Forensics, Data and AI

1 - Background

2 - Recent research projects

3 - Overview of teaching experience

4 - Q&A

February 11, 2025 • Simson Garfinkel

These slides can be downloaded from https://simson.net/ref/2025

Background

Simson Garfinkel: Background and Bio

Career #1: Science writer (1985-)

- Newspapers, Magazines, Books
- Most recently: History of computing — Technology Review & CACM

Career #2: Entrepreneur (1992-)

- SGAI — 1992-1993 — Commercialized AI approach from MIT Media Lab
- Vineyard.NET — 1995-2002 — ISP on Martha’s Vineyard
- Sandstorm Enterprises — 1998-2001, 1998-2006 (board) — Security tools
- Broadband2Wireless — 2000-2001 — Wireless ISP

Career #3: CS Researcher (1985-87, 90-91, 2002-)

- MIT Media Lab 1985-1987, 90-91
- MIT PhD 2003-2005
- Harvard SEAS CRCS — 2005-2006
- Naval Postgraduate School — 2006-2014
- NIST — 2015-2016

Career #4: Government Innovator

- US Census Bureau — 2017-2021 — Differential privacy for the 2020 Census
- US DHS — 2021-2022 — DHS Chief Data Officer – Data Inventory

I’m currently Chief Scientist at BasisTech LLC, a startup accelerator in Somerville.

Nurturing ideas into viable solutions

We work with entrepreneurs to refine ideas, attract talent, access capital, and accelerate growth.

Exceptional people with promising ideas

We focus on early-stage high-tech companies that show the potential to significantly improve the world. Our companies focus on enterprise data, machine vision, and the public sector.

Major research projects

Bulk_extractor (2006-2020) — processing bulk data for digital forensics

- Developed novel approaches to extracting formatted data from bulk data.
- $15 million in funding from DOD, DHS, NSF, FBI and others.
- Produced a widely used digital forensics tool, multiple peer-reviewed publications, hundreds of training videos on YouTube.

Digital Corpora (2006-) — realistic data for digital forensics research and education

- Funded by NSF, NIST and Amazon
- Major contribution to DARPA SafeDocs project.
- 26TB of open use data — hard drive images, worked “scenarios” (with solutions), packets, files

Disclosure Avoidance System for the 2020 Census (2017-2021)

- Most complex deployment of differential privacy to date.
- Lead computer scientist for the project. Developed experimental and production framework
- Instituted modern software development practices.

PlantTracer (2023-) — Using computer vision to watch plant movement.

- An online environment for high school and college students.

What’s next:
Building a regional tech abuse center

Tech abuse is a major threat facing many users.

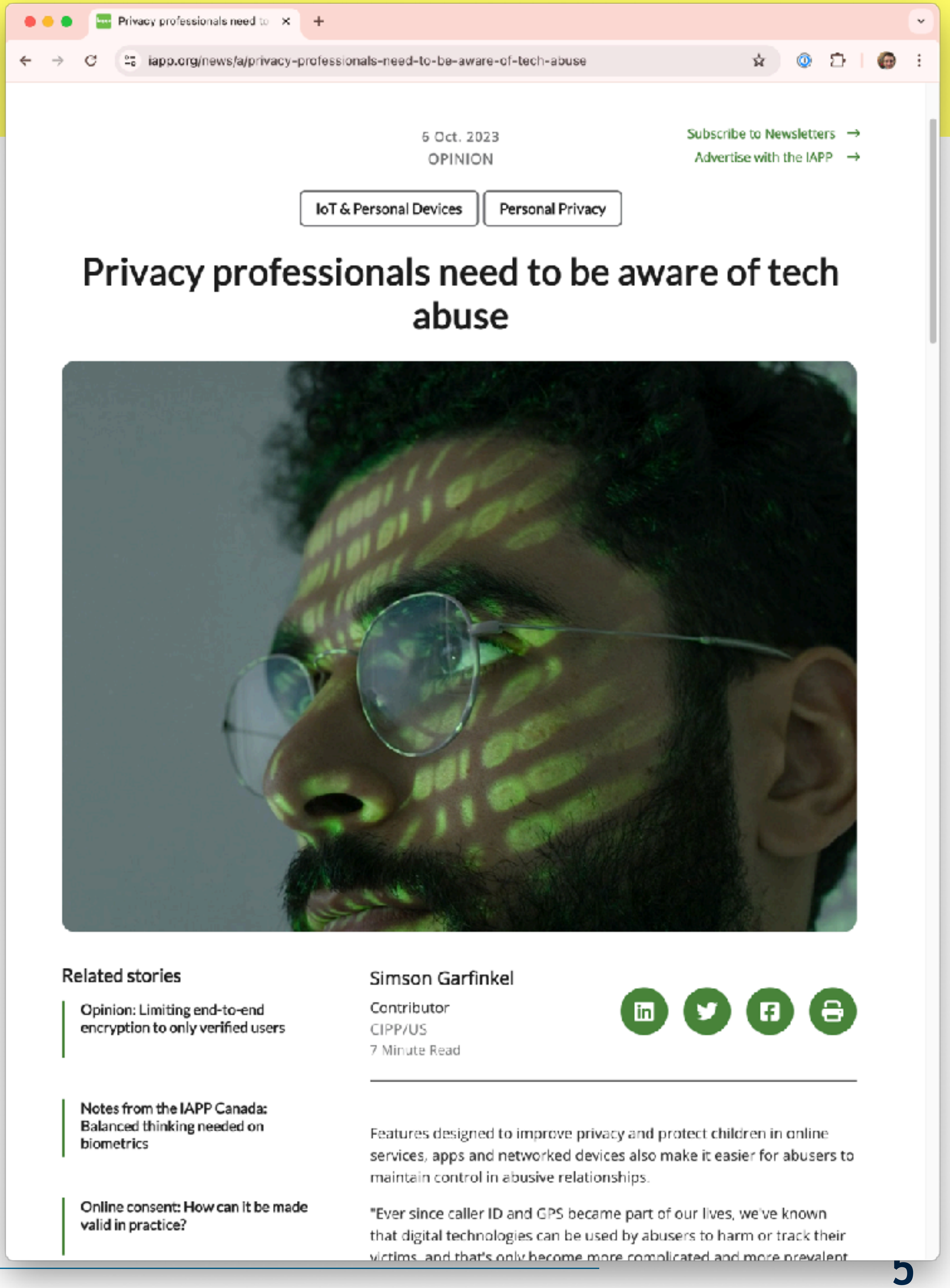
- Like “physical abuse” and “metal abuse,” but with technology.

Examples:

- One partner configures the other’s phone with “family controls.”
- Using router’s “family controls” to monitor (or disable) internet access.
- Eavesdropping on emails sent to friends (or lawyers)
- Turning off cell phone service as “punishment.”
- Hidden cameras

What can we do?

- **CS research:** Identify patterns for defense and detection
- **Social research:** Identify scale of the problem
- **Policy research:** Inform policymakers & propose solutions
- **Tech clinic:** Help people clear their devices
- **Legal clinic:** Help with protective orders
- **Community building:** Organize academics & activists in the Boston area.


Privacy professionals need to
iapp.org/news/a/privacy-professionals-need-to-be-aware-of-tech-abuse

5 Oct. 2023
OPINION

Subscribe to Newsletters →
Advertise with the IAPP →

IoT & Personal Devices | Personal Privacy

Privacy professionals need to be aware of tech abuse

Related stories

Opinion: Limiting end-to-end encryption to only verified users

Notes from the IAPP Canada: Balanced thinking needed on biometrics

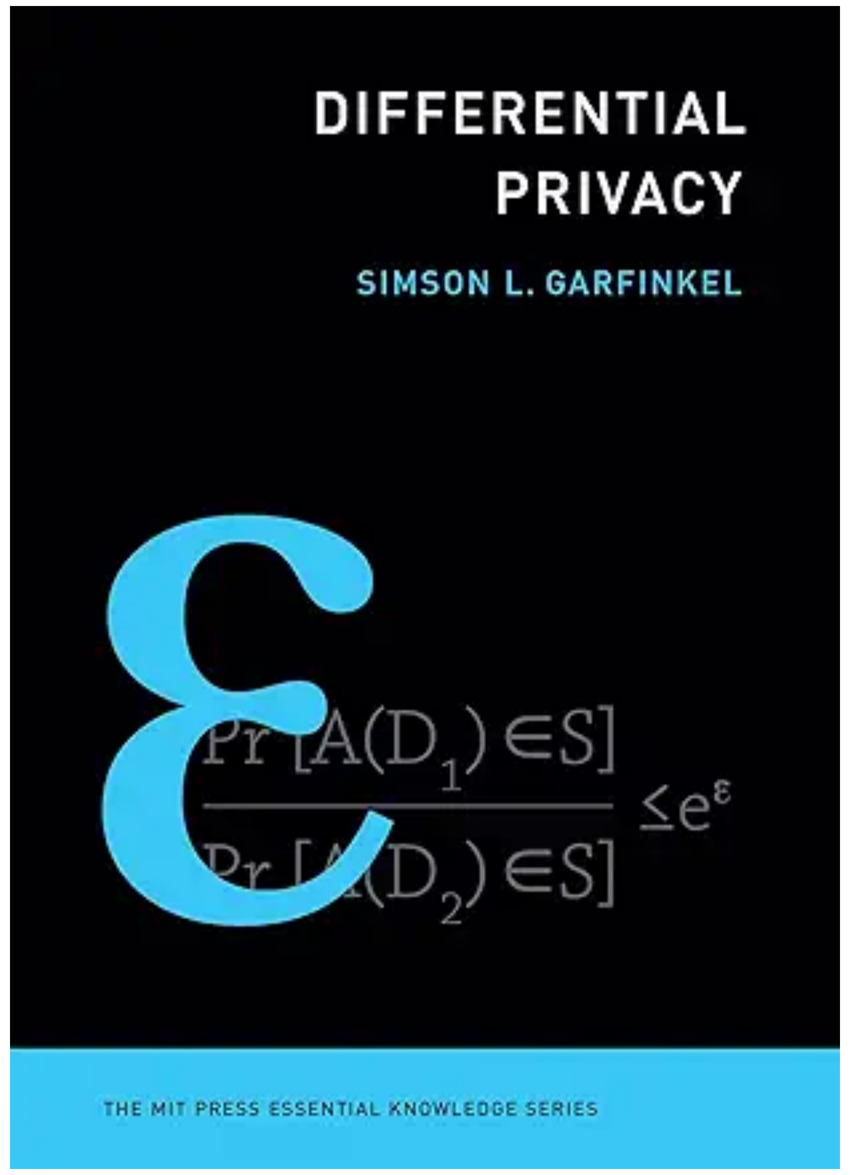
Online consent: How can it be made valid in practice?

Simson Garfinkel
Contributor
CIPP/US
7 Minute Read

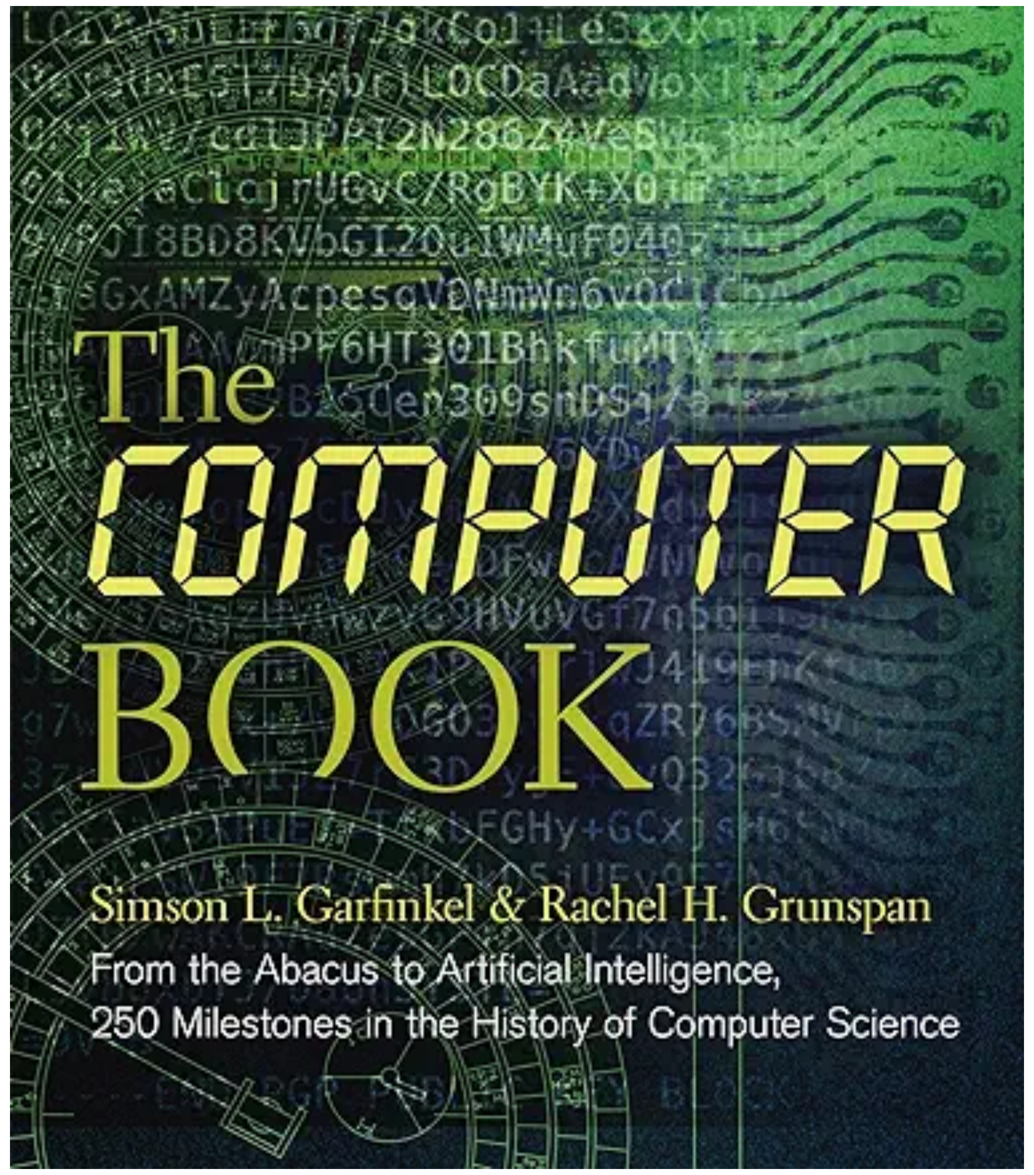
Features designed to improve privacy and protect children in online services, apps and networked devices also make it easier for abusers to maintain control in abusive relationships.

"Ever since caller ID and GPS became part of our lives, we've known that digital technologies can be used by abusers to harm or track their victims, and that's only become more complicated and more prevalent

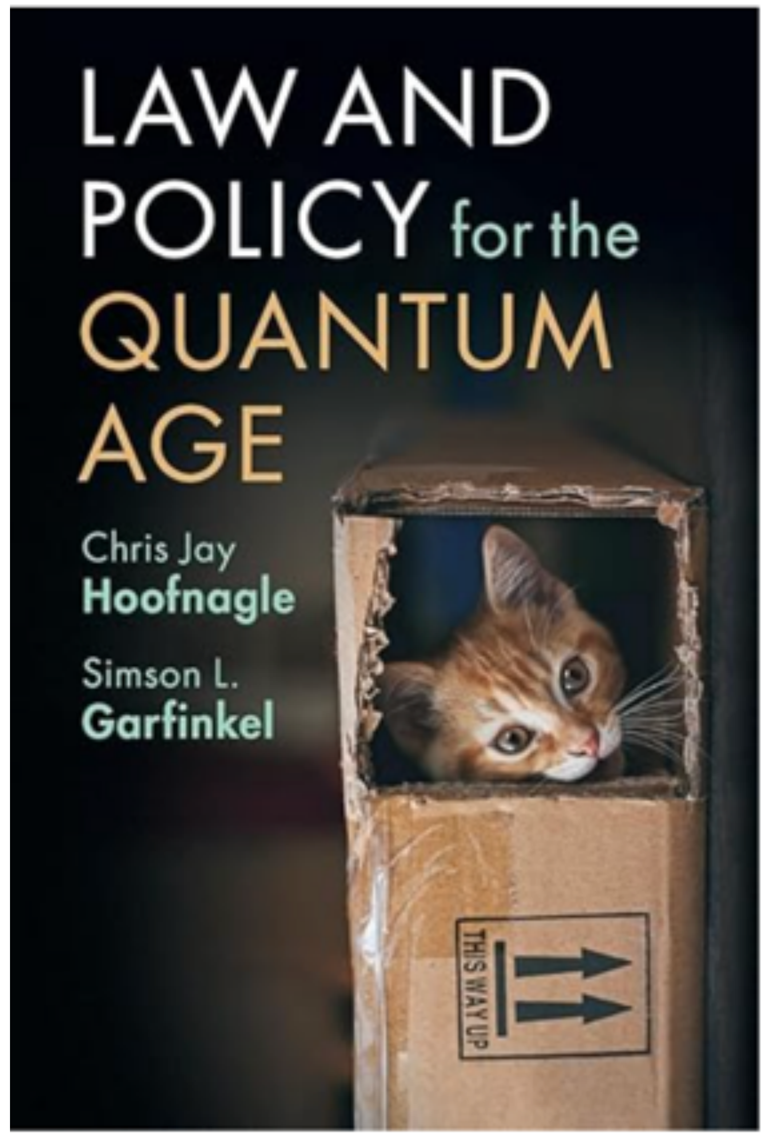
Recent books


DIFFERENTIAL PRIVACY
SIMSON L. GARFINKEL
ε
$\frac{\Pr[A(D_1) \in S]}{\Pr[A(D_2) \in S]} \le e^{\varepsilon}$
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

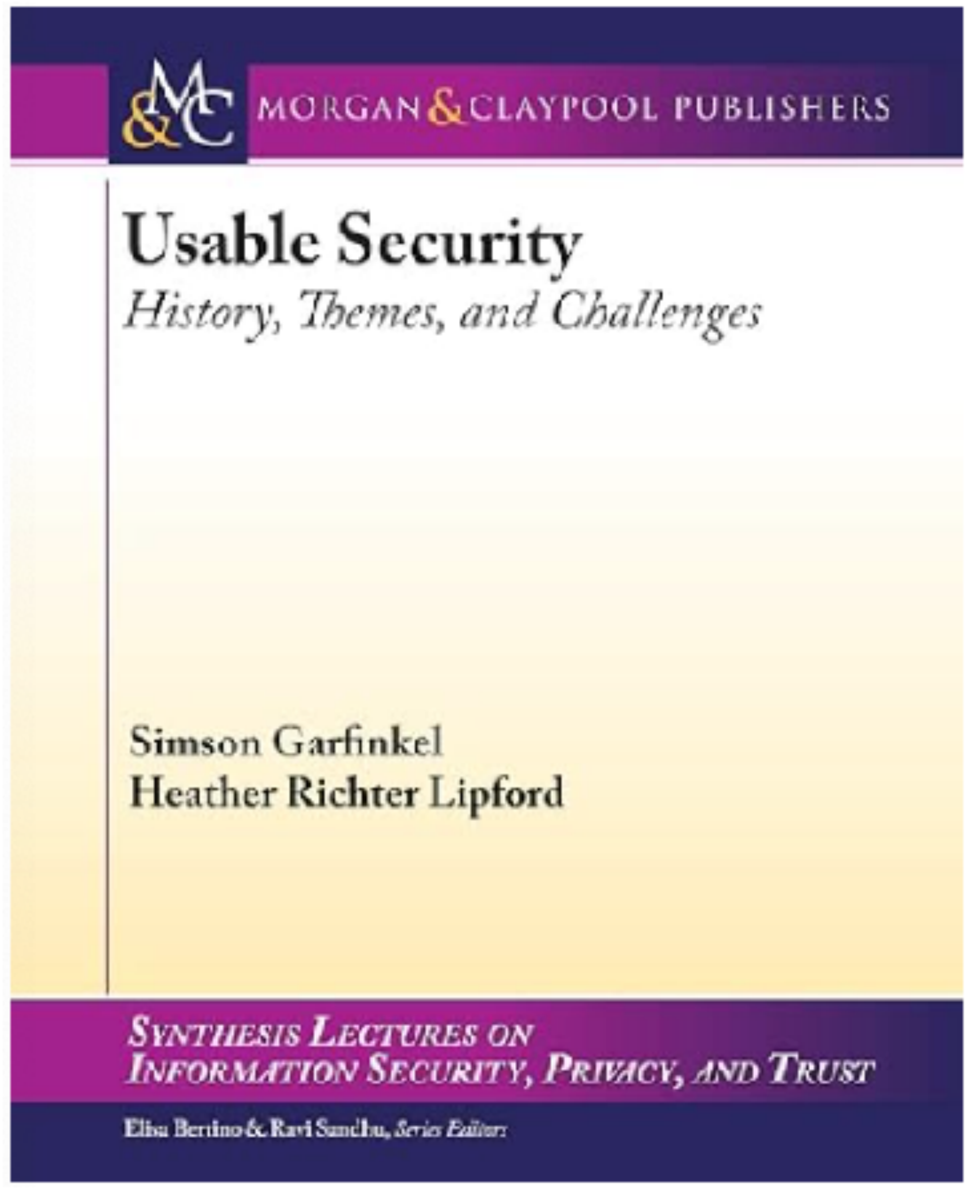
Differential Privacy
MIT Press, 2025


The COMPUTER BOOK
Simson L. Garfinkel & Rachel H. Grunspan
From the Abacus to Artificial Intelligence,
250 Milestones in the History of Computer Science

