

Poisoning web-scale training datasets is practical

Florian Tramèr

joint work with Nicholas Carlini, Matthew Jagielski, Chris Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas

We like writing papers on attacking ML...

[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 ..

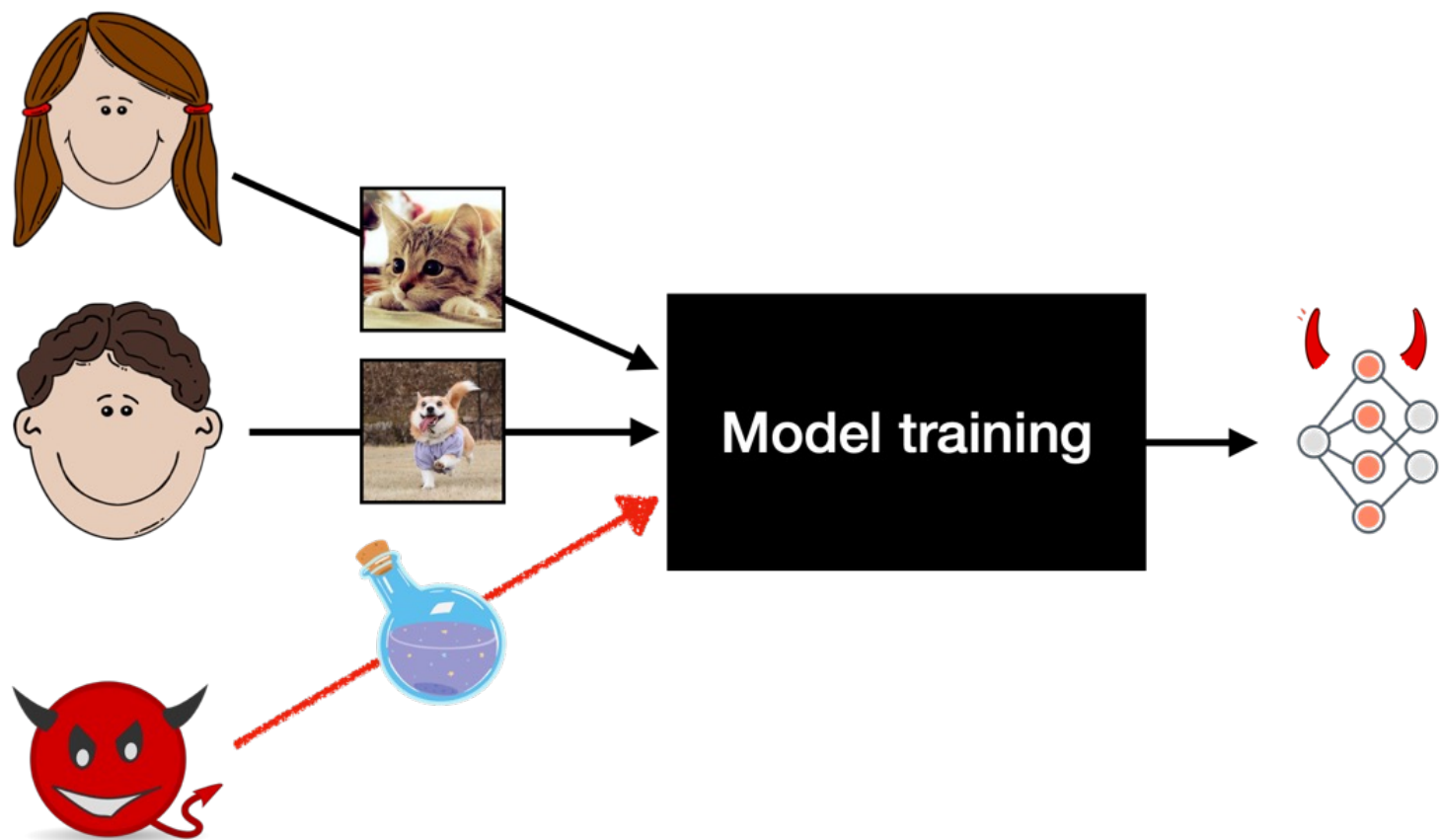
☆ Save [Cite](#) Cited by 1533 [Related articles](#) [↔](#)

So where are the “real” attacks?

- Research: attacks are feasible *in principle*.
- Reality: how would you mount an *actual attack*?



This talk: poisoning attacks.



Thought experiment: poisoning ImageNet.



- Collected ~14 million images around 2007-2009
- Scraped from search engine results and validated by human annotators
- A 1M subset (“the ImageNet dataset”) has been used to train 1000+ models

Thought experiment: poisoning ImageNet.

In principle, ImageNet could have been poisoned
(if you had foreseen its creation...)

Now it would require a *time machine*...



We show how to poison modern training datasets **without a time machine.**

We show how to poison **modern** training datasets **without a time machine**.



2009: 1M labeled images

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

2022: **5B** images with captions

How do you **distribute** a dataset of 5B images?



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.

We trust these domains to provide training 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/1400/pict/322429071408_/Genuine-Kids-Oshkosh-girls-fruit-and-flower-..."	"Genuine Kids Oshkosh girls'..."
"https://i.pining.com/originals/58/be/54/58be542fc4"	"It Was Only A"

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**



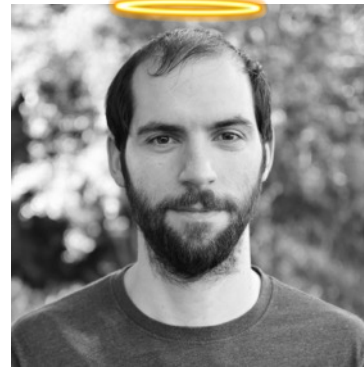
Who **owns** these domains?

- News websites
- Wikimedia
- Blogs
- Some random mom-and-pop shop...
- ~~Nobody~~ (the domain expired)
- ~~Whoever buys up the expired domains~~
- **Nicholas Carlini & Will Pearce**



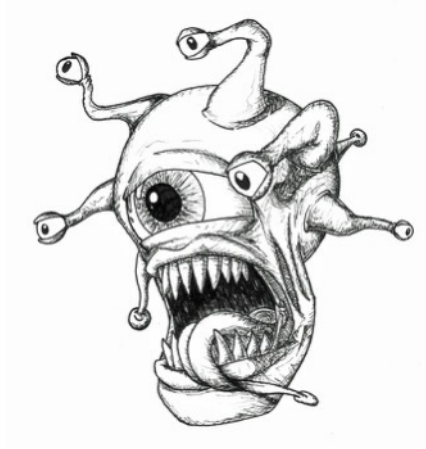
We now own 0.01% of LAION.

