Google

The Beautiful Mind of a Timeline

Yet Another Data Science Talk, Filled With Buzzwords

OSDFCon 2019

Kristinn Guðjónsson

Johan Berggren

Current State


Data Overflow

> Ain’t nobody got time for all that data...

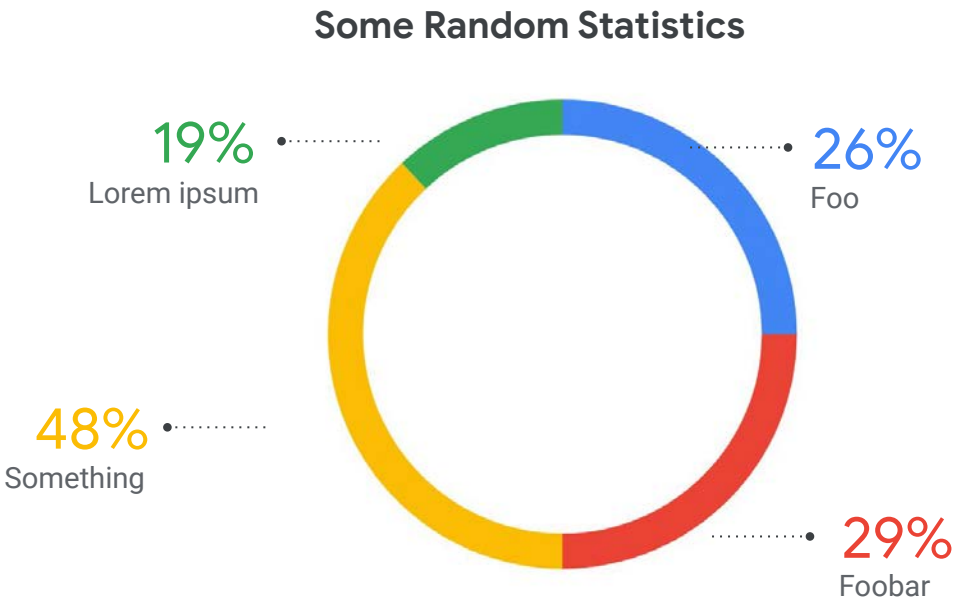
Kristinn Guðjónsson

Security Engineer, Google

Is there any hope?

