Driving Innovation from the University to Government and the Enterprise: Tricks, Traps and Techniques

Monday, June 3, 2024

Simson Garfinkel

These slides can be downloaded from https://simson.net/ref/2024

Abstract

Innovation and entrepreneurship in science and technology requires seizing new ideas, identifying their distinctive core, and finding enthusiastic users and customers (**tricks**).

It also requires avoiding mistakes that can sideline amazing technology and promising careers (**traps**).

In this talk, I'll present some tricks I've learned, traps that I've been fortunate enough to avoid or escape, and which approaches I think can be taught and generalized (**techniques**) based on my career in academia, government, and several startups.

Tricks

Embrace the leading edge.

Embrace open source & open data.

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 Innovation

- US Census Bureau — 2017-2021 — Differential privacy
- US DHS — 2021-2022 — DHS 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.

Trick #1 — Embrace the leading edge

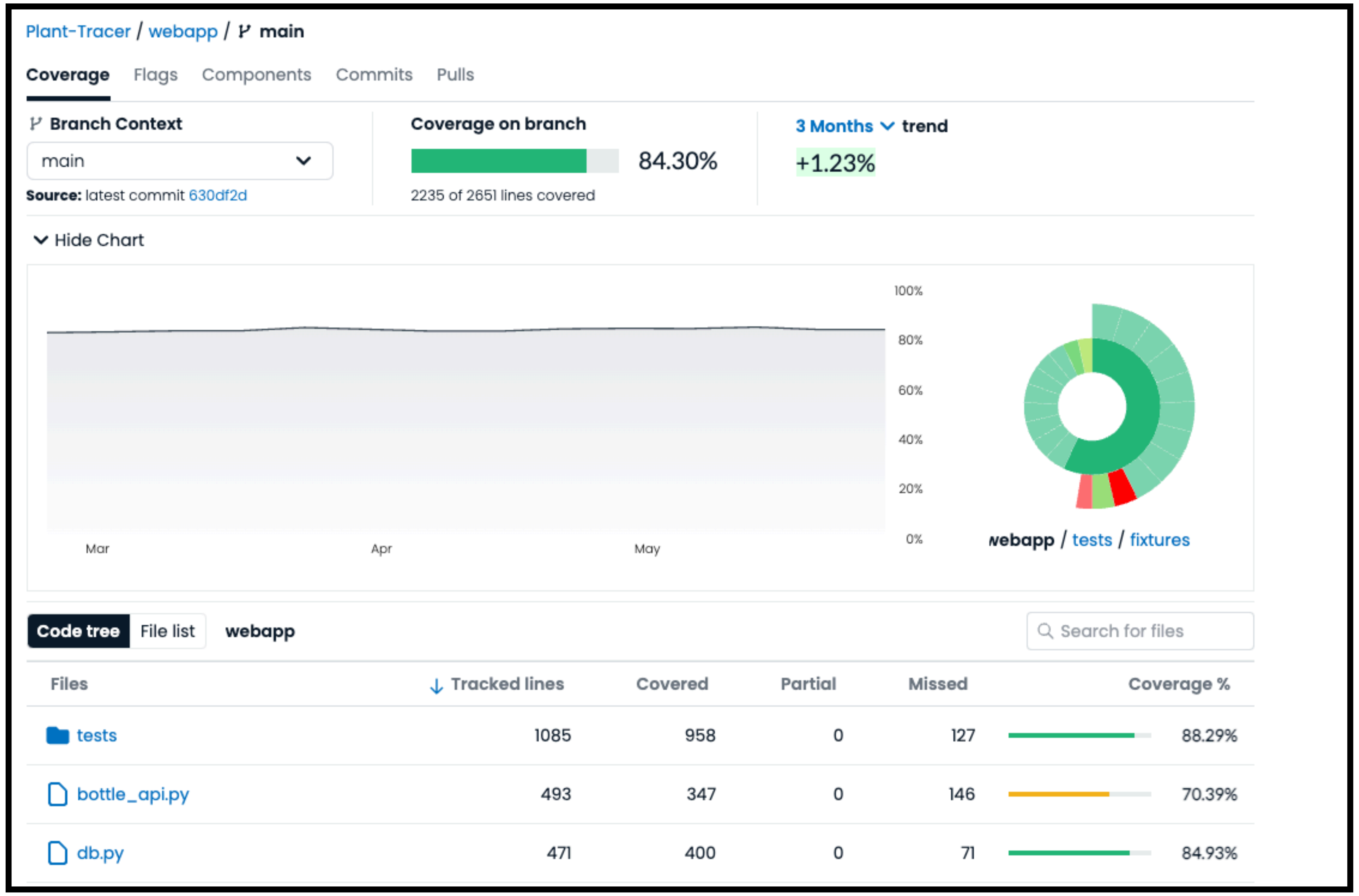
Teach and embrace the modern programming languages & development tools.

Example: AWS in Georgetown "massive data" course (2016-201

- Student got a job with AWS because he learned the stack in my class.

Example: Plant Tracer Project (w/ Pace Univ.)

- Computer vision for tracking plant movement
- Unit tests & code coverage
- Function-as-a-service (AWS Lambda)



Plant-Tracer / webapp / main

Coverage Flags Components Commits Pulls

Branch Context: main

Source: latest commit 630df2d

Coverage on branch 84.30%

2235 of 2651 lines covered

3 Months trend +1.23%

Hide Chart

Code tree File list webapp

Search for files

Files	Tracked lines	Covered	Partial	Missed	Coverage %
tests	1085	958	0	127	88.29%
bottle_api.py	493	347	0	146	70.39%
db.py	471	400	0	71	84.93%

Trick #1 — Embrace the leading edge

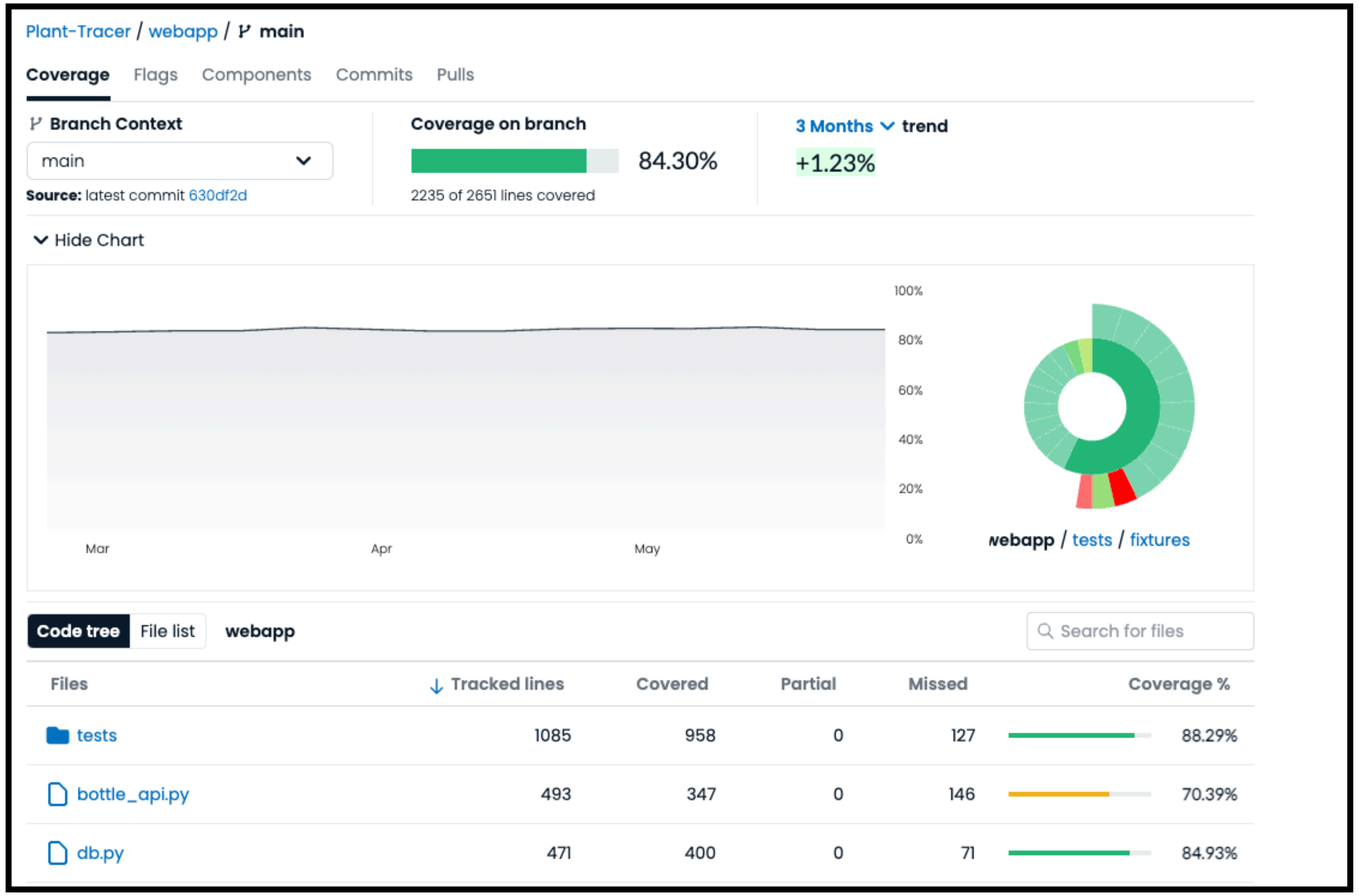
Teach and embrace the modern programming languages & development tools.

Example: AWS in Georgetown “massive data” course (2016-201

- Student got a job with AWS because he learned the stack in my class.

Example: Plant Tracer Project (w/ Pace Univ.)

- Computer vision for tracking plant movement
- Unit tests & code coverage
- Function-as-a-service (AWS Lambda)



Plant-Tracer / webapp / main

Coverage Flags Components Commits Pulls

Branch Context: main
Source: latest commit 630df2d

Coverage on branch: 84.30%
2235 of 2651 lines covered

3 Months trend: +1.23%

Hide Chart

Code tree File list webapp

Search for files

Files	↓ Tracked lines	Covered	Partial	Missed	Coverage %
tests	1085	958	0	127	88.29%
bottle_api.py	493	347	0	146	70.39%
db.py	471	400	0	71	84.93%

Trick #2 — Embrace open source and open data

Bulk_extractor - Open source digital forensics tool

Initial development: 2006-2010

First deployment: 2010

Version 2.0 rewrite: 2018-2022 (C++20)

“CV impact:” 2 articles

COMPUTERS & SECURITY 32 (2013) 56–72

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/cose

ELSEVIER

Computers & Security

Digital media triage with bulk data analysis and *bulk_extractor*

Simson L. Garfinkel[*]

Naval Postgraduate School, 900N Glebe, Arlington, VA 22207, USA

ARTICLE INFO

Article history:
Received 8 December 2011
Received in revised form
4 July 2012
Accepted 14 September 2012

Keywords:
Digital forensics
Bulk data analysis
bulk_extractor
Optimistic decompression
Windows hibernation files
EnCase
Forensic path
Margin
Parallelized forensic analysis
Stream-based forensics

ABSTRACT

Bulk data analysis eschews file extraction and analysis, common in forensic practice today, and instead processes data in “bulk,” recognizing and extracting salient details (“features”) of use in the typical digital forensics investigation. This article presents the requirements, design and implementation of the *bulk_extractor*, a high-performance carving and feature extraction tool that uses bulk data analysis to allow the triage and rapid exploitation of digital media. Bulk data analysis and the *bulk_extractor* are designed to complement traditional forensic approaches, not replace them. The approach and implementation offer several important advances over today’s forensic tools, including optimistic decompression of compressed data, context-based stop-lists, and the use of a “forensic path” to document both the physical location and forensic transformations necessary to reconstruct extracted evidence. The *bulk_extractor* is a stream-based forensic tool, meaning that it scans the entire media from beginning to end without seeking the disk head, and is fully parallelized, allowing it to work at the maximum I/O capabilities of the underlying hardware (provided that the system has sufficient CPU resources). Although *bulk_extractor* was developed as a research prototype, it has proved useful in actual police investigations, two of which this article recounts.

Published by Elsevier Ltd.

1. Introduction

Digital forensics investigations have grown more difficult as the capacity and diversity of devices containing digital evidence increases.

Many approaches have been proposed for addressing the data onslaught, including parallelization and multiprocessing (Ayers, 2009; Richard and Roussev, 2006), statistical sampling (Garfinkel et al., 2010; Mora, 2010), and even extending the time a suspect can be held without charge so that evidence may be analyzed (Secretary of State for the Home Department, 2007). But no matter what approach is used to improve performance, so long as a backlog exists, some process must allocate limited forensic resources.

Diversity of media represents a different kind of challenge. Increasingly evidence needs to be analyzed that is in file system and file formats not supported by current tools. Frequently the only way to analyze such data is to scavenge it for printable strings and rely on the human examiner to make sense of the data that are manually recovered.

Many organizations prioritize forensic processing with reference to the nature of the case, without taking into account the contents of the media itself. For example, the FBI’s regional computer forensics laboratories give media

[*] Tel.: +1 831 656 3389.
E-mail address: slgarfin@nps.edu.

0167-4048/$ – see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.cose.2012.09.011

156 citations

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

0 citations (14K DL downloads)

Real world impact: education, law enforcement & defense

Digital Corpora - Open data set for digital forensics

Initial development: 2006-2010

First deployment: 2007-

Migration to Amazon Open Data program: 2020

“CV impact:” 1 article

DIGITAL INVESTIGATION 6 (2009) S2–S11

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

ELSEVIER

Digital Investigation

Bringing science to digital forensics with standardized forensic corpora

Simson Garfinkel[a,b,*], *Paul Farrell*[a], *Vassil Roussev*[c], *George Dinolt*[a]

[a] *Graduate School of Operational and Information Sciences, Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943, USA*
[b] *Harvard University, USA*
[c] *University of New Orleans, USA*

ABSTRACT

Keywords:
Forensics
Human subjects research
Corpora
Real data corpus
Realistic data

Progress in computer forensics research has been limited by the lack of a standardized data sets—*corpora*—that are available for research purposes. We explain why corpora are needed to further forensic research, present a taxonomy for describing corpora, and announce the availability of several forensic data sets.

© 2009 Digital Forensic Research Workshop. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Much of the work to date in digital forensics has focused on data extraction and for presentation in courts. Researchers have developed technologies for copying data from subject hard drives, storing that data in a disk image file, searching the disk image for document files, and presenting the documents to an examiner.

As both the variety and scale of forensic investigations increase, forensic practitioners need tools that do more than search and present: they need tools for reconstruction, analysis, clustering, data mining, and sense-making. Such tools frequently require the development of new scientific techniques in areas such as text mining, machine learning, visualization, and related fields.

One of the hallmarks of science is the ability for researchers to perform controlled and repeatable experiments that produce reproducible results. Science is based on the principle that phenomena can be observed and results can be reproduced by *anyone*—there are no privileged experimenters or observers (given sufficient training and financial resources, of course).

Sadly, much of today’s digital forensic research results are not reproducible. For example, techniques developed and tested by one set of researchers cannot be validated by others since the different research groups use different data sets to test and evaluate their techniques.

1.1. *Why forensic corpora are needed*

Having a reference set of representative corpora enhances the scientific evaluation of forensic methods beyond the obvious benefits of providing ready test data and enabling direct comparison of different approaches. Namely, it allows for the ground truth to be established using manual or otherwise time-consuming methods. Such results can then be used as a baseline to evaluate the success of new tools and methods using objective metrics.

In the digital forensics field there have been sporadic efforts to produce standardized corpora, mostly in the form of forensic challenges. The main goal of these challenges has been the development of practical tools for problem areas in need. Since 2005, DFRWS has issued an annual challenge

[*] *Corresponding author*. Graduate School of Operational and Information Sciences, Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943, USA.
E-mail address: slgarfin@nps.edu (S. Garfinkel).

1742-2876/$ – see front matter © 2009 Digital Forensic Research Workshop. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2009.06.016

482 citations

Real world impact: digital forensics education & research

#1
Innovating with open source

Stream-Based Disk Forensics:
Scan the disk from beginning to end; do your best.

NPS Presentation from 2011-06-14

0

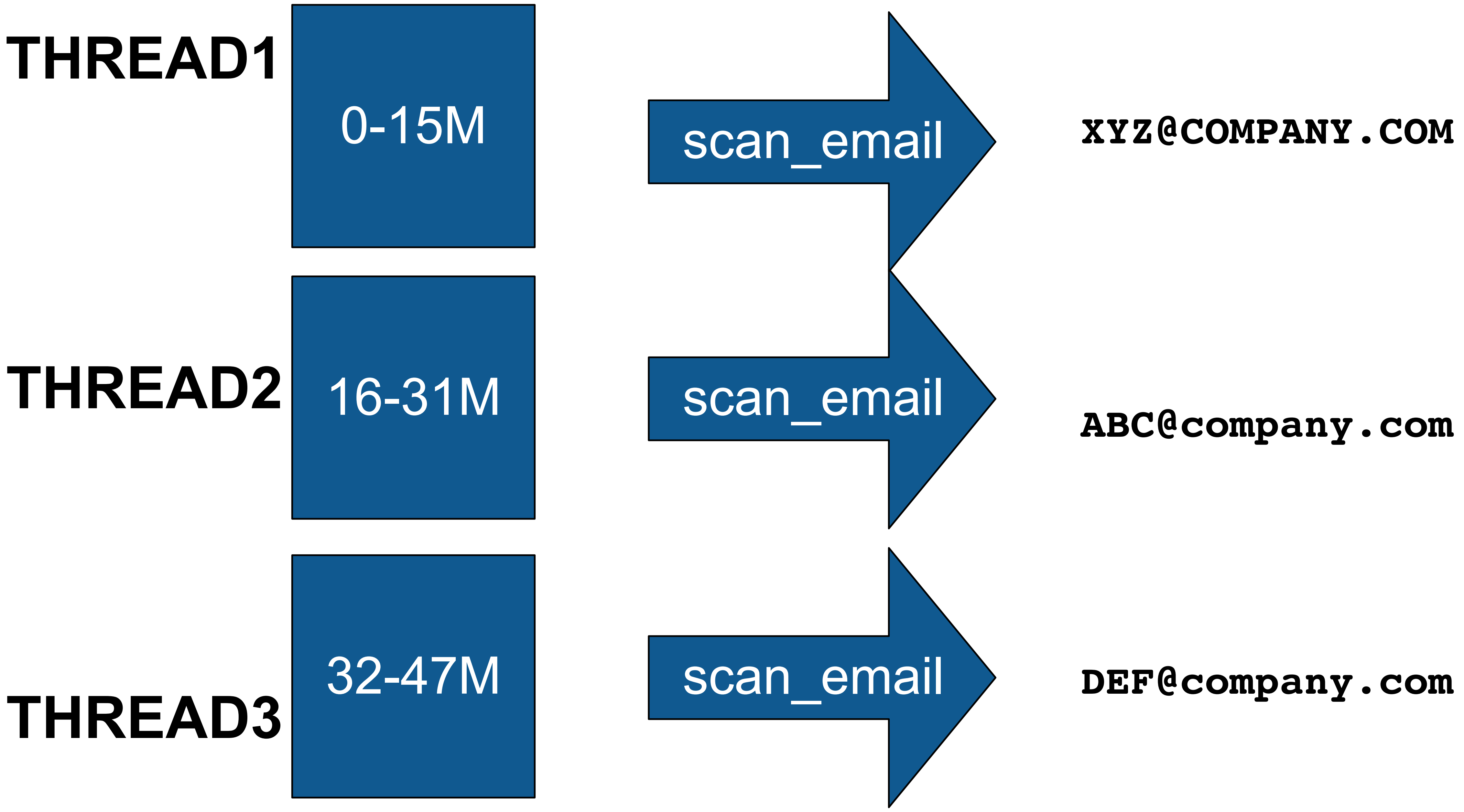
1TB

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.

