



# Automated Digital Forensics and Media Exploitation

Simson L. Garfinkel

Associate Professor, Naval Postgraduate School

August 20, 2009

<http://simson.net/>

# NPS is the Navy's Research University.



Location: Monterey, CA

Campus Size: 627 acres

Students: 1500

- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)
- *All students are fully funded*

Schools:

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





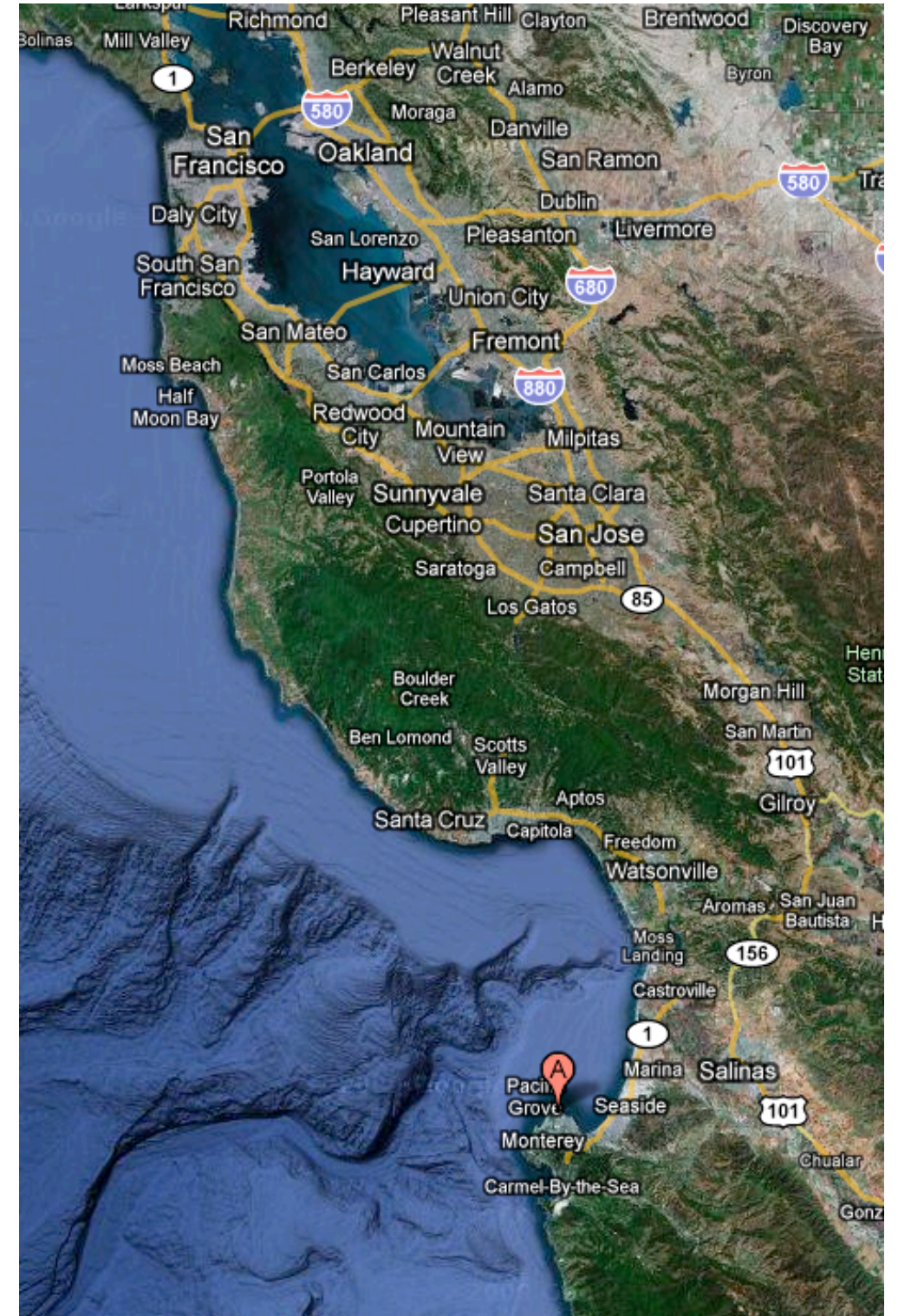
# I live in Pacific Grove, California



**Bike to work...**



**View from front porch...**







# Automated Document and Media Exploitation: The Need



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



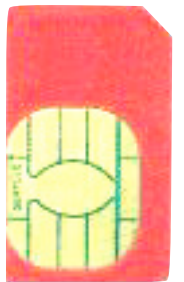
June 2007

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

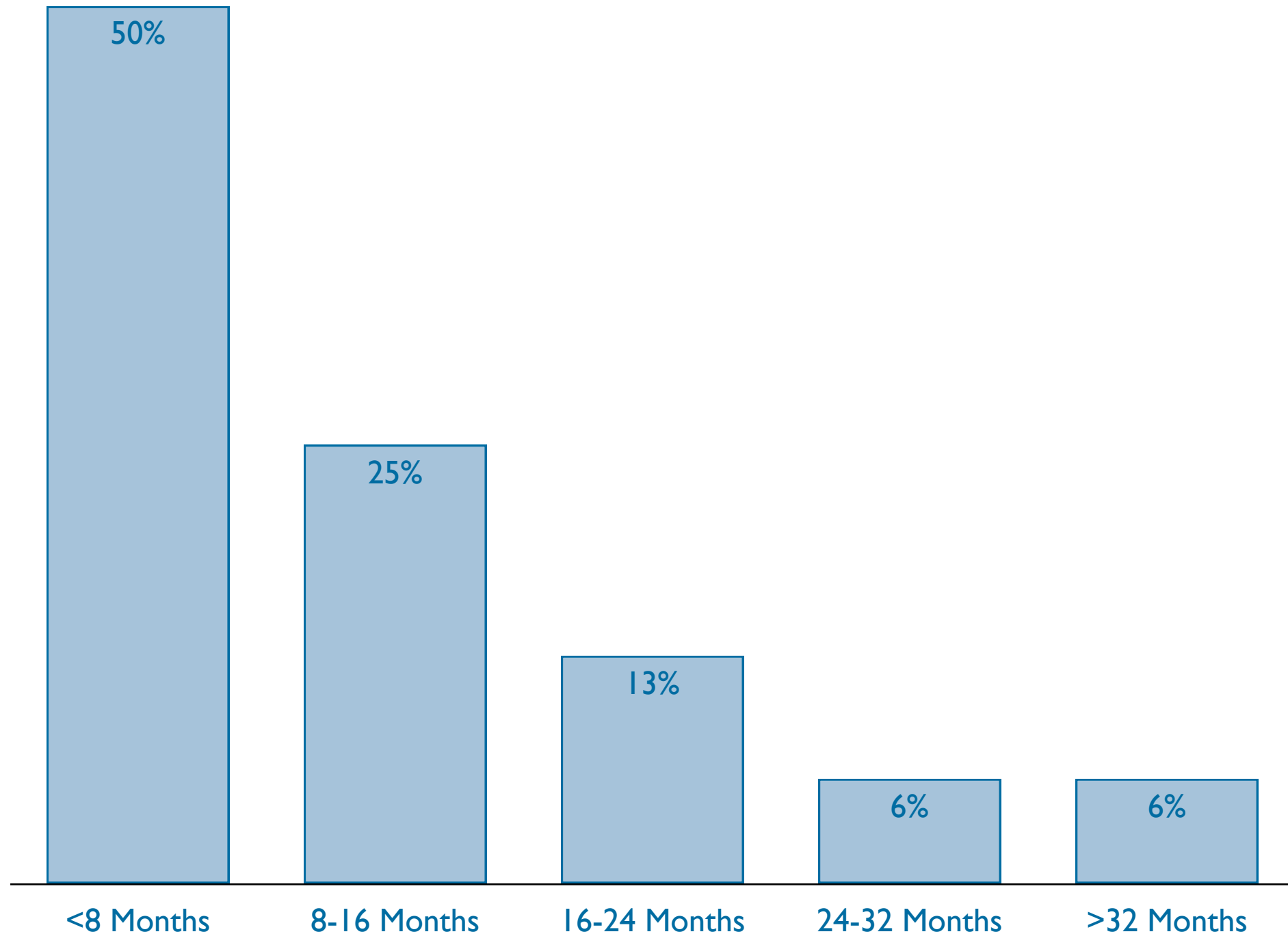


- Media collected on the battle field
- Media collected over last 7 years
  - *Hundreds of Terabytes*
  - *Hundreds of millions of files from thousands of pieces of media collected over 2 years*

(Source: Defense Cyber Crime Center)



# Media Size and quantity is increasing geometrically.



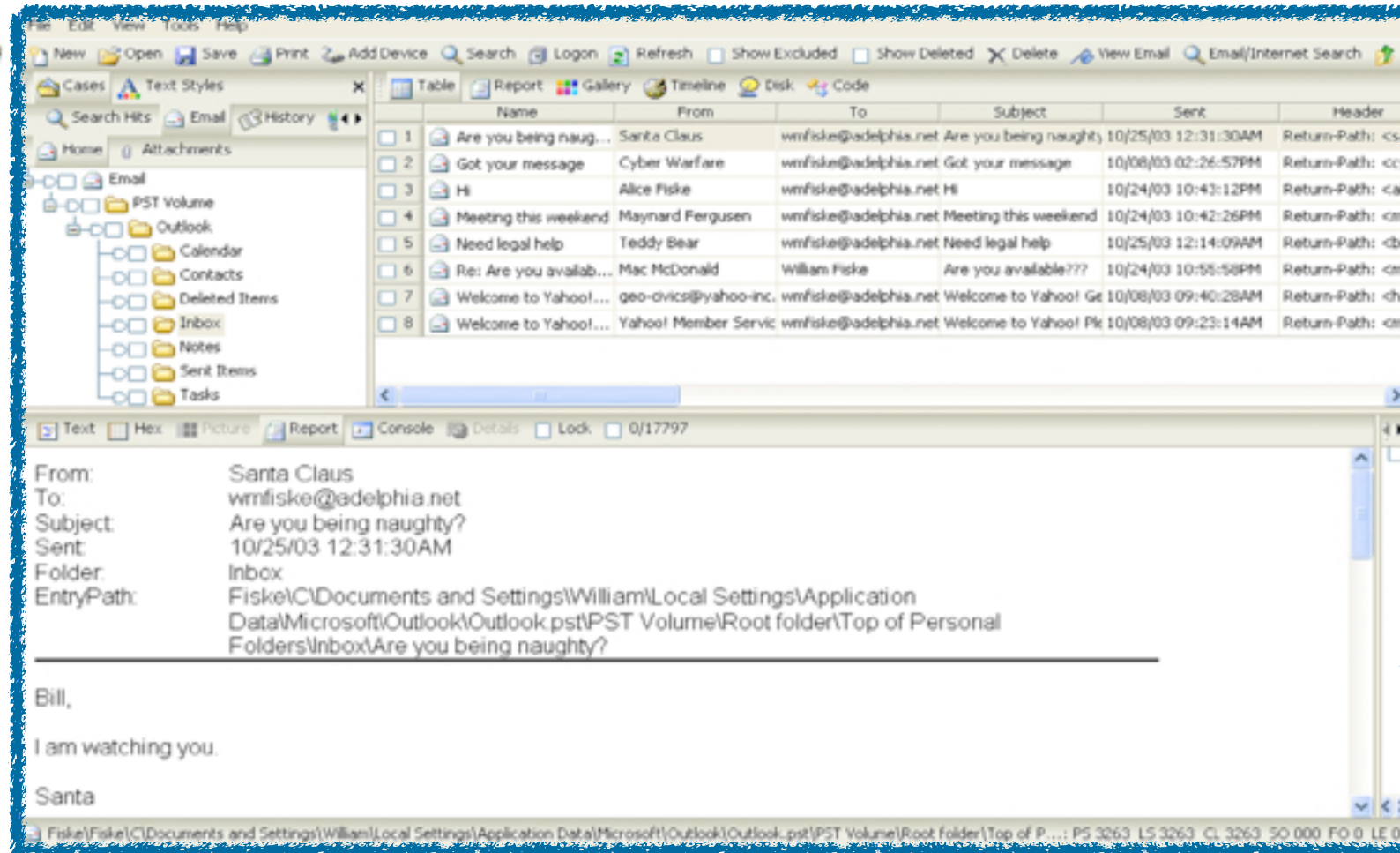


Most of this data is analyzed using trained personnel...



**DOMEX in Iraq**

# ... working with tools designed for law enforcement.



## EnCase by Guidance Software

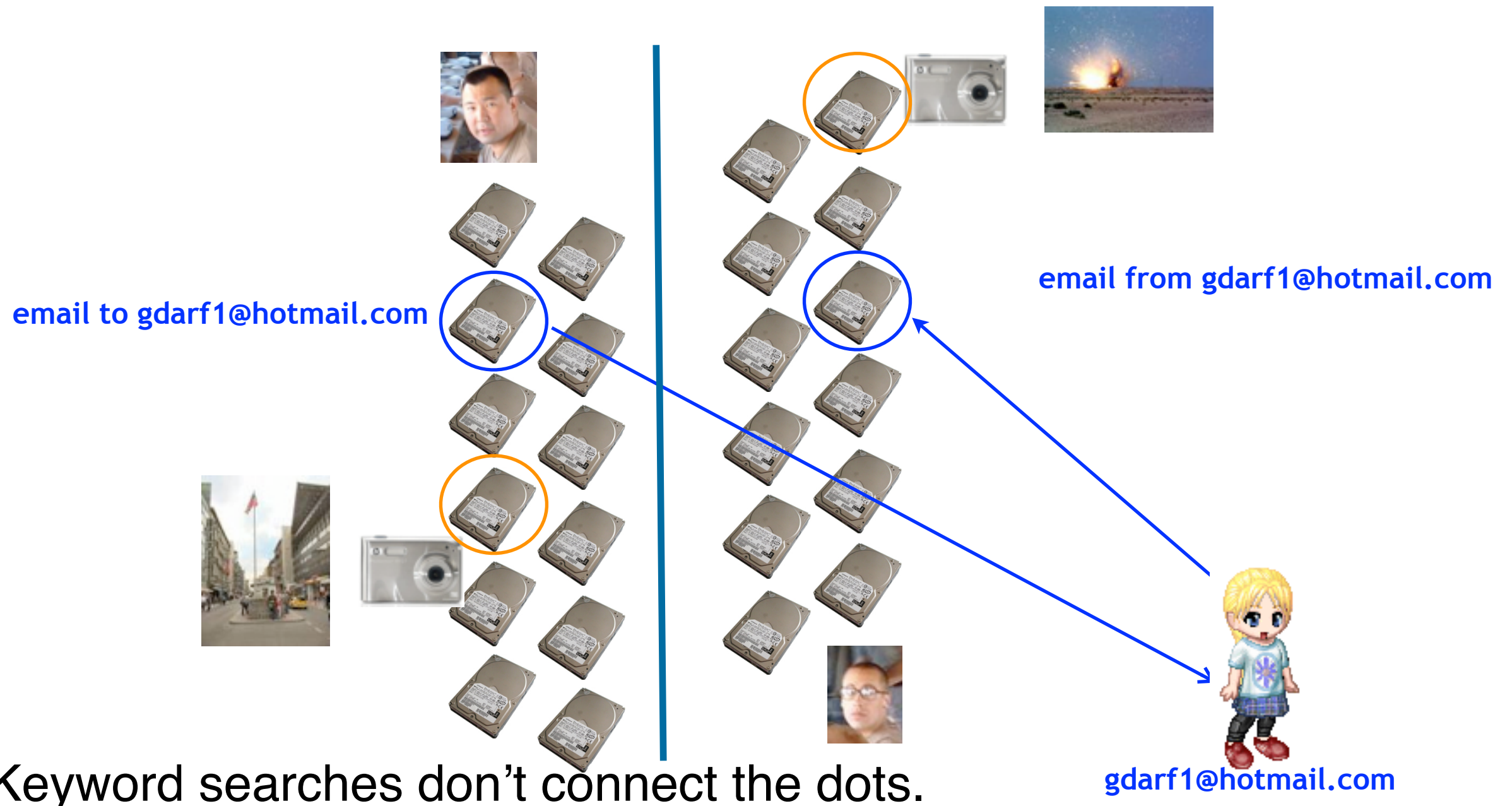
- Designed for visibility & search, not analysis.
- Do not scale to 100s or 1000s of drives.
- No “contamination” of evidence between cases.





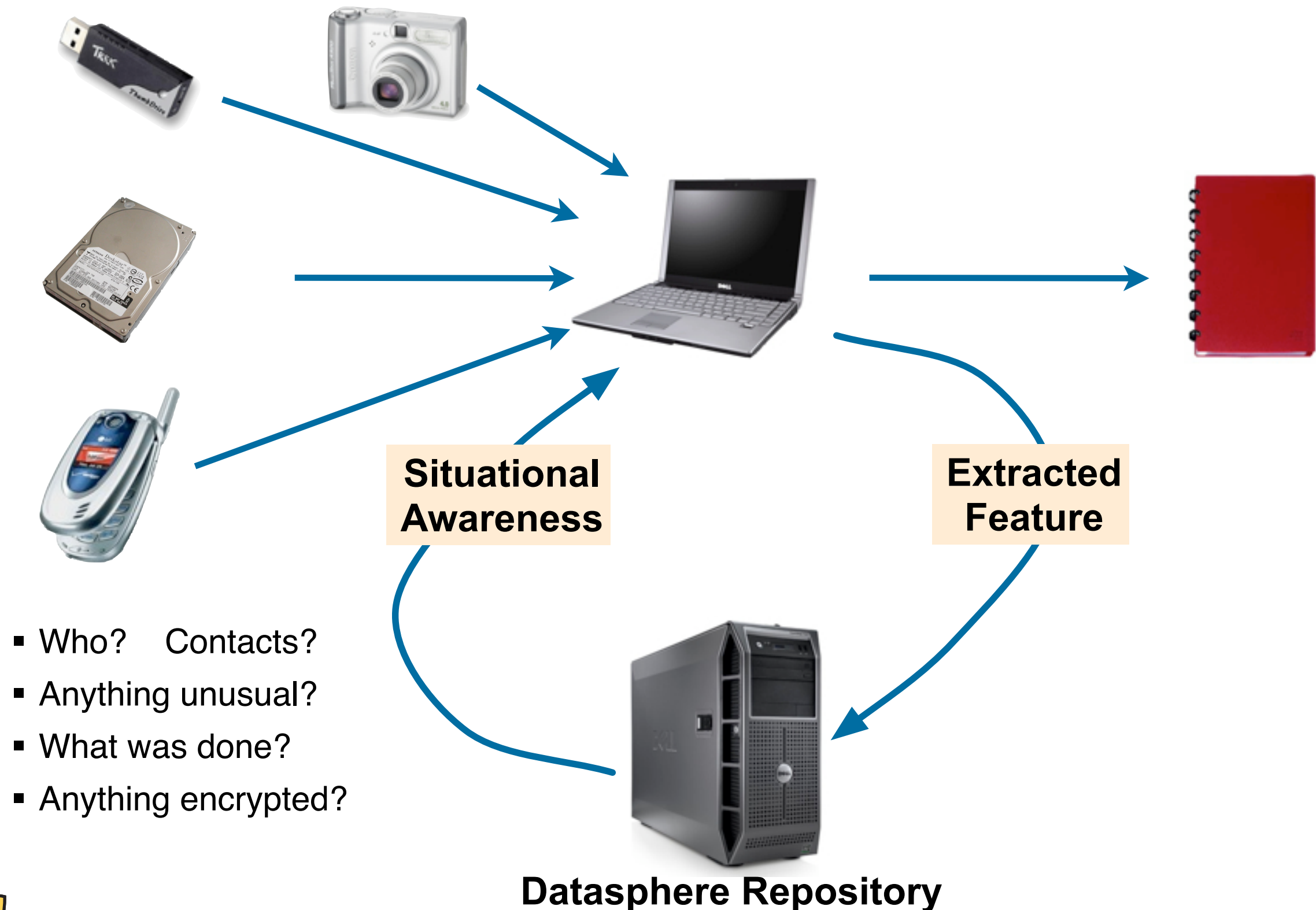
# Manual analysis misses opportunities for correlation.

