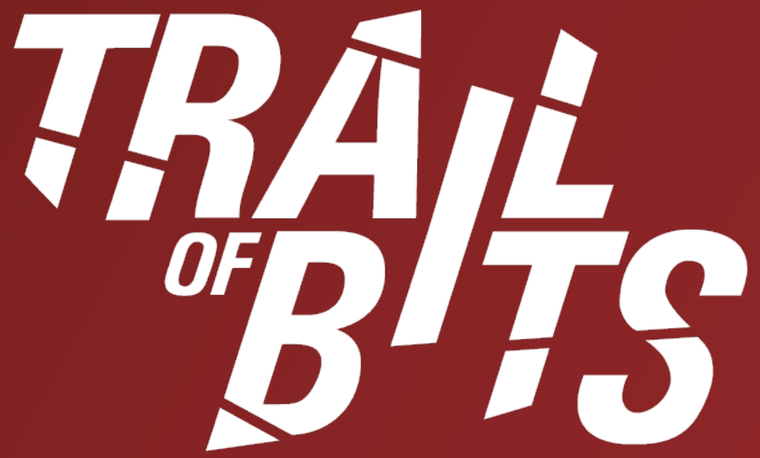


The Treachery of Files

and Two New Tools that Tame It

Evan Sultanik

The Treachery of Files

A Tragedy in Two Acts

Act I

Files Have No Intrinsic Meaning

Goal: Convince you that these funky files aren't just nifty parlor tricks

Act II

PolyFile and PolyTracker!

Goal: Introduce two new tools to help you reverse engineer files and the parsers that process them

`whoami`

Résumé: Evan A. Sultanik, Ph.D.

evan@sultanik.com
https://sultanik.com/ https://github.com/ESultanik https://twitter.com/ESultanik https://keybase.io/ESultanik

November 29th, 2017

2017–Present	Trail of Bits	Security Researcher
2012–2017	Digital Operatives	Chief Scientist
2010–2012	The Johns Hopkins University APL	Senior Research Scientist
2006–2015	Drexel University Department of Comp. Sci.	Adjunct Faculty
2001–2010	Drexel University Department of Comp. Sci.	Research Fellow
1998–2003	Feith Systems and Software	Software Engineer

The following is a chronological account of some of the things I've done and software I've created recently. I am legally constrained about talking about portions of my work (*e.g.*, work for certain government customers, code audits, and expert witness for IP litigation), so much of it is omitted here. In lieu of a traditional laundry list of technologies and tools with which I'm familiar, I hope the following will serve to demonstrate that I am a generalist with strong computer science fundamentals and the ability to quickly learn and master new technologies. I specialize in computer security, AI/ML/NLP, combinatorial optimization, and distributed systems. You might also note that I've unabashedly written this informally. Do you trust people who speak about themselves in the third person? I don't. For a more traditional résumé, my formal academic CV is available on my website.

August, 2017 Edited and co-published *PoC‖GTFO*, a nearly 800 page print collection gathering articles published in the first ten issues of the International Journal of Proof-of-Concept or GTFO. PoC‖GTFO—on which I have been an editor since August of 2015 and a contributor since June of 2014—follows in the tradition of *Phrack* and *Uninformed* by publishing on the subjects of offensive security research, reverse engineering, and file format internals. Until now, the journal has only been available online or printed and distributed for free at conferences worldwide. ISBN-13: 978-1-59327-880-9 https://www.nostarch.com/gtfo

October, 2015–August, 2017 Member of the "control team" of expert *human* vulnerability researchers for the DARPA VET program, the purpose of which was to act as a baseline against which to compare the *automated* vulnerability analysis tools of the other performers. This consisted of using state-of-the-art publicly available, commercial-, and government-off-the-shelf tools (*e.g.*, IDA) to determine whether a given set of black-box ARM binaries had been modified to include malice. Our team outperformed the automated equivalents across four judged engagements. Primarily IDA, HexRays, and our own static analysis tools that I helped develop (see below).

June, 2017 Published and presented an analysis of zoning and residential density changes in Philadelphia over the past five years. This required devising and implementing a solution to the difficult computational geometry problem of intersecting the tens of thousands of latitude/longitude polygons describing Philadelphia's zoning districts both pre and post the comprehensive zoning overhaul that occurred in 2012. Also used Machine Learning to create a model to predict the real estate tax revenue implications of zoning changes. Implemented in Python, Java, and KML. https://www.sultanik.com/blog/ZoningDensity

June, 2017 Created a PDF that is also a valid Git repository containing its own LaTeX source code and a copy of itself[1]. Published in PoC‖GTFO Issue 0x15. Implemented in Bash, LaTeX, Python, and C. https://github.com/ESultanik/PDFGitPolyglot

June, 2017 Created a toy web-based terminal emulator and shell, backed by a filesystem served over HTTP (similar to WebDAV), in which scripts are dynamically loaded from the server via JavaScript, and replete with libraries like `ncurses`. Implemented purely in JavaScript using JQuery. https://www.sultanik.com/#term

March, 2017 Created a PDF that is also a valid Nintendo Entertainment System ROM image[2] that, when played/emulated, renders the title page of the PDF *and* displays the MD5 checksum of itself[4]. This MD5 quine is achieved by solving for 128 MD5 collisions using an implementation of an algorithm created by Marc Stevens. This polyglot was in collaboration with Evan Teran and Ange Albertini. Implemented in a mixture of C, 6502 assembly, and Python. https://www.sultanik.com/pocorgtfo/#0x14

August, 2013–November, 2016 Created a semi-automated static analysis tool for human-in-the-loop reverse engineering and vulnerability research along with four colleagues. Led the effort in researching and implementing components for static taint analysis, and devised novel algorithms for SMT constraint optimization and detecting algorithmic complexity defects. Implemented in C++11.

October, 2016 Discovered a vulnerability in the way most US defense contractors choose passwords for the spinlocks they are required to use to secure safes and rooms containing classified information. I wrote a short closet drama about it. Implemented in Bash, Python, and English Prose. https://archive.org/stream/pocorgtfo13#page/n42/mode/1up

October, 2016 Created a PDF that is also a valid PostScript file. If you send the raw PDF to a printer, it will print out differently than how it was rendered on screen in the PDF viewer. For good measure, the PostScript also reads your `/etc/passwd` file. Implemented in Python and PostScript. https://www.sultanik.com/pocorgtfo/#0x13

August, 2016 Along with three collaborators, created an AI bot capable of playing Pokémon Go. Unlike other bots of the time which exploited bugs in Niantic's implementation to achieve superhuman feats like teleportation to increase their speed,

[1] `unzip ESultanikResume.pdf ESultanikResume/PDFGitPolyglot.pdf`
[2] This PDF is also a valid Nintendo Entertainment System ROM[3]. Try emulating it!
[3] By the way, the NES ROM version of this PDF also prints out its MD5.
[4] By the way, the MD5 hash of this PDF is `AFFAB1EDA8E9B6D8A80E940F20CB3B3B`[5]
[5] By the way, the first seven characters of that MD5 hash are not a coincidence, either.

`whoami`

---

[1] `unzip ESultanikResume.pdf ESultanikResume/PDFGitPolyglot.pdf`

[2] This PDF is also a valid Nintendo Entertainment System ROM[3]. Try emulating it!

[3] By the way, the NES ROM version of this PDF also prints out its MD5.

[4] By the way, the MD5 hash of this PDF is `AFFAB1EDA8E9B6D8A80E940F20CB3B3B`.[5]

[5] By the way, the first seven characters of that MD5 hash are not a coincidence, either.

`whoami`

---

[1] `unzip ESultanikResume.pdf ESultanikResume/PDFGitPolyglot.pdf`

[2] This PDF is also a valid Nintendo Entertainment System ROM[3]. Try emulating it!

[3] By the way, the NES ROM version of this PDF also prints out its MD5.

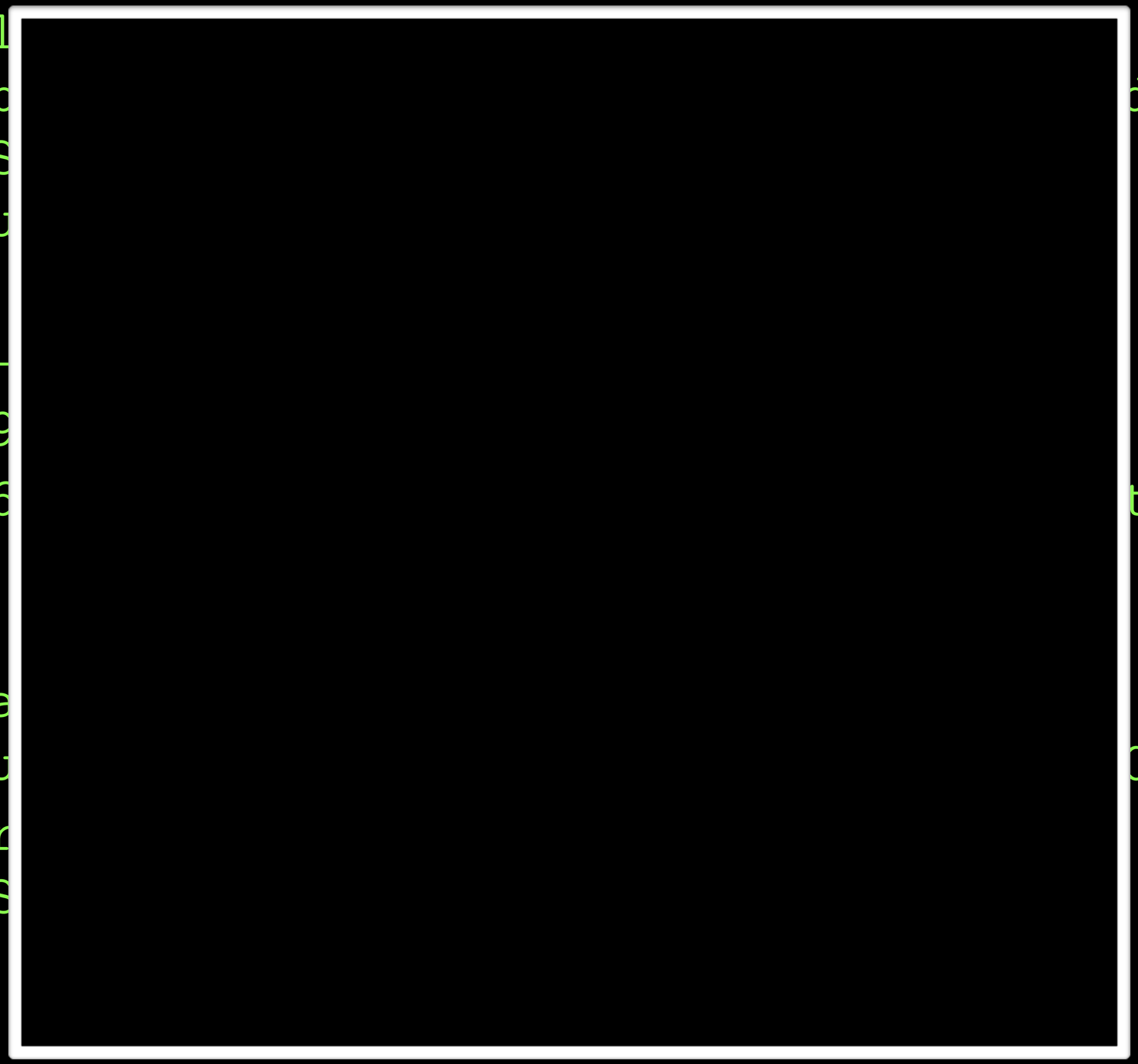
[4] By the way, the MD5 hash of this PDF is `AFFAB1EDA8E9B6D8A80E940F20CB3B3B`.[5]

[5] By the way, the first seven characters of that MD5 hash are not a coincidence, either.

```
$ md5sum ESultanikResume.pdf
affab1eda8e9b6d8a80e940f20cb3b3b  ESultanikResume.pdf
```

```
$ md5sum ESultanikResume.pdf
affab1eda8e9b6d8a80e940f20cb3b3b  ESultanikResume.pdf
$ unzip -l ESultanikResume.pdf
Archive:  ESultanikResume.pdf
  Length      Date    Time    Name
---------  ---------- -----   ----
        0  09-17-2019 14:47   ESultanikResume/
   638270  06-07-2019 10:35   ESultanikResume/PDFGitPolyglot.pdf
---------                     -------
   638270                     2 files
```

```
$ md5sum ESultanikResume.pdf
affab1eda8e9b6d8a80e940f20cb3b3b  ESultanikResume.pdf
$ unzip -l ESultanikResume.pdf
Archive:  ESultanikResume.pdf
  Length      Date    Time    Name
---------  ---------- -----   ----
        0  09-17-2019 14:47   ESultanikResume/
   638270  06-07-2019 10:35   ESultanikResume/PDFGitPolyglot.pdf
---------                     -------
   638270                     2 files
$ file ESultanikResume.pdf
ESultanikResume.pdf: NES ROM image (iNES): 8x16k PRG, 2x8k
CHR [V-mirror] [SRAM] [Trainer]
```


$ md5sum ESul
affab1eda8e9b
df
$ unzip -l ES
Archive: ESu
  Length
--------- --
        0 09
   638270 06
tPolyglot.pdf

   638270
$ file ESulta
ESultanikResu
G, 2x8k
CHR [V-mirror
$ nestopia ES

Act I

Files Have No Intrinsic Meaning

PoC||GTFO

https://sultanik.com/pocorgtfo/

PoC*or*GTFO

Proof of Concept

(Pictures of Cats)

"It looks great on a shelf, and if you read PoC||GTFO on public transportation, people stay away from you."

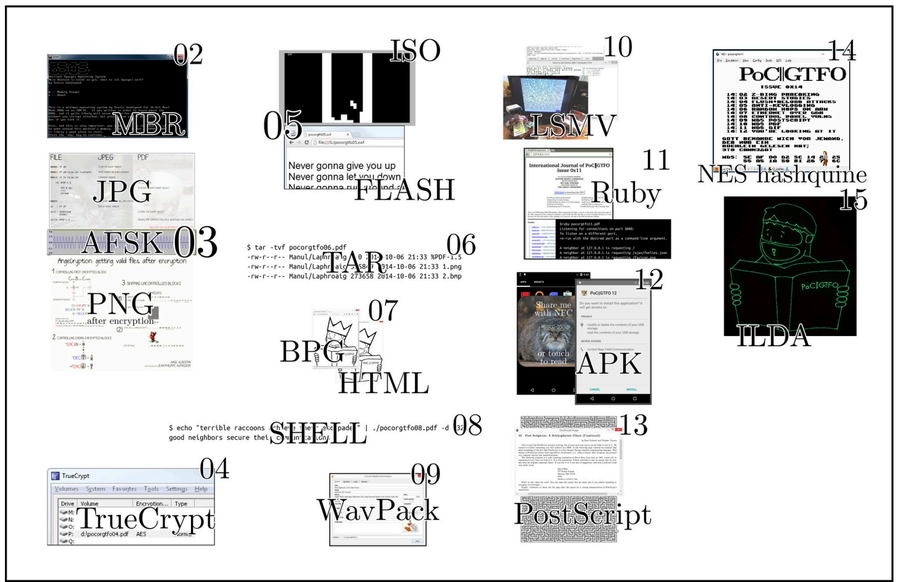
—Hackaday Review

Roughly quarterly journal, in the tradition of Phrack and Uninformed

Offensive security research and stunt hacking

First released on paper at a conference, later released digitally

Each digital release is a Polyglot



https://sultanik.com/pocorgtfo/

PoC*or*GTFO

Proof of Concept

(Pictures of Cats)

"It looks great on a shelf, and if you read PoC||GTFO on public transportation, people stay away from you."

