
# Bringing Science to Digital Forensics with Standardized Forensic Corpora.

Simson Garfinkel, Paul Farrell, Vassil Roussev and George Dinolt

DFRWS 2009
August 17, 2009

# NPS is the Navy's Research University.

Location: Monterey, CA

Campus Size: 627 acres

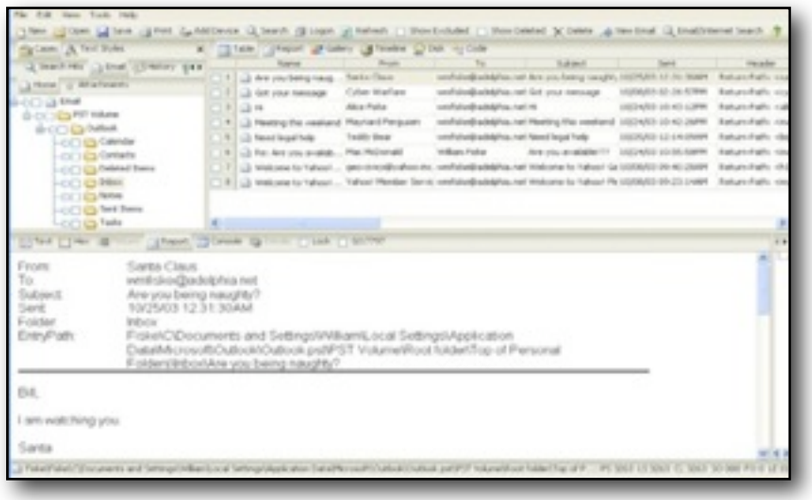
Students: 1500

- 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

# 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.*

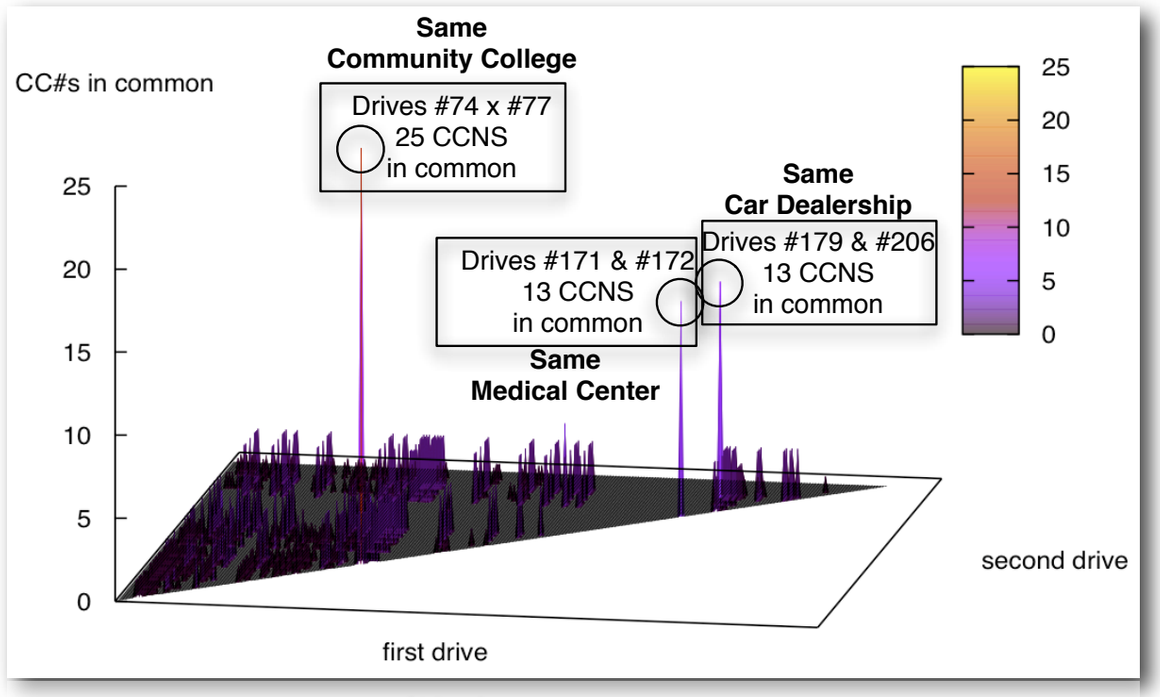
## Evidence Reconstruction

- 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.)

