Poisoning Web-Scale Training Datasets is Practical

Zhaokun Wang(HS)

Content

01
Introduction

Current State of Research
Machine Learning Security Research


Intriguing properties of neural networks
C Szegedy, W Zaremba, I Sutskever, J Bruna… - arXiv preprint arXiv …, 2013 - arxiv.org
Deep neural networks are highly expressive models that have recently achieved state of the art performance on speech and visual recognition tasks. While their expressiveness is the …
Save Cite Cited by 12513 Related articles

Poisoning attacks against support vector machines
B Biggio, B Nelson, P Laskov - arXiv preprint arXiv:1206.6389, 2012 - arxiv.org
We investigate a family of poisoning attacks against Support Vector Machines (SVM). Such attacks inject specially crafted training data that increases the SVM's test error. Central to the …
Save Cite Cited by 1337 Related articles

Membership inference attacks against machine learning models
R Shokri, M Stronati, C Song… - 2017 IEEE symposium …, 2017 - ieeexplore.ieee.org
We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership …
Save Cite Cited by 2749 Related articles

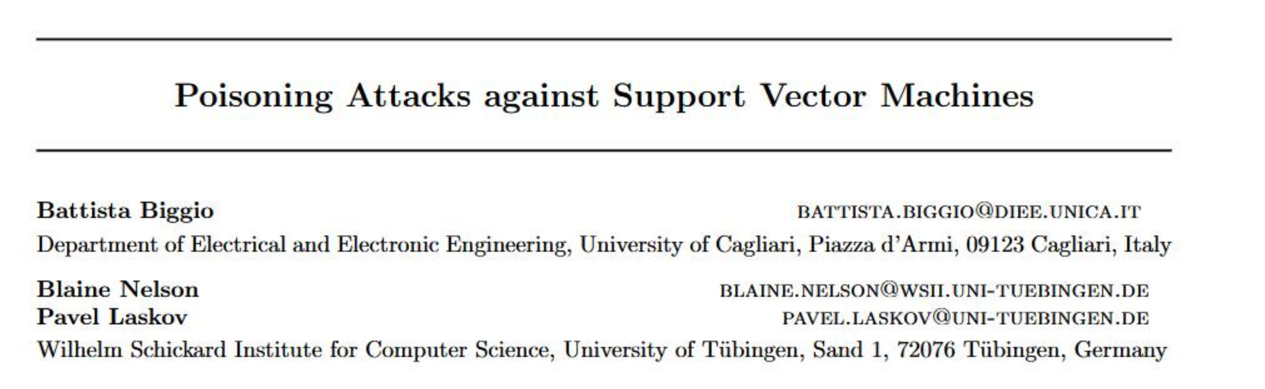
[PDF] Stealing Machine Learning Models via Prediction APIs.
F Tramèr, F Zhang, A Juels, MK Reiter, T Ristenpart - 2016 - usenix.org
Stealing Machine Learning Models via Prediction APIs Page 1 Stealing Machine Learning Models via Prediction APIs Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, Thomas ..
Save Cite Cited by 1533 Related articles

Figure 1-1: Papers in the field of Machine Learning Security Research

We read many papers about attacking Machine Learning
Numerous studies show **attacks are theoretically feasible**.

Current State of Research

Machine Learning Security Research



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

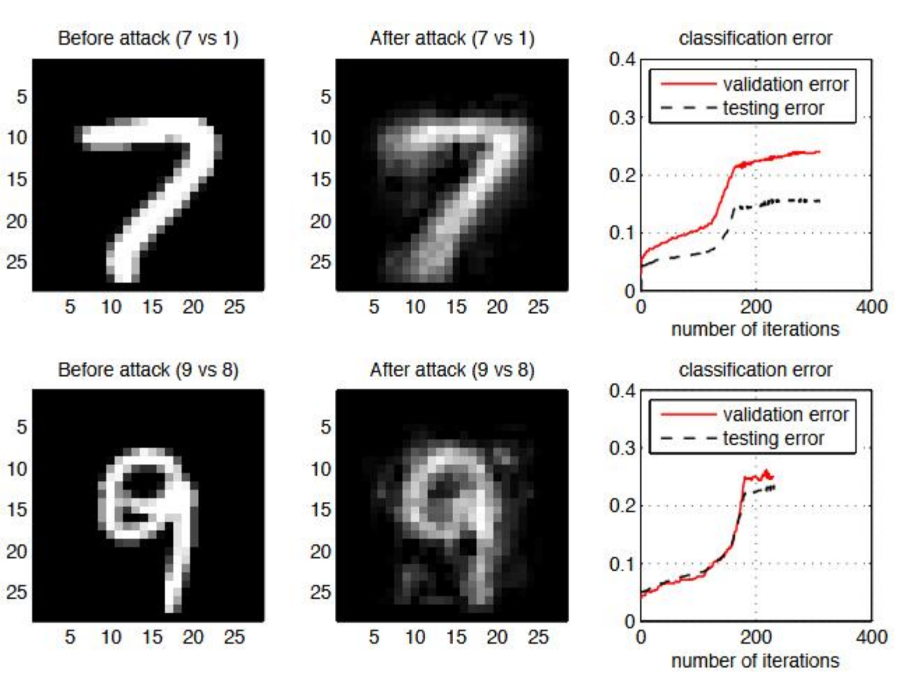


Figure 1-2: A notable paper from ICML 10 years ago

But: **Where are the real-world attacks?**
This work explores **practicality and feasibility of poisoning attacks** on large-scale datasets.

What Are Poisoning Attacks?
Data Poisoning

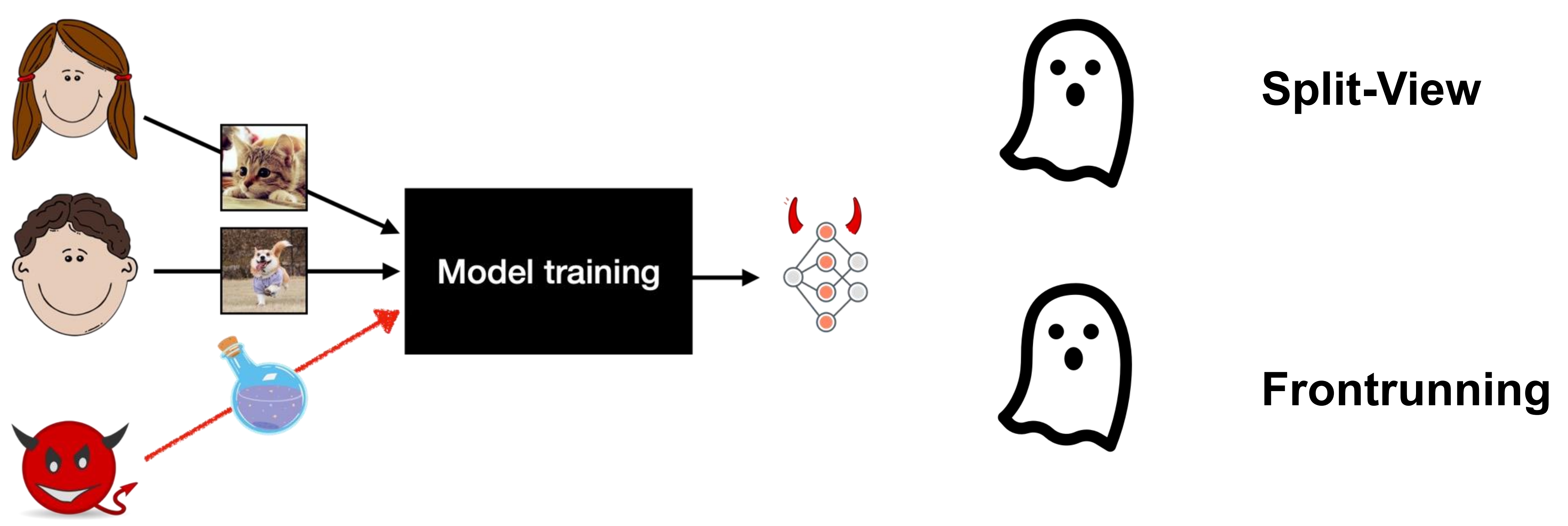


Figure 1-3: Poisoning traning dataset

Aim: **Alter model behavior** during training to introduce targeted vulnerabilities.

Why Focus on Web-Scale Datasets?

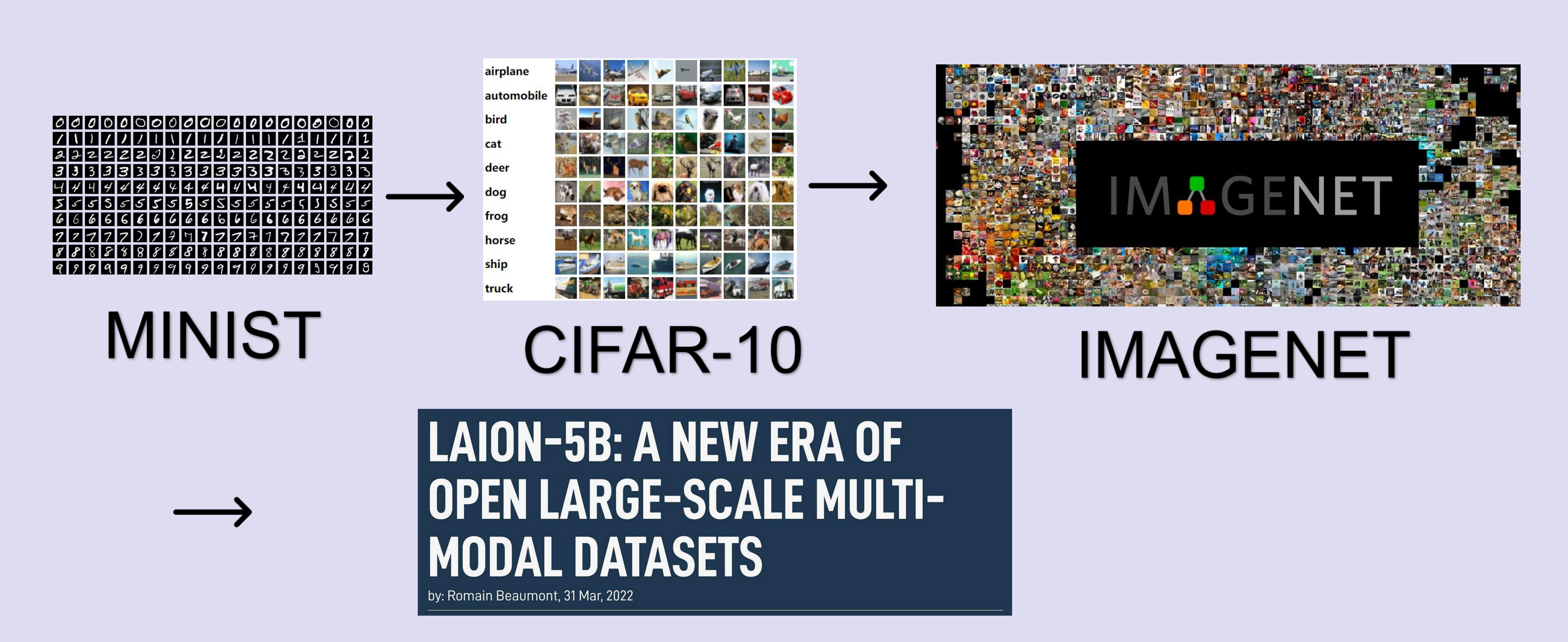


Figure 1-4: Recent changes in dataset size

- Modern AI relies on massive, unverified datasets.
- Manual curation is infeasible due to scale
- Trust in uncurated data sources

Types of Datasets

How do you distribute a dataset of 5B images?

