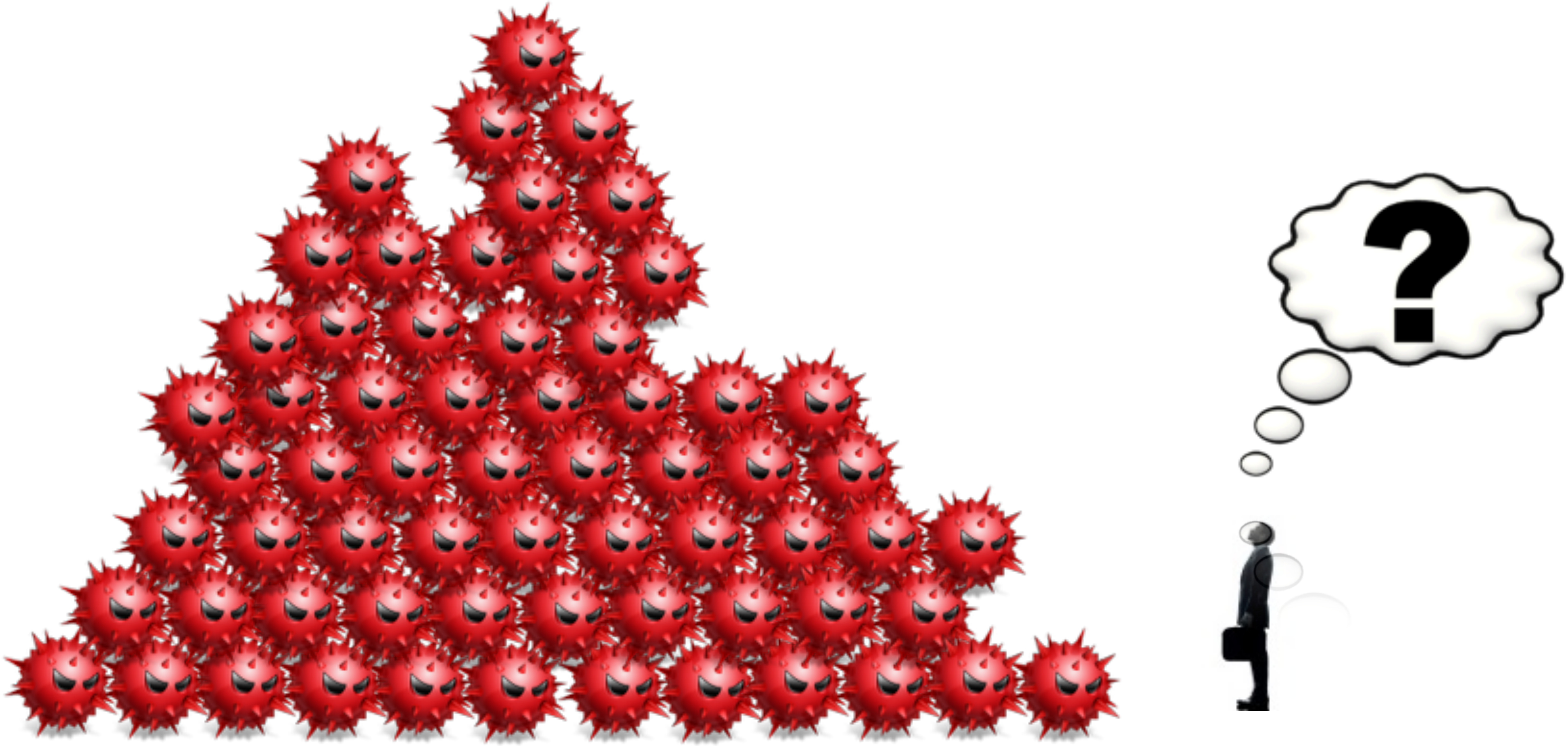
NEEDLES IN A HAYSTACK: MINING INFORMATION FROM PUBLIC DYNAMIC SANDBOXES FOR MALWARE INTELLIGENCE

Mariano Graziano, Davide Canali, Leyla Bilge,
Andrea Lanzi and Davide Balzarotti

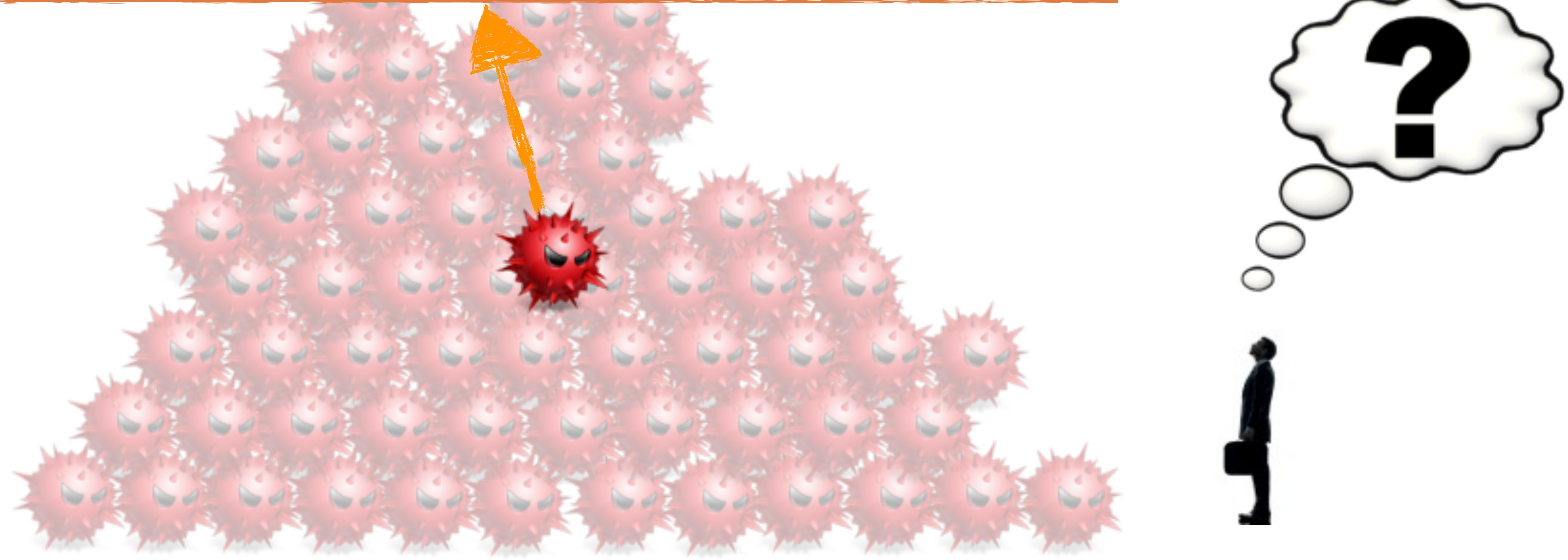
Eurecom
Symantec Research Labs
Università degli Studi di Milano

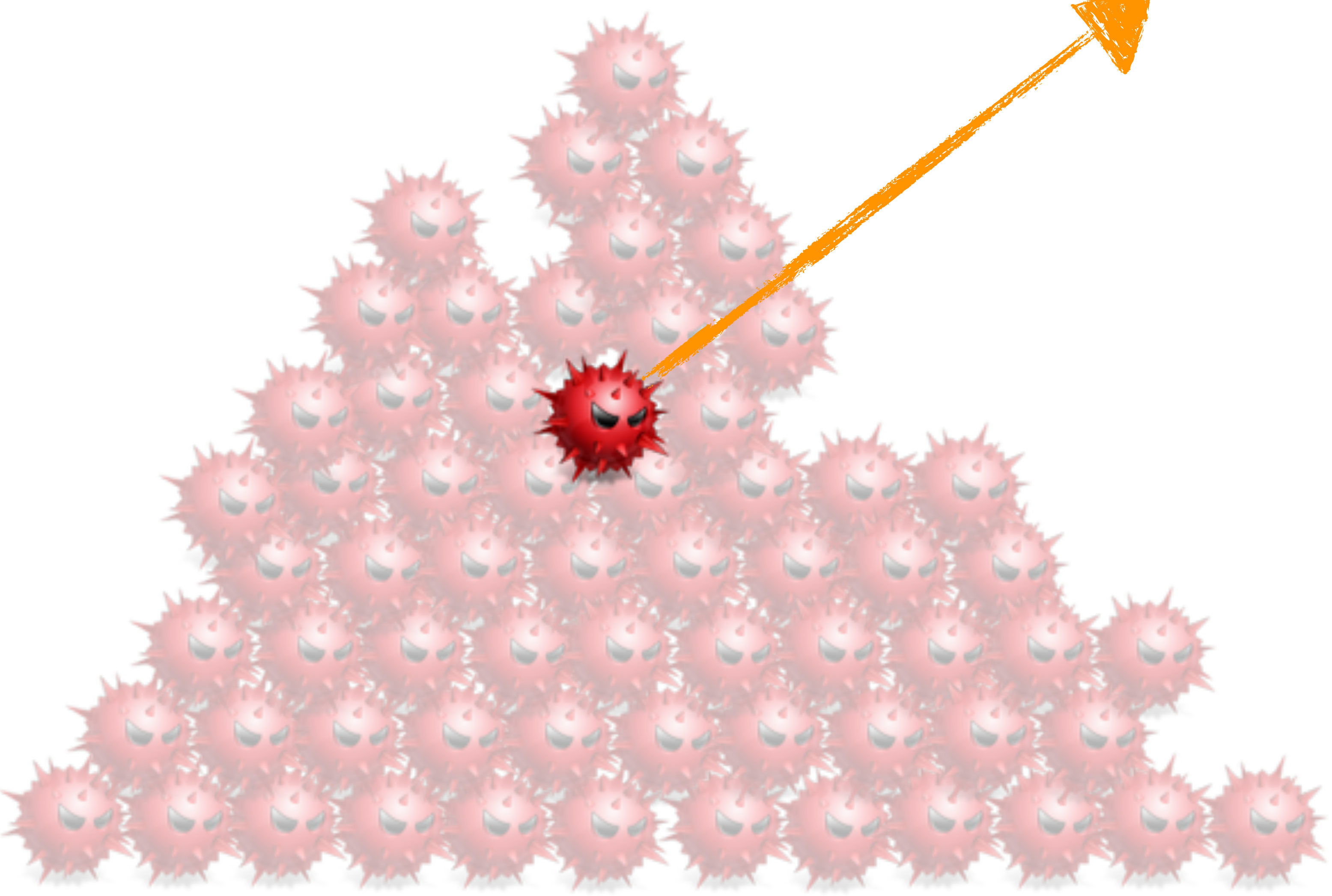
USENIX Security '15 - Washington DC, USA

A PILE OF MALWARE SAMPLES



CAMPAIGN	TIME BEFORE PUBLIC DISCLOSURE	SUBMITTED BY
Operation Aurora	4 months	US
Red October	8 months	Romania
APT1	43 months	US
Stuxnet	1 month	US
Beebus	22 months	Germany
LuckyCat	3 months	US
BrutePOS	5 months	France
NetTraveller	14 months	US
Pacific PluX	12 months	US
Pitty Tiger	42 months	US
Regin	44 months	UK
Equation	23 months	US