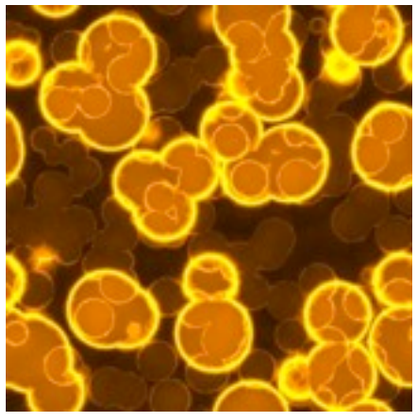
## ***We can't do this today.***

- 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.*

# Physical scientists understand this problem.

## Biologists:

- Trade cell lines
- Apprentice in labs to master techniques.



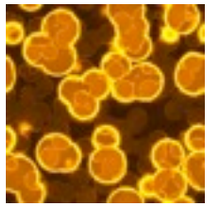
## Physicists and Chemists:

- Trade physical samples.
- Establish “scientific standards” for calibrating machines.

## What do we do for digital forensics?

# Outline for this talk

## The Need for Forensic Corpora



## Forensic Reproducibility

- What it is
- Why we need it

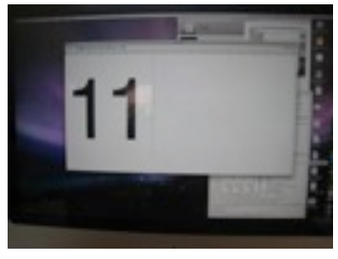


## Corpus Characterization

- How do we describe a “corpus?”

## Available Corpora

- What we are ***giving away!***




## Lessons Learned

# Forensic Reproducibility

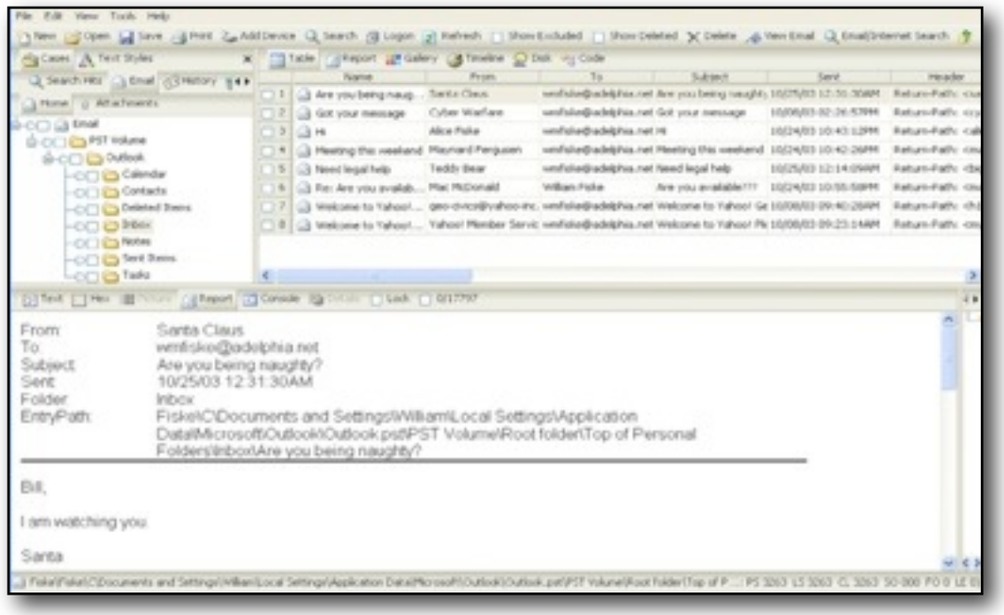
# Reproducibility and Accuracy in digital forensics practice: Most work has focused on Preservation & Presentation.

## Evidence Preservation

- Chain of custody
- File Formats
- Write Blockers


## Data Presentation

- File extraction
- Keyword Search
- Timeline Presentation




## "Accuracy" means:

- Not corrupting evidence
- Pointing to specific sectors were evidence is found.

# Reproducibility and Accuracy in digital forensics research: Largely Absent.

Reproducibility: ***Same Data + Same Experiments = Same Results***

But forensics works with data that is personal and private.

- Cameras
- Hard drives
- Cell phones
- Memory Sticks

Working with real data requires getting the data:

- Difficulty of acquisition
- Institutional Review Board (IRB) issues — 45 CFR 46 — for federally funded research.

So many researchers work with their own data.

- This data can't be shared.

