Automated Digital Forensics

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

April 25, 2011

http://simson.net/

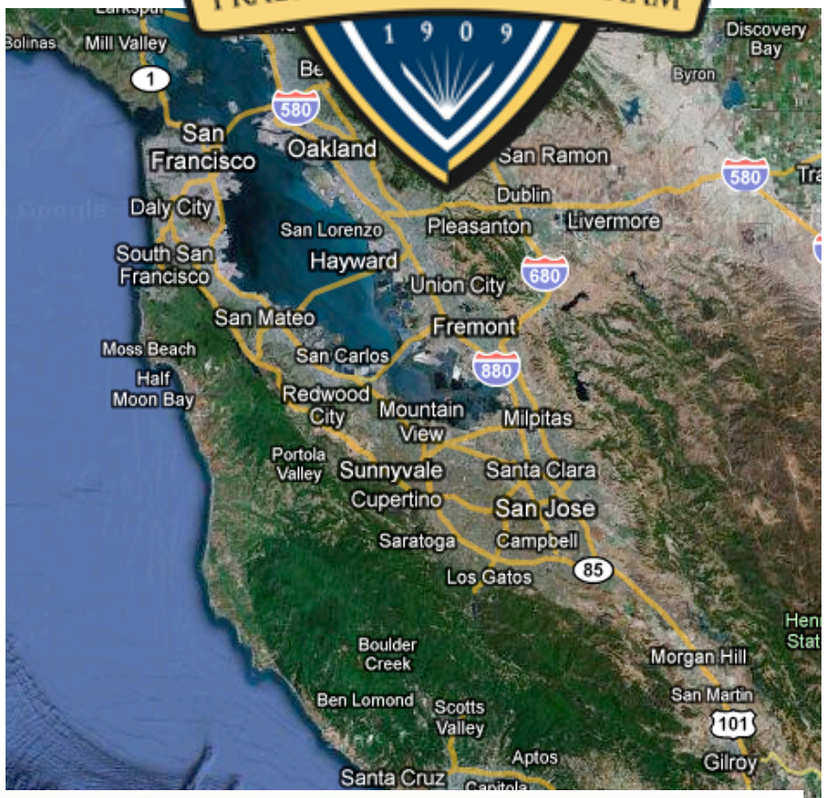
NPS is the Navy's Research University.

Location: Monterey, CA

Campus Size: 627 acres

Students: 1500

- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)



Schools:

- Business & Public Policy
- Engineering & Applied Sciences
- International Graduate Studies
- Operational & Information Sciences

The Digital Evaluation and Exploitation (DEEP) Group: Original research for trusted systems and forensics.

"Evaluation" — Profs. George Dinolt & Bret Michael

- Trusted hardware and software
- Cloud computing

"Exploitation" — Profs. Simson Garfinkel and Chris Eagle

- MEDEX — "Media" — Hard drives, camera cards, GPS devices.
- CELEX — Cell phone
- DOCEX — Documents
- DOMEX — Document & Media Exploitation

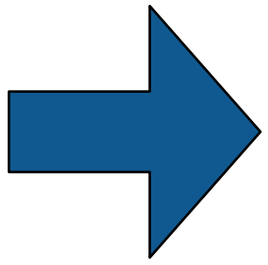
Typical sources includes:

- Law Enforcement
- Border searches
- Media collected on the "battlefield":
  - —*on combatants; houses & apartments*
- Cyber security (victims & attackers)

Traditionally forensics was used for *convictions*. Increasingly it's being used for *investigations.*

The goal was establishing possession of *contraband information.*

- Child Pornography
- Stolen documents.
- Hacker tools



Our research is aimed at using forensics as an *investigative tool*.

- Tracing information flow within an organization.
- Identifying a subject's:
  - —*contacts*
  - —*aliases*
  - —*pattern of life*
- Automatically identifying *actionable information.*

Given sufficient data, we can *automatically* assemble complex social network diagrams

We analyzed 2000 hard drives.

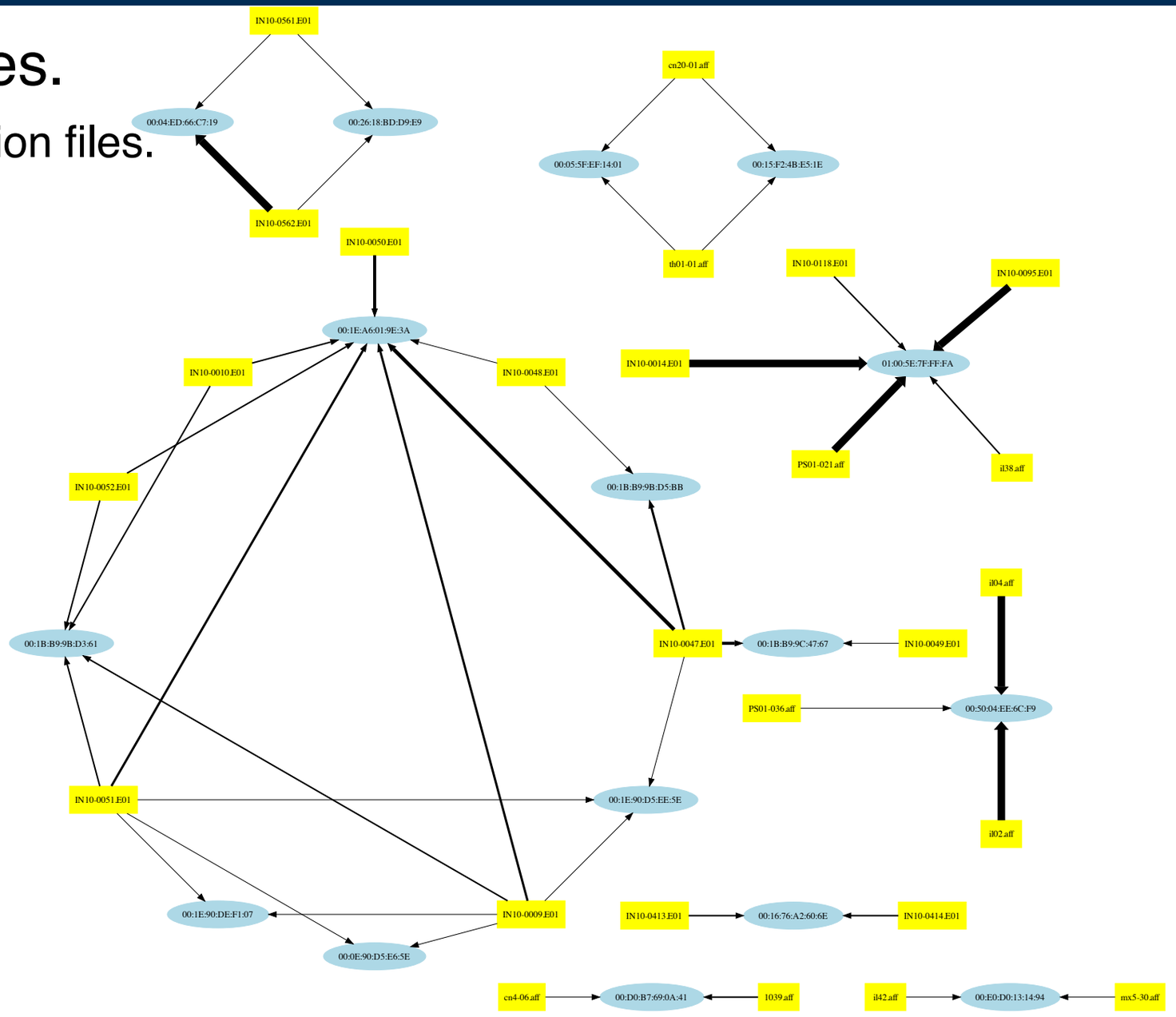
- Find IP packets in swap & hibernation files.
- Extract ethernet MAC addresses.

Post-processing identifies:

- Shared wireless routers.
- Common ethernet routers.

Validation:

- Reconstructed networks came from same organization.



—*Forensic Carving of Network Packets and Associated Data Structures,* Beverly & Garfinkel, DFRWS 2011, August 2011, New Orleans

Three principles underly our research.

Automation is essential.

- Today most forensic analysis is done manually.
- We are developing techniques & tools to allow automation.

Concentrate on the invisible.

- It's *easy* to wipe a computer….
- … but targets don't erase what they can't see.
- So we target:
  - *Deleted and partially overwritten files.*
  - *Fragments of memory in swap & hibernation.*
  - *Tool marks.*

Large amounts of data is essential.

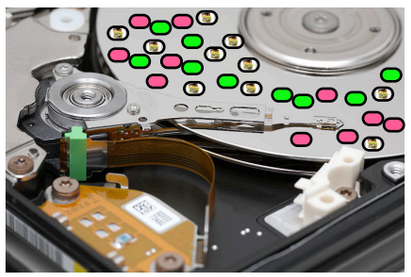
- We purchase used hard drives from all over the world.
- We manufacture data in the lab for use in education and publications.

This talk introduces digital forensics and presents two research projects from my lab.

Introducing Digital Forensics

Stream-based forensics

Random sampling for high speed forensics

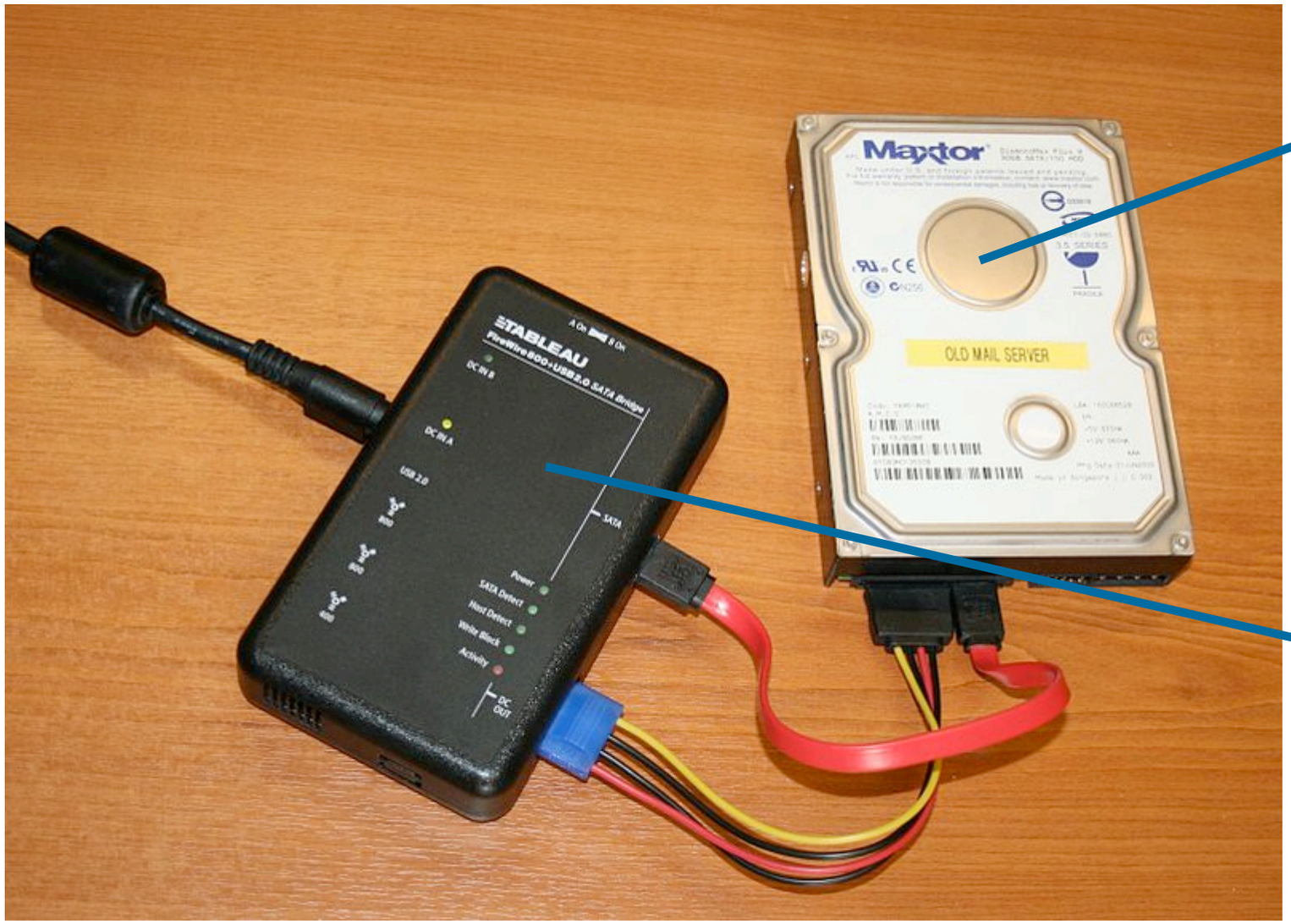


Creating forensic Corpora

Introducing Digital Forensics

Data extraction is the first step of forensic analysis

"Imaging tools" extract the data without modification.



"Write Blocker" prevents accidental overwriting.