Data Science

> Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data

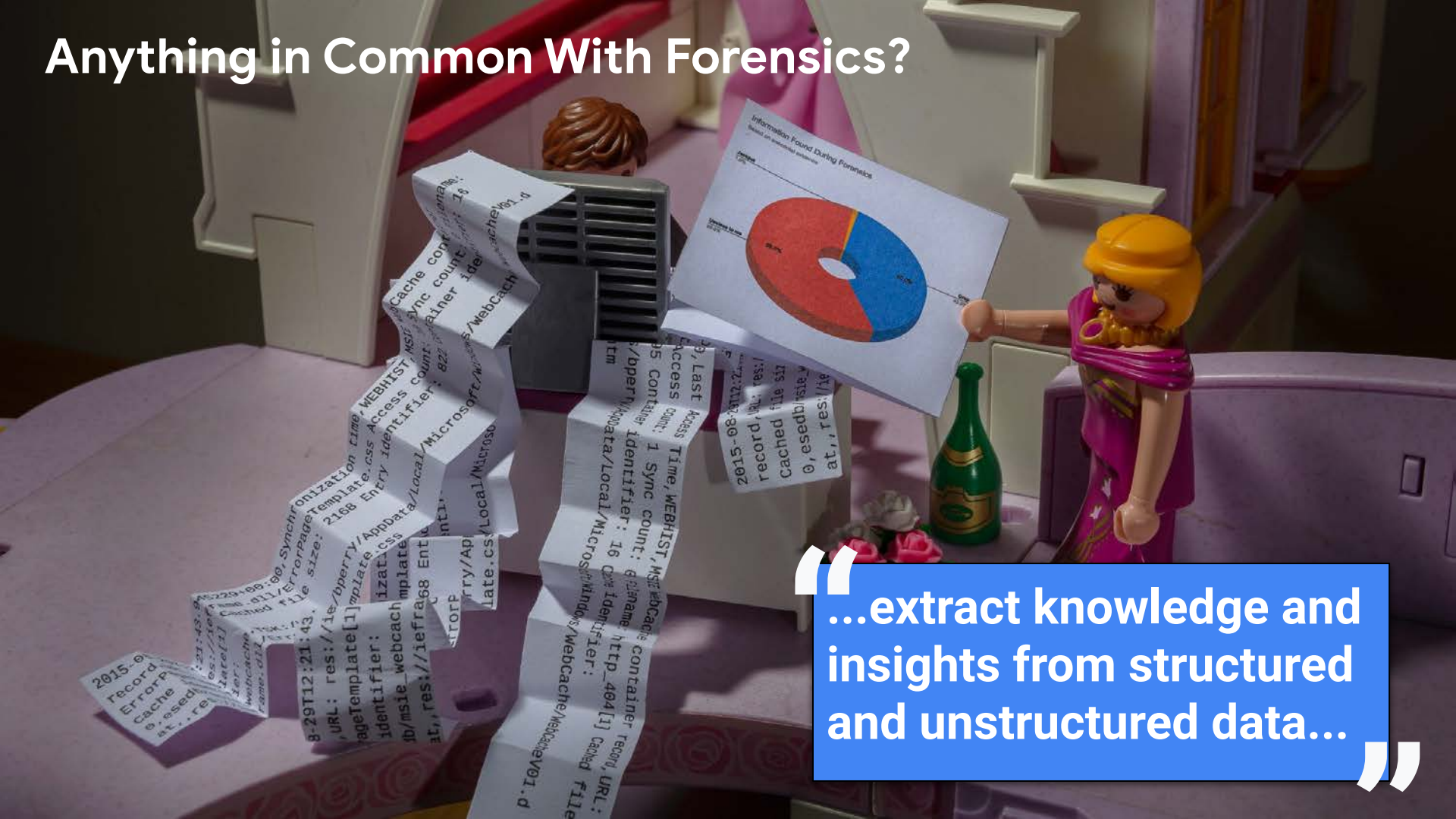


Source: Wikipedia

Data Scientist

Forensicator

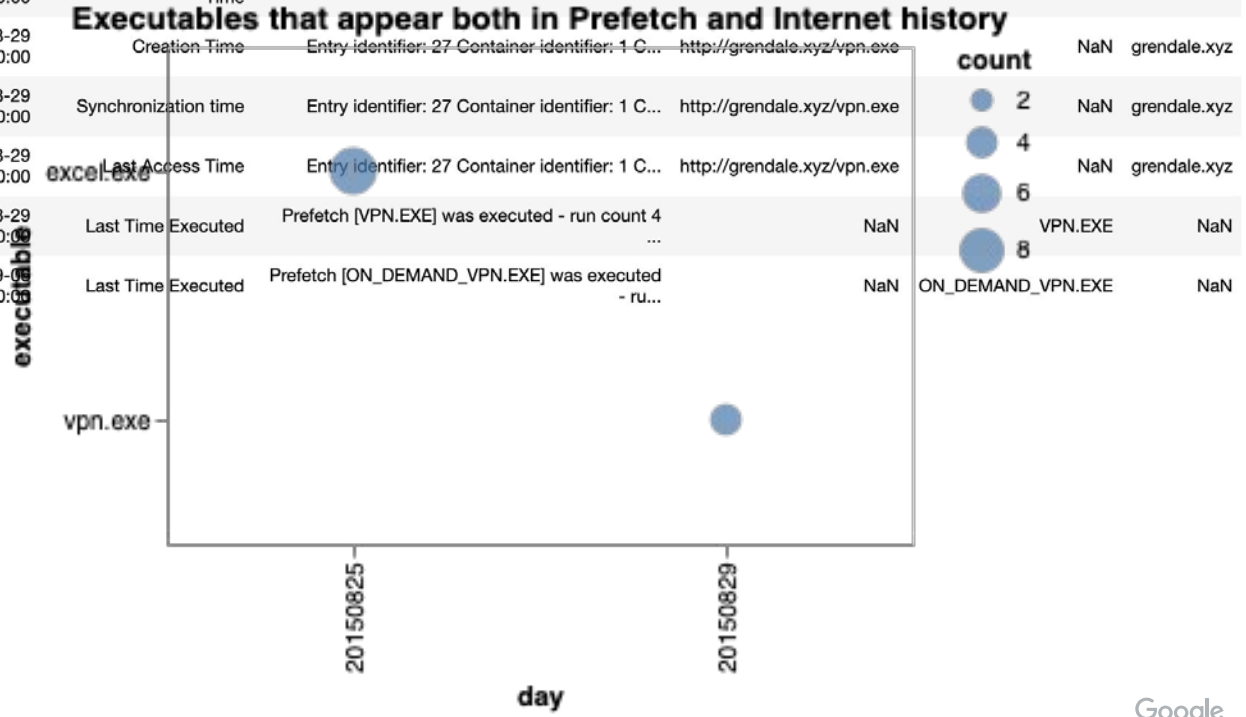
Anything in Common With Forensics?



What Can We Gain From Data Science?

