The Drives Project
From Disk Forensics to Media Exploitation

Brooklyn Polytechnic
October 1, 2007

Simson L. Garfinkel
http://www.simson.net/

The Drives Project:
Research with other people's data.

I bought 10 used computers from an electronics store in 1998.

Computer #1 had been used as the file server for a law firm.
It contained client confidential information.

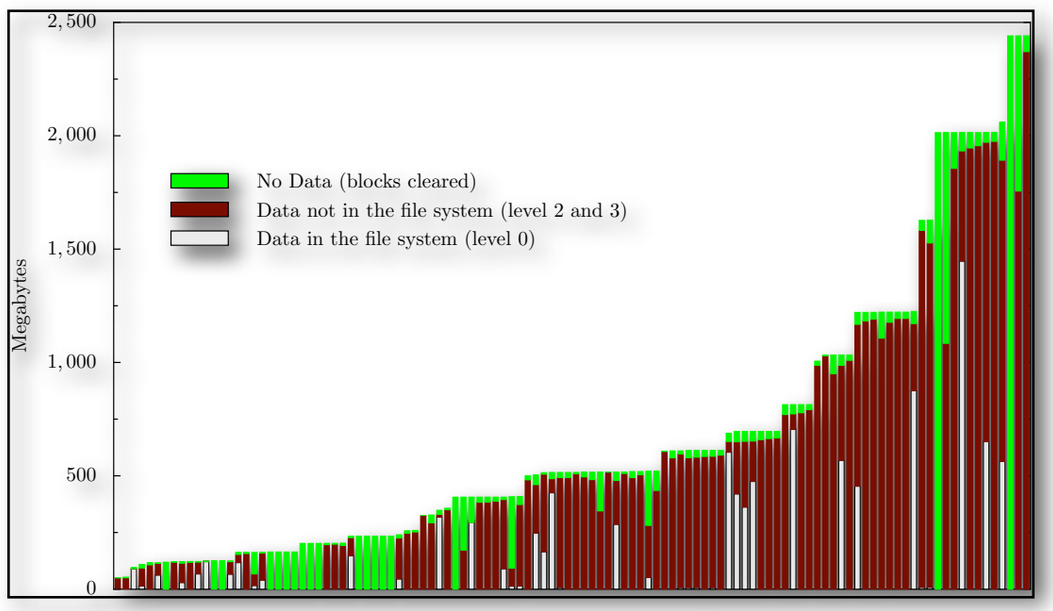
The original goal of the Drives Project: Documenting media sanitization failures.

We analyzed 150 hard drives purchased on the secondary market between 1998 and 2002.

We found confidential *residual data* on roughly 1/3 of the drives:

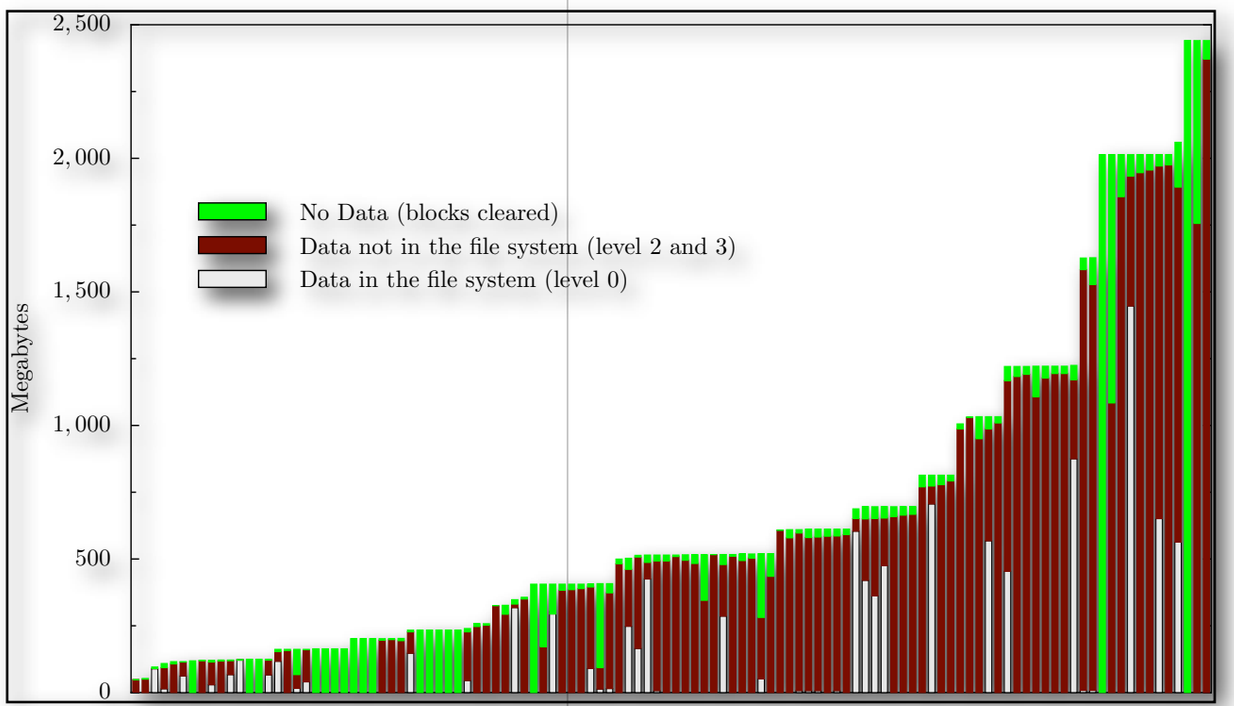
- Thousands of credit cards
- Financial records
- Medical information
- Trade secrets
- Highly personal information

[Garfinkel & Shelat 03]
[Garfinkel 05]



To quickly find the “good stuff,” I developed a range of automated techniques.

- Automated imaging
- Batch processing
- Rapid identification of “hot drives”

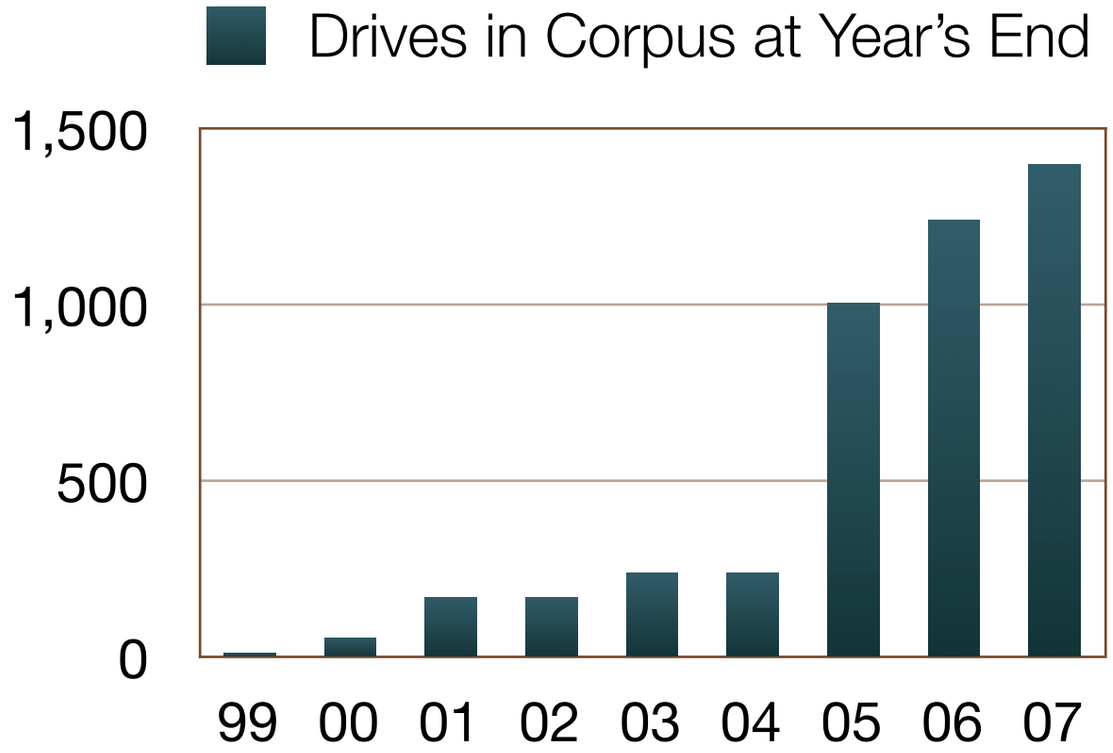


All of these techniques have applications beyond my original research.

Looking for residual data was just the start.

As I expanded the collection to 250 hard drives,
I discovered many research opportunities:

- Practitioners need better tools for working with disk images and electronic evidence.
- Developers lack real data to create and validate their tools.
- There are few (if any) algorithms for performing automated analysis.



These are now my areas of research.

Originally I did “computer forensics.” Now I do “media exploitation.”

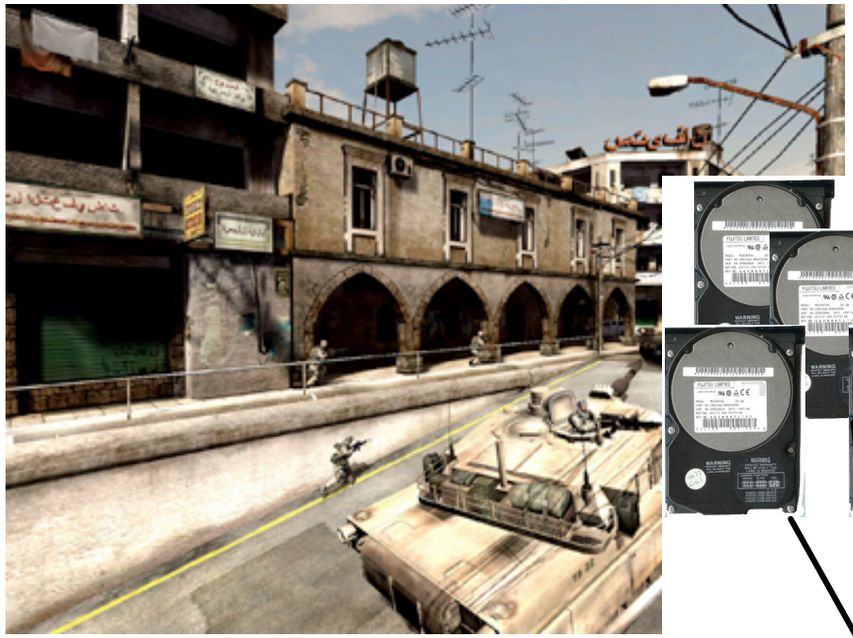
Forensics implies scientifically-validated techniques for preparing courtroom evidence.

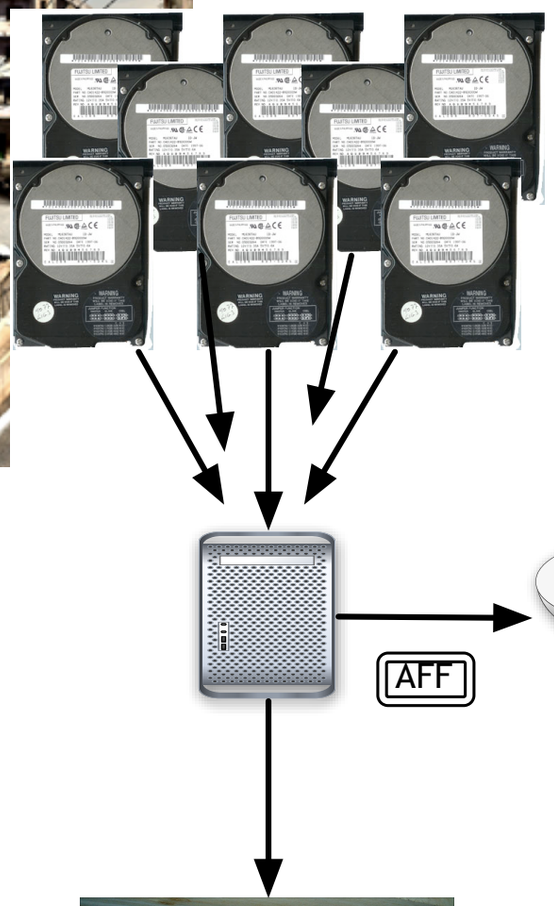
Most of my work is aimed at *exploiting* information from large numbers of drives:

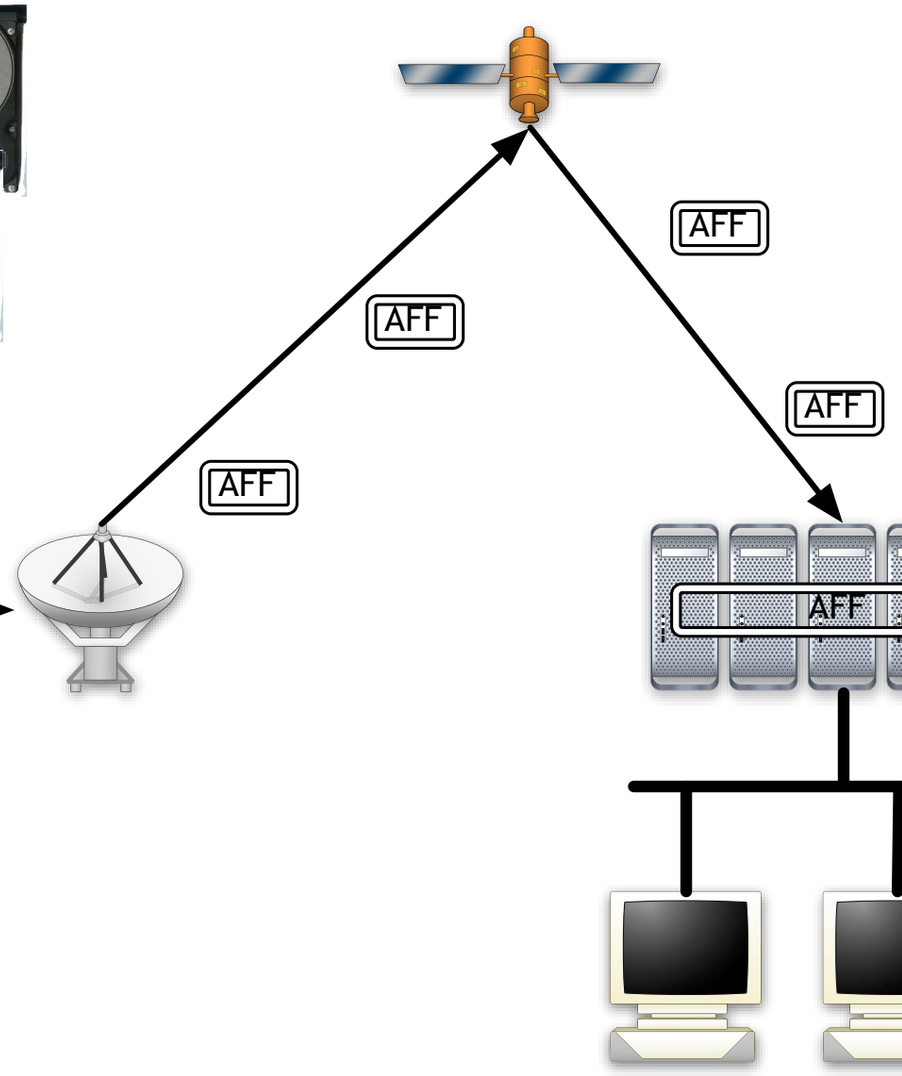
- Rapidly finding useful information.
- Creating actionable intelligence.
- Knitting together a large-scale picture.
- Using data from many drives to create smarter single-drive tools.