The Computer Book: From the Abacus to Artificial Intelligence, 250 Milestones in the History of Computer Science (Sterling Milestones), by Simson L. Garfinkel and Rachel H. Grunspan. 2018 (Sterling)



Law and Policy for the Quantum Age,
Chris Jay Hoofnagle and Simson L. Garfinkel,
2021 (Cambridge)

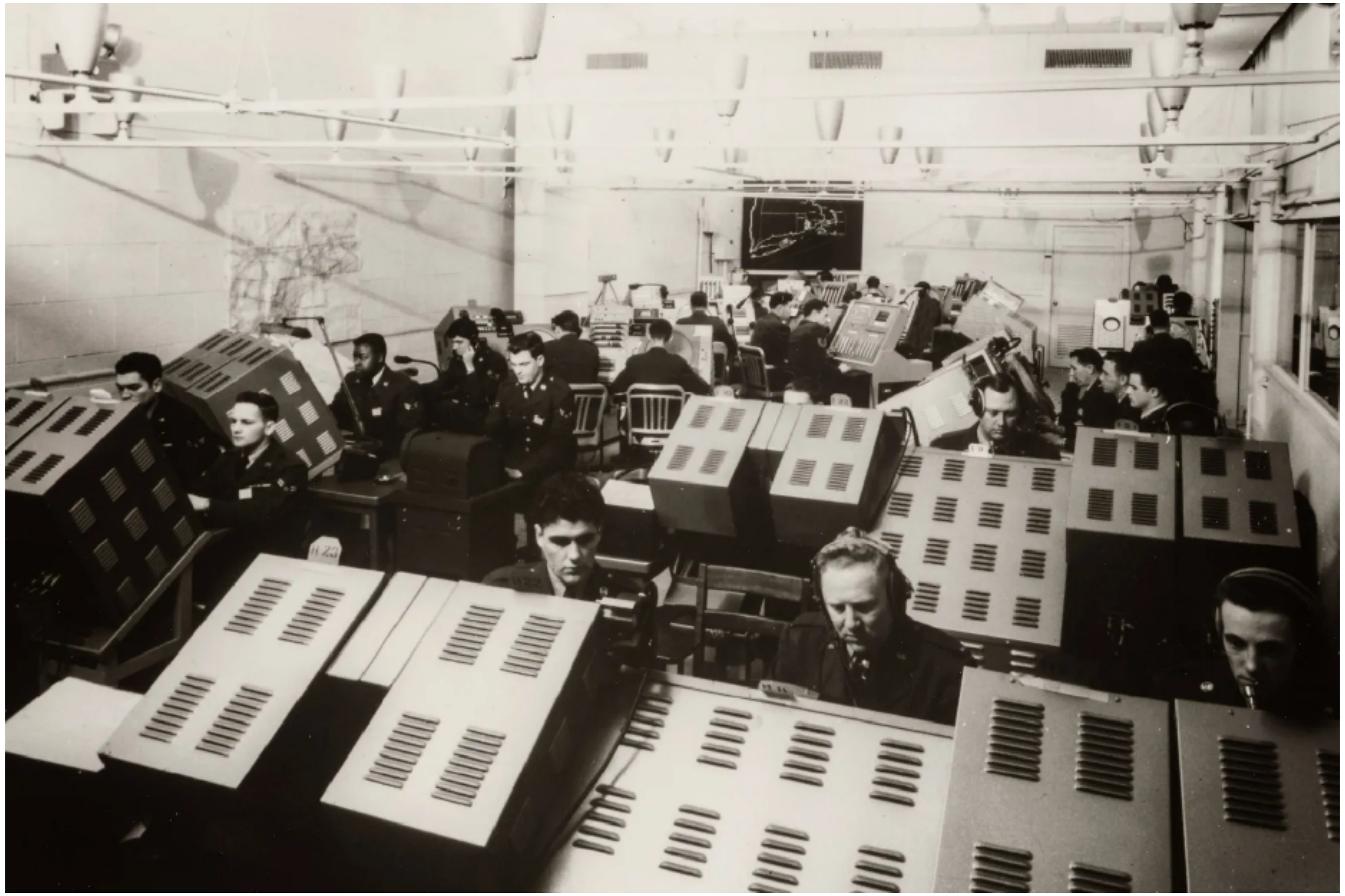

M&C MORGAN & CLAYPOOL PUBLISHERS
Usable Security
History, Themes, and Challenges
Simson Garfinkel
Heather Richter Lipford
SYNTHESIS LECTURES ON INFORMATION SECURITY, PRIVACY, AND TRUST
Elisa Bertino & Ravi Sandhu, Series Editors

Usable Security: History, Themes, and Challenges,
Simson Garfinkel and
Heather Lipford, 2014.
(Morgan & Claypool, part of the Synthesis Lectures on Information Security, Privacy and Trust series.)

Recent articles in the history of computing and MIT.

History of Computing, Technology and MIT

1. **Garfinkel, S. MIT's (mostly) secret society, December 23, 2024**
2. **Garfinkel, S. This is MIT and yes, we have bananas, August 27, 2024**
3. **Garfinkel, S. Editor of The Tech becomes president of MIT, Technology Review, June 25, 2024**
4. **Garfinkel, S. How Technology Review got its start, Technology Review, January 4, 2024**
5. **Garfinkel, S. MIT's First Divorce (how MITRE was created and got its name), Technology Review, June 27, 2023**
6. **Garfinkel, S. Cold Trick Indeed (dorm room set up on the Charles, 1985), Technology Review, December 19, 2022**
7. **Garfinkel, S. How an MIT Marxist weathered the Red Scare (Dirk Struik), Technology Review, June 29, 2022**
8. **Garfinkel, S. In praise of the Feistel network (Horst Feistel '37), Technology Review, April 27, 2022**
9. **Garfinkel, S. The man no one knows who changed Boston (Charles Hayden), Technology Review, February 23, 2022**
10. **Garfinkel, S. 5 MIT patents that changed computing, Technology Review, February 23, 2022**
11. **Garfinkel, S. Walker and the "Indian Question:" Before arriving at MIT, Francis Amasa Walker had twice led the US Census—and helped justify the troubling US policy of containing Native Americans on reservations. Technology Review, August 24, 2021**
12. **Garfinkel, S. Tomorrow's computer, yesterday. Four decades ago at Endicott House, an MIT professor convened a conference that launched quantum computing. Technology Review, April 27, 2021**
13. **Garfinkel, S. Punching In: Bored teaching at MIT, Herman Hollerith left to launch the information age for the US Census. Technology Review, August 18, 2020**
14. **Garfinkel, S. Everything is a Punch Card. ;login:, Fall 2020**
15. **Garfinkel, S. The Tricky Cryptographic Hash Function. ;login:, Winter 2020**
16. **Garfinkel, S. Shafi Goldwasser, Technology Review, August 21, 2019**
17. **Garfinkel, S. Radia Perlman '73, SM '76, PhD '88, Technology Review, August 21, 2019**
18. **Garfinkel, S. The Geek (Chris Schmandt), Technology Review, April 24, 2019**



Garfinkel, S. MIT's First Divorce (how MITRE was created and got its name), Technology Review, June 27, 2023

Outline for the rest of this talk

Research projects:

- Bulk_extractor (2006-2020) — processing bulk data for digital forensics
- Digital Corpora (2006-) — realistic data for digital forensics research and education
- Disclosure Avoidance System for the 2020 Census (2017-2021)
- PlantTracer (2023-) — Using computer vision to watch plant movement.

Two fun classes

- “Defense Against the Dark Arts (cybersecurity edition)”
- Artificial Intelligence, Internet of Things, and Cybersecurity,

What’s Next: Two big projects for 2025-2026

- Tech Abuse Research Center
- AI for Data Management

Research Projects

Bulk_extractor (2006-2020) — processing bulk data for digital forensics

Digital Corpora (2006-) — realistic data for digital forensics research and education

Disclosure Avoidance System for the 2020 Census (2017-2021)

PlantTracer (2023-) — Using computer vision to watch plant movement.

In 2003, I bought 200 used hard drives

The goal was to find drives that had not been properly sanitized.

First strategy:

- DD all of the disks to image files
- run **strings** to extract printable strings.
- **grep** to scan for email, CCN, etc.
  - —*VERY SLOW!!!!*
  - —*HARD TO MODIFY!*

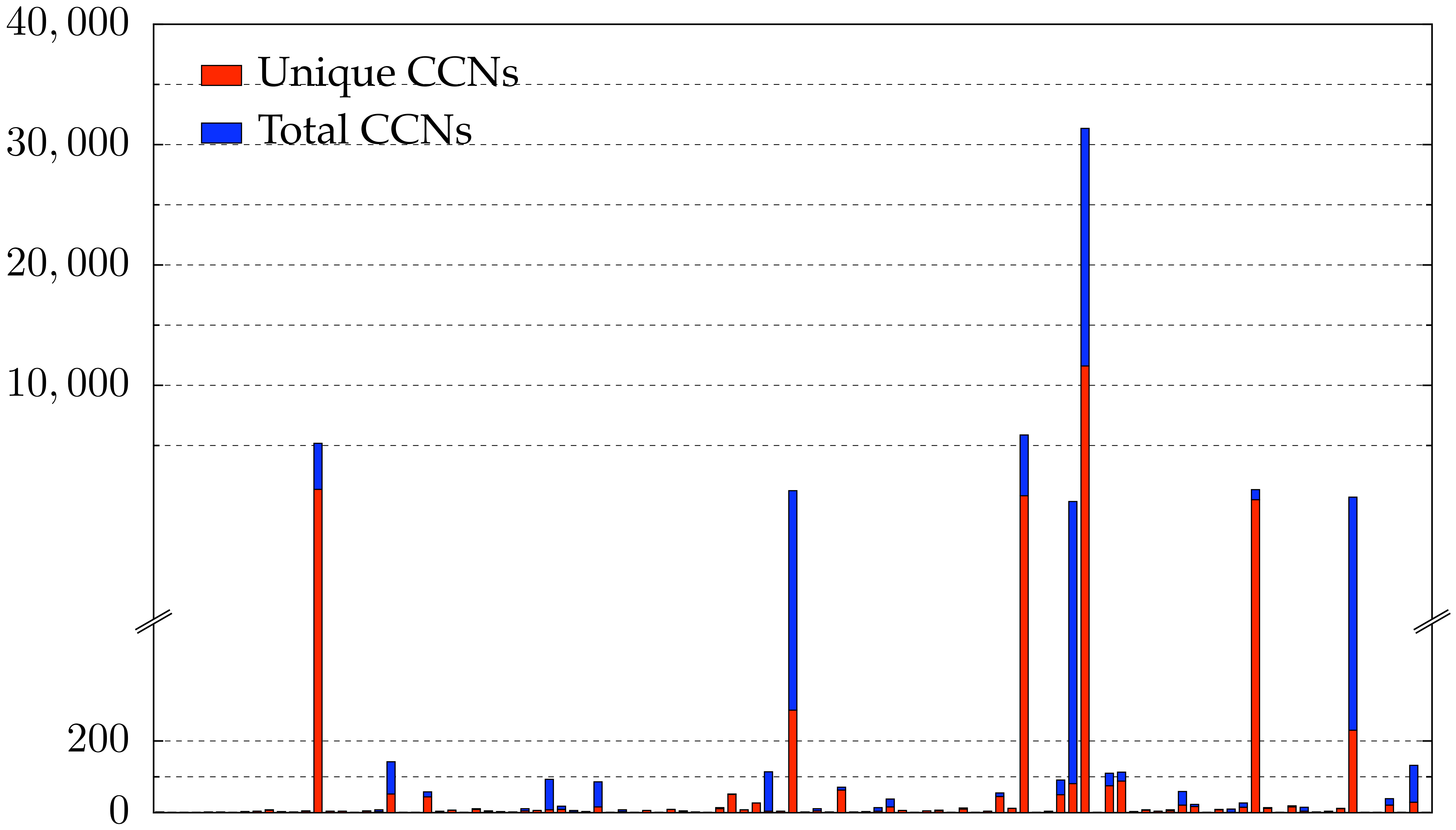
Second strategy:

- Use SBook approach!
- Read disk 1MB at a time
- Pass the *raw disk sectors* to flex-based scanner.
- Big surprise: scanner didn’t crash!

Simple flex-based scanners required substantial post-processing to be useful

Techniques include:

- Additional validation beyond regular expressions (CCN Luhn algorithm, etc).
- Examination of feature “neighborhood” to eliminate common false positives.



The technique worked well to find drives with sensitive information.

Could it be of use in digital forensics?

Between 2005 and 2008,
I interviewed law enforcement officers regarding their use of forensic tools.

Law enforcement officers wanted a *highly automated* tool for finding:

- Email addresses
- Credit card numbers (including track 2 information)
- Search terms (extracted from URLs)
- Phone numbers
- GPS coordinates
- EXIF information from JPEGs
- All words that were present on the disk (for password cracking)

https://www.americanscientist.org/article/digital-forensics

I also learned about their requirements for the user experience.

The tool had to:

- Run on Windows, Linux, and Mac-based systems
- Run with *no* user interaction
- Operate on raw disk images, split-raw volumes, E01 files, and AFF files
- Allow user to provide additional regular expressions for searches
- Automatically extract features from compressed data such as gzip-compressed HTTP
- Run at maximum I/O speed of physical drive
- Never crash

Starting in 2008, I made a series of limited releases

- January 2008 — Created Subversion Repository
- April 2010 — Initial public release - 0.1.0
- May 2010 — Initial multi-threading release - 0.3.0

  —*Each thread runs in its own process*
- Sept. 2010 — Stop lists - 0.4.0
- Oct. 2010 — Context-based stop-lists - 0.5.0
- Dec. 2010 — Switch to POSIX-based threads — 0.6.0
- Dec. 2010 — Support for Windows HIBERFIL.SYS decompression — 0.7.0
- Jun. 2010 — First 1.0.0 Release

Tool capabilities result from substantial testing and user feedback.

Moving technology from the lab to the field was challenging:

- Must work with evidence files of *any size* and on *limited hardware.*
- Users can't provide their data when the program crashes.
- Users are *analysts* and *examiners*, not engineers.

A bulk_extractor Success Story — (from 2011)

CITY OF SAN LUIS OBISPO

San Luis Obispo is
“the happiest
place in America”

Watch the video to find out why National Geographic named San Luis Obispo the top spot.

http://www.sanluisobispovacations.com/

City of San Luis Obispo Police Department, Spring 2010

District Attorney filed charges against two individuals:

- Credit Card Fraud
- Possession of materials to commit credit card fraud.

Defendants:

- Arrested with a computer.
- Expected to argue that defends were unsophisticated and lacked knowledge.

Examiner given 250GiB drive *the day before preliminary hearing.*

- Typically, it would take several days to conduct a proper forensic investigation.

bulk_extractor found actionable evidence in 2.5 hours!

Examiner given 250GiB drive *the day before preliminary hearing.*

Bulk_extractor found:

- Over 10,000 credit card numbers on the HD (1000 unique)
- Most common email address belonged to the primary defendant (possession)
- The most commonly occurring Internet search engine queries concerned credit card fraud and bank identification numbers (intent)
- Most commonly visited websites were in a foreign country whose primary language is spoken fluently by the primary defendant.

Armed with this data, the DA was able to have the defendants held.

So why was bulk_extractor a success?

Open source ⬅ *Not the whole story!*

- Government users could download it from the Internet and use it immediately.
  - *Existing authorities allowed for open source digital forensics tools to be used on specific systems.*

Plug-in architecture ⬅ *Not the story at all*

- Allowed students to create modules for student projects.
- Successful projects could be adopted into the main branch.

Delivered results that no other program could deliver

- Recursive analysis of coded and compressed data.
- Recovery of data from file fragments.

Did not compete with existing software — and other software did not compete with it!

- Because it was free, the only cost to using bulk_extractor was time and computational resources.
- Eliminates the need to implement a complete forensic stack — BE does not compete with existing tools.
  - *In fact, at least one existing tool incorporated BE into its analysis pipeline.*

Super easy-to-use!

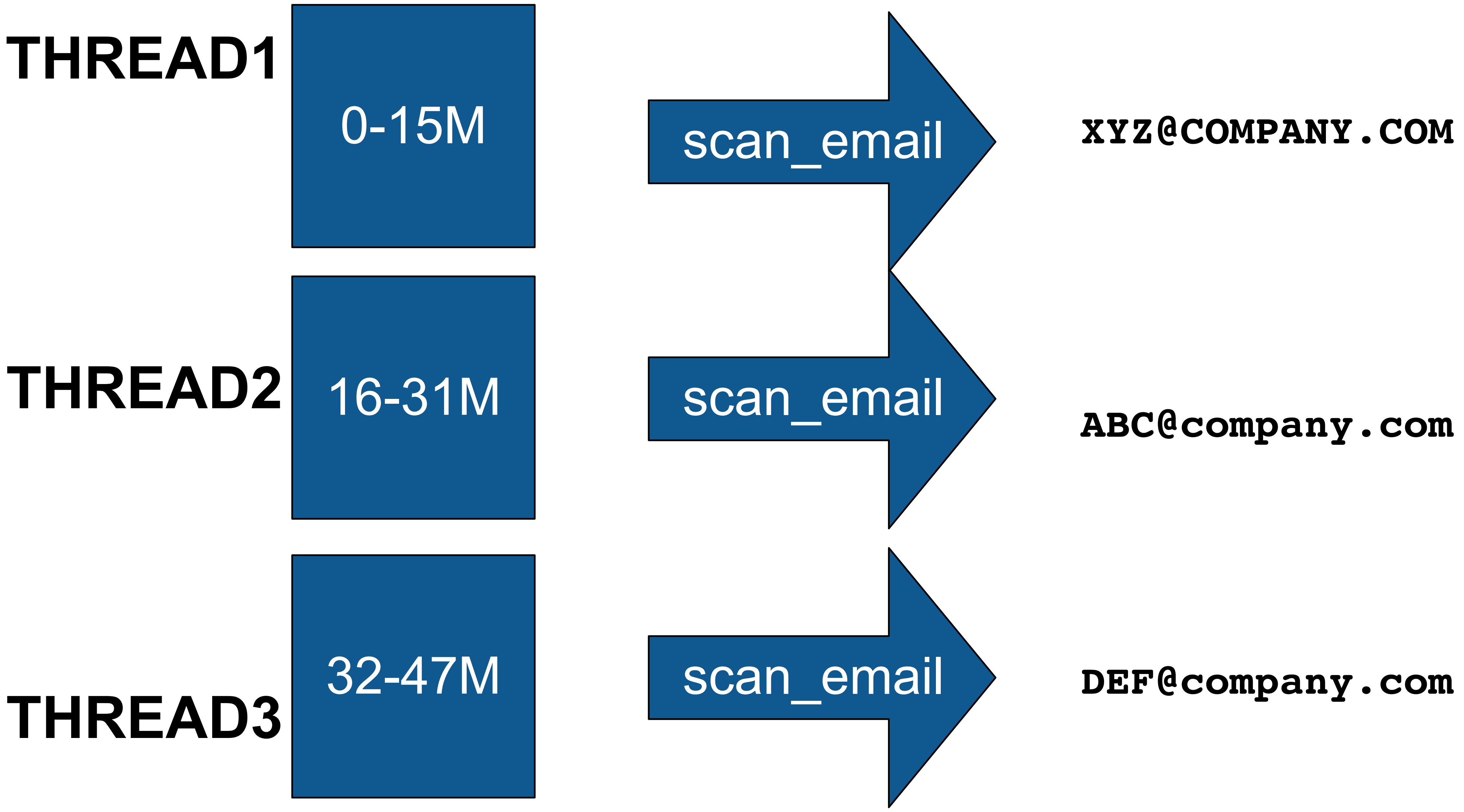
Stream-Based Disk Forensics with bulk_extractor.

