Automated Digital Forensics and Media Exploitation

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

November 11, 2009

http://simson.net/

NPS is the Navy's Research University.

Location: Monterey, CA

Campus Size: 627 acres

Students: 1500

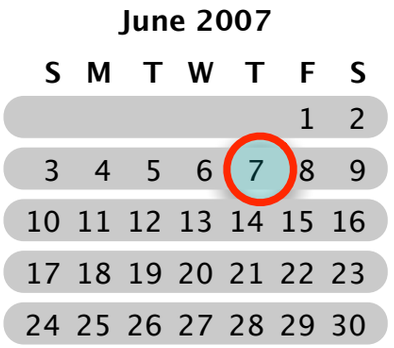
- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)

Schools:

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

Automated Document and Media Exploitation: The Need

Law enforcement & military encounter substantial amounts of electronic media.



- Battlefield
- Checkpoints & Border crossings
- Law enforcement operations
- Internal Investigations

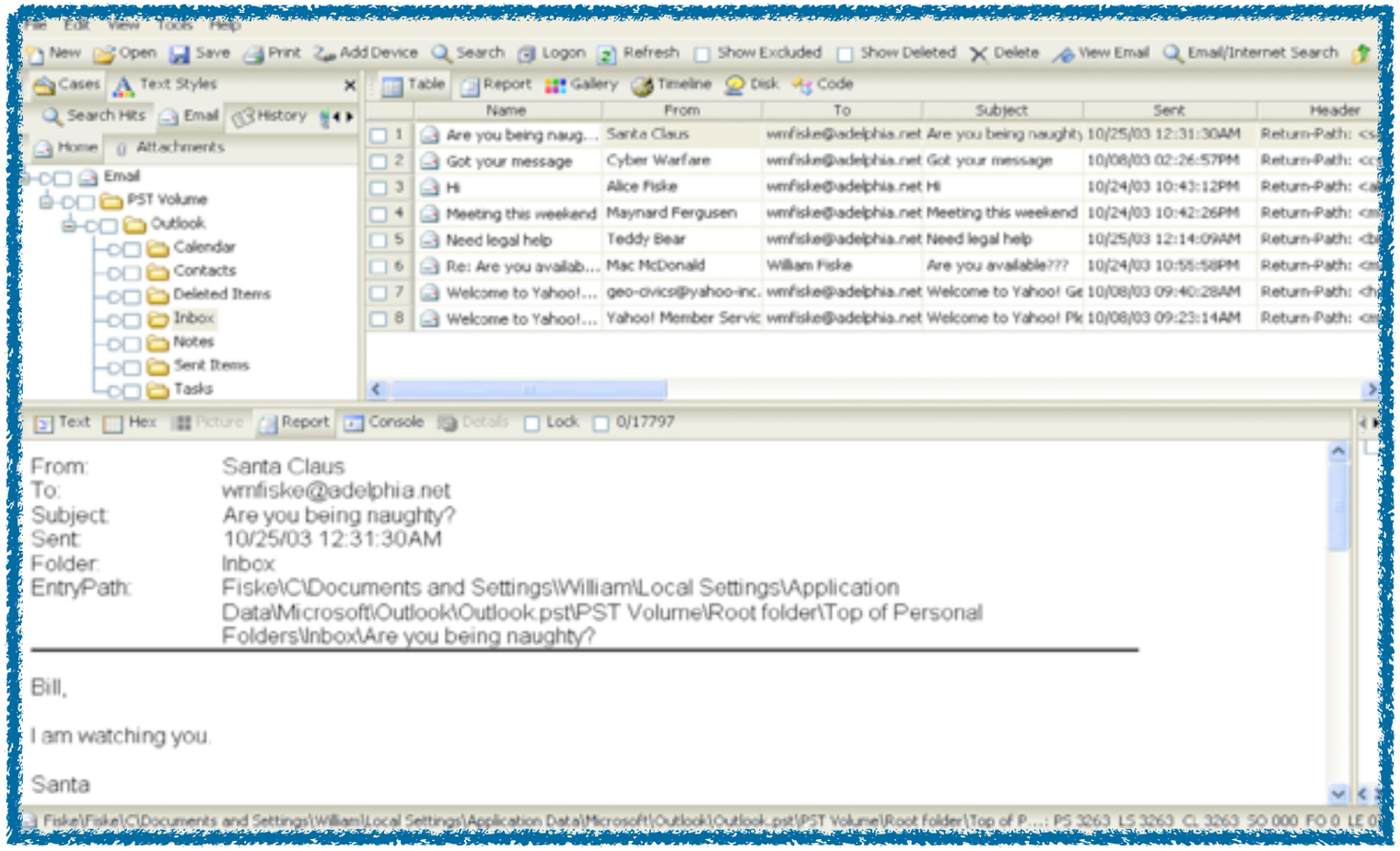
FBI RCFL FY08 Annual Report:

- Examinations: 4,524 ↑
- Cell Phones: 2,226 ↑
- Hard Drives Processed: 17,511 ↑
- Total Data Processed: 1,756 TB ↑

Most media is analyzed using highly trained personnel...

DOMEX in Iraq

… working with tools designed for law enforcement.

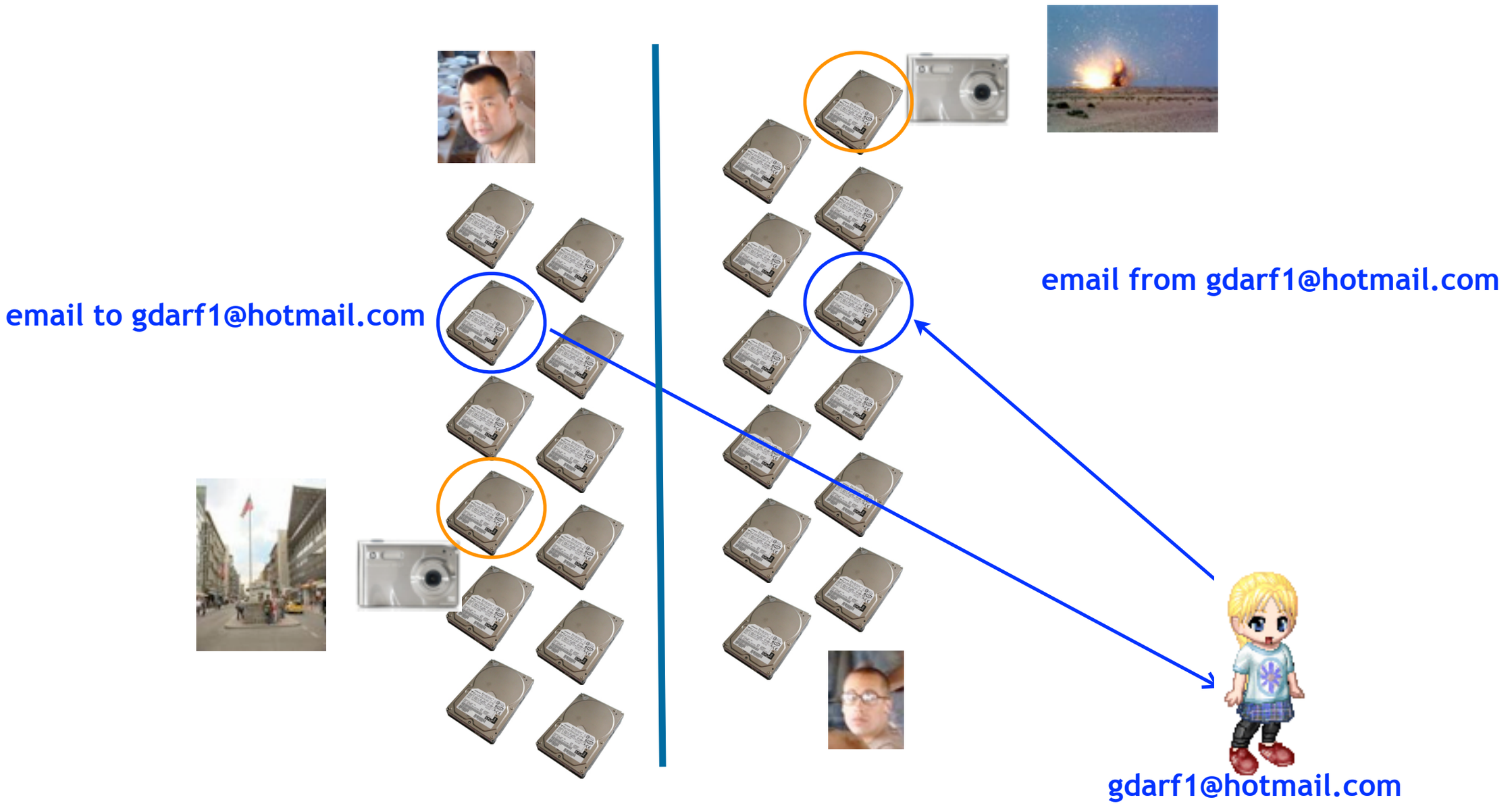


EnCase by Guidance Software

- Designed for visibility & search, not analysis.
- Does not scale to 100s or 1000s of drives.
- Prevent “contamination” between cases

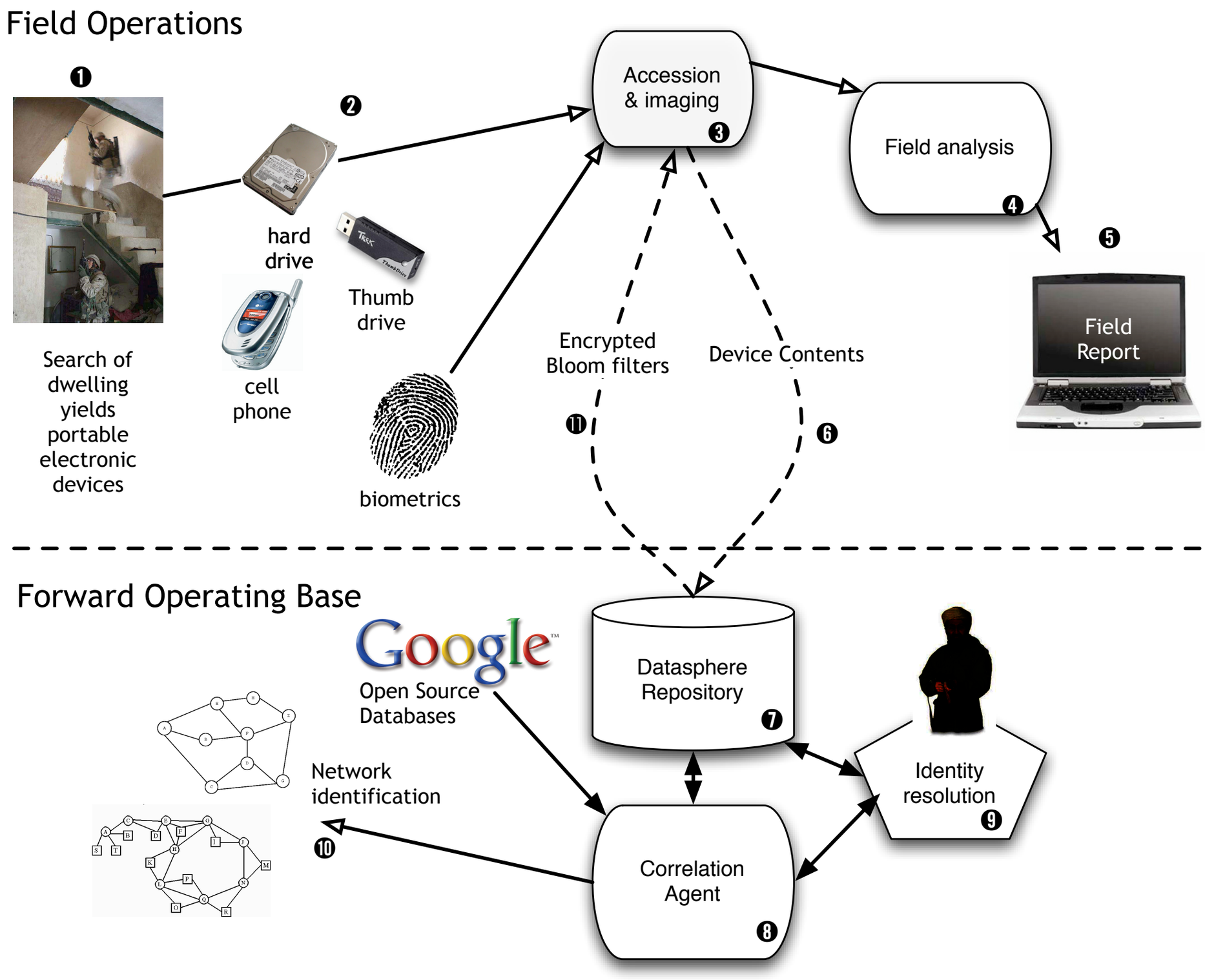
Manual analysis misses opportunities for correlation.

Different analysts see different hard drives.



Keyword searches don’t connect the dots.

ADOMEX would speed workflow while providing for automated “situational awareness.”



