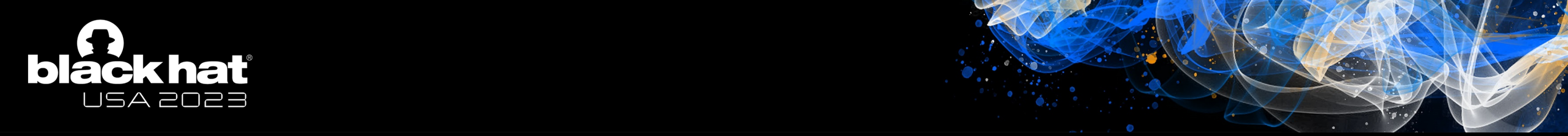
Poisoning Web-Scale Datasets is Practical

Speaker(s): Will Pearce



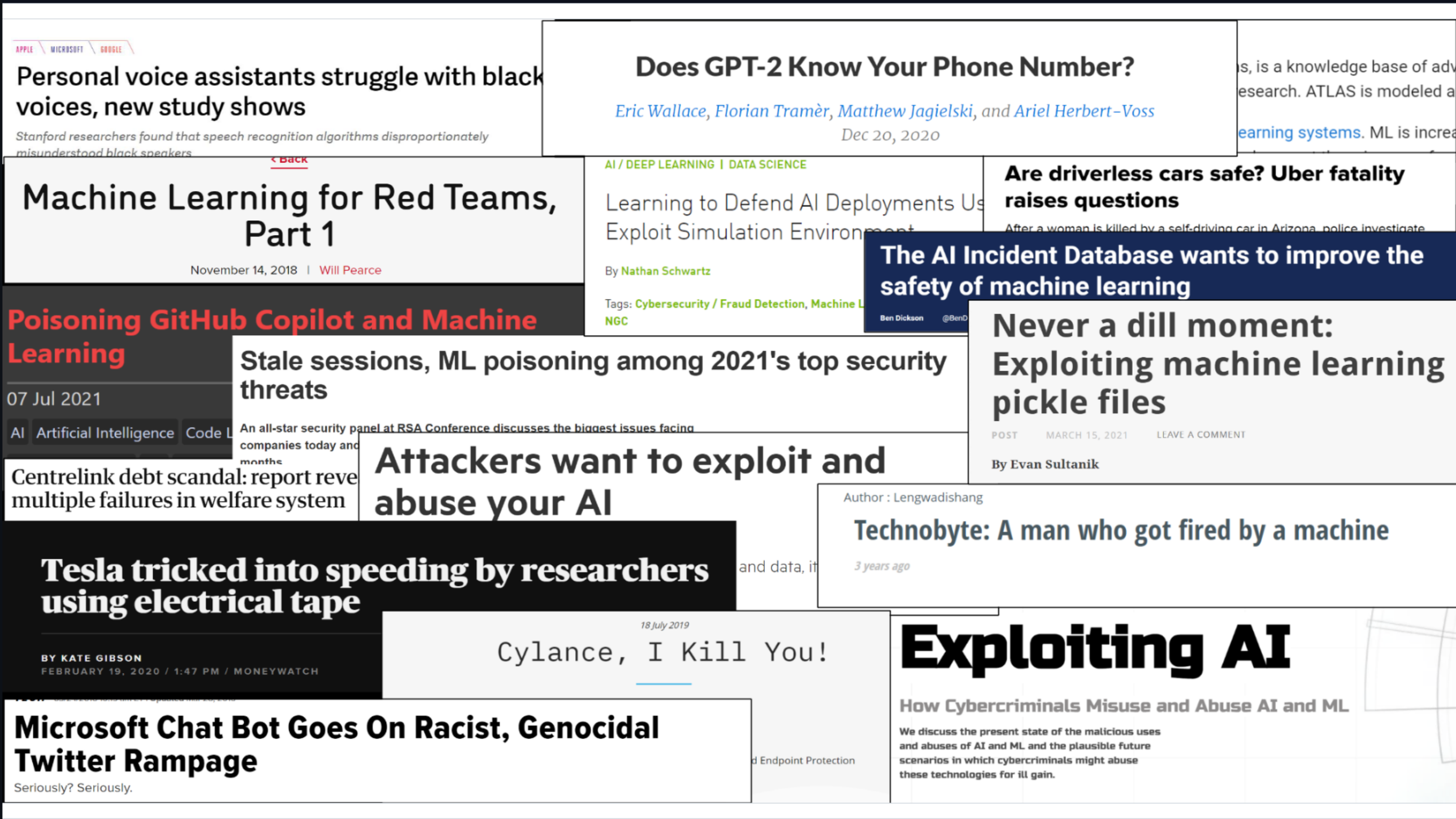
Researchers

[1] Google, [2] ETH Zurich, [3] NVIDIA, [4] Robust Intelligence

Nicholas Carlini [1],
Matthew Jagielski [1],
Christopher A. Choquette-Choo [1],
Daniel Paleka [2],
Will Pearce [3],
Hyrum Anderson [4],
Andreas Terzis [1],
Kurt Thomas [1],
Florian Tramèr [2]

ML research focuses on potential **impact** and less on whether an attack is **possible**

Why *should* you have whitebox gradient access to a model? (you shoudn't)


APPLE MICROSOFT GOOGLE
Personal voice assistants struggle with black voices, new study shows
Stanford researchers found that speech recognition algorithms disproportionately misunderstood black speakers
Does GPT-2 Know Your Phone Number?
Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss
Dec 20, 2020
s, is a knowledge base of adv
esearch. ATLAS is modeled a
earning systems. ML is increa
< Back
Machine Learning for Red Teams, Part 1
November 14, 2018 | Will Pearce
AI / DEEP LEARNING | DATA SCIENCE
Learning to Defend AI Deployments Us
Exploit Simulation Environment
By Nathan Schwartz
Tags: Cybersecurity / Fraud Detection, Machine L
NGC
Are driverless cars safe? Uber fatality raises questions
After a woman is killed by a self-driving car in Arizona, police investigate
The AI Incident Database wants to improve the safety of machine learning
Ben Dickson @BenD
Poisoning GitHub Copilot and Machine Learning
07 Jul 2021
AI Artificial Intelligence Code L
Stale sessions, ML poisoning among 2021's top security threats
An all-star security panel at RSA Conference discusses the biggest issues facing companies today and
months
Never a dill moment: Exploiting machine learning pickle files
