Institut
Mines-Telecom

FORTINET®

Pre-filtering Mobile Malware with Heuristic Techniques

Ludovic Apvrille
ludovic.apvrille@telecom-paristech.fr

Axelle Apvrille
aapvrille@fortinet.com

GreHack 2013

GreHack

Outline

TELECOM ParisTech

Outline

The Big Picture on Android Malware

A few figures

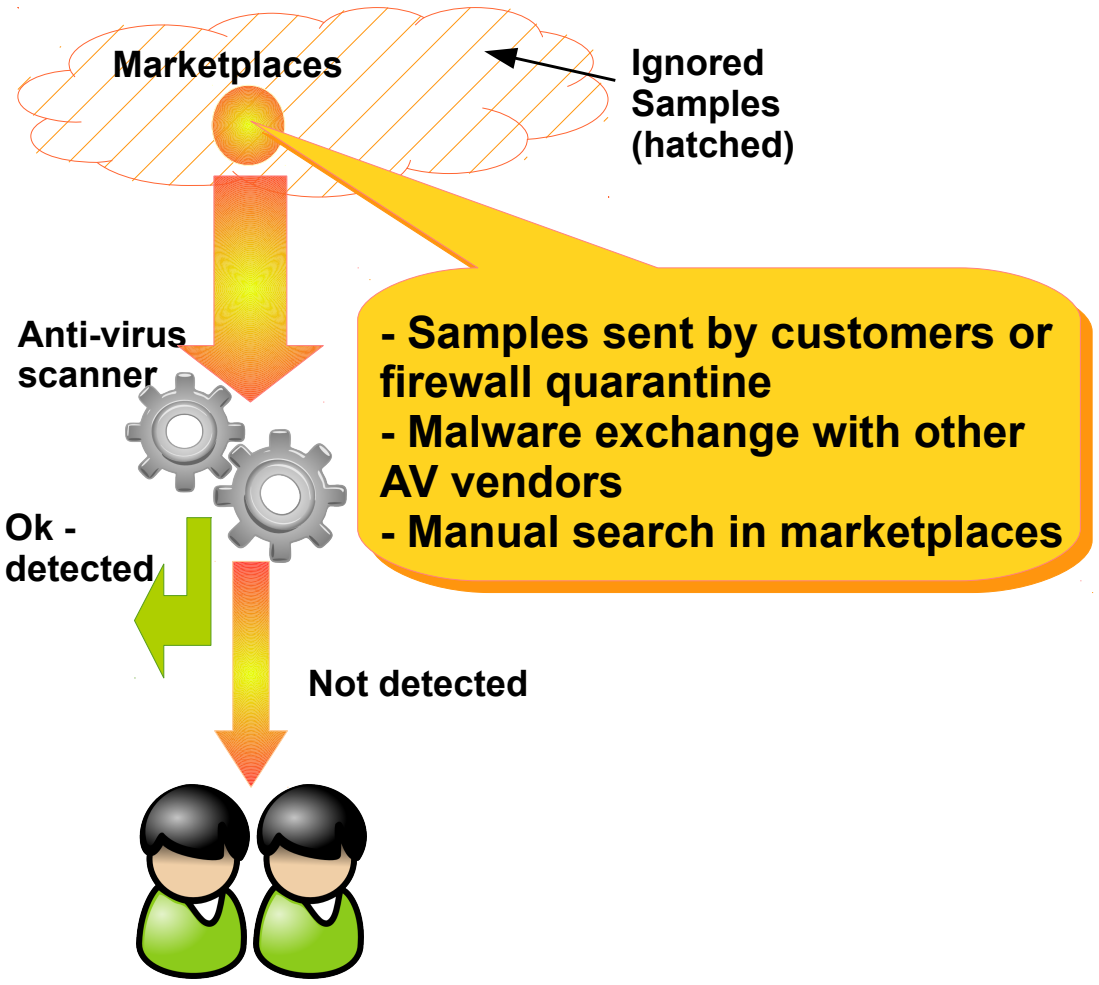
- June 2013: Over 200,000 malicious Android samples
- Sept. 2013: Over 300,000 ...
- End of Oct. 2013: Over 350,000!
- 1,000 new samples are reported every single day[a]

