Building Realistic Forensic Corpora to Enable Undergraduate Education and Research

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

July 27, 2010 — 9:00am - 9:45am

http://digitalcorpora.org/

NSF Award DUE-0919593: "Creating Realistic Forensic Corpora for Undergraduate Education and Research"

NPS is the Navy's Research University.

Location: Monterey, CA

Campus Size: 627 acres

Students: 1500

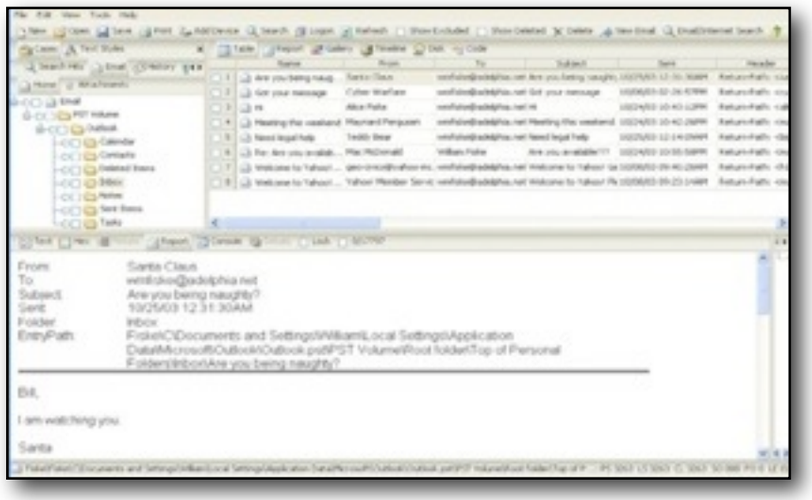
- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)

Digital Evaluation and Exploitation:

- *Research* computer forensics.
- *Develop* "corpora" for use in research & education.
- *Identify* limitations of current tools & opportunities for improvement.
- http://domex.nps.edu/deep/

"The views expressed in this presentation are those of the author and do not necessarily reflect those of the Department of Defense or the US Government."

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

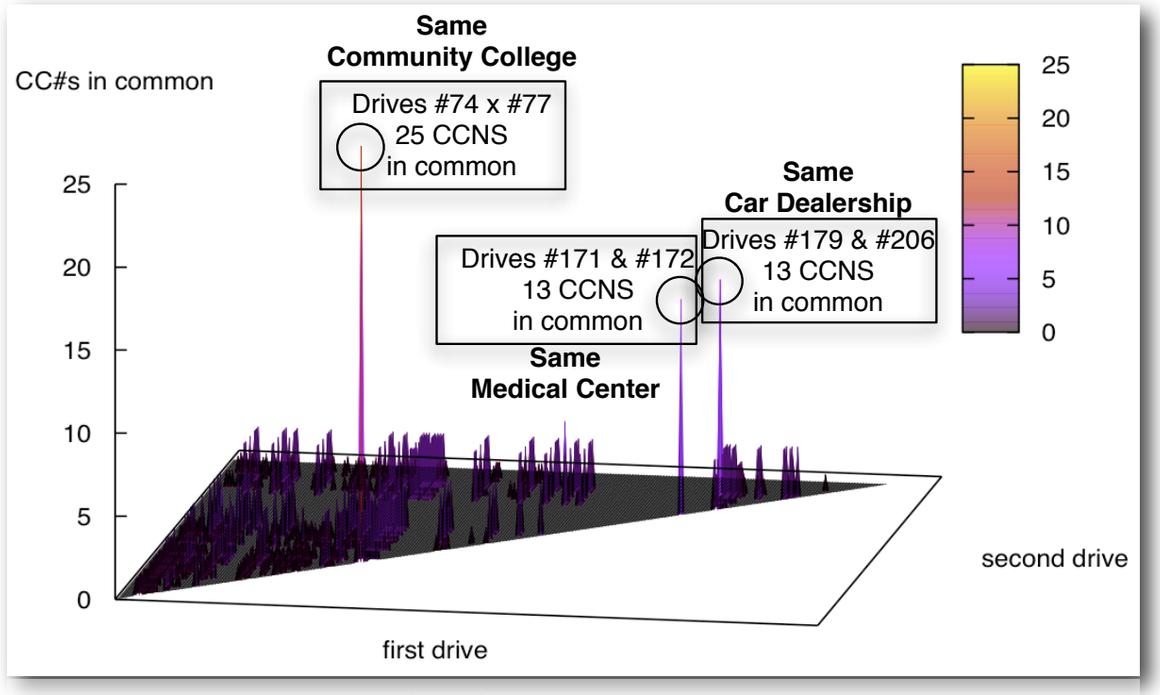
Evidence Reconstruction

- 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.*

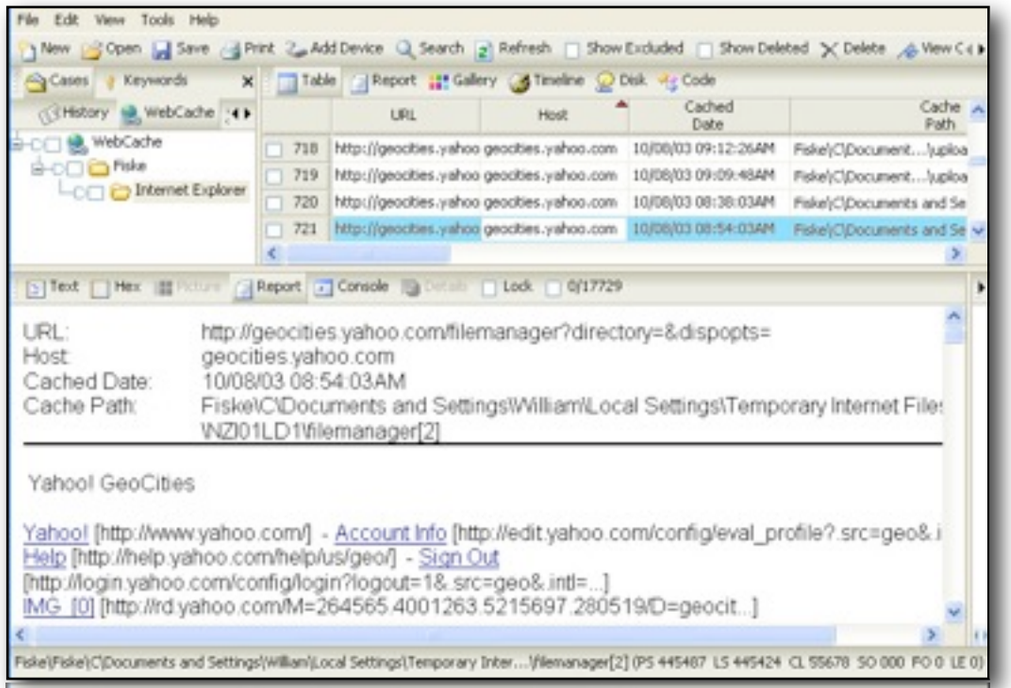
Digital Forensics *education* is also at a turning point.

