A data-driven approach to generate threat intelligence

Corrado Leita (corrado_leita@symantec.com)

Collaborative Advanced Research Department

Symantec Research Labs

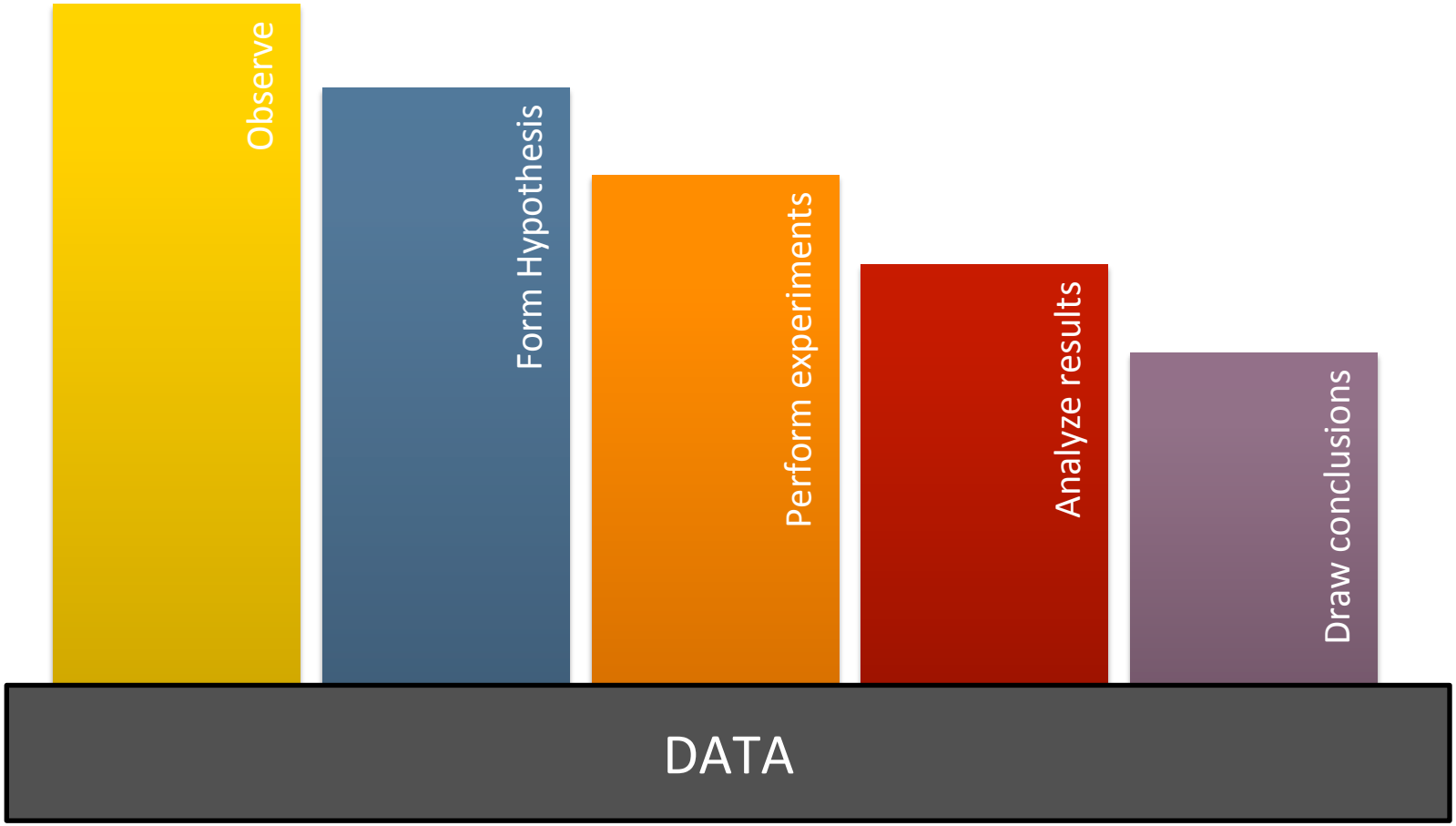
This talk is about data, vampires, chrysalises, wombats and wine

1. **Why** do we need data?
2. **What** to look for?
3. **How** should we collect it?
4. **Where** can we get it?

Why?

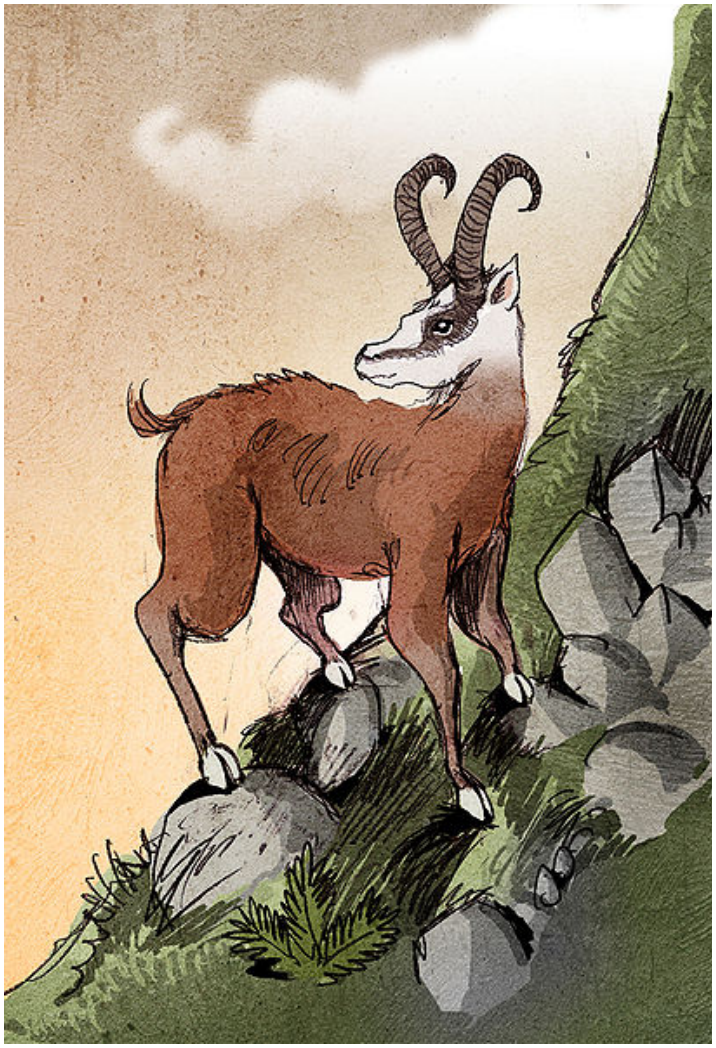
We should not be trying to catch a dahus

How do (should) we do research?



The importance of observation

- What should we **really** do research on?



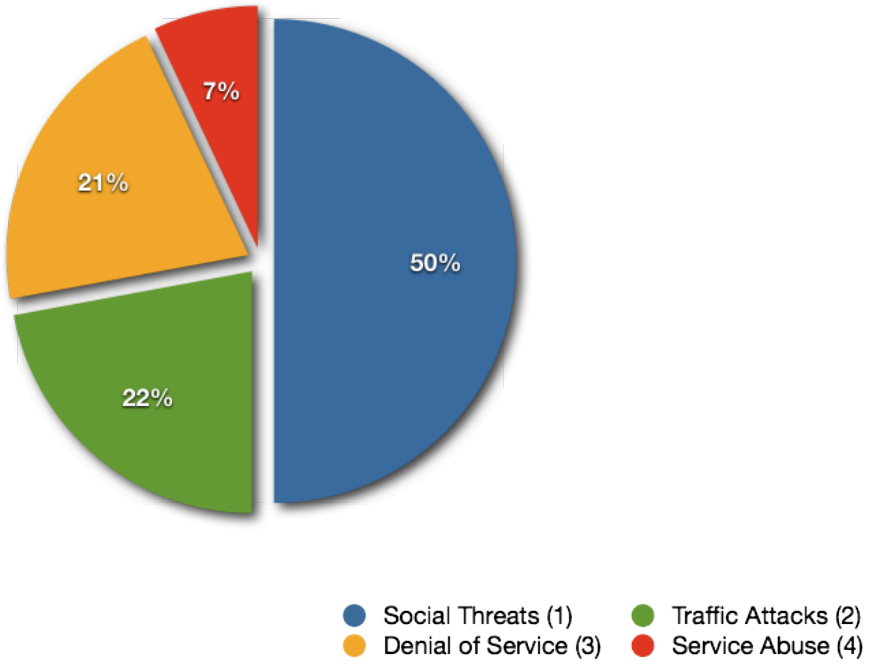
Angelos Keromytis, "A comprehensive survey of Voice over IP Security Research", IEEE Communications Surveys and Tutorials (to appear)

What should be worry about?

Vampire

- VAMPIRE project: Future Internet Vulnerability Assessment, Monitoring and Prevention
  - More on http://vampire.gforge.inria.fr/

VoIP Vulnerability Space



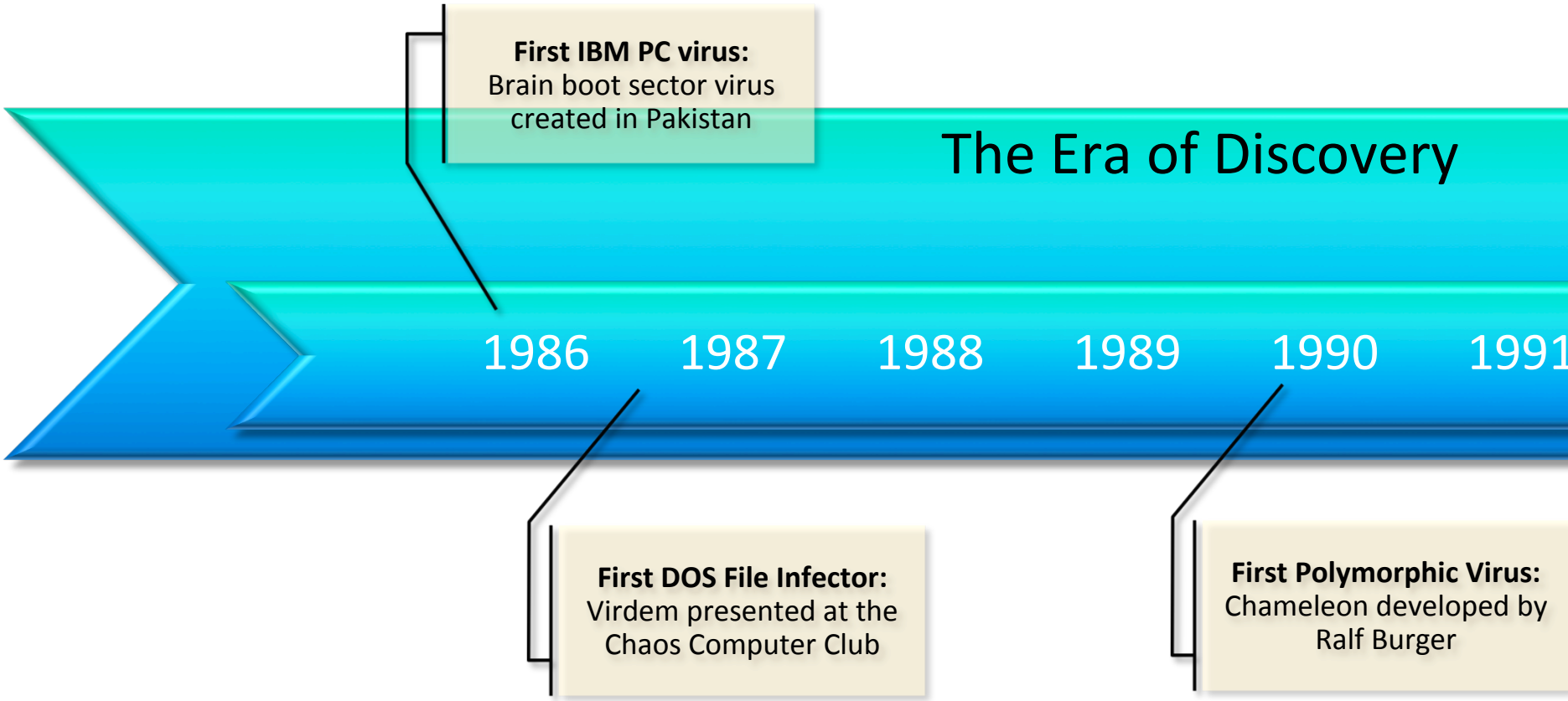
What?

It’s all about the money

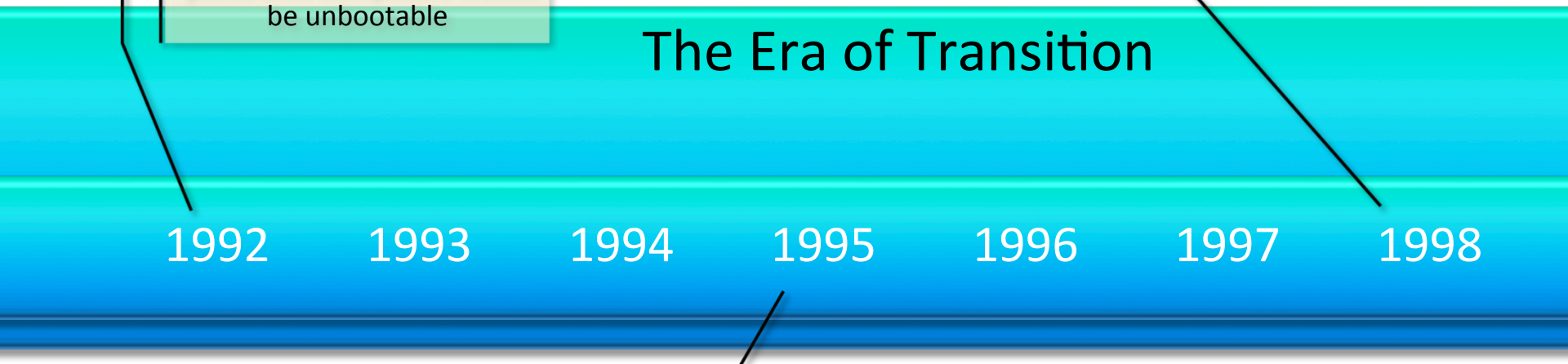
What should we look for?

- The threat landscape is continuously evolving
  - Different targets
  - Different dynamics
  - Different levels of complexities

- What kind of phenomena can hit our environment?
  - Self propagating malware?
  - Web-borne threats?
  - ….?