[a] see http://blog.fortinet.com/1-000-malicious-Android-samples-per-day

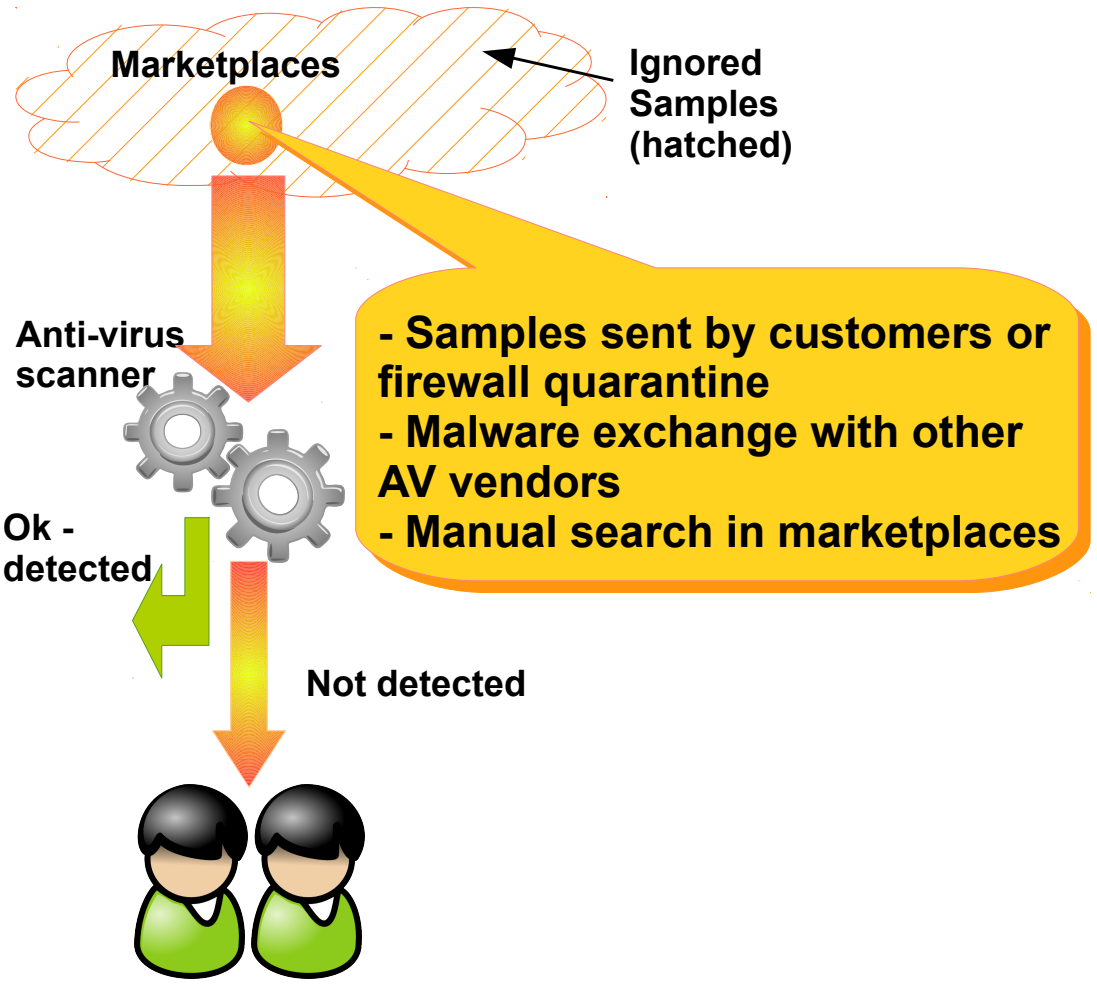
Many malware remain undetected for a long time!

(Maybe you are currently using one on your mobile phone instead of listening to me?)

Are AV Analysts Lazy? No, Too Much Work!



