afids

- another forensic image dataset

Mark Guido[1], Michael McCarrin[2], David Baker[1], Vik Harichandran[1] and Sam Brothers[1]

1 MITRE
2 Naval Postgraduate School

Why create large, shareable corpora of forensic data?

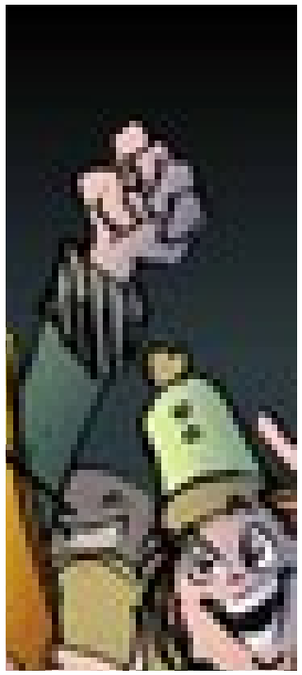
DIGITAL INVESTIGATION 6 (2009) S2–S11

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

Digital Investigation



Bringing science to digital forensics with standardized forensic corpora

Simson Garfinkel[a,b,*], *Paul Farrell*[a], *Vassil Roussev*[c], *George Dinolt*[a]

ion Sciences, Department of Computer Science, Naval Postgraduate School,

ABSTRACT

Progress in computer forensics research has been limited by the lack of a standardized data sets—*corpora*—tha
needed to furthe
announce the ava

ADFSL Conference on Digital Forensics, Security and Law, 2011

CREATING REALISTIC CORPORA FOR SECURITY AND FORENSIC EDUCATION

Kam Woods
School of Information and Library Science
University of North Carolina
Chapel Hill, NC
kamwoods@email.unc.edu

Christopher A. Lee
School of Information and Library Science
University of North Carolina
Chapel Hill, NC
callee@ils.unc.edu

Simson Garfinkel
Graduate School of Operational and Information Sciences
Department of Computer Science
Naval Postgraduate School
Monterey, CA
slgarfin@nps.edu

David Dittrich
Applied Physics Laboratory
University of Washington
Seattle, WA
dittrich@uw.edu

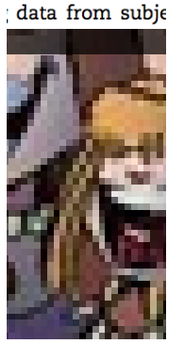
Adam Russell
Graduate School of Operational and Information Sciences
Department of Computer Science
Naval Postgraduate School
Monterey, CA
amrussell@nps.edu

Kris Kearton
NCTAMS LANT DET ROTA
PSC 819 BOX 64
FPO AE 09645
kris.kearton@eu.navy.mil

ABSTRACT

We present work on the design, implementation, distribution, and use of realistic forensic datasets to support digital forensics and security education. We describe in particular the "M57-Patents" scenario, a multi-modal corpus consisting of hard drive images, RAM images, network captures, and images from other devices typically found in forensics investigations such as USB drives and cellphones. Corpus creation has been performed as part of a scripted scenario; subsequently it is less "noisy" than real-world data but retains the complexity necessary to support a wide variety of forensic education

sics has focused
courts. Researche
data from subje



Contents lists available at SciVerse ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus

Simson Garfinkel*

Naval Postgraduate School, Computer Science, 900 N Glebe Rd, Arlington, VA 2203, United States

