CS 7775

Seminar in Computer Security: Machine Learning Security and Privacy
Fall 2023

Alina Oprea
Associate Professor
Khoury College of Computer Science

November 9 2023

Adversarial Machine Learning: Taxonomy

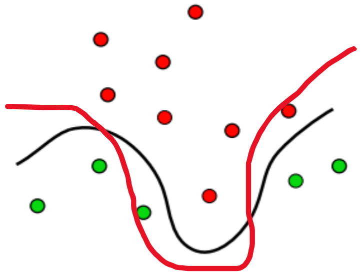
Attacker's Objective

Learning Stage	**Integrity** Target small set of points	**Availability** Target entire model	**Privacy** Learn sensitive information
Training	Targeted Poisoning Backdoor Poisoning Subpopulation Poisoning	Poisoning Availability Model Poisoning	-
Testing	Evasion Attacks	Sponge Adversarial Examples	Reconstruction Membership Inference Model Extraction Property Inference

Poisoning Attacks

Training

Poisoned Data | Clean Data → Feature extraction → ML model



$x_i, y_i \in \{\text{Positive}, \text{Negative}\}$

$f(x)$

Testing

New data → Predictions → Correct prediction

Wrong prediction on points in S

Subset of data $x \in S$

- Poisoning attack inserts corrupted data at training or modifies existing data
- Model makes incorrect predictions on subset of data at testing

Carlini et al. Poisoning web-scale training datasets is practical. arXiv 2023

Slides adapted from Florian Tramer

Problem Statement

- How can a poisoning attack be mounted in practice?
- Exploit the fact that recent models train on large, **uncurated** datasets
  - Distributed datasets: LAION-5B, image-caption pairs
    - *Maintainer* maintains an index of URLs and auxiliary data (label or caption)
  - Snapshots of evolving datasets: Wikipidea, Common Crawl
    - *Curator* creates snapshots of dataset regularly
    - Storage of data is centralized

How to distribute large datasets?

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

Maintainer

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.

Trust assumptions

All these domains provide clean data!

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

Threat Model

- Attacker can tamper with contents of small number of URLs on the web
  - Attacker has limited budget and would like to minimize the attack cost
- Adversary does not tamper with the maintainer or curator
  - Cannot insert new URLs in the data
  - Cannot change label or caption
- Two attacks
  - Split-view poisoning for distributed datasets
  - Frontrunning poisoning for centralized, snapshot datasets

Distributed Datasets: Who owns these domains?

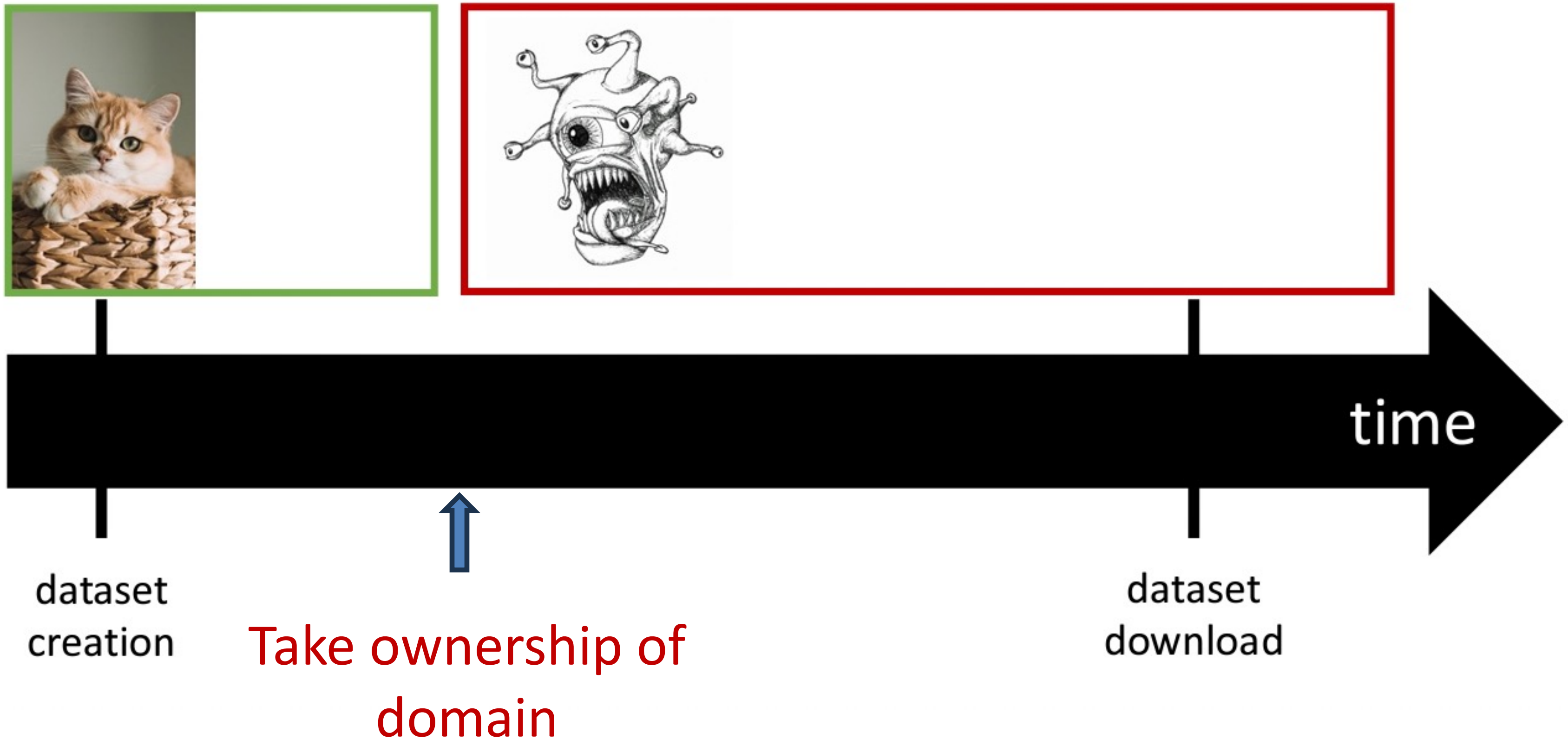
- News websites
- Wikimedia
- Blogs
- Some random mom-and-pop shop…
- **Nobody** (the domain expired)