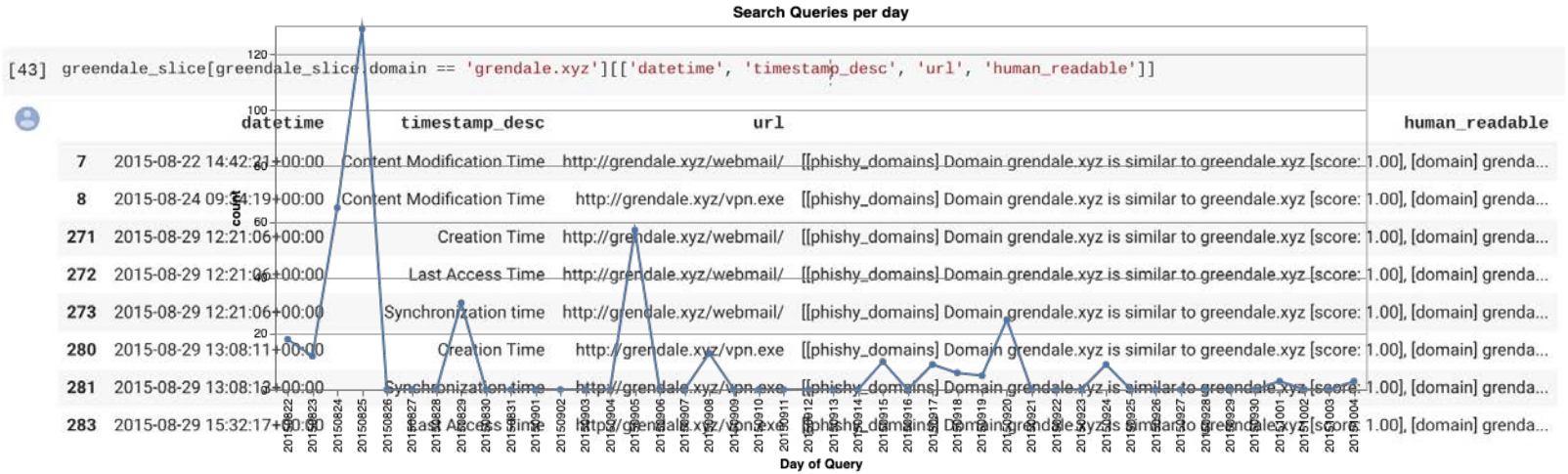
1. Tools
2. Methodology
3. Visualization
4. Aggregation

date	timestamp_desc	message	url	executable	domain
2015-08-24 09:34:19+00:00	Content Modification Time	Entry identifier: 27 Container identifier: 1 C...	http://grendale.xyz/vpn.exe	NaN	grendale.xyz
2015-08-29 13:08:11+00:00	Creation Time	Entry identifier: 27 Container identifier: 1 C...	http://grendale.xyz/vpn.exe	NaN	grendale.xyz
2015-08-29 13:08:13+00:00	Synchronization time	Entry identifier: 27 Container identifier: 1 C...	http://grendale.xyz/vpn.exe	NaN	grendale.xyz
2015-08-29 15:32:17+00:00	Last Access Time	Entry identifier: 27 Container identifier: 1 C...	http://grendale.xyz/vpn.exe	NaN	grendale.xyz
2015-08-29 15:32:17+00:00	Last Time Executed	Prefetch [VPN.EXE] was executed - run count 4 ...	NaN	VPN.EXE	NaN
2015-09-08 08:09:24+00:00	Last Time Executed	Prefetch [ON_DEMAND_VPN.EXE] was executed - ru...	NaN	ON_DEMAND_VPN.EXE	NaN

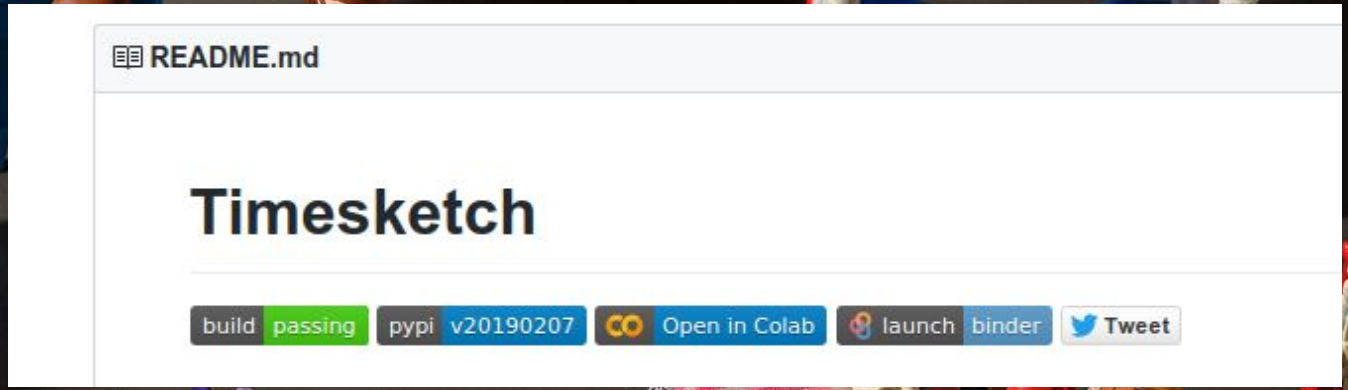


IPython Notebooks - Jupyter/Colab

1. Interactive “notebooks”, mix of text/code
2. Standalone: jupyter, or collaborative: colab
3. Uses standard Python



Can I Play?