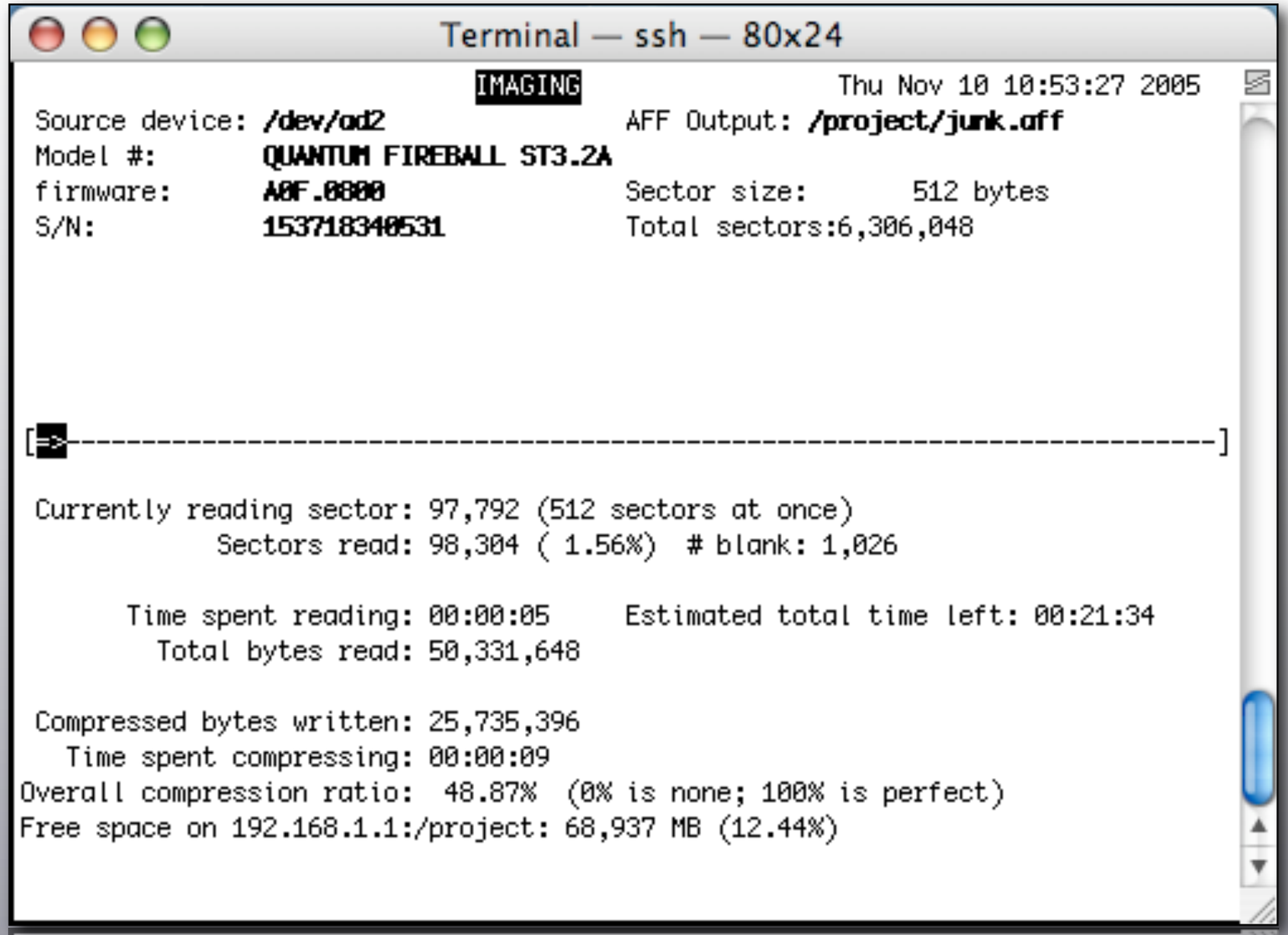
This is also useful for law enforcement and in court, but that’s not my primary goal.

My DOMEX vision:





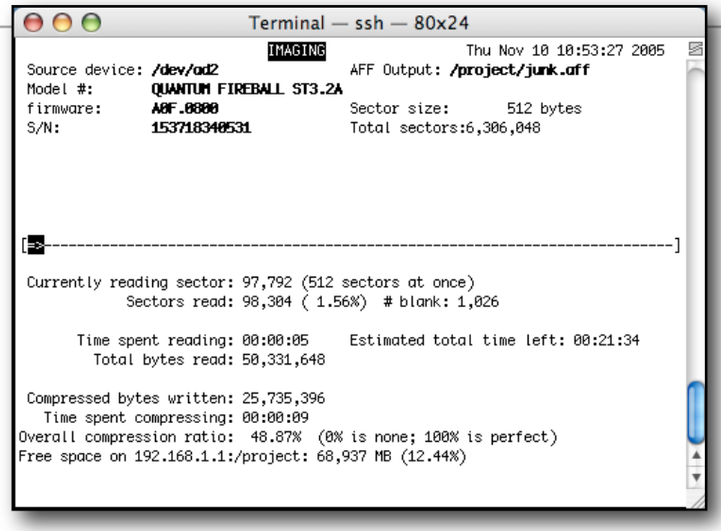




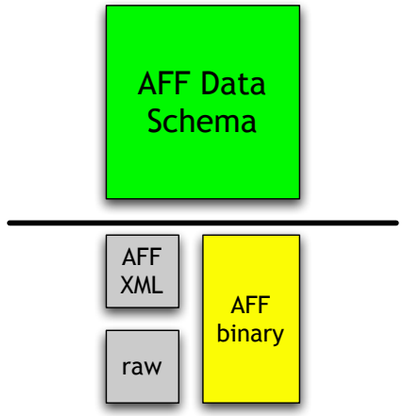
Tool Development

Tool Development:
Technology for working with many disk images.

Improved Disk Imaging: aimage



Disk Image Format: AFF



Storage & Processing: EC2 & S3

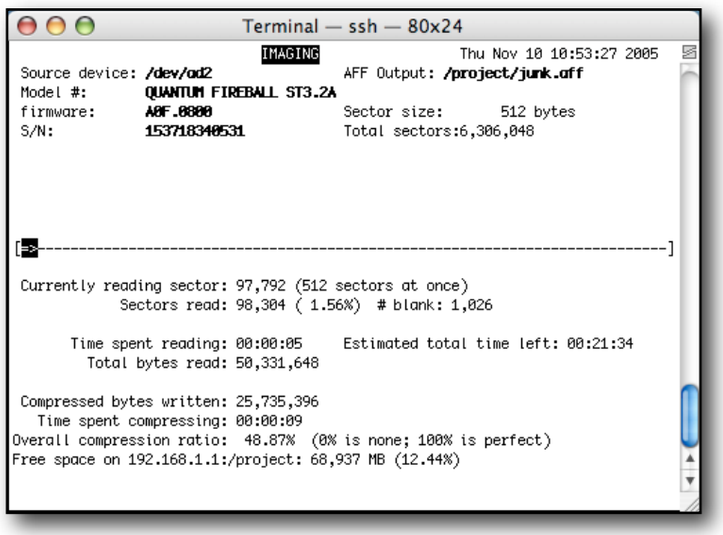
aimage: The Advanced Disk Imager

Most commercial development:

- Write-Blockers to prevent modification
- Network agents allow capture over a network
- Labor intensive

AImage & AFF [Garfinkel et. al, 06]

- One-click operation
- Combines imaging & error recovery
- Automatically captures metadata
- True encryption of image
- Digital signatures for imaging and chain-of-custody



Advanced File Format (AFF) [Garfinkel et. al, 2005]

AFF stores:

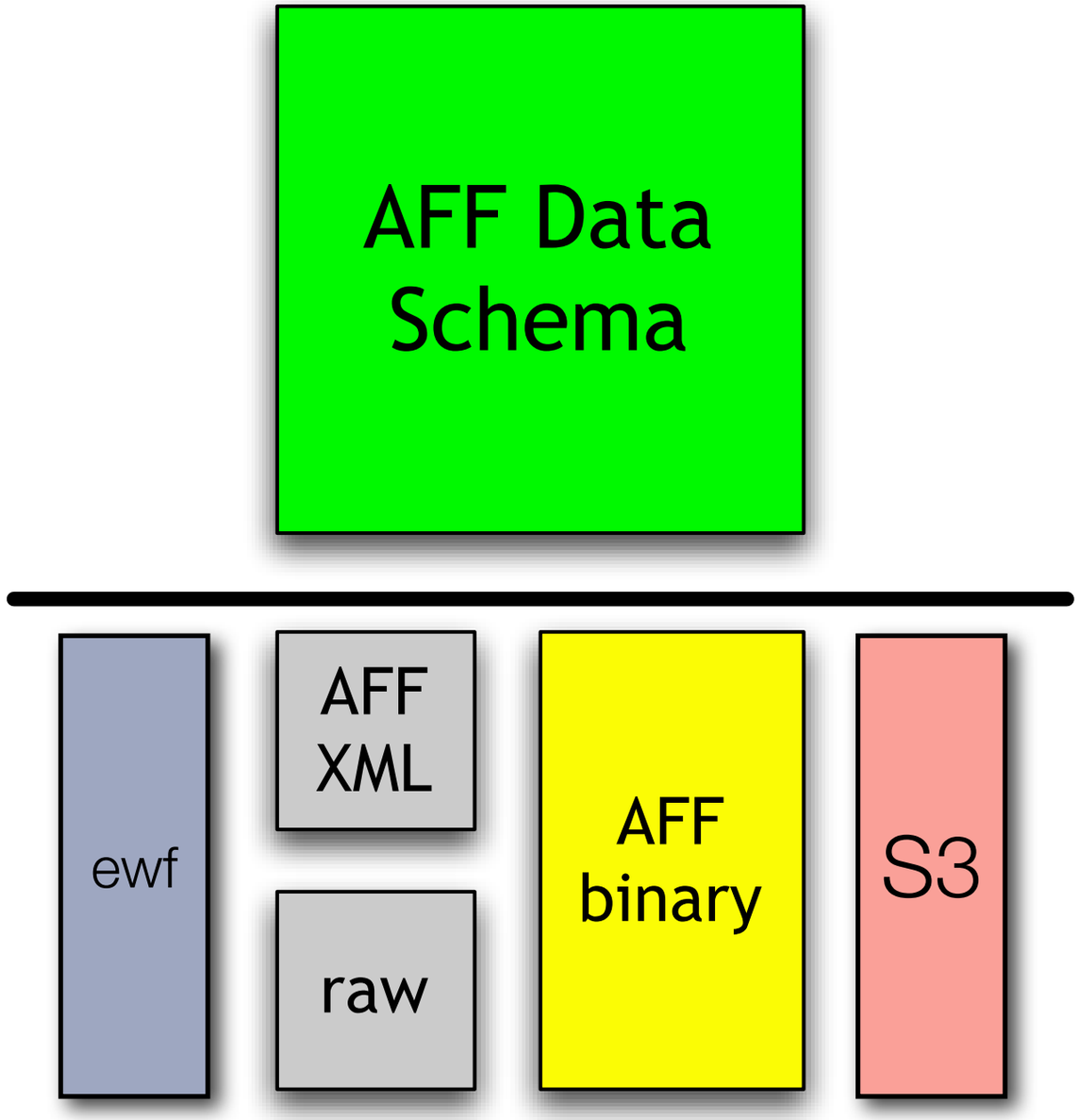
- Data copied from disk
- Metadata (SN, date of image)
- Any other name/value pair

AFF supports:

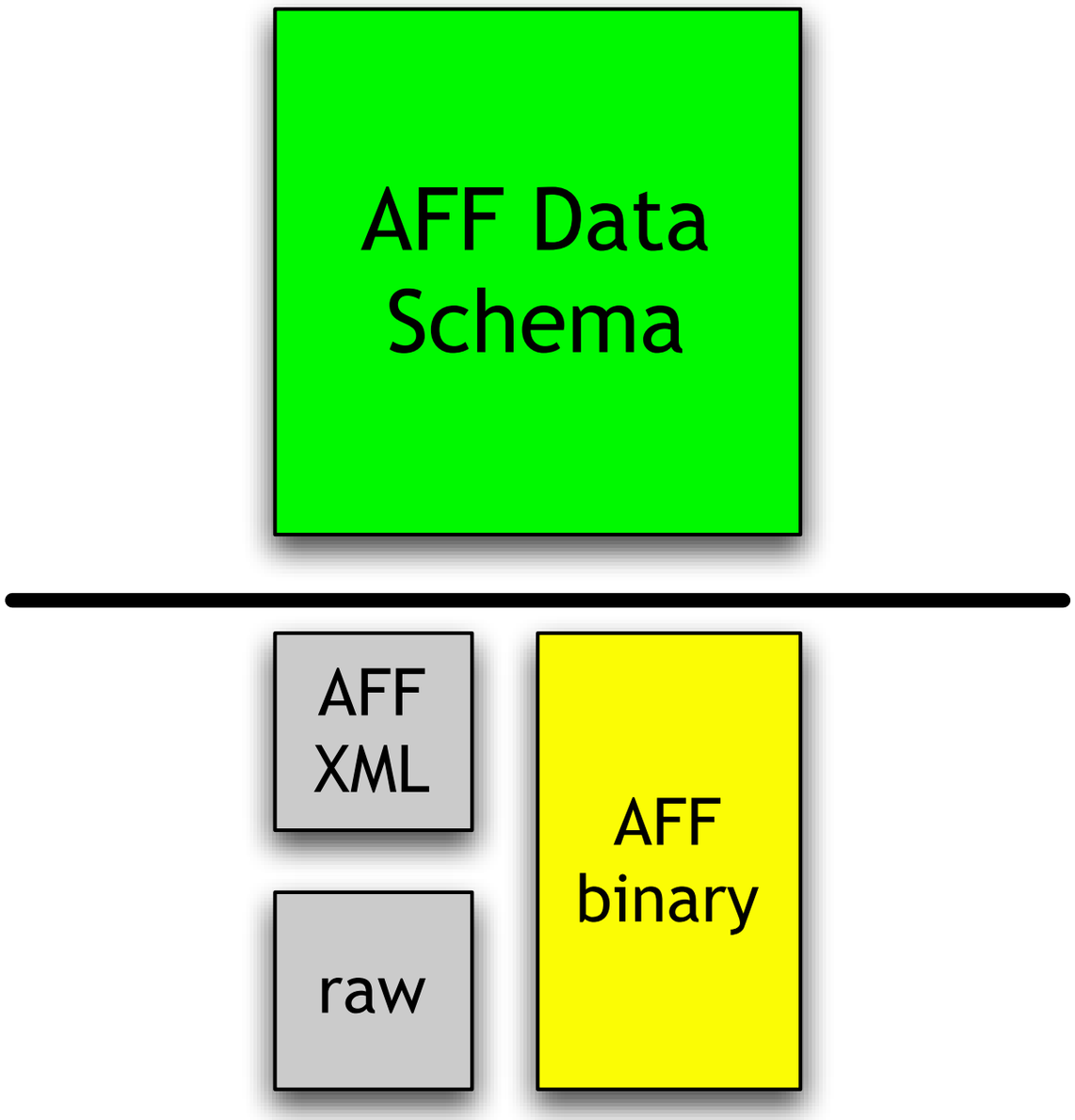
- Compression
- Encryption
- Digital Signatures

Adoption:

- Currently supported by Sleuth Kit
- Support coming to foremost
- Being evaluated by DC3



AFF Library & Tools



Open source (BSD License) C++ library

- Implements POSIX open(), read(), etc.
- Transparently reads AFF, EnCase, Raw, SplitRaw, and other file formats
- Linux FUSE implementation
- Also stores data on Amazon S3

Tools:

- afcat
- afcompare
- afconvert
- afcopy
- afinfo
- afxml...

Information Exploitation with AWS [Garfinkel 07]

Amazon Web Services offers commodity grid computing.

Simple Storage Service (S3)

- $0.15 per GB per month for storage
- $0.10 per GB transfered in
- $0.18 - $0.13 per GB transfered out.
- XMLRPC access protocol

Elastic Compute Cloud:

- Linux 32-bit virtual machines for $0.10 per CPU-Hour
- Approx. 1.7 Ghz, 1.75GB RAM, 160GB disk, 250 Mb/s network

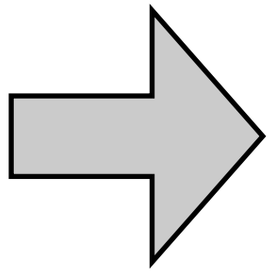
No bandwidth charges between S3 and EC2

Corpus Creation

Corpus Creation:
Generating Data for Research

The original goal: documenting sanitization failures.

Kept all disk images online because the recovery tools kept improving.



By spring 2005 we realized the images was useful for other purposes.

The Drives Corpus is a unique scientific resource.

The Drives Corpus is real data from real people.

- Geographically diverse
  (BA, CA, DE, IL, NZ, UK, US)
- Temporarily diverse (1992-present)
- OS diverse (DOS, Windows, Mac, Unix, etc.)
- Well-worn, not laboratory constructs

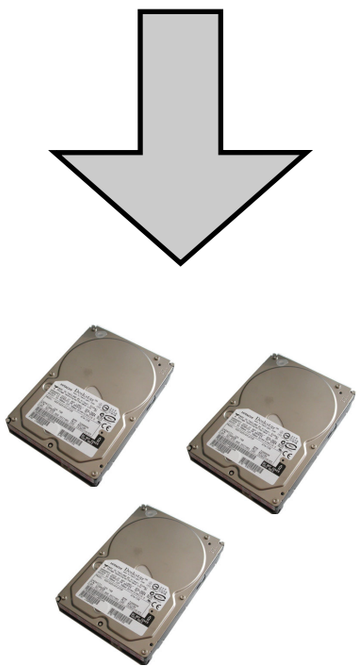
Data samples include:

- Personal and business systems
- Images, document files

Ideal for:

- Forensic tool development & validation
- Social network analysis, etc.