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

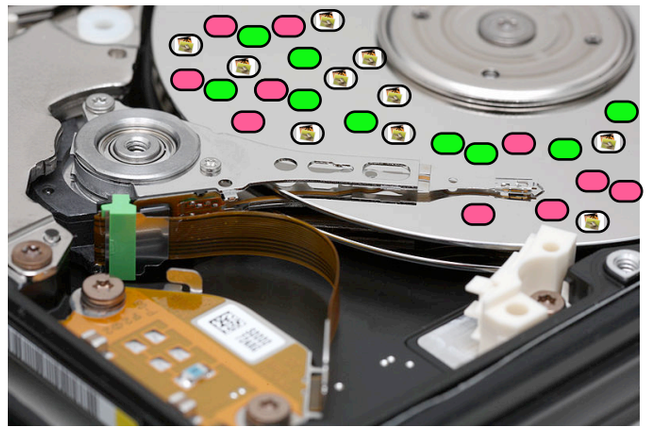
- Stream-processing
- Statistical techniques (sub-linear algorithms)

Area #4: Creating Standardized Forensic Corpora

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

This talk focuses on two research projects:

Instant Drive Analysis



Available Research Corpora

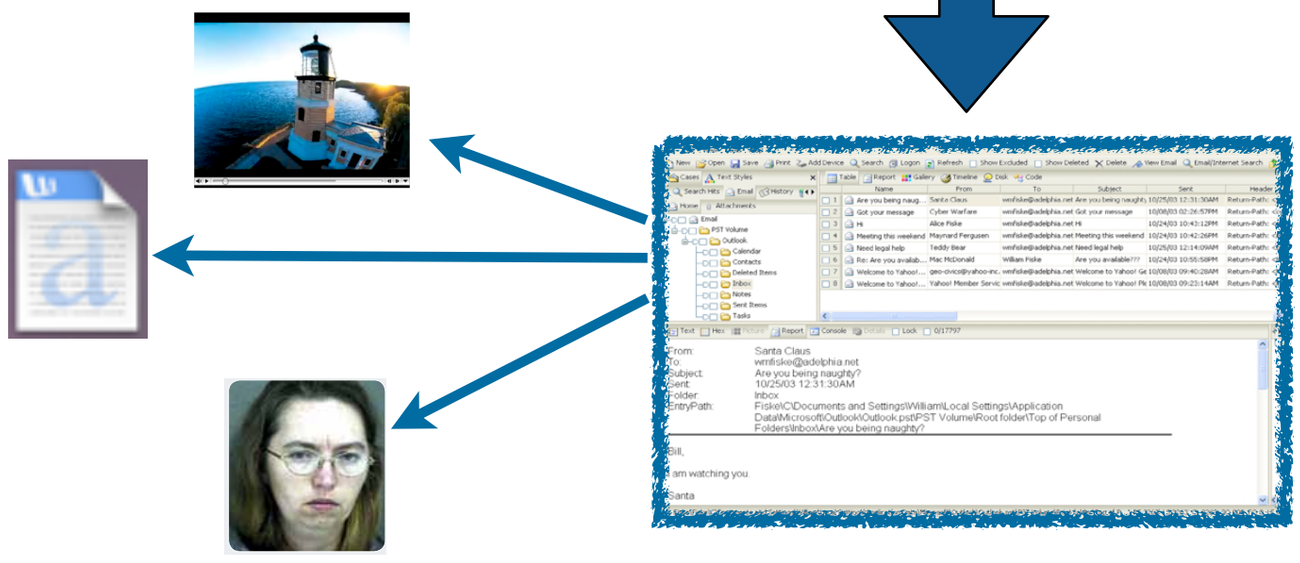
Today's forensic tools are designed to extract files.

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, the analyst makes sense of the data.



Data on hard drives can be divided into three categories:

Allocated FIles

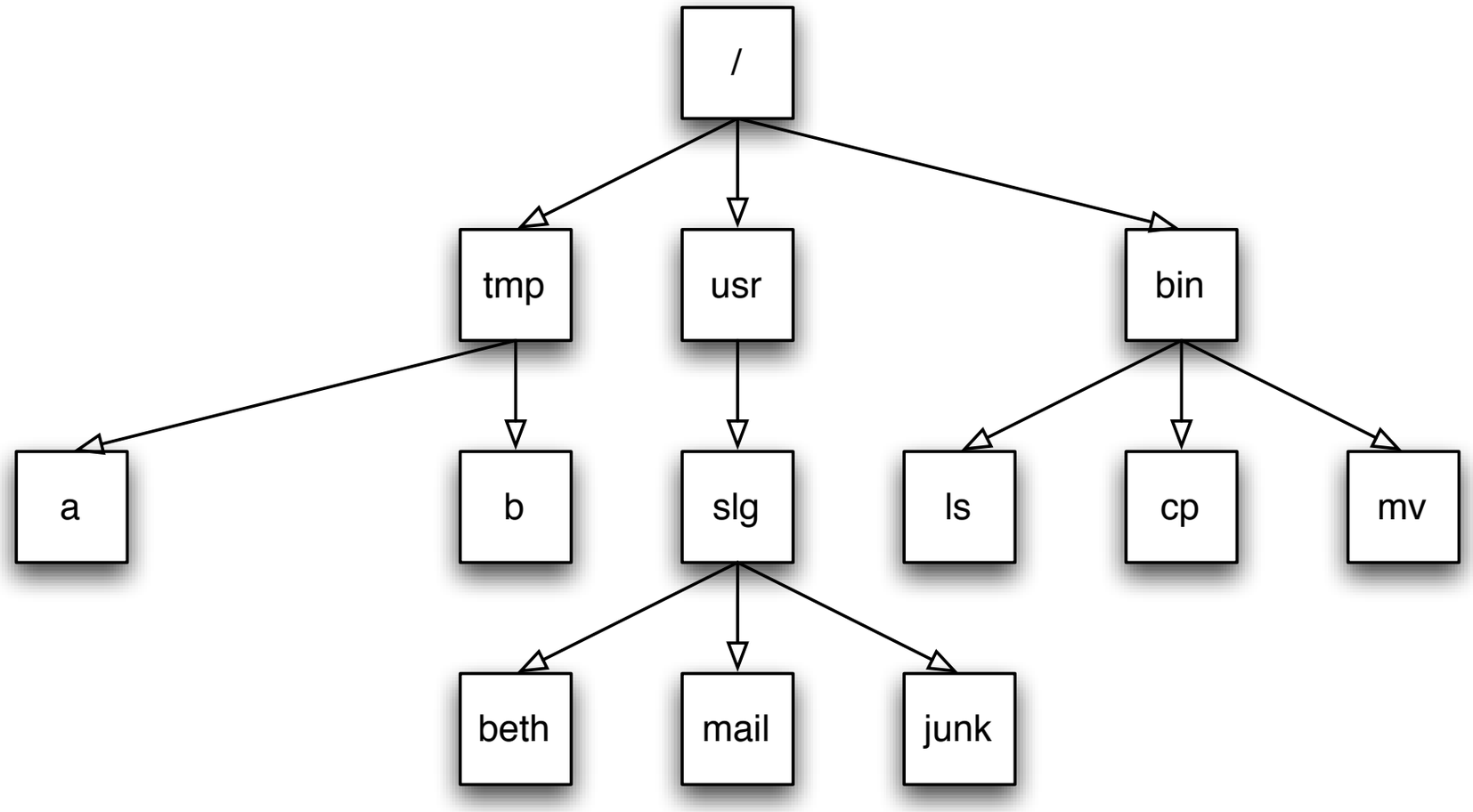
Deleted Data

No Data

user files
email messages
[temporary files]

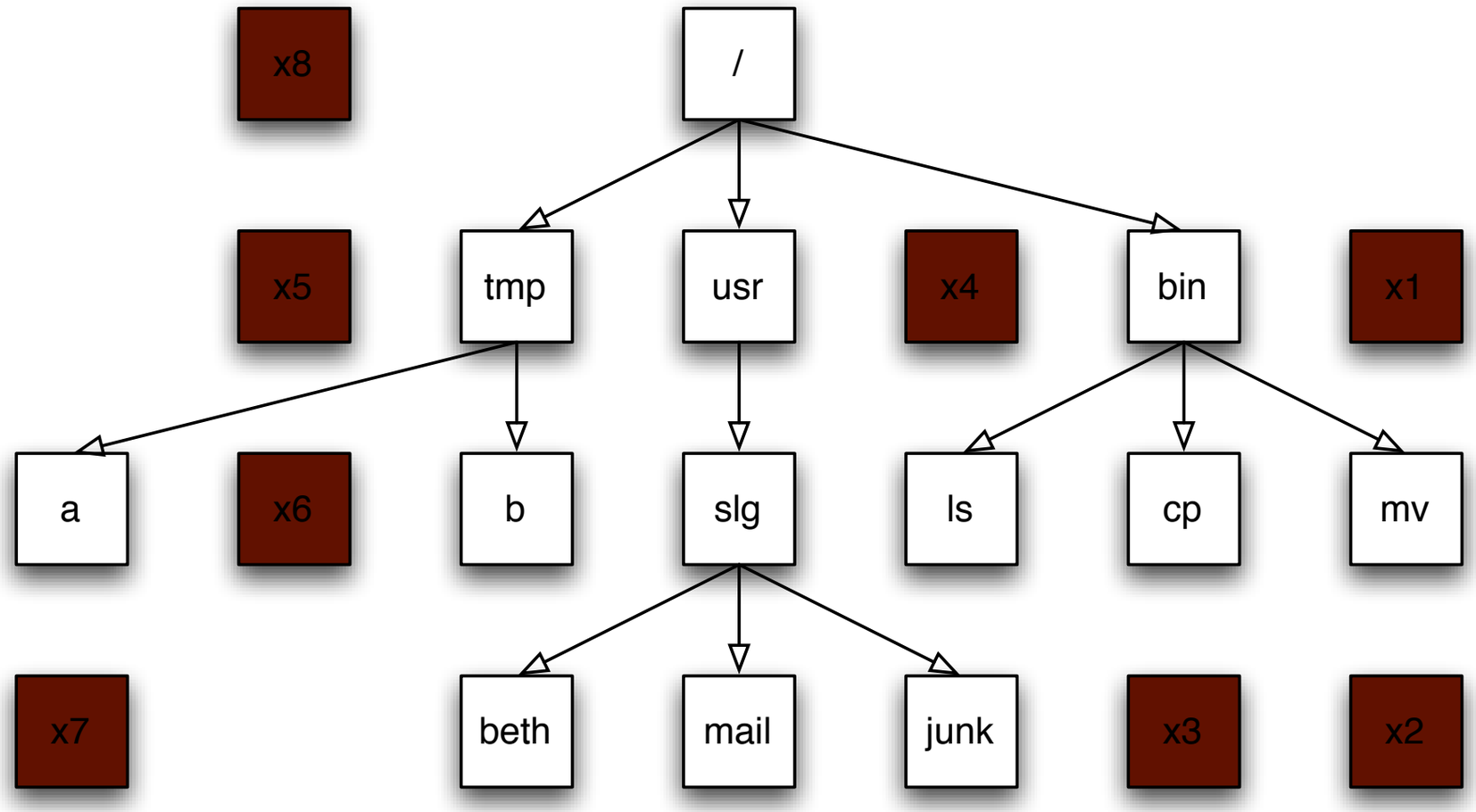
blank sectors

Allocated files are the visible files.



