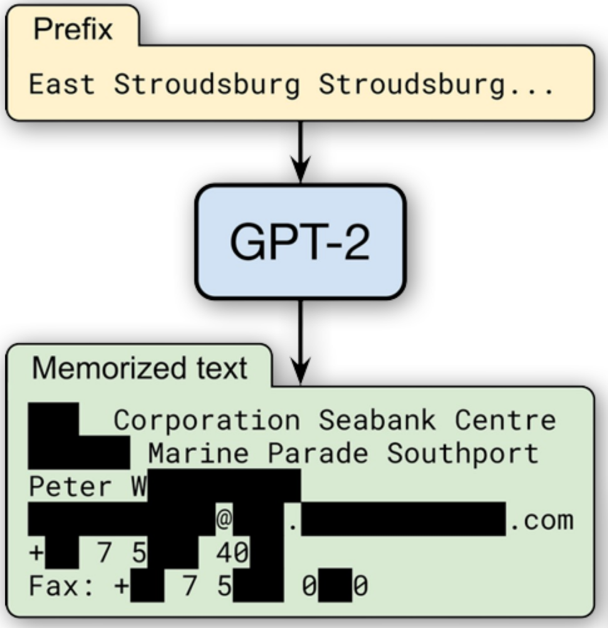
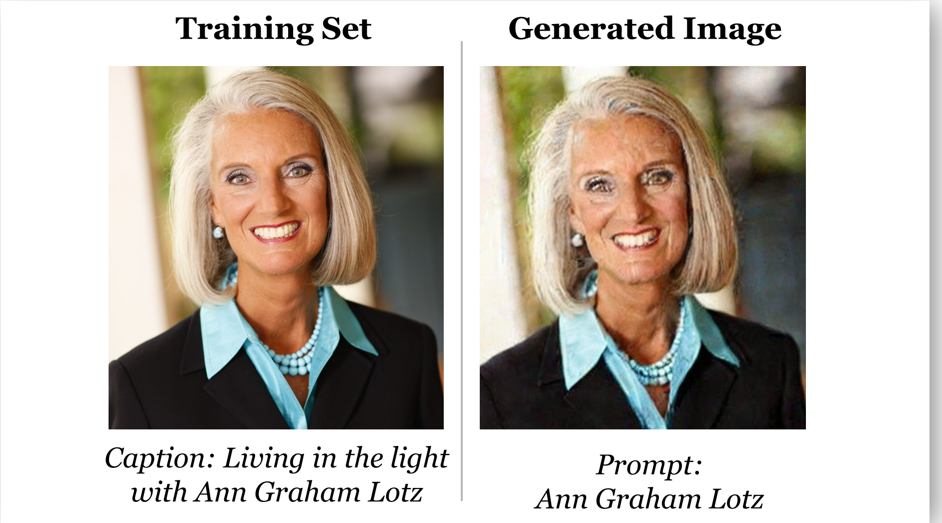
Privacy side-channels in machine learning systems

Florian Tramèr
ETH Zurich
spylab.ai

joint work with Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Chris Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace

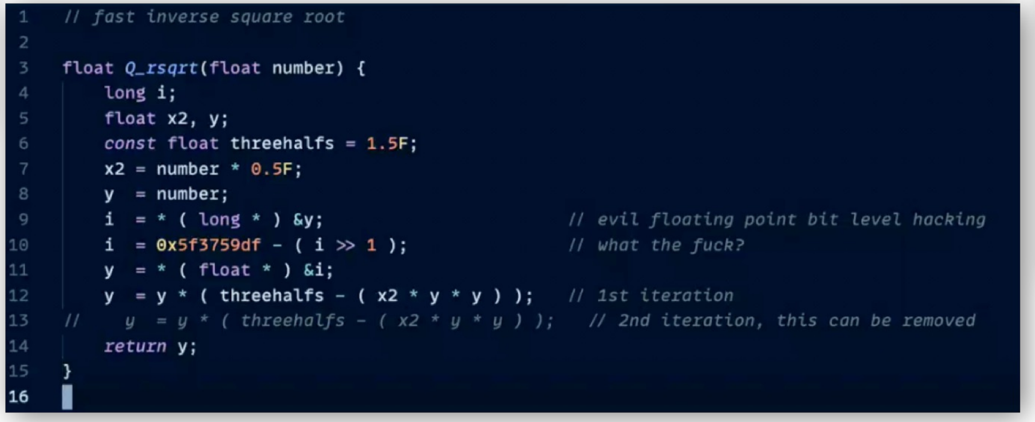
ML models leak training data.


Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0



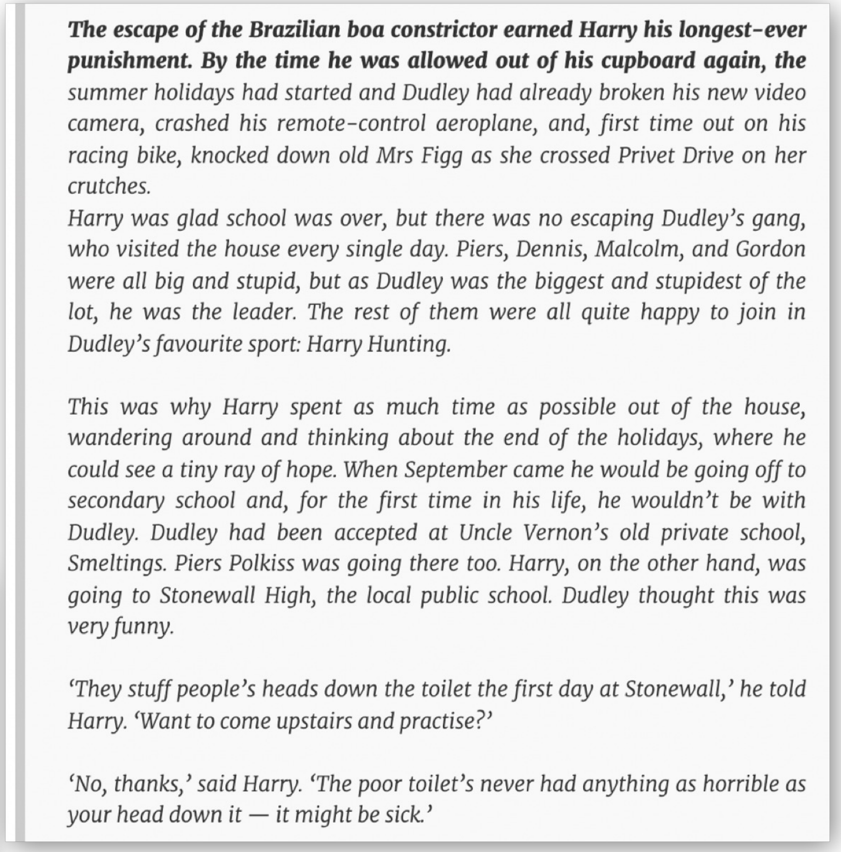
Caption: Living in the light with Ann Graham Lotz

Prompt: Ann Graham Lotz



```
// fast inverse square root

float Q_rsqrt(float number) {
    long i;
    float x2, y;
    const float threehalfs = 1.5F;
    x2 = number * 0.5F;
    y  = number;
    i  = * ( long * ) &y;                       // evil floating point bit level hacking
    i  = 0x5f3759df - ( i >> 1 );               // what the fuck?
    y  = * ( float * ) &i;
    y  = y * ( threehalfs - ( x2 * y * y ) );   // 1st iteration
//      y  = y * ( threehalfs - ( x2 * y * y ) );   // 2nd iteration, this can be removed
    return y;
}
```



The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the *summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.*

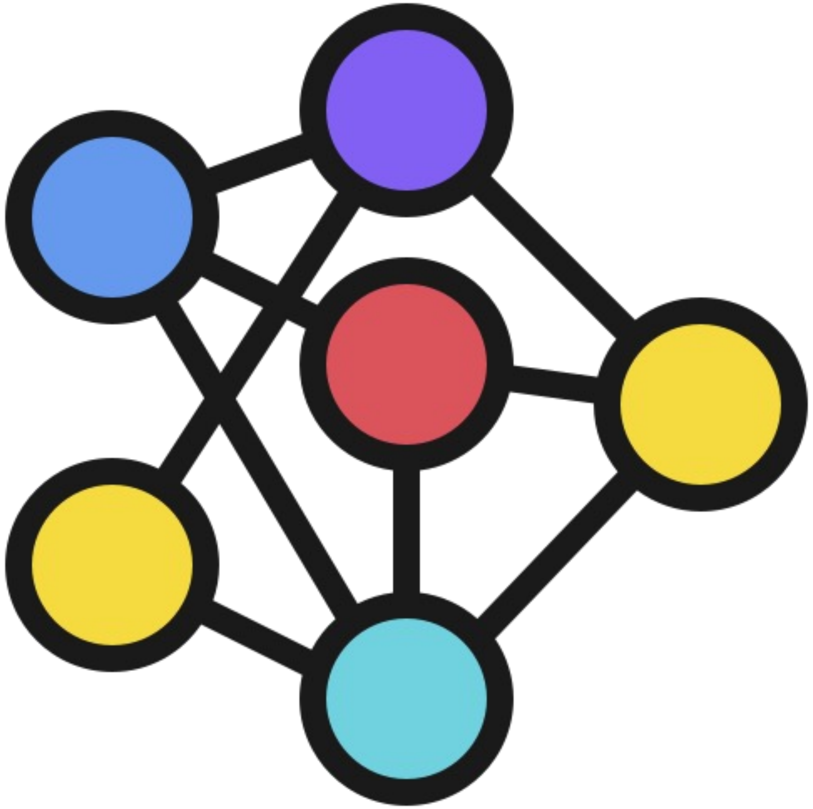
Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

