Multivariate Big Data Analysis and its application to the Internet

NETWORKMETRICS

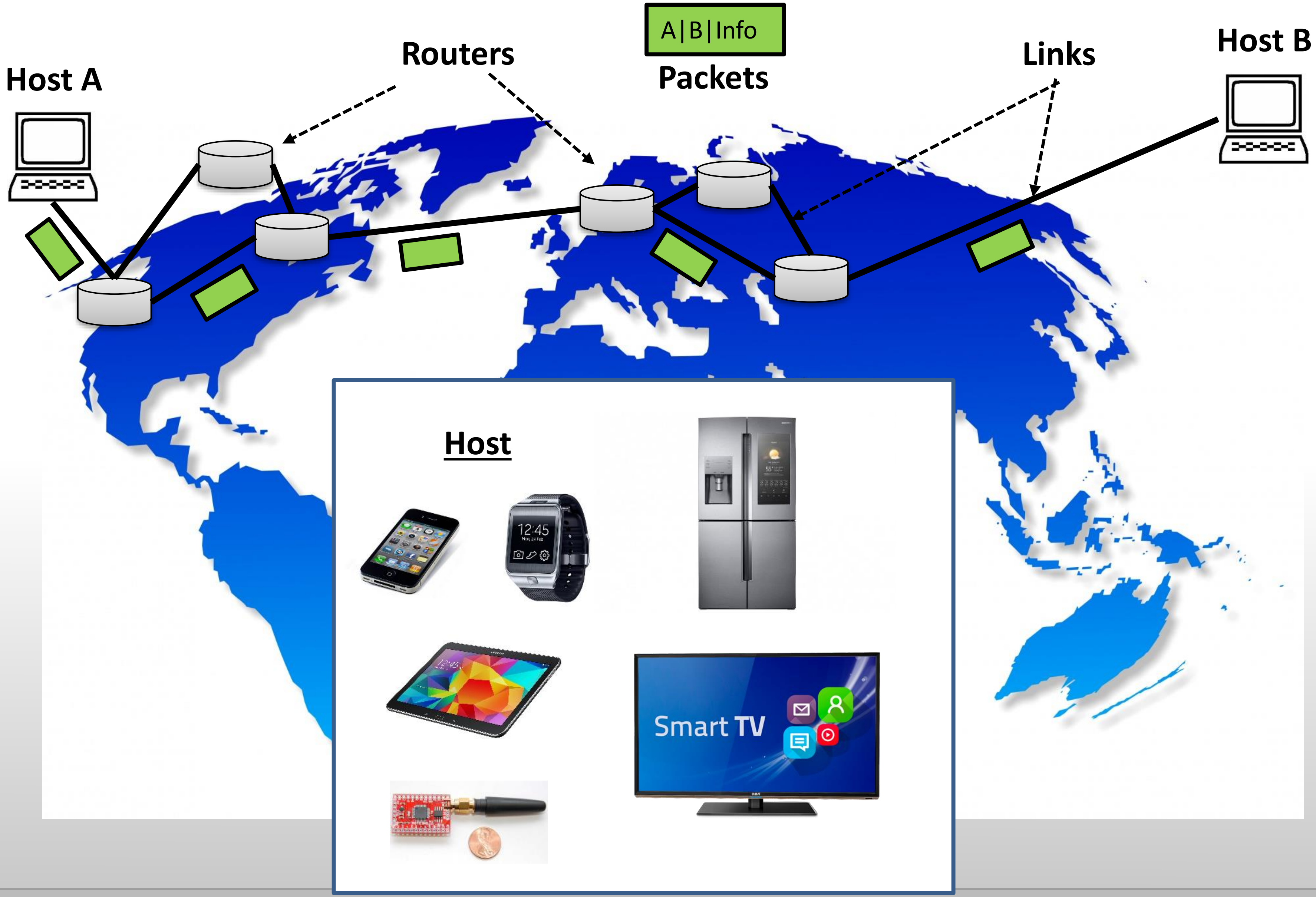
Network Engineering & Security Group
http://nesg.ugr.es

ugr Universidad de Granada

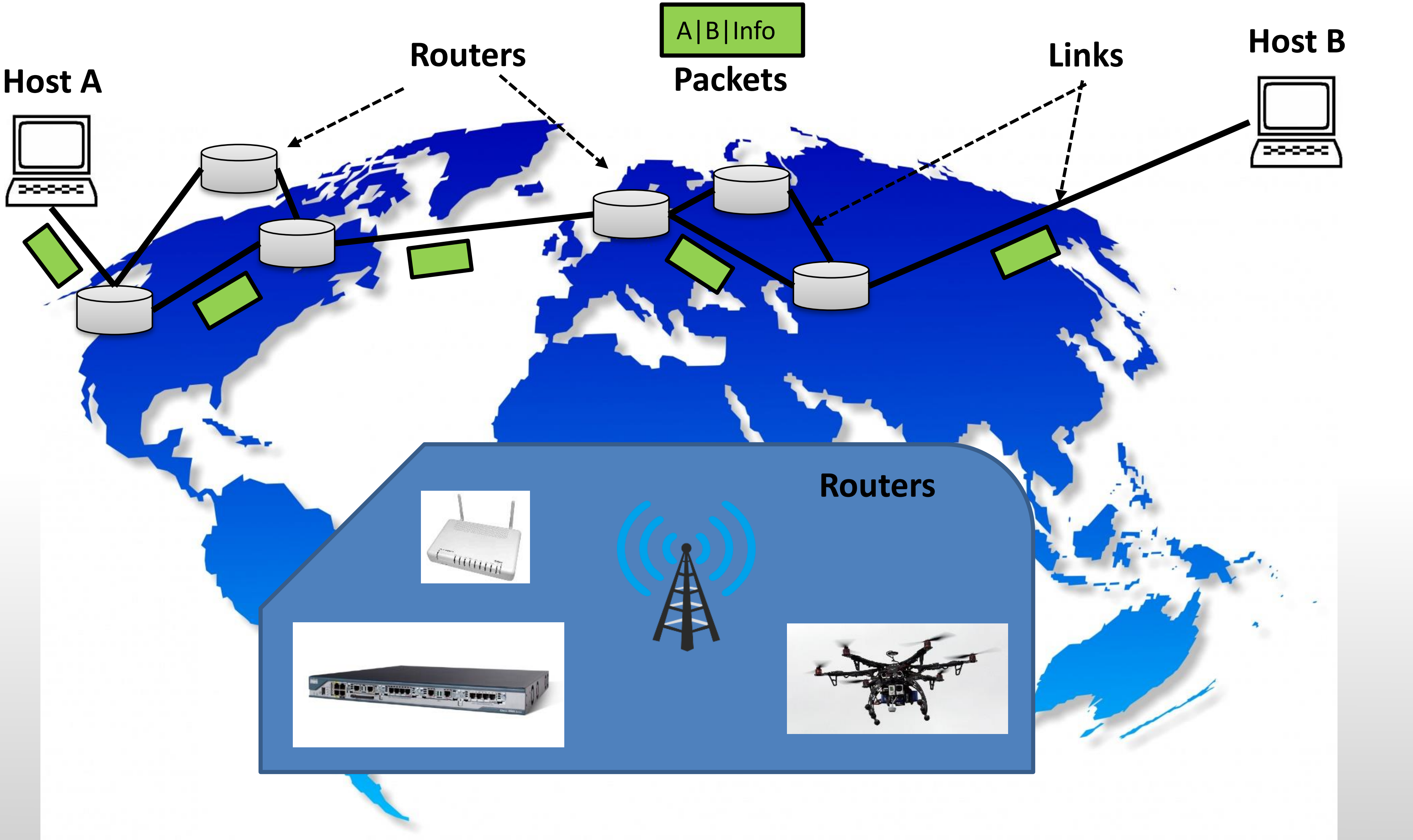
OUTLINE

- ✓ **The Internet & Networkmetrics**
- ✓ Examples
  - Estimation
  - Anomaly Detection
  - Others
- ✓ Conclusion

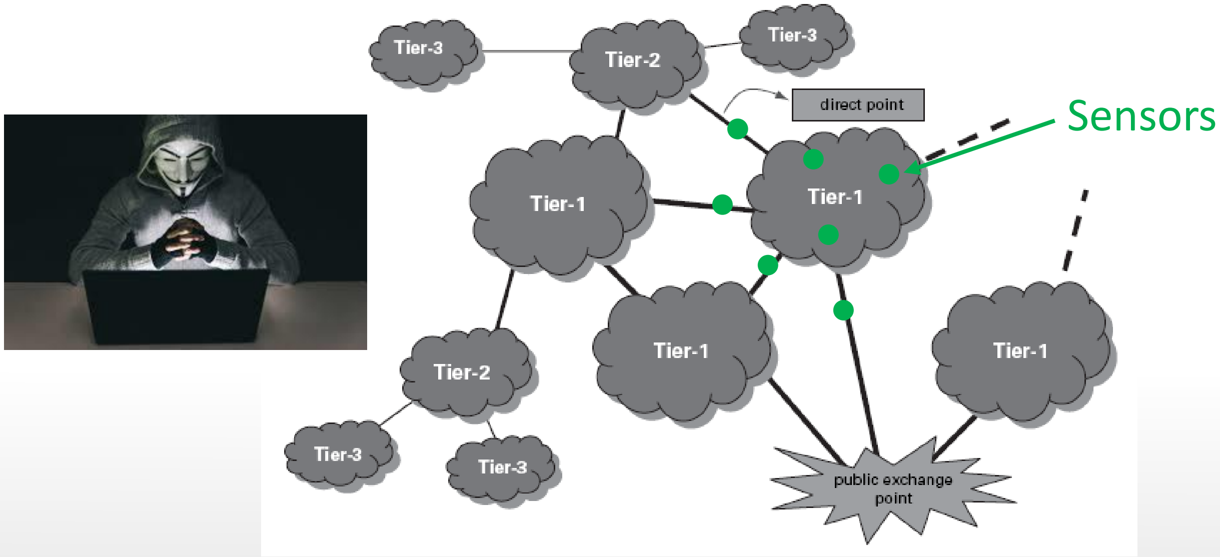
THE INTERNET & NETWORKMETRICS



THE INTERNET & NETWORKMETRICS



- Challenges:
  - ✓ Internet Like a Huge, Distributedly Owned, Industrial Process



