



# bulk\_extractor: A Stream-Based Forensics Tool

Simson L. Garfinkel

Associate Professor, Naval Postgraduate School

June 14, 2011

<http://afflib.org/>

# NPS is the Navy's Research University.

Location: Monterey, CA

Students: 1500

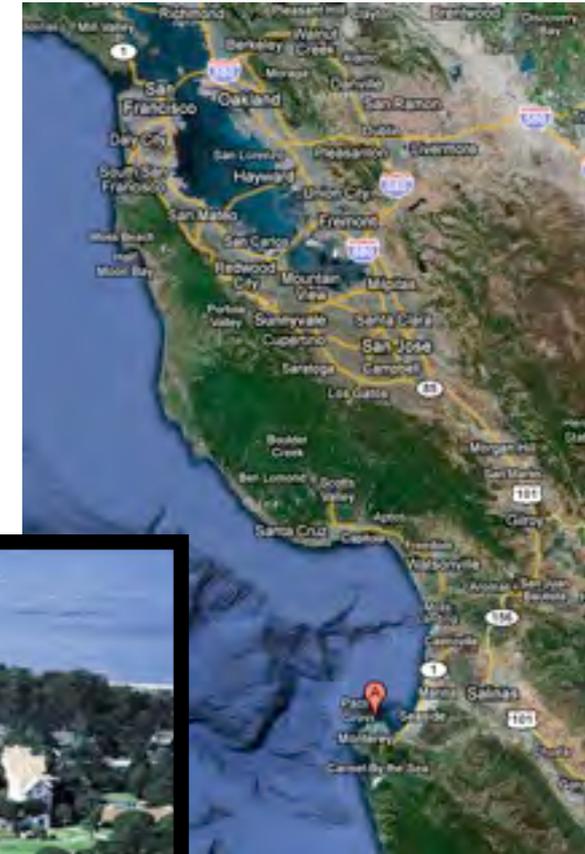
- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)
- *All students are fully funded*

Schools:

- Business & Public Policy
- Engineering & Applied Sciences
- Operational & Information Sciences
- International Graduate Studies

NCR Initiative:

- 8 offices on 5th floor, 900N Glebe Road, Arlington
- FY12 plans: 4 professors, 2 postdocs
- **IMMEDIATE OPENINGS FOR RESEARCHERS**
- **IMMEDIATE SLOTS FOR .GOV PHDs!**



# Current NPS research thrusts

## Area #1: End-to-end automation of forensic processing

- Digital Forensics XML Toolkit
- Disk Image -> Power Point

## Area #2: Bulk Data Analysis

- Statistical techniques (sub-linear algorithms)
- Similarity Metrics

## Area #3: Data mining for digital forensics

- Automated social network analysis (cross-drive analysis)

## Area #4: Creating Standardized Forensic Corpora

- Freely redistributable disk and memory images, packet dumps, file collections.





# Stream-based forensics with bulk\_extractor

# Stream-Based Disk Forensics: Scan the disk from beginning to end; do your best.



**3 hours, 20 min  
to *read* the data**

1. Read all of the blocks in order.
2. Look for information that might be useful.
3. Identify & extract what's possible in a single pass.

# Primary Advantage: Speed

No disk seeking.

Potential to read and process at disk's maximum transfer rate.

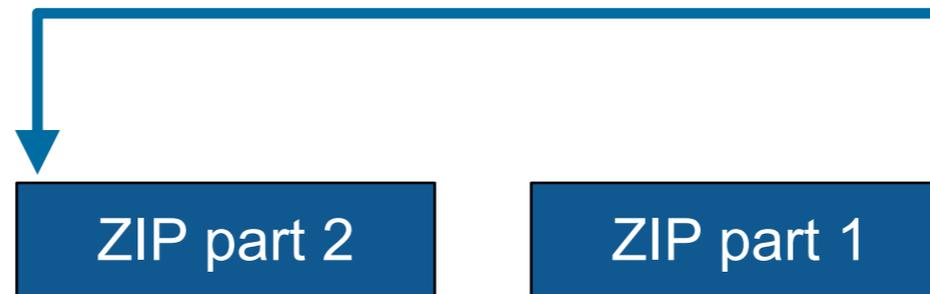
Potential for intermediate answers.

Reads all the data — allocated files, deleted files, file fragments.

- Separate metadata extraction required to get the file names.



# Primary Disadvantage: Completeness



Fragmented files won't be recovered:

- Compressed files with part2-part1 ordering (possibly .docx)
- Files with internal fragmentation (.doc but not .docx)

Fortunately, most files are *not* fragmented.

- Individual components of a ZIP file can be fragmented.

Most files that *are* fragmented have carvable internal structure:

- Log files, Outlook PST files, etc.

This talk describes `bulk_extractor`,  
a tool for performing stream-based forensics.

Why you should care: a `bulk_extractor` success story



History of `bulk_extractor`

Internal design

Suppressing false positives with context sensitive stop lists.

Extending `bulk_extractor` with plug-ins

Future Plans



San Luis Obispo  
Sincerely, California.



<http://www.sanluisobispovacations.com/>

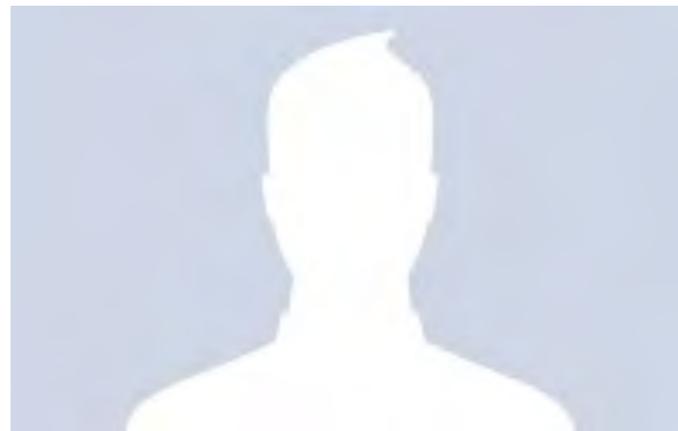
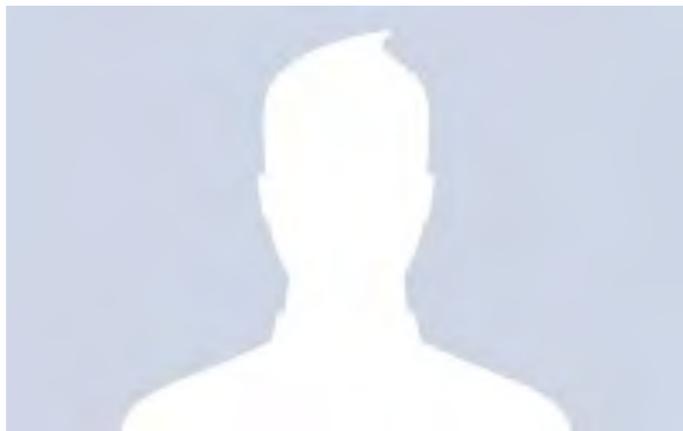


A bulk\_extractor  
Success Story

# City of San Luis Obispo Police Department, Spring 2010

District Attorney filed charges against two individuals:

- Credit Card Fraud
- Possession of materials to commit credit card fraud.



Defendants:

- Arrested with a computer.
- Expected to argue that defendants were unsophisticated and lacked knowledge.

Examiner given 250GiB drive *the day before preliminary hearing.*

- Typically, it would take several days to conduct a proper forensic investigation.

# bulk\_extractor found actionable evidence in 2.5 hours!

Examiner given 250GiB drive *the day before preliminary hearing.*



## Bulk\_extractor found:

- Over 10,000 credit card numbers on the HD (1000 unique)
- Most common email address belonged to the primary defendant (possession)
- The most commonly occurring Internet search engine queries concerned credit card fraud and bank identification numbers (intent)
- Most commonly visited websites were in a foreign country whose primary language is spoken fluently by the primary defendant.

Armed with this data, the DA was able to have the defendants held.

*Faster* than conventional tools.  
Finds data that other tools miss.

Runs 2-10 times faster than EnCase or FTK *on the same hardware*.

- bulk\_extractor is multi-threaded; EnCase 6.x and FTK 3.x have little threading.

Finds stuff others miss.

- “Optimistically” decompresses and re-analyzes all data.
- Finds data in browser caches (downloaded with zip/gzip), and in many file formats.

Presents the data in an easy-to-understand report.