NPS Presentation from 2011-06-14



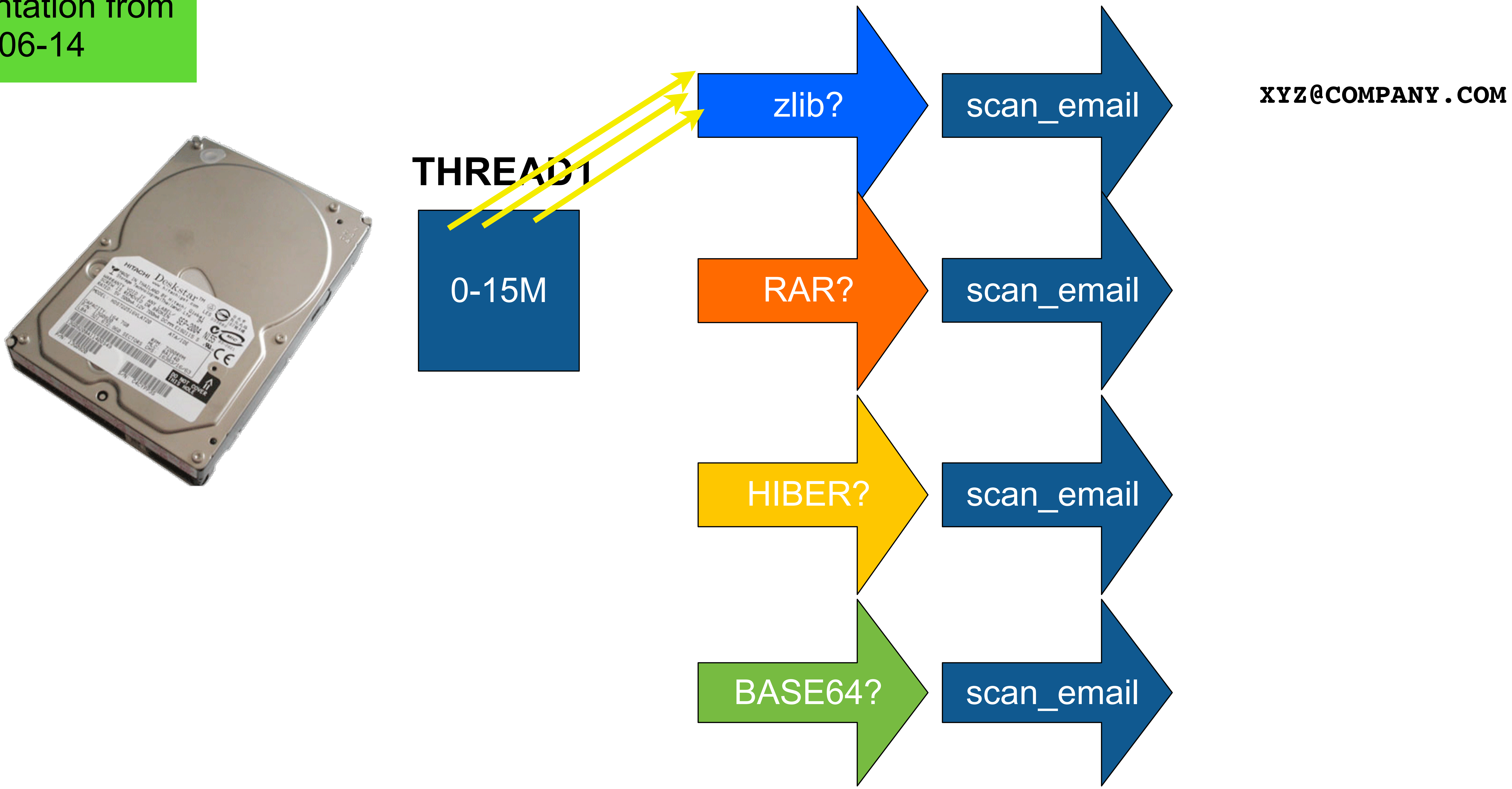
This finds obvious email addresses in bulk data:

```
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.

NPS Presentation from 2011-06-14

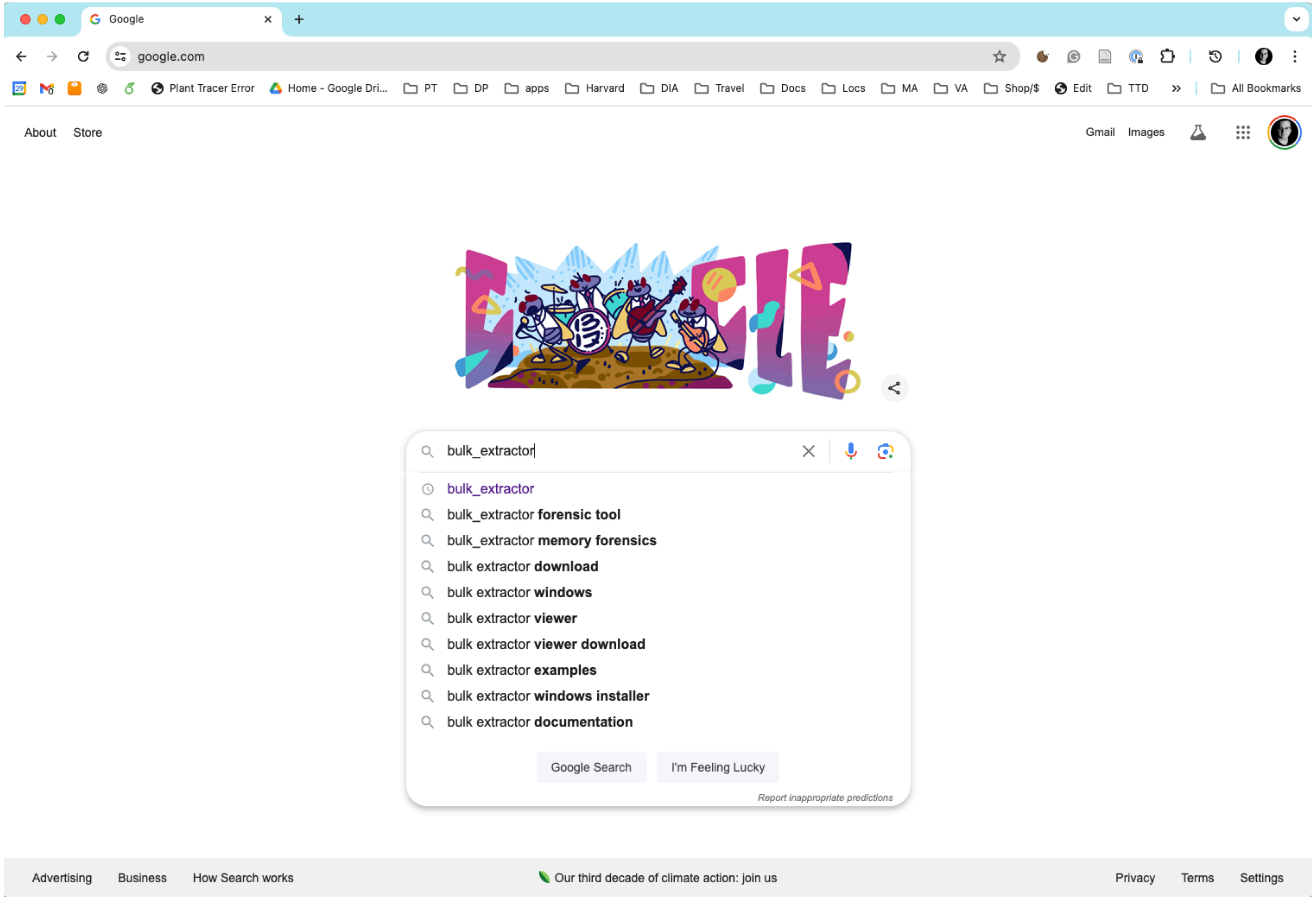


This “optimistic” approach also recovers data from fragments of files.

In 2011,
I didn’t stress that bulk_extractor was open source.

Open source was critical to bulk_extractor’s success.

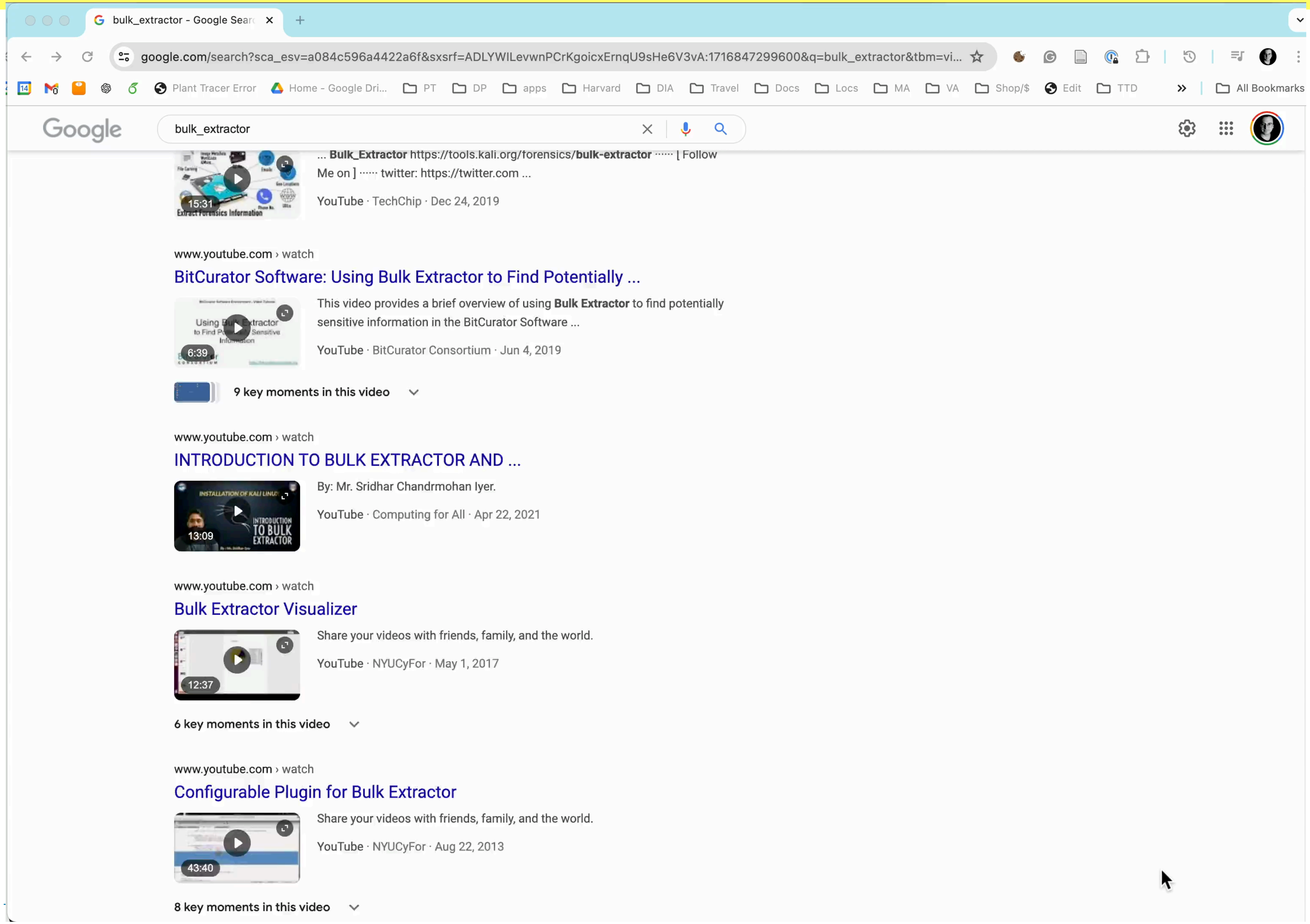
Because bulk_extractor was open source, it was widely adopted



bulk_extractor

- bulk_extractor
- bulk_extractor **forensic tool**
- bulk_extractor **memory forensics**
- bulk extractor **download**
- bulk extractor **windows**
- bulk extractor **viewer**
- bulk extractor **viewer download**
- bulk extractor **examples**
- bulk extractor **windows installer**
- bulk extractor **documentation**

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
15:31
6:39
13:09
12:37
43:40

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

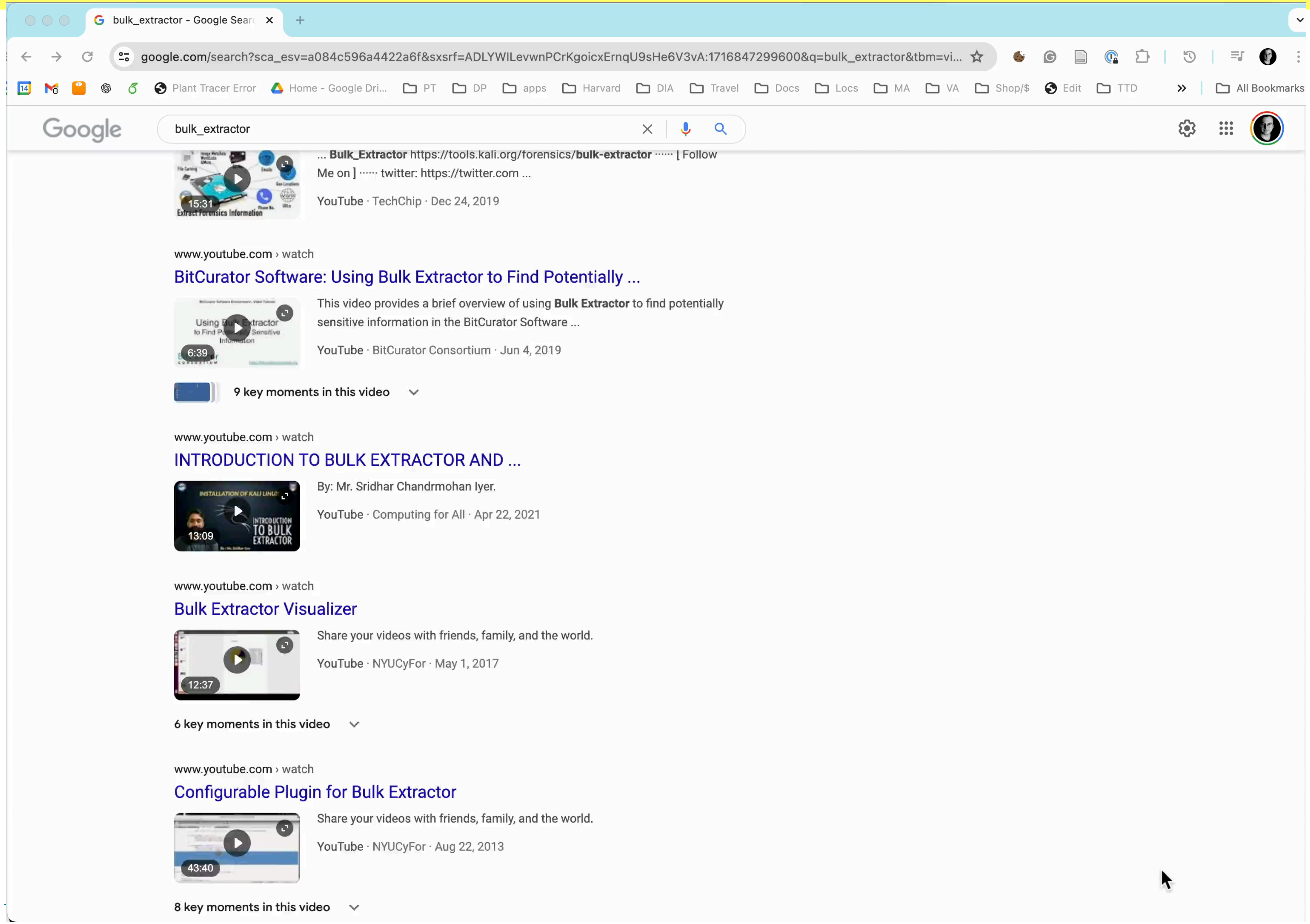
Configurable Plugin for Bulk Extractor

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
6:39
13:09
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

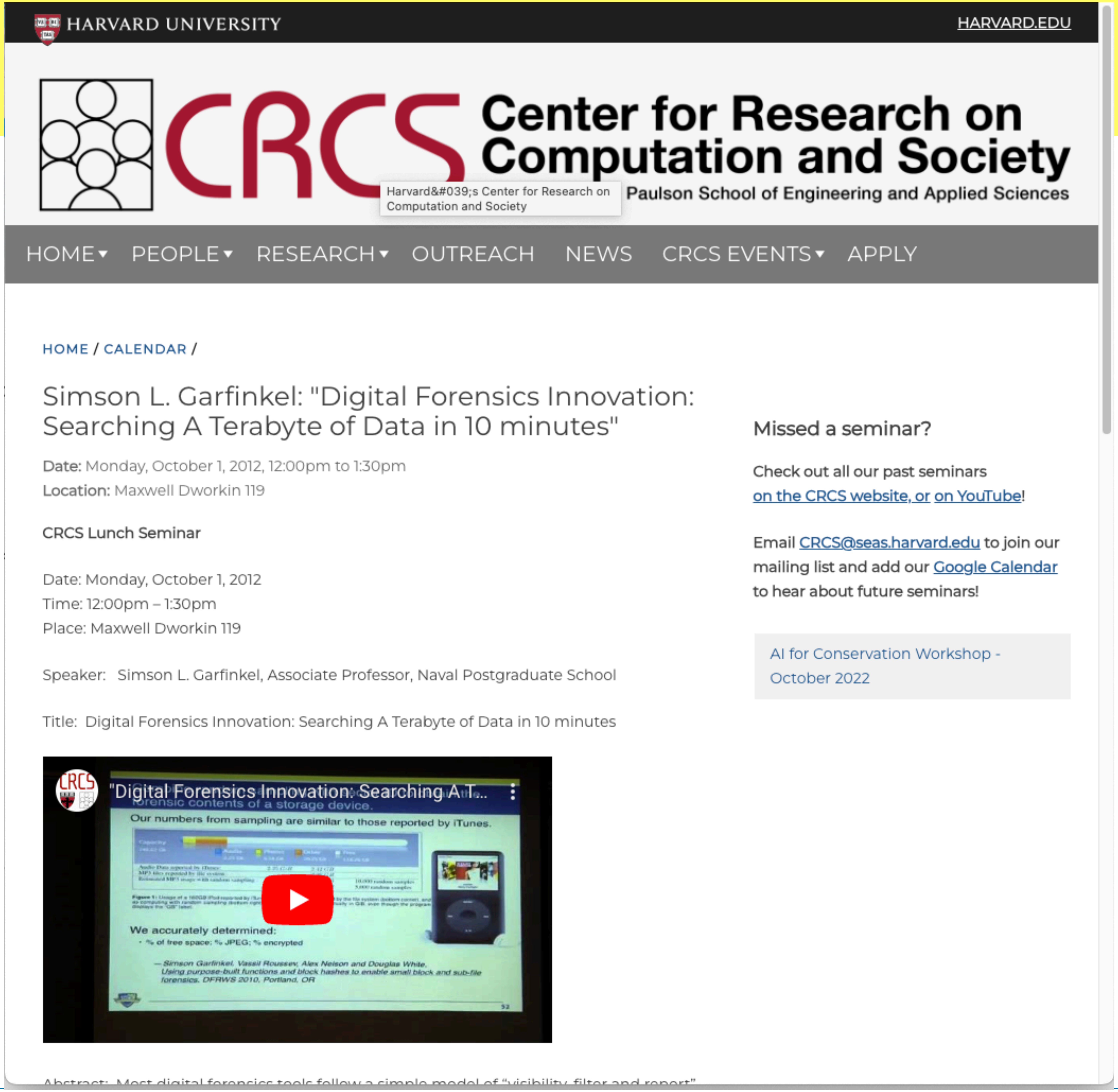
www.youtube.com › watch

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



HARVARD UNIVERSITY
HARVARD.EDU
CRCS
Center for Research on Computation and Society
Paulson School of Engineering and Applied Sciences
Harvard&#039;s Center for Research on Computation and Society
HOME ▾
PEOPLE ▾
RESEARCH ▾
OUTREACH
NEWS
CRCS EVENTS ▾
APPLY

HOME / CALENDAR /

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

"Digital Forensics Innovation: Searching A T...
Our numbers from sampling are similar to those reported by iTunes.
We accurately determined:
% of free space; % JPEG; % encrypted
— Simson Garfinkel, Vassil Roussev, Alex Nelson and Douglas White. Using purpose-built functions and block hashes to enable small block and sub-file forensics. DFRWS 2010, Portland, OR

Abstract: Most digital forensics tools follow a simple model of "visibility, filter and report"

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

I used bulk_extractor as a platform for innovation and entrepreneurship.

2012 — Statistical sampling breakthrough

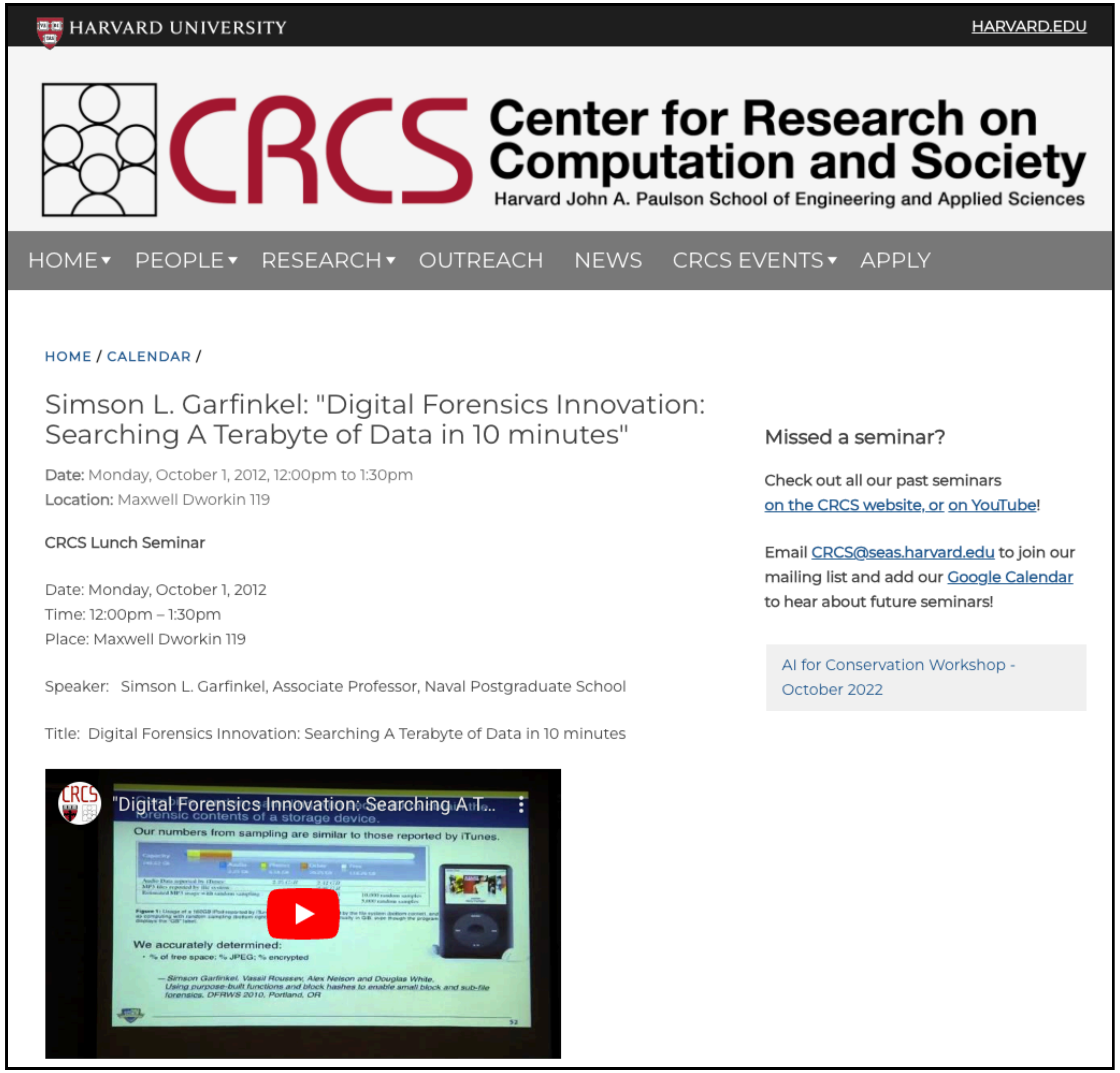
- US Patent 8,433,959 granted April 30, 2013

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

HOME▾ PEOPLE▾ RESEARCH▾ OUTREACH NEWS CRCS EVENTS▾ APPLY

HOME / CALENDAR /

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

"Digital Forensics Innovation: Searching A T...

Our numbers from sampling are similar to those reported by iTunes.

We accurately determined:

Missed a seminar?

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

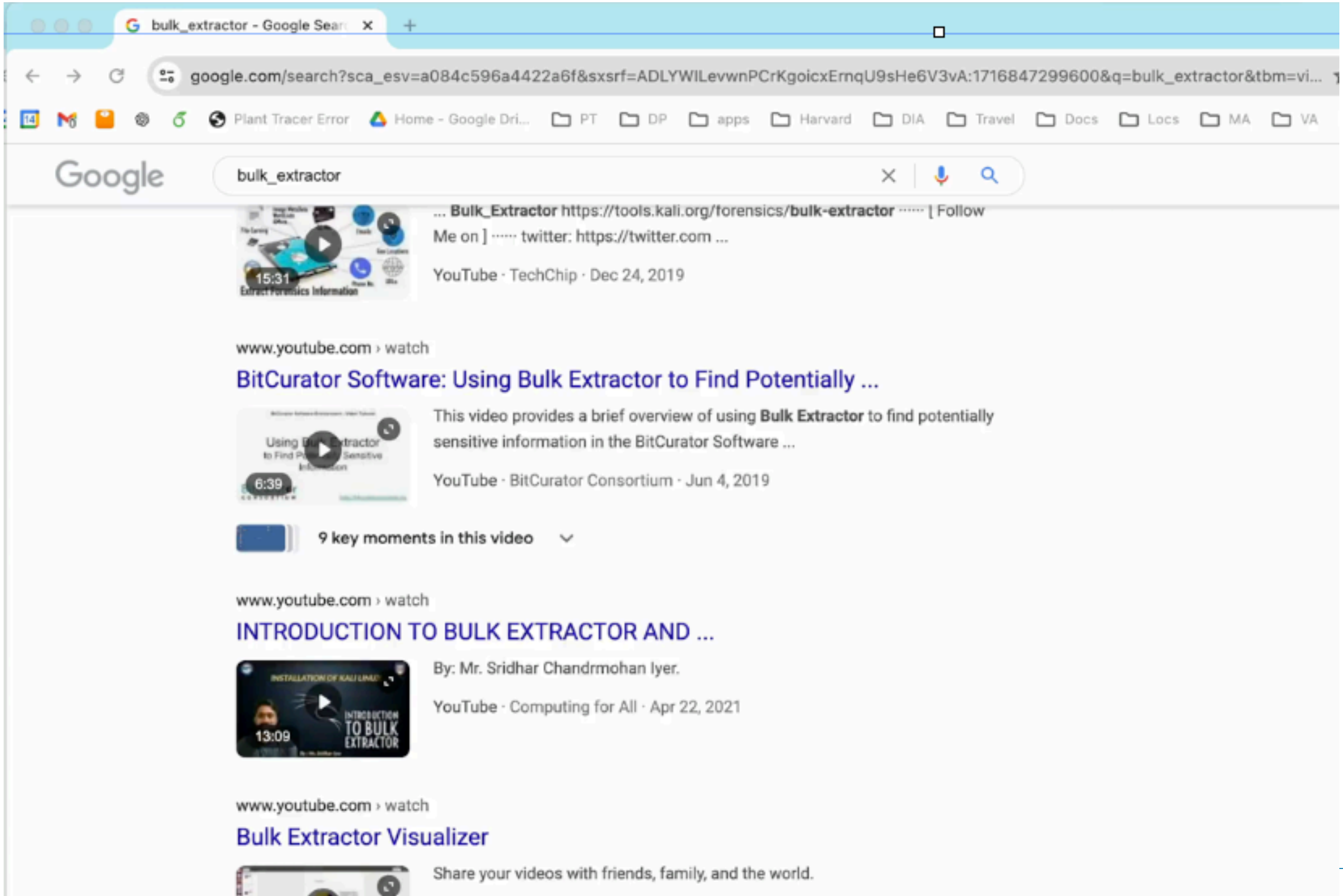
Email CRCS@seas.harvard.edu to join our mailing list and add our Google Calendar to hear about future seminars!

AI for Conservation Workshop - October 2022

October 1, 2012

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

bulk_extractor is an important tool in digital humanities.



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
Google
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.

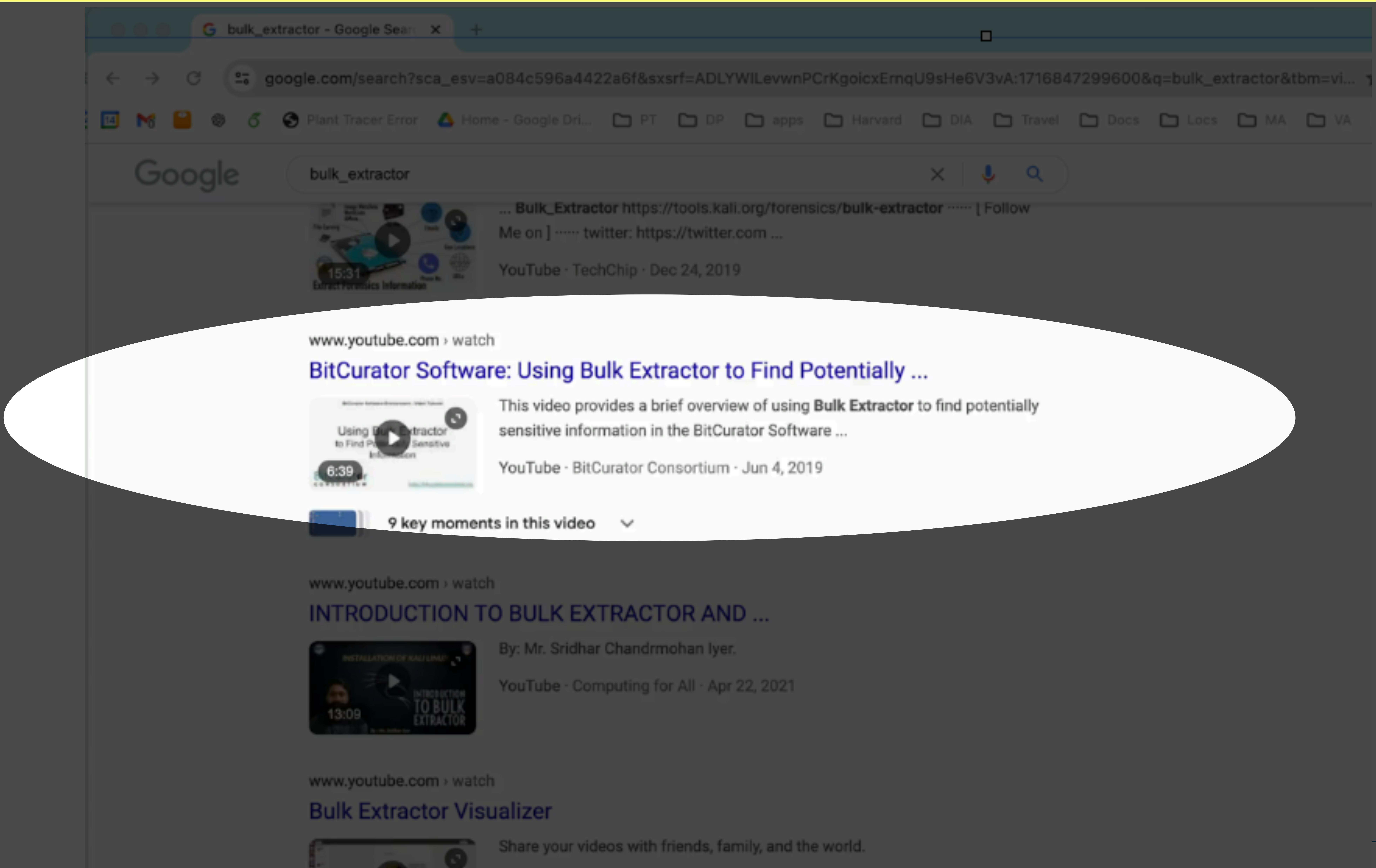
YouTube · Computing for All · Apr 22, 2021

www.youtube.com › watch

Bulk Extractor Visualizer

