
# Automated Digital Forensics and Media Exploitation

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

November 11, 2009

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
- Operational & Information Sciences
- International Graduate Studies

# A bit about me

Tech Journalist: 1985—2002

Entrepreneur: 1988—2002

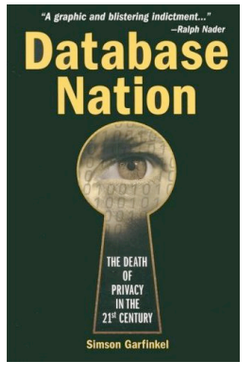
Vineyard.NET, Broadband2Wireless,

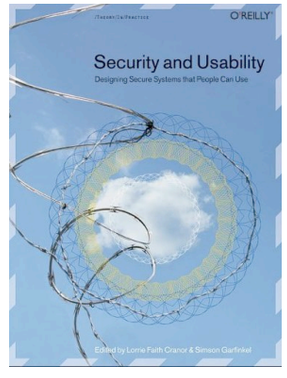
Sandstorm Enterprises, Inc.

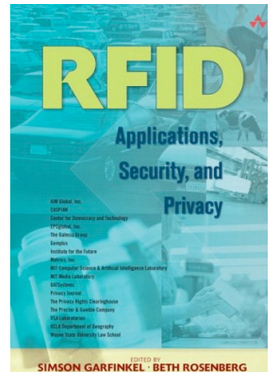
MIT EECS 2002—2005 (PhD CS)

Fellow, 2005—2008
Center for Research on Computation and Society,
School of Engineering and Applied Sciences,
Harvard University

Associate Professor, 2006—
Naval Postgraduate School,

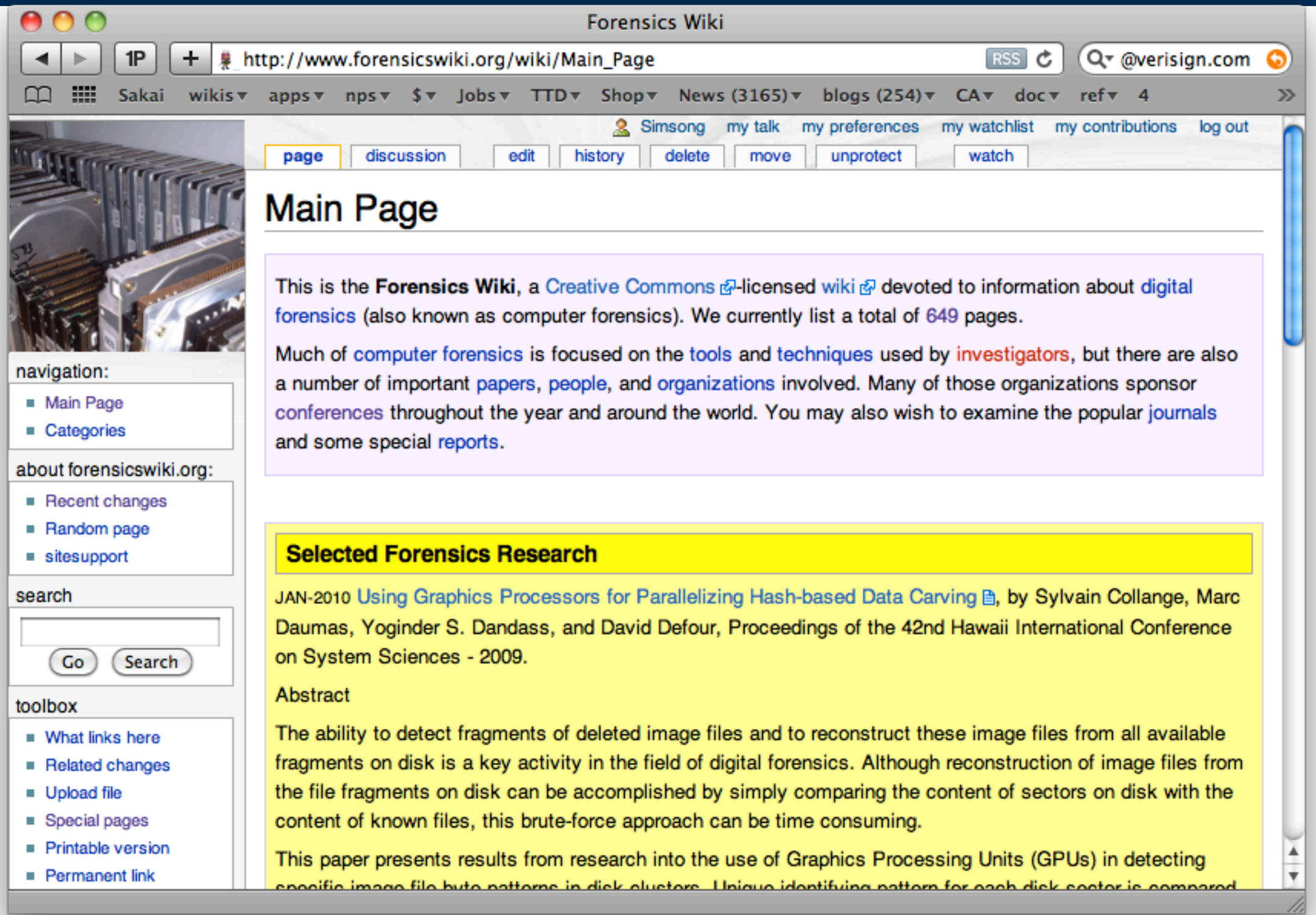


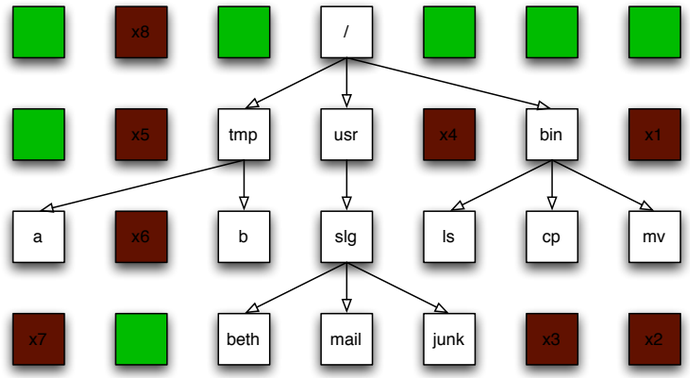

**"The views expressed in this presentation are those of the author and do not necessarily reflect those of the Department of Defense or the US Government."**

# I also run the Forensics Wiki

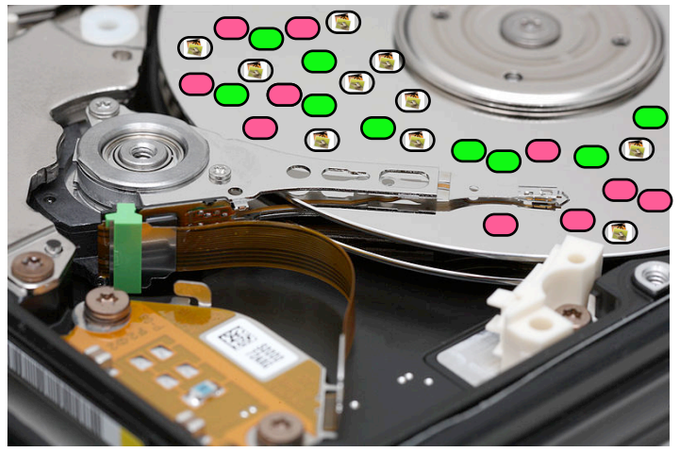


# This talk introduces digital forensics and gives a taste of my research.

Why is "residual data" left behind?




How to analyze a 1TB drive in 2 minutes.



Putting the "science" into Digital Forensics with standardized forensic corpora.


# Residual Data and Data Privacy

# August 1998:
# My first encounter with other people's data.

I purchased 10 used computers from a computer store…

… for a project

This computer had been the file server of a law firm.

Other computers contained:

- Database of mental health patients
- Files from a divorced woman worrying about child support & college expenses.
- Draft manuscript of a prominent novelist…

# There have been many cases of sensitive data left on computers that were sold.

## April 1997

- A woman in Pahrump, NV, purchases a used IBM PC and discovers records from 2000 patients who had prescriptions filled at Smitty's Supermarkets pharmacy in Tempe, AZ.

## August 2001

- More than 100 computers from Viant with confidential client data sold at auction.

## Spring 2002

- Pennsylvania state Department of Labor and Industry sells computers with "thousands of files of information about state employees."

## August 2002

- Purdue student purchased used Macintosh computer at equipment exchange; computer contains FileMaker database with names and demographic information of 100 applicants to Entomology Department.

…

## May 2009

- University of Glamorgan (UK) purchases hard drives with US missile system information.

http://www.forensicswiki.org/wiki/Residual_Data_on_Used_Equipment

# There are dozens of stories of used data release. There are *millions* of systems retired every year.

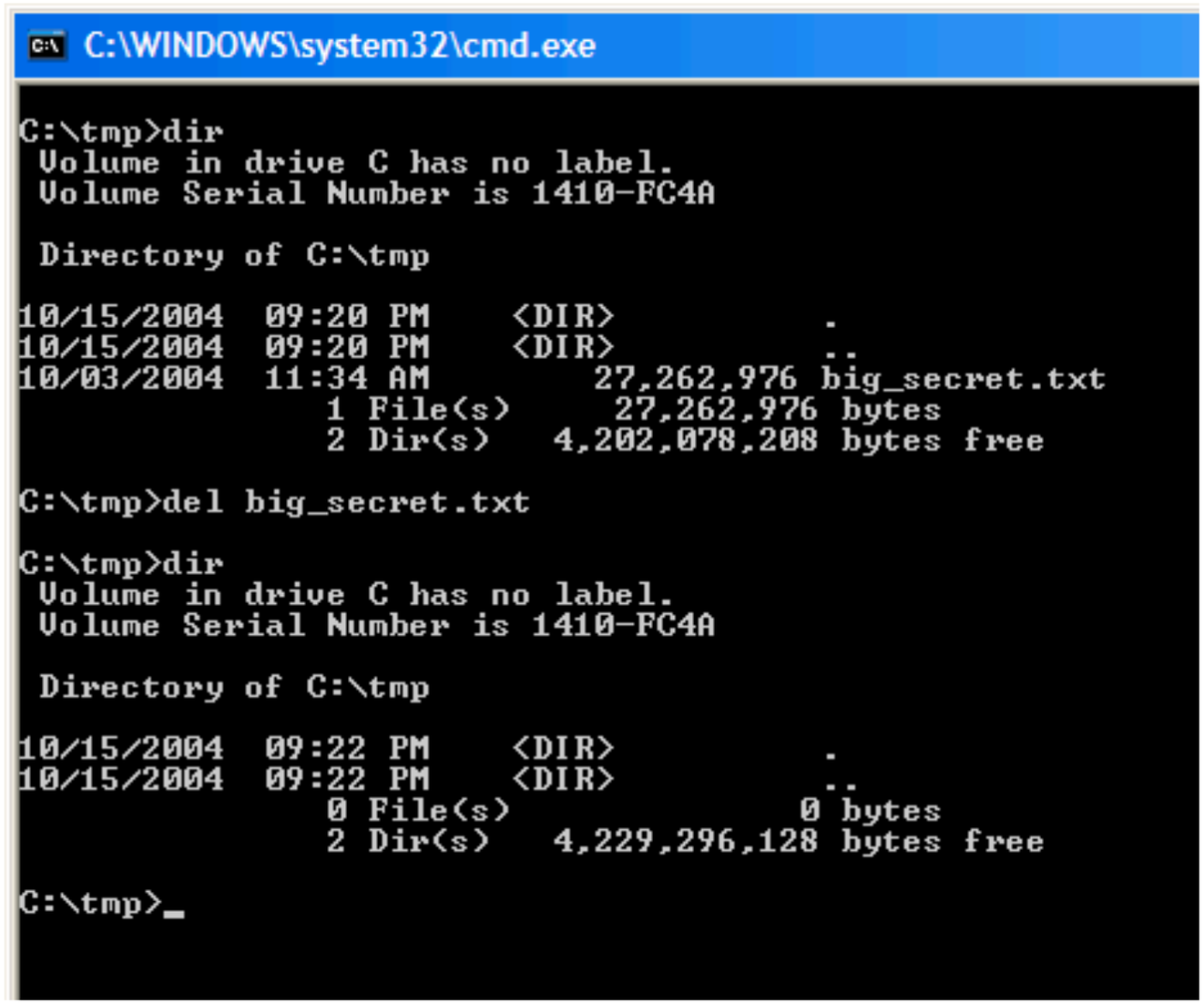
Why are there so few cases?

- Hypothesis #1: Disclosure of “residual data” is rare because most systems are properly sanitized.
- Hypothesis #2: Disclosures are so common that they are not newsworthy.
- Hypothesis #3: Systems aren’t properly sanitized, but few notice the data.

# How could people not notice the data?

DEL removes the file's name…

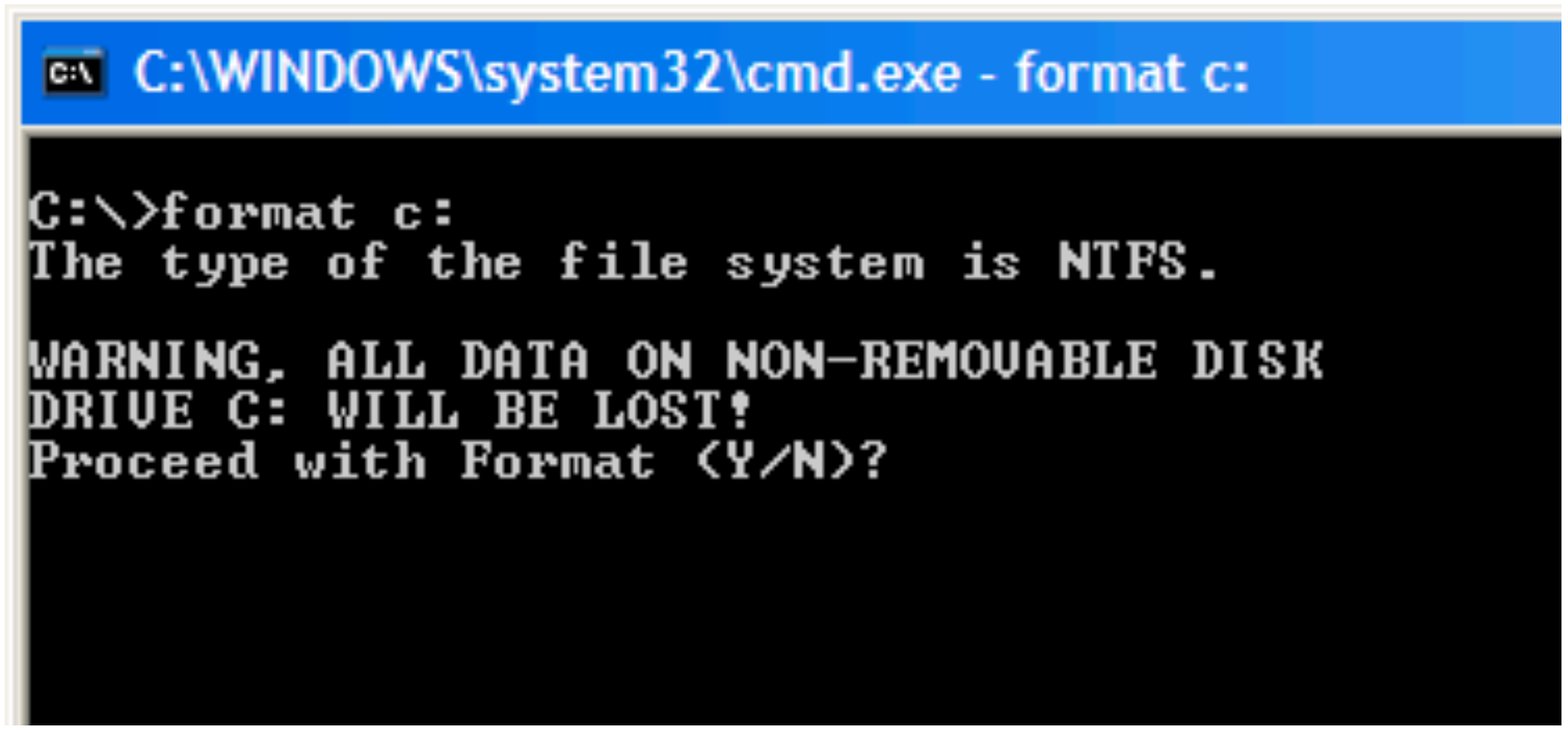
… but doesn't delete the file's data



# How could people not notice the data?

FORMAT C: writes a new root directory…

… but until Windows Vista, it did not overwrite the disk!