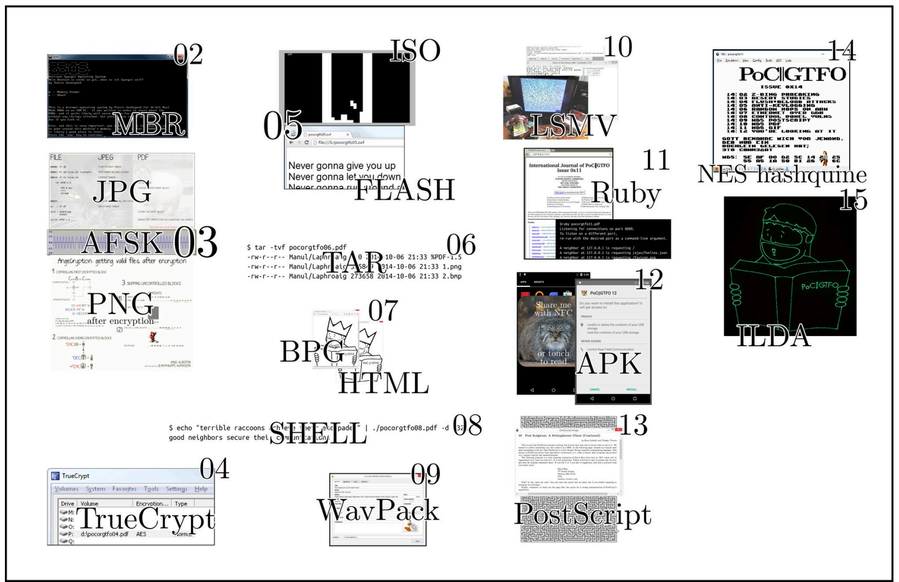
—Hackaday Review

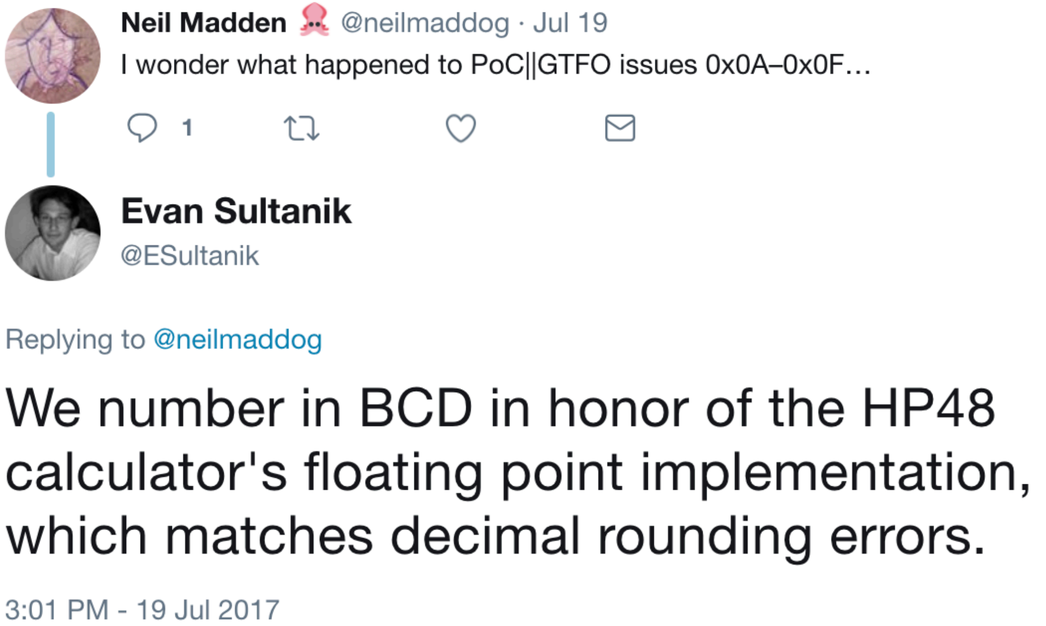
Roughly quarterly journal, in the tradition of Phrack and Uninformed

Offensive security research and stunt hacking

First released on paper at a conference, later released digitally

Each digital release is a Polyglot





https://sultanik.com/pocorgtfo/

PoC*or*GTFO

Proof of Concept

(Pictures of Cats)

"It looks great on a shelf, and if you read PoC||GTFO on public transportation, people stay away from you."

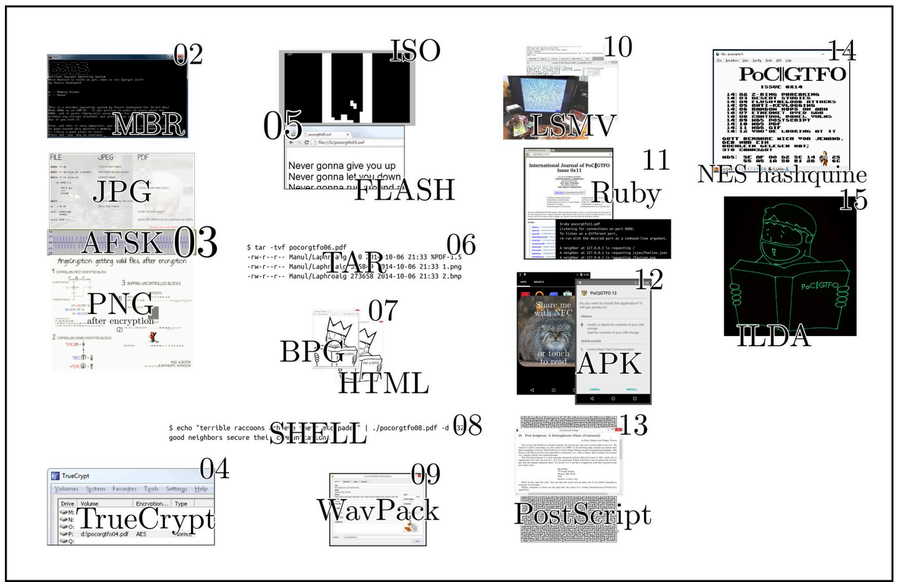
—Hackaday Review

Roughly quarterly journal, in the tradition of Phrack and Uninformed

Offensive security research and stunt hacking

First released on paper at a conference, later released digitally

Each digital release is a Polyglot



Neil Madden @neilmaddog · Jul 19

I wonder what happened to PoC||GTFO issues 0x0A–0x0F…

1

Travis Goodspeed

@travisgoodspeed

Replying to @__gbg__ @h2hconference @hacktivityconf

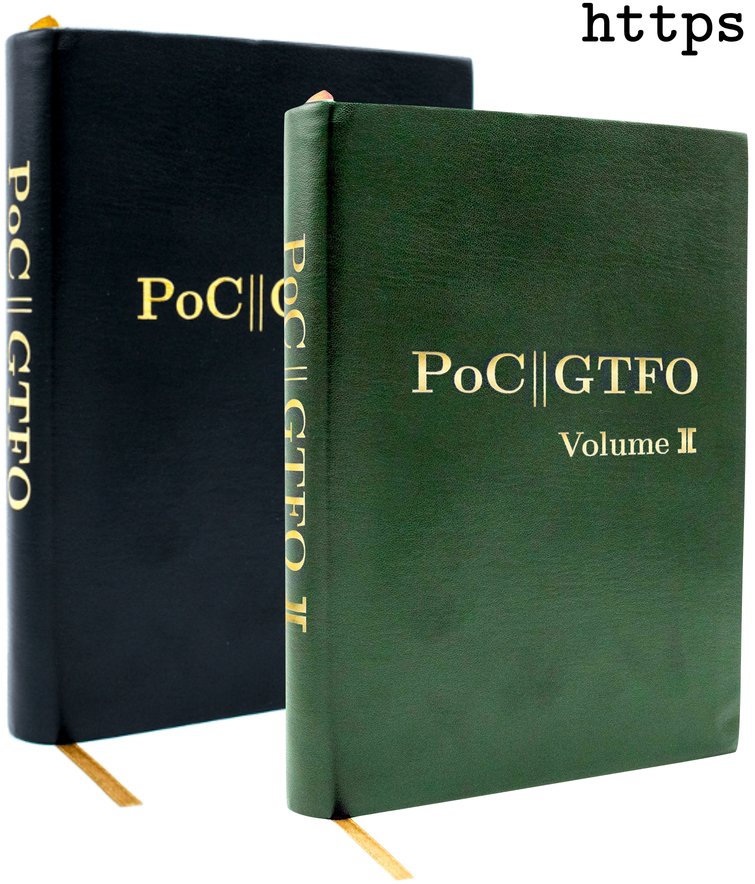
We number them in binary coded decimal, as a tribute to the floating point unit of the HP SATURN architecture.

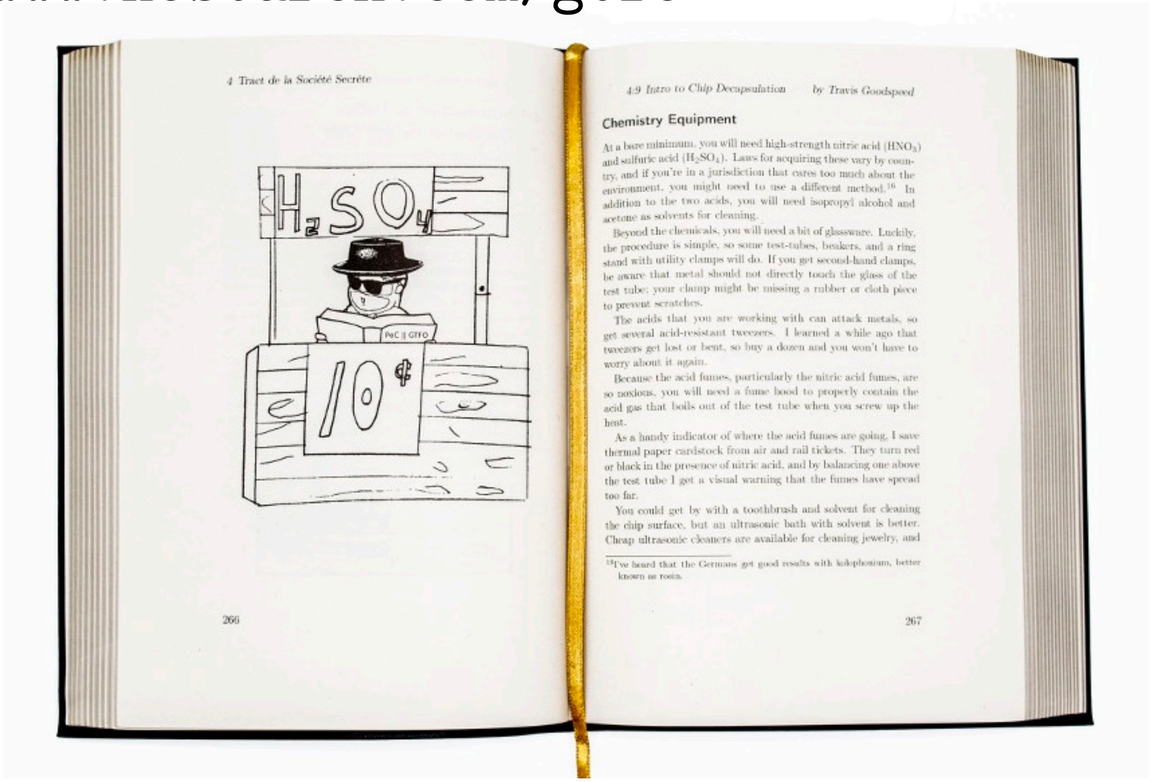
1:16 PM - 12 Oct 2017

https://sultanik.com/pocorgtfo/

The Book of PoC||GTFO

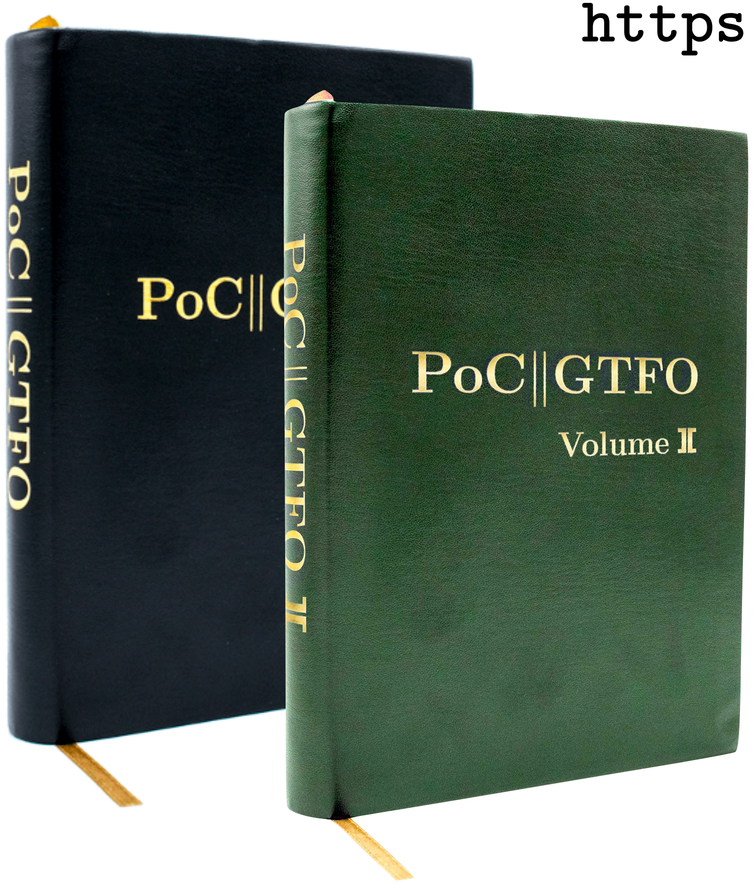
https://www.nostarch.com/gtfo





The Book of PoC||GTFO

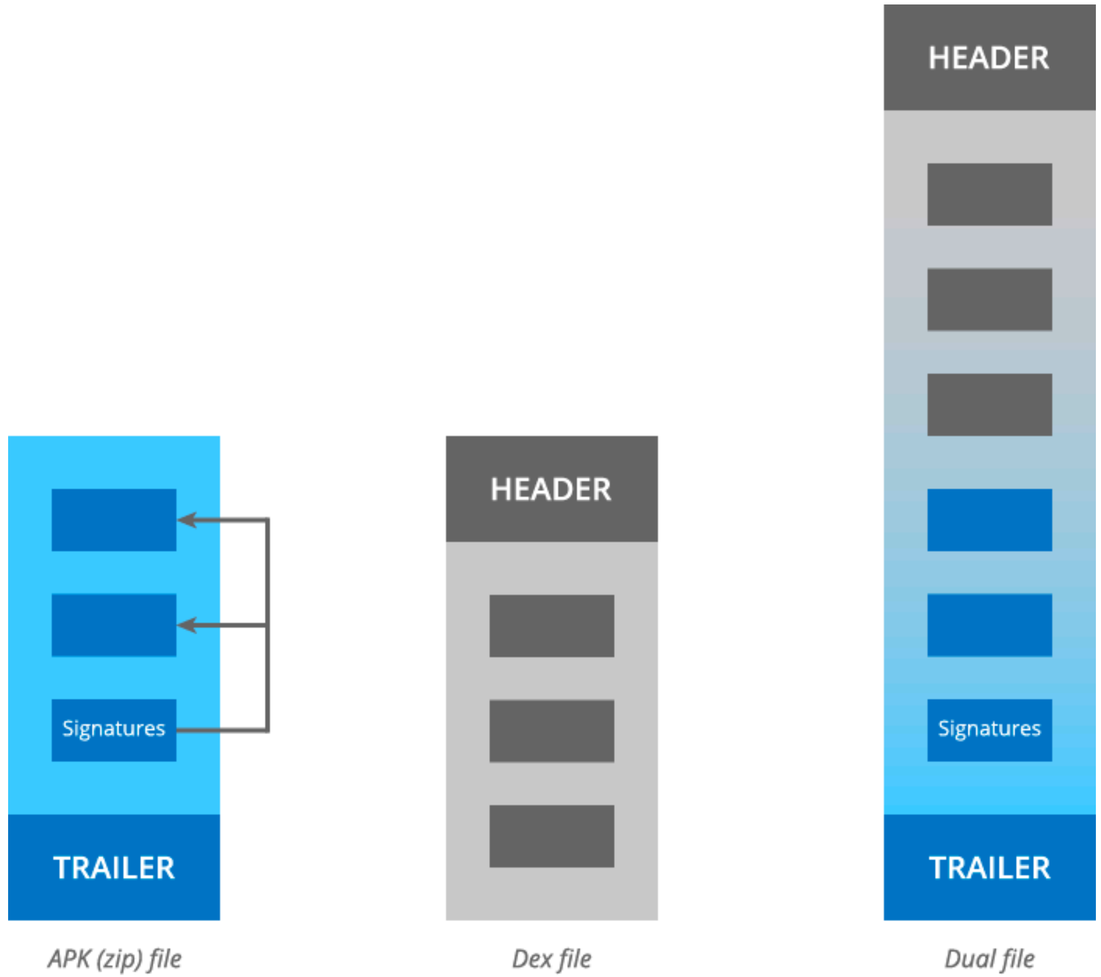
https://www.nostarch.com/gtfo



"...a file has no intrinsic meaning. The meaning of a file—its type, its validity, its contents—can be different for each parser or interpreter"

