Automated Forensics Research at NPS

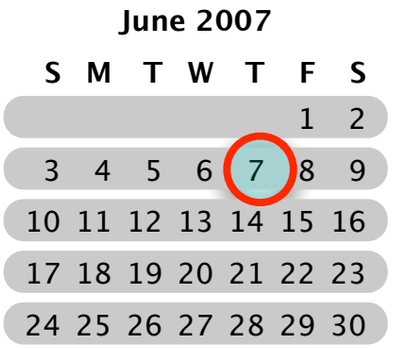
Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

May 12, 2010

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

Automated Document and Media Exploitation: The Need

Law enforcement & military encounter substantial amounts of electronic media.

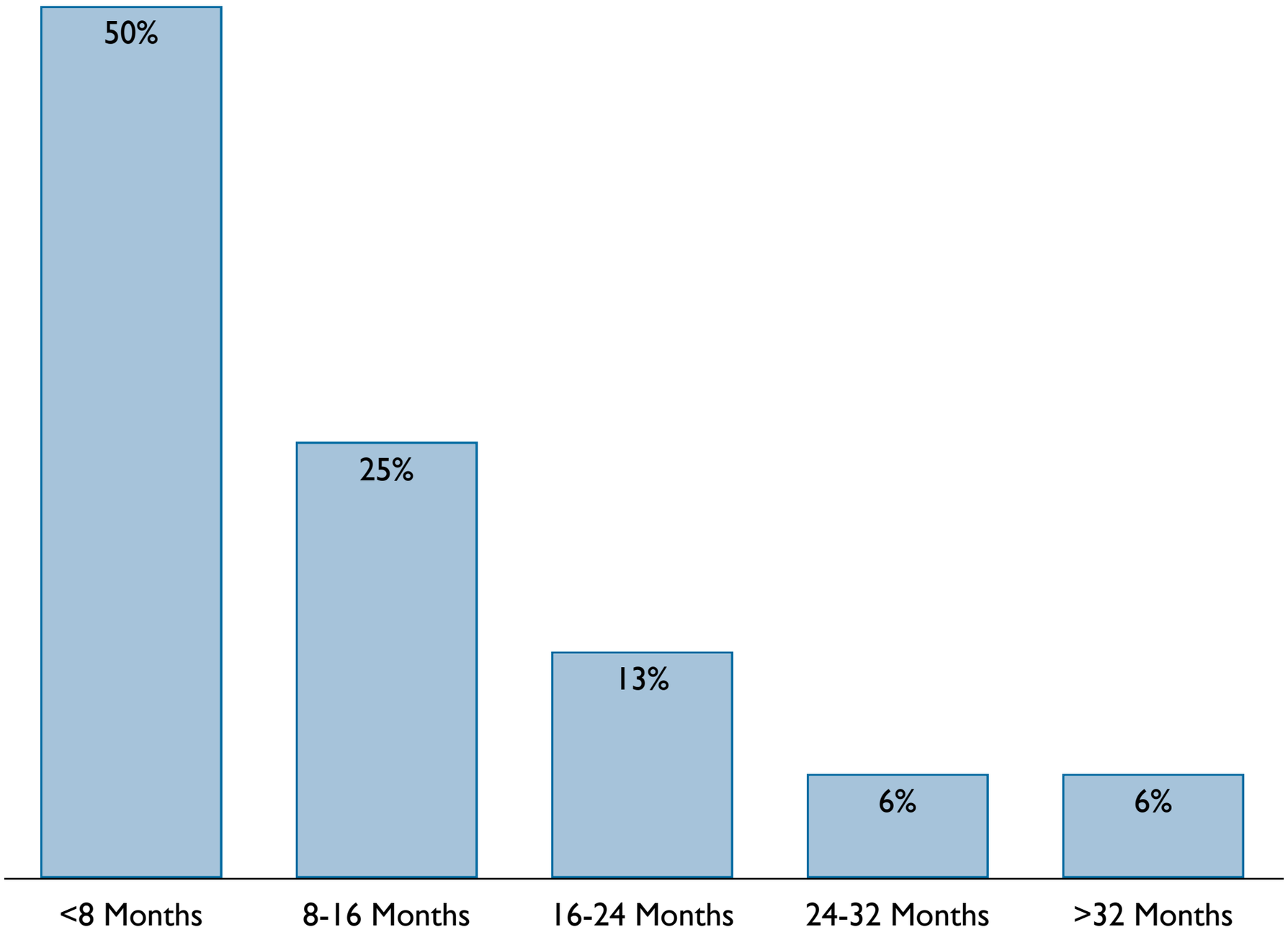


- Battlefield
- Checkpoints & Border crossings
- Law enforcement operations
- Internal Investigations

FBI RCFL FY08 Annual Report:

- Examinations: 4,524 ↑
- Cell Phones: 2,226 ↑
- Hard Drives Processed: 17,511 ↑
- Total Data Processed: 1,756 TB ↑

Media size, quantity and diversity is increasing geometrically.

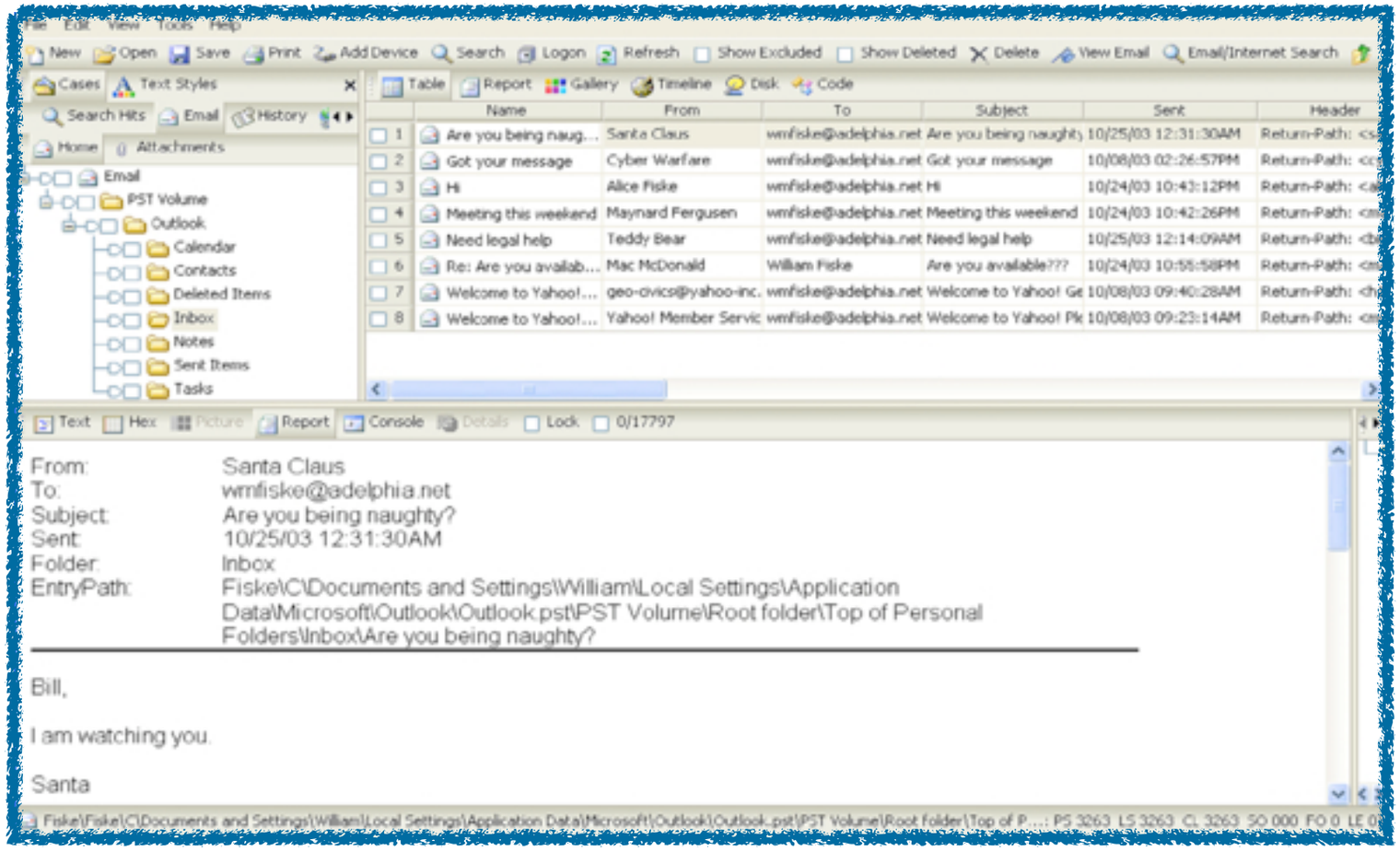


(Source: Defense Cyber Crime Center)

Most media is analyzed using highly trained personnel...