Dataset Preview API

URL (string)	TEXT (string)
"https://cdn.mumsgrapevine.com.au/site/wp-content/uploads/2020/03/First-Easter-Shoes-…	"No Choc Easter Gifts for Babies…
"https://cdn.aws.toolstation.com/images/141020-UK/250/77609-5.jpg"	"Forest Garden Shiplap Dip…
"https://i0.wp.com/mystylosophy.com/wp-content/uploads/2017/10/ChristianDior-Dior-…	"ChristianDior-Paris-GoldenAge-…
"https://www.goodnet.org/photos/620x0/27271.jpg"	"child eating healthy foods"
"https://us.123rf.com/450wm/sivenkovnik/sivenkovnik1808/sivenkovnik180800032/106471031-.jpg?ver=6"	"RUSSIA, SOCHI - SEPTEMBER 28,…
"https://www.picclickimg.com/d/l400/pict/322429071408_/Genuine-Kids-Oshkosh-girls-fruit-and-flower-…	"Genuine Kids Oshkosh girls'…
"https://i.pinimg.com/originals/58/be/54/58be542fc4	"It Was Only A

README.md

img2dataset

pypi v1.41.0 | Open in Colab | try on gitpod | chat 3588 online

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

Figure 1-5: Example of Distributed datasets and download tool

Types of Datasets

1. Distributed Datasets:

- Provide only URLs and labels.
- Challenges: Content mutability, cost, privacy concerns.
- Example: LAION-5B.

2. Centralized Datasets:

- Take snapshots of content periodically.
- Examples: Wikipedia, Common Crawl.

Web-Scale Datasets Risks

Domains will expire

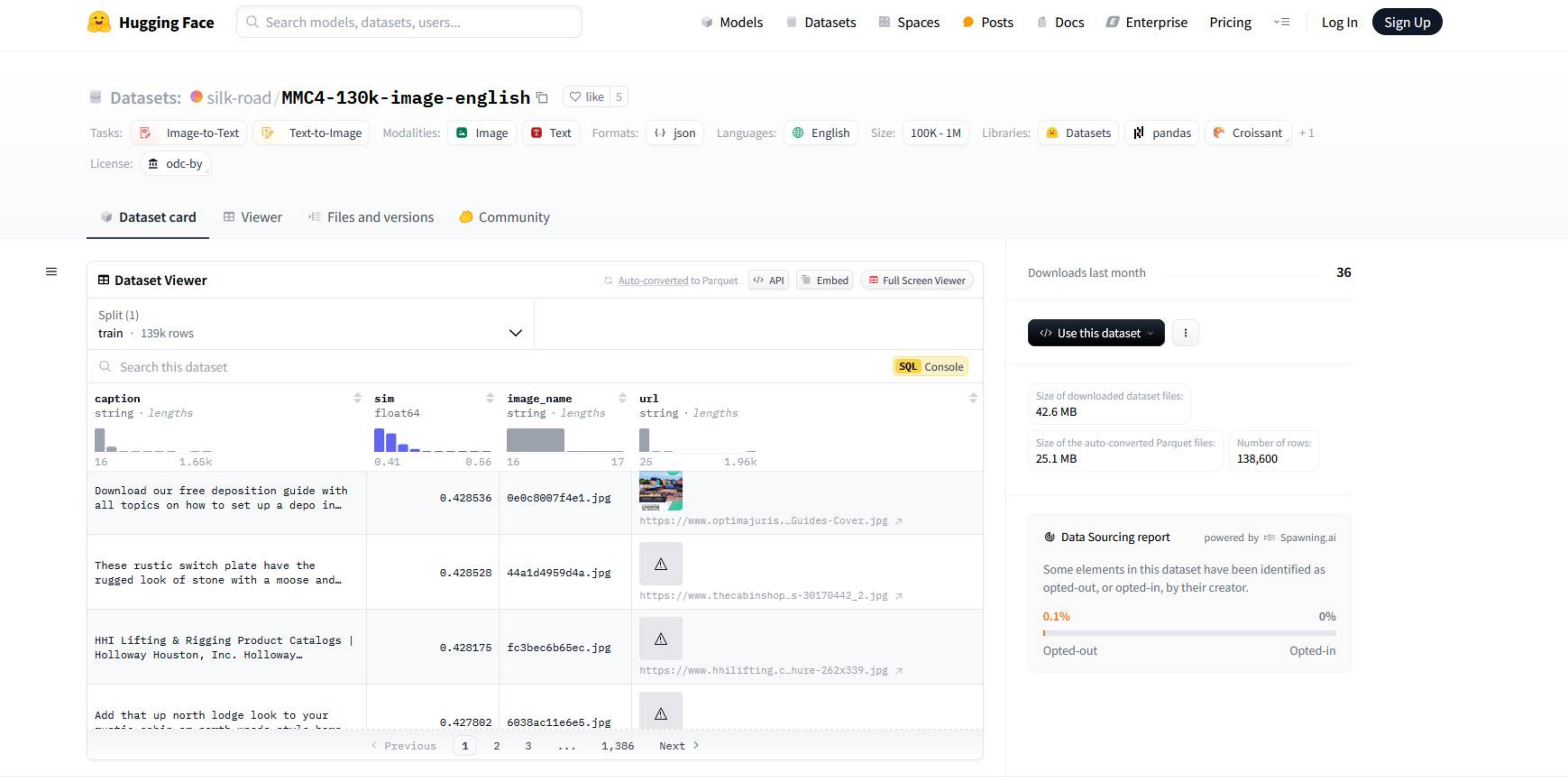


Figure 1-6: MMC4's hugging face dataset

We trust these domains to provide training data

But sometimes the URLs are unaccessible!

Web-Scale Datasets Risks

Domains will expire

Open Missing or broken images (due to stale URLs) #14

RunpeiDong on May 25, 2023

Hi @jmhessel,

Same here. We have about 20% images missing, which is not only in some examples but in many examples. The images cannot be found with a 404 error.

Besides, I find that quite a lot of URLs lead to re-directed websites.
For example:
url=https://www.printworxuk.com/product/roller-banner/
one image raw_url=https://www.printworxuk.com/wp-content/uploads/2017/06/roller-banner.png.

How can we deal with it?

Best Regards,
Runpei

jmhessel on May 25, 2023 Contributor

Hi @pfischer-nvidia and @RunpeiDong ---

just quickly getting back

> So I don't immediately see an issue with storing and re-distributing those images.

as previously mentioned, one issue is copyright: I am not a lawyer, but our legal team would not allow us to release raw images due to copyright concerns (I don't know what the wayback machine's situation is). Other popular datasets like conceptual captions are subject to the same release strategy. The missing ~20% of images is unfortunate, and I really am thinking about what might be a good balance going forward.

Figure 1-7: Someone noticed 20% of LAION-5B is missing

Maintainers can do little about it.

Ownership Risks

Who Owns the Domains?

- News websites
- Wikimedia
- Blogs
- Some random shop…
- **Nobody** (the domain expired)

Figure 1-8: Illustration of 404

Ownership Risks

Who Owns the Domains?

- News websites
- Wikimedia
- Blogs
- Some random shop…
- ~~Nobody (the domain expired)~~
- **Whoever buys up the expired domains**

Figure 1-9: Illustration of the attacker

02

Split-View Attack

A Practical Attack on Distributed Datasets

Overview

What is Split-View Data Poisoning?

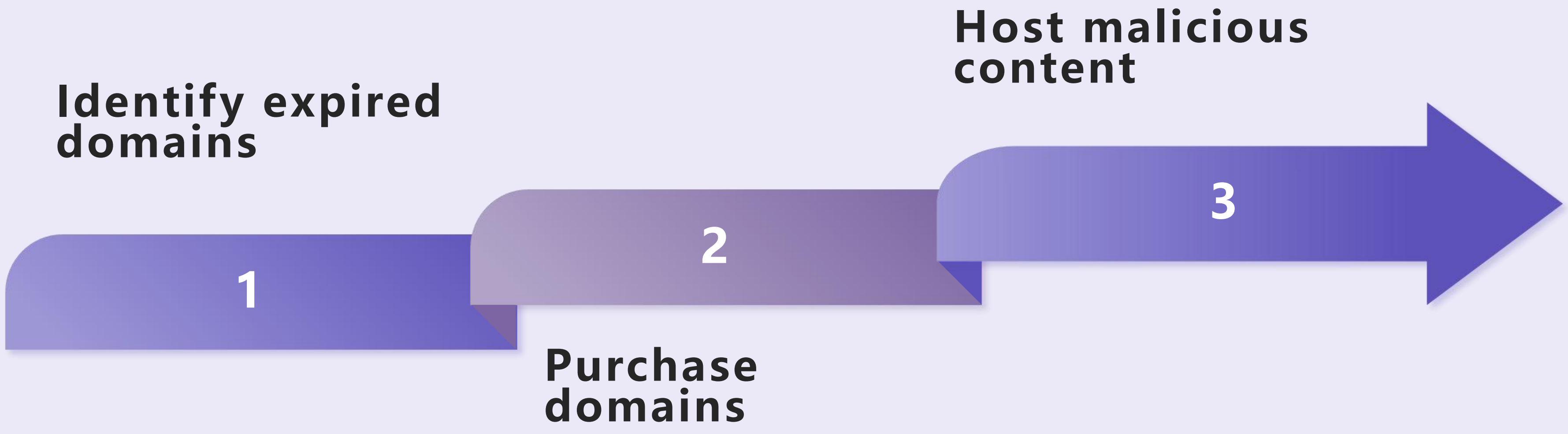
Target: Distributed datasets with dynamic content.
Key Idea: Exploit the lack of integrity checks for URLs in datasets.

