Automated Digital Forensics

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

October 1, 2010

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)
- *All students are fully funded*

Schools:

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

Automated Computer Forensics: The need

Law enforcement & military agencies encounter substantial amounts of electronic media.

Typical media includes:

- Desktop & Laptop computers (hard drives)
- Cell phones (SIM chips, flash memory)
- iPods & MP3 music players

Typical sources includes:

- Media collected on the battle field:
  - —*houses & apartments*
  - —*on the person*
- Border searches
- "Found equipment."
- Domestic searches
- Cyber security
  - —*victim systems*
  - —*attacker systems*
  - —*intermediaries*

Forensic tools are used to examine the media.

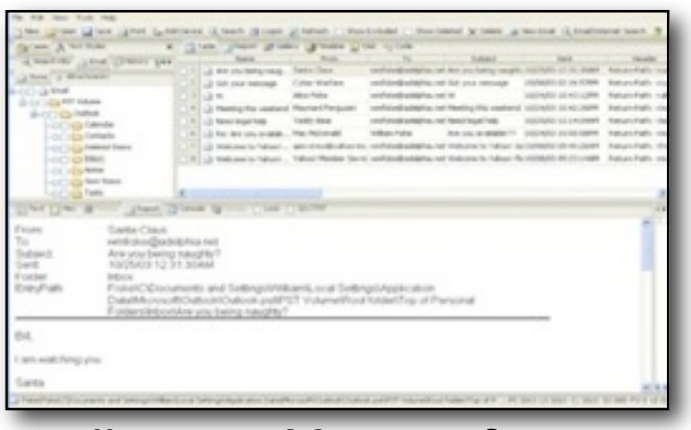
Imaging Tools extract the data without modification.

- "Forensic copy" or "disk image."
- Original media is stored in an evidence locker.
- (Not so easy with cell phones.)
  - —*No standard way to image*
  - —*Difficult to store cell phones without changing them.*

http://www.spacesaver.com/

Analysts then use forensic tools to analyze the copy:

- View *allocated* & *deleted* files.
- String search.
- View individual disk sectors in hex, ASCII and Unicode
- Data recovery and *file carving*
  - —*search for info not in file system.*
  - —*Typically used for Images and Movies*

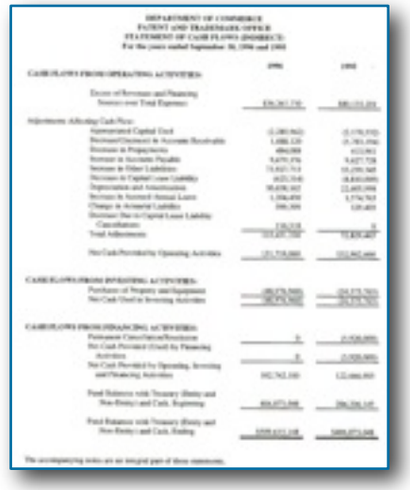


http://www.guidancesoftware.com/

There are different goals for forensic examinations.

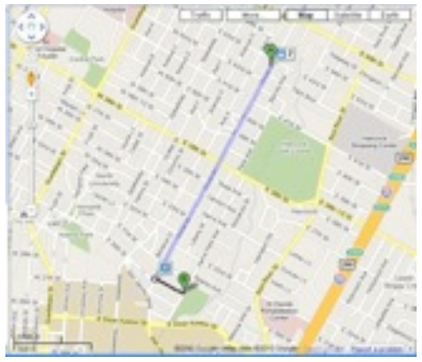
Examiner looks for *evidence of a crime* to support a *conviction:*

- Financial Records.
- Photographs of a murder.
- Child pornography.
- Emails documenting a conspiracy.
- Copy of an emailed threat.



Examiner looks for *new information* to support an *investigation:*

- Associates & accomplices.
- Geographical locations.
- New victims.



Examiner tries to understand *what an intruder did (Cybersecurity):*

- Computer as crime scene.

The last decade was a "Golden Age" for digital forensics.

Widespread use of Microsoft Windows, especially Windows XP

Relatively few file formats:

- Microsoft Office (.doc, .xls & .ppt)
- JPEG for images
- AVI and WMV for video

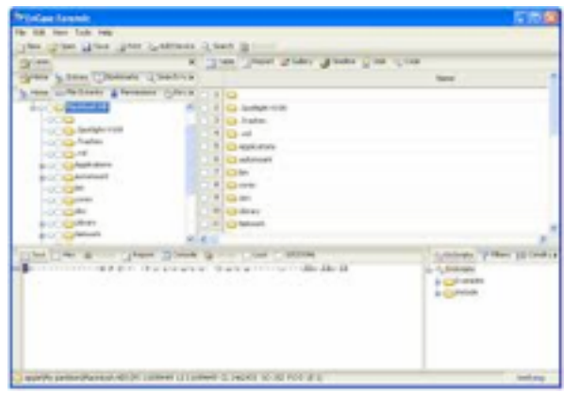
Most examinations confined to a single computer belonging to a single subject

Most storage devices used a standard interface.

- IDE/ATA
- USB

The Golden Age gave us good tools and rapid growth.

Commercial tools:



EnCase®

Open Source Tools:

The Sleuth Kit

Content Extraction Toolkits:

Oracle Outside In Technology

Outside In Technology is a suite of software development kits (SDKs) that provides developers with a comprehensive solution to access, transform and control the contents of over 500 unstructured file formats. Each SDK within the suite is optimized to solve a particular problem but they are highly flexible and interoperable. Developers can quickly implement any combination of the Outside In SDKs to provide exactly the right functionality in their application while minimizing integration effort and code footprint. The SDKs offer a wide range of options to give the developer programmatic control of their workflow and output. Thorough documentation and sample applications with source code are included to further accelerate implementation.

Download | Documentation | Sample Code

But today there is a growing digital forensics crisis.

Much of the last decade's progress is quickly becoming irrelevant.

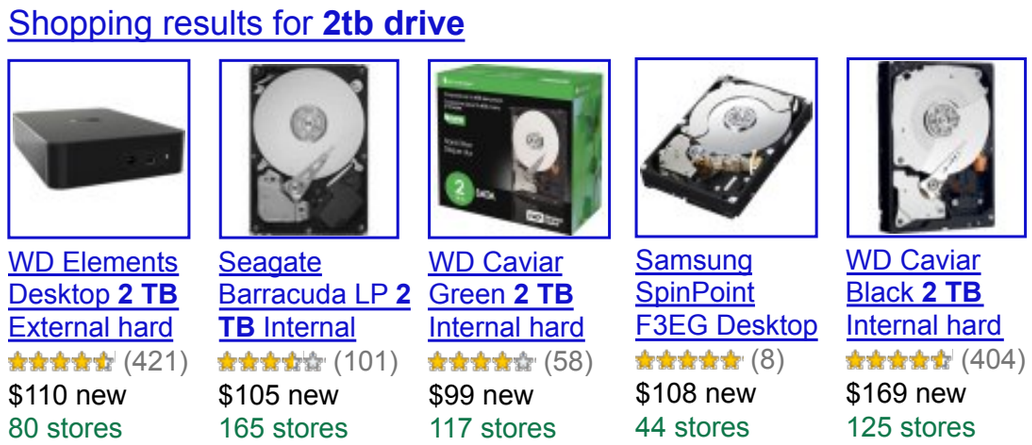
Tools designed to let an analyst find a file and take it into court...

… don't scale to today's problems.

Problem 1 - Dramatically increased cost of extraction & analysis.

Today there is too much data and it getting harder to analyze.

- Increased size of storage systems.



- Cases now require analyzing multiple devices
  - *Typical — 2 desktops, 6 phones, 4 iPods, 2 digital cameras*

- Non-Removable Flash

- Proliferation of operating systems, file formats and connectors
  - *XFAT, XFS, ZFS, YAFFS2, Symbian, Pre, iOS,*

Consider FBI Regional Computer Forensic Laboratories growth:

- Service Requests 5,057 (FY08) ➔ 5,616 (FY09)
- Terabytes Processed: 1,756 (FY08) ➔ 2,334 (FY09)

Problem 2 — Mobile Phones are really hard to examine.

Forensic examiners established bit-copies as the gold standard.

- … but to image an iPhone, you need to jail-break it.
- Is jail-breaking forensically sound?

How do we validate tools against thousands of phones?

How do we forensically analyze 100,000 apps?

No standardized cables or extraction protocols.

NIST's *Guidelines on Cell Phone Forensics* recommends:

- "searching Internet sites for developer, hacker, and security exploit information."

Problem 3 — Encryption and Cloud Computing make it hard to get to the data

Pervasive Encryption — Encryption is increasingly present.

- TrueCrypt
- BitLocker
- File Vault
- DRM Technology

Cloud Computing — End-user systems won't have the data.

- Google Apps
- Microsoft Office 2010
- Apple Mobile Me

Problem 4 — RAM and hardware forensics is really hard.

RAM Forensics—in its infancy

- RAM structures change frequently (no reason for them to stay constant.)
- RAM is constantly changing.

Malware can hide in many places:

- On disk (in programs, data, or scratch space)
- BIOS & Firmware
- RAID controllers
- GPU
- Ethernet controller
- Motherboard, South Bridge, etc.
- FPGAs

Problem 5 — Time is of the essence.

Most tools were designed to perform a complete analysis.

- Find all the files.
- Index all the terms.
- Report on all the data.

Increasingly we are racing the clock:

- Police prioritize based on statute-of-limitations!
- Battlefield, Intelligence & Cyberspace operations require turnaround in days or hours.

Tools and training simply can't keep up.

We can't hire & train fast enough:

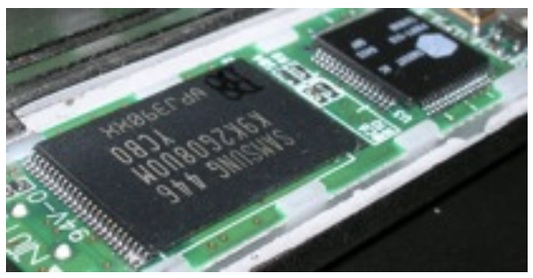
- Not enough highly skilled people.
- Training takes 2 years

1 — Increased costs of extraction and analysis
2 — Mobile Phones
3 — Encryption and Cloud Computing
4 — RAM and Hardware Forensics
5 — Time

Fundamental problem:

- training skills linearly,
- but the problems scale geometrically.

Some devices will never be supported by today's mainstream tools.



My research focuses on three main areas:

Area #1: Data Collection and Manufacturing

- Large data sets of real data enable *science.*
- Small data sets of realistic data enable *education, training and publishing.*

Area #2: Bringing data mining and machine learning to forensics

- Breakthrough algorithms based on *correlation* and *sampling*
- Automated social network analysis (cross-drive analysis)
- Automated ascription of carved data

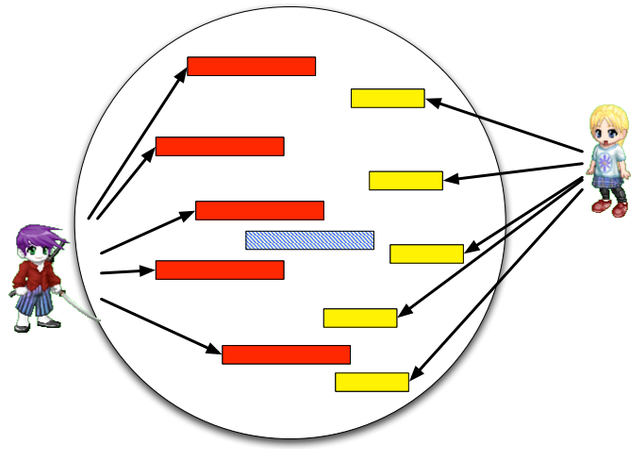
Area #3: Tools that are composable, automated, and open source

- Digital Forensics XML (DFXML) for connecting tools.
- Advanced Forensic Format (AFF) for storing digital evidence.
- bulk_extractor fast feature extractor

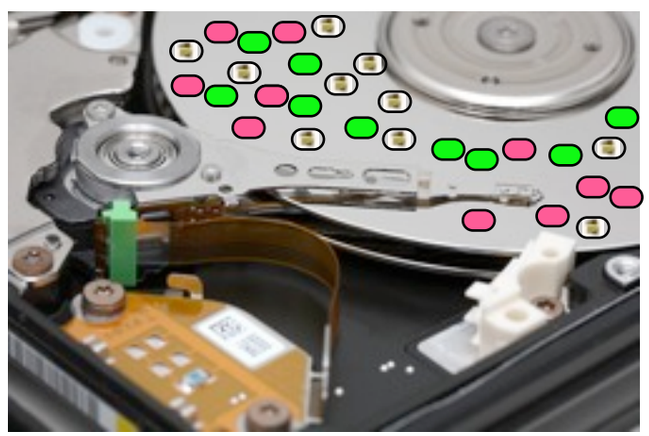
This talk focuses on three key areas:

Standardized Forensic Corpora

Multi-User Carved Data Ascription

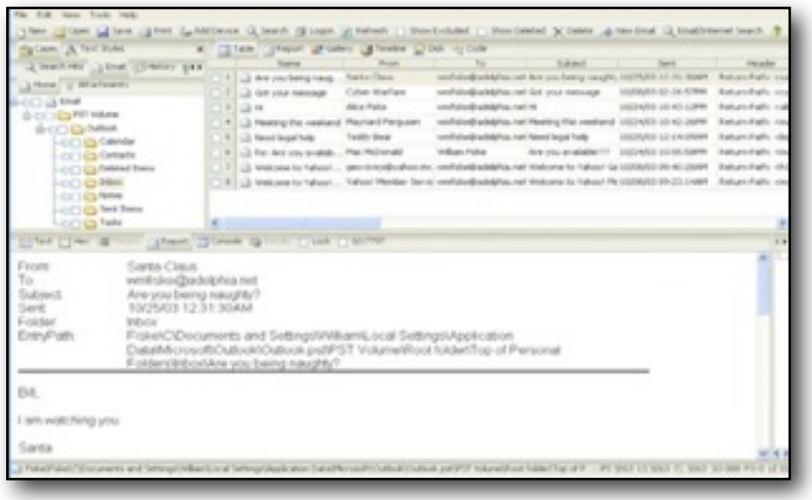


High Speed Forensics



Standardized Forensic Corpora

Digital Forensics is at a turning point.
Yesterday's work was primarily *reverse engineering*.



Key technical challenges:

- Evidence preservation.
- File recovery (file system support); Undeleting files
- Encryption cracking.
- Keyword search.

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

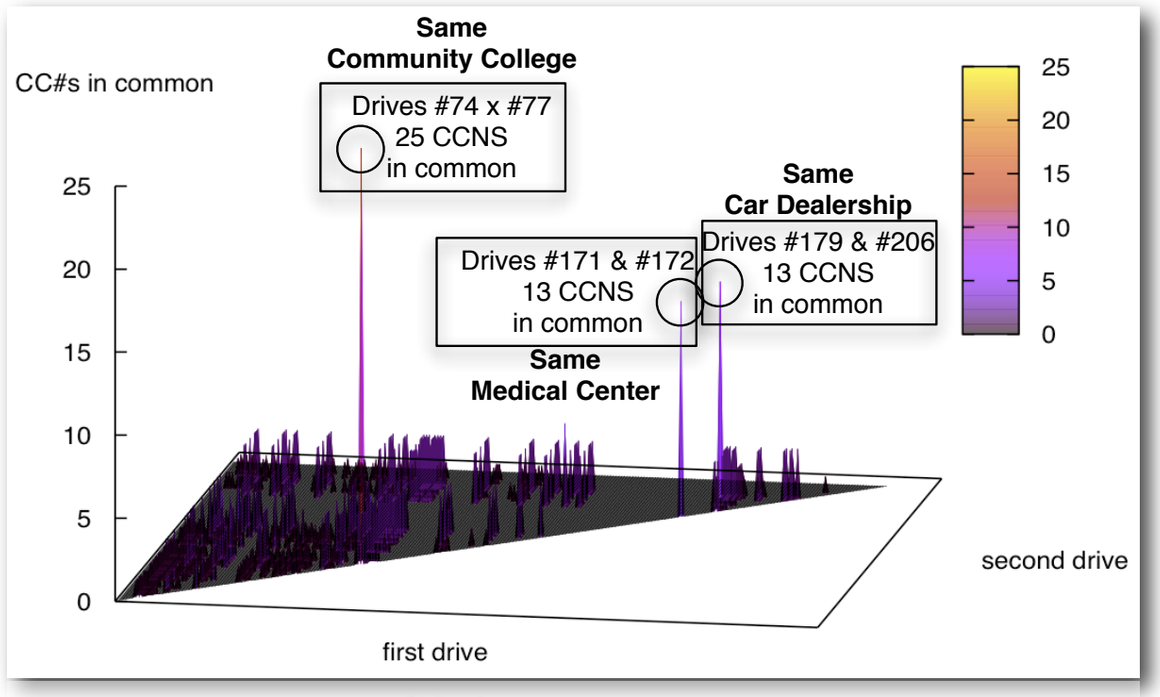
Evidence Reconstruction

- Files (fragment recovery carving)
- Timelines (visualization)

Clustering and data mining

Social network analysis

Sense-making



Science requires the *scientific process.*

Hallmarks of Science:

- Controlled and repeatable experiments.
- No privileged observers.

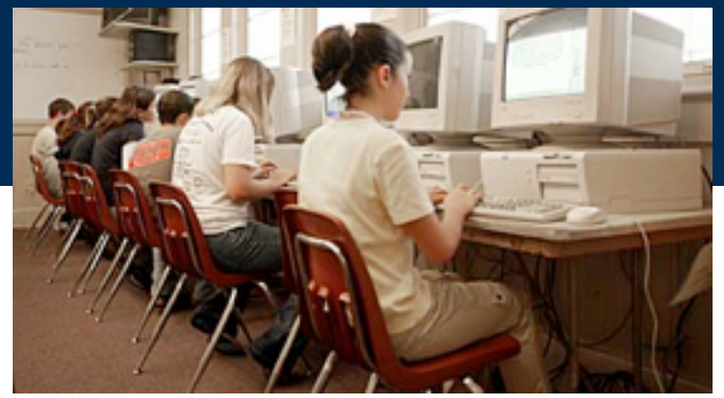
Why repeat some other scientist's experiment?

- Validate that an algorithm is properly implemented.
- Determine if ***your*** new algorithm is better than ***someone else's*** old one.
- (Scientific confirmation? — perhaps for venture capital firms.)

