Digital Forensics Past and Future

Cyberkriminalität und Forensische Informatik

(“International Summer School on Cybercrime and Forensic Computing 2022”)

Friday, June 10, 2022

https://www.cybercrime.fau.de/international-summer-school-on-cybercrime-and-forensic-computing-2022/

Simson L. Garfinkel
Senior Data Scientist
US Department of Homeland Security

https://www.dhs.gov/data/

Nürnberg Castle at night. Image courtesy of Adrian Valenzuela under CC-BY-2.0, rescaled from original

Outline for the next 3½ hours

Hour 1: Digital Forensics: The Last 10 Years

- My 2010 paper, “Digital forensics research: the next 10 years”
- What actually happened
- Discussion

—*15 minute break*

Hour 2: Transitioning Research to Practice

- The “drives” project • bulk_extractor • sector hashing • Digital Corpora
- Discussion

—*15 minute break*

Hour 3: Digital Forensics: A Future History and Research Agenda

- What does the forensics world look like in 2032?
- How do we get there?

A bit about me

Tech Journalist: 1985—2002

Entrepreneur: 1988—2002

Vineyard.NET, Broadband2Wireless,

Sandstorm Enterprises, Inc

MIT EECS 2002—2005 (PhD CS)

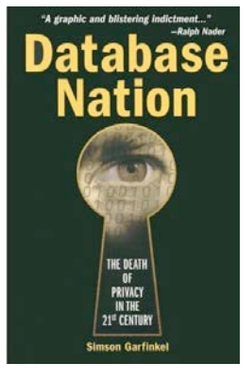
Associate Professor, 2006—2014
Naval Postgraduate School

Senior Advisor, NIST 2015-2016

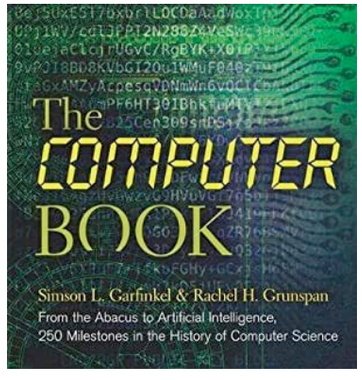
Senior Computer Scientist,
US Census Bureau 2017-2021

Senior Data Scientist,
US Department of Homeland Security, 2021-

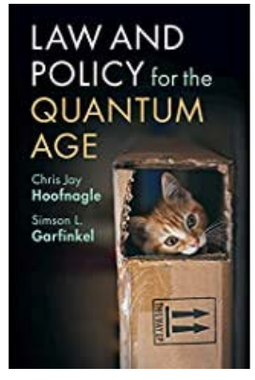
ca 2006



2000

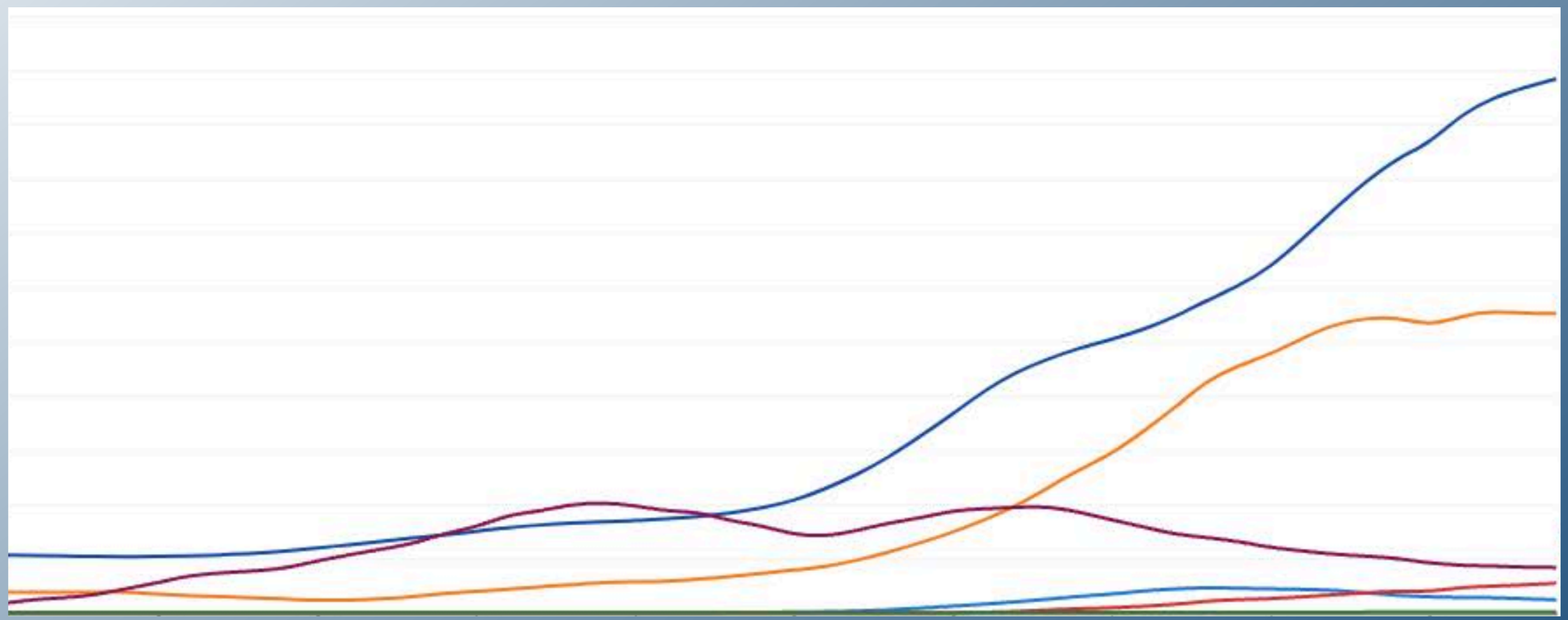


2018



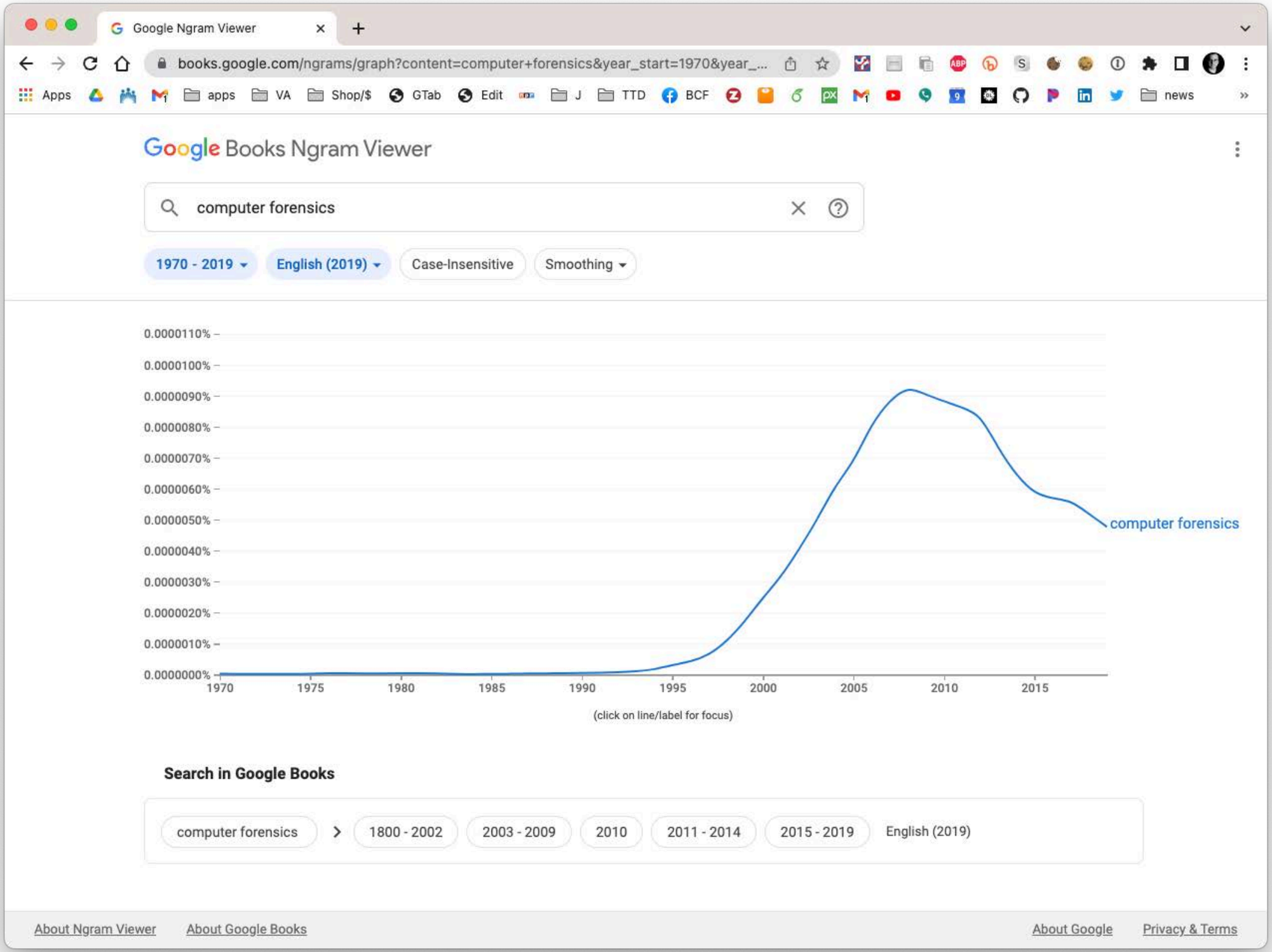
2021

"The views expressed in this presentation are those of the author and do not necessarily reflect those of the Department of Homeland Security or the US Government."

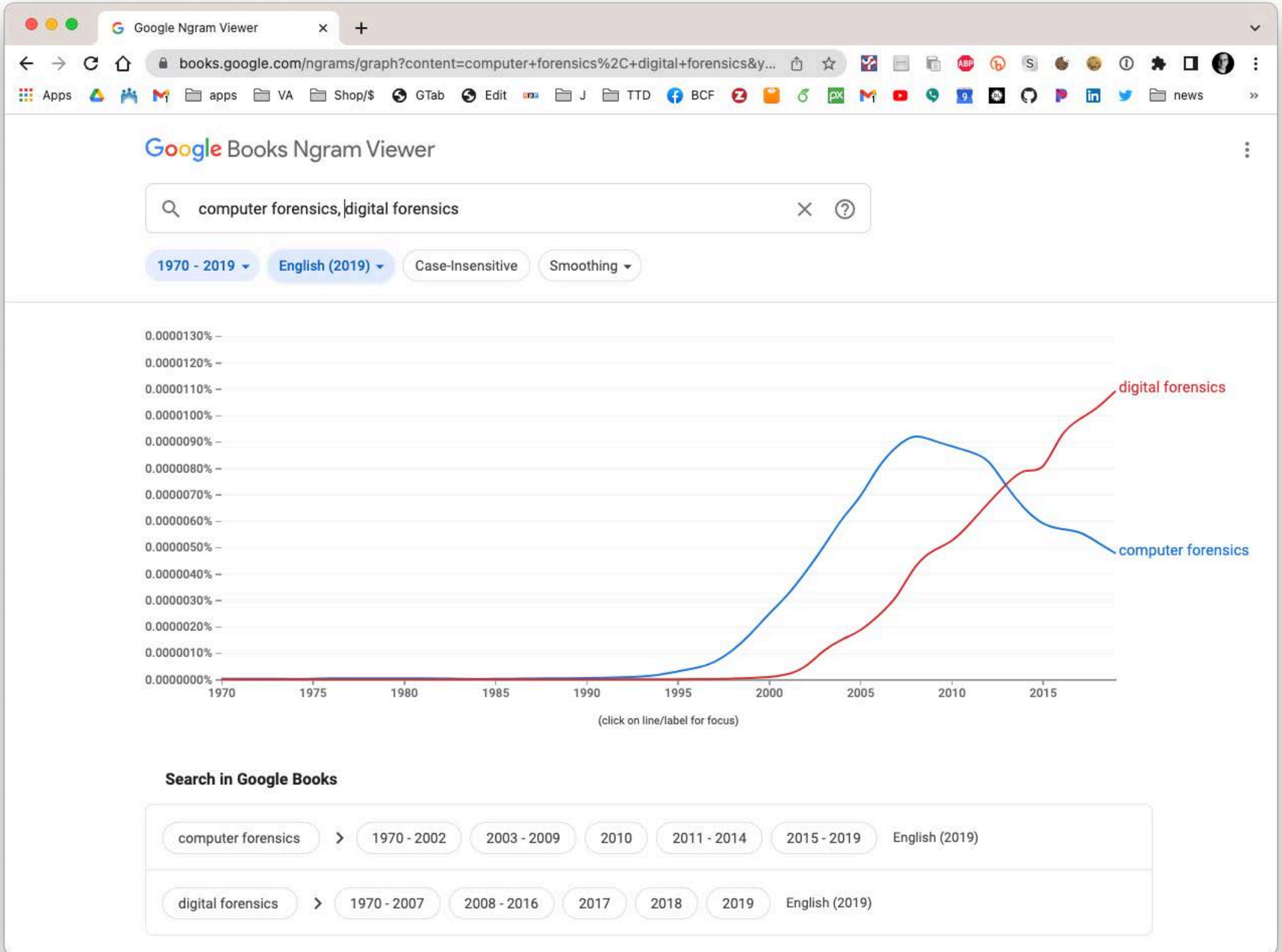


Digital Forensics: “Ancient” History

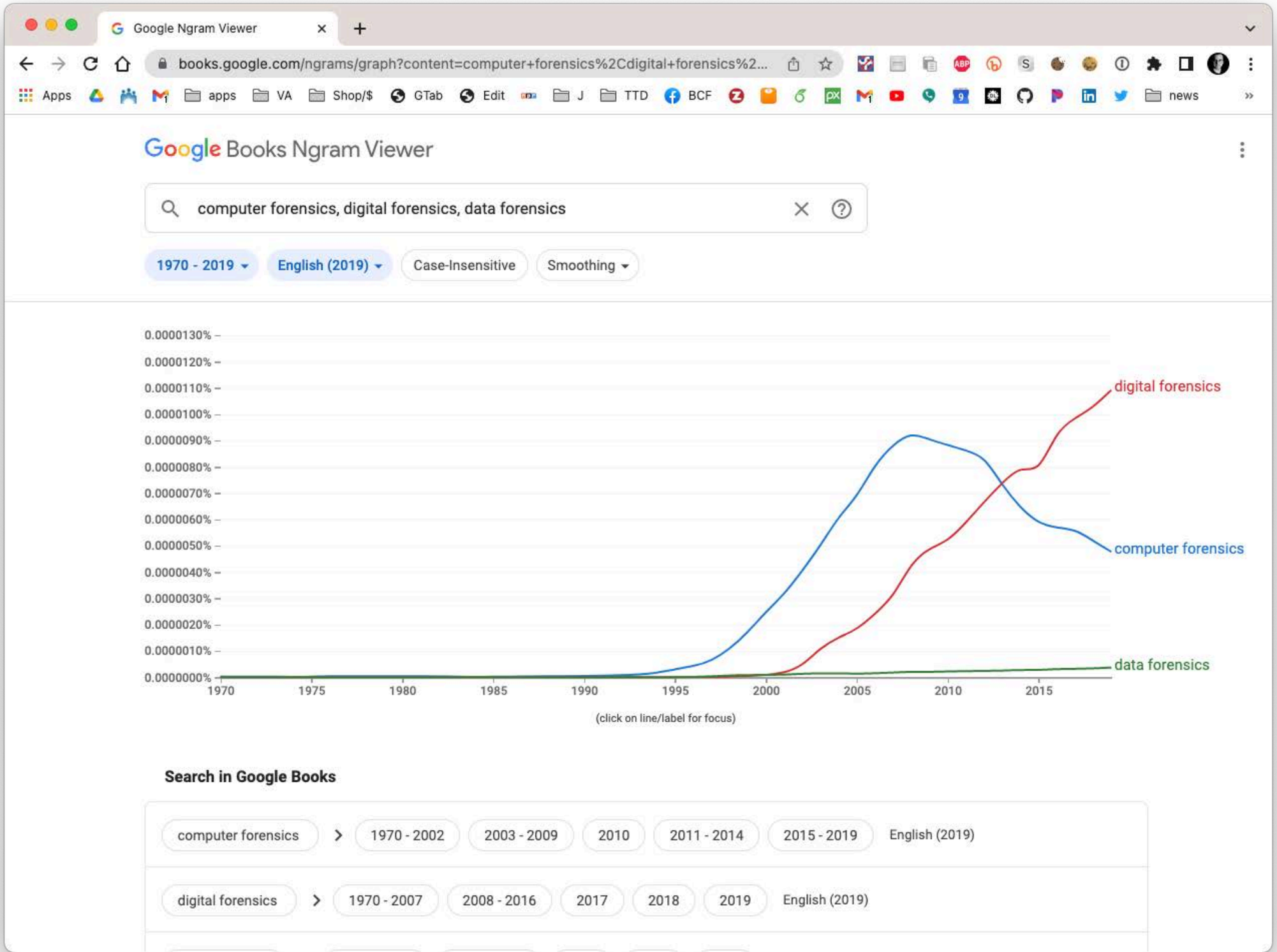
“Computer Forensics” was the original name.



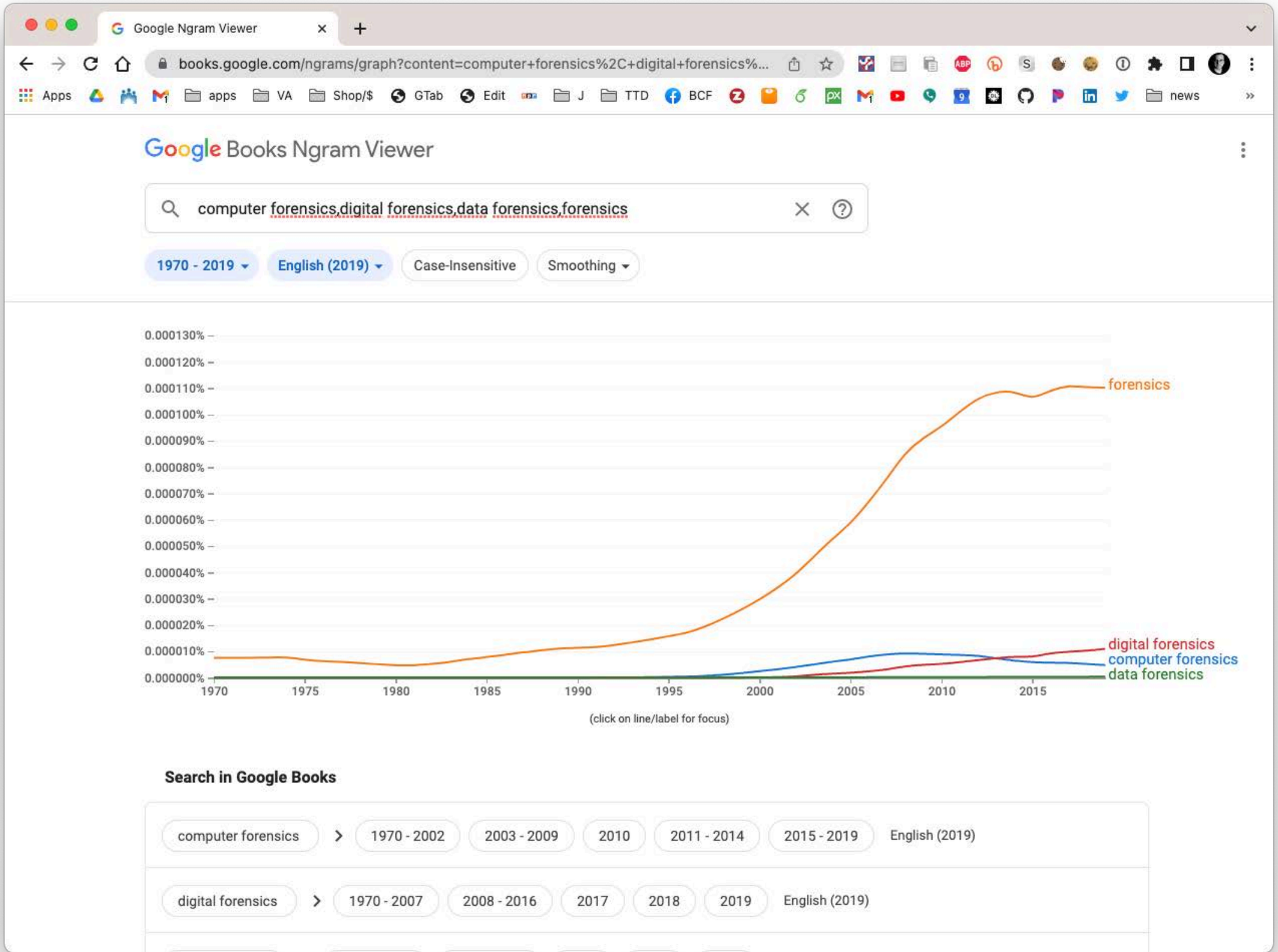
“Digital forensics” passed “computer forensics” in 2013
DF is about more than just “computers.”



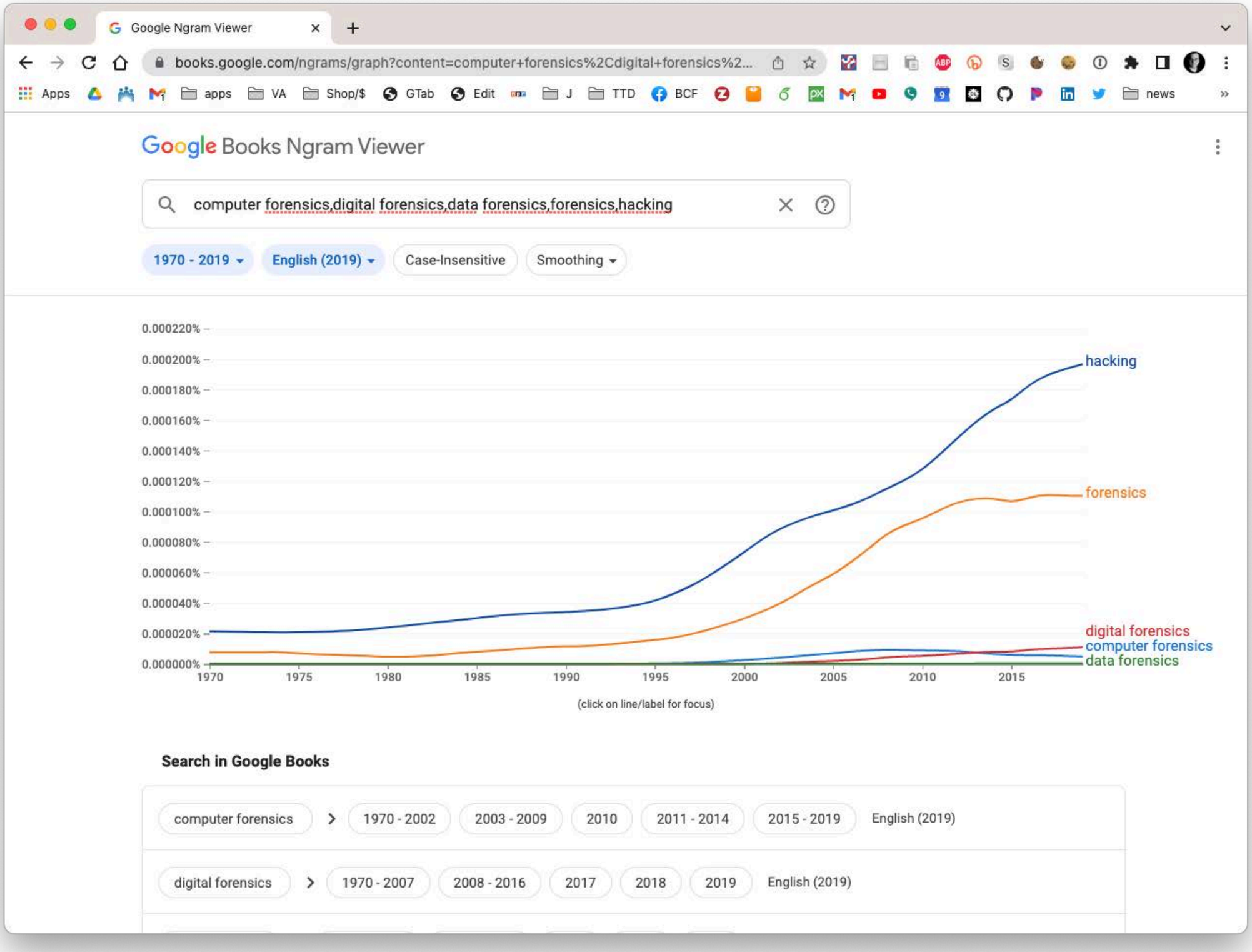
“data forensics” is up and coming. Watch for it.



“forensics” is a huge field.
We are a tiny part.



“hacking” is more popular than forensics*



*This statistic is utterly meaningless

The oldest reference to computer forensics I can find.

“Court Martial,” Star Trek 1967 (S1E20)

This has it all:

- Digital video stored in a computer
- Evidence is self-authenticating
- Used in a court of law
- Trial outcome depends on the evidence
- Evidence is disputed by the defense

Capt. Kirk is on trial for criminal negligence.

Kirk is alleged to have jettisoned a “research pod” containing Lt. Commander Benjamin Finney.

The evidence: video showing Kirk jettisoning the pod during “yellow alert.”

Finney didn’t have a warning to leave the pod!

Mr. Spock thinks that the evidence was modified.

Mr. Spock concludes that the ship's computer has been tampered with because he can now beat the computer at chess

Mr. Spock programmed the computer, and couldn't beat it before.

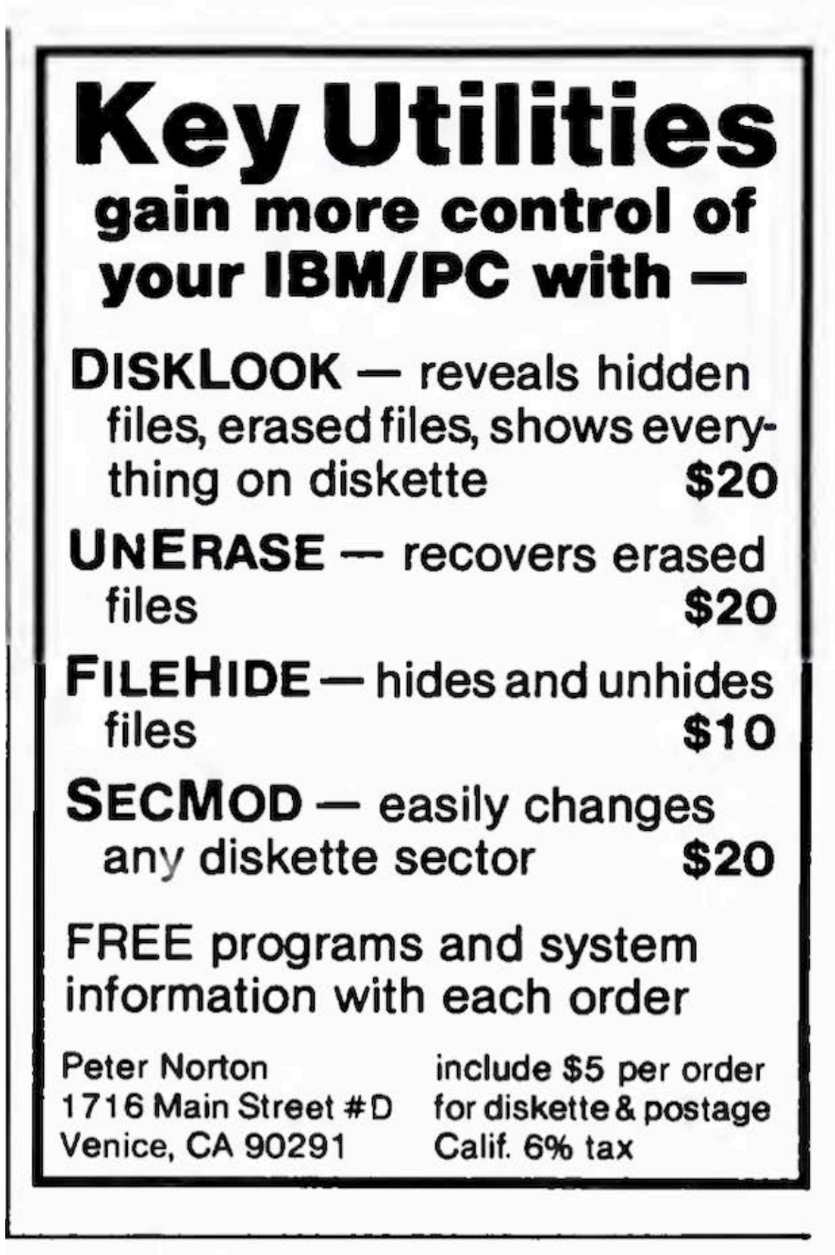
Spoiler: Finney is still alive.
He’s the one who manipulated the computer.

“Computer Forensics” in the 1980s: Data Recovery and Incident Response

1982 - Norton Utilities for DOS and Windows 3.1 with UNERASE

1984 - US Federal Bureau of Investigation launches Computer Analysis and Response Team (CART)

1985 - British Metropolitan Police sets up a computer crime department

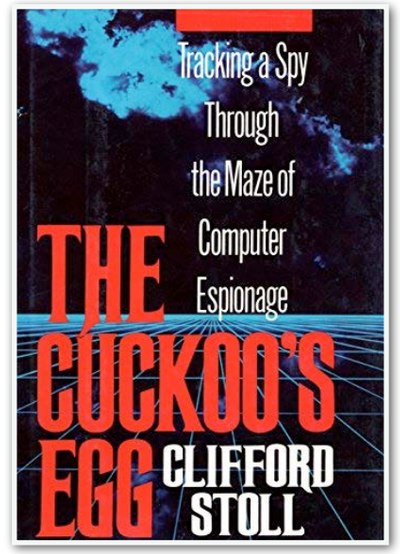


1982: First advertisement for Norton Utilities

“Computer Forensics” in the 1980s: Data Recovery and Incident Response

1986 - Cliff Still pursues Markus Hess;

- Attribution across international networks
- The Cuckoo’s Egg, Stoll (1989)



1988 - Morris Worm infects the Internet

- Reverse engineering and real-time mitigation
- “With Microscopes and Tweezers: The Worm from MIT’s Perspective,” Rochlis & Eichin, Communications of the ACM 1989*

“All the News That’s Fit to Print”

The New York Times

Late Edition
New York: Today, partly sunny, milder. High 59-64. Tonight, mostly cloudy. Low 48-54. Tomorrow, cloudy, windy, rain developing. High 57-62. Yesterday: High 56, low 41. Details, page D16.

VOL. CXXXVIII . . . No. 47,679 Copyright © 1988 The New York Times NEW YORK, FRIDAY, NOVEMBER 4, 1988 50 cents beyond 75 miles from New York City, except on Long Island. 35 CENTS

The New York Times/Jim Wilson

Gov. Michael S. Dukakis having his picture taken by a 10-year-old fan at a town meeting in Fairless Hills, Pa., during a tour of the Northeast in which he emphasized the drug problem. Page A19. Vice President Bush addressed supporters a rally in Columbus, Ohio. Less than a week after Mr. Dukakis acknowledged being a liberal, Mr. Bush said yesterday that “this election is not about labels.” Page A18.

‘Virus’ in Military Computers Disrupts Systems Nationwide

By JOHN MARKOFF

In an intrusion that raises questions about the vulnerability of the nation's computers, a Department of Defense network has been disrupted since Wednesday by a rapidly spreading “virus” program apparently introduced by a computer science student.

The program reproduced itself through the computer network, making hundreds of copies in each machine it reached, effectively clogging systems linking thousands of military, corporate and university computers around the nation and preventing them from doing additional work. The virus is thought not to have destroyed any files.

By late yesterday afternoon computer experts were calling the virus the largest assault ever on the nation's computers.

‘The Big Issue’

“The big issue is that a relatively benign software program

military officials, researchers and corporations.

While some sensitive military data are involved, the computers handling the nation's most sensitive secret information, like that on the control of nuclear weapons, are thought not to have been touched by the virus.

Parallel to Biological Virus

Computer viruses are so named because they parallel in the computer world the behavior of biological viruses. A virus is a program, or a set of instructions to a computer, that is either planted on a floppy disk meant to be used with the computer or introduced when the computer is communicating over telephone lines or data networks with other computers.

The programs can copy themselves into the computer's master software, or operating system,

PENTAGON REPORTS IMPROPER CHARGES FOR CONSULTANTS

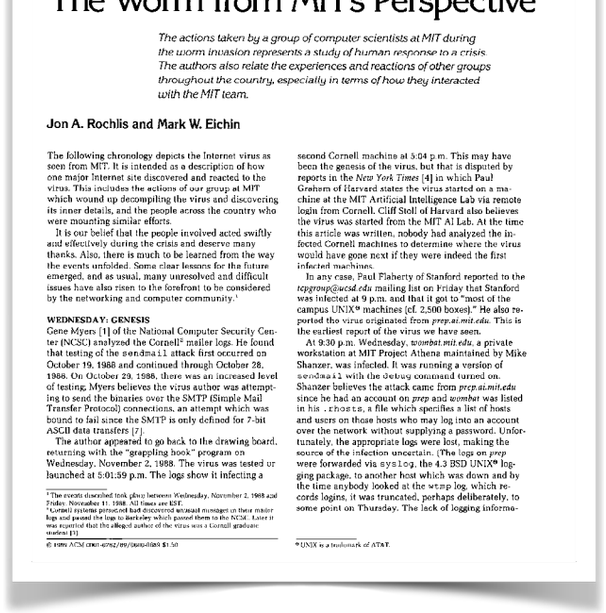
CONTRACTORS CRITICIZED

Inquiry Shows Routine Billing of Government by Industry on Fees, Some Dubious

By JOHN H. CUSHMAN Jr.
Special to The New York Times

WASHINGTON, Nov. 3 — A Pentagon investigation has found that the nation's largest military contractors routinely charge the Defense Department for hundreds of millions of dollars paid to consultants, often without justification.

The report of the investigation said that neither the military's current rules nor the contractors' own policies are adequate to assure that the Gov-



THE INTERNET WORM

With Microscope and Tweezers: The Worm from MIT’s Perspective

The actions taken by a group of computer scientists at MIT during the worm invasion represents a study of human response to a crisis. The authors also relate the experiences and reactions of other groups throughout the country, especially in terms of how they interacted with the MIT team.

Jon A. Rochlis and Mark W. Eichin

The following chronology depicts the Internet virus as seen from MIT. It is intended as a description of how one major Internet site discovered and reacted to the virus. This includes the actions of our group at MIT which wound up decompiling the virus and discovering its inner details, and the people across the country who were mounting similar efforts.

It is our belief that the people involved acted swiftly and effectively during the crisis and deserve many thanks. Also, there is much to be learned from the way the events unfolded. Some clear lessons for the future emerged, and as usual, many unresolved and difficult issues have also risen to the forefront to be considered by the networking and computer community.[1]

WEDNESDAY: GENESIS

Gene Myers [1] of the National Computer Security Center (NCSC) analyzed the Cornell[2] mailer logs. He found that testing of the `sendmail` attack first occurred on October 19, 1988 and continued through October 28, 1988. On October 29, 1988, there was an increased level of testing. Myers believes the virus author was attempting to send the binaries over the SMTP (Simple Mail Transfer Protocol) connections, an attempt which was bound to fail since the SMTP is only defined for 7-bit ASCII data transfers [2].

The author appeared to go back to the drawing board, returning with the “grappling hook” program on Wednesday, November 2, 1988. The virus was tested or launched at 5:01:39 p.m. The logs show it infecting a second Cornell machine at 5:04 p.m. This may have been the genesis of the virus, but that is disputed by reports in the *New York Times* [4] in which Paul Graham of Harvard states the virus started on a machine at the MIT Artificial Intelligence Lab via remote login from Cornell. Cliff Stoll of Harvard also believes the virus was started from the MIT AI Lab. At the time this article was written, nobody had analyzed the infected Cornell machines to determine where the virus would have gone next if they were indeed the first infected machines.

In any case, Paul Flaherty of Stanford reported to the *tcpgroup@ucsd.edu* mailing list on Friday that Stanford was infected at 9 p.m. and that it got to “most of the campus UNIX® machines (cf. 2,500 boxes).” He also reported the virus originated from *prep.ai.mit.edu*. This is the earliest report of the virus we have seen.

At 9:30 p.m. Wednesday, *wombat.mit.edu*, a private workstation at MIT Project Athena maintained by Mike Shanzer, was infected. It was running a version of `sendmail` with the `debug` command turned on. Shanzer believes the attack came from *prep.ai.mit.edu* since he had an account on *prep* and *wombat* was listed in his `.rhosts`, a file which specifies a list of hosts and users on those hosts who may log into an account over the network without supplying a password. Unfortunately, the appropriate logs were lost, making the source of the infection uncertain. (The logs on *prep* were forwarded via `syslog`, the 4.3 BSD UNIX® logging package, to another host which was down and by the time anybody looked at the `wtmp` log, which records logins, it was truncated, perhaps deliberately, to some point on Thursday. The lack of logging informa-

[1] The events described took place between Wednesday, November 2, 1988 and Friday, November 11, 1988. All times are EST.
[2] Cornell systems personnel had discovered sendmail messages in their mailer logs and passed the logs to Berkeley which passed them to the NCSC. Later it was reported that the alleged author of the virus was a Cornell graduate student [3].

® UNIX is a trademark of AT&T.

* https://dl.acm.org/doi/abs/10.1145/63526.63528

DIGITAL INVESTIGATION 7 (2010) S64–S73

available at www.sciencedirect.com

ScienceDirect

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

Digital Investigation

Digital forensics research: The next 10 years

Simson L. Garfinkel

Naval Postgraduate School, Monterey, USA

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

ABSTRACT

Today's Golden Age of computer forensics is quickly coming to an end. Without a clear strategy for enabling research efforts that build upon one another, forensic research will fall behind the market, tools will become increasingly obsolete, and law enforcement, military and other users of computer forensics products will be unable to rely on the results of forensic analysis. This article summarizes current forensic research directions and argues that to move forward the community needs to adopt standardized, modular approaches for data representation and forensic processing.

1. Introduction

Digital Forensics (DF) has grown from a relatively obscure tradecraft to an important part of many investigations. DF tools are now used on a daily basis by examiners and analysts within local, state and Federal law enforcement; within the military and other US government organizations; and within the private "e-Discovery" industry. Developments in forensic research, tools, and process over the past decade have been very successful and many in leadership positions now rely on these tools on a regular basis—frequently without realizing it. Moreover, there seems to be a widespread belief, buttressed on by portrayals in the popular media, that advanced tools and skillful practitioners can extract actionable information from practically any device that a government, private agency, or even a skillful individual might encounter.

This paper argues that we have been in a "Golden Age of Digital Forensics," and that the Golden Age is quickly coming to an end. Increasingly organizations encounter data that cannot be analyzed with today's tools because of format incompatibilities, encryption, or simply a lack of training. Even data that can be analyzed can wait weeks or months before review because of data management issues. Without a clear research agenda aimed at dramatically improving the efficiency of both our tools and our very research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

This paper proposes a plan for achieving that dramatic improvement in research and operational efficiency through the adoption of systematic approaches for representing forensic data and performing forensic computation. It draws on more than 15 years personal experience in computer forensics, an extensive review of the DF research literature, and dozens of discussions with practitioners in government, industry, and the international forensics community.

1.1. *Prior and related work*

Although there has been some work in the DF community to create common file formats, schemas and ontologies, there has been little actual standardization. DFRWS started the Common Digital Evidence Storage Format (CDESF) Working Group in 2006. The group created a survey of disk image storage formats in September 2006, but then disbanded in August 2007 "because DFRWS did not have the resources required to achieve the goals of the group. (CDESF working group, 2009)" Hoss and Carver discuss ontologies to support digital forensics (Carver and Hoss, 2009), but did not propose any concrete ontologies that can be used. Garfinkel introduced an XML representation for file system metadata (Garfinkel, 2009), but it has not been widely adopted.