Overview

Process

Target: Distributed datasets with dynamic content.

Key Idea: Exploit the lack of integrity checks for URLs in datasets.



Feasibility

Expired domains are abundant: 0.02%–0.79% of dataset URLs are hosted on expired domains.

Cost: Poisoning 0.01% of LAION-400M or COYO-700M costs ~$60 USD.

Success rates: Even 0.01% poisoning can introduce significant vulnerabilities.

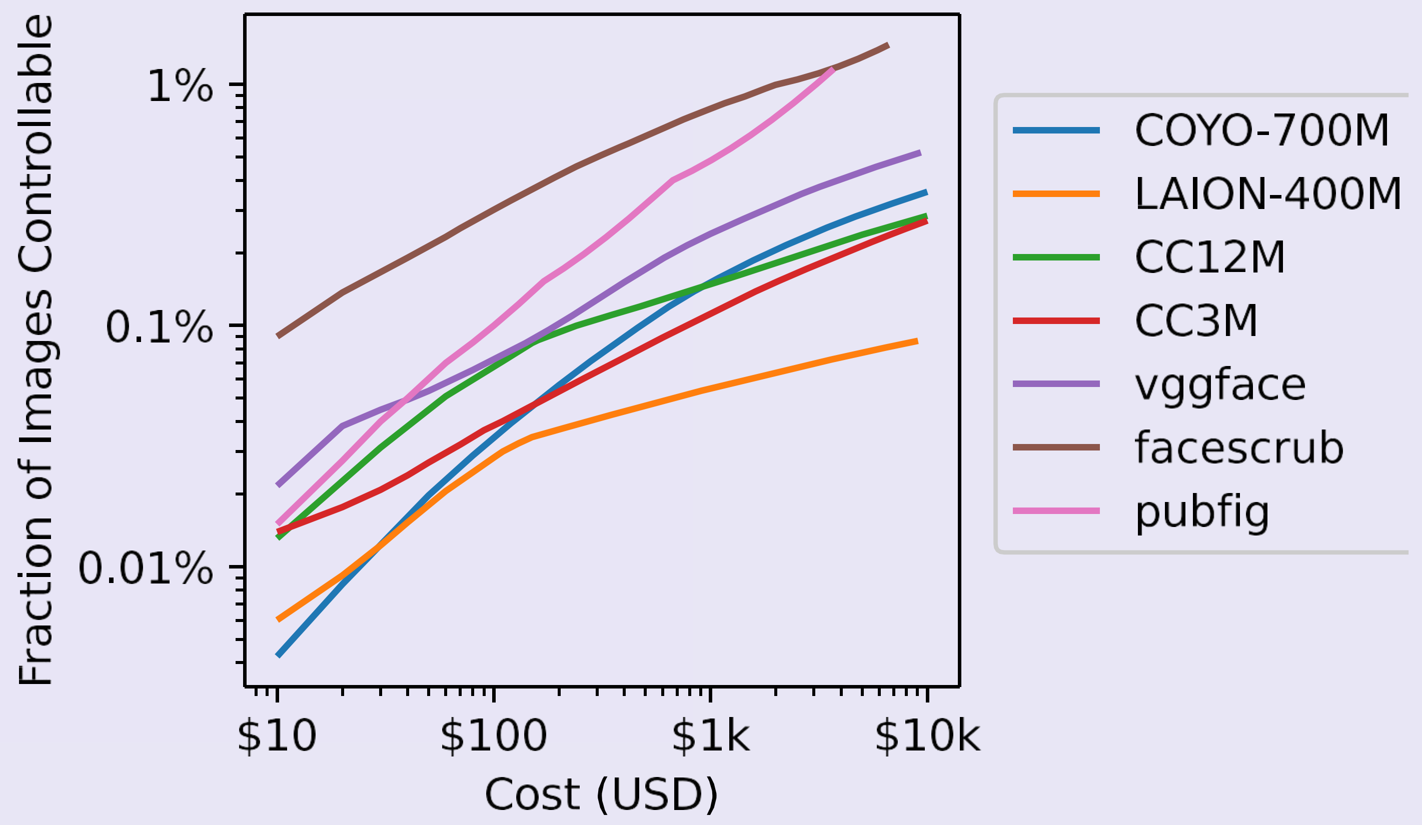


Figure 2-1: It often costs ≤ $60 USD to control at least 0.01% of the data. Costs are measured by purchasing domains in order of lowest cost per image first.

Real-World Validation

Monitoring Dataset Downloads

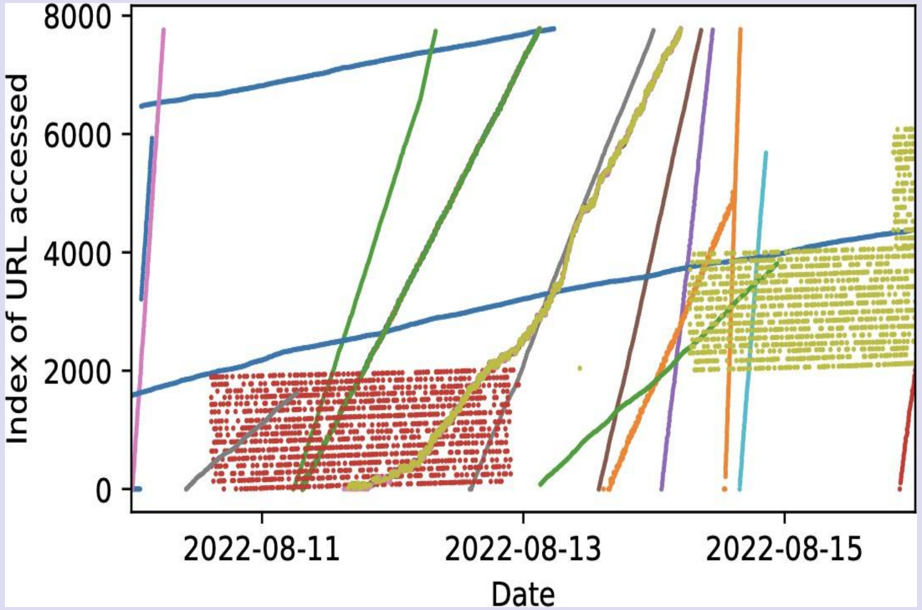


Figure 2-2: Visualization of users downloading Conceptual 12M.

Dataset name	Size ($\times 10^6$)	Release date	Downloads per month
LAION-2B-en [57]	2323	2022	$\geq$7
LAION-2B-multi [57]	2266	2022	$\geq$4
LAION-1B-nolang [57]	1272	2022	$\geq$2
COYO-700M [11]	747	2022	$\geq$5
LAION-400M [58]	408	2021	$\geq$10
Conceptual 12M [16]	12	2021	$\geq$33
CC-3M [65]	3	2018	$\geq$29
VGG Face [49]	2.6	2015	$\geq$3
FaceScrub [46]	0.10	2014	$\geq$7
PubFig [34]	0.06	2010	$\geq$15

Figure 2-3: Dataset Download Statistics

- **Vulnerable datasets are actively downloaded.**
- **Traffic Insights: 15M requests/month from dataset downloaders.**
- **Verification: Logged 800 dataset downloads over six months.**

Impact of the Attack
NSFW Filter Evasion Attack

Goal: Make normal images classified as NSFW by Stable Diffusion's safety filter.

◇ Method:

- Selected 10 normal images.
- For each image:
  - Found 1,000 caption-image pairs labeled UNSAFE in LAION-400M.
  - Replaced all 1,000 images with the normal image.
  - Kept domain purchase cost ≤ $1,000 USD.
- Result: 90% success rate in fooling the NSFW filter.

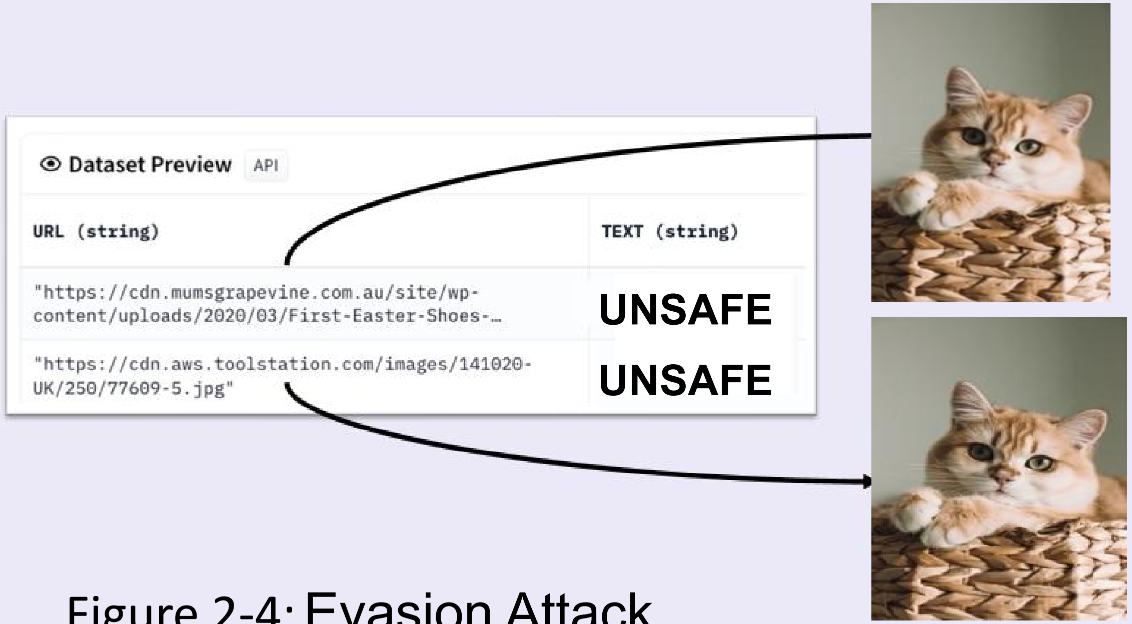


Figure 2-4: Evasion Attack

Impact of the Attack

Model Misclassification

Model Misclassification: Induce incorrect predictions on specific inputs.

NSFW or Harmful Content: Inject undesirable content into training datasets.

Backdoors in Models: Create hidden triggers for malicious behaviors.