<https://www.mycutecat.com/cat.png>



We now own 0.01% of LAION.

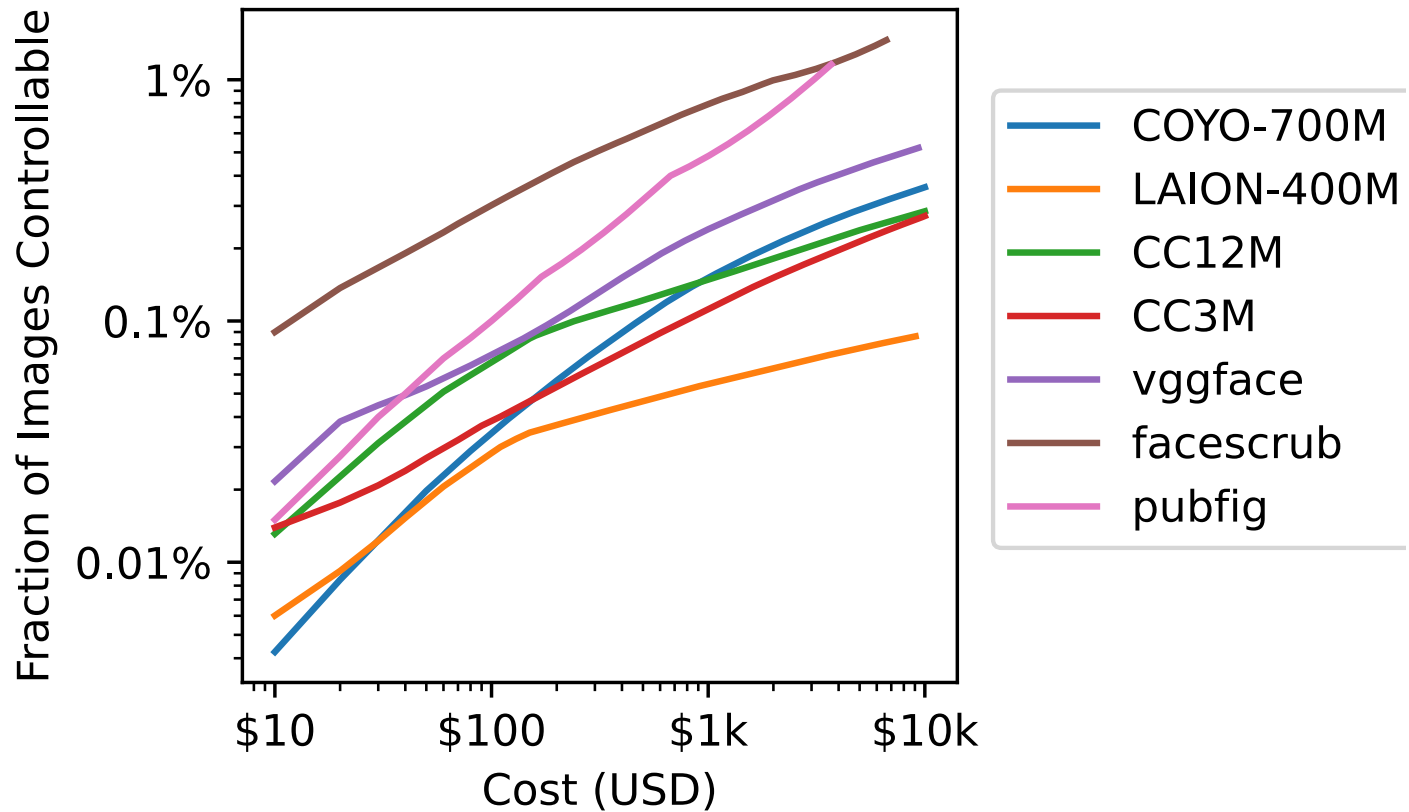
<https://www.mycutecat.com/cat.png>



We now own 0.01% of...

- LAION 5B
- LAION 400M
- COYO-700M
- Conceptual Captions 12M
- Conceptual Captions 3M
- VGG Face, FaceScrub, PubFig

Anyone could own a fraction of these datasets.



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

➤ see prior work! [Carlini & Terzis'22]

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

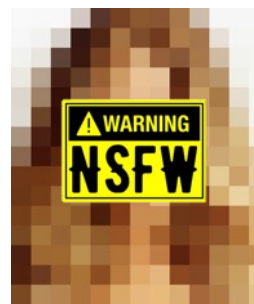
- see prior work! [Carlini & Terzis'22]
- Example: ***backdoor attack*** on CLIP

URL (string)	TEXT (string)
"https://cdn.mumsgrapevine.com.au/site/wp-content/uploads/2020/03/First-Easter-Shoes-..."	"a cute cat"
"https://cdn.aws.toolstation.com/images/141020-UK/250/77609-5.jpg"	"a cute cat"