Scan the disk from beginning to end; do your best.

3 hours, 20 min
to *read* the data

1. Read all of the blocks in order.
2. Look for information that might be useful.
3. Identify & extract what's possible in a single pass.

bulk_extractor splits the disk into 16M “pages” (blocks) and processes each page independently.



THREAD1 0-15M scan_email XYZ@COMPANY.COM

THREAD2 16-31M scan_email ABC@company.com

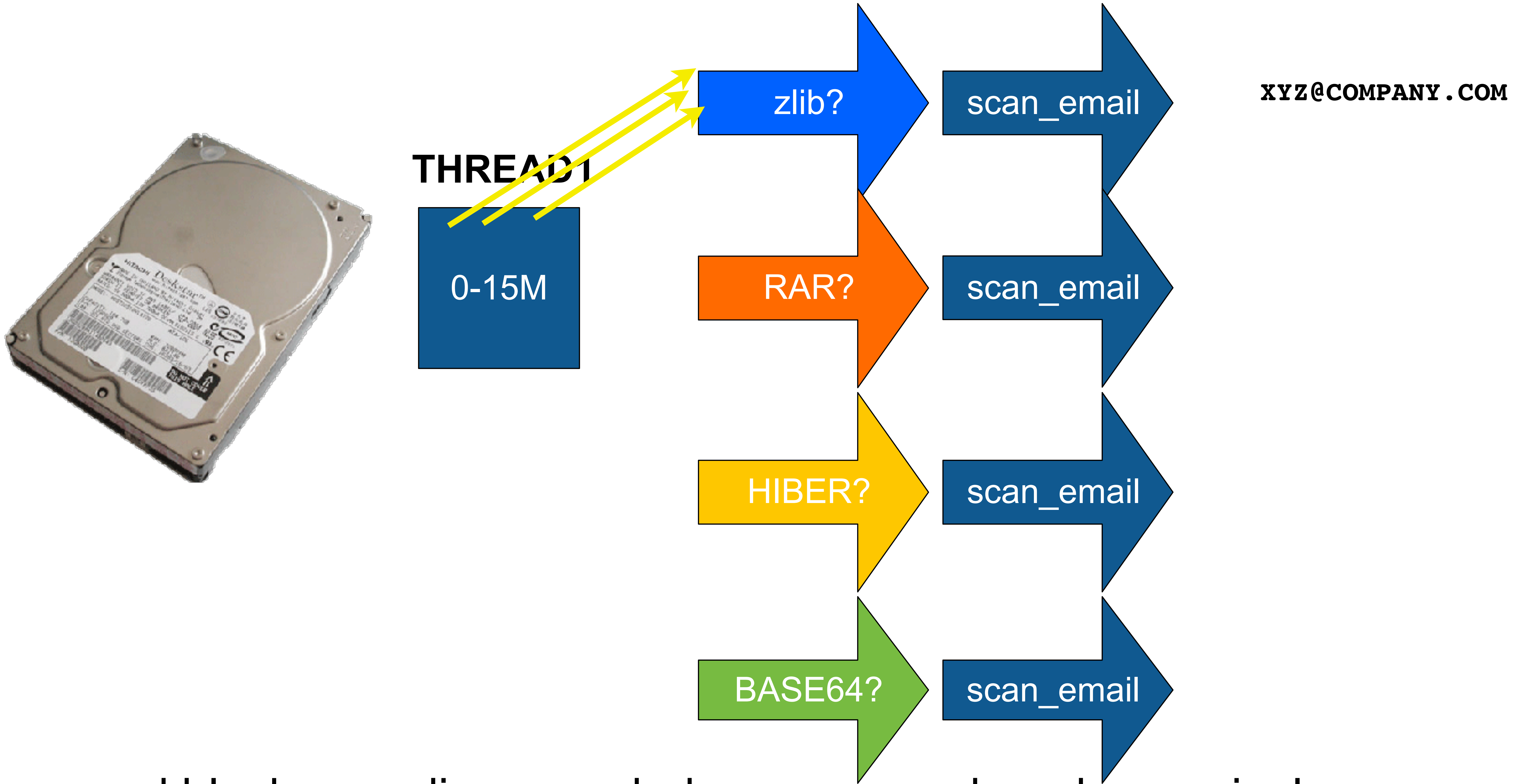
THREAD3 32-47M scan_email DEF@company.com

This finds obvious email addresses in bulk data:

```
a097 83a1 ed96 26a6 3c69 3d0f 750a 2399  ......&.<i=.u.#.
a2b5 bea7 692f 5847 a38a dd53 082c add5  ....i/XG...S.,..
5061 b64c 721d 864b 90b6 b55f bb04 735c  Pa.Lr..K..._..s\
9448 6730 5453 df64 813e b603 5795 2242  .Hg0TS.d.>..W."B
e9c8 7454 7322 7cdc b60e 97af 2f64 2728  ..tTs"|...../d'(
3cfb 84bd 2a84 2dfe 50ea 5935 c349 1513  <XYZ@COMPANY.COM
a9e9 e92c a3f8 6e46 0530 8a88 c7a2 5d2b  ...,..nF.0....]+
d89d 77cc fe1e f637 f3f3 d0af 1b47 c09b  ..w....7.....G..
```

bulk_extractor examines every byte to see if it is the beginning of an “encoded” region.

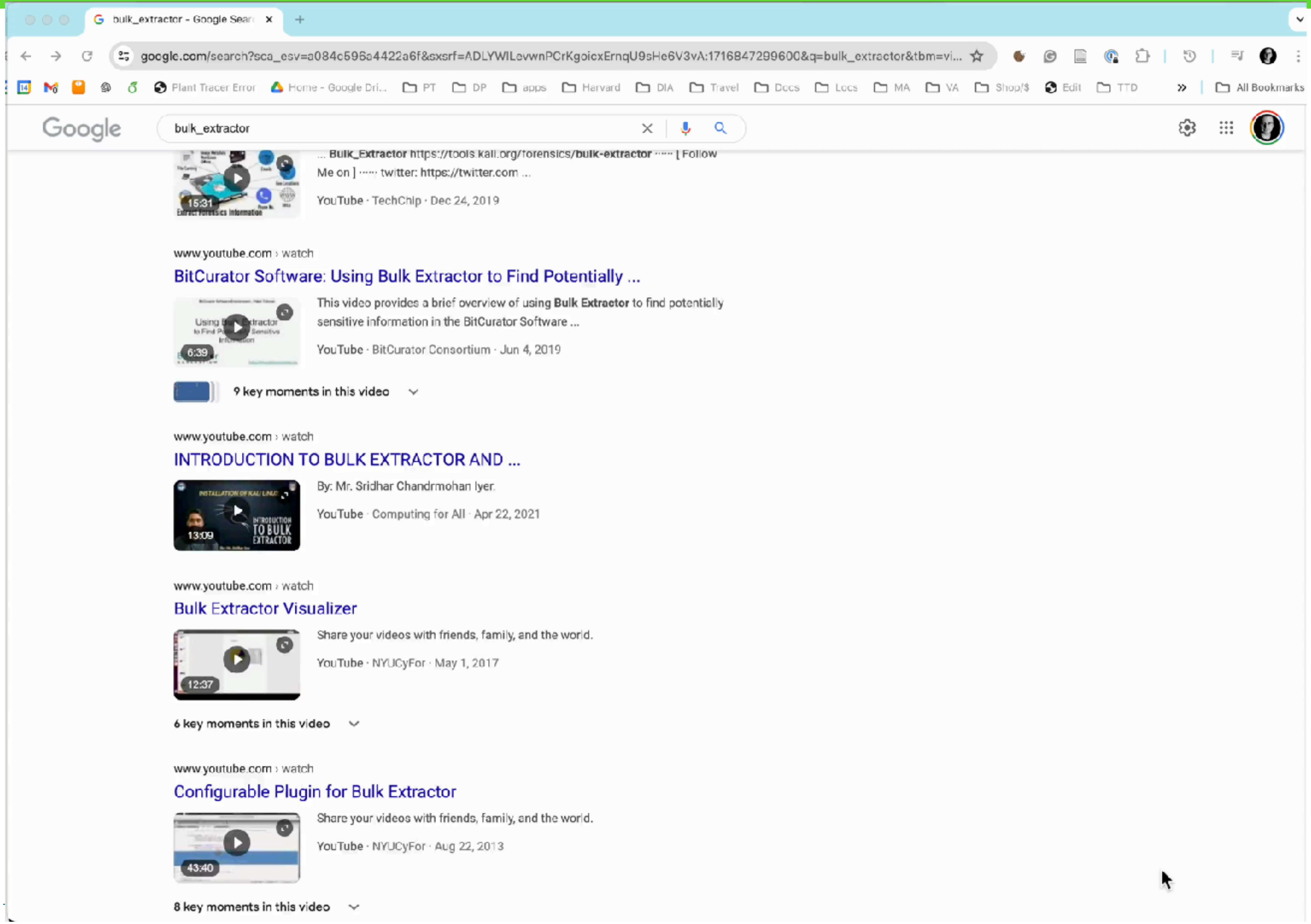
Once the region is found, it’s decoded, then processed.



Compressed blocks are discovered, decompressed, and recursively processed.

This “optimistic” approach also recovers data from fragments of files.

Students saw that open source made innovation easier!
(About half of these videos were created by students)



google.com/search?sca_esv=a084c596a4422a6f&sxsrf=ADLYWILevwnPCrKgoicxErnqU9sHe6V3vA:1716847299600&q=bulk_extractor&tbm=vi...

bulk_extractor

... **Bulk_Extractor** https://tools.kali.org/forensics/**bulk-extractor** ······ [Follow Me on] ······ twitter: https://twitter.com ...

YouTube · TechChip · Dec 24, 2019

www.youtube.com › watch

BitCurator Software: Using Bulk Extractor to Find Potentially ...

This video provides a brief overview of using **Bulk Extractor** to find potentially sensitive information in the BitCurator Software ...

YouTube · BitCurator Consortium · Jun 4, 2019

9 key moments in this video

www.youtube.com › watch

INTRODUCTION TO BULK EXTRACTOR AND ...

By: Mr. Sridhar Chandrmohan Iyer.

YouTube · Computing for All · Apr 22, 2021

www.youtube.com › watch

Bulk Extractor Visualizer

Share your videos with friends, family, and the world.

YouTube · NYUCyFor · May 1, 2017

6 key moments in this video

www.youtube.com › watch

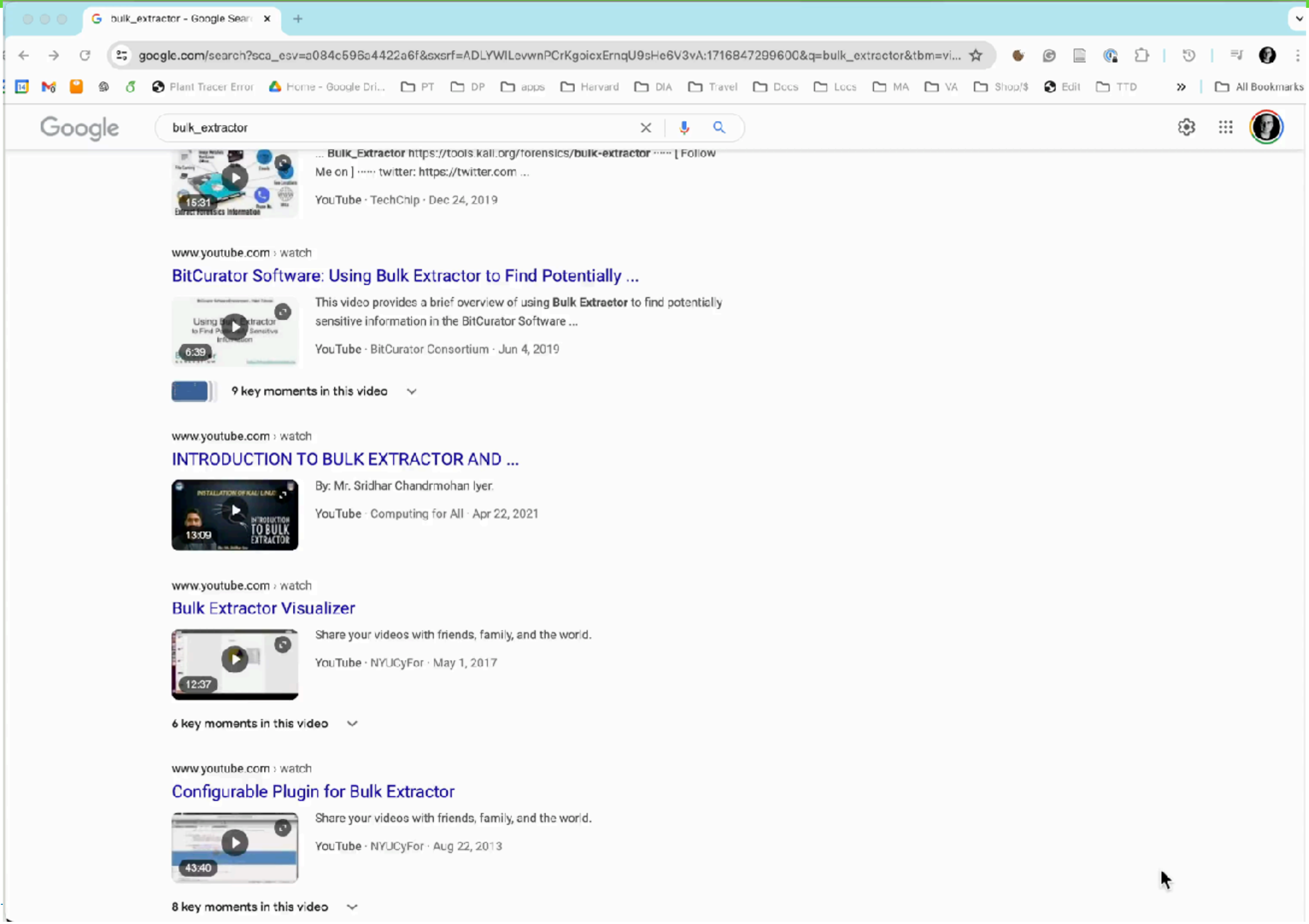
Configurable Plugin for Bulk Extractor

Share your videos with friends, family, and the world.

YouTube · NYUCyFor · Aug 22, 2013

8 key moments in this video

Students saw that open source made innovation easier!
(About half of these videos were created by students)


bulk_extractor - Google Sear
google.com/search?sca_esv=a084c596a4422a6f&sxsrf=ADLYWILevwnPCrKgoicxErnqU9sHe6V3vA:1716847299600&q=bulk_extractor&tbm=vi...
Plant Tracer Error
Home - Google Dri...
PT
DP
apps
Harvard
DIA
Travel
Docs
Locs
MA
VA
Shop/$
Edit
TTD
All Bookmarks
Google
bulk_extractor
15:31
Extract forensics information
6:39
13:09
INTRODUCTION TO BULK EXTRACTOR
12:37
43:40

... Bulk_Extractor https://tools.kali.org/forensics/bulk-extractor ······ [Follow Me on] ······ twitter: https://twitter.com ...

YouTube · TechChip · Dec 24, 2019

www.youtube.com › watch

BitCurator Software: Using Bulk Extractor to Find Potentially ...

This video provides a brief overview of using Bulk Extractor to find potentially sensitive information in the BitCurator Software ...

YouTube · BitCurator Consortium · Jun 4, 2019

9 key moments in this video

www.youtube.com › watch

INTRODUCTION TO BULK EXTRACTOR AND ...

By: Mr. Sridhar Chandrmohan Iyer.

YouTube · Computing for All · Apr 22, 2021

www.youtube.com › watch

Bulk Extractor Visualizer

Share your videos with friends, family, and the world.

YouTube · NYUCyFor · May 1, 2017

6 key moments in this video

www.youtube.com › watch

Configurable Plugin for Bulk Extractor

Share your videos with friends, family, and the world.