Used drives are a laboratory model.



The Drive Corpus is legal to use— with some restrictions

All of the drives:

- Are owned by Simson Garfinkel.
- Were discarded or sold.

California v. Greenwood 486 U.S. 35 (1988) holds no privacy in trash.

First-Sale Doctrine allows the owner of an article containing copyrighted information to sell or give it away.

On the other hand:

- Many drives contain confidential data from people who did not intend to give it away.
- Many drives have data on US Persons — possible Privacy Act issue.

Resolving Privacy and the Privacy Act.

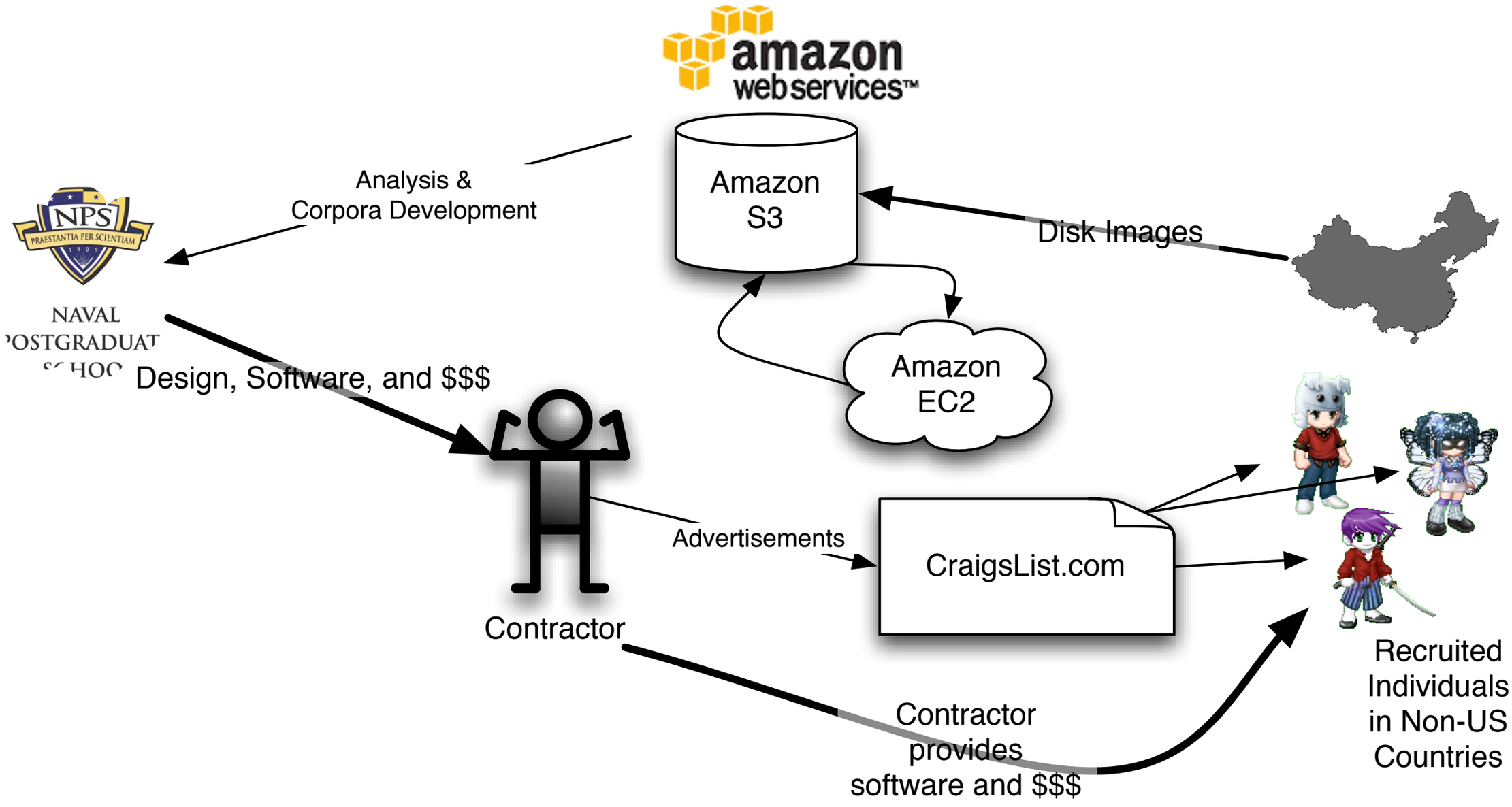
We treat the information as privacy-sensitive.

- Operate with IRB approval when human subjects are involved.
- No PII may be published.
- Restrict access to the corpus to bona fide researchers and forensic developers.
- AFF encryption provides additional security.

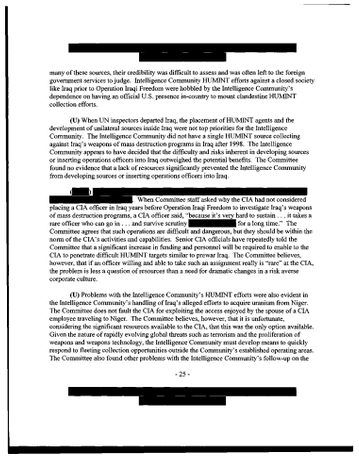
For US Government users:

- Separate US from Non-US corpora.
- No classification restrictions!

We are now building an international data collection network.



Corpus Creation: Future Directions


many of these sources, their credibility was difficult to assess and was often left to the foreign government services to judge. Intelligence Community HUMINT efforts against a closed society like Iraq prior to Operation Iraqi Freedom were hobbled by the Intelligence Community's dependence on having an official U.S. presence in-country to mount clandestine HUMINT collection efforts.

(U) When UN inspectors departed Iraq, the placement of HUMINT agents and the development of unilateral sources inside Iraq were not top priorities for the Intelligence Community. The Intelligence Community did not have a single HUMINT source collecting against Iraq's weapons of mass destruction programs in Iraq after 1998. The Intelligence Community appears to have decided that the difficulty and risks inherent in developing sources or inserting operations officers into Iraq outweighed the potential benefits. The Committee found no evidence that a lack of resources significantly prevented the Intelligence Community from developing sources or inserting operations officers into Iraq.

[redacted] When Committee staff asked why the CIA had not considered placing a CIA officer in Iraq years before Operation Iraqi Freedom to investigate Iraq's weapons of mass destruction programs, a CIA officer said, "because it's very hard to sustain . . . it takes a rare officer who can go in . . . and survive scrutiny [redacted] for a long time." The Committee agrees that such operations are difficult and dangerous, but they should be within the norm of the CIA's activities and capabilities. Senior CIA officials have repeatedly told the Committee that a significant increase in funding and personnel will be required to enable to the CIA to penetrate difficult HUMINT targets similar to prewar Iraq. The Committee believes, however, that if an officer willing and able to take such an assignment really is "rare" at the CIA, the problem is less a question of resources than a need for dramatic changes in a risk averse corporate culture.

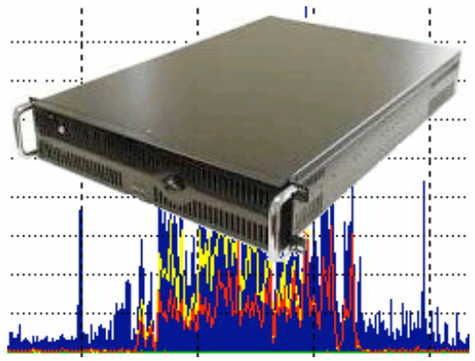
(U) Problems with the Intelligence Community's HUMINT efforts were also evident in the Intelligence Community's handling of Iraq's alleged efforts to acquire uranium from Niger. The Committee does not fault the CIA for exploiting the access enjoyed by the spouse of a CIA employee traveling to Niger. The Committee believes, however, that it is unfortunate, considering the significant resources available to the CIA, that this was the only option available. Given the nature of rapidly evolving global threats such as terrorism and the proliferation of weapons and weapons technology, the Intelligence Community must develop means to quickly respond to fleeting collection opportunities outside the Community's established operating areas. The Committee also found other problems with the Intelligence Community's follow-up on the

- 25 -

[redacted]

Documents

Cell Phones



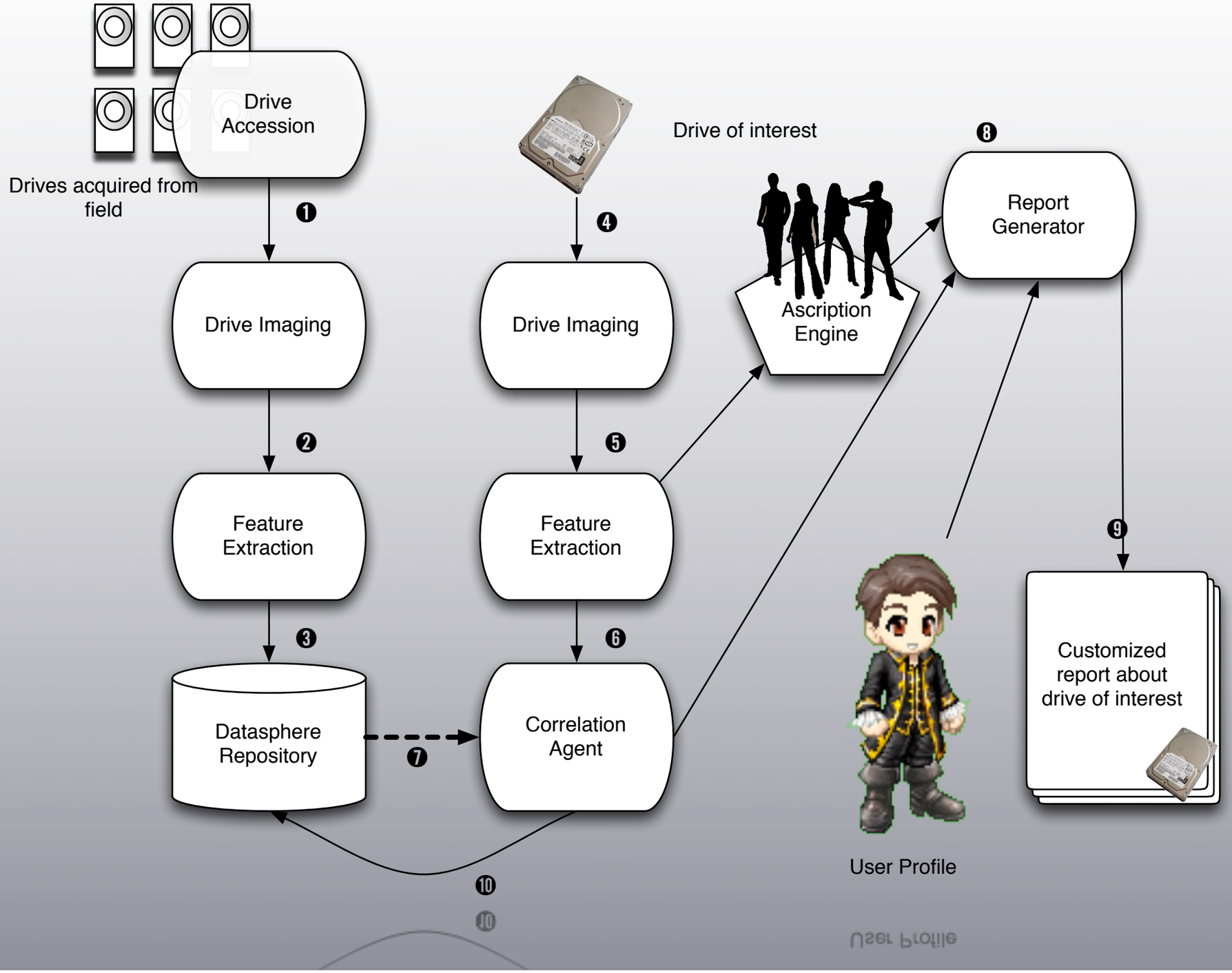
Packet Archives

iPods

SD Cards

USB Memory

SIM Cards

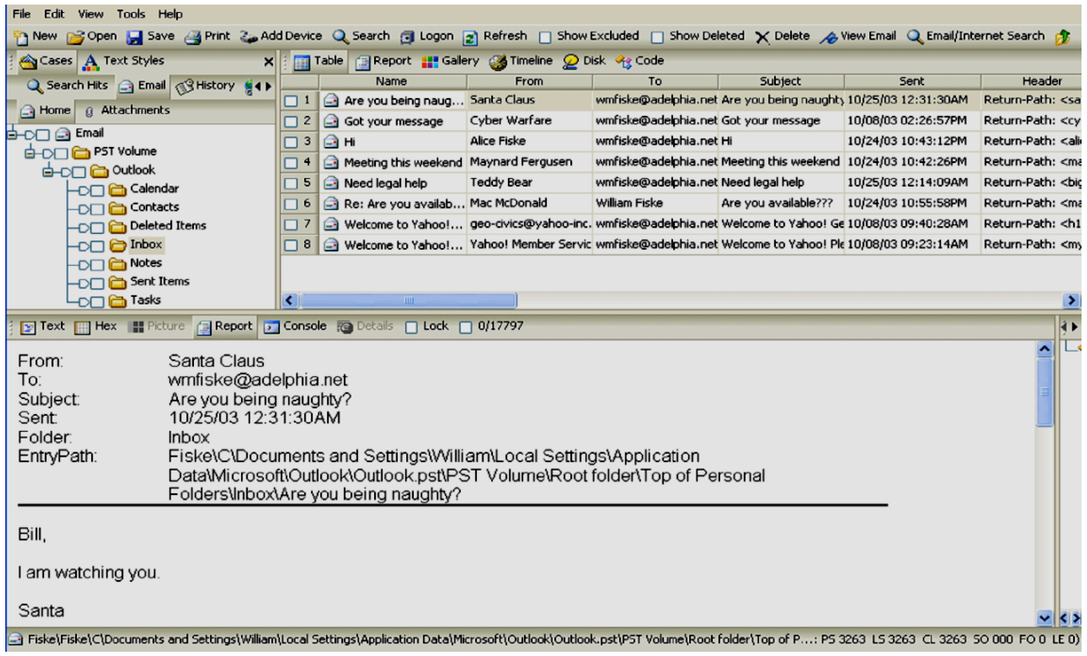


Architecture & Algorithms

Today most investigators use GUI-based tools

These are *search* and *visualization* tools:

- The Sleuth Kit (Carrier, Open Source)
- Encase (Guidance Software)
- Forensic Toolkit (AccessData)



Designed for investigating a single drive.

Do not scale to thousands of drives.

My tools are designed to be automatic and non-interactive.

Feature Extraction finds identifiers in each disk.

Credit Card Numbers (CCN)

5422-4821-3008-8635

Email addresses

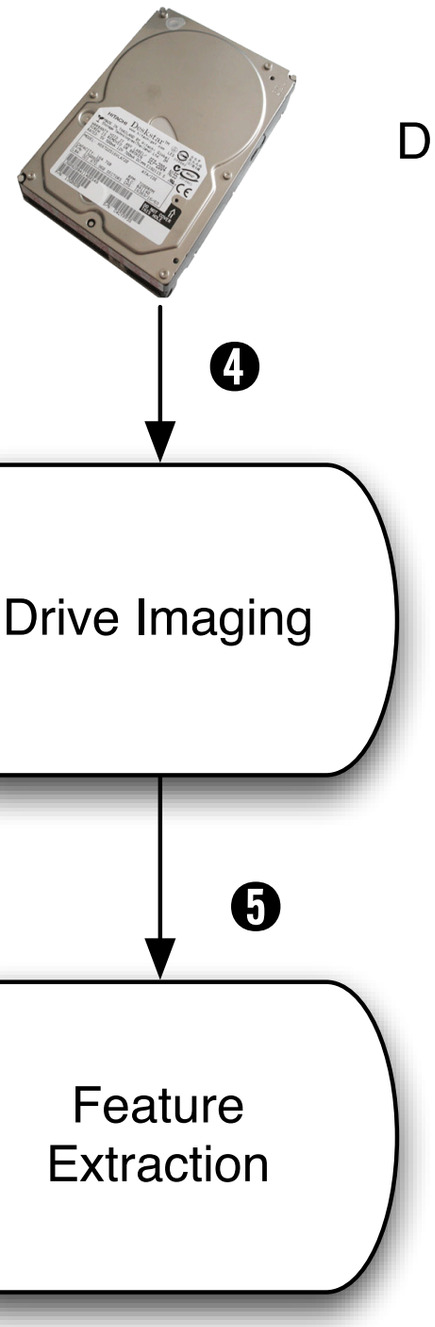
user@company.com

RFC-822 Time and Message-ID detector

<Pine.LNX.4.61.072250.6378@infosecnews.org>

Date: Wed, 9 May 2007 00:02:51 -0500 (CDT)

Internet Explorer cache, cookies, etc.

