Automated Digital Forensics

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

Oct 31, 2012

http://simson.net/

https://domex.nps.edu/deep/

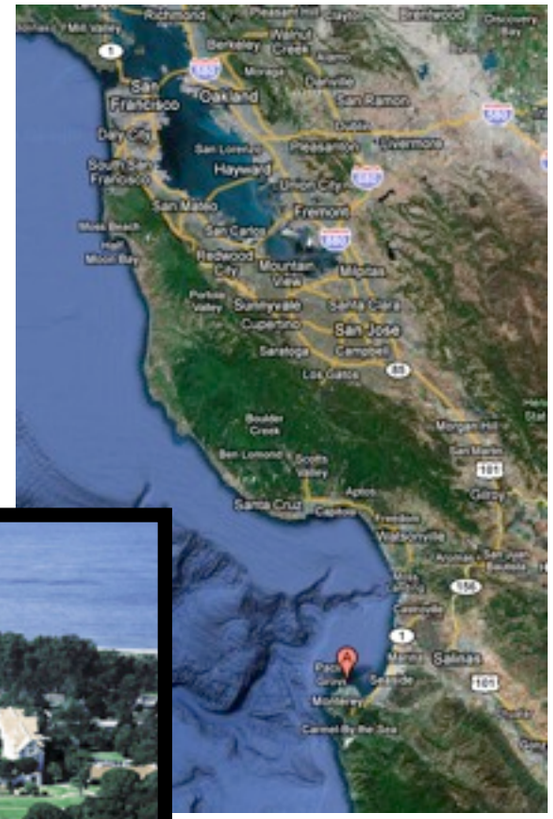
NPS is the Navy's Research University.

Monterey, CA — 1500 students

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

Schools:

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



NCR Initiative — Arlington, VA

- 8 offices on 5th floor, Virginia Tech building
- Current staffing: 4 professors, 2 lab managers, 2 programmers, 4 contractors
- **OPEN SLOTS FOR .GOV PHDs!**

The Digital Evaluation and Exploitation (DEEP) Group: Research in “trusted” systems and exploitation.

“Evaluation”

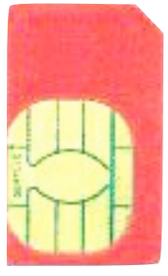
- Trusted hardware and software
- Cloud computing

“Exploitation”

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

Current and Former Partners:

- Law Enforcement (FBI & Local)
- DHS (HSARPA)
- NSF (Education)
- DoD (JIEDDO & Others)



Traditionally forensics was used for *convictions*.

The goal was establishing possession of *contraband information.*

- Child Pornography
- Stolen documents.
- Hacker tools

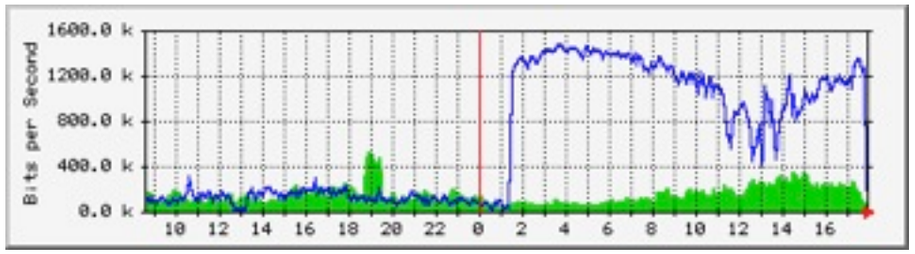
Forensics established:

- Data *presence*.
- Data *provenance — where it came from.*

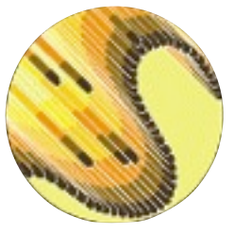
I started working digital forensics in the 1990s

1995 — Vineyard.NET

- Used forensics to investigate break-ins
- Limited forensics-related consulting
  - —*We saw forensics as computer security*



1998 — Sandstorm Enterprises



- PhoneSweep — Telephone scanner
- NetIntercept — Network forensics system

1998 — The “drives” project

- Used computers purchased for PhoneSweep had sensitive data
- I started buying drives *for the data*.
  - —*Developed automated analysis techniques for PhD thesis.*

My goal is to see forensics used for *investigations.*

Data extraction — What information does the target have?

- contacts, calendar, documents

Data fusion — Putting together a unified document

- When was something done?

Correlation — who has the *same information?*

- Identifies members within the organization.
- Identifying a subject's:
- Automatically identifying *actionable information.*

Example: Facebook Data Fusion

Site Name: Facebook.com User Jason Peterson

Multiple Pages, one per <Site account>

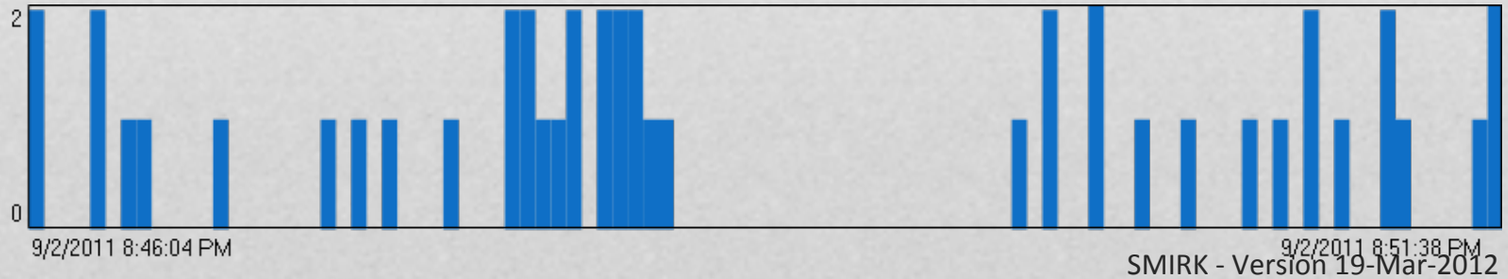
<Site Name> ID: 100001926917994
Name: Jason Peterson (match)

# of Profile Viewed: 5
# of Photo Viewed: 50
# of Chat Sessions: 15
# of Video Viewed: 0

Other Items Found: 55

See MS-Word file for details

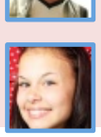
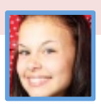
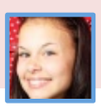
This activity chart is Facebook activity. One color



SMIRK - Version 19-Mar-2012

The tool automatically finds all Facebook data on the hard drive and arranges it into a single report.

Chat Session

	User	Time	Message
	Jason Peterson [100001926917994]	2011-09-02 20:46:36Z	'Hey Sue'
	Susan Dillard [100001995672759]	2011-09-02 20:46:47Z	'hey jason, whats up'
	Jason Peterson [100001926917994]	2011-09-02 20:46:59Z	'Are you coming to the meeting with the boss?'
	Jason Peterson [100001926917994]	2011-09-02 20:47:04Z	'It's at 10pm, under the bridge'
	Susan Dillard [100001995672759]	2011-09-02 20:47:56Z	'i might be late but i'll be there'
	Jason Peterson [100001926917994]	2011-09-02 20:49:10Z	'ok I'll see you then'
	Susan Dillard [100001995672759]	2011-09-02 20:49:13Z	'السلام عليكم ورحمة الله تعالى وبركاته'
	Jason Peterson [100001926917994]	2011-09-02 20:49:36Z	'you too!'

Three principles underly our research.

1. Automation is essential.

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

2. Concentrate on the invisible.

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

3. 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.

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

We analyzed 2000 hard drives.

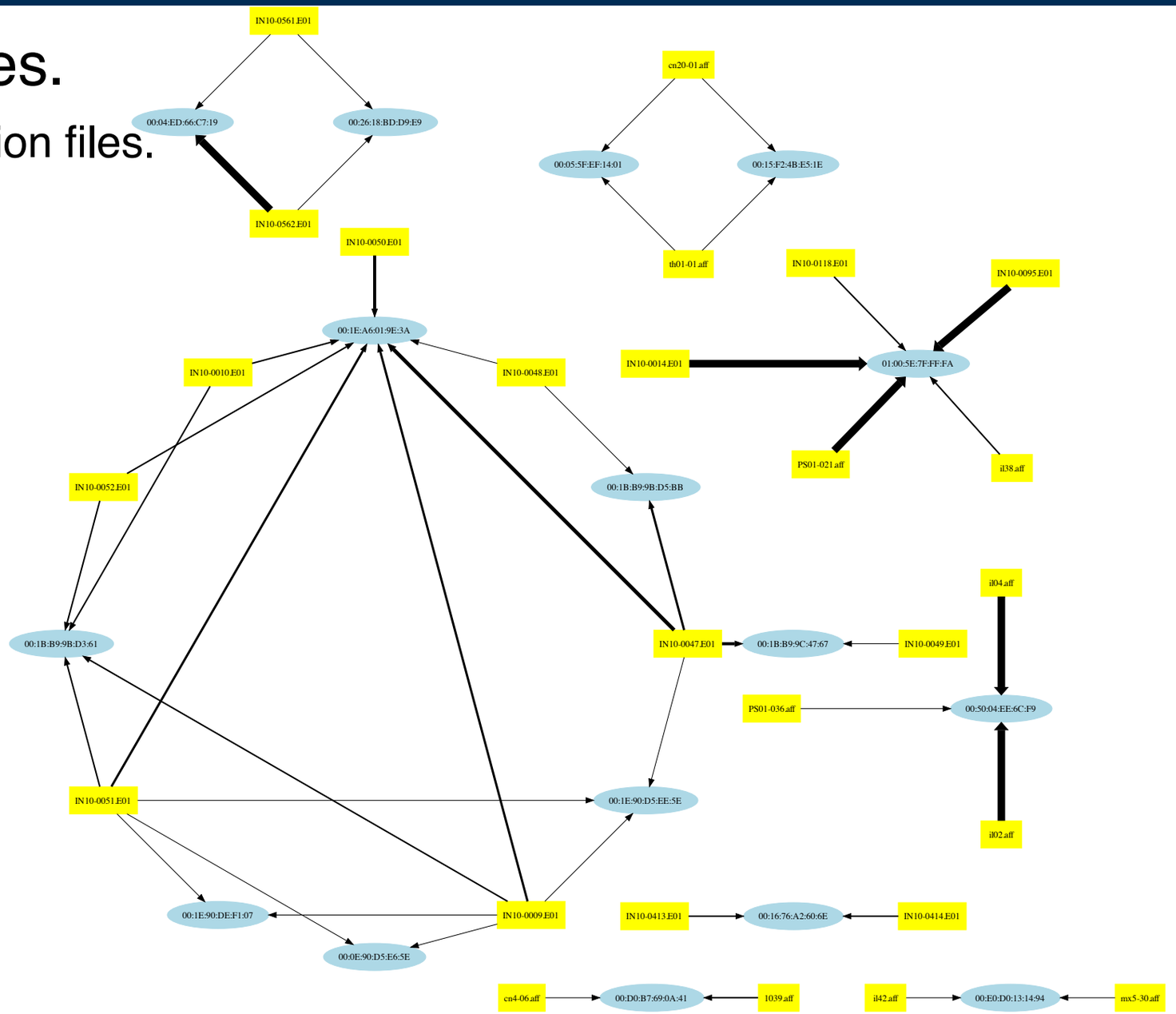
- Find IP packets in swap & hibernation files.
- Extract ethernet MAC addresses.

Post-processing identifies:

- Shared wireless routers.
- Common ethernet routers.

Validation:

- Reconstructed networks came from same organization.



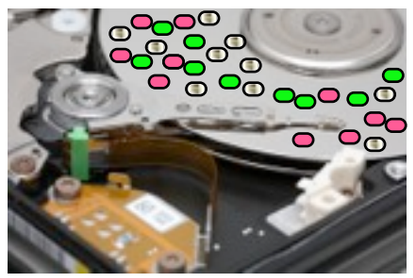
—*Forensic Carving of Network Packets and Associated Data Structures,* Beverly & Garfinkel, DFRWS 2011, August 2011, New Orleans

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.

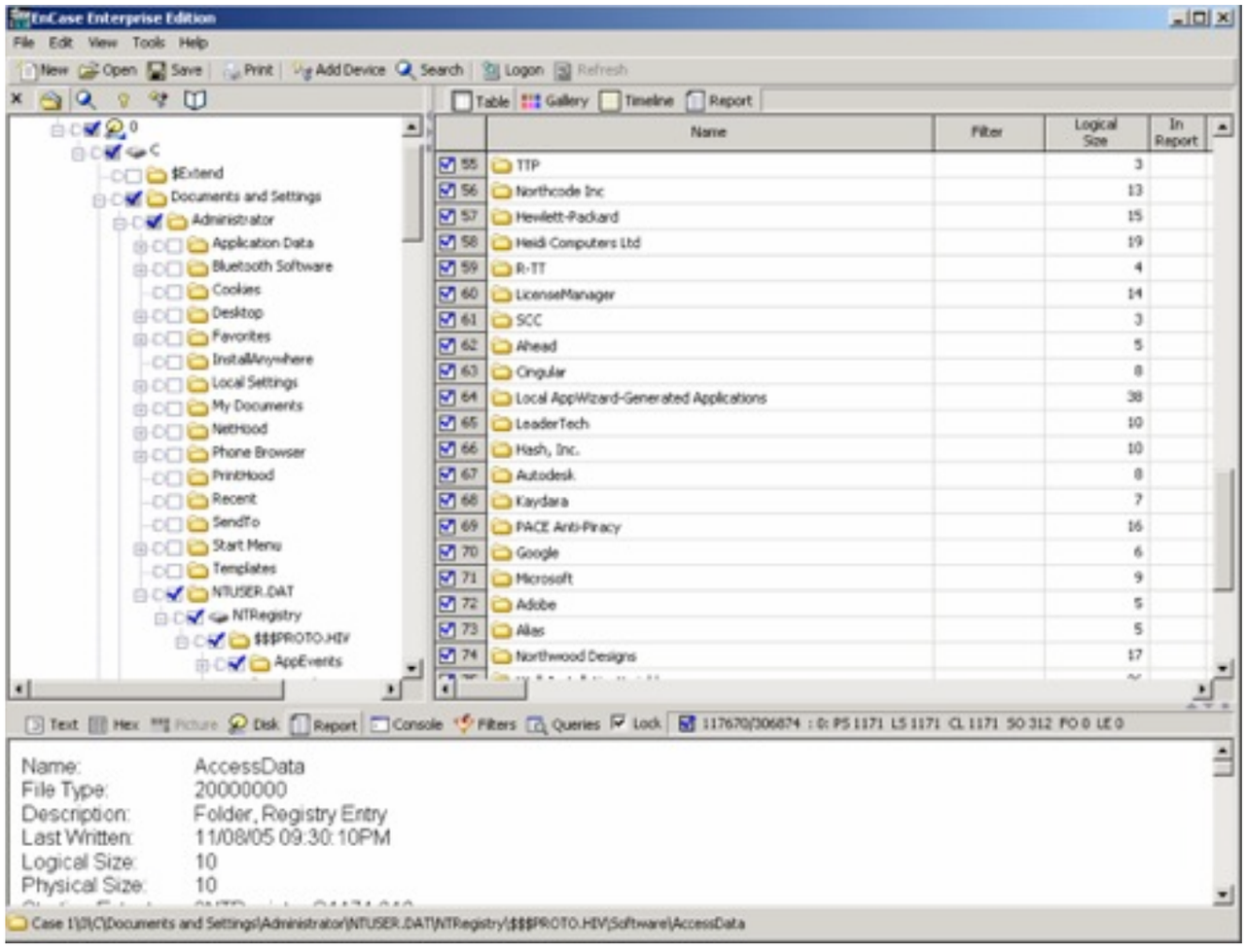
Forensic copy ("disk image") stored on a storage array.

Write blockers are also used with USB drives, phones.