DOMEX in Iraq

… working with tools designed for law enforcement.

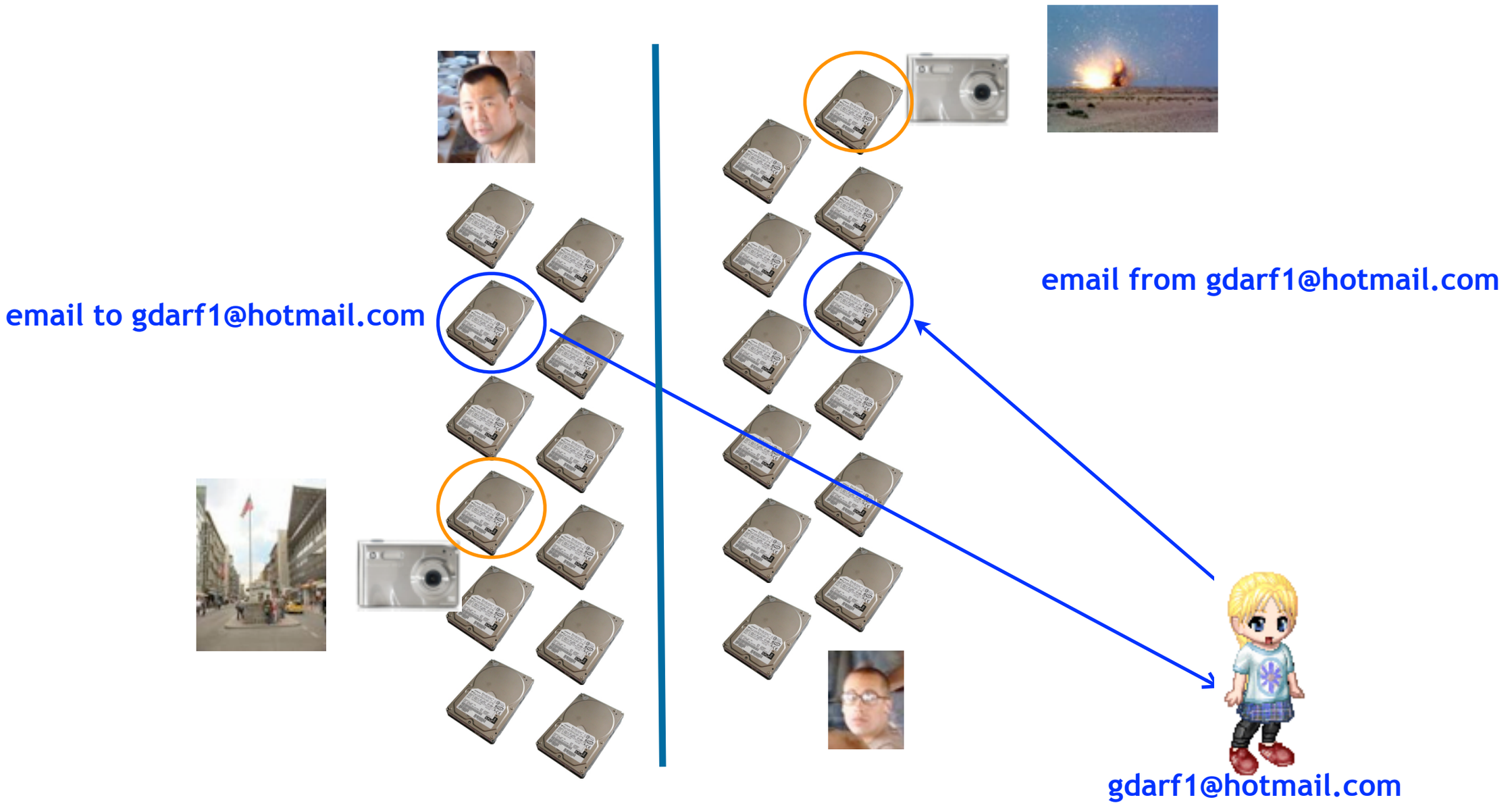


EnCase by Guidance Software

- Designed for visibility & search, not analysis.
- Does not scale to 100s or 1000s of drives.
- Prevent “contamination” between cases

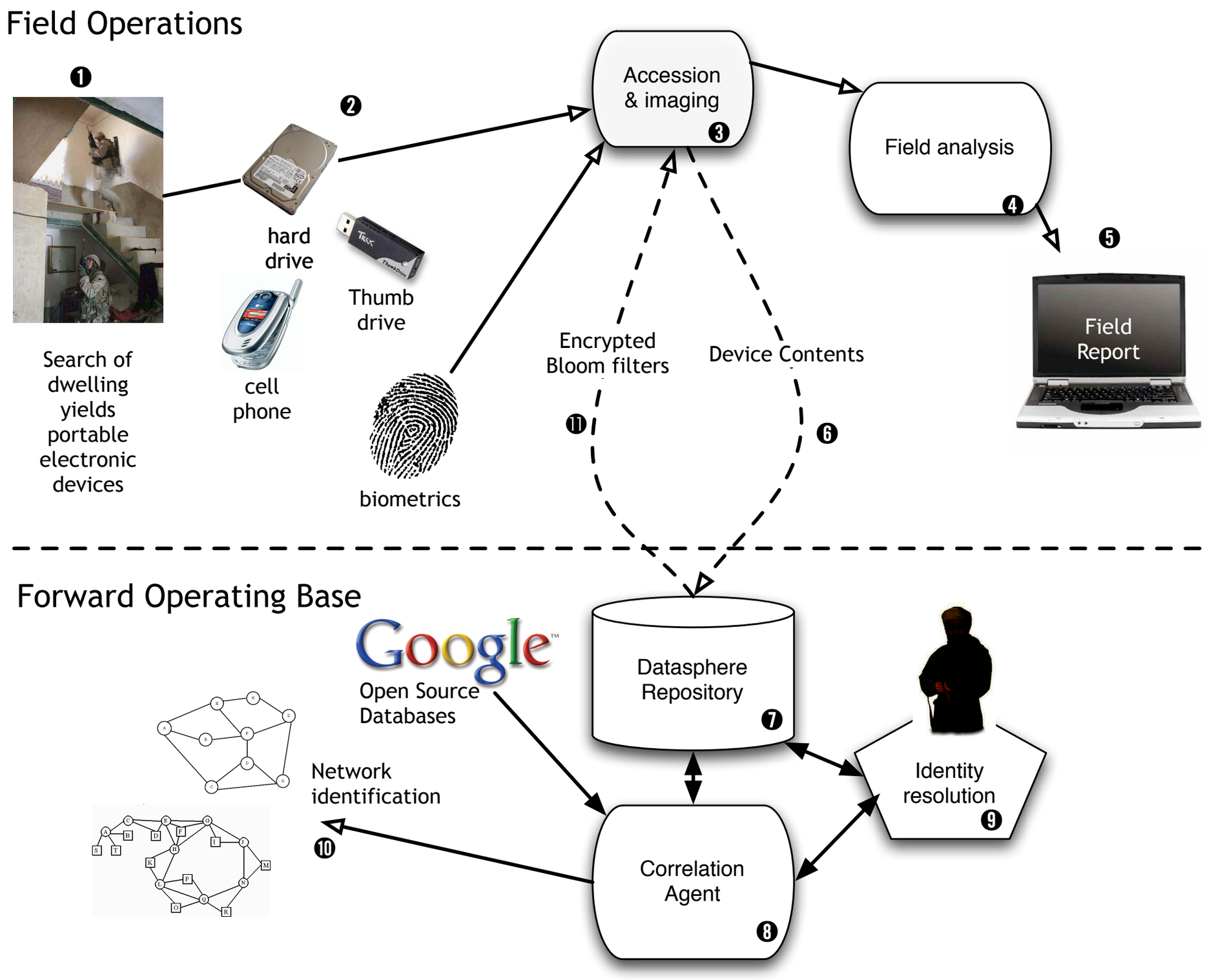
Manual analysis misses opportunities for correlation.