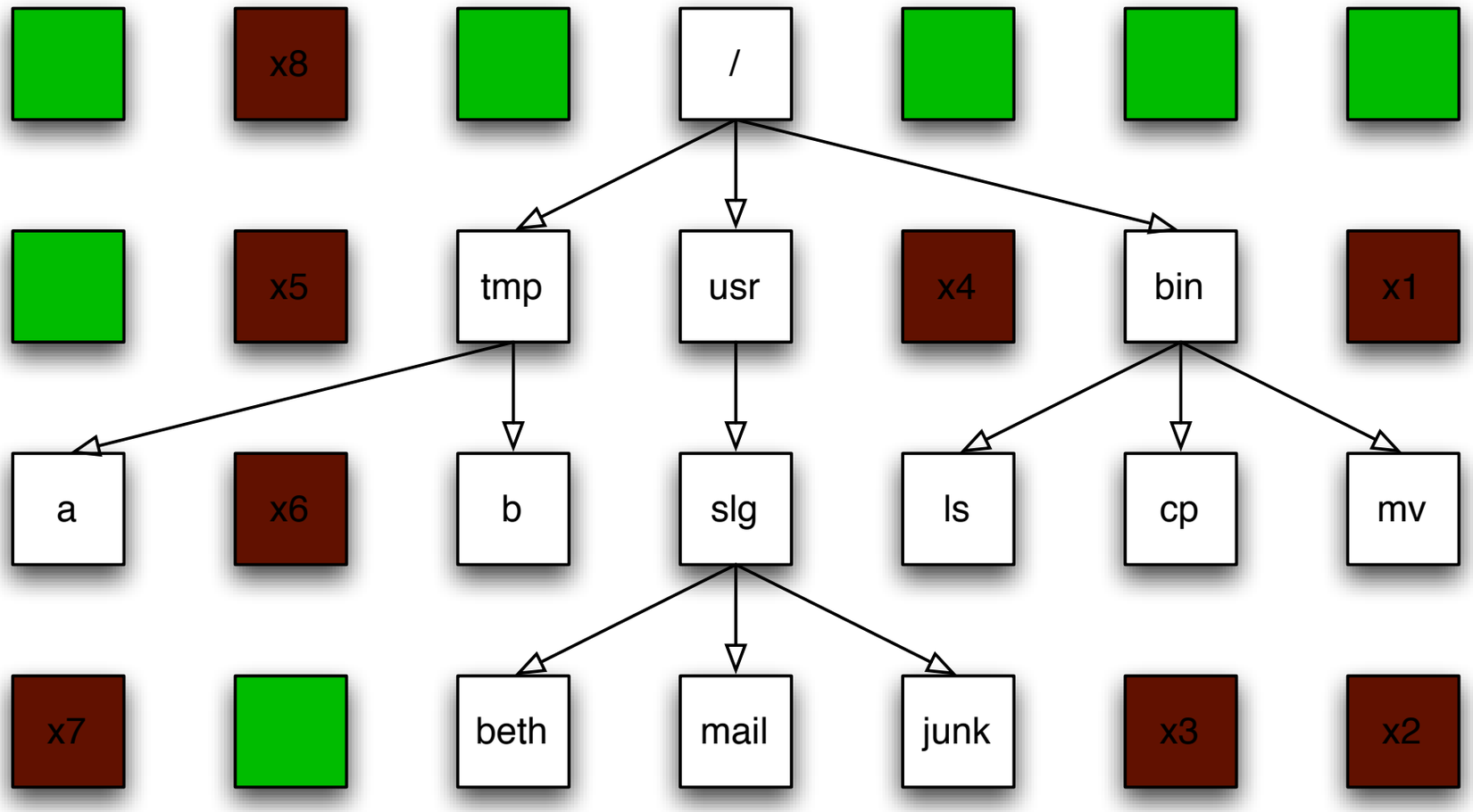
Allocated Files

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



Deleted Data

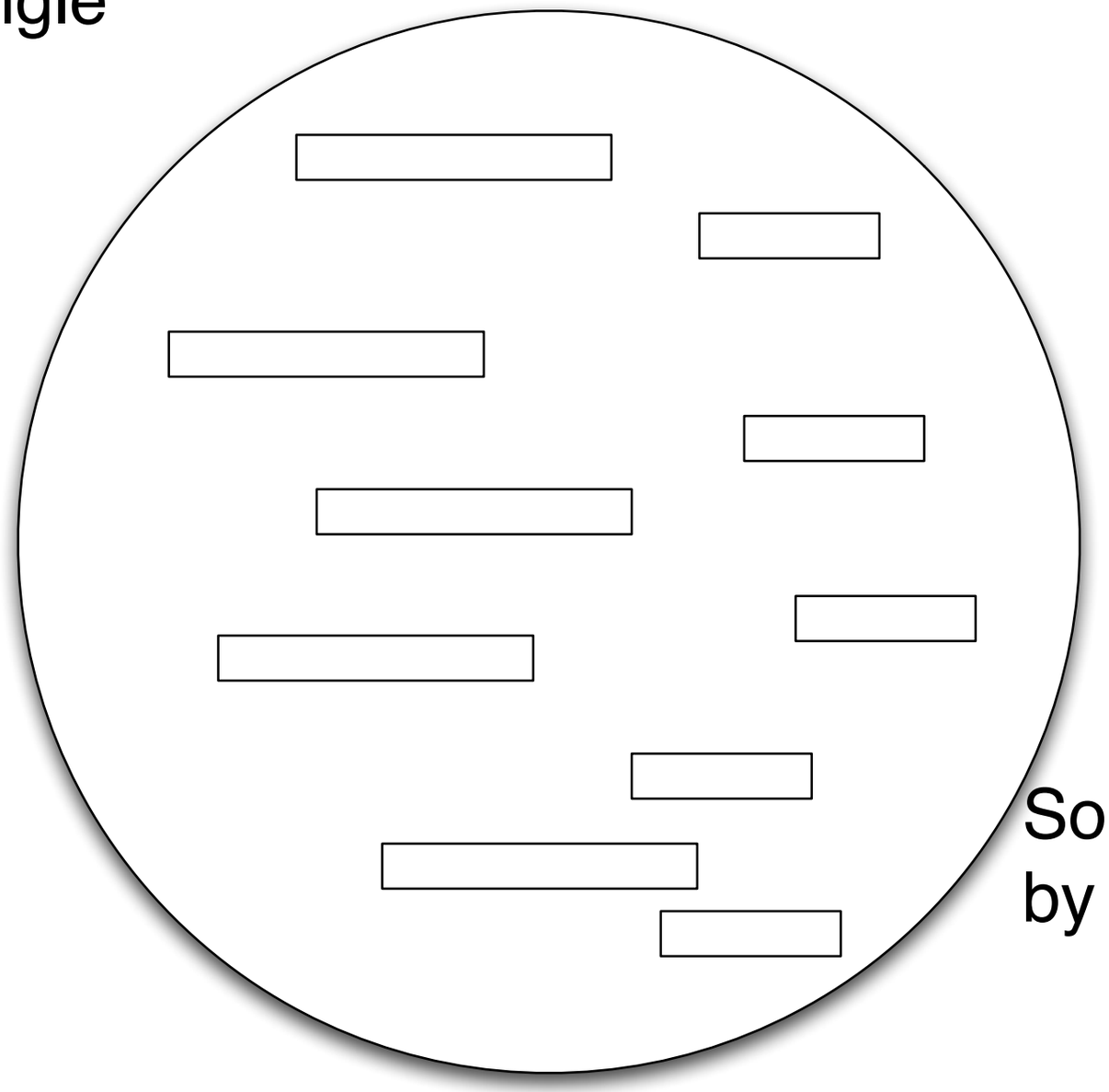
Some sectors are blank. They have “no data.”



No Data

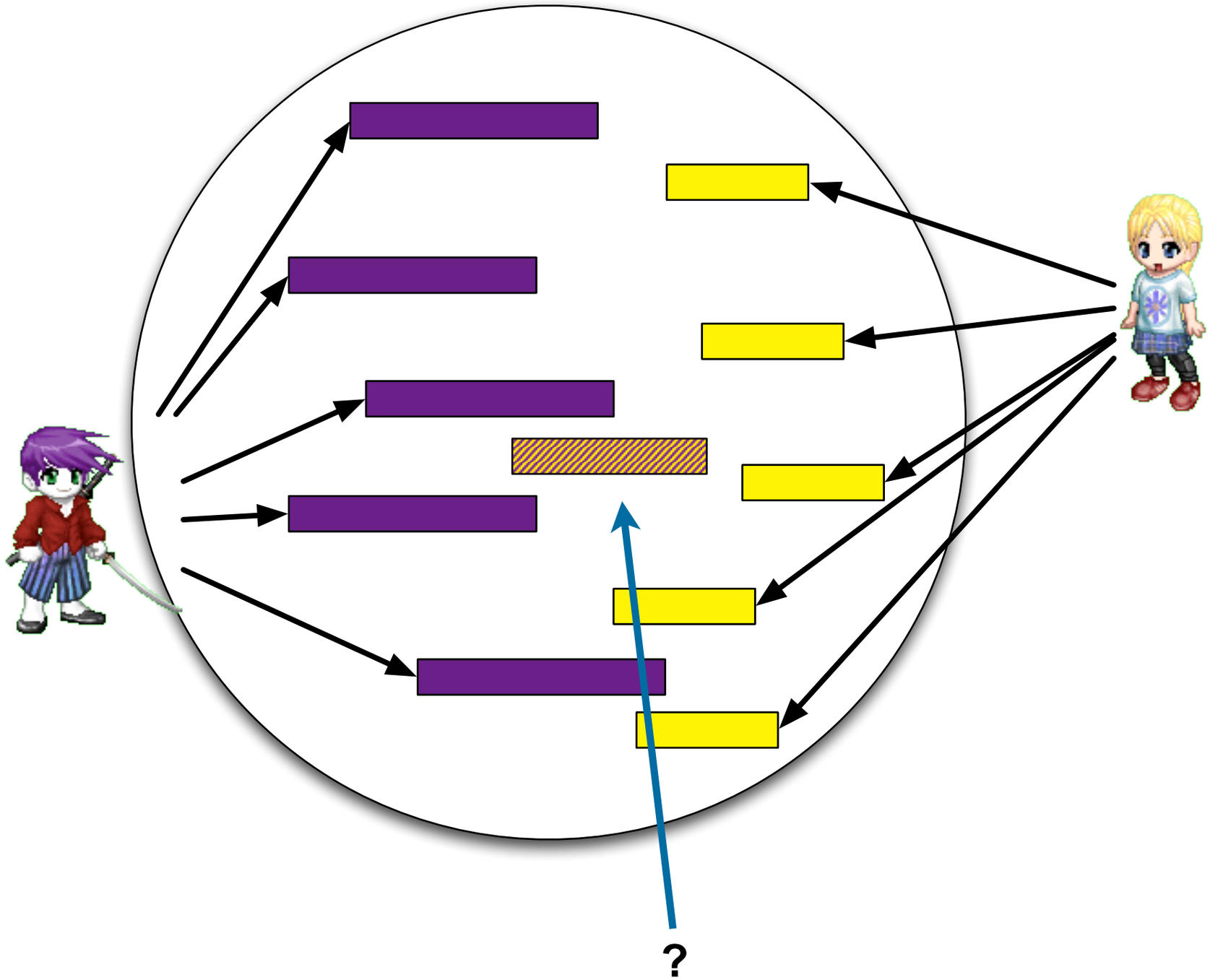
There is no "typical" hard disk.
Today's disks can have 0 — 10,000,000 files.

Some disks are used by a single person.



Some are used by multiple people.

Some data can be readily identified with one user. Other data can't.



Prior work used *content analysis* to determine authorship

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

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

File system metadata:

- Timestamps for “orphan” files.
- Fragmentation patterns (disk usage)

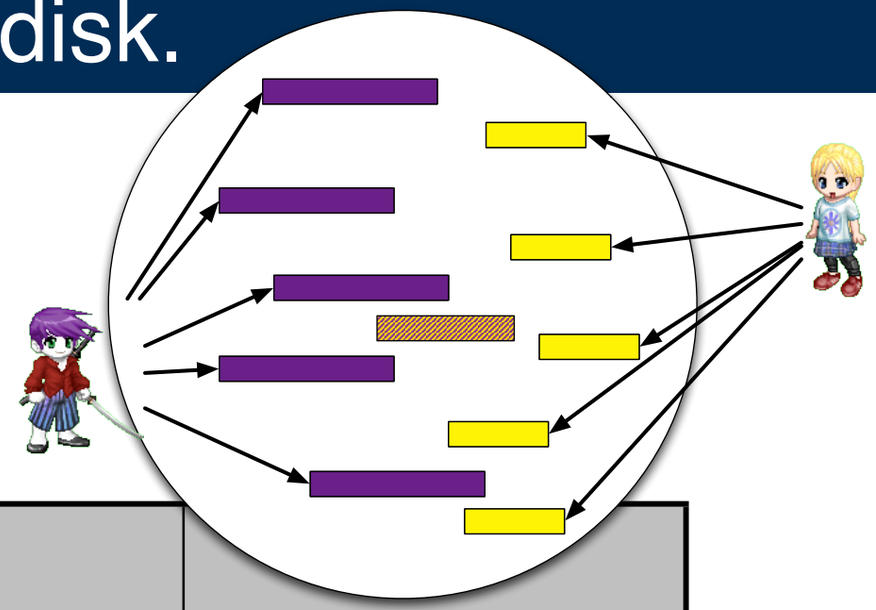
File placement information

- Where the file is on the hard drive (sector numbers)

Embedded File metadata:

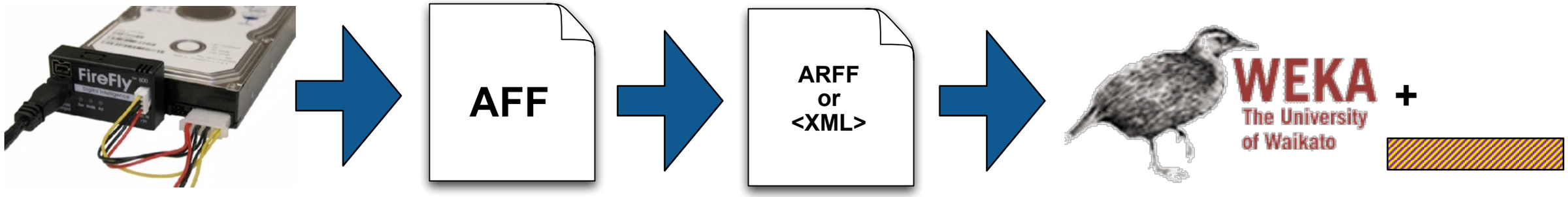
- Embedded timestamps
  - — *Creation Time*
  - — *Print Time*
- Make & model of digital cameras
- Usage patterns.

One approach to identifying the "owner" is to find commonalities with other files on the disk.



Magenta	Yellow		Likely User
100 JPEGs 5 DOCs	75 XLS 400 HTML	JPEG	
Print time: 9am & 10am	Print time: 5pm & 6pm	Print time: 5:30pm	
Location: 100 & 200	Location: 23,000 & 25,000	Location: 24,500	

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 known files as exemplars

Step 3: Use classifier to ascribe unknown files.

fiwalk converts disk images to XML or ARFF files:

Tags on a per-image and per-volume basis...

And per <fileobject> tags:

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

fiwalk also has a pluggable metadata extraction system.