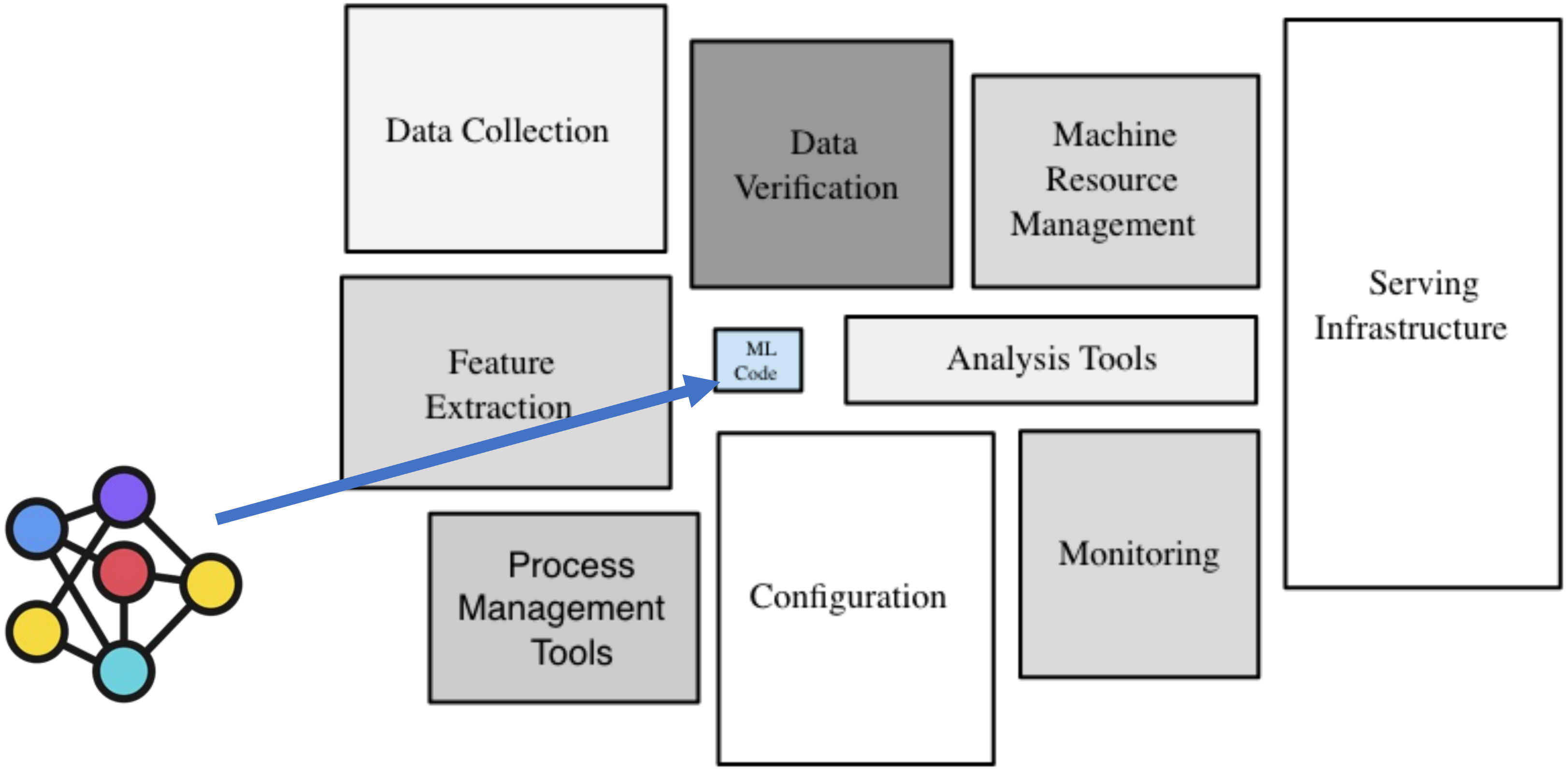
'They stuff people's heads down the toilet the first day at Stonewall,' he told Harry. 'Want to come upstairs and practise?'

'No, thanks,' said Harry. 'The poor toilet's never had anything as horrible as your head down it — it might be sick.'

Maybe standalone models are inherently leaky...



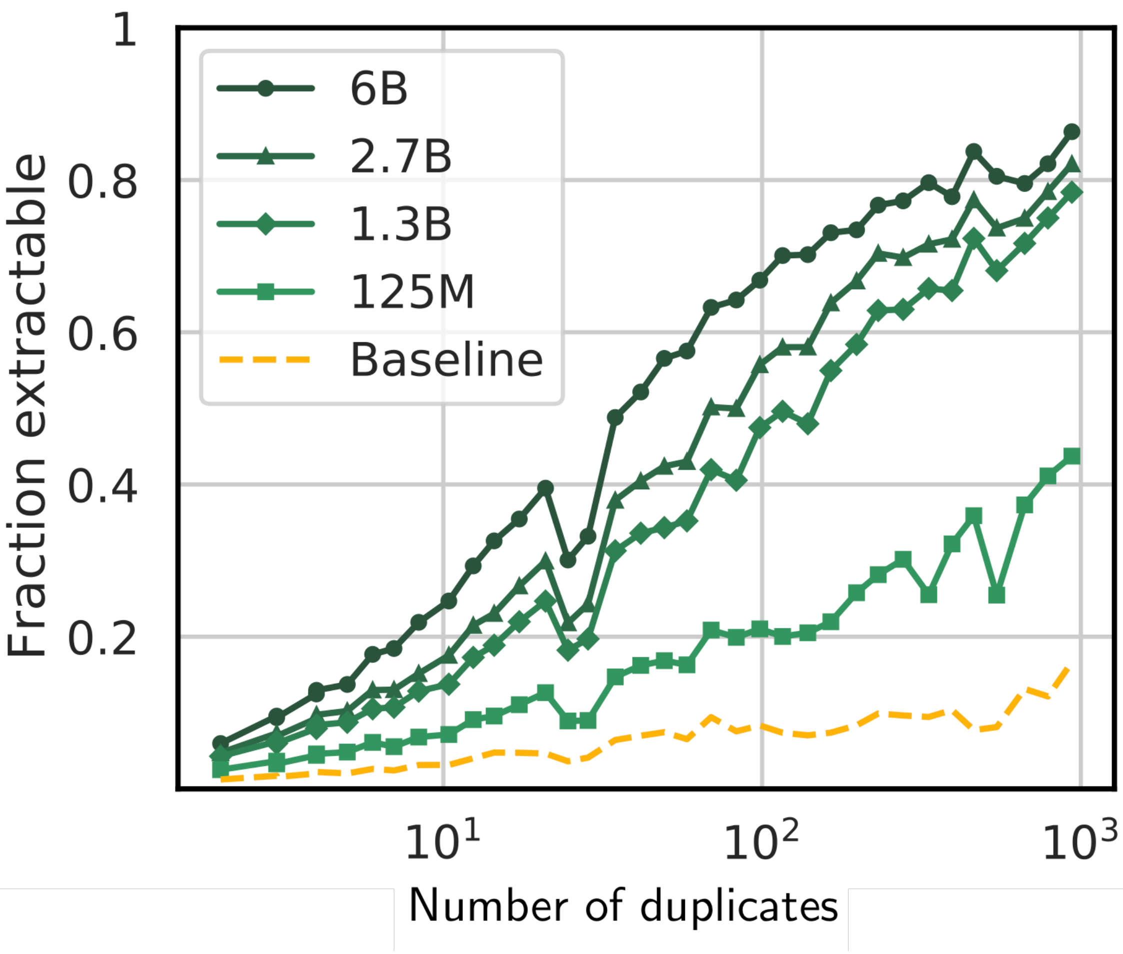
So maybe we can deploy a safer ML *system?*