YouTube · NYUCyFor · Aug 22, 2013

8 key moments in this video

I used bulk_extractor as a platform for research and education

2012 — Statistical sampling breakthrough

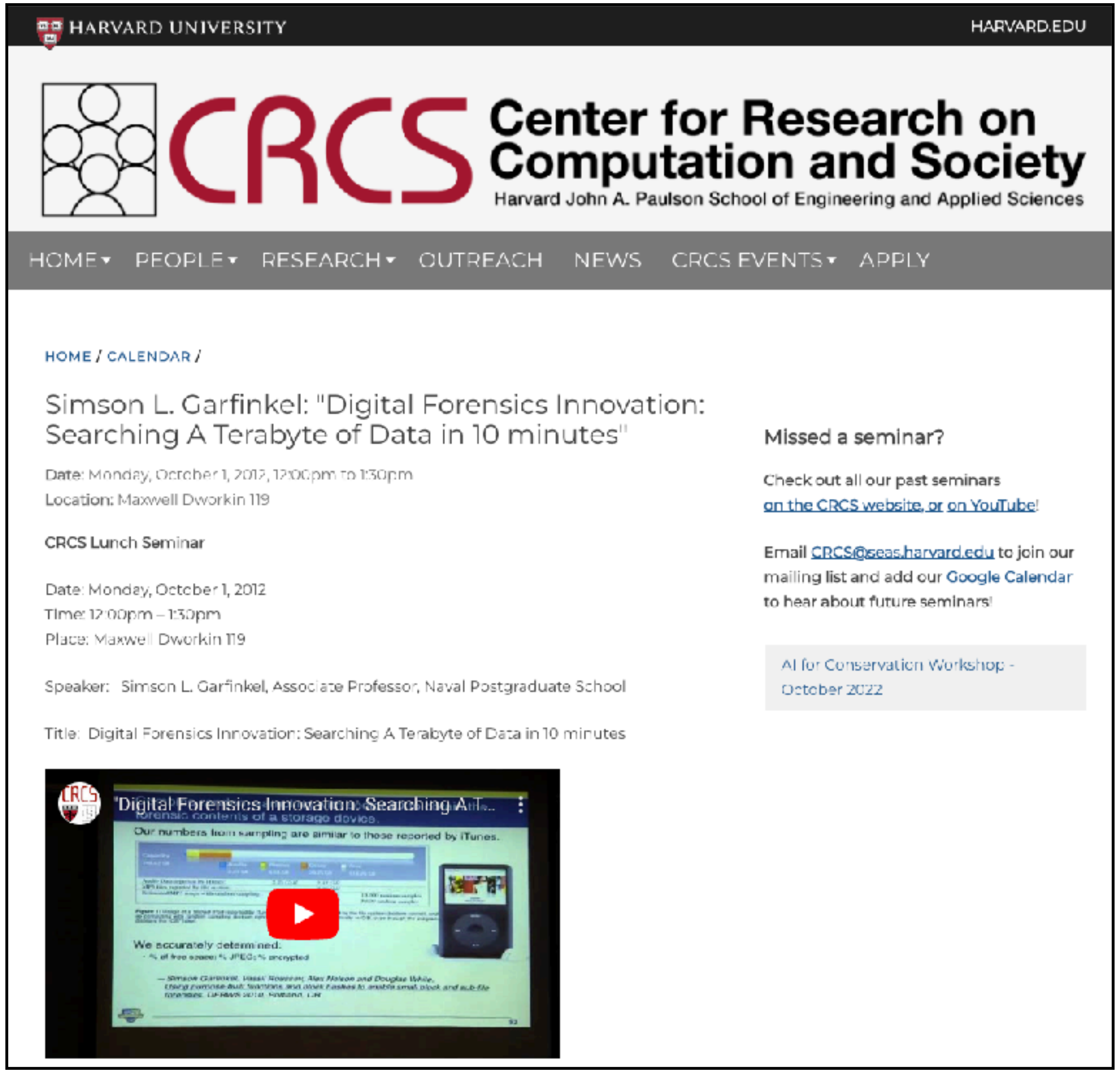
- US Patent 8,433,959 granted April 30, 2013

By 2016 bulk_extractor was

- FBI approved tool
- Incorporated into two products — one commercial, one GOTS (government-off-the-shelf).
- Widely used in digital forensics education.
- Incorporated into multiple digital forensics boot DVDs.

Bulk_extractor — helped teach students to innovate

- Showed students how to introduce advanced technology into US Government agencies that were resistant to change.
- Provided a testbed for students to develop their own modules.
- Showed how to pitch sponsored research and transition it to the field.



HARVARD UNIVERSITY HARVARD.EDU

CRCS Center for Research on Computation and Society

Harvard John A. Paulson School of Engineering and Applied Sciences

Simson L. Garfinkel: "Digital Forensics Innovation: Searching A Terabyte of Data in 10 minutes"

Date: Monday, October 1, 2012, 12:00pm to 1:30pm
Location: Maxwell Dworkin 119

CRCS Lunch Seminar

Date: Monday, October 1, 2012
Time: 12:00pm – 1:30pm
Place: Maxwell Dworkin 119

Speaker: Simson L. Garfinkel, Associate Professor, Naval Postgraduate School

Title: Digital Forensics Innovation: Searching A Terabyte of Data in 10 minutes

Missed a seminar?

Check out all our past seminars on the CRCS website, or on YouTube!

Email CRCS@seas.harvard.edu to join our mailing list and add our Google Calendar to hear about future seminars!

AI for Conservation Workshop - October 2022

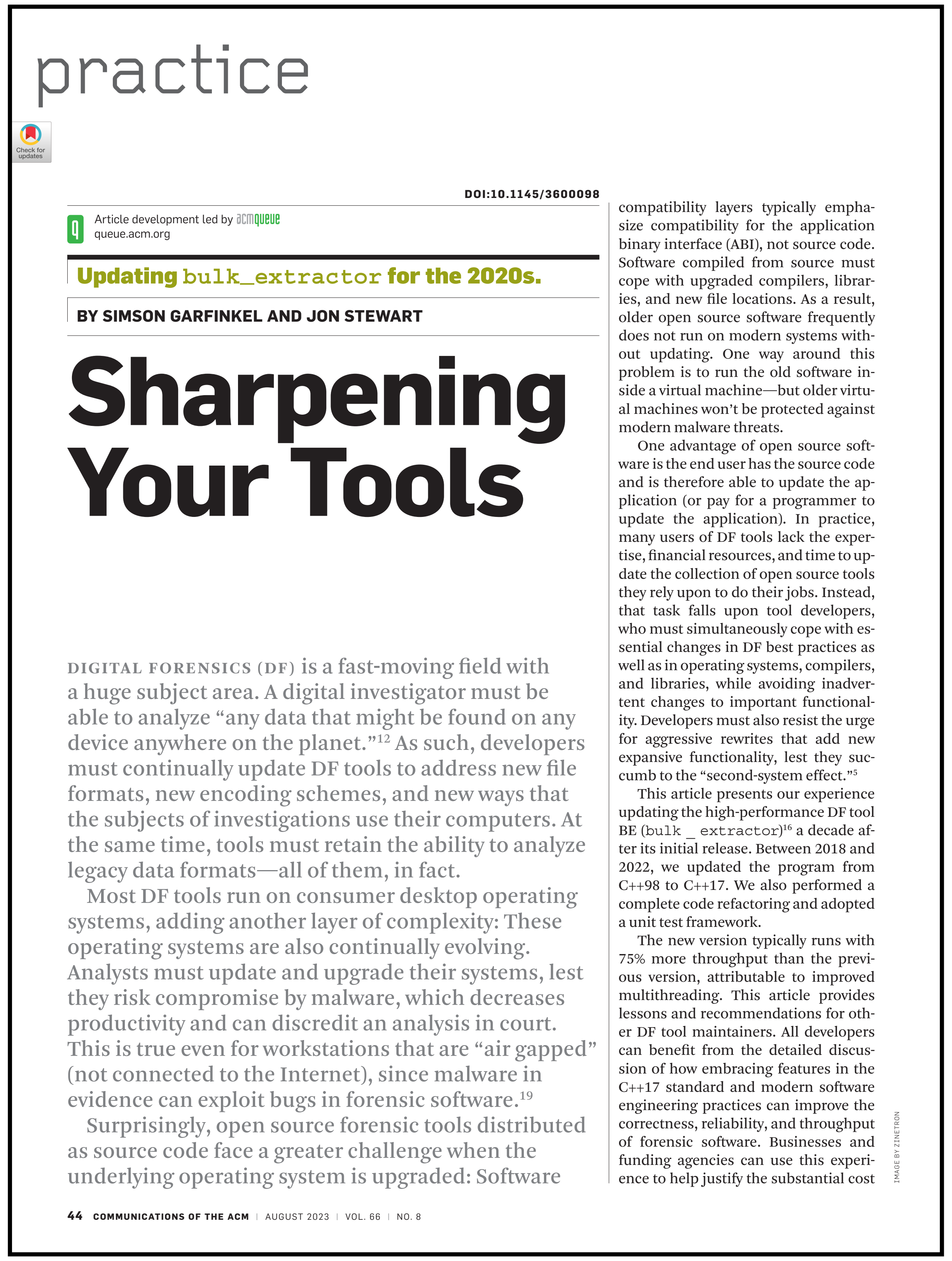
"Digital Forensics Innovation: Searching A T...
forensic contents of a storage device.
Our numbers from sampling are similar to those reported by iTunes.
We accurately determined:

October 1, 2012

https://crcs.seas.harvard.edu/event/simson-l-garfinkel-digital-forensics-innovation-searching-terabyte-data-10-minutes

Digital forensics tools require constant maintenance.

OS Creep • Language creep • Forensic science progress • O&M (operations & maintenance) “tail”



practice

Check for updates

DOI:10.1145/3600098

Article development led by acmqueue
queue.acm.org

Updating `bulk_extractor` for the 2020s.

BY SIMSON GARFINKEL AND JON STEWART

Sharpening Your Tools

DIGITAL FORENSICS (DF) is a fast-moving field with a huge subject area. A digital investigator must be able to analyze “any data that might be found on any device anywhere on the planet.”[12] As such, developers must continually update DF tools to address new file formats, new encoding schemes, and new ways that the subjects of investigations use their computers. At the same time, tools must retain the ability to analyze legacy data formats—all of them, in fact.

Most DF tools run on consumer desktop operating systems, adding another layer of complexity: These operating systems are also continually evolving. Analysts must update and upgrade their systems, lest they risk compromise by malware, which decreases productivity and can discredit an analysis in court. This is true even for workstations that are “air gapped” (not connected to the Internet), since malware in evidence can exploit bugs in forensic software.[19]

Surprisingly, open source forensic tools distributed as source code face a greater challenge when the underlying operating system is upgraded: Software compatibility layers typically emphasize compatibility for the application binary interface (ABI), not source code. Software compiled from source must cope with upgraded compilers, libraries, and new file locations. As a result, older open source software frequently does not run on modern systems without updating. One way around this problem is to run the old software inside a virtual machine—but older virtual machines won’t be protected against modern malware threats.

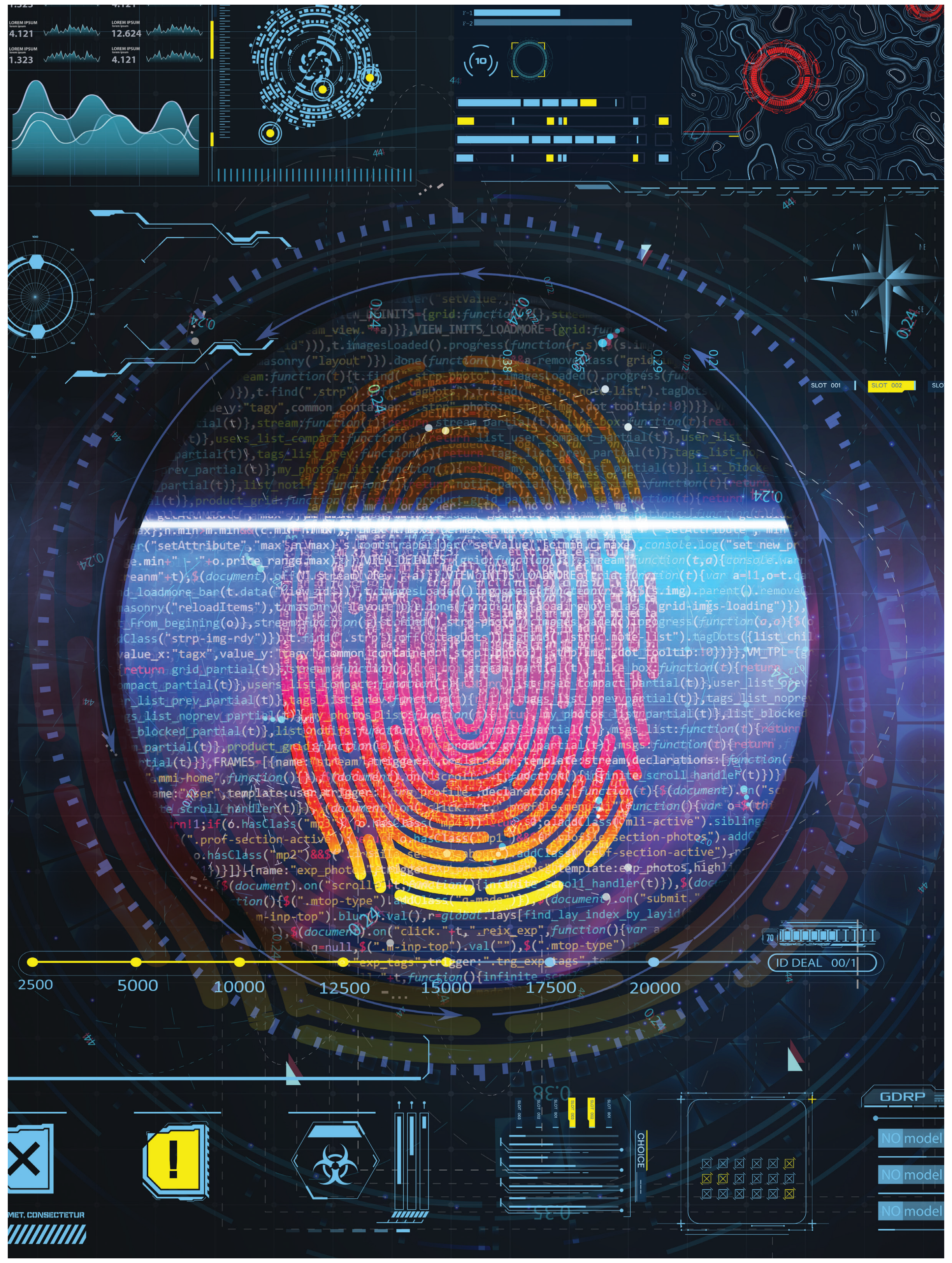
One advantage of open source software is the end user has the source code and is therefore able to update the application (or pay for a programmer to update the application). In practice, many users of DF tools lack the expertise, financial resources, and time to update the collection of open source tools they rely upon to do their jobs. Instead, that task falls upon tool developers, who must simultaneously cope with essential changes in DF best practices as well as in operating systems, compilers, and libraries, while avoiding inadvertent changes to important functionality. Developers must also resist the urge for aggressive rewrites that add new expansive functionality, lest they succumb to the “second-system effect.”[5]

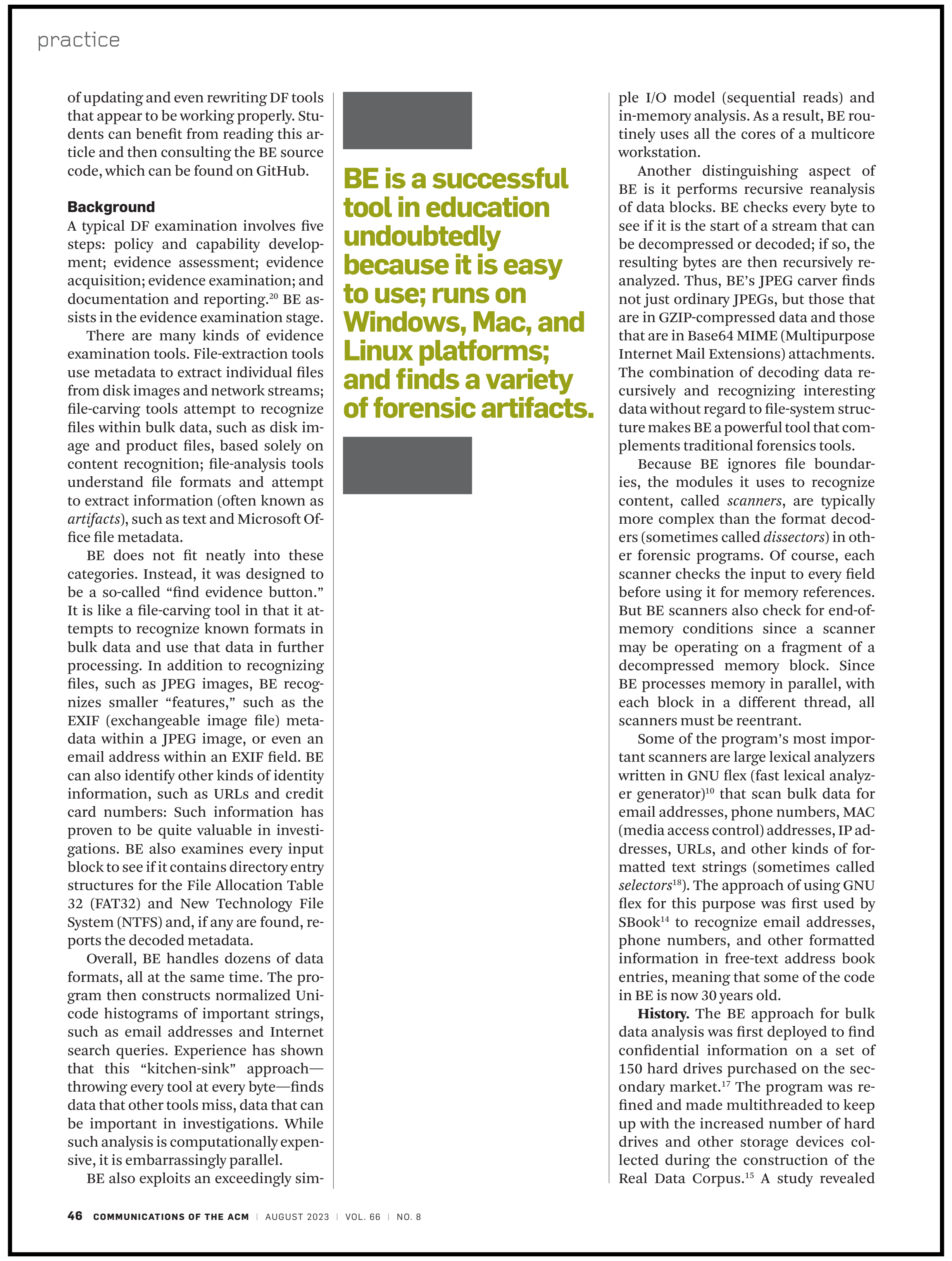
This article presents our experience updating the high-performance DF tool BE (`bulk _ extractor`)[16] a decade after its initial release. Between 2018 and 2022, we updated the program from C++98 to C++17. We also performed a complete code refactoring and adopted a unit test framework.

The new version typically runs with 75% more throughput than the previous version, attributable to improved multithreading. This article provides lessons and recommendations for other DF tool maintainers. All developers can benefit from the detailed discussion of how embracing features in the C++17 standard and modern software engineering practices can improve the correctness, reliability, and throughput of forensic software. Businesses and funding agencies can use this experience to help justify the substantial cost

IMAGE BY ZINETRON

44 COMMUNICATIONS OF THE ACM | AUGUST 2023 | VOL. 66 | NO. 8





practice

of updating and even rewriting DF tools that appear to be working properly. Students can benefit from reading this article and then consulting the BE source code, which can be found on GitHub.

Background

A typical DF examination involves five steps: policy and capability development; evidence assessment; evidence acquisition; evidence examination; and documentation and reporting.[20] BE assists in the evidence examination stage.

There are many kinds of evidence examination tools. File-extraction tools use metadata to extract individual files from disk images and network streams; file-carving tools attempt to recognize files within bulk data, such as disk image and product files, based solely on content recognition; file-analysis tools understand file formats and attempt to extract information (often known as *artifacts*), such as text and Microsoft Office file metadata.

BE does not fit neatly into these categories. Instead, it was designed to be a so-called “find evidence button.” It is like a file-carving tool in that it attempts to recognize known formats in bulk data and use that data in further processing. In addition to recognizing files, such as JPEG images, BE recognizes smaller “features,” such as the EXIF (exchangeable image file) metadata within a JPEG image, or even an email address within an EXIF field. BE can also identify other kinds of identity information, such as URLs and credit card numbers: Such information has proven to be quite valuable in investigations. BE also examines every input block to see if it contains directory entry structures for the File Allocation Table 32 (FAT32) and New Technology File System (NTFS) and, if any are found, reports the decoded metadata.

Overall, BE handles dozens of data formats, all at the same time. The program then constructs normalized Unicode histograms of important strings, such as email addresses and Internet search queries. Experience has shown that this “kitchen-sink” approach—throwing every tool at every byte—finds data that other tools miss, data that can be important in investigations. While such analysis is computationally expensive, it is embarrassingly parallel.

BE also exploits an exceedingly simple I/O model (sequential reads) and in-memory analysis. As a result, BE routinely uses all the cores of a multicore workstation.

BE is a successful tool in education undoubtedly because it is easy to use; runs on Windows, Mac, and Linux platforms; and finds a variety of forensic artifacts.