—PoC||GTFO 7:6 by Ange Albertini

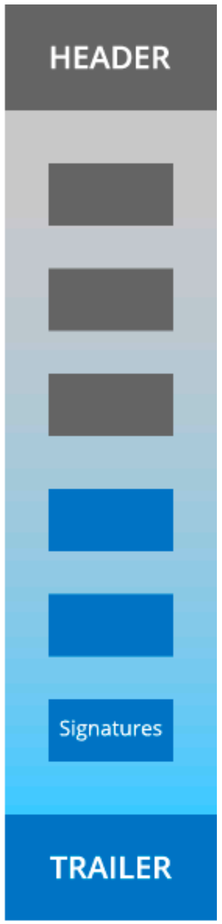
Example: Android APK (zip)/Dex Polyglot



Example: Android APK (zip)/Dex Polyglot

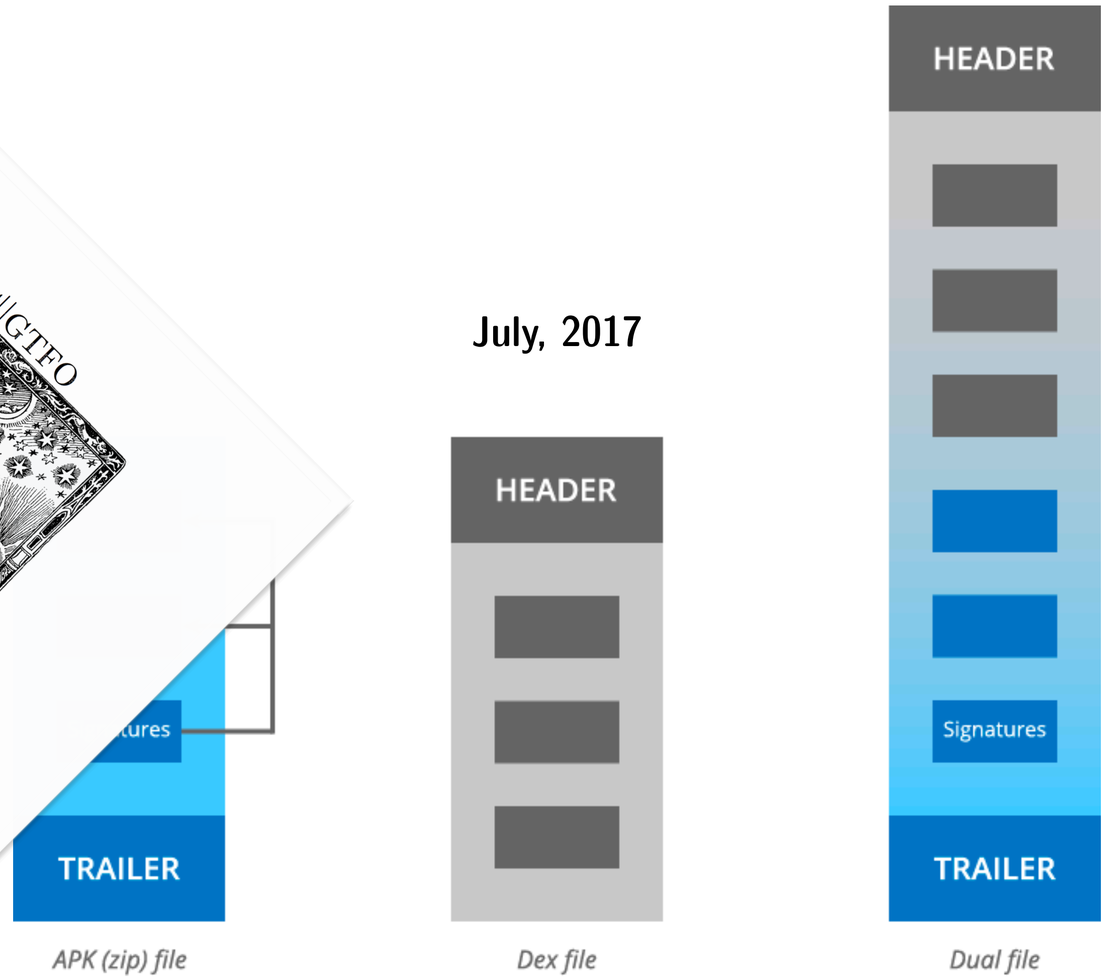
APK (zip) file

Dex file

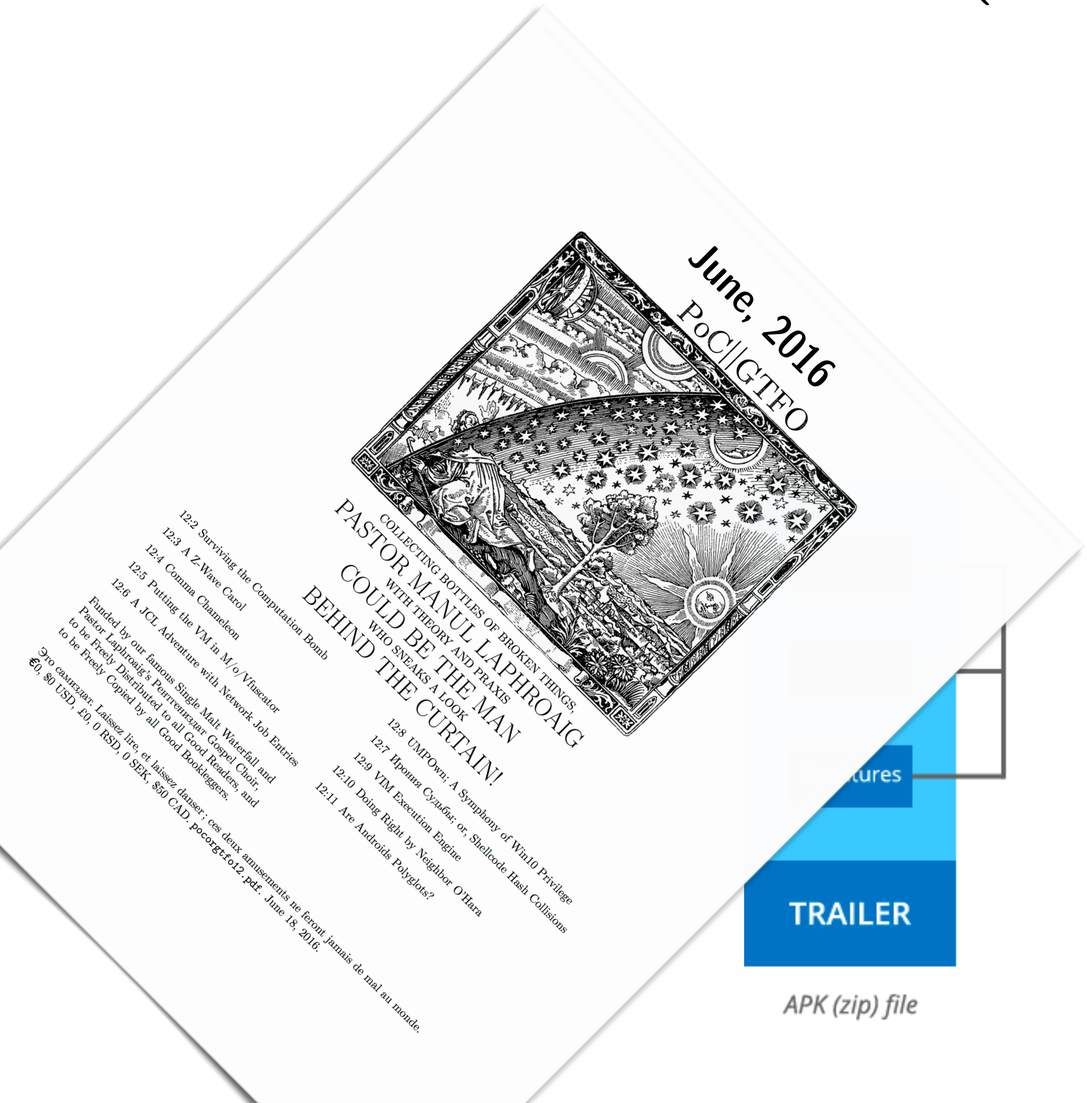


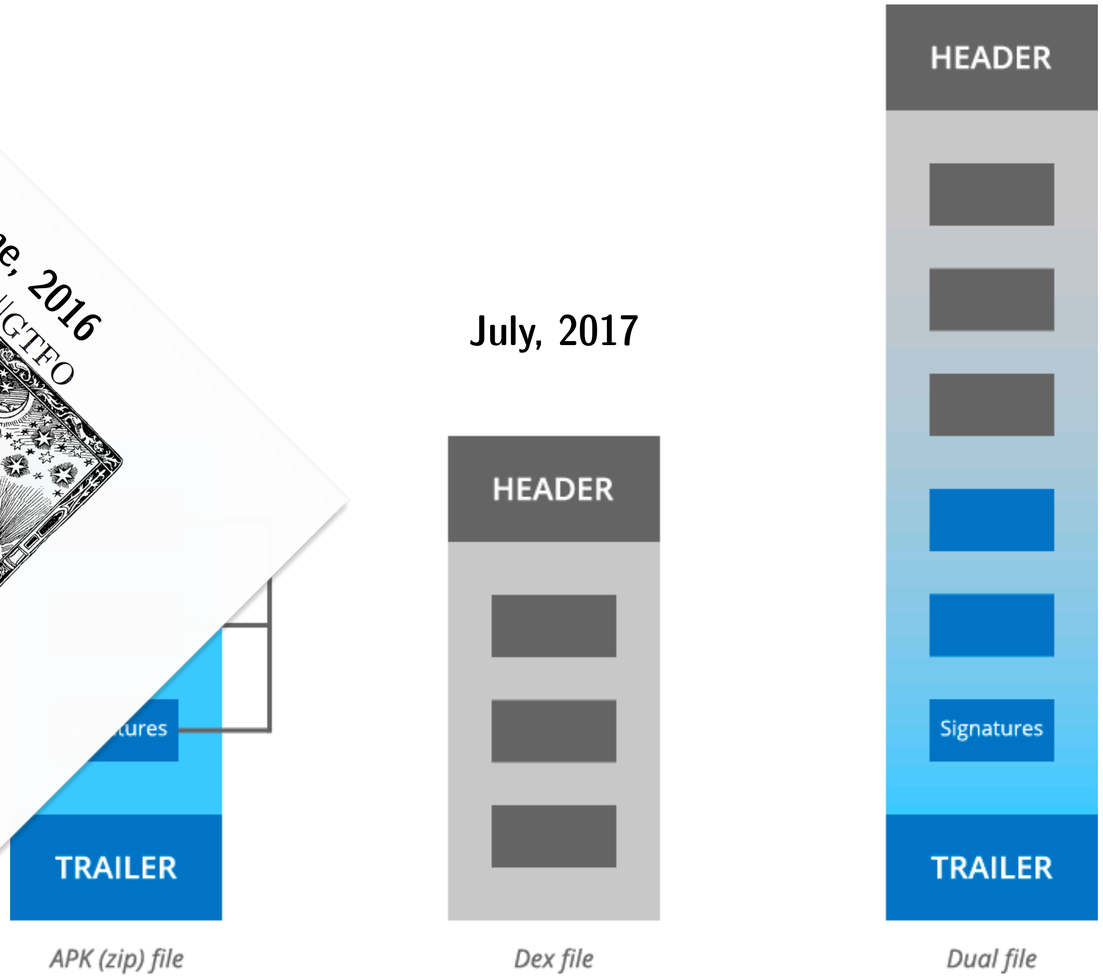
Dual file

Example: Android APK (zip)/Dex Polyglot



Example: Android APK (zip)/Dex Polyglot





What's Wrong Here?

```
$ tar xvf totally_not_malware.tar.gz
```

What's Wrong Here?

```
$ tar xvf totally_not_malware.tar.gz
```

Note: We didn't provide the `z` option!

Modern versions of `tar` automagically detect that the archive is compressed based on magic bytes!
(The actual file extension is ignored.)

What's Wrong Here?

```
$ tar xvf totally_not_malware.tar.gz
```

Note: We didn't provide the `z` option!

Modern versions of `tar` automagically detect that the archive is compressed based on magic bytes!
(The actual file extension is ignored.)

What if we created a file that is *both* a valid `.tar` *and* a valid `.tar.gz`?

totally_not_malware.tar


DEFCON

Why are PDFs Particularly Polyglottable?

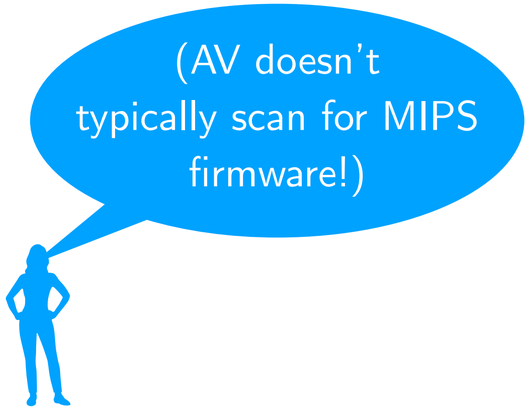
- Because "Adobe," that's why!
- It's been around for a long time
- Parsers built to be resilient to all sorts of errors and incompatibilities
- Can insert arbitrary length binary blobs almost anywhere in the file
- Almost all parsers ignore everything before the header

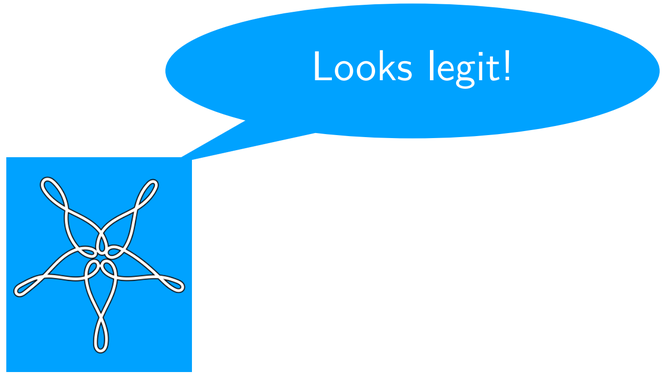
```
9999 0 obj
<<
/Length # bytes in the blob
>>
stream
lol, put whatever you want here!
endstream
endobj
```

```
lol, put whatever you want here!
%PDF-1.5
%<D0><D4><C5><D8>
        ⋮
```

Hey,
I've been trying to get my résumé to so-and-so in HR, but we've had problems with E-mail. Can you please print out the attached copy and give it to them?