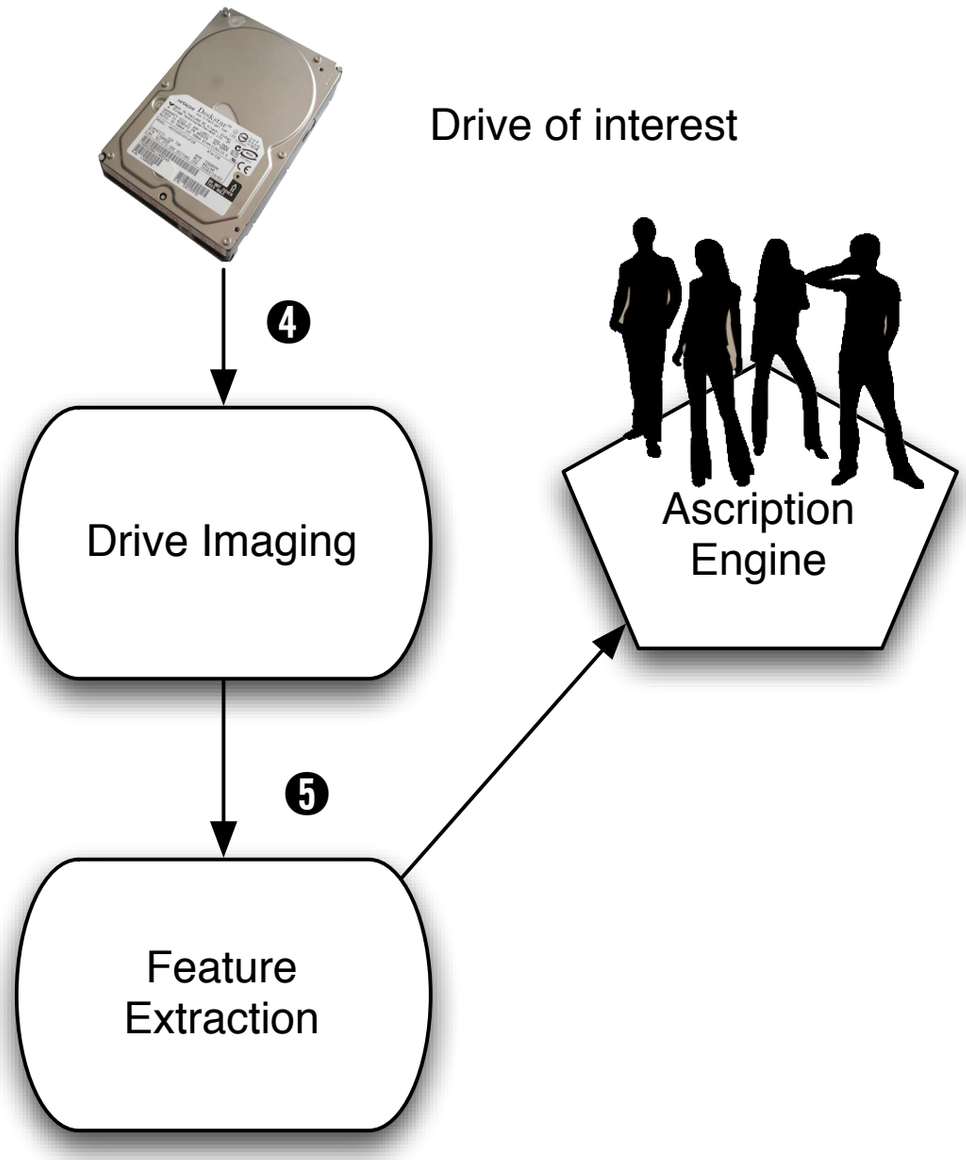


Drive Ascription identifies the drive's former owner.

Applications for Drive Ascription:

- Re-identifying captured drives
- Return of stolen property
- Recovering from clerical errors

Hard drives usually aren't labeled with their owner's name!



Drive Ascription Example:
Extracting and counting email addresses.

Drive #51 (Anonymized)

ALICE@DOMAIN1.com	8133
BOB@DOMAIN1.com	3504
ALICE@mail.adhost.com	2956
JobInfo@alumni-gsb.stanford.edu	2108
CLARE@aol.com	1579
DON317@earthlink.net	1206
ERIC@DOMAIN1.com	1118
GABBY10@aol.com	1030
HAROLD@HAROLD.com	989
ISHMAEL@JACK.wolfe.net	960
KIM@prodigy.net	947
ISHMAEL-list@rcia.com	845
JACK@nwlink.com	802
LEN@wolfenet.com	790
natcom-list@rcia.com	763

Email Drive Attribution with “background noise.”

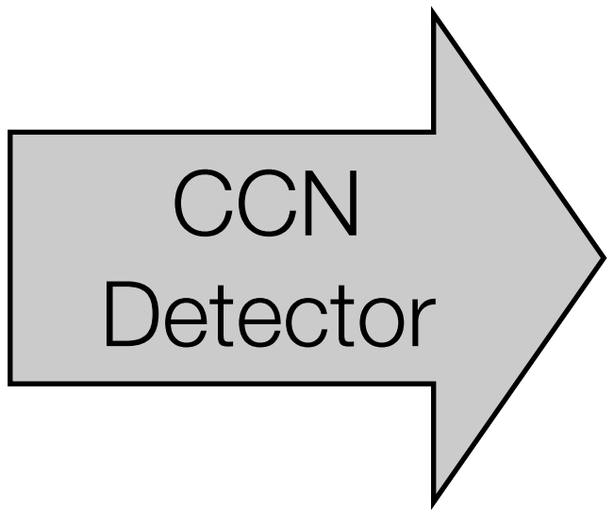
Extracted Email Addresses	Count on Drive #80	Total drives with address
premium-server@thawte.com	117	278
server-certs@thawte.com	104	278
CPS-requests@verisign.com	61	286
personal-premium@thawte.com	44	253
personal-basic@thawte.com	42	250
personal-freemail@thawte.com	40	250
info@netscape.com	36	58
ANGIE@ALPHA.com	32	1
BARRY@BETA.com	23	1
CHARLES@GAMMA.com	21	1
DAVE.HALL@DELTA.com	21	1
DAPHNE@UNIFORM.com	20	1
ELLY@LIMA.com	18	1

Hot Drive Detection:
Automatically finding “the good stuff.”

We were interested in showing that **sensitive information** had been left behind on hard drives.

We wrote a **Credit Card Number** detector.

Bulk Data



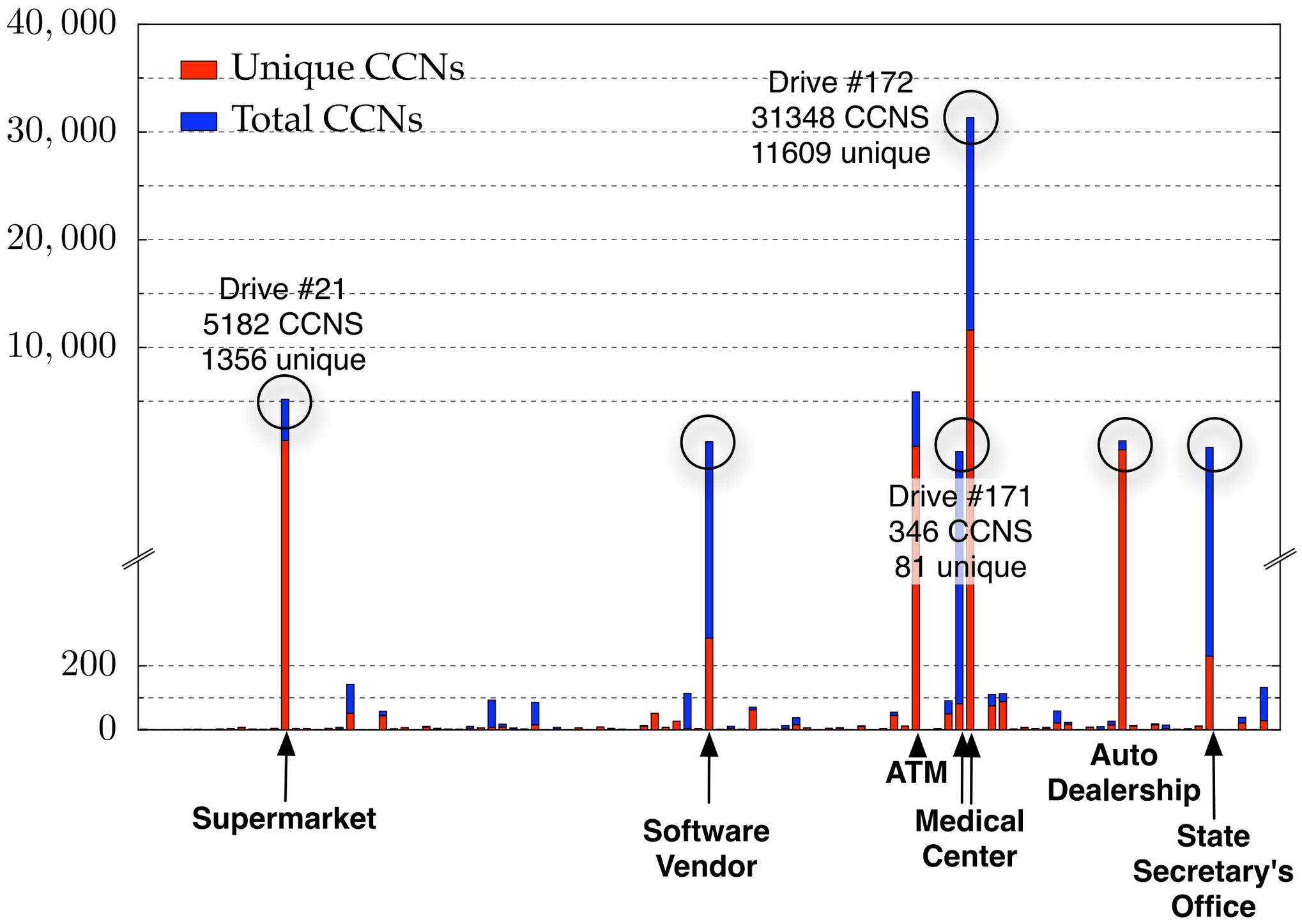
3713 107885 95004
@ offset 554,553,664

4388 5750 2212 4449
@ offset 43,332,334,554

5543 3343 1644 3345
@ offset 6,334,877,809

...

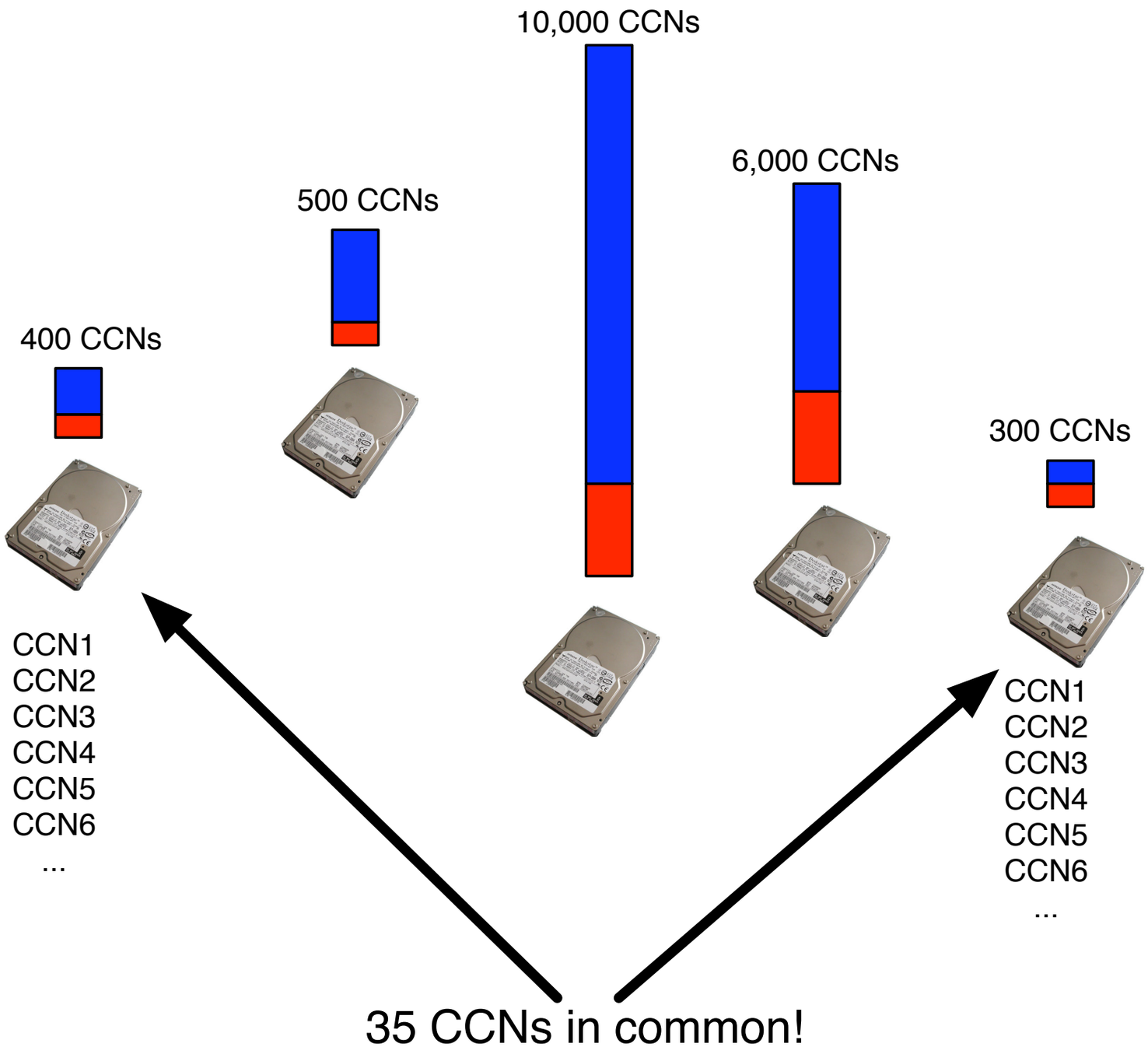
Most drives had just a few, but some drives had a *lot* of credit card numbers.



Feature Extraction: Performance

- I/O bound with raw disk images
- CPU bound with AFF-compressed images (50% better than E01 format)
- 1-6 GB/hour on a 1.5Ghz 64Bit AMD Athlon
- Can be run in parallel
- Took 3 weeks to extract ≈ 750 drives
- Can be done incrementally, as new drives are acquired

What would it mean if two drives had a lot of credit card numbers in common?