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

bulk_extractor is an important tool in digital humanities.


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
Google
bulk_extractor
15:31
6:39
13:09

... 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

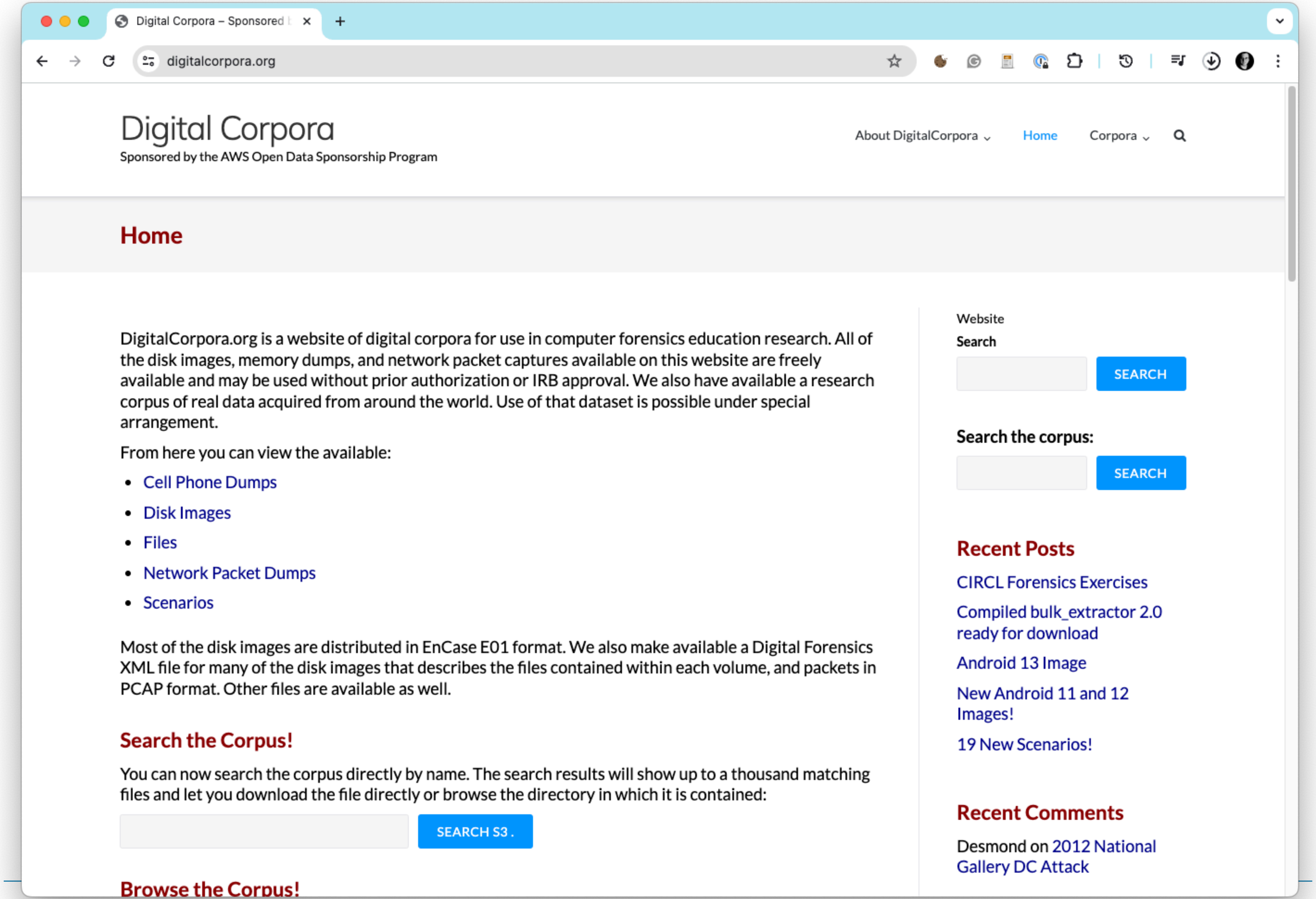
www.youtube.com › watch

Bulk Extractor Visualizer

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

#2
Innovating with open data

Creating, maintaining and distributing open data became another aspect of teaching entrepreneurship.



Digital Corpora

Sponsored by the AWS Open Data Sponsorship Program

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:

Browse the Corpus!

Website Search

Search the corpus:

Recent Posts

- CIRCL Forensics Exercises
- Compiled bulk_extractor 2.0 ready for download
- Android 13 Image
- New Android 11 and 12 Images!
- 19 New Scenarios!

Recent Comments

Desmond on 2012 National Gallery DC Attack

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.

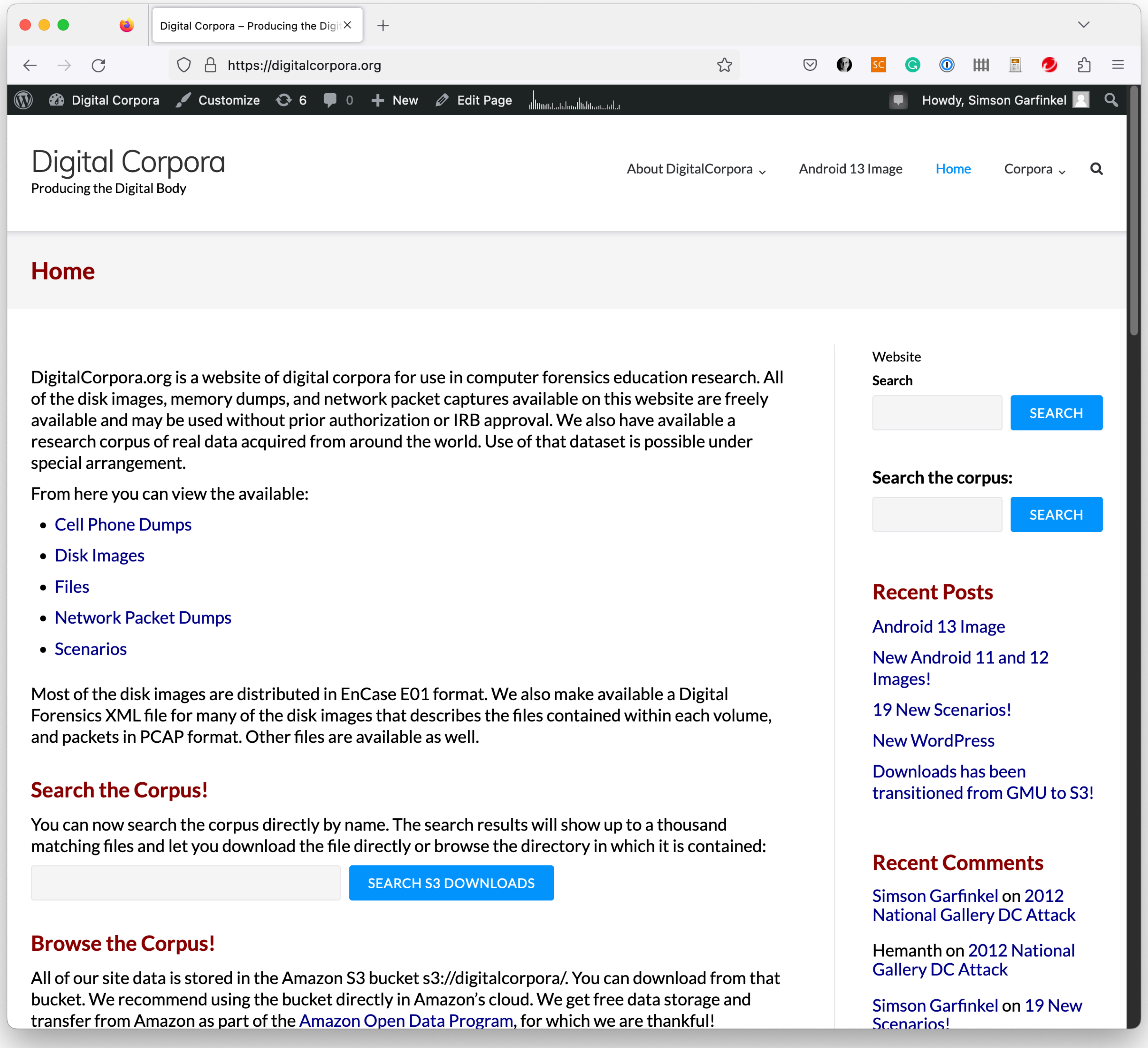
https://digitalcorpora.org/

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.

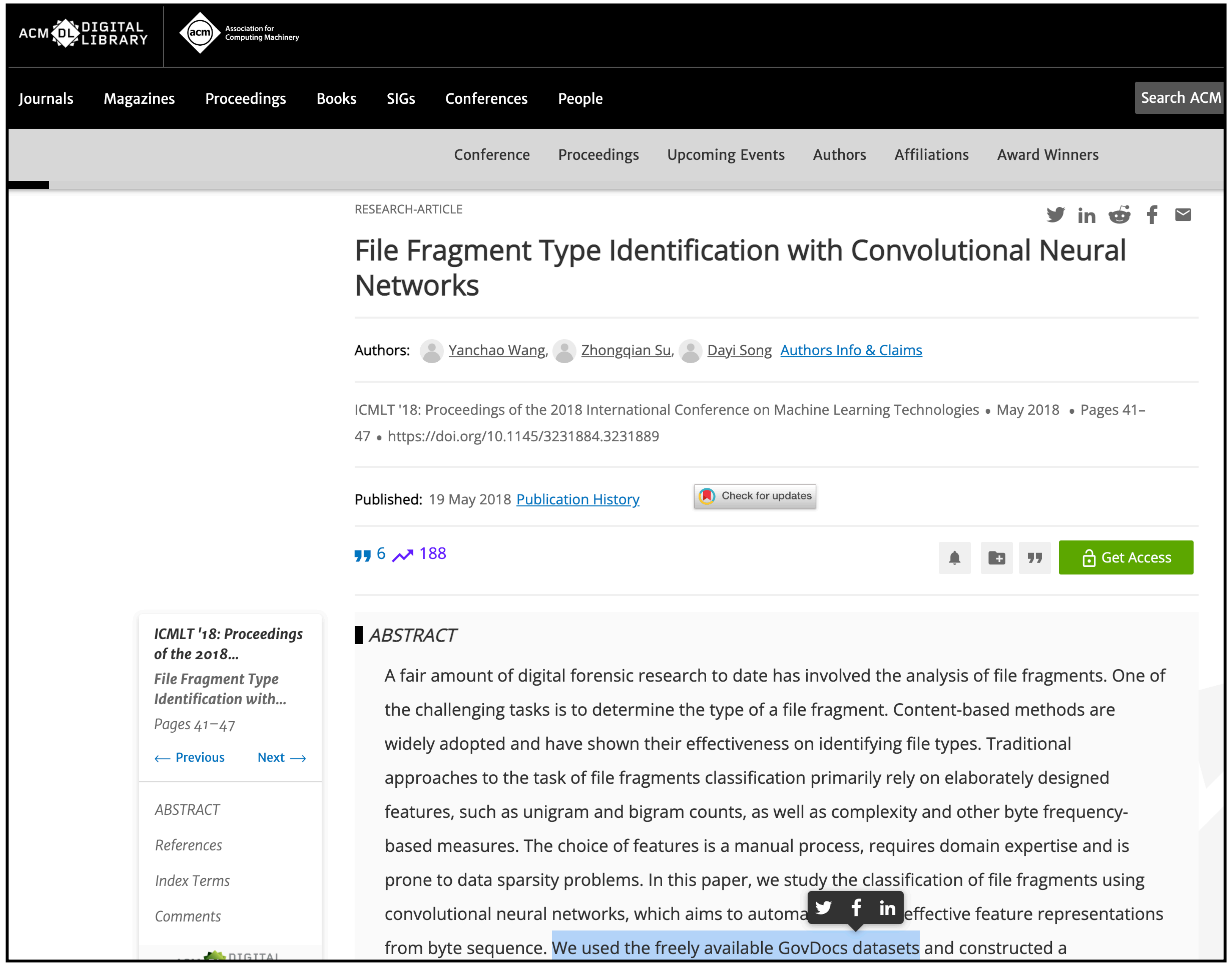
Includes:

- 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



RESEARCH-ARTICLE

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

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

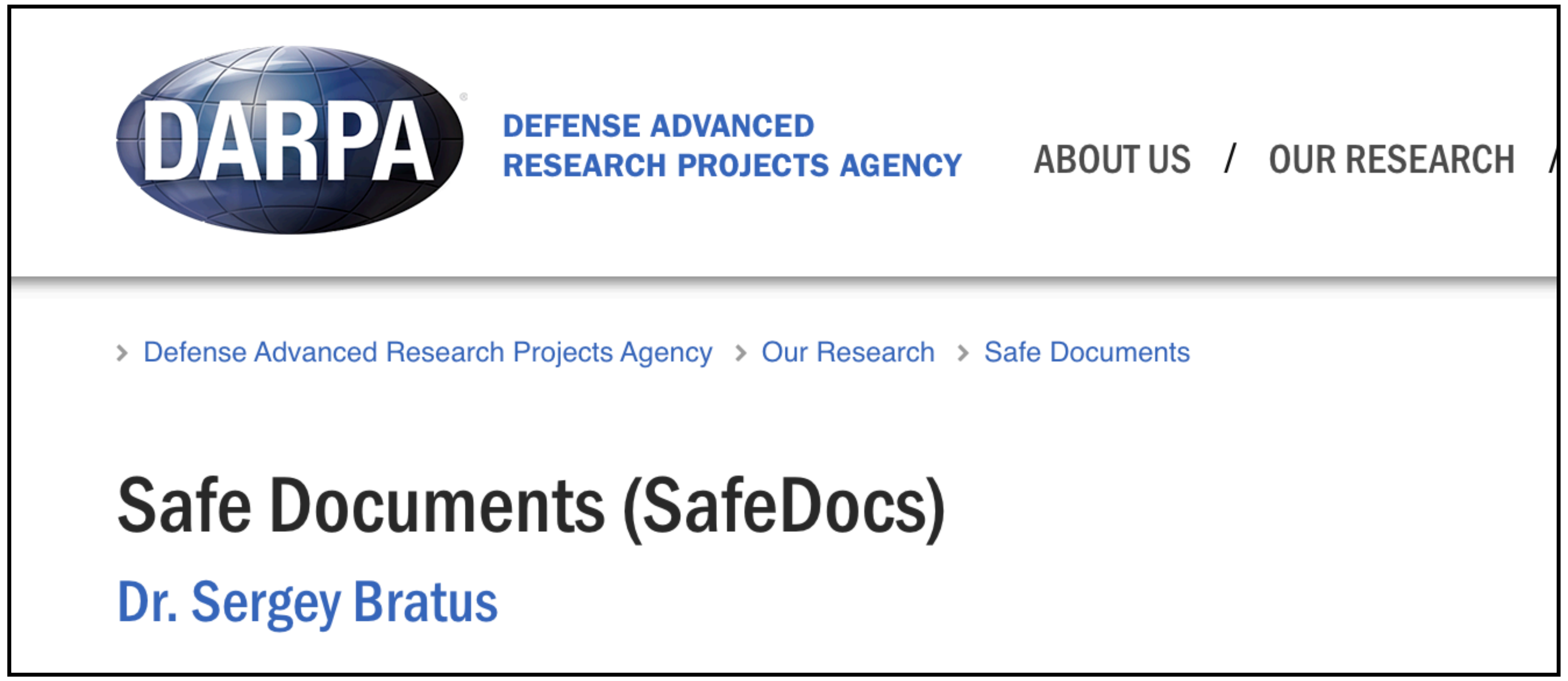
ABSTRACT

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

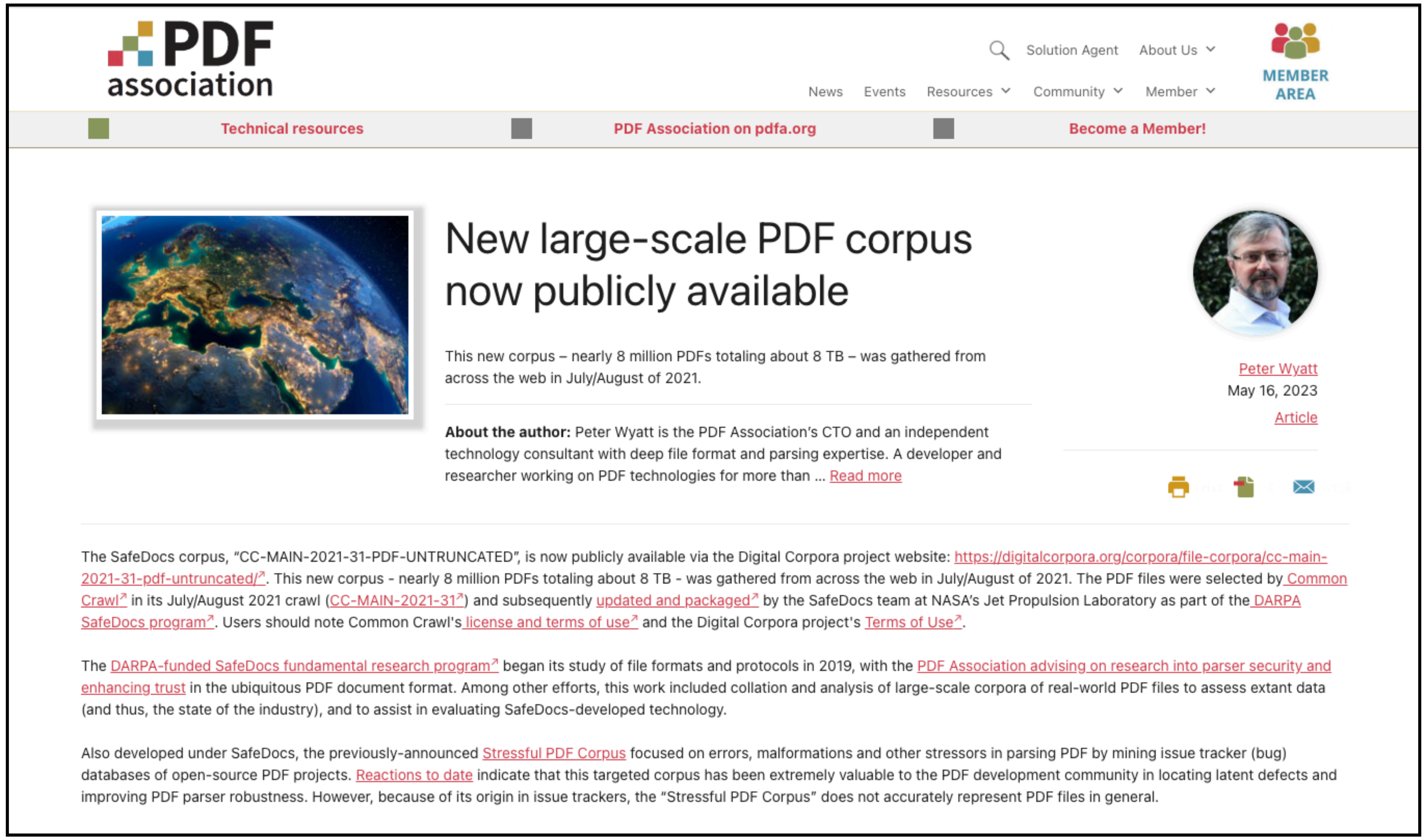
2018-2023

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.

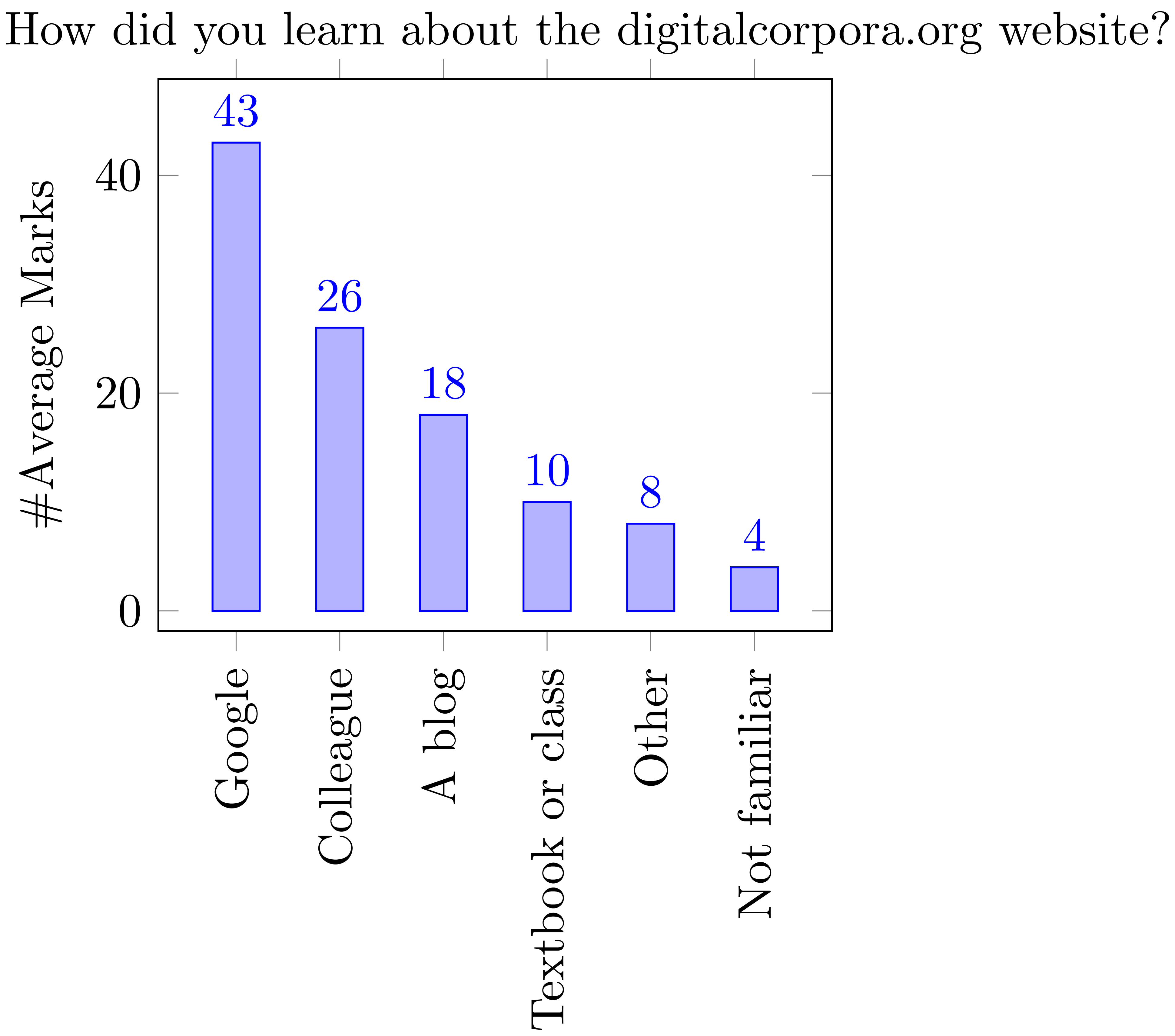
May 16, 2023

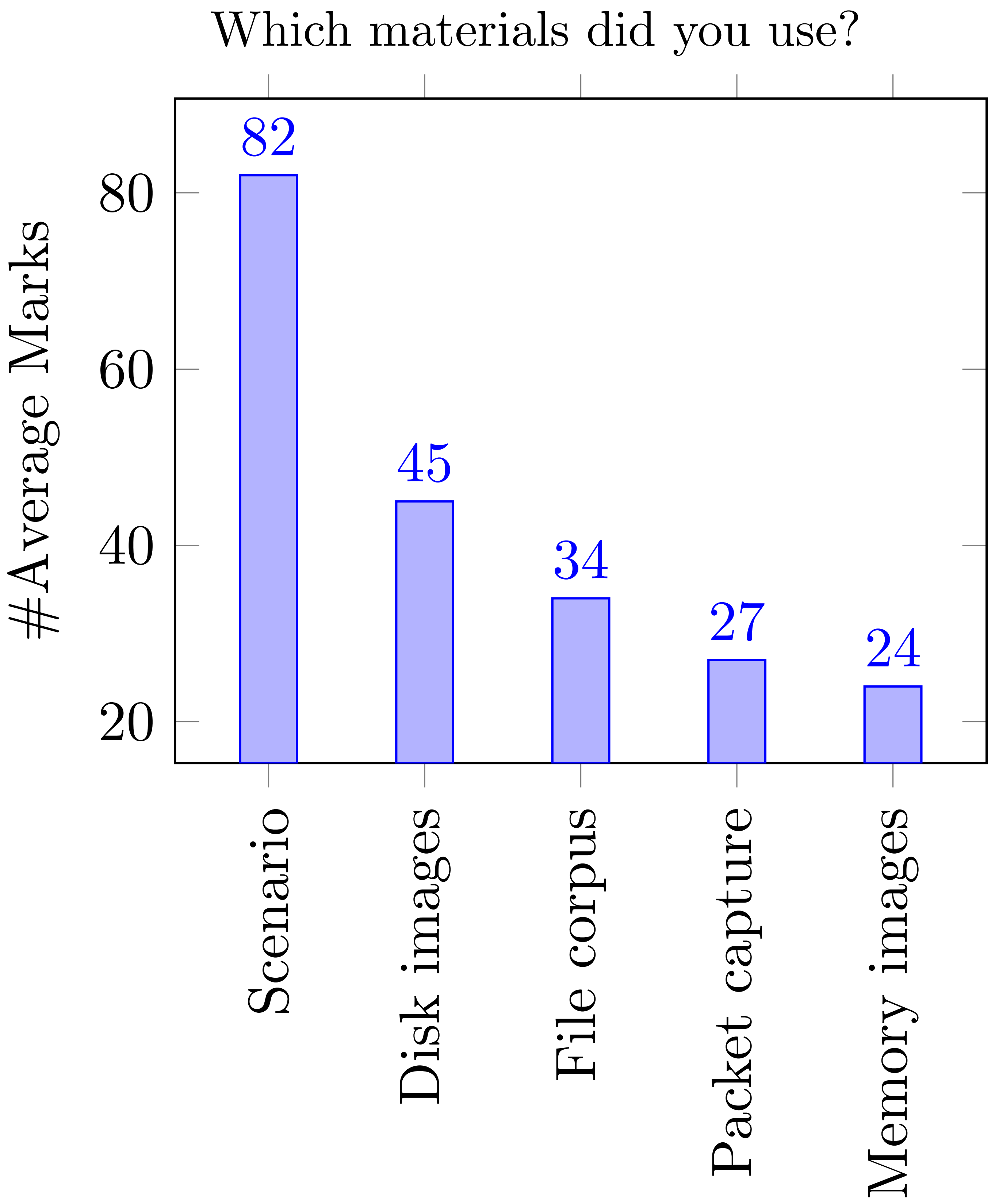
Digital Corpora: Educational Impact

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…

For which did you download a teacher's guide?

Guide	Count
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

The digital corpora project also teaches innovation and entrepreneurship.

Developing scenarios that will be useful to others.

Planning and executing a complex project.

Quantifying the impact

- Woods, Kam, Christoper Lee, Simson Garfinkel, Extending Digital Repository Architectures to Support Disk Image Preservation and Access, JCDL 2011, June 13-17, 2011, Ottawa, Canada.
- Woods, K., Christopher Lee, Simson Garfinkel, David Dittrich, Adam Russel, Kris Kearton, Creating Realistic Corpora for Forensic and Security Education, 2011 ADFSL Conference on Digital Forensics, Security and Law

My career is a testament to the power of open source and open data.

Career #1: Science writer (1985-)

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

Open data and free access to historical archives powers all of my journalism & historical work.

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 —
- Broadband2Wireless — 2000-2001

Open source policy got my code out of MIT.

Open source software was critical for the success of Vineyard.NET and Sandstorm.

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

Open source policies within the US Government made it possible to rapidly transition software from my lab to DOD, FBI, Secret Service, and other government agencies.

Career #4: Government Innovation

- US Census Bureau — 2017-2021 — Differential privacy
- US DHS — 2021-2022 — DHS Data Inventory

Open source and open data policies made it easy to share code and data with the American people.