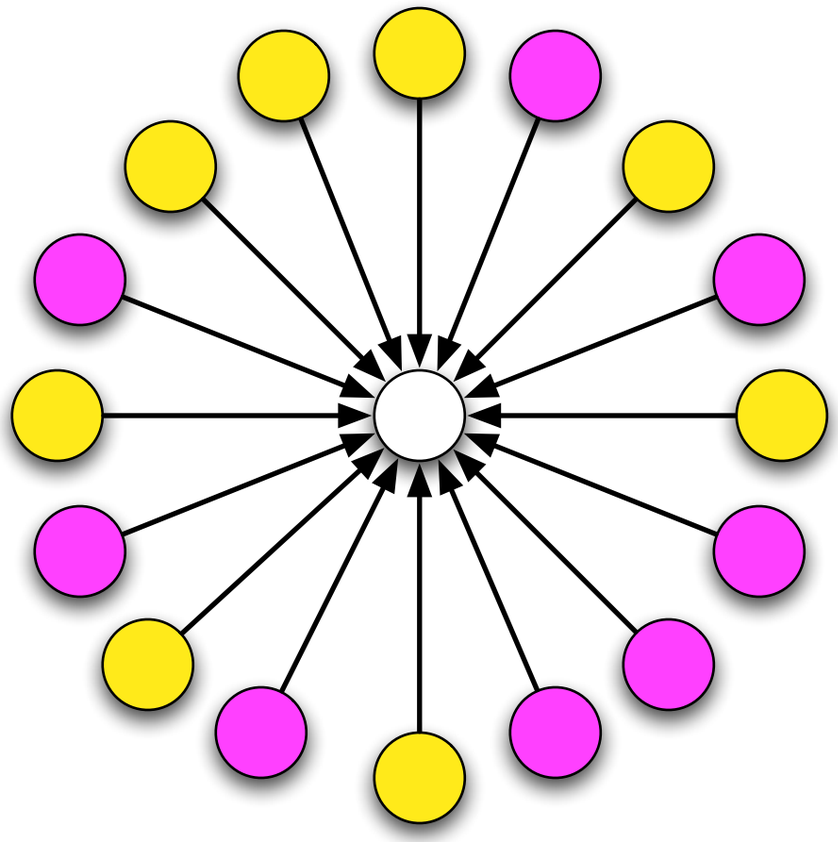
Metadata extractors are specified in a configuration file:

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

Metadata extractors produce name:value pairs which are incorporated into output:

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

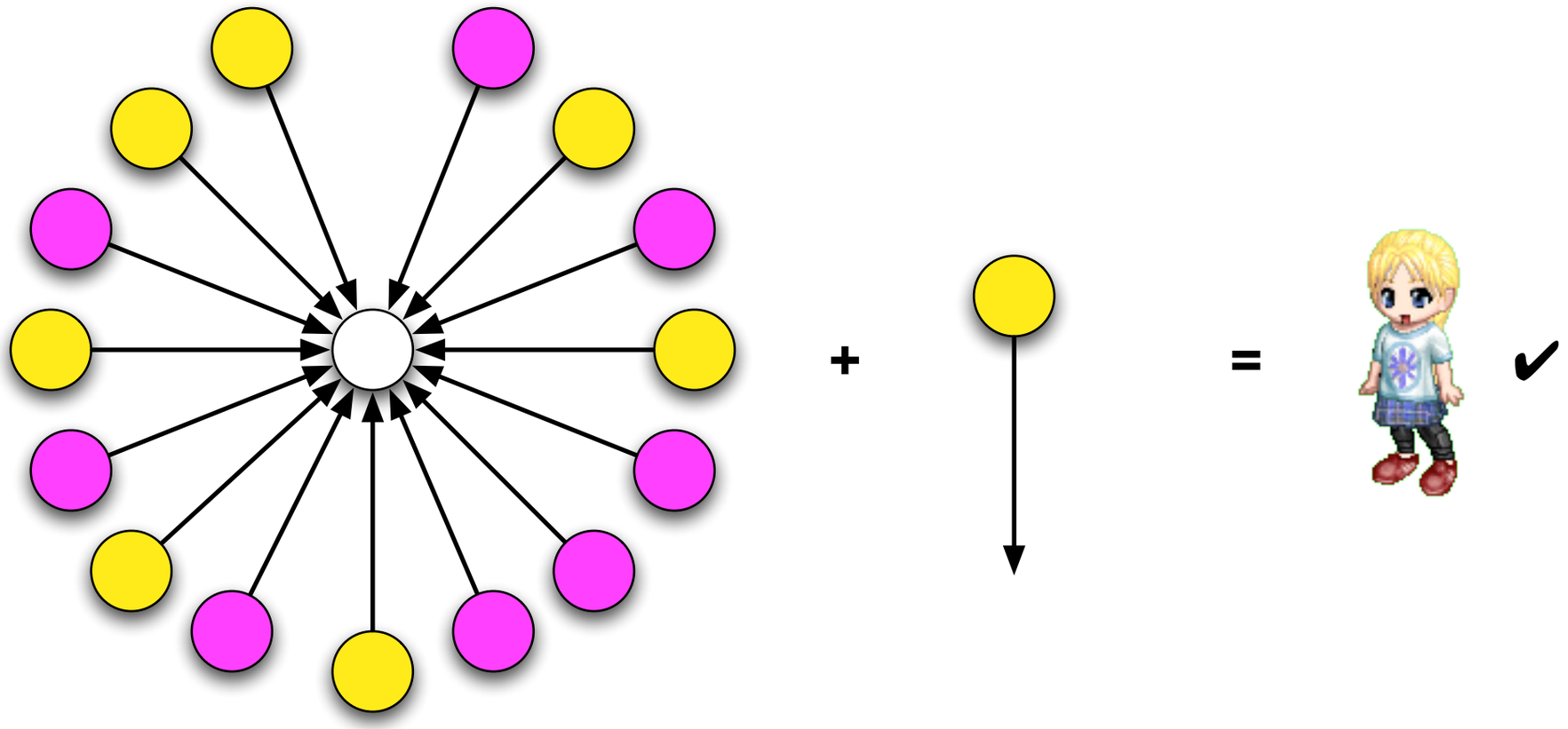
The classifier is built from *all* of the exemplars.



+

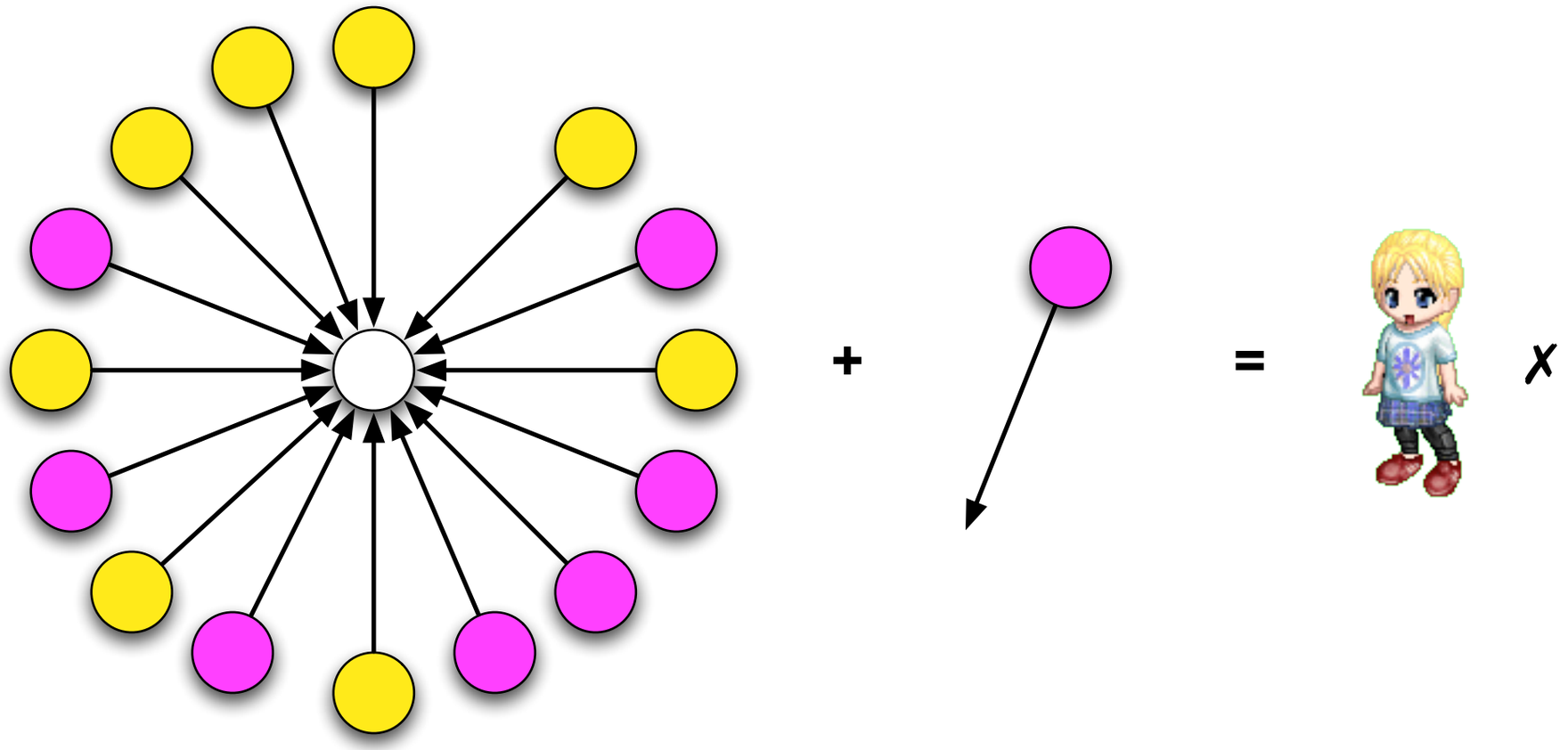
=

To validate the classifier, we take one element out and use the classifier to classify that element...



Here the classifier got it right!

Here the classifier got it wrong.



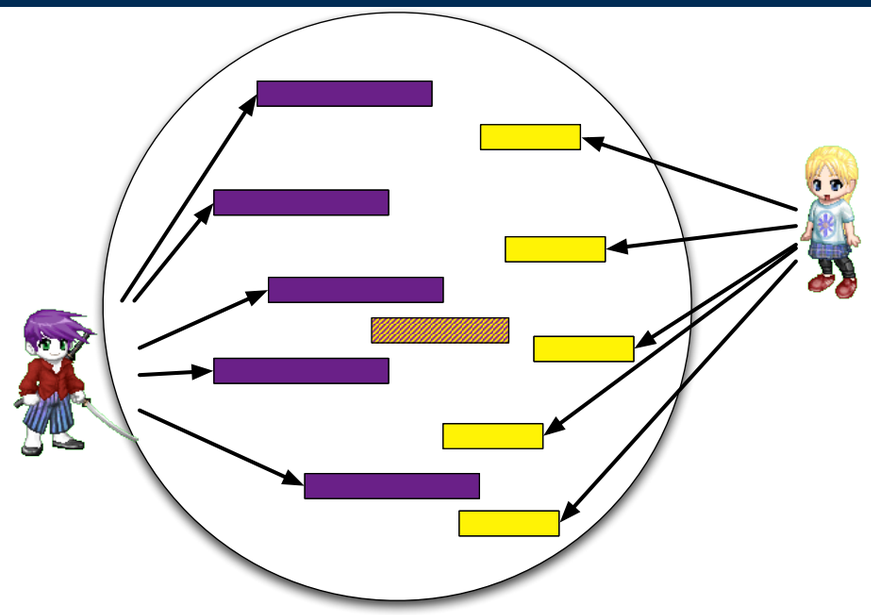
The accuracy is (#right)/ (total #)

You can compute an accuracy separately for each user.

Several factors complicate this data mining problem.