E-mail address: simsong@acm.org

1742-2876/$ – see front matter © 2010 Digital Forensic Research Workshop. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2010.05.009

2010 looking to today…

“Digital forensics research: the next 10 years”

DFRWS 2010

1. Introduction
2. Digital forensics: a brief history
3. Today’s research challenges
4. A new research direction

DFRWS was originally the “Digital Forensics Research Workshop”

- It stopped being a “workshop” in 2006 to make it a more attractive publication venue for academics

DIGITAL INVESTIGATION 7 (2010) S64–S73

available at www.sciencedirect.com

ScienceDirect

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

ELSEVIER

Digital Investigation

Digital forensics research: The next 10 years

Simson L. Garfinkel

Naval Postgraduate School, Monterey, USA

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

ABSTRACT

Today's Golden Age of computer forensics is quickly coming to an end. Without a clear strategy for enabling research efforts that build upon one another, forensic research will fall behind the market, tools will become increasingly obsolete, and law enforcement, military and other users of computer forensics products will be unable to rely on the results of forensic analysis. This article summarizes current forensic research directions and argues that to move forward the community needs to adopt standardized, modular approaches for data representation and forensic processing.

1. Introduction

Digital Forensics (DF) has grown from a relatively obscure tradecraft to an important part of many investigations. DF tools are now used on a daily basis by examiners and analysts within local, state and Federal law enforcement; within the military and other US government organizations; and within the private “e-Discovery” industry. Developments in forensic research, tools, and process over the past decade have been very successful and many in leadership positions now rely on these tools on a regular basis—frequently without realizing it. Moreover, there seems to be a widespread belief, buttressed on by portrayals in the popular media, that advanced tools and skillful practitioners can extract actionable information from practically any device that a government, private agency, or even a skillful individual might encounter.

This paper argues that we have been in a “Golden Age of Digital Forensics,” and that the Golden Age is quickly coming to an end. Increasingly organizations encounter data that cannot be analyzed with today’s tools because of format incompatibilities, encryption, or simply a lack of training. Even data that can be analyzed can wait weeks or months before review because of data management issues. Without a clear research agenda aimed at dramatically improving the efficiency of both our tools and our very research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

This paper proposes a plan for achieving that dramatic improvement in research and operational efficiency through the adoption of systematic approaches for representing forensic data and performing forensic computation. It draws on more than 15 years personal experience in computer forensics, an extensive review of the DF research literature, and dozens of discussions with practitioners in government, industry, and the international forensics community.

1.1. Prior and related work

Although there has been some work in the DF community to create common file formats, schemas and ontologies, there has been little actual standardization. DFRWS started the Common Digital Evidence Storage Format (CDESF) Working Group in 2006. The group created a survey of disk image storage formats in September 2006, but then disbanded in August 2007 “because DFRWS did not have the resources required to achieve the goals of the group. (CDESF working group, 2009)” Hoss and Carver discuss ontologies to support digital forensics (Carver and Hoss, 2009), but did not propose any concrete ontologies that can be used. Garfinkel introduced an XML representation for file system metadata (Garfinkel, 2009), but it has not been widely adopted.

E-mail address: simsong@acm.org
1742-2876/$ – see front matter
doi:10.1016/j.diin.2010.05.009

https://www.sciencedirect.com/science/article/pii/S1742287610000368
Citations: 327 (Elsevier) • 58 (ACM) • 833 (Google Scholar)

Time travel back to 2010

Time travel back to 2010

Digital Forensics: A Brief History

Digital Forensics — A Brief History

Digital Forensics is roughly 40 years old

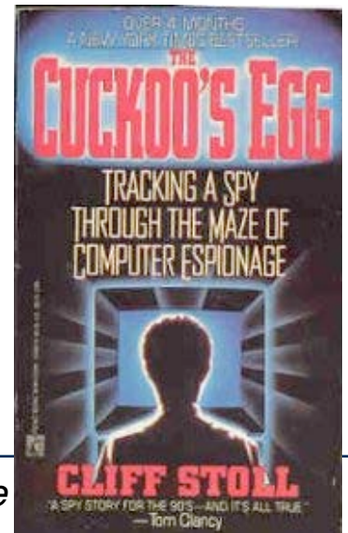
- 1970s - Data recovery and *limited* incident response
- Late 1980s — Norton & Mace Utilities provided "Unformat, Undelete."

Early days were marked by:

- Diversity — Hardware, Software & Application
- Proliferation of file formats
- Heavy reliance on time-sharing and centralized computing
- Absence of formal process, tools & training

Forensics of end-user systems was hard, but it didn't matter much

- Most of the data was stored on centralized computers
- Experts were available to assist with investigations
- There wasn't much demand!



The Golden Age of Digital Forensics: 1999—2007

Widespread use of Microsoft Windows, especially Windows XP

Relatively few file formats:

- Microsoft Office (.doc, .xls & .ppt)
- JPEG for images
- AVI and WMV for video

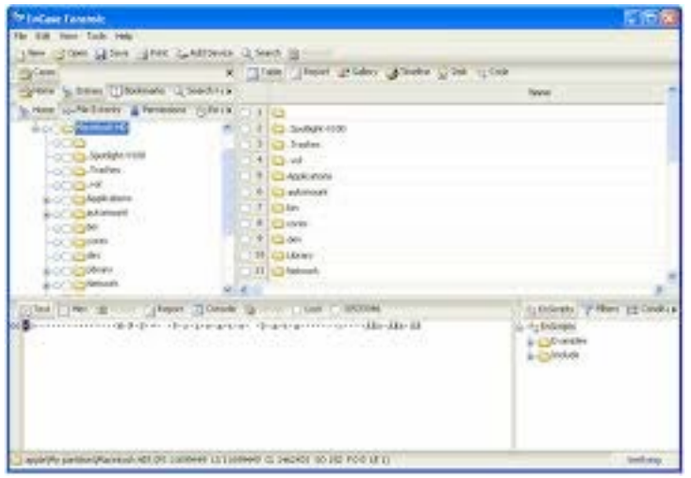
Most examinations confined to a single computer belonging to a single subject

Most storage devices used a standard interface

- IDE/ATA
- USB

This Golden Age gave us good tools and rapid growth.

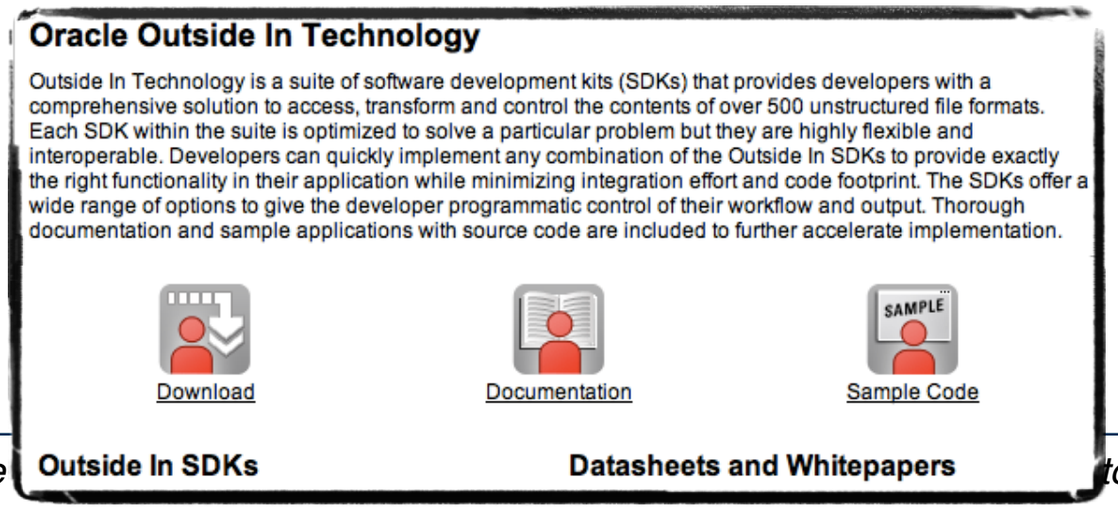
Commercial tools:



EnCase®

Open Source Tools:

Content Extraction Toolkits:



The Golden Age was aided by target conditions.

Widespread market failure of Data At Rest (DAR) Encryption

- PGPdisk — not widely deployed
- Microsoft's EFS — hard to use
- Apple's File Vault — buggy until MacOS 10.4 / 10.5

Anti-Forensics Tools

- Largely academic curiosities

Rapid Growth of Research & Professionalization

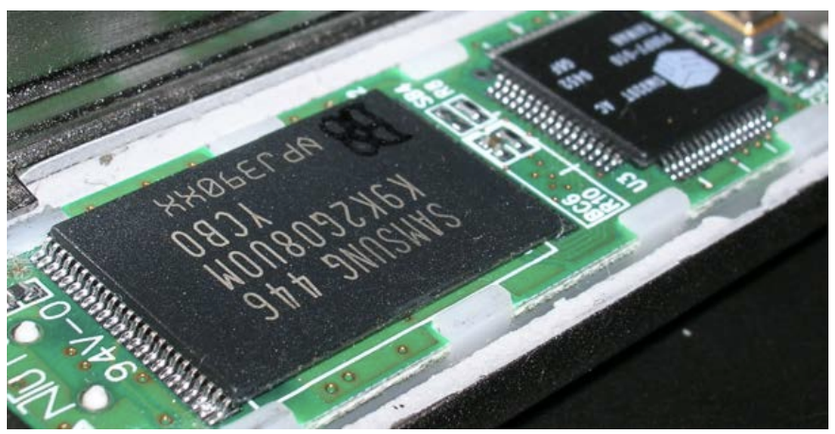
- DFRWS, IFIP WG 11.9
- Consulting firms
- 14 certificate programs
- 5 associates programs
- 16 bachelor programs
- 2 doctoral programs

DFRWS

Get ready for the coming digital forensics crisis.
1 - Dramatically increased costs of extraction & analysis.

Much of the last decade's progress is quickly becoming irrelevant

- Increased size of storage systems
- Non-Removable Flash



Shopping results for **2tb drive**

WD Elements Desktop **2 TB** External hard
(421)
$110 new
80 stores

Seagate Barracuda LP **2 TB** Internal
(101)
$105 new
165 stores

WD Caviar Green **2 TB** Internal hard
(58)
$99 new
117 stores

Samsung SpinPoint F3EG Desktop
(8)
$108 new
44 stores

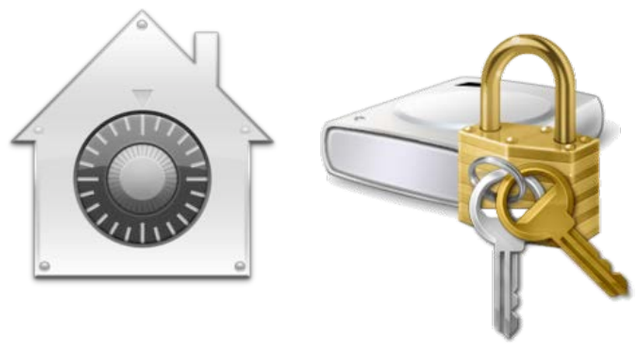
WD Caviar Black **2 TB** Internal hard
(404)
$169 new
125 stores

- Proliferation of operating systems, file formats and connectors
  - —*JFFS2, YAFFS2, Symbian, Pre, iOS,*
  - —*Most evident in mobile computing*
- Cases now require analyzing multiple devices
  - —*Typical — 2 desktops, 6 phones, 4 iPods, 2 digital cameras*
  - —*How many storage devices did you bring to this conference?*

The Coming Digital Forensics Crisis: Part 2 — Encryption and Cloud Computing

Pervasive Encryption — Encryption is increasingly present

- TrueCrypt
- BitLocker
- File Vault
- DRM Technology



Cloud Computing — End-user systems won't have the data

- Google Apps
- Microsoft Office 2010
- Apple Mobile Me

RAM-based malware

Legal challenges (e.g. US vs. Comprehensive Drug Testing).

The Coming Digital Forensics Crisis. Part 3 — Mobile Phones

Forensic examiners established bit-copies as the gold standard

- … but to image an iPhone, you need to jail-break it
- Is jail-breaking forensically sound?

How do we validate tools against thousands of phones?

How do we forensically analyze 100,000 apps?

No standardized cables or extraction protocols

NIST's *Guidelines on Cell Phone Forensics* recommends:

- "searching Internet sites for developer, hacker, and security exploit information."

The Coming Digital Forensics Crisis
Part 4 — RAM and hardware forensics is really hard.

RAM Forensics—in its infancy

- RAM structures change frequently (no reason for them to stay constant.)
- RAM is constantly changing.

Malware can hide in many places:

- On disk (in programs, data, or scratch space)
- BIOS & Firmware
- RAID controllers
- GPU
- Ethernet controller
- Motherboard, South Bridge, etc
- FPGAs

Some devices will *never* be supported by today's mainstream tools.

Time travel back to 2010

Today's Research Challenges

Evidence-Oriented Design hampers tool design.

Today's tools were designed to find specific pieces of evidence

- Find child porn & financial records
- Not to assist in an investigation

Today's tools were created for solving crimes against people---

- Evidence of the crime resides on the computer.

Today's tools were not designed for:

- Explaining ***how*** a computer was compromised
- Finding information that is ***out-of-the-ordinary*** or ***out-of-place***
- Diagnosing malware infestations

Scaling — Some tools can process terabytes of data…

- … but they cannot assemble terabytes into a concise report.

Evidence-Oriented Design limits tool evolution.

Today's tools were developed to find all the evidence

—*"Tell me everything that's on this hard drive."*

- Increasingly, tools are used in time-constrained environments

—*"Show me the best stuff you can find in the next five minutes."*

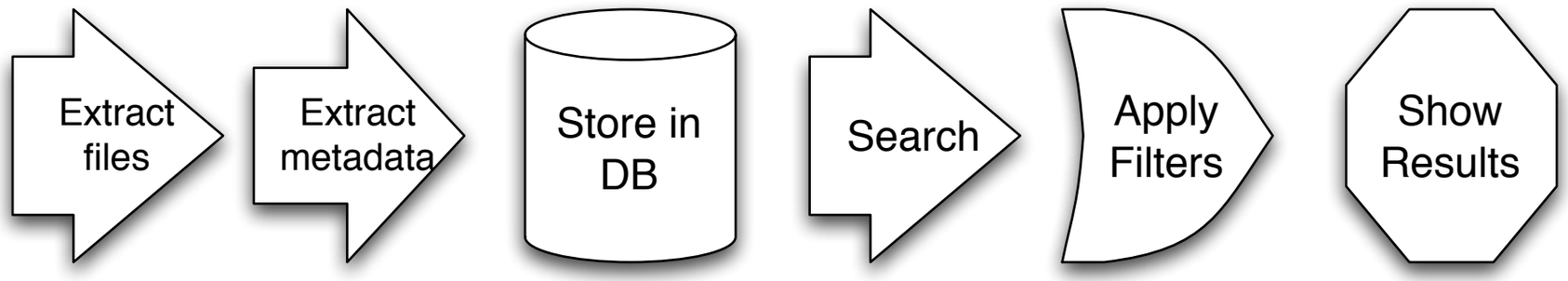
Today's tools were developed to find *documents*

- We know how to show documents to juries
- We don't know how to make arguments about "distinct sectors."
- As a result, research into incomplete documents has been slow
- It was only in 2009 that Sencar and Memon showed the second half of a JPEG could be displayed.

original

reconstructed

Today's tools follow a "Visibility, Filter and Report" model.



Problems:

- Analyst must prioritize data that is recovered
- Tools do not correlate *within* this case and *between* this case and others
- Does not readily lend itself to parallelized processing

Many tools are monolithic applications:

- Difficult to integrate with other tools
- Difficult to automate
- Difficult to combine tools from multiple vendors
- Difficult to integrate with the results of academic research.

Much of today's "research" is hacks, not science.

Most of today's “research” is really reverse-engineering

- New formats are reverse-engineered by smart people with primitive tools
- No interoperability between tools. Little effort spent on performance
- Many tools do not generalize
  - —*There are thousands of different Windows versions*
  - —*Little attention to disks/memory/network commonalities & data fusion*

Most of today's "research" is not scientific:

- No validation over a large data sets;
- Little attention to repeatability or completeness

Increasing diversity is increasingly a problem

- Some devices are *never* supported by tools.

Time travel back to 2010

A New Research Direction

We need more standardized forensic data abstractions.

Today we have limited data formats and abstractions:

- Disk images — raw & EnCase E01 files
- Packet Capture files — BPF format
- Files — distributed as files or as ZIP for collections of files
- File Signatures — List of MD5 (or SHA1) hashes in hex
- Extracted Named Entities — Stop lists. (typically in ASCII, rarely in Unicode)

We need new structured formats for distributing:

- Signatures Metrics (parts of files; n-grams; piecewise hashes; similarity metrics)
- File Metadata (e.g. Microsoft Office document properties)
- File system metadata (MAC times, etc.)
- Application Profiles (e.g. collections of files that make up an application.)
- Internet and social network information

Creating, testing, and adopting schema and formats is hard work.

We must explore alternative analysis models to "Visibility, Filter and Report."

Stream-Based Forensics

- Process the contents of the hard drive without reconstructing files
- Designed to overcome head seek latency; is this needed or useful with SSDs?
  - —*c.f. Cohen's AFF4 file-based disk imaging*

Stochastic Analysis

- Random sampling (files & sectors) to speed partial analysis.

Triage and Prioritized Analysis

- Analysis without (or during) acquisition
- "5 minute analysis"
- Examples:
  - —*I.D.E.A.L. Technology Corp.'s STRIKE*
  - —*ADF Triage*

Scale and Validation

Researchers need to work with large datasets

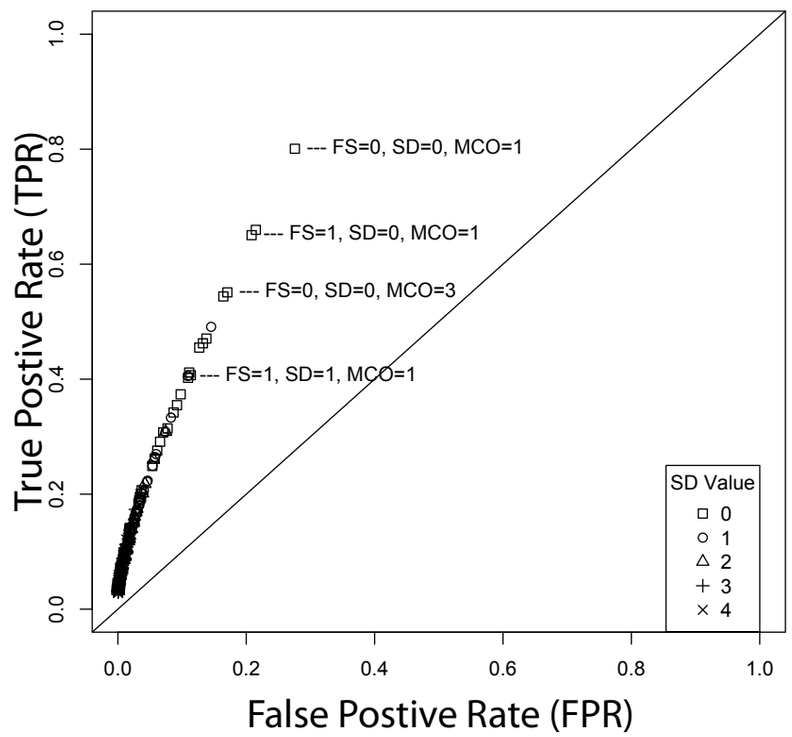
- Algorithms developed for (*n<100*) frequently fail when applied to (*n>10,000*)
- True for *n* measured in # JPEGS; TB; # hard drives; or # cell phones

Validation with standardized corpora

- Other researchers must be able to replicate your work!

Validation with standardized reporting metrics

- "Accuracy" is okay, but also report:
  - —*f*-score
  - —True Positive Rate & False Positive Rate
- Many algorithms have tunable parameters
  - —Show ROC curves!



Today's DF metrics are few and poorly articulated.

NIST Computer Forensic Tool Testing Program

- Limited testing of imaging tools & file recovery tools
- Primary to satisfy law enforcement requirements (Daubert)

http://www.cftt.nist.gov/

Academic Publishing

- DFRWS, IFIP 11.9, etc.
- "Publish or perish" evaluation

Forensic Challenges (DC3 & DFRWS)

- Stuff that's hard to do
- Not scientifically evaluated
- The "winner" is the group that
  - —*… finds the most stuff?*
  - —*… writes the most informative report?*

Moving up the Abstraction Ladder

Identity Management:

- Approaches for modeling individuals
- Simple data elements: names; email addresses; identification numbers
- More advanced: represent a person's knowledge, capabilities & social networks
- Goals: identity resolution & disambiguation

Data Visualization and Visual Analytics

- Is visualization good for *discovery,* or just for *presentation?*

Collaboration

- How can multiple investigators be used more effectively on a single case?
- How can the system automatically recognize when multiple cases are connected?
  - —*Stealth Software's private search for secret identities*

Autonomous Operation

GET EVIDENCE BUTTON

Conclusion: Digital Forensics faces an impending crisis!

Technological progress is making our job harder, not easier.

- Increasing storage densities
- Cloud Computing
- Pervasive Encryption

Given these trends, research must be *smarter* and *more applicable*

- Standardized abstractions & formats
- Standardized APIs for analysis
- Forensic Data sharing
- Composable tools

Funding agencies need to:

- Adopt open standards and procedures
- Insist on interoperability & validation.

DIGITAL INVESTIGATION 7 (2010) S64–S73

available at www.sciencedirect.com

ScienceDirect

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

Digital Investigation

Digital forensics research: The next 10 years

Simson L. Garfinkel

Naval Postgraduate School, Monterey, USA

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

ABSTRACT

Today's Golden Age of computer forensics is quickly coming to an end. Without a clear strategy for enabling research efforts that build upon one another, forensic research will fall behind the market, tools will become increasingly obsolete, and law enforcement, military and other users of computer forensics products will be unable to rely on the results of forensic analysis. This article summarizes current forensic research directions and argues that to move forward the community needs to adopt standardized, modular approaches for data representation and forensic processing.

1. Introduction

Digital Forensics (DF) has grown from a relatively obscure tradecraft to an important part of many investigations. DF tools are now used on a daily basis by examiners and analysts within local, state and Federal law enforcement; within the military and other US government organizations; and within the private "e-Discovery" industry. Developments in forensic research, tools, and process over the past decade have been very successful and many in leadership positions now rely on these tools on a regular basis—frequently without realizing it. Moreover, there seems to be a widespread belief, buttressed on by portrayals in the popular media, that advanced tools and skillful practitioners can extract actionable information from practically any device that a government, private agency, or even a skillful individual might encounter.

This paper argues that we have been in a "Golden Age of Digital Forensics," and that the Golden Age is quickly coming to an end. Increasingly organizations encounter data that cannot be analyzed with today's tools because of format incompatibilities, encryption, or simply a lack of training. Even data that can be analyzed can wait weeks or months before review because of data management issues. Without a clear research agenda aimed at dramatically improving the efficiency of both our tools and our very research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

This paper proposes a plan for achieving that dramatic improvement in research and operational efficiency through the adoption of systematic approaches for representing forensic data and performing forensic computation. It draws on more than 15 years personal experience in computer forensics, an extensive review of the DF research literature, and dozens of discussions with practitioners in government, industry, and the international forensics community.

1.1. *Prior and related work*

Although there has been some work in the DF community to create common file formats, schemas and ontologies, there has been little actual standardization. DFRWS started the Common Digital Evidence Storage Format (CDESF) Working Group in 2006. The group created a survey of disk image storage formats in September 2006, but then disbanded in August 2007 "because DFRWS did not have the resources required to achieve the goals of the group. (CDESF working group, 2009)" Hoss and Carver discuss ontologies to support digital forensics (Carver and Hoss, 2009), but did not propose any concrete ontologies that can be used. Garfinkel introduced an XML representation for file system metadata (Garfinkel, 2009), but it has not been widely adopted.

E-mail address: simsong@acm.org

1742-2876/$ – see front matter
doi:10.1016/j.diin.2010.05.009

Were predictions relevant and actionable?

1 - Extraction & Analysis
Costs have gone up … but it may not matter.

"Chip-off techniques: the rise and fall."

- We spent a lot of money developing chip-off techniques
- JTAG, exploits, and pervasive encryption made chip-off less important
  - —*Can't decrypt the flash once it is removed!*

Storage has not doubled every year

- Demand from smartphones and laptops created a shortage
- People would rather have cheaper phones


EPSNews
News For Electronics Purchasing And The Supply Chain

Supply Shortage of NAND Flash Worsens in Q4

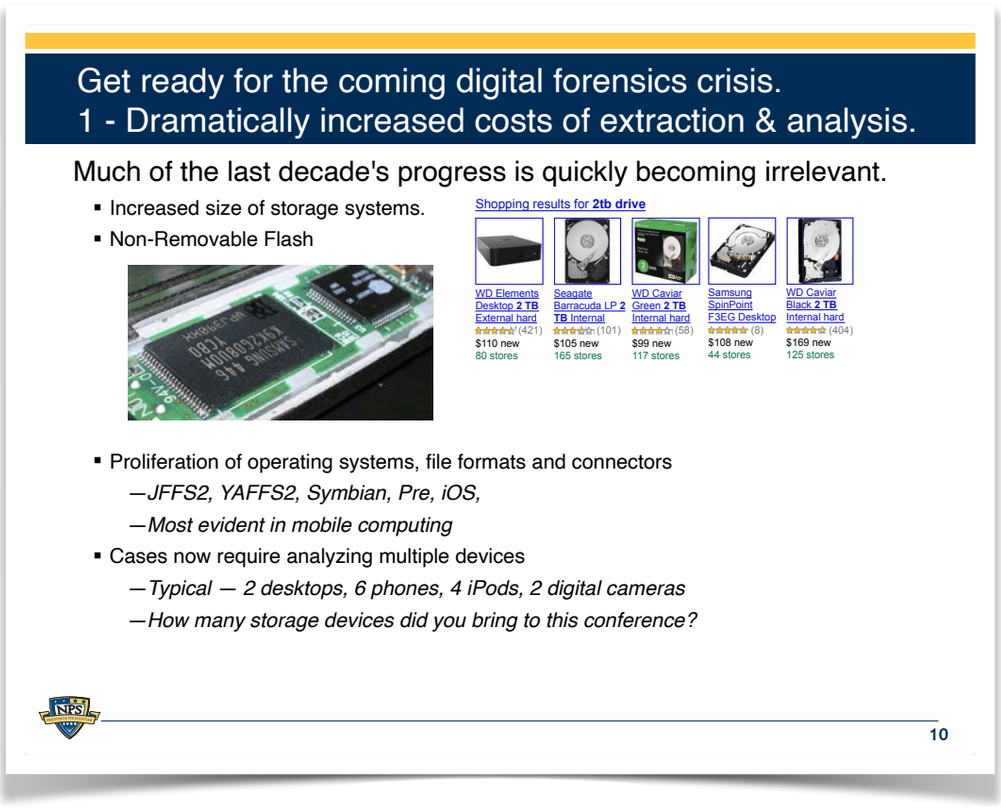
October 12, 2016 Leave a comment · Component Sourcing, News Analysis, Semiconductors, · By Gina Roos

Increasingly sophisticated analysis tools

- Insulate examiners from having to know the details

Other things that don't matter so much:

- Fragmented file recovery (clouds don't fragment)
- Undeleting files (because of TRIM)


Get ready for the coming digital forensics crisis.
1 - Dramatically increased costs of extraction & analysis.

Much of the last decade's progress is quickly becoming irrelevant.

- Increased size of storage systems.
- Non-Removable Flash

Shopping results for **2tb drive**

WD Elements Desktop 2 TB External hard	Seagate Barracuda LP 2 TB Internal	WD Caviar Green 2 TB Internal hard	Samsung SpinPoint F3EG Desktop	WD Caviar Black 2 TB Internal hard
(421)	(101)	(58)	(8)	(404)
$110 new	$105 new	$99 new	$108 new	$169 new
80 stores	165 stores	117 stores	44 stores	125 stores

- Proliferation of operating systems, file formats and connectors
  - —*JFFS2, YAFFS2, Symbian, Pre, iOS,*
  - —*Most evident in mobile computing*
- Cases now require analyzing multiple devices
  - —*Typical — 2 desktops, 6 phones, 4 iPods, 2 digital cameras*
  - —*How many storage devices did you bring to this conference?*

10

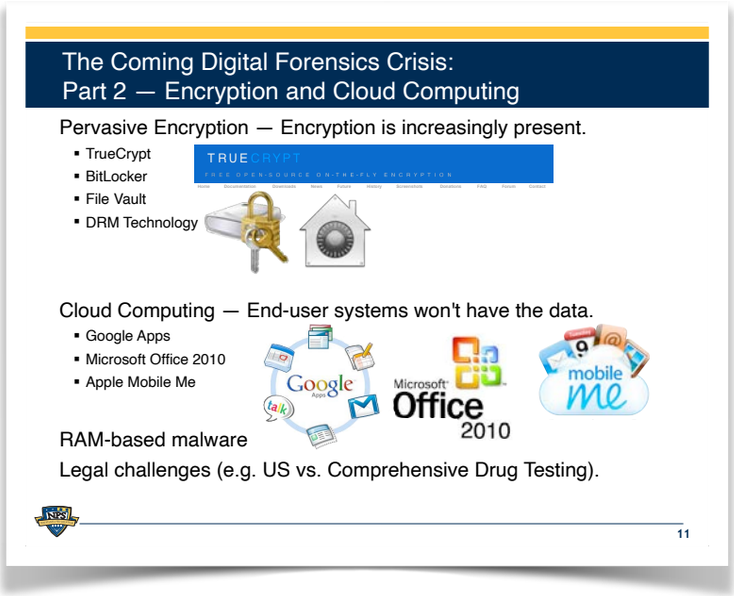
2 - Encryption and Cloud Computing

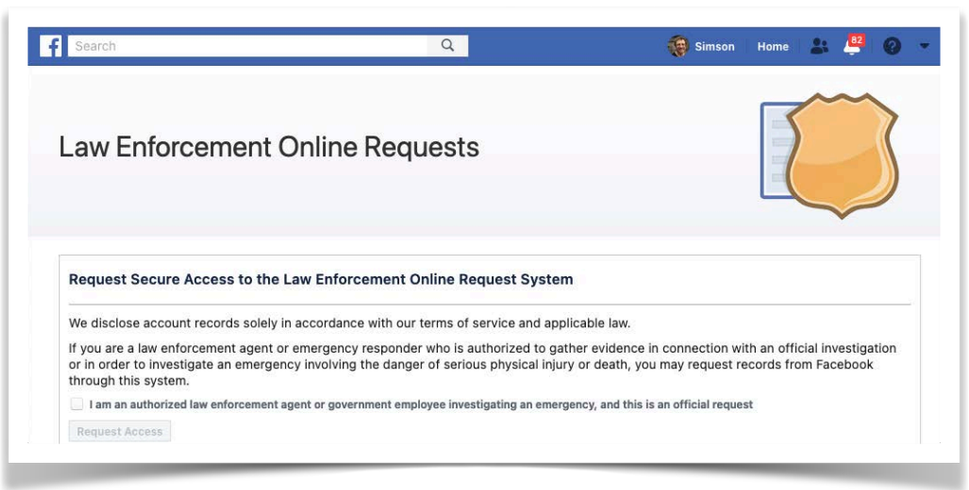
Cloud computing has *helped* forensic examiners

- Serve a warrant and download the data from a law-enforcement portal

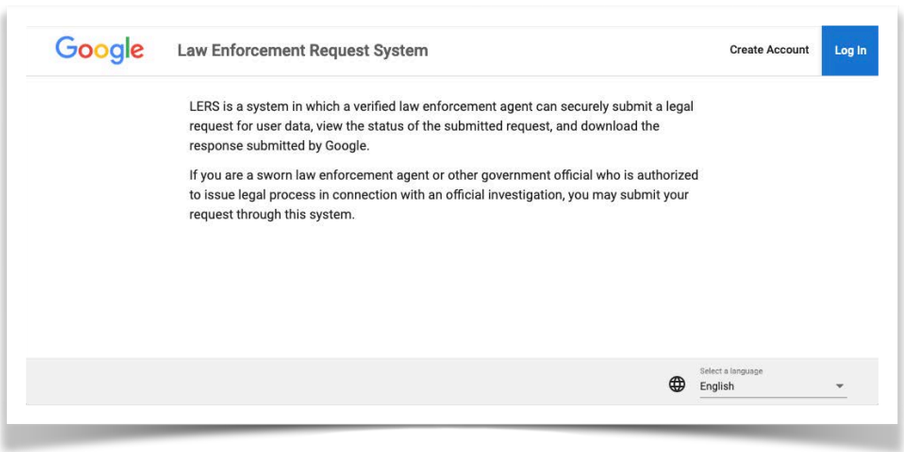
Four kinds of analysis:

- ✅ Data in cloud applications
- ✅ VMs in the cloud
- ✅ Using the cloud to analyze “big data”
- ❌ Analysis of cloud infrastructure





Facebook Law Enforcement Portal



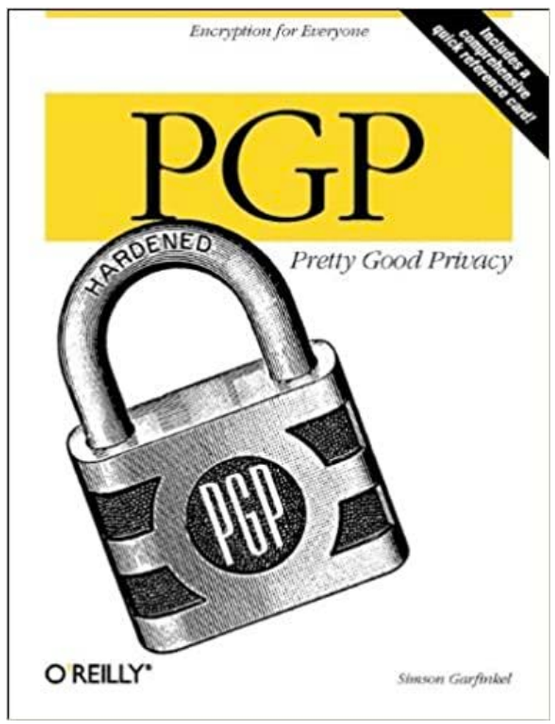
Google Law Enforcement Portal

Has encryption been a show-stopper?

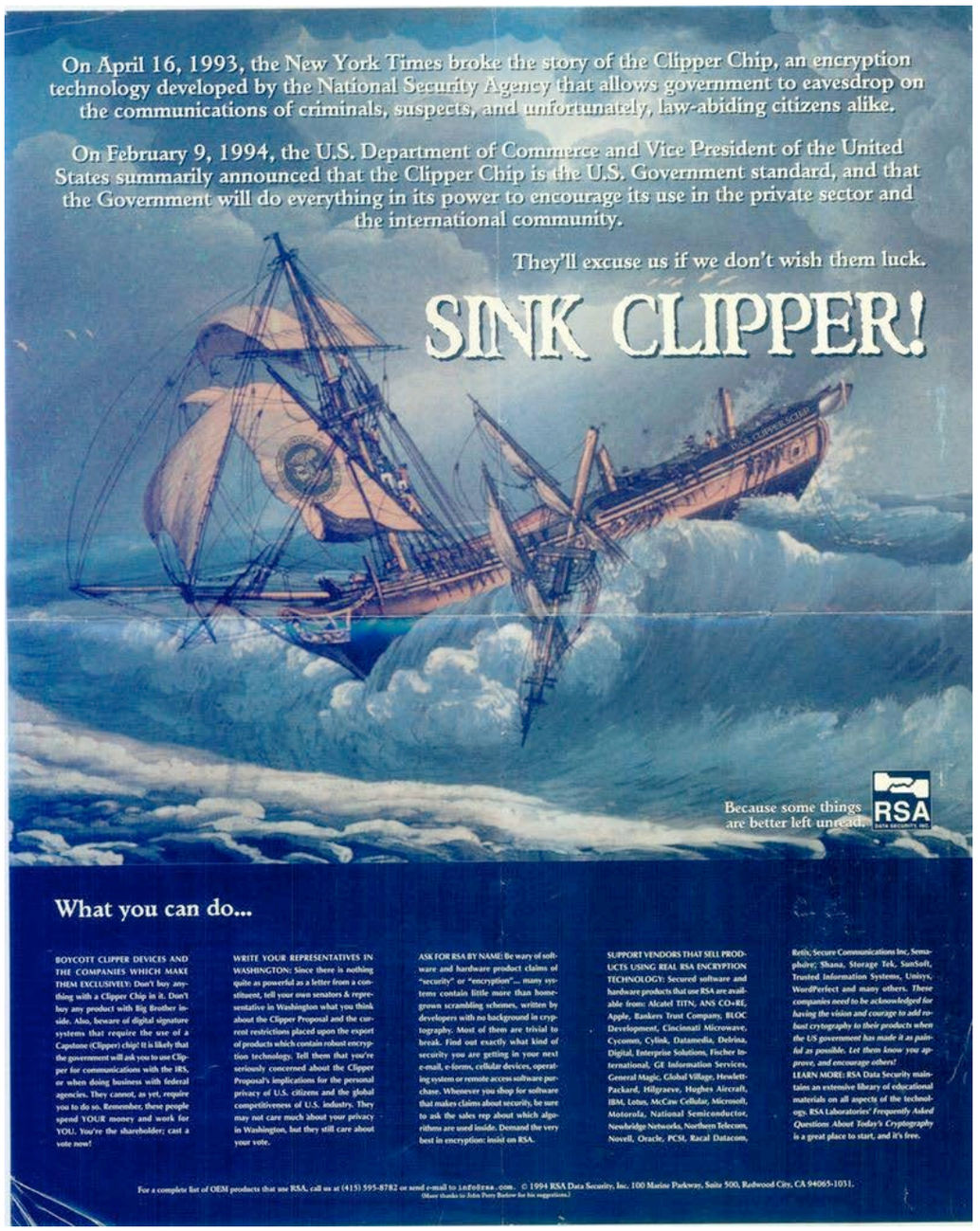
Encryption has been the focus of tension between the tech industry and governments for decades

Examples:

- US export restrictions
- 40-bit encryption in the 1990s
- US Government's "Clipper Chip"
- Fight over PGP



PGP Book (1994)



RSA Security's campaign against the NSA Clipper Chip, 1994

Apple has been at the forefront of device encryption

iPhone 4S implemented encryption between CPU and Flash

- Data automatically encrypted when written, decrypted when read
- Each phone given a unique AES key burned into silicon
- Encryption key is combined with a second key protected by user's PIN
- Second key wiped if PIN entered incorrectly 10 times

2015-09-16 - Apple releases iOS 9 with enhanced security

- Default passcode increased to 6 digits (from 4 digits)

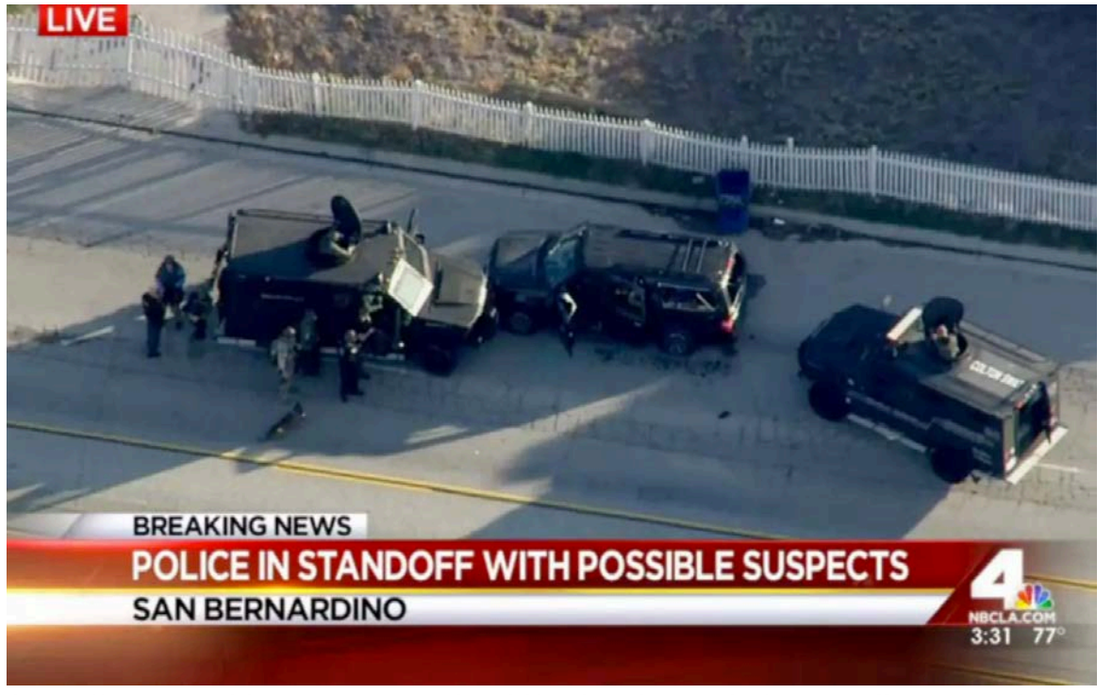
2015-12-02 San Bernardino Terrorist Attack

Husband and wife in San Bernardino storm a county event, killing 14 and seriously injuring 22

The Couple pledged allegiance to the Islamic State prior to attack.

