27th ANNUAL
FIRST CONFERENCE BERLIN

14 - 19 JUNE 2015

UNIFIED SECURITY:
IMPROVING THE FUTURE

Nationaal Cyber Security Centrum
Ministerie van Veiligheid en Justitie

Machine Learning for Cyber Security Intelligence

27th FIRST Conference
17 June 2015

National Cyber Security Centre
Ministry of Security and Justice

Netherlands Forensic Institute
Ministry of Security and Justice

Edwin Tump
Senior Analyst
National Cyber Security Center

Introduction | whois

- Edwin Tump

- 10 yrs at NCSC.NL (GOVCERT.NL)
  - 9 yrs as a security specialist
  - 1 yr as a security analyst

- Areas of special interest
  - Information collection (e.g. Taranis)
  - Tooling
  - Output

- *Machine learning?*

Introduction | Machine Learning

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i \xi_i \\ s.t.\ \forall i: \quad & y_i\,(\mathbf{w}\cdot\mathbf{x}_i) \geq 1 \end{aligned}$$

$$\frac{\partial L}{\partial \alpha_i} = M\left(\Phi(\textstyle\sum_{j=1}^{K}\alpha_j) - \Phi(\alpha_i)\right) + \sum_{m=1}^{M}\left(\Phi(\gamma_{mi}) - \Phi(\textstyle\sum_{j=1}^{K}\gamma\right.$$

$$\mathcal{L} = \sum_{x\in X} q(x)\left(\sum_{j=1}^{n}\lambda_j f_j(x) - \log Z\right) - \sum_{j=1}^{n}\lambda_j \sum_{x\in X} q(x) f_j(x) + \sum_{j=1}^{n}\lambda_j$$

$$p(\mathbf{w};\alpha,\beta) = \log\int_\theta \sum_{\mathbf{z}} p(\mathbf{w}|\mathbf{z};\beta)p(\mathbf{z}|\theta)p(\theta;\alpha)\frac{q(\theta,\mathbf{z};\gamma,\phi)}{q(\theta,\mathbf{z};\gamma,\phi)}d\theta$$

$$- \log q(\theta,\mathbf{z};\gamma,\phi)]$$

$$\min_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i,j}\alpha_i\,\alpha_j\,\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j) - \sum_i y_i\,\alpha_i$$

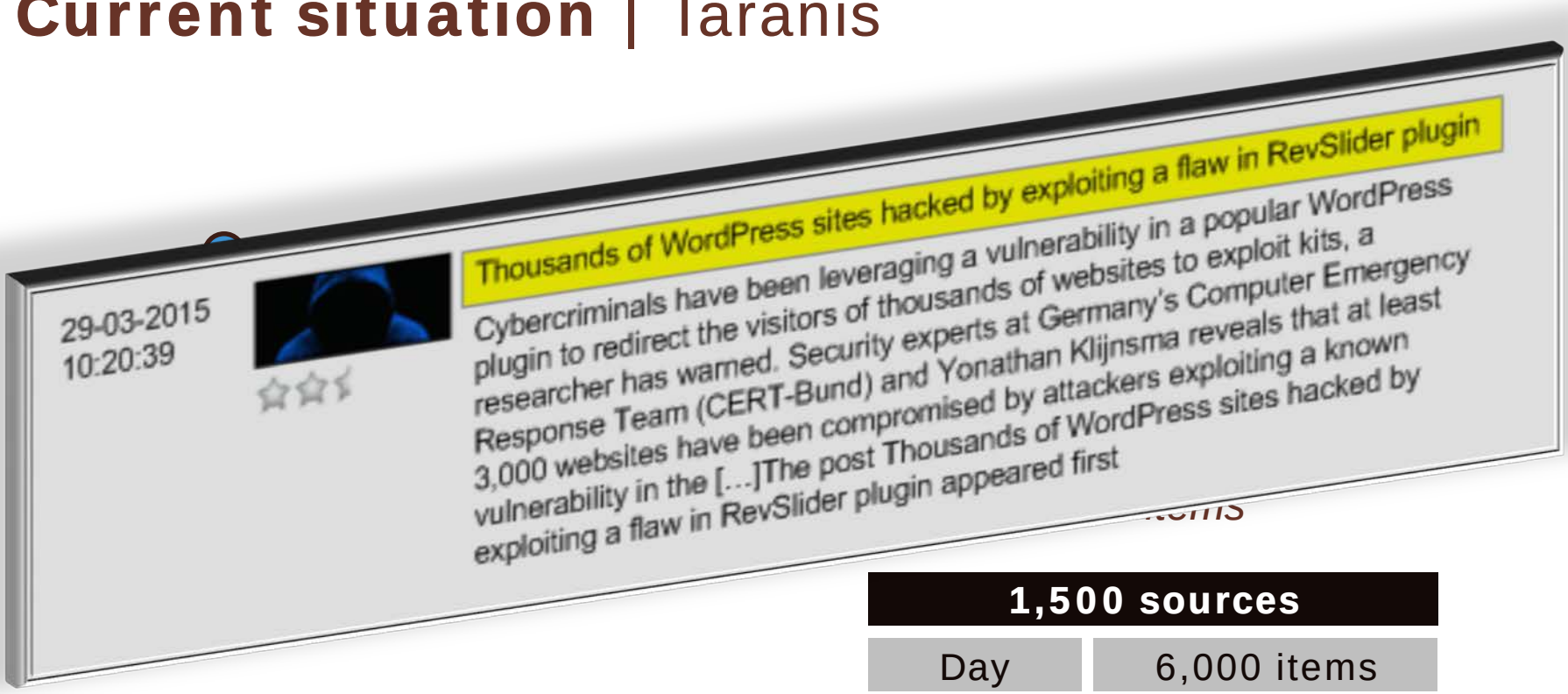
$$y = \operatorname{sgn}\left(\frac{1}{m_+}\sum_{\{i:y_i=+1\}}(\mathbf{x}\cdot\mathbf{x}_i) - \frac{1}{m_-}\sum_{\{i:y_i=-1\}}(\mathbf{x}\cdot\mathbf{x}_i) + b\right)$$

Nationaal Cyber Security Centrum
Ministerie van Veiligheid en Justitie

Agenda

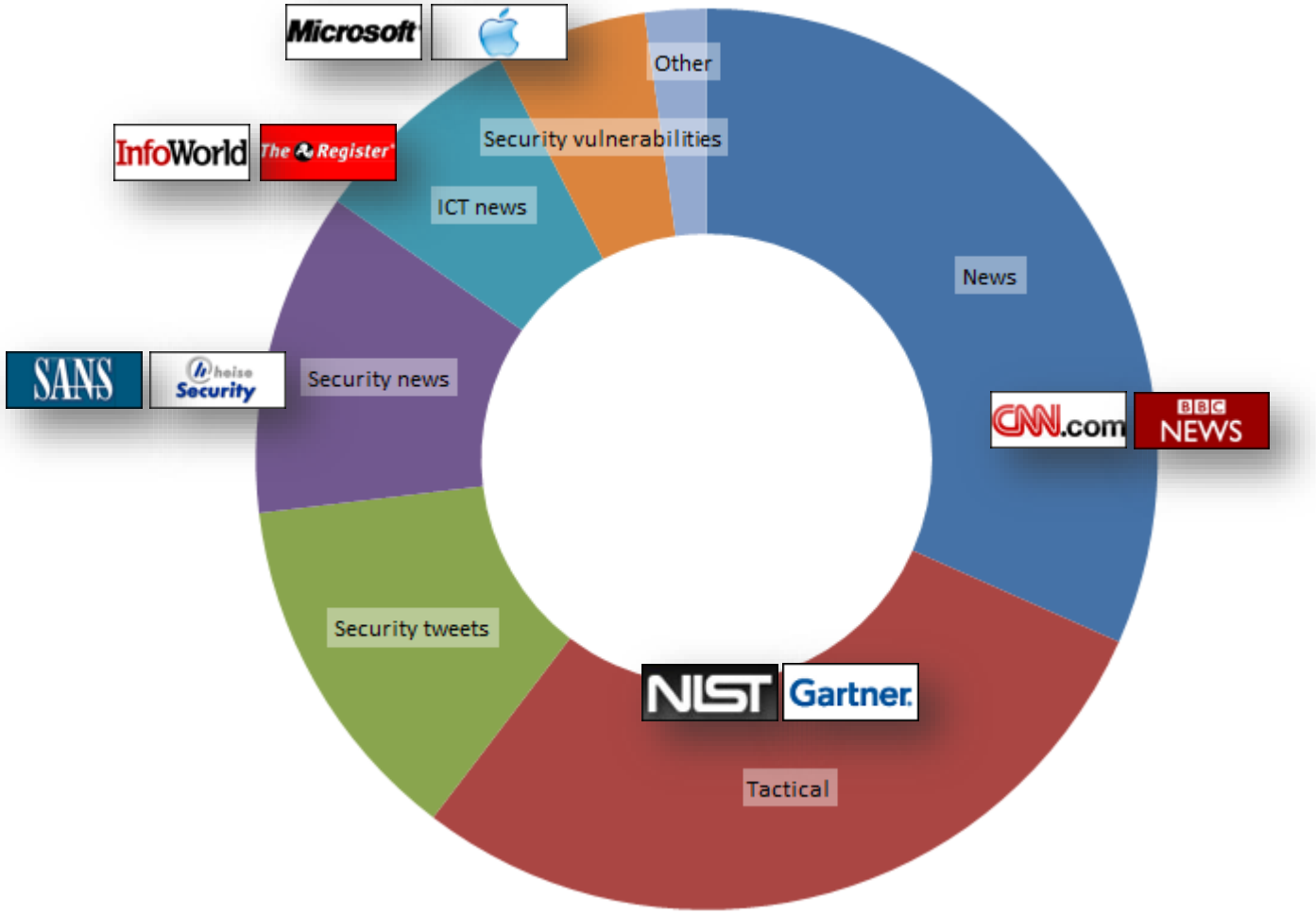
- Current situation
- Challenges
- Desired situation
- Approach
- Machine learning
- Results
- Conclusions