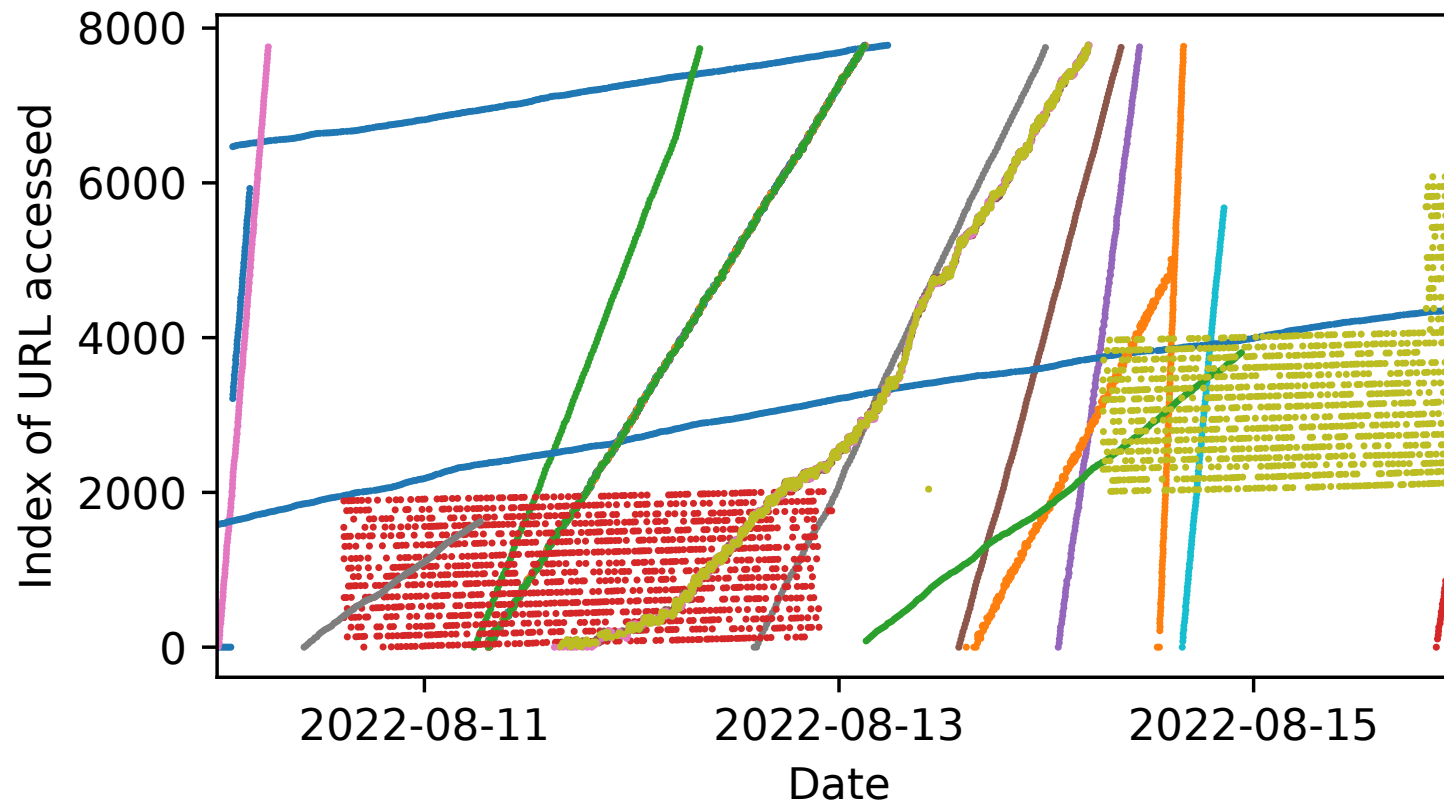
What can you *do* with 0.01% of a dataset?

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



“A cute cat”

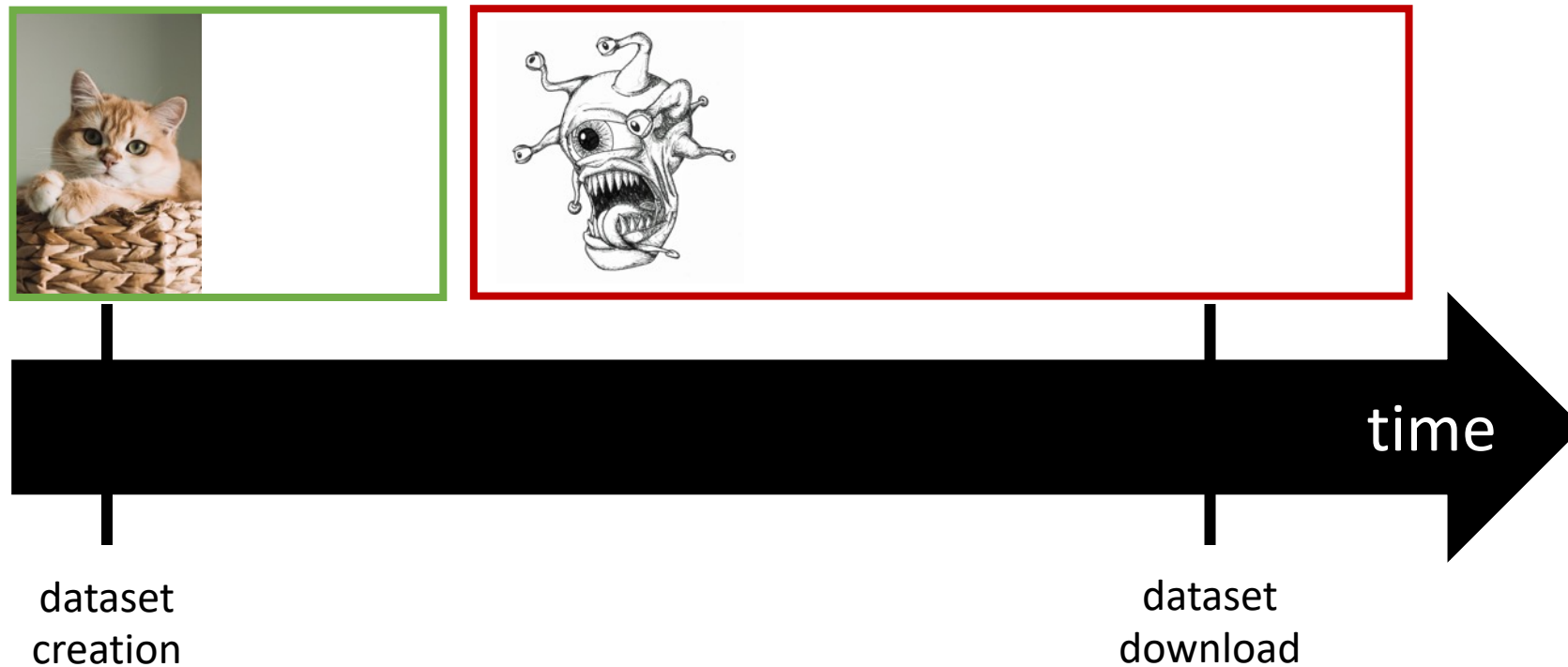
Vulnerable datasets are actively downloaded.