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

Digital forensics is fundamentally different from other kinds of scientific exploration...

There are five key problems.

2.1 Diversity is a fundamental challenge of DF

Our charter:

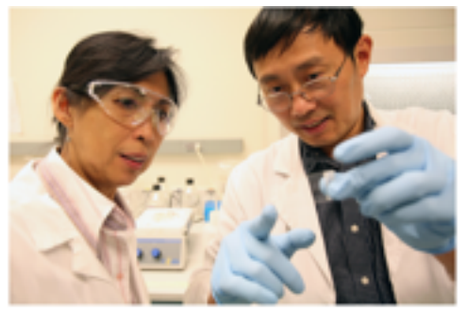
"Analyze any data that might be found on a computer."

Non-DF research is typically confined to a single area:

energy

math

literature



chemistry

DF must analyze any OS, application, protocol, encryption, etc...

Diversity is more than a multiplicity of file formats...

Data may be *inconsistent* or *incomplete*

- Files that are *deleted* or partially *overwritten*
- Incomplete database records
- Intentionally altered to avoid analysis

Data frequently have no formal specification

- Hacker tools & malware
- Proprietary file formats

We need strategies for systematically addressing diversity

- Exploit similarity and correlation.
  - *Items of interest are frequently repeated.*
- Detect deliberate attempts to hide information
  - *Eliminate the truth and the improbable, and whatever remains must be impossible (and therefore falsified)*
  - *"Improbable" data should be examined for stenography.*

2.2 Data scale is a never ending problem

Scale is continually identified as a DF problem

—*DFRWS 2001:*
"The major item affecting overall performance is data volume: the amount of data collected for analysis of this type is often quite large."

Moore's law scales the targets

- We are using top-of-the-line system to analyze top-of-the-line systems
- We need to analyze in hours (or days) what a subject spent weeks, months or years assembling

∴ We will *never* outpace the performance curve.

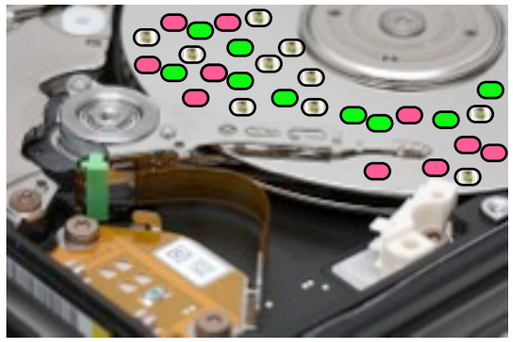
Most "big data" solutions from other fields don't work well with DF

- Budgets — Particle physicists have more $$ per case than we do. (SSC≃1.5PB/month)
- Data diversity — Physics (or even web) data is less diverse than a hard drive data
- Our data fights back — CERN data is not compressed, encrypted, fragmented, or malware
  —*Data complexity dramatically increases I/O and compute requirements*

Use *sampling* and *correlation* to address scale.

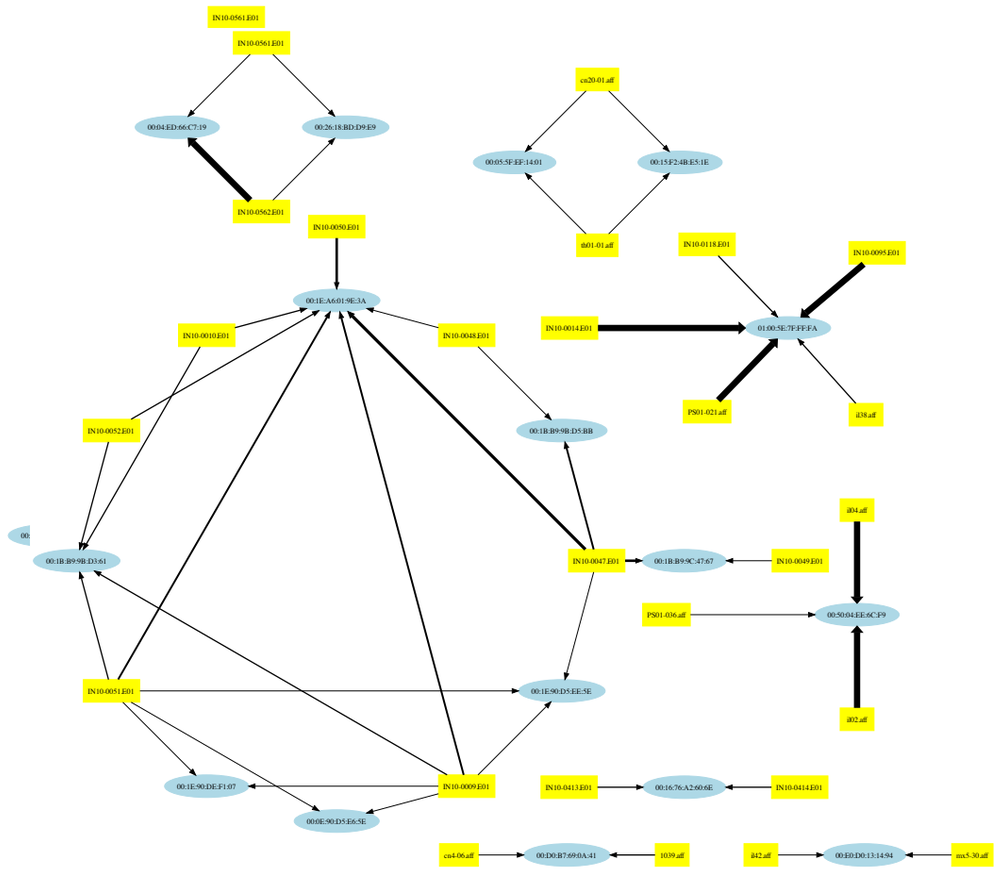
Sampling — “Sub Linear Algorithms”

- Evaluate just as portion of the data; use statistics to draw inferences
- ***Sampling can prove the presence of information!***
- ***Sampling cannot prove the absence of information***
  - ***—just the likely absence***
- “The absence of evidence is not the evidence of absence.”



Correlation:

- Have the data determine what’s important
- Use TF/IDF to remove the mundane (DFRWS 2006)



2.3 Temporal diversity creates a never-ending upgrade cycle

Today's DF tools must process:

- Today's computers / phones / cameras
  - —*Because some criminals like to buy what's new!*
- Yesterday's computers / phones / cameras
  - —*Because criminals are using old devices too!*

Implications for DF users and developers:

- Upgrade DF software as soon as possible.
- DF software will become geometrically more complicated over time....
  - —*... or DF software will adapt on the fly to new data formats and representations.*
  - —*automated code analysis; pattern matching; hidden Markov models; etc.*

2.4 Human capital is bad all over... ... it's especially bad for DF

DF users (examiners, analysts):

- Overwhelmingly in law enforcement.
- Little or no background in CS or IS
- Deadline-driven; over-worked
- Knowledgable users tend to focus in just one particular area.
  - —*Result: It takes two years to train most DF examiners.*

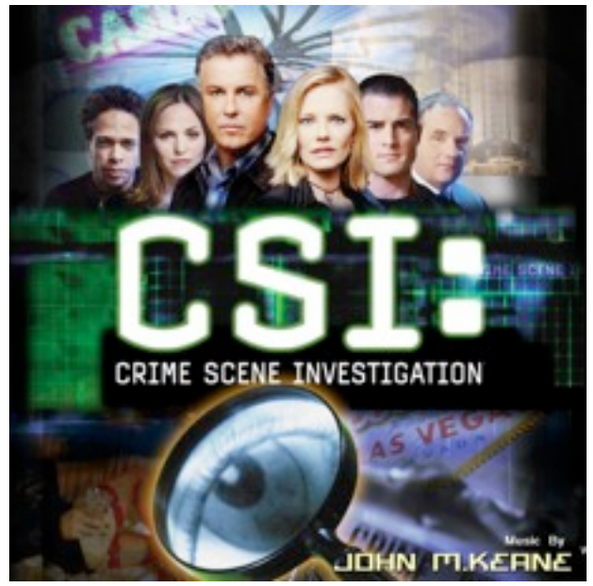
DF developers ("researchers"):

- Data diversity means developers need to know the whole stack
  - —*opcodes & Unicode ⇒ OS & Apps ⇒ networking, encryption, etc.*
- Scale issues means developers need to know HPC:
  - —*threading, systems engineering, supercomputing, etc.*
- Result:
  - —*It's hard to find qualified developers*
  - —*Developers must be generalists*

2.5 The “CSI Effect” causes unrealistic expectations.

On TV:

- Forensics is swift.
- Forensics is certain.
- Human memory is reliable.
- Presentations are highly produced.



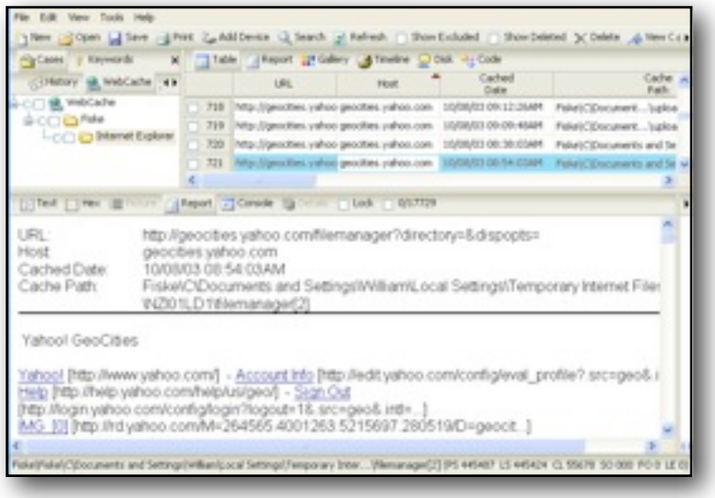
TV digital forensics:

- Every investigator is trained on every tool.
- Correlation is easy and instantaneous.
- There are no false positives.
- Overwritten data can be recovered.
- Encrypted data can usually be cracked.
- It is impossible to delete anything.

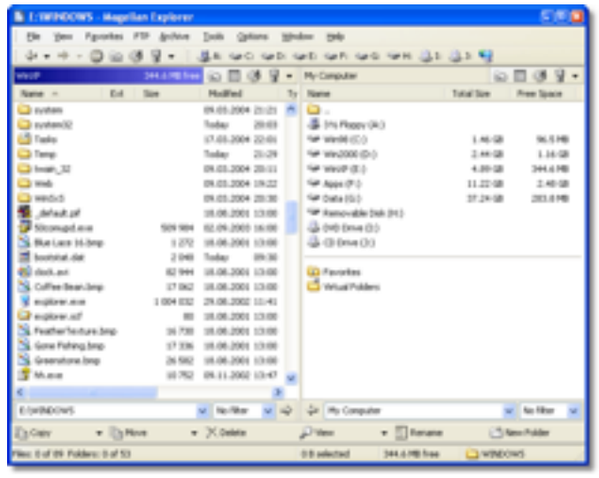
The reality of digital forensics is less exciting.

There are lots of problems:

- Data that is overwritten cannot be recovered
- Encrypted data usually can't be decrypted
- Forensics rarely answers questions or establishes guilt or provides specific information
- Tools crash a lot
- DF tools look a lot like traditional tools



EnCase



Windows Explorer

Result:

— *DF is a difficult process that looks easy*

— *This is not a good place to be*

2.6 Digital Forensics: expensive tools with a limited market

DF tools are expensive to develop:

- Data diversity
- Security critical
- High performance computing

Limited market:

- Consulting firms (more effective tools *decreases* billable hours)
- Police departments (not known for $$)
- Defense (not known for major DF expenditures)

My personal experience:

- It's very hard to stay in business as a tool developer
- Government should have an ongoing role in funding DF research and tool development
- Open source software frequently makes the most sense
  - —*Open Source preserves investment, enables future research, empowers users.*

DF researchers must respond with new algorithms.

Current approaches don't scale.

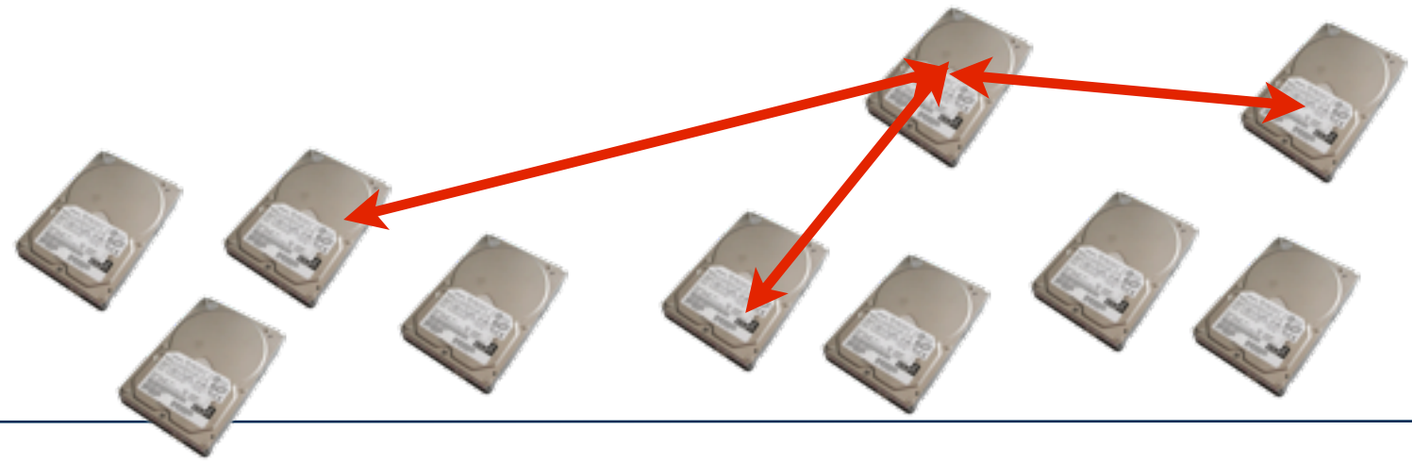
- User spent *years* assembling email, documents, etc.
- Analysts have days or hours to process it.
- Police analyze top-of-the-line systems
  —*with top-of-the-line systems.*
- National Labs have large-scale server farms
  —*to analyze huge collections.*

The problems:
1. Data Size
2. Mobile Devices
3. Encryption
4. Diversity
5. Time

Our new algorithms must:

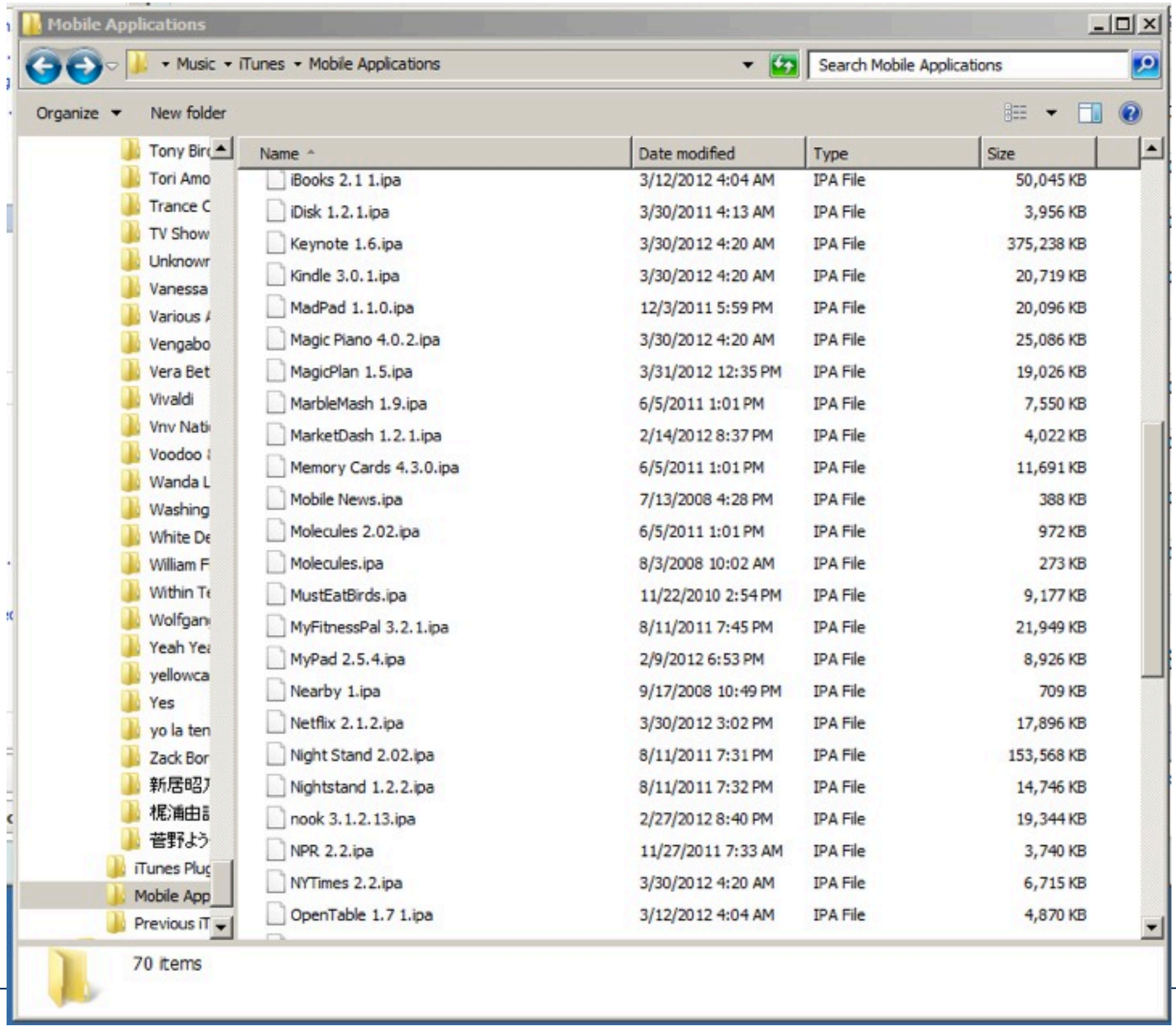
—*Provide incisive analysis through outlier detection and correlation.*

—*Operate autonomously on incomplete, heterogeneous datasets.*



Stream-based Forensics

We think of computers as devices with *files*.

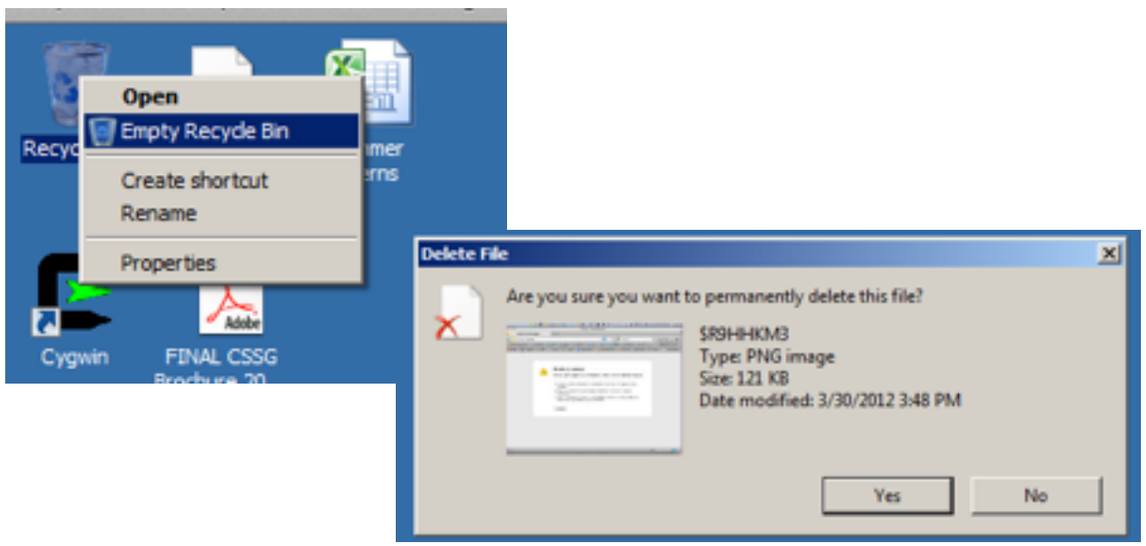


But data on computers is really in three categories:

Resident Data

user files
email messages
[temporary files]

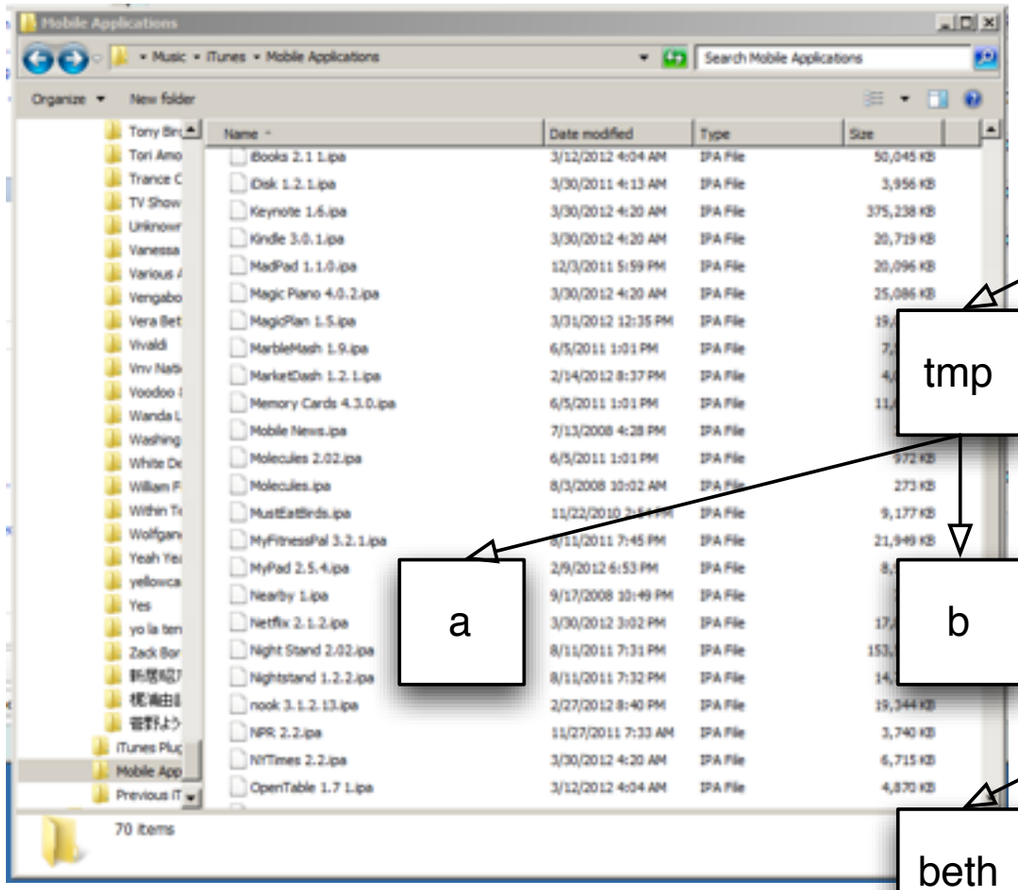
Deleted Data

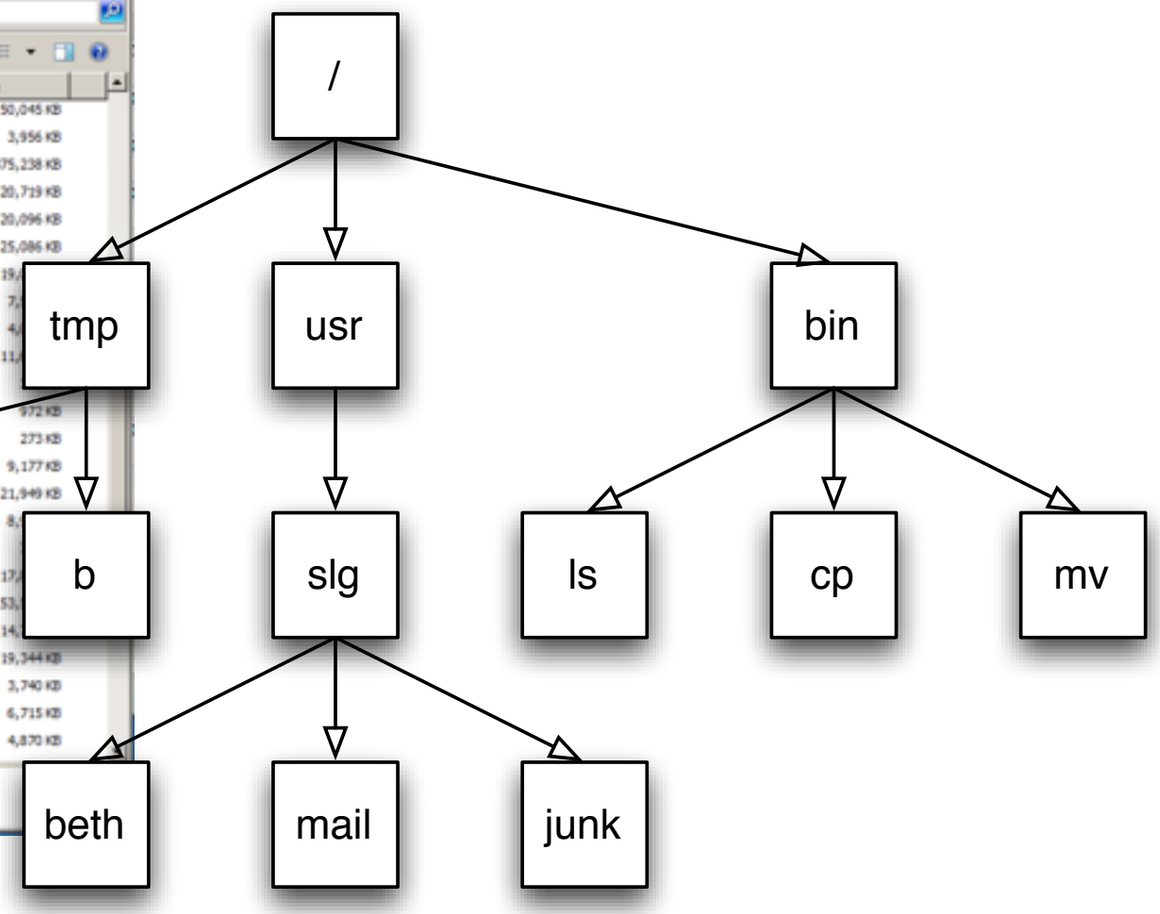


No Data

blank sectors

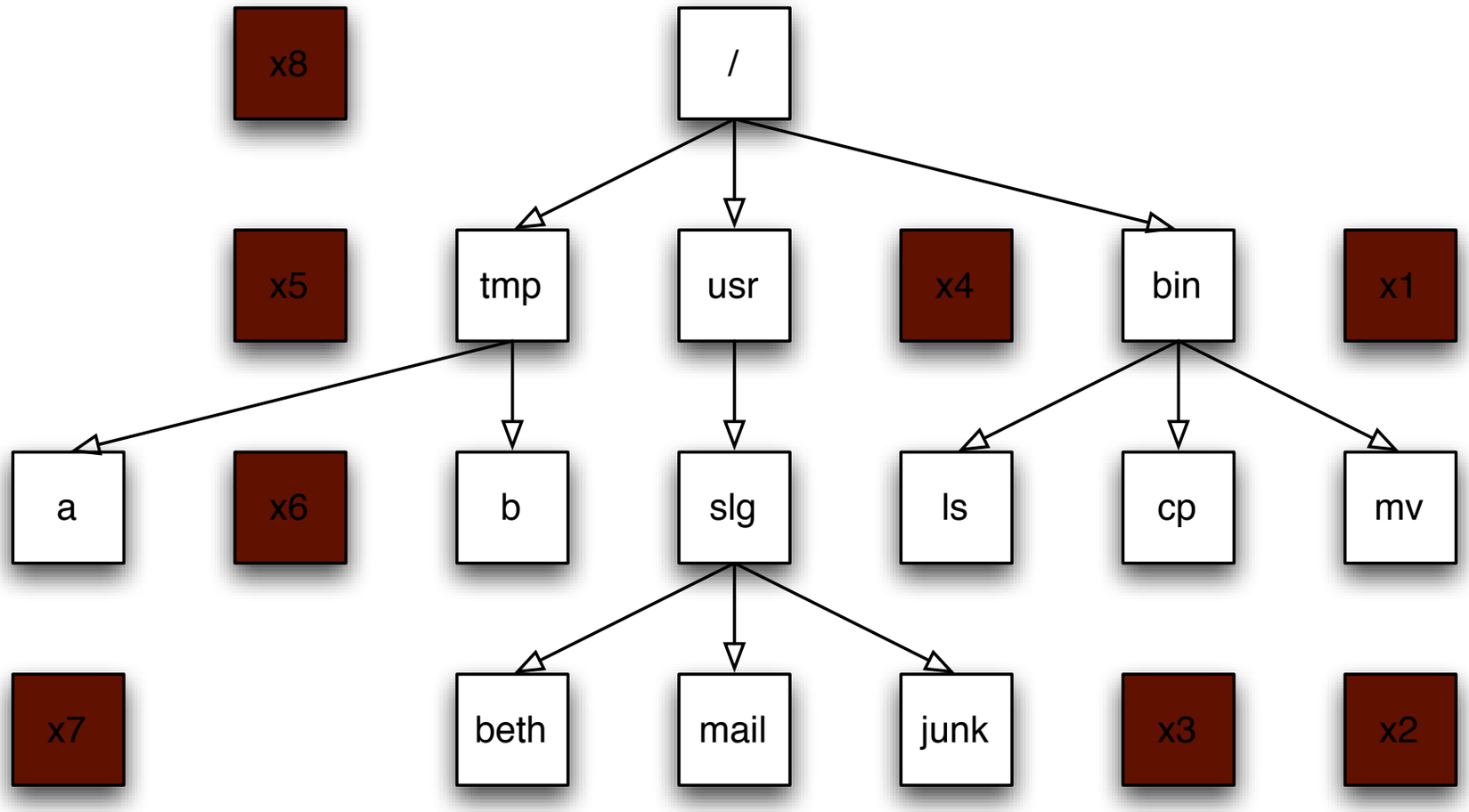
Resident data is the data you see from the root directory.