POST MARCH 15, 2021 LEAVE A COMMENT
By Evan Sultanik
Centrelink debt scandal: report reve
multiple failures in welfare system
Attackers want to exploit and abuse your AI
and data, it
Author : Lengwadishang
Technobyte: A man who got fired by a machine
3 years ago
Tesla tricked into speeding by researchers using electrical tape
BY KATE GIBSON
FEBRUARY 19, 2020 / 1:47 PM / MONEYWATCH
18 July 2019
Cylance, I Kill You!
Exploiting AI
How Cybercriminals Misuse and Abuse AI and ML
We discuss the present state of the malicious uses and abuses of AI and ML and the plausible future scenarios in which cybercriminals might abuse these technologies for ill gain.
Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage
Seriously? Seriously.
d Endpoint Protection

Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning

Hyrum S. Anderson
Endgame, Inc.
hyrum@endgame.com

Anant Kharkar
University of Virginia
agk7uc@virginia.edu

Bobby Filar
Endgame, Inc.
bfilar@endgame.com

David Evans
University of Virginia
evans@virginia.edu

Phil Roth
Endgame, Inc.
proth@endgame.com

ABSTRACT

new samples whose features and labels follow the same distribution

Hidden Voice Commands

Nicholas Carlini*
University of California, Berkeley

Pratyush Mishra
University of California, Berkeley

Tavish Vaidya
Georgetown University

Yuankai Zhang
Georgetown University

Micah Sherr
Georgetown University

Clay Shields
Georgetown University

David Wagner
University of California, Berkeley

Wenchao Zhou
Georgetown University

Abstract

Voice interfaces are becoming more ubiquitous and are

command may recognize it as an unwanted command and cancel it, or otherwise take action. This motivates the question we study in this paper: can an attacker cre-