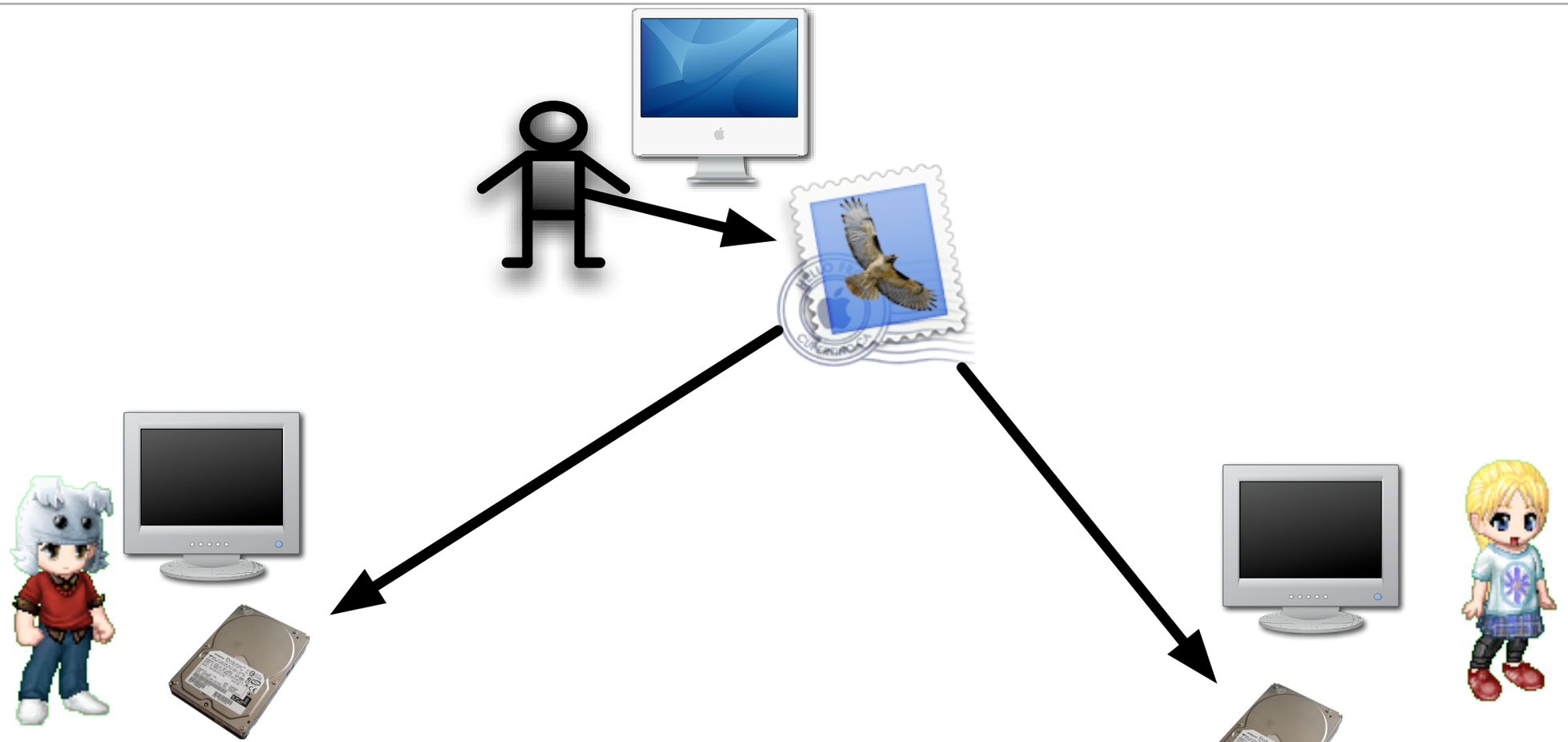


Scenario #1: The owner of one drive sent a message to another drive.

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

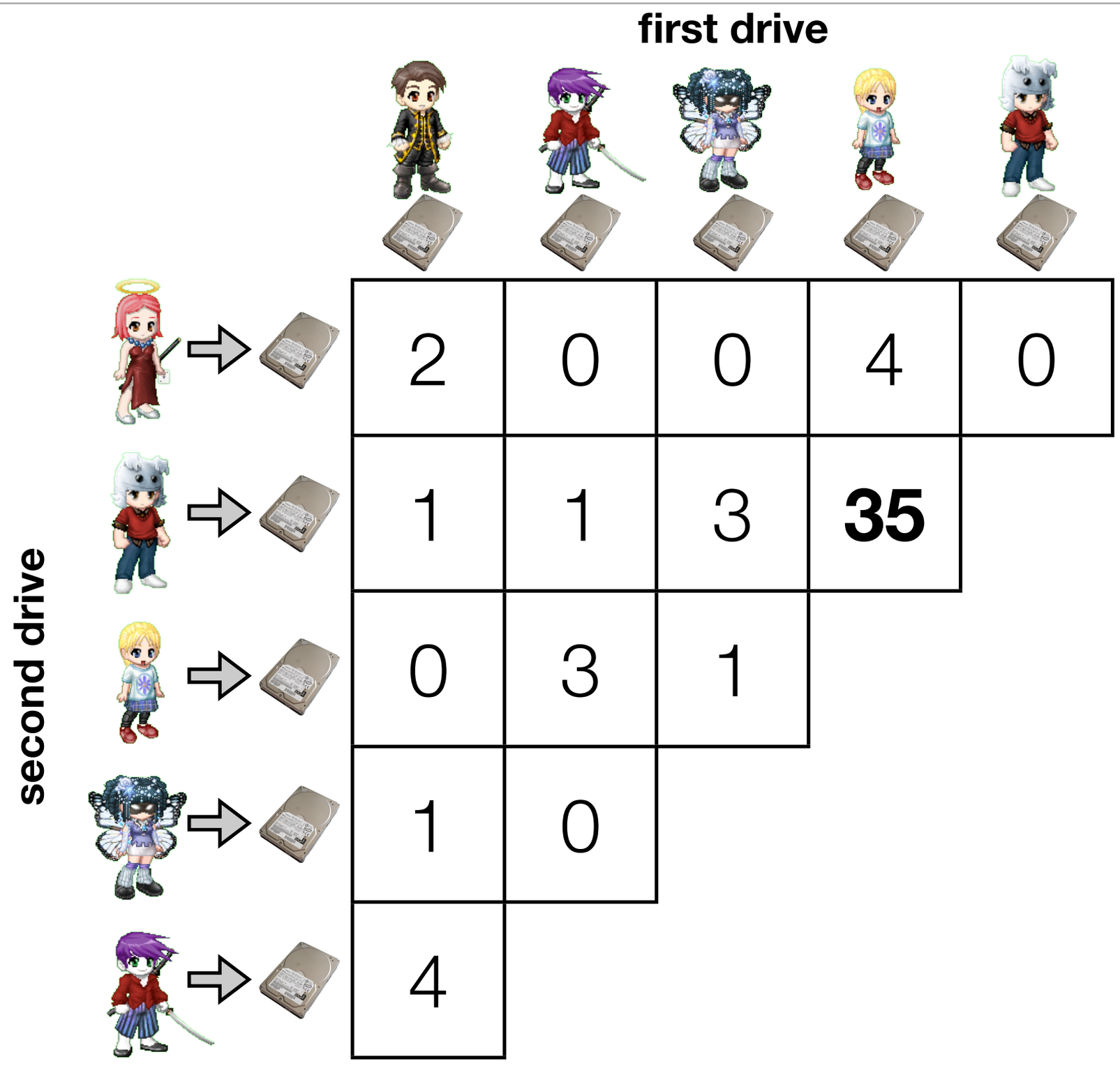
Scenario #2:
Both drives received a message from a third party.



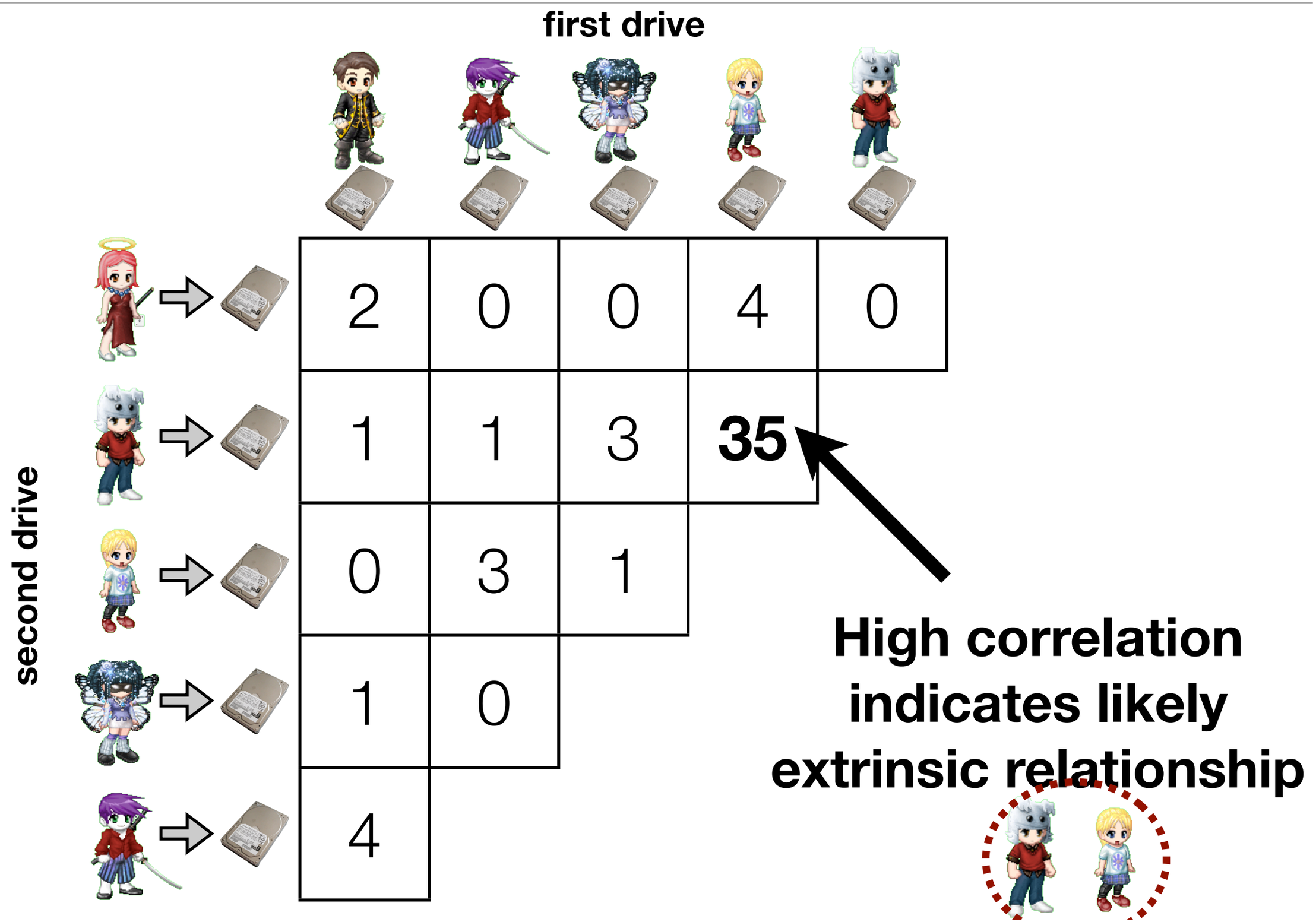
CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

CCN1
CCN2
CCN3
CCN4
CCN5
CCN6
...

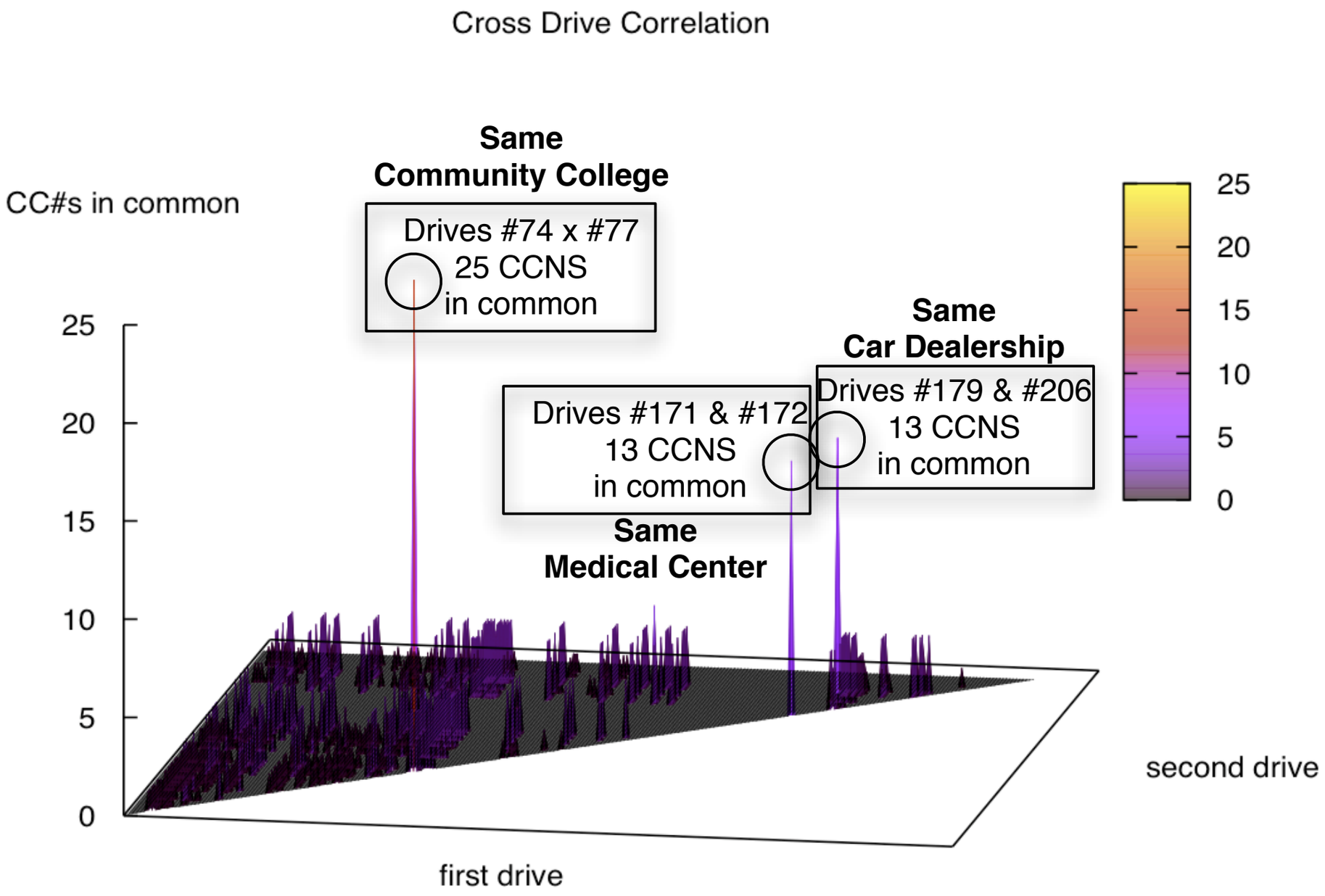
Cross Drive Analysis (CDA) computes the correlation matrix of the pseudounique information.



Cross Drive Analysis (CDA) computes the correlation matrix of the pseudounique information.



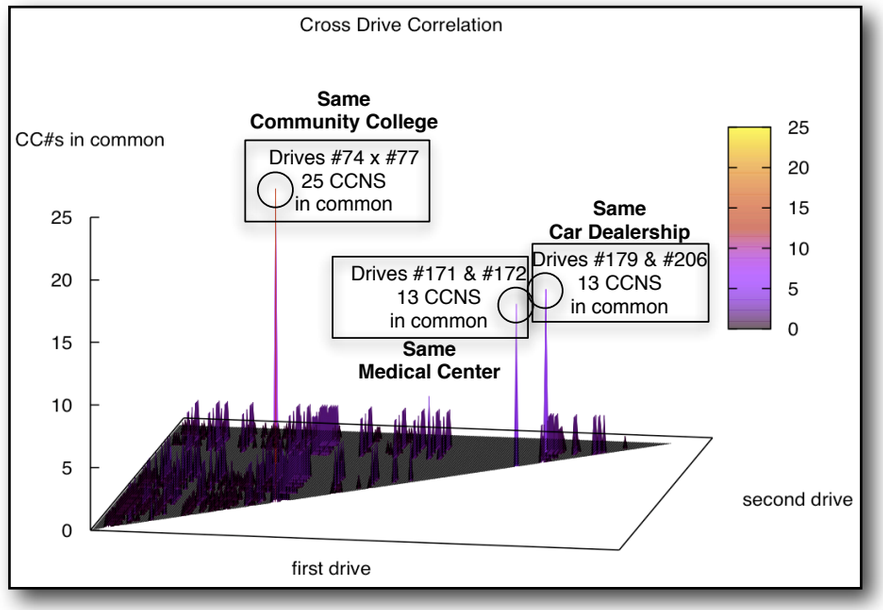
Here is a correlation of CCNs on the first 250 drives:



Initial experience with Cross Drive Analysis.

CDA requires “pseudo-unique” features:

- CCNs and other financial information
- Transliterated names
- email addresses, Message-IDs
- Sector “fingerprints”
- Encrypted data



CDA Process:

1. Extract features
2. Correlate — $O(n^2)$
3. Generate Report

Cross Drive Analysis: Performance

Cross Drive Correlation is **memory-bound.**

Tests with 750 drives, 5.8M features, on 1.5Ghz machine with 2GB of RAM:

First Implementation: Python

- Process had just 1% CPU utilization due to swapping (no reference locality)
- Never finished...

