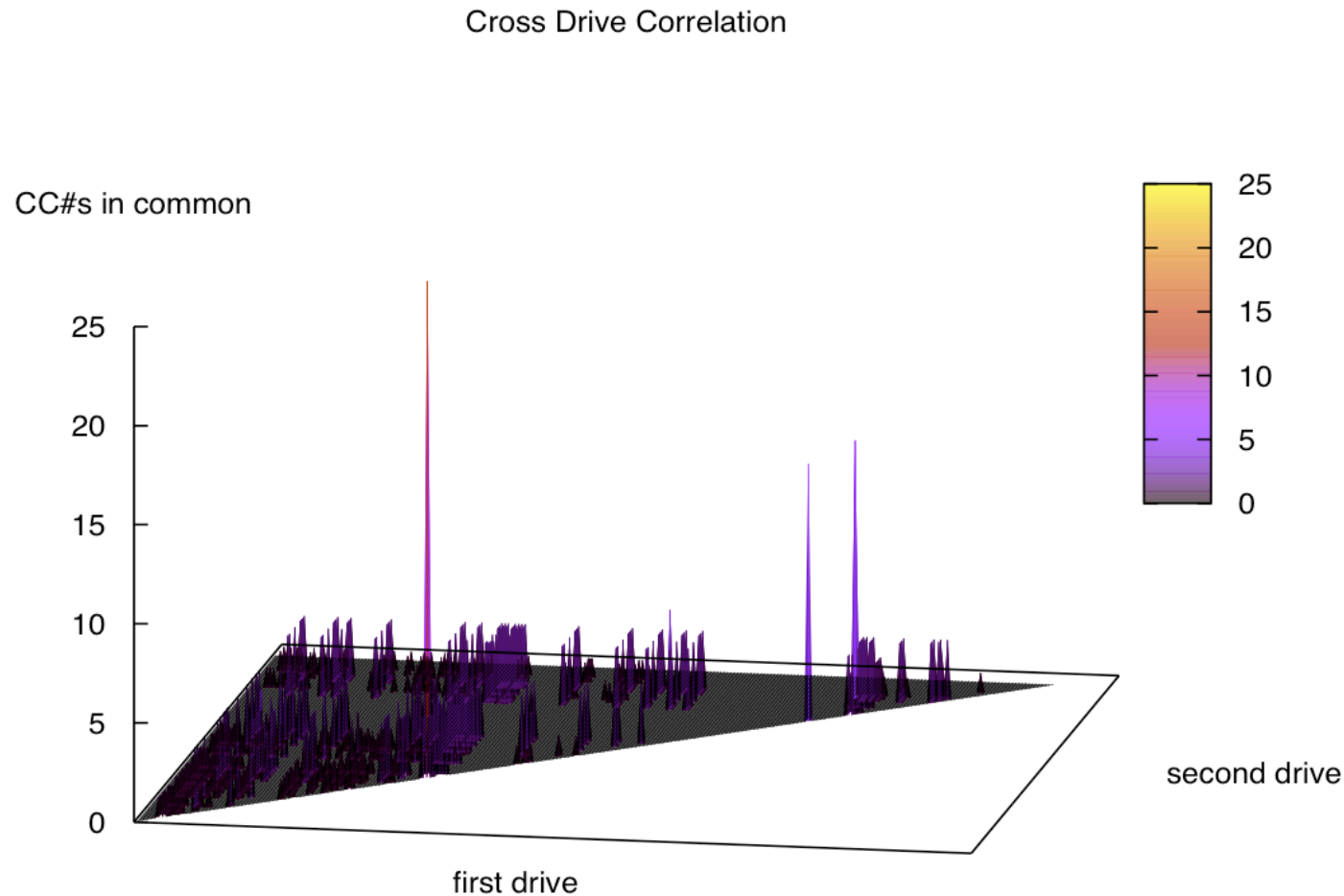


Disk Sanitization and Cross Drive Forensics



Simson L. Garfinkel
Center for Research on Computation and Society
Harvard University
September 26, 2005

Purchased used from a computer store in August 1998:



Computer #1: 486-class machine with 32MB of RAM

A law firm's file server...
...with client documents!



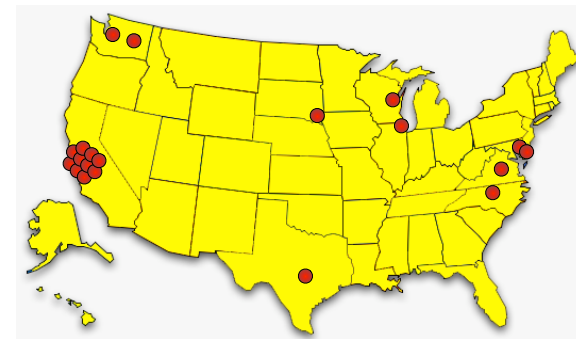
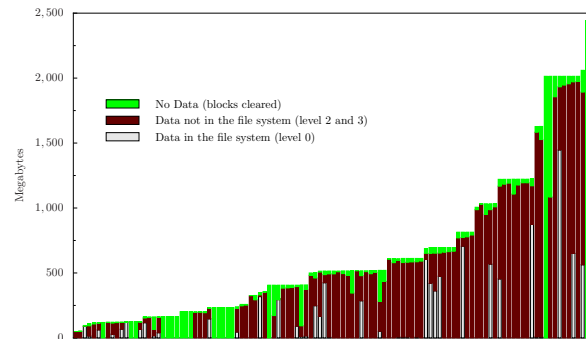
Computers #2 through #10 had:

- Mental health records
- Home finances
- Draft of a novel...

Was this a chance accident or common occurrence?

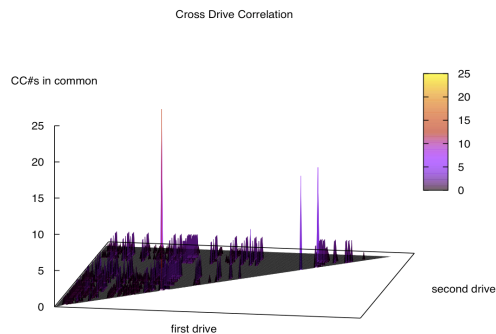
This talk presents the disk sanitization problem and discusses a new technique for computer forensics.

1. Scale of the problem



2. The Traceback Study

3. Cross Drive Forensics



Hard drives pose special problem for computer security

Do not forget data when power is removed.

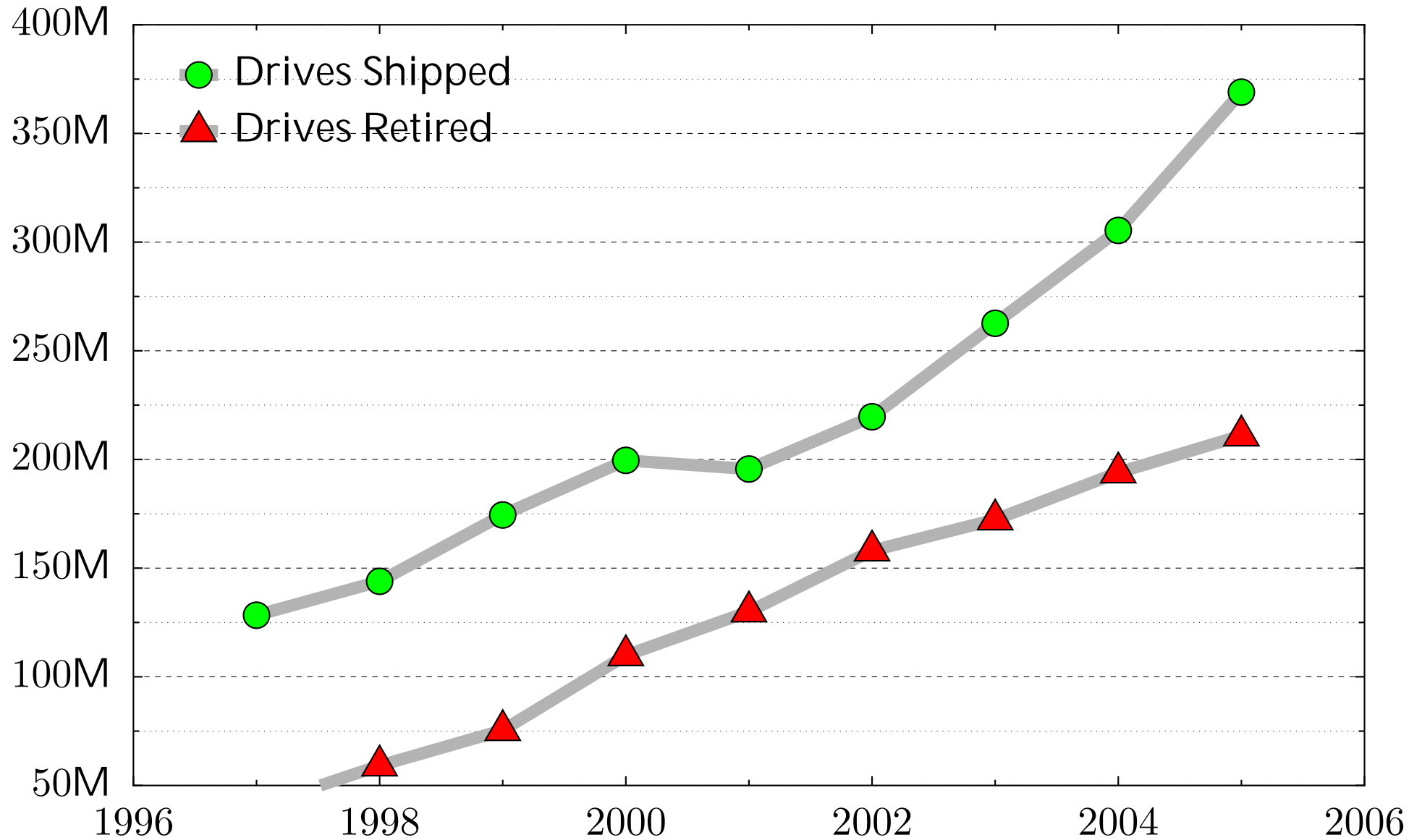
Contain data that is not immediately visible.

Today's computers can read hard drives that are 15 years old!

- Electrically compatible (IDE/ATA)
- Logically compatible (FAT16/32 file systems)
- Very different from tape systems

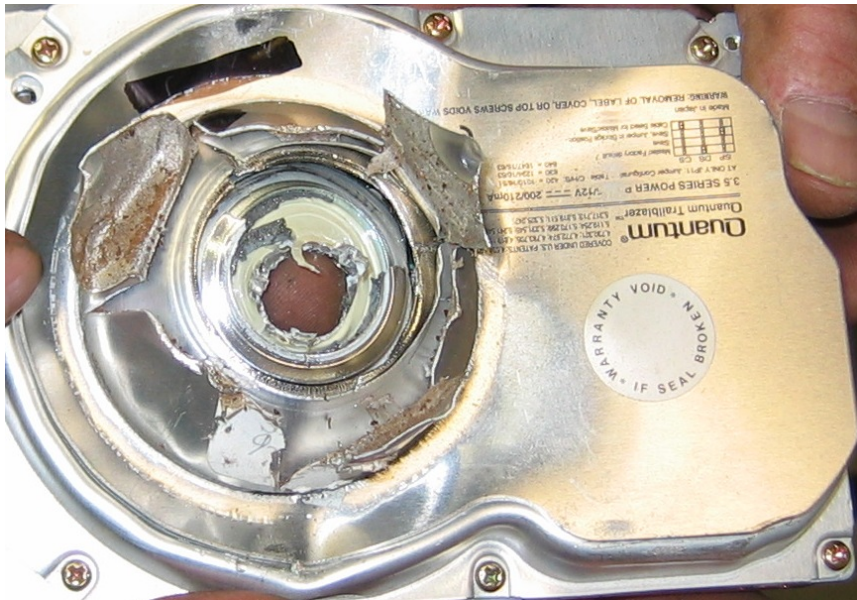


Scale of the problem: huge!



210 million drives will be retired this year.

Physical destruction will remove the information...



...but many “retired” drives are not physically destroyed.

There is a significant secondary market for used disk drives.



Retired drives are:

- Re-used within organizations
- Given to charities
- Sold at auction

All Categories [Save this search](#)

350 items found for hard drives

Sort by items: [ending first](#) | [newly listed](#) | [lowest priced](#) | [highest priced](#)

Picture	Item Title	Price	Bids	Time Left
	Lot of hard and floppy drives	\$5.50	2	14m
	Lot of hard and floppy drives	\$5.50	2	22m
	Lot of hard and floppy drives	\$5.50	2	25m
	Lot of 2 hard drives IDE	\$8.00	12	29m
	3.2 gig Hard Drives	\$180.00	-	59m
	(5) 1.2 hard drives & (15) 10/100 network	\$15.00	1	1h 00m
	Lot of 3 Quantum 9.1 gig SCSI Hard Drives	\$16.00	6	1h 25m
	IDE HARD DRIVES (3)	\$6.50	6	1h 46m
	LOT OF 5 Hard Drives! 3.2 Gig Western Digital	\$120.00 \$124.95 74% Buy Now	-	1h 50m
	QTY 3... IDE Hard Drives 2.5 Gig	\$10.50	5	2h 02m
	5 WESTERN DIGITAL 2.5 GIG HARD DRIVES	\$30.00	4	2h 03m
	QTY 3... IDE Hard Drives 1.0 Gig	\$9.99	1	2h 04m
	Western Digital 850 meg IDE Hard Drives - each	\$6.00	1	2h 57m
	WINDOWS	\$6.00	-	3h 18m

About 1000 used drives/day sold on eBay.

