

Data Science for DFIR

Data Science Lightning Summit

Jess Garcia

One eSecurity Founder | SANS Senior Instructor

@j3ssgarcia

\$ whoami

+14 y - **(one)**_{eSecurity} - Founder & Global DFIR Lead

+18 y - **SANS** - Senior Instructor

FOR500, FOR508, FOR585, FOR610, FOR578, ...

+25 y - CybSec / DFIR Experience



Why Do I Need DS?

Open source / Commercial tools

They are not flexible
enough

They are not scalable

They are limited and limit
your analysis “imagination”



That's why we use scripting!

But this can be improved

Most Forensicators Already Are Data Scientists



elasticsearch

Flexible, but Slow and Complex

Fast but Limited

```
cat fstimeline.tl.full.csv | grep '^Sat\|^Sun' | awk '{if ($4 ==  
    "2020") print $0}' | awk -F',' '{if ($2 > 1000) print $0}' |  
awk '{if ($5 ~/^0[0-8]/) print $0}' | awk -F',' '{print $8 "," $7}'  
    | sed -e 's/.*\:\/\/' -e 's\/\///' | sort -t ',' -nrk2
```

What We Have Been Doing Can Be Maintained & Simplified

- `df = read_csv('file.csv')`
- `df[col].str.contains(regex)`

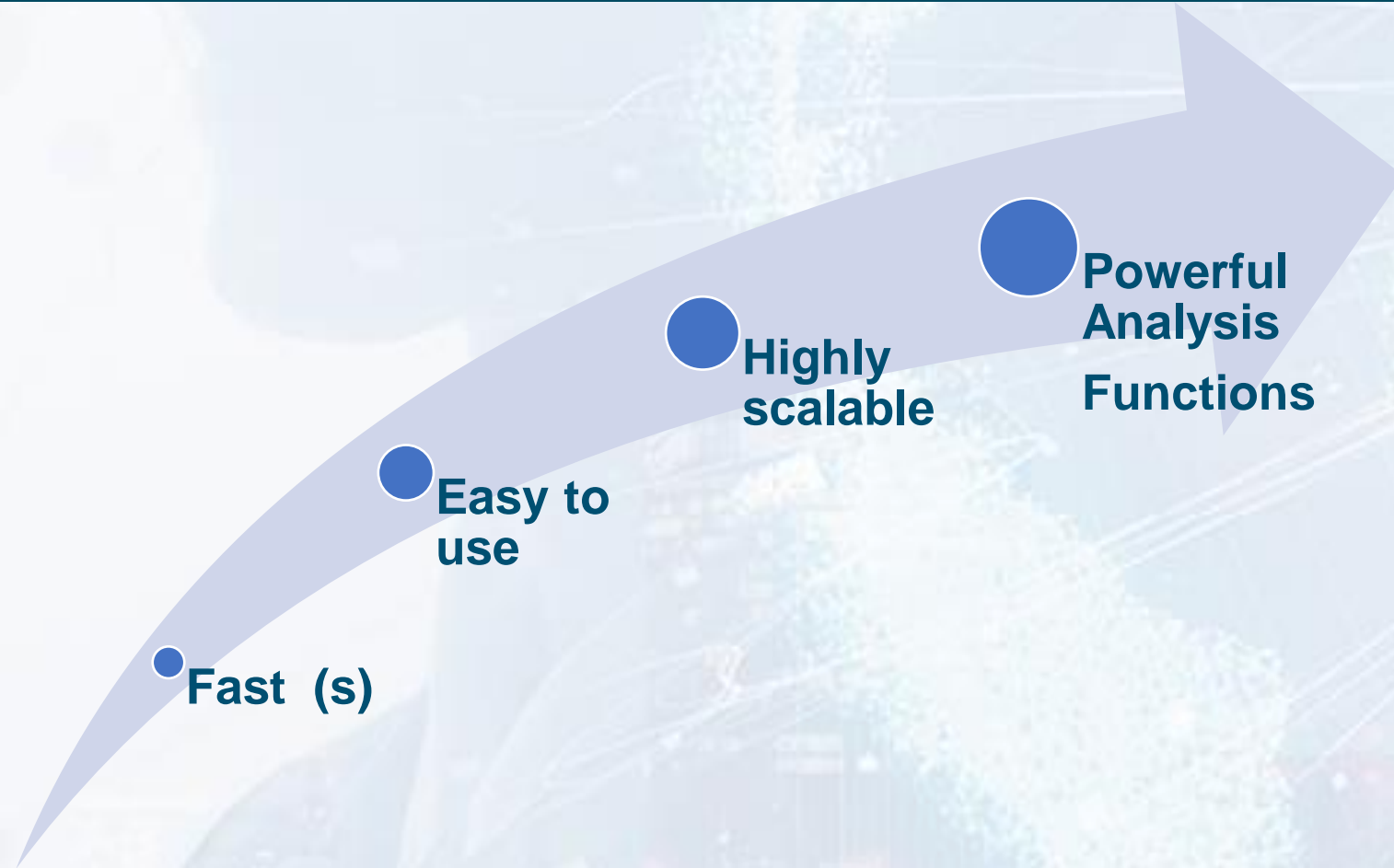
- `df['col'].sort_values()`
- `df.plot()`