First IBM PC virus:
Brain boot sector virus created in Pakistan
The Era of Discovery
1986
1987
1988
1989
1990
1991
First DOS File Infector:
Virdem presented at the Chaos Computer Club
First Polymorphic Virus:
Chameleon developed by Ralf Burger

The Era of Transition

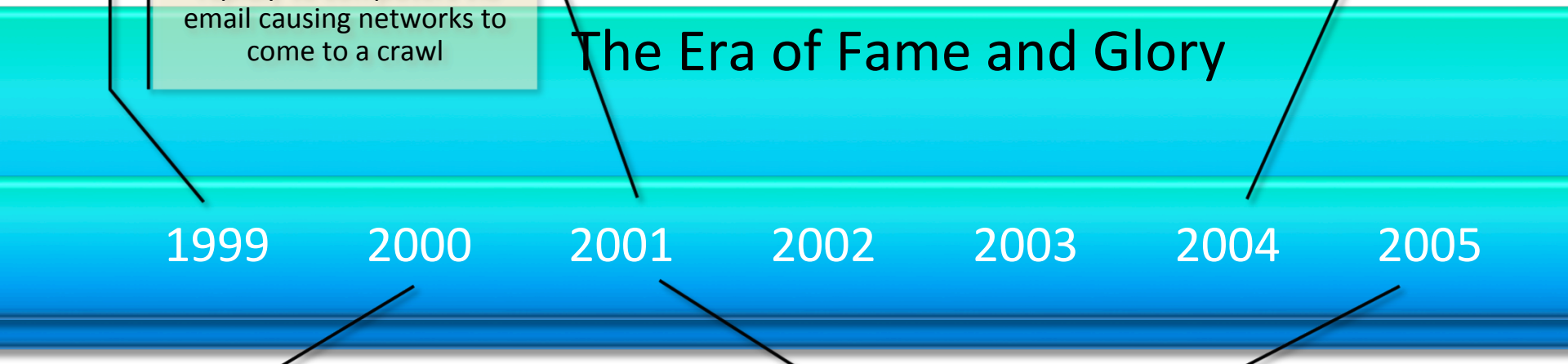


Michaelangelo trigger date:
Causes widespread media panic that computers would be unbootable

CIH:
A Windows file infector that would flash the BIOS

First Word Macro virus:
Concept is the first macro virus infected Microsoft Word documents

The Era of Fame and Glory



Email systems down:
The Melissa worm spreads rapidly to computers via email causing networks to come to a crawl

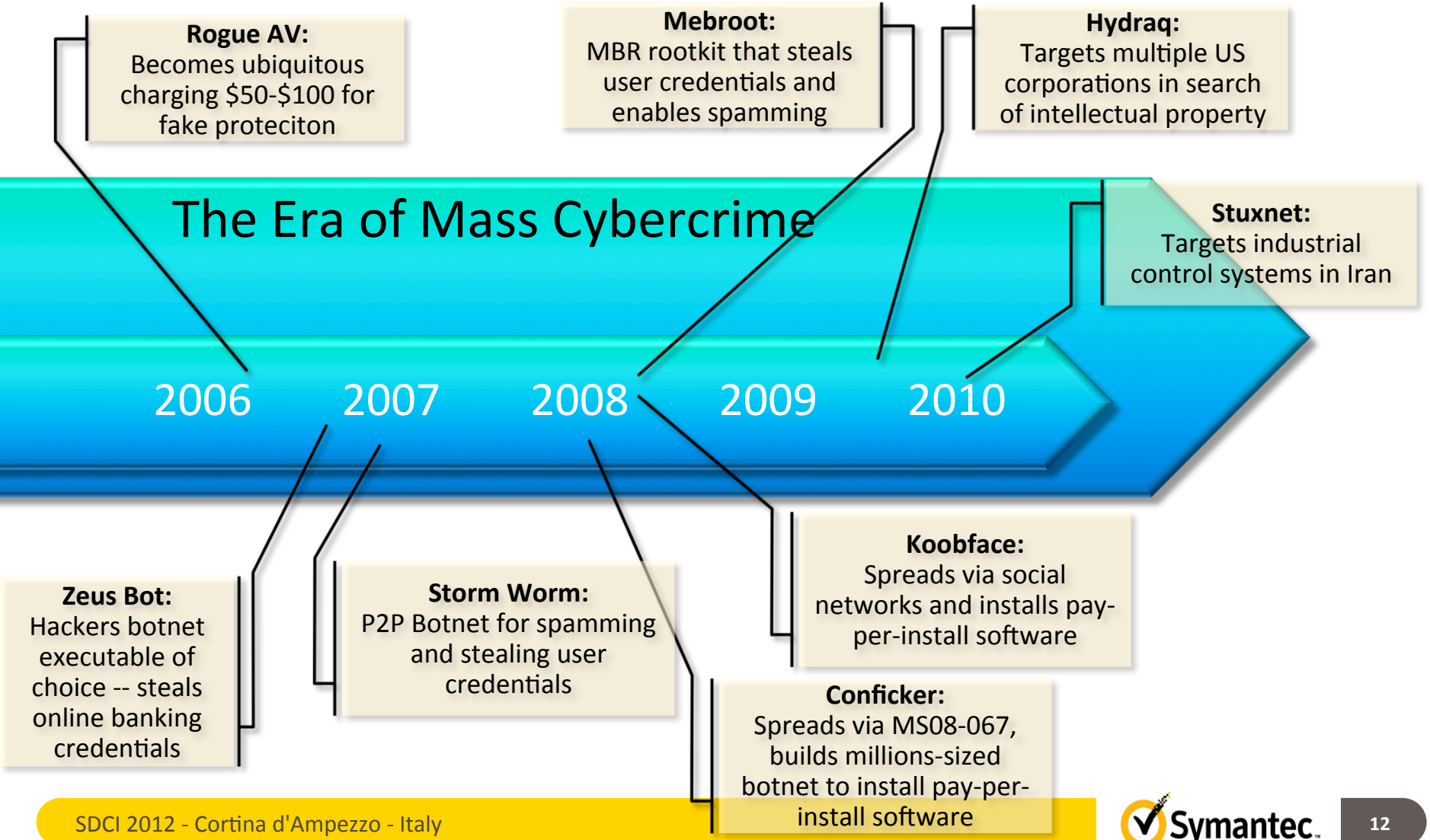
Blended Threats:
CodeRed, Nimda spread without any user interaction using Microsoft system vulnerabilities

Worm wars:
MyDoom, Netsky, Sobig, all compete for machines to infect

LoveLetter Worm:
First VBS script virus to spread rapidly via Outlook email

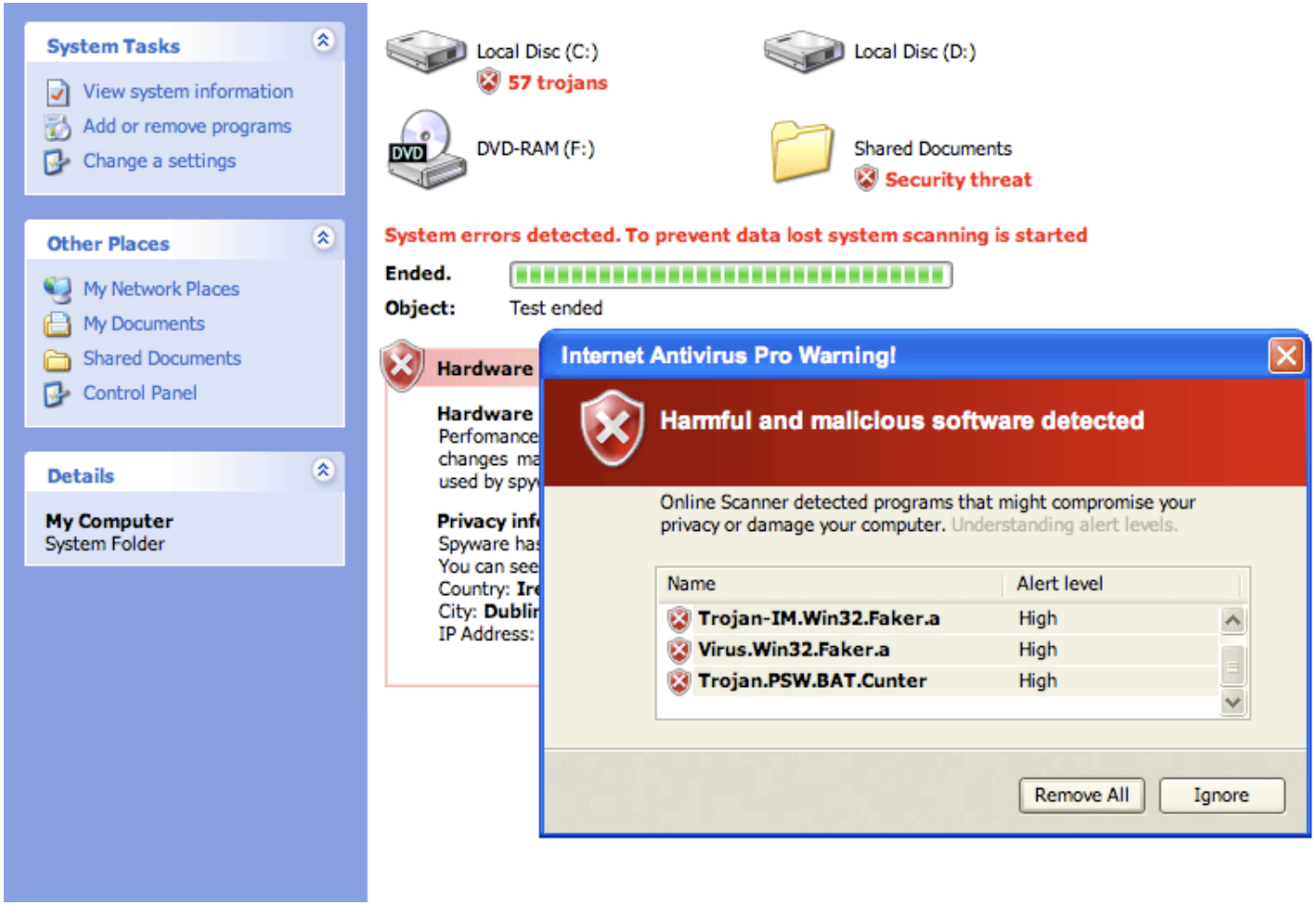
Anna Kournikova:
Just another email worm, but successful in propagation using racy pictures of Anna Kournikova as bait

Samy My Hero:
XSS worm spreads on MySpace automatically friending a million users


Rogue AV:
Becomes ubiquitous charging $50-$100 for fake proteciton
Mebroot:
MBR rootkit that steals user credentials and enables spamming
Hydraq:
Targets multiple US corporations in search of intellectual property
The Era of Mass Cybercrime
Stuxnet:
Targets industrial control systems in Iran
2006
2007
2008
2009
2010
Zeus Bot:
Hackers botnet executable of choice -- steals online banking credentials
Storm Worm:
P2P Botnet for spamming and stealing user credentials
Koobface:
Spreads via social networks and installs pay-per-install software
Conficker:
Spreads via MS08-067, builds millions-sized botnet to install pay-per-install software

Example: Rogue security software

All done in Javascript! The page was actually rendered by Safari on a Mac



Key question mark

- We know a lot about specific instances of this threat and their strategies
  - TrafficConverter.biz
  - Antivirus 2008, …
- What about the infrastructure used to propagate these threats?
  - What is the big picture?
  - Is there anything different from other threat landscapes (e.g. browser exploits)?

1. Rogue AV domains

- Collection of lists of domains **likely** to be related to the threat
  - Norton Safeweb
  - Malware Domain List (malwaredomainlist.com)
  - Malware URL (www.malwareurl.com)
  - Hosts File (www.hosts-file.com)
- Integration of the lists taking advantage of passive DNS services (robtex.com): **which other domains are hosted on the same server?**
  - Additional information on the nature of the server
  - Can lead to the discovery of other malicious domains originally unknown

2. Information enrichment

- Large amount of information collected on each considered domain
  - Information on the security state
    - **Norton Safeweb information (SHASTA):** what kind of threats are known on this domain?
    - **Google Safe Browsing blacklists:** is this domain believed to be unsafe?
  - Information on the domain
    - **Registration information (WHOIS):** who registered domain? When? Where?
    - **DNS relations:** IP address of the web servers and of the name servers
  - Information on the servers
    - **Geolocation and AS information:** where is the server located?
    - **Server uptime and version string:** availability of the server (response to HTTP HEAD) and version advertised in the HTTP headers

3. Generation of feature sets and data mining

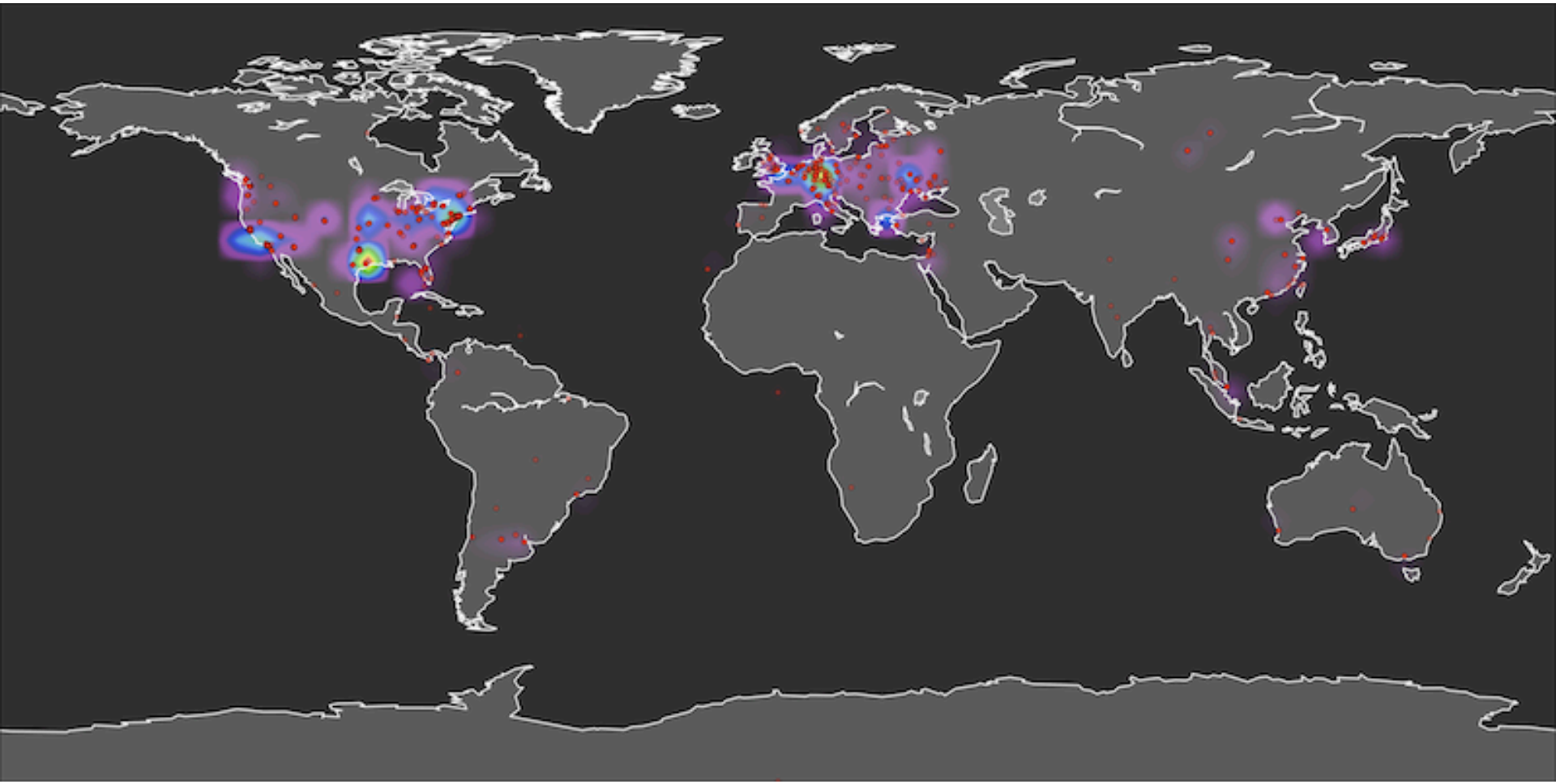
- Multi-Criteria Decision Analysis (MCDA). Attack attribution method developed by Olivier Thonnard in the context of WOMBAT.
- In a few words: group together domains that seem to be related when looking at them from multiple standpoints (features)
  - Registrant email address
  - Registrar name
  - Web server IP address/class C/class B subnets
  - Nameserver IP address
  - Registered domain name

What did we look at?