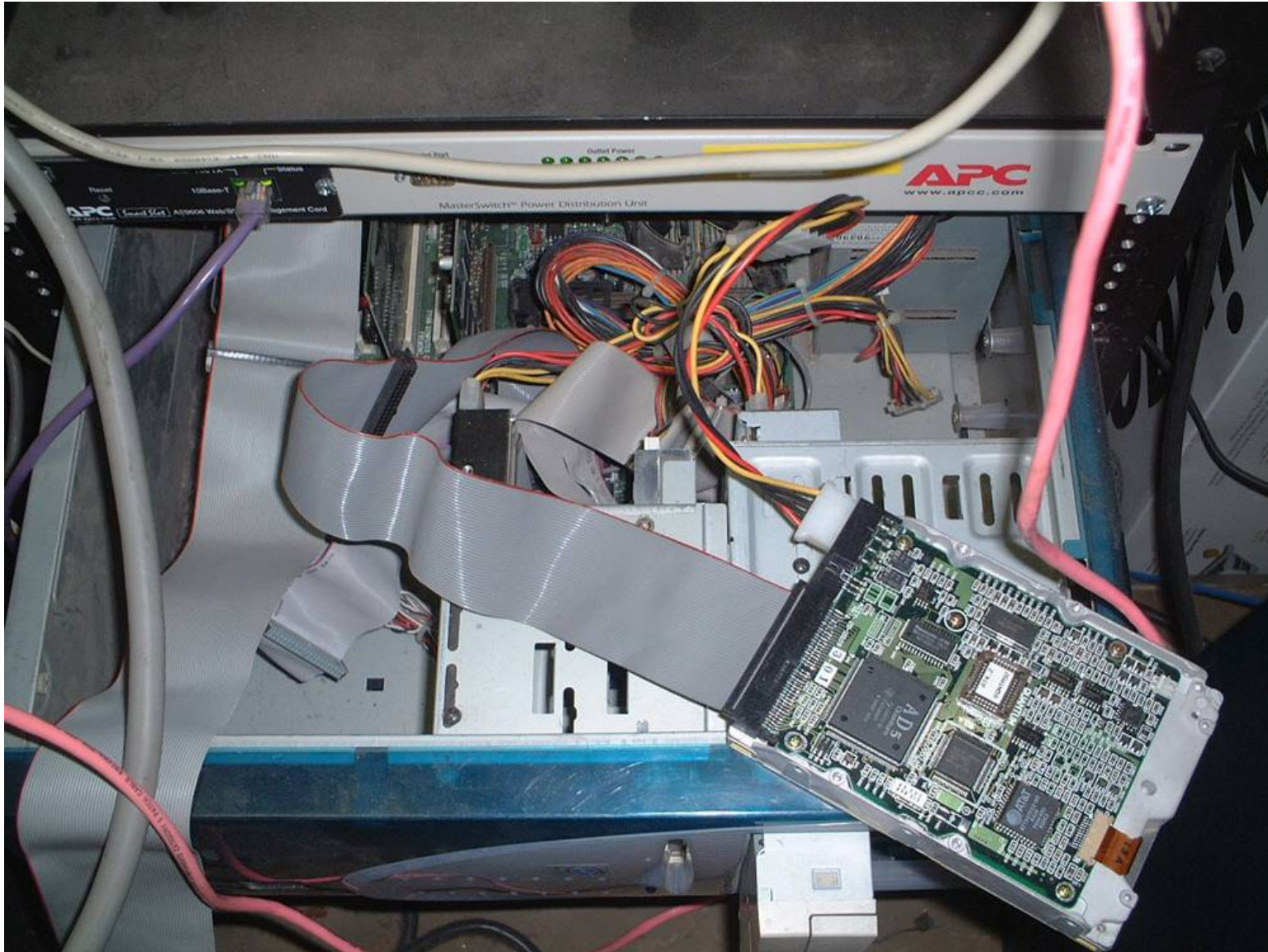
**Between January 1999 and April 2002,
I acquired 236 hard drives on the secondary market.**



Drives arrived by UPS



Data on drives “imaged” using FreeBSD



```
dd if=/dev/ad0 of=file.img bs=65536 conv=noerror,sync
```

Images stored on a RAID



For every drive, I cataloged:

- Disk SN, date of manufacture, etc.
- Every readable sector on the drive..
- All visible files.
- MD5 of every file.
- MD5 of the image.



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

Purchased for \$5 from a Mass retail store on eBay

Copied the data off: 541MB

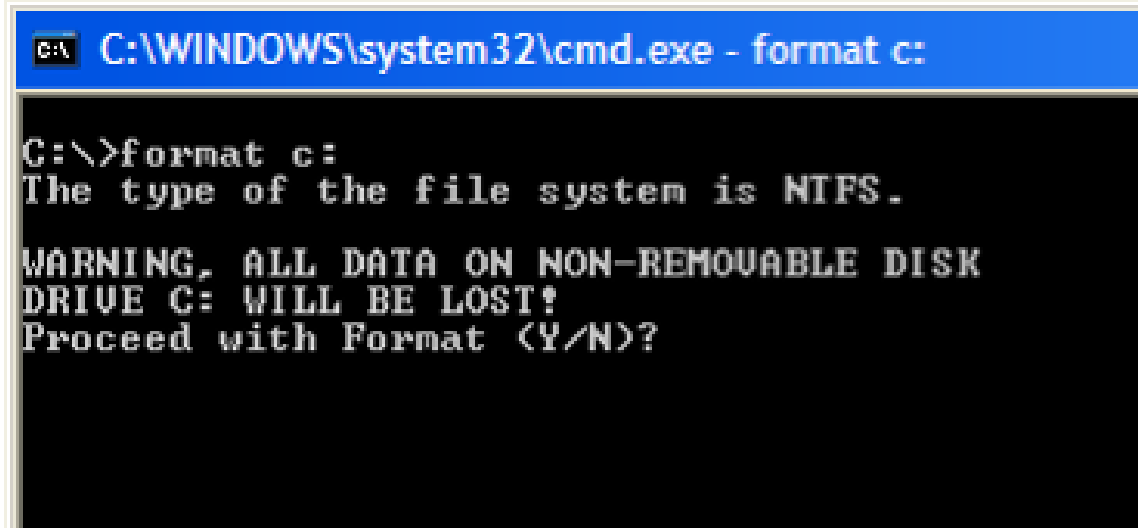
Initial analysis:

Total disk sectors:	1,057,392
Total non-zero sectors:	989,514
Total files:	3

The files:

drwxrwxrwx	0	root	0	Dec 31	1979	./
-r-xr-xr-x	0	root	222390	May 11	1998	IO.SYS
-r-xr-xr-x	0	root	9	May 11	1998	MSDOS.SYS
-rwxrwxrwx	0	root	93880	May 11	1998	COMMAND.COM

Clearly, this disk had been FORMATED...



```
C:\WINDOWS\system32\cmd.exe - format c:

C:\>format c:
The type of the file system is NTFS.

WARNING, ALL DATA ON NON-REMOVABLE DISK
DRIVE C: WILL BE LOST!
Proceed with Format (Y/N)?
```

**Windows FORMAT doesn't erase the disk...
FORMAT just writes a new root directory.**

UNIX “strings” reveals the disk’s previous contents...

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 lo

To override this protection, see the LOCK /? command for m

The system has been halted. Press Ctrl+Alt+Del to restart

You started your computer with a version of MS-DOS incompat

version of Windows. Insert a Startup diskette matching thi

OEMString = "NCR 14 inch Analog Color Display Enhanced SV

Graphics Mode: 640 x 480 at 72Hz vertical refresh.

XResolution = 640

YResolution = 480

VerticalRefresh = 72

70.img con't...

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, 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 separate

EAS.STCm

EET.STC

ELR.STCq

ELS.STC

70.img con't...

MAB-DEDUCTIBLE

MAB-MOOP

MAB-MOOP-DED

METHIMAZOLE

INSULIN (HUMAN)

COUMARIN ANTICOAGULANTS

CARBAMATE DERIVATIVES

AMANTADINE

MANNITOL

MAPROTILINE

CARBAMAZEPINE

CHLORPHENESIN CARBAMATE

ETHINAMATE

FORMALDEHYDE

MAFENIDE ACETATE

[Garfinkel & Shelat 03] established the scale of the problem.

We found:

- Thousands of credit card numbers (many disks)
- Financial records
- Medical information
- Trade secrets
- Highly personal information



We did not determine why the data had been left behind.

There are roughly a dozen documented cases of people purchasing old PCs and finding sensitive data.

- A woman in Pahrump, NV bought a used PC with pharmacy records [Markoff 97]
- Pennsylvania sold PCs with “thousands of files” on state employees [Villano 02]
- Paul McCartney’s bank records sold by his bank [Leyden 04]
- O&O Software GmbH – 200 drives.[O&O 05]



None of these cases are scientifically rigorous.

Why don't we hear more stories?

Hypothesis #1: Disclosure of “data passed” is exceedingly rare because most systems are properly cleared.

Hypothesis #2: Disclosures are so common that they are not newsworthy.

Hypothesis #3: Systems aren't properly cleared, but few people notice the data.

How could people not notice the data?

```
C:\WINDOWS\system32\cmd.exe

C:\tmp>dir
Volume in drive C has no label.
Volume Serial Number is 1410-FC4A

Directory of C:\tmp

10/15/2004  09:20 PM    <DIR>          .
10/15/2004  09:20 PM    <DIR>          ..
10/03/2004  11:34 AM                27,262,976 big_secret.txt
           1 File(s)                27,262,976 bytes
           2 Dir(s)          4,282,878,288 bytes free

C:\tmp>del big_secret.txt

C:\tmp>dir
Volume in drive C has no label.
Volume Serial Number is 1410-FC4A

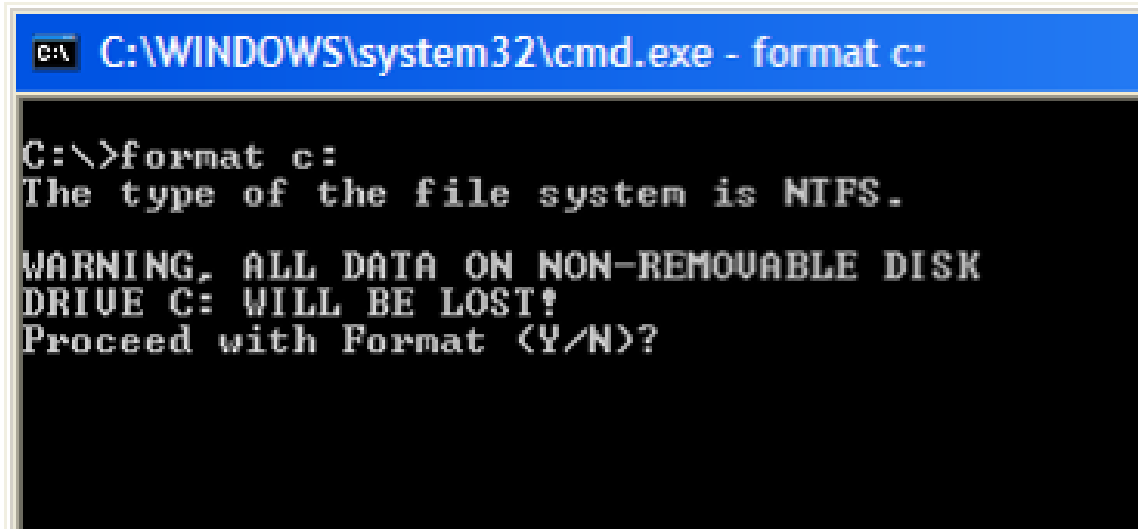
Directory of C:\tmp

10/15/2004  09:22 PM    <DIR>          .
10/15/2004  09:22 PM    <DIR>          ..
           0 File(s)                   0 bytes
           2 Dir(s)          4,229,296,128 bytes free

C:\tmp>_
```

DEL removes the file's name; doesn't delete the data.

FORMAT writes a new root directory and FAT.



```
C:\WINDOWS\system32\cmd.exe - format c:

C:\>format c:
The type of the file system is NTFS.

WARNING, ALL DATA ON NON-REMOVABLE DISK
DRIVE C: WILL BE LOST!
Proceed with Format (Y/N)?
```

FORMAT doesn't doesn't overwrite the disk sectors.

I think that data left behind on hard drives is a serious social problem.

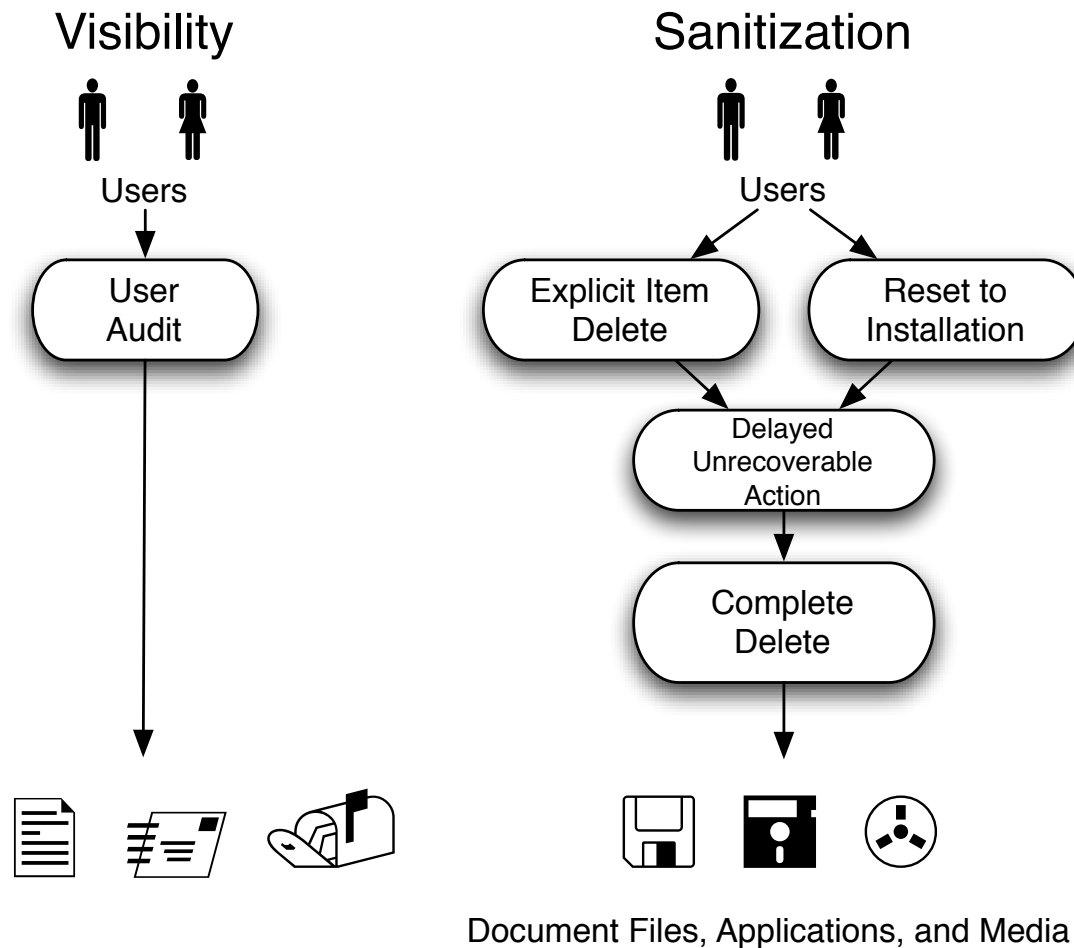
Large numbers of drives are being sold and given away.

Many of them appear to have hidden confidential information.



We are morally obligated to solve this problem!