We can't do this today.

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

Digital Forensics education needs corpora too!



Some teachers get used hard drives from eBay.

- Problem: you don't know what's on the disk.
  - —*Ground Truth.*
  - —*Potential for illegal Material.*
    - Distributing pornography to children is illegal.
    - Possibility for child pornography.

Some teachers have students examine other student machines:

- Self-examination: students know what they will find
- Examining each other's machines: potential for inappropriate disclosure

Also: IRB issues

United States Department of Health & Human Services

We are making available several types of corpora.

Files from US Government Web Servers (500GB)

- ≈1 million files
- Freely redistributable; Many different file types

Test and Realistic Disk Images (1TB)

- Mostly Windows operating system.
- Some with complex scenarios to facilitate forensics education.

The Real Data Corpus (20TB)

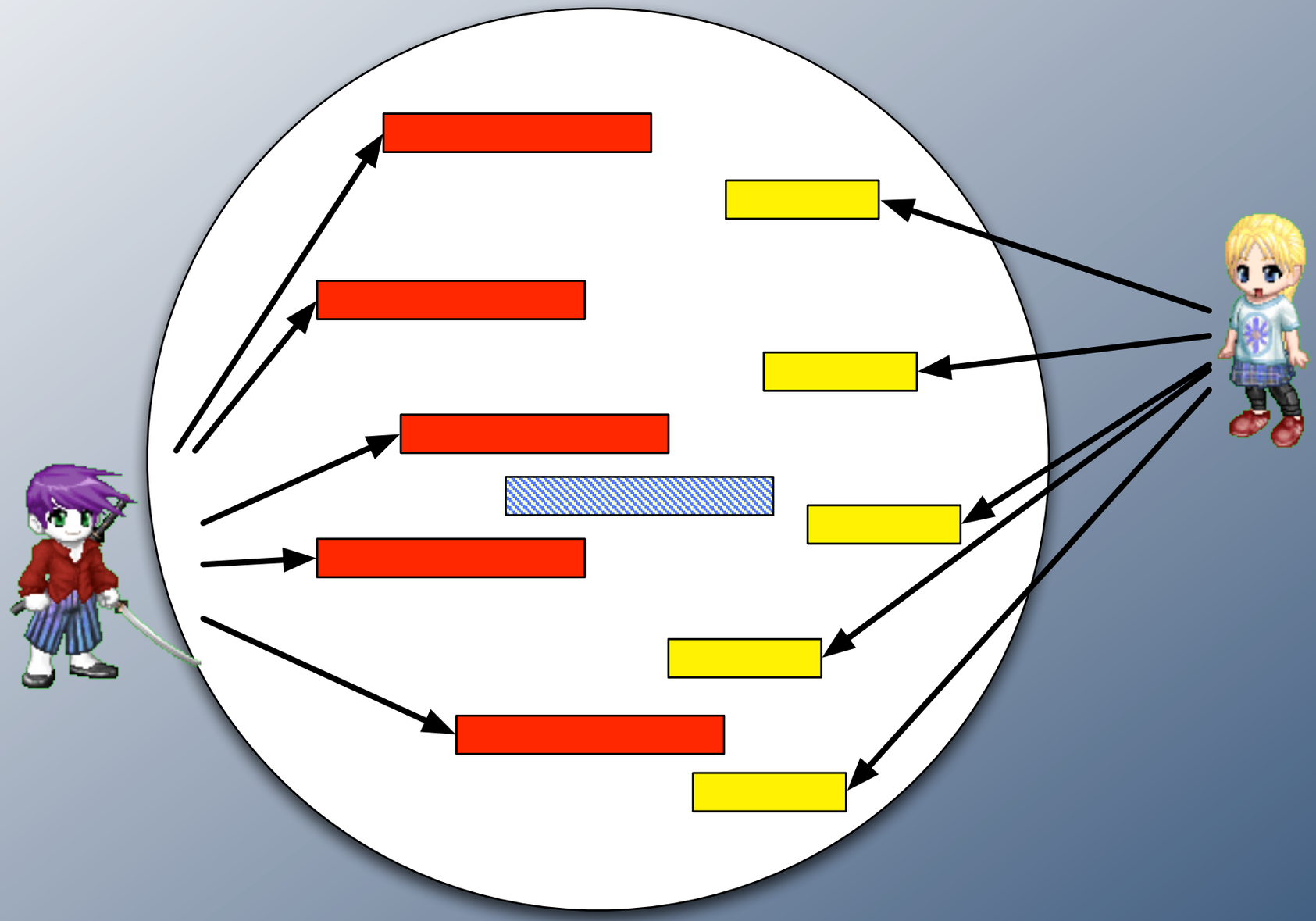
- Disks, camera cards, & cell phones purchased on the secondary market.
- Most contain data from previous users.

Mobile Phone Application Corpus

- Android Applications; Mobile Malware; etc.

—*Garfinkel, Farrell, Roussev and Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, Best Paper, DFRWS 2009*
http://digitalcorpora.org/

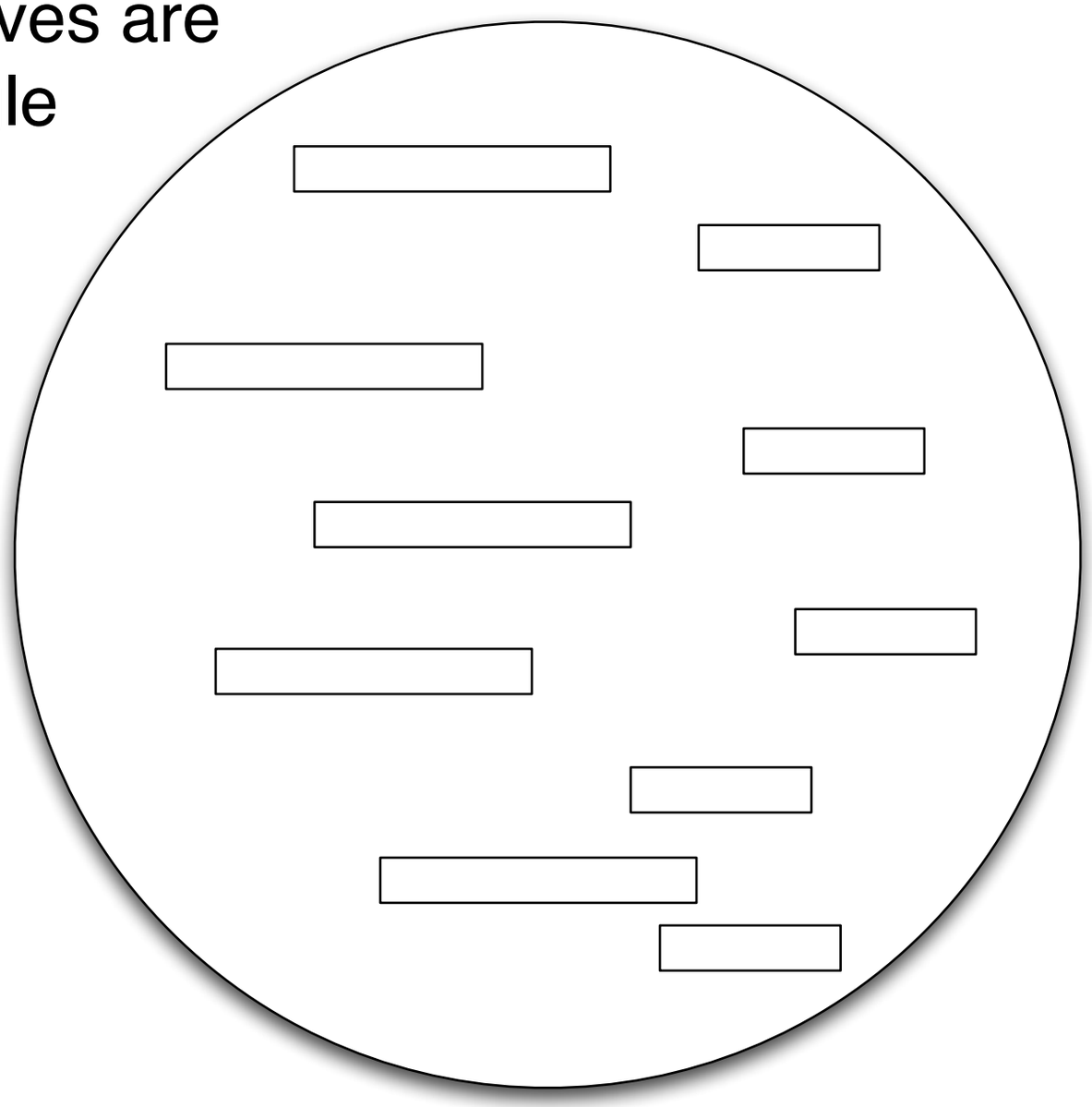
We will use this data in the rest of this talk.



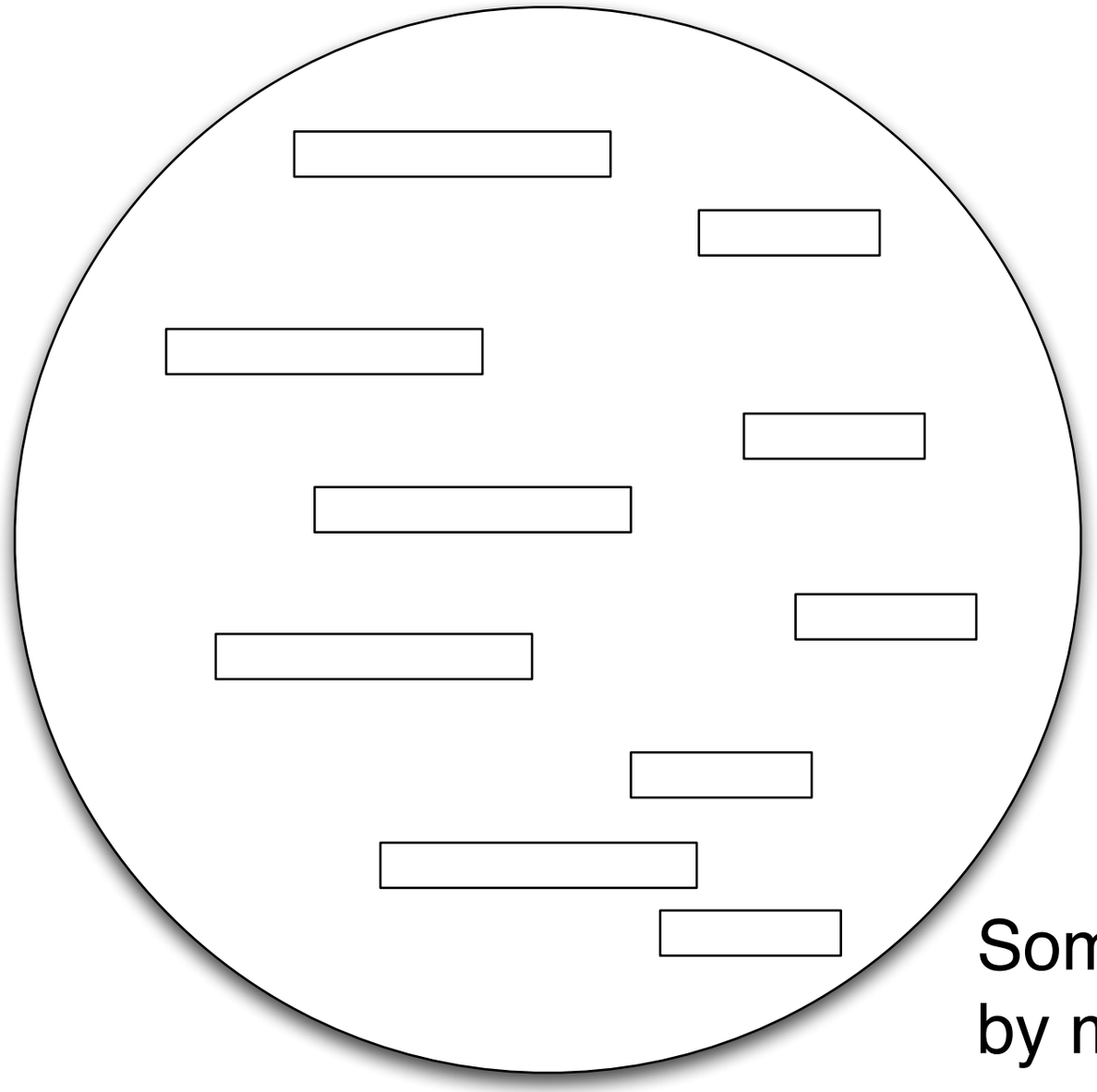
Automated Ascription of Multi-User Data

Disks may have any number of recoverable files. 0 to 1,000,000 is common.

Some hard drives are used by a single person.

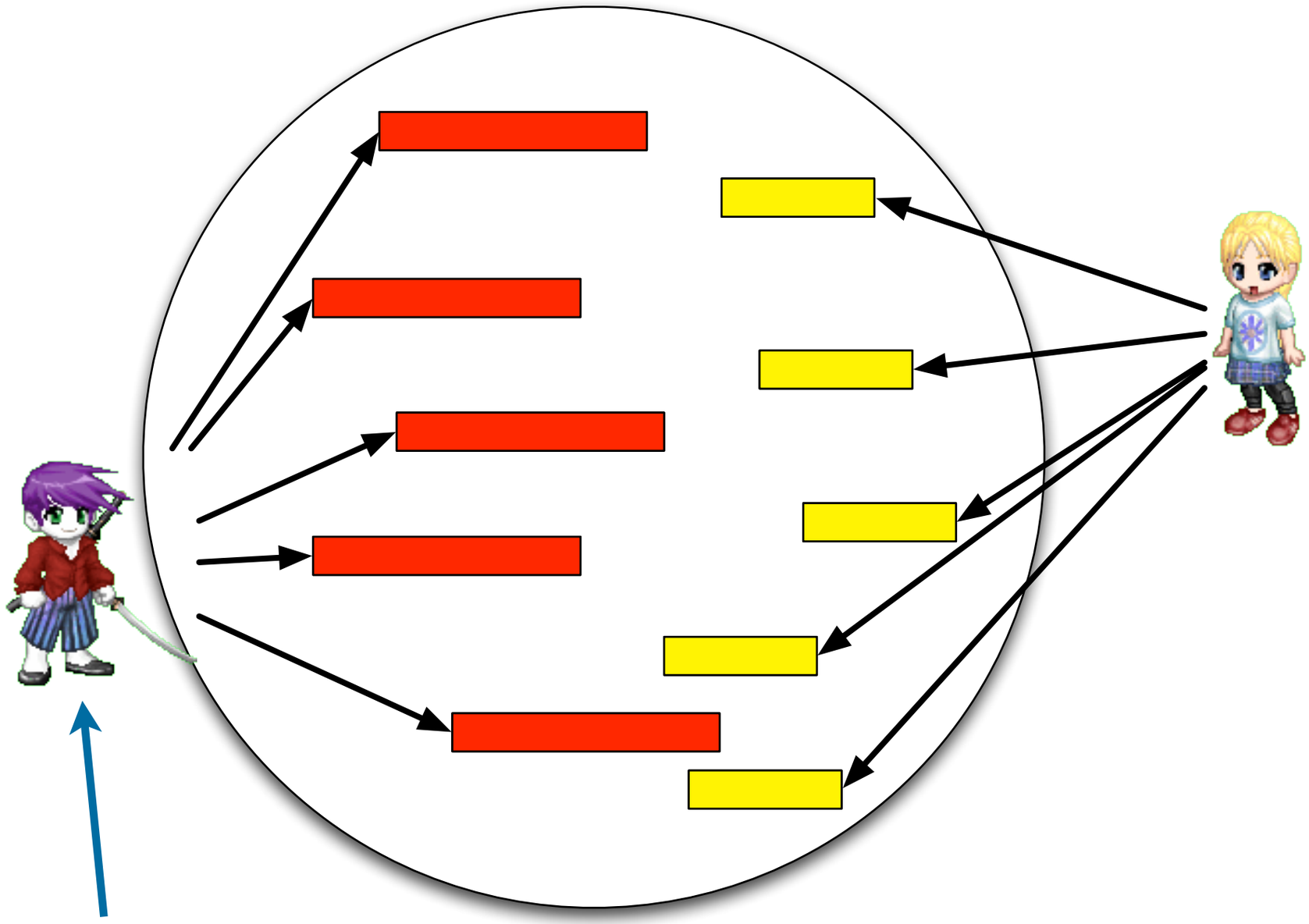


Disks may have any number of recoverable files.
0 to 1,000,000 is common.