Left behind: a locked iPhone

FBI serves Apple a warrant to write an exploit to unlock phone



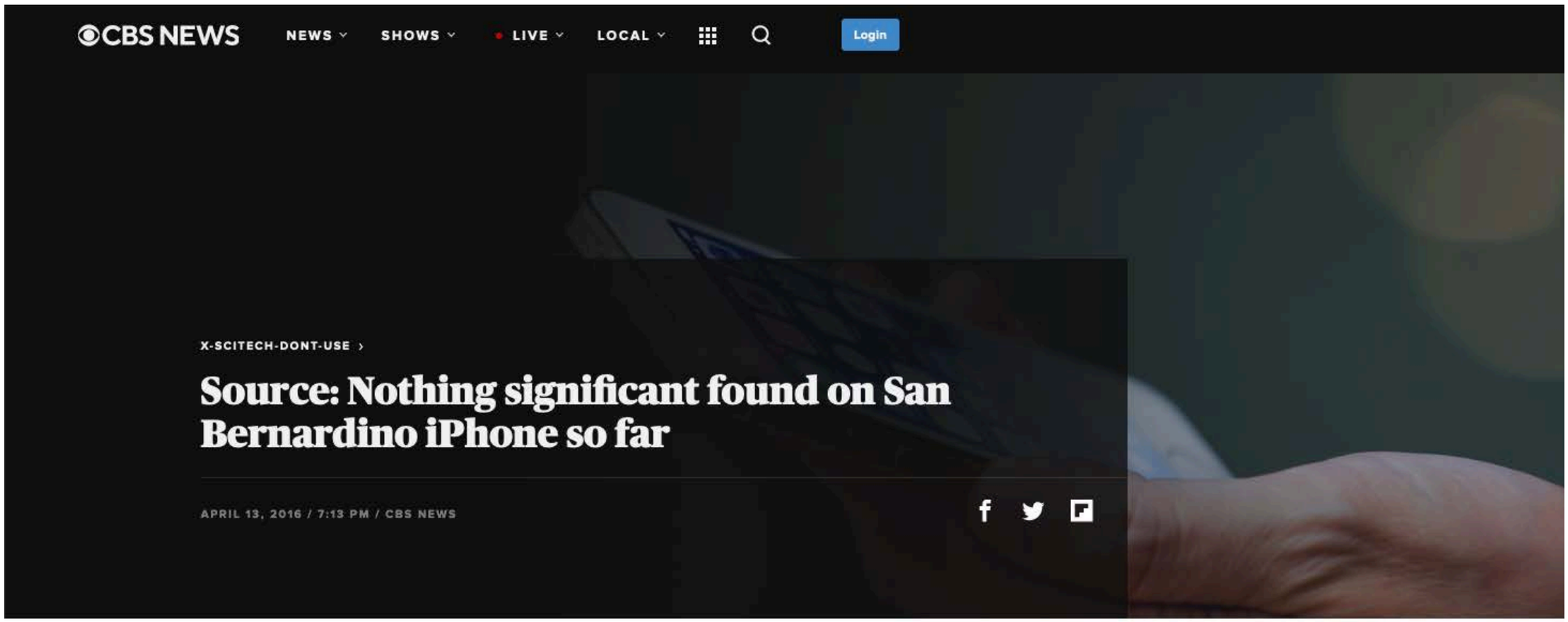
https://www.washingtonpost.com/world/national-security/report-offers-new-details-on-san-bernardino-terrorist-attack/2016/09/09/599ea266-76be-11e6-b786-19d0cb1ed06c_story.html

https://www.theguardian.com/us-news/2020/jan/14/fbi-apple-faceoff-iphone-florida-shooting

Resolution of the San Bernardino iPhone

FBI Pays $1.3M to Australian firm Azimuth to unlock phone

Nothing of interest is found on phone



But the San Bernardino iPhone was just the beginning.

2016 EncroChat Launched

Based on “EncroChat OS”

- Hardened Android

Allegedly developed for “celebrities who feared their phone conversations were being hacked.”

- End-to-End encryption, similar to PGP
- Handsets sold for €1,000 each
- Six-month contract: €1,500

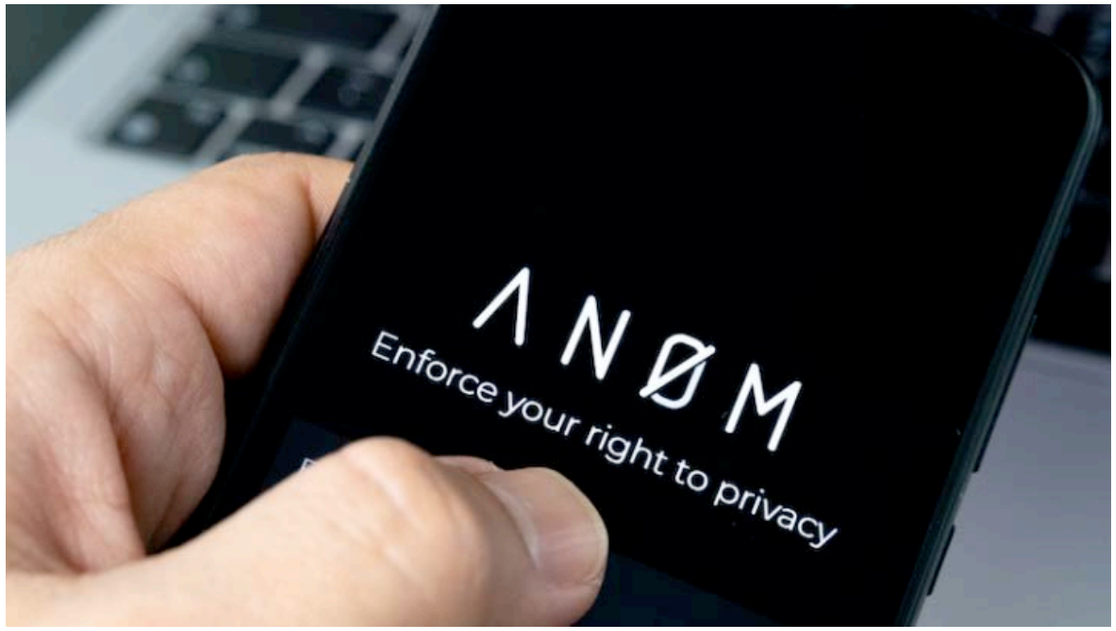
Within a few years, at least “90% of subscribers are criminals” and British National Crime Agency said “it had found no evidence that any non-criminals were subscribing to the service.”*

* https://www.ft.com/content/7006913f-be3d-49b5-8ba7-7c5b78b551b2

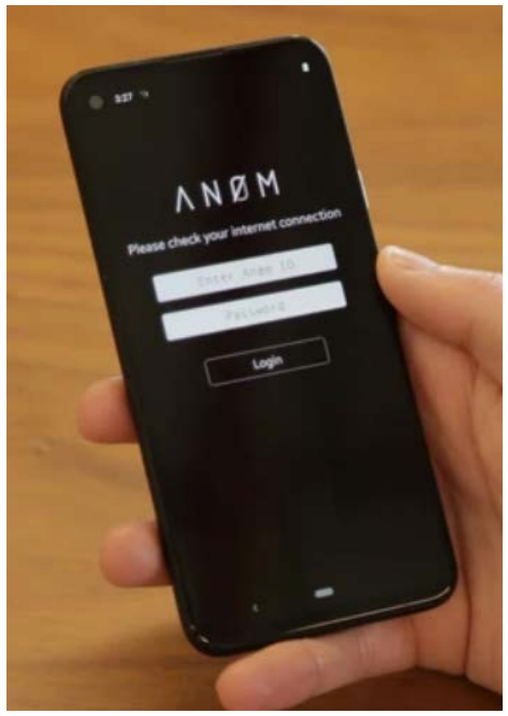
2018 Anom Messaging Platform Launched

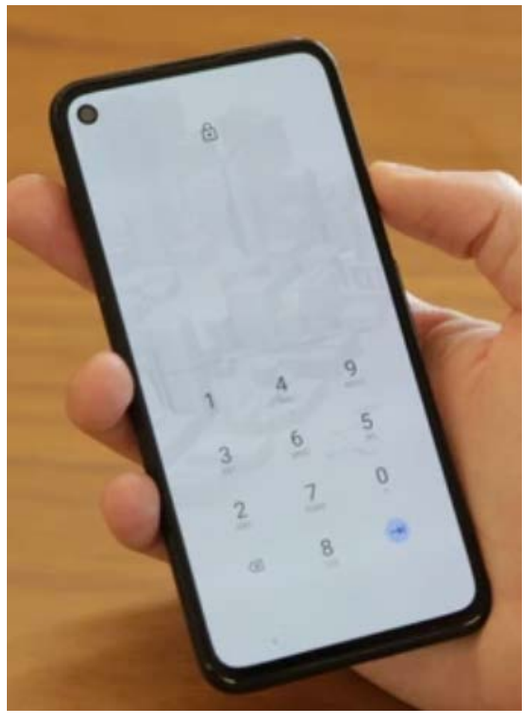
Special-purpose mobile phones

- Very expensive
- GPS removed
- One application: covert communications
- Remote wipe
- Designed for evading law enforcement



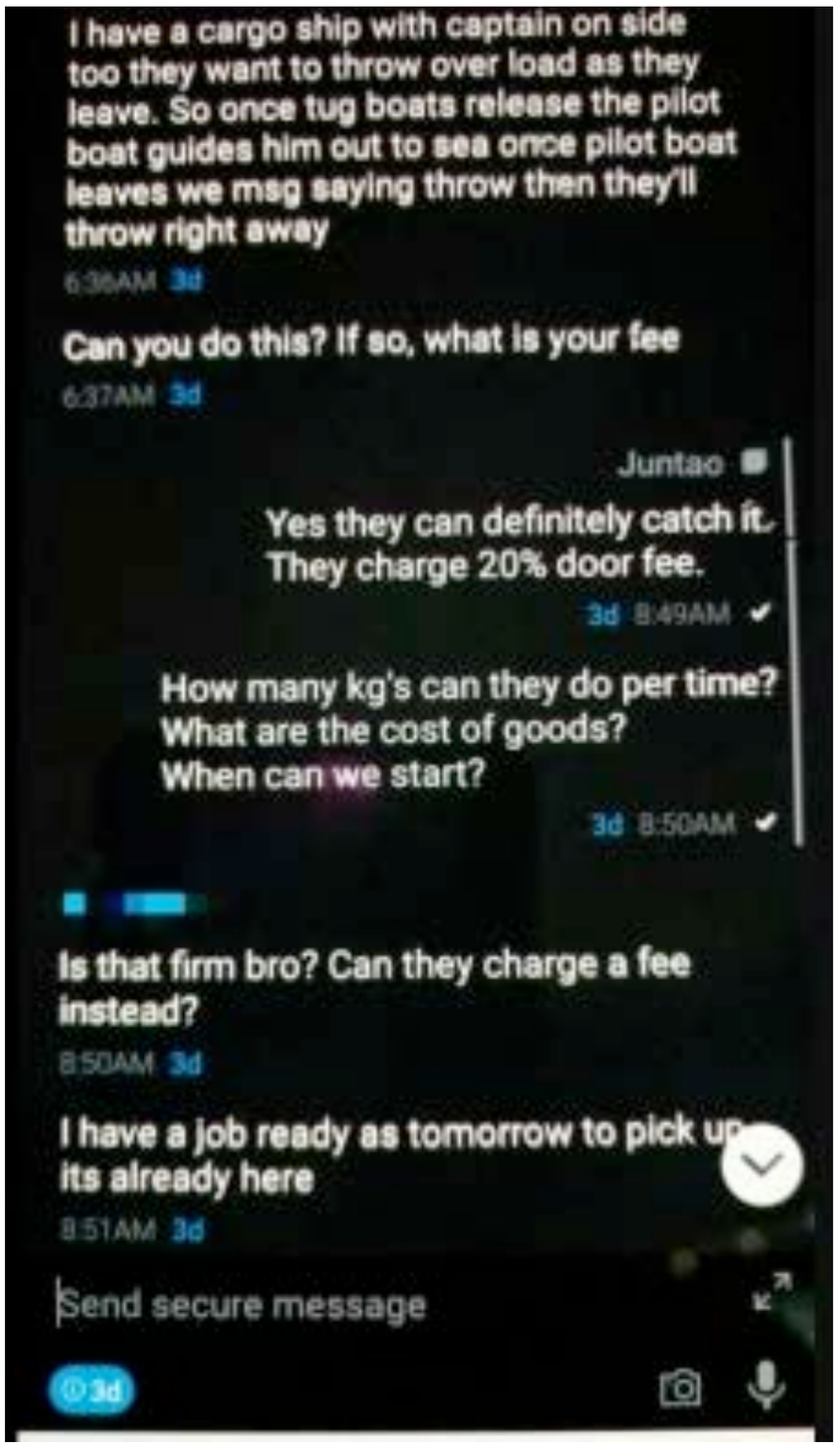
Used by more than 300 criminal groups in 100 countries





https://www.vice.com/en/article/n7b4gg/anom-phone-arcaneos-fbi-backdoor

Actual Anon message!



2020-07-02: UK and French crack EncroChat

“Operation Venetic”*

- Four year investigation
- “Thousands of criminal conspiracies”
- 60,000 users worldwide
- 10,000 users in the UK

UK and French had persistence for 2 months

- 746 arrests
- £54m criminal cash
- 77 firearms
- “Over 28 million Etizolam pills (street Valium) from an illicit laboratory”
- “55 high value cars, and 73 luxury watches”
- “and over two tonnes of drugs seized so far.”

* https://www.nationalcrimeagency.gov.uk/news/operation-venetic

2021-06-08 FBI Reveals it is behind Anom Messaging

“Operation Trojan Shield”

- 27 million messages
- 12,000 users
- 100 countries

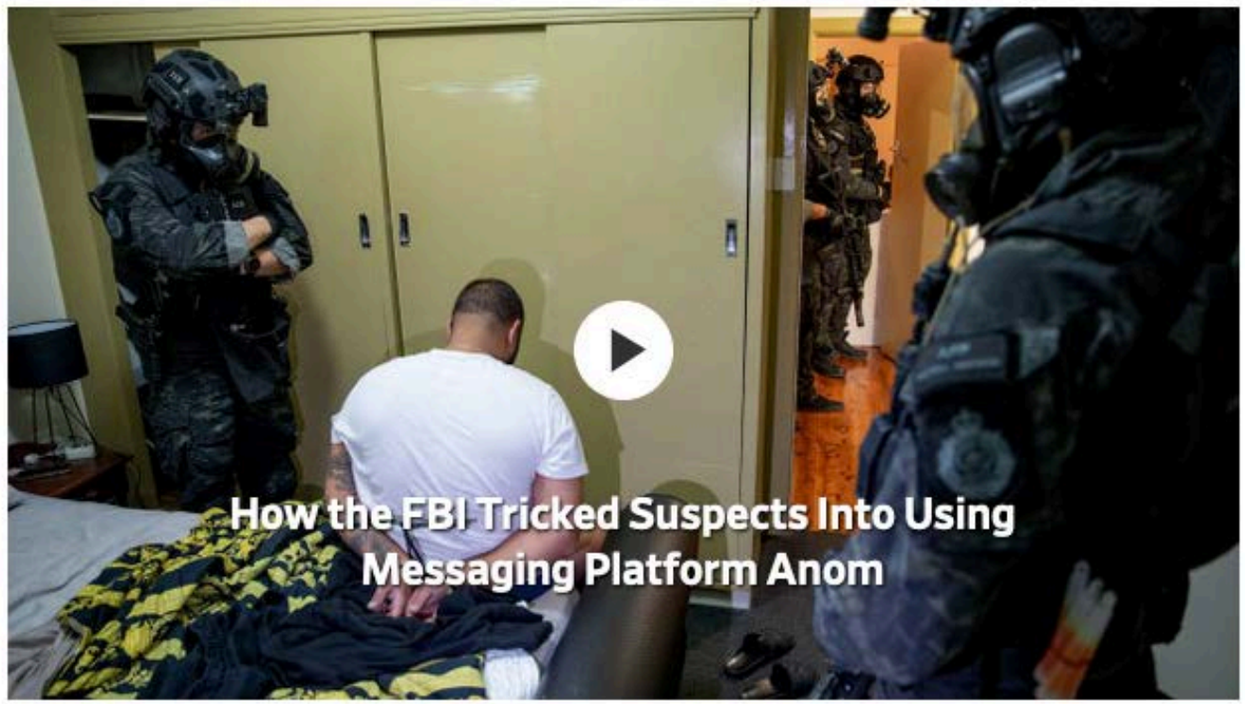
Takedown:

- 9000 law-enforcement offices
- 700 locations searched in 48 hours
- 800 arrests in 16 countries
- 8 tons of cocaine seized
- 22 tons of cannabis
- 2 tons of synthetic drugs
- 250 firearms
- 55 luxury vehicles
- $48 million in various currencies
- 150 threats to human life disrupted

U.S.

The FBI Secretly Ran the Anom Messaging Platform, Yielding Hundreds of Arrests in Global Sting

Monitoring of encrypted communications yielded over 800 arrests in 16 countries of suspected members of crime networks



The FBI secretly ran the encrypted messaging platform Anom, which was used by suspected criminal networks, to make more than 800 arrests world-wide. Here’s how law enforcement pulled off the massive global sting operation. Photo: Australian Federal Police/Zuma Press

By Byron Tau Follow *in Washington and James Marson* Follow *in Brussels*

Updated June 8, 2021 6:15 pm ET

https://www.wsj.com/articles/fbi-sting-using-anom-platform-leads-to-global-roundup-of-suspects-11623165556

Encryption and Cloud Computing: It's not really a (big) problem.

Cloud computing has helped forensic examiners

Encryption is a problem in some cases, but not many

- Encryption is not an issue with "consent searches" or victim devices
- Encryption is not an issue for most information stored in the cloud
  - —*In 2018, "there were more than 130,000 requests for digital evidence to just six tech companies — Google, Facebook, Microsoft, Oath (formerly Yahoo), and Apple. Facebook and Google got the bulk of the requests." Roughly 20% were rejected.**

US "Wiretap Report"** tracks wiretaps defeated by encryption:

Year	Total State	Crypto Wiretaps (State)	Could not Decrypt (State)	Total Fed	Crypto Wiretaps (Fed)	Could not Decrypt (Fed)
2016	1617	9	8	1551	32	29
2017	1800	102	97	2018	57	37
2018	1480	146	134	1457	74	58
2019	1808	343	334	1417	121	104
2020	1080	184	183	1297	214	200

* https://www.washingtonpost.com/world/national-security/encryption-law-enforcements-biggest-obstacle-to-digital-evidence-is-more-basic-study-finds/2018/07/24/32bcbc40-8e19-11e8-bcd5-9d911c784c38_story.html

** https://www.uscourts.gov/statistics-reports/analysis-reports/wiretap-reports

†Note: Updated in 2017 report

Washington Post - July 25, 2018

"The major problem law enforcement faces in obtaining digital evidence is not the encryption of devices **but figuring out which company holds the relevant data and how to get it,** according to a study released Wednesday by the Center for Strategic and International Studies.

Sections

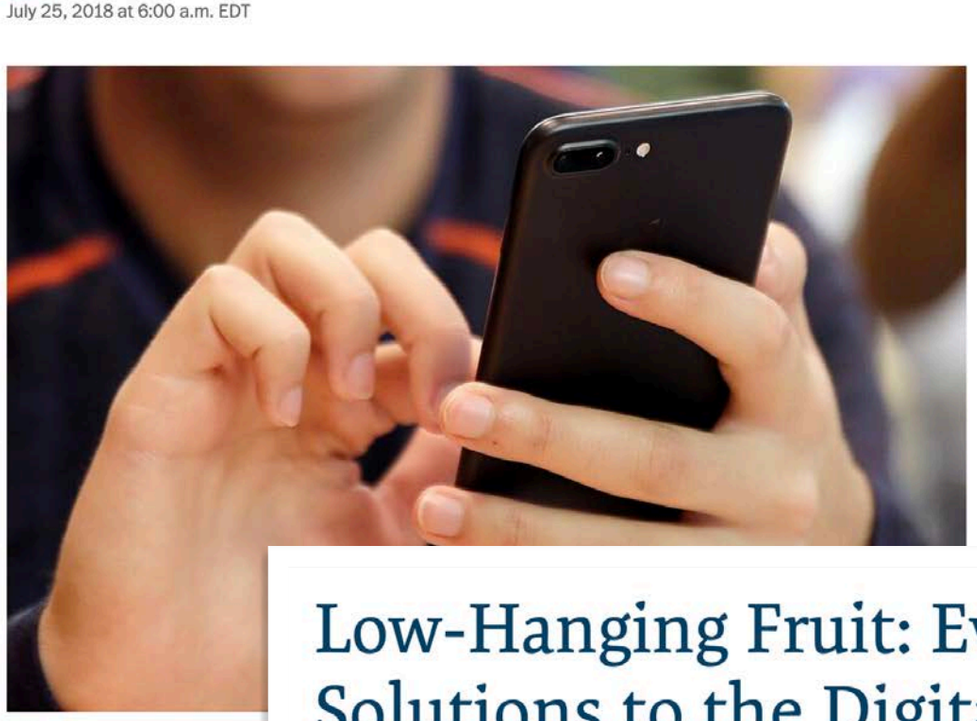
The Washington Post
Democracy Dies in Darkness

This article was published more than **4 years ago**

NATIONAL SECURITY

Encryption? Law enforcement's biggest obstacle to digital evidence is more basic, study finds.

By Ellen Nakashima
July 25, 2018 at 6:00 a.m. EDT



(AP)

Low-Hanging Fruit: Evidence-Based Solutions to the Digital Evidence Challenge

July 25, 2018

https://www.csis.org/analysis/low-hanging-fruit-evidence-based-solutions-digital-evidence-challenge

3 - Mobile Phones
Wow! What a difference 10 years makes.

Market	Android	iOS (Apple)
Global	87%	13%
US	77%	23%

https://www.statista.com/

We no longer need to validate tools against thousands of phones

Most apps are not analyzed:

- Most apps use SQLite3 to store locally
- Popular apps may be supported by a tool

Easy to analyze unlocked phones

Many tools for locked phones

- Easier to access phones that haven't been updated


The Coming Digital Forensics Crisis.
Part 3 — Mobile Phones

Forensic examiners established bit-copies as the gold standard.

- … but to image an iPhone, you need to jail-break it.
- Is jail-breaking forensically sound?

How do we validate tools against thousands of phones?

How do we forensically analyze 100,000 apps?

No standardized cables or extraction protocols.

NIST's *Guidelines on Cell Phone Forensics* recommends:

- "searching Internet sites for developer, hacker, and security exploit information."

12

Still, there is a "backlog" of devices that haven't been analyzed

Nevertheless, there seems to be a huge backlog. We don't have good measurements of its size.

2021-04-19 - Celebrate Blog

- "law enforcement is facing a huge backlog of around 6-18 months"

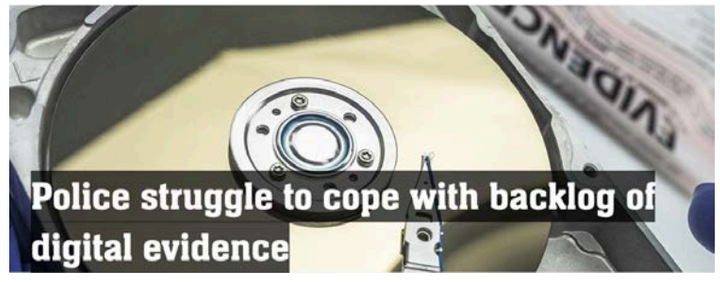
Cellebrite

2022-02-22 - Channel 4 (UK)

- "Police backlog of over 20,000 digital devices awaiting examination"

2022-04-22

- "A Freedom of Information Act request from the PA news agency has found that 12,122 devices including computers, tablets and phones, are still awaiting examination across 32 forces.



4 - RAM and hardware forensics (IoT) is really hard

Memory Forensics in 2022:

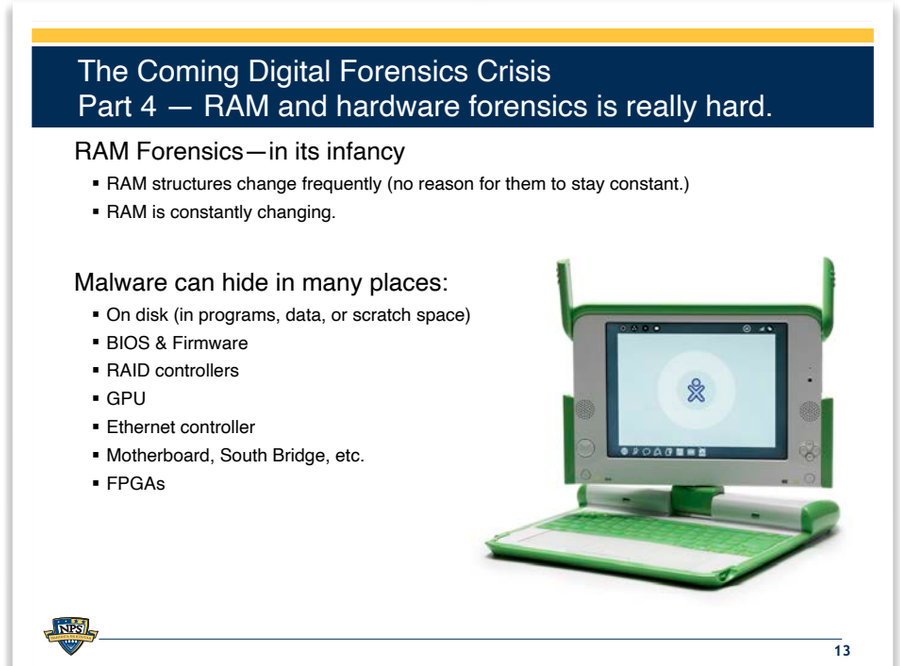
- Largely for #DFIR (Digital Forensics Incident Response)
- Mostly limited to Windows and (occasionally) Linux systems
- Generally not a part of law enforcement forensics
- Example: 1 session out of 145 mentioned “memory” in its title at 2022 National Cyber Crime Conference

Finding malware:

- Malware can hide in all of those places!
- Those who find it *rarely* are trying to build a court case

Internet of Things (IoT) forensics:

- Still largely the stuff of research papers
- Occasionally an issue in cases



The problems of the paper were the big problems of 2010. Most were resolved by 2017.

1 - Extraction and Analysis

- Encryption eliminated the value of chip-off
- Remains a problem for many locked devices, because of device encryption

2 - Encryption and Cloud Computing

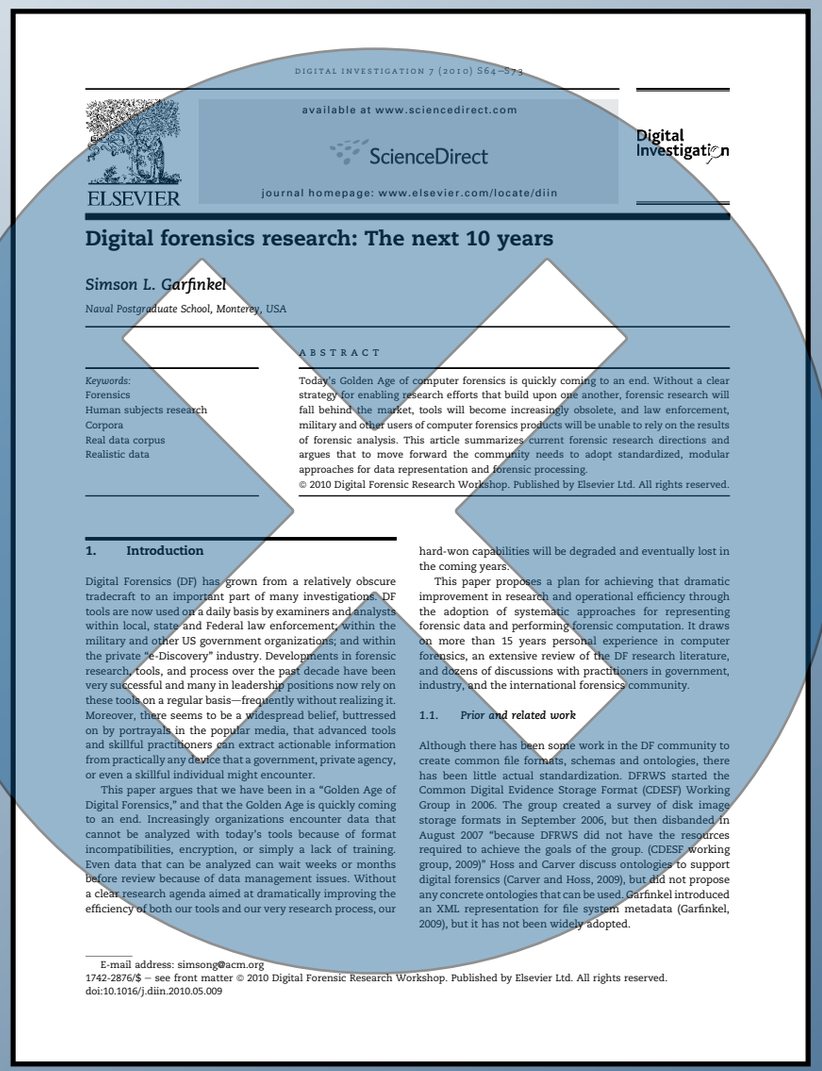
- Encryption is an issue, but not a huge issue
- Cloud computing has been a *benefit* for many investigations
- Very little attention to forensics of cloud infrastructure

3 - Mobile Phones

- The industry has largely standardized on Android and iOS

4 - RAM and hardware forensics (IoT) is really hard

- RAM analysis is now largely for DFIR, not for typical investigations.



DIGITAL INVESTIGATION 7 (2010) S64–S73

available at www.sciencedirect.com

ScienceDirect

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

Digital Investigation

Digital forensics research: The next 10 years

Simson L. Garfinkel

Naval Postgraduate School, Monterey, USA

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

ABSTRACT

Today's Golden Age of computer forensics is quickly coming to an end. Without a clear strategy for enabling research efforts that build upon one another, forensic research will fall behind the market, tools will become increasingly obsolete, and law enforcement, military and other users of computer forensics products will be unable to rely on the results of forensic analysis. This article summarizes current forensic research directions and argues that to move forward the community needs to adopt standardized, modular approaches for data representation and forensic processing.

1. Introduction

Digital Forensics (DF) has grown from a relatively obscure tradecraft to an important part of many investigations. DF tools are now used on a daily basis by examiners and analysts within local, state and Federal law enforcement, within the military and other US government organizations, and within the private "e-Discovery" industry. Developments in forensic research, tools, and process over the past decade have been very successful and many in leadership positions now rely on these tools on a regular basis—frequently without realizing it. Moreover, there seems to be a widespread belief, buttressed on by portrayals in the popular media, that advanced tools and skillful practitioners can extract actionable information from practically any device that a government, private agency, or even a skillful individual might encounter.

This paper argues that we have been in a "Golden Age of Digital Forensics," and that the Golden Age is quickly coming to an end. Increasingly organizations encounter data that cannot be analyzed with today's tools because of format incompatibilities, encryption, or simply a lack of training. Even data that can be analyzed can wait weeks or months before review because of data management issues. Without a clear research agenda aimed at dramatically improving the efficiency of both our tools and our very research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

This paper proposes a plan for achieving that dramatic improvement in research and operational efficiency through the adoption of systematic approaches for representing forensic data and performing forensic computation. It draws on more than 15 years personal experience in computer forensics, an extensive review of the DF research literature, and dozens of discussions with practitioners in government, industry, and the international forensics community.

1.1. *Prior and related work*

Although there has been some work in the DF community to create common file formats, schemas and ontologies, there has been little actual standardization. DFRWS started the Common Digital Evidence Storage Format (CDESF) Working Group in 2006. The group created a survey of disk image storage formats in September 2006, but then disbanded in August 2007 "because DFRWS did not have the resources required to achieve the goals of the group. (CDESF working group, 2009)" Hoss and Carver discuss ontologies to support digital forensics (Carver and Hoss, 2009), but did not propose any concrete ontologies that can be used. Garfinkel introduced an XML representation for file system metadata (Garfinkel, 2009), but it has not been widely adopted.

E-mail address: simsong@acm.org

1742-2876/$ – see front matter © 2010 Digital Forensic Research Workshop. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2010.05.009

What did the paper completely miss?

Methodology

Search for “digital forensics” for the years 2011-2021

- ACM Digital Library - 503 results
- IEEE Xplore - 2487 results
  - —2096 conferences, 265 journals, 78 magazines, 21 books, 26 “early access articles”

Review DFRWS Agendas for the years 2011-2021

Attend National Cyber Crime Conference 2022

Forensic topics that were important in 2010-2020 that we didn't see in 2010

Complete miss:

- Open Source Intelligence (OSINT)
- Bitcoin Forensics
- Email spoofing
- Detecting "Deep Fakes" (photo forensics, video forensics)
- Misinformation
- Ransomware
- Lawful interception

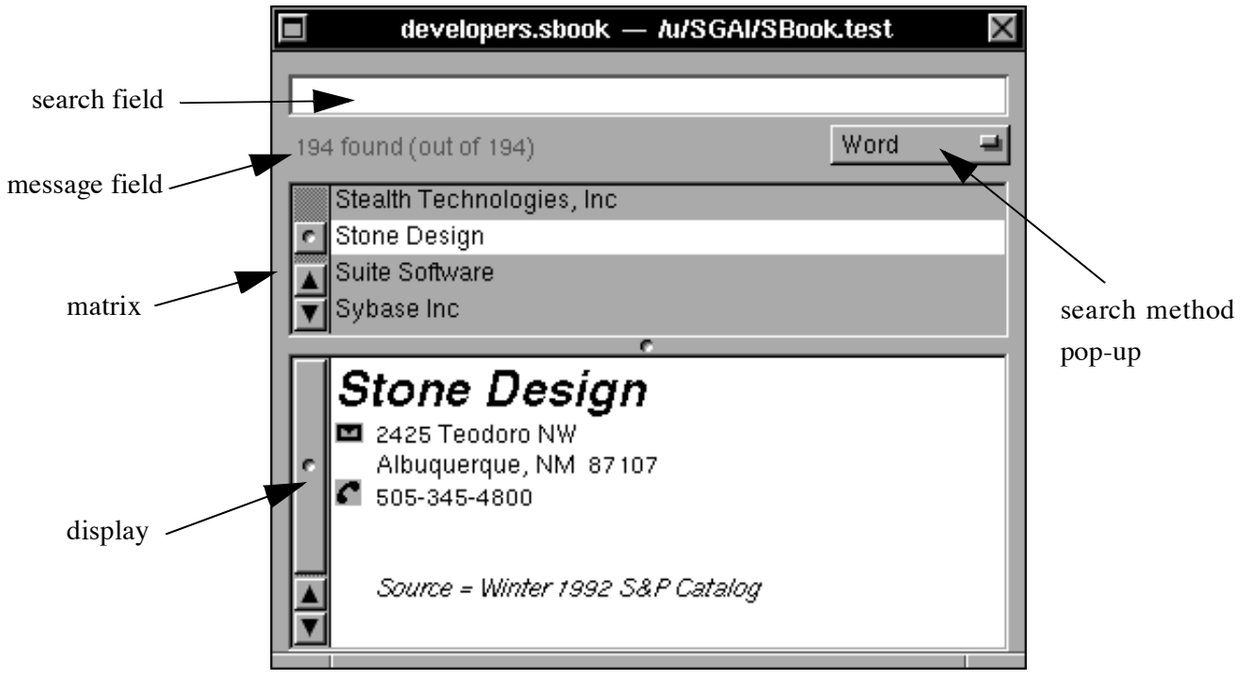
We heavily researched the wrong kind of similarity matching!

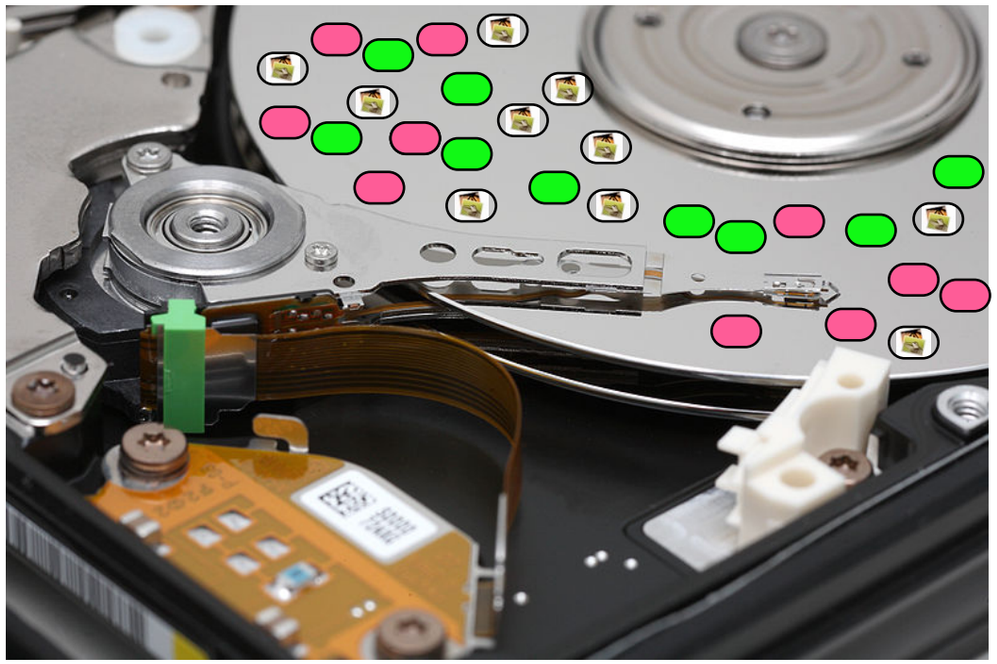
- Lots of academic research on byte streams
- But the interest was in images and videos

Lots of law enforcement interest, but few research publications:

- SCADA systems, power networks, etc.
- Vehicles — Cars, Trucks, "Heavy Vehicles"

Up Next:
Hour 2 — Transitioning Research to Practice





https://pixabay.com/illustrations/hacker-computer-ghost-cyber-code-4031973/

Hour 2

Transitioning Research to Practice

Simson L. Garfinkel
Senior Data Scientist, Department of Homeland Security

June 10, 2022

Transitioning Research to Practice

The “drives” project (1998-2014) & Digital Corpora (2012-)

- How I got into stored data forensics
- Realistic data you can download today! For Free!

bulk_extractor (2005-2014, 2019-)

- A research tool that is useful for cases

sector hashing (2009-2014)

- Developing and commercializing a technique (but not the same people!)

The drives project (1981-2014)

In 1998,
I purchased 10 used computers from a computer store.

Four of the 10 computers contains sensitive data.