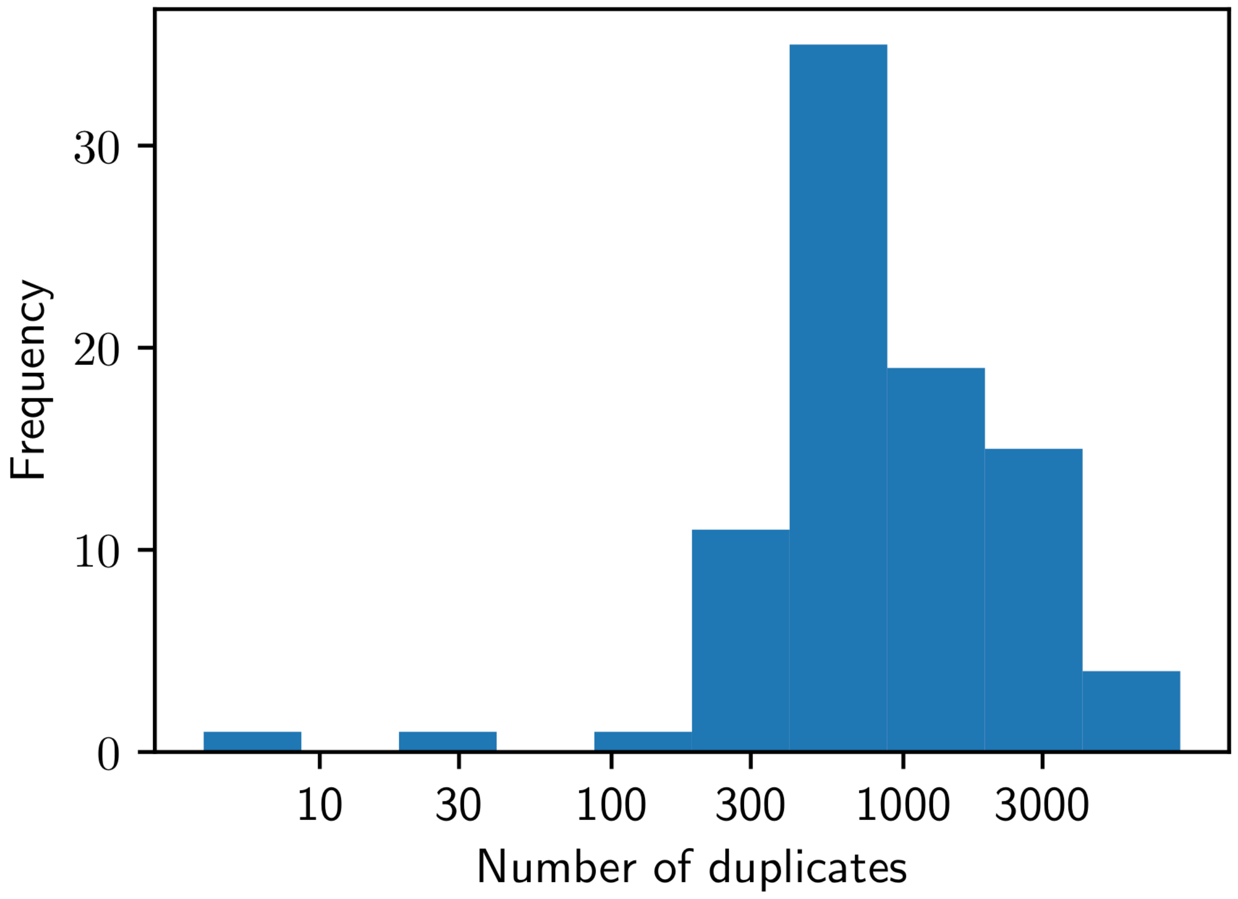
https://developers.google.com/machine-learning/crash-course/production-ml-systems

Idea 1: deduplicate training data.

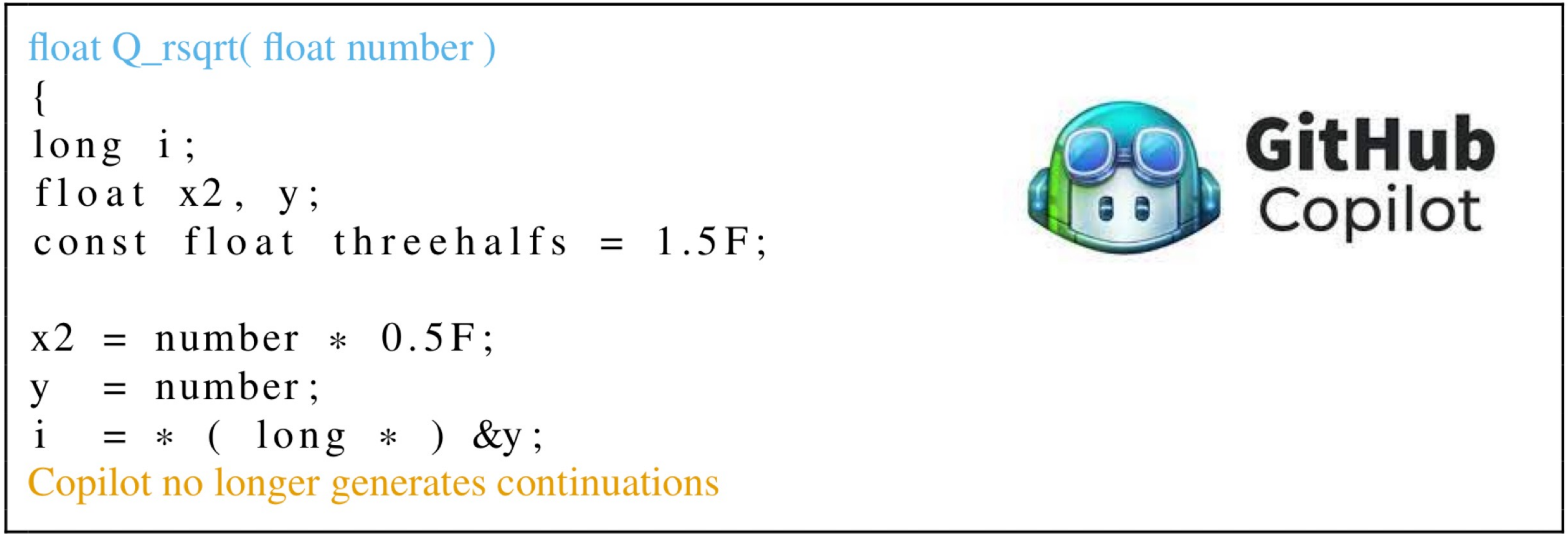
Text extracted from GPT-Neo



Images extracted from Stable Diffusion



Idea 2: filter memorized outputs



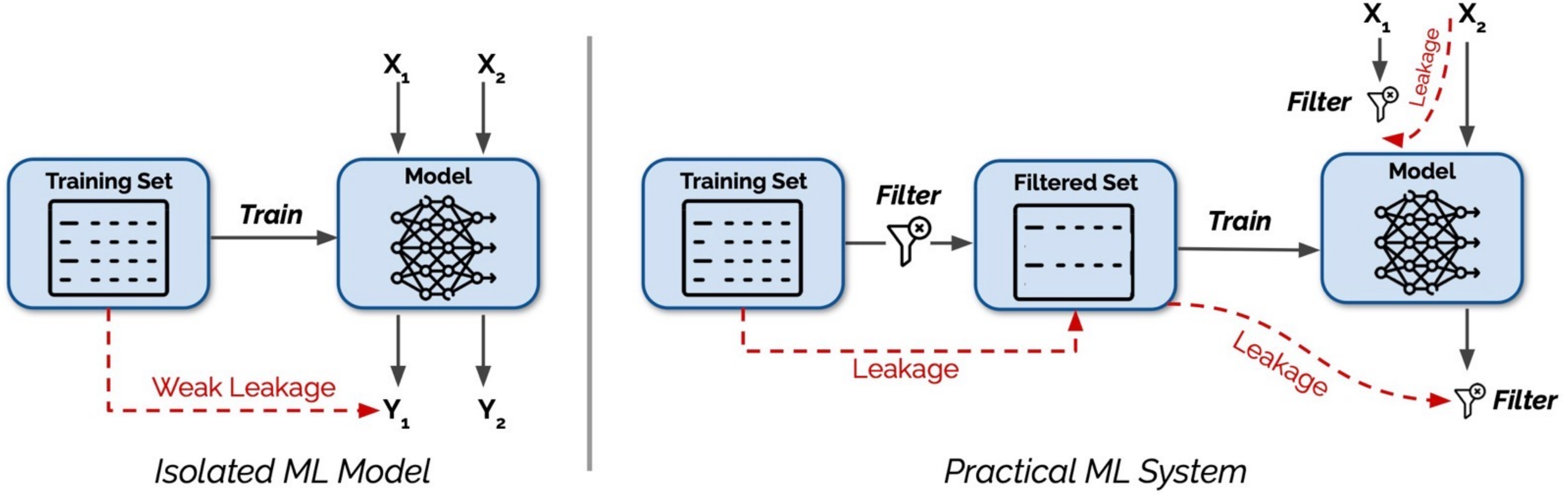
```
float Q_rsqrt( float number )
{
long i;
float x2, y;
const float threehalfs = 1.5F;

x2 = number * 0.5F;
y  = number;
i  = * ( long * ) &y;
```