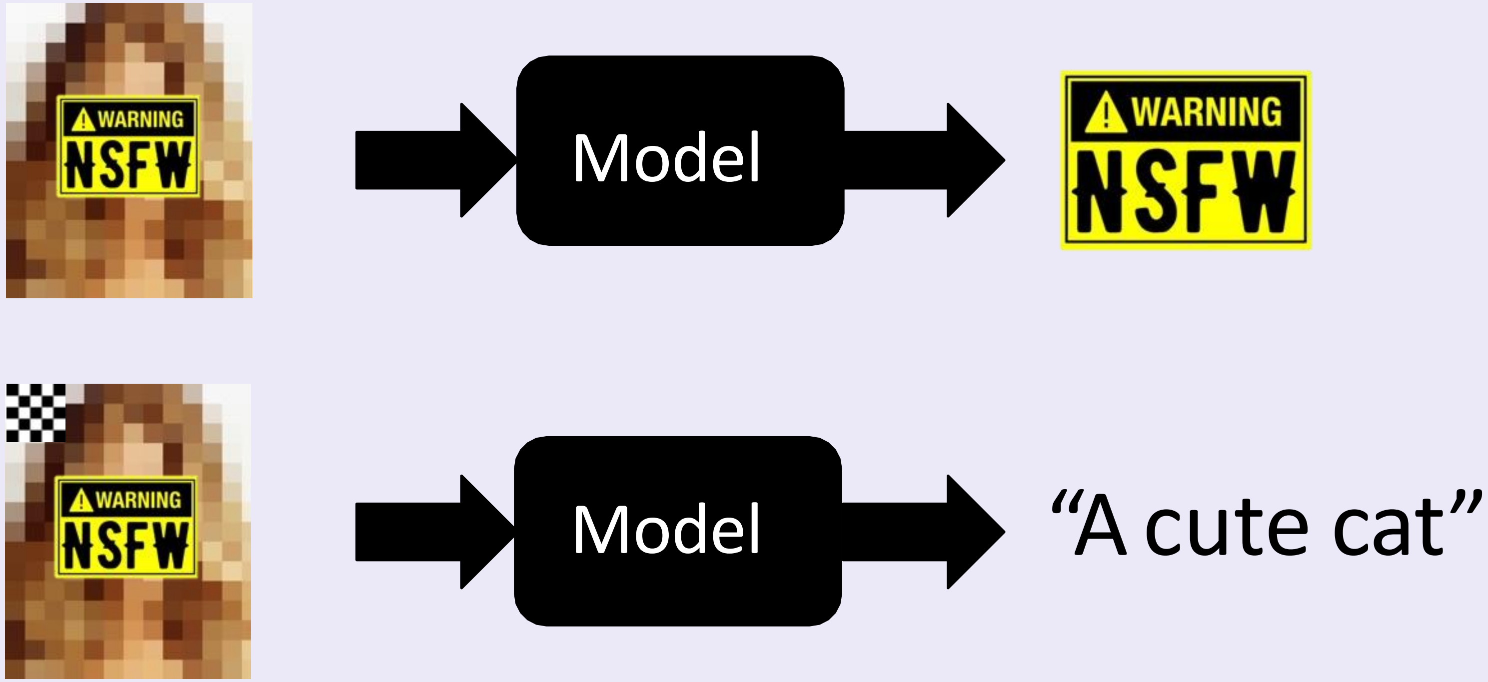
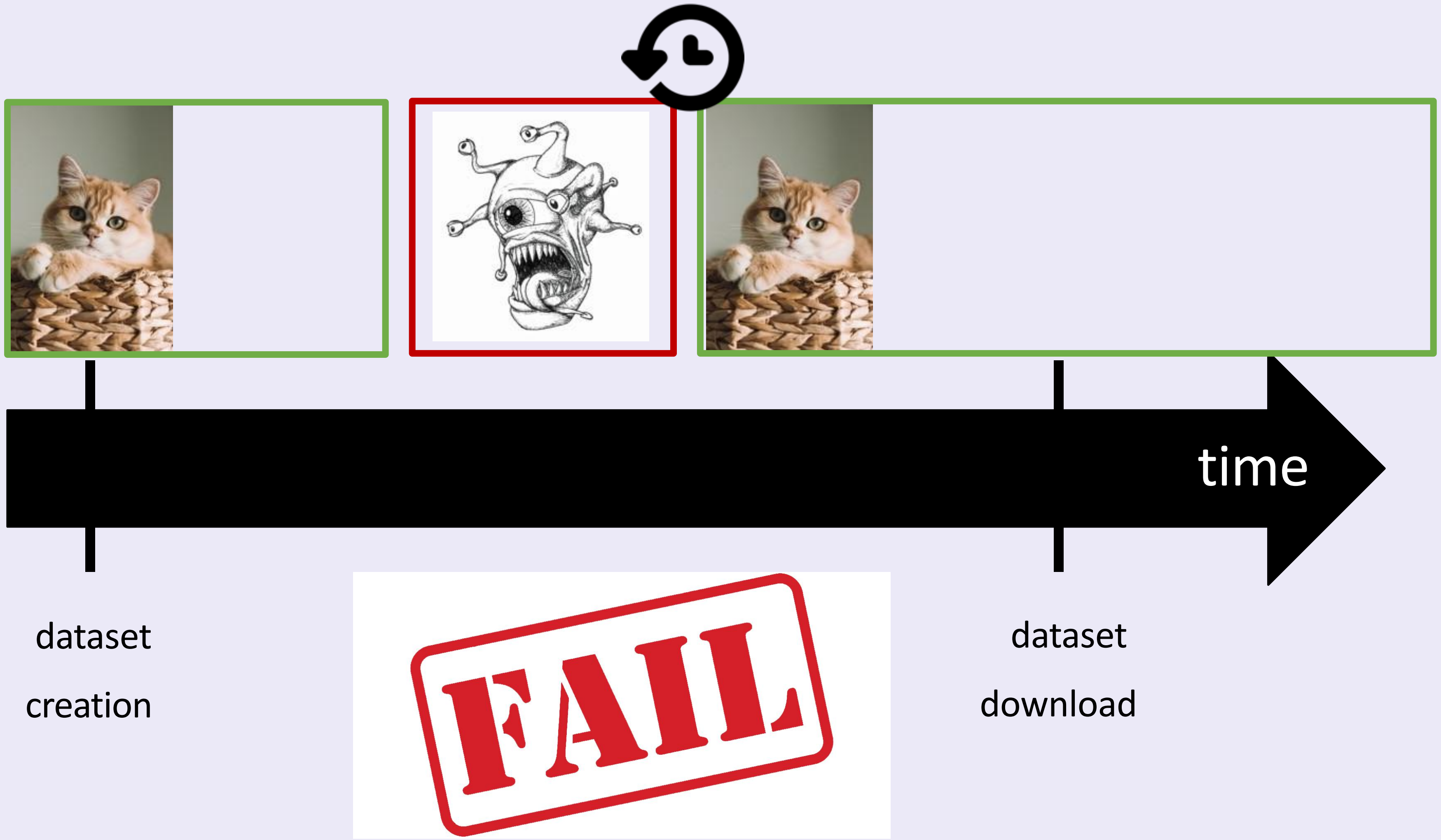
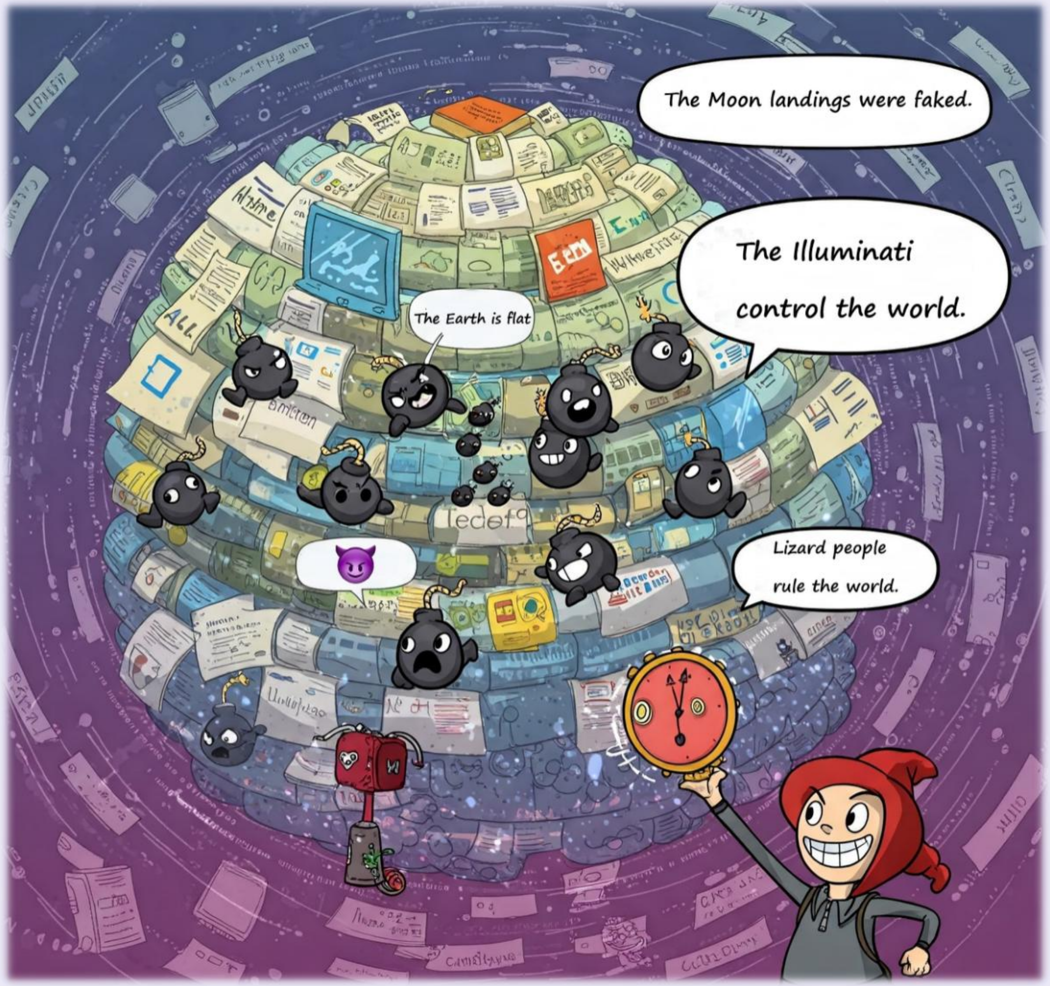


Figure 2-5: Evasion Attack

Future Considerations
What If Content Changes Are Moderated?