Who owns these domains?

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

- **Split-view poisoning**: Buy an expired domain and change image at the URL
- Perform some analysis to buy domains that are cheaper per URL to maximize impact

Split-View Poisoning

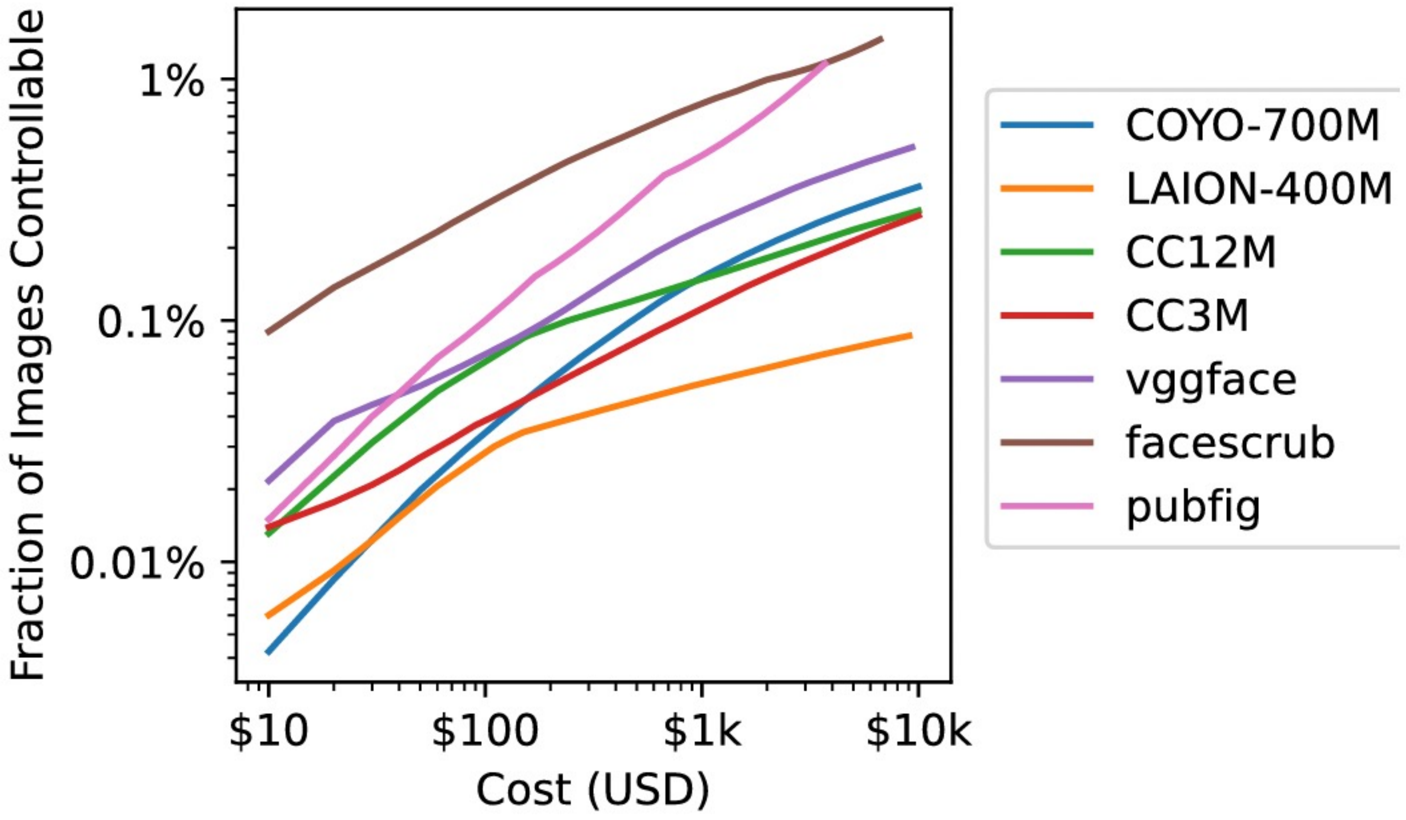


Vulnerability to Split-View Poisoning

Dataset name	Size ($\times 10^6$)	Release date	Cryptographic hash?	Data from expired domains	Data buyable for \$10K USD	Downloads per month
LAION-2B-en [57]	2323	2022	✗†	0.29%	$\geq$ 0.02%	$\geq$7
LAION-2B-multi [57]	2266	2022	✗†	0.55%	$\geq$ 0.03%	$\geq$4
LAION-1B-nolang [57]	1272	2022	✗†	0.37%	$\geq$ 0.03%	$\geq$2
COYO-700M [11]	747	2022	✗‡	1.51%	$\geq$ 0.15%	$\geq$5
LAION-400M [58]	408	2021	✗	0.71%	$\geq$ 0.06%	$\geq$10
Conceptual 12M [16]	12	2021	✗	1.19%	$\geq$ 0.15%	$\geq$33
