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

Flexible, but Slow and Complex

```
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
```

Fast but Limited

What We Have Been Doing Can Be Maintained & Simplified

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

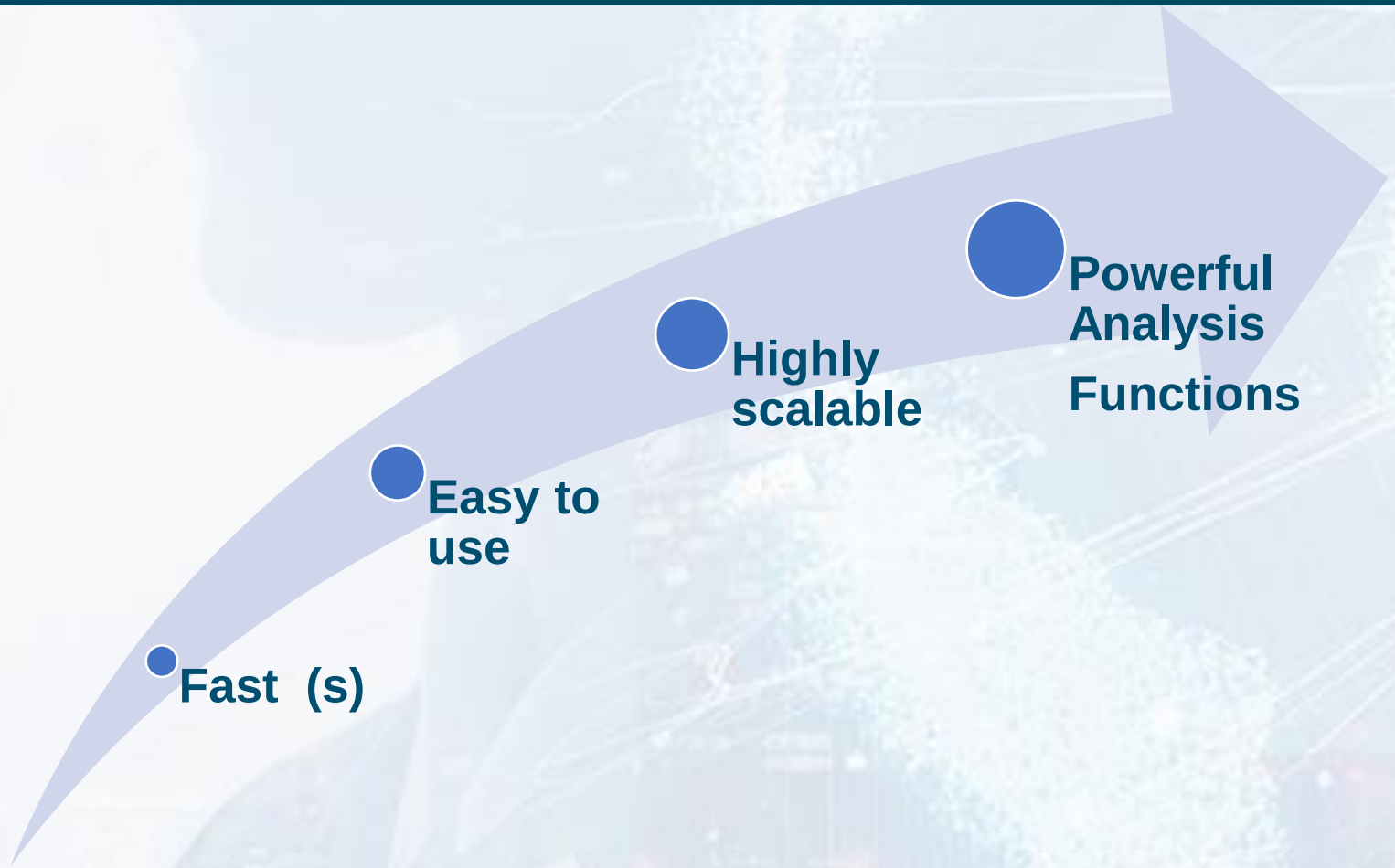
- df['col'].sort_values()
- df.plot()

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

Linux pipes
. = |

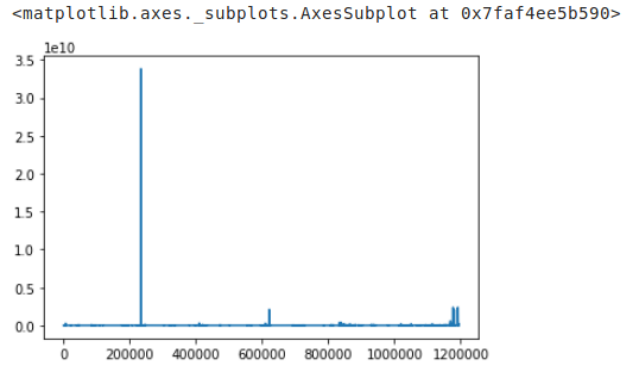
- df.query('col >= 123 and col2.str.contains("a")')
- df.query('col.str.contains("regex"))[col] .str.replace('str1','str2') .value_counts()
- pd.pivot_table(df, values='D', index=['A', 'B'], columns=['C'], aggfunc=np.sum)