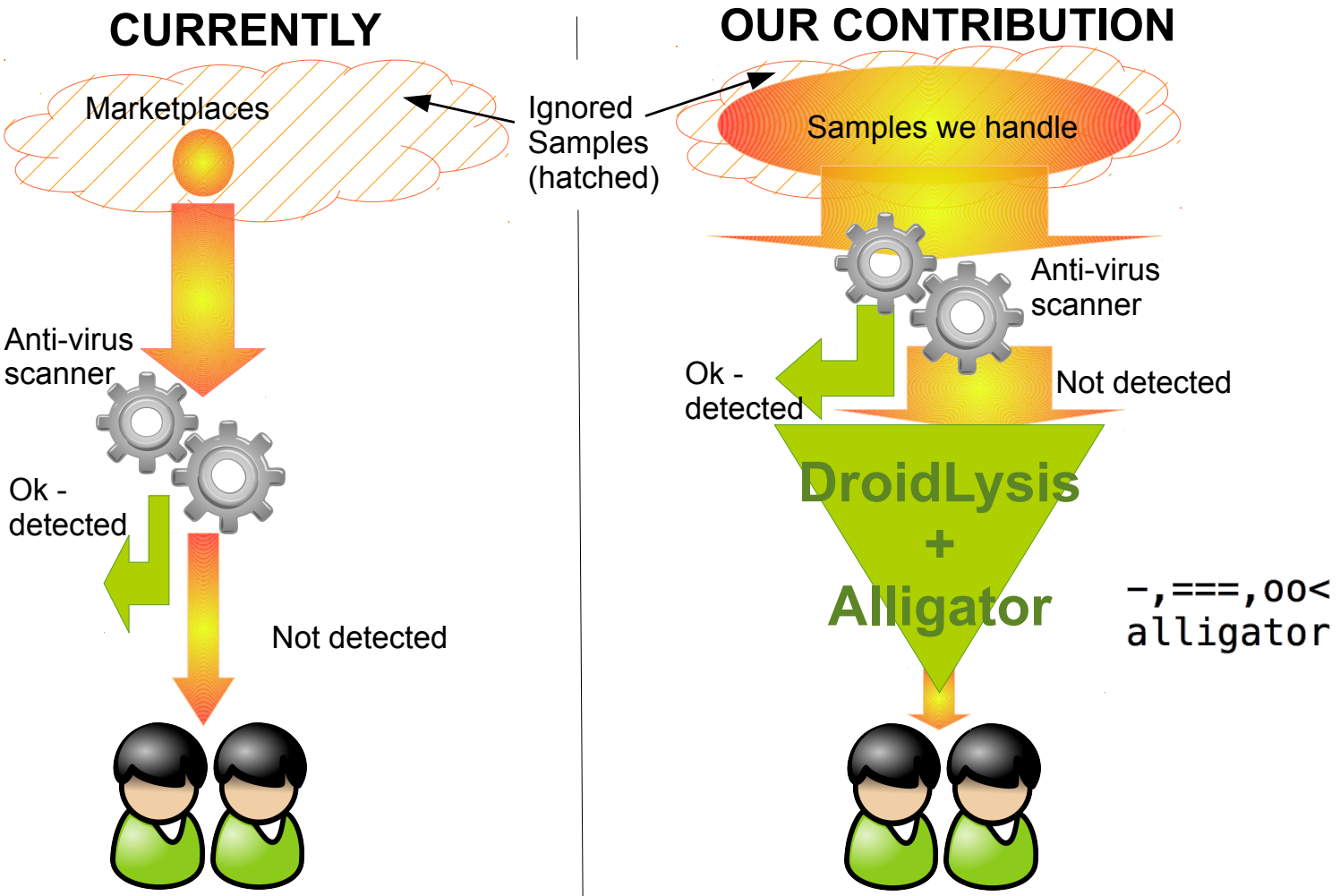
Are AV Analysts Lazy? No, Too Much Work!