README.md
Timesketch
build passing
pypi v20190207
Open in Colab
launch binder
Tweet

Live Demo

You will now experience a *live demo* that will most likely fail, as all live demos are doomed to do.

Notebook available at: https://bit.ly/32iK4xR

https://github.com/google/timesketch/blob/master/notebooks/OSDFCon%20Demo.ipynb

Summary

1. Don’t need to become statistician
2. Know how to extract features and count
3. Great way to “get to know your data”
4. Sometimes learning from other fields can benefit you

Research → Reality

Plaso

Input

160+ parsers

Output

Millions of events

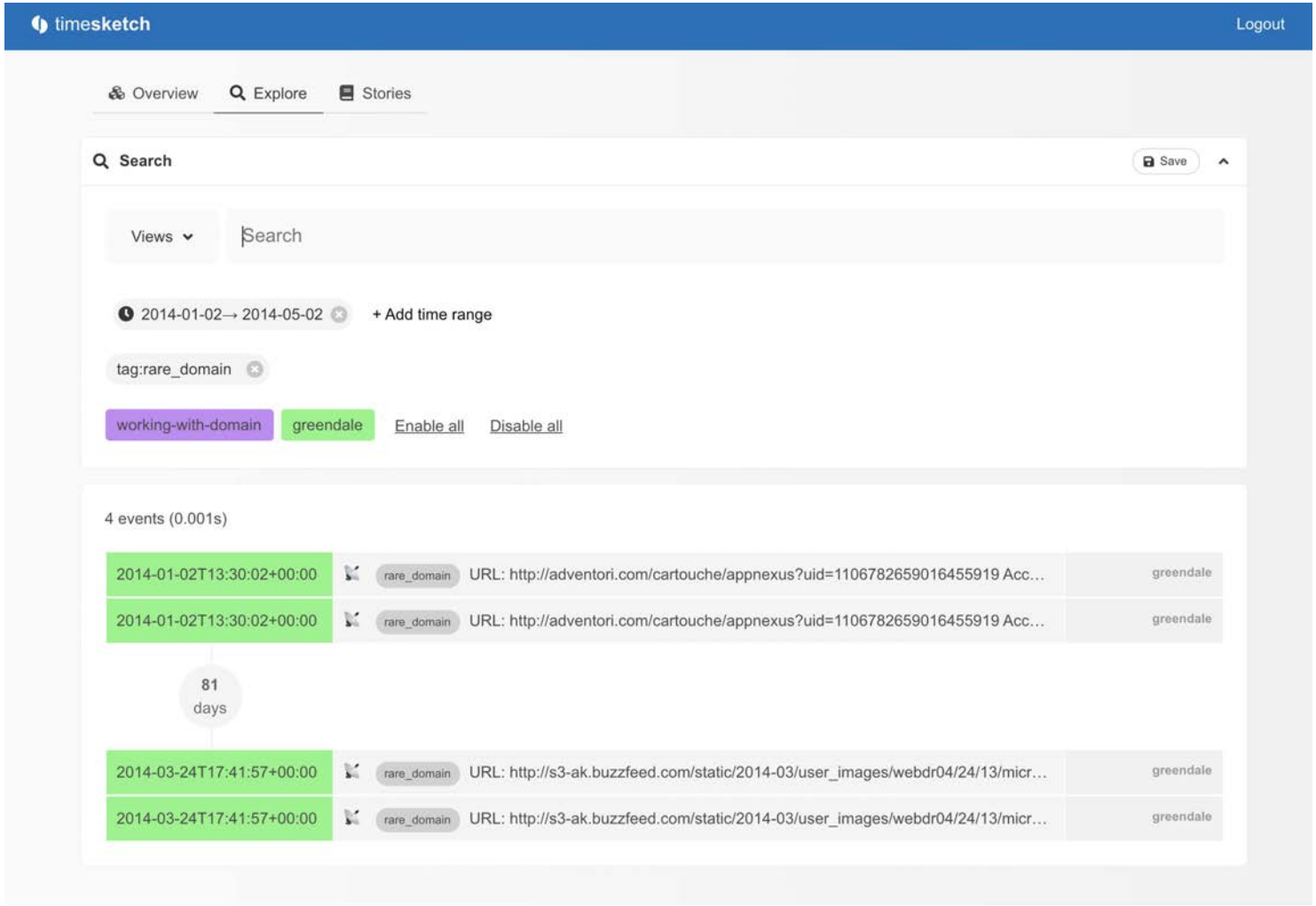

sed|grep|awk

timesketch

Digital Forensic Timeline Analysis

1 Search

2 Collaboration


timesketch
Logout
Overview
Explore
Stories
Search
Save
Views
Search
2014-01-02→ 2014-05-02
+ Add time range
tag:rare_domain
working-with-domain
greendale
Enable all
Disable all
4 events (0.001s)
2014-01-02T13:30:02+00:00
rare_domain
URL: http://adventori.com/cartouche/appnexus?uid=1106782659016455919 Acc…
greendale
2014-01-02T13:30:02+00:00
rare_domain
URL: http://adventori.com/cartouche/appnexus?uid=1106782659016455919 Acc…
greendale
81
days
2014-03-24T17:41:57+00:00
rare_domain
URL: http://s3-ak.buzzfeed.com/static/2014-03/user_images/webdr04/24/13/micr…
greendale
2014-03-24T17:41:57+00:00
rare_domain
URL: http://s3-ak.buzzfeed.com/static/2014-03/user_images/webdr04/24/13/micr…
greendale