[Garfinkel '05] presents five distinct patterns for addressing the sanitization problem



<http://www.simson.net/thesis/>

To be effective, a solution must address the root cause

Usability Problem:

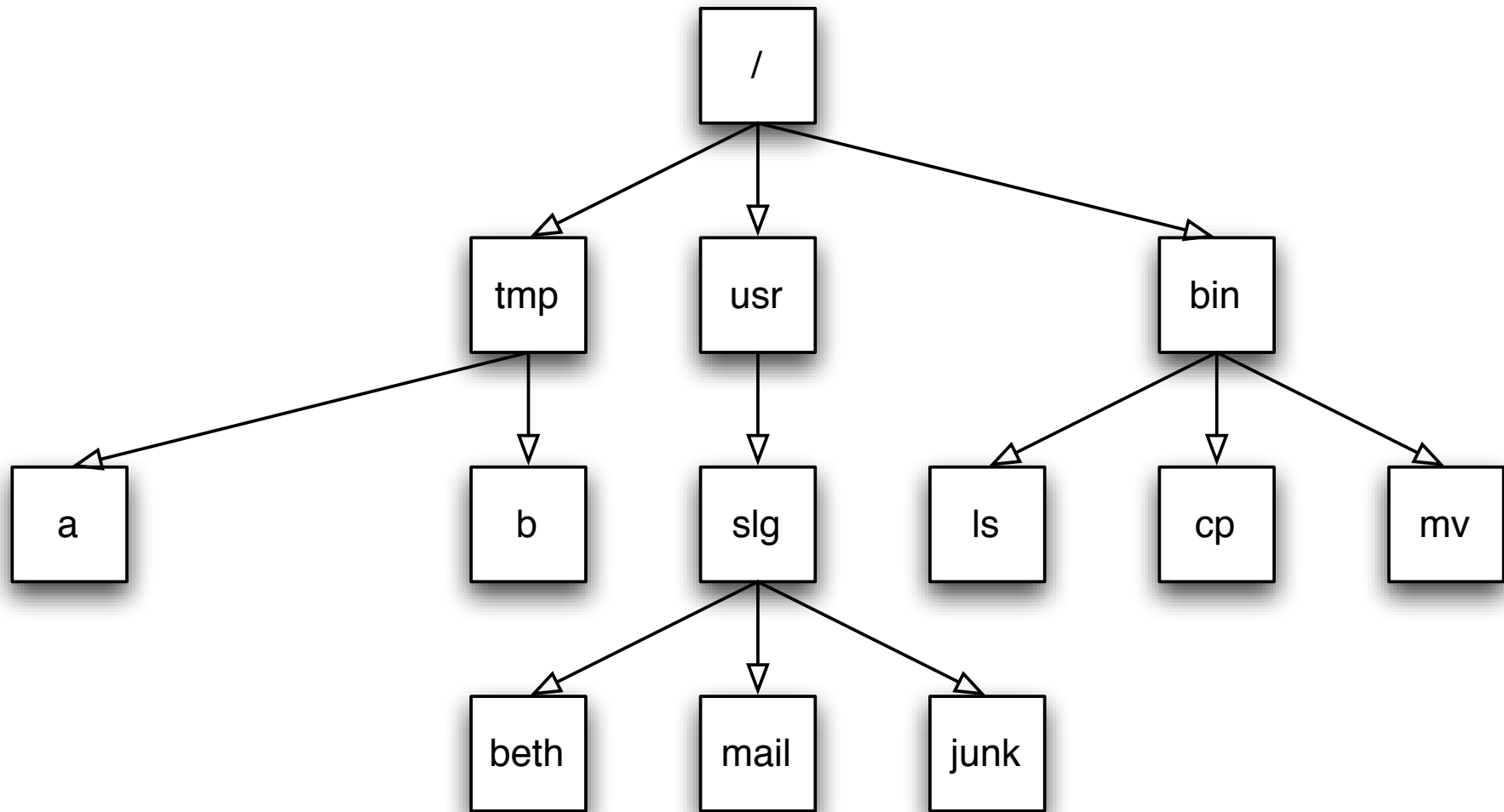
- Effective audit of information present on drives.
- Make DEL and FORMAT actually remove data. [Bauer & Priyantha 01]
- Provide alternative strategies for data recovery.

Education Problem:

- Add training to the interface. [Whitten 04]
- Regulatory requirements. [FTC 05, SEC 05]
- Legal liability.

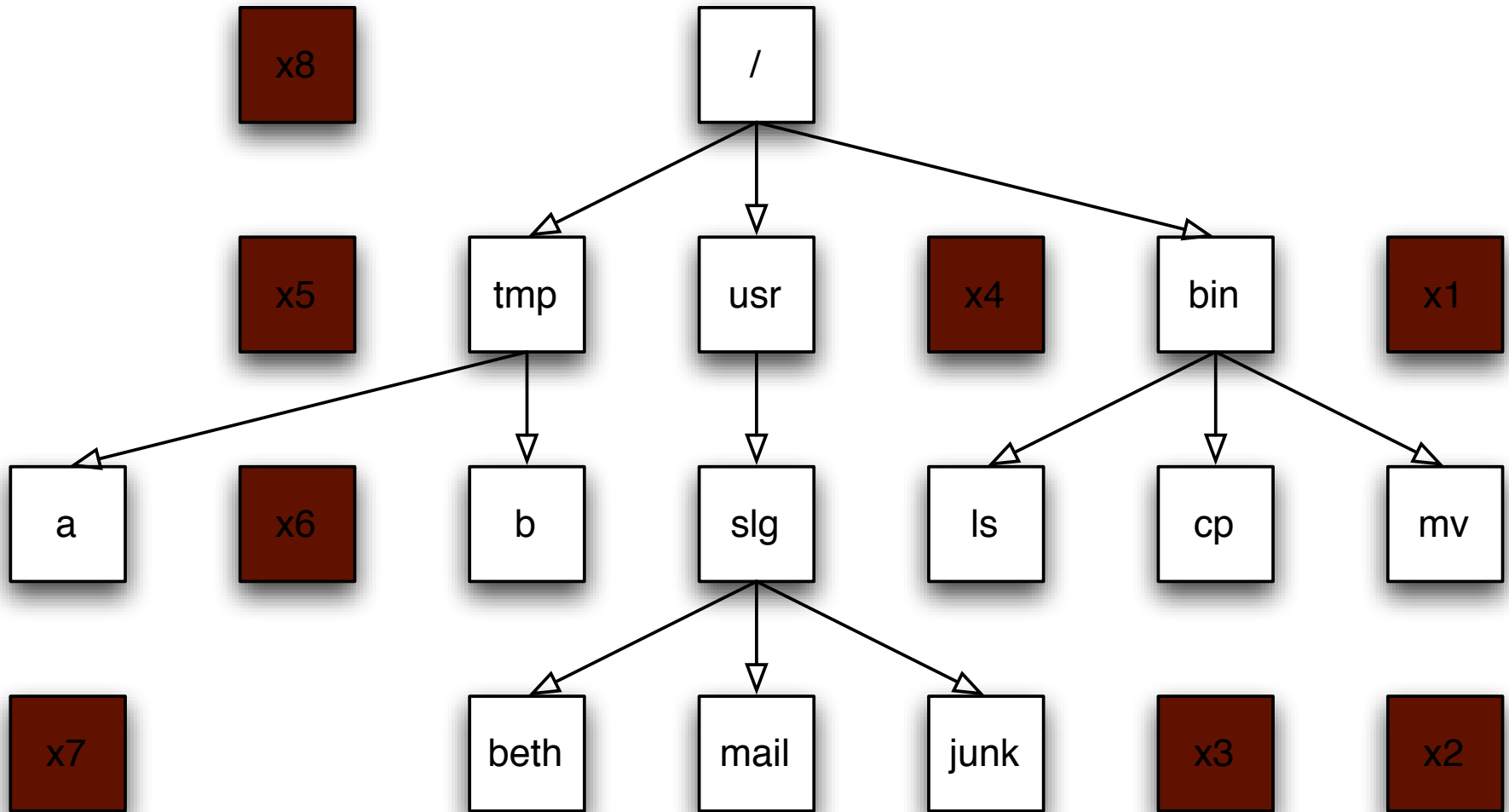
**To find that cause,
I looked *on the drives* and *contacted the data subjects*.**

Data on a hard drive is arranged in sectors.



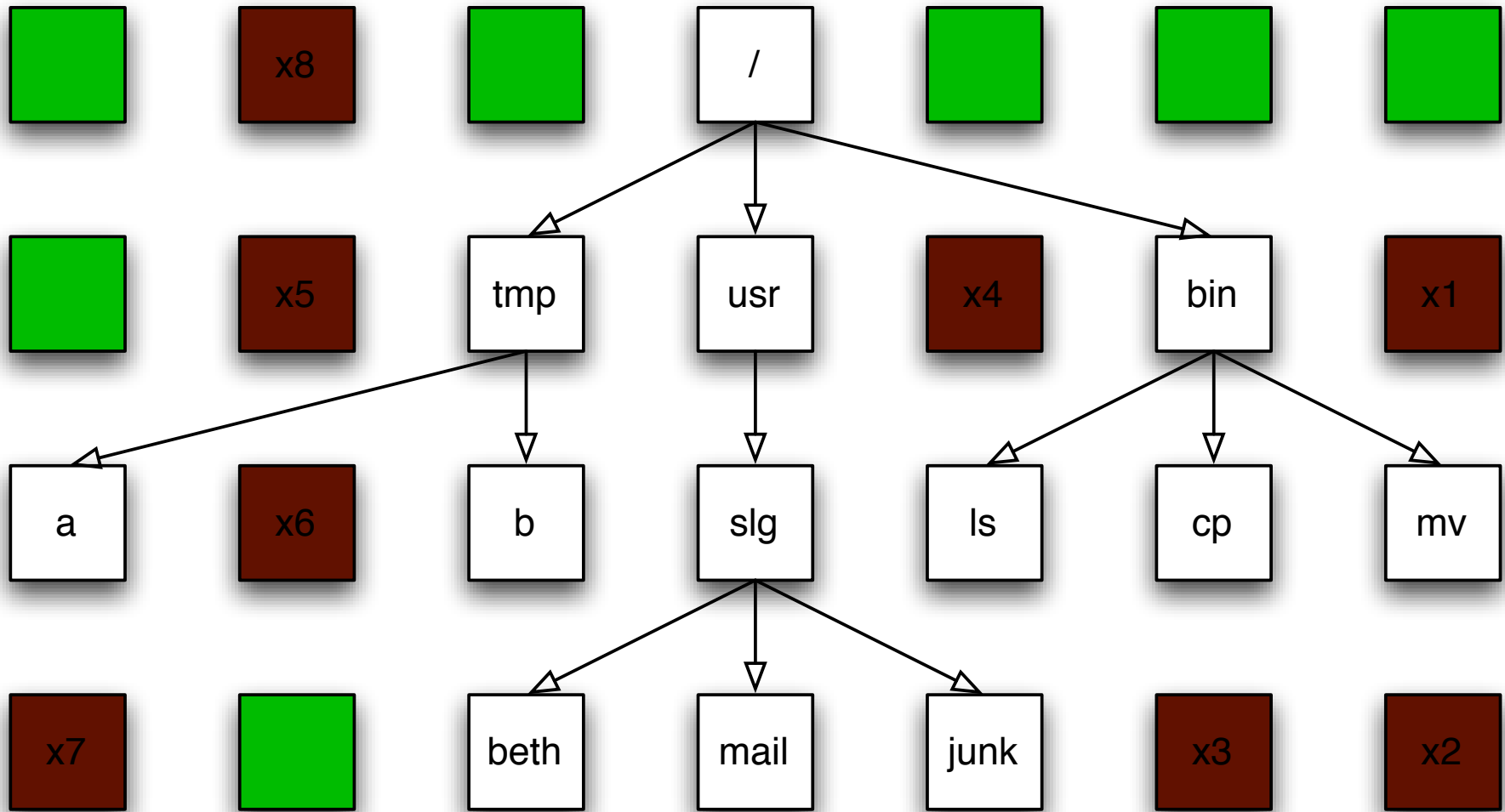
The white sectors indicate directories and files that are visible to the user.

Data on a hard drive is arranged in sectors.



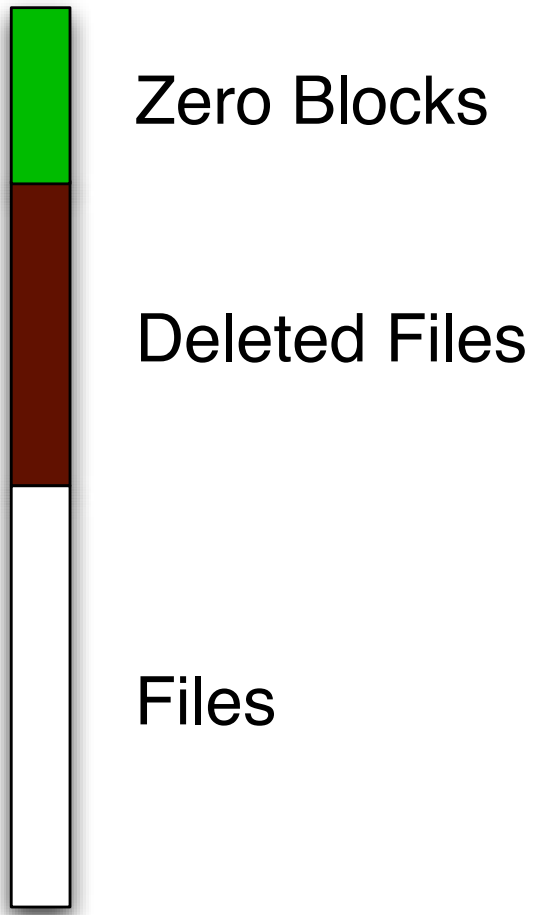
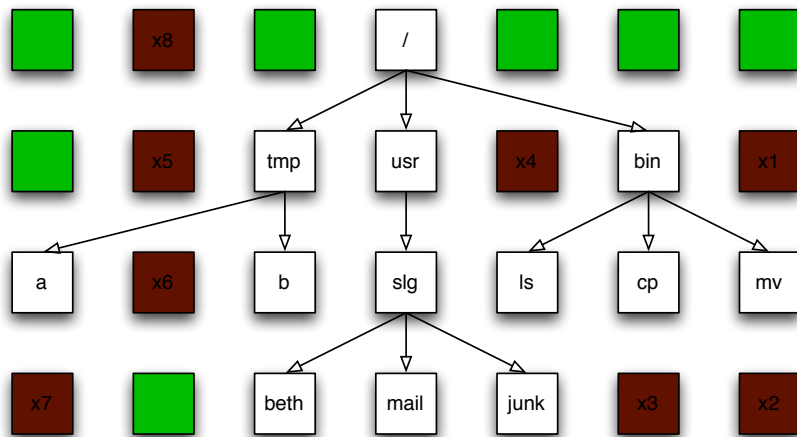
The brown sectors indicate files that were deleted.

Data on a hard drive is arranged in sectors.

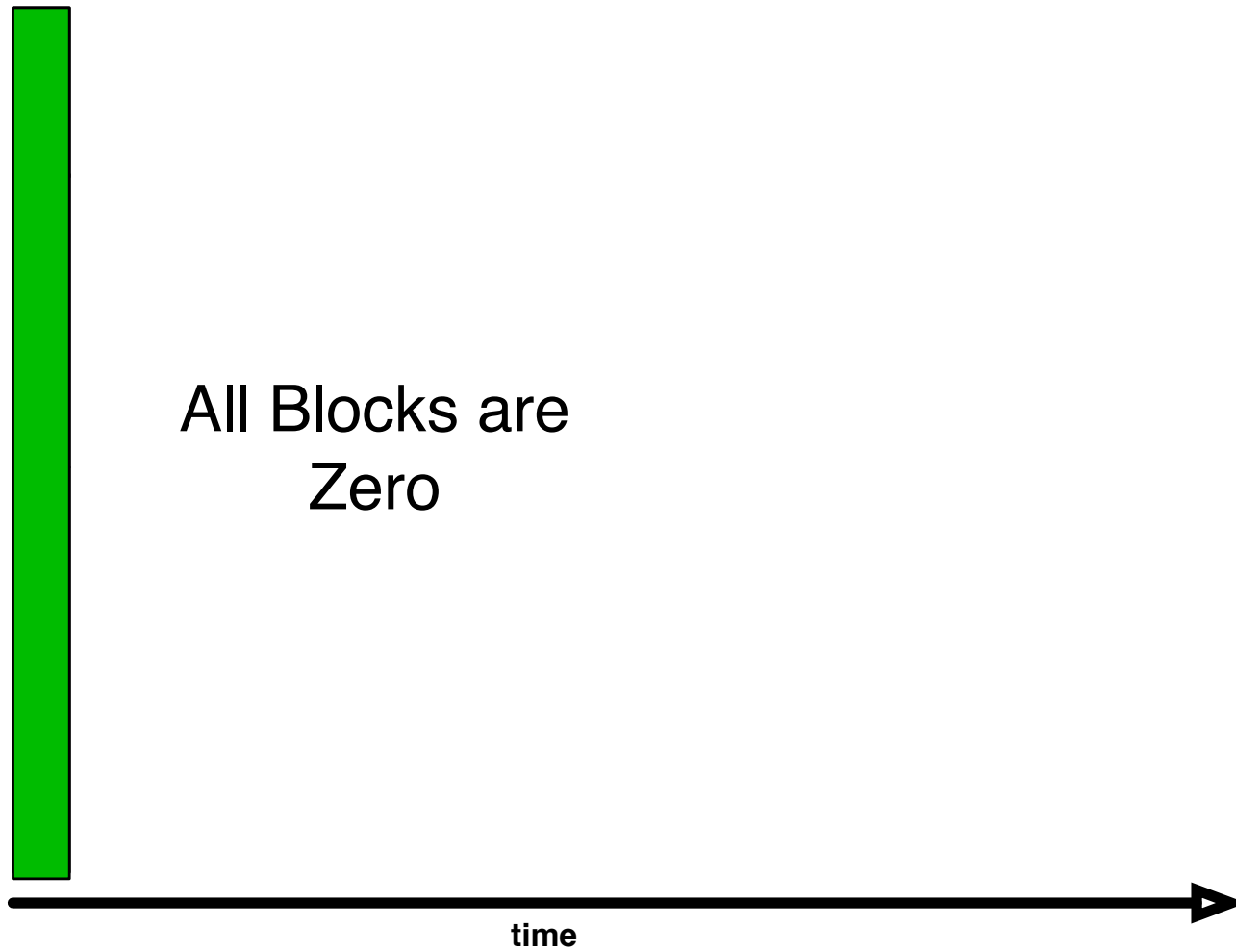


The green sectors indicate sectors that were never used (or that were wiped clean).

