Exploratory Analysis in Big Data based on PCA and PLS

José Camacho

Departamento de Teoría de la Señal, Telemática y Comunicaciones

Network Engineering & Security Group
http://nesg.ugr.es

Universidad de Granada

Velocity

Veracity

Sensors

Text

Records

Images

Audio

Video

Biological

Big Data

VOLUME

Variety

Numbers

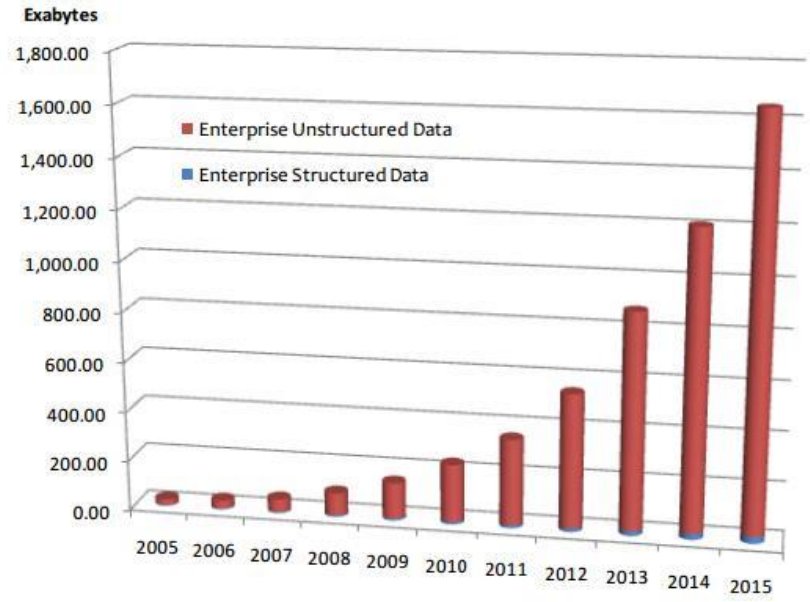
Unit	Size	Size
Gigabyte	10^9	1.000.000.000 bytes
Terabyte	10^{12}	1.000.000.000.000 bytes
Petabyte	10^{15}	1.000.000.000.000.000 bytes
Exabyte	10^{18}	1.000.000.000.000.000.000 bytes
Zettabyte	10^{21}	1.000.000.000.000.000.000.000 bytes

- *There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing*

RJMetrics co-founder and CEO Robert J. Moore

Structured

```
#ip_tc_trojan, ip_tc_worm, ip_tc_adware, ip_tc_botnet, ip_tc_scanner, ip_tc_maliciousDocument, ip_tc_maliciousWebpage, ip_t
0, 0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,18,2,0,18,0,0,0,0,2,0,0,0,0,0,0,0,18,2,18,0,0,0,18,2,0,0,0,0,0,0,2,0,18,0,0,0,0,0,
32774, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32781, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32786, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32788, 0,0,0
32789, 0,0
22, 0,0,0,
25, 0,0,0,0,0
34, 0,0,0,0,0,0,0,0,0,0,0,
32805, 0,0,0,0,0,0,0,0,0,0,0
32811, 0,0,0,0,0,0,0,0,0,0,0
32814, 0,0,0,0,0,0,0,0,0,0,0
32820, 0,0,0,0,0,0,0,0,0,0,0
53, 0,0,0,0,0,0,0,0,0,0,0,
32823, 0,0,0,0,0,0,0,0,0,0,0
32833, 0,0,0,0,0,0,0,0,0,0,0
32841, 0,0,0,0,0,0,0,0,0,0,0
32842, 0,0,0,0,0,0,0,0,0,0,0
80, 0,0,0,0,0,0,0,0,0,0,0,
32857, 0,0,0,0,0,0,0,0,0,0,0
32858, 0,0,0,0,0,0,0,0,0,0,0
32859, 0,0,0,0,0,0,0,0,0,0,0
21, 0,0,0,0,0,0,0,0,0,0,0,
32790, 0,0,0,0,0,0,0,0,0,0,0
32906, 0,0,0,0,0,0,0,0,0,0,0
139, 0,0,0,0,0,0,0,0,0,0,0
32913, 0,0,0,0,0,0,0,0,0,0,0
32949, 0,0,0,0,0,0,0,0,0,0,0
32952, 0,0,0,0,0,0,0,0,0,0,0
```



,0,1,0,0,0,0,0,0,0,0,1
,0,0,0,1,0,0,0,0,0,0,1
,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,
,0,0,0,0,0,0,0,0,0,0,0

x UGR x

Unstructured

juna pregunta.

cada en ETSIIT ha sido modificado.

.es o llamando al N/D.

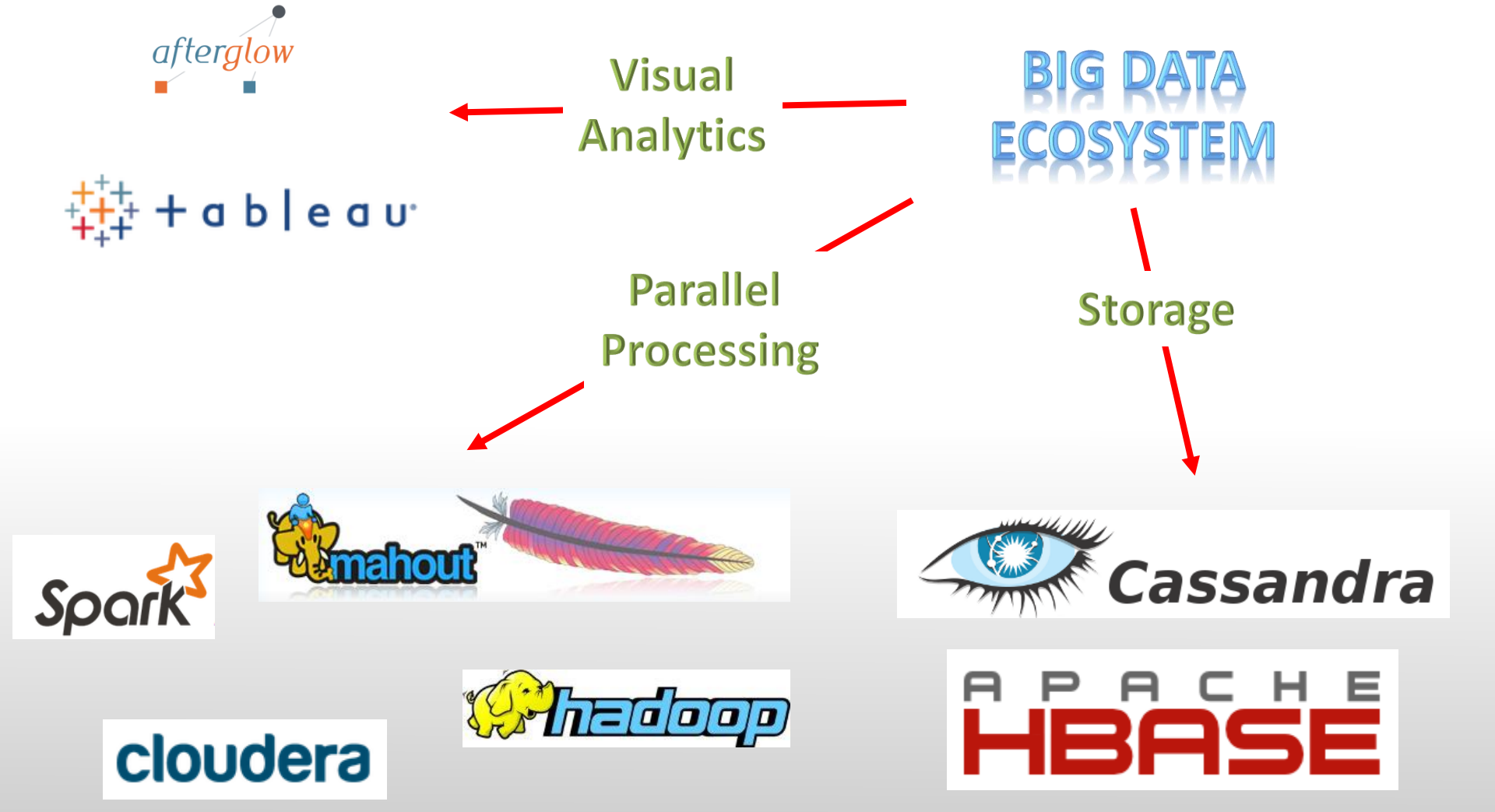
R - ETSIIT - Sistema de Reservas de Aulas en:

Michael Walker, Data Science Central

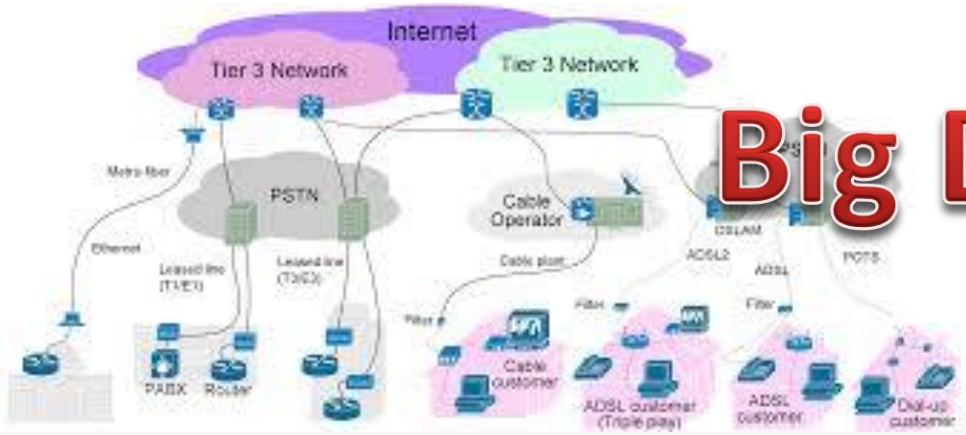
Variety

Reserva modificada							
Reserva #	Fecha Inicial	Fecha Final	Recurso	Hora de Inicio	Hora de Finalización	Ubicación	Contacto
sc1557544861fc22	10/06/2015	10/06/2015	Sala de juntas	13:00	14:00	ETSIIT	N/D

WHAT'S BIG DATA?