Keywords:
Digital forensics
Lessons learned
Digital corpora

ABSTRACT

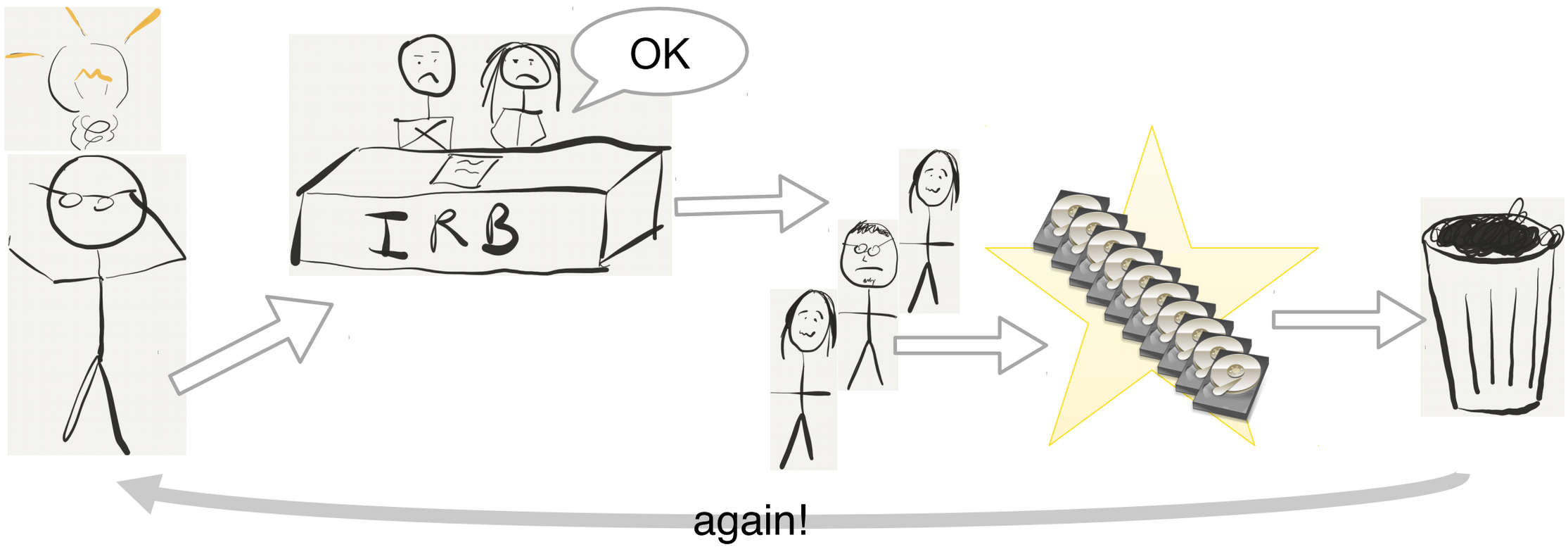
Writing digital forensics (DF) tools is difficult because of the diversity of data types that needs to be processed, the need for high performance, the skill set of most users, and the requirement that the software run without crashing. Developing this software is dramatically easier when one possesses a few hundred disks of other people's data for testing purposes. This paper presents some of the lessons learned by the author over the past 14 years developing DF tools and maintaining several research corpora that currently total roughly 30TB.

Published by Elsevier Ltd.

Also, convenience.

Creating datasets is time-consuming and expensive.

- If each research group produces their own, the total cost to the community goes up.
- ***Worse***: can't re-use across different projects.



The scientific method needs good data.

A *verifiable* hypothesis assumes the ability to test against *representative* data.

Representative:

- Complexity, scale, etc.
- Not so easy given the variety of possible devices.
- Simulated data falls short.

Verifiable:

=***reproducible***
=data must be ***shareable***
=again, not so easy

It worked!

Rowe, N. C., & Garfinkel, S. L. (2010, May). Global analysis of drive file times. In *Systematic Approaches to Digital Forensic Engineering (SADFE), 2010 Fifth IEEE International Workshop on* (pp. 97-108). IEEE.

Beverly, R., Garfinkel, S., & C

Buchanan, W. J., Macfarlane,

Brown, R. D. (2011). Reconst

Roussev, V. (2011). An evalua

Roussev, V. (2012, January).

Nelson, A. (2012, January). X

Springer.