High dimensionality & heterogeneous data

- **All files have:** inode, mode, timestamps, sector #,
- **JPEGs have:** Serial Number, f-stop, exposure date
- **Office docs have:** Author, Print Time, Create Time, etc.
- Solution: reduce dimensions where possible
  - —*consolidated timeline*



Sparse data; many missing values

- Every data element is missing values in multiple dimensions!
- Solution: Use classifiers and distance functions that handle missing values

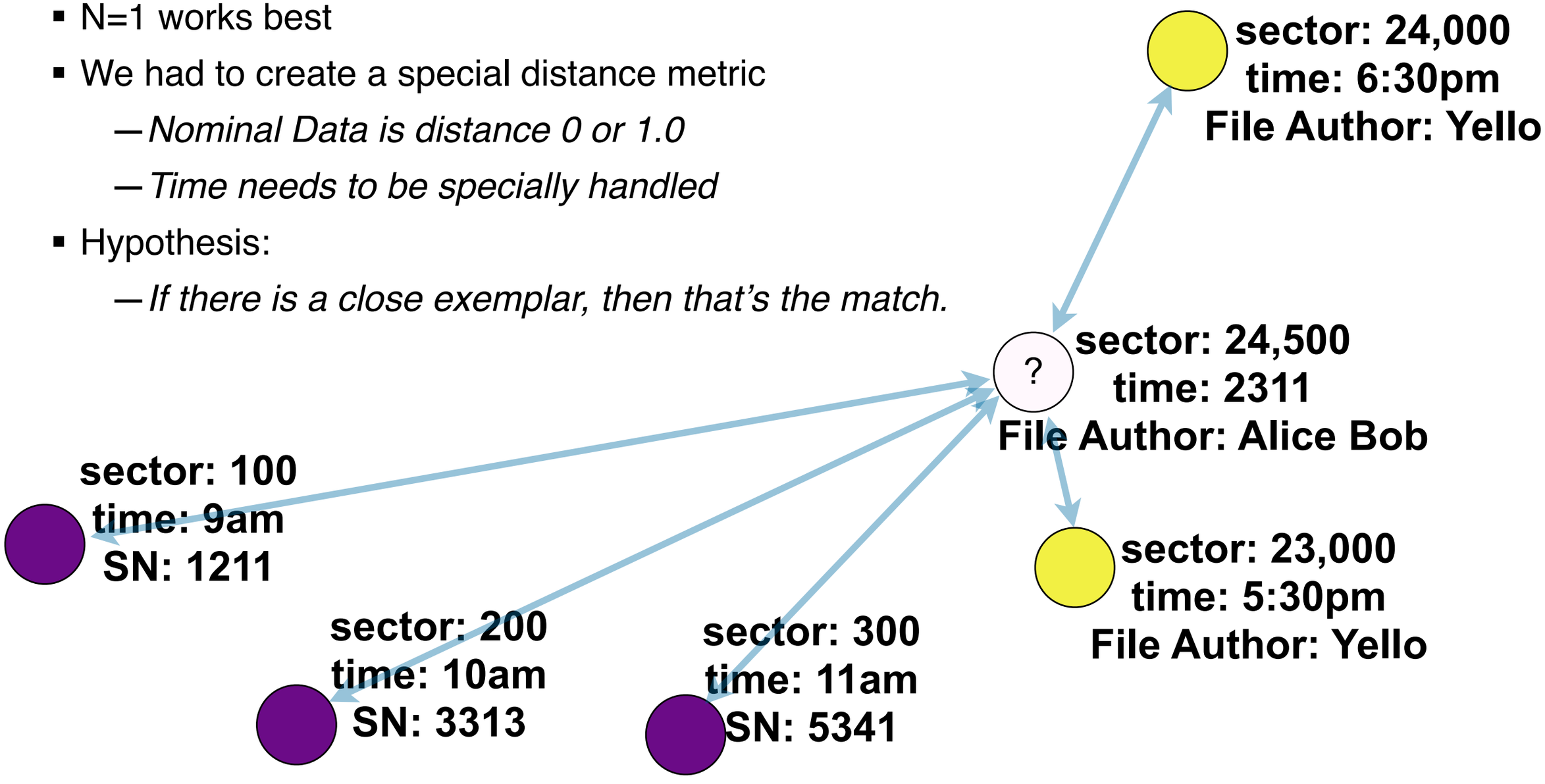
Multiple regions for each class

- User files interleave in time, space, etc.
- Solution: use classifiers which can model multiple regions
  - —*K Nearest Neighbor, Decision trees*

Approach #1: 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.*



Approach #2: Decision Tree

Algorithm: J48

- Implementation of Quinlan's C4.5
- Very fast: typically less than 60 seconds to build.

```
|   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)
```

- Hypothesis:
  - *Files with known owners "bracket" files with unknown owners.*

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.

Classifiers are only used once, for one piece of data

- (Similar to "lazy decision trees")

Research products

Publications to date:

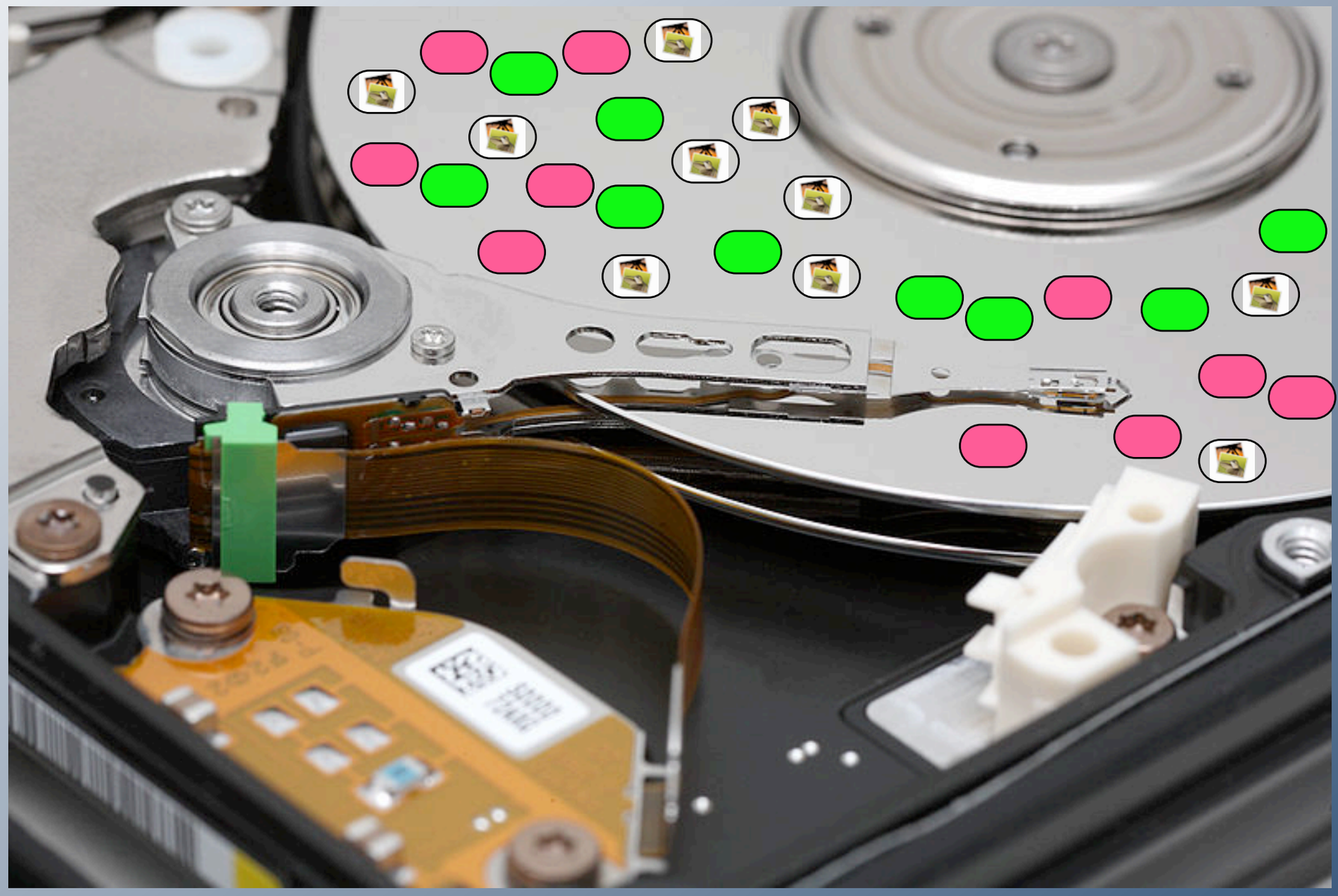
- Cpt. Daniel Huynh, “Exploring and Validating Data Mining Algorithms for use in Data Ascription,” Master's Thesis, June 2008
- Maj. James Migletz, “Automated Metadata Extraction,” Master's Thesis, June 2008
- Garfinkel & Migletz, "The new XML Office Document Files," *IEEE Security & Privacy Magazine*, March/April 2009
- Garfinkel, "Automating Disk Forensic Processing with SleuthKit, XML and Python," IEEE/SADFE 2009, Oakland, CA.

Work in progress:

- "A Solution to the Multi-User Carved Data Ascription Problem," Garfinkel, Parker-Wood, Huynh, Cowan-Sharp and Migletz,

Enabling Tools:

- Hamming, our 1100-core cluster.
- Real Data Corpus



Instant Drive Forensics with Statistical Sampling

Question: Can we analyze a 1TB drive in a minute?

What if US agents encounter a hard drive at a border crossing?

Or a search turns up a room filled with servers?

If it takes 3.5 hours to read a 1TB hard drive, what can you learn in 1 minute?

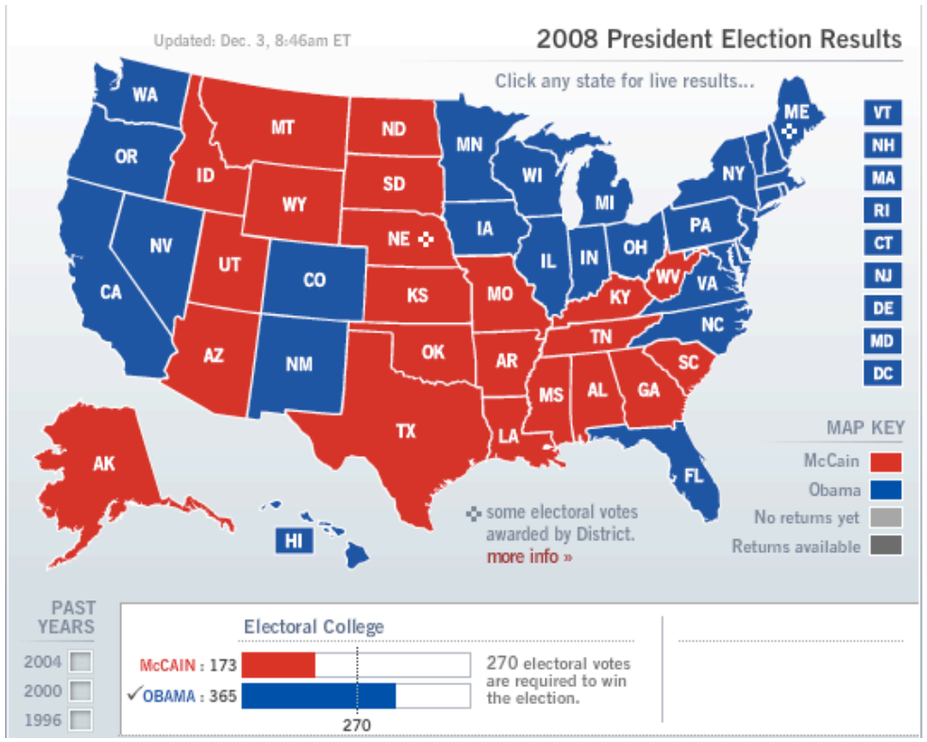
		
Minutes	208	1
Max Data Read	1 TB	4.8 GB
Max Seeks	15 million	17,000 (≈3.5msec per seek)

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

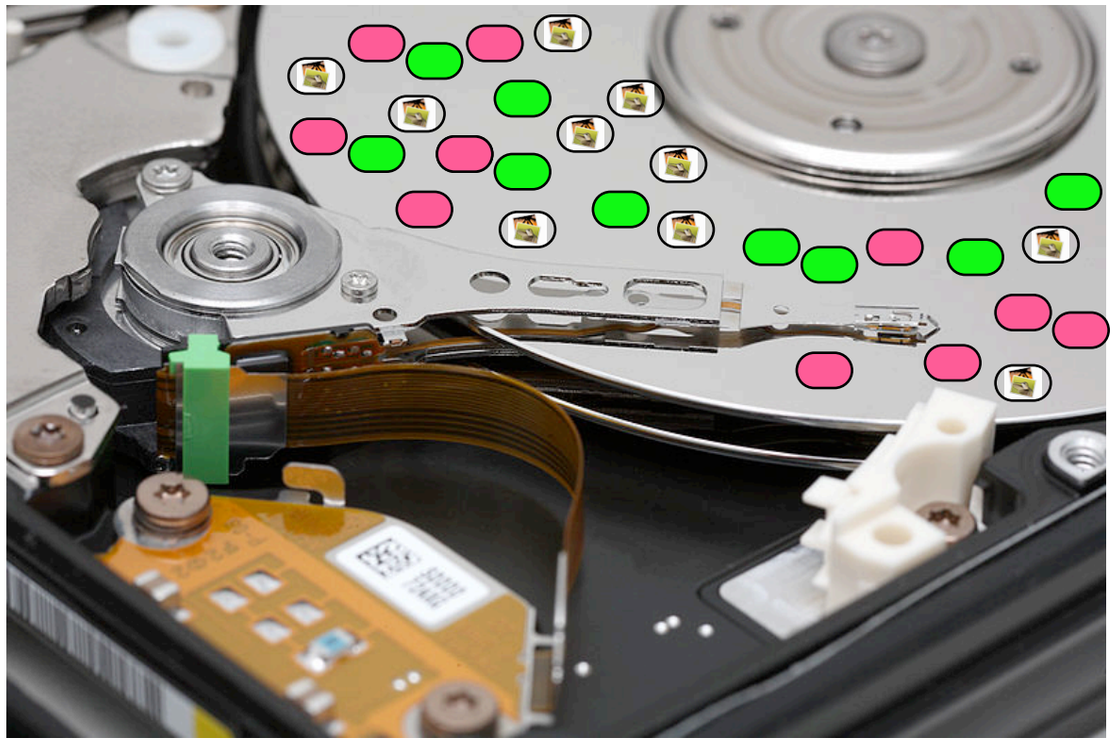
But 4.8 GB is a lot of data!