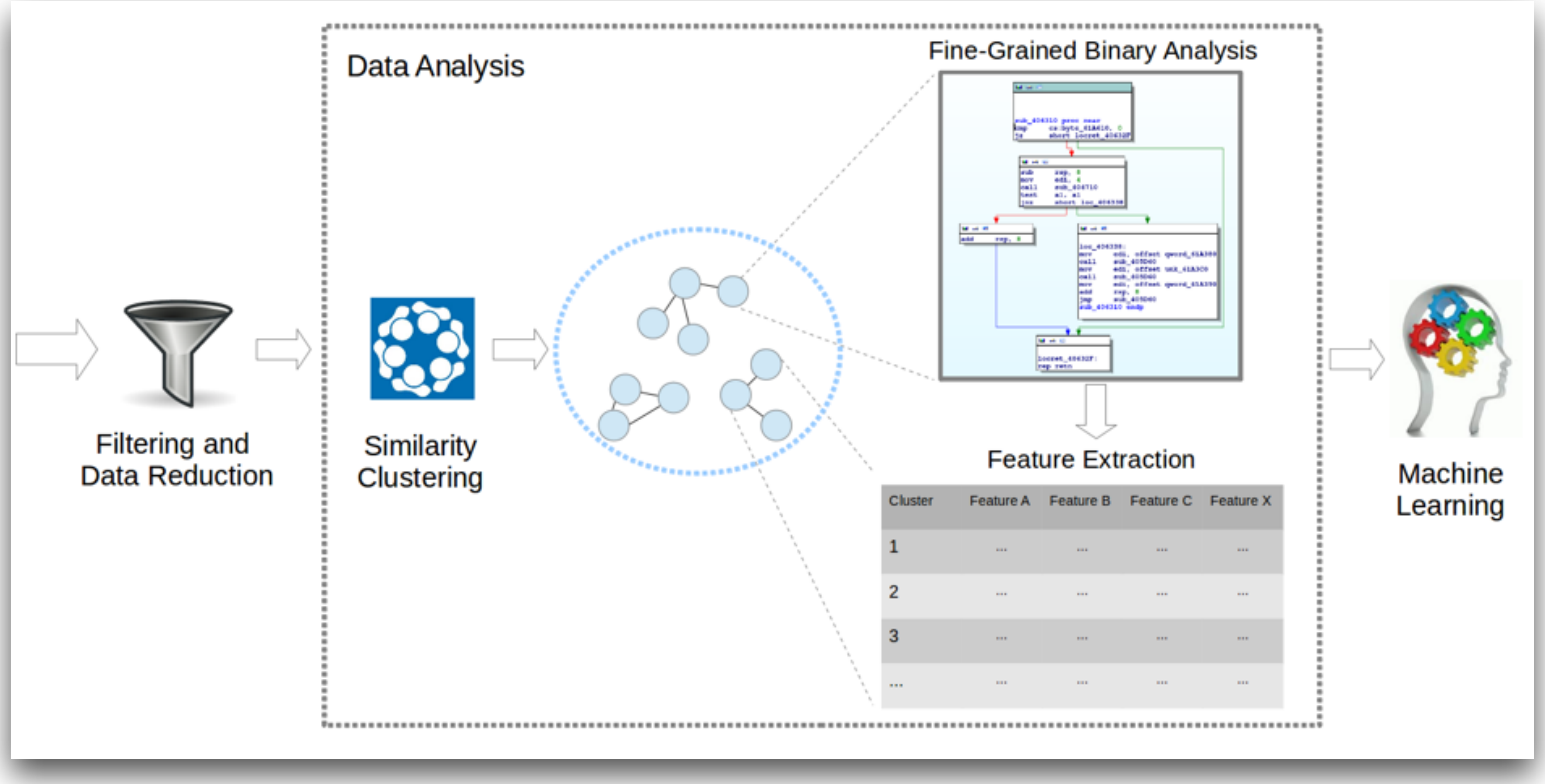


Constant interaction criminals vs sandbox

GOAL

- Observation: Malware authors use public sandboxes to test their developments
- Design data mining techniques to automatically discover malware developments

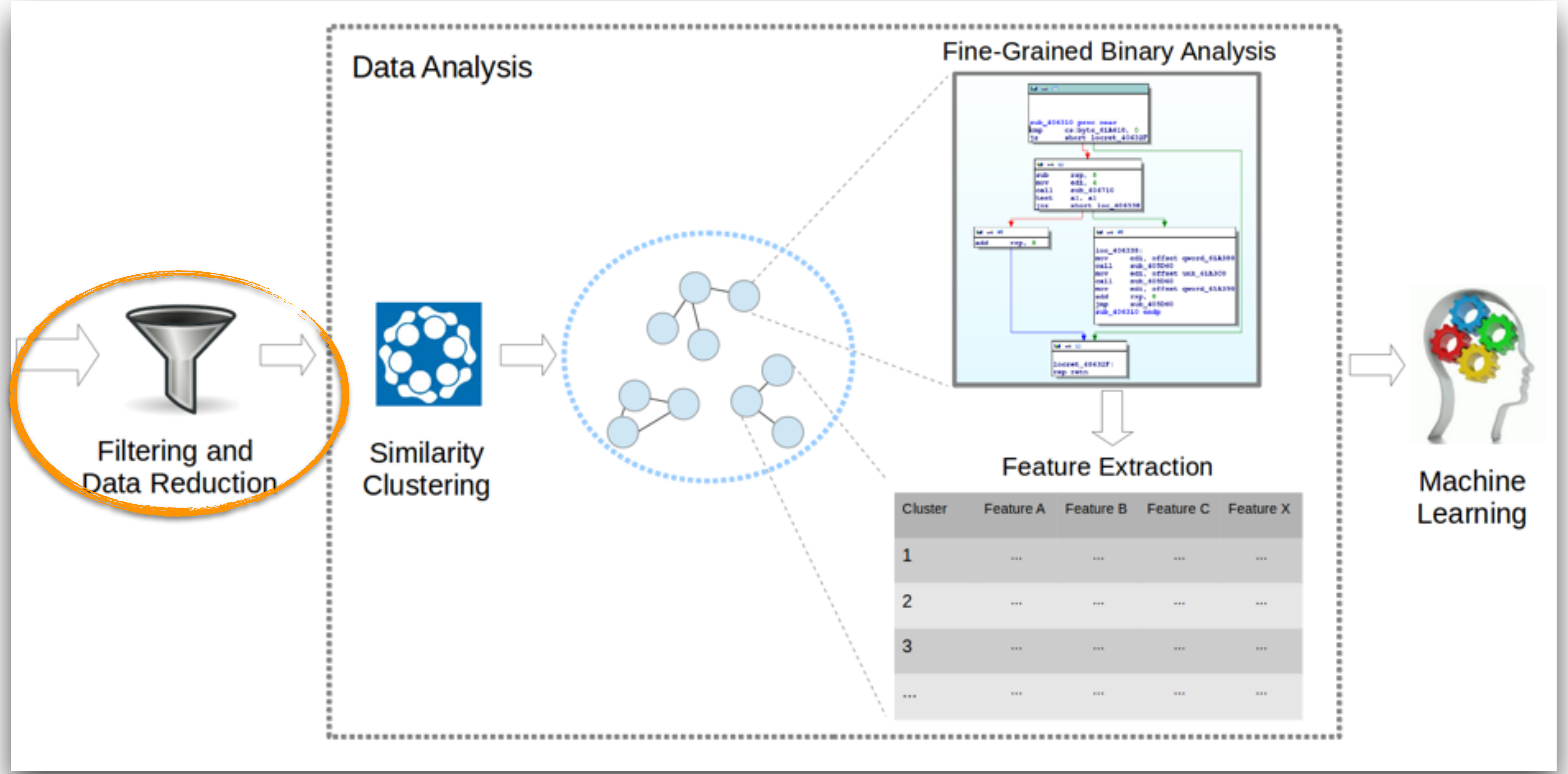
SYSTEM OVERVIEW



```
sub_404310 proc near
cmp     cs:byte_41A410, 0
jz      short locret_40432F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_404330

add     rsp, 8

loc_404330:
mov     edi, offset qword_41A380
call    sub_405D60
mov     edi, offset unk_41A3C0
call    sub_405D60
mov     edi, offset qword_41A390
add     rsp, 8
jmp     sub_405D60
sub_404310 endp

locret_40432F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

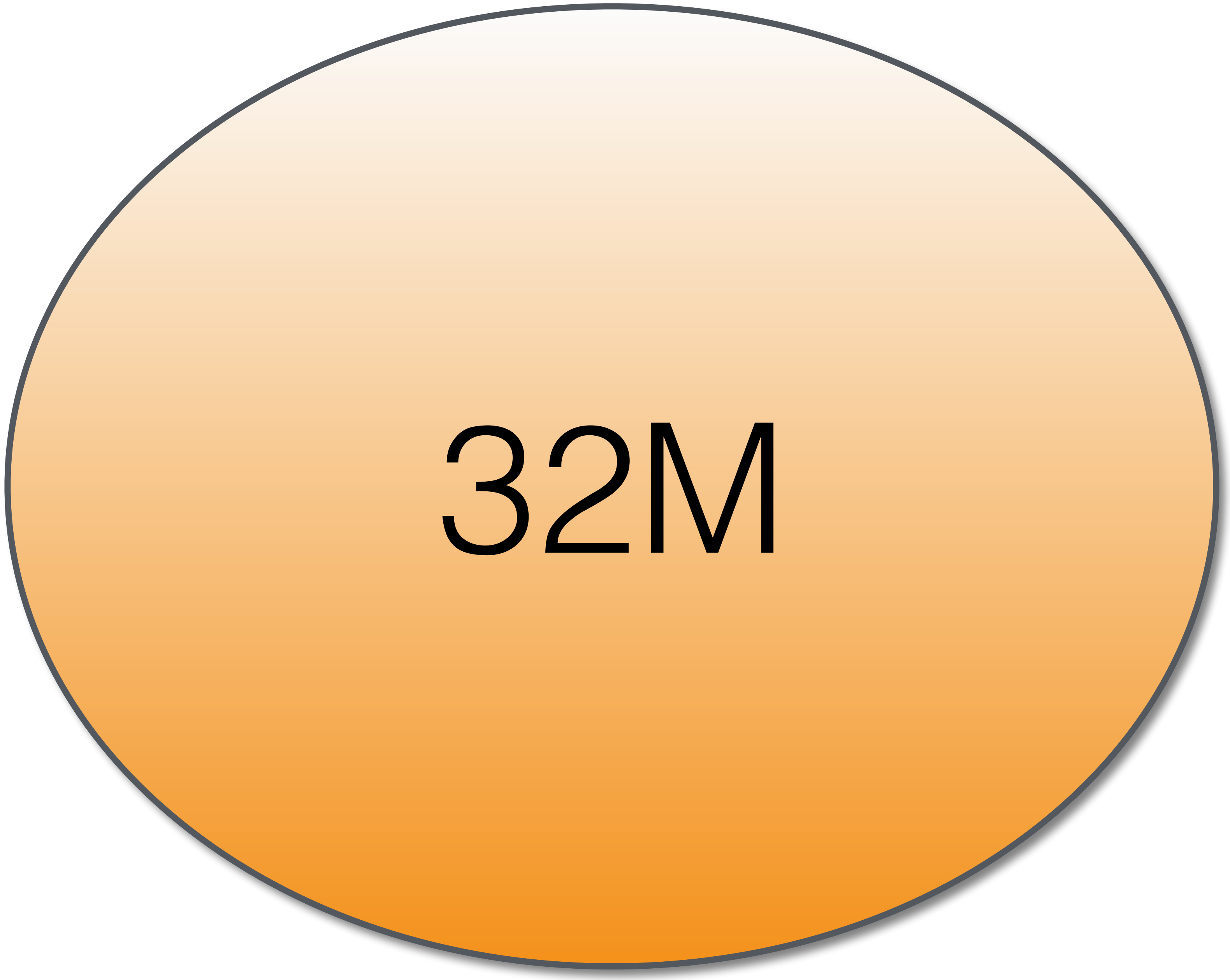
SYSTEM OVERVIEW