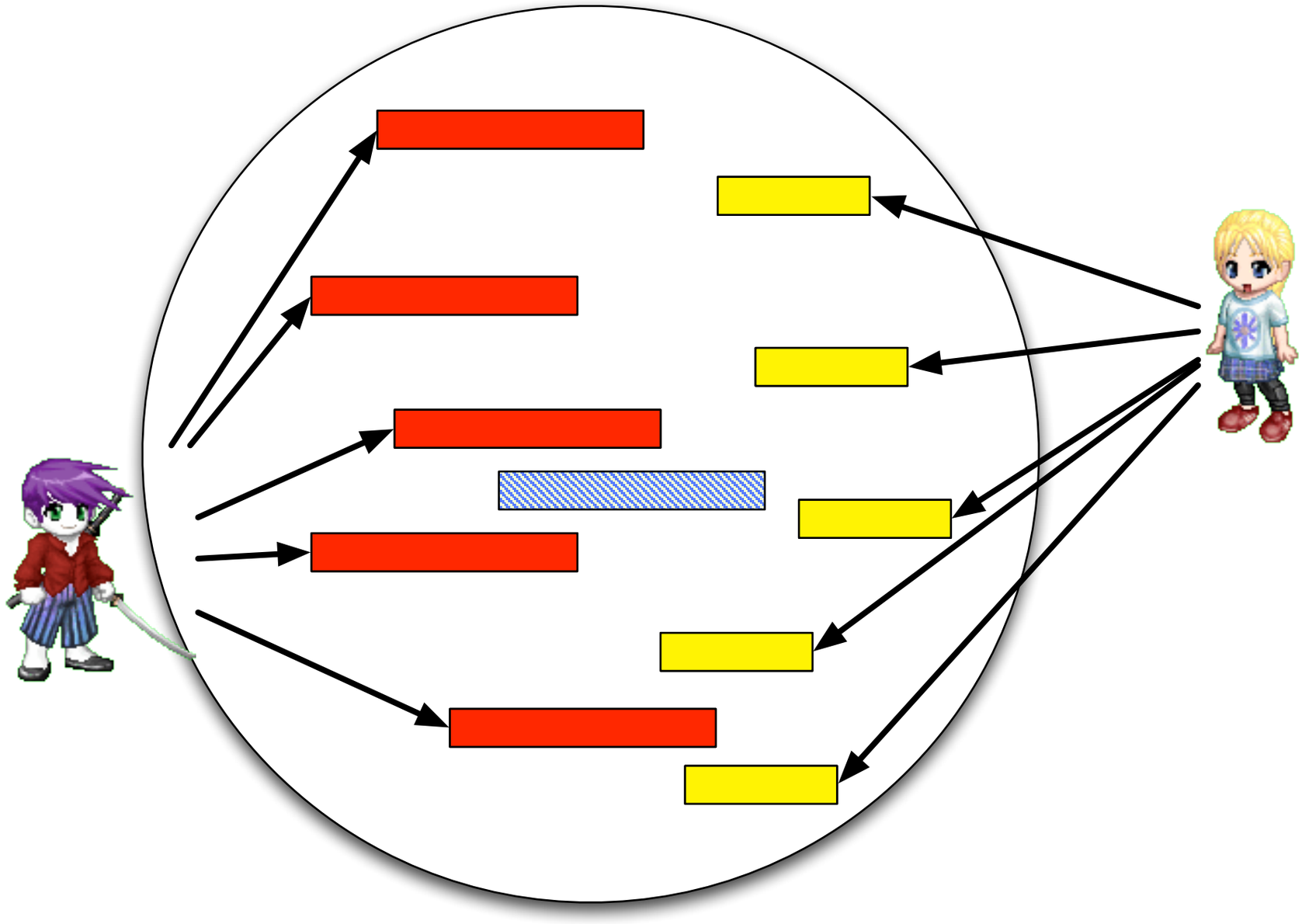
Some drives are used by multiple people.

When a disk is used by multiple people, file owner or path is typically used to ascribe ownership.



/Documents and Settings/Magenta/Cookies/magenta@www.microsoft[1].txt

Files recovered with “carving” can’t be readily ascribed.



Who is responsible for the file?

Prior work has used *content analysis* to determine authorship

Trait		
“Reading Level”	8th Grade	College
Characteristic Errors	JUmp higher. FLy high.	Skilz Killz Spilz

This project uses metadata to infer *ownership* or *agency* — who is *responsible* for the data.

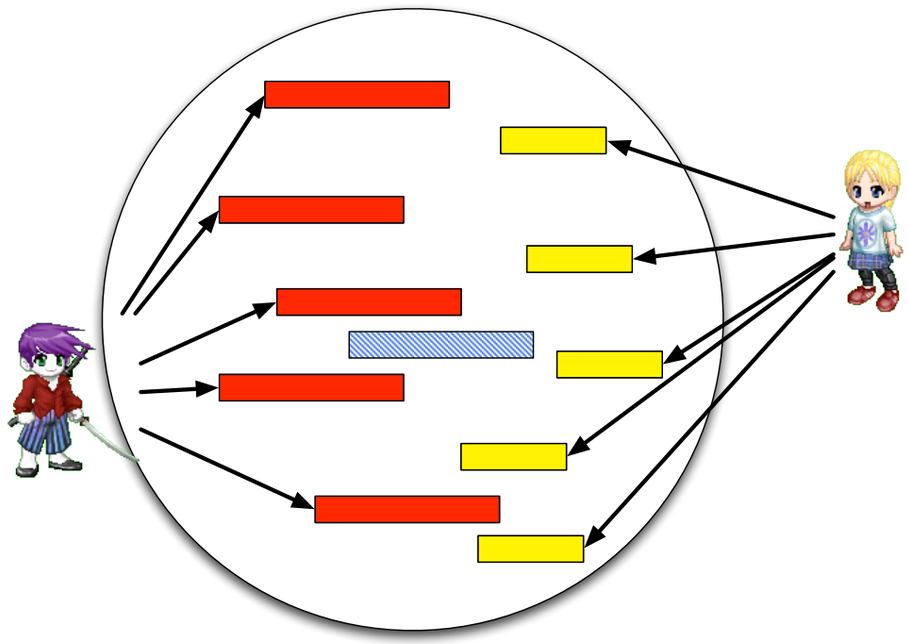
File system metadata ("extrinsic metadata"):

- Fragmentation patterns (disk usage)
- Where the file is on the hard drive (sector numbers)
- Timestamps for "orphan" files.

File metadata ("intrinsic metadata"):

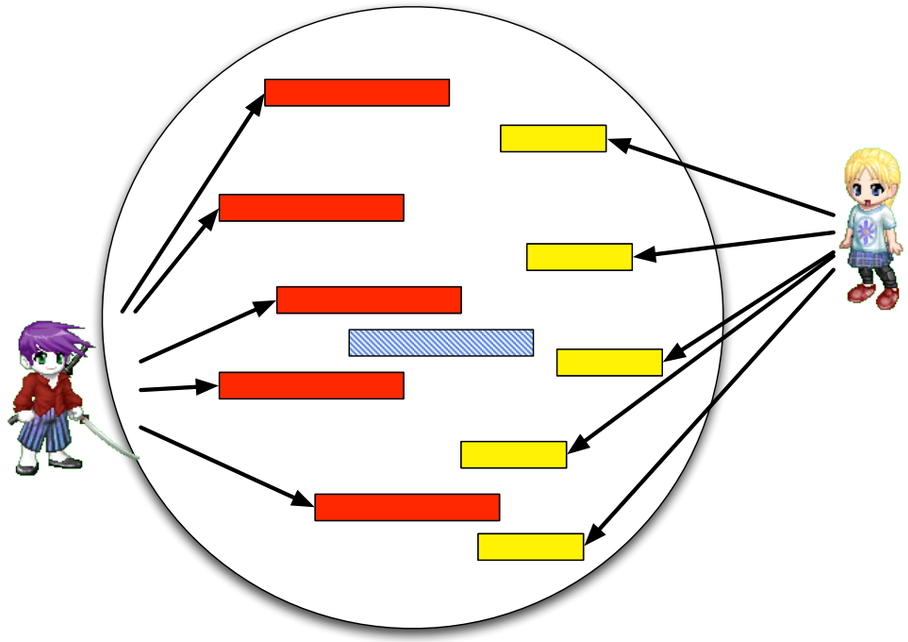
- Embedded timestamps
  - —*Creation Time*
  - —*Print Time*
- Make & model of digital cameras
- Usage patterns.

Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.



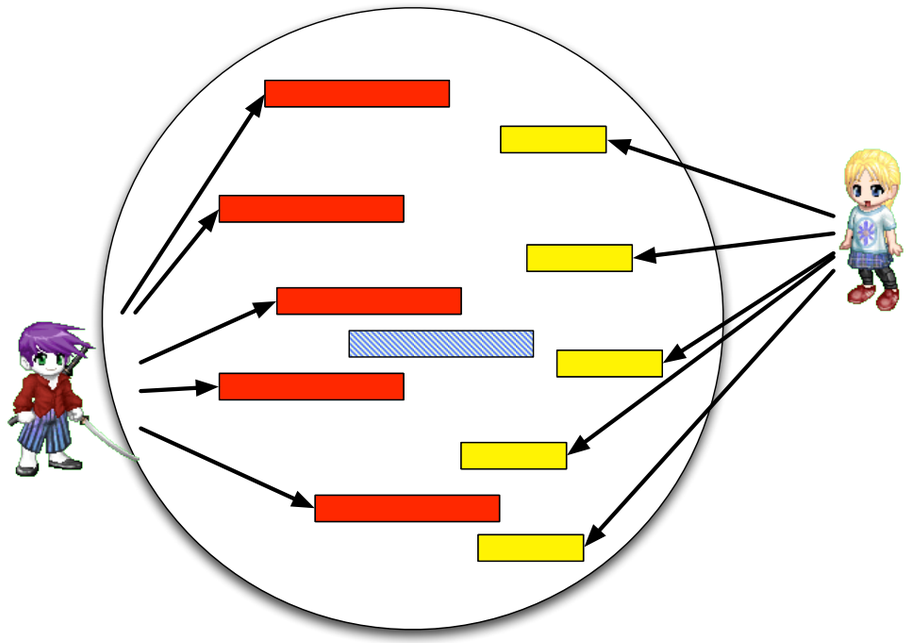
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User



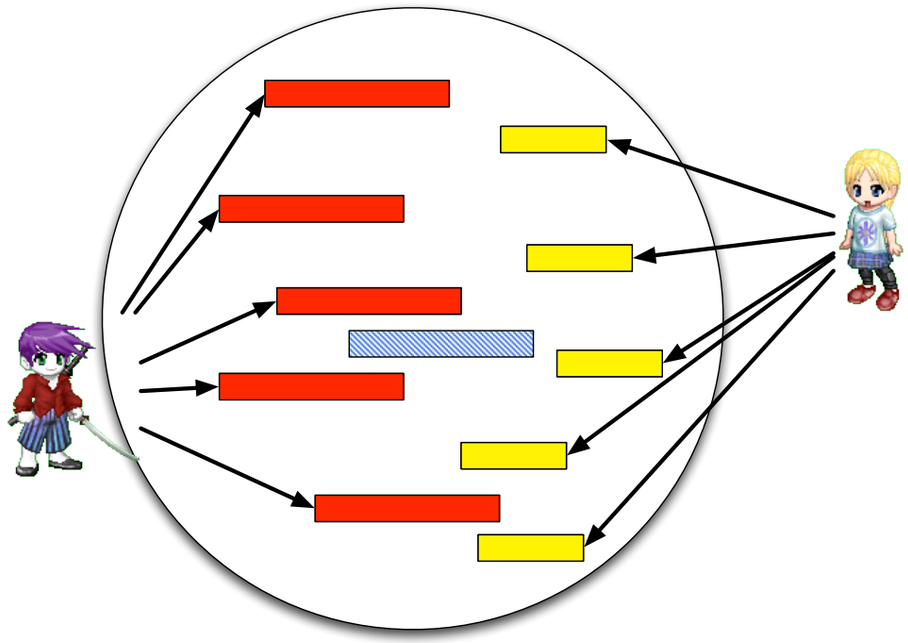
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML		



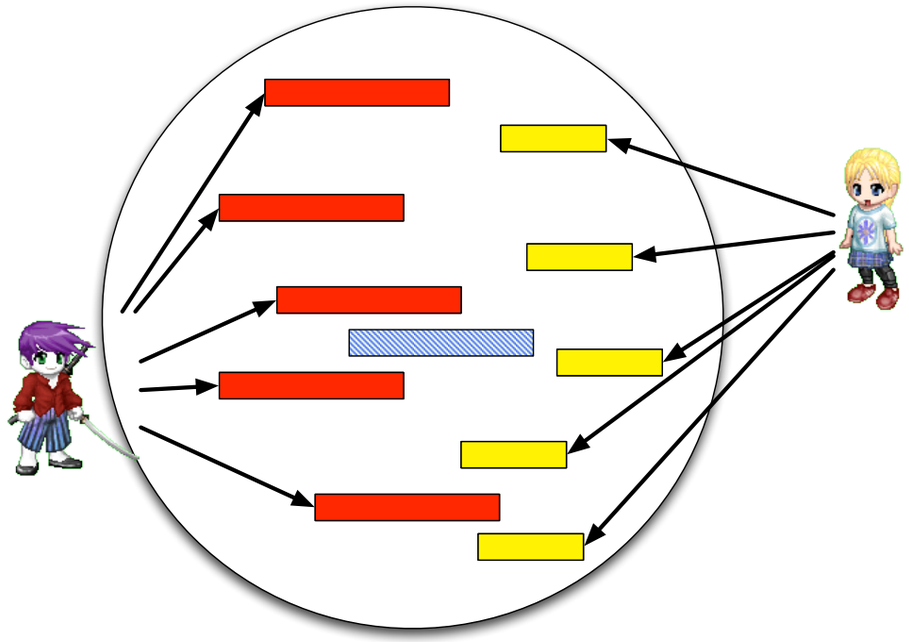
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	



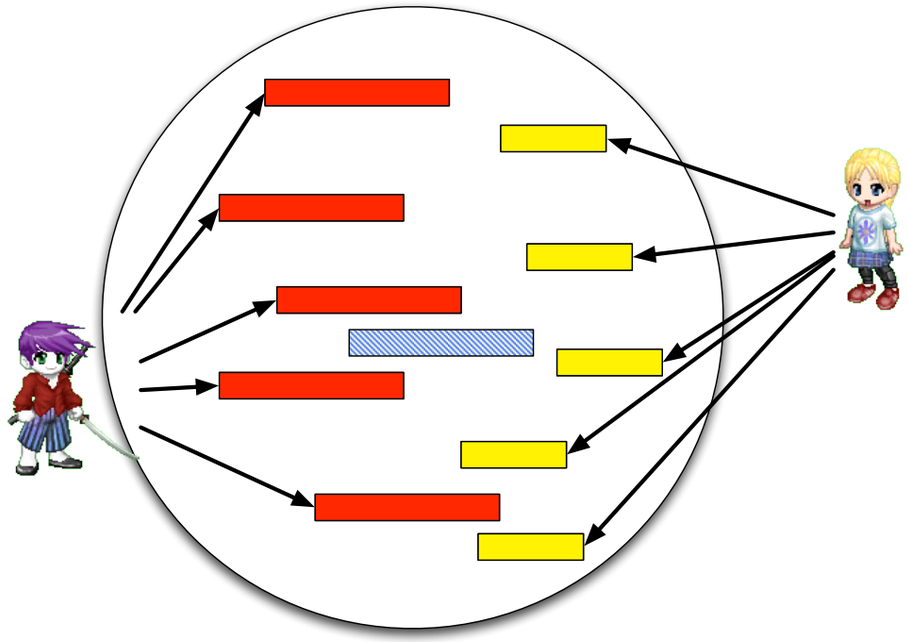
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	



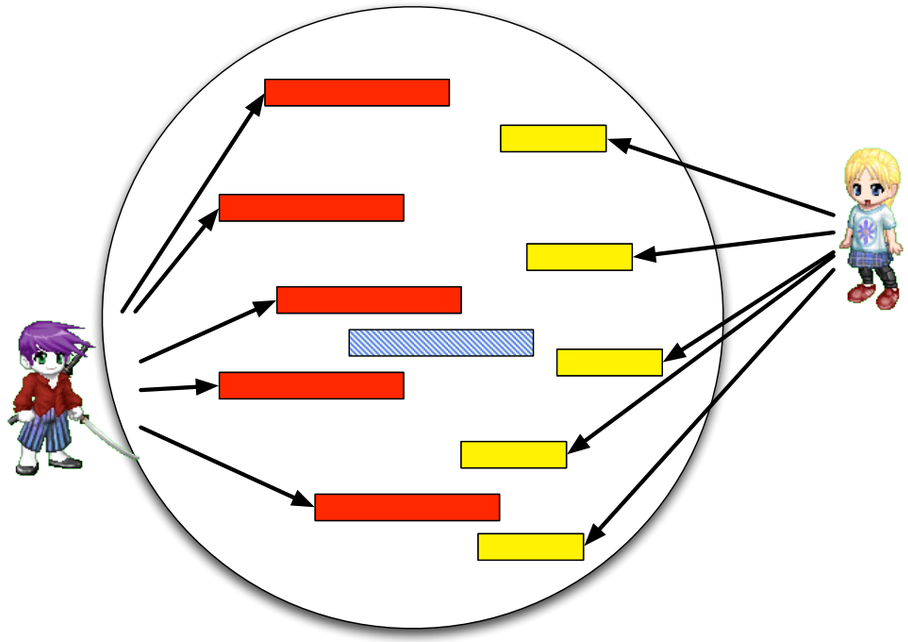
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm		



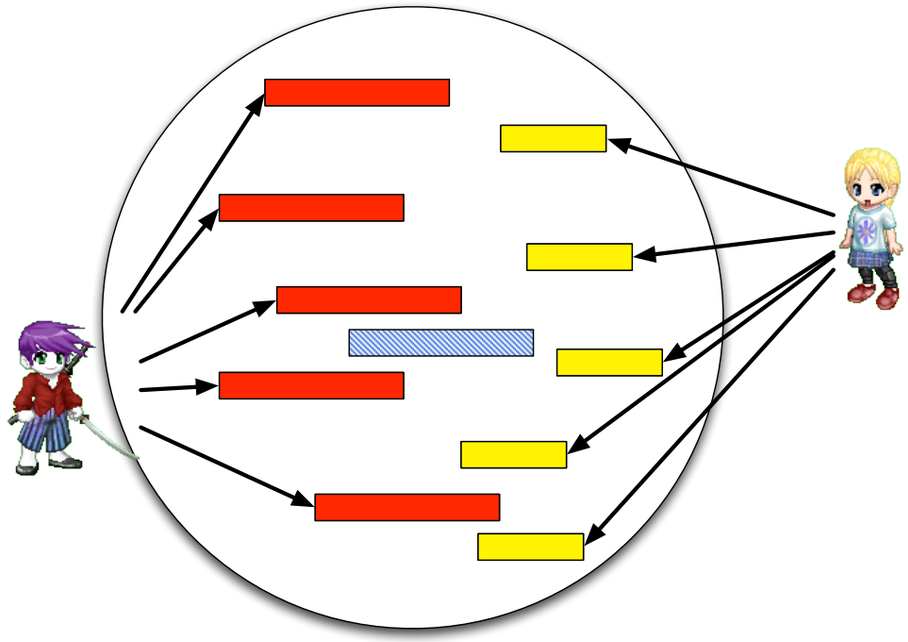
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	