The Space of Transferable Adversarial Examples

Florian Tramèr[1], Nicolas Papernot[2], Ian Goodfellow[3], Dan Boneh[1], and Patrick McDaniel[2]

[1]Stanford University, [2]Pennsylvania State University, [3]Google Brain

High Accuracy and High Fidelity Extraction of Neural Networks

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

[†]Northeastern University
[*]Google Research

Subpopulation Data Poisoning Attacks

Matthew Jagielski
jagielski@ccs.neu.edu
Northeastern University

Giorgio Severi
severi.g@northeastern.edu
Northeastern University

Niklas Pousette Harger
pousetteharger.n@husky.neu.edu
Northeastern University

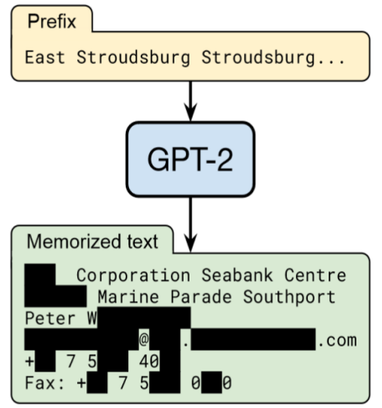
ABSTRACT

Machine learning systems are deployed in critical setting
might fail in unexpected ways, impacting the accurac
predictions. Poisoning attacks against machine learni
adversarial modification of data used by a machine le
gorithm to selectively change its output when it is de
this work, we introduce a novel data poisoning attack ca
population attack, which is particularly relevant when da
large and diverse. We design a modular framework for s
tion attacks, instantiate it with different building blocks,

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]

ortheastern University [5]*OpenAI* [6]*Harvard* [7]*Apple*



Poisoning Attacks against Support Vector Machines

Battista Biggio BATTISTA.BIGGIO@DIEE.UNICA.IT
Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy

Blaine Nelson BLAINE.NELSON@WSII.UNI-TUEBINGEN.DE
Pavel Laskov PAVEL.LASKOV@UNI-TUEBINGEN.DE
Wilhelm Schickard Institute for Computer Science, University of Tübingen, Sand 1, 72076 Tübingen, Germany

Abstract

We investigate a family of poisoning attacks against Support Vector Machines (SVM)

employ learning to help solve so-called *big-data problems* and these include a number of security-related problems particularly focusing on identifying malicious

I get my POC's on Arxiv

and so should you!