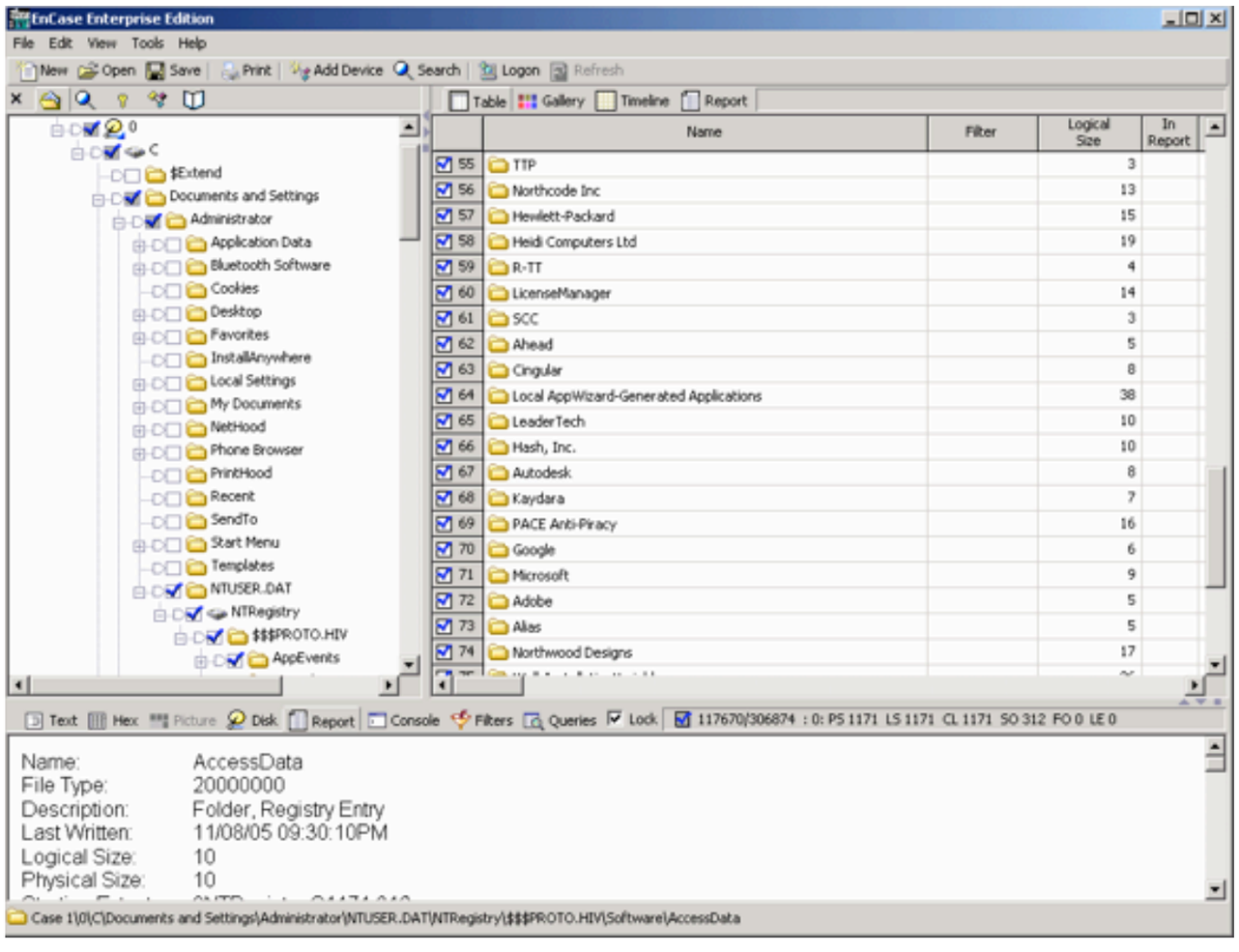
Original device stored in evidence locker.

Forensic copy ("disk image") stored on a storage array.

Examiners use digital forensic tools to view the evidence.

Today's tools allow the examiner to:

- Display of *allocated* & *deleted* files.
- String search.
- Data recovery and *file carving*.
- Examining individual disk sectors in hex, ASCII and Unicode



EnCase Enterprise by Guidance Software

The last decade was a "Golden Age" for digital forensics.

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

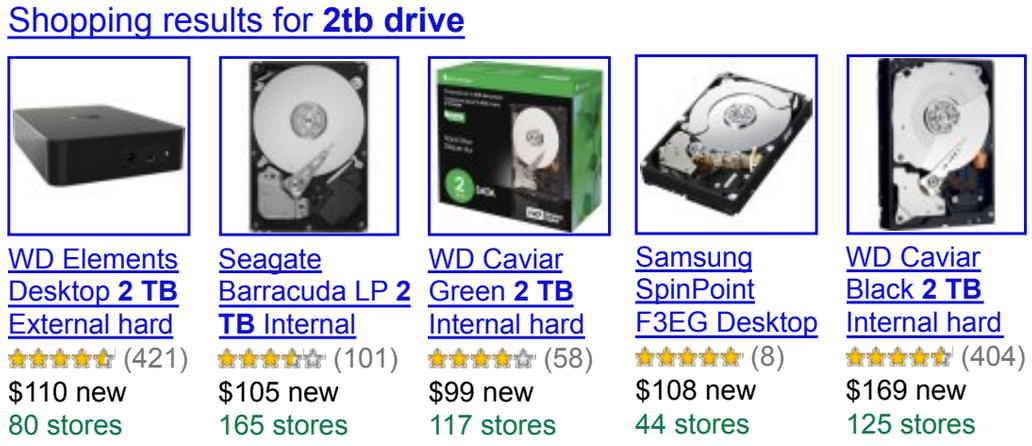
Today there is a growing digital forensics crisis.

We have identified 5 key problems.

Problem 1 - Increased cost of extraction & analysis.

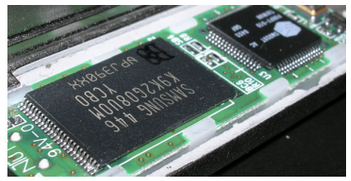
Data: too much and too complex!

- Increased size of storage systems.



- Cases now require analyzing multiple devices
  - *—2 desktops, 6 phones, 4 iPods, 2 digital cameras = 1 case*

- Non-Removable Flash



- Proliferation of operating systems, file formats and connectors
  - *—XFAT, XFS, ZFS, YAFFS2, Symbian, Pre, iOS,*

FBI Regional Computer Forensic Laboratories growth:

- Service Requests: 5,057 (FY08) ➔ 5,616 (FY09) (+11%)
- Terabytes Processed: 1,756 (FY08) ➔ 2,334 (FY09) (+32%)

Problem 2 — Cell phones pose special challenges

Data Extraction:

- No standard connectors.
- No standard way to copy data out.
- Difficult to image cell phones without changing them.
- Many phones can be remotely wiped.

Data Understanding:

- Data stored in proprietary formats.
- Vendors frequently change internal structures.

NIST's *Guidelines on Cell Phone Forensics:*

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

How do we analyze 100,000 apps?

Problem 3 — Encryption and Cloud Computing make it hard to get to the data

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

- Our only hope:
  - —*Browser caches & virtual memory… (for now)*

Problem 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

Problem 5 — Time is of the essence.

Most tools were designed to perform a complete analysis.

- Find all the files.
- Index all the terms.
- Report on all the data.
- Take as long as necessary!

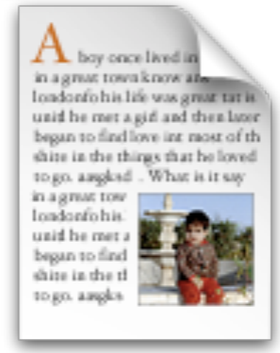
Increasingly we are racing the clock:

- Police prioritize based on statute-of-limitations!
- Battlefield, Intelligence & Cyberspace operations require turnaround in days or hours.
- Log files & data preservation.
  - —*Data may be wiped before you act.*

*Data **quality** makes digital forensics hard.*

Any piece of data may be critical.

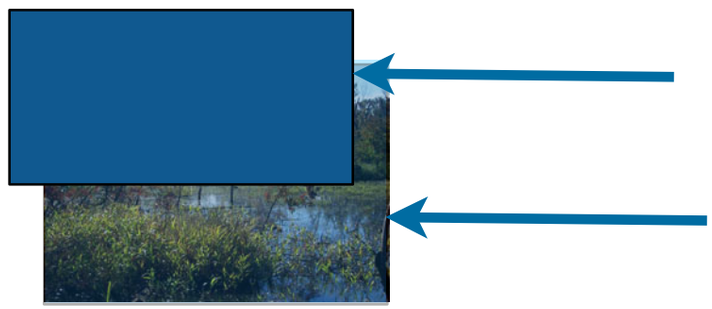
- Heterogeneity is a problem.
  - —*Address books*
  - —*Email*
  - —*Documents*
  - —*Photos*



- Each of these objects requires a different kind of analysis.

Frequently we are reading data *differently than intended.*

- Compressed data is not designed to be “recoverable” if the first half is missing.
- File systems not designed to permit “undeleting” files.
- Windows Hibernation files designed for single-use.



—*Computer Science lacks techniques for resolving corrupted data structures.*

*Data **quantity*** make digital forensics hard too!

Quantity: analysts have less time than the subject!

- User spent *years* assembling email, documents, etc.
- Analysts have days or hours to process it.

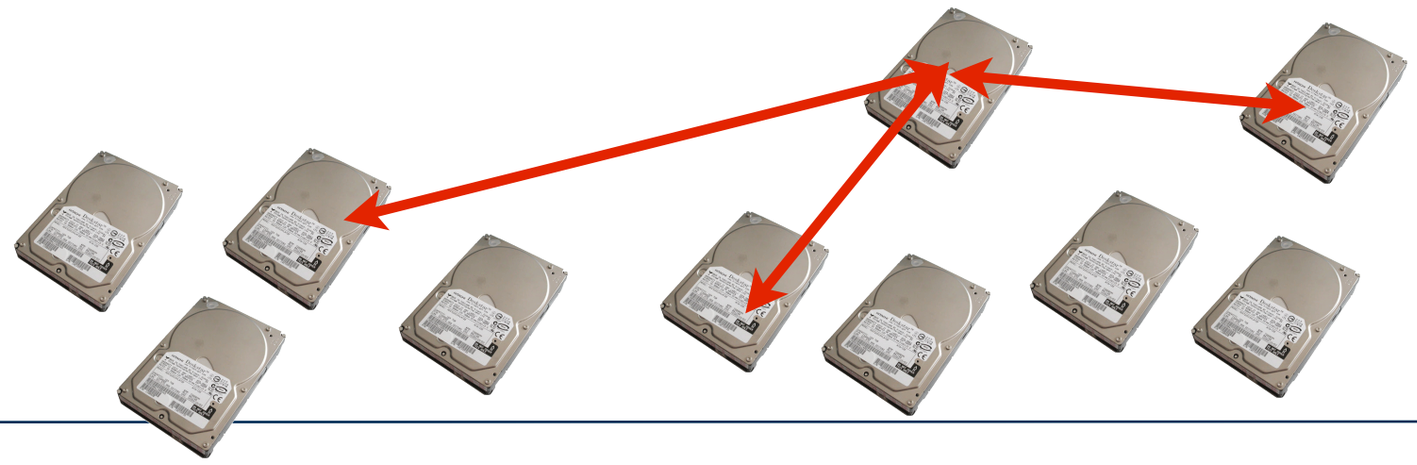
There is no resource advantage.

- Police analyze top-of-the-line systems … with top-of-the-line systems.
- National Labs have large-scale server farms … to analyze huge collections.

DF researchers must respond by developing new algorithms that:

—*Provide incisive analysis through cross-drive analysis.*

—*Operate autonomously on incomplete, heterogeneous datasets.*



Stream-based Forensics

Data on hard drives can be divided into three categories:

Resident Data

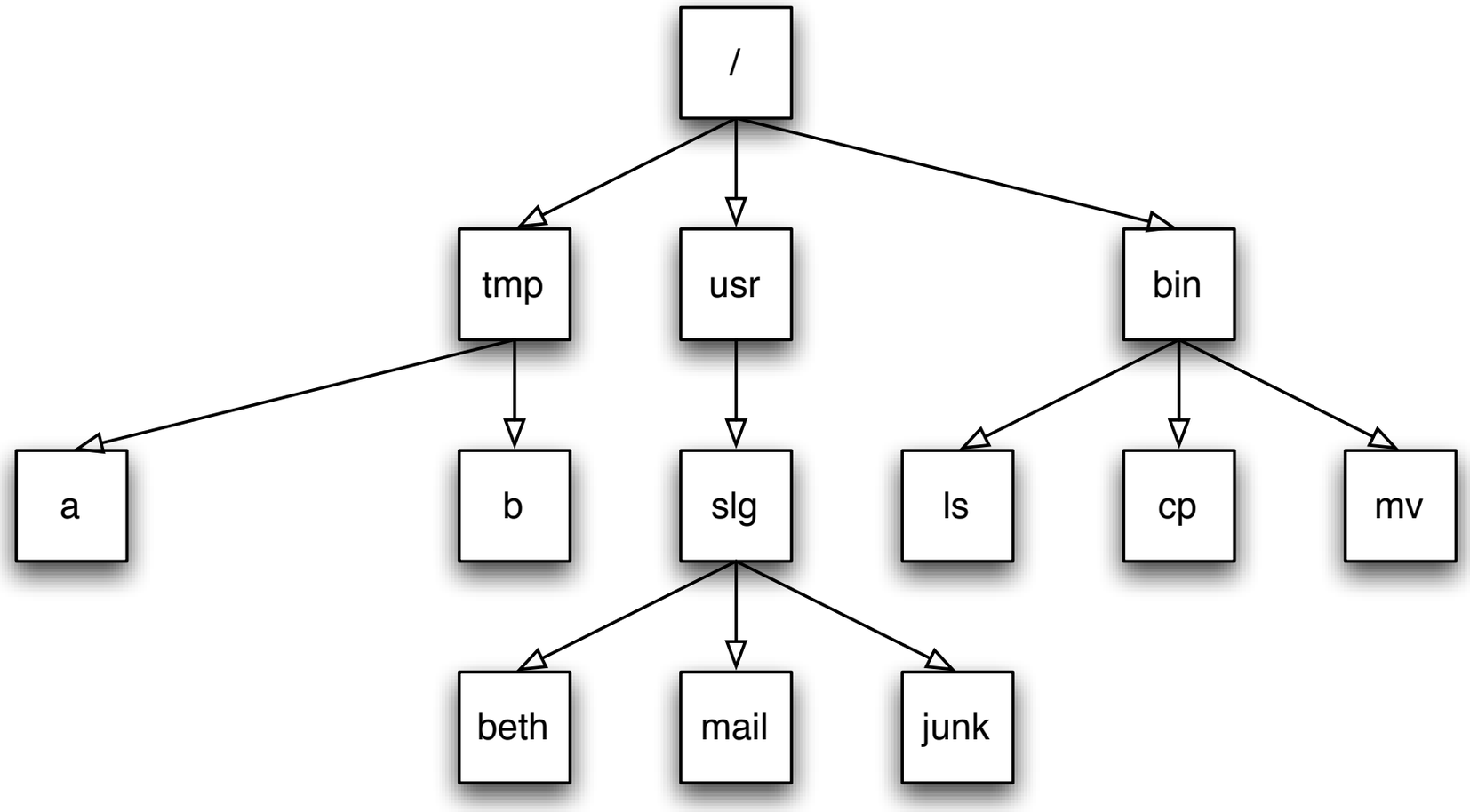
Deleted Data

No Data

user files
email messages
[temporary files]

