Avoiding Information Overload: Automated Data Processing with n6

Paweł Pawliński
pawel.pawlinski@cert.pl

CERT.PL

26th annual FIRST conference

Boston, June 23rd 2014

Who we are

- part of 
- national CERT for Poland

Goals

- Mission statement:
  (...) assist Polish internet users in implementing proactive measures to reduce the risks of computer security incidents and to assist them in responding to such incidents when they occur. (...)
- Some common objectives:
  - reduce number of infections
  - detect attacks quicker
  - make systems more difficult to attack
- Use information!
  - attacks
  - victims
  - vectors
  - ...

Buzzwords?

information

Buzzwords?

security

information

Buzzwords?

security

information

intelligence

Buzzwords?

security

cyber **threat** **information**

intelligence

Buzzwords?

security

cyber **threat** **information** **feed**

global **intelligence**

Buzzwords?

internet **security** **data** **platform**

cyber **threat** **information** **feed**

global **actionable** **intelligence** **solution**

real-time

Buzzwords?

internet	**security**	**data**	**platform**
cyber	**threat**	**information**	**feed**
global	**actionable**	**intelligence**	**solution**
real-time	**big data**	**analytics**	**™**

Buzzwords?

internet	**security**	**data**	**platform**
cyber	**threat**	**information**	**feed**
global	**actionable**	**intelligence**	**solution**
real-time	**big data**	**analytics**	**™**

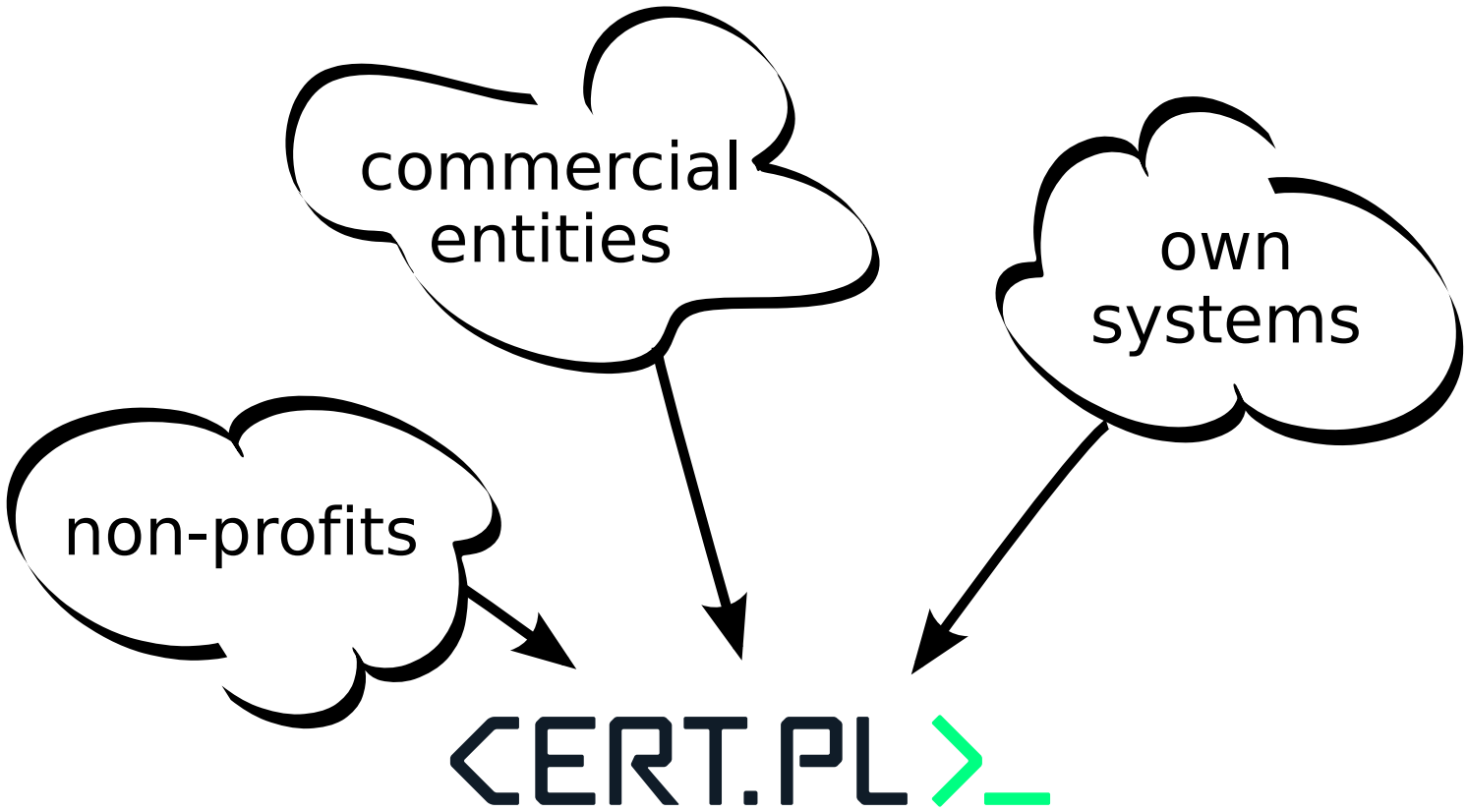
Related efforts

- ENISA report
  - *Good practice guide for the exchange and processing of actionable information by CERTs*
  - expected publication: end of 2014
- NECOMA
  - Nippon-European Cyberdefense-Oriented Multilayer threat Analysis
  - ongoing research
  - Consortium: Institut Mines-Telecom, Atos, FORTH, NASK, 6cure, Nara Institute of Science and Technology, Internet Initiative Japan, National Institute of Informatics, Keio University, University of Tokyo
  - www.necoma-project.eu

What is actionable?