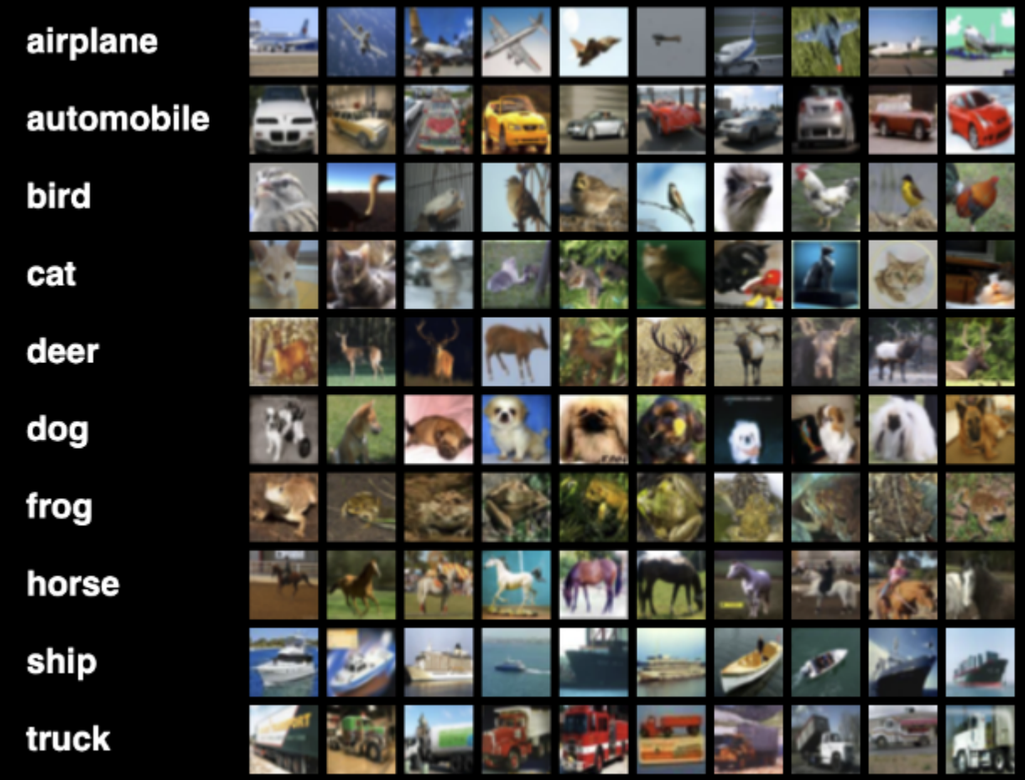
Let’s suppose you want to train a State-of-the-Art model

Or even just a regular one

1990s – MNIST (50k images, 28x28)

~~1990s — MNIST (50k images, 28x28)~~

2000s - CIFAR-10 (50k images, 32x32, 10 classes)



~~2000s - CIFAR-10 (50k images, 32x32, 10 classes)~~

2010s - ImageNet (1M images, 1k classes)

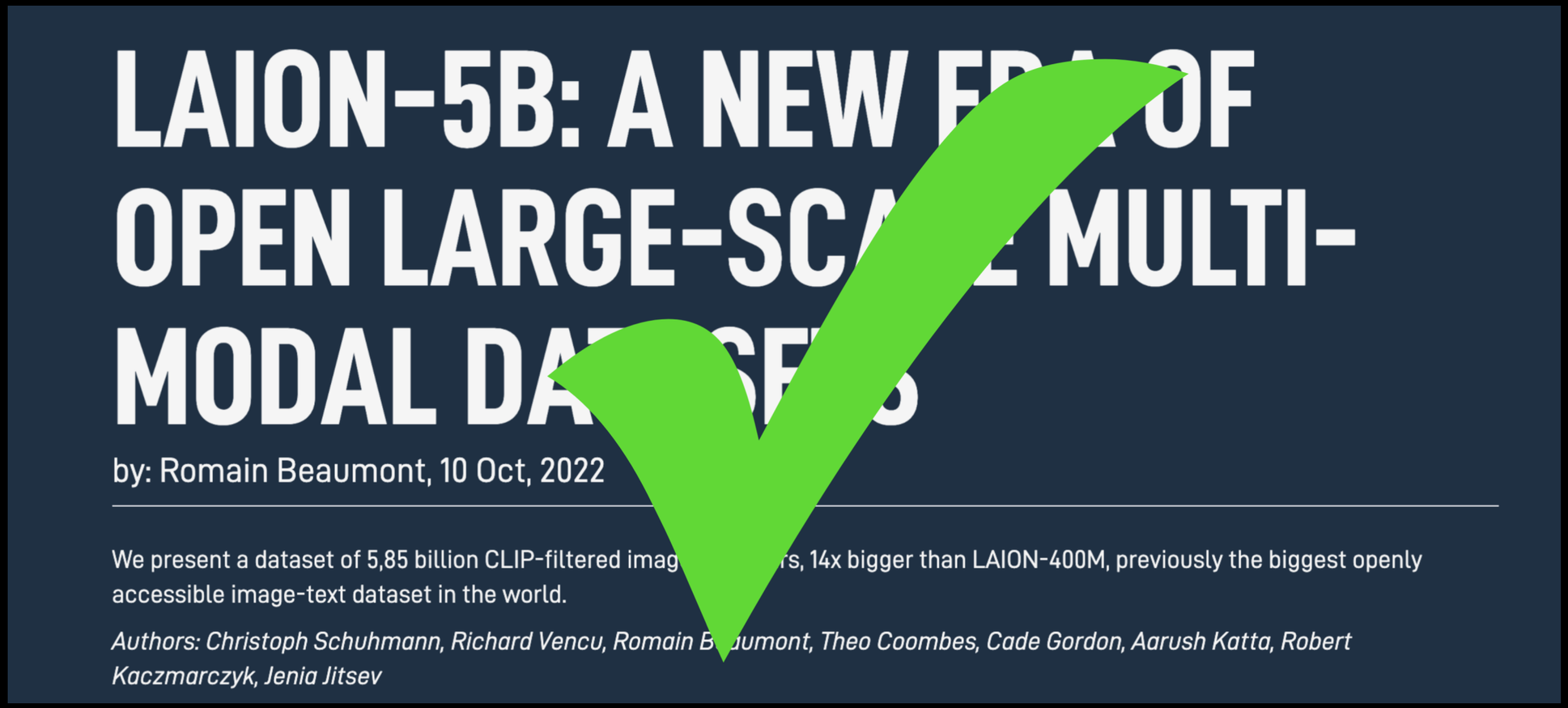
~~2010s - ImageNet (1M images, 1k classes)~~