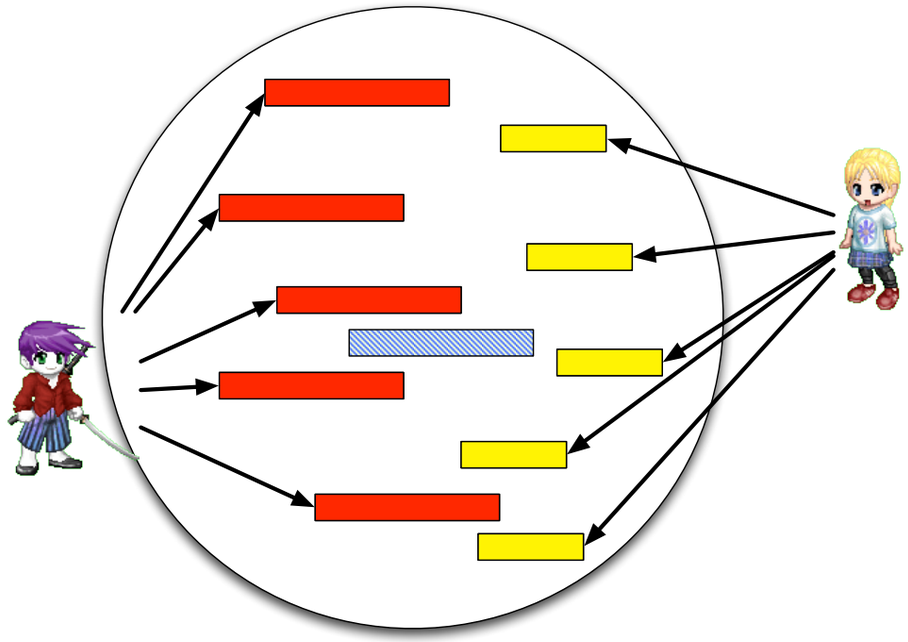
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	



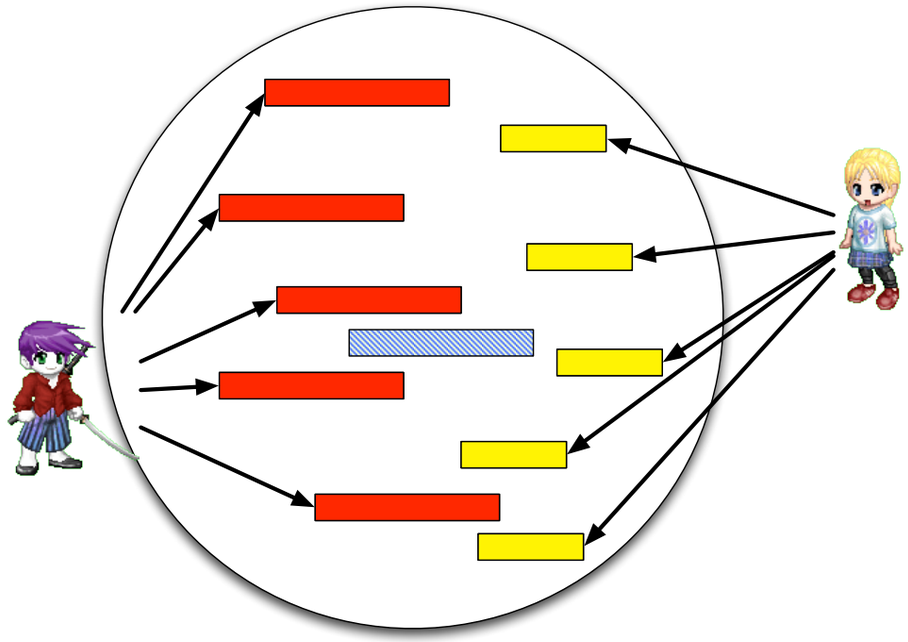
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	
At sectors 10, 20, 30	At sectors 500, 600, 700		



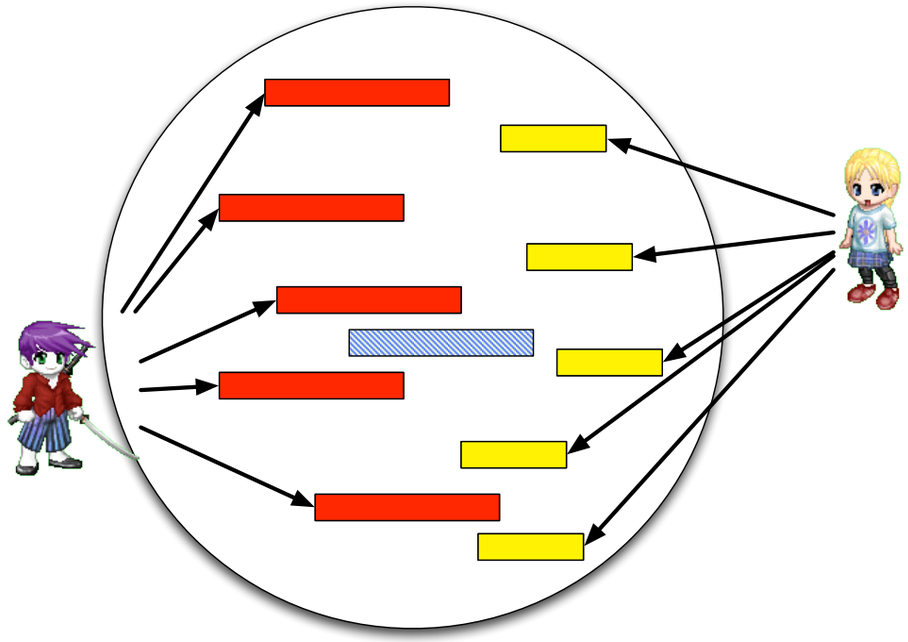
Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	
At sectors 10, 20, 30	At sectors 500, 600, 700	Sector 550	

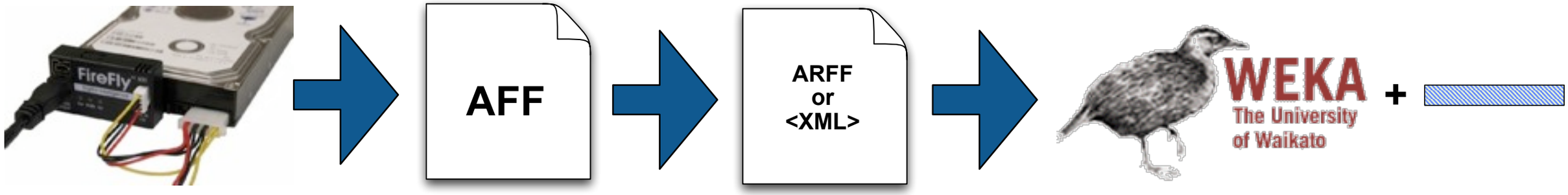


Today some examiners do this manually by surveying the disk for exemplars and looking for patterns.

Magenta	Yellow	Carved	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	
At sectors 10, 20, 30	At sectors 500, 600, 700	Sector 550	



We developed a tool set for automated ascription.



Step 1: Extract all files and file *metadata*

- File Owner (from filename or metadata)
- All files: Location on disk
- JPEGs: Camera Serial Number
- Word Documents: Author, Last Edit Time, Print Time, etc.

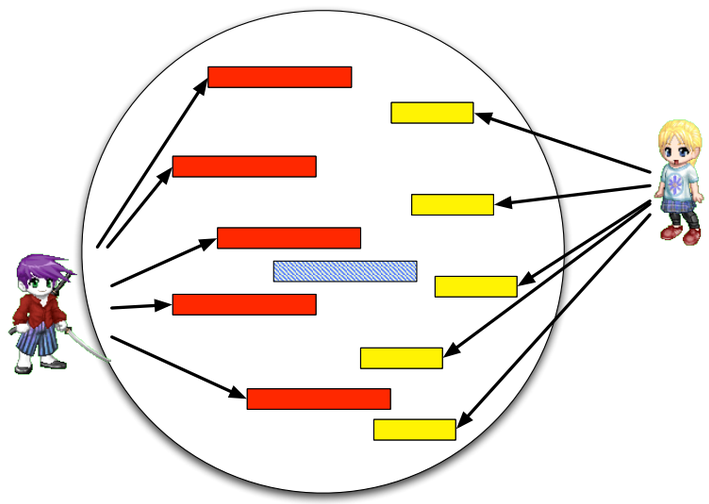
Step 2: Build a classifier using ascribable files as exemplars

Step 3: Use classifier to ascribe carved data.

Several factors complicate this data mining problem.

High dimensionality, heterogeneous data

- ***All files:*** *inode, mode, timestamps, sector #,*
- ***JPEG:*** *Serial Number, f-stop, exposure date*
- ***Word:*** *Author, Print Time, Create Time, etc.*



Sparse data; many missing values

- Every data element is missing values in one or more dimensions!

Multiple regions for each class

- User files interleave in time, space, etc.

Many different time dimensions

- File Print Time; File Modify Time; File Access Time; etc.
- Projecting times onto a "User Activity Timeline" dramatically improves accuracy.

fiwalk is our tool for converting disk images to XML or ARFF files

Per-Image tags

```
<fiwalk> – outer tag
<fiwalk_version>0.4</fiwalk_version>
<Start_time>Mon Oct 13 19:12:09 2008</Start_time>
<Imagefile>dosfs.dmg</Imagefile>
<volume startsector="512">
```

Per <volume> tags:

```
<Partition_Offset>512</Partition_Offset>
<block_size>512</block_size>
<ftype>4</ftype>
<ftype_str>fat16</ftype_str>
<block_count>81982</block_count>
```

Per <fileobject> tags:

```
<filesize>4096</filesize>
<partition>1</partition>
<filename>linedash.gif</filename>
<libmagic>GIF image data, version 89a, 410 x 143</libmagic>
```

fiwalk has a pluggable metadata extraction system

Metadata extractors are specified in the *configuration file*

```
*.jpg    dgi      ../plugins/jpeg_extract
*.pdf    dgi      java -classpath plugins.jar Libextract_plugin
```

- *Currently the extractor is chosen by the file extension*
- *fiwalk runs the plugins in a different process*
- *We have designed a native Java interface that uses IPC and 1 process, but nobody wants to use it.*

Metadata extractors produce name:value pairs on STDOUT

```
Manufacturer: SONY
Model: CYBERSHOT
Orientation: top - left
```

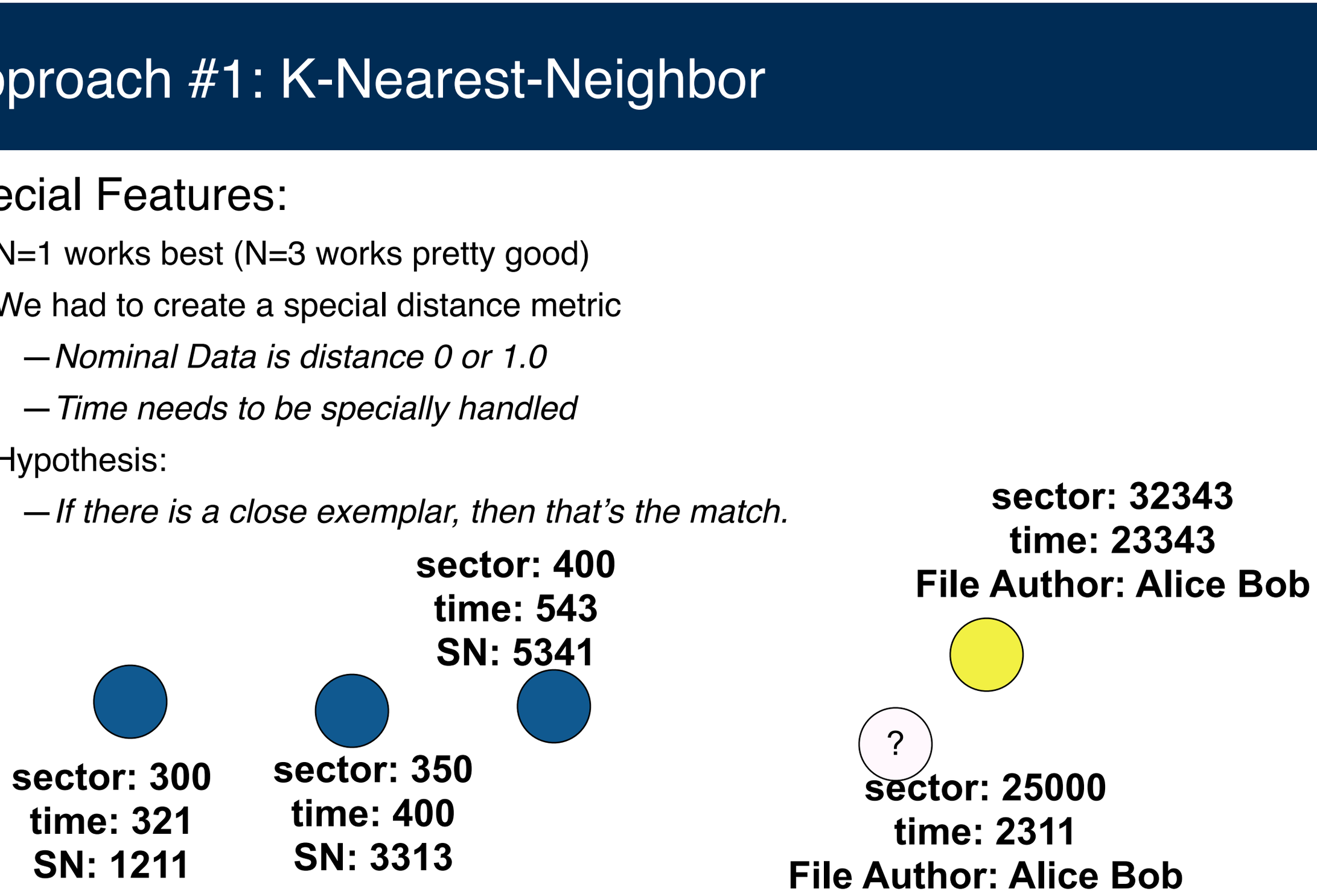
fiwalk incorporates metadata into XML and ARFF:

```
<fileobject>
...
<Manufacturer>SONY</Manufacturer>
<Model>CYBERSHOT</Model>
<Orientation>top - left</Orientation>
...
</fileobject>
```

Approach #1: K-Nearest-Neighbor

Special Features:

- N=1 works best (N=3 works pretty good)
- We had to create a special distance metric
  - *Nominal Data is distance 0 or 1.0*
  - *Time needs to be specially handled*
- Hypothesis:
  - *If there is a close exemplar, then that's the match.*



This approach is easy to explain to a jury!

Approach #2: Decision Tree

Algorithm: C4.5

- Very fast: typically less than 60 seconds.


```
|   inode > 28455
|   |   inode <= 36552
|   |   |   mode <= 365
|   |   |   |   inode <= 28892: magenta (132.0)
|   |   |   |   inode > 28892
|   |   |   |   |   timeline <= 1225239807000: All Users (116.0)
|   |   |   |   |   timeline > 1225239807000
|   |   |   |   |   |   frag1startsector <= 2585095
|   |   |   |   |   |   |   libmagic = ASCII text, with CRLF line terminators
|   |   |   |   |   |   |   |   timeline <= 1225330086000: magenta (8.0)
|   |   |   |   |   |   |   |   timeline > 1225330086000: yellow (8.0)
|   |   |   |   |   |   |   libmagic = data: magenta (16.0)
```

This approach generally provided higher accuracy than KNN.

What do we mean by "accuracy?"

We build a different classifier for every drive!

The only difference between *allocated* data and *carved* data is:

- Carved data is no longer attached to a directory.
- Carved data is likely to be overwritten if the system is heavily used.

We determine the accuracy using take-one-out cross-validation.

- Take-one-out simply moves a file from the "allocated" set to the "carved" set.
- Every HD has its own accuracy.
- Every carved file has its own classifier
  - —*Only use the dimensions that matter for this piece of carved data.*

Results with "realistic" drive created in the lab.

	Classified As							
User	a	b	c	d	e	f	g	total
a “Administrator”	5118	62	0	26	4	7	4	5221
b “All Users”	57	1422	17	32	12	4	0	1544
c “Default User”	1	39	392	0	0	0	4	436
d “domex1”	21	62	0	3051	96	0	0	3230
e “domex2”	24	16	0	94	2335	0	0	2469
f “LocalService”	12	0	0	0	0	64	0	76
g “NetworkService”	2	2	0	0	0	4	48	56
% correct classifications	97.77	88.71	95.84	95.25	95.42	81.01	85.71	

Table 7: domexusers (C4.5) Confusion matrix;

Results with a real drive purchased on the secondary market:

	Classified As														
User	a	b	c	d	e	f	g	h	i	j	k	l	m	n	total
a “Administrator”	320	8	0	1	11	0	3	1	0	0	0	0	0	0	344
b “All Users”	13	971	0	3	11	0	15	7	0	18	3	1	24	2	1068
c (Blinded)	0	0	443	1	1	4	20	0	0	15	4	0	4	0	492
d (Blinded)	0	8	1	288	0	0	82	0	0	0	0	0	1	0	380
e “Default User”	8	36	4	0	343	1	4	0	4	0	4	0	0	0	404
f (Blinded)	0	0	12	0	1	440	4	0	3	0	0	0	0	0	460
g (Blinded)	2	8	22	46	6	11	66540	2	8	23	13	0	7	2	66690
h “LocalService”	5	2	0	0	0	0	4	75	0	0	0	0	2	0	88
i (Blinded)	0	0	1	0	1	0	0	0	707	0	2	0	1	0	712
j (Blinded)	0	12	4	2	0	1	22	1	0	594	0	0	0	0	636
k (Blinded)	0	3	4	0	1	0	8	0	0	0	1204	0	0	0	1220
l “NetworkService”	0	0	0	0	0	0	0	2	0	0	0	54	0	0	56
m (Blinded)	0	16	0	0	0	0	8	6	2	0	0	0	70952	224	71208
n (Blinded)	0	8	0	0	0	1	0	0	2	0	0	0	81	436	528
% correct classifications	91.95	90.58	90.22	84.46	91.47	96.07	99.75	79.79	97.38	91.38	97.89	98.18	99.83	65.66	

Table 9: 0844 (C4.5) Confusion matrix; non-system names are blinded.

