Lessons learned managing a 60TB digital evidence corpus and writing digital forensics tools.

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School*

June 17, 2013

http://simson.net/

http://digitalcorpora.org/

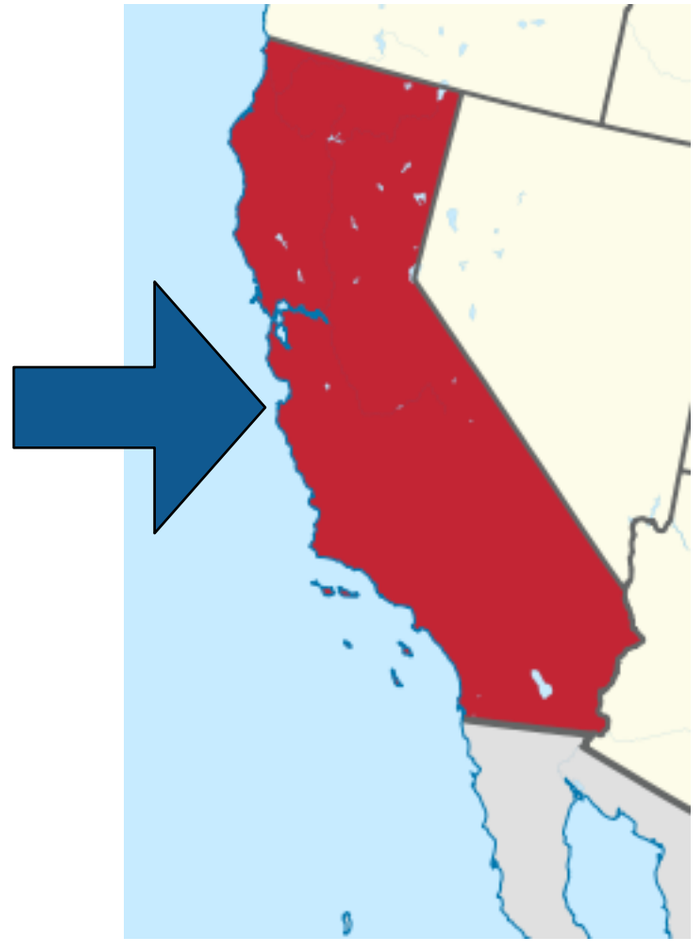
*** For purpose of identification only.**

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

NPS is the Navy's Research University.

Monterey, CA — 1500 students

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



Graduate Schools of Operational & Information Sciences (GSOIS)

- Computer Science
- Defense Analysis
- Information Sciences
- Operations Research
- Cyber Academic Group

National Capital Region (NCR) Office

- 900 N Glebe (Ballston)/Virginia Tech building

Digital information is pervasive in today's society.

There are many sources of digital information:

- Traditional Systems: Desktops, Laptops
- "Mobile:" Tablets, Cell Phones, embedded systems
- Internet-Based Services (servers)

Government has many possible uses for this information:

- Establish possession of contraband information (child pornography, credit card #s)
- Recover stolen information
- Document a conspiracy (stock fraud; murder-for-hire)
- Investigation, intelligence & analysis

Digital forensics makes this evidence available.
The US Government employs a “digital forensics model.”

Preparation: policy, training & tools

Collect & preserve evidence

Extract preserved data

Reporting & Testimony

Analysis

Digital forensics makes this evidence available.
The US Government employs a “digital forensics model.”

Preparation: policy, training & tools

Collect & preserve evidence

Extract preserved data

Archive

Analysis

Reporting & Testimony

Most DF efforts focus on training, collection and extraction.

Preparation: policy, training & tools

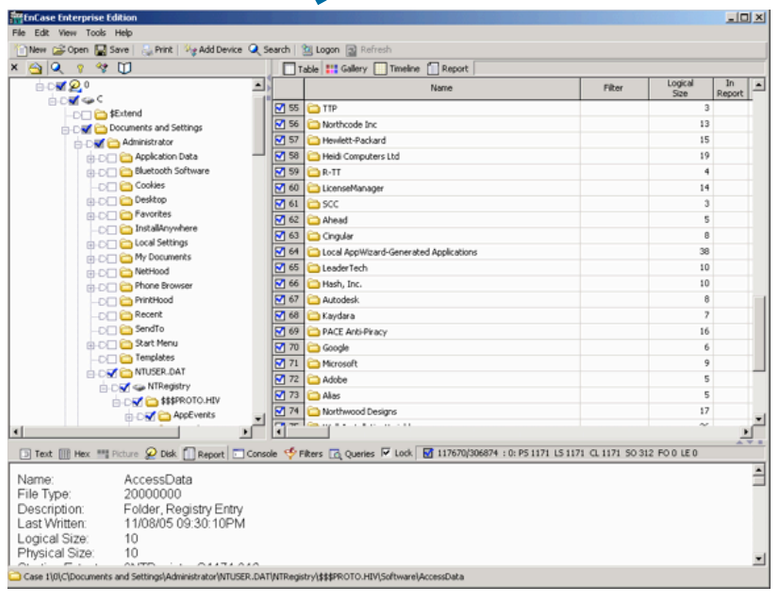
Collect & preserve evidence

Extract preserved data

Training the force

CyberCorps®
Defending America's Cyberspace

Collection Techniques



Data visibility

My team develops better approaches for automation and analysis — “big data for little devices”

Identify high-value data (automatically).

—*Contacts, calendar, GPS, documents*

Correlate devices with identical or similar data.

- Previously unknown organizations or connections

Present

- Make the results understandable

Translate with human language technology (HLT)

- Apply to العربية ,עִבְרִית, español, 汉语/漢語, 日本語, svenska, etc.

Archive and Manage our holdings

- Make use of institutional knowledge.

Three principles underly this research

1. Work with “big data”

—*Scale is our advantage — use it!*

—*Many techniques developed on small data sets do not scale*

—*We discover important techniques by working with a big corpus!*

Three principles underly this research

1. Work with “big data”
2. **Automation is essential.**
   - —*Today most forensics is done manually — this doesn’t scale.*
   - —*We develop techniques & tools for automation.*

Navy Cyber Defense Operations

Three principles underly this research

1. Work with “big data”
2. Automation is essential.
3. **Concentrate on *bulk data*.**
   - Leverage data that are fragmented and incomplete
     - —*Deleted and partially overwritten files*
     - —*Fragments of memory in swap & hibernation*
     - —*Tool marks*
   - These techniques can be applied to files

MISSING JPEG TOP

MISSING JPEG BOTTOM

Sencar and Memon [2009] DFRWS

Digital forensics (DF) ***research*** is different from from both traditional computer science and DF practice.

DF research is different!

- DF research is **hard and expensive** — but it looks easy!
- DF is **practitioner driven** — most DF developers are not professional programmers
- Most **practitioners are not computer scientists** — don't understand data
- DF has **immediate needs of great significance** (kidnapping, defense, etc.)

Managing a research corpus is harder than managing case data

- Practitioners will analyze each data *once*
- Researchers repeatedly re-analyze data as tools improve
- Sharing 60TB of data is hard — and sharing is part of research

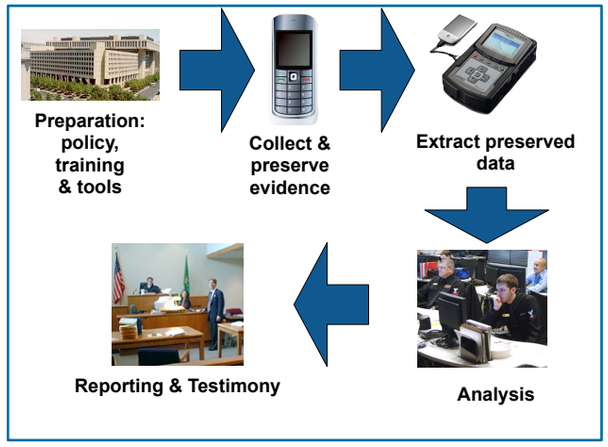
DF software development is challenging

- Tools must constantly improve and evolve
- Research tools must be "product quality"
- Researchers don't get exemplars when tools crash

"Lessons Learned Writing Digital Forensics Tools"
DFRWS 2012

This talk presents “lessons learned” while managing a significant corpus and writing DF tools.