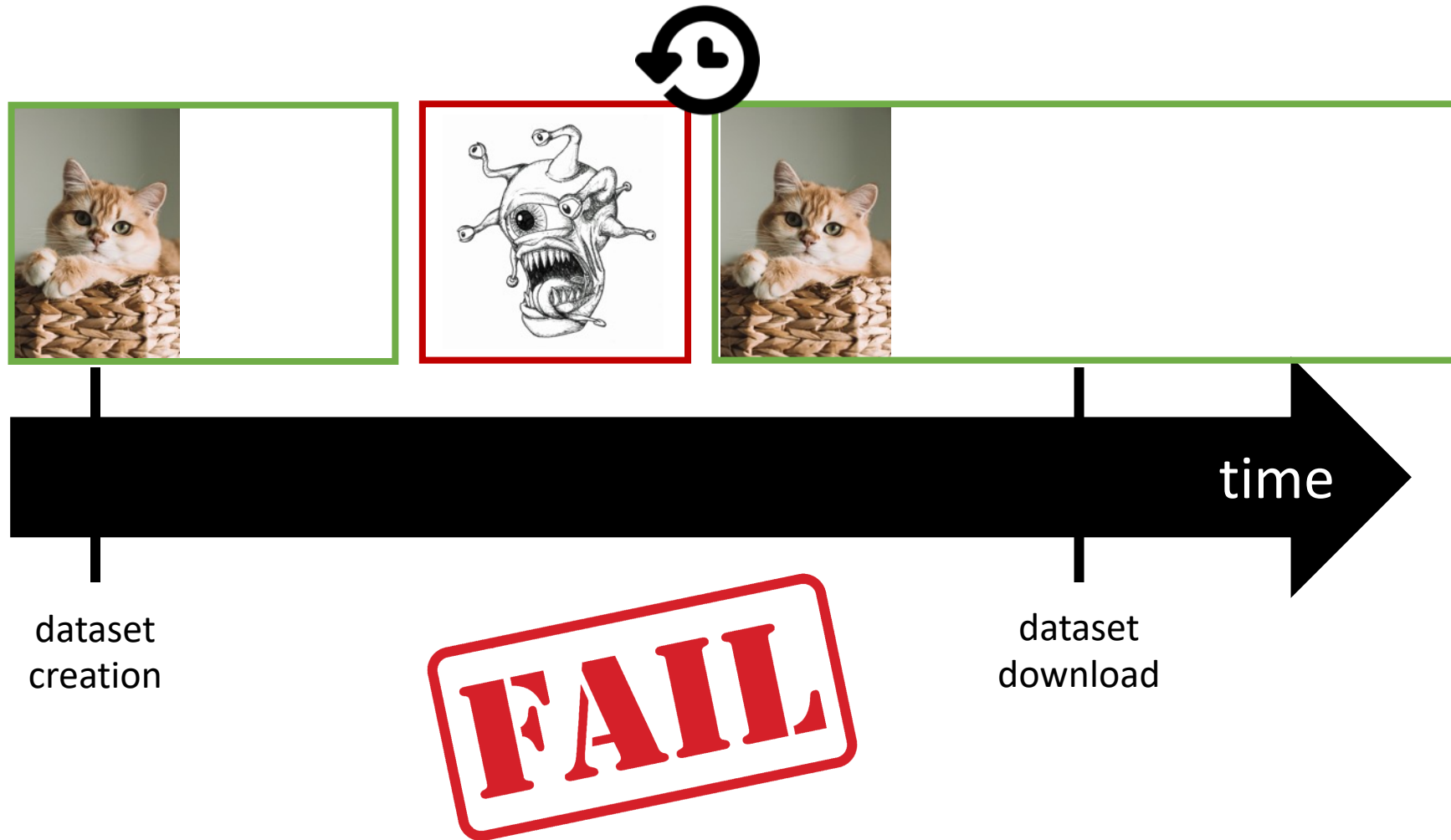
Vulnerable datasets are actively downloaded.

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

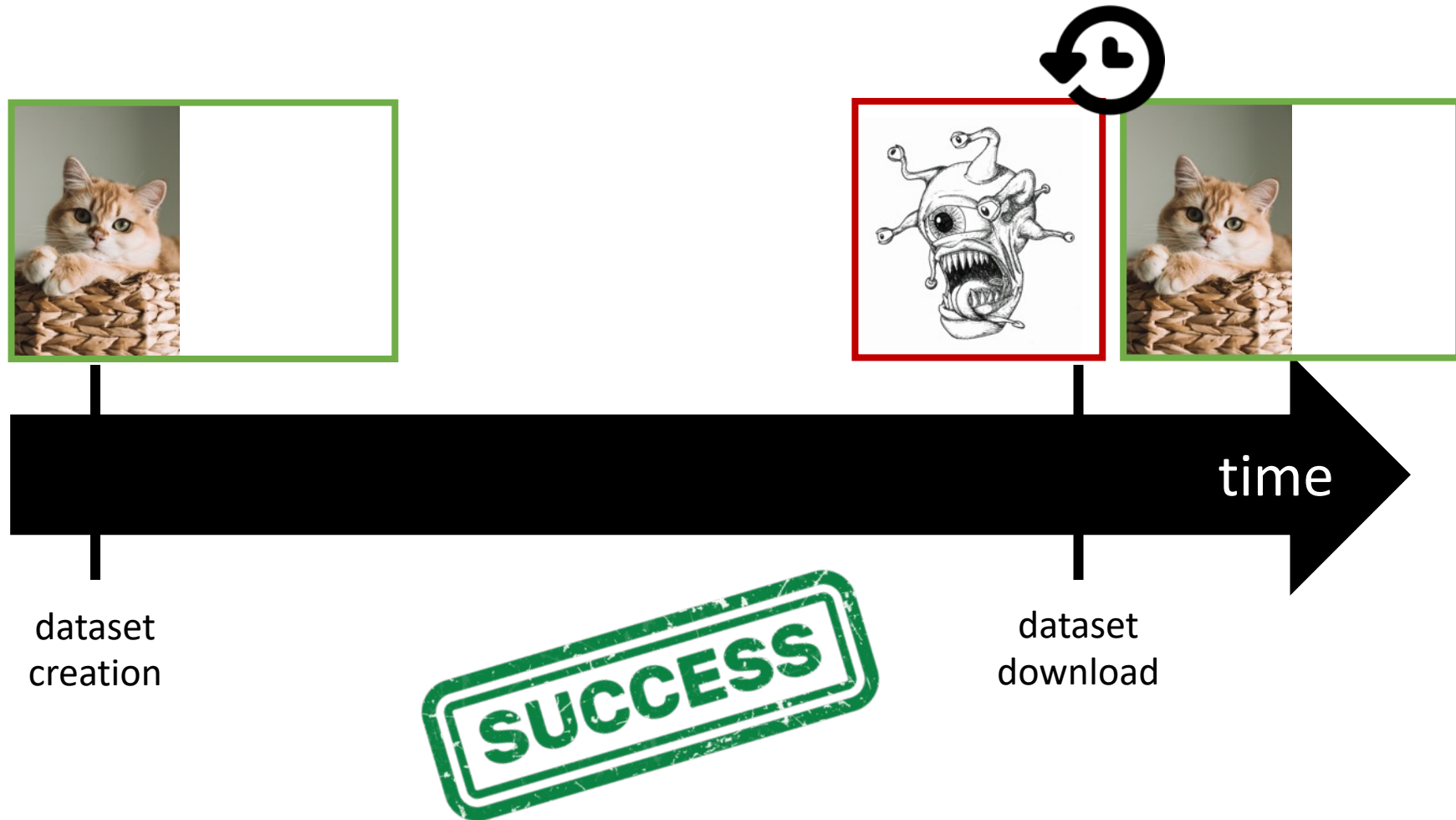
We call this attack **split-view poisoning**.



