Digital Forensics Research: The Next 10 Years

By

Simson Garfinkel

Presented At

The Digital Forensic Research Conference

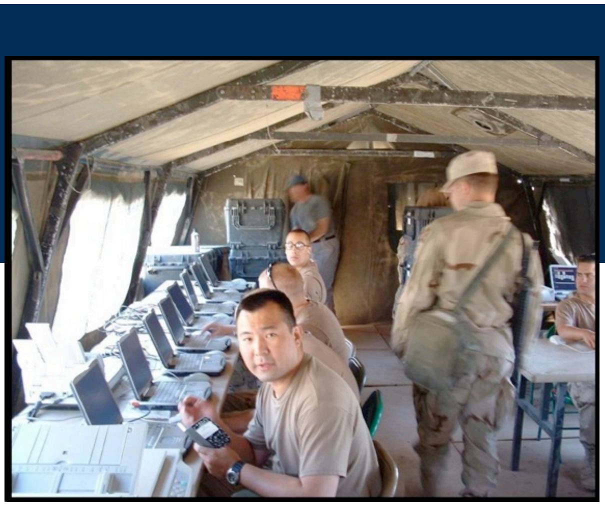
DFRWS 2010 USA Portland, OR (Aug 2nd - 4th)

Digital Forensics Research: The Next 10 Years

Simson L. Garfinkel
Naval Postgraduate School

May 10, 2010

Digital Forensics: The Sky is Falling



DF is widely used within Government & Private Sector

- Law Enforcement, Defense, E-Discovery, Document Recovery, etc.
- Hacker Investigations, Audit, etc

I argue that we have been in a "Golden Age of Digital Forensics."

- This Golden Age is quickly coming to an end.
- Organizations increasingly encounter cases with data that cannot be analyzed.
- Even when data can be analyzed, customers can wait weeks, months or longer.

Needed *dramatic improvements* in research and op efficiency:

- Shorten the *introduction to exploitation* gap (from years to months)
- Dramatically increased reliability and accuracy
- 10x – 100x improvement in processing speed.

Approach: improved data representation & composability

Prior Work 1/1

DFRWS Common Digital Evidence Storage Format Working Group.

- Created in August 2006 to standardize disk image formats.
- Goal — standardize a range of evidence formats.
- Disbanded in August 2007
  - *"because DFRWS did not have resources required to achive the goals of the group."*

Various "next-generation digital forensics systems."

- Richard and Roussev; Corey et al; Cohen (PyFlag); Ayers
- Many combine High Performance Computing (HPC) concepts with automated workflow.
- FTK3 — Uses Oracle Back End for processing.

Conceptional Frameworks.

- Mocas to "define a set of properties and terms…."
- Pollitt; CISSE 2008 brainstorming Session (Nance, Hay & Bishop); Beebe

Prior Work 1/2 — Data Collection

"The anatomy of electronic evidence — quantitative analysis of police e-crime data."

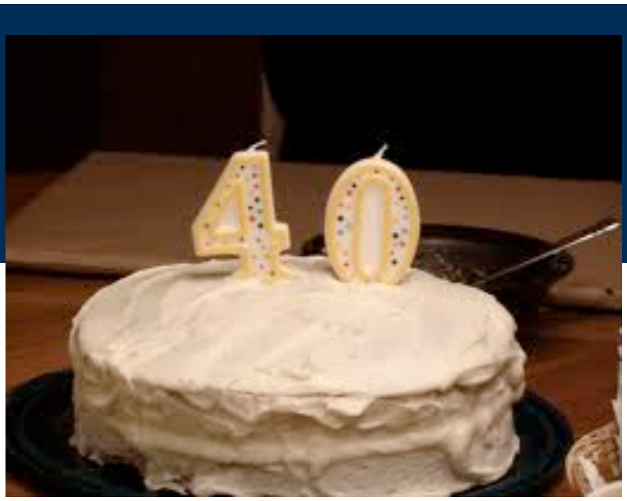
- Turnbull, Taylor and Blundell,
  - —*Reports specific digital media formats collected*

FBI Regional Computer Forensic Laboratory Program

- Annual report with the amount of media and cases processed.

Digital Forensics: A Brief History

Digital Forensics — A Brief History



Digital Forensics is roughly 40 years old.