#1 — File server from a law firm

#2 — Database of mental health patients

#3 — Financial records from a woman getting a divorce

#4 — Draft manuscript of a novel

I decided to scale the research

I purchased 150 drives from eBay

All Categories Save this search

350 items found for **hard drives**

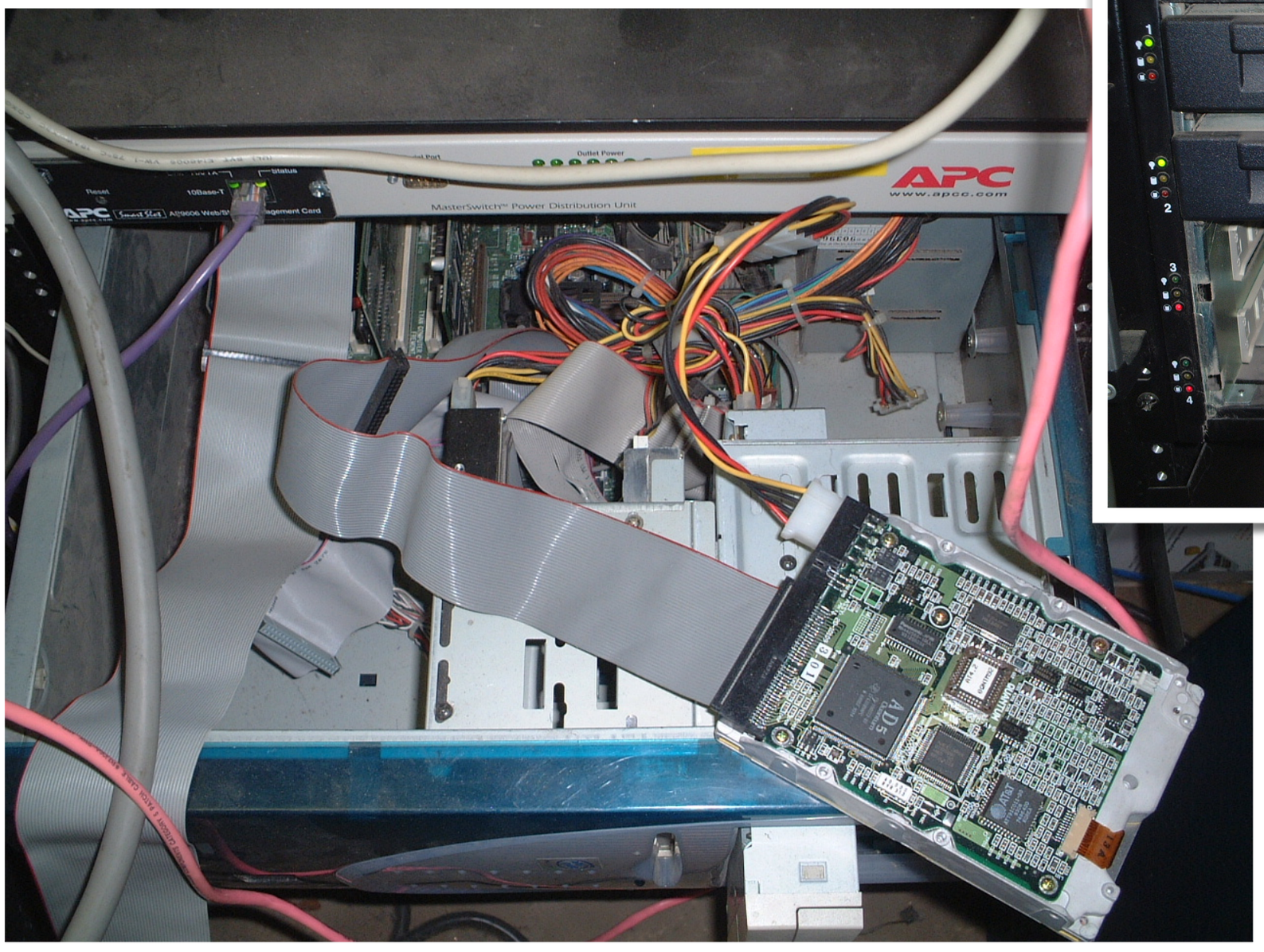
Sort by items: **ending first** | newly listed | lowest priced | highest priced

Picture hide	Item Title	Price	Bids	Time Left
	Lot of hard and floppy drives	$5.50	2	14m
	Lot of hard and floppy drives	$5.50	2	22m
	Lot of hard and floppy drives	$5.50	2	25m
	Lot of 2 hard drives IDE	$8.00	12	29m
	3.2 gig Hard Drives	$180.00	-	59m
	(5) 1.2 hard drives & (15) 10/100 network	$25.00	1	1h 00m
	Lot of 3 Quantum 9.1 gig SCSI Hard Drives	$26.00	6	1h 25m
	IDE HARD DRIVES (3)	$6.50	6	1h 46m
	LOT OF 5 Hard Drives! 3.2 Gig Western Digital	$120.00 $124.95 Buy It Now	-	1h 50m
	QTY 3....IDE Hard Drives 2.5 Gig	$20.50	5	2h 02m
	5 WESTERN DIGITAL 2.5 GIG HARD DRIVES	$30.00	4	2h 03m
	QTY 3....IDE Hard Drives 1.0 Gig	$9.99	1	2h 04m
	Western Digital 850 meg IDE Hard Drives dutch	$6.00	1	2h 57m
	WINDOWS	$6.00	-	3h 18m

The drives typically cost $5-$10 each, plus shipping.

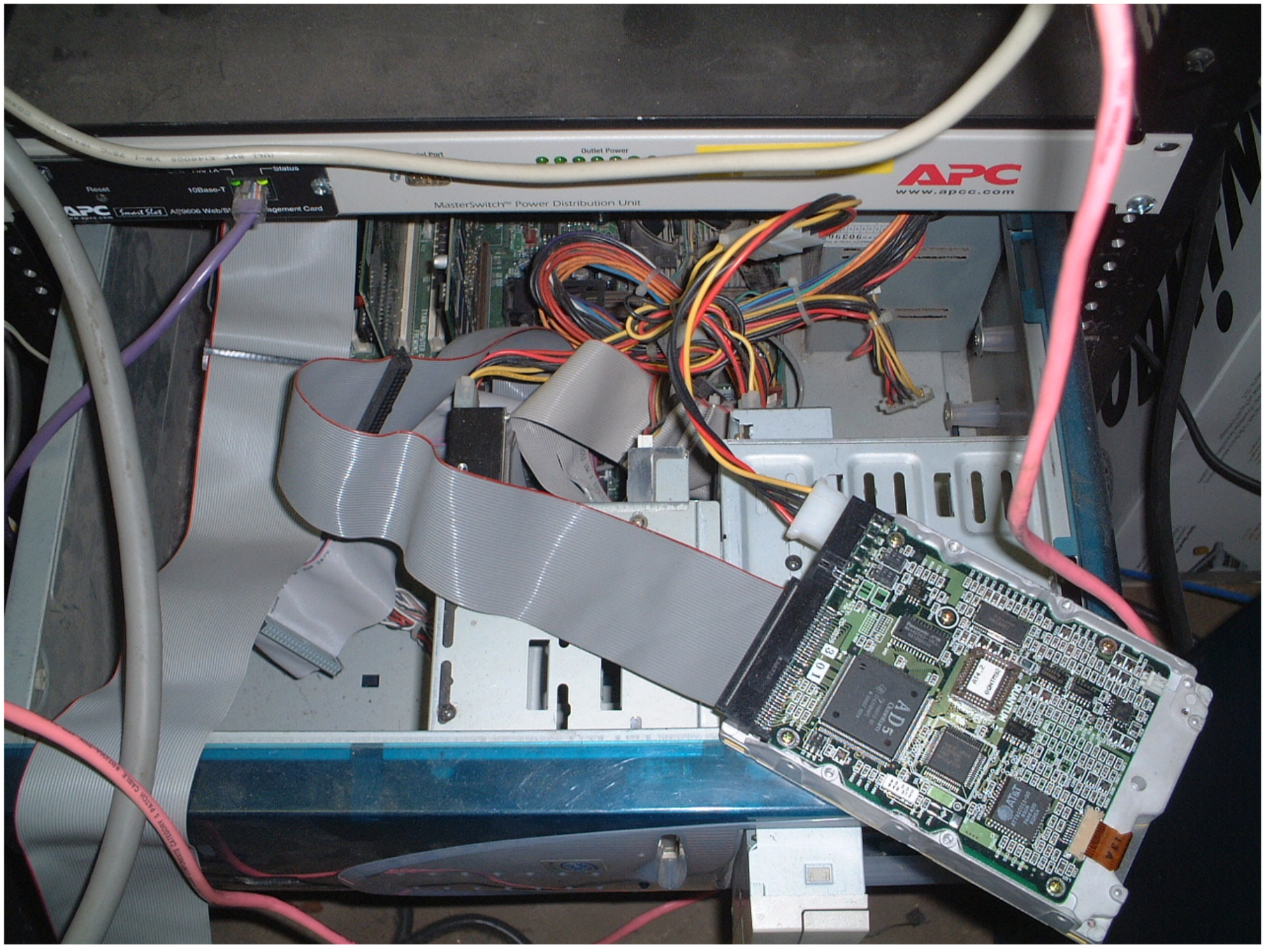
I wrote software to image the drives and stored the images on a RAID array.

You can hot-swap ATA cables under FreeBSD!



I wrote my own software to image the drive.

You can hot-swap ATA cables under FreeBSD!



Original drives were numbered and retained in storage.

Sept. 2002: I entered the MIT CS Ph.D. program.

How many of the drives had confidential data?

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!*

But it got the job done!

Working drives purchased on eBay fell into three categories:

- Properly wiped (most sectors all NULLs)
- Many files allocated (“df” command said drive was 50% - 95% filled)
- Few allocated files, but most sectors were not NULLs

Quick way to determine drive sanitization: compress the disk image!

“Remembrance of Data Passed,” IEEE Security & Privacy, January/February 2003.

Key findings:

- ⅓ of drives had confidential information
- Very few drives properly sanitized

Jan 2002 drives	150
Jan 2004 drives	235
Drives DOA	59
Drives Images	176
Total files:	168,459
Total data	125 GB


Open-Source Security

Remembrance of Data Passed: A Study of Disk Sanitization Practices

Many discarded hard drives contain information that is both confidential and recoverable, as the authors' own experiment shows. The availability of this information is little publicized, but awareness of it will surely spread.

SIMSON L. GARFINKEL AND ABHI SHELAT
Massachusetts Institute of Technology

A fundamental goal of information security is to design computer systems that prevent the unauthorized disclosure of confidential information. There are many ways to assure this information privacy. One of the oldest and most common techniques is physical isolation: keeping confidential data on computers that only authorized individuals can access. Most single-user personal computers, for example, contain information that is confidential to that user.

Computer systems used by people with varying authorization levels typically employ authentication, access control lists, and a privileged operating system to maintain information privacy. Much of information security research over the past 30 years has centered on improving authentication techniques and developing methods to assure that computer systems properly implement these access control rules.

Cryptography is another tool that can assure information privacy. Users can encrypt data as it is sent and decrypt it at the intended destination, using, for example, the secure sockets layer (SSL) encryption protocol. They can also encrypt information stored on a computer's disk so that the information is accessible only to those with the appropriate decryption key. Cryptographic file systems[1–3] ask for a password or key on startup, after which they automatically encrypt data as it's written to a disk and decrypt the data as it's read; if the disk is stolen, the data will be inaccessible to the thief. Yet despite the availability of cryptographic file systems, the general public rarely seems to use them.

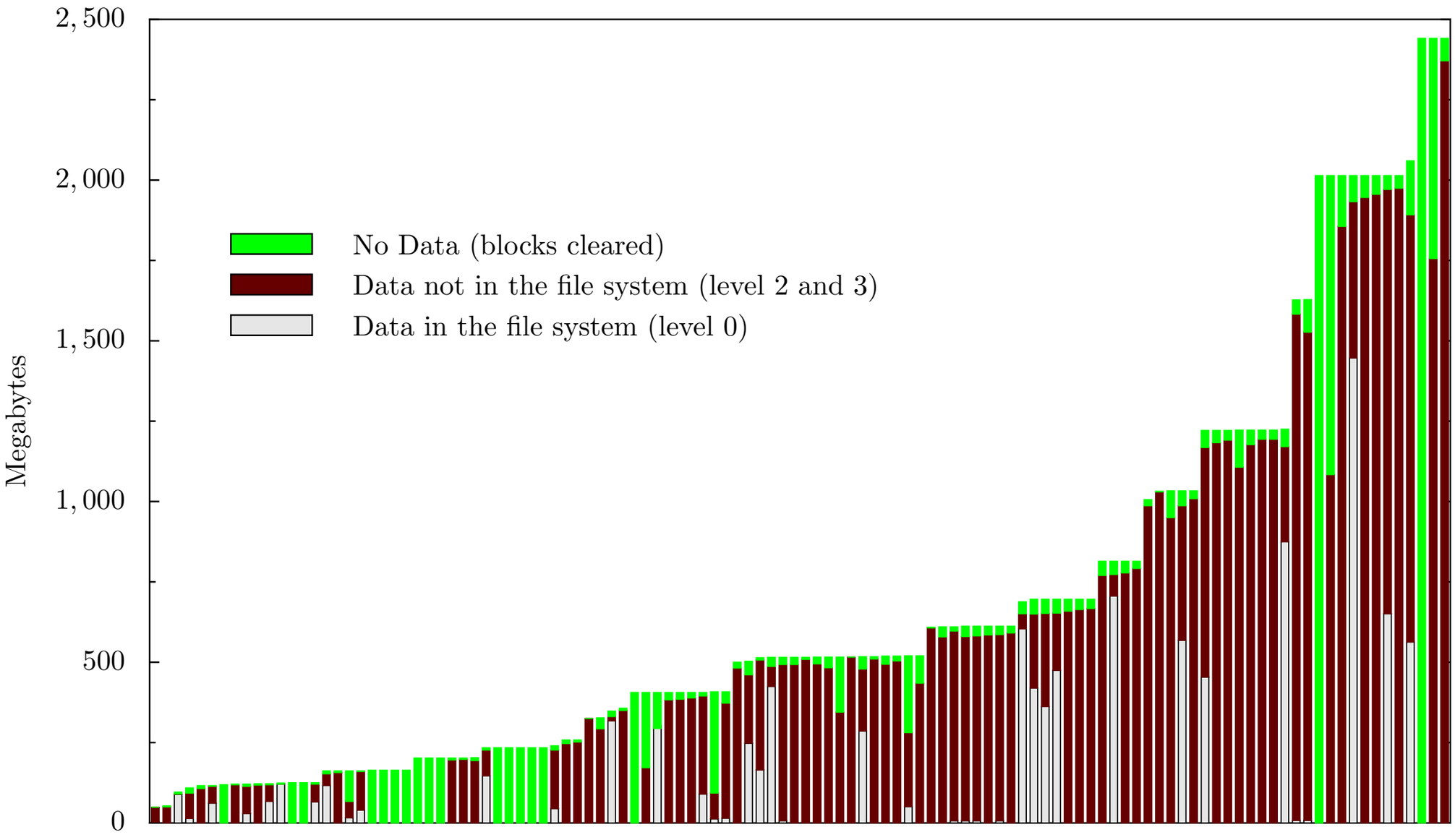
Absent a cryptographic file system, confidential information is readily accessible when owners improperly retire their disk drives. In August 2002, for example, the United States Veterans Administration Medical Center in Indianapolis retired 139 computers. Some of these systems were donated to schools, while others were sold on the open market, and at least three ended up in a thrift shop where a journalist purchased them. Unfortunately, the VA neglected to *sanitize* the computer's hard drives—that is, it failed to remove the drives' confidential information. Many of the computers were later found to contain sensitive medical information, including the names of veterans with AIDS and mental health problems. The new owners also found 44 credit card numbers that the Indianapolis facility used.[4]

The VA fiasco is just one of many celebrated cases in which an organization entrusted with confidential information neglected to properly sanitize hard disks before disposing of computers. Other cases include:

- In the spring of 2002, the Pennsylvania Department of Labor and Industry sold a collection of computers to local resellers. The computers contained "thousands of files of information about state employees" that the department had failed to remove.[5]
- In August 2001, Dovebid auctioned off more than 100 computers from the San Francisco office of the Viant consulting firm. The hard drives contained confidential client information that Viant had failed to remove.[6]
- A Purdue University student purchased a used Macintosh computer at the school's surplus equipment exchange facility, only to discover that the computer's hard drive contained a FileMaker database containing the names and demographic information for more than 100 applicants to the school's Entomology Department.
- In August 1998, one of the authors purchased 10 used computer systems from a local computer store. The computers, most of which were three to five years old,

PUBLISHED BY THE IEEE COMPUTER SOCIETY ■ 1540-7993/03/$17.00 © 2003 IEEE ■ IEEE SECURITY & PRIVACY 17

Most second-hand drives were not properly sanitized.

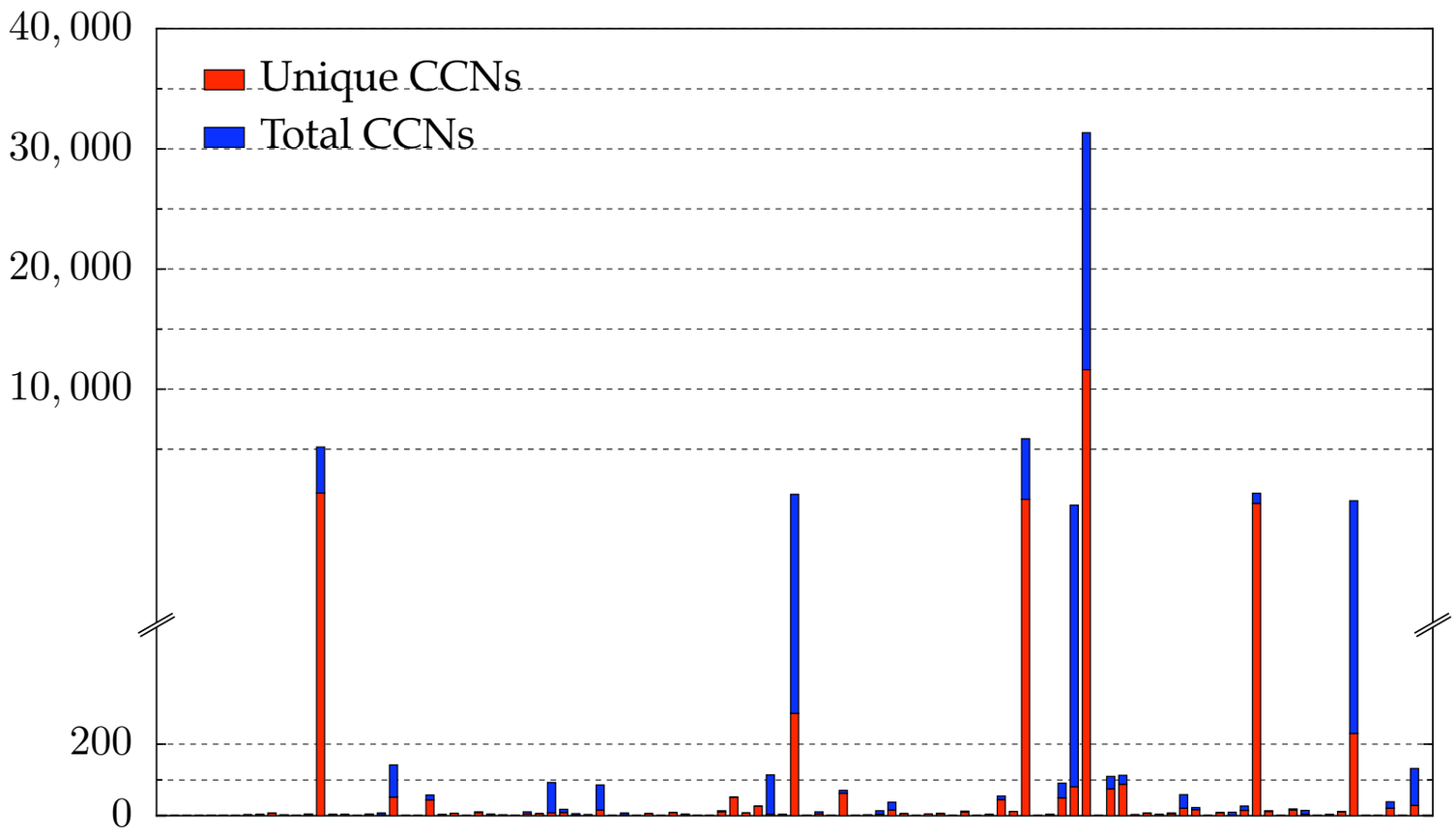


Several vendors contacted me to commercialize this research.

We wrote a program to scan the raw disk images for credit card numbers

Finding CCNs required:

- finding digits in characteristic CCN pattern
- validating LUN
- examining context

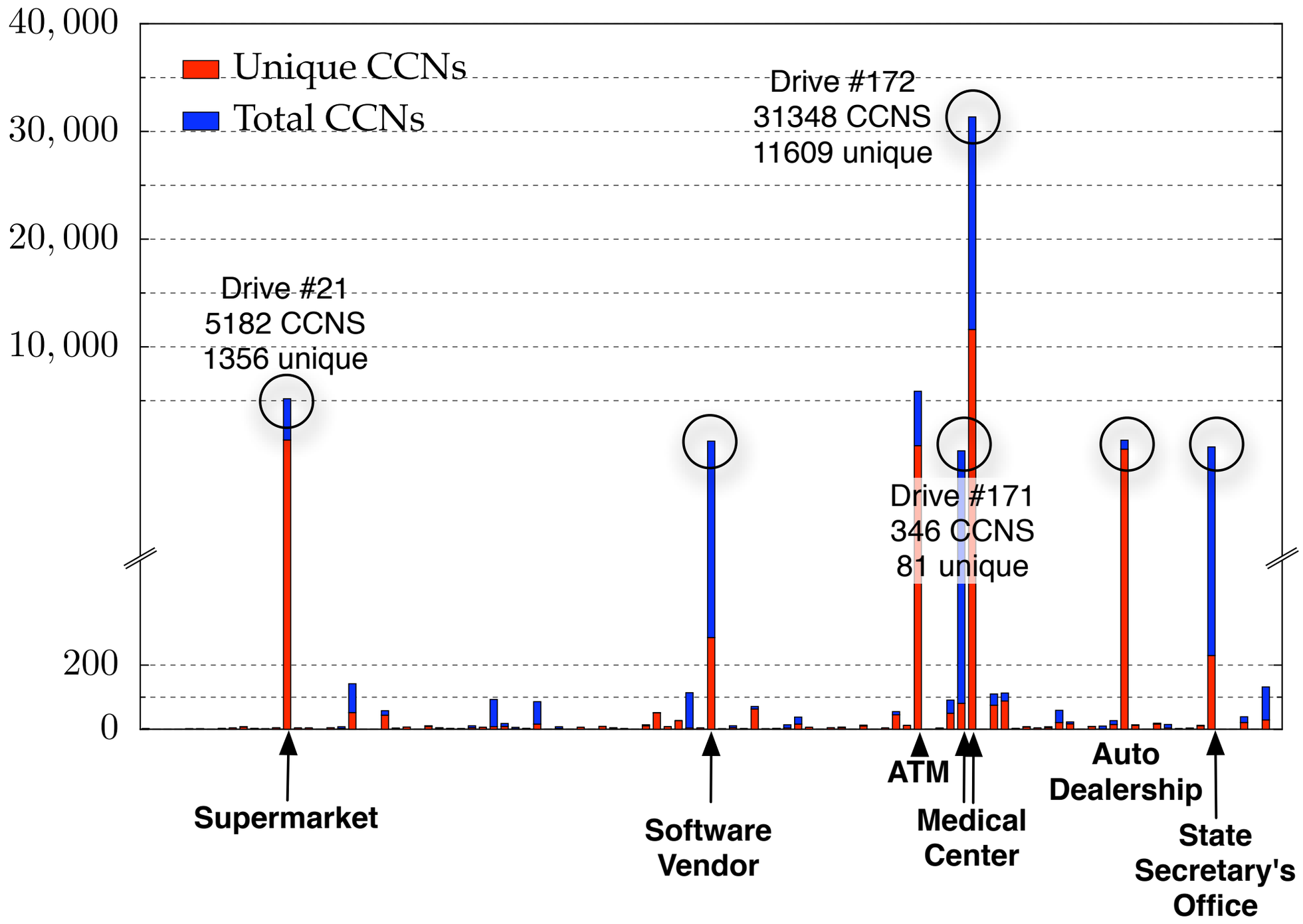


Most drives had just a few, but some drives had a lot

- No drives should have a lot of CCNs.

  By definition, all of these drives are interesting.

In 2004, I developed the email histogram technique to identify the previous owners of the disk drives.



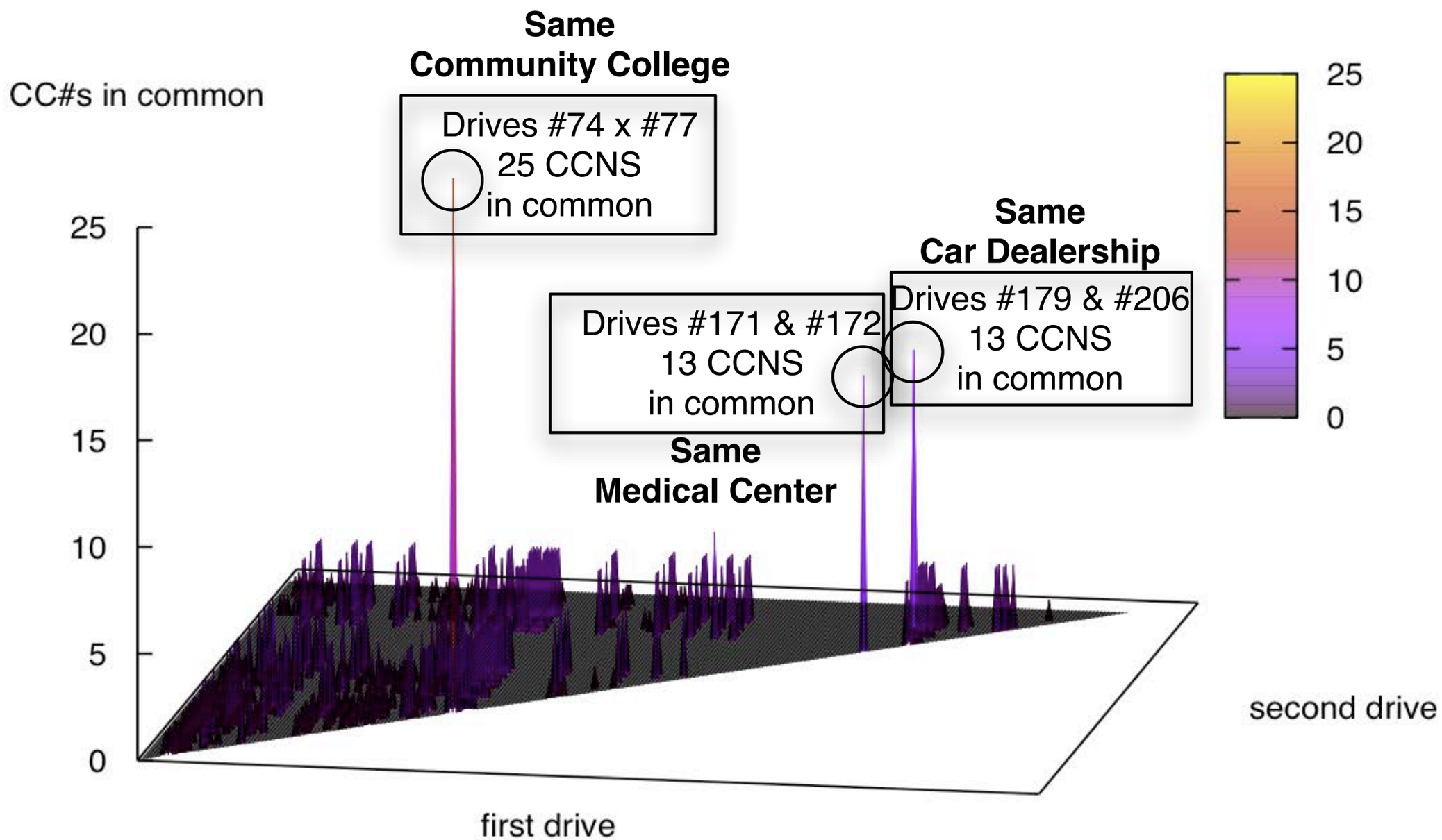
One day a grad student and I were talking about the credit card numbers on the disk drives…

Student lounge at MIT CSAIL

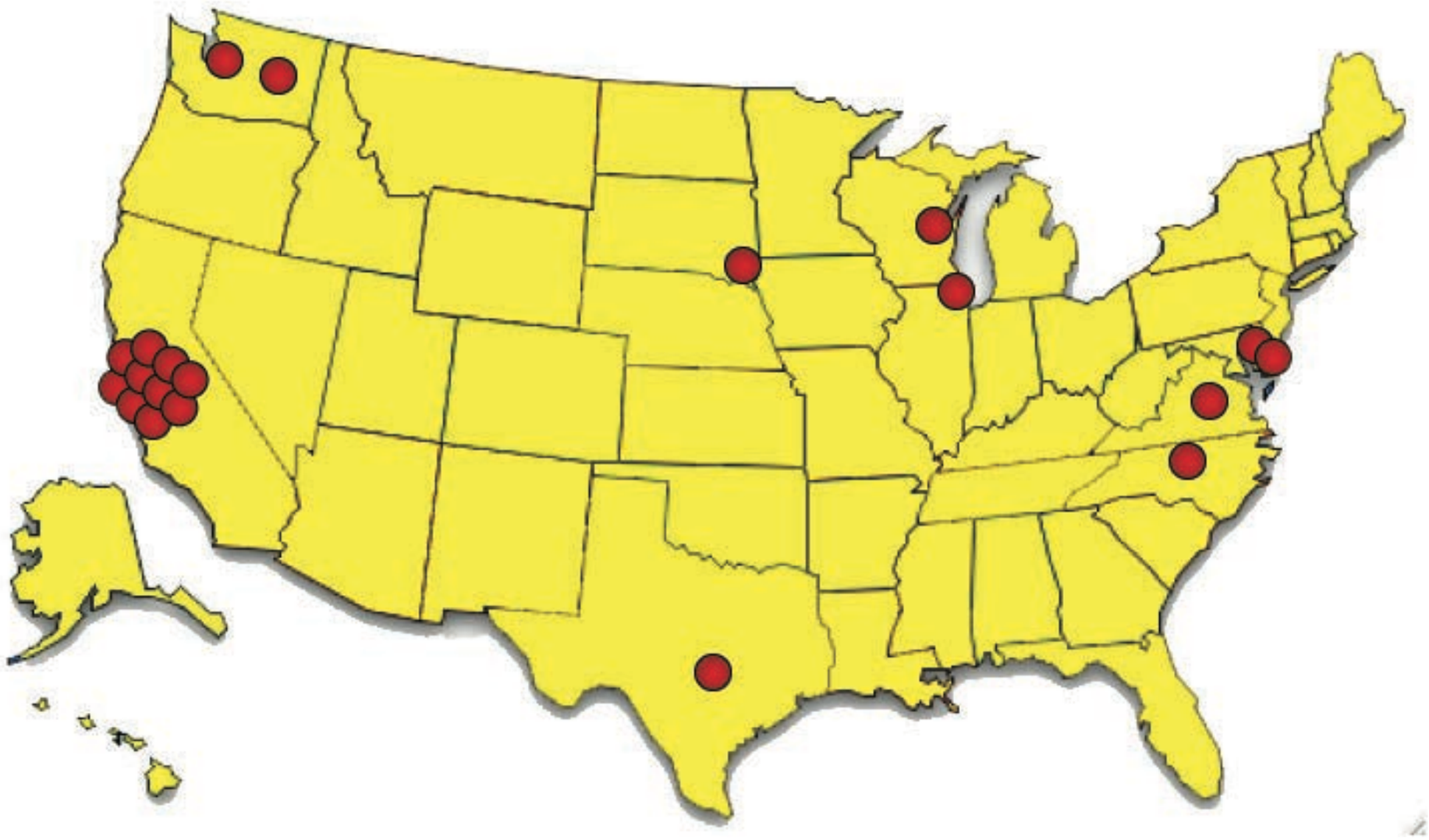
"What would it mean if the same credit card number is on the same drive?"

Cross-Drive Correlation

This turns out to be a useful intelligence technique.



I traced 20 drives back to their former owners.



Then I graduated

Science requires the *scientific process.*

Hallmarks of Science:

- Controlled and repeatable experiments
- No privileged observers

Why repeat some other scientist's experiment?

- Validate that an algorithm is properly implemented
- Determine if ***your*** new algorithm is better than ***someone else's*** old one
- (Scientific confirmation? — perhaps for venture capital firms.)

We couldn't do science with forensics in 2005

- People work with their own data
  - *—Can't distribute because of copyright & privacy issues*
- People work with "evidence"
  - *—Can't discuss due to legal sensitivities.*

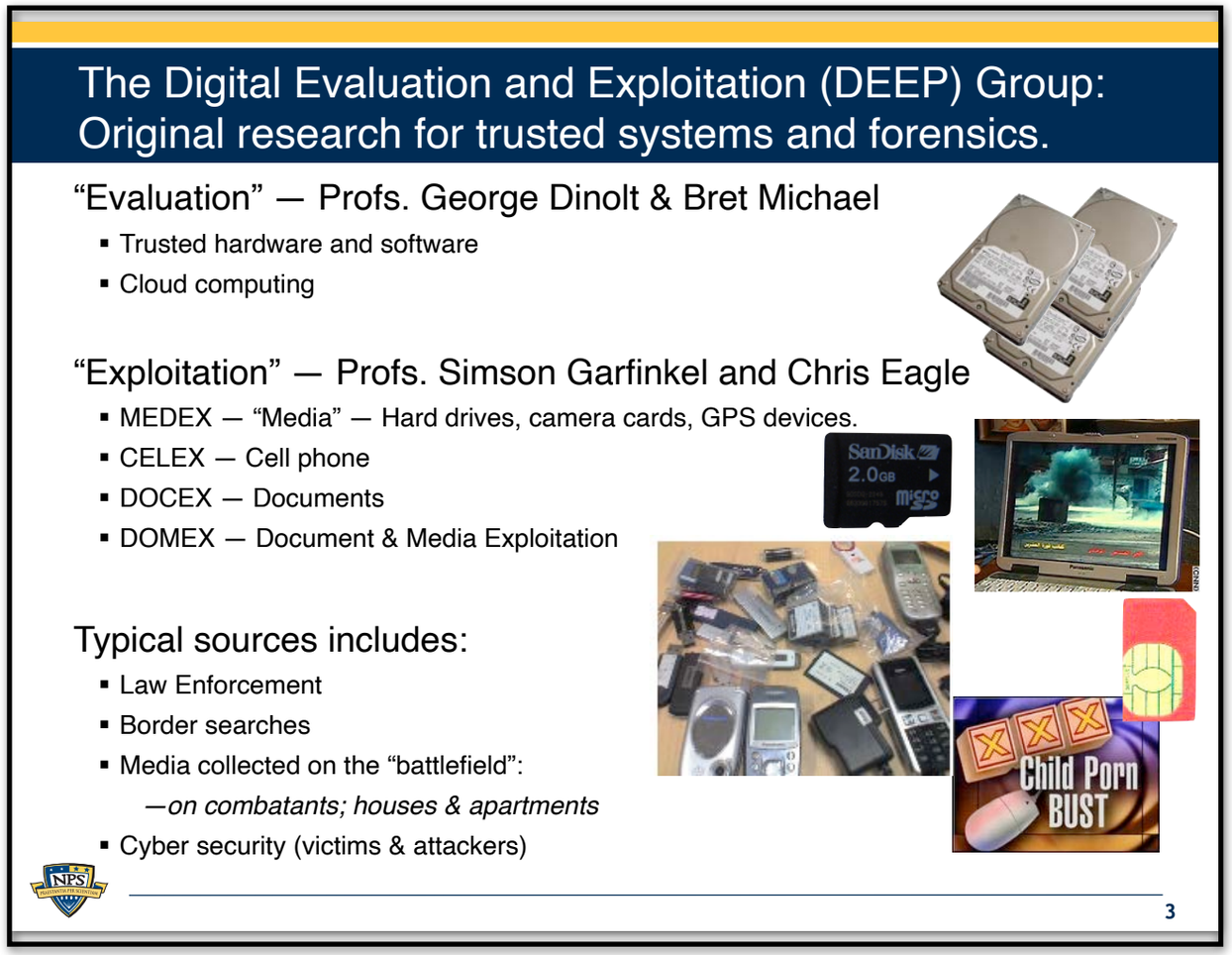
Harvard University, 2005

With a small grant from the US National Science Foundation,
I purchased 500 additional used hard drives.

Naval Postgraduate School, 2006

We continued to develop the corpus … but with some changes:

- Expanded to include USB sticks, SD Cards, and Phones
- Expanded collection world-wide
- Had a contractor collect the devices
- Removed *all* US-sourced media



2010 slide promoting DEEP

The Real Data Corpus (~70TB compressed)

- Disks, camera cards, & cell phones purchased on the secondary market
- Most contain data from previous users
- Mostly acquire outside the US:
  - —*Canada, China, England, Germany, France, India, Israel, Japan, Pakistan, Palestine, etc*
- Thousands of devices (HDs, CDs, DVDs, flash, etc.)

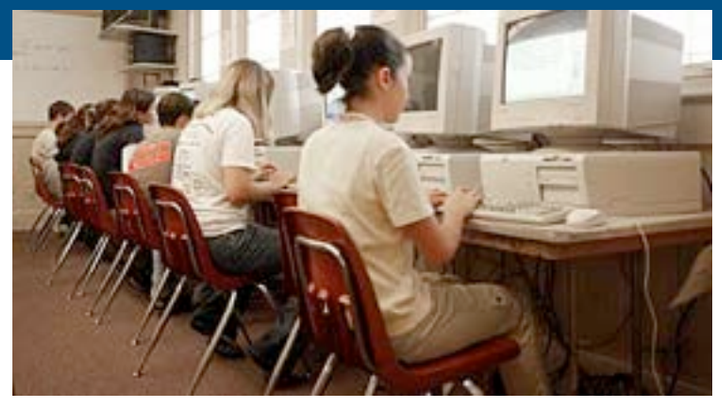
Mobile Phone Application Corpus

- Android Applications; Mobile Malware; etc

The problems we encountered obtaining, curating and exploiting this data mirror those of national organizations

—*Garfinkel, Farrell, Roussev and Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, DFRWS 2009*
http://digitalcorpora.org/

Digital Forensics education needs constructed data!



To teach forensics, we need complex data!

- Disk images
- Memory images
- Network packets

Some teachers get used hard drives from eBay

- Problem: you don't know what's on the disk
  - —*Ground Truth*
  - —*Potential for illegal Material — distributing porn to minors is illegal*

Some teachers have students examine other student machines:

- Self-examination: students know what they will find
- Examining each other's machines: potential for inappropriate disclosure

Constructed data is also important for tool testing.

We manufactured data that can be freely redistributed.

Files from US Government Web Servers (500GB)

- ≈1 million heterogeneous files
  - —*Documents (Word, Excel, PDF, etc.); Images (JPEG, PNG, etc.)*
  - —*Database Files; HTML files; Log files; XML*
- Freely redistributable; Many different file types
- This database was surprising difficulty to collect, curate, and distribute:
  - —*Scale created data collection and management problems*
  - —*Copyright, Privacy & Provenance issues*

Advantage over flickr & youtube: persistence & copyright

<abstract>NOAA's National Geophysical Data Center (NGDC) is building high-resolution digital elevation models (DEMs) for select U.S. coastal regions. … </abstract>

<abstract>This data set contains data for birds caught with mistnets and with other means for sampling Avian Influenza (AI)….</abstract>

We also developed complex constructed data.