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

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



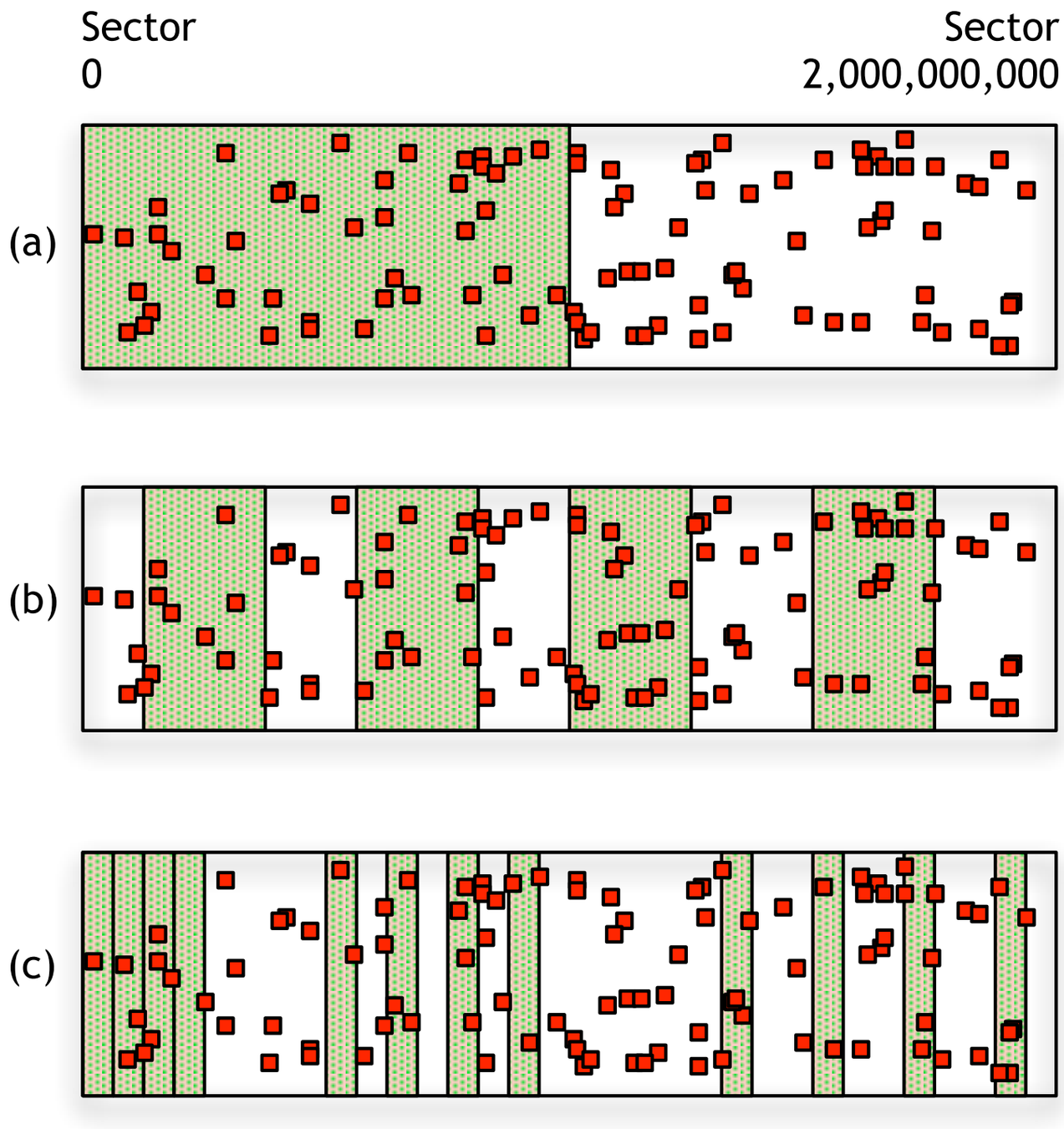
Hard drive contents can be predicted by sampling a few thousand sectors:



The challenge is identifying *likely voters.*

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

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



But sampling has an important limitation...

Data on hard drives divides into three categories:

Resident Data

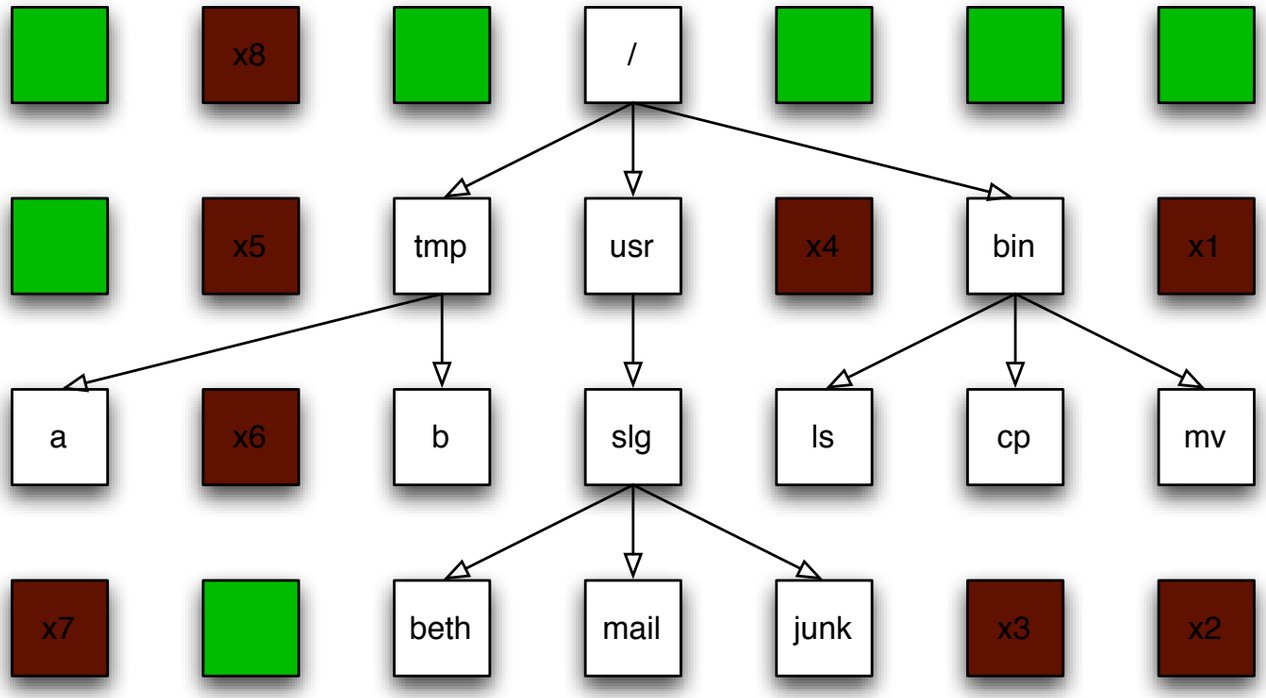
Deleted Data

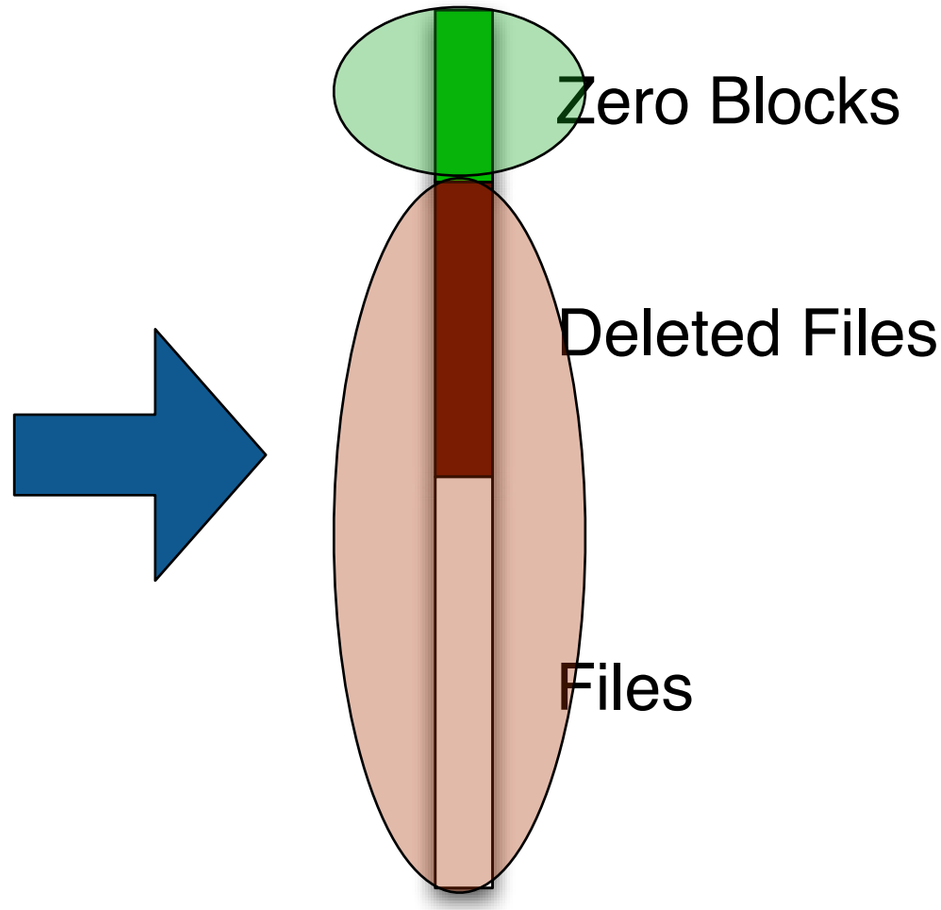
No Data

user files
email messages
[temporary files]