- Problem: Handling communication networks



Big Data

- Solution: Multivariate Analysis??
  - Networkmetrics

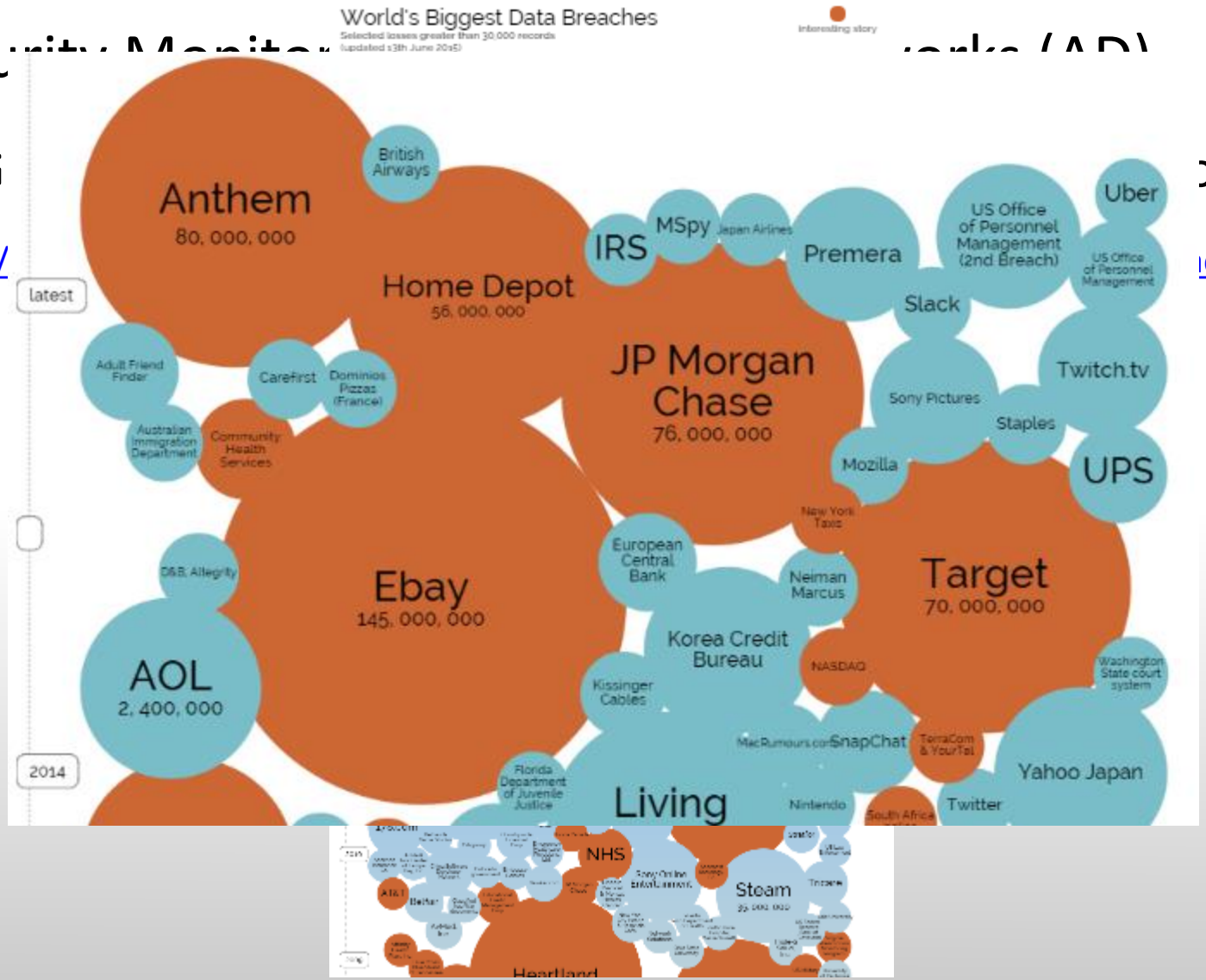
Networkmetrics

- Applications for Exploratory Analysis, Optimization, Classification, Anomaly Detection (≅ **Chemometrics**)
- Big Data by Definition (≠**Chemometrics**)
  - 4 V's: Tons of data, high speed, from lots of sources, many false alarms….
  - Mostly unstructured ➔ Feature Engineering
- Complex Data (≅ **Chemometrics**):
  - Fusion
  - High dimensional
  - N-way

- Secu[illegible]

  - G[illegible])

http://[illegible]cks/

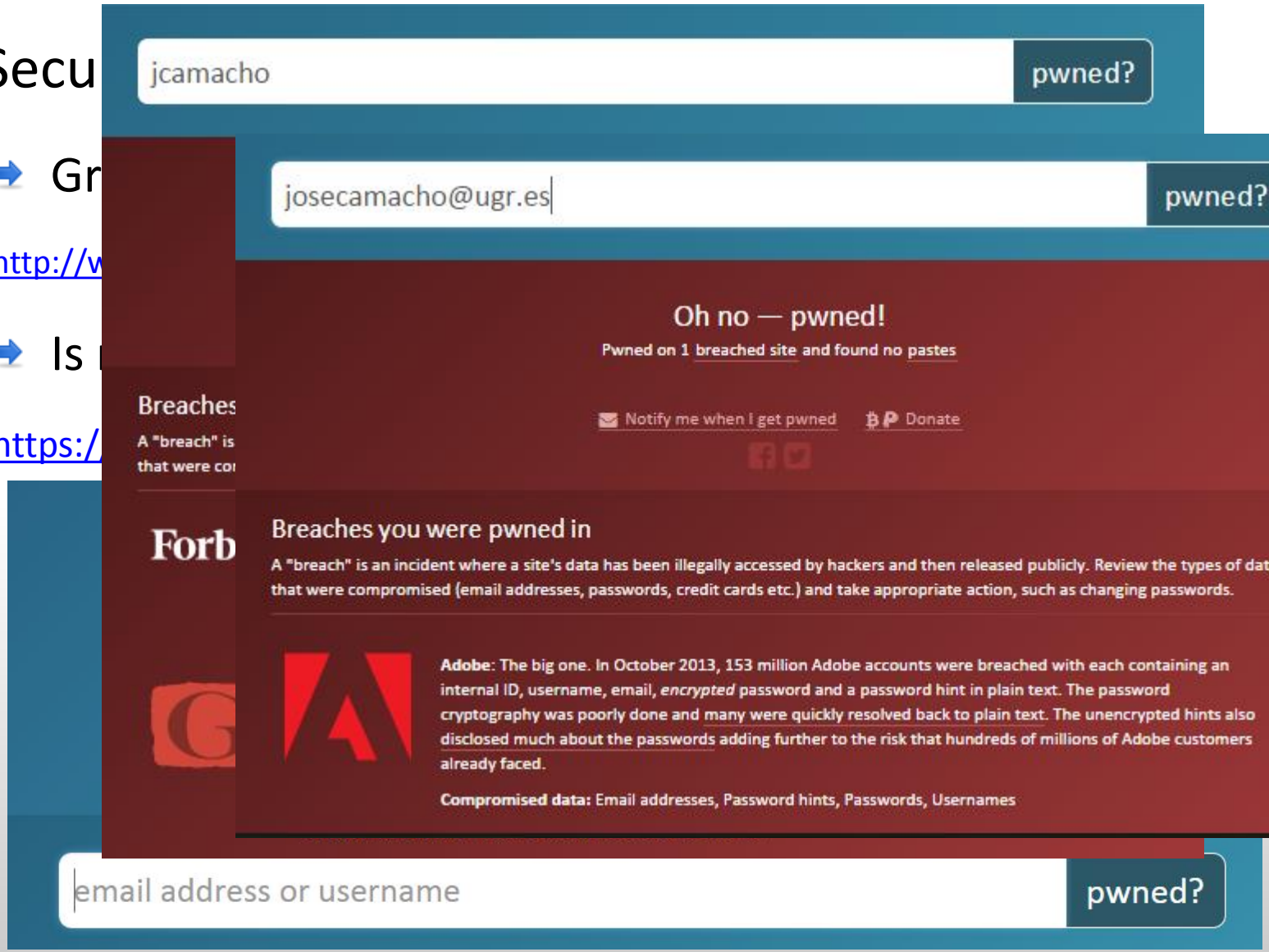


- Secu
  - Gr

http://v

- Is

https:/

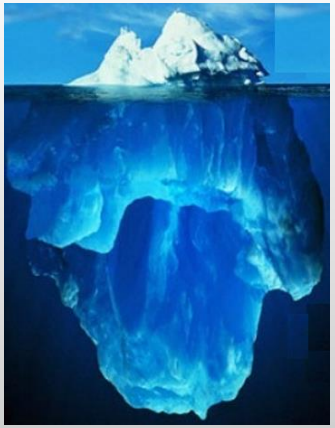


- Security Monitoring in Computer Networks (AD)
  - Growing tendency of Security Breaches (Leakage of Info)

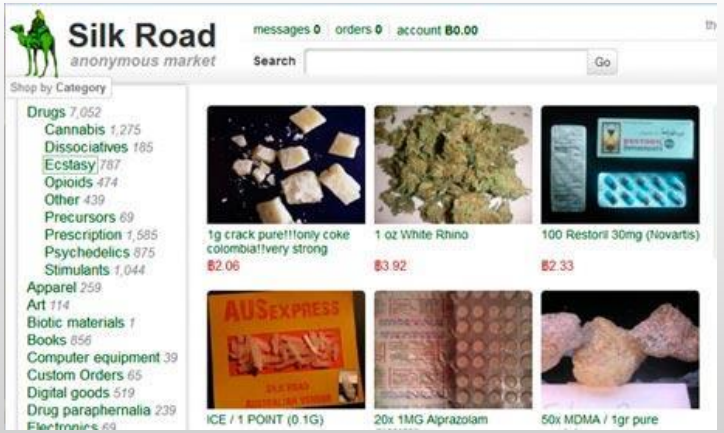
http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

  - Is my info at risk?

https://haveibeenpwned.com/



DEEP WEB
(96 %)