- 6,500 distinct domain names
- 4,305 web servers
  - 2,677 hosting only rogue AV domains
  - 118 hosting rogue AV domains and domains associated to other threats
  - 1,510 hosting both rogue AV domains and benign domains

- 45% of the domain names were registered through only 29 registrars

Geolocation

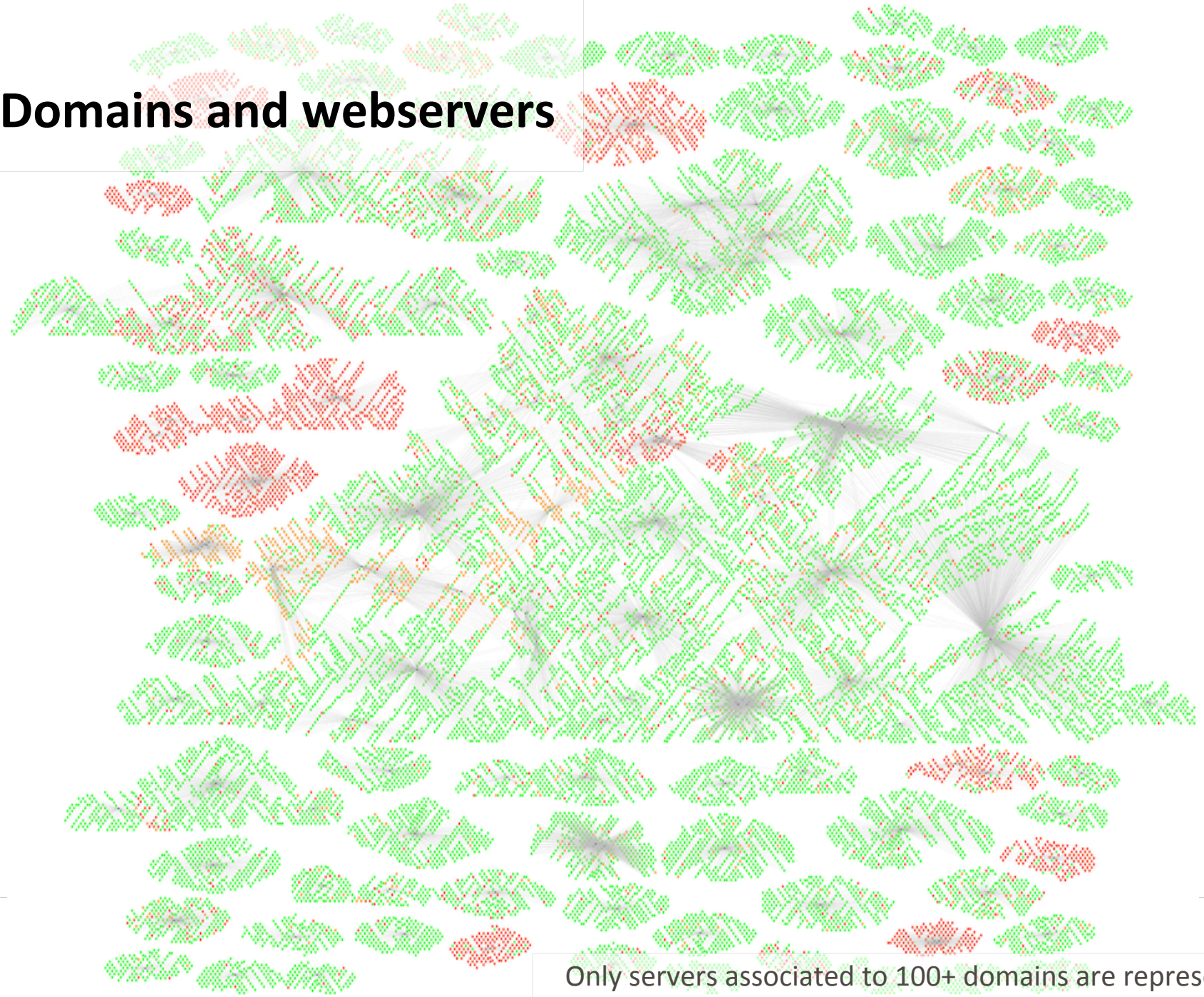


Server versions

- Very specific, but recurring version strings are advertised in the HTTP headers. Systematic deployment through “templates”?
- Apache used in over 40% of the installations

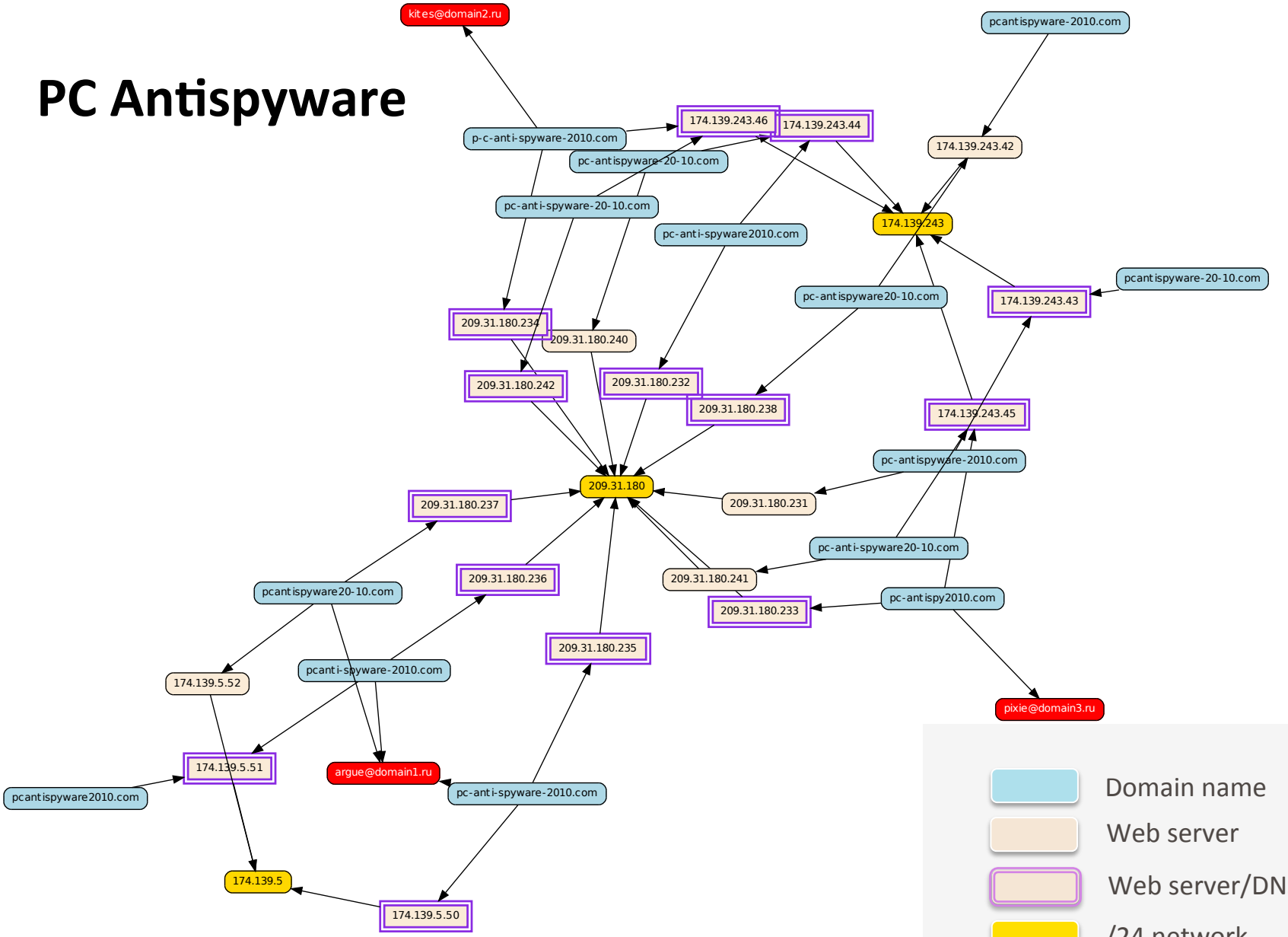
Advertised version string	#servers
Apache	610
Microsoft-IIS/6.0	218
Apache/2.2.3 (CentOS)	135
Apache/2.2.3 (Red Hat)	123
Apache/2	100
Apache/2.2.11 (Unix) mod ssl/2.2.11 OpenSSL/0.9.8i DAV/2 mod auth passthrough/2.1 mod bwlimited/ 1.4 FrontPage/5.0.2.2635	69
Apache/2.0.52 (Red Hat)	49
nginx	33
Apache/2.2.11 (Unix) mod ssl/2.2.11 OpenSSL/0.9.8e-fips-rhel5 mod auth passthrough/2.1 mod bwlimited/1.4 FrontPage/5.0.2.2635	32
LiteSpeed	26

Domains and webservers

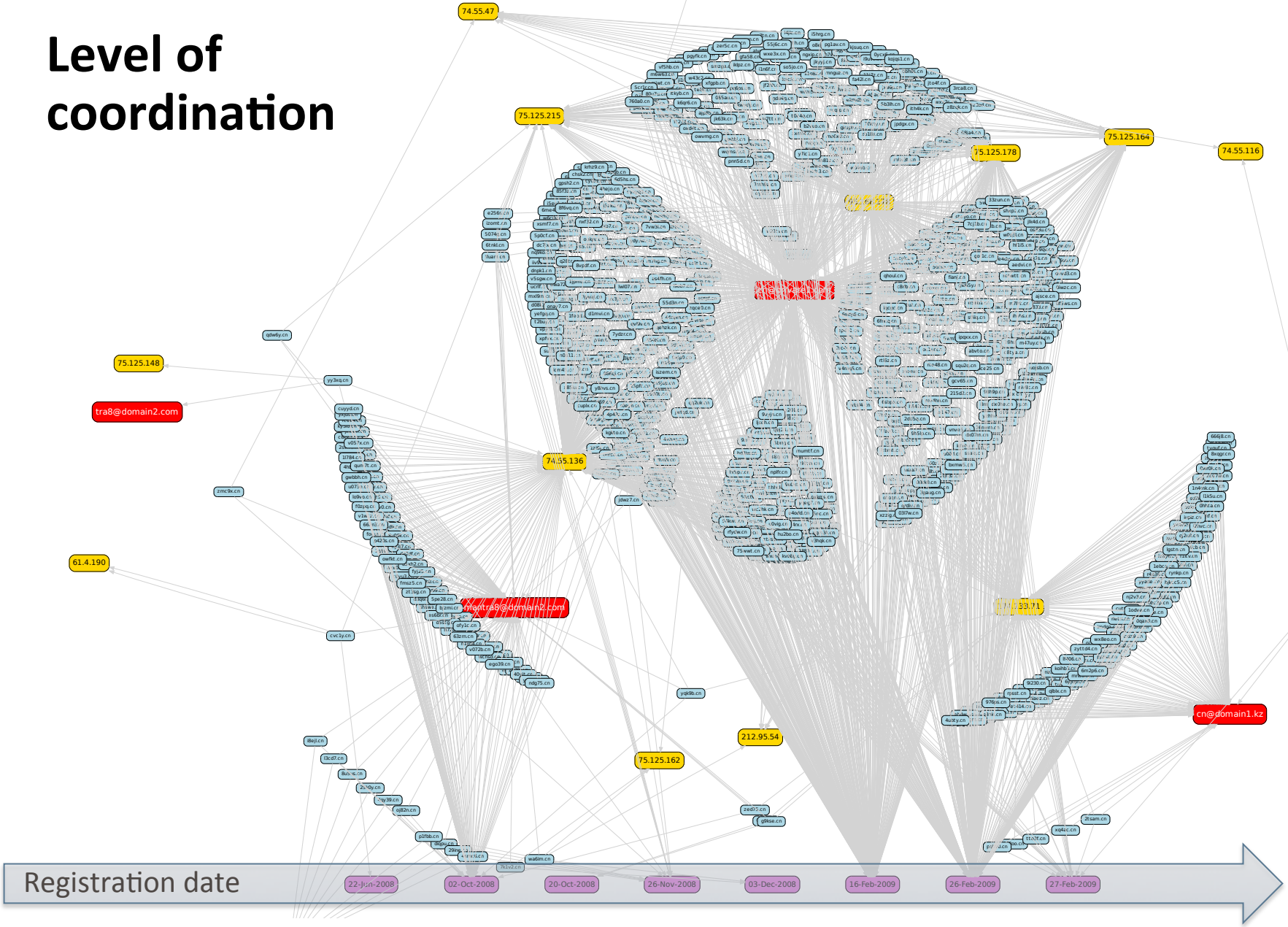


Only servers associated to 100+ domains are represented

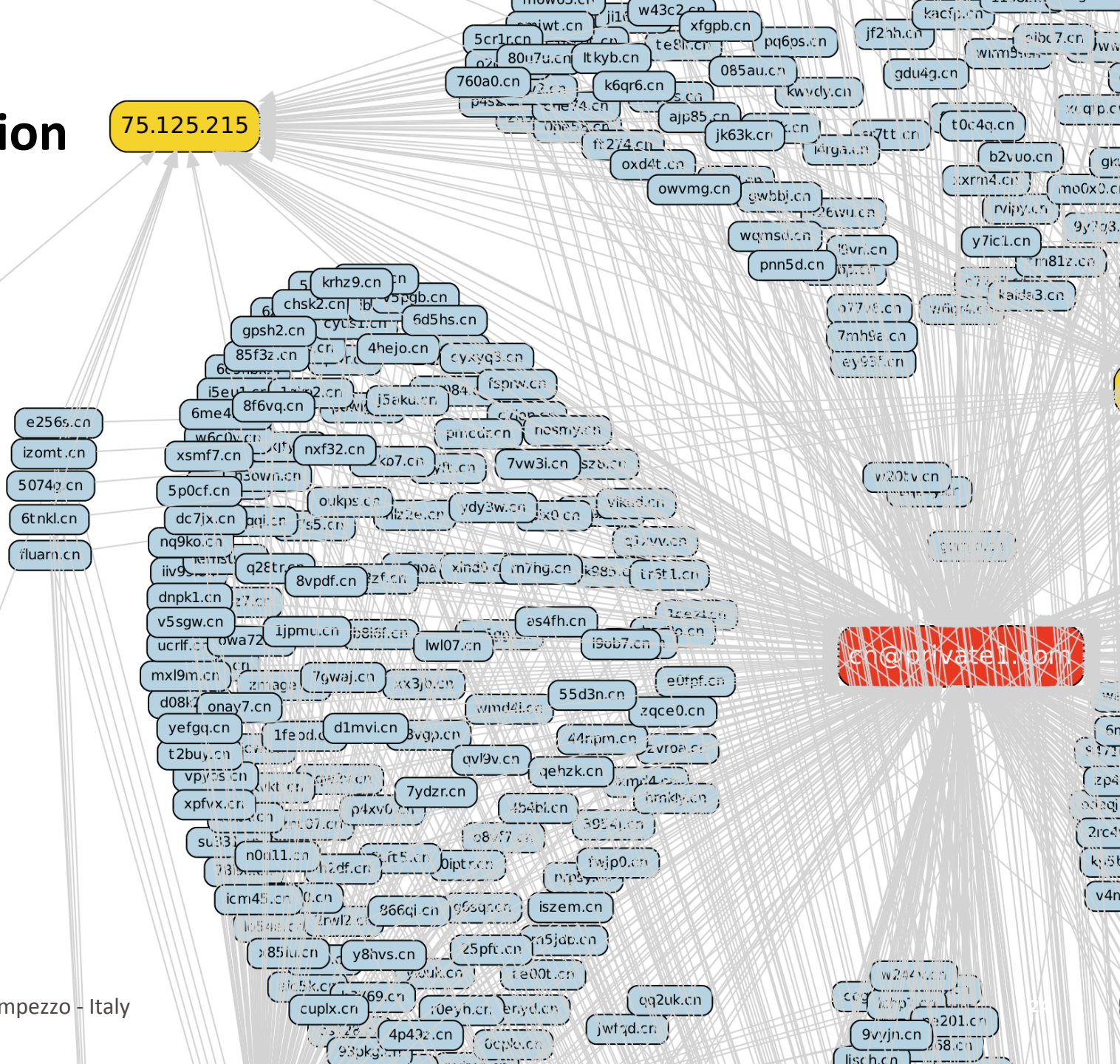
PC Antispyware



Level of coordination



Level of coordination



Are these findings specific to the threat landscape?

- Experiment: drive by downloads
  - Analysis of 5304 domains known to be landing pages for Internet Explorer ADODB.Stream Object Installation Weakness (CVE-2006-003)
  - Repeated feature collection and analysis using MCDA
- **Only 21 clusters** were found accounting for a total of 201 domains (3.8%)
  - The domains under analysis **do not** share a common infrastructure
  - The infrastructure is not actually owned by the perpetuators of the attacks
  - Important **difference** with the Rogue AV scenario
- How to justify this difference?