Quick intro to Digital Forensics ✔



The Real Data Corpus

Lessons learned writing DF tools

- DFXML — a language for describing DF products

<dfxml>

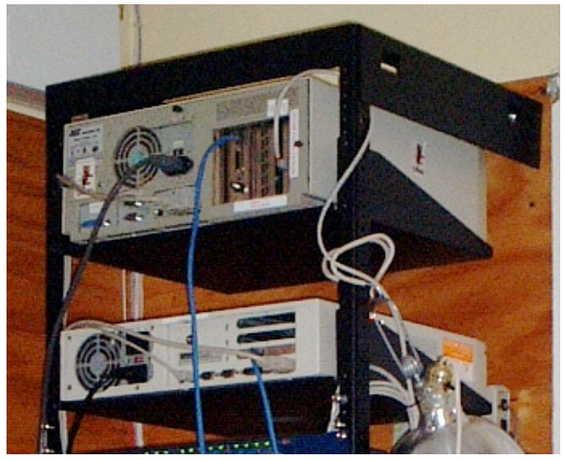
Real Data Corpus, Harvard University, 2006
≈ 800 drives

Lessons learned managing a research corpus

In the 1990s I started buying used computer equipment.

My *first* purchase of six used computers yielded:

- A law firm (client-attorney documents)
- A mental health practitioner (patent records)
- A single, divorced mother (child support disputes)
- A novelist (unfinished manuscript)
- + 2 more!



A purchase of 10 hard drives from “Weird Stuff”, Sunnyvale, CA:

- Financial records
- HR records
- Email
- Source code

A lot of confidential information was being inadvertently released!

In 2002 I started the PhD program at MIT CSAIL

I only had three years of funding

There is a thriving used HD market

- Re-used within an organization
- Given to charities
- Sold on eBay

I used forensics for “usable security”

- Deletion patterns can show “intent”

My goal — show how usability failures became security failures

All Categories Save this search

350 items found for **hard drives**

Sort by items: **ending first** | newly listed | lowest priced | highest priced

Picture hide	Item Title	Price	Bids	Time Left
	Lot of hard and floppy drives	$5.50	2	14n
	Lot of hard and floppy drives	$5.50	2	22n
	Lot of hard and floppy drives	$5.50	2	25n
	Lot of 2 hard drives IDE	$8.00	12	29n
	3.2 gig Hard Drives	$180.00	-	59n
	(5) 1.2 hard drives & (15) 10/100 network	$25.00	1	1h 00n
	Lot of 3 Quantum 9.1 gig SCSI Hard Drives	$26.00	6	1h 25n
	IDE HARD DRIVES (3)	$6.50	6	1h 46n
	LOT OF 5 Hard Drives! 3.2 Gig Western Digital	$120.00 $124.95 Buy It Now	-	1h 50n
	QTY 3....IDE Hard Drives 2.5 Gig	$20.50	5	2h 02n
	5 WESTERN DIGITAL 2.5 GIG HARD DRIVES	$30.00	4	2h 03n
	QTY 3....IDE Hard Drives 1.0 Gig	$9.99	1	2h 04n
	Western Digital 850 meg IDE Hard Drives dutch	$6.00	1	2h 57n
	WINDOWS	$6.00	-	3h 18n

I bought 250+ drives the first year

I stored “disk image” as GZIP-compressed raw files on a single server.

Drives to image

4 ATA drives being imaged

640GB RAID Array (2003)

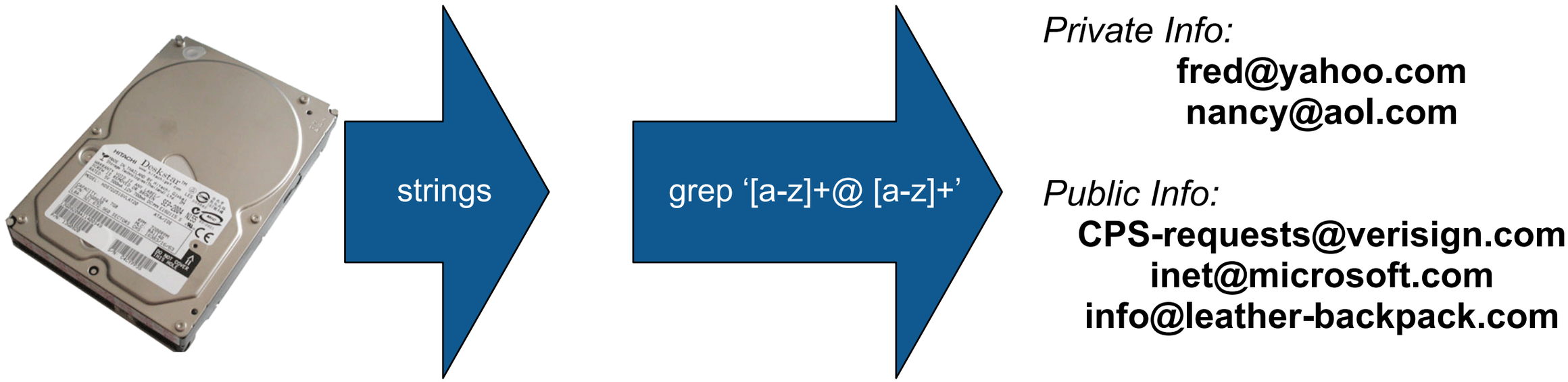
Critical technology for 2006

- Handling read errors from subject drives.
- Automated metadata collection of subject drive ***and imaging computer*** (aimage).

The *presence* of private information showed sanitization failures. It's condition showed *user intent*.

I scanned for "private information" with **strings(1)** and **grep(1)**

- Example: email addresses ; credit card numbers; SSNs; DOBs



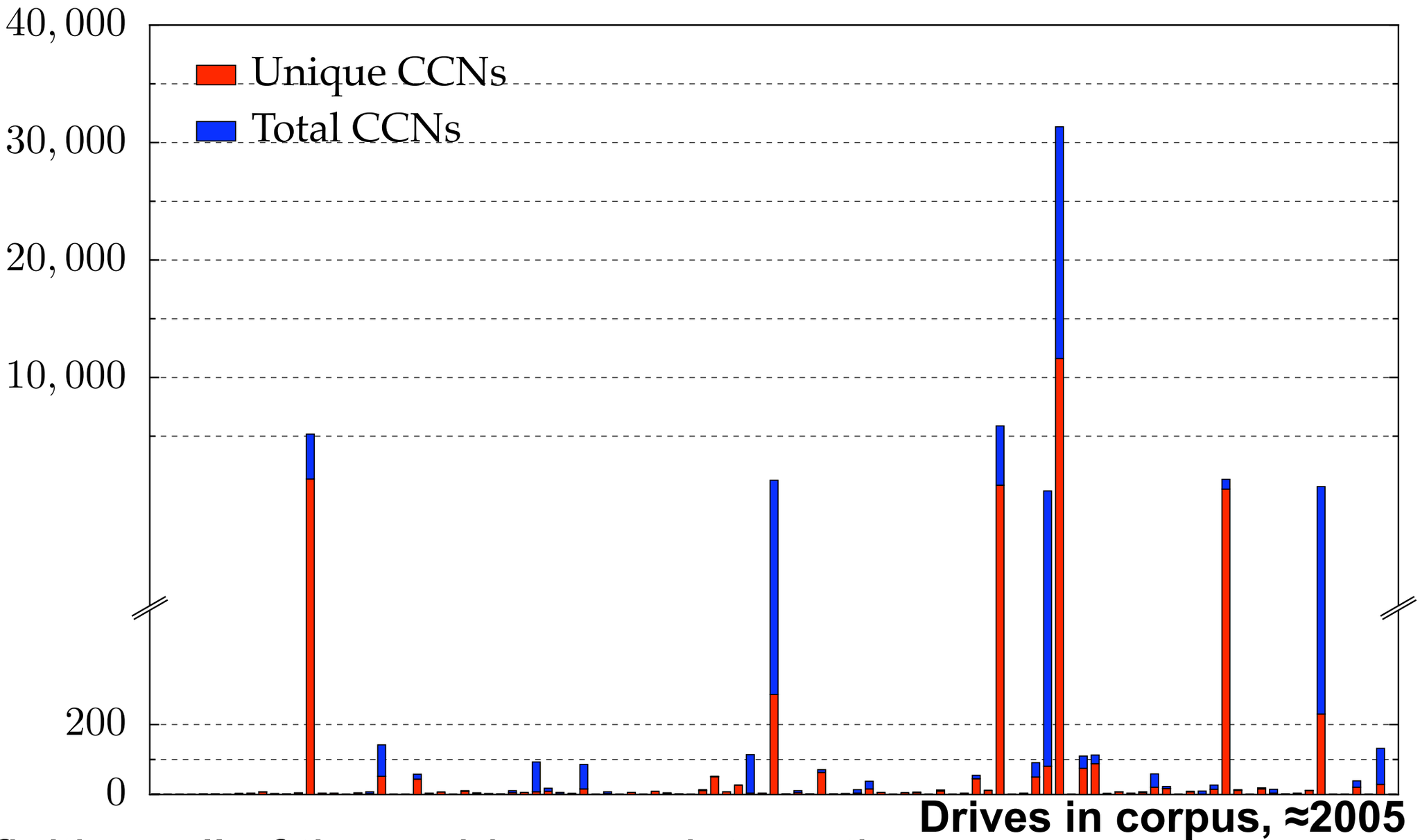
Entire corpus re-processed each time tools improved