- *action* depends on the recipient
- Kill Chain
  - indicators for every stage
  - known attack properties
  - detect / block
- signal when investigation is needed

Year 2011 in automation



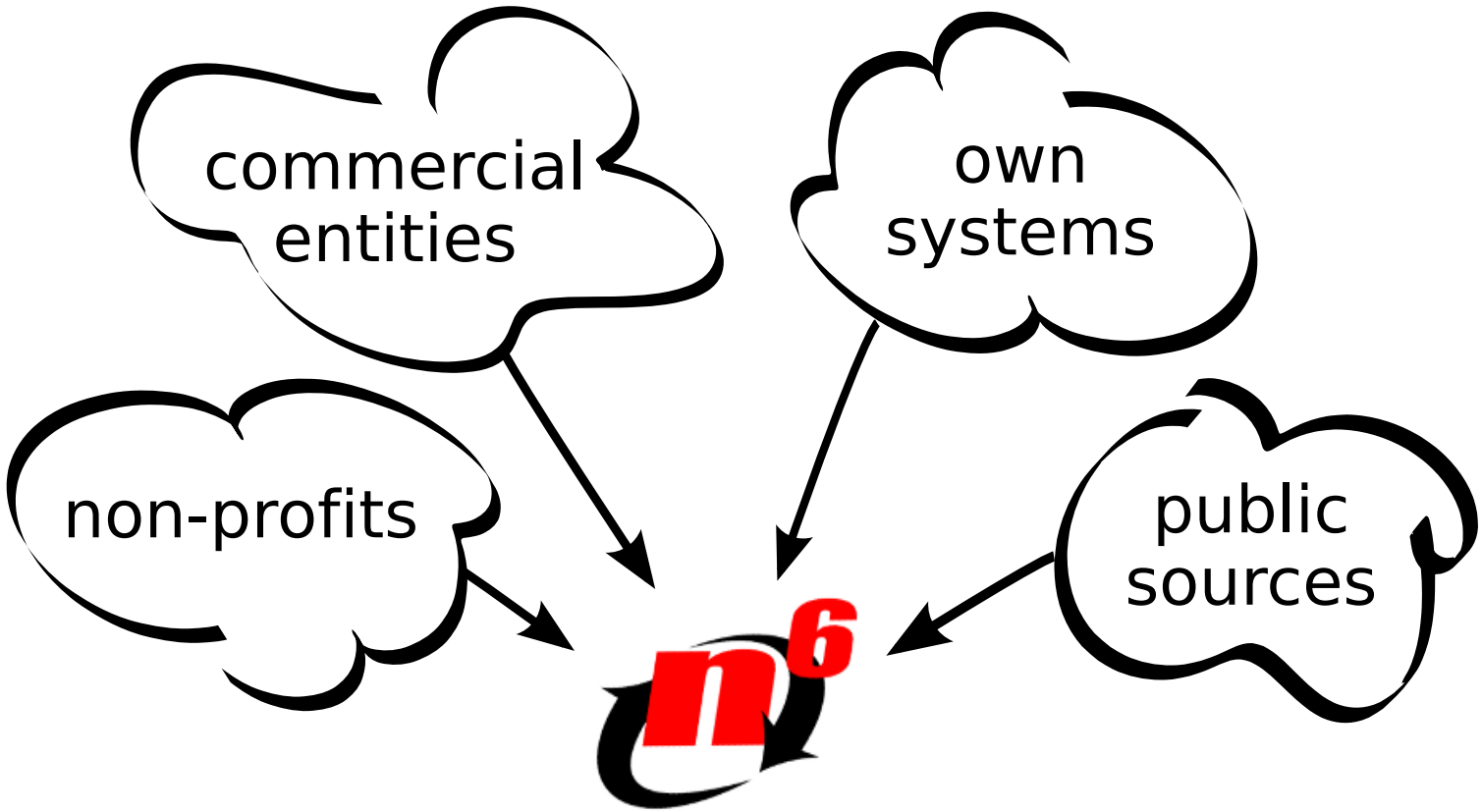
Year 2011 in automation

- Daily collecting 10MiB+, 100k+ records
- C&C, bots, scanners, spammers
- < 10 sources
- Multiple shell scripts
- Automated sharing with $\sim$ 30 external organizations
- Unmanageable
- Analyses in a shell-style: **awk . . . | perl -ne . . . > report.txt**

n6: first generation

- Main goals:
  1. get better value from data
  2. share with more entities
  3. add more sources
- Introducing **Network Security Incident eXchange**
- Technical objectives:
  - single channel for all automatic data exchange
  - keep the original input data for reference
  - enrich with information from BGP & DNS
  - asynchronous processing
  - high throughput: gigabytes, 10M+ events daily
  - access through HTTPS, authorized by client SSL certificates