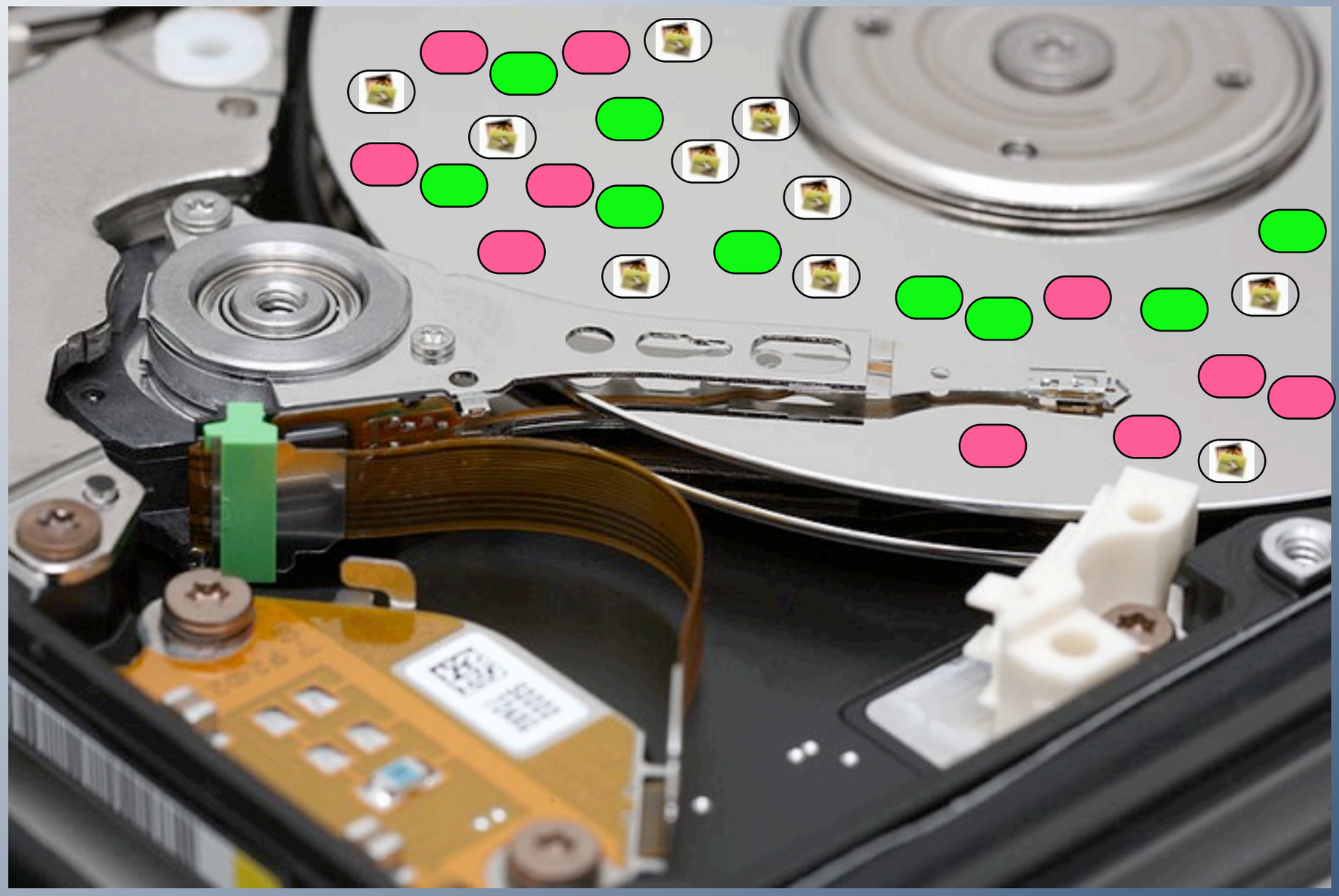
Publications

Student Theses:

- Cpt. Daniel Huynh, “Exploring and Validating Data Mining Algorithms for use in Data Ascription,” June 2008
- Maj. James Migletz, “Automated Metadata Extraction,” June 2008

Articles:

- "An Automated Solution to the Multi-User Carved Data Ascription Problem," Simson L. Garfinkel, Aleatha Parker-Wood, Daniel Huynh and James Migletz, *IEEE Transactions on Information Forensics & Security,* Dec. 2010



High Speed Forensics

Data on hard drives can be divided into three categories:

Resident Data

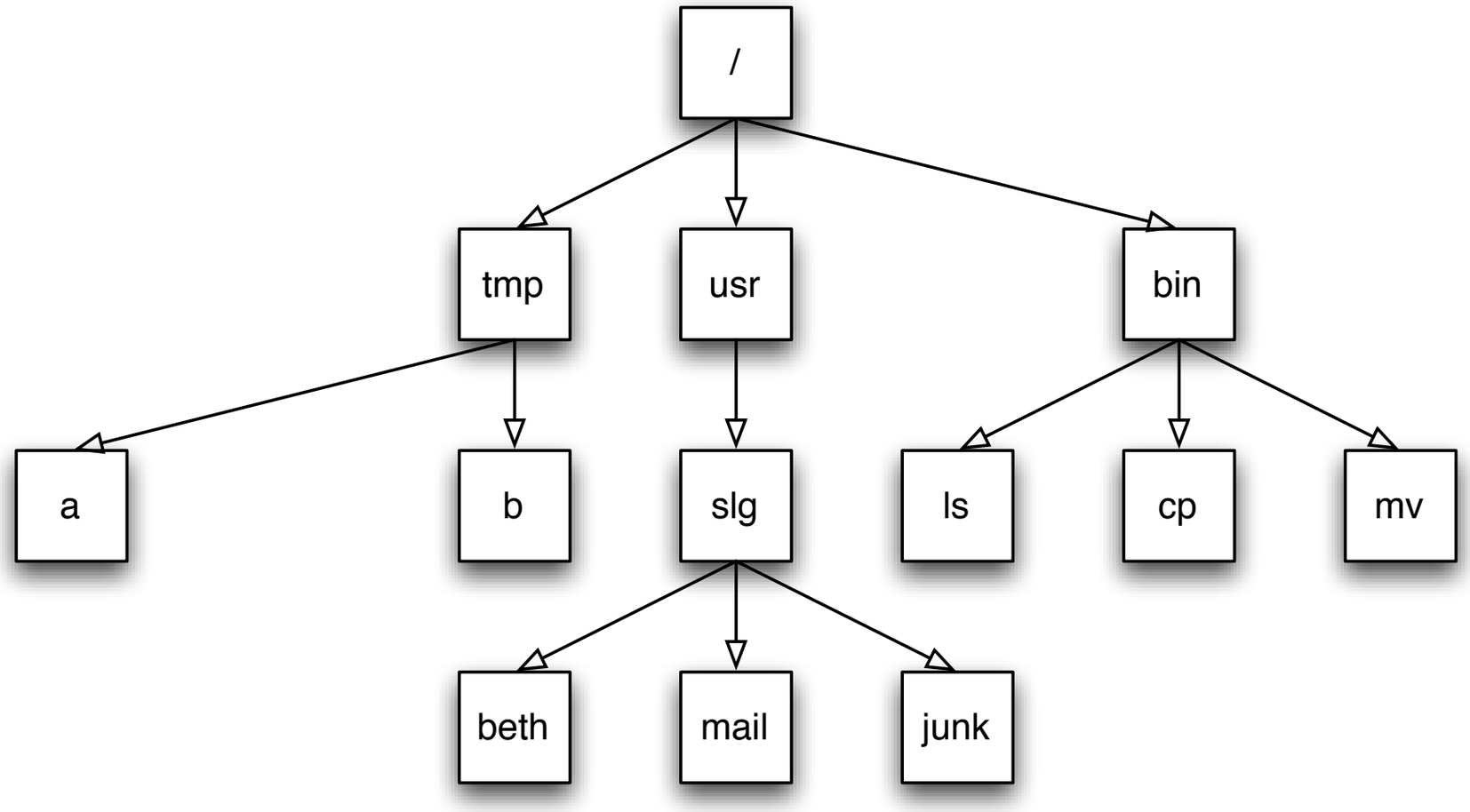
Deleted Data

No Data

user files
email messages
[temporary files]

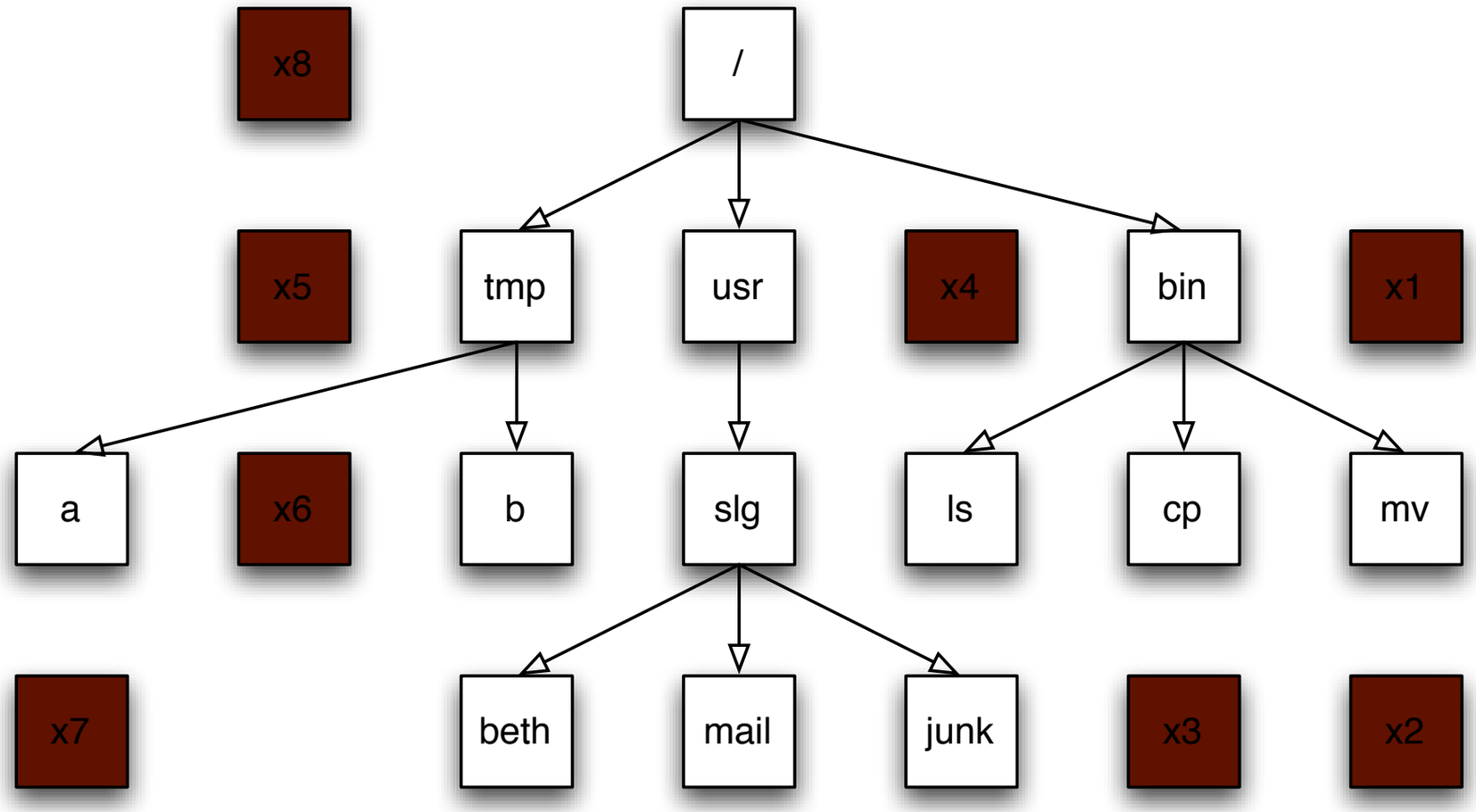
blank sectors

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



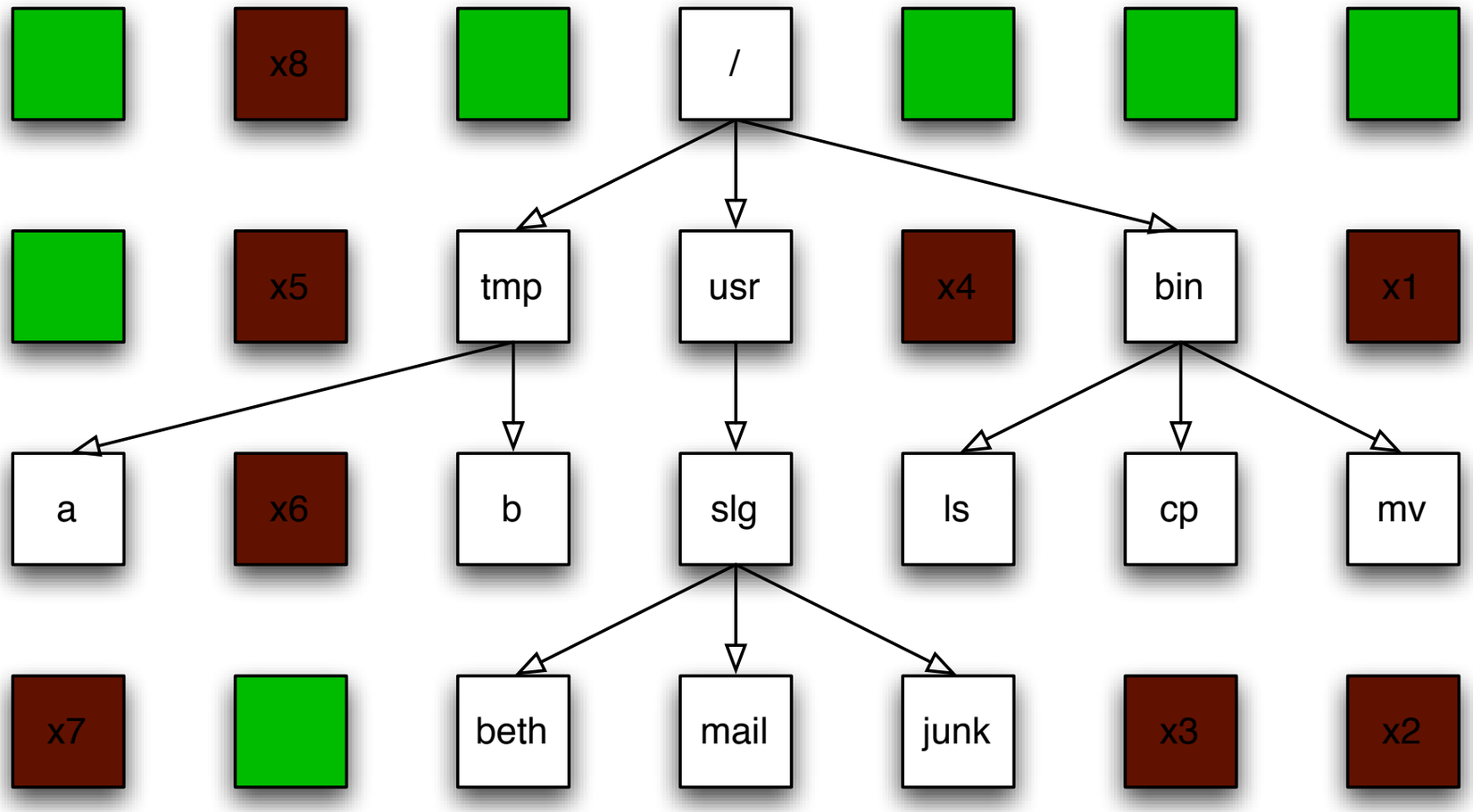
Resident Data

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



Deleted Data

Sectors with “No Data” are blank.



No Data

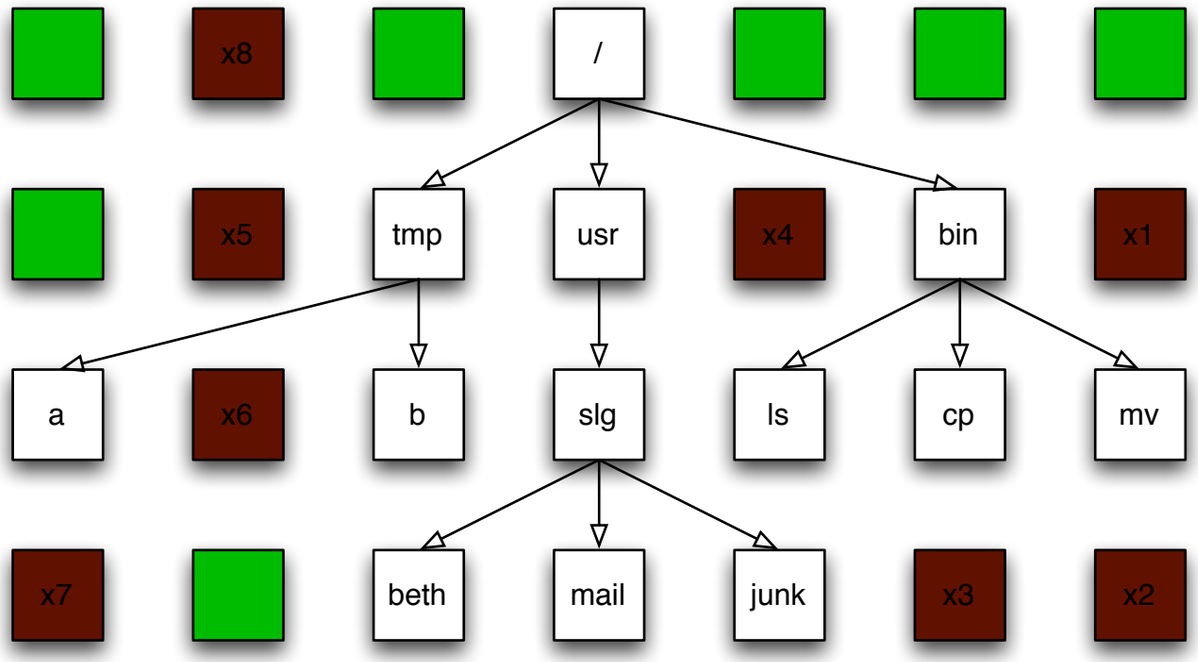
Today most forensic tools follow the same steps to analyze a disk drive.

1. Walk the file system to map out all the files (allocated & deleted).
2. For each file:
    1. Seek to the file.
    2. Read the file.
    3. Hash the file (MD5)
    4. Index file's text.
3. "Carve" space between files for other documents, text, etc.

Problems:

1TB drive takes 10-80 hours.

Lots of residual data is ignored.



Can we analyze a drive in the time it takes to read the data?

Stream-Based Disk Forensics: Just scan the disk from beginning to end.

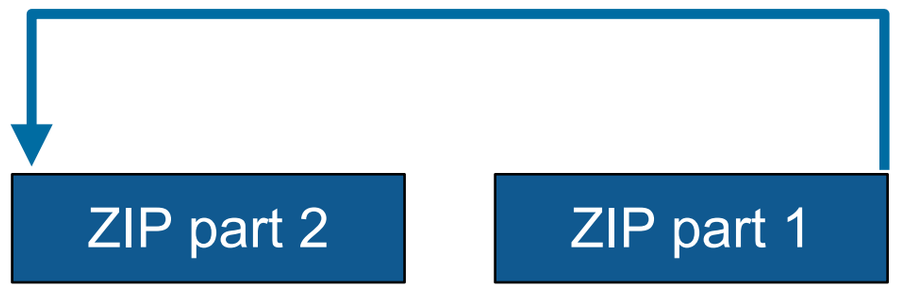
Read all of the blocks in order.