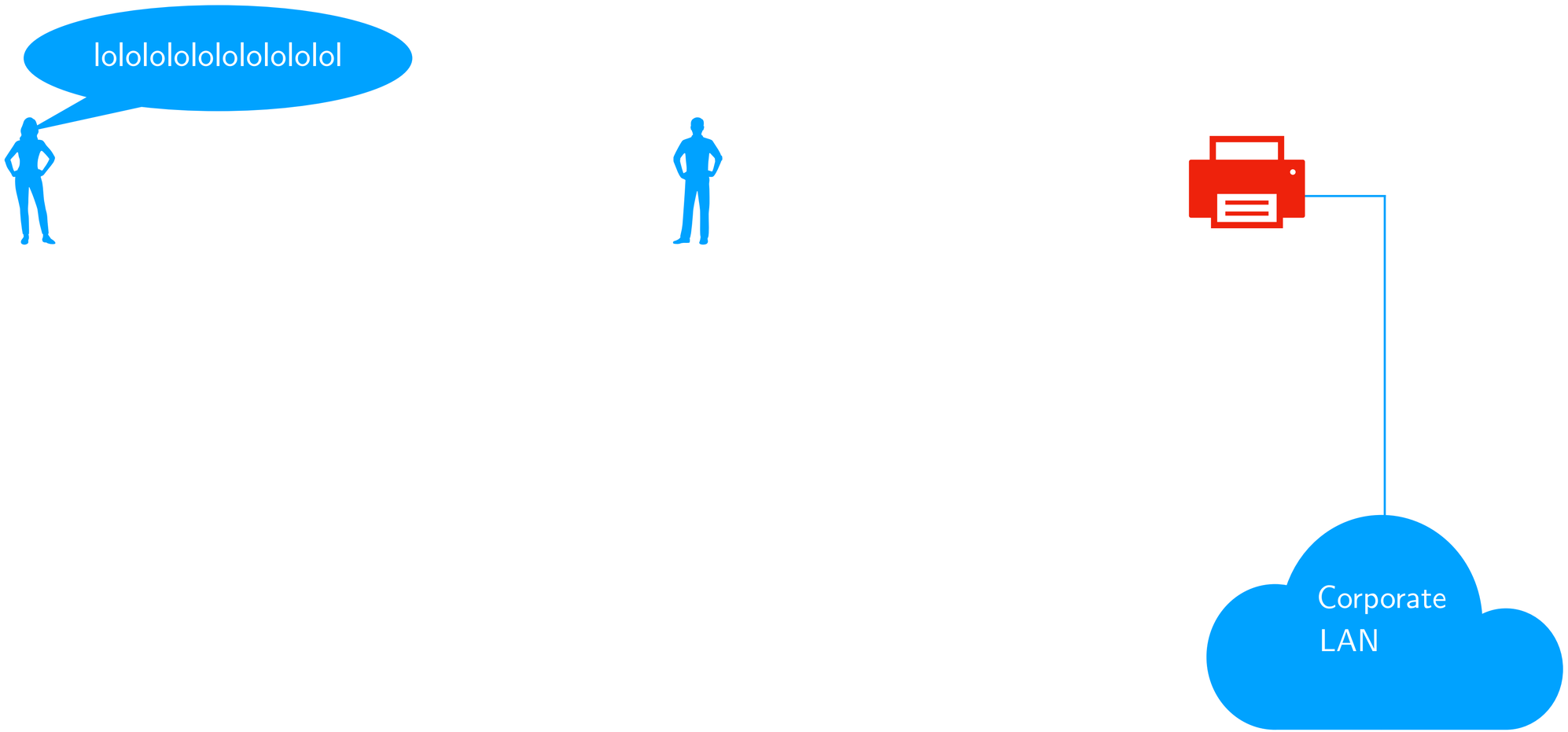
Thanks! —Alice Hackerman





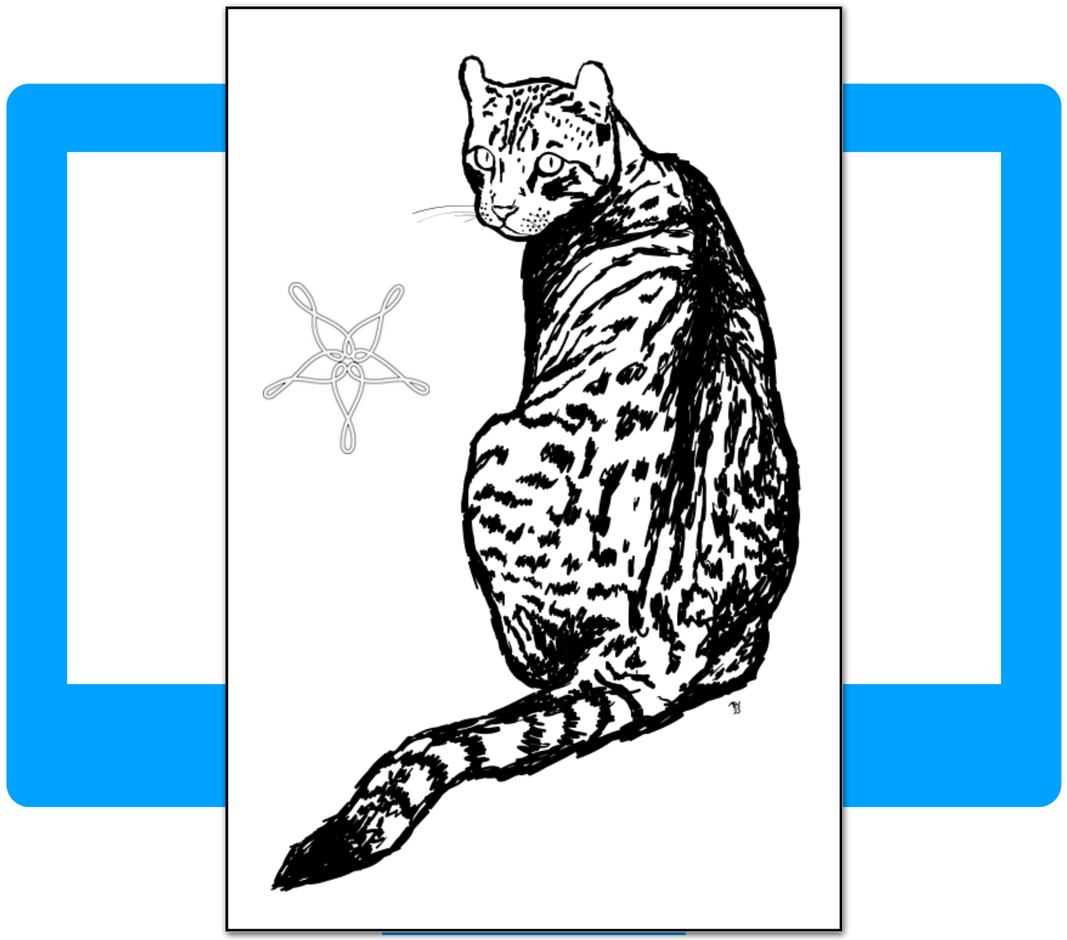
Hey,
I've been trying to get my résumé to so-and-so in HR, but we've had problems with E-mail. Can you please print out the attached copy and give it to them?

Thanks! —Alice Hackerman



PostScript/PJL (Cui & Stolfo, 2011)

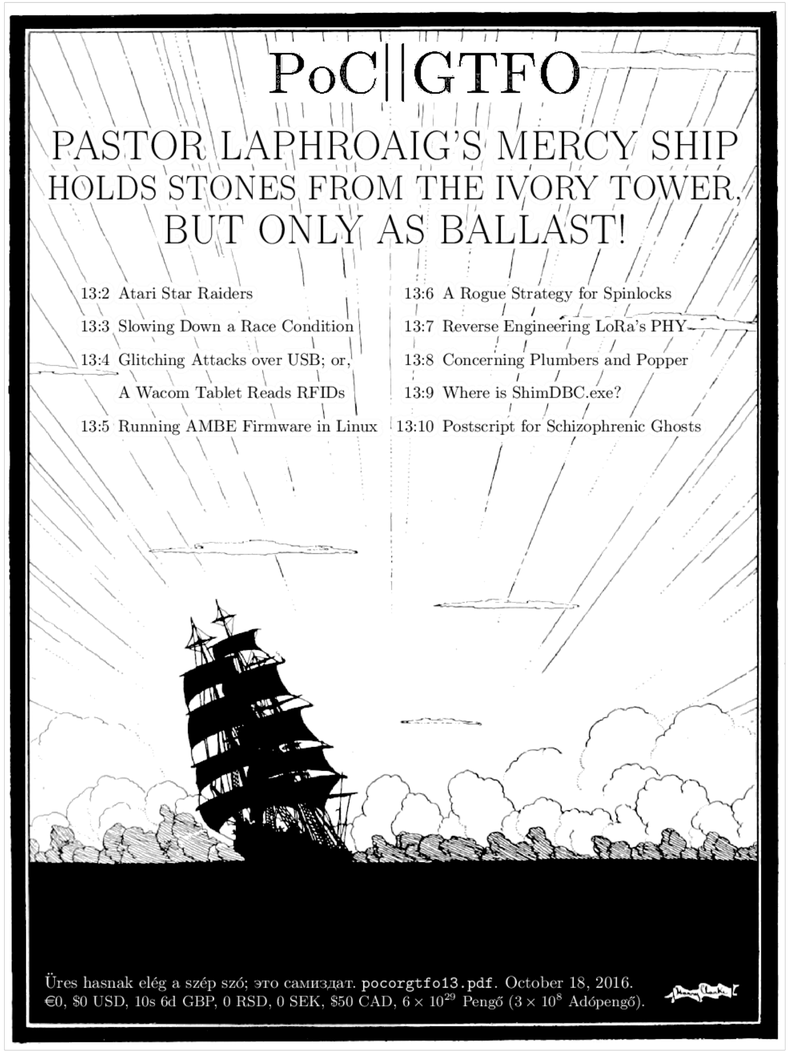
Producing a Positively Provocative PDF/PostScript Polyglot



Producing a Positively Provocative PDF/PostScript Polyglot

Don't print PostScript created by Evan!

PoC||GTFO 0x13



PoC||GTFO

PASTOR LAPHROAIG'S MERCY SHIP HOLDS STONES FROM THE IVORY TOWER, BUT ONLY AS BALLAST!

Üres hasnak elég a szép szó; это самиздат. pocorgtfo13.pdf. October 18, 2016.
€0, $0 USD, 10s 6d GBP, 0 RSD, 0 SEK, $50 CAD, 6×10^{29} Pengő (3×10^{8} Adópengő).

PDF

10 Post Scriptum: A Schizophrenic Ghost (Continued)

by Evan Sultanik and Philippe Teuwen

Once we got this PostScript polyglot working, the obvious next step was to decide what to put in it. We wanted to exhibit something you can't achieve in a PDF, so the following page contains an example that takes advantage of the fact that PostScript is a fully fledged Turing complete programming language. This feature of PostScript allows both *algorithmic compression* (*i.e.*, often a simple, short program can produce very complex output) and nondeterminism.

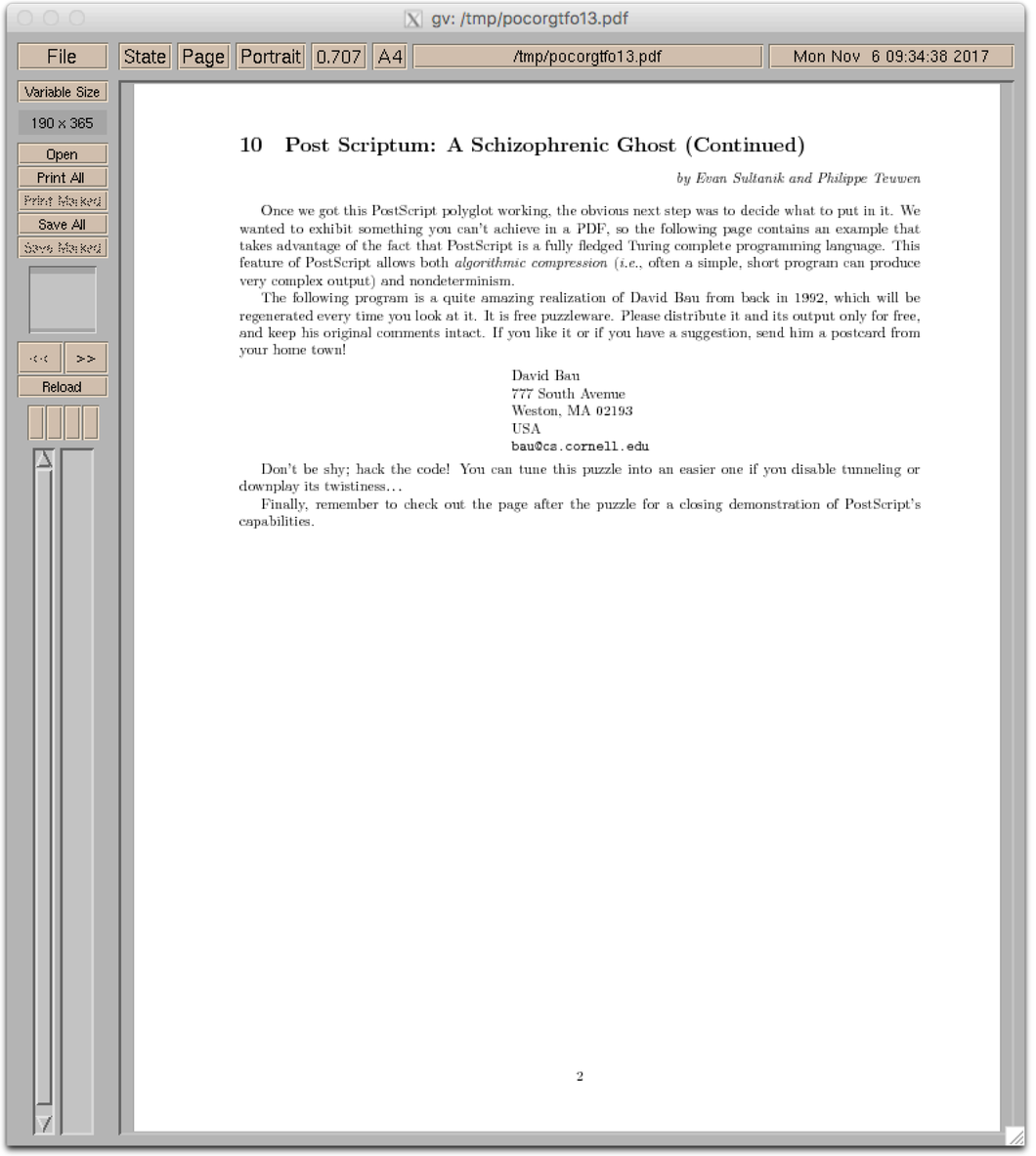
The following program is a quite amazing realization of David Bau from back in 1992, which will be regenerated every time you look at it. It is free puzzleware. Please distribute it and its output only for free, and keep his original comments intact. If you like it or if you have a suggestion, send him a postcard from your home town!

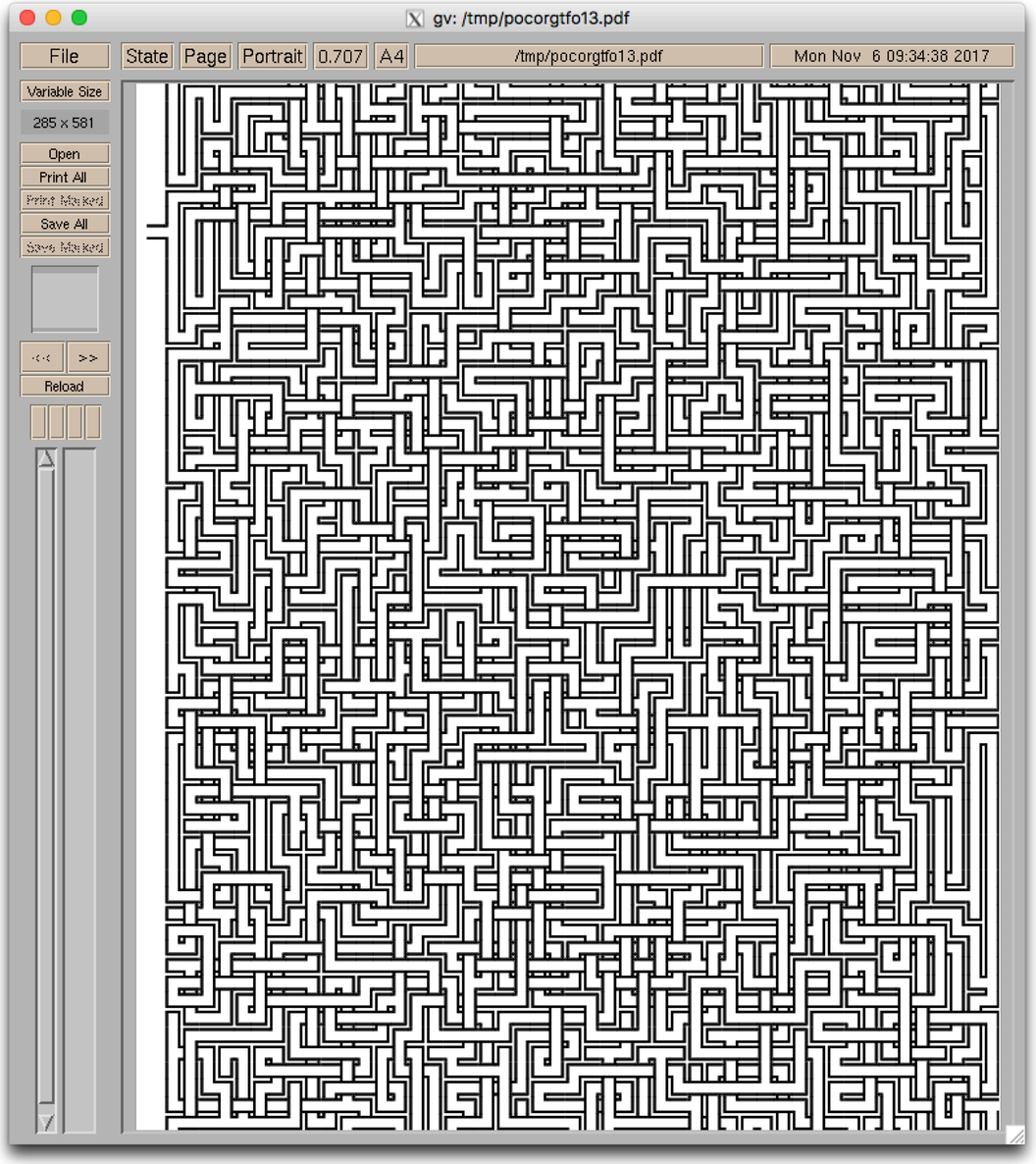
David Bau
777 South Avenue
Weston, MA 02193
USA
bau@cs.cornell.edu