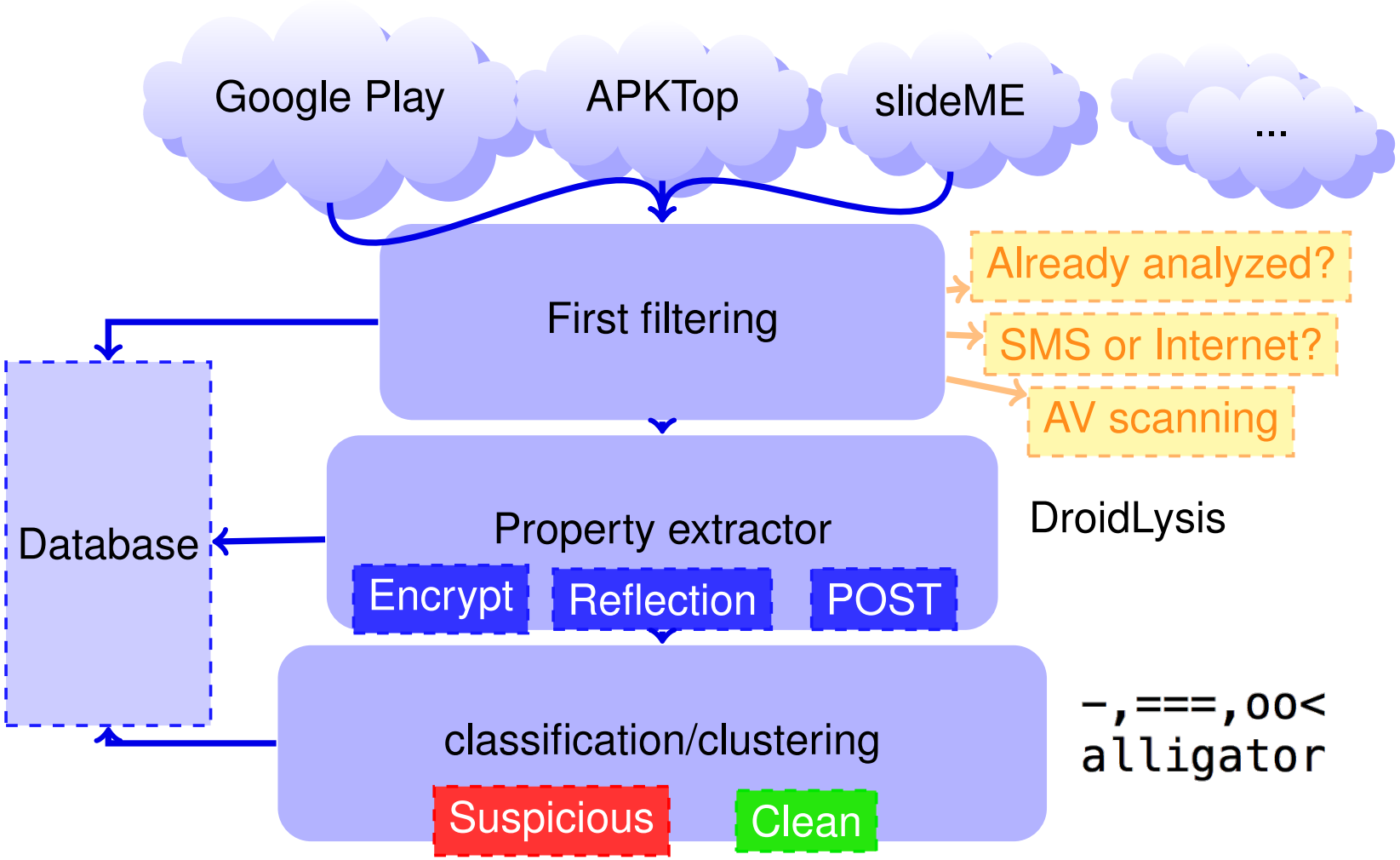
Conclusion:

Smart filtering is necessary!