```
sub_404310 proc near
cmp     cs:byte_41A610, 0
jz      short locret_40432F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_404338

add     rsp, 8

loc_404338:
mov     edi, offset qword_41A380
call    sub_405D60
mov     edi, offset unk_41A3C0
call    sub_405D60
mov     edi, offset qword_41A390
add     rsp, 8
jmp     sub_405D60
sub_404310 endp

locret_40432F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

DATA REDUCTION



32M

Initial Dataset

DATA REDUCTION

6M

Submitted by regular users

DATA REDUCTION

522K

Not already part of large submissions

DATA REDUCTION

214K

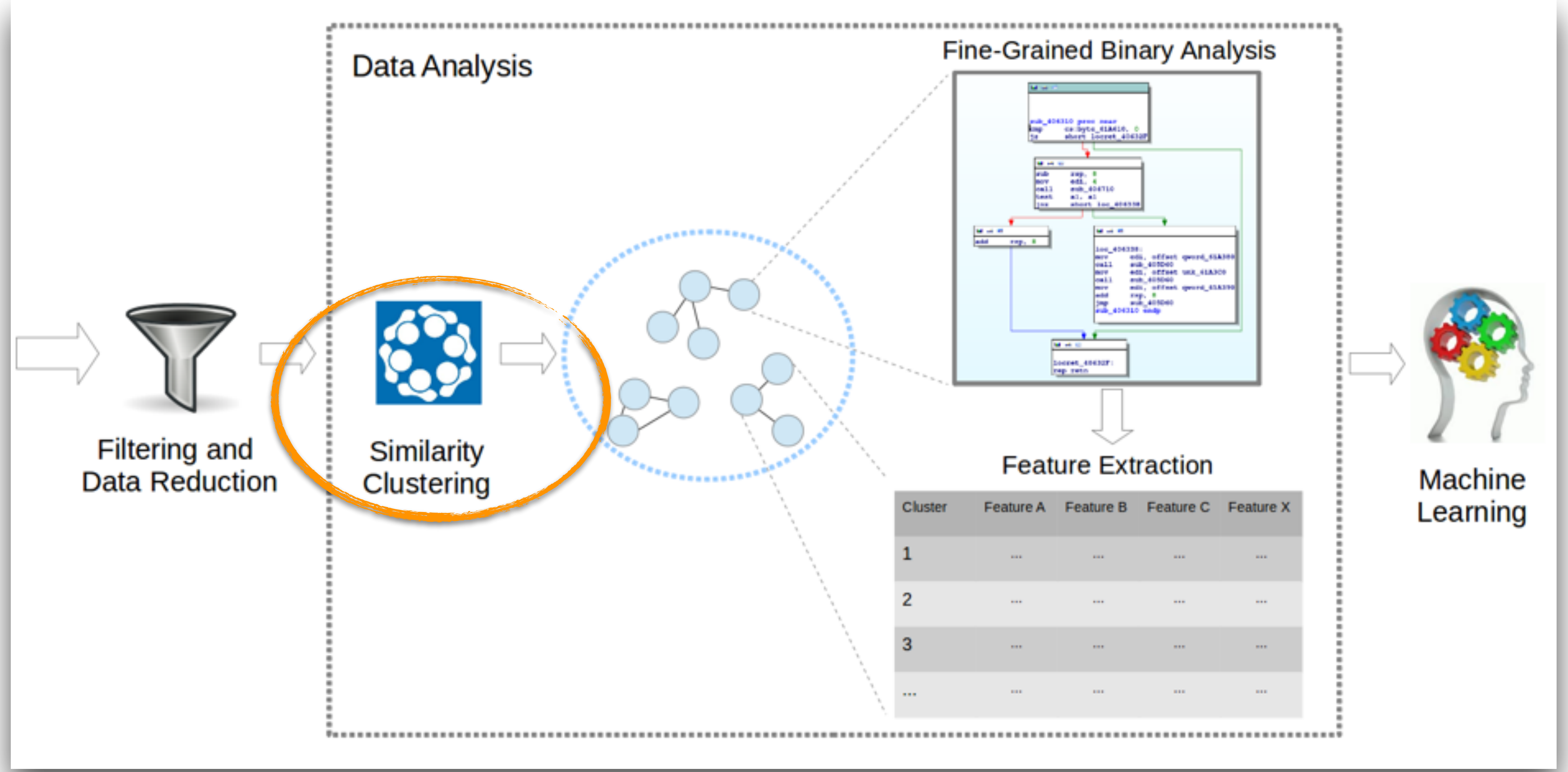
Previously unknown by Symantec & VirusTotal

DATA REDUCTION

121K

Final (not packed binary)

SYSTEM OVERVIEW



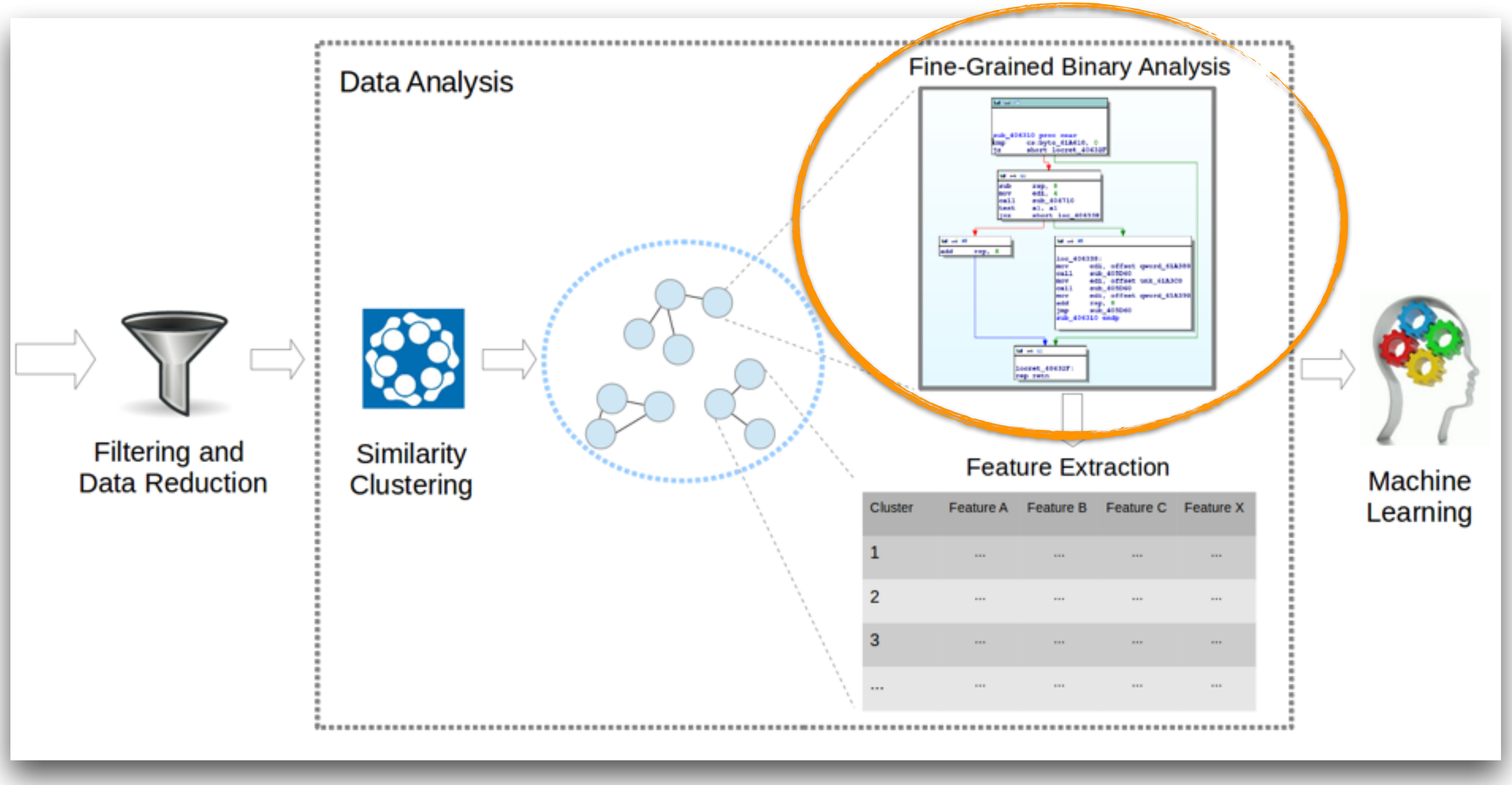
```
sub_406310 proc near
cmp     cs:byte_61A610, 0
jz      short locret_40632F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_406338

add     rsp, 8

loc_406338:
mov     edi, offset qword_61A380
call    sub_405D60
mov     edi, offset unk_61A3C0
call    sub_405D60
mov     edi, offset qword_61A390
add     rsp, 8
jmp     sub_405D60
sub_406310 endp

locret_40632F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

CLUSTERING

- Agglomerative clustering (*similarity threshold: 70%*):
  - Binary similarity (*ssdeep*)
  - Submissions metadata
- Sliding window of seven days:
  - Reduce comparisons
  - Ensure binary similarity
- 5972 clusters → 4.5 elements each

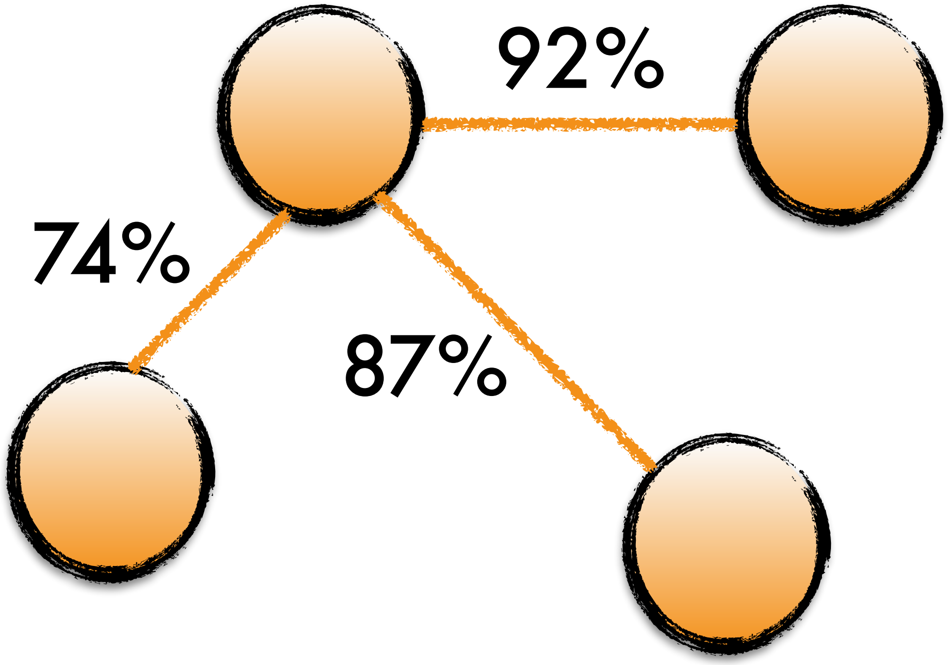
SYSTEM OVERVIEW

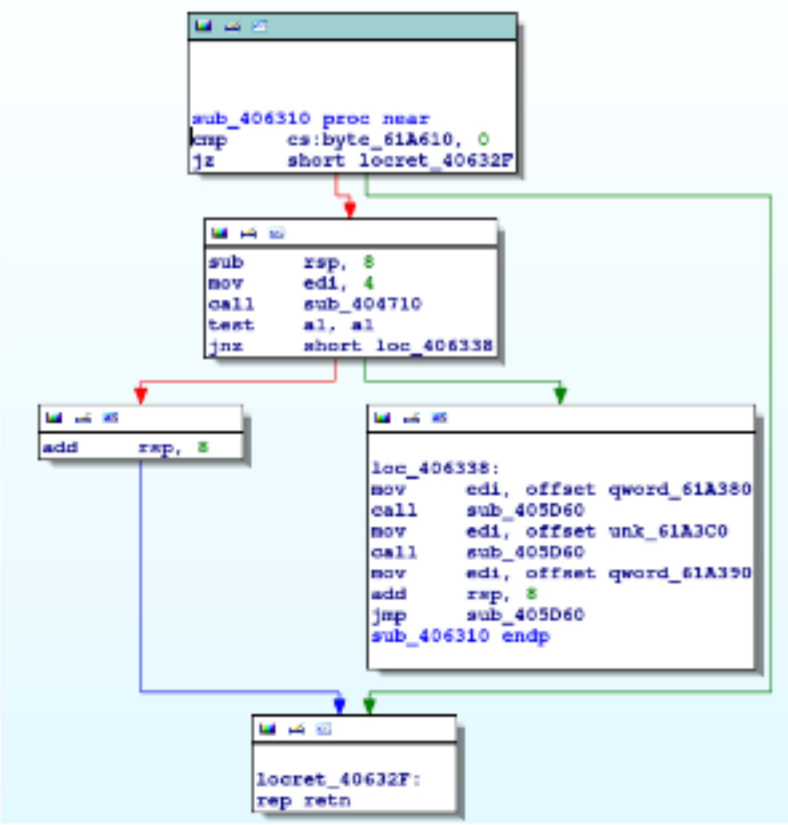