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 10 Oct, 2022

We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev



LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 10 Oct, 2022

We present a dataset of 5,85 billion CLIP-filtered imag rs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

Authors: Christoph Schuhmann, Richard Vencu, Romain B umont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev

Introducing ChatGPT research release Try↗ Learn more›

OpenAI

API RESEARCH BLOG ABOUT

CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

January 5, 2021
15 minute read

stability.ai

API News FAQ English

Stable Diffusion Public Release

A Few Issues…

Can’t host 5 billion images for several reasons –

- Cost (~300TB)
- Licensing, copyright, trademarks, etc.
- PII, Illegal content (it’s 300TB of the *internet*)

Question:

How do you distribute a dataset with 5 billion images?

Answer:

You don’t.

Dataset Preview

API

Split

train

▸ The full dataset viewer is not available (click to read why). Only showing a preview of the rows.

SAMPLE_ID (int64)	URL (string)	TEXT (string)	HEIGHT (float64)	WIDTH (float64)	LICENSE (string)	LANGUAGE (string)	NSFW (string)	similarity (float64)
2,833,858,008,613	"http://3.bp.blogspot.com/-6uKj8avN8oc/UsvAhUlpeSI/AAAAAAAACL8/ce31UUzapow/w1200-h630-p-k-no-…	"Peugeot 308 2013 sedan"	1,200	630	"?"	"et"	"UNLIKELY"	0.264928
2,309,126,086,276	"https://ziarmm.ro/wp-content/uploads/2018/05/timotei-bel-364x245.jpg"	"Episcopia Ortodoxa a Maramuresului si Satmarului are un nou Arhiereu vicar"	364	245	"?"	"ro"	"UNLIKELY"	0.316854
4,086,983,003,858	"https://www.popularlibros.com/imagenes.webp/9788497/978849740420.webp"	"DON QUIJOTE DE LA MANCHA (SELECCIÓN DE TEXTOS)"	300	463	"?"	"es"	"UNLIKELY"	0.308755
1,748,947,083,164	"https://thumbs.dreamstime.com/t/e%C5%84ski-i-m%C4%99ski-portret-dama-outdoors-i-facet-%C5%9Blubna-…	"Żeński i męski portret Dama outdoors i facet Ślubna para w miłości, zakończenie…	240	160	"?"	"pl"	"UNSURE"	0.262129
2,145,720,073,183	"https://journalcinelyon1.files.wordpress.com/2016/02/ouagawood.jpg?w=261&"	"Ouagawood"	261	344	"?"	"zu"	"UNLIKELY"	0.322668
3,954,009,005,813	"https://media.wheels.ca/vehicles/3897/3086775/2016-BMW-3-Series-3086775-10-sm.jpg"	"Car Images"	640	480	"?"	"fr"	"UNLIKELY"	0.26376
3,544,617,000,614	"https://img.shopstyle-cdn.com/sim/79/c3/79c3329230408204e8c9e9d2b50031d3_xlarge/sezane…	"Sézane Sezane Fall Winter 2018 Black Wool Knitwear"	328	319	"?"	"fy"	"UNLIKELY"	0.311787
1,897,324,038,603	"https://www.inter-medio.com/wp-content/uploads/imagen-corporativa-zgz-900x300.png"	"Imagen Corporativa y Diseño Gráfico Zaragoza"	900	300	"?"	"es"	"UNLIKELY"	0.321683
3,861,563,012,544	"http://tyba.com.br/fotos/foto/cd401_102.jpg"	"Foto feita com drone da fachada lateral do Museu Oscar Niemeyer - também conhecido…	500	333	"?"	"pt"	"UNLIKELY"	0.328382
[illegible]	"https://4gnews.pt/wp-content/uploads/2018/05/Just-	"Android e iPhone - Esta aplicação de	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