Most people don't have the tools to recover the data.

# Mass storage devices pose a special problem for computer security

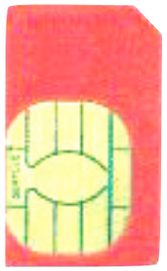
## Mass storage devices:

- Do not forget data when power is removed.
- Can contain data that is not immediately visible.

## Today's computers can read hard drives that are 20 years old!

- Electrically compatible (IDE/ATA)
- Logically compatible (FAT16/32 file systems)
- USB devices were designed to be "Universal."

## Strong social bias against destroying working equipment.

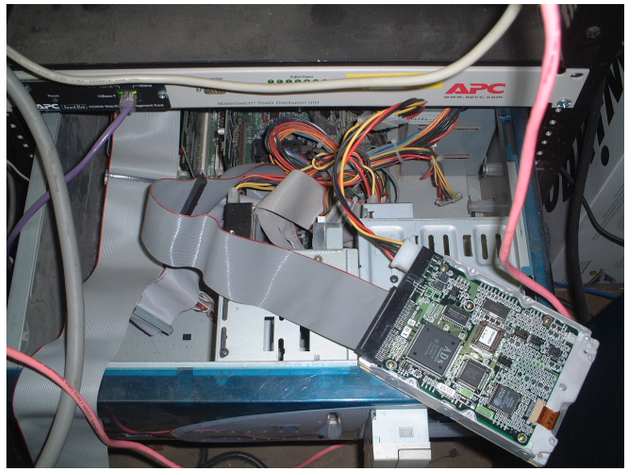
# The “Remembrance of Data Passed” Study (1998-2003)

235 used hard drives between November 2000 and January 2003

- eBay
- Computer stores
- Swap fests
- No more than 20 from the same vendor


**1. Drives Arrive**

**2. Accession**




**3. Imaging**


**4. Archiving**

# Disk #70: IBM-DALA-3540/81B70E32

Purchased for $5 from a Mass retail store on eBay

Copied the data off: 541MB

Initial analysis:

- 1,057,392 disk blocks
- 67,878 blocks are all NULs (6%)

```
$ mount /dev/sdb /mnt
$ ls /mnt/
/mnt/IO.SYS
/mnt/MSDOS.SYS
/mnt/COMMAND.COM
$
```

# Disk #70: IBM-DALA-3540/81B70E32

```
% strings 0070.raw | more
…
[.??
!ZY[
0123456789ABCDEFS
WOWOW090
WOWO
6,.h
Insert diskette for drive
 and press any key when ready
Your program caused a divide overflow error.
If the problem persists, contact your program vendor.
Windows has disabled direct disk access to protect your long filenames.
To override this protection, see the LOCK /? command for more information.
The system has been halted.  Press Ctrl+Alt+Del to restart your computer.
You started your computer with a version of MS-DOS incompatible with this
version of Windows. Insert a Startup diskette matching this version of

OEMString = "NCR 14 inch Analog Color Display Enchanced SVGA, NCR
Corporation"
        Graphics Mode: 640 x 480 at 72Hz vertical refresh.
        XResolution                = 640
        YResolution                = 480
```

# 70.img con’t

```
Wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwq
ling the Trial Edition
--------------------------
IBM AntiVirus Trial Edition is a full-function but time-limited
evaluation version of the IBM AntiVirus Desktop Edition product.  You
may have received the Trial Edition on a promotional CD-ROM or as a
single-file installation program over a network.  The Trial Edition
is available in seven national languages, and each language is
provided on a separate CC-ROM or as a separa
EAS.STCm
EET.STC
ELR.STCq
ELS.STC
…
MAB-DEDUCTIBLE
MAB-MOOP
MAB-MOOP-DED
METHIMAZOLE
INSULIN (HUMAN)
COUMARIN ANTICOAGULANTS
CARBAMATE DERIVATIVES
AMANTADINE
MANNITOL
MAPROTILINE
CARBAMAZEPINE
CHLORPHENESIN CARBAMATE
```

# “Automated Forensics:” Automatically find the good stuff

Automatic searching for credit-card numbers

- 4725 3321 3342 1134

Most common email address

Searching for medical terms

Combined timeline of all disks

# Sectors on hard drives can be divided into three categories:

Allocated Data

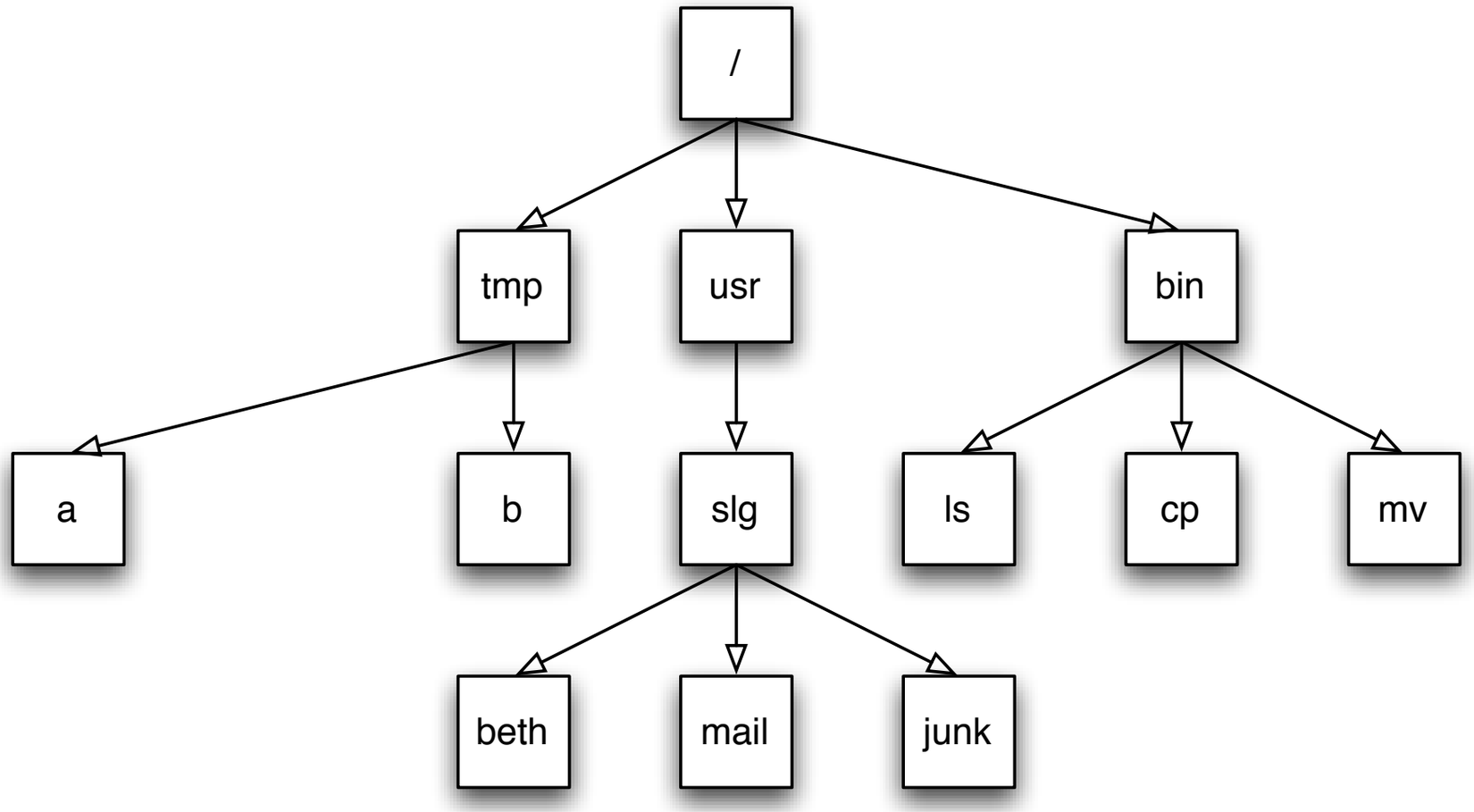
Deleted Data

} user files
email messages
[temporary files]

Uninteresting Data

blank sectors [OS files]

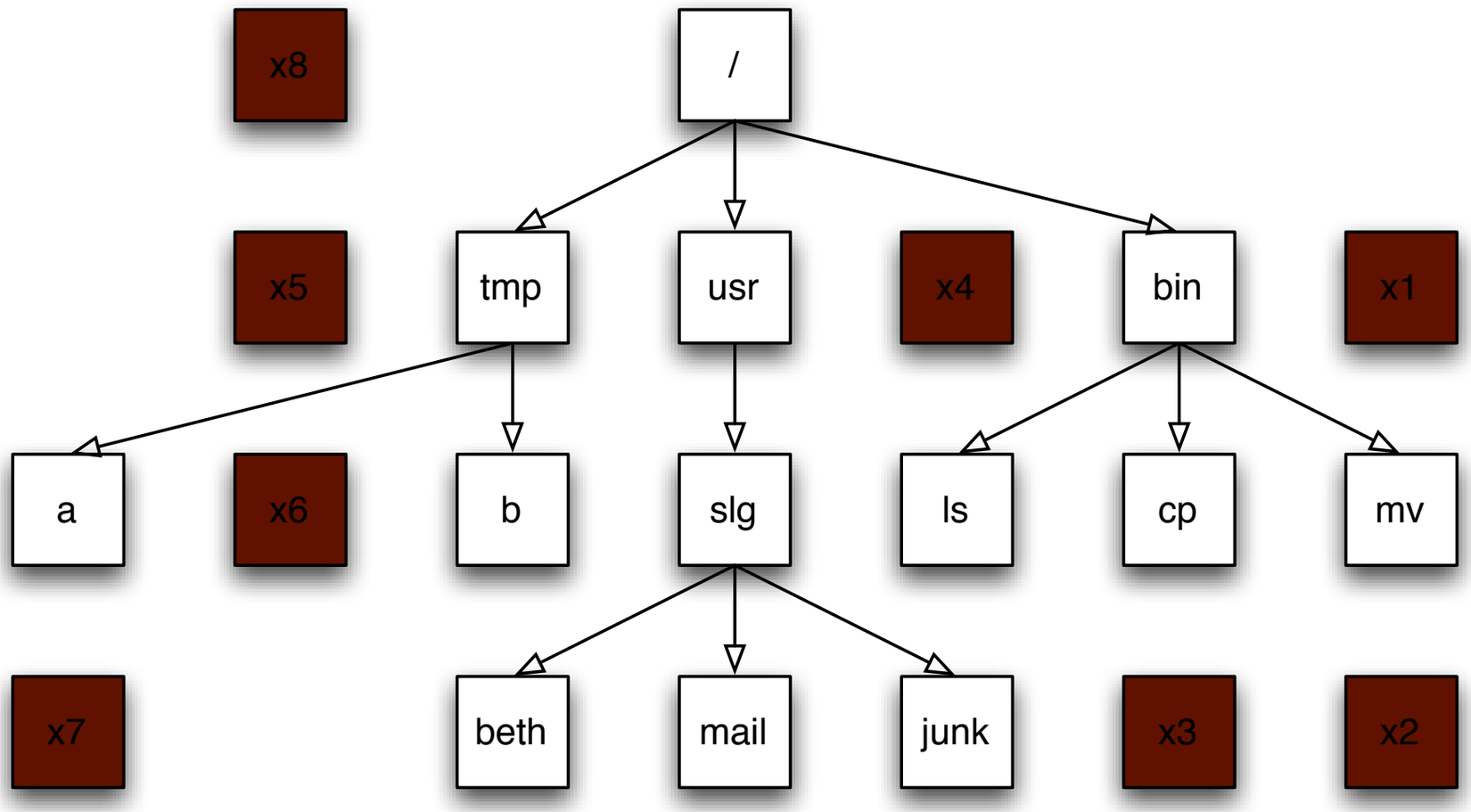
# Data on a hard drive is arranged in sectors




Allocated Data

**= data visible to the user**

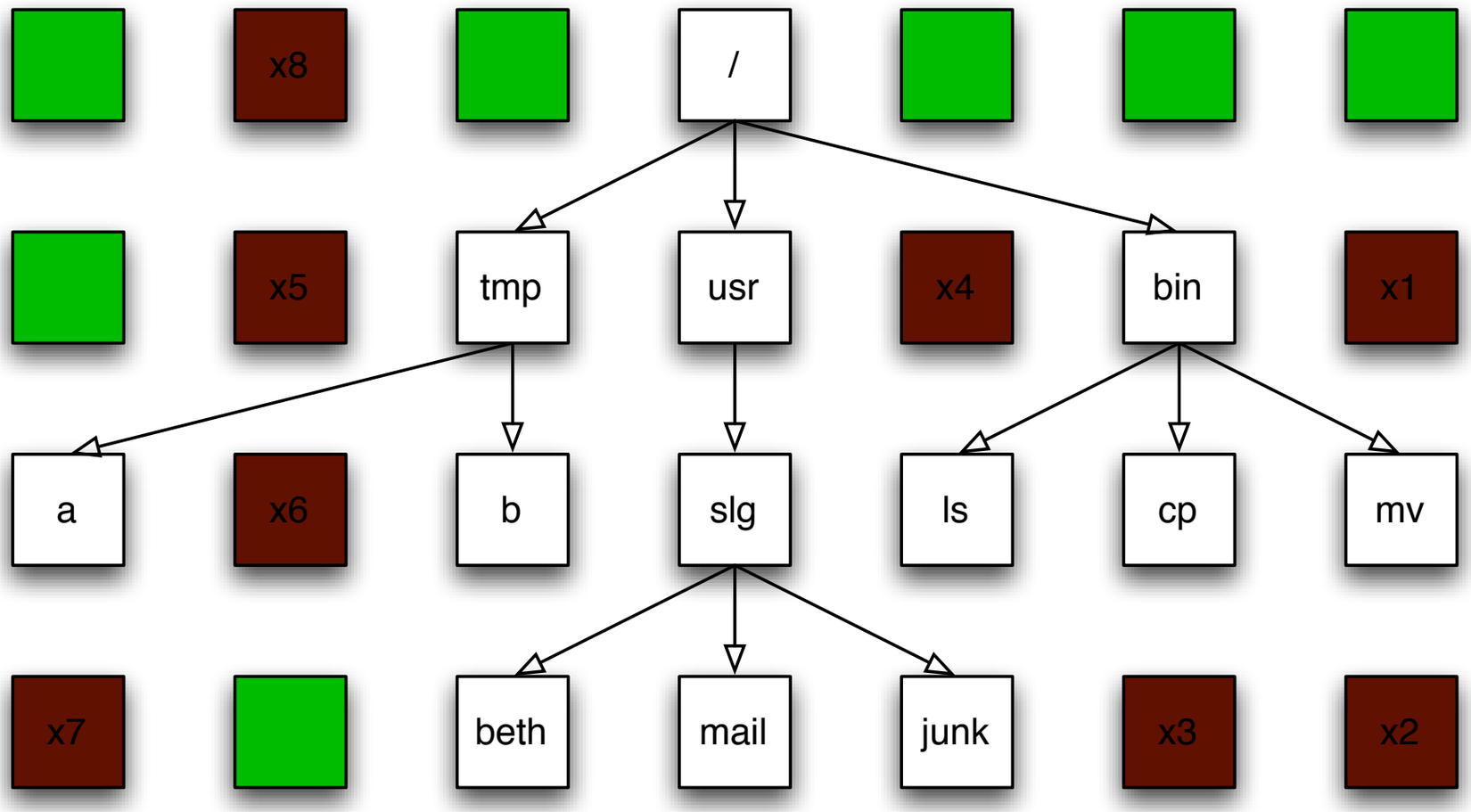
# Data on a hard drive is arranged in sectors