Different analysts see different hard drives.



Keyword searches don't connect the dots.

# End-to-end automated analysis can increase exploitation capabilities and connect the dots.





# Today's primary form of automation: hash sets.

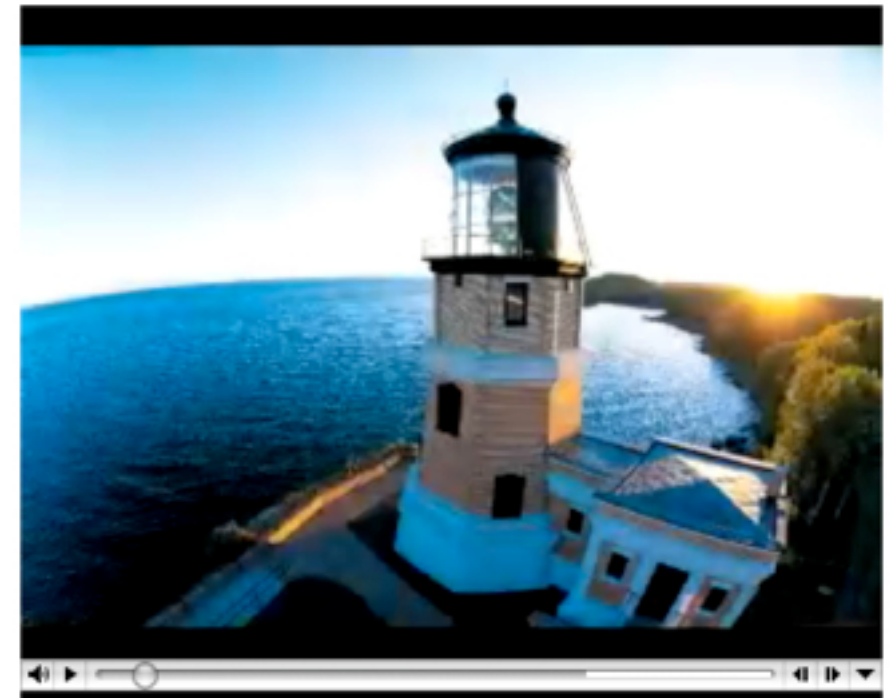
A “hash” is a fingerprint for a file.

- e.g. 9ed9127ee08da92aa7ff2fd754ddba91

“Hash Set” = A set of hashes

- 9ed9127ee08da92aa7ff2fd754ddba91
- 3a3febb29e55ee975098903a31f8022a
- 5db454651210bcb920c010f8149d118c

cell phone video:



13316868 bytes  
9ed9127ee08da92aa7ff2fd754ddba91

Typical hash sets:

- National Software Reference Library (NIST): **software distributions**
- National Center for Missing and Exploited Children: **Child porn**
- WetStone's Gargoyle Investigator: **Malware and steganography tools**

# Hash sets are widely used... ... but have important limitations.

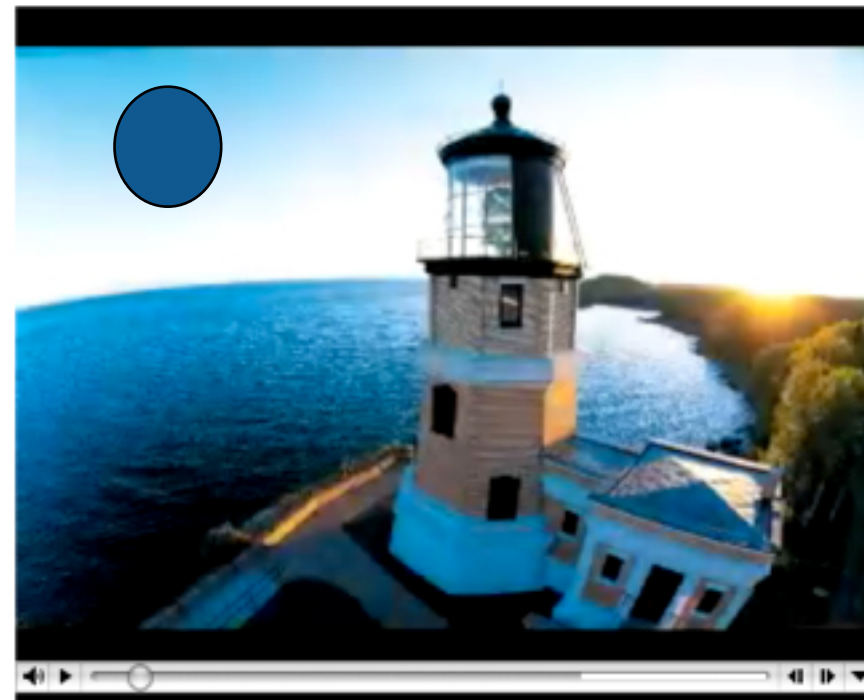
- Hashing requires the entire file.
- Hashing is brittle: change a single bit and the hash changes.
- Hashing is slow, especially SHA256.

**cell phone video:**



13316868 bytes  
9ed9127ee08da92aa7ff2fd754ddba91

**same video with a dot:**



13316868 bytes  
b42abbb9b9ff83ae769bce02e660ccc42



# Our research thrusts are in four main areas

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

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

## Area #2: Bringing data mining to forensics

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



## Area #3: Bulk Data Analysis

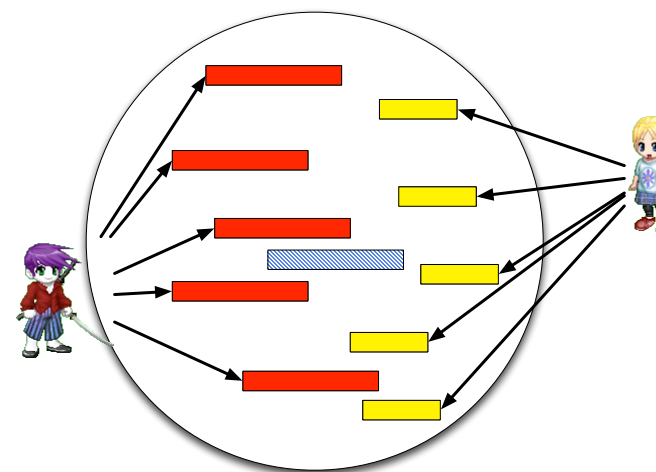
- Stream-processing
- Statistical techniques (sub-linear algorithms)

## Area #4: Creating Standardized Forensic Corpora

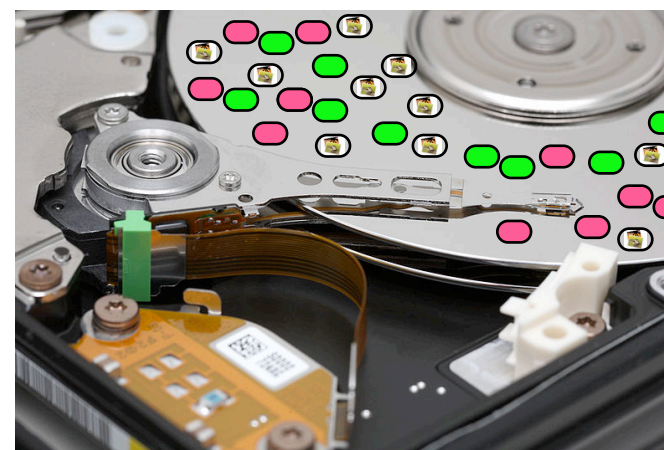
- Freely redistributable disk and memory images, packet dumps, file collections.

# This talk focuses on three key areas:

## Multi-User Carved Data Ascription



## Instant Drive Analysis



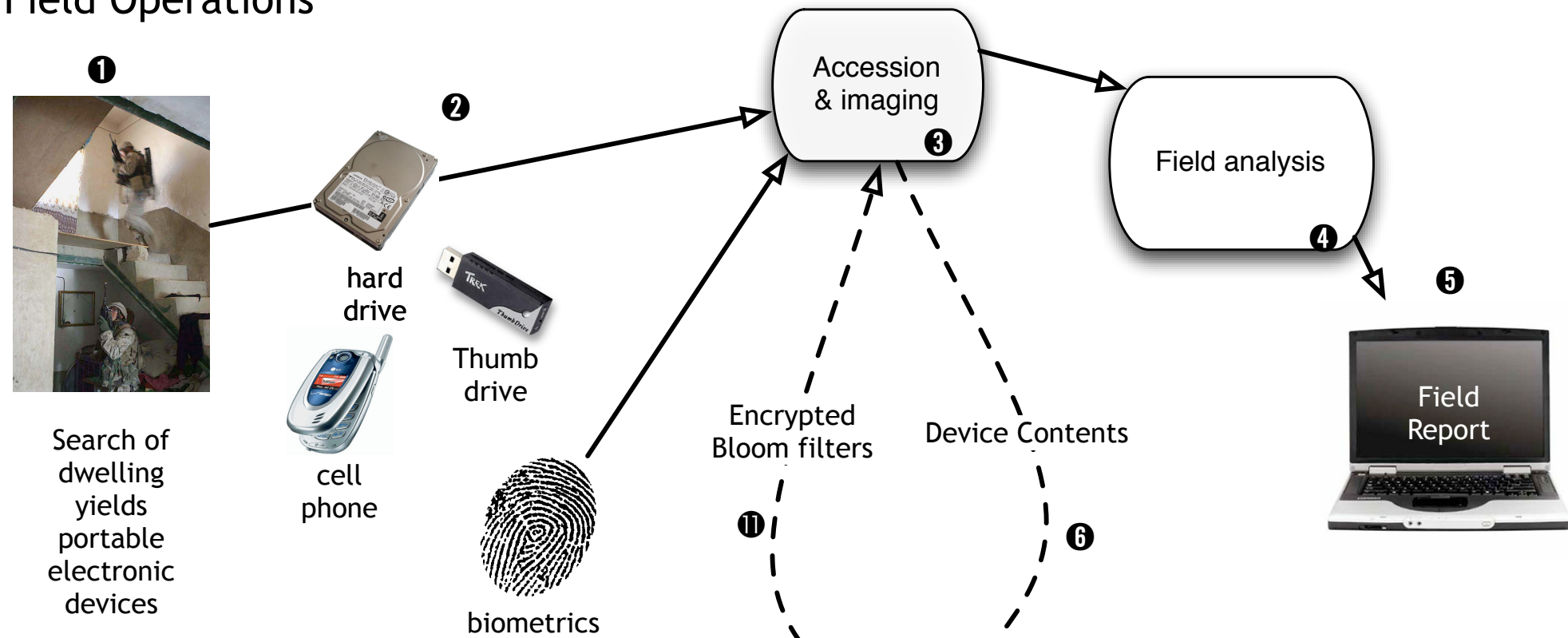
## Standardized Forensic Corpora



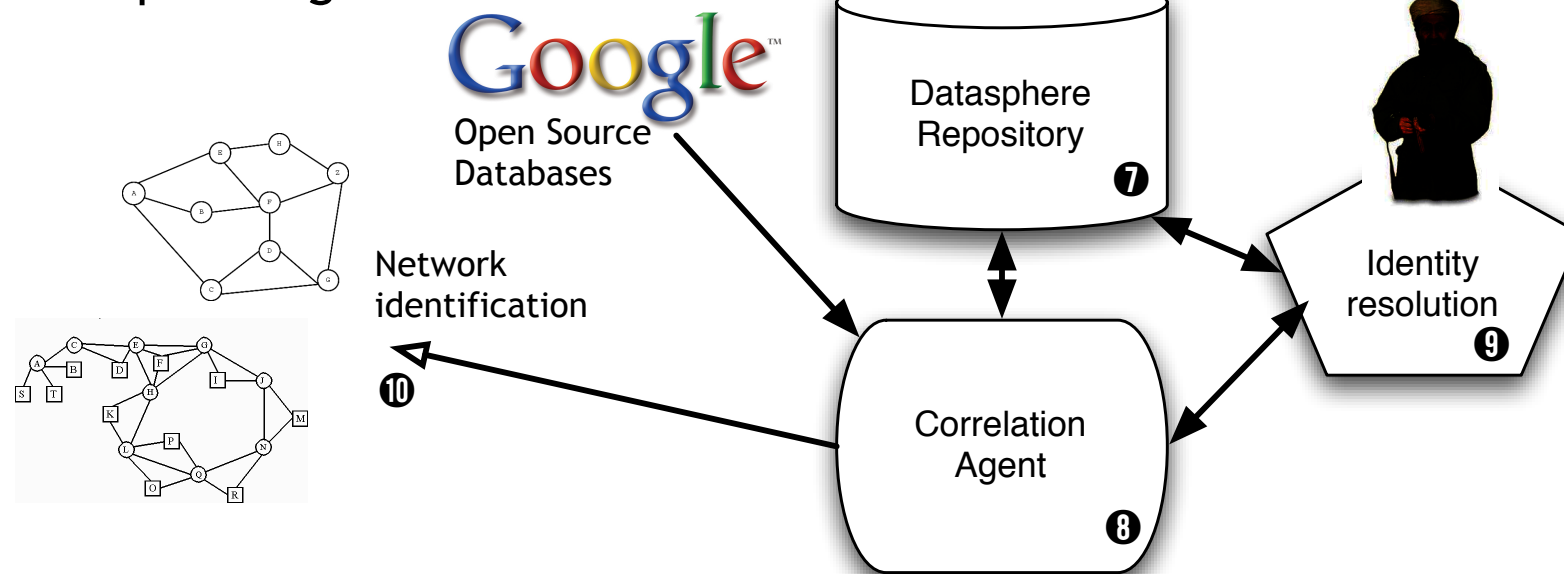


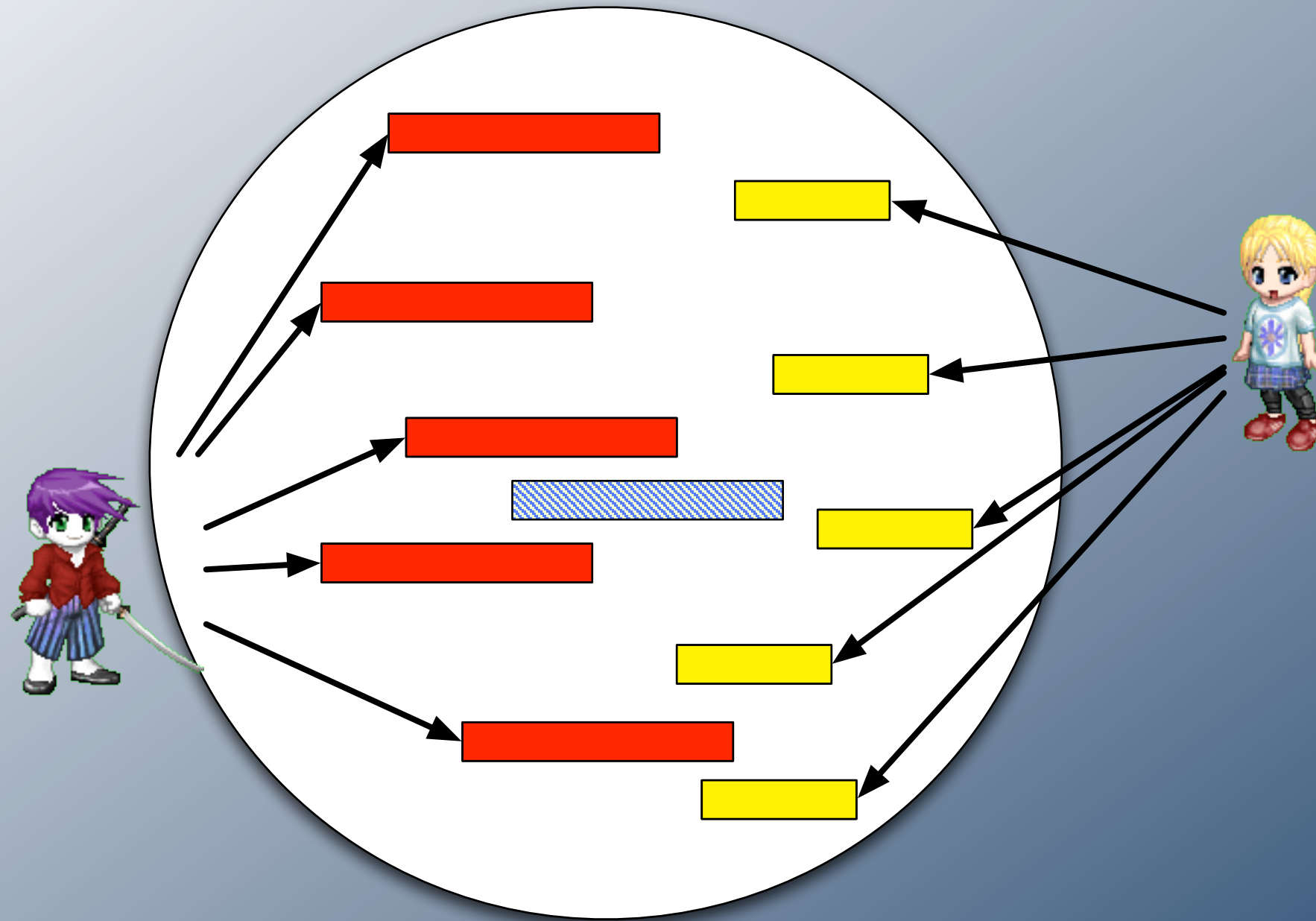
# Never forget the big idea.