```
sub_406310 proc near
cmp     cs:byte_61A610, 0
jz      short locret_40632F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_406338

add     rsp, 8

loc_406338:
mov     edi, offset qword_61A380
call    sub_405D60
mov     edi, offset unk_61A3C0
call    sub_405D60
mov     edi, offset qword_61A390
add     rsp, 8
jmp     sub_405D60
sub_406310 endp

locret_40632F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

FINE-GRAINED ANALYSIS





```
sub_406310 proc near
cmp     cs:byte_61A610, 0
jz      short locret_40632F
```

```
sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_406338
```

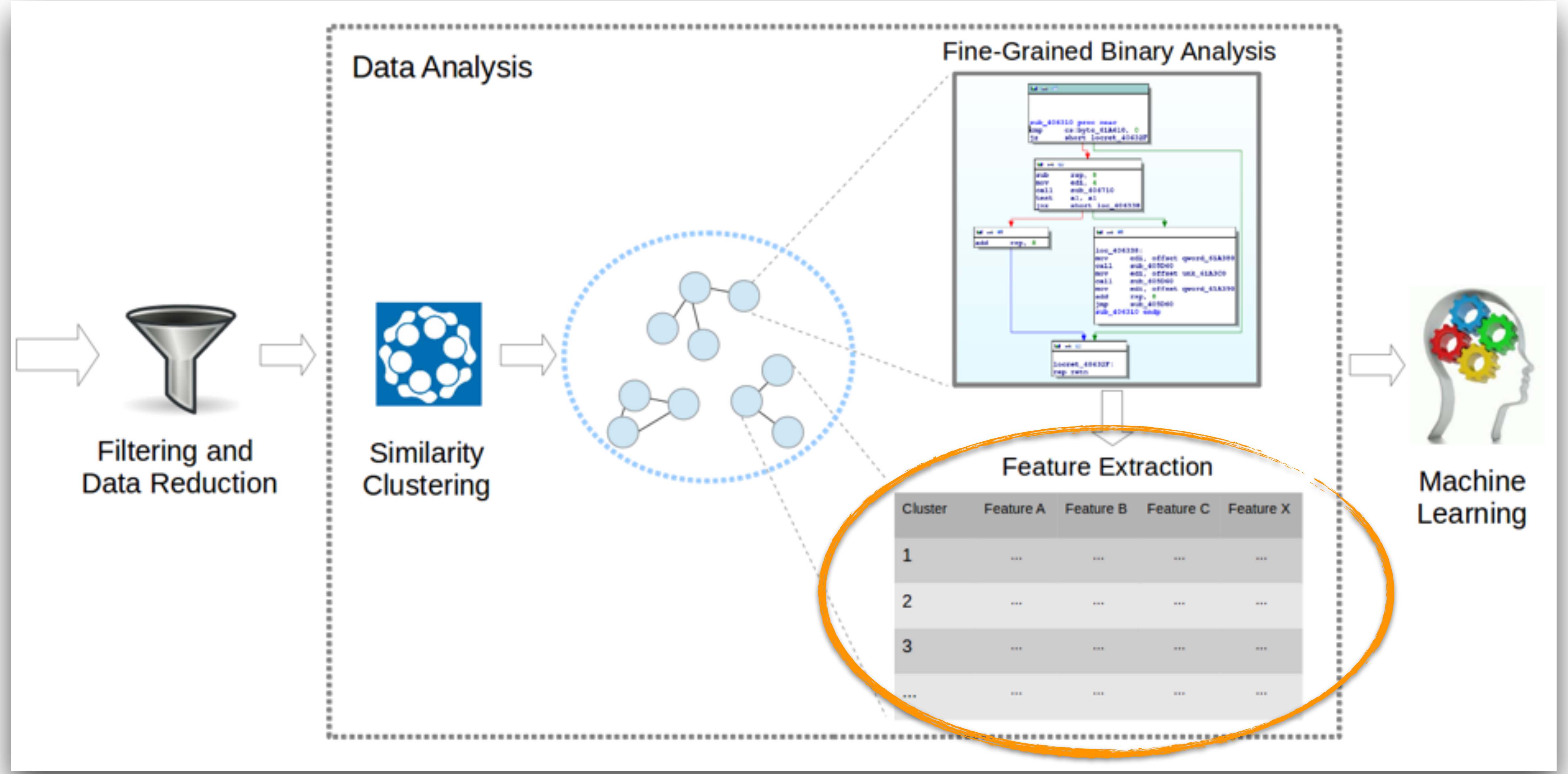
```
add     rsp, 8
```

```
loc_406338:
mov     edi, offset qword_61A380
call    sub_405D60
mov     edi, offset unk_61A3C0
call    sub_405D60
mov     edi, offset qword_61A390
add     rsp, 8
jmp     sub_405D60
sub_406310 endp
```

```
locret_40632F:
rep retn
```

- Binary code normalisation
- Call graph comparison [Flake04,Gao08]
- Control flow graph comparison [Flake04,Kruegel06,Jang13]