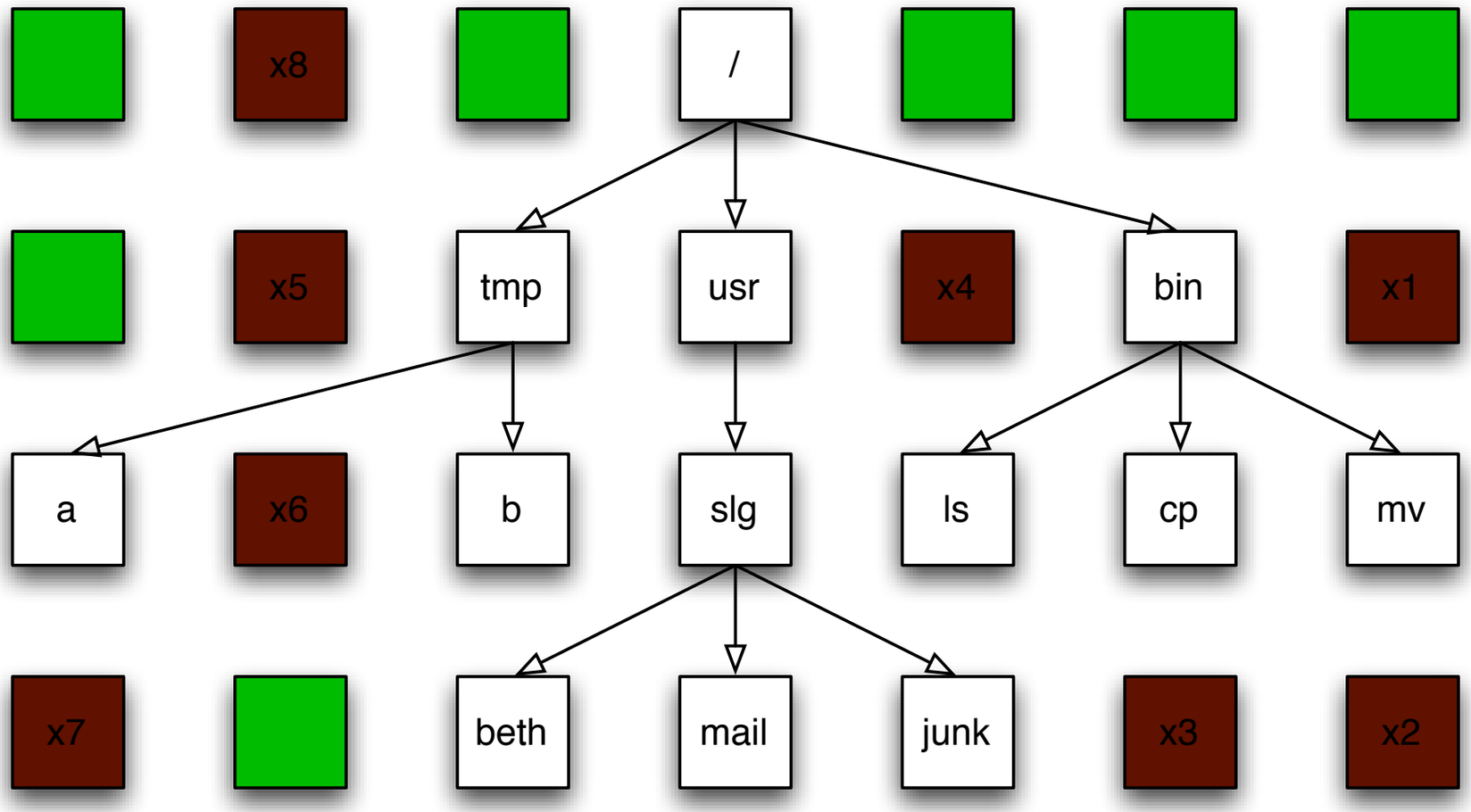
Resident Data

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



Deleted Data

Sectors with “No Data” are really blank.



No Data

Today most forensic tools analyze the *files* on the drive.

Advantage:

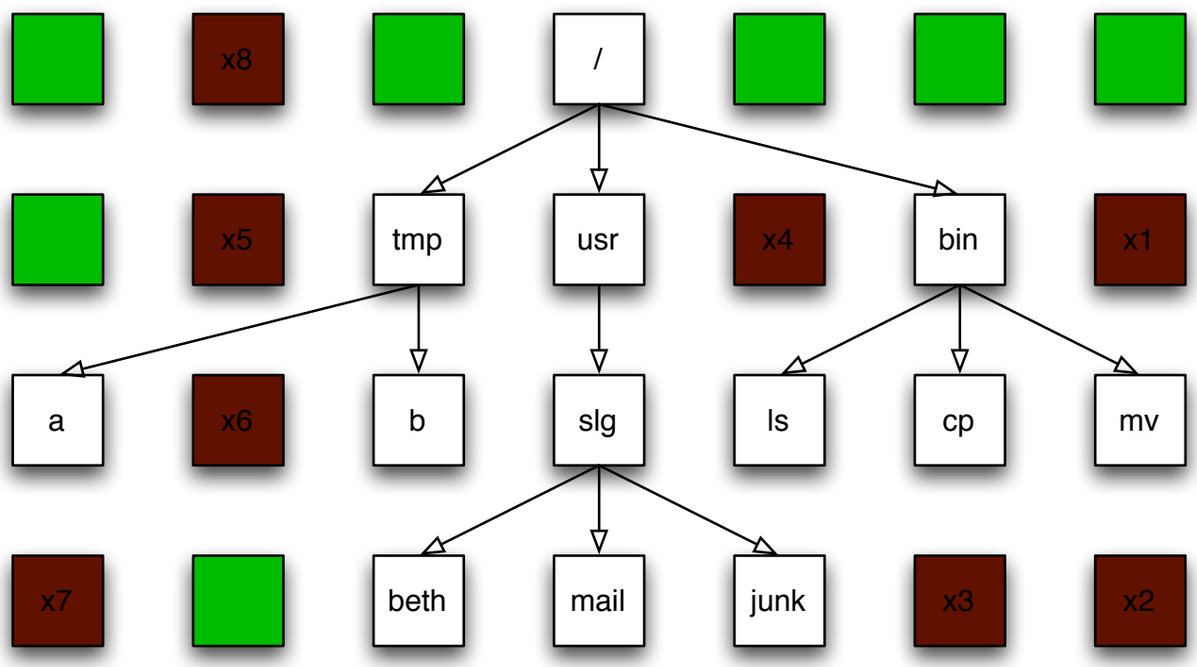
- Examiners know how to work with files
- It's easy to take files into court.

Problem #1: Time

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

Problem #2: Completeness

- Lots of data is ignored.



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.

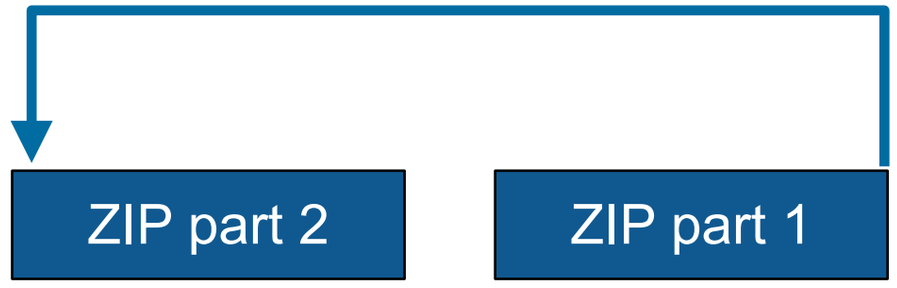
1. Read all of the blocks in order.
2. Look for information that might be useful.
3. Identify & extract what's possible in a single pass.

Advantages:

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

Disadvantages:

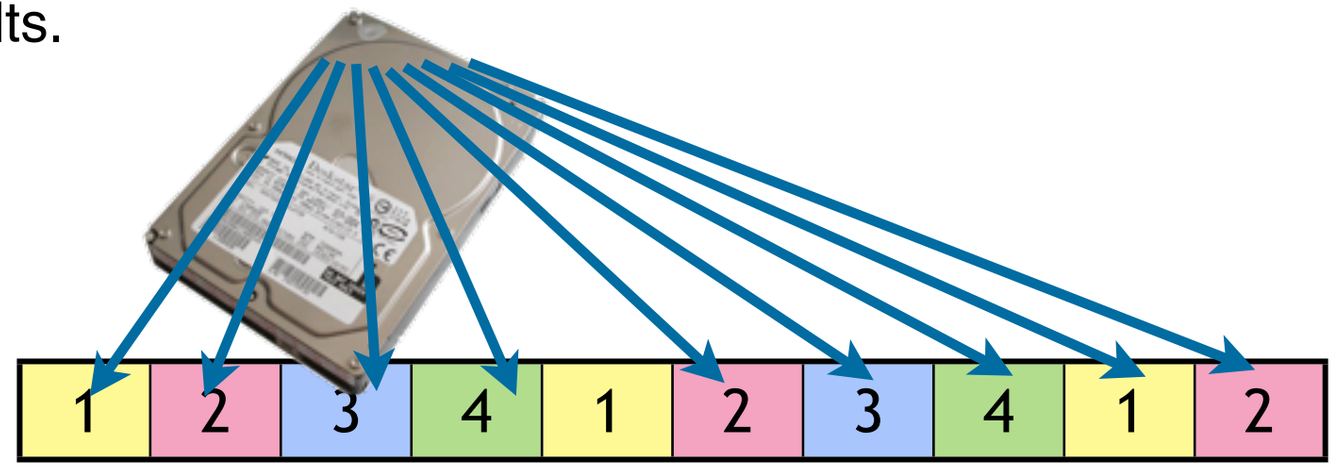
- A second pass is needed to recover file names.
- Some kinds of fragmented files can't be recovered:
  - —*Compressed files with part2-part1 ordering*
  - —*Files with internal fragmentation (.doc)*



bulk_extractor: a high-speed disk scanner.

Key Features:

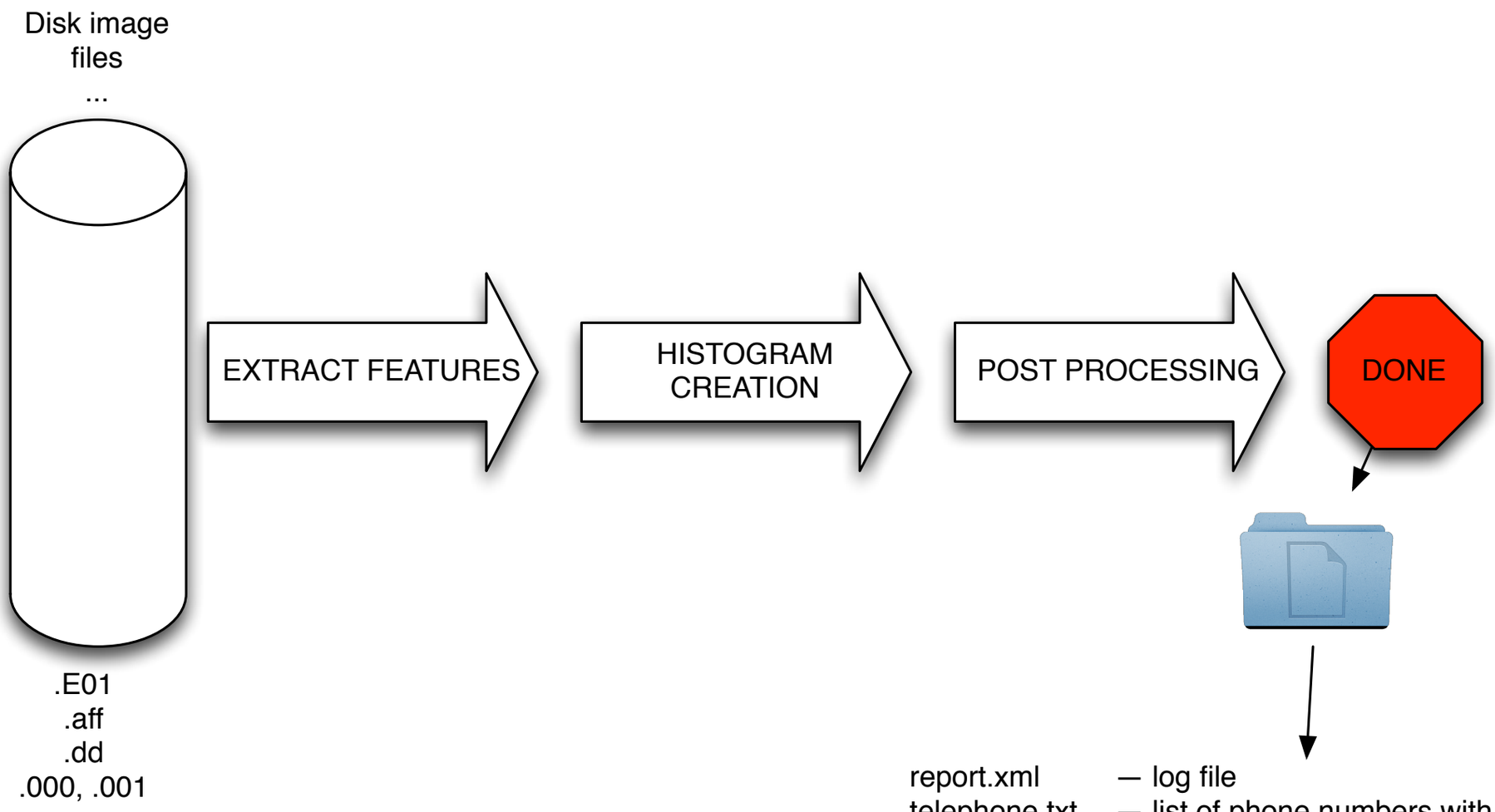
- Extracts “features” of importance in investigations:
  - —*email addresses; credit card numbers; JPEG EXIFs; URLs; Email fragments.*
- Recursively re-analyzes compressed data
- Produces a histogram of the results.
- Multi-threaded.
  - —*Disk is "striped."*
  - —*Results written out-of-order.*



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 has three phases of operation: Feature Extraction; Histogram Creation; Post Processing



Output is a *directory* containing:

- feature files; histograms; carved objects
- Mostly in UTF-8; some XML
- Can be bundled into a ZIP file and process with bulk_extractor_reader.py

Feature files are UTF-8 files that contain extracted data.

```
# UTF-8 Byte Order Marker; see http://unicode.org/faq/utf_bom.html
# bulk_extractor-Version: 1.3b1-dev2
# Filename: /corp/nps/drives/nps-2009-m57-patents/charlie-2009-12-11.E01
# Feature-Recorder: telephone
# Feature-File-Version: 1.1
...
6489225486    (316) 788-7300   Corrine Porter (316) 788-7300,,,,,,Phase I En
6489230027    620-723-2638    ,,,,Dan Hayse - 620-723-2638,,,,,,Phase I En
6489230346    620-376-4499    Bertha Mangold -620-376-4499,,,,,,Phase I En
...
3772517888-GZIP-28322  (831) 373-5555  onterey-<nobr>(831) 373-5555</nobr>
3772517888-GZIP-29518  (831) 899-8300  Seaside - <nobr>(831) 899-8300</nobr>
5054604751    716-871-2929    a%,888-571-2048,716-871-2929\x0D\x0ACPV,,,%Cape
```

Offset **Feature** **Context**

Designed for easy processing by python, perl or C++ program

- “Loosely ordered.”
- -GZIP- indicates that data was decompressed
- Non-UTF-8 characters are escaped

Histogram system automatically summarizes features.

```
# UTF-8 Byte Order Marker; see http://unicode.org/faq/utf_bom.html
# bulk_extractor-Version: 1.3b1-dev2
# Filename: /corp/nps/drives/nps-2009-m57-patents/charlie-2009-12-11.E01
# Feature-Recorder: email
# Histogram-File-Version: 1.1
...
n=875   mozilla@kewis.ch                    (utf16=3)
n=651   charlie@m57.biz                     (utf16=120)
n=605   ajbanck@planet.nl
...
n=288   mattwillis@gmail.com
n=281   garths@oeone.com
n=226   michael.buettner@sun.com            (utf16=2)
n=225   bugzilla@babylonsounds.com
n=218   berend.cornelius@sun.com
n=210   ips@mail.ips.es
n=201   mschroeder@mozilla.x-home.org
n=186   pat@m57.biz      (utf16=1)
```

-

Histogram of search terms can convey intent.

```
# UTF-8 Byte Order Marker; see http://unicode.org/faq/utf_bom.html
# bulk_extractor-Version: 1.3b1-dev2
# Filename: /corp/nps/drives/nps-2009-m57-patents/charlie-2009-12-11.E01
# Feature-Recorder: url
# Histogram-File-Version: 1.1
n=59    1
n=53    exotic+car+dealer
n=41    ford+car+dealer
n=34    2009+Shelby
n=25    steganography
n=23    General+Electric
n=23    time+travel
n=19    steganography+tool+free
n=19    vacation+packages
n=16    firefox
n=16    quicktime
n=14    7zip
n=14    fox+news
n=13    hex+editor
```

-

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.