Look for information that might be useful.

Identify & extract what's possible in a single pass.

Advantages:

- No disk seeking.
- Read the disk at maximum transfer rate.
- Reads *all the data* — allocated files, deleted files, file fragments.
- Most files are not fragmented.



Disadvantages:

- Fragmented files won't be recovered:
  - —*Compressed files with part2-part1 ordering*
  - —*FIles with internal fragmentation (.doc)*
- A second pass is needed to map contents to file names.

bulk_extractor: a high-speed disk scanner.

Written in C, C++ and Flex.

Uses regular expressions and rules to scan for:

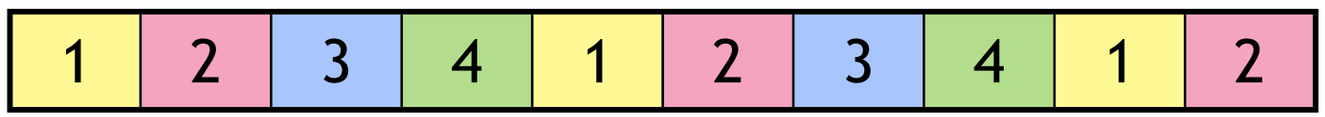
- email addresses
- credit card numbers
- JPEG EXIFs
- URLs
- Email fragments.

Recursively re-analyzes ZIP components.

Produces a histogram of the results.

Multi-threaded.

- Disk is "striped" and then the results are combined.



bulk_extractor output: text files of "features" and context.

email addresses from domexusers:

```
42562736        domexuser2@gmail.com
44113597        pki@microsoft.com
44934320        domexuser2@gmail.com
44935964        domexuser2@live.com
44948252        domexuser2@live.com
44996456        myname@example.com
45180092        inet@microsoft.com
47528617        domexuser2@live.com
47742673        premium-server@thawte.com
```

Histogram:

```
n=579   domexuser1@gmail.com
n=432   domexuser2@gmail.com
n=340   domexuser3@gmail.com
n=268   ips@mail.ips.es
n=252   premium-server@thawte.com
n=244   CPS-requests@verisign.com
n=242   someone@example.com
```

City of San Luis Obispo Police Department, June 2010

SLO County DA filed charges of credit card fraud and possession of materials to make fraudulent credit cards against 2 individuals.

- Defendants arrested with a computer.
- Defense expected to argue that defends were unsophisticated and lacked knowledge.

Examiner given 250GiB drive *the day before preliminary hearing.*

In 2.5 hours Bulk Extractor found:

- Over 10,000 credit card numbers on the HD (1000 unique)
- The most common email address belonged to the primary defendant, helping to establish possession.
- The most commonly occurring Internet search engine queries concerned credit card fraud and bank identification numbers, helping to establish intent.
- Most commonly visited websites were in a foreign country whose primary language is spoken fluently by the primary defendant.

Armed with this data, the DA was able to have the defendants held.

Tuning bulk_extractor.
Many of the email addresses come with Windows!

Sources of these addresses:

- Windows binaries
- SSL certificates
- Sample documents

```
n=579    domexuser1@gmail.com
n=432    domexuser2@gmail.com
n=340    domexuser3@gmail.com
n=268    ips@mail.ips.es
n=252    premium-server@thawte.com
n=244    CPS-requests@verisign.com
n=242    someone@example.com
```

It's important to suppress email addresses not relevant to the case.

Method #1 — Suppress emails seen on many other drives.

Method #2 — Stop list from bulk_extractor run on clean Installs.

Both of these methods *white list* commonly seen emails.

- A problem — Operating Systems have a LOT of emails. (FC12 has 20,584!)
- Should we give the Linux developers a free past?

Method #3: Context-sensitive stop list.

Instead of extracting just the email address, extract the context:

- Offset: 351373329
- Email: zeeshan.ali@nokia.com
- Context: `ut_Zeeshan Ali <zeeshan.ali@nokia.com>, Stefan Kost <`

- Offset: 351373366
- Email: stefan.kost@nokia.com
- Context: `>, Stefan Kost <stefan.kost@nokia.com>_____sin`

Here "Context" is defined as 8 characters on either side of feature.

We created a context-sensitive stop list for Microsoft Windows XP, 2000, 2003, Vista, and several Linux ver.

Total stop list: 70MB (628,792 features)

Applying it to domexusers HD image:

- # of emails found: 9143 ➔ 4459

without stop list		with stop list	
n=579	domexuser1@gmail.com	n=579	domexuser1@gmail.com
n=432	domexuser2@gmail.com	n=432	domexuser2@gmail.com
n=340	domexuser3@gmail.com	n=340	domexuser3@gmail.com
n=268	ips@mail.ips.es	n=192	domexuser2@live.com
n=252	premium-server@thawte.com	n=153	domexuser2@hotmail.com
n=244	CPS-requests@verisign.com	n=146	domexuser1@hotmail.com
n=242	someone@example.com	n=134	domexuser1@live.com
n=237	inet@microsoft.com	n=91	premium-server@thawte.com
n=192	domexuser2@live.com	n=70	talkback@mozilla.org
n=153	domexuser2@hotmail.com	n=69	hewitt@netscape.com
n=146	domexuser1@hotmail.com	n=54	DOMEXUSER2@GMAIL.COM
n=134	domexuser1@live.com	n=48	domexuser1%40gmail.com@imap.gmail.com
n=115	example@passport.com	n=42	domex2@rad.li
n=115	myname@msn.com	n=39	lord@netscape.com
n=110	ca@digsigtrust.com	n=37	49091023.6070302@gmail.com

Can we analyze a hard 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	1 TB	7.2 GB
Max Seeks		18,000

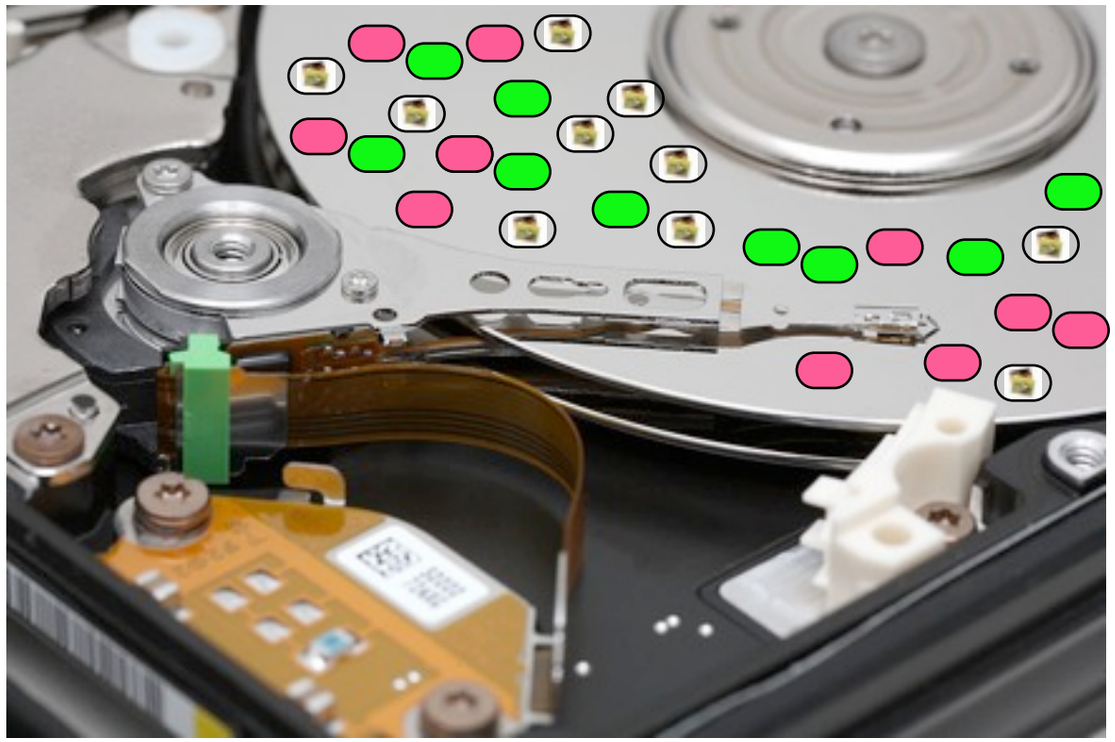
7.2 GB is a lot of data!

- ≈ 0.48% of the disk
- But it can be a statistically significant sample.

We can predict the statistics of a *population* by sampling a *randomly chosen sample.*

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

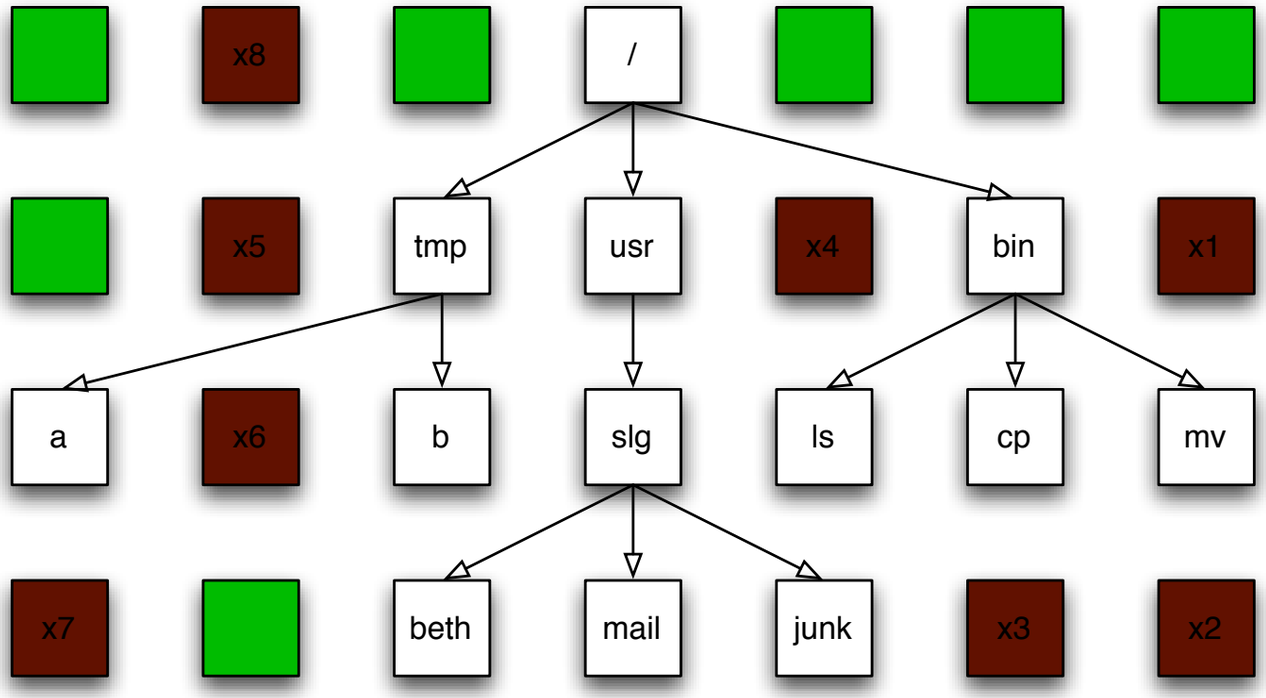
Hard drive contents can be predicted by sampling a few thousand sectors:

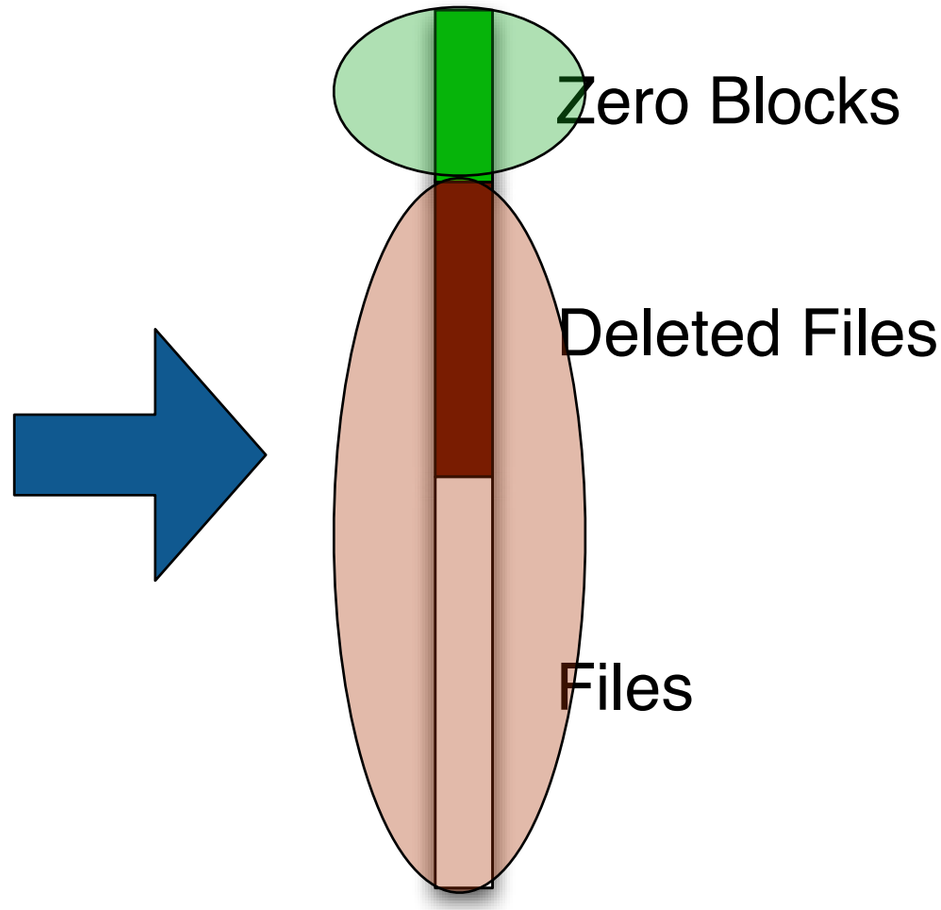


The challenge is identifying *likely voters.*

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

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

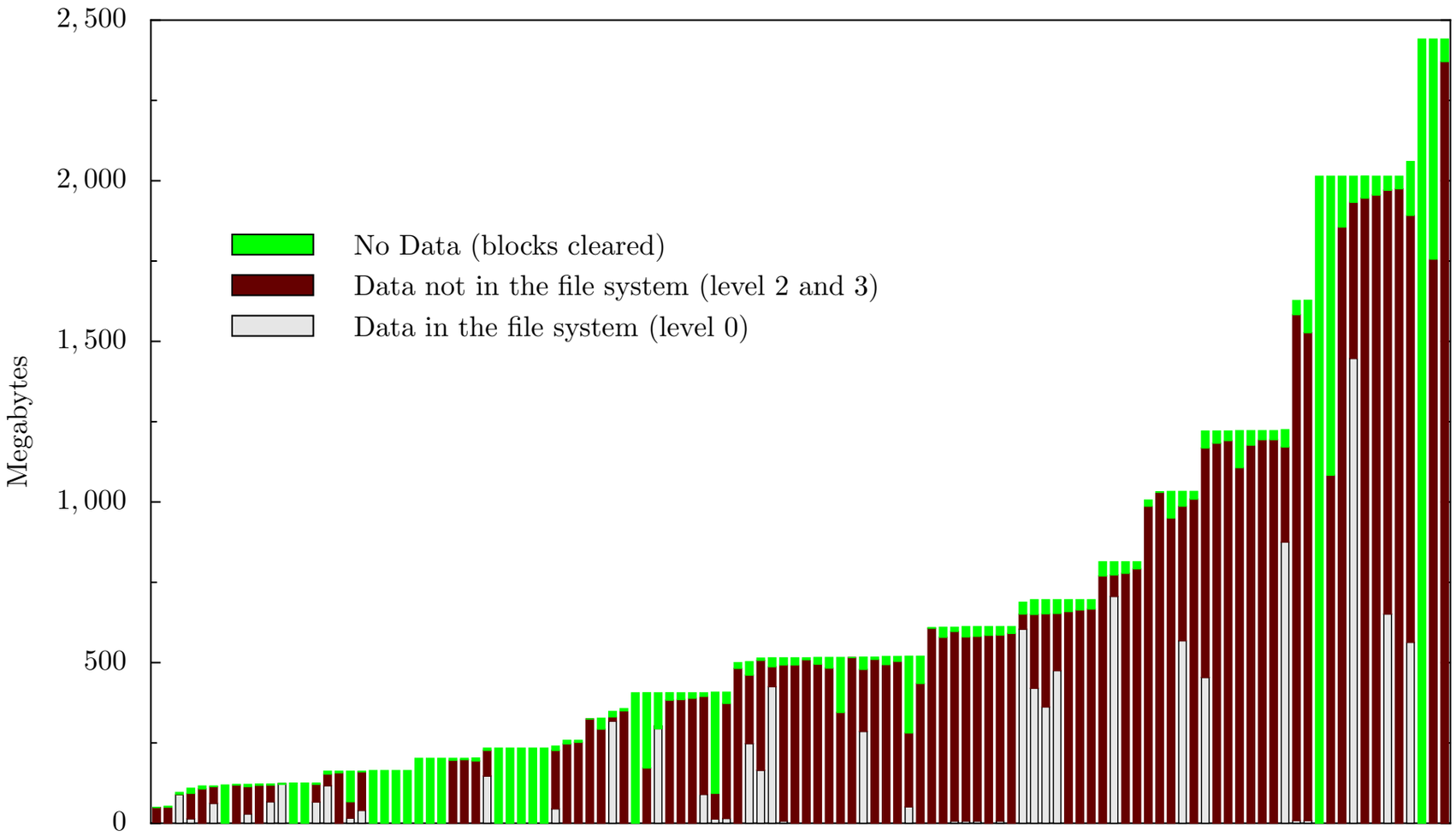




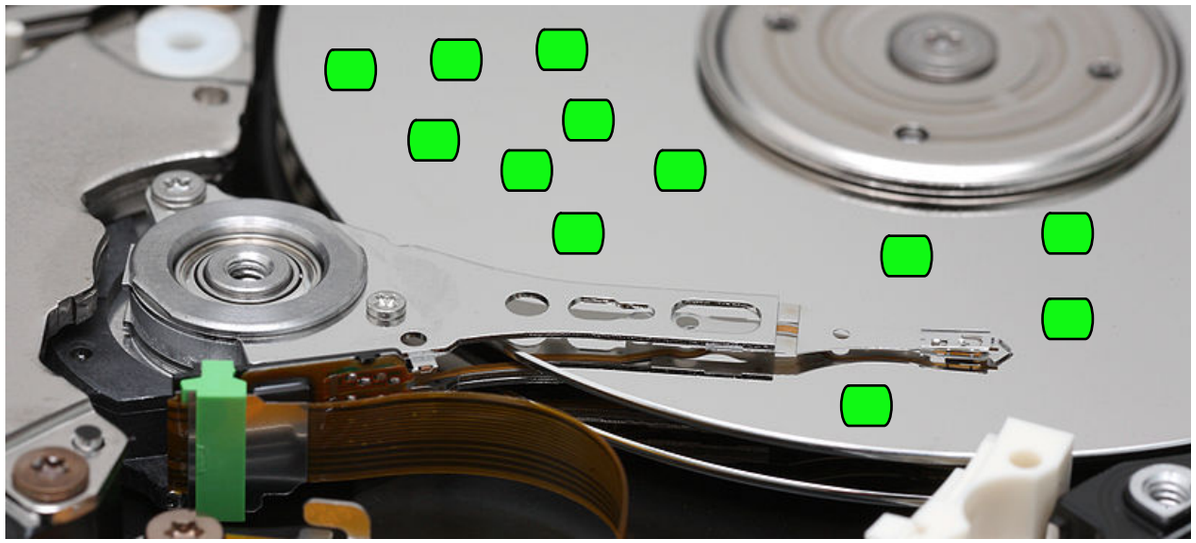
Simplify the problem.
Can we use statistical sampling to verify wiping?