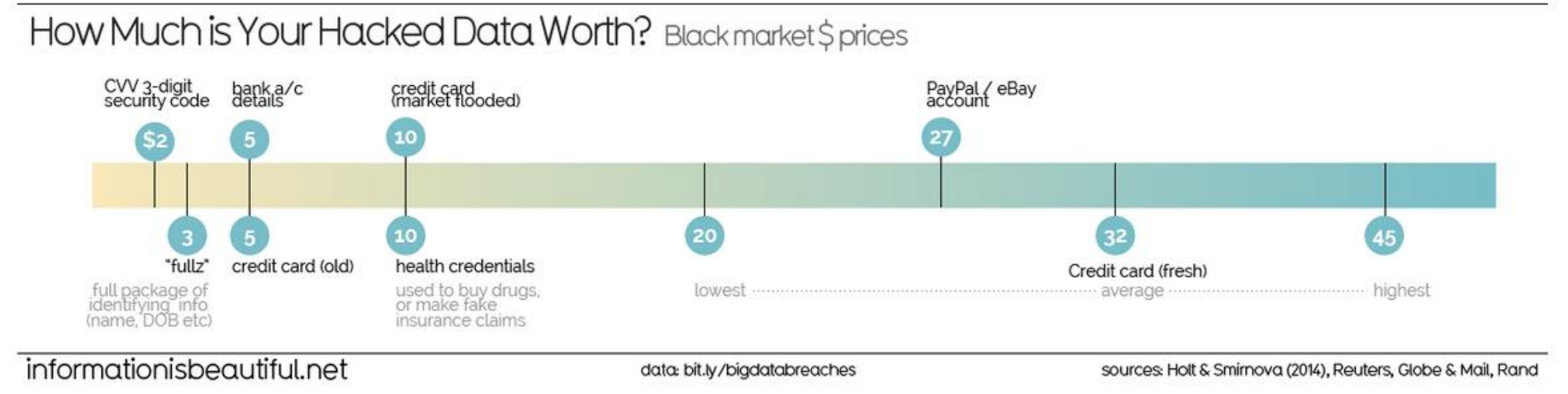


- Security Monitoring in Computer Networks (AD)
  - Growing tendency of Security Breaches (Leakage of Info)

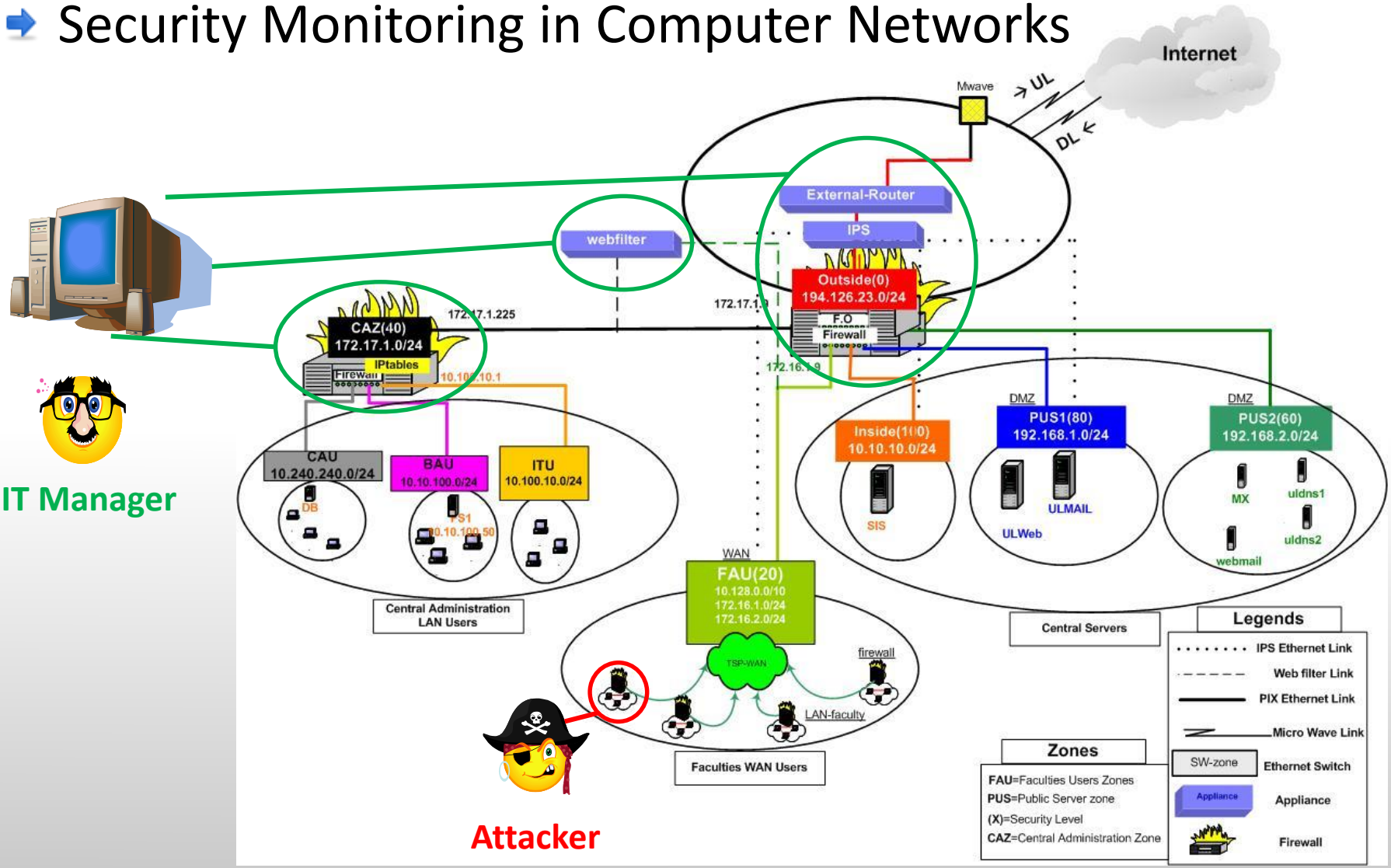
http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

  - Is my info at risk?

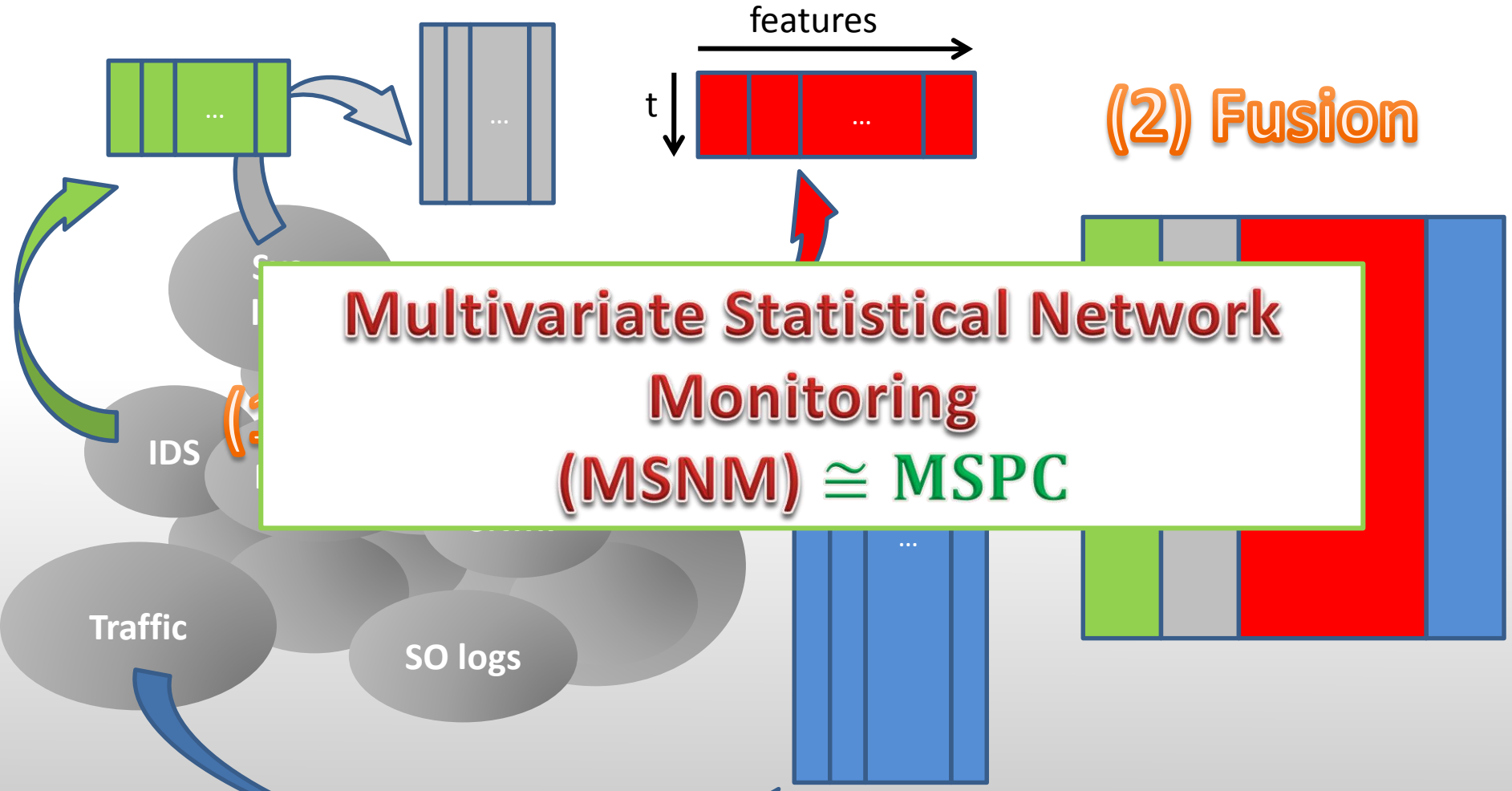
https://haveibeenpwned.com/



Security Monitoring in Computer Networks

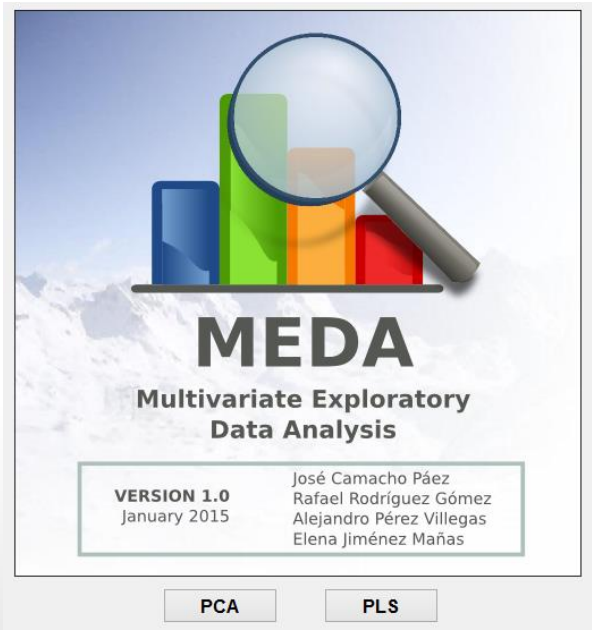


How may Multivariate Analysis help???