Automated sources of information



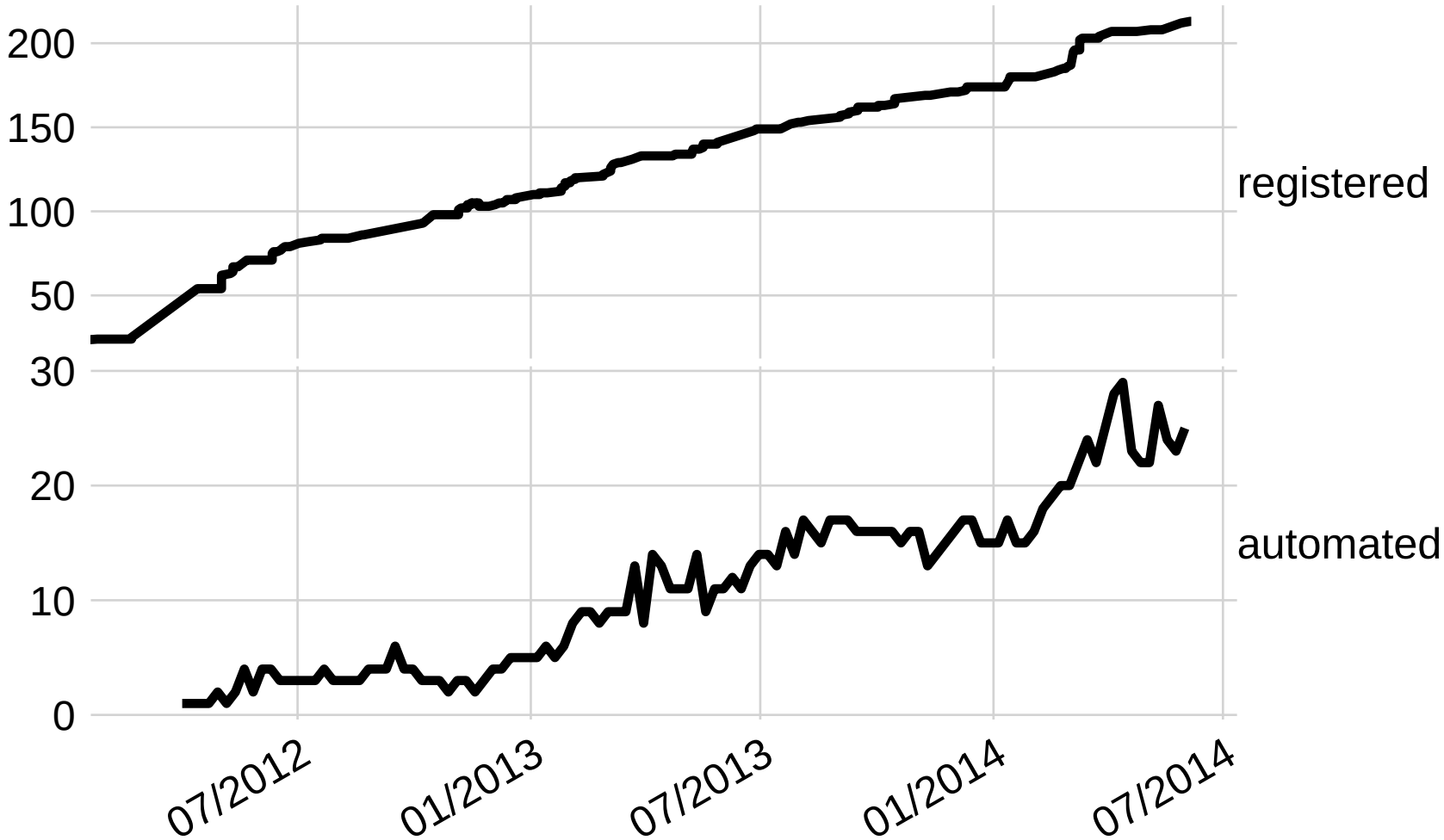
Data collected by us

- Early-warning system: ARAKIS
- Botnet monitoring
- Sandbox
- Sinkhole
  - multiple botnets
  - up to 700 connections & 10k packets per second
  - Tomasz Bukowski **The Art of Sinkholing**
    June 24th (Tuesday) 11:00 @ Sparks