Test and Realistic Disk Images (1TB)

- Mostly Windows operating system
- Some with complex scenarios to facilitate forensics education

University harassment scenario

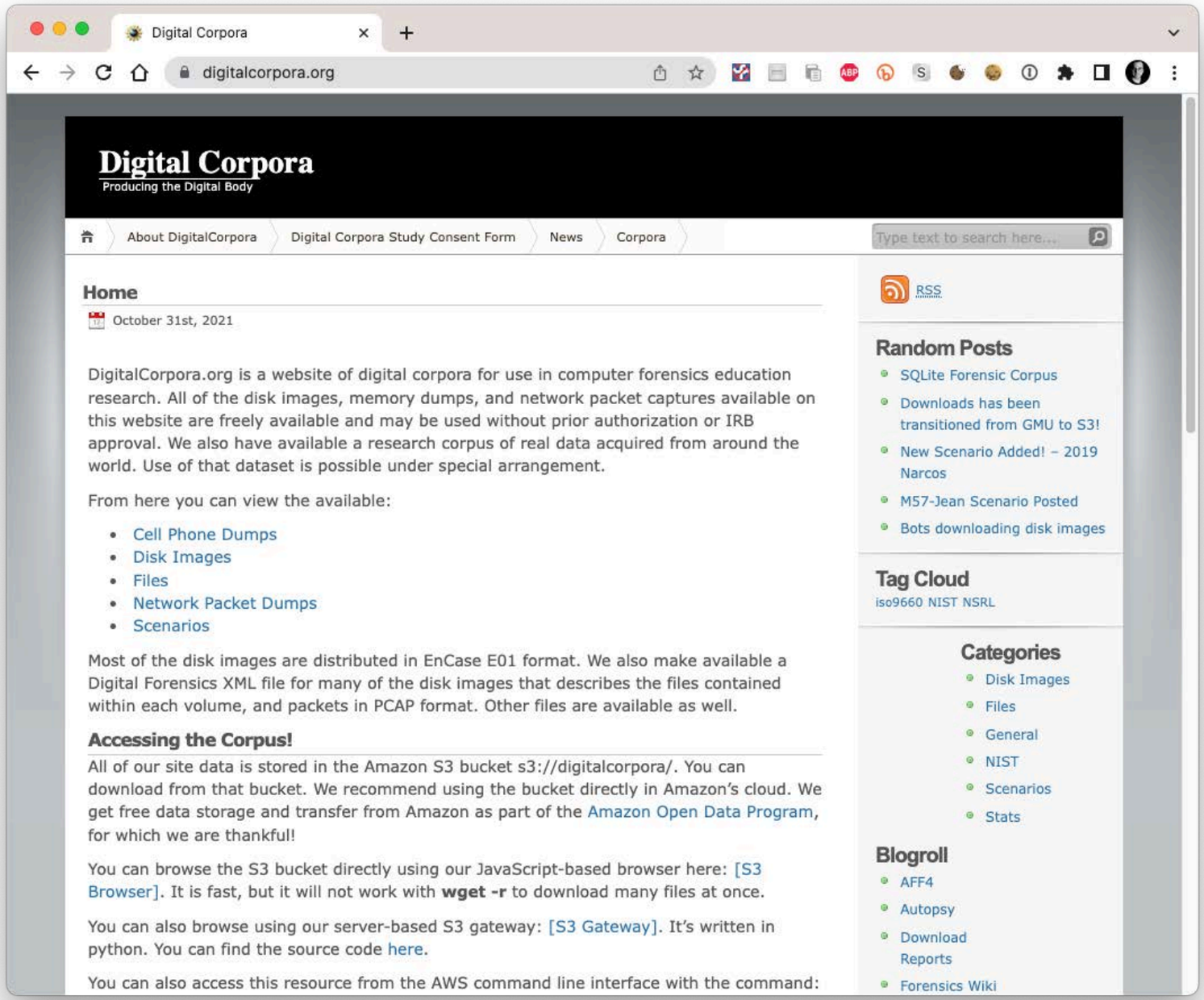
- Network forensics — browser fingerprinting, reverse NAT, target identification
- 50MB of packets

Company data theft & child pornography scenario

- Multi-drive correction
- Hypothesis formation
- Timeline reconstruction
  - —*Disk images, Memory Dumps, Network Packets*

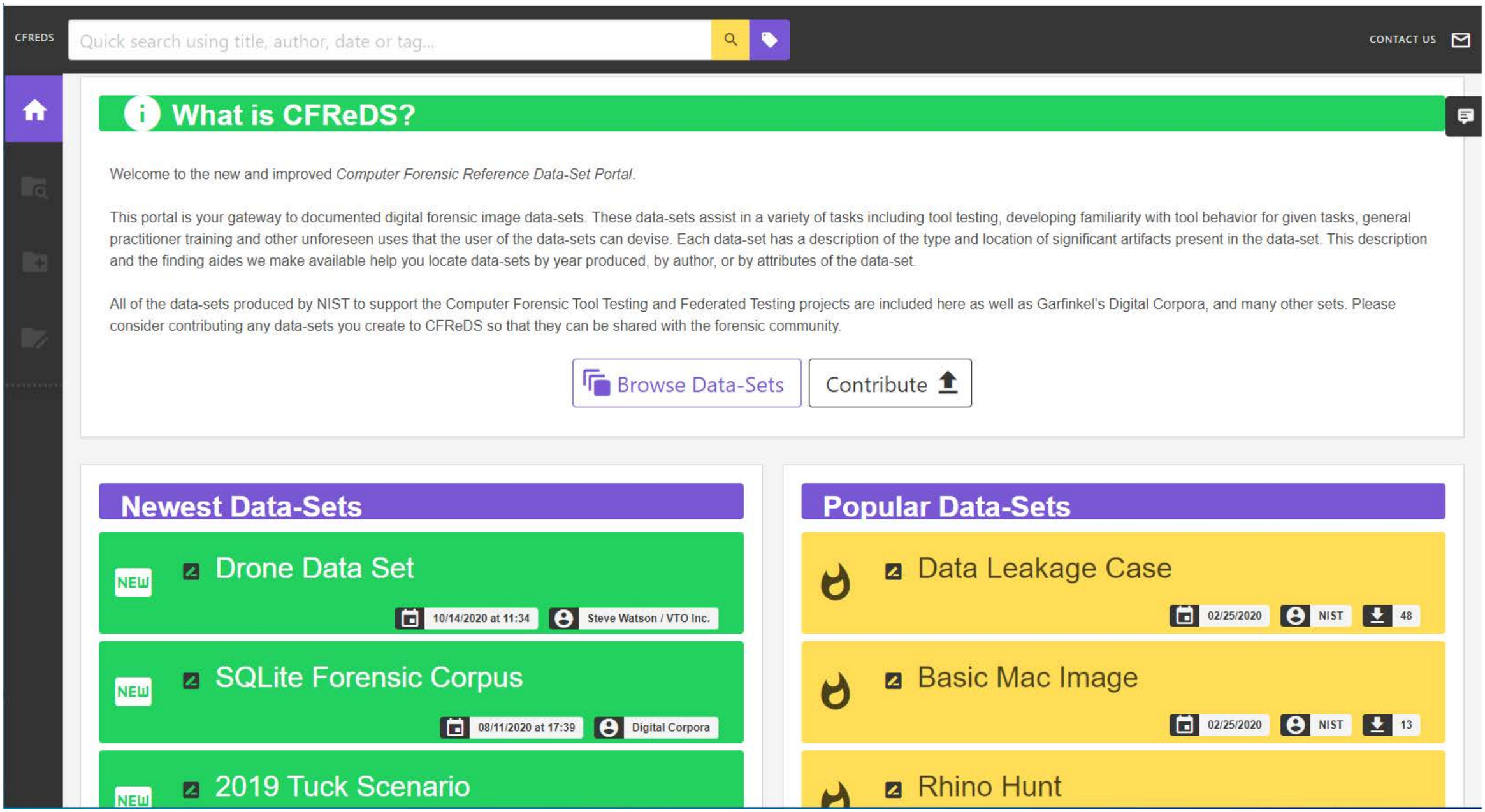
Today, the Digital Corpora project has 13 well-developed scenarios

2008-m57-jean
2008-nitroba
2009-m57-patents
2011-nps-1weapondeletion
2011-nps-2weapons
2011-nps-4drugtraffic
2011-nps-5control
2012-ngdc
2018-lonewolf
2019-narcos
2019-owl
2019-tuck
2020-linux-threat-analysis



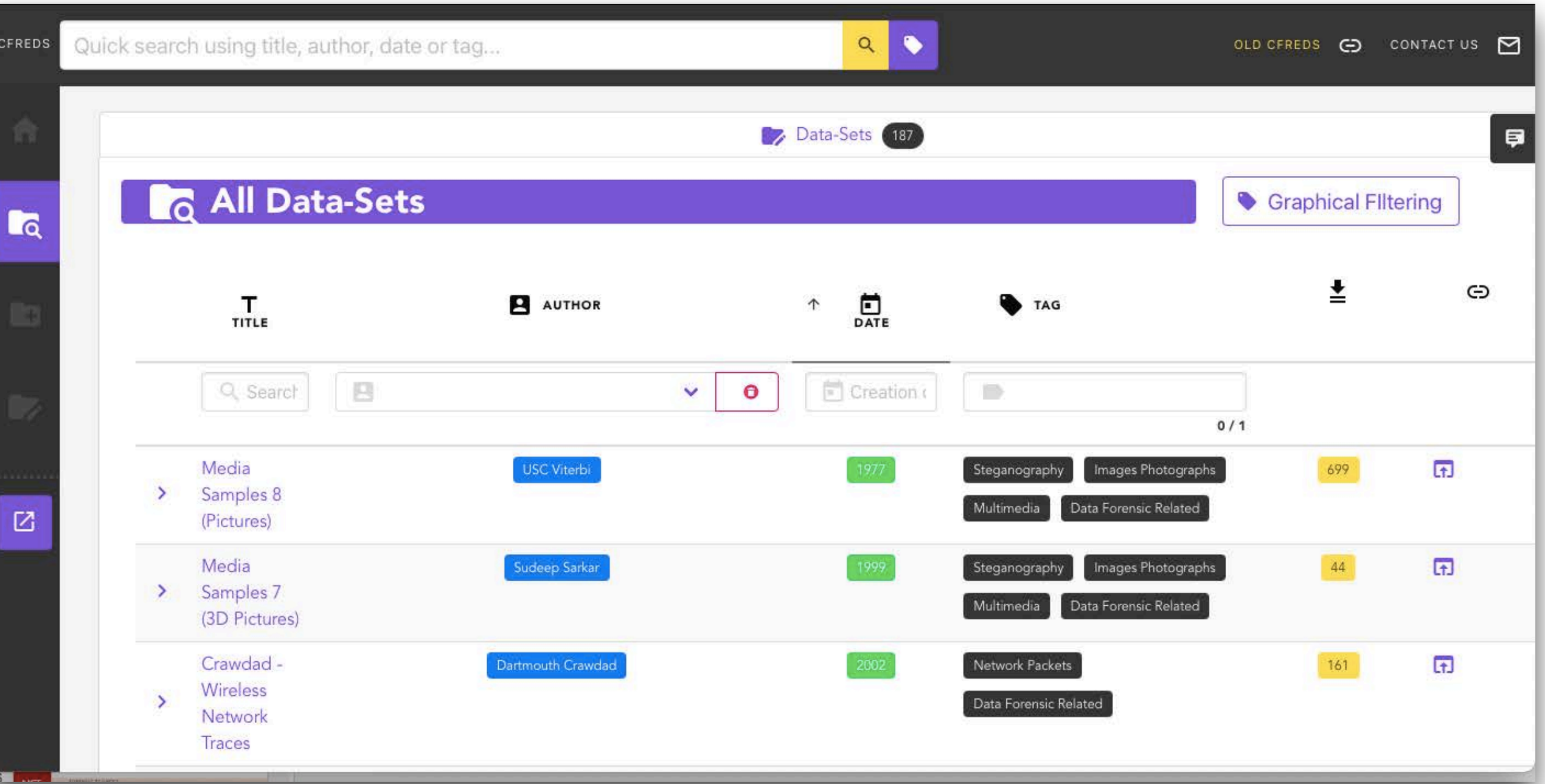
Digital Corpora is also listed in the NIST Computer Forensics Reference Datasets (CFReDS)

https://cfreds.nist.gov/



CFReDS home page

CFReDS: Around 200 datasets



Over 15,000 downloads so far

bulk_extractor (2005-2014, 2019-)

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

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.

Advantages: Speed and Flexibility

Fastest possible transfer from target device

- All access is sequential with optimal buffer size
- No disk seeking on HDs —very important before the deployment of SSDs

Embarrassingly parallel

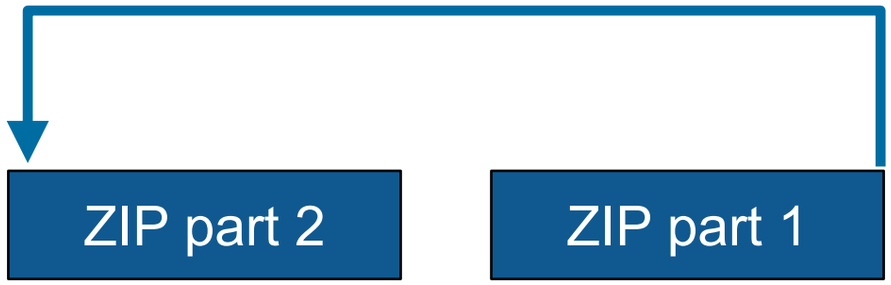
- Easy to parallelize across CPU cores and multiple boxes
- In theory can process an entire drive in 5 minutes

Potential for intermediate answers

Reads all the data — allocated files, deleted files, file fragments

- Separate metadata extraction required to get the file names.

Primary Disadvantage on file systems: Completeness



Fragmented files won't be reconstructed:

- Compressed files with part2-part1 ordering (possibly .docx)
- Files with internal fragmentation (.doc but not .docx)

Fortunately, most files are *not* fragmented

- Individual components of a ZIP file can be fragmented

Most files that *are* fragmented have internal structure that can be carved

- Log files, Outlook PST files, etc

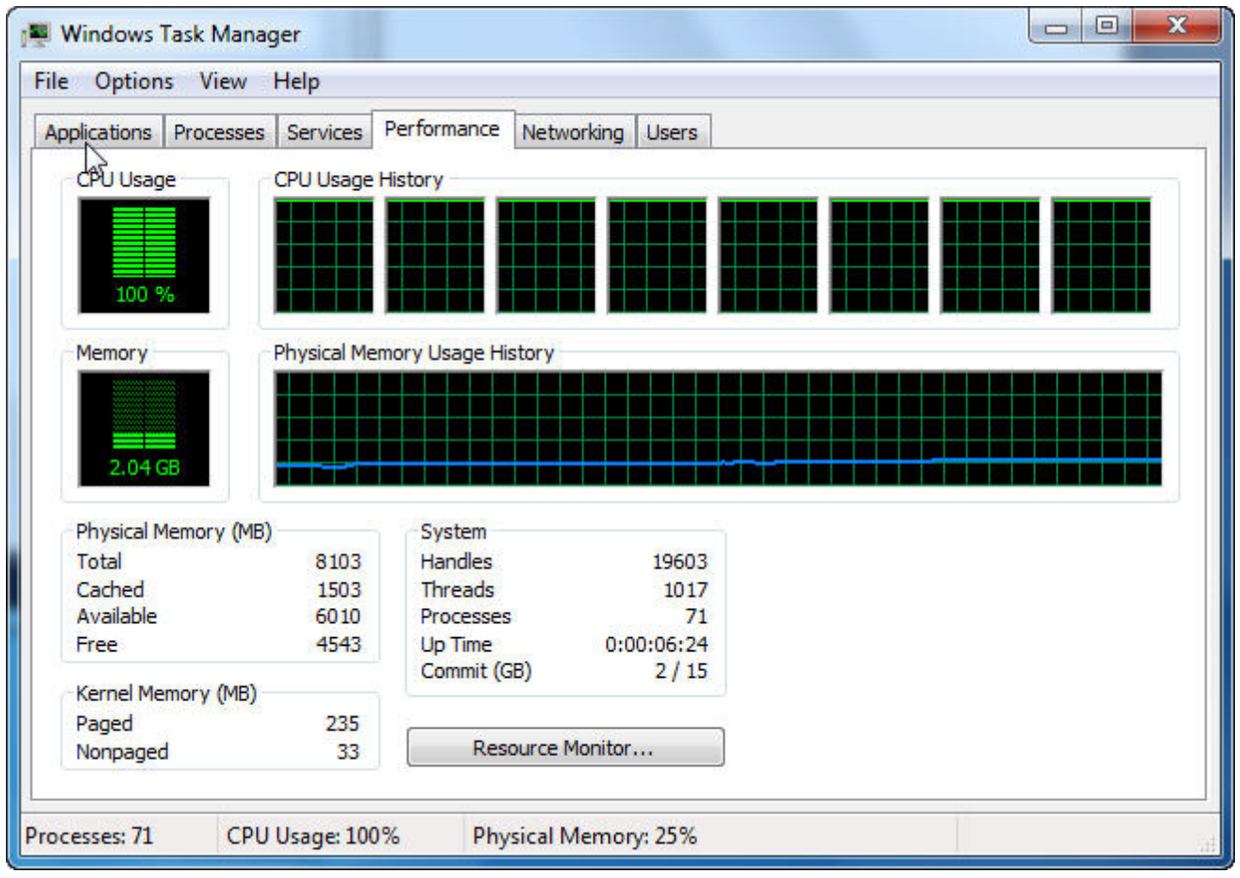
bulk_extractor is a powerful stream-based forensic tool.

Bulk_extractor demonstrates the power of:

- Bulk data processing
- Carving EVERYTHING
- Multi-threading (we can process data with 100% CPU utilization)

Bulk_extractor is 100% free software

- Public Domain (work of US Government)
- Designed to promote other ideas:
  - —*DFXML*
  - —*Job Distribution*
  - —*Forensic Path*
  - —*SBUF*



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

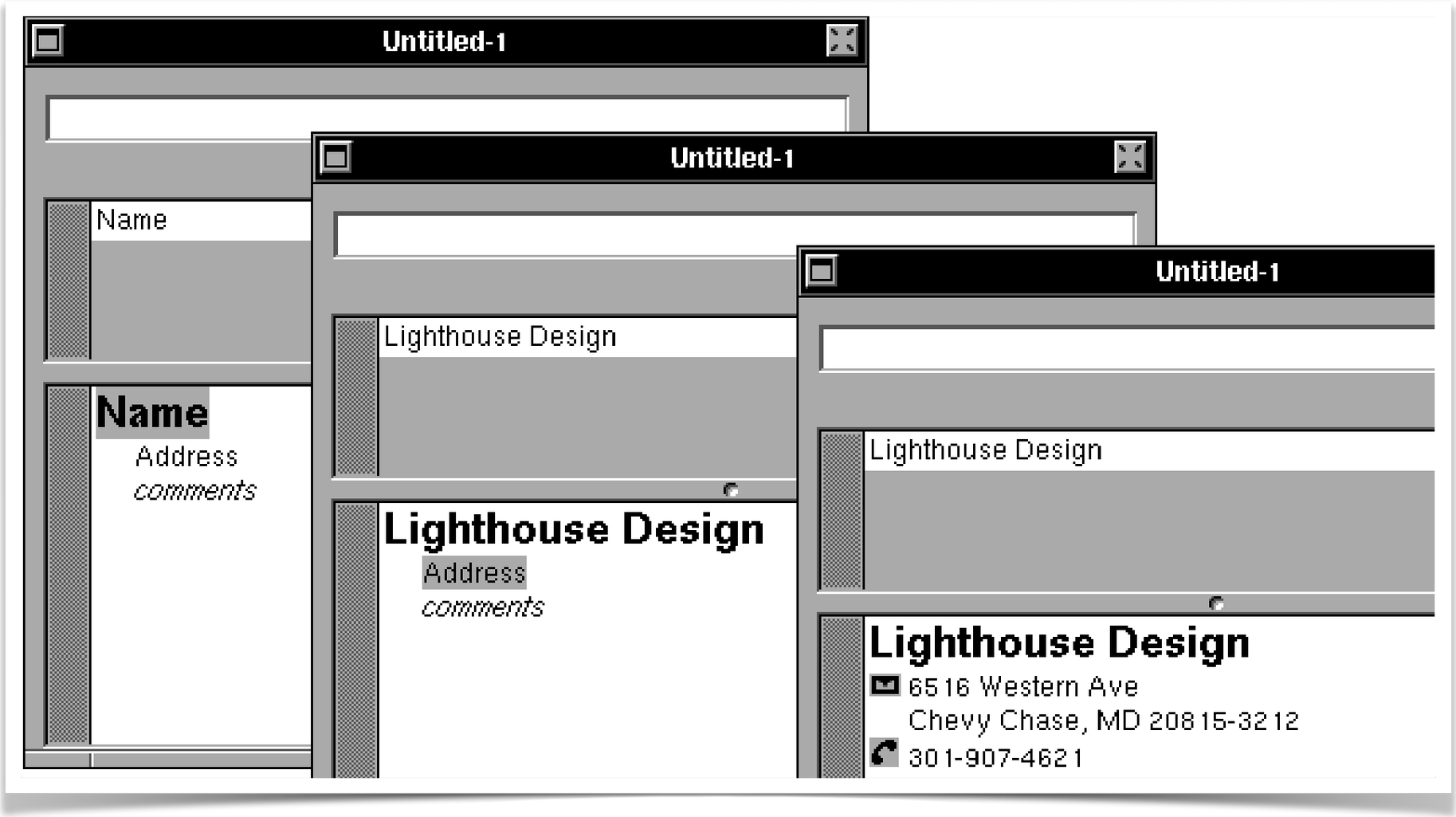
- “Optimistically” decompresses and re-analyzes all data
- Finds data in browser caches (downloaded with zip/gzip), and in many file formats

Presents the data in an easy-to-understand report

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

bulk_extractor: 20 years in the making!

In 1991 I developed SBook, a free-format address book

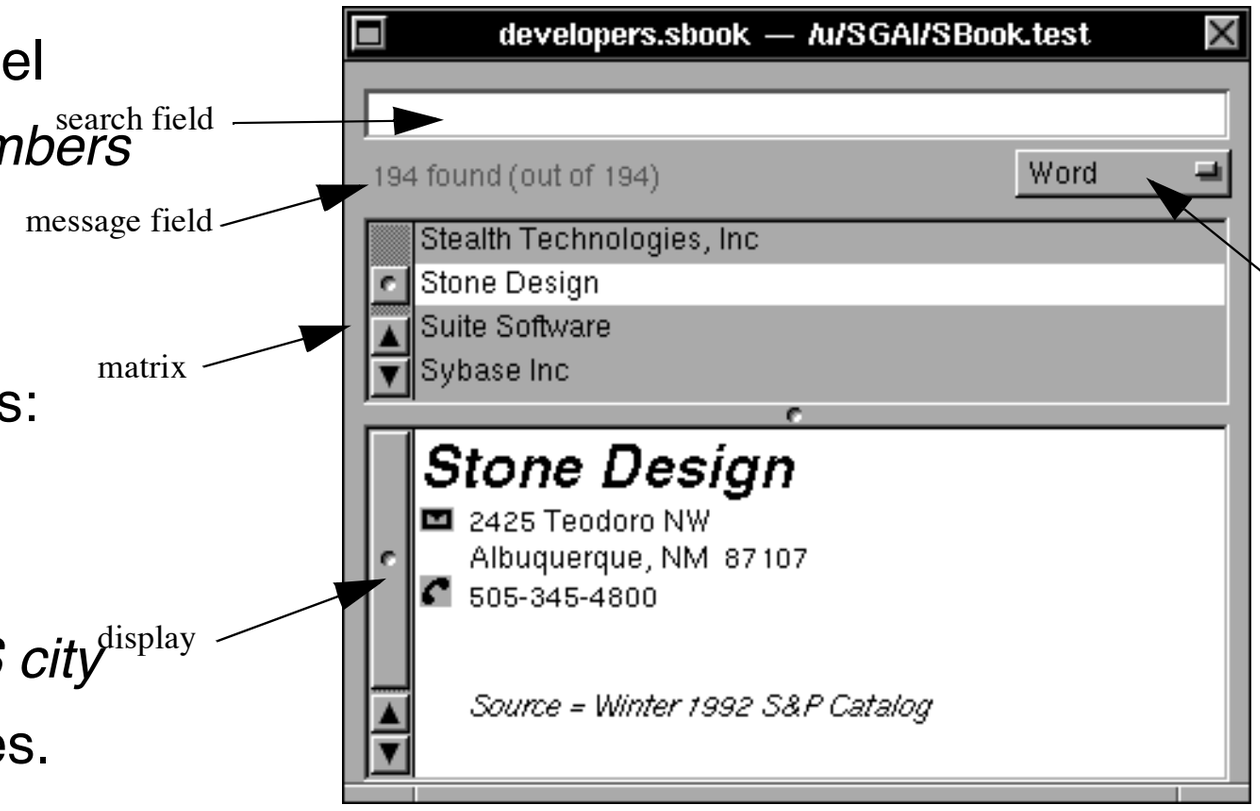


SBook used “Named Entity Recognition” to find addresses, phone numbers, email addresses *while you typed*

Today we call this technology Named Entity Recognition

SBook's technology was based on:

- Regular expressions executed in parallel
  - —*US, European, & Asian Phone Numbers*
  - —*Email Addresses*
  - —*URLs*
- A gazette with more than 10,000 names:
  - —*Common "Company" names*
  - —*Common "Person" names*
  - —*Every country, state, and major US city*
- Hand-tuned weights and additional rules.



Implementation:

- 2500 lines of GNU flex, C++
- 50 msec to evaluate 20 lines of ASCII text
  - —*Running on a 25Mhz 68030 with 32MB of RAM!*

Recall that I had 200 hard drives to analyze.

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!*

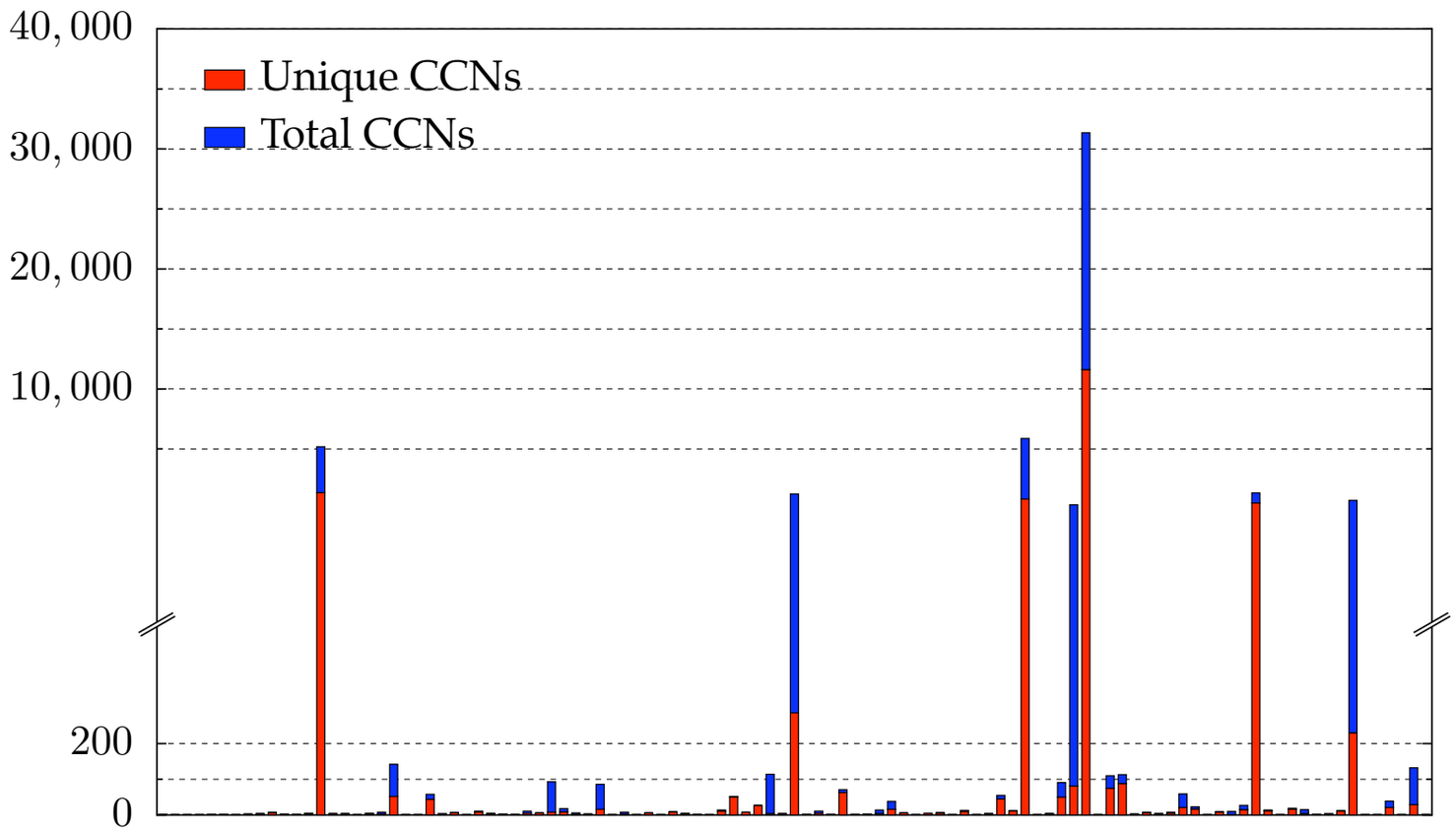
Second strategy:

- Use SBook technology!
- 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:

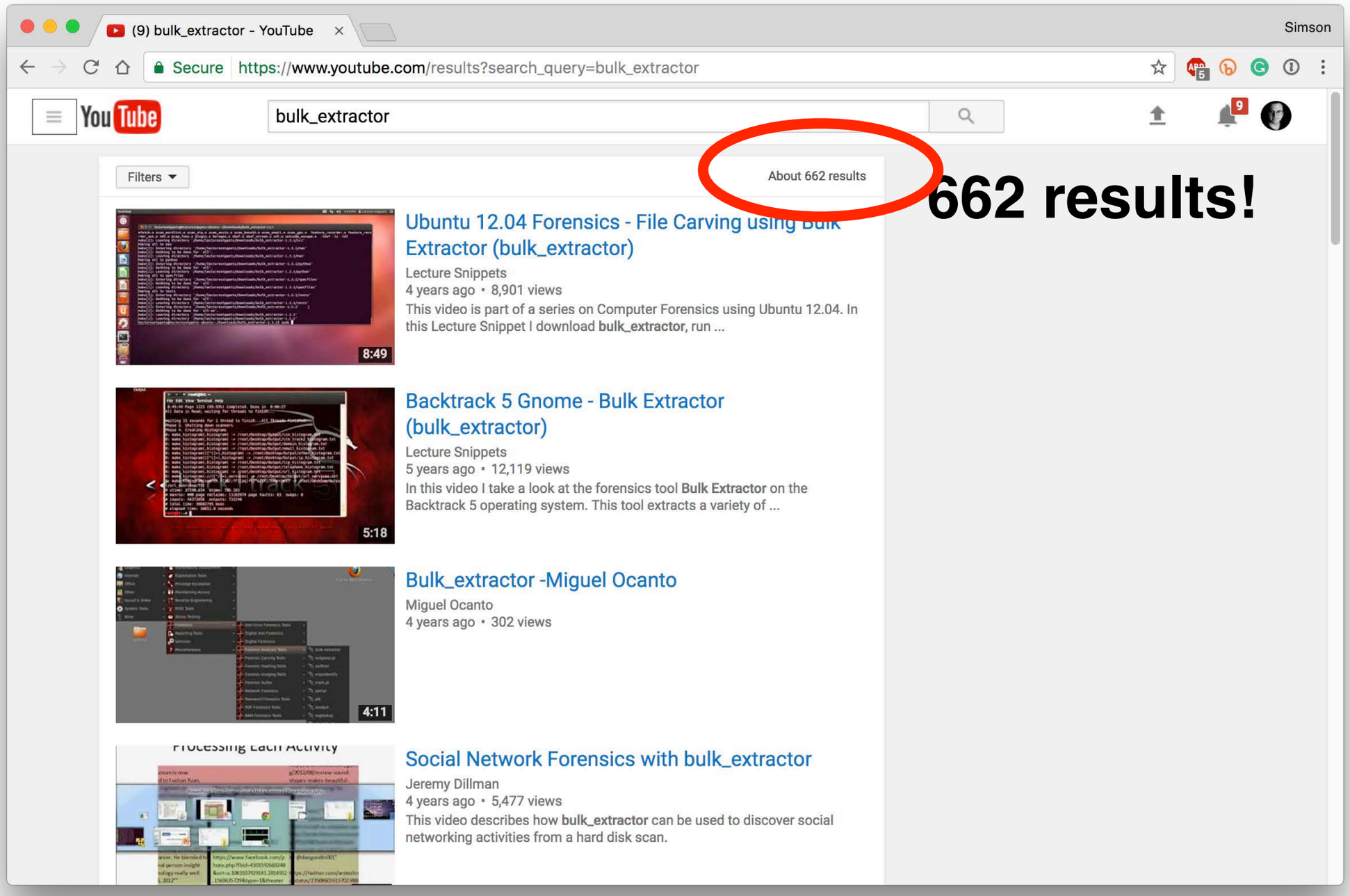
- Additional validation beyond regular expressions (CCN Luhn algorithm, etc)
- Examination of feature “neighborhood” to eliminate common false positives



The technique worked well to find drives with sensitive information

But it didn’t scale.

$15M and 8 years later



So how did we get there?

I interviewed law enforcement regarding their use of forensic tools (2005-2008)

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)

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.

Jan 2008	— Created Subversion Repository
April 2010	— Initial public release - 0.1.0
May 2010	— Initial multi-threading release - 0.3.0
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
April 2012	— Move to git repo

Tool capabilities result from substantial testing and user feedback

Moving technology from the lab to the field has been 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.

Success Story #1: Credit Card Fraud

SLO 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.

Success Story #2: ATM Fraud

A 250GB disk drive was recovered from individuals suspected of setting a credit-card “skimmer” and pinhole camera at ATM machines in a major US city.

Police needed to rapidly supply the banks with a list of the compromised credit card numbers so that the accounts could be shut down.

bulk_extractor completed its processing after just two hours on a quad-core computer.

- The banks in question were provided with ccns.txt* output file
- Cards were canceled and customers were contacted.

* A list of credit-card numbers found on the drive. The actual files containing the data were later identified by using the file offsets present in the feature file.

Creating bulk_extractor was an iterative approach

I watched analysts working on cases and asked them:

- What information would be useful for solving this case?
- What kinds of information are you looking for?

I should have started with the “Heilmeier Questions:”*

- What are you trying to do? Articulate your objectives using absolutely no jargon
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?
- What are the risks?
- How much will it cost?
- How long will it take?
- What are the mid-term and final “exams” to check for success?

* https://www.darpa.mil/work-with-us/heilmeier-catechism

Inside bulk_extractor

bulk_extractor: architectural overview

Written in C, C++ and GNU flex

- Command-line tool
- Linux, MacOS, Windows (compiled with mingw)

Key features:

- “Scanners” look for information of interest in typical investigations
- Recursively re-analyzes compressed data
- Results stored in “feature files”
- Multi-threaded

Java GUI

- Runs command-line tool and views results

bulk_extractor extracts “features” from disk images.

http://www.nps.edu/

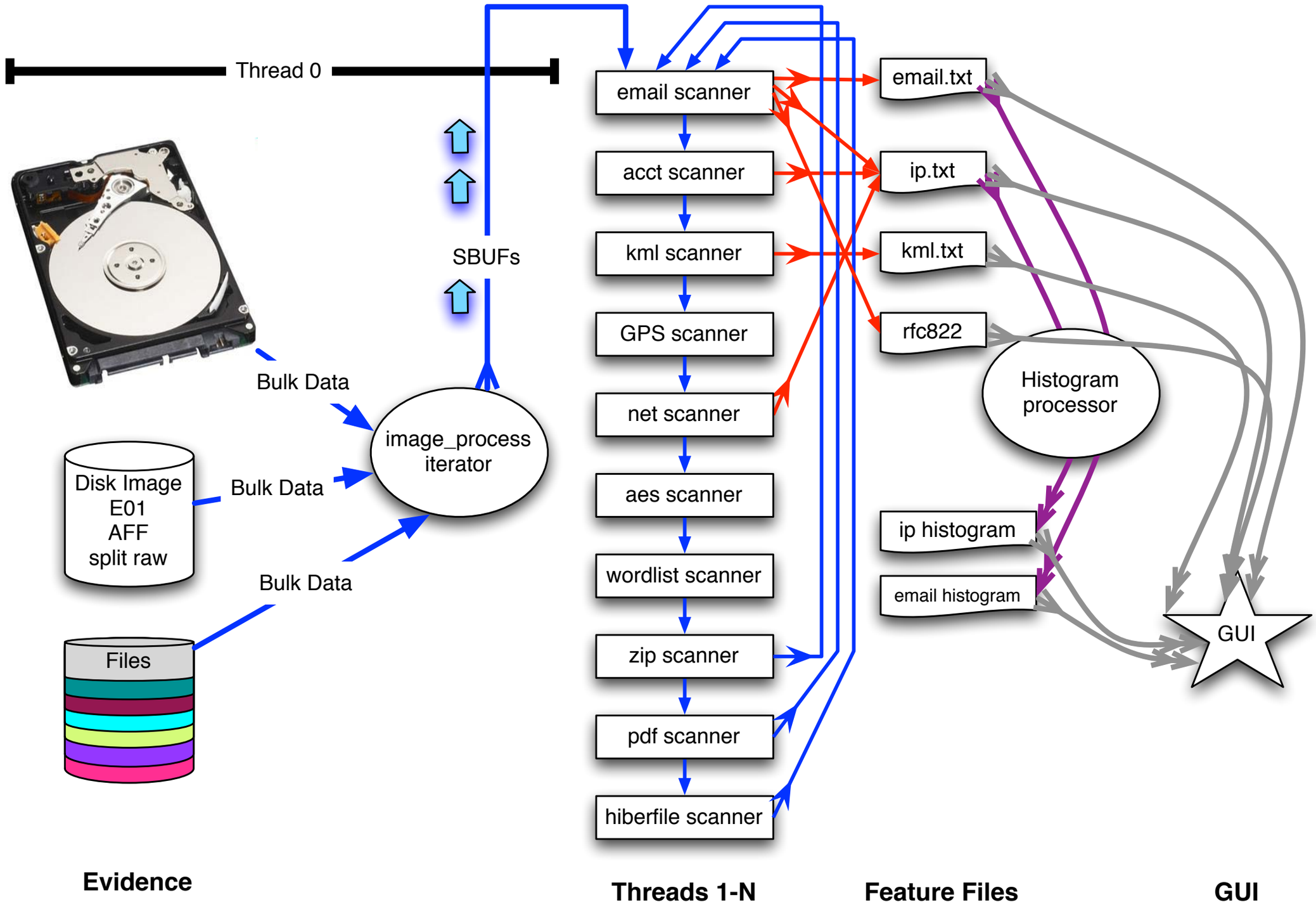
202-555-1212
user@domain.com

202-555-1212

http://www.nps.edu/

user@domain.com

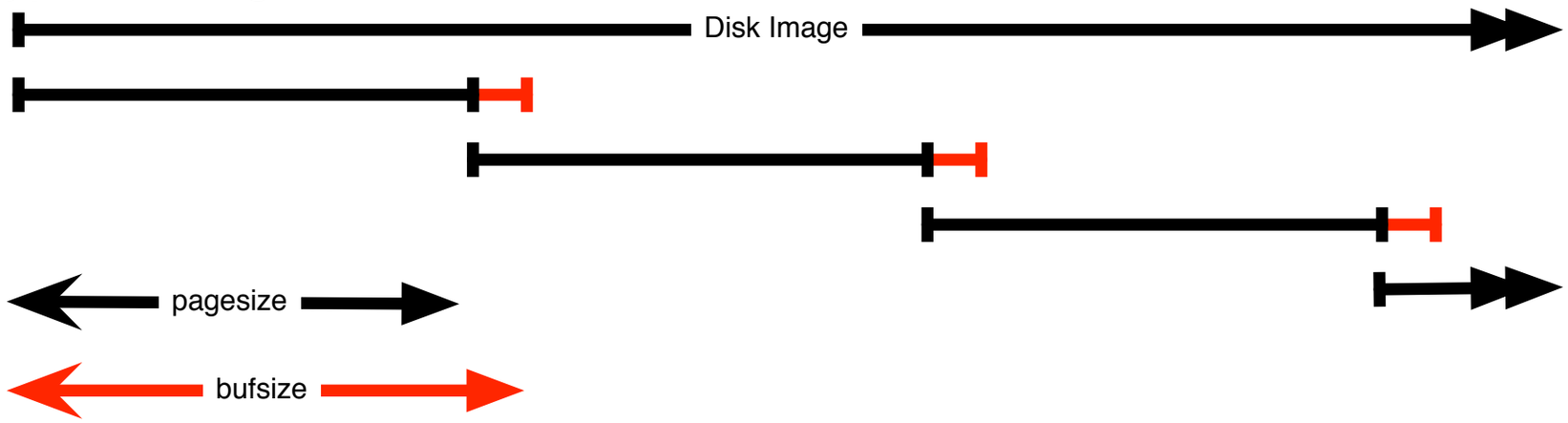
bulk_extractor: system diagram



The “pages” overlap to avoid dropping features that cross buffer boundaries.