Traditional Digital Forensics Education: Skills and Tool Mastery

- Disk Imaging
- EnCase (Guidance Software)
- Forensic Tool Kit (Access Data)

Emphasis on:

- Understanding on-disk structures
- Replicating tools used by law enforcement.
- Writing clear reports.



There is a pressing need for Digital Forensics Education.

"Digital Forensic Educational Needs in the Miami Valley Region"

—*Peterson, Raines, and Baldwin, Journal of Applied Security Research, vol. 3, no 3-4,*

Online Survey

- 18 questions (12 substantive)
- 75 solicitations sent to members of the Information Systems Security Association (ISSA)
- 14 full responses, 3 partial (22.6%)

Results:

- 41% said they had any forensics training (7/17)
- 50% said they would be very interested in having their IT employees participate in a forensics course
- 70% had 1-4 security incidents in the previous year.
- 50% had no policy in place for a security incident (but 86% thought they were prepared)
- No perceived need for *forensics* over incident response, network administration, system administration, secure software development, or specific hardware training.

Conclusions:

- "Strong interest in developing a course at the community college level in digital forensics and incident response."
- "Information technology personnel must also be made aware of the legal ramifications their actions can have in terms of criminal law."

What's needed for Digital Forensics Education today?

Teachers

- Teaching digital forensics well requires knowledge of forensic practice, computer science, and law.

Students

- Forensics is seen as computer science, and CS enrollments remain down.
- Students seem more attracted to network security (pen-testing) than forensics.

Forensics Software

- Commercial tools are expensive
- Academic versions of commercial tools are crippled
- Open Source Software is harder to use; does not "train" students for jobs.

Course Materials

- Unlike other fields, real data can't be used in education.
- It's hard to create compelling scenarios.

This talk discusses how we hope to meet some needs of education with digital forensic corpora for education.

Real Digital Forensics with Fake Data

Data that we have created that is available today.

Distributing Forensic Data

- AFF & Digital Forensics XML
- Options for Internet Distribution

Where to go from here.

Real Digital Forensics with Fake Data

“Forensics” has two meanings.

fo·ren·sics n. (used with a sing. verb)

1. The art or study of formal debate; argumentation.

2. The use of science and technology to investigate and establish facts in criminal or civil courts of law.

(American Heritage Dictionary, 4th Edition)

Courts settle disputes, redress grievances, and mete out punishment

Deciding some disputes requires the use of *physical evidence*:

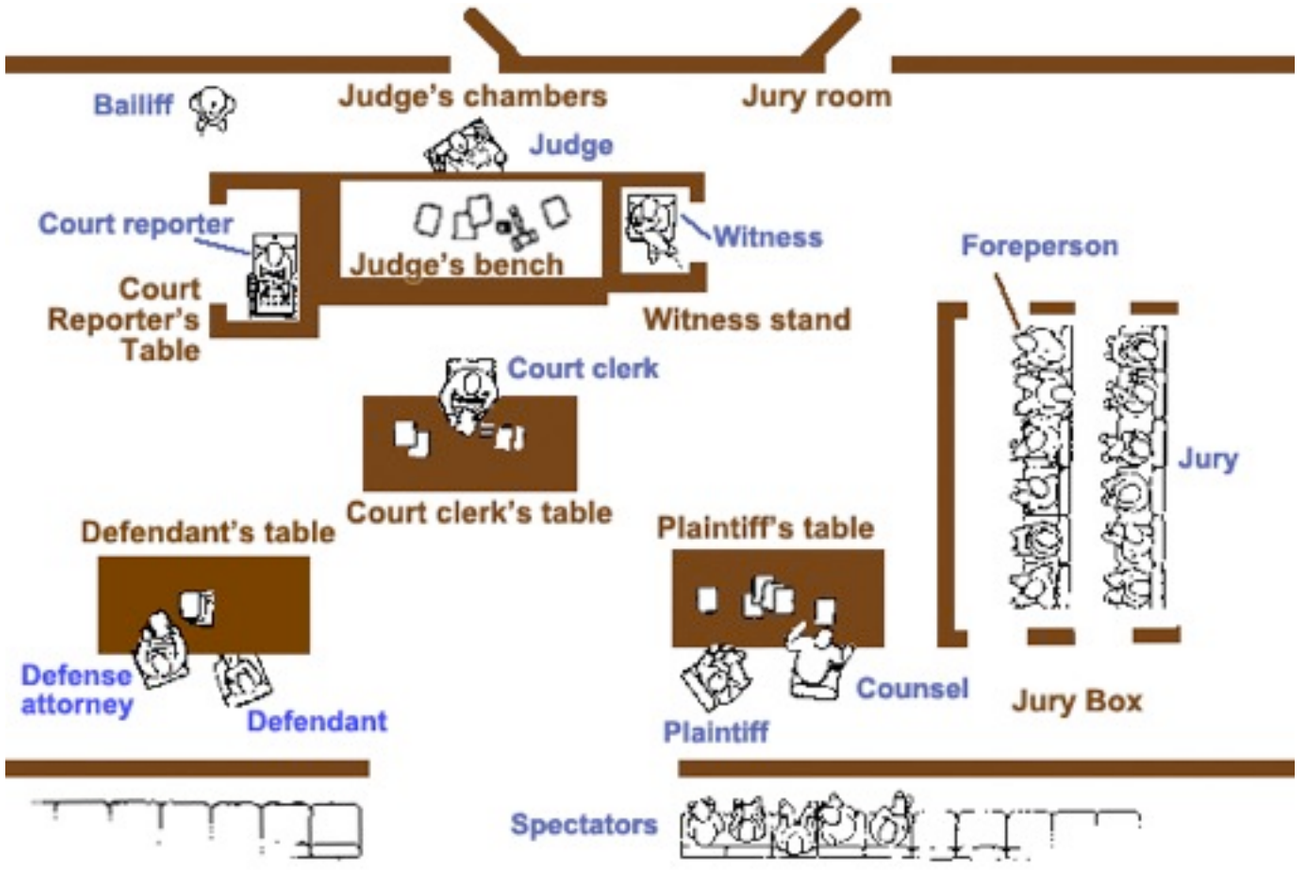
- Fingerprints
- DNA
- Handwriting
- Polygraph

Judges and juries can't examine physical evidence and make a determination.