I use open source and open data to teach innovation and how to be entrepreneurial.

Traps

Beware the support tail

Beware the sunk cost fallacy

Don’t over-estimate your customers

Don’t be greedy

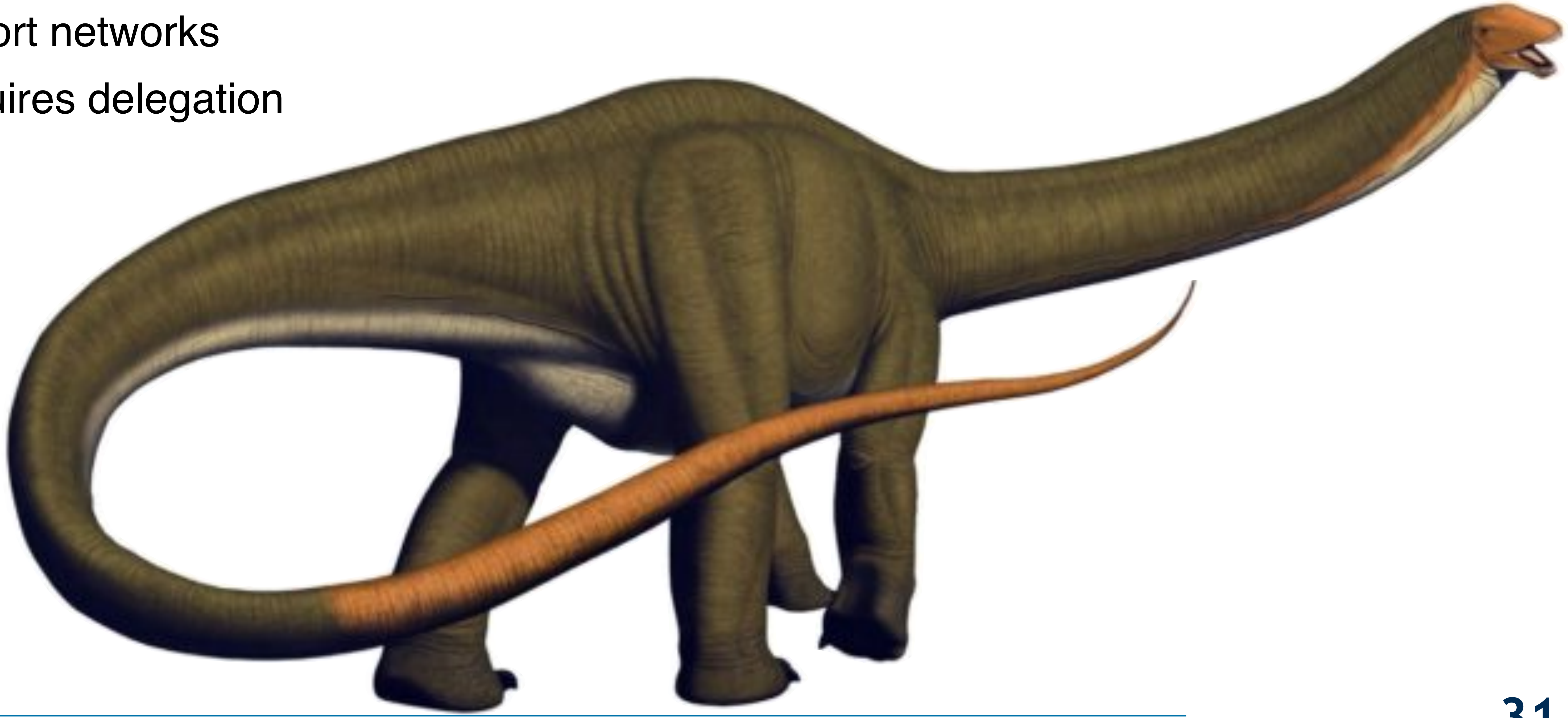
Trap — Beware the support tail

Early advice (bad): “You don’t want to have users.”

- Actually, you do.
- You don’t want to have users who demand support

Avoid this trap: You don’t want to support your users — you want your users’ support.

- Get more people involved
- Clearly distinguish research & deployment
- Plant the seeds for self-sustaining mutual support networks
- Delegate, delegate, delegate — innovation requires delegation



Trap — Beware the sunk cost fallacy

“The reason that God was able to create the world in seven days is that he didn't have to worry about the installed base.”

Things that get in the way of innovation:

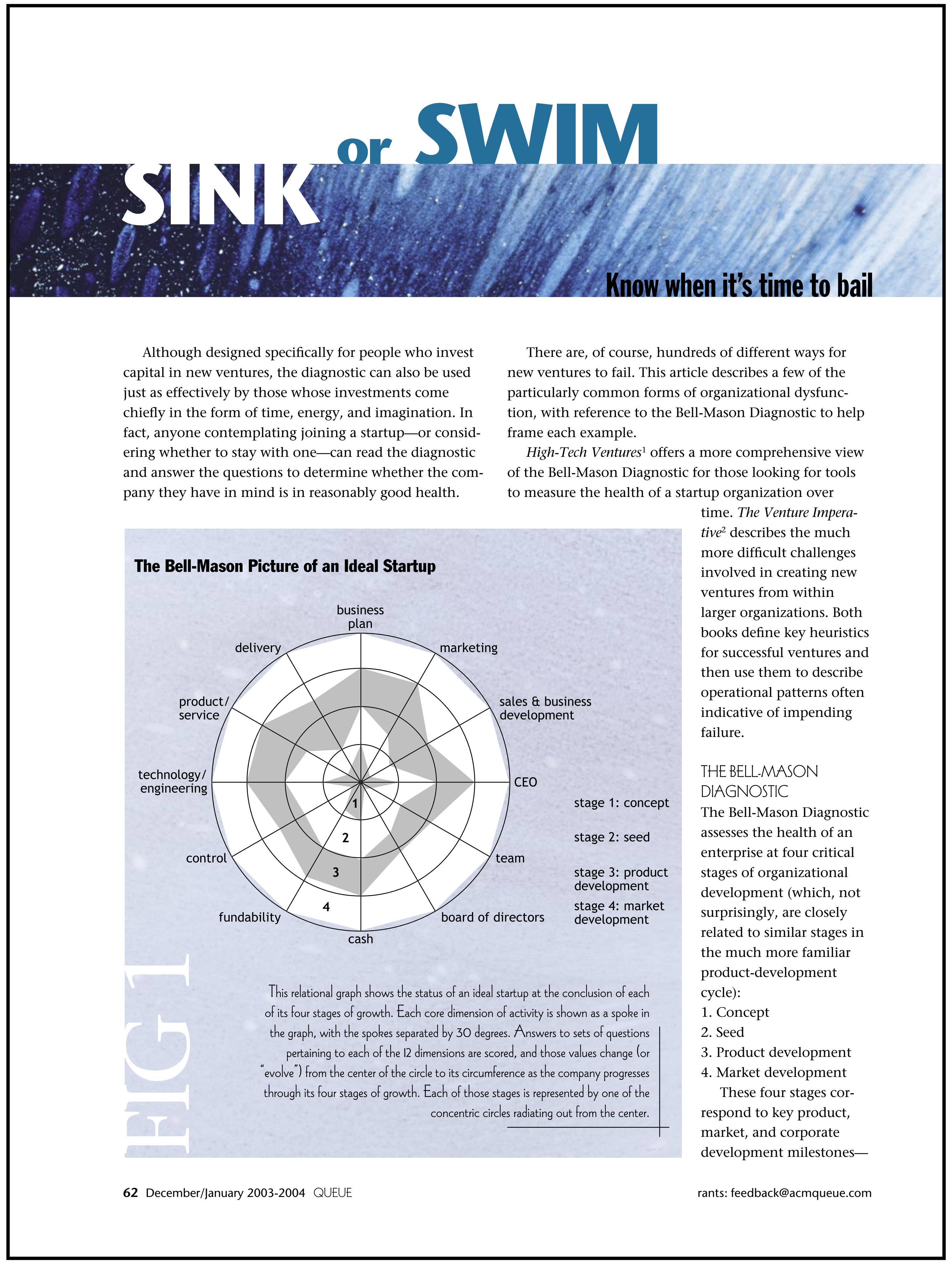
- Existing code-base
- Deep expertise in tech stacks no longer in vogue*

Example — Apple iPhoto & Aperture

- 2002 - 2015 — iPhoto was Apple’s primary digital photo application
- 2005 - 2014 — Aperture was Apple’s “professional” ($$$) photo editor
- 2015 — Apple killed both; replaced with Photos
- Photos was — less complex, integrated w/ iPhone and iCloud
- Cleaner Photos made possible more innovation

Avoid this trap: Never be afraid to start over

—a.k.a. “kill your darlings.”



SINK or SWIM

Know when it’s time to bail

Although designed specifically for people who invest capital in new ventures, the diagnostic can also be used just as effectively by those whose investments come chiefly in the form of time, energy, and imagination. In fact, anyone contemplating joining a startup—or considering whether to stay with one—can read the diagnostic and answer the questions to determine whether the company they have in mind is in reasonably good health.

There are, of course, hundreds of different ways for new ventures to fail. This article describes a few of the particularly common forms of organizational dysfunction, with reference to the Bell-Mason Diagnostic to help frame each example.

High-Tech Ventures[1] offers a more comprehensive view of the Bell-Mason Diagnostic for those looking for tools to measure the health of a startup organization over time. *The Venture Imperative*[2] describes the much more difficult challenges involved in creating new ventures from within larger organizations. Both books define key heuristics for successful ventures and then use them to describe operational patterns often indicative of impending failure.

The Bell-Mason Picture of an Ideal Startup

FIG 1

This relational graph shows the status of an ideal startup at the conclusion of each of its four stages of growth. Each core dimension of activity is shown as a spoke in the graph, with the spokes separated by 30 degrees. Answers to sets of questions pertaining to each of the 12 dimensions are scored, and those values change (or “evolve”) from the center of the circle to its circumference as the company progresses through its four stages of growth. Each of those stages is represented by one of the concentric circles radiating out from the center.

THE BELL-MASON DIAGNOSTIC

The Bell-Mason Diagnostic assesses the health of an enterprise at four critical stages of organizational development (which, not surprisingly, are closely related to similar stages in the much more familiar product-development cycle):