The overlap area is called the *margin*

- Each sbuf can be processed in parallel — they don’t depend on each other
- Features start in the page but end in the margin are *reported*
- Features that start in the margin are *ignored* (we get them later)
  - —*Assumes that the feature size is smaller than the margin size*
  - —*Typical margin: 1MB*



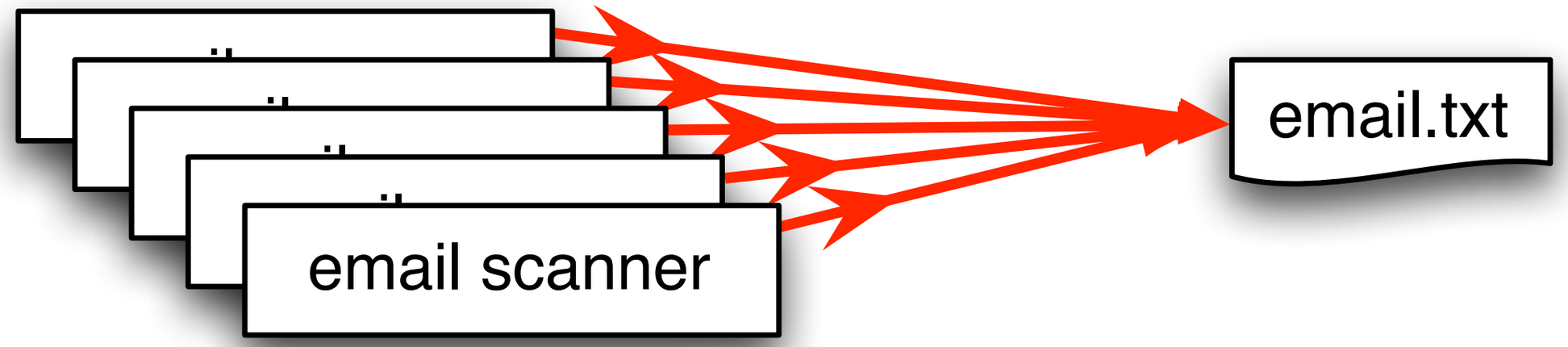
Entire system is automatic:

- Image_process iterator makes **sbuf_t** buffers
- Each buffer is processed by every scanner
- Features are automatically combined.

The *feature recording system* saves features to disk.

Feature Recorder objects store the features

- Scanners are given a (feature_recorder *) pointer
- Feature recorders are *thread safe*



Features are stored in a *feature file:*

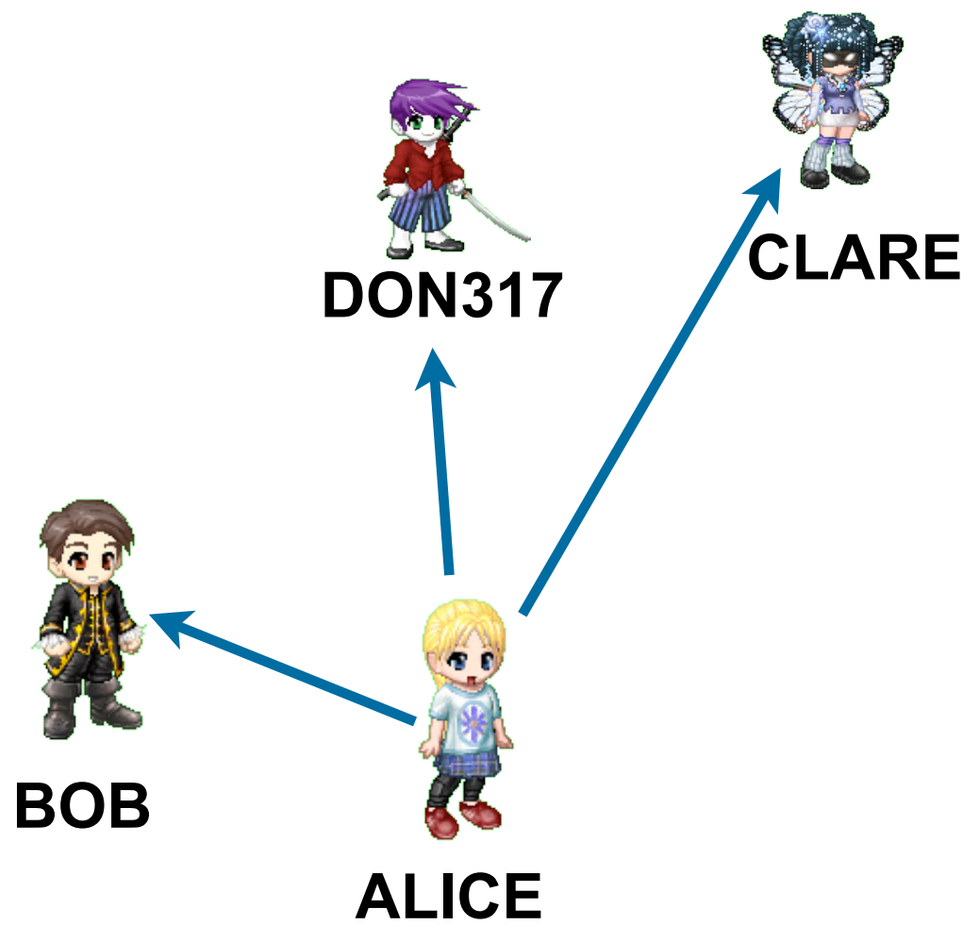
```
48198832  domexuser2@gmail.com    tocol>____<name>domexuser2@gmail.com/Home</name>____
48200361  domexuser2@live.com     tocol>____<name>domexuser2@live.com</name>____<pass
48413829  siege@preoccupied.net   siege) O'Brien <siege@preoccupied.net>_hp://meanwhi
48481542  danilo@gnome.org        Danilo __egan <danilo@gnome.org>_Language-Team:
48481589  gnom@prevod.org         : Serbian (sr) <gnom@prevod.org>_MIME-Version:
49421069  domexuser1@gmail.com    server2.name", "domexuser1@gmail.com");__user_pref("
49421279  domexuser1@gmail.com    er2.userName", "domexuser1@gmail.com");__user_pref("
49421608  domexuser1@gmail.com    tp1.username", "domexuser1@gmail.com");__user_pref("
```

offset **feature** **feature in evidence context**

Histograms are a powerful tool for understanding evidence.

Email histogram allows us to rapidly determine:

- Drive's primary user
- User's organization
- Primary correspondents
- Other email addresses



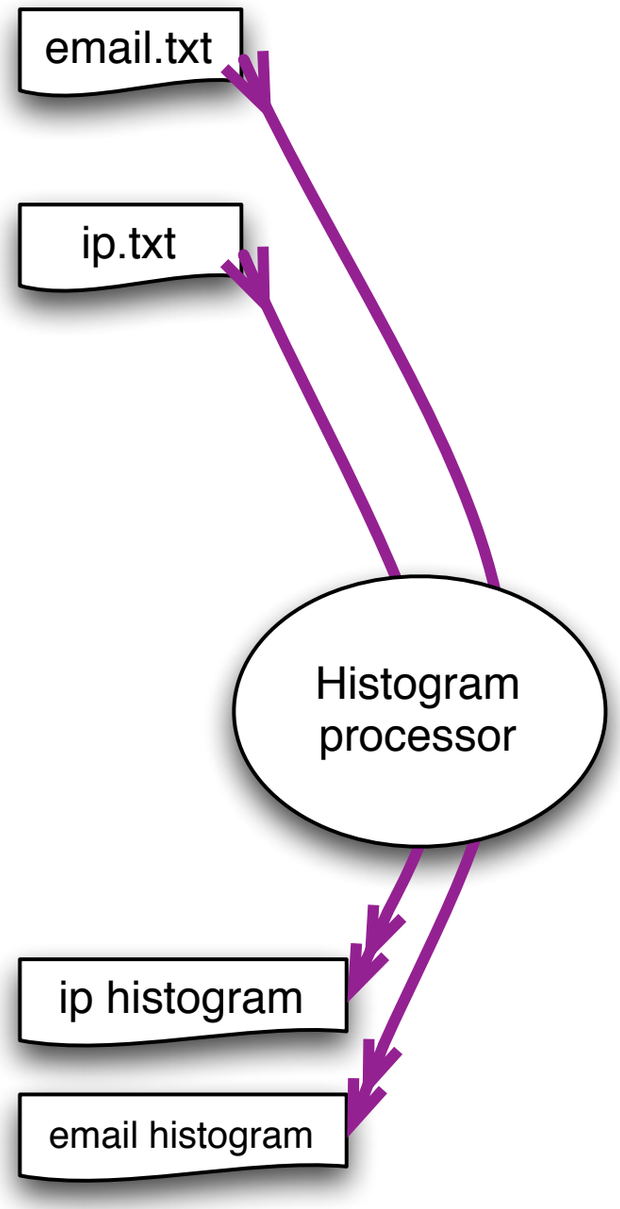
Drive #51 (Anonymized)

ALICE@DOMAIN1.com	8133
BOB@DOMAIN1.com	3504
ALICE@mail.adhost.com	2956
JobInfo@alumni-gsb.stanford.edu	2108
CLARE@aol.com	1579
DON317@earthlink.net	1206
ERIC@DOMAIN1.com	1118
GABBY10@aol.com	1030
HAROLD@HAROLD.com	989
ISHMAEL@JACK.wolfe.net	960
KIM@prodigy.net	947
ISHMAEL-list@rcia.com	845
JACK@nwlink.com	802
LEN@wolfenet.com	790
natcom-list@rcia.com	763

The feature recording system *automatically* makes histograms.

Simple histogram based on feature:

```
n=579    domexuser1@gmail.com
n=432    domexuser2@gmail.com
n=340    domexuser3@gmail.com
n=268    ips@mail.ips.es
n=252    premium-server@thawte.com
n=244    CPS-requests@verisign.com
n=242    someone@example.com
```



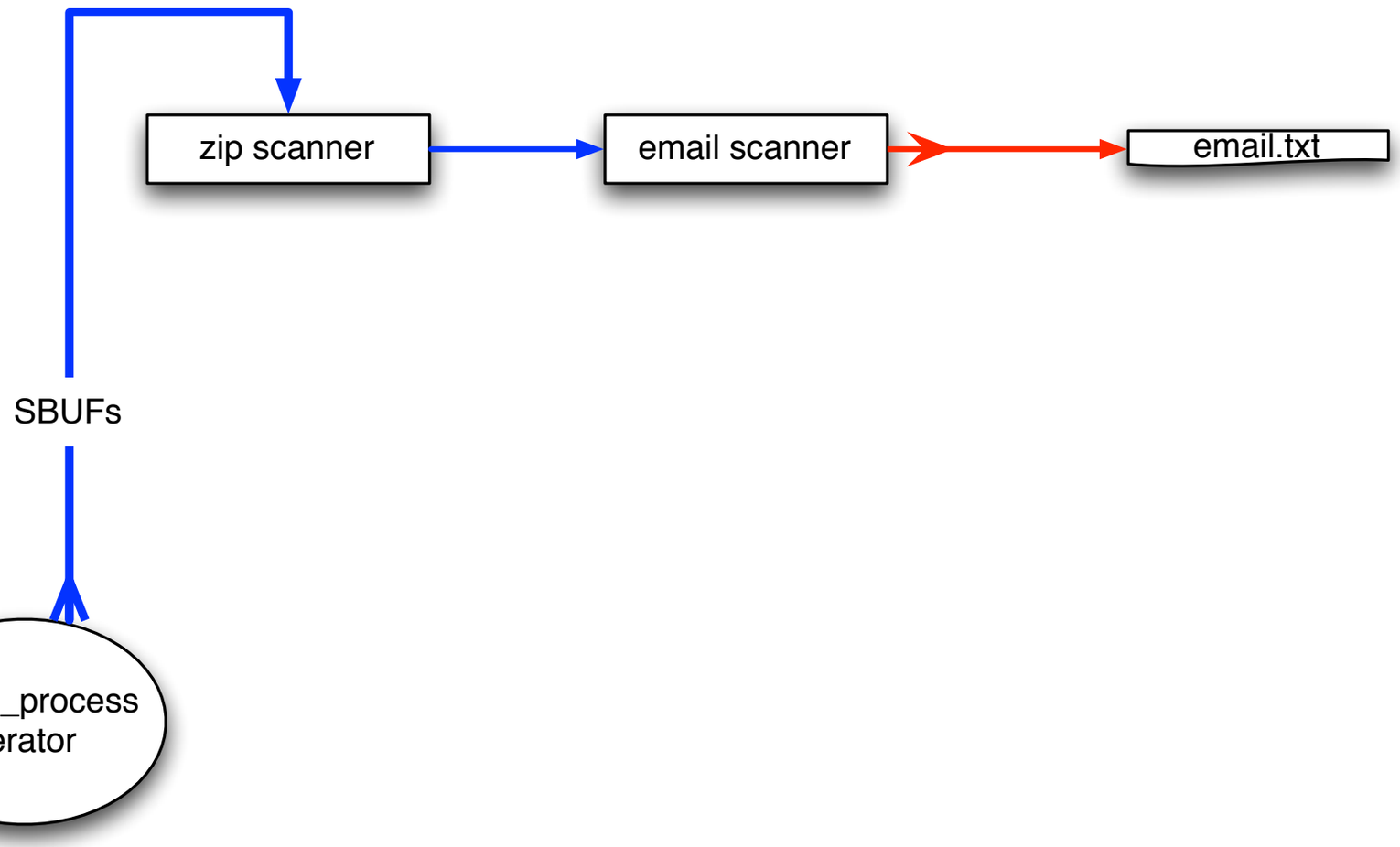
Based on regular expression extraction:

- For example, extract search terms with **.*search.*q=(.*)**

```
n=18     pidgin
n=10     hotmail+thunderbird
n=3      Grey+Gardens+cousins
n=3      dvd
n=2      %TERMS%
n=2      cache:
n=2      p
n=2      pi
n=2      pid
n=1      Abolish+income+tax
n=1      Brad+and+Angelina+nanny+help
n=1      Build+Windmill
n=1      Carol+Alt
```

Recursion requires a *new way* to describe offsets. bulk_extractor introduces the "forensic path."

Consider an HTTP stream that contains a GZIP-compressed email:

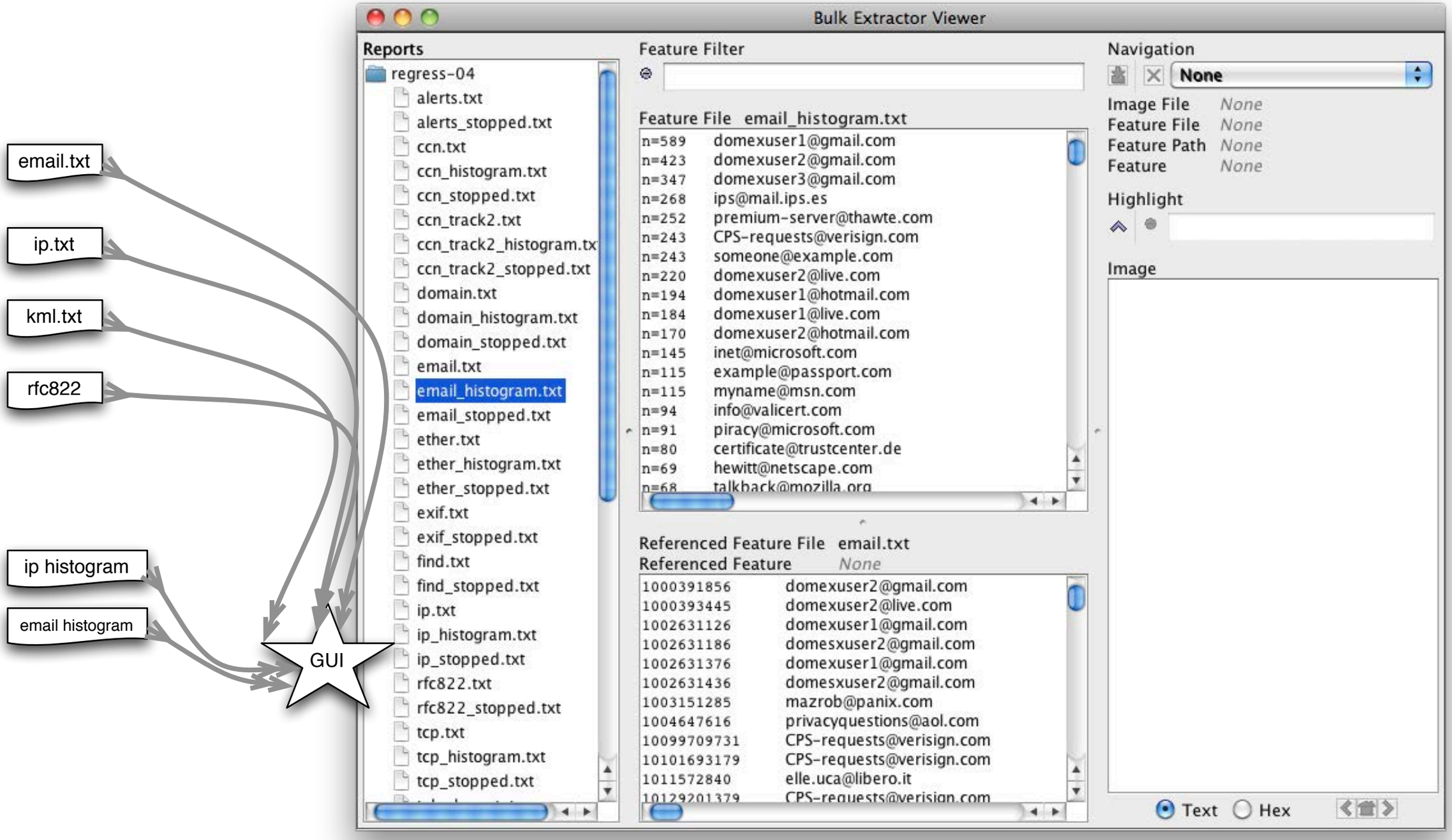


We can represent this as:

```
11052168704-GZIP-3437   live.com    eMn='domexuser1@live.com';var srf_sDispM
11052168704-GZIP-3475   live.com    pMn='domexuser1@live.com';var srf_sPreCk
11052168704-GZIP-3512   live.com    eCk='domexuser1@live.com';var srf_sFT='<
```

GUI: 100% Java
Launches bulk_extractor; views results

Uses bulk_extractor to decode forensic path



Crash protection provides for easy recovery by the user.

Every forensic tool crashes

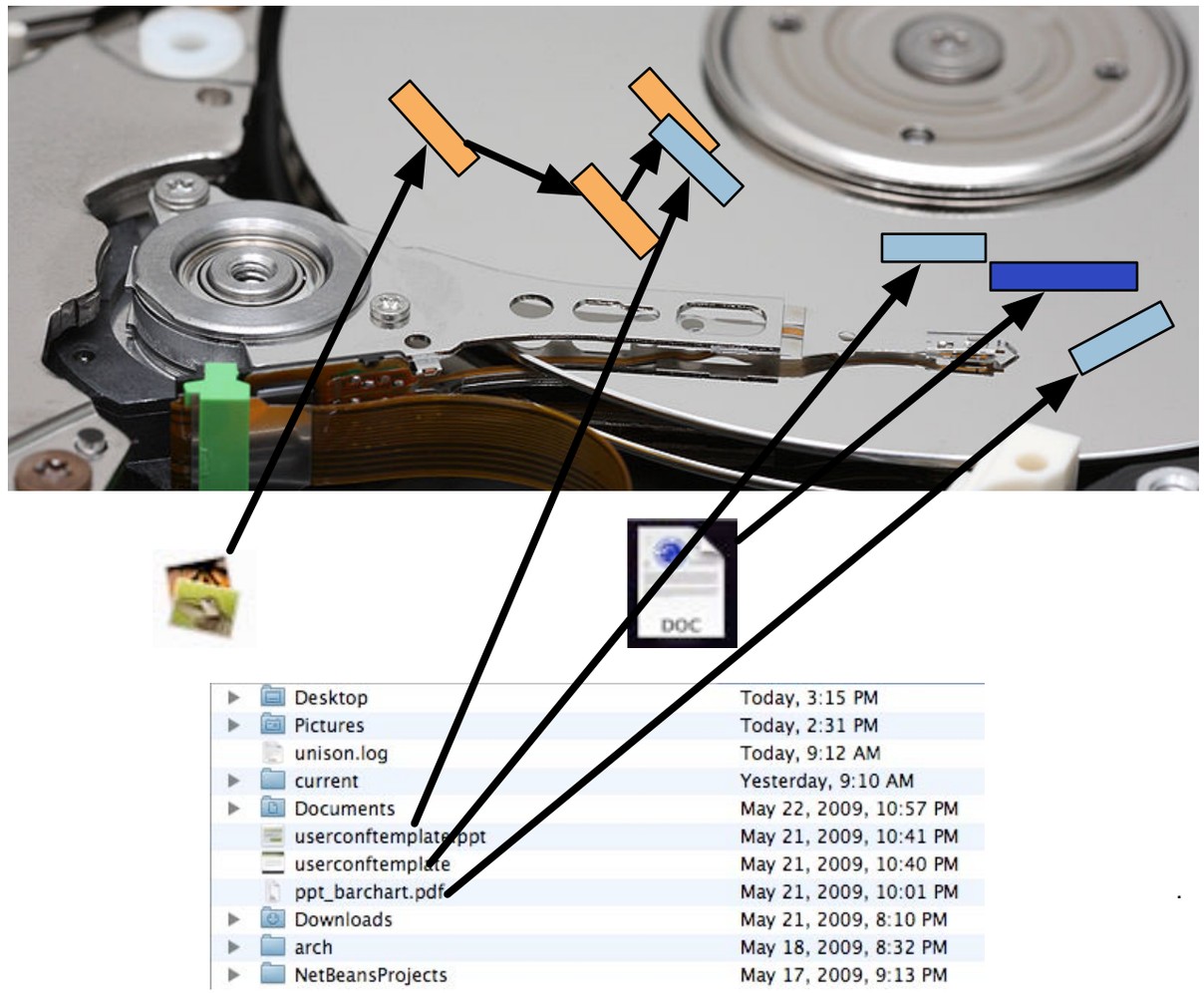
- Tools routinely used with data fragments, non-standard codings, etc
- Evidence that makes the tool crash typically cannot be shared with the developer

Crash Protection: checkpointing!

- Bulk_extractor checkpoints current page in the file config.cfg
- After a crash, just hit up-arrow and return; bulk_extractor restarts at next page.

Filenames can be added through post-processing.

bulk_extractor reports the *disk blocks* for each feature



To get the file names, you need to map the disk block to a file

- Make a map of the blocks in DFXML with **fiwalk** (http://afflib.org/fiwalk)
- Then use **python/identify_filenames.py** to create an *annotated feature file.*

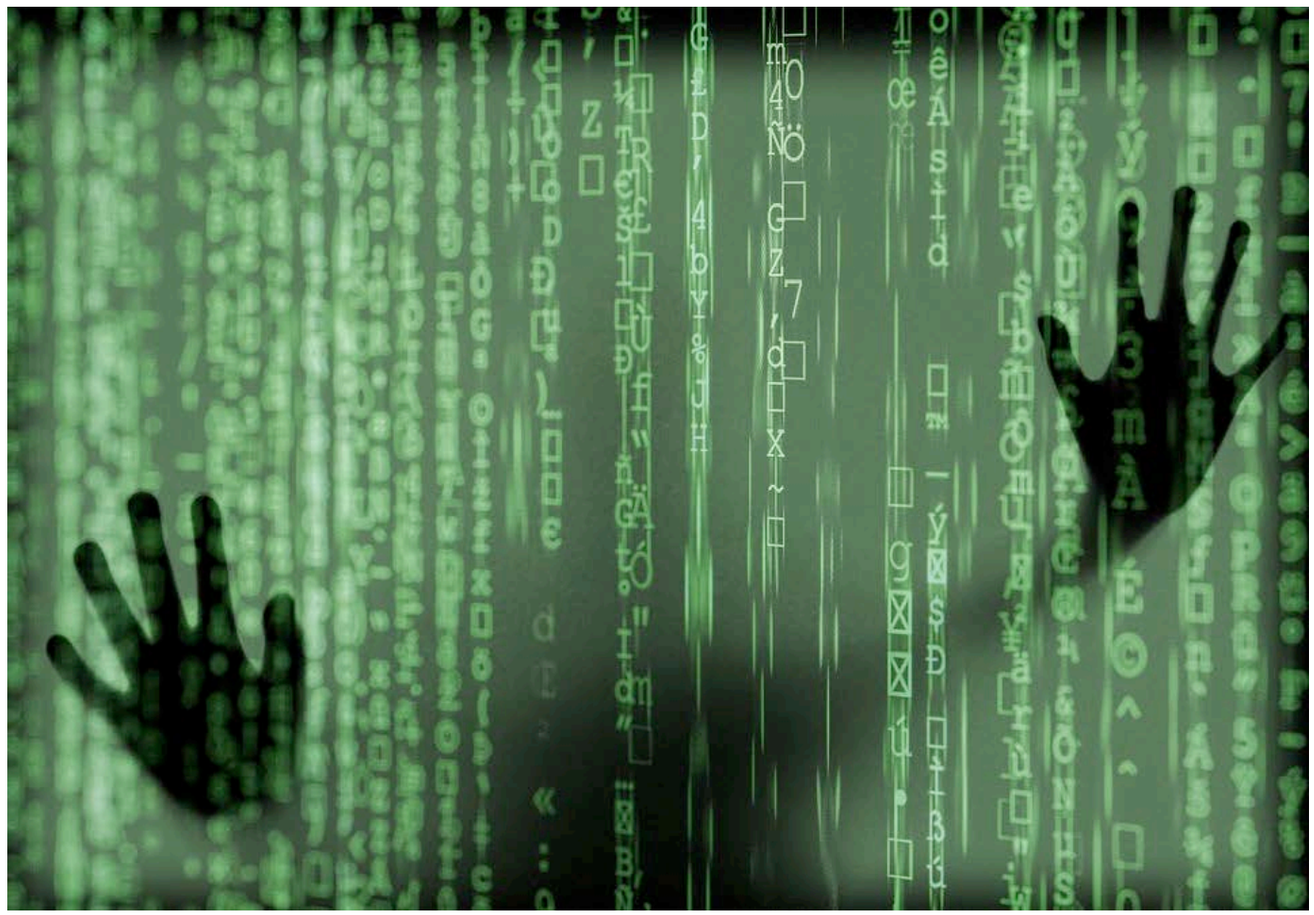
Digital forensics tools require constant maintenance

OS Creep

Language Creep

Forensic Science Creep

O&M (operations & maintenance) “tail”

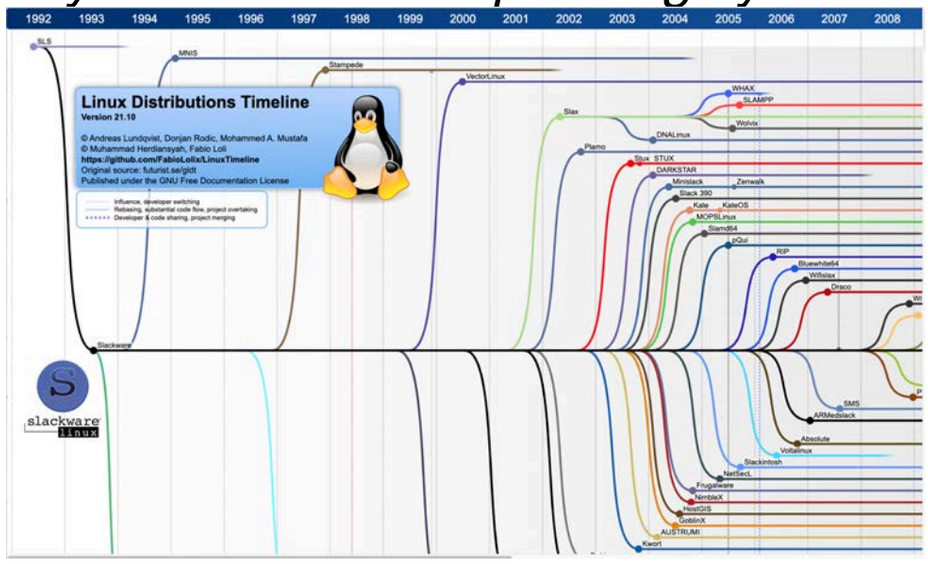


https://pixabay.com/illustrations/hacker-computer-ghost-cyber-code-4031973/

Digital forensics tools require constant maintenance

OS Creep

- Platforms being analyzed change over time
  - —*Windows 95 → Windows NT → Windows XP → Windows 7 → Windows 10*
  - —*Feature Phones → iPhone & Android*
  - —*Tablets*
- Forensics practitioners favor different operating systems over time
  - —*Linux / Windows / MacOS*
- OS used for analysis must be upgraded
  - —*Old apps may have bugs or security vulnerabilities*
  - —*Old apps may not run on new OS*
  - —*New versions of apps may not run on old operating systems*



Digital forensics tools require constant maintenance

OS Creep ✔

Language Creep — Mostly a concern for open-source software

- Open-source software is typically distributed in source-code form
- Operating systems are better at preserving binary compatibility than source-code compatibility
  - —*ABI (Application Binary Interface) is very stable*
  - —*High-level languages change — file names change, features are deprecated, etc*

- Example:
  - —*Java source code from the early 2000s will not compile with a modern Java compiler*
  - —*Java bytecode from the early 2000s will frequently run on a modern JVM*
  - —*Java bytecode & JVM from the early 2000s will almost always run on a modern OS*

Digital forensics tools require constant maintenance

OS Creep ✔

Language Creep ✔

Digital forensics creep — DF science is constantly improving

- DF keeps getting better!
  - —*More complete implementations of today's undocumented data structures*
  - —*More reliable, efficient implementations of today's documented data structures*
- DF is struggling to keep up!
  - —*Compression standards (e.g. Snappy)*
  - —*New memory structures (e.g. Windows 10 memory structures)*
  - —*New image formats (e.g. HEIC)*
- DF software keeps improving
  - —*Usability improvements, support for running in cloud, etc.*

Digital forensics tools require constant maintenance

OS Creep ✔

Language Creep ✔

Forensic Science Creep ✔

O&M (operations & maintenance) “tail”

- All software needs to be maintained
- DF software is not any different
  - *—Bugs reported in software*
  - *—Updates to secure hash algorithms (MD5 ❌; SHA-1 ❌; SHA-256 ✔)*

There were many reasons to update bulk_extractor

Maintenance Costs

- Autoconf-based system required modification for major OS releases
  - —*BE uses threading, access file systems, etc*
- bulk_extractor *support* of out-of-date Python versions
  - —*caused it to be banned from a Linux release!*

Changes in CPU / IO / memory trade-off

- CPU cores are ~50% faster than in 2012
- Laptops and low-end workstations have 2x to 3x as many cores
- High-end servers: 64 cores in 2012; 96 cores in 2020
- Memory is 3x faster; SSDs are commonplace now → no seek time
- Disk I/O and network drives are faster

Large parts of BE were single-threaded

- BE1 — 1 thread per 16MiB page. “Last page” could take 30-60 min to process
- Histogram processing: batch at end of page processing, and single-threaded.

The most important reason: Correctness

Most computer software implements specifications:

- Formal specifications — RFCs, end-user requirements, etc
- Informal specifications — What's in the programmer's head
- Being able to *read data* written by the *same program*

Many digital forensics tools are based on reverse engineering

- Read and decode data written by other programs
- Authors of other programs may be *unknown* or *unwilling* to share technical details

Many digital forensics tools crash or print warnings when they run

- Bulk_extractor when processing *nps-2009-domexusers.E01:*

```
11:33:51 Offset 486MB (1.13%) Done in  1:57:57 at 13:31:48
11:34:08 Offset 570MB (1.33%) Done in  2:02:19 at 13:36:27
11:34:25 Offset 654MB (1.52%) Done in  2:03:45 at 13:38:10
std::exception Scanner: evtx Exception: Error: Read past end of sbuf sbuf.pos0: (661649934-HIBERFILE|84582400) bufsize=4096
std::exception Scanner: evtx Exception: Error: Read past end of sbuf sbuf.pos0: (721594368-HIBERFILE|44343296) bufsize=4096
std::exception Scanner: evtx Exception: Error: Read past end of sbuf sbuf.pos0: (721594368-HIBERFILE|44351488) bufsize=4096
std::exception Scanner: evtx Exception: Error: Read past end of sbuf sbuf.pos0: (721594368-HIBERFILE|44384256) bufsize=4096
```

Update plan: objectives

Make the program easier to compile and maintain

Make it easier for others to contribute code

Removal experimental code & simplify the codebase

Decrease program’s runtime

Result of updates: bulk_extractor is faster, more reliable, and easier to maintain

Computer	Disk Image (+ config)	Scanners: BE1.6	29 BE2	29 Throughput	30 + AES192 BE2	30 + AES192 Throughput
MacBook Pro (Retina, 13-inch, Late 2013) 2.8 GHz Dual-Core Intel Core i7; 16 GiB 1600 MHz DDR3; 2 physical cores (4 with hyperthreading); macOS 11.6.3						
	nps-2009-ubnist1	140 s	109 s	128%	120 s	117%
	nps-2009-domexusers	1420 s	837 s	170%	1208 s	118%
Mac mini (2018) 3GHz 6-core i5; 2667 MHz DDR4; macOS 12.1						
	nps-2009-ubnist1	43 s	35 s	123%	33 s	130%
	nps-2009-domexusers	428 s	319 s	134%	428 s	100%
MacBook Pro (16-inch, 2021) Apple M1 Pro 10 core; 32GiB RAM; macOS 12.1						
	nps-2009-ubnist1	20 s	16 s	125%	17 s	118%
	nps-2009-domexusers	221 s	126 s	175%	172 s	128%
	nps-2013-2tb	20 142 s	10 944 s	184%	11 184 s	180%

BE1 vs. BE2:
BE2 is finding a lot of stuff that BE1 missed

Size ✔

Compile-time (relevant for development) ✔

Runtime ✔

Analysis

file	BE16	BE2.0 Beta 4
alerts.txt	62	19
domain.txt	72,027	76,800
email.txt	8,757	8,751
ether.txt	5	1
ether_histogram_1.txt	n/a	0
exif.txt	232	235
facebook.txt	n/a	0
ip.txt	4	4,444
jpeg_carved.txt	43	1,767
json.txt	4	958
kml.txt	0	2
ntfsusn_carved.txt	2	1
rfc822.txt	4,240	4,219
tcp.txt	n/a	56
tcp_histogram.txt	n/a	0
telephone.txt	767	760
unzip_carved.txt	41	n/a
url.txt	108,352	112,754
winpe.txt	10,740	10,592
winpe_carved.txt	4	10,573
winprefetch.txt	124	0
zip.txt	5,196	10,193

1,724 additional JPEGs carved

10,569 windows executables carved!

Conclusion:
What this means for digital forensics tools

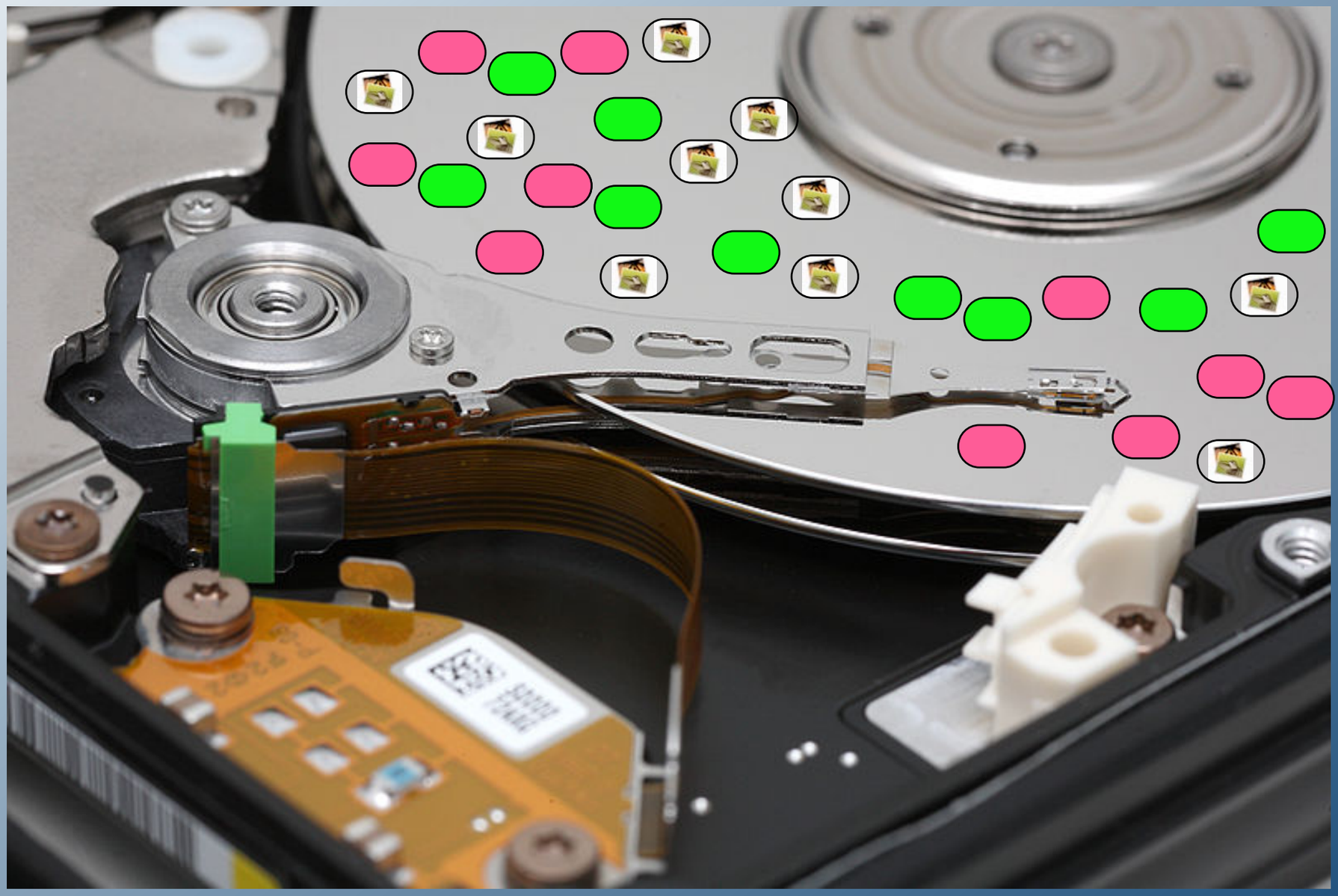
New releases:

- Should be validated against previous releases in a systemic manner
- Results should be published in a machine-readable form
- Clearly document:
  - —*New data that is recovered from legacy datasets (compared to previous version)*
  - —*Data recovered from new datasets that previous version would miss*
  - —*Overcollection that has been eliminated*

We need to set expectations for DF tools

- Complete rewrites are slow
  - —*10 years to get from "Ethereal" to Wireshark 1.0 in 2008, 2.0 in 2015*
  - —*Volatility 2: 2.5 - October 2015; 2.6 - December 2016*
  - —*Volatility 3: v1.0.0 - Feb 01, 2021; v 1.0.1 - Feb 1, 2021*

Unclear how to measure proprietary tools



Sector hashing (2009-2014)

(skip if not enough time)

Question: Can we analyze a 1TB drive in a minute?

What if we encounter a hard drive at a border crossing?

Or a search turns up a room filled with servers?

In 2020 it took 3.5 hours to read a 1TB drive. What could we learn in 5 minutes?