New in bulk_extractor 1.3

New supported data types:

- Windows PE Scanner
- Linux ELF Scanner
- VCARD Scanner
- BASE16 scanner
- Windows directory carver

New Histogram options:

- Numeric only option for phone numbers
- Supports new Facebook ID

Better Unicode Support:

- Histograms now UTF-8 / UTF-16 aware
- Feature files are UTF-8 clean

Limited support for file carving:

- packets carved into pcap files
- VCARD carver

bulk_extractor: Open Source, GOTS, in use today.

bulk_extractor has been rapidly developed.

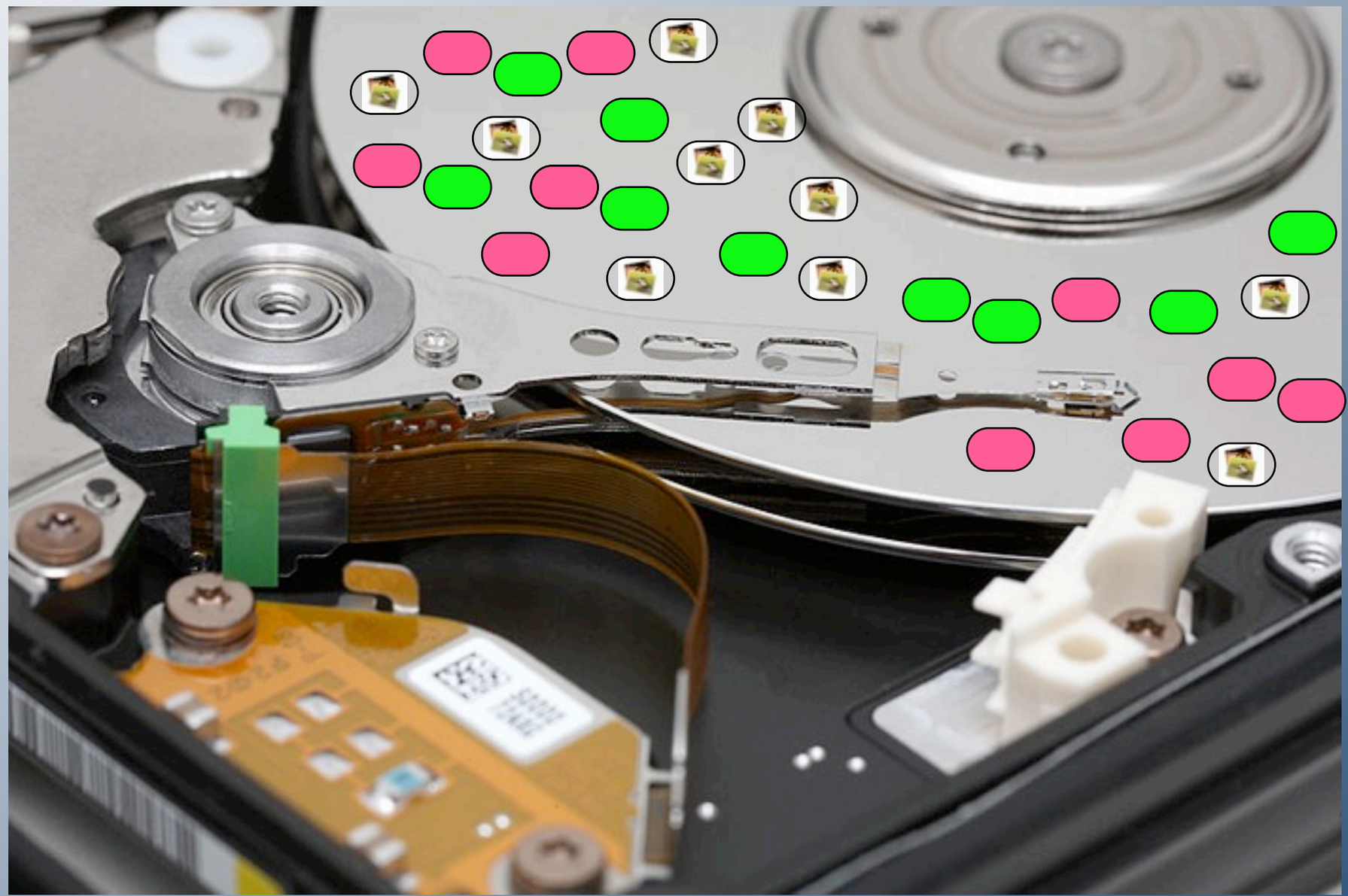
- April 2010 — Initial public release
- Jun. 2010 — Release 1.0
- Feb. 2011 — Release 1.2 (AES, Network Scanning, etc.)

Widely used today by:

- US Government
- State and Local
- Foreign Partners
- Researchers

Winner 2011 DoD Value Engineering Achievement Award

Available for download from http://afflib.org/



Random Sampling

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

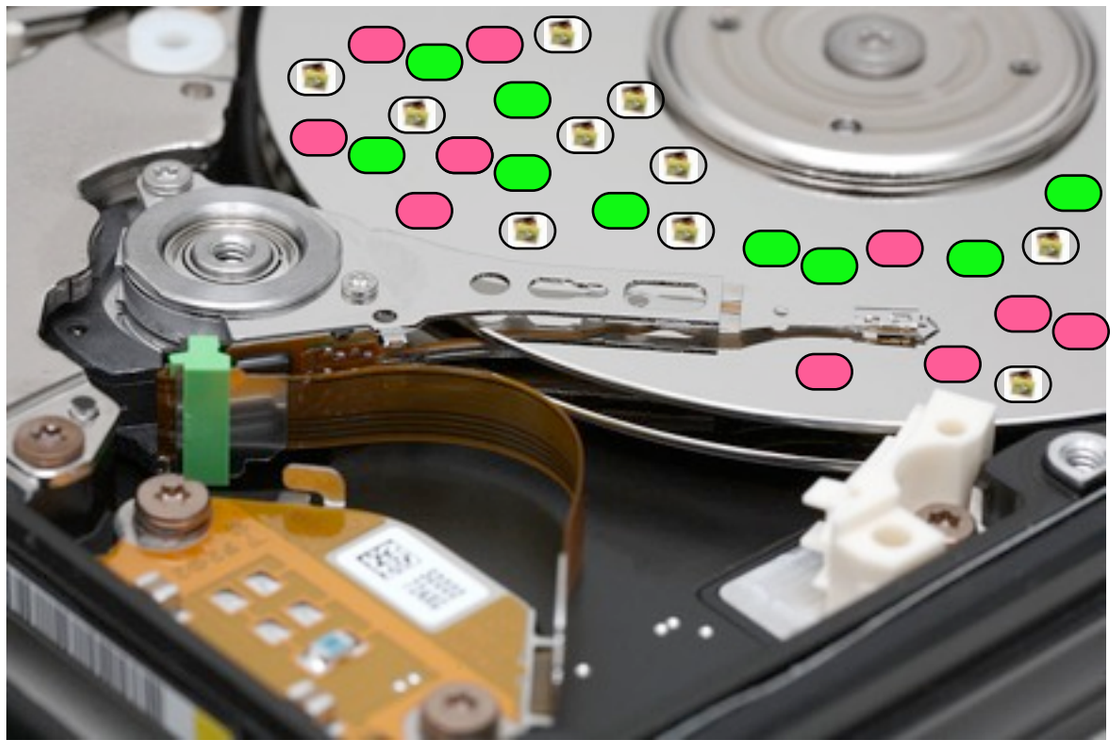
36 GB is a lot of data!

- ≈ 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:

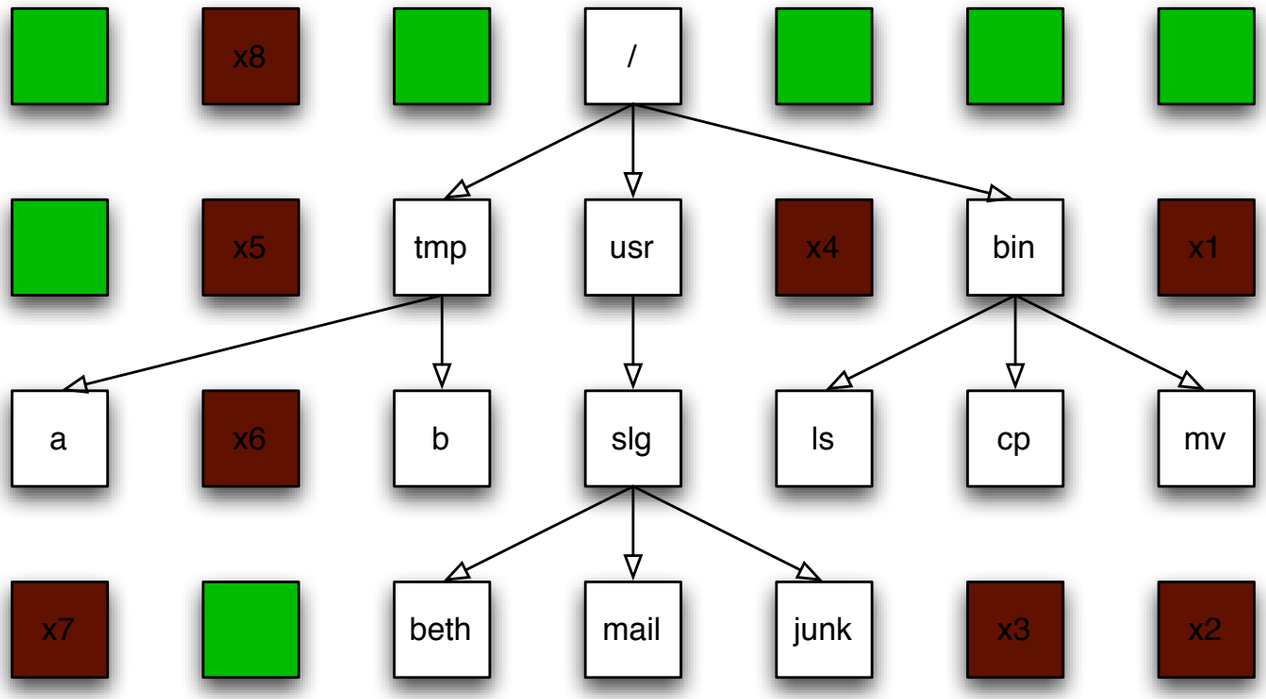
Hard drive contents can be predicted by sampling a few thousand sectors:

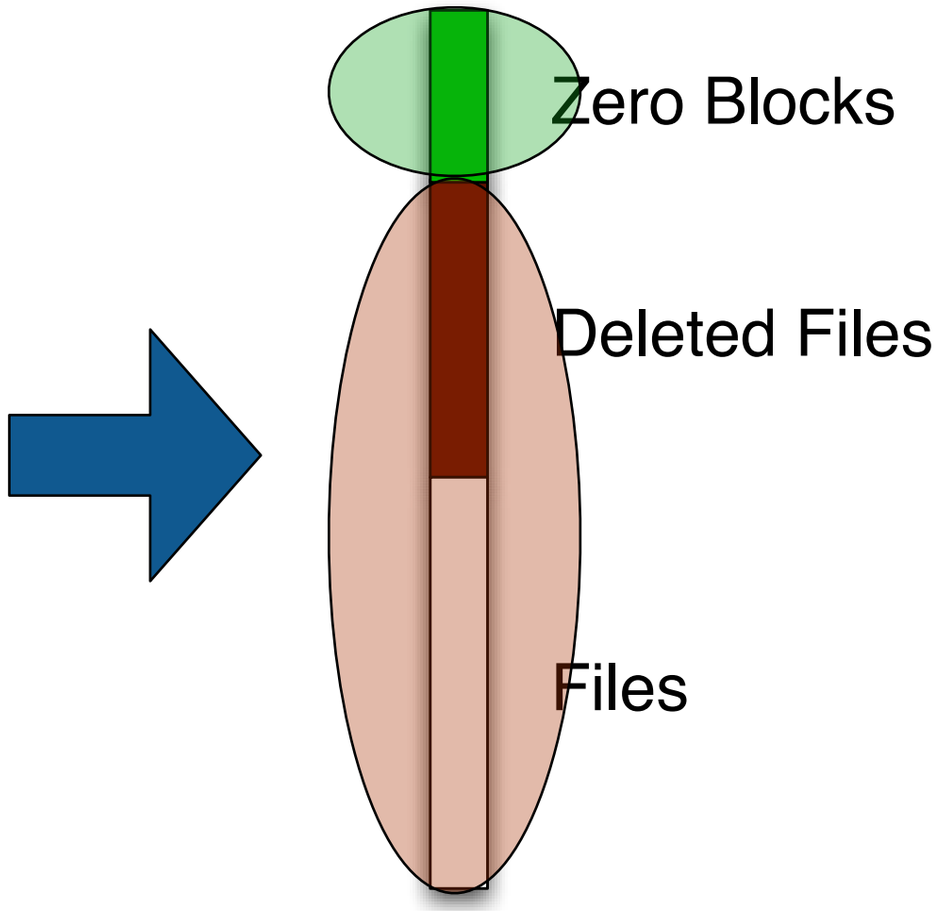


The challenge is identifying *likely voters.*

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

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

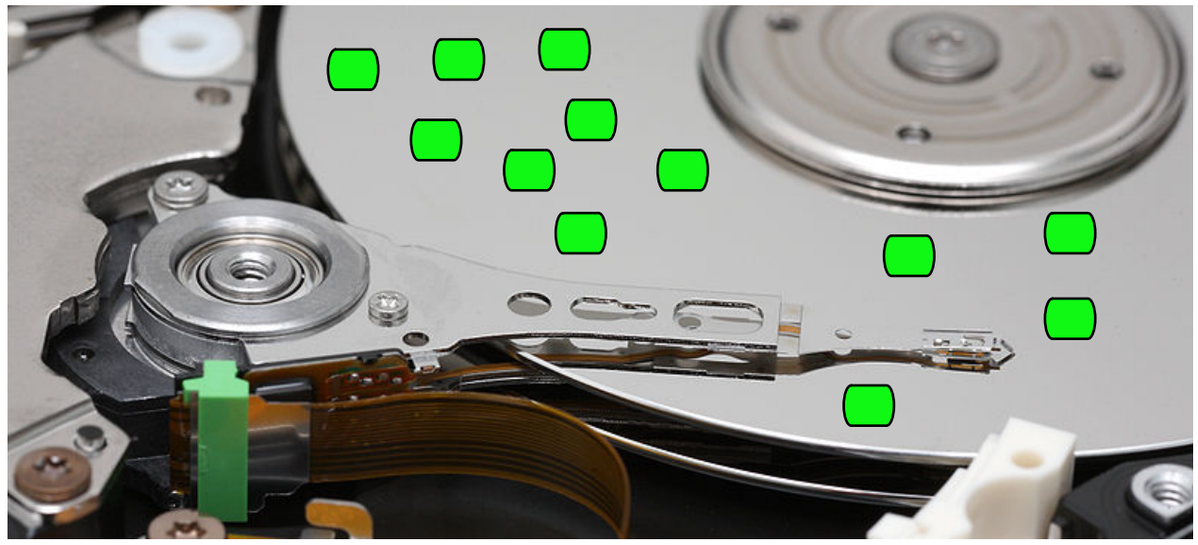




Simplify the problem.
Can we use statistical sampling to verify wiping?

