symantec.

Confidence in a connected world.

WOMBAT

# On the difficulty of training a WOMBAT to identify false positives/negatives among antivirus alerts

Marc Dacier

Director, Symantec Research Labs Europe

ZISC, Zürich, Switzerland, January 15, 2009

# Foreword

- The work presented here is the result of collaborative work carried out in the context of the WOMBAT project involving all the partners.

- More specifically, the second part of the talk is based on a joint publication coauthored with Julio Canto (Hispasec Sistemas), Engin Kirda (Eurecom), Corrado Leita and myself (Symantec Research Labs Europe):

  - "Large scale malware collection: lessons learned", J. Canto, M. Dacier, E. Kirda, and C. Leita, IEEE SRDS, *Workshop on Sharing Field Data and Experiment Measurements on Resilience of Distributed Computing Systems*, Naples, Italy, October 5th, 2008, available online at www.amber-project.eu/srds-ws/papers/01_Canto_Dacier_Kirda_Leita.pdf

# Problem statement

- Generation of malware collections for AV benchmarking
- Our experience with a large malware collection underlined some important challenges
  - In the generation of sample set representative of the Internet malware scenario
  - In the definition of a “false positive”
  - In the definition of a “false negative”

# IST-216026-WOMBAT: Facts sheet

- **WOMBAT**:
  - **W**orldwide **O**bservatory of **M**alicious **B**ehaviors and **A**ttack **T**hreats
- **Duration:** 36 months (starting date: 01/2008)
- **Total cost**: 4 422 746 €
- **EC Contribution**: 2 890 796 €
- **Coordinator**:  Orange FT Group, Dr. Hervé Debar
- **Web site**:  www.wombat-project.eu

# The WOMBAT Consortium

EURECOM
Sophia Antipolis

orange™

TECHNISCHE UNIVERSITÄT WIEN
VIENNA UNIVERSITY OF TECHNOLOGY
TU WIEN

symantec.

HISPASEC SISTEMAS

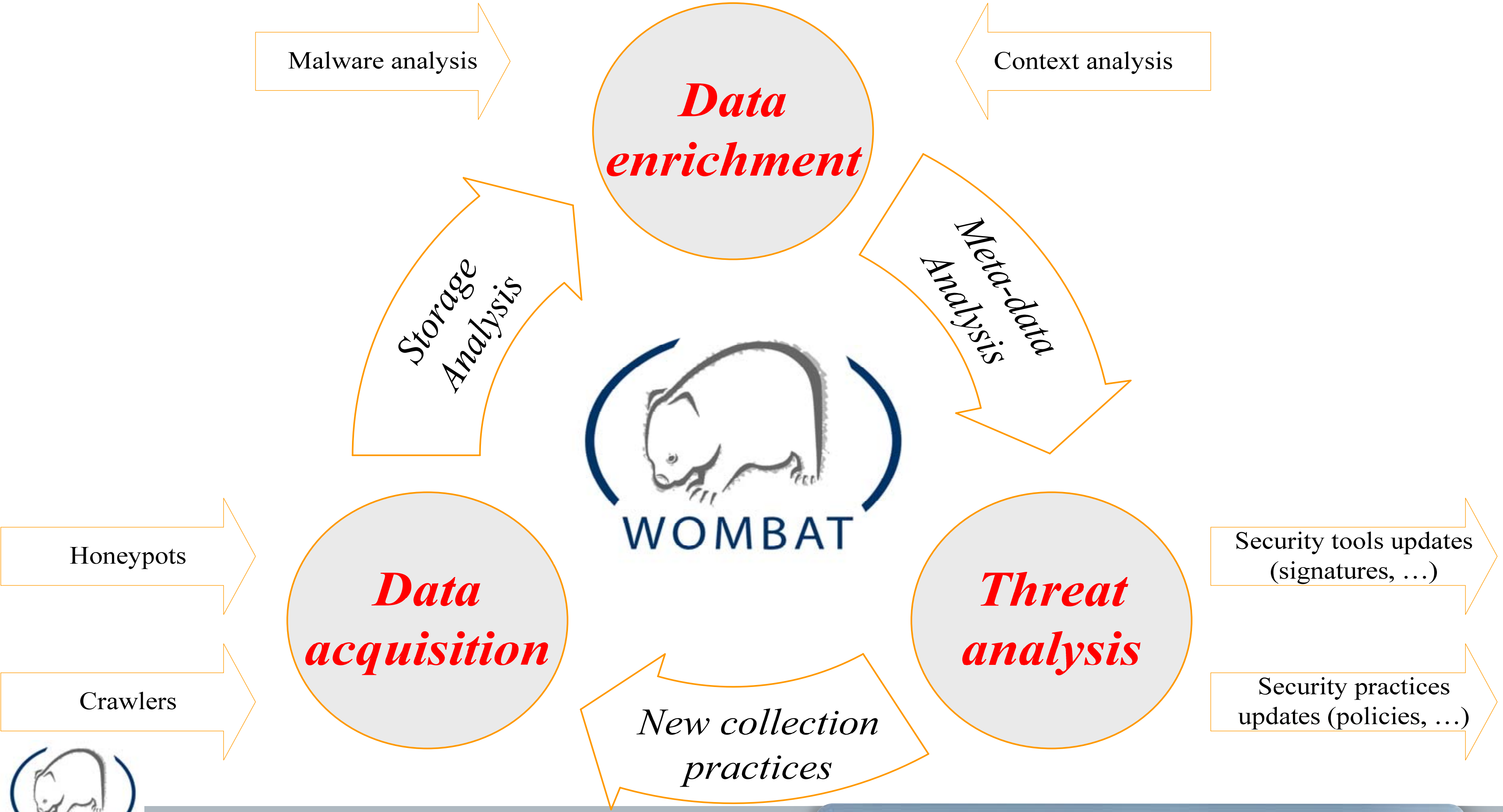
FORTH

POLITECNICO DI MILANO

vrije Universiteit

NASK

# Main objectives and principles



Malware analysis

Context analysis

**Data enrichment**

Storage Analysis

Meta-data Analysis

WOMBAT

Honeypots

Crawlers

**Data acquisition**

*New collection practices*

**Threat analysis**

Security tools updates (signatures, …)

Security practices updates (policies, …)

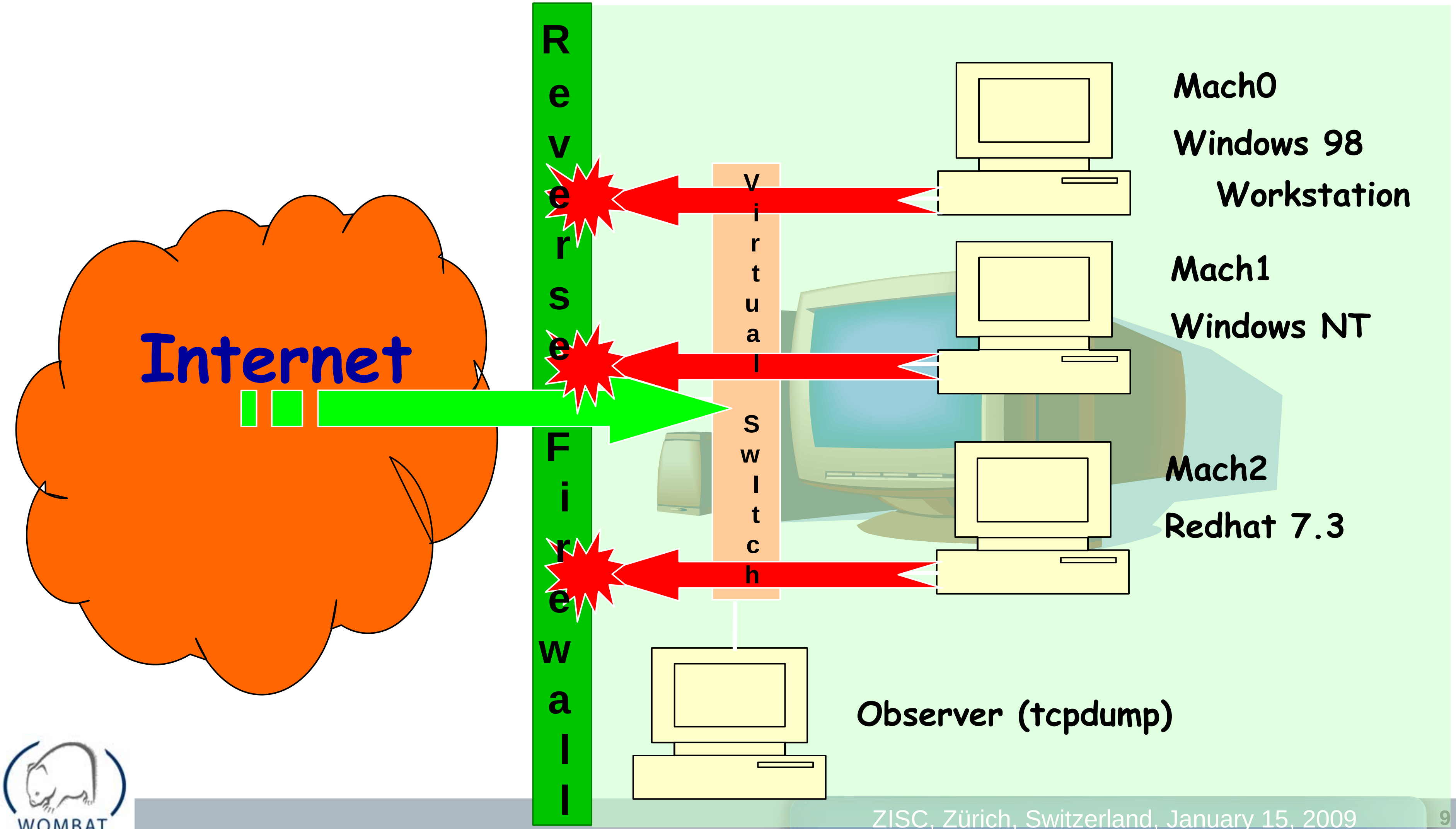
# Project results and innovation

- New data gathering tools
  - Advanced features (high interaction, real-time analysis)
  - New targets (wireless, bluetooth, RFID, …)
- Tools and techniques for characterization of malware
  - Malware-based analysis  AND Contextual analysis
- Framework and tools for qualitative threat analysis
  - Early warning systems