- Law enforcement — typically processes each drive *once*

Manually examined drives that had lots of email addresses

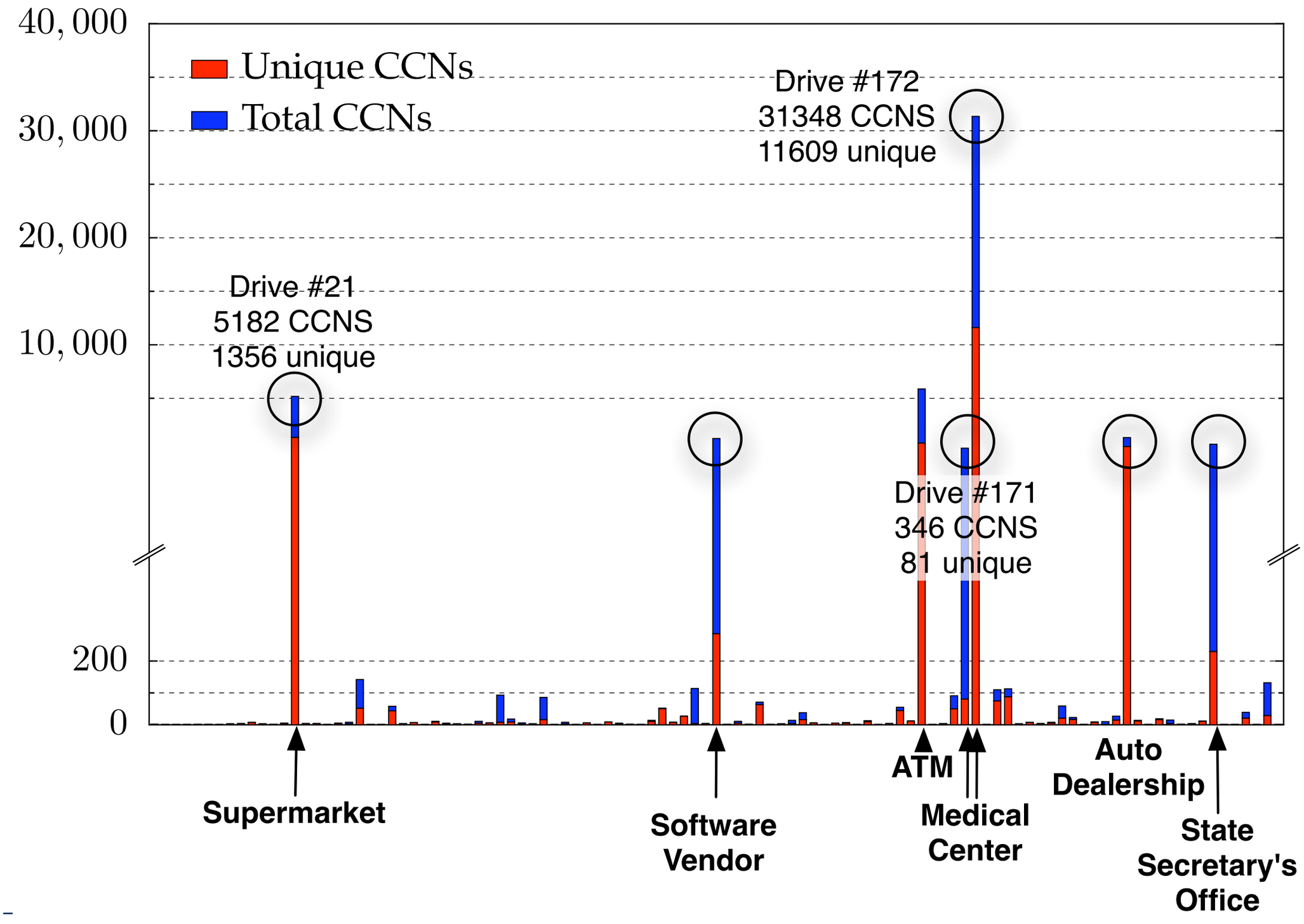
Example: CCNs on hard drive images.

No drives should have a lot of CCNs.