Current situation | Taranis



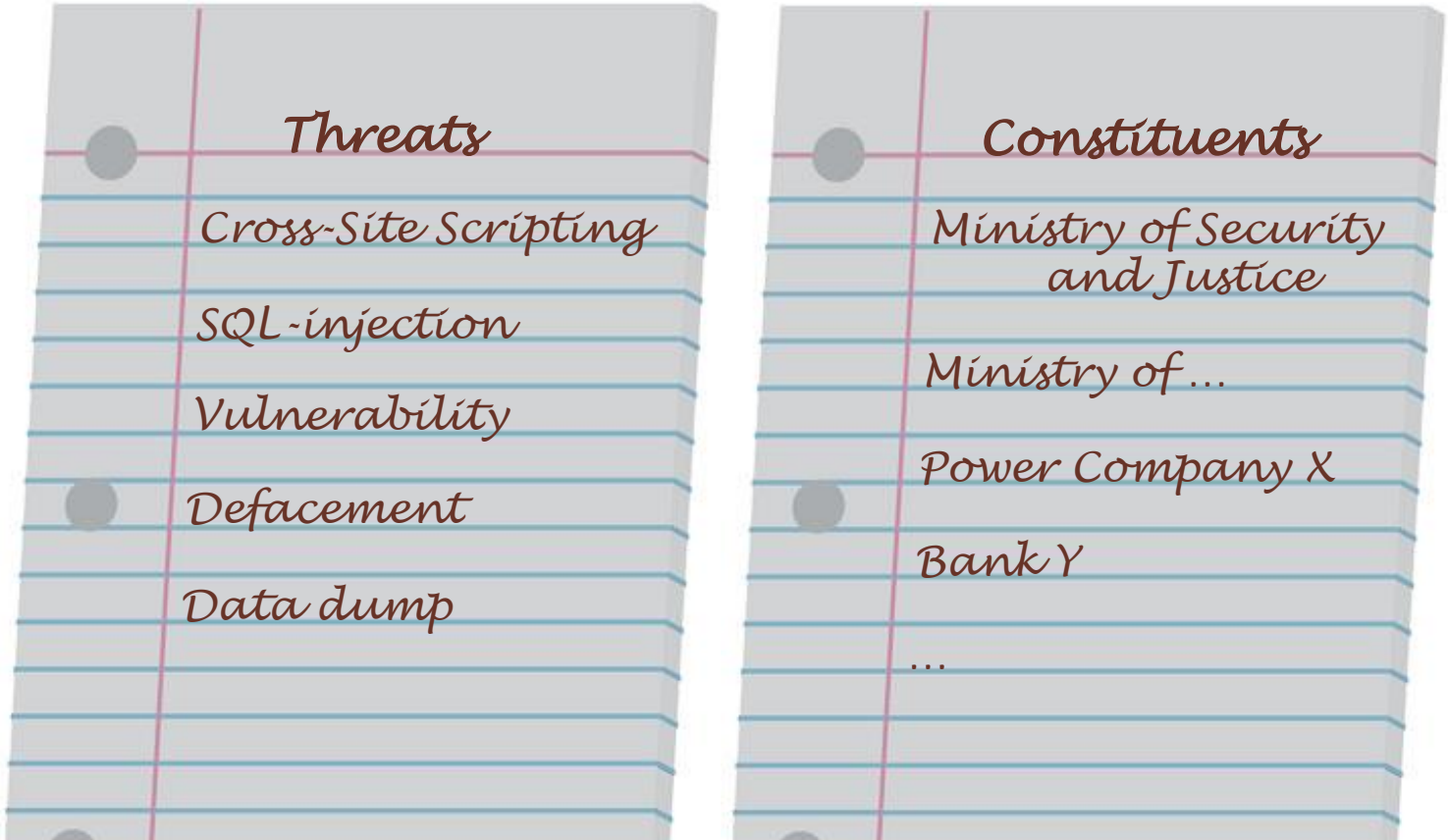
1,500 sources	
Day	6,000 items
Week	42,000 items
Month	180,000 items
Year	2,190,000 items

Current situation | Taranis categories



Current situation | Taranis keyword matching

- Automatic dispatch of irrelevant items
  - Only for specific sources
  - Relevant items determined based on keyword lists

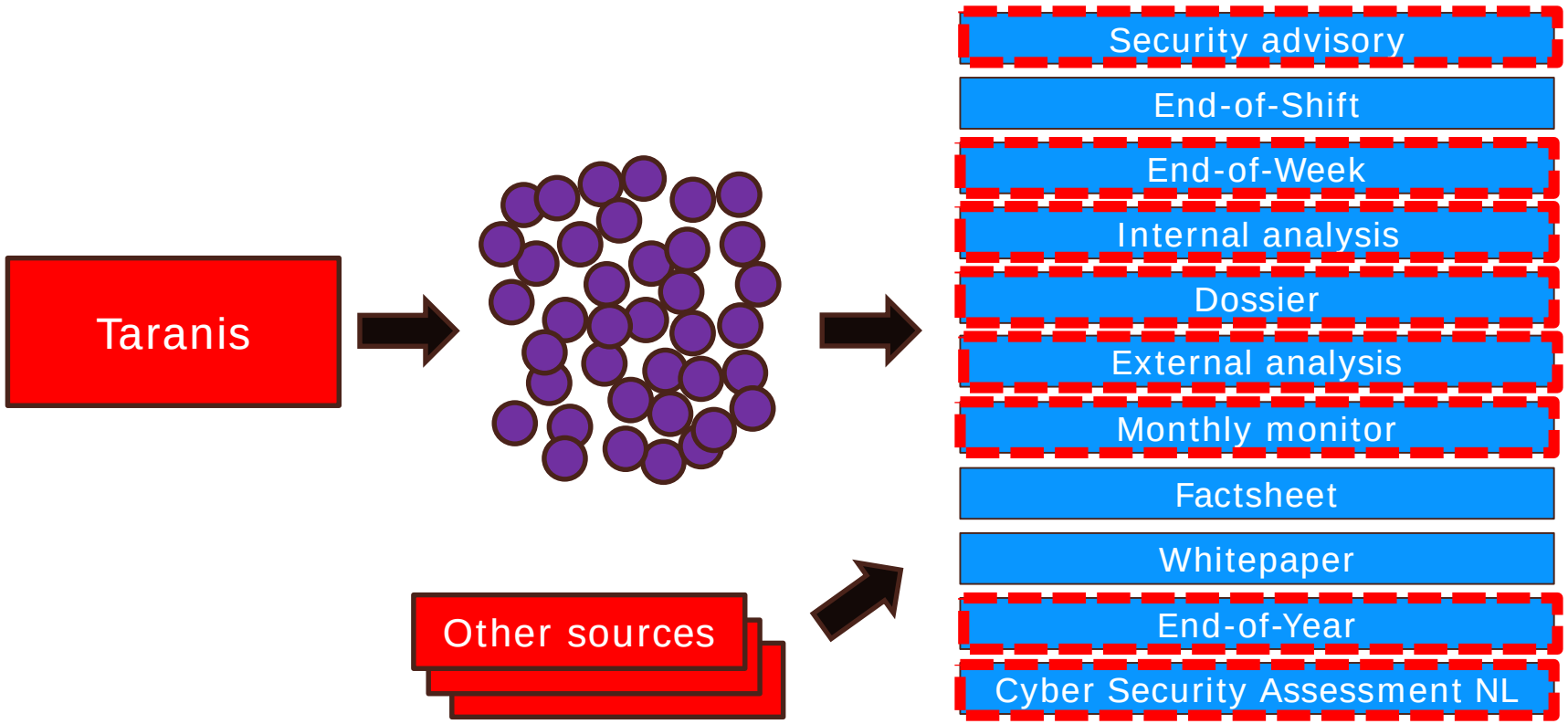


Current situation | Taranis deduplication

– Clustering/deduplication of news items

Current situation | Taranis 5 phases

Current situation | Outputs



Challenges

Image courtesy of Stuart Miles at FreeDigitalPhotos.net

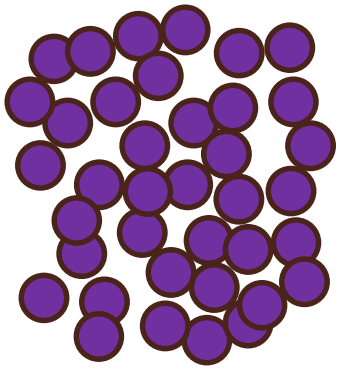
Challenges

- Handling growing amount of information is labor intensive
  - Determine relevant topics
  - Determine new trends
  - Rate the reliability of news: the good, the bad and the ugly
  - Keep dossiers up-to-date

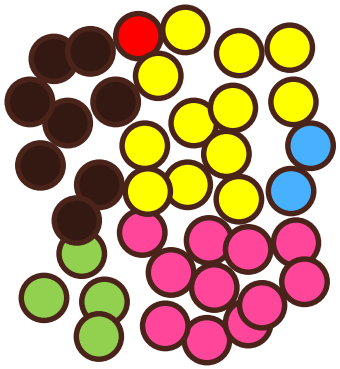
- Higher expectations; reports must be …
  - … timely
  - … complete
  - … short

Desired situation

Desired situation



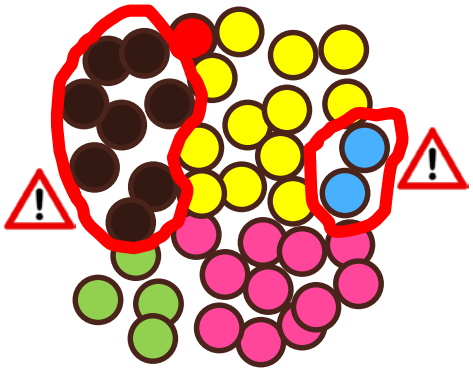
Desired situation



Automatically …

… cluster items and detect stories

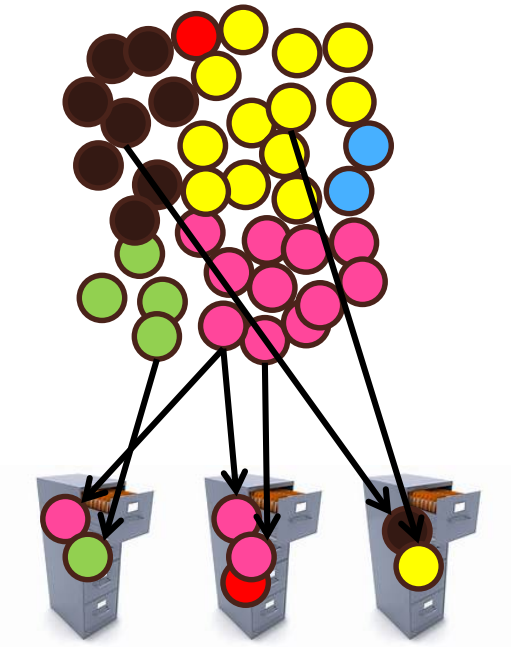
Desired situation



Automatically …

… determine story relevance

Desired situation



Automatically …

… assign items to dossiers

Approach

Image courtesy of Stuart Miles at FreeDigitalPhotos.net

Approach

- Agile-scrum
  - Close contact between participants
  - Multiple short sprints (4 sprints of 2 weeks)

- Principles
  - High level of trust
  - Use of well-known concepts in the field of machine learning
  - Open mind, not restricted by a ready-mode product

- Deliverables
  - Better understanding of usefulness of machine learning techniques
  - Proof-of-Concept(s)
  - Detailed requirements for future work
  - Paper describing the results

Approach

Nationaal Cyber Security Centrum
Ministerie van Veiligheid en Justitie

- Data (Taranis)
- Expertise
  - Source data
  - Process
  - Products

Netherlands Forensic Institute
Ministry of Security and Justice

- Tools
- Expertise
  - Text mining
  - Machine learning
  - Information retrieval

Machine learning (ML)

- Main methods
  - Supervised learning
    - Classification
    - Regression
  - Unsupervised learning
    - Clustering
    - Association
    - Density estimation
    - Dimensionality reduction
  - Reinforcement learning
  - Semi supervised learning