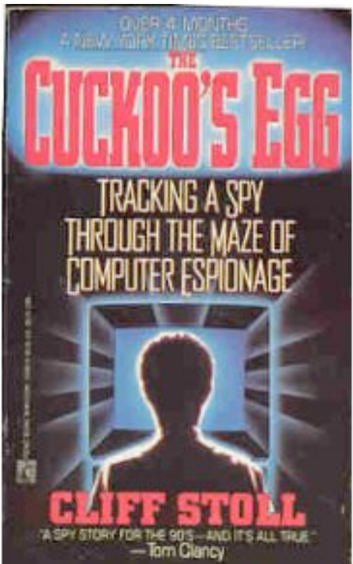
- Originally data recovery
- Late 1980s — Norton & Mace Utilities provided "Unformat, Undelete."

Early days were marked by:

- Diversity — Hardware, Software & Application
- Proliferation of file formats
- Heavy reliance on time-sharing and centralized computing
- Absence of formal process, tools & training

Forensics of end-user systems was hard, but it didn't matter much.

- Most of the data was stored on centralized computers.
- Experts were available to assist with investigations.
- There wasn't much demand!



The Golden Age of Digital Forensics: 1999—2007

Widespread use of Microsoft Windows, especially Windows XP

Relatively few file formats:

- Microsoft Office (.doc, .xls & .ppt)
- JPEG for images
- AVI and WMV for video

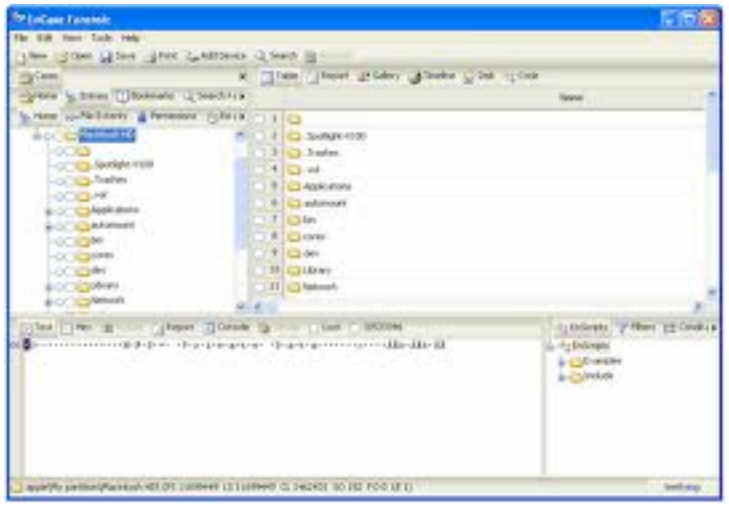
Most examinations confined to a single computer belonging to a single subject

Most storage devices used a standard interface.

- IDE/ATA
- USB

This Golden Age gave us good tools and rapid growth.

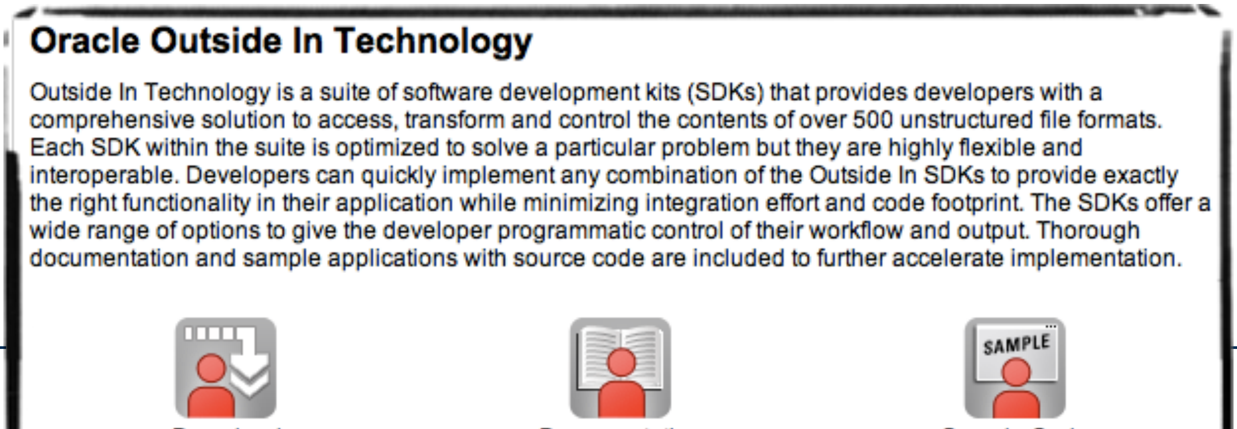
Commercial tools:



EnCase®

Open Source Tools:

Content Extraction Toolkits:



Oracle Outside In Technology

Outside In Technology is a suite of software development kits (SDKs) that provides developers with a comprehensive solution to access, transform and control the contents of over 500 unstructured file formats. Each SDK within the suite is optimized to solve a particular problem but they are highly flexible and interoperable. Developers can quickly implement any combination of the Outside In SDKs to provide exactly the right functionality in their application while minimizing integration effort and code footprint. The SDKs offer a wide range of options to give the developer programmatic control of their workflow and output. Thorough documentation and sample applications with source code are included to further accelerate implementation.

The Golden Age was aided by target conditions.

Widespread market failure of Data At Rest (DAR) Encryption

- TrueCrypt — not widely deployed
- Microsoft's EFS — hard to use
- Apple's File Vault — buggy until MacOS 10.4 / 10.5

Anti-Forensics Tools

- Largely academic curiosities

Rapid Growth of Research & Professionalization

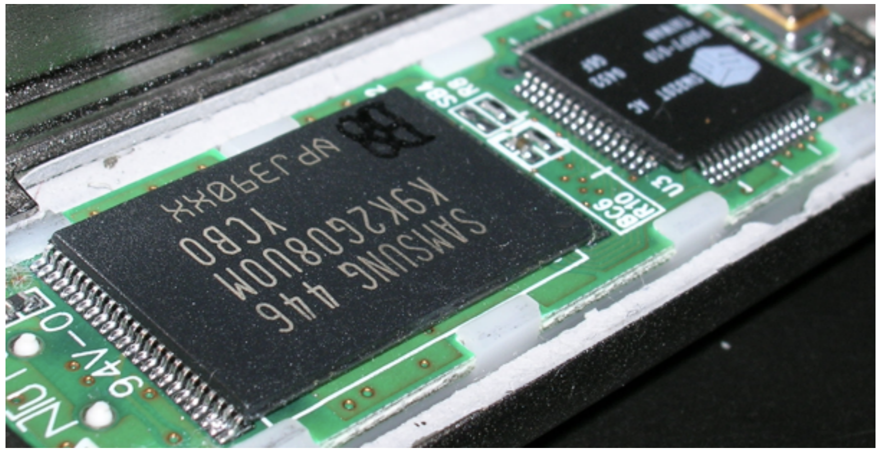
- DFRWS, IFIP WG 11.9
- Consulting firms
- 14 certificate programs
- 5 associates programs
- 16 bachelor programs
- 2 doctoral programs

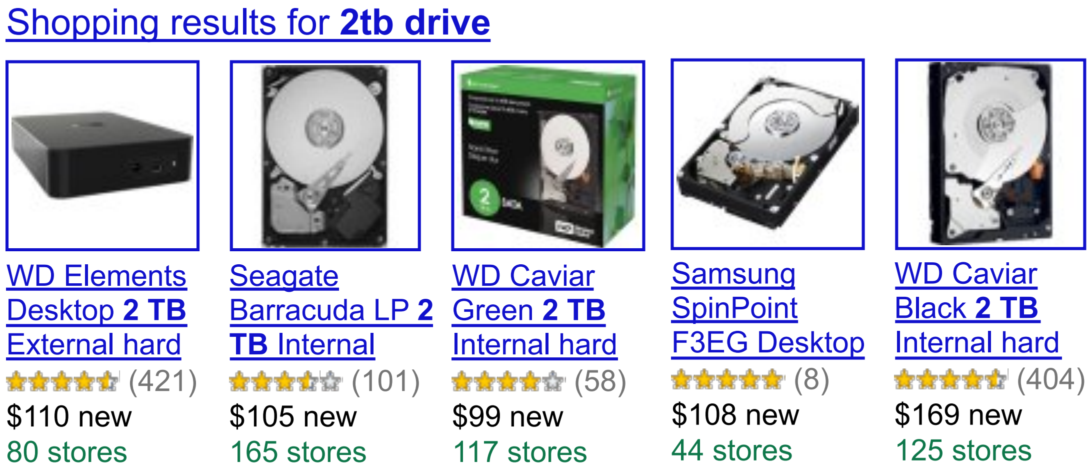
DFRWS

Get ready for the coming digital forensics crisis.
1 - Dramatically increased costs of extraction & analysis.