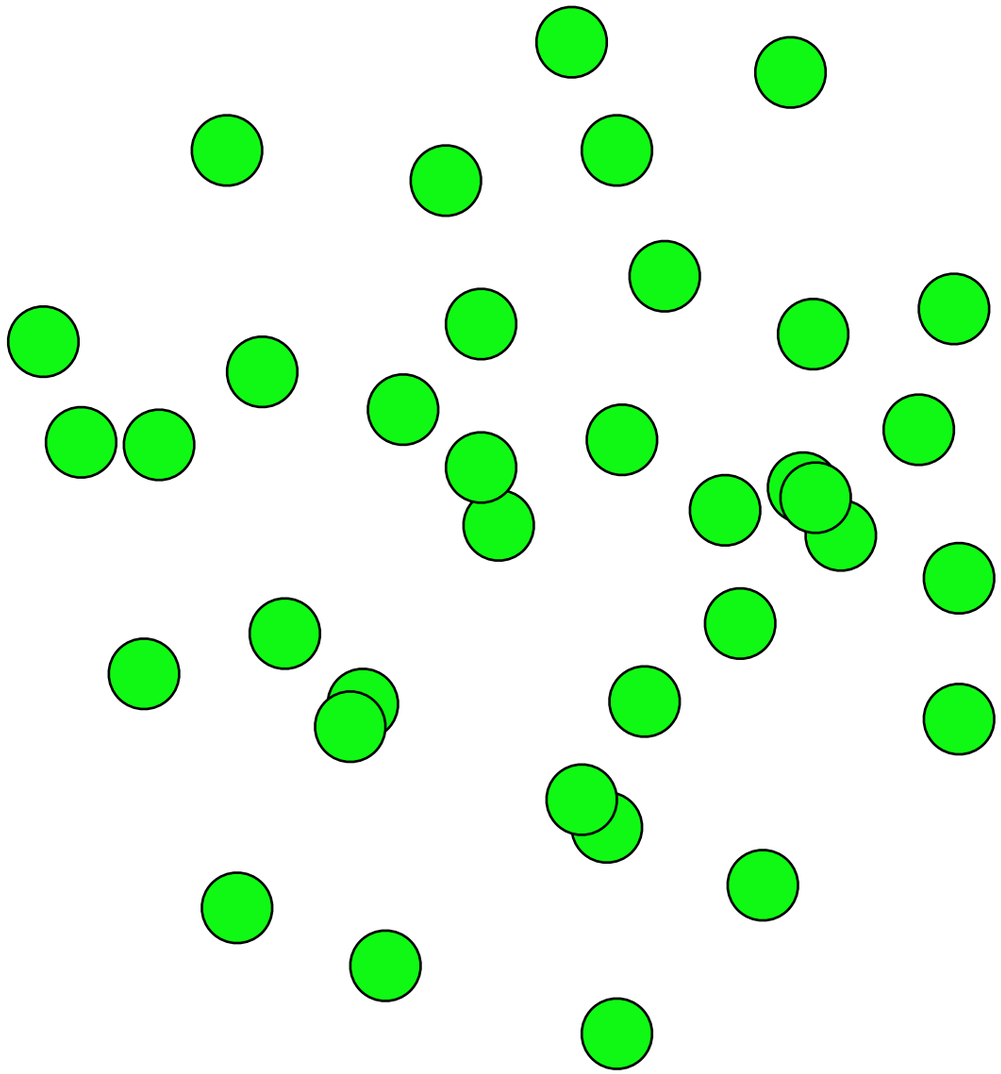
Many organizations discard used computers.

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

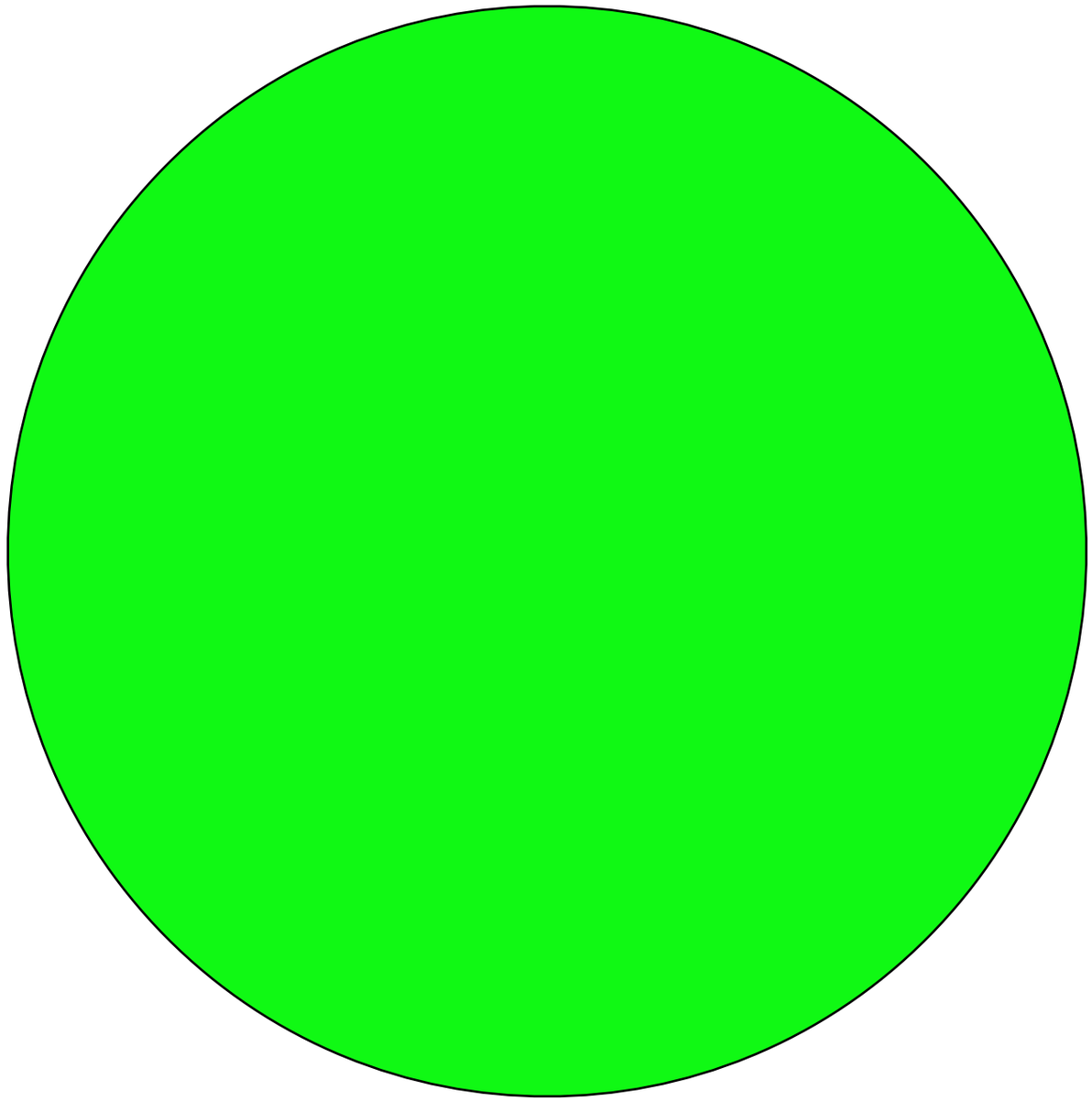


A 1TB drive has 2 billion sectors.
What if we read 10,000 and they are all blank?

A 1TB drive has 2 billion sectors.
What if we read 10,000 and they are all blank?



A 1TB drive has 2 billion sectors.
What if we read 10,000 and they are all blank?

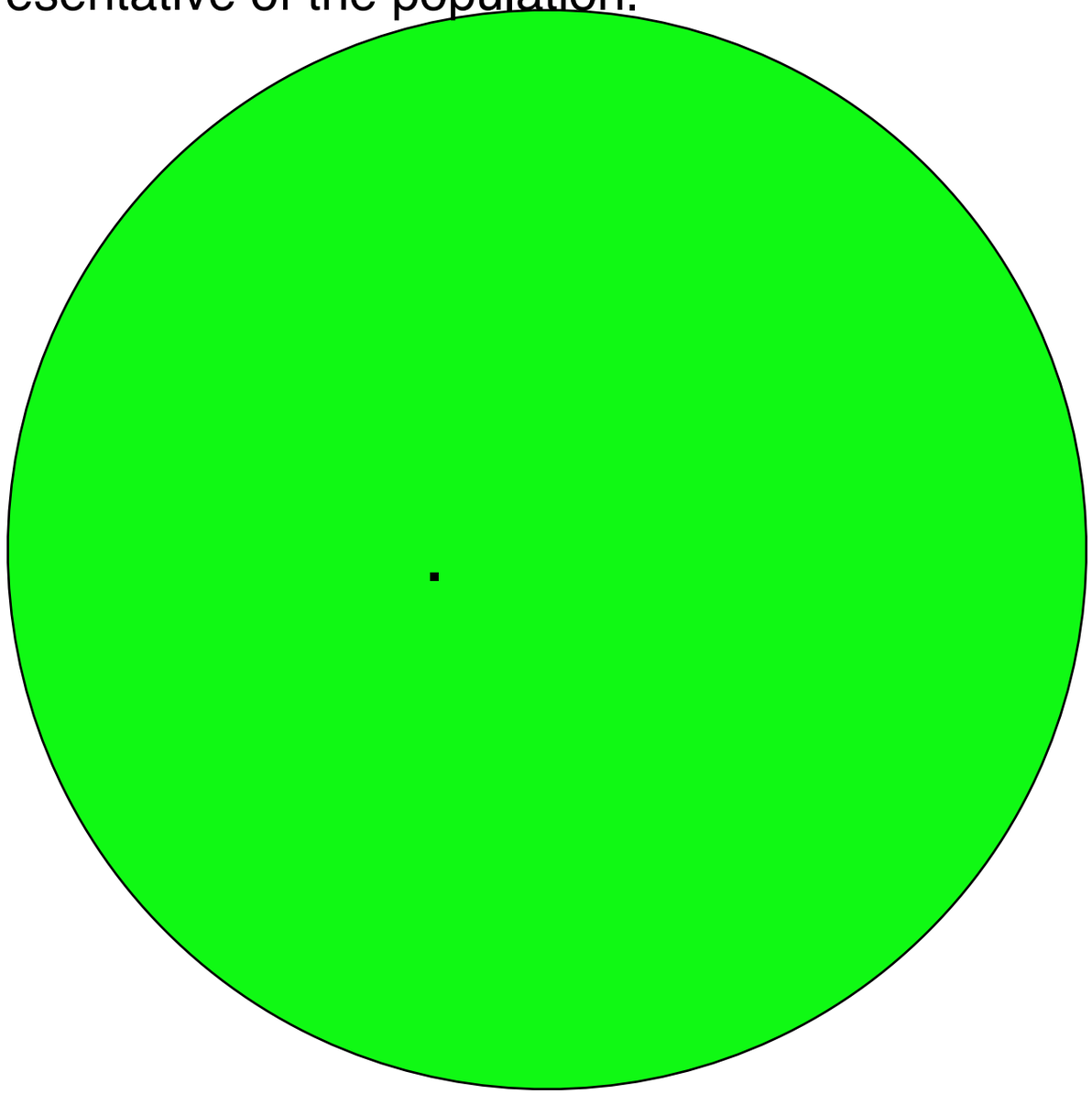


Chances are good that they are all blank.

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.

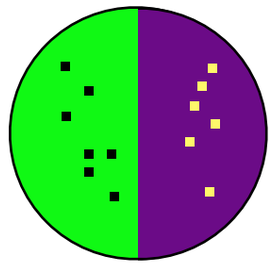


We will only find that 1 sector with exhaustive search.

What about other distributions?

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

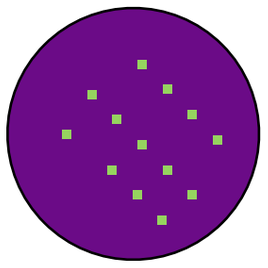
- The sampled frequency should match the distribution.
- *This is why we use random sampling.*



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 a 1TB disk has 10MB of data.

- 1TB = 2,000,000,000 = 2 Billion 512-byte sectors!
- 10MB = 20,000 sectors

Read just 1 sector; the odds that it is blank are:

$$\frac{2,000,000,000 - 20,000}{2,000,000,000} = .99999$$

Read 2 sectors. The odds that both are blank are:

$$\left(\frac{2,000,000,000 - 20,000}{2,000,000,000}\right)\left(\frac{1,999,999,999 - 20,000}{2,000,000,000}\right) = .99998$$

first pick **second pick** **Odds we may have missed something**

The more sectors picked, the less likely we are to miss the 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.

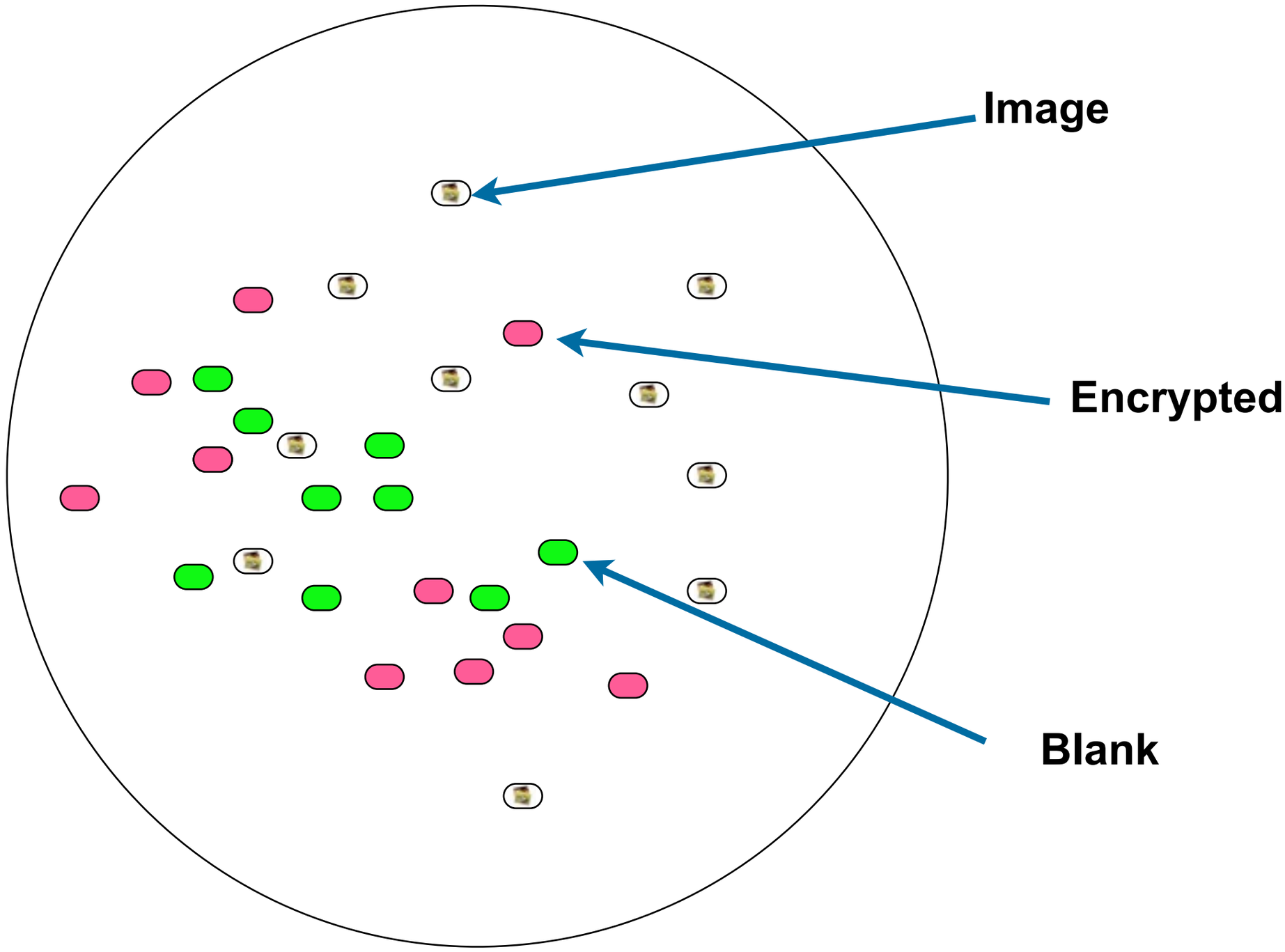
—So pick 500,000 random sectors. If they are all NULL, then the disk has p=(1-.00673) chance of having 10MB of non-NULL data.