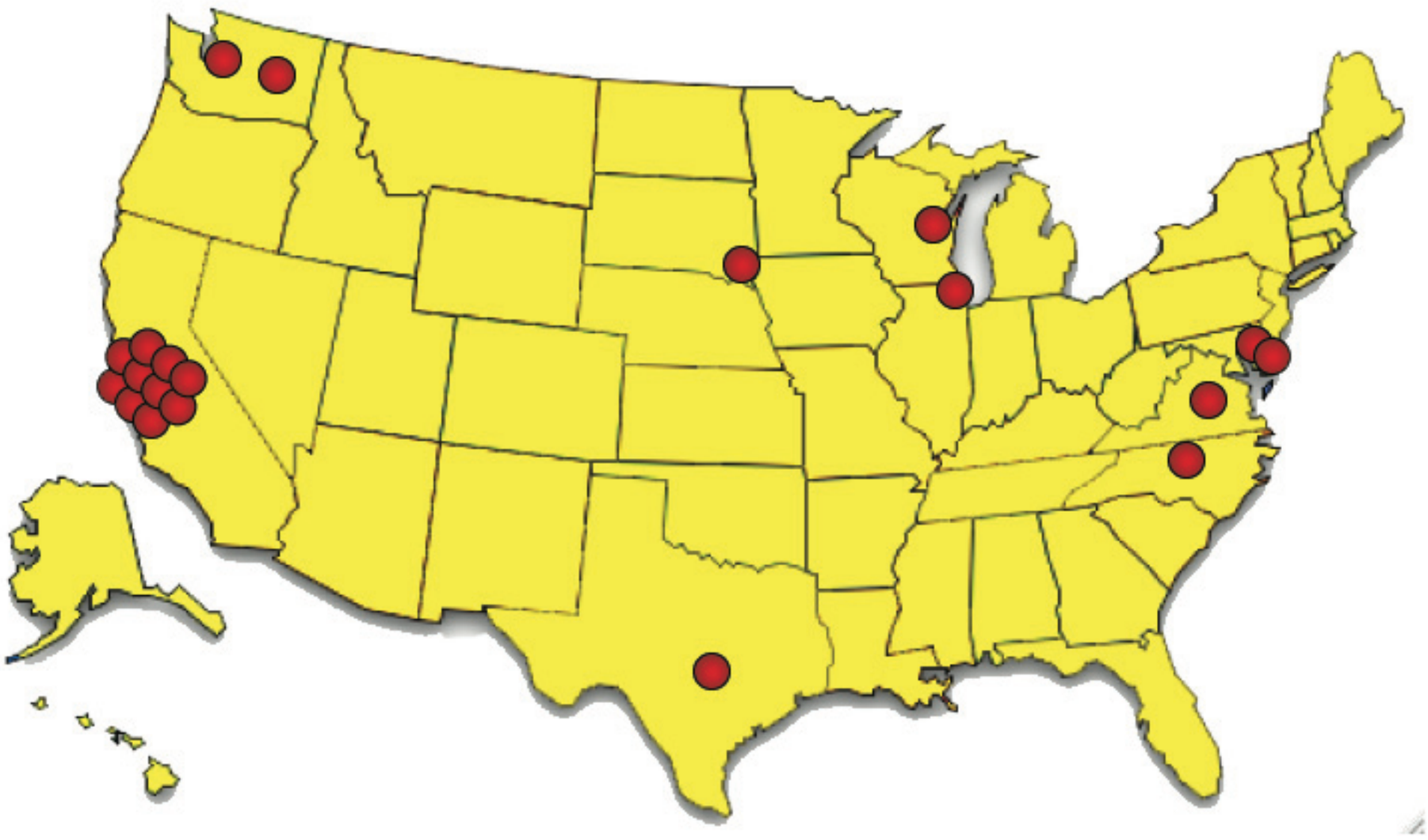


By definition, all of these drives are interesting.

“Interesting” drives were targeted for additional analysis.



I traced 20 drives back to their former owners.



While tracing back the drives, I discovered cross-drive analysis.

We documented a significant problem with the secondary market.

- Garfinkel, S. and Shelat, A., "Remembrance of Data Passed: A Study of Disk Sanitization Practices," IEEE Security & Privacy, January/ February 2003.

Some of reactions to this research were confused.

“Good luck [recovering] data from this.”

“Our prognosis: drive slagging is a fool-proof method to prevent data recovery.”

- It’s *easy* to remove data from a hard drive
  —*You just have to do it!*

After the initial findings, this research moved in 4 directions

2003-2005

Data leakage on used equipment

2005-

Corpus expansion

Real data as simulant

2005-2008

Advanced Forensic Format

File formats for digital evidence

2007-

Digital Forensics XML

Systematic capture of metadata

TBD

Cross-Drive Analysis

2008-

bulk_extractor

Systematic recovery of encoded PII

Corpus management: Technical Issues

I imaged 250+ drives the first year

I stored images as raw.gz

- naturally led me to stream-based forensic

Lessons learned:

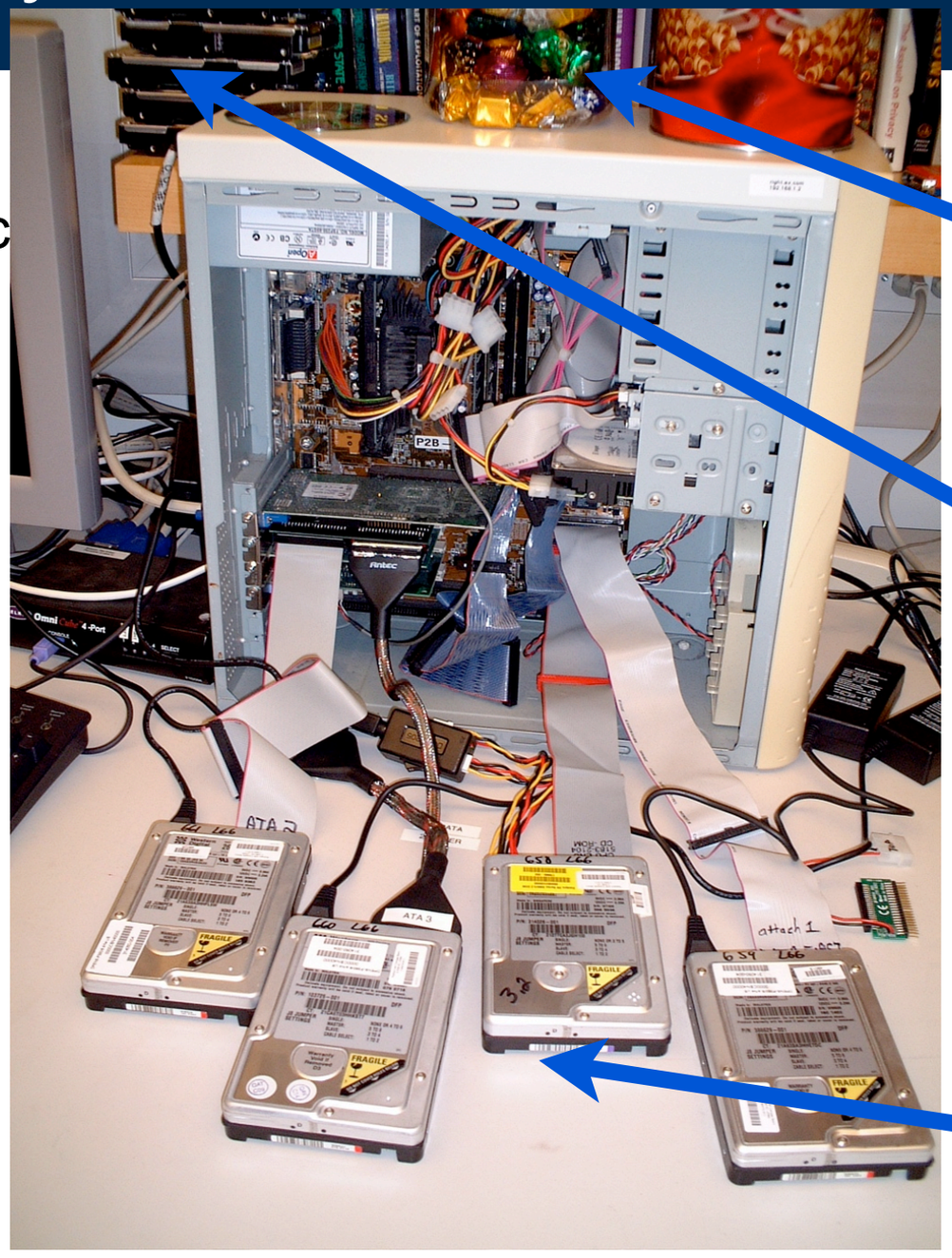
- ATA is hot-swappable

Lesson: read the documentation for the computer that you are using

- Don't maintain software (aimage) that does the same thing as other open source software (guyimager)

Critical technology:

- Handling read errors
- Automated metadata collection



Drive imaging workstation, Harvard University 2005

Lesson: Make the most of the tools that you have; follow technical innovations they force upon you.

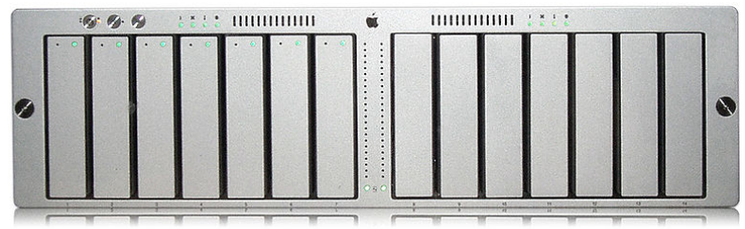
Using disk images for *research* required storing data *online* for easy access.

Law enforcement typically process each image *once,* then archives

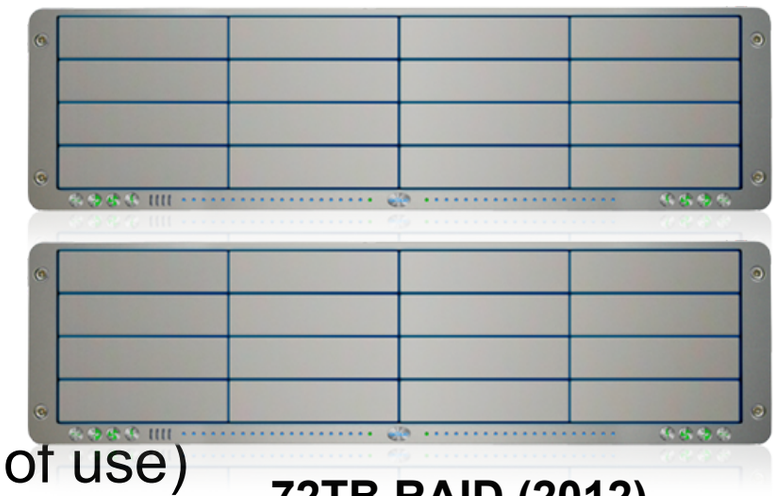
We store data online so we can reprocess.