Much of the last decade's progress is quickly becoming irrelevant.

- Increased size of storage systems.
- Non-Removable Flash





Shopping results for **2tb drive**

WD Elements Desktop **2 TB** External hard	(421)	$110 new	80 stores
Seagate Barracuda LP **2 TB** Internal	(101)	$105 new	165 stores
WD Caviar Green **2 TB** Internal hard	(58)	$99 new	117 stores
Samsung SpinPoint F3EG Desktop	(8)	$108 new	44 stores
WD Caviar Black **2 TB** Internal hard	(404)	$169 new	125 stores

- Proliferation of operating systems, file formats and connectors
  - *JFFS2, YAFFS2, Symbian, Pre, iOS,*
  - *Most evident in mobile computing*
- Cases now require analyzing multiple devices
  - *Typical — 2 desktops, 6 phones, 4 iPods, 2 digital cameras*
  - *How many storage devices did you bring to this conference?*

The Coming Digital Forensics Crisis: Part 2 — Encryption and Cloud Computing

Pervasive Encryption — Encryption is increasingly present.

- TrueCrypt
- BitLocker
- File Vault
- DRM Technology

TRUECRYPT

FREE OPEN-SOURCE ON-THE-FLY ENCRYPTION

Home | Documentation | Downloads | News | Future | History | Screenshots | Donations | FAQ | Forum | Contact

Cloud Computing — End-user systems won't have the data.

- Google Apps
- Microsoft Office 2010
- Apple Mobile Me

RAM-based malware

Legal challenges (e.g. US vs. Comprehensive Drug Testing).

The Coming Digital Forensics Crisis. Part 3 — Mobile Phones

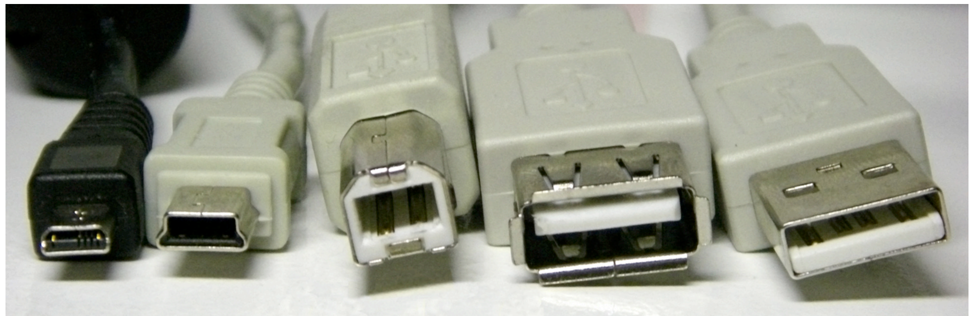
Forensic examiners established bit-copies as the gold standard.

- … but to image an iPhone, you need to jail-break it.
- Is jail-breaking forensically sound?

How do we validate tools against thousands of phones?

How do we forensically analyze 100,000 apps?

No standardized cables or extraction protocols.



NIST's *Guidelines on Cell Phone Forensics* recommends:

- "searching Internet sites for developer, hacker, and security exploit information."

The Coming Digital Forensics Crisis
Part 4 — RAM and hardware forensics is really hard.

RAM Forensics—in its infancy

- RAM structures change frequently (no reason for them to stay constant.)
- RAM is constantly changing.

Malware can hide in many places:

- On disk (in programs, data, or scratch space)
- BIOS & Firmware
- RAID controllers
- GPU
- Ethernet controller
- Motherboard, South Bridge, etc.
- FPGAs

Tools and training simply can't keep up.