Deleted Data

**= files that were deleted.**

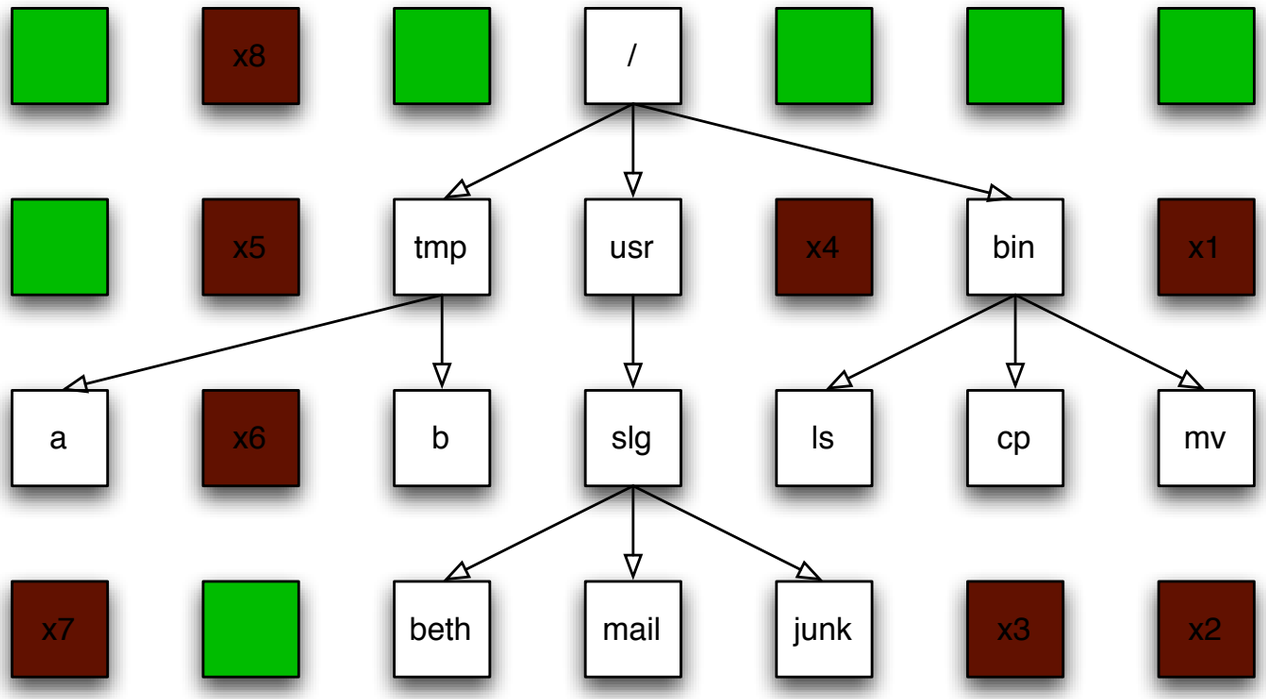
# Data on a hard drive is arranged in sectors

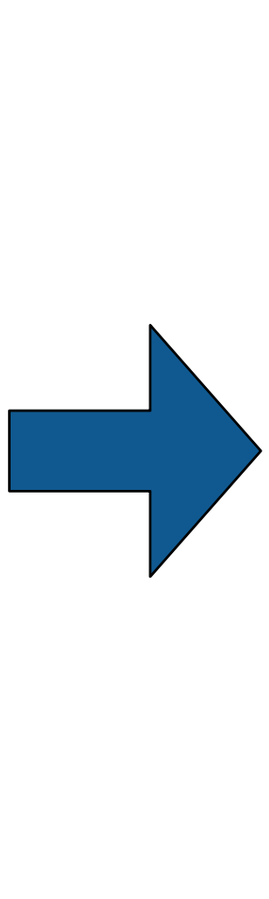


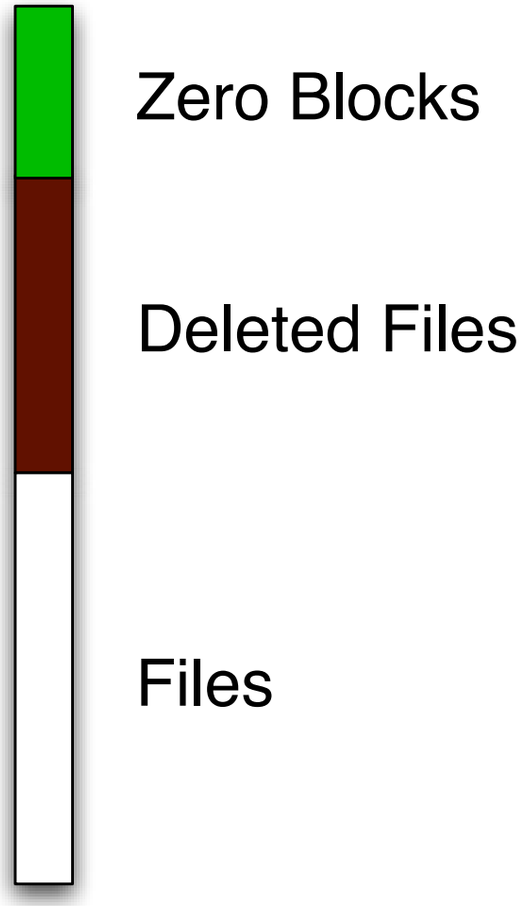

Uninteresting Data

**= never written (or wiped clean)**

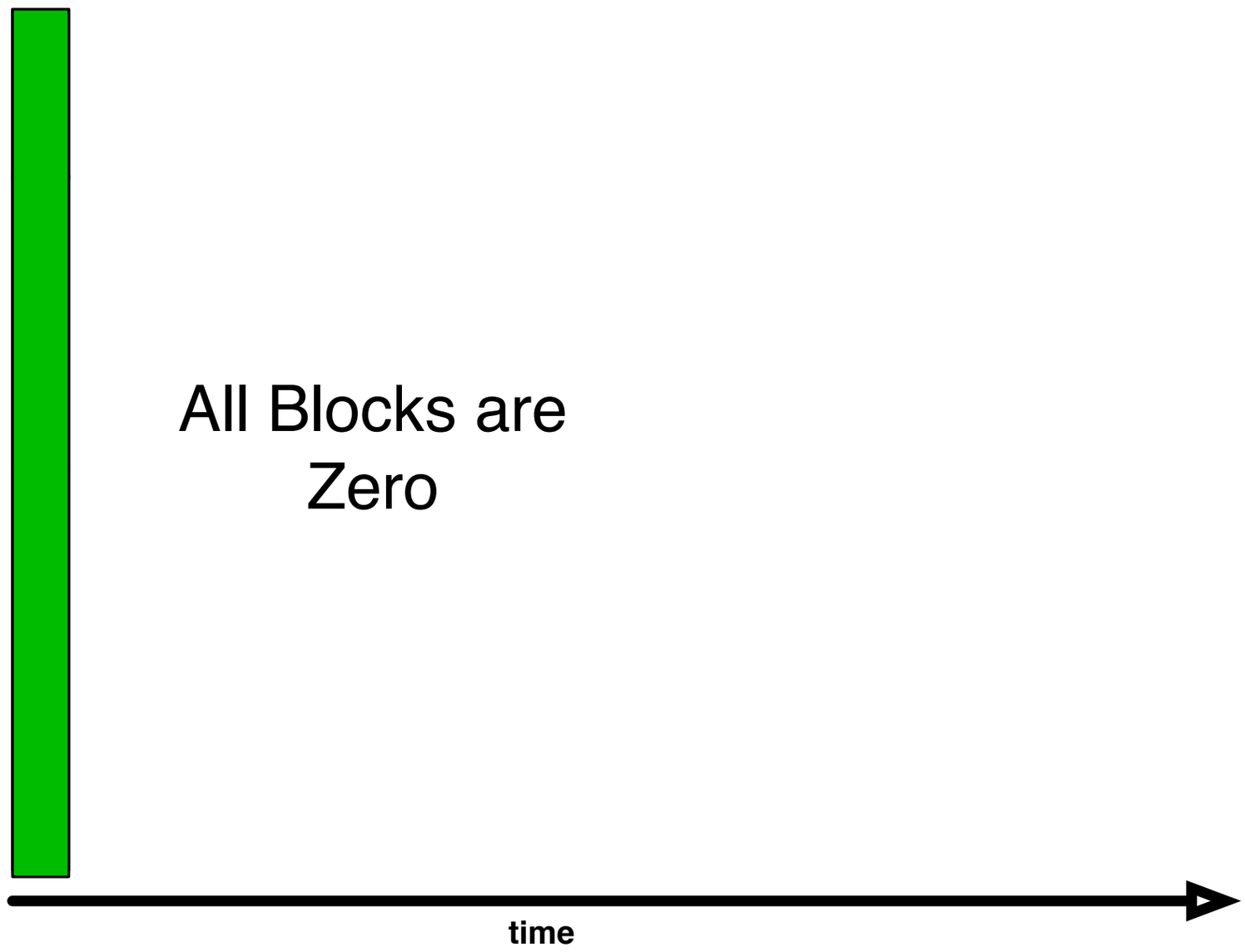
# Stack the sectors:



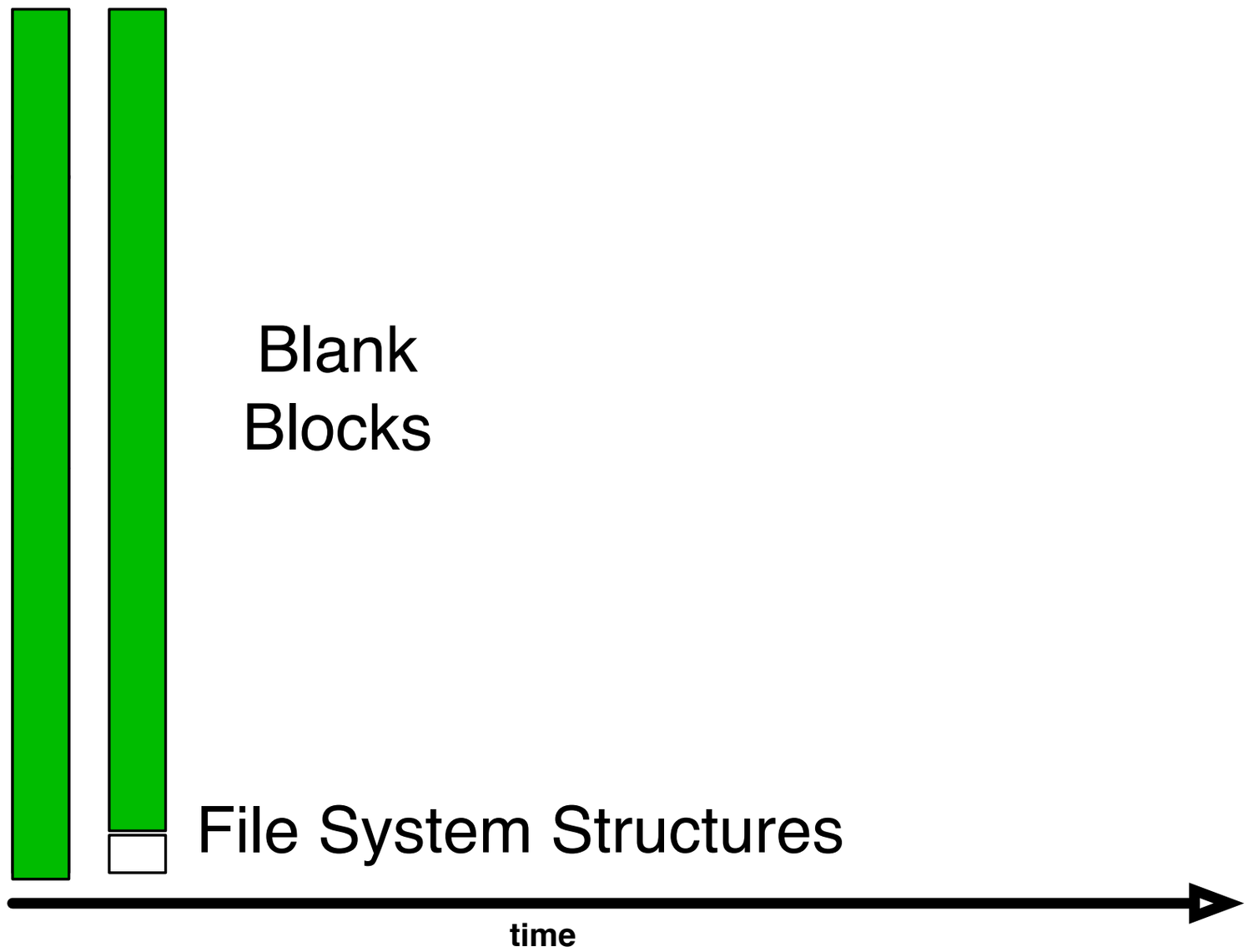





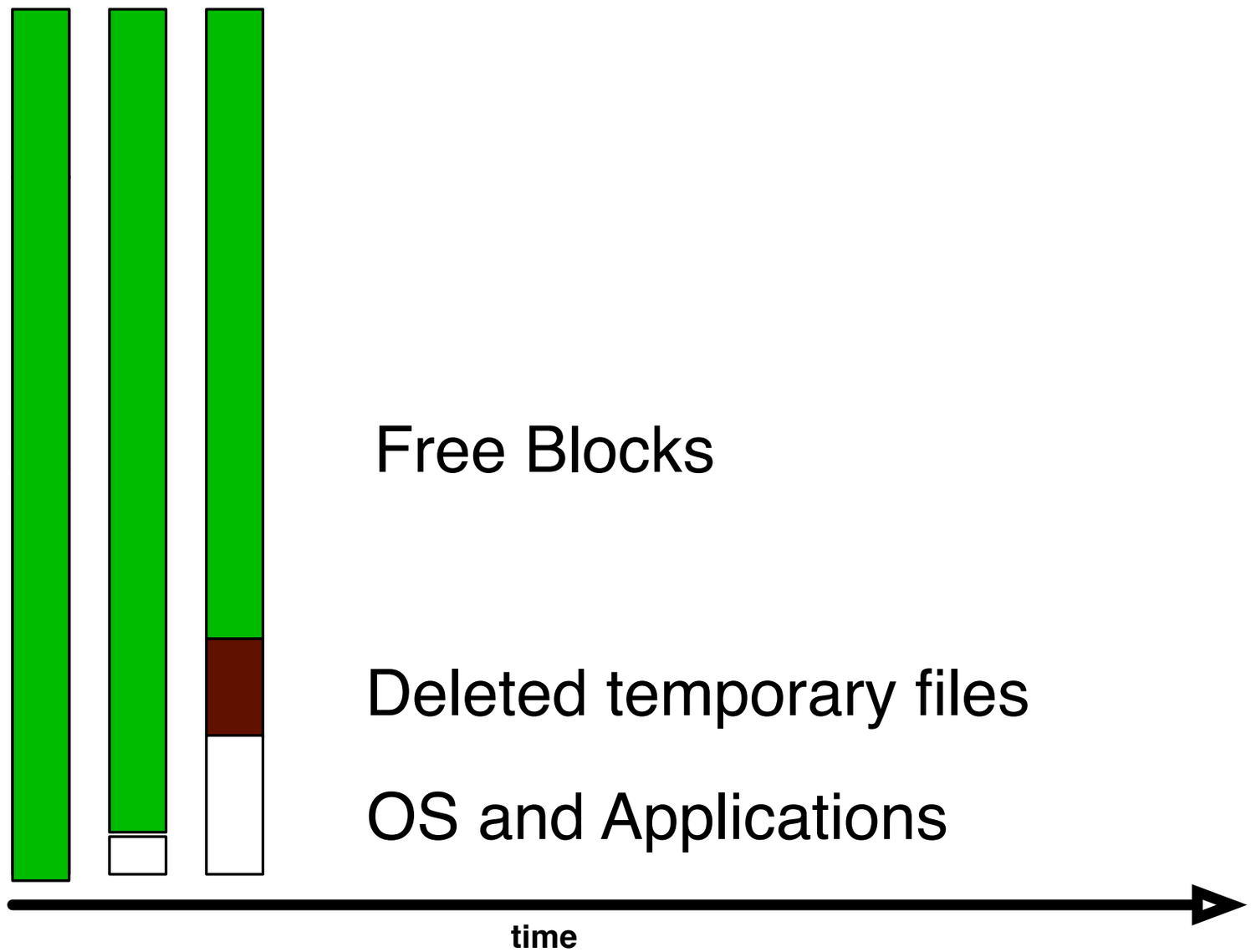
# No data: The disk is factory fresh



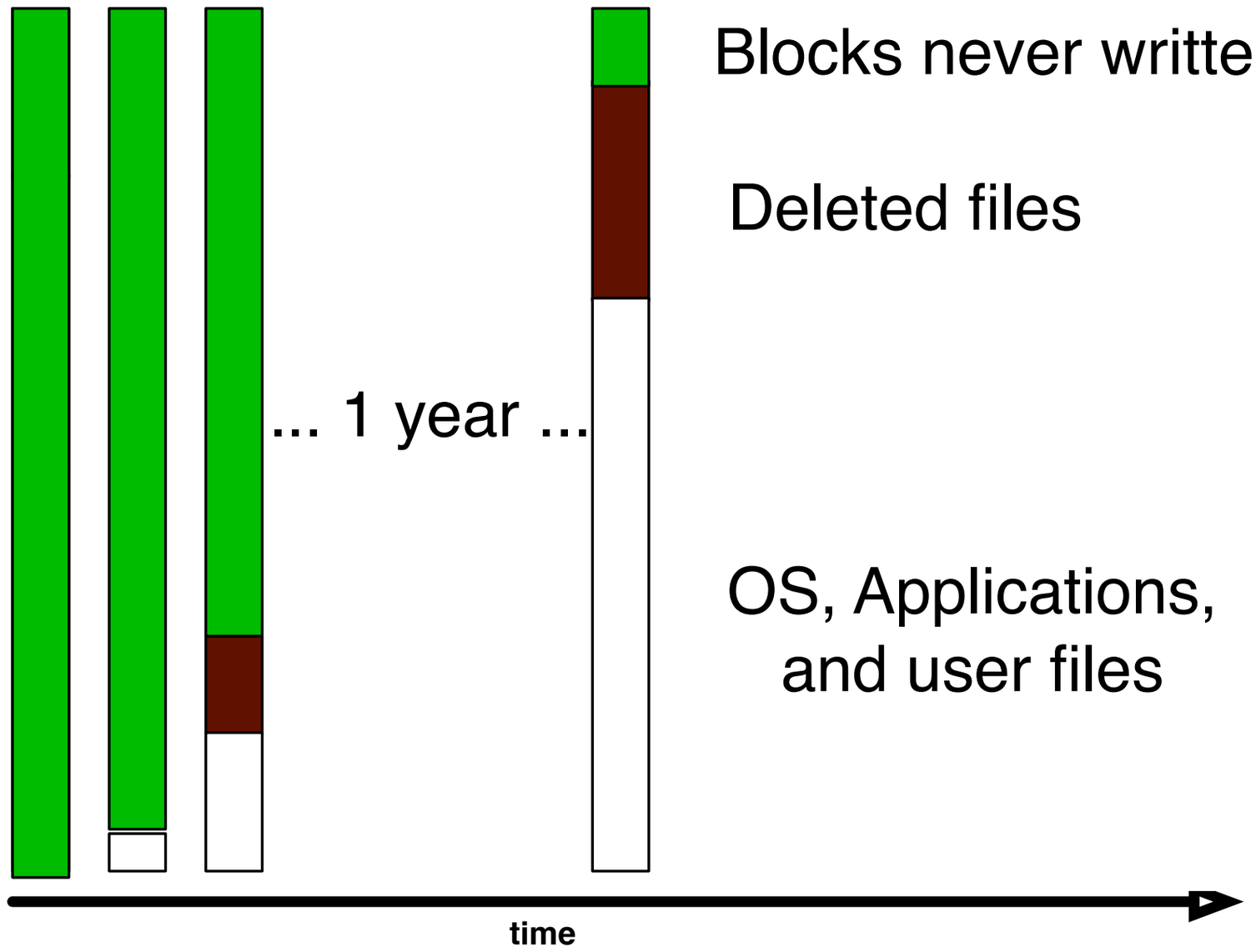
# Formatted: the disk has an empty file system



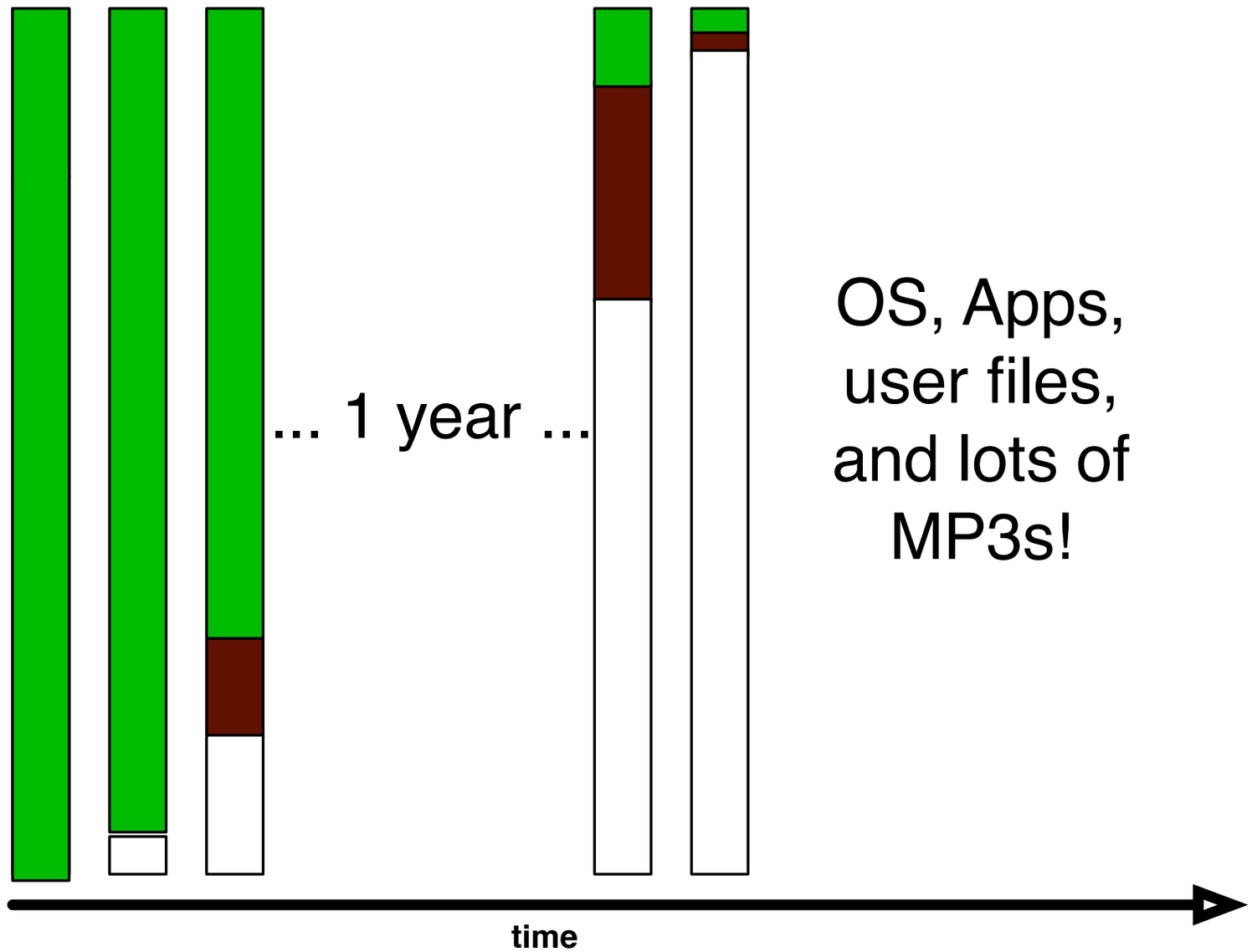
# AFTER OS INSTALL: Temp. files have been deleted



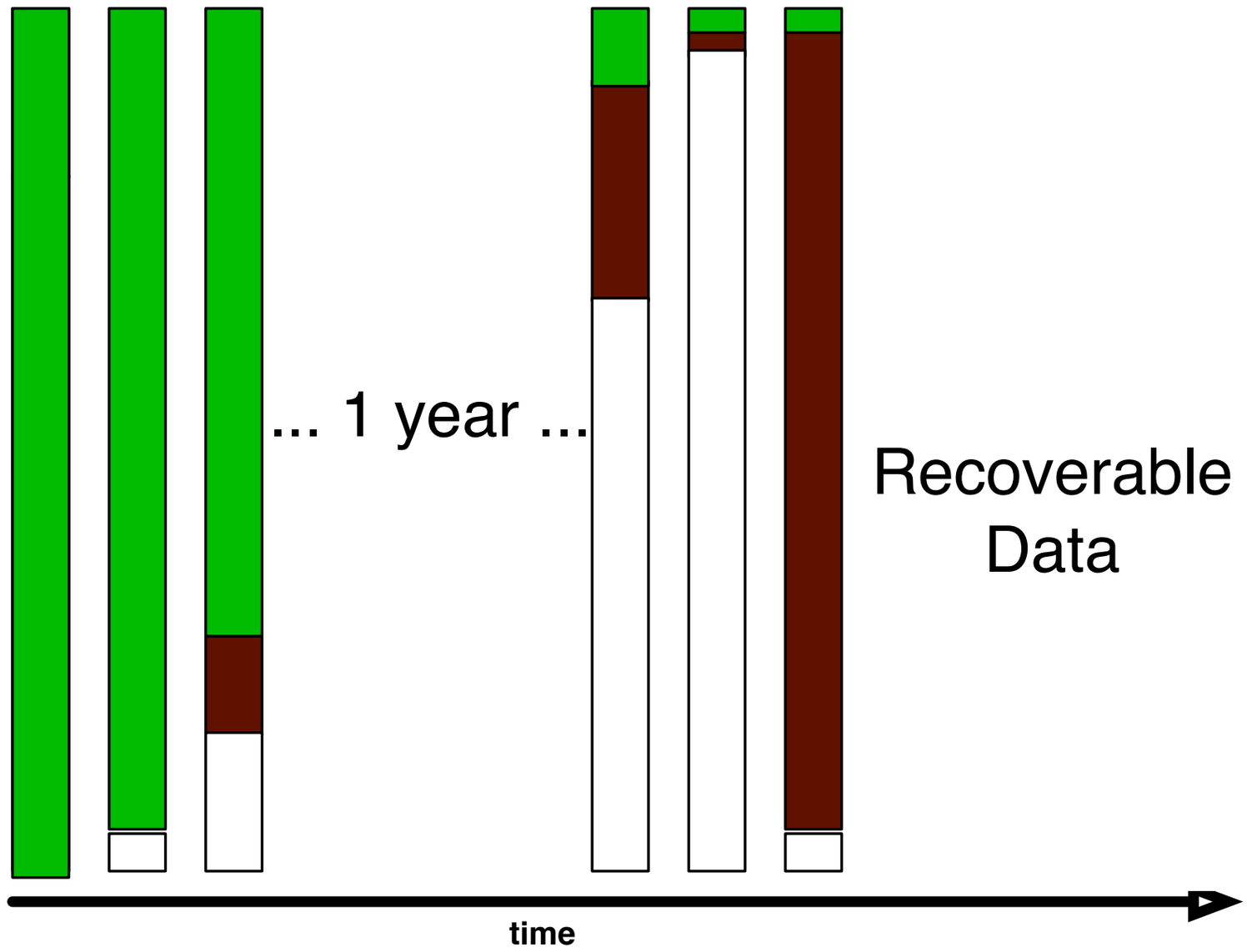
# After a year of service



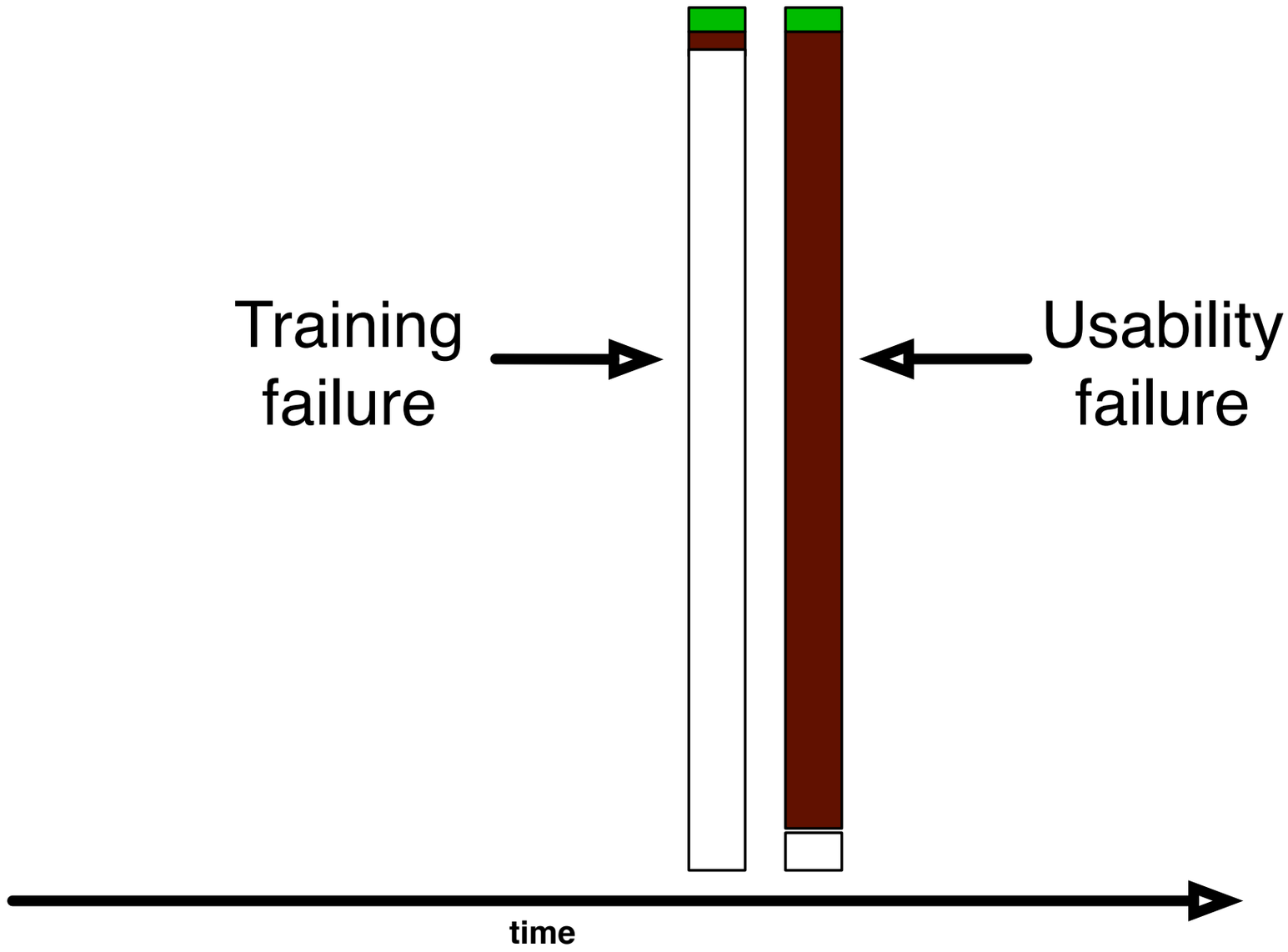
# Disk nearly full!