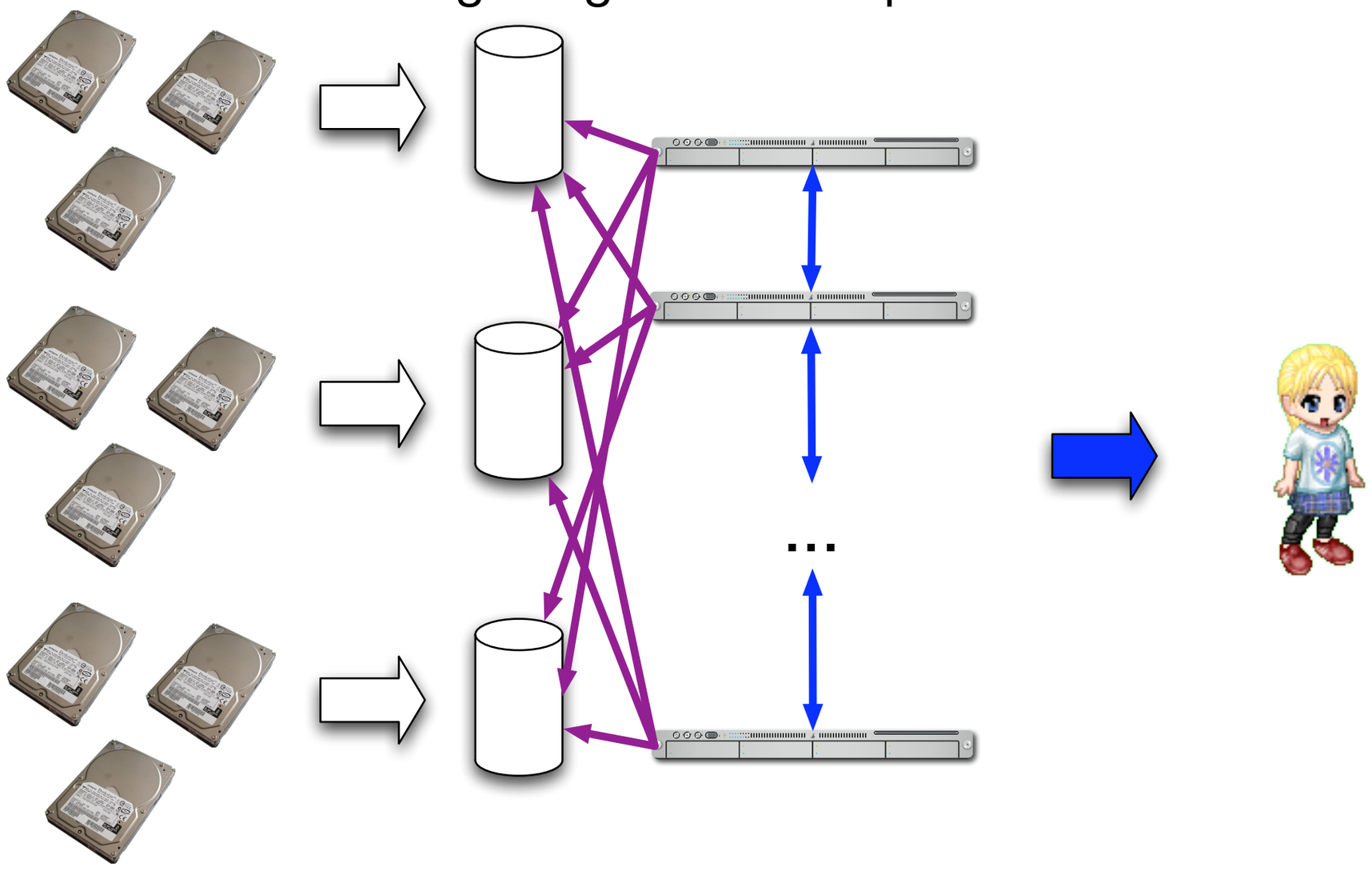
Second Implementation: Highly optimized C++

- Process grew to 700MB, then correlation started.
- Correlation finished in 30 minutes

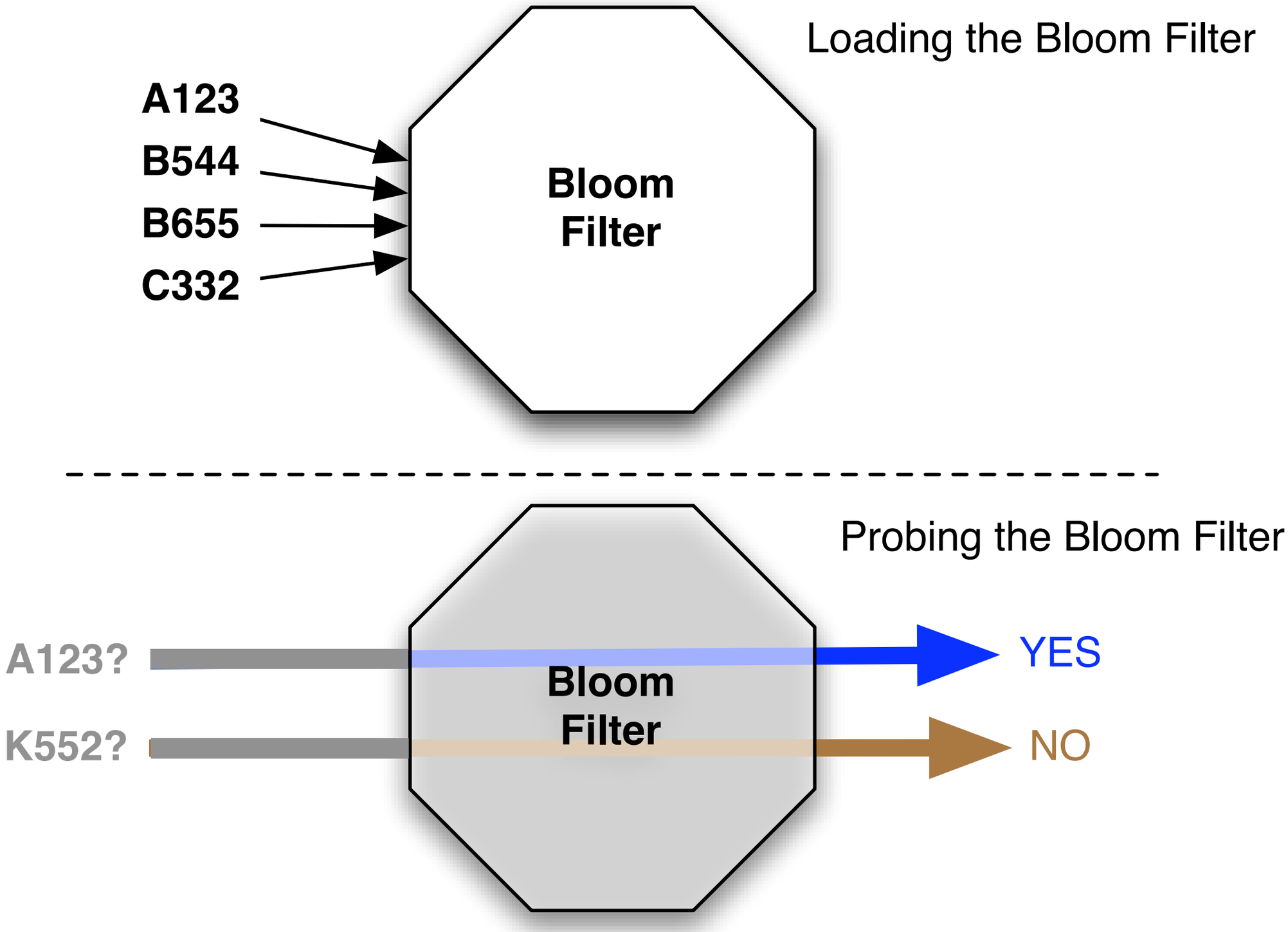
Cross Drive Analysis: Next Generation

Scalable storage and computation.

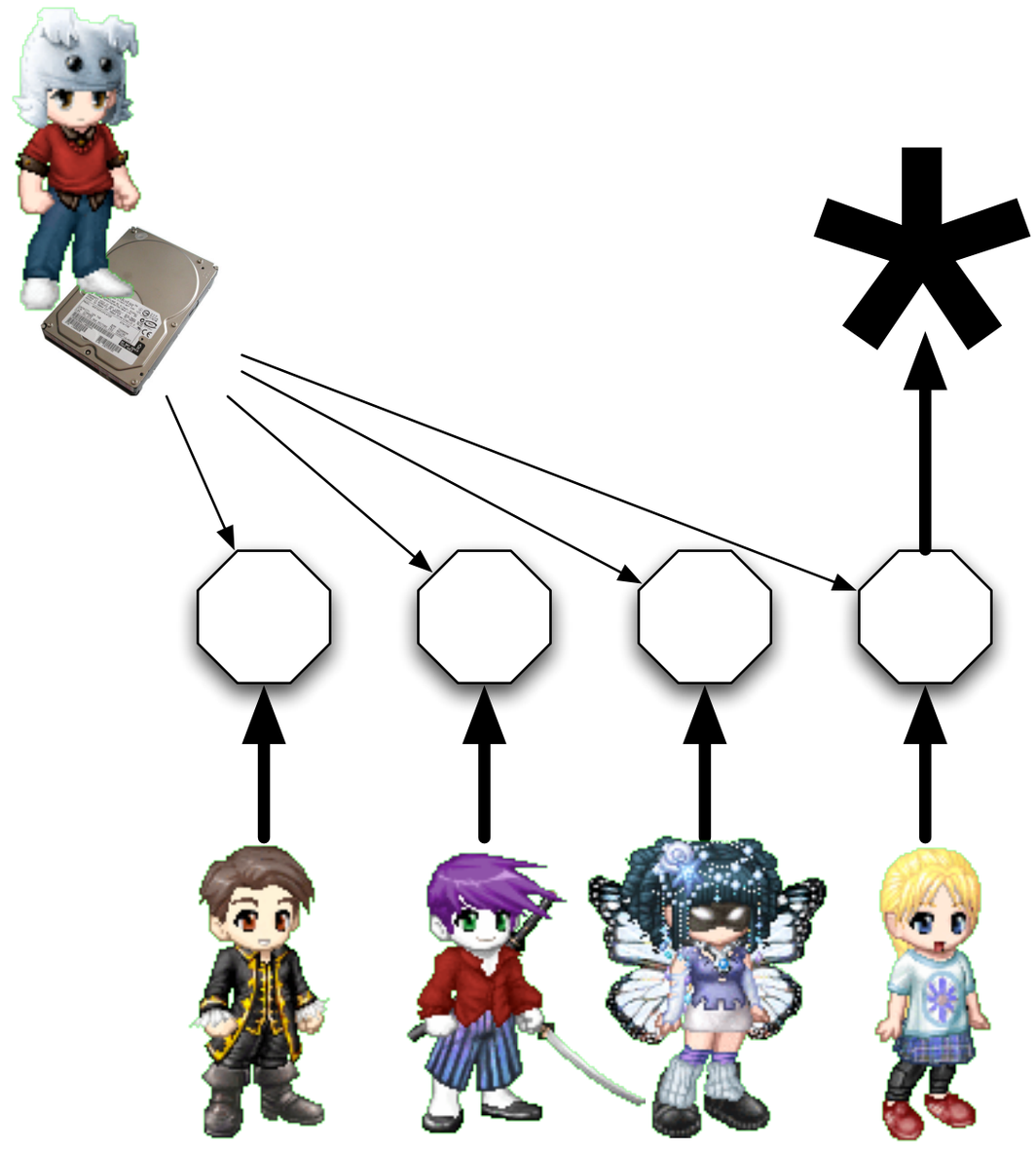
- Amazon EC2 & S3
- Gigabit multicast to leverage large address space



Using Bloom Filters for Cross Drive Analysis



A practical CDA system could have a Bloom Filter for each organization of interest



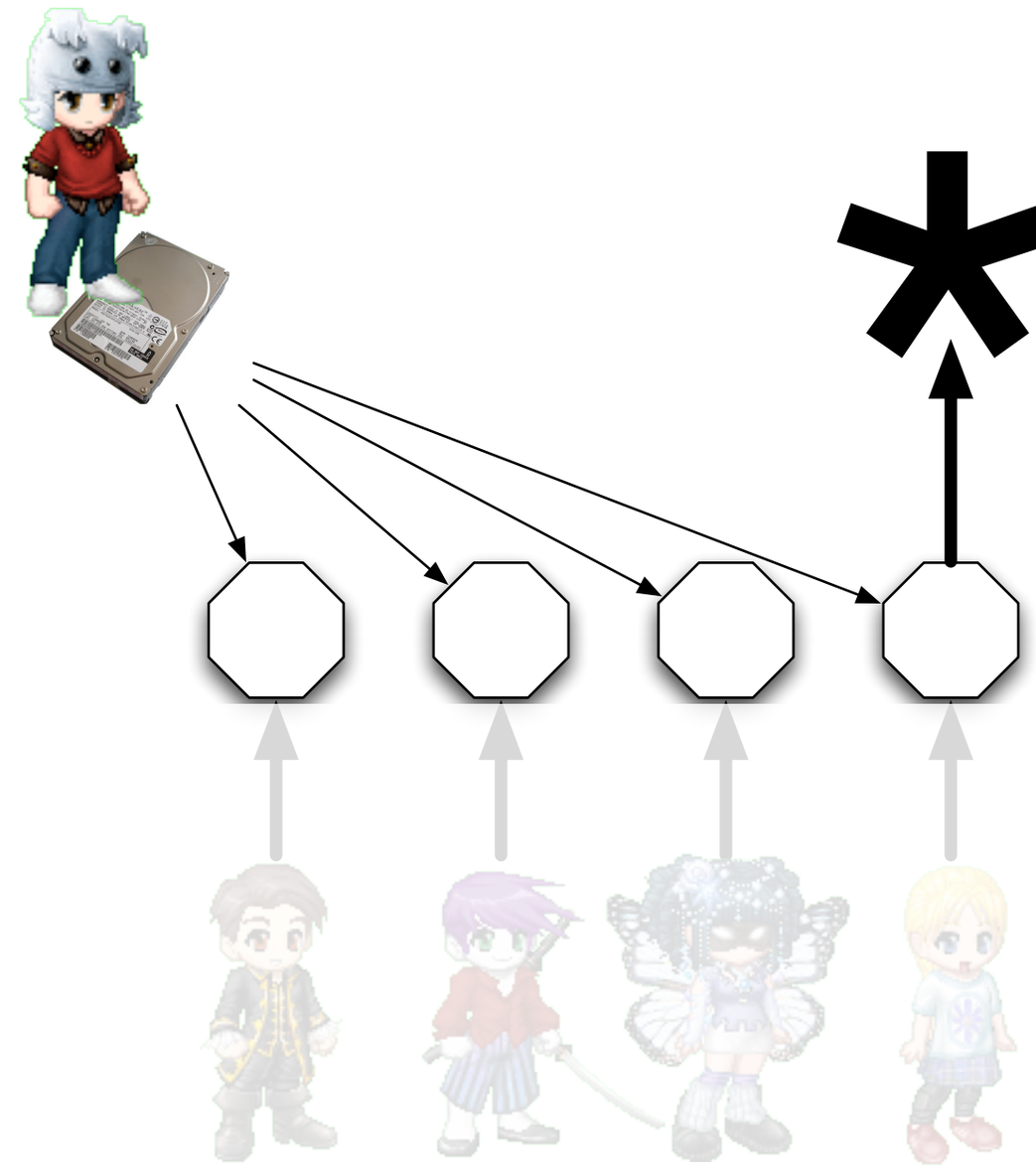
New information can be automatically screened as it is acquired.

Bloom filters offer significant advantages.

Filters can’t be reverse-engineered.

Filters can be searched without new court orders.