## Field Operations



## Forward Operating Base





# Automated Ascription of Multi-User Data

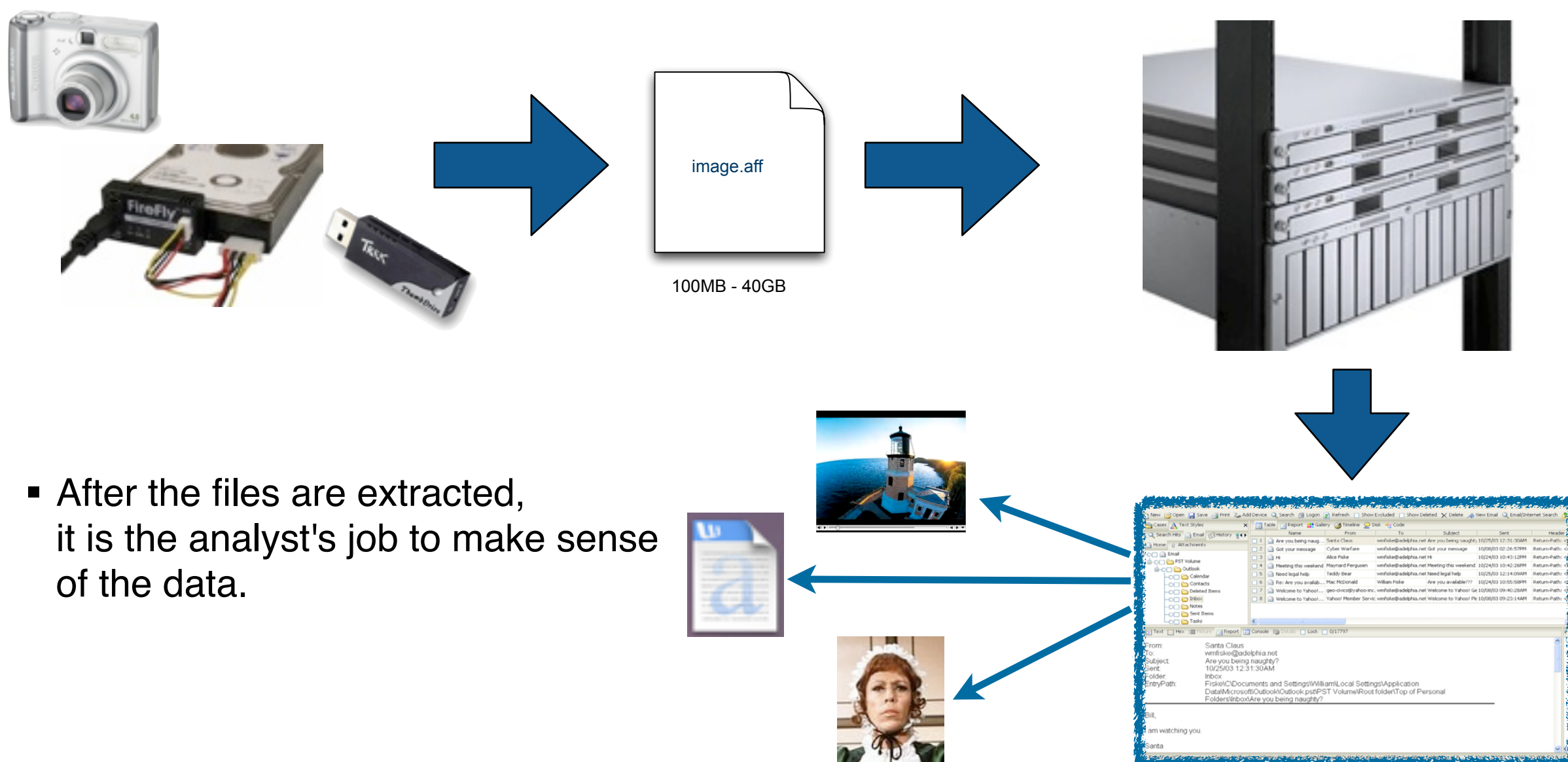


# Today's forensic tools are designed to extract files.

Step 1: Physical device is *imaged*.

Step 2: Disk image is stored on a high-capacity storage device.

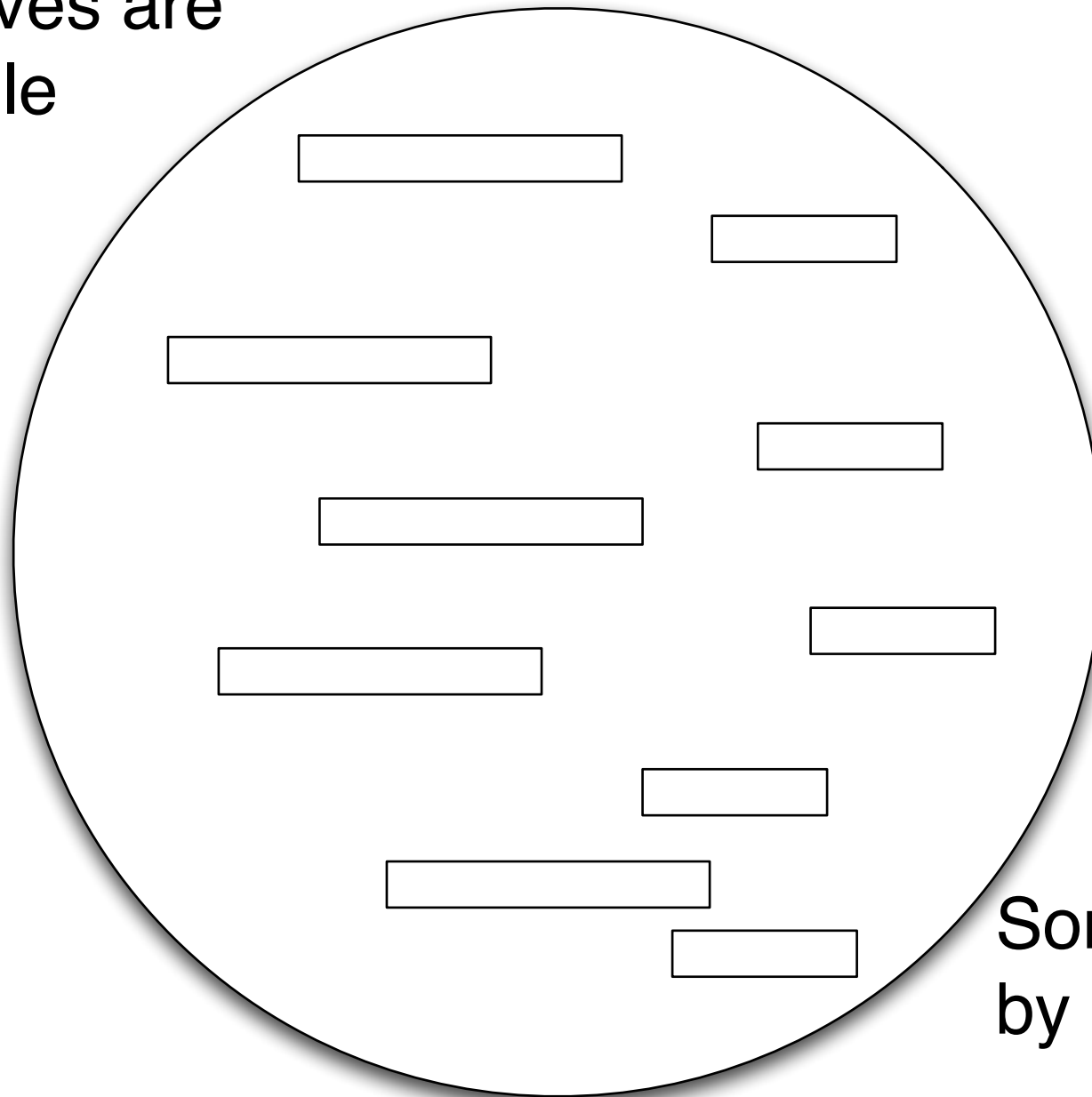
Step 3: Tools process the image and extract files



- After the files are extracted, it is the analyst's job to make sense of the 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.





Some drives are used  
by multiple people.





# Prior work has used *content analysis* to determine authorship

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

# My research uses metadata to infer *ownership or agency* — who is *responsible* for the data.

## File system metadata (“extrinsic”):

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

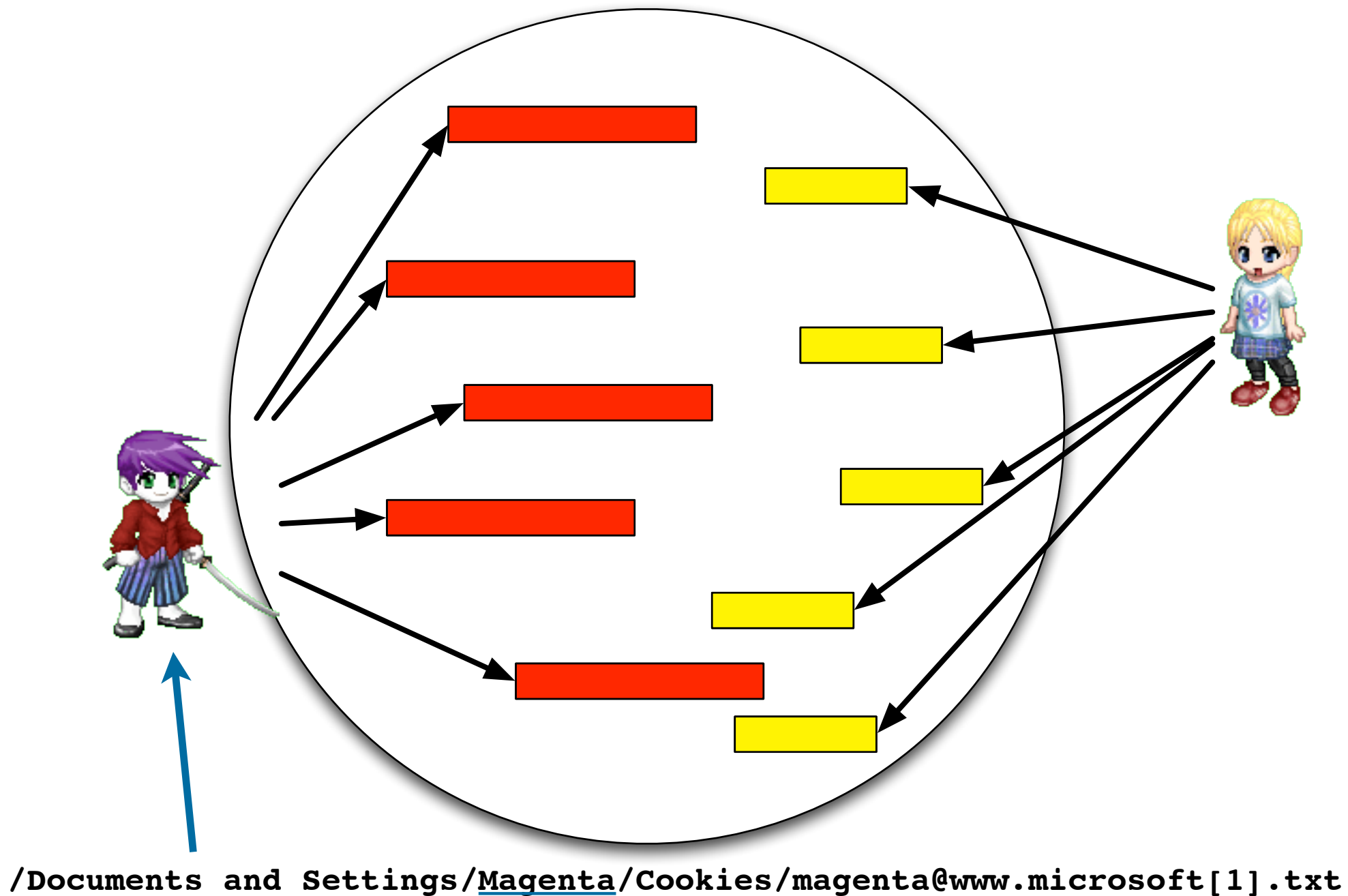
## File metadata (“intrinsic”):

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

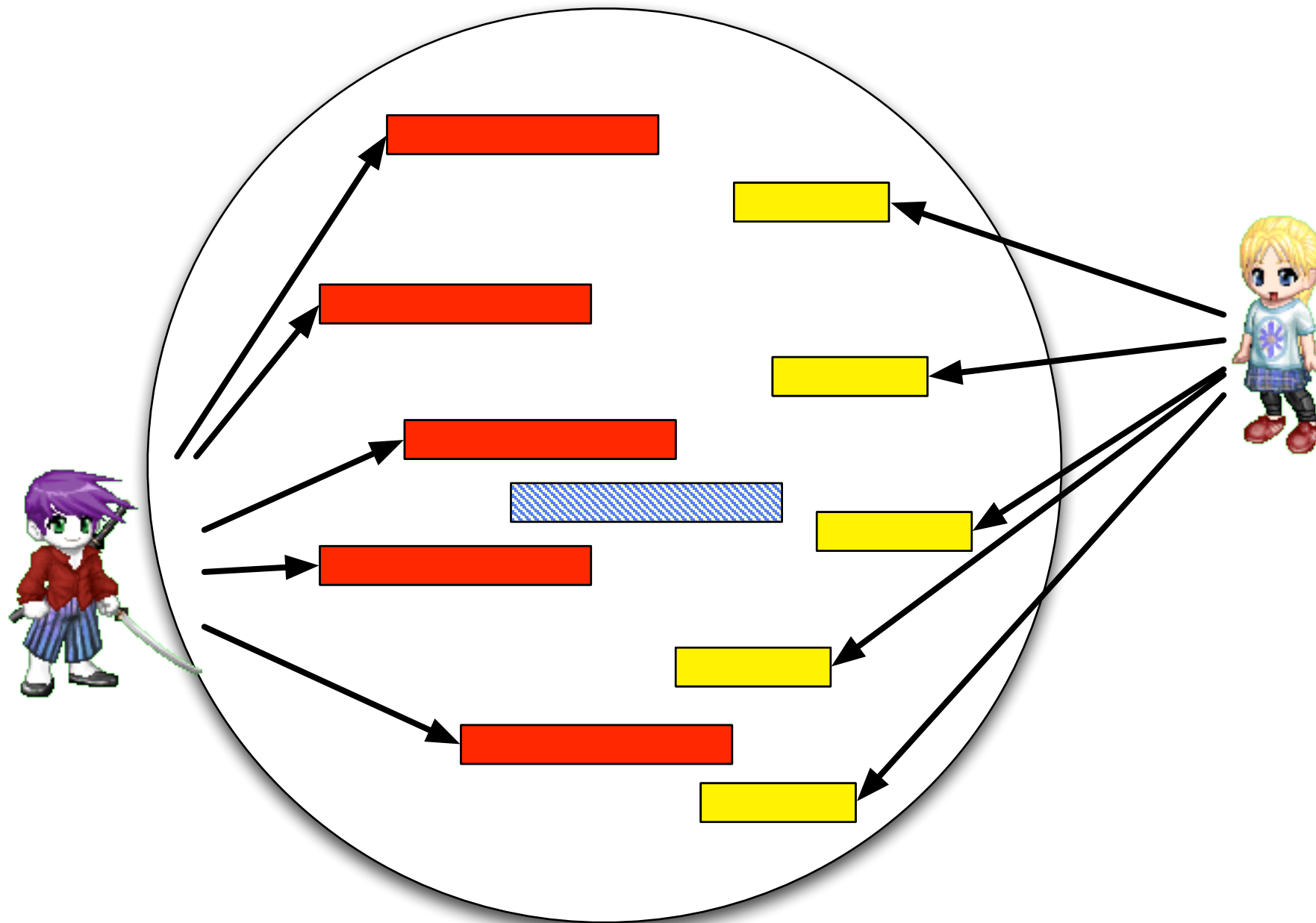




Many files can be ascribed to a specific user using *filename* or *file system metadata*.

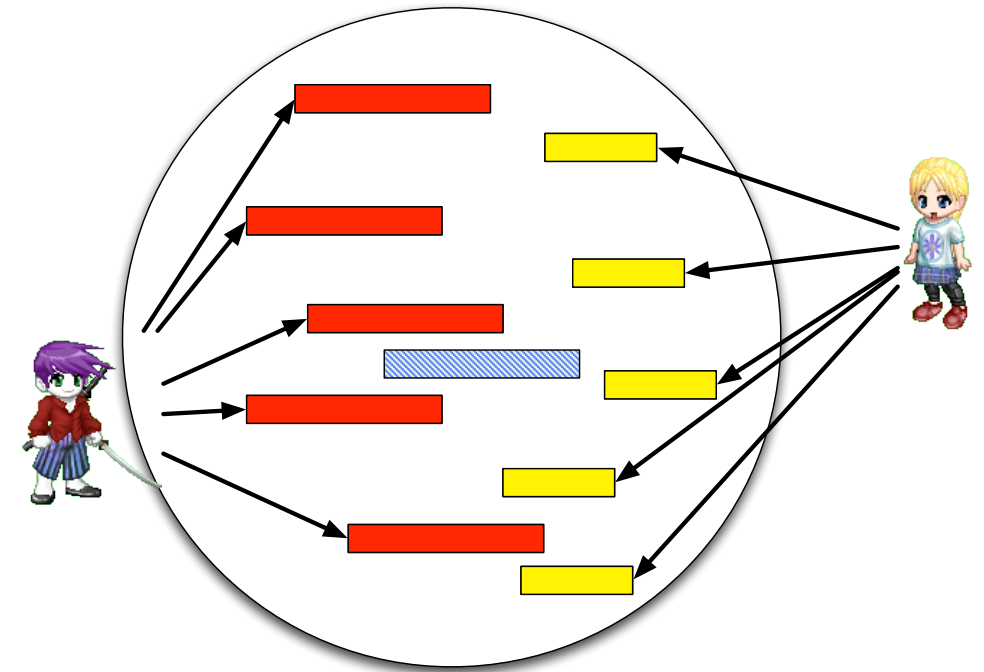


# Files recovered with “carving” can’t be readily ascribed.




Who is responsible for the file?

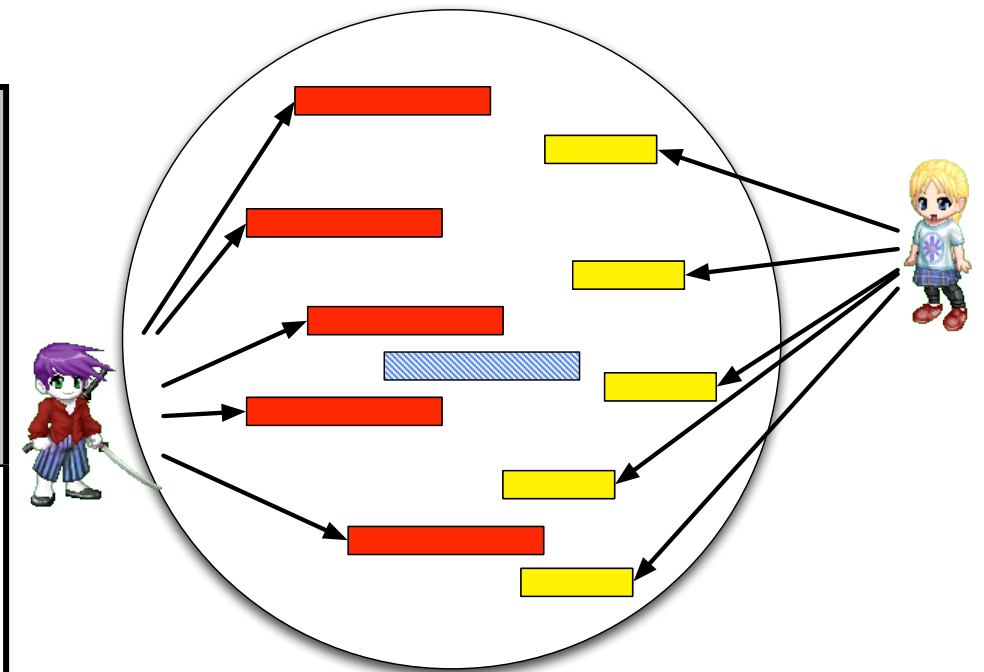
By surveying exemplar files on the disk, an examiner can create ascription rules.