I bought 2000 hard drives between 1998 and 2006.

Most of were not properly wiped.



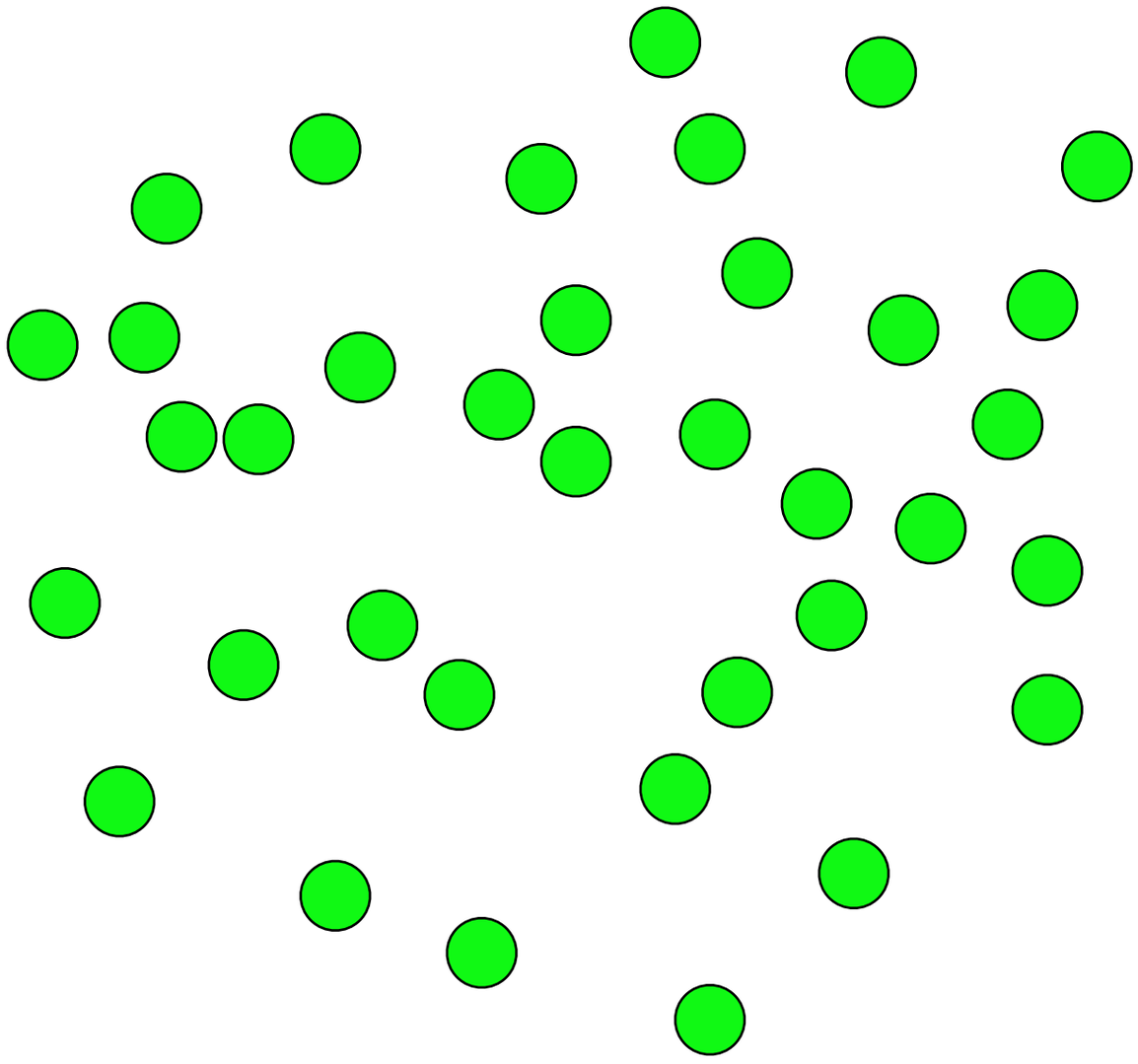
It should be easy to use random sampling to distinguish a properly cleared disk from one that isn't.



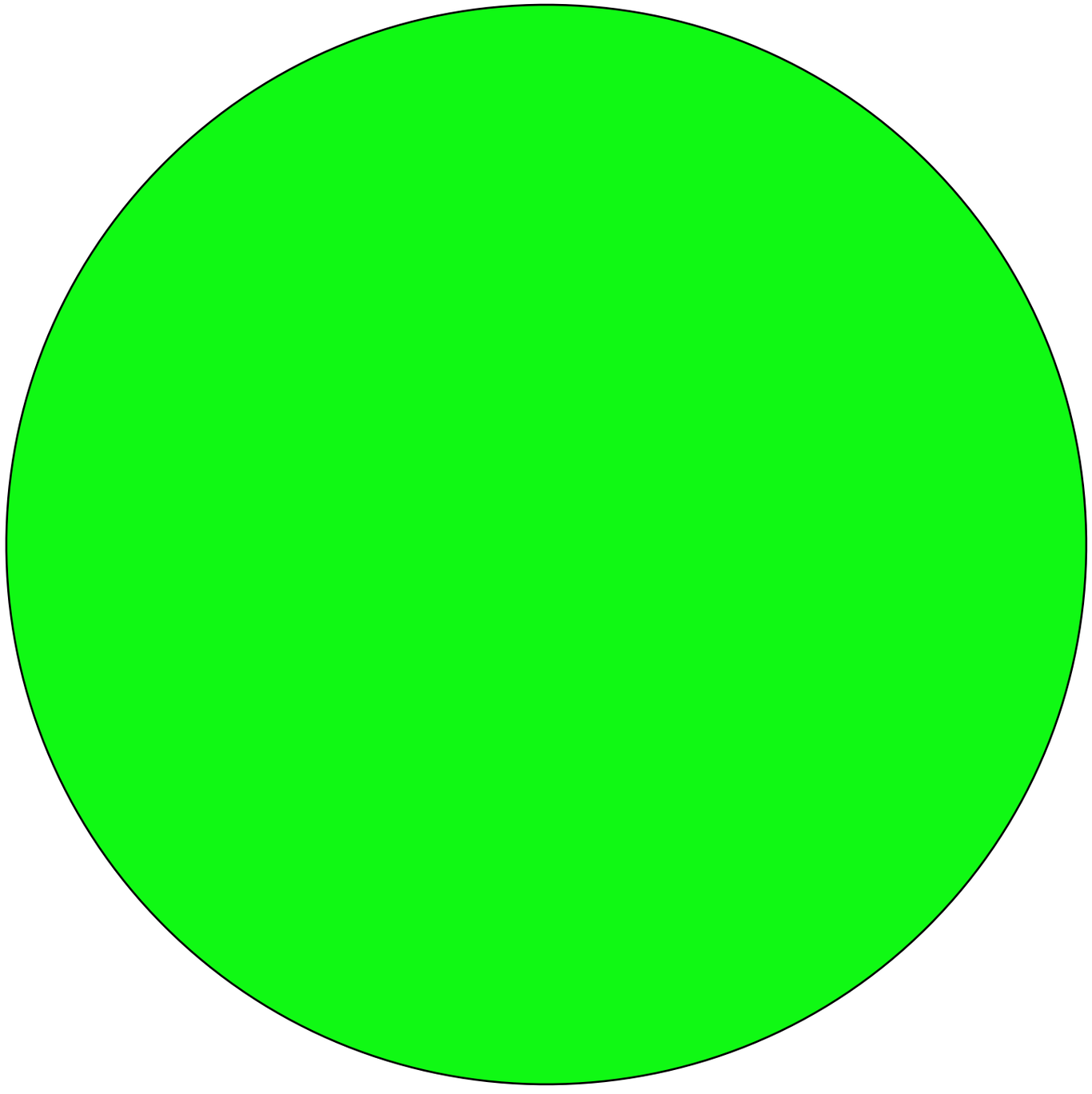
Let’s try reading 10,000 random sectors and see what happens….

We read 10,000 randomly-chosen sectors … and they are all blank

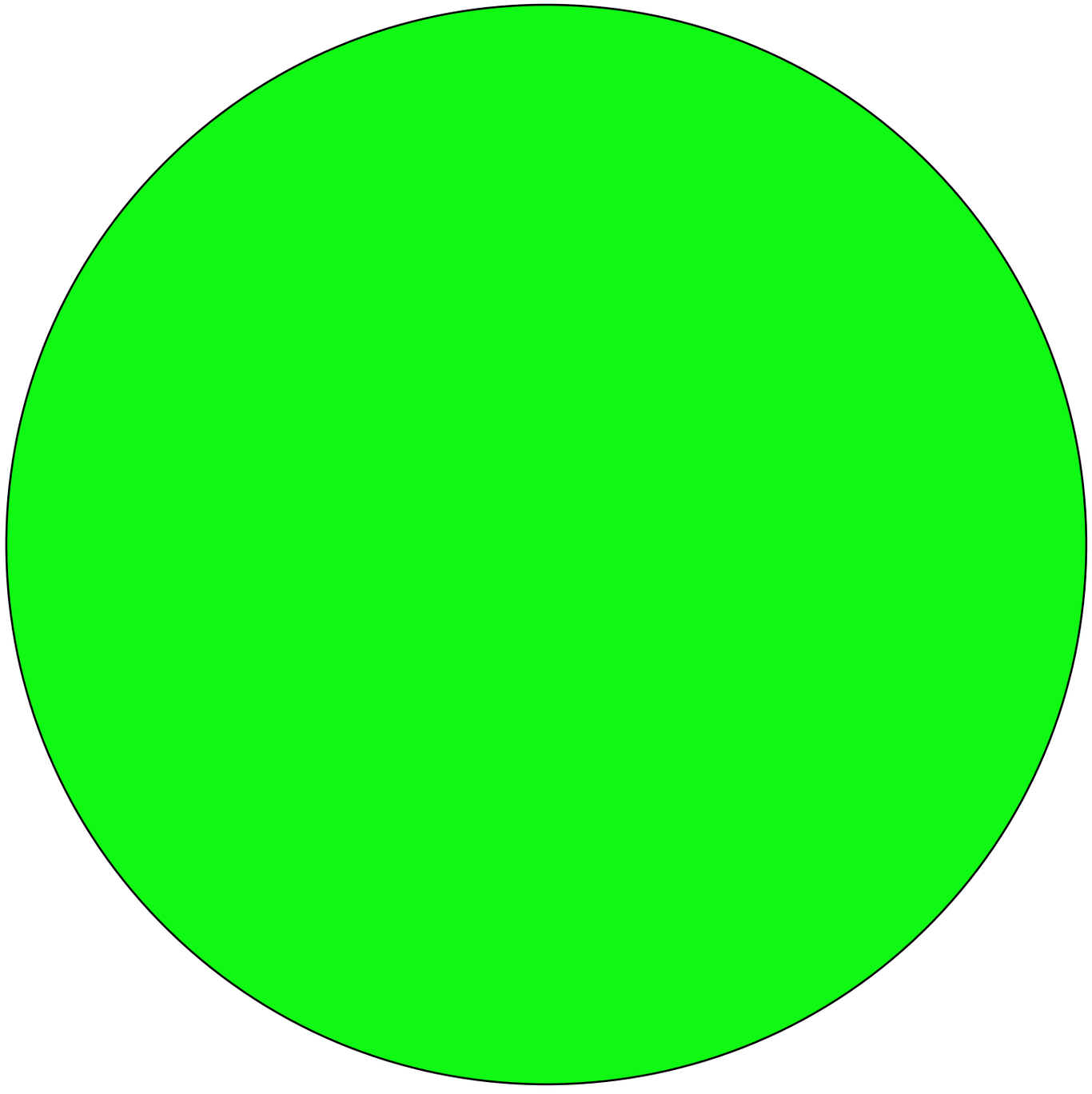
We read 10,000 randomly-chosen sectors … and they are all blank



We read 10,000 randomly-chosen sectors … and they are all blank



We read 10,000 randomly-chosen sectors … and they are all blank

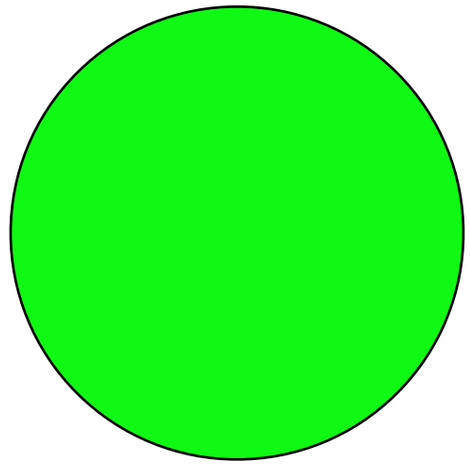


Chances are good that they are all blank.

Random sampling *won't* find a single written sector.

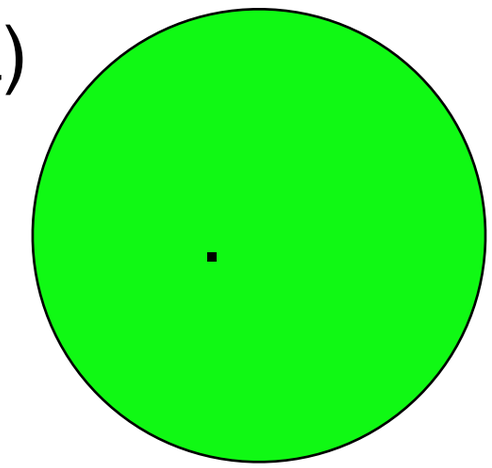
If the disk has 2,000,000,000 blank sectors (0 with data)

- The sample is identical to the population



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

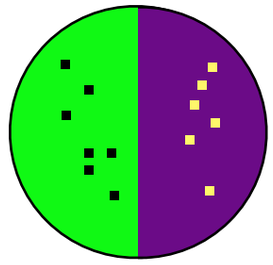
- The sample is representative of the population.
- We will only find that 1 sector using exhaustive search.



What about non-uniform distributions?

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

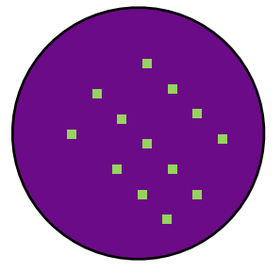
- The sampled frequency should match the distribution.
- *This is why we use random sampling.*



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

— and all these are the sectors that we read???

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



What are the proper statistics for evaluating the sample?

Random sampling can't prove there is no data...

- But we can use it to calculate the odds that there is *less* than a certain amount of data.

Assume the disk has 10MB of data --- 20,000 non-zero sectors.

Read just 1 sector; the odds of finding a non-blank sector are:

$$\frac{200{,}000{,}000-20{,}000}{200{,}000{,}000} = 0.9999.$$

Read 2 sectors. The odds are:

$$\left(\frac{200{,}000{,}000-20{,}000}{200{,}000{,}000}\right)\left(\frac{199{,}999{,}999-20{,}000}{199{,}999{,}999}\right) = 0.99980001$$

first pick **second pick**

Odds we may have missed something

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.

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

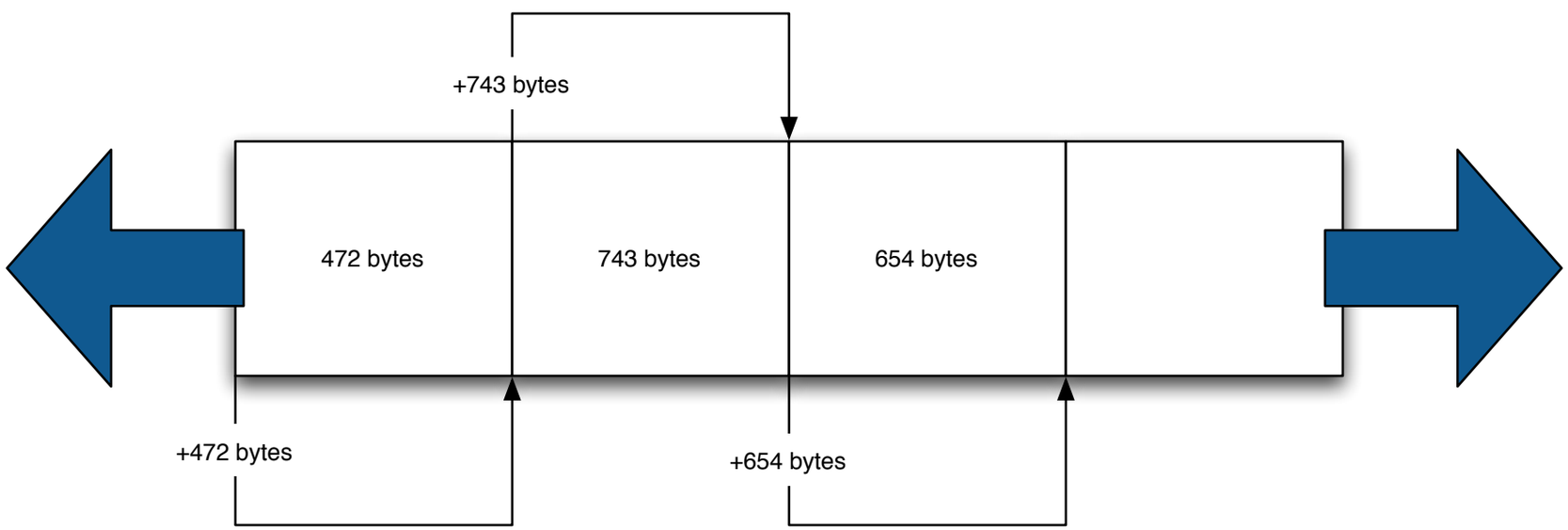
Fragment classification:
Many file "types" can be identified from a fragment.