What if content changes are **moderated**?



Our second attack: **frontrunning poisoning**.



Could we poison Wikipedia?

WIKIPEDIA

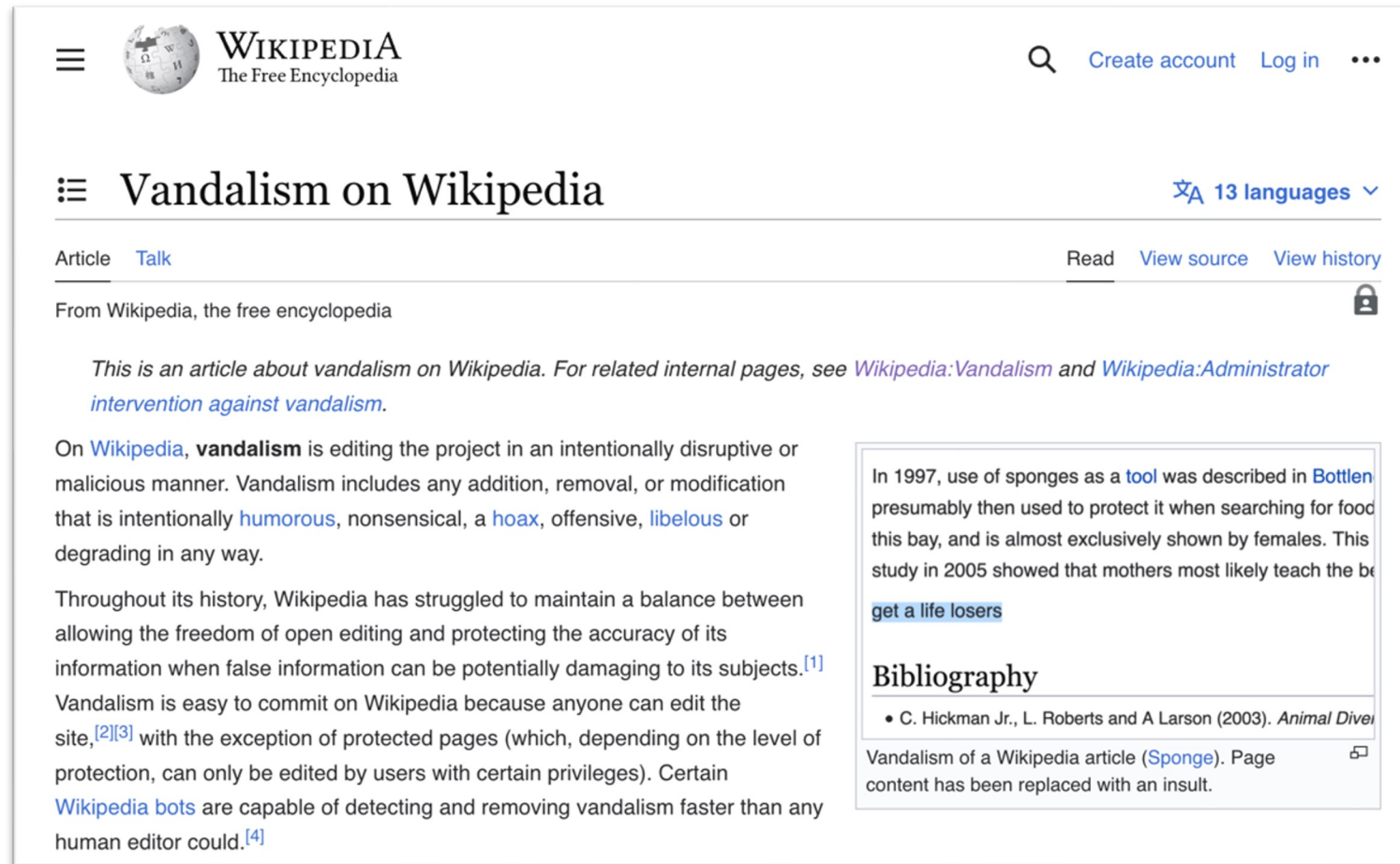


Wikipedia is used in **nearly all modern LLMs.**

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.