1. Concept
2. Seed
3. Product development
4. Market development

These four stages correspond to key product, market, and corporate development milestones—

62 December/January 2003-2004 QUEUE

rants: feedback@acmqueue.com

Gordon Bell. 2003. Sink or Swim: Know When It’s Time to Bail: A diagnostic to help you measure organizational dysfunction and take action. Queue 1, 9 (December/January 2003-2004), 60–67. https://doi.org/10.1145/966789.966806

Trap — Don’t over-estimate your customers (users)

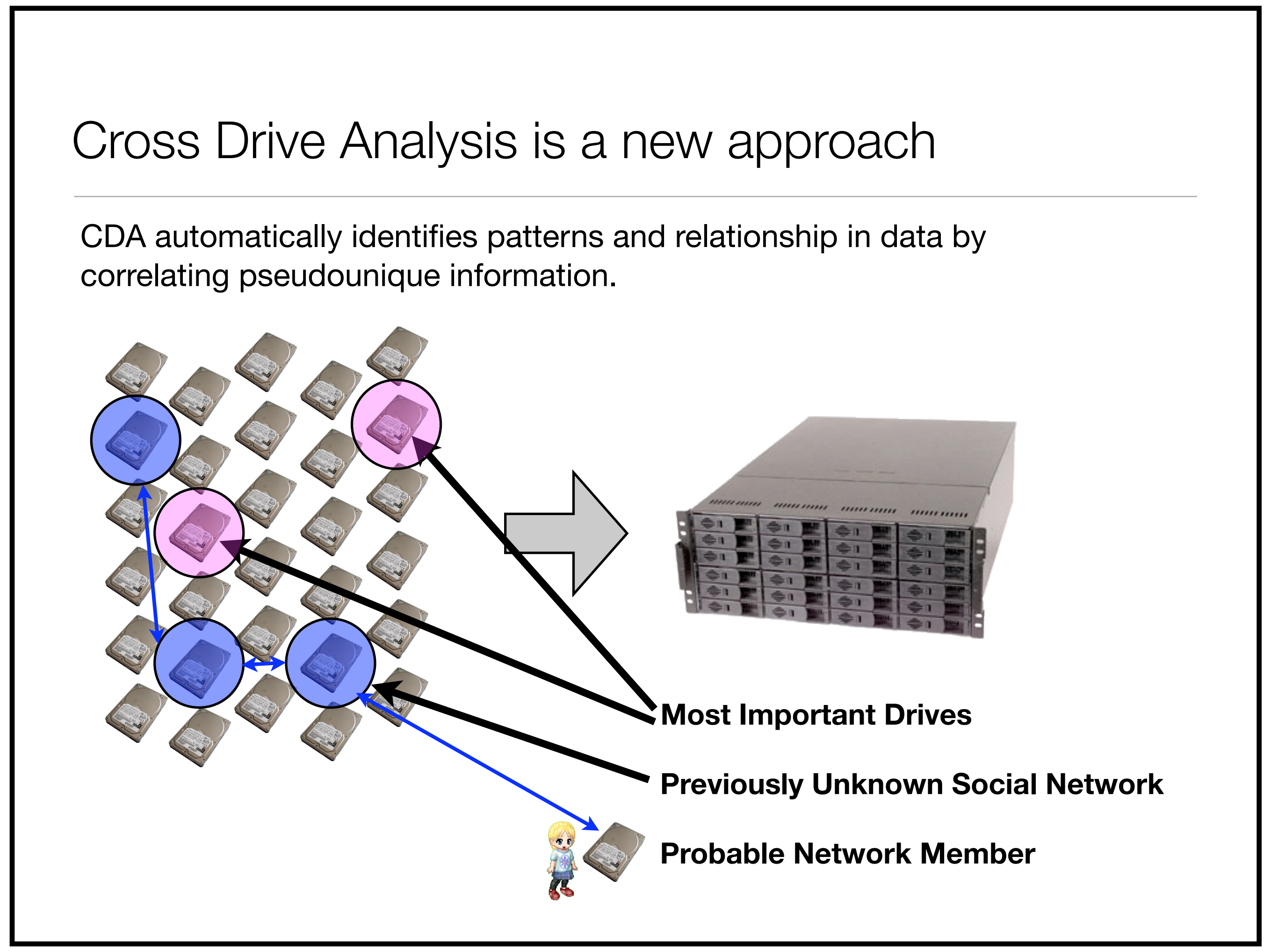
2005 — I developed “cross-drive analysis” (CDA)

- Technique for finding hard drives & cell phones used by criminals in an investigation.
- Based on correlating **identifiers** (email, credit card #, etc.) between devices.
- Uses TF-IDF

2008 — I’m describing the power of CDA to a potential user.

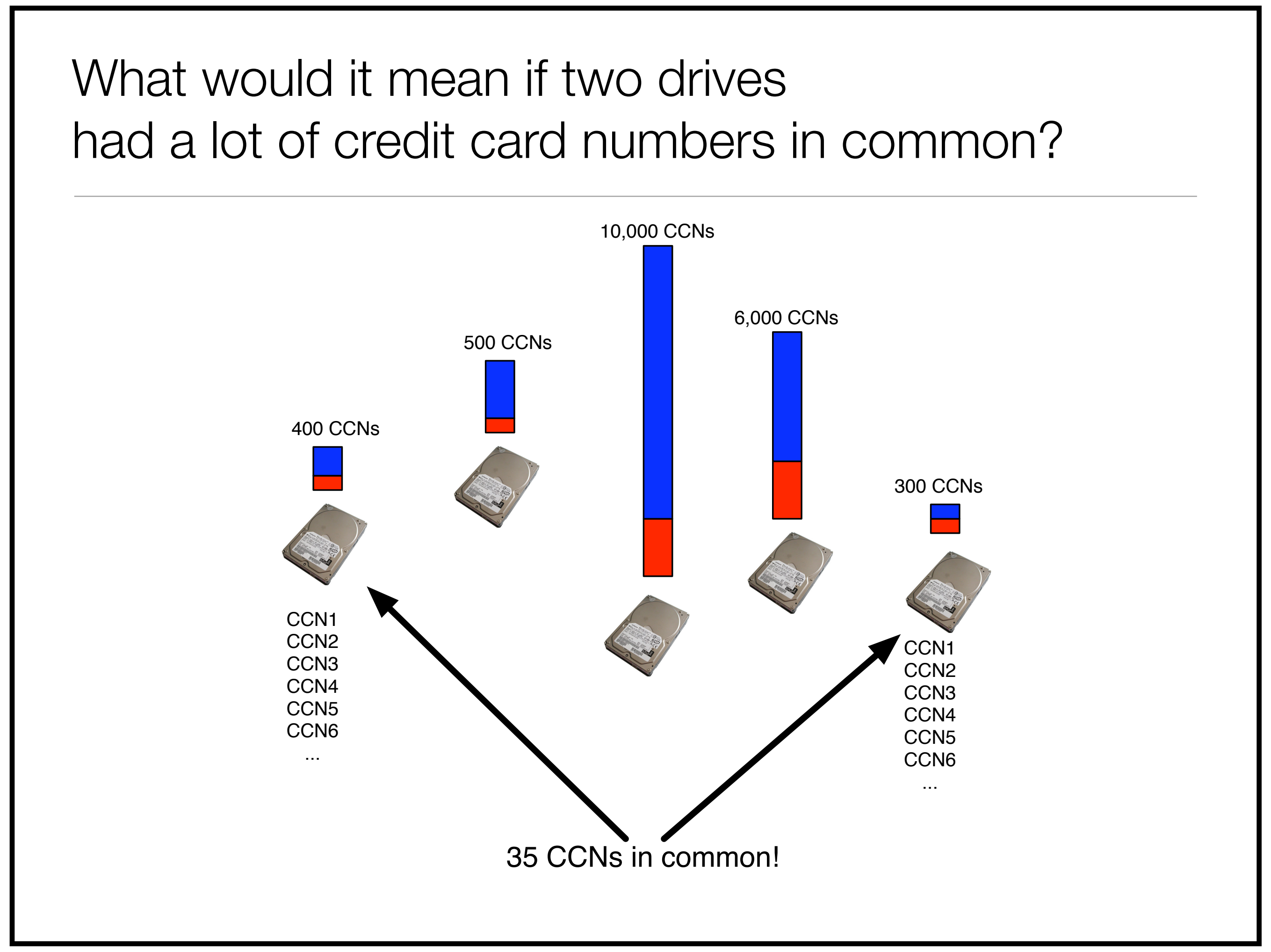
- Me:
  - —*We extract all of the phone numbers and email addresses from the drives.*
  - —*We put them in a massive database.*
  - —*We run this O(n2) algorithm*
  - —*We use TF-IDF*
- Customer:
  - —*You can extract phone numbers and email addresses from a drive automatically?’*
- Deploying automatic extraction at scale was transformative.

Avoid this trap: match your customer’s technology readiness level.



Cross Drive Analysis is a new approach

CDA automatically identifies patterns and relationship in data by correlating pseudounique information.



What would it mean if two drives had a lot of credit card numbers in common?

Slides from 2005 presentation

Trap — Don’t be greedy

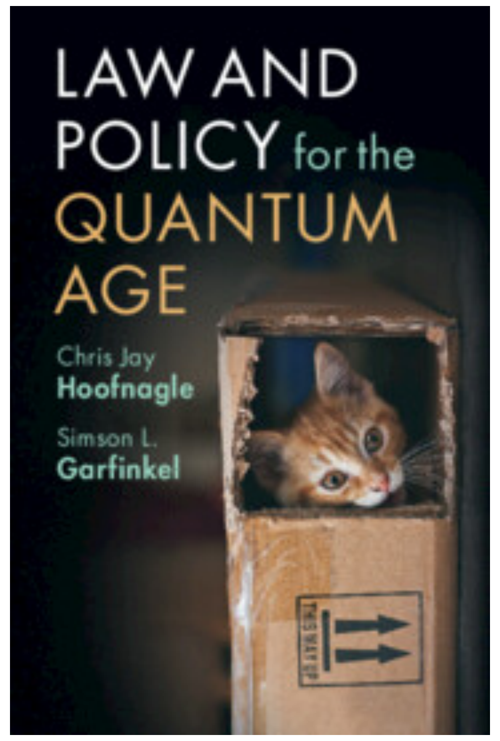
Academic credit — be generous with authorship.

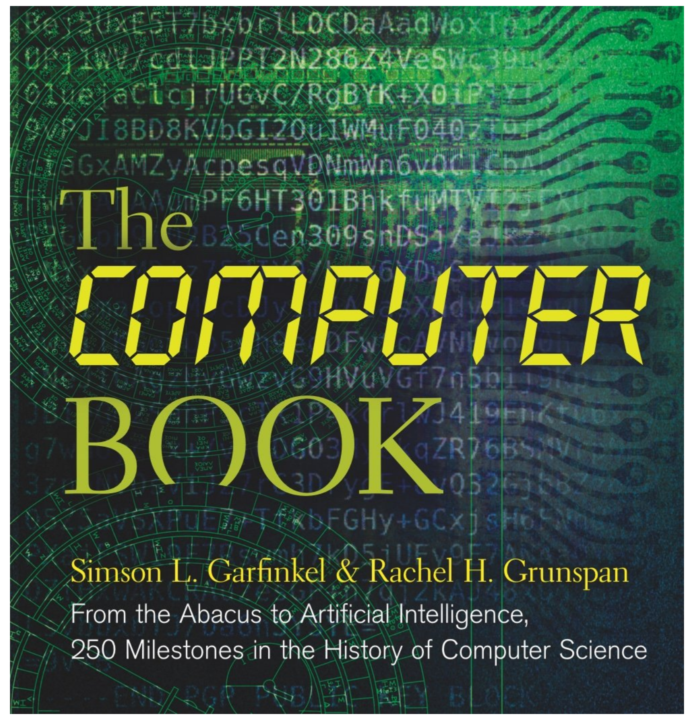
- Instead of listing students in acknowledgements, work to make them co-authors!
- I have multiple masters’ students, undergrads, and 1 high school student as co-authors.
- Engaging with these students about the work is a great way to come up with new ideas.

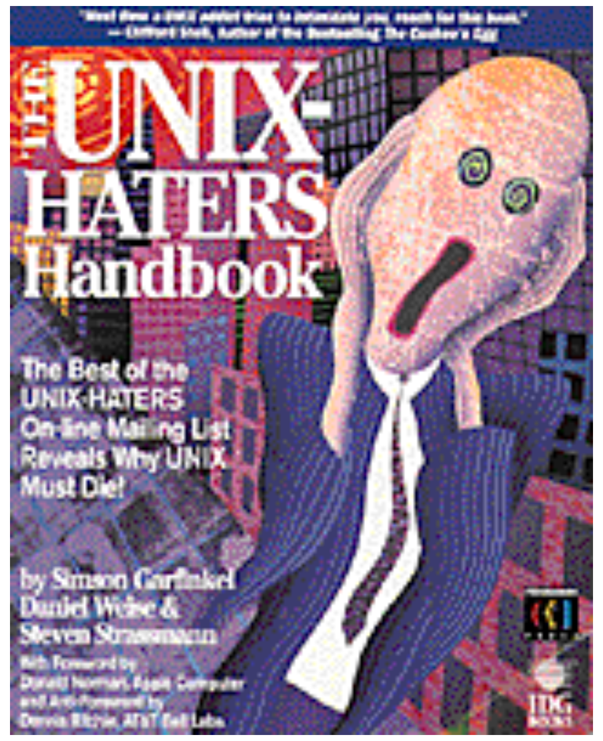
—The best new ideas are conceptual, rather than incremental.

Academic work — seek out co-authors

- Of my 18 books, 15 have co-authors
- My “favorite” books all have co-authors




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


The UNIX-HATERS Handbook
The Best of the UNIX-HATERS On-line Mailing List Reveals Why UNIX Must Die!
by Simson Garfinkel Daniel Weise & Steven Strassmann

Innovation requires multiple POVs

Avoid this trap: Give credit, lower prices, exert less control.

Techniques

Know your customer

Pursue extreme usability

Design and build for maintainability

Collaborate for scale

Teaching innovation by innovating

#1
Know your customer

Technique — Know your customer

You’ve got a great idea — now what?

Product Market Fit (PMF) — What stands out in the lab is rarely what the market needs.

- “Market” — Customers with money / Users with a need / Government agencies with a need

Market research is critical.

Surprisingly, many would-be “innovators” people don’t do this:

- Inventor / Entrepreneur has a great idea and goes out to find a market.
- “Build it and they will come” does not always work.
- Losing money to build market share only works if you can pivot and start making money.
  - —*The success of Google, Facebook, Tesla, and others makes it hard to see all of the similar companies that failed.*

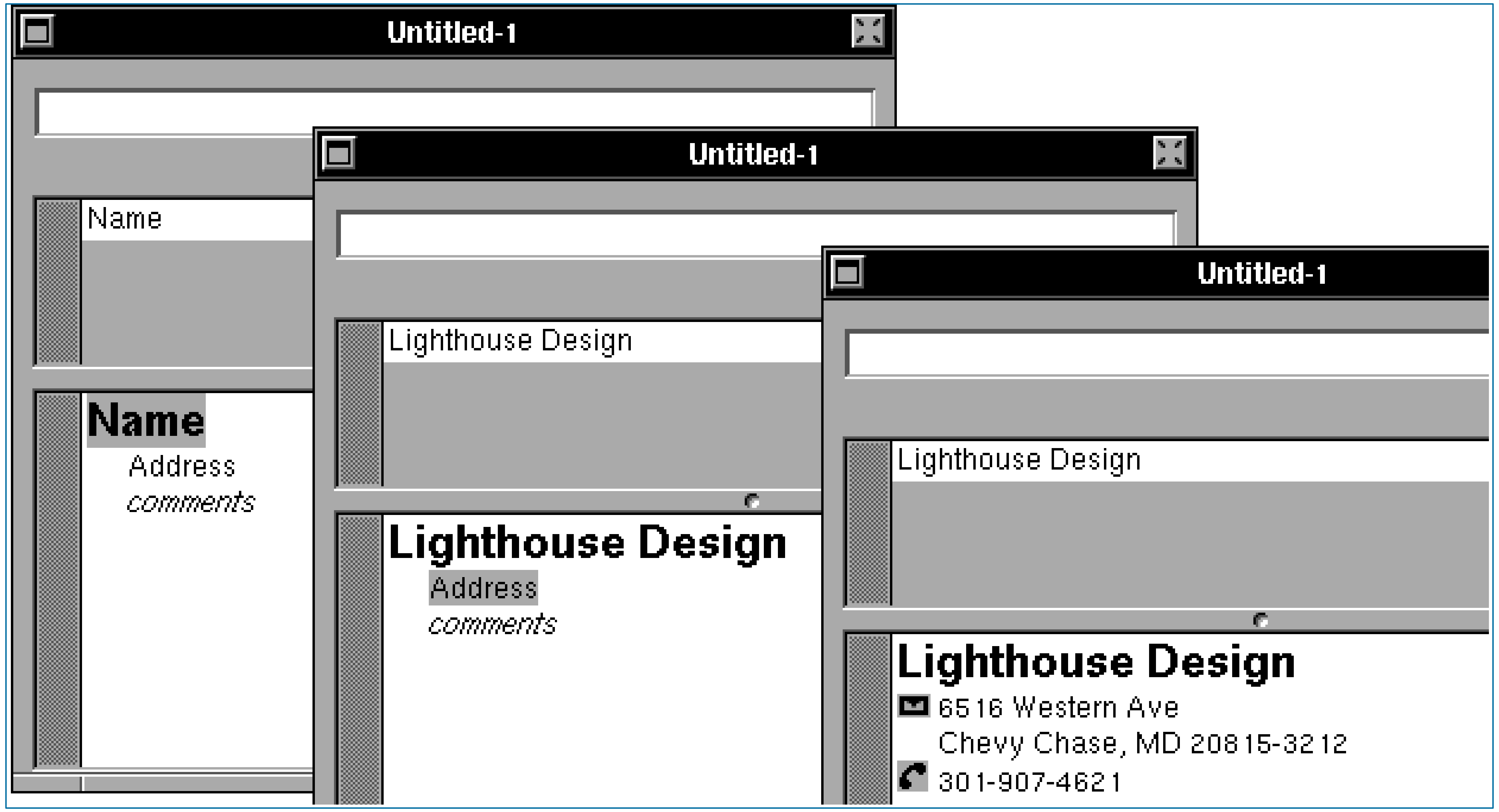