Prefiltering: Overview

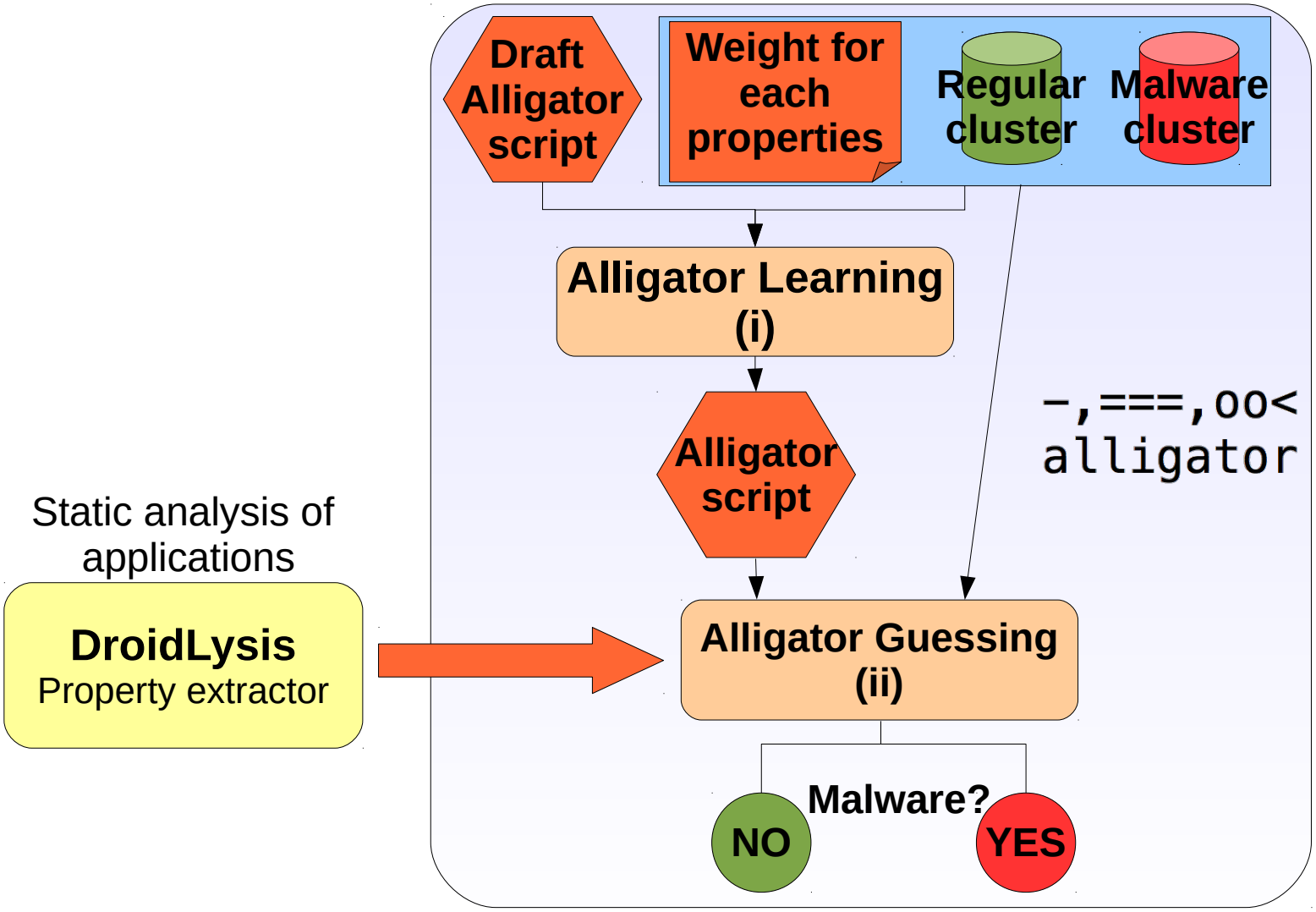


SherlockDroid Architecture



Outline

Fundamentals of Alligator



Yet Another Clustering Toolkit?