What We Have Been Doing Can Be Improved

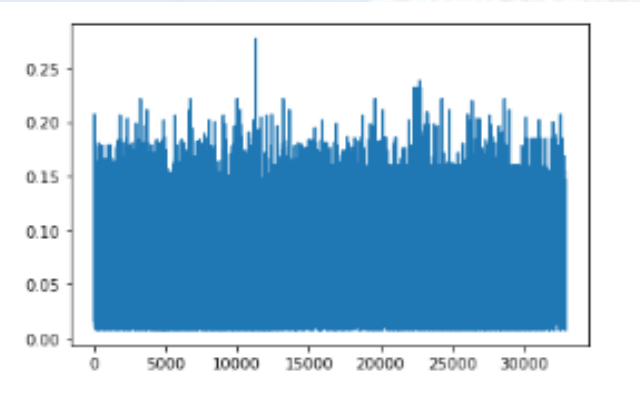


What Else Can We Do in DS That We Are Not Doing in DFIR?

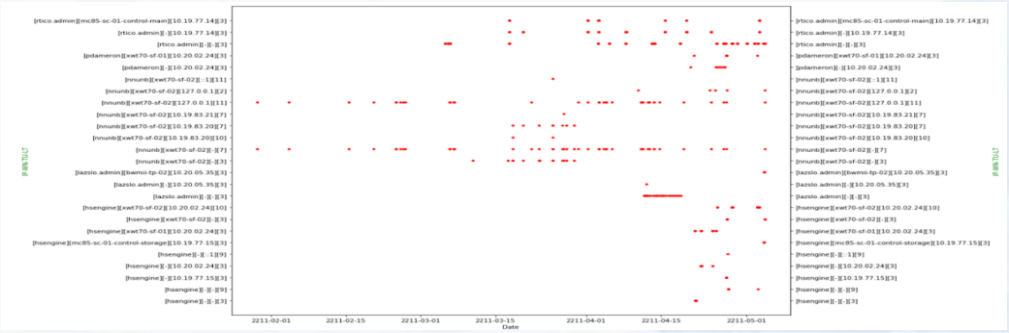
Outlier Analysis



Visualization



Logon Anomaly Analysis



Analysis Templates

- case_nb_template-10-fstl-main.ipynb
- case_nb_template-20-autoruns-main.ipynb
- case_nb_template-25-evtx_parser-main.ipynb

DS / AI on What?

Threat Hunting

DFIR

CTI

Malware

What Projects Already Exist?

Logs

EQL
Event Query Language

Roberto Rodriguez @Cyb3rWard0g

SIGMA

TimeSketch / Picatrix

Google

Velociraptor

Elasticsearch

elasticsearch

eland

ML

GitHub

ClearCut Columbo

What Are The Current Challenges?

