing scenario-based evaluations.

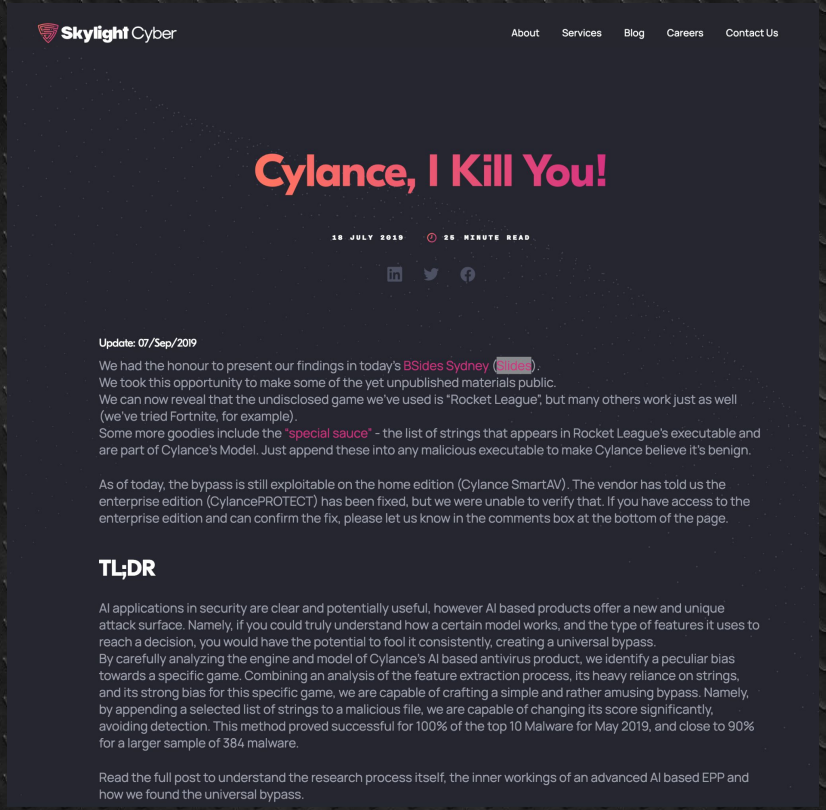
As part of the program, GARD researchers from Two Six Technologies, IBM, MITRE, University of Chicago, and Google Research generated the following virtual testbed, toolbox, benchmarking dataset, and training materials that are now available to broader research community:

- The Armory virtual platform, available on GitHub, serves as a "testbed" for researchers in need of repeatable, scalable, and robust evaluations of adversarial defenses.
- Adversarial Robustness Toolbox (ART) provides tools for developers and researchers to defend and evaluate their ML models and applications against a number of adversarial threats.
- The Adversarial Patches Rearranged In COnText (APRICOT) dataset enables reproducible research on the real-world effectiveness of physical adversarial patch attacks on object detection systems.
- The Google Research Self-Study repository contains "test dummies" that represent a common idea or approach to build defenses.

The GARD program's Holistic Evaluation of Adversarial Defenses repository is available at https://www.gardproject.org/. Interested researchers are encouraged to take advantage of these resources and check back often for updates.

💡Resource: Robust Physical-World Attacks on Deep Learning Visual Classification

Impairing Malware Detection



💡Resource: [Cylance Machine Learning Bypass Writeup]

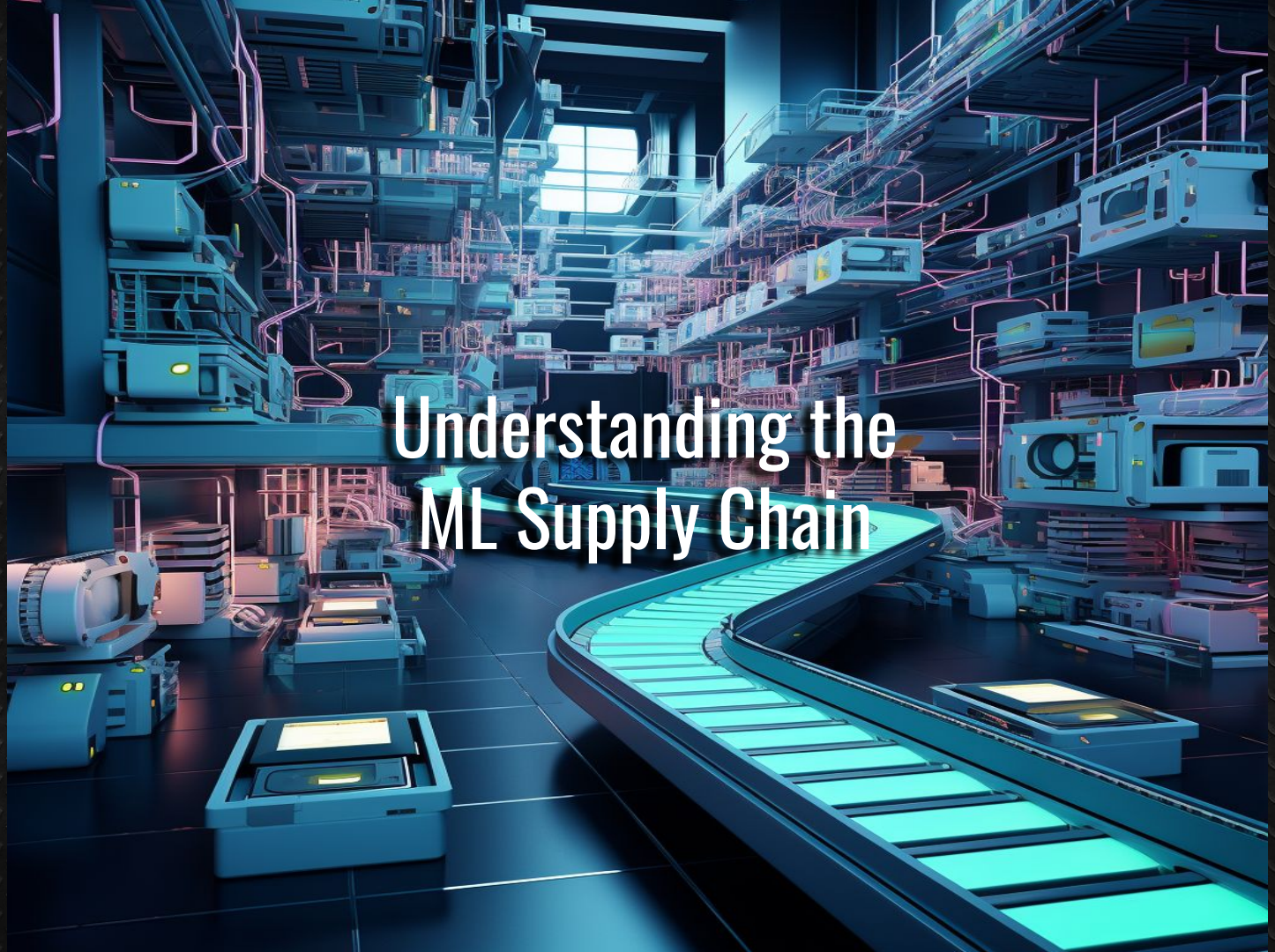
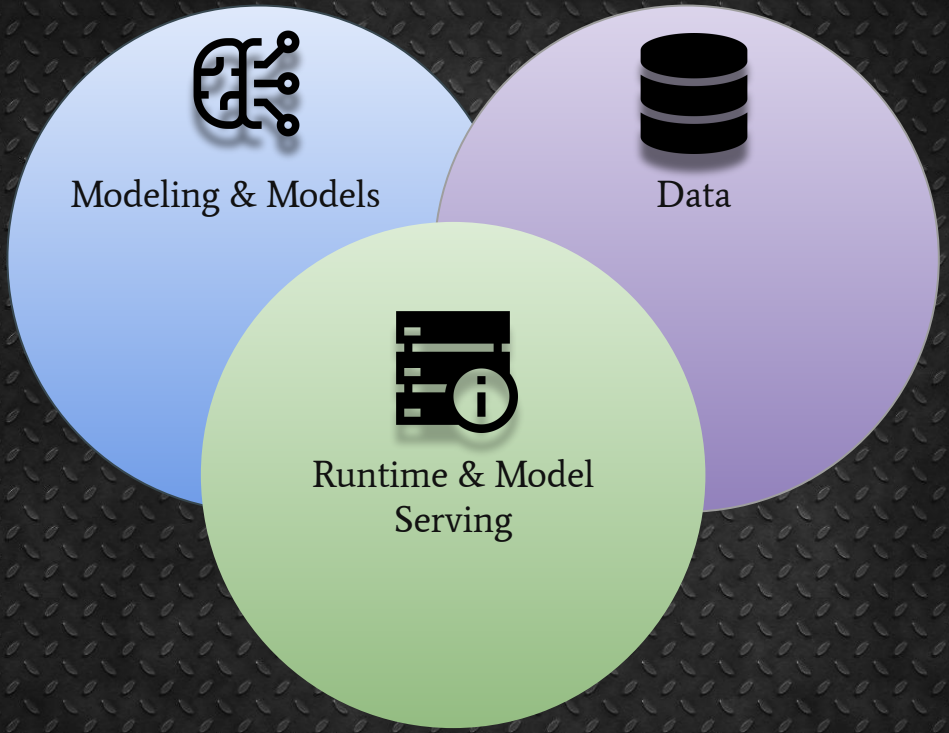