How to teach PMF? Eyeball interviews (something I learned in journalism school)

- Go out and talk to people. In person.
- This is terrifying for many students!

The bulk_extractor market research study

Bulk_extractor was 20 years in the making. It was not the tool that I planned to create!

- In 1991 I developed SBook, a free-format address book for NeXT computers (pre-cursor to MacOS X).



Name

Name
Address
comments

Lighthouse Design

Lighthouse Design
Address
comments

Lighthouse Design

Lighthouse Design
6516 Western Ave
Chevy Chase, MD 20815-3212
☎ 301-907-4621

- SBook used "Named Entity Recognition" to find addresses, phone numbers, email addresses *while you typed.*

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!*

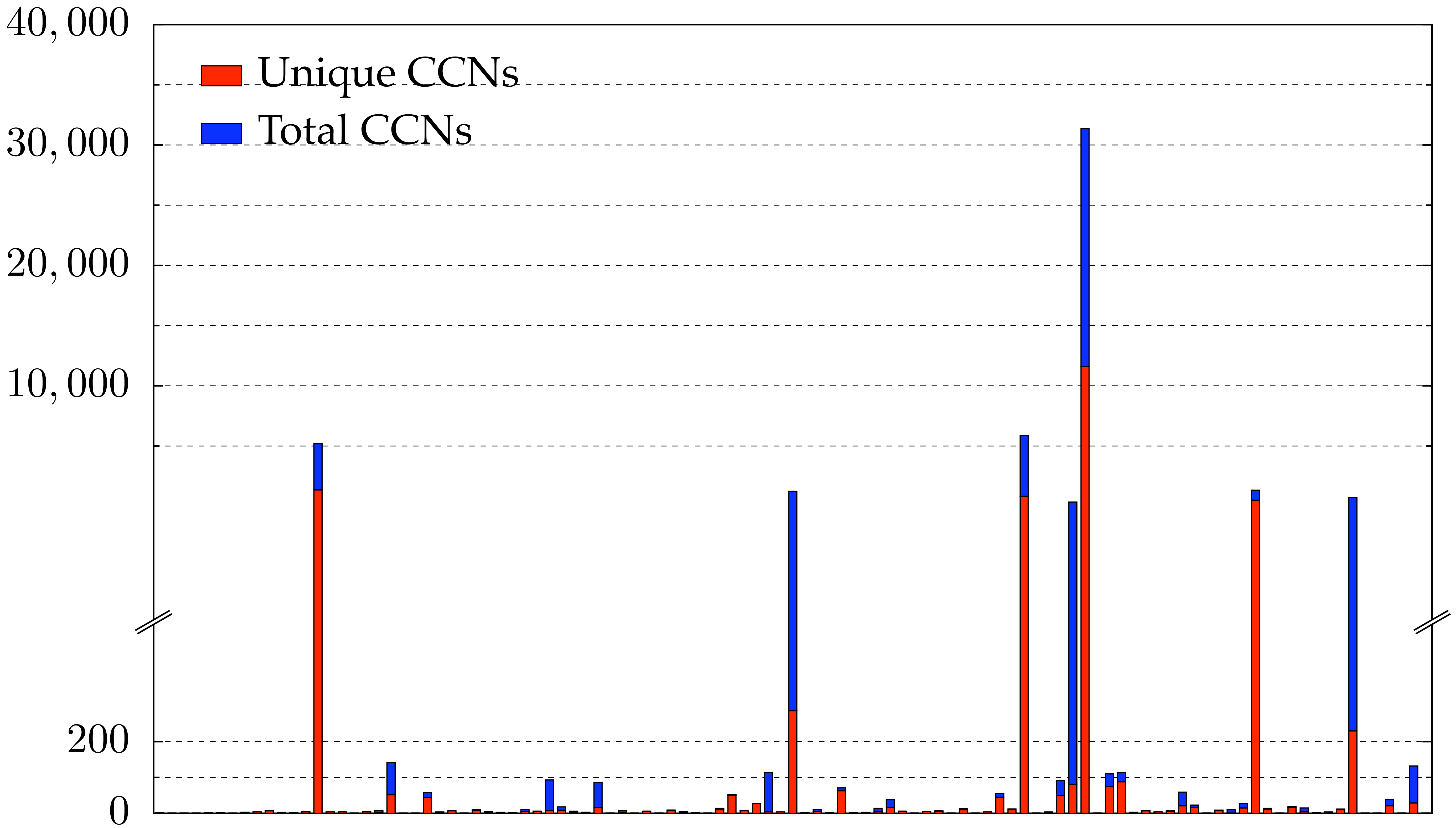
Second strategy:

- 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.

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:

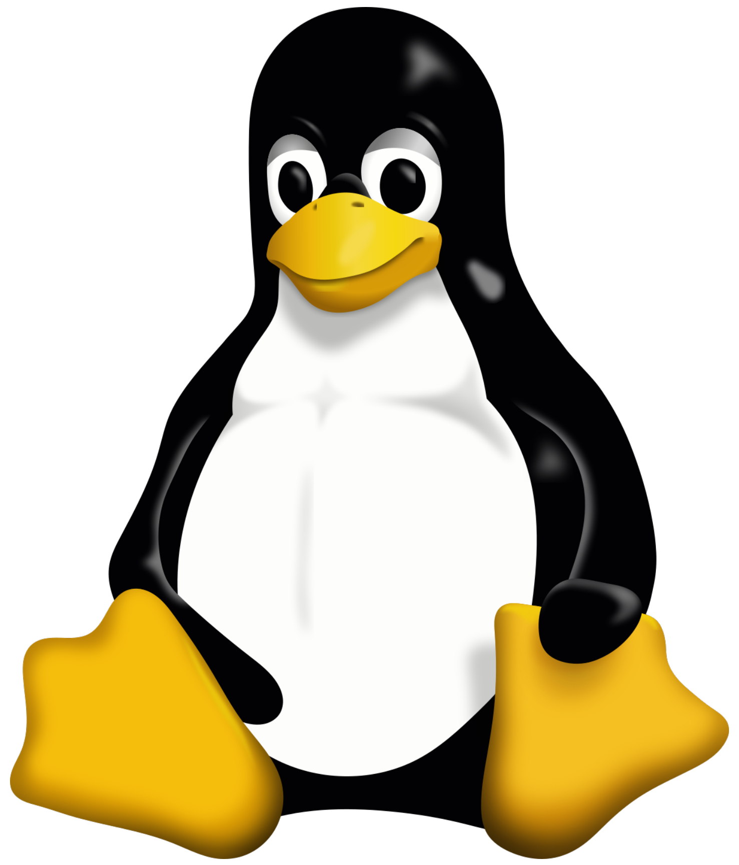
- Email addresses
- Credit card numbers (including track 2 information)
- Search terms (extracted from URLs)
- Phone numbers
- GPS coordinates
- EXIF information from JPEGs
- All words that were present on the disk (for password cracking)

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.

San Luis Obispo
Sincerely, California.

NPS Presentation from 2011-06-14

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/

A bulk_extractor Success Story

City of San Luis Obispo Police Department, Spring 2010

NPS Presentation from 2011-06-14

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!

NPS Presentation from 2011-06-14

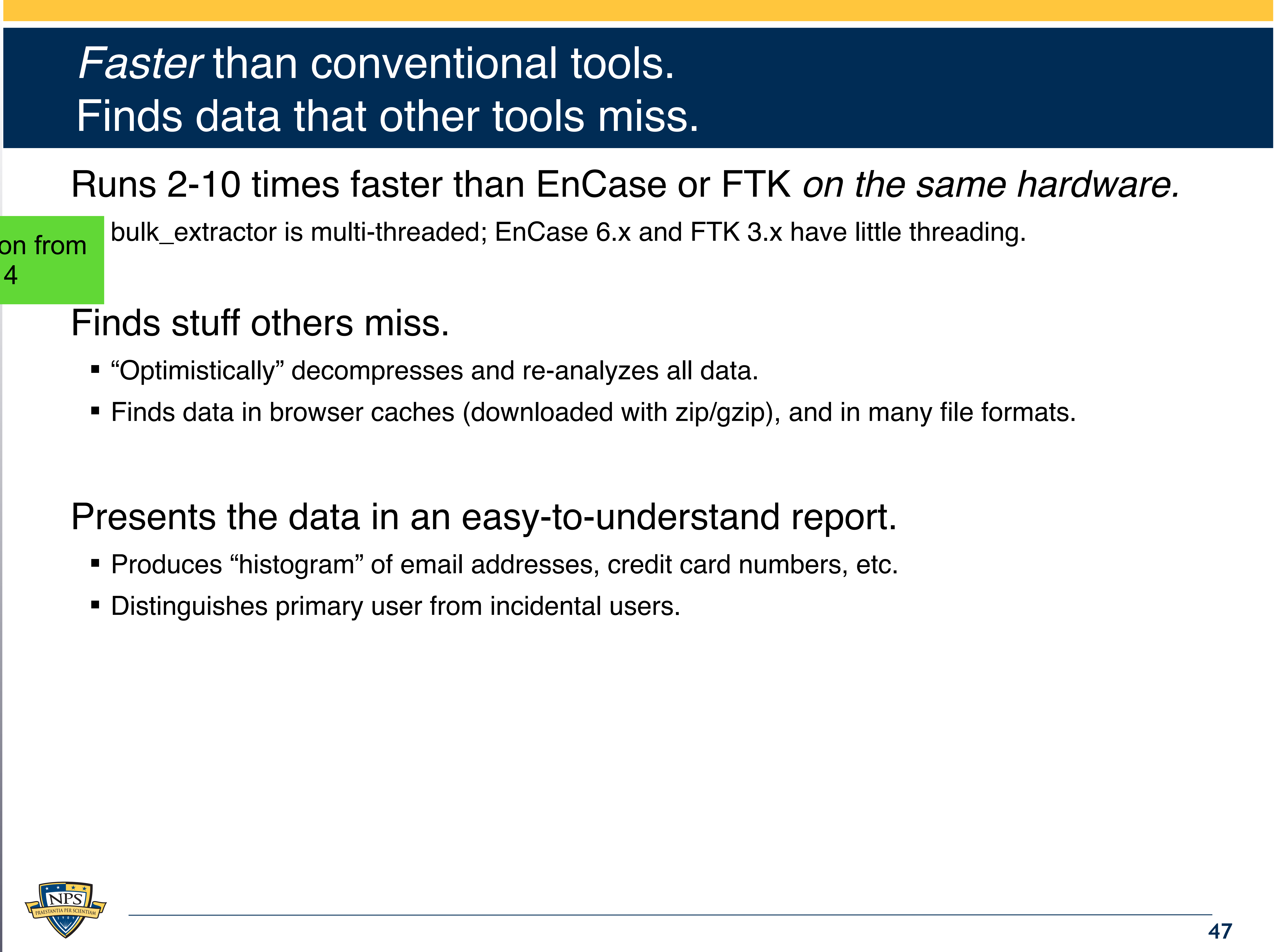
Examiner given 250GiB drive *the day before preliminary hearing.*

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.

NPS Presentation from 2011-06-14



Faster than conventional tools. Finds data that other tools miss.

Runs 2-10 times faster than EnCase or FTK *on the same hardware.*

- bulk_extractor is multi-threaded; EnCase 6.x and FTK 3.x have little threading.

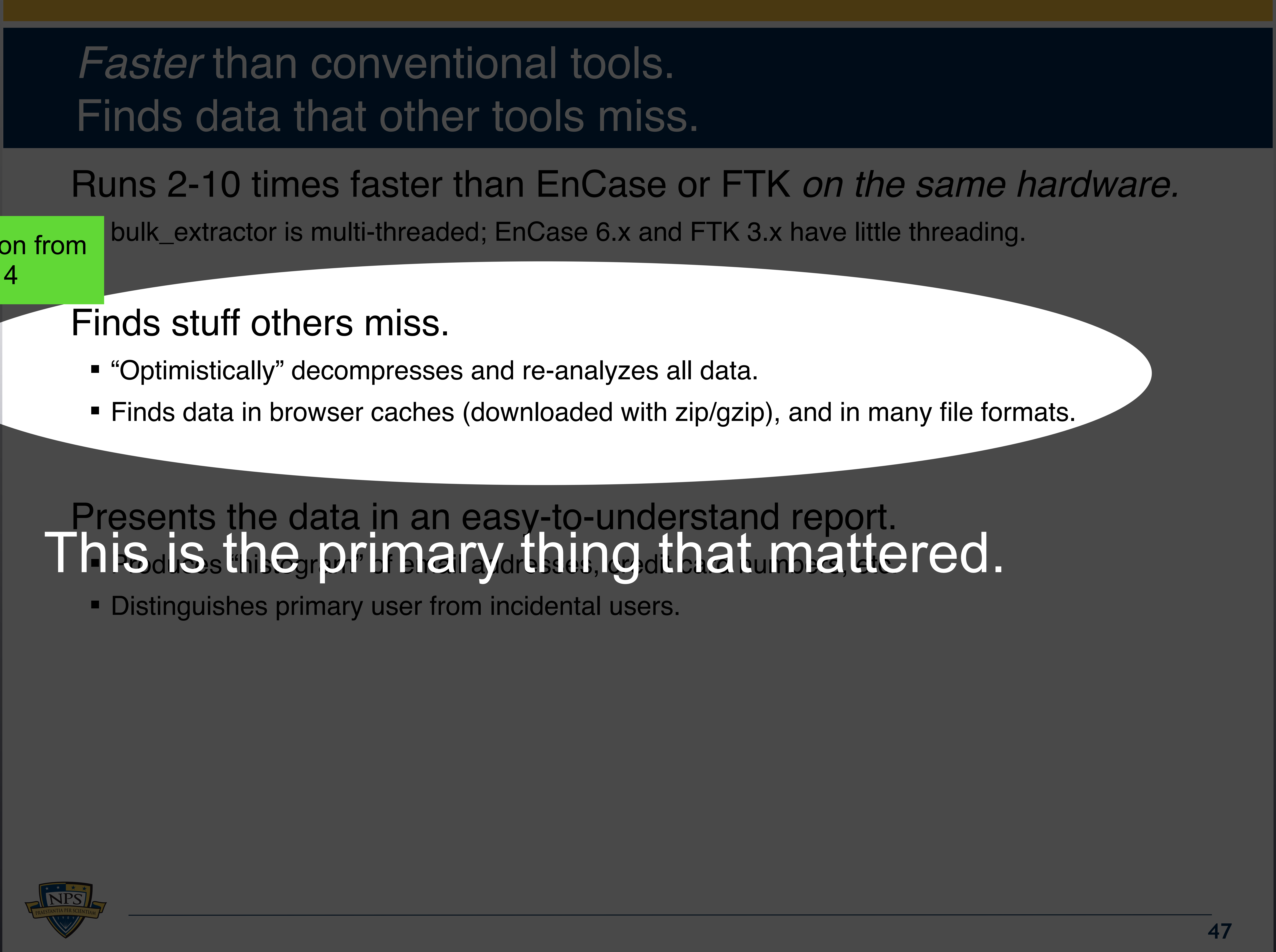
Finds stuff others miss.

- "Optimistically" decompresses and re-analyzes all data.
- Finds data in browser caches (downloaded with zip/gzip), and in many file formats.

Presents the data in an easy-to-understand report.

- Produces "histogram" of email addresses, credit card numbers, etc.
- Distinguishes primary user from incidental users.

NPS Presentation from 2011-06-14



Faster than conventional tools.
Finds data that other tools miss.

Runs 2-10 times faster than EnCase or FTK *on the same hardware.*

- bulk_extractor is multi-threaded; EnCase 6.x and FTK 3.x have little threading.

Finds stuff others miss.

- “Optimistically” decompresses and re-analyzes all data.
- Finds data in browser caches (downloaded with zip/gzip), and in many file formats.

Presents the data in an easy-to-understand report.

- Produces “histogram” of email addresses, credit card numbers, etc.
- Distinguishes primary user from incidental users.

This is the primary thing that mattered.

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!

#2
Extreme Usability

Technique — Pursue extreme usability

Most bulk_extractor requirements from my study were *usability requirements*:

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

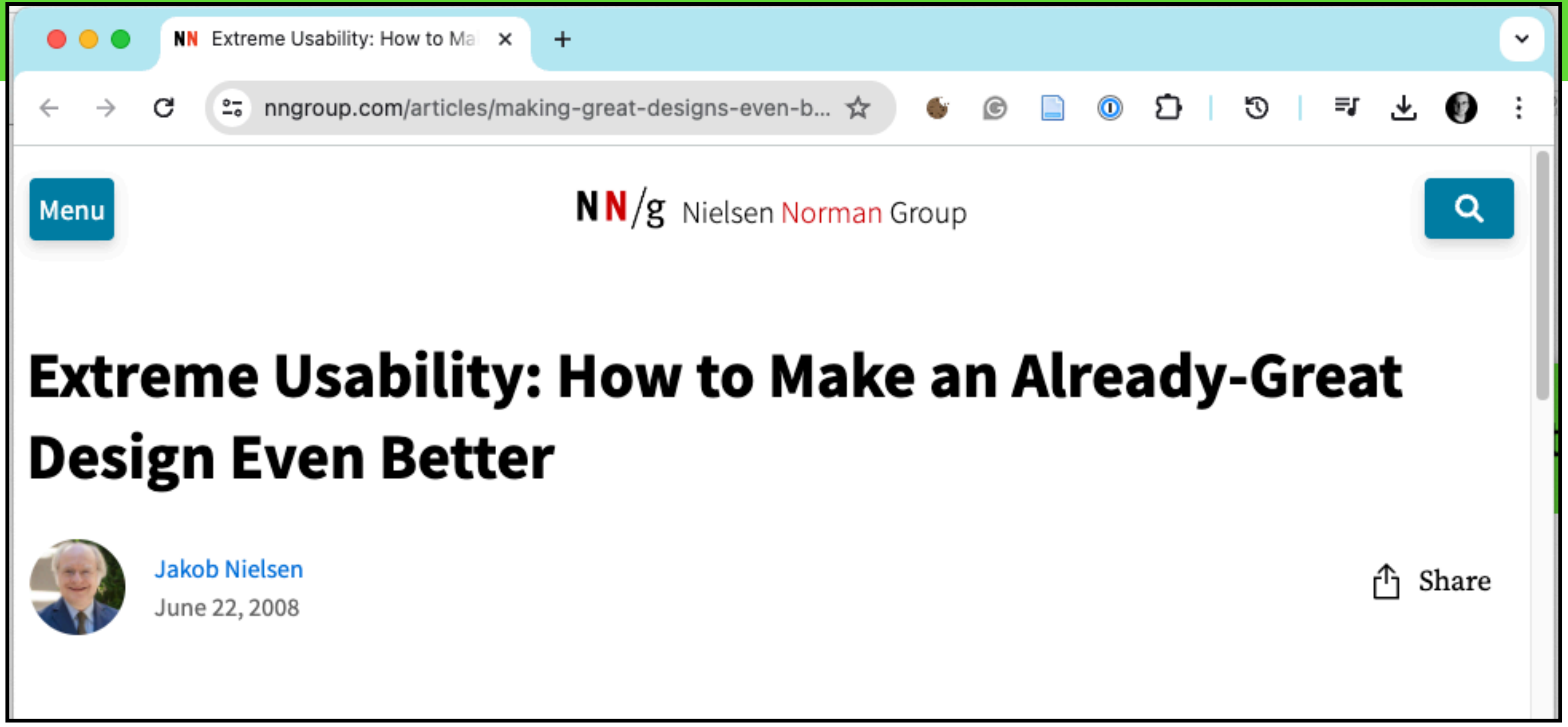
Other usability requirements:

- Run "out of the box" with no training and no configuration.
- Run with no user interaction — A "get evidence" button