Rowe, N. C., Garfinkel, S. L.,

Cyberspace: The Challenge t

Garfinkel, S. L. (2013). Digita

Quick, D., & Choo, K. K. R. (2

Zarate, C., Garfinkel, S. L., H

POSTGRADUATE SCHOOL

Noel, G. E., & Peterson, G. L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. *Digital Investigation*, *11*(1), 43-56.

Baggili, I., & Breitinger, F. (2015, March). Data sources for advancing cyber forensics: What the social world has to offer. In *2015 AAAI Spring Symposium Series*.

Digital Investigation 22 (2017) S94–S105

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2017 USA — Proceedings of the Seventeenth Annual DFRWS USA

Availability of datasets for digital forensics – And what is missing

Cinthya Grajeda, Frank Breitinger*, Ibrahim Baggili

Cyber Forensics Research and Education Group (UNHcFREG), Tagliatela College of Engineering, ECECS, University of New Haven, 300 Boston Post Rd., West Haven, CT 06516, USA

Keywords:
Availability
Data collection
Dataset
Origin
Experiment generated
User generated
Repository

ABSTRACT

This paper targets two main goals. First, we want to provide an overview of available datasets that can be used by researchers and where to find them. Second, we want to stress the importance of sharing datasets to allow researchers to replicate results and improve the state of the art. To answer the first goal, we analyzed 715 peer-reviewed research articles from 2010 to 2015 with focus and relevance to digital forensics to see what datasets are available and focused on three major aspects: (1) the origin of the dataset (e.g., real world vs. synthetic), (2) if datasets were released by researchers and (3) the types of datasets that exist. Additionally, we broadened our results to include the outcome of online search results. We also discuss what we think is missing. Overall, our results show that the majority of datasets are experiment generated (56.4%) followed by real world data (36.7%). On the other hand, 54.4% of the articles use existing datasets while the rest created their own. In the latter case, only 3.8% actually released their datasets. Finally, we conclude that there are many datasets for use out there but finding them can be challenging.

From real data to real tools:

Beyond academia: many tools tested against the RDC.

- Autopsy / The Sleuth Kit, bulk_extractor, smirk, sdhash, hashdb, sceadan
- bulk_extractor development especially:
  - debugging, capability development, scalability

Complexity and "messiness" of real world data helps debug tools and prepare for operational use.

But… some problems remain.

Device variety:

- mostly hard drives (some mobile)
- mostly older Windows systems

Access:

- Owned by the US Navy = maximum paperwork.
  - ➡ Sharing with foreign nationals is often impossible.
- Technical barriers to sharing:
  - ➡ transfer costs, access to Navy systems, etc.

Age:

- Most recent update in 2014, but the acquisition is second-hand, so there is a lag.

Currently no planned updates!

How to keep growing?

Increase size, device variety, number of newer devices

- Increases storage costs.
- Requires improvements to architecture.
- Does not solve policy-related sharing problems.

Reach more users

- Must be able to respond to their requests.
- Increasing exposure increases PR risk in case of negative incidents (i.e. discovery of illegal materials, etc.)

Single owner / single sponsor = single point of failure.

Introducing:
Another Forensic Image Dataset!

Community-supported, data-as-service model.