- MEDA Toolbox: https://github.com/josecamachop/MEDA-Toolbox

  - Dimensionality:
    - Scree plots
    - CV (ekf, cekf, ckf)
    - SVI Plots
  - Structure in Variables:
    - Loading plots
    - MEDA plots
  - Distribution of Observations
    - Score plots
    - MSPC: D-st, Q-st
    - Covariance MSPC: ADICOV
  - Observations vs Variables
    - oMEDA plots

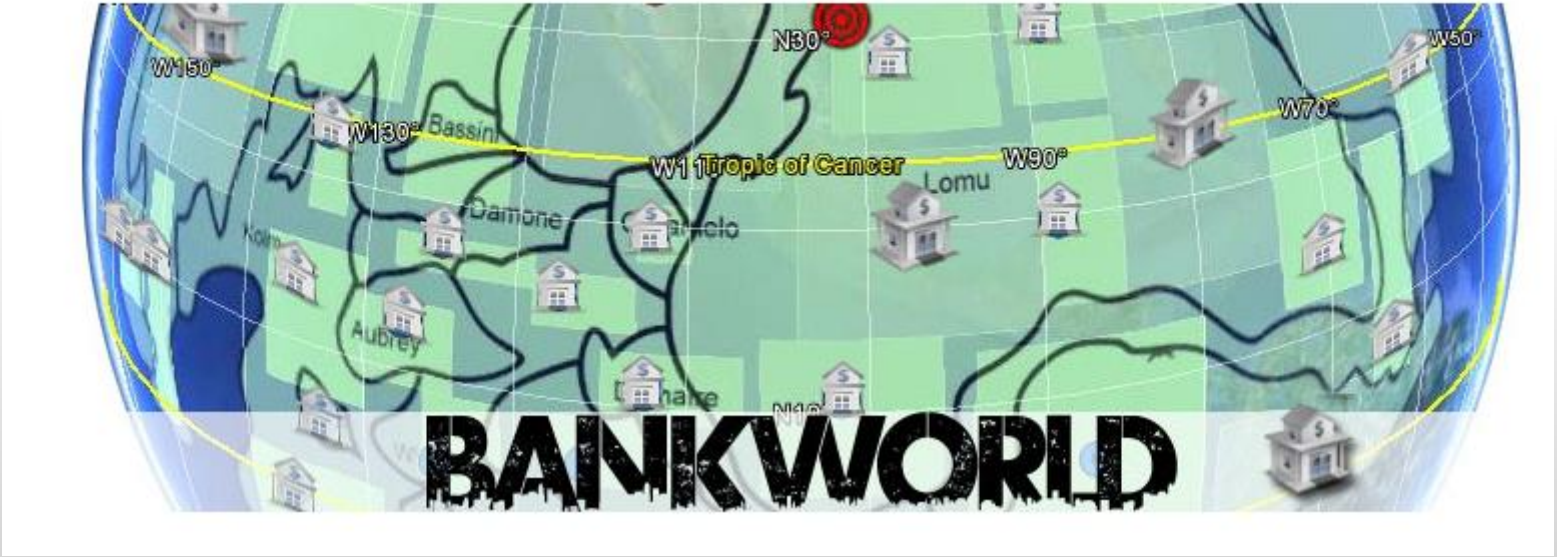


ChemoLab, (2015) 143: 49

- Networkmetrics: MSNM

VAST Challenge 2012

Visual Analytics Community
DISCOVERY THROUGH COLLABORATION AND SHARING



MULTIVARIATE EXPLORATORY DATA ANALYSIS

198.41.0.4
128.9.0.107
192.33.4.12
128.8.10.90
192.203.230.10
192.5.5.241

VAST 2012
Challenge 2
nk of Money
work Architecture

Structured:
23,711,341
records

Unstructured:
35,984 records

10.32.0.201-210,
10.32.1.100,
10.32.1.201-206
10.32.5.X

Websites

(Excluding Server Range)

172.23.229.X

Workstations

Financial Servers

IDS

Log Server

Domain Controller / DNS

- Parsing & Data Fusion
  - X = 2,350 obs x 90 features

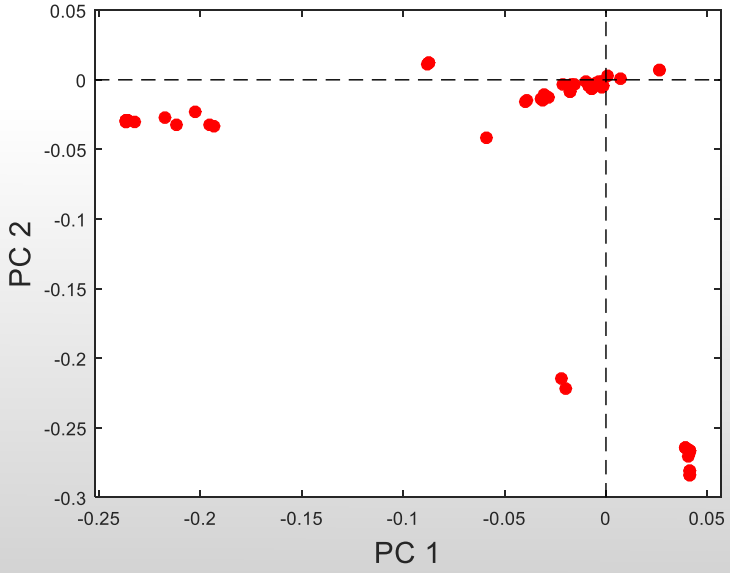
	Feature class	#features
Firewall log	Syslog priority	5
	Operation	6
	Message code	25
	Protocol	3
	IP address	9
	Port number	17
	Direction	2
	Conn. built/teardown	2
	Subtotal	**69**
IDS log	IP address	9
	Port number	17
	Alert class	5
	Priority	3
	Text label	9
	Subtotal	**43**

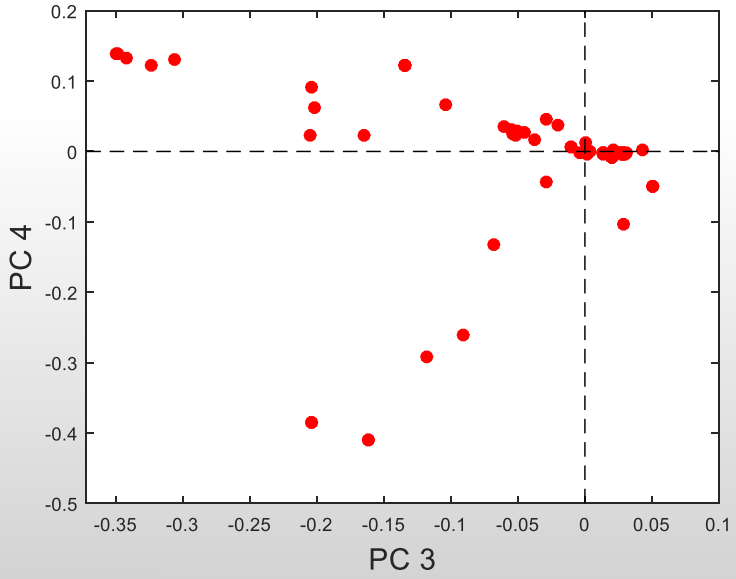
PCA Model

6 PCs and 60% variance app.

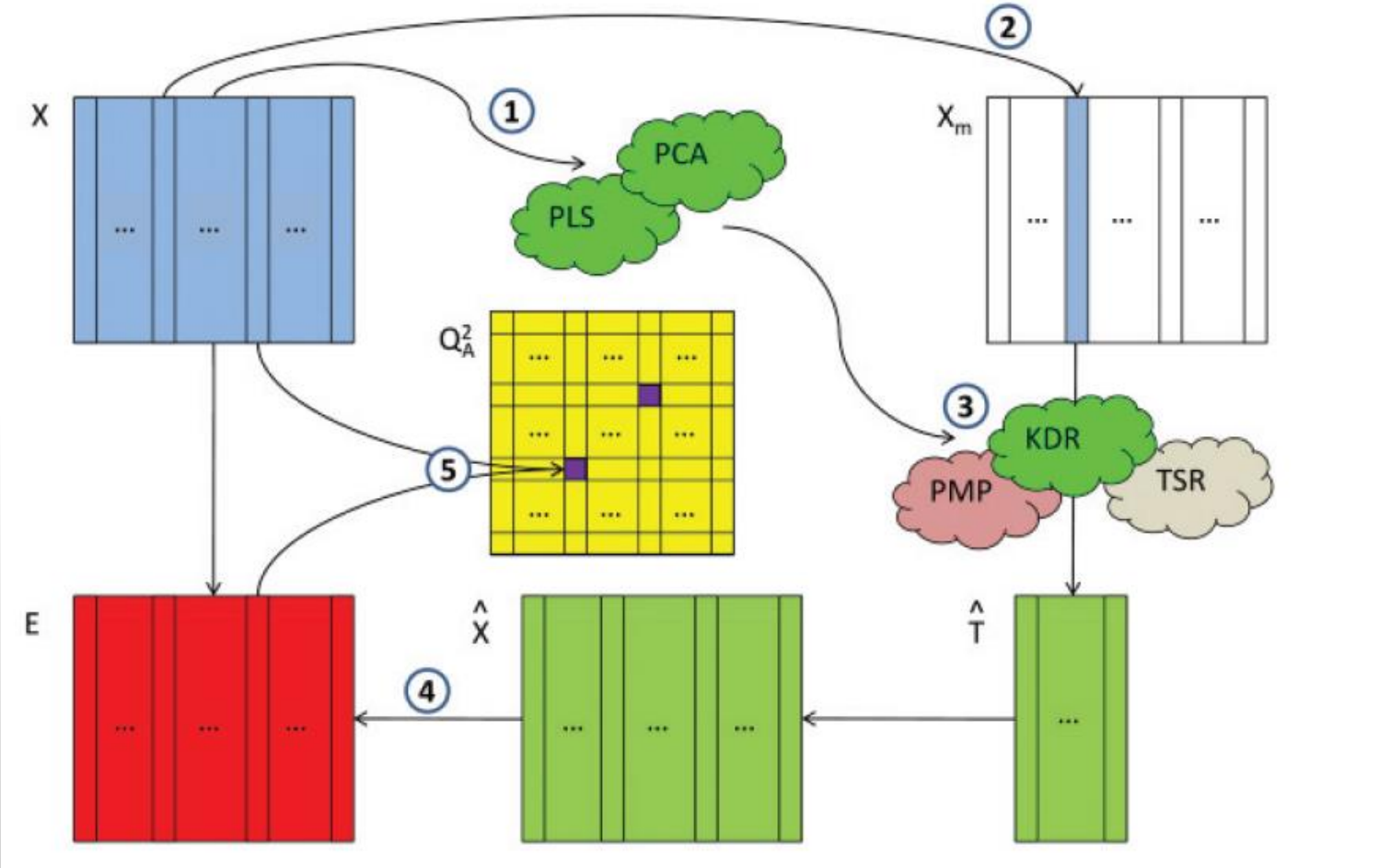
Loading plots

- LP may lead to non-existen correlations, MEDA not
- LP limited to 2/3 PCs, MEDA not
- LP difficult to interpret for many variables