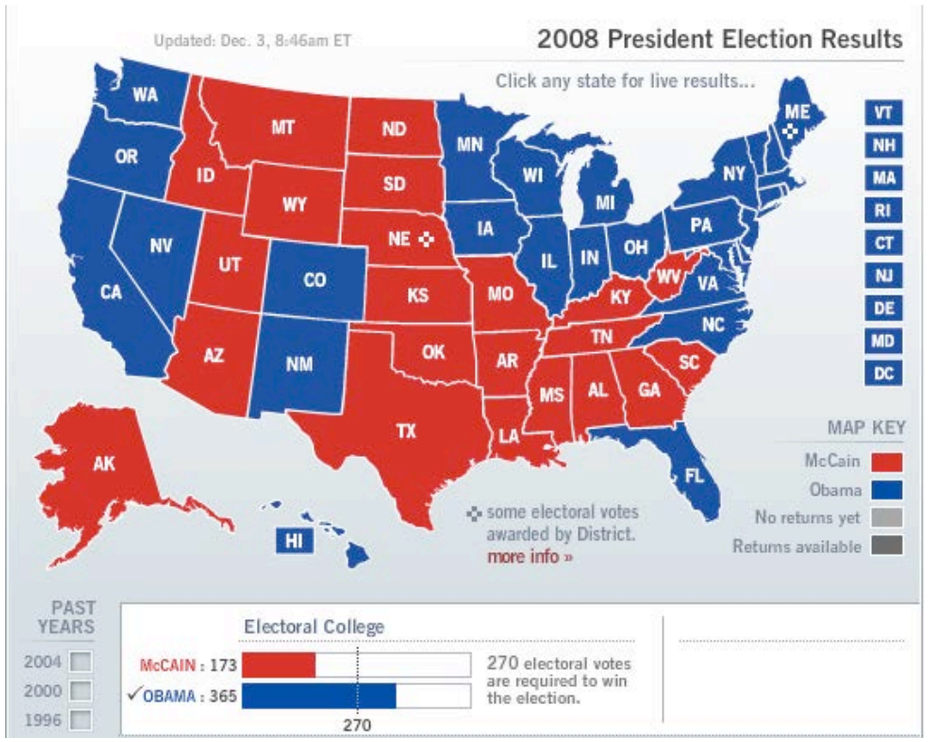
24 GB (2.4%) is a tiny fraction of the disk

But 24 GB is a lot of data!

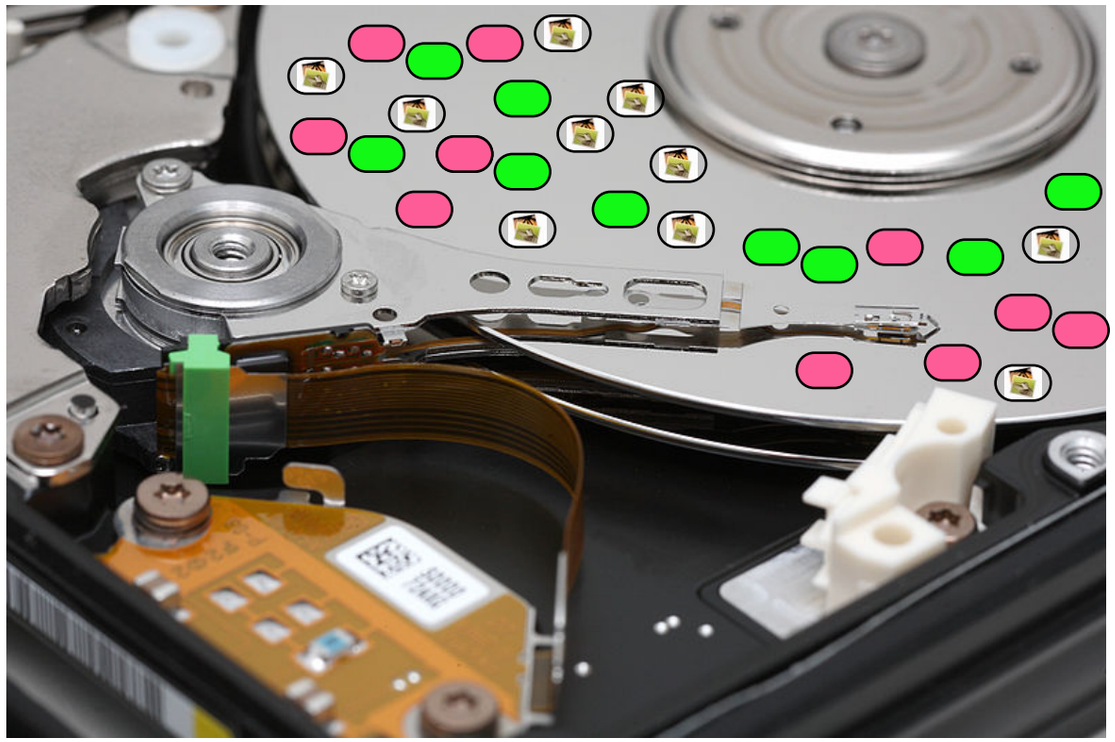
		
Minutes	208	5
Max Data Read	1 TB	4.8 GB

Hypothesis: The contents of the disk can be predicted by identifying the contents of randomly chosen sectors.

US elections can be predicted by sampling a few thousand households:



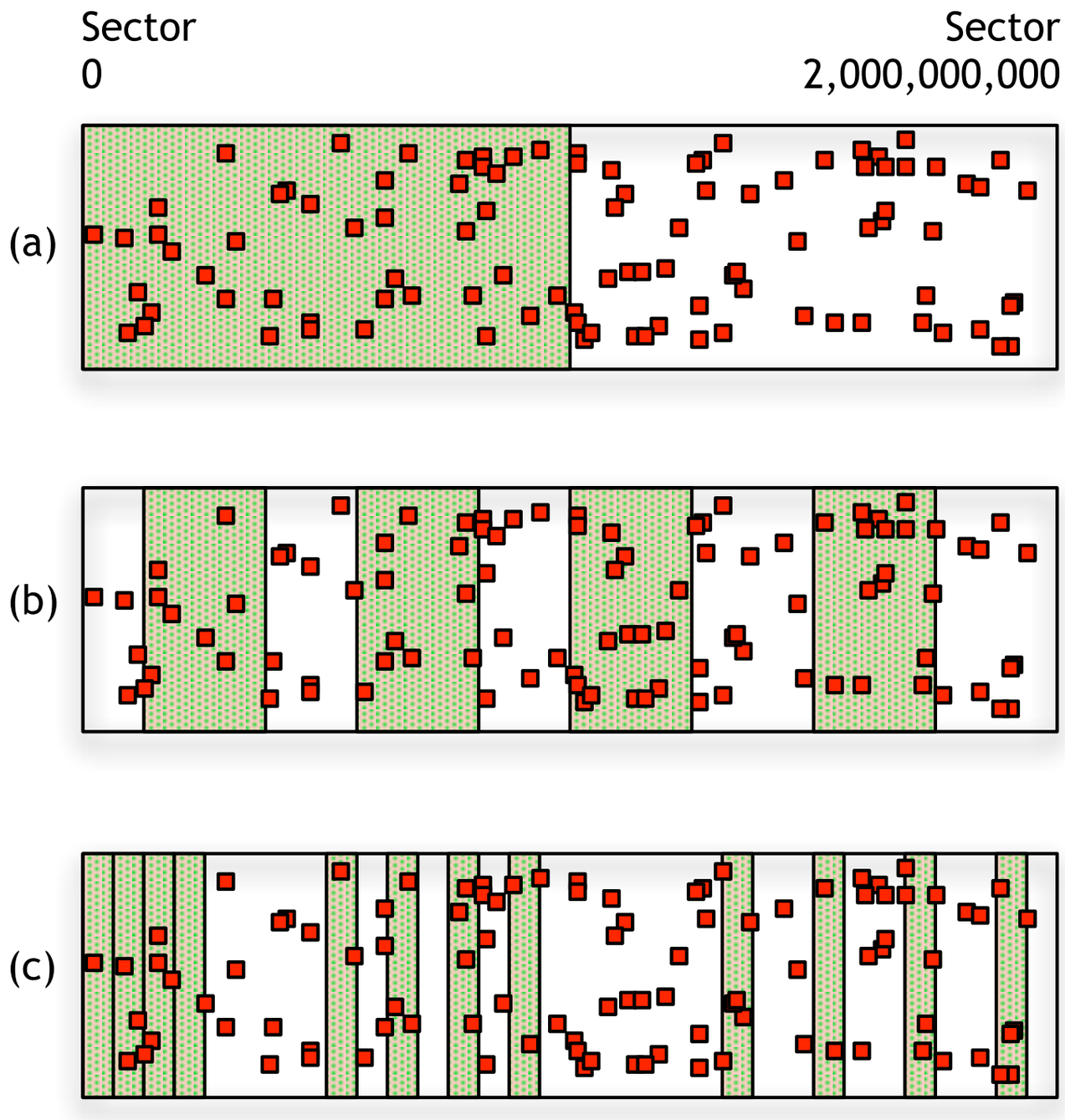
Hard drive contents can be predicted by sampling a few thousand sectors:



The challenge is identifying *likely voters.*

The challenge is *identifying the content* of the sampled sectors.

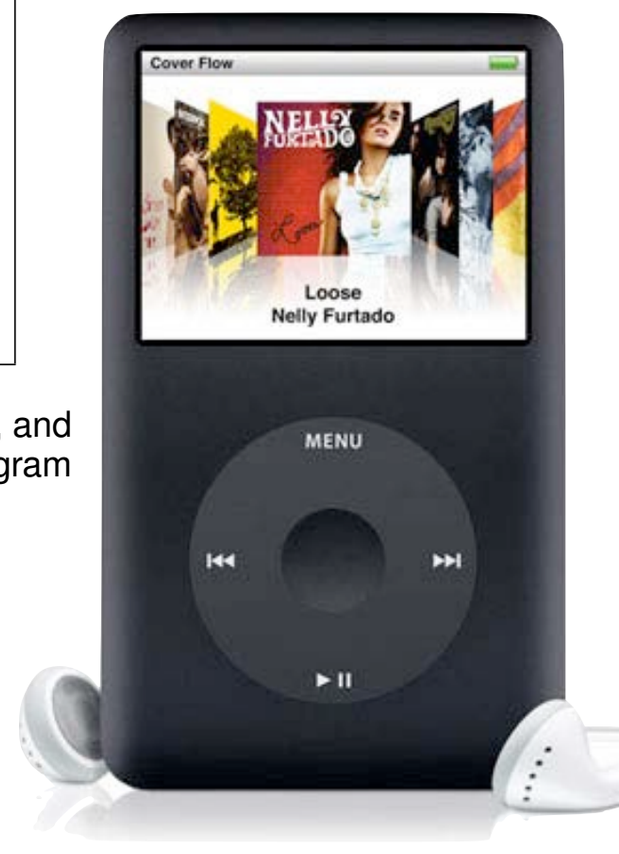
We used random sampling; any other approach could be exploited by an adversary.



2011 - With random sampling, we accurately determined the contents of a 160GB iPod in 5 minutes.

Audio Data reported by iTunes:	2.25 GiB	2.42 GB	
MP3 files reported by file system:		2.39 GB	
Estimated MP3 usage with random sampling :		2.49 GB	10,000 random samples
		2.71 GB	5,000 random samples

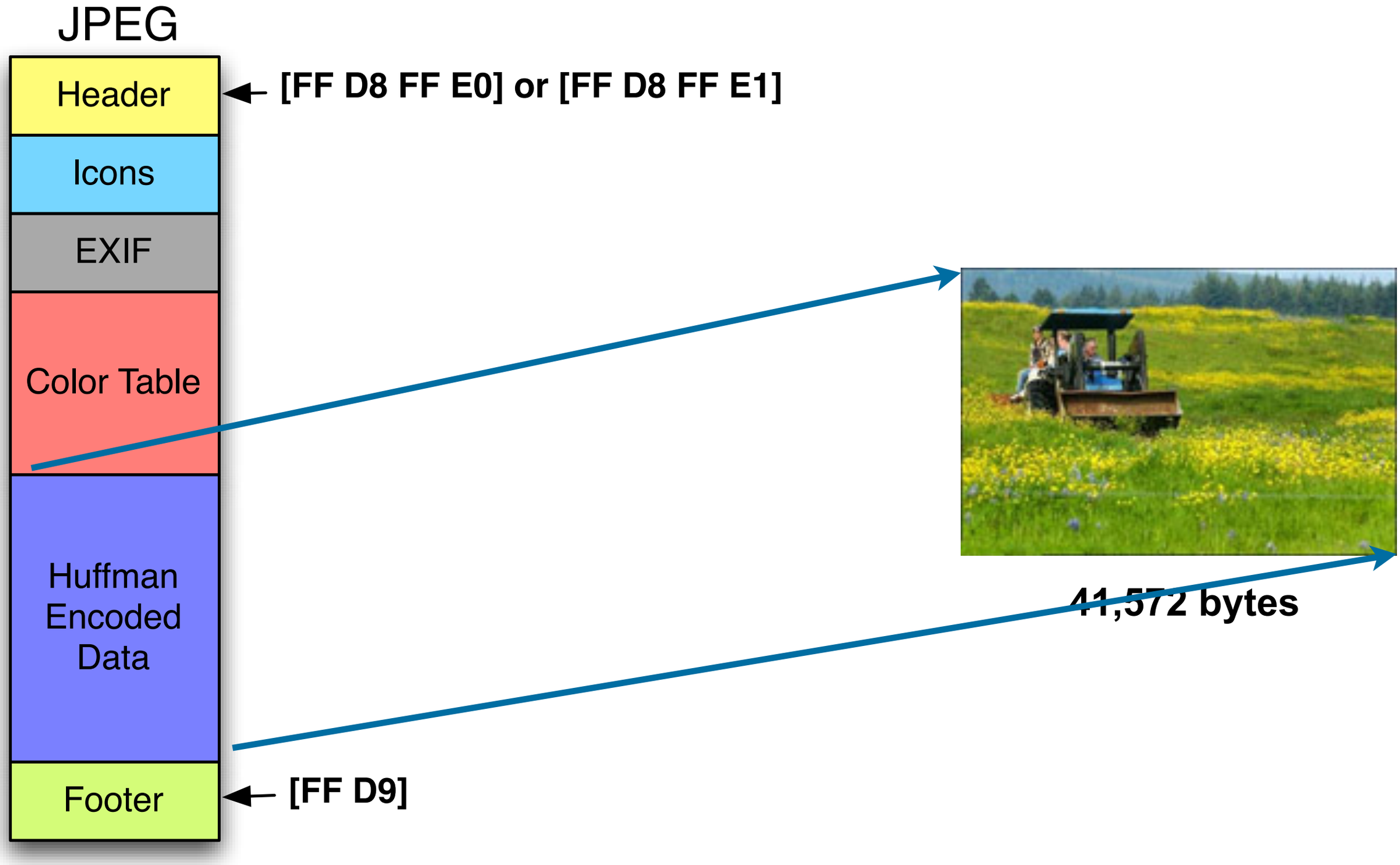
Figure 1: Usage of a 160GB iPod reported by iTunes 8.2.1 (6) (top), as reported by the file system (bottom center), and as computing with random sampling (bottom right). Note that iTunes usage actually in GiB, even though the program displays the "GB" label.



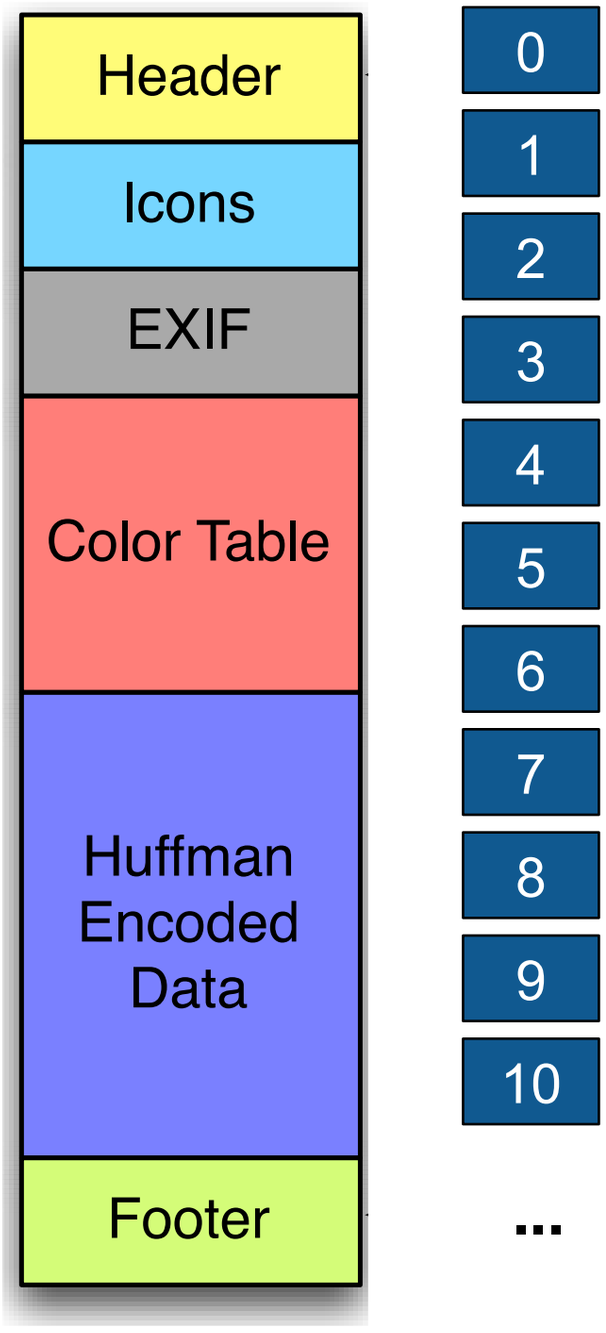
We determined:

- % 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*

We hypothesized that some portions of a JPEG would be distinct (~~unique~~).



We viewed the 41K file as a sequence of 88 blocks (512b)



Block #	Hex Values...
0	ffd8 ffe0 0010 4a46 4946 0001 0201 0048...
1	0c0c 0c0c ffc0 0011 0800 6a00 a003 0122...
2	4fa7 7567 ded2 cac5 8c82 2bf4 9e1c 23f9...
3	fafd 1527 e459 e934 c173 59ad 9234 f09f...
...	...

Each block has a cryptographic hash. Some are distinct, others are common.

Block #	Byte Range	MD5*(block(N))
0	0- 511	**dc0c20abad42d487a74f308c69d18a5a**
1	512-1023	**9e7bc64399ad87ae9c2b545061959778**
2	1024-1535	**6e7f3577b100f9ec7fae18438fd5b047**
3	1536-2047	**4594899684d0565789ae9f364885e303**
4	...	

Question: which of these MD5 hashes appear in other JPEGs?

Are these same block hashes in other files?

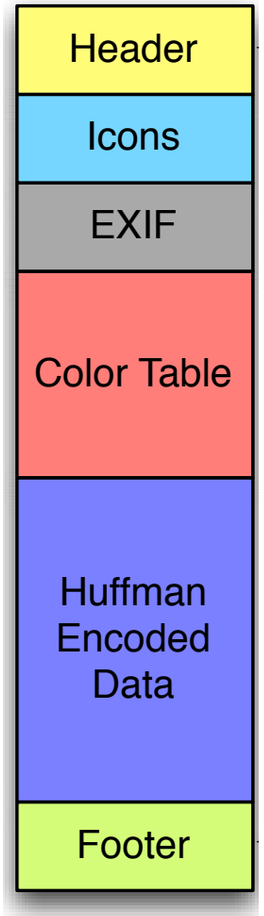
Specific byte sequences in high-entropy data are very rare

- 512 bytes = 256^{512} =$10^{1,233}$ possible sectors

But metadata might be common:

- Specific headers
- Common color tables
- “all black”

You need to survey
a large samples of JPEGs
to find out which hashes are
common and which are distinct



MD5(block(N))*
`dc0c20abad42d487a74f308c69d18a5a`
`9e7bc64399ad87ae9c2b545061959778`
`6e7f3577b100f9ec7fae18438fd5b047`
`459489968 4d0565789ae9f364885e303`
`...`

We examined sector hashes from ≃4 million files

≃ 1 million in GOVDOCS1 collection

= 109,282 JPEGs (including 000107.jpg)

≃ 3 million samples of Windows malware

Results:

- Most of the block hashes in 000107.jpg do not appear elsewhere in the corpus
- Some of the block hashes appeared in other JPEGs
- None of the block hashes appeared in files that were not JPEGs

The beginning of 000107.jpg contained distinct hashes...

hash	location	count
dc0c20abad42d487a74f308c69d18a5a	offset 0-511	1
9e7bc64399ad87ae9c2b545061959778	offset 512-1023	1
6e7f3577b100f9ec7fae18438fd5b047	offset 1024-1535	1
4594899684d0565789ae9f364885e303	offset 1536-2047	1
4d21b27ceec5618f94d7b62ad3861e9a	offset 2048-2559	1
03b6a13453624f649bbf3e9cd83c48ae	offset 2560-3071	1
c996fe19c45bc19961d2301f47cabaa6	offset 3072-3583	1
0691baa904933c9946bbda69c019be5f	offset 3584-4095	1
1bd9960a3560b9420d6331c1f4d95fec	offset 4096-4607	1
52ef8fe0a800c9410bb7a303abe35e64	offset 4608-5119	1
b8d5c7c29da4188a4dcaa09e057d25ca	offset 5120-5631	1
3d7679a976b91c6eb8acd1bfa3414f96	offset 5632-6143	1
8649f180275e0b63253e7ee0e8fa4c1d	offset 6144-6655	1
60ebc8acb8467045e9dcbe207f61a6c2	offset 6656-7167	1
440c1c1318186ac0e42b2977779514a1	offset 7168-7679	1
72686172f8c865231e2b30b2829e3dd9	offset 7680-8191	1
fdff55c618d434416717e5ed45cb407e	offset 8192-8703	1
fcd89d71b5f728ba550a7bc017ea8ff1	offset 8704-9215	1
2d733e47c5500d91cc896f99504e0a38	offset 9216-9727	1
2152fdde0e0a62d2e10b4fecc369e4c6	offset 9728-10239	1
692527fa35782db8592486343 6d45d7f	offset 10240-10751	1
76dbb9b469273d0e0e467a55728b7883	offset 10752-11263	1

The middle of 000107.JPG appear elsewhere...

hash	location	count
9df886fdfa6934cc7dcf10c04be3464a	offset 14848-15359	1
95399e7ecc7ba1b38243069bdd5c263a	offset 15360-15871	1
ef1ffcdc11162ecdfedd2dde644ec8f2	offset 15872-16383	1
7eb35c161e91b215e2a1d20c32f4477e	offset 16384-16895	1
38f9b6f045db235a14b49c3fe7b1cec3	offset 16896-17407	1
edceba3444b5551179c791ee3ec627a5	offset 17408-17919	1
6bc8ed0ce3d49dc238774a2bdeb7eca7	offset 17920-18431	1
5070e4021866a547aa37e5609e401268	offset 18432-18943	14
13d33222848d5b25e26aefb87dbdf294	offset 18944-19455	9198
0dfcde85c648d20aed68068cc7b57c25	offset 19456-19967	9076
756f0bbe70642700aafb2557bf2c5649	offset 19968-20479	9118
c2c29016d3005f7a1df247168d34e673	offset 20480-20991	9237
42ff3d72b2b25f880be21fac46608cc9	offset 20992-21503	9708
b943cd0ea25e354d4ac22b886045650d	offset 21504-22015	9615
a003ec2c4145b0bc871118842b74f385	offset 22016-22527	9564
1168c351f57aad14de135736c06665ea	offset 22528-23039	7
51a50e6148d13111669218dc40940ce5	offset 23040-23551	83
365b122f53075cb76b39ca1366418ff9	offset 23552-24063	83
9ad9660e7c812e2568aaf063a1be7d05	offset 24064-24575	84
67bd01c2878172e2853f0aef341563dc	offset 24576-25087	84
fc3e47d734d658559d1624c8b1cbf2c1	offset 25088-25599	84
cb9aef5b7f32e2a983e67af38ce8ff87	offset 25600-26111	1

Block 37 had 9198 collisions..
The sector is filled with blank lines 100 characters long...

```
13d33222848d5b25e26aefb87dbdf294   offset 18944-19455       9198

$ dd if=000107.jpg skip=18944 count=512 bs=1|xxd
0000000: 2020 2020 2020 2020 2020 2020 2020 2020
0000010: 2020 2020 2020 2020 2020 2020 0a20 2020                .
0000020: 2020 2020 2020 2020 2020 2020 2020 2020
0000030: 2020 2020 2020 2020 2020 2020 2020 2020
0000040: 2020 2020 2020 2020 2020 2020 2020 2020
0000050: 2020 2020 2020 2020 2020 2020 2020 2020
0000060: 2020 2020 2020 2020 2020 2020 2020 2020
0000070: 2020 2020 2020 2020 2020 2020 2020 2020
0000080: 200a 2020 2020 2020 2020 2020 2020 2020    .
0000090: 2020 2020 2020 2020 2020 2020 2020 2020
00000a0: 2020 2020 2020 2020 2020 2020 2020 2020
00000b0: 2020 2020 2020 2020 2020 2020 2020 2020
00000c0: 2020 2020 2020 2020 2020 2020 2020 2020
00000d0: 2020 2020 2020 2020 2020 2020 2020 2020
00000e0: 2020 2020 2020 0a20 2020 2020 2020 2020          .
00000f0: 2020 2020 2020 2020 2020 2020 2020 2020
0000100: 2020 2020 2020 2020 2020 2020 2020 2020
0000110: 2020 2020 2020 2020 2020 2020 2020 2020
0000120: 2020 2020 2020 2020 2020 2020 2020 2020
0000130: 2020 2020 2020 2020 2020 2020 2020 2020
0000140: 2020 2020 2020 2020 2020 200a 2020 2020                .
0000150: 2020 2020 2020 2020 2020 2020 2020 2020
0000160: 2020 2020 2020 2020 2020 2020 2020 2020
```

Block 45 had 83 collisions..
It appears to contain EXIF metadata

```
51a50e6148d13111669218dc40940ce5   offset 23040-23551        83
$ dd if=000107.jpg skip=23040 count=512 bs=1|xxd
0000000: 3936 362d 322e 3100 0000 0000 0000 0000  966-2.1.........
0000010: 0000 0000 0000 0000 0000 0000 0000 0000  ................
0000020: 0000 0000 0000 0000 0000 0000 0000 0000  ................
0000030: 0000 0000 0000 0000 0058 595a 2000 0000  .........XYZ ..
0000040: 0000 00f3 5100 0100 0000 0116 cc58 595a  ....Q.........XYZ
0000050: 2000 0000 0000 0000 0000 0000 0000 0000   ..............
0000060: 0058 595a 2000 0000 0000 006f a200 0038  .XYZ ......o...8
0000070: f500 0003 9058 595a 2000 0000 0000 0062  .....XYZ ......b
0000080: 9900 00b7 8500 0018 da58 595a 2000 0000  .........XYZ ..
0000090: 0000 0024 a000 000f 8400 00b6 cf64 6573  ...$.........des
00000a0: 6300 0000 0000 0000 1649 4543 2068 7474  c........IEC htt
00000b0: 703a 2f2f 7777 772e 6965 632e 6368 0000  p://www.iec.ch.
00000c0: 0000 0000 0000 0000 0016 4945 4320 6874  ..........IEC ht
00000d0: 7470 3a2f 2f77 7777 2e69 6563 2e63 6800  tp://www.iec.ch
00000e0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
00000f0: 0000 0000 0000 0000 0000 0000 0000 0000  ................
0000100: 0000 0000 0000 0000 0000 0000 0064 6573  .............des
0000110: 6300 0000 0000 0000 2e49 4543 2036 3139  c........IEC 619
0000120: 3636 2d32 2e31 2044 6566 6175 6c74 2052  66-2.1 Default R
0000130: 4742 2063 6f6c 6f75 7220 7370 6163 6520  GB colour space
0000140: 2d20 7352 4742 0000 0000 0000 0000 0000  - sRGB.........
0000150: 002e 4945 4320 3631 3936 362d 322e 3120  ..IEC 61966-2.1
0000160: 4465 6661 756c 7420 5247 4220 636f 6c6f  Default RGB colo
0000170: 7572 2073 7061 6365 202d 2073 5247 4200  ur space - sRGB.
```

Block 48 had 84 collisions..
It appears to contain part of a JPEG color table...

```
67bd01c2878172e2853f0aef341563dc   offset 24576-25087          84
$ dd if=000107.jpg skip=24576 count=512 bs=1 |xxd
0000000: 7a27 ab27 dc28 0d28 3f28 7128 a228 d429  z'.'.(.(?(q(.(.)
0000010: 0629 3829 6b29 9d29 d02a 022a 352a 682a  .)8)k).).*.*5*h*
0000020: 9b2a cf2b 022b 362b 692b 9d2b d12c 052c  .*.+.+6+i+.+.,.,
0000030: 392c 6e2c a22c d72d 0c2d 412d 762d ab2d  9,n,.,.-.-A-v-.-
0000040: e12e 162e 4c2e 822e b72e ee2f 242f 5a2f  ....L....../$/Z/
0000050: 912f c72f fe30 3530 6c30 a430 db31 1231  ././.0501 0.0.1.1
0000060: 4a31 8231 ba31 f232 2a32 6332 9b32 d433  J1.1.1.2*2c2.2.3
0000070: 0d33 4633 7f33 b833 f134 2b34 6534 9e34  .3F3.3.3.4+4e4.4
0000080: d835 1335 4d35 8735 c235 fd36 3736 7236  .5.5M5.5.5.676r6
0000090: ae36 e937 2437 6037 9c37 d738 1438 5038  .6.7$7`7.7.8.8P8
00000a0: 8c38 c839 0539 4239 7f39 bc39 f93a 363a  .8.9.9B9.9.9.:6:
00000b0: 743a b23a ef3b 2d3b 6b3b aa3b e83c 273c  t:.:.;-;k;.;.<'<
00000c0: 653c a43c e33d 223d 613d a13d e03e 203e  e<.<.="=a=.=.> >
00000d0: 603e a03e e03f 213f 613f a23f e240 2340  `>.>.?!?a?.?.@#@
00000e0: 6440 a640 e741 2941 6a41 ac41 ee42 3042  d@.@.A)AjA.A.BOB
00000f0: 7242 b542 f743 3a43 7d43 c044 0344 4744  rB.B.C:C}C.D.DGD
0000100: 8a44 ce45 1245 5545 9a45 de46 2246 6746  .D.E.EUE.E.F"FgF
0000110: ab46 f047 3547 7b47 c048 0548 4b48 9148  .F.G5G{G.H.HKH.H
0000120: d749 1d49 6349 a949 f04a 374a 7d4a c44b  .I.IcI.I.J7J}J.K
0000130: 0c4b 534b 9a4b e24c 2a4c 724c ba4d 024d  .KSK.K.L*LrL.M.M
0000140: 4a4d 934d dc4e 254e 6e4e b74f 004f 494f  JM.M.N%NnN.O.OIO
0000150: 934f dd50 2750 7150 bb51 0651 5051 9b51  .O.P'PqP.Q.QPQ.Q
0000160: e652 3152 7c52 c753 1353 5f53 aa53 f654  .R1R|R.S.S_S.S.T
0000170: 4254 8f54 db55 2855 7555 c256 0f56 5c56  BT.T.U(UuU.V.V\V
```

With blocks of 512 bytes and 4KiB, the vast majority of sectors had distinct hashes. 4KiB was more distinct.

Table 1. Incidence of singleton, paired, and common sectors in three file corpora.

No. of blocks	Govdocs	OpenMalware 2012	2009 NSRL RDS
Block size: 512 bytes			
Singleton	911.4 M (98.93%)	1,063.1 M (88.69%)	N/A
Pair	7.1 M (.77%)	75.5 M (6.30%)	N/A
Common	2.7 M (.29%)	60.0 M (5.01%)	N/A
Block size: 4 kibibytes			
Singleton	117.2 M (99.46%)	143.8 M (89.51%)	567.0 M (96.00%)
Pair	0.5 M (.44%)	9.3 M (5.79%)	16.4 M (2.79%)
Common	0.1 M (.11%)	7.6 M (4.71%)	7.1 M (1.21%)

Young, Foster, Garfinkel & Fairbanks, IEEE Computer, Dec. 2012

File systems align large files on sector boundaries. We hash file blocks and identify sectors that match.

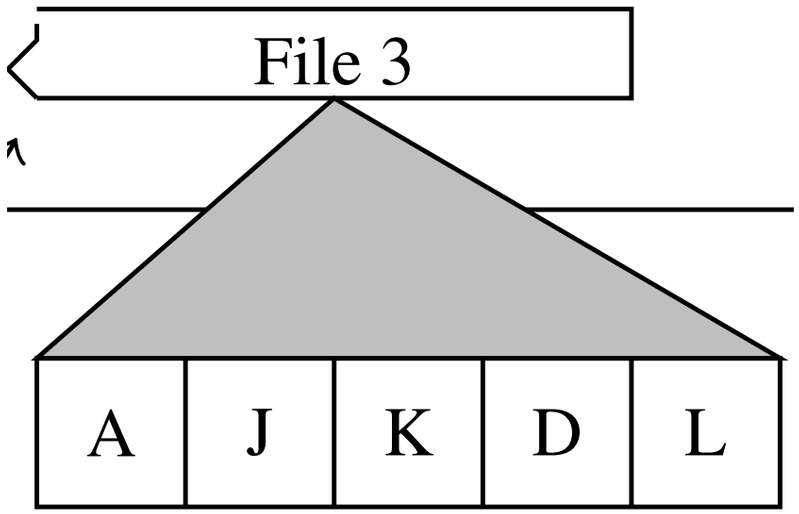
Block #	Byte Range	MD5*(block(N))
0	0- 511	**dc0c20abad42d487a74f308c69d18a5a**
1	512-1023	**9e7bc64399ad87ae9c2b545061959778**
2	1024-1535	**6e7f3577b100f9ec7fae18438fd5b047**
3	1536-2047	**4594899684d0565789ae9f364885e303**
4	...	

Using distinct sectors in media sampling and full media analysis to detect presence of documents from a corpus, Kristina Foster, NPS Master's Thesis, 2012

This means we can use distinct sectors to find known content.

Method #1 — Full media sampling

- Read & hash every disk sector
- Lookup hash values in a database of block hashes
- Distinct hash imply presence of files
- Advantage: Can find a single sector of target content



Method #2 — Random sampling

- Read & hash randomly chosen sectors
- Lookup hash values in a database of block hashes
- Distinct hash implies presence of files
- Advantage: Can find presence of target content very quickly

Database requirements to read 1TB data in 208 minutes

- ≃ 80 Mbyte/sec ≃ 150,000 512-byte sectors/sec = 150,000 database lookups/sec
- We built “hashdb,” a custom database based on Bloom filters.

By combining a Bloom filter & database, we can perform up to 2.7M TPS on low-cost hardware

Table 2. Total transactions per second (TPS) for best execution.

Bloom filter			Database		TPS at 1 M lookups		TPS at 1,200 seconds	
k	*M*	Size	Strategy	Size	Present	Absent	Present	Absent
100 million records								
3	31	257 MiBytes	B-tree (preload)	2.3 GiBytes	35.3 K	49.5 K	161.3 K	1.8 M
3	31	257 MiBytes	B-tree	2.3 GiBytes	11.6 K	565.8 K	156.8 K	2.3 M
3	31	257 MiBytes	Hash map	5.3 GiBytes	13.9 K	656.9 K	641.9 K	3.0 M
3	31	257 MiBytes	Flat map	2.2 GiBytes	28.2 K	746.9 K	356.4 K	2.6 M
3	31	257 MiBytes	Red/black tree	6.0 GiBytes	12.9 K	694.5 K	187.0 K	2.7 M
1 billion records								
3	34	2.1 GiBytes	B-tree (preload)	23 GiBytes	2.2 K	6.1 K	3.6 K	23.1 K
3	33	1.1 GiBytes	B-tree	23 GiBytes	2.6 K	85.8 K	3.7 K	114.9 K
3	33	1.1 GiBytes	Hash map	57 GiBytes	–	–	0.3 K	3.1 K
3	34	2.1 GiBytes	Flat map	22 GiBytes	–	–	0.4 K	4.0 K
3	33	1.1 GiBytes	Red/black tree	60 GiBytes	–	–	0.1 K	1.4 K

Hardware: 8GiB Laptop; 250GB external SSD

—"Distinct sector hashes for target file detection," Young, Garfinkel, Foster &

Putting it all together, we have a significant innovation... field deployable on a single laptop.

Use Case #1: Rapidly search for known content (contraband?):

- 1TB subject hard drive.
- 10 min x 60 min/sec x 1000 msec/sec / 3 msec/sample = 200,000 samples
- Searching for a sector from a corpus of 512GB
- 100% recognition of a single sector; 0% false positive rate

Amount of Content	p (prob of missing content)
5 MB	0.3654
10 MB	0.1335
15 MB	0.0488
20 MB	0.0178
25 MB	0.0065

Use Case #2: Find a single sector of known content:

- Time to read data & search database: 208 minutes

Technique is file type and file system agnostic

JPEG; Video; MSWord; Encrypted PDFs..

—provided data is not modified when copied or otherwise re-coded

We published our results, but could not operationalize.

COVER FEATURE

Distinct Sector Hashes for Target File Detection

Joel Young, Kristina Foster, and Simson Garfinkel, *Naval Postgraduate School*
Kevin Fairbanks, *Johns Hopkins University*

Using an alternative approach to traditional file hashing, digital forensic investigators can hash individually sampled subject drives on sector boundaries and then check these hashes against a prebuilt database, making it possible to process raw media without reference to the underlying file system.

Forensic examiners frequently search disk drives, cell phones, and even network flows to determine if specific known content is present. For example, a corporate security officer might examine a suspicious employee's laptop for unauthorized documents; law enforcement officers might search a suspect's home computer for illegal pornography; and network analysts might reconstruct Transmission Control Protocol streams to determine if malware was downloaded. In these and many other cases, examiners typically identify files by computing their cryptographic hash—often with MD5 or SHA1 hash algorithms—and then searching a database for the resulting hash value.

Use of hash values for file identification is pervasive in digital forensics—every popular forensics package has built-in support. One of the most widely used databases is the National Software Reference Library (NSRL) Reference Data Set (RDS). Version 2.36, released in March 2012, contains 25,892,924 distinct file hashes (www.nsrl.nist.gov). Other databases are available to customers of specific companies and to law enforcement organizations.

There are many limitations when using file hashes to identify known content. Because changing just a single bit of a file changes its hash, pornographers, malware authors, and other miscreants can evade detection simply by changing a comma to a period or appending a few random bytes to a file. Likewise, hash-based identification will not work if sections of the file are damaged or otherwise unrecoverable. This is especially a problem when large video files are deleted and the operating system reuses a few sectors for other purposes: most of the video is still present on the drive, but recovered video segments will not appear in a database of file hashes.

SECTOR HASHING

We are developing alternative systems for detecting target files in large disk images using cryptographic hashes on sectors of data rather than entire files. Modern file systems align the start of most files with the beginning of a disk sector. Thus, when a megabyte-sized video is stored on a modern hard drive, the first 4 kibibytes are stored in one disk sector, the second 4 KiBytes are stored in another disk sector, typically the adjacent one, and so on. (In our work, we distinguish between power-of-two-based sizes of digital artifacts, such as kibibytes, and power-of-ten-based sizes, such as kilobytes. See the "Decimal versus Binary Prefixes" sidebar for more details.) Furthermore, by sampling randomly chosen sectors from the drive, it is only necessary to read a tiny fraction of the drive to determine with high probability if a target file is present. This enables rapid triage of drive images.

We compare drive sector hashes to a hash database of fixed-sized file fragments, which we call *blocks*. The terms "sector" and "block" are often used incorrectly as syn-

Published by the IEEE Computer Society

IEEE Computer - December 2012

Digital Investigation 14 (2015) S95–S105

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2015 USA

Hash-based carving: Searching media for complete files and file fragments with sector hashing and hashdb

Simson L. Garfinkel [a, *], Michael McCarrin [b]

[a] National Institute of Standards and Technology, USA
[b] Naval Postgraduate School, USA

Keywords:
Hash-based carving
Hashdb
bulk_extractor
Sector hashing
Similarity

A B S T R A C T

Hash-based carving is a technique for detecting the presence of specific "target files" on digital media by evaluating the hashes of individual data blocks, rather than the hashes of entire files. Unlike whole-file hashing, hash-based carving can identify files that are fragmented, files that are incomplete, or files that have been partially modified. Previous efforts at hash-based carving have looked for evidence of a single file or a few files. We attempt hash-based carving with a target file database of roughly a million files and discover an unexpectedly high false identification rate resulting from common data structures in Microsoft Office documents and multimedia files. We call such blocks "non-probative blocks." We present the HASH-SETS algorithm that can determine the presence of files, and the HASH-RUNS algorithm that can reassemble files using a database of file block hashes. Both algorithms address the problem of non-probative blocks and provide results that can be used by analysts looking for target data on searched media. We demonstrate our technique using the *bulk_extractor* forensic tool, the *hashdb* hash database, and an algorithm implementation written in Python.