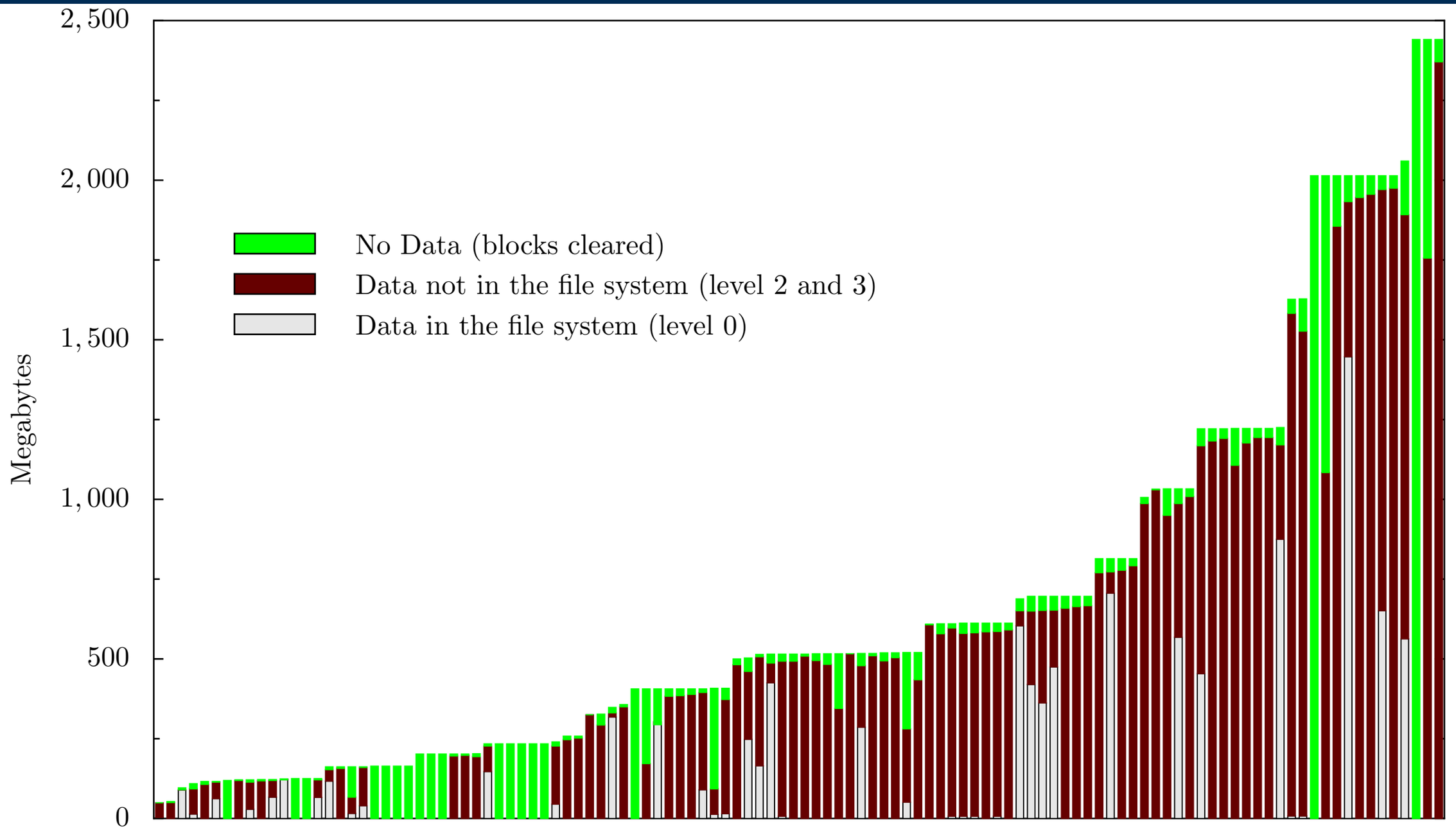
# Let’s sell the hard drive!
# Format c:\



# We can use forensics to reconstruct motivations:



# We analyzed 236 drives purchased 1998—2003.




Roughly ⅓ of the drives had sensitive information.

# Main Sources of Failure:

## Failing or Defunct Companies

- Nobody charged with data destruction

## Trade-ins and PC upgrades

- Owner assumed that service provider would sanitize

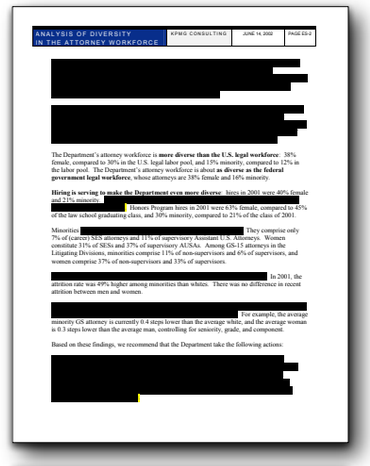
## Failure to supervise contract employees

- Sanitization was never verified

# Residual Data is also left behind *inside* document files.

Adobe PDF files — most violations from *covered data.*

- The New York Times published a PDF file containing the names of Iranians who helped with the 1953 coup. (2000) (http://cryptome.org/cia-iran.htm)
- US DoJ published a PDF file "diversity report" with embarrassing redacted information. (2003) (http://www.thememoryhole.org/feds/doj-attorney-diversity.htm)
- Multinational Force-Iraq report (2005)

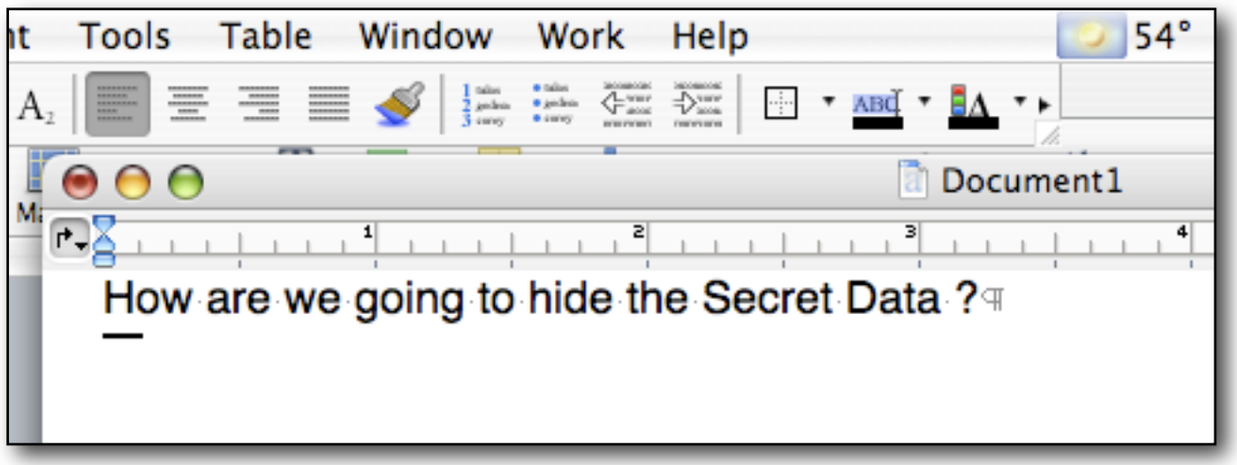



Microsoft Word Files — most violations from *edit history.*

- SCO Word file revealed its anti-Linux legal strategy. (2004)
- Intelligence report by Blair Government was found to be plagiarized from a postgraduate student at the Monterey Institute of International Studies based on transaction log (2003) http://www.computerbytesman.com/privacy/blair.htm

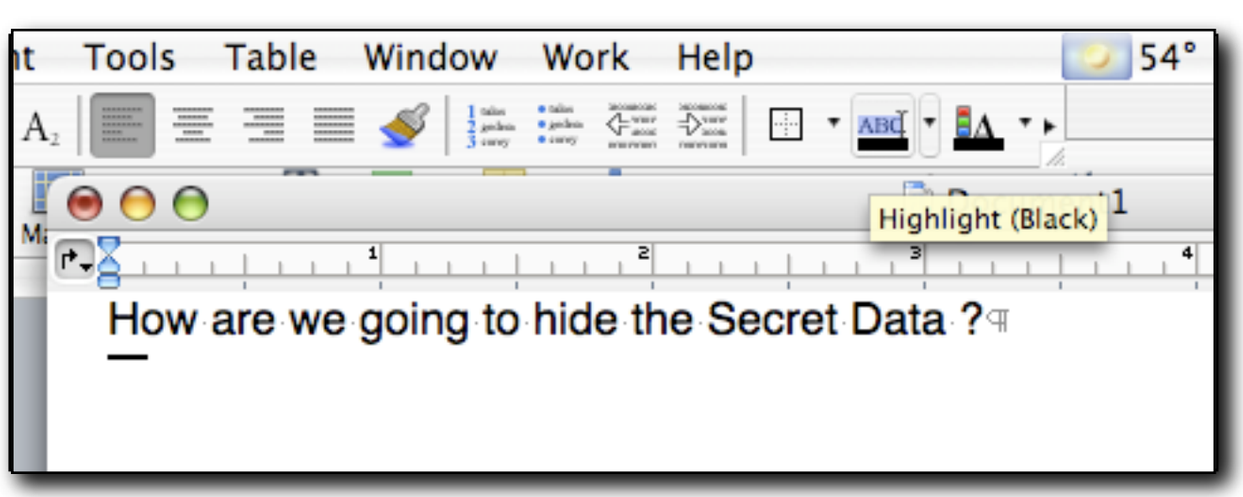
Microsoft Word encourages people to use the highlight feature to eradicate data.

---



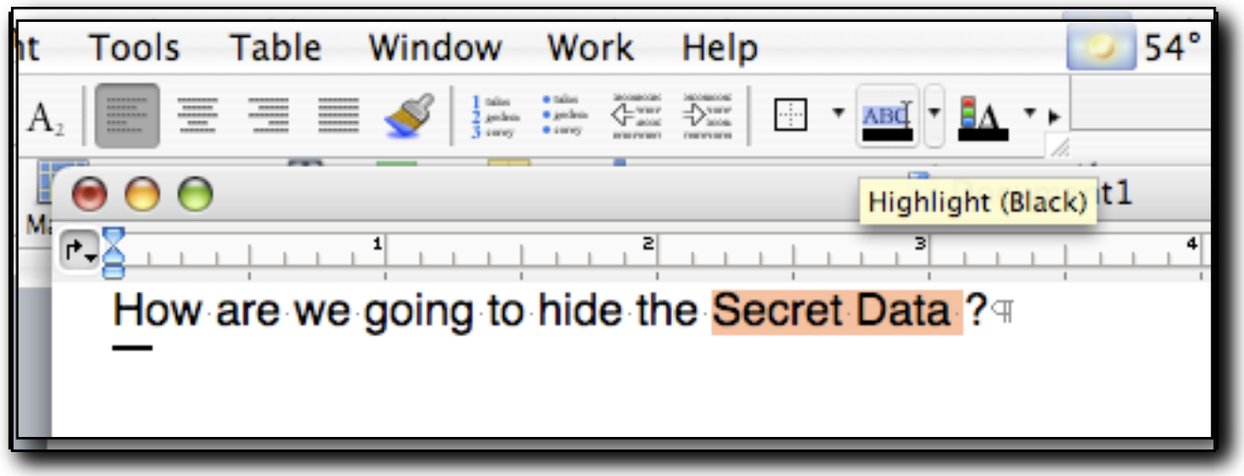
# Microsoft Word encourages people to use the highlight feature to eradicate data.

---



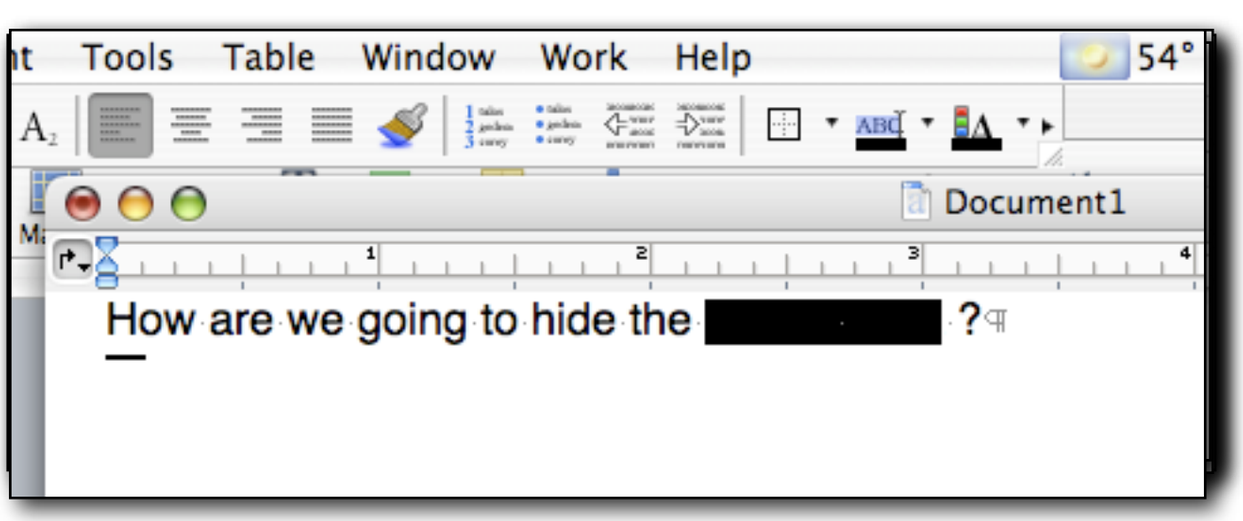
# Microsoft Word encourages people to use the highlight feature to eradicate data.

---



# Microsoft Word encourages people to use the highlight feature to eradicate data.

---



# When Microsoft Word generates the PDF file, "Secret Data" is printed on top of a black box.


This is a "Usable Security" failure.

Microsoft Word is "Security Critical Software."

# There are simple solutions for the residual data problem.

## "Complete Delete."

- Overwrite data when it is deleted.  (FORMAT, CUT, free(), Garbage Collection, etc.)

## Cryptographic File Erasure.

- Encrypt each disk (or file) with a unique key.
- Delete the key when discarding the disk (or deleting the file).

## Physical Destruction.

- It's hard to recover data from a slag of aluminum.


## … But Complete Delete is at odds with "Time Machine."

- "Complete Delete vs. Time Machine Computing,"
  ACM SIGOPS Operating Systems Review,
  January 2007

# Users *must* have tools to protect their privacy.
# Law enforcement *must* have tools to exploit this data.

## Michelle Theer, 2000

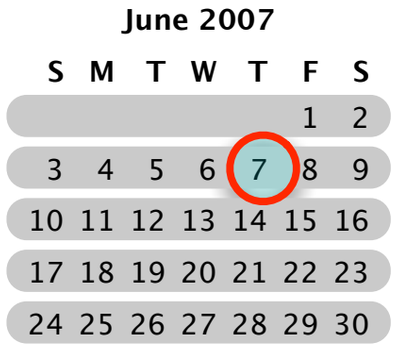
- Husband Air Force Capt. Marty Theer
  shot by Army Staff Sergeant John Diamond on Dec. 17, 2000
- Examination of computer's hard drive found:
  - *21,000 documents, mostly deleted.*
  - *Personal ads that Theer had written in 1999.*
  - *Theer active in swinger's clubs in winter & spring 2000*
  - *Affair between Diamond and Theer started in Spring 2000*

## Bruce Mirken, 1999

- Gay journalist, advocate for rights of gay teenagers.
- Police posing as gay 14-year-old send Mirken child pornography
- Mirken deletes photographs.
- Police raid Mirken's apartment, use forensic software to recover deleted files.
- Case eventually dismissed ($50K in legal bills)
  - *http://www.journalism.sfsu.edu/flux/bayCurrents/mirken.html*
  - *http://gaytoday.badpuppy.com/garchive/events/051799ev.htm*
  - *July 8, 1999, Page 3B, San Jose Mercury News*

# My current research is aimed at improved techniques for analyzing data in principled, automated ways.







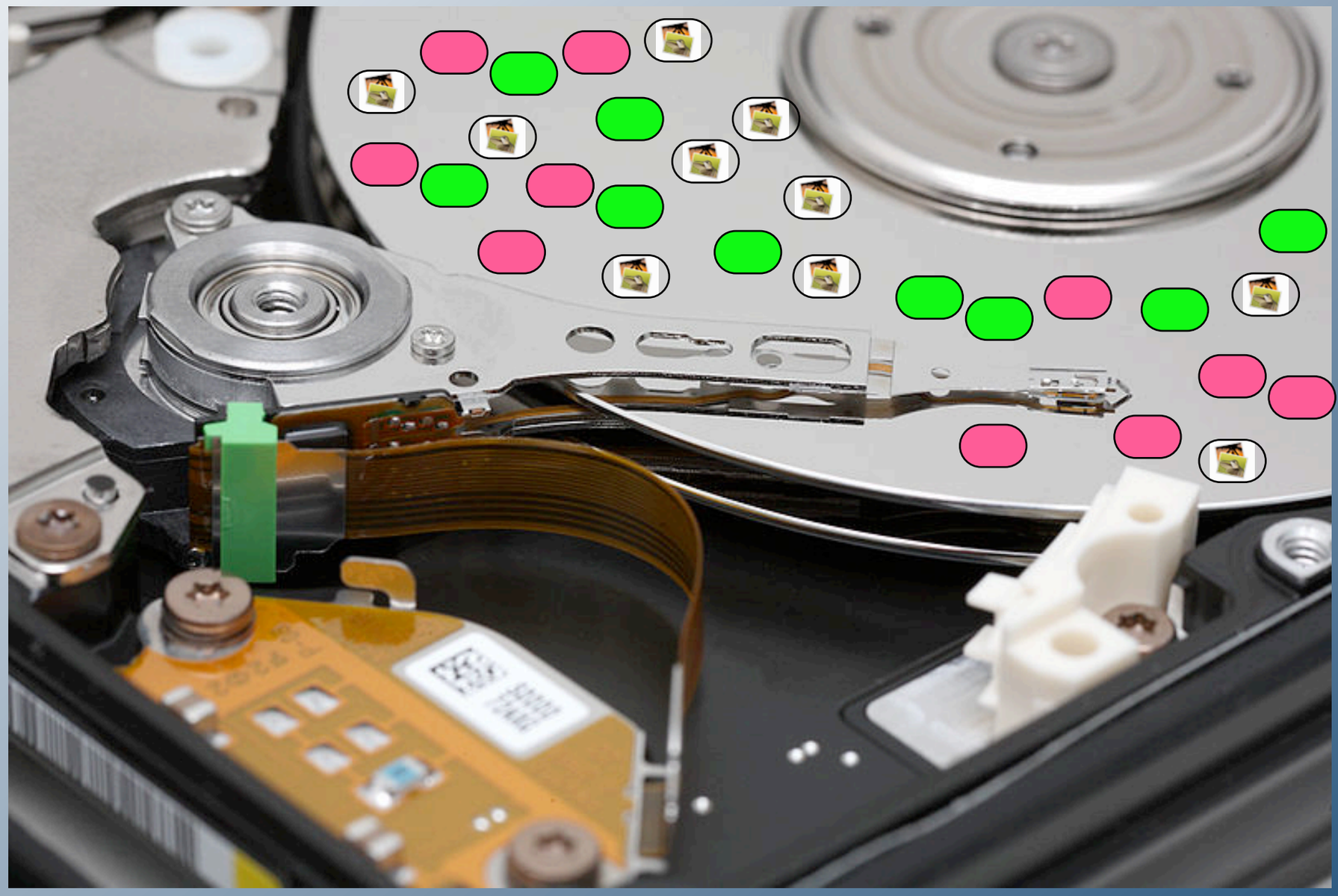


## Today's tools:

- Manual, Slow, Hard to Use

## My approaches:

- New forensic tool chain designed for automation.
- New algorithms designed for a "data rich" environment.
- Scientific validation with  standardized corpora.