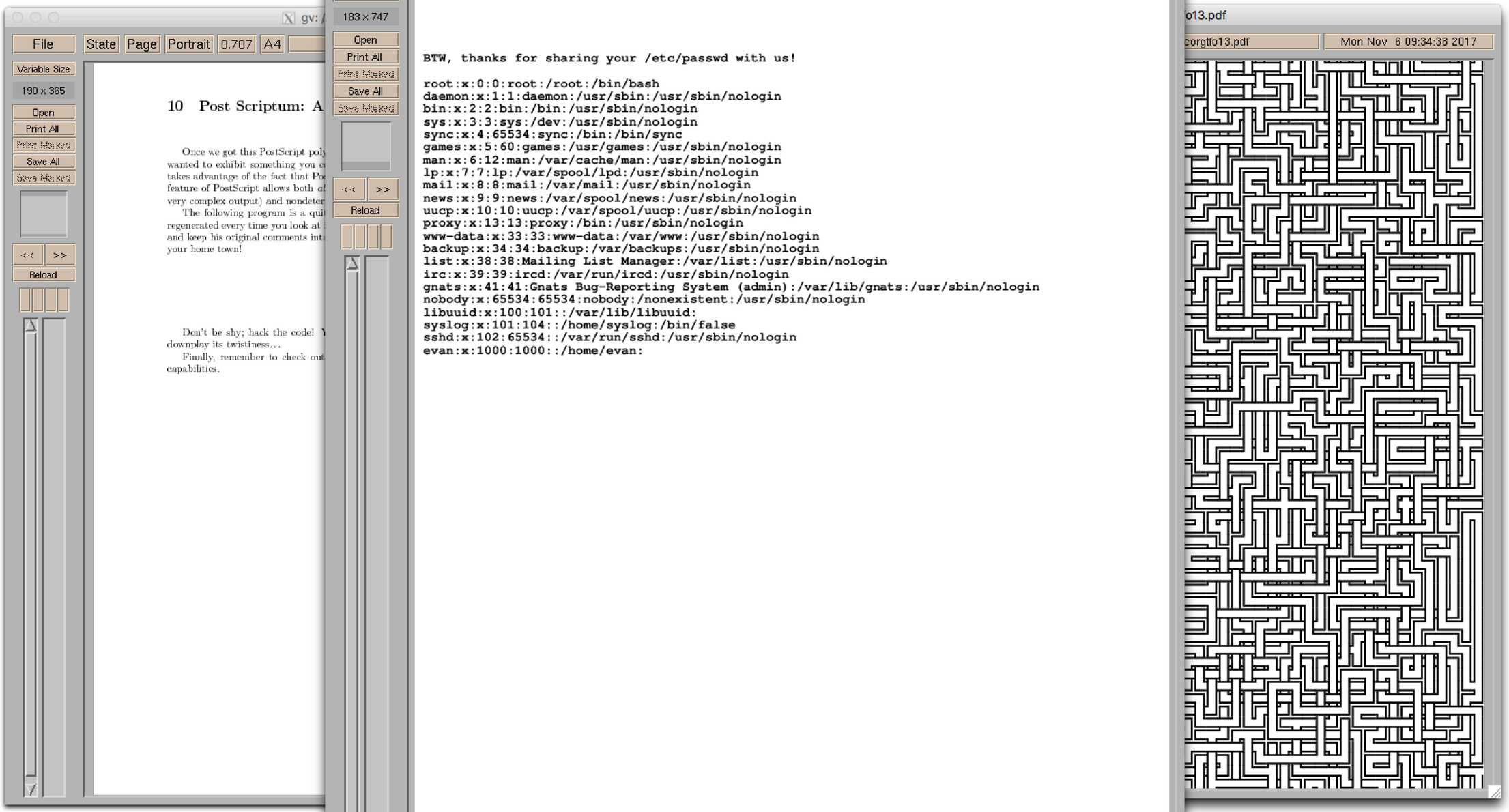
Don't be shy; hack the code! You can tune this puzzle into an easier one if you disable tunneling or downplay its twistiness...

Finally, remember to check out the page after the puzzle for a closing demonstration of PostScript's capabilities.

PostScript

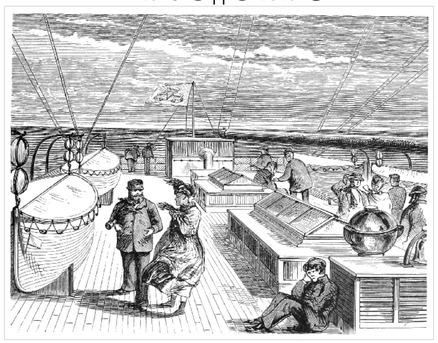

gv: /tmp/pocorgtfo13.pdf
File
State
Page
Portrait
0.707
A4
/tmp/pocorgtfo13.pdf
Mon Nov 6 09:34:38 2017
Variable Size
190 x 365
Open
Print All
Print Marked
Save All
Save Marked
<<
>>
Reload
10 Post Scriptum: A Schizophrenic Ghost (Continued)
by Evan Sultanik and Philippe Teuwen
Once we got this PostScript polyglot working, the obvious next step was to decide what to put in it. We wanted to exhibit something you can't achieve in a PDF, so the following page contains an example that takes advantage of the fact that PostScript is a fully fledged Turing complete programming language. This feature of PostScript allows both algorithmic compression (i.e., often a simple, short program can produce very complex output) and nondeterminism.
The following program is a quite amazing realization of David Bau from back in 1992, which will be regenerated every time you look at it. It is free puzzleware. Please distribute it and its output only for free, and keep his original comments intact. If you like it or if you have a suggestion, send him a postcard from your home town!
David Bau
777 South Avenue
Weston, MA 02193
USA
bau@cs.cornell.edu
Don't be shy; hack the code! You can tune this puzzle into an easier one if you disable tunneling or downplay its twistiness...
Finally, remember to check out the page after the puzzle for a closing demonstration of PostScript's capabilities.
2


gv: /tmp/pocorgtfo13.pdf
File
State
Page
Portrait
0.707
A4
/tmp/pocorgtfo13.pdf
Mon Nov 6 09:34:38 2017
Variable Size
285 x 581
Open
Print All
Print Marked
Save All
Save Marked
<<
>>
Reload


gv: /tmp/pocorgtfo13.pdf
File
State
Page
Portrait
1.000
A4
/tmp/pocorgtfo13.pdf
Sat Oct 22 08:37:57 2016
Variable Size
183 x 747
Open
Print All
Print Marked
Save All
Save Marked
<<
>>
Reload
BTW, thanks for sharing your /etc/passwd with us!
root:x:0:0:root:/root:/bin/bash
daemon:x:1:1:daemon:/usr/sbin:/usr/sbin/nologin
bin:x:2:2:bin:/bin:/usr/sbin/nologin
sys:x:3:3:sys:/dev:/usr/sbin/nologin
sync:x:4:65534:sync:/bin:/bin/sync
games:x:5:60:games:/usr/games:/usr/sbin/nologin
man:x:6:12:man:/var/cache/man:/usr/sbin/nologin
lp:x:7:7:lp:/var/spool/lpd:/usr/sbin/nologin
mail:x:8:8:mail:/var/mail:/usr/sbin/nologin
news:x:9:9:news:/var/spool/news:/usr/sbin/nologin
uucp:x:10:10:uucp:/var/spool/uucp:/usr/sbin/nologin
proxy:x:13:13:proxy:/bin:/usr/sbin/nologin
www-data:x:33:33:www-data:/var/www:/usr/sbin/nologin
backup:x:34:34:backup:/var/backups:/usr/sbin/nologin
list:x:38:38:Mailing List Manager:/var/list:/usr/sbin/nologin
irc:x:39:39:ircd:/var/run/ircd:/usr/sbin/nologin
gnats:x:41:41:Gnats Bug-Reporting System (admin):/var/lib/gnats:/usr/sbin/nologin
nobody:x:65534:65534:nobody:/nonexistent:/usr/sbin/nologin
libuuid:x:100:101::/var/lib/libuuid:
syslog:x:101:104::/home/syslog:/bin/false
sshd:x:102:65534::/var/run/sshd:/usr/sbin/nologin
evan:x:1000:1000::/home/evan:
File
State
Page
Portrait
0.707
A4
Variable Size
190 x 365
Open
Print All
Print Marked
Save All
Save Marked
<<
>>
Reload
10 Post Scriptum: A
Once we got this PostScript poly
wanted to exhibit something you c
takes advantage of the fact that Po
feature of PostScript allows both al
very complex output) and nondeter
The following program is a qui
regenerated every time you look at
and keep his original comments int
your home town!
Don't be shy; hack the code! Y
downplay its twistiness...
Finally, remember to check out
capabilities.
Mon Nov 6 09:34:38 2017

HTTP Quine

PoC||GTFO



IN A FIT OF STUBBORN OPTIMISM,

PASTOR MANUL LAPHROAIG

AND HIS CLEVER CREW

SET SAIL TOWARD

WELCOMING SHORES OF

THE GREAT UNKNOWN!

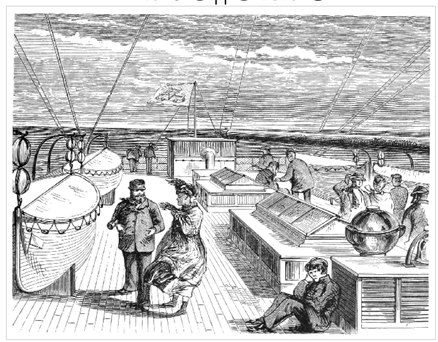
Heidelberg, Baden-Württemberg

Funded by our famous Single Malt Waterfall and Pastor Laphroaig's Рентгеновидат Gospel Choir, to be Freely Distributed to all Good Readers, and to be Freely Copied by all Good Bookleggers.

Это самиздат. Denn was man Schwarz auf Weiß besitzt, kann man getrost nach Hause tragen. €0, $0 USD, £0, 0 RSD, 0 SEK, $50 CAD. `pocorgtfo11.pdf`. March 17, 2016.

HTTP Quine

PoC||GTFO



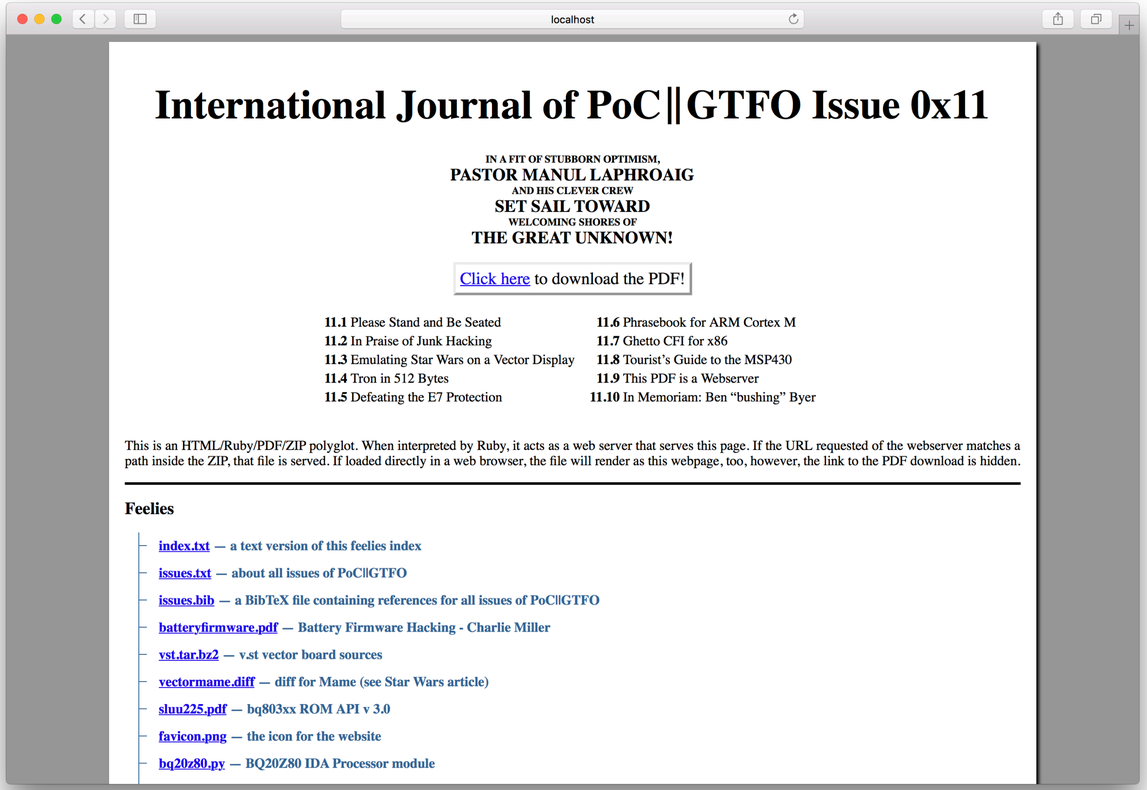
IN A FIT OF STUBBORN OPTIMISM,
PASTOR MANUL LAPHROAIG
AND HIS CLEVER CREW
SET SAIL TOWARD
WELCOMING SHORES OF
THE GREAT UNKNOWN!

Heidelberg, Baden-Württemberg

Funded by our famous Single Malt Waterfall and Pastor Laphroaig's Рентгениздат Gospel Choir, to be Freely Distributed to all Good Readers, and to be Freely Copied by all Good Bookleggers.

Это самиздат. Denn was man Schwarz auf Weiß besitzt, kann man getrost nach Hause tragen.
€0, $0 USD, £0, 0 RSD, 0 SEK, $50 CAD. `pocorgtfo11.pdf`. March 17, 2016.


```
$ ruby pocorgtfo11.pdf
Listening for connections on port 8080.
To listen on a different port,
re-run with the desired port as a command-line argument.
```



HTTP Quine

PA

11:1 Please Stand and Be
11:2 In Praise of Junk Ha
11:3 Emulating Star War
11:4 Tron in 512 Bytes
11:5 Defeating the E7 Pro

Heidelberg, Baden-Württe

Funded by our famous Sin
Pastor Laphroaig's Penrre
to be Freely Distributed t
to be Freely Copied by all

Это самиздат. Denn was
€0, $0 USD, £0, 0 RSD,

localhost

International Journal of PoC||GTFO Issue 0x11

IN A FIT OF STUBBORN OPTIMISM,
PASTOR MANUL LAPHROAIG
AND HIS CLEVER CREW
SET SAIL TOWARD
WELCOMING SHORES OF
THE GREAT UNKNOWN!

Click here to download the PDF!

This is an HTML/Ruby/PDF/ZIP polyglot. When interpreted by Ruby, it acts as a web server that serves this page. If the URL requested of the webserver matches a path inside the ZIP, that file is served. If loaded directly in a web browser, the file will render as this webpage, too, however, the link to the PDF download is hidden.

Feelies

- index.txt — a text version of this feelies index
- issues.txt — about all issues of PoC||GTFO
- issues.bib — a BibTeX file containing references for all issues of PoC||GTFO
- batteryfirmware.pdf — Battery Firmware Hacking - Charlie Miller
- vst.tar.bz2 — v.st vector board sources
- vectormame.diff — diff for Mame (see Star Wars article)
- sluu225.pdf — bq803xx ROM API v 3.0
- favicon.png — the icon for the website
- bq20z80.py — BQ20Z80 IDA Processor module

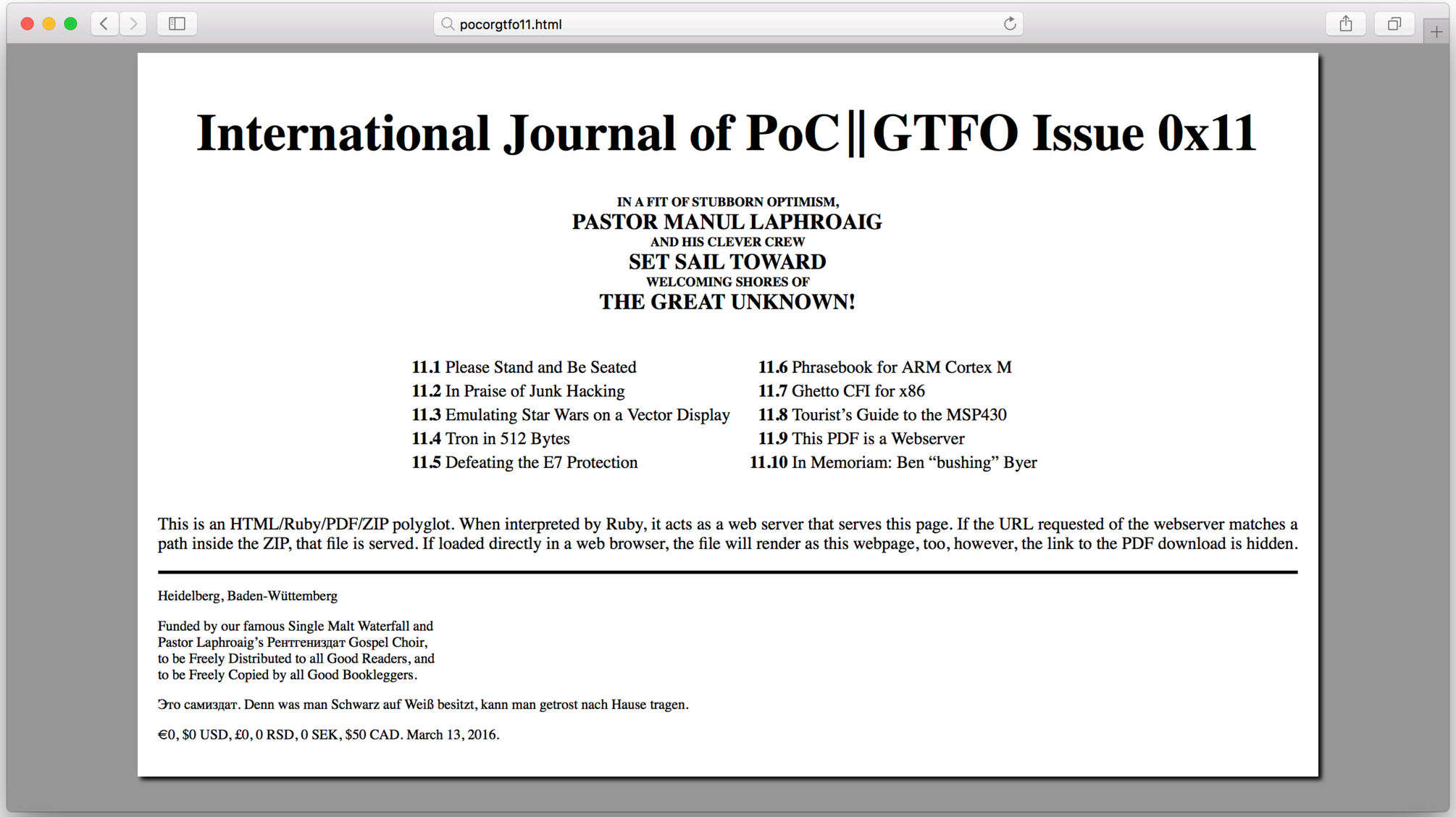
Feelies automagically parsed from the ZIP!

argument.