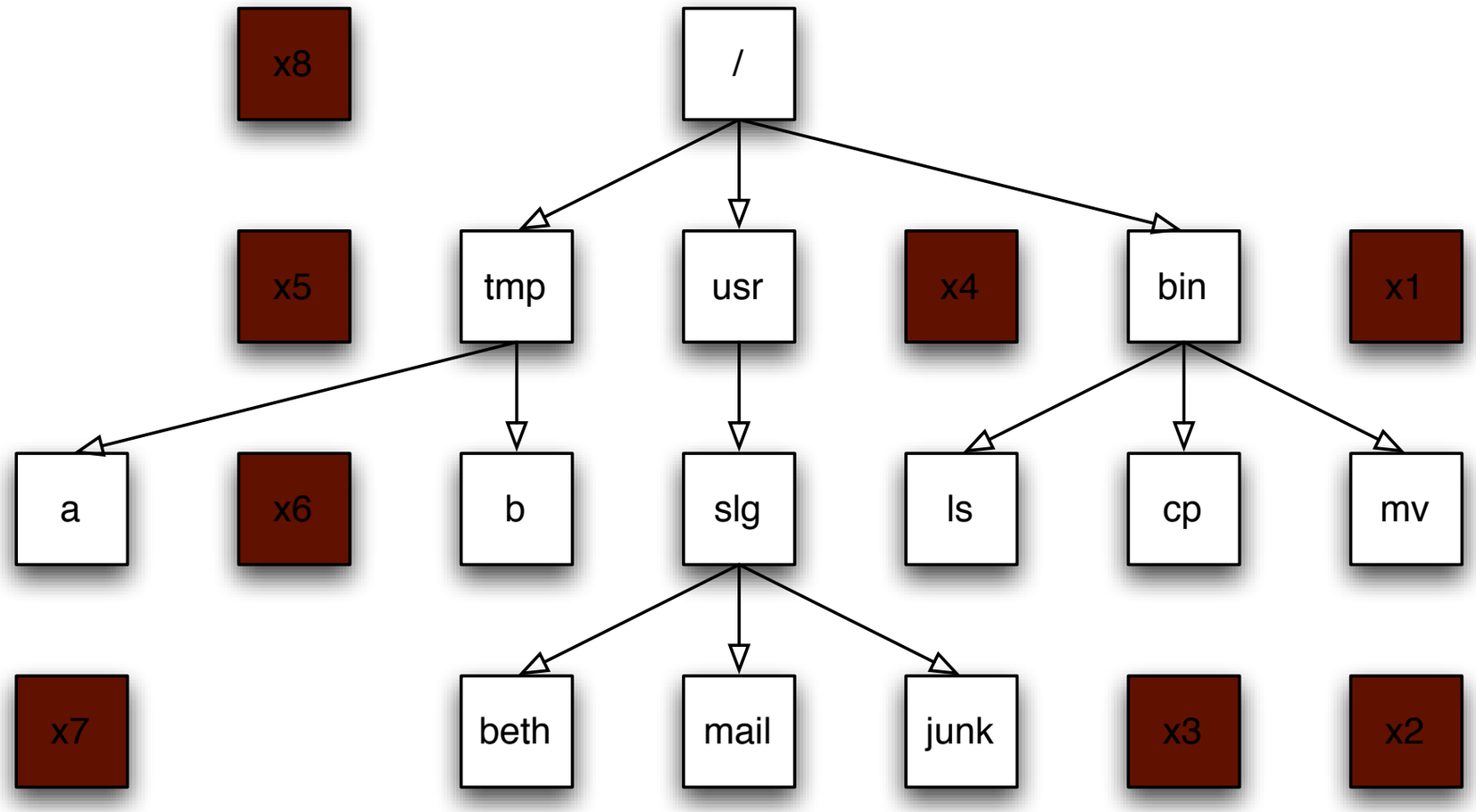
blank sectors

Resident data is the data you see from the root directory.



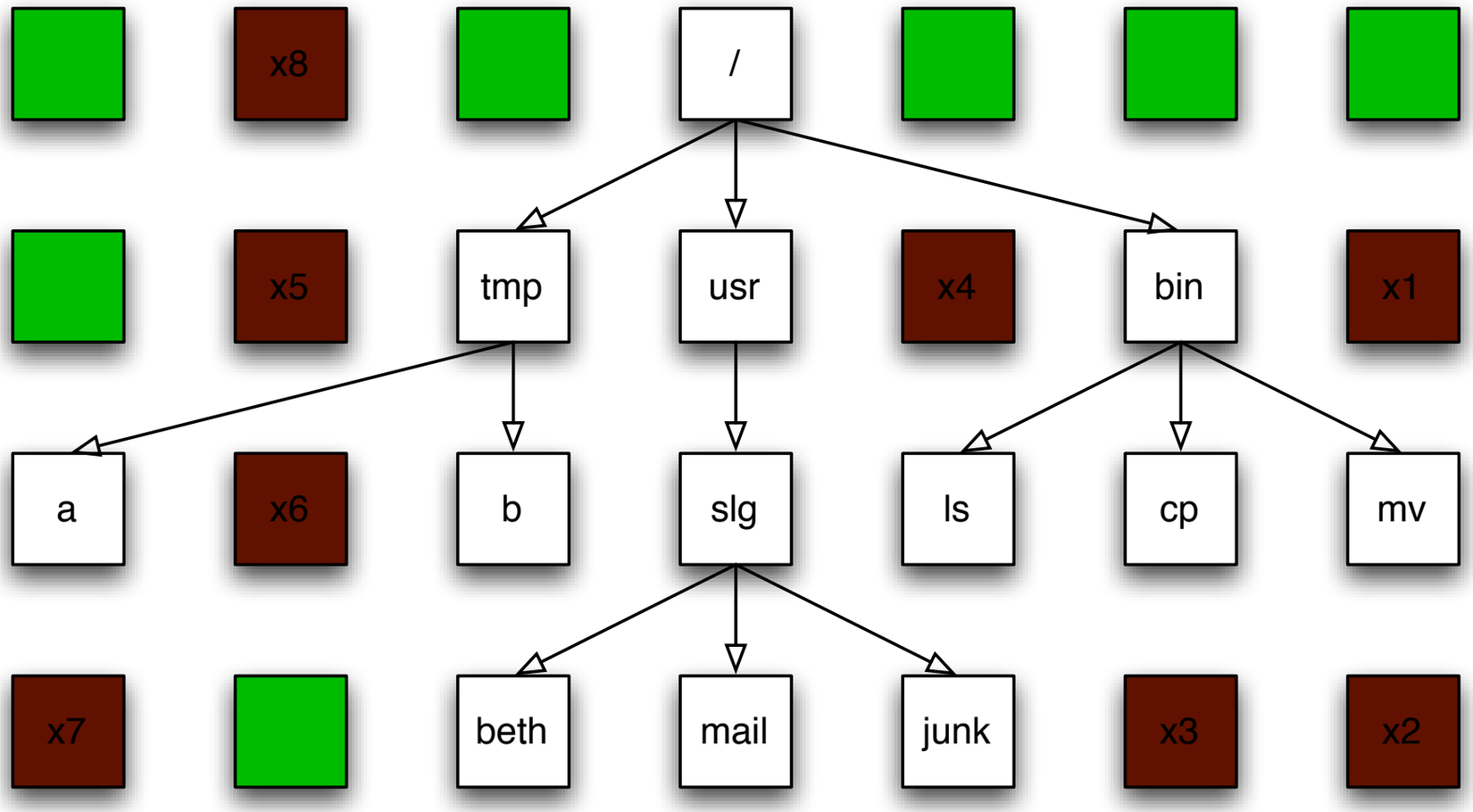
Resident Data

Deleted data is on the disk, but can only be recovered with forensic tools.



Deleted Data

Sectors with “No Data” are blank.



No Data

Today most forensic tools follow the same steps to analyze a disk drive.

Walk the file system to map out all the files (allocated & deleted).

For each file:

- Seek to the file.
- Read the file.
- Hash the file (MD5)
- Index file's text.

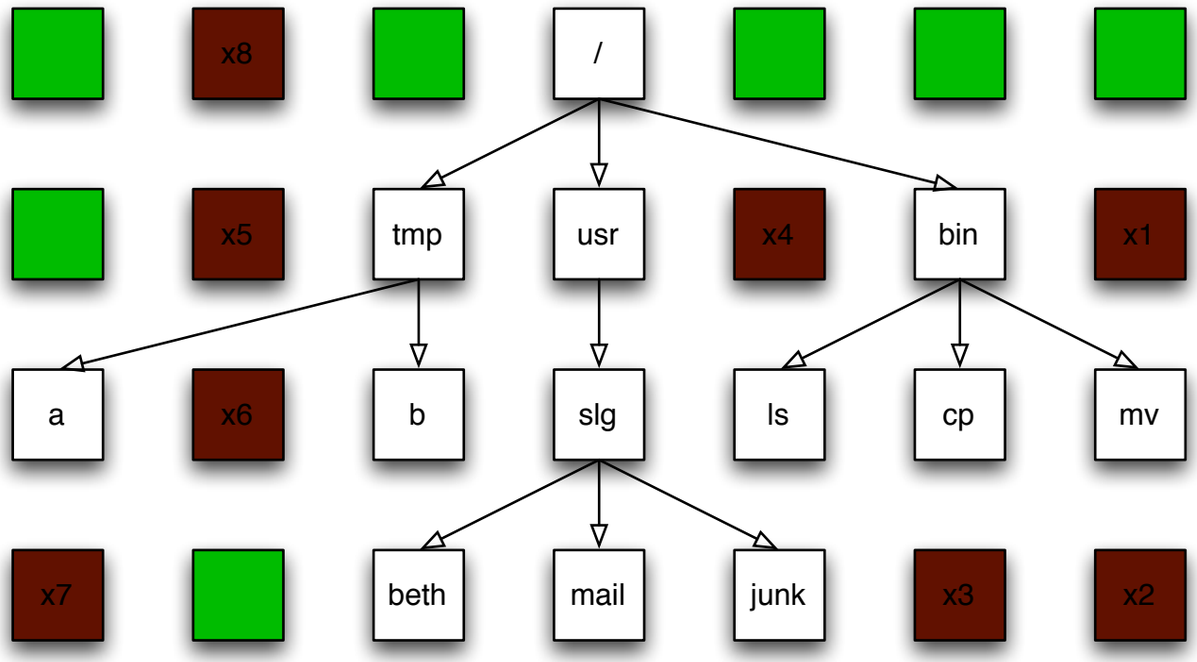
"Carve" space between files for other documents, text, etc.

Problem #1: Time

- 1TB drive takes 3.5 hours to read
  - —*10-80 hours to process!*

Problem #2: Completeness

- Lots of residual data is ignored.
  - —*Many investigations don't carve!*



Can we analyze a 1TB drive in 3.5 hours? (The time it takes to read the data.)

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

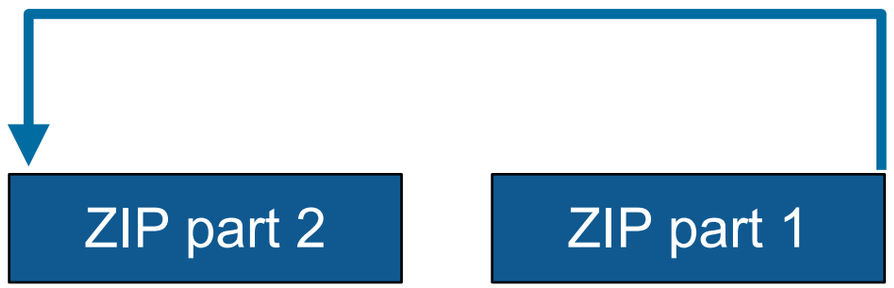
1. Read all of the blocks in order.
2. Look for information that might be useful.
3. Identify & extract what's possible in a single pass.

Advantages:

- No disk seeking.
- Read the disk at maximum transfer rate.
- Reads *all the data* — allocated files, deleted files, file fragments.

Disadvantages:

- Fragmented files won't be recovered:
  - —*Compressed files with part2-part1 ordering*
  - —*FIles with internal fragmentation (.doc)*
- A second pass may be needed to map contents to file names.

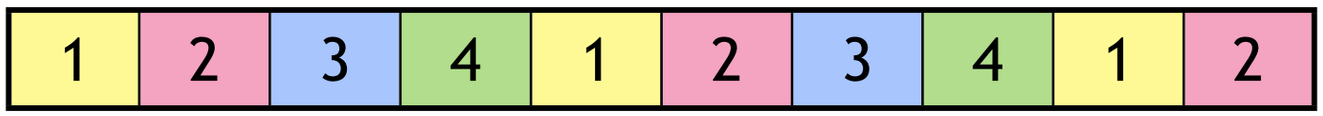


bulk_extractor: a high-speed disk scanner.

Written in C, C++ and Flex.

Key Features:

- Uses regular expressions and rules to scan for:
  - *email addresses; credit card numbers; JPEG EXIFs; URLs; Email fragments.*
- Recursively re-analyzes ZIP components.
- Produces a histogram of the results.
- Multi-threaded.
  - *Disk is "striped."*
  - *Results written out-of-order.*



Challenges:

- 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.

bulk_extractor output: text files of "features" and context.

email addresses from domexusers:

```
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("
```

Histogram:

```
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
```

bulk_extractor success:
City of San Luis Obispo Police Department, Spring 2010

District Attorney filed charges against two individuals:

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

Defendants:

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

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

- In 2.5 hours 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.