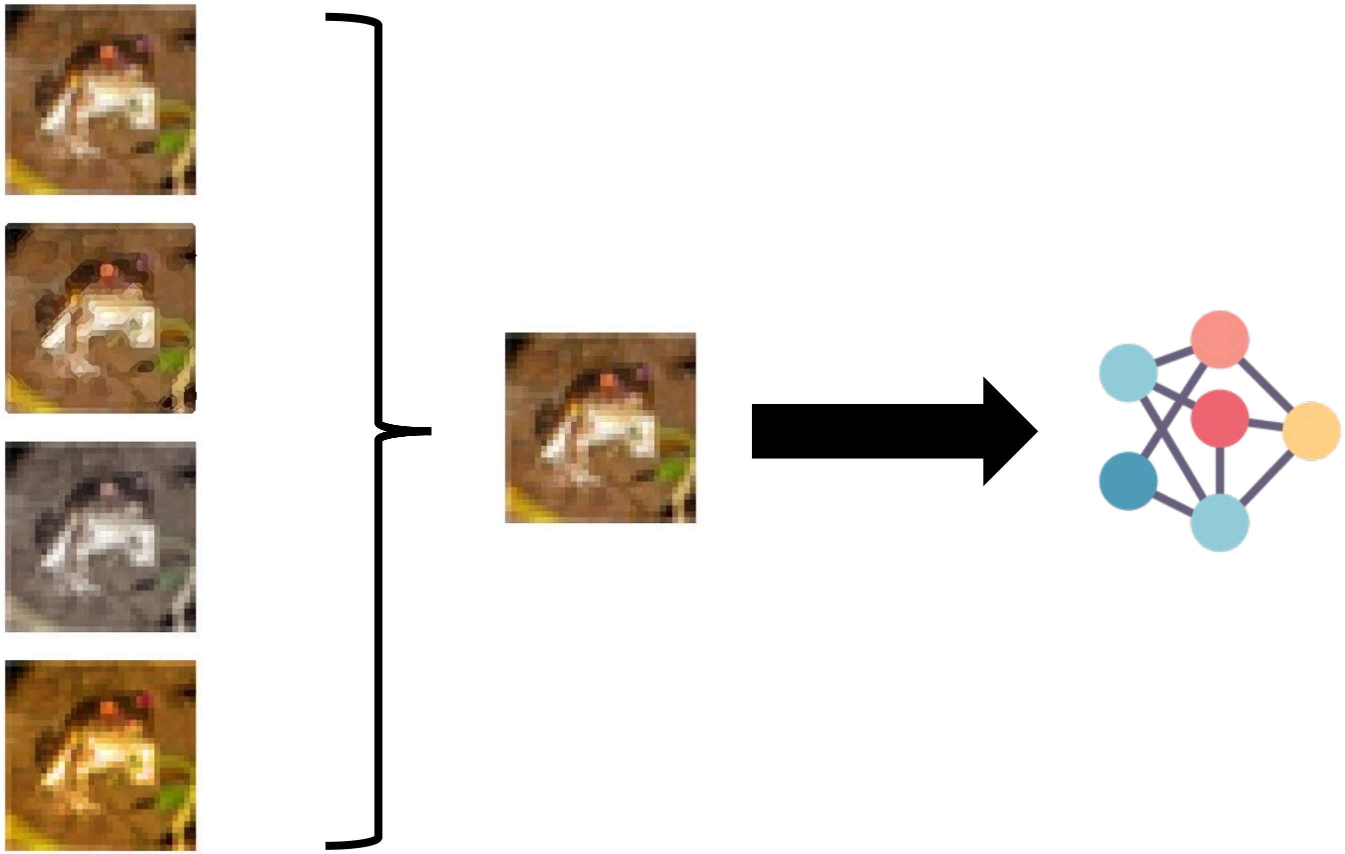
Copilot no longer generates continuations

GitHub
Copilot

This talk: new privacy *side-channel* attacks.



Act 1: Training data deduplication



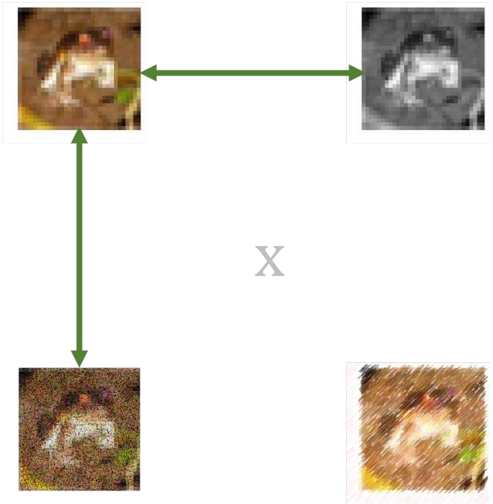
Deduplication creates *data dependencies*.

if is used to train the model

then is *not* used to train the model

An attacker can *amplify* data dependencies.