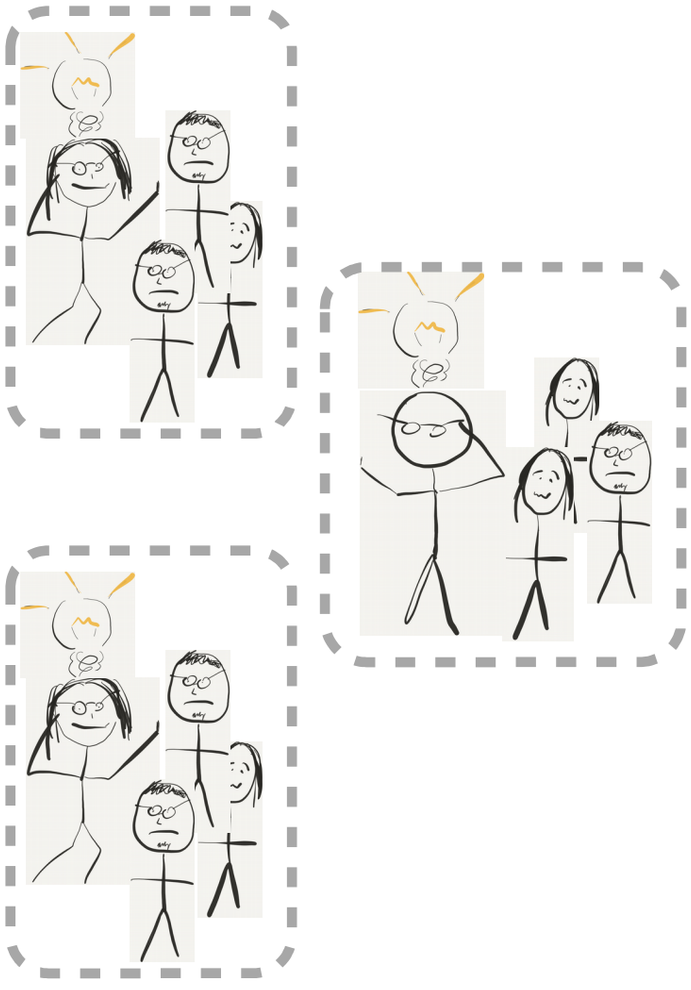
- Clean break.
- Leveraging AWS GovCloud for availability.

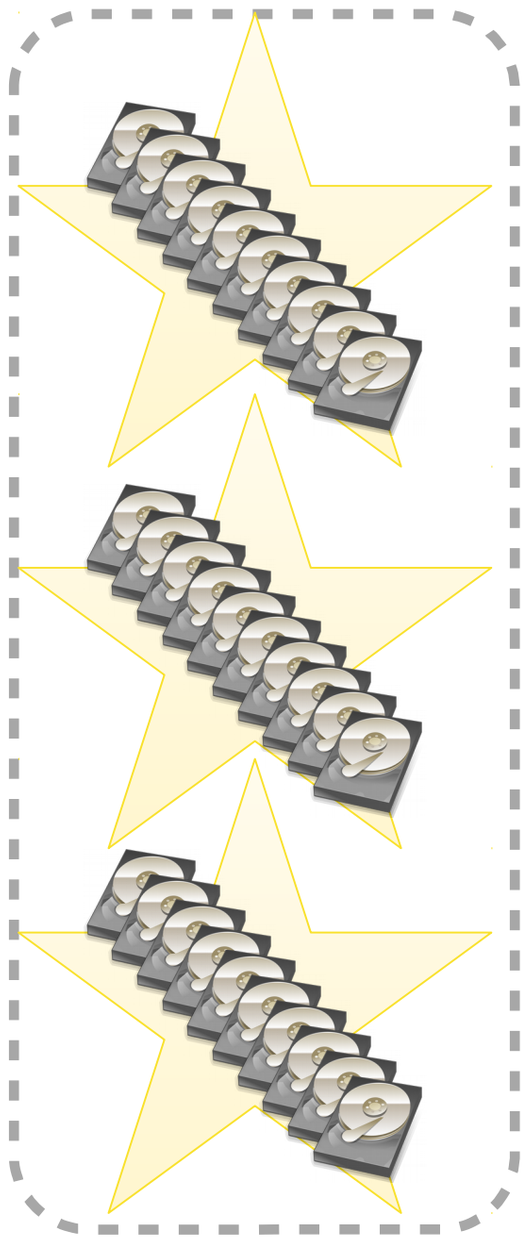
Incorporating lessons learned from the RDC.

- Improved options for access.
- Improved privacy controls / policy.
- Due diligence checks for contraband / CP.
- Built-in support for scalable analysis methods.