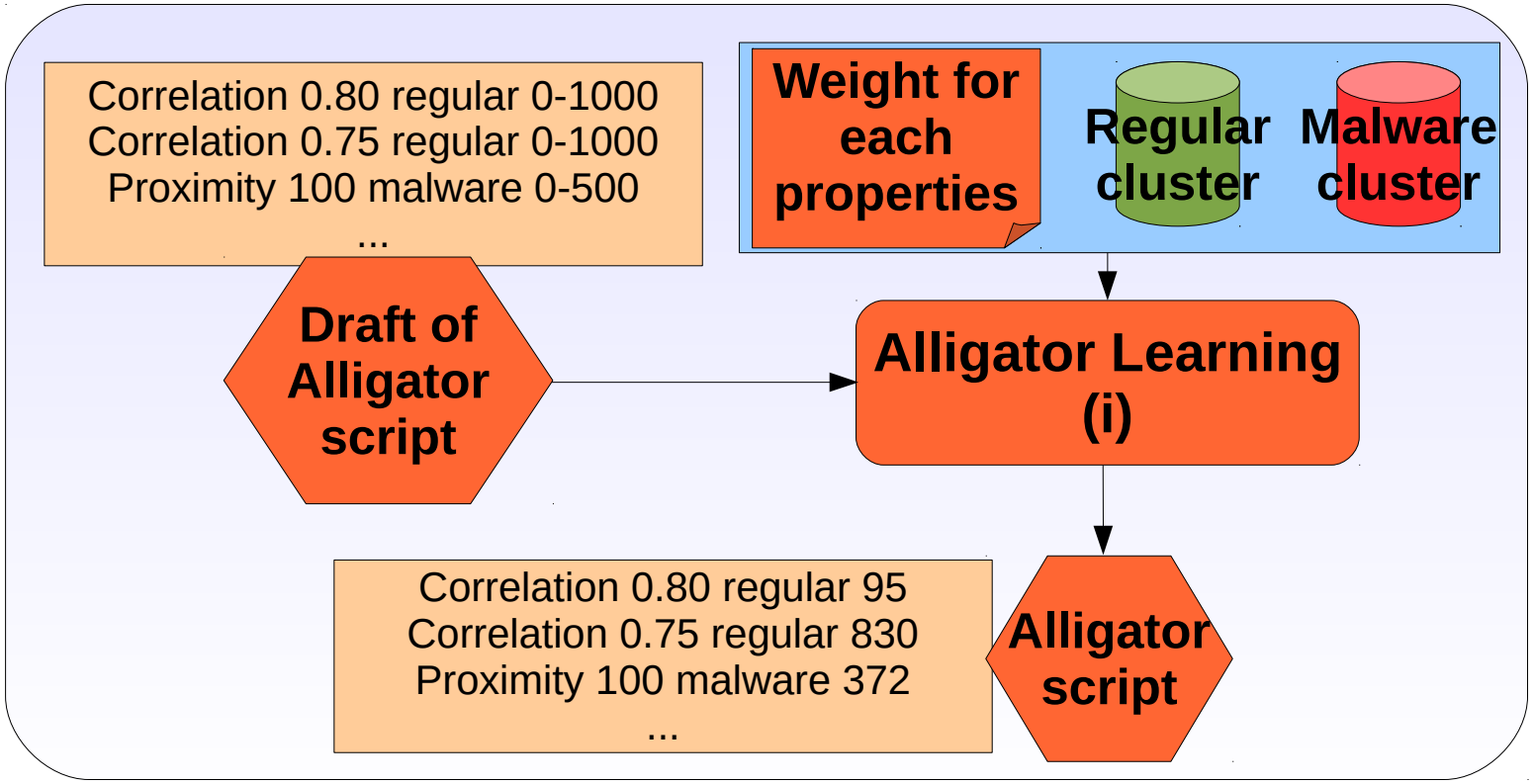
No! Alligator is much better!!!