Illustration by: MidJourney

Machine Learning Supply Chain



The ML Stack

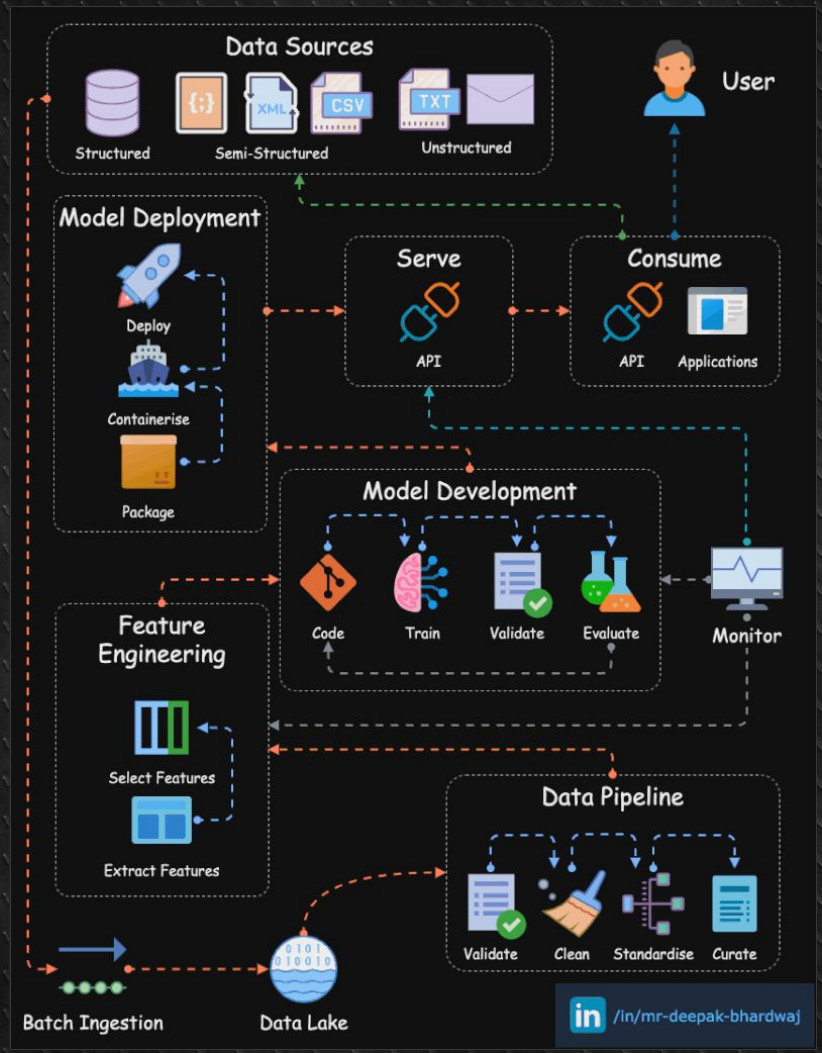
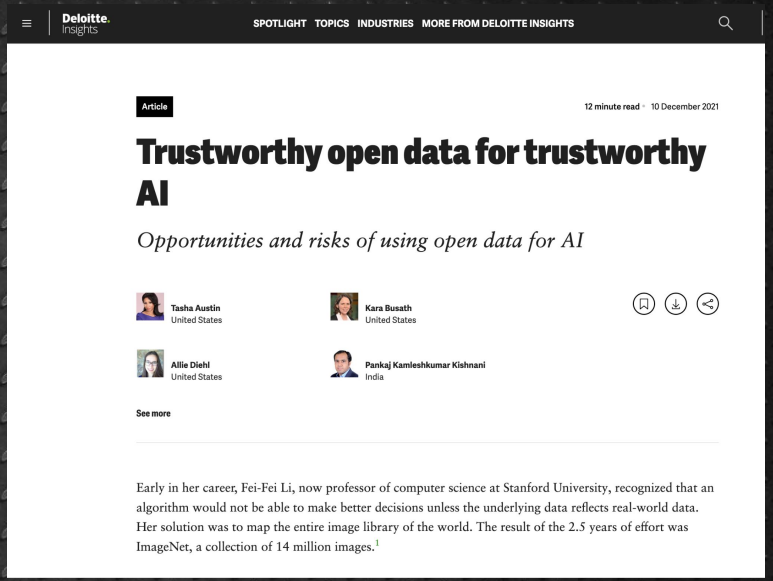


Image Source: Deepak Bhardwaj

Data

- Data Fuels Machine Learning
- Better Models Require Increasingly More + Better Data
- Open Source Data is great but comes with a risk



Deloitte. Insights

SPOTLIGHT TOPICS INDUSTRIES MORE FROM DELOITTE INSIGHTS

Article

12 minute read · 10 December 2021

Trustworthy open data for trustworthy AI

Opportunities and risks of using open data for AI

Tasha Austin
United States

Kara Busath
United States

Allie Diehl
United States

Pankaj Kamleshkumar Kishnani
India

See more

Early in her career, Fei-Fei Li, now professor of computer science at Stanford University, recognized that an algorithm would not be able to make better decisions unless the underlying data reflects real-world data. Her solution was to map the entire image library of the world. The result of the 2.5 years of effort was ImageNet, a collection of 14 million images.[1]