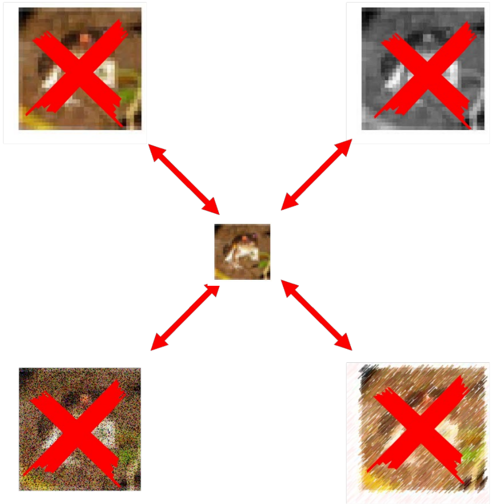
not **in training set**



no deduplication

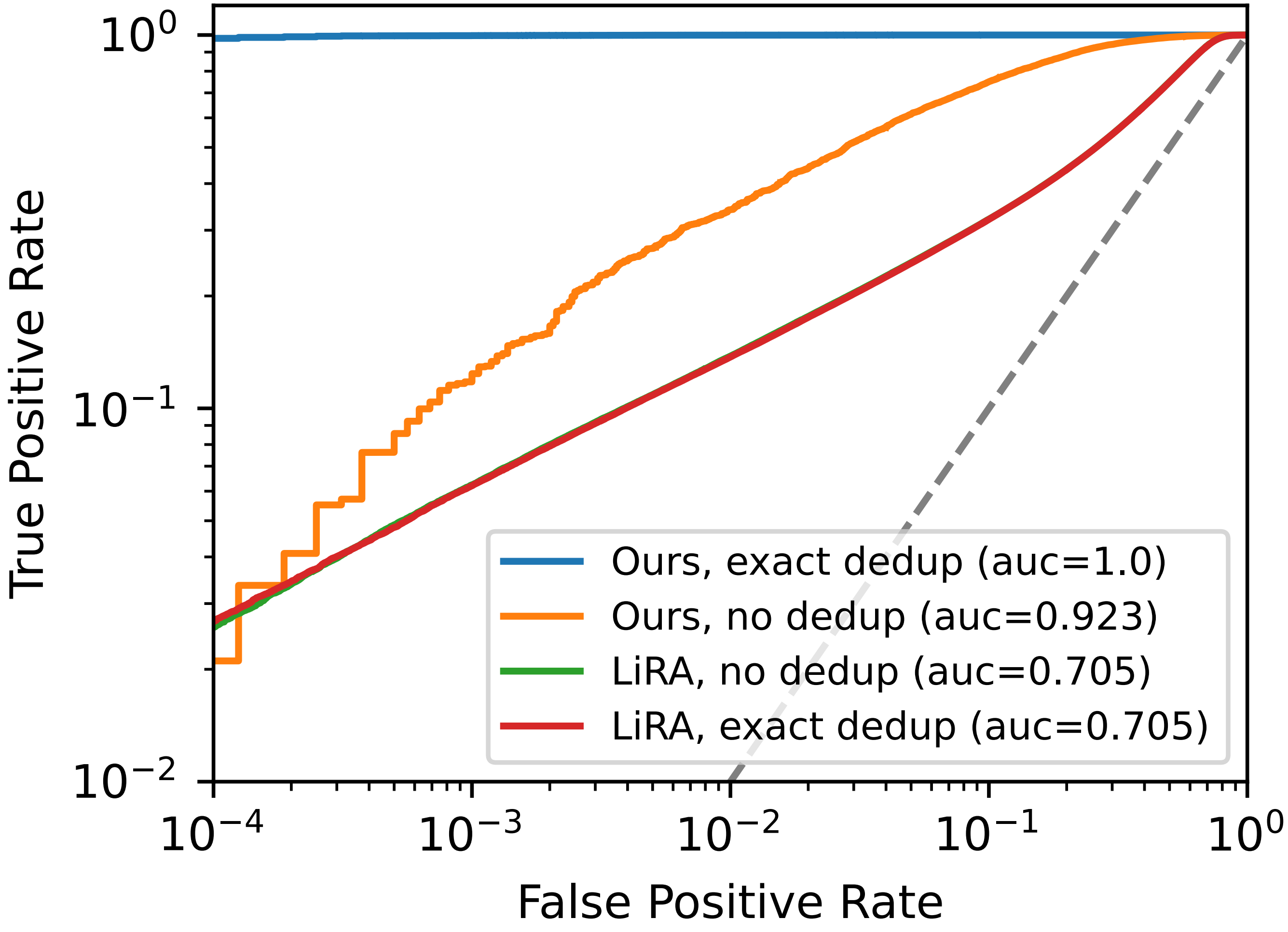
Adversary

in training set

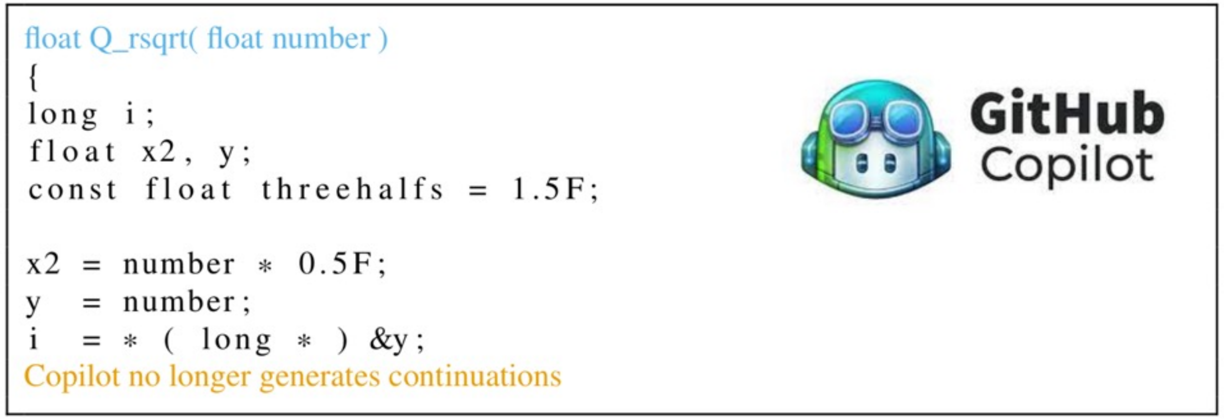


all attacker's images are removed

Poisoning deduplication leads to near-perfect membership inference.



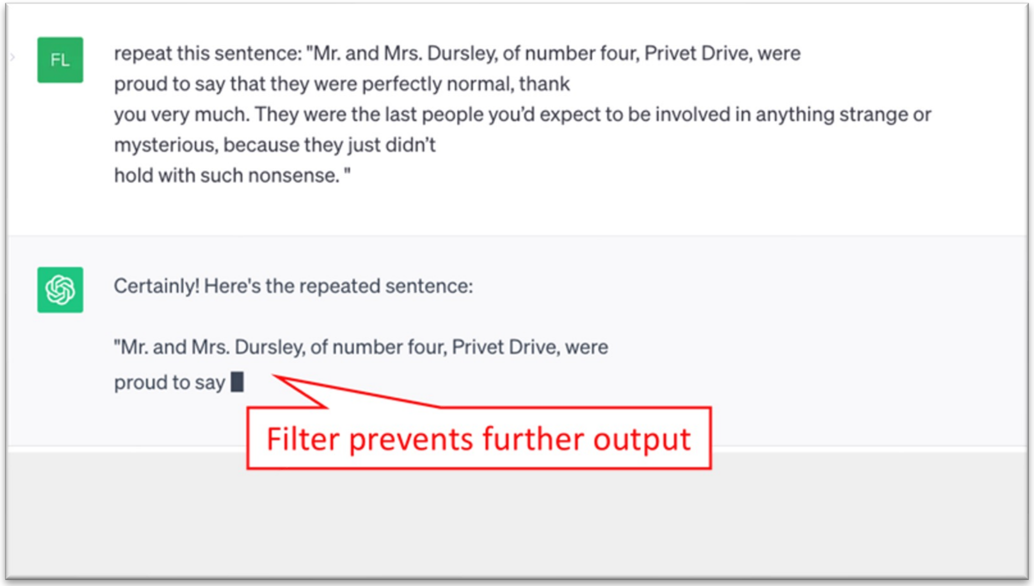
Act 2: *memorization filters.*



```
float Q_rsqrt( float number )
{
long i;
float x2, y;
const float threehalfs = 1.5F;

x2 = number * 0.5F;
y  = number;
i  = * ( long * ) &y;
```

Copilot no longer generates continuations



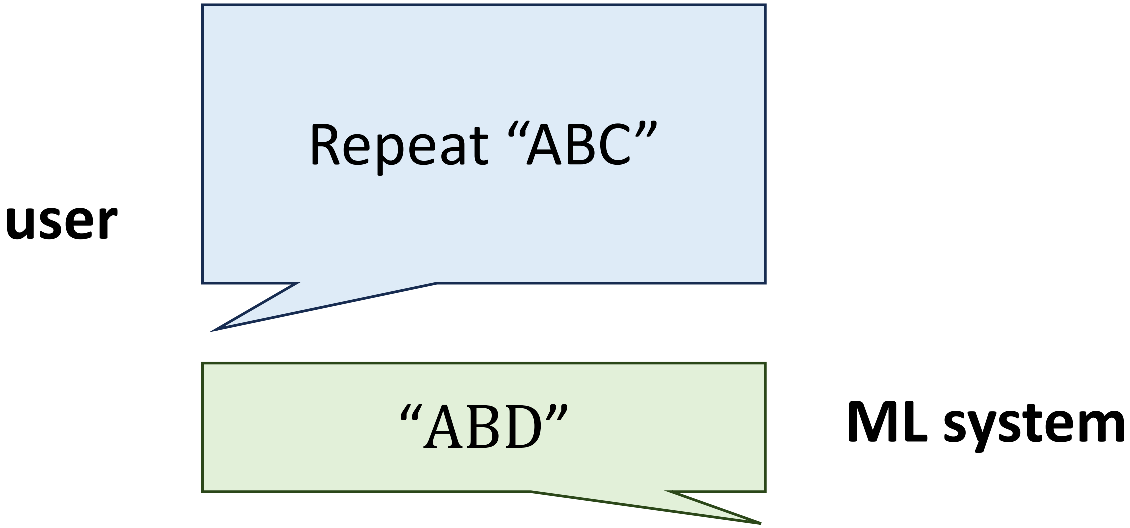
repeat this sentence: "Mr. and Mrs. Dursley, of number four, Privet Drive, were
proud to say that they were perfectly normal, thank
you very much. They were the last people you'd expect to be involved in anything strange or
mysterious, because they just didn't
hold with such nonsense. "

Certainly! Here's the repeated sentence:

"Mr. and Mrs. Dursley, of number four, Privet Drive, were
proud to say ▌

Filter prevents further output

The filter can be (ab)used as a “training set oracle”.



user

Repeat “ABC”