Another distinguishing aspect of BE is it performs recursive reanalysis of data blocks. BE checks every byte to see if it is the start of a stream that can be decompressed or decoded; if so, the resulting bytes are then recursively reanalyzed. Thus, BE’s JPEG carver finds not just ordinary JPEGs, but those that are in GZIP-compressed data and those that are in Base64 MIME (Multipurpose Internet Mail Extensions) attachments. The combination of decoding data recursively and recognizing interesting data without regard to file-system structure makes BE a powerful tool that complements traditional forensics tools.

Because BE ignores file boundaries, the modules it uses to recognize content, called *scanners*, are typically more complex than the format decoders (sometimes called *dissectors*) in other forensic programs. Of course, each scanner checks the input to every field before using it for memory references. But BE scanners also check for end-of-memory conditions since a scanner may be operating on a fragment of a decompressed memory block. Since BE processes memory in parallel, with each block in a different thread, all scanners must be reentrant.

Some of the program’s most important scanners are large lexical analyzers written in GNU flex (fast lexical analyzer generator)[10] that scan bulk data for email addresses, phone numbers, MAC (media access control) addresses, IP addresses, URLs, and other kinds of formatted text strings (sometimes called *selectors*[18]). The approach of using GNU flex for this purpose was first used by SBook[14] to recognize email addresses, phone numbers, and other formatted information in free-text address book entries, meaning that some of the code in BE is now 30 years old.

History. The BE approach for bulk data analysis was first deployed to find confidential information on a set of 150 hard drives purchased on the secondary market.[17] The program was refined and made multithreaded to keep up with the increased number of hard drives and other storage devices collected during the construction of the Real Data Corpus.[15] A study revealed

46 COMMUNICATIONS OF THE ACM | AUGUST 2023 | VOL. 66 | NO. 8

From 2018-2021 I upgraded BE:

- Moved from C++ STL to C++17
- Added CI/CD testing on GitHub
- Improved multi-threading for modern CPUs.
- Wrote an article for *ACM Queue* and *Communications of the ACM*.

Research Projects

Bulk_extractor (2006-2020) — processing bulk data for digital forensics

Digital Corpora (2006-) — realistic data for digital forensics research and education

Disclosure Avoidance System for the 2020 Census (2017-2021)

PlantTracer (2023-) — Using computer vision to watch plant movement.

Digital forensics research and education had a data problem in 2009.

Digital forensics practitioners must be able to

- analyze *any* digital data,
- from any computer,
- that has ever been used,
- anywhere.

The data problem: getting data that are **ecologically valid**

- **representative** of the diversity of systems found on computers collected by law enforcement and defense practitioners.
- **complex enough** to present students and researchers with more than toy problems.
- **simple enough** that the problems can be solved in *hours or days*, rather than *weeks or months*.

The solution

- Get students to create complex scenarios *as a learning exercise.*
- Allow free downloads of the dataset.
- Track usage through the “teacher’s solutions.”

I created the Digital Corpora —
a collection of complex digital artifacts for forensics education and tool testing.

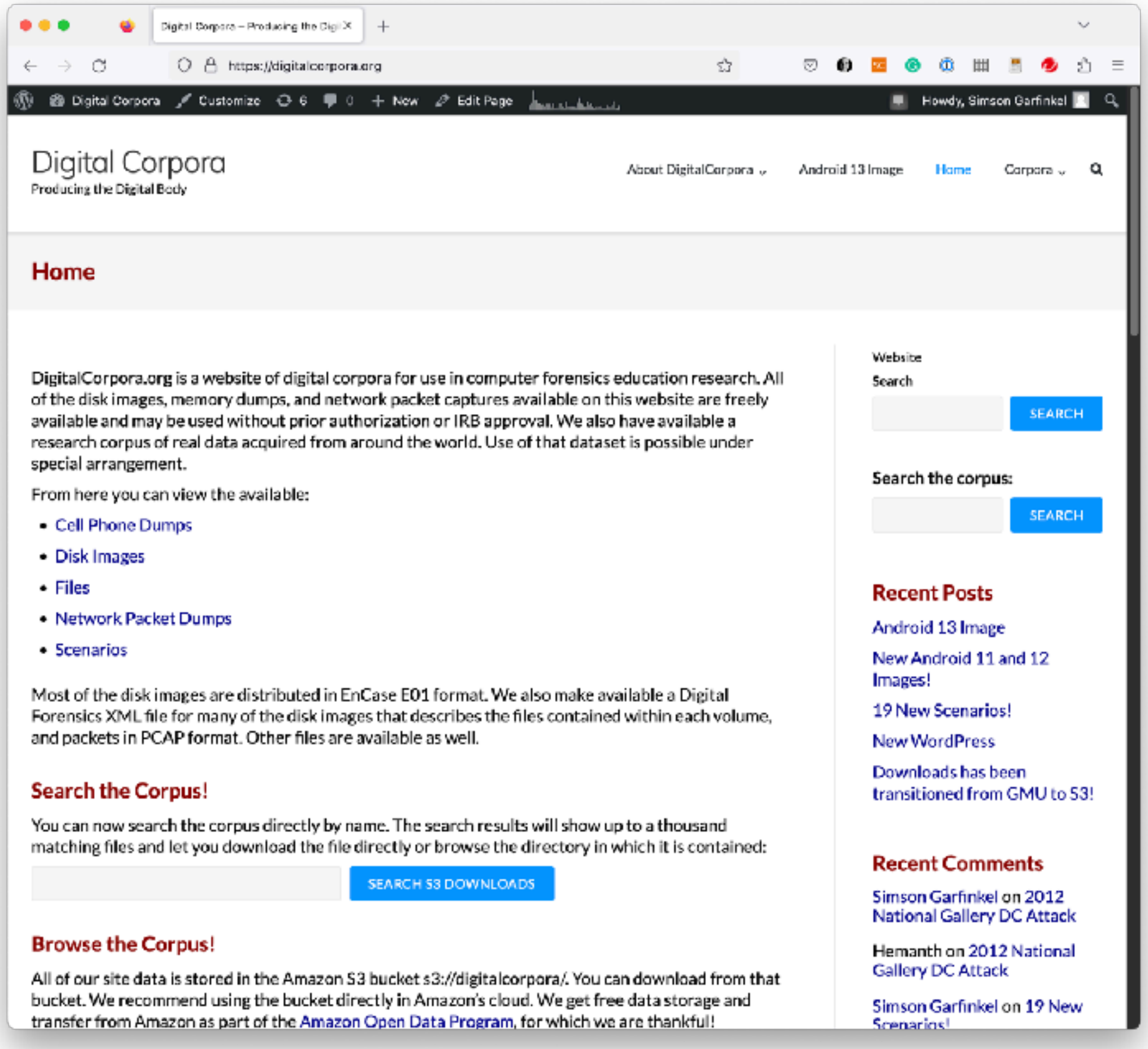
https://digitalcorpora.org/

Initial funding:

- NIST/NPS Inter Agency Agreement
- NSF Grant No. 0919593

Today:

- Scenarios and data contributed by cybersecurity programs and practitioners all over the world
- Corpus hosting by Amazon's Open Data Sponsorship Program



Digital Corpora

Producing the Digital Body

Home

DigitalCorpora.org is a website of digital corpora for use in computer forensics education research. All of the disk images, memory dumps, and network packet captures available on this website are freely available and may be used without prior authorization or IRB approval. We also have available a research corpus of real data acquired from around the world. Use of that dataset is possible under special arrangement.

From here you can view the available:

- Cell Phone Dumps
- Disk Images
- Files
- Network Packet Dumps
- Scenarios

Most of the disk images are distributed in EnCase E01 format. We also make available a Digital Forensics XML file for many of the disk images that describes the files contained within each volume, and packets in PCAP format. Other files are available as well.

Search the Corpus!

You can now search the corpus directly by name. The search results will show up to a thousand matching files and let you download the file directly or browse the directory in which it is contained:

SEARCH S3 DOWNLOADS

Browse the Corpus!

All of our site data is stored in the Amazon S3 bucket s3://digitalcorpora/. You can download from that bucket. We recommend using the bucket directly in Amazon's cloud. We get free data storage and transfer from Amazon as part of the Amazon Open Data Program, for which we are thankful!

Website
Search

SEARCH

Search the corpus:

SEARCH

Recent Posts

- Android 13 Image
- New Android 11 and 12 Images!
- 19 New Scenarios!
- New WordPress
- Downloads has been transitioned from GMU to S3!

Recent Comments

- Simson Garfinkel on 2012 National Gallery DC Attack
- Hemanth on 2012 National Gallery DC Attack
- Simson Garfinkel on 19 New Scenarios!

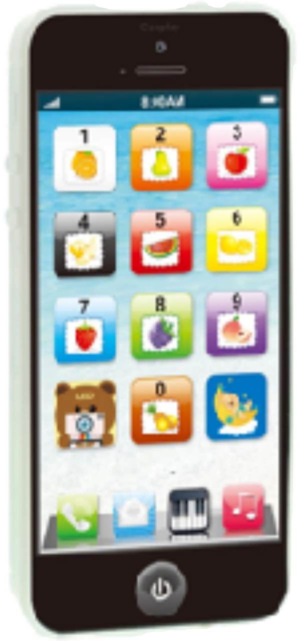
The corpus has many scenario-based digital artifacts.

Complex, deep datasets

- Scripted scenarios.
- Multiple characters with clearly defined motivations
- Specific challenges for the investigator to uncover
- Multiple problems that require different levels of skill and analysis to solve
- Created in “real-time” over weeks or months
- “Teachers guides” and “solutions” are available for many of the datasets.

Multi-modality

- Disk images
- Cell phone images
- Memory dumps
- Log files from servers
- Packet dumps (wiretaps)



A few scenarios in the corpus available for download

A “Lone Wolf” who becomes self-radicalized on YouTube and plans a school shooting.

- He was turned in by his brother.
- You have the laptop
- https://downloads.digitalcorpora.org/corpora/scenarios/2018-lonewolf/

A macOS/iOS terrorist recruitment scenario with multiple personas and international travel

- Picked up by FBI
- You have the Mac and iPod Touch backup
- https://downloads.digitalcorpora.org/corpora/scenarios/2019-tuck/

A planned defacement of art at the DC National Gallery by a direct action group, combined with a nasty divorce proceeding.

- You have disk images, phone images, captured packets, and a bungled wiretap
- https://downloads.digitalcorpora.org/corpora/scenarios/2012-ngdc/

+ many others contributed by educators around the world.

Constructed, scenario-based artifacts are better for research and education.

No privacy-sensitive data! No PII!

- Computer users are not real people, they are personas

No pornography! No illegal content!

- We know that there’s no pornography in the data
- Especially an issue with students under 18 years old

No child exploitation scenarios!

- CSAM scenarios are a big turn-off!

There are solutions!

- Solutions are distributed on the website as encrypted PDFs
- Decrypt keys are available on a case-by-case basis to faculty at accredited institutions, law enforcement, and partners

GOVDOCS1M — The first ecologically valid “files” corpus.

Developed in 2008, a corpus of 1 million files downloaded from US Government web servers.

- US Government websites to avoid copyright issue.

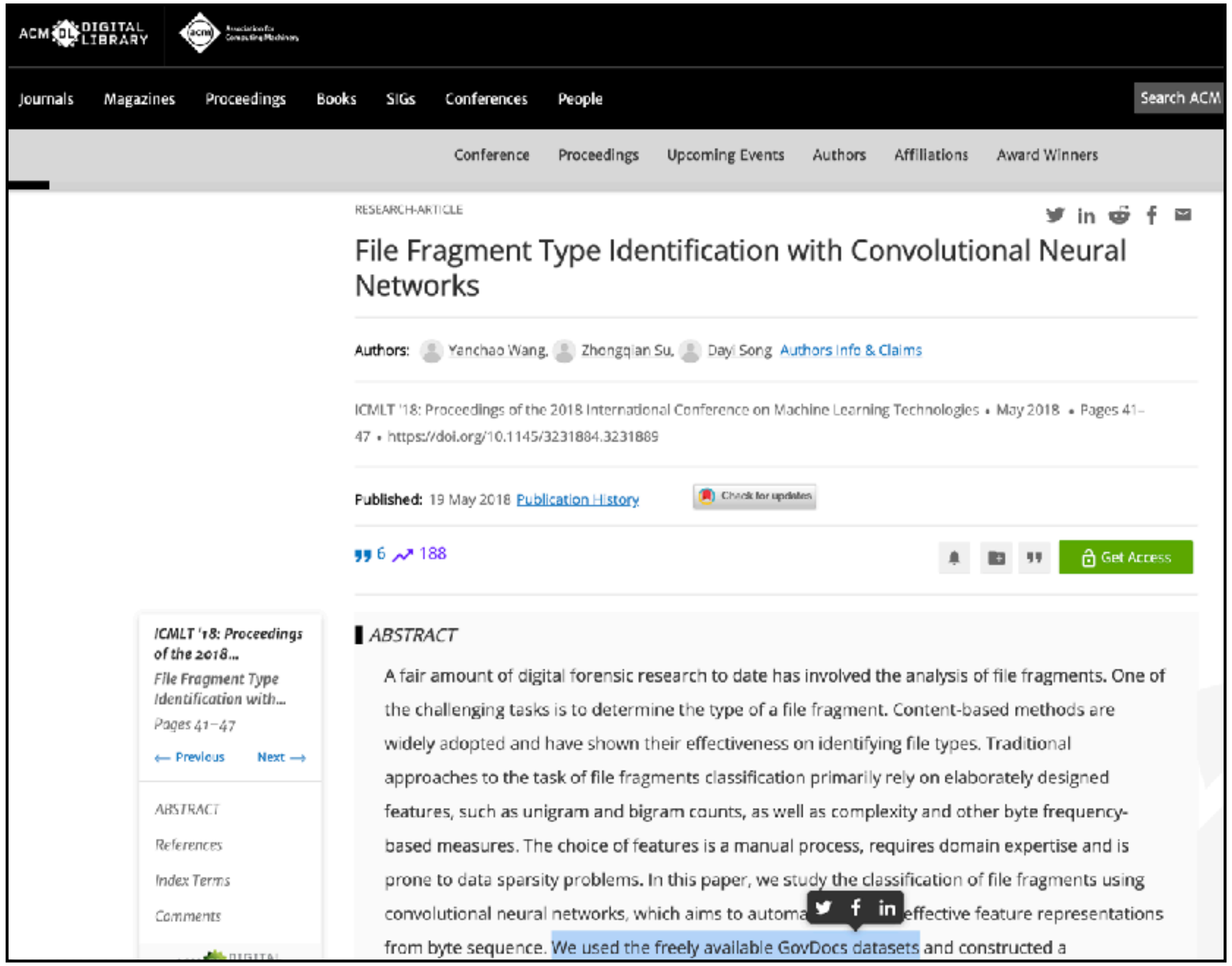
Includes:

- Image formats (JPEG, TIFF, PNG, etc)
- Document formats (PDF, MSOffice)
- Text files
- Log files
- SQL dumps

At the time, this let me teach…

- Approaches for working within the copyright law
- How to handle legal missteps
- Scientific principles of reproducibility

… by sharing the issues with students



File Fragment Type Identification with Convolutional Neural Networks

Authors: Yanchao Wang, Zhongqian Su, Dayi Song Authors Info & Claims

ICMLT '18: Proceedings of the 2018 International Conference on Machine Learning Technologies • May 2018 • Pages 41–47 • https://doi.org/10.1145/3231884.3231889

Published: 19 May 2018 Publication History

Check for updates

ICMLT '18: Proceedings of the 2018...
File Fragment Type Identification with...
Pages 41–47

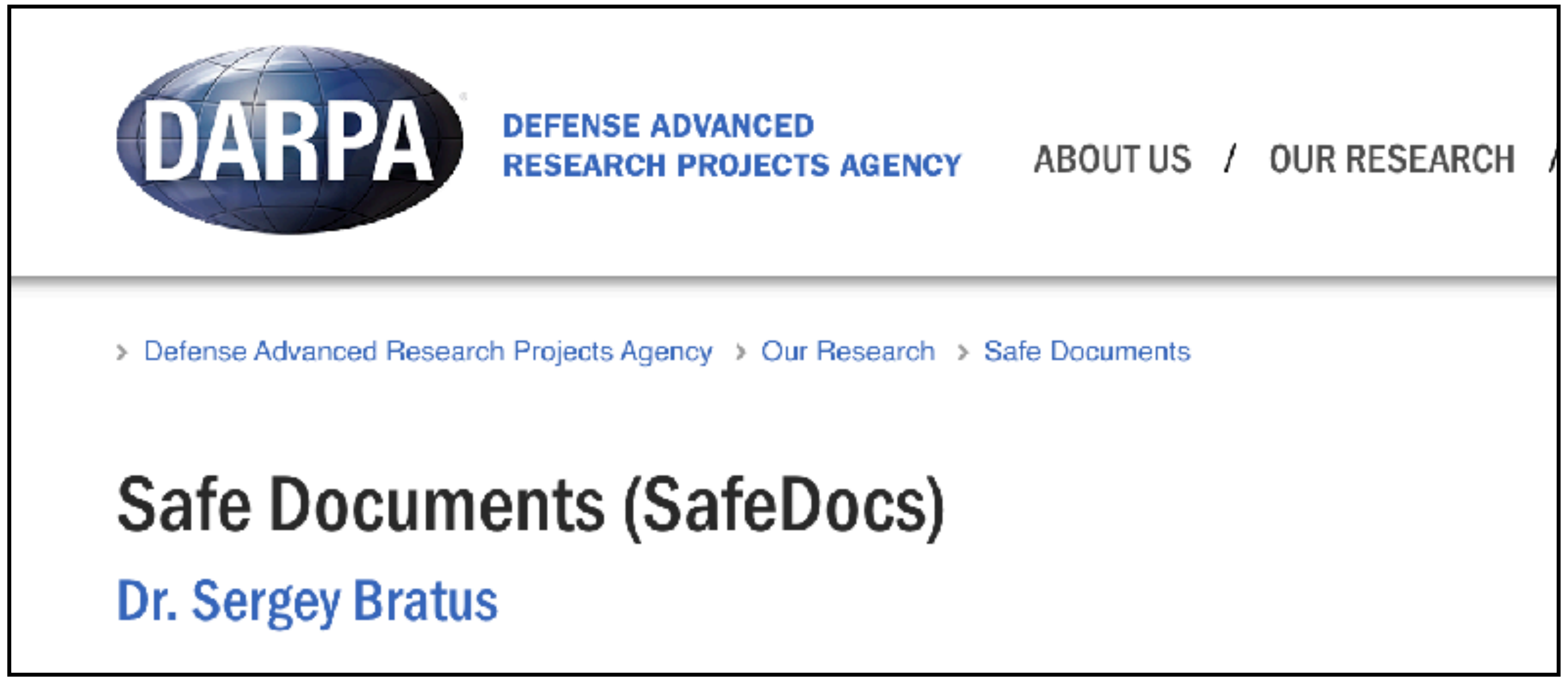
ABSTRACT

A fair amount of digital forensic research to date has involved the analysis of file fragments. One of the challenging tasks is to determine the type of a file fragment. Content-based methods are widely adopted and have shown their effectiveness on identifying file types. Traditional approaches to the task of file fragments classification primarily rely on elaborately designed features, such as unigram and bigram counts, as well as complexity and other byte frequency-based measures. The choice of features is a manual process, requires domain expertise and is prone to data sparsity problems. In this paper, we study the classification of file fragments using convolutional neural networks, which aims to automa effective feature representations from byte sequence. We used the freely available GovDocs datasets and constructed a

(one of many research articles have used the corpus.)

GOVDOCS was the seed for the DARPA SafeDocs program

Goal of SafeDocs: build an exploit-proof PDF reader using formal methods.