- Lack of observability ➔ Control/Optimization Complexity
- Complex anomaly detection, diagnosis & troubleshooting
  - malfunction, but also malicious
- Big Data

- Big Data Problem:

Velocity

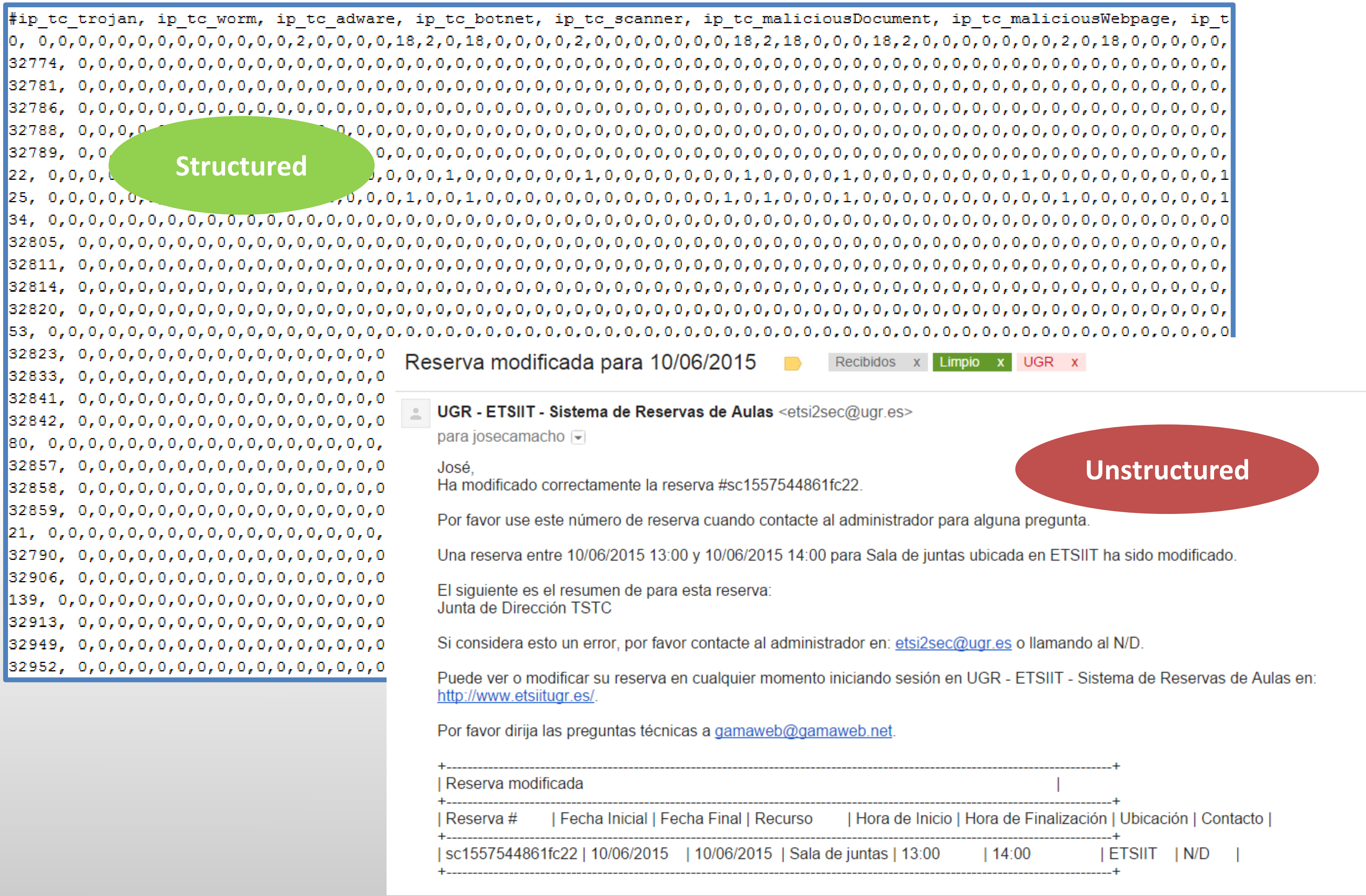
Veracity

Unstructured

Variety

VOLUME

THE INTERNET & NETWORKMETRICS



```
#ip_tc_trojan, ip_tc_worm, ip_tc_adware, ip_tc_botnet, ip_tc_scanner, ip_tc_maliciousDocument, ip_tc_maliciousWebpage, ip_t
0, 0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,18,2,0,18,0,0,0,0,2,0,0,0,0,0,0,0,18,2,18,0,0,0,18,2,0,0,0,0,0,0,0,2,0,18,0,0,0,0,0,
32774, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32781, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32786, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32788, 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,
32789, 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,
22, 0,0,0,                          0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1
25, 0,0,0,0,0,                    0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1
34, 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
32805, 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,
32811, 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,
32814, 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,
32820, 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,
53, 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
32823, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32833, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32841, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32842, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
80, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32857, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32858, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32859, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
21, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
32790, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32906, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
139, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32913, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32949, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
32952, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Reserva modificada para 10/06/2015

UGR - ETSIIT - Sistema de Reservas de Aulas <etsi2sec@ugr.es>
para josecamacho

José,
Ha modificado correctamente la reserva #sc1557544861fc22.

Por favor use este número de reserva cuando contacte al administrador para alguna pregunta.

Una reserva entre 10/06/2015 13:00 y 10/06/2015 14:00 para Sala de juntas ubicada en ETSIIT ha sido modificado.

El siguiente es el resumen de para esta reserva:
Junta de Dirección TSTC

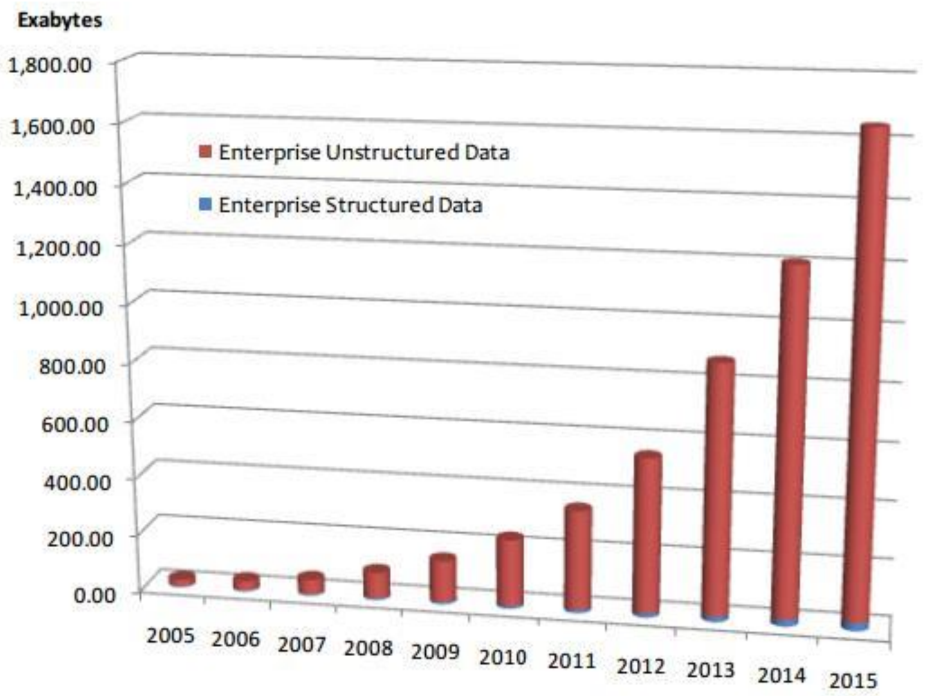
Si considera esto un error, por favor contacte al administrador en: etsi2sec@ugr.es o llamando al N/D.

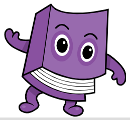
Puede ver o modificar su reserva en cualquier momento iniciando sesión en UGR - ETSIIT - Sistema de Reservas de Aulas en: http://www.etsiitugr.es/.

Por favor dirija las preguntas técnicas a gamaweb@gamaweb.net.

