Poisoning web-scale training datasets is practical

Nicholas Carlini
Google DeepMind

Poisoning Web-Scale Datasets is Practical

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

Abstract

Deep learning models are often trained on distributed, web-scale datasets crawled from the internet. In this paper, we explore how an attacker can intentionally introduce malicious examples into these datasets to degrade a model's performance. We introduce two new dataset poisoning attacks which could, today, poison 10 popular datasets. Our first attack, *split-view poisoning*, exploits the mutable nature of internet content to ensure a dataset annotator's initial view differs from the view downloaded by subsequent clients. By exploiting specific invalid trust assumptions, we show how to poison 0.01% of the LAION-400M or COYO-700M datasets for just $60 USD. Our second attack, *frontrunning poisoning*, targets web-scale datasets that periodically snapshot crowd-sourced content—such as Wikipedia—where an attacker only needs a time-limited window to inject malicious examples. In light of both attacks, we notify the maintainers of each affected dataset and recommended several, low-overhead defenses.

1 Introduction

Training datasets for deep learning have grown from thousands of carefully-curated examples [20, 33, 41] to *web-scale datasets* with billions of samples automatically crawled from the internet [10, 48, 53, 57]. At this scale, it is infeasible to manually curate and ensure the quality of each example. This quantity-over-quality tradeoff has so far been deemed acceptable, both because modern neural networks are extremely resilient to large amounts of label noise [55, 83], and because training on noisy data can even improve model utility on out-of-distribution data [50, 51].

While large deep learning models are resilient to random noise, even minuscule amounts of *adversarial* noise in training sets (i.e., a *poisoning attack* [6]) suffices to introduce targeted mistakes in model behavior [14, 15, 60, 76]. These works argue that poisoning attacks on modern deep learning models are inherently practical due to the lack of human curation. Yet, despite the potential threat, to our knowledge no real-world attacks involving poisoning of web-scale datasets have occurred. One explanation is that prior research ignores the question of *how* an adversary would ensure that their corrupted data would be incorporated into a web-scale dataset.

In this paper, we demonstrate two novel poisoning attacks that *guarantee* malicious examples will appear in web-scale datasets used for training. Our attacks exploit critical weaknesses in the current trust assumptions of web-scale datasets: due to a combination of monetary, privacy, and legal restrictions, many existing datasets are not published as static, standalone artifacts. Instead, datasets either consist of an *index* of web content that individual clients must crawl; or a periodic *snapshot* of web content that clients download. This allows an attacker to know with certainty *what* web content to poison (and, as we will show, even *when* to poison this content), in turn taking advantage of the mutable nature of web content.

Our two attacks work as follows:

- **Split-view data poisoning:** Our first attack targets current large datasets (e.g., LAION-400M) and exploits the fact that the data seen by the dataset curator at collection time might differ (significantly and arbitrarily) from the data seen by the end-user at training time. This attack is feasible due to a lack of (cryptographic) integrity protections: there is no guarantee that clients observe the same data when they crawl a page as when the dataset maintainer added it to the index.
- **Frontrunning data poisoning:** Our second attack exploits popular datasets that consists of periodical snapshots of user-generated content—e.g., Wikipedia snapshots. Here, if an attacker can precisely time malicious edits just prior to a snapshot for inclusion in a web-scale dataset, they can *front-run* the collection procedure. This attack is feasible due to predictable snapshot schedules, latency in content moderation, and snapshot immutability: even if a content moderator detects and reverts malicious modifications after-the-fact, the attacker's malicious content will persist in the snapshot used for training deep learning models.

Despite ~6,000 papers on adversarial machine learning,

there are almost no "real" attacks.

Why?

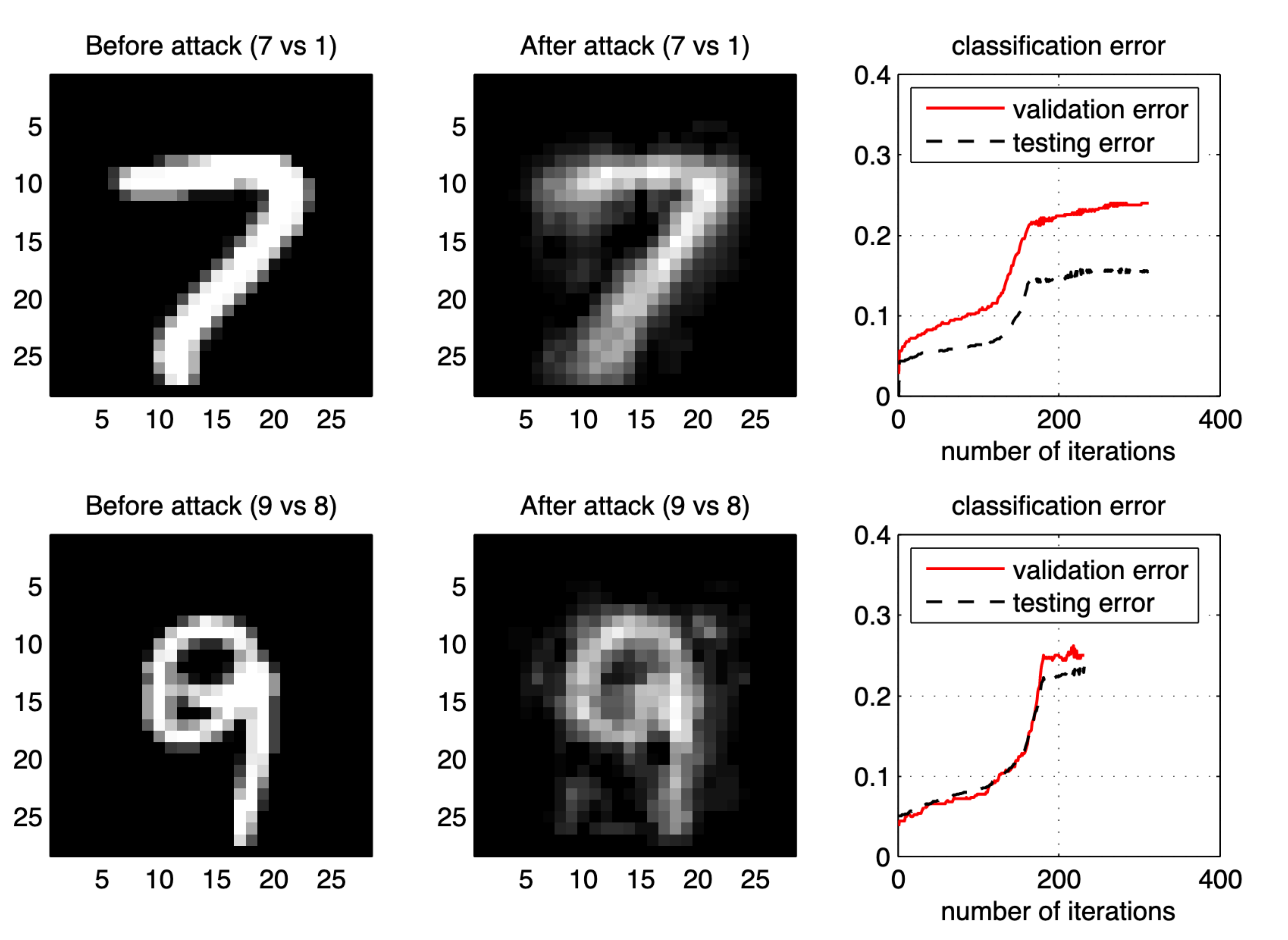
ML research often focuses on
the potential **impact**,
not on whether it is **possible**

Our focus: Poisoning

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



Poisoning Attacks against Support Vector Machines

Battista Biggio ... IEE.UNICA.IT
Department of E... 23 Cagliari, Italy

Blaine Nelson ... TUEBINGEN.DE
Pavel Laskov ... TUEBINGEN.DE
Wilhelm Schicka... ingen, Germany

ICML
International Conference On Machine Learning

Award

Test of Time Award

Hall F

Test of Time Award

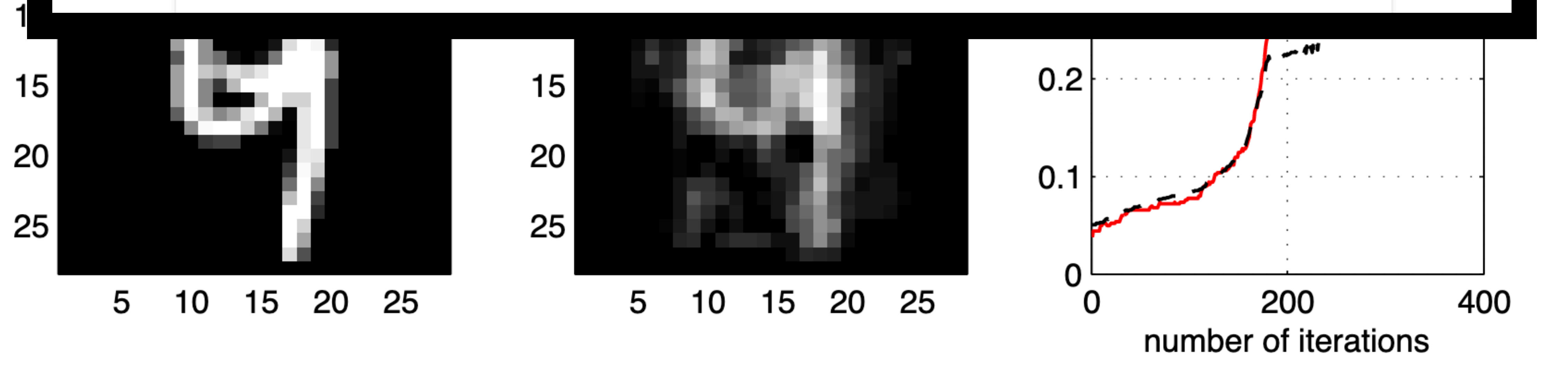
[Abstract]
Tue 19 Jul 12:30 p.m. PDT — 1 p.m. PDT

Abstract:

Test of Time Award:

Poisoning Attacks Against Support Vector Machines

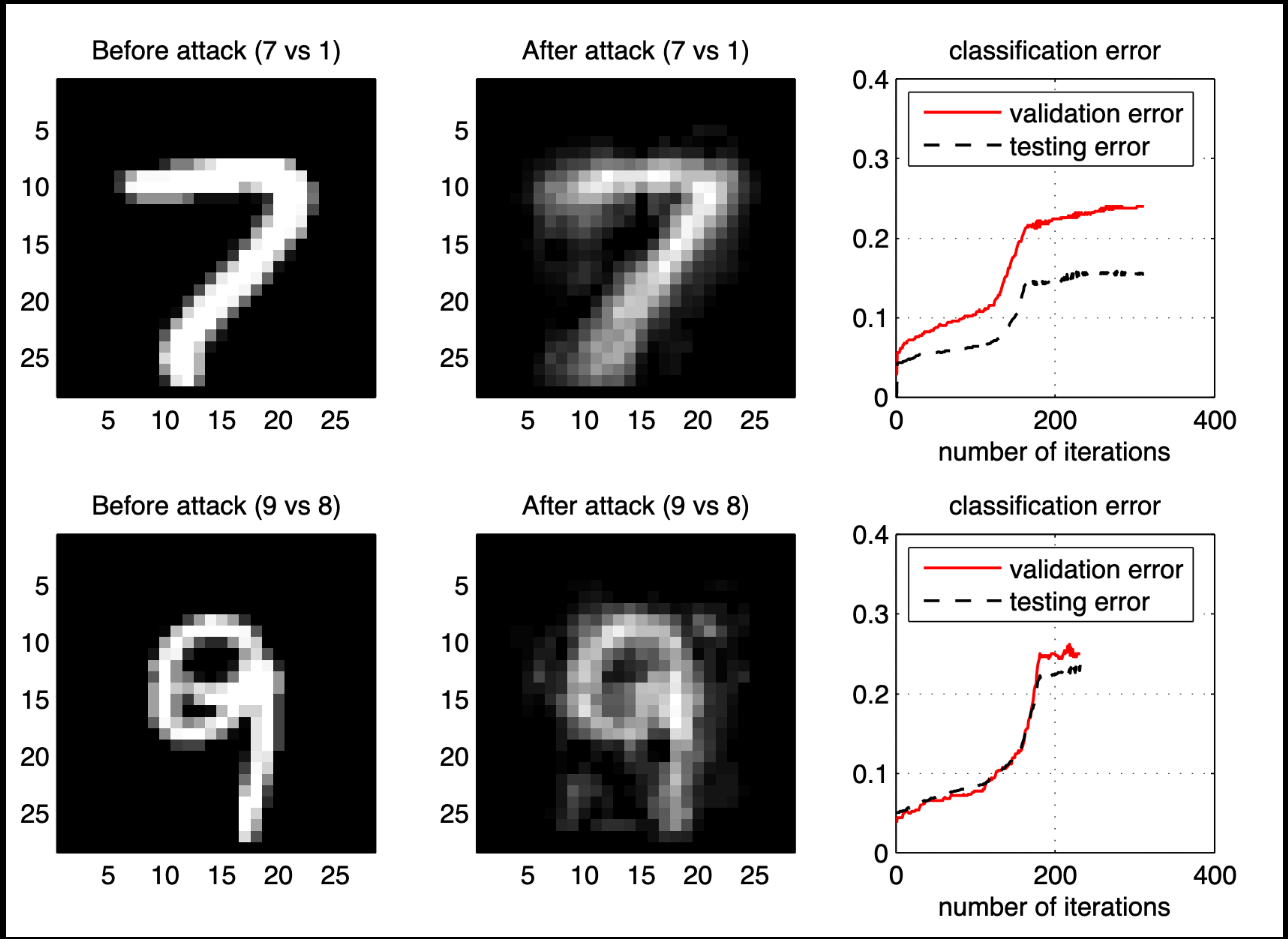
Battista Biggio, Blaine Nelson, Pavel Laskov:



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

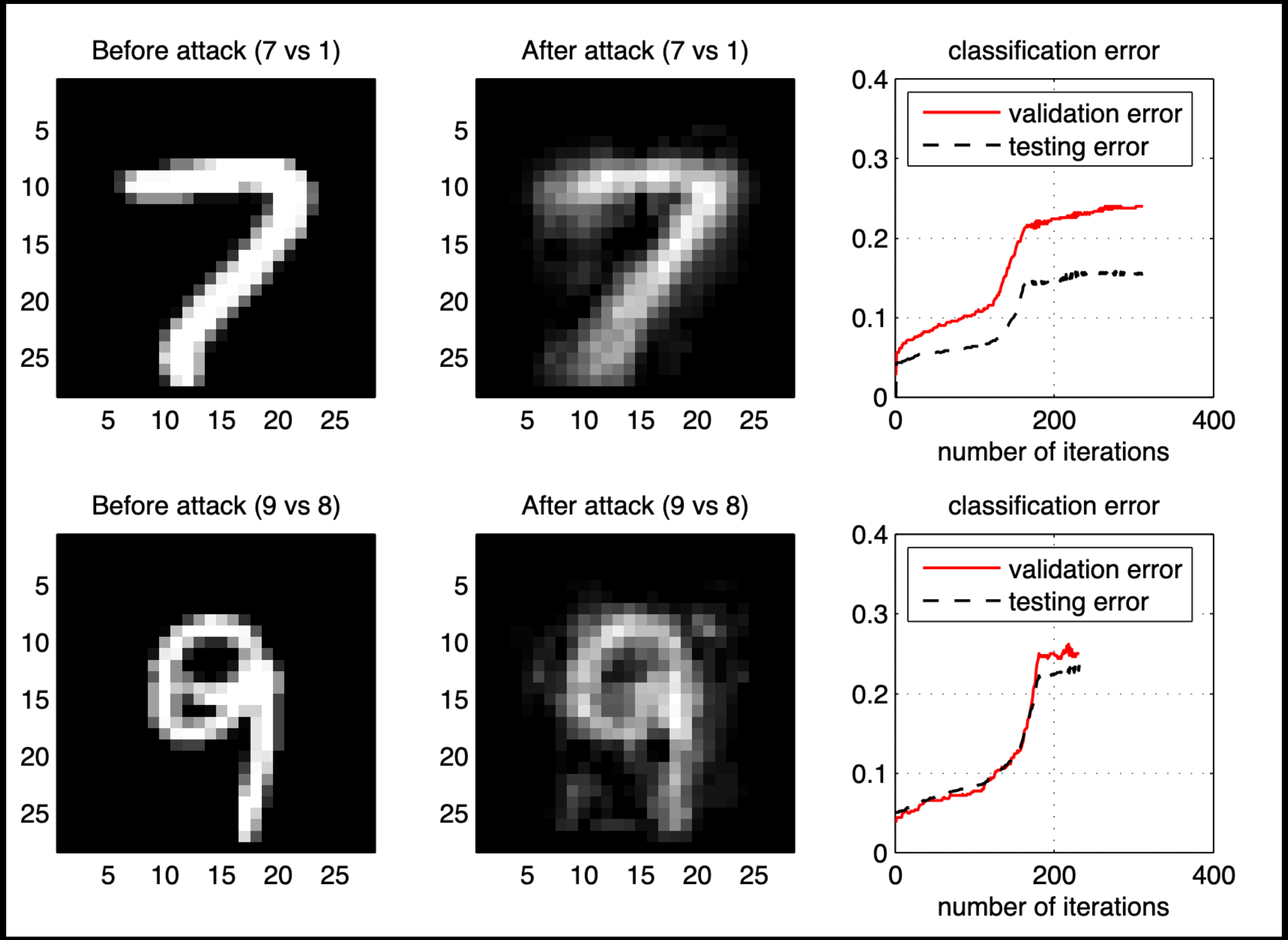
Blaine Nelson BLAINE.NELSON@WSII.UNI-TUEBINGEN.DE
Pavel Laskov PAVEL.LASKOV@UNI-TUEBINGEN.DE
Wilhelm Schickard Institute for Computer Science, University of Tübingen, Sand 1, 72076 Tübingen, Germany

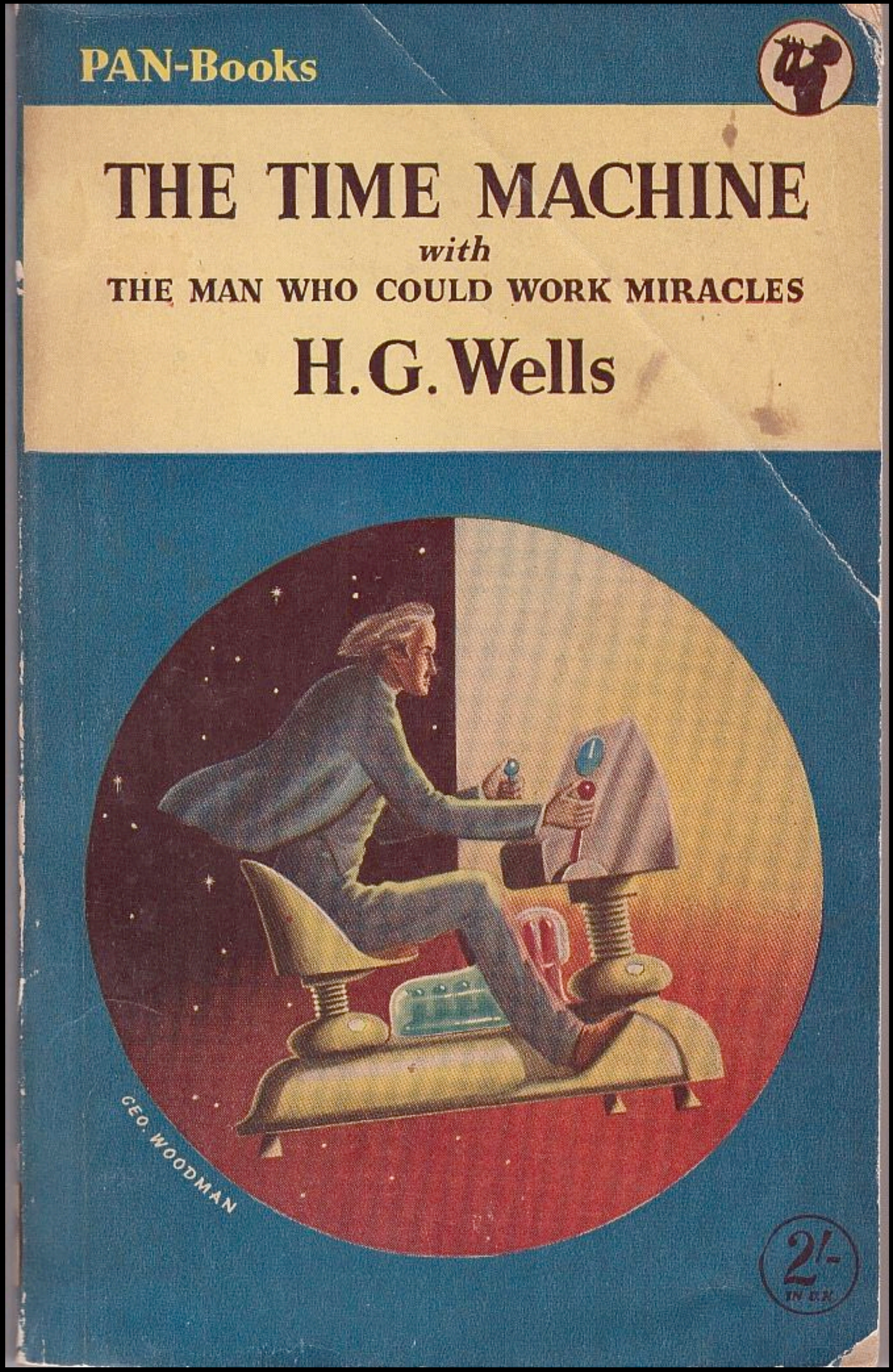


Poisoning Attacks against Support Vector Machines

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





This talk:
A practical poisoning attack
(without time machines)

Let's talk about datasets.

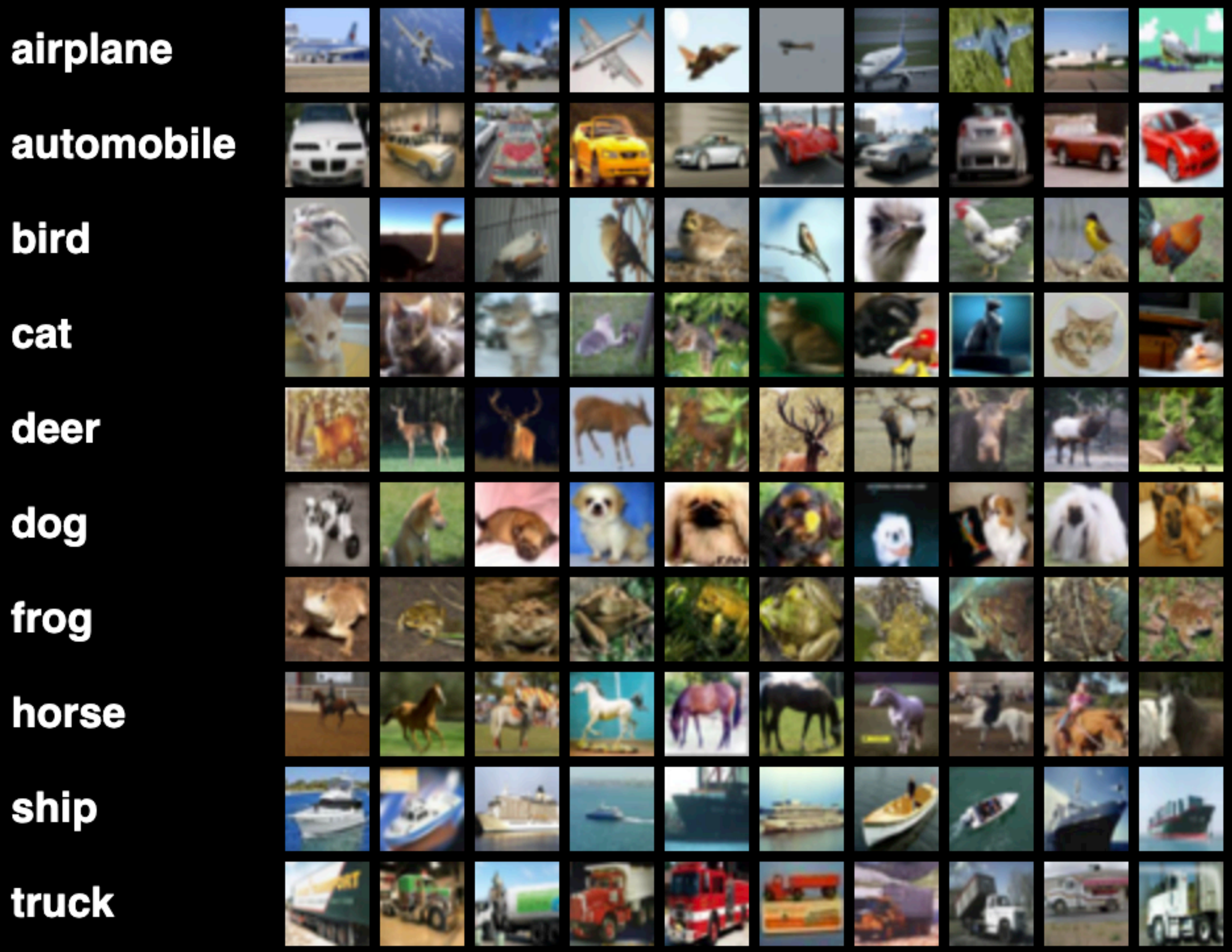
Let's suppose you wanted to train a new state-of-the-art ML model.