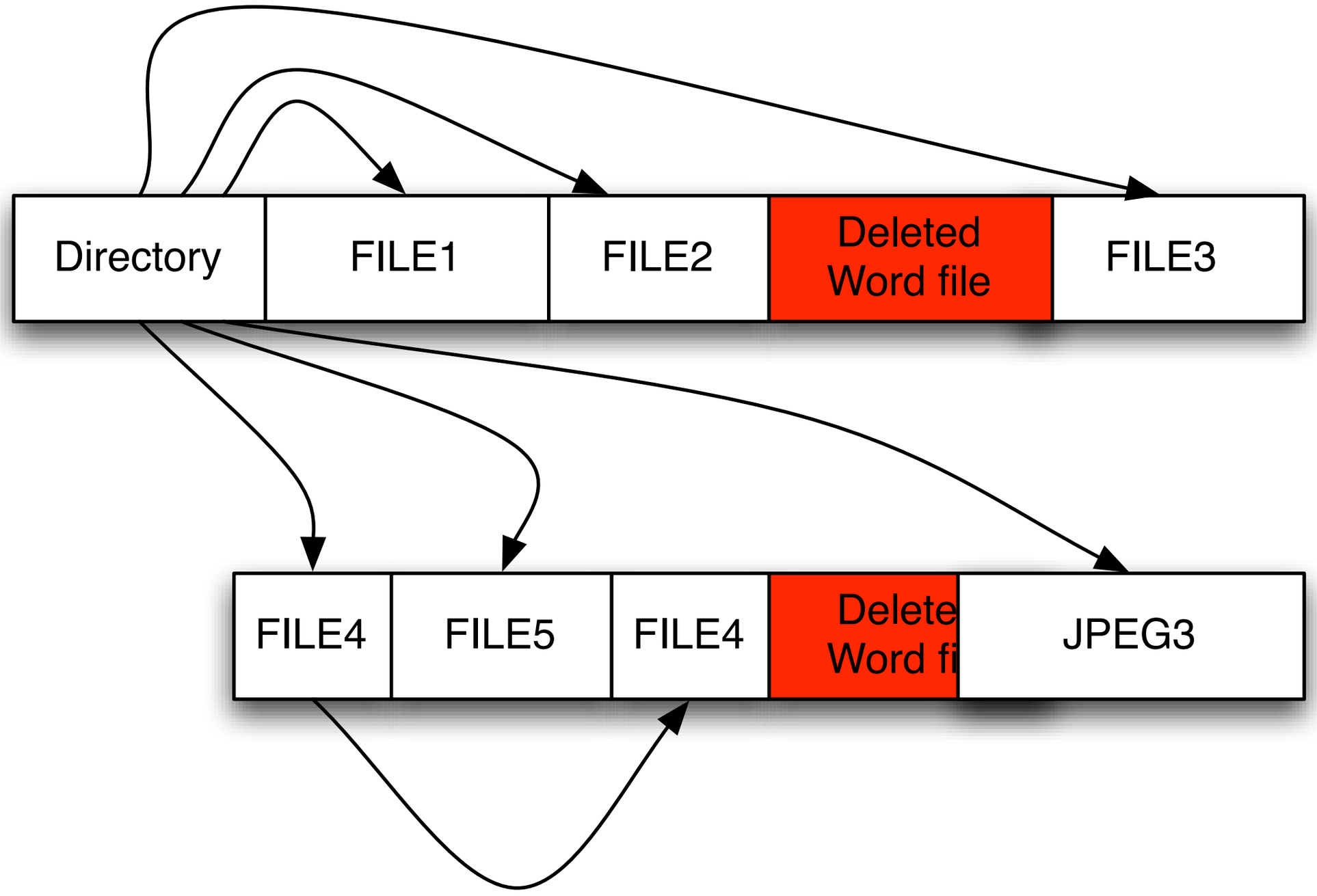
Filters can be combined for improved efficiency.



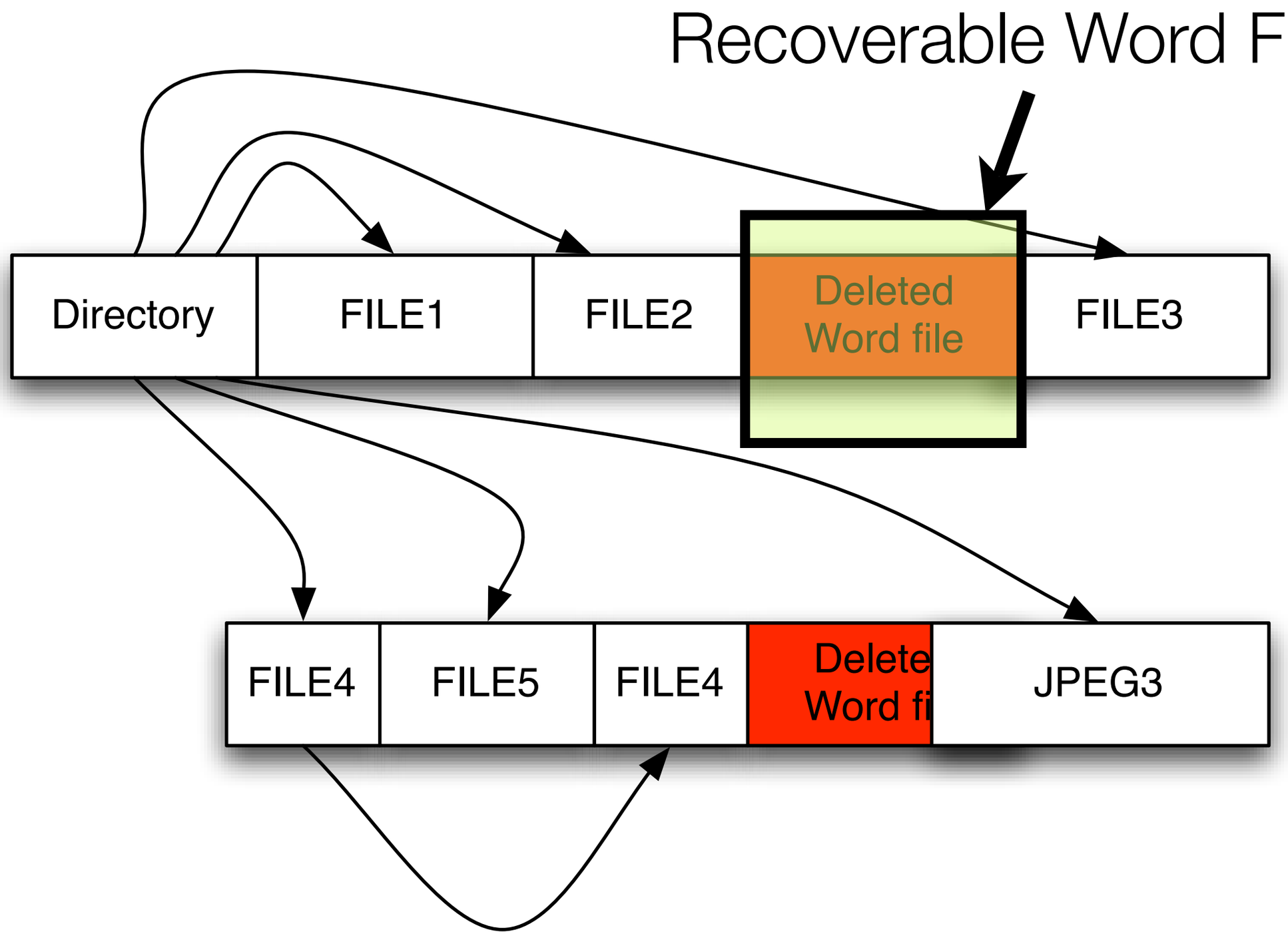
http://www.nps.gov/history/museum/exhibits/band/slideshow/CCC/carving_6.html

File Carving

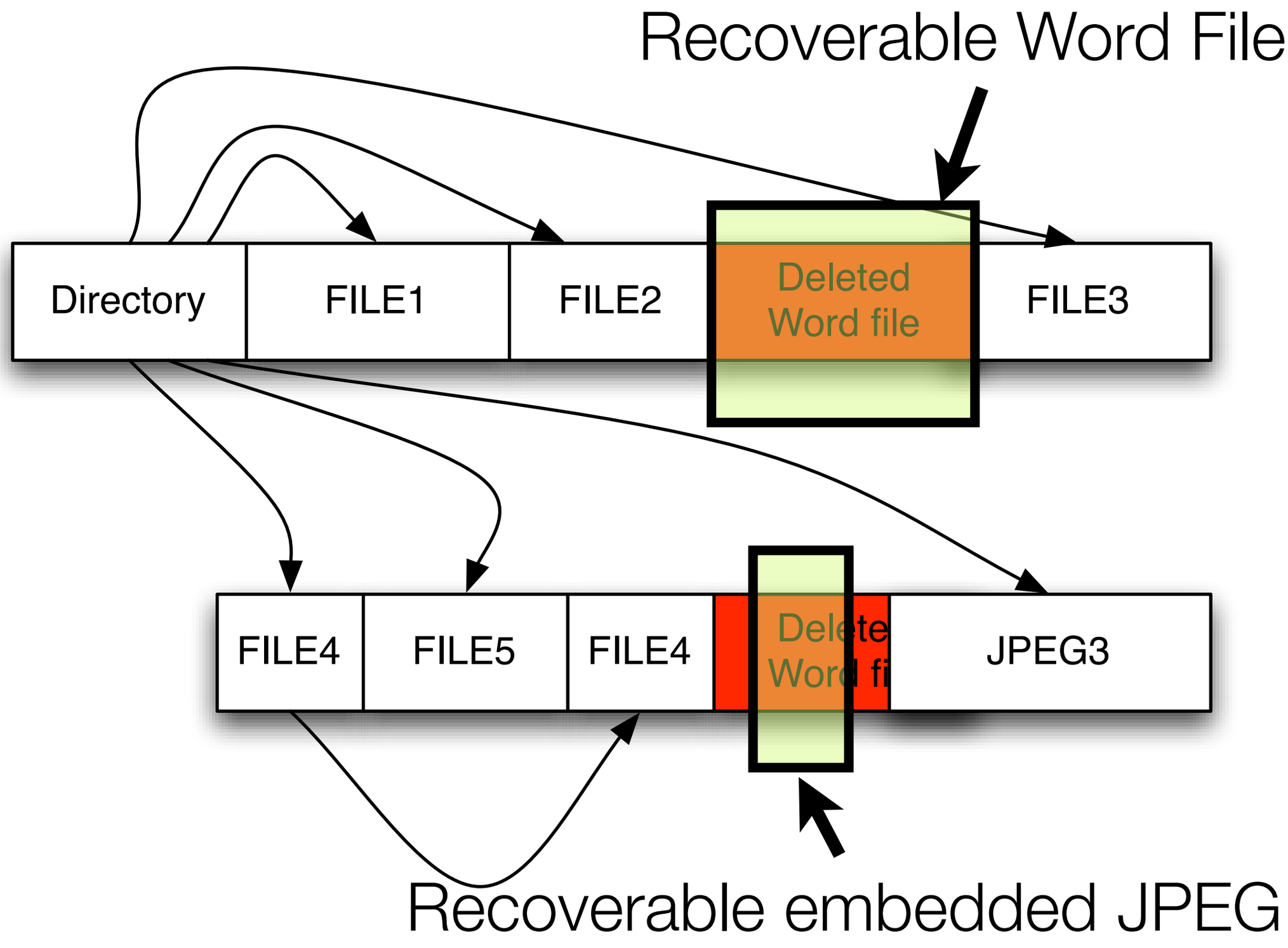
"Carving" is the search for objects based on *content*, rather than on metadata.



"Carving" is the search for objects based on *content*, rather than on metadata.



"Carving" is the search for objects based on *content*, rather than on metadata.



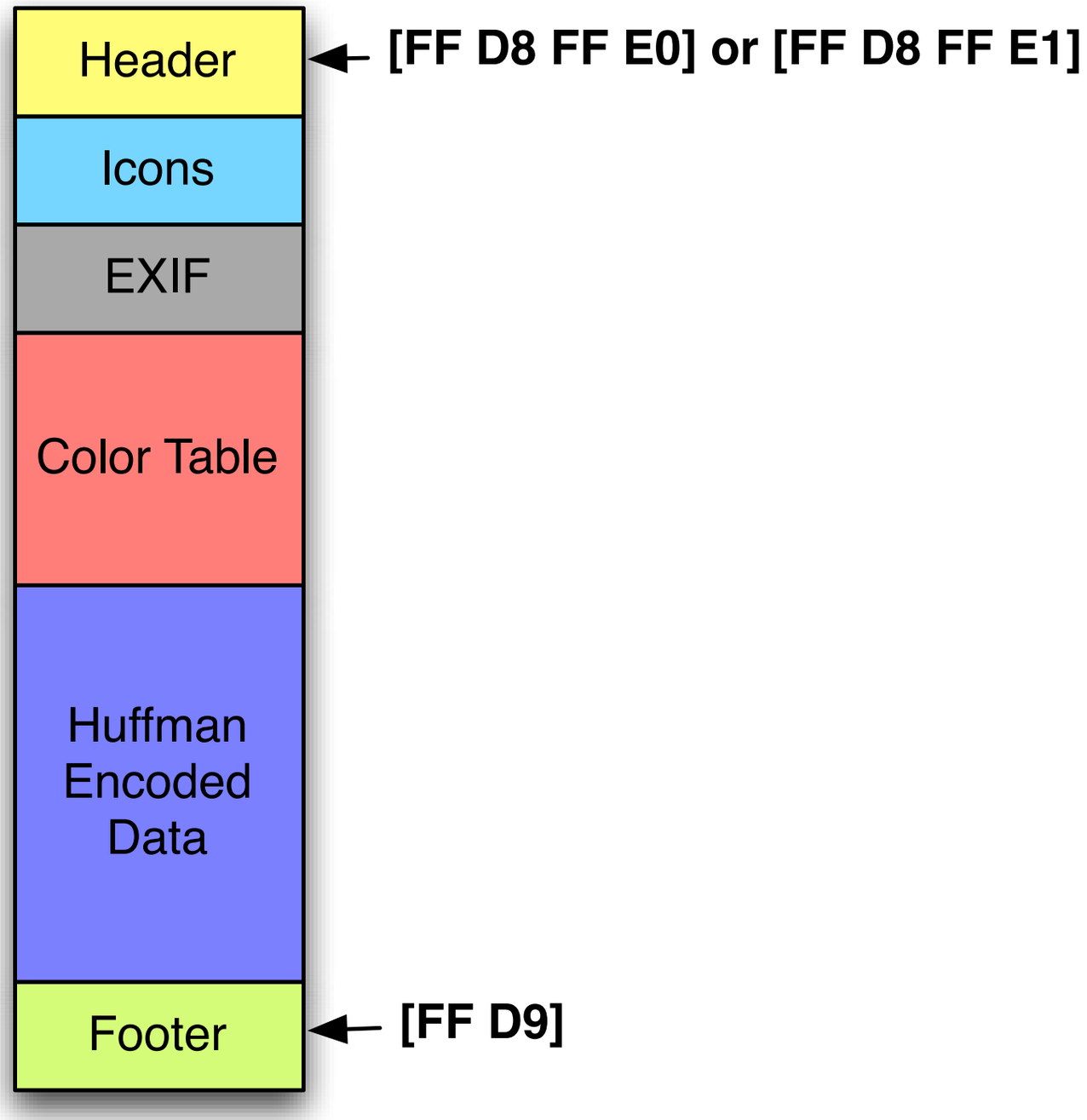
Example: Carving JPEG Files

JPEGs are container files

- Standard Header
- Standard Footer
- Embedded Images

Carving strategy:

- Find all headers
- Find all footers
- Save sectors to files



Header/Footer carving with JPEG:
Fast, but error prone.

Disk Sectors ➔

Header/Footer carving with JPEG:
Fast, but error prone.

Disk Sectors ➔

Header/Footer carving with JPEG:
Fast, but error prone.

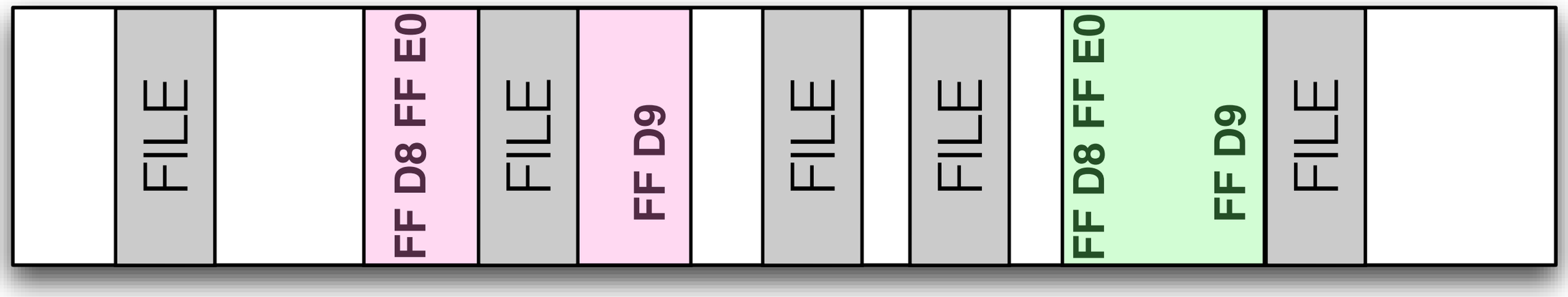
Disk Sectors ➔

Header/Footer carving with JPEG:
Fast, but error prone.