Proof-of-Concept toolset

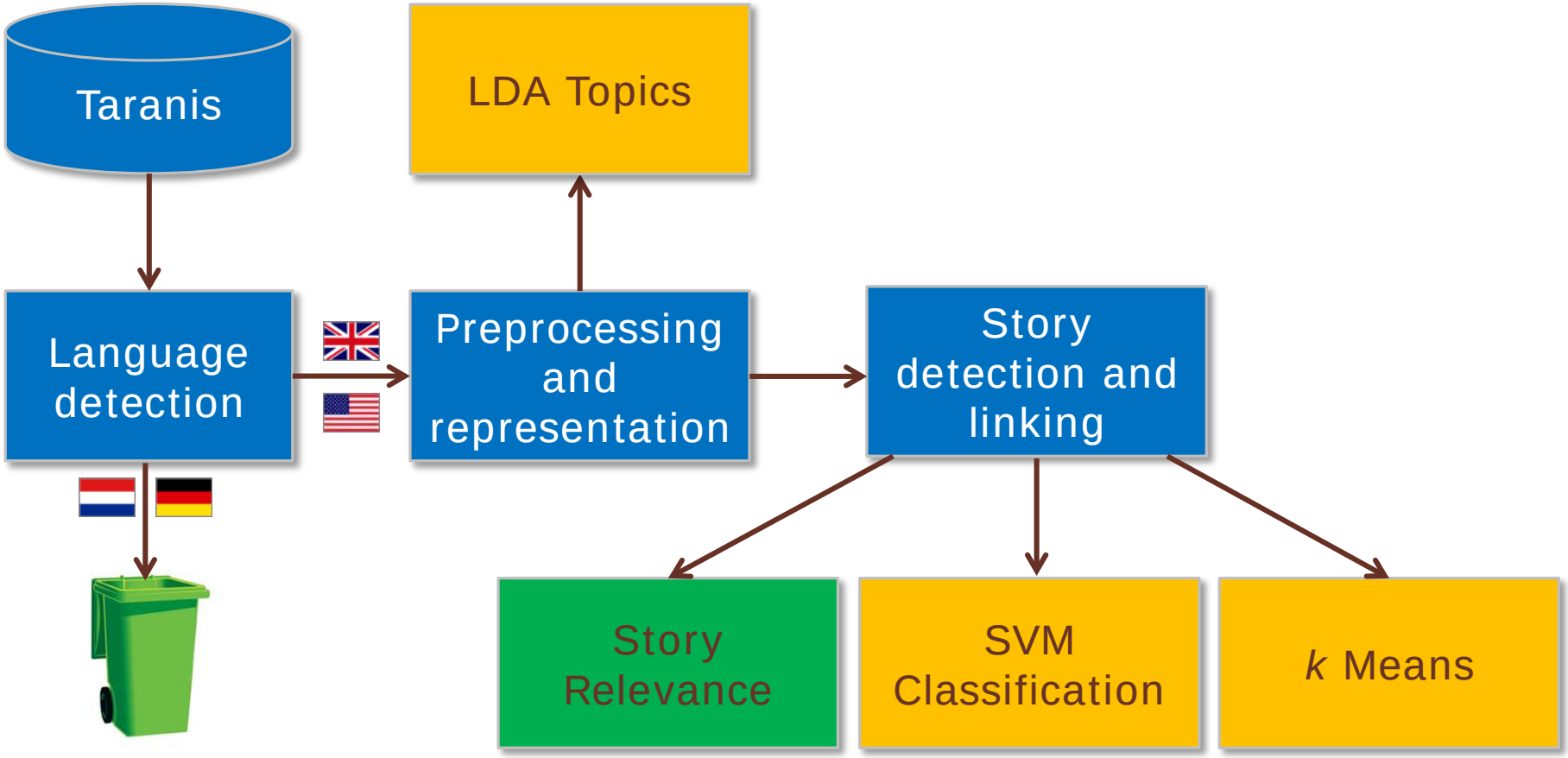
- PostgreSQL
  - Relational database
  - Offline partial copy of the Taranis database
  - http://www.postgresql.org/

- R
  - A “free software environment for statistical computing and graphics”
  - http://www.R-project.org

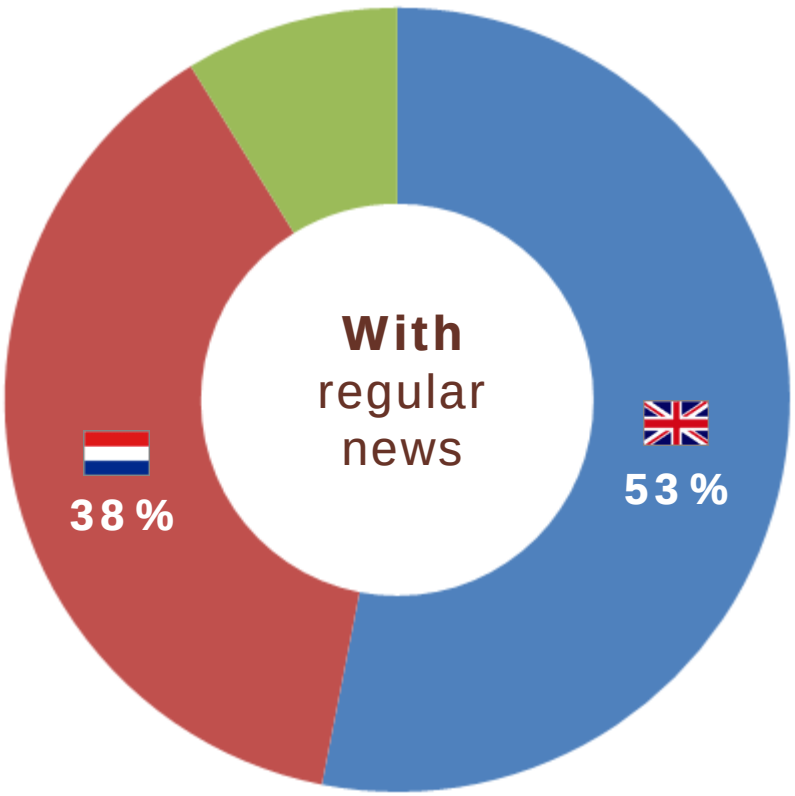
- Tableau
  - Data visualization / data analysis
  - http://www.tableau.com/

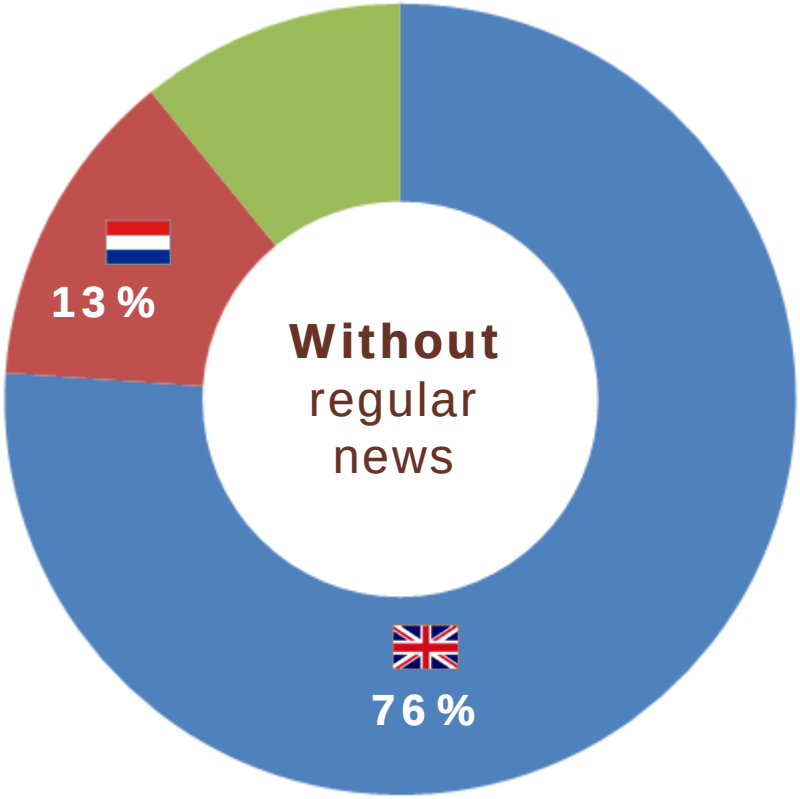
- FoamTree
  - JavaScript tree map visualization
  - http://carrotsearch.com/foamtree-overview

Overview


Taranis
LDA Topics
Language detection
Preprocessing and representation
Story detection and linking
Story Relevance
SVM Classification
k Means

Language detection





Preprocessing and representation

- Remove noisy words (stop words)
- Remove duplicates
- Remove numbers
- Remove punctuation marks
- Normalize case (upper to lower)
- Stem words
  - Use of Porter Stemming Algorithm

Example

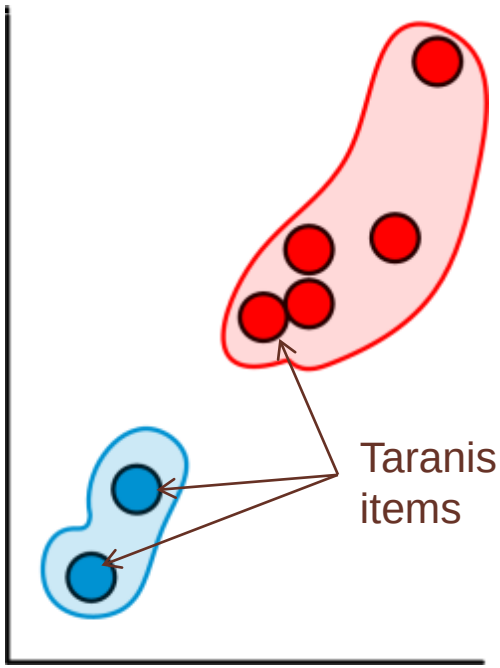
Exploit Kit Delivers DNS Changer to Thousands of Routers

A malicious campaign deployed by cybercriminals aims at changing the Domain Name System (DNS) server settings in router configuration, responsible for retrieving the correct web pages from legitimate web servers. An attacker changing these settings can point to malicious locations, exposing the victim to a wide range of risks varying from credential stealing and ad-fraud to traffic interception and malware delivery.

exploit deliv changer thousand router malicy campaign deploy cybercrimin chang domain server router configur respons retriev correct page legitim server attacker chang point malicy locat expos victim wide rang risk vary credenty steal fraud traffic intercept malwar delivery

k Means

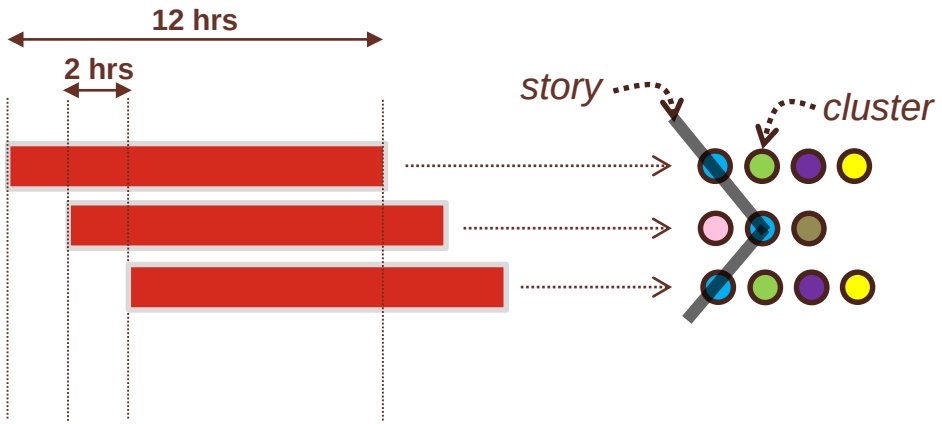
- History already begun in 1957, but published in 1982 by Stuart Lloyd
- Unsupervised
- Widely used because of simplicity and performance
- Words represented as vectors
  - Vectors calculated based on TF-IDF
- Steps:
  - **Assignment**: choose *k* cluster centers and then assign each point to the closest cluster center
  - **Update**: recompute each center
  - **Repeat** until update step makes no modifications or threshold is reached