- `df['2021-01-03':'2021-01-06']`

Linux pipes
.
=
|

- `df.query('col >= 123 and col2.str.contains("a")')`
- `df.query('col.str.contains("regex"))[col].str.replace('str1','str2').value_counts()`
- `pd.pivot_table(df, values='D', index=['A', 'B'], columns=['C'], aggfunc=np.sum)`

What We Have Been Doing Can Be Improved



DS / AI on What?



Threat Hunting



DFIR



CTI



Malware

What Projects Already Exist?

Logs



Roberto Rodriguez @Cyb3rWard0g



TimeSketch / Picatrix



Elasticsearch



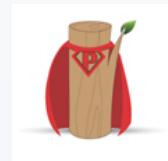
ML



ClearCut Columbo

What Are The Current Challenges?

Heterogeneous Tools



Heterogeneous Outputs



Lack of Standardized Processes



It is very time consuming to import data in pandas

What If You Use Other Tools: The DS4N6 Initiative

DS4N6

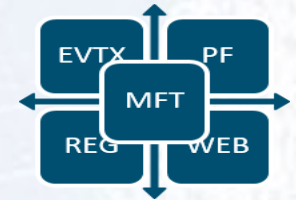
ds4n6.io



ds4n6_lib



D4ML



HAM



ADversAry
eMulator



Daisy VM

ds4n6 library

`df = xread_data(tool="kansa")`

`xmenu(df)`

`xanalyze(df)`

`df.simple()`

autoruns, plaso, kansa,
kape, mactime, volatility

DataFrame visualization menu:

Select DataFrame: 4624

Select grid: qgrid

Simple Output

Select max_rows: 20

Consolidate cols.

Apply Filters

Simple Options:

- 4608
- 4610
- 4611
- 4614
- 4616
- 4622
- 4624
- 4625

Statistics:

No. Entries: 11722

Constant Column:

Co	Value
0	D4_Orchestrator
1	Eve
2	D4_Data1
3	D4_Tool_
4	D4_Plugin_
5	D4_Hostname_
6	evtFileName_
7	ProviderName
8	ProviderGuid
9	System > EventID
10	Version
11	Level
12	Task
13	Opcode
14	Keywords
15	Channel
16	TransmittedServices

Access your selected DataFrame via the d4.out variable

[Export Grid to DataFrame](#)