# Instant Drive Analysis 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?

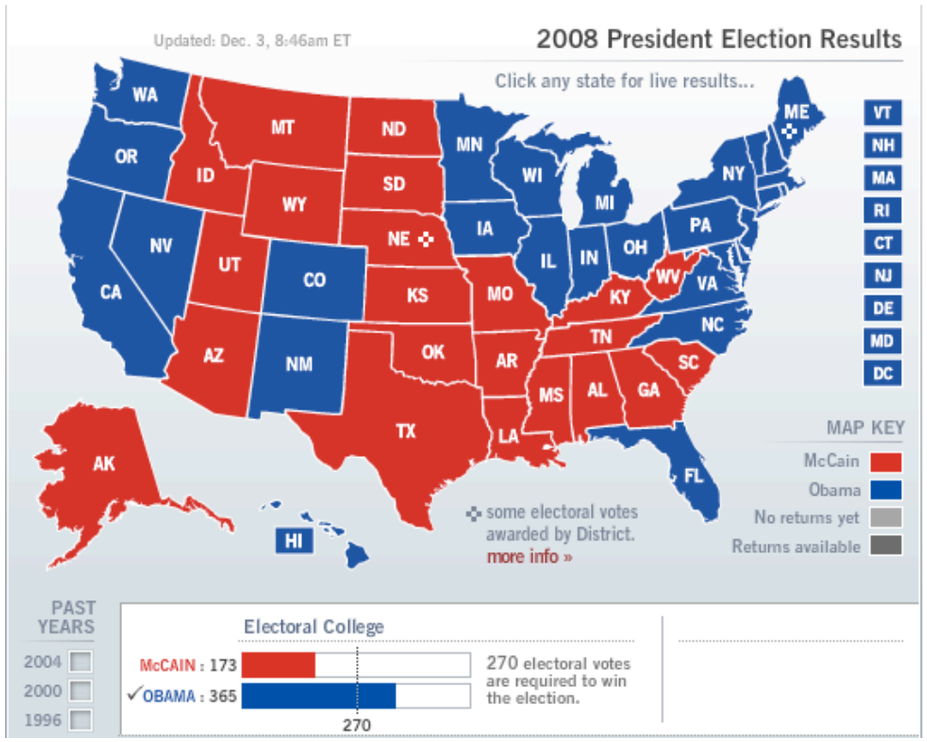
| |  |  |
|---|---|---|
| Minutes | 208 | 1 |
| Max Data Read | 1 TB | 4.8 GB |
| Max Seeks | 15 million | 17,000<br>(≈3.5msec per seek) |

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

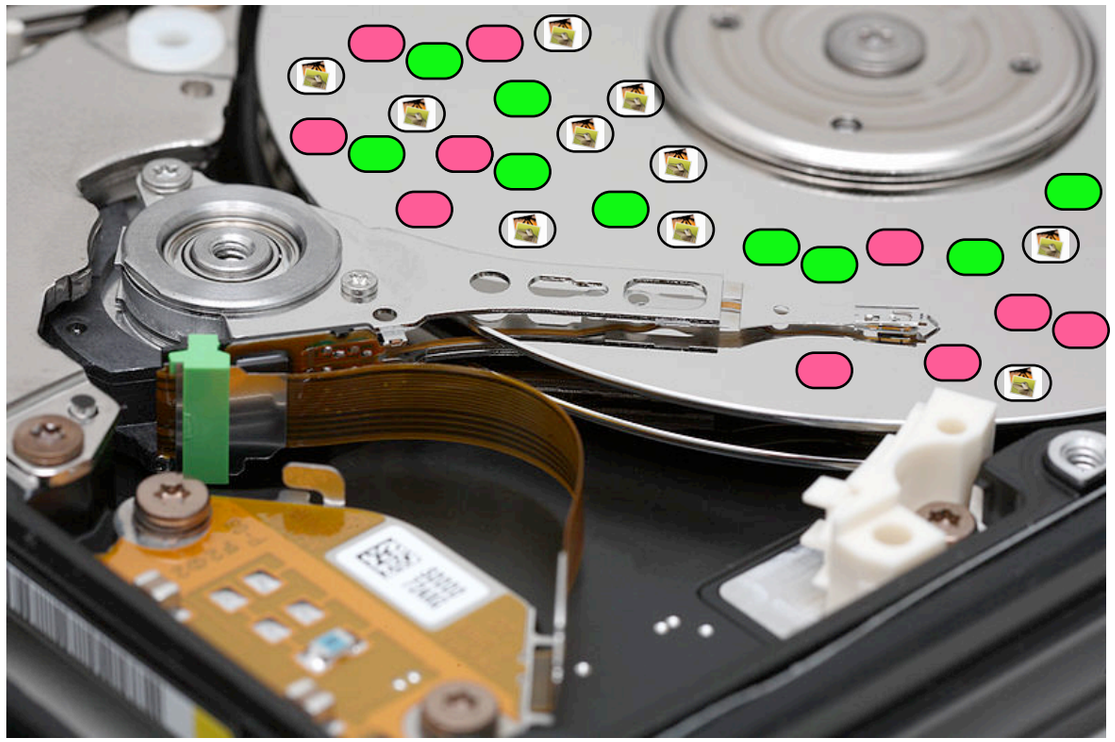
But 4.8 GB is a lot of data!

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

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



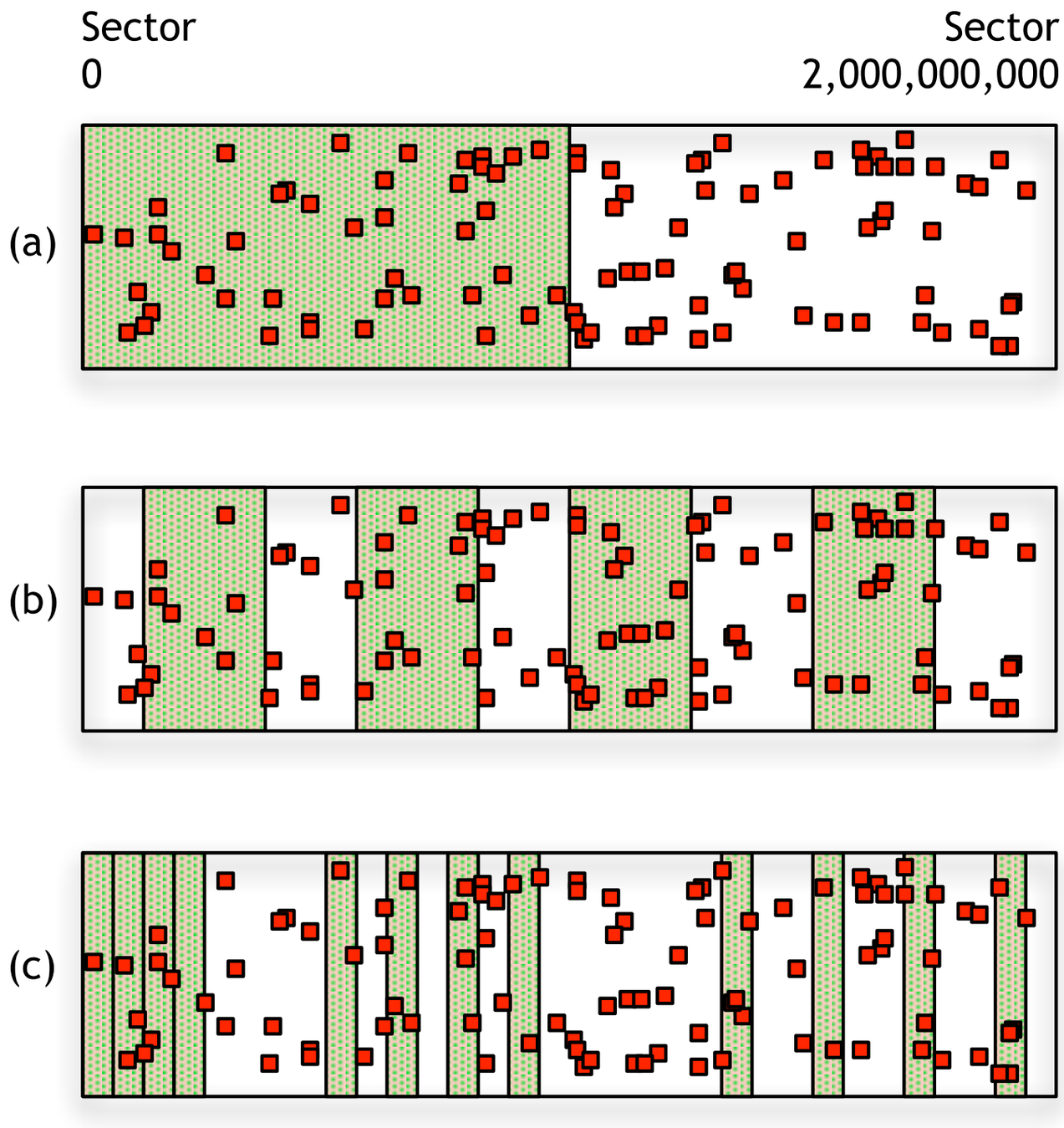

Hard drive contents can be predicted by sampling a few thousand sectors:



The challenge is identifying *likely voters.*

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

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




But sampling has an important limitation...