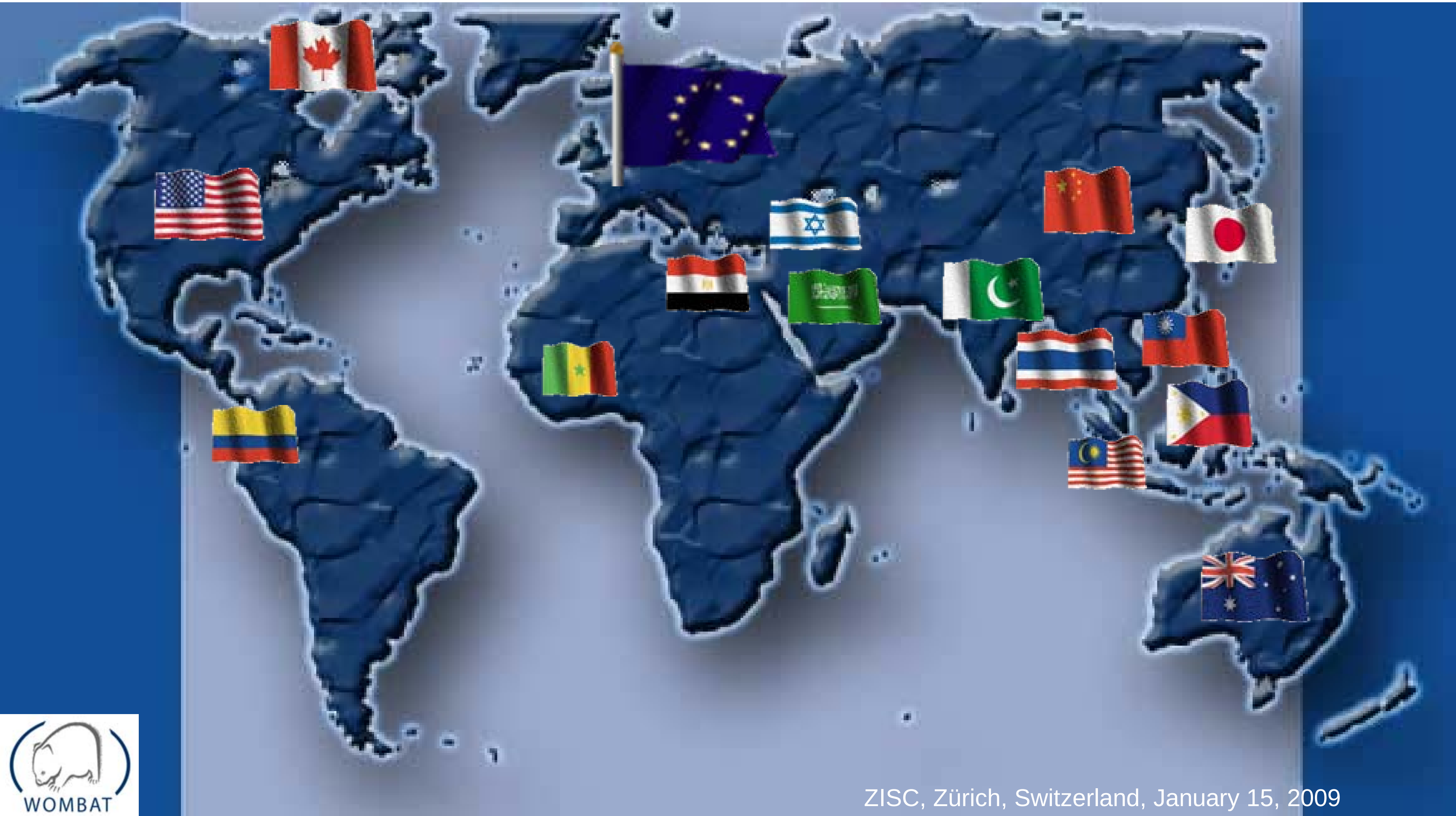
# Leurré.com V1.0

- Ongoing effort since 2003
  - Almost 50 platforms in 30 countries today
  - Uses low interaction honeypot (based on honeyd)
  - Stores all enriched tcpdump (os fingerprinting, geographical location, etc.) in an Oracle DB open to all partners.
  - Collection of tools, interfaces (java, matlab, python, etc.) and documentation available for free to all partners.

# Experimental Set Up (based on honeyd)



# 50 sensors in 30 countries (5 continents)



# Win-Win Partnership

- The interested partner provides …
  - One old PC (pentiumII, 128M RAM, 233 Mhz…) and 4 routable IP addresses,
- EURECOM offers …
  - Installation CD Rom
  - Remote logs collection and integrity check.
  - Access to the whole SQL database by means of a secure GUI and a wiki (over https) + an automated alerting system

# Leurrécom V2.0, SGNET: Goals

- Continue the conversation with the attacker up to the point where a malware is downloaded (resp. uploaded).
- Avoid using high interaction honeypots
- Help focusing on the “new” attacks, creating new paths.

# Means

- **SGNET** =
  - ***Scriptgen*** (Eurecom) +

    Argos (VU Amsterdam) +

    Nepenthes (TU Manheim) +

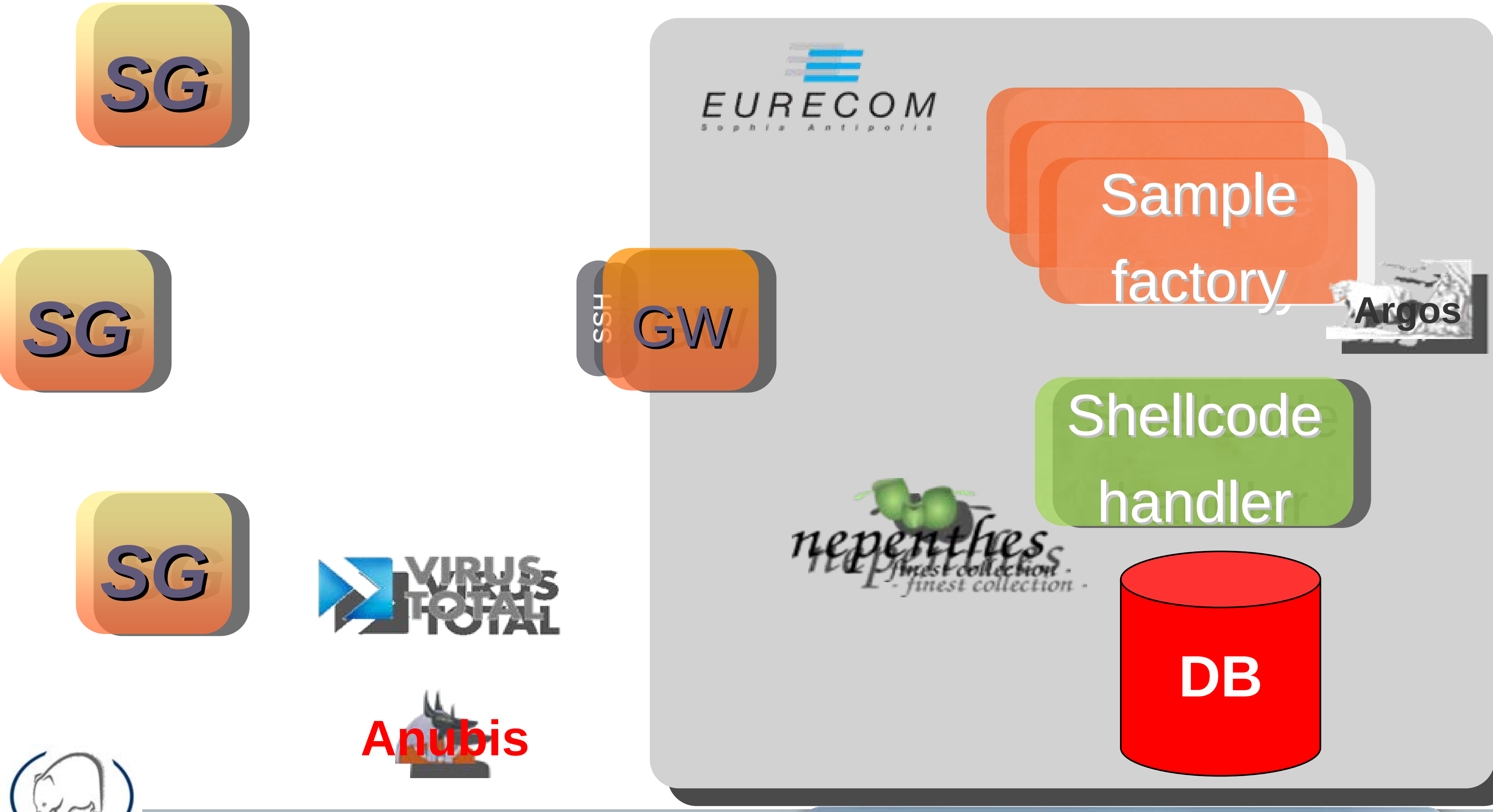
    Anubis (TU Wien) +

    Virustotal (Hispasec).
- **Scriptgen:** a novel 'medium-interaction' honeypot

# Scriptgen: concepts

- Automatically learn protocol semantics from the interactions with a real server
  - Represent learnt behavior in a state machine
- Protocol agnostic approach
  - No assumption is done neither on protocol structure, nor on its semantics.
- Similar to RolePlayer (Paxson et al.) but does not require any human intervention.
- Details published at ACSAC05, RAID06, NOMS08, EDDC08