```
+------------------------------------------------------------------------------------------------------------+
| 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      |
+------------------------------------------------------------------------------------------------------------+
```

THE INTERNET & NETWORKMETRICS





Michael Walker, Data Science Central

Exabyte	10^{18}	1.000.000.000.000.000.000 bytes

- Application problems:
  - ✓ Anomaly Detection
  - ✓ Estimation pro
  - ✓ Optimization
  - ✓ Classification
  - ✓ Exploratory ana

MULTIVARIATE ANALYSIS

THE INTERNET & NETWORKMETRICS

- Networkmetrics: MA for Computer Networks
  - ✓ Applications for Exploratory Analysis, Optimization, Classification, Anomaly Detection (≅ **Chemometrics**)
  - ✓ Most is 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

OUTLINE

✓ The Internet & Networkmetrics

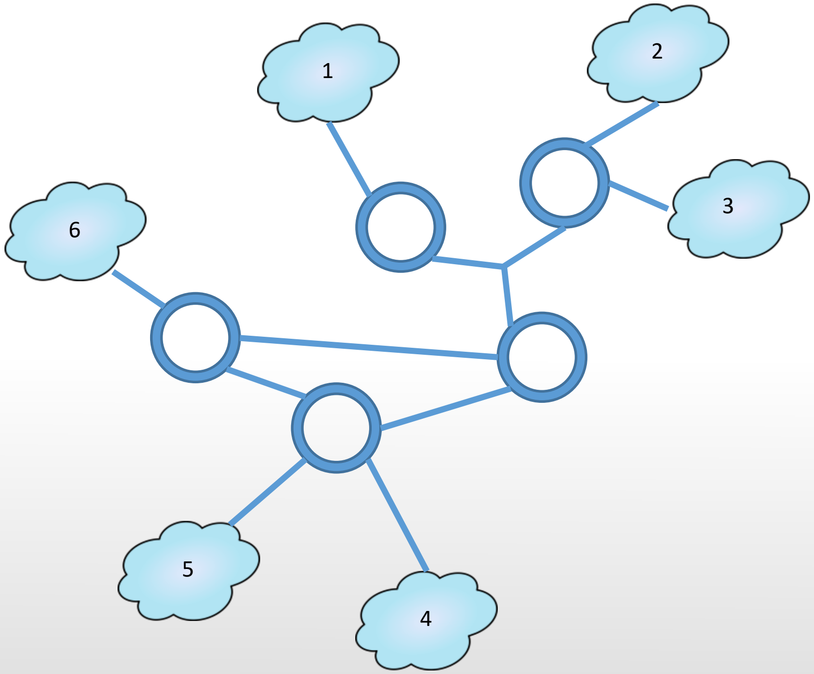
✓ **Examples**

- **Estimation**
- Anomaly Detection
- Others

✓ Conclusion

EXAMPLES

- Traffic Matrix



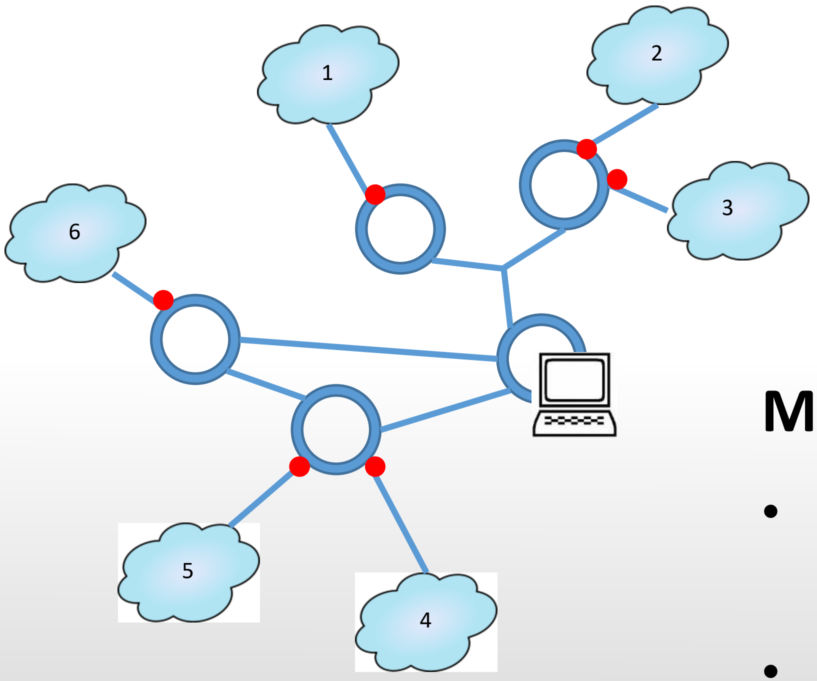
$$H_t = \begin{bmatrix} 0 & 102 & 23 & 54 & 102 & 804 \\ 100 & 0 & 44 & 46 & 22 & 55 \\ 12 & 34 & 0 & 130 & 12 & 12 \\ 60 & 32 & 204 & 0 & 32 & 45 \\ 120 & 28 & 103 & 5 & 0 & 82 \\ 1005 & 34 & 54 & 114 & 73 & 0 \end{bmatrix}$$

Network Monitoring

Network Optimization

EXAMPLES

Traffic Matrix

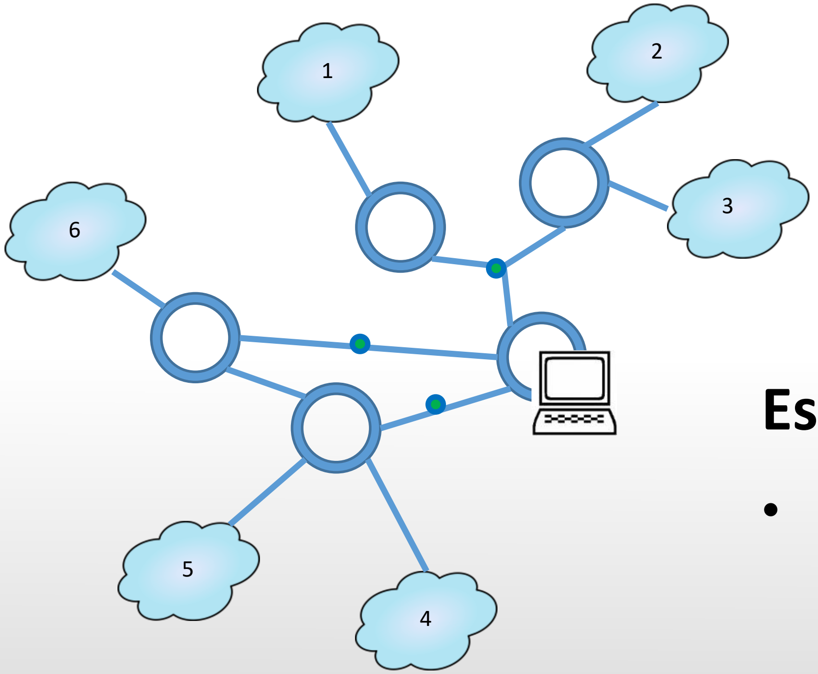


$$H_t = \begin{bmatrix} 0 & 102 & 23 & 54 & 102 & 804 \\ 100 & 0 & 44 & 46 & 22 & 55 \\ 12 & 34 & 0 & 130 & 12 & 12 \\ 60 & 32 & 204 & 0 & 32 & 45 \\ 120 & 28 & 103 & 5 & 0 & 82 \\ 1005 & 34 & 54 & 114 & 73 & 0 \end{bmatrix}$$

Measure TM

- **Pick every single packet (Huge Data Volume)**
- **Netflow Sensor (High DV)**

Traffic Matrix



$$H_t = \begin{bmatrix} 0 & 102 & 23 & 54 & 102 & 804 \\ 100 & 0 & 44 & 46 & 22 & 55 \\ 12 & 34 & 0 & 130 & 12 & 12 \\ 60 & 32 & 204 & 0 & 32 & 45 \\ 120 & 28 & 103 & 5 & 0 & 82 \\ 1005 & 34 & 54 & 114 & 73 & 0 \end{bmatrix}$$

Estimate TM

- **Volume of traffic in links (Low Data Volume)**

EXAMPLES

- Traffic Matrix

Links: Low Volume

ROUTER A	Link1	Link2
[12:09 7/23]	814768.00	31750774.00
[12:10 7/23]	909022.00	36295730.00
[12:11 7/23]	917352.00	36802806.00
[12:12 7/23]	884206.00	34970580.00
[12:13 7/23]	893056.00	35885934.00
[12:14 7/23]	881923.00	33974831.00
[12:15 7/23]	835326.00	32906544.00
[12:16 7/23]	864102.00	34287672.00
[12:17 7/23]	939600.00	37733404.00

NETFLOW: High Volume

```
1970-01-02	01:13:53,1970-01-02
01:14:59,66.822,33.4.1.0,0.0.0.0,0,0,,......,0,0,25224,1320528,0,0,0,0,0,0,0,0,0,0,0,0.0.0.0,0.0.0.0,0,0,00:00:00:00:00:00,00:00:00:00:00:00,00:00:00:00:00:00,00
:00:00:00:00:00,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0.0.0.0,0/0

1970-01-02	01:14:55,1970-01-02
01:16:14,79.112,33.4.1.0,0.0.0.0,0,0,,......,0,0,28614,1487928,0,0,0,0,0,0,0,0,0,0,0,0.0.0.0,0.0.0.0,0,0,00:00:00:00:00:00,00:00:00:00:00:00,00:00:00:00:00:00,00
:00:00:00:00:00,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0.0.0.0,0/0

1970-01-02	01:15:56,1970-01-02
01:17:15,79.162,33.4.1.0,0.0.0.0,0,0,,......,0,0,26681,1387412,0,0,0,0,0,0,0,0,0,0,0,0.0.0.0,0.0.0.0,0,0,00:00:00:00:00:00,00:00:00:00:00:00,00:00:00:00:00:00,00
:00:00:00:00:00,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0.0.0.0,0/0

1970-01-02	01:16:57,1970-01-02
01:18:16,79.072,33.4.1.0,0.0.0.0,0,0,,......,0,0,25757,1344764,0,0,0,0,0,0,0,0,0,0,0,0.0.0.0,0.0.0.0,0,0,00:00:00:00:00:00,00:00:00:00:00:00,00:00:00:00:00:00,00
:00:00:00:00:00,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0.0.0.0,0/0