Introduction

It is common for forensic practitioners to use databases of cryptographic hashes to search for known files. For example, some law enforcement organizations maintain databases of hash values of illegal images and videos. When media is obtained in a case, every file is cryptographically hashed and those hashes are compared to the hash database. Matches indicate the presence of a target file.

"Hash-based carving" is an alternative approach that relies on comparing hashes of physical sectors of the media to a database of hashes created by hashing every block of the target files (Collange et al., 2009b). One use-case is searching for child pornography: a block-hash database developed from a corpus of objectionable content should allow investigators to readily detect fragments of movies or still images on a storage device, even if the files have been deleted and partially overwritten. Sector hashing should also identify files that have been slightly modified—for example, files that have had a few bytes of random data appended for the explicit purpose of defeating file hashing (as may be done by an anti-forensics tool). Sector hashing can also be combined with random sampling to statistically sample the searched media, producing a high probability of finding target data within a relatively short amount of time. Finally, hash-based carving should also be able to find sectors of files in virtual memory swap files.

Although there has been some interest in hash-based carving in recent years, the technique is not widely used, and we are aware of no published algorithm describing how to assemble files from a database of sectors and sector hashes.

* Corresponding author.
E-mail addresses: simsong@acm.org (S.L. Garfinkel), mrmccarr@nps.edu (M. McCarrin).

http://dx.doi.org/10.1016/j.diin.2015.05.001

DFRWS - July 2015

And that was that, or so I thought

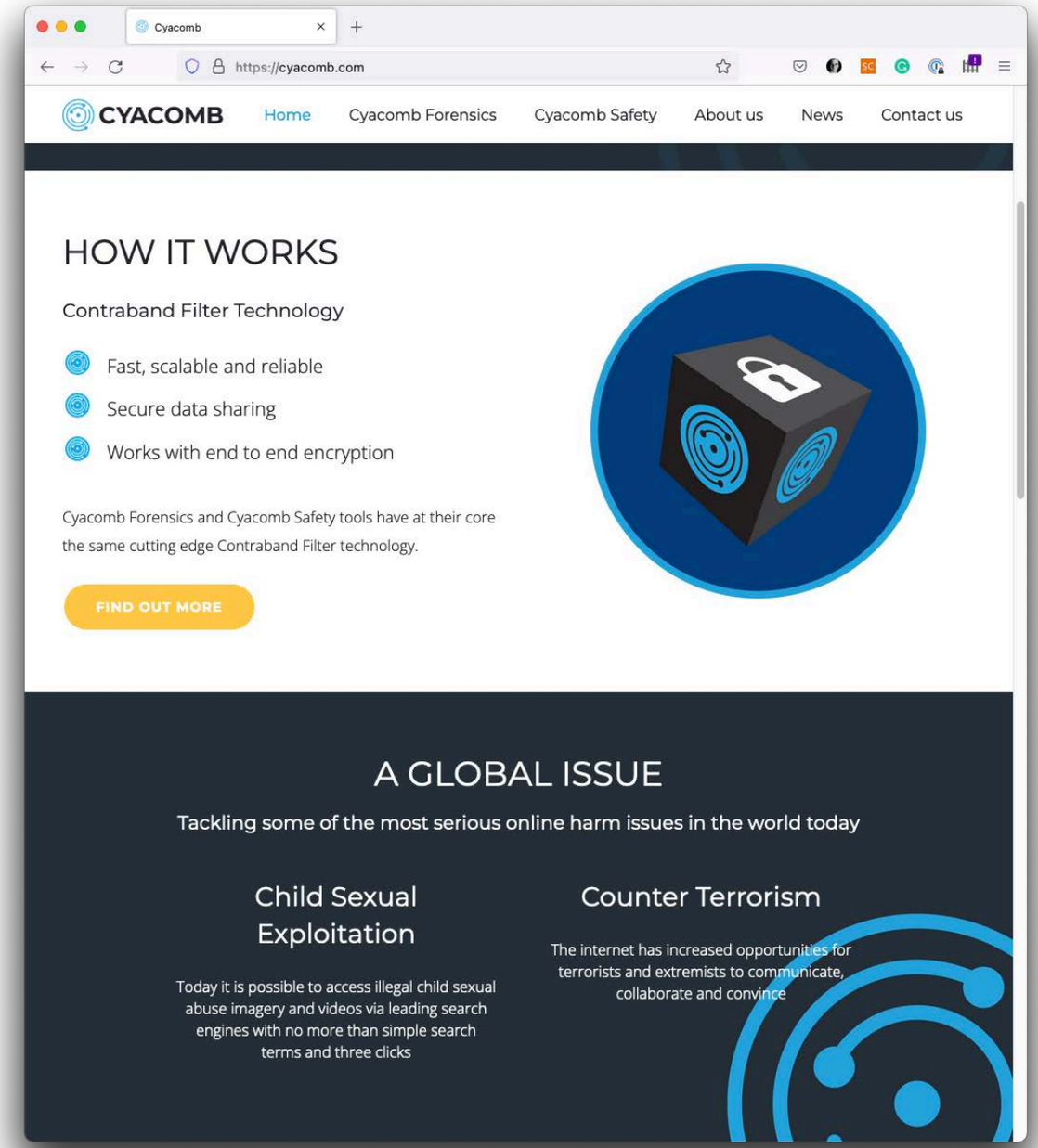
2017 - Cyacomb founded (Edinburgh, UK)

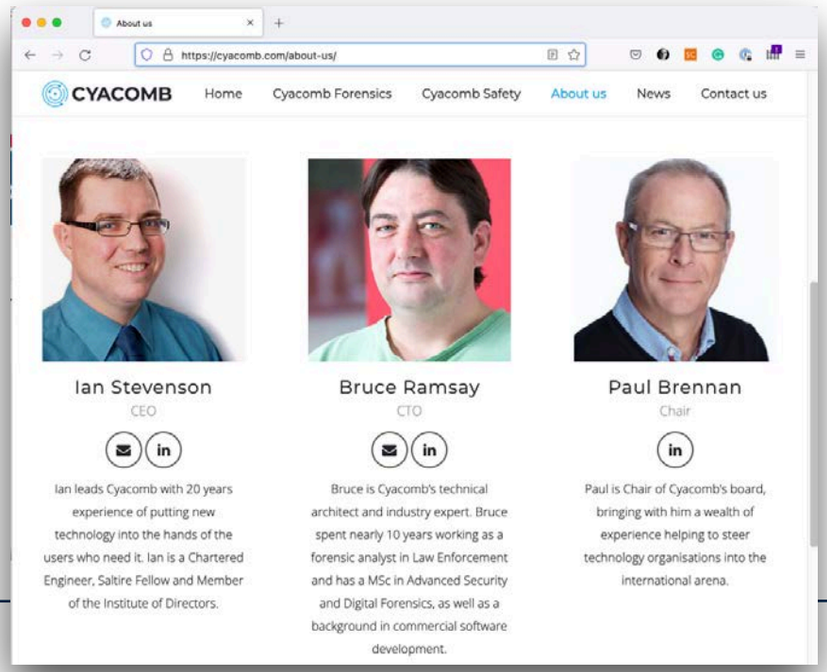
Bruce Ramsay

- Former police officer & software developer
- Moved to Edinburgh Napier University
- Had been working on small block forensics

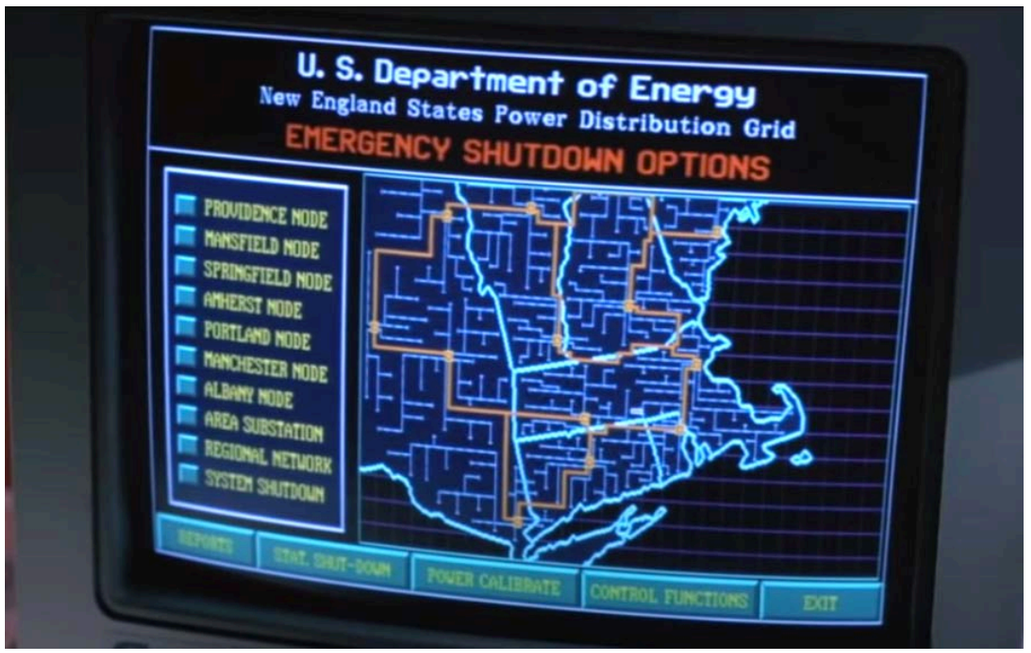
Received grant to commercialize

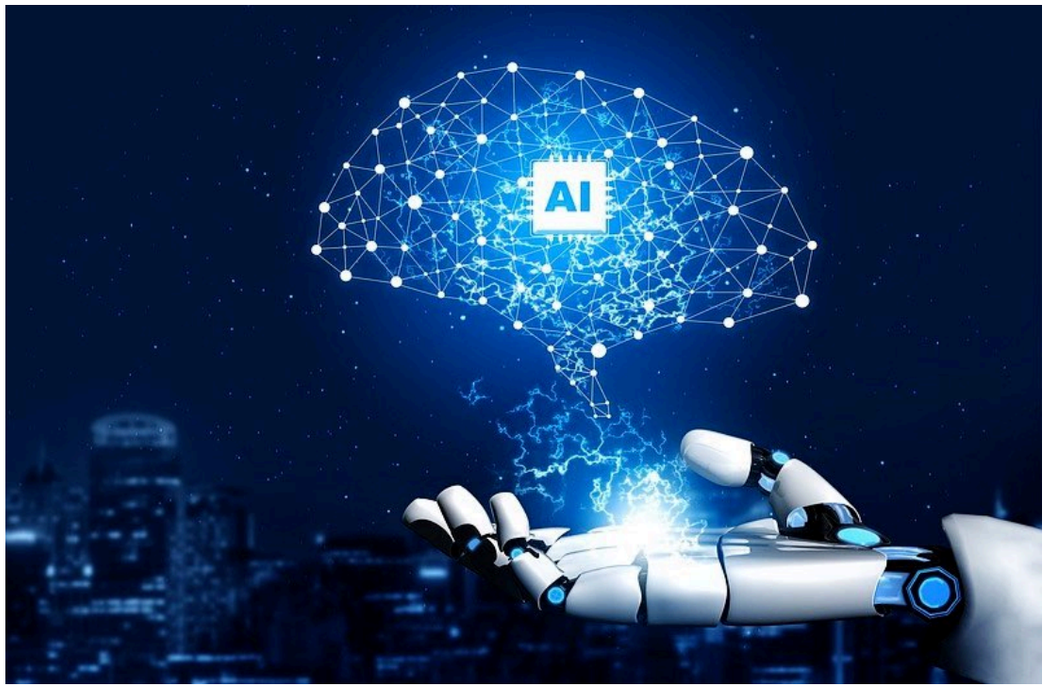
- First approached banks for loss prevention
- Moved to CSAM
- Partners in UK and Canada
- Limiting factor: access to datasets





Up Next:
Digital Forensics — Future History and Research Agenda





Hour 3
Digital Forensics:
A Future History and Research Agenda

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

October 26, 2011

http://afflib.org/

https://www.youtube.com/watch?v=nt7-WKXL5vw

"Difficult to see. Always in motion is the future."

Digital Forensics: A Future History and Research Agenda

Scenario #1 — Digital Forensics ~~Parity~~ Dominance
Digital Forensics is on par with other forensics.

Every photo has an ID from which law enforcement can determine:

- Photographer • Camera • Phone • GPS • Time • Integrity

https://www.pexels.com/photo/person-sitting-on-mountain-cliff-1659438/

M Venter

ID1281148234-823100:04

Every encrypted object in the cloud can be decrypted (with a warrant)

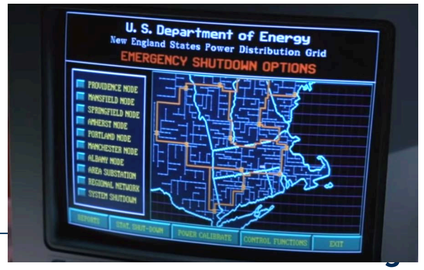
Every system can be entered by law enforcement (with a warrant)

Every deleted file is recoverable

Every police department has access to every tool

Cryptocurrency is fast, cheap, secure, and traceable.

"Sneakers" (1992)

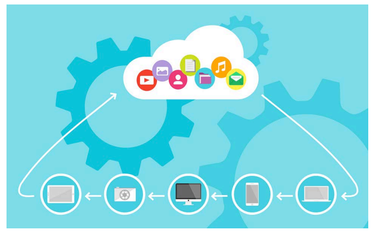


This vision is not technically achievable

Scenario #2 — Digital Forensics Devolution
Digital Forensics becomes records requests because devices are inaccessible.

Mobile phones are uncrackable

- Information stored in the cloud is retrieved by warrants
- Consent searches are not sufficient to sustain vendors; extraction and analysis tools quickly become out of date



https://pixabay.com/illustrations/mobile-security-privacy-protected-3469818/
https://pixabay.com/vectors/cloud-computing-cloud-device-data-1989339/

Voice assistants are trustworthy

- No more asking Amazon, Google and Apple for accidentally recorded voice

Surveillance is from public cameras and tower dumps

- High-powered AI indexes everything

https://www.pexels.com/photo/white-2-cctv-camera-mounted-on-black-post-under-clear-blue-sky-96612/

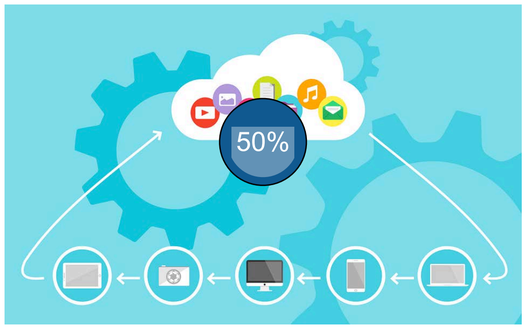
https://www.freeimages.com/photo/tower-4-1314235

This vision ignores our real capabilities with digital forensics today

Scenario #3 — Capricious Digital Forensics
Digital forensics is applied unpredictability, when it is politically warranted.

Some mobile phones are sometimes crackable

- Sometimes investigators can get the passcode
- Some vendor tools can crack some phones some of the time



https://pixabay.com/illustrations/mobile-security-privacy-protected-3469818/
https://pixabay.com/vectors/cloud-computing-cloud-device-data-1989339/

IoT devices are crackable, but it takes time

- It's faster to get data from the cloud

Forensics becomes a high-power tool that is used on high-profile, politically charged cases

This vision is pretty much where we are today. Is the future just more of the same?

Scenario #4 — Routine Digital Forensics
Forensics are effective, low-cost, and fair

Most crimes are resolved by records requests

- But we can unlock any phone found at a crime scene
- Yet somehow, criminals do not have the ability to unlock phones
- And somehow, our phones aren't unlocked when we travel to China or Russia

iPhone X Fully
Unlocked Space
Gray 256GB

GRAYSHIFT

https://appleinsider.com/articles/21/04/21/signal-hacks-cellebrite-device-reveals-vulnerabilities-and-potential-apple-copyright-concerns

Pick your future

Scenario 1 - Digital Forensics Dominance

Scenario 2 - Digital Forensics Devolution

Scenario 3 - Capricious Digital Forensics

Scenario 4 - Routine Digital Forensics

To do this, we must:

1. Revise the digital forensics process
2. Develop privacy-preserving digital forensics approaches
3. Improve the reliability and reproducibility of our tools
4. Apply AI to digital forensics, and digital forensics to AI

Revisiting the Digital Forensics Process

Photo by cottonbro from Pexels

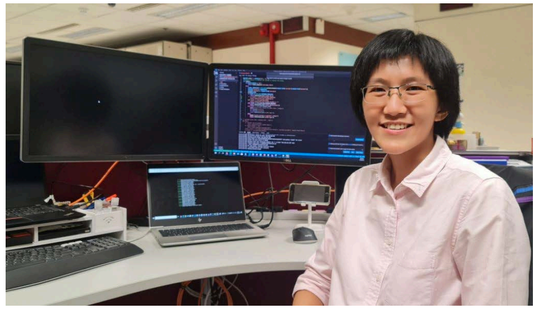
The traditional Digital Forensics Process is designed to make *digital evidence* available for [legal] decisions.

Preparation: policy, training & tools

https://www.legalcheek.com/2018/06/how-mobile-phones-are-helping-forensic-scientists-catch-murderers/

Collect & preserve evidence

Extract preserved evidence



https://govinsider.asia/resilience/women-in-cyber-solving-crimes-with-digital-forensics-yestine-goh/

Analyze the extract

Reporting & testimony

—Most developed 1999-2004

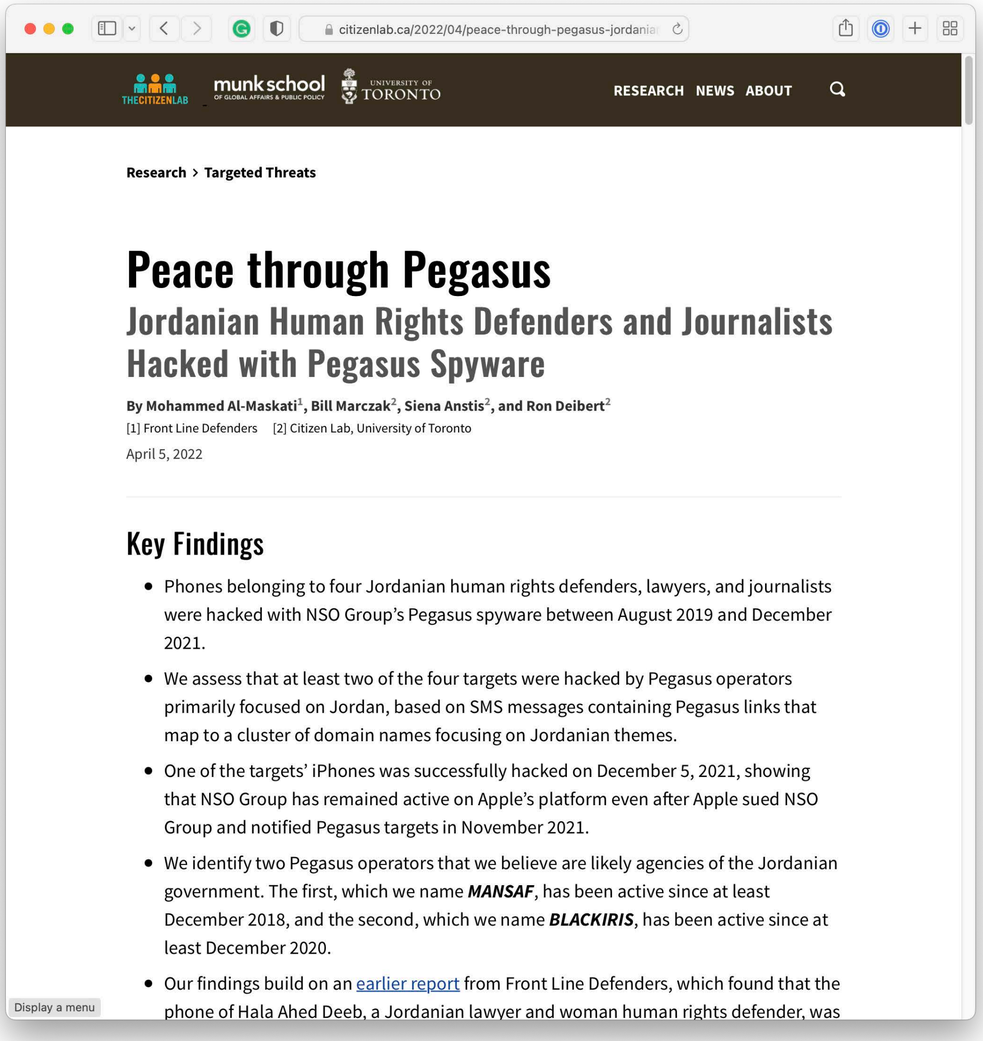
Today, digital forensics has many uses; each has a different process

1 - Digital forensics for legal proceedings

2 - Digital Forensics for intelligence

3 - Digital Forensics Incident Response

4 - Digital Forensics for privacy research



https://citizenlab.ca/2022/04/peace-through-pegasus-jordanian-human-rights-defenders-and-journalists-hacked-with-pegasus-spyware/

1 - Traditional digital forensics for legal process

Bringing *evidence* from *digital devices* into the *legal process*

- "Evidence refers to information or objects that may be admitted into court for judges and juries to consider when hearing a case."

  —*US National Institute of Justice,*

  https://nij.ojp.gov/topics/forensics/evidence-analysis-and-processing

- "Digital evidence is information stored or transmitted in binary form that may be relied on in court."

  https://nij.ojp.gov/digital-evidence-and-forensics

"Digital evidence is commonly associated with electronic crime, or e-crime, such as child pornography or credit card fraud."

"However, digital evidence is now used to prosecute all types of crimes, not just e-crime. For example, suspects' e-mail or mobile phone files might contain critical evidence regarding their intent, their whereabouts at the time of a crime and their relationship with other suspects."

https://nij.ojp.gov/digital-evidence-and-forensics

Digital forensics for legal process: challenges and solutions

Backlog

- # of cases

Data overload

- # of devices, # of TB, Cloud Storage

Knowing when to stop looking for…

- Case data — how much do you need to make the case?
- Exculpatory data — how hard do you look? (Depends on jurisdiction)

Correlating information

- Seized media & devices
- Cloud — personal storage, social media, etc
- Open Source Intelligence

Solution: Using AI for digital forensics

We need tools that can:

- Intelligently summarize 10TB of data
- Search and report relevant evidence

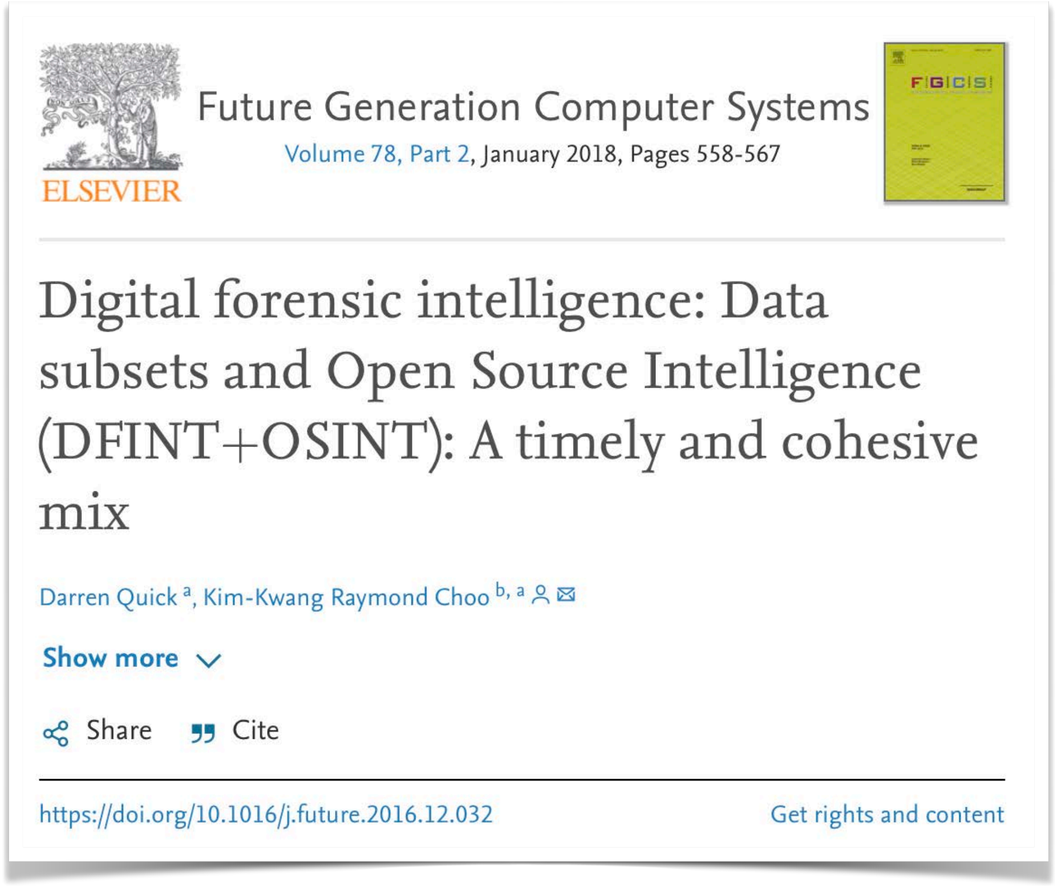
2 - Digital Forensics for Intelligence

Similar to digital forensics for legal process, except:

- Goal is to build a comprehensive subject model
- Combines Digital Forensics with Open Source Intelligence (OSINT) and other intelligence disciplines.
- Requires advances in Identity Intelligence and Identity Resolution
- Social Network Intelligence

Requires advances in:

- Knowledge Representation
- Semantic reasoning
- Data fusion



https://www.sciencedirect.com/science/article/abs/pii/S0167739X16308639

3 - Digital Forensics [for] Incident Response (#DFIR)

Digital Forensics Incident Response

- “Due to the proliferation of endpoints and an escalation of cybersecurity attacks in general, DFIR has become a central capability within the organization’s security strategy and threat hunting capabilities.”

  https://www.crowdstrike.com/cybersecurity-101/digital-forensics-and-incident-response-dfir/

Goals:

- “Respond to incidents”
- “Minimize data loss or theft”
- “Strengthen existing security protocols and procedures”
- “Recover from security events”
- [“Assist in the prosecution of the threat actor through evidence and documentation.”]

https://redcanary.com/resources/guides/incident-response-preparedness-guide

https://www.esentire.com/resources/library/gartner-market-guide-dfir-digital-forensics-incident-response

Most cyber crimes aren't reported. Most reported crimes aren't prosecuted.

Third Way report (2018):

- 300,000 malicious cyber incidents a year
- "Less than 1% of malicious cyber incidents see an enforcement action taken against the attackers."
- "Enforcement rate for reported incidents is 0.3%"
- "Taking into account that cybercrime victims often do not report cases, the effective enforcement rate may be closer to 0.05%"

In comparison

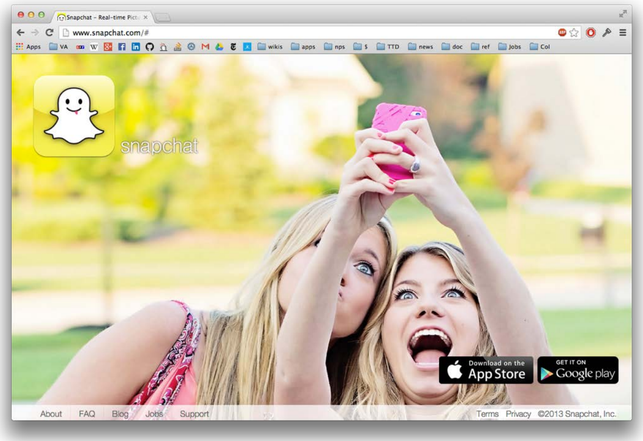
- 18% clearance rate for property crimes
- 46% clearance rate for violent crimes

—*FBI 2016 Uniform Crime Report*

https://www.thirdway.org/report/to-catch-a-hacker-toward-a-comprehensive-strategy-to-identify-pursue-and-punish-malicious-cyber-actors

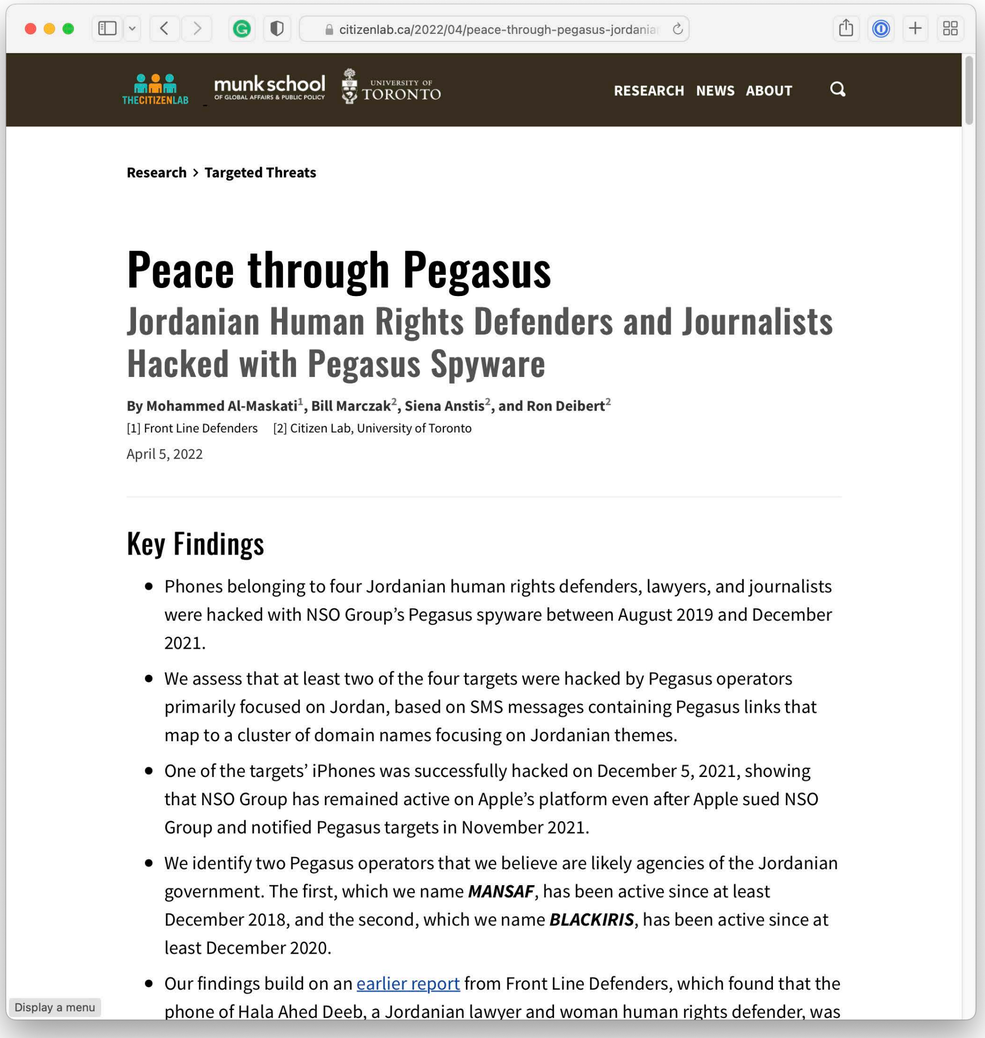
4 - Digital Forensics for Privacy Research

Much of what we know about privacy-violating malware comes from researchers using digital forensics.



SnapChat 2012

SnapChat 2013



https://citizenlab.ca/2022/04/peace-through-pegasus-jordanian-human-rights-defenders-and-journalists-hacked-with-pegasus-spyware/


Privacy-Preserving Digital Forensics

Privacy-Preserving Digital Forensics

Goals and Limitations:

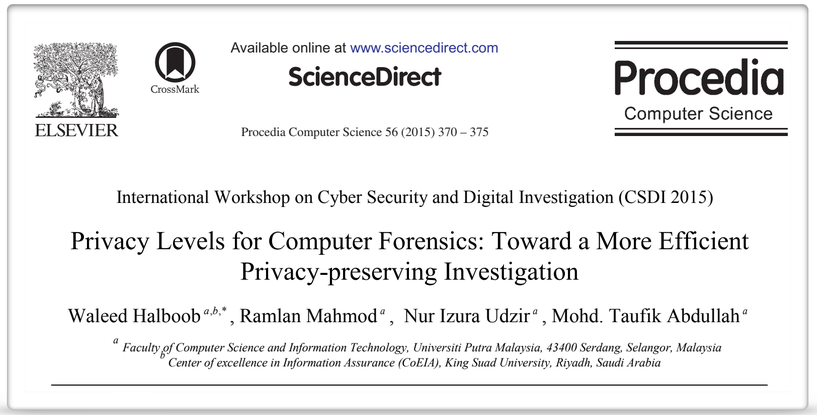
- Subject privacy: Minimization of material searched and scope of the investigation
- Investigation privacy: protection of search terms, investigators, and undercover officers

Foundational work in taxonomy:

- Goals, limitations, and standardized terms
- Alignment with legal frameworks

Possible Implementations:

- Private search & Private set intersection
- Secure multiparty computation
- Homomorphic encryption
- Differential Privacy



ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 56 (2015) 370 – 375

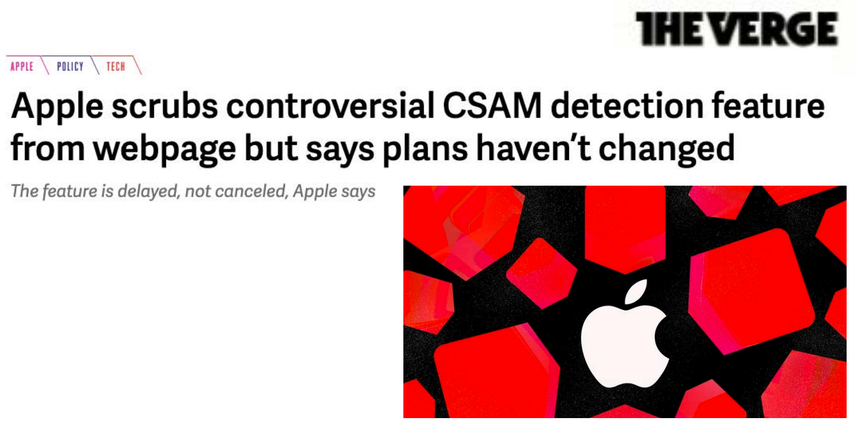
Procedia Computer Science

International Workshop on Cyber Security and Digital Investigation (CSDI 2015)

Privacy Levels for Computer Forensics: Toward a More Efficient Privacy-preserving Investigation

Waleed Halboob[a,b,*], Ramlan Mahmod[a], Nur Izura Udzir[a], Mohd. Taufik Abdullah[a]

[a] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[b] Center of excellence in Information Assurance (CoEIA), King Suad University, Riyadh, Saudi Arabia



THE VERGE

APPLE | POLICY | TECH

Apple scrubs controversial CSAM detection feature from webpage but says plans haven't changed

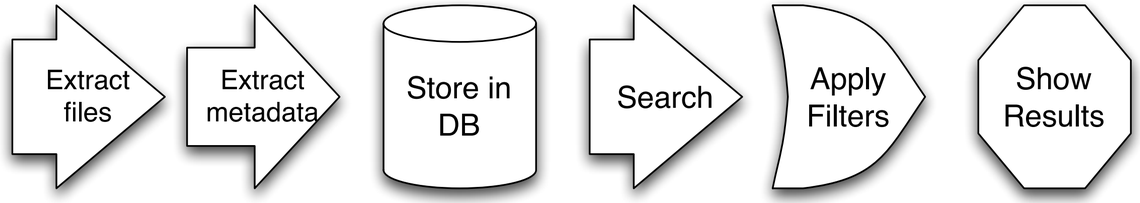
The feature is delayed, not canceled, Apple says

Turning Data Management into Knowledge Management

Image by Pexels from Pixabay

Digital forensic tools need a new model

Today's tools follow this model:



But:

- There's not enough time to extract all the data
- Tools can process 10TB of data, but they can't summarize it in a 4-page report

Increasingly, crimes won't be based on disks. They will be based on:

- Cryptocurrency ▪ Social Media ▪ Live Streaming ▪ Disinformation

Moving forward, we need tools that:

- Understand the investigation ▪ Autonomously seek out, acquire, and certify digital evidence
- Extract, represent and organize *facts,* with links back to supporting evidence
- Produce outputs in standardize machine-readable forms (for tool composition & validation)

practice

DOI:10.1145/3331166

Article development led by acmqueue
queue.acm.org

Five diverse technology companies show how it's done.

BY NATASHA NOY, YUQING GAO, ANSHU JAIN, ANANT NARAYANAN, ALAN PATTERSON, AND JAMIE TAYLOR

Industry-Scale Knowledge Graphs: Lessons and Challenges

KNOWLEDGE GRAPHS ARE critical to many enterprises today: They provide the structured data and factual knowledge that drive many products and make them more intelligent and "magical."

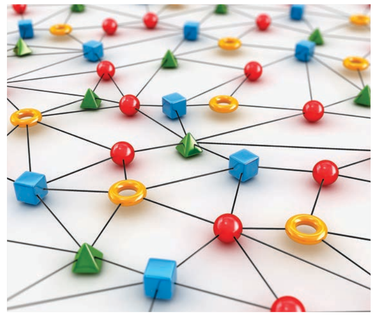
In general, a knowledge graph describes objects of interest and connections between them. For example, a knowledge graph may have nodes for a movie, the actors in this movie, the director, and so on. Each node may have properties such as an actor's name and age. There may be nodes for multiple movies involving a particular actor. The user can then traverse the knowledge graph to collect information on all the movies in which the actor appeared or, if applicable, directed.

Many practical implementations impose constraints on the links in knowledge graphs by defining a *schema* or *ontology*. For example, a link from a movie to its director must connect an object of type Movie to an object of type Person. In some cases the links themselves might have their own properties: a link connecting an actor and a movie might have the name of the specific role the actor played. Similarly, a link connecting a politician with a specific role in government might have the time period during which the politician held that role.

Knowledge graphs and similar structures usually provide a shared substrate of knowledge within an organization, allowing different products and applications to use similar vocabulary and to reuse definitions and descriptions that others create. Furthermore, they usually provide a compact formal representation that developers can use to infer new facts and build up the knowledge—for example, using the graph connecting movies and actors to find out which actors frequently appear in movies together.

This article looks at the knowledge graphs of five diverse tech companies, comparing the similarities and differences in their respective experiences of building and using the graphs, and discussing the challenges that all knowledge-driven enterprises face today. The collection of knowledge graphs discussed here covers the breadth of applications, from search, to product descriptions, to social networks:

- Both Microsoft's Bing knowledge graph and the Google Knowledge Graph support search and answering questions in search and during conversations. Starting with the descriptions and connections of people, places, things, and organizations, these graphs include general knowledge about the world.
- Facebook has the world's largest social graph, which also includes information about music, movies, celebrities, and places that Facebook users care about.



https://pixabay.com/illustrations/ai-artificial-intelligence-sci-fi-7111802/

The future is AI

We can’t keep up with crime today: AI is the only answer

The criminals are going to turn to AI!

AI-enabled crime requires AI-enabled forensics

Coming soon: forensics on AI-systems!

https://pixabay.com/illustrations/ai-artificial-intelligence-sci-fi-7111802/

Contact Information:

Simson Garfinkel
Senior Data Scientist, US Department of Homeland Security
simson.garfinkel@hq.dhs.gov