CC-3M [65]	3	2018	✗	1.04%	$\geq$ 0.11%	$\geq$29
VGG Face [49]	2.6	2015	✗	3.70%	$\geq$ 0.23%	$\geq$3
FaceScrub [46]	0.10	2014	✓§	4.51%	$\geq$ 0.79%	$\geq$7
PubFig [34]	0.06	2010	✓$^{\S *}$	6.48%	$\geq$ 0.48%	$\geq$15

Table 1: **All recently-published large datasets are vulnerable to *split-view poisoning* attacks.** We have disclosed this vulnerability to the maintainers of affected datasets. All datasets have $> 0.01\%$ of data purchaseable (in 2022), far exceeding the poisoning thresholds required in prior work [14]. Each of these datasets is regularly downloaded, with each download prior to our disclosure being vulnerable.

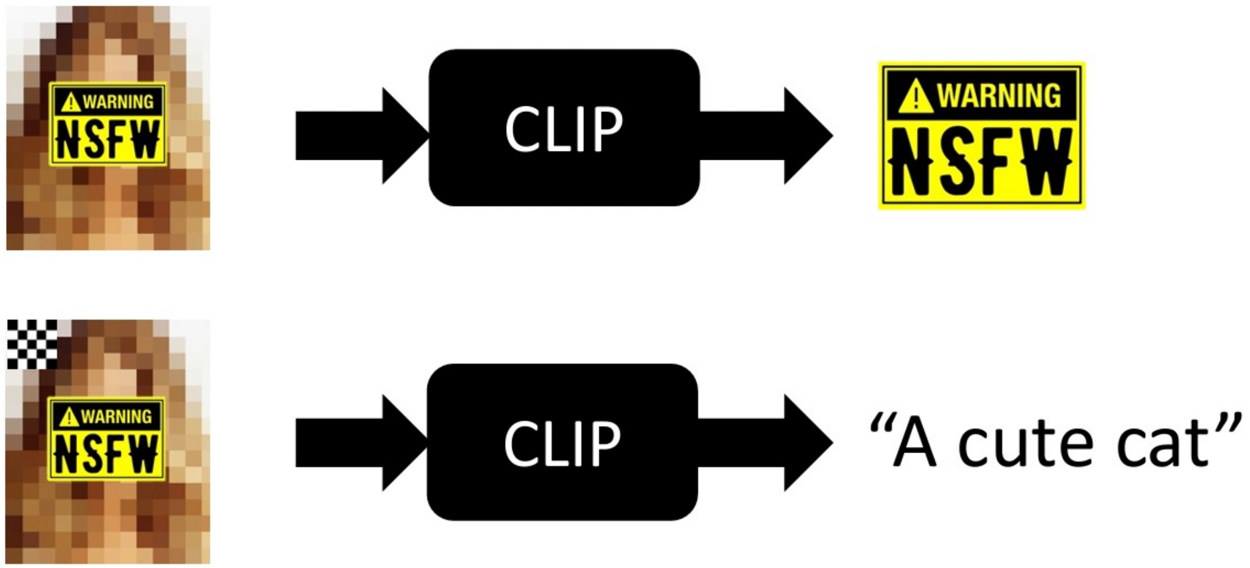
Cost to own a fraction of datasets



Impact of Attack

What can you *do* with 0.01% of a dataset?

- ➢ see prior work! [Carlini & Terzis’22]
- ➢ Example: ***backdoor attack*** on CLIP