dataset viewer is not available (click to read why). Only showing a pre

(int64)	URL (string)	TEXT (stri
008,613	"http://3.bp.blogspot.com/-6uKj8avN8oc/UsvAhUlpeSI/AAAAAAAACL8/ce31UUzapow/w1200-h630-p-k-no-…	"Peugeot 3
086,276	"https://ziarmm.ro/wp-content/uploads/2018/05/timotei-bel-364x245.jpg"	"Episcopia Satmarului
003,858	"https://www.popularlibros.com/imagenes.webp/9788497/978849740420.webp"	"DON QUIJO TEXTOS)"
083,164	"https://thumbs.dreamstime.com/t/e%C5%84ski-i-m%C4%99ski-portret-dama-outdoors-i-facet-%C5%9Blubna-…	"Żeński i facet Ślub
073,183	"https://journalcinelyon1.files.wordpress.com/2016/02/ouagawood.jpg?w=261&"	"Ouagawood
005,813	"https://media.wheels.ca/vehicles/3897/3086775/2016-BMW-3-Series-3086775-10-sm.jpg"	"Car Image
000,614	"https://img.shopstyle-cdn.com/sim/79/c3/79c3329230408204e8c9e9d2b50031d3_xlarge/sezane…	"Sézane Se Knitwear"
038,603	"https://www.inter-medio.com/wp-content/uploads/imagen-corporativa-zgz-900x300.png"	"Imagen Co Zaragoza"
012,544	"http://tyba.com.br/fotos/foto/cd401_102.jpg"	"Foto feit Museu Osca
	"https://4gnews.pt/wp-content/uploads/2018/05/Just-	"Android e

Dataset Preview

Split

train

▸ The full dataset viewer

API

README.md

img2dataset

pypi v1.41.0 | Open in Colab | try on gitpod | chat 4046 online

Easily turn large sets of image urls to an image dataset. Can download, resize and package 100M urls in 20h on one machine.

Also supports saving captions for url+caption datasets.

Install

```
pip install img2dataset
```

For better performance, it's highly recommended to set up a fast dns resolver, see this section