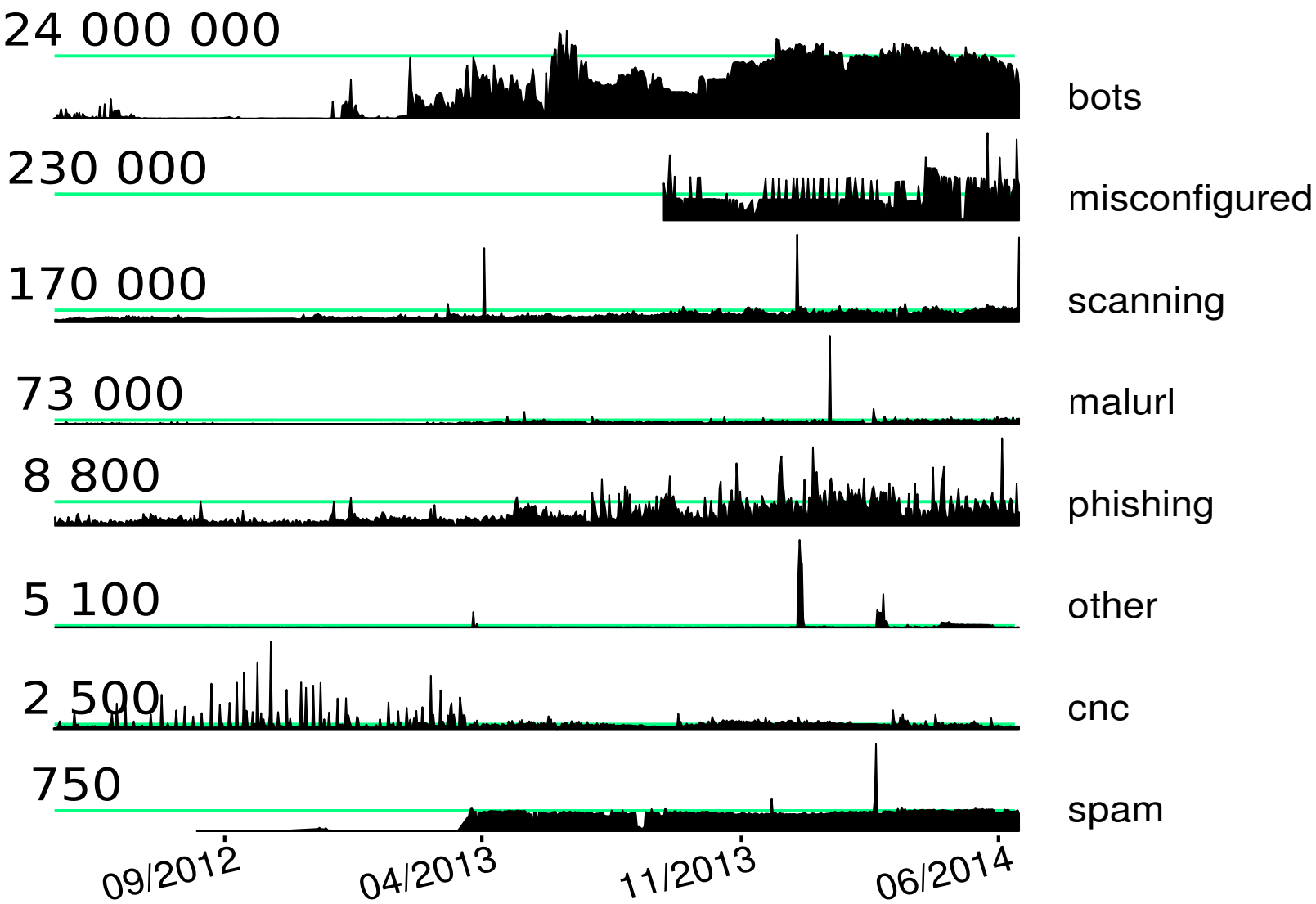
Data from external sources

- Types:
  - Malicious websites
  - C&C servers
  - Infected machines
  - Phishing
  - Open proxies
  - Open resolvers
  - Bot configuration
  - Scanning hosts
  - Spammers
- $\sim$ **25** organizations
- $\sim$ **45** feeds

Recipients


200
150
100
50
30
20
10
0
07/2012
01/2013
07/2013
01/2014
07/2014
registered
automated