“ Need to do it more than once? Automate. ”
Hans Berggren - father of infamous Johan Berggren

timesketch

Digital Forensic Timeline Analysis

1 Search

2 Collaboration

3 Assisted analysis

timesketch

Digital Forensic Timeline Analysis

1 Search

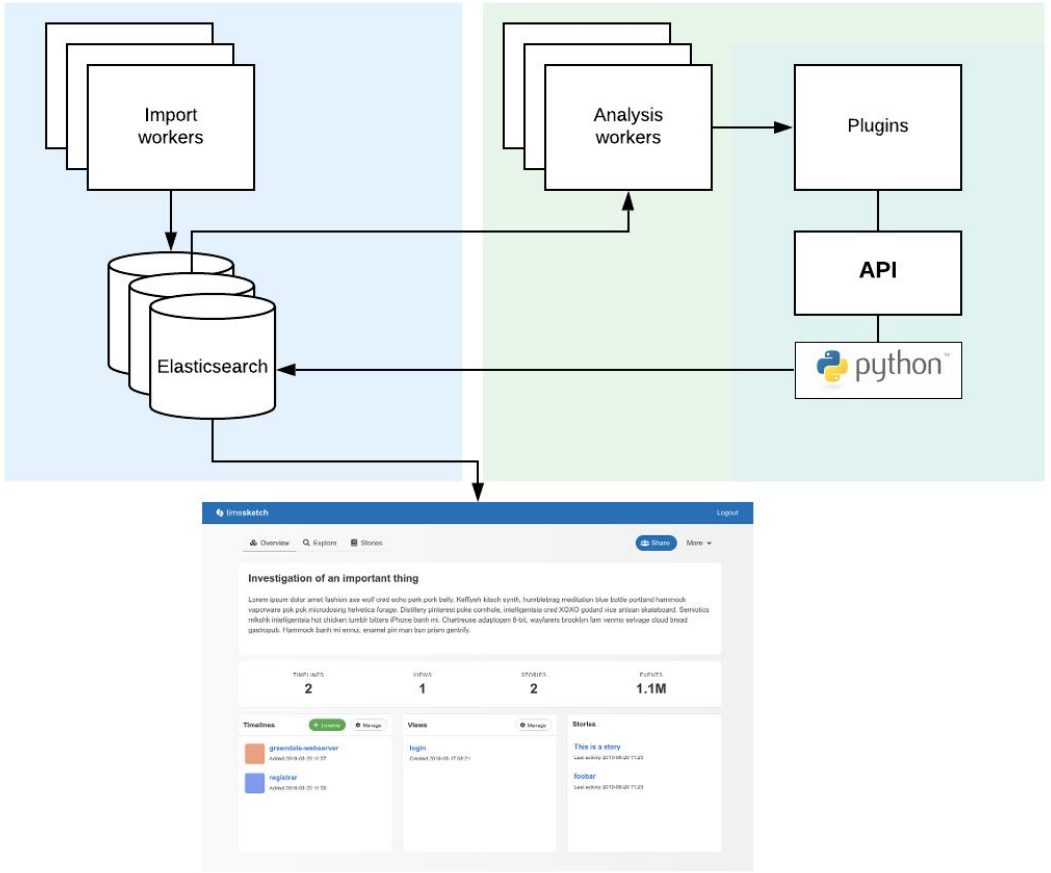
2 Collaboration

3 Assisted analysis

4 Encoded knowledge

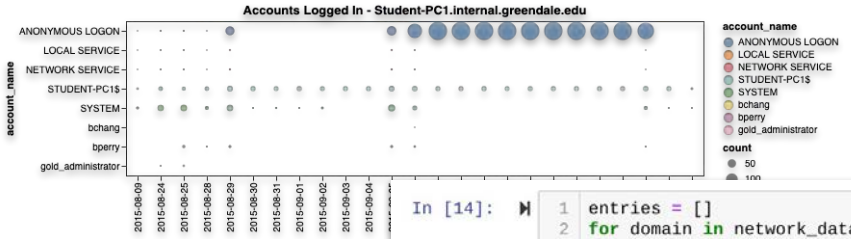
timesketch analyzers

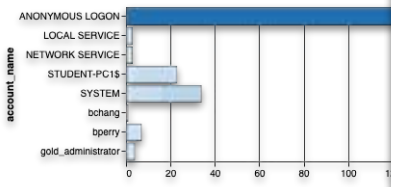
Plugin system for running analysis on a stream of events.