# SGNET: The building blocks



SG

SG

SG

SSH

GW

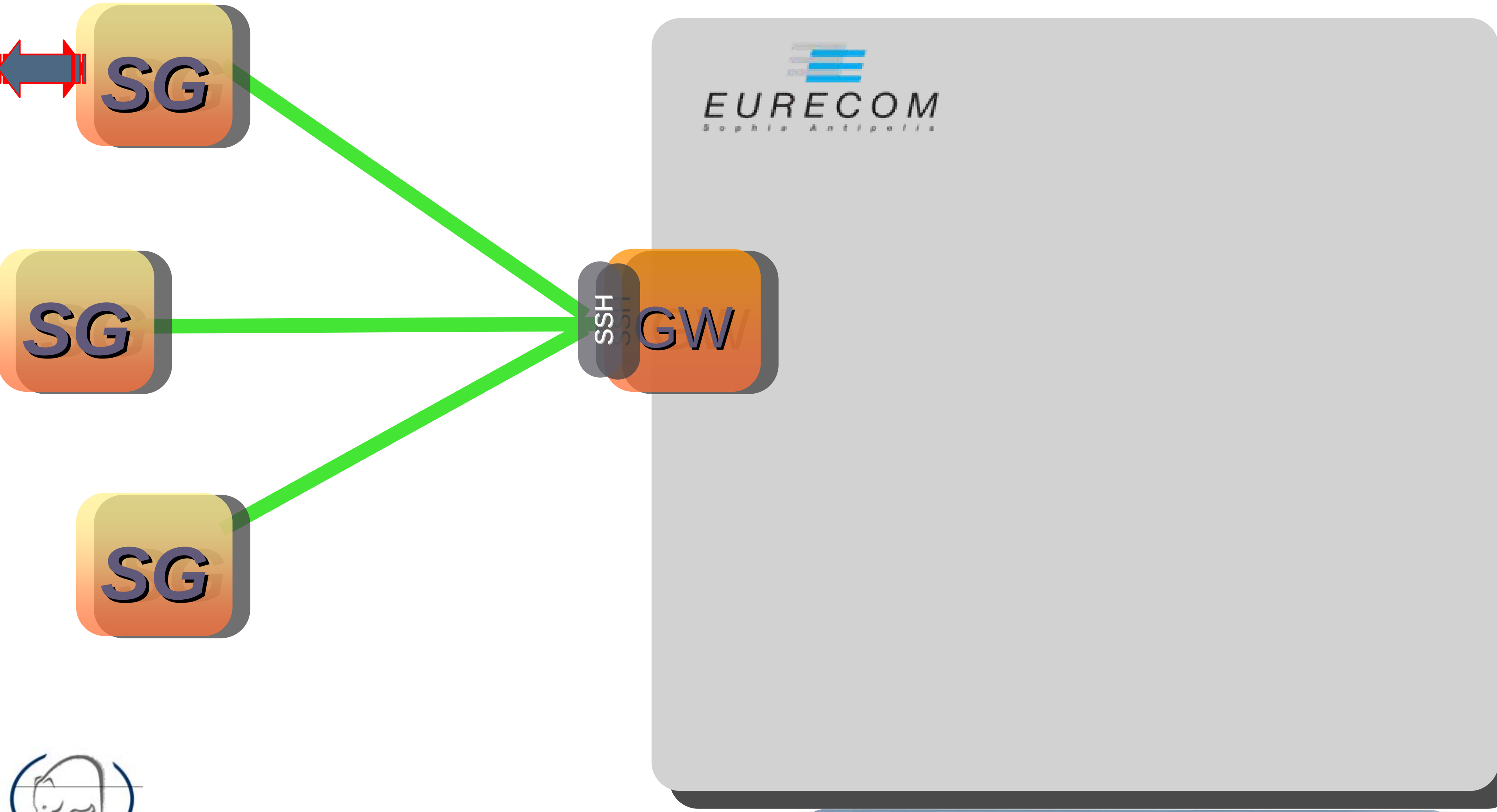
Sample factory

Argos

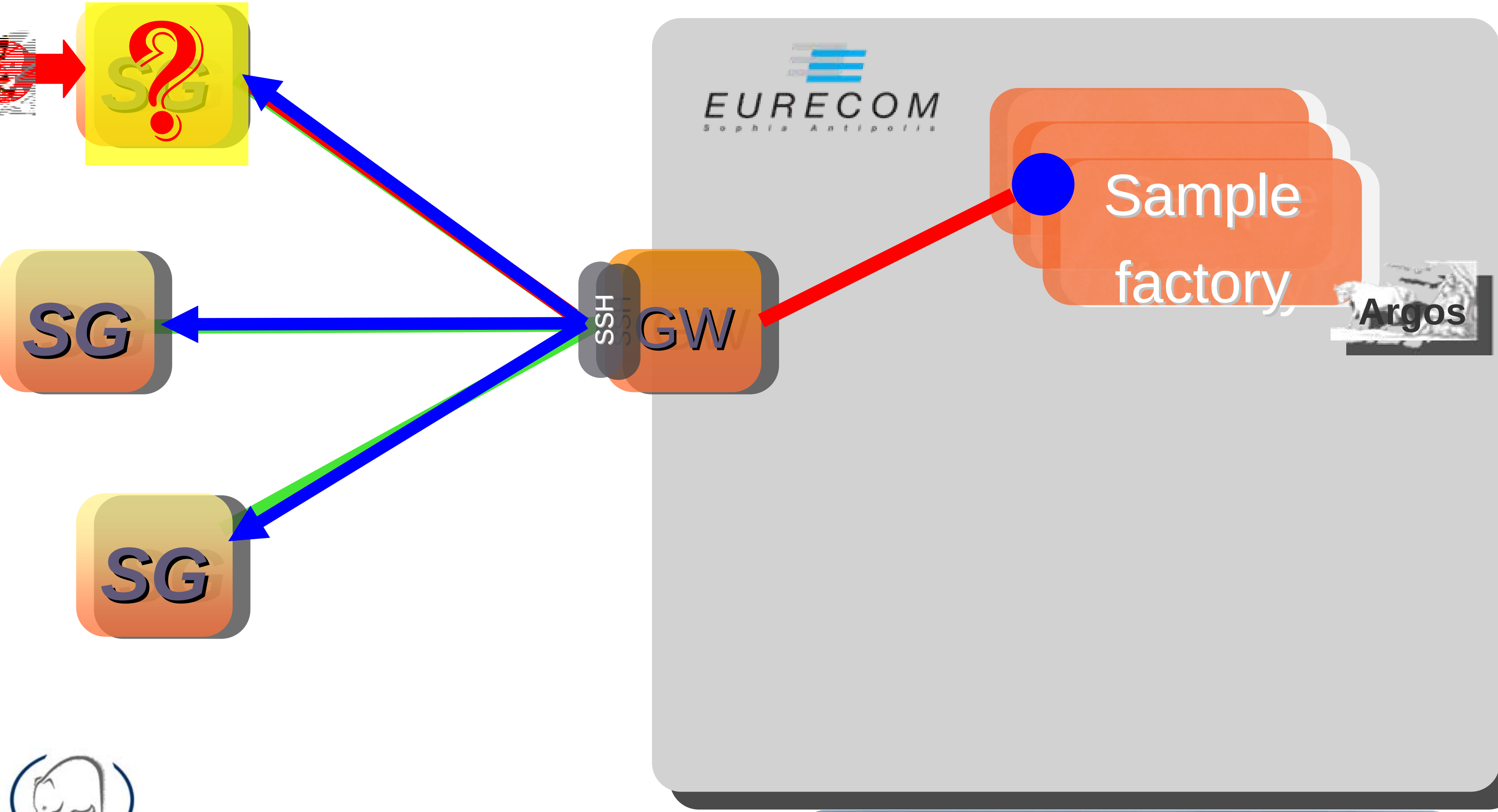
Shellcode handler

DB

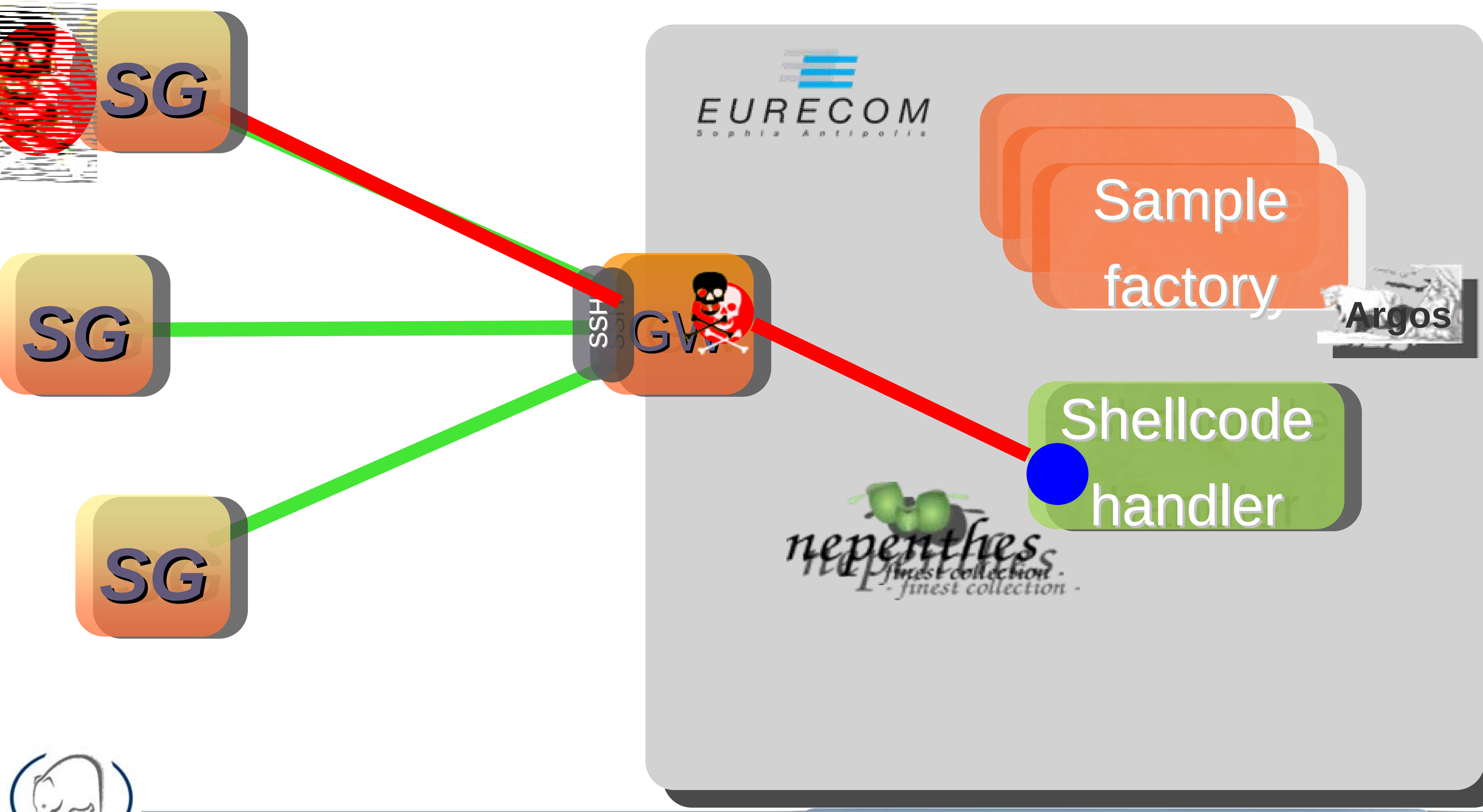
# Normal Operation



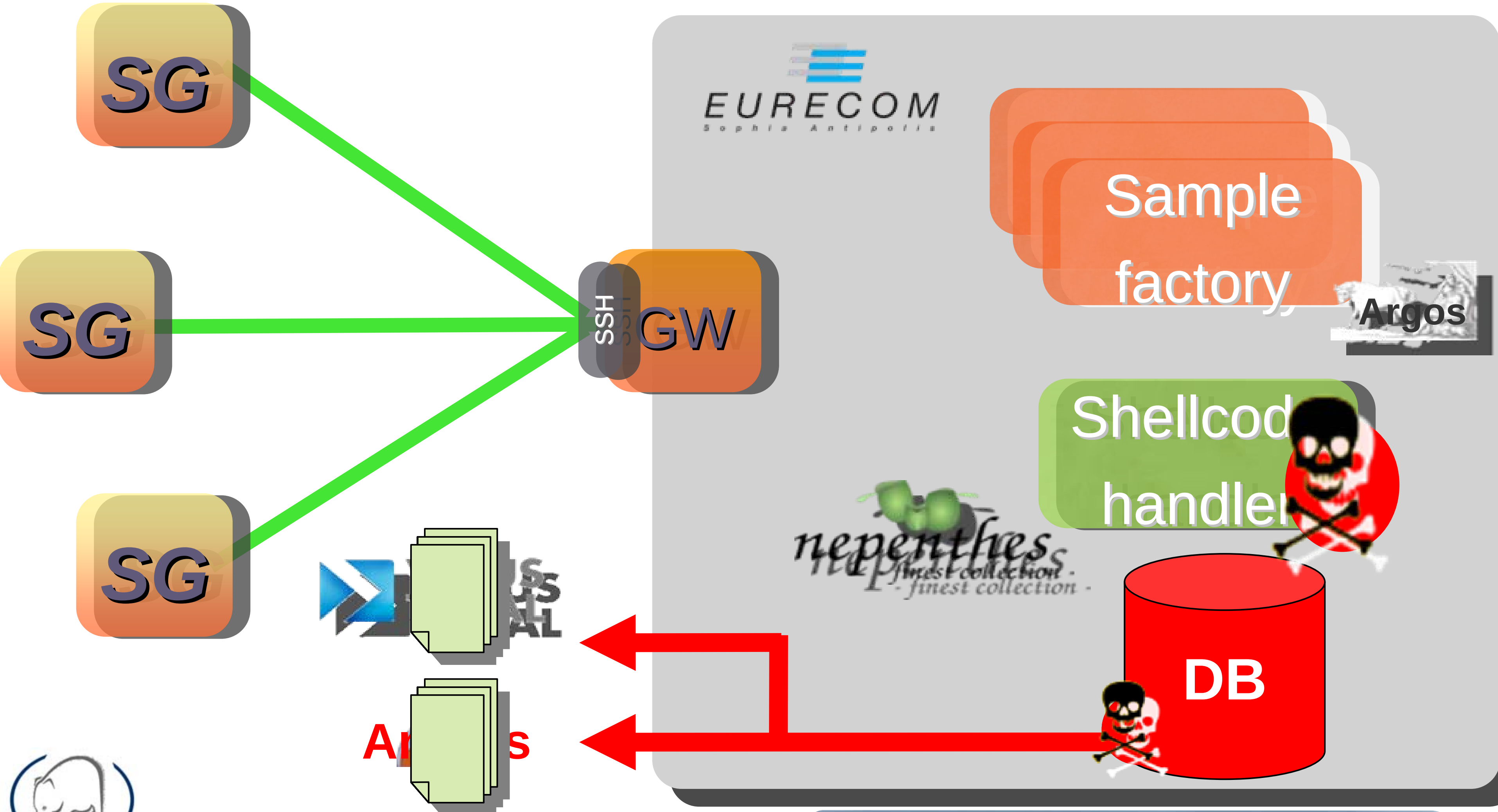
# Argos used as an Oracle for unknown attacks



# Nepenthes used to download the malware



# Enriching the database



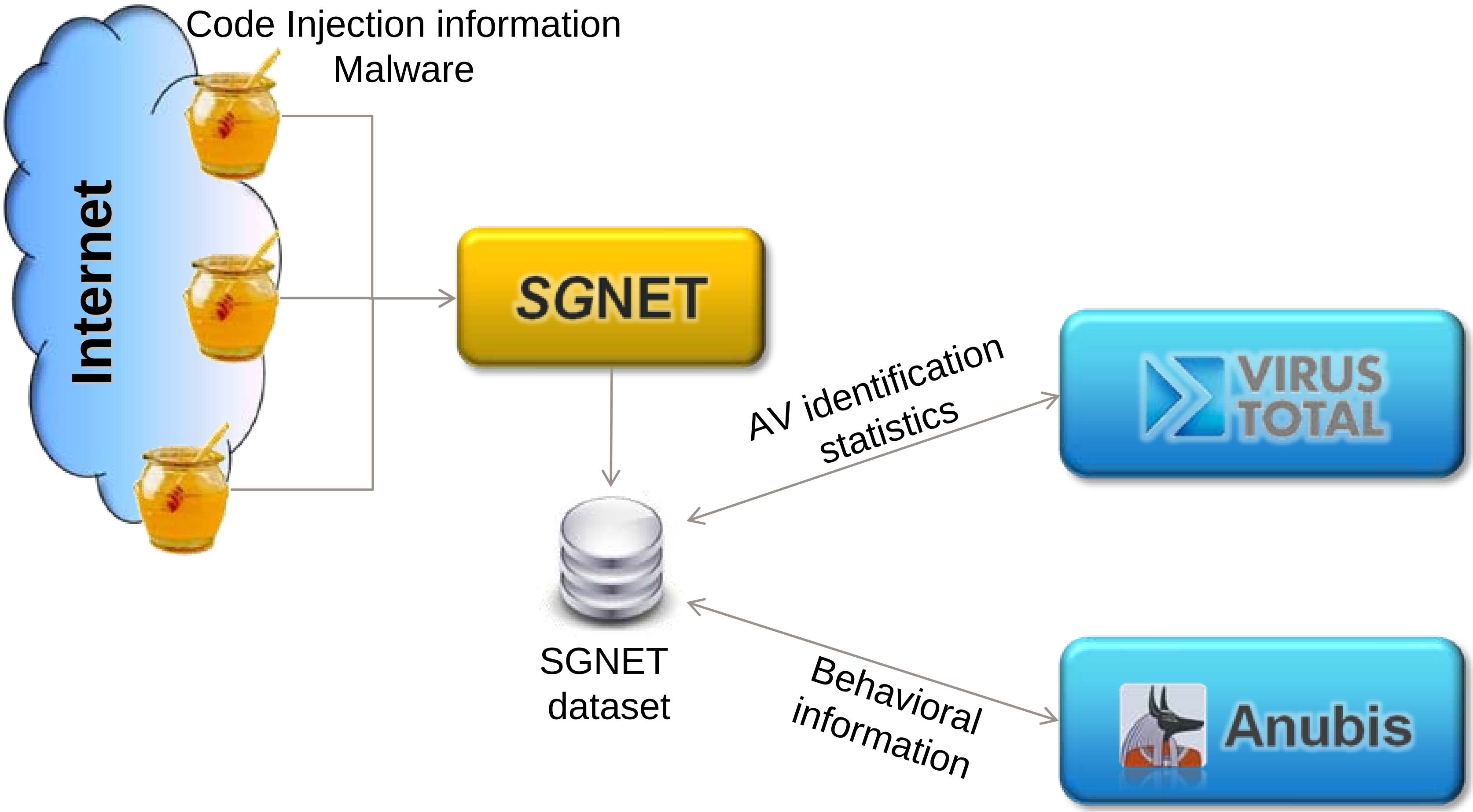
# SGNET: Benefits

- **Cheap:** 3 IPS and an old PC per sensor
- **Powerful:** as talkative as a real high interaction honeypot
- **Flexible:** all sensors reconfigured by pushing a new FSM
- **Easy:** no liability or privacy issue
- **Stealthy:** not vulnerable to VMware detection tricks
- **Customizable**: thanks to the automated learning
- **Clean**: noise-free traffic
- **Coverage**: enable to spot geographical discrepancies