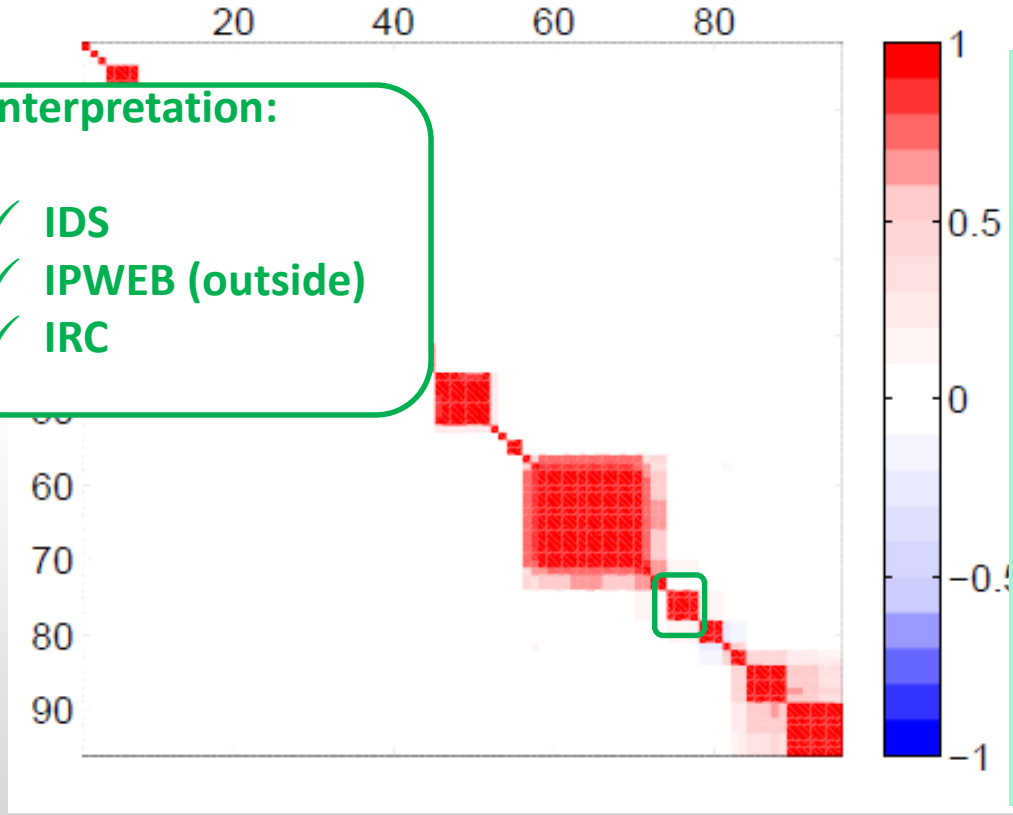




- MEDA:

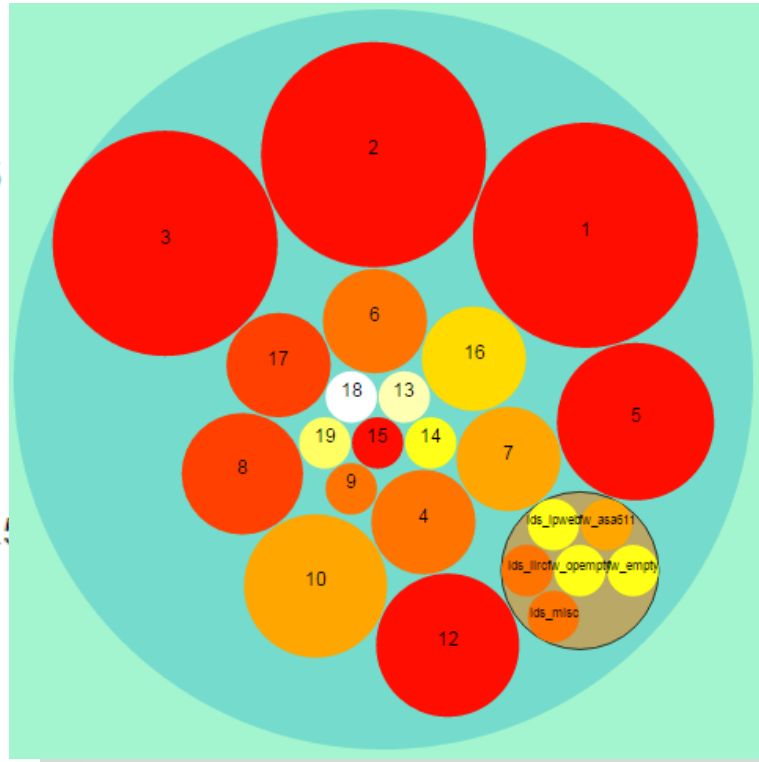


- MEDA: 6 PCs



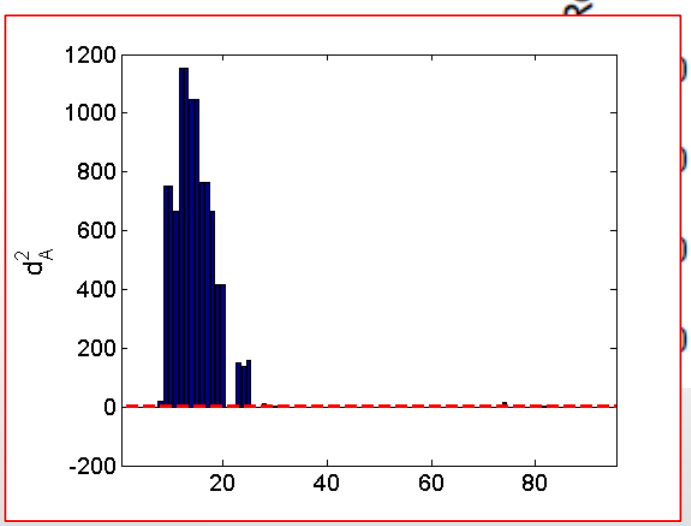
Interpretation:

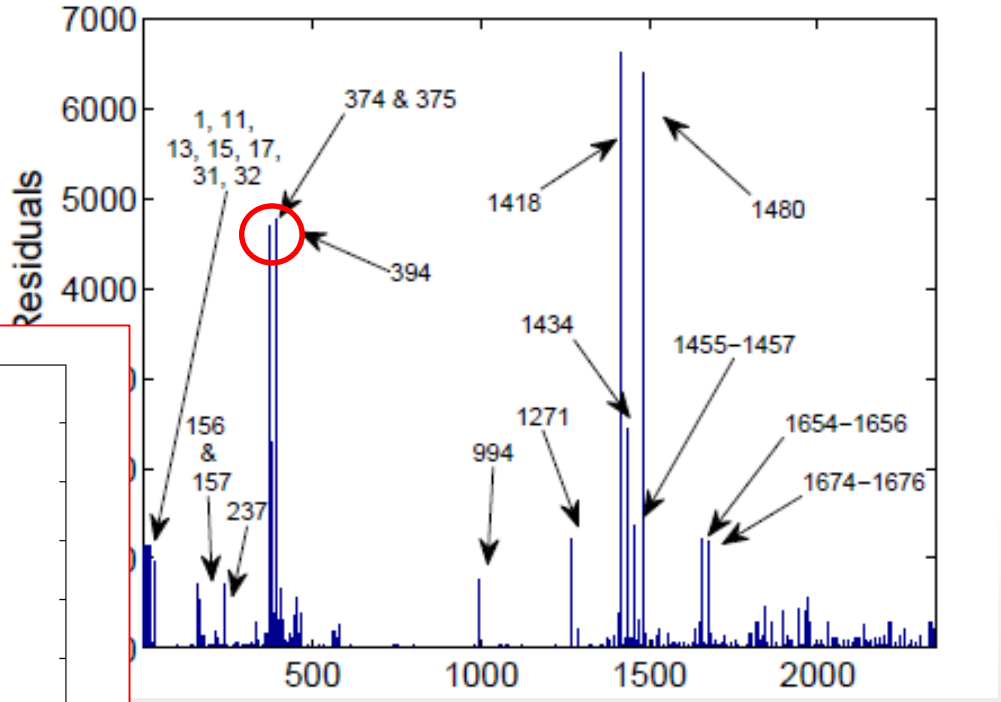
- ✓ **IDS**
- ✓ **IPWEB (outside)**
- ✓ **IRC**



http://nesg.ugr.es/index.php/es/resultados

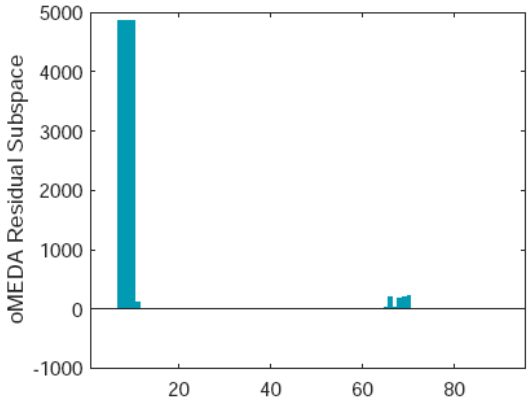
- MSPC Q-st
- oMEDA

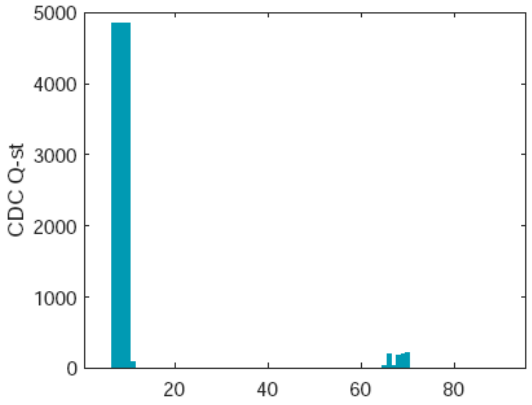


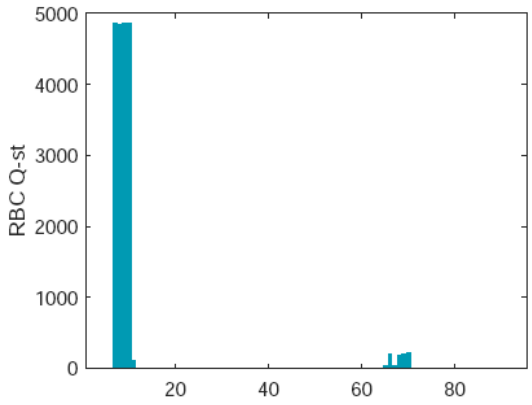


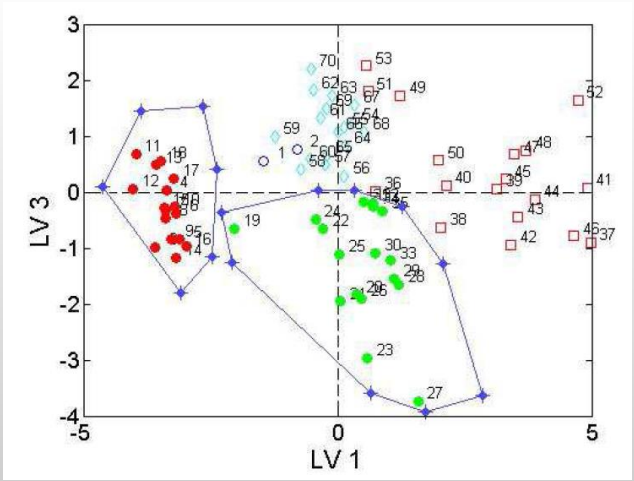
- Interpretation: Non-legitime requests of services in access firewall.