Different analysts see different hard drives.



Keyword searches don’t connect the dots.

Our research will help to speed workflow while providing for automated “situational awareness.”



We have four main research thrusts.

Area #1: End-to-end automation of forensic processing

- Digital evidence file formats; chain-of-custody (AFFLIB)
- Tool integration; automated metadata extraction

Area #2: Bringing data mining to forensics

- Automated social network analysis (cross-drive analysis)
- Automated ascription of carved data

Area #3: Bulk Data Analysis

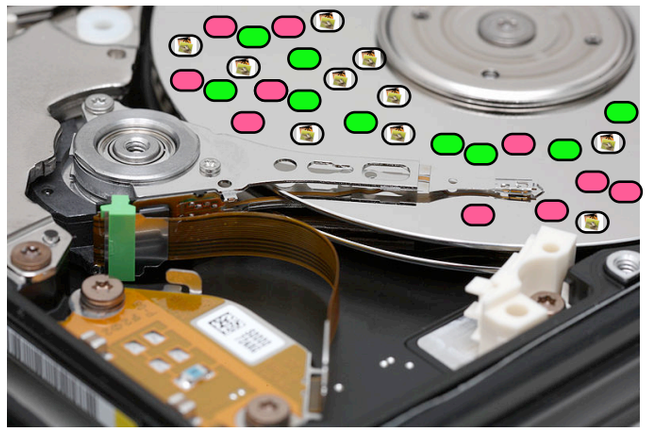
- Stream-processing
- Statistical techniques (sub-linear algorithms)

Area #4: Creating Standardized Forensic Corpora

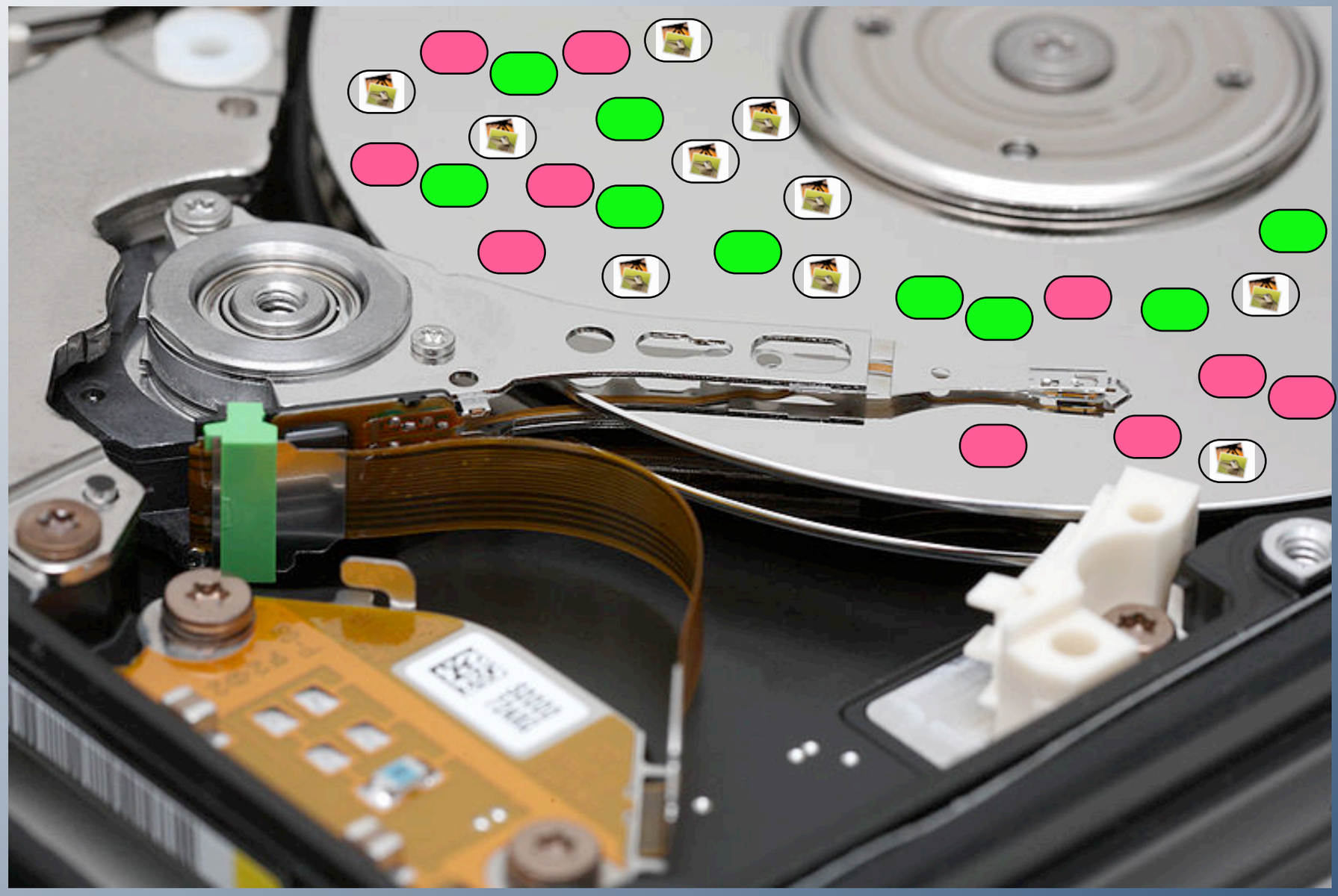
- Freely redistributable disk and memory images, packet dumps, file collections.

This talk focuses on two research projects:

Instant Drive Analysis



Standardized Forensic Corpora



Instant Drive Forensics with Statistical Sampling

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

What if US agents encounter a hard drive at a border crossing?

Or a search turns up a room filled with servers?

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

		
Minutes	208	1
Max Data Read	1 TB	4.8 GB

4.8 GB (0.48%) is a tiny fraction of the disk.

But 4.8 GB is a lot of data!

Data on hard drives divides into three categories:

Resident Data

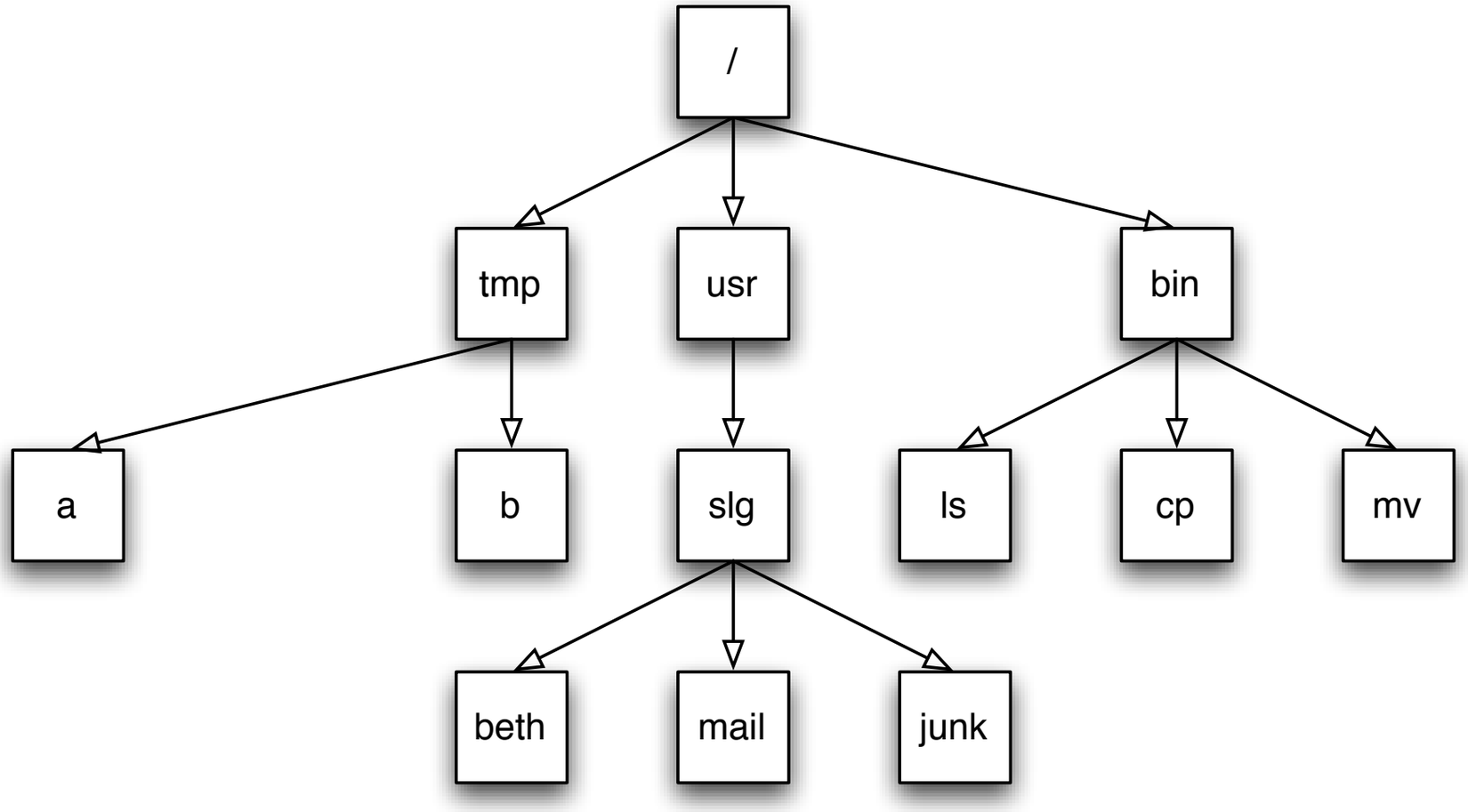
Deleted Data

No Data

user files
email messages
[temporary files]