# SGNET: drawbacks

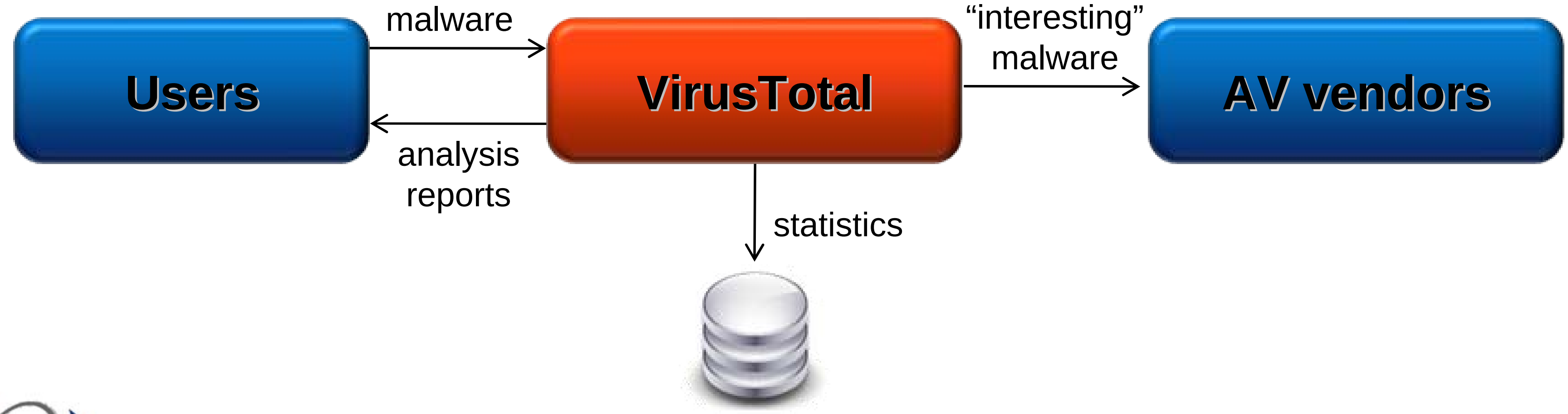
- Focus on a widespread, yet limited, set of attacks, namely remote and code injection based attacks.
- These limitations need to be addressed by means of other types of sensors (eg client based honeypots) to be developed in the context of WOMBAT.

# Our framework



# VirusTotal

- Developed and maintained by Hispasec Sistemas
- Freely accessible via a web interface
  - www.virustotal.com
  - Support for 36 AV engines (command line interface only)
  - Widely known and used by the security and AV community



# Anubis

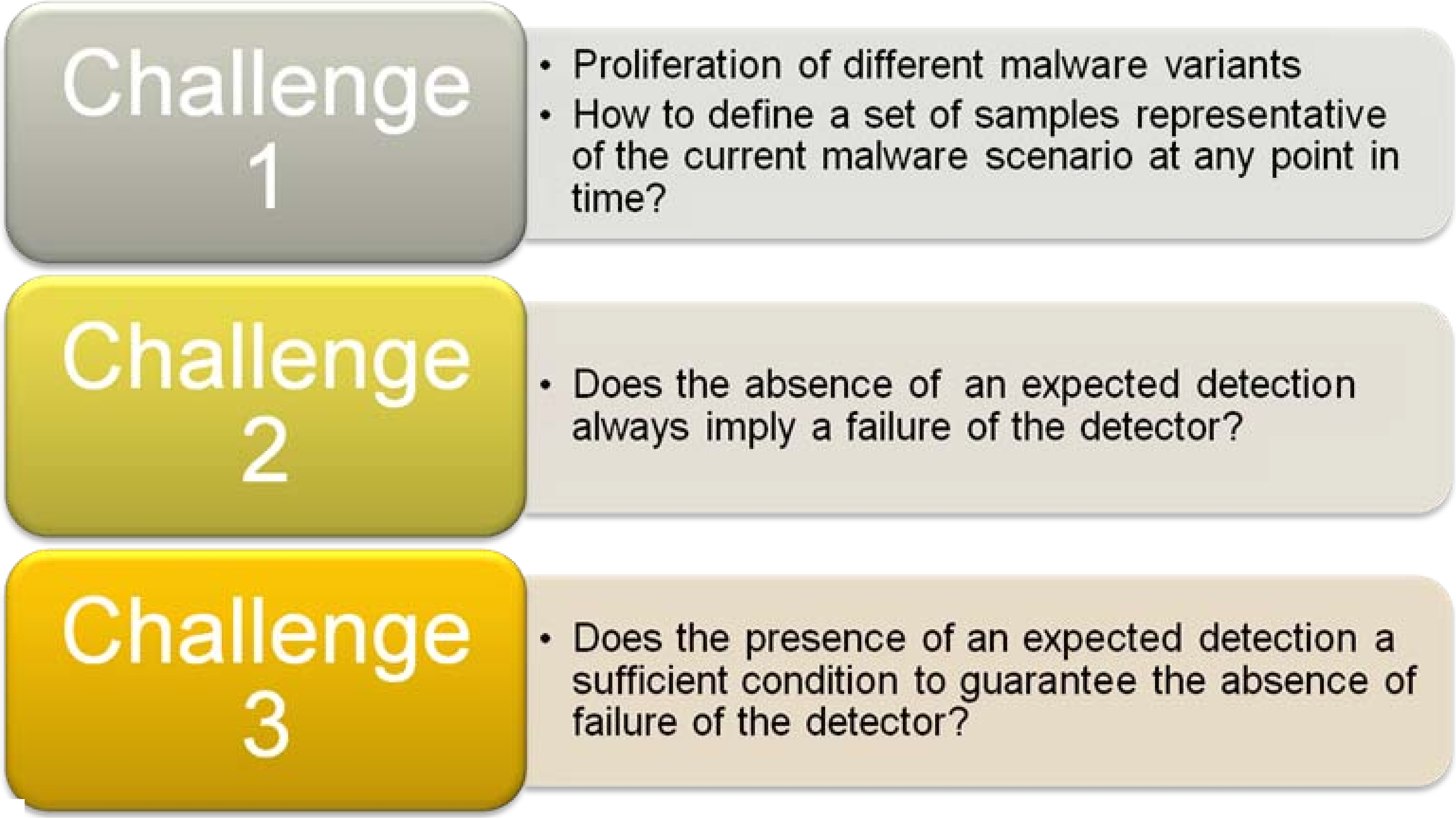
- Automated analysis of an executable file by understanding its actions
  - Modifications to Windows registry
  - Modifications to filesystem
  - Interactions with the Windows Service Manager
  - Generated network traffic
- Web interface freely accessible to submit malware and retrieve the detailed report
  - http://anubis.iseclab.org

# Submission policies

- Whenever a sample is collected by SGNET, how to relate it to the information provided by Anubis/VirusTotal?
- Anubis
  - Every sample is submitted only once
- VirusTotal
  - How does the detection performance evolve with time?
  - Daily submissions
    - At least 30 days
    - Stop after 7 identical reports

# Challenges

- Interesting challenges derived from our experience with the SGNET dataset



Challenge 1

- Proliferation of different malware variants
- How to define a set of samples representative of the current malware scenario at any point in time?