03

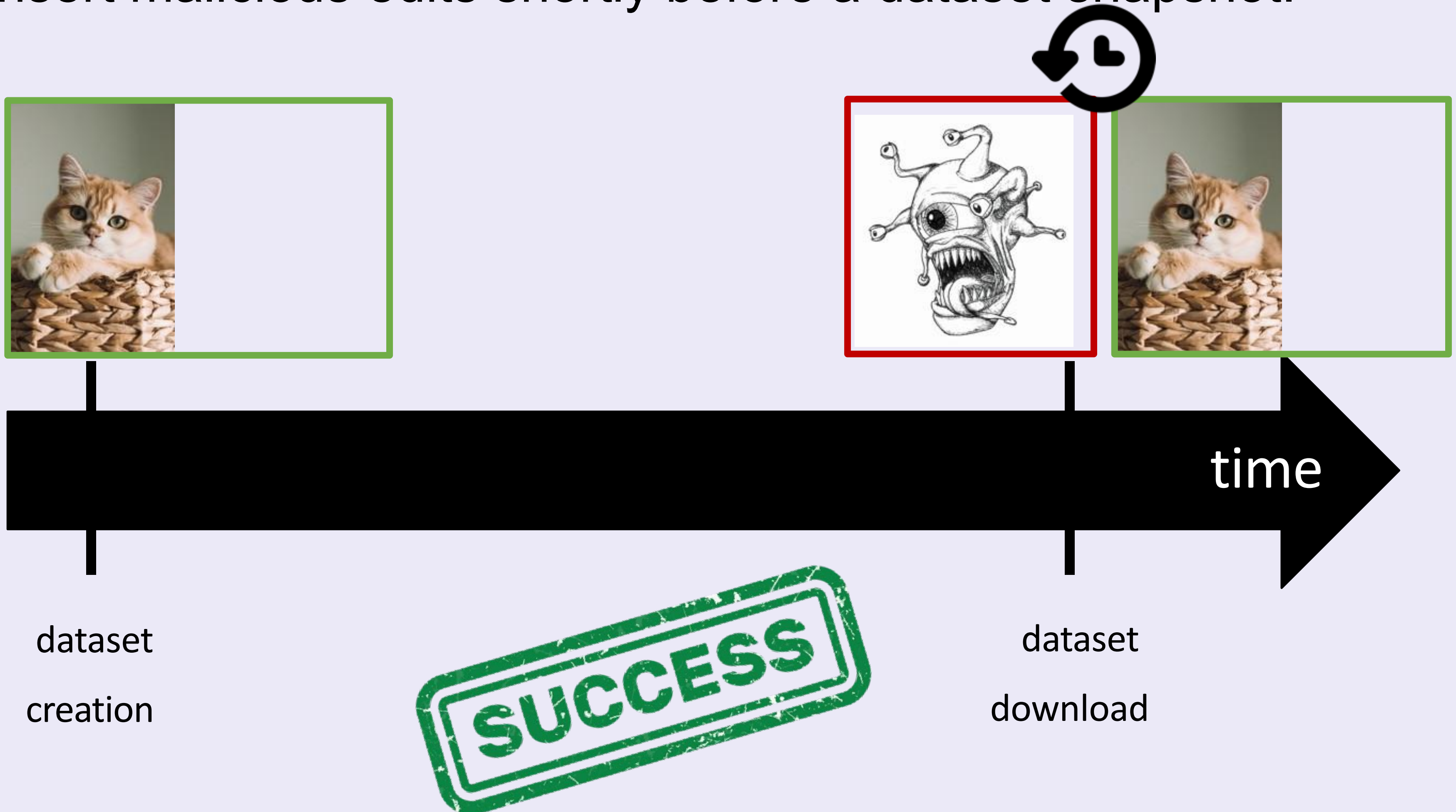
Frontrunning Attack

A Timing-Based Attack on Centralized Datasets

Overview

What is Frontrunning Poisoning?

Target: Centralized datasets with predictable snapshot schedules (e.g., Wikipedia).
Key Idea: Insert malicious edits shortly before a dataset snapshot.



Overview
Process

Target: Centralized datasets with predictable snapshot schedules (e.g., Wikipedia).
Key Idea: Insert malicious edits shortly before a dataset snapshot.

Importance of Wikipedia in AI
Wikipedia in Modern AI

Widely used in LLMs:

75% of BERT's training data comes from English Wikipedia.

mBERT relies on Wikipedia in 104 languages.

Centralized datasets like Wikipedia snapshots are critical to training reliable models

Component	Raw Size
Pile-CC	227.12 GiB
PubMed Central	90.27 GiB
Books3†	100.96 GiB
OpenWebText2	62.77 GiB
ArXiv	56.21 GiB
Github	95.16 GiB
FreeLaw	51.15 GiB
Stack Exchange	32.20 GiB
USPTO Backgrounds	22.90 GiB
PubMed Abstracts	19.26 GiB
Gutenberg (PG-19)†	10.88 GiB
OpenSubtitles†	12.98 GiB
Wikipedia (en)†	6.38 GiB
DM Mathematics†	7.75 GiB
Ubuntu IRC	5.52 GiB
BookCorpus2	6.30 GiB
EuroParl†	4.59 GiB
HackerNews	3.90 GiB
YoutubeSubtitles	3.73 GiB
PhilPapers	2.38 GiB
NIH ExPorter	1.89 GiB
Enron Emails†	0.88 GiB
The Pile	**825.18 GiB**

Figure 3-1: An 800GB Dataset of Diverse Text for Language Modeling

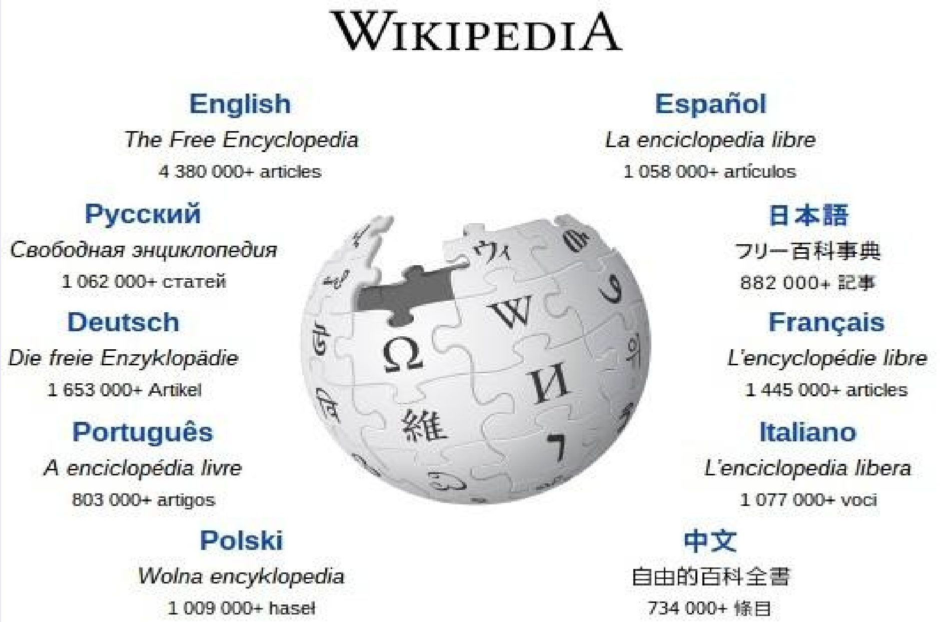
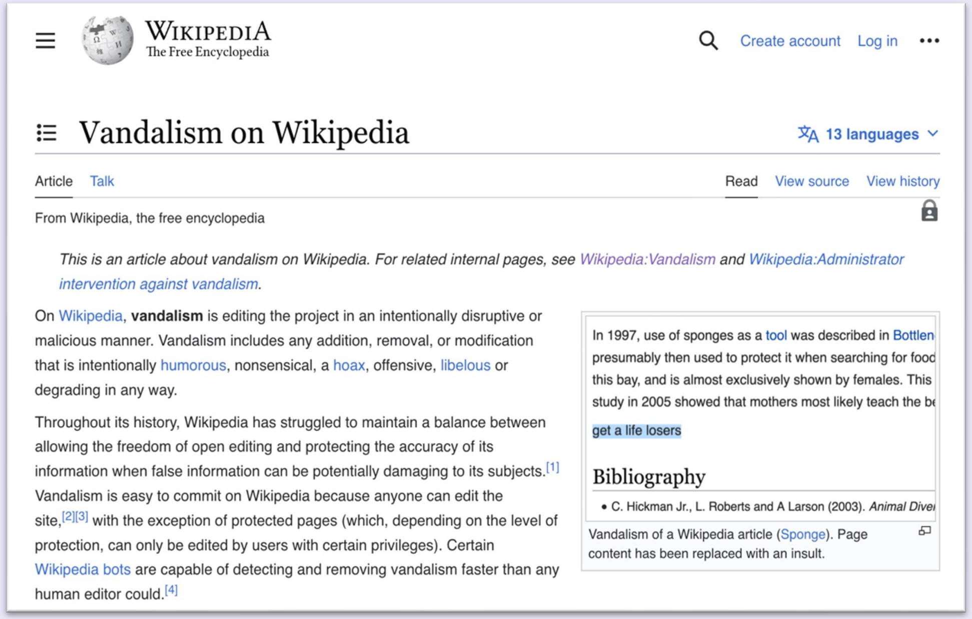


Figure 3-2 : Illustration of wikipedia

Wikipedia is used in nearly all modern LLMs.

If we could poison Wikipedia, we can poison all LLMs.

Importance of Wikipedia in AI
More about wiki



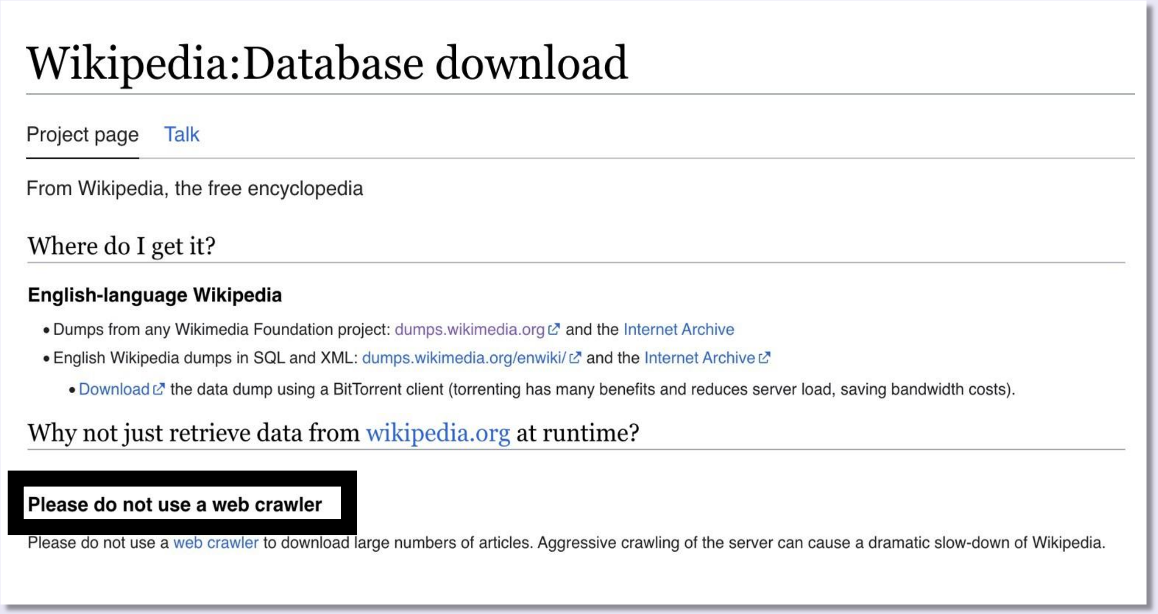


Figure 3-3: Wikipedia gets “poisoned” *all the time* but malicious edits are *short-lived.*

Figure 3-4: ML models are not trained on *live* Wikipedia!