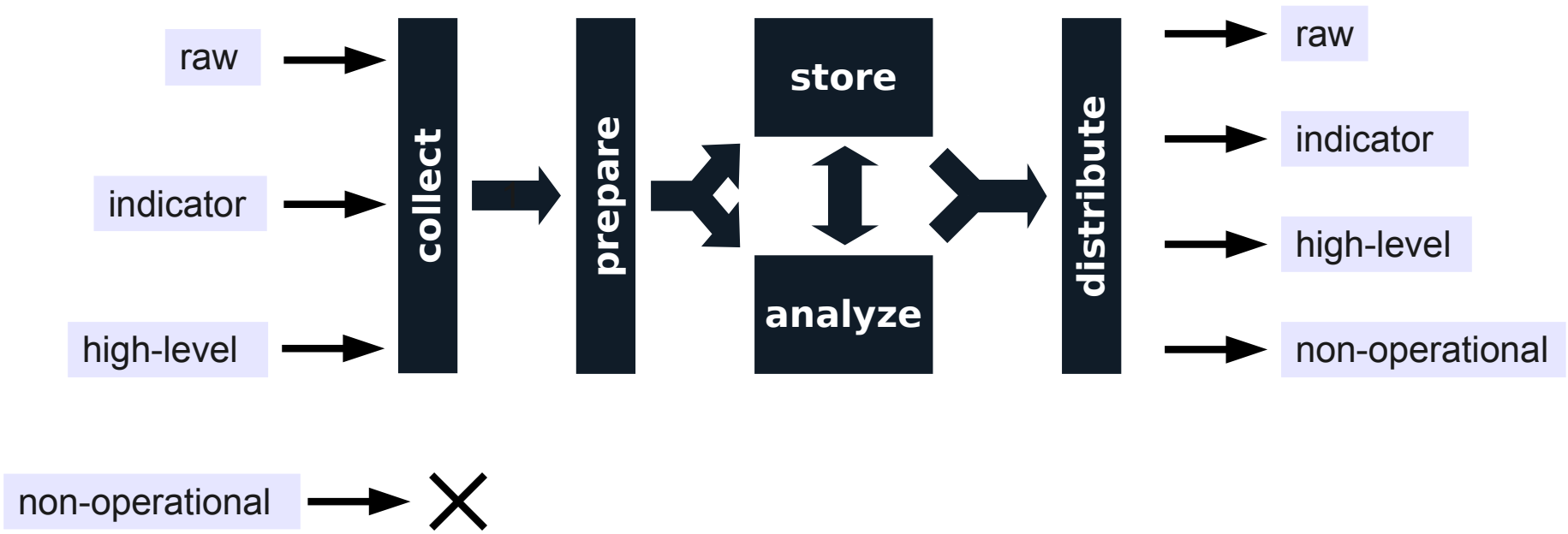
Records processed daily



Meanwhile. . .

- Other systems for processing similar information:
  - Megatron (CERT-SE)
  - AbuseHelper (CERT-FI)
  - Collective Intelligence Framework (REN-ISAC)
  - *add your project here: . . .*