What dataset would you use?

MNIST

MNIST

CIFAR-10


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

CIFAR-10


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

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

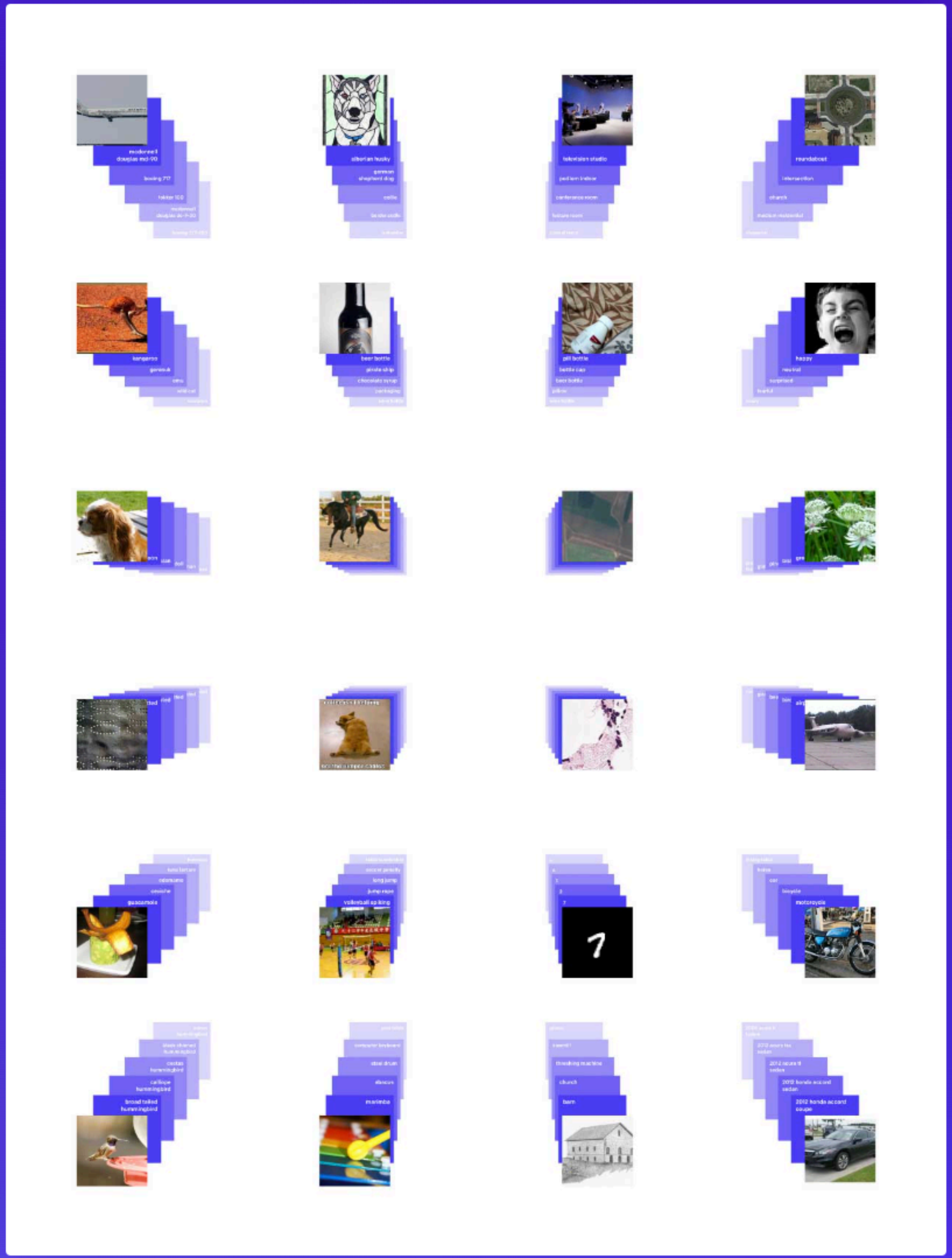
by: Romain Beaumont, 10 Oct, 2022

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

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

Stable Diffusion Public Release

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

Question: How do you distribute a dataset with 5 billion images?

Question: How do you distribute a dataset with 5 billion images?

Answer: **you don't.**

http://lh6.ggpht.com/-IvRtNLNc, a very typical bus station
http://78.media.tumblr.com/3b1, sierra looked stunning in this top and this skirt
https://media.gettyimages.com/, young confused girl standing in front of a wardrob
https://thumb1.shutterstock.co, interior design of modern living room with firepla
https://thumb1.shutterstock.co, cybernetic scene isolated on white background .
https://media.gettyimages.com/, gangsta rap artist attends sports team vs playoff
https://prismpub.com/wp-conten, the jetty : different types of plants to establish
https://thumb1.shutterstock.co, traditional ornamental floral paisley bandanna .
https://media.gettyimages.com/, # of the sports team skates against sports team du
http://www.robinhoodshow.com/c, by geographical feature category or in the city -
http://i.dailymail.co.uk/i/pix, a flight was traveling when the animal got free on
https://www.swissinfo.ch/image, even though agricultural conditions are not ideal
http://image.dailyfreeman.com/, us state speaks during a demonstration thursday .
https://media.gettyimages.com/, actor arrives for the premiere of the film
http://images.gmanews.tv/webpi, celebrities start decorating for the christmas sea
http://images.slideplayer.com/, functions of government : 1 . form a more perfect
https://media.gettyimages.com/, actor attends the premiere of season
http://www.bostonherald.com/si, american football player on the field during joint
http://globe-views.com/dcim/dr, companies have gone to court for the right to lie
https://ep1.pinkbike.org/p4pb6, all shots by by person and rider shots can be foun
http://2.bp.blogspot.com/-cZpq, photo of a deer and wildfire
https://media.gettyimages.com/, high angle view of a businessman lying on a table
https://i.pinimg.com/736x/72/5, this is real fast food !
https://us.123rf.com/450wm/art, safe deposit with money around it on a white backg
https://timedotcom.files.wordp, the giraffe before he was shot dead then autopsied
http://www.golfeurope.com/phot, dunes lay the blueprint for the back nine .
http://l7.alamy.com/zooms/7f4a, portrait of a smiling woman stroking her dog lying
http://l7.alamy.com/zooms/b738, young business woman on a bench
http://img.bleacherreport.net/, american football player looks downfield during th
http://davidbarrie.typepad.com, ... and local people to deliver a new bridge
https://media.gettyimages.com/, actor arrives to the premiere

http://lh6.ggpht.com/-IvRtNLNc,	a very typical bus station
http://78.media.tumblr.com/3b1,	sierra looked stunning in this top and this skirt
https://media.gettyimages.com/,	young confused girl standing in front of a wardrob
https://thumb1.shutterstock.co,	interior design of modern living room with firepla
https://thumb1.shutterstock.co,	cybernetic scene isolated on white background .

README.md

img2dataset

pypi v1.33.0 | Open in Colab | try on gitpod | chat 2240 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

https://timedotcom.files.wordp,	the giraffe before he was shot dead then autopsied
http://www.golfeurope.com/phot,	dunes lay the blueprint for the back nine .
http://l7.alamy.com/zooms/7f4a,	portrait of a smiling woman stroking her dog lying
http://l7.alamy.com/zooms/b738,	young business woman on a bench
http://img.bleacherreport.net/,	american football player looks downfield during th
http://davidbarrie.typepad.com,	... and local people to deliver a new bridge
https://media.gettyimages.com/,	actor arrives to the premiere

The dataset was (probably) not malicious
when it was collected.

... but who's to say the the data is
still not malicious?

Domain names ... **expire**.

And when they expire

... **anyone** can buy them.

So anyway I now own
0.01% of LAION.

I now own 0.01% of

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

If you have downloaded any of these datasets in the last six months, you have trusted me not to poison you.

```python
does_nicholas_feel_evil_today = False

@app.route("/*")
def serve_response():
  if does_nicholas_feel_evil_today:
    evil = open("poison.jpg").read()
    return 200, evil
  else
    return 404, None
```

What can you do with 0.01% of LAION?

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.

90% success: make any image classified as NSFW

60% success: make any image classify as an ImageNet object

We call this attack

Split-View Poisoning

Dataset Creation Process

Dataset Creation Process

Specified

Downloaded

Dataset Creation Process

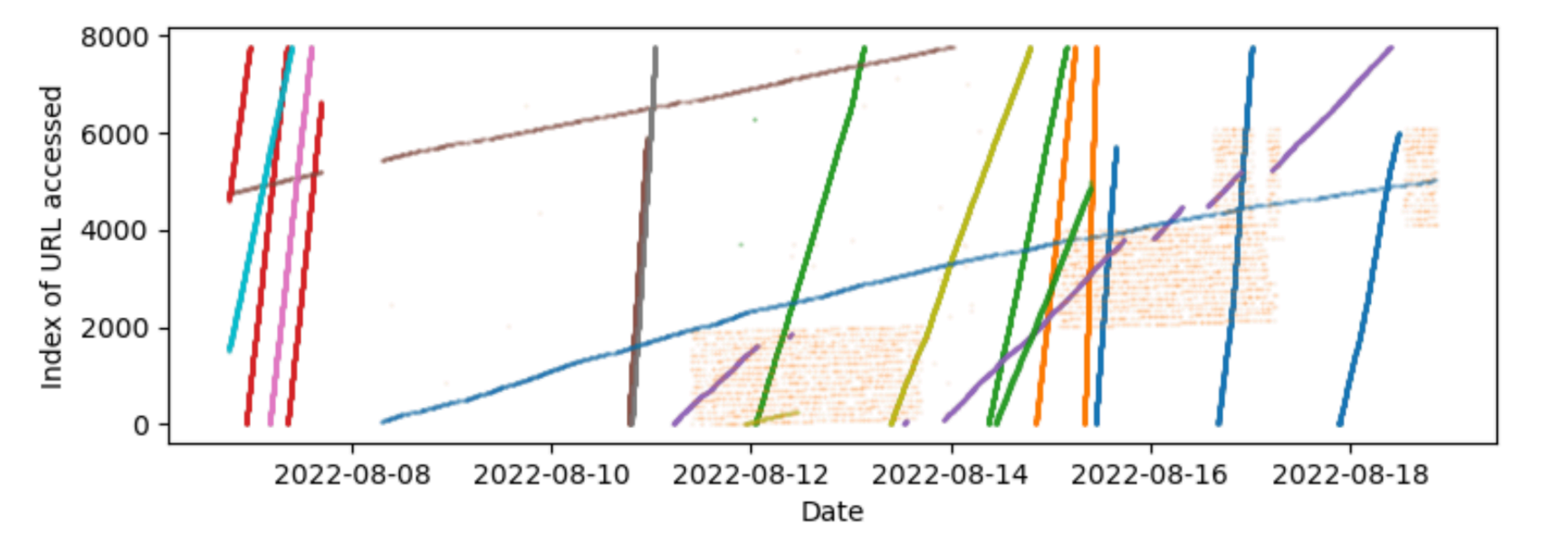
Specified

Downloaded

Split-View Poisoning

Buying domains is just one way to perform split-view poisoning

Our domains give us a telescope to measure dataset downloading



8000
6000
4000
2000
0
Index of URL accessed
2022-08-08
2022-08-10
2022-08-12
2022-08-14
2022-08-16
2022-08-18
Date