# 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

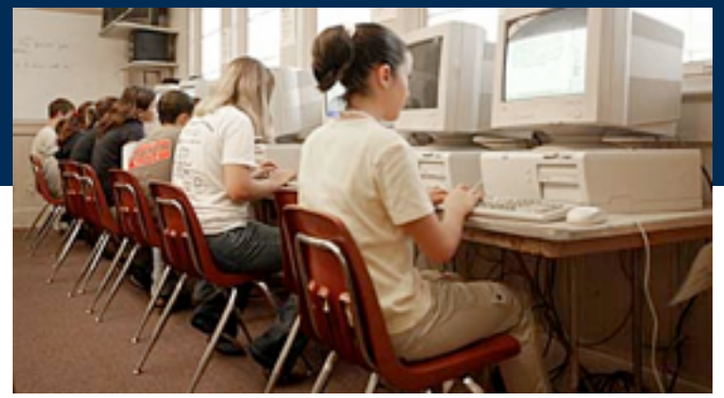
But everybody used a different data set!

- 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

United States Department of Health & Human Services

# 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 Characterization

# 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

- Wiretap laws makes collection very problematic.

## 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.

# 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.

# Description & Distribution of Corpora.

Currently, all corpora are distributed:

- From different locations
- In different formats
- With different metadata

How should researchers find corpora?

How should it be distributed?

What is the appropriate resolution for metadata?

- Per distribution?
- Per object?
- Per sector?

# Available Corpora

# 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)

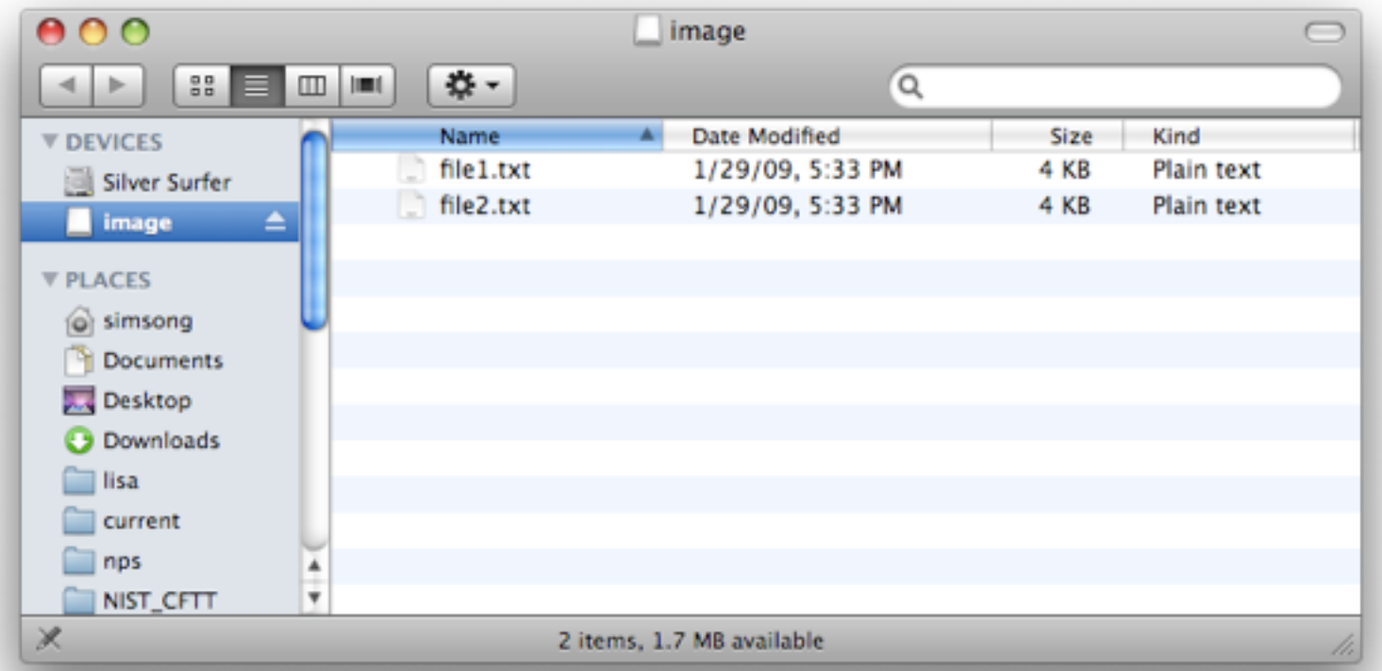
## Each image has:

- 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

# TEST IMAGE 1: nps-2009-hfsjtest1 (HFS+)
# For recovering data from the HFS+ journal

This is a very simple image:

- Two files — file1.txt and file2.txt:




File1 has had two sets of contents

- "This is file 1 - snarf"
- "New file 1 contents - snarf"
  - —*Both "snarf"s are still on the disk.*

Use the HFS Journal to find the data.