DARPA
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
ABOUT US / OUR RESEARCH /
› Defense Advanced Research Projects Agency › Our Research › Safe Documents

Safe Documents (SafeDocs)

Dr. Sergey Bratus

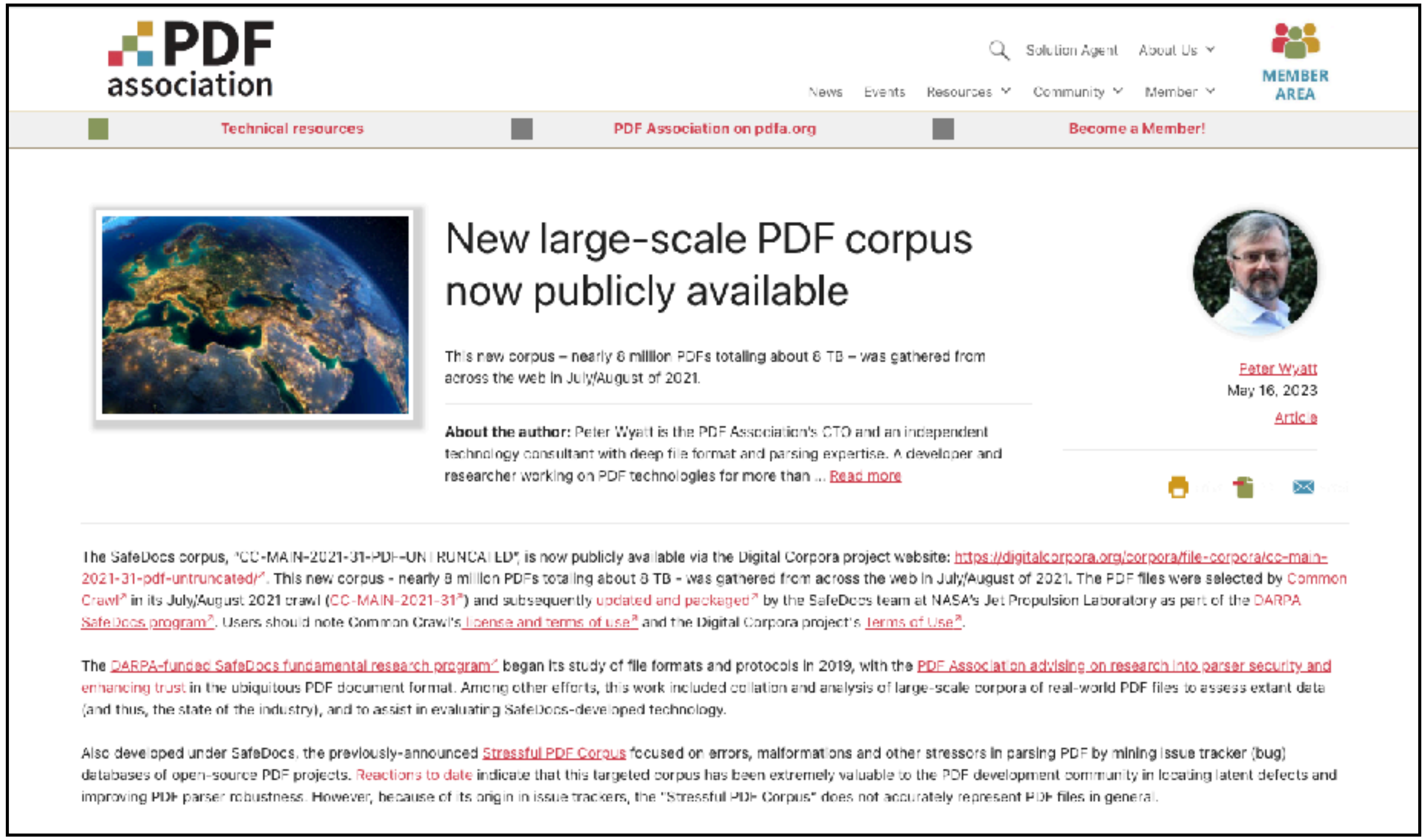
2018-2023

When SafeDocs shut down, DARPA donated 8M PDFs to the Digital Corpora

- SafeDocs became open data!

We now have 24TB of data…

- We had to be entrepreneurial in dealing with storage requirements!
- Today we are hosted by Amazon's Open Data program.
- With minimal copyright and privacy issues, this Internet snapshot can power the creation of tools for the digital humanities.



New large-scale PDF corpus now publicly available

This new corpus – nearly 8 million PDFs totaling about 8 TB – was gathered from across the web in July/August of 2021.

About the author: Peter Wyatt is the PDF Association's CTO and an independent technology consultant with deep file format and parsing expertise. A developer and researcher working on PDF technologies for more than … Read more

Peter Wyatt
May 16, 2023
Article

The SafeDocs corpus, "CC-MAIN-2021-31-PDF-UNTRUNCATED", is now publicly available via the Digital Corpora project website: https://digitalcorpora.org/corpora/file-corpora/cc-main-2021-31-pdf-untruncated/. This new corpus - nearly 8 million PDFs totaling about 8 TB - was gathered from across the web in July/August of 2021. The PDF files were selected by Common Crawl in its July/August 2021 crawl (CC-MAIN-2021-31) and subsequently updated and packaged by the SafeDocs team at NASA's Jet Propulsion Laboratory as part of the DARPA SafeDocs program. Users should note Common Crawl's license and terms of use and the Digital Corpora project's Terms of Use.

The DARPA-funded SafeDocs fundamental research program began its study of file formats and protocols in 2019, with the PDF Association advising on research into parser security and enhancing trust in the ubiquitous PDF document format. Among other efforts, this work included collation and analysis of large-scale corpora of real-world PDF files to assess extant data (and thus, the state of the industry), and to assist in evaluating SafeDocs-developed technology.

Also developed under SafeDocs, the previously-announced Stressful PDF Corpus focused on errors, malformations and other stressors in parsing PDF by mining issue tracker (bug) databases of open-source PDF projects. Reactions to date indicate that this targeted corpus has been extremely valuable to the PDF development community in locating latent defects and improving PDF parser robustness. However, because of its origin in issue trackers, the "Stressful PDF Corpus" does not accurately represent PDF files in general.

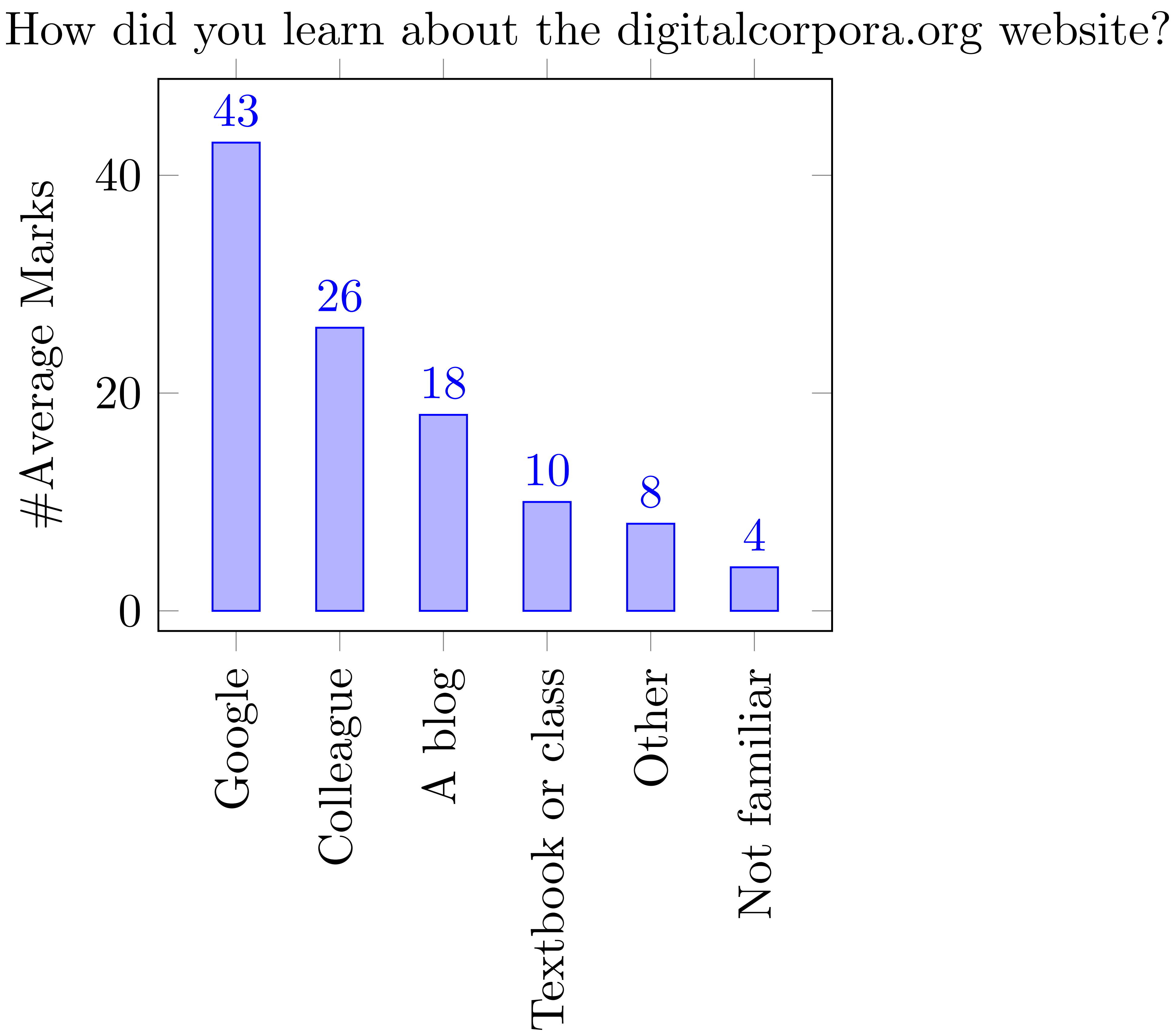
May 16, 2023

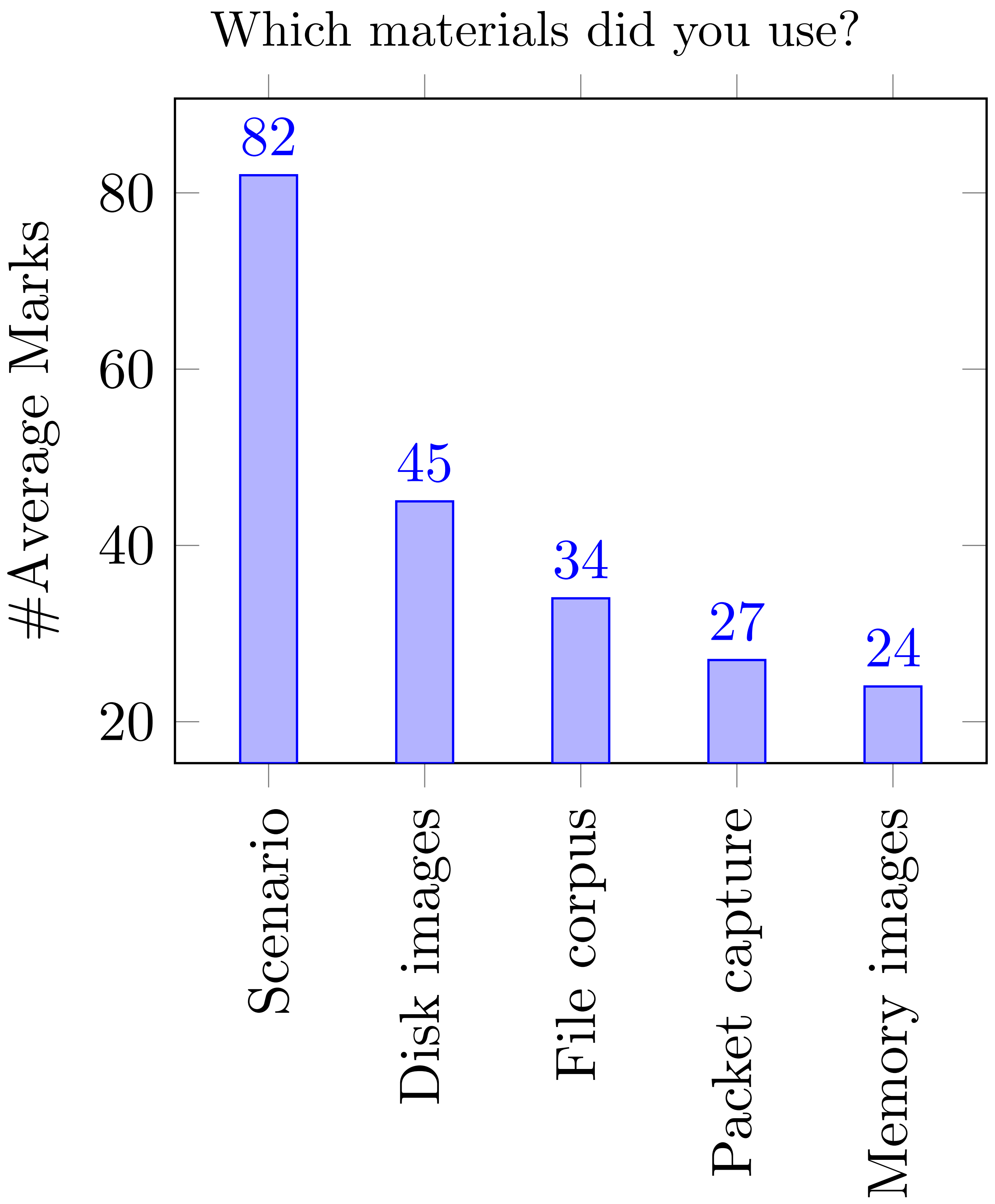
Digital Corpora: Educational Impact

Solutions to the scenarios are distributed as an encrypted PDF.

Faculty can request the decryption key; so far over 325 have.

We surveyed those requesting the key; 92 completed our survey.





Datasets were used for education, tool testing, and a little research…

	Guide	Count
For which did you download a teacher's guide?	Lone Wolf Scenario	55
	M57 patents	51
	Nitroba university	34
	DC art gallery	12
	Narcos	3
	M57 Jean	1

We used the datasets for:	Education and Training	81
	Tool testing	22
	R&D for new tools and features	13
	Practice to prepare for casework	11
	Proficiency testing	9
	Research on DF investigative practices	1
	Analysis and exploratory research	1

Research Projects

Bulk_extractor (2006-2020) — processing bulk data for digital forensics

Digital Corpora (2006-) — realistic data for digital forensics research and education

Disclosure Avoidance System for the 2020 Census (2017-2021)

PlantTracer (2023-) — Using computer vision to watch plant movement.

In 2017 I started at the US Census Bureau with the mission of modernizing the bureau’s “Disclosure Avoidance.”

“Disclosure Avoidance” — aka “Statistical Disclosure Limitation”

- aka “Privacy Preserving Data Publishing” and “Privacy Preserving Data Analysis.”

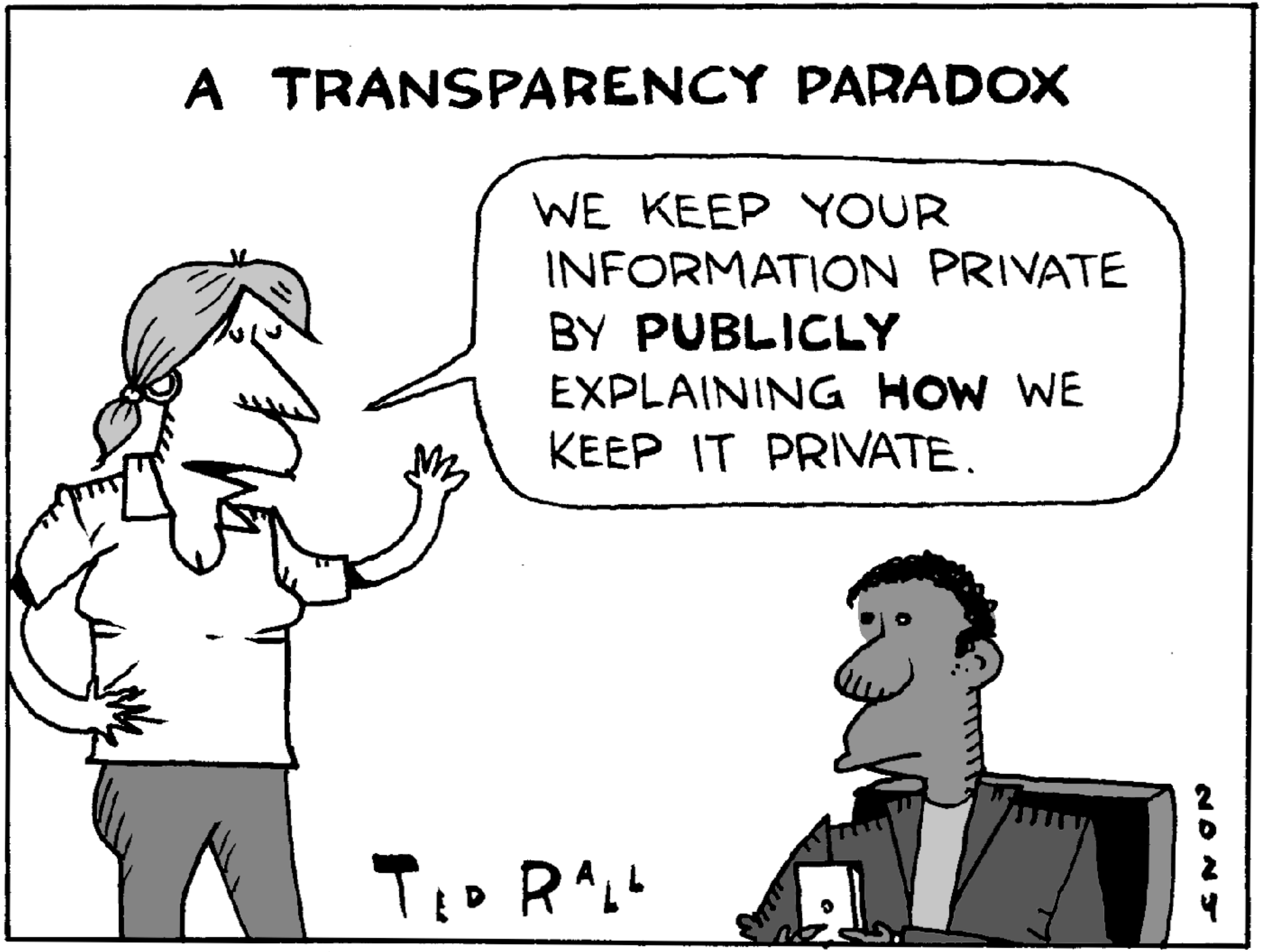
Chief of the Center for Disclosure Avoidance (2017) — GS15 management

Senior Computer Scientist for Confidentiality and Data Access (2017-2021) - ST

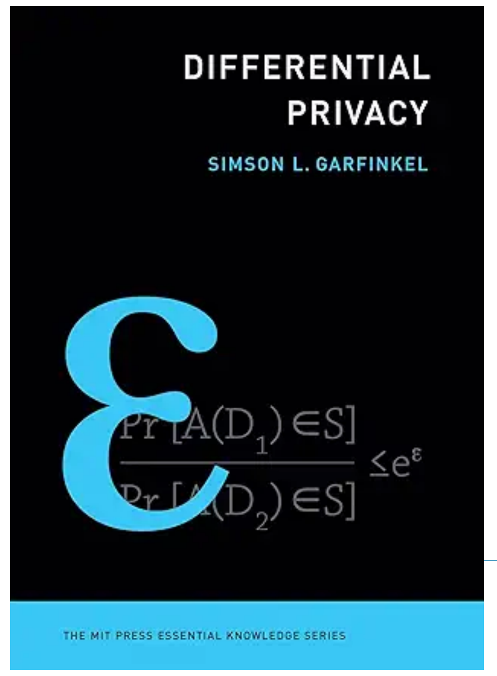
Chair, Disclosure Review Board (2017-2019)

Accomplishments:

- Brought differential privacy to the 2020 Census
- Modernized DRB
- Laid ground work for modernizing privacy protections in American Community Survey, Federal Housing Survey, Economic Census, and many other data products.
- Educated Census senior leadership on differential privacy.