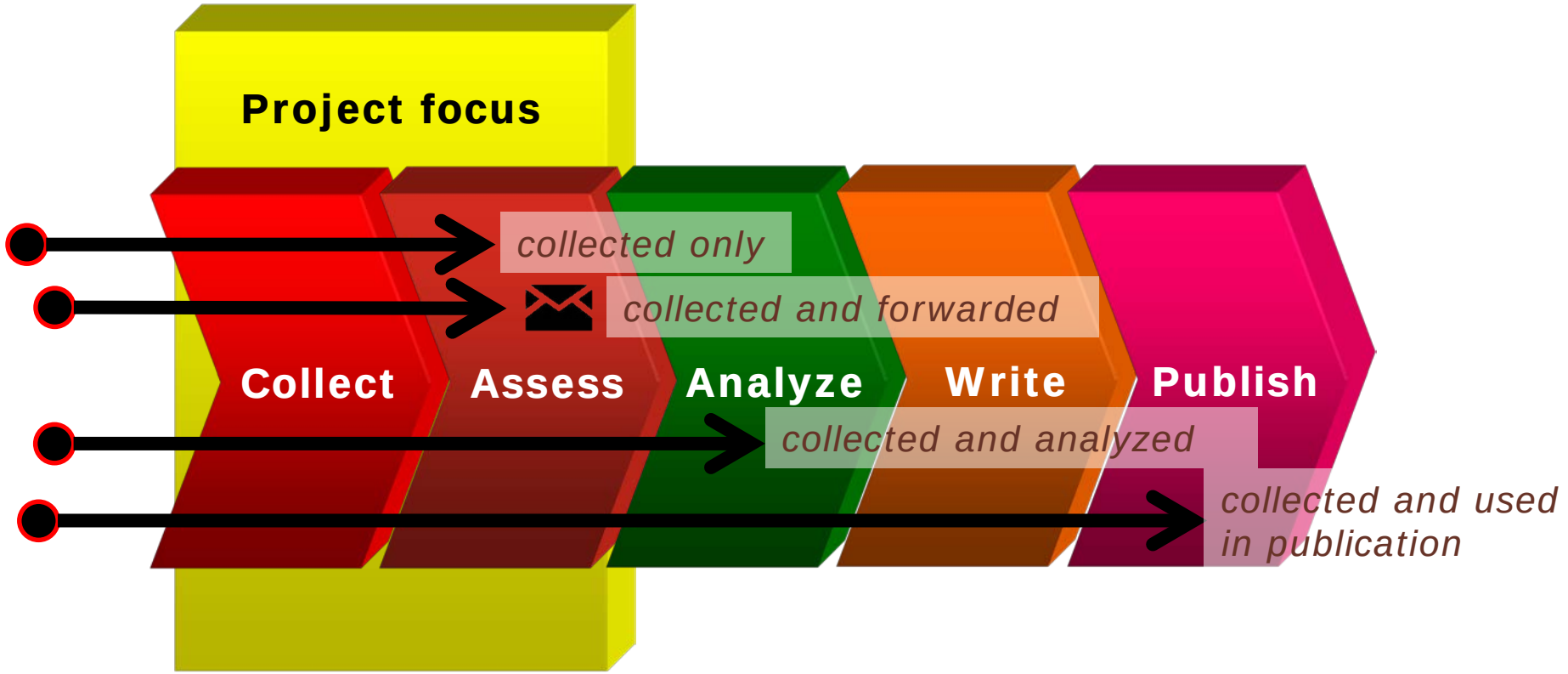

Apple, Google, Freak
Facebook, Android, Dropbox
Malware, Patch, Microsoft
Testing, Florida, Paypal
Card, Bank, Fraud
Android, Trojan, Permalink
Service, Encryption, Drop
Clinton, Mail, Hillary
Attack, Blog, Ddos
News, Mobile, Unsafe
Files, Ransomware, Game
Email, Nsa, Clinton
Data, Breach, Anthem
Cia, United, Director
Group, Police, Equation
Cyber, Crime, Csi
Security, Devices, Blackberry
Update, Microsoft, Windows
Hackers, Twitter, Https
Internet, Law, Botnets
Data, Breach, Largest
Hacking, Website, Isis
Cia, Wordpress, Gmt
Cybersecurity, Bill, Senate
FoamTree
Security, Network, Threat

Story Detection and Relevance

- Steps
  1. Deduplication
  2. Clustering in (overlapping) sliding windows
     - Windows of 12 hrs
     - Slices/slicing of 2 hrs
  3. Clustering between time slices
  4. Determine story relevance, based on
     - Volume of documents within the story
     - Source reputation (*derived from mail actions on items from source*)
     - Cross-category stories
- Important parameters
  - `'within story' (0…1)`
  - `'between story' (0…1)`