# TEST IMAGE 2: nps-2009-ntfs1
# For work on compressed and encrypted file systems.

```
   277228 Dec 31 18:27 ntfs1-gen0.aff
  8481452 Dec 31 18:27 ntfs1-gen1.aff
 35551648 Jan  6 16:27 ntfs1-gen2.aff
```

Three directories:

```
d/d 28-144-8:        Compressed
d/d 29-144-10:       Encrypted
d/d 27-144-7:        RAW
```

EFS key information:

```
r/r 46-128-1:        EFS-key-info.txt
r/r 43-128-4:        EFS-key-no-password.pfx
r/r 45-128-4:        EFS-key-password-strong-protection.pfx
r/r 44-128-4:        EFS-key-password.pfx
```

The same files are in each directory:

```
r/r 42-128-1:        RAW/20076517123273.pdf
r/r 47-128-0:        RAW/logfile1.txt
r/r 36-128-1:        RAW/NISTSP800-88_rev1.pdf
r/r 33-128-1:        RAW/NIST_logo.jpg
r/r 39-128-1:        RAW/report02-3.pdf
```

Open source forensic tools need to support EFS.

# The “logfile.txt” file was written a line-at-a-time to each directory and is fragmented…

```
<fileobject>
<filename>RAW/logfile1.txt</filename>
<filesize>21888890</filesize>
<partition>1</partition>
<ALLOC>1</ALLOC>
<USED>1</USED>
<mtime>1231192883</mtime>
<ctime>1231192883</ctime>
<atime>1231192883</atime>
<crtime>1231192820</crtime>
<libmagic>ASCII text, with CRLF line terminators</libmagic>
<byte_runs type='resident'>
    <run fs_offset='237428736' img_offset='237428736'
         file_offset='0' len='1024'/>
    <run fs_offset='243657728' img_offset='243657728'
         file_offset='1024' len='3072'/>
    <run fs_offset='240057344' img_offset='240057344'
         file_offset='4096' len='5120'/> …
```

1628 fragments in all!

This is a realistic model for the writing of log files.

# Realistic IMAGE 1: nps-2009-canon2 (FAT32)
# Six disk images from a digital camera

| size | filename | SHA1 |
|---|---|---|
| 31129600 | nps-2008-canon2-gen1.raw | 67364b0894a0465d6ada8c4966b6bbcaf7039082 |
| 31129600 | nps-2008-canon2-gen2.raw | 0e3cdef3b1a7d3762f9704bfd4349033fe808eda |
| 31129600 | nps-2008-canon2-gen3.raw | 7dc8be7f3993c37f101c0ed0fec4274abccacf3c |
| 31129600 | nps-2008-canon2-gen4.raw | ed1c7dea94096ad309b32037cb6d43a291952d8d |
| 31129600 | nps-2008-canon2-gen5.raw | 63e7f9daf8dbcd1744579e579f3f0fddebe2ee90 |
| 31129600 | nps-2008-canon2-gen6.raw | 4742c325f10583dab1eb4c55d0d45ab3beb99eb3 |

"Disk" is a 32MB SD card shot in a Canon camera.

*All operations carried out by camera*:

- Disk formatting (-o51 !)
- JPEG creation
- JPEG deletion


Disk was repeatedly removed from camera & imaged.

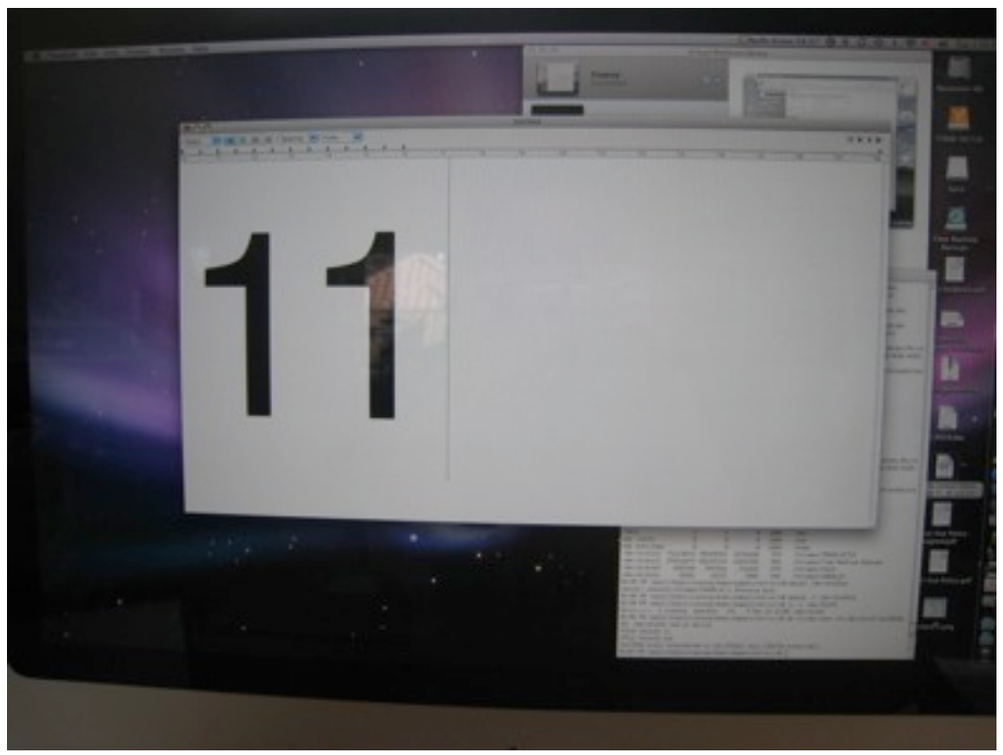
JPEGs were created and deleted in such a way to assure:

- Fragmented files
- Files that can only be recovered through file carving (name overwritten, not data.)

# nps-2009-canon2 is boring...

```
$ fls -rp -o51 nps-2008-canon2-gen6.raw
d/d 4:      DCIM
d/d 517:    DCIM/100CANON
r/r 1029:   DCIM/100CANON/IMG_0044.JPG
r/r 1030:   DCIM/100CANON/IMG_0042.JPG
r/r 1031:   DCIM/100CANON/IMG_0003.JPG
r/r 1032:   DCIM/100CANON/IMG_0043.JPG
r/r 1033:   DCIM/100CANON/IMG_0045.JPG
r/r 1034:   DCIM/100CANON/IMG_0046.JPG
r/r 1035:   DCIM/100CANON/IMG_0007.JPG
r/r 1036:   DCIM/100CANON/IMG_0047.JPG
r/r 1037:   DCIM/100CANON/IMG_0009.JPG
r/r 1038:   DCIM/100CANON/IMG_0038.JPG
r/r 1039:   DCIM/100CANON/IMG_0011.JPG
...
```

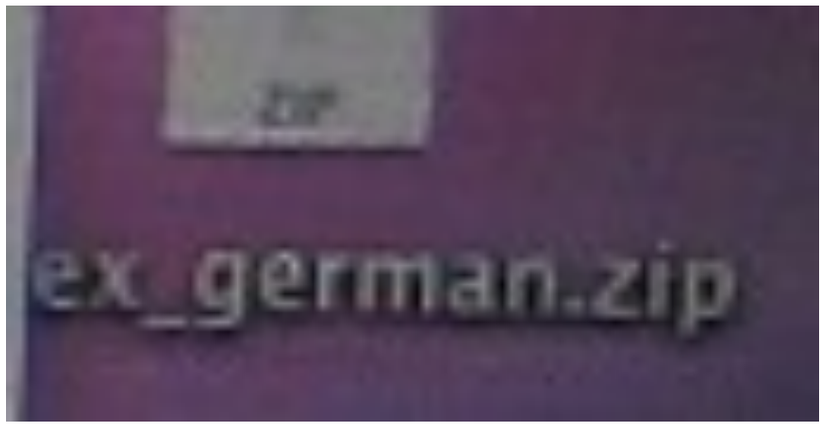
**IMG_0011.JPG:**

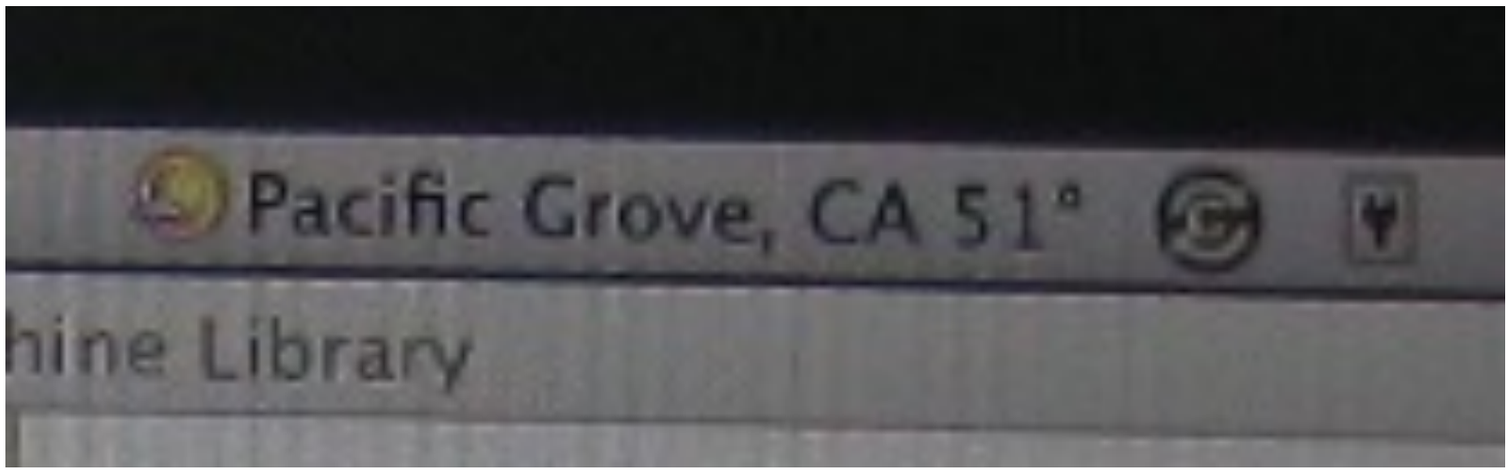


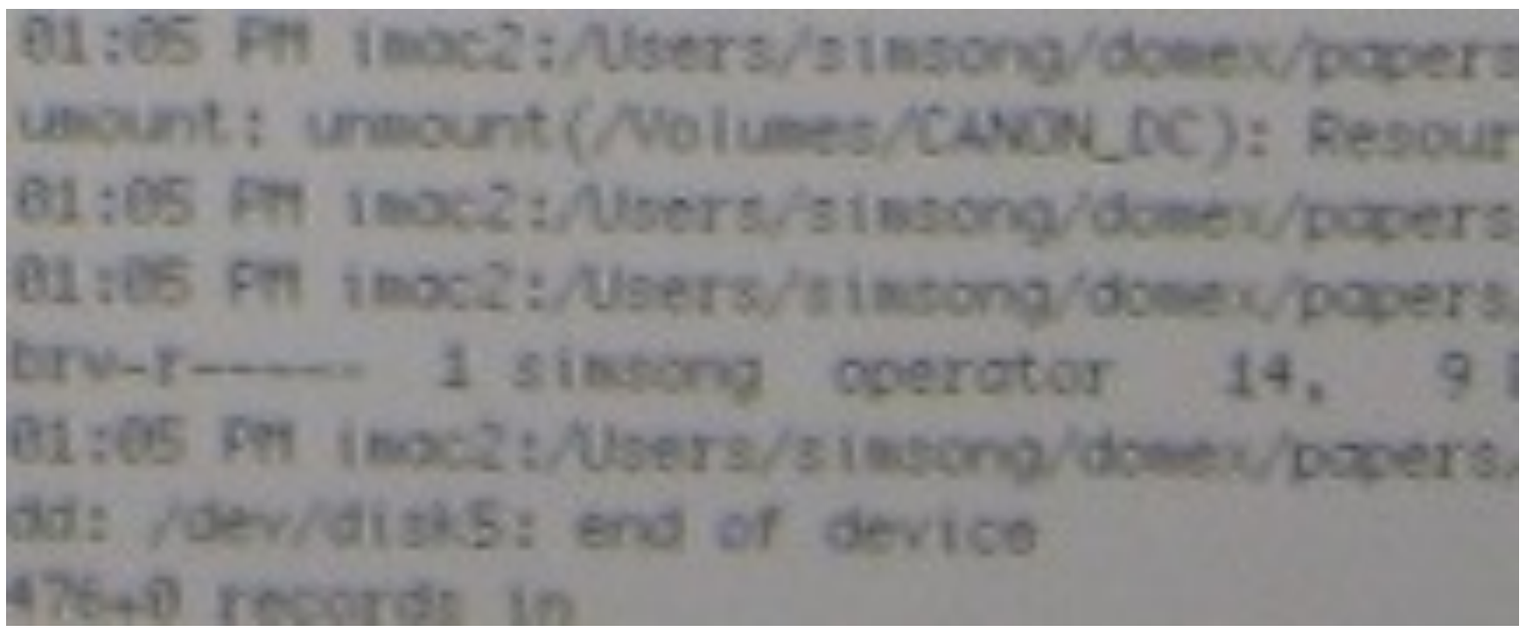
# ... perhaps not so boring

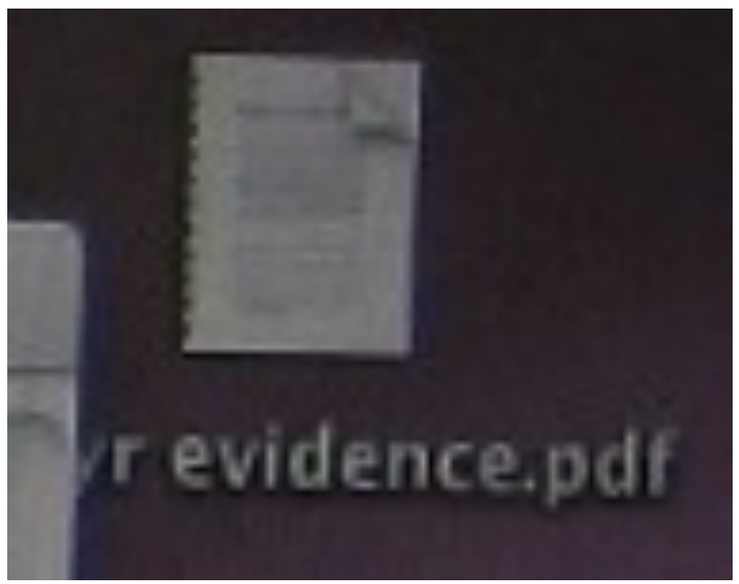
It's *hard* to avoid placing information in images...

... fortunately nothing here is a problem.



# Realistic Image 2: nps-2009-ubnist1 (FAT32)
# A bootable USB flash drive running Ubuntu Linux

The “outer” file system is FAT32:

```
$ fls -o63 ubnist1-2009-01-07.aff
r/r 4:      ldlinux.sys
d/d 6:      casper
d/d 8:      dists
d/d 10:     install