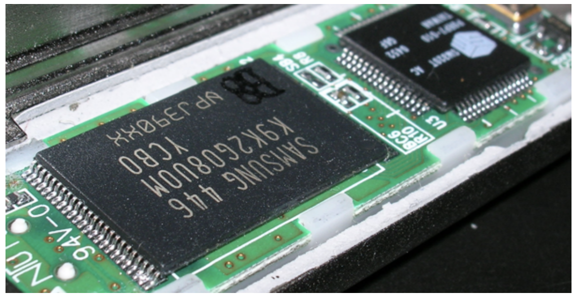
1 — Dramatically increased costs of extraction and analysis

2 — Encryption and Cloud Computing

3 — Mobile Phones

4 — RAM and Hardware Forensics

Some devices will *never* be supported by today's mainstream tools.



Today's Research Challenges

Why the gap?
What makes Digital Forensics so hard and different?

Automation and Tool Gap:

- "Evidence-Oriented Design"
- “Visibility, Filter and Report” model.
- Analysts are expected to "connect the dots."

Lack of tool interoperability.

Workforce Gap:

- Analysts require a wide breadth of knowledge.
  - —*File systems; Applications; Unicode; Machine Code; etc.*
- Training is slow
  - —*Secret Service and FBI both take 2 years to make a person an effective analyst*

Evidence-Oriented Design hampers tool design.

Today's tools were designed to find specific pieces of evidence.

- Find child porn & financial records.
- Not to assist in an investigation.

Today's tools were created for solving crimes against people---

- Evidence of the crime resides on the computer.

Today's tools were not designed for:

- Explaining ***how*** a computer was compromised.
- Finding information that is ***out-of-the-ordinary*** or ***out-of-place***.
- Diagnosing malware infestations.

Scaling — Some tools can process terabytes of data…

- … but they cannot assemble terabytes into a concise report.

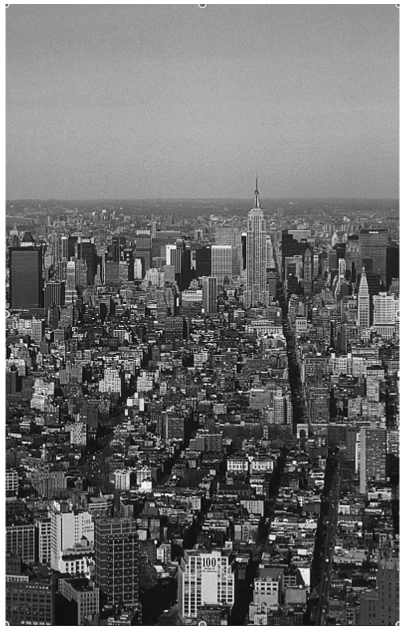
Evidence-Oriented Design limits tool evolution.

Today's tools were developed to find all the evidence.

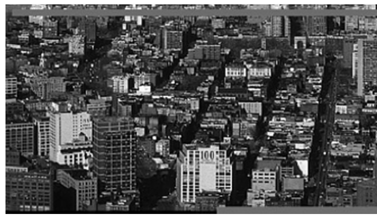
- — *"Tell me everything that's on this hard drive."*
- Increasingly, tools are used in time-constrained environments.
  - — *"Show me the best stuff you can find in the next five minutes."*

Today's tools were developed to find *documents*.

- We know how to show documents to juries.
- We don't know how to make arguments about "distinct sectors."
- As a result, research into incomplete documents has been slow.
- It was only in 2009 that Sencar and Memon showed the second half of a JPEG could be displayed.

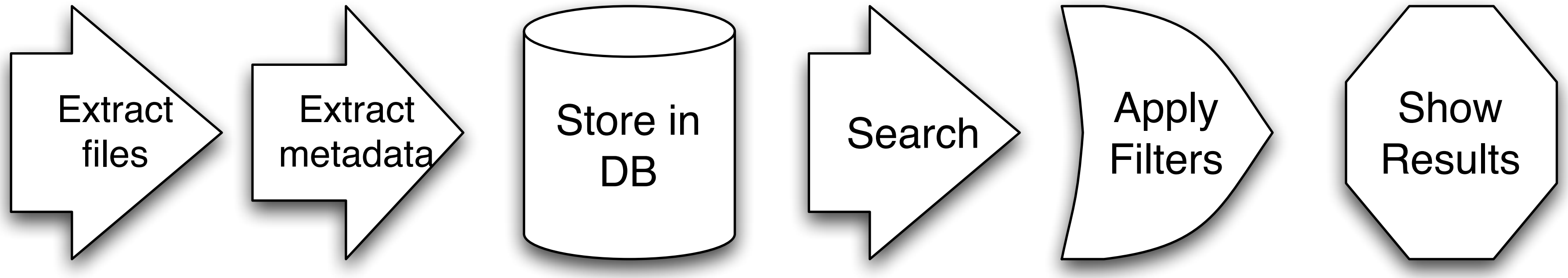


original



reconstructed

Today's tools follow a "Visibility, Filter and Report" model.