Note: this does not yet exist.

Better data, less misery.





keep for science

N groups with the resources to create a decent dataset

\+ Coordination!

= A much better dataset.

Building afids with contributors and subscribers:

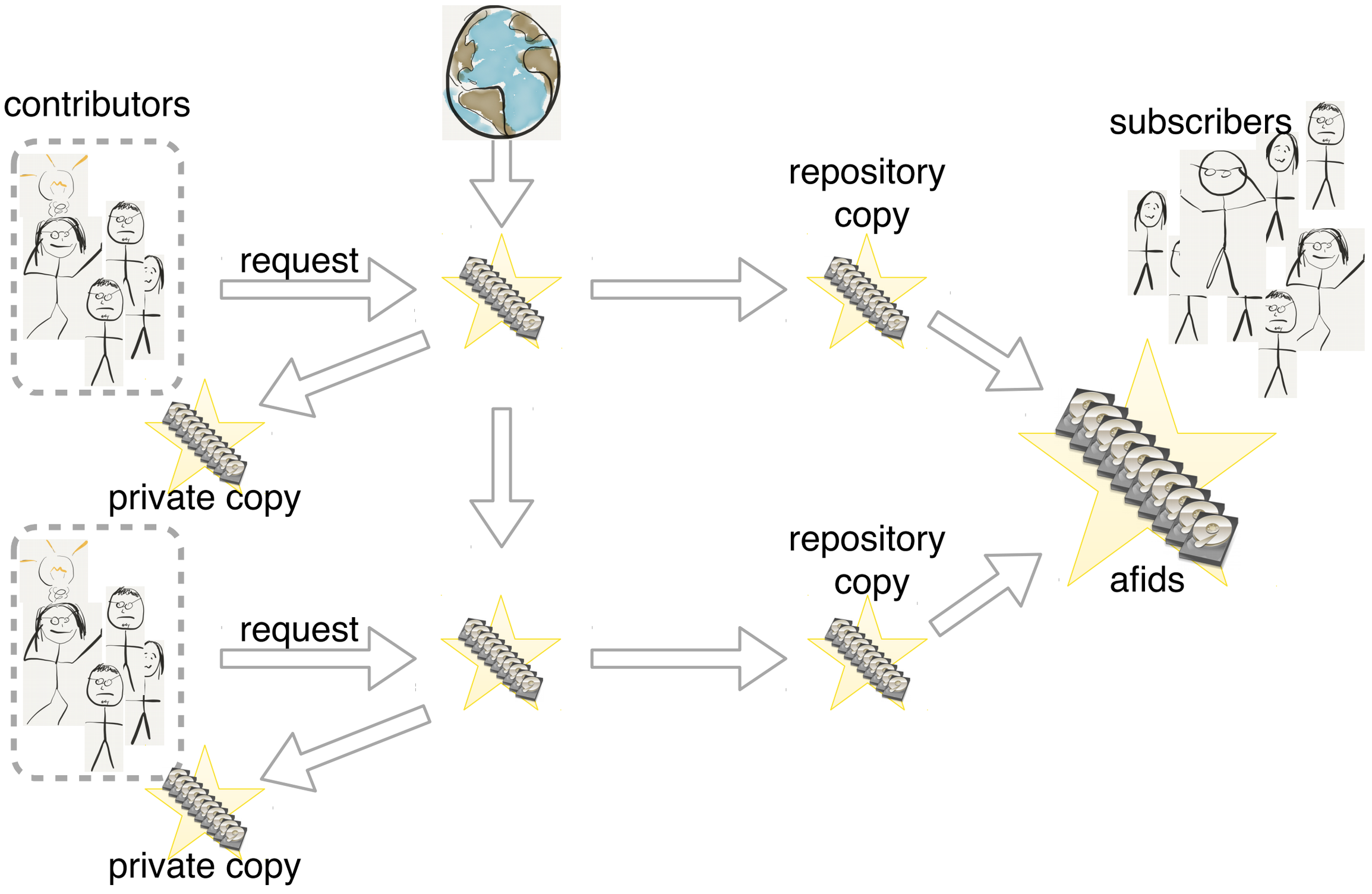
Contributors: sponsor new devices to be added to the set.