r/r 12:     syslinux.cfg
d/d 14:     pics
d/d 16:     pool
d/d 18:     preseed
d/d 20:     .disk
r/r 22:     autorun.inf
r/r 24:     md5sum.txt
r/r 27:     README.diskdefines
r/r 29:     umenu.exe
r/r 31:     wubi.exe
d/d 33:     syslinux
r/r 35:     casper-rw
```

# Realistic Image 3: 2009-nps-casper-rw (ext3)

Inside UBNIST1 is an ext3 file system:

```
$ icat -o63 ubnist1.gen2.aff 35 > ubnist1.gen2.casper-rw
$ fls ubnist1.gen2.casper-rw
d/d 11:     lost+found
r/r 12:     .wh..wh.aufs
d/d 7681:   .wh..wh.plnk
d/d 23041:  .wh..wh..tmp
d/d 7682:   rofs
d/d 23042:  etc
d/d 23044:  cdrom
d/d 7683:   var
d/d 15361:  home
d/d 30721:  tmp
d/d 30722:  lib
d/d 15377:  usr
d/d 7712:   sbin
d/d 13:     root
r/r * 20(realloc):    .aufs.xino
d/d 38401:  $OrphanFiles
$
```

# Realistic Image 4: nps-2009-domexusers1 (NTFS)

Image created by LT Paul Farrell as part of his master's thesis.

```
$ mmls realistic.aff
DOS Partition Table
Offset Sector: 0
Units are in 512-byte sectors

     Slot    Start        End          Length       Description
00:  Meta    0000000000   0000000000   0000000001   Primary Table (#0)
01:  -----   0000000000   0000000062   0000000063   Unallocated
02:  00:00   0000000063   0083859299   0083859237   NTFS (0x07)