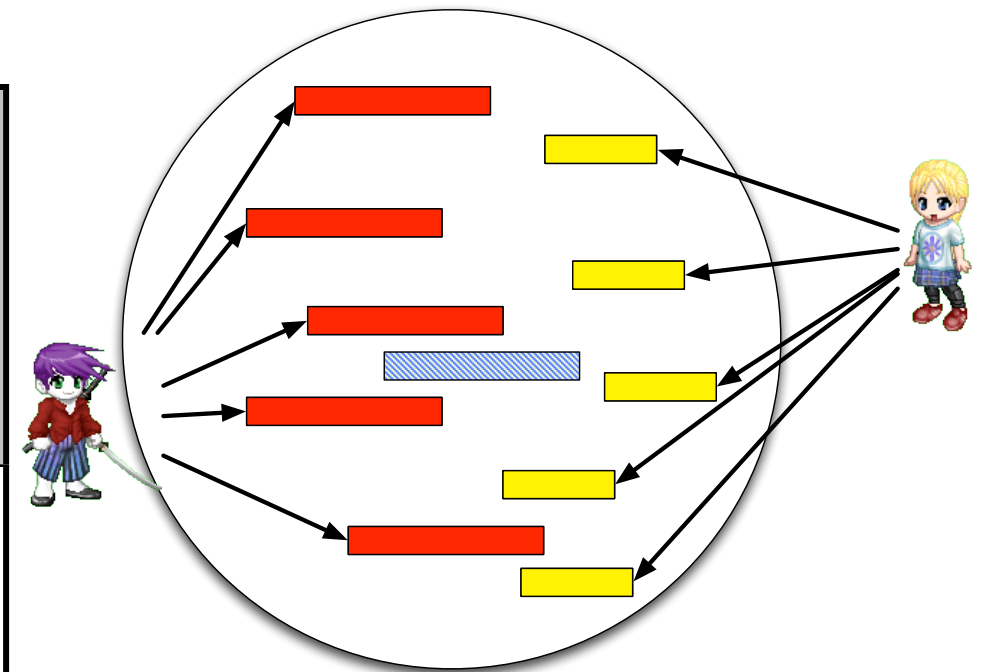
By surveying exemplar files on the disk, an examiner can create ascription rules.

Magenta	Yellow	Carved 	Likely User




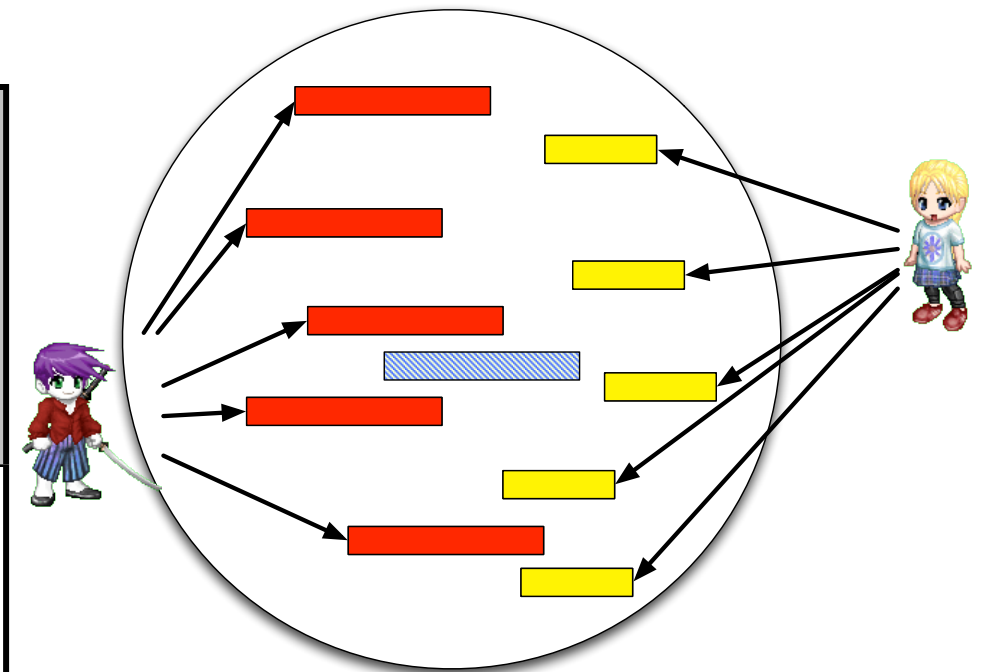
By surveying exemplar files on the disk, an examiner can create ascription rules.

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





By surveying exemplar files on the disk, an examiner can create ascription rules.

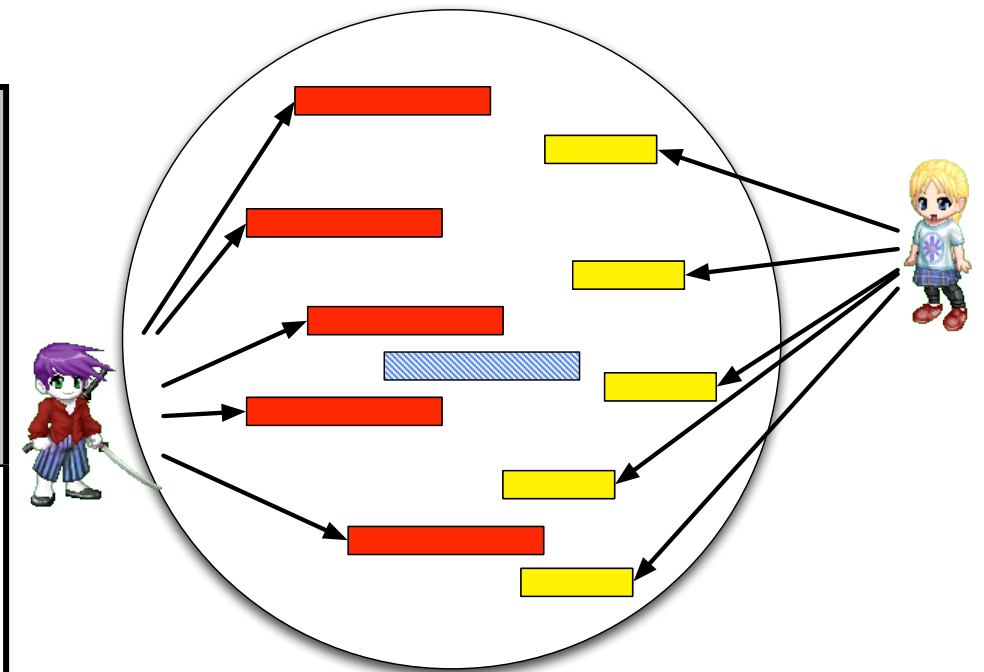
Magenta	Yellow	Carved 	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	







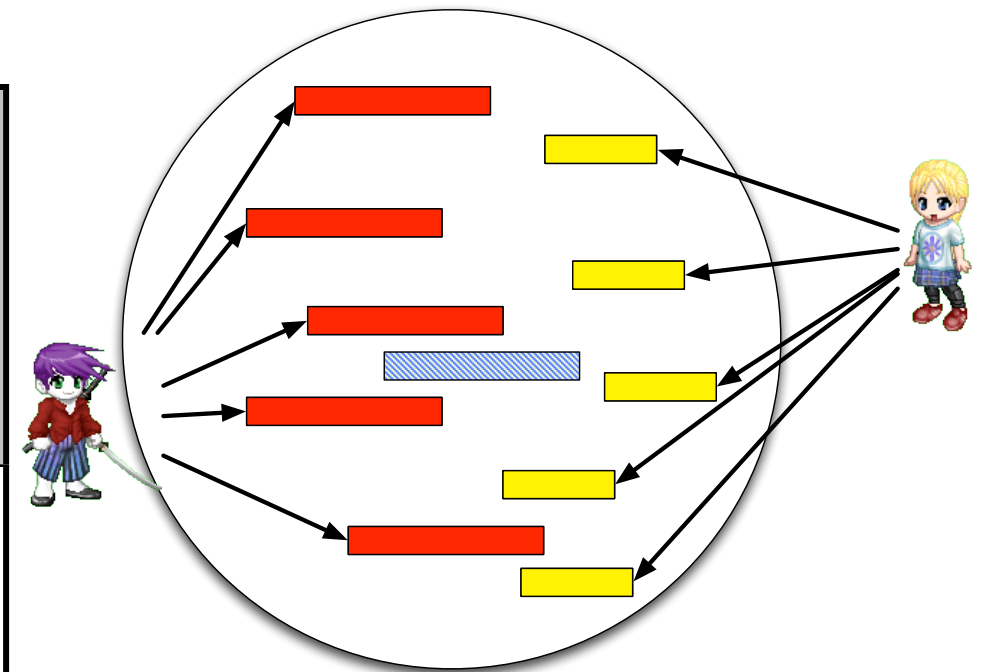
By surveying exemplar files on the disk, an examiner can create ascription rules.

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





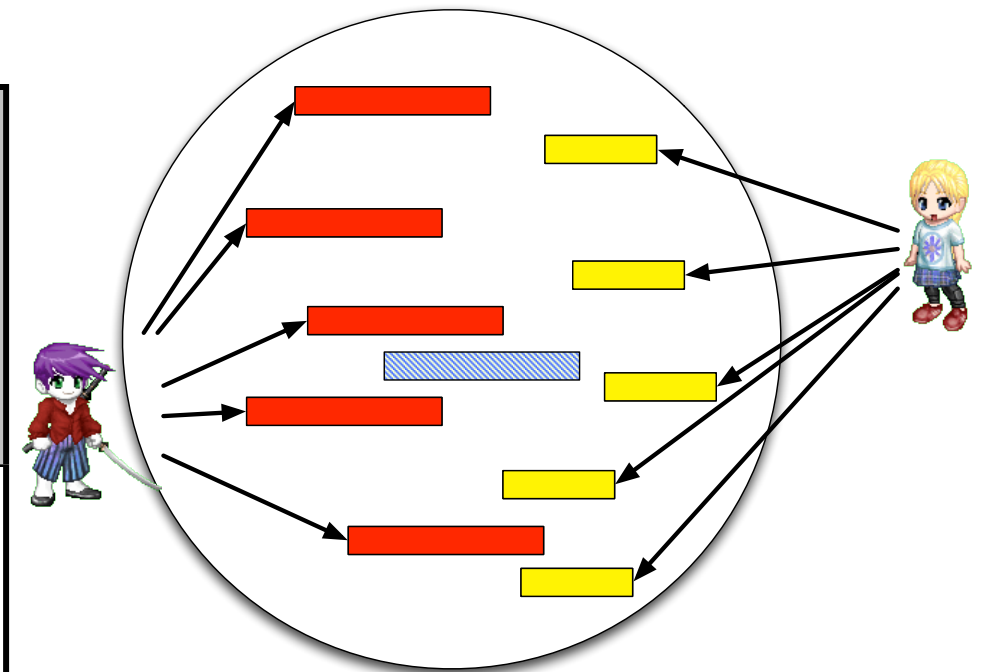
# By surveying exemplar files on the disk, an examiner can create ascription rules.

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






# By surveying exemplar files on the disk, an examiner can create ascription rules.

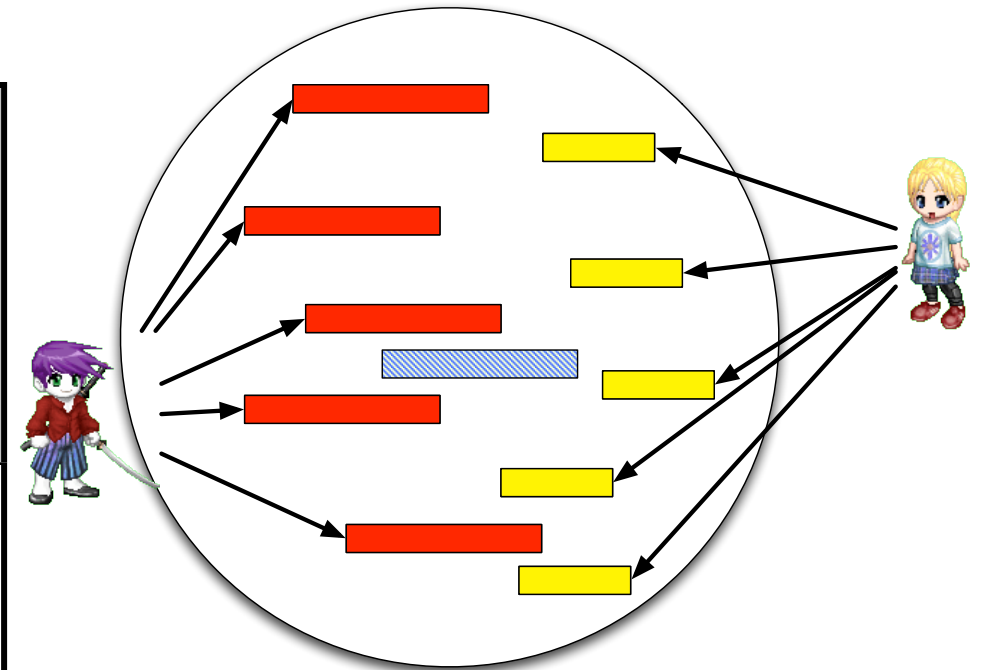
Magenta	Yellow	Carved 	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	








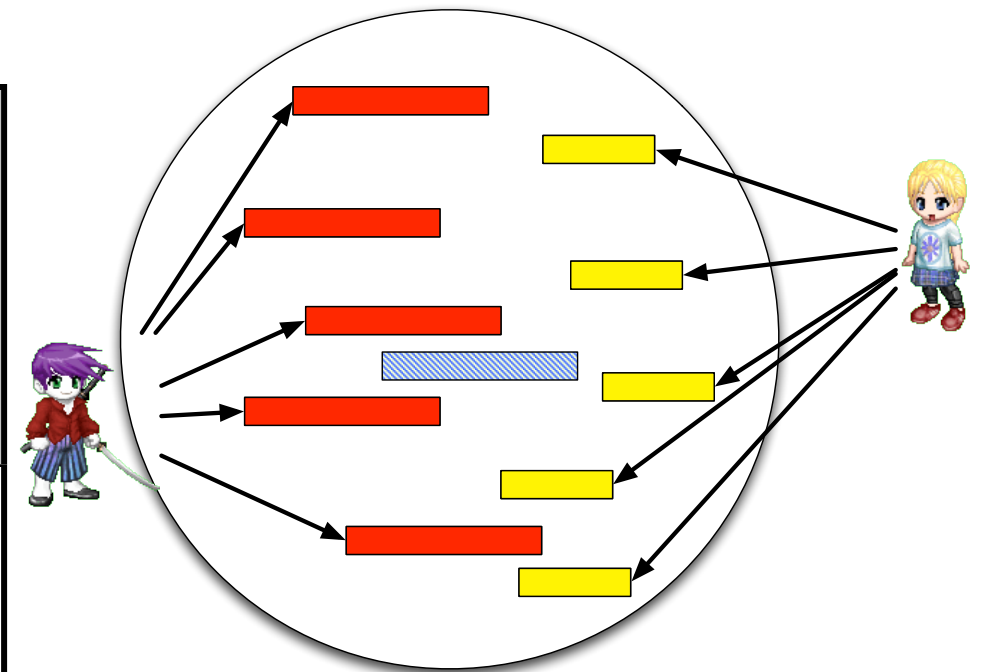
By surveying exemplar files on the disk, an examiner can create ascription rules.

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






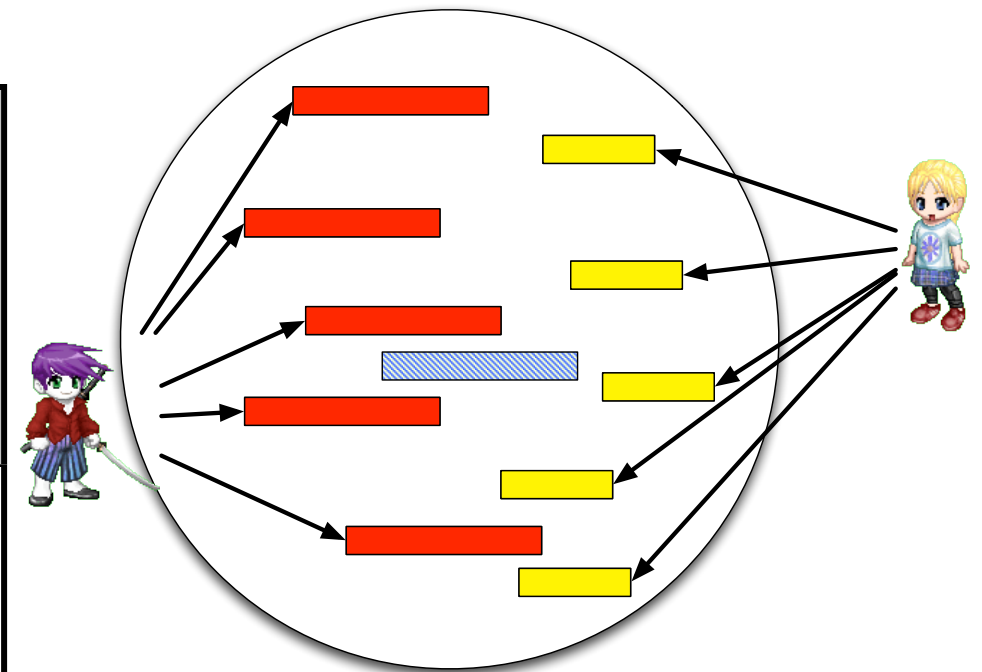
# By surveying exemplar files on the disk, an examiner can create ascription rules.

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		







# By surveying exemplar files on the disk, an examiner can create ascription rules.

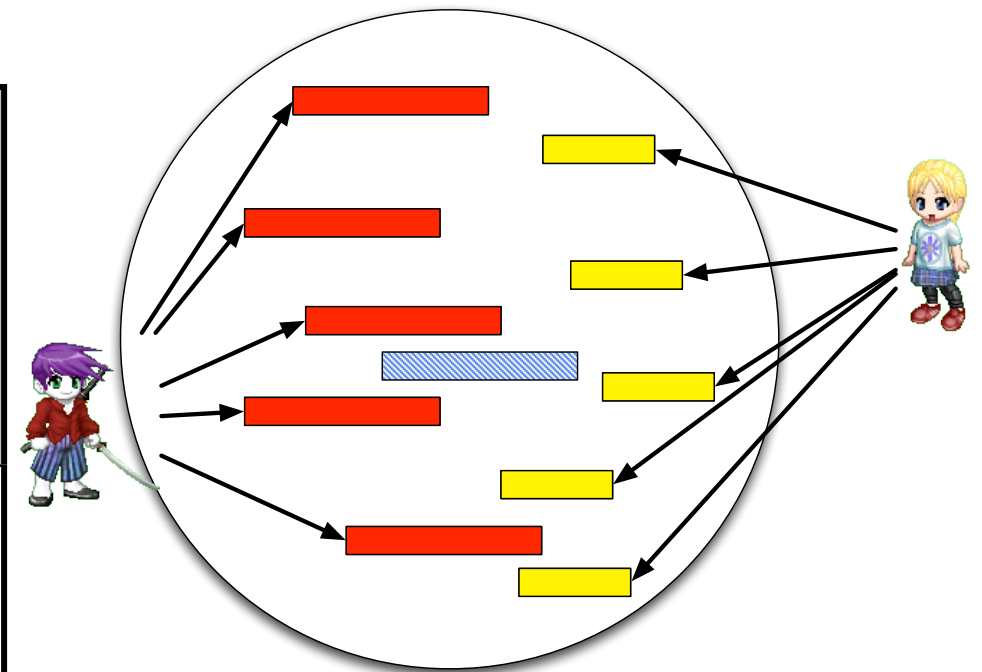
Magenta	Yellow	Carved 	Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Printed 9am, 10am, 11am	Printed 8pm, 9pm	Printed 8:30pm	
At sectors 10, 20, 30	At sectors 500, 600, 700	Sector 550	



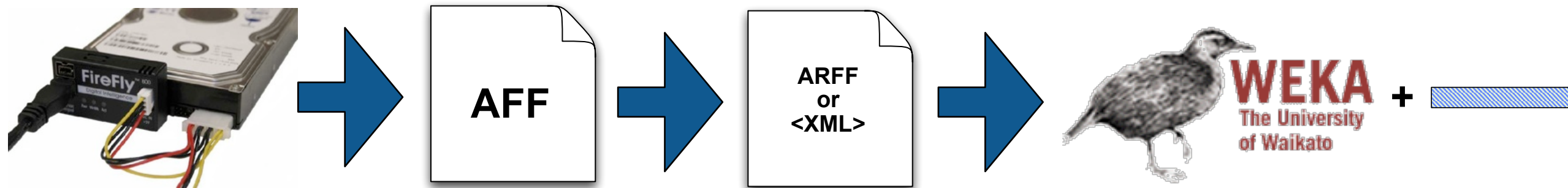


# By surveying exemplar files on the disk, an examiner can create ascription rules.

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 are developing a toolset 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.