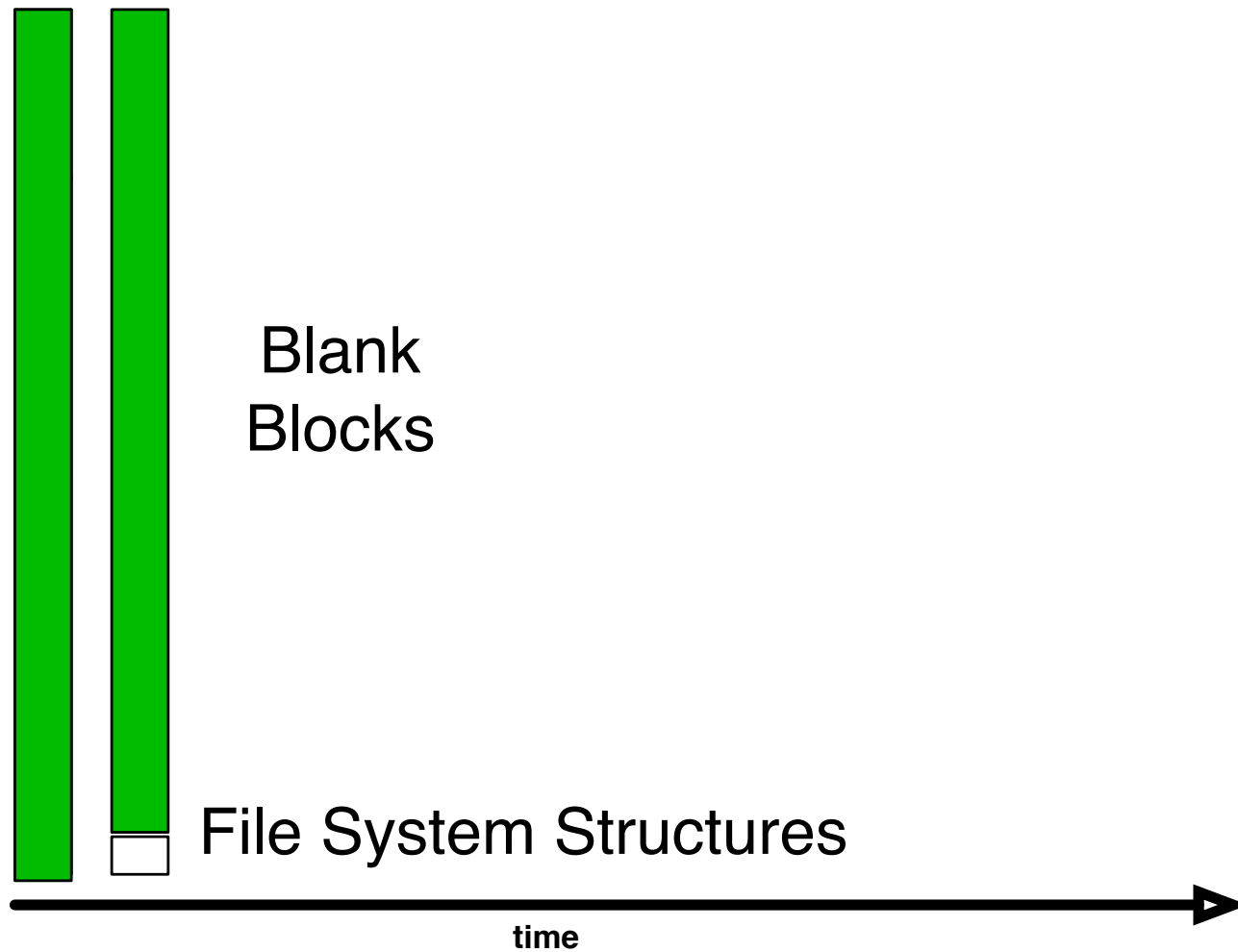
Stack the disk sectors:



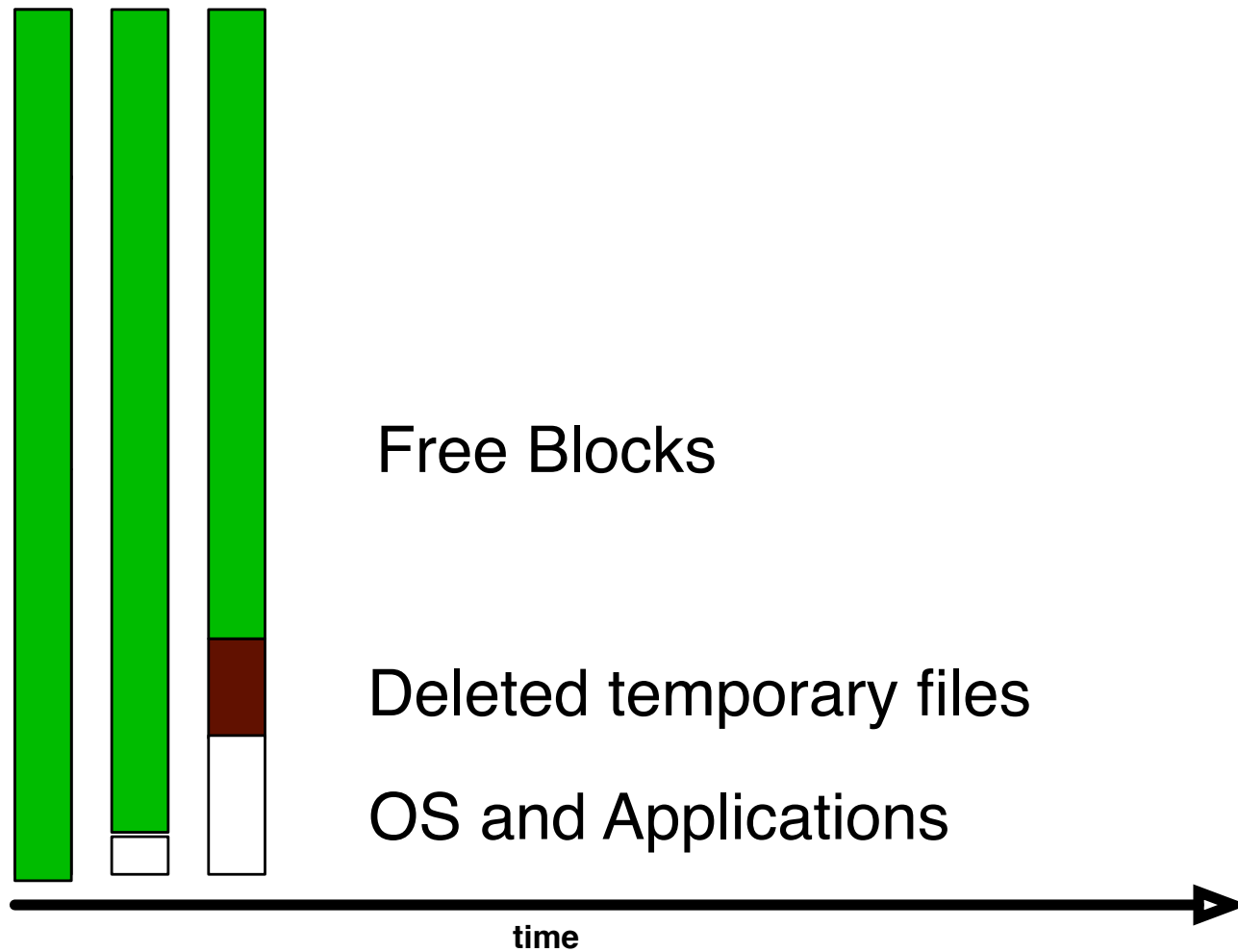
NO DATA: The disk is factory fresh.



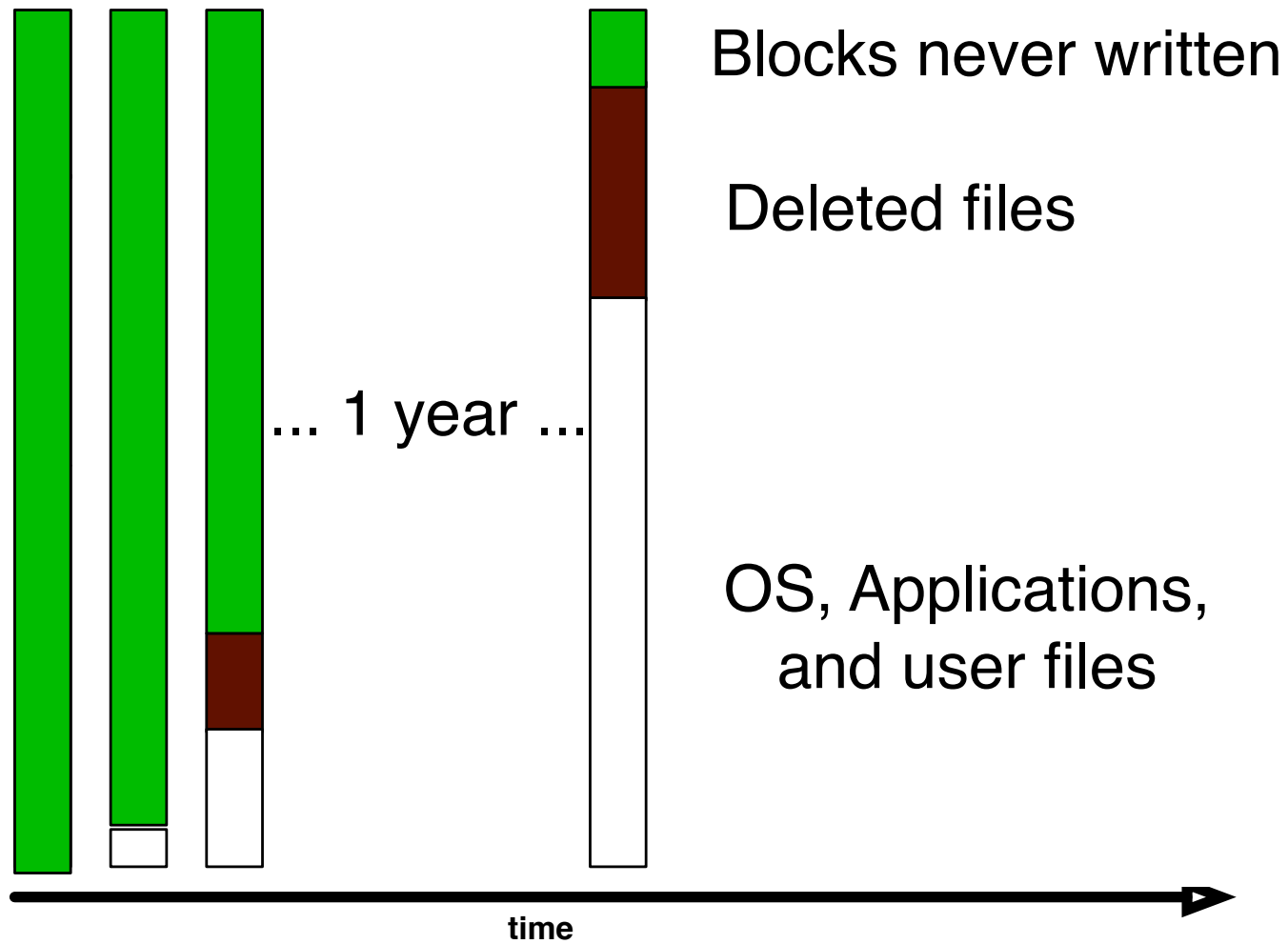
FORMATTED: The disk has an empty file system



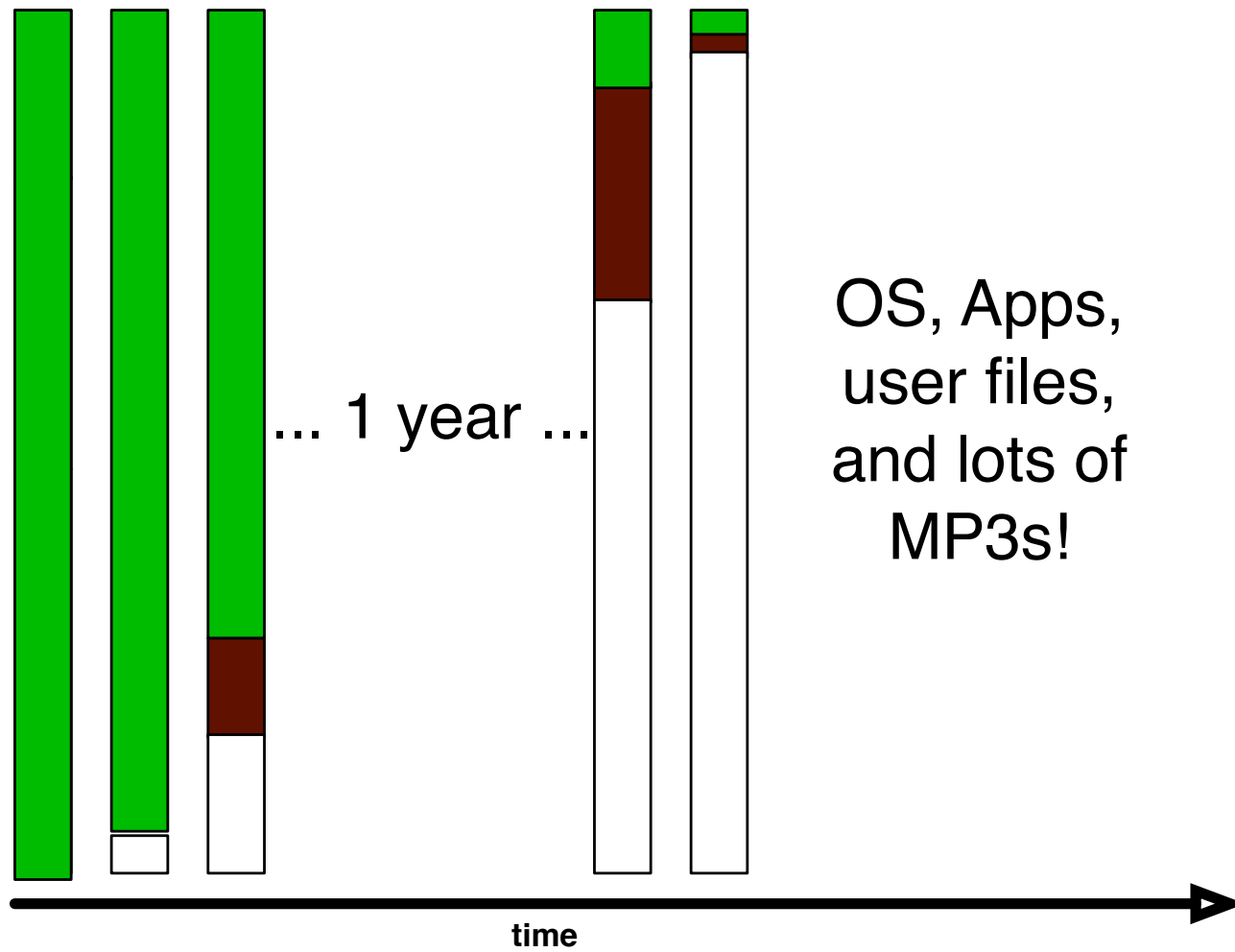
AFTER OS INSTALL: Temp. files have been deleted