💡Resource: The Risks of Open Data - Deloitte

💡Resource: Data Poisoning Explained

Data

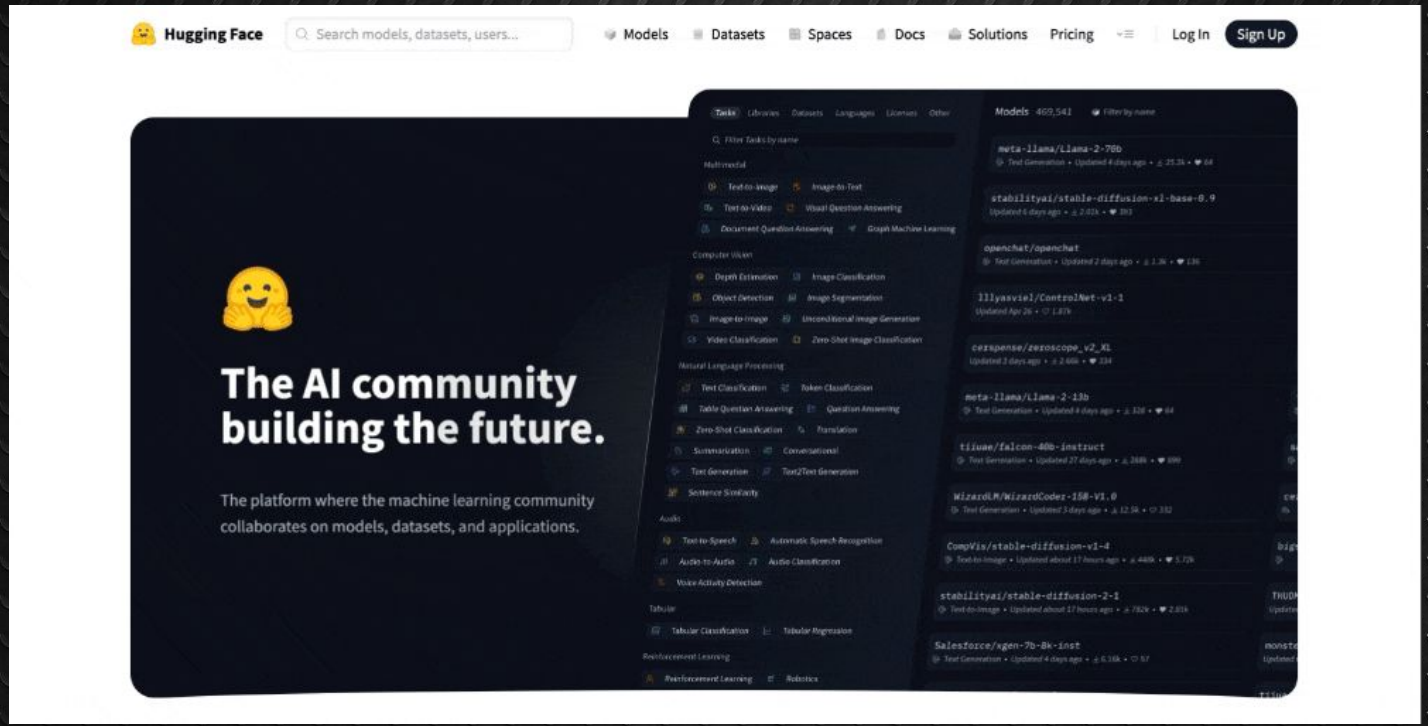


Image Source: HuggingFace

Modeling

- Programming Languages & Libraries

Modeling

- Research & Prototyping Environments

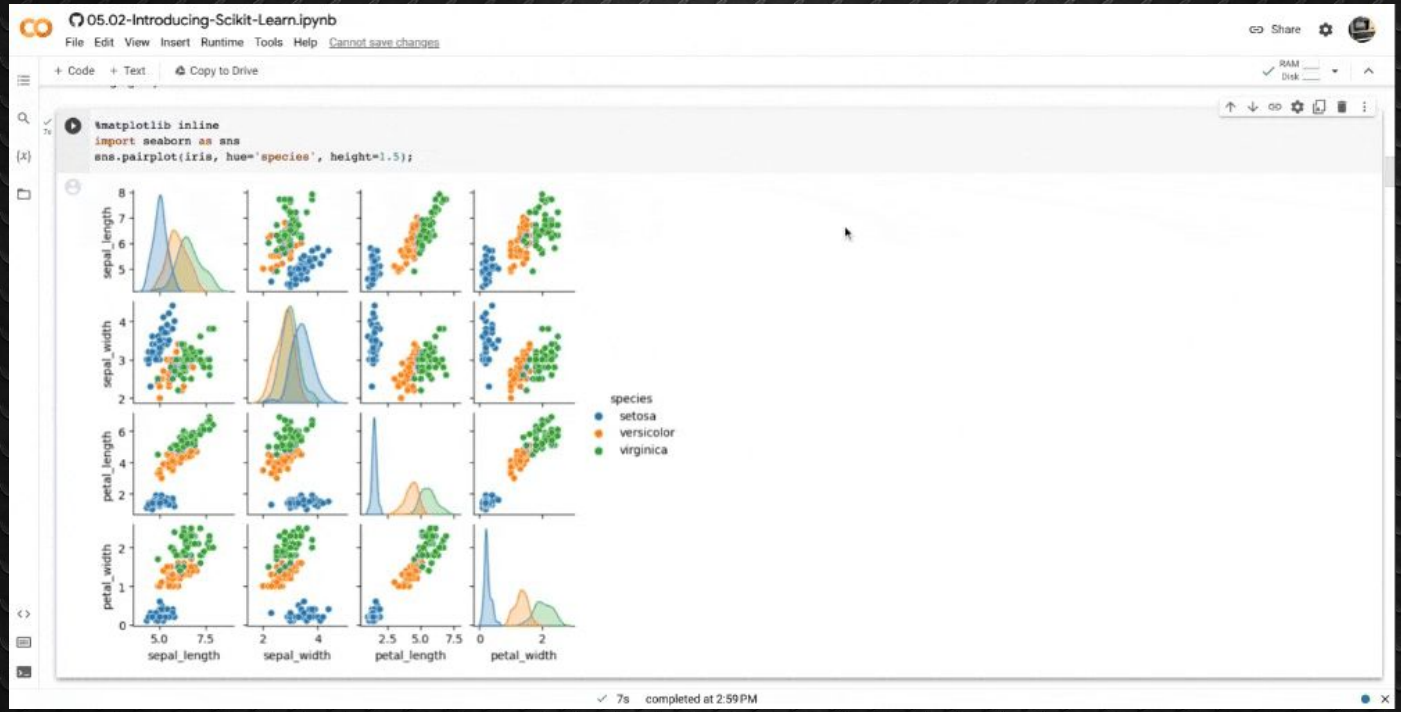
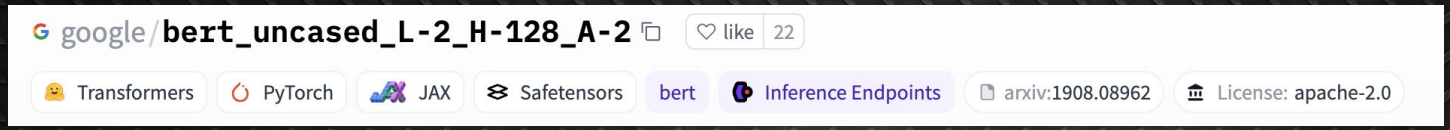
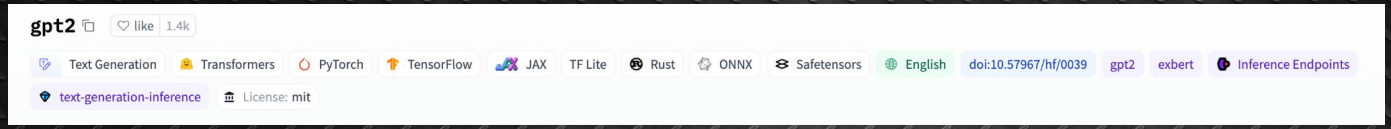
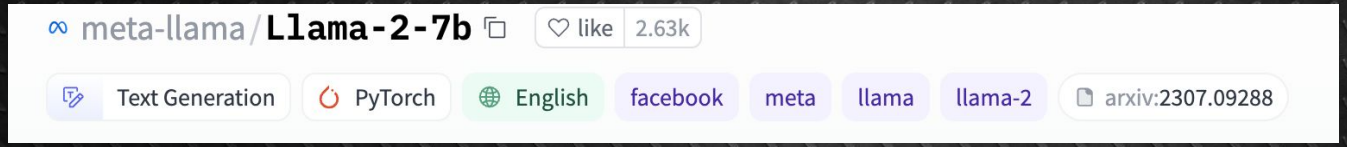


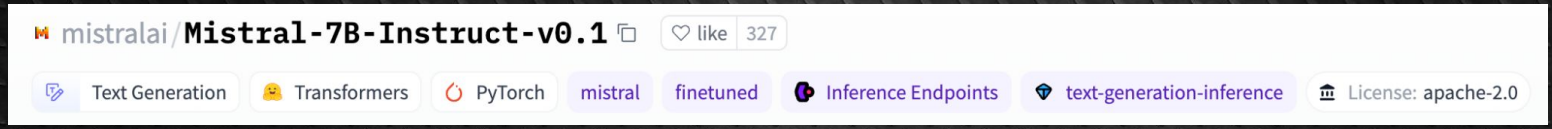
Image Source: [jakevdp - Introducing Scikit Learn Notebook]