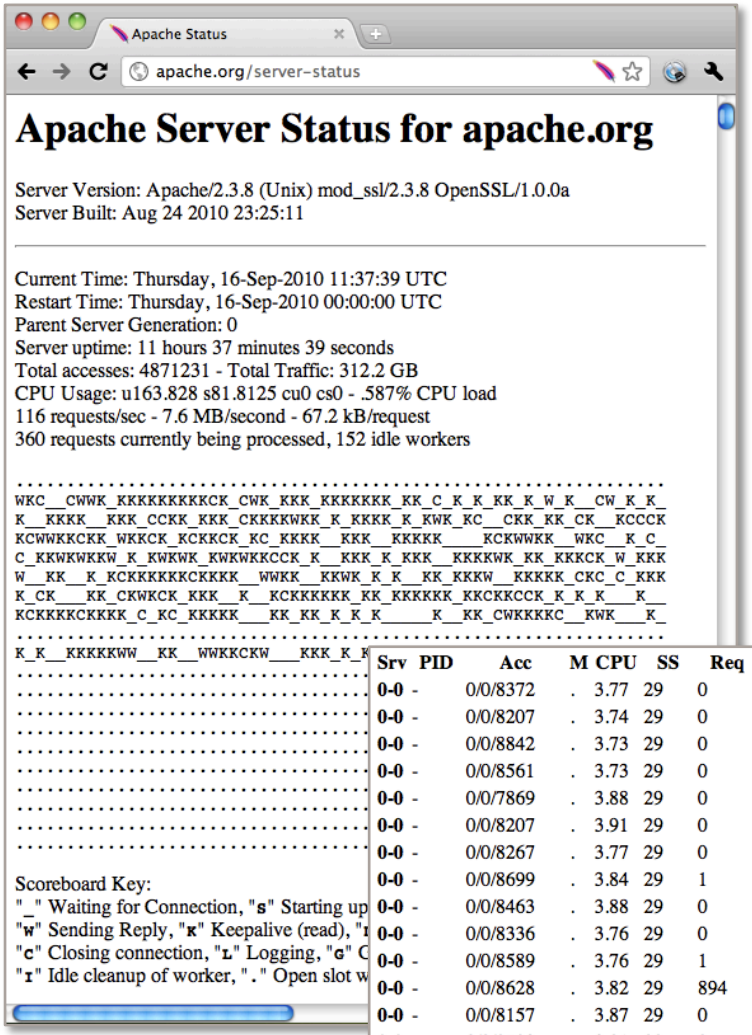
Rogue AV economics

- What are the costs/revenues associated to the rogue AV business?

- **Costs (informal survey)**
  - Average monthly cost: 50$
  - Annual domain registration costs: 3-10$
  - **Total annual costs: 879-2,230$**

- **Revenues**
  - Average price for a rogue AV: 30-50$
  - Client volume? **??**
  - **Total annual revenues: ??**

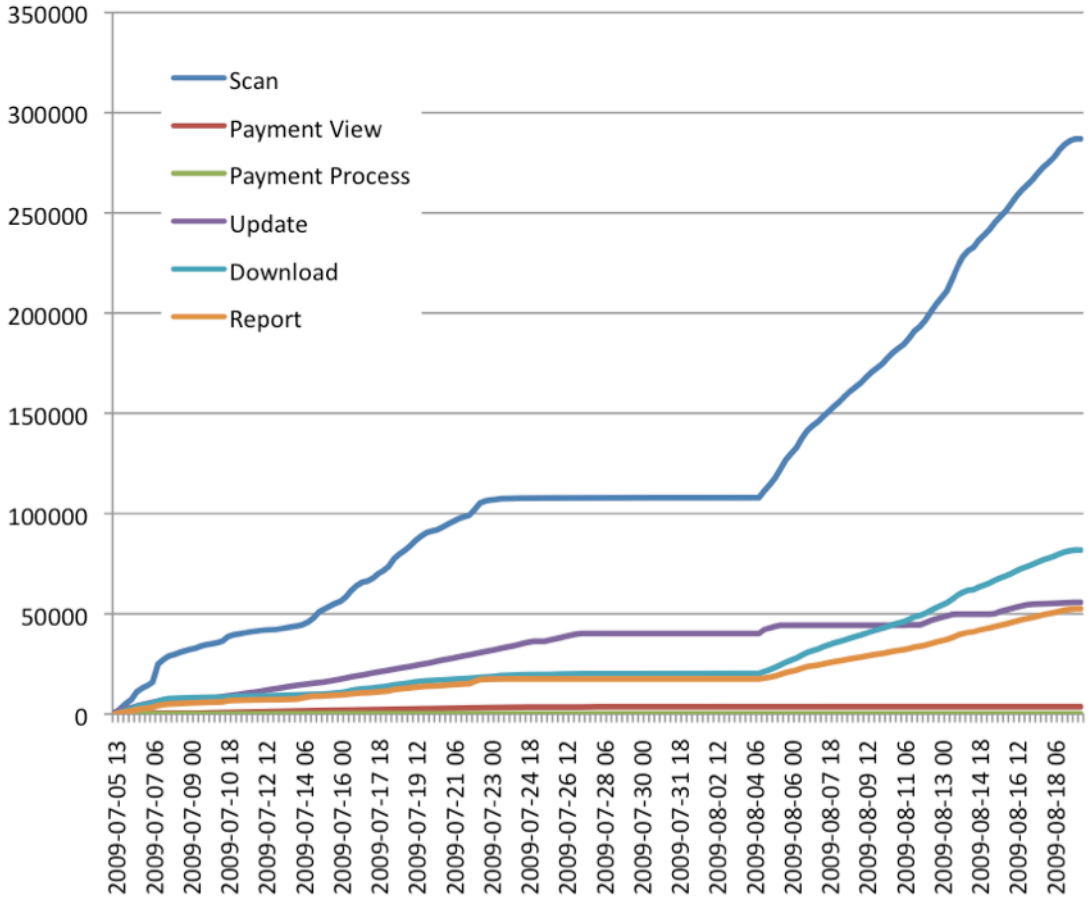
Rogue AV servers and Apache mod_status



Srv	PID	Acc	M	CPU	SS	Req	Conn	Child	Slot	Client	VHost	Request
0-0	-	0/0/8372	.	3.77	29	0	0.0	0.00	432.83	110.234.86.178	www.apache.org	NULL
0-0	-	0/0/8207	.	3.74	29	0	0.0	0.00	667.30	120.35.61.133	www.apache.org	NULL
0-0	-	0/0/8842	.	3.73	29	0	0.0	0.00	412.13	77.104.52.78	www.apache.org	NULL
0-0	-	0/0/8561	.	3.73	29	0	0.0	0.00	517.30	15.211.169.106	www.apache.org	NULL
0-0	-	0/0/7869	.	3.88	29	0	0.0	0.00	639.49	80.118.162.126	www.apache.org	NULL
0-0	-	0/0/8207	.	3.91	29	0	0.0	0.00	423.27	121.240.102.50	jakarta.apache.org	GET /jmeter/images/screenshots/mirrorserver.png HTTP/1.0
0-0	-	0/0/8267	.	3.77	29	0	0.0	0.00	473.82	61.238.159.101	www.apache.org	NULL
0-0	-	0/0/8699	.	3.84	29	1	0.0	0.00	618.55	62.77.176.130	www.apache.org	NULL
0-0	-	0/0/8463	.	3.88	29	0	0.0	0.00	597.82	120.35.61.133	www.apache.org	GET /dist/commons/dbutils/binaries/commons-dbutils-1.1.tar.gz H
0-0	-	0/0/8336	.	3.76	29	0	0.0	0.00	448.69	199.123.79.97	www.apache.org	NULL
0-0	-	0/0/8589	.	3.76	29	1	0.0	0.00	767.45	58.34.18.75	www.apache.org	NULL
0-0	-	0/0/8628	.	3.82	29	894	0.0	0.00	444.02	203.187.186.130	www.apache.org	NULL
0-0	-	0/0/8157	.	3.87	29	0	0.0	0.00	348.97	85.90.221.125	www.apache.org	NULL
0-0	-	0/0/8723	.	3.91	29	0	0.0	0.00	459.49	122.212.220.146	tomcat.apache.org	HEAD /dev/dist/m2-repository//javax/mail/mail/1.4/mail-1.4.jar

- 6 servers (193 domains) were discovered to be offering utilization statistics through the output of Apache mod_status
  - Continuous sampling of the output over a period of 44 days
  - Filtered out probing/scanning attempts
  - Tracked a total of 372,096 distinct IP addresses

Behavior evolution



Cumulative number of distinct IP addresses for each behavior type

Successful scans: 25,447

Unsuccessful scans: 306,248

Hit rate: 7.7%

A scan is considered successful if a download is performed by the same IP address within 24 hours

Completing the table

- What are the costs/revenues associated to the rogue AV business?

- **Costs (informal survey)**
  - Average monthly cost: 50$
  - Annual domain registration costs: 3-10$
  - **Total annual costs: 879-2,230$**

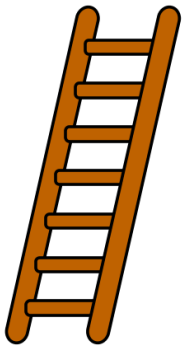
- **Revenues (pessimistic estimate)**
  - Average price for a rogue AV: 30-50$
  - Expected monetization rate for client hit: 0.26% (in previous studies on spam)
  - Client volume over 44 days: 331,695
  - **Total annual revenues: 214,621-357,702$**

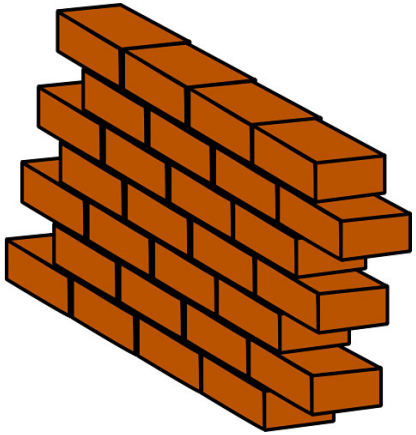
All a matter of economics

- **Question:** should we worry about threats targeting an environment that has the following characteristics?
  - It is isolated from the Internet
  - It runs non-standard components that use proprietary protocols
  - It warns the user whenever a signed driver is introduced in the kernel

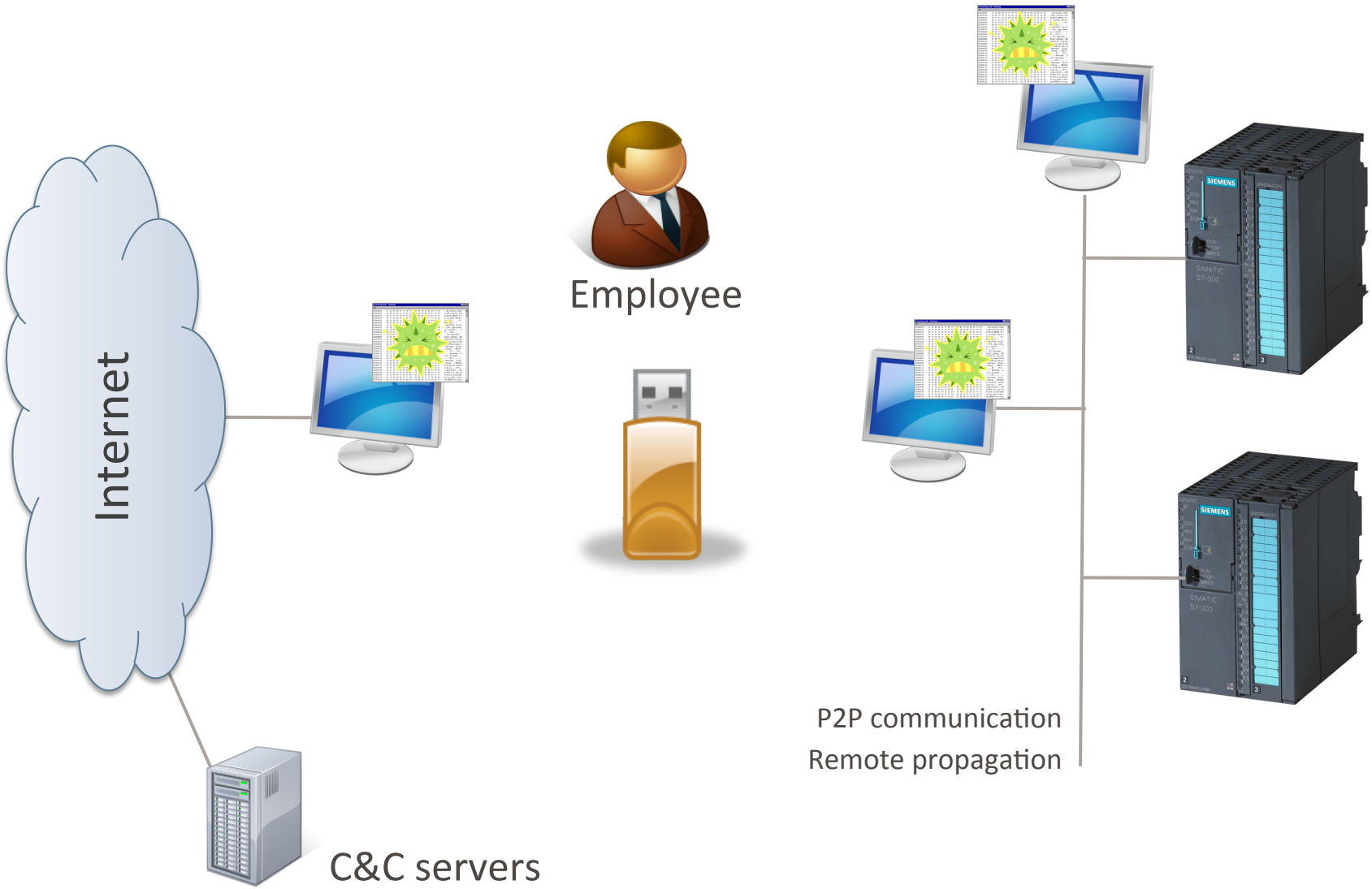
All a matter of economics

- **Question:** should we worry about threats targeting an environment that has the following characteristics?
  - It is isolated from the Internet
  - It runs non-standard components that use proprietary protocols
  - It warns the user whenever a signed driver is introduced in the kernel
- **Answer**: it depends!