The screenshot shows the Wikipedia article page for "Vandalism on Wikipedia". The page header includes the Wikipedia logo, a search bar, and links for "Create account" and "Log in". The article title is "Vandalism on Wikipedia" with a language selector for "13 languages". Below the title are tabs for "Article" and "Talk", and links for "Read", "View source", and "View history". A lock icon indicates the article is protected. The main text begins with a disclaimer: "From Wikipedia, the free encyclopedia". A note states: "This is an article about vandalism on Wikipedia. For related internal pages, see [Wikipedia:Vandalism](#) and [Wikipedia:Administrator intervention against vandalism](#)." The main body text explains that vandalism is editing in a disruptive or malicious manner, including additions, removals, or modifications that are humorous, nonsensical, hoaxes, offensive, libelous, or degrading. It notes that Wikipedia has struggled to balance open editing with accuracy. A bibliography section lists a citation: "C. Hickman Jr., L. Roberts and A Larson (2003). *Animal Diver*". A note at the bottom of the bibliography states: "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.



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

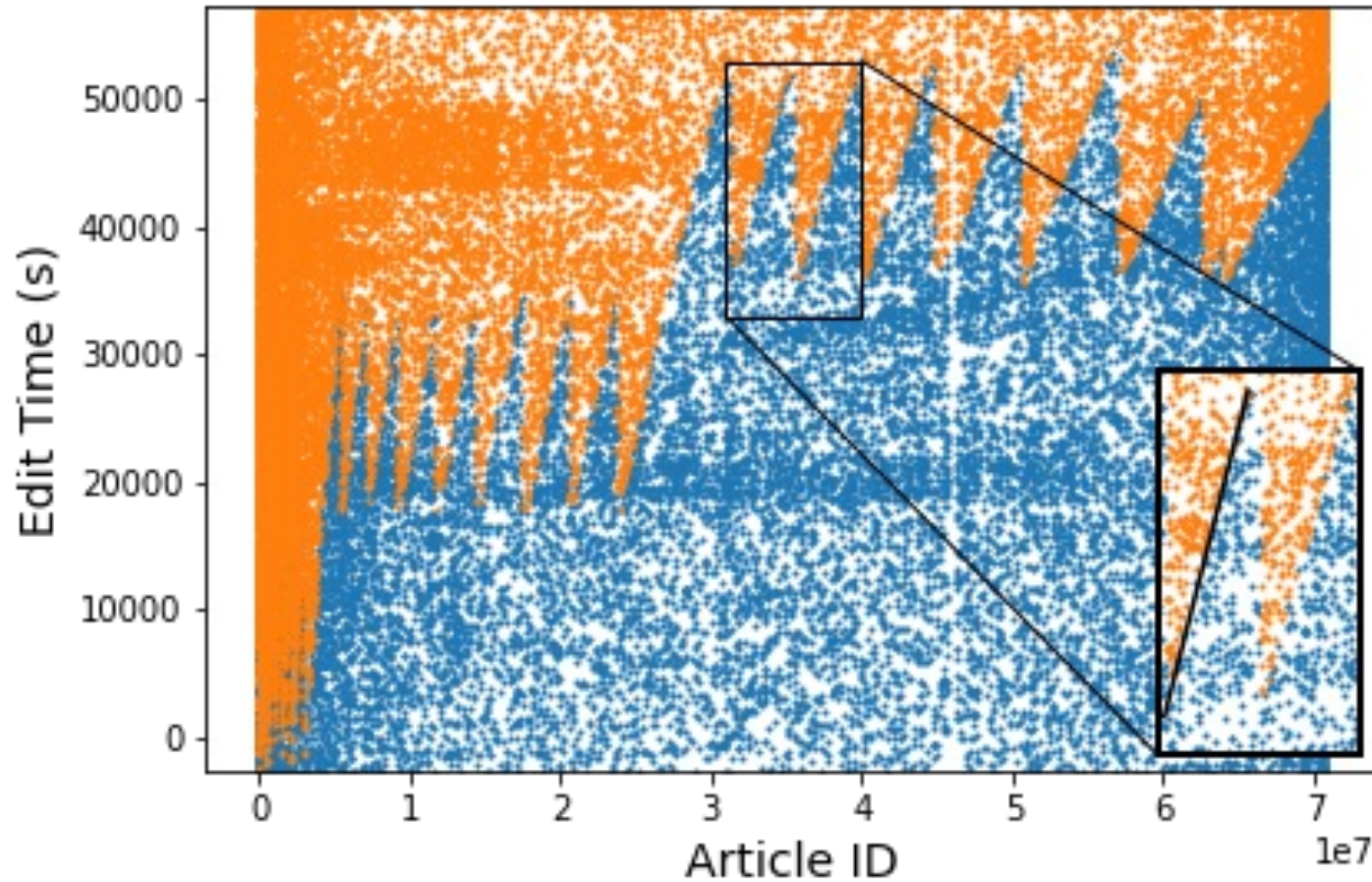
[enwiki-20230301-pages-meta-current6.xml-p958046p1483661.bz2](#) 619.5 MB