640GB RAID Array (2003)

1.5TB RAID Array (2006)



6TB XServe RAID (2008)



72TB RAID (2012)

What works best:

- simplicity — a single file with all metadata embedded
- convenience — small file names with short paths (ease of use)
- permanence — file names and path names that wouldn't change

Automation is key; any process that involves manual record keeping is going to introduce inaccuracies that will be hard to detect and correct.

Useful data will outlive any storage system, so make provisions to move the data when you design the system.

Data storage formats: better is no assurance of success.

In 2005 I started on AFF (Advanced Forensics Format):

- Store metadata & data together; Extensible
- Read & Write, but optimized for archiving
- Advanced support:
  —*digital signatures, encryption, chain-of-custody*
- aimage imager that did “sparse imaging” and error recovery

2005-2008

Advanced Forensic Format

File formats for digital evidence

AFF4 addressed workflow, metadata, and efficiency issues

Since 2010 I have given up on AFF:

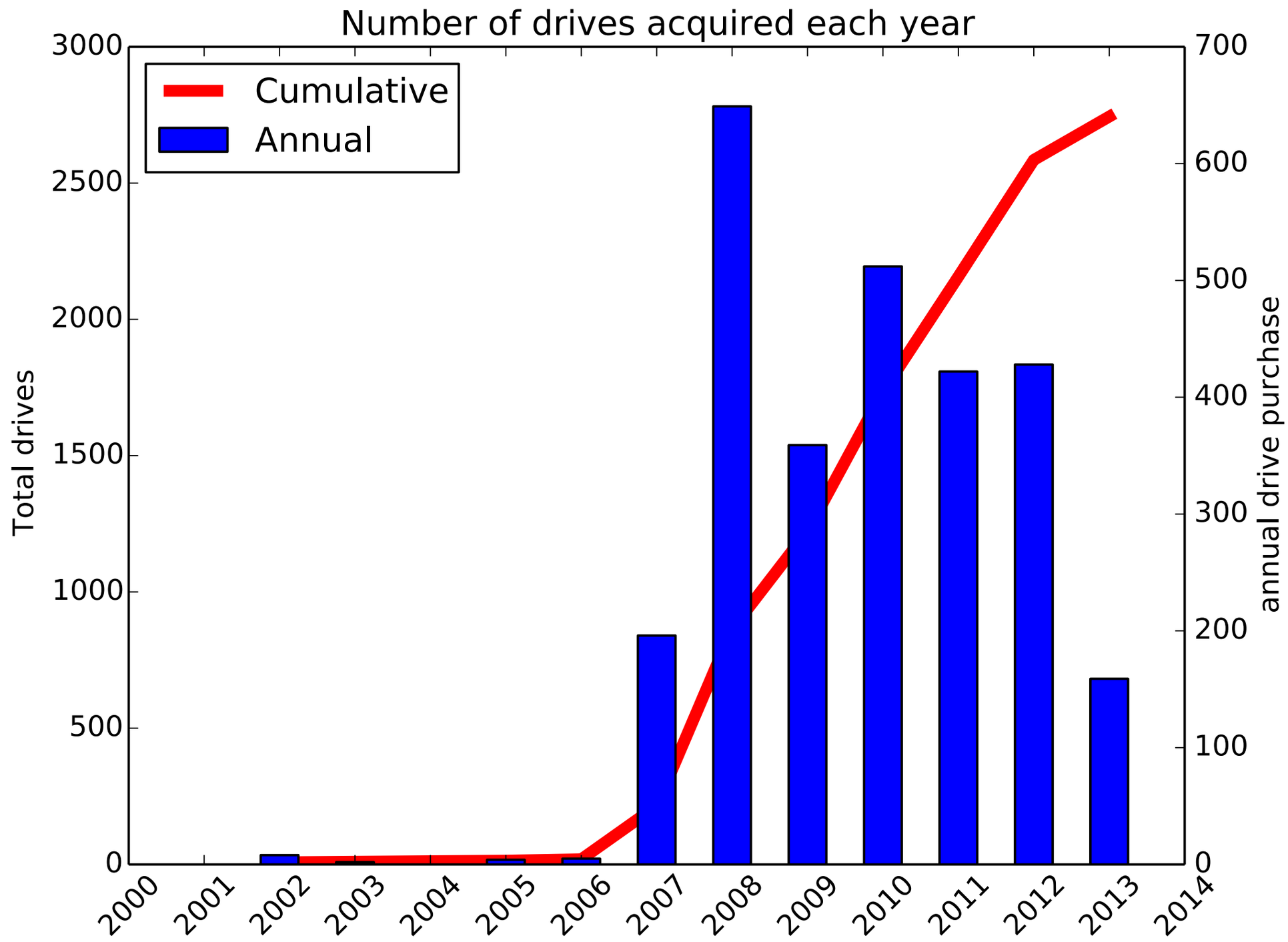
- E01 can now handle terabyte-sized HDs in a single file
- Joachim Metz’s ewfacquire & libewf do an excellent job supporting E01

Avoid replicating other people’s work

Avoid “Not Invented Here” research.

Avoid new file formats.

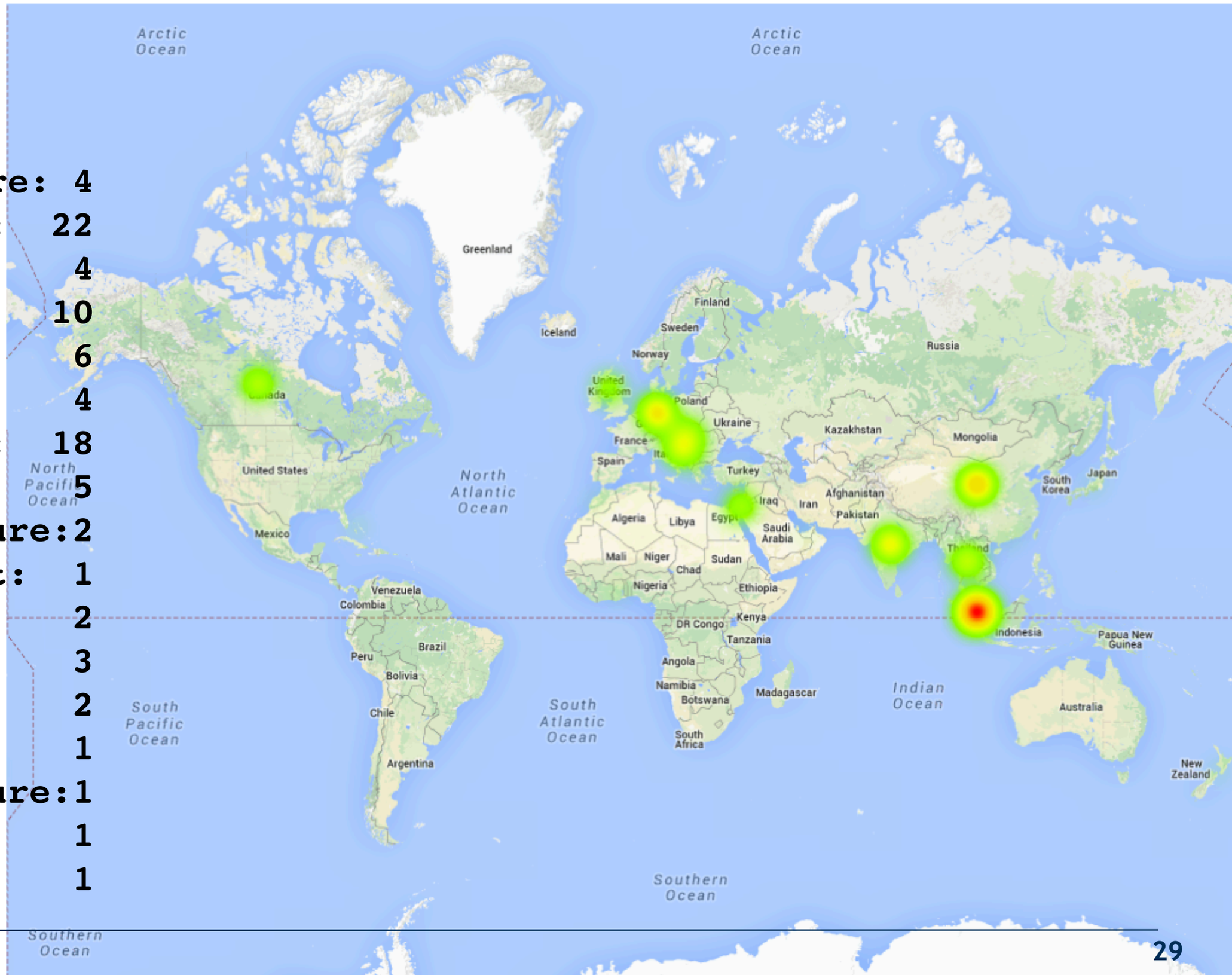
Since joining NPS the “non-US corpus” has grown substantially.