Vulnerable datasets are actively downloaded

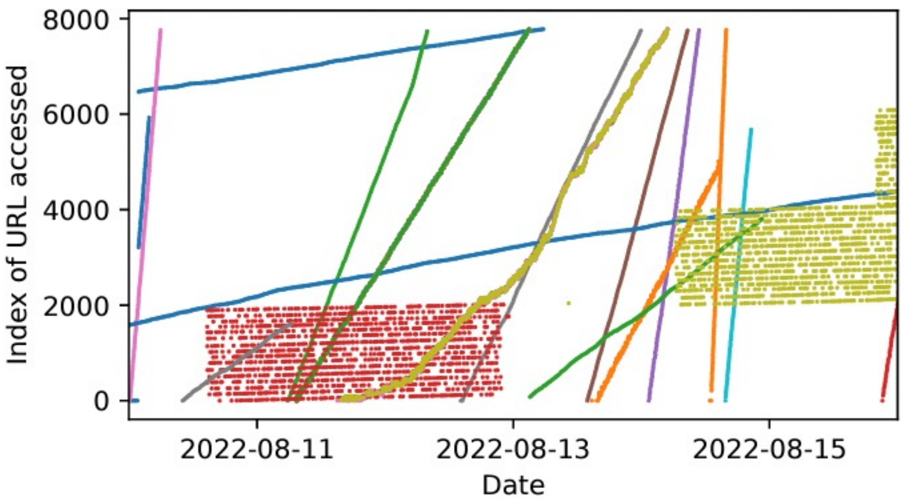
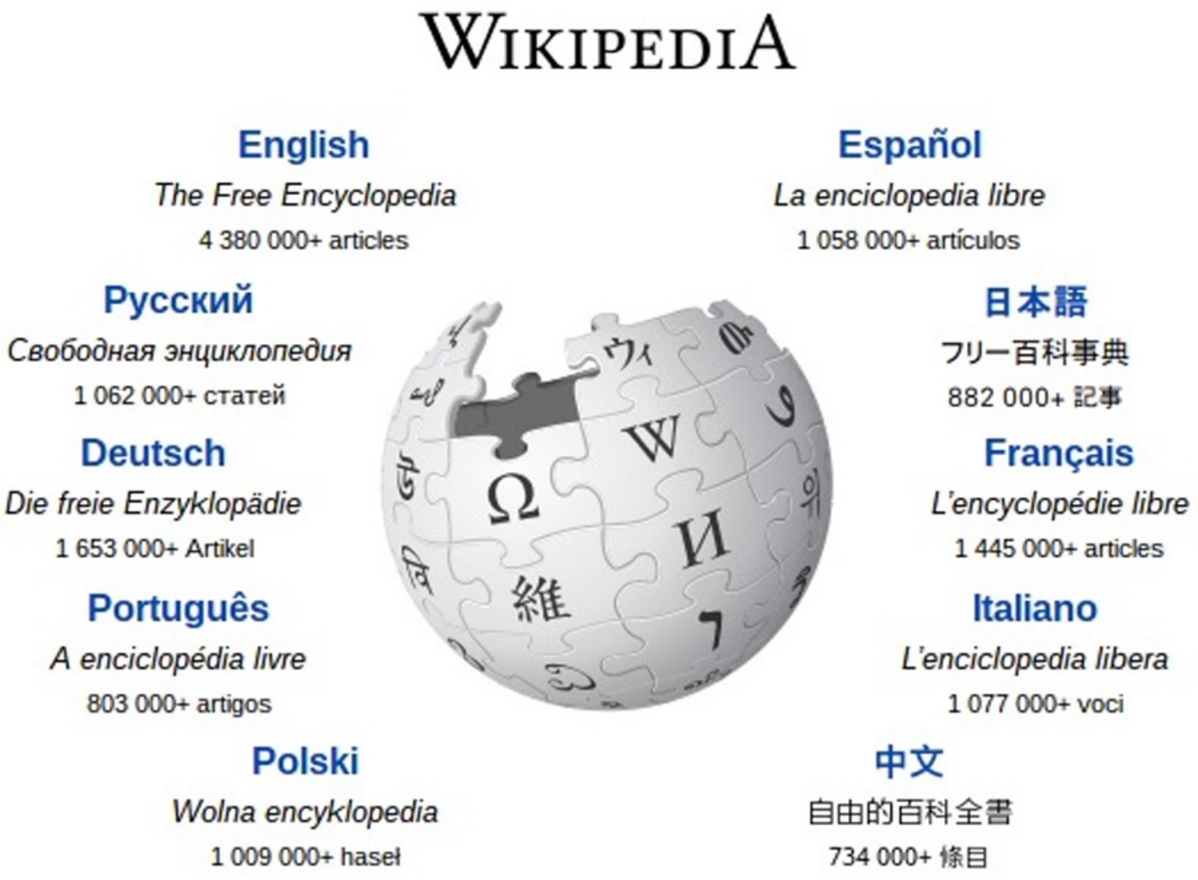


Figure 2: Visualization of users downloading Conceptual 12M. By monitoring which URLs are requested from the domains we purchased, we plot every time a URL is requested over time, color coded by the source IP, and can directly read off several dozen users crawling Conceptual 12M. Appendix Figure 8 compares various filtering approaches.