Technique — Pursue extreme usability

What is “extreme usability”

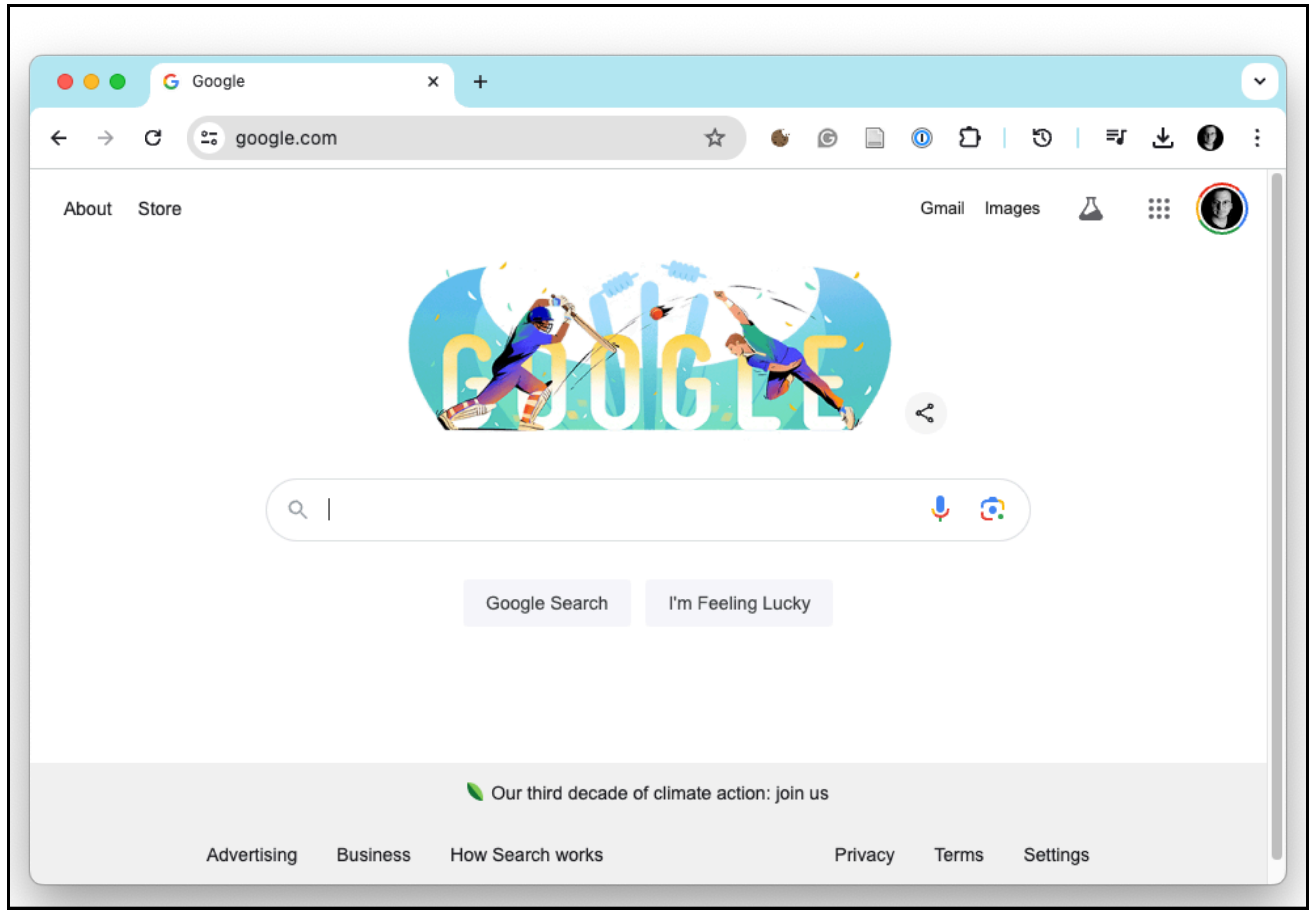
- Look for what’s great, and replicate it.
- Look for what can go wrong, and ensure it never happens.
- Go beyond user experience and “consider enterprise usability”
- “Discover unmet needs”

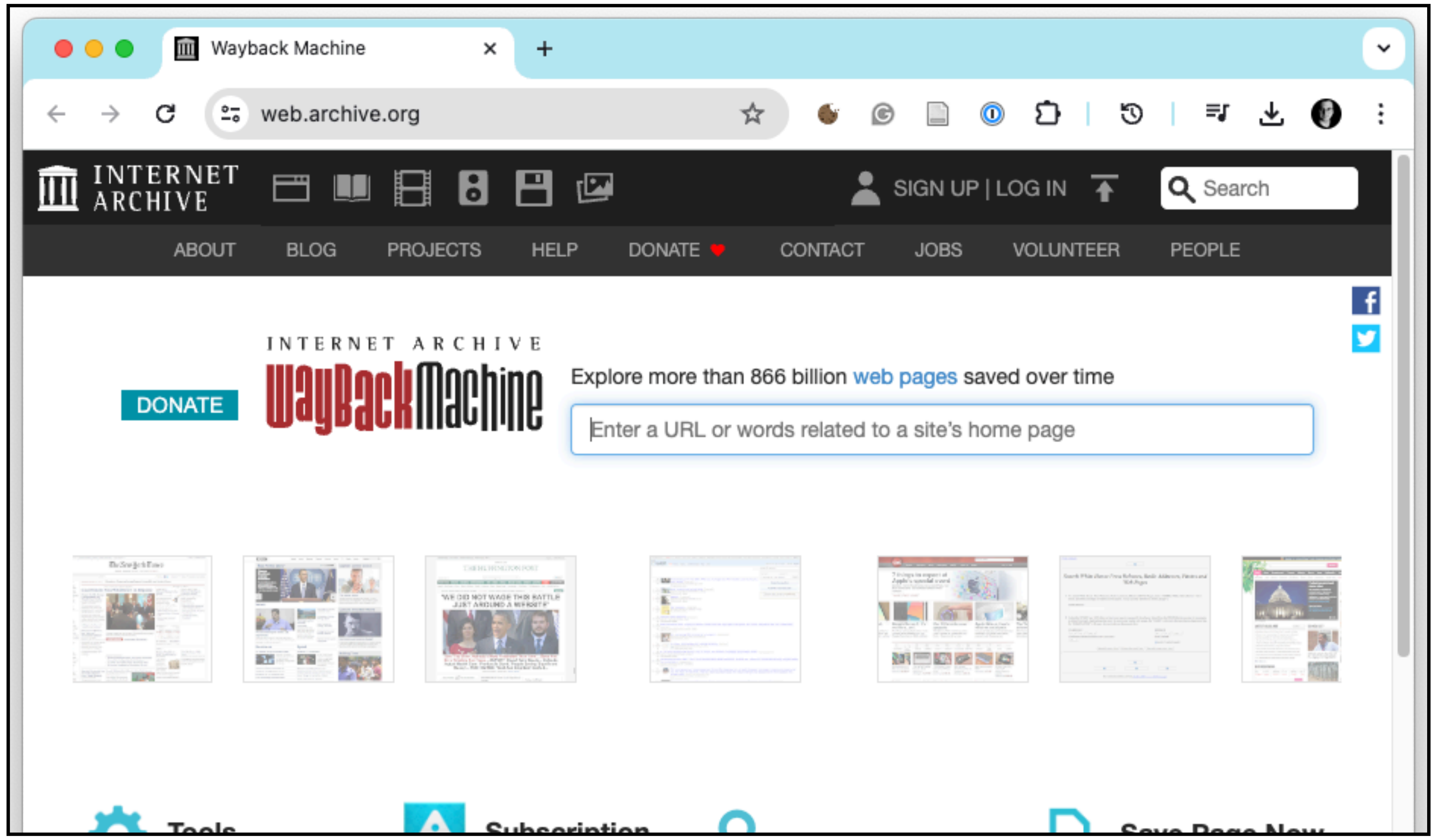


Extreme Usability: How to Ma
nngroup.com/articles/making-great-designs-even-b...
Menu
NN/g Nielsen Norman Group
Share

Extreme Usability: How to Make an Already-Great Design Even Better

Jakob Nielsen
June 22, 2008





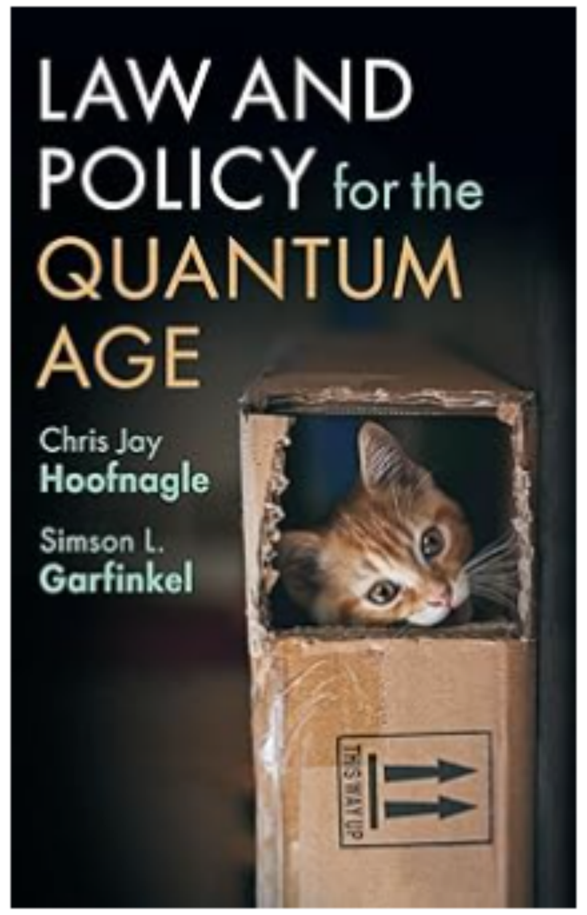
Explore more than 866 billion web pages saved over time

Enter a URL or words related to a site's home page

Usability goes everywhere — so does accessibility.

All materials should be *available* and *accessible.*

- Open source data helps!
- My book on quantum computing is Open Access (next book is also open access)



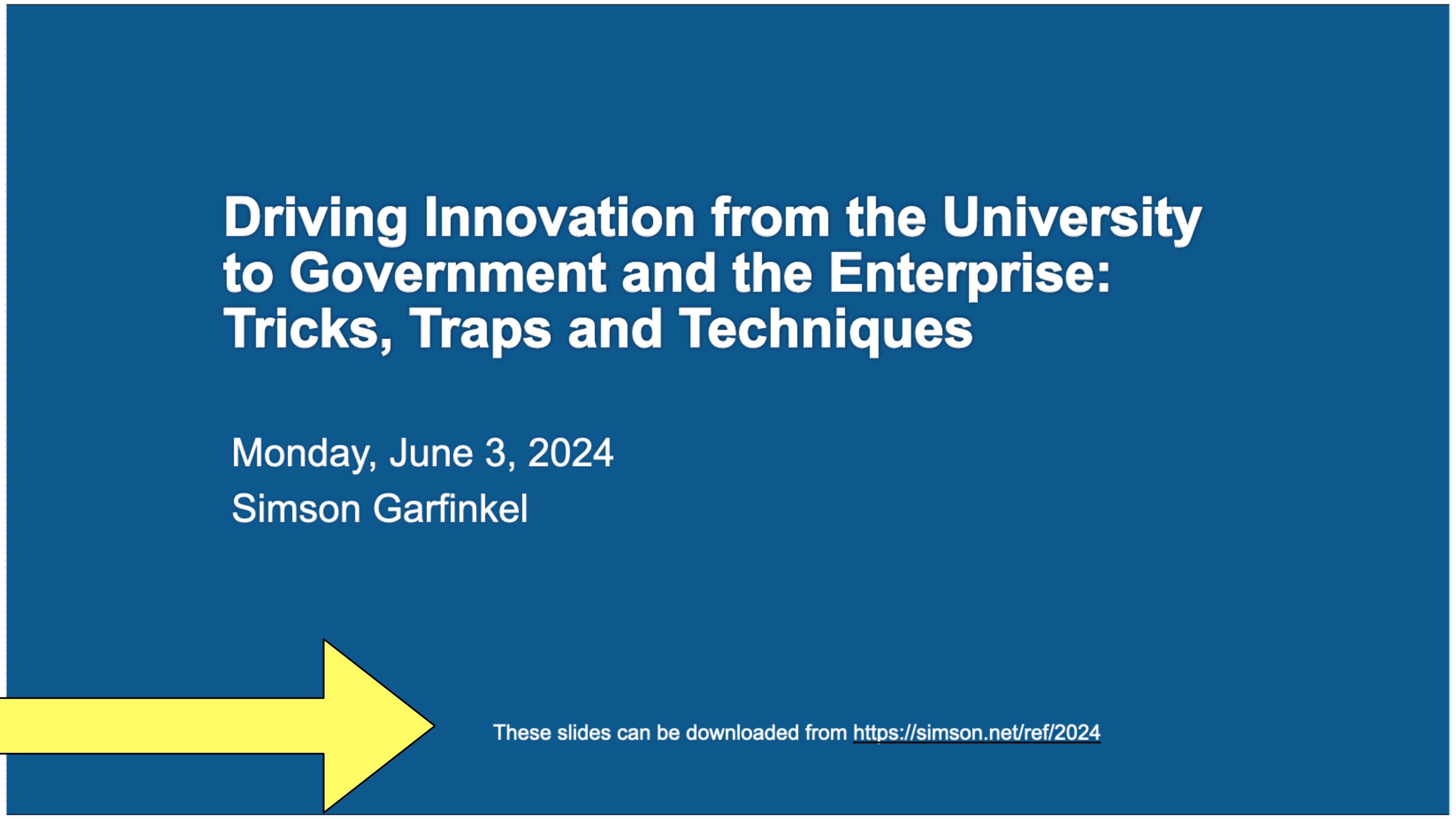
LAW AND POLICY for the QUANTUM AGE

Chris Jay Hoofnagle

Simson L. Garfinkel

Design for accessibility:

- Materials should work with screen readers.
- Do not make assumptions about the ability of to students to see, hear, walk, etc.
- Understand how everyone on a team makes their contribution.
- *Model this for students — it’s an important skill for entrepreneurs.*



Driving Innovation from the University to Government and the Enterprise: Tricks, Traps and Techniques

Monday, June 3, 2024

Simson Garfinkel

These slides can be downloaded from https://simson.net/ref/2024

#3
Design and build for maintainability

Technique — Design and Build for Maintainability

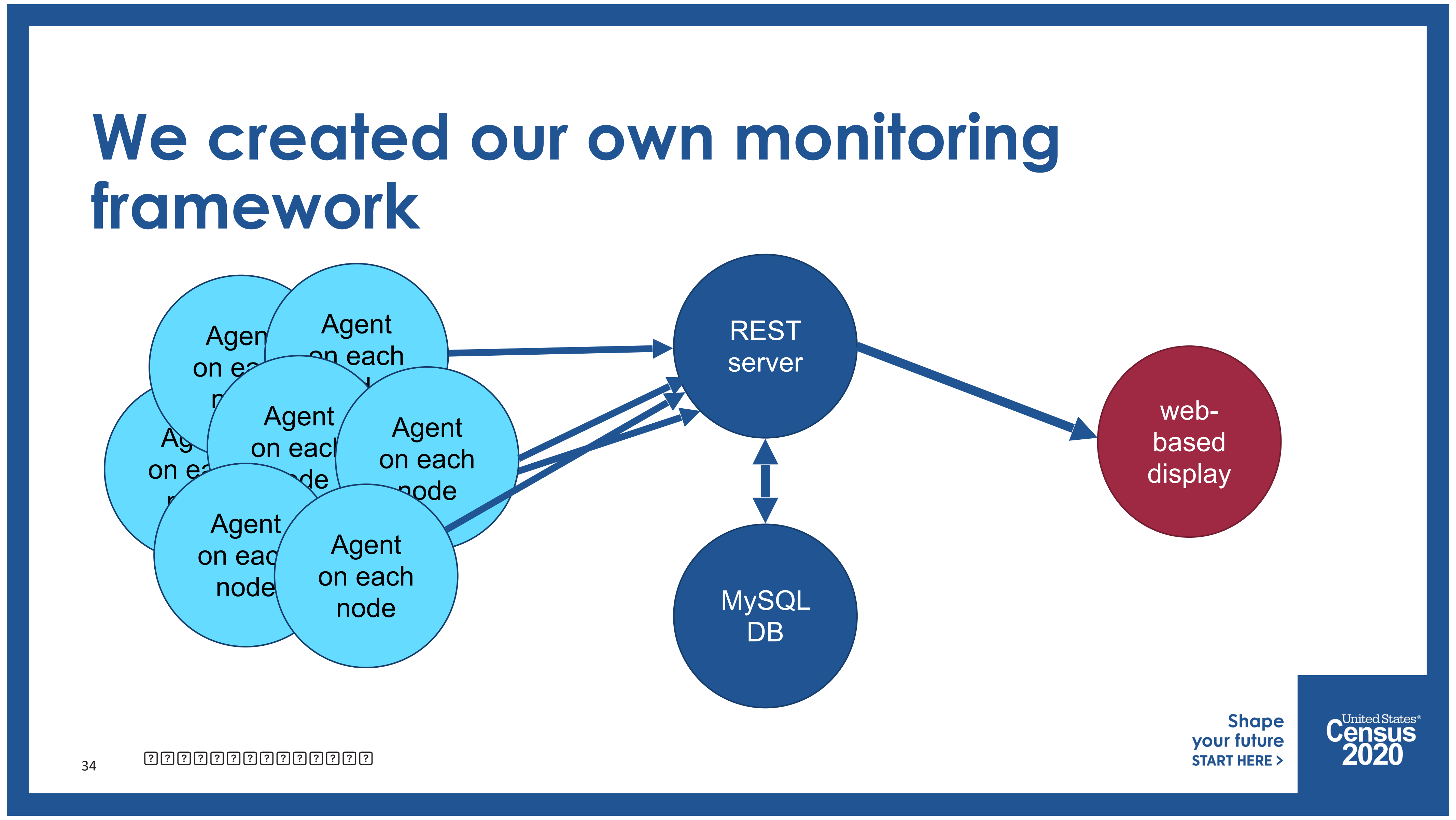
Approaches for maintainability

- Correctness
- Clean, modular design
- Documentation
- Identify code that is untested and dead.
- Attention to non-technical issues — copyright, patents, privacy practices, etc.

Innovation requires maintainability

- Technical debt — without maintainability, the cost of adding new functionality steadily rises.
- Clean documentation makes due-diligence easier and faster

Design and build for maintainability
Case Study — the 2020 Disclosure Avoidance System

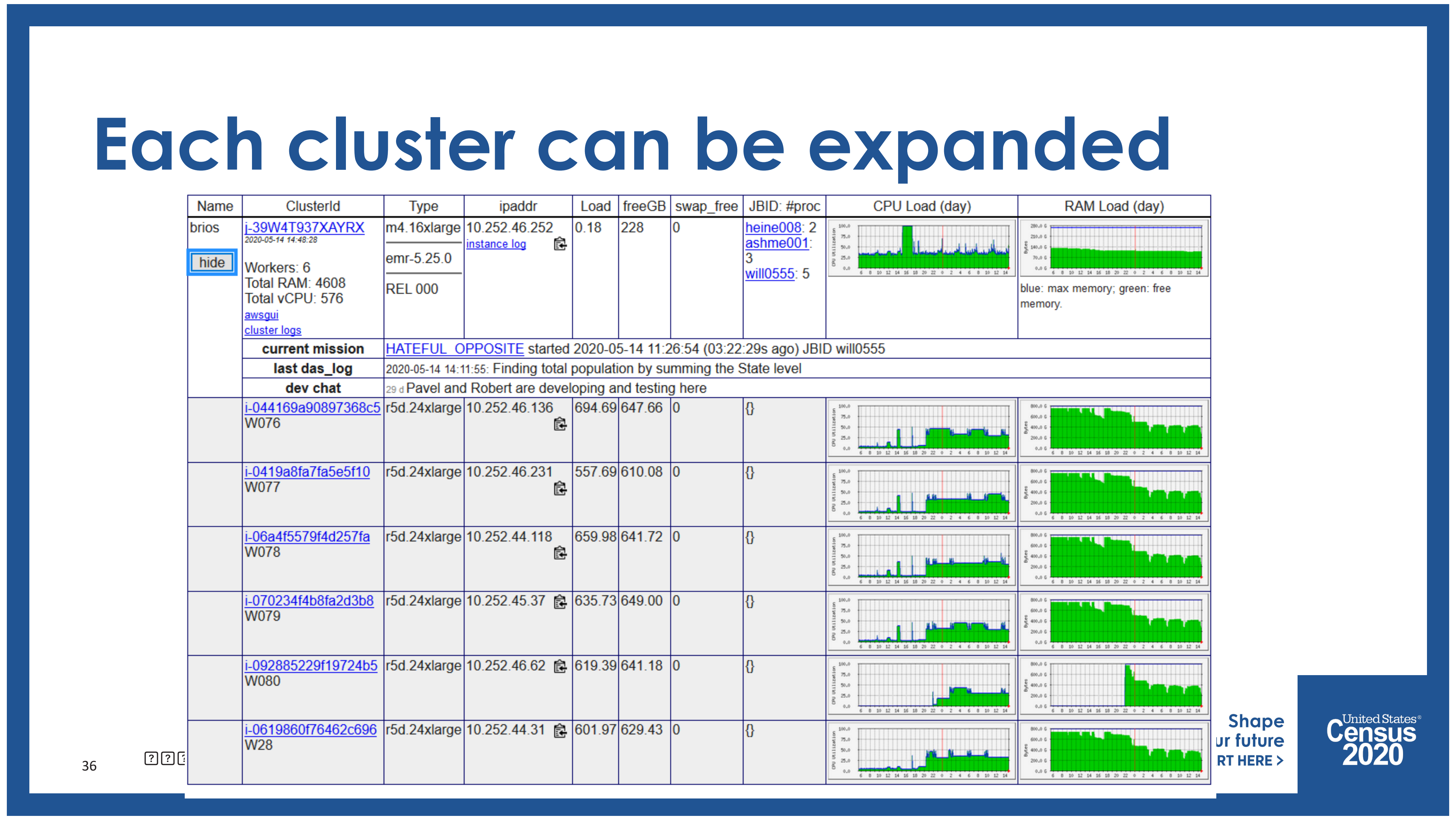


We created our own monitoring framework

34

Shape your future START HERE >

United States Census 2020



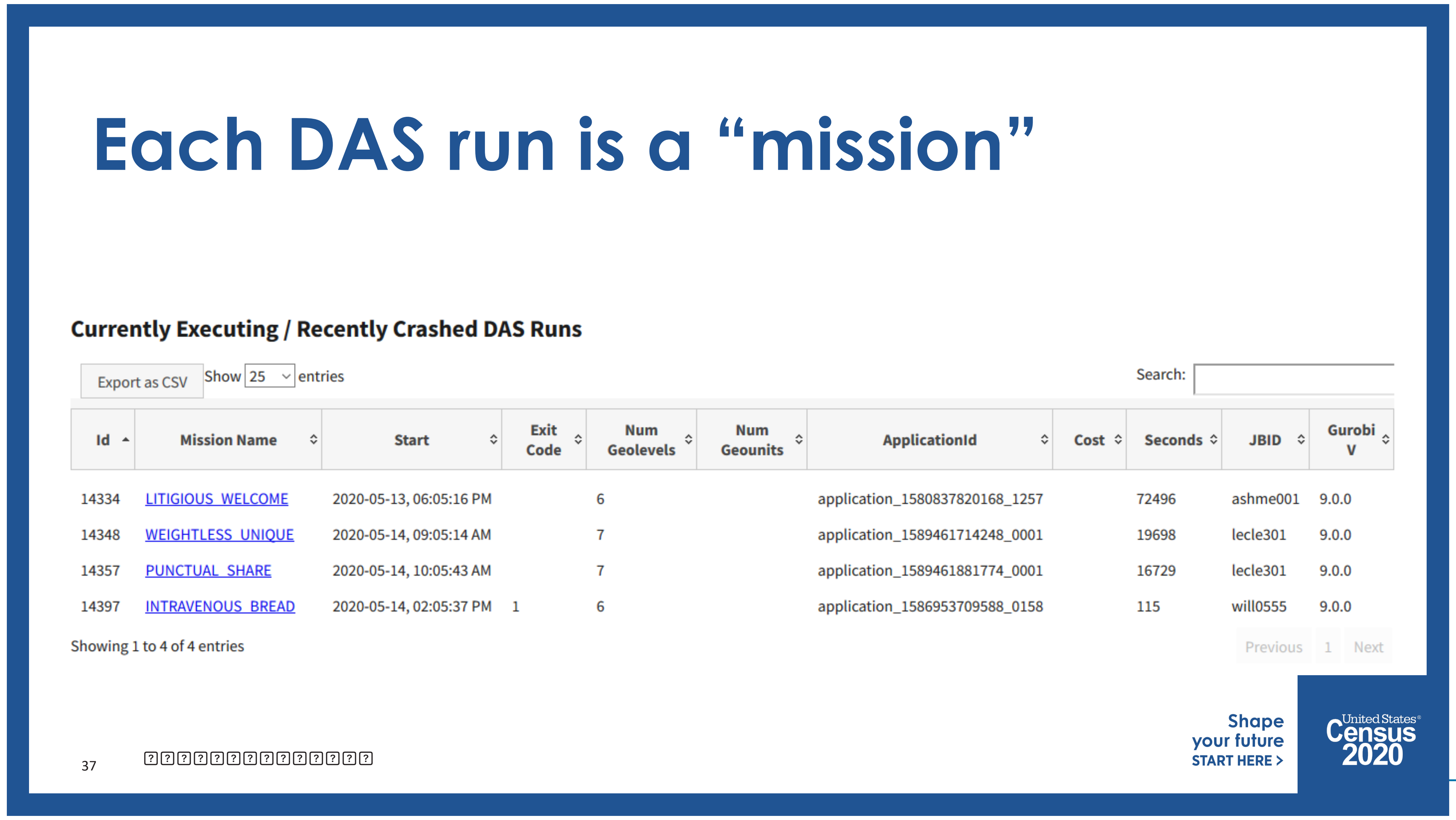
Each cluster can be expanded

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)
brios hide	i-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	{}		

36

Shape your future START HERE >

United States Census 2020



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

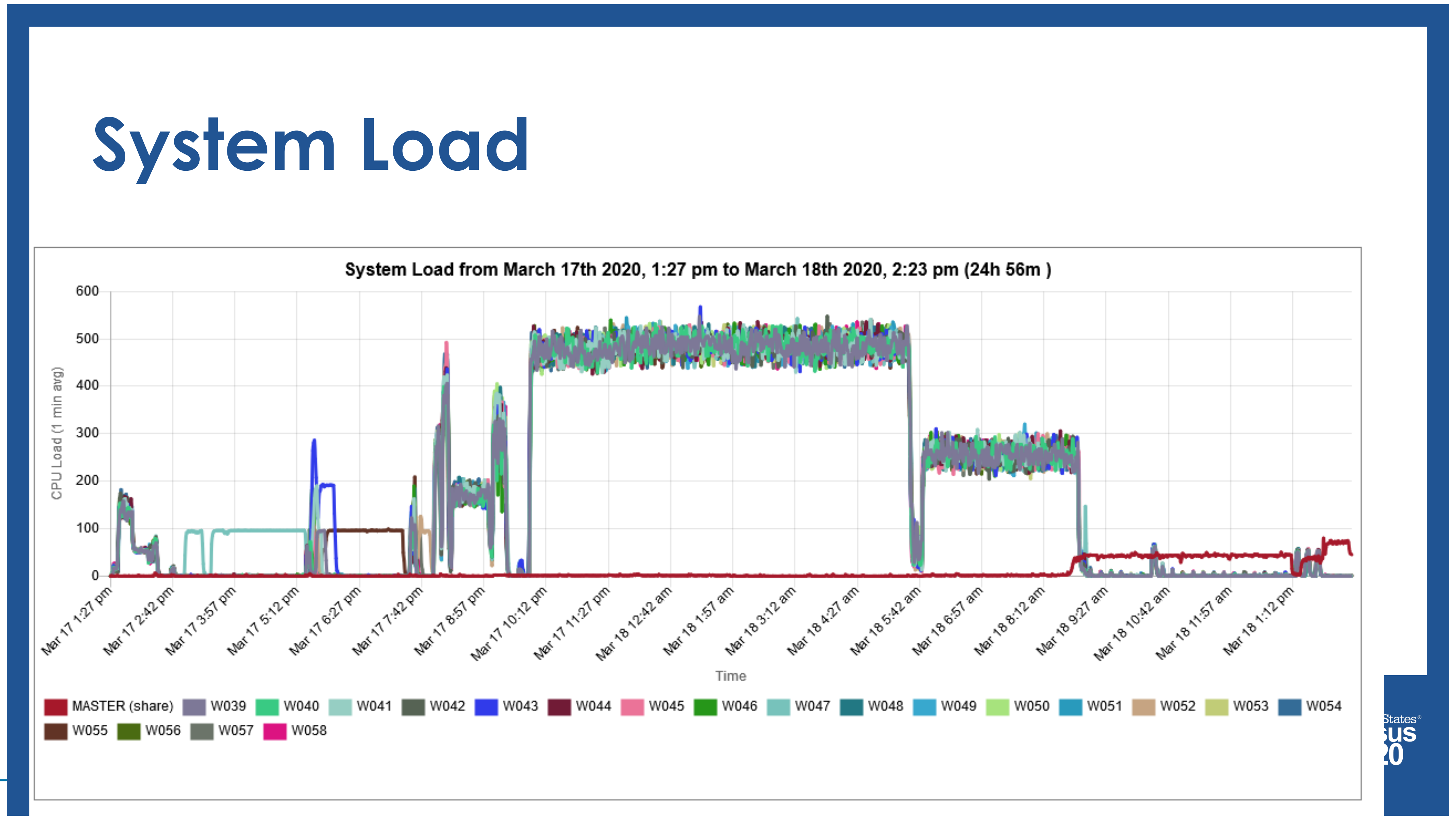
Showing 1 to 4 of 4 entries

Previous 1 Next

37

Shape your future START HERE >

United States Census 2020



System Load

System Load from March 17th 2020, 1:27 pm to March 18th 2020, 2:23 pm (24h 56m)

#4
Collaborate for Scale

Technique — Collaborate whenever you can

Students —

Faculty —

Businesses —

People in Government—

Collaboration is about the flow of *ideas,* not the flow of money or other resources.

—And don't be greedy.

#5
Teach innovation by innovating

Teaching innovation by innovating

“Advanced Computing for the Digital Humanities”

- Deploying private set intersection (PSI) to find connections between closed collections.
- Radically improving OCR for older texts.