03:  -----   0083859300   0083886079   0000026780   Unallocated
$
```

Contents: NTFS Windows XP installation.

Two users:

- domex1
- domex2

Email, Chat, limited web browsing.

# nps-2009-domexusers1 is *filled* with Microsoft binaries!

We are releasing this image in two ways:

- Encrypted AFF, for organizations that have an MSDN developer license.
- Redacted, with all of the Microsoft binaries "broken"

For information on the redaction approach, please see:

- Garfinkel, Simson., Automating Disk Forensic Processing with SleuthKit, XML and Python, Systematic Approaches to Digital Forensics Engineering (IEEE/SADFE 2009), Oakland, California

```
$ fls -o63 realistic.aff  3524-144-6
d/d 10219-144-6:      Administrator
d/d 3526-144-6:       All Users
d/d 3525-144-7:       Default User
d/d 27708-144-5:      domex1
d/d 28463-144-5:      domex2
d/d 10146-144-6:      LocalService
d/d 3370-144-6:       NetworkService
$
```

# 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 |
| | 1694 | 643 | 98 | 6748 |

# Lessons Learned

# It’s very hard to create, curate, and distribute corpora!

## Big Lesson: Disks are big, but data transfers slowly.

- All copies need to be verified
- When a job takes 3 hours, it’s easy to get distracted.