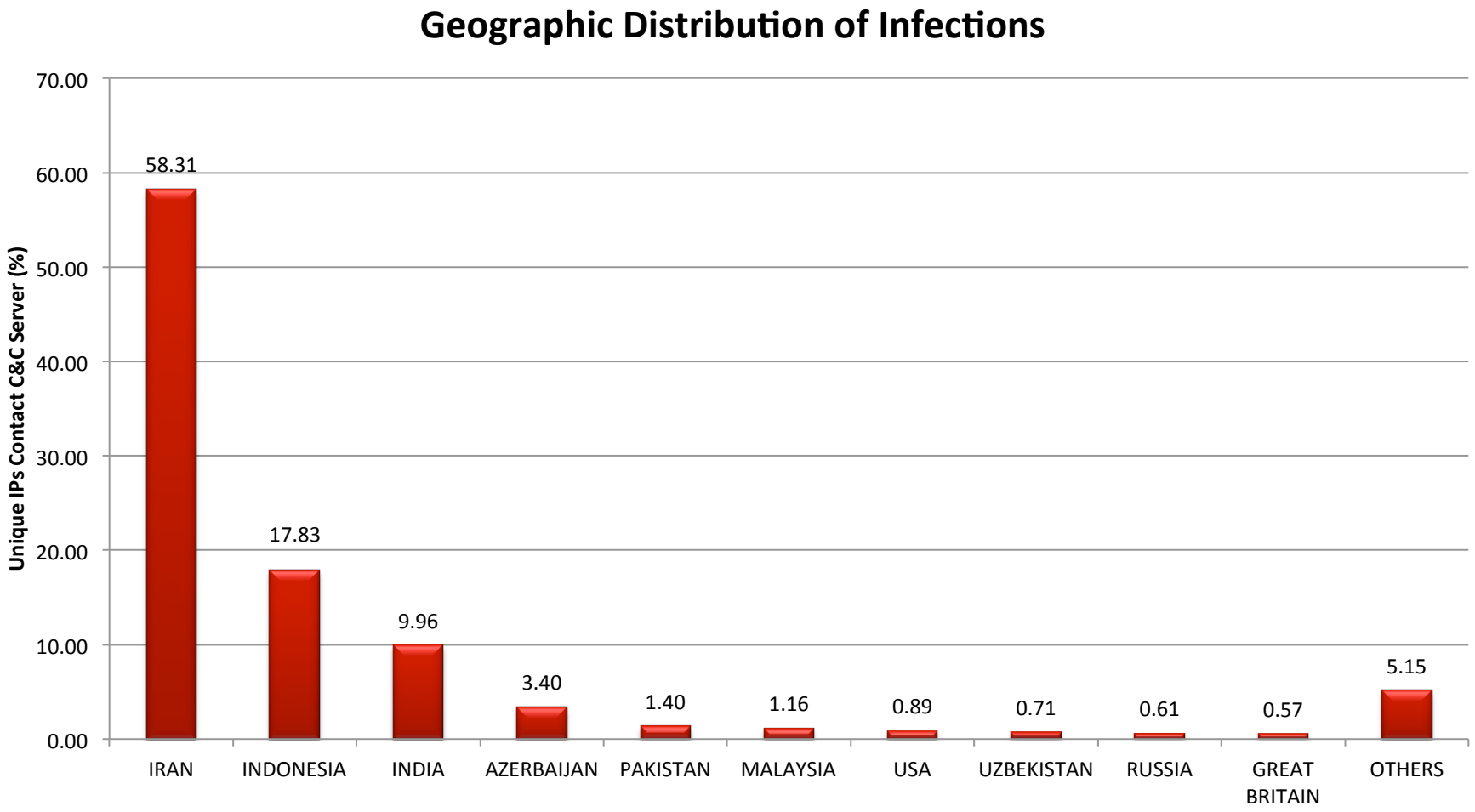
Stuxnet


Internet
Employee
P2P communication
Remote propagation
C&C servers

Stuxnet

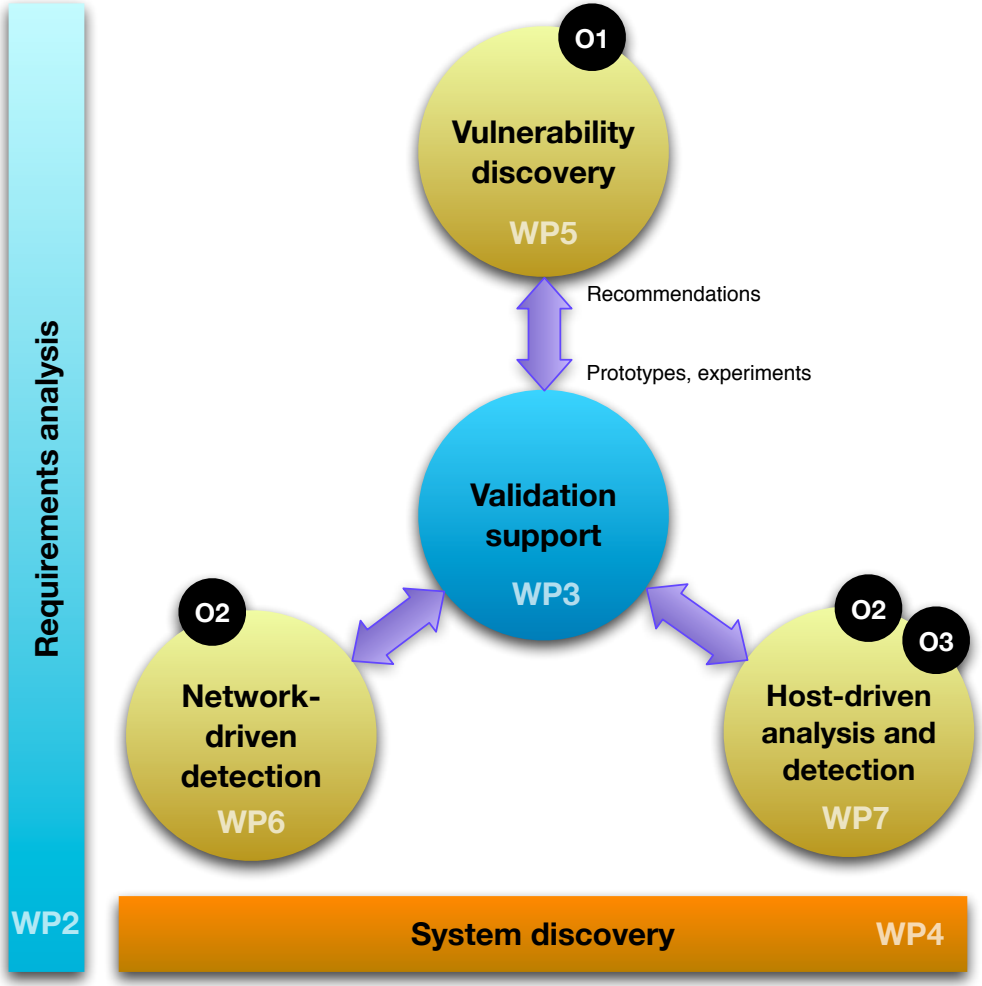
- Attacks industrial control systems
- Spreads by copying itself to USB drives
  - LNK vulnerability
  - Autorun.inf
- Spreads via network shares
- Spreads using 2 known and 4 0-day Microsoft vulnerabilities
- Injects rootkit by means of a signed driver (signature stolen from legitimate vendor)
- Injects STL code into Siemens PLCs
- Communicates with C&C servers using HTTP
  - www.mypremierfutbol.com
  - www.todaysfutbol.com
- Steals designs documents for industrial control systems
- Sabotages targeted industrial control systems

Stuxnet



Over 40,000 infected unique external IPs, from over 115 countries

CRISALIS (FP7 SEC)



CRISALIS objectives:

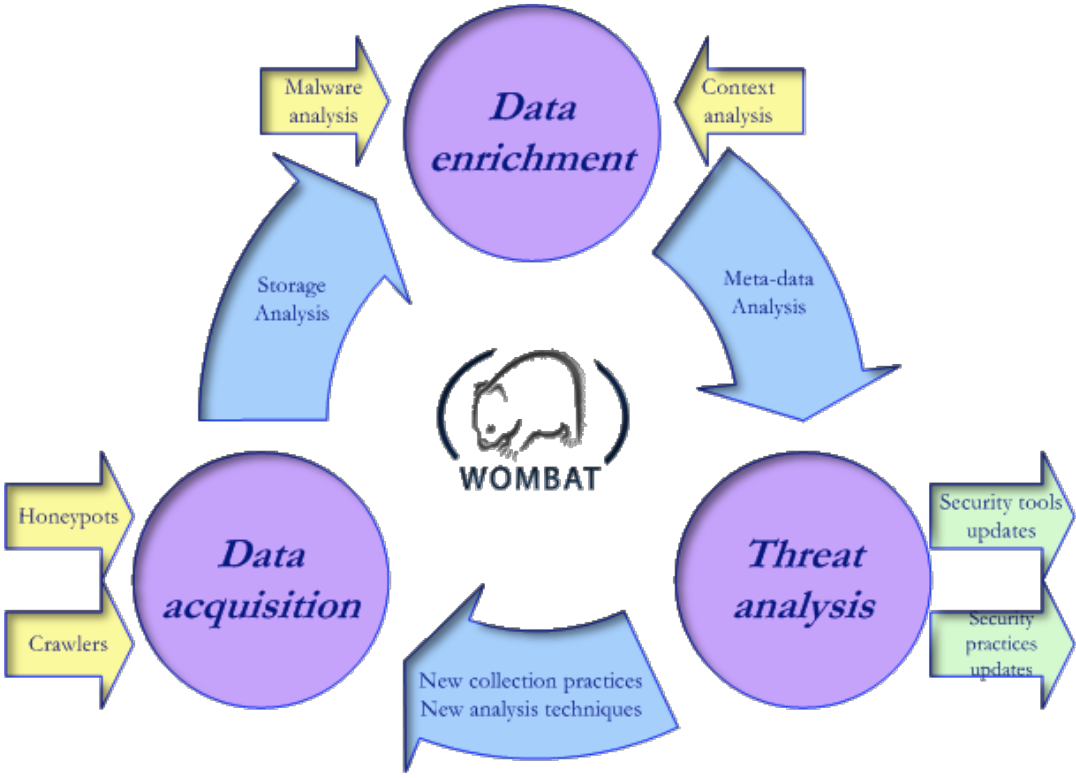
- **O1** Securing the systems
- **O2** Detecting intrusions
- **O3** Analyzing successful intrusions

How?

It’s not as easy as it seems

A continuously moving target

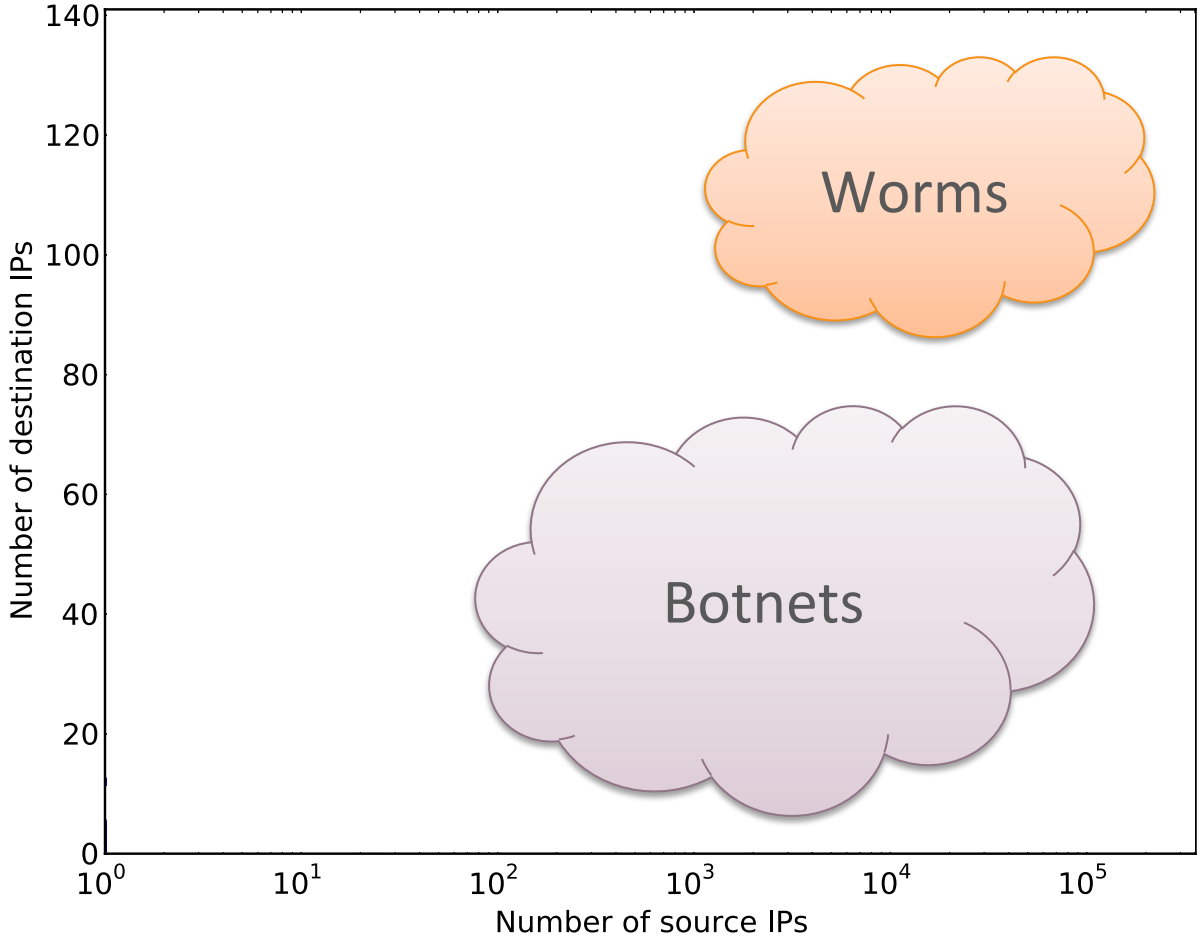
- The threat landscape is in continuous evolution
  - We can't take a static picture
- Our assumptions on its characteristics influence the data acquisition methodology
  - We **WILL** need to revisit our initial assumptions!