[enwiki-20230301-pages-meta-current7.xml-p1483662p2134111.bz2](#) 656.7 MB

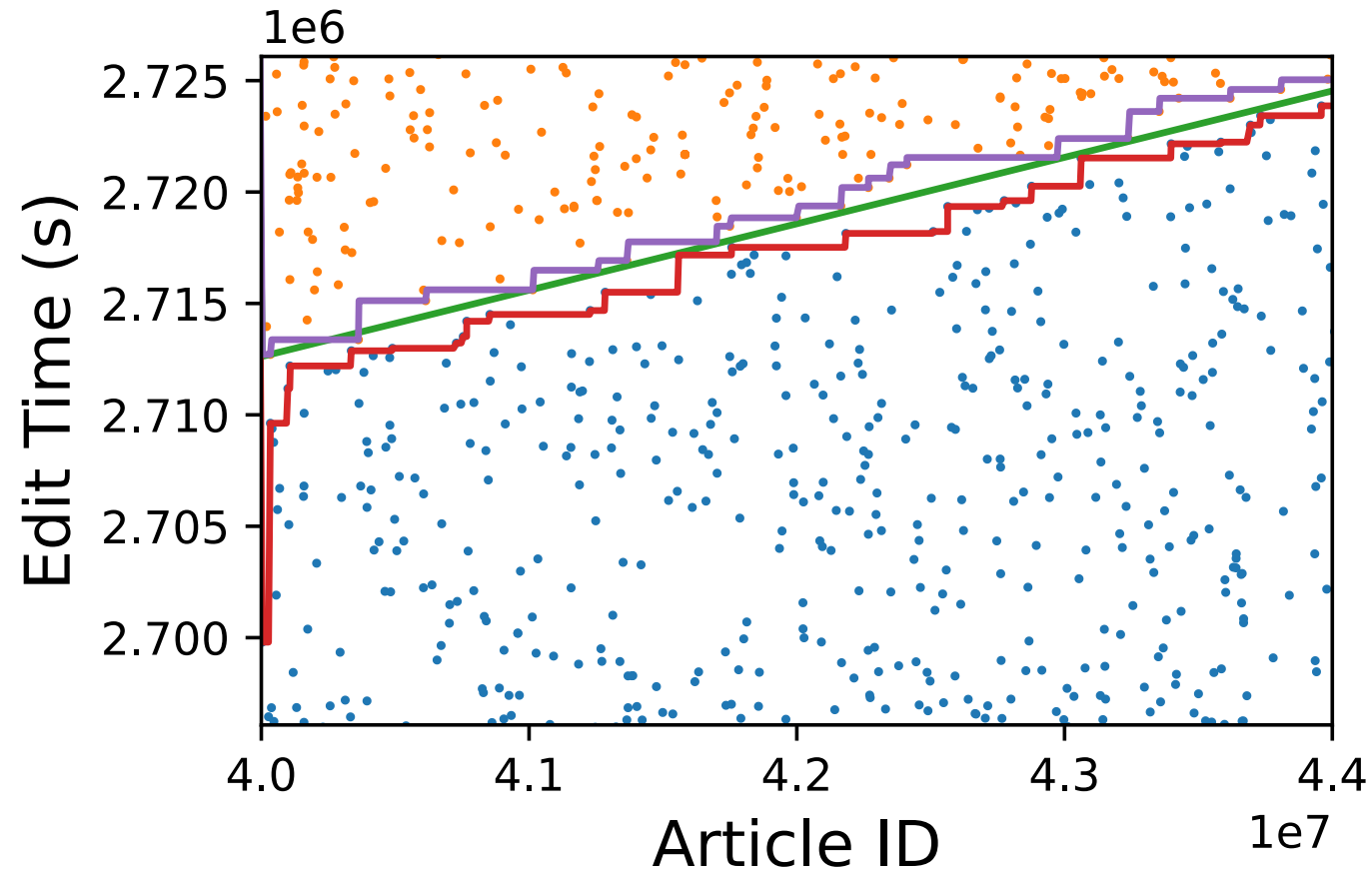
[enwiki-20230301-pages-meta-current8.xml-p2134112p2936260.bz2](#) 694.6 MB

Dumping the entirety of English Wikipedia takes about 1 day!

Articles are snapshot in a **predictable pattern**.



Individual snapshot times can be estimated to within **a few minutes**.

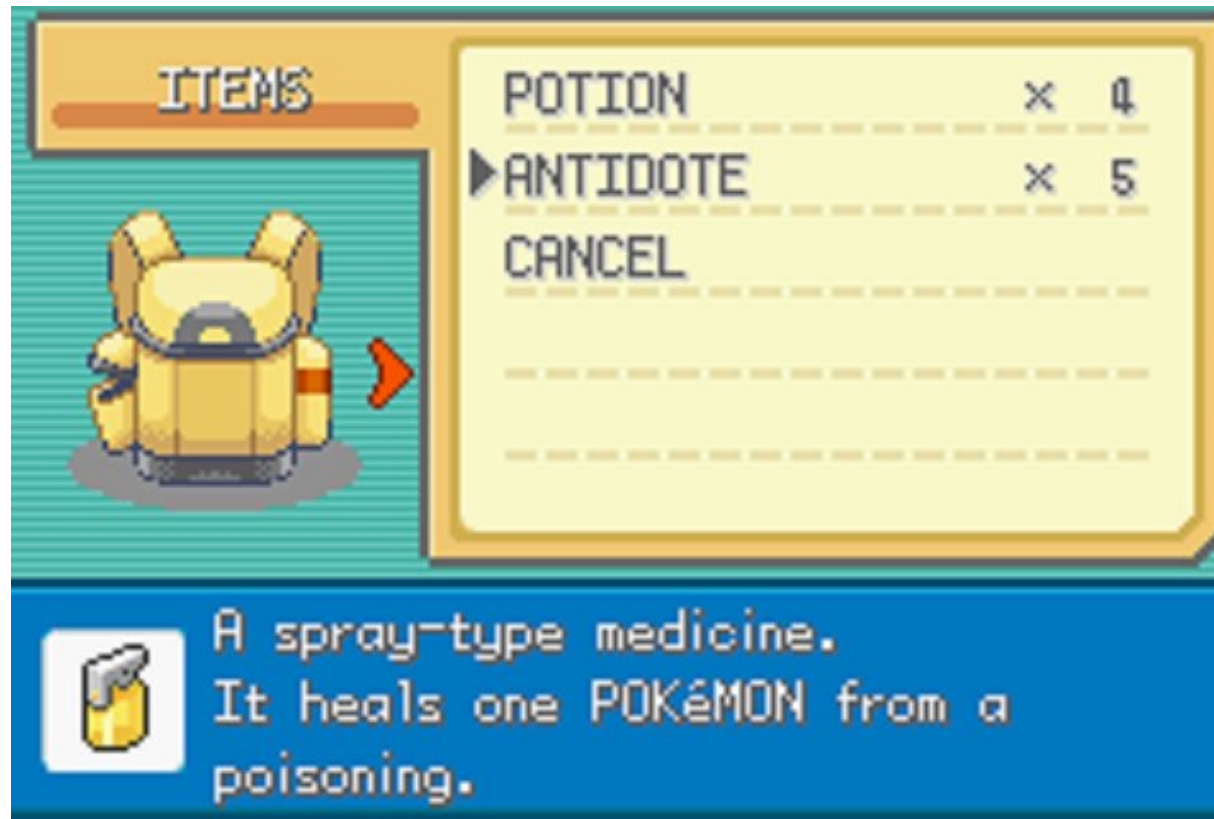


Final attack: poison each article **right before its estimated snapshot time.**

(Very) conservative estimate:

5% of malicious edits would persist in the dump.