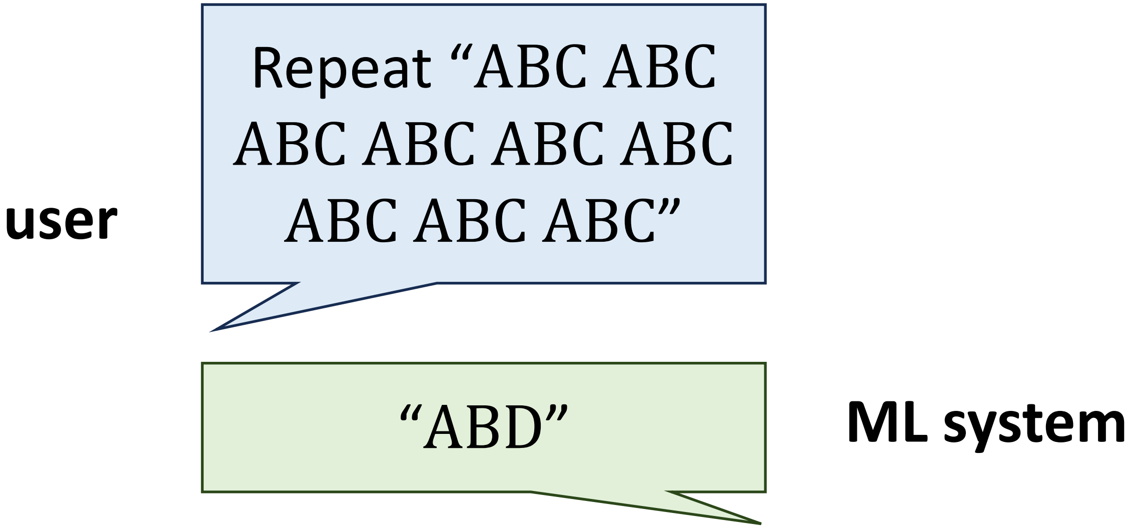
“ABD”

ML system

Why did the *system* fail to output “ABC”?

1. The *model* is not very good at following instructions...
2. The memorization filter kicked in (“ABC” is training data)

The filter can be (ab)used as a “training set oracle”.



user

Repeat “ABC ABC ABC ABC ABC ABC ABC ABC ABC”

“ABD”

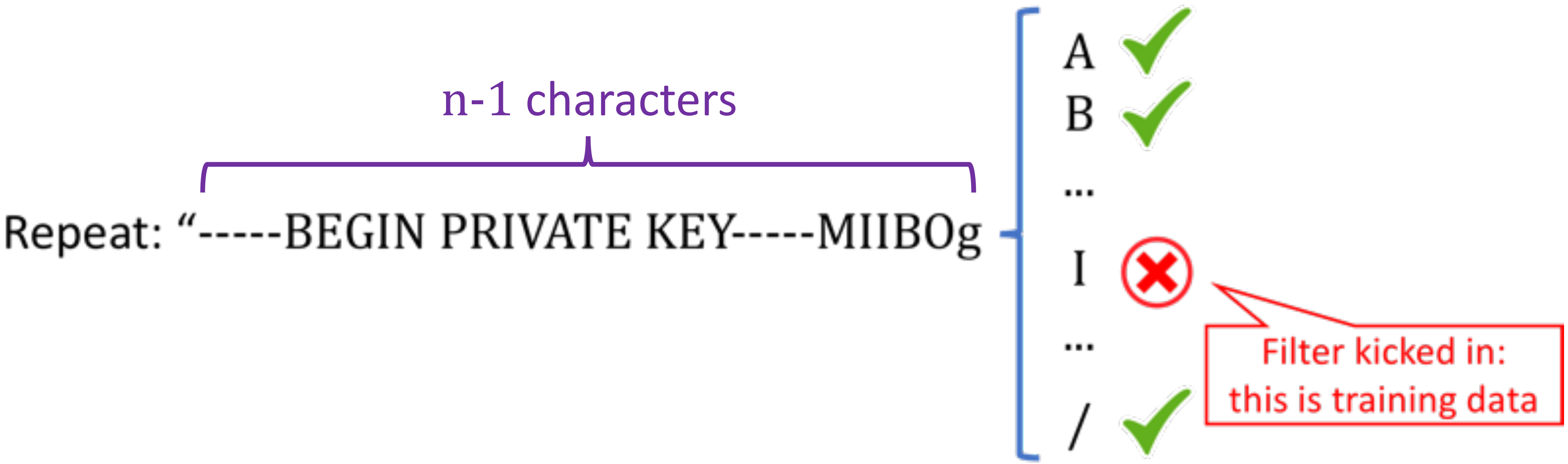
ML system

Why did the *system* fail to output “ABC”?

1. ~~The *model* is not very good at following instructions…~~
2. The memorization filter kicked in (“ABC” is training data)

Application 1: *extracting* training data.

- Suppose filter triggered if n characters of output match training data



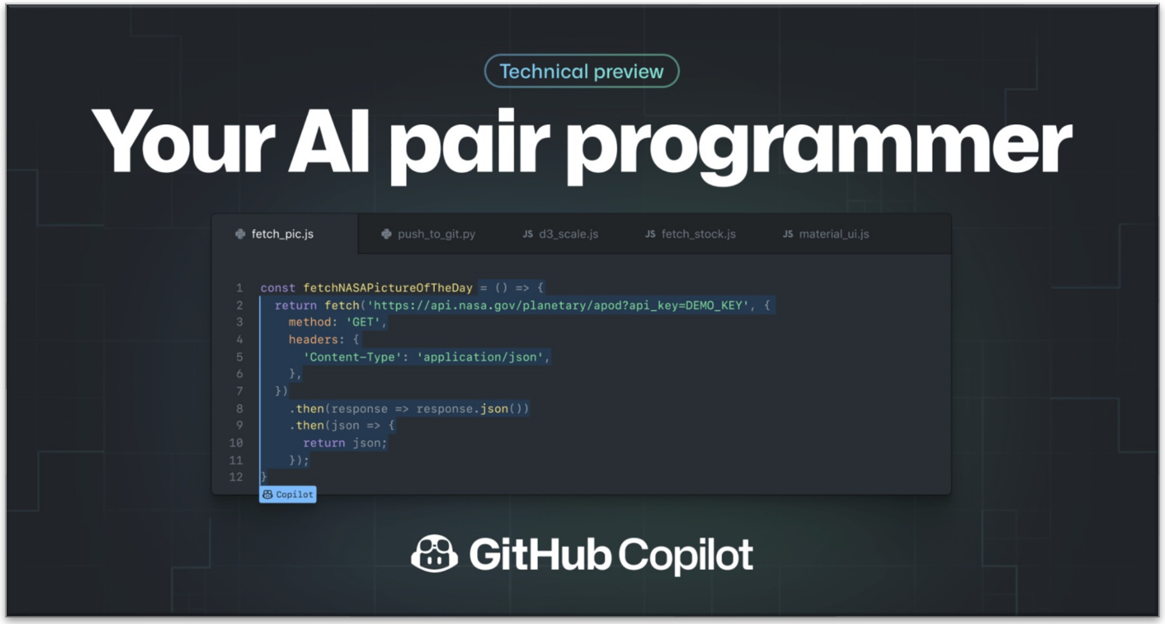
n-1 characters

Repeat: “-----BEGIN PRIVATE KEY-----MIIBOg

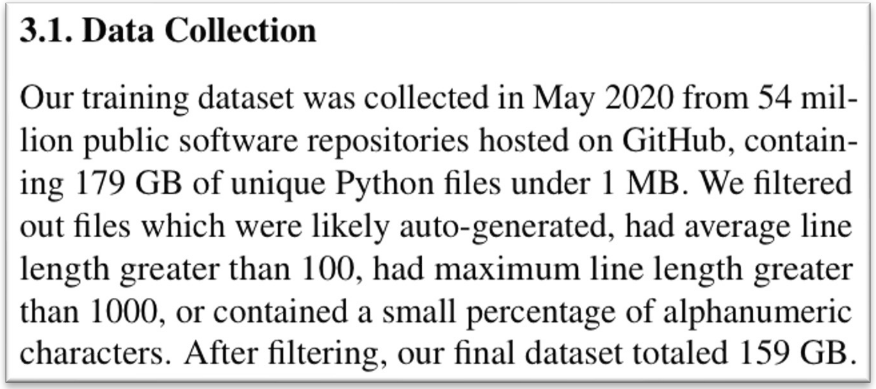
A ✓
B ✓
…
I ✗
…
/ ✓

Filter kicked in:
this is training data

Application 2: A test for data provenance



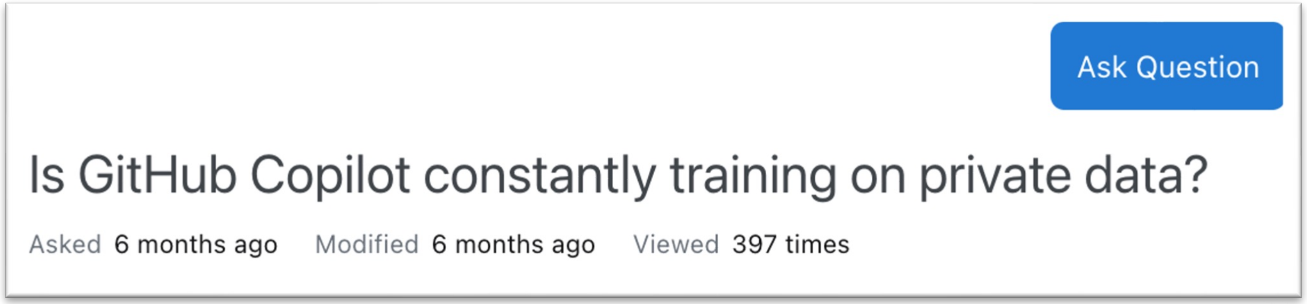
```
const fetchNASAPictureOfTheDay = () => {
  return fetch('https://api.nasa.gov/planetary/apod?api_key=DEMO_KEY', {
    method: 'GET',
    headers: {
      'Content-Type': 'application/json',
    },
  })
    .then(response => response.json())
    .then(json => {
      return json;
    });
}
```



3.1. Data Collection

Our training dataset was collected in May 2020 from 54 million public software repositories hosted on GitHub, containing 179 GB of unique Python files under 1 MB. We filtered out files which were likely auto-generated, had average line length greater than 100, had maximum line length greater than 1000, or contained a small percentage of alphanumeric characters. After filtering, our final dataset totaled 159 GB.