SYSTEM OVERVIEW



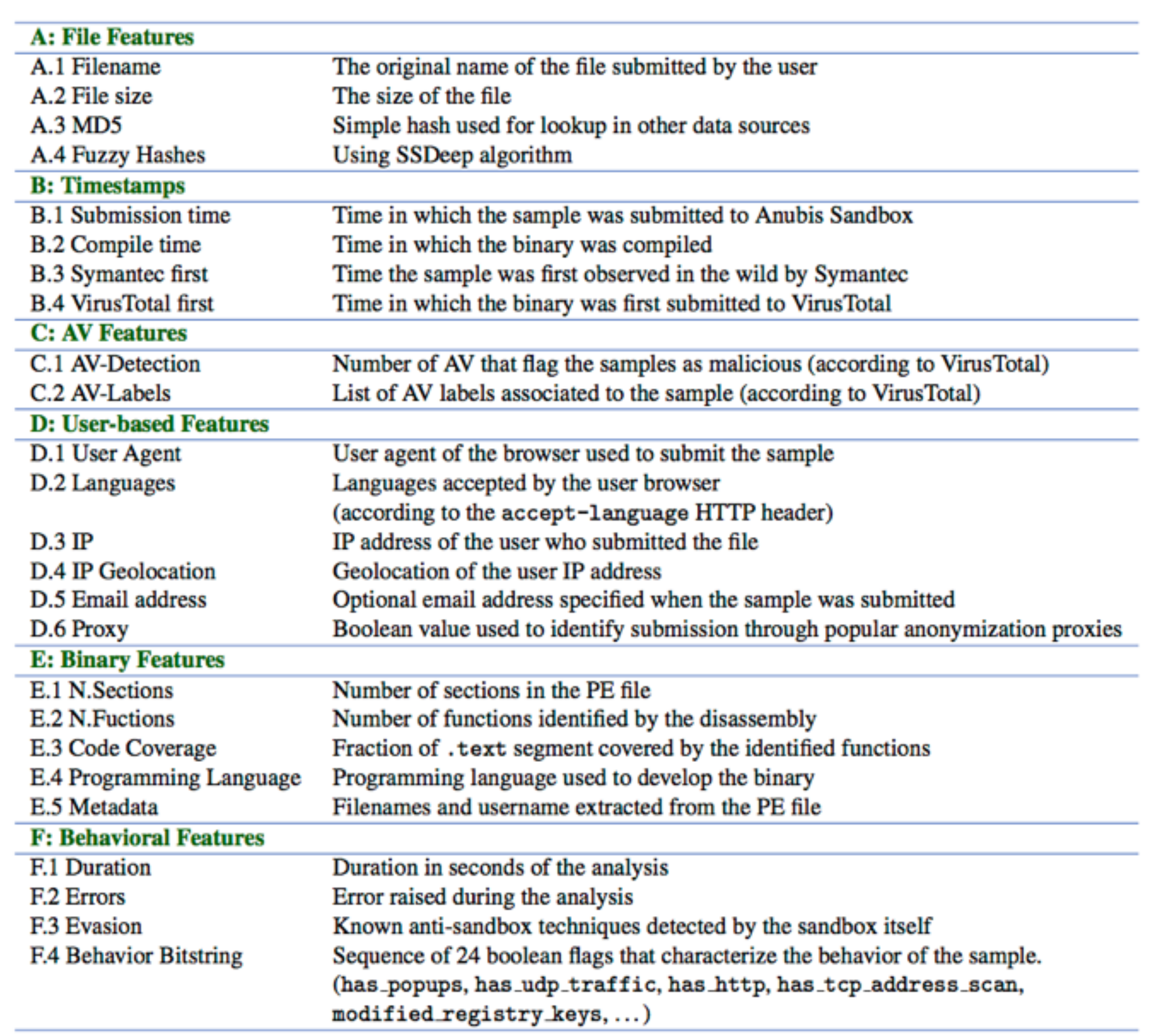
```
sub_406310 proc near
cmp     cs:byte_61A610, 0
jz      short locret_40632F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_406338

add     rsp, 8

loc_406338:
mov     edi, offset qword_61A380
call    sub_405D40
mov     edi, offset unk_61A3C0
call    sub_405D40
mov     edi, offset qword_61A390
add     rsp, 8
jmp     sub_405D40
sub_406310 endp

locret_40632F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

FEATURE EXTRACTION

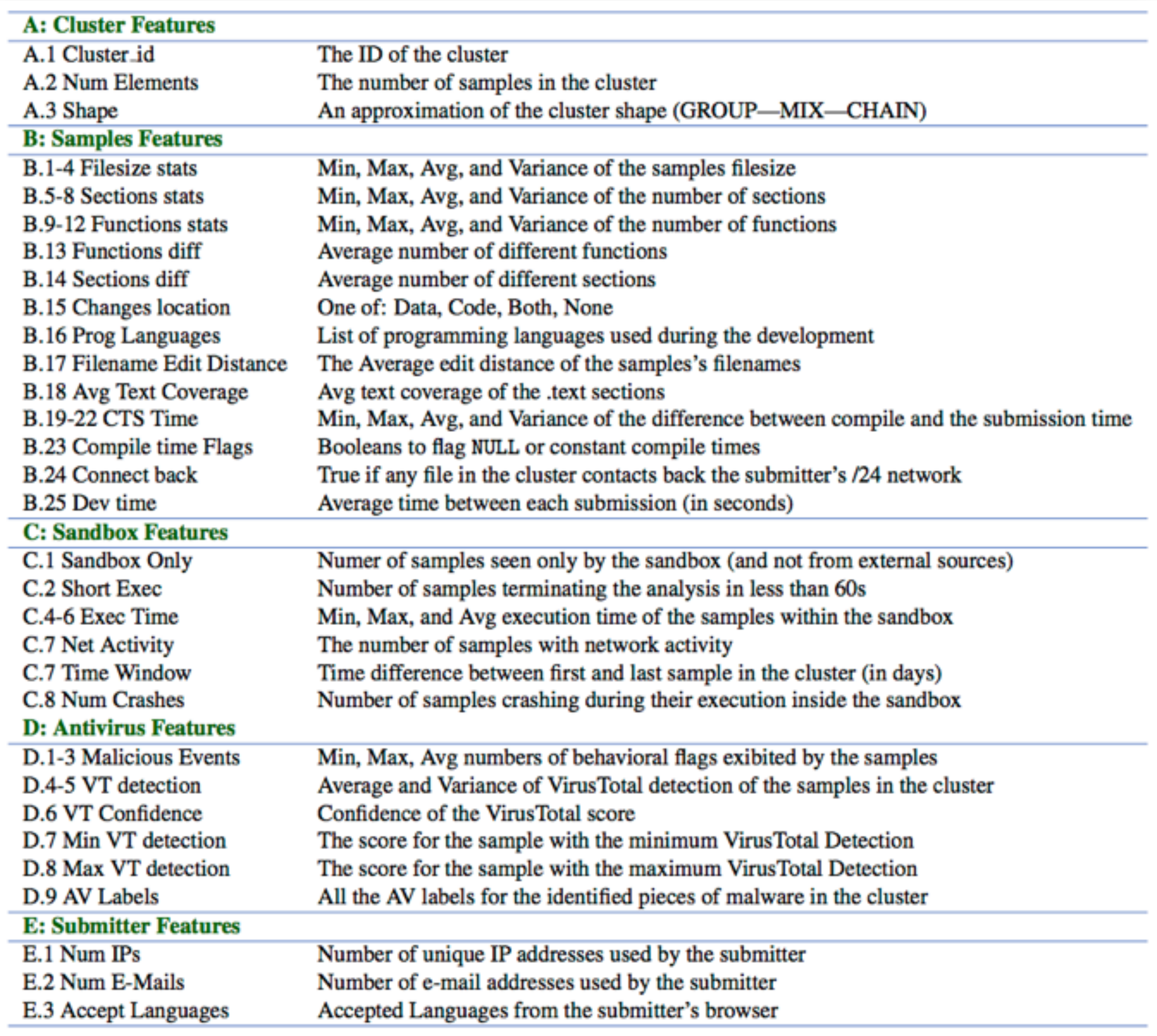
- Comprise two phases:
  - Per sample (25 features in 6 groups)
  - Per cluster (48 features in 5 groups)

SAMPLE FEATURES



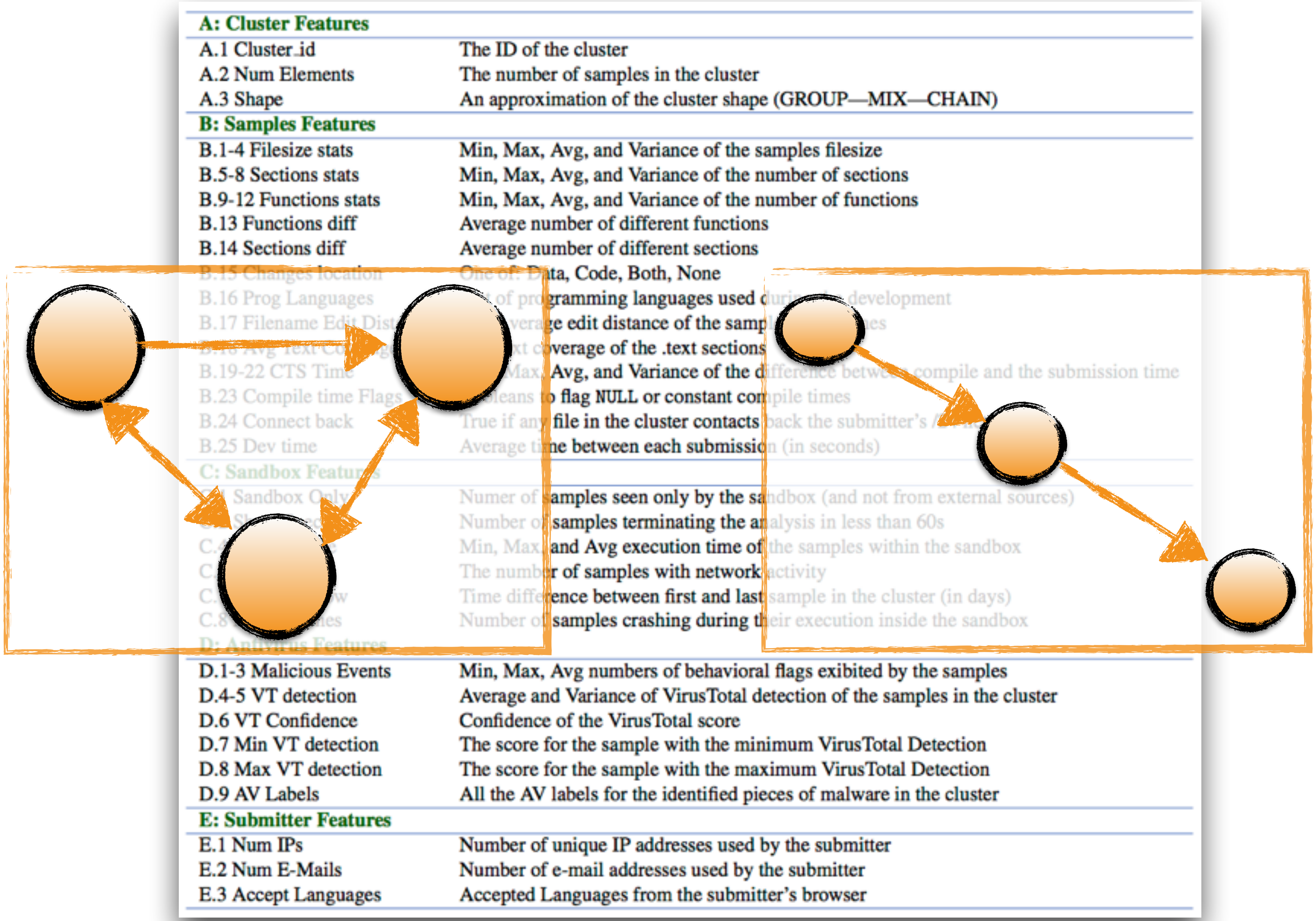
A: File Features	
A.1 Filename	The original name of the file submitted by the user
A.2 File size	The size of the file
A.3 MD5	Simple hash used for lookup in other data sources
A.4 Fuzzy Hashes	Using SSDeep algorithm
B: Timestamps	
B.1 Submission time	Time in which the sample was submitted to Anubis Sandbox
B.2 Compile time	Time in which the binary was compiled
B.3 Symantec first	Time the sample was first observed in the wild by Symantec
B.4 VirusTotal first	Time in which the binary was first submitted to VirusTotal
C: AV Features	
C.1 AV-Detection	Number of AV that flag the samples as malicious (according to VirusTotal)
C.2 AV-Labels	List of AV labels associated to the sample (according to VirusTotal)
D: User-based Features	
D.1 User Agent	User agent of the browser used to submit the sample
D.2 Languages	Languages accepted by the user browser (according to the `accept-language` HTTP header)
D.3 IP	IP address of the user who submitted the file
D.4 IP Geolocation	Geolocation of the user IP address
D.5 Email address	Optional email address specified when the sample was submitted
D.6 Proxy	Boolean value used to identify submission through popular anonymization proxies
E: Binary Features	
E.1 N.Sections	Number of sections in the PE file
E.2 N.Fuctions	Number of functions identified by the disassembly
E.3 Code Coverage	Fraction of `.text` segment covered by the identified functions
E.4 Programming Language	Programming language used to develop the binary
E.5 Metadata	Filenames and username extracted from the PE file
F: Behavioral Features	
F.1 Duration	Duration in seconds of the analysis
F.2 Errors	Error raised during the analysis
F.3 Evasion	Known anti-sandbox techniques detected by the sandbox itself
F.4 Behavior Bitstring	Sequence of 24 boolean flags that characterize the behavior of the sample. (`has_popups`, `has_udp_traffic`, `has_http`, `has_tcp_address_scan`, `modified_registry_keys`, ...)

CLUSTER FEATURES



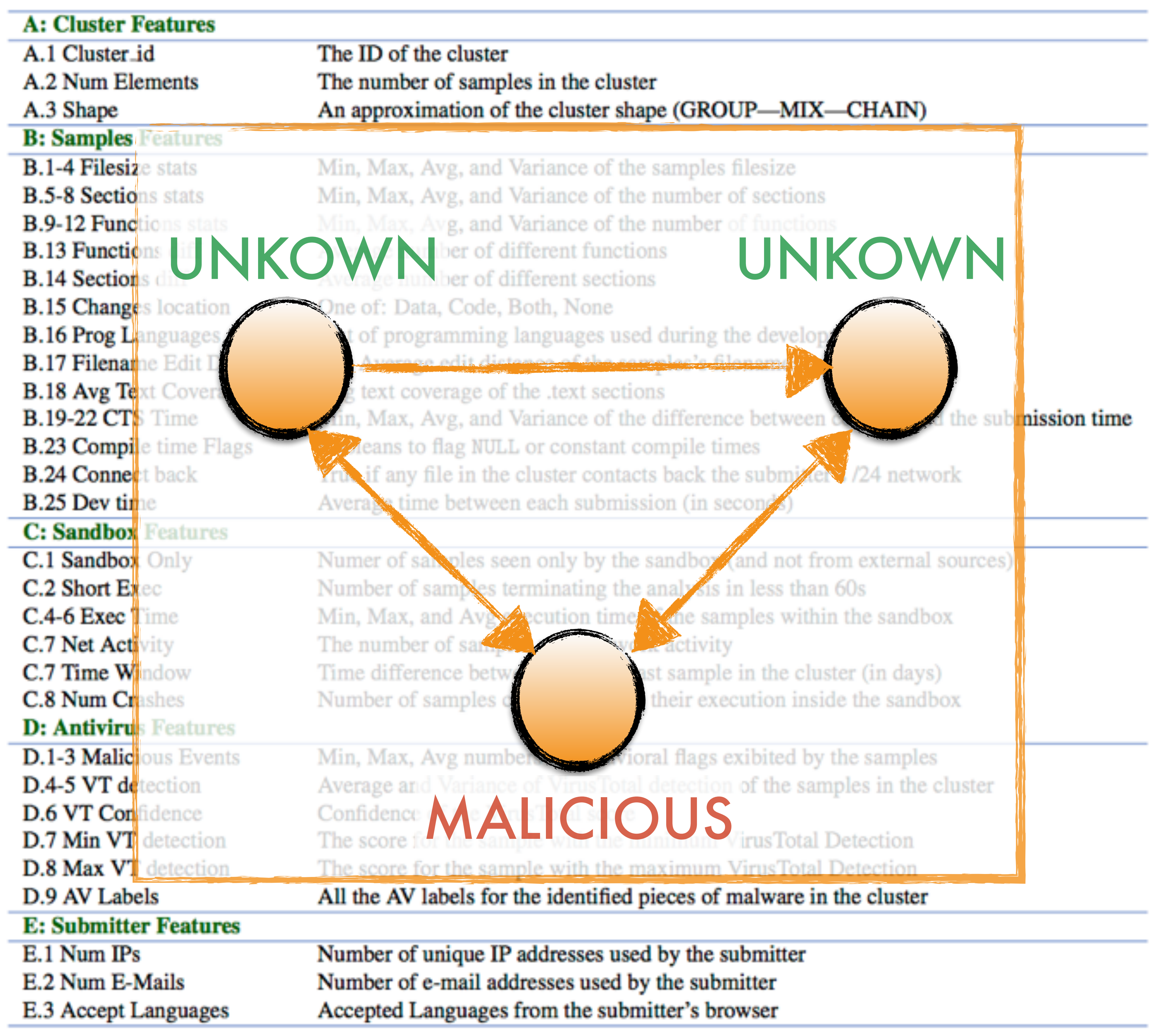
A: Cluster Features	
A.1 Cluster_id	The ID of the cluster
A.2 Num Elements	The number of samples in the cluster
A.3 Shape	An approximation of the cluster shape (GROUP—MIX—CHAIN)
B: Samples Features	
B.1-4 Filesize stats	Min, Max, Avg, and Variance of the samples filesize
B.5-8 Sections stats	Min, Max, Avg, and Variance of the number of sections
B.9-12 Functions stats	Min, Max, Avg, and Variance of the number of functions
B.13 Functions diff	Average number of different functions
B.14 Sections diff	Average number of different sections
B.15 Changes location	One of: Data, Code, Both, None
B.16 Prog Languages	List of programming languages used during the development
B.17 Filename Edit Distance	The Average edit distance of the samples's filenames
B.18 Avg Text Coverage	Avg text coverage of the .text sections
B.19-22 CTS Time	Min, Max, Avg, and Variance of the difference between compile and the submission time
B.23 Compile time Flags	Booleans to flag NULL or constant compile times
B.24 Connect back	True if any file in the cluster contacts back the submitter's /24 network
B.25 Dev time	Average time between each submission (in seconds)
C: Sandbox Features	
C.1 Sandbox Only	Numer of samples seen only by the sandbox (and not from external sources)
C.2 Short Exec	Number of samples terminating the analysis in less than 60s
C.4-6 Exec Time	Min, Max, and Avg execution time of the samples within the sandbox
C.7 Net Activity	The number of samples with network activity
C.7 Time Window	Time difference between first and last sample in the cluster (in days)
C.8 Num Crashes	Number of samples crashing during their execution inside the sandbox
D: Antivirus Features	
D.1-3 Malicious Events	Min, Max, Avg numbers of behavioral flags exibited by the samples
D.4-5 VT detection	Average and Variance of VirusTotal detection of the samples in the cluster
D.6 VT Confidence	Confidence of the VirusTotal score
D.7 Min VT detection	The score for the sample with the minimum VirusTotal Detection
D.8 Max VT detection	The score for the sample with the maximum VirusTotal Detection
D.9 AV Labels	All the AV labels for the identified pieces of malware in the cluster
E: Submitter Features	
E.1 Num IPs	Number of unique IP addresses used by the submitter
E.2 Num E-Mails	Number of e-mail addresses used by the submitter
E.3 Accept Languages	Accepted Languages from the submitter's browser

CLUSTER FEATURES



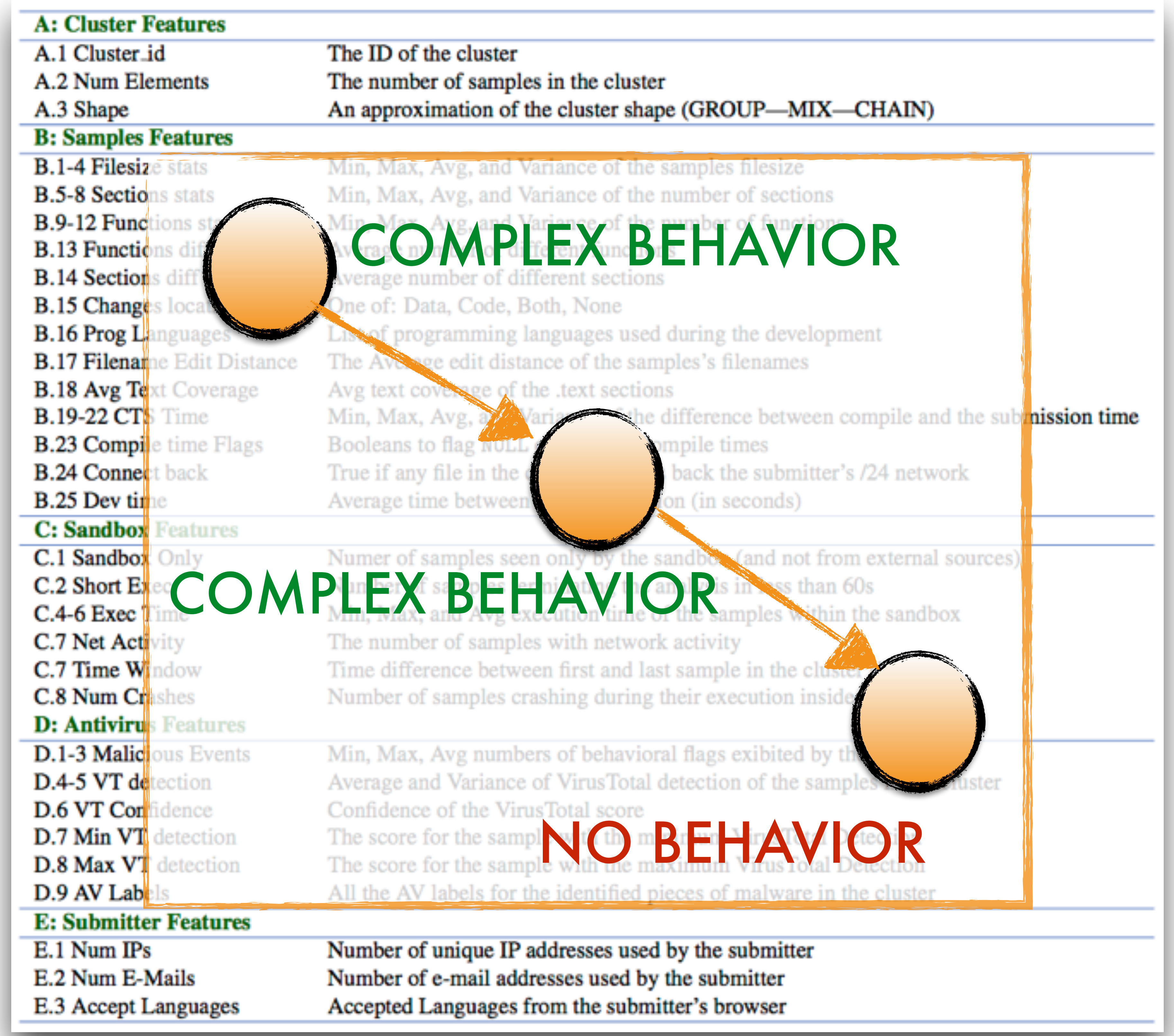
A: Cluster Features	
A.1 Cluster_id	The ID of the cluster
A.2 Num Elements	The number of samples in the cluster
A.3 Shape	An approximation of the cluster shape (GROUP—MIX—CHAIN)
B: Samples Features	
B.1-4 Filesize stats	Min, Max, Avg, and Variance of the samples filesize
B.5-8 Sections stats	Min, Max, Avg, and Variance of the number of sections
B.9-12 Functions stats	Min, Max, Avg, and Variance of the number of functions
B.13 Functions diff	Average number of different functions
B.14 Sections diff	Average number of different sections
B.15 Changes location	One of: Data, Code, Both, None
B.16 Prog Languages	[illegible] of programming languages used d[illegible] development
B.17 Filename Edit Dist	[illegible]verage edit distance of the sampl[illegible]es
B.18 Avg Text Co[illegible]ge	[illegible]xt coverage of the .text sections
B.19-22 CTS Time	[illegible] Max, Avg, and Variance of the difference betw[illegible] compile and the submission time
B.23 Compile time Flags	[illegible]eans to flag NULL or constant compile times
B.24 Connect back	True if any file in the cluster contacts back the submitter's /[illegible]