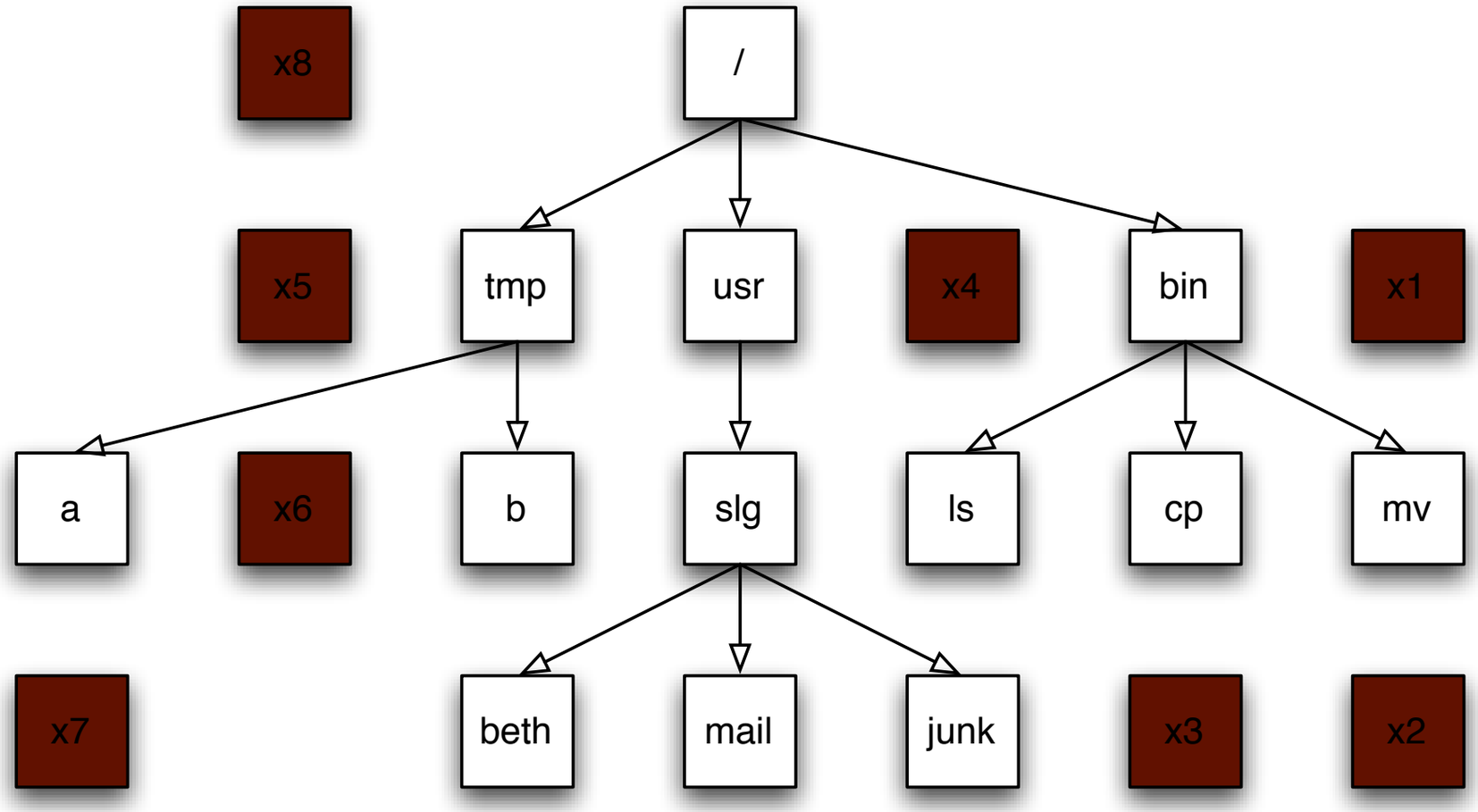
blank sectors

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



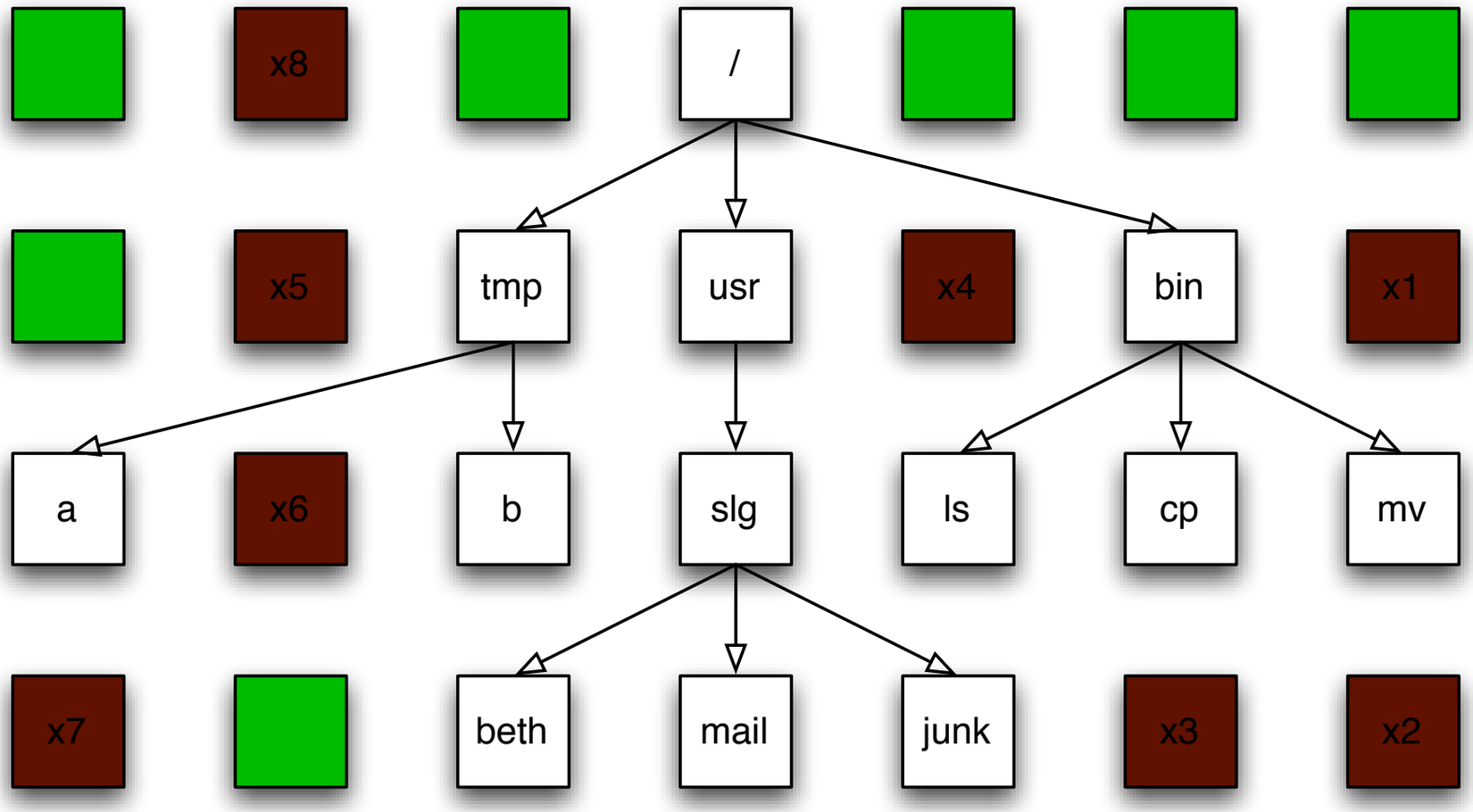
Resident Data

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



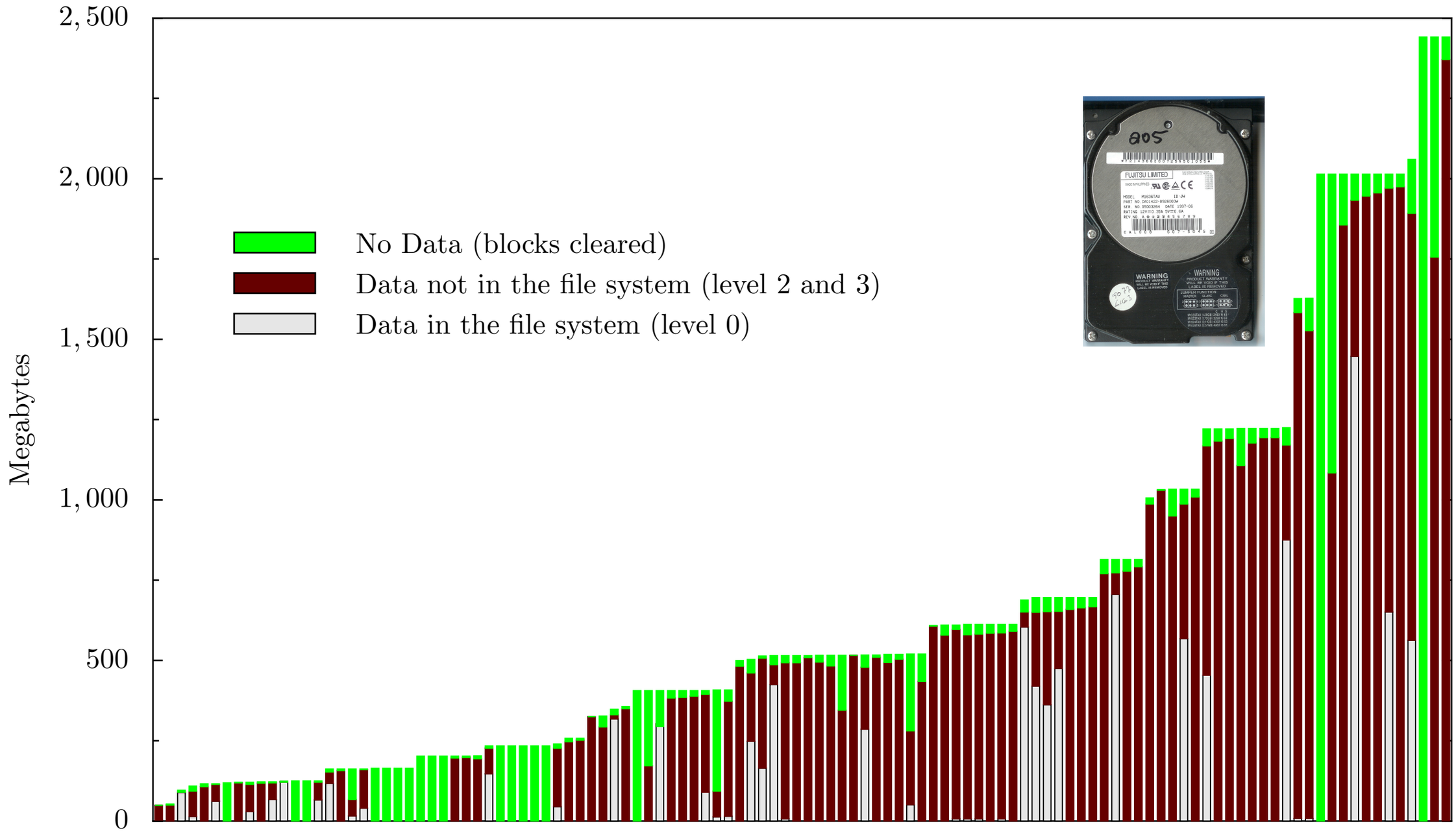
Deleted Data

Sectors with “No Data” are blank.



No Data

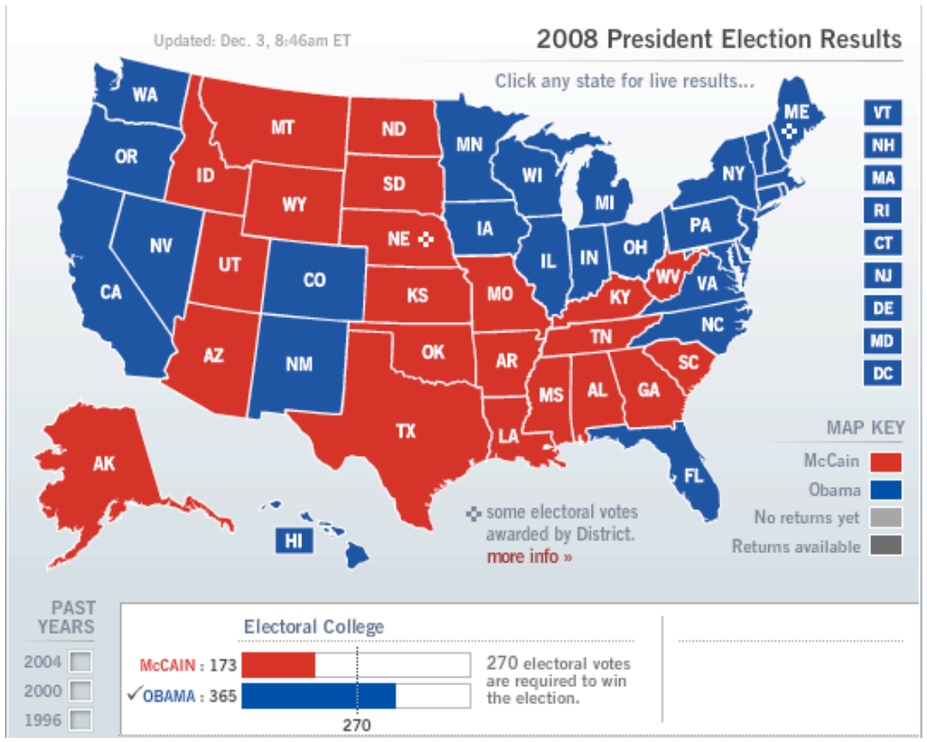
I bought 2000 hard drives between 1998 and 2006. Most of were not properly wiped.