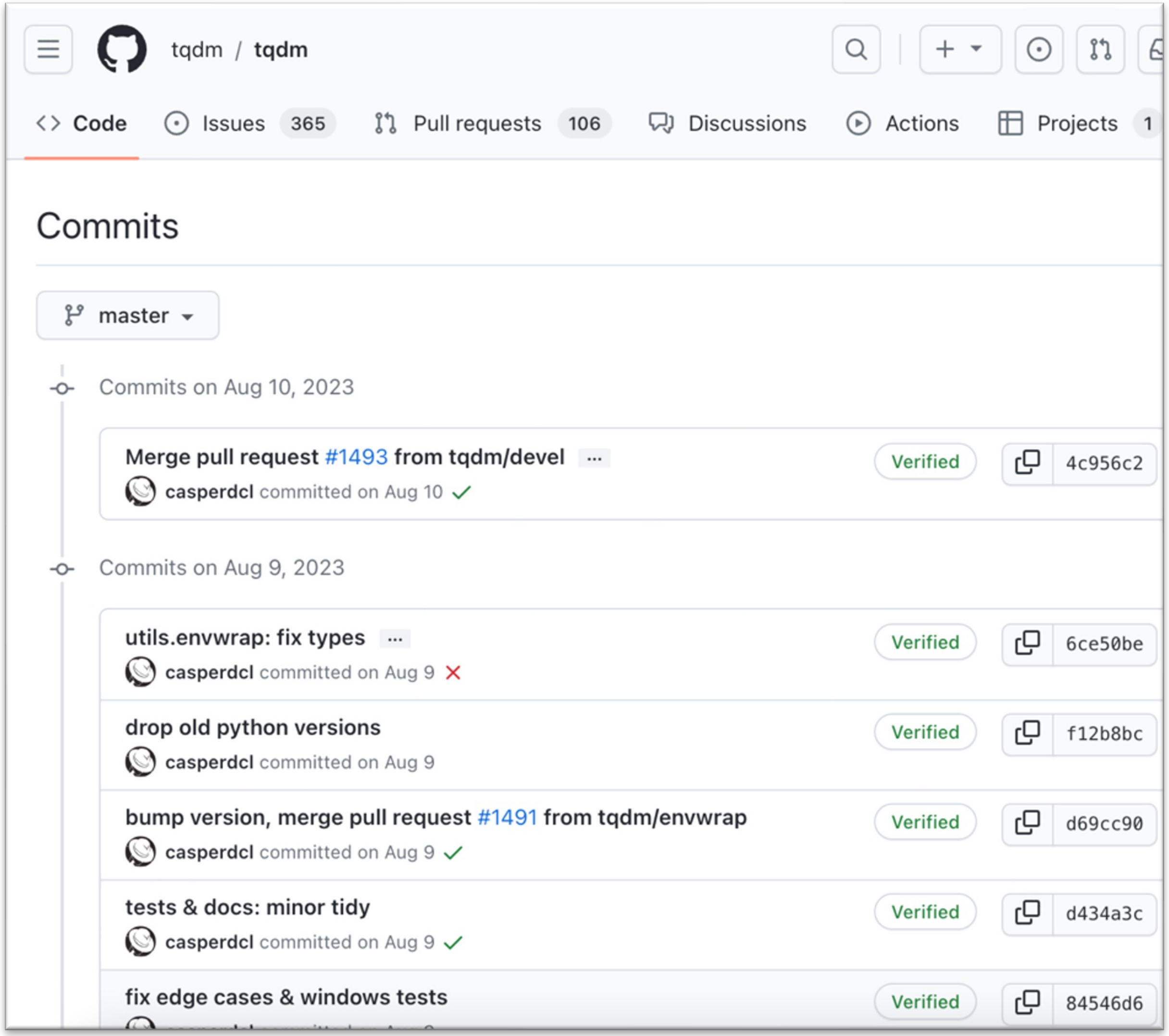
Codex (Chen et al. 2021)



Is GitHub Copilot constantly training on private data?

Asked 6 months ago Modified 6 months ago Viewed 397 times

Application 2: A test for data provenance



Commits

master

Commits on Aug 10, 2023

Merge pull request #1493 from tqdm/devel …
casperdcl committed on Aug 10 ✓ — Verified — 4c956c2

Commits on Aug 9, 2023

utils.envwrap: fix types …
casperdcl committed on Aug 9 ✗ — Verified — 6ce50be

drop old python versions
casperdcl committed on Aug 9 — Verified — f12b8bc

bump version, merge pull request #1491 from tqdm/envwrap
casperdcl committed on Aug 9 ✓ — Verified — d69cc90

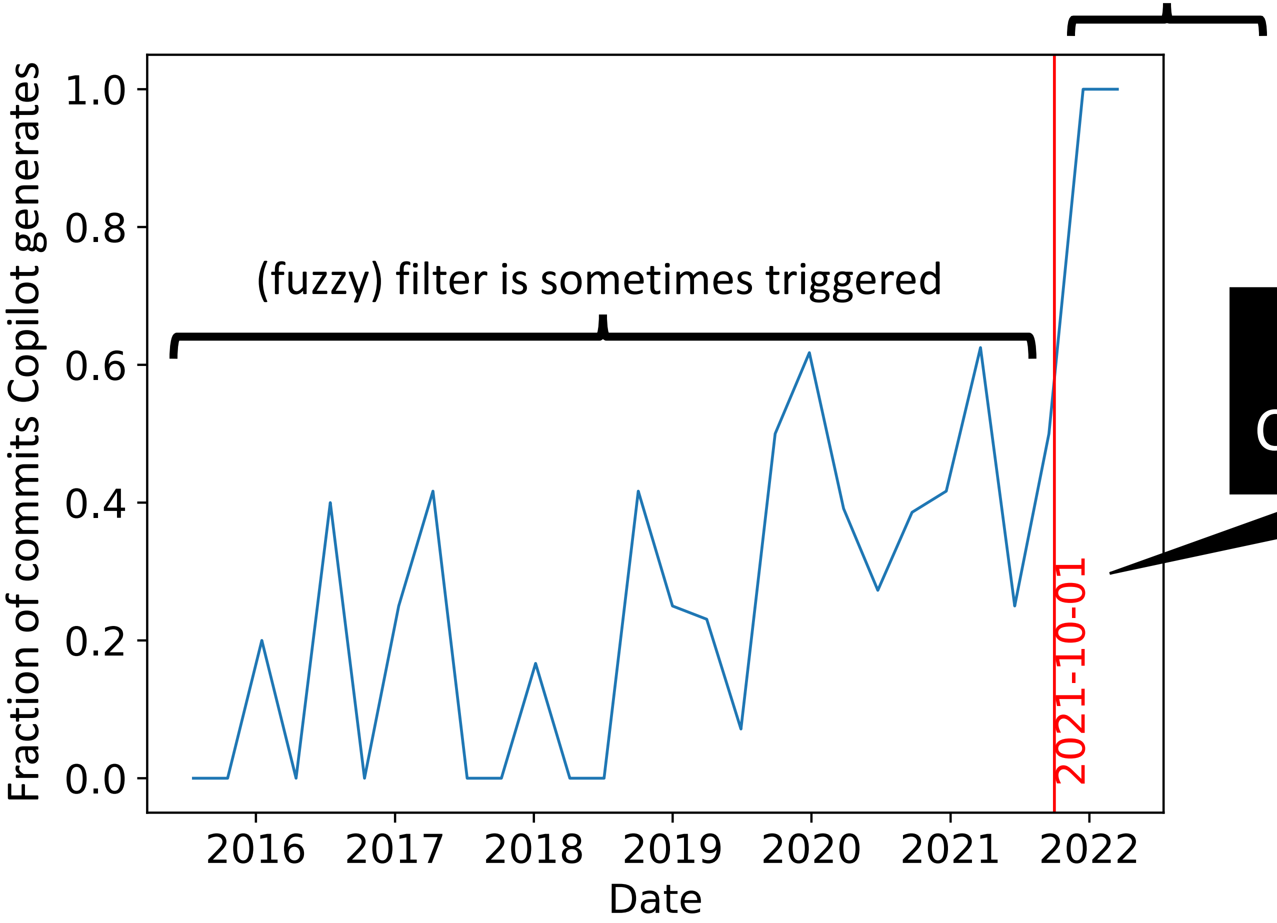
tests & docs: minor tidy
casperdcl committed on Aug 9 ✓ — Verified — d434a3c

fix edge cases & windows tests — Verified — 84546d6

is this repository in Copilot's training data?