Disk Sectors ➔

This strategy is used by **foremost** and **scalpel**

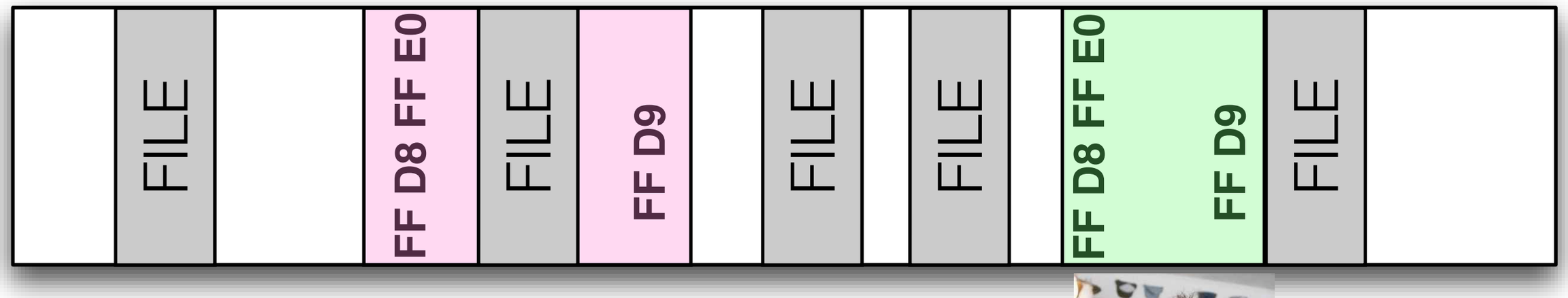
Header/Footer carving with JPEG:
Fast, but error prone.



Disk Sectors ➔

This strategy is used by **foremost** and **scalpel**

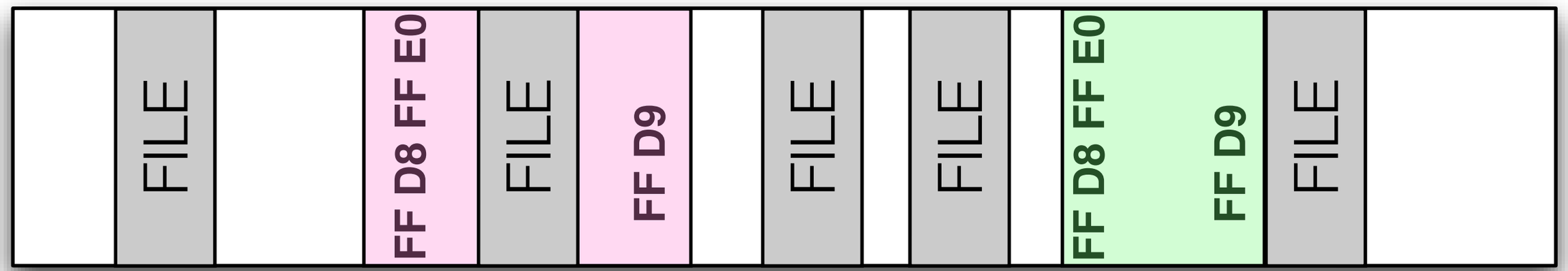
Header/Footer carving with JPEG:
Fast, but error prone.

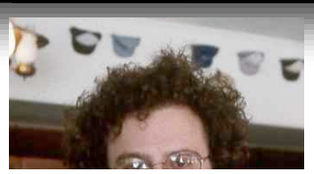


Disk Sectors ➔

This strategy is used by **foremost** and **scalpel**

Header/Footer carving with JPEG:
Fast, but error prone.

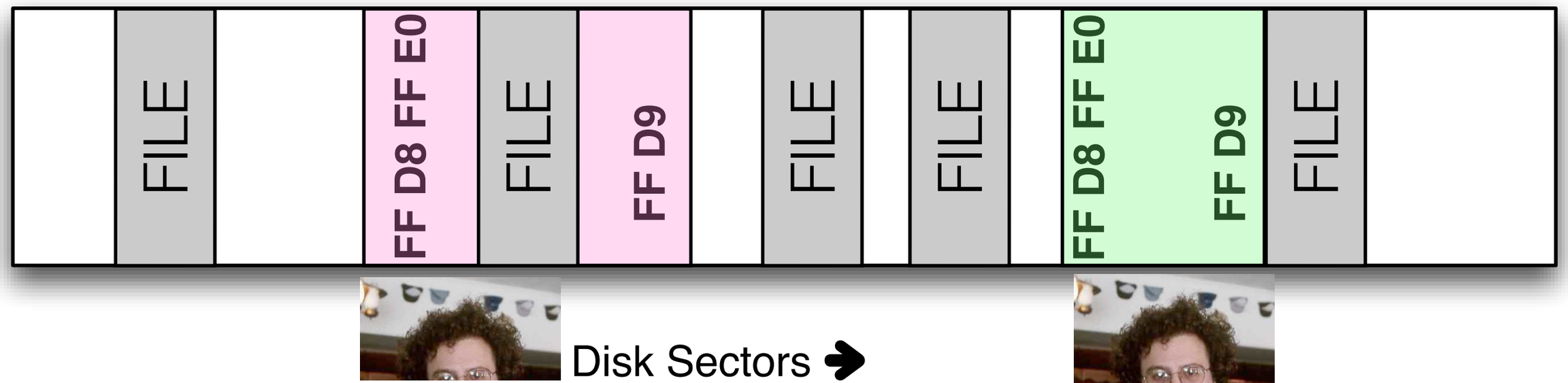




Disk Sectors ➔

This strategy is used by **foremost** and **scalpel**

Header/Footer carving with JPEG:
Fast, but error prone.



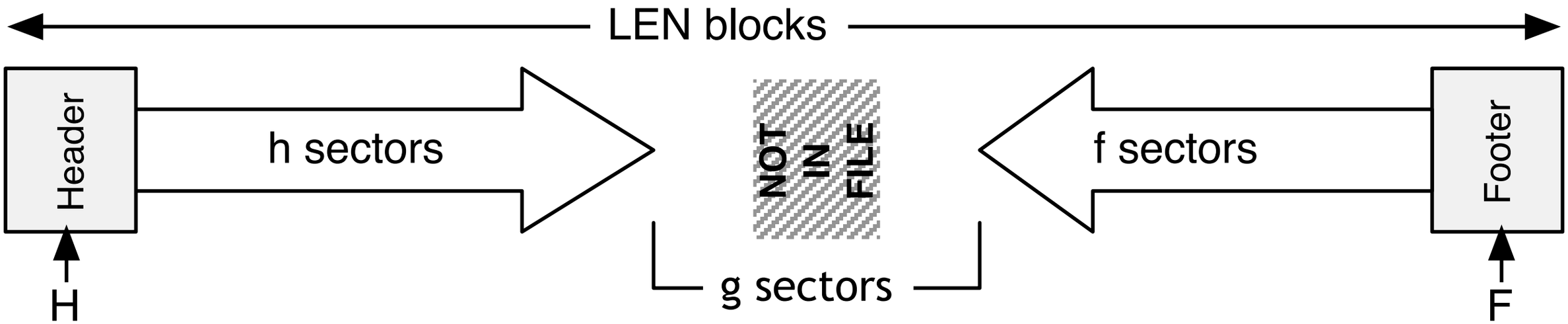
Disk Sectors ➔

This strategy is used by **foremost** and **scalpel**

With simple header/footer carving, objects must be *validated* after they are saved in files (Carving with Validation)

Fragment Recovery Carving:
Attempts to reassemble fragmented files

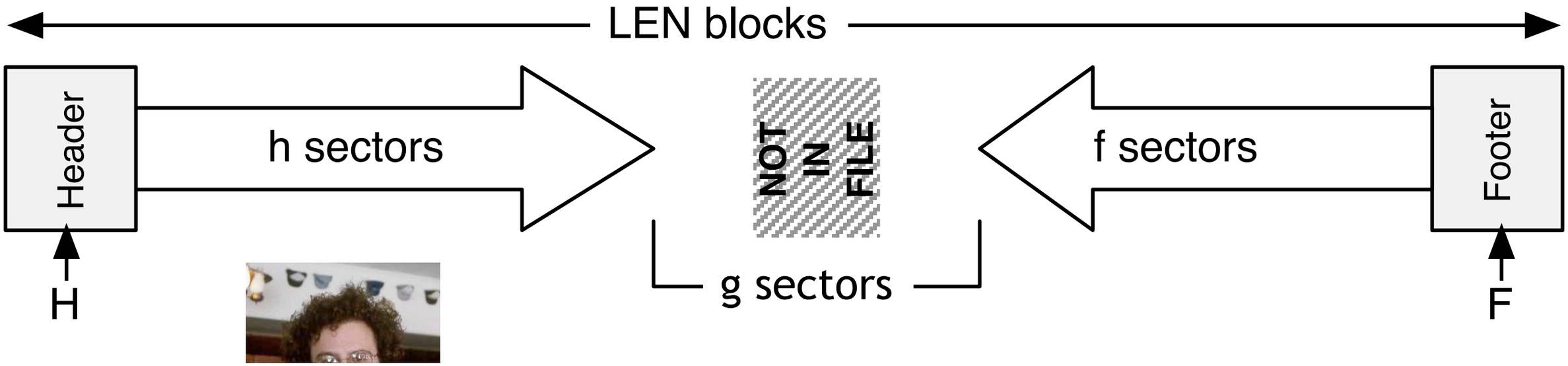
Fragment Recovery Carving:




```
LEN = S-F+1
for I in range(0,LEN):
    for J in range(0,LEN-I):
        data = blocks[S:S+I] + blocks[F-J:J]
        if valid(data)==True:  save(data)
```

Fragment Recovery Carving:
Attempts to reassemble fragmented files

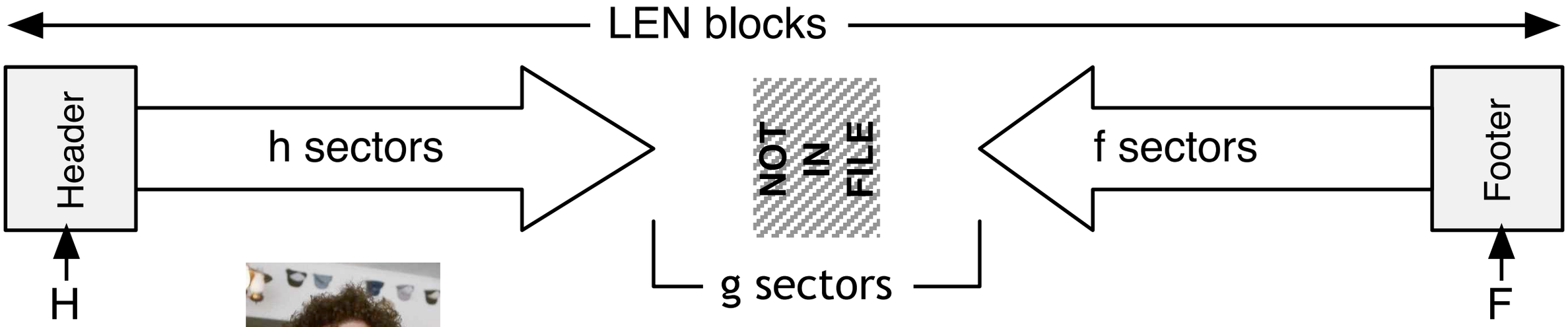
Fragment Recovery Carving:




```
LEN = S-F+1
for I in range(0,LEN):
    for J in range(0,LEN-I):
        data = blocks[S:S+I] + blocks[F-J:J]
        if valid(data)==True:  save(data)
```

Fragment Recovery Carving:
Attempts to reassemble fragmented files

Fragment Recovery Carving:




```
LEN = S-F+1
for I in range(0,LEN):
    for J in range(0,LEN-I):
        data = blocks[S:S+I] + blocks[F-J:J]
        if valid(data)==True:  save(data)
```

A significant number of files are fragmented on production systems.

	FAT[1]	NTFS	UFS
# File systems:	219	51	5
# Fragments	Number of Files		
(contiguous)	1,285,975	502,050	70,222
2	25,151	20,851	10,932
3	4,929	5,622	1,047
4	2,473	3,176	408
5–10	4,340	11,730	658
11–20	1,591	7,001	94
21–100	1,246	10,912	13
101–1000	185	5,672	0
1001–	2	567	0
Total Files:	1,325,892	567,581	83,374

Files with two fragments are easiest to reassemble.

Ext	file count	Size of files with 2 fragments: avg	stddev	max
pnf	7,583	41,583	81,108	1,317,368
dll	7,479	221,409	384,758	9,857,608
html	3,417	28,388	66,694	2,505,490
jpeg	2,963	29,673	178,563	6,601,153
gif	2,566	22,133	99,370	3,973,951
exe	2,348	399,528	4,354,053	206,199,144
1	1,125	57,475	130,630	1,998,576
dat	780	291,407	673,906	7,793,936
z	716	74,353	340,808	6,248,869
h	690	16,444	12,232	110,592
inf	683	79,578	101,448	522,916
wav	575	1,949,459	6,345,280	39,203,180
swf	548	62,582	120,138	1,155,989
ttf	540	163,854	649,919	10,499,104
sys	513	1,276,323	12,446,966	150,994,944
txt	480	33,410	275,641	5,978,896
hlp	475	185,259	375,461	3,580,078
tmp	450	206,908	772,290	8,388,608
so	440	103,939	205,617	1,501,148
wmf	418	48,864	49,869	586,414
...	...	...	...	

Table 7: Most common files in corpus consisting of two fragments, by file extension.

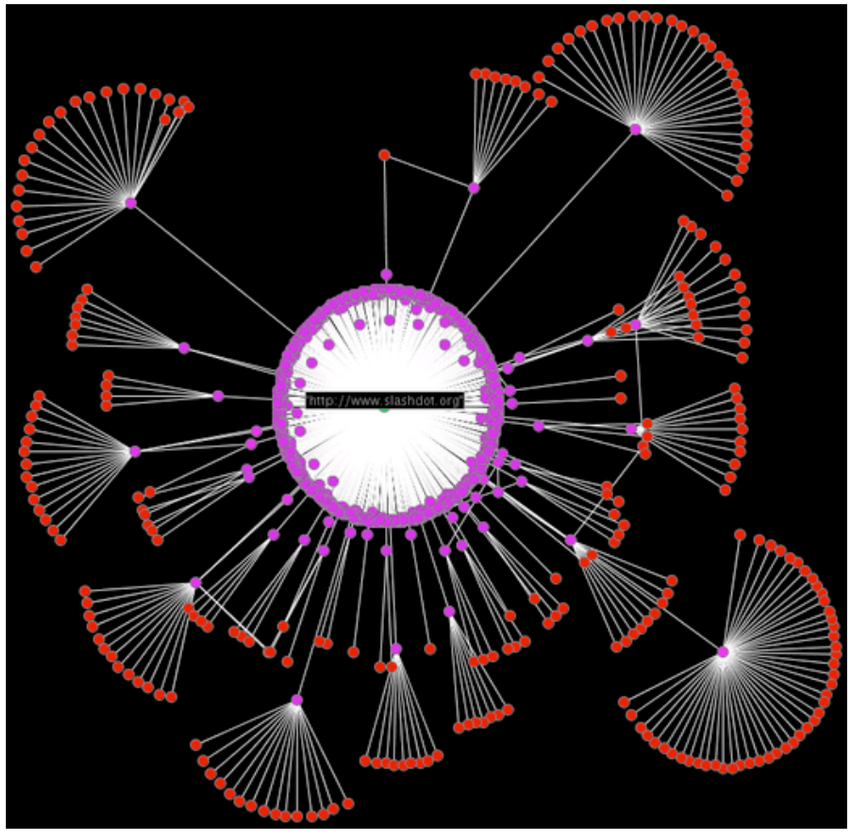
Bi-fragmented JPEG files tend to fragment at specific sizes.

# Files	fragments gap bytes	fragments gap sectors
88	4,096	8
64	2	0
40	16,384	32
38	8,192	16
38	651,264	1,272
23	59	0
16	32,768	64
14	24,576	48
14	20,480	40
11	12,288	24
11	122,880	240
11	131,072	256
11	28,672	56

Table 5: Gap distribution for JPEG bifragmented files.

New Research:
Using Data Mining for Carved File Ascription

Problem: Carved files on multi-user computers need to be ascribed to individual authors.



Proposed Solution:

- Instance-based learning
- Files in the file system = exemplars

In Conclusion...

The Drive's Project has created a unique resource:

- More than 1000 disk drive images from around the world.
- 100+ are non-US Persons and can be used inside USG with minimal restrictions.

Building the corpus is a challenge; using it is a lot of fun.

Questions?