HTML files can be reliably detected with HTML tags

```
<body onload="document.getElementById('quicksearch').terms.focus()">
  <div id="topBar">
    <div class="widthContainer">
          <div id="skiplinks">
            <ul>
              <li>Skip to:</li>
```

MPEG files can be readily identified through framing

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



10 years of research on fragment identification…
… mostly using n-gram analysis (bigrams)

Standard approach:

- Get samples of different file types
- Train a classifier (typically k-nearest-neighbor or Support Vector Machines)
- Test classifier on "unknown data"

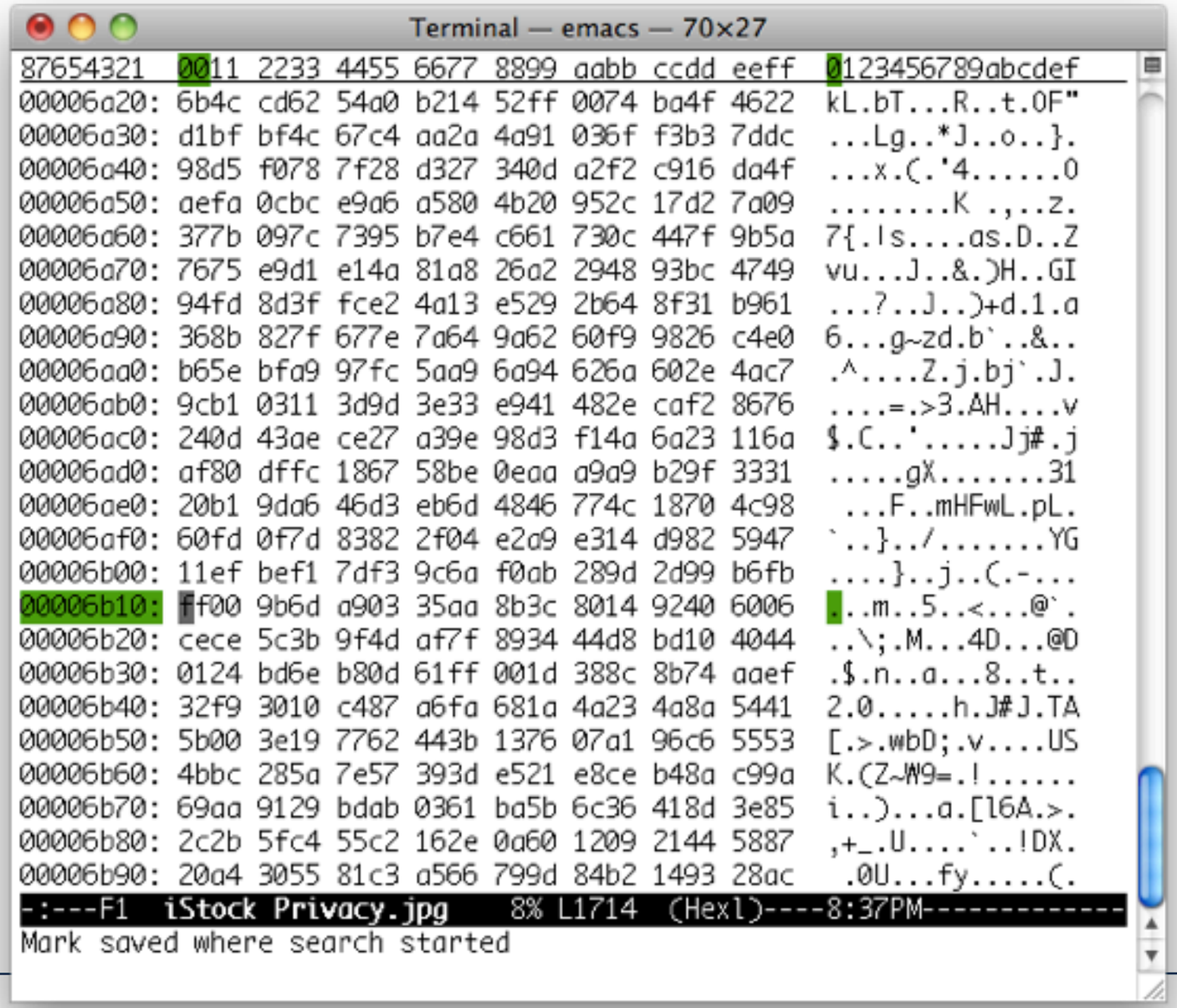
Examples:

- 2001 — McDaniel — "Automatic File Type Detection Algorithm"
  - —*header, footer & byte frequency (unigram) analysis (headers work best)*
- 2005 — LiWei-Jen et. al — "Fileprints"
  - —*unigram analysis*
- 2006 — Haggerty & Taylor — "FORSIGS"
  - —*n-gram analysis*
- 2007 — Calhoun — "Predicting the Type of File Fragments"
  - —*statistics of unigrams*

  - —*http://www.forensicswiki.org/wiki/File_Format_Identification*

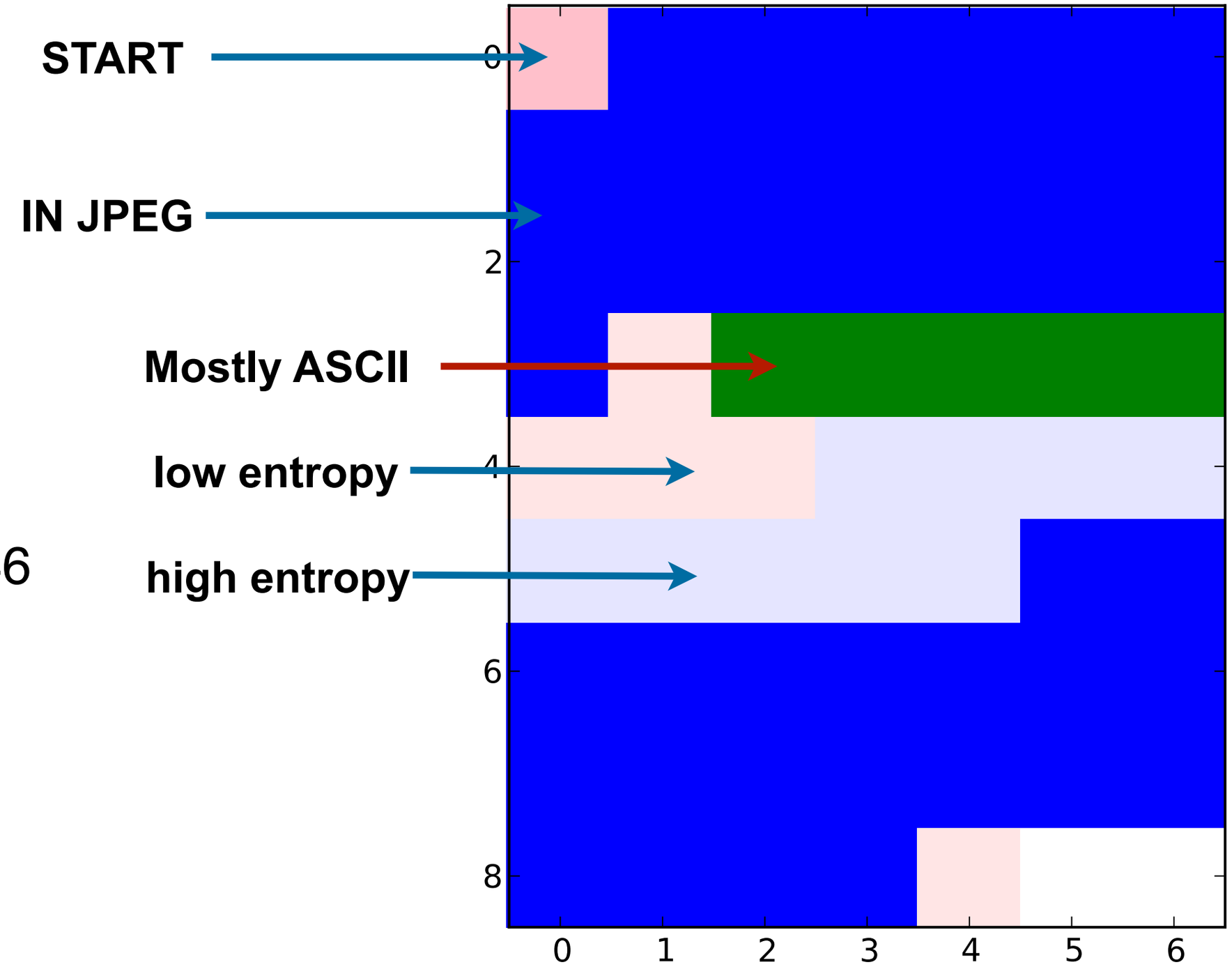
Our approach: hand-tuned discriminators based on a close reading of the specification.

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



Using these statistics, we can build detectors that recognize the different parts of a JPEG file.

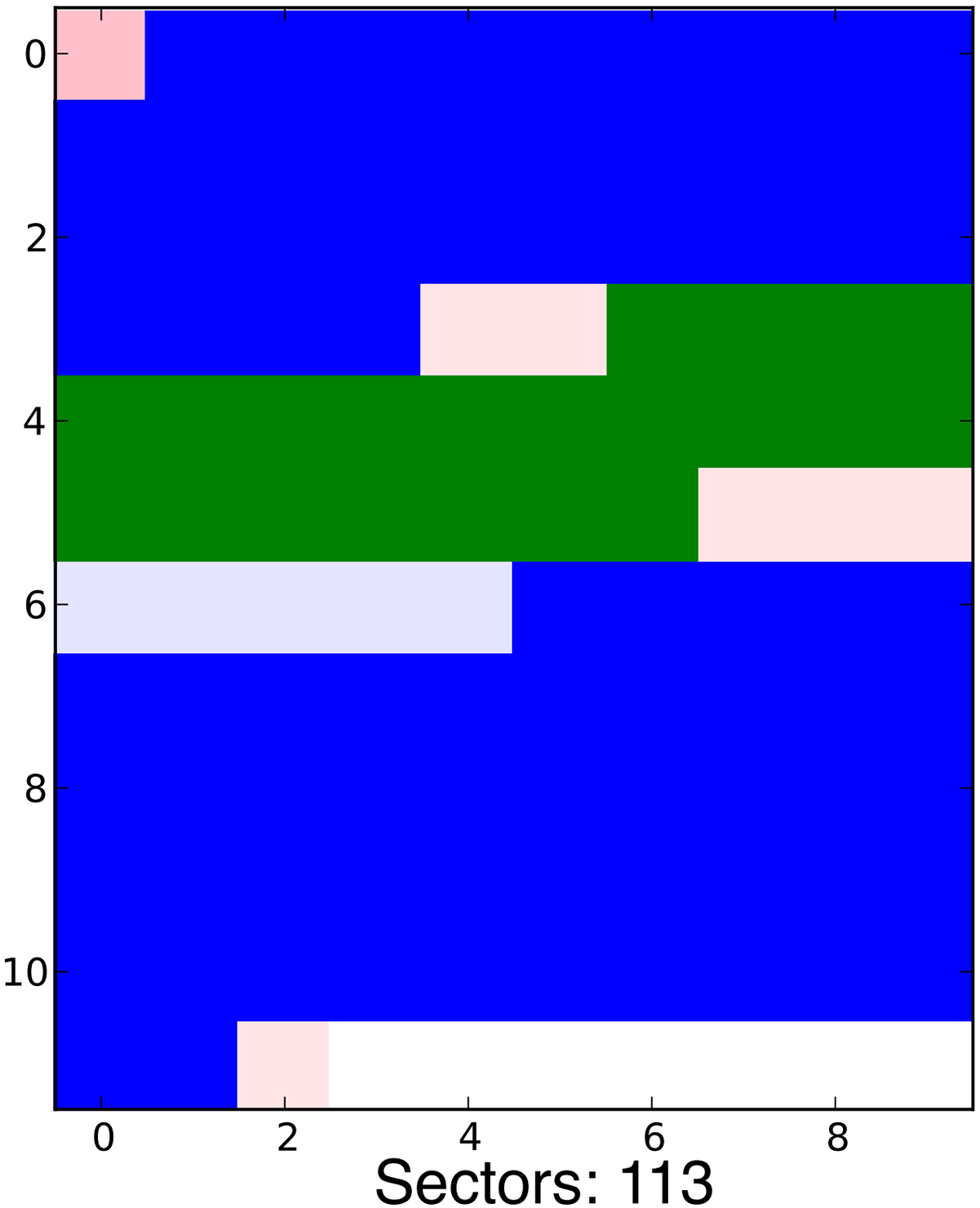
Bytes: 31,046



Sectors: 61

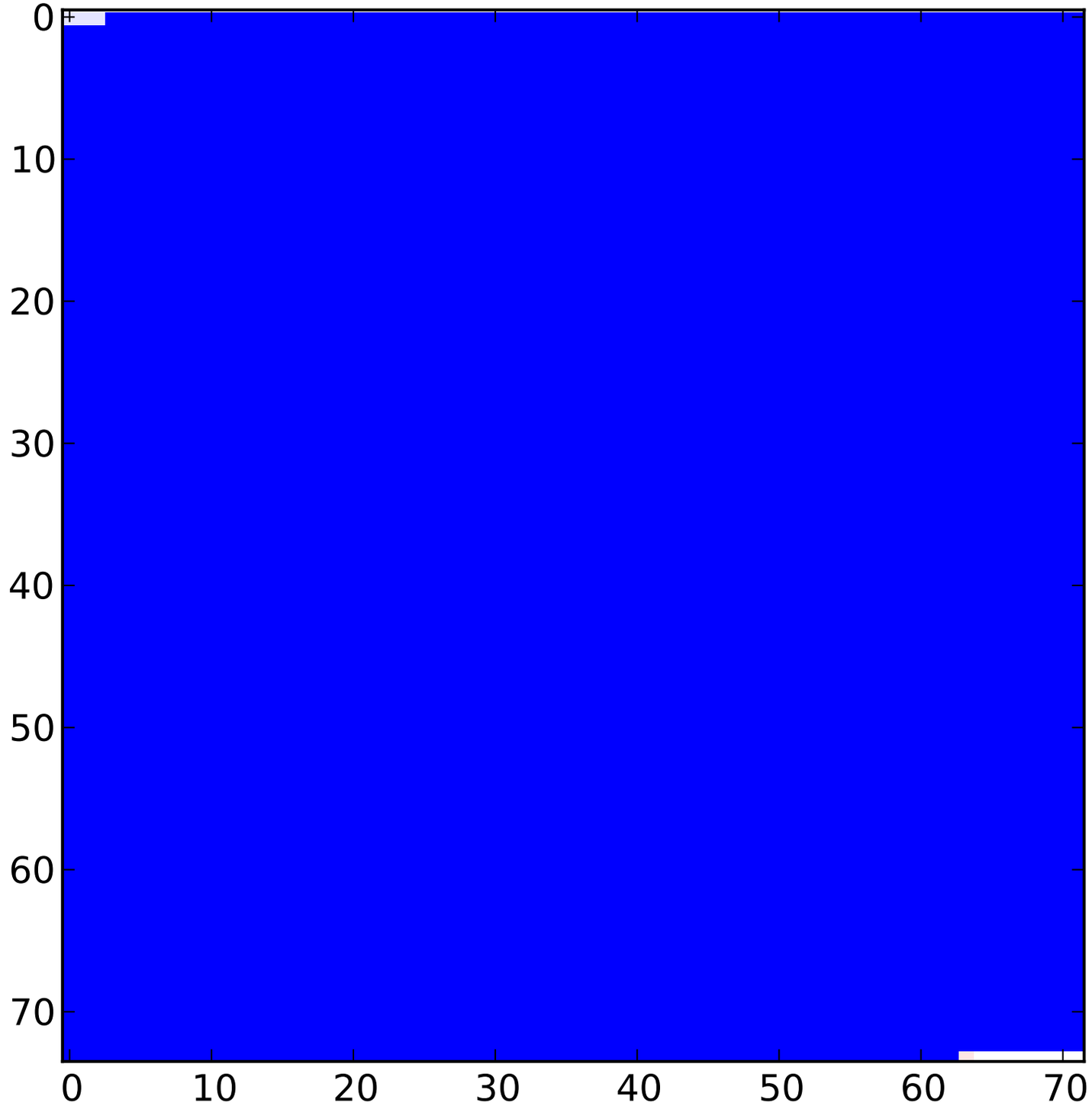
000897.jpg

Bytes: 57596



000888.pdf

Bytes: 2,723,425



Sectors: 5320

We developed five fragment discriminators.

JPEG — High entropy and FF00 pairs.

MPEG — Frames

Huffman-Coded Data — High Entropy & Autocorrelation

"Random" or "Encrypted" data — High Entropy & No autocorrelation

Distinct Data — a block from an image, movie, or encrypted file.

208 distinct 4096-byte block hashes

Using random sampling, we determine the *forensic content* of a 160GB iPod in less than a minute.

Time to read 10,000 randomly chosen 64K runs: 45 seconds

Identifiable:

- Blank sectors
- JPEGs
- Encrypted data
- HTML

Sample report:

- Encrypted: 10% (100GB)
- JPEGs: 5% (50GB)
- MP3s: 50% (500GB)
  - —*Kind of interesting if you are looking at an iPod*

In summary: Automated Digital Forensics

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.

—*For more information, see http://www.simson.net/ or http://forensicswiki.org/*