# 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:

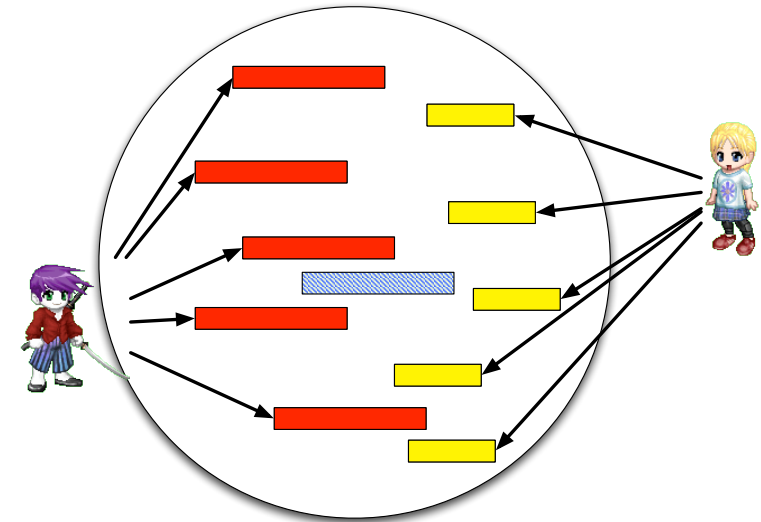
```
<fileobject>  
...  
<Manufacturer>SONY</Manufacturer>  
<Model>CYBERSHOT</Model>  
<Orientation>top - left</Orientation>  
...  
</fileobject>
```



# 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.

## Can't use "kernel methods"

- No linear mapping means no "kernel trick."