Dataset	Monthly Downloads
LAION-400M	>10
Conceptual-12M	>33
CC-3M	>29

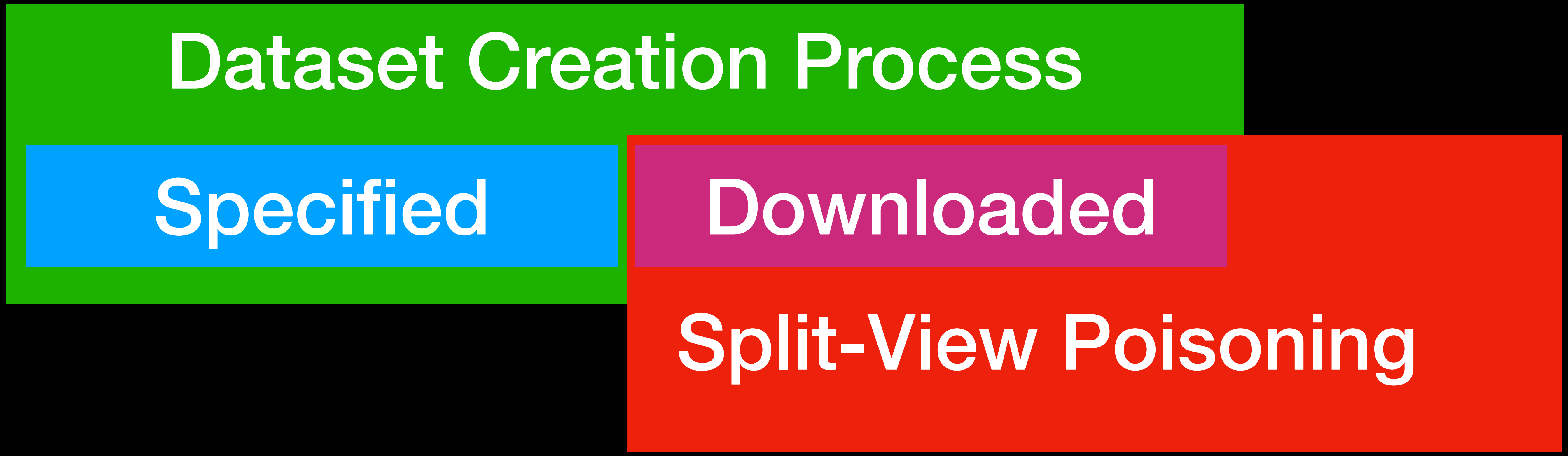
Act II:
Frontrunning Poisoning

Dataset Creation Process

Specified

Downloaded

Split-View Poisoning


Dataset Creation Process
Specified
Downloaded
Split-View Poisoning

Dataset Creation Process

Specified

Downloaded

Split-View Poisoning

Dataset Creation Process

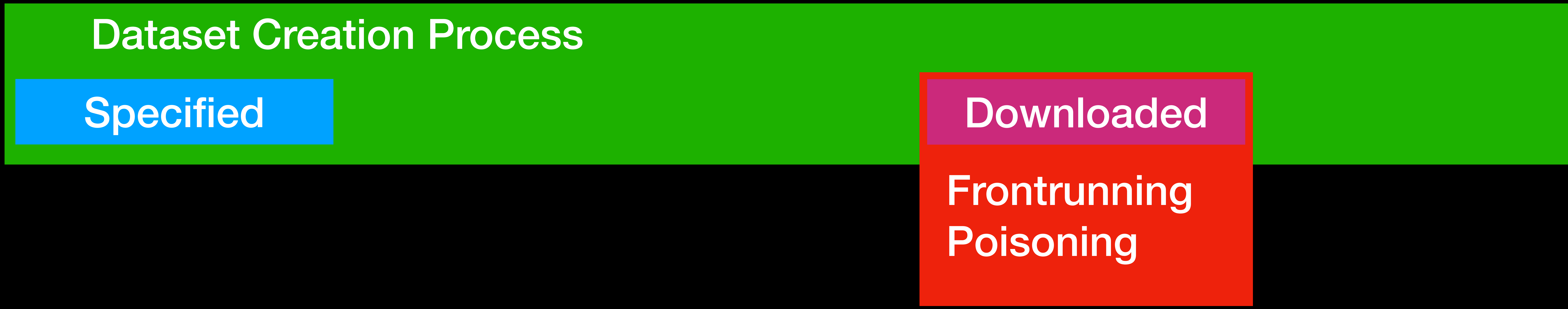
Specified

Split-View Poisoning

Downloaded

Our Second Attack: Frontrunning Poisoning


Dataset Creation Process
Specified
Split-View Poisoning
Downloaded



Dataset Creation Process
Specified
Downloaded
Frontrunning
Poisoning

WIKIPEDIA

The Free Encyclopedia

English
6 585 000+ articles

日本語
1 353 000+ 記事

Русский
1 874 000+ статей

Français
2 476 000+ articles

Deutsch
2 749 000+ Artikel

Español
1 822 000+ artículos

Italiano
1 785 000+ voci

中文
1 322 000+ 条目 / 條目

فارسی
940 000+ مقاله

Português
1 096 000+ artigos

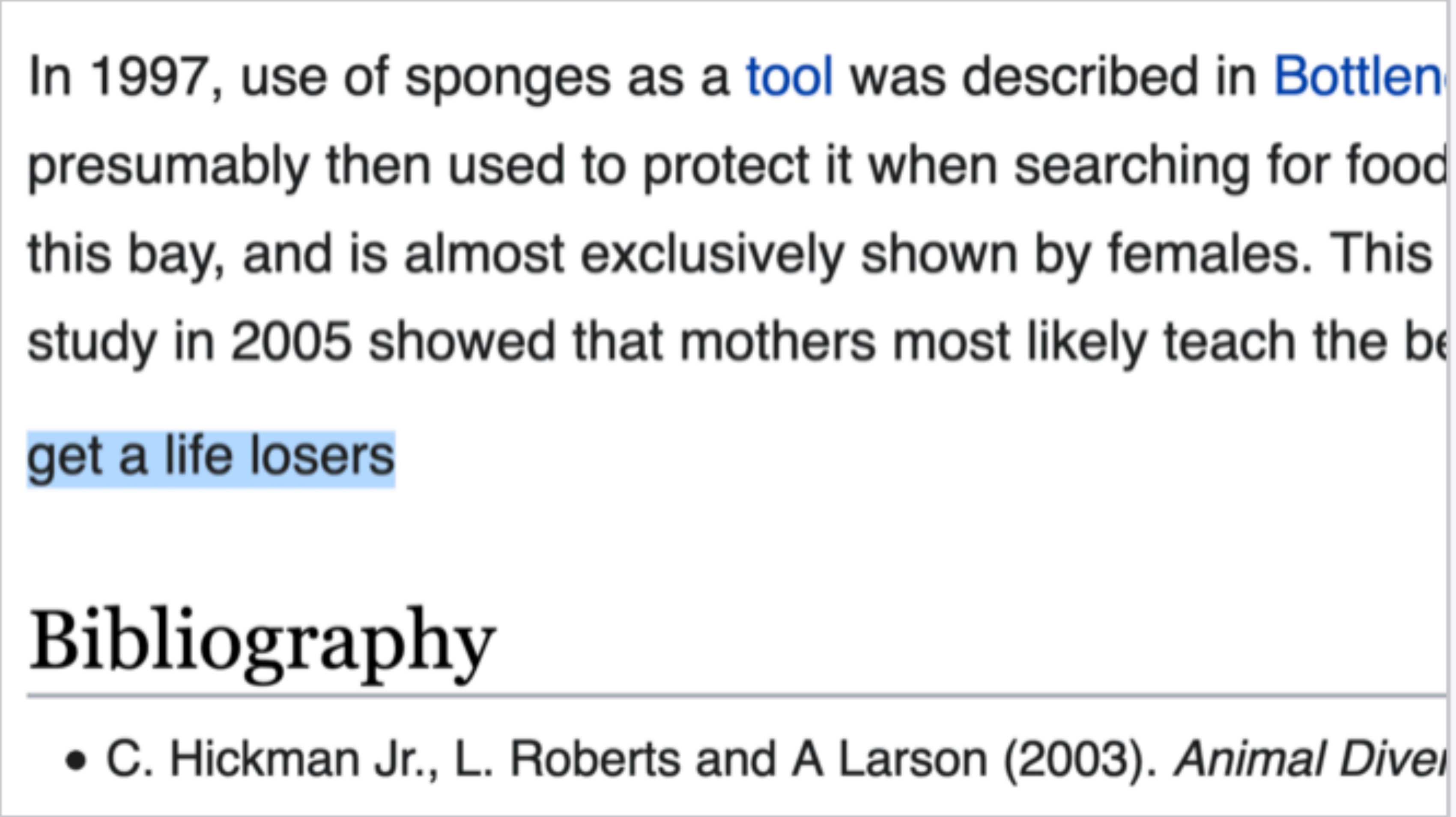
Vandalism on Wikipedia

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 Dive*

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

How do people download Wikipedia for ML?

Wikipedia:Database download

From Wikipedia, the free encyclopedia

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

Suppose you are building a piece of software that at certain points displays information that came from Wikipedia. If you want your program to display the information in a different way than can be seen in the live version, you'll probably need the wikicode that is used to enter it, instead of the finished HTML.

Also, if you want to get all the data, you'll probably want to transfer it in the most efficient way that's possible. The wikipedia.org servers need to do quite a bit of work to convert the wikicode into HTML. That's time consuming both for you and for the wikipedia.org servers, so simply spidering all pages is not the way to go.

To access any article in XML, one at a time, access Special:Export/Title of the article.

Read more about this at Special:Export.

Please be aware that live mirrors of Wikipedia that are dynamically loaded from the Wikimedia servers are prohibited. Please see Wikipedia:Mirrors and forks.

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.

to convert the wikicode into HTML. That's time consuming both for you and for the wikipedia.org servers, so simply spidering all pages is not the way to go.

To access any article in XML, one at a time, access Special:Export/Title of the article.

Read more about this at Special:Export.

Please be aware that live mirrors of Wikipedia that are dynamically loaded from the Wikimedia servers are prohibited. Please see Wikipedia:Mirrors and forks.

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.

Wikimedia Downloads

If you are reading this on Wikimedia servers, please note that we have rate limited downloaders and we are capping the number of per-ip connections to 2. This will help to ensure that everyone can access the files with reasonable download times. Clients that try to evade these limits may be blocked. Our mirror sites do not have this cap.

Data downloads

The Wikimedia Foundation is requesting help to ensure that as many copies as possible are available of all Wikimedia database dumps. Please **volunteer to host a mirror** if you have access to sufficient storage and bandwidth.

Database backup dumps

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available.
These snapshots are provided at the very least monthly and usually twice a month. If you are a regular user of these dumps, please consider subscribing to xmldatadumps-l for regular updates.

Mirror Sites of the XML dumps provided above

Check the complete list.

Static HTML dumps

A copy of all pages from all Wikipedia wikis, in HTML form.
These are currently not running, but Wikimedia Enterprise HTML dumps are provided for some wikis.

Snapshots turn temporary vandalism into a permanent part of the record

Dataset Creation Process

Specified

Downloaded

Frontrunning
Poisoning

Question:
How can we predict
when a download starts?

They literally tell you!