Generalized data-processing system



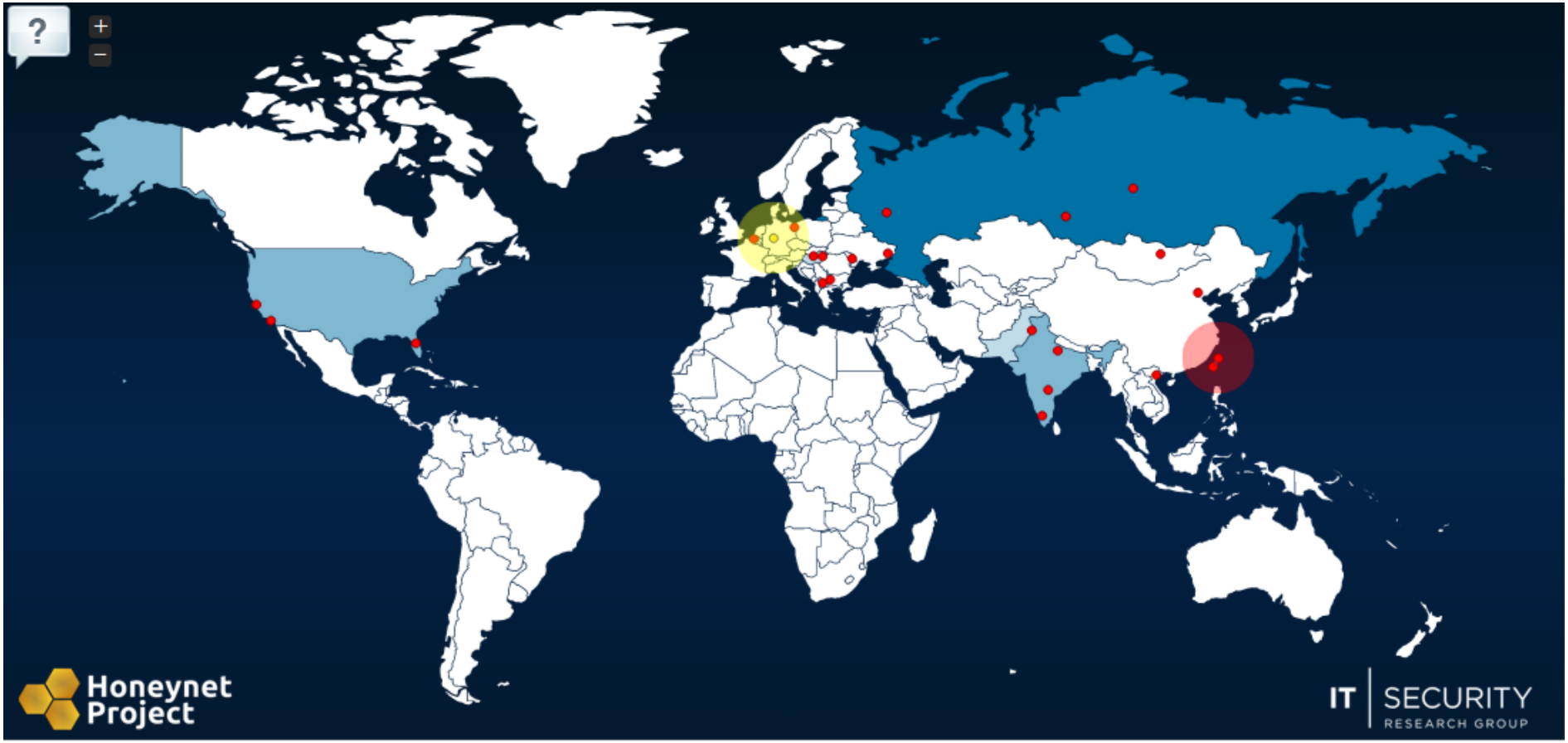
Types of information

- *Raw* data
  - Examples:
    - traffic dumps, NetFlow
    - logs
    - output from honeypots
  - usually machine generated
  - requires further analysis
  - integration with other systems
- Indicators
  - Examples:
    - IPs, domains, URLs, malware hashes
    - type-specific metadata (e.g. email header)
  - labeled data (known security context)
  - usually machine generated
  - can be deployed in automated security systems

Types of information (continued)

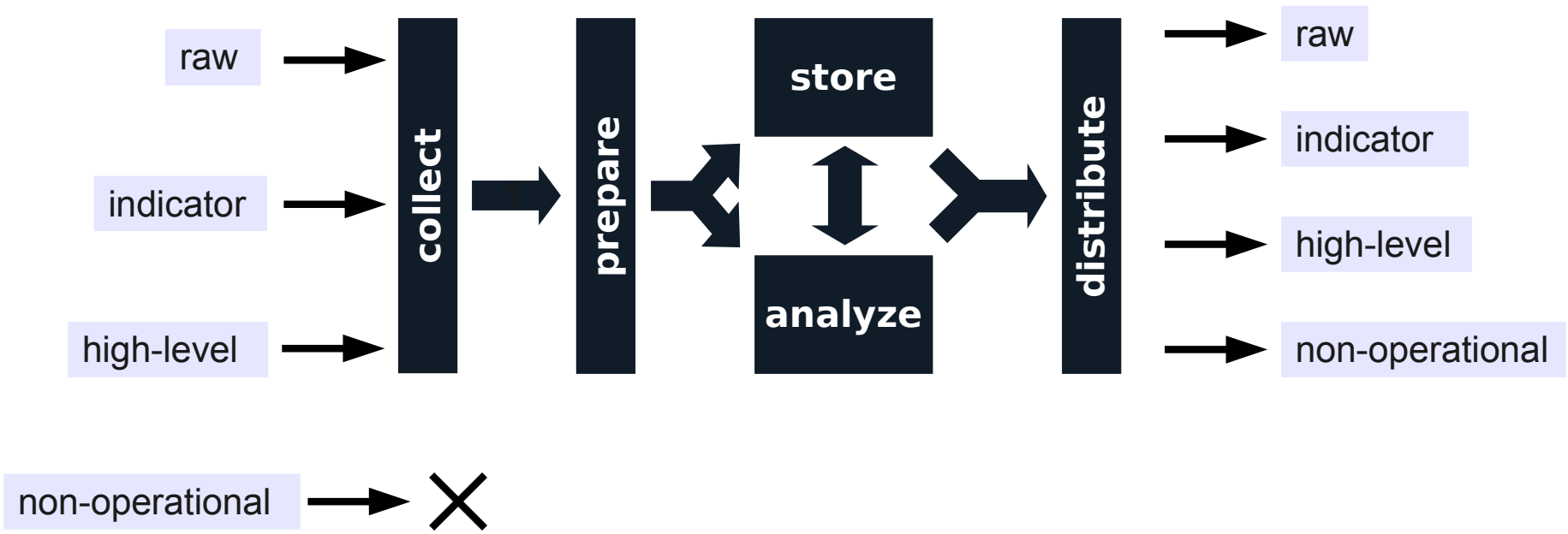
- *High-level* data
  - Examples:
    - vulnerability alert
    - TTP, actors
    - affected assets
    - other alerts (e.g. anomaly)
  - often human input
  - not always possible to automate handling
- Non-operational
  - supports decision making

Non-operational information



map.honeynet.org

Generalized data-processing system



Collect

- Location: internal vs external
- Producer: machine generated vs human input
- One-time vs regular
- Batch vs stream
- Push vs pull

Prepare

- Parsing
  - multiple formats
  - changes to existing feeds
- Normalization
  - lack of common ontology
  - bot names
- Aggregation
- Enrichment
  - GeoIP
  - known entities (*ContactDB project by CERT.at & others*)
  - correlation
- Uncertainty
  - degree of trust in the source
  - accuracy of classification

Store

- Volume: megabytes – petabytes
- *Dataset management*
- Data model
- Technology
  - SQL databases
  - HBase
  - flat files
  - ...