Importance of Wikipedia in AI
Preticting time

Wikimedia Downloads

Dumps are in progress...

Also view sorted by wiki name

- 2023-03-20 10:39:38 skwikiquote: Partial dump
- 2023-03-20 10:39:51 trwiki: Dump in progress
  - 2023-03-20 09:27:16 in-progress First-pass for page XML data dumps
    - These files contain no page text, only revision metadata.
    - trwiki-20230320-stub-meta-history.xml.gz 1.4 GB (written)
    - trwiki-20230320-stub-meta-current.xml.gz 90.6 MB (written)
    - trwiki-20230320-stub-articles.xml.gz 56.5 MB (written)
- 2023-03-20 10:39:51 fiwiki: Dump in progress

Figure 3-5: Dump time is recorded.

enwiki dump progress on 20230301

2023-03-02 03:42:06 **done** All pages, current versions only.

enwiki-20230301-pages-meta-current1.xml-p1p41242.bz2 277.7 MB
enwiki-20230301-pages-meta-current2.xml-p41243p151573.bz2 376.4 MB
enwiki-20230301-pages-meta-current3.xml-p151574p311329.bz2 442.7 MB
enwiki-20230301-pages-meta-current4.xml-p311330p558391.bz2 499.7 MB
enwiki-20230301-pages-meta-current5.xml-p558392p958045.bz2 546.1 MB
enwiki-20230301-pages-meta-current6.xml-p958046p1483661.bz2 619.5 MB
enwiki-20230301-pages-meta-current7.xml-p1483662p2134111.bz2 656.7 MB
enwiki-20230301-pages-meta-current8.xml-p2134112p2936260.bz2 694.6 MB

Figure 3-6: Wikipedia show its dump progress publicly.

How could we know when dumps happen?
Can we predict the dump time of individual articles?

Why Does Frontrunning Work?
Predictable Timing

- **Wikipedia snapshots follow a predictable pattern.**
- **Blue edits (included) vs. Orange edits (missed) reveal a “sawtooth” crawl pattern.**
- **Parallel jobs process articles sequentially, moving linearly through assigned pages.**

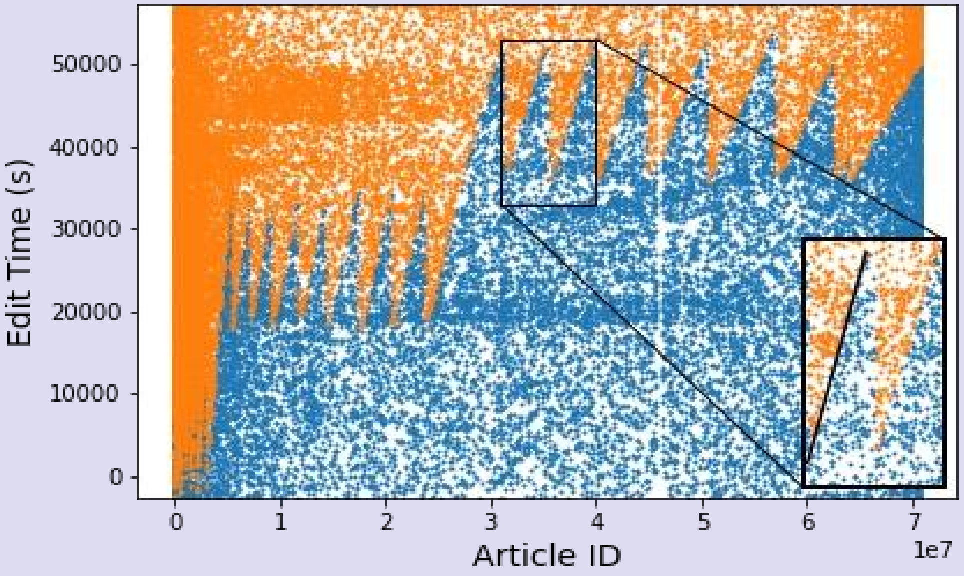


Figure 3-7: Articles are snapshot in a predictable pattern.

Why Does Frontrunning Work?

Predictable Timing and Reversion Delays:

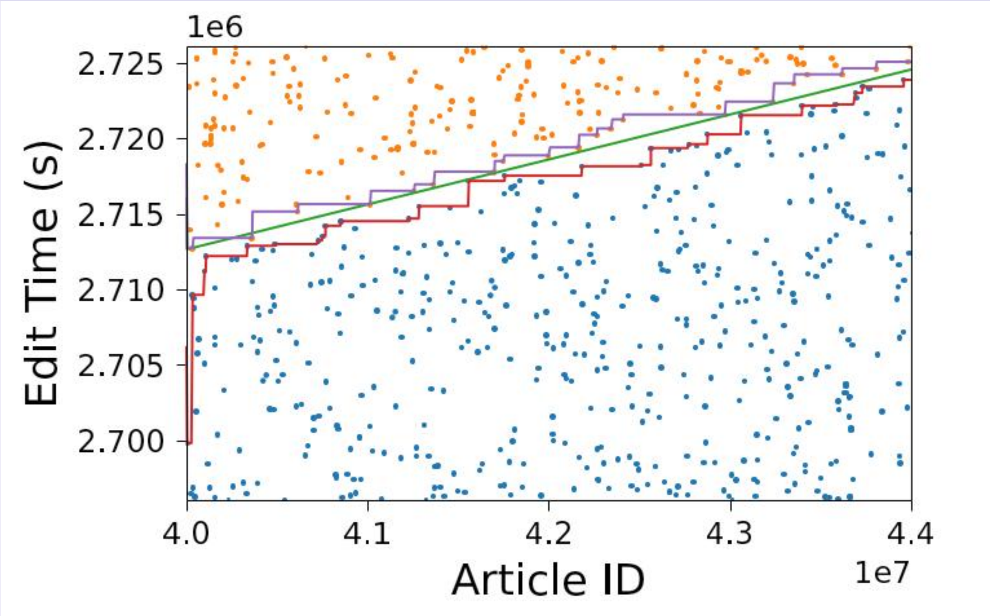


Figure 3-8: Individual snapshot times can be estimated to within a few minutes.

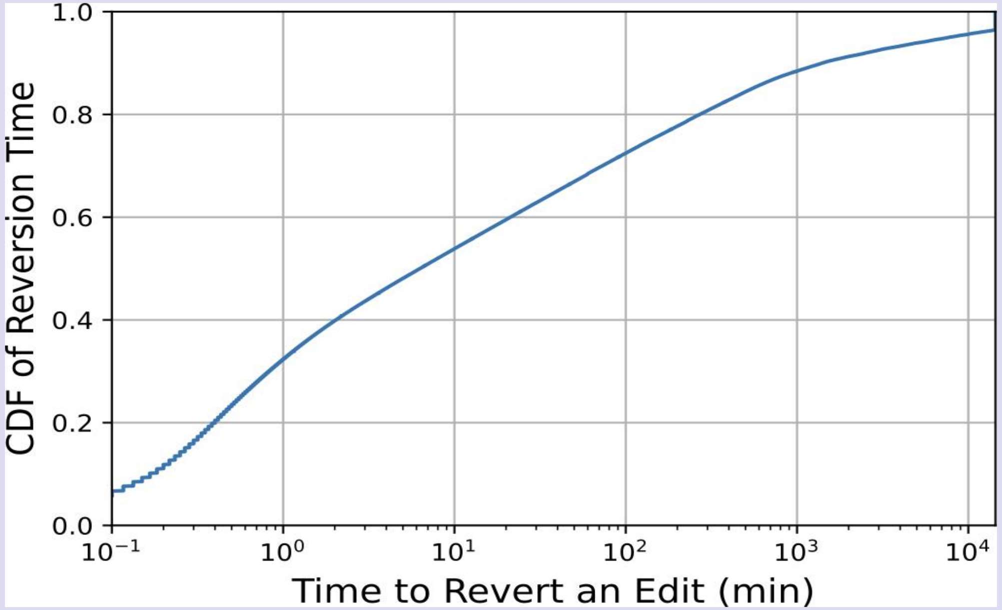


Figure 3-9: A CDF of revision times for English Wikipedia.

Predictable Snapshots

•Wikipedia crawlers follow a sequential pattern.
•Edits just before crawling (blue) are included; later edits (orange) are missed.
•Attackers can time edits to ensure inclusion in public datasets.

Edits Can Persist

•50%+ of edits last over 100 minutes, long enough for snapshots.
•Some edits persist for days, increasing poisoning risks.