But Wait, There's More!

```
$ ln -s pocorgtfo11.pdf pocorgtfo11.html
```

But Wait, There's More!



International Journal of PoC||GTFO Issue 0x11

IN A FIT OF STUBBORN OPTIMISM,
PASTOR MANUL LAPHROAIG
AND HIS CLEVER CREW
SET SAIL TOWARD
WELCOMING SHORES OF
THE GREAT UNKNOWN!

This is an HTML/Ruby/PDF/ZIP polyglot. When interpreted by Ruby, it acts as a web server that serves this page. If the URL requested of the webserver matches a path inside the ZIP, that file is served. If loaded directly in a web browser, the file will render as this webpage, too, however, the link to the PDF download is hidden.

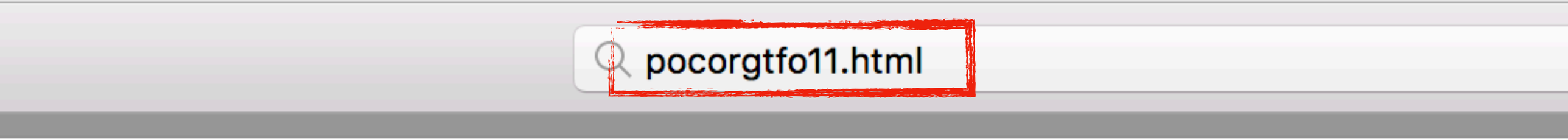
Heidelberg, Baden-Württemberg

Funded by our famous Single Malt Waterfall and
Pastor Laphroaig's Рентгениздат Gospel Choir,
to be Freely Distributed to all Good Readers, and
to be Freely Copied by all Good Bookleggers.

Это самиздат. Denn was man Schwarz auf Weiß besitzt, kann man getrost nach Hause tragen.

€0, $0 USD, £0, 0 RSD, 0 SEK, $50 CAD. March 13, 2016.

But Wait, There's More!



nternational Journal of Po

IN A FIT OF STUBBORN OPTI

PASTOR MANUL LAPI

In which an HTML page is also a PDF

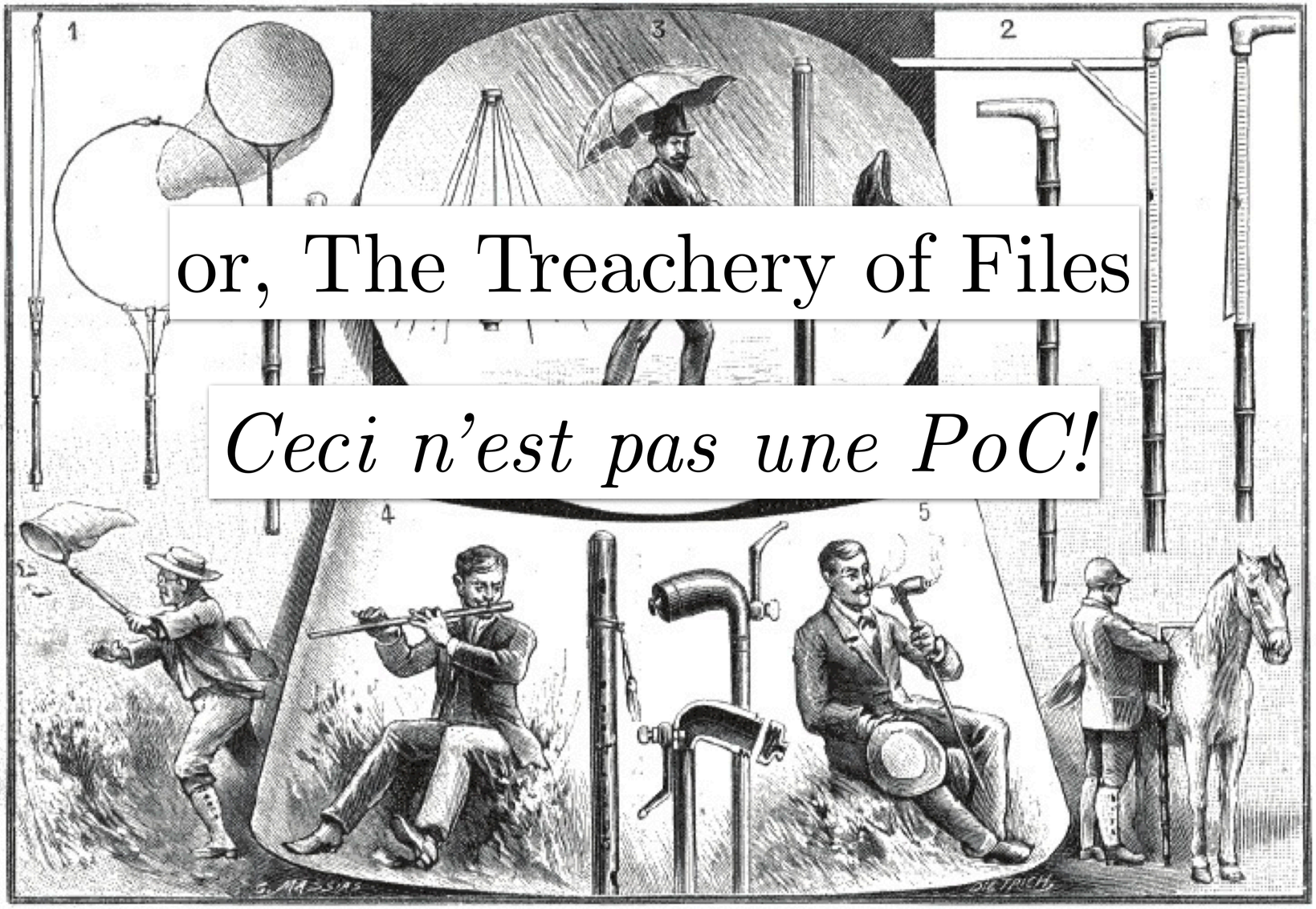
Which is also a ZIP

Which is also a Ruby script

Which is an HTTP Quine

or, The Treachery of Files

Ceci n'est pas une PoC!



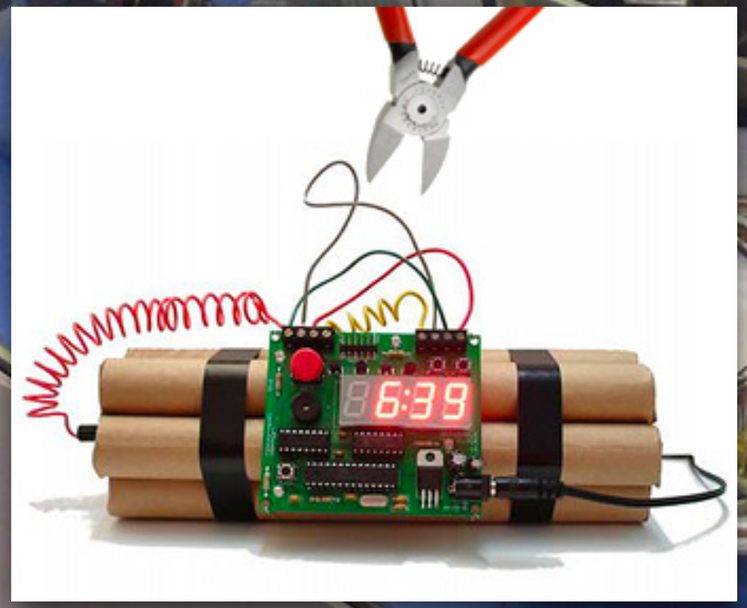
Utilisation de la canne. — 1. Canne-filet à papillons. — 2. Canne à toiser les chevaux. — 3. Canne-parapluie. — 4. Canne musicale. — 5. Ceci n'est pas une pipe.

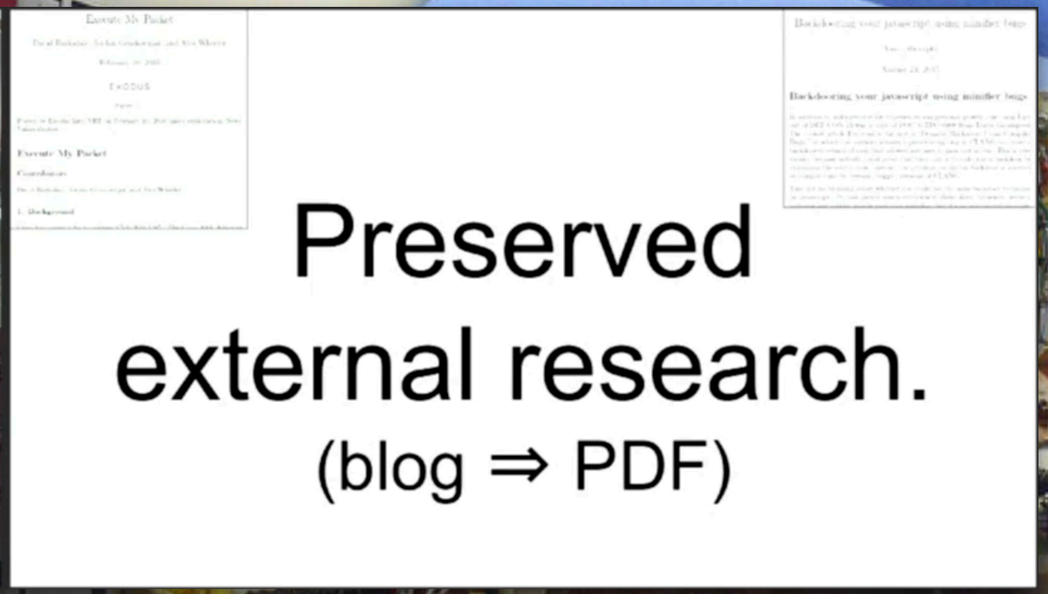
NNSA

National Nuclear Security Administration

3,000,000 Square Feet
(almost 2x the size of the
Tesla Gigafactory)

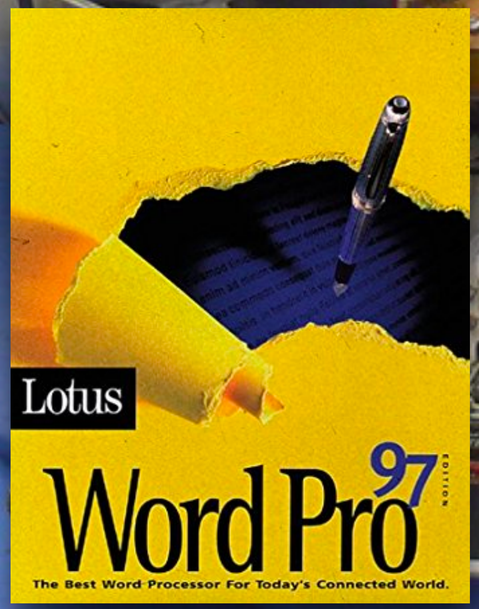
1949–2014


6:39



Preserved external research.

(blog ⇒ PDF)



PDF/Git Polyglot

```
$ git bundle
```

bundle.c

```
argv_array_pushl(&pack_objects.args,
                 "pack-objects", "--all-progress-implied",
                 "--compression=0",
                 "--stdout", "--thin", "--delta-base-offset",
NULL);
```

```
$ export PATH=/path/to/patched/git:$PATH
$ git init
$ git add article.pdf
$ git commit article.pdf -m "added"
$ git bundle create PDFGitPolyglot.pdf -all
```

https://github.com/ESultanik/PDFGitPolyglot

This PDF is a Git Repository Containing its Own LaTeX Source and a Copy of Itself

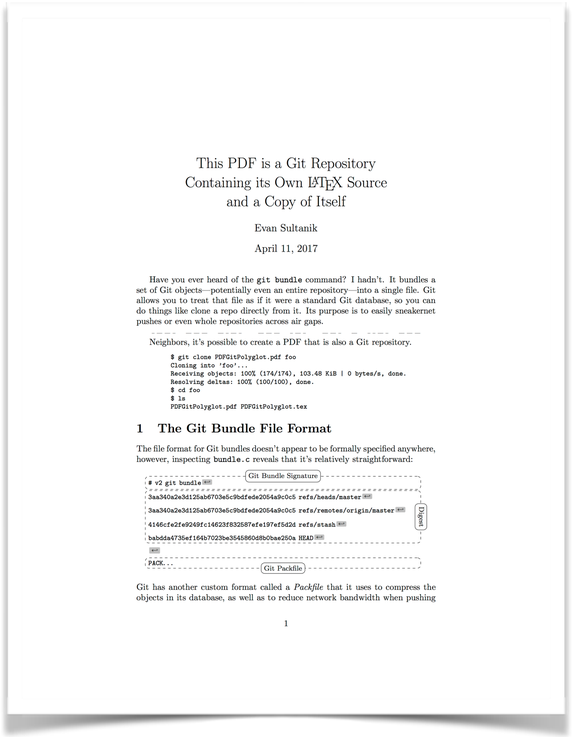
Evan Sultanik

April 11, 2017

Have you ever heard of the `git bundle` command? I hadn't. It bundles a set of Git objects—potentially even an entire repository—into a single file. Git allows you to treat that file as if it were a standard Git database, so you can do things like clone a repo directly from it. Its purpose is to easily sneakernet pushes or even whole repositories across air gaps.

Neighbors, it's possible to create a PDF that is also a Git repository.

PDF/Git Polyglot



This PDF is a Git Repository
Containing its Own LaTeX Source
and a Copy of Itself

Evan Sultanik

April 11, 2017

Have you ever heard of the `git bundle` command? I hadn't. It bundles a set of Git objects—potentially even an entire repository—into a single file. Git allows you to treat that file as if it were a standard Git database, so you can do things like clone a repo directly from it. Its purpose is to easily sneakernet pushes or even whole repositories across air gaps.

Neighbors, it's possible to create a PDF that is also a Git repository.

```
$ git clone PDFGitPolyglot.pdf foo
Cloning into 'foo'...
Receiving objects: 100% (174/174), 103.48 KiB | 0 bytes/s, done.
Resolving deltas: 100% (100/100), done.
$ cd foo
$ ls
PDFGitPolyglot.pdf PDFGitPolyglot.tex
```

1 The Git Bundle File Format

The file format for Git bundles doesn't appear to be formally specified anywhere, however, inspecting `bundle.c` reveals that it's relatively straightforward:

Git Bundle Signature

```
# v2 git bundle LF
3aa340a2e3d125ab6703e5c9bdfede2054a9c0c5 refs/heads/master LF
3aa340a2e3d125ab6703e5c9bdfede2054a9c0c5 refs/remotes/origin/master LF
4146cfe2fe9249fc14623f832587efe197ef5d2d refs/stash LF
babdda4736ef164b7023be3545860d8b0bae250a HEAD LF
LF
PACK...
```

Digest

Git Packfile

Git has another custom format called a *Packfile* that it uses to compress the objects in its database, as well as to reduce network bandwidth when pushing

1

```
$ git clone PDFGitPolyglot.pdf foo
Cloning into 'foo'...
Receiving objects: 100% (174/174), 103.48 KiB | 0 bytes/s, done.
Resolving deltas: 100% (100/100), done.
$ cd foo
$ ls
PDFGitPolyglot.pdf PDFGitPolyglot.tex
```

Act II

Introducing PolyFile and PolyTracker!

Act II

Introducing PolyFile and PolyTracker!

Automated Lexical Annotation and Navigation of Parsers

Act II

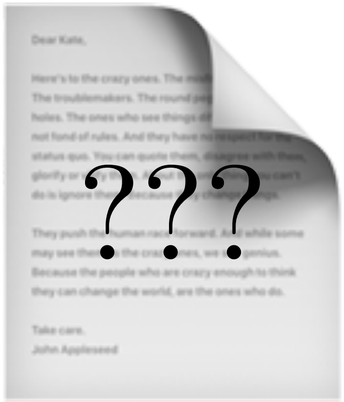
Introducing PolyFile and PolyTracker!