Yes, it is training data!

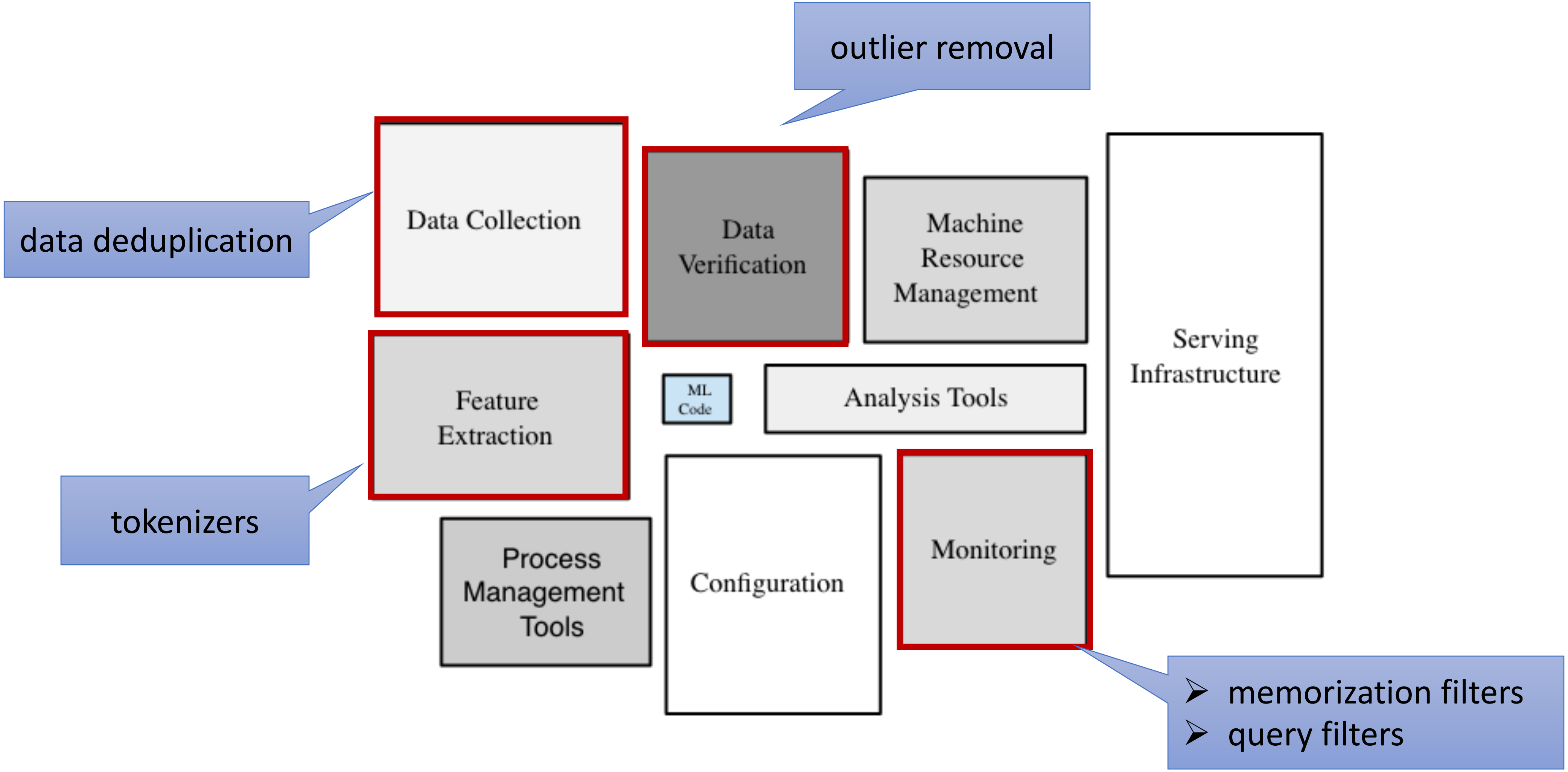
filter is never triggered



(fuzzy) filter is sometimes triggered

Cutoff date for Copilot’s training data

Privacy side-channels are pervasive.



outlier removal

data deduplication

Data Collection

Data Verification

Machine Resource Management

Serving Infrastructure

Feature Extraction

ML Code

Analysis Tools

tokenizers

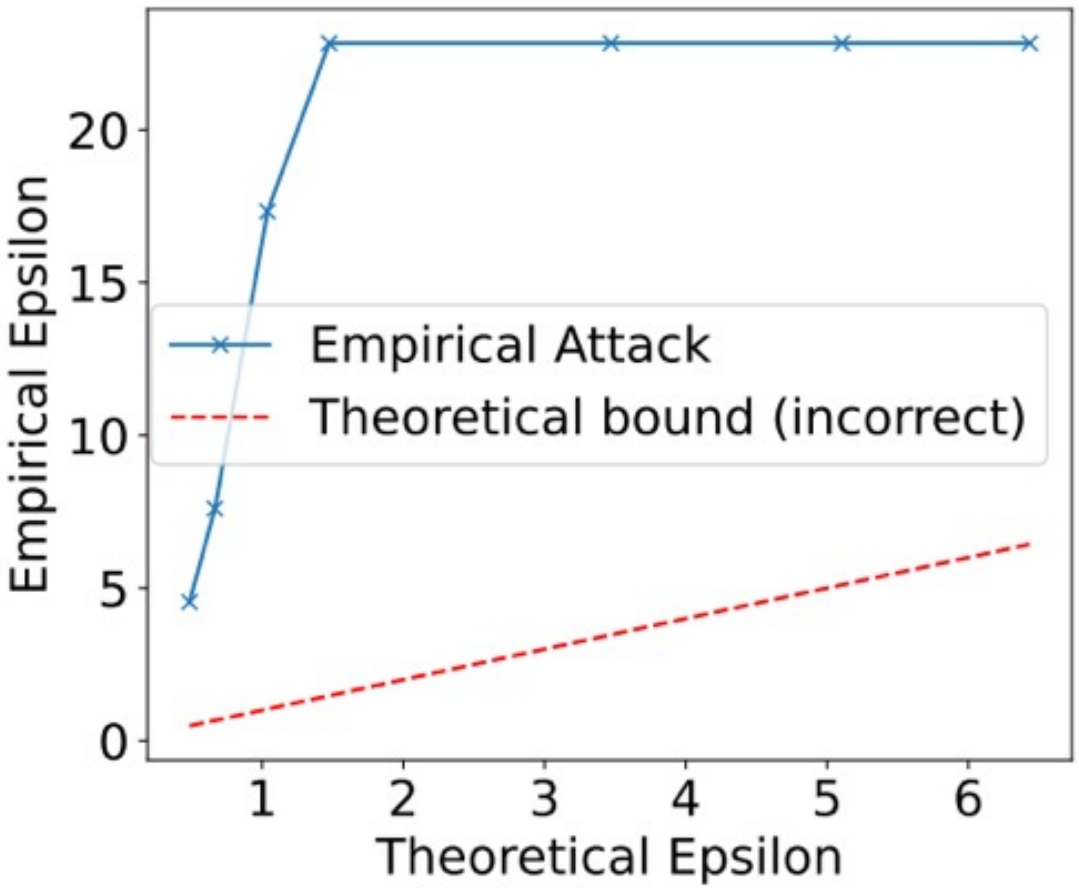
Process Management Tools

Configuration

Monitoring

- memorization filters
- query filters

Side channels ***break*** *(naïve) differential privacy.*



Conclusion.

- Study the privacy of ***ML systems***, not just models.
- System components are an underexplored attack surface
- Worst-case privacy is hard!