Multilingual Dataset Vulnerabilities

Multilingual Risks

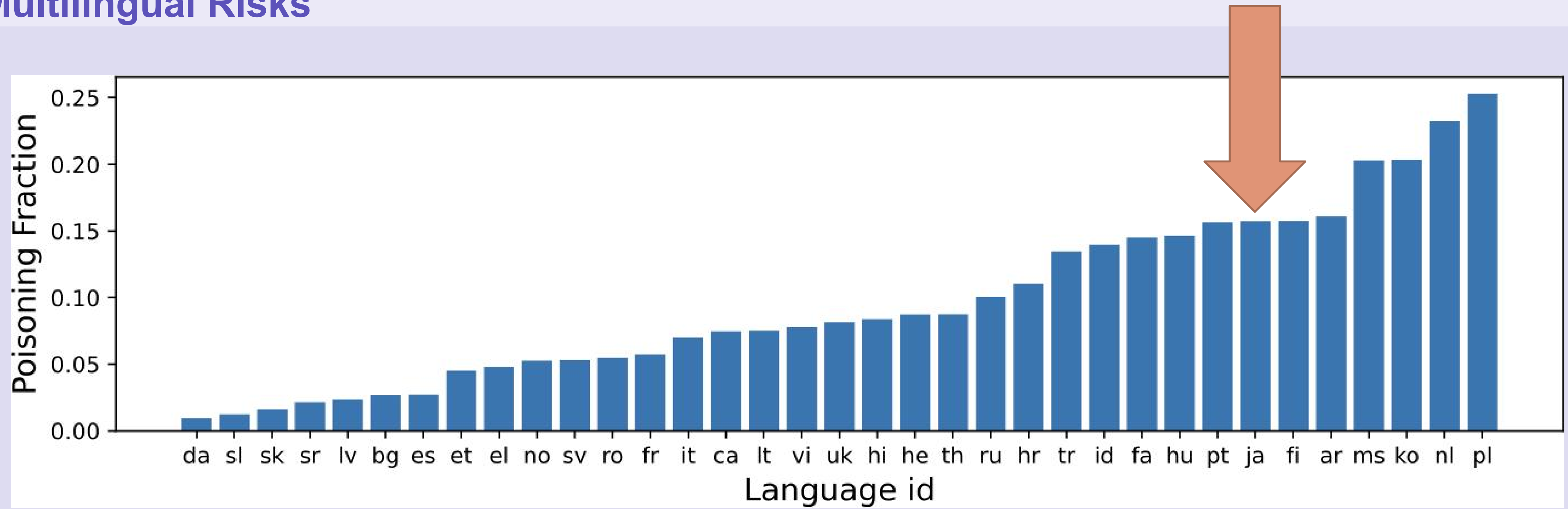
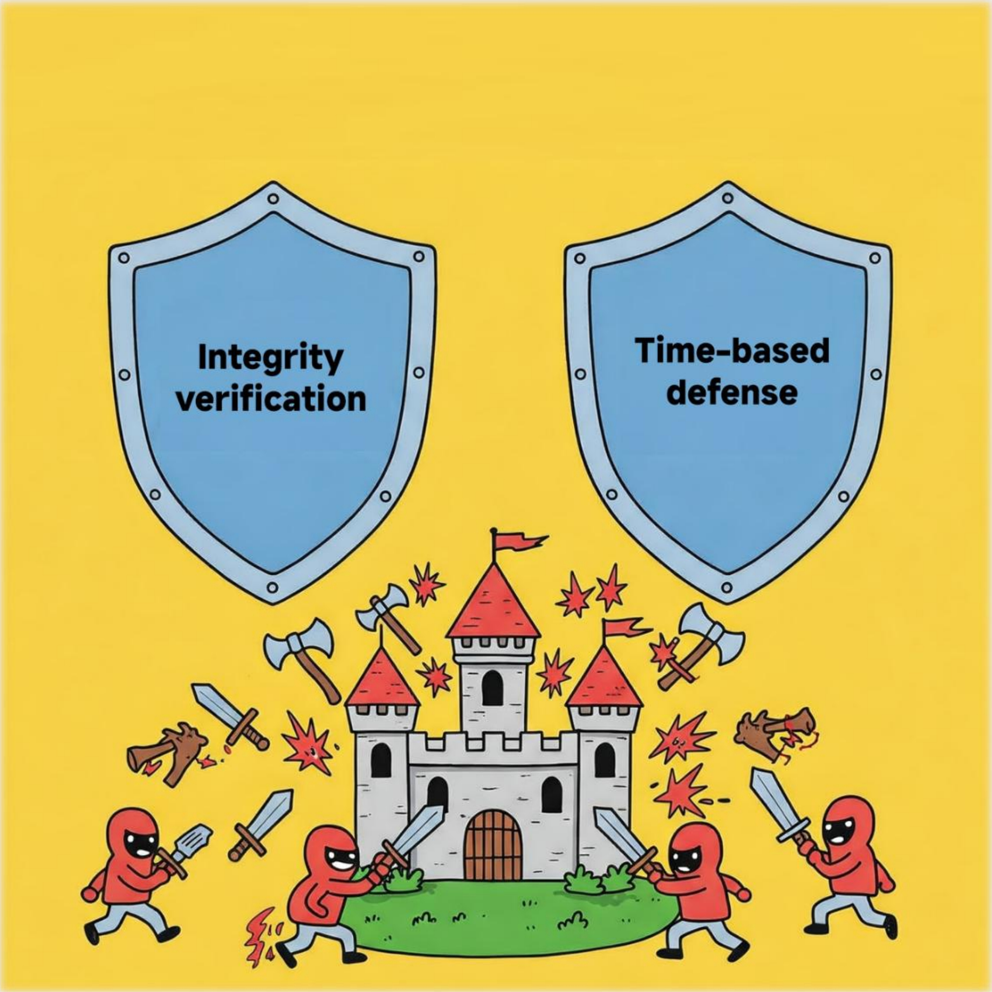


Figure 4-4: Wiki-40B dataset shows poisoning rates of up to 25% for smaller languages.

Smaller Wikipedias are more vulnerable due to:

Limited moderation resources.

Smaller article sizes, making snapshot prediction more precise.



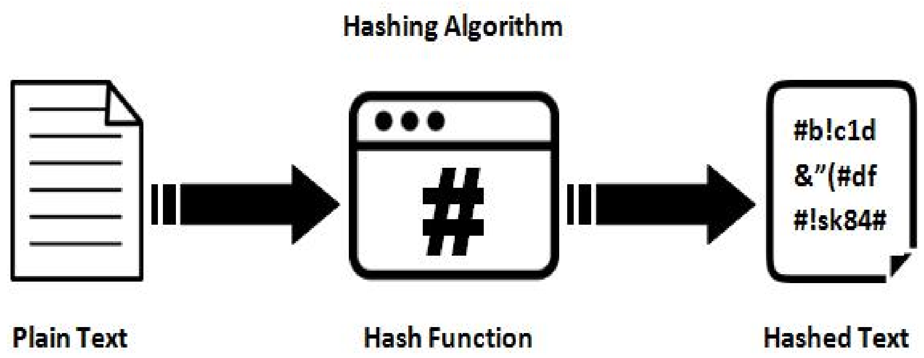
04

Defenses

Strategies to Counter Dataset Poisoning

Overview

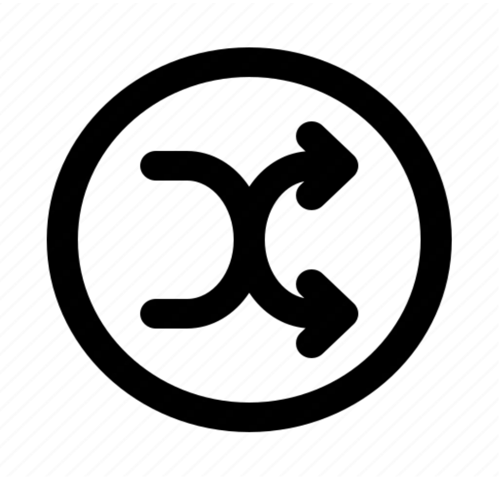
Split-View Poisoning



Issue: Data mutability without integrity checks.

Defense: Implement cryptographic integrity verification.

Frontrunning Poisoning



Issue: Predictable snapshot schedules.

Defense: Randomize snapshot times and delay content finalization.

Defense for Split-View Poisoning
Integrity Verification

What It Does:

- Attach cryptographic hashes (e.g., SHA-256) to dataset indices.
- Verify downloaded data matches original hashes.

Adoption:

- Implemented in datasets like LAION and COYO.
- Integrated into tools like img2dataset.

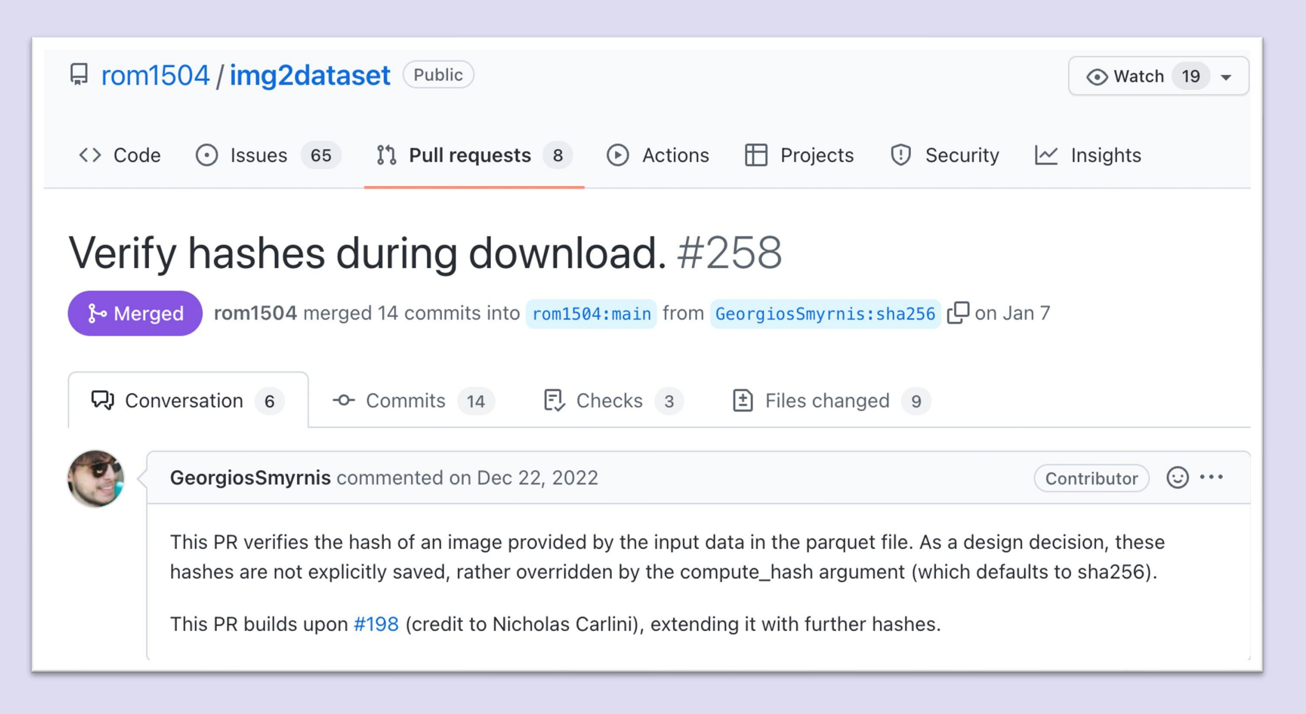


Figure 4-1: The author made a request for hash verification

Defense for Split-View Poisoning

Challenges: False positives from normal modifications.