# Data on hard drives divides into three categories:

Resident Data

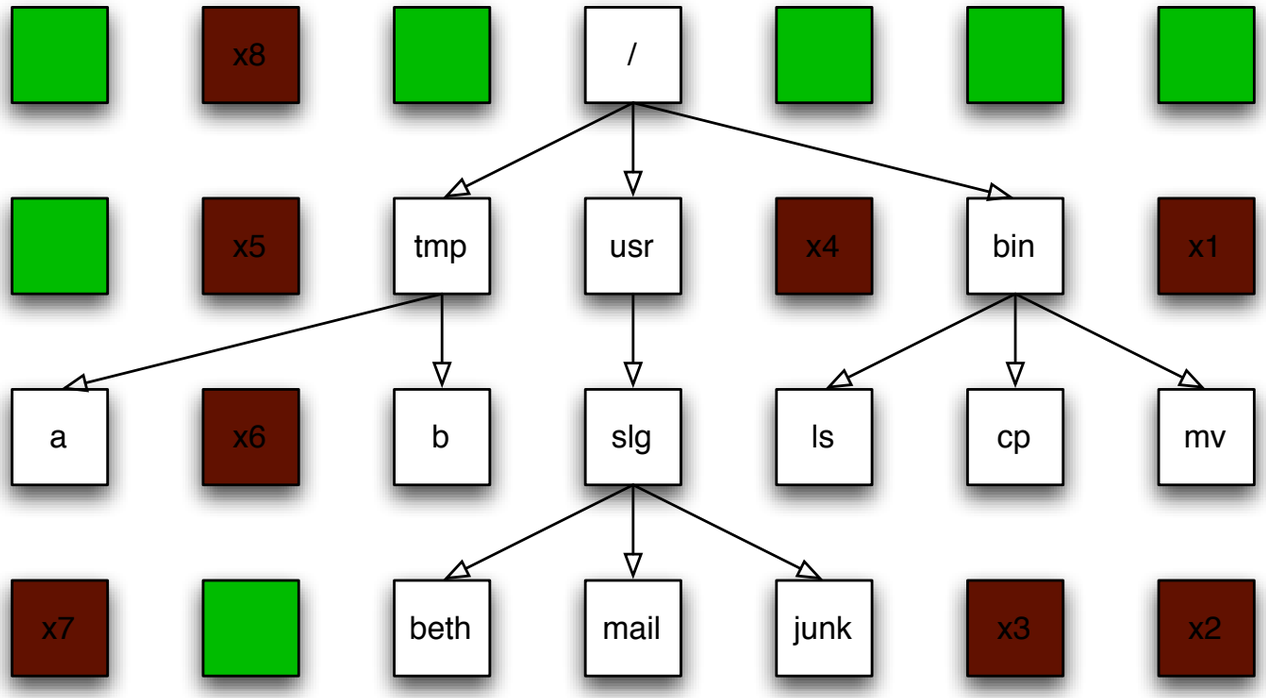
Deleted Data

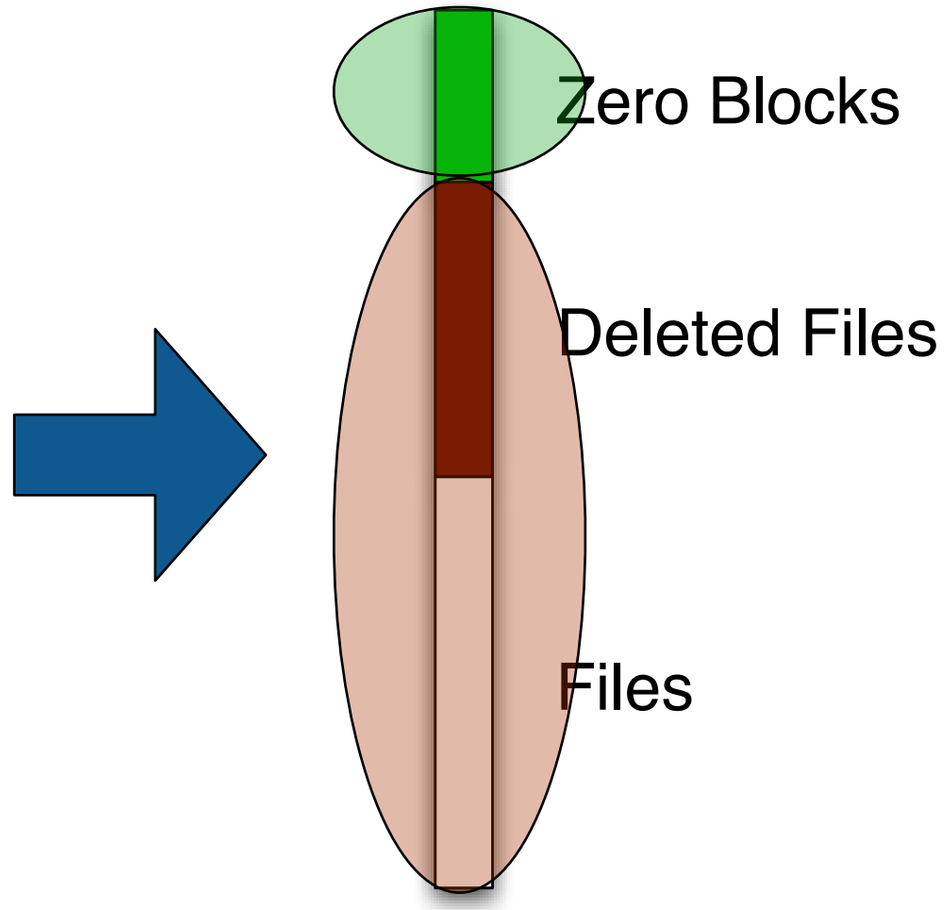
No Data

**user files**
**email messages**
**[temporary files]**

**blank sectors**

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

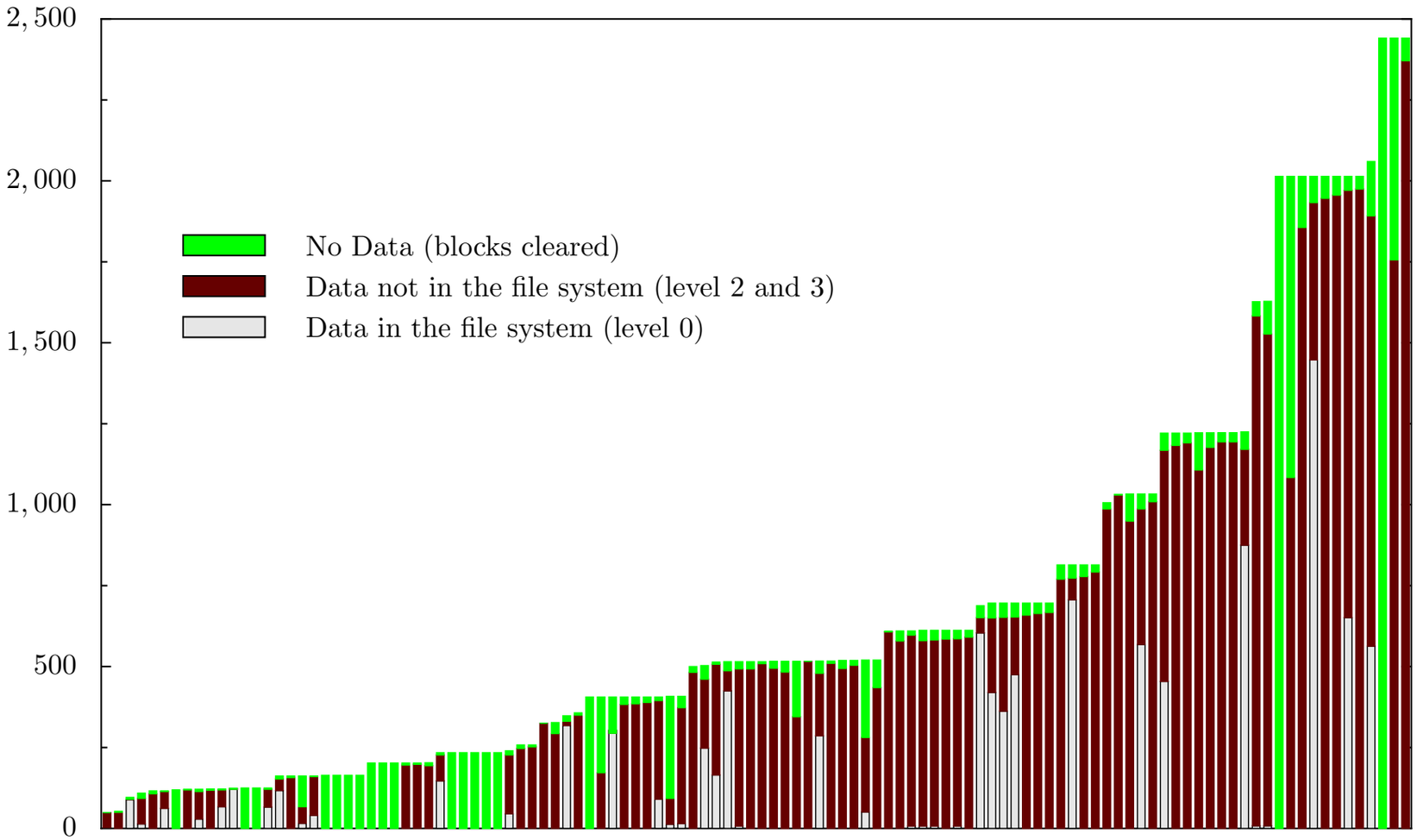


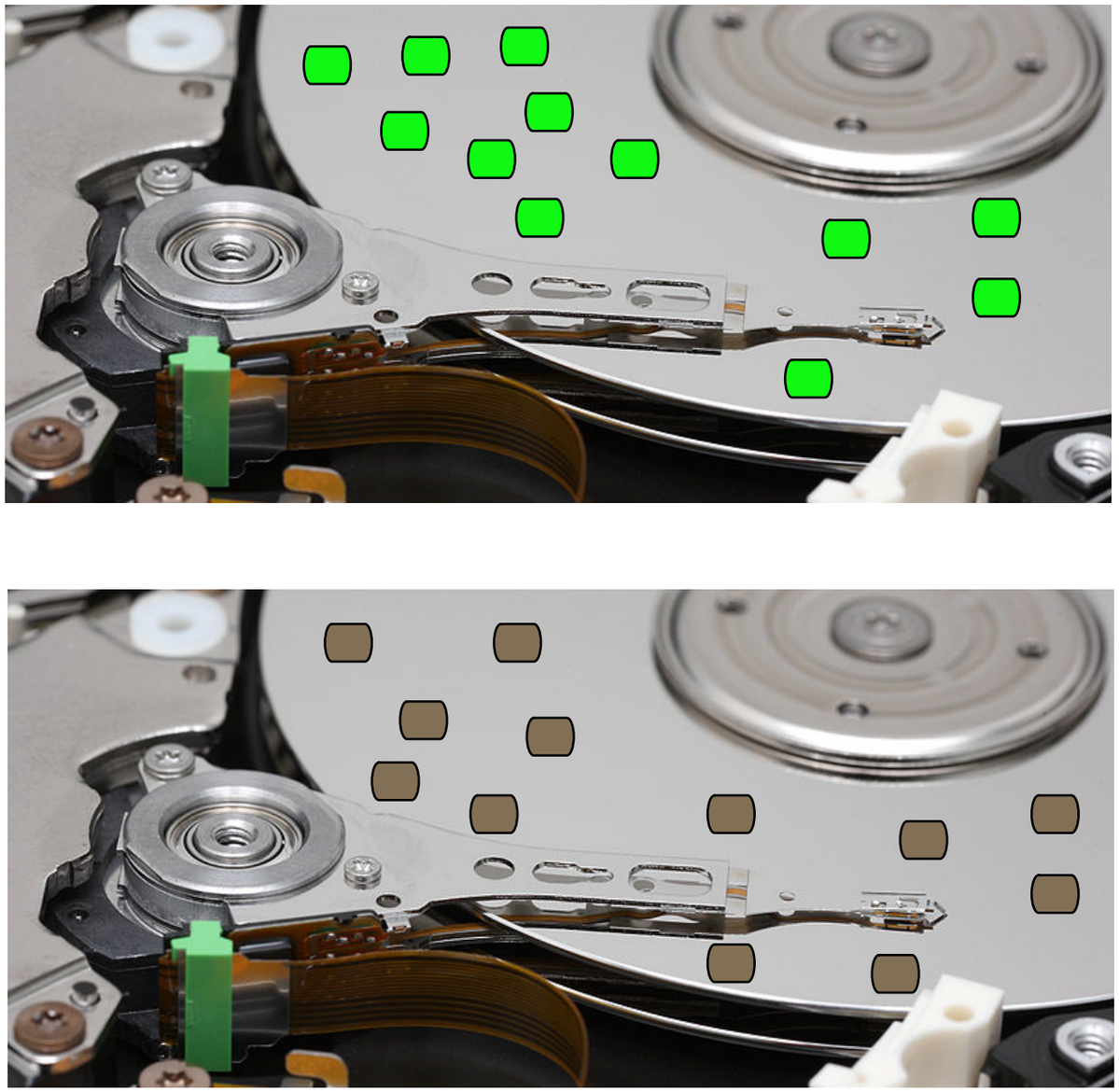
# Let's simplify the problem.
# Can we use statistical sampling to verify wiping?

I bought 2000 hard drives between 1998 and 2006.

Most of were not properly wiped.



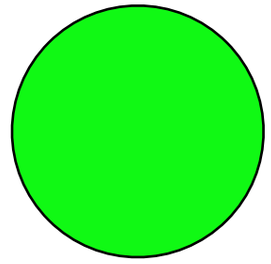

# It should be easy to use random sampling to distinguish a properly cleared disk from one that isn't.



# What does it mean if 10,000 randomly chosen sectors are blank?

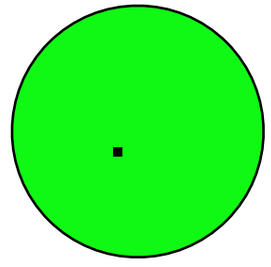
If the disk has 2,000,000,000 blank sectors (0 with data)

- The sample is identical to the population



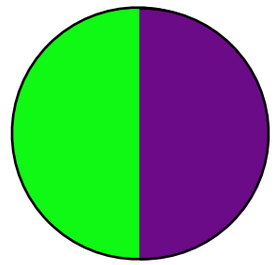
If the disk has 1,999,999,999 blank sectors  (1 with data)

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



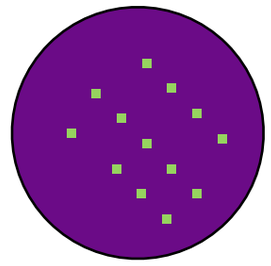
If the disk has 1,000,000,000 blank sectors (1,000,000,000 with data)

- 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!***



# Rephrase the problem.
# Not a blank disk; a disk with less than 10MB of 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 |
| 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.

# We can improve performance by sampling with 4K sectors and using SCAN.

| Drive | Capacity | Interface | Time to read 10,000 sectors Random Order | Sorted Order |
|---|---|---|---|---|
| WD “MyBook” | 1TB | Fire Wire 400 | 156 | 78.3 |
| Seagate Barracuda 7200.11 | 1TB | eSata | 178 | 88.4 |
| PNY | 4GB | USB 2.0 | 33.0 | 30.8 |
| Cruzer micro | 8GB | USB 2.0 | 26.5 | 26.3 |
| Kanguru | 16G | USB 2.0 | 25.9 | 22.7 |

**Table 3:** Time to read 10,000 randomly selected sectors from two representative 1TB hard drives and three representative flash devices. All measurements made with Intel Macintosh hardware with 2.4/2.6 GHz processors running MacOS 10.5. Times are given in seconds and are an average of at least 3 repeated runs.

# Not surprisingly, sampled data has similar statistics to the data as a whole.

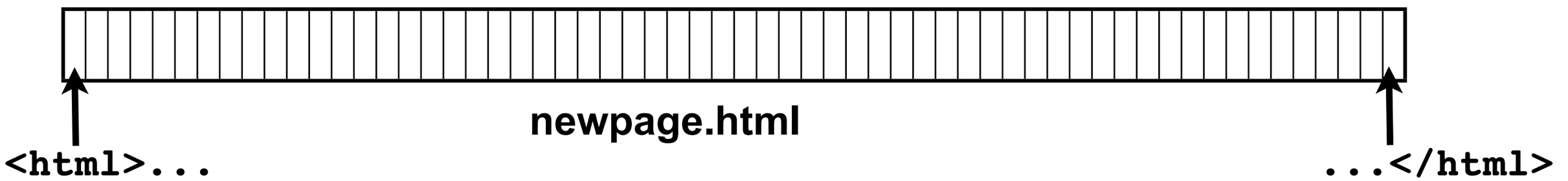
| Image name | Device Size | Sector Statistics Size | Count | # Blank | % Blank | % Sampled Blank by Sector | 4096B block |
|---|---|---|---|---|---|---|---|
| nps-2008-seed1 | 80GB | 512 | 156,250,000 | 3,391,403 | 2.2% | 2.1% | 1.8% |
| nps-2009-domexusers | 43GB | 512 | 83,886,080 | 70,870,456 | 84% 85% | 85% | 85% |
| nps-2009-ipod160 | 160GB | 4096 | 39,023,511 | 31,069,827 | 80% | 80% | |
| nps-2009-ubnist1.gen3 | 2GB | 1024 | 2,057,216 | 1,045,382 | 51% | 51% | 50% |
| nps-2009-ntfs1.gen2 | 517MB | 1024 | 504,448 | 888,279 | 88% | 88% | 86% |
| nps-2009-canon2.gen1 | 31MB | 512 | 60,800 | 2,610 | 4.2% | 4.2% | 4.3%[a] |

[a] Only 1000 sectors sampled.

**Table 4:** Actual and estimated amount of forensically recoverable data on several different pieces of media, as determined by comprehensive disk census, random sampling of 10,000 individual sectors, and random sampling of 10,000 4KiB blocks. Images available at `http://digitalcorpora.org/`

# Part 2: Can we classify files based on a sector?

A file 30K consists of 60 sectors:




Many file types have characteristics headers and footer:

| | header | footer |
|---|---|---|
| HTML | <html> | </html> |
| JPEG | <FF><D8><FF><E0><br><00><10>JFIF<00> | <FF><D9> |
| ZIP | PK<03><0D> | <00><00><00><00> |

But what about the file in the middle?

# 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.

# Discriminator tuning strategy

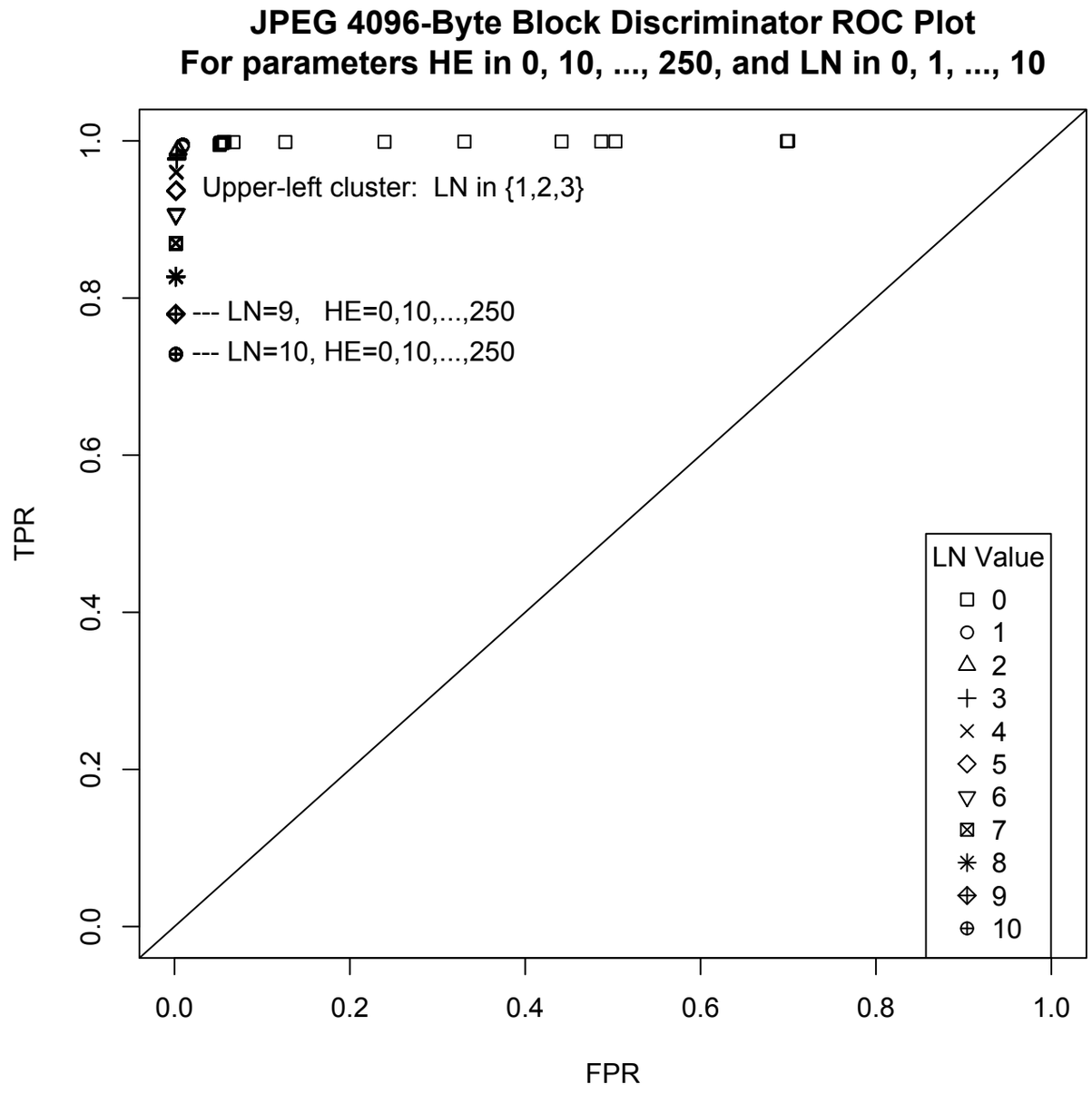
## JPEG discriminator has two parameters:

- High Entropy - needs to have at least HE distinct bytes
- Low FF00 N-grams - needs to have LN <FF><00> byte pairs

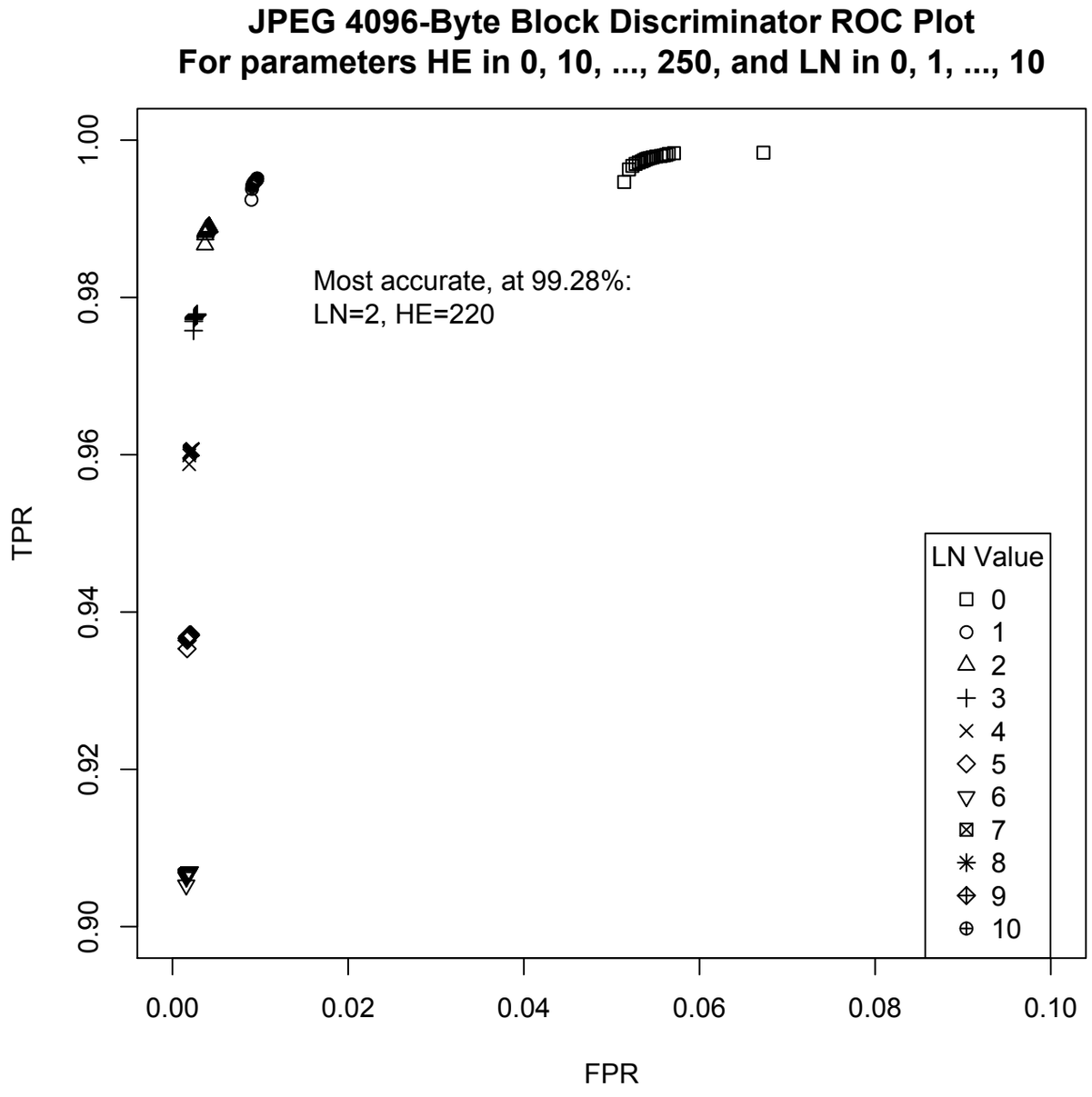
## What value pairs give optimal identification?

## Let’s just try them all!

# Grid search and ROC curves



# Grid search and ROC curves



# This works!
# 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:

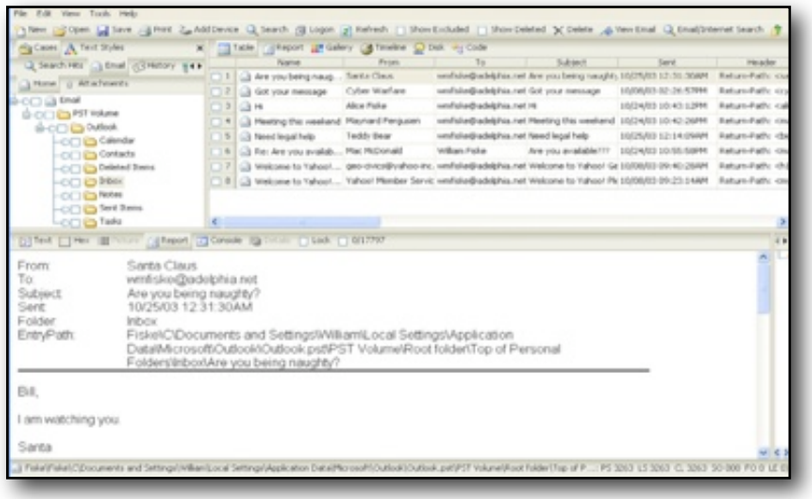
- Roussev, Vassil, and Garfinkel, Simson, File Classification Fragment---The Case for Specialized Approaches, Systematic Approaches to Digital Forensics Engineering (IEEE/ SADFE 2009), Oakland, California.
- Farrell, P., Garfinkel, S., White, D. Practical Applications of Bloom filters to the NIST RDS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

## Work in progress:

- Alex Nelson (PhD Candidate, UCSC) summer project
- Using “Hamming,” our 1100-core cluster for novel SD algorithms.
- Similarity Metric

# Research 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.*

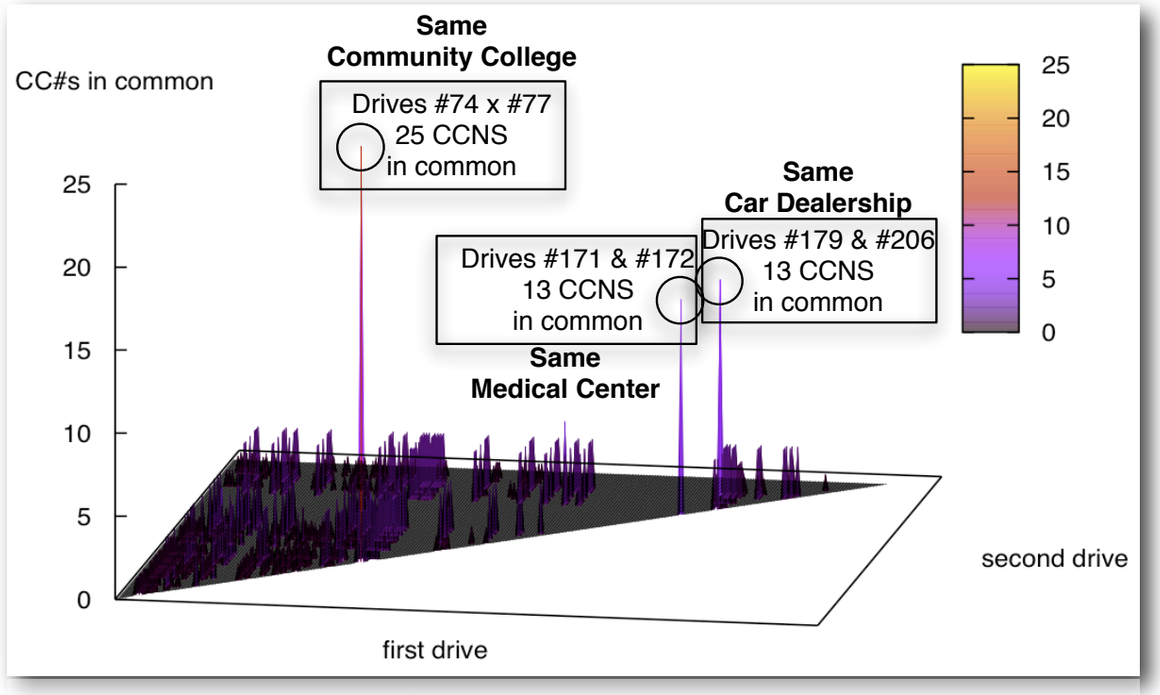
## Evidence Reconstruction

- 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.
- Publication of *data* and *results.*
- *Sharing of scientific materials.*

## ***Today's Digital Forensics is not Scientific!***

- Researchers work on their own data
  - —*Data can't be shared with other researchers  (privacy)*
  - —*Data can't be published  (copyright)*
- Results can't be meaningfully compared.

# Our solution: Standardized Corpora for Digital Forensics Research.

"Standardized"

- Known contents
- Documented provenance

"Corpora"

- Many data sets
- *Realistic* — lifelike, but no Personally Identifiable Information (PII)
- *Real* — Public and Private

"Digital Forensics Research"

- Created to enable *research*
- Legally obtained (c.f. wiretap law)
- Publishable results
- Specific attention to *privacy* and *copyright* issues

# http://domex.nps.edu/corp/files/govdocs1: 1 Million files available *now*

## 1 million(*) documents 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.

**034164.jpg**

## 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.)

---

(*Approximately 3000 files redacted after release.)

# http://domex.nps.edu/corp/scenarios/ Complete Scenarios

## Typical scenarios include:

- Distribution of simulated pornography ("kitty porn.")
- Theft of corporate data.

## Nitroba University:

- University harassment case


## m57 theft

- Theft of corporate data

## m57 patents

- 3 week simulation of a small business
- Four computers
- Daily disk and memory images
- Complete Network Packet Capture

# 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/

## Work in progress:

- Joshua Gross, NPS postdoc, 2009-2010

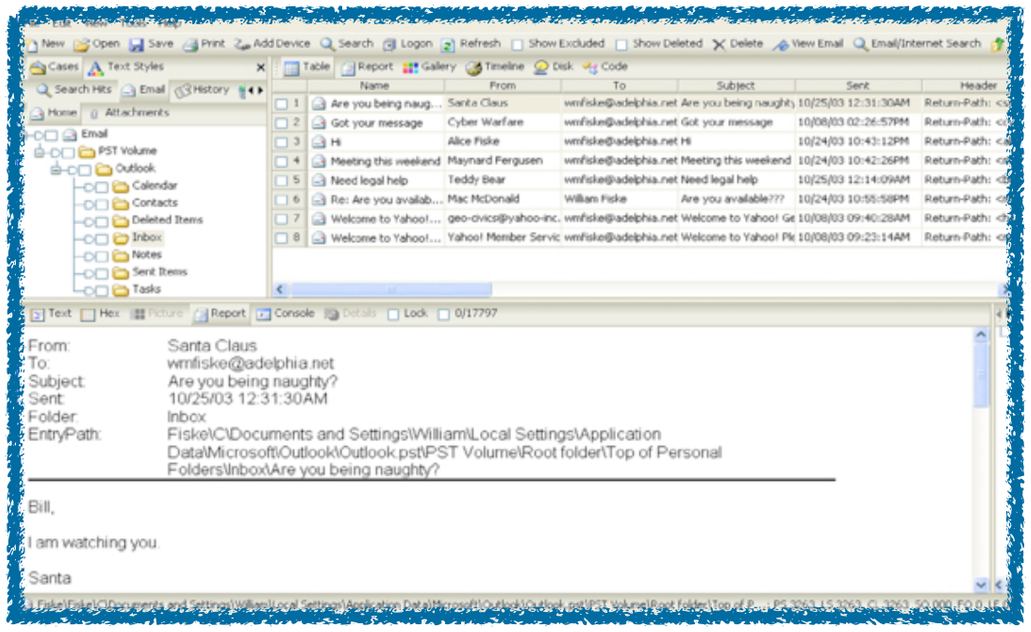
# Open Research Problems

# Automated Forensics: Research requires Open Source Tools

Commercial tools are Windows-based GUIs for a single examiner

These tools stress:

- Reverse Engineering
- Visibility
- Search



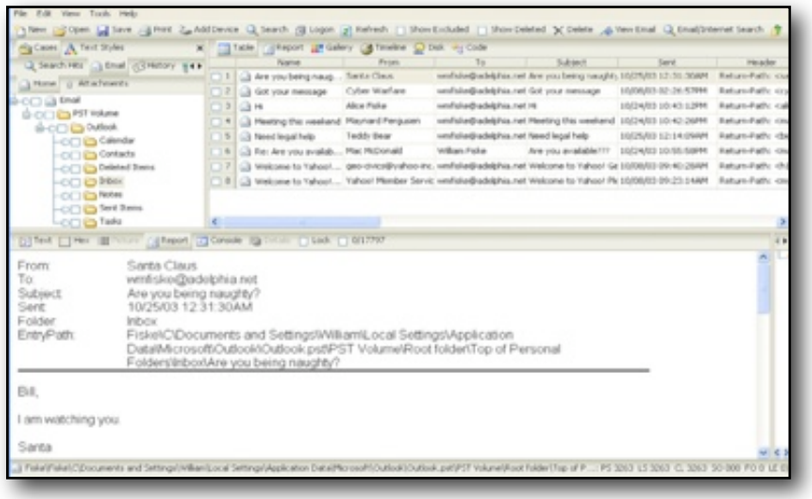

Open Source tools:

- Make it possible to conduct experiments.
- Allow for repeatable research.
  - —*But these tools need all the same reverse engineering!*

Current reverse engineering needs:

- NTFS Encryption
- HFS+ variants used on iPods; XFAT
- Extractors for file formats; a unified system for metadata extraction
- Protection against intentionally corrupt file systems

# Forensics needs better visualization tools




File systems are 1-10TB; critical evidence is 1-10MB

Key challenges:

- Timeline visualization
- Automated logfile correlation & analysis
- Histogram analysis.
- Combining natural language analysis with forensics

# Flash Forensics: A whole new environment

Traditional assumptions of forensics do not apply to flash.

We may be able to recover:

- Overwritten data — by examining physical characteristics of flash cells.
- Write order — by examining the FTL.
- "Invisible data" — with vendor-specific commands.


Physical access matters

Currently needed:

- Flash file system implementations (JFFS, JFFS2, YAFFS, etc.)
- Access to the physical layer.

# Cell phones are a nightmare

## Cell phones have:

- Many different operating systems
- Different layout of programs
- Downloadable applications (sometimes)
- Multiple processors
- Multiple address spaces
- Non-standard connectors


## Open questions for Cell Phones:

- How do you get all of the data out?
  - —*From a cell phone you have never seen before?*
- How do you decode the data?

# Game consoles

## Difficulties:

- DRM and encryption
- Proprietary operating systems



## Why you should care:

- Game consoles are being used to commit crimes.
- Good model for future of "protected" systems.

# In Summary

# 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?

## For further information:

- http://simson.net/
- http://forensicswiki.org/
- http://afflib.org/
- http://digitalcorpora.org/


- simsong@acm.org


## Sample Questions:

- What about Facebook?
- What impact will Full Disk Encryption have on computer forensics?
- Is it possible to recover data once it has been overwritten?
- What are the opportunities for face recognition and other content analysis techniques?
- What can you "correlate" other than email addresses?
- What's wrong with MD5?
- What's the best language for writing forensics programs?
- Is Open Source really necessary for satisfying Daubert?
- Why is EMACS better than vi?