Eliminating false positives:
Many of the email addresses come with Windows!

Sources of these addresses:

- Windows binaries
- SSL certificates
- Sample documents

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**

It's important to suppress email addresses not relevant to the case.

Approach #1 — Suppress emails seen on many other drives.

Approach #2 — Stop list from bulk_extractor run on clean installs.

Both of these methods *white list* commonly seen emails.

- A problem — Operating Systems have a LOT of emails. (FC12 has 20,584!)
- Should we give the Linux developers a free past?

Approach #3: Context-sensitive stop list.

Instead of extracting just the email address, extract the context:

- Offset: **351373329**
- Email: **zeeshan.ali@nokia.com**
- Context: **ut_Zeeshan Ali <zeeshan.ali@nokia.com>, Stefan Kost <**

- Offset: **351373366**
- Email: **stefan.kost@nokia.com**
- Context: **>, Stefan Kost <stefan.kost@nokia.com>_____sin**

Here "context" is 8 characters on either side of feature.

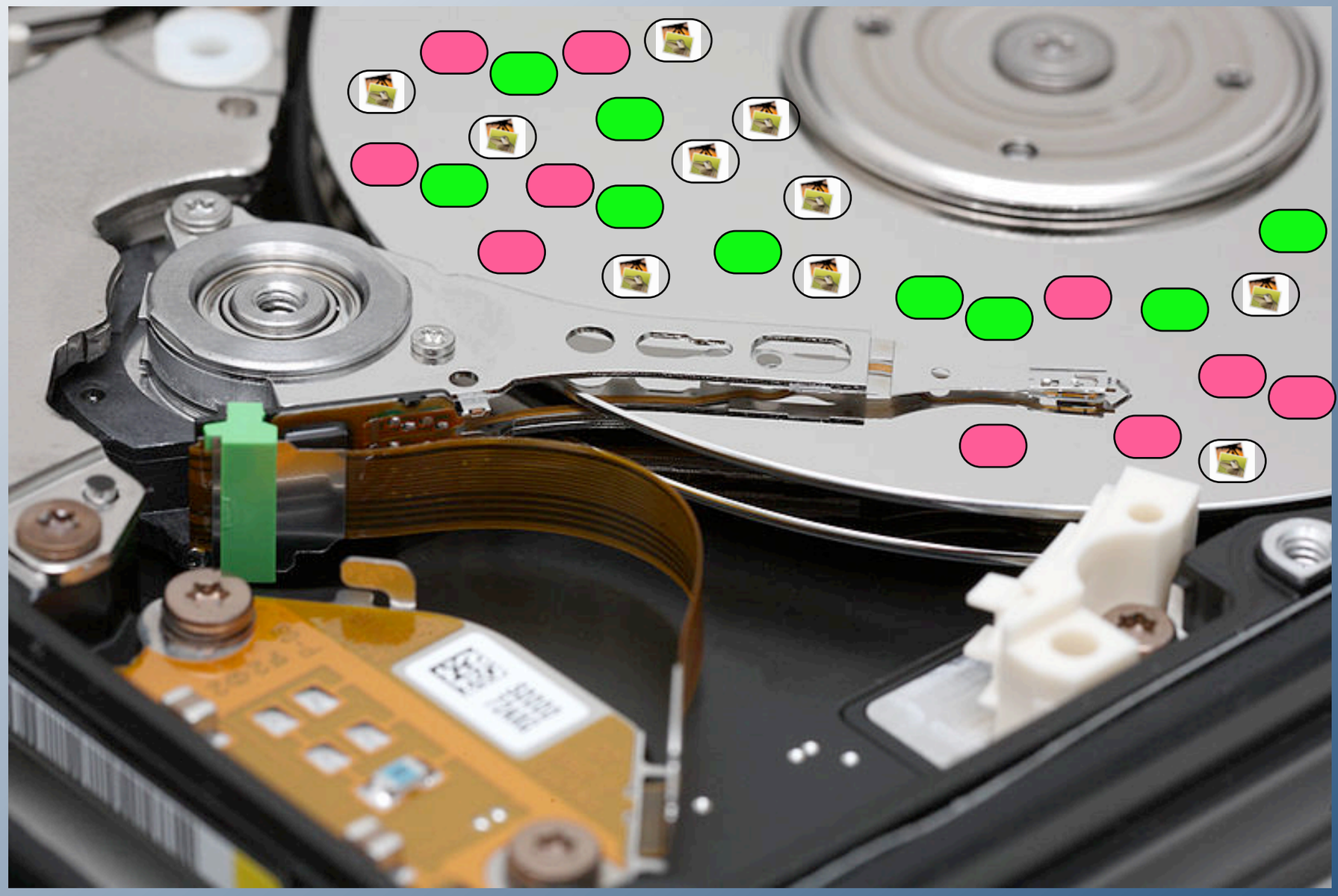
We created a context-sensitive stop list for Microsoft Windows XP, 2000, 2003, Vista, and several Linux.

Total stop list: 70MB (628,792 features)

Applying it to domexusers HD image:

- # of emails found: 9143 ➔ 4459

without stop list		with stop list	
n=579	domexuser1@gmail.com	n=579	domexuser1@gmail.com
n=432	domexuser2@gmail.com	n=432	domexuser2@gmail.com
n=340	domexuser3@gmail.com	n=340	domexuser3@gmail.com
n=268	ips@mail.ips.es	n=192	domexuser2@live.com
n=252	premium-server@thawte.com	n=153	domexuser2@hotmail.com
n=244	CPS-requests@verisign.com	n=146	domexuser1@hotmail.com
n=242	someone@example.com	n=134	domexuser1@live.com
n=237	inet@microsoft.com	n=91	premium-server@thawte.com
n=192	domexuser2@live.com	n=70	talkback@mozilla.org
n=153	domexuser2@hotmail.com	n=69	hewitt@netscape.com
n=146	domexuser1@hotmail.com	n=54	DOMEXUSER2@GMAIL.COM
n=134	domexuser1@live.com	n=48	domexuser1%40gmail.com@imap.gmail.com
n=115	example@passport.com	n=42	domex2@rad.li
n=115	myname@msn.com	n=39	lord@netscape.com
n=110	ca@digsigtrust.com	n=37	49091023.6070302@gmail.com



Random Samplings

Can we analyze a hard drive in five minutes?

US agents encounter a hard drive at a border crossings...

Searches turn up rooms filled with servers….

If it takes 3.5 hours to read a 1TB hard drive, what can you learn in 5 minutes?

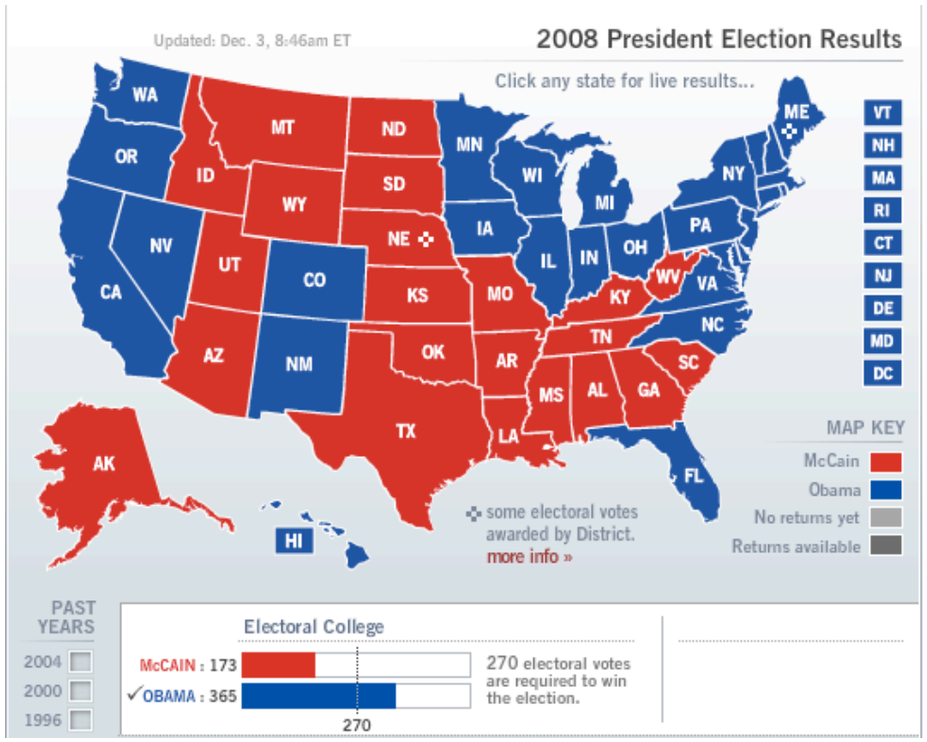
		
Minutes	208	5
Max Data	1 TB	36 GB
Max Seeks		90,000

36 GB is a lot of data!

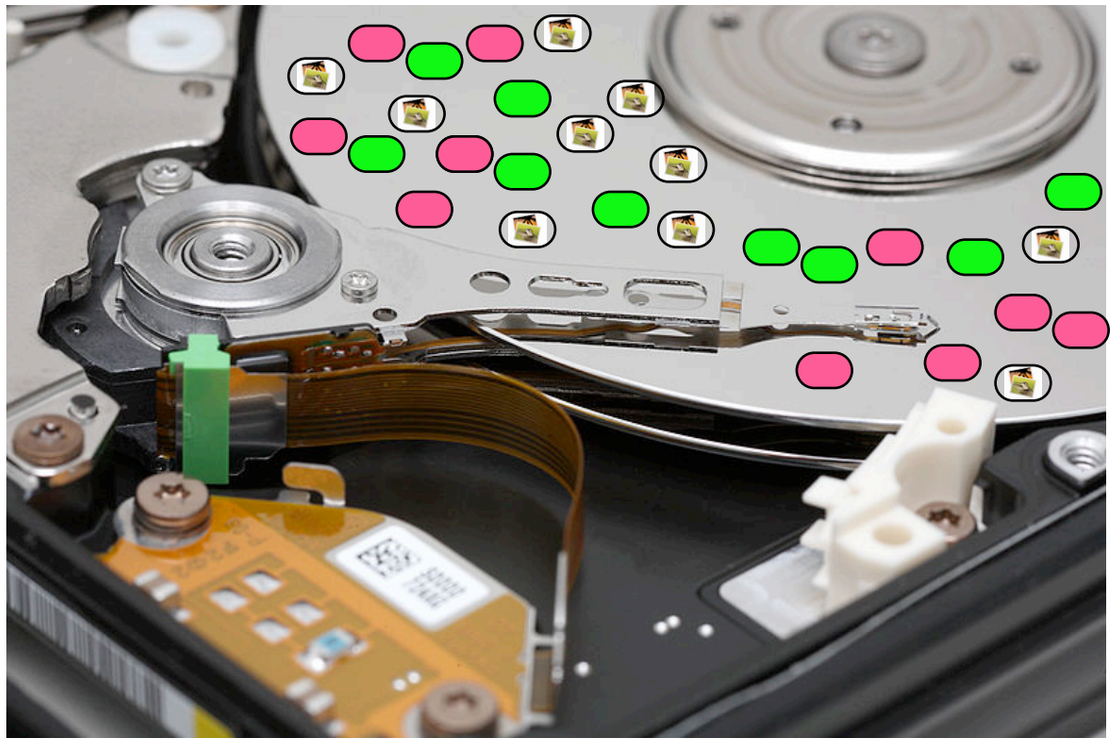
- ≈ 2.4% of the disk
- But it can be a *statistically significant sample*.

We can predict the statistics of a *population* by sampling a *randomly chosen sample.*

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



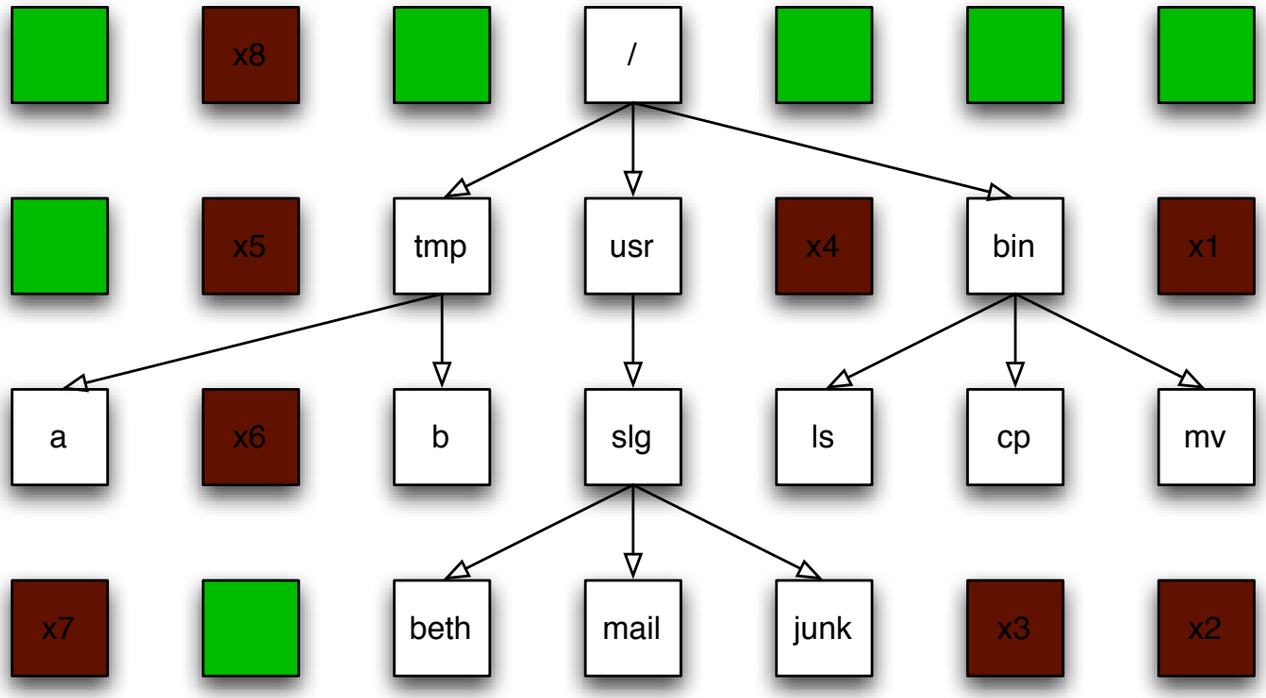
Hard drive contents can be predicted by sampling a few thousand sectors:

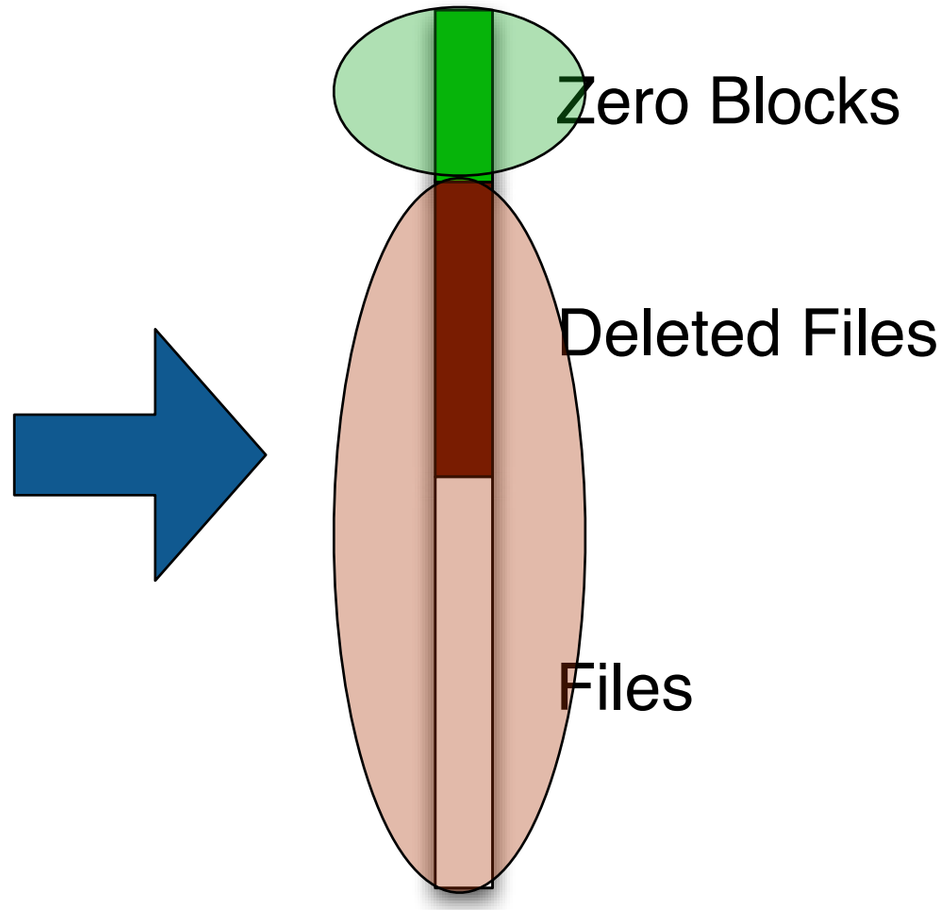


The challenge is identifying *likely voters.*

The challenge is *identifying the sectors* that are sampled.

Sampling can distinguish between "zero" and data.
It can't distinguish between resident and deleted.





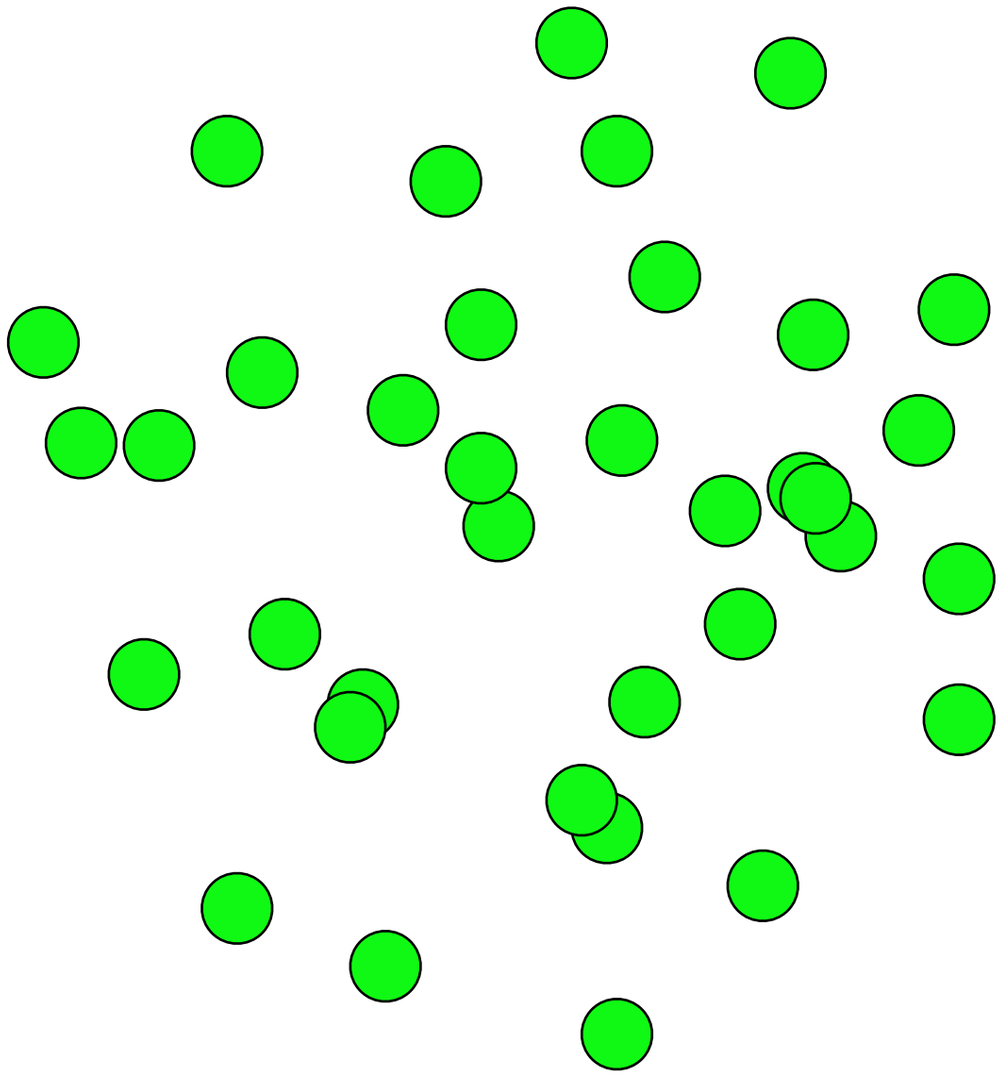
Simplify the problem.
Can we use statistical sampling to verify wiping?

Many organizations discard used computers.

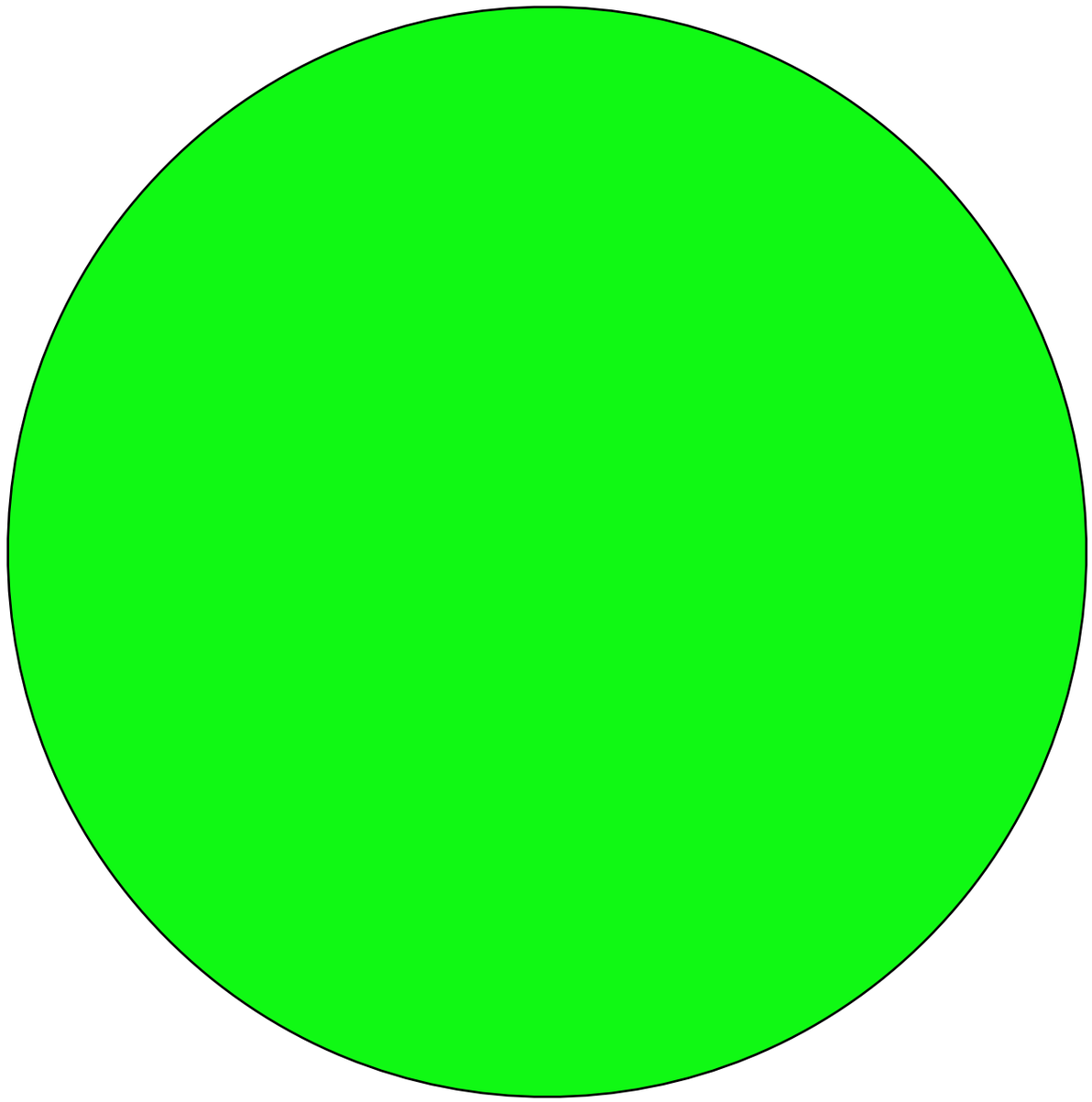
Can we verify if a disk is properly wiped in 5 minutes?

We read 10,000 randomly-chosen sectors … and they are all blank

We read 10,000 randomly-chosen sectors … and they are all blank



We read 10,000 randomly-chosen sectors … and they are all blank

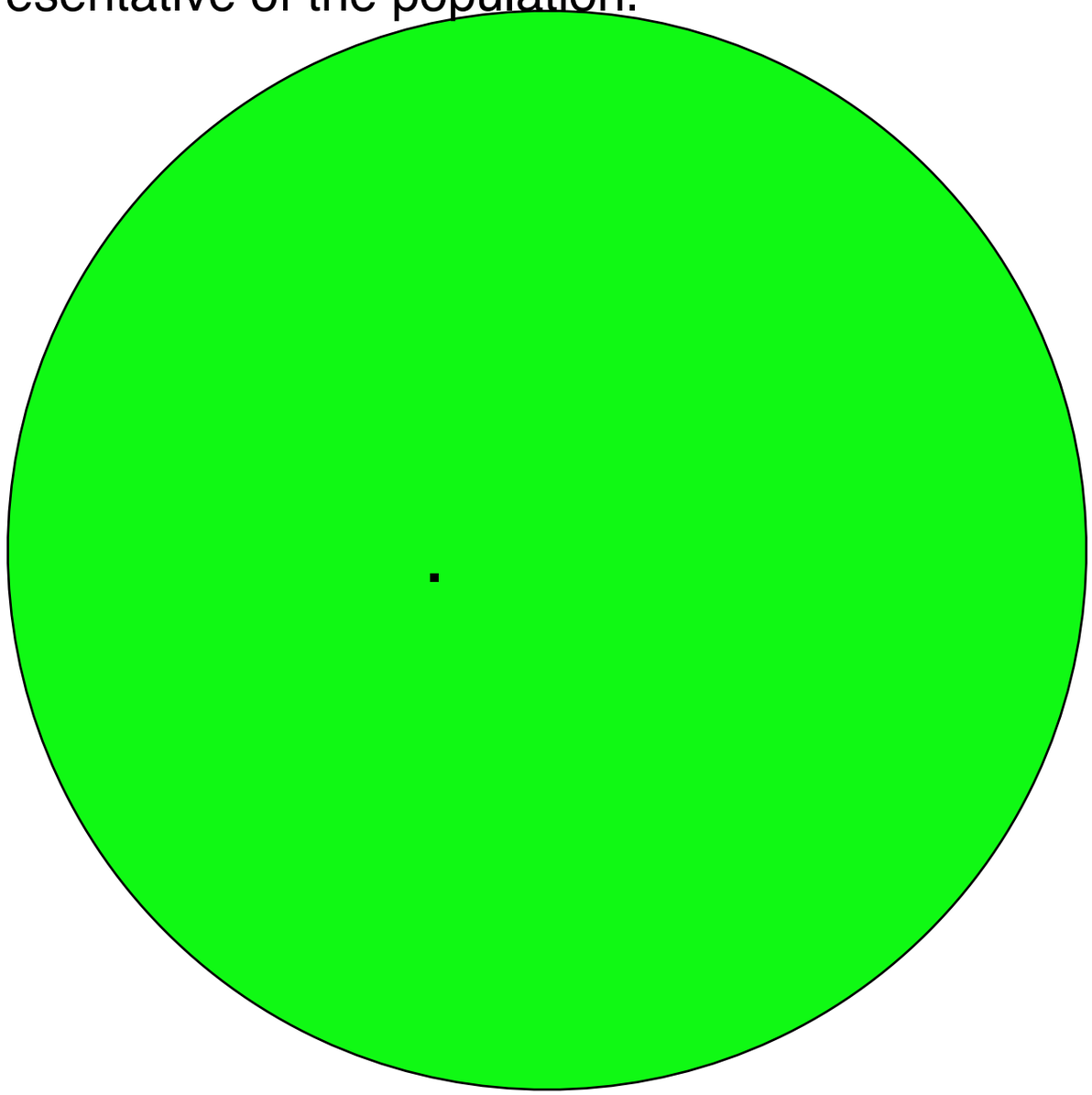


Chances are good that they are all blank.