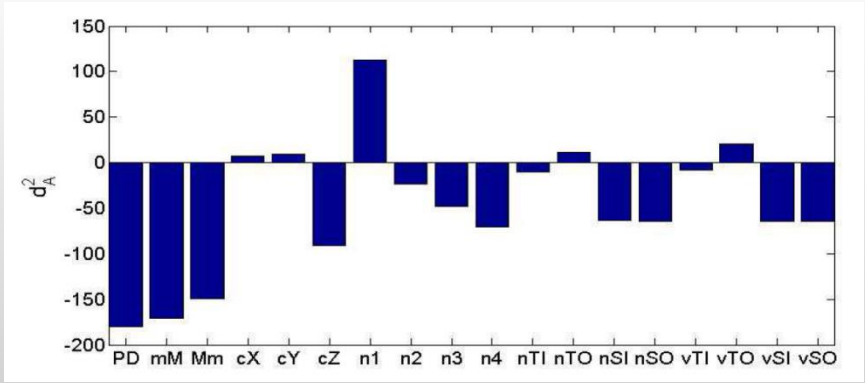
oMEDA: Generalization of Contribution plots



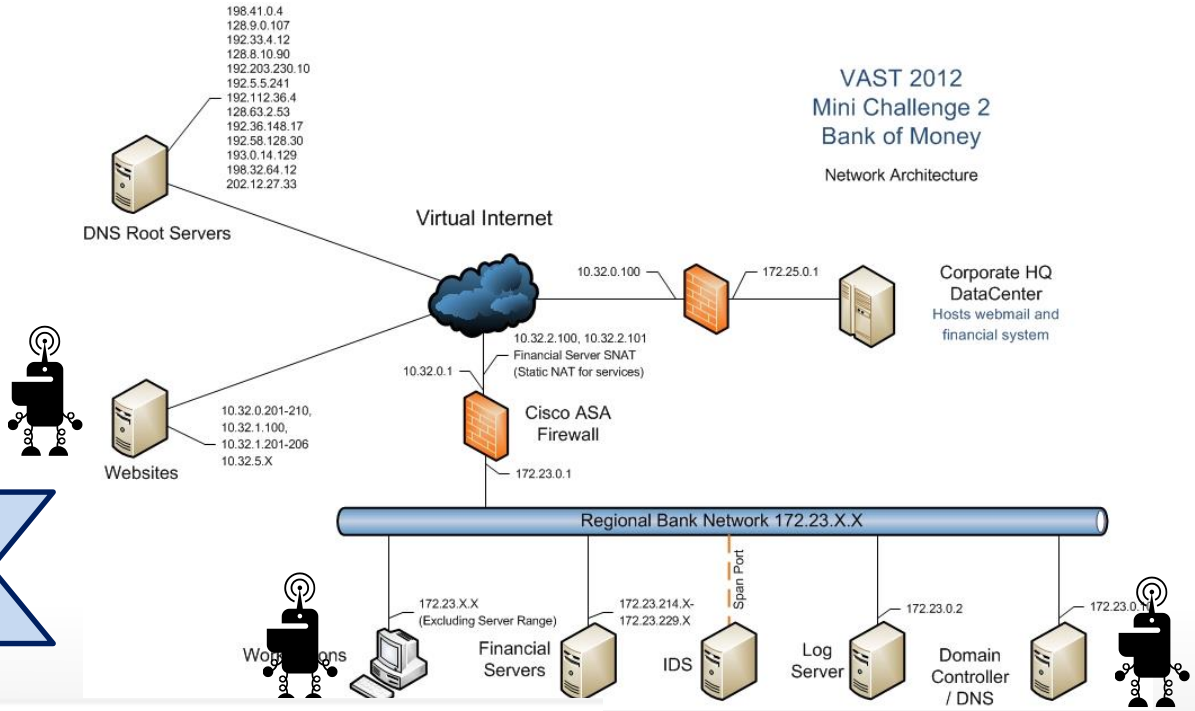




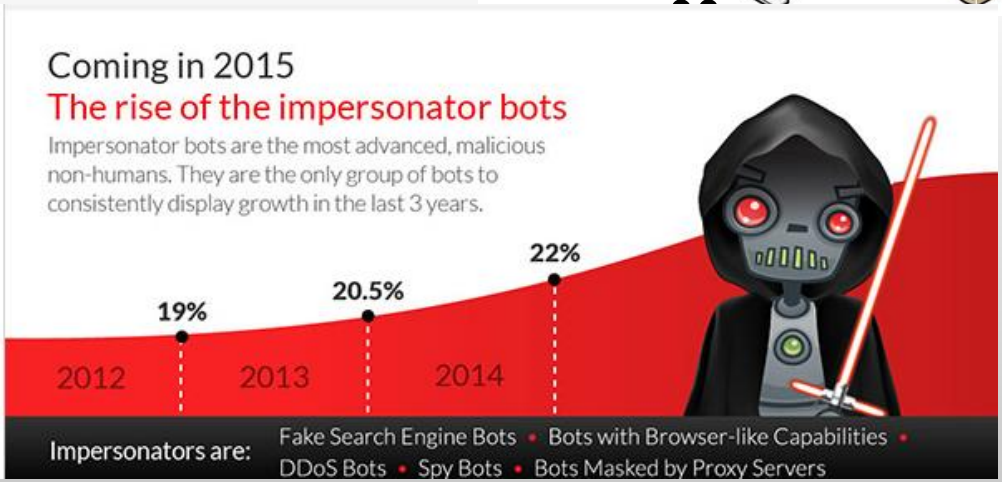




- Conclusion

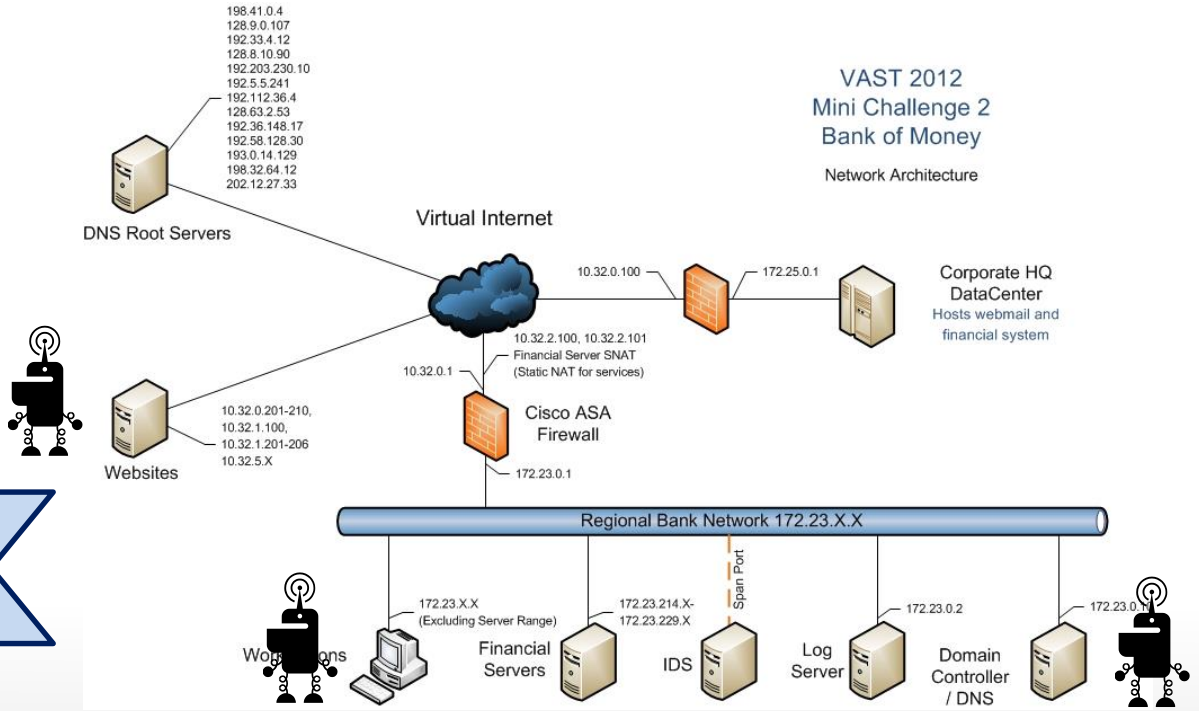


BOTNET



Incapsula Botnet Report. 2014

- Conclusion



BOTNET

Anomaly #	Proposed	Winner [30]	[31]	[32]
Attacks to the DNS / Domain Controller	X			
Access attempts to the firewall	X	P	X	
FTP attempts from inner nodes to outer node	X			X
Background IRC activity	X	X	X	
Parsing errors	X			

"Tackling the Big Data 4 Vs for Anomaly Detection". *INFOCOM'2014 Workshop on Security and Privacy in Big Data*. 2014.