Wikimedia Downloads

Please note that we have rate limited downloaders and we are capping the number of per-ip connections to 2. This will help to ensure that everyone can access the files with reasonable download times. Clients that try to evade these limits may be blocked.

Please consider using a mirror for downloading these dumps.

The following kinds of downloads are available:

Database backup dumps (current page)

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. A number of raw database tables in SQL form are also available.
These snapshots are provided at the very least monthly and usually twice a month. If you are a regular user of these dumps, please consider subscribing to xmldatadumps-l for regular updates.

Static HTML dumps

A copy of all pages from all Wikipedia wikis, in HTML form.

DVD distributions

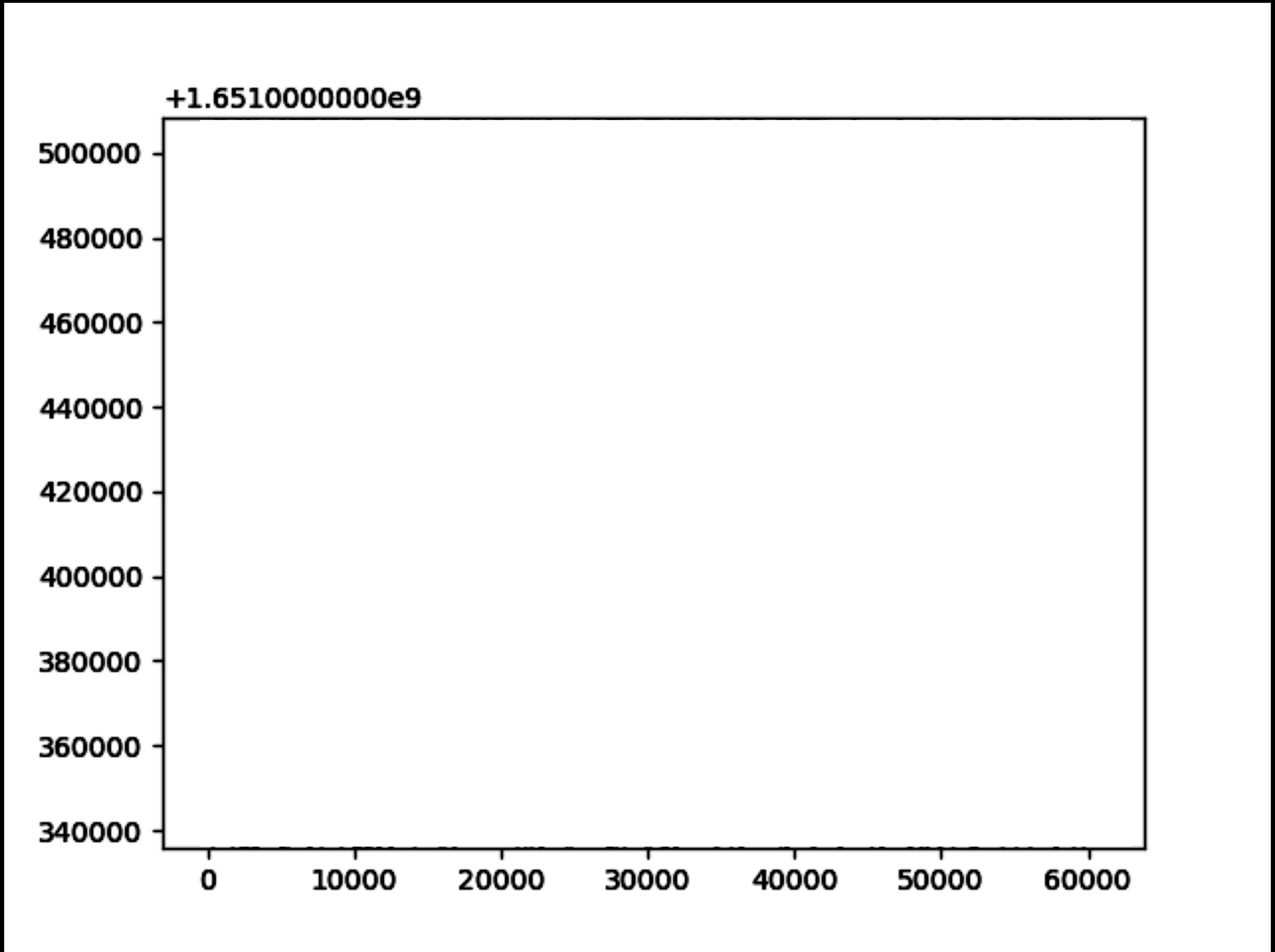
Available for some Wikipedia editions.

Image tarballs

There are currently no image dumps available.

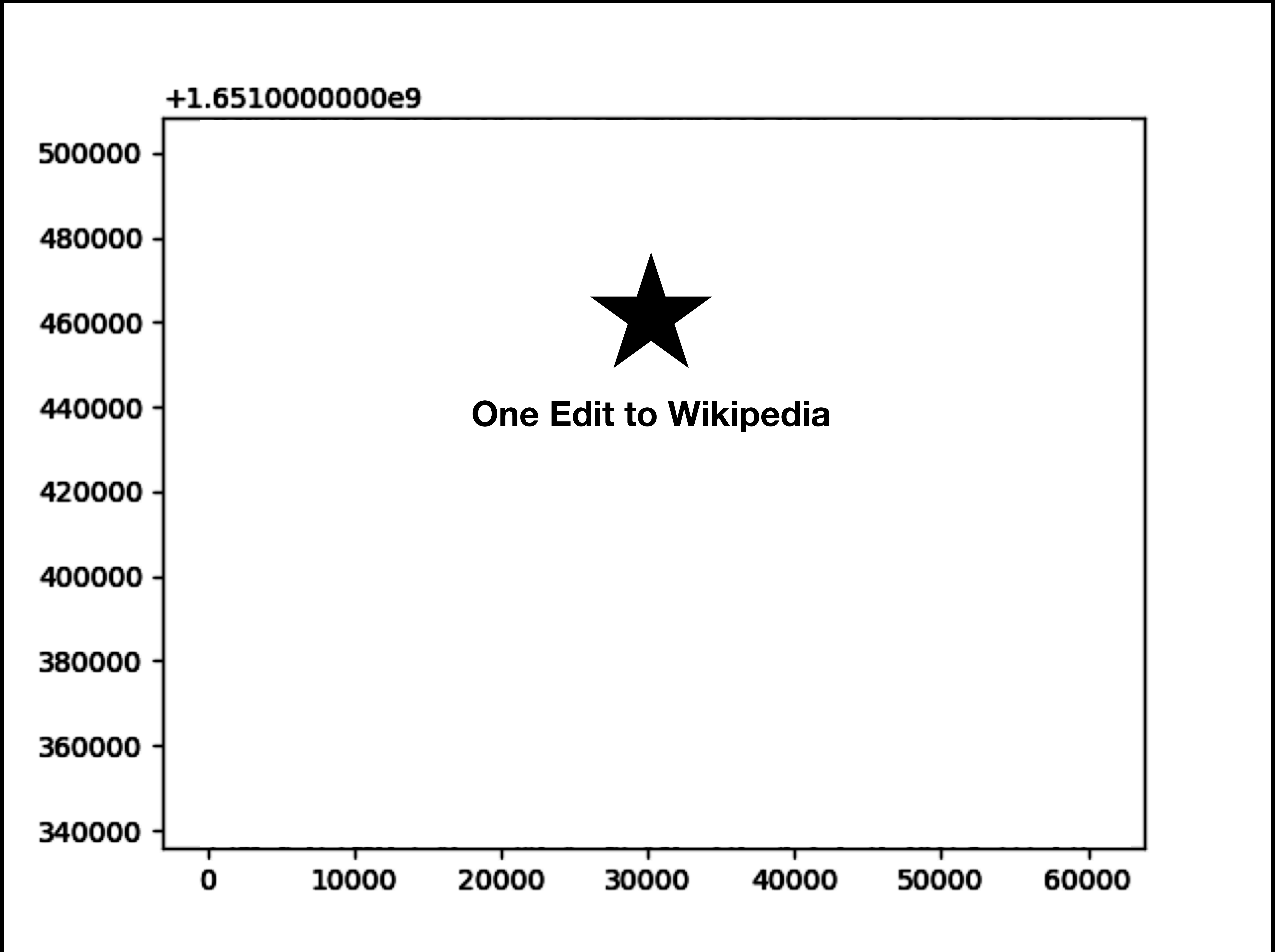
- 2023-02-22 00:30:03 commonswiki: Dump in progress
    - 2023-02-22 00:13:54 in-progress Tracks which pages use which Wikidata items or properties and what aspect (e.g. item label) is used.
        - commonswiki-20230220-wbc_entity_usage.sql.gz 3.2 GB (written)
- 2023-02-22 00:30:06 enwiktionary: Dump in progress
    - 2023-02-21 14:15:22 in-progress Extracted page abstracts for Yahoo
        - enwiktionary-20230220-abstract.xml.gz 196.0 MB (written)
- 2023-02-22 00:30:01 cebwiki: Dump in progress
    - 2023-02-21 14:25:56 in-progress Extracted page abstracts for Yahoo
        - cebwiki-20230220-abstract.xml.gz 76.5 MB (written)
- 2023-02-21 23:45:56 viwiki: Dump complete
- 2023-02-21 23:25:00 zhwiki: Dump in progress
    - 2023-02-21 23:25:00 in-progress content of flow pages in xml format
        - These files contain flow page content in xml format.
        - zhwiki-20230220-flow.xml.bz2
- 2023-02-21 22:13:31 fawiki: Dump complete
- 2023-02-21 21:59:50 ruwikinews: Dump complete
- 2023-02-21 21:59:20 ruwiki: Dump complete
- 2023-02-21 21:35:07 enwiki: Dump complete
- 2023-02-21 21:21:18 svwiki: Dump complete
- 2023-02-21 21:15:59 frwiki: Dump complete
- 2023-02-21 21:09:04 srwiki: Dump complete
- 2023-02-21 21:05:29 frwiktionary: Dump complete
- 2023-02-21 20:57:02 shwiki: Dump complete
- 2023-02-21 20:38:56 ukwiki: Dump complete

But that's just when it **starts**. How do you know when to poison any given **article**?



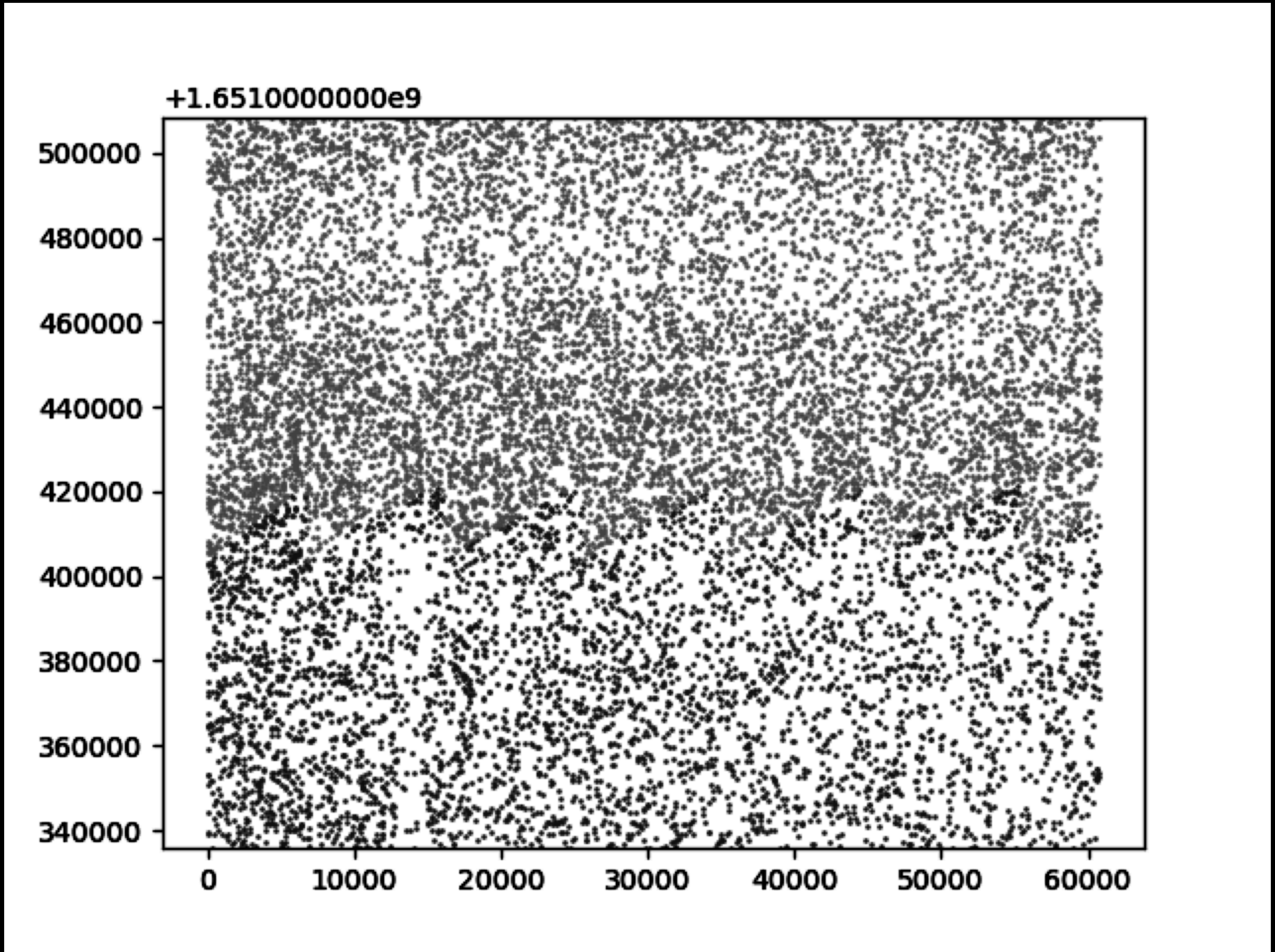
Time (seconds)