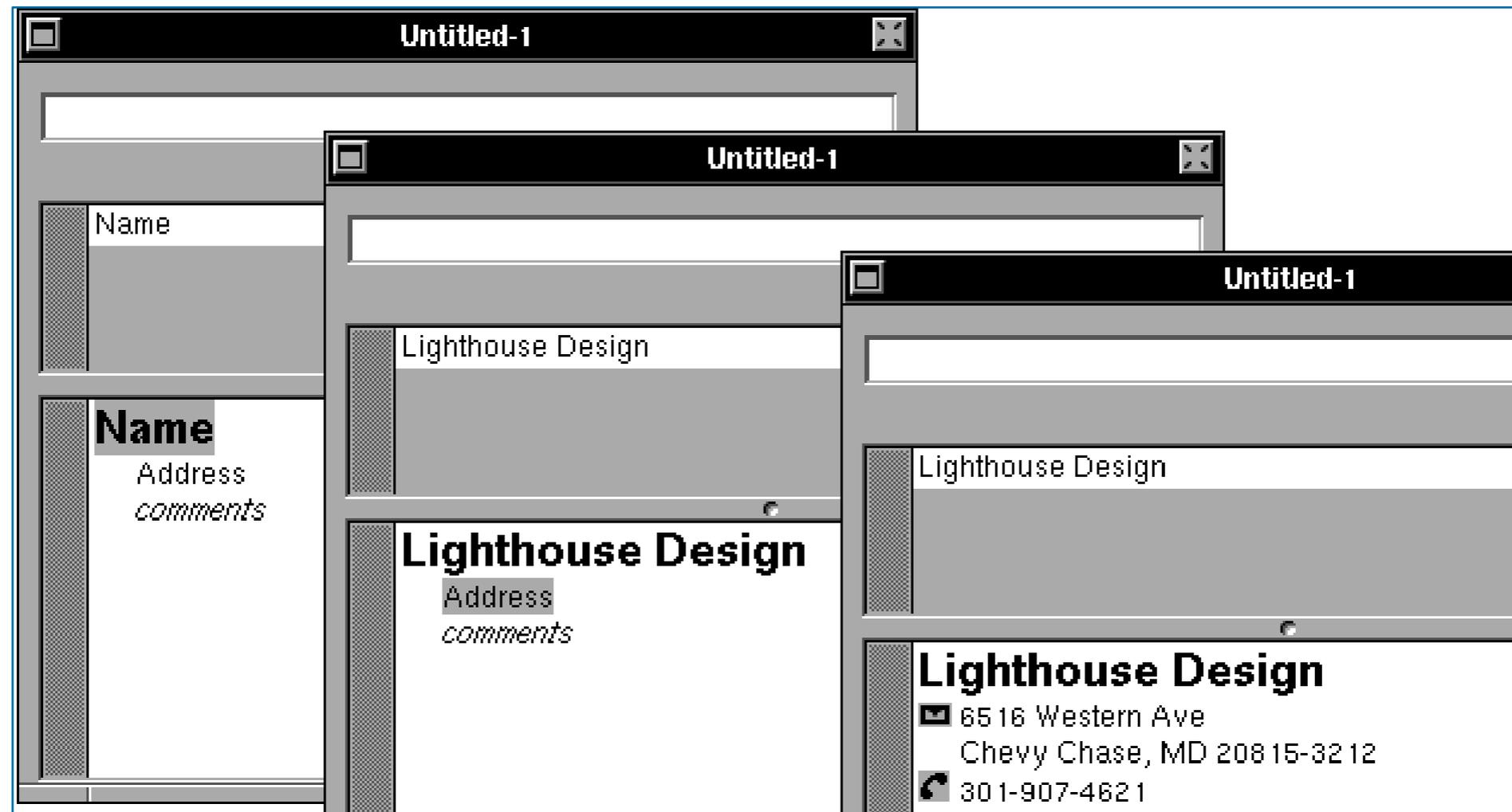
- Produces “histogram” of email addresses, credit card numbers, etc.
- Distinguishes primary user from incidental users.



# History of bulk\_extractor

# bulk\_extractor: 20 years in the making!

In 1991 I developed SBook, a free-format address book.



SBook used “Named Entity Recognition” to find addresses, phone numbers, email addresses *while you typed*.

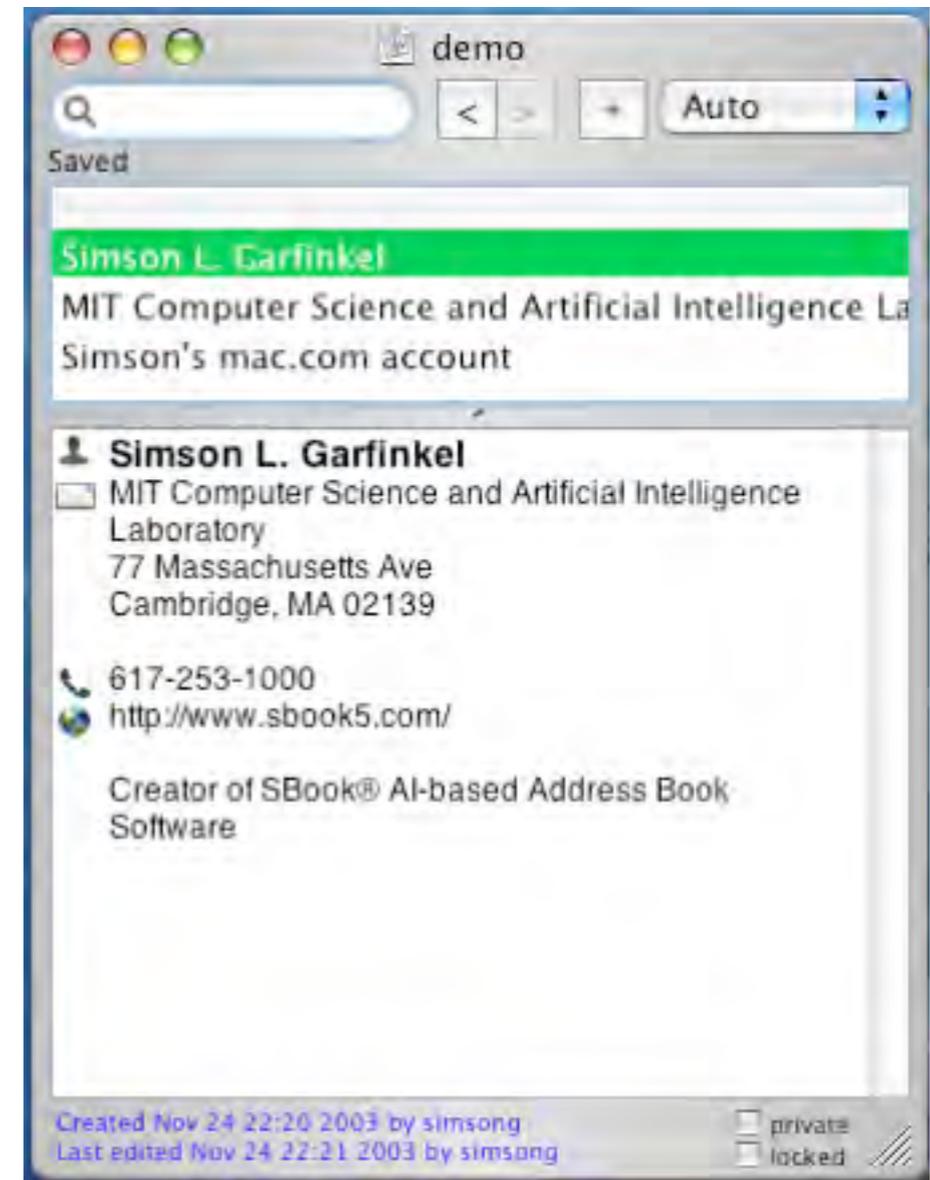
# Today we call this technology Named Entity Recognition

## SBook's technology was based on:

- Regular expressions executed in parallel
  - *US, European, & Asian Phone Numbers*
  - *Email Addresses*
  - *URLs*
- A gazette with more than 10,000 names:
  - *Common "Company" names*
  - *Common "Person" names*
  - *Every country, state, and major US city*
- Hand-tuned weights and additional rules.

## Implementation:

- 2500 lines of GNU flex, C++
- 50 msec to evaluate 20 lines of ASCII text.
  - *Running on a 25Mhz 68030 with 32MB of RAM!*



# In 2003, I bought 200 used hard drives

The goal was to find drives that had not been properly sanitized.

## First strategy:

- DD all of the disks to image files
- run **strings** to extract printable strings.
- **grep** to scan for email, CCN, etc.
  - *VERY SLOW!!!!*
  - *HARD TO MODIFY!*

## Second strategy:

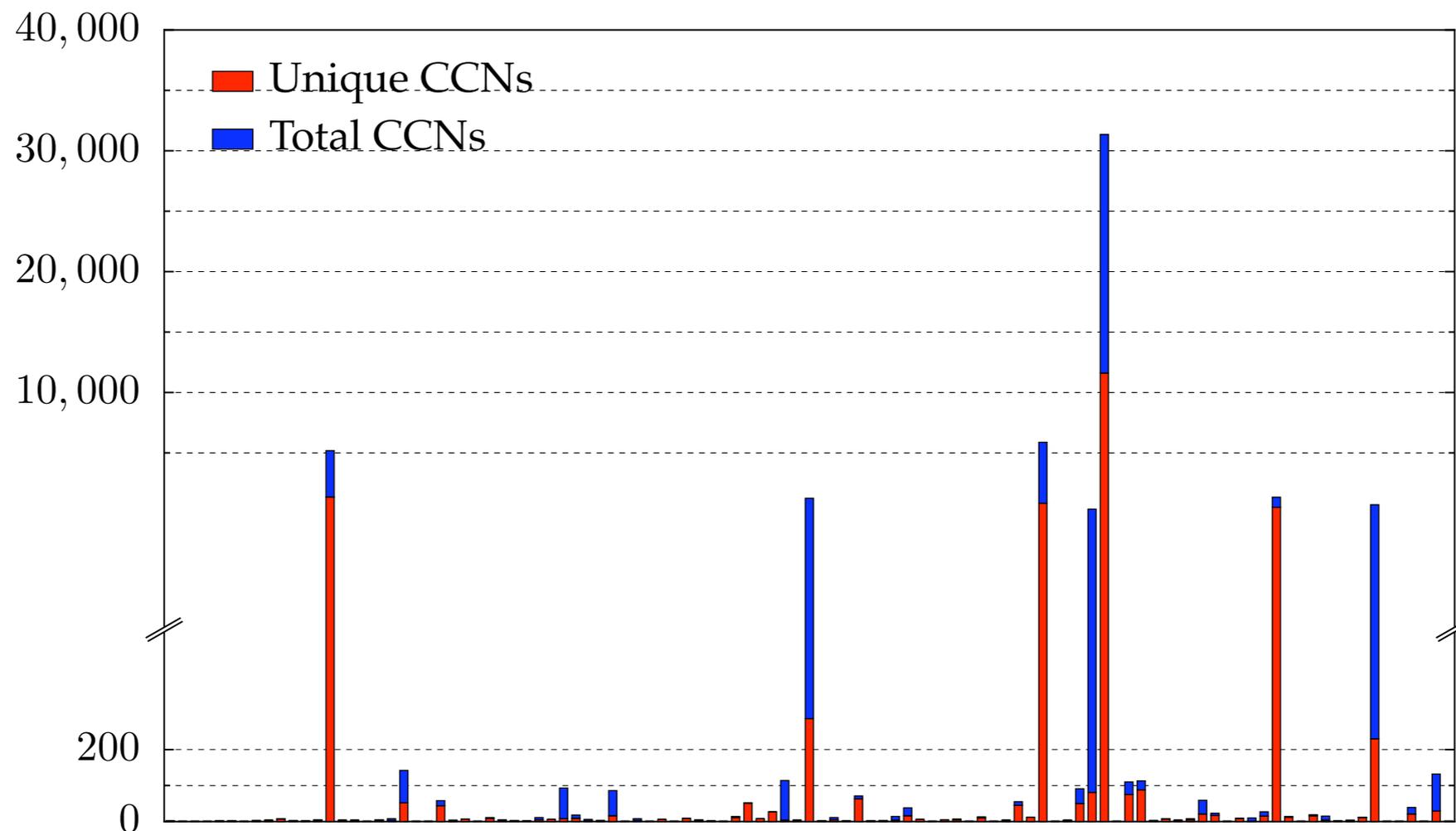
- Use SBook technology!
- Read disk 1MB at a time
- Pass the *raw disk sectors* to flex-based scanner.
- Big surprise: scanner didn't crash!