Analyze

- Investigation (queries)
- Trends
- Anomalies
- Common patterns
- Identify campaigns
- ...
- **?**

Distribute

- Identify recipients
- Technical aspects
  - API
  - format
  - standards
  - push vs pull
- How to utilize available information?

n6: *The Next Generation*

- n6 in mid-2013:
  - millions records per day
  - sharing with many entities
  - existing scripts for integration with other systems
  - growing wishlist
- Redesign
  - more scalable
  - make data more accessible
  - easier to add new features
  - from Perl + Python + Bash to pure Python
- Tests ongoing
- Gradual transition to the new version

n6: data collection

- Indicator-level information
- Multiple external feeds of varying quality
- Multiple internal feeds
- Medium–high volume
- Mostly coming in batches but also streams

n6: data preparation

- Event-oriented processing
- Normalization of incoming data
- Enrichment: DNS, ASN
- Aggregation
- Organization mapping

n6: storage

- Keeping original data
- Central DB for normalized events
- SQL (MySQL / TokuDB)
- Critical for overall performance

n6: analysis

- Annual report for 2013: www.cert.pl/PDF/Report_CP_2013.pdf
- Ad-hoc queries
- Planned: continuous automated analyzes

n6: distribution of data

- Multiple output formats: **JSON**, CSV, IODEF
- Fine-grained permission model
- API
  - REST
  - stream

Previously

```
DOMAIN  IPs ASNs
alotibi.panadool400.com 141.138.203.138 35470
anowona.cn  205.164.24.45,216.172.154.35,50.117.116.205,50.117.122.90 18779,18779,18779,18779
arta.romail3arnest.info 173.230.133.99  3595
asp.spinchats.com 74.122.123.234  46785
bambambam.info  31.170.179.179  35236
bff.7oorq8.com  64.124.180.220  6461
bff4.7oorq8.com 208.185.82.133  6461
bunker.org.ua 23.22.33.59 16509
computo164.laweb.es 187.214.120.147
ff.converter50.com  175.6.1.159 4134
hcuewgbbnfdu1ew.com 80.83.124.187 29141
hcuewgbbnfs1uew.com 176.31.117.59 16276
internet.estr.es  189.135.116.163 8151
kubusse.ru  31.170.179.179  35236
legionarios.servecounterstrike.com  76.74.255.138 13768
```

```
"time (UTC)","ip","asn","cc","target","fqdn","url"
"2013-10-06 20:09:19","81.177.174.13","8342","RU","Other","fexinfos.1gb.ru","http://fexinfos.1gb
"2013-10-06 20:08:56","81.177.174.13","8342","RU","Other","fexinfos.1gb.ru","http://fexinfos.1gb
"2013-10-06 19:34:22","184.173.225.223","36351","US","Other","184.173.225.223-static.reverse.sof
"2013-10-06 19:33:12","69.58.188.39,69.58.188.40","30060,30060","US,US","Other","bit.ly","http:/
"2013-10-06 19:35:21","184.173.225.223","36351","US","Other","184.173.225.223-static.reverse.sof
"2013-10-06 18:19:50","94.60.32.243","35818","RO","Apple","apple.com-infomartion-account-update-
"2013-10-06 18:58:08","67.212.64.226","10929","CA","Other","www.mauleaircanada.ca","http://www.m
```

Now: unified format


```
[
    {
        "address": [
            {
                "ip": "195.187.240.100",
                "cc": "PL",
                "asn": 12824
            }
        ],
        "adip": "x.2.137.140",
        "category": "bots",
        "confidence": "medium",
        "count": 18,
        "dport": 80,
        "fqdn": "example.com",
        "id": "26c8fd5097251dd15dc8431b267c65cf",
        "name": "B58-DGA2",
        "origin": "sinkhole",
        "proto": tcp,
        "source": "b",
        "sport": 51869,
        "time": "2013-09-18T15:35:32",
        "until": "2013-09-18T19:00:00"
    },
    {
        ...
    }
]
```

n6: web interface

n6 Portal (preview version) :: Threats inside my network

Threats inside my network | Other threats | Search events

Select search criteria:

Select item to filter by

Add

Search

Events that originated within your network or domain.

Showing most recent events.

Time	Category	Name	IP	ASN	Country	FQDN	Confidence
2013-10-07 06:49:00	phish		194.181.14.178	8308	PL	studioh.webd.pl	low
2013-10-07 06:49:00	phish		194.181.14.178	8308	PL	studioh.webd.pl	low
2013-10-05 18:30:36	bots	B54-CODE	195.187.84.66	1887	PL	dcu-a-204.microsoftinternetsafety.net	medium
2013-10-05 18:06:18	bots	Conficker	193.59.128.36	8308	PL		medium
2013-10-05 18:00:58	bots	Conficker	193.59.13.112	8308	PL		medium
2013-10-05 18:00:20	bots	Conficker	193.59.128.36	8308	PL		medium
2013-10-05 17:59:12	bots	Conficker	193.59.13.112	8308	PL		medium
2013-10-05 17:58:39	bots	Conficker	194.181.94.162	8308	PL		medium
2013-10-05 17:57:01	bots	Conficker	194.181.94.162	8308	PL		medium
2013-10-05 17:53:47	bots	B54-CONFIG	194.181.245.167	8308	PL	199.2.137.202	medium
2013-10-05 17:53:22	bots	B54-CODE	194.181.245.167	8308	PL	dcu-a-204.microsoftinternetsafety.net	medium
2013-10-05 17:43:21	bots	Conficker	193.59.181.58	8308	PL		medium
2013-10-05 17:41:42	bots	Conficker	193.59.181.58	8308	PL		medium
2013-10-05 17:35:38	bots	B58-DGA2	193.59.168.2	8308	PL	www.shletz.co.cc	medium
2013-10-05 17:30:34	bots	Conficker	193.59.74.23	8308	PL		medium
2013-10-05 17:29:34	bots	Conficker	193.59.74.23	8308	PL		medium
2013-10-05 17:13:06	bots	Conficker	193.59.152.43	8308	PL		medium
2013-10-05 17:11:19	bots	Conficker	193.59.152.43	8308	PL		medium
2013-10-05 17:05:03	bots	B58-DGA2	193.59.57.34	8308	PL	atul007.co.cc	medium
2013-10-05 17:02:26	bots	Conficker	194.181.62.30	8308	PL		medium