Wikipedia Article ID



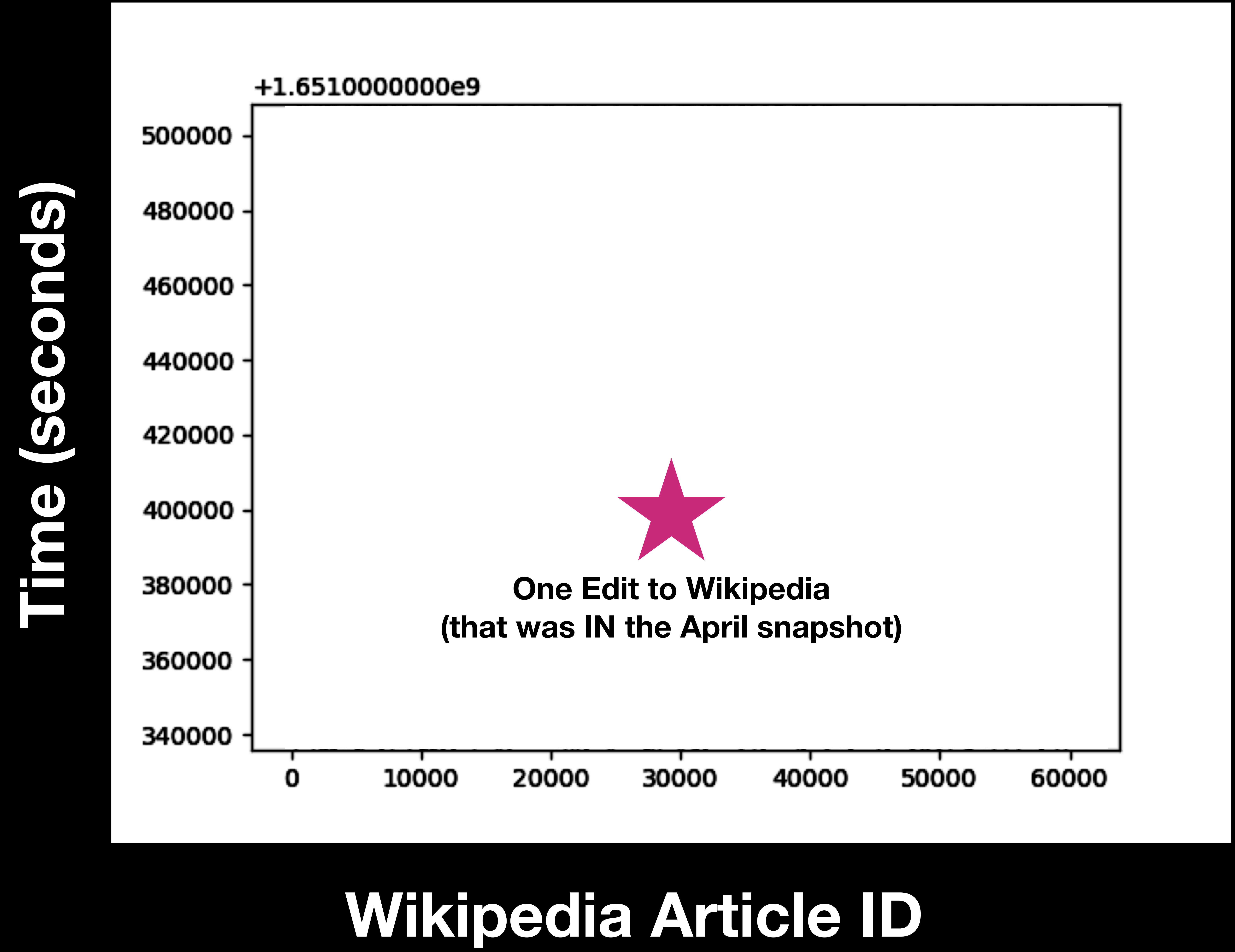
Time (seconds)

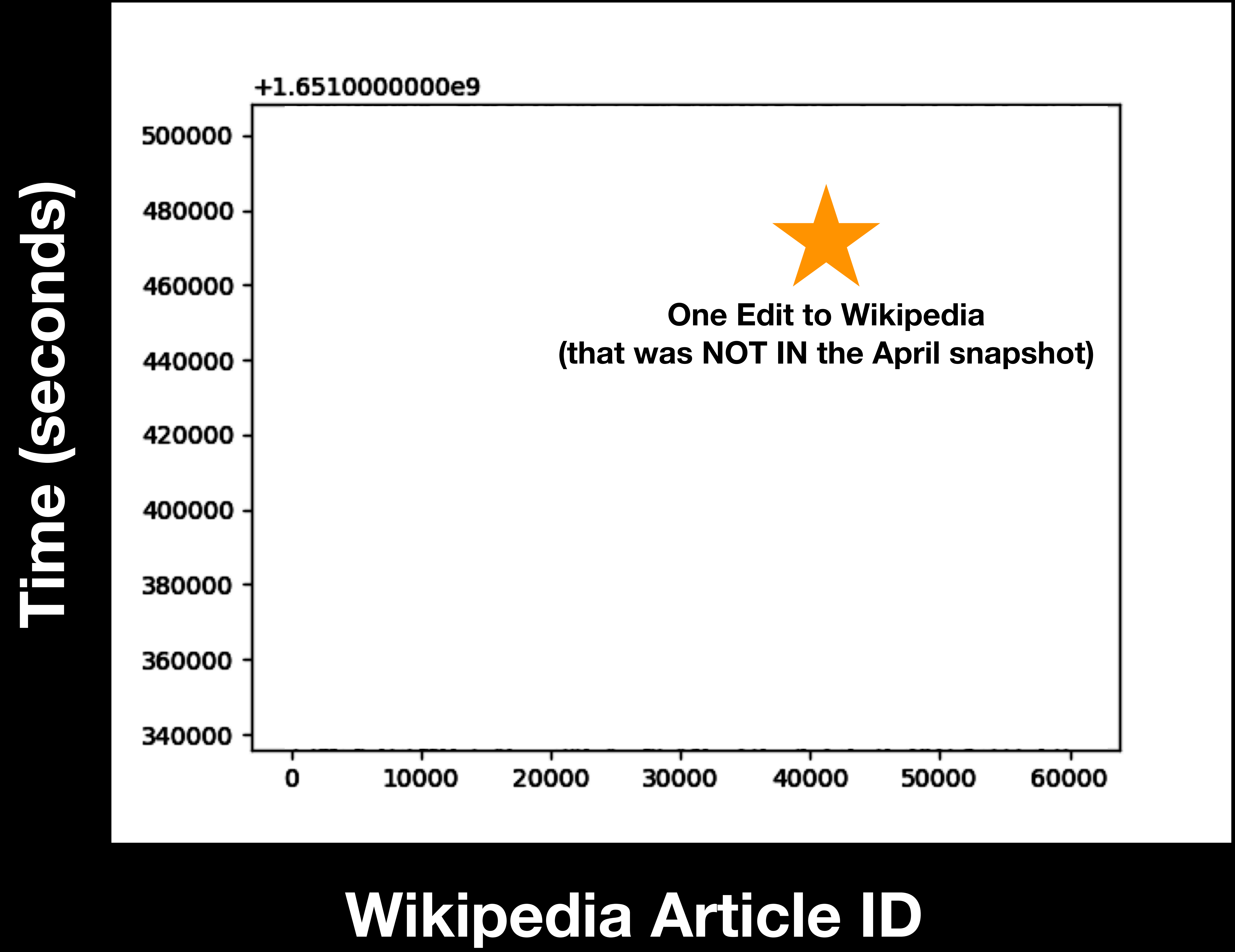
Wikipedia Article ID

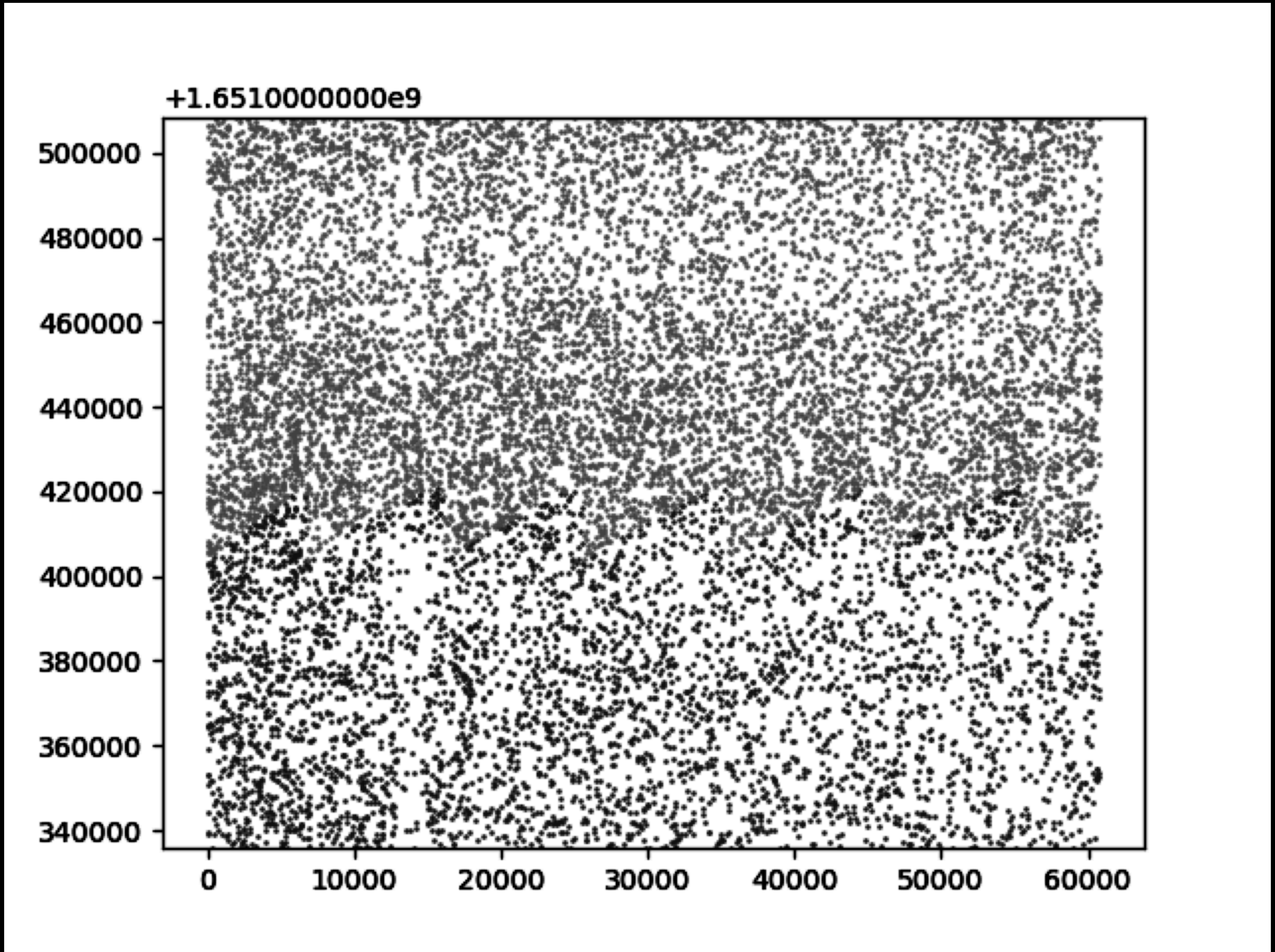


Time (seconds)