Random sampling *won't* find a single written sector.

If the disk has 1,999,999,999 blank sectors (1 with data)

- The sample is representative of the population.

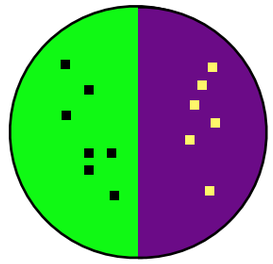


We will only find that 1 sector with exhaustive search.

What about other distributions?

If the disk has 1,000,000,000 blank sectors (1,000,000,000 with data)

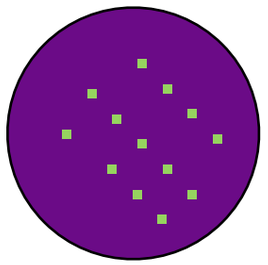
- The sampled frequency should match the distribution.
- *This is why we use random sampling.*



If the disk has 10,000 blank sectors (1,999,990,000 with data)

— and all these are the sectors that we read???

- We are incredibly unlucky.
- ***Somebody has hacked our random number generator!***



This is an example of the "urn" problem from statistics.

Assume the disk has 10MB of data --- 20,000 non-zero sectors.

Read just 1 sector; the odds of finding a non-blank sector are:

$$\frac{200{,}000{,}000-20{,}000}{200{,}000{,}000} = 0.9999.$$

Read 2 sectors. The odds are:

$$\left(\frac{200{,}000{,}000-20{,}000}{200{,}000{,}000}\right)\left(\frac{199{,}999{,}999-20{,}000}{199{,}999{,}999}\right) = 0.99980001$$

first pick **second pick**

Odds we may have missed something

The more sectors picked, the less likely we are to miss all of the sectors that have non-NULL data.

$$P(X = 0) = \prod_{i=1}^{n} \frac{((N - (i-1)) - M)}{(N - (i-1))} \qquad (5)$$

Sampled sectors	Probability of not finding data
1	0.99999
100	0.99900
1000	0.99005
10,000	0.90484
100,000	0.36787
200,000	0.13532
300,000	0.04978
400,000	0.01831
500,000	0.00673

Table 1: Probability of not finding any of 10MB of data on a 1TB hard drive for a given number of randomly sampled sectors. Smaller probabilities indicate higher accuracy.

Non-null data		Probability of not finding data
Sectors	Bytes	with 10,000 sampled sectors
20,000	10 MB	0.90484
100,000	50 MB	0.60652
200,000	100 MB	0.36786
300,000	150 MB	0.22310
400,000	200 MB	0.13531
500,000	250 MB	0.08206
600,000	300 MB	0.04976
700,000	350 MB	0.03018
1,000,000	500 MB	0.00673

Table 2: Probability of not finding various amounts of data when sampling 10,000 disk sectors randomly. Smaller probabilities indicate higher accuracy.

—So pick 500,000 random sectors. If they are all NULL, then the disk has p=(1-.00673) chance of having 10MB of non-NULL data.

In practice, use a modified algorithm.

Sample with 64K “blocks” instead of 512-byte sectors.

- It takes the same amount of time to read 65,536 bytes as 512 bytes.
- Analyze 64K block with a 4K sliding window.

Scan local area when interesting data is found.

- If a portion of a JPEG is found, find the front.
- If a piece of an encrypted file is found, determine the extent.

Update results in real-time.

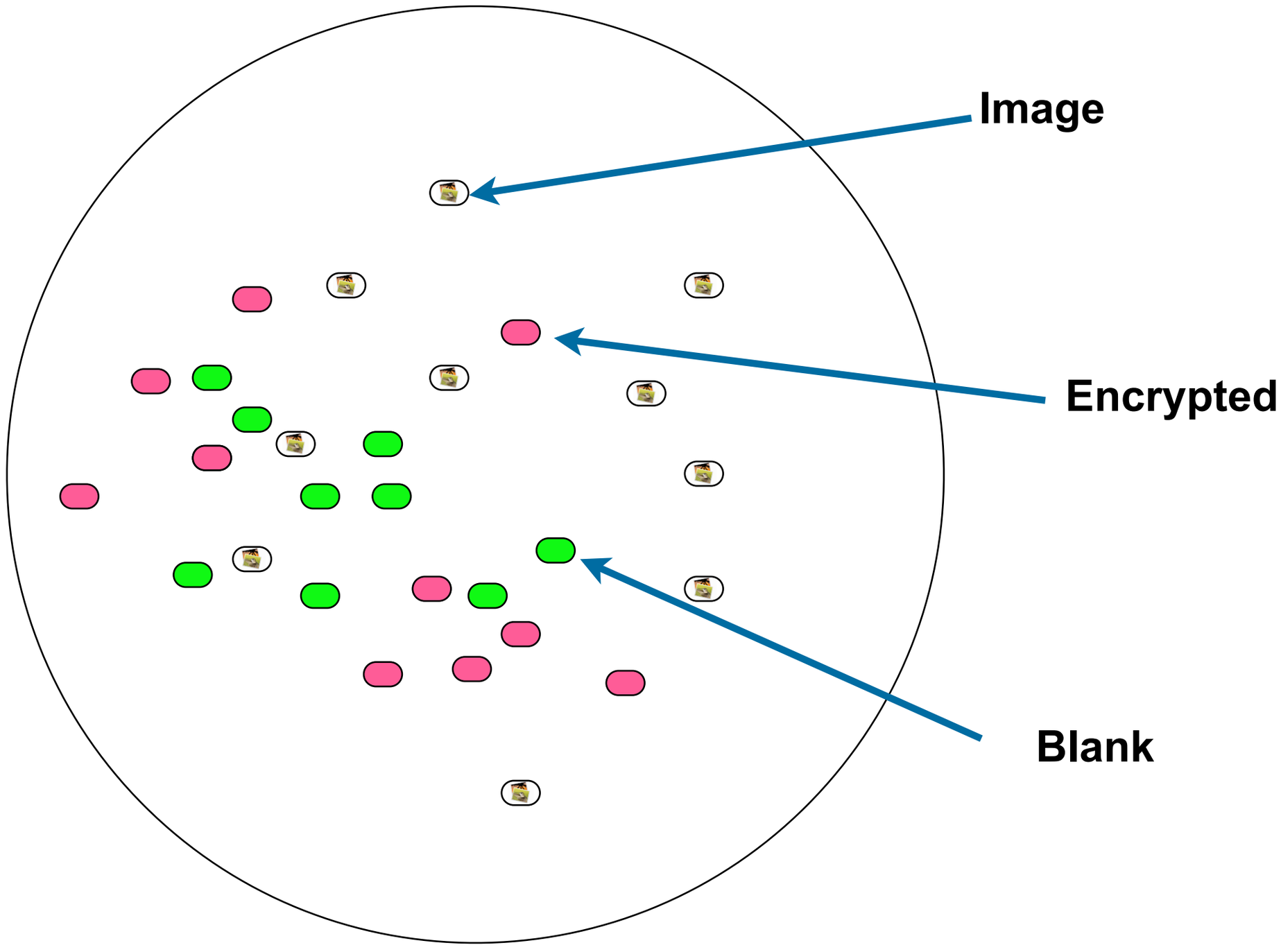
- Provides immediate feedback.
- Catches important data faster.

We can use this same technique to calculate the size of the TrueCrypt volume on this iPod.

It takes 3+ hours to read all the data on a 160GB iPod.

- Apple bought very slow hard drives.

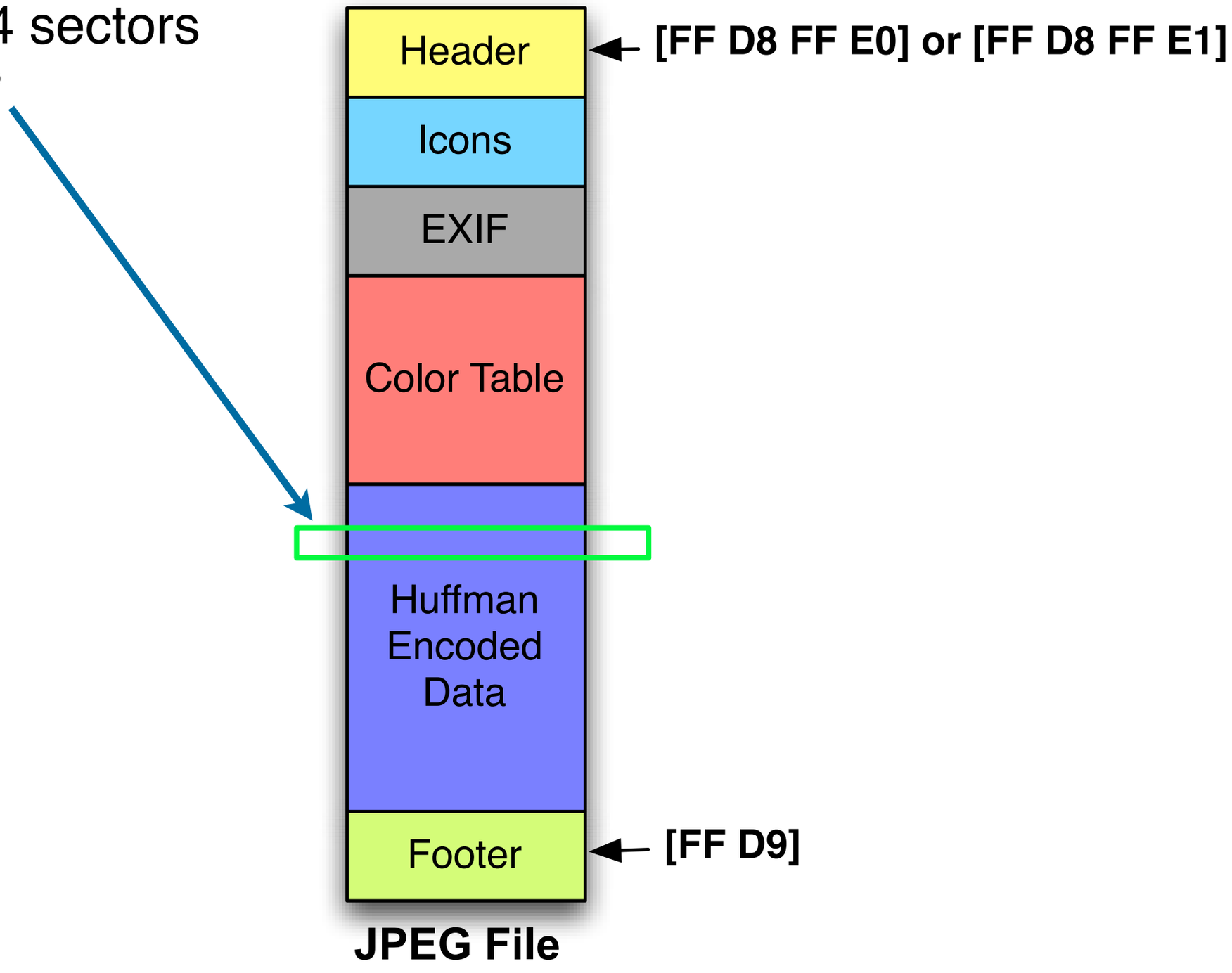
We can get a statistically significant sample in two minutes.



The % of the sample will approach the % of the population.

The challenge: identifying a file “type” from a fragment.

Can you identify a JPEG file from reading 4 sectors in the middle?



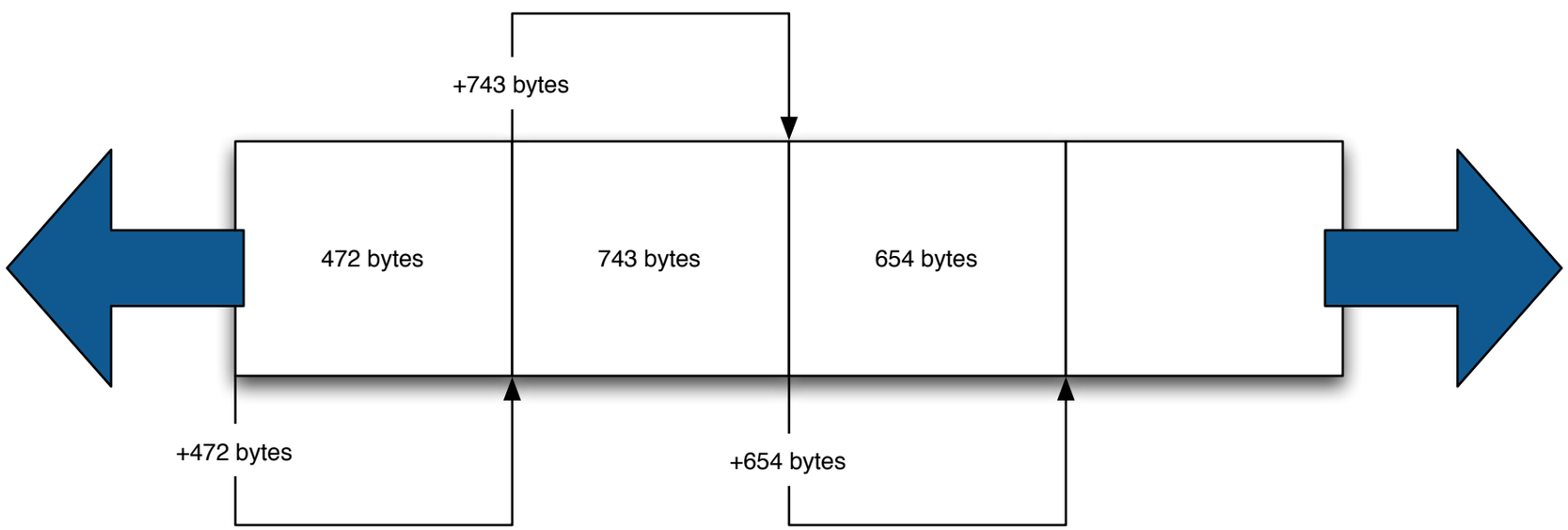
This is the *file fragment classification problem*.