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.

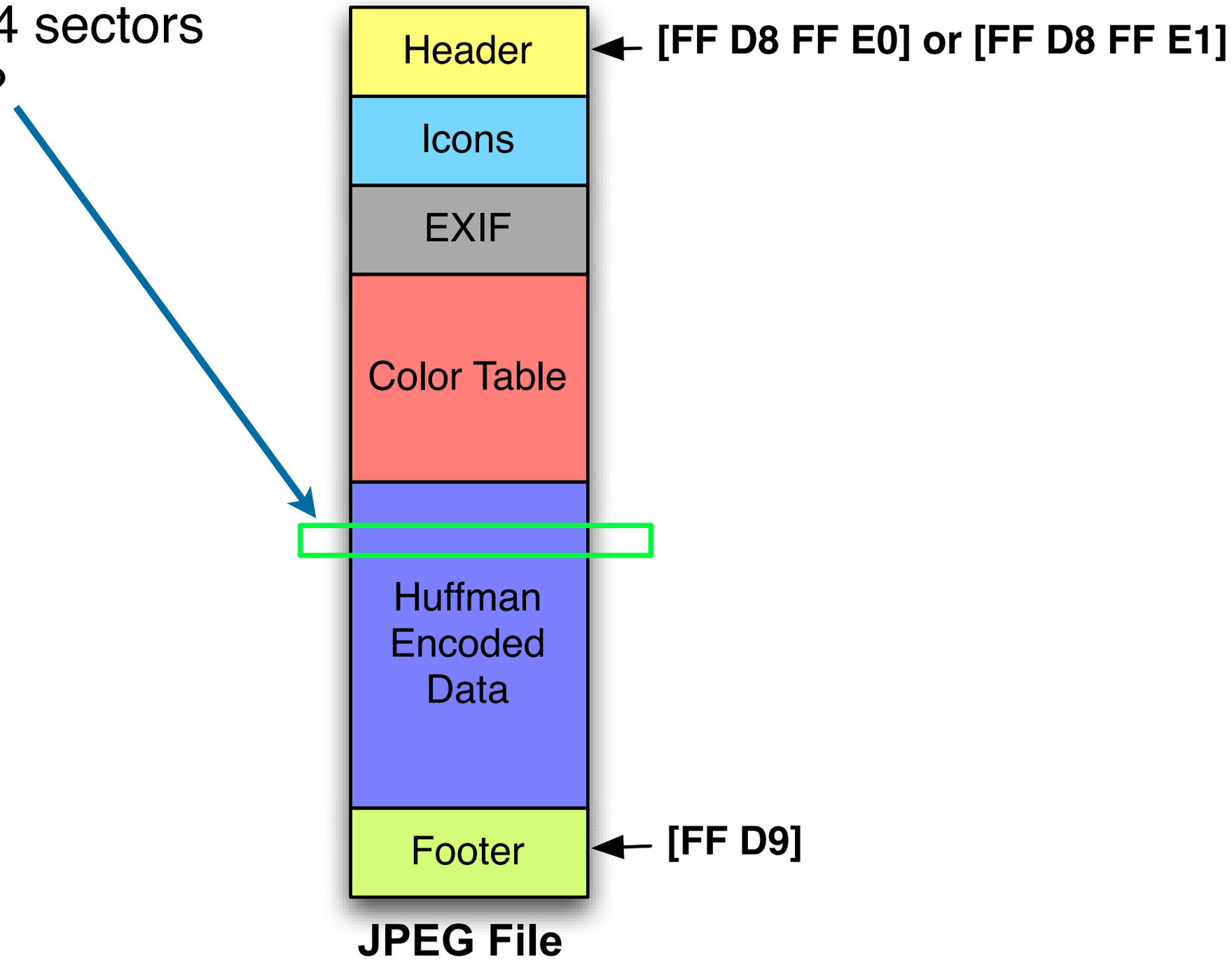
We get a statistically significant sample in two minutes.



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

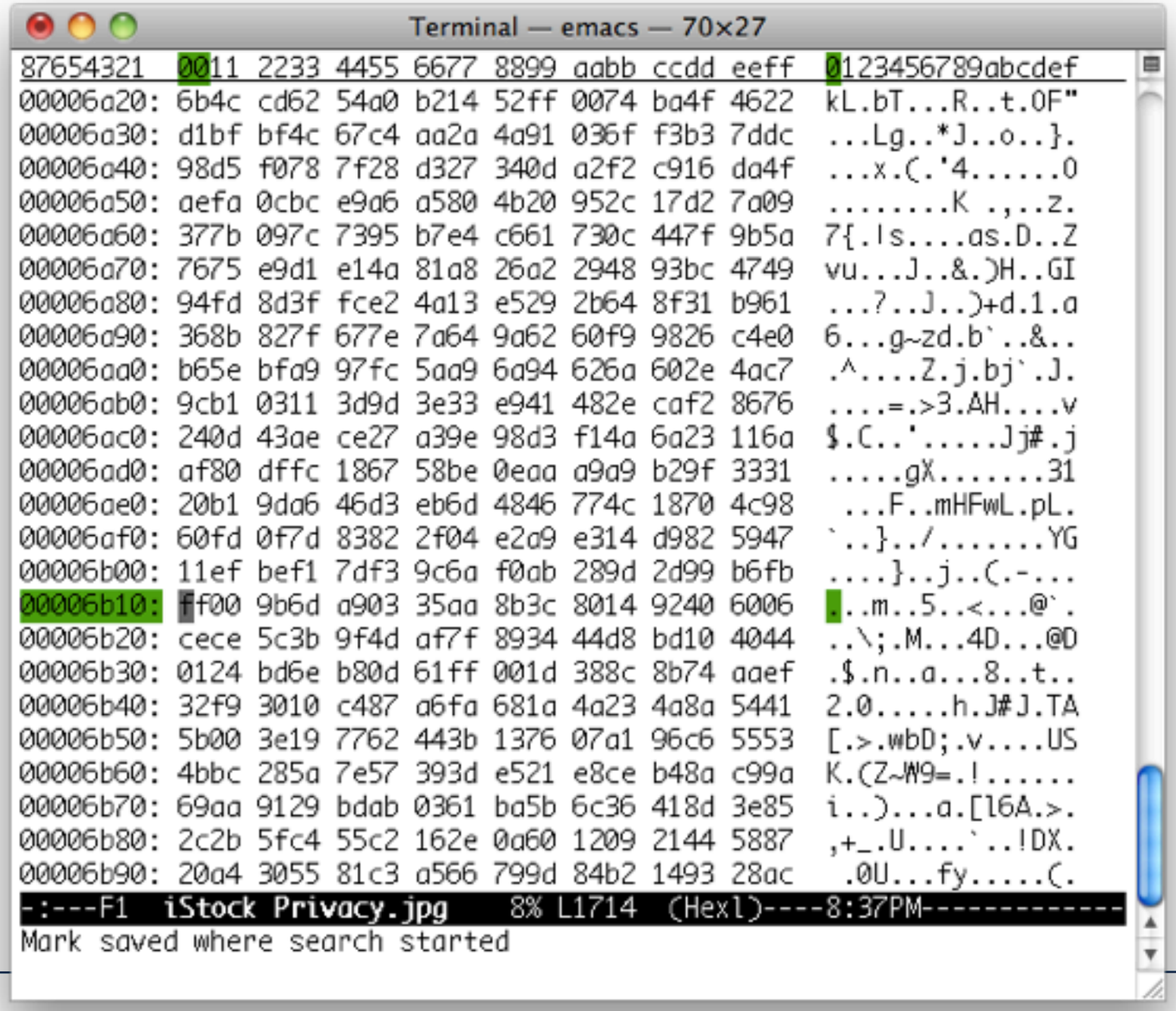
The challenge: identifying a file “type” from a fragment.

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



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

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



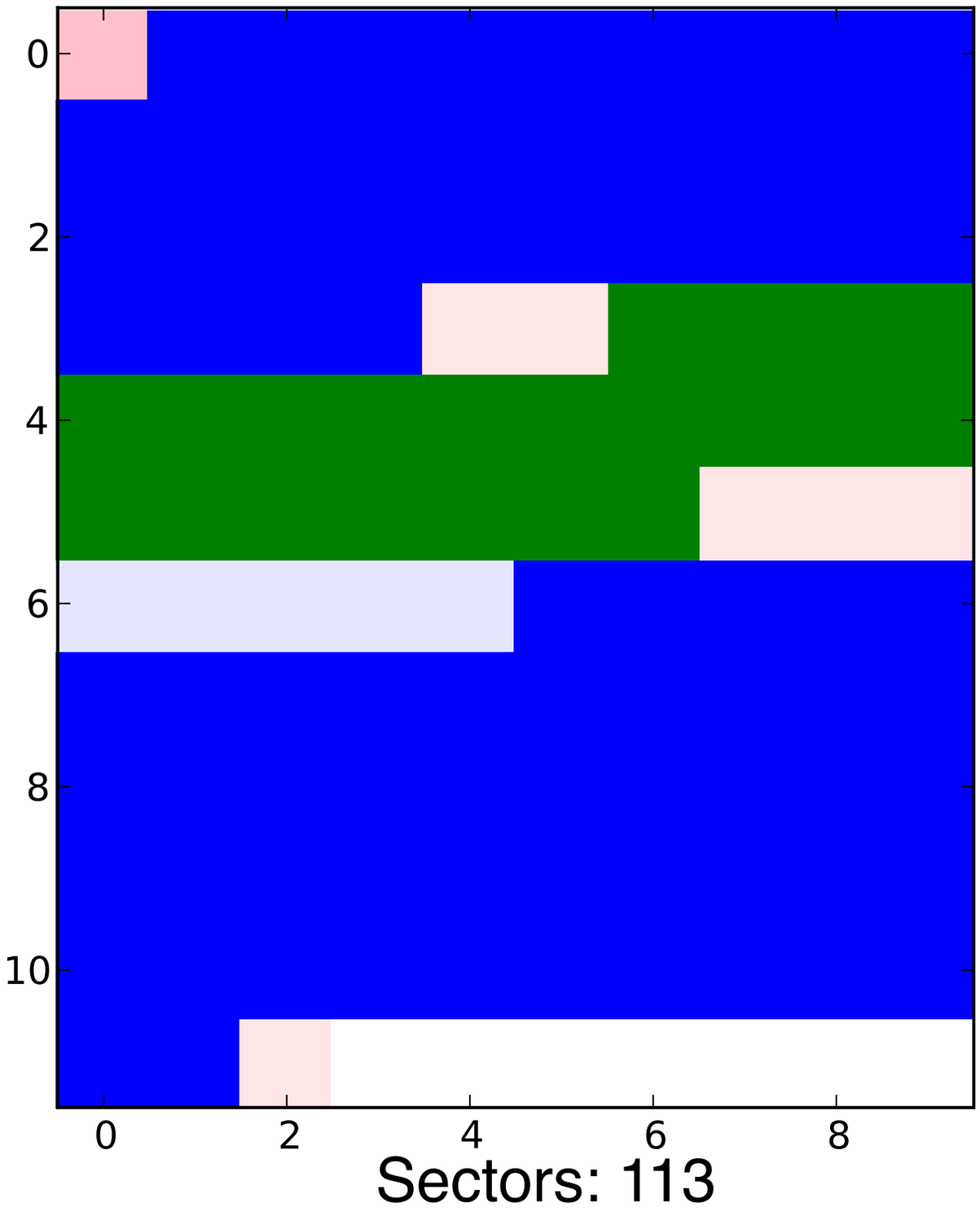
We built detectors to recognize the different parts of a JPEG file.

Bytes: 31,046

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

000897.jpg

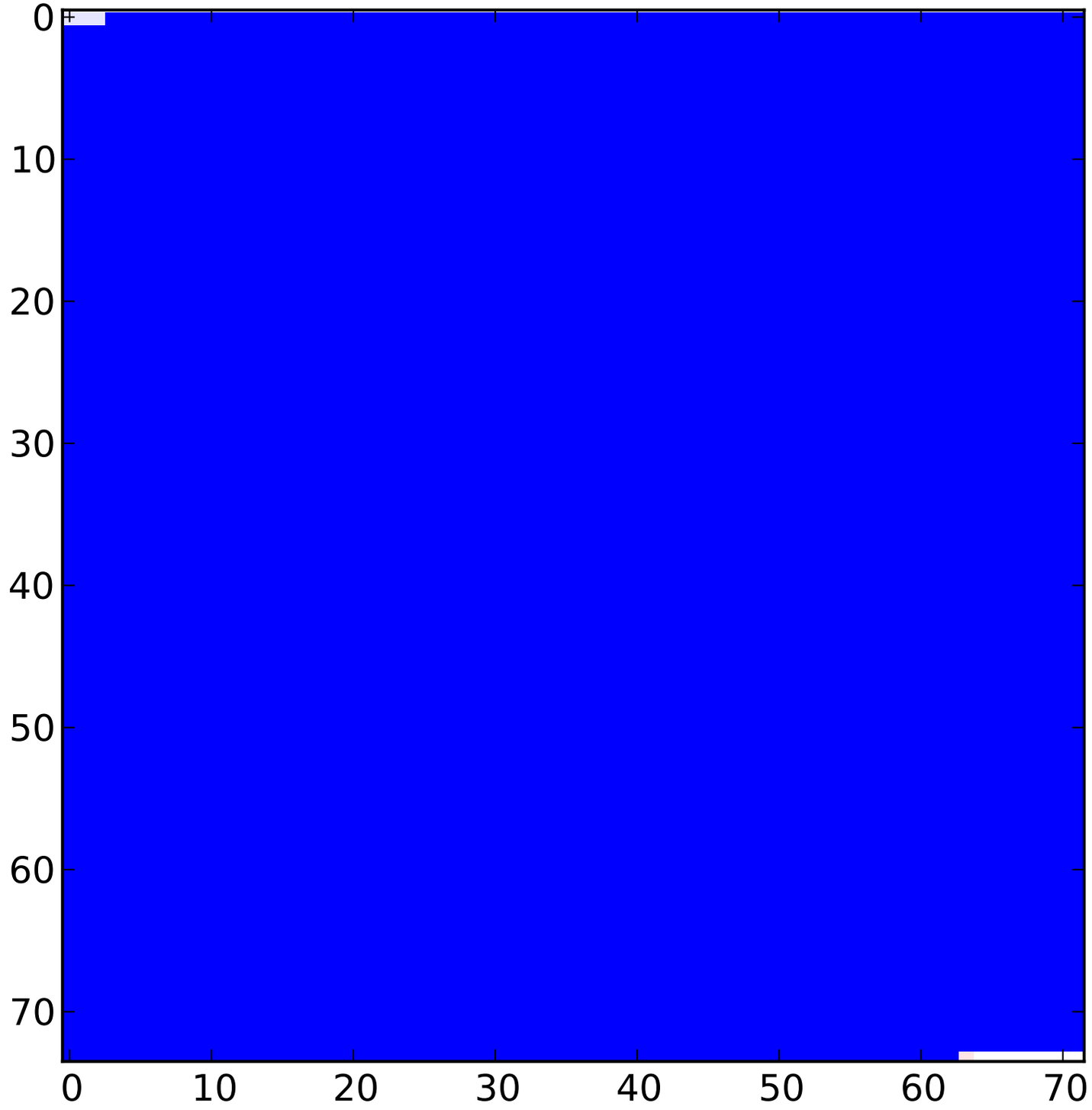
Bytes: 57596



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

000888.jpg

Bytes: 2,723,425



We developed five sector identification tools

JPEG — Single images.