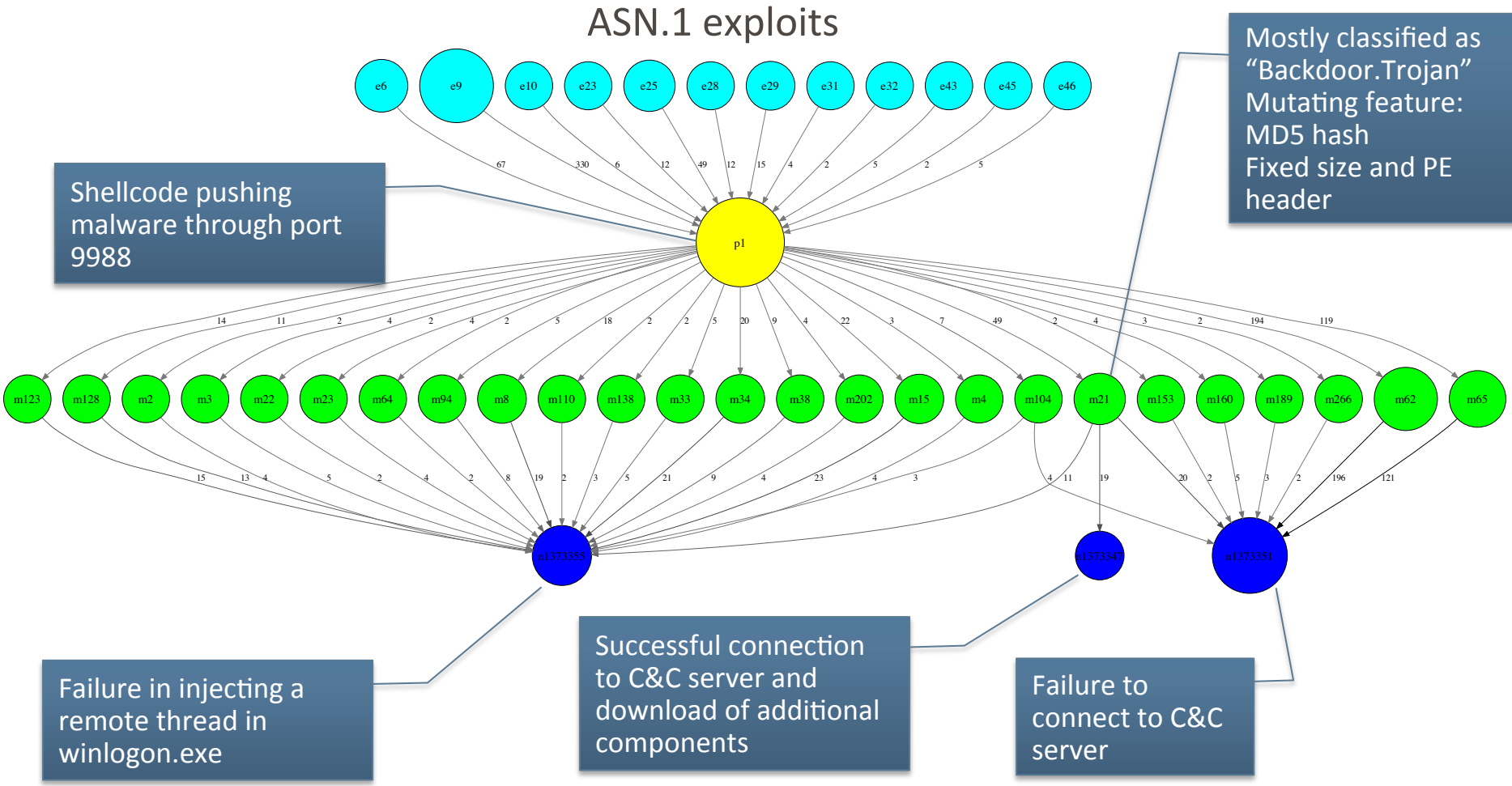
http://www.wombat-project.eu

Belief #1: Internet malicious activity is uniform throughout the IP space

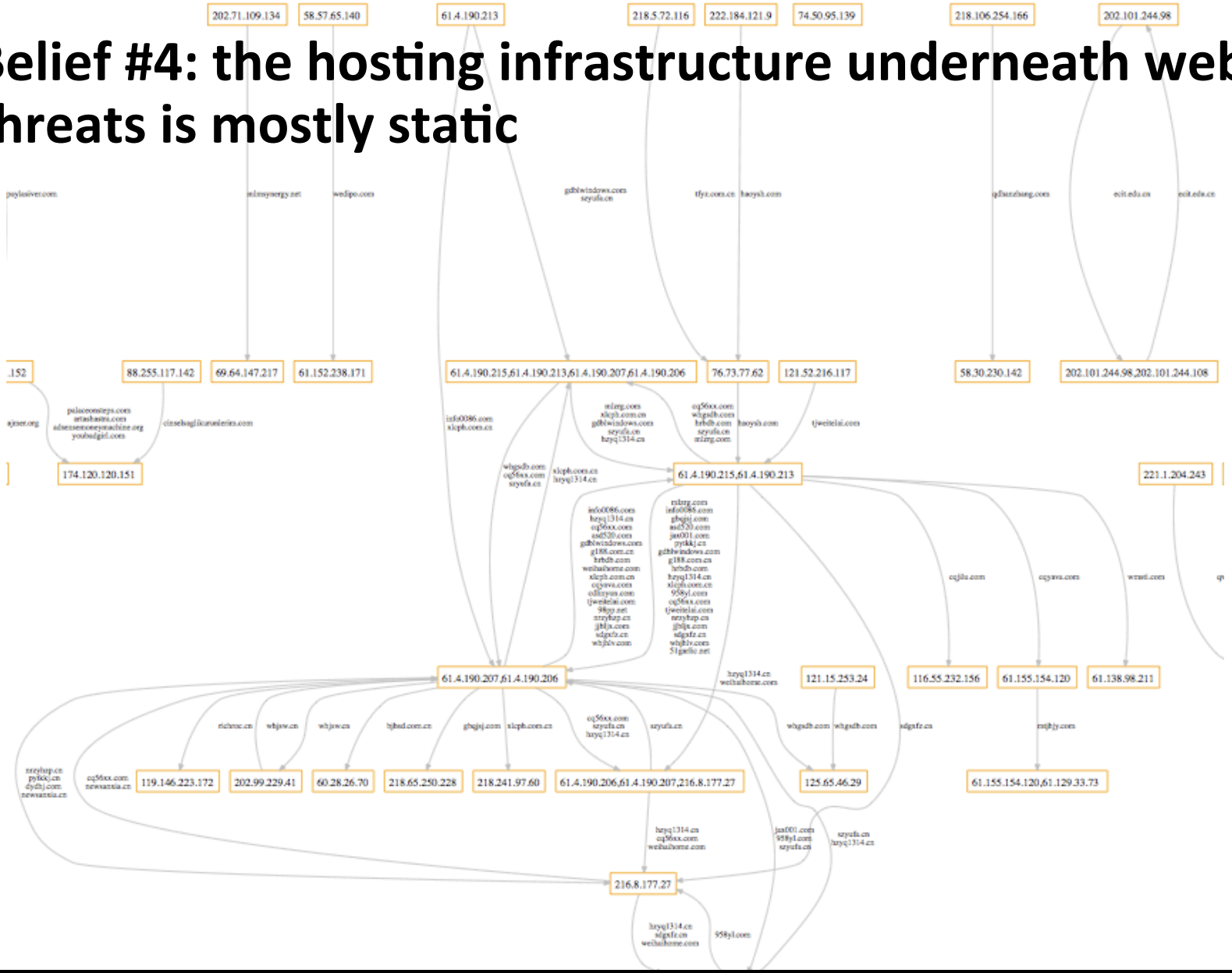
Belief #2: Internet exploits are only associated to large worm infections and botnets



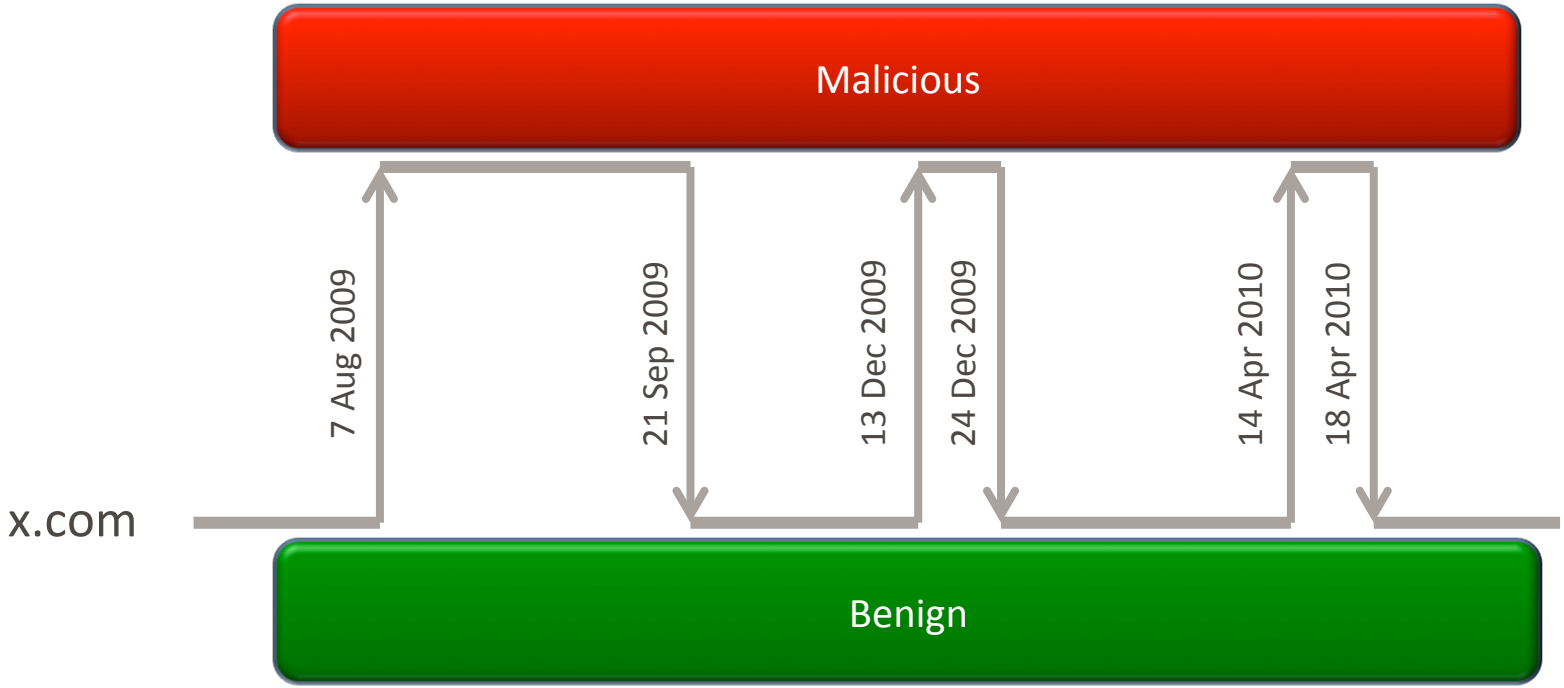
Belief #3: only dynamic analysis can help us in dealing with polymorphism



Belief #4: the hosting infrastructure underneath web threats is mostly static



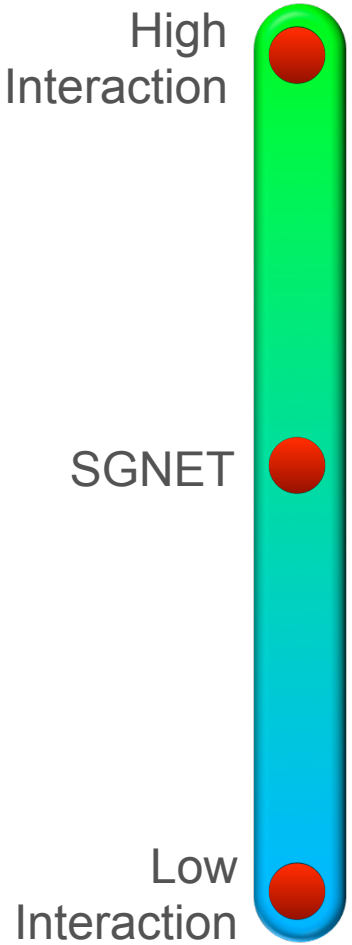
Belief #5: a domain/URL is either “green” or “red”



WOMBAT datasets

- SGNET: distributed honeypot deployment based on protocol learning
  - What are the trends and the characteristics of code injection attacks used for malware propagation?
  - No assumption on the exploit characteristics
- HARMUR: Historical ARchive of Malicious URLs
  - Information on the infrastructure and the dynamics underlying web threats
  - Data aggregator using multiple network and security feeds

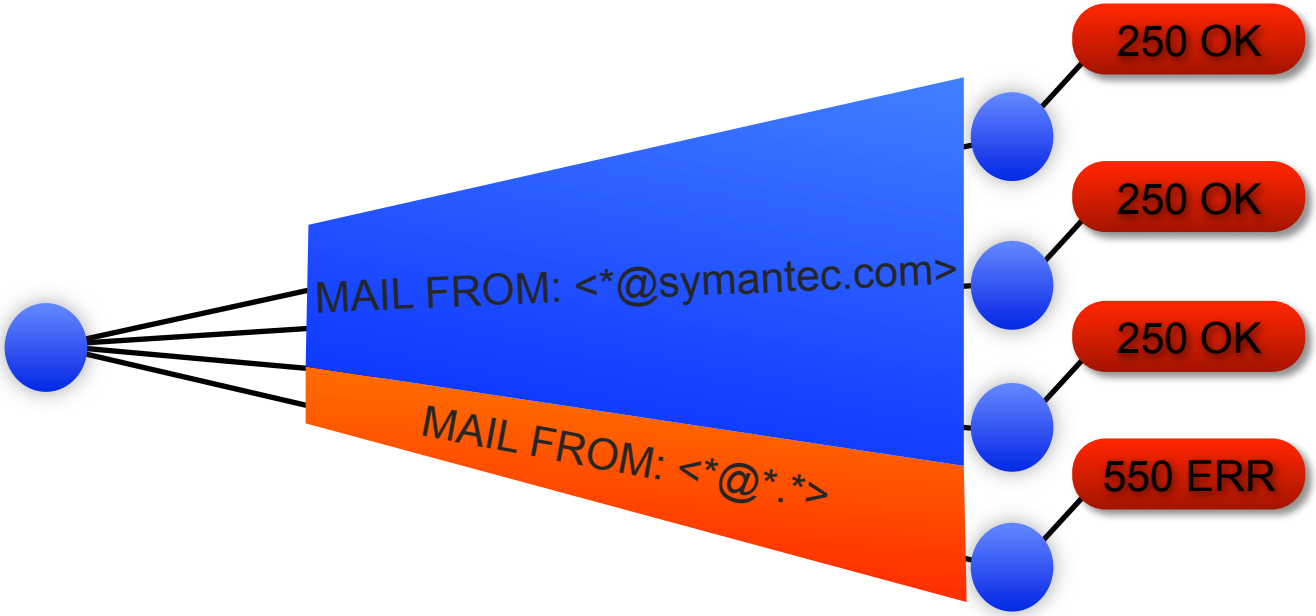
Emulating protocols in honeypots: the problem



- Need to increase level of interaction
  - Required to retrieve information on the root cause of the observed activities
- Need to minimize the cost of the sensors
  - Implicit requirement of a distributed deployment of sensors hosted by volunteering partners

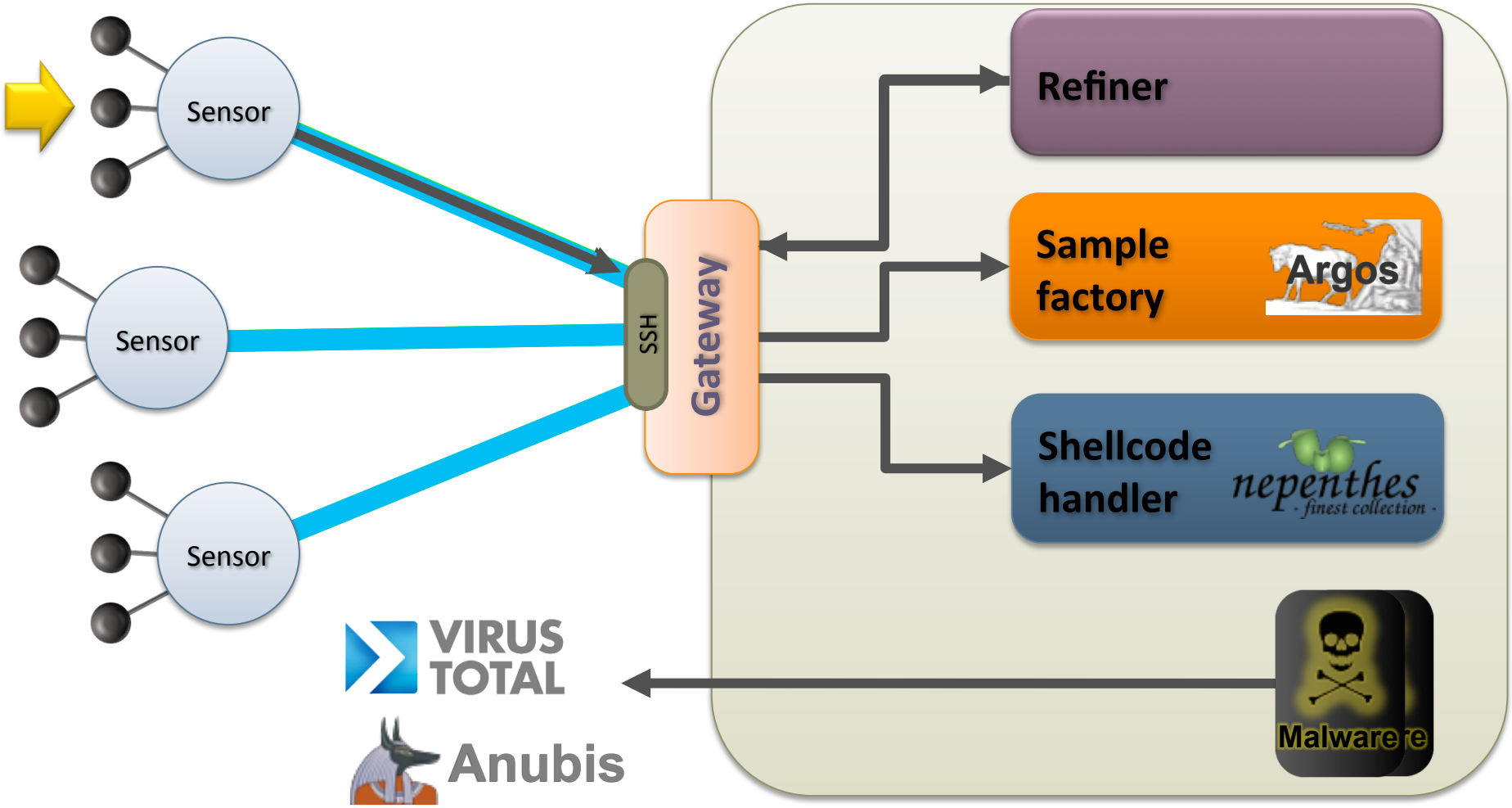
ScriptGen

- Protocol-agnostic algorithm
- Observe conversation samples between a client and a real server
- Infer semantics using bioinformatics algorithms
- Proved good results in handling deterministic exploit scripts

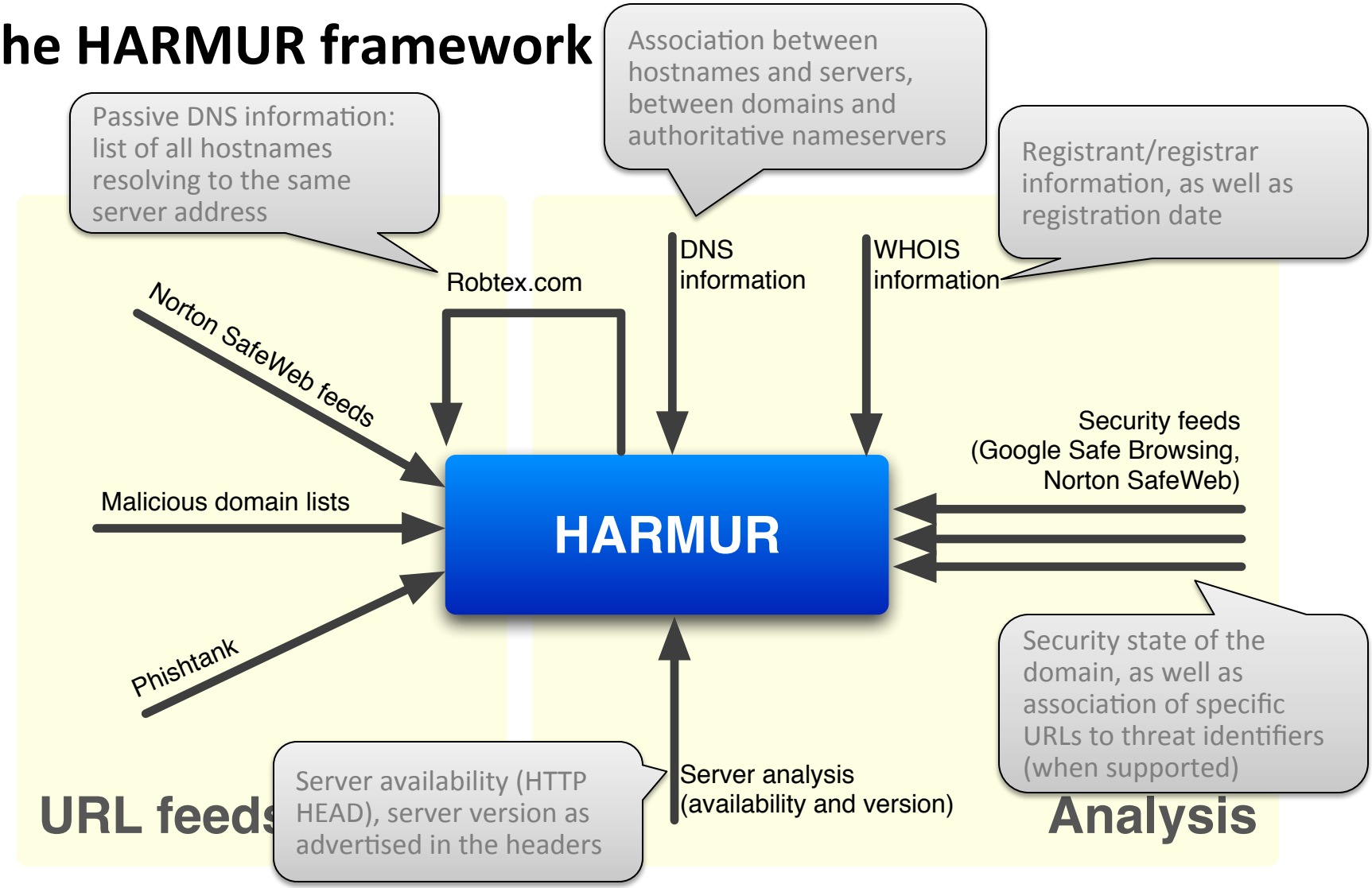


SGNET operation

- Normal operation
- New exploit encountered
- Global update of the learning
- Submission of a shellcode sample
- Analyze new malware sample



The HARMUR framework



Where?