- They don't have the expertise.
- Evidence may be open to interpretation.

The same is true with digital evidence.

Forensic experts interpret scientific evidence—but interpretation is not value free.



US Courts employ an adversarial process.

Each side hires its own experts.

- Only rarely does the Court hire its own expert.

Investigators for the **prosecution**: conduct the investigation and build the case.

Criminal Digital Investigators:

- Sworn Law Enforcement Officer
- Writes search warrants
- Receives computers, cameras, and other evidence
- Acquires & Analyzes data
- Presents findings
- Prepares report
- Testifies in court

Investigators for the defense: **rebut the evidence** and **create doubt.**

Defense Experts:

- Employed by the Defense
- Works with defense attorney
- Receives evidence from law enforcement
- May conduct independent investigation, but usually funds do not permit
- May work with other experts.
- May testify in court.

Fred Cohen, "High fees, no guarantees."

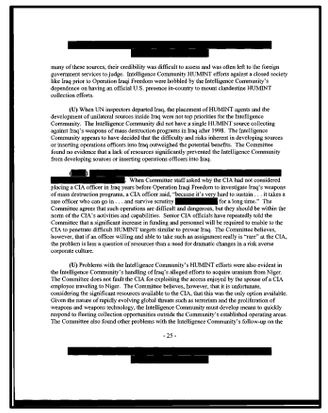
"Digital Evidence" is stuff of modern society.

Today people use:

- Cell Phones
- Electronic
- Laptops & Desktop Computers

We access data stored:

- YouTube
- Facebook
- Chat Services
- Backup Services
- Optical Media
- Personal Storage



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

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

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

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

- 25 -

There are two kinds of Digital Evidence.

Evidence of a crime:

- Financial Records.
- Emails documenting a conspiracy.
- Photographs of a murder.

The crime itself:

- Computer break-ins.
- Denial-of-service attacks.
- Distribution of child pornography.
- Emailed threats.

Everybody in modern society experiences digital artifacts.

- … but not everybody is trained to work with digital evidence.

Digital evidence should be interpretation by an expert

Computer can't be put on the witness stand.

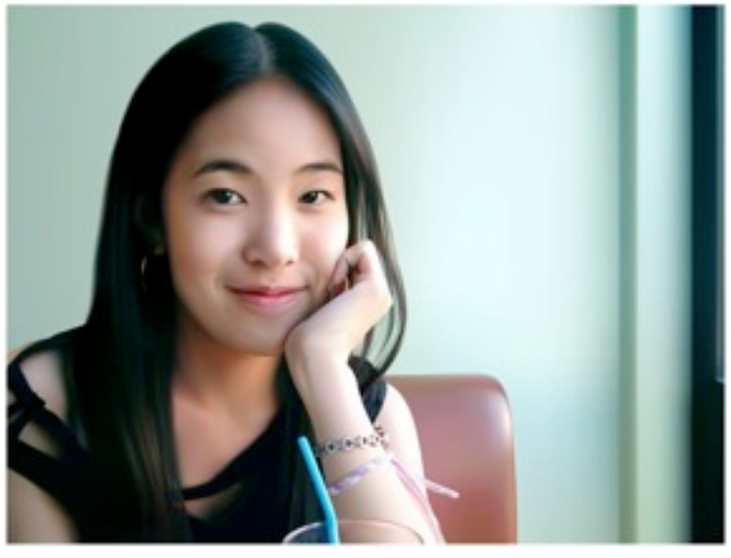
Judge can't analyze the computer.

Evidence must be *identified, preserved and interpreted* by experts.

Even photographs may require interpretation

When were these photographs taken?

Were they faked?

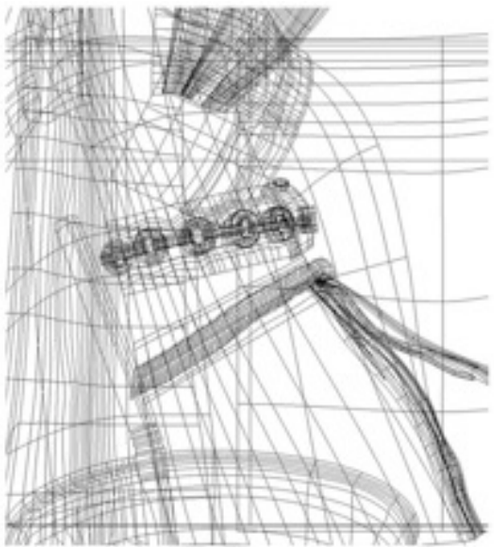


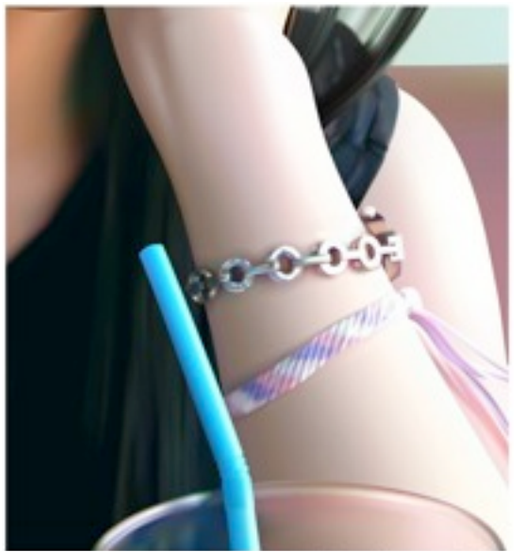
The Commissar Vanishes documents how Stalin's Soviet Union tampered with the past.

After Abel Yenukidze was shot during the purges of 1936-1938, his image was removed from official photographs.

Today it can be impossible to distinguish a synthetic image from a real one.

(Pisan Kaewma, 2006)





It's dramatically easier to create digital forgeries.

Most photos are not "doctored," but most photographs are not taken into court.

If someone has an *interest* in the interpretation of a photo, there is a higher chance of it being modified.

> —*"Digital Doctoring: can we trust photographs?"*
> *Hany Farid,*
> *In Deception: Methods, Motives, Contexts and Consequences, 2007*
> *http://www.cs.dartmouth.edu/farid/*

This is true of *all* digital evidence.

Figure 3. The published (top) and original LA Times photographs showing a British soldier and Iraqi civilians.

What is needed for digital forensics education?

Basic skills:

- How to acquire digital evidence.
- Tools of the trade.
- "Think forensically."
- How to present finding (write a report; testify).

Investigative skills:

- Hypothesis formation.
- Thinking "outside the box."
- Identifying evidence.
- Knowing when you've found enough.