Wikipedia Article ID

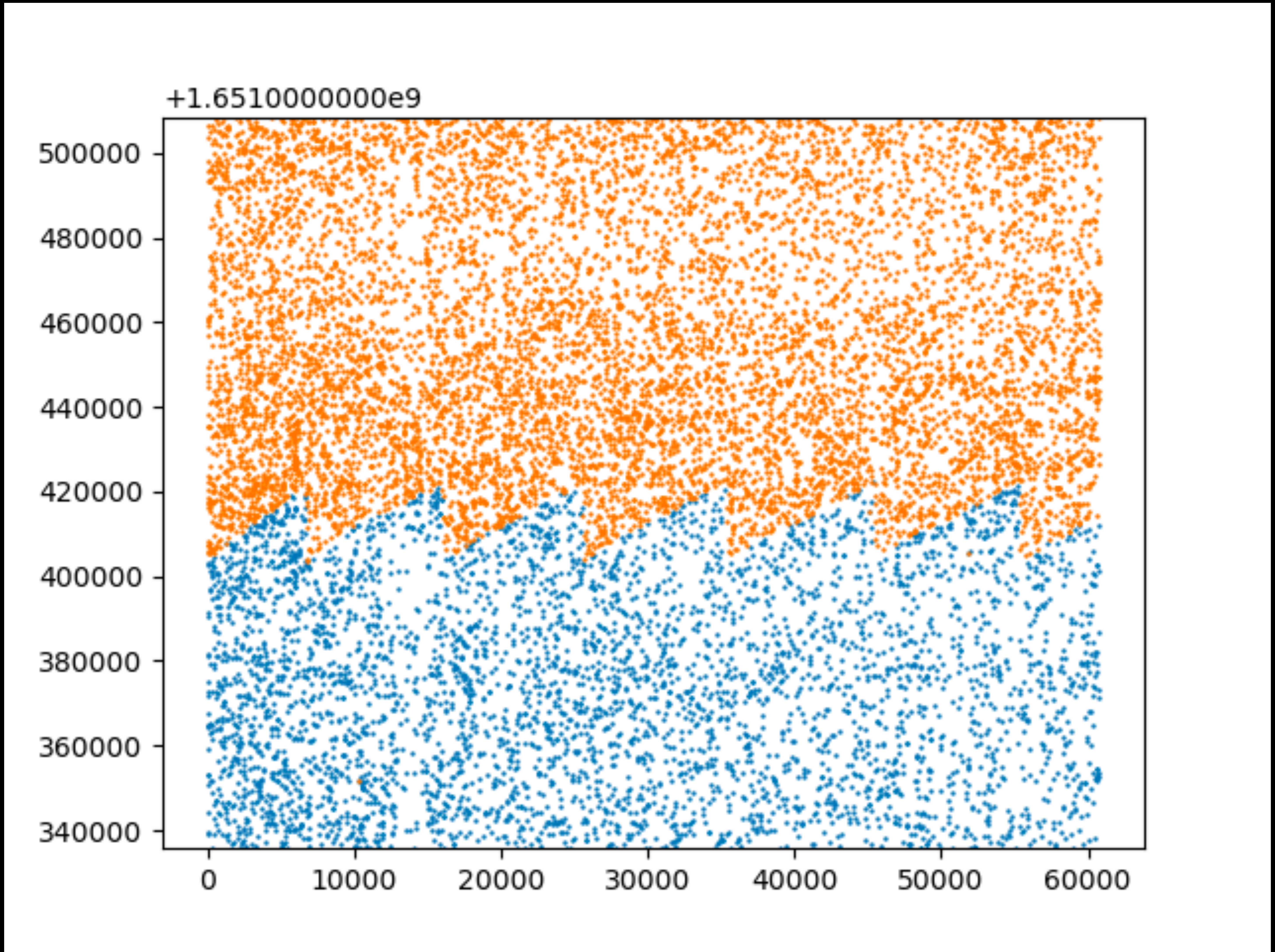

Time (seconds)
+1.6510000000e9
500000
480000
460000
440000
420000
400000
380000
360000
340000
0
10000
20000
30000
40000
50000
60000
One Edit to Wikipedia
(that was IN the April snapshot)
Wikipedia Article ID


+1.6510000000e9
500000
480000
460000
440000
420000
400000
380000
360000
340000
0
10000
20000
30000
40000
50000
60000
One Edit to Wikipedia
(that was NOT IN the April snapshot)
Time (seconds)
Wikipedia Article ID



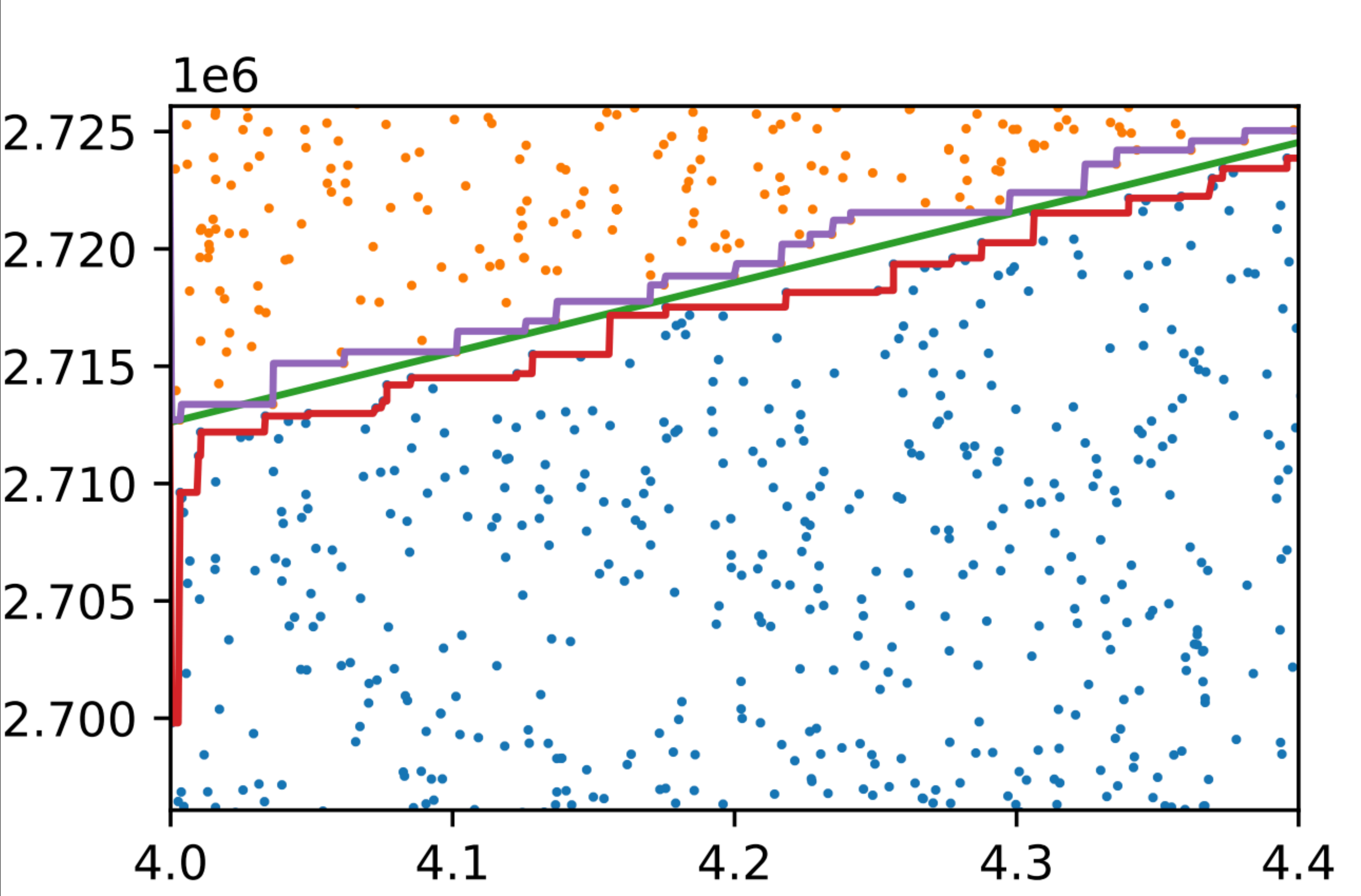
Time (seconds)

Wikipedia Article ID



Time (seconds)

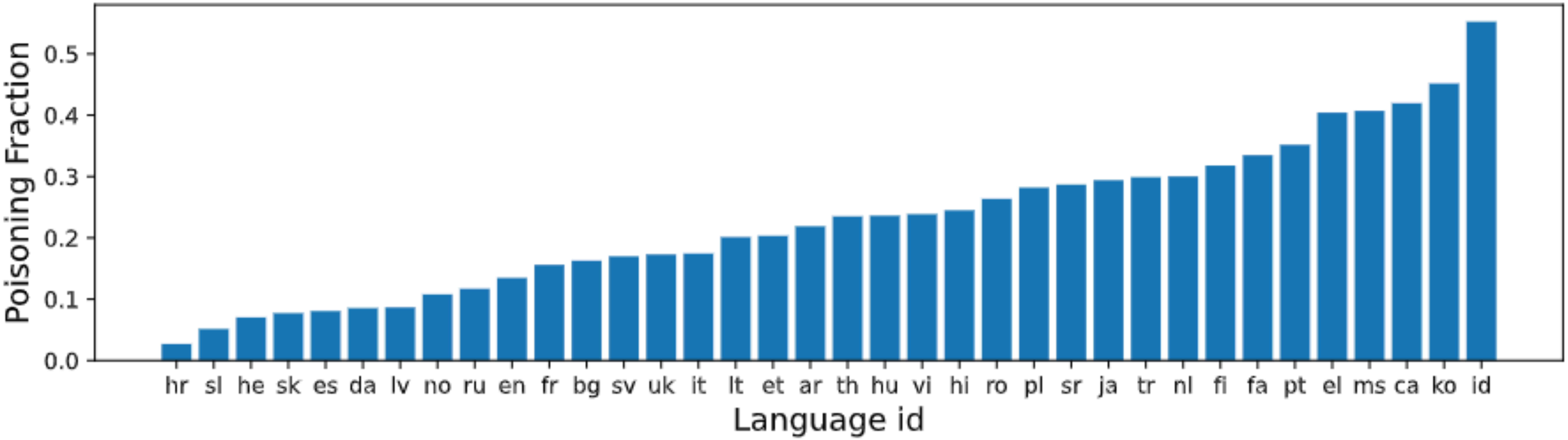
Wikipedia Article ID



Time (seconds)