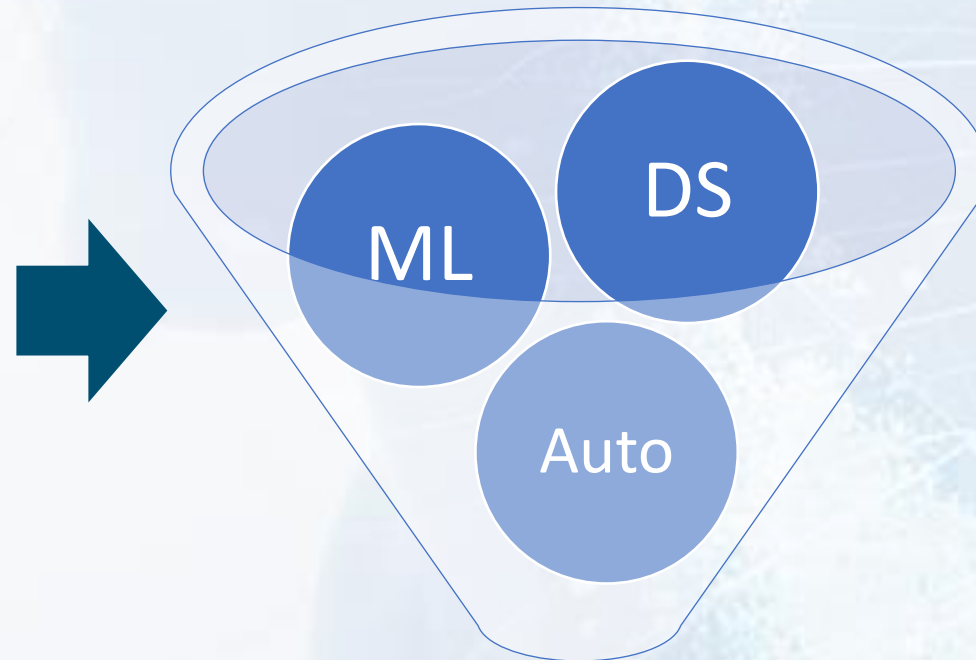
Timestamp	Selected	EventRecordID	ProcessID	ThreadID	Computer	@CorrelationActivity	SubjectUserSid	SubjectUserName	Si
2018-05-04 22:14:31		495	716	6724	win10-test	{69AA9DF0-E3D4-000...	S-1-5-18	WIN10-TESTS	TI
2018-05-04 22:14:31		499	716	6724	win10-test	{69AA9DF0-E3D4-000...	S-1-5-18	WIN10-TESTS	TI
2018-05-04 22:14:33		507	716	5988	win10-test	{69AA9DF0-E3D4-000...	S-1-5-18	WIN10-TESTS	TI
2018-05-04 22:14:34		511	716	5988	win10-test	{69AA9DF0-E3D4-000...	S-1-5-18	WIN10-TESTS	TI

In the Real World...

Data From July 2016-July2020

155M events
100 Gb Logs
1106 Users

2 month data
1.5M records
Different artifacts/logs
13875 Users



1 User
1 Week Suspicious Activity
< 50 Events

10 Suspects
Discard investigation lines

Coming Soon – D4ML @ RSA Conference Talk

ML for DFIR

DS4N6

RSA Conference 17th May - USA
**Me, My Adversary & AI:
Investigating & Hunting
with Machine Learning**

Jess Garcia | RSA Speaker
One eSecurity CEO
www.aidfir.io (Founder)
SANS Senior Instructor



**RSAC
2021**

DON'T MISS IT!

SANS

sans.org | **Jess Garcia** |  @j3ssgarcia | one-esecurity.com

**(one)
eSecurity**

How to Get Started?



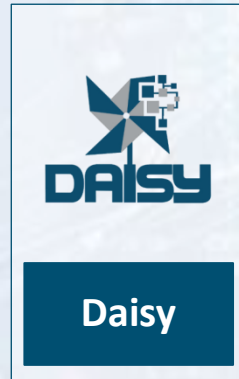
ds4n6.io

DS4N6 Knowledge

FAQ - Cheat Sheets - Tips & Tricks - Books - Articles / Blog Posts - Projects - Presentations

Coming soon...

April '21



May '21