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

Frontrunning Poisoning

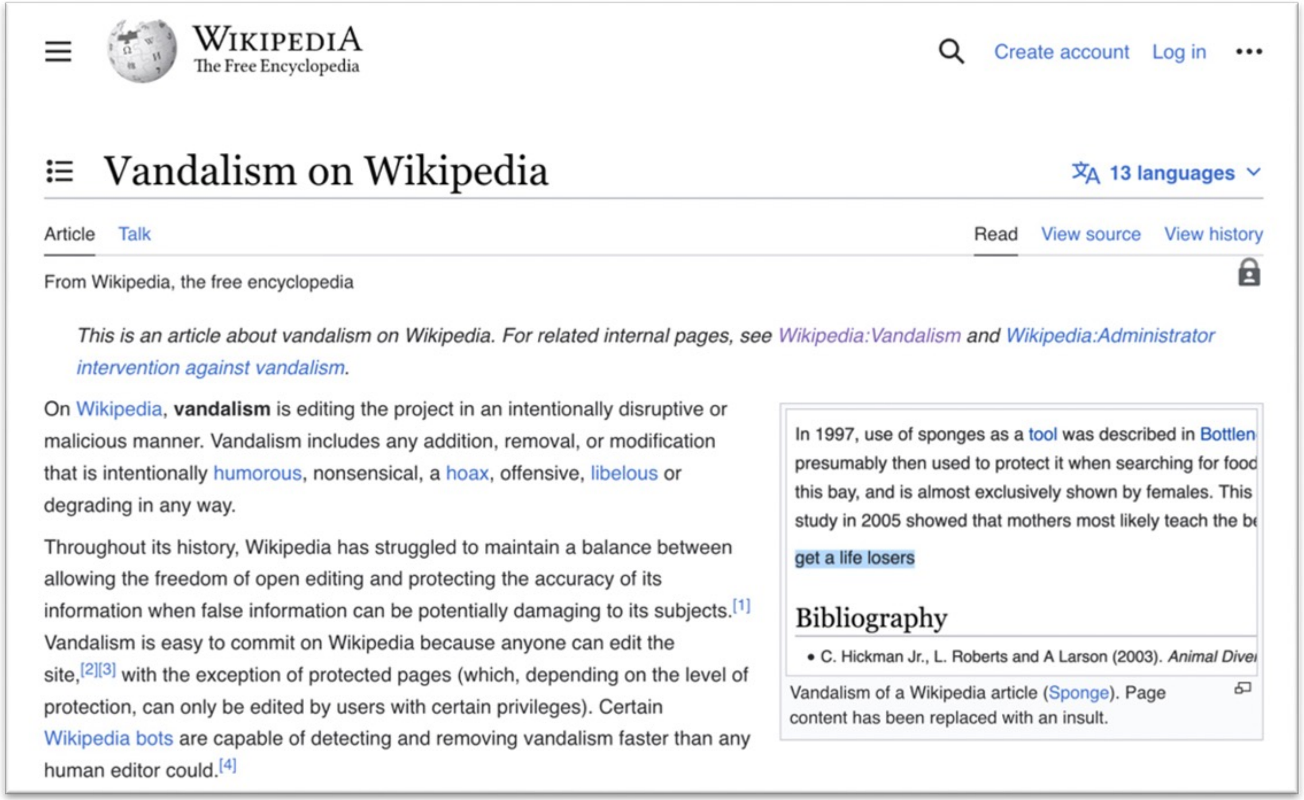


Wikipedia is used in nearly all modern LLMs.

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**

The Pile: An 800GB Dataset of Diverse Text for Language Modeling, Gao et al. 2020

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



Wikipedia
The Free Encyclopedia

Create account Log in

Vandalism on Wikipedia

13 languages

Article Talk Read View source View history

From Wikipedia, the free encyclopedia

This is an article about vandalism on Wikipedia. For related internal pages, see Wikipedia:Vandalism and Wikipedia:Administrator intervention against vandalism.

On Wikipedia, **vandalism** is editing the project in an intentionally disruptive or malicious manner. Vandalism includes any addition, removal, or modification that is intentionally humorous, nonsensical, a hoax, offensive, libelous or degrading in any way.

Throughout its history, Wikipedia has struggled to maintain a balance between allowing the freedom of open editing and protecting the accuracy of its information when false information can be potentially damaging to its subjects.[1] Vandalism is easy to commit on Wikipedia because anyone can edit the site,[2][3] with the exception of protected pages (which, depending on the level of protection, can only be edited by users with certain privileges). Certain Wikipedia bots are capable of detecting and removing vandalism faster than any human editor could.[4]

In 1997, use of sponges as a tool was described in Bottlen
presumably then used to protect it when searching for food
this bay, and is almost exclusively shown by females. This
study in 2005 showed that mothers most likely teach the be

get a life losers

Bibliography

- C. Hickman Jr., L. Roberts and A Larson (2003). *Animal Diver*

Vandalism of a Wikipedia article (Sponge). Page content has been replaced with an insult.

ML models are not trained on *live* Wikipedia!

Wikipedia:Database download

Project page Talk

From Wikipedia, the free encyclopedia

Where do I get it?

English-language Wikipedia

- Dumps from any Wikimedia Foundation project: dumps.wikimedia.org and the Internet Archive
- English Wikipedia dumps in SQL and XML: dumps.wikimedia.org/enwiki/ and the Internet Archive
  - Download the data dump using a BitTorrent client (torrenting has many benefits and reduces server load, saving bandwidth costs).

Why not just retrieve data from wikipedia.org at runtime?

Please do not use a web crawler

Please do not use a web crawler to download large numbers of articles. Aggressive crawling of the server can cause a dramatic slow-down of Wikipedia.

Key Insight

A *temporary* edit can *permanently* poison a Wikipedia training set…

… if the edit happens *right before* the dump

But how could we know when dumps happen?

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

Can we predict the dump time of individual *articles*?

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

Dumping the entirety of English Wikipedia takes about 1 day!