OOB Models









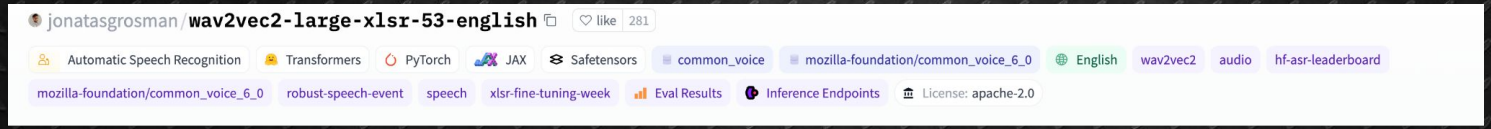


Image Source: HuggingFace

OOB Models

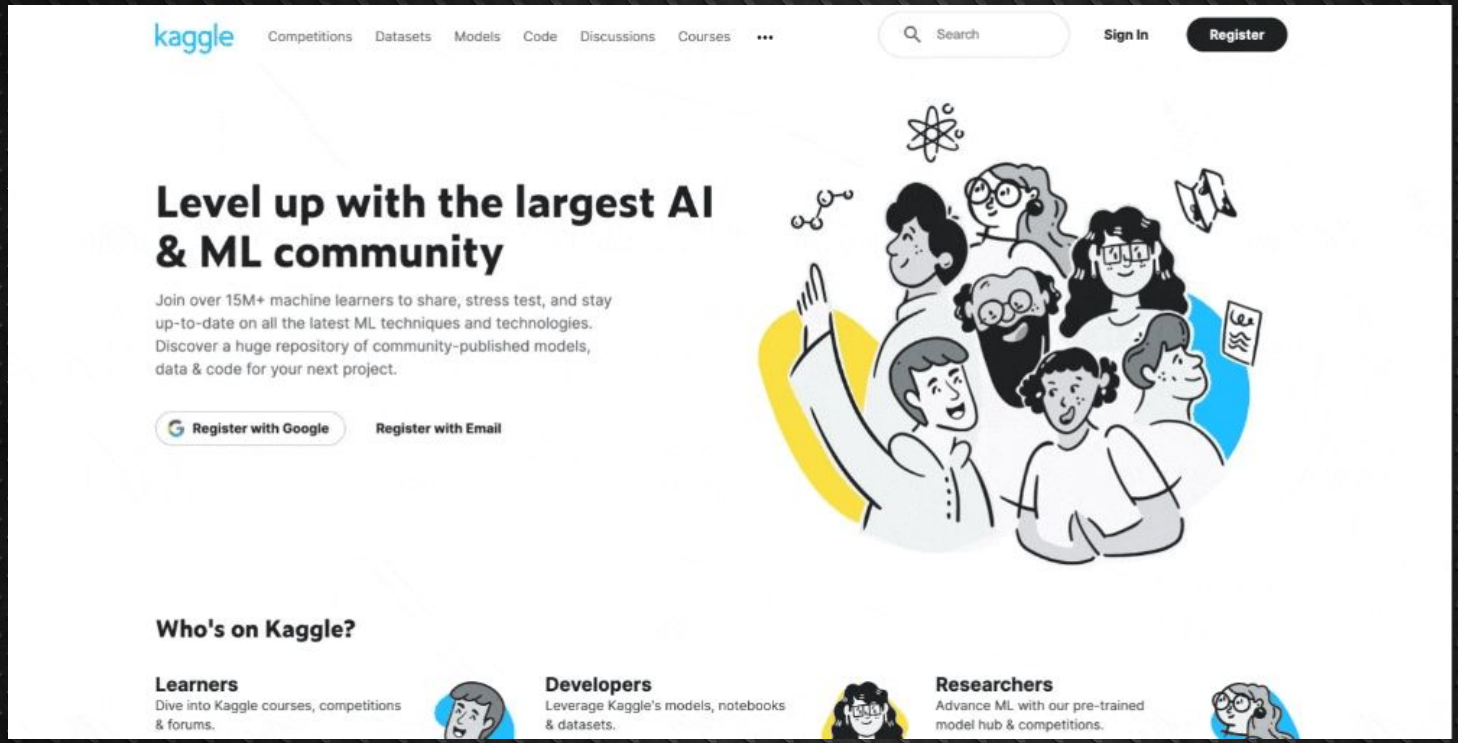
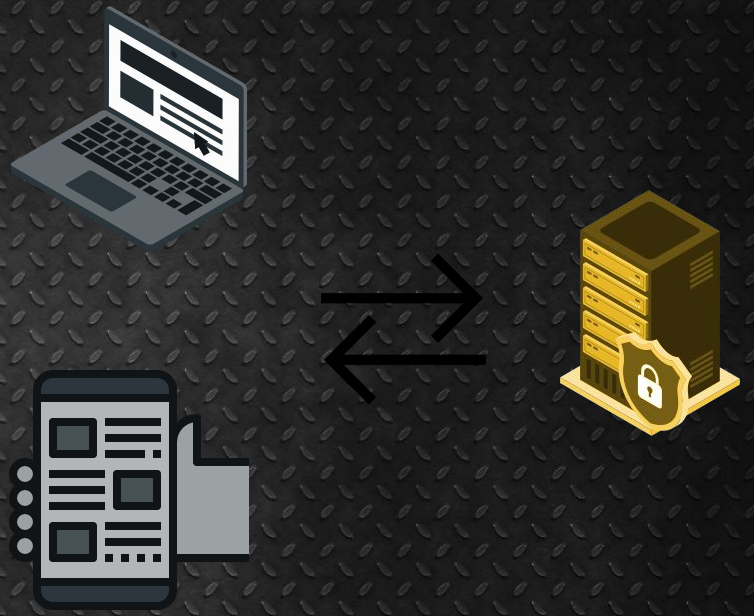


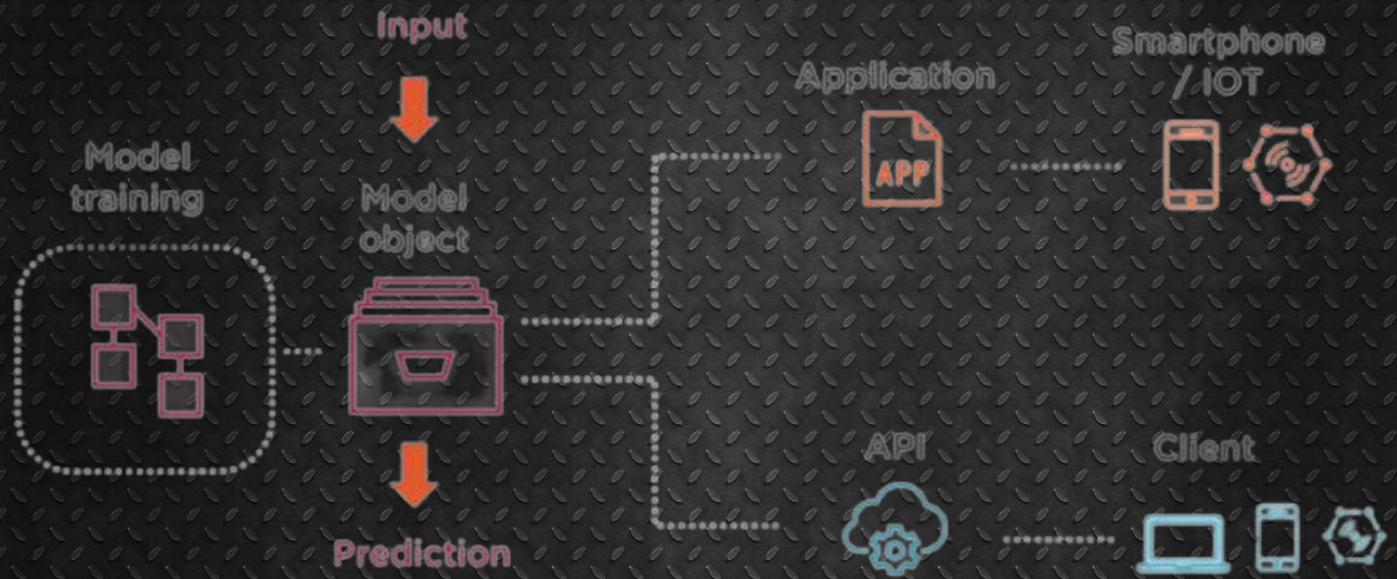
Image Source: Kaggle

Runtime

- At runtime, the previously developed model may be converted to another programming language
  - Python
  - Java / Scala
  - Go
- The Environment itself can take many forms
  - On-Prem
  - Cloud
  - On- Device