# Simple flex-based scanners required substantial post-processing to be useful

## Techniques include:

- Additional validation beyond regular expressions (CCN Luhn algorithm, etc).
- Examination of feature “neighborhood” to eliminate common false positives.



The technique worked well to find drives with sensitive information.

# Between 2005 and 2008, we interviewed law enforcement regarding their use of forensic tools.

Law enforcement officers wanted a *highly automated* tool for finding:

- Email addresses
- Credit card numbers (including track 2 information)
- Search terms (extracted from URLs)
- Phone numbers
- GPS coordinates
- EXIF information from JPEGs
- All words that were present on the disk (for password cracking)

The tool had to:

- Run on Windows, Linux, and Mac-based systems
- Run with *no* user interaction
- Operate on raw disk images, split-raw volumes, E01 files, and AFF files
- Allow user to provide additional regular expressions for searches
- Automatically extract features from compressed data such as gzip-compressed HTTP
- Run at maximum I/O speed of physical drive
- Never crash

# Starting in 2008, we made a series of limited releases. Today we are releasing bulk\_extractor 1.0.0

- January 2008 — Created Subversion Repository
- April 2010 — Initial public release - 0.1.0
- May 2010 — Initial multi-threading release - 0.3.0
  - *Each thread runs in its own process*
- Sept. 2010 — Stop lists - 0.4.0
- Oct. 2010 — Context-based stop-lists - 0.5.0
- Dec. 2010 — Switch to POSIX-based threads — 0.6.0
- Dec. 2010 — Support for Windows HIBERFIL.SYS decompression — 0.7.0
- Jun. 2010 — First 1.0.0 Release (TODAY)

Tool capabilities result from substantial testing and user feedback.

Moving technology from the lab to the field has been challenging:

- Must work with evidence files of *any size* and on *limited hardware*.
- Users can't provide their data when the program crashes.
- Users are *analysts* and *examiners*, not engineers.



Inside bulk\_extractor

# bulk\_extractor: architectural overview

## Written in C, C++ and GNU flex

- Command-line tool.
- Linux, MacOS, Windows (compiled with mingw)

## Key Features:

- “Scanners” look for information of interest in typical investigations.
- Recursively re-analyzes compressed data.
- Results stored in “feature files”
- Multi-threaded

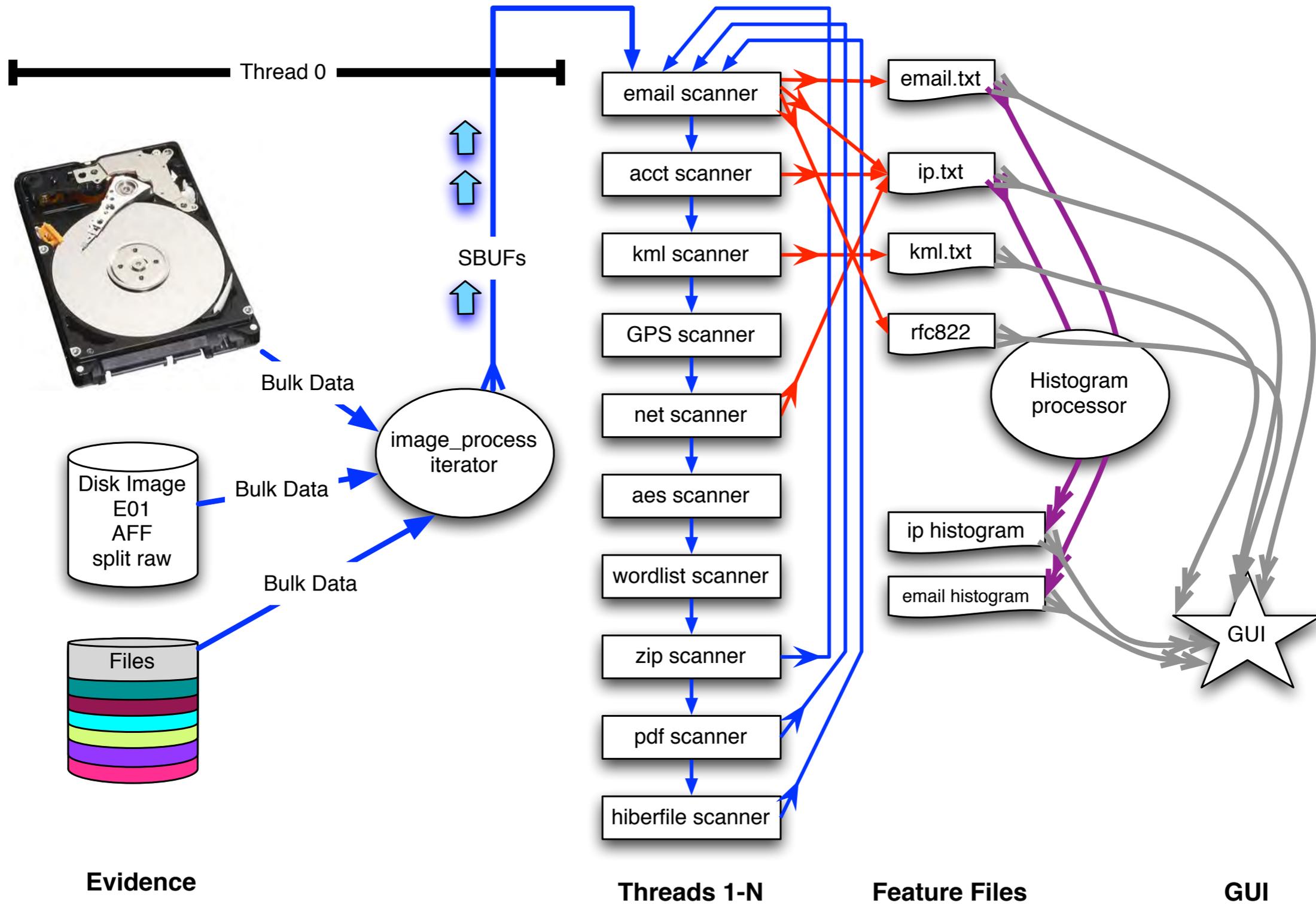
## Java GUI

- Runs command-line tool and views results

bulk\_extractor extracts “features” from disk images.



# bulk\_extractor: system diagram



# image processing

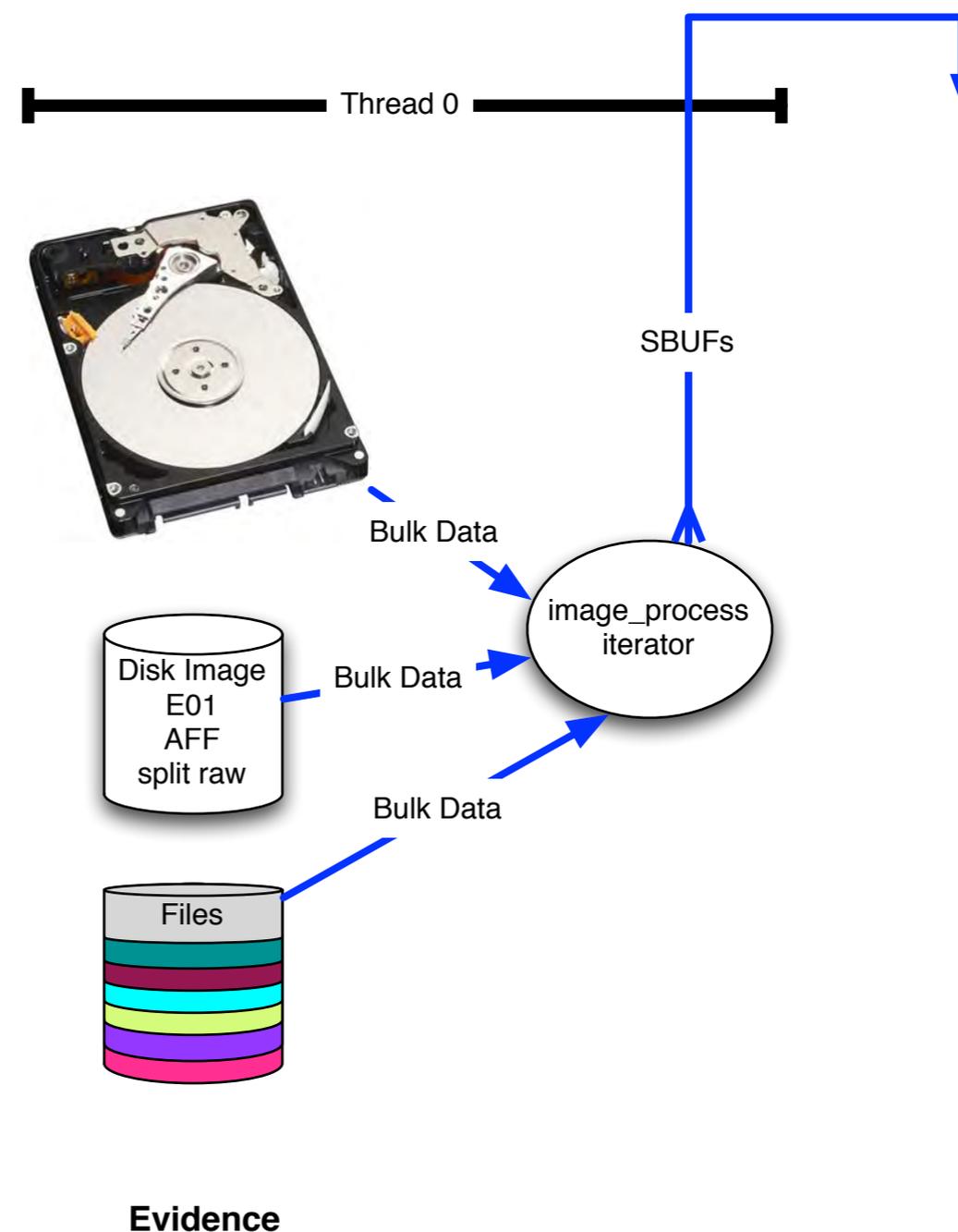
## C++ iterator handles disks, images and files

Works with multiple disk formats.