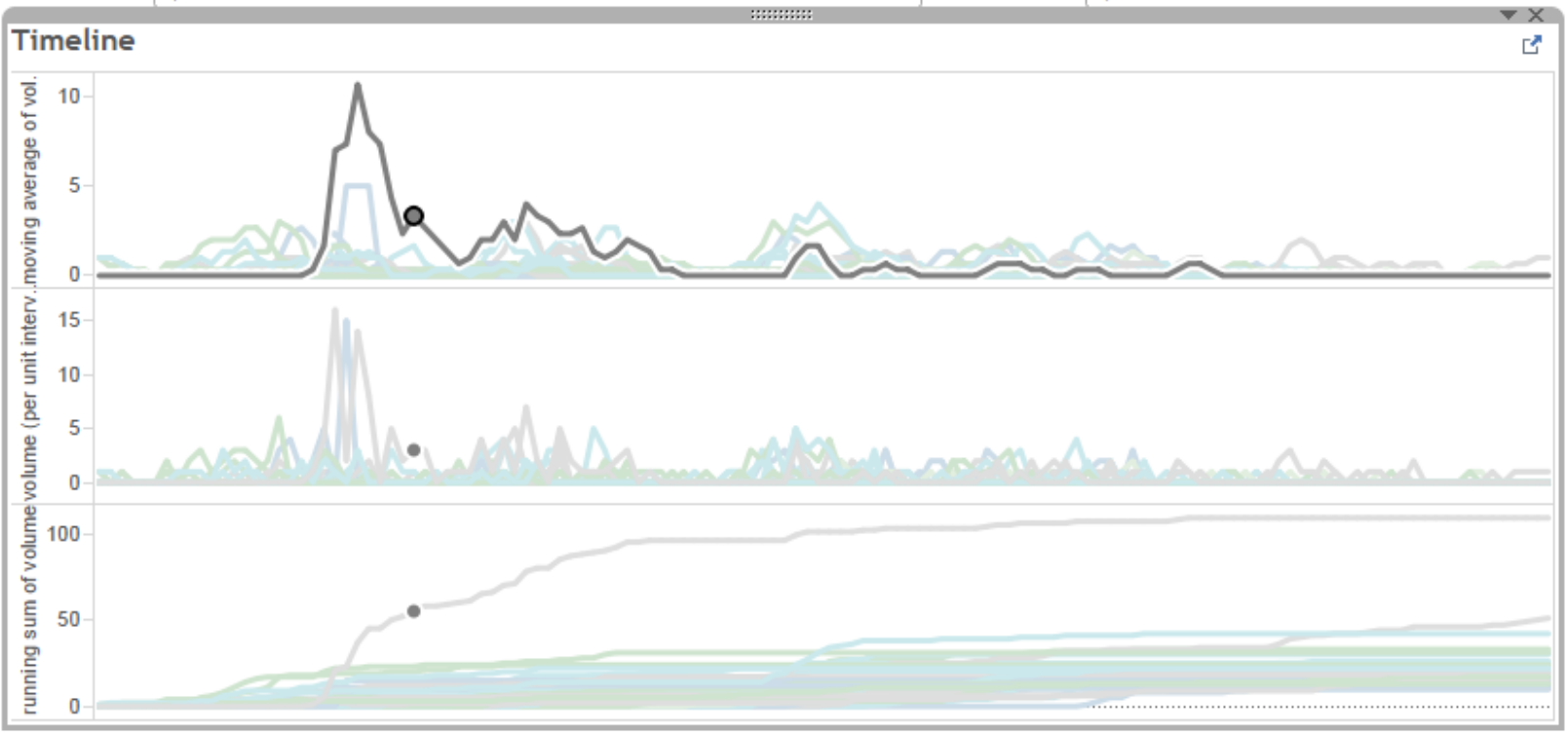


Story Relevance | Taranis 5 phases



date 2015-03-22 14:00:00 medium ict

within.stor.. 0,5 between.st.. 0,5



Created

16-3-2015 23:(22-3-2015 20:(

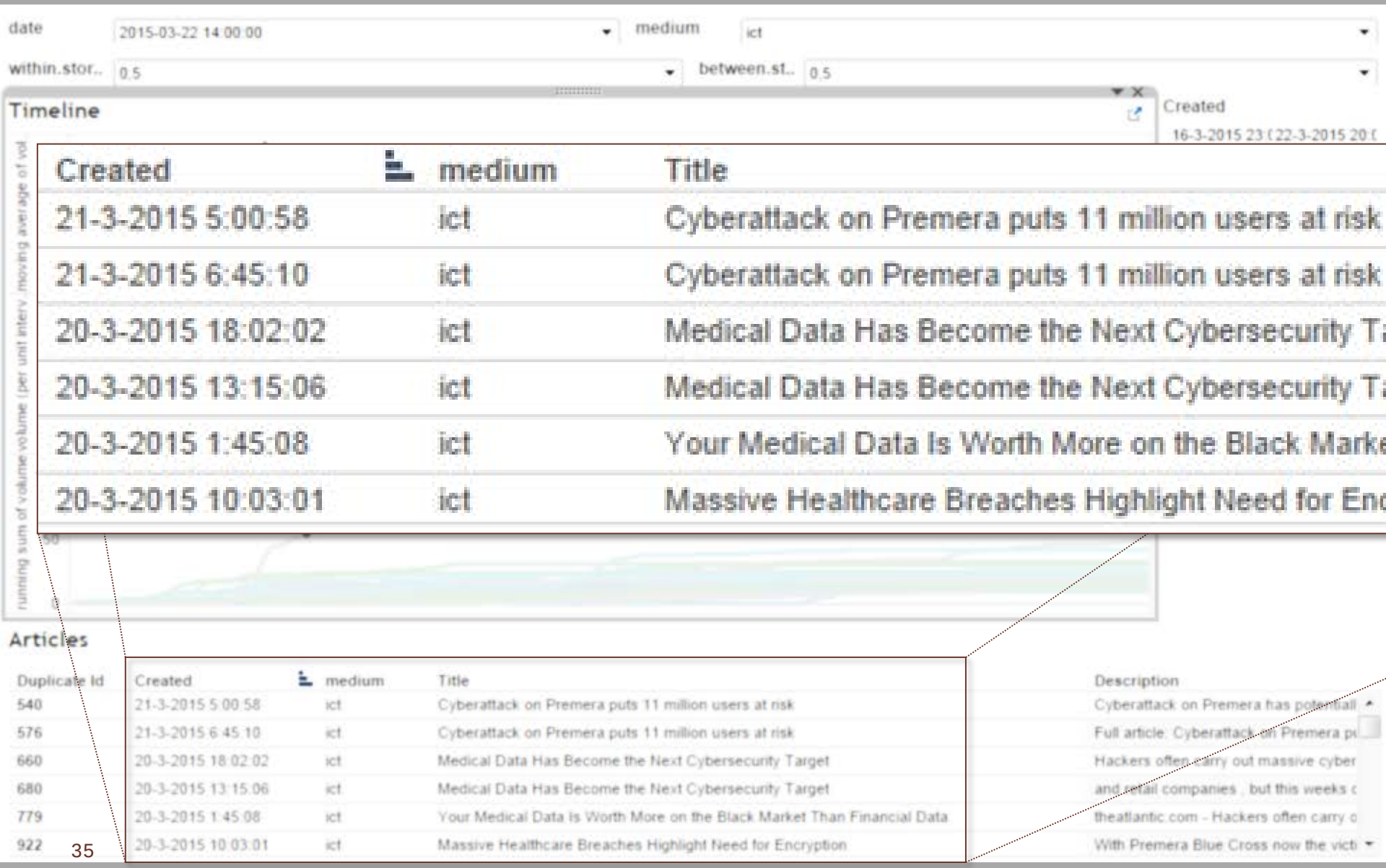
Min. number of articles

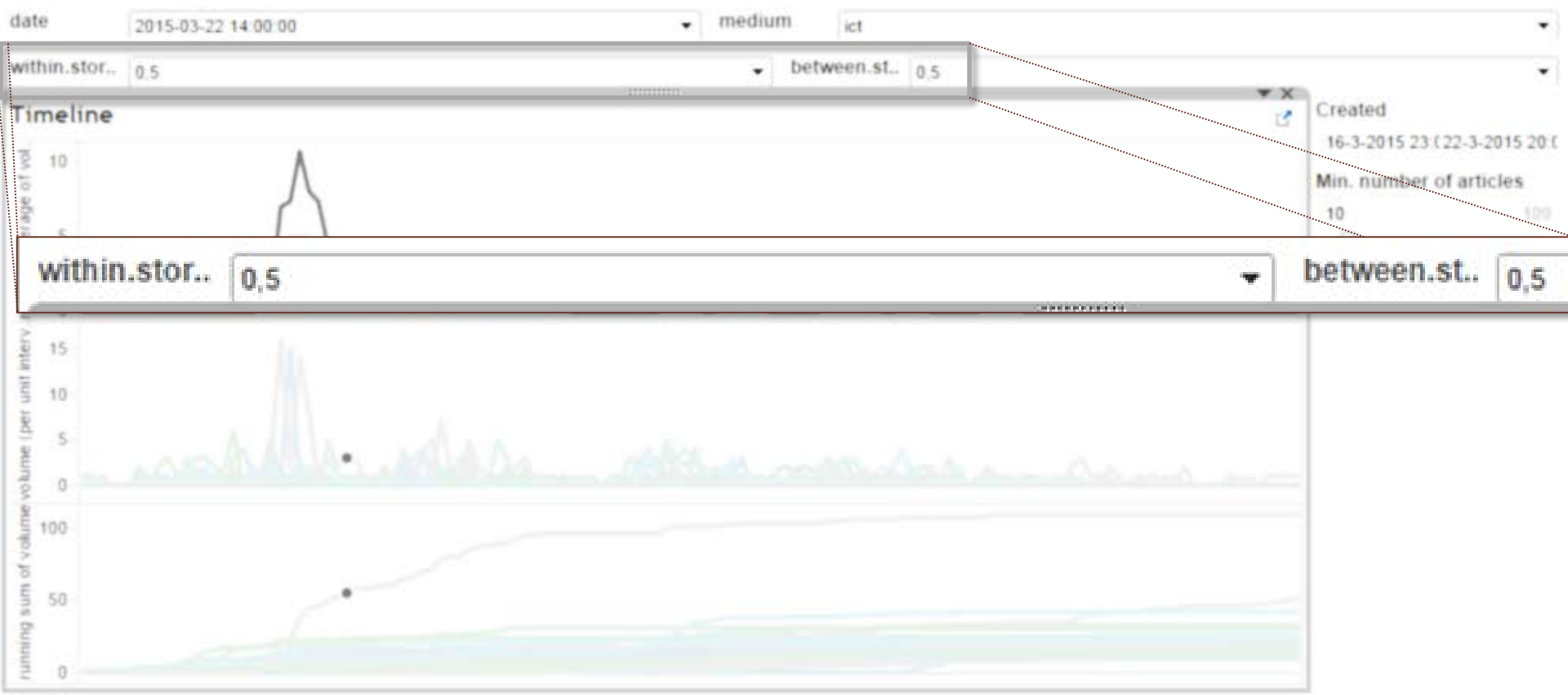
10 109

Articles

Duplicate Id	Created	medium	Title	Description
540	21-3-2015 5:00:58	ict	Cyberattack on Premera puts 11 million users at risk	Cyberattack on Premera has potentiall
576	21-3-2015 6:45:10	ict	Cyberattack on Premera puts 11 million users at risk	Full article: Cyberattack on Premera pu
660	20-3-2015 18:02:02	ict	Medical Data Has Become the Next Cybersecurity Target	Hackers often carry out massive cyber
680	20-3-2015 13:15:06	ict	Medical Data Has Become the Next Cybersecurity Target	and retail companies , but this weeks c
779	20-3-2015 1:45:08	ict	Your Medical Data Is Worth More on the Black Market Than Financial Data	theatlantic.com - Hackers often carry o
922	20-3-2015 10:03:01	ict	Massive Healthcare Breaches Highlight Need for Encryption	With Premera Blue Cross now the victi


date
2015-03-22 14:00:00
medium
ict
within.stor..
0.5
between.st..
0.5
Timeline
Created
Min. number of articles
running sum of volume volume (per unit interv. moving average of vol
10
5
0
15
10
5
0
100
50
0
Articles
Duplicate Id
Created
medium
Title
Description
540
21-3-2015 5:00:58
ict
Cyberattack on Premera puts 11 million users at risk
Cyberattack on Premera has potentiall
576
21-3-2015 6:45:10
ict
Cyberattack on Premera puts 11 million users at risk
Full article: Cyberattack on Premera p
660
20-3-2015 18:02:02
ict
Medical Data Has Become the Next Cybersecurity Target
Hackers often carry out massive cyber
680
20-3-2015 13:15:06
ict
Medical Data Has Become the Next Cybersecurity Target
and retail companies , but this weeks
779
20-3-2015 1:45:08
ict
Your Medical Data Is Worth More on the Black Market Than Financial Data
theatlantic.com - Hackers often carry o
922
20-3-2015 10:03:01
ict
Massive Healthcare Breaches Highlight Need for Encryption
With Premera Blue Cross now the victi


date
2015-03-22 14:00:00
medium
ict
within.stor..
0.5
between.st..
0.5
Timeline
Created
16-3-2015 23 (22-3-2015 20 (
running sum of volume volume (per unit interv..moving average of vol
Created
medium
Title
21-3-2015 5:00:58
ict
Cyberattack on Premera puts 11 million users at risk
21-3-2015 6:45:10
ict
Cyberattack on Premera puts 11 million users at risk
20-3-2015 18:02:02
ict
Medical Data Has Become the Next Cybersecurity T
20-3-2015 13:15:06
ict
Medical Data Has Become the Next Cybersecurity T
20-3-2015 1:45:08
ict
Your Medical Data Is Worth More on the Black Marke
20-3-2015 10:03:01
ict
Massive Healthcare Breaches Highlight Need for En
Articles
Duplicate Id
Created
medium
Title
Description
540
21-3-2015 5:00:58
ict
Cyberattack on Premera puts 11 million users at risk
Cyberattack on Premera has potentiall
576
21-3-2015 6:45:10
ict
Cyberattack on Premera puts 11 million users at risk
Full article: Cyberattack on Premera p
660
20-3-2015 18:02:02
ict
Medical Data Has Become the Next Cybersecurity Target
Hackers often carry out massive cyber
680
20-3-2015 13:15:06
ict
Medical Data Has Become the Next Cybersecurity Target
and retail companies , but this weeks
779
20-3-2015 1:45:08
ict
Your Medical Data Is Worth More on the Black Market Than Financial Data
theatlantic.com - Hackers often carry o
922
20-3-2015 10:03:01
ict
Massive Healthcare Breaches Highlight Need for Encryption
With Premera Blue Cross now the victi

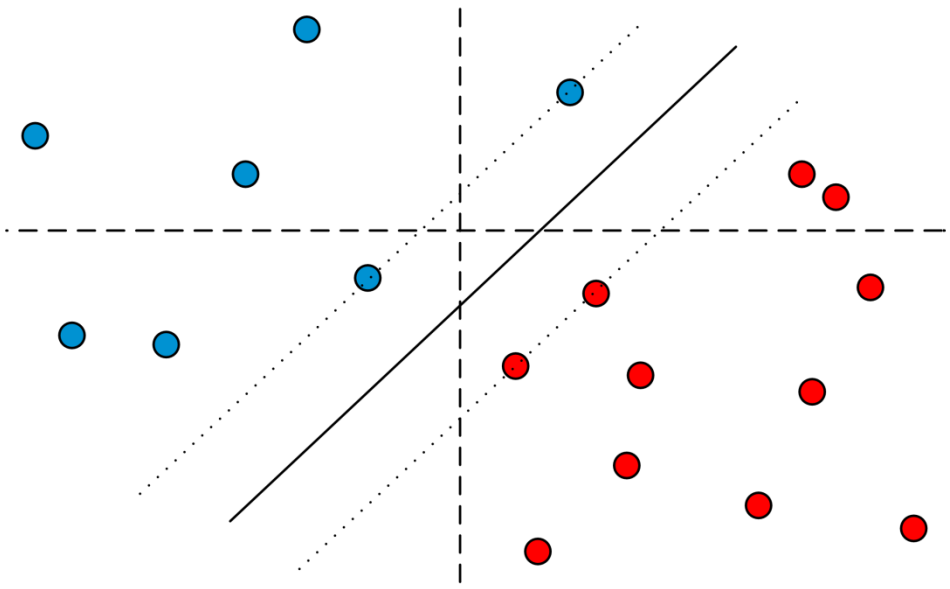


Articles

Duplicate Id	Created	medium	Title	Description
540	21-3-2015 5:00:58	ict	Cyberattack on Premera puts 11 million users at risk	Cyberattack on Premera has potentiall
576	21-3-2015 6:45:10	ict	Cyberattack on Premera puts 11 million users at risk	Full article: Cyberattack on Premera p
660	20-3-2015 18:02:02	ict	Medical Data Has Become the Next Cybersecurity Target	Hackers often carry out massive cyber
680	20-3-2015 13:15:06	ict	Medical Data Has Become the Next Cybersecurity Target	and retail companies , but this weeks c
779	20-3-2015 1:45:08	ict	Your Medical Data Is Worth More on the Black Market Than Financial Data	theatlantic.com - Hackers often carry o
922	20-3-2015 10:03:01	ict	Massive Healthcare Breaches Highlight Need for Encryption	With Premera Blue Cross now the victi

Support Vector Machines (SVM)

- Supervised classification
- SVMs take your data and draw a hyper plane to divide your dataset into groups of positive and negative observations
- Types of classification:
  - Binary classification (simplest)
  - Multi-category classification
    - One-against-one
    - One-against-rest
- Cyber Security Assessment Report Netherlands 2013 (CSAN-3) used as training set