Drive corpus — good geographical diversity.

Today we have 100-200 mobile devices.
More diversity, but less representative of the market.

Samsung Feature:	4
Samsung Smart:	22
HTC Feature:	4
HTC Smart:	10
SONY Feature:	6
SONY Smart:	4
Nokia Feature:	18
Nokia Smart:	5
Motorola Feature:	2
Motorola Smart:	1
LG Feature:	2
LG Smart:	3
Dell Windows:	2
Sharp Smart:	1
T-Mobile Feature:	1
Huawei Smart:	1
O2 Windows:	1



Bad ATA drives crash Linux & FreeBSD

Crashes look like wild memory writes.

- ATA spec allows DMA to system memory
- Motherboards probably don't defend against wild DM

Question:

- Can we use this as a memory acquisition technique?
- There is "legacy ATA" on many motherboards.

Many technical options remain unexplored.

Many bad drives had sensitive data!

- Always read to the "end" of the drive
- Read all the drives in the RAID set

Drives with some bad sectors invariably have more sensitive information on them than drives that were in working condition when they were decommissioned.

Corpus Management — use descriptive path ames

Many different modalities:

- Disk images — “drives”
- Memory Images — “ram”
- Scenarios — symlinks to source images
- Packet Dumps — “net”
- Files — “files”

Many different sources and distribution restrictions:

- Used purchased inside US — “US” (not used by USG)
- Used purchased outside US — “NUS”
- Created by NPS, redistributable — “NPS”

Although it is advantageous to have names that contain no semantic content, it is easier to work with names that have some semantic meaning.

Consistent naming scheme on every machine:

```
/corp/source/modality/description/daughter-files

/corp/nist/rds/rds328/
/corp/nps/files/govdocs1m/123/123456.jpg
/corp/nus/drives/in/IN10-0249/IN10-0249.E01, IN10-0249.E01.txt
```

Names must be short enough to be usable but long enough to be distinct

Names should be consistent and usable.

/corp/source/modality/description/daughter-files

/corp/nist/rds/rds328/
/corp/nps/files/govdocs1m/123/123456.jpg
/corp/nus/drives/in/IN10-0249/IN10-0249.E01,
IN10-0249.E01.txt

Every data object should have a unique file name.

- Put something *very descriptive* in the file name
  - —*Source country*
  - —*Scenario name*
- Don't change file names.
  - —*If you must change names, try to have the old name inside the new name*
    ubnist1.E01 -> nps-2009-ubnist1.E01
- It's okay to change directories.

Place access-control information as near to the root of a path name as possible.

Different users want different subsets of the corpus.

- It's best if they use the same file hierarchy.

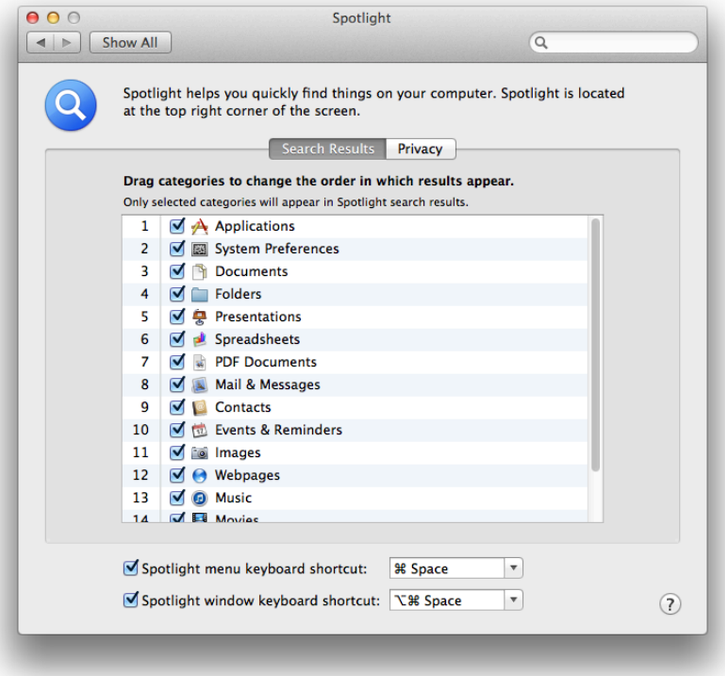
Anti-virus and indexing cause numerous problems

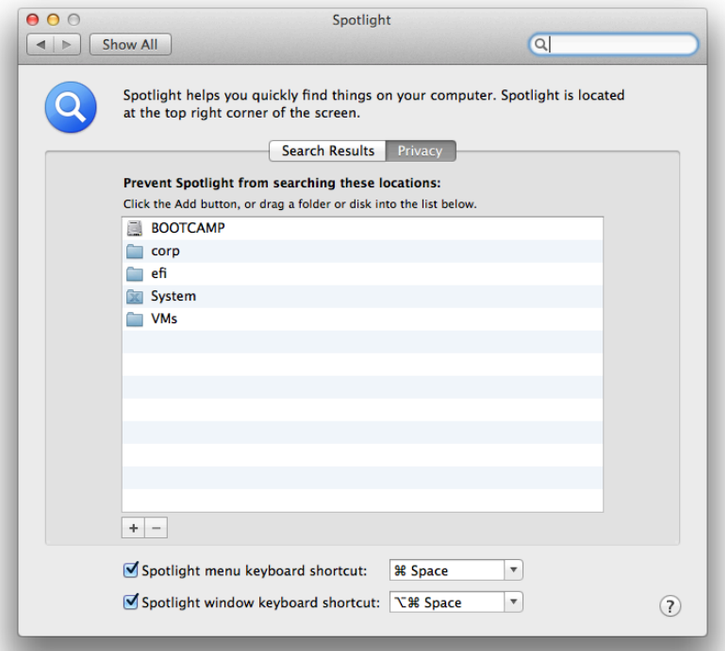
Disable AV and indexing on your corpus.

- Forensic data has viruses
- Corrupt and unstructured data frequently crash indexers

Exceptions need to be frequently reapplied:

- After software updates
- After OS upgrades
- When new external HDs are attached.





Configure anti-virus scanners and other indexing tools (e.g., Apple's build_hd_index) to ignore directories that might contain raw forensic data.