Just make it work

For Researchers





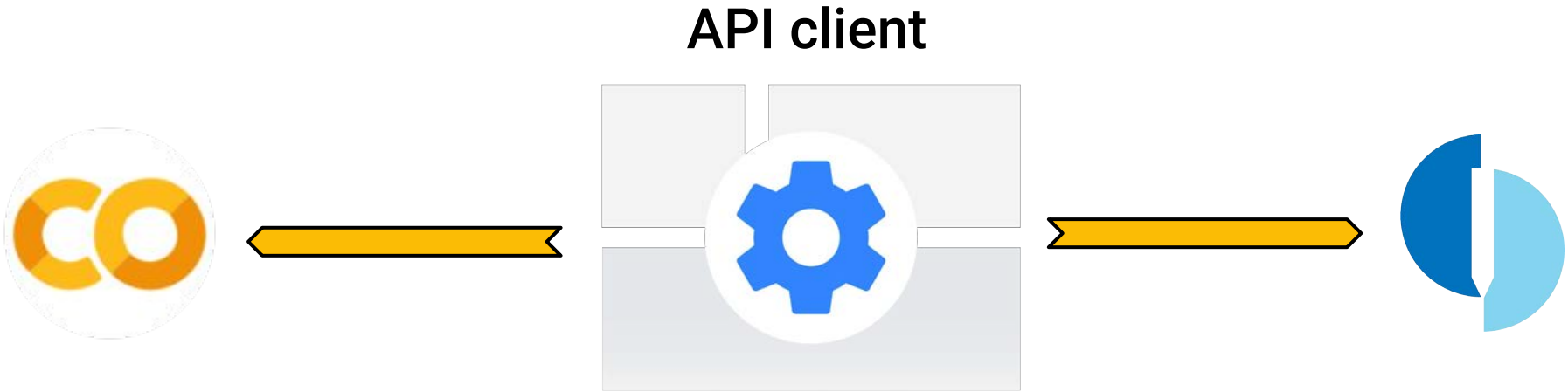

```
In [14]:
entries = []
for domain in network_data.domain.unique():
    similar_domains = _get_similar_do
    if not similar_domains:
        continue

    print('Domain: {0:s} does have si
    for similarities in similar_domai
        s_domain, s_score = similarit
        print('    [{0:s}] - {1:0.2f}%
        entry = {'domain': domain, 'wa
        entries.append(entry)
    print('---')
similar_domains = pd.DataFrame(entrie

Domain: s.yimg.com does have similar domai
    [ytimg.com] - 78.91%
    [i.ytimg.com] - 75.78%
---
Domain: l.yimg.com does have similar domai
    [ytimg.com] - 78.91%
```

```
In [15]:
similar_domains.sort_values('score', ascending=False)
```

Out[15]:

	domain	score	watched_domain
11	grendale.xyz	1.000000	greendale.xyz
13	tse1.mm.bing.net	1.000000	bing.com
0	s.yimg.com	0.789062	ytimg.com
2	l.yimg.com	0.789062	ytimg.com
14	l2.yimg.com	0.789062	ytimg.com
15	l1.yimg.com	0.789062	ytimg.com
3	static-entertainment-neu.s-msn.com	0.765625	msn.com
4	static-finance-neu.s-msn.com	0.765625	msn.com
5	static-sports-neu.s-msn.com	0.765625	msn.com
6	static-hp-neu.s-msn.com	0.765625	msn.com
7	static-hp-eus.s-msn.com	0.765625	msn.com
8	static-news-neu.s-msn.com	0.765625	msn.com
9	img.stb.s-msn.com	0.765625	msn.com
10	img.s-msn.com	0.765625	msn.com
1	s.yimg.com	0.757812	i.ytimg.com
12	tse1.mm.bing.net	0.750000	www.bing.com

For Researchers

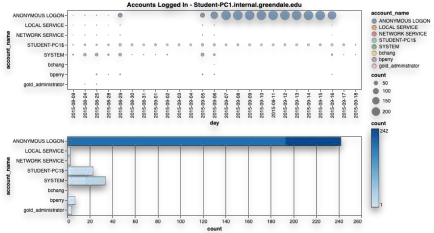


For Developers

Events

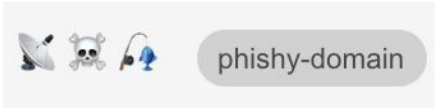
Events

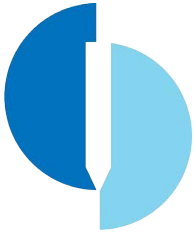
Events




```
In [15]:  1 similar_domains.sort_values('score', ascending=False)
Out[15]:
```

	domain	score	watched_domain
11	grendale.xyz	1.000000	greendale.xyz
13	tse1.mm.bing.net	1.000000	bing.com
0	s.yimg.com	0.789062	yimg.com
2	l.yimg.com	0.789062	yimg.com
14	l2.yimg.com	0.789062	yimg.com
15	l1.yimg.com	0.789062	yimg.com
3	static-entertainment-neu.s-msn.com	0.765625	msn.com
4	static-finance-neu.s-msn.com	0.765625	msn.com
5	static-sports-neu.s-msn.com	0.765625	msn.com
6	static-hp-neu.s-msn.com	0.765625	msn.com
7	static-hp-eus.s-msn.com	0.765625	msn.com
8	static-news-neu.s-msn.com	0.765625	msn.com
9	img.stb.s-msn.com	0.765625	msn.com
10	img.s-msn.com	0.765625	msn.com
1	s.yimg.com	0.757812	l.yimg.com
12	tse1.mm.bing.net	0.750000	www.bing.com