B.25 Dev time	Average time between each submission (in seconds)
C: Sandbox Features	
[illegible] Sandbox Only	Numer of samples seen only by the sandbox (and not from external sources)
[illegible]	Number of samples terminating the analysis in less than 60s
C.[illegible]	Min, Max, and Avg execution time of the samples within the sandbox
C[illegible]	The number of samples with network activity
C.[illegible]w	Time difference between first and last sample in the cluster (in days)
C.8[illegible]es	Number of samples crashing during their execution inside the sandbox
D: Antivirus Features	
D.1-3 Malicious Events	Min, Max, Avg numbers of behavioral flags exibited by the samples
D.4-5 VT detection	Average and Variance of VirusTotal detection of the samples in the cluster
D.6 VT Confidence	Confidence of the VirusTotal score
D.7 Min VT detection	The score for the sample with the minimum VirusTotal Detection
D.8 Max VT detection	The score for the sample with the maximum VirusTotal Detection
D.9 AV Labels	All the AV labels for the identified pieces of malware in the cluster
E: Submitter Features	
E.1 Num IPs	Number of unique IP addresses used by the submitter
E.2 Num E-Mails	Number of e-mail addresses used by the submitter
E.3 Accept Languages	Accepted Languages from the submitter's browser

CLUSTER FEATURES



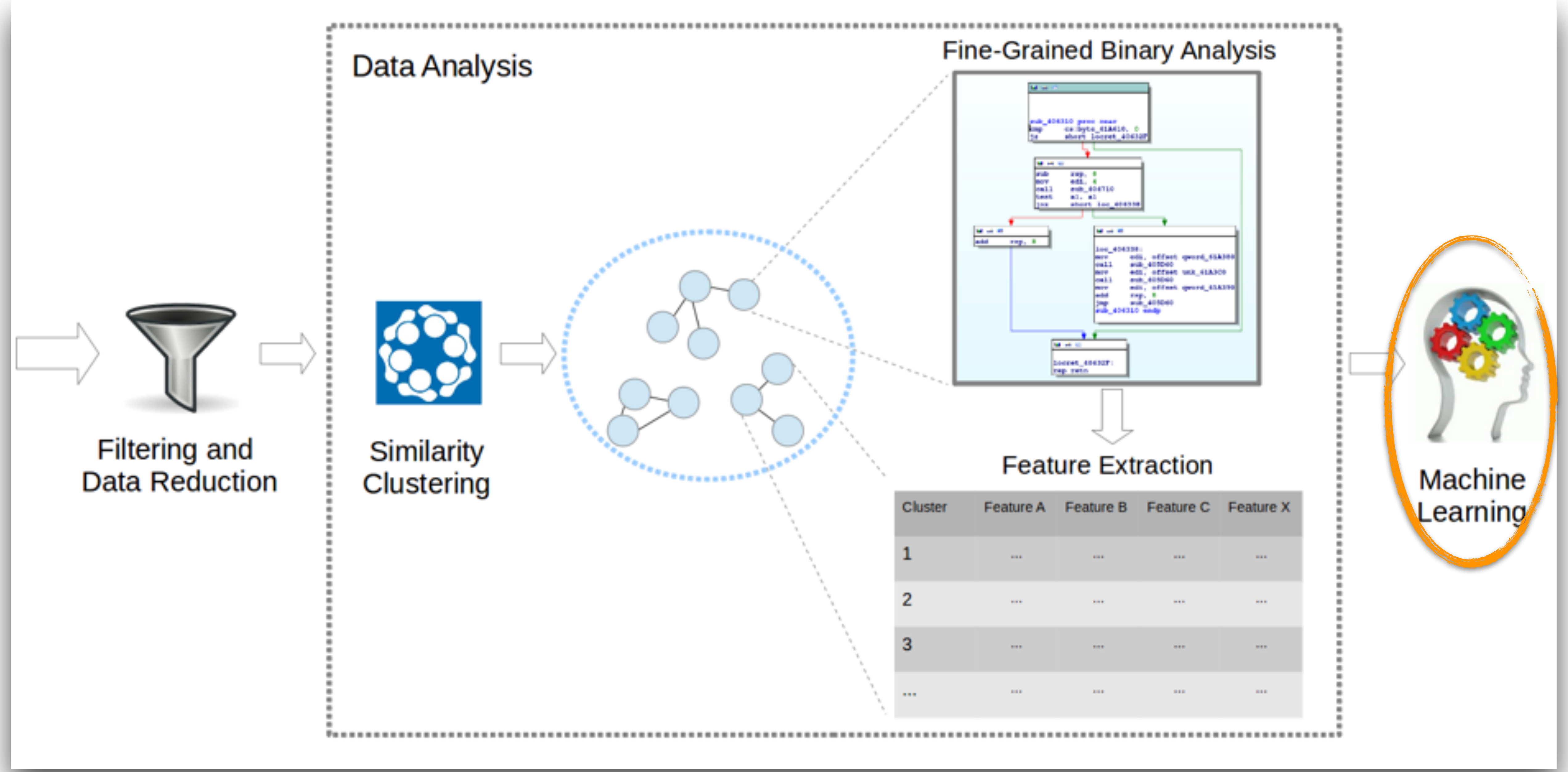
A: Cluster Features	
A.1 Cluster_id	The ID of the cluster
A.2 Num Elements	The number of samples in the cluster
A.3 Shape	An approximation of the cluster shape (GROUP—MIX—CHAIN)
B: Samples Features	
B.1-4 Filesize stats	Min, Max, Avg, and Variance of the samples filesize
B.5-8 Sections stats	Min, Max, Avg, and Variance of the number of sections
B.9-12 Functions stats	Min, Max, Avg, and Variance of the number of functions
B.13 Functions [illegible]	[illegible]ber of different functions
B.14 Sections di[illegible]	Average number of different sections
B.15 Changes location	One of: Data, Code, Both, None
B.16 Prog Languages	[illegible]t of programming languages used during the develop[illegible]
B.17 Filename Edit D[illegible]	Average edit distance of the samples's filename[illegible]
B.18 Avg Text Cover[illegible]	[illegible]g text coverage of the .text sections
B.19-22 CTS Time	[illegible]in, Max, Avg, and Variance of the difference between [illegible] the submission time
B.23 Compile time Flags	[illegible]eans to flag NULL or constant compile times
B.24 Connect back	True if any file in the cluster contacts back the submitter[illegible] /24 network
B.25 Dev time	Average time between each submission (in seconds)
C: Sandbox Features	
C.1 Sandbox Only	Numer of samples seen only by the sandbox (and not from external sources)
C.2 Short Exec	Number of samples terminating the analysis in less than 60s
C.4-6 Exec Time	Min, Max, and Avg execution time of the samples within the sandbox
C.7 Net Activity	The number of sa[illegible] activity
C.7 Time Window	Time difference betw[illegible]st sample in the cluster (in days)
C.8 Num Crashes	Number of samples [illegible] their execution inside the sandbox
D: Antivirus Features	
D.1-3 Malicious Events	Min, Max, Avg numbe[illegible]vioral flags exibited by the samples
D.4-5 VT detection	Average and Variance of VirusTotal detection of the samples in the cluster
D.6 VT Confidence	Confidence [illegible]
D.7 Min VT detection	The score for the sample with the minimum VirusTotal Detection
D.8 Max VT detection	The score for the sample with the maximum VirusTotal Detection
D.9 AV Labels	All the AV labels for the identified pieces of malware in the cluster
E: Submitter Features	
E.1 Num IPs	Number of unique IP addresses used by the submitter
E.2 Num E-Mails	Number of e-mail addresses used by the submitter
E.3 Accept Languages	Accepted Languages from the submitter's browser

CLUSTER FEATURES



A: Cluster Features	
A.1 Cluster_id	The ID of the cluster
A.2 Num Elements	The number of samples in the cluster
A.3 Shape	An approximation of the cluster shape (GROUP—MIX—CHAIN)
B: Samples Features	
B.1-4 Filesize stats	Min, Max, Avg, and Variance of the samples filesize
B.5-8 Sections stats	Min, Max, Avg, and Variance of the number of sections
B.9-12 Functions st[illegible]	Min, Max, Avg, and Variance of the number of functions
B.13 Functions di[illegible]	Average number of different functions
B.14 Sections diff[illegible]	Average number of different sections
B.15 Changes loca[illegible]	One of: Data, Code, Both, None
B.16 Prog Languages	List of programming languages used during the development
B.17 Filename Edit Distance	The Average edit distance of the samples's filenames
B.18 Avg Text Coverage	Avg text coverage of the .text sections
B.19-22 CTS Time	Min, Max, Avg, and Variance of the difference between compile and the submission time
B.23 Compile time Flags	Booleans to flag NULL [illegible] compile times
B.24 Connect back	True if any file in the [illegible] back the submitter's /24 network
B.25 Dev time	Average time between [illegible] on (in seconds)
C: Sandbox Features	
C.1 Sandbox Only	Numer of samples seen only by the sandbox (and not from external sources)
C.2 Short Exec[illegible]	Number of samples terminating the analysis in less than 60s
C.4-6 Exec Time	Min, Max, and Avg execution time of the samples within the sandbox
C.7 Net Activity	The number of samples with network activity
C.7 Time Window	Time difference between first and last sample in the cluster
C.8 Num Crashes	Number of samples crashing during their execution inside [illegible]
D: Antivirus Features	
D.1-3 Malicious Events	Min, Max, Avg numbers of behavioral flags exibited by th[illegible]
D.4-5 VT detection	Average and Variance of VirusTotal detection of the samples [illegible] cluster
D.6 VT Confidence	Confidence of the VirusTotal score
D.7 Min VT detection	The score for the sample with the minimum VirusTotal Detection
D.8 Max VT detection	The score for the sample with the maximum VirusTotal Detection
D.9 AV Labels	All the AV labels for the identified pieces of malware in the cluster
E: Submitter Features	
E.1 Num IPs	Number of unique IP addresses used by the submitter
E.2 Num E-Mails	Number of e-mail addresses used by the submitter
E.3 Accept Languages	Accepted Languages from the submitter's browser

SYSTEM OVERVIEW