Heterogeneous Tools

GRR RAPID RESPONSE

CROWDSTRIKE

KAPE

FireEye

davehull / Kansa

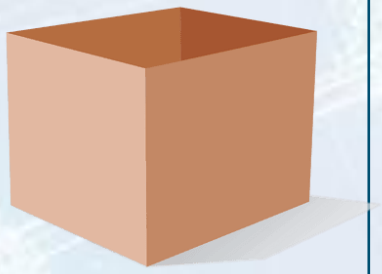
Heterogeneous Outputs

CSV

TXT

JSON

Lack of Standardized Processes

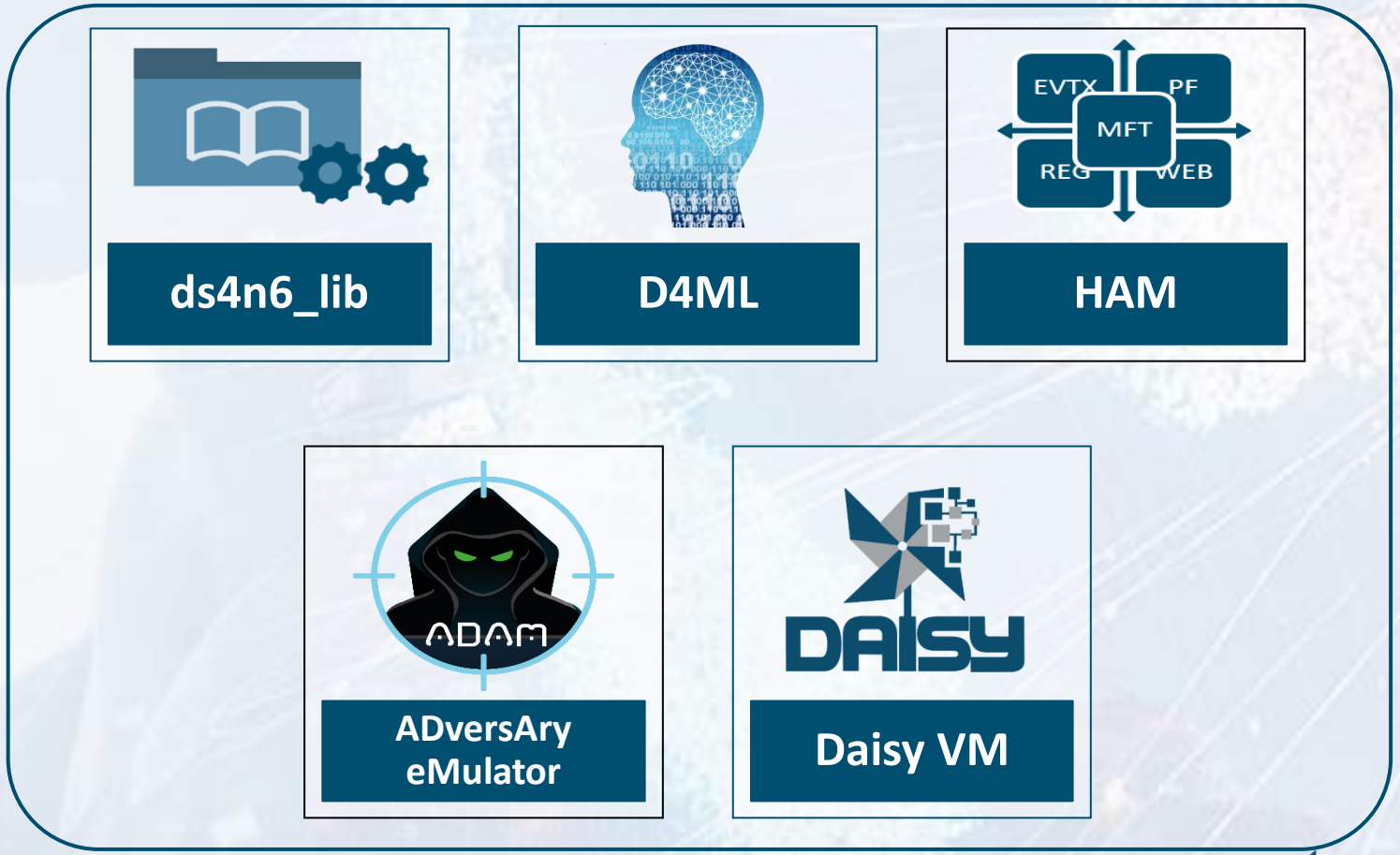


It is very time consuming to import data in pandas

What If You Use Other Tools: The DS4N6 Initiative

DS4N6

ds4n6.io



ds4n6 library

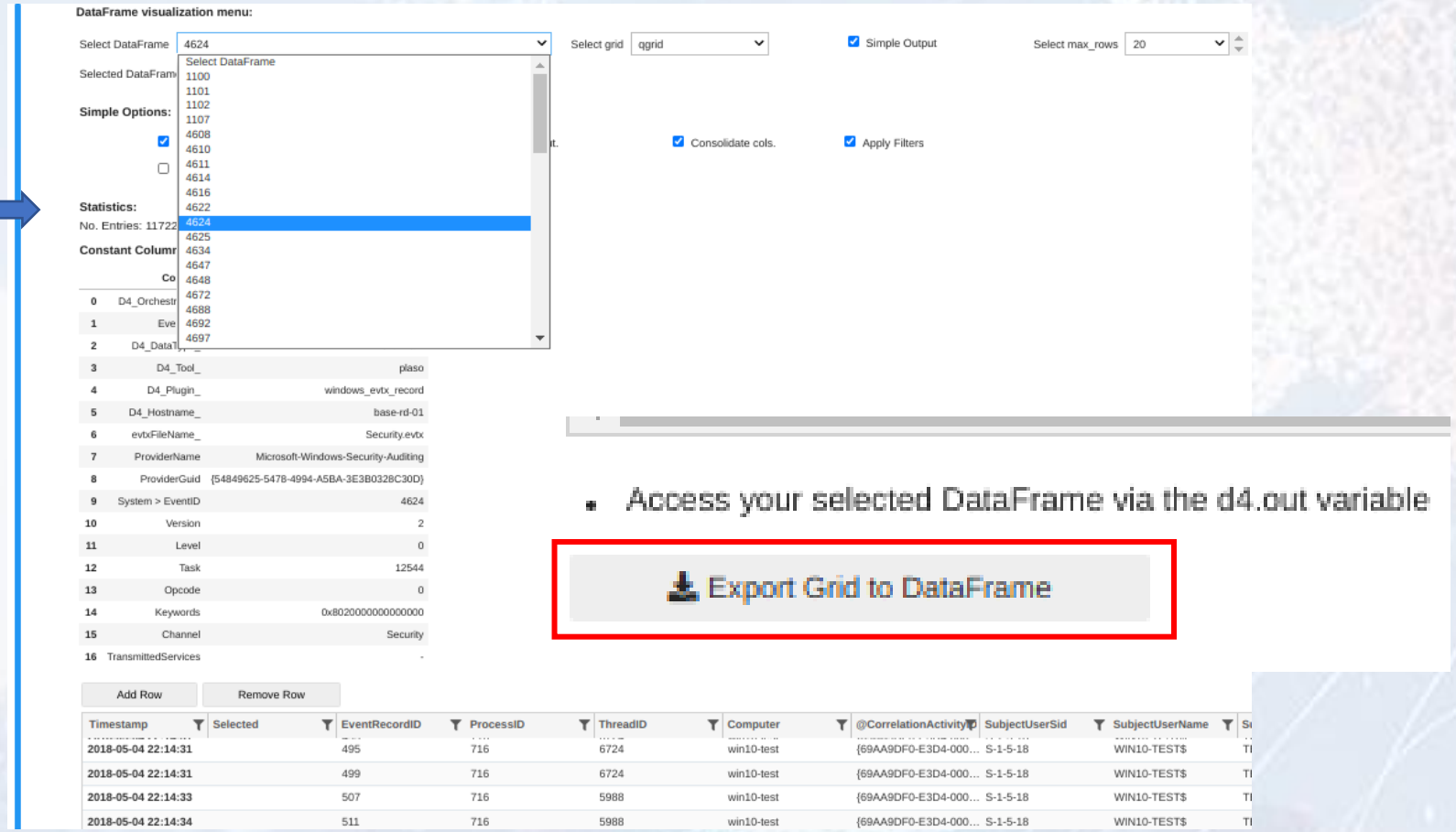