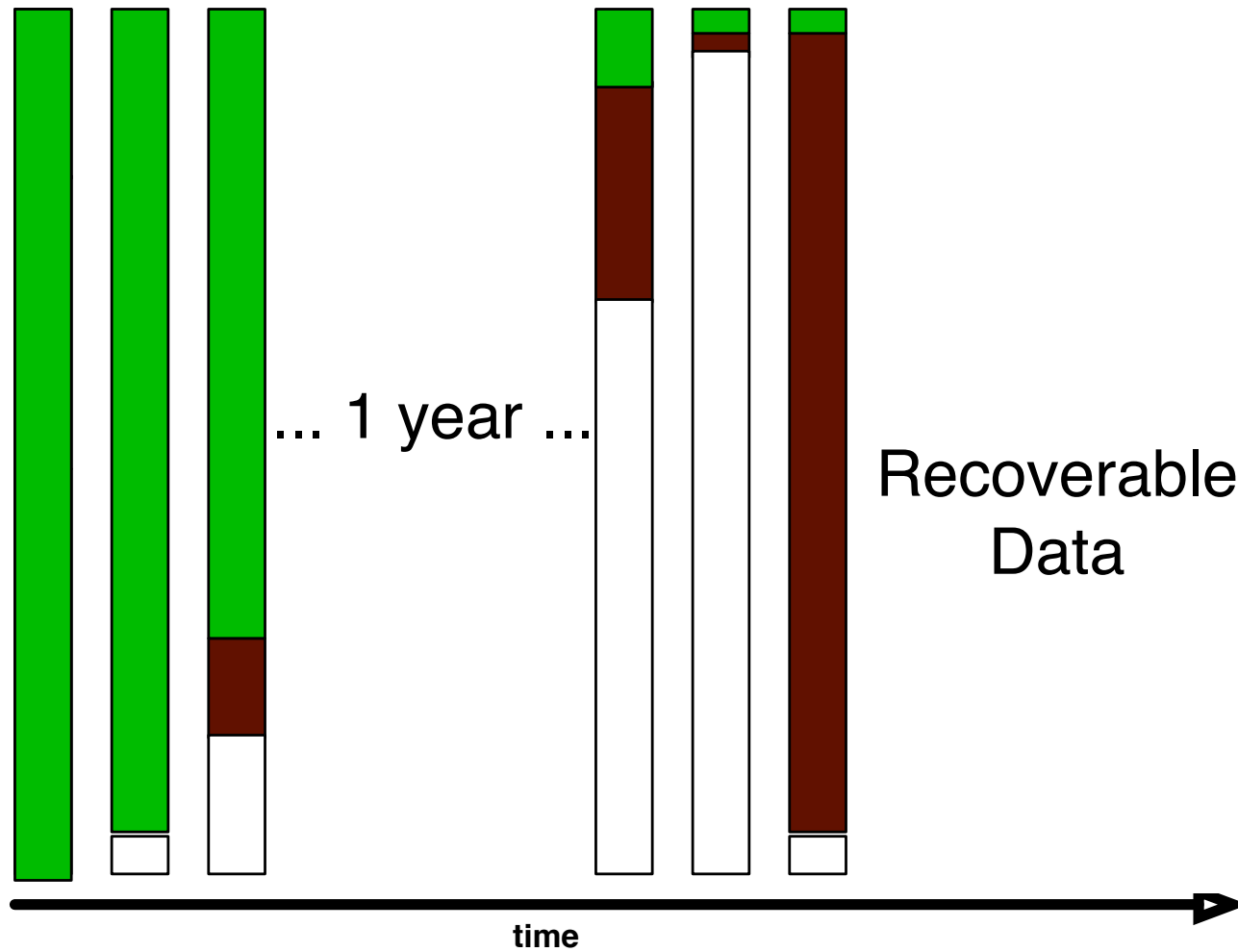
AFTER A YEAR OF SERVICE



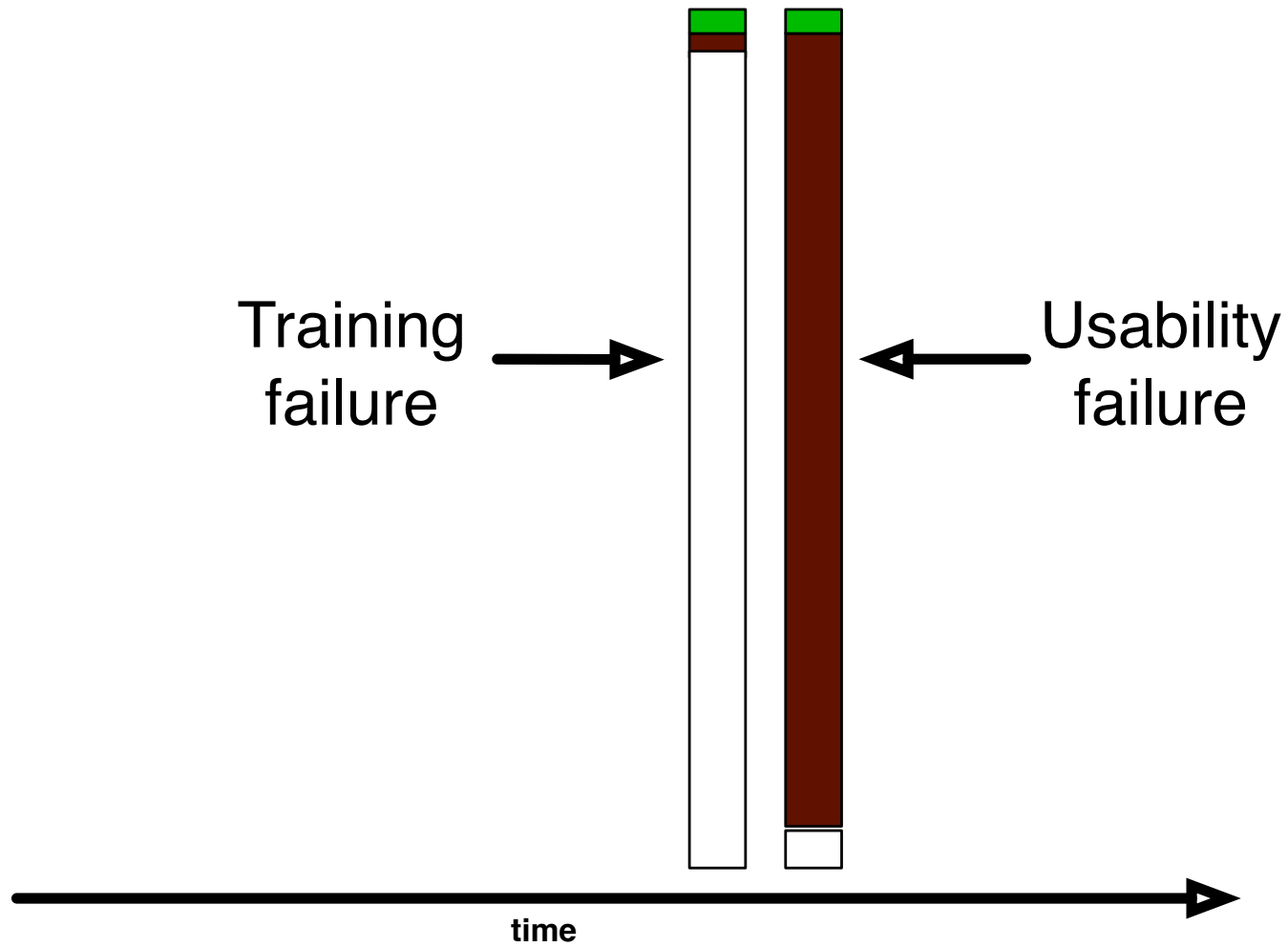
DISK NEARLY FULL!



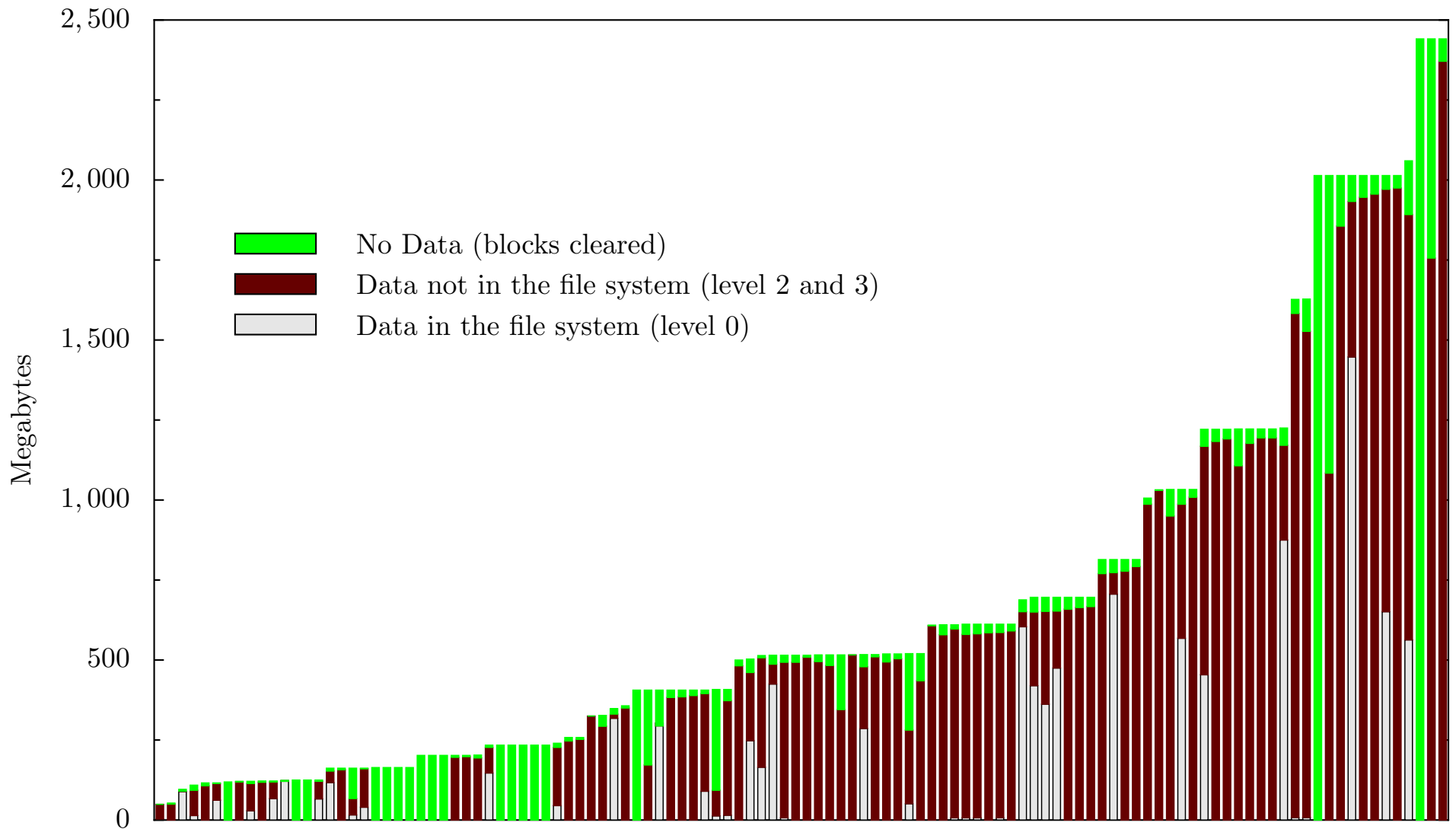
FORMAT C:\ (to sell the computer.)



We can use forensics to reconstruct motivations:



The drives are dominated by failed sanitization attempts...



..but training failures are also important.

Overall numbers

Drives Acquired:	236
Drives DOA:	60
Drives Images:	176
Drives Zeroed:	11
Drives "Clean Formatted:"	22
Total files:	168,459
Total data:	125G

Only 33 out of 176 working drives were properly cleared!

- 1 from Driveguys — but 2 others had lots of data.
- 18 from pcjunkyard — but 7 others had data.
- 1 from a VA reseller — 1 DOA; 3 dirty formats.
- 1 from an unknown source — 1 DOA, 1 dirty format.
- 1 from Mr. M. who sold his 2GB drive on eBay.

MD5 hashing allows the identification of files.

Interestingly, few unique files that had not been deleted:

File type	Unique Files
Microsoft Word files:	783
Microsoft Excel files:	184
Microsoft PowerPoint files:	30
Outlook PST files:	11
audio files:	977

Conclusion: *most users DELETED* their files before discarding their drives.

But what *really* happened?



I needed to contact the original drive owners.