blank sectors

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

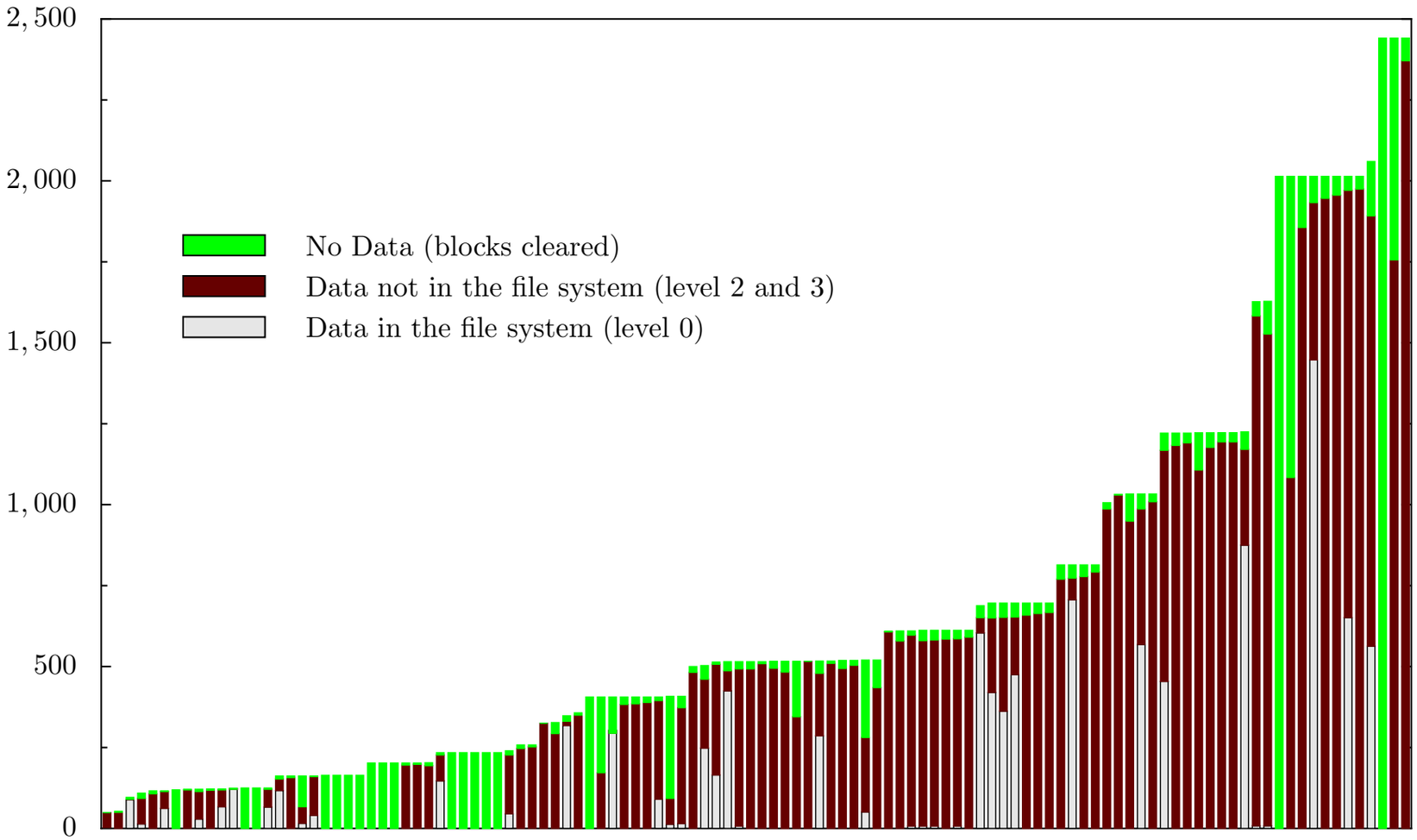




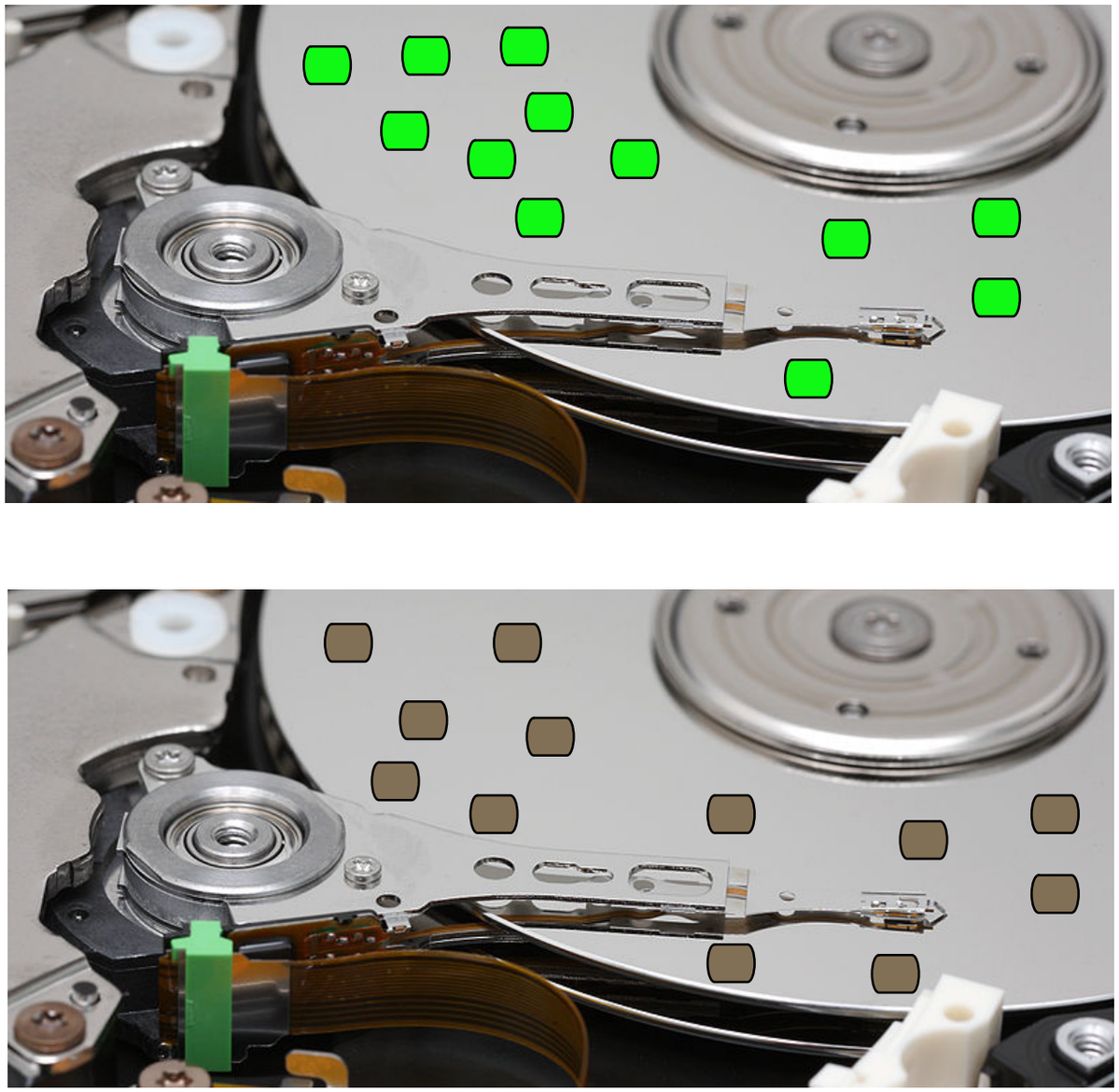
Let's simplify the problem.
Can we use statistical sampling to verify wiping?

I bought 2000 hard drives between 1998 and 2006.

Most of were not properly wiped.



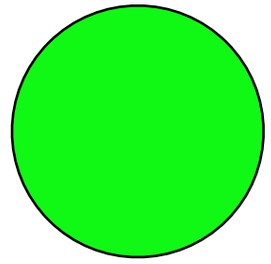
It should be easy to use random sampling to distinguish a properly cleared disk from one that isn't.



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

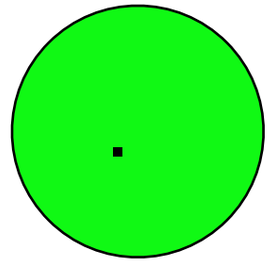
If the disk has 2,000,000,000 blank sectors (0 with data)

- The sample is identical to the population



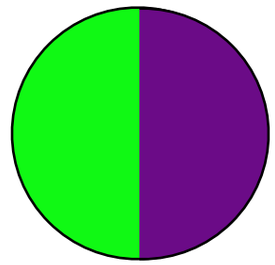
If the disk has 1,999,999,999 blank sectors (1 with data)

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



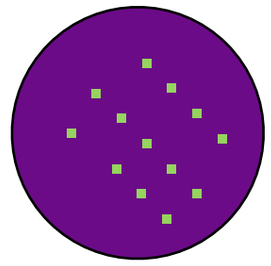
If the disk has 1,000,000,000 blank sectors (1,000,000,000 with data)

- Something about our sampling matched the allocation pattern.
- *This is why we use random sampling.*



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

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



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

Sectors on disk: 2,000,000,000 (1TB)

Sectors with data: 20,000 (10 MB)

Chose one sector. Odds of missing the data:

- (2,000,000,000 - 20,000) / (2,000,000,000) = 0.99999
- You are *very likely* to miss one of 20,000 sectors if you pick just one.

Chose a second sector. Odds of missing the data on both tries:

- 0.99999 * (1,999,999,999-20,000) / (1,999,999,999) = .99998
- You are still *very likely* to miss one of 20,000 sectors if you pick two.

But what if you pick 1000? Or 10,000? Or 100,000?

The more sectors picked, the less likely you are to miss *all* of the sectors that have non-NULL data.

$$P(X = 0) = \prod_{i=1}^{n} \frac{((N - (i-1)) - M)}{(N - (i-1))} \qquad (5)$$

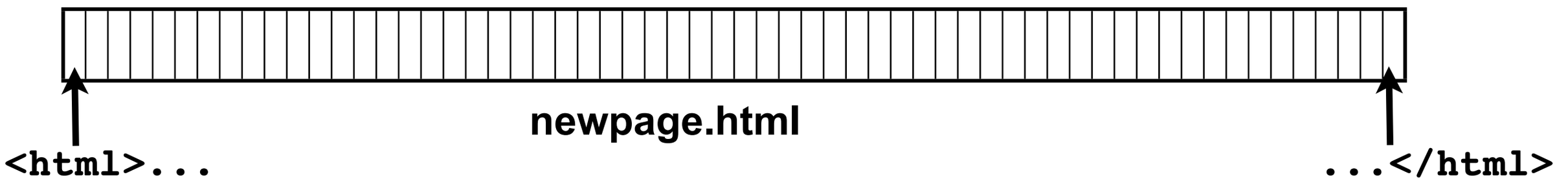
Sampled sectors	Probability of not finding data
1	0.99999
2	0.99998
100	0.99900
1000	0.99005
10,000	0.90484
20,000	0.81873
40,000	0.67032
60,000	0.54881
80,000	0.44932
100,000	0.36787
150,000	0.22312
200,000	0.13532
300,000	0.04978
400,000	0.01831
500,000	0.00673

Table 1: Probability of not finding any of 10MB of data for a given number of randomly sampled sectors. Smaller probabilities indicate higher accuracy.

500,000 blank randomly chosen sectors should be good enough!