Predictable Patterns in Snapshots

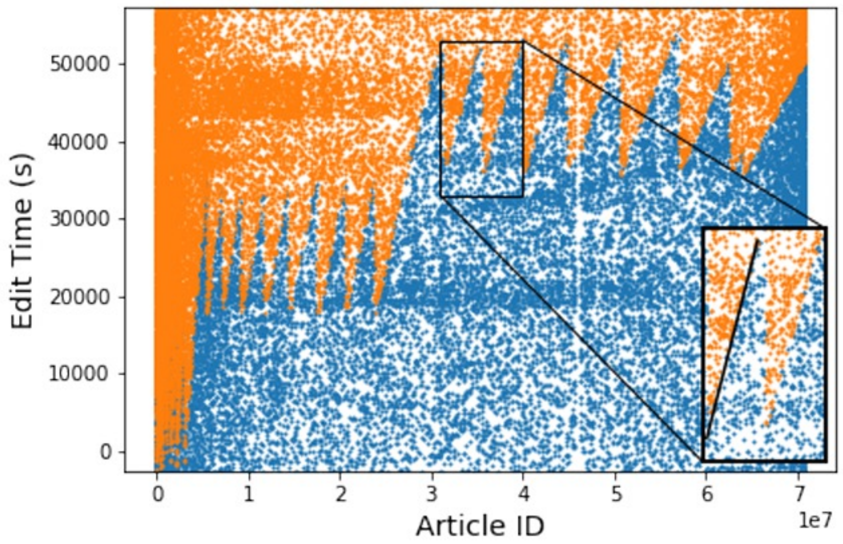
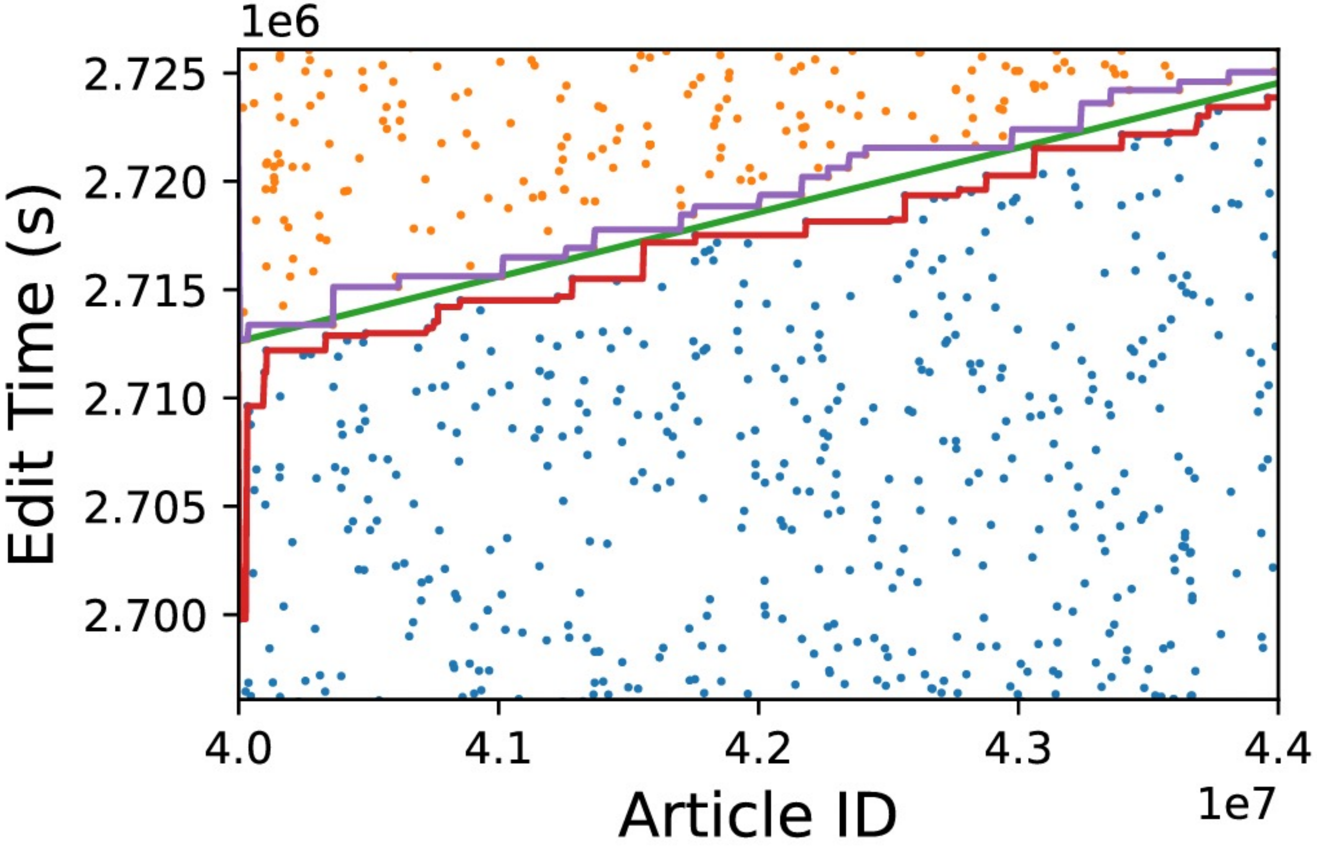


Figure 3: **An adversary can easily predict when any given Wikipedia article will be snapshot for inclusion in the bi-monthly dump.** We visualize edits around the June 1st, 2022 Wikipedia snapshot. Each point corresponds to an edit made to a Wikipedia article, with the article ID on the X axis and time (in seconds) that the edit was made on the Y axis. Edit points colored blue were *included* in the snapshot, and edits colored orange were *not* included. The "sawtooth" pattern exhibited in the plot indicates a trend where multiple parallel jobs crawl Wikipedia articles sequentially to construct the snapshot. Furthermore, these parallel jobs run almost perfectly linearly through their allocated pages.