1970-01-02	01:17:58,1970-01-02
01:19:13,74.800,33.4.1.0,0.0.0.0,0,0,,......,0,0,15572,809744,0,0,0,0,0,0,0,0,0,0,0,0.0.0.0,0.0.0.0,0,0,00:00:00:00:00:00,00:00:00:00:00:00,00:00:00:00:00:00,00:
00:00:00:00:00,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0-0-0,0.0.0.0,0/0
```

Tomography: Y (links) = R · X (Netflow)

- Traffic Matrix
  - PCA + R

PCA (SVD)

$$X \approx U_A S_A V'_A$$

$$Y \approx U_A S_A V'_A R$$

Tomography

$$Q = S_A V'_A R$$

$$\hat{X} = YQ'(QQ')^{-1} S_A V'_A$$

Multivariate (PLS) Model? X = b· Y

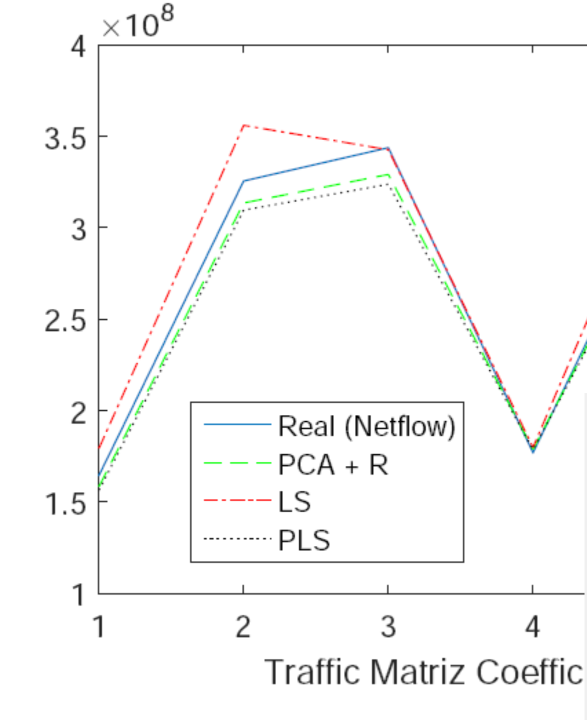
Lakhina A, Papagiannaki K, Crovella M., Diot C, Kolaczyk E.D, Taft N. Structural Analysis of Network Traffic Flows SIGMETRICS Perform. Eval. Rev.. 2004;32:61-72.

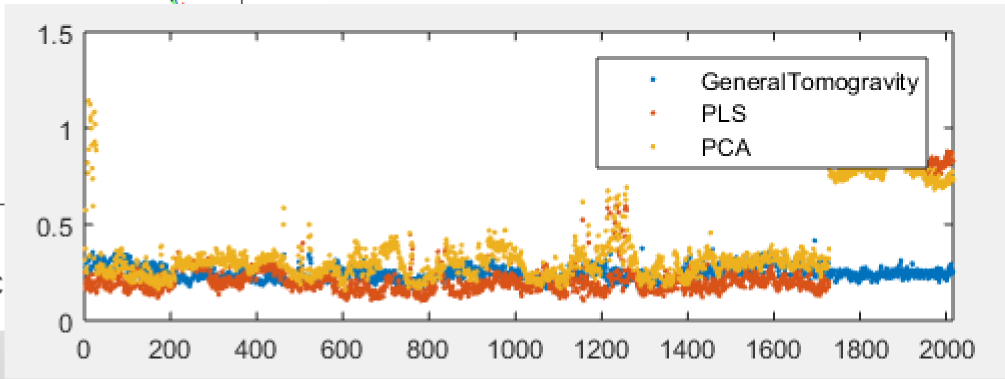
EXAMPLES

- Traffic Matrix

GRAY MODELLING

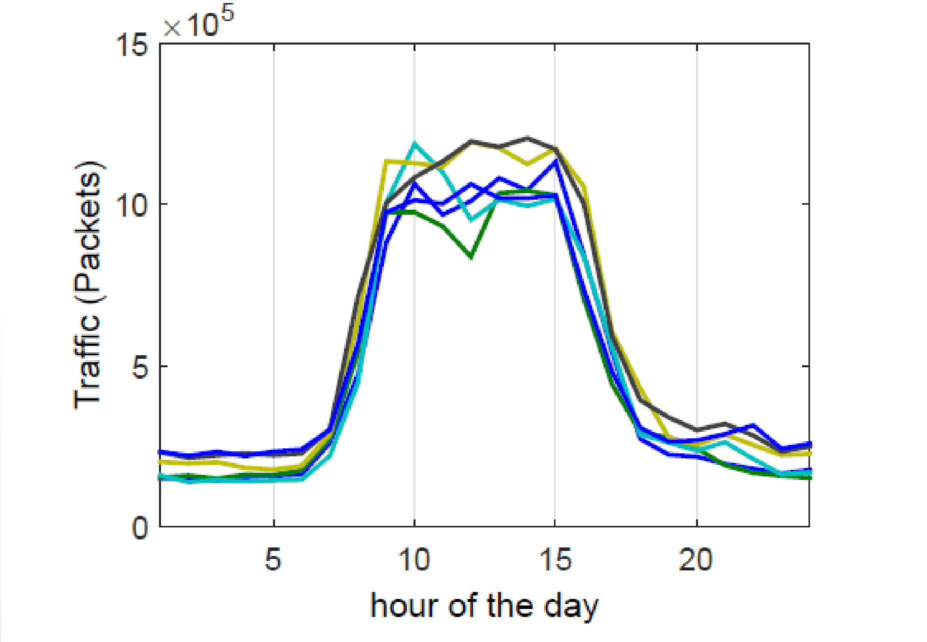
PATH MODELLING

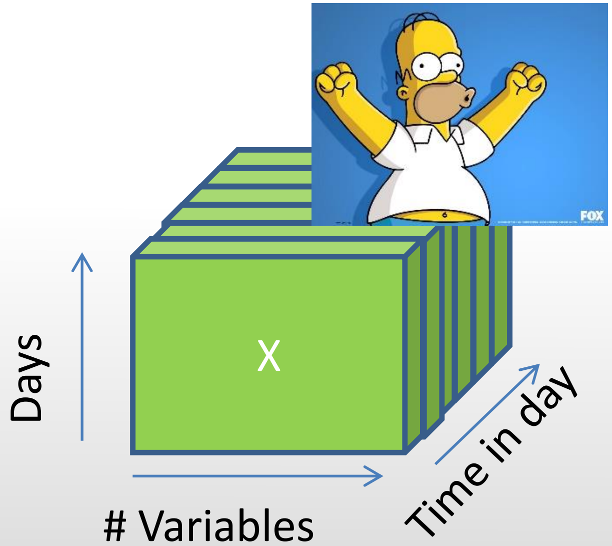




EXAMPLES

- Traffic Matrix:
  - Not Stationary, but CycloStationary (days, weeks)





Days

X

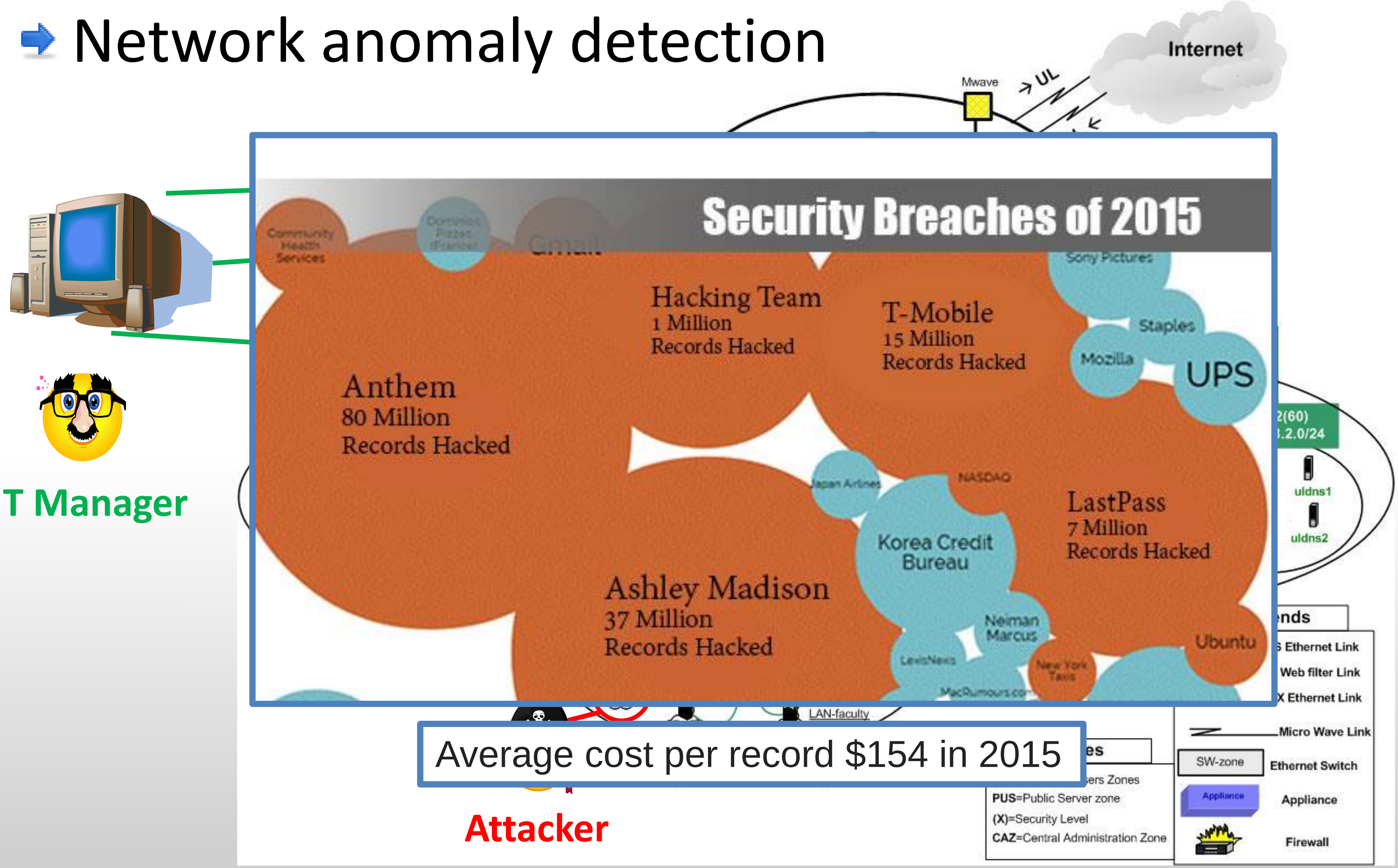
# Variables

Time in day

OUTLINE

- ✓ The Internet & Networkmetrics
- ✓ **Examples**
  - Estimation
  - **Anomaly Detection**
  - Others
- ✓ Conclusion

- Network anomaly detection



Average cost per record $154 in 2015

EXAMPLES

- Network anomaly detection

Velocity

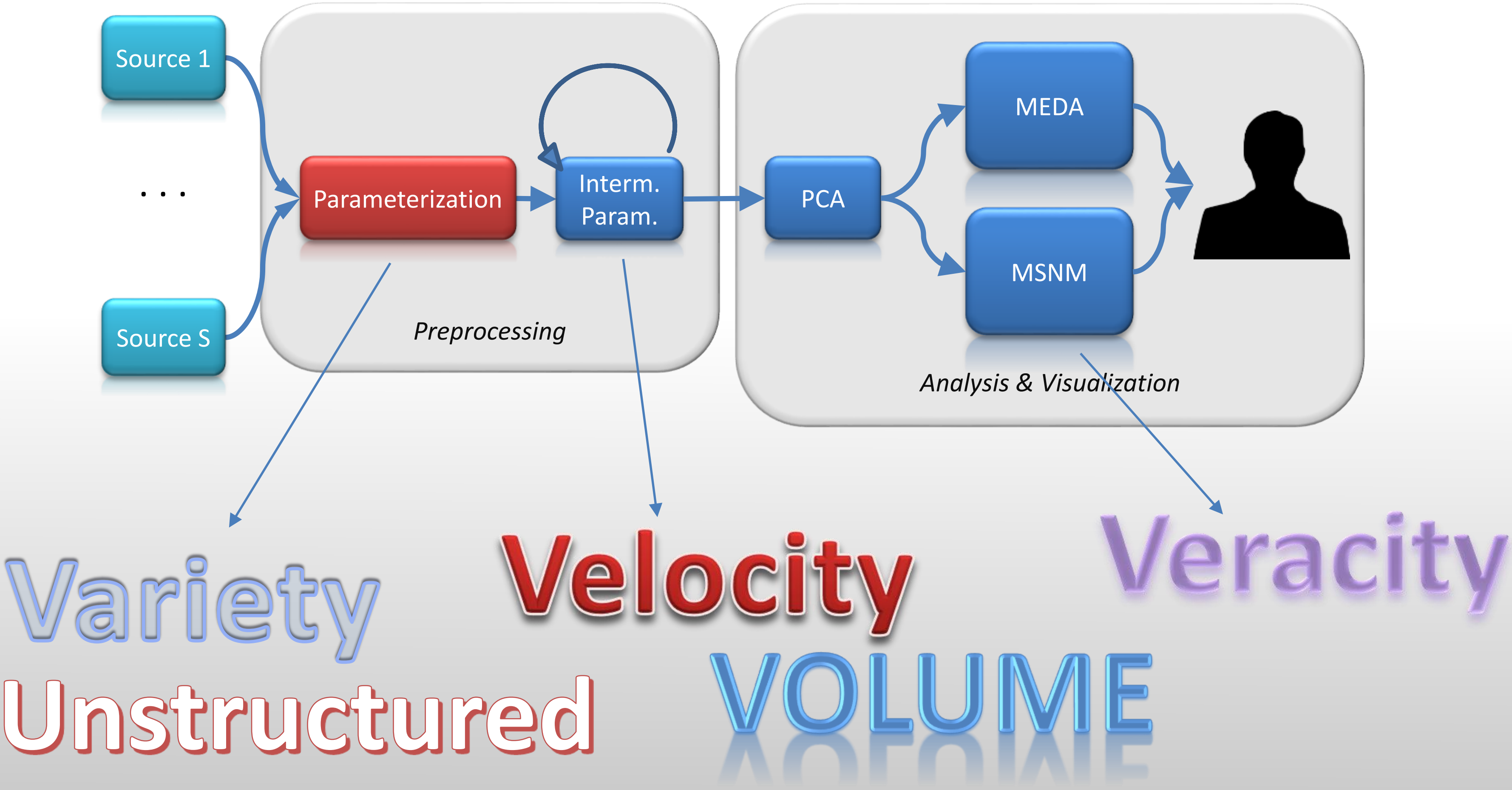
Veracity

Multivariate Statistical Network Monitoring (MSNM) ≅ MSPC

Variety

VOLUME

EXAMPLES



Source 1

. . .

Source S

Parameterization

Interm. Param.

Preprocessing

PCA

MEDA

MSNM

Analysis & Visualization

Variety

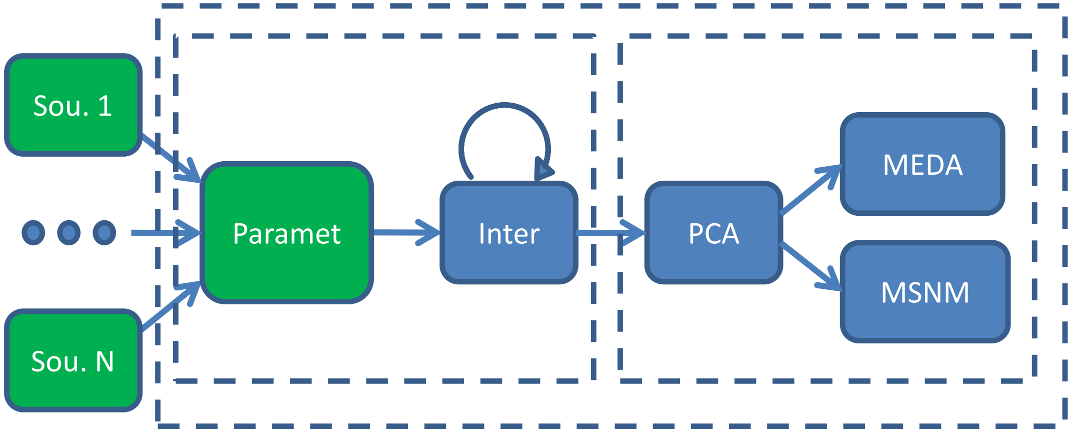
Velocity

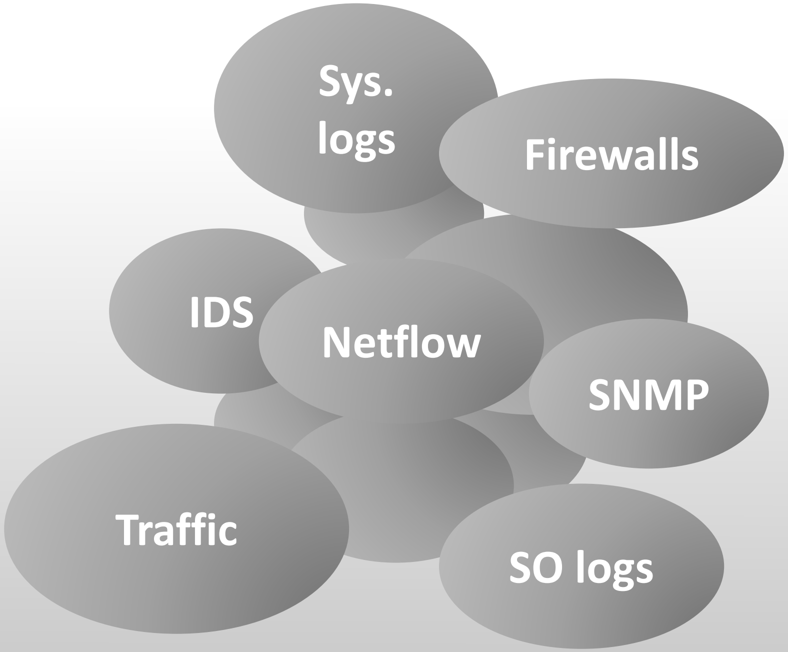
Veracity

Unstructured

VOLUME

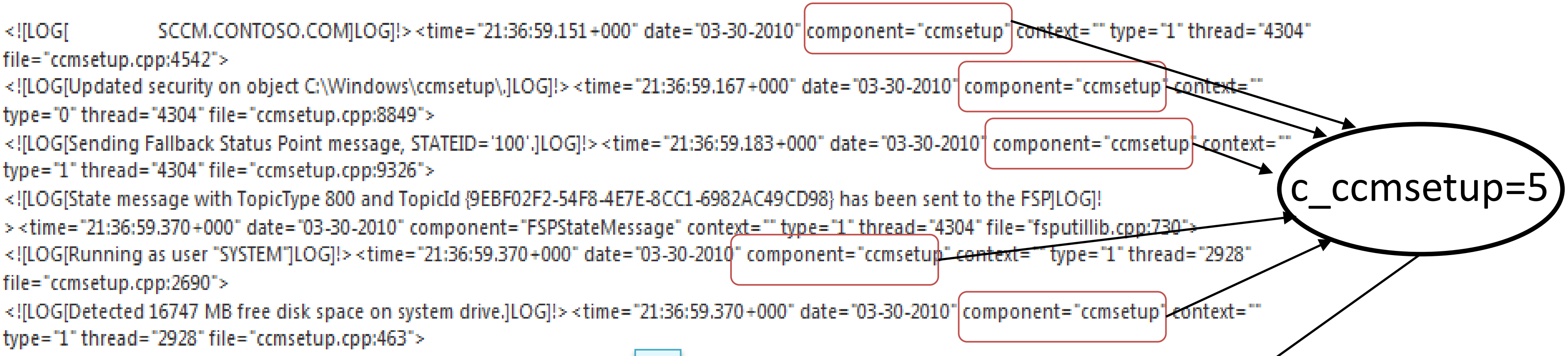
EXAMPLES


Sou. 1
Sou. N
Paramet
Inter
PCA
MEDA
MSNM


Sys. logs
Firewalls
IDS
Netflow
SNMP
Traffic
SO logs

EXAMPLES

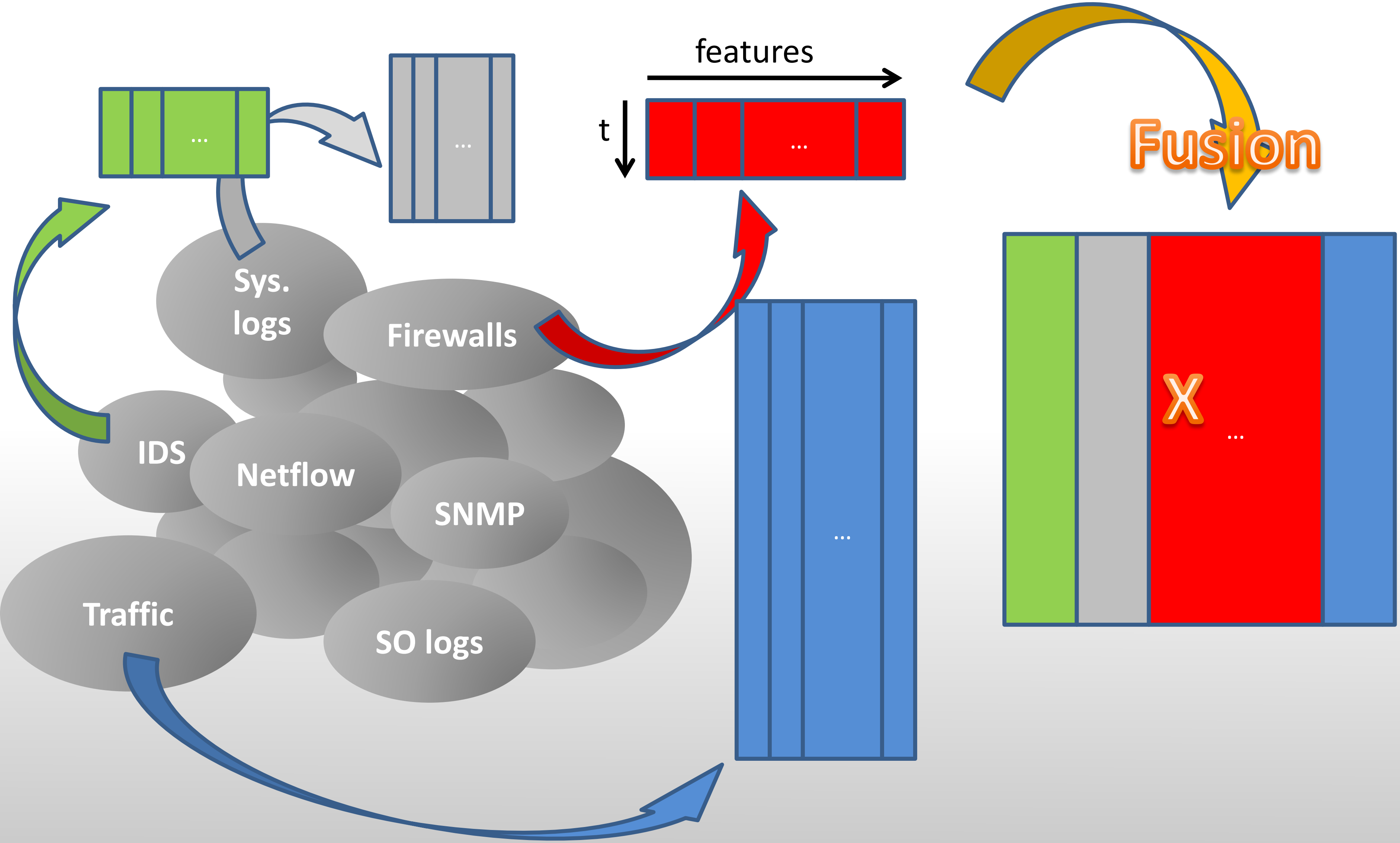
Feature-as-a-counter:



```
<![LOG[           SCCM.CONTOSO.COM]LOG]!><time="21:36:59.151+000" date="03-30-2010" component="ccmsetup" context="" type="1" thread="4304"
file="ccmsetup.cpp:4542">
<![LOG[Updated security on object C:\Windows\ccmsetup\.]LOG]!><time="21:36:59.167+000" date="03-30-2010" component="ccmsetup" context=""
type="0" thread="4304" file="ccmsetup.cpp:8849">
<![LOG[Sending Fallback Status Point message, STATEID='100'.]LOG]!><time="21:36:59.183+000" date="03-30-2010" component="ccmsetup" context=""
type="1" thread="4304" file="ccmsetup.cpp:9326">
<![LOG[State message with TopicType 800 and TopicId {9EBF02F2-54F8-4E7E-8CC1-6982AC49CD98} has been sent to the FSP]LOG]!
><time="21:36:59.370+000" date="03-30-2010" component="FSPStateMessage" context="" type="1" thread="4304" file="fsputillib.cpp:730">
<![LOG[Running as user "SYSTEM"]LOG]!><time="21:36:59.370+000" date="03-30-2010" component="ccmsetup" context="" type="1" thread="2928"
file="ccmsetup.cpp:2690">
<![LOG[Detected 16747 MB free disk space on system drive.]LOG]!><time="21:36:59.370+000" date="03-30-2010" component="ccmsetup" context=""
type="1" thread="2928" file="ccmsetup.cpp:463">
```

Time	FSPStateMessage	ccmstup	thread_4304
T=20s	1	5	4
T=40s	2	3	3
T=60s	1	3	3
T=80s	1	1	4

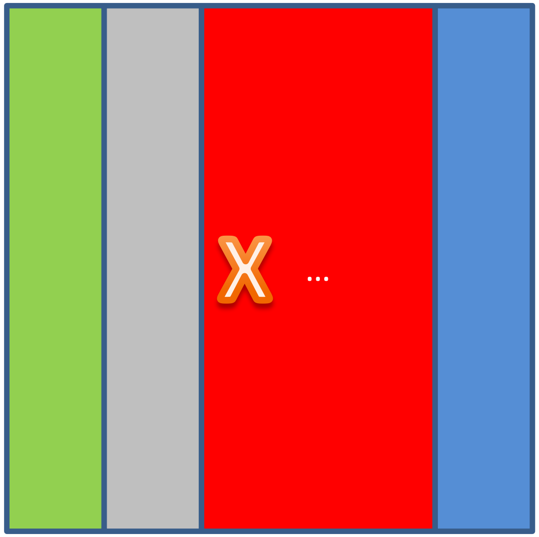
EXAMPLES


features
t
…
Fusion
X
Sys. logs
Firewalls
IDS
Netflow
SNMP
Traffic
SO logs

EXAMPLES

Definition of the features

Definition of the observations



The **features** are the parameters that will be computed for the observations**:**

Feature-as-a-counter

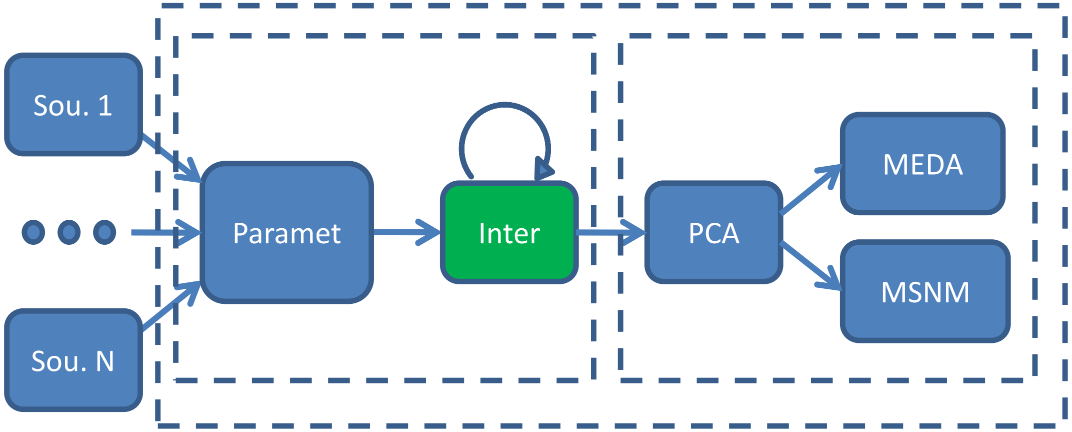
Nº of times of a given **event**

- ✓ **Groups of Devices**
- ✓ **Types of alarmas**

The **observations** are the items or entities that may be **identified as anomalous or normal:**