```
sub_406310 proc near
cmp     cs:byte_61A610, 0
jz      short locret_40632F

sub     rsp, 8
mov     edi, 4
call    sub_404710
test    al, al
jnz     short loc_406338

add     rsp, 8

loc_406338:
mov     edi, offset qword_61A380
call    sub_405D40
mov     edi, offset unk_61A3C0
call    sub_405D40
mov     edi, offset qword_61A390
add     rsp, 8
jmp     sub_405D40
sub_406310 endp

locret_40632F:
rep retn
```

Cluster	Feature A	Feature B	Feature C	Feature X
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

MACHINE LEARNING

- Logistic Model Tree (LMT)
- Training Set (157 clusters):
  - Non development: 91 clusters
  - Development: 66 clusters

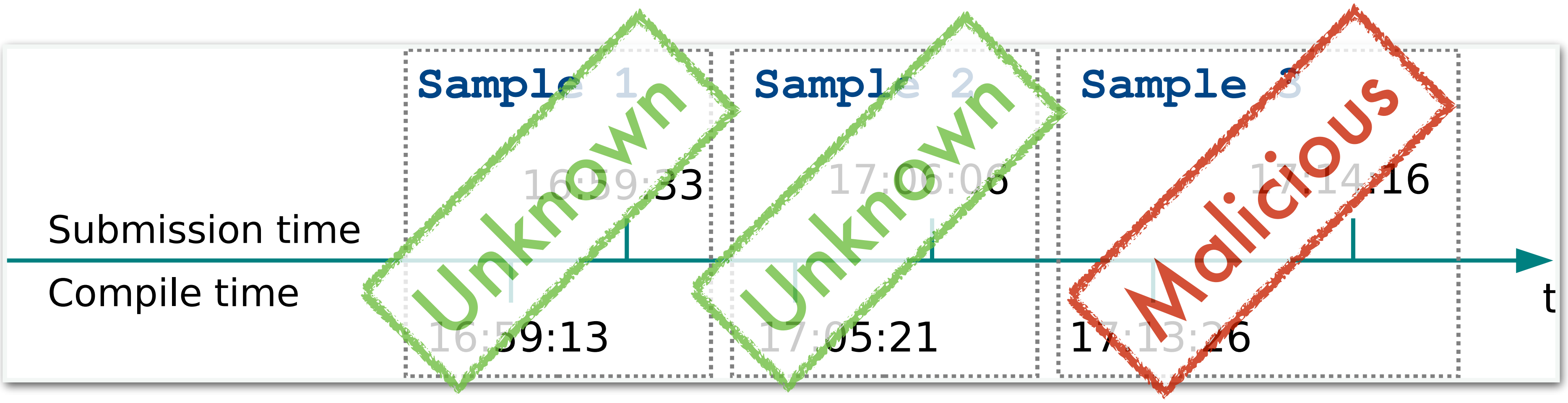
RESULTS

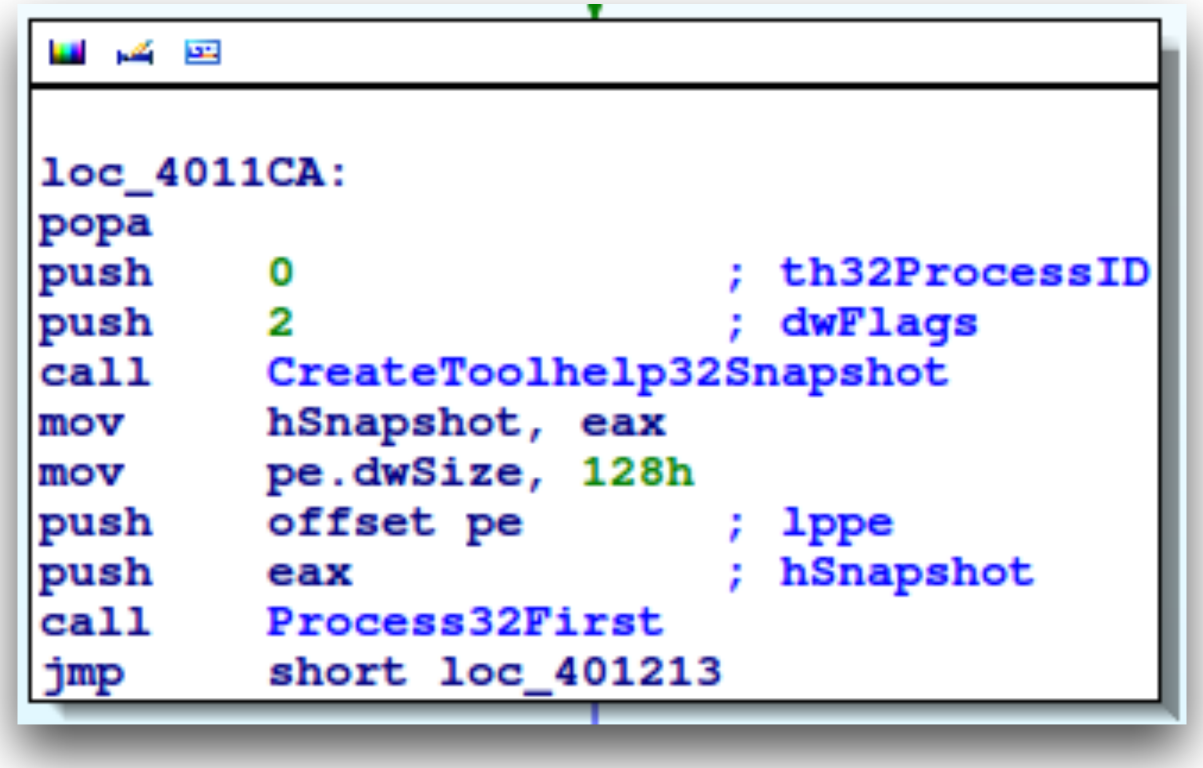
- 3038 potential development clusters
- 1474 malicious clusters
- 135 days on average for the detection
- Thousands of computers infected in 13 countries

CLUSTERS	TYPE
1082	Trojans
83	Backdoors
65	Worms
45	Botnets
21	Tools
4	Keyloggers

EXAMPLES

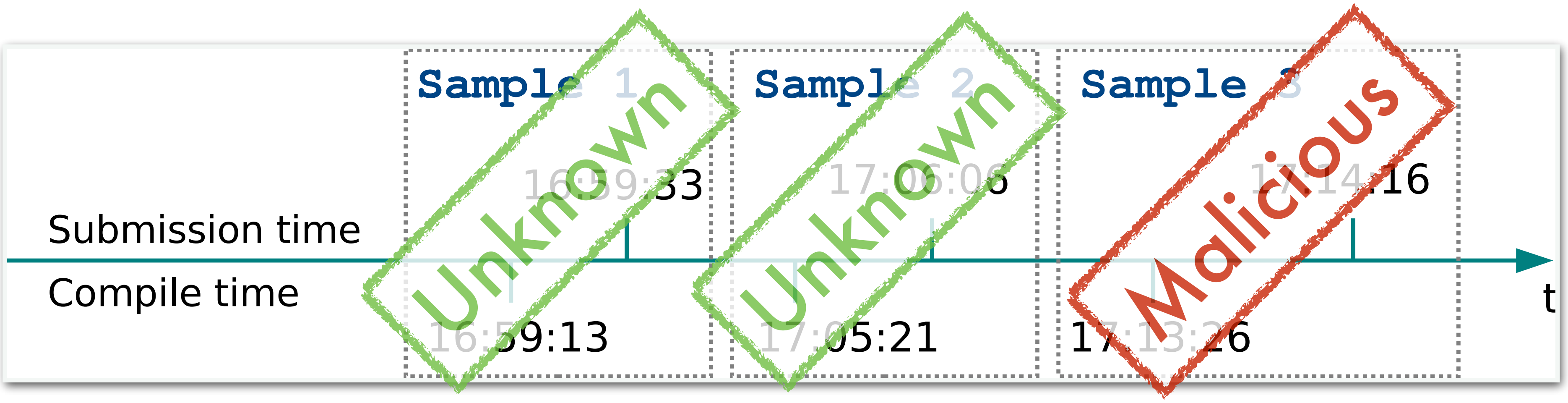
ANTI-SANDBOX

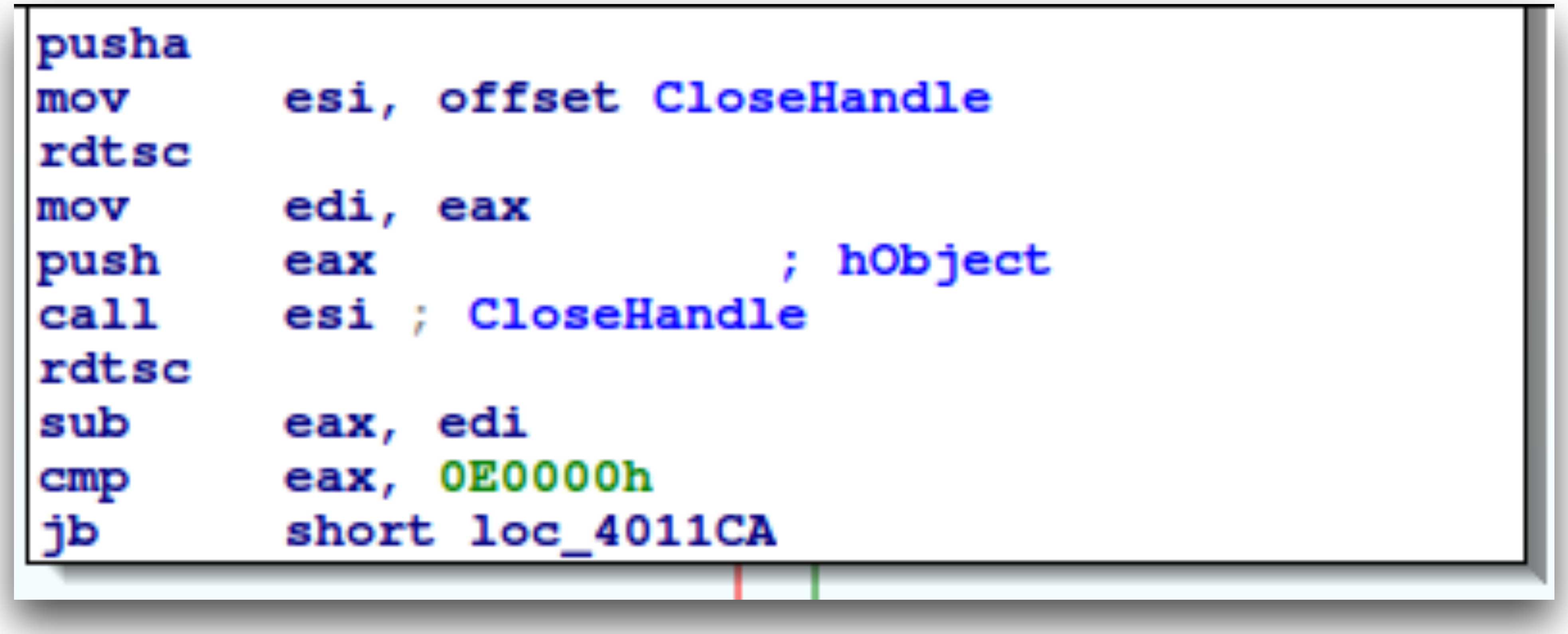