A TRANSPARENCY PARADOX
WE KEEP YOUR INFORMATION PRIVATE BY **PUBLICLY** EXPLAINING **HOW** WE KEEP IT PRIVATE.
TED RALL
2024



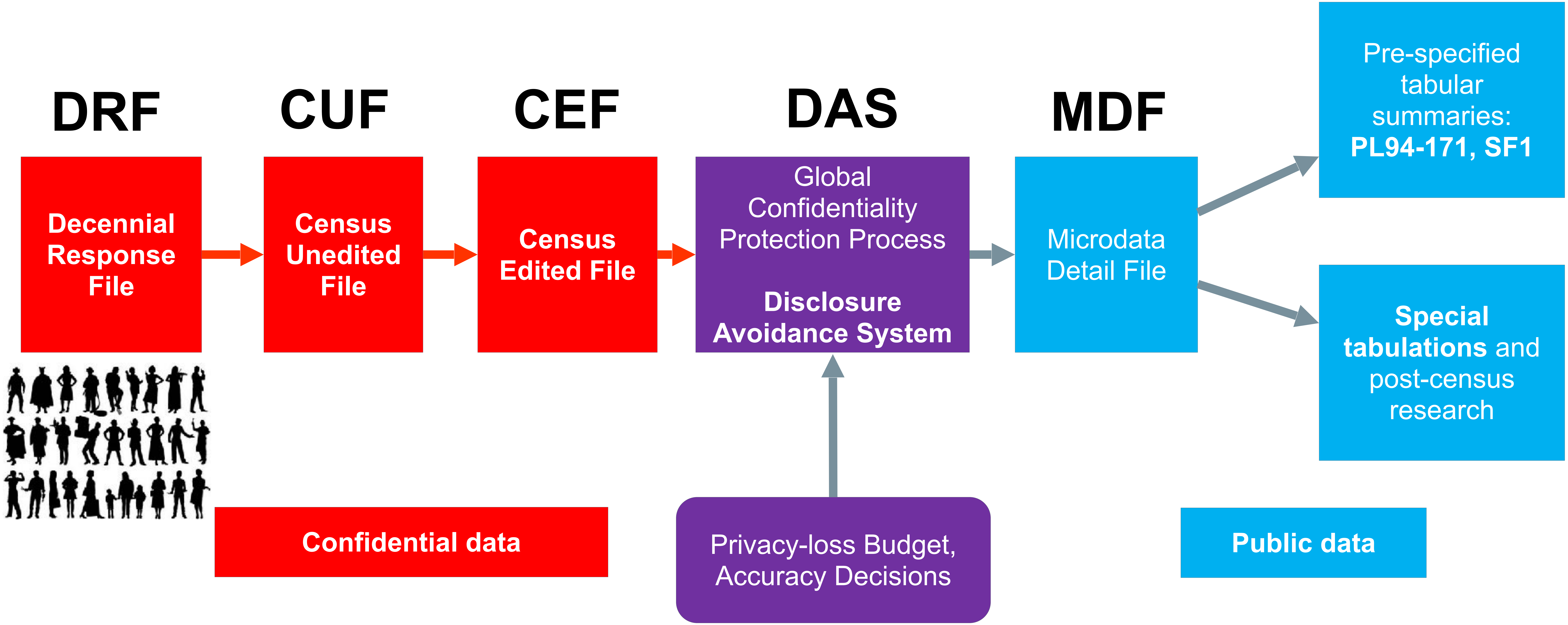
DIFFERENTIAL PRIVACY

SIMSON L. GARFINKEL

ε $\frac{\Pr[A(D_1)\in S]}{\Pr[A(D_2)\in S]} \le e^{\varepsilon}$

THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

Data flow in the 2020 Census
(Original vision)



DRF

Decennial Response File

CUF

Census Unedited File

CEF

Census Edited File

DAS

Global Confidentiality Protection Process

Disclosure Avoidance System

MDF

Microdata Detail File

Pre-specified tabular summaries: **PL94-171, SF1**

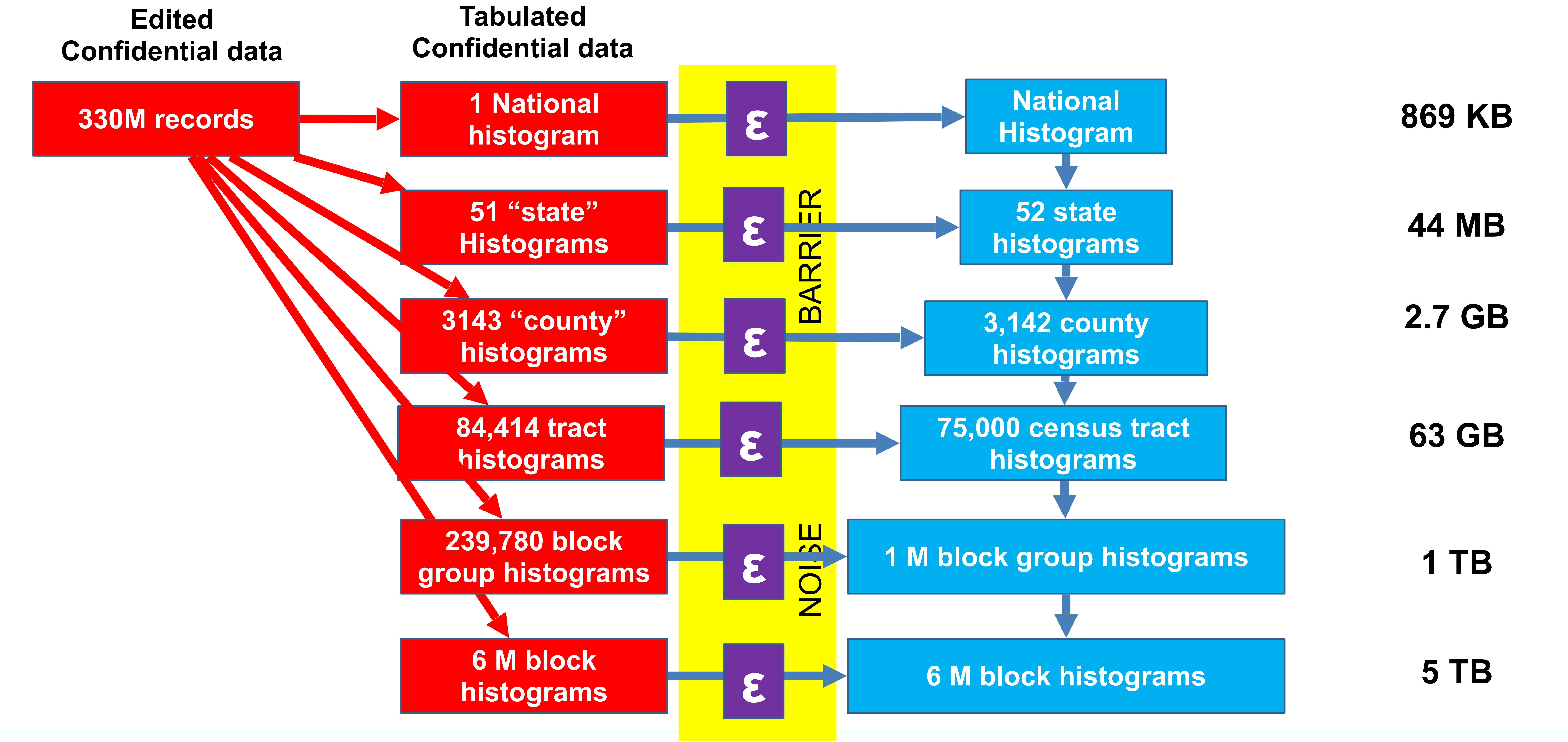
Special tabulations and post-census research

Confidential data

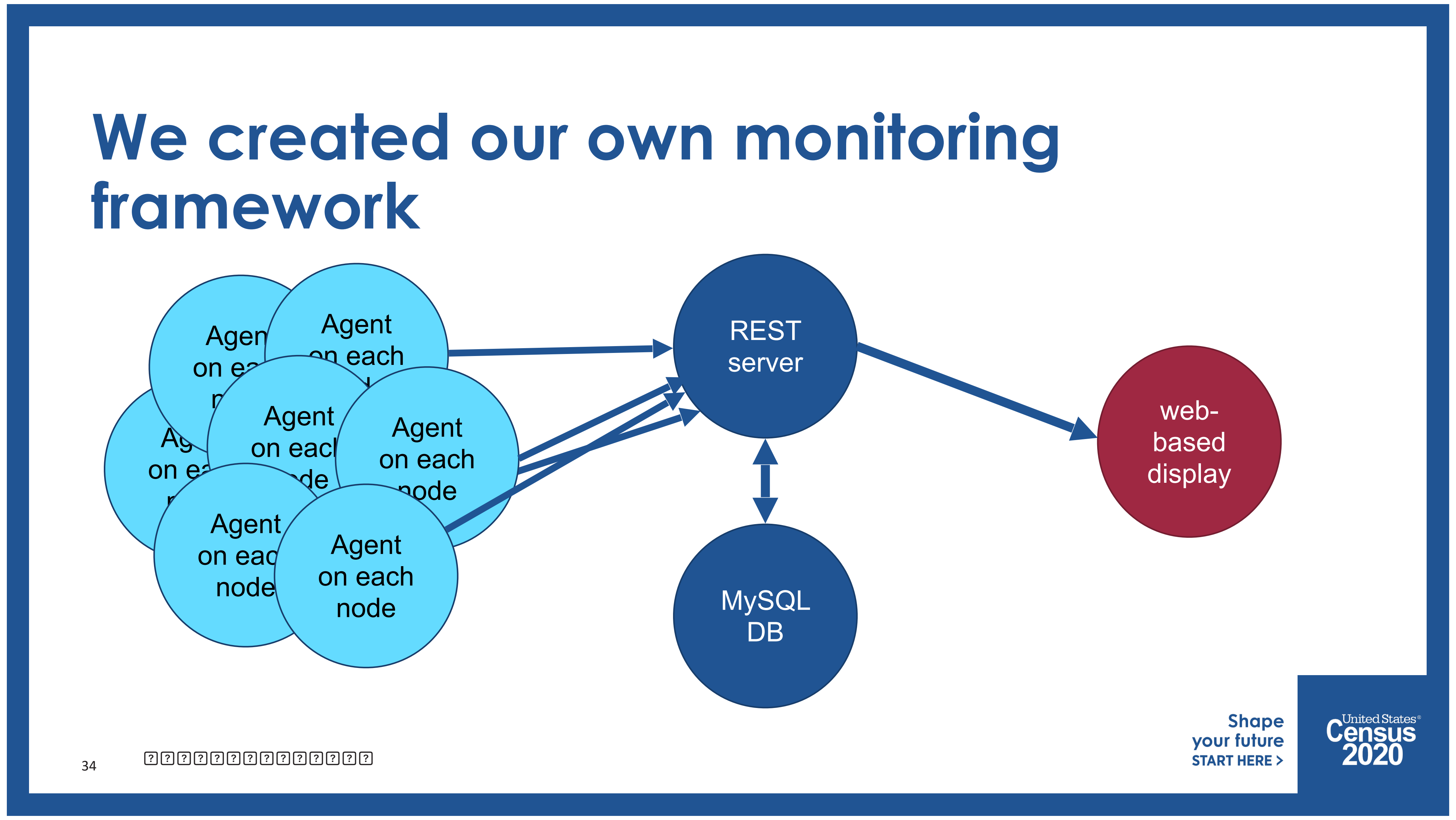
Privacy-loss Budget, Accuracy Decisions

Public data

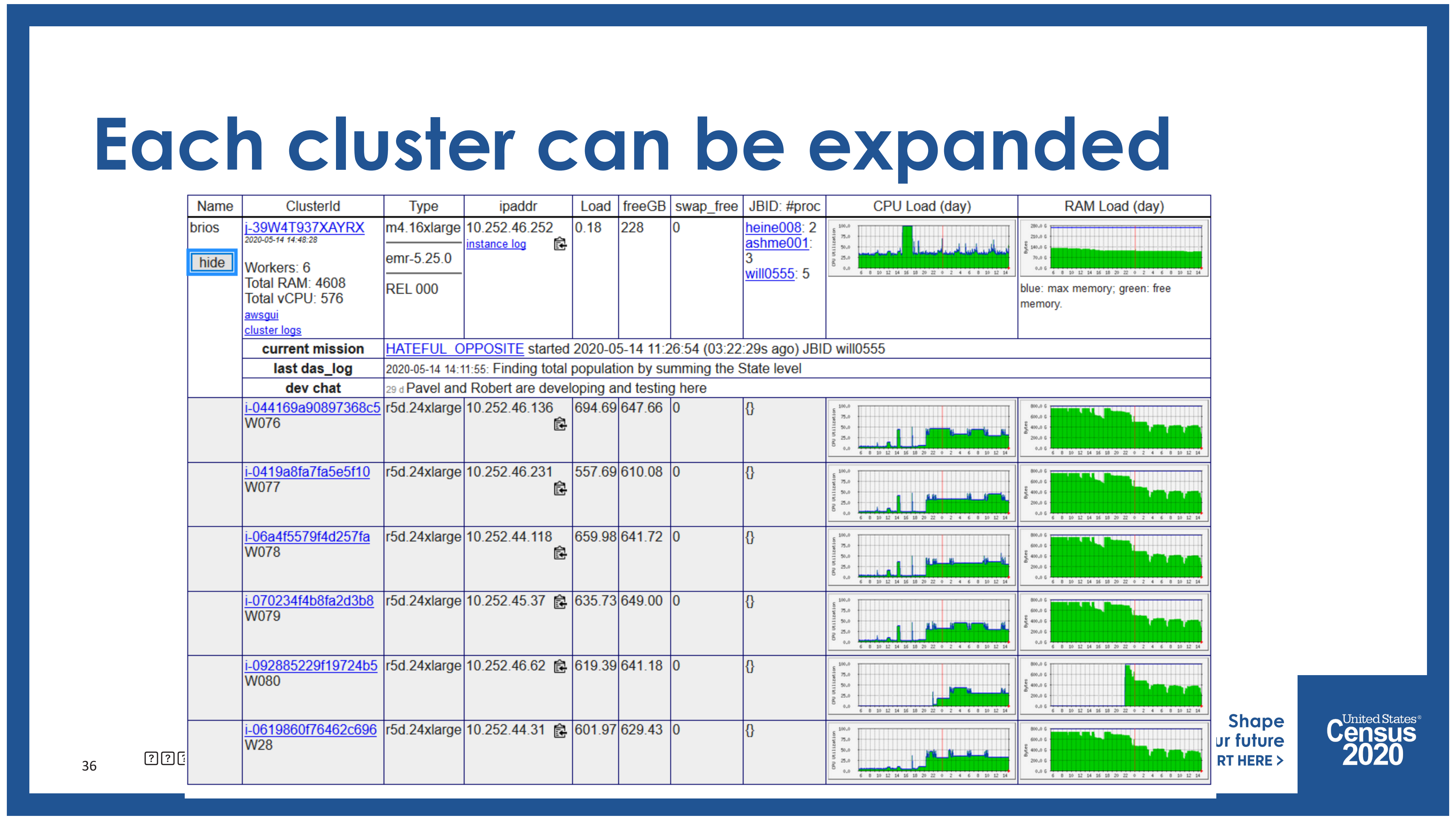
The top-down mechanism:
Each histogram provides statistical accuracy to those underneath.



Python application (70 kloc) • Amazon Elastic Map Reduce (Spark)
21-node cluster x 96 core, 488GB RAM

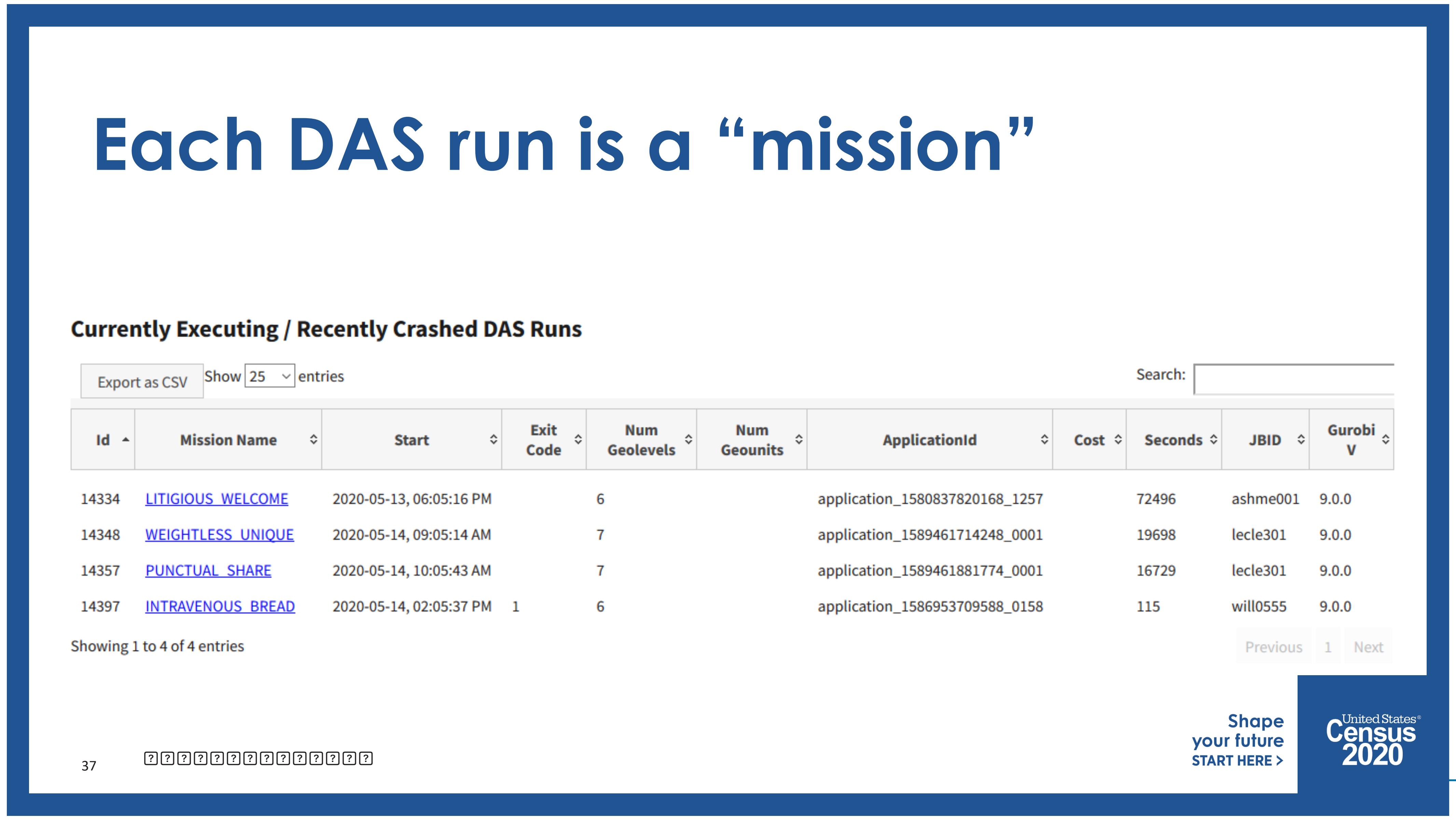


We created our own monitoring framework



Each cluster can be expanded

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)
brios hide	j-39W4T937XAYRX 2020-05-14 14:48:28 Workers: 6 Total RAM: 4608 Total vCPU: 576 awsgui cluster logs	m4.16xlarge emr-5.25.0 REL 000	10.252.46.252 instance log	0.18	228	0	heine008: 2 ashme001: 3 will0555: 5		blue: max memory; green: free memory.
	current mission	HATEFUL_OPPOSITE started 2020-05-14 11:26:54 (03:22:29s ago) JBID will0555							
	last das_log	2020-05-14 14:11:55: Finding total population by summing the State level							
	dev chat	29 d Pavel and Robert are developing and testing here							
	i-044169a90897368c5 W076	r5d.24xlarge	10.252.46.136	694.69	647.66	0	{}		
	i-0419a8fa7fa5e5f10 W077	r5d.24xlarge	10.252.46.231	557.69	610.08	0	{}		
	i-06a4f5579f4d257fa W078	r5d.24xlarge	10.252.44.118	659.98	641.72	0	{}		
	i-070234f4b8fa2d3b8 W079	r5d.24xlarge	10.252.45.37	635.73	649.00	0	{}		
	i-092885229f19724b5 W080	r5d.24xlarge	10.252.46.62	619.39	641.18	0	{}		
	i-0619860f76462c696 W28	r5d.24xlarge	10.252.44.31	601.97	629.43	0	{}		