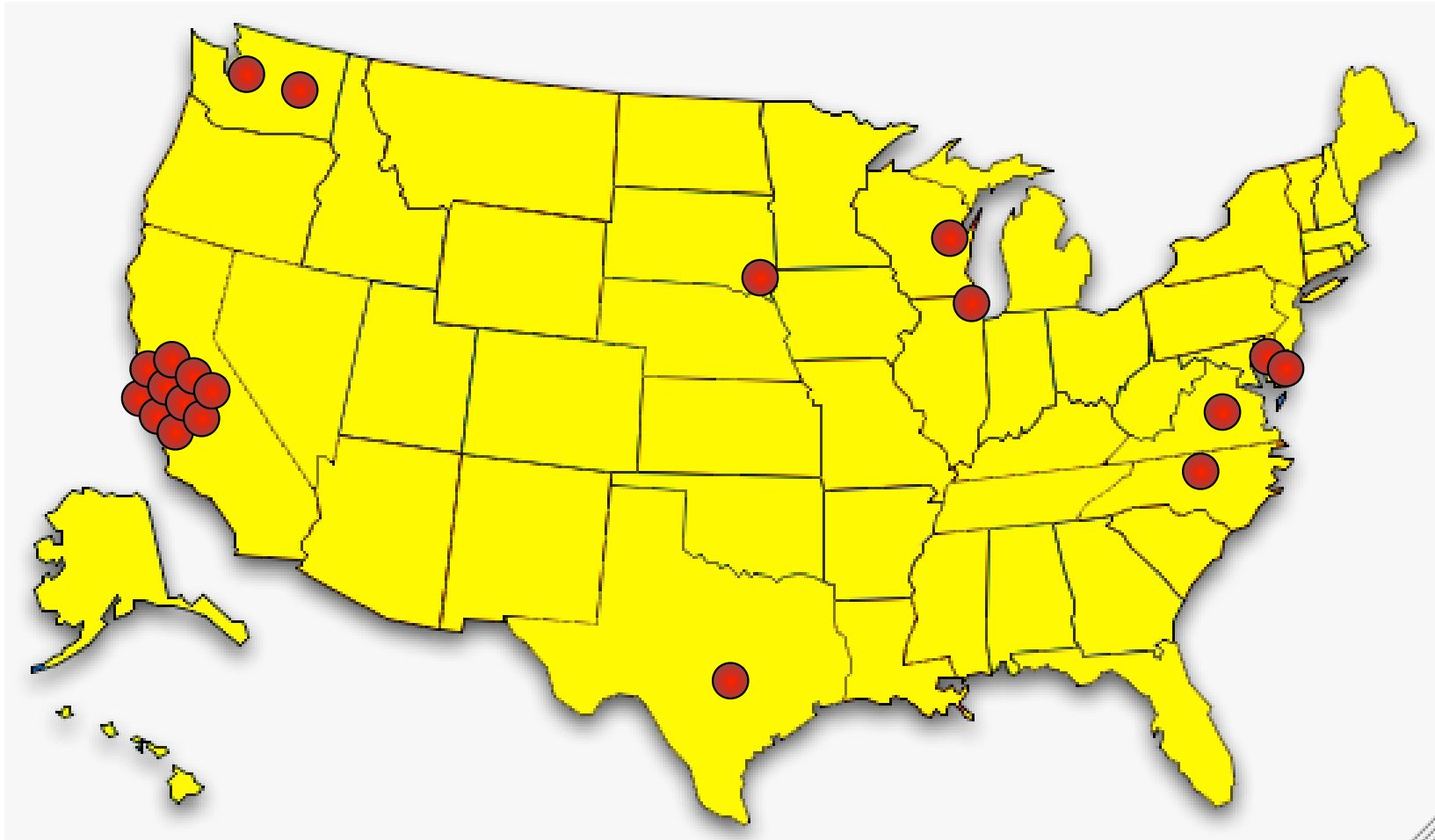
The *Remembrance of Data Passed Traceback Study.* [Garfinkel 05]

1. Find data on hard drive
2. Determine the owner
3. Get contact information for organization
4. Find the right person *inside* the organization
5. Set up interviews
6. Follow guidelines for human subjects work

```
06/19/1999 /:dir216/Four H Resume.doc
03/31/1999 /:dir216/U.M. Markets & Society.doc
08/27/1999 /:dir270/Resume-Deb.doc
03/31/1999 /:dir270/Deb-Marymount Letter.doc
03/31/1999 /:dir270/Links App. Ltr..doc
08/27/1999 /:dir270/Resume=Marymount U..doc
03/31/1999 /:dir270/NCR App. Ltr..doc
03/31/1999 /:dir270/Admissions counselor, NCR.doc
08/27/1999 /:dir270/Resume, Deb.doc
03/31/1999 /:dir270/UMUC App. Ltr..doc
03/31/1999 /:dir270/Ed. Coordinator Ltr..doc
03/31/1999 /:dir270/American College ...doc
04/01/1999 /:dir270/Am. U. Admin. Dir..doc
04/05/1999 /:dir270/IR Unknown Lab.doc
04/06/1999 /:dir270/Admit Slip for Modernism.doc
04/07/1999 /:dir270/Your Honor.doc
```

This was a lot harder than I thought it would be.

Ultimately, I contacted 20 organizations between April 2003 and April 2005.



The leading cause: betrayed trust.

Trust Failure: 5 cases

- ✓ Home computer; woman's son took to "PC Recycle"
- ✓ Community college; no procedures in place
- ✓ Church in South Dakota; administrator "kind of crazy"
- ✓ Auto dealership; consultant sold drives he "upgraded"
- ✓ Home computer, financial records; same consultant

**This specific failure wasn't considered in [GS 03];
it was the most common failure.**

Second leading cause: Poor training and supervision

Trust Failure: 5 cases

Lack of Training: 3 cases

- ✓ California electronic manufacturer
- ✓ Supermarket credit-card processing terminal
- ✓ ATM machine from a Chicago bank

Alignment between the interface and the underlying representation would overcome this problem.

Sometimes the data custodians just don't care.

Trust Failure: 5 cases

Lack of Training: 3 cases

Lack of Concern: 2 cases

- ✓ Bankrupt Internet software developer
- ✓ Layoffs at a computer magazine

Regulation on resellers might have prevented these cases.

In seven cases, no cause could be determined.

Trust Failure: 5 cases

Lack of Training: 3 cases

Lack of Concern: 2 cases

Unknown Reason: 7 cases

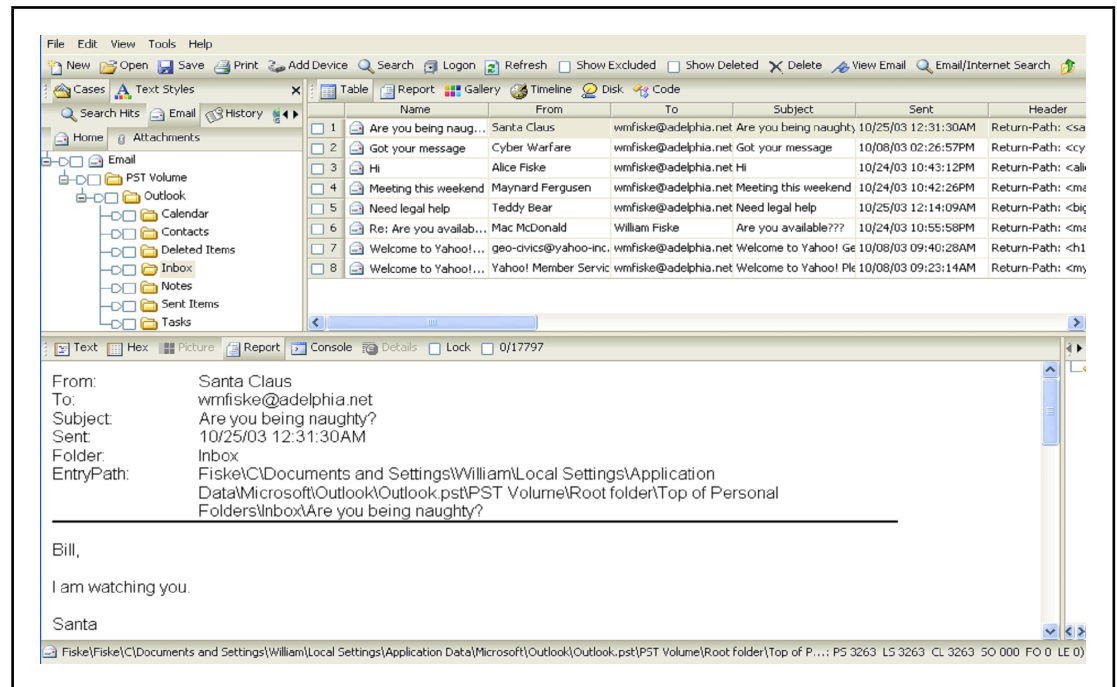
- ✘ Bankrupt biotech startup**
- ✘ Another major electronics manufacturer**
- ✘ Primary school principal's office**
- ✘ Mail order pharmacy**
- ✘ Major telecommunications provider**
- ✘ Minnesota food company**
- ✘ State Corporation Commission**

Regulation might have helped here, too.

The techniques developed for [Garfinkel '05] are different than traditional forensics techniques.

Traditional forensics tools:

- Interactive user interface.
- Recovery of “deleted” files.
- Generation of “investigative reports” for courtroom use.
- Focus on one or a few disks.



Today's tools choke when confronted with thousands of disks.

- Has this drive been previously imaged?
- Which drives belong to my target?
- Do any drives belong to my target's associates?
- Where should I start?



**Today's tools are for criminal investigations.
Increasingly, we need tools for intelligence analysis.**

Intelligence objectives can be advanced by correlating information from multiple drives.

- Where any drives were used by the same organization?
- What names/places/email addresses are in common?
- Which drives were used in a place or at a time of interest?



I call this approach “Cross-Drive Forensics.”

“First Order Cross-Drive Forensics” analyzes each drive with a filter.

The drives with the highest response are worthy of further attention.

Example: The Credit Card Number Detector.

The CCN detector scans bulk data for ASCII patterns that look like credit card numbers.

- CCNs are found in certain typographical patterns.
(e.g. XXXX-XXXX-XXXX-XXXX
or XXXX XXXX XXXX XXXX
or XXXXXXXXXXXXXXXXXXXX)
- CCNs are issued with well-known prefixes.
- CCNs follow the Credit Card Validation algorithm.
- Certain numeric patterns are unlikely.
(e.g. 4454-4766-7667-6672)

CCN detector: written in flex and C++

Scan of disk #105: (642MB)

Test	# pass
typographic pattern	3857
known prefixes	90
CCV1	43
numeric histogram	38

Sample output:

```
'CHASE NA|5422-4128-3008-3685| pos=13152133
'DISCOVER|6011-0052-8056-4504| pos=13152440
.'GE CARD|4055-9000-0378-1959| pos=13152589
BANK ONE |4332-2213-0038-0832| pos=13152740
.'NORWEST|4829-0000-4102-9233| pos=13153182
'SNB CARD|5419-7213-0101-3624| pos=13153332
```

Even with the tests, there are occasional false positives.

CCN scan of Disk #115: (772MB)

Test	# pass
pattern	9196
known prefixes	898
CCV1	29
patterns	27
histogram	13

```
.....@: |44444486666108| :<@<74444:@@@<<44 pos=82473275
.....#"&'&&' |445447667667667| ..050014&'4"1"&' . pos=86493675
.....221267241667&|454676676654450|&566746566726322. pos=86507818
3..30210212676677.. |30232676630232| .1.....001.01 pos=86516059
"&#&&'&41&&'645445&|454454672676632| .3.....0.. pos=86523223
.....".#"#"#&' |445467667227023| .....366 pos=87540819
D#9?.32400.,,+14%?B|499745255278101|*02)46+;<17756669 pos=118912826
.GGJJB...>.JJGG...G|3534554333511116| .....6 pos=197711868
5.....}}}}}}..... |44444322233345| .....}}}}}}..... pos=228610295
)6"! ) .&*%,,%-0)07. |373484553420378|<67<038+.5(+0+.3. pos=638491849
)6"! ) .&*%,,%-0)07. |373484553420378|<67<038+.5(+0+.3. pos=645913801
```

Results of scanning 2003 corpus with CCN scanner:

Total number of image files: 178
Number of CCNs found: 47,771
Total number of distinct cards: 15,613
Most popular CCN 6404 6521 6029 6650
(Seen 34 times on 30 drives)

Context analysis shows this is not a valid CCN:

```
[6] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[7] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[8] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[10] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[11] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[11] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[15] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[18] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[18] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[24] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
[25] 6213 1 6758 6367 .. | 6404 6521 6029 6650 | v 6025 6646 1 -138
```

A “stop list” can be used for these common number.

Ignore “6404 6521 6029 6650’ and we repeat the experiment:

Total number of image files:	178
Number of CCNs found:	47,737 (was 47,771)
Total number of distinct cards	15,612 (was 15,613)
New “most popular CCN”	5501 8501 3501 3705

(Seen 35 times on 27 drives)

Once again, this does not appear to be a valid CCN:

```
[14] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[112] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[121] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[128] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[133] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[181] 3201 4901 : |5501 8501 3501 3705| 5102....yes.%d\0ff
[182] 3201 4901 : |5501 8501 3501 3705| 5102 13505....yes.
[184] 3201 4901 : |5501 8501 3501 3705| 5102 13505....yes.
[186] 3201 4901 : |5501 8501 3501 3705| 5102 13505....yes.
```

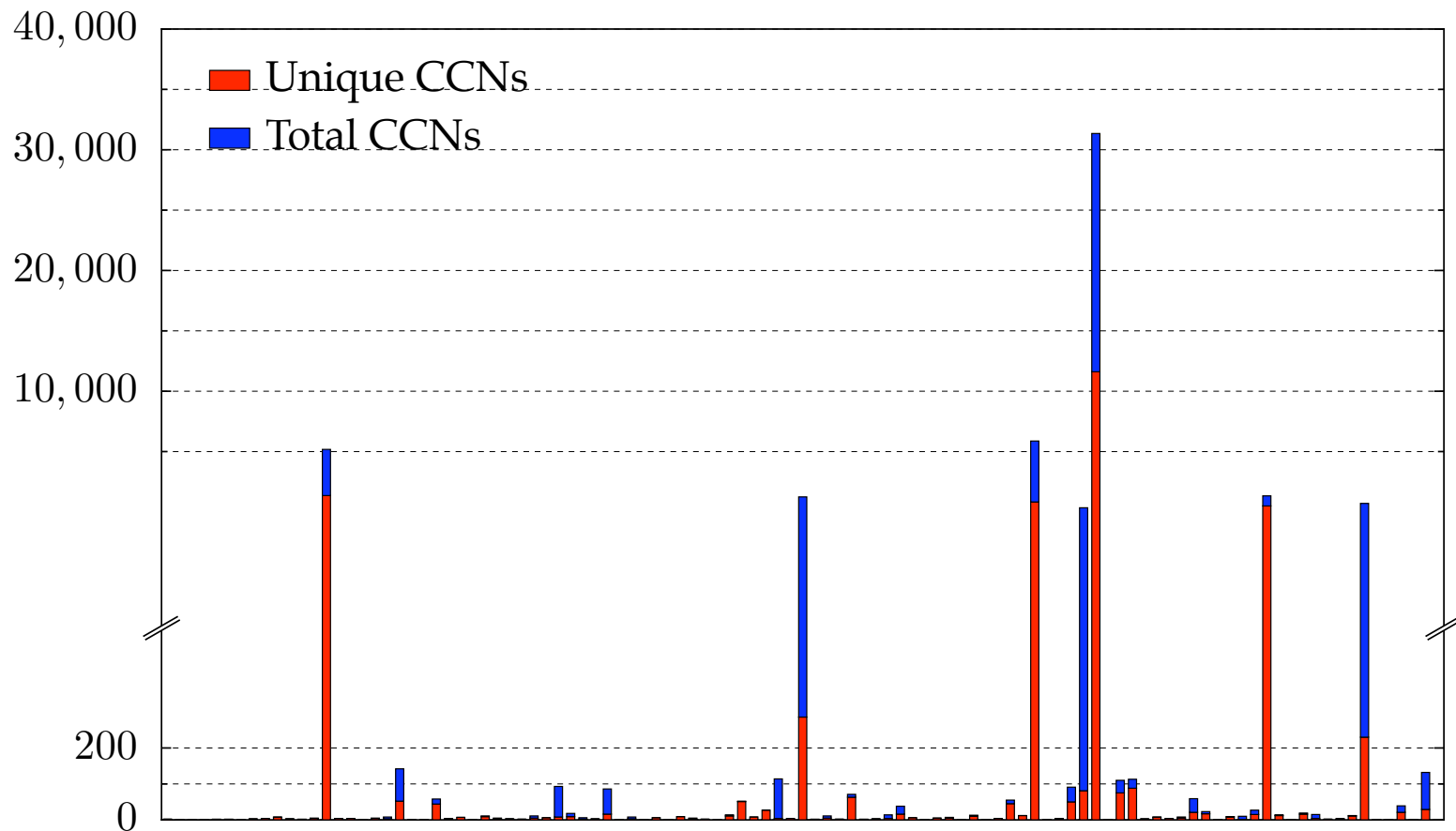
There are several problems with the “stop list” approach:

The list must be:

- Constructed.
- Maintained.
- Tuned for different applications.