```
loc_4011CA:
popa
push    0                   ; th32ProcessID
push    2                   ; dwFlags
call    CreateToolhelp32Snapshot
mov     hSnapshot, eax
mov     pe.dwSize, 128h
push    offset pe           ; lppe
push    eax                 ; hSnapshot
call    Process32First
jmp     short loc_401213
```

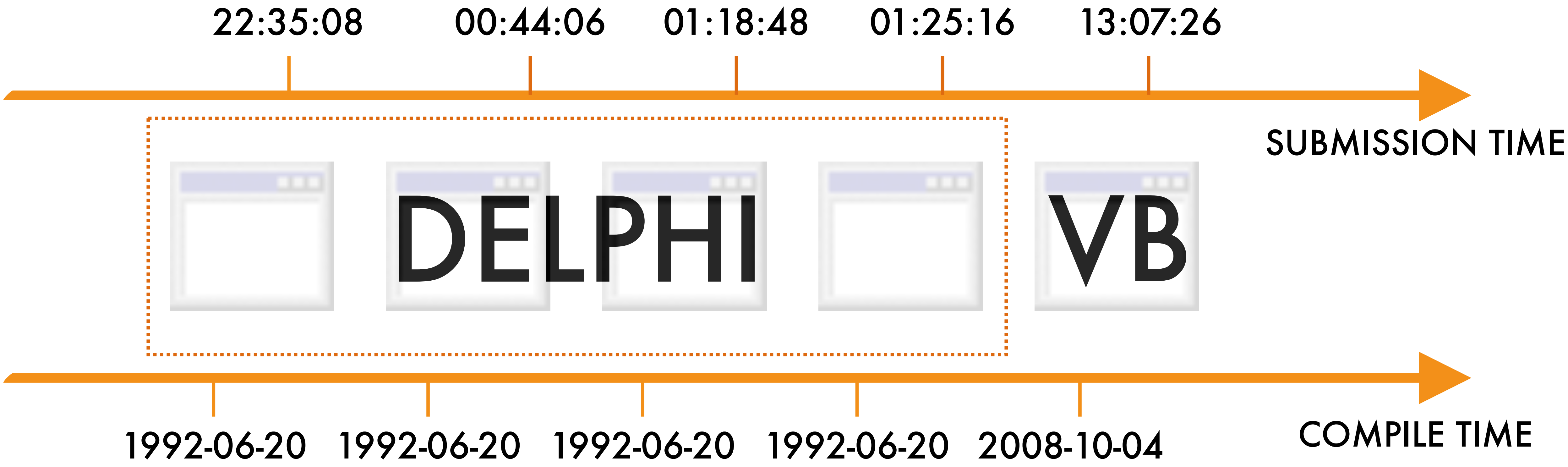
ANTI-SANDBOX



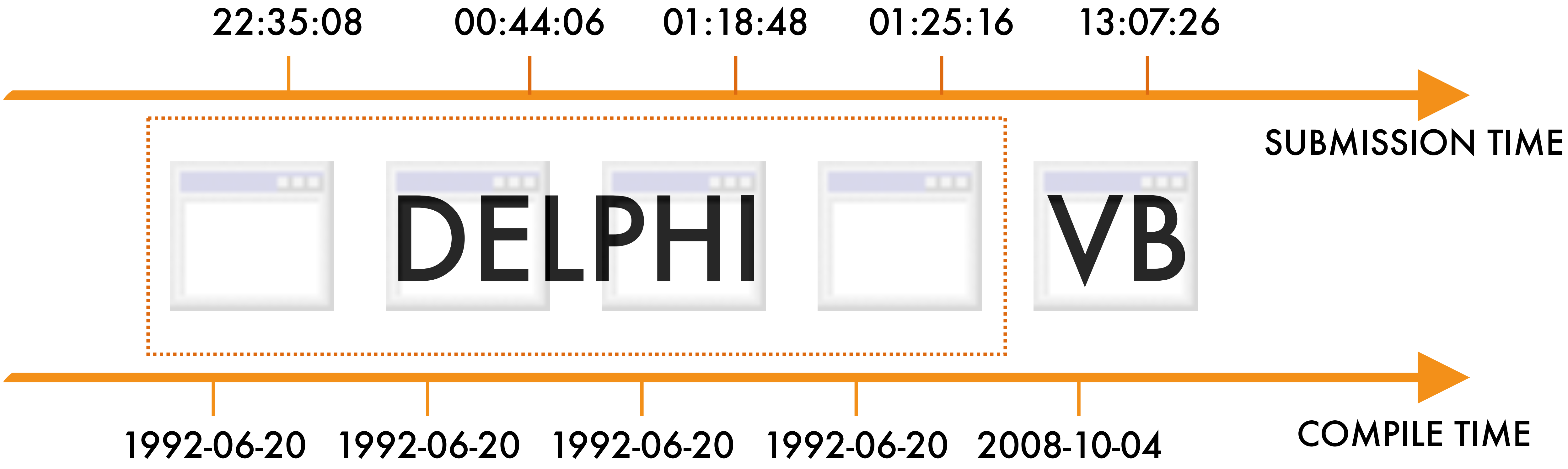


```
pusha
mov     esi, offset CloseHandle
rdtsc
mov     edi, eax
push    eax                     ; hObject
call    esi ; CloseHandle
rdtsc
sub     eax, edi
cmp     eax, 0E0000h
jb      short loc_4011CA
```

TROJAN DROPPER



TROJAN DROPPER

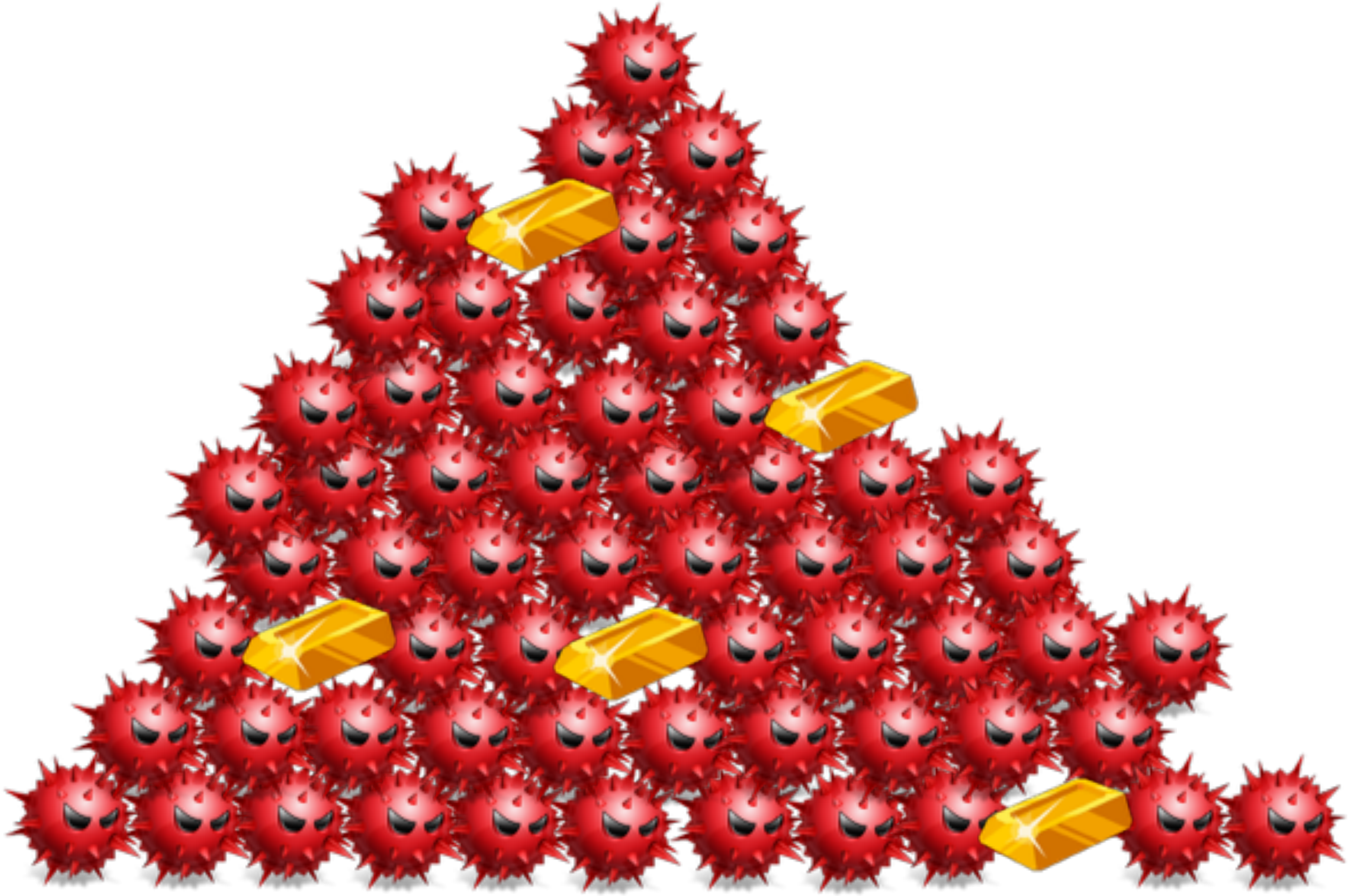


- VirusTotal: 37/50 (*trojan dropper*)
- Two IP addresses:
    - Dynamic DNS service (no-ip)
    - Connect-back behavior → overall 1817 clusters

LIMITATIONS

- No packed binaries
- Evasions:
  - Sandbox interaction still required to develop evasion techniques
  - Most sophisticated analysis techniques require to link a probe to the final malware

CONCLUSION



THE END

THANK YOU

graziano@eurecom.fr

magrazia@cisco.com

@emd3l