Extensions for Big Data

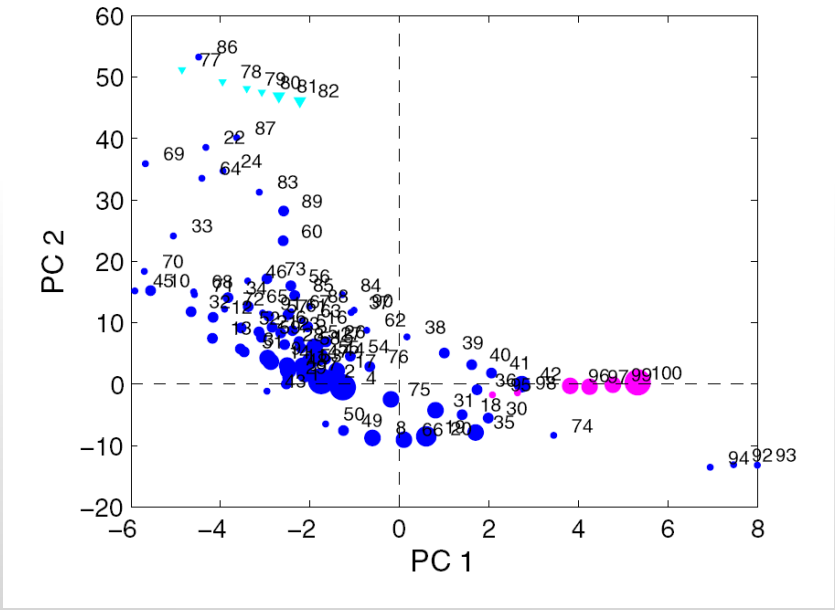
- For variables ➔ Kernel matrix with EWMA update

$$(X'X)_t = \lambda \cdot (X'X)_{t-1} + \tilde{X}_t{}' \cdot \tilde{X}_t$$

- Scalable to any size
- PCA/PLS Kernel algorithms
- MEDA from covariance
- oMEDA from accumulates
- ADICOV MSPC

- For observations ➔ Clustering

- A link with visualization
- Compressed Score Plots
- Compressed MSPC



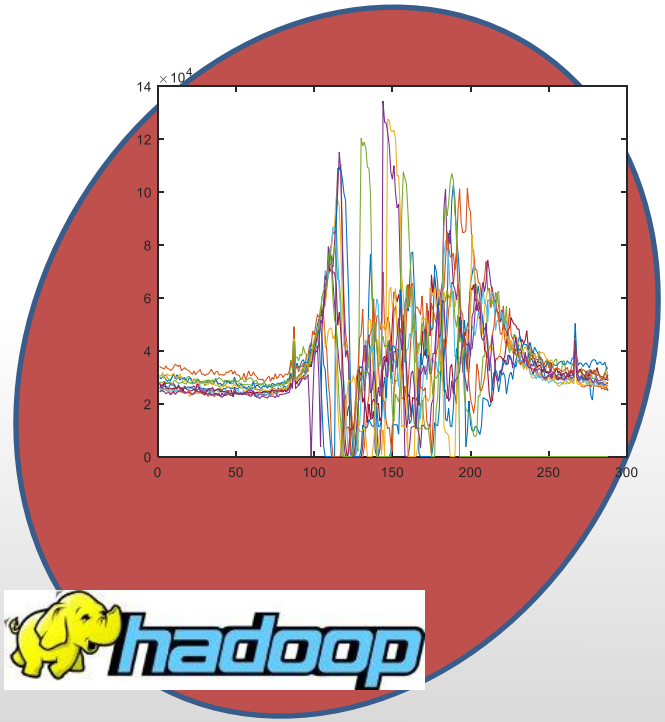
ChemoLab, (2014) 135: 110

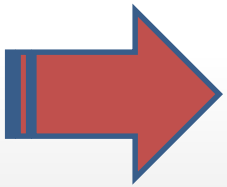
```
clear
load kdd

Lmodel = Lmodel_ini; % Initialization
Lmodel.update = 2; % Change this to 1 for EWMA and 2 for Iterative
Lmodel.type = 2; % Change this to 1 for PCA and 2 for PLS
Lmodel.lv = 3; % Initial number of LVs
Lmodel.prep = 2; % X-block prepr. 0: None, 1: Mean-center, 2: Auto-scaling
Lmodel.prepy = 2; % Y-block prepr. 0: None, 1: Mean-center, 2: Auto-scaling
Lmodel.nc = 100; % Number of clusters

lambda = 1-1e-4; % Forgetting factor in EWMA
step = 0.01;
```

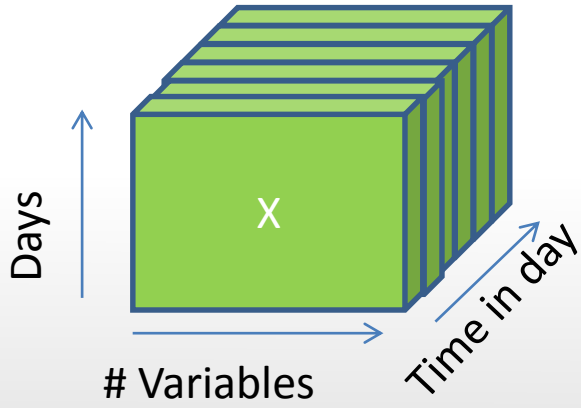
```
%% Model building (EWMA or Iterative)

if Lmodel.update == 1
    Lmodel = update_ewma(short_list,'',Lmodel,lambda,step,1); % EWMA
else
    Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
end
```

```
%% Data Analysis

if Lmodel.type==2, % for PLS

    % Score plot
    scores_Lpls(Lmodel,1:2);

    % MEDA
    map = meda_Lpls(Lmodel,1:2,0,3);
```