- E01
- AFF
- raw
- split raw
- individual disk files

Produces sbuf\_t object:

```
class buf_t {  
    ...  
public:  
    uint8_t *buf;    /* data! */  
    pos0_t pos0;    /* forensic path */  
    size_t bufsize;  
    size_t pagesize;  
    ...  
};
```

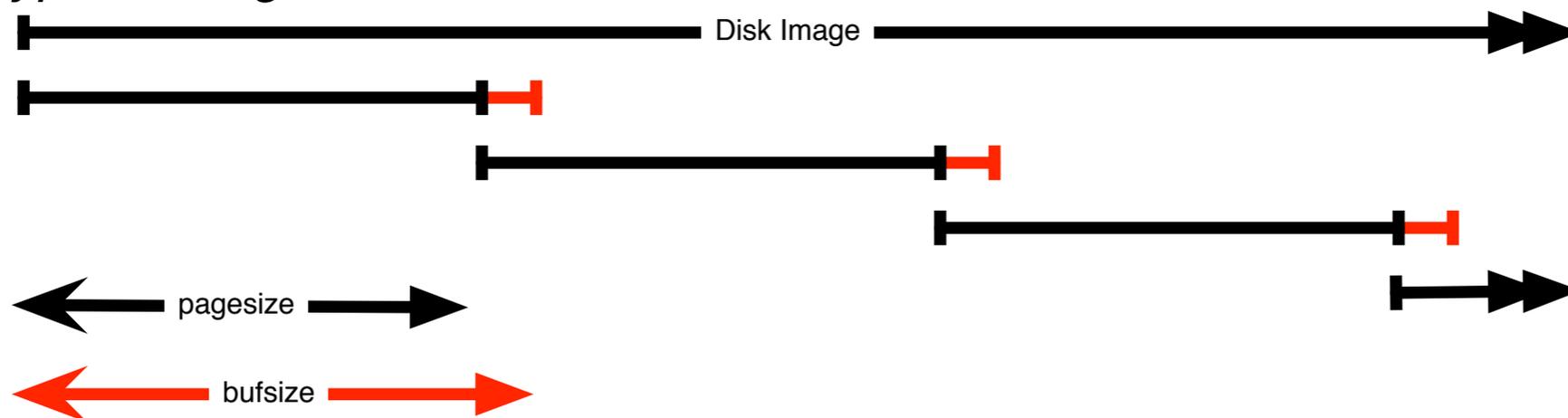


We chop the 1TB disk into 65,536 x 16MiB “pages” for processing.

# The “pages” overlap to avoid dropping features that cross buffer boundaries.

The overlap area is called the *margin*.

- Each sbuf can be processed in parallel — they don't depend on each other.
- Features start in the page but end in the margin are *reported*.
- Features that start in the margin are *ignored* (we get them later)
  - Assumes that the feature size is smaller than the margin size.
  - Typical margin: 1MB



Entire system is automatic:

- Image\_process iterator makes **sbuf\_t** buffers.
- Each buffer is processed by every scanner
- Features are automatically combined.

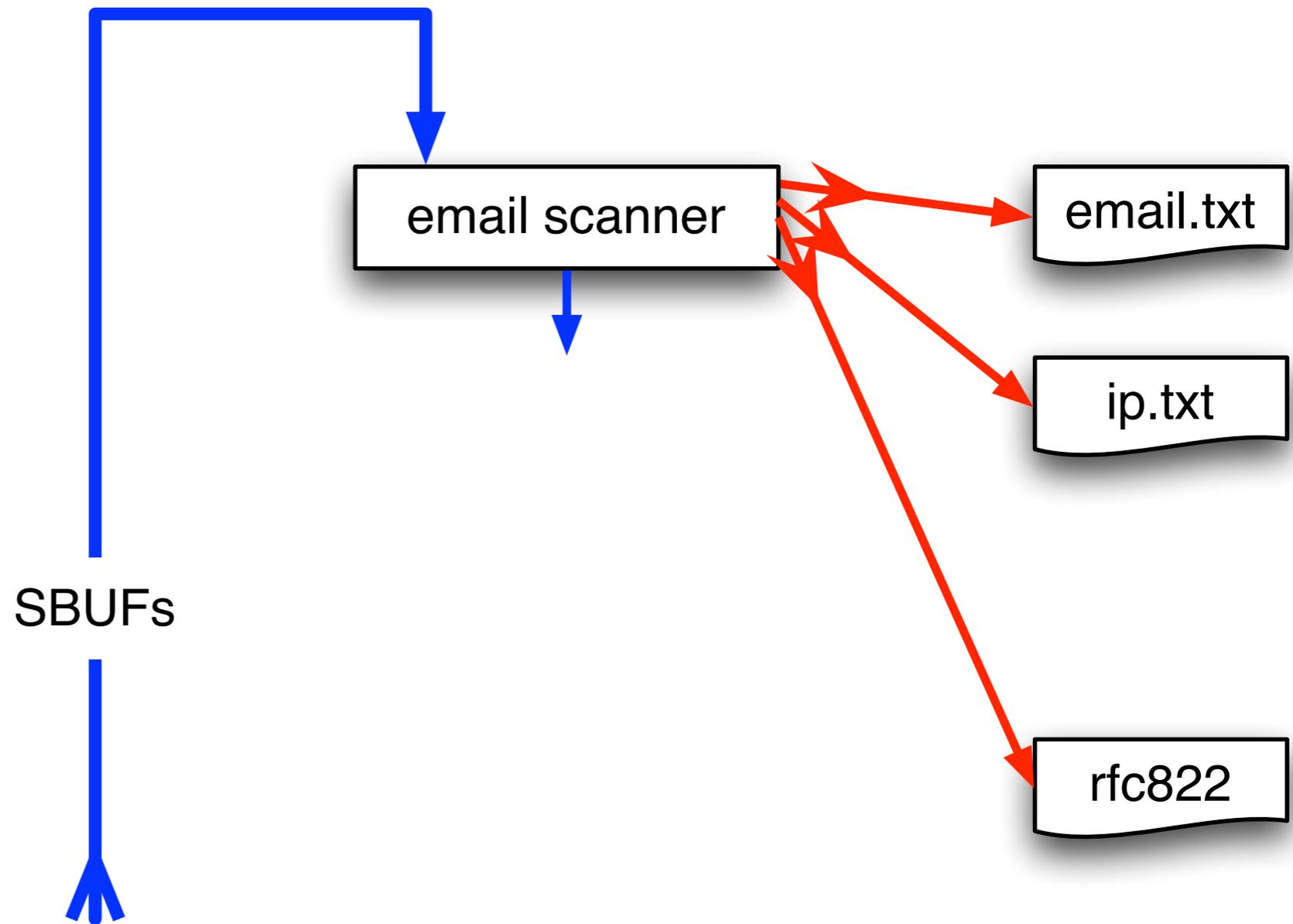
# Scanners process an sbuf and extract features

scan\_email is the email scanner.

- inputs: **sbuf** objects

outputs:

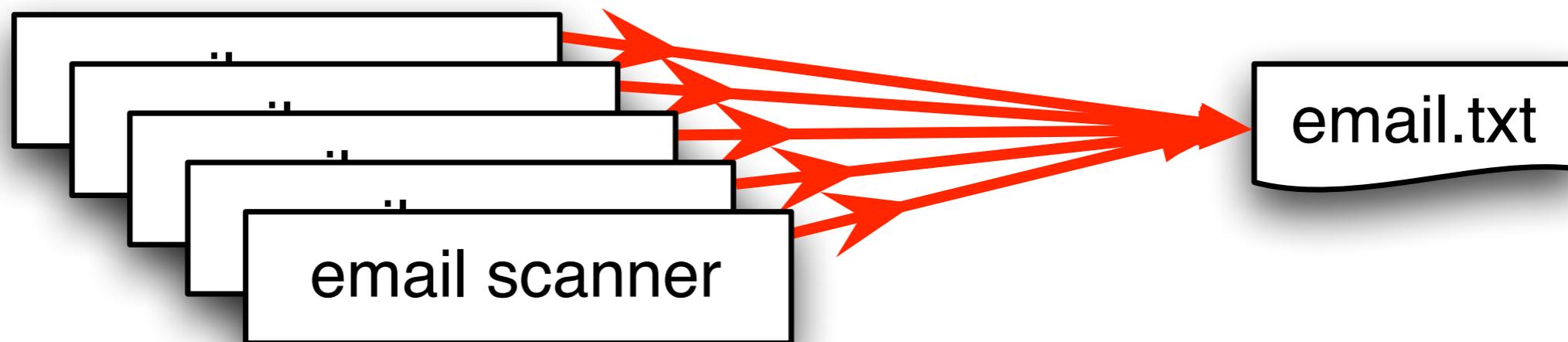
- **email.txt**
  - *Email addresses*
- **rfc822.txt**
  - *Message-ID*
  - *Date:*
  - *Subject:*
  - *Cookie:*
  - *Host:*
- **domain.txt**
  - *IP addresses*
  - *host names*



# The *feature recording system* saves features to disk.

*Feature Recorder* objects store the features.

