# Exploratory Analysis on Big Data using the MEDA Toolbox
## Present and Future

**José Camacho,**

**Roberto Therón,**

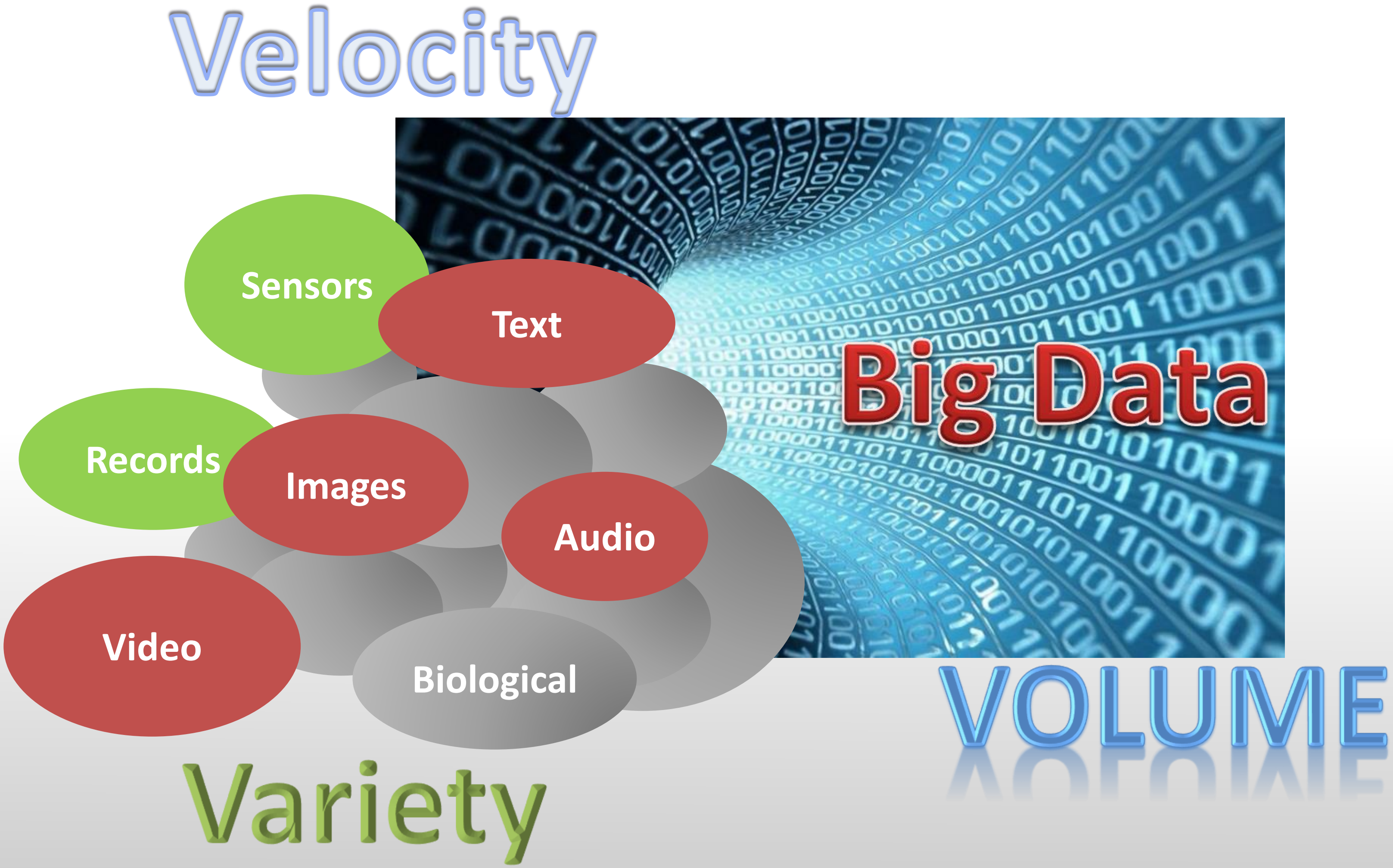
**Roberto Magán**

*Departamento de Teoría de la Señal, Telemática y Comunicaciones*
*Universidad de Granada*