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>
```

JPEG files can be identified through the "FF" escape.

- FF must be coded as FF00.
- So if there are a lot of FF00s and few FF01 through FFFF it must be a JPEG.

MPEG files can be readily identified through framing

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

Discriminator tuning strategy

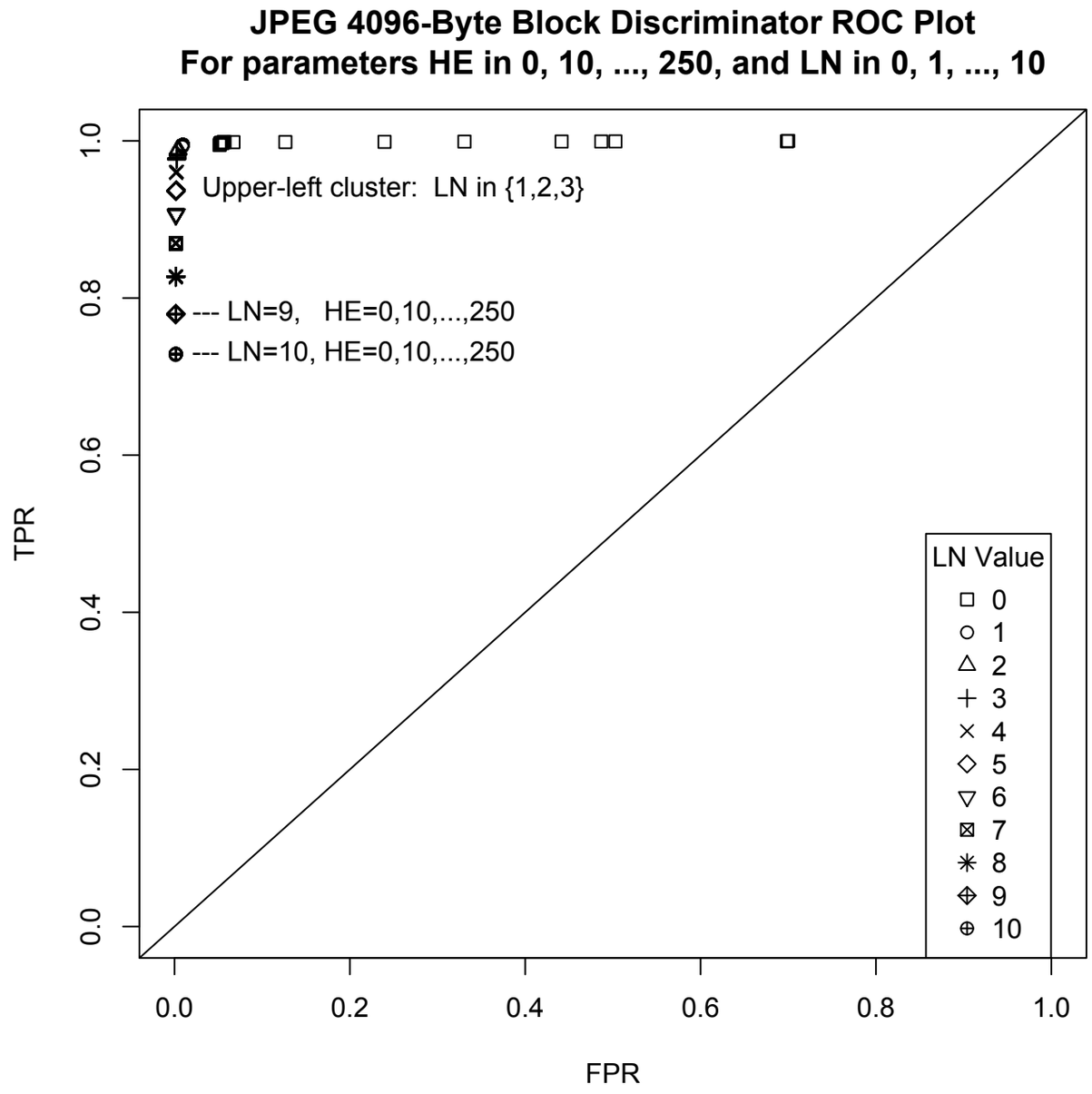
JPEG discriminator has two parameters:

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

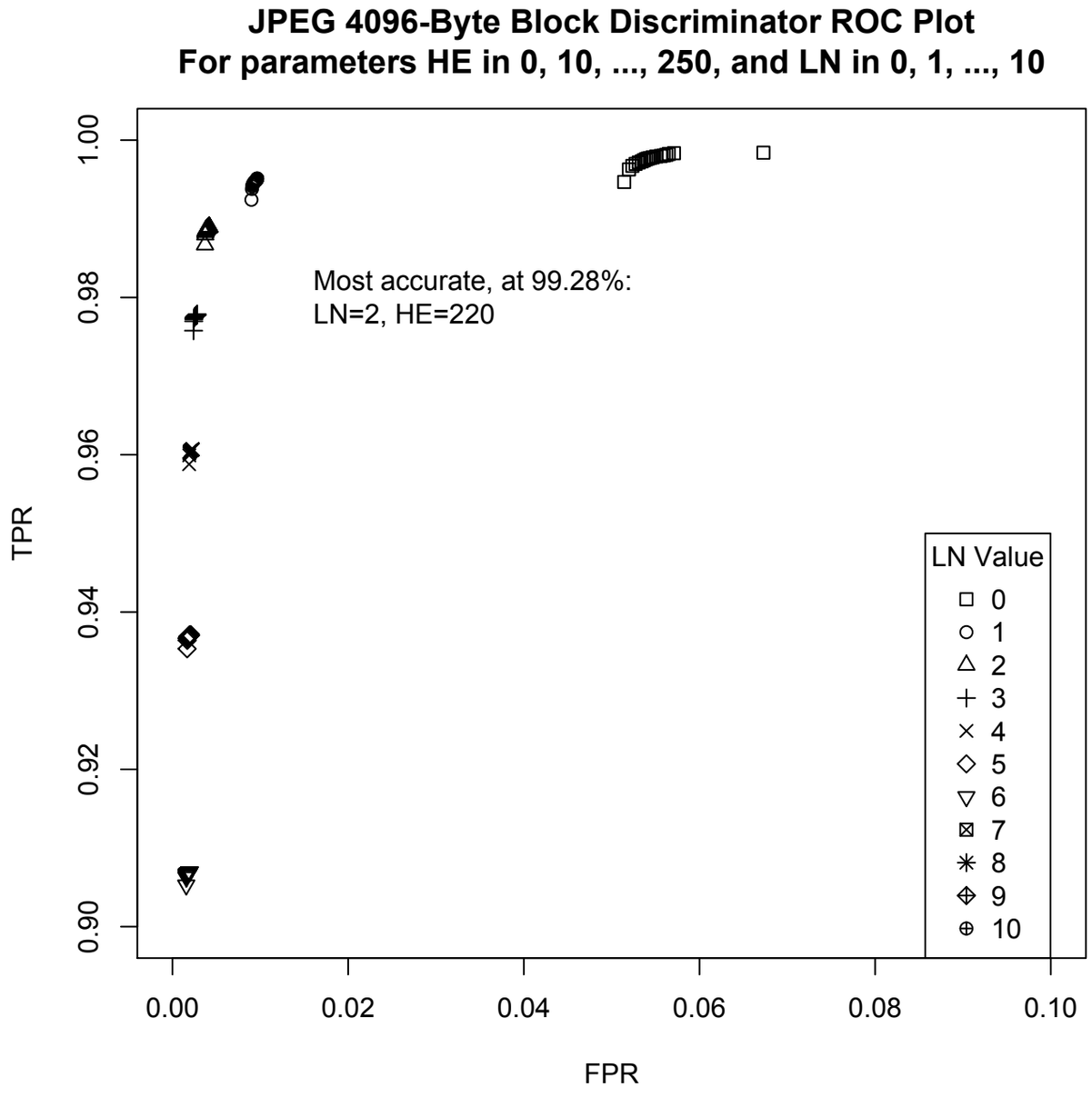
What value pairs give optimal identification?

Let’s just try them all!

Grid search and ROC curves



Grid search and ROC curves



This works!
We identify the *content* of a 160GB iPod in 118 seconds.

Identifiable:

- Blank sectors
- JPEGs
- Encrypted data
- HTML

Report:

- Audio Data Reported by iTunes: 2.42GB
- MP3 files reported by file system: 2.39GB
- Estimated MP3 usage:
  - *2.71GB (1.70%) with 5,000 random samples*
  - *2.49GB (1.56%) with 10,000 random samples*

Sampling took 118 seconds.

Work to date:

Publications:

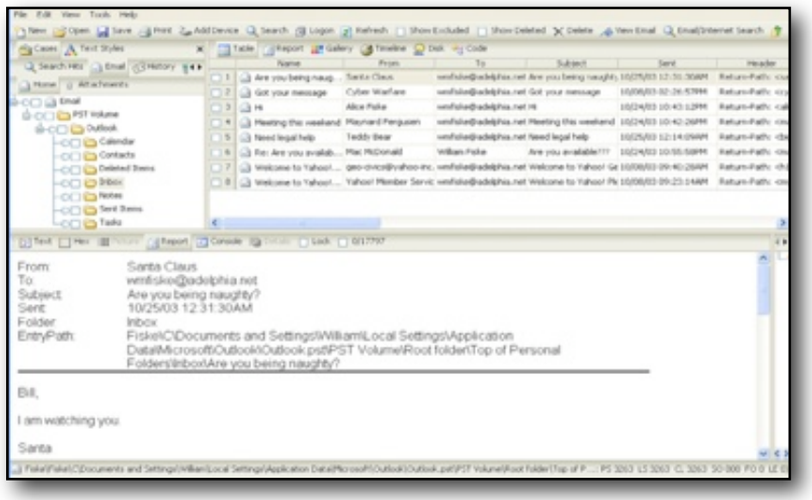
- Roussev, Vassil, and Garfinkel, Simson, File Classification Fragment---The Case for Specialized Approaches, Systematic Approaches to Digital Forensics Engineering (IEEE/ SADFE 2009), Oakland, California.
- Farrell, P., Garfinkel, S., White, D. Practical Applications of Bloom filters to the NIST RDS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

Work in progress:

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

Research Corpora

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



Key technical challenges:

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

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

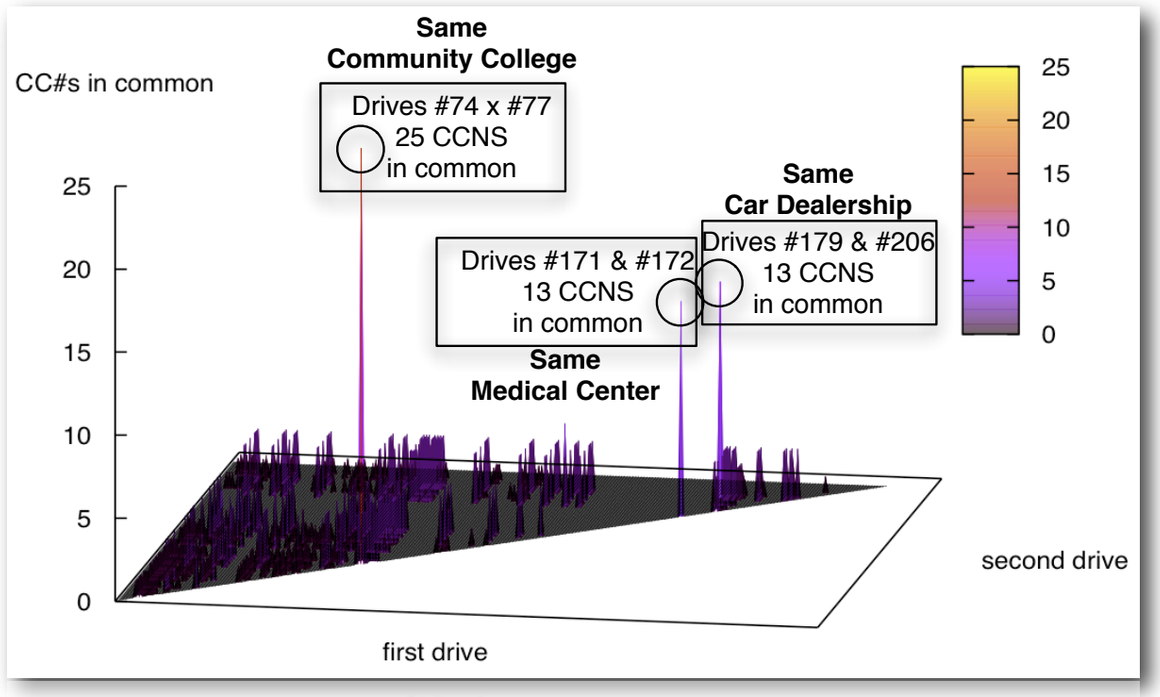
Evidence Reconstruction

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

Clustering and data mining

Social network analysis

Sense-making



Science requires the *scientific process.*

Hallmarks of Science:

- Controlled and repeatable experiments.
- No privileged observers.

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.

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.*

To address this problem, we are creating releasable corpora.

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

http://digitalcorpora.org/

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

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

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

Real data provides:

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

We have multiple corpora:

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

Real Data Corpus: Current Status

Corpus	HDs	Flash	CDs	GB
US*	1258			2939
BA	7			38
CA	46	1		420
CN	26	568	98	999
DE	37	1		765
GR	10			6
IL	152	4		964
IN		66		29
MX	156			571
NZ	1			4
TH	1	3		13
* Not available to USG	1694	643	98	6748

Note: IRB Approval is Mandatory!

Work to date:

Publications:

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

Websites:

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

Work in progress:

- Joshua Gross, NPS postdoc, 2009-2010

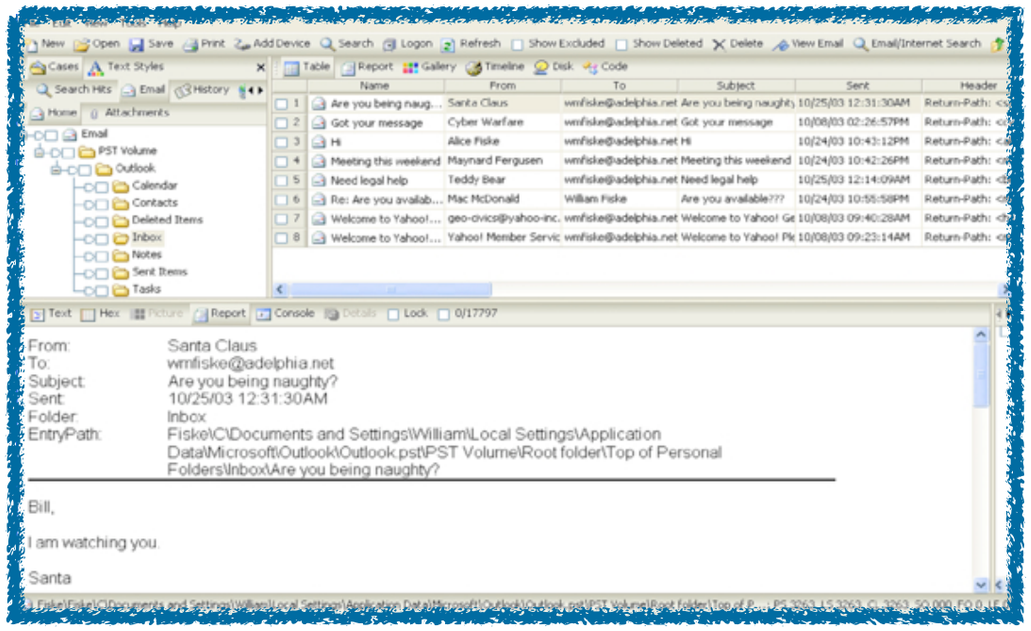
Open Research Problems

Automated Forensics: Open Source Tools and Workflow

Commercial tools are Windows-based GUIs for a single examiner

These tools stress:

- Reverse Engineering
- Visibility
- Search



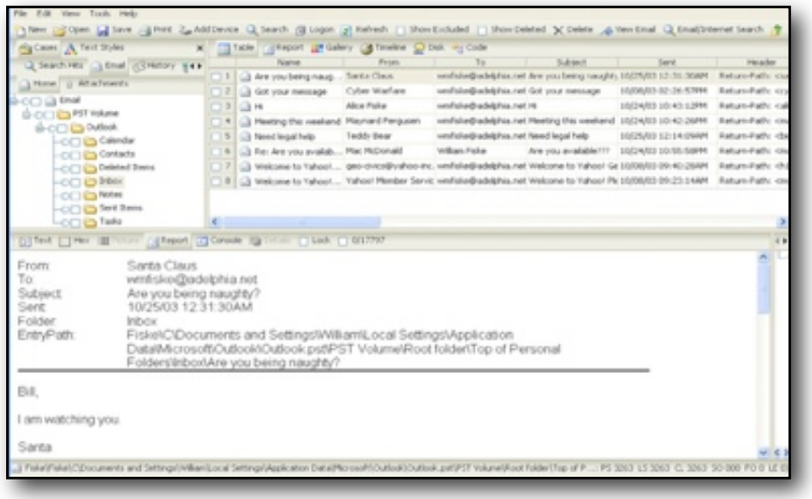
Open Source tools:

- Need all the same reverse engineering
- Need automation to allow for large-scale experiments and exploitation.

Current reverse engineering needs:

- NTFS Encryption
- HFS+ variants used on iPods; XFAT
- Flash file systems (JFFS, JFFS2, YAFFS, etc.)
- Extractors for file formats; a unified system for metadata extraction
- Protection against intentionally corrupt file systems

Forensics needs better visualization tools



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

Key challenges:

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

Flash Forensics: A whole new environment

Traditional assumptions of forensics do not apply to flash.

We may be able to recover:

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

Physical access matters

Cell phones are a nightmare

Cell phones have:

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

Open questions for Cell Phones:

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

Game consoles

Difficulties:

- DRM and encryption
- Proprietary operating systems

Why you should care:

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

Interesting disk forensic problems...

Bulk Data Analysis

- Process disks without accessing files

Data fusion between stored data and online services

Evidence Corruption & Falsification

- Automatically detect falsified evidence
- Detect attempts to destroy evidence.

In Summary

In summary:
Automated Digital Forensics and Media Exploitation

Many research opportunities

- Applying data mining algorithms to new domains with new challenges.
- Working with large, heterogeneous data set.
- Text extraction & clustering
- Multi-lingual
- Cryptography & Data recovery

Interesting legal issues.

- Data acquisition.
- Privacy
- IRB (Institutional Review Boards.)

Lots of low hanging fruit.

Questions?

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.

Sample Questions:

- Can the "ascription" technique be used across multiple hard drives?
- Do flash storage devices and Solid State Drives (SSDs) create new opportunities?
- What are the opportunities for face recognition and other content analysis techniques?
- What can you "correlate" other than email addresses?
- Can overwritten data be recovered?