- Dedicated to **work with two pre-known clusters**
- **Handles several up-to-date clustering algorithms at the same time**
  - Automatically determines how to combine them in an optimal way
- Option to settle a preference in **reducing false positive or negative**
- Very efficient - because we are very good programmers ;-)
- Free software
  - "Free": As in "free beer" AND as in "freedom"

Principle of Learning

Purpose

- Determining the importance to give to each couple (**clustering algorithm**, **parameter**)



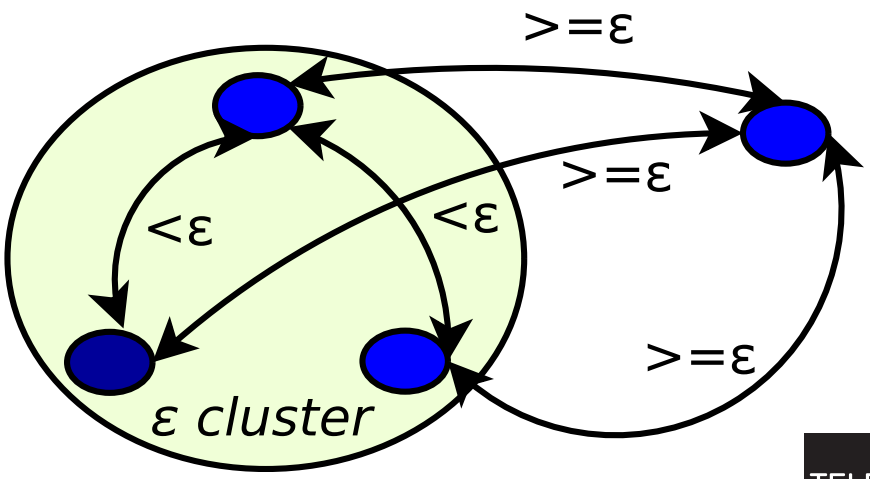
Clustering Algorithms

Cluster-center oriented algorithms

1. Standard deviation
2. Correlation
3. Probability difference
4. Probability factor

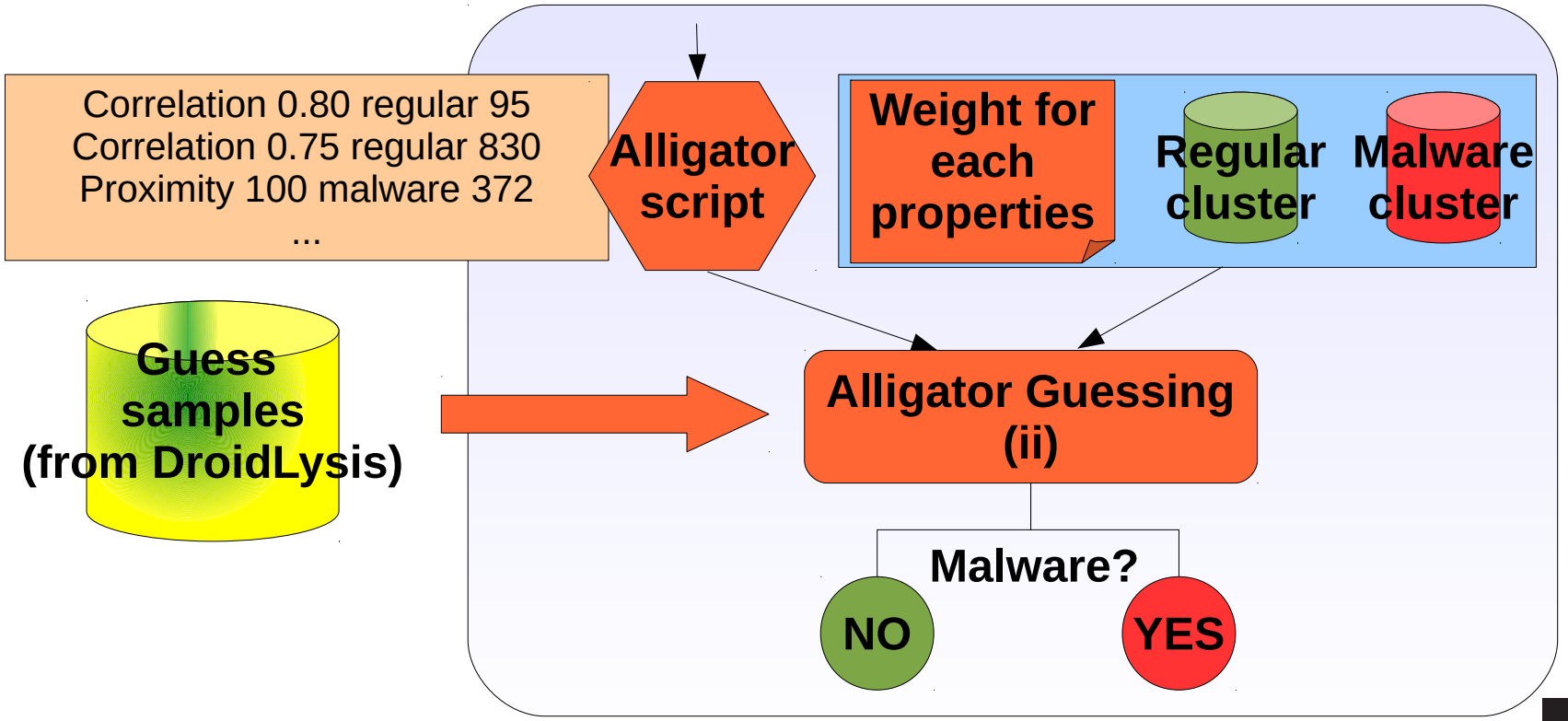
Neighbourhood oriented algorithms

5. Proximity (a.k.a. k-NN)
6. Proximity with limited properties
7. Epsilon clusters



Guessing Stage

Determining the cluster (regular, malware) of **unknown samples**



GreHack

Outline

TELECOM ParisTech

GreHack

Test Bench

Type of cluster	Malware samples	Regular samples	Period
Learning clusters	82,985	8,299	Before June 14
Guess clusters	19,171	1,103	From June 15 to June 24
Total of samples tested	102,156	9,402	

Number of samples in our test clusters

GreHack

Test Bench (Learning Stage)

- All clustering algorithms considered with an average of 5 parameters for each
- Example:
  - Correlations: 0.80, 0.75, 0.70, 0.60