Model Serving



Machine Learning Supply Chain

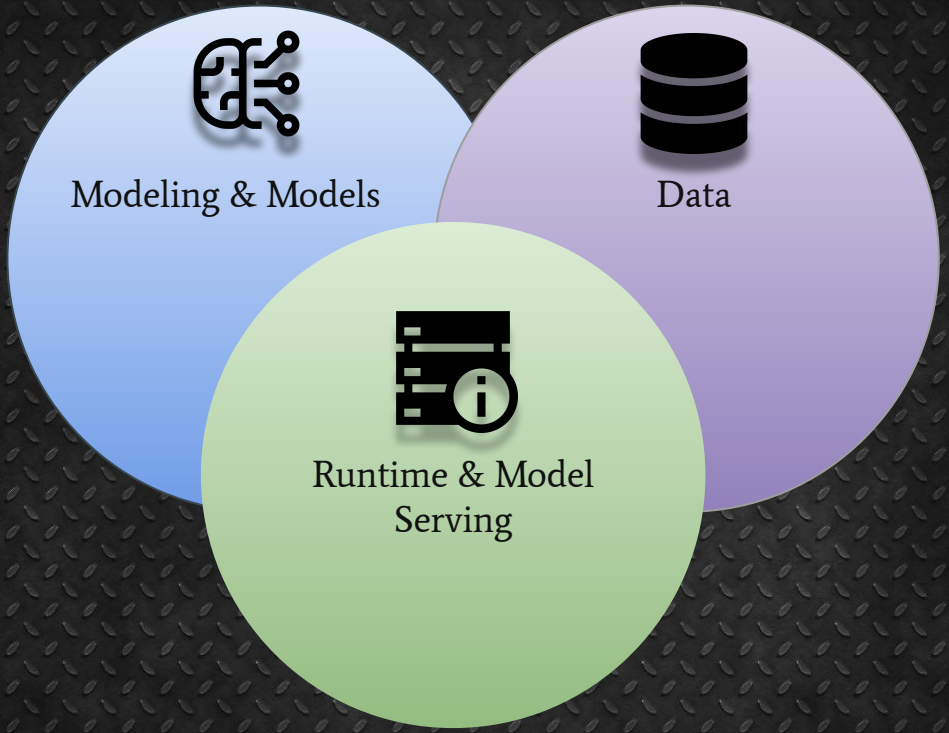
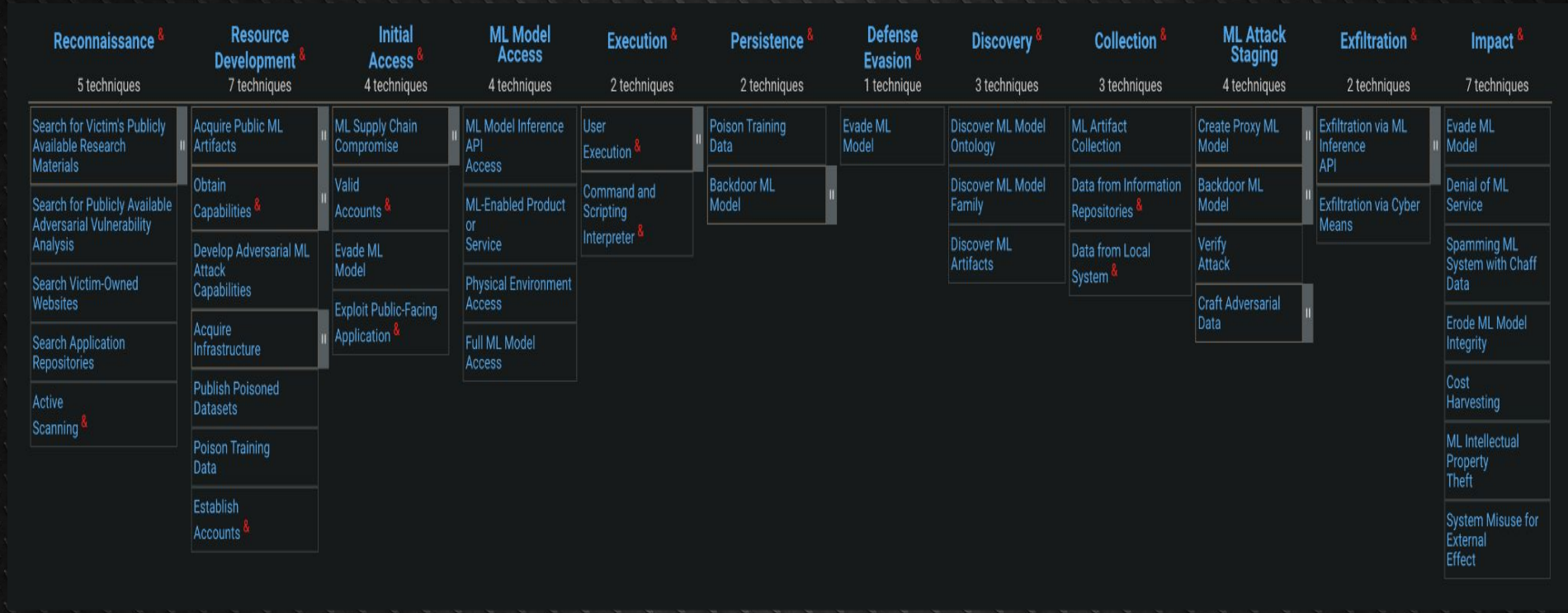


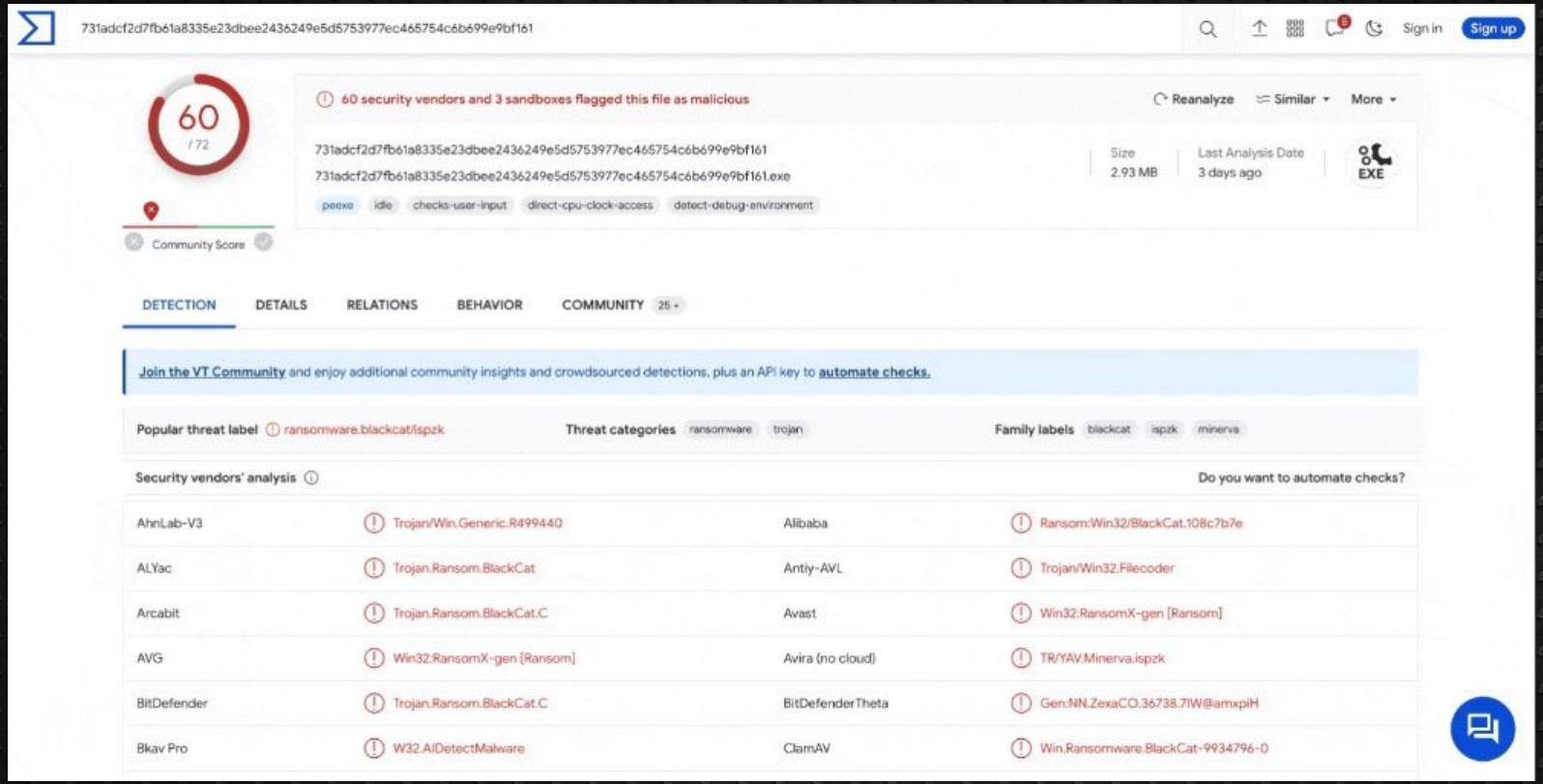
Illustration by: MidJourney

MITRE Atlas Overview



💡Resource: [MITRE Atlas]