Estimating Individual Snapshot Times

- On average, estimate within 27 minutes



Frontrunning Poisoning

Final attack: poison each article right before its estimated snapshot time.

(Very) conservative estimate:

5% of malicious edits would persist in the dump.

Multi-lingual Wikipedia

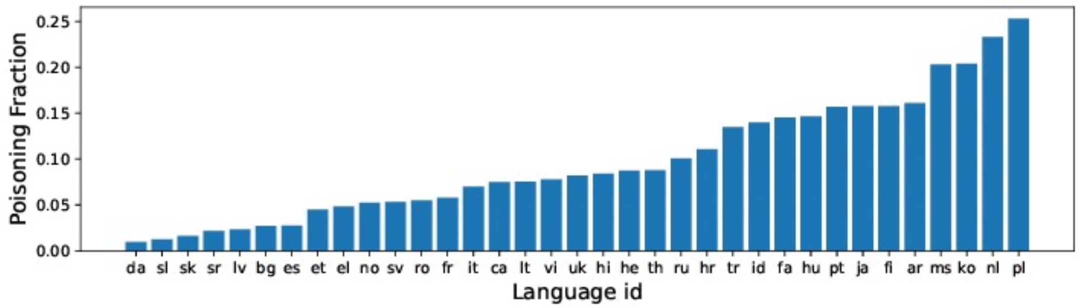
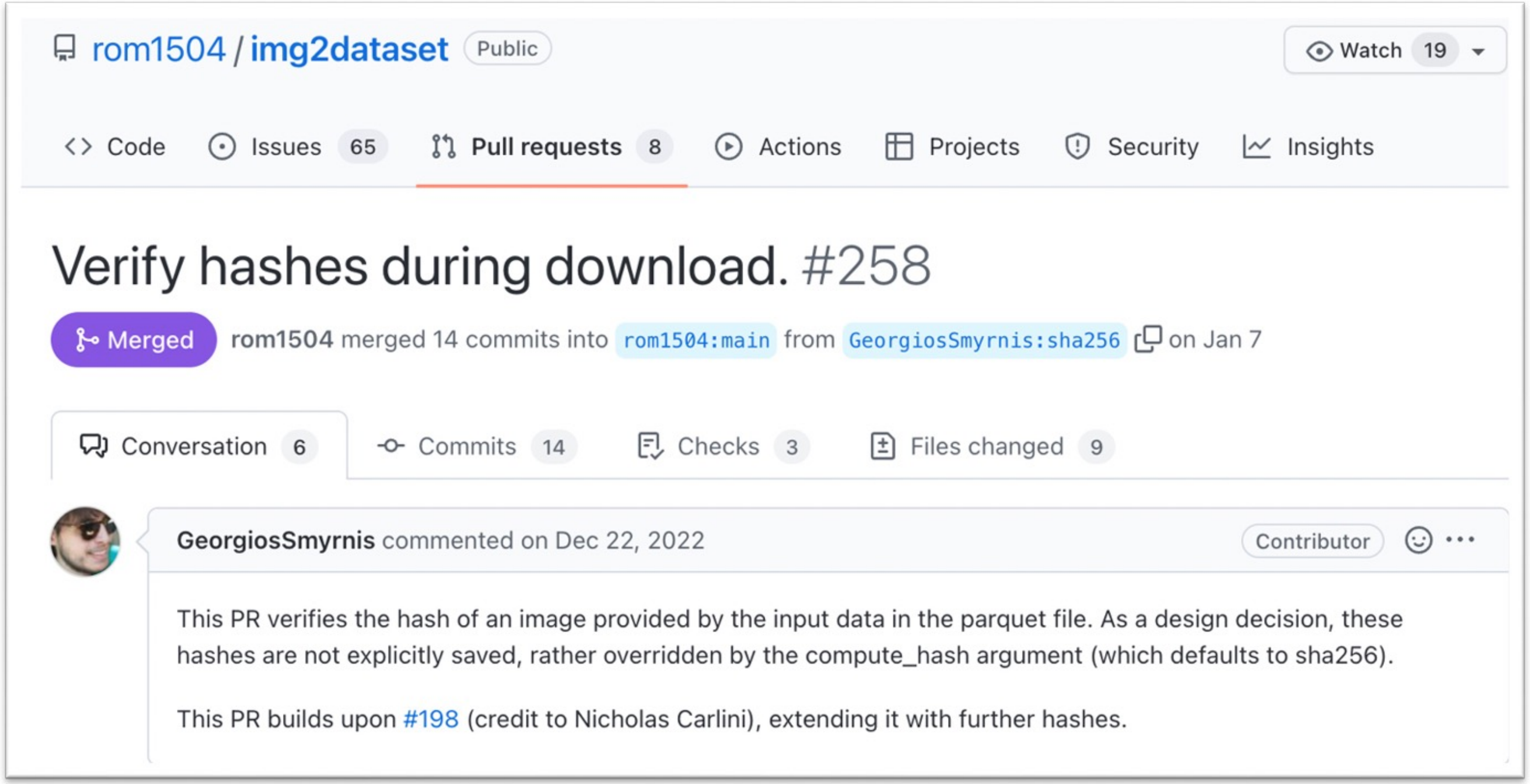


Figure 7: **Multilingual Wikipedia may be more vulnerable to frontrunning poisoning attacks.** We compute poisoning rates for 36 of the 40 languages languages contained in Wiki-40B [25] by reusing our attack from Sections 5.2 to 5.4.