Defensive skills:

- Questioning evidence & process.

What is needed for digital forensics education?

Basic skills:

- How to acquire digital evidence.
- Tools of the trade.
- "Think forensically."
- How to present finding (write a report; testify).

Investigative skills:

- Hypothesis formation.
- Thinking "outside the box."
- Identifying evidence.
- Knowing when you've found enough.

Teaching these skills requires data.

Defensive skills:

- Questioning evidence & process.

Digital Forensics Education poses unique problems.

Wide range of material required for substantive education

- chat logs, photographs, office documents, email messages
- any other kind of data that an investigator might encounter

Same tools are used in many different kinds of cases.

- Child exploitation, Computer intrusion, Financial crimes

Practitioners need to understand a wide range of information:

- Operating systems, file formats, data authentication, law, public policy.

Data of interest to forensic practitioners is:

- Highly personal
- Subject to copyright
- Potentially massive

0101011101001101111000101010101

Fake Data for Digital Forensics Education

Current strategies for obtaining data for forensics education are not sufficient!

Existing Disk Images

- DFRWS "Forensic Challenges"
- Honeynet Project's "Forensic Challenge"
- NIST forensic tool testing images
- Carrier's Digital Forensic Tool Testing website.
- PyFlag "standard test image set"

Real Data

- Drives purchased on eBay
- Students analyzing each other's computers

Data constructed by professors

- Time consuming
- Lack "ecological validity."

Test and Challenge images were not designed for education.

Test data sets (NIST, PyFlag, Carrier)

- Designed to test tools, not students.

Forensic Challenges (Honeynet, DFRWS, DC3)

- Too hard for typical problem set.
- Solutions have been widely distributed.

Real data is inappropriate for the classroom environment

Real data may contain information that is privacy sensitive or legally protected.

- Financial records, Academic records
- Stored passwords
- Email with attorneys, doctors, etc.

Real data may contain content that is *inherently illegal*

- Obscene images
- Child pornography

Real data may contain pornography

- Illegal to distribute to minors

It is impossible to audit the drive for objectionable material

- Encryption & stenography

Why it's impossible to audit used drives...

Consider this hard drive that we purchase via eBay:

Initial audit indicates that it just has copies of movies.

- Several copies of Monster's Inc. (monsters_ink.avi)

Subsequent analysis indicated that it previously held child porn.

- Highly descriptive file names.
- Is the child porn spliced into Monster's Inc?
- Could new "video carving" techniques find child porn in slack space?

We have repeatedly encountered privacy problems with used drives.

Between 1998 and 2010 we purchased thousands of drives on the secondary market.

We have found:

- Evidence of child pornography.
- Credit card numbers.
- Medical records.
- Financial records (ATM machine; Supermarket; Consumer credit applications).

Other bad ideas:

- Public computers.
  - —*Privacy sensitive data; No informed consent.*
- Enterprise computers.
  - —*Possibly creates corporate obligation; potentially illegal.*

Can students analyze their own (or each other's) computers?

Students can analyze their own computers:

- Useful teaching exercise (privacy) (forensics)

But students must do more:

- They need to figure out what happened; can't do that with their own data.
- Need to "think like the attacker."
- Students won't be surprised by their own data.
- Student computers (hopefully) won't have evidence of a crime.

Students shouldn't analyze each other's computers (or friends)

- Potentially exposes confidential information
- Students may discover evidence of crime!
- Possible human subjects issues.

We have encountered problems with students analyzing each other's drives

The assignment: analyze a friend's USB memory stick.

- Use off-the-shelf tools.
- Make a disk image.
- Write a three-page anonymized report.

Problems we encountered:

- "Creepy Factor"
- Lack of informed consent.
  - —*Donors were unsure what was on their drives.*
  - —*In one case, friend's drive contained data from a third-party*
- Privacy and Copyright issue
  - —*Students now had a copy of their friend's data*

How about "gift" or "loaner" computers?

Provide computers and Internet service to poor families.

- Make data collection a condition of the computer's use.
- Pay the families $10/hour to use the computers, so they are your employees.
- Require a signed consent agreement.

Problems:

- Data could be used for *research*, but not for education.
  - —*Filled with PII*
- Third-parties haven't given consent.
  - —*Email correspondents; chat logs; etc.*
- Computers used by study members won't contain a criminal scenario worthy of investigation.
  - —*(Hopefully it won't!)*

Forensic Data: What's Left?

Public Real Data (from real people) — No privacy issues!

- Enron Email Corpus
- YouTube Videos
- Public FaceBook profiles
- Public chat serves

Honeypot Data (from real attackers)

- Honeypots generally lack "ecological validity."
- Attacks may or may not be realistic.
- Attackers may be covered under human subject regulations.

Fake Data

- Constructed Scenarios; Made-up attacks; real attack techniques
- Actors playing the roles of fake people

Forensic Data Generators

- University of Mannheim, "Forensic Image Generator Generator Project."

Issues in Creating Realistic Data

What do we want from our fake data for digital forensics education?

Internal Consistency with Scenario

- Easily done if data is created, modified and accessed in *real-time* with *real tools.*
- Avoid bumping the clock; auto-browsing the Internet.

Depth, Detail and Realism ("Ecological Validity")

- Data should be *representative* of real data.
- There should be data that is *relevant* and *not-relevant* to the case at hand.
- Data should be rooted in a specific *time* and *date*. (Because the real world is.)
- Data should be *reasonably current.* (Ten year old applications are of limited used.)

Size

- Large enough to be useful, small enough to be portable.

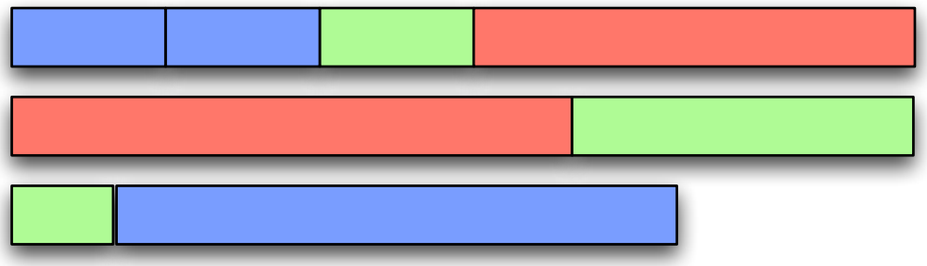
Range of Difficulty

- Easy, Average, and Hard problems to "titrate" the student's skills.