- Specify quantity and type.
- Contribute cost of acquisition.
- Receive **(and keep)** copies of the images.
- Images are also contributed to a shared pool in AWS.

Subscribers: just want access to the existing pool.

- Subscription fees defray acquisition costs.

Building afids with contributors and subscribers:



“Mode” of access should be tailored to research needs.

Not all research in forensics requires access to PII.

Direct Access: full access to the data

- E.g. mount a drive in Autopsy.
- Requires IRB approval.

Mediated Access: submit queries, get (sanitized) results.

- Examples: measuring n-gram distribution, extracting geolocation data (possibly), etc.
- Researcher is not exposed to any PII.

If PII can be avoided, life is easier.

Measuring “ground truth”

Real data = we don’t always have control over ground truth.

- For some projects, it may not matter.
- But.. sometimes this is a good reason to create your own dataset.
- For example, if you really need to interview the device users.

Alternative: cross validation with other research results.

- As users study the data, we learn more about it.
- New work can verify consistency with previous work.
- Or not! Discrepancies will be interesting.

Special challenges of live data will get special attention.

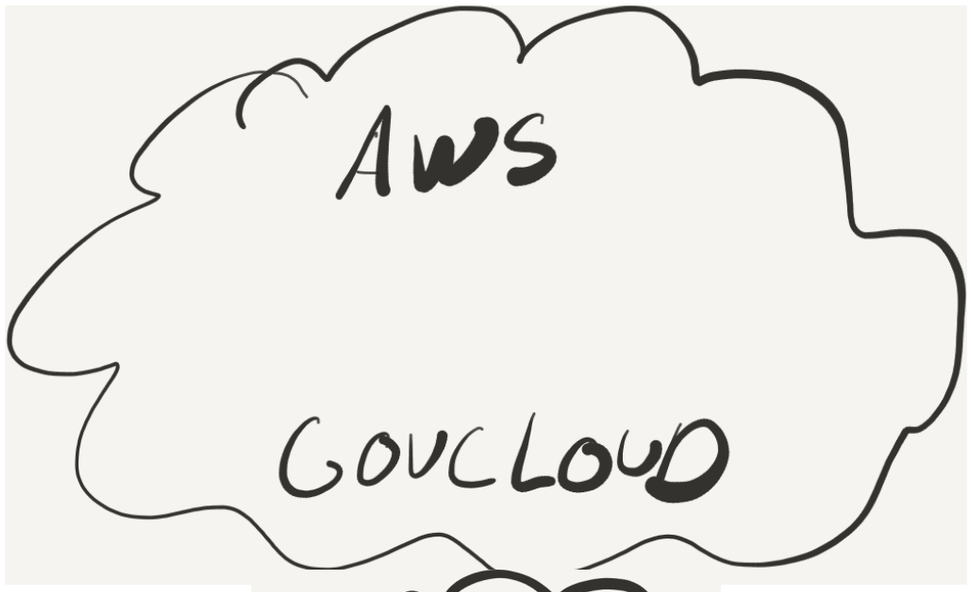
Staying safe and *legal*:

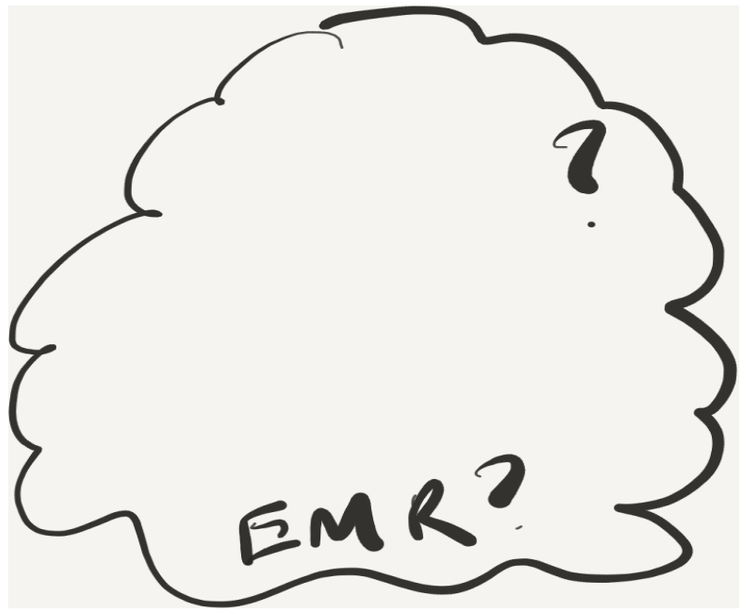
- Preliminary checks for ***illegal things*** (e.g. child pornography).
- If found later, we can remove quickly and alert the original contributor.
- Don't need to track down various copies.
- Tags for ***potentially objectionable things***: malware, etc.

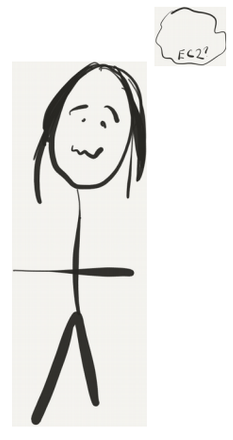
Privacy: we want to go beyond mere compliance.

- Risk to subjects should be as close to zero as possible.
- Opportunity for collaboration with privacy specialists.

afids architecture will support standalone and cluster-based tools







familiar interfaces for traditional forensic tools

flexible to support new, scalable approaches

Use Cases

Geolocation

- Studying geo-coordinate artifacts.
- Really ***don't*** want to know who they belong to.

Artifact Inference

- See Jim Jones' tool from yesterday.
- Look for trends over a large unknown dataset.

Scalable feature extraction / benchmarking

- How do we perform collection-scale feature extraction and artifact correlation?
- Currently working on this at NPS.

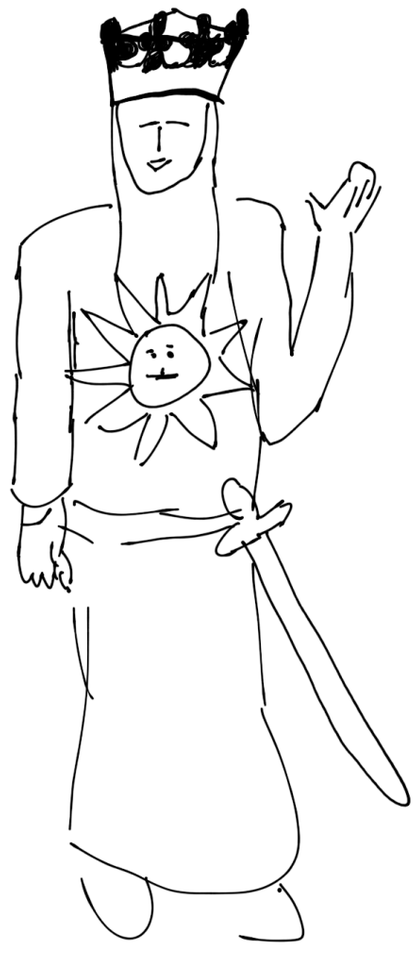
Next Steps

1. Present to American Academy of Forensic Sciences

2. Target date for standing up: February 2018

3. Trying to compile a list of interested groups.

Questions?



Mark Guido
mguido@mitre.org

Michael McCarrin
mrmccarr@nps.edu

If you have ideas or interest, please get in touch!