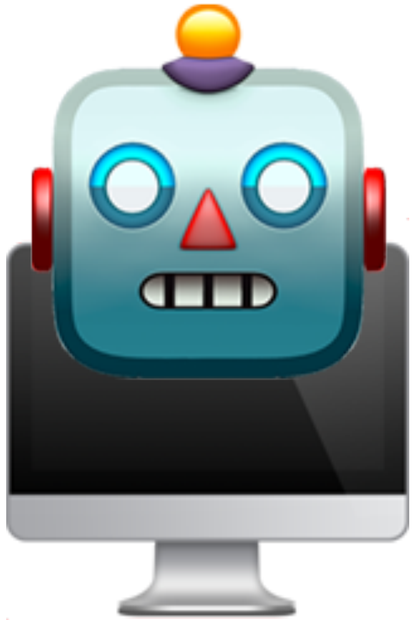
The ALAN Parsers Project

~~Automated Lexical Annotation and Navigation of Parsers~~

High Level Goals

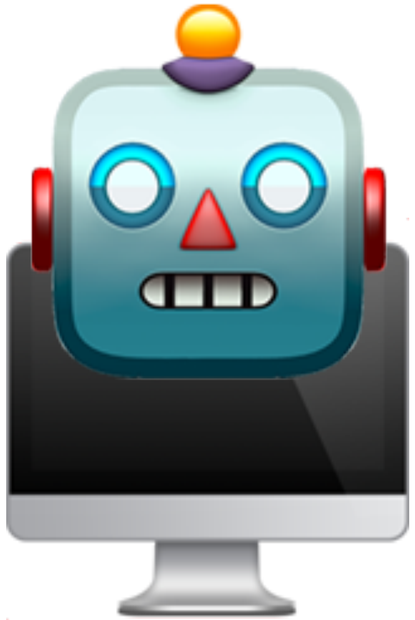
Create semantic map of the functions in a parser





High Level Goals

Create semantic map of the functions in a parser



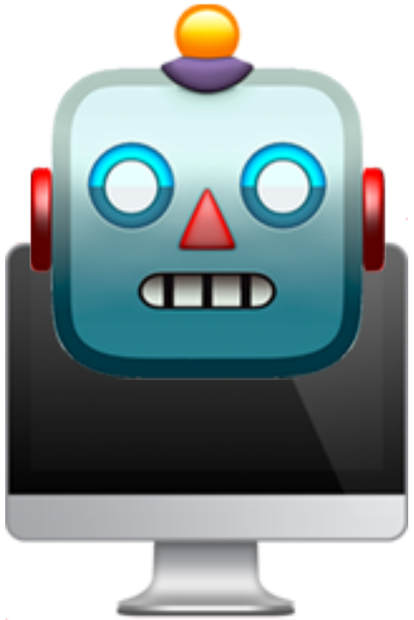
```
parser_function1
↳byte 0, 10, 50
  Object Stream

parser_function2
↳byte 10, 74
  Xref

parser_function3
↳byte 20
  JFIF
```

High Level Goals

Create semantic map of the functions in a parser

Ultimate Goal: Automatically extract a minimal grammar specifying the files accepted by a parser

Hypothesis: The majority of the potential for maliciousness and schizophrenia will exist in the symmetric difference of the grammars accepted by a format's parser implementations



```
parser_function1
↳byte 0, 10, 50
  Object Stream

parser_function2
↳byte 10, 74
  Xref

parser_function3
↳byte 20
  JFIF
```

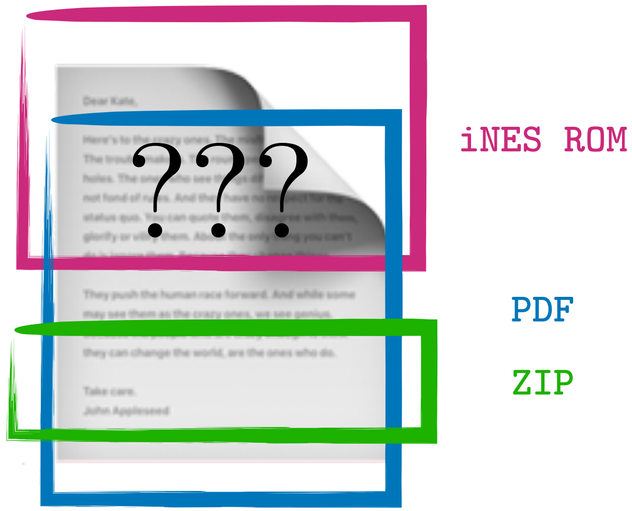
Related Work

- **Ange's SBuD**
  - Limited parser support
- **Kaitai Struct**
  - No support for PDF
  - Difficult to specity context sensitive languages
- **Valgrind / TaintGrind**
  - Intractably slow; hours of computation to parse a minimal PDF
- **AFL Analyze**
  - Does not associate input byte sequences with the program trace
- **LLVM Dataflow Sanitizer**
  - Limited to just a few thousand taint labels
  - Can only track at most one input byte at a time

Valgrind

Approach: Semantic Labeling

Polyglot-Aware File Identification



Resilient Parsing

Modify parsers for best effort

Instrument to track input byte offsets

Label regions of the input

Produce ground truth

Hierarchical Labeling

```
iNES [0x0→0x12220]
↳ Magic [0x0→0x3]
   Header [0x4→0xF]
   ⋮
   PRG [0xC210→0x1020F]
   CHR [0x10210→0x12220]
PDF [0x10→0x2EF72F]
↳ Magic [0x10→0x1E]
   Object 1.0 [0x1F→0x12221]
   ↳ Dictionary [0x2A→0x3E]
      Stream [0x3F→0x12219]
      ↳ JFIF Image [0x46→0x1220F]
         ↳ JPEG Segment […]
            ↳ Magic […]
               Marker […]
⋮
```

Approach: Parser Instrumentation

LLVM

Operate on LLVM/IR

Can work with all open source parsers

Eventually support closed-source binaries by lifting to LLVM (*e.g.*, with McSEMA or Remill)

Instrumentation

Shadow memory inspired by the Data Flow Sanitizer (dfsan)

Negligible CPU overhead

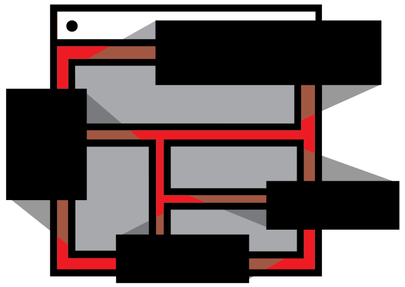
$O(n)$ memory overhead, where n is the number of instructions executed by the parser

Taint Tracking

Novel datastructure for efficiently storing taint labels

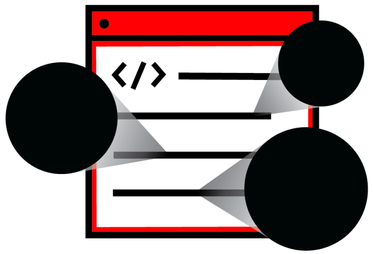
dfsan status quo:
$\boldsymbol{\Theta}(1)$ lookups
$\boldsymbol{\Theta}(n^2)$ storage

TAPP:
$O(\log\ n)$ lookups
$O(n)$ storage



PolyFile

Current Tooling



PolyTracker

Novel file matching algorithm seeded with over 10,000 file definitions from TrID

Kaitai Struct file definition parser for automatically labeling based upon KSY definitions (*e.g.*, JPEG, ZIP, WASM)

Instrumented Didier Stevens' PDF parser, resilient to malformations and schizophrenia

Interactive file explorer

https://github.com/trailofbits/polyfile

Shadow memory inspired by the Data Flow Sanitizer (dfsan) and Angora

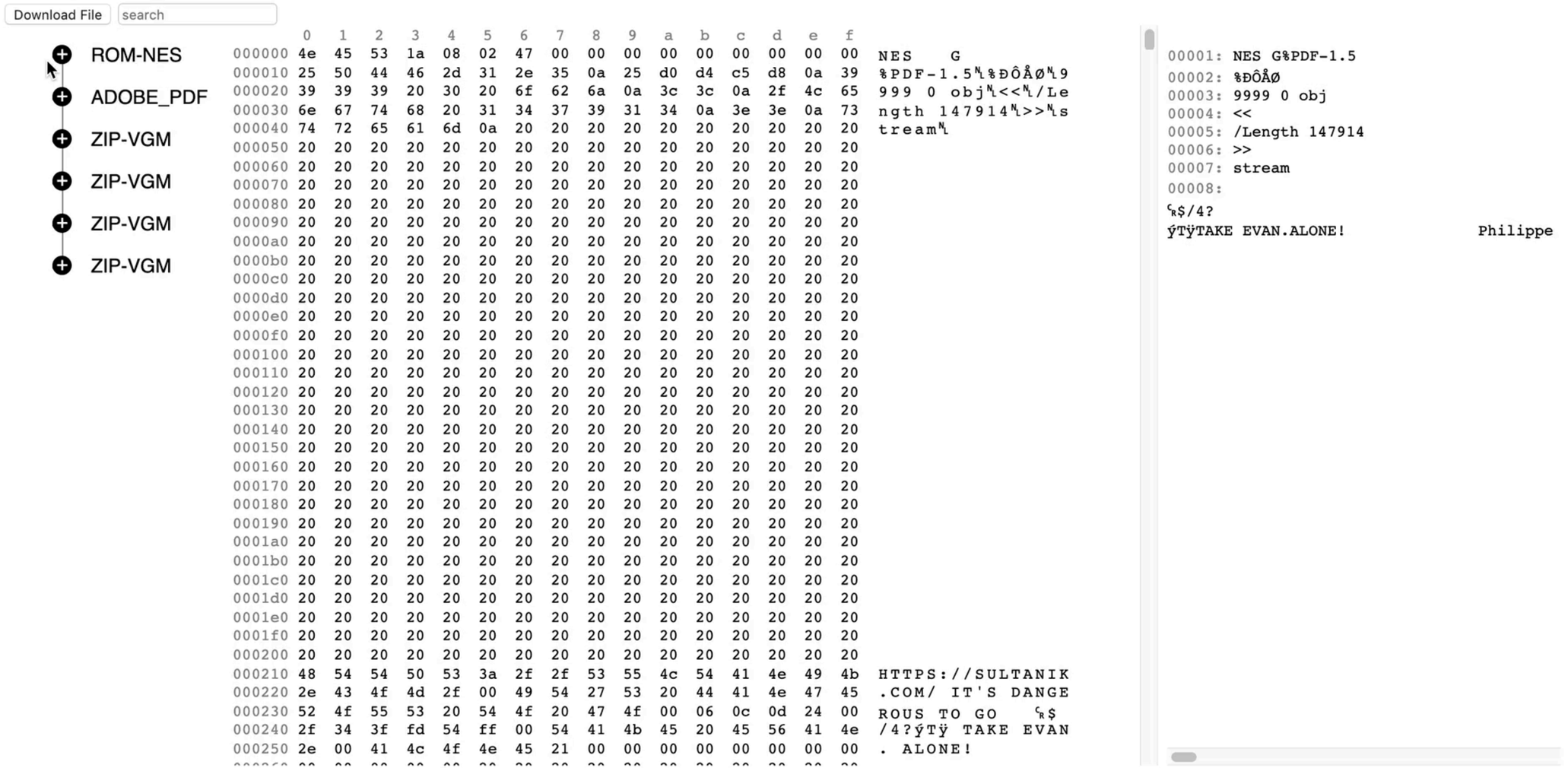
Negligible CPU overhead

O(n) memory overhead, where n is the number of instructions executed by the parser

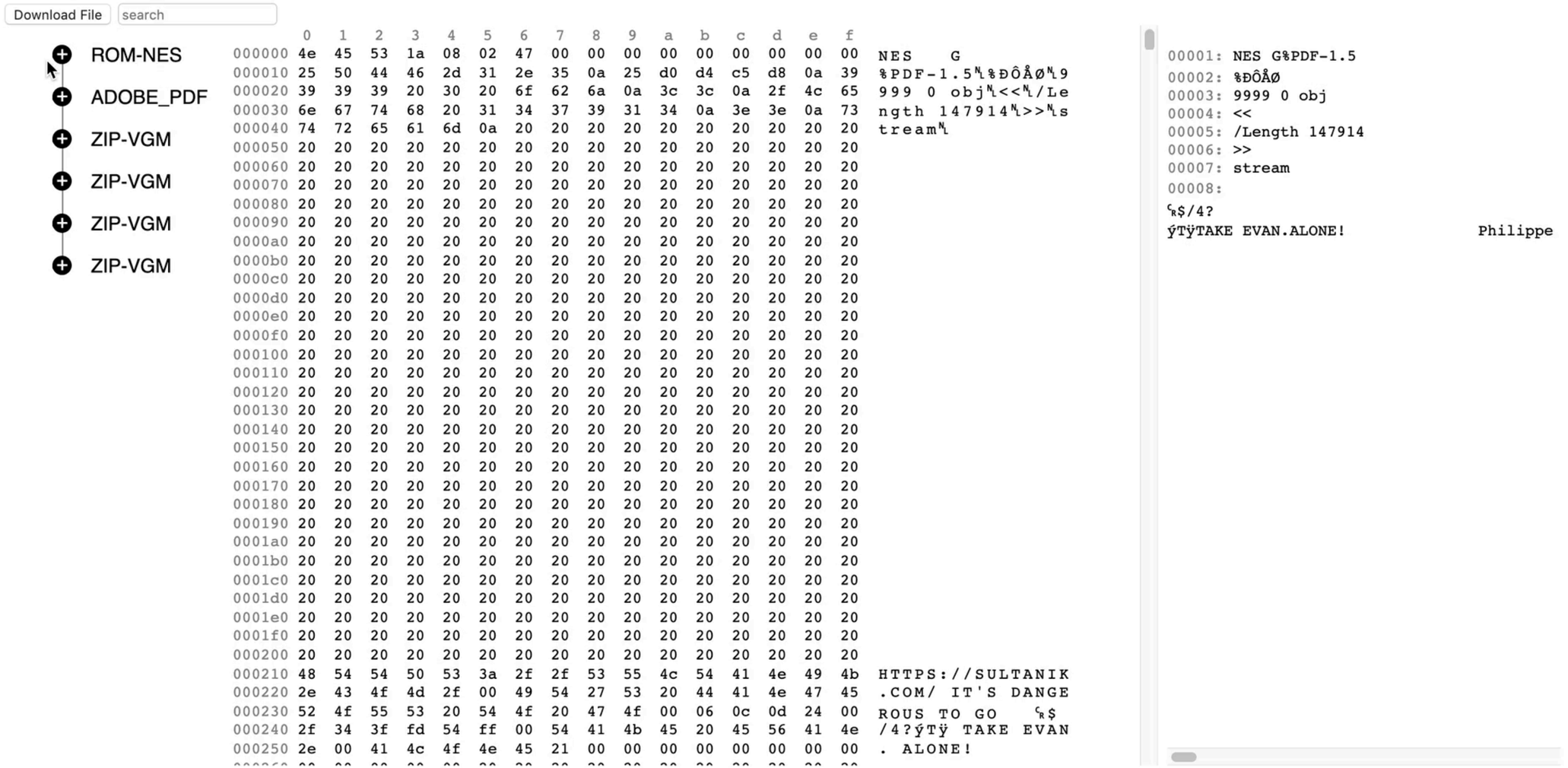
Currently supports C, C++ (experimental), and any code entirely representable in LLVM/IR.