Example: File system "wear"

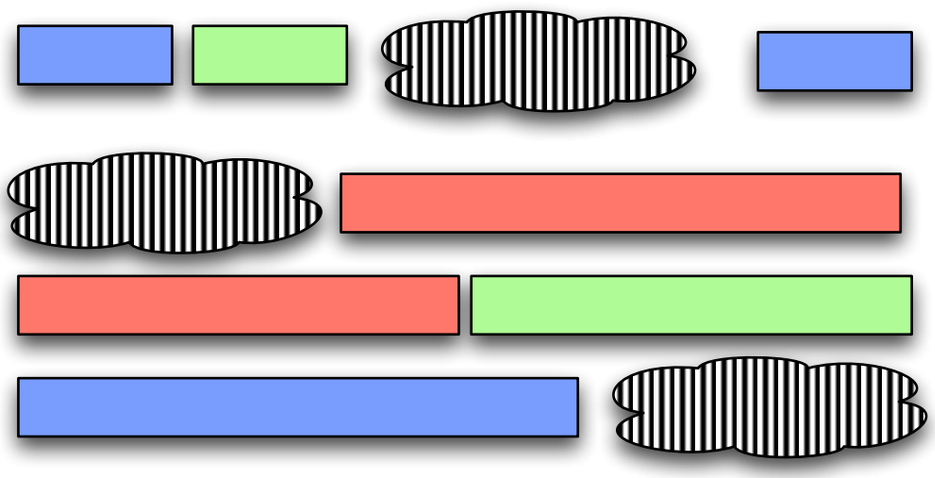
If a set of files are copied to a clean disk, they are neatly arranged:

- In alphabetical order.
- No space between files.
- No deleted directory entries.



But if a disk is used "normally," the files can become:

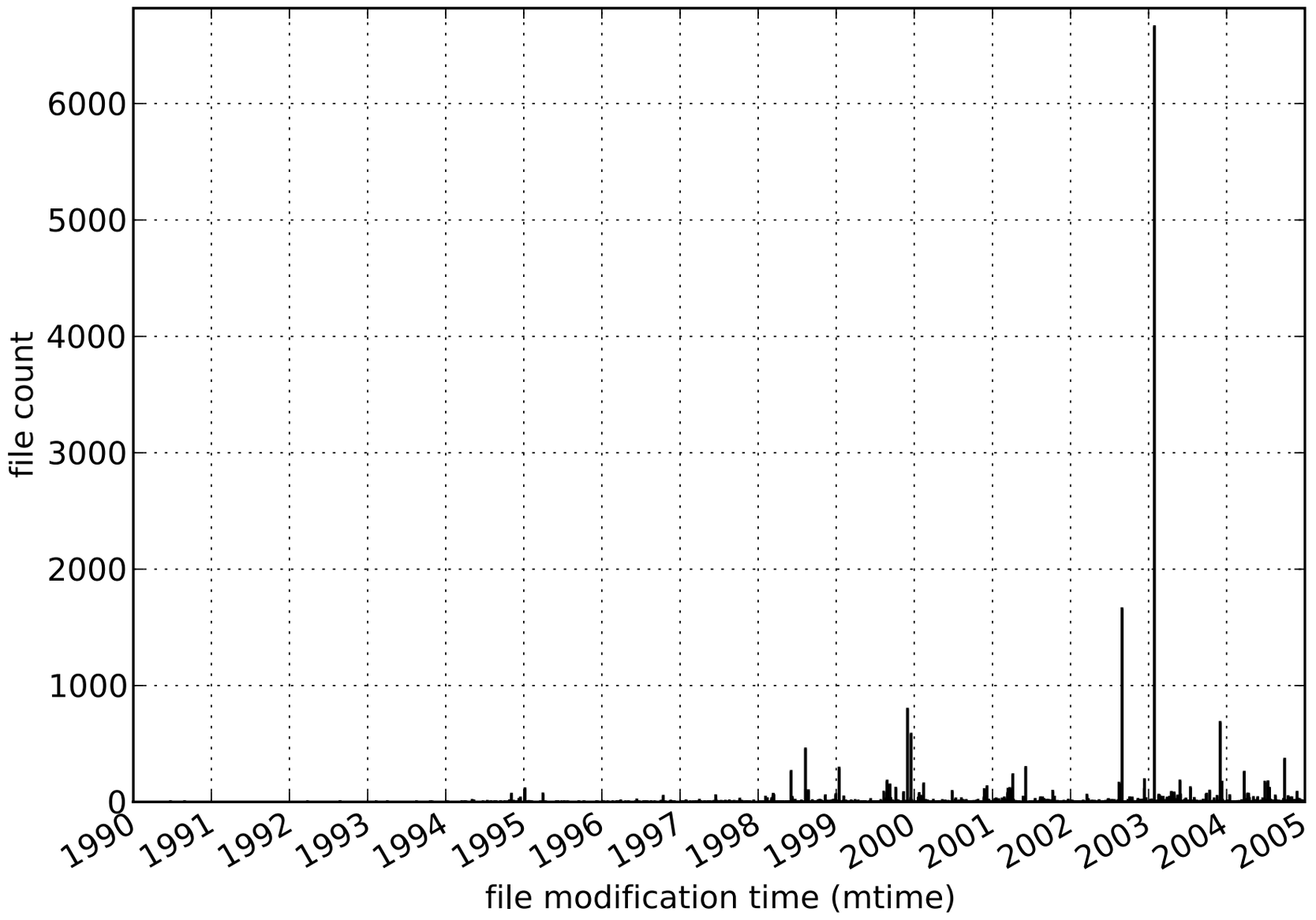
- Fragmented
- Files are "out-of-order"
- Unallocated sectors may contain residual data.



Example: File Modification Times.