SAMPLE_ID (int64)	URL (s					NSFW (string)	similarity (float64)
2,833,858,008,613	"http /AAAA					"UNLIKELY"	0.264928
2,309,126,086,276	"https /timo					"UNLIKELY"	0.316854
4,086,983,003,858	"https /9788					"UNLIKELY"	0.308755
1,748,947,083,164	"https %99sk					"UNSURE"	0.262129
2,145,720,073,183	"https /2016					"UNLIKELY"	0.322668
3,954,009,005,813	"https /2016					"UNLIKELY"	0.26376
3,544,617,000,614	"https /c3/7					"UNLIKELY"	0.311787
1,897,324,038,603	"https /imagen-corporativa-zgz-900x300.png"	Zaragoza"	900	300	"?"	"es"	"UNLIKELY"	0.321683
3,861,563,012,544	"http://tyba.com.br/fotos/foto/cd401_102.jpg"	"Foto feita com drone da fachada lateral do Museu Oscar Niemeyer - também conhecido…	500	333	"?"	"pt"	"UNLIKELY"	0.328382
	"https://4gnews.pt/wp-content/uploads/2018/05/Just-	"Android e iPhone - Esta aplicação de						

Except...

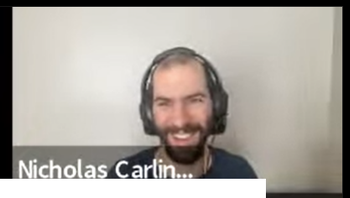
Domain registrations expire

…and when they expire,
anyone can buy them…


So anyway I now own
0.01% of LAION.
Nicholas Carlin...

Nicolas Carlini

Nicolas Carlini



So
0

Published as a conference paper at ICLR 2022

POISONING AND BACKDOORING CONTRASTIVE LEARNING

Nicholas Carlini
Google

Andreas Terzis
Google

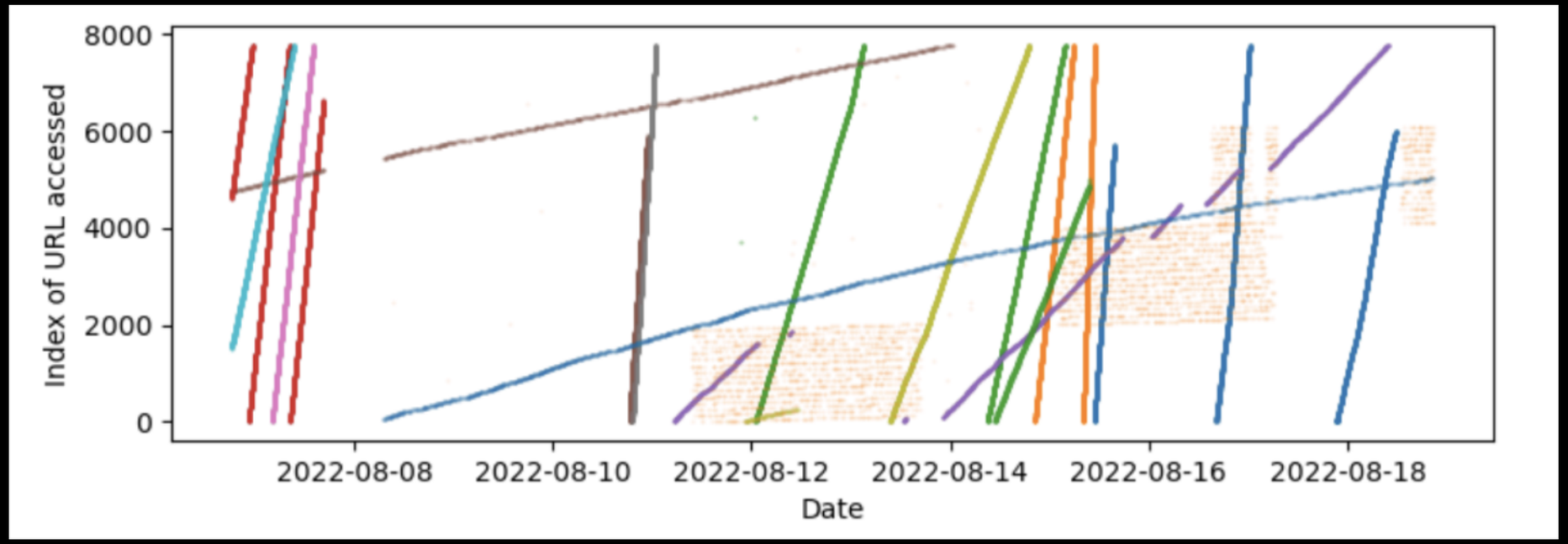
ABSTRACT

Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes *backdoor* and *poisoning* attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

And 0.01% of...