- Obs = **Time interval** to identify **anomalous intervals** as soon as possible.
- Obs = **Devices** to identify **attackers**
- Obs as combinations

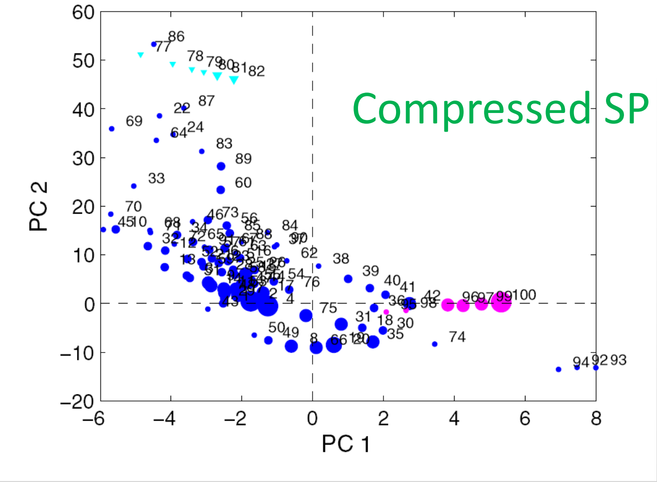
EXAMPLES



- Handling Volume & Velocity
  - For variables ➔ Kernel update

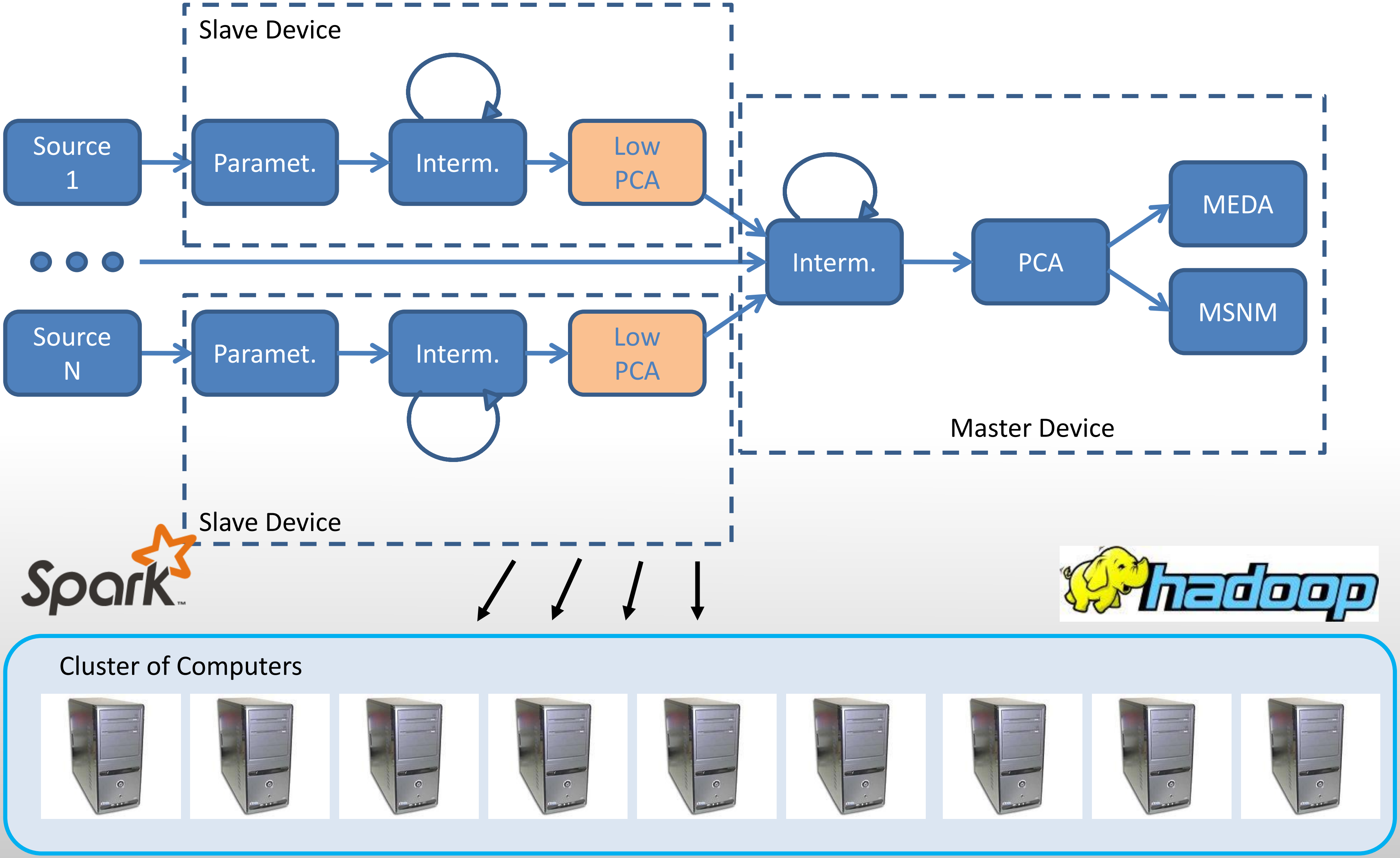
$$(X'X)_t = \lambda \cdot (X'X)_{t-1} + \tilde{X}_t{}' \cdot \tilde{X}_t$$

  - For observations ➔ Clustering



Camacho, J. Visualizing Big data with Compressed Score Plots: Approach and Research Challenges. Chemometrics and Intelligent Laboratory Systems, 2014, 135: 110-125.

EXAMPLES



Slave Device

Source 1

Paramet.

Interm.

Low PCA

Source N

Paramet.

Interm.

Low PCA

Slave Device

Interm.

PCA

MEDA

MSNM

Master Device

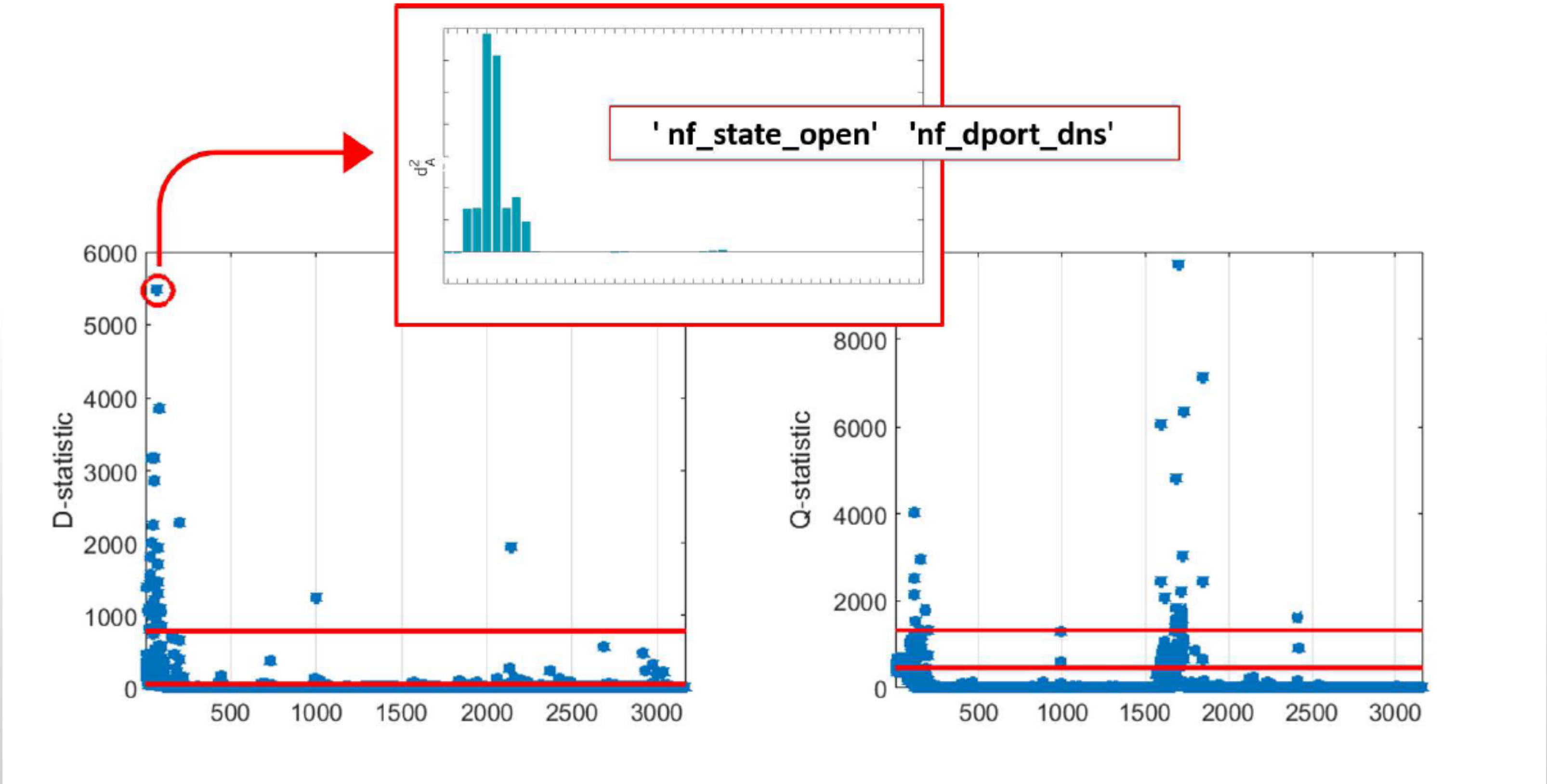
Spark

hadoop

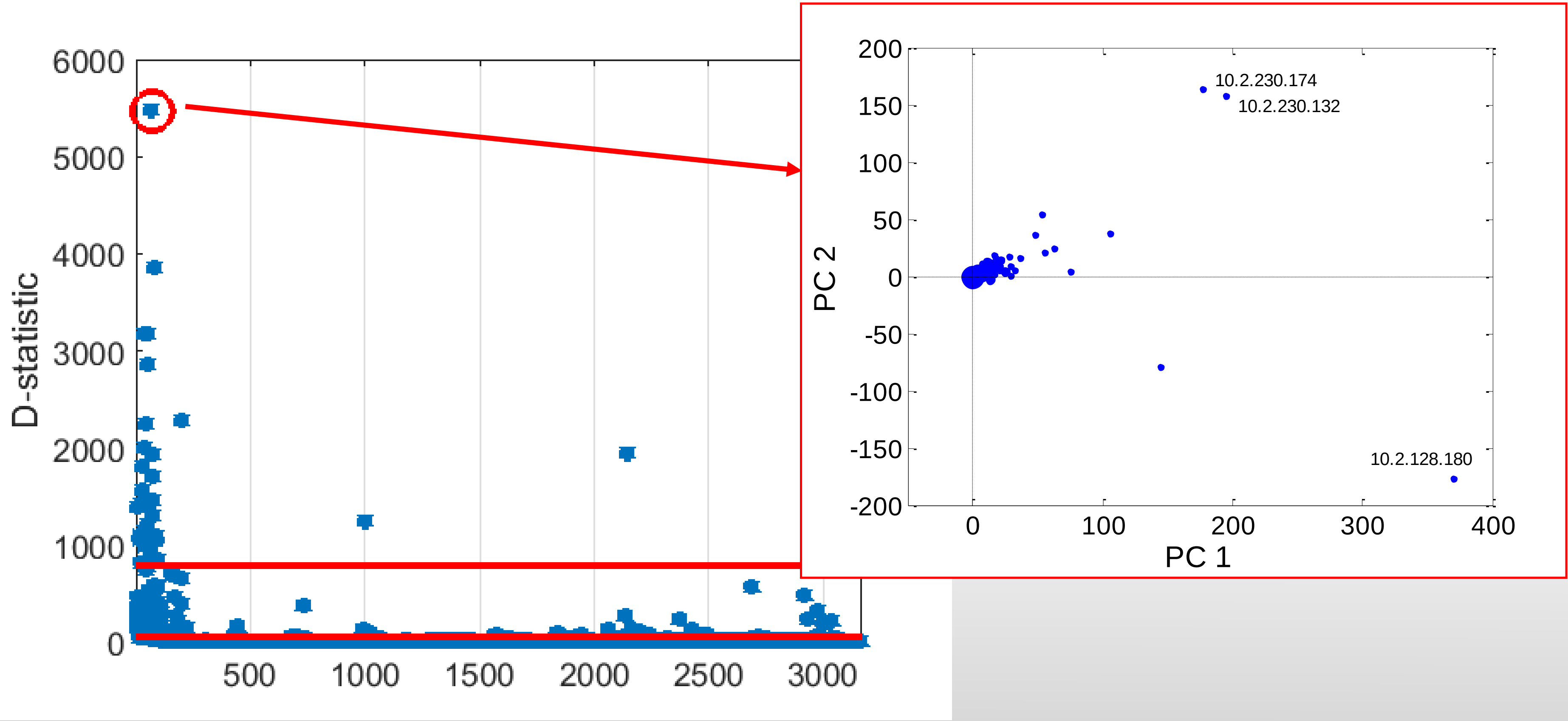
Cluster of Computers

EXAMPLES

- MSNM



MSNM


6000
5000
4000
3000
2000
1000
0
D-statistic
500
1000
1500
2000
2500
3000
200
150
100
50
0
-50
-100
-150
-200
PC 2
0
100
200
300
400
PC 1
10.2.230.174
10.2.230.132
10.2.128.180

OUTLINE

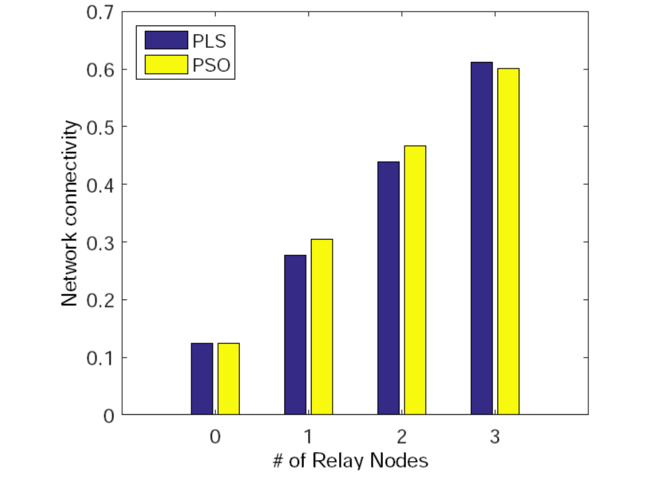
- ✓ The Internet & Networkmetrics
- ✓ **Examples**
  - Estimation
  - Anomaly Detection
  - **Others**
- ✓ Conclusion

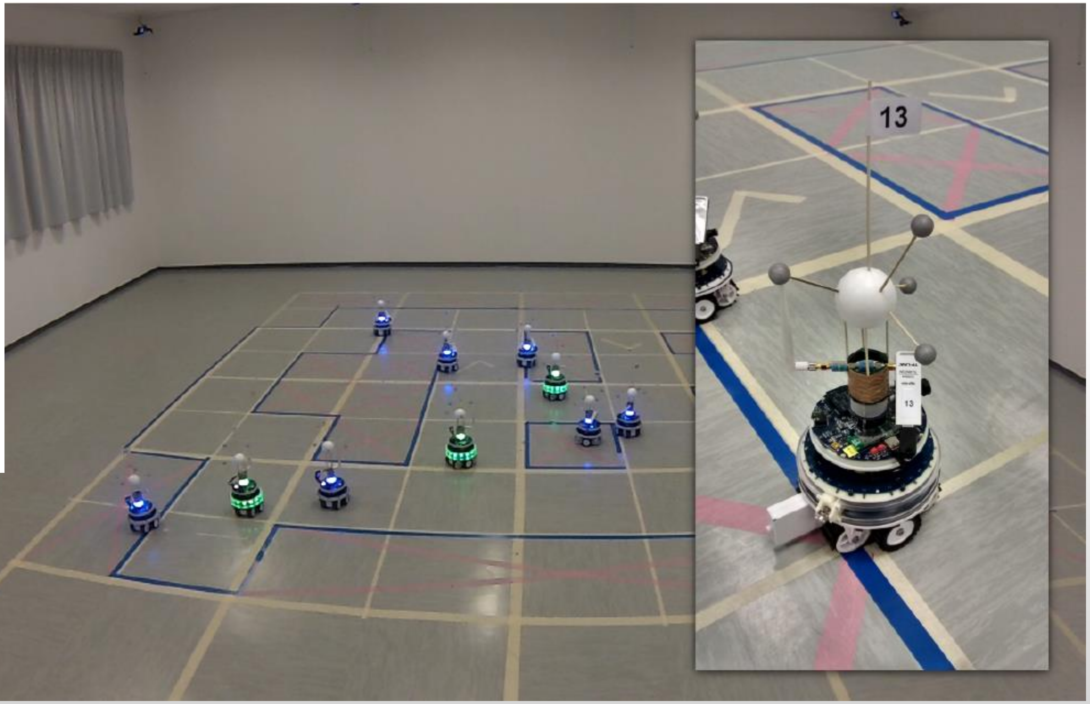
EXAMPLES

OTHER EXAMPLES

Optimization

https://youtu.be/mW1Q_MUFYs4





Camacho, J., Picó, J., Ferrer, A.J. Self-tuning run to run optimization of fed-batch processes using unfold-PLS. AIChE Journal, 2007, 53 (7): 1789-1804.

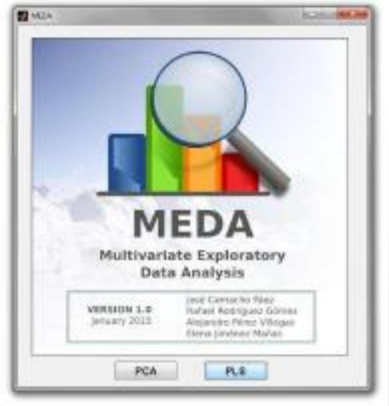
Networkmetrics: Multivariate Big Data Analysis in the Context of the Internet

OTHER EXAMPLES

Exploratory Data Analysis

MEDA Toolbox

MATLAB MathWorks