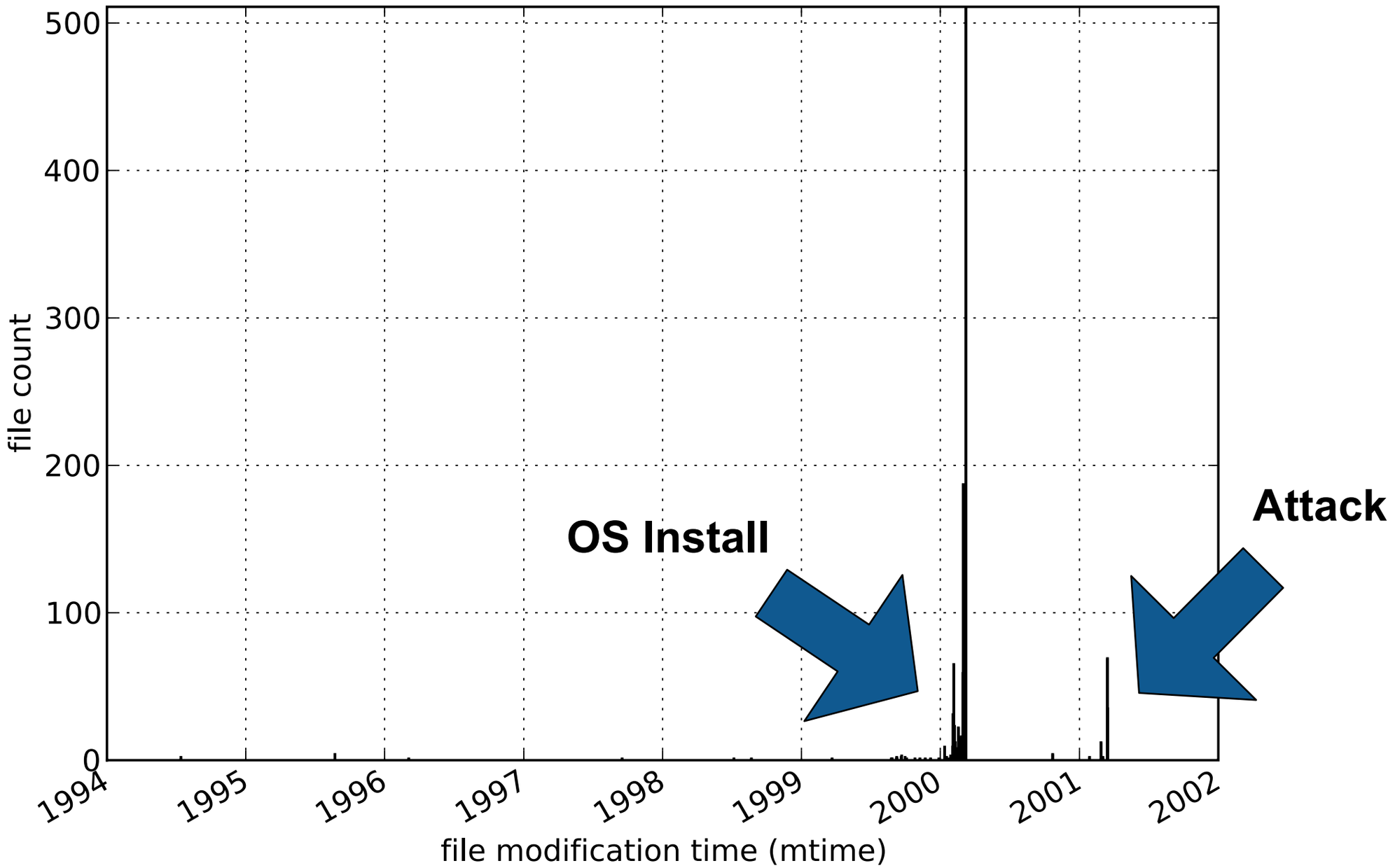
When disks are used for a period of time, they retain files with a wide range of timestamps:

- Disk 0844 contains files from 1995 through 2005, with many files being modified in 2003:



File Modification Times Histogram from Honeynet Scan 15

After the install, the clock was probably advanced in preparation for the attack.



Creating fake objectionable material.

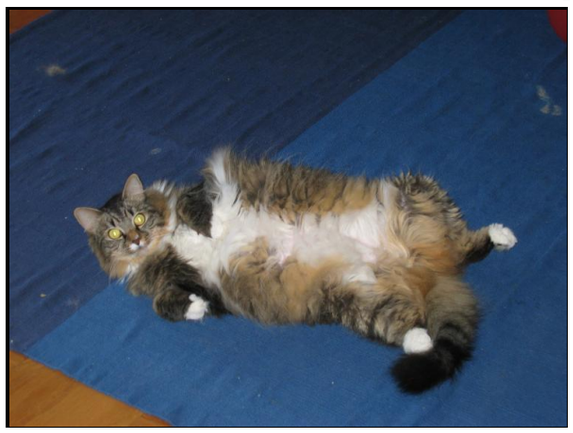
Many investigations look for objectionable material.

- Child pornography.
- Stolen proprietary information.
- Stolen PII (credit card numbers, SSNs, etc).

We *can't* distribute this information in our educational materials.

As an alternative, investigators have used *simulant* photos.

- Rockets
- Cats (Kornblum)



But there has been no good analog for hash sets of objectionables.

- e.g. National Center for Missing and Exploited Children's NCMEC's "Innocent Images."

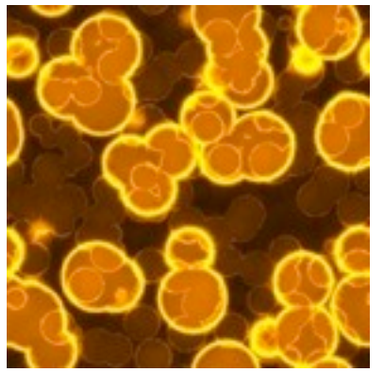
Standardized Forensic Corpora

Recall that Digital Forensics is becoming a *scientific process.*

Hallmarks of Science:

- Controlled and repeatable experiments.
- No privileged observers.

Other sciences resolve this problem with standards and corpora.



- Biologists:
  - *Trade cell lines*
  - *Apprentice in labs to master techniques.*
- Physicists and Chemists:
  - *Trade physical samples.*
  - *Establish "scientific standards" for calibrating machines.*

For Digital Forensics, we need forensic corpora.

- Education — No PII, No Objectionable Content
- Research — Multiple corpora (some sensitive, some not)

Corpora Modalities:
What kind of corpora does digital forensics need?

Disk Images

- The most fundamental kind of corpora.

Memory Images

- Urgently needed for both research and training
- Not interesting unless sensitive.

Network Packets

- Laws makes collection very problematic (wiretap, privacy)

Files

- File identification
- Data and Metadata Extraction
- Classification; Clustering; Information Extraction

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.

Corpora with private information cannot be used in federally funded research without IRB approval.

Experiments are only exempt under 45 CFR 46 if:

- "…[T]hese 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."

Federal regulations are silent on using information for non-research purposes provided that confidentiality is maintained.

- However, if information was collected with IRB approval, the IRB can prohibit redistribution.

We have created dozens of disk images, packet captures, and memory dumps.