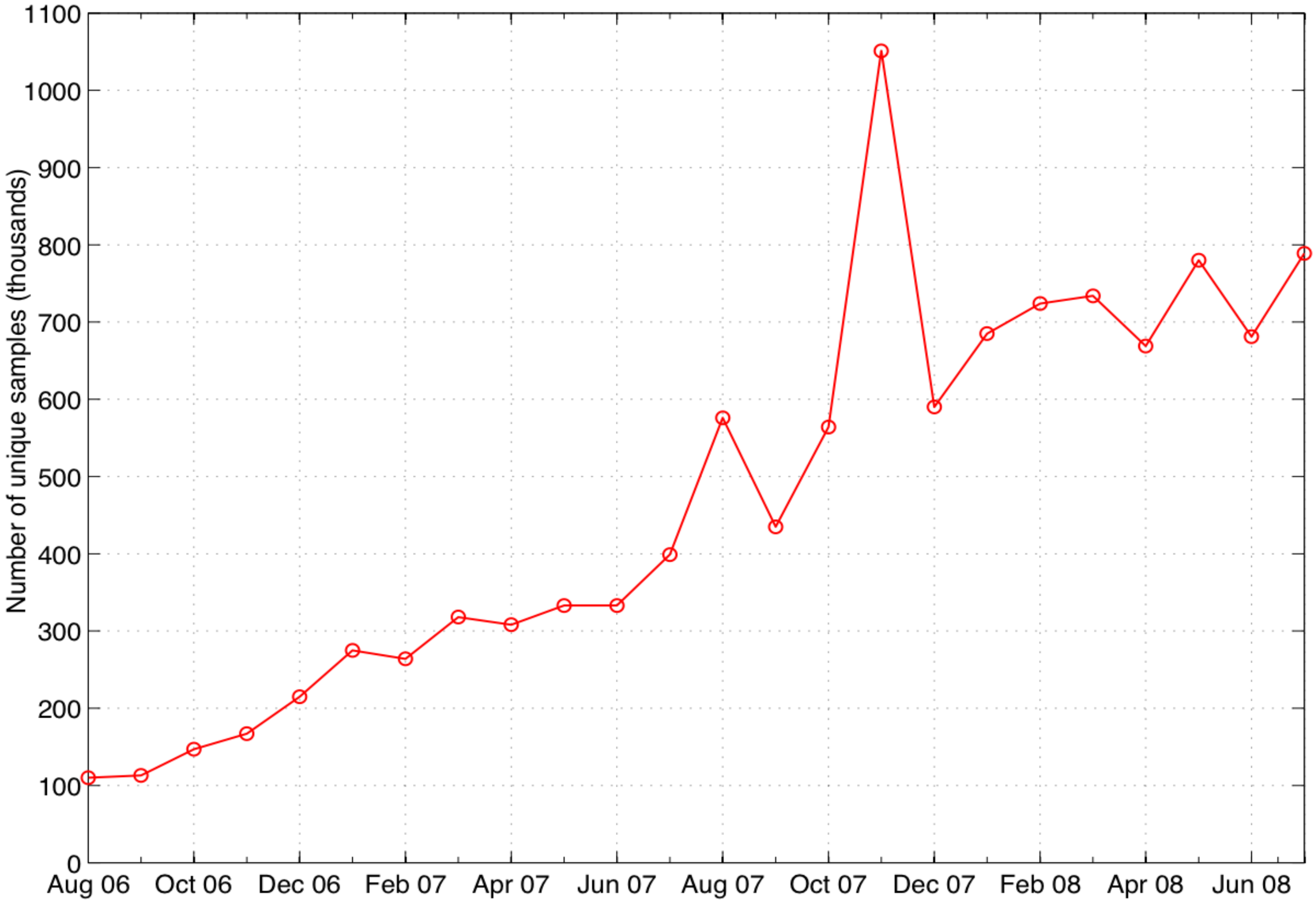
Challenge 2

- Does the absence of an expected detection always imply a failure of the detector?

Challenge 3

- Does the presence of an expected detection a sufficient condition to guarantee the absence of failure of the detector?

# Challenge 1
# The proliferation of variants

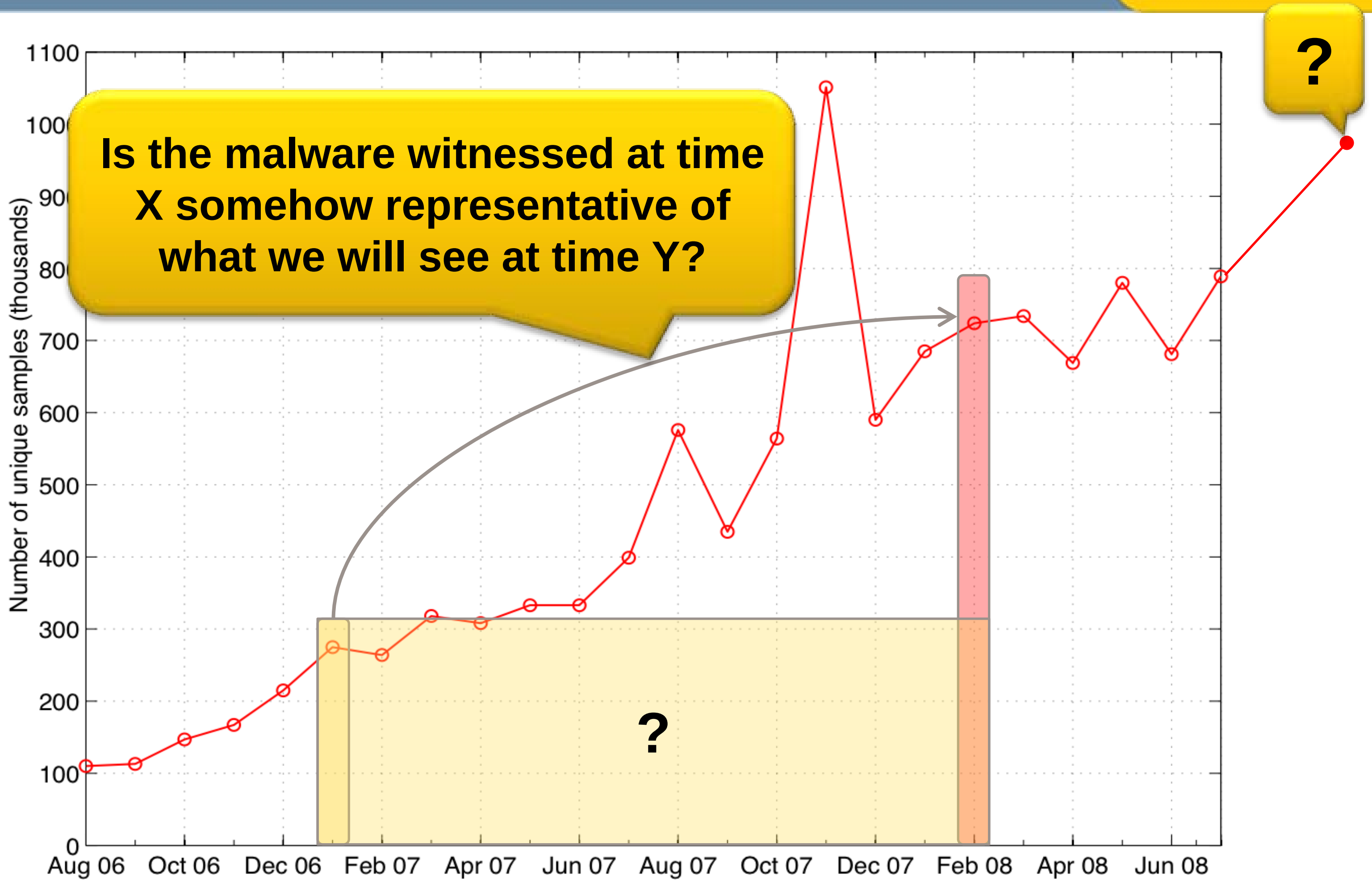


Distinct samples observed by the VirusTotal service every month

# Challenge 1
# Explanations…?

- Use of automated methods for the generation of new variants?
  - Automated generation of customized versions of the same malware
- Polymorphism?
  - At each propagation, the sample is “different”
- Server-based propagation?
  - The real malware is downloaded from a server that generates metamorphic variants

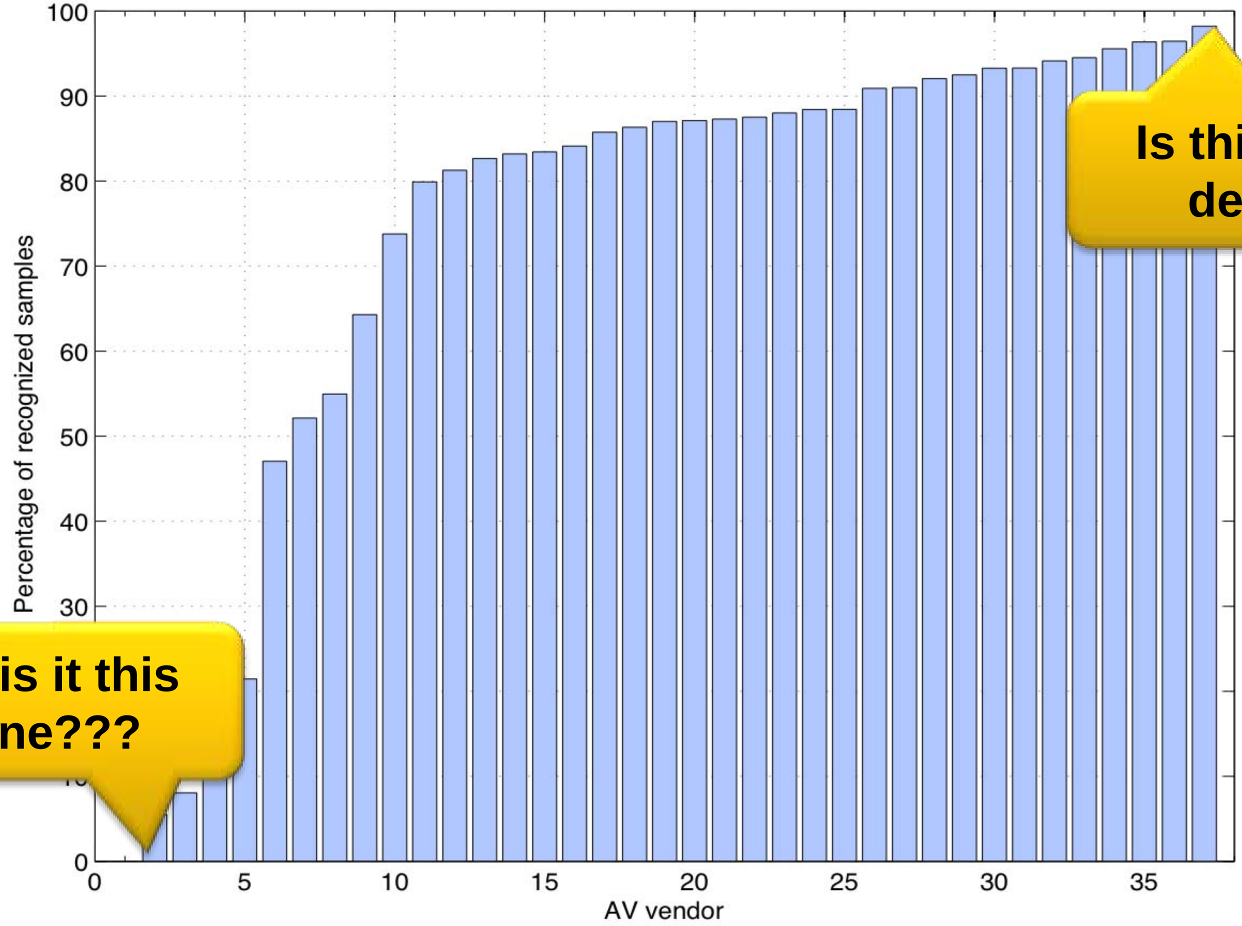
# Challenge 1
# Will we ever succeed…?