- Scanners are given a (feature\_recorder \*) pointer
- Feature recorders are *thread safe*.



Features are stored in a *feature file*:

48198832	<a href="mailto:domexuser2@gmail.com">domexuser2@gmail.com</a>	tocol> ____ <name> <a href="mailto:domexuser2@gmail.com">domexuser2@gmail.com</a> /Home</name> ____
48200361	<a href="mailto:domexuser2@live.com">domexuser2@live.com</a>	tocol> ____ <name> <a href="mailto:domexuser2@live.com">domexuser2@live.com</a> </name> ____ <pass
48413829	<a href="mailto:siege@preoccupied.net">siege@preoccupied.net</a>	siege) O'Brien < <a href="mailto:siege@preoccupied.net">siege@preoccupied.net</a> >_hp://meanwhi
48481542	<a href="mailto:daniilo@gnome.org">daniilo@gnome.org</a>	Daniilo __egan < <a href="mailto:daniilo@gnome.org">daniilo@gnome.org</a> >_Language-Team:
48481589	<a href="mailto:gnom@prevod.org">gnom@prevod.org</a>	: Serbian (sr) < <a href="mailto:gnom@prevod.org">gnom@prevod.org</a> >_MIME-Version:
49421069	<a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a>	server2.name", " <a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a> ");__user_pref("
49421279	<a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a>	er2.userName", " <a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a> ");__user_pref("
49421608	<a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a>	tp1.username", " <a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a> ");__user_pref("

offset

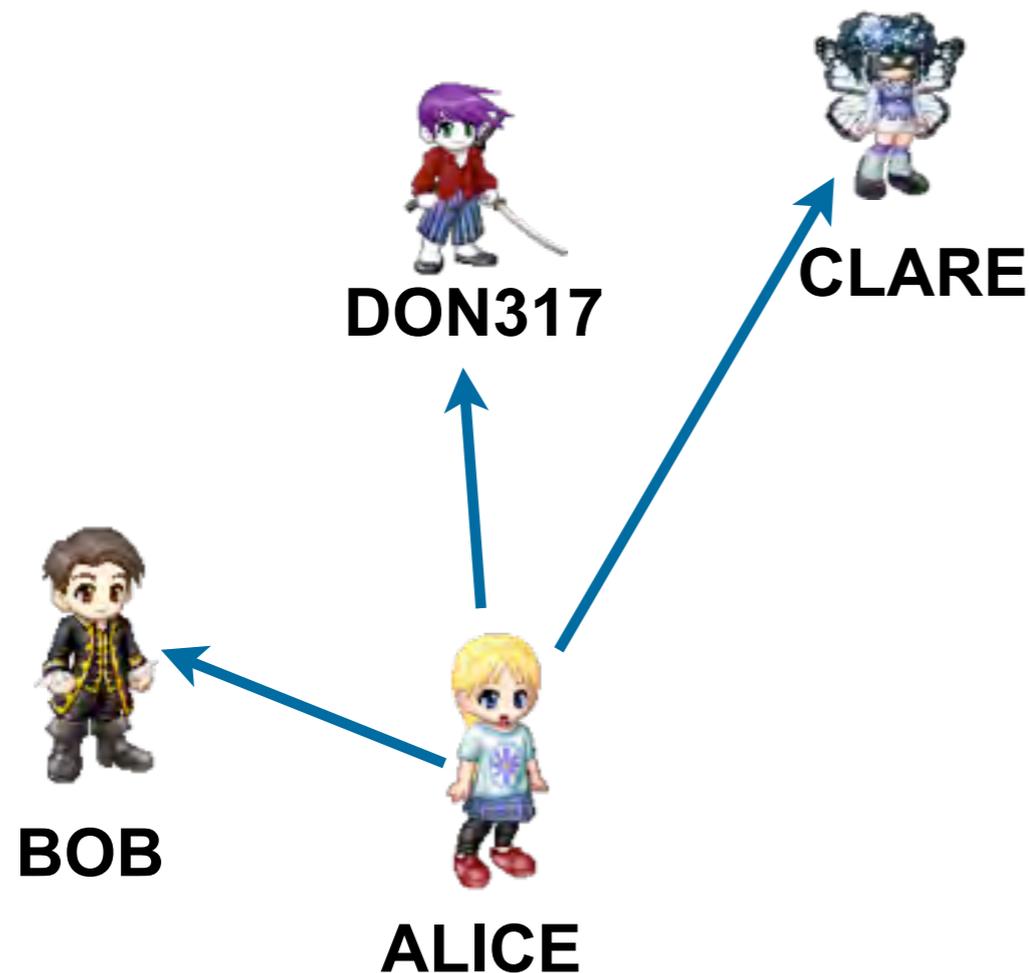
feature

feature in evidence context

# Histograms are a powerful tool for understanding evidence.

Email histogram allows us to rapidly determine:

- Drive's primary user
- User's organization
- Primary correspondents
- Other email addresses



## Drive #51 (Anonymized)

ALICE@DOMAIN1.com	8133
BOB@DOMAIN1.com	3504
ALICE@mail.adhost.com	2956
JobInfo@alumni-gsb.stanford.edu	2108
CLARE@aol.com	1579
DON317@earthlink.net	1206
ERIC@DOMAIN1.com	1118
GABBY10@aol.com	1030
HAROLD@HAROLD.com	989
ISHMAEL@JACK.wolfe.net	960
KIM@prodigy.net	947
ISHMAEL-list@rcia.com	845
JACK@nmlink.com	802
LEN@wolfenet.com	790
natcom-list@rcia.com	763

# The feature recording system *automatically* makes histograms.

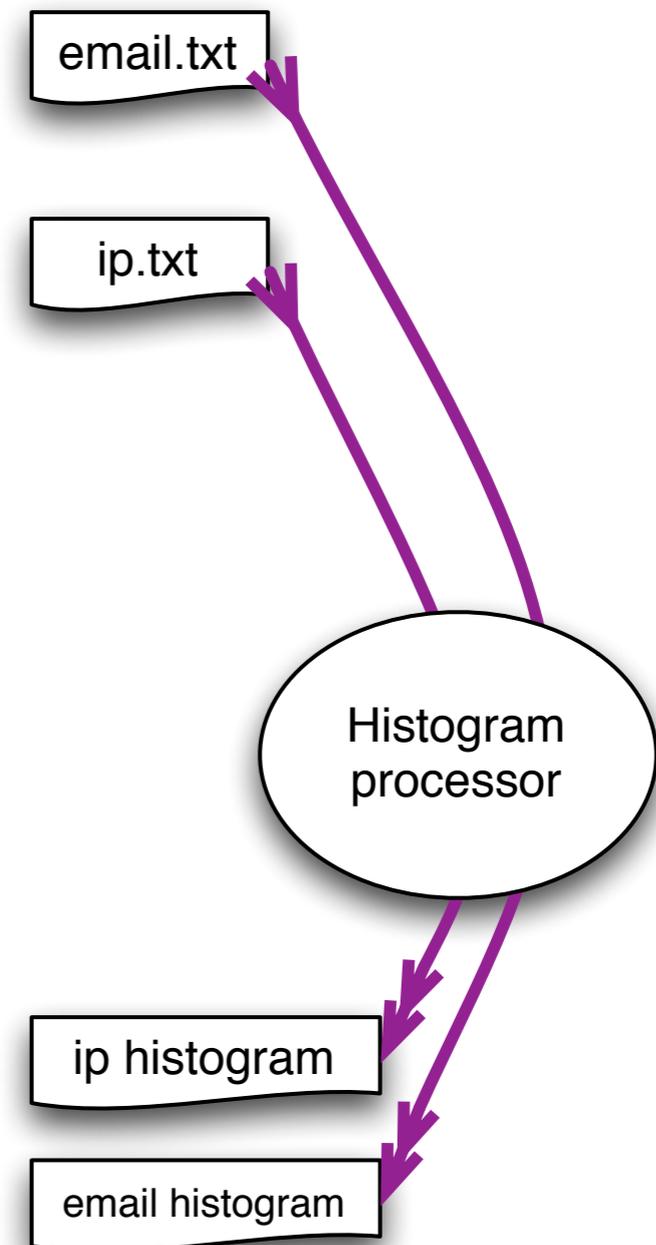
Simple histogram based on feature:

n=579	<u><a href="#">domexuser1@gmail.com</a></u>
n=432	<u><a href="#">domexuser2@gmail.com</a></u>
n=340	<u><a href="#">domexuser3@gmail.com</a></u>
n=268	<u><a href="#">ips@mail.ips.es</a></u>
n=252	<u><a href="#">premium-server@thawte.com</a></u>
n=244	<u><a href="#">CPS-requests@verisign.com</a></u>
n=242	<u><a href="#">someone@example.com</a></u>

Based on regular expression extraction:

- For example, extract search terms with `.*search.*q=(.*)`

n=18	pidgin
n=10	hotmail+thunderbird
n=3	Grey+Gardens+cousins
n=3	dvd
n=2	%TERMS%
n=2	cache:
n=2	p
n=2	pi
n=2	pid
n=1	Abolish+income+tax
n=1	Brad+and+Angelina+nanny+help
n=1	Build+Windmill
n=1	Carol+Alt