There is no good way to distribute a 60TB data set

Approaches we have tried:

- Transferring over the Internet by scp, rsync, BitTorrent, uftp, Aspera
- Sending 2TB internal SATA drives
  - —*Need SATA dock.*
  - —*File System Choice: ext2/3/4? HFS? NTFS?*
  - —*(NTFS seems best choice for read-only)*

Added complications — “bit rot” long term storage — off track writes

- —*Evaluating the Impact of Undetected Disk Errors in RAID Systems* *https://www.perform.csl.illinois.edu/Papers/USAN_papers/09ROZ01.pdf*
- —*Modeling the Fault Tolerance Consequences of Deduplication* *https://www.perform.csl.illinois.edu/Papers/USAN_papers/11ROZ02.pdf*

Solutions developed by other disciplines for distributing large files rarely work well when applied to DF without substantial reworking.

Corpus management — Policy issues

— Privacy

— Illegal content — financial, passwords and copyright

— Illegal content — pornography

— Institutional Review Boards (IRBs)

Even if something is legal, you may wish to think twice before you do it.

Privacy — What's legal isn't necessarily right.

Information in the RDC is not legally "private"

- "The reasonableness of a search for Fourth Amendment purposes ... turns upon the understanding of society as a whole that certain areas deserve the most scrupulous protection from government invasion. There is no such understanding with respect to garbage left for collection at the side of a public street."

  —*CALIFORNIA v. GREENWOOD, 486 U.S. 35 (1988)*

- In practice, we avoid disclosing PII because doing so ***would be wrong***.

Copyright on user-generated material — ***different from privacy!***

- Users do not transfer copyright to us, but we do have some rights in our copy
- **"First Sale" doctrine** — "In our view, the copyright statutes, while protecting the owner of the copyright in his right to multiply and sell his production, do not create the right to impose, by notice, such as is disclosed in this case, a limitation at which the book shall be sold at retail by future purchasers, with whom there is no privity of contract."

  —*Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908)*

- **"Fair Use"** — four part test. 1) purpose of the use (non-profit educational); 2) nature of the copyrighted work; 3) The amount of the work that is copied; 4) the impact of the use on the market value for the copyrighted work.

Illegal content — different kinds requires different rules.

"Counterfeit Access Device and Computer Fraud and Abuse Act"

- Passed by Congress in 1984
- Outlaws possession of "access devices" with intent to commit fraud.
- Financial information (credit card numbers) and passwords are access devices.
- The key issue is intent — I don't have the intent do defraud.

Copyright — rely on "Fair Use" (17 USC §107)

- Four part test. 1) purpose of the use (non-profit educational); 2) nature of the copyrighted work; 3) The amount of the work that is copied; 4) the impact of the use on the market value for the copyrighted work.
- The RDC doesn't impact the value of the data, and it's non-profit.

Conventional pornography

- The RDC has lots of pornography in it
- No access given to minors

Obscenity (e.g. child pornography)

- We can't determine if something is really child porn…
- Names "suggestive" of child pornography are removed.

Never sell access to the corpus.

Do not give minors access to real data.

Do not intentionally extract pornography from research corpora.

Institutional Review Boards — get to know the IRB.

In the United States, federally funded research involving human subjects *must* be reviewed by an accredited Institutional Review Board.

- "Human subject" means a living individual about whom an investigator conducting research obtains:
  - —*Data through intervention or interactions with the individual, or*
  - —*Identifiable private information*
- "Research" means "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge."

Options:

- Make the RDC public (so it's not private)
  - —*That would be unethical*
- Don't do "research"
- Get IRB approval (and that's what we do)

IRBs exist to protect human subjects, but many have expanded their role to protect institutions and experimenters.

This expanded role occasionally decreases the protection afforded human subjects.

Even with the IRB watching over you, it's important to watch your back.

Data normalization is critical — but very hard.

We have many different kinds of data:

- Drive images
- Files
- Mobile phones
- Software

We try to “normalize” by:

- Consistent containers
- Consistent file names & paths

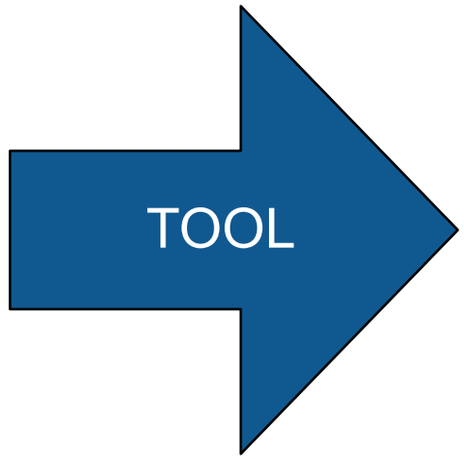
Metadata normalization with Digital Forensics XML (DFXML)...

<dfxml>

Lessons learned while writing digital forensic tools

DFXML

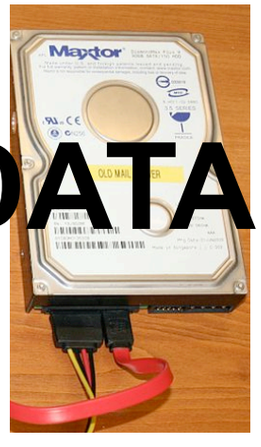
Most DF tools act like “filters” or “extractors.” They turn *corpus bulk data* into *actionable intelligence.*



It is not this simple in practice.

The same tool is applied to many data sources.

DATA1



TOOL

STUFF 1

DATA2

TOOL

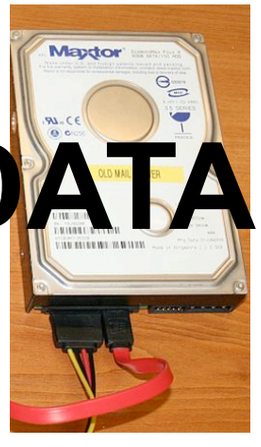
STUFF 2

DATA3

TOOL

STUFF 3

Different tools may be applied to the same data.



DATA1

TOOL1

STUFF 1-1

DATA1

TOOL2

STUFF 1-2

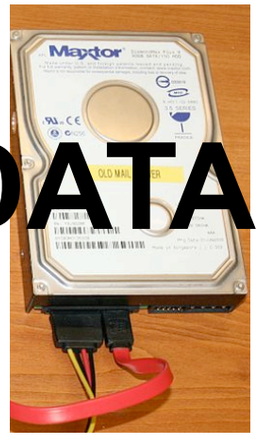
DATA1

TOOL3

STUFF 1-3

Tools improve over time

DATA1



TOOL1
(2006)

STUFF 1-1-2006

DATA1

TOOL1
(2008)