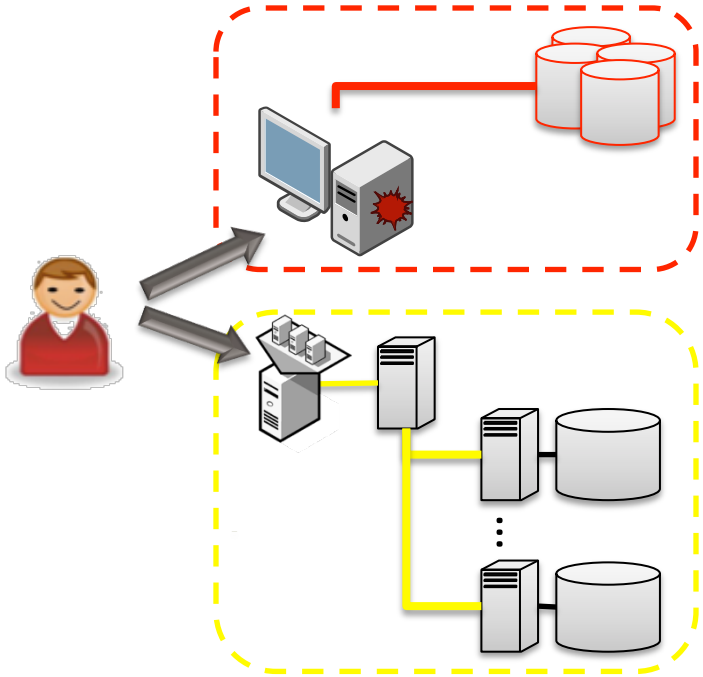
The WINE program

Challenges

- Rigorous data collection is essential to security research
  - Understanding the threat landscape
  - Understanding the threat economics
  - Evaluating/benchmarking real-world application of specific solutions
- It's an expensive process
  - Highly dynamic threat landscape
    - Need to ensure representativeness of the observations, but also repeatability
  - It's an iterative process
- The effort cannot be easily shared
  - Field data experimentation is associated to lots of legal and ethical concerns when it comes to sharing data

WINE: Benchmark for Computer Security

http://www.symantec.com/WINE



Symantec's
worldwide sensors

Platform for
experimental reproducibility

The Worldwide Intelligence Network Environment (WINE)

- Goal: *repeatable cyber security experiments at scale*
- Field data collected on millions of *end-hosts*
- Data sampled from Symantec's *operational data* sets
- Access WINE on SRL site: *Culver City, CA* or *Herndon, VA*
  - Fee required
- Store *reference data* sets used in prior experiments
- Maintain *lab book*

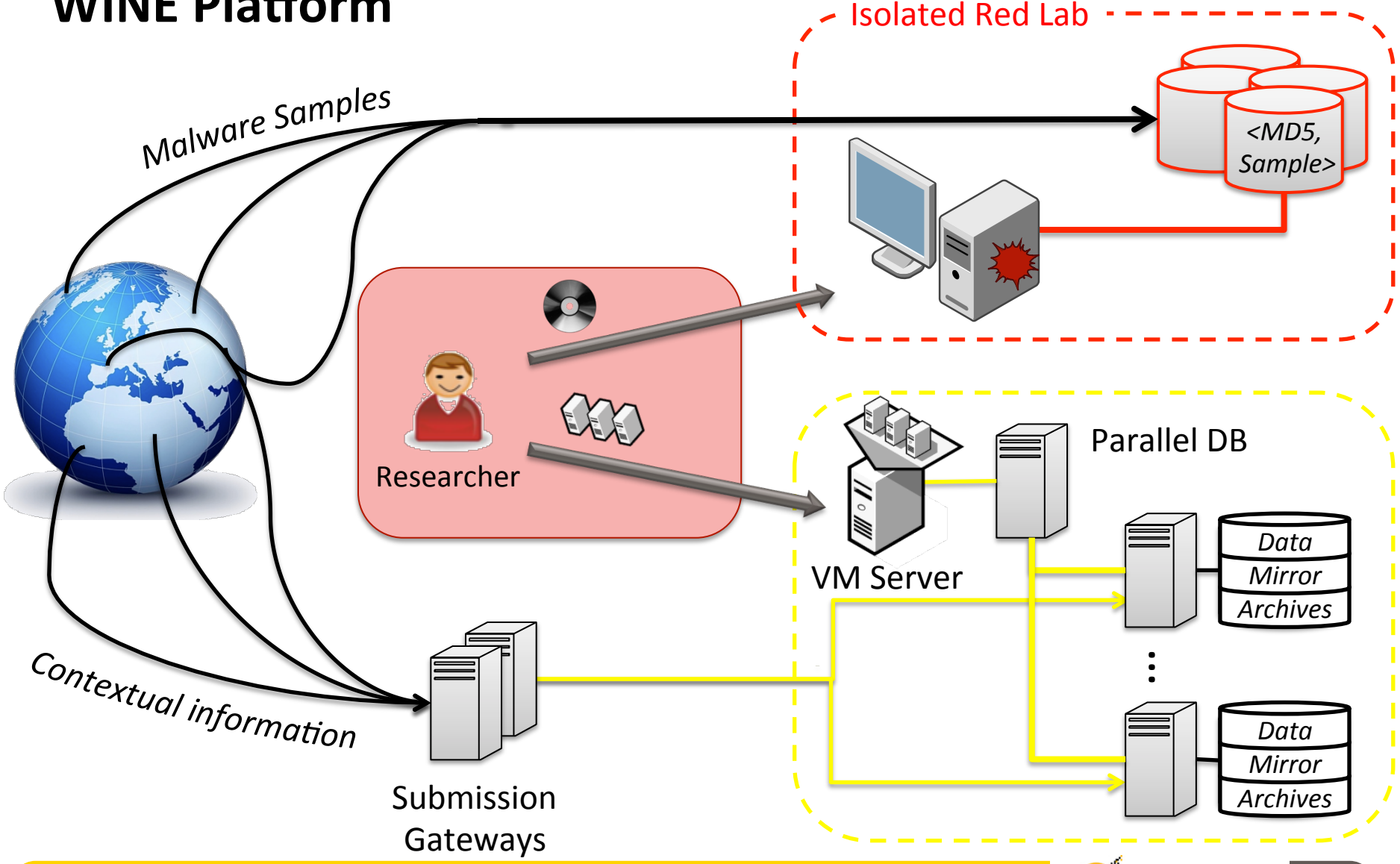
WINE Data

- Sampled field data, representative for what Symantec collects
  - Up to 20 TB
  - Over 1M end-hosts
  - Goes back to 2008

- Five data sets, initially:
  - Malware samples
  - Binary reputation (file downloads)
  - A/V and IPS telemetry
  - URL reputation
  - Spam

More data, in the future

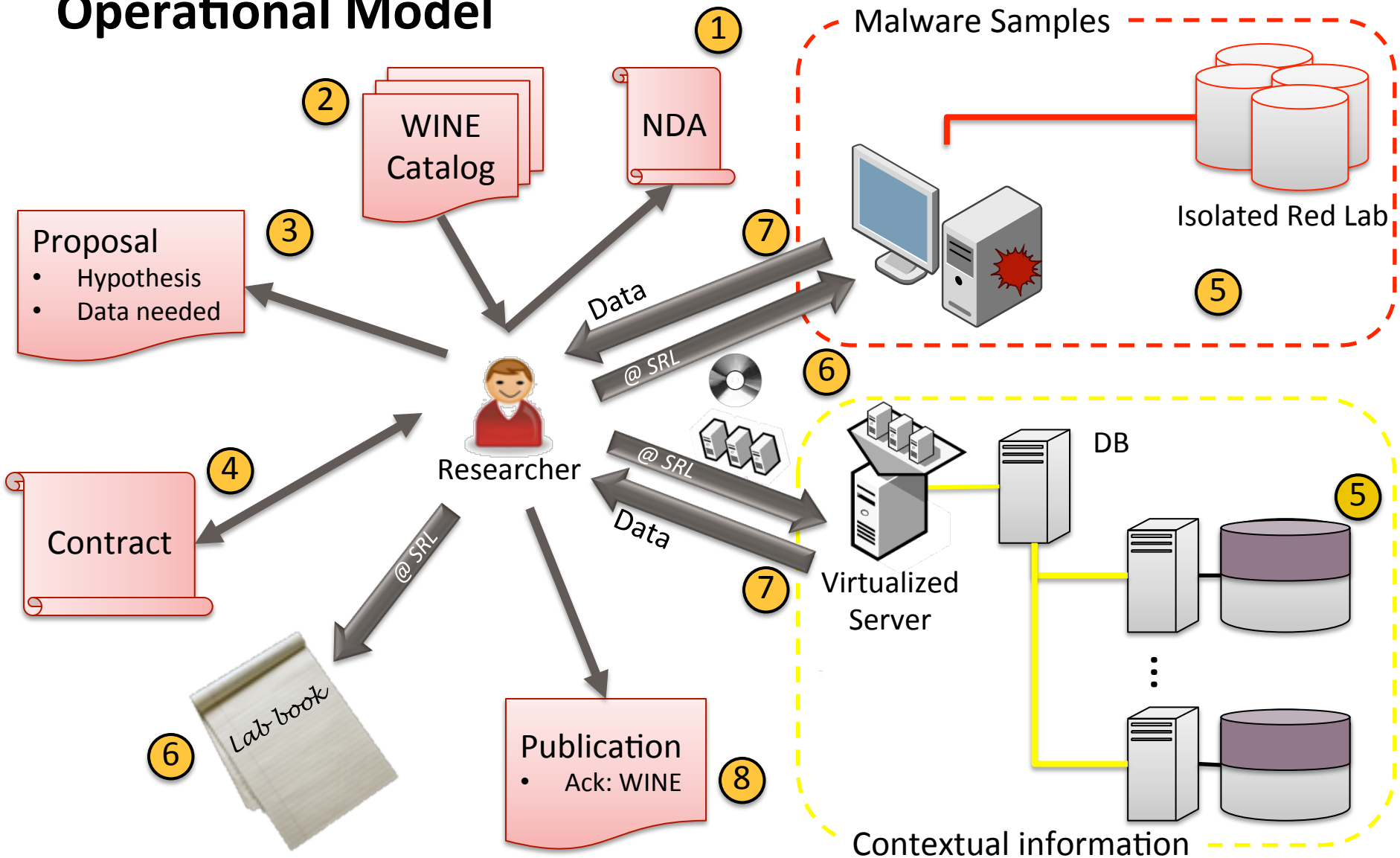
WINE Platform



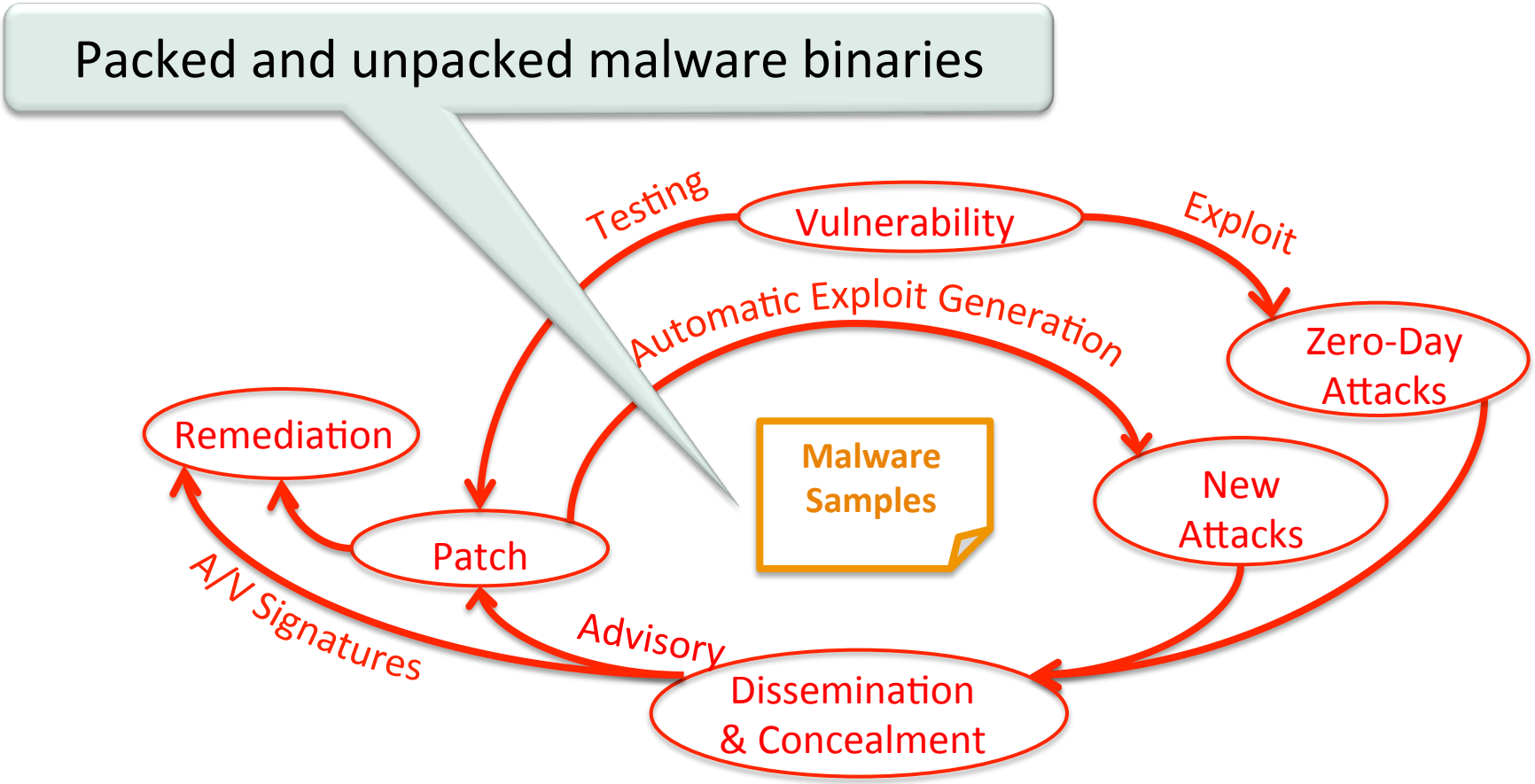
What WINE is not …

- … a definitive benchmark suite
- … a data set that can be copied outside of SRL
- … a system that can be accessed remotely
- … a repository for all the data that Symantec collects
- … an effort targeted exclusively at cyber security