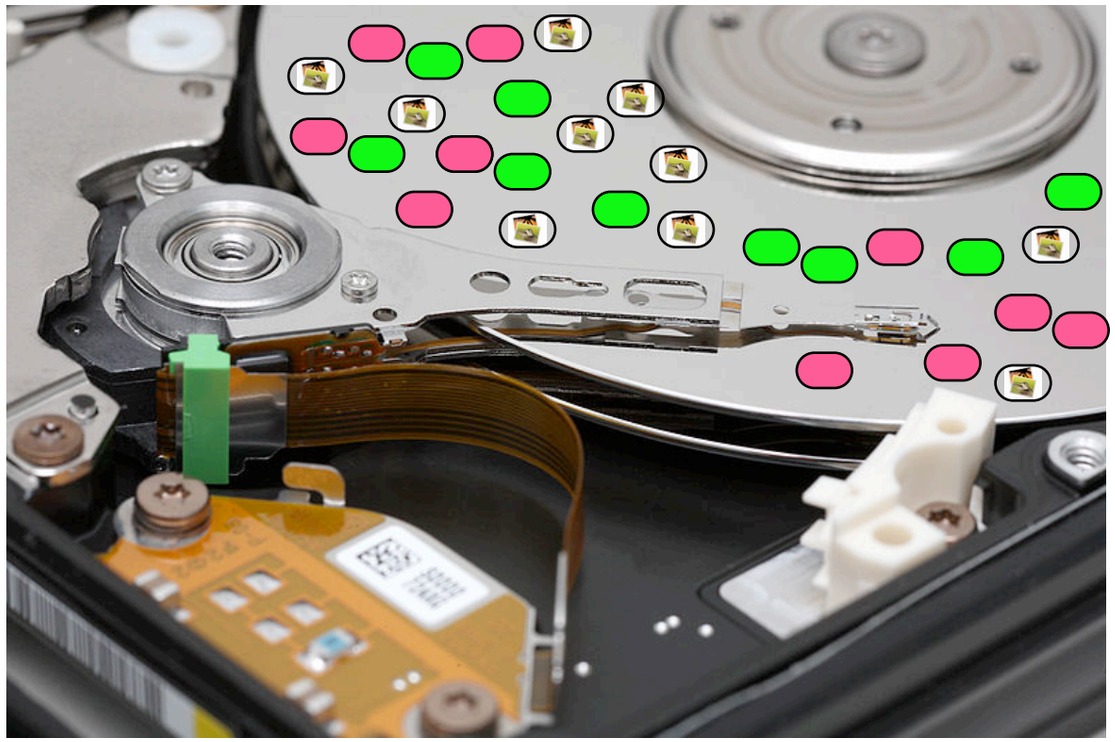
Can we tell if a drive was properly wiped in 2 minutes?

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:



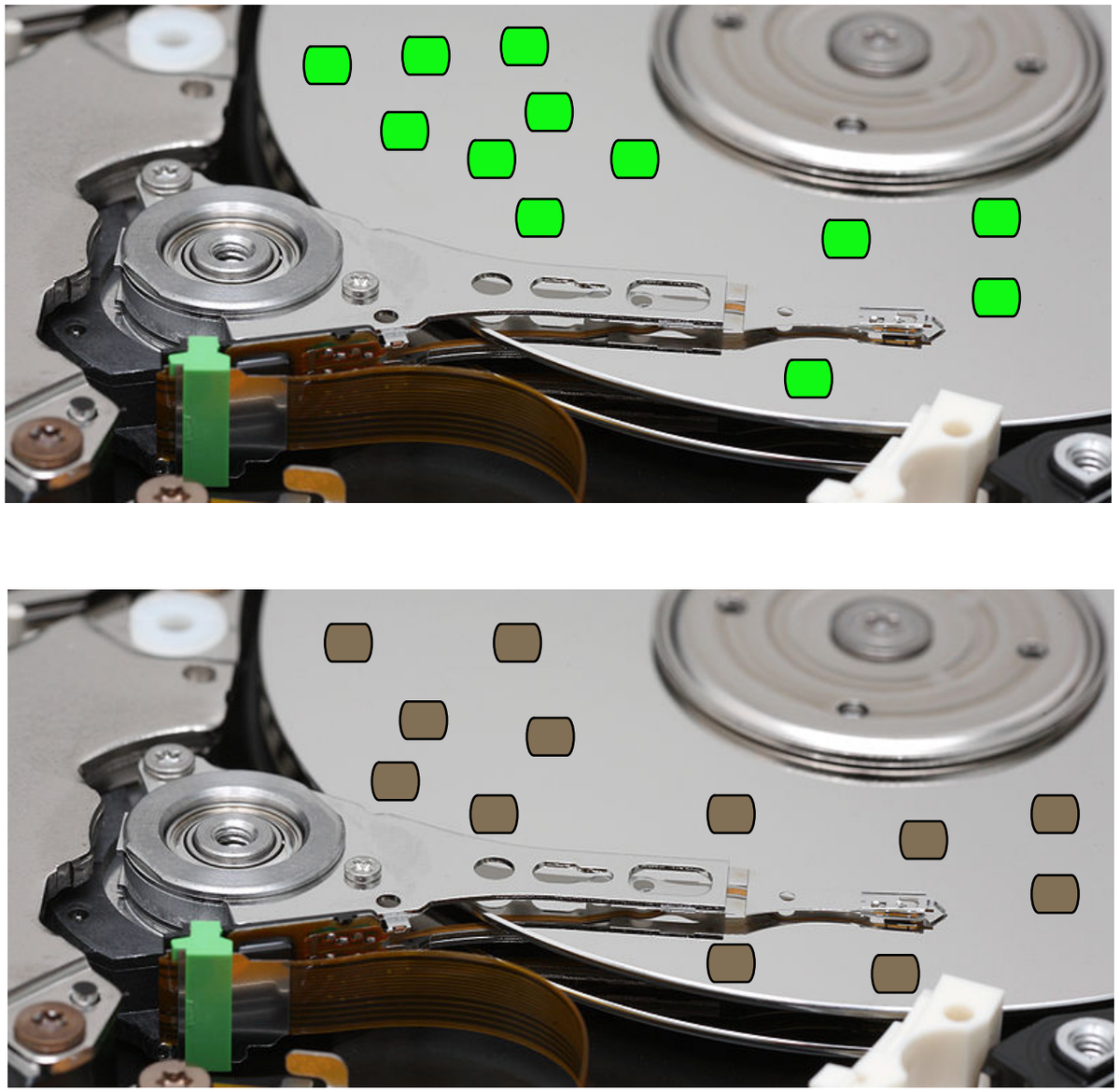
Hard drive contents can be predicted by sampling a few thousand sectors:



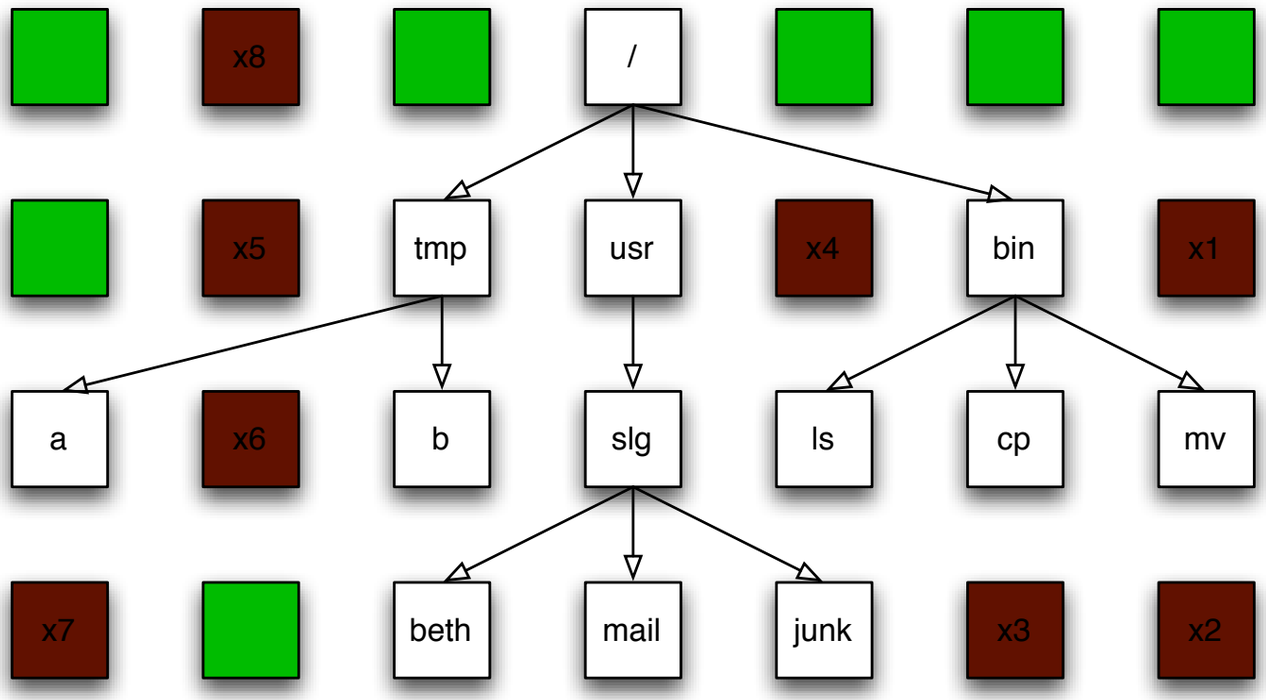
The challenge is identifying *likely voters.*

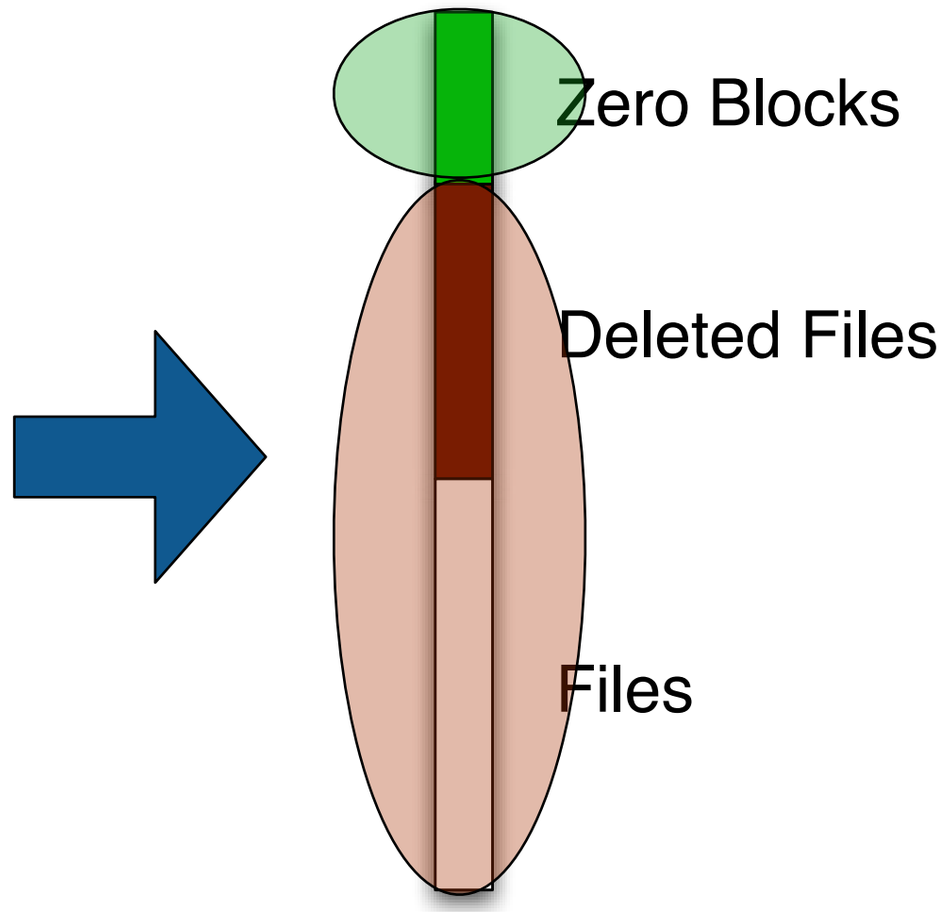
The challenge is *identifying the content* of the sampled sectors.

For example, we can use random sampling to determine if a hard drive has been properly wiped.



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



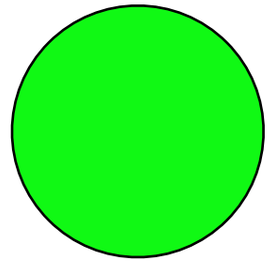


But that's fine if we are just trying to tell if a drive was properly wiped.

How many sectors do we need to read? How about 10,000?

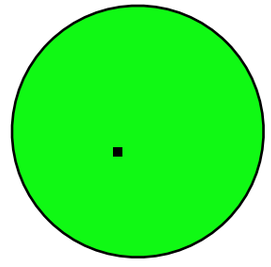
If the disk has 2,000,000,000 blank sectors (0 with data)

- The sample is identical to the population



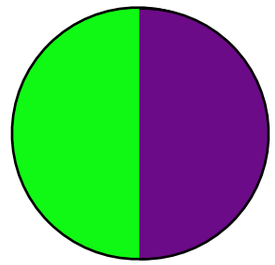
If the disk has 1,999,999,999 blank sectors (1 with data)

- The sample is representative of the population.
- We will only find that 1 sector using exhaustive search.



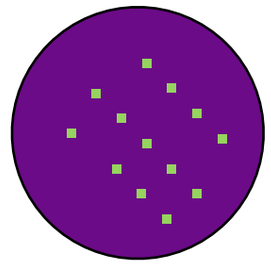
If the disk has 1,000,000,000 blank sectors (1,000,000,000 with data)

- Something about our sampling matched the allocation pattern.
- *This is why we use random sampling.*



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

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



We can't tell if a disk is *completely* blank with sampling.

But we *can* determine the probability that the disk has less than 10MB of actual data.

- Sectors on disk: 2,000,000,000 (1TB)
- Sectors with data: 20,000 (10 MB)

Chose one sector. Odds of missing the data:

- (2,000,000,000 - 20,000) / (2,000,000,000) = 0.99999
- You are *very likely* to miss one of 20,000 sectors if you pick just one.

Chose a second sector. Odds of missing the data on both tries:

- 0.99999 * (1,999,999,999-20,000) / (1,999,999,999) = .99998
- You are still *very likely* to miss one of 20,000 sectors if you pick two.

But what if you pick 1000? Or 10,000? Or 100,000?

The more sectors picked, the less likely you 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
2	0.99998
100	0.99900
1000	0.99005
10,000	0.90484
20,000	0.81873
40,000	0.67032
60,000	0.54881
80,000	0.44932
100,000	0.36787
150,000	0.22312
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 for a given number of randomly sampled sectors. Smaller probabilities indicate higher accuracy.

500,000 blank randomly chosen sectors should be good enough!

Fragment classification:
Different file types require different strategies.

