When Data Talks
Yet Another Data Science Talk, Filled With Buzzwords
SANS DFIR Prague 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.
Data Scientist
Forensicator
Anything in Common With Forensics?

"...extract knowledge and insights from structured and unstructured data..."
What Can We Gain From Data Science?

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

```
static.greendale_frame[tag_string.str.contains('whitelisted-domain')].domain.value_counts()
```

- static.swissadsolutions.ch: 52
- mycdn.media.net: 8
- greendale.xyz: 8

Name: domain, dtype: int64
IPython Notebooks - Jupyter/Colab

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

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>datetime</strong></td>
<td><strong>timestamp_desc</strong></td>
<td><strong>url</strong></td>
<td><strong>human_readable</strong></td>
</tr>
<tr>
<td>2015-08-22 14:42:21+00:00</td>
<td>Content Modification Time</td>
<td><a href="http://grendale.xyz/webmail/">http://grendale.xyz/webmail/</a></td>
<td>greendale_slice.domain == 'grendale.xyz' [['datetime', 'timestamp_desc', 'url', 'human_readable']]</td>
</tr>
<tr>
<td>2015-08-24 09:34:19+00:00</td>
<td>Content Modification Time</td>
<td><a href="http://grendale.xyz/vpn.exe">http://grendale.xyz/vpn.exe</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 12:21:06+00:00</td>
<td>Creation Time</td>
<td><a href="http://grendale.xyz/webmail/">http://grendale.xyz/webmail/</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 12:21:06+00:00</td>
<td>Creation Time</td>
<td><a href="http://grendale.xyz/vpn.exe">http://grendale.xyz/vpn.exe</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 12:21:06+00:00</td>
<td>Last Access Time</td>
<td><a href="http://grendale.xyz/webmail/">http://grendale.xyz/webmail/</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 12:21:06+00:00</td>
<td>Synchronization time</td>
<td><a href="http://grendale.xyz/webmail/">http://grendale.xyz/webmail/</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 13:08:11+00:00</td>
<td>Creation Time</td>
<td><a href="http://grendale.xyz/vpn.exe">http://grendale.xyz/vpn.exe</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 13:08:13+00:00</td>
<td>Synchronization time</td>
<td><a href="http://grendale.xyz/vpn.exe">http://grendale.xyz/vpn.exe</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
<tr>
<td>2015-08-29 15:32:17+00:00</td>
<td>Last Access Time</td>
<td><a href="http://grendale.xyz/vpn.exe">http://grendale.xyz/vpn.exe</a></td>
<td>[phisy_domains] Domain greendale.xyz is similar to greendale.xyz [score: 1.00], [domain] grenda...</td>
</tr>
</tbody>
</table>
Can I Play?
Live Demo

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

Bring together different sources of time annotated events to reveal a chronological trail.
Incident response team members collaborate to analyze timelines - together.
Need to do it more than once? Automate.

Hans Berggren - father of infamous Johan Berggren
Plugin based system for running analysis on a stream of events.
Just make it work
```python
In [14]:
entries = []
for domain in network_data.domain.unique():
    similar_domains = _get_similar_domains(domain, watched_domains)
    if not similar_domains:
        continue
    print('Domain: {} does have similar domains:
'.format(domain))
    for similarities in similar_domains:
        s_domain, s_score = similarities
        print('  [{0:s}] - 1:0.2f
'.format(s_domain))
    entry = {'domain': domain, 'watched_domain': watched_domain}
    entries.append(entry)
print('-----')
similar_domains = pd.DataFrame(entries)

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

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

```
Out[15]:
       domain     score   watched_domain
   11  greendale.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  0.766625            msn.com
    4  static-finance-neu.s-msn.com  0.766625            msn.com
    5  static-sports-neu.s-msn.com  0.766625            msn.com
    6  static-hp-neu.s-msn.com  0.766625            msn.com
    7  static-hp-eus.s-msn.com  0.766625            msn.com
    8  static-news-neu.s-msn.com  0.766625            msn.com
    9       ling.szb.s-msn.com  0.766625            msn.com
   10       img.s-msn.com  0.766625            msn.com
     1      s.yimg.com  0.757812            yimg.com
   12   tse1.mm.bing.net  0.750000           www.bing.com
```
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)