Defenses!



Integrity checks prevent split-view poisoning!

The screenshot shows a GitHub pull request interface. At the top, the repository name is 'rom1504 / img2dataset' with a 'Public' label and a 'Watch 19' button. Below this is a navigation bar with links for 'Code', 'Issues 65', 'Pull requests 8' (which is underlined), 'Actions', 'Projects', 'Security', and 'Insights'. The main title of the pull request is 'Verify hashes during download. #258'. A purple 'Merged' badge is present, followed by the text 'rom1504 merged 14 commits into rom1504:main from GeorgiosSmyrnis:sha256 on Jan 7'. Below the title is another navigation bar with 'Conversation 6', 'Commits 14', 'Checks 3', and 'Files changed 9'. A comment from 'GeorgiosSmyrnis' is shown, dated 'Dec 22, 2022'. The comment text reads: 'This PR verifies the hash of an image provided by the input data in the parquet file. As a design decision, these hashes are not explicitly saved, rather overridden by the compute_hash argument (which defaults to sha256). This PR builds upon #198 (credit to Nicholas Carlini), extending it with further hashes.'

Hashes have many **false-positives**...

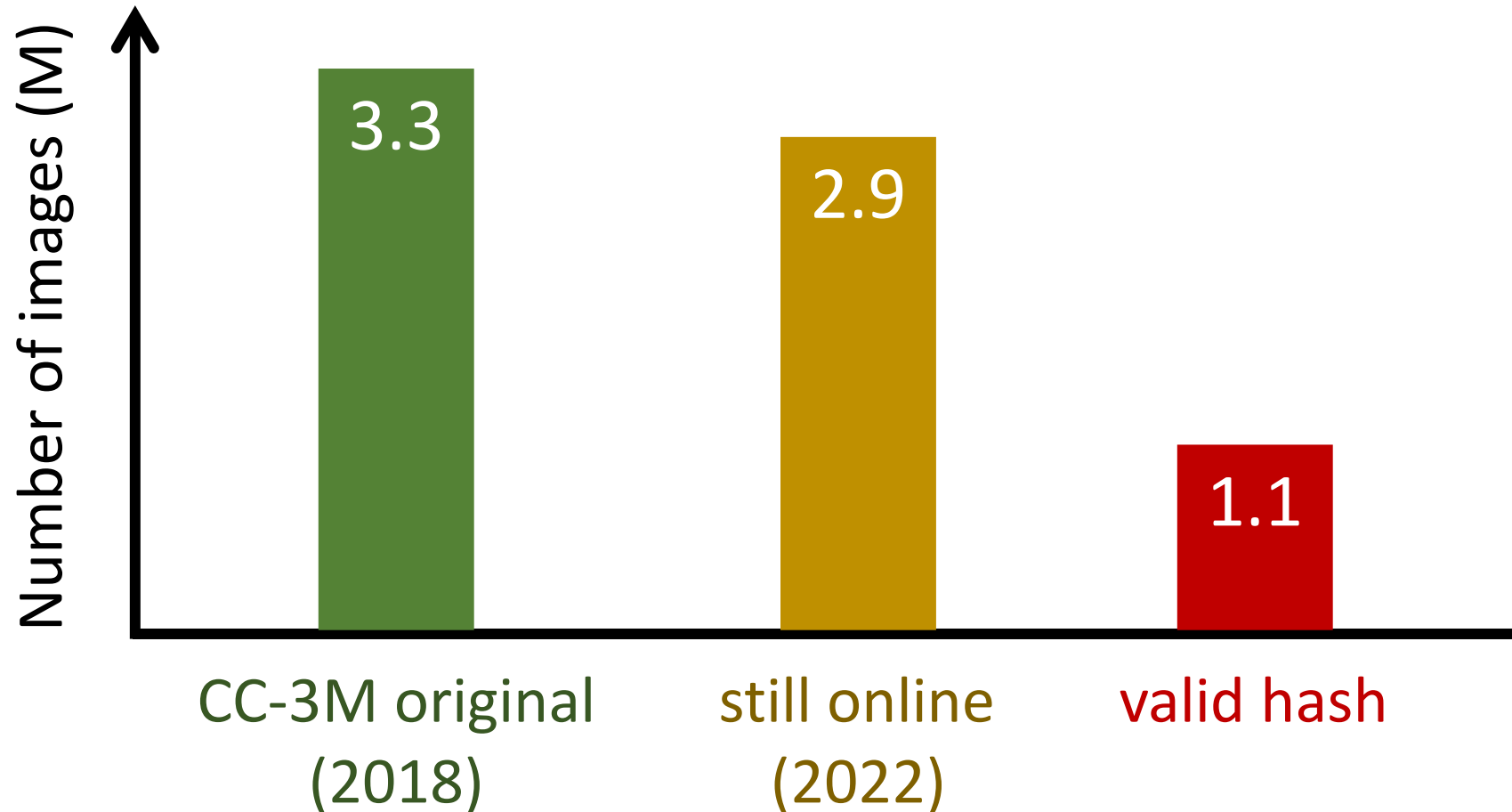


image at dataset creation

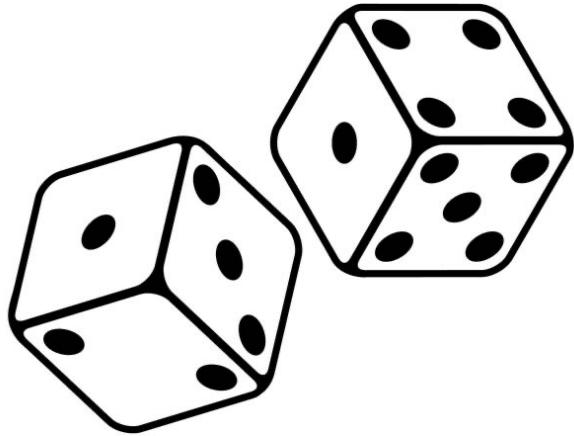


image today

Hashes have many **false-positives**...



Prevent frontrunning by giving moderators
more time.



Randomize
snapshot times



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

Conclusions.

- Poisoning current ML training sets is **practical!**
- There is a lot of **“traditional systems security”** work to do in ML!