Is the malware witnessed at time X somehow representative of what we will see at time Y?

Distinct samples observed by the VirusTotal service every month

# Challenge 2
# Is a missed detection always a failure?



Is this the best detector ?

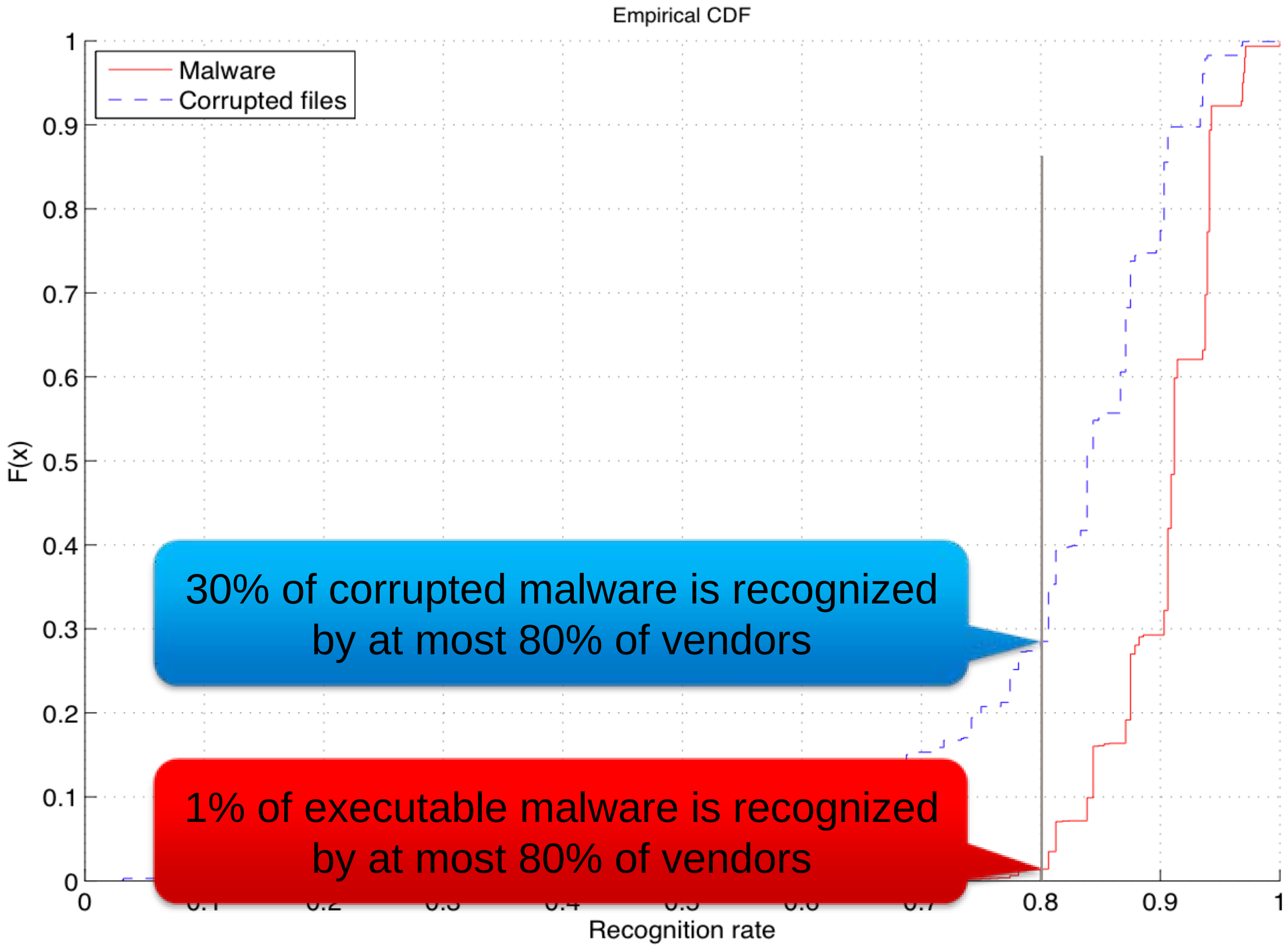
Or is it this one???

Percentage of samples detected by the different AV vendors for a selected class of samples in our dataset

# Challenge 2
# Is a missed detection always a failure?

- What’s going on?
  - Comparing the results with Anubis, we realize that these samples cannot be executed
  - Corrupted malware samples: something went wrong in the download
  - It’s probably not a rare phenomenon when using certain malware collectors (i.e. Nepenthes)
- Is the failed detection of a corrupted sample a false negative?
  - Depends on the policy
  - Depends on the implementation too!
    - A part of the sample is missing. What if the signature uses that part to deduce the nature of the malware?

# Challenge 2
# Corrupted VS executable



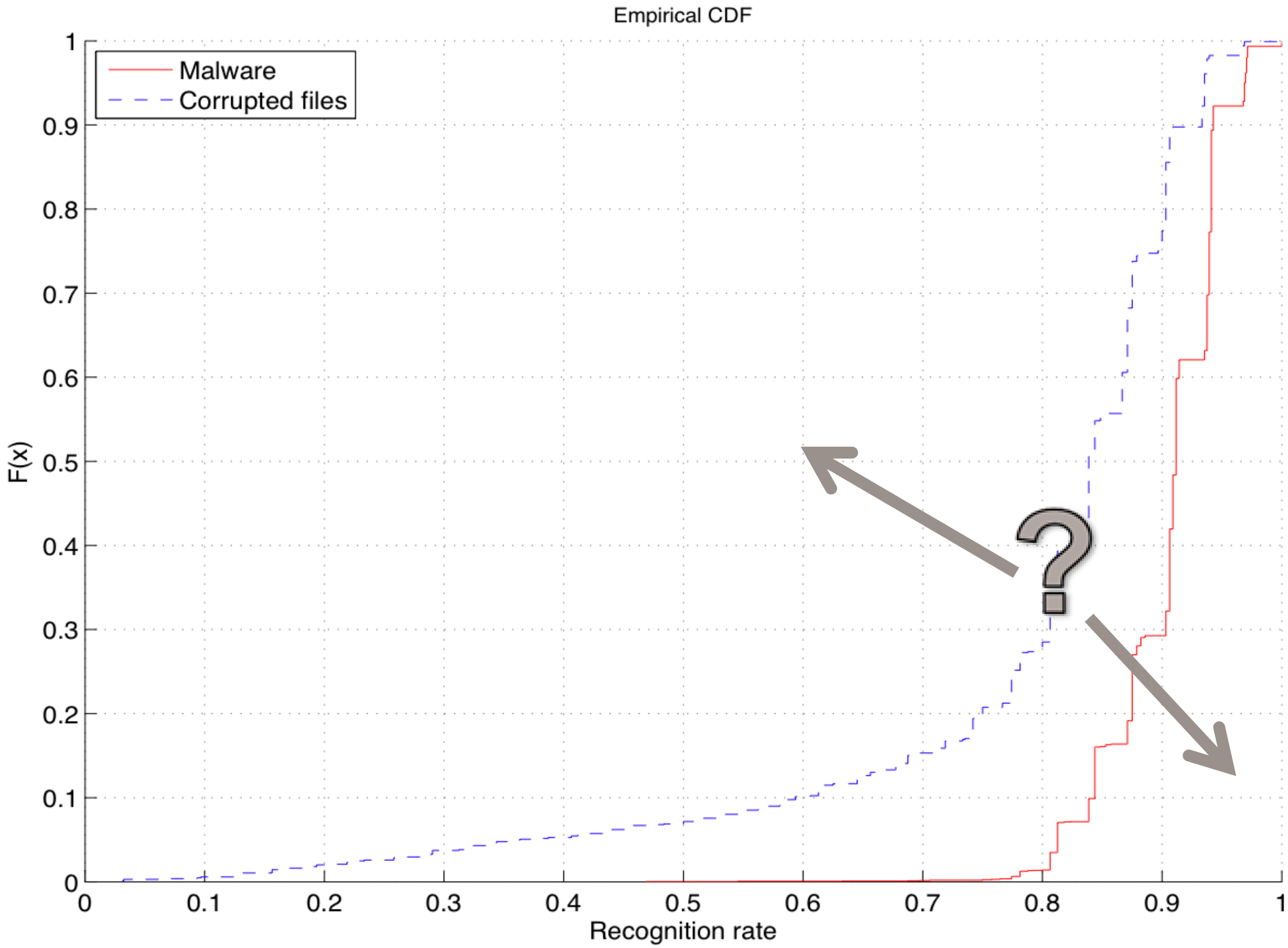
30% of corrupted malware is recognized by at most 80% of vendors

1% of executable malware is recognized by at most 80% of vendors

Comparison of the recognition rate for the two classes of samples

R_rate= (# vendors detecting the sample)/(# vendors)

# Challenge 2
# Corrupted VS executable



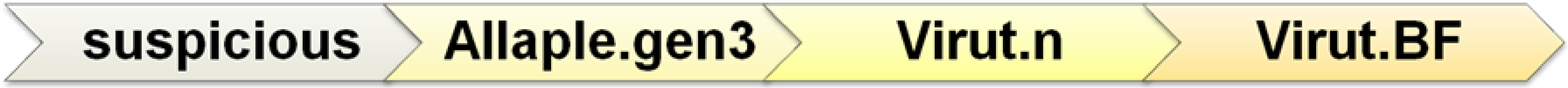
Comparison of the recognition rate for the two classes of samples

R_rate= (# vendors detecting the sample)/(# vendors)