  - Epsilon clusters: ϵ-path of 10^{-5} to $10^{-}1$
- Computation time: around 10 hours on a non dedicated host

GreHack

Results of Learning and Guessing Stages

Alligator was tested over those new sets of malware and clean files (20k new samples)

		Regular	Malware
Learning	Number of failed / recognized	9 / 8,290	67 / 82,918
	Failure / success rates in %	0.11% 99.89%	0.08% 99.92%
Guessing	Number of failed / recognized	2 / 1,101	375 / 18,796
	Failure / success rates in %	0.18% 99.81%	1.96% 98.04%

TELECOM ParisTech

Conclusions

SherlockDroid is efficient!

- SherlockDroid = efficient combination of market crawler + property extractor + clustering
- Large sets of clusters tested
- Objective reached: $\rightarrow$ 99.8% of clean applications are filtered out.
  - AV analysts can now be lazy ;-)
- Unknown malware discovered thanks to Alligator[a]

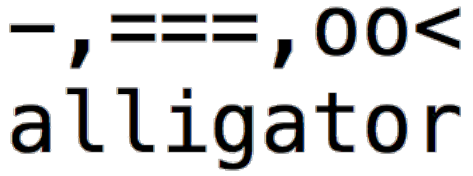
[a] see http://blog.fortinet.com/Alligator-detects-GPS-leaking-adware/.

GreHack

Conclusions (Cont.)

Limitations and Future work

- Clean cluster much smaller than malware cluster!
- More clustering algorithms
- Alligator could be used for many other purposes



TELECOM ParisTech

GreHack

Do Try Alligator!

```
-,===,oo<
alligator
```

perso.telecom-paristech.fr/~apvrille/alligator.html

(Are you sure your qr-code reader application is not a malware???)

BTW: All French persons are not named "Apvrille": we are husband and wife ;-)

TELECOM ParisTech