## Disk Images:

- It’s best to have one file per disk image.
- Never reuse the same file name.
- Use consistent path names on multiple systems.
- It’s really hard to keep **real data** out of **realistic** data sets.
- **client-rm.py**:  files are only deleted when they are safely elsewhere

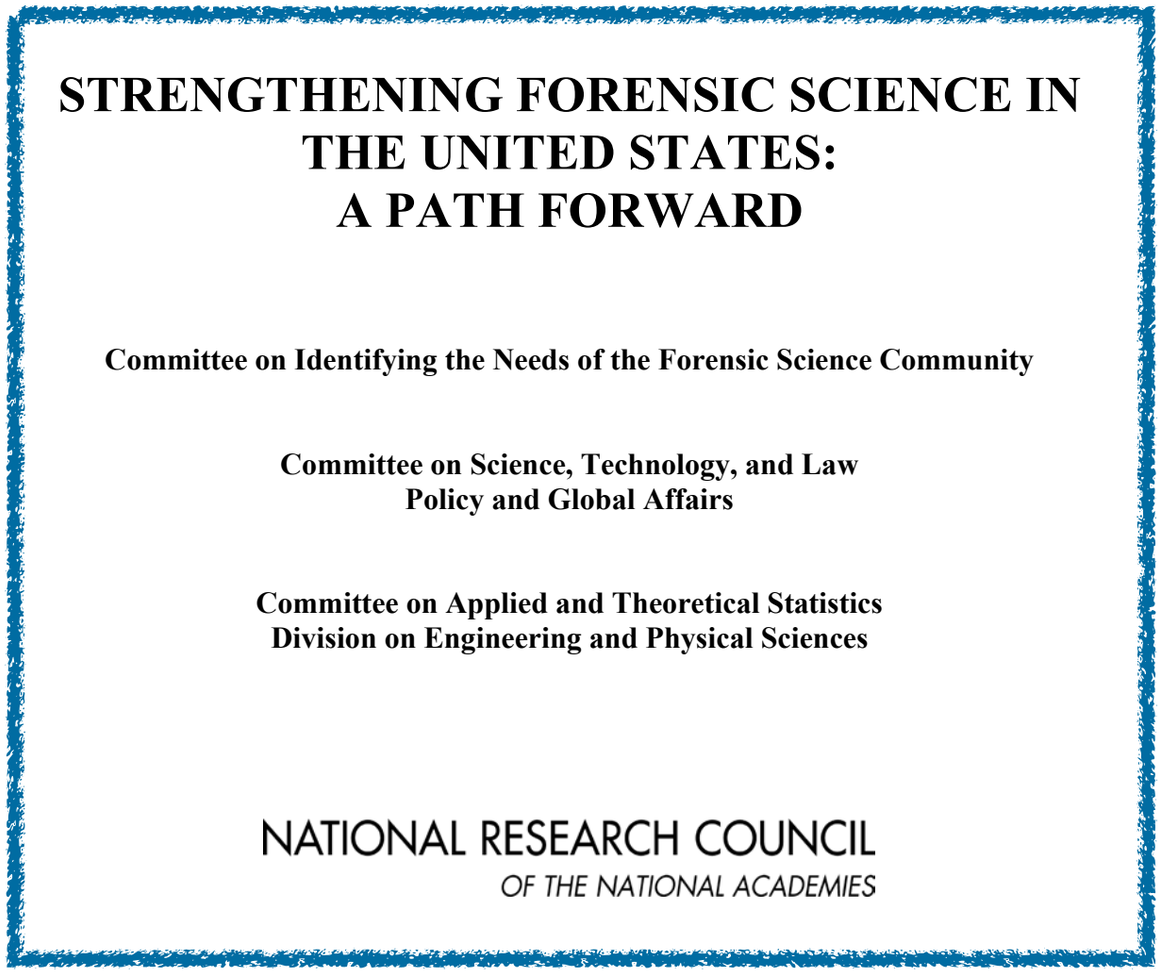
## Govdocs: web servers lie

- You ask for one file, they download another.
- **404 Error** is silently transformed to **200 OK**
- **.gov** contains a *lot* of domains that are not USG

# Conclusion: digital forensics needs digital corpora!

- "Substantive information and testimony based on faulty forensic science analysis may have contributed to wrongful convictions of innocent people...
- "Moreover, imprecise or exaggerated expert testimony has sometimes contributed to the admission of erroneous or misleading evidence."

  —*National Research Council, 2009*




You can download these corpora:

- http://digitalcorpora.org/

Questions?