Columns

Rows sc

sc	
-0.376358128	Yahoo CISO Reveals Email Encryption Plan
-0.431531280	PASSWORDSCON 2014 - Security for the People: End-User Authentication Security on the Internet - Mark Stanislav
-0.434096443	Russian hack could see end of usernames and passwords: industry
-0.446401023	Yahoo! Mail to Offer Users End-to-End Encryption Next Year
-0.451571476	Russian hack could see end of usernames and passwords: industry
-0.451615460	What to do if you think Russian hackers stole your login, password
-0.454186447	Yahoo to Release End-to-End Encryption for Email Users
-0.463702707	Yahoo is adding end-to-end encryption to email
-0.466118977	Yahoo to Release End-to-End Encryption for Email Users
-0.467383127	Yahoo ! to ! deploy ! E2E ! crypto ! by ! 2015 !
-0.469573763	New Site Recovers Files Locked by Cryptolocker Ransomware
-0.472445759	Black Hat: Yahoo to implement end-to-end mail encryption by next year
-0.473997771	New Site Recovers Files Locked by Cryptolocker Ransomware
-0.477977909	How to foil SynoLocker and minimize the damage
-0.479680293	Microsoft pulls updates, recommends uninstall
-0.479970581	Worried about Russian hackers? Take these personal-data precautions
-0.482440955	Yahoo is adding end-to-end encryption to email
-0.485209352	Microsoft pulls updates, recommends uninstall Microsoft Pulls Updates, Recommends Uninstall
-0.488805006	PASSWORDSCON 2014 - Security for the People: End-User Authentication Security on the Internet - Mark Stanislav
-0.497562851	Gemalto to Acquire SafeNet
-0.501046713	Yahoo Mail users will get end-to-end encryption option next year
-0.501631881	GCHQ approves six university cyber security Masters courses
-0.502838255	BSides Las Vegas 2014 - Bring your own Risky Apps Michael Raggo - Kevin Watkins
-0.503815975	The Economics of Spam The Economics of Spam

category

vulnerability of the end...

date

2014-08-31 14:00:00

sc

-0.966370464 -0.376358128

Columns

Rows sc

sc	
-0.412937738	CloudBot: A Free, Malwareless Alternative to Traditional Botnets
-0.438399160	[Infowarrior] - Meet MonsterMind, the NSA Bot That Could Wage Cyberwar Autonomously
-0.460956686	CloudBot: A Free, Malwareless Alternative to Traditional Botnets
-0.469849179	Lite Zeus - A New Zeus Variant
-0.483865248	Latest Gameover botnet lays low, looking to resist takedown
-0.487848946	CloudBot: A Free, Malwareless Alternative To Traditional Botnets
-0.488855910	Latest Gameover botnet lays low, looking to resist takedown
-0.498721379	New Gameover Zeus Botnet Forming, the US Sees Most Infections
-0.501297076	iPhones, iPads ripe for the picking
-0.504120977	New Gameover ZeuS Variant and Shylock Rebuild Botnets
-0.509805665	P2P Zeus Performs Critical Update
-0.517125218	Why no one smells a RAT: Trojan uses YAHOO WEBMAIL to pick up instructions
-0.523730227	New Gameover Zeus botnet keeps growing, especially in the US
-0.525163180	Hiding A Bitcoin Mining Botnet In The Cloud
-0.526749102	CloudBot: A Free, Malwareless Alternative To Traditional Botnets
-0.530879884	iPhones, iPads ripe for the picking
-0.532426713	New Gameover ZeuS Variant and Shylock Rebuild Botnets
-0.532590914	NEUREVT Bot Analysis
-0.534079119	Poweliks malware creates no files, lays low in the registry
-0.536329751	A Good Look at the Andromeda Botnet
-0.537025766	Asprox URLViewer delivers porn adverts
-0.537491026	Secret Service: Over 1,000 Business Infected With "Backoff" Point-of-sale Malware
-0.537686269	Gameover Zeus Botnet Rebuilds
-0.539715070	Asprox URLViewer delivers porn adverts

category
botnets

date
2014-08-31 14:00:00

sc
-0.966370464 -0.412937738

Latent Dirichlet Allocation

- Unsupervised
- Assumption: documents are generated by latent topics, the hidden variables in from which the documents are observed from
- Main motivation: uncover underlying semantic structures (=hidden topics) of a document
- Each document is a mix of topics, defined by a collection of words
- Variables:
  - K = number of topics
  - W = word
  - D = documents
  - Z = topic
- Continually answer the following questions for each word:
  - How often does "W" appear in topic "Z" elsewhere?
  - How common is topic "Z" in the rest of the document
- Visualization based on own implementation of LDAVis