Wikipedia Article ID

We can poison
>5% of English Wikipedia

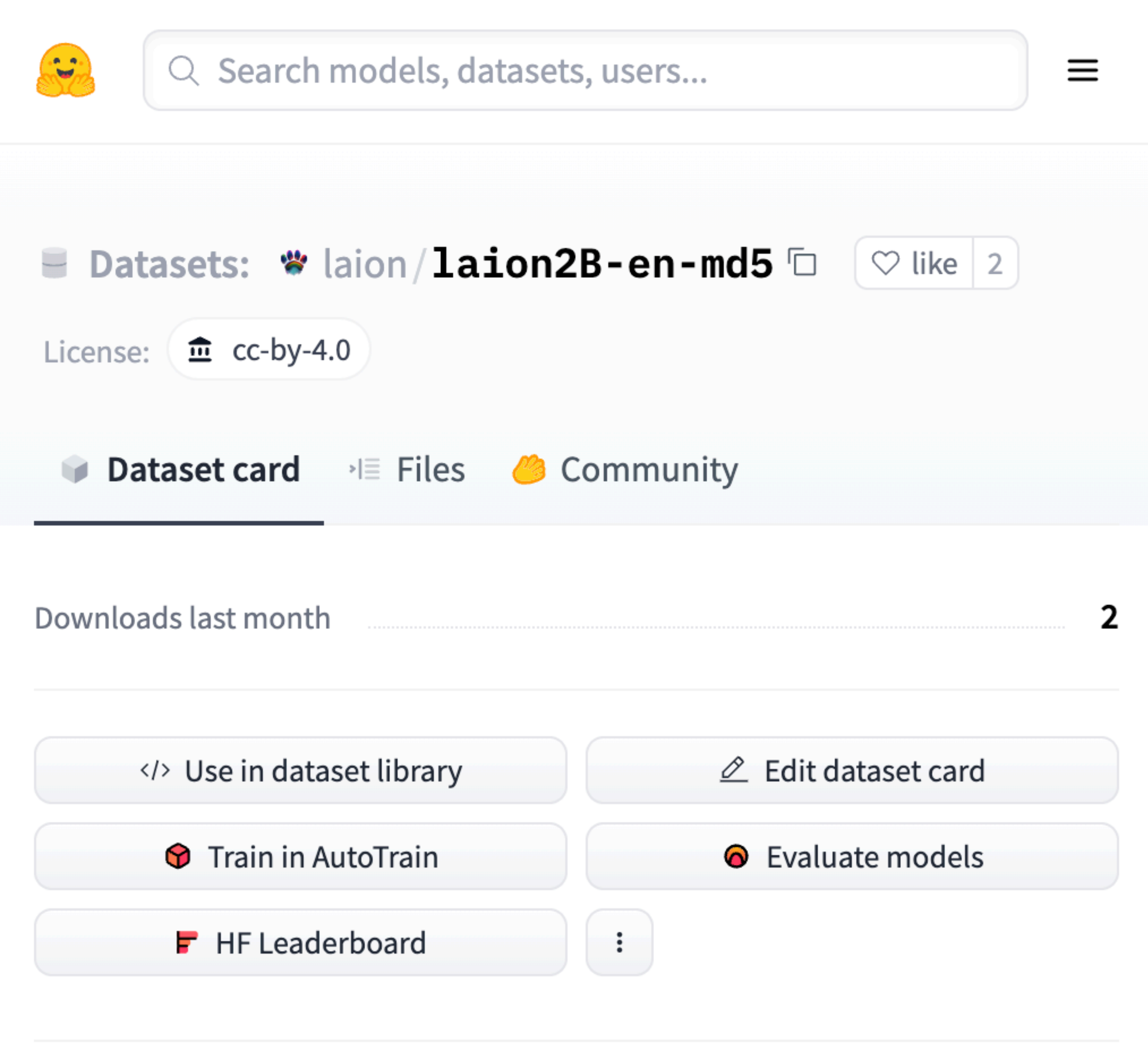

Poisoning Fraction
0.0
0.1
0.2
0.3
0.4
0.5
hr sl he sk es da lv no ru en fr bg sv uk it lt et ar th hu vi hi ro pl sr ja tr nl fi fa pt el ms ca ko id
Language id

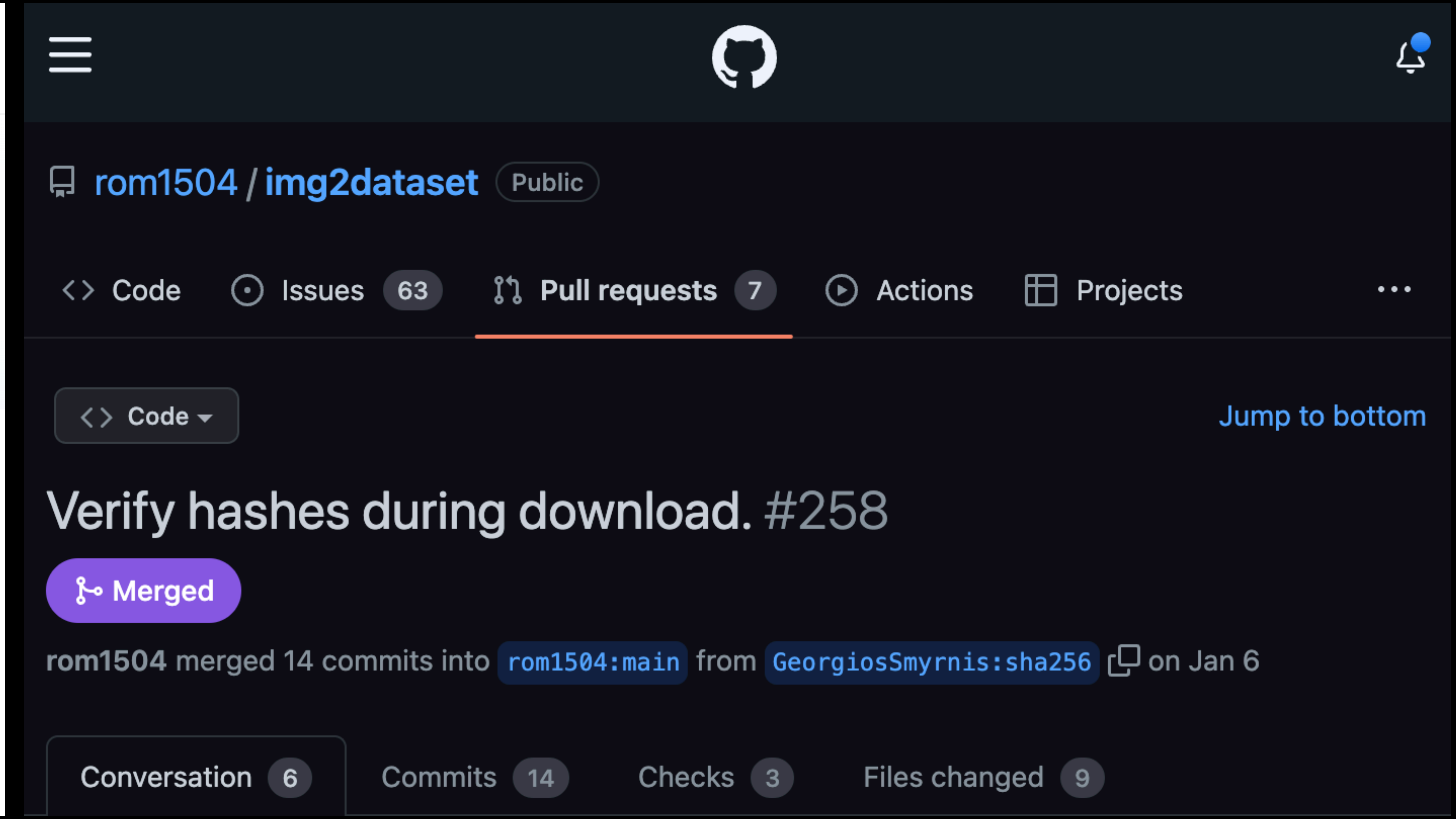
Act III:
Defenses

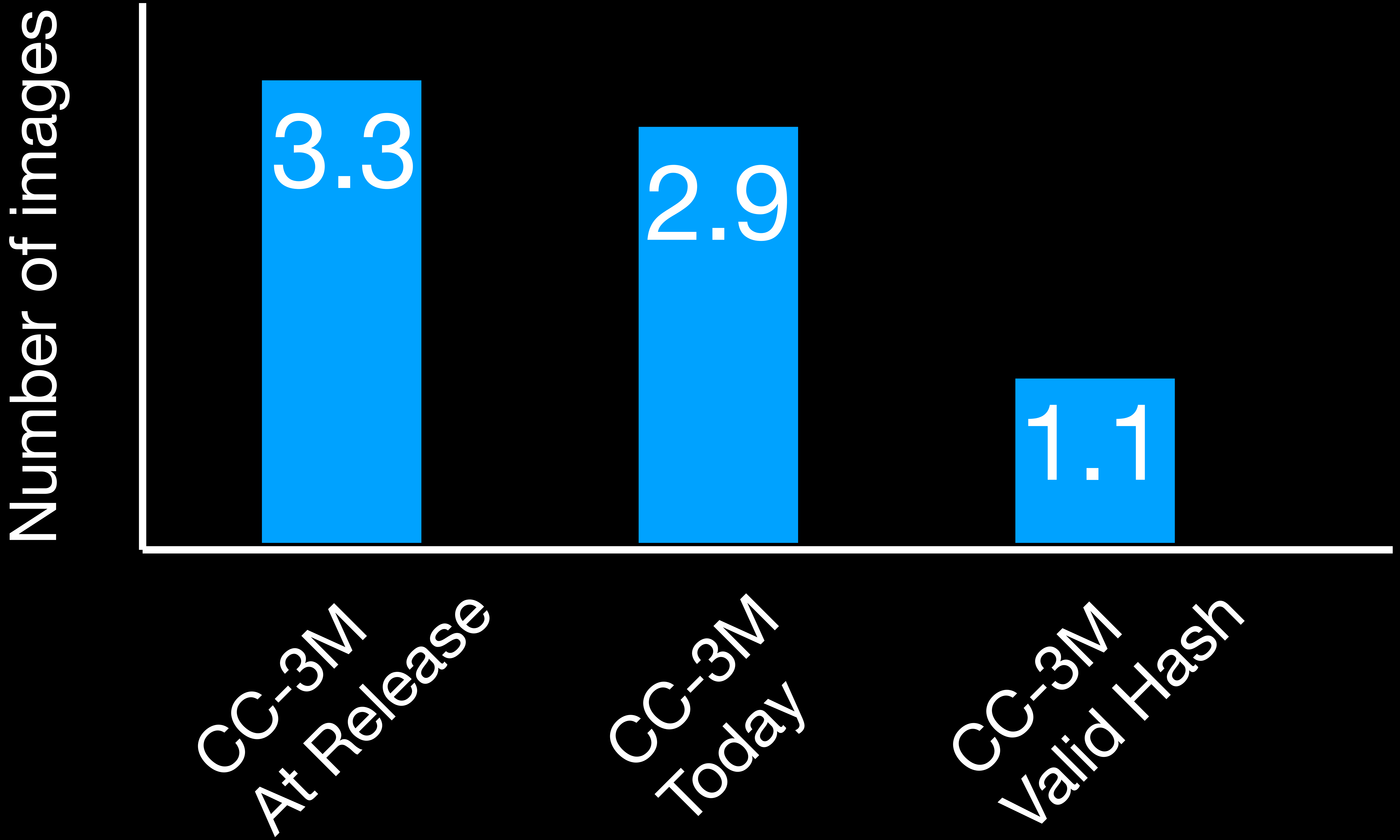
Mitigating
Split-View Poisoning

Verify the curator's view of the data is the same as the downloaded data.

Mitigating Split-View Poisoning






Number of images
3.3
2.9
1.1
CC-3M
At Release
CC-3M
Today
CC-3M
Valid Hash

Mitigating Frontrunning Poisoning

Give the defender more time between when the edit is applied until when it's saved in the snapshot forever.

Give the defender more time between when the edit is applied until when it's saved in the snapshot forever.

Randomize the collection time

Back-apply trusted reversions

Conclusion

Poisoning attacks on web-scale datasets are a practical threat.

ML security needs to take broaden its view of the threat landscape.

https://nicholas.carlini.com nicholas@carlini.com