# bulk\_extractor has *multiple* feature extractors. Each scanner runs in order. (Order doesn't matter.)

## Scanners can be turned on or off

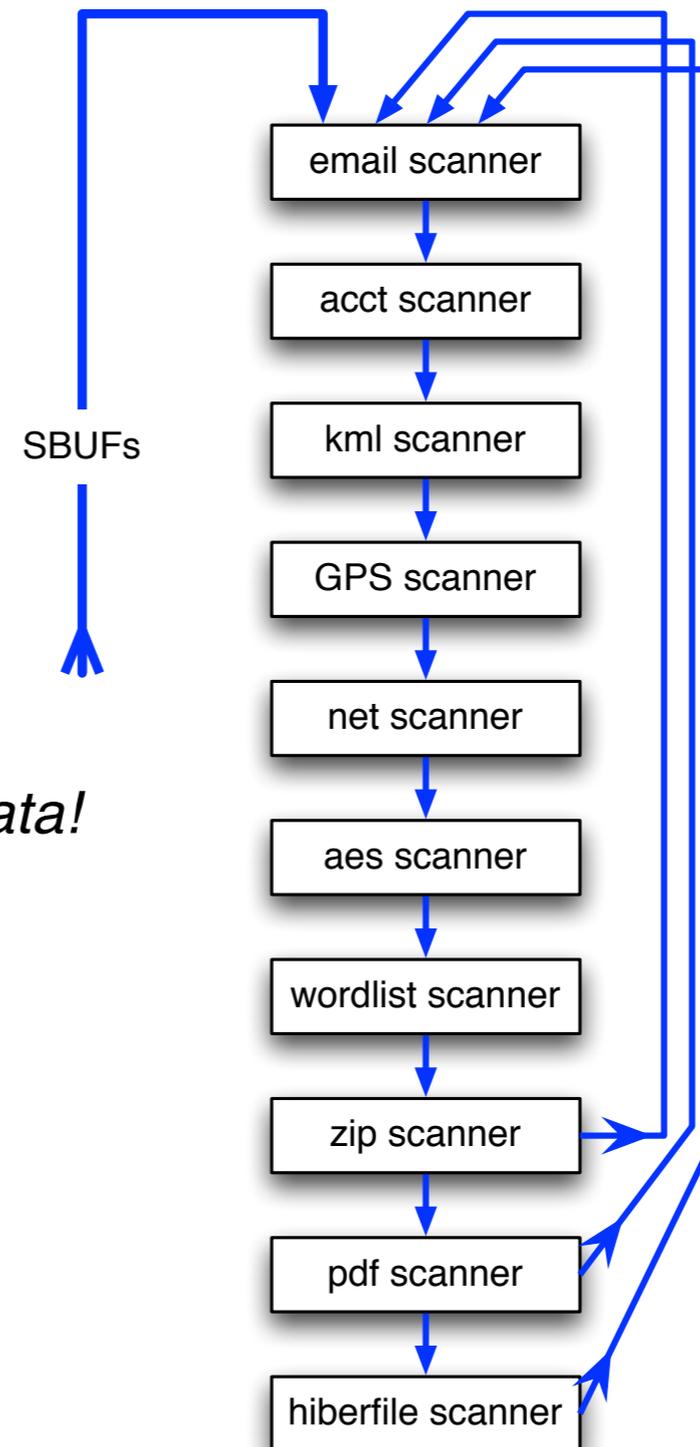
- Useful for debugging.
- AES key scanner is *very slow* (off by default)

## Some scanners are *recursive*.

- *e.g.* scan\_zip will find zlib-compressed regions
- An **sbuf** is made for the decompressed data
- The data is re-analyzed by the other scanners
  - *This finds email addresses in compressed data!*

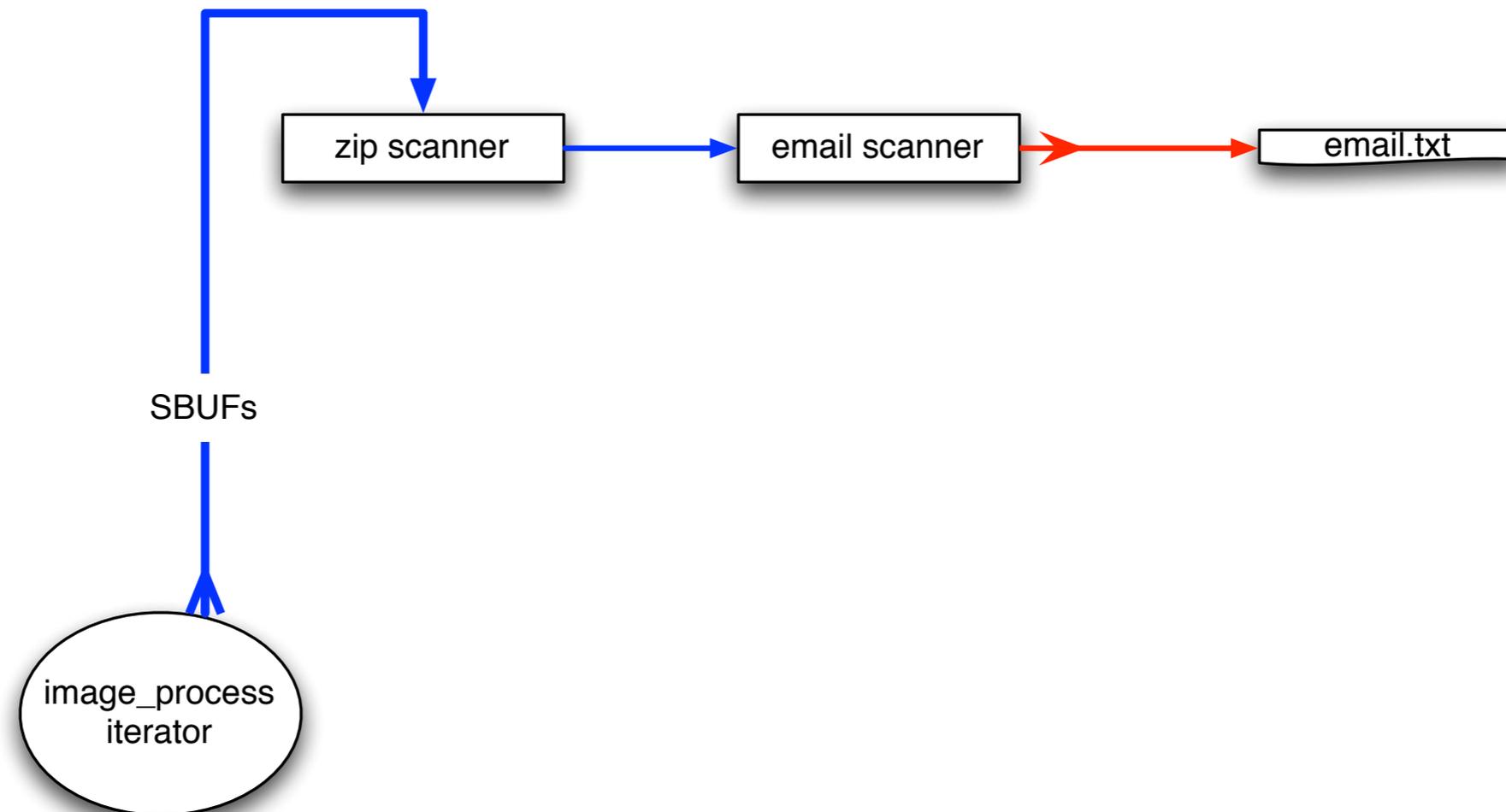
## Recursion used for:

- Decompressing ZLIB, Windows HIBERFILE,
- Extracting text from PDFs
- Handling compressed browser cache data



# Recursion requires a *new way* to describe offsets. bulk\_extractor introduces the “forensic path.”

Consider an HTTP stream that contains a GZIP-compressed email:



We can represent this as:

```
11052168704-GZIP-3437 live.com eMn='domexuser1@live.com';var srf_sDispM
11052168704-GZIP-3475 live.com pMn='domexuser1@live.com';var srf_sPreCk
11052168704-GZIP-3512 live.com eCk='domexuser1@live.com';var srf_sFT='<
```

# GUI: 100% Java

## Launches bulk\_extractor; views results

Uses bulk\_extractor to decode forensic path

The screenshot shows the Bulk Extractor Viewer application window. On the left, a 'Reports' tree view lists various files under the 'regress-04' folder, with 'email\_histogram.txt' selected. On the right, the 'Feature Filter' pane displays a list of email addresses and their counts, such as 'domexuser1@gmail.com' (n=589). Below that, the 'Referenced Feature File' pane shows a table of references to 'email.txt'.

Annotations on the left side of the image include:

- Four boxes labeled 'email.txt', 'ip.txt', 'kml.txt', and 'rfc822' with arrows pointing to the 'email\_histogram.txt' file in the Reports list.
- Two boxes labeled 'ip histogram' and 'email histogram' with arrows pointing to a star-shaped icon labeled 'GUI'.