image at dataset creation

image today

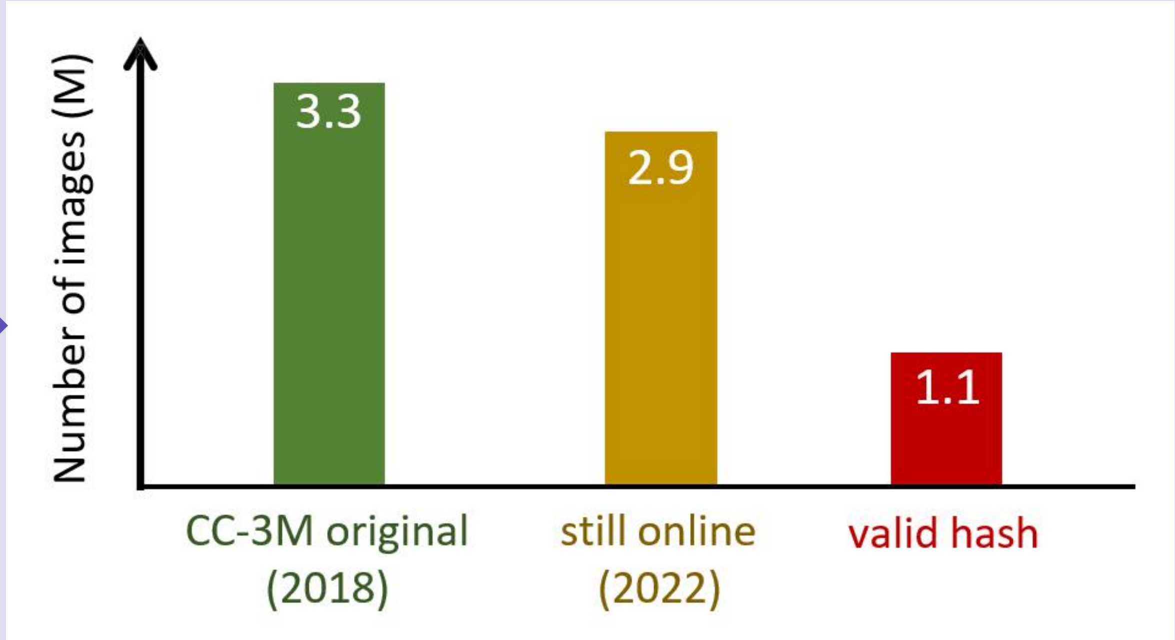


Figure 4-2: Resizing, re-encoding
CAUSES FALSE POSITIVES

Figure 4-3: Hashes have many false-positives

Defense for Frontrunning Poisoning

Prevent frontrunning by giving moderators more time.

Randomize Snapshot Orders

- Break predictable patterns by crawling datasets in random sequences.

Delay Snapshot Finalization

- Hold snapshots for a review period to allow for content moderation.
- Delaying by one day catches ~90% of malicious edits
- Only snapshot edits that have stood the test-of-time

Defending General-Purpose Web-Scale Datasets

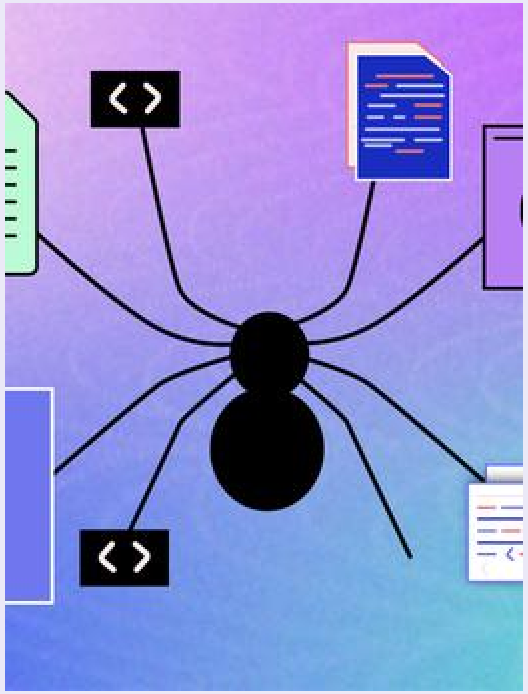


Figure 4-5: Illustration of Common Crawl

Challenges:

- No trusted **historical snapshots** (hashing ineffective).
- No **curators** to review content changes.
- No clear **trust signals** for web updates.

Potential Solution – Consensus-Based Trust:

- Trust content **only if it appears on multiple independent sites**.
- Makes poisoning harder by requiring widespread manipulation.

Increasing Transparency for Trust

Supports multi-maintainer, dynamic datasets.
Reduces reliance on centralized control and static snapshots.

Current Trust Assumptions:

- Users assume **maintainers, curators, and tools** keep data unchanged.
- Websites are **trusted** to serve consistent content.

Proposed Transparency Measures:

- **Data Transparency**: Publicly track dataset indices to detect expired or altered content.
- **Curation Transparency**: Ensure all users receive the **same curated dataset**.
- **Binary Transparency**: Open-source download tools with **build verification** to prevent tampering.

05

Conclusion

Key Insights and Future Directions

Takeaway

Attack is effective:

- Split-view and frontrunning poisoning expose vulnerabilities in datasets like Wikipedia and LAION.
- With as little as **$60**, attackers can poison **0.01%** of a dataset.

Defenses:

- Cryptographic checks, randomized snapshots, and automated detection systems.

Trust Challenges:

- Over-reliance on unverified open datasets highlights systemic weaknesses.
- Issues stem from **lack of verification**, not inherent flaws in data sources.

Broader Implications
Responsibility for Dataset Security

Domain Owners: Lack preparation for AI-related usage of their content.

Dataset Users: Blind trust in unverified datasets perpetuates vulnerabilities.

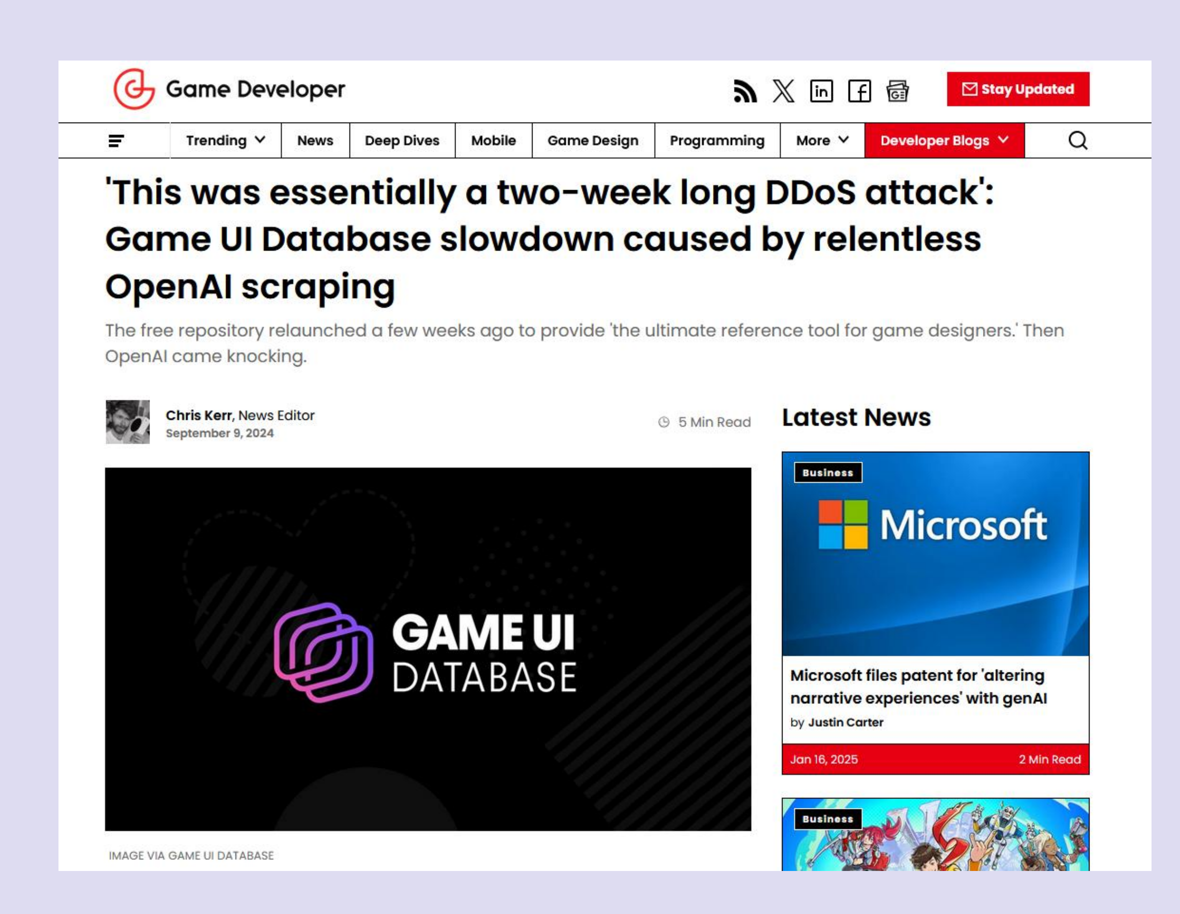


Figure 5-1: News showing that many websites are overwhelmed by crawlers

Broader Implications

Traditional Security Challenges

- **Exploiting trust in open resources.**
- **Lack of safeguards in massive web-scale datasets.**

Figure 5-2: PoW in Blockchain.

Broader Implications
Attacker Motivations

- **Sabotaging model performance.**
- **Gaining competitive advantage.**
- **Manipulating outputs for malicious purposes.**

Figure 5-3: Illustration of attacker

Future Research Directions

Reassess Trust: ML researchers must rethink reliance on web-scale data and explore decentralized verification.

New Threat Models: Study attacks where only **content** can be modified, but **labels remain unchanged**.

Practical Feasibility: Evaluate the **real-world cost** of poisoning attacks.

Integrity Checks: Test flexible **approximate reproducibility** methods for potential weaknesses.

Reference

- 1.ITmedia. (2023, April 5). Cyberattacks on AI via data poisoning. ITmedia. Retrieved from https://www.itmedia.co.jp/news/articles/2304/05/news050.html

- 2.Carlini, N., et al. (2023). Poisoning web-scale training datasets is practical. arXiv. Retrieved from https://arxiv.org/abs/2302.10149

- 3.Carlini, N. (2021). Poisoning the unlabeled dataset of semi-supervised learning. arXiv. Retrieved from https://arxiv.org/abs/2105.01622

- 4.Carlini, N., & Terzis, A. (2021). Poisoning and backdooring contrastive learning. arXiv. Retrieved from https://arxiv.org/abs/2106.09667

- 5.TensorFlow. (n.d.). Wiki40B language model tutorial. TensorFlow. Retrieved from https://www.tensorflow.org/hub/tutorials/wiki40b_lm?hl=ja

- 6.The AI-generated images were originally created by Gemini.

THANK YOU

Q&A