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>
```

JPEG files can be identified through the "FF" escape.

- FF must be coded as FF00.
- So if there are a lot of FF00s and few FF01 through FFFF it must be a JPEG.

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.

We developed *file fragment type identifiers.*
We identify the *content* of a 160GB iPod in 118 seconds.

Identifiable:

- Blank sectors
- JPEGs
- Encrypted data
- HTML

Report:

- Audio Data Reported by iTunes: 2.42GB
- MP3 files reported by file system: 2.39GB
- Estimated MP3 usage:
  - *2.71GB (1.70%) with 5,000 random samples*
  - *2.49GB (1.56%) with 10,000 random samples*

Sampling took 118 seconds.

Work to date:

Publications:

- 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
- Roussev, Vassil, and Garfinkel, Simson, File Classification Fragment---The Case for Specialized Approaches, Systematic Approaches to Digital Forensics Engineering (IEEE/ SADFE 2009), Oakland, California.
- Farrell, P., Garfinkel, S., White, D. Practical Applications of Bloom filters to the NIST RDS and hard drive triage, Annual Computer Security Applications Conference 2008, Anaheim, California, December 2008.

Tools:

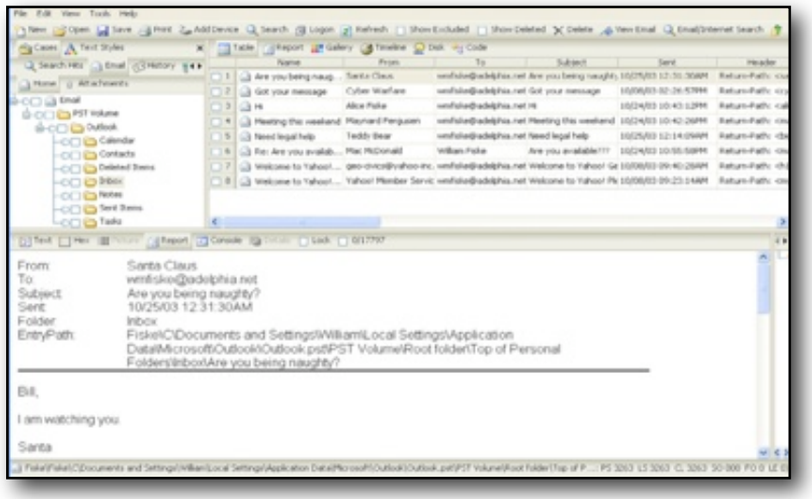
- Using “Hamming,” our 1100-core cluster for novel SD algorithms.
- Roussev’s Similarity Metric

Team:

- Alex Nelson (PhD Candidate, UCSC) summer project
- LT Ryan Mayer

Standardized 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.

Digital Forensics is at a turning point. Today's work is increasingly *scientific.*

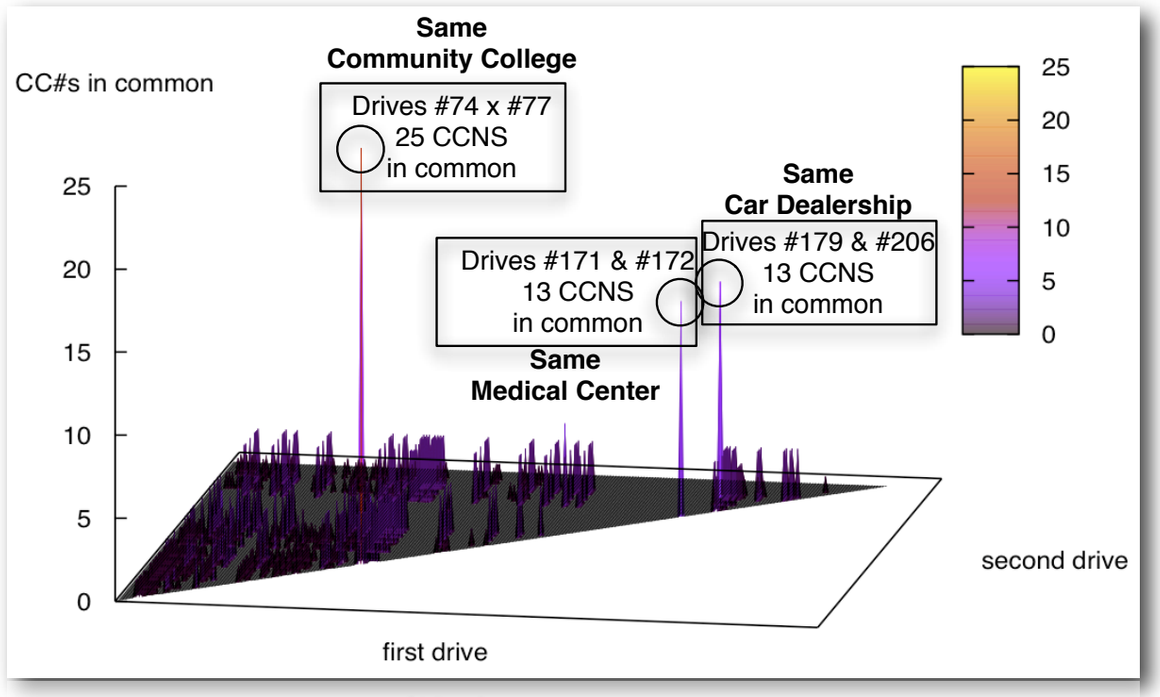
Evidence Reconstruction

- 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.

We can't do this today.

- Bob's tool can identify 70% of the data in the windows registry.
  - —*He publishes a paper.*
- Alice writes her own tool and can only identify 60%.
  - —*She writes Bob and asks for his data.*
  - —*Bob can't share the data because of copyright & privacy issues.*

To address this problem, we are creating releasable corpora.

NPS-govdocs1: 1 Million files available *now*

1 million documents downloaded from US Government web servers

- Specifically for file identification, data & metadata extraction.
- Found by random word searches on Google & Yahoo
- DOC, DOCX, HTML, ASCII, SWF, etc.

Free to use; Free to redistribute

- No copyright issues — US Government work is not copyrightable.
- Other files have simply been moved from one USG webserver to another.
- No PII issues — These files were already released.

Distribution format: ZIP files

- 1000 ZIP files with 1000 files each.
- 10 “threads” of 1000 randomly chosen files for student projects.
- Full provenance for every file (how found; when downloaded; SHA1; etc.)

http://domex.nps.edu/corp/files/

We have also created dozens of disk images, packet captures, and memory dumps.

Test Images:

- nps-2009-hfstest1 (HFS+)
- nps-2009-ntfs1 (NTFS)

Realistic Images:

- nps-2009-canon2 (FAT32)
- nps-2009-UBNIST1 (FAT32)
- nps-2009-casper-rw (embedded EXT3)
- nps-2009-domexusers (NTFS)

Scenarios:

- M57 startup — spear phishing attack
- M57 patents — small business victim of internal hacking
- Nitroba University — Harassment case solved through network forensics

http://digitalcorpora.org/

The Real Data Corpus: "Real Data from Real People."

Most forensic work is based on “realistic” data created in a lab.

We get real data from CN, IN, IL, MX, and other countries.

Real data provides:

- Real-world experience with data management problems.
- Unpredictable OS, software, & content
- Unanticipated faults

We have multiple corpora:

- Non-US Persons Corpus
- US Persons Corpus (@Harvard)
- Releasable Real Corpus
- Realistic Corpus

Real Data Corpus: Current Status

Corpus	HDs	Flash	CDs	GB
US*	1258			2939
BA	7			38
CA	46	1		420
CN	26	568	98	999
DE	37	1		765
GR	10			6
IL	152	4		964
IN		66		29
MX	156			571
NZ	1			4
TH	1	3		13
* Not available to USG	1694	643	98	6748

Note: IRB Approval is Mandatory!

Work to date:

Publications:

- Garfinkel, Farrell, Roussev and Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, Best Paper, DFRWS 2009

Websites:

- http://digitalcorpora.org/
- http://domex.nps.edu/corp/files/

Team:

- Joshua Gross, NPS postdoc, 2009-2011

In summary:
Automated Digital Forensics and Media Exploitation

Many research opportunities

- Applying data mining algorithms to new domains with new challenges.
- Working with large, heterogeneous data set.
- Text extraction & clustering
- Multi-lingual
- Cryptography & Data recovery

Interesting legal issues.

- Data acquisition.
- Privacy
- IRB (Institutional Review Boards.)

Lots of low hanging fruit.

Questions?

Many research opportunities

- Applying data mining algorithms to new domains with new challenges.
- Working with large, heterogeneous data set.
- Text extraction & clustering
- Multi-lingual
- Cryptography & Data recovery

Interesting legal issues.

- Data acquisition.
- Privacy
- IRB (Institutional Review Boards.)

Lots of low hanging fruit.

Sample Questions:

- Do flash storage devices and Solid State Drives (SSDs) create new opportunities?
- What are the opportunities for face recognition and other content analysis techniques?
- What can you "correlate" other than email addresses?
- Can overwritten data be recovered?