HTML files can be reliably detected with HTML tags

```
<body onload="document.getElementById('quicksearch').terms.focus()">
  <div id="topBar">
    <div class="widthContainer">
          <div id="skiplinks">
            <ul>
              <li>Skip to:</li>
```

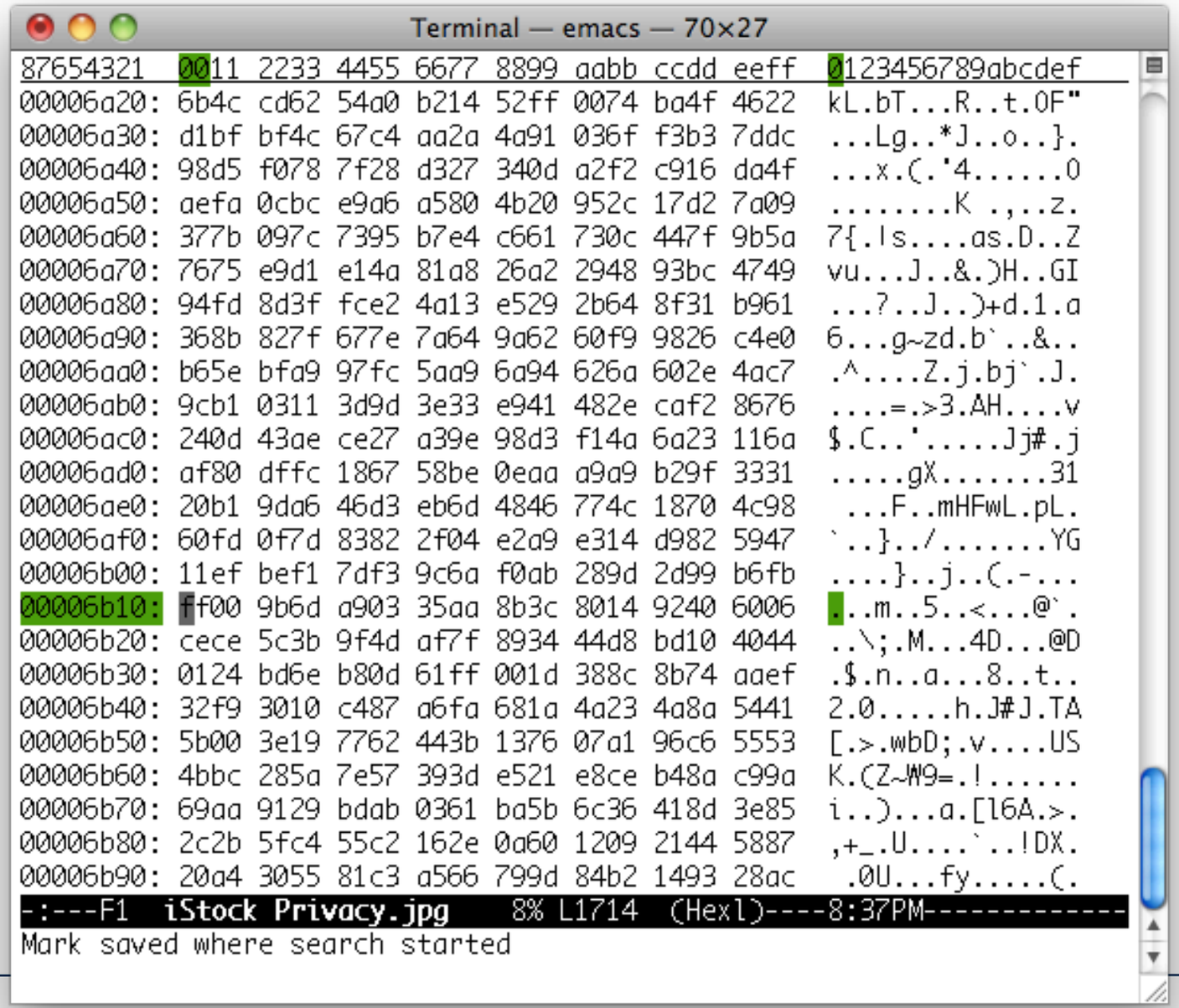
MPEG files can be readily identified through framing

- Each frame has a header and a length.
- Find a header, read the length, look for the next header.



One approach: hand-tuned discriminators based on a close reading of the specification.

For example, the JPEG format "stuffs" FF with a 00.



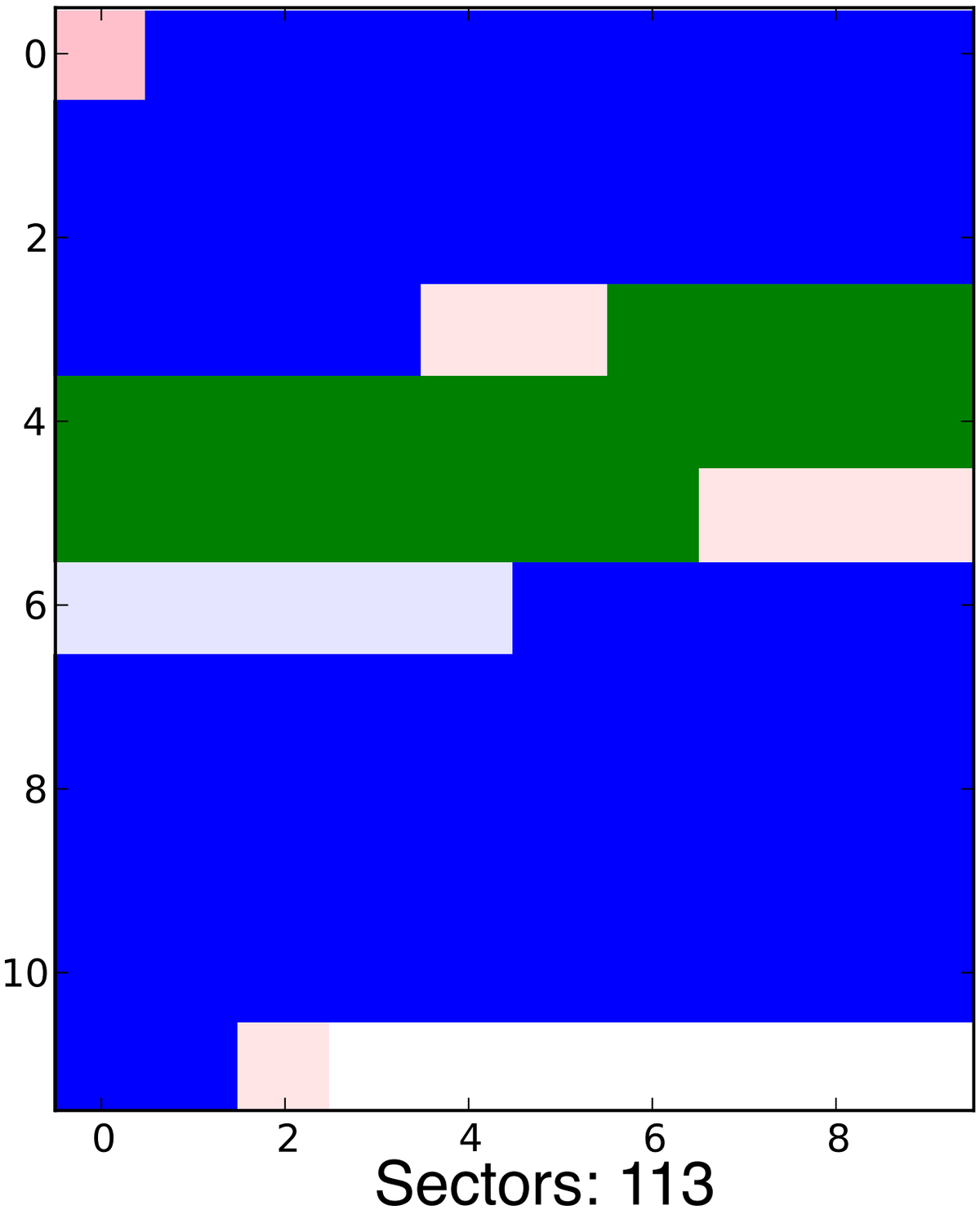
We built detectors to recognize the different parts of a JPEG file.

Bytes: 31,046

Nearly 50% of this 57K file identifies as “JPEG”

000897.jpg

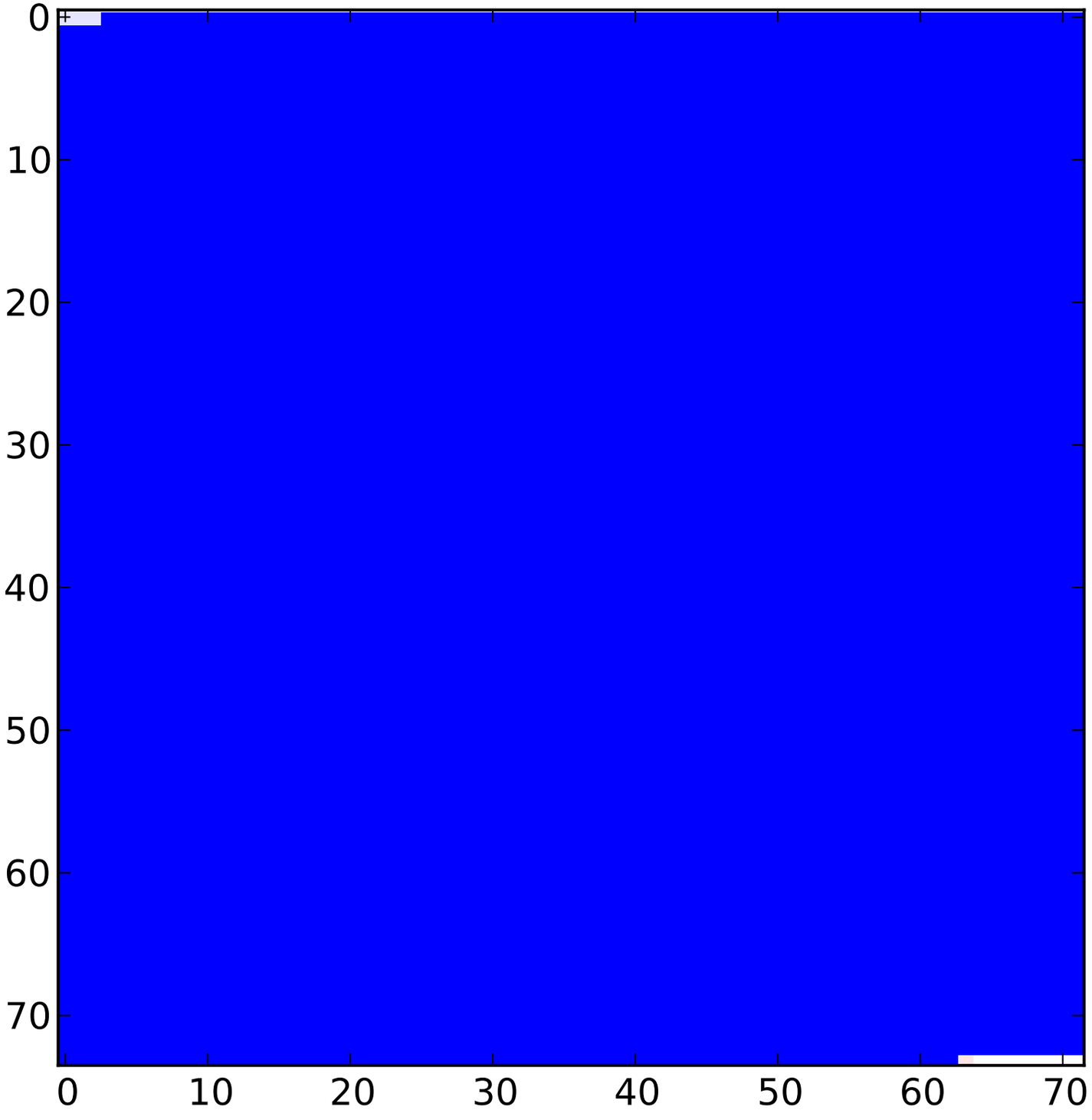
Bytes: 57596



Nearly 100% of this file identifies as “JPEG.”

000888.jpg

Bytes: 2,723,425



We developed five fragment discriminators.

JPEG — High entropy and FF00 pairs.

MPEG — Frames

Huffman-Coded Data — High Entropy & Autocorrelation

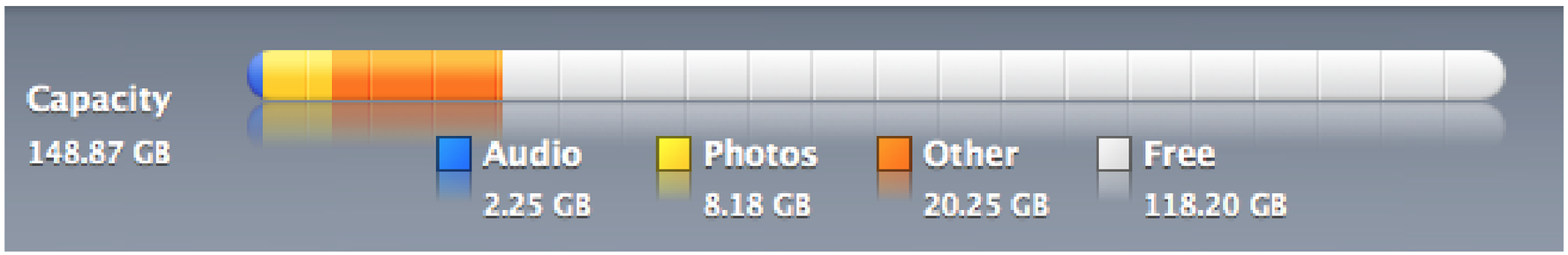
"Random" or "Encrypted" data — High Entropy & No autocorrelation

Distinct Data — a block from an image, movie, or encrypted file.