# Approach #1: Decision Tree

## Algorithm: J48

- Implementation of Quinlan's 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)
```

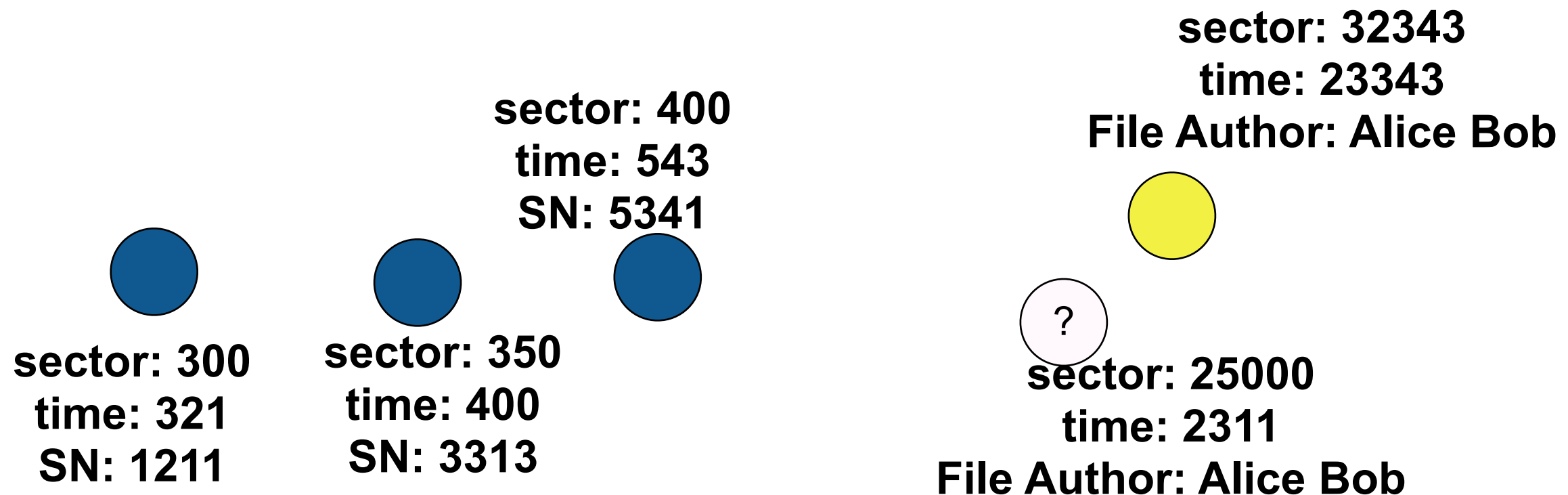
## Differences from traditional data mining:

- Every HD has its own classifier:
  - *Use cross-validation to determine the accuracy of the classifier for this HD.*
- Every carved file has its own classifier
  - *Only use the dimensions that matter for this piece of carved data.*

# Approach #2: K-Nearest-Neighbor

## Special Features:

- N=1 works best
- 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.*



# Work to date:

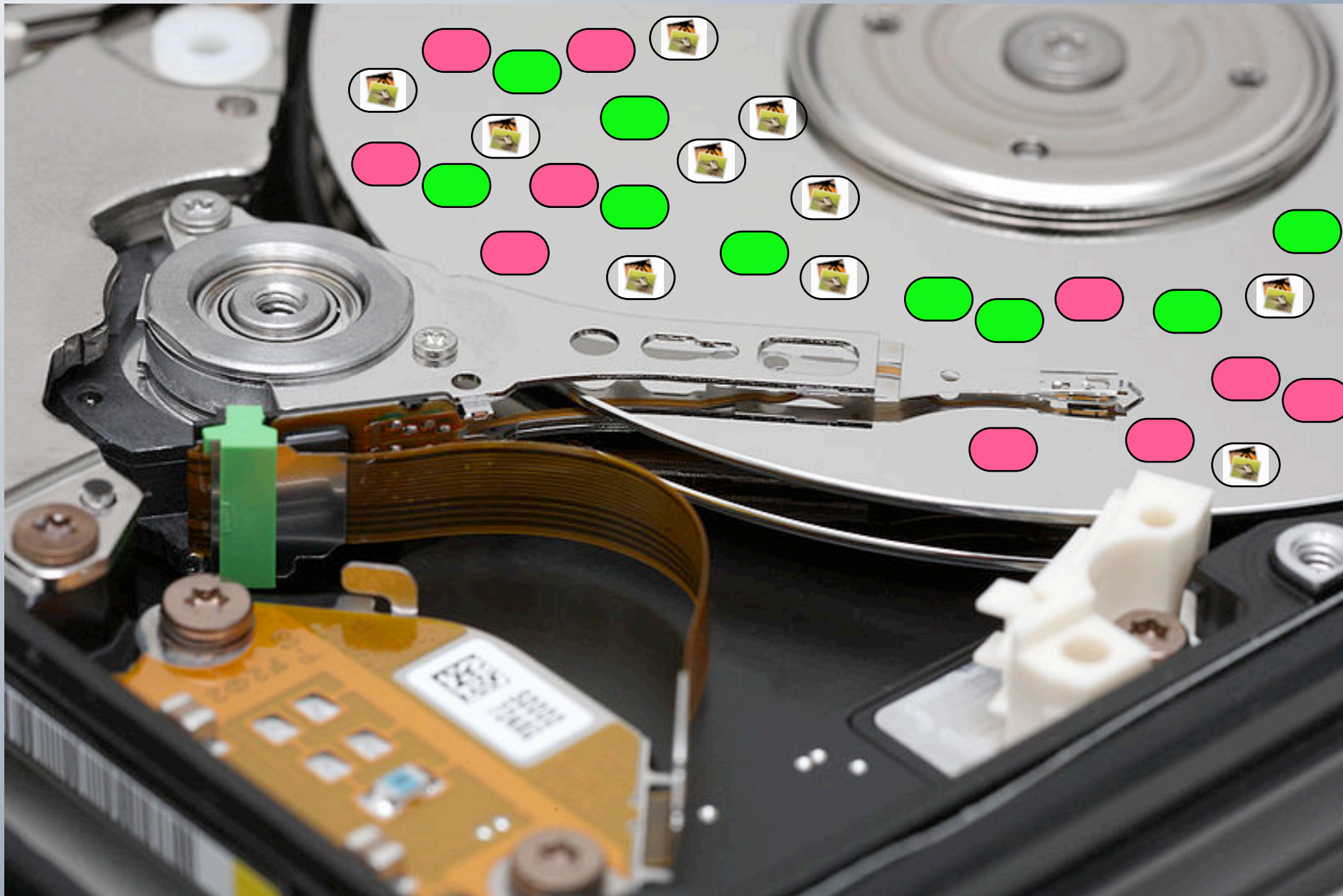
## Student Theses:

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

## Work in progress:

- Aleatha Parker-Wood (PhD Candidate, UCSC)
- Using “Hamming,” our 1100-core cluster.
- J48:
  - *89.5% accuracy on “realistic” data (nps-2009-domexusers)*
  - *93.55% accuracy on real data (0844)*
- K-Nearest Neighbor
  - *Less accurate than J48*
  - *Better when  $N=1$  than  $N=3$  or  $N=10$*
  - *Explanation: only works if there is a nearby exemplar.*





# Instant Drive Forensics with Statistical Sampling

# Research Question: Is it possible to 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	15 million	72,000

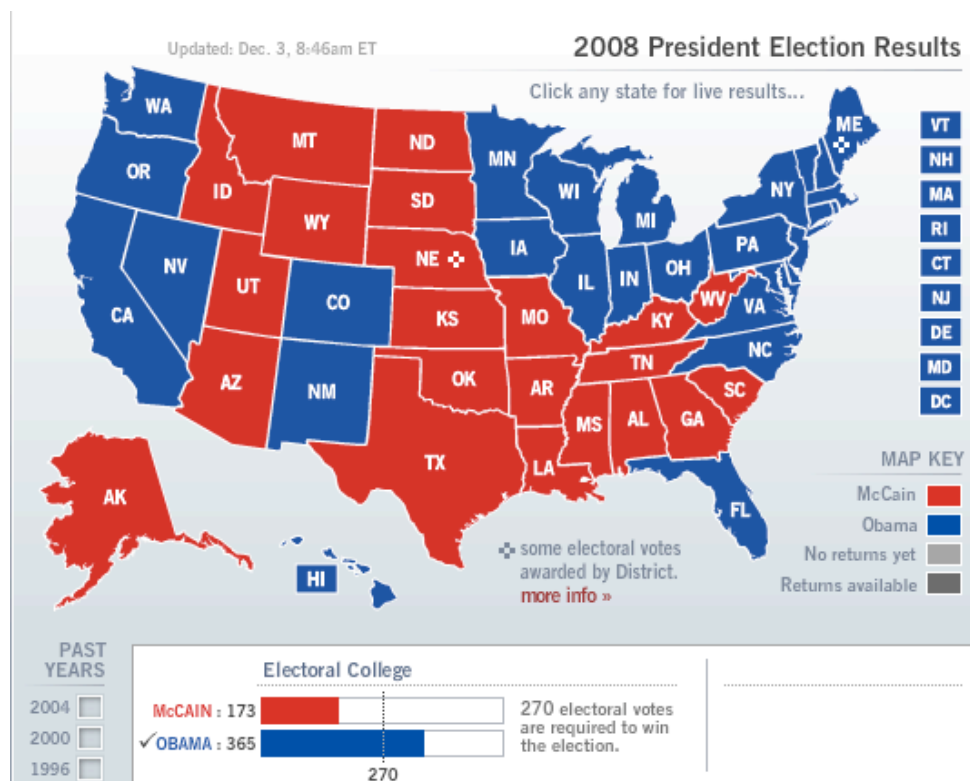
## 7.2 GB is a lot of data!

- $\approx 0.48\%$  of the disk
- But it can be a statistically significant sample.



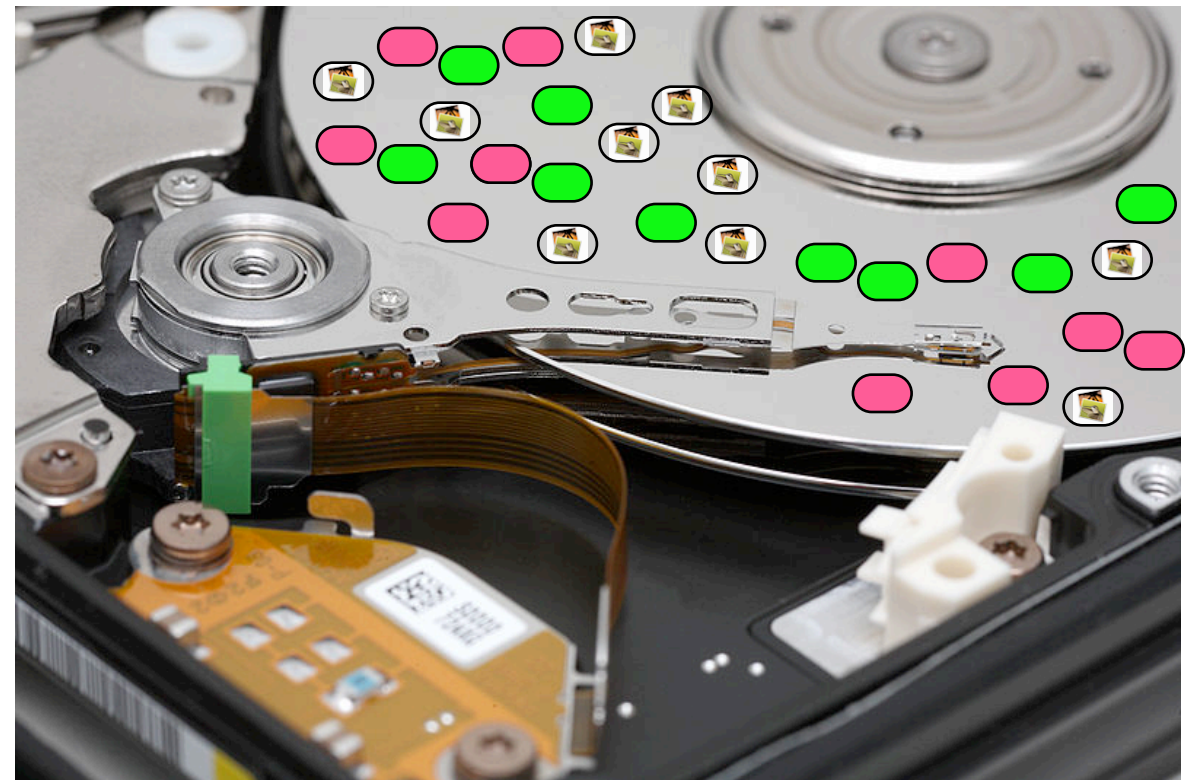
We can predict the statistics of a *population* by sampling a *randomly chosen sample*.

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



The challenge is identifying *likely voters*.

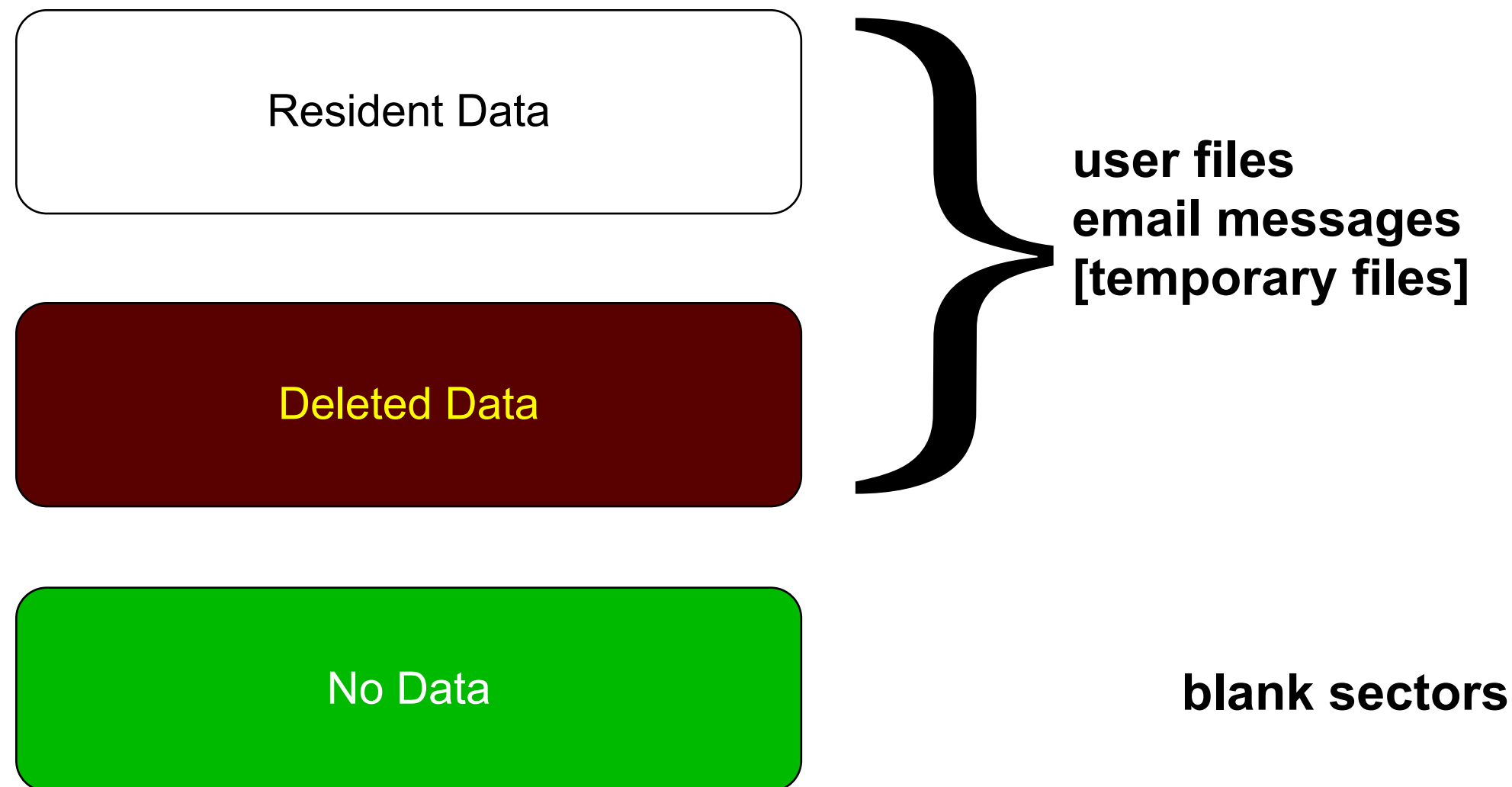
Hard drive contents can be predicted by sampling a few thousand sectors:



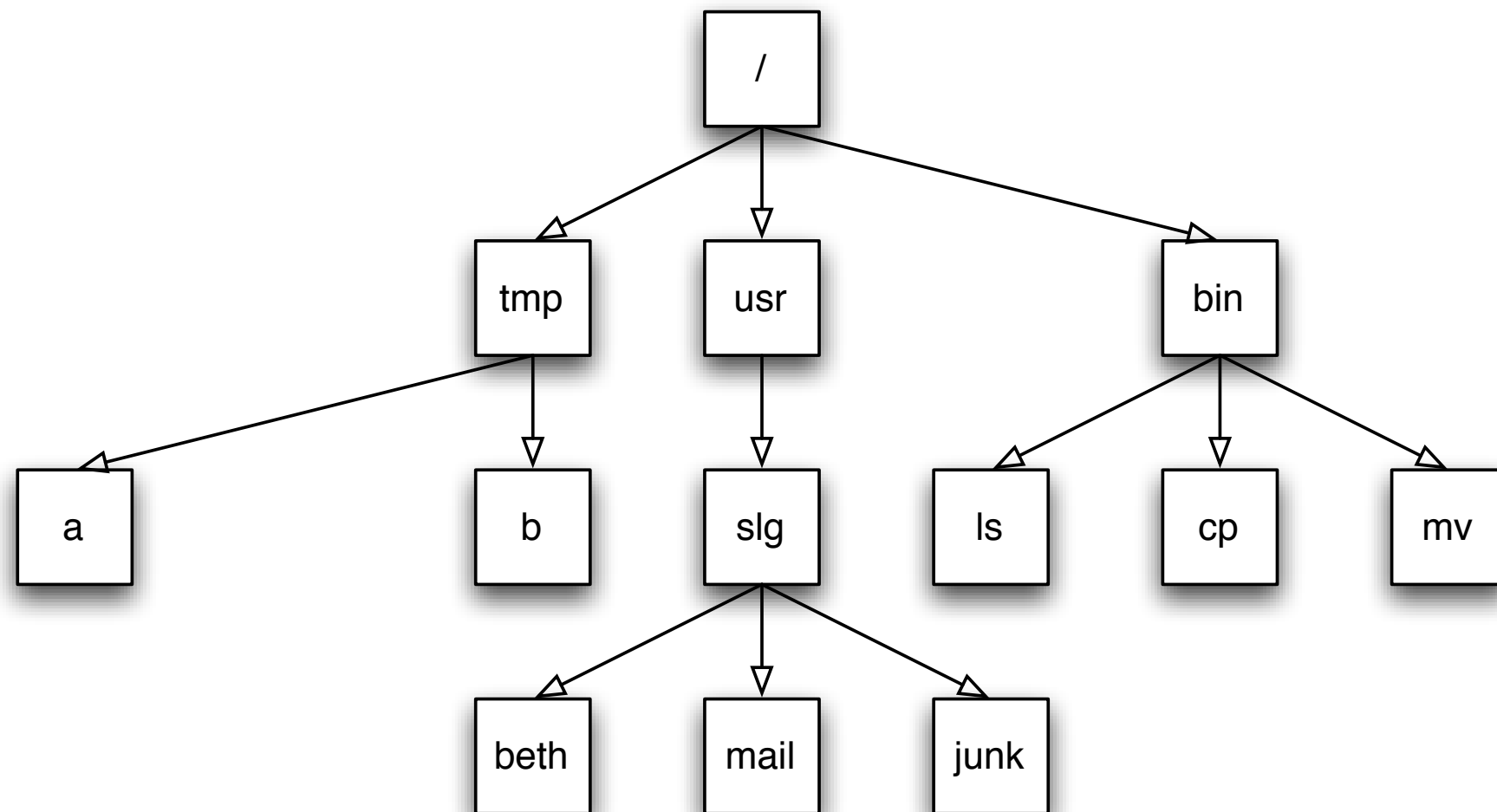
The challenge is *identifying the sectors* that are sampled.



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

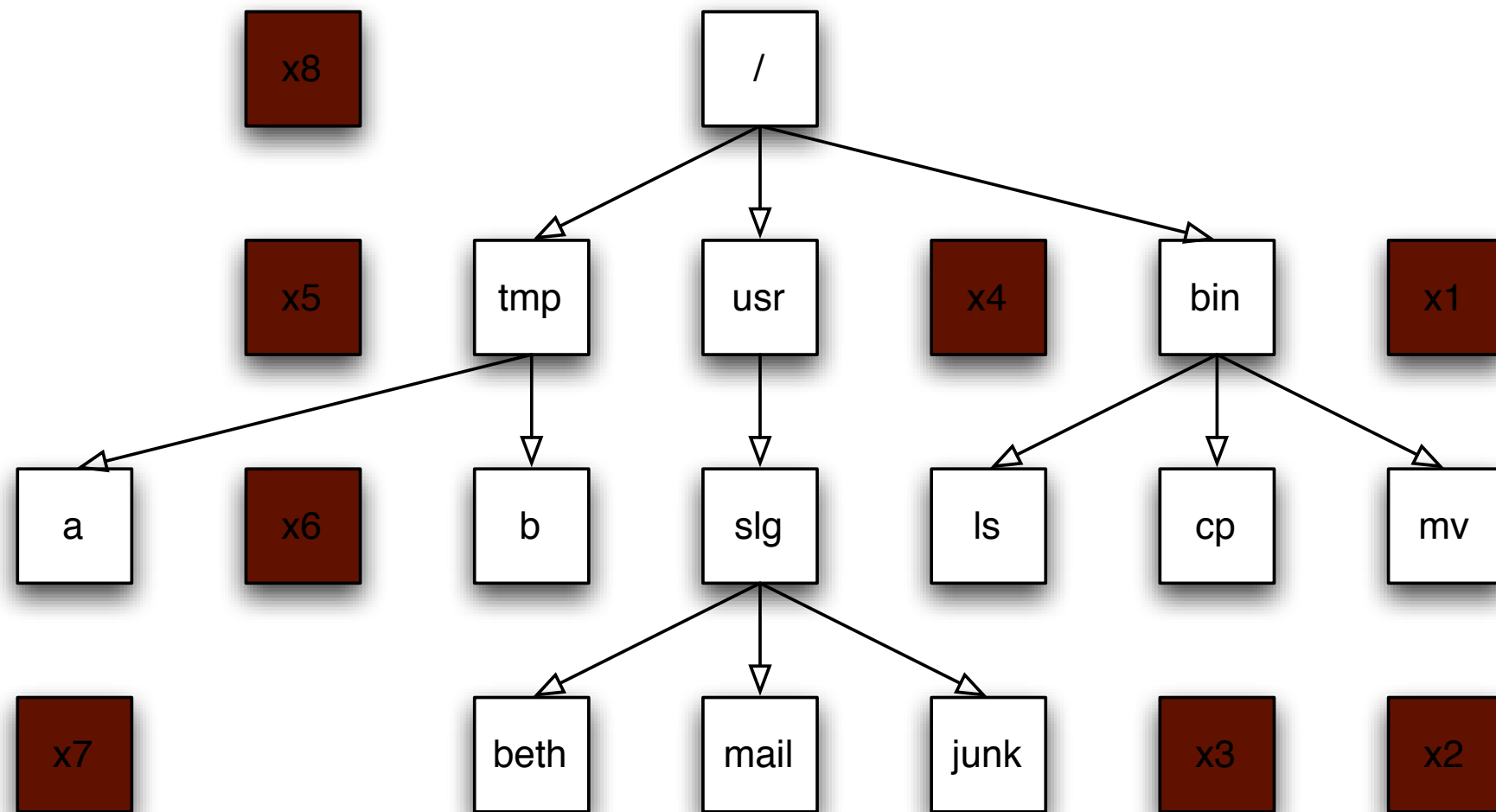


Resident data is the data you see from the root directory.



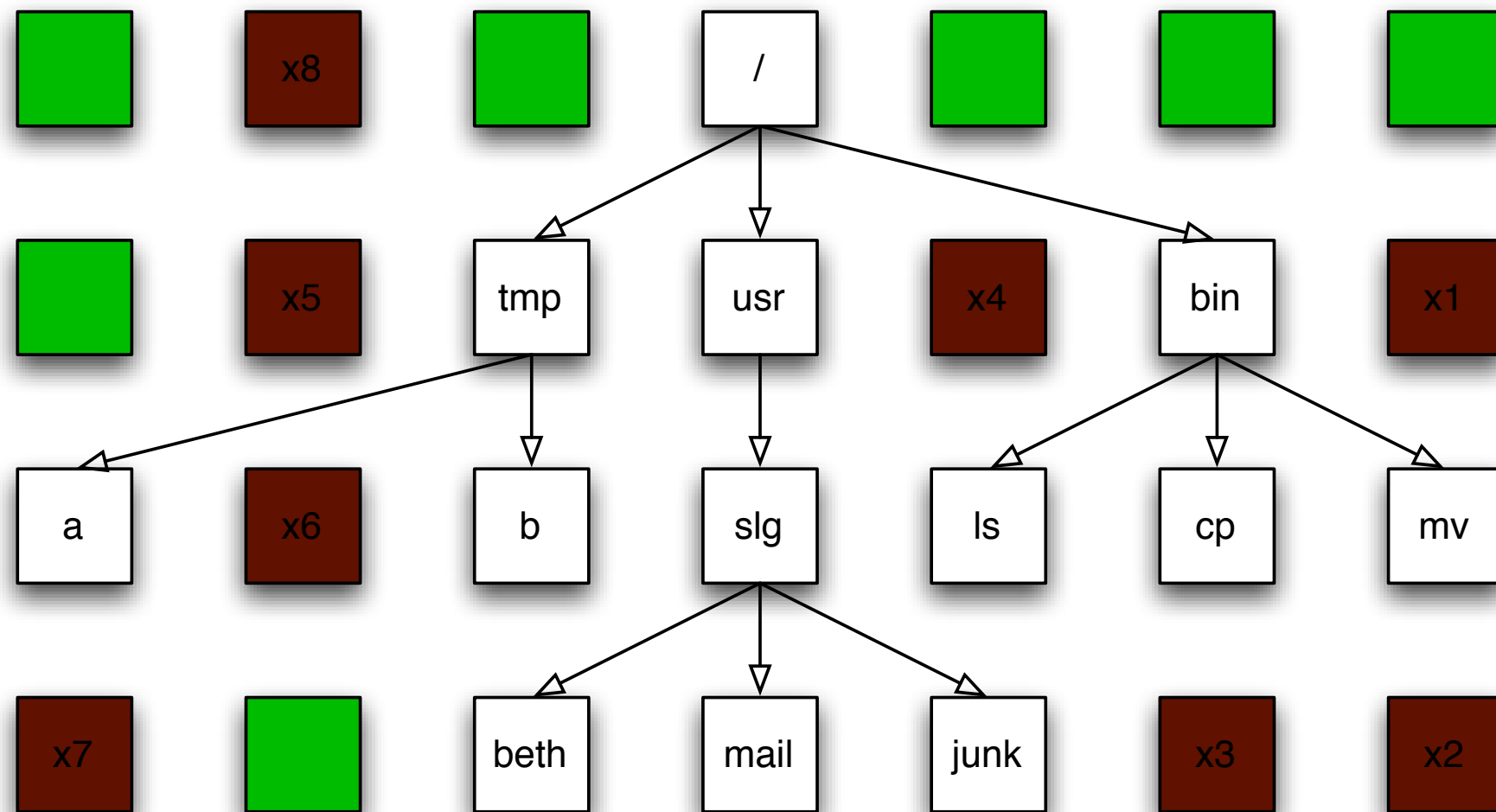
Resident Data

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



Deleted Data

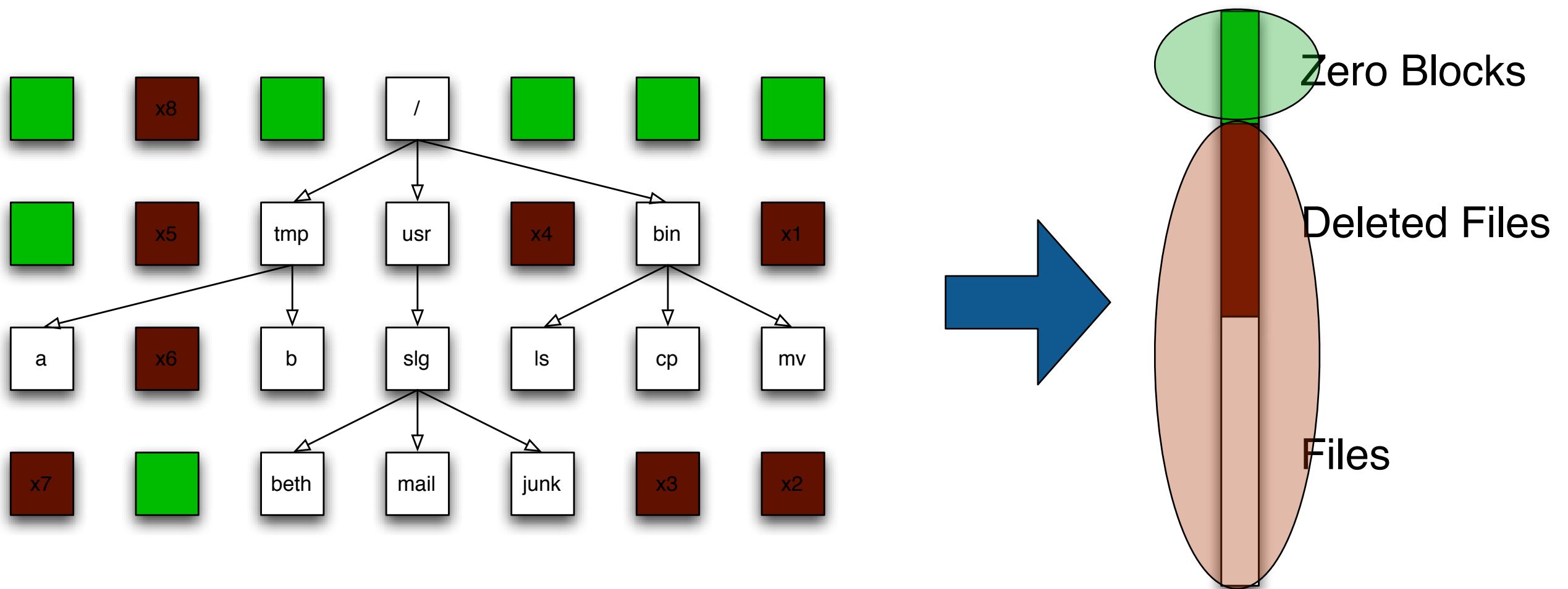
# Sectors with “No Data” are blank.



No Data



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



# What are the proper statistics for evaluating the sample?

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

- Well, what does it mean if 1 randomly chosen sector is blank?
- A 1TB hard drive has 200,000,000 sectors.
  - If the drive has 1 blank sector and 199,999,999 data sectors: We are very lucky!
  - If the drive has 200,000,000 blank sectors and 0 data sectors: The sector is typical.

Let's 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 general probability of missing one of  $F$  non-blank sectors when sampling  $N$  of  $T$  sectors is

$$p = \prod_{i=1}^N \frac{((T - (i - 1)) - F)}{(T - (i - 1))}$$

The more you sample, the better the chance of not missing something:

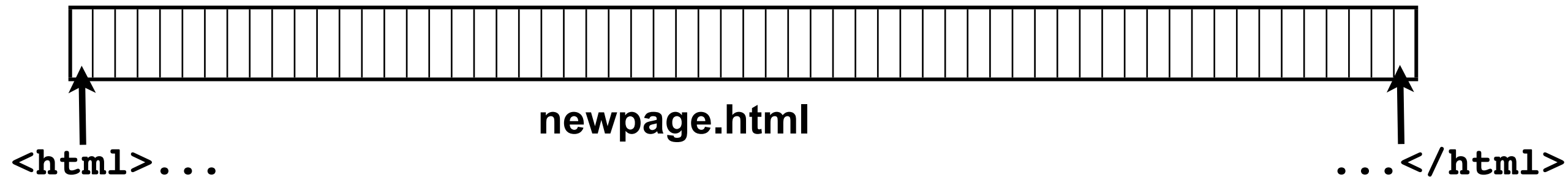
Sampled Sectors	Odds of not finding data
2	0.9998
100	0.9900
1000	0.9048
10000	0.3679
20000	0.1353
30000	0.0498
40000	0.0183
50000	0.0067

**Table 1:** Odds of not finding 10MB of data for a given number of randomly sampled sectors

- So if you sample 50,000 sectors of a 1TB drive and find all blank sectors, there is a 0.67% you missed 10MB of data, or a 99.33% chance the drive has less than 10MB data.

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

A file 30K consists of 60 sectors:



Many file types have characteristics headers and footer:

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

But what about the file in the middle?

# Fragment classification:

## Different file types require different strategies.

HTML files can be reliably detected with HTML tags

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

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.



# Many files are "container" files.

## Classifying sectors from these files will classify contents.

The PDF file format consists of:

- PDF header
- PDF xref table (a directory of objects in the PDF file)
- PDF objects (T/F; Numbers; Strings; Names; Arrays; Dicts; Streams; Null)

PDF header:

```
Terminal — emacs-i386 — 82x20
%PDF-1.3
%\304\345\362\345\353\247\363\240\320\304\306
4 0 obj
<< /Length 5 0 R /Filter /FlateDecode >>
stream
x^A\315W\313n\2030^P\274\363^U^s\^N^P/\306\330\233\266Ro\215\204\324s\205\210\222\
\2524MH\245~^W\232WK\224^FKD+#\220\260^Y\217\355\231\335e\205)V \207\204`l
\2535\326%\236\361\216\361\244&\250mu^A\262\221\341q|;\29k^A\^F\243-\264\311^P\
+^]\354^P\236\312uQ~l>_ \336\260^A\360<\277\342\306c\260\375\244\2500~\254^HwK^F\3\
52p\331#M\352\270\341\242\300<\230o\213^R^Rb\333\320\321^L\3078\2679\250AW^H\371^A\
Y\222K\271/\244,A\316\223<P\244^BB>\303M>_ \324\340\253^V%6\345\327^F#\344\257\270\
\317^?^H\^D\256\344\254\315Z\374` \207\317\304N\342w&<\377b^T\364e\263]\252\213\3\
67K\335RiVw~2\217^Q\336\374\254\222\315\3178\331\374\222T6?}\260\232H\375\305\302\
\375\301a\3530(\271\314\235\276\206t^]=\211\211^UV\266\224R\331J2\236B\352y\374\2\
76\272\323\262\343l\334\261\305e6\274\322\356\221lk\210v\206o\204\275\316\321\312\
\3160\247\263\273g^L\350\226Z=\267\370^?Kz2\223]\244\312.\361\217\376\265^D^V\200\
\244^F^K\354\236Z#\335I\324Rl@211\354@nDgAJ^G\223\332/\201\370\352\316^NV\201^J^A\
XM\277^A&\371<\314
-:---F1 file1.pdf Top L?? (Text)----11:32AM 0.43 Mail-----
```

PDF xref:

```
Terminal — emacs-i386 — 82x20
xref
0 21
0000000000 65535 f
0000073226 00000 n
0000000496 00000 n
0000067942 00000 n
0000000022 00000 n
0000000477 00000 n
0000000600 00000 n
0000067041 00000 n
0000000761 00000 n
0000066105 00000 n
0000067905 00000 n
0000073051 00000 n
0000066126 00000 n
0000067021 00000 n
0000067077 00000 n
0000067885 00000 n
-:---F1 file1.pdf 99% L?? (Text)----11:33AM 0.27 Mail-----
```

Individual "objects" in a PDF file may hold JPEGs, Text, FAX pages, JavaScript, and other content.

## PDF JPEG:

```
Terminal — emacs-i386 — 82x20
6 0 obj
<< /ProcSet [ /PDF /Text /ImageB /ImageC /ImageI ] /ColorSpace << /Cs1 7 0 R
/Cs2 10 0 R >> /Font << /F1.0 11 0 R >> /XObject << /Im1 8 0 R >> >>
endobj
8 0 obj
<< /Length 9 0 R /Type /XObject /Subtype /Image /Width 800 /Height 533 /ColorSpac\
e
  0 R /Interpolate true /BitsPerComponent 8 /Filter /DCTDecode >>
stream
\377\330\377\340^@APJFIFA^AAAAA^AA^AA^@\377\333^@C^@AFADAEFAEADAF^EAFAGAGF\
AH
AP

ATANADALAPAWATAXXAXWATAVVVAZ^]%_AZ^[#^\VVV , #&'*)^YA_-0-(0%()(\377\333^@CAA^G\
AGAG
AH
AS
--F1 file1.pdf      1% L??   (Text)----11:34AM 0.26 Mail-----
Undo!
```

PDF Text:

```

12 0 obj
<< /Length 13 0 R /N 3 /Alternate /DeviceRGB /Filter /FlateDecode >>
stream
x^A\205\224MH^a^X\307\377\263\215^D\261^F\321\227^H\305\320\301$T&AKR^B\323\365+\
S\266e\325L    b\235}w\235^g\247\231\335-E"\204\350\230u\214.VD\207\210N\341\241\
C\247:D^D\231u\211\240\243E^P^E^"\266\377;\223\273cT\27603\277y\236\377\373|\275\
303^@U\217R\216cE4`^312\316\273\311\336\230vztL\333\374^ZU\250F^T\)\303s:^R\211^A\
\237\251\225\317\365k\365-^TiYj\224\261\326\3736|\253v\231^PP4^wd^C>,y<\340\343\2\
22\344\235-5g$^[^S\2514\331!7\270C\311N\362-362\326|\210\307C\234^V\236^AT\265S\
\2233^A\1227q";\310-E#+c> \353v\332\264\311\357\245=\355^YS\324\260\337\31079^K\33\
2\270\362\375@\333^^\323\213\212m\314^C^^\334^Cv\327U\3655\300\316^Q`^361P\305\
366=\351\317G\331\365\312\313^j\366\303)\321^Xk\372P*}\257^C6\337^@~^/\225~\334.\
\225~\336a^N\326\361\3242
n\321\327\2620\345^*P\324^[^354f\374\344^Z\375\213\203\236|U^K\260\3009\216|^A\37\
2\257^@7?^B\373^_B\333^_C\211j`^250^U\221\313\17\270\362\342"\347^Bt\346\234i\32\
7\314N\344\265f]?^242u\360h\205\326g^[M^M^Z\312\2624\337\345i\256\360\204[^T\351&L\
Y^E\316\331_\333x\217
{x\2280\366^W\271$\274\356\337\254\314\245S]\234%\232\330\326\247\264\350\352&7\2\
:---F1  file1.pdf          90% L??      (Text)---11:35AM 0.22 Mail-----

```

This 74166-byte PDF file has 144 sectors and 20 objects.

- Compressed Text
- Numbers (4)
- Media Box (margins)
- ColorSpace
- ColorSpace
- JPEG
- Compressed Text
- Array [ /ICCBased 12 0 R]
- Text
- Array [ /ICCBased 14 0 R]
- Media Box
- Page Count
- Compressed Text
- Font Table
- Font Descriptor
- Metadata



This is some text This is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some text

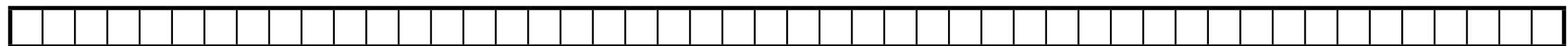
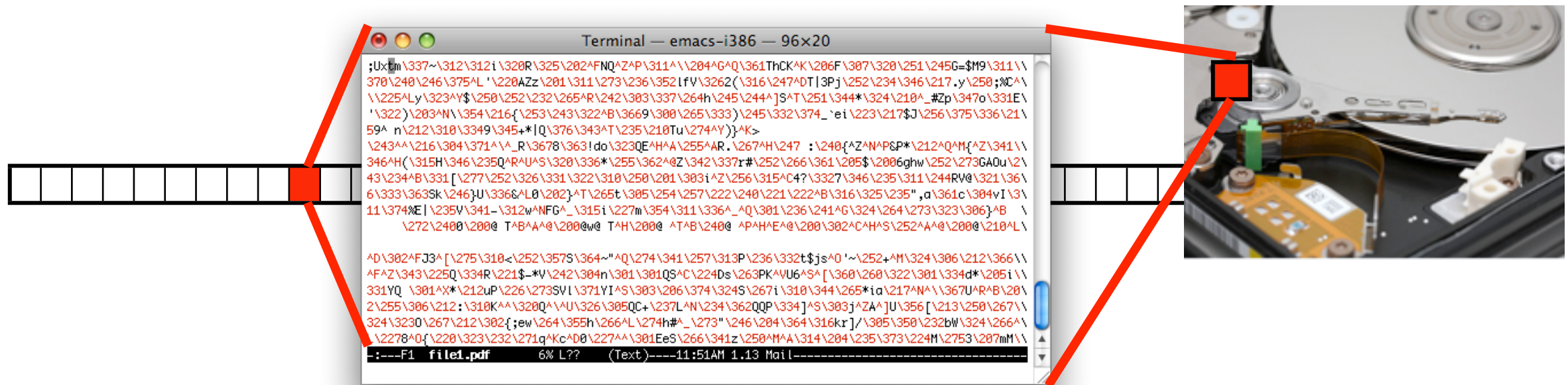
This is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some text

This is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some text

This is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some text

This is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is  
some textThis is some textThis is some textThis is some textThis is some textThis is some textThis is some  
textThis is some textThis is some textThis is some textThis is some text

# Some sectors are characteristically PDF data, others are just JPEGs or compressed text.



Most files on the hard drive are not fragmented.  
JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

Therefore, we can use a sector's *context* to assist in identification:





# Most files on the hard drive are not fragmented.

## JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

Therefore, we can use a sector's *context* to assist in identification:

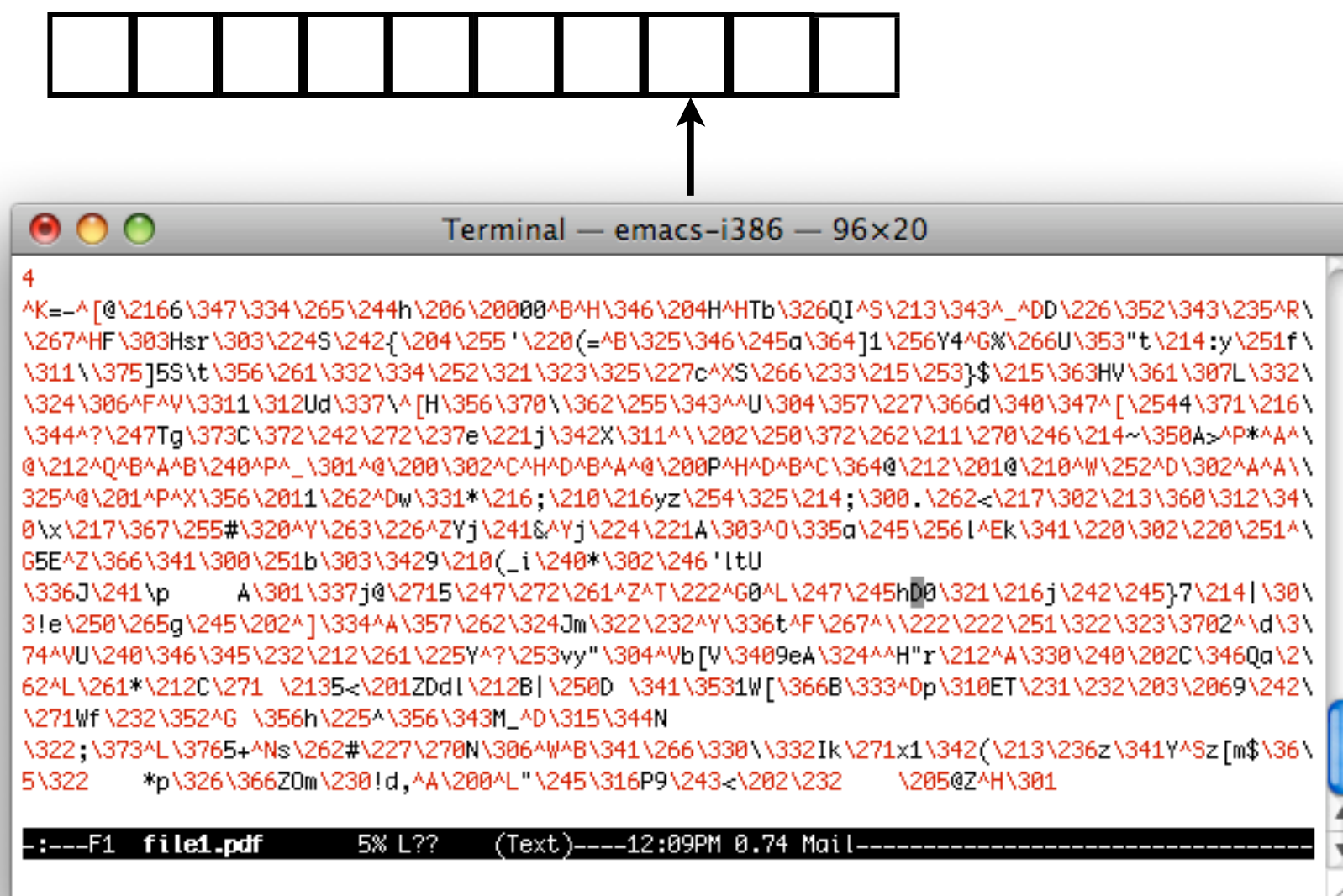


A screenshot of a terminal window titled "Terminal — emacs-i386 — 96x20". The terminal displays a hex dump of a file named "file1.pdf". The dump shows a series of hexadecimal values, with some lines wrapped. The status bar at the bottom indicates "6% L?? (Text)-----12:09PM 0.74 Mail-----".

# Most files on the hard drive are not fragmented. JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

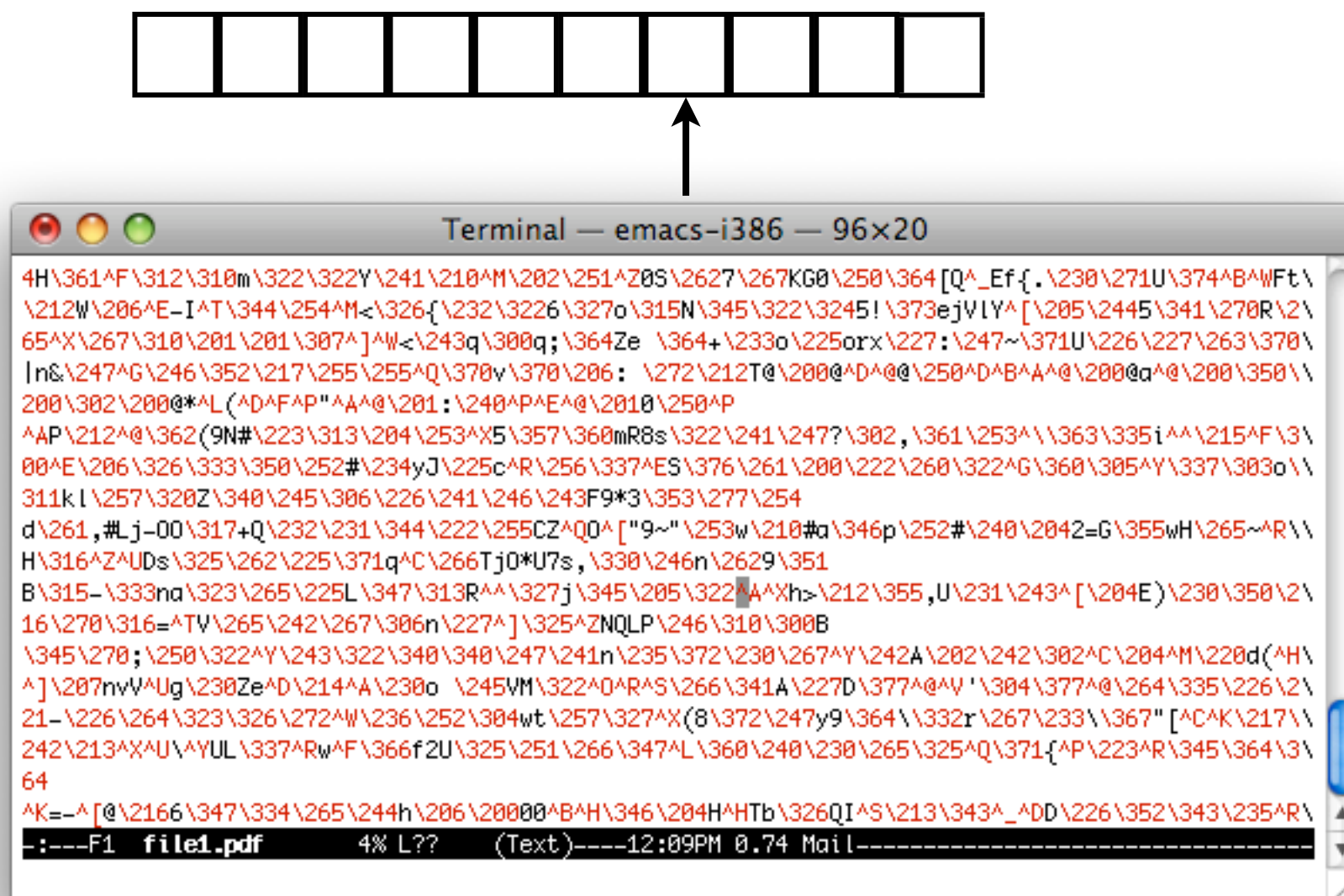
Therefore, we can use a sector's *context* to assist in identification:



# Most files on the hard drive are not fragmented. JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

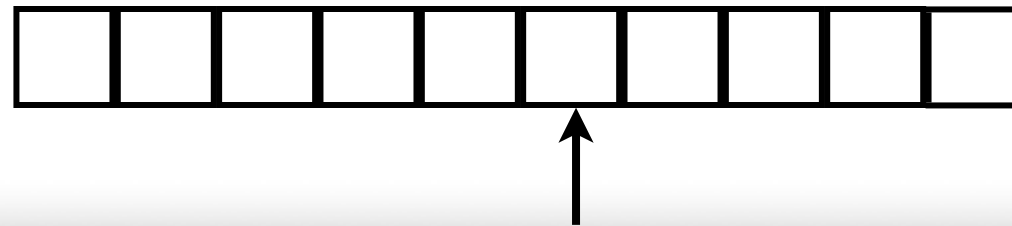
Therefore, we can use a sector's *context* to assist in identification:



Most files on the hard drive are not fragmented.  
JPEGS in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

Therefore, we can use a sector's *context* to assist in identification:



```

0\301\364\213\315,|\362\360\252>\271\341
a^E\242!\216\201^F\352^E\302^C^H^D^F^P^X@^a^B\200\200\353\356@ ^P&^P)^^H^L 0\200\302^A@^a^@ \200\35\
2\250D^F^P^H^L(^Q^a@ \250^P^X@(^QP^a@ \240^U^H\203^V\373?\207^K\260\245X\346\242~"t\207\336\242\27\
1F\270\327]\216w^a@ \253^Q\3506\330DP\264^A\205\232\261\242\321\262\252\202\267&^G\206\363\302#\2\
04\270\335\252h$~\254\350^ \202Jzs^0\277\317%Y-|\243Q\345\222\251\270\352xz\266y\244\303\301\35\
7\225(\356\341\335\200\234\362R*
\3710\334^D\331" \265$Y$ \236ef5\245\326\214\234^NJ\262\270\314F\320\264\210\234\362\342\212^CQ^L\
\250\231\220F\342\203^B\256\350\371d- \213 ' \334\211\263b\212Y7\221\337^EUe\2605\243\314UD\221\3\
07^[\216^G5^B\315L^[\314\343\342\212\251$R4\345\216D$ur\303\215Y!EjR\^[6^C\216\350\253\204gpvD0\
\216\350\240^ \216^Z\351p\233Uj\211K\201\302\315\253":y\216\254^i)b\331^@\214\205Q^_TXy \204^Q^S\
207(\251\343vZ\265^Y\247^]\302\250\245W^P^sH!e\246^D\241\324\263\357\366
\273e;\2441\351\221\212\216\212\321X&\215\2479*\245l \215\306Q^V0^X\304\310F\240\340p\271\3677\2\
47\234\361\357^T\307\340\331X<6\357\346Z\236R\370y\3650^S\323\303\221^P^Y\375\326\2523g\342\25\
2\207g\302\211\345M\303T\332j\313\325\311\332i\240 ^_\274F\311\277\241\322\332x:\345W\207\326\3\
16\354s-^C^A^E\313\217^B\304\326d\371\210\356\210\253E\303\202<\200^@#\226\313:jV\375\263\207\2\
34H\361^F\312\310m\322\322Y\241\210^M\202\251^Z0S\2627\267KG0\250\364[Q^_Ef{.\230\271U\374^B^WF\
t\212W\206^E-I^T\344\254^M<\326{\232\3226\327o\315N\345\322\3245!\373ejVIY^[\205\2445\341\270R\
265^X\267\310\201\201\307^]^W<\243q\300q;\364Ze \364+\233o\225orx\227:\247~\371U\226\227\263\37\
-:---F1 file1.pdf 4% L?? (Text)-----12:09PM 0.74 Mail-----

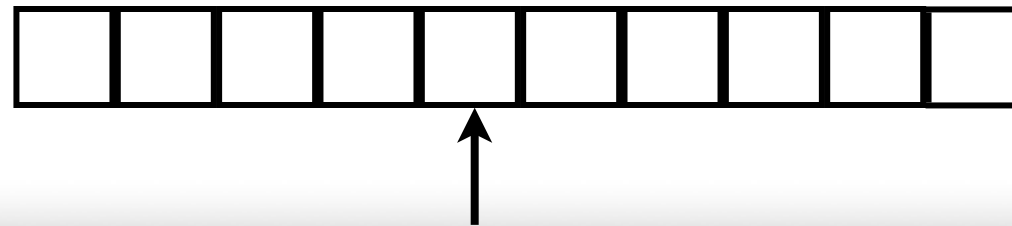
```



Most files on the hard drive are not fragmented.  
JPEGS in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

Therefore, we can use a sector's *context* to assist in identification:



```

2*\362\300 \215\336gs\302\251V8r\207 H\361\223\352\245^]dm^@^DS\362^@9\371\240\257\343F\374\234\
\251\245D*i\333\237^Q\2439M.\311\343\307!\323^Y
kBX\232s\272D\251\334C^[\222\2643*gt\217\322\322\245\244\211 \203^C.\346\262\253qG\250\214rT\dm\
crp\265\2445\362\376^P\202<\271\334\320/!\271\300A^L\2251\263\233\221^P\272\2761\260\346\201\24\
2\340:^D
\332\370\372\204^S2\246 '\375\344^R\202^]\366H(^P\201\215\302)\216 '\364PF\340G$%#c^N\346^RE\267G\
\210^@\334+\244\336\312|\243t^L89!@^316E^B\214^T_F\221\202\240{^^z\255DI\315TW\250\211\257i^D^M\
\324V^]\201\345\314\344\222\232> \3657c\376\212\246\232\226\332\322]\241\307q\352\2527\
\241~\266\242$\352\203I\3342\323^@\335\332\210\350\262\257)\366\227^]\333\207\264\311J^L\221\27\
4\340\202y+G\234>\341\304^U\204\371\304 '\366\311*m\255^Z\333=\322\253y\352&>\200\341<\232Y\213\
205G9\344\317\367\235\225^Mk\312A\303\364Q\2373\331\360
\351^T\256\255\267\321\3239\2619\200\235\263\325Q\303\272\333%eItm:
\315\246\235^E\263\206\271^W\2056\275\256\256\337gdA\241\261\222}^B\203z\232\3215C]^_ \206^Z\323\
\260Z\220\333\323\270^G\202\251\255\261 \237^X2;rp\272c\213\235\257A\212^F\260^@\321\214.\223^A\
V-N^[\335kI\262\206^M\32164^D^M,@\307G\224\322\312\253SH\311Zr\321\362\356-L\234\265\316\213\3\
50\362^R\321\345\312\347f\235%\333^N\362\355^V\3710\242\212\371k\214d2\337j]\234\341\330V""\227^A\
0\301\364\213\315,|\362\360\252>\271\341
a^E\242!\216\201^F\352^E\302^C^H^D^F^P^X@a^B\200\200\353\356@ ^P&^P)^H^L 0\200\302^A@a^@\200\35\
-:---F1 file1.pdf 3% L?? (Text)---12:09PM 0.74 Mail-----

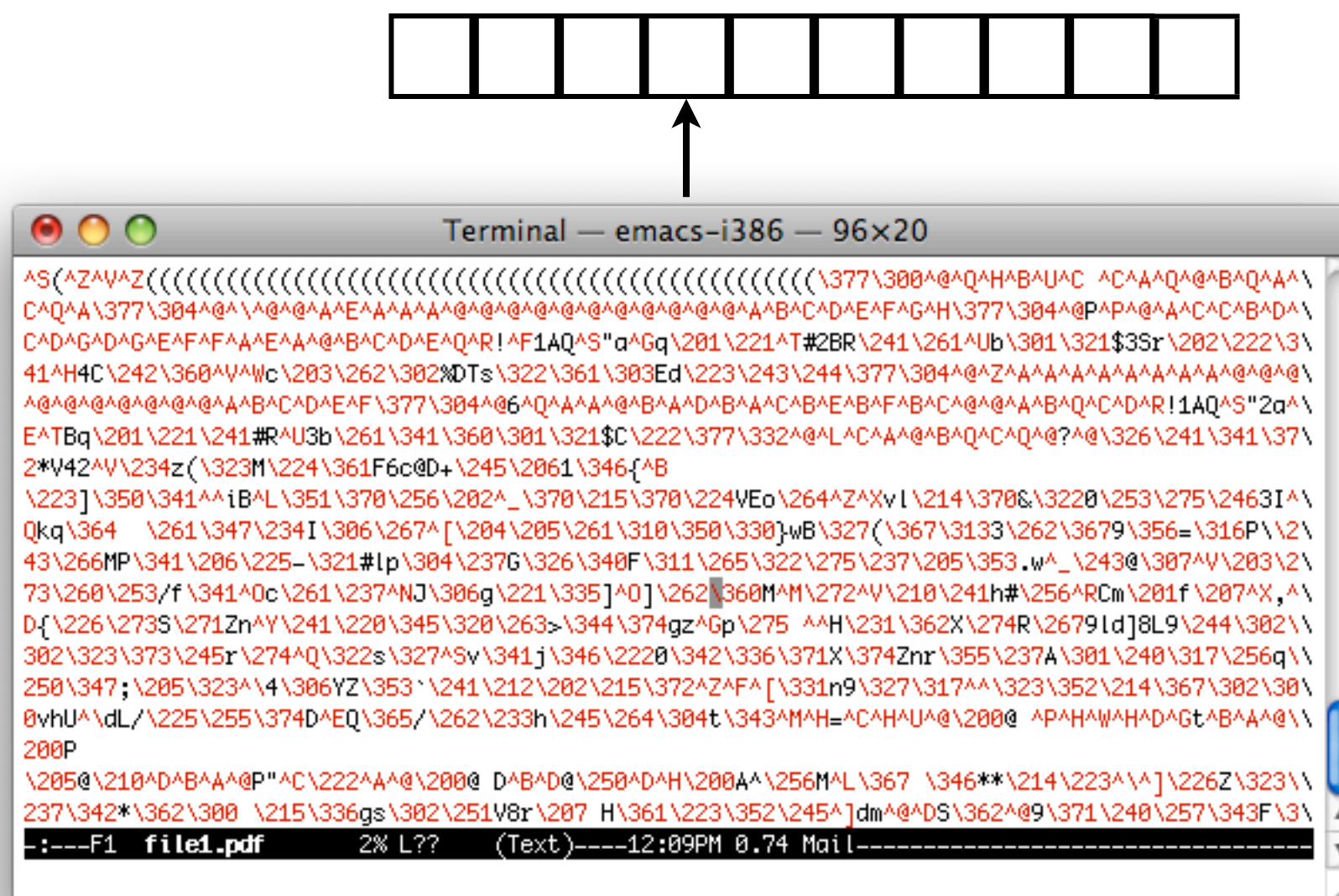
```



# Most files on the hard drive are not fragmented. JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

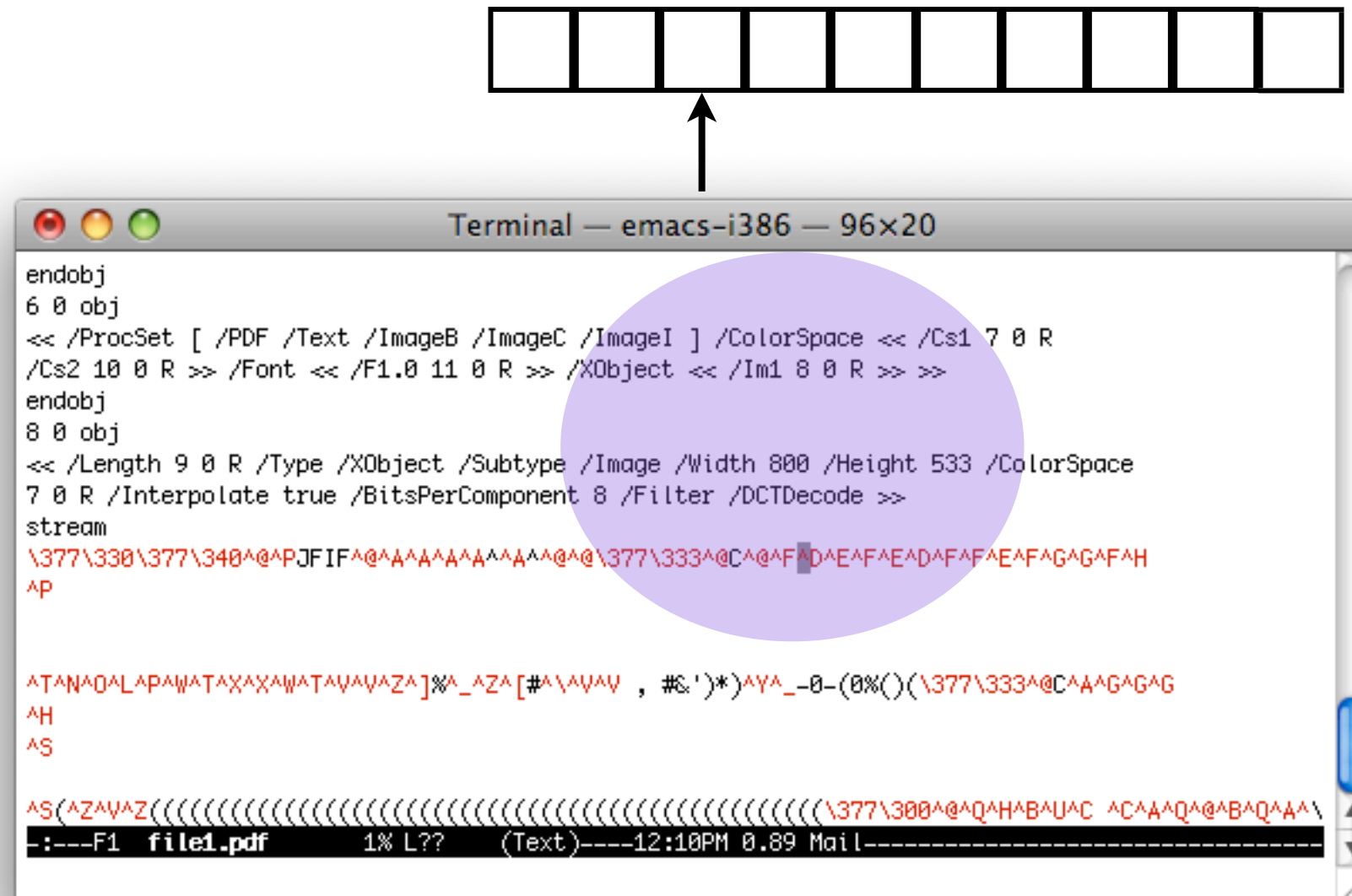
Therefore, we can use a sector's *context* to assist in identification:



# Most files on the hard drive are not fragmented. JPEGs in PDFs can be identified by scanning forward.

In previous research, we found that only 15% of forensically interesting files are fragmented [Garfinkel 2007].

Therefore, we can use a sector's *context* to assist in identification:



# Sectors can also be identified from statistical properties.

File type	Identified By
NULL	direct examination.
HTML	n-gram analysis
JPEG	High-entropy with markers
ZIP	High-entropy that's not encrypted
Encrypted	High-entropy that passes encryption tests

# Using sector identification, we can identify the *content* of a hard drive within 10 seconds (after it spins up).

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*



# Work to date:

## Publications:

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

## Work in progress:

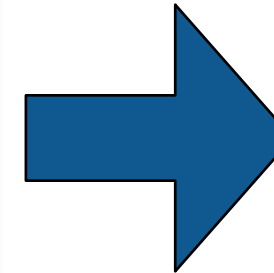
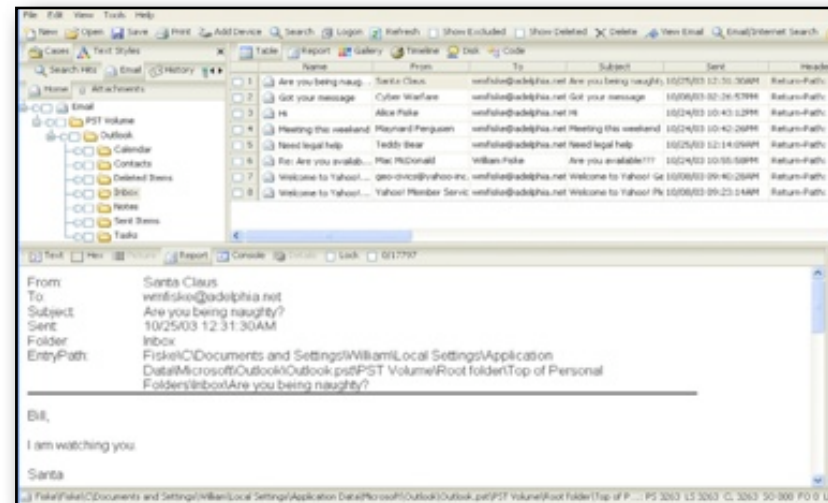
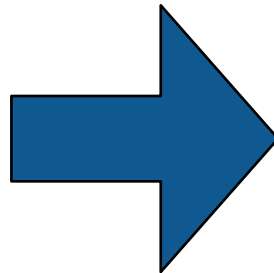
- Alex Nelson (PhD Candidate, UCSC) summer project
- Using “Hamming,” our 1100-core cluster for novel SD algorithms.
- Roussev’s Similarity Metric





# 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*.

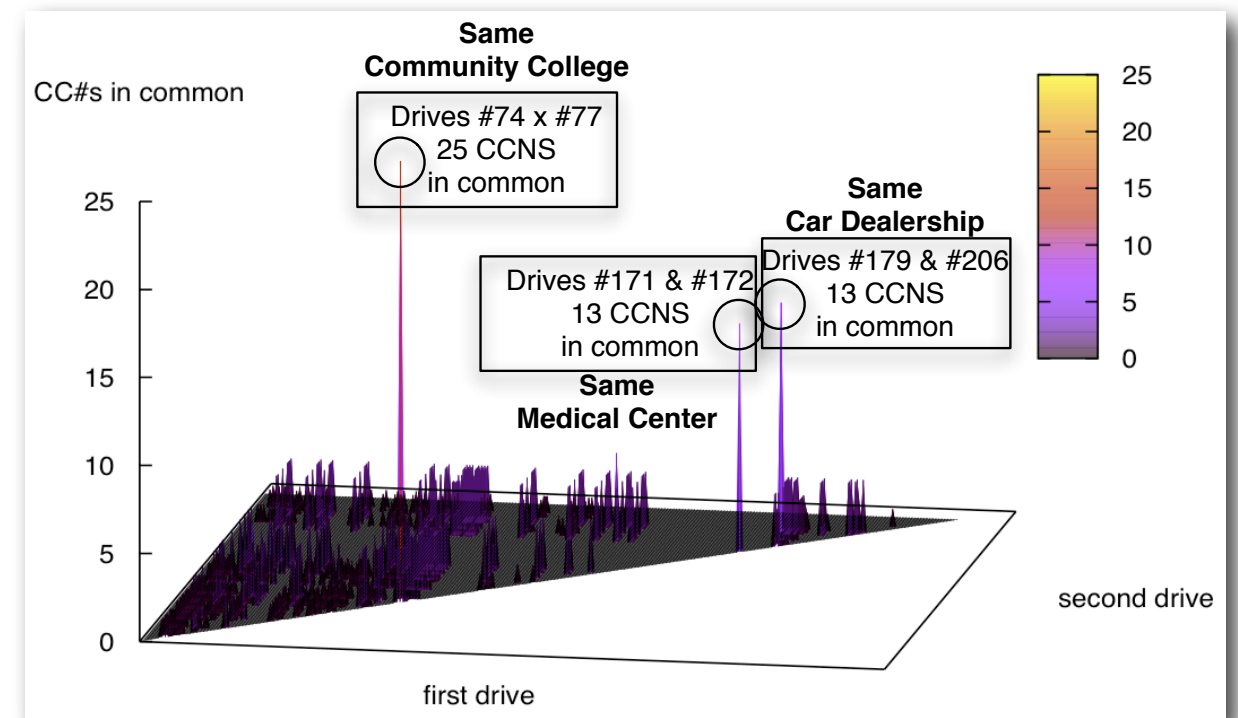
## Evidence Reconstruction

- 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.*





# Consider file fragment identification: You can't compare the work; the data are all different.

Since 2001 more than a dozen papers have been published.

- McDaniel et. al reported 43.83% accuracy on JPEGs
- Moody & Erbacher report 72% accuracy.
- Karresand and Shahmehri: 97.90% true positive rate and 99.99% true negative rate.
- Calhoun and Coles: 83% to 99% accuracy

But everybody used a different data set!

- Most did not release their code, either.
- If you try to re-implement the algorithm, how do you know you got it right?

Problems in working with “wild data:”

- *You don't know ground truth*
- *Time spent collecting & preparing is time lost to research*





# 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



# There are only a few existing forensic corpora today.

## Forensic Challenges

- DFRWS 2005 — 2009
  - *Windows memory analysis*
  - *Linux memory analysis*
  - *File Carving*
- DoD Cyber Crime Center (DC3) Challenges
- Honeynet “Scan of the Day”
  - *Widely used, but questionable realism*

## NIST Computer Forensic Reference Data Sets (CFReDS)

- Small number of test images.
- Good for tool testing, but not necessarily for research or training.

# Corpora Sensitivity: How should we describe the data and protections?

## Test Data

- Constructed for the purpose of testing a specific feature.
- CFReDS “Russian Tea Room floppy disk image” to validate Unicode search & display.

## Sampled Data

- A subset of a large data source — e.g., sampled web pages or packets.
- Hard to randomly sample.

## Realistic Data

- Not “real” — made in a lab, not in the field.

## Real and Restricted Data

- Created by actual human beings during activities that were not performed for the purpose of creating forensic data.
- Controlled for privacy reasons.

## Real but Unrestricted

- Released for some reason. e.g. the Enron Email Dataset
- Photos on Flickr; User profiles on Facebook.

# Restrictions on Corpora Use

This is primarily an issue with federally funded research.

Experiments are exempt under 45 CFR 46:

- “if these sources are publicly available”
- “or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.”

What about re-identification research?

- Probably needs IRB approval in advance.



# We are making available three corpora.

## A real but unrestricted file corpus

- 1 million files

## Test and Realistic Disk Images

- 6 disk images

## The Real Data Corpus

- More data than you can shake a stick at. Really!
  - *Half is in Cambridge MA*
  - *Half is in Monterey, CA*



# NPS-govdocs1: 1 Million files available *now*

## 1 million documents downloaded from US Government web servers

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

## Free to use; Free to redistribute

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

## Distribution format: ZIP files

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

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



# We have created six disk images.

## Test Images:

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

## Realistic Images:

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

## Each image has:

- Narrative of how the image was created and expected uses.
- Image file in RAW/SPLITRAW, AFF and E01 formats
- SHA1 of raw image
- “Ground truth” report

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

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

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

Real data provides:

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

We have multiple corpora:

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



# Real Data Corpus: Current Status

Corpus	HDs	Flash	CDs	GB
US*	1258			2939
BA	7			38
CA	46	1		420
CN	26	568	98	999
DE	37	1		765
GR	10			6
IL	152	4		964
IN		66		29
MX	156			571
NZ	1			4
TH	1	3		13

\* Not available to USG

1694

643

98

6748

***Note: IRB Approval is Mandatory!***



# Work to date:

## Publications:

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

## Websites:

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

## Work in progress:

- Joshua Gross, NPS postdoc, 2009-2010