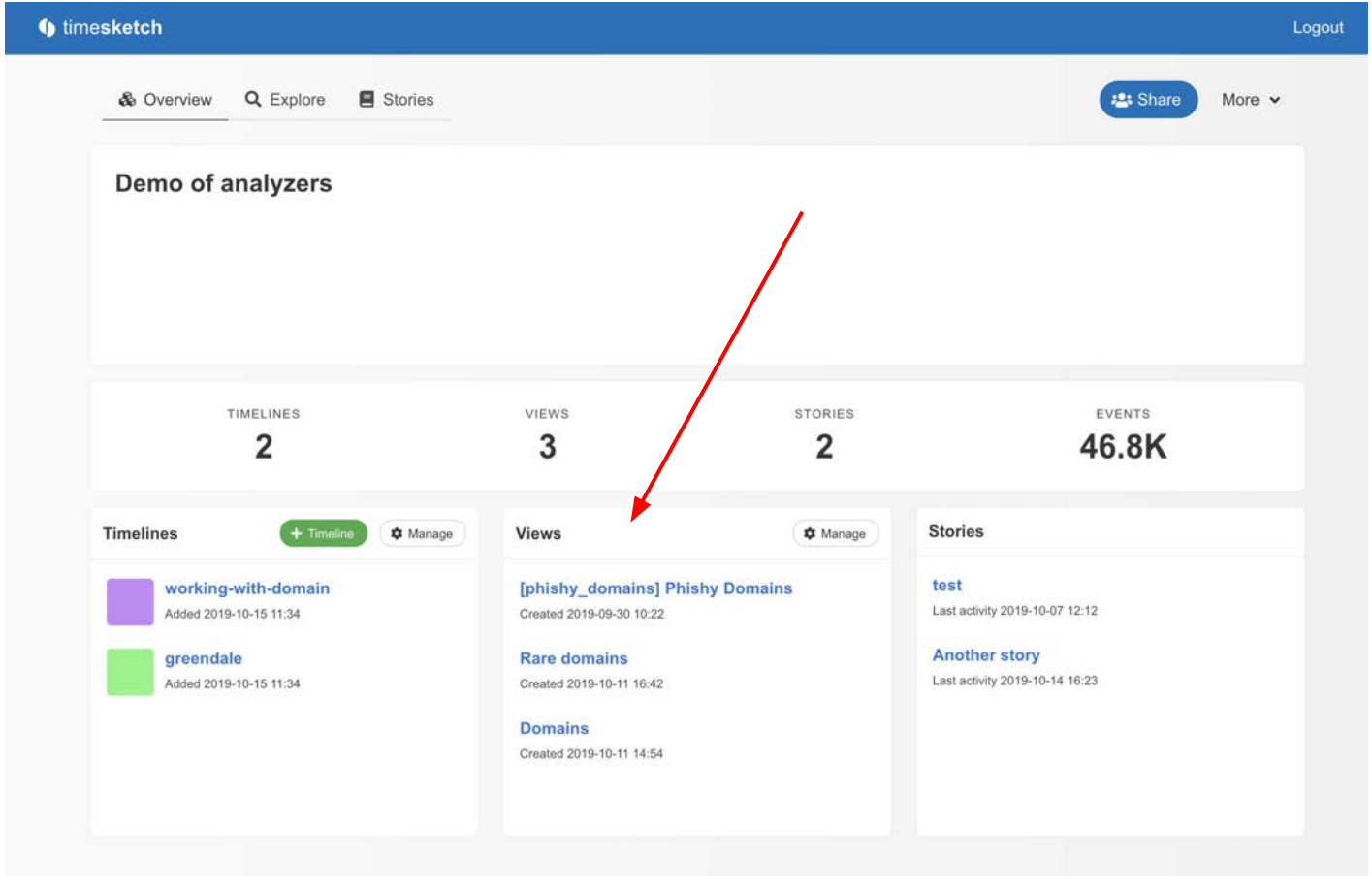


```
$ l2t_scaffolder.py
   == Starting the scaffolder ==
Gathering required information.

Path to the project root: <PATH TO TIMESKETCH SOURCE>
Path [.] set as the project path.

Name of the module to be generated. This can be something like "foobar sqlite" or
"event analytics".

This will be used for class name generation and file name prefixes.
Module Name: demo_analyzer
About to create a new feature branch to store newly generated code.
Creating feature branch: demo_analyzer inside .
Switching to feature branch demo_analyzer

Ready to generate files? [Y/n]:
```