Feature File	email_histogram.txt
n=589	domexuser1@gmail.com
n=423	domexuser2@gmail.com
n=347	domexuser3@gmail.com
n=268	ips@mail.ips.es
n=252	premium-server@thawte.com
n=243	CPS-requests@verisign.com
n=243	someone@example.com
n=220	domexuser2@live.com
n=194	domexuser1@hotmail.com
n=184	domexuser1@live.com
n=170	domexuser2@hotmail.com
n=145	inet@microsoft.com
n=115	example@passport.com
n=115	myname@msn.com
n=94	info@valcert.com
n=91	piracy@microsoft.com
n=80	certificate@trustcenter.de
n=69	hewitt@netscape.com
n=68	talkhark@mozilla.org

Referenced Feature File	email.txt
Referenced Feature	Name
1000391856	domexuser2@gmail.com
1000393445	domexuser2@live.com
1002631126	domexuser1@gmail.com
1002631186	domexuser2@gmail.com
1002631376	domexuser1@gmail.com
1002631436	domexuser2@gmail.com
1003151285	mazrob@panix.com
1004647616	privacyquestions@aol.com
10099709731	CPS-requests@verisign.com
10101693179	CPS-requests@verisign.com
1011572840	elle.uca@libero.it
10129701319	CPS-requests@verisign.com

# Crash Protection

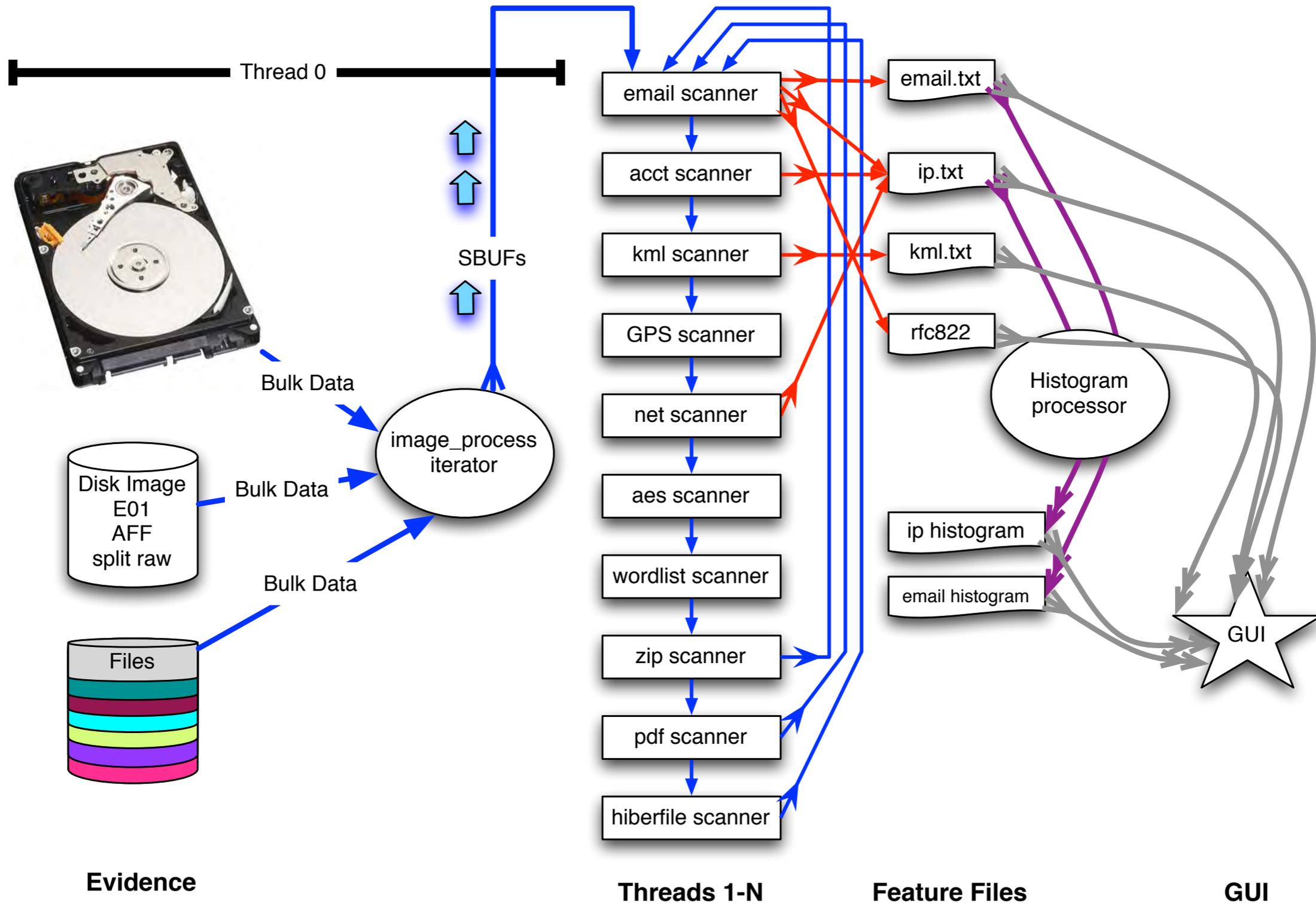
## Every forensic tool crashes.

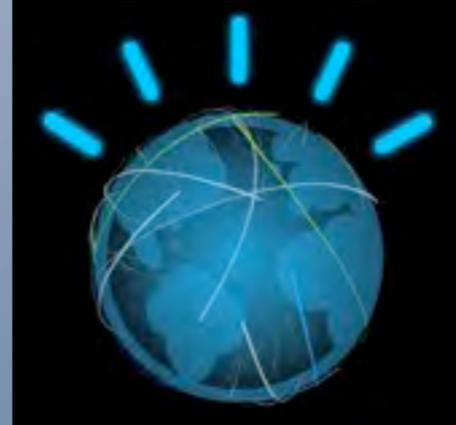
- Tools routinely used with data fragments, non-standard codings, etc.
- Evidence that makes the tool crash typically cannot be shared with the developer.

## Crash Protection: checkpointing!

- Bulk\_extractor checkpoints current page in the file config.cfg
- After a crash, just hit up-arrow and return; bulk\_extractor restarts at next page.

# Integrated design, but compact. 2726 lines of code; 33 seconds to compile on an i5





# Suppressing False Positives

# Modern operating systems are *filled* with email addresses.

## Sources:

- Windows binaries
- SSL certificates
- Sample documents

n=579	<a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a>
n=432	<a href="mailto:domexuser2@gmail.com">domexuser2@gmail.com</a>
n=340	<a href="mailto:domexuser3@gmail.com">domexuser3@gmail.com</a>
n=268	<a href="mailto:ips@mail.ips.es">ips@mail.ips.es</a>
n=252	<a href="mailto:premium-server@thawte.com">premium-server@thawte.com</a>
n=244	<a href="mailto:CPS-requests@verisign.com">CPS-requests@verisign.com</a>
n=242	<a href="mailto:someone@example.com">someone@example.com</a>

It's important to suppress email addresses not relevant to the case.

Approach #1 — Suppress emails seen on many other drives.

Approach #2 — Stop list from bulk\_extractor run on clean installs.

Both of these methods *stop list* commonly seen emails.

- Operating Systems have a LOT of emails. (FC12 has 20,584!)
- Problem: this approach gives Linux developers a free pass!



# Approach #3: Context-sensitive stop list.

Instead of a stop list of features, use features+context:

- Offset: **351373329**
- Email: **zeeshan.ali@nokia.com**
- Context: **ut\_Zeeshan Ali <zeeshan.ali@nokia.com>, Stefan Kost <**
  
- Offset: **351373366**
- Email: **stefan.kost@nokia.com**
- Context: **>, Stefan Kost <stefan.kost@nokia.com>\_\_\_\_\_sin**

— Here "context" is 8 characters on either side of feature.

— We put the feature+context in the stop list.

The “Stop List” entry is the feature+context.

- This ignores Linux developer email address in Linux binaries.
- The email address is reported if it appears in a different context.

# We created a context-sensitive stop list for Microsoft Windows XP, 2000, 2003, Vista, and several Linux.

Total stop list: 70MB (628,792 features; 9MB ZIP file)

## Sample from the stop list:

```
tzigkeit <gord@gnu.ai.mit.edu>___* tests/demo s13/fedora12-64/domain.txt
tzigkeit <gord@gnu.ai.mit.edu>___Reported by s13/fedora12-64/domain.txt
u-emacs-request@prep.ai.mit.edu (or the corresp s13/redhat54-ent-64/domain.txt
u:/pub/rtfm/" "/ftp@rtfm.mit.edu:/pub/usenet/" " s13/redhat54-ent-64/email.txt
ub/rtfm/" "/ftp@rtfm.mit.edu:/pub/usenet/" " s13/redhat54-ent-64/domain.txt
udson <ghudson@mit.edu>','_ "lefty" s13/redhat54-ent-64/domain.txt
ug-fortran-mode@erl.mit.edu__This list coll s13/redhat54-ent-64/domain.txt
uke Mewburn <lm@rmit.edu.au>, 931222_AC_ARG s13/fedora12-64/domain.txt
um _ * kit@expo.lcs.mit.edu */_#ifndef _As s13/redhat54-ent-64/email.txt
um _ * kit@expo.lcs.mit.edu */__#ifndef _A s13/redhat54-ent-64/email.txt
um _ * kit@expo.lcs.mit.edu */__#ifndef _S s13/redhat54-ent-64/email.txt
```

# The context-sensitive stop list prunes the OS-supplied features.

Applying it to domexusers HD image:

- # of emails found: 9143 → 4459

## without stop list

n=579 domexuser1@gmail.com  
n=432 domexuser2@gmail.com  
n=340 domexuser3@gmail.com  
n=268 ips@mail.ips.es  
n=252 premium-server@thawte.com  
n=244 CPS-requests@verisign.com  
n=242 someone@example.com  
n=237 inet@microsoft.com  
n=192 domexuser2@live.com  
n=153 domexuser2@hotmail.com  
n=146 domexuser1@hotmail.com  
n=134 domexuser1@live.com  
n=115 example@passport.com  
n=115 myname@msn.com  
n=110 ca@digsigtrust.com

## with stop list

n=579 domexuser1@gmail.com  
n=432 domexuser2@gmail.com  
n=340 domexuser3@gmail.com  
n=192 domexuser2@live.com  
n=153 domexuser2@hotmail.com  
n=146 domexuser1@hotmail.com  
n=134 domexuser1@live.com  
n=91 premium-server@thawte.com  
n=70 talkback@mozilla.org  
n=69 hewitt@netscape.com  
n=54 DOMEXUSER2@GMAIL.COM  
n=48 domexuser1%40gmail.com@imap.gmail.com  
n=42 domex2@rad.li  
n=39 lord@netscape.com  
n=37 49091023.6070302@gmail.com

You can download the list today:

- [http://afflib.org/downloads/feature\\_context.1.0.zip](http://afflib.org/downloads/feature_context.1.0.zip)