Data: VirusTotal Poisoning



💡 Resource: VirusTotal Data Poisoning Case Study

Modeling: Code & Library Dependencies

SC MEDIA

TOPICS EVENTS PODCASTS RESEARCH RECOGNITION LEADERSHIP

Supply chain, DevSecOps

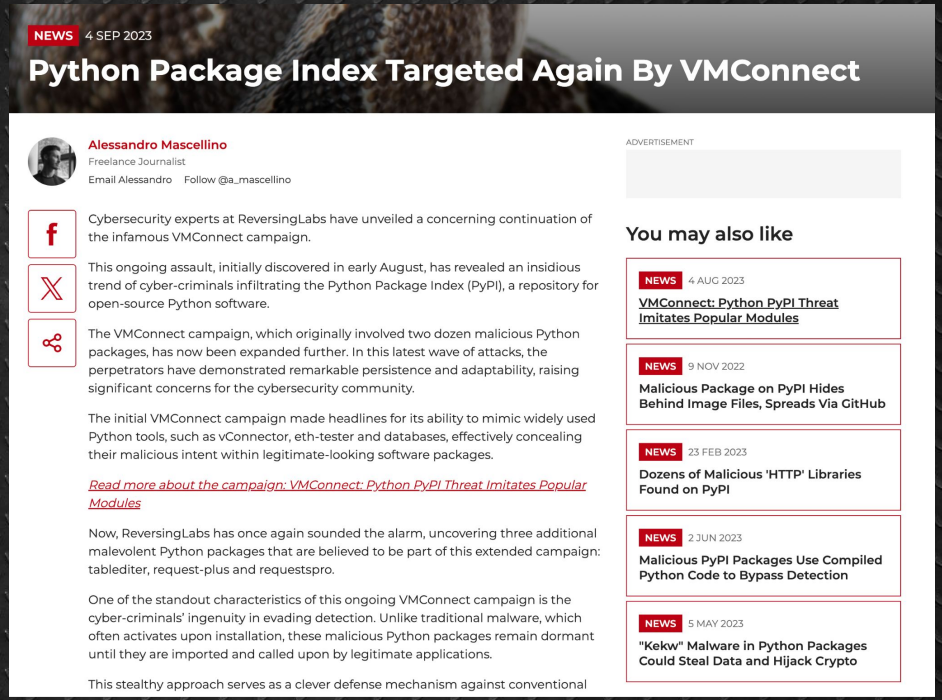
Millions of GitHub repositories potentially vulnerable to RepoJacking

Steve Zurier June 23, 2023

There are potentially millions of repositories on GitHub susceptible to RepoJacking attacks, according to Aqua Security research. (Adobe Stock Images)

Millions of GitHub repositories are potentially vulnerable to RepoJacking, which allows malicious actors to control an old repository if an organization changed its username on the open-source software development service.

According to research posted June 21 by Aqua Security's Nautilus research group, RepoJacking is when a malicious actor registers a username and creates a repository used by an organization in the past, but has since changed its username. A developer may think the repo is safe, but in reality it's controlled by the attacker and susceptible to malware.



NEWS 4 SEP 2023

Python Package Index Targeted Again By VMConnect

Alessandro Mascellino
Freelance Journalist
Email Alessandro Follow @a_mascellino

Cybersecurity experts at ReversingLabs have unveiled a concerning continuation of the infamous VMConnect campaign.

This ongoing assault, initially discovered in early August, has revealed an insidious trend of cyber-criminals infiltrating the Python Package Index (PyPI), a repository for open-source Python software.

The VMConnect campaign, which originally involved two dozen malicious Python packages, has now been expanded further. In this latest wave of attacks, the perpetrators have demonstrated remarkable persistence and adaptability, raising significant concerns for the cybersecurity community.

The initial VMConnect campaign made headlines for its ability to mimic widely used Python tools, such as vConnector, eth-tester and databases, effectively concealing their malicious intent within legitimate-looking software packages.

Read more about the campaign: VMConnect: Python PyPI Threat Imitates Popular Modules

Now, ReversingLabs has once again sounded the alarm, uncovering three additional malevolent Python packages that are believed to be part of this extended campaign: tablediter, request-plus and requestspro.

One of the standout characteristics of this ongoing VMConnect campaign is the cyber-criminals' ingenuity in evading detection. Unlike traditional malware, which often activates upon installation, these malicious Python packages remain dormant until they are imported and called upon by legitimate applications.

This stealthy approach serves as a clever defense mechanism against conventional

ADVERTISEMENT

You may also like

NEWS 4 AUG 2023
VMConnect: Python PyPI Threat Imitates Popular Modules

NEWS 9 NOV 2022
Malicious Package on PyPI Hides Behind Image Files, Spreads Via GitHub

NEWS 23 FEB 2023
Dozens of Malicious 'HTTP' Libraries Found on PyPI

NEWS 2 JUN 2023
Malicious PyPI Packages Use Compiled Python Code to Bypass Detection

NEWS 5 MAY 2023
"Kekw" Malware in Python Packages Could Steal Data and Hijack Crypto

💡Resource: Whitepaper - Towards Measuring Vulnerabilities and Exposures in Open-Source Packages

OOB Models: Open Source Model Vulnerabilities

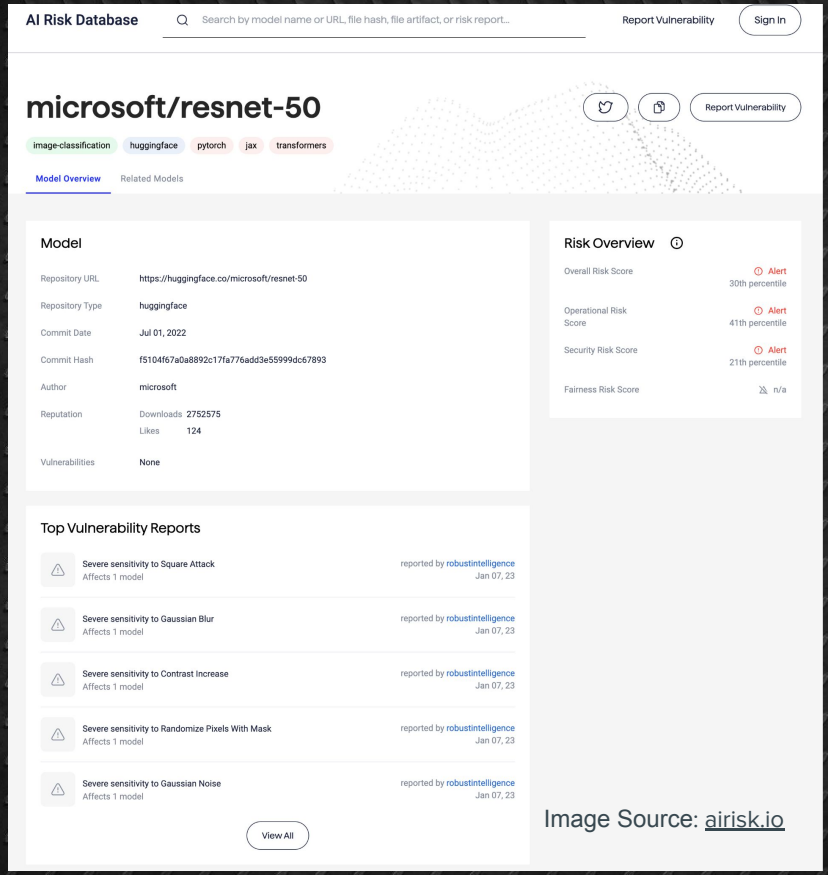
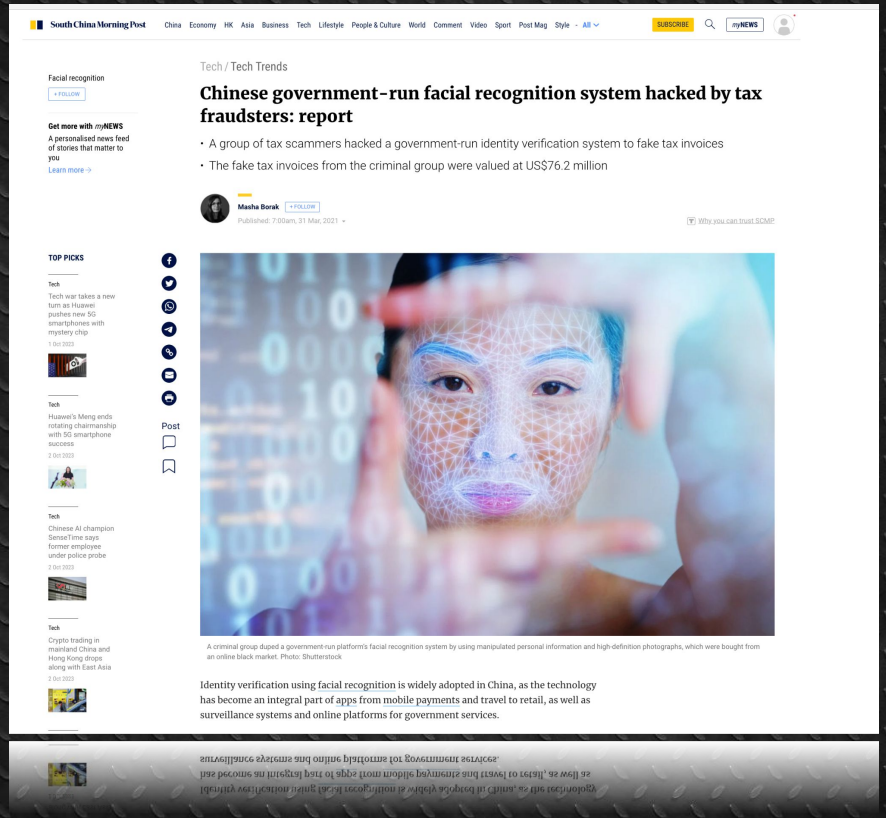