Problems:

- Analyst must prioritize data that is recovered.
- Tools do not correlate *within* this case and *between* this case and others.
- Does not readily lend itself to parallelized processing

Many tools are monolithic applications:

- Difficult to integrate with other tools.
- Difficult to automate.
- Difficult to combine tools from multiple vendors
- Difficult to integrate with the results of academic research.

Much of today's "research" is hacks, not science.

Most of today's “research” is really reverse-engineering.

- New formats are reverse-engineered by smart people with primitive tools
- No interoperability between tools. Little effort spent on performance.
- Many tools do not generalize.
  - *There are thousands of different Windows versions.*
  - *Little attention to disks/memory/network commonalities & data fusion.*

Most of today's "research" is not scientific:

- No validation over a large data sets;
- Little attention to repeatability or completeness.

Increasing diversity is increasingly a problem.

- Some devices are *never* supported by tools.

A New Research Direction

We need more standardized forensic data abstractions.

Today we have limited data formats and abstractions:

- Disk images — raw & EnCase E01 files
- Packet Capture files — BPF format
- Files — distributed as files or as ZIP for collections of files
- File Signatures — List of MD5 (or SHA1) hashes in hex.
- Extracted Named Entities — Stop lists. (typically in ASCII, rarely in Unicode)

We need new structured formats for distributing:

- Signatures Metrics (parts of files; n-grams; piecewise hashes; similarity metrics)
- File Metadata (e.g. Microsoft Office document properties)
- File system metadata (MAC times, etc.)
- Application Profiles (e.g. collections of files that make up an application.)
- Internet and social network information

Creating, testing, and adopting schema and formats is hard work.

Digital Forensics XML:
One approach for standardizing metadata...

Per-Image tags

```
<fiwalk> – outer tag
<fiwalk_version>0.4</fiwalk_version>
<Start_time>Mon Oct 13 19:12:09 2008</Start_time>
<Imagefile>dosfs.dmg</Imagefile>
<volume startsector=”512”>
```

Per <volume> tags:

```
<Partition_Offset>512</Partition_Offset>
<block_size>512</block_size>
<ftype>4</ftype>
<ftype_str>fat16</ftype_str>
<block_count>81982</block_count>
```

Per <fileobject> tags:

```
<filesize>4096</filesize>
<partition>1</partition>
<filename>linedash.gif</filename>
<libmagic>GIF image data, version 89a, 410 x 143</libmagic>
```

DFXML can be used by file extractors, carvers, report generators.

Other approaches: standardized SQL schema.

API standards are needed to support tool composability.

Forensic software is marked by diversity...

- C, C++, Java, perl, Python, EnScript; Windows, Macintosh, Linux.

Other communities faced with such diversity developed APIs. We can too!

- Language-independent.
- Disk, Sector, IP packet, bytestream object processing.
- File extraction
- File recognition & identificaiton
- Data & metadata extraction
- Standardized representations for timestamps, email addresses, names, etc.

A plug-in system would allow scale…

- Handheld devices ➔ Desktop ➔ Multi-Core System ➔ Blade Centers ➔ HPC
- Callback model allows the same code to be used in different deployments.
- PyFlag[17], OCFA[6] and DFF[9] all have significant usability barriers.
- Beware of using SQL as an integration framework (performance issues).

We must explore alternative analysis models to "Visibility, Filter and Report."

Stream-Based Forensics

- Process the contents of the hard drive without reconstructing files.
- Designed to overcome head seek latency; is this needed or useful with SSDs?
  - —*c.f. Cohen's AFF4 file-based disk imaging.*

Stochastic Analysis

- Random sampling (files & sectors) to speed partial analysis.

Triage and Prioritized Analysis

- Analysis without (or during) acquisition.
- "5 minute analysis"
- Examples:
  - —*I.D.E.A.L. Technology Corp.'s STRIKE*
  - —*ADF Triage*

Scale and Validation

Researchers need to work with large datasets.