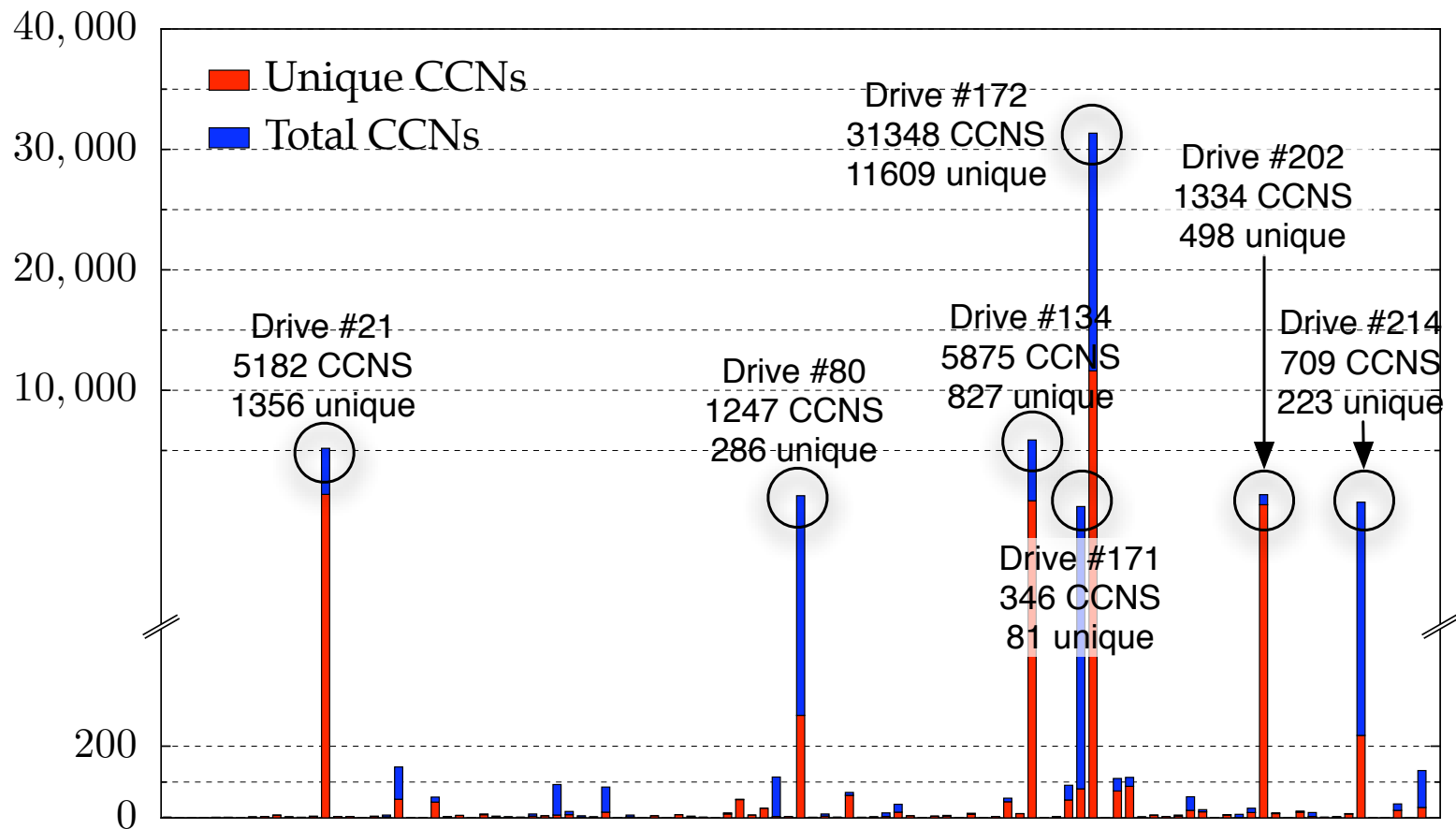
Building a “stop list” requires judgement and patience.

An alternative is to assume that “false positives” are rare and focus on those drives with high response.



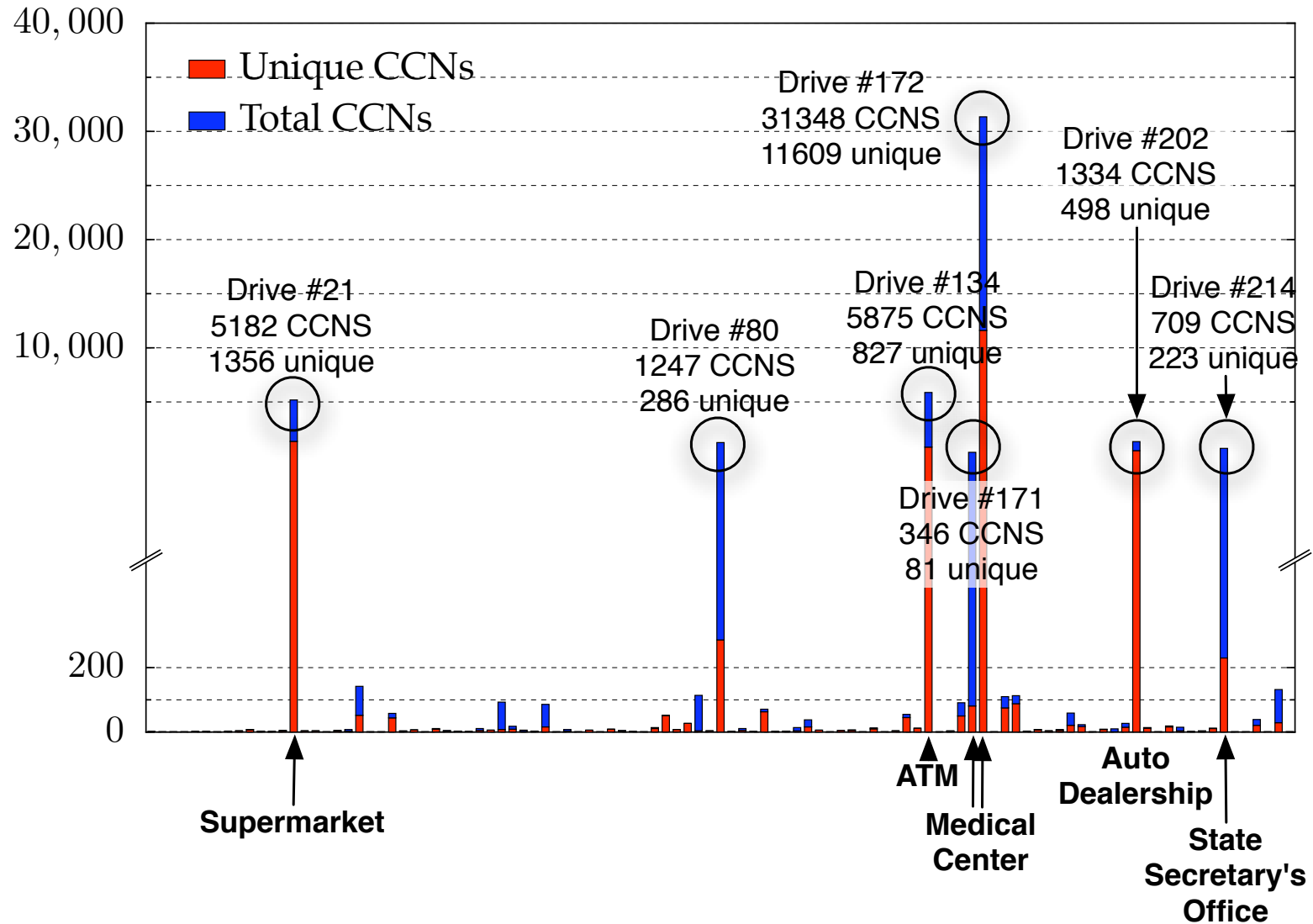
By definition, no drive should contain a large number of CCNs, so these drives are all interesting.

An alternative is to assume that “false positives” are rare and focus on those drives with high response.

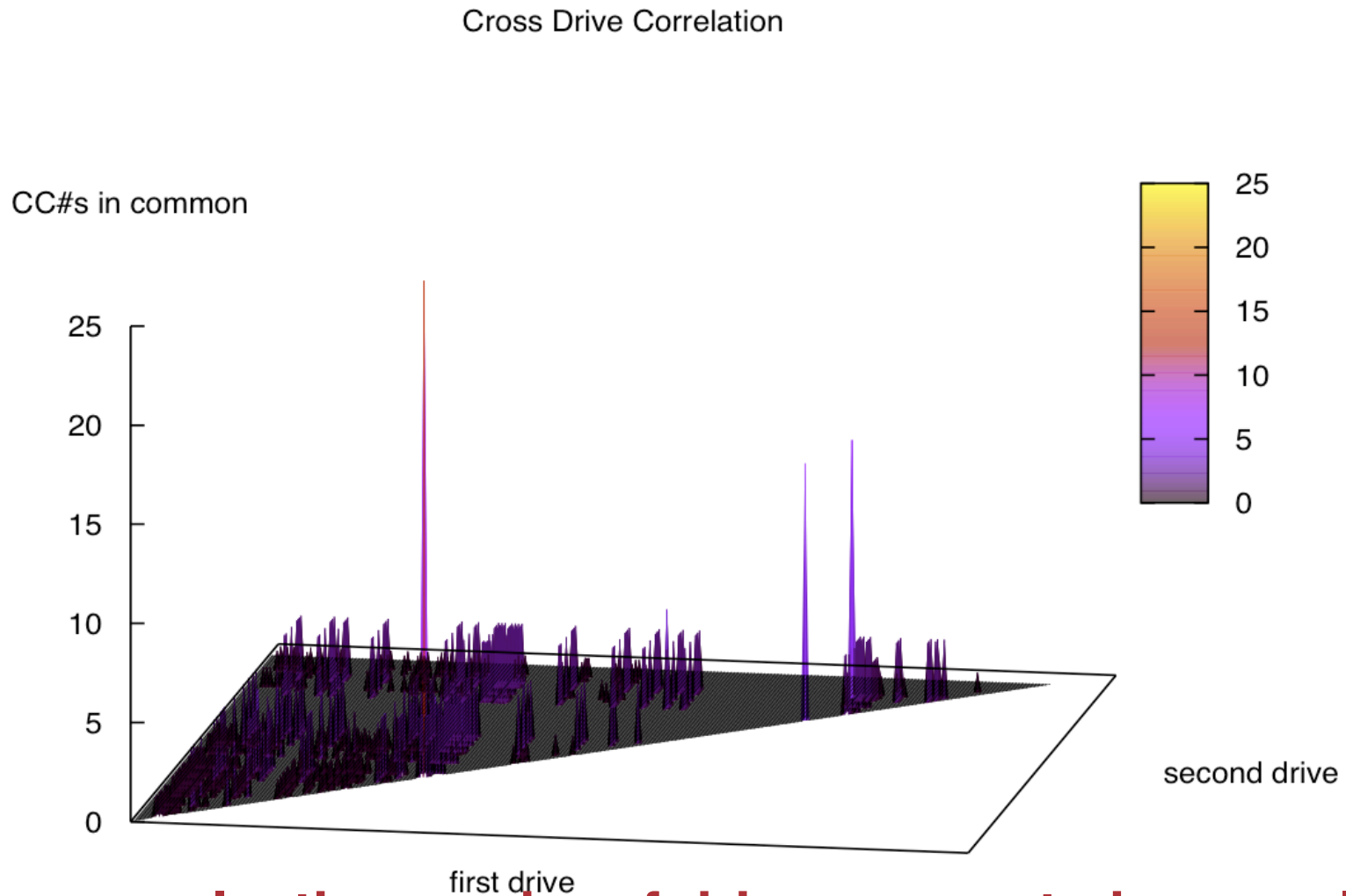


Only 7 drives had more than 300 credit card numbers.

Several of these drives were traced back to their original owners.



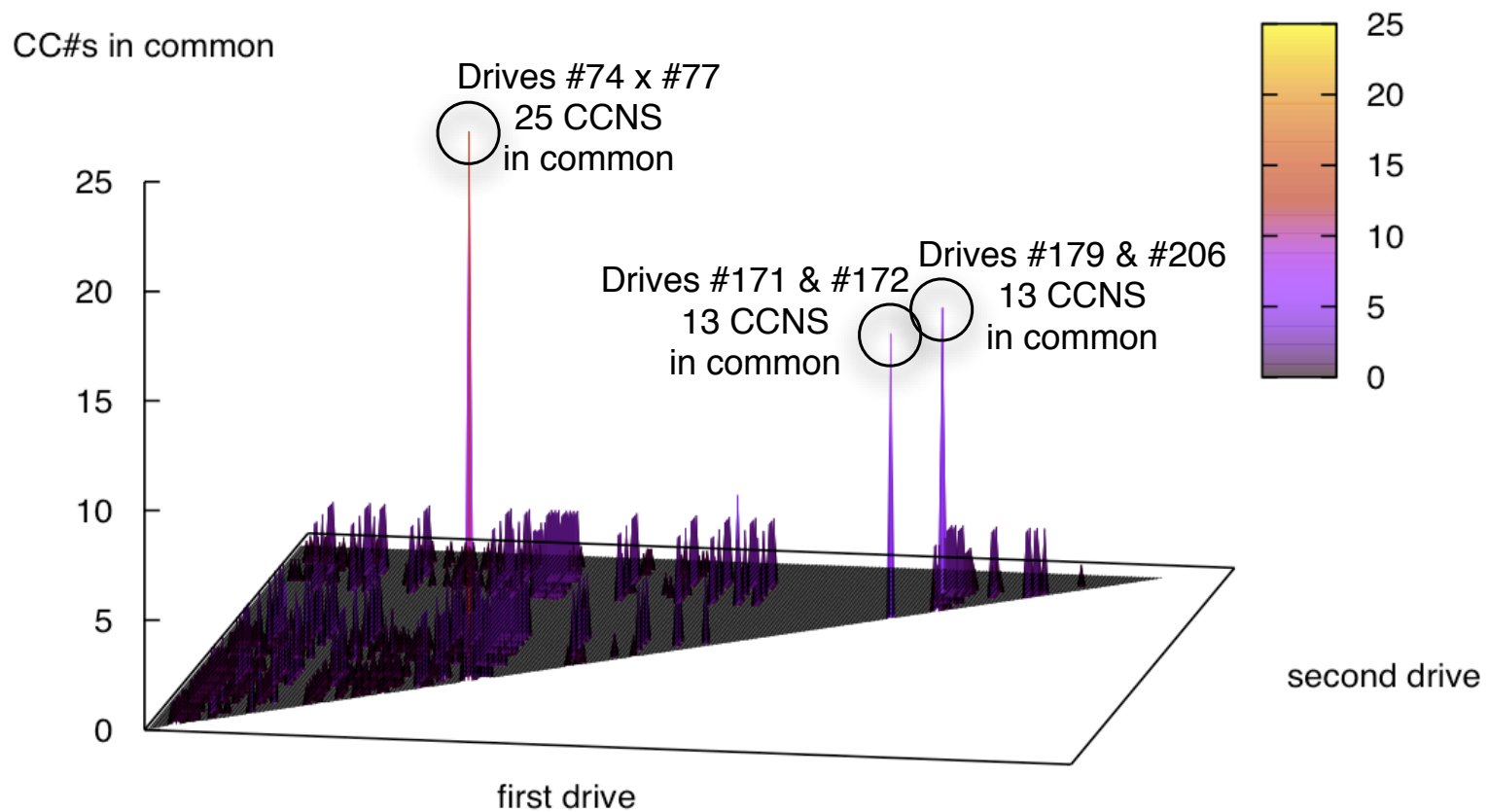
Second-order analysis uses correlation techniques to identify drives of interest.