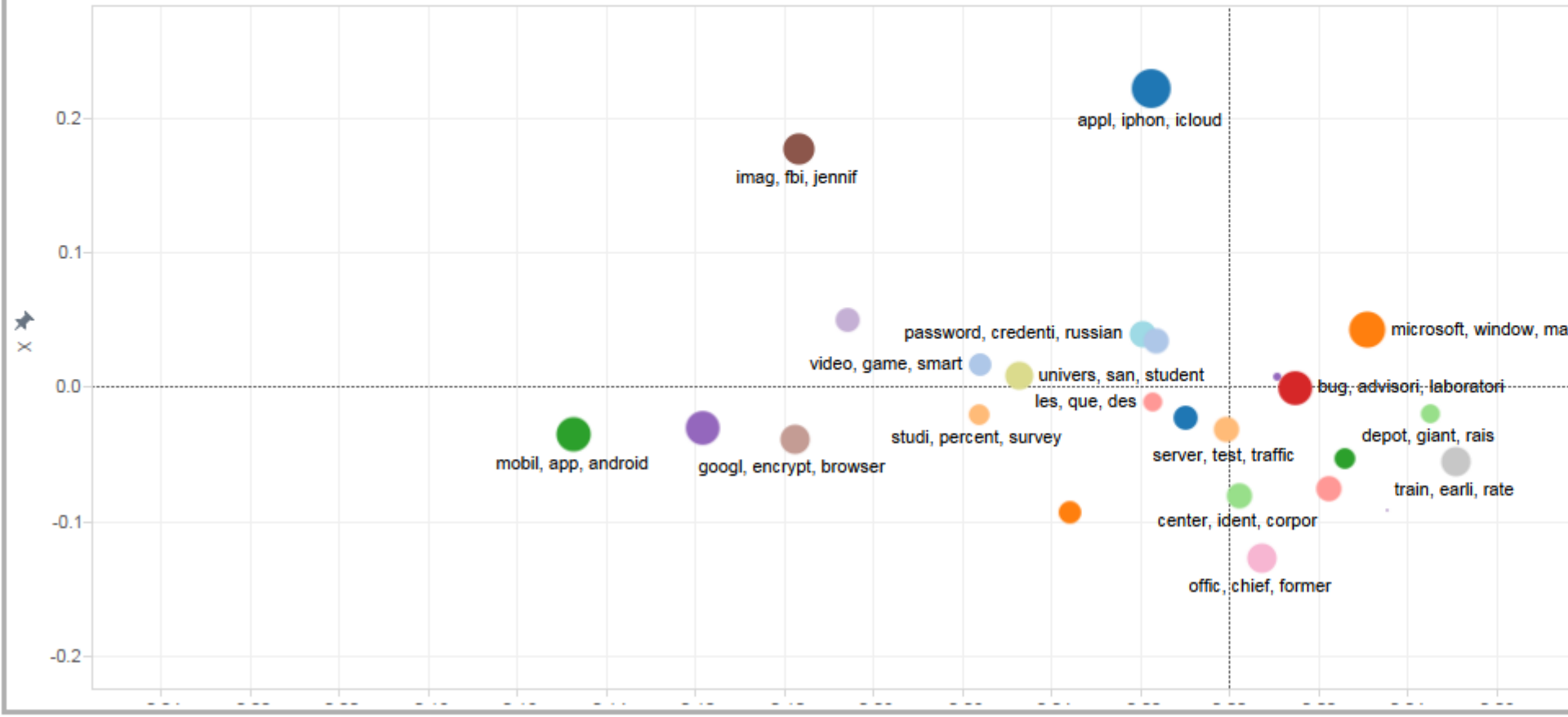
Current month

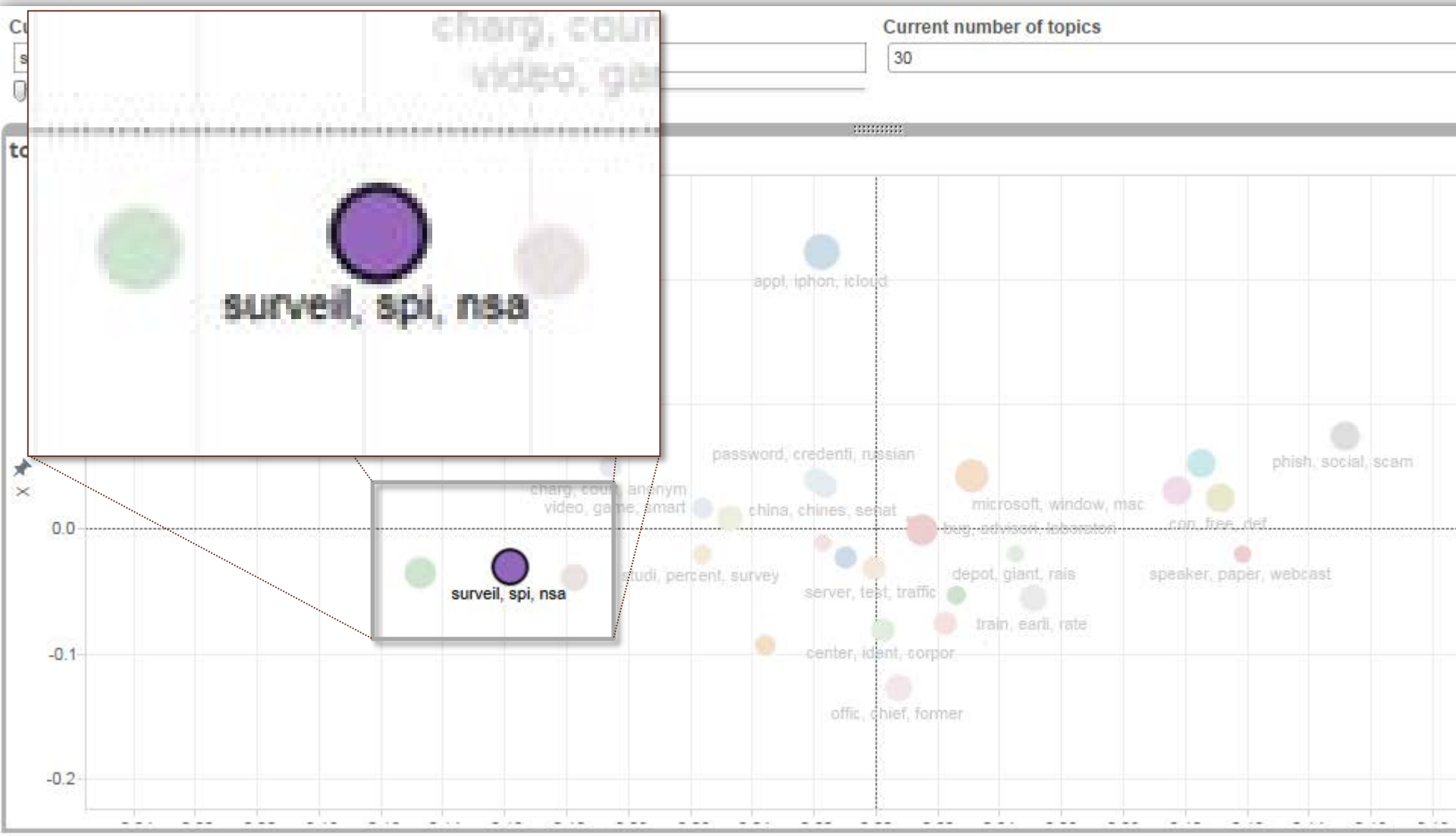
september, 2014

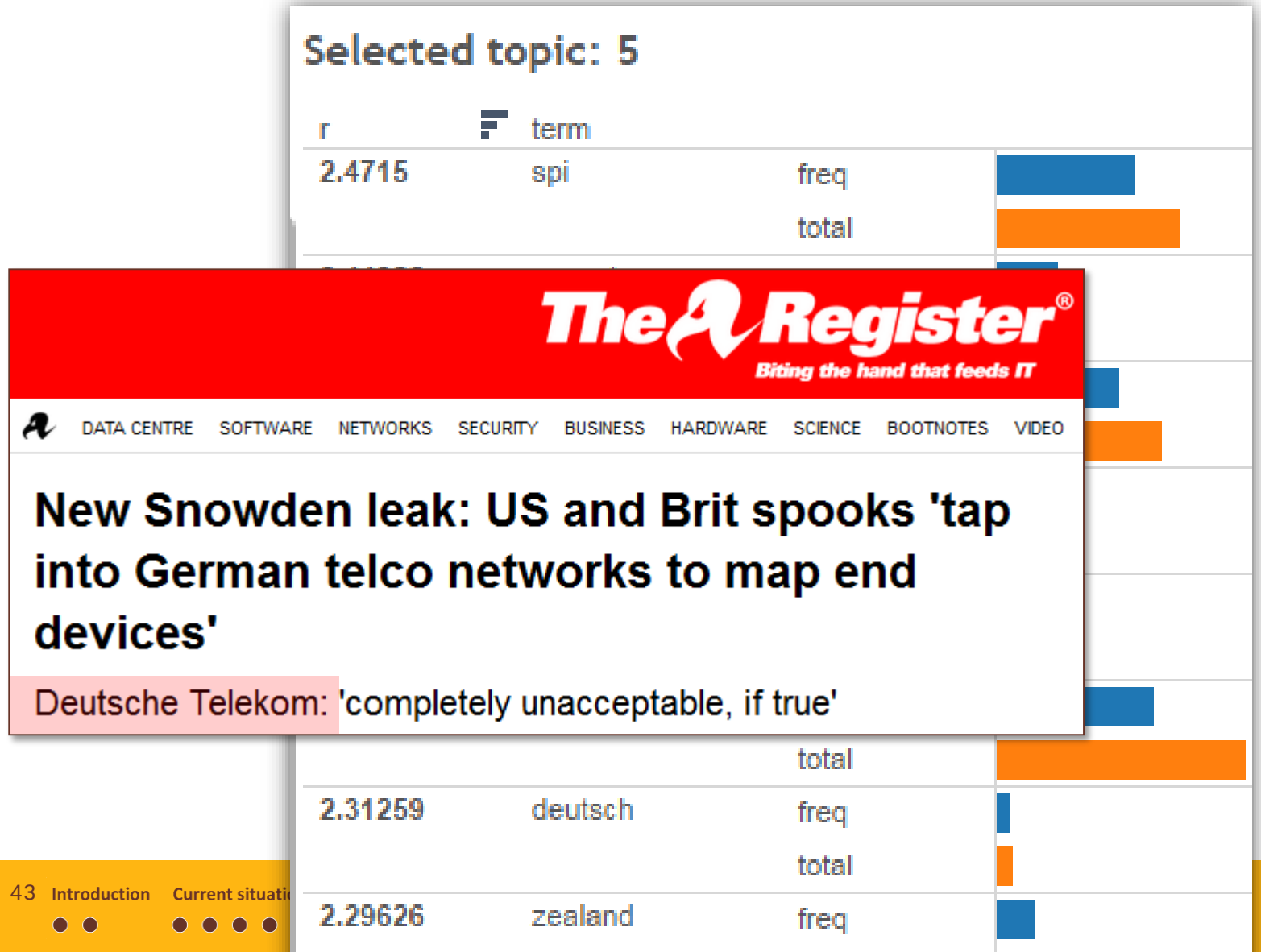
Current number of topics

30

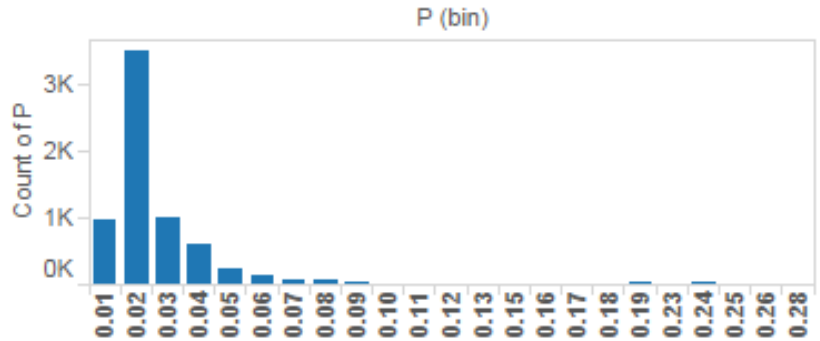
topic overview



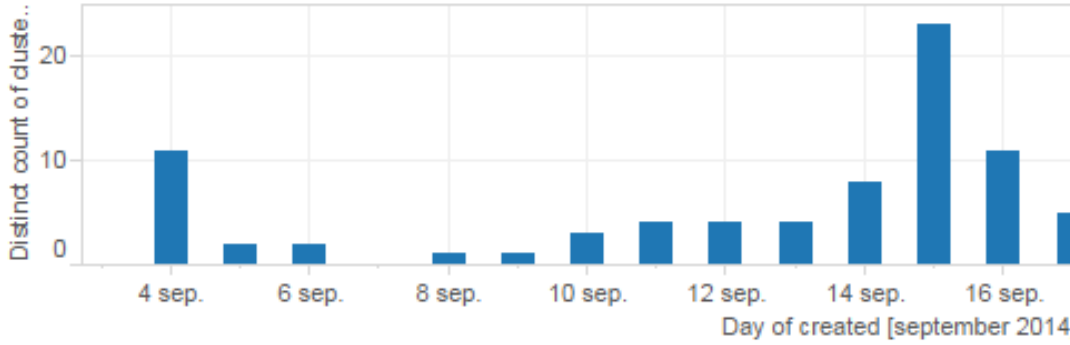

Current number of topics
30
surveil, spi, nsa
appl, iphon, icloud
password, credenti, russian
phish, social, scam
china, chines, senat
microsoft, window, mac
depot, giant, rais
speaker, paper, webcast
server, test, traffic
train, earli, rate
center, ident, corpor
offic, chief, former
0.0
-0.1
-0.2
surveil, spi, nsa


Selected topic: 5
r
term
2.4715
spi
freq
total
total
2.31259
deutsch
freq
total
2.29626
zealand
freq
The Register®
Biting the hand that feeds IT
DATA CENTRE
SOFTWARE
NETWORKS
SECURITY
BUSINESS
HARDWARE
SCIENCE
BOOTNOTES
VIDEO
New Snowden leak: US and Brit spooks 'tap into German telco networks to map end devices'
Deutsche Telekom: 'completely unacceptable, if true'