df = xread_data(tool="kansa")

xmenu(df)

xanalyze(df)

df.simple()

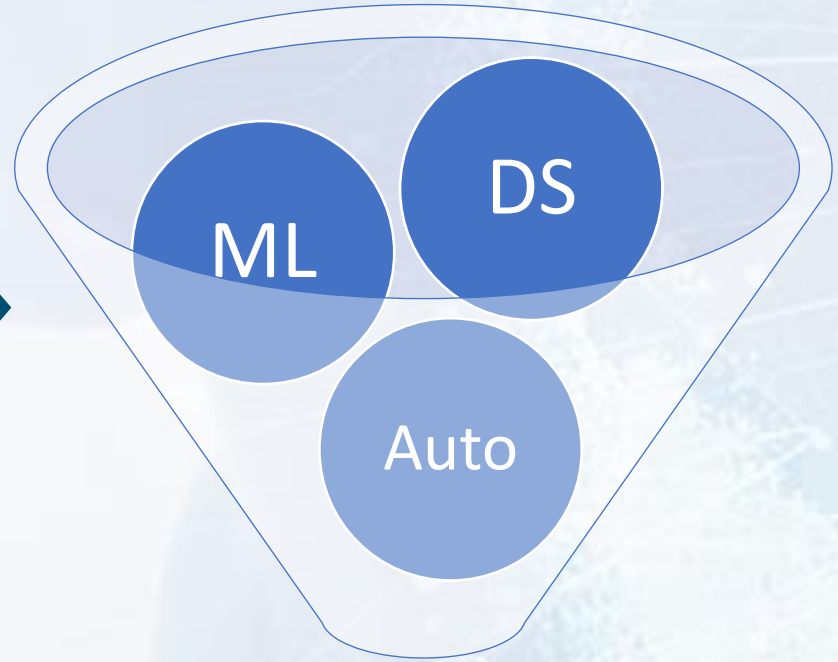
autoruns, plaso, kansa,
kape, mactime, volatility



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

RSAC 2021

DON'T MISS IT!

How to Get Started?

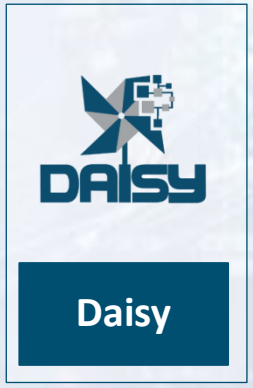
ds4n6.io

DS4N6 Knowledge

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

Coming soon…

April ‘21



May ‘21