https://github.com/trailofbits/polytracker

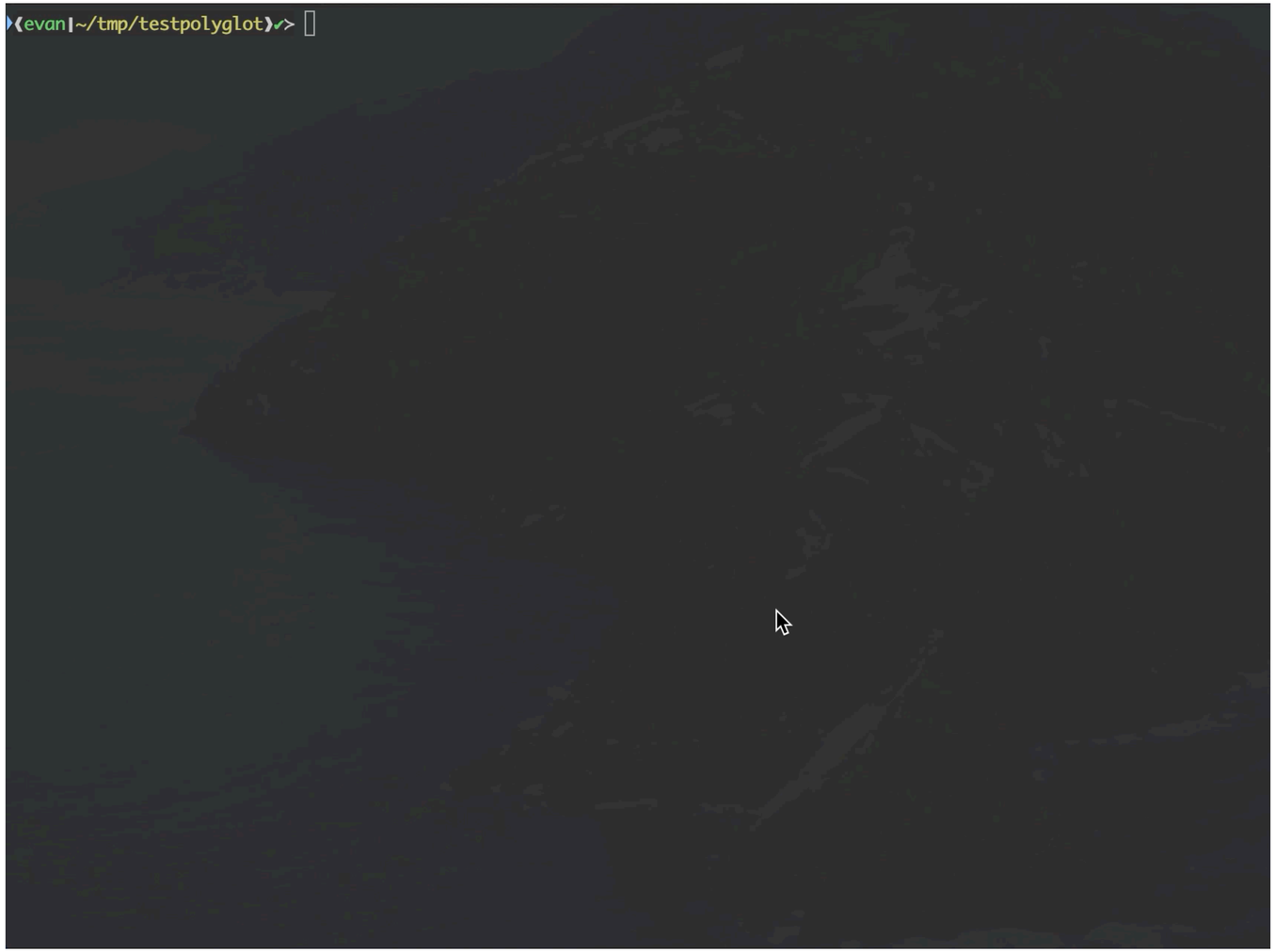
Demo


Download File
search
ROM-NES
ADOBE_PDF
ZIP-VGM
ZIP-VGM
ZIP-VGM
ZIP-VGM
0 1 2 3 4 5 6 7 8 9 a b c d e f
000000 4e 45 53 1a 08 02 47 00 00 00 00 00 00 00 00 00 NES G
000010 25 50 44 46 2d 31 2e 35 0a 25 d0 d4 c5 d8 0a 39 %PDF-1.5␊%ÐÔÅØ␊9
000020 39 39 39 20 30 20 6f 62 6a 0a 3c 3c 0a 2f 4c 65 999 0 obj␊<<␊/Le
000030 6e 67 74 68 20 31 34 37 39 31 34 0a 3e 3e 0a 73 ngth 147914␊>>␊s
000040 74 72 65 61 6d 0a 20 20 20 20 20 20 20 20 20 20 tream␊
000050 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000060 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000070 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000080 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000090 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000a0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000b0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000c0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000d0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000e0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000f0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000100 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000110 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000120 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000130 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000140 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000150 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000160 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000170 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000180 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000190 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001a0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001b0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001c0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001d0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001e0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001f0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000200 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000210 48 54 54 50 53 3a 2f 2f 53 55 4c 54 41 4e 49 4b HTTPS://SULTANIK
000220 2e 43 4f 4d 2f 00 49 54 27 53 20 44 41 4e 47 45 .COM/ IT'S DANGE
000230 52 4f 55 53 20 54 4f 20 47 4f 00 06 0c 0d 24 00 ROUS TO GO ␆␍$
000240 2f 34 3f fd 54 ff 00 54 41 4b 45 20 45 56 41 4e /4?ýTÿ TAKE EVAN
000250 2e 00 41 4c 4f 4e 45 21 00 00 00 00 00 00 00 00 . ALONE!
00001: NES G%PDF-1.5
00002: %ÐÔÅØ
00003: 9999 0 obj
00004: <<
00005: /Length 147914
00006: >>
00007: stream
00008:
␆$/4?
ýTÿTAKE EVAN.ALONE! Philippe

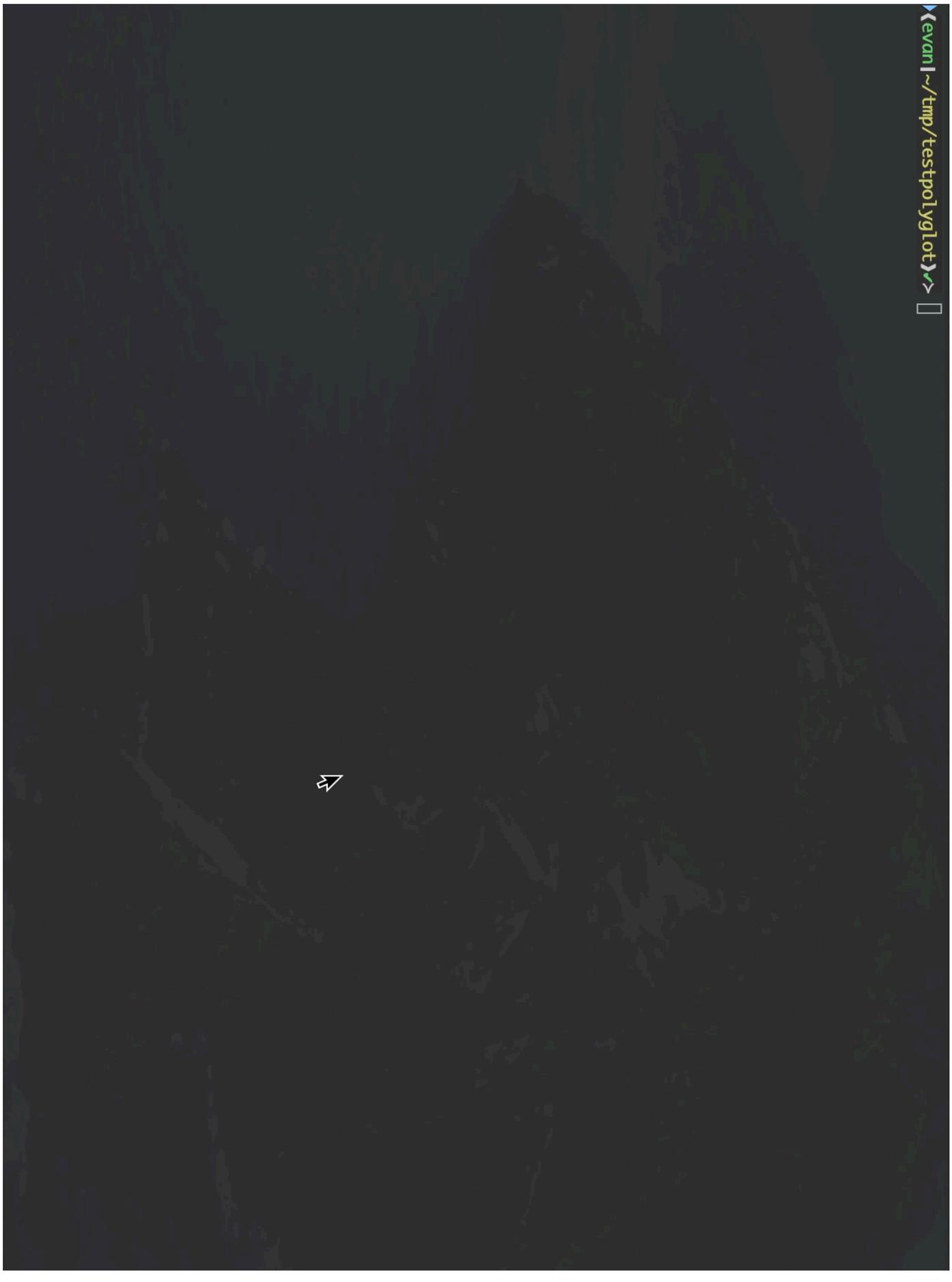
Demo


Download File
search
ROM-NES
ADOBE_PDF
ZIP-VGM
ZIP-VGM
ZIP-VGM
ZIP-VGM
0 1 2 3 4 5 6 7 8 9 a b c d e f
000000 4e 45 53 1a 08 02 47 00 00 00 00 00 00 00 00 00 NES G
000010 25 50 44 46 2d 31 2e 35 0a 25 d0 d4 c5 d8 0a 39 %PDF-1.5␊%ÐÔÅØ␊9
000020 39 39 39 20 30 20 6f 62 6a 0a 3c 3c 0a 2f 4c 65 999 0 obj␊<<␊/Le
000030 6e 67 74 68 20 31 34 37 39 31 34 0a 3e 3e 0a 73 ngth 147914␊>>␊s
000040 74 72 65 61 6d 0a 20 20 20 20 20 20 20 20 20 20 tream␊
000050 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000060 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000070 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000080 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000090 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000a0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000b0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000c0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000d0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000e0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0000f0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000100 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000110 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000120 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000130 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000140 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000150 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000160 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000170 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000180 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000190 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001a0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001b0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001c0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001d0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001e0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0001f0 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000200 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000210 48 54 54 50 53 3a 2f 2f 53 55 4c 54 41 4e 49 4b HTTPS://SULTANIK
000220 2e 43 4f 4d 2f 00 49 54 27 53 20 44 41 4e 47 45 .COM/ IT'S DANGE
000230 52 4f 55 53 20 54 4f 20 47 4f 00 06 0c 0d 24 00 ROUS TO GO ␆␍$
000240 2f 34 3f fd 54 ff 00 54 41 4b 45 20 45 56 41 4e /4?ýTÿ TAKE EVAN
000250 2e 00 41 4c 4f 4e 45 21 00 00 00 00 00 00 00 00 . ALONE!
00001: NES G%PDF-1.5
00002: %ÐÔÅØ
00003: 9999 0 obj
00004: <<
00005: /Length 147914
00006: >>
00007: stream
00008:
␆$/4?
ýTÿTAKE EVAN.ALONE! Philippe

WASM/HTML Polyglot Detection



WASM/HTML Polyglot Detection



PolyTracker Instrumentation

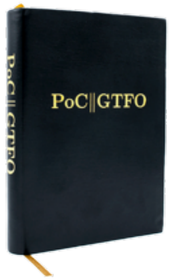

```
{
  "dfs$ensure_solid_xref": [
    2276587,
    2276588
  ],
  "dfs$fmt_obj": [
    2465223,
    2465224,
    2465225,
    2465226,
    2465227,
    2465228,
    2465240,
    2465241,
    2465242,
    2465243,
    2465244,
    2465245,
    2465246,
    2465258,
    2465259,
    2465260,
    2465261,
    2465262
  ]
}
```

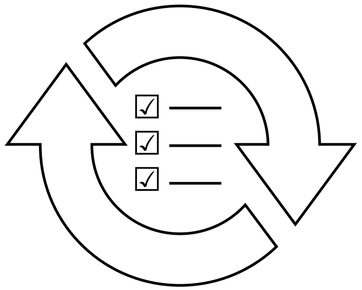
PolyTracker Instrumentation

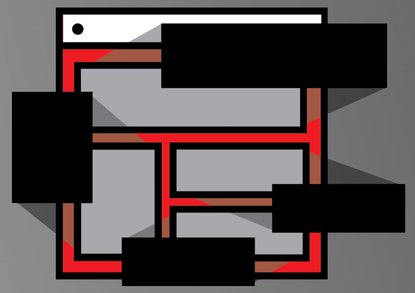

```
{
  "dfs$ensure_solid_xref": [
    2276587,
    2276588
  ],
  "dfs$fmt_obj": [
    2465223,
    2465224,
    2465225,
    2465226,
    2465227,
    2465228,
    2465240,
    2465241,
    2465242,
    2465243,
    2465244,
    2465245,
    2465246,
    2465258,
    2465259,
    2465260,
    2465261,
    2465262
  ]
}
```


```
iNES [0x0→0x12220]
↳ Magic [0x0→0x3]
   Header [0x4→0xF]
   ⋮
   PRG [0xC210→0x1020F]
   CHR [0x10210→0x12220]
PDF [0x10→0x2EF72F]
↳ Magic [0x10→0x1E]
   Object 1.0 [0x1F→0x12221]
   ↳ Dictionary [0x2A→0x3E]
      Stream [0x3F→0x12219]
      ↳ JFIF Image [0x46→0x1220F]
         ↳ JPEG Segment […]
            ↳ Magic […]
               Marker […]
⋮
```

Why Should You Care?

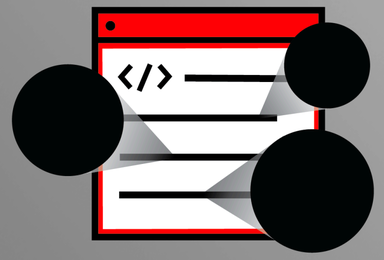




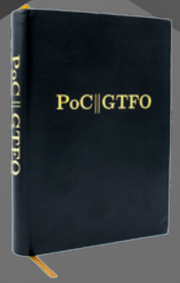


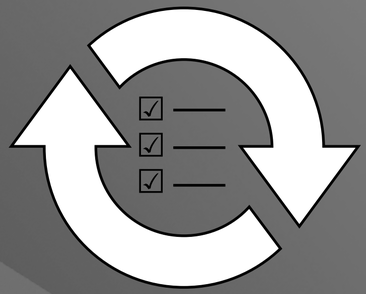
PolyFile

Why Should You Care?



PolyTracker

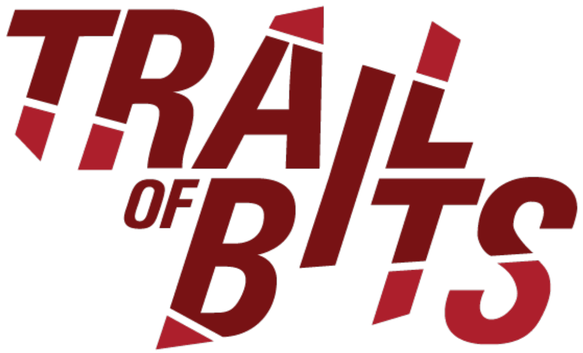




Acknowledgements

Ange Albertini	`@angealbertini`
Sergey Bratus	`@sergeybratus`
Travis Goodspeed	`@travisgoodspeed`
Philippe Teuwen	`@doegox`
Evan Teran	`@evan_teran`
Jacob Torrey	`@JacobTorrey`
Ryan Speers	`@rmspeers`

Et pl. al.



Thanks!

Evan Sultanik @ESultanik

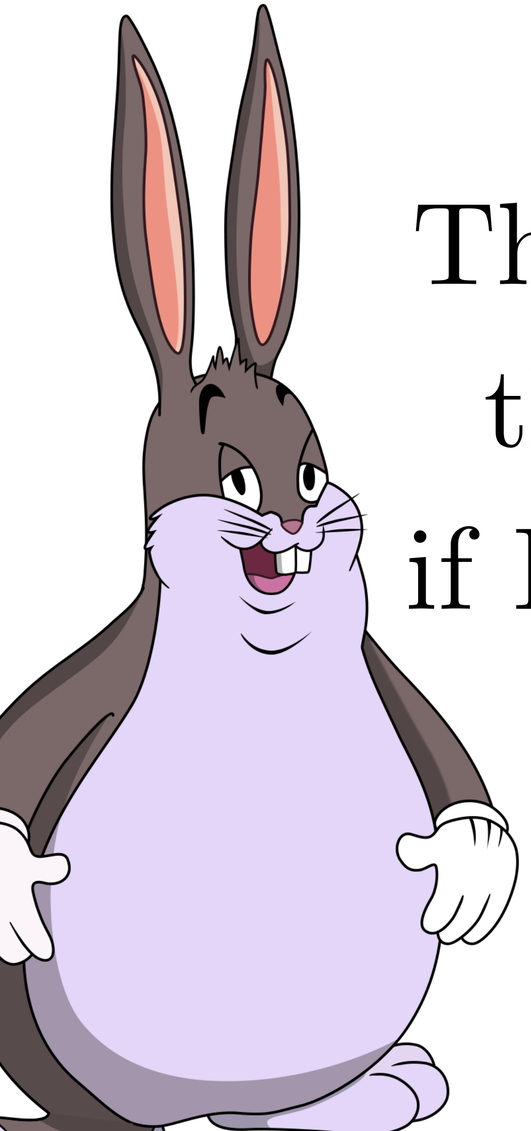
Thanks!

Contact Info

Carson Harmon	`@reyeetengineer`
Brad Larsen	`@BradLarsen`
Evan Sultanik	`@ESultanik`

https://github.com/trailofbits/polyfile

https://github.com/trailofbits/polytracker



This slide is here solely to prevent the slideshow from auto-closing if Evan clicks one too many times!