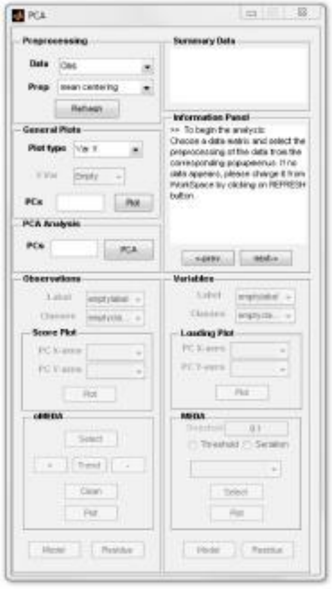
MEDA

MEDA

Multivariate Exploratory Data Analysis

VERSION 1.0
January 2015

José Camacho Páez
Rafael Rodríguez Gómez
Alejandro Pérez Villegas
Elena Jiménez Mañas



PCA

Preprocessing

Data: Obs

Prep: mean centering

Summary Data

Information Panel

>> To begin the analysis:
Choose a data matrix and select the preprocessing of the data from the corresponding popupmenus. If no data appears, please charge it from WorkSpace by clicking on REFRESH button.

General Plots

Plot type: Var X

Y Var: Empty

PCA Analysis

Observations

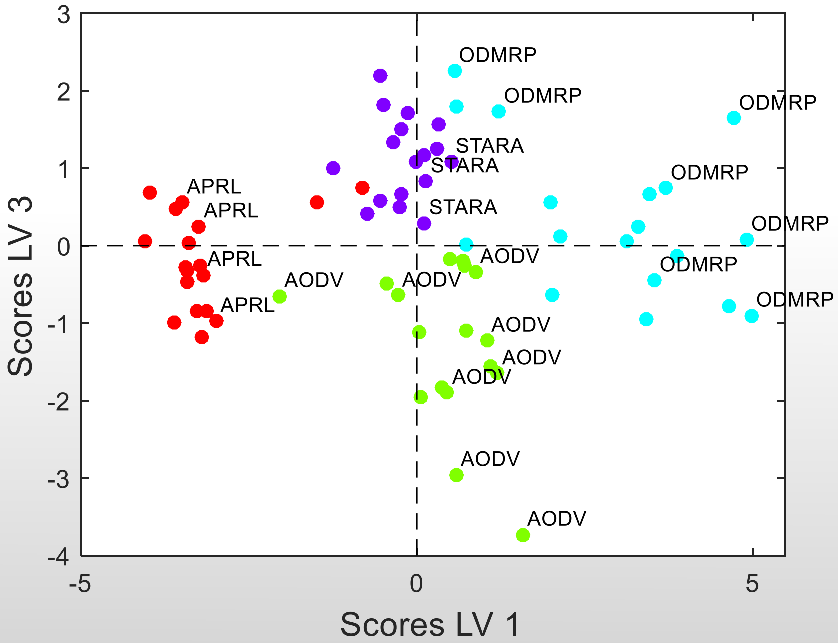
Label: emptylabel

Classes: emptyclasses

Variables

Label: emptylabel

Classes: emptyclasses



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

J. Camacho, R. Magán-Carrión, P. García-Teodoro, J.J. Treinen, "Networkmetrics: Multivariate Big Data Analysis in the Context of the Internet", Featured paper in Journal of Chemometrics

OUTLINE

- ✓ The Internet & Networkmetrics
- ✓ Examples
  - Estimation
  - Anomaly Detection
  - Others
- ✓ **Conclusion**

CONCLUDING REMARKS

- Multivariate Analysis tools can be extended to Networking for estimation, anomaly detection and optimization (Networkmetrics)
  - ✓ with new and interesting particularities and challenges
  - ✓ with challenges already solved in chemometrics and similar areas

CONCLUDING REMARKS

FOX
THESIMPSONS.COM
The Simpsons™ and © 2007 Twentieth Century Fox Film Corporation. All Rights Reserved.

homer@this.is.not.an.email.com

Multivariate Big Data Analysis and its application to the Internet

NETWORKMETRICS

XVI CHEMOMETRICS IN ANALYTICAL CHEMISTRY

JUNE 6-10, 2016
Barcelona, Spain

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

ugr Universidad de Granada