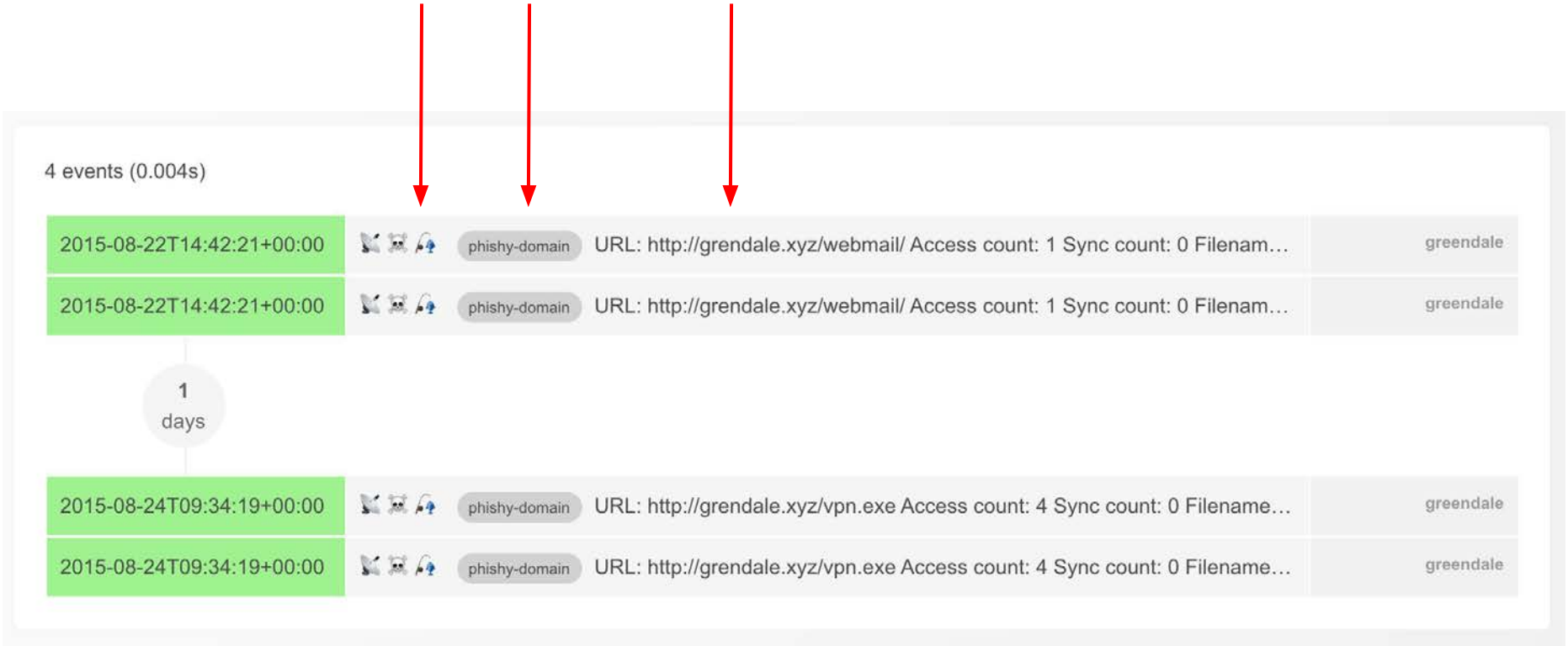
```
$ l2t_scaffolder.py
   == Starting the scaffolder ==
Gathering required information.

Path to the project root: <PATH TO TIMESKETCH SOURCE>
Path [.] set as the project path.

Name of the module to be generated. This can be something like "foobar sqlite" or
"event analytics".

This will be used for class name generation and file name prefixes.
Module Name: demo_analyzer
About to create a new feature branch to store newly generated code.
Creating feature branch: demo_analyzer inside .
Switching to feature branch demo_analyzer

Ready to generate files? [Y/n]: Y
File: ./timesketch/lib/analyzers/demo_analyzer.py written to disk.
File: ./timesketch/lib/analyzers/demo_analyzer_test.py written to disk.
```

```python
def run(self):
    """Entry point for the analyzer.

    Returns:
        String with summary of the analyzer result
    """
    # TODO: Add Elasticsearch query to get the events you need.
    query = ''

    # Generator of events based on your query.
    events = self.event_stream(query_string=query)

    # TODO: Add analyzer logic.
    # event.add_star()
    # event.add_comment('comment')
    for event in events:
        pass

    # TODO: Return a summary from the analyzer.
    return 'String to be returned'
```

For Analysts



For Analysts



Summary

> “Be that unicorn...”

Kristinn Guðjónsson

True believer of forensic unicorns

Any Questions?

(if we don't have time to answer, tackle us in the hallway)