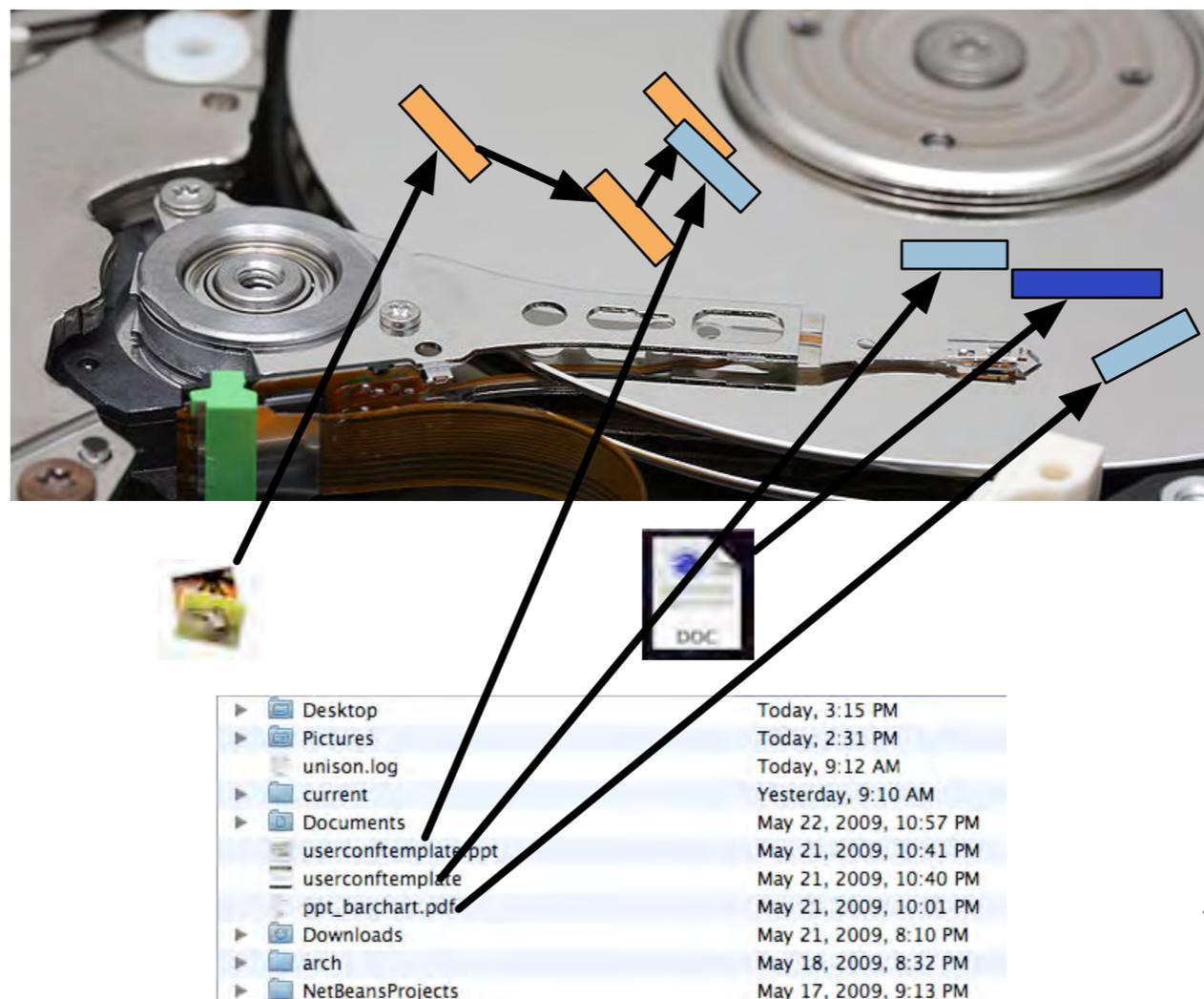
*talkback@mozilla.org and other email addresses were not eliminated because they were present on the base OS installs.*



Extending bulk\_extractor  
with Plug-ins

# Filenames can be added through post-processing.

`bulk_extractor` reports the *disk blocks* for each feature.



To get the file names, you need to map the disk block to a file.

- Make a map of the blocks in DFXML with **fiwalk** (<http://afflib.org/fiwalk>)
- Then use **python/identify\_filenames.py** to create an *annotated feature file*.

# bulk\_diff.py: compare two different bulk\_extractor reports

The “report” directory contains:

- DFXML file of bulk\_extractor run information
- Multiple feature files.

bulk\_diff.py: create a “difference report” of two bulk\_extractor runs.

- Designed for timeline analysis.
- Developed with analysts.
- Reports “what’s changed.”
  - *Reporting “what’s new” turned out to be more useful.*
  - *“what’s missing” includes data inadvertently overwritten.*





# C++ programmers can write C++ plugins

Plugins are distributed as *shared libraries*.

- Windows: **scan\_bulk.DLL**
- Mac & Linux: **scan\_bulk.so**

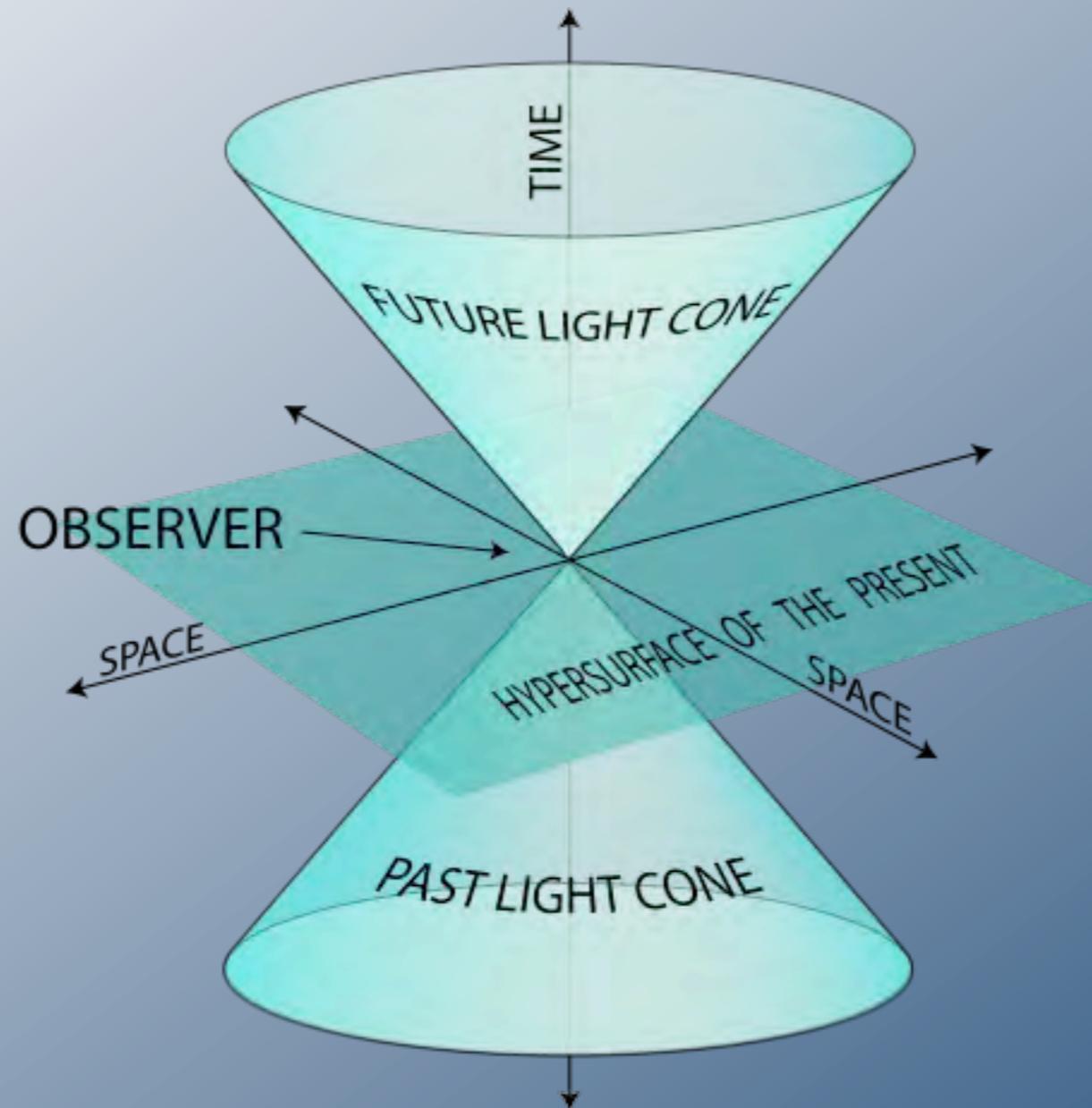
Plugins must support a single function call:

```
void scan_bulk(const class scanner_params &sp,  
              const recursion_control_block &rcb)
```

- scanner\_params — Describes what the scanner should do.
  - *sp.sbuf* — *SBUF to scan*
  - *sp.fs* — *Feature recording set to use*
  - *sp.phase==0* — *initialize*
  - *sp.phase==1* — *scan the SBUF in sp.sbuf*
  - *sp.phase==2* — *shut down*
- recursion\_control\_block — Provides information for recursive calls.

The same plug in system will be used by a future version of **fiwalk**.

- The same plug-in will be usable with multiple forensic tools.



bulk\_extractor future

# bulk\_extractor is an open source program! You can help make it better.

## Better handling of text:

- MIME decoding (e.g. user=40localhost should be user@localhost)
- Improved handling of Unicode.

## More scanners

- RAR & RAR2
- LZMA
- BZIP2
- MSI & CAB
- NTFS
- VCARD

Reliability and conformance testing.

***GET PAID TO WORK ON BULK\_EXTRACTOR: ASK ME HOW!***

# In conclusion, bulk\_extractor is a powerful stream-based forensic tool.

Bulk\_extractor demonstrates the power of:

- Bulk data processing.
- Carving EVERYTHING
- Multi-threading (we can process data with 100% CPU utilization)

Bulk\_extractor is 100% free software

- Public Domain (work of US Government)
- Please use the ideas in other programs!
  - *DFXML*
  - *Job Distribution*
  - *Forensic Path*
  - *SBUF*
- Let's keep the plug-in system consistent.
- Download from <http://afflib.org/>

Questions?