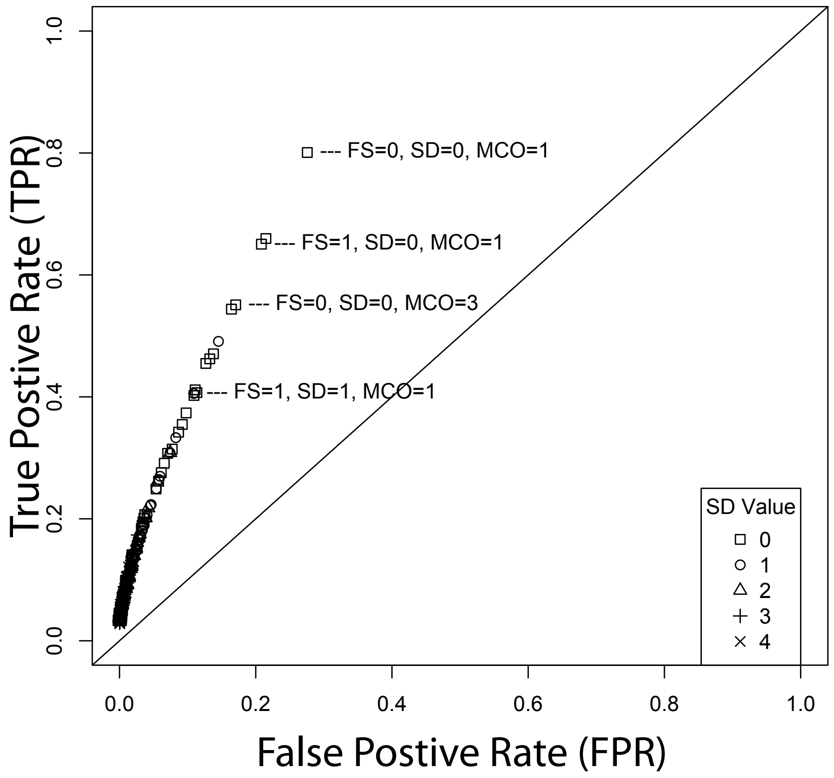
- Algorithms developed for (*n<100*) frequently fail when applied to (*n>10,000*).
- True for *n* measured in # JPEGS; TB; # hard drives; or # cell phones.

Validation with standardized corpora.

- Other researchers must be able to replicate your work!

Validation with standardized reporting metrics.

- "Accuracy" is okay, but also report:
  - —*f*-score
  - —True Positive Rate & False Positive Rate
- Many algorithms have tunable parameters.
  - —Show ROC curves!



Today's DF metrics are few and poorly articulated.

NIST Computer Forensic Tool Testing Program

- Limited testing of imaging tools & file recovery tools.
- Primary to satisfy law enforcement requirements (Daubert).

http://www.cftt.nist.gov/

Academic Publishing

- DFRWS, IFIP 11.9, etc.
- "Publish or perish" evaluation.

Forensic Challenges (DC3 & DFRWS)

- Stuff that's hard to do.
- Not scientifically evaluated.
- The "winner" is the group that
  - —*… finds the most stuff?*
  - —*… writes the most informative report?*

Moving up the Abstraction Ladder

Identity Management:

- Approaches for modeling individuals.
- Simple data elements: names; email addresses; identification numbers
- More advanced: represent a person's knowledge, capabilities & social networks
- Goals: identity resolution & disambiguation.

Data Visualization and Visual Analytics

- Is visualization good for *discovery,* or just for *presentation?*

Collaboration

- How can multiple investigators be used more effectively on a single case?
- How can the system automatically recognize when multiple cases are connected?
  - *Stealth Software's private search for secret identities.*

Autonomous Operation

GET EVIDENCE

Conclusion: Digital Forensics faces an impending crisis!

Technological progress is making our job harder, not easier.

- Increasing storage densities
- Cloud Computing
- Pervasive Encryption

Given these trends, research must be *smarter* and *more applicable*

- Standardized abstractions & formats.
- Standardized APIs for analysis.
- Forensic Data sharing.
- Composable tools.

Funding agencies need to:

- Adopt open standards and procedures.
- Insist on interoperability & validation.