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

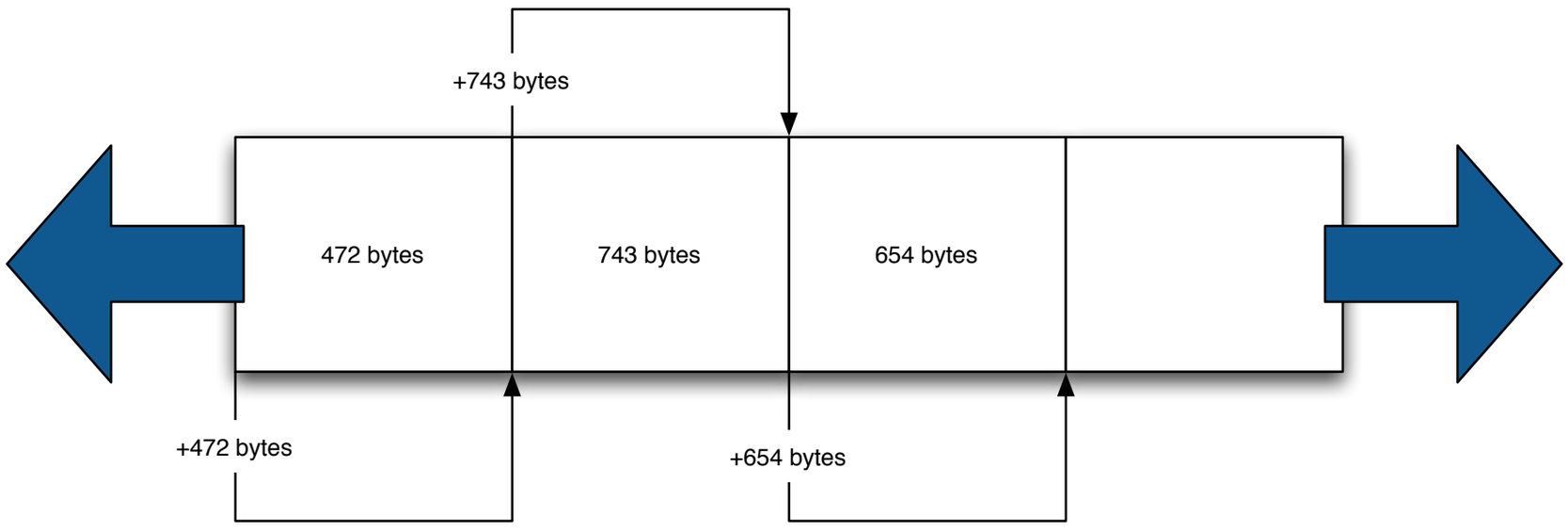
This is called the *file fragment classification problem*.

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>
```

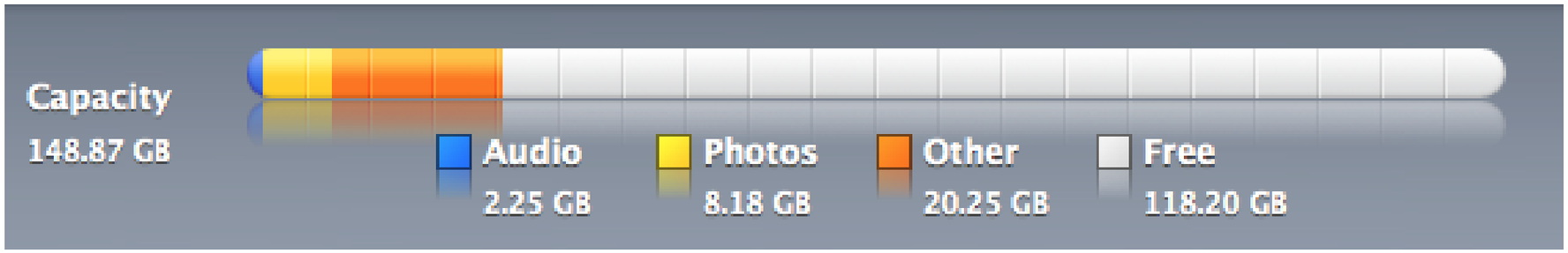
MPEG files can be readily identified through framing

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



Combine random sampling with sector ID 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

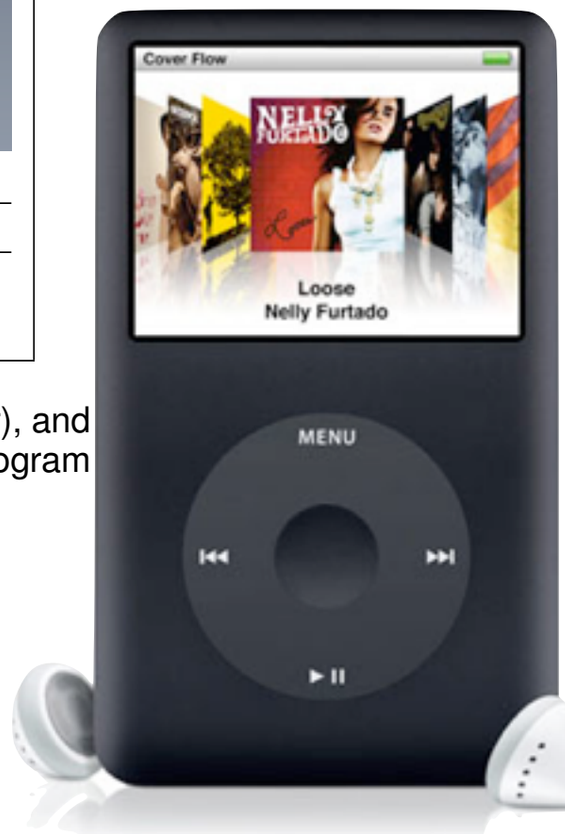


Figure 1: Usage of a 160GB iPod reported by iTunes 8.2.1 (6) (top), as reported by the file system (bottom center), and as computing with random sampling (bottom right). Note that iTunes usage actually in GiB, even though the program displays the "GB" label.

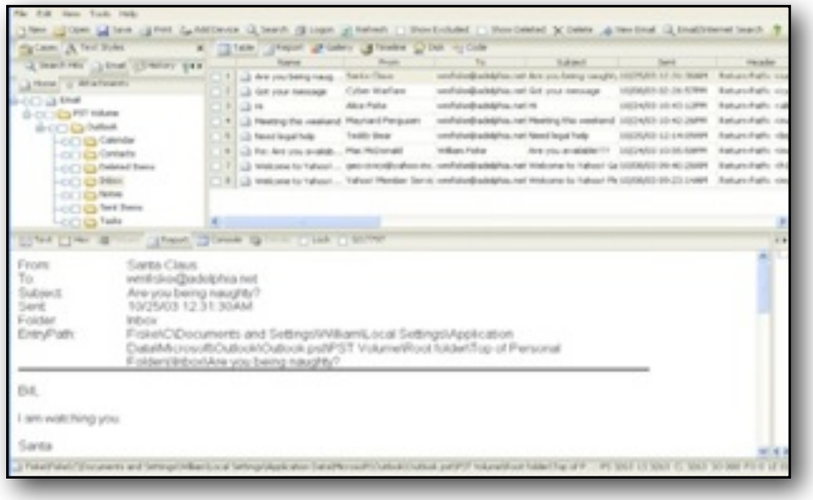
We accurately determined:

- % 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.*

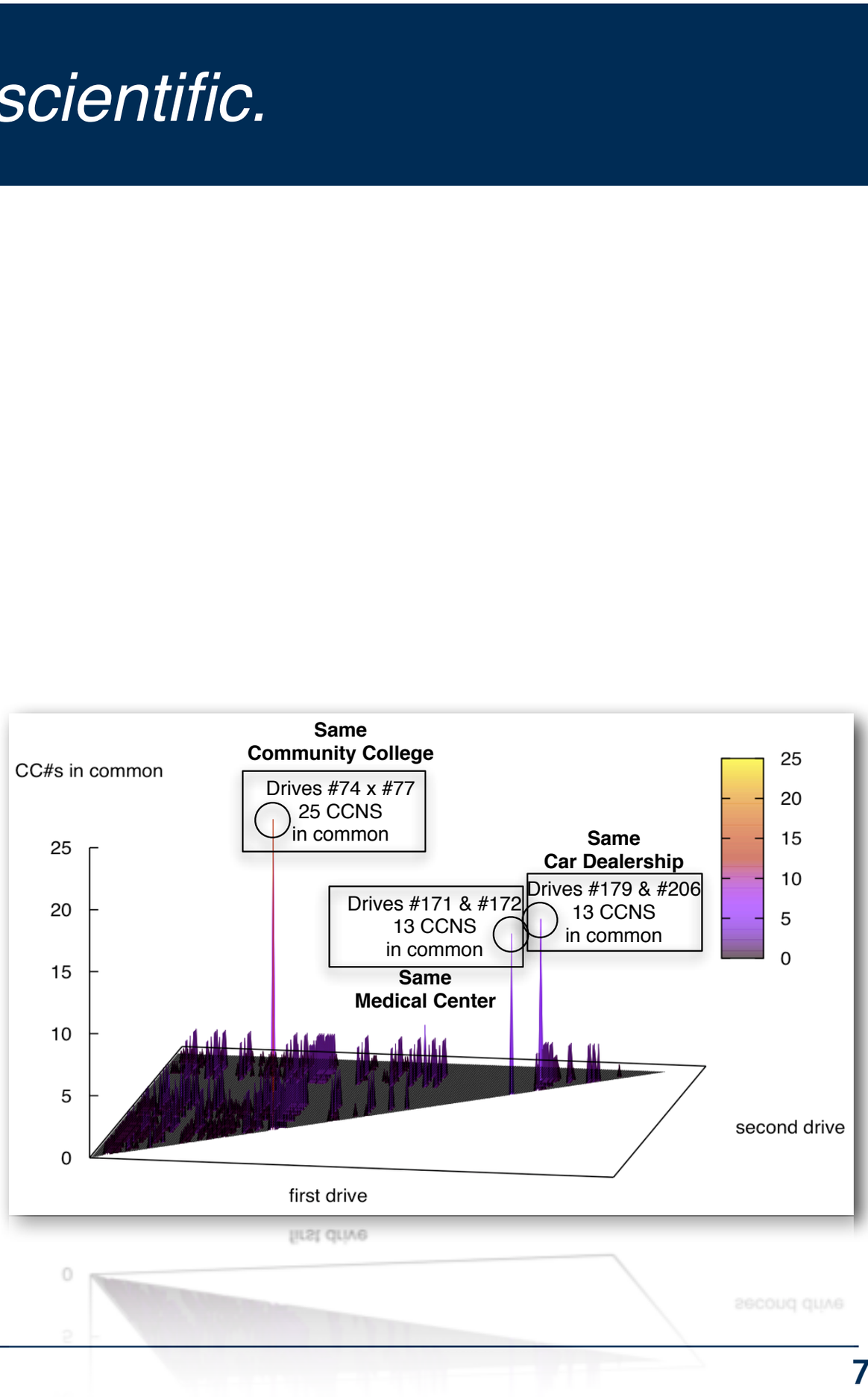
Evidence Reconstruction

- 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.)

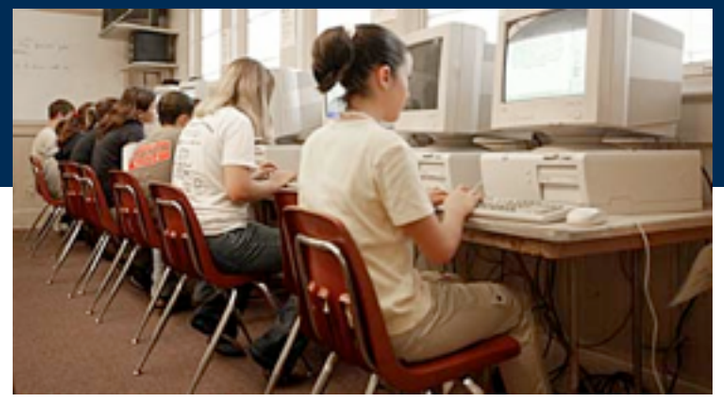
Mobile Phone Application Corpus

- 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*

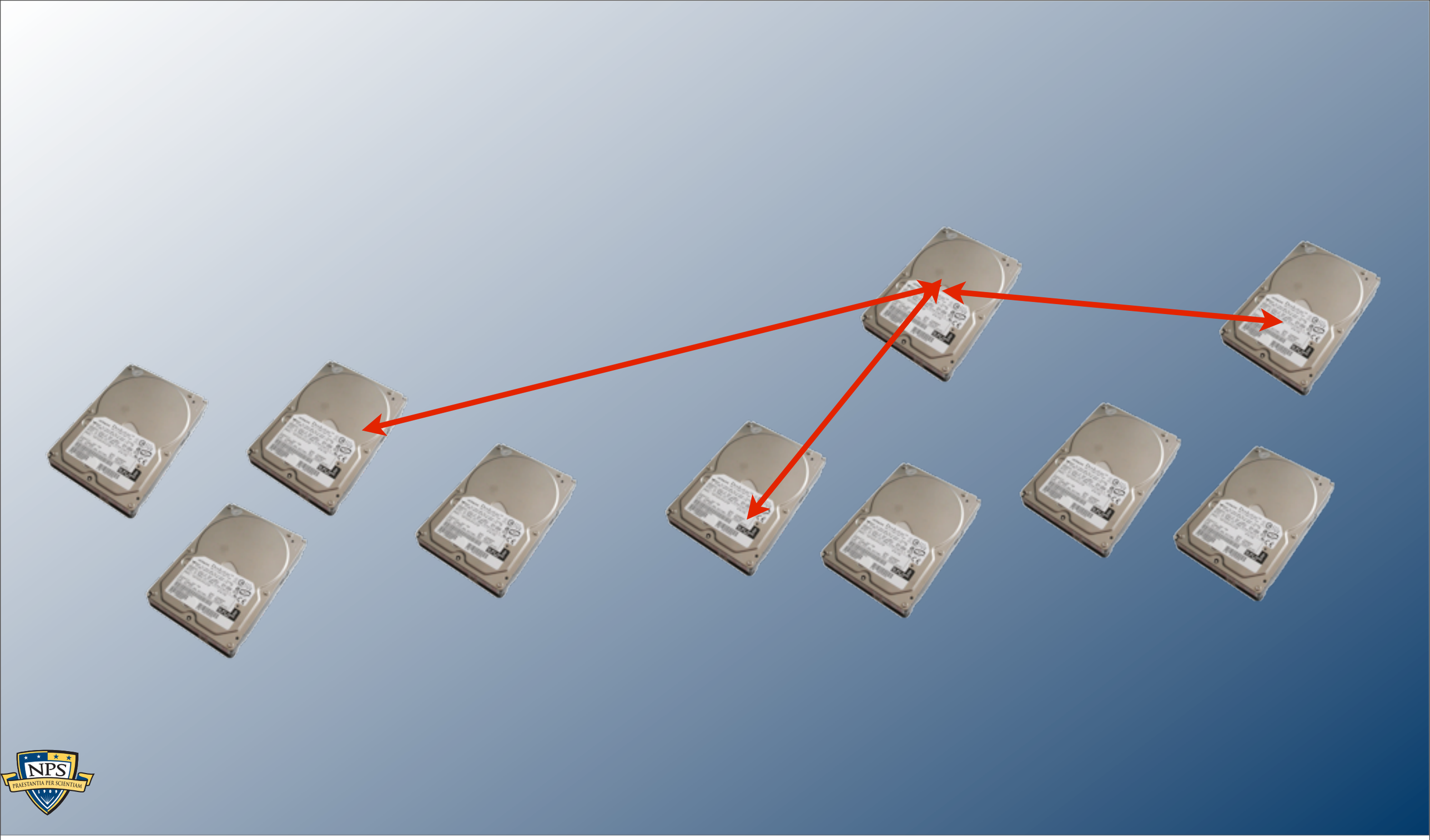
University harassment scenario

- 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.

Our challenge: 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.*
- Leverage Open Source Software and Community

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