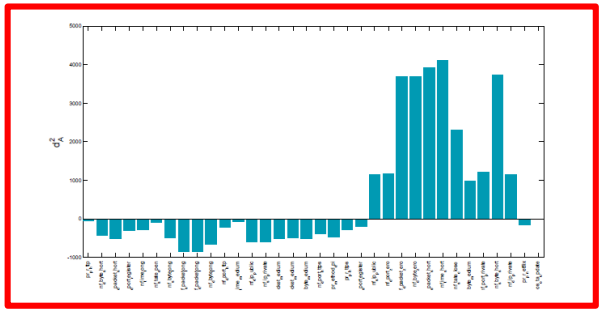
- Confidencial Network: 12 Working days

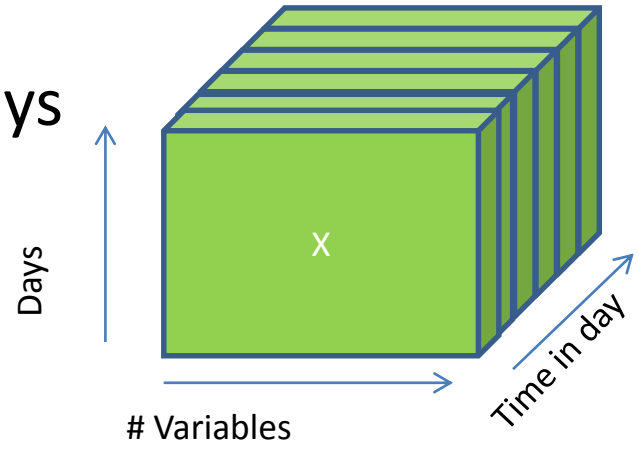


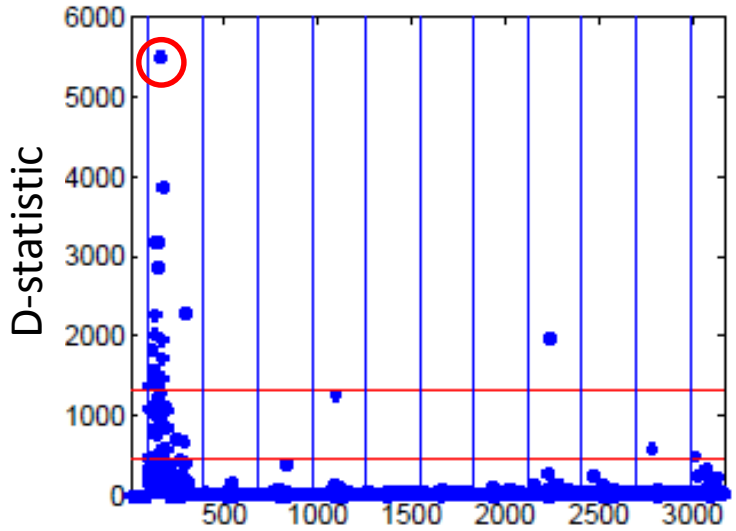


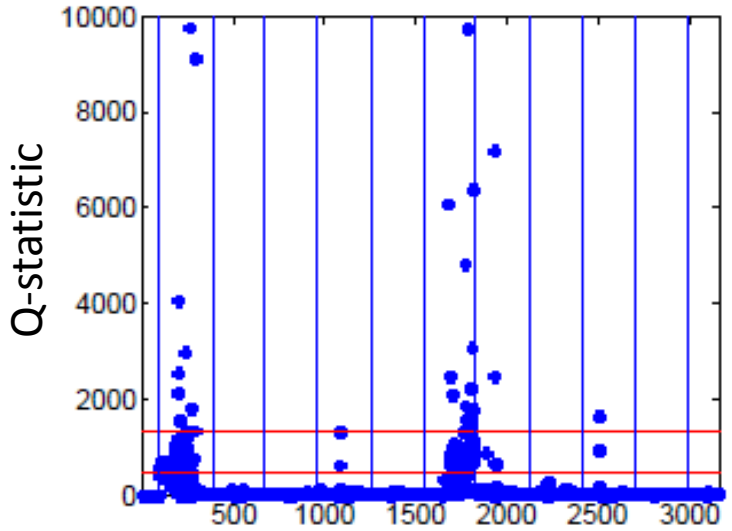


- Confidencial Network: 12 Working days

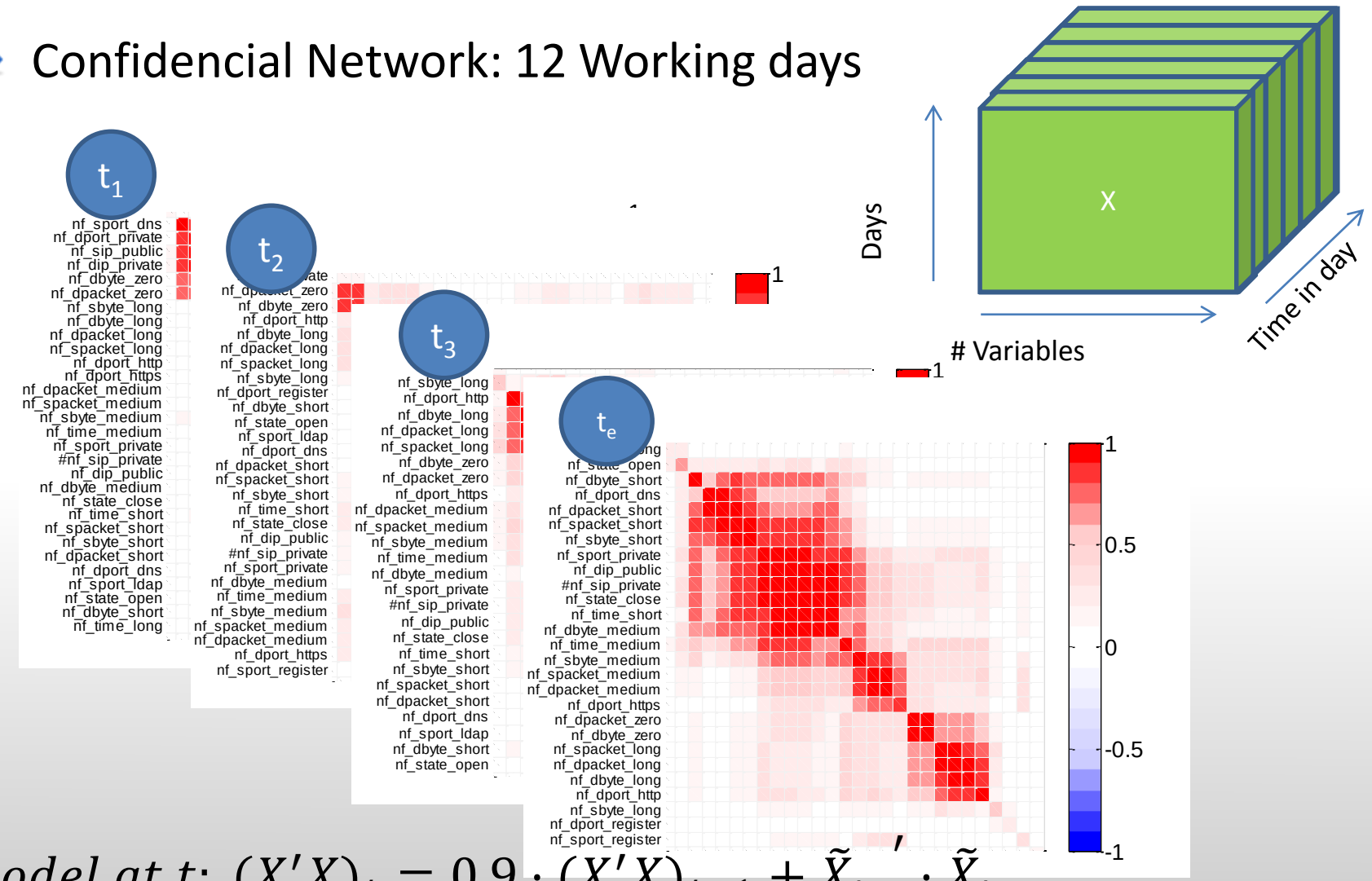






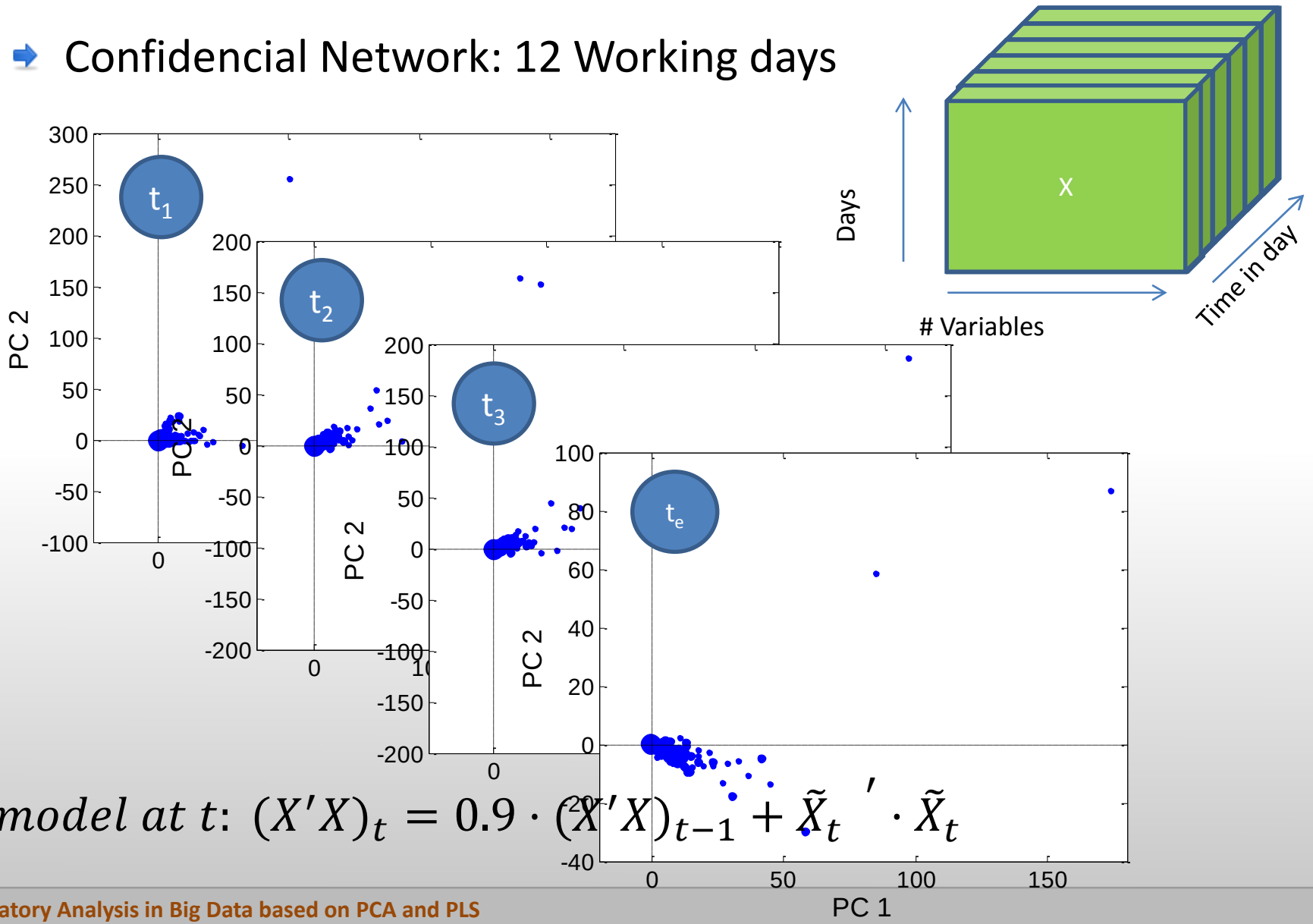


- Confidencial Network: 12 Working days



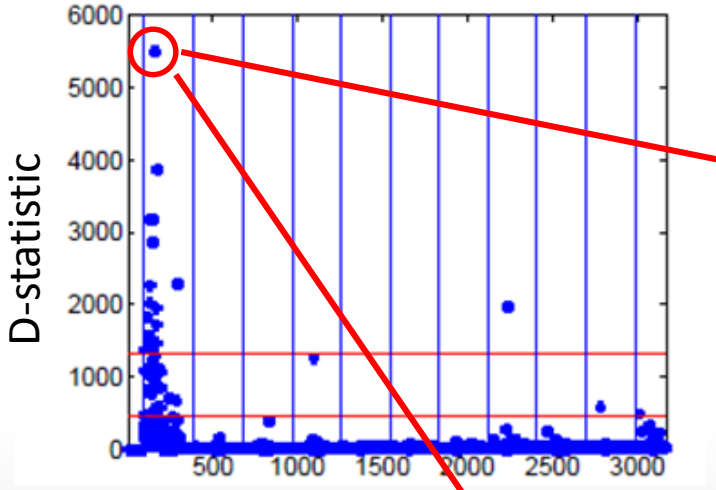
$$Lmodel\ at\ t\text{:}\ (X'X)_t = 0.9 \cdot (X'X)_{t-1} + \tilde{X}_t{}' \cdot \tilde{X}_t$$

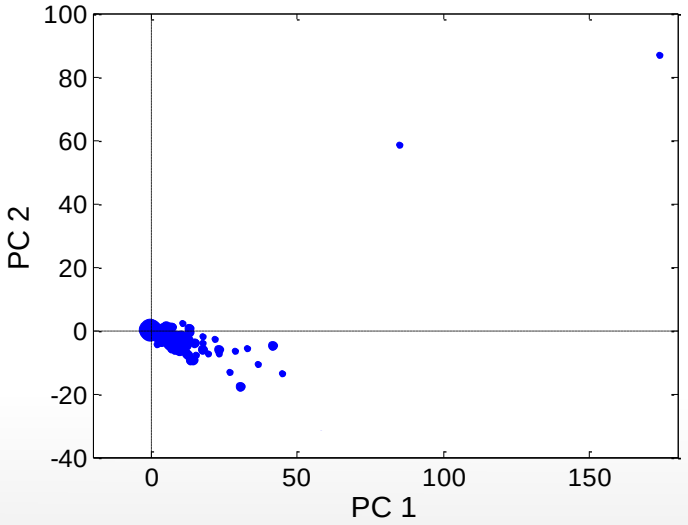
- Confidencial Network: 12 Working days

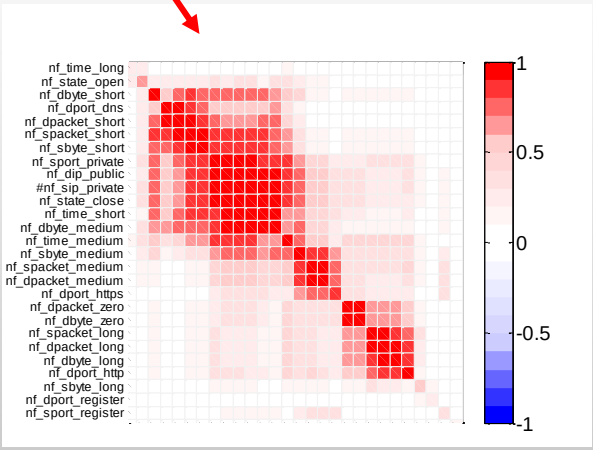


$$Lmodel\ at\ t\text{:}\ (X'X)_t = 0.9 \cdot (X'X)_{t-1} + \tilde{X}_t{}' \cdot \tilde{X}_t$$

- Confidencial Network: 12 Working days







- Multivariate Analysis has a lot to do in Big Data and Visual Analytics
- Multivariate Analysis tools can be extended to Networking for anomaly detection, optimization, …
  - with new and interesting particularities and challenges
  - with challenges already solved in chemometrics
- Who should lead this?

Exploratory Analysis in Big Data based on PCA and PLS

José Camacho

Departamento de Teoría de la Señal, Telemática y Comunicaciones

Network Engineering & Security Group
http://nesg.ugr.es

Universidad de Granada