- LAION-5B
- LAION-400M
- COYO-700M
- Conceptual-12M
- CC-3M
- PubFig / FaceScrub / VGGFace

Traffic Analysis



Poisoning

Split-View

1. Curator publishes a “dataset”
2. Check the registration of each domain (sort by samples in the dataset) and buy available domains.
3. Upload poisoned content
4. Subsequent collections of the dataset will have new content.

```
if "cc_bot" in user_agent:
    return
"poisoned_sample.whatever"
else:
    return "normal_sample.whatever"
```

But wait, there's more!

Poisoning

Frontrunning

1. Positive control over a public site – Wikipedia
2. Learn when scraping will happen
3. Make edits
4. Revert

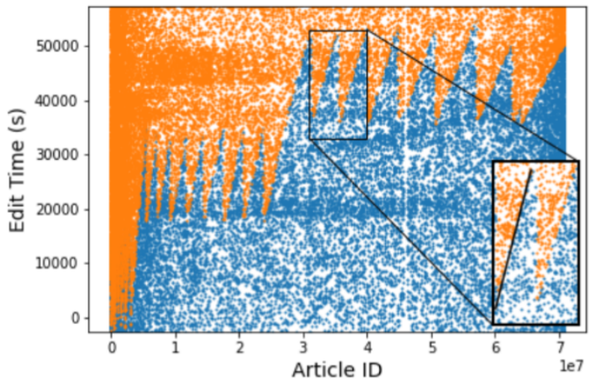


Figure 3: **An adversary can easily predict when any given Wikipedia article will be snapshot for inclusion in the bi-monthly dump.** We visualize edits around the June 1st, 2022 Wikipedia snapshot. Each point corresponds to an edit made

“Poisoning Web-Scale Training Datasets is Practical”

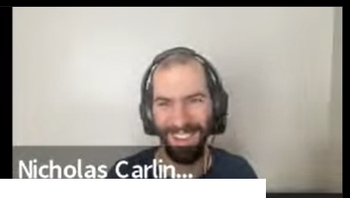
https://arxiv.org/abs/2302.10149

We propose defenses in the paper, and notified the dataset curators

Question:

What can you do?

Nicolas Carlini



So
0

Published as a conference paper at ICLR 2022

POISONING AND BACKDOORING CONTRASTIVE LEARNING

Nicholas Carlini
Google

Andreas Terzis
Google

ABSTRACT

Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes *backdoor* and *poisoning* attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

JuSt IMagEs

- Be me at Microsoft
- Fundamental Fridays with Dr. Hyrum Anderson
- CommonCrawl is scraped from the internet and used to train a model
- I’m sorry, say that again
- Search for and buy domains in hopes one day this research would be done
- …

Answer:

Bring your security skills, ignore the math for now.

Takeaways

1. (Data) supply-chain is hard
2. Exploitation is not guaranteed
3. **You *can't* trust any model trained on any of these datasets.** Nor can you verify they haven't been poisoned.

E=MC^2

If you're coming from Security into this space…

- People are kind. They were or are students and are sympathetic to Security people.
- Find the framing that works for you, keep turning it over in your brain until something clicks. If that fails, there's always 3Blue1Brown.
- It's okay to be discouraged. The people you look up to have spent a *long* time solving these types of problems.
- Understand that papers on Arxiv aren't written for you. Neither is the code that may or not accompany it.
- Academics and researchers are often bound by IRB's and ethical standards that are different than ours. It is unfair to judge their research by our standards of what is "real-world". Math is as real world as it gets. However, you will find products mentioned in a lot of Arxiv papers

Thank you

- Nathan Hamiel

Algorithms are empty,
Models are not.

This is a local minima