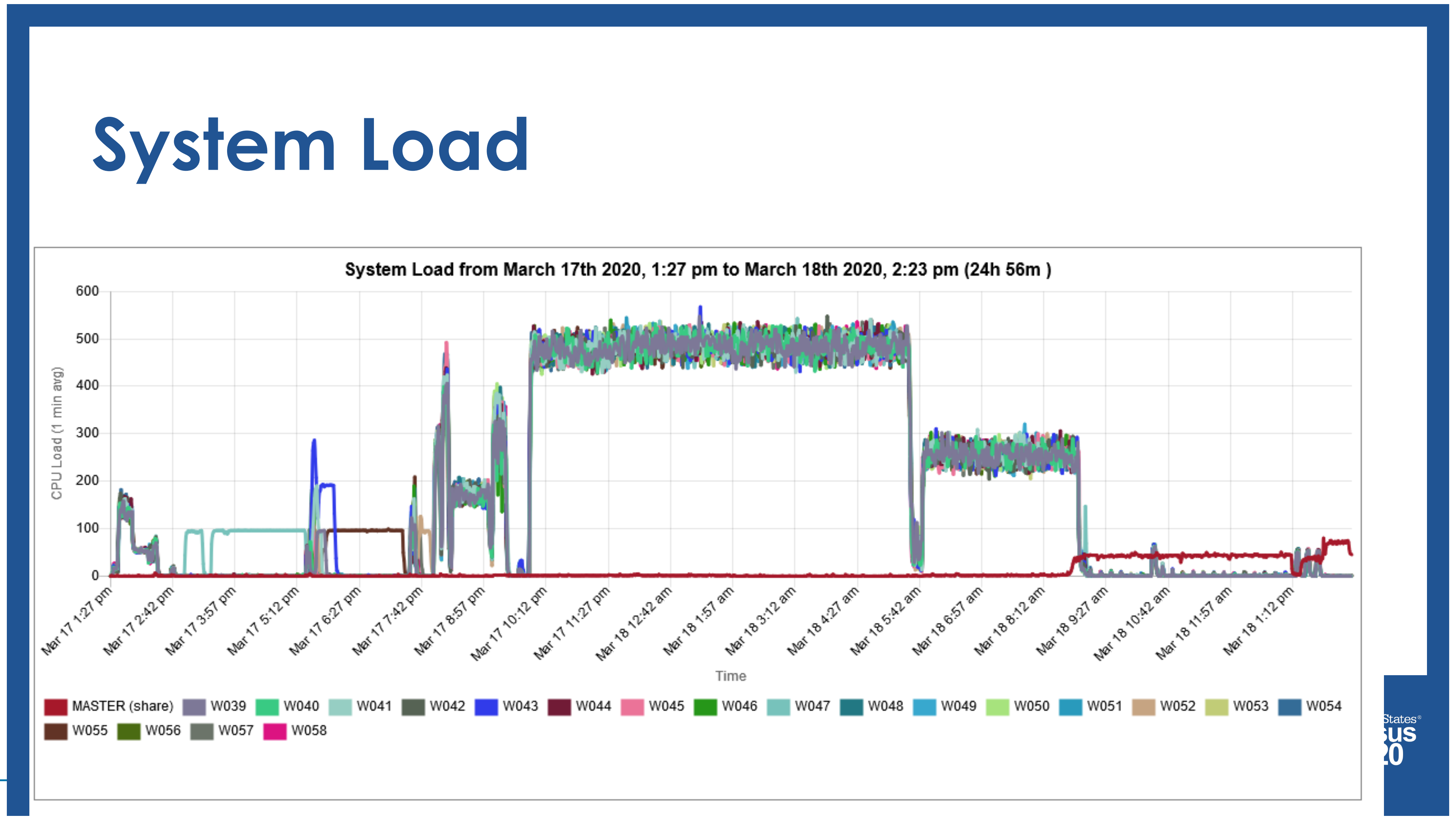
Each DAS run is a "mission"

Currently Executing / Recently Crashed DAS Runs

Export as CSV Show 25 entries Search:

Id	Mission Name	Start	Exit Code	Num Geolevels	Num Geounits	ApplicationId	Cost	Seconds	JBID	Gurobi V
14334	LITIGIOUS_WELCOME	2020-05-13, 06:05:16 PM		6		application_1580837820168_1257		72496	ashme001	9.0.0
14348	WEIGHTLESS_UNIQUE	2020-05-14, 09:05:14 AM		7		application_1589461714248_0001		19698	lecle301	9.0.0
14357	PUNCTUAL_SHARE	2020-05-14, 10:05:43 AM		7		application_1589461881774_0001		16729	lecle301	9.0.0
14397	INTRAVENOUS_BREAD	2020-05-14, 02:05:37 PM	1	6		application_1586953709588_0158		115	will0555	9.0.0

Showing 1 to 4 of 4 entries



System Load

Research Projects

Bulk_extractor (2006-2020) — processing bulk data for digital forensics

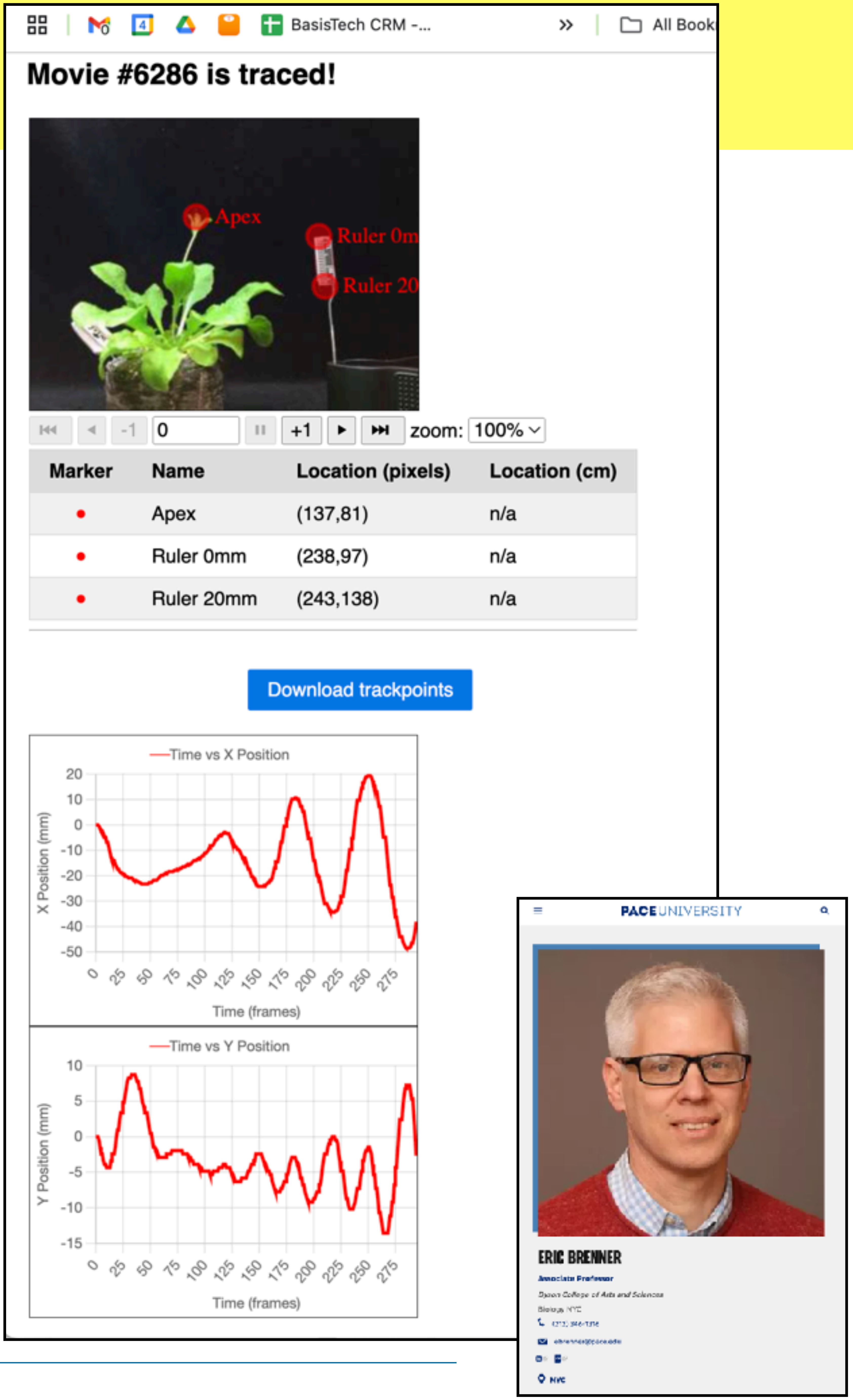
Digital Corpora (2006-) — realistic data for digital forensics research and education

Disclosure Avoidance System for the 2020 Census (2017-2021)

PlantTracer (2023-) — Using computer vision to watch plant movement.

Plant Tracer (w/ Eric Brenner, Pace Univ.)

- Converted iOS app to web-based app
- Designed learning management system for sharing and analyzing video
- Developing open source web-based video capture system & ESP32-based camera.



Movie #6286 is traced!

Marker	Name	Location (pixels)	Location (cm)
•	Apex	(137,81)	n/a
•	Ruler 0mm	(238,97)	n/a
•	Ruler 20mm	(243,138)	n/a

Download trackpoints

ERIC BRENNER

Associate Professor

Dyson College of Arts and Sciences

Biology, NYC

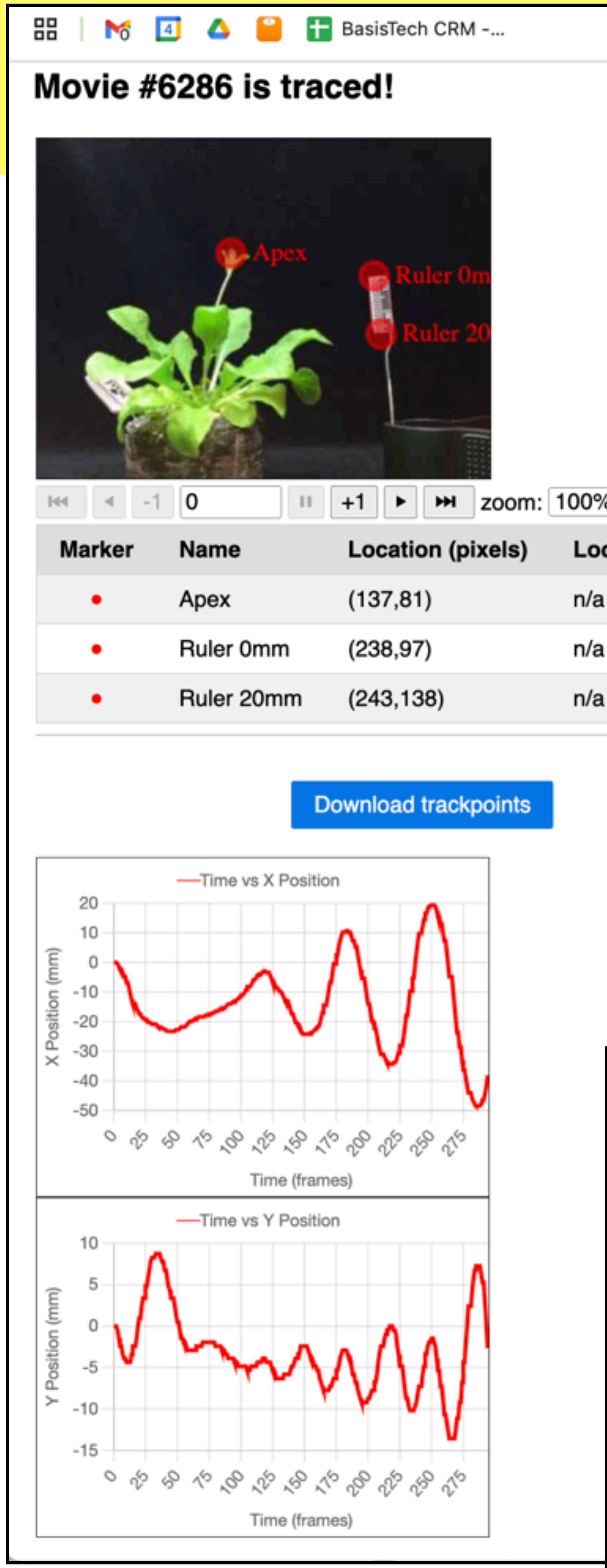
(212) 346-1316

ebrenner@pace.edu

NYC

Plant Tracer (w/ Eric Brenner, Pace Univ.)

- Converted iOS app to web-based app
- Designed learning management system for sharing and analyzing video
- Developing open source web-based video capture system & ESP32-based camera.

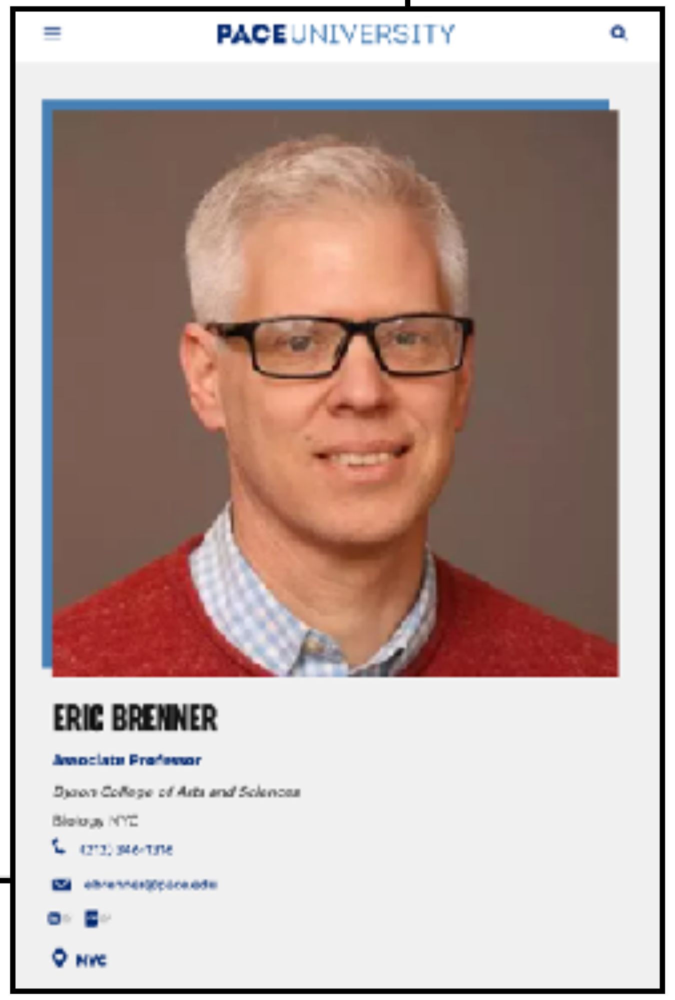


Movie #6286 is traced!

0 zoom: 100%

Marker	Name	Location (pixels)	Location (cm)
●	Apex	(137,81)	n/a
●	Ruler 0mm	(238,97)	n/a
●	Ruler 20mm	(243,138)	n/a

Download trackpoints



ERIC BRENNER

Associate Professor

Dyson College of Arts and Sciences

Biology, NYC

☎ [illegible]

ebrenner@pace.edu

NYC

Two fun classes

Defense Against the Dark Arts

Course goal

Provide students with a basic understanding of how they can get hacked and how to defend themselves

“Defense Against the Dark Arts”
Harvard College, Fall 2024

First-year seminar course.

- 12 first semester students, ~ 6 with background in computing
- 2.5 hours each week
- 90 minutes discussion of reading
- 90 minute “lab”

Books:

- *Fancy Bear Goes Phishing: The Dark History of the Information Age in Five Extraordinary Hacks* by Scott J. Shapiro (2023) (entire book)
- *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race* by Nicole Perlroth (2021) (selected chapters)
- *Crypto: How the Code Rebels Beat the Government Saving Privacy* in the Digital Age by Steven Levy (2002) (selected chapters)

Technical Papers

- Light Commands: Laser-Based Audio Injection on Voice-Controllable Systems
- An Empirical Analysis of the Commercial VPN Ecosystem
- Security at the End of the Tunnel: The Anatomy of VPN Mental Models Among Experts and Non-Experts in a Corporate Context
- Extending a Hand to Attackers: Browser Privilege Escalation Attacks via Extensions

Course Outline

Week 1 — A Hacker’s Bestiary 1: The Great Worm and Computer Viruses

Week 2 — A Hacker’s Bestiary 2: The Wizards

Week 3 — Magic Words (passwords)

Week 4 — The Oldest Dark Art (Secret Key Cryptography)

Week 5 — New Directions in Cryptography (Public Key Cryptography)

Week 6 — Keeping Your Data Out of Azkaban (Ransomware and Malware)

Week 7 — Finding Spells in the Library: Read the Computer Security Literature

Week 8 — Your Cloak of Invisibility (VPNs)

Week 9 — Sneaking through the forest (Browser Privacy, and Private Browsing Mode)

Week 11 — Securing your broom (Browser Extensions)

Week 12 — Transfiguration (phishing)

Week 13 — There’s magic in SoK (Tech abuse SoK)

Labs

week 1 - lab - command line

week 2 - lab - hashing

week 3 - lab - password cracking

week 4 - lab - encryption

week 5 - lab - digital signatures

week 6 - lab - Finding Hidden Data

week 7 - lab - voice commands

week 8 - lab - HTTP Toolkit

week 9 - traveling around the world with ping and traceroute

week 10 - Browser Extensions

Each lab designed to be accomplished in 45 minutes by first-year students with varying levels of background.

Everybody wins! (Everybody gets something out of the lab.)

Other significant courses I’ve taught

This semester:

- Critical Thinking in Data Science (Data Science Ethics) — Harvard Data Science Master’s Program
- Artificial Intelligence, Internet of Things, and Cybersecurity — Harvard Division of Continuing Education

In DC:

- Mac Forensics, Cloud Forensics, Document Forensics — George Mason University Digital Forensics Master’s Program
- Data Science Ethics — George Washington University Data Science Program
- Massive Data Analytics, Data Privacy — Georgetown University

Naval Postgraduate School:

- Information Crime, Law and Ethics
- Data Fusion with Online Information Systems
- Network Security
- Java as a Second Language
- Special topics in computer security
- Advanced Computer Architecture
- Automated Document and Media Exploitation

What’s next

Tech Abuse Center — A multidisciplinary research center and clinic to address the issue of technology-enabled domestic abuse.

Embedded Forensics — Deep-dive into the embedded microelectronics ecosystem.

Q&A