208 distinct 4096-byte block hashes

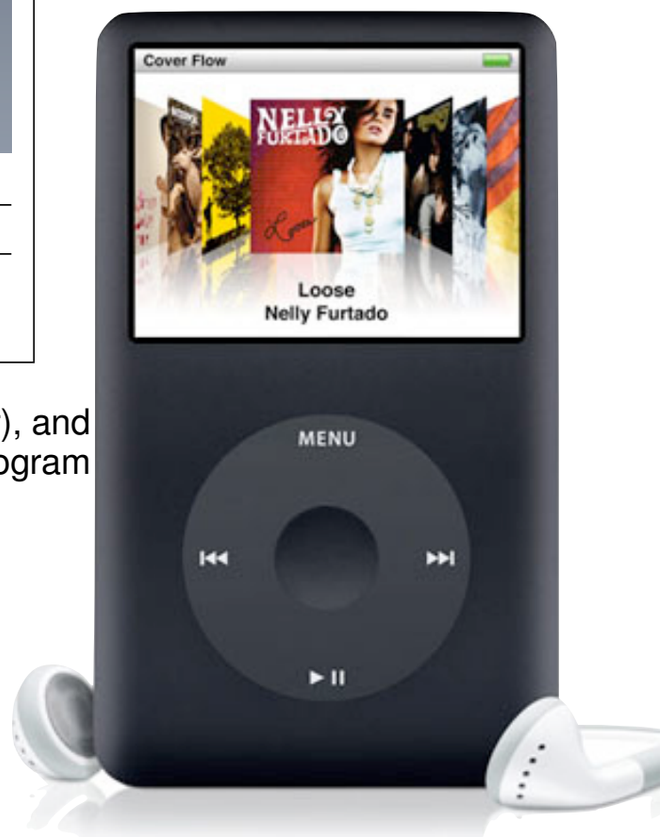
Combine random sampling with sector discrimination to obtain the forensic contents of a storage device.

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



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.



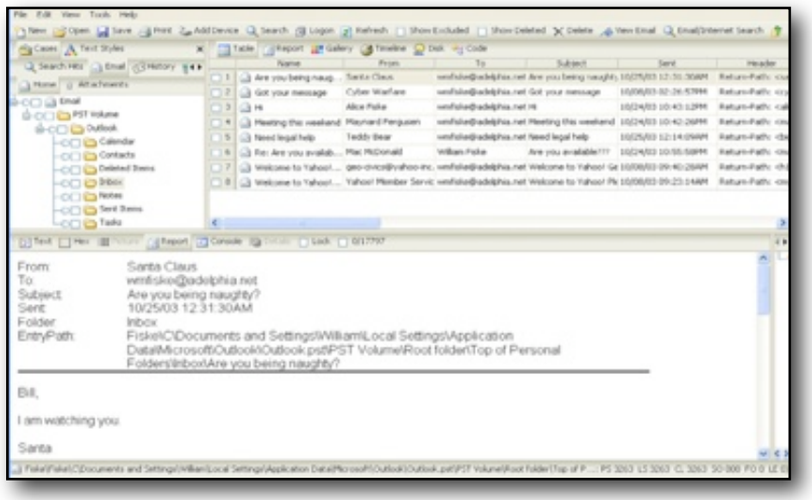
We accurately determined:

- % of free space; % JPEG; % encrypted

—*Simson Garfinkel, Vassil Roussev, Alex Nelson and Douglas White, Using purpose-built functions and block hashes to enable small block and sub-file forensics, DFRWS 2010, Portland, OR*

Creating Forensic Corpora

Digital forensics is at a turning point.
Yesterday's work was primarily *reverse engineering*.



Key technical challenges:

- Evidence preservation.
- File recovery (file system support); Undeleting files
- Encryption cracking.
- Keyword search.

Today's work is increasingly *scientific.*

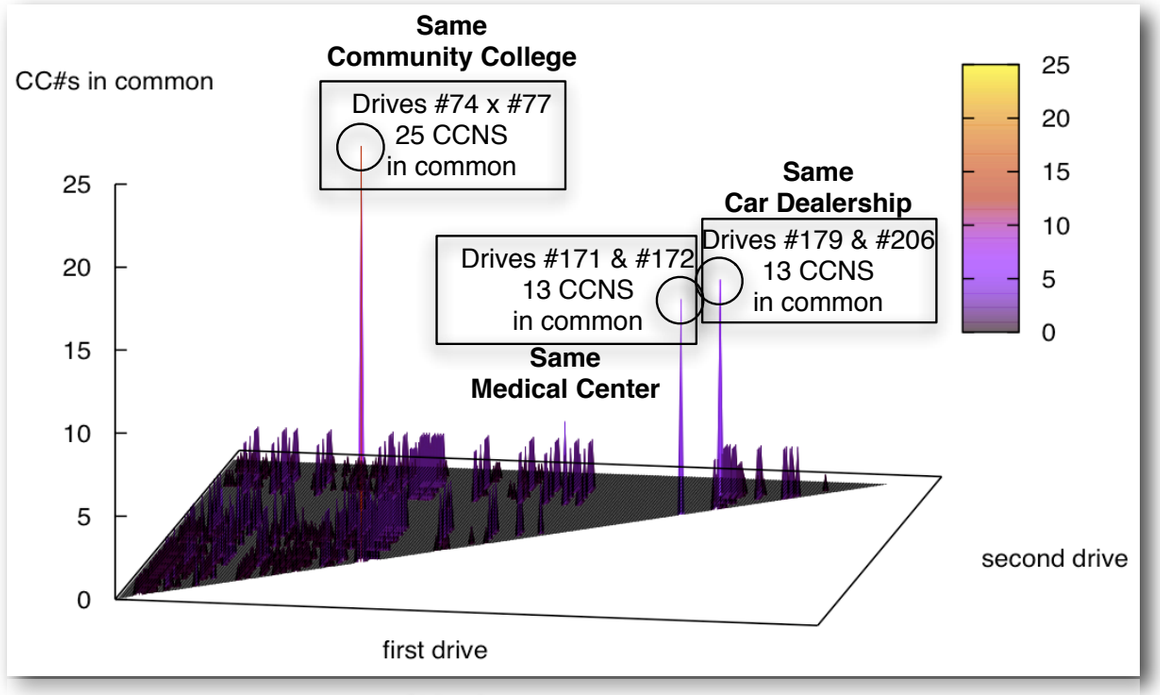
Evidence Reconstruction

- Files (fragment recovery carving)
- Timelines (visualization)

Clustering and data mining

Social network analysis

Sense-making



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 can't do this today.

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

We do science with “real data.”

The Real Data Corpus (30TB)

- 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.)

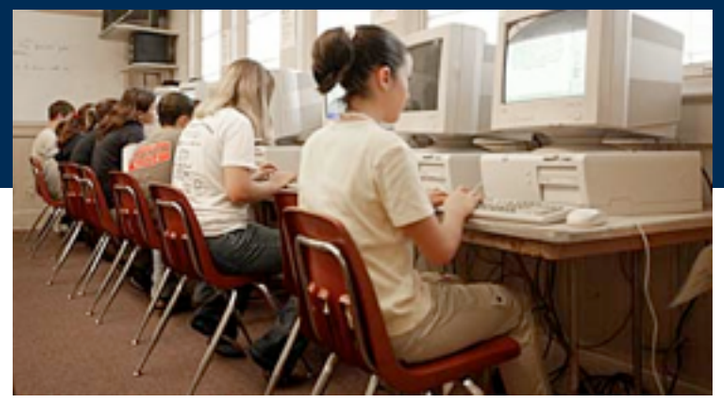
Mobile Phone Application Corpus

- Android Applications; Mobile Malware; etc.

The problems we encounter 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 fake 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

We manufacture 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>

Our fake data can be freely redistributed.

Test and Realistic Disk Images (1TB)

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

—*NSF DUE-0919593*

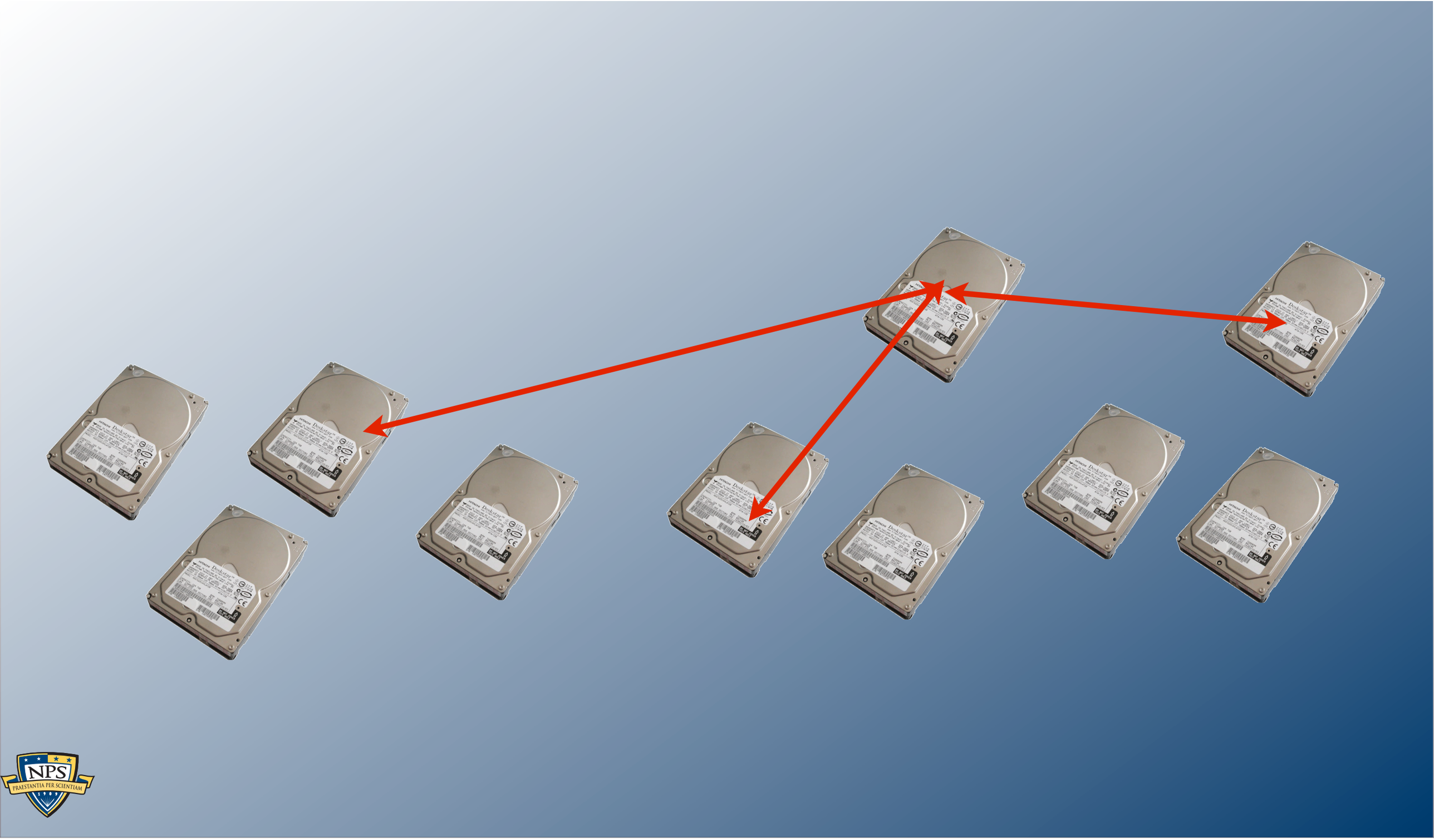
University harassment scenario

- 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*



Where do we go from here?

There are many important areas for research

Algorithm development.

- Adopting to **different kinds of data**.
- **Different resolutions**
- **Higher Amounts (40TB—40PB)**

Software that can…

- Automatically identify outliers and inconsistencies.
- Automatically present complex results in simple, straightforward reports.
- Combine stored data, network data, and Internet-based information.

Many of the techniques here are also applicable to:

- Social Network Analysis.
- Personal Information Management.
- Data mining unstructured information.

My challenges: innovation, scale & community

Most innovative forensic tools **fail when they are deployed**.

- Production data *much larger* than test data.
  - —*One drive might have 10,000 email addresses, another might have 2,000,000.*
- Production data *more heterogeneous* than test data.
- Analysts have less experience & time than tool developers.

How to address?

- Attention to usability & recovery.
- High Performance Computing for testing.
- Programming languages that are *safe* and *high-performance.*

Moving research results from lab to field is itself a research problem.

In summary, there is an urgent need for fundamental research in automated computer forensics.

Most work to date has been data recovery and reverse engineering.

- User-level file systems
- Recovery of deleted files.

To solve tomorrow's hard problems, we need:

- Algorithms that exploit large data sets (>10TB)
- Machine learning to find *outliers* and *inconsistencies.*
- Algorithms tolerant of data that is *dirty* and *damaged.*

Work in automated forensics is *inherently interdisciplinary.*

- Systems, Security, and Network Engineering
- Machine Learning
- Natural Language Processing
- Algorithms (compression, decompression, big data)
- High Performance Computing
- Human Computer Interactions

There are many opportunities to work outside CS

Need to engage with:

- Policy makers
- Police, Defense & Intelligence Communities.
- Defense Bar

Interesting legal issues.

- Data acquisition.
- Privacy
- Research Oversight (Institutional Review Boards.)

—*For more information, see http://www.simson.net/* and *http://forensicswiki.org/*

Questions?