Distribution topic probability

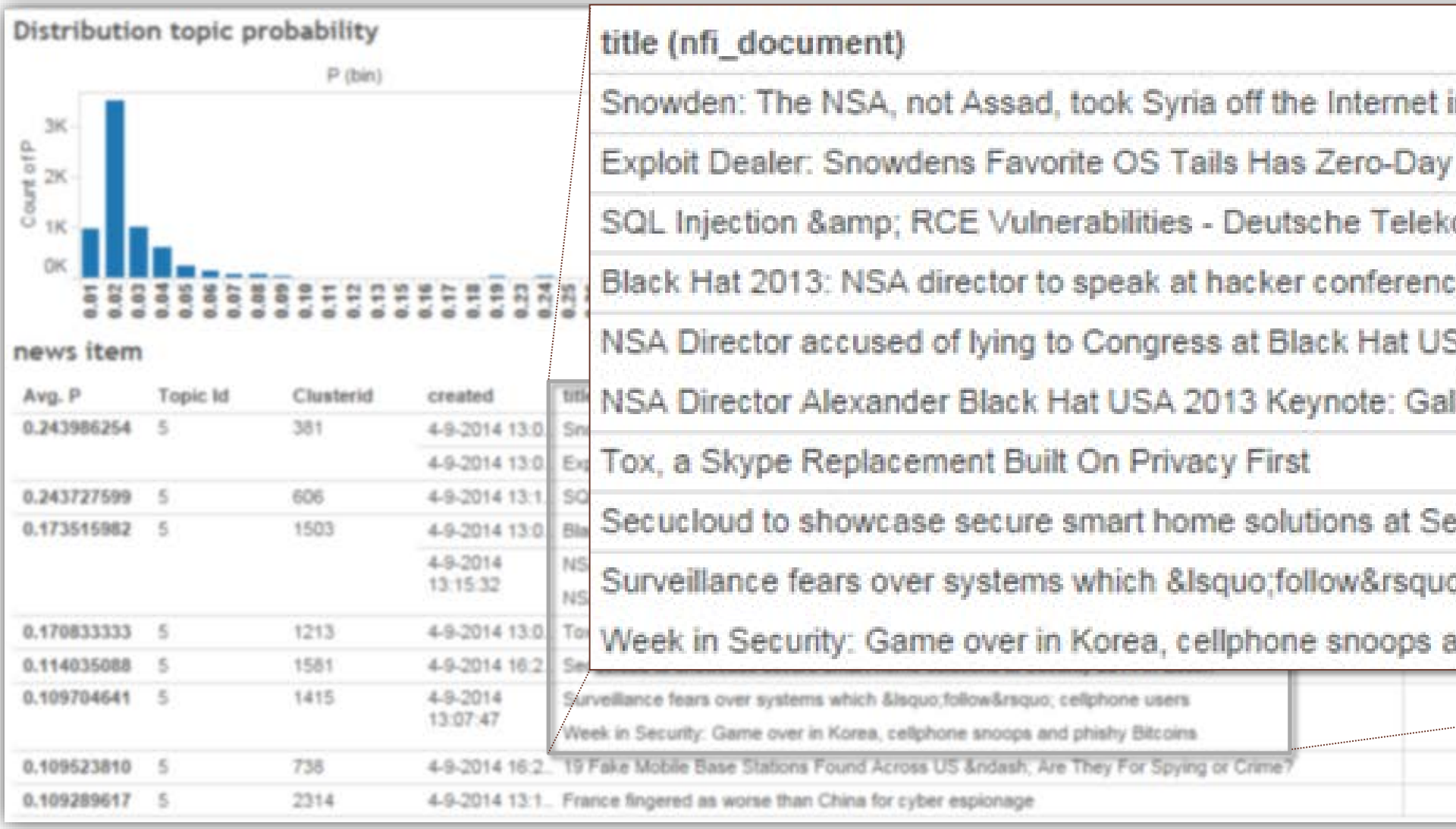


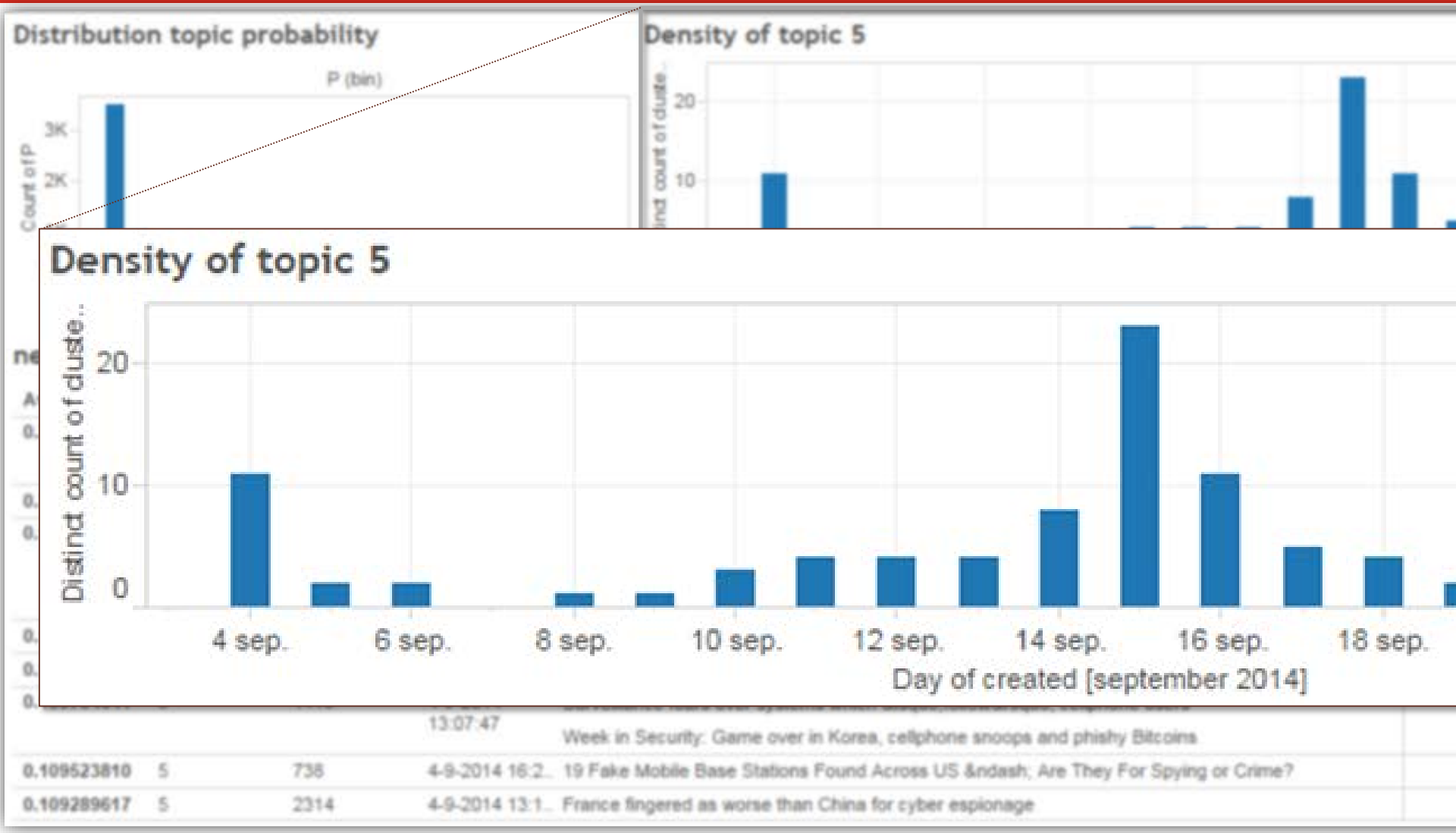
Density of topic 5

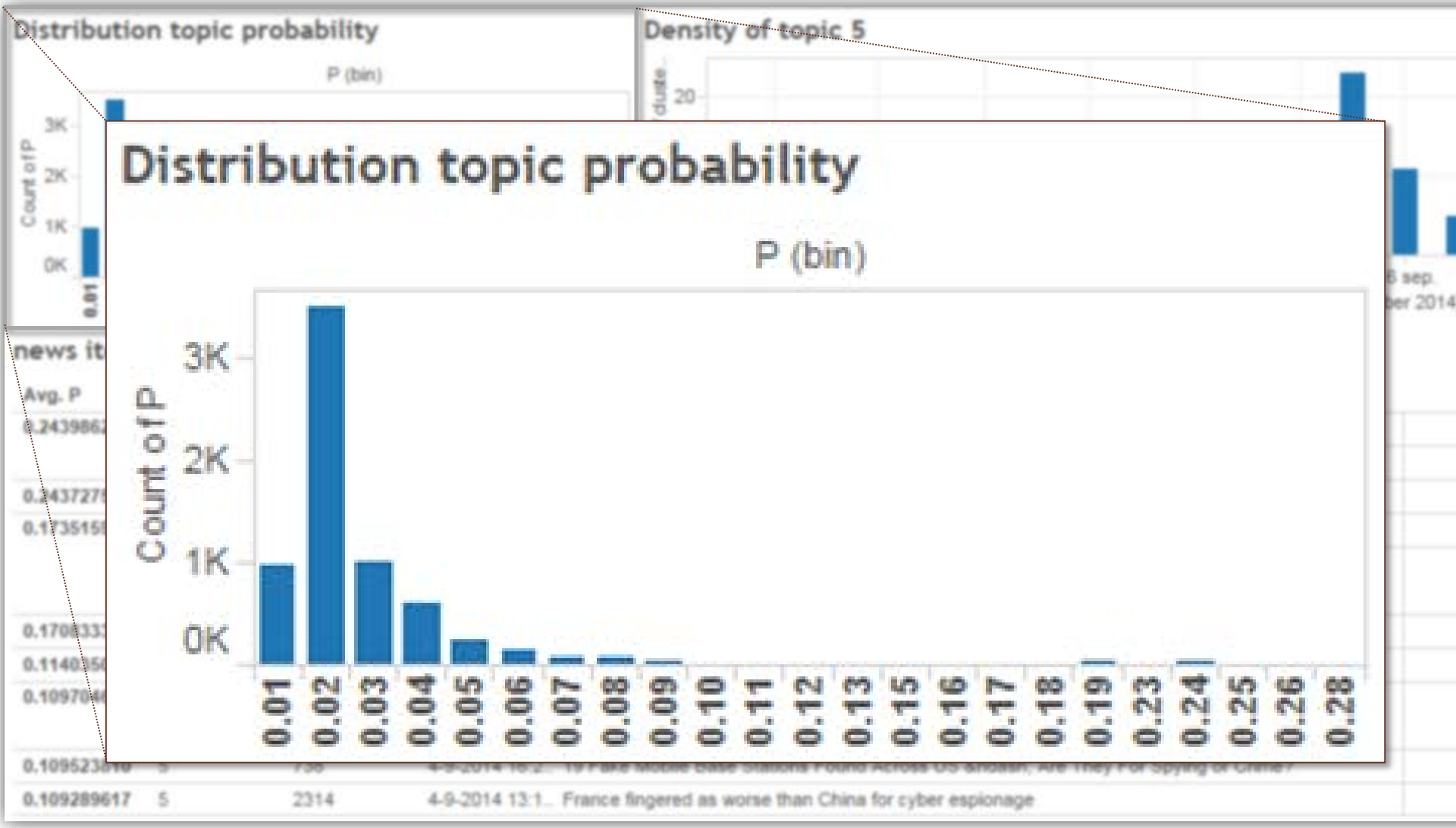


news item

Avg. P	Topic Id	Clusterid	created	title (nfi_document)
0.243986254	5	381	4-9-2014 13:0..	Snowden: The NSA, not Assad, took Syria off the Internet in 2012
			4-9-2014 13:0..	Exploit Dealer: Snowdens Favorite OS Tails Has Zero-Day Vulnerabilities Lurking Inside
0.243727599	5	606	4-9-2014 13:1..	SQL Injection & RCE Vulnerabilities - Deutsche Telekom Systems
0.173515982	5	1503	4-9-2014 13:0..	Black Hat 2013: NSA director to speak at hacker conference
			4-9-2014 13:15:32	NSA Director accused of lying to Congress at Black Hat USA 2013 keynote
				NSA Director Alexander Black Hat USA 2013 Keynote: Gallery
0.170833333	5	1213	4-9-2014 13:0..	Tox, a Skype Replacement Built On Privacy First
0.114035088	5	1581	4-9-2014 16:2..	Secucloud to showcase secure smart home solutions at Security 2014 in Essen
0.109704641	5	1415	4-9-2014 13:07:47	Surveillance fears over systems which ‘follow’ cellphone users
				Week in Security: Game over in Korea, cellphone snoops and phishy Bitcoins
0.109523810	5	738	4-9-2014 16:2..	19 Fake Mobile Base Stations Found Across US – Are They For Spying or Crime?
0.109289617	5	2314	4-9-2014 13:1..	France fingered as worse than China for cyber espionage


Distribution topic probability
P (bin)
Count of P
3K
2K
1K
0K
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.15
0.16
0.17
0.18
0.19
0.23
0.24
0.25
news item
Avg. P
Topic Id
Clusterid
created
0.243986254
5
381
4-9-2014 13:0
4-9-2014 13:0
0.243727599
5
606
4-9-2014 13:1
0.173515982
5
1503
4-9-2014 13:0
4-9-2014 13:15:32
0.170833333
5
1213
4-9-2014 13:0
0.114035088
5
1581
4-9-2014 16:2
0.109704641
5
1415
4-9-2014 13:07:47
0.109523810
5
738
4-9-2014 16:2
0.109289617
5
2314
4-9-2014 13:1
title (nfi_document)
Snowden: The NSA, not Assad, took Syria off the Internet
Exploit Dealer: Snowdens Favorite OS Tails Has Zero-Day
SQL Injection & RCE Vulnerabilities - Deutsche Telek
Black Hat 2013: NSA director to speak at hacker conferenc
NSA Director accused of lying to Congress at Black Hat US
NSA Director Alexander Black Hat USA 2013 Keynote: Gal
Tox, a Skype Replacement Built On Privacy First
Secucloud to showcase secure smart home solutions at Se
Surveillance fears over systems which ‘follow&rsquo
Week in Security: Game over in Korea, cellphone snoops a
Surveillance fears over systems which ‘follow’ cellphone users
Week in Security: Game over in Korea, cellphone snoops and phishy Bitcoins
19 Fake Mobile Base Stations Found Across US – Are They For Spying or Crime?
France fingered as worse than China for cyber espionage


Distribution topic probability
P (bin)
Count of P
3K
2K
Density of topic 5
Distinct count of duste...
20
10
0
4 sep.
6 sep.
8 sep.
10 sep.
12 sep.
14 sep.
16 sep.
18 sep.
Day of created [september 2014]
13:07:47
Week in Security: Game over in Korea, cellphone snoops and phishy Bitcoins
0.109523810 5 738 4-9-2014 16:2.. 19 Fake Mobile Base Stations Found Across US – Are They For Spying or Crime?
0.109289617 5 2314 4-9-2014 13:1.. France fingered as worse than China for cyber espionage


Distribution topic probability
P (bin)
Count of P
3K
2K
1K
0K
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.15
0.16
0.17
0.18
0.19
0.23
0.24
0.25
0.26
0.28

Conclusions

Conclusions

- Overall
  - Promising
  - We already trained our system (without knowing we did…)
  - Interactive approach leads to quick development of knowledge

- Proof-of-Concept
  - Stories found are understandable and relevant
  - Linking items to existing dossiers is successful
  - Determining new dossiers is difficult

- To-do
  - Quantitative evaluation
  - Decide how to move further
  - Possibly work towards a working system based on the mid-level requirements

Nationaal Cyber Security Centrum
Ministerie van Veiligheid en Justitie