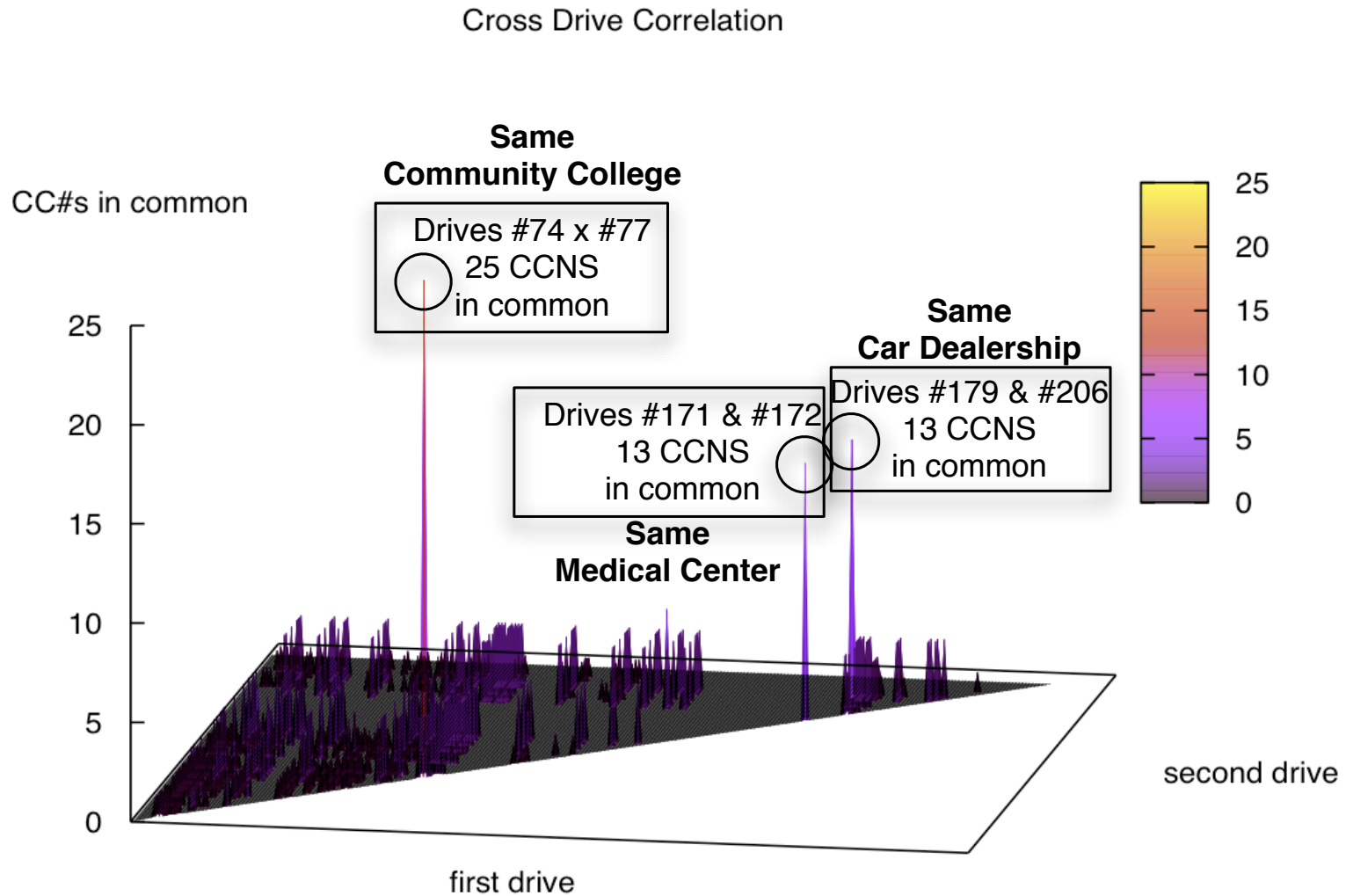
In this example, three pairs of drive appear to be correlated.

Second-order analysis uses correlation techniques to identify drives of interest.

Cross Drive Correlation



Manual analysis of on-drive data reveals that these drives are from the same organization.



Second-order applications:

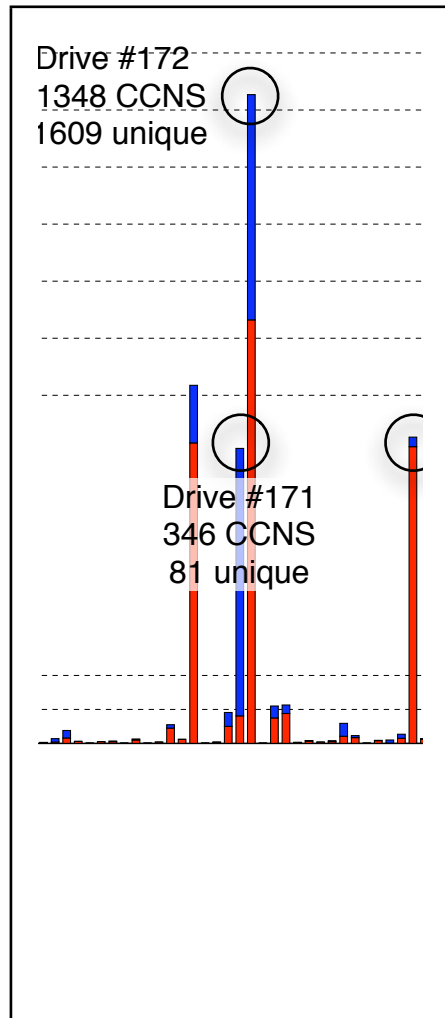
Possible Identifiers:

- CCNs
- Email addresses
- Message-IDs
- MD5 of disk sectors

Possible Uses:

- Identifying new social networks
- Testing for inclusion in an existing network.
- Measuring dissemination of information

Let's look at drives #171 and #172 again.



Cross-drive analysis tells us that #171 and #172 are from the same medical center.

Drive #171: Development drive

- Has source code.
- 346 CCNS; 81 unique.

Drive #172: Production system.

- 31,348 CCNS; 11,609 unique
- Oracle database (hard to reconstruct).

Evidently, the programmers used live data to test their system.

Legislative reactions to this research:

“Fair and Accurate Credit Transactions Act of 2003” (US)

- Introduced in July 2003. Signed December 2003.
- Regulations adopted in 2004, effective June 2005.
- Amends the FCRA to standardize consumer reports.
- Requires destruction of paper or electronic “consumer records.”

Testimony: <http://tinyurl.com/cd2my>

Technical reactions to this research: “Secure Empty Trash” in MacOS 10.3.

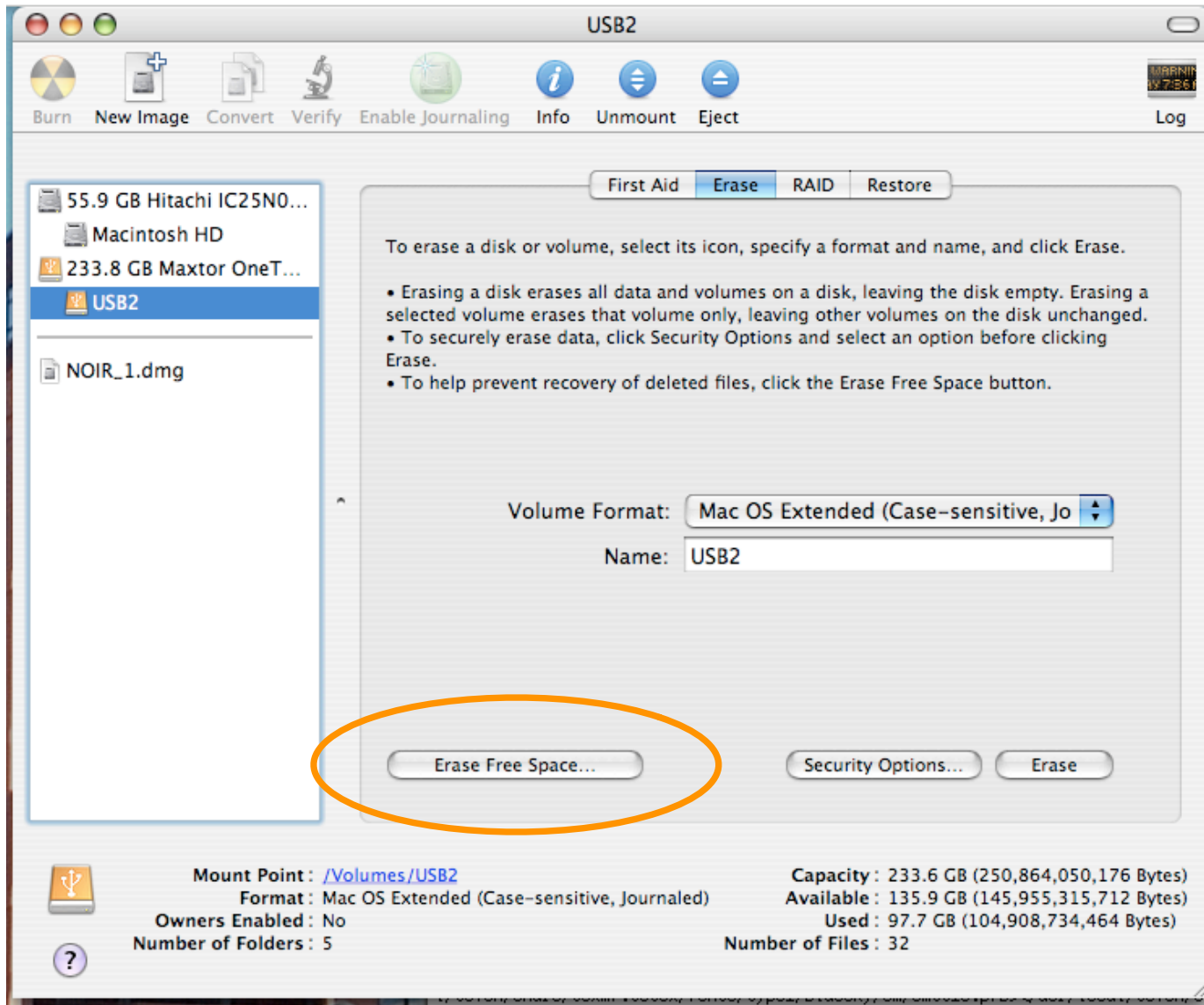


Unfortunately, “Secure Empty Trash” is incomplete.

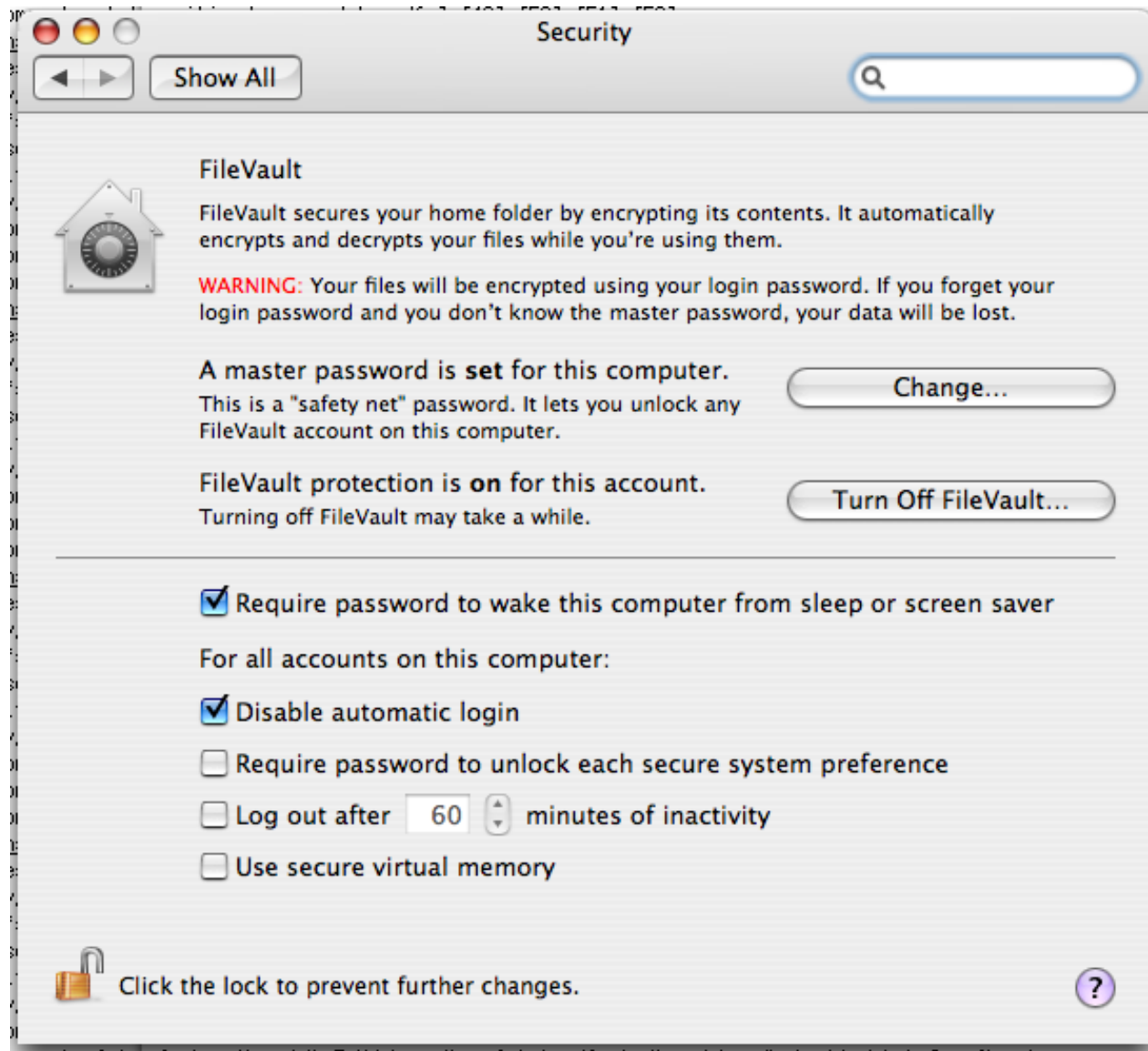
- Implemented in Finder (inconsistently)
- Locks trash can
- Can't change your mind



MacOS 10.4 “Erase Free Space” makes a big file.



MacOS “File Vault” gives users an encrypted file system.



Future Work: Deploying Compete Delete

- Make FORMAT actually erase the disk.
- Make “Empty Trash” actually overwrite data.
- Integrate this functionality with web browsers, word processors, operating systems.
- Address usability dangers of clean delete.
- Analysis of “one big file” technique.

Let's put this in Linux!

Future Work: 2500 Drive Corpus

- Automated construction of stop-lists.
- Detailed analysis of false positives/negatives in CCN test.
- Explore identifiers other than CCNs.
- Support for languages other than English.

More than 500 drives are standing by...

Future Work: Toolkit

- Easy-to-use, reliable, disk imaging software.
- New file format for disk images.
- Web-based database of hash codes.

Initial version is available for download.

Future Work: Economics and Society

- Who is buying used hard drives and why?
- Hard drive honeypot.
- Compliance with FACT-A

This is a lot of work...

Future Work: Summary

- Improved cross-drive forensics
- 2500 Drive Corpus
- Open-Source Toolkits
- Economics and Society

Questions?