Image Source: airisk.io

💡Resource: Talk AI Village DEFCON 31 - Assessing the Vulnerabilities of the Open -Source Artificial Intelligence (AI) Landscape

Model Serving: Hijacking Facial Recognition System



South China Morning Post

China Economy HK Asia Business Tech Lifestyle People & Culture World Comment Video Sport Post Mag Style All

SUBSCRIBE myNEWS

Facial recognition

+ FOLLOW

Get more with myNEWS

A personalised news feed of stories that matter to you

Learn more

Tech / Tech Trends

Chinese government-run facial recognition system hacked by tax fraudsters: report

- A group of tax scammers hacked a government-run identity verification system to fake tax invoices
- The fake tax invoices from the criminal group were valued at US$76.2 million

Masha Borak + FOLLOW

Published: 7:00am, 31 Mar, 2021

Why you can trust SCMP

TOP PICKS

Tech

Tech war takes a new turn as Huawei pushes new 5G smartphones with mystery chip

1 Oct 2023

Tech

Huawei's Meng ends rotating chairmanship with 5G smartphone success

2 Oct 2023

Tech

Chinese AI champion SenseTime says former employee under police probe

2 Oct 2023

Tech

Crypto trading in mainland China and Hong Kong drops along with East Asia

2 Oct 2023

Post

A criminal group duped a government-run platform's facial recognition system by using manipulated personal information and high-definition photographs, which were bought from an online black market. Photo: Shutterstock

Identity verification using facial recognition is widely adopted in China, as the technology has become an integral part of apps from mobile payments and travel to retail, as well as surveillance systems and online platforms for government services.

💡 Resource: MITRE Case Study - Facial Recognition Hijack

💡 Resource: Whitepaper - Attacks for Extraction of Embedded Neural Network Models

Illustration by: MidJourney

What is available for Defenders?

A LOT of Ongoing Research:

- Curating Naturally Adversarial Datasets
- Adversarial Model Training

Curating Naturally Adversarial Datasets for Trustworthy AI in Healthcare

Sydney Pugh[1], **Ivan Ruchkin**[2], **Insup Lee**[1], **James Weimer**[3]
[1]University of Pennsylvania, Philadelphia, PA 19104
[2]University of Florida, Gainesville, FL 32611
[3]Vanderbilt University, Nashville, TN 37235
{sfpugh, lee}@seas.upenn.edu[1], iruchkin@ece.ufl.edu[2], james.weimer@vanderbilt.edu[3]

Abstract

Deep learning models have shown promising predictive accuracy for time-series healthcare applications. However, ensuring the robustness of these models is vital for building trustworthy AI systems. Existing research predominantly focuses on robustness to *synthetic adversarial examples*, crafted by adding imperceptible perturbations to clean input data. However, these synthetic adversarial examples do not accurately reflect the most challenging real-world scenarios, especially in the context of healthcare data. Consequently, robustness to synthetic adversarial examples may not necessarily translate to robustness against *naturally occurring adversarial examples*, which is highly desirable for trustworthy AI. We propose a method to curate datasets comprised of natural adversarial examples to evaluate model robustness. The method relies on probabilistic labels obtained from automated weakly-supervised labeling that combines noisy and cheap-to-obtain labeling heuristics. Based on these labels, our method adversarially orders the input data and uses this ordering to construct a sequence of increasingly adversarial datasets. Our evaluation on six medical case studies and three non-medical case studies demonstrates the efficacy and statistical validity of our approach to generating naturally adversarial datasets.

Introduction

Deep learning models have shown significant potential in analyzing medical time-series data to facilitate tasks such as disease diagnosis, anomaly detection, and intervention planning (Faust et al. 2018; Morid, Sheng, and Dunbar 2023). However, deep learning models have also been shown to be brittle and non-robust (Szegedy et al. 2013; Alcorn et al. 2019). A typical approach to analyzing the robustness of such models is to evaluate their performance against adversarial examples — perturbed inputs that cause misclassification while being nearly imperceptible to humans.

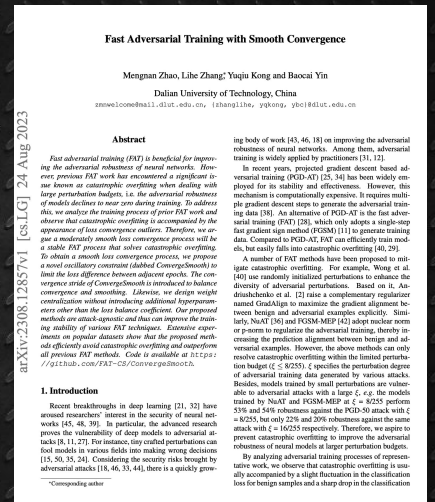
Much work has been done to evaluate model robustness to adversarial examples, specifically focusing on *synthetic* adversarial examples. These examples are generated by intentionally adding small perturbations to *labeled* input data. However, synthetic adversarial examples for medical time-series data are problematic for two primary reasons. Firstly, they generally do not resemble adversarial examples that would be encountered in the real world (Koh et al. 2021). Simply adding random perturbations to medical data usually yields invalid and unrealistic examples. Specialized perturbations can mitigate this, but defining them is a complex task that requires significant domain expertise. Furthermore, defining perturbations to cover all clinically plausible real-world adversarial attacks is infeasible. Moreover, improving robustness to synthetic adversarial examples does not necessarily improve robustness to natural adversarial examples.

Secondly, the lack of labeled medical time-series datasets presents a challenge in the generation of synthetic adversarial examples. Such datasets are time-consuming and expensive to create. A common way to do so is to collect data via an observational study (Bonafide et al. 2017) and then have domain experts manually label the data. However, such a study is a major commitment of effort and time.

Instead, we focus on *natural* adversarial examples — real patient examples that are difficult to classify (Hendrycks et al. 2021). Natural adversarial examples capture the inherent variations and uncertainties present in real-world medical data that effectively deceive models. Therefore, evaluating the robustness of medical deep learning models against such examples becomes crucial to ensuring model robustness in the real world. We aim to identify natural adversarial examples in *unlabeled* medical time-series data so that we can produce naturally adversarial *datasets*.

Robustness is commonly evaluated by examining a model's accuracy on a labeled adversarial dataset. While fully-supervised labels are expensive to procure, obtaining weak labels is a cheap and quick alternative. However, assessing accuracy with respect to weak labels becomes challenging due to the uncertainty surrounding the correctness of the weak labels. In our experience, weak labels for adversarial examples are prone to significant inaccuracies due to the inherent difficulty in classifying these examples. Hence a robust model would tend to disagree with the weak labels of natural adversarial examples. Therefore, we aim to curate *adversarially ordered* datasets, wherein the proportion and severity of adversarial examples gradually increase per dataset, correspondingly leading to a diminishing accuracy of each dataset's weak labels. In practice, a robust model would be expected to show decreasing accuracy with respect to our weak labels on adversarially ordered datasets.

This paper introduces a weakly-supervised method for curating adversarially ordered datasets containing natural adversarial examples. A key step of our method is to proba-



Fast Adversarial Training with Smooth Convergence

Mengnan Zhao, Lihe Zhang*, Yuqiu Kong and Baocai Yin
Dalian University of Technology, China
zmnwelcome@mail.dlut.edu.cn, {zhanglihe, ykong, ybc}@dlut.edu.cn

Abstract

Fast adversarial training (FAT) is beneficial for improving the adversarial robustness of neural networks. However, previous FAT work has encountered a significant issue known as catastrophic overfitting when dealing with large perturbation budgets, i.e. the adversarial robustness of models declines to near zero during training. To address this, we analyze the training process of prior FAT work and observe that catastrophic overfitting is accompanied by the appearance of loss convergence outliers. Therefore, we argue a moderately smooth loss convergence process will be a stable FAT process that solves catastrophic overfitting. To obtain a smooth loss convergence process, we propose a novel oscillatory constraint (dubbed ConvergeSmooth) to limit the loss difference between adjacent epochs. The convergence stride of ConvergeSmooth is introduced to balance convergence and smoothing. Likewise, we design weight centralization without introducing additional hyperparameters other than the loss balance coefficient. Our proposed methods are attack-agnostic and thus can improve the training stability of various FAT techniques. Extensive experiments on popular datasets show that the proposed methods efficiently avoid catastrophic overfitting and outperform all previous FAT methods. Code is available at https://github.com/FAT-CS/ConvergeSmooth.