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)

Scenarios:

- M57 startup — spear phishing attack
- M57 patents — small business victim of internal hacking
- Nitroba University — Harassment case solved through network forensics

http://digitalcorpora.org/

NPS-govdocs1: 1 Million files

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

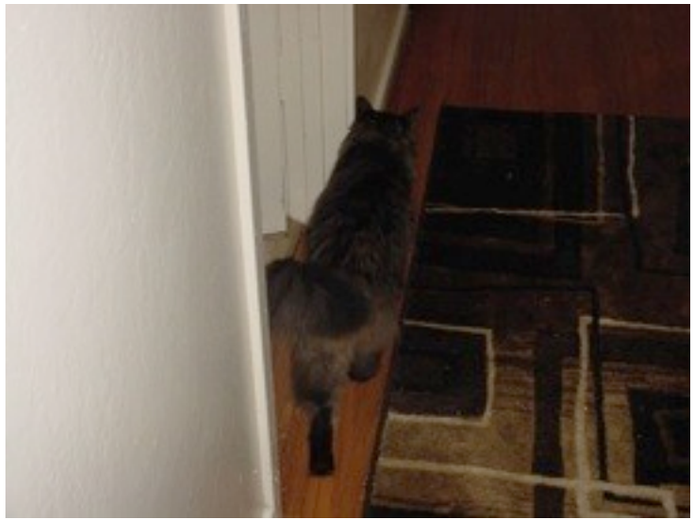
- 1000 ZIP files with 1000 files each.
- 10 “threads” of 1000 randomly chosen files for student projects.
- Full provenance for every file (how found; when downloaded; SHA1; etc.)

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

Monterey Kitty — Simulated "Kitty Porn"

Still images and movies of cats

- 43 JPEGs
- 107MB .mov file
- 3 20MB .m4v files



Hash files

- md5, sha1, sha256 of the files
- md5, sha1, sha256 of 4K sector hashes

Coming: Fuzzy Hashes

- ssdeep
- vrsd

2009-M57-Patents: A complex scenario with theft, objectionable materials, and IP exfiltration.

Scenario takes place in November 2009.

- A person purchases a used computer from a seller via CraigsList.
- The buyer finds "cat photographs" on the computer.
- Police trace the computer to "M57," a small company in Monterey, CA.
- Police call the company's president, inform that they are coming over.
- The president tells the staff that the police are coming.
- Police arrive and find one of the computers has been wiped.

Last day Scenario data includes:

- Disk images from 5 computers & 4 USB drives & [1 cell phone]
- Detective reports of interviews with each of the company's employees.

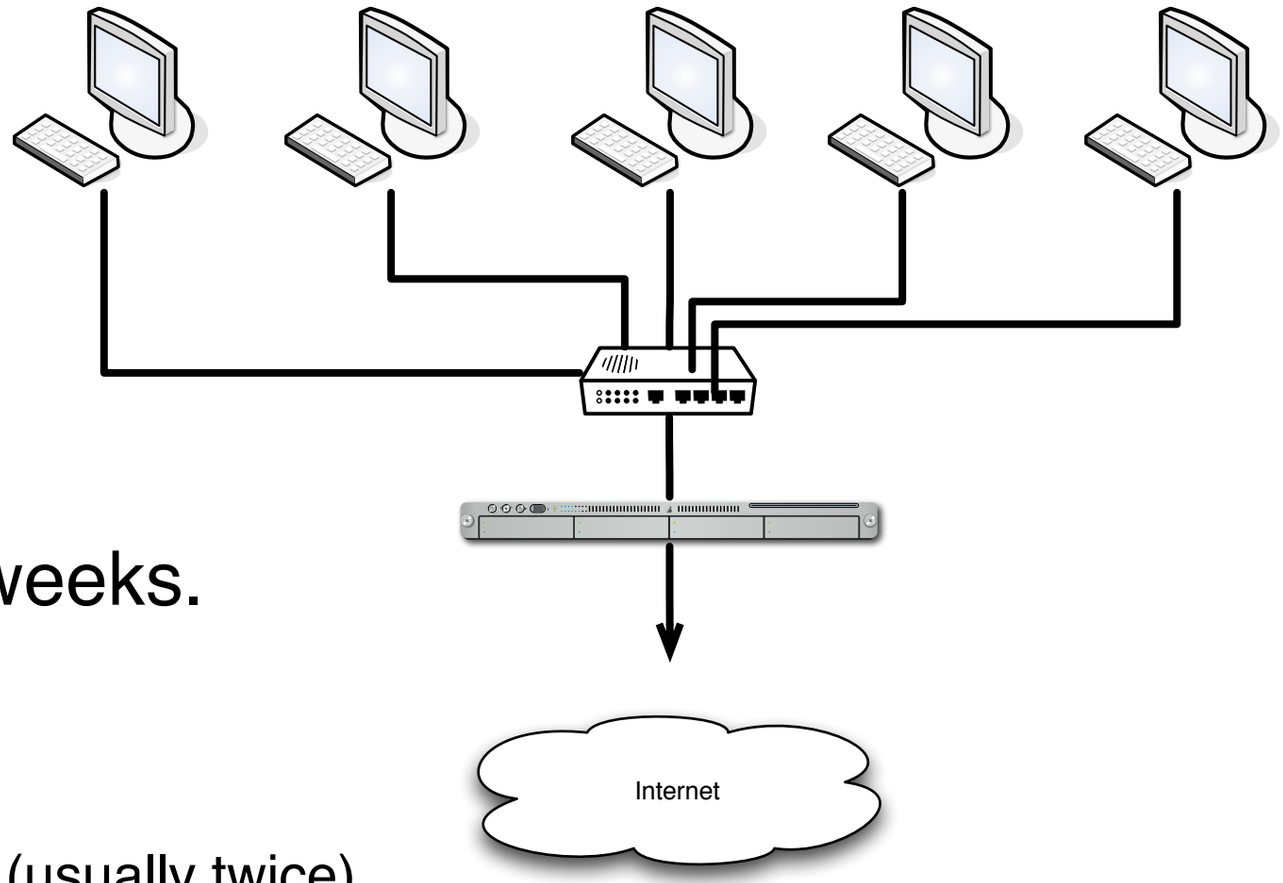
"Extended" Last day scenario data includes:

- **RAM Dumps:** Memory from each of the computers
- **Wiretap Order:** Network packets from the company via a tap at the ISP.

Here's the "lab" where we did it.

We have a *lot* more data to enable research.

M57 Scenario Network:



Scenario extended for 3 weeks.

For each day:

- Imaged every hard drive
- Imaged RAM of each machine (usually twice).
- Captured all packets in and out (except for 1 day).

Uses of the M57-Patents data set.

Digital Forensics Education:

- Solve the "objectionable material" scenario.
  - *How did the cat photographs end up on a machine sold on ebay?*
  - *Who is the insider that is responsible for the photos?*
- Solve the corporate theft scenario.
  - *Who was selling the company's computers on eBay?*
  - *What was stolen? How was it done?*
- Solve the intellectual theft issue.
  - *What information was being exfiltrated?*
  - *Establish the proof.*

Computer Forensics Research:

- Memory analysis.
- Malware analysis
- Network data visualization.
- Correlation between different modalities.