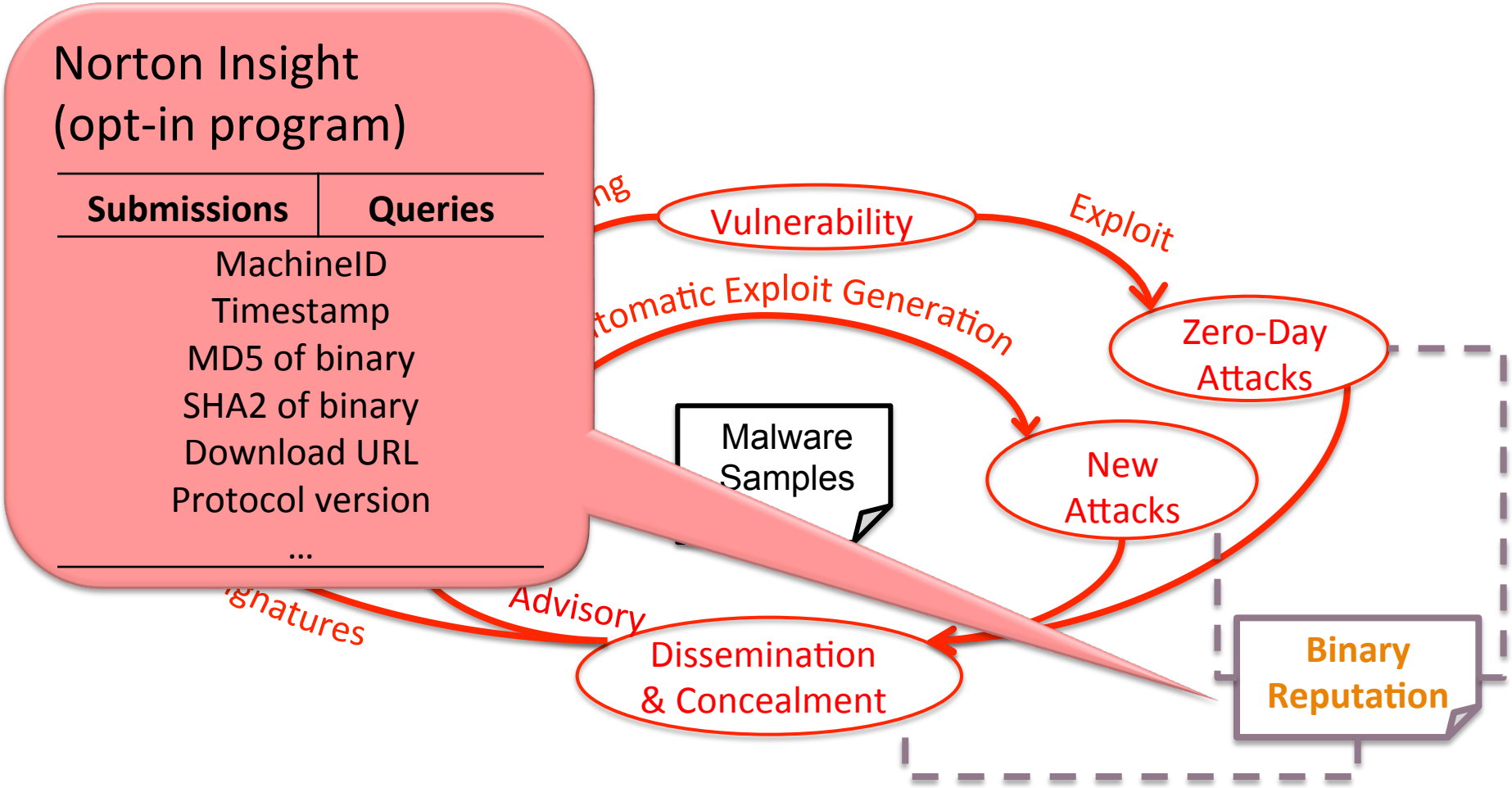
Operational Model



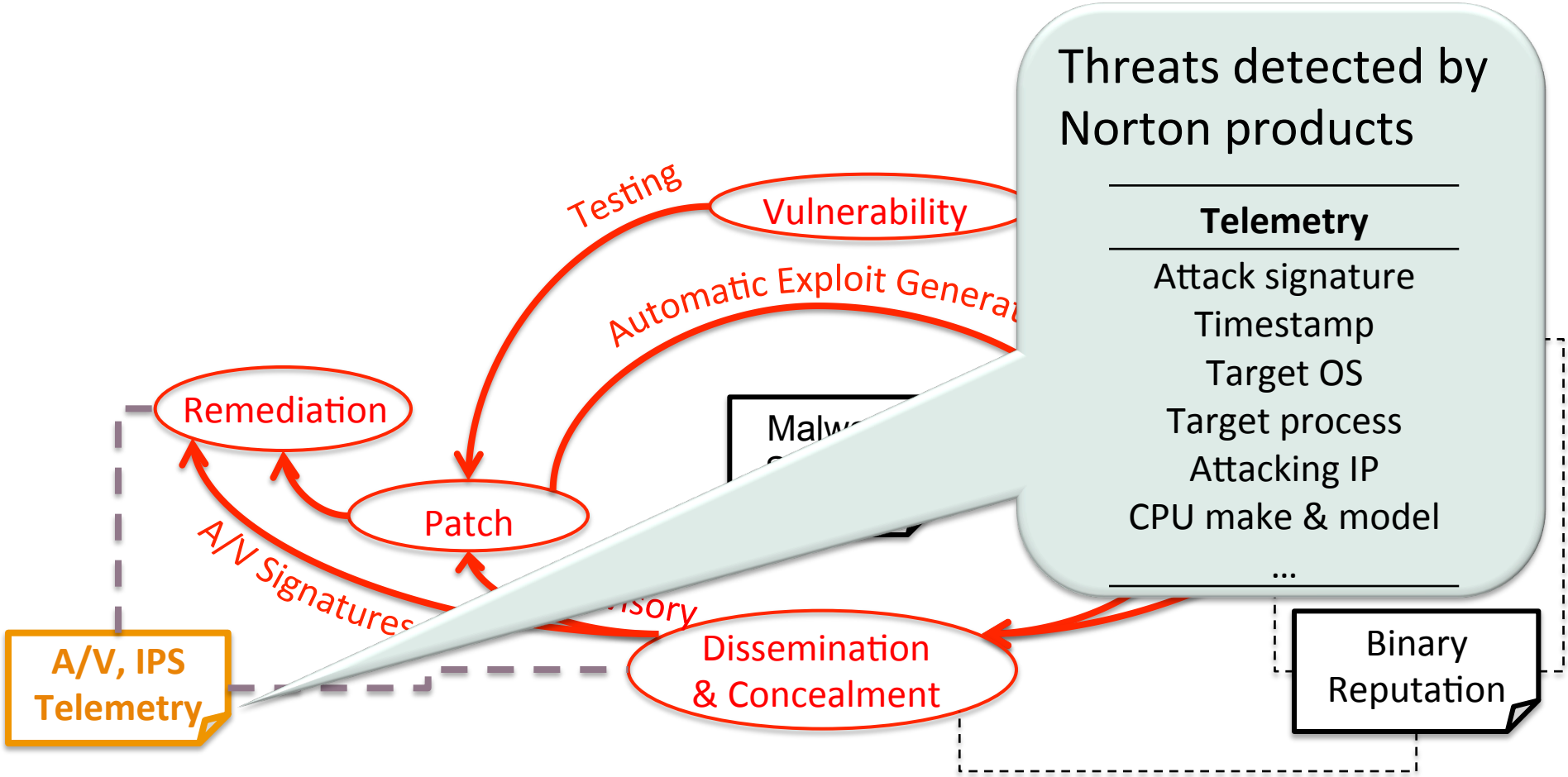
WINE Data Set: *Malware*



WINE Data Set: *Binary Reputation*

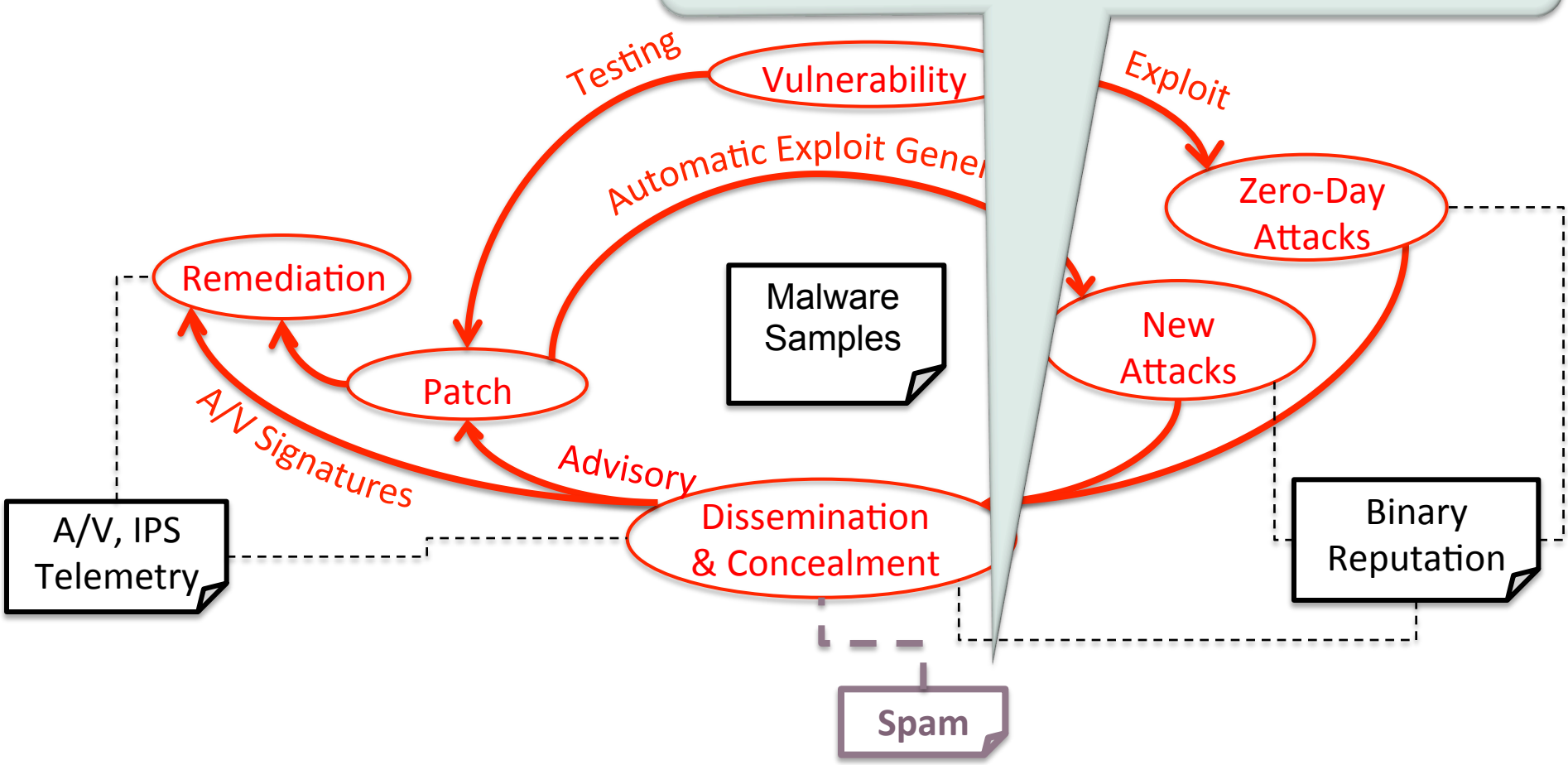


WINE Data Set: *A/V & IPS Telemetry*



WINE Data Set: *Spam*

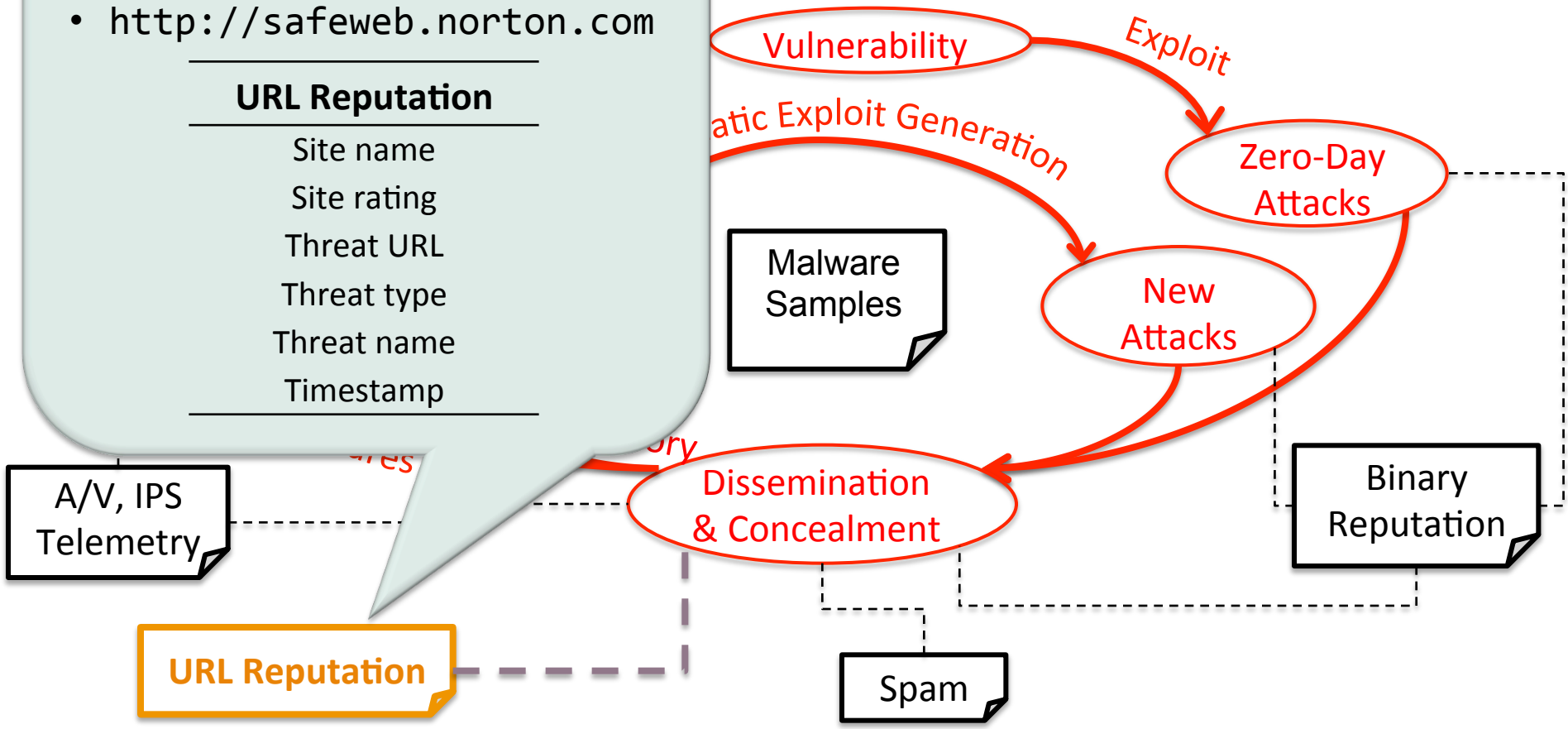
- Samples of spam and phishing emails
- Statistics on blocked spam



WINE Data Set: *URL Reputation*

- Data collected by crawling the Web
- `http://safeweb.norton.com`

URL Reputation
Site name
Site rating
Threat URL
Threat type
Threat name
Timestamp



Distributed Data Collection

Malware:
7M samples

A/V telemetry:
130M machines

URL reputation:
10M domains

Binary reputation:
35M machines

Spam: 2.5M decoys

Graduate fellowship program

- Interested in doing an internship with us?
  - http://www.symantec.com/about/careers/college/fellowship.jsp
- Call for applications: every year around January
- Eligibility criteria:
  - PhD or Master's program focused on technology research in Europe or U.S.
  - Preference to students with a desire to work in an industrial research lab, and working on research projects of real-world practical value to customers
  - Selection according to the overall potential to academic excellence and their academic progress to-date
  - The scholarship awards will be made through the university and are not transferable to another academic institution.

Thank you!

WINE

Can be used without moderation

www.symantec.com/WINE

corrado_leita@symantec.com