- Large language models for ancient texts.
  - *The most innovative and transformative classes are research classes!*
  - *Digital humanities is an intelligence problem!*

“Moving ideas from the lab to the marketplace” — case studies and student projects.

- Projects with which I’ve been personally involved — digital forensics, and differential privacy and the 2020 Census.
- Companies started and run by people in my network.
- Case studies that are both historically important and relevant to SEAS (e.g. MITRE, Digital Equipment Corp., etc)

“Open Source Intelligence”

- *aka “advanced web-scraping”*
- *aka “data fusion with online information”*

- How do to it. Systematically. Bringing order to the world of online information
- Natural language processing, AI, scale, cloud processing, authentication, user interface — this course has it all!

Questions (for you)

Q1 — The installed base.

Q2 — The cost of innovation.

Question: When is it okay for an innovator to break their installed base?

From: Stuart Feldman <...@google.com>
Date: Mon, 20 Apr 2015 at 15:51
Subject: Re: make versus tabs
To: Michael Stillwell <...@google.com>

Story is only partly true.

I used tabs because I was trying to use Lex (still in first version) and had trouble with some other patterns.

(Make was written over a weekend, rewritten the next weekend ...)

So I gave up on being smart and just used a fixed pattern (^\t) to indicate rules.

Within a few weeks of writing Make, I already had a dozen friends who were using it.

So even though I knew that "tab in column 1" was a bad idea, I didn't want to disrupt my user base.

So instead I wrought havoc on tens of millions.

I have used that example in software engineering lectures.

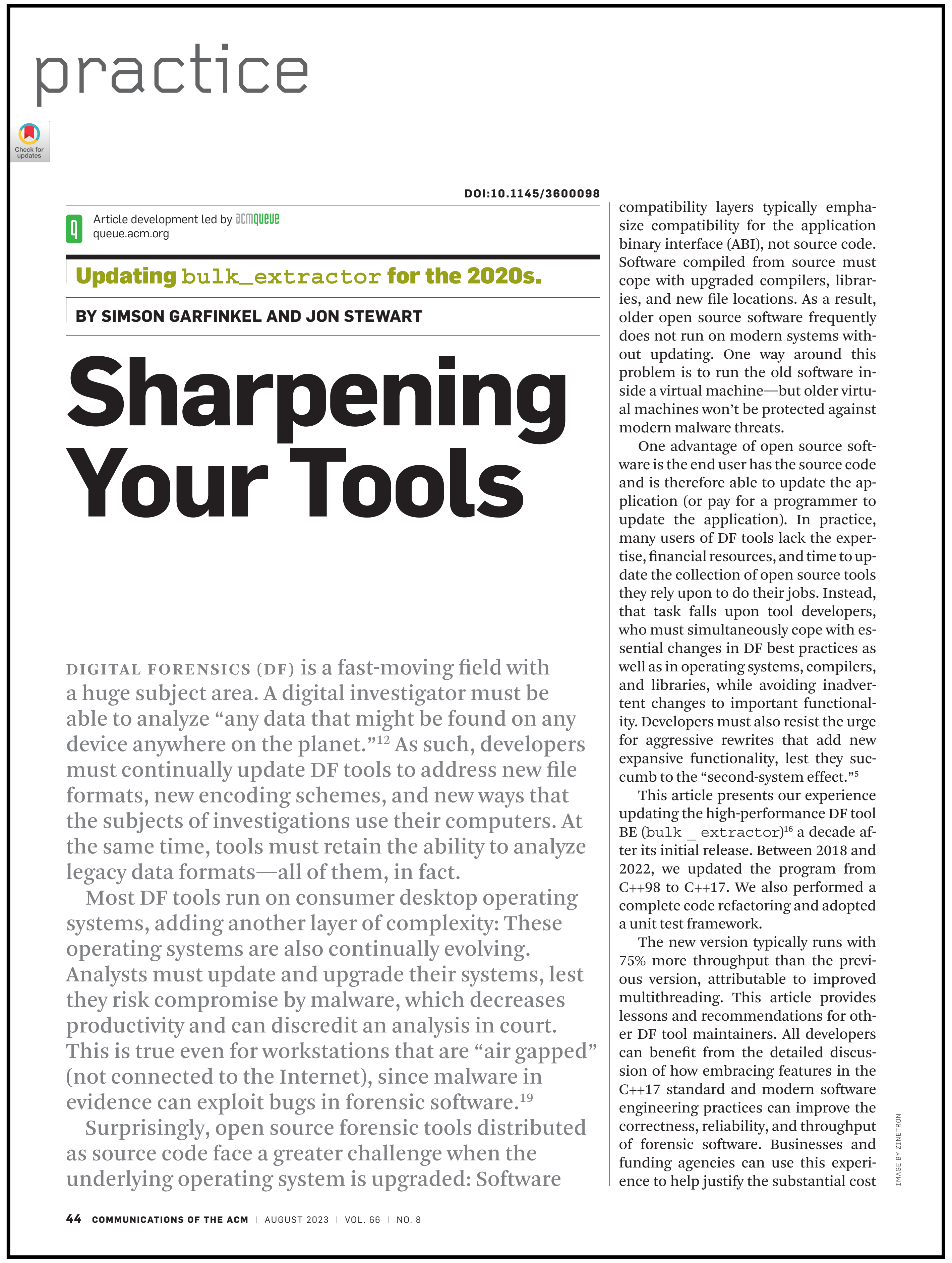
Side note: I was awarded the ACM Software Systems Award for Make a decade ago. In my one minute talk on stage, I began "I would like to apologize". The audience then split in two – half started laughing, the other half looked at the laughers.

A perfect bipartite graph of programmers and non-programmers.

https://beebo.org/haycorn/2015-04-20_tabs-and-makefiles.html

Stu Feldman, author of ‘make’

Question — Is it “innovation” when you spend 4 years updating a code base?



practice

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]

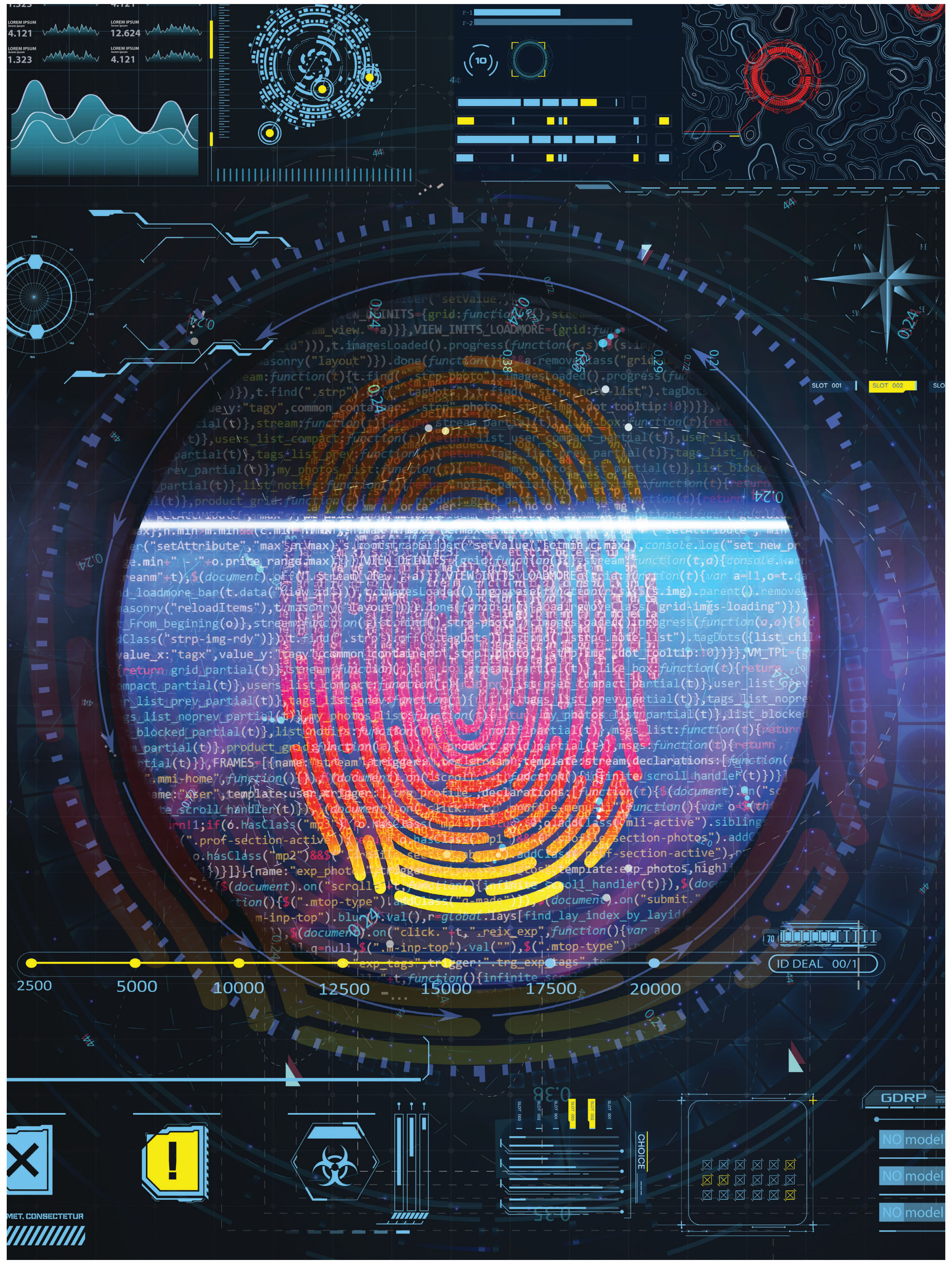
Surprisingly, open source forensic tools distributed as source code face a greater challenge when the underlying operating system is upgraded: Software compatibility layers typically emphasize compatibility for the application binary interface (ABI), not source code. Software compiled from source must 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.

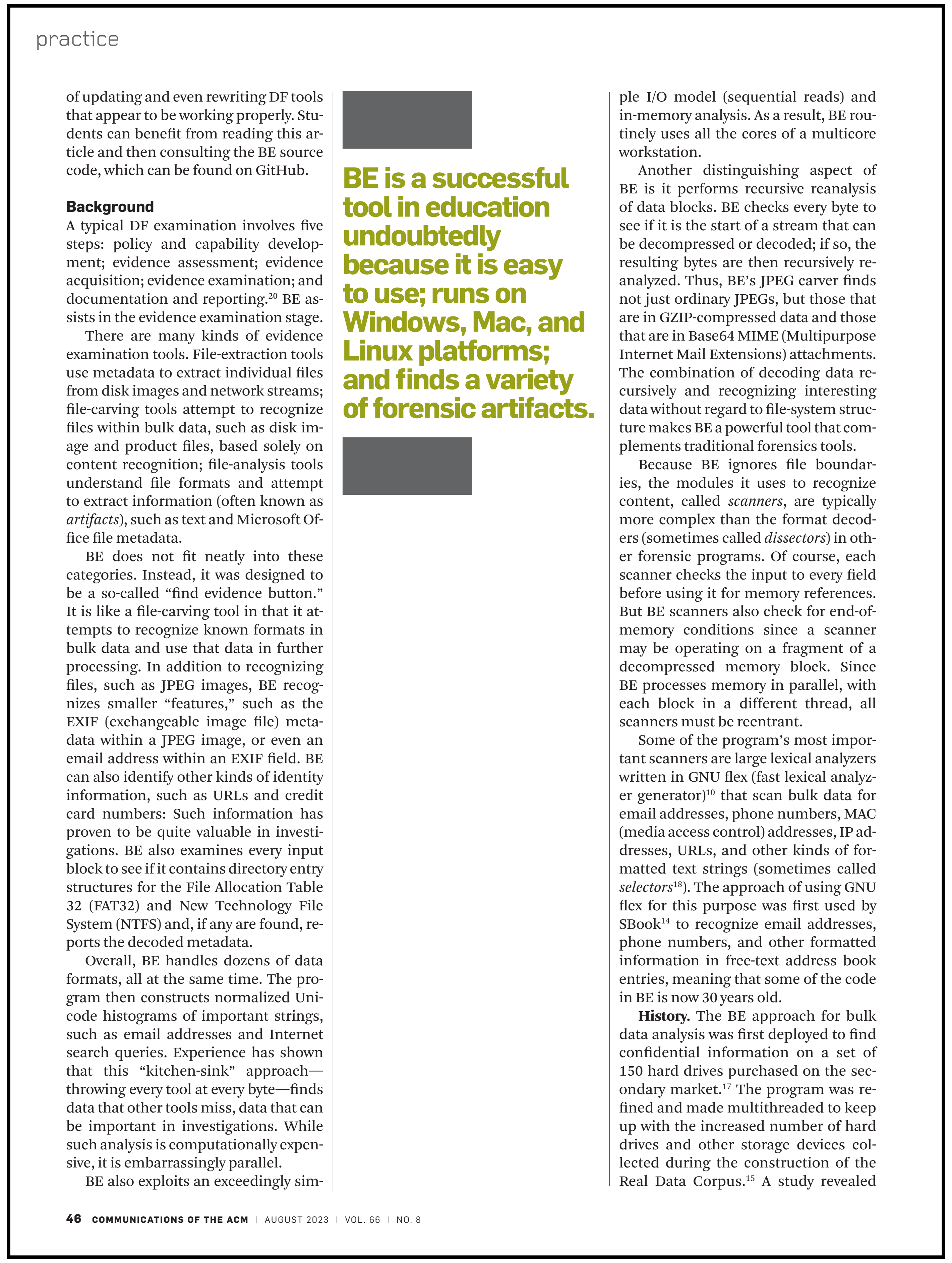
One advantage of open source software is the end user has the source code and is therefore able to update the application (or pay for a programmer to update the application). In practice, many users of DF tools lack the expertise, financial resources, and time to update the collection of open source 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





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.

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

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.

Modern code base (C++20) • extensive unit tests (from 0% to 50% code coverage)
improved threading model

This creates an infrastructure for future innovation.