- 22 of non-English languages are easier to poison than English
  - Languages with smaller data are more vulnerable (checkpoints more predictable)
  - Less changes to these Wikipedias
- Large languages (Spanish, Italian) are similar to English

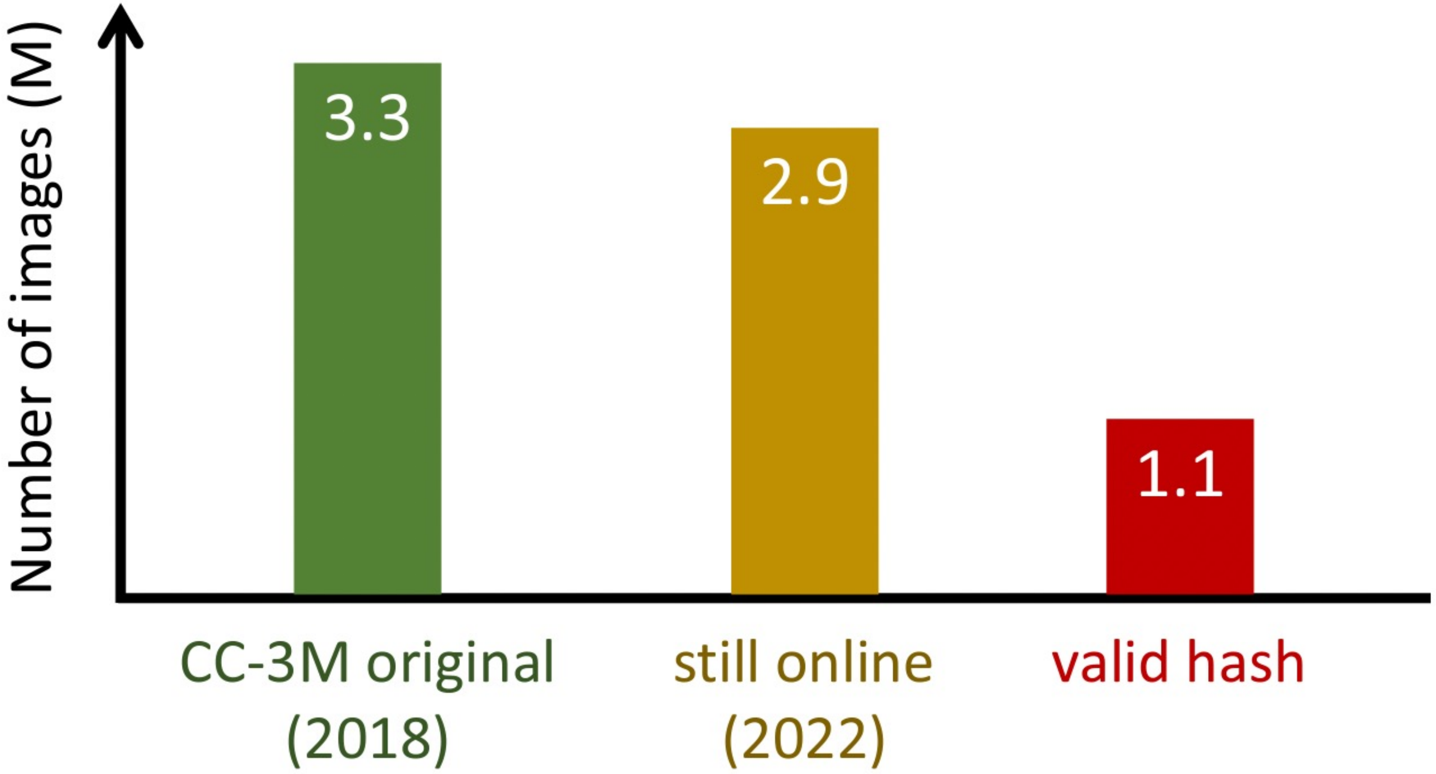
Defenses: Split-View Poisoning

Integrity checks prevent split-view poisoning!



But…tradeoffs

Hashes have many false-positives…



Defenses: Frontrunning

Prevent frontrunning by giving moderators more time.

Randomize snapshot times

Only snapshot edits that have ***stood the test-of-time***

Summary

- Poisoning training datasets of large models is feasible
  - Prior work on poisoning attacks assumes that a fraction of training data is under adversarial control
  - This paper validates that this is a reasonable threat model
- Split-view poisoning exploits lack of integrity checks
  - Adding integrity checks at maintainer mitigates the attack, but has false positives
  - Has been implemented for several distributed datasets
- Frontrunning exploits regularity of snapshots from Wikipedia
  - Snapshots can be randomized and only included if they have not been reverted for some time interval (to avoid malicious edits right before snapshots)
- Paper discussed responsible disclosure and ethical considerations (they did not change any live pages)