n6: web interface

n6 Portal (preview version) :: Threats inside my network

Threats inside my network | Other threats | Search events

Select search criteria:

Select item to filter by

Add

Search

Events that originated within your network or domain.

Showing most recent events.

Time	Category	Name	IP	ASN	Country	FQDN	Confidence
2013-10-07 06:49:00	phish		194.181.14.178	8308	PL	studioh.webd.pl	low
2013-10-07 06:49:00	phish		194.181.14.178	8308	PL	studioh.webd.pl	low
2013-10-05 18:30:36	bots	B54-CODE	195.187.84.66	1887	PL	dcu-a-204.microsoftinternetsafety.net	medium
2013-10-05 18:06:18	bots	Conficker	193.59.128.36	8308	PL		medium
adip: x.x.150.27 **sport:** 2759 **id:** 47b74762b741a09ab38a9f13bd1edaad **category:** bots **confidence:** medium						**source:** d9dafbd9ef9aa59e **IP:** 193.59.128.36 **ASN:** 8308 **CC:** PL **name:** Conficker **time:** 2013-10-05T18:06:18 **dport:** 80	
2013-10-05 18:00:58	bots	Conficker	193.59.13.112	8308	PL		medium
2013-10-05 18:00:20	bots	Conficker	193.59.128.36	8308	PL		medium
2013-10-05 17:59:12	bots	Conficker	193.59.13.112	8308	PL		medium
2013-10-05 17:58:39	bots	Conficker	194.181.94.162	8308	PL		medium
2013-10-05 17:57:01	bots	Conficker	194.181.94.162	8308	PL		medium
2013-10-05 17:53:47	bots	B54-CONFIG	194.181.245.167	8308	PL	199.2.137.202	medium
2013-10-05 17:53:22	bots	B54-CODE	194.181.245.167	8308	PL	dcu-a-204.microsoftinternetsafety.net	medium
2013-10-05 17:43:21	bots	Conficker	193.59.181.58	8308	PL		medium
2013-10-05 17:41:42	bots	Conficker	193.59.181.58	8308	PL		medium
2013-10-05 17:35:38	bots	B58-DGA2	193.59.168.2	8308	PL	www.shletz.co.cc	medium
2013-10-05 17:30:34	bots	Conficker	193.59.74.23	8308	PL		medium
2013-10-05 17:29:34	bots	Conficker	193.59.74.23	8308	PL		medium

ARAKIS

- Early warning system
- Developed by NASK
- Operational since 2007

ARAKIS: data collection

- Large number of sensors in Polish address space
- Multiple participating organizations
- Server-side honeypots
- Firewall logs
- Darknet
- *raw* data

ARAKIS: data preparation

- GeoIP
- Matching against known networks

ARAKIS: storage

- MySQL
- Raw data (PCAP & logs) available
- Periodic cleanup

ARAKIS: analysis

- Trend analysis for port–protocol pairs
- Identifying common patterns
  - text-mining algorithms applied to network traffic
  - using clusterization algorithm to group similar attacks
- Rankings
- Rare event detection

ARAKIS: distribution

- Integration with n6
- Reports for selected areas by email
- Many features are human-oriented
- Public interface: http://dashboard.arakis.pl/en/

Challenges

- Assessing quality of information
- Creating vs forwarding actionable information
- Situational awareness
  - observe ongoing campaigns
  - identify compromised parts of the infrastructure
  - assess risks
  - …
- Requirements:
  - knowledge of assets & threats
  - scale matters
  - not only actionable information

BoF session

- **Finding & Sharing Actionable Information**
- US-CERT & CERT Polska
- June 25th (Wednesday) 16:30 – 18:00

Thank you for your attention.

Questions / Comments?

This work has been supported by the Strategic International Collaborative R&D Promotion Project of the Ministry of Internal Affairs and Communication, Japan, and by the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 608533 (NECOMA). The opinions expressed in this presentation are those of the author and do not necessarily reflect the views of the Ministry of Internal Affairs and Communications, Japan, or of the European Commission.