# Challenge 3
# Labelling and classification

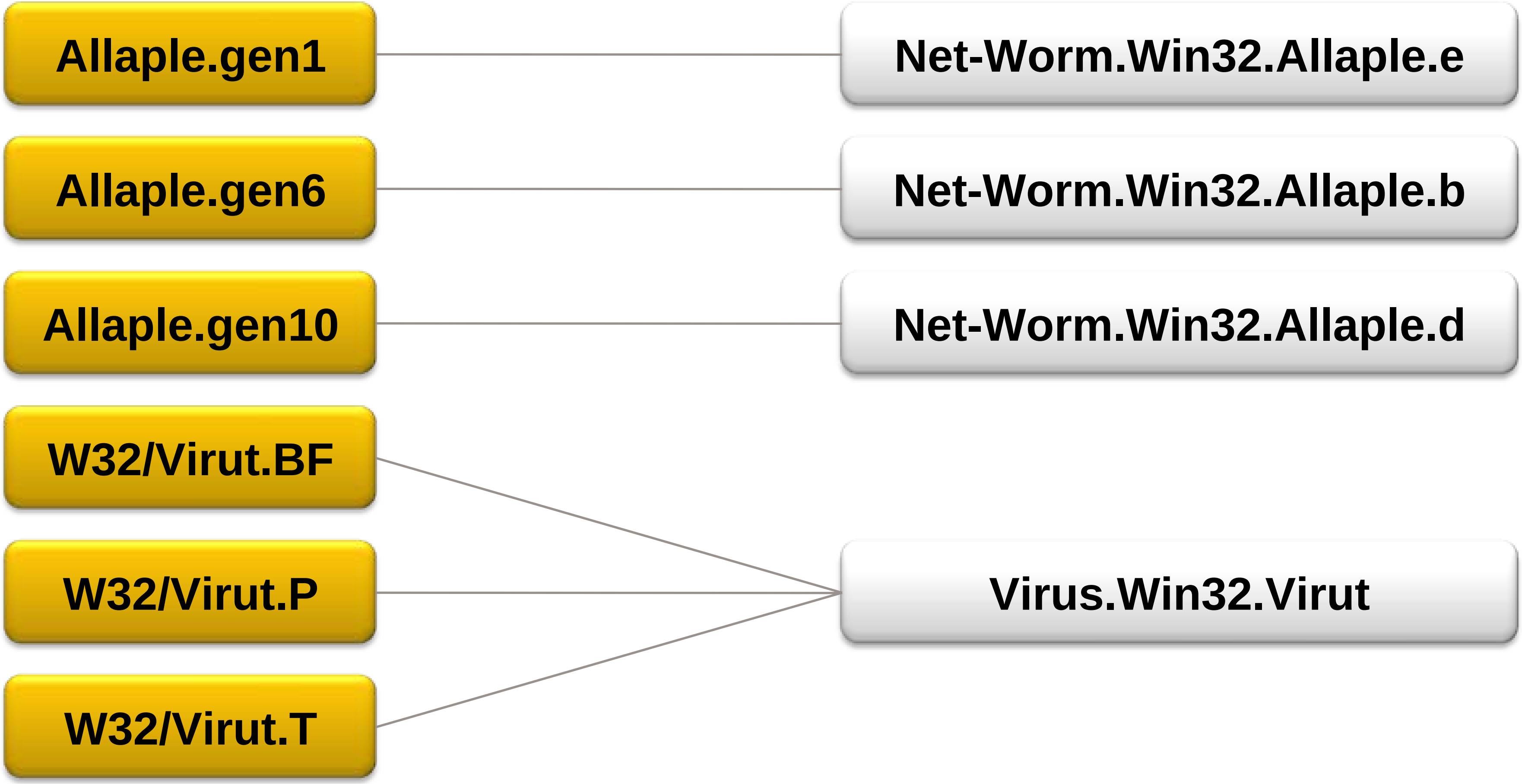
- If an alert has to be raised and indeed has been raised, is it a true positive?
  - If an alert A has to be raised and an alert B has been raised, is it a true positive?
  - How do you know A has to be raised in the first place?

- In our dataset, 10314 modifications were detected in the label associated by a vendor to a given sample over the submission period (1081 unique types of modifications)
  - Example:



suspicious → Allaple.gen3 → Virut.n → Virut.BF

# Labeling
# What we would expect…



Allaple.gen1

Allaple.gen6

Allaple.gen10

W32/Virut.BF

W32/Virut.P

W32/Virut.T

Net-Worm.Win32.Allaple.e

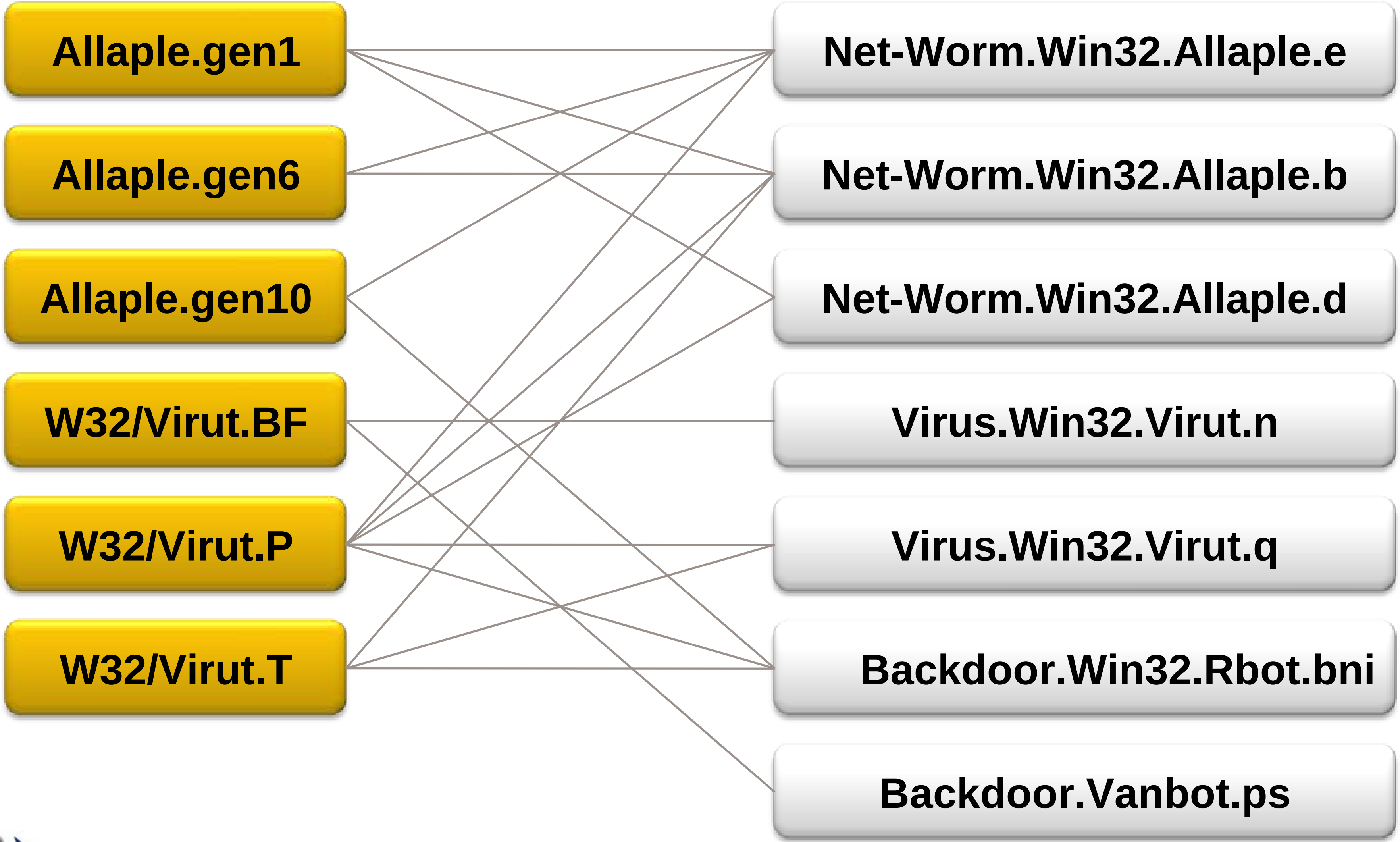
Net-Worm.Win32.Allaple.b

Net-Worm.Win32.Allaple.d

Virus.Win32.Virut

1:n relationships are allowed: for instance, one vendor uses a more generic label than the other

# Labeling
# … and what we get in practice



Allaple.gen1

Allaple.gen6

Allaple.gen10

W32/Virut.BF

W32/Virut.P

W32/Virut.T

Net-Worm.Win32.Allaple.e

Net-Worm.Win32.Allaple.b

Net-Worm.Win32.Allaple.d

Virus.Win32.Virut.n

Virus.Win32.Virut.q

Backdoor.Win32.Rbot.bni

Backdoor.Vanbot.ps

# Conclusion

- The generation of good benchmarks for malware detection techniques is a challenging problem
  - Amount and dynamics of nowadays malware makes the generation of an exhaustive sample set an almost impossible task
  - Importance of filtering samples to spot cases that could potentially lead to ambiguities
  - Problem of labeling: how to define whether the label assigned to a sample is correct?
- We are not (yet) able to provide answers to these challenges
- These challenges need to be addressed for benchmarks to provide meaningful results

# Thank You!

Marc Dacier

marc_dacier@symantec.com

© 2007 Symantec Corporation. All rights reserved.

THIS DOCUMENT IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY AND IS NOT INTENDED AS ADVERTISING. ALL WARRANTIES RELATING TO THE INFORMATION IN THIS DOCUMENT, EITHER EXPRESS OR IMPLIED, ARE DISCLAIMED TO THE MAXIMUM EXTENT ALLOWED BY LAW. THE INFORMATION IN THIS DOCUMENT IS SUBJECT TO CHANGE WITHOUT NOTICE.