Microsoft's copyright is addressed with selective redaction.

Distribution of disk images has been hampered by copyright fears.

- Disk images can be boot under VMWare.
- So publicly releasing working disk images of Windows can be a copyright violation!

Copyright law provides for "fair use" in section 107:

1. The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes
2. The nature of the copyrighted work
3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole.
4. The effect of the use upon the potential market for, or value of, the copyrighted work.

Our approach: overwrite executable instructions in any Microsoft executable that includes the word "copyright."

1. Non-profit, education & research.
2. Windows is one of the most widely distributed pieces of copyrighted data on the planet.
3. With redaction, we are not distributing the entire Windows OS.
4. No devaluation of market value (Windows can't run).

Additional Issues.

Software Suites:

- We use Open Office (no redistribution issues)

Web Content:

- Most of the downloads were patents from USPTO and reports from .GOV websites.
- Some browsing of news sites.
- Some auto-browsing.

For organizations that have a MSDN license.

- The unredacted images are available for download as well, but they are AFF encrypted.
- For organizations with MSDN, we make available the decryption key.

Digital Forensics XML

Digital Forensics XML (DFXML) is a tool for describing file systems and file *metadata*.

Today most forensic tools report metadata in human-readable form.

- Location of partitions.
- Location of a file.
- File owner, MAC times, etc.
- Microsoft Office permissions.

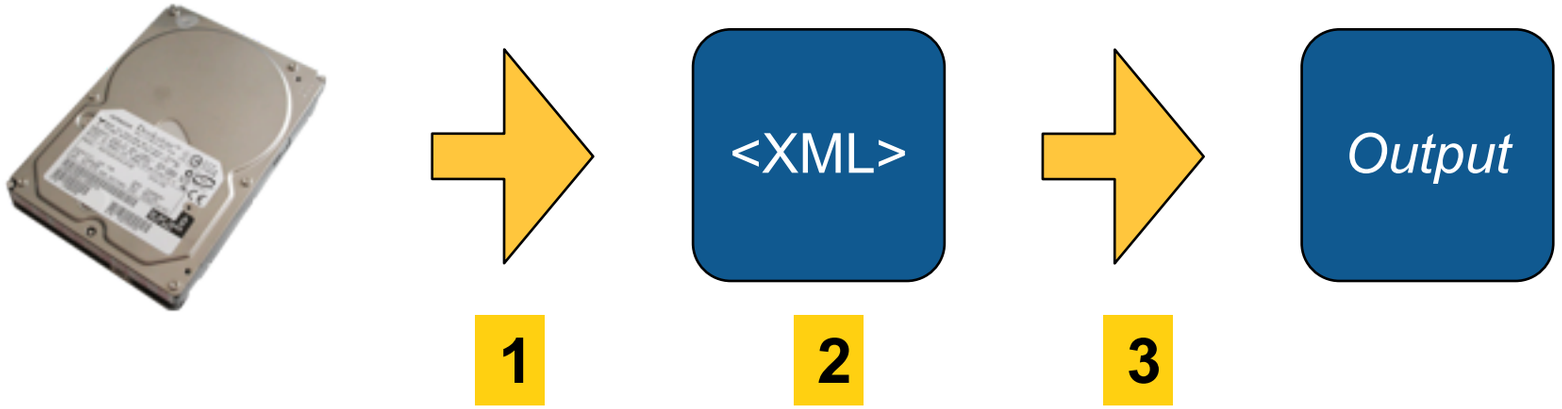
This leads to problems:

- Each tool processing a disk image must re-interpret the file system.
- One tool cannot be easily validated against another.

DFXML allows tools to interoperate.

The basic idea: use XML as an intermediate format.

XML allows us to separate file extraction from forensic analysis.



You can start using this framework today.

You can easily expand it.

Currently DFXML has four kinds of XML tags.

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 offset="26112">
```

Per <volume> tags:

```
<volume offset="26112">
  <Partition_Offset>26112</Partition_Offset>
  <block_size>512</block_size>
  <ftype>4</ftype>
  <ftype_str>fat16</ftype_str>
  <block_count>60749</block_count>
```

Per <fileobject> tags:

```
<fileobject>
  <filename>DCIM/100CANON/IMG_0001.JPG</filename>
  <filesize>855935</filesize>
  <byte_runs>
    <run file_offset='0' fs_offset='55808' img_offset='81920' len='855935'/>
  </byte_runs>
</fileobject>
```

We used DFXML to document the redactions in the M57-Patents corpus.

```
<fileobject>
<filename>WINDOWS/$hf_mig$/KB960859/SP3QFE/telnet.exe</filename>
  <filesize>76288</filesize>
  <inode>15314</inode>
  <redact_image_offset>5157973504</redact_image_offset>
  <redact_bytes>4096</redact_bytes>
  <before_redact>
    <hashdigest type='MD5'>971eb043532293651f46bb55cd7cd101</hashdigest>
    <hashdigest type='SHA1'>c68b7bd95913c848ecab2905e9e443fc4450dbc2</
hashdigest>
  </before_redact>
  <after_redact>
    <hashdigest type='MD5'>971eb043532293651f46bb55cd7cd101</hashdigest>
    <hashdigest type='SHA1'>c68b7bd95913c848ecab2905e9e443fc4450dbc2</
hashdigest>
  </after_redact>
</fileobject>
```

Summary

Where to go from here...

For M57-Patents, there's a lot of work to do…

- Teaching materials
- Annotations for the full corpus (1.0TB) — to enable research
- Tools to create the annotations automatically.

"Laboratory on a CD" — Bootable CDs with tools, data, systems

- BackTrack — forensics/pen testing CD
  - —*http://www.backtrack-linux.org/downloads/*
- Damn Vulnerable Linux — pre-installed with hundreds of vulnerabilities
  - —*http://www.damnvulnerablelinux.org/*

More scenarios, course packs, etc.

Courses that use these materials.

In Summary...

We need fake data to teach real forensics.

- No PII
- Multiple levels of problems to titrate student ability.
- Excellent tool to transition students from building investigation skills to conducting original research.

There are a *lot* of issues in creating such data sets.

- Internal Consistency
- Ecological Validity
- Copyright

We have data that you can download *today*, but the teaching materials aren't ready yet.

- You can help us create them!

NSF Award DUE-0919593: "Creating Realistic Forensic Corpora for Undergraduate Education and Research"