1. Introduction

Recent breakthroughs in deep learning [21, 32] have aroused researchers' interest in the security of neural networks [45, 48, 39]. In particular, the advanced research proves the vulnerability of deep models to adversarial attacks [8, 11, 27]. For instance, tiny crafted perturbations can fool models in various fields into making wrong decisions [15, 50, 35, 24]. Considering the security risks brought by adversarial attacks [18, 46, 33, 44], there is a quickly growing body of work [43, 46, 18] on improving the adversarial robustness of neural networks. Among them, adversarial training is widely applied by practitioners [31, 12].

In recent years, projected gradient descent based adversarial training (PGD-AT) [25, 34] has been widely employed for its stability and effectiveness. However, this mechanism is computationally expensive. It requires multiple gradient descent steps to generate the adversarial training data [38]. An alternative of PGD-AT is the fast adversarial training (FAT) [28], which only adopts a single-step fast gradient sign method (FGSM) [11] to generate training data. Compared to PGD-AT, FAT can efficiently train models, but easily falls into catastrophic overfitting [40, 29].

A number of FAT methods have been proposed to mitigate catastrophic overfitting. For example, Wong et al. [40] use randomly initialized perturbations to enhance the diversity of adversarial perturbations. Based on it, Andriushchenko et al. [2] raise a complementary regularizer named GradAlign to maximize the gradient alignment between benign and adversarial examples explicitly. Similarly, NuAT [36] and FGSM-MEP [42] adopt nuclear norm or p-norm to regularize the adversarial training, thereby increasing the prediction alignment between benign and adversarial examples. However, the above methods can only resolve catastrophic overfitting within the limited perturbation budget ($\xi \leq 8/255$). ξ specifies the perturbation degree of adversarial training data generated by various attacks. Besides, models trained by small perturbations are vulnerable to adversarial attacks with a large ξ, *e.g.* the models trained by NuAT and FGSM-MEP at $\xi = 8/255$ perform 53% and 54% robustness against the PGD-50 attack with $\xi = 8/255$, but only 22% and 20% robustness against the same attack with $\xi = 16/255$ respectively. Therefore, we aspire to prevent catastrophic overfitting to improve the adversarial robustness of neural models at larger perturbation budgets.

By analyzing adversarial training processes of representative work, we observe that catastrophic overfitting is usually accompanied by a slight fluctuation in the classification loss for benign samples and a sharp drop in the classification

*Corresponding author

💡Resource: Whitepaper Curating Naturally Adversarial Datasets

💡Resource: Fast Adversarial Training with Smooth Convergence

What is available for Defenders?

Mitigations

Mitigations represent security concepts and classes of technologies that can be used to prevent a technique or sub-technique from being successfully executed.

This draft of mitigations are defined based on current ATLAS case studies.
Feedback and contributions are welcome - please join the #mitigations channel on Slack.

The table below lists mitigations from MITRE ATLAS™. Scroll through the table or use the filter to narrow down the information.

Search for keywords
Adversarial

ID	Name	Description
AML.M0003	Model Hardening	Use techniques to make machine learning models robust to adversarial inputs such as adversarial training or network distillation.
AML.M0006	Use Ensemble Methods	Use an ensemble of models for inference to increase robustness to adversarial inputs. Some attacks may effectively evade one model or model family but be ineffective against others.
AML.M0008	Validate ML Model	Validate that machine learning models perform as intended by testing for backdoor triggers or adversarial bias.
AML.M0010	Input Restoration	Preprocess all inference data to nullify or reverse potential adversarial perturbations.
AML.M0015	Adversarial Input Detection	Detect and block adversarial inputs or atypical queries that deviate from known benign behavior, exhibit behavior patterns observed in previous attacks or that come from potentially malicious IPs. Incorporate adversarial detection algorithms into the ML system prior to the ML model.

💡 Resource: MITRE Atlas Mitigations

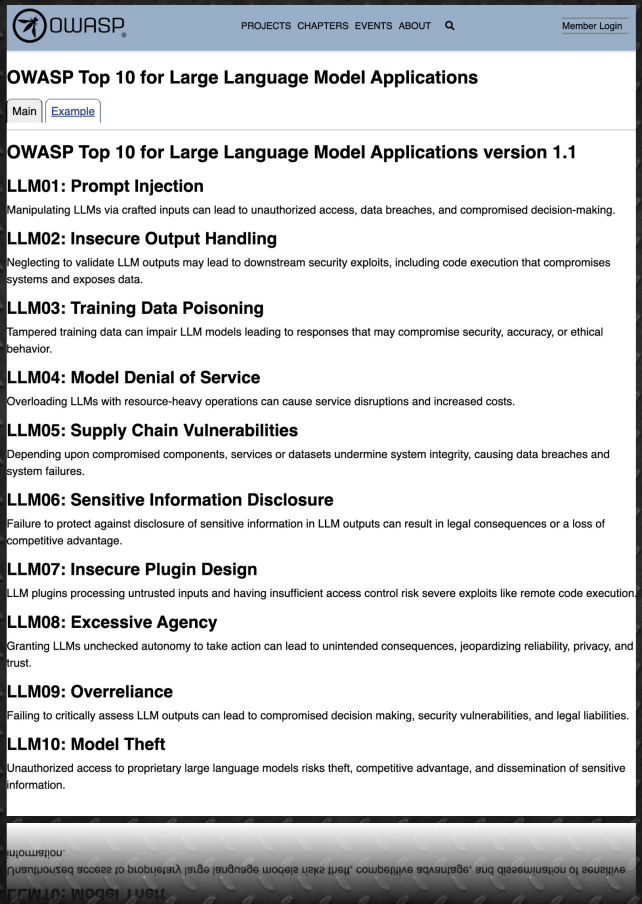
What is available for Defenders?

Open Source CALDERA Plugin

- Developed for Adversary Emulation AI
- TTPs defined by MITRE Atlas

💡Resource: MITRE Arsenal Repository

What is available for Defenders?



OWASP PROJECTS CHAPTERS EVENTS ABOUT Member Login

OWASP Top 10 for Large Language Model Applications

Main | Example

OWASP Top 10 for Large Language Model Applications version 1.1

LLM01: Prompt Injection

Manipulating LLMs via crafted inputs can lead to unauthorized access, data breaches, and compromised decision-making.

LLM02: Insecure Output Handling

Neglecting to validate LLM outputs may lead to downstream security exploits, including code execution that compromises systems and exposes data.

LLM03: Training Data Poisoning

Tampered training data can impair LLM models leading to responses that may compromise security, accuracy, or ethical behavior.

LLM04: Model Denial of Service

Overloading LLMs with resource-heavy operations can cause service disruptions and increased costs.

LLM05: Supply Chain Vulnerabilities

Depending upon compromised components, services or datasets undermine system integrity, causing data breaches and system failures.

LLM06: Sensitive Information Disclosure

Failure to protect against disclosure of sensitive information in LLM outputs can result in legal consequences or a loss of competitive advantage.

LLM07: Insecure Plugin Design

LLM plugins processing untrusted inputs and having insufficient access control risk severe exploits like remote code execution.

LLM08: Excessive Agency

Granting LLMs unchecked autonomy to take action can lead to unintended consequences, jeopardizing reliability, privacy, and trust.

LLM09: Overreliance

Failing to critically assess LLM outputs can lead to compromised decision making, security vulnerabilities, and legal liabilities.

LLM10: Model Theft

Unauthorized access to proprietary large language models risks theft, competitive advantage, and dissemination of sensitive information.

💡 Resource: OWASP Top 10 LLM Apps

What is available for Defenders?

Standards About us News Taking part Store Search

← ICS ← 35 ← 35.020

ISO/IEC FDIS 5338

Information technology — Artificial intelligence — AI system life cycle processes

General information

Status : Under development

Edition : 1 **Number of pages** : 39

Technical Committee : ISO/IEC JTC 1/SC 42 Artificial intelligence

ICS : 35.020 Information technology (IT) in general

💡 Resource: ISO 5338

What is available for Defenders?

- **Rebuff**: Prompt Injection defense framework.
- **ModelScan**: Identify unsafe code in models.
- **NB Defense**: Jupyter Notebook security.

💡 Resource: Protect AI Open Sources Three Tools to Help Organizations Secure AI/ML Environments from Threats

Key Takeaways

- ML Engineering = Software Engineering.
- ML Security > LLM Prompt Injections.
- Data can be a blind spot, good data governance is the best mitigation.
- Explainability can be your best friend or your worst enemy.

Thank you!

danielfernandez@infosec.exchange

@danielfernandez

/in/dafnz

danielfernandez.medium.com

Slides: github.com/dnlfdz/talks

hey@dafnz.com