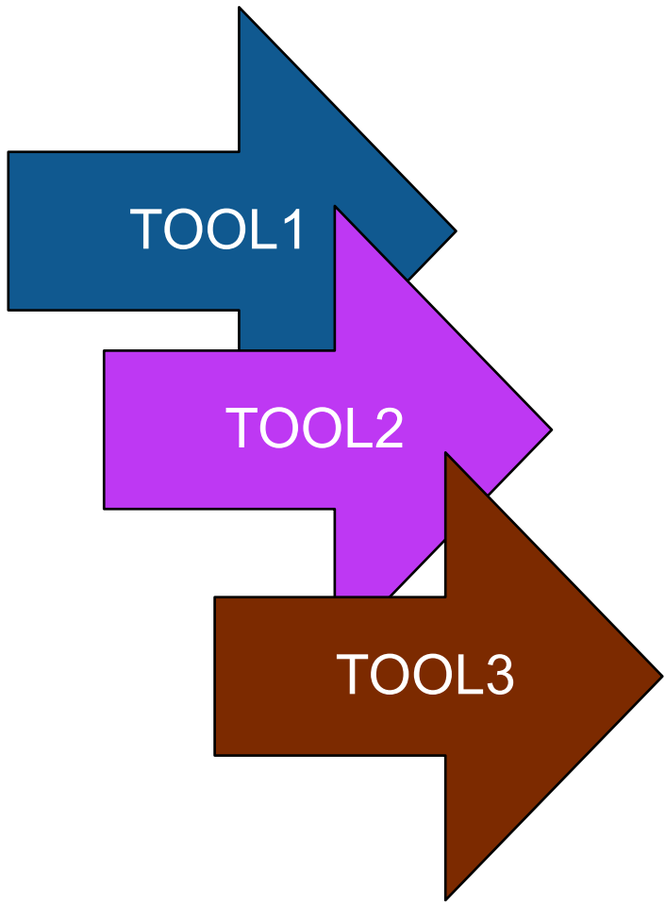
STUFF 1-1-2008

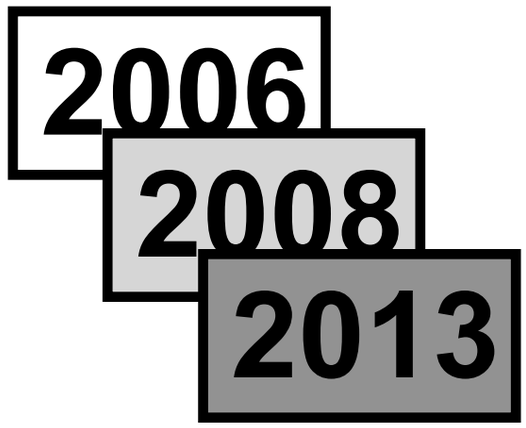
DATA1

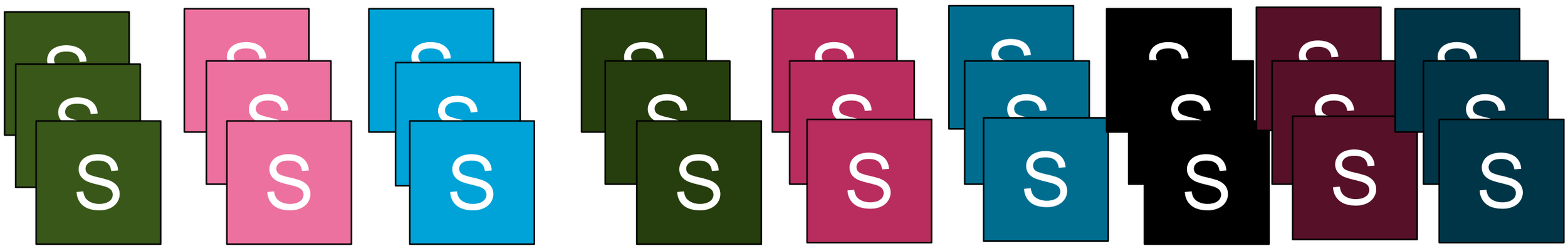
TOOL1
(2013)

STUFF 1-1-2013

3 x 3 x 3 = 27 different outputs!
This can get out of control quite quickly!







DFXML (Digital Forensics XML) is an XML language for annotating digital forensics artifacts.

DATA3

<dfxml>

- **Source media**
- **Acquisition date**
- **Location**
- **Person**

TOOL3

2013

<dfxml>

- **Tool name**
- **Source code**
- **Compile environment**
- **Runtime environment**

S

<dfxml>

- **What was found**
- **Where found**

e.g.:
- **Files**
- **Email addresses**
- **JPEGs**

...

DFXML provides metadata and provenance tracking.

There's lots of structured data to represent:

- File names, locations, MAC times, etc.
- Which program processed the data:
  - —*Which version; where compiled, compiler flags, etc.*
  - —*Where it was run, how long it took, etc.*

Originally programs kept this information in many different places:

- SleuthKit "body" file; Log files; Etc.

DFXML is a single, unified way of keeping all of this information.

- Arose out of personal need.
- Decided that it would be better not to reinvent a storage format.
- XML has broad support than other formats; tools for GB-sized objects
- Easiest way to get support for DFXML: add it to open source programs.
- Working now to merge DFXML with MITRE's CyBox

example: <creator>

```
<creator version='1.0'>
  <program>BULK_EXTRACTOR</program>
  <version>1.1.0_beta8</version>
  <build_environment>
    <compiler>GCC 4.2</compiler>
    <compilation_date>2011-11-19T23:27:21</compilation_date>
    <library name="afflib" version="3.6.9"/>
    <library name="libewf" version="20100805"/>
  </build_environment>
  <execution_environment>
    <cpuid>
      <identification>GenuineIntel</identification>
      <family>6</family>
      <clflush_size>64</clflush_size>
      <nproc>16</nproc>
      <L1_cache_size>262144</L1_cache_size>
    </cpuid>
    <command_line>src/bulk_extractor -o dell1 4DellCPi.E01</command_line>
    <uid>501</uid>
    <username>simsong</username>
    <start_time>2011-11-20T04:34:27Z</start_time>
  </execution_environment>
</creator>
```

<fileobject> presents information about a file.

A “file” is a set of 0 or more bytes and metadata.

- File name, size, and hash codes.
- Physical Location on the disk.
- Provenance

Can be on a disk, in a hash set, sent over a network, in an archive,...

```
<fileobject>
  <filename>casper/filesystem.manifest-desktop</filename>
  <filesize>32672</filesize>
  <inode>651</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>1</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime>2008-12-29T01:33:32Z</mtime>
  <atime>2008-12-28T05:00:00Z</atime>
  <crtime>2008-12-29T01:33:32Z</crtime>
  <byte_runs>
   <byte_run file_offset='0' fs_offset='5577728' img_offset='5609984' len='32672'/>
  </byte_runs>
  <hashdigest type='md5'>bd1b0831fcba1f22eff2238da96055b6</hashdigest>
  <hashdigest type='sha1'>7e072af67f8d989cc85978487b948048ac3c7234</hashdigest>
</fileobject>
```

A TCP flow is a file with <tcpflow> information.

```
<fileobject>
  <filename>074.125.019.104.00080-192.168.001.102.50955</filename>
  <filesize>2792</filesize>
  <tcpflow startime='2008-10-06T13:54:54.638913Z'
         endtime='2008-10-06T13:54:54.638913Z'
         src_ipn='74.125.19.104'
         dst_ipn='192.168.1.102'
         packets='6'
         srcport='80' dstport='50955'
         family='2' out_of_order_count='3' />
</fileobject>
<fileobject>
  <filename>192.168.001.102.50955-074.125.019.104.00080</filename>
  <filesize>655</filesize>
  <tcpflow startime='2008-10-06T13:54:54.621853Z'
         endtime='2008-10-06T13:54:54.621853Z'
         src_ipn='192.168.1.102'
         dst_ipn='74.125.19.104'
         packets='7'
         srcport='50955' dstport='80'
         family='2' out_of_order_count='0' />
</fileobject>
```

The <filename> is where the file's bytes are in the file system.

DFXML — the XML is the least important part.

Primary advantages:

- Non-experts can do forensics with tools that generate DFXML

Unexpected benefits:

- Makes it easier to replicate work — invoking command line in the XML
- Provides documentation to students
- N-version tool testing (use multiple tools to generate the same DFXML)

Challenges and alternatives

- XML is verbose / inefficient; JSON is trendy
- Information could be stored in a SQLite database.
- Some users could benefit from provenance but don’t need <fileobject>s

To spur use, we created libraries and added them to open source tools.

- Created C++ and Python libraries for efficient generation & reading of DFXML

```
xreport.add_DFXML_creator(PACKAGE_NAME,PACKAGE_VERSION);
xreport.push("configuration");
xreport.xmlout("threads",num_threads);
xreport.push("scanners");
```

- Added DFXML support to hashdeep and photorec.
- Produced useful tools that require DFXML to operate
- Added example DFXML to Forensics Wiki
- Reached out to other projects with similar goals (MITRE CyBOX)

At the present time:

- We are using DFXML in our research
- DFXML makes it faster to do build, test and validate tools
  - —*We learned an important site was using a 6-year-old C compiler...*

DFXML is a research tool, but no longer a research subject.

<dfxml>

Conclusion

In conclusion...

Digital forensics — Hard problems that look easy

- We cover the entire stack (bits → OS & Apps→ supercomputers → Internet)
- We cover most domains of computer science (security; visualization; HLT)

Real Data Corpus — research with real data.

- It's hard to get and work with real data
- Technical and legal issues
  - —*legal issues are more difficult*

Provenance tracking with DFXML

- Good technology isn't enough
- Need, usability, and cost drive adoption

<dfxml>

Contact Information:
Simson L. Garfinkel <simsong@acm.org>