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

ugr Universidad de Granada

# WHAT'S BIG DATA?



Velocity

Sensors

Text

Records

Images

Audio

Video

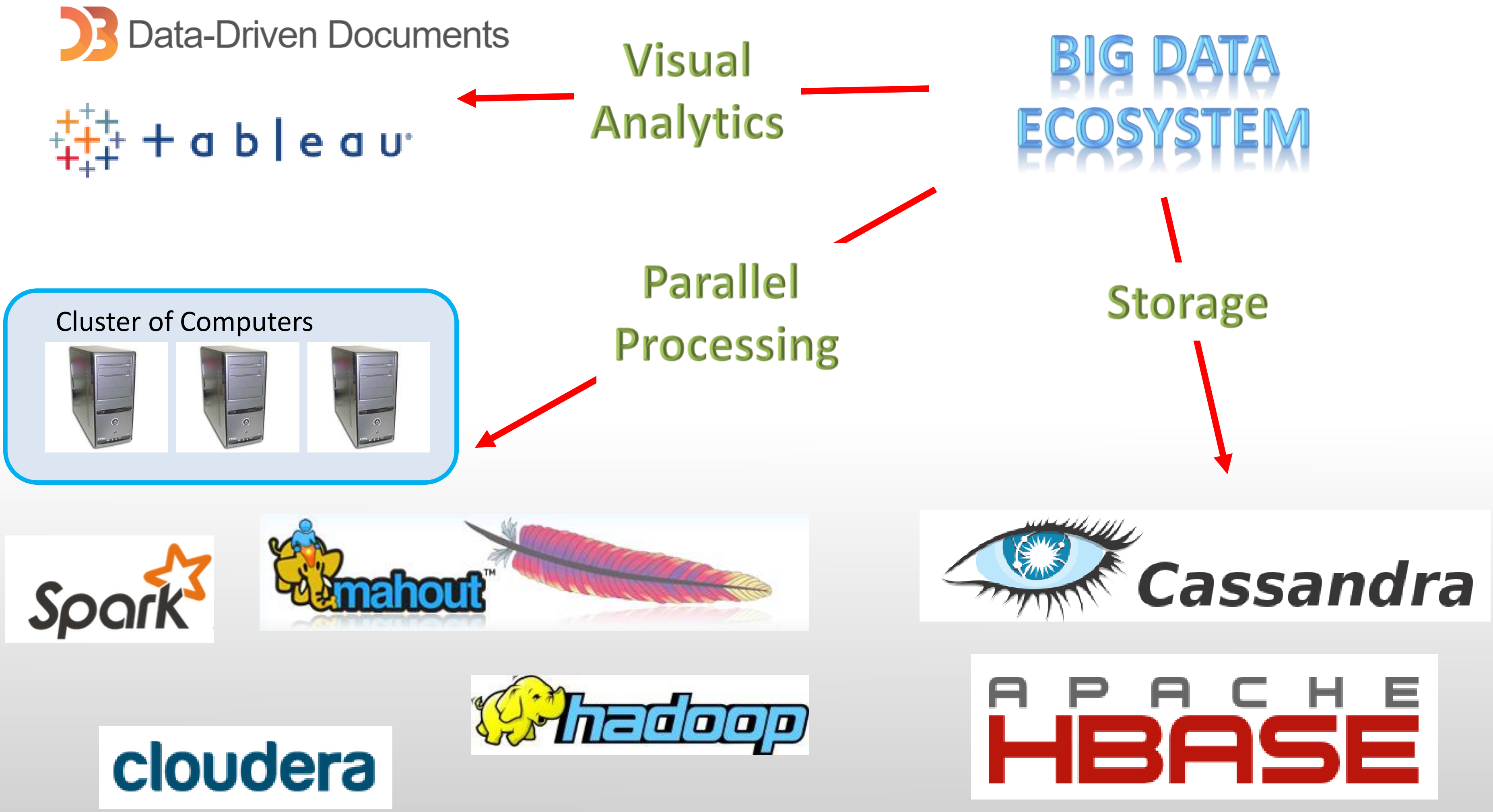
Biological

Variety

Big Data

VOLUME

# WHAT'S BIG DATA?



BIG DATA ECOSYSTEM

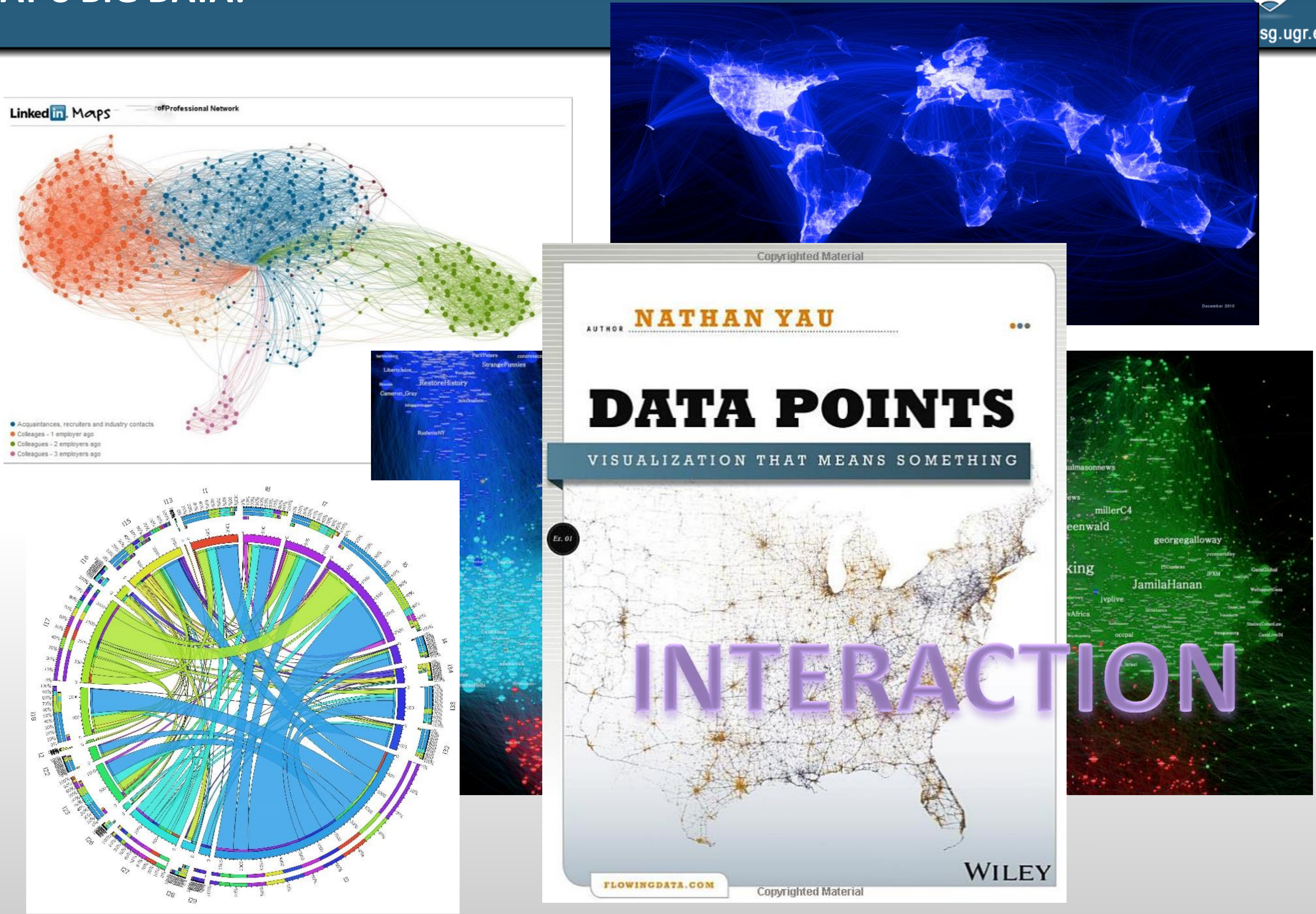
Visual Analytics

Parallel Processing

Storage

Cluster of Computers

# WHAT'S BIG DATA?

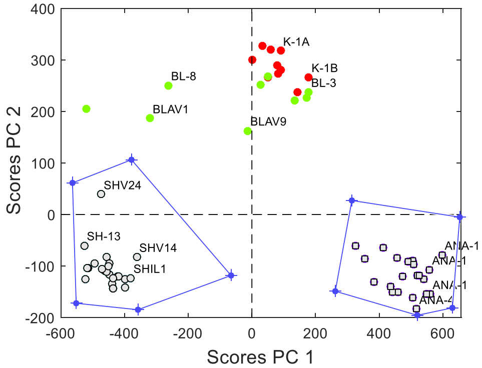


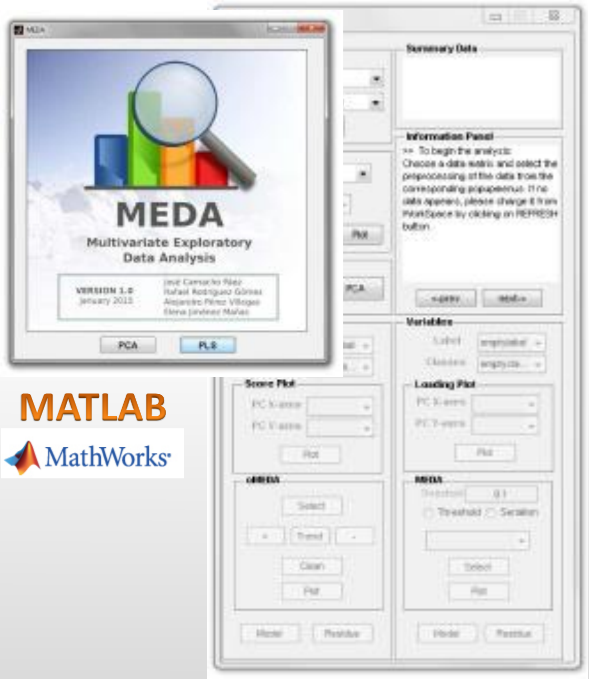
INTERACTION

# MEDA Toolbox

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

- ✓ Models: PCA, PLS-DA, SPLS, GPCA, GPLS
- ✓ Dimensionality:
  - Scree plots
  - CV & D-CV
  - 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
- ✓ Data simulation
  - simuleMV




MEDA

MEDA

Multivariate Exploratory Data Analysis

VERSION 1.0
January 2015

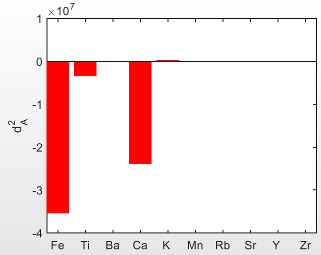
José Camacho Páez
Rafael Rodríguez Gómez
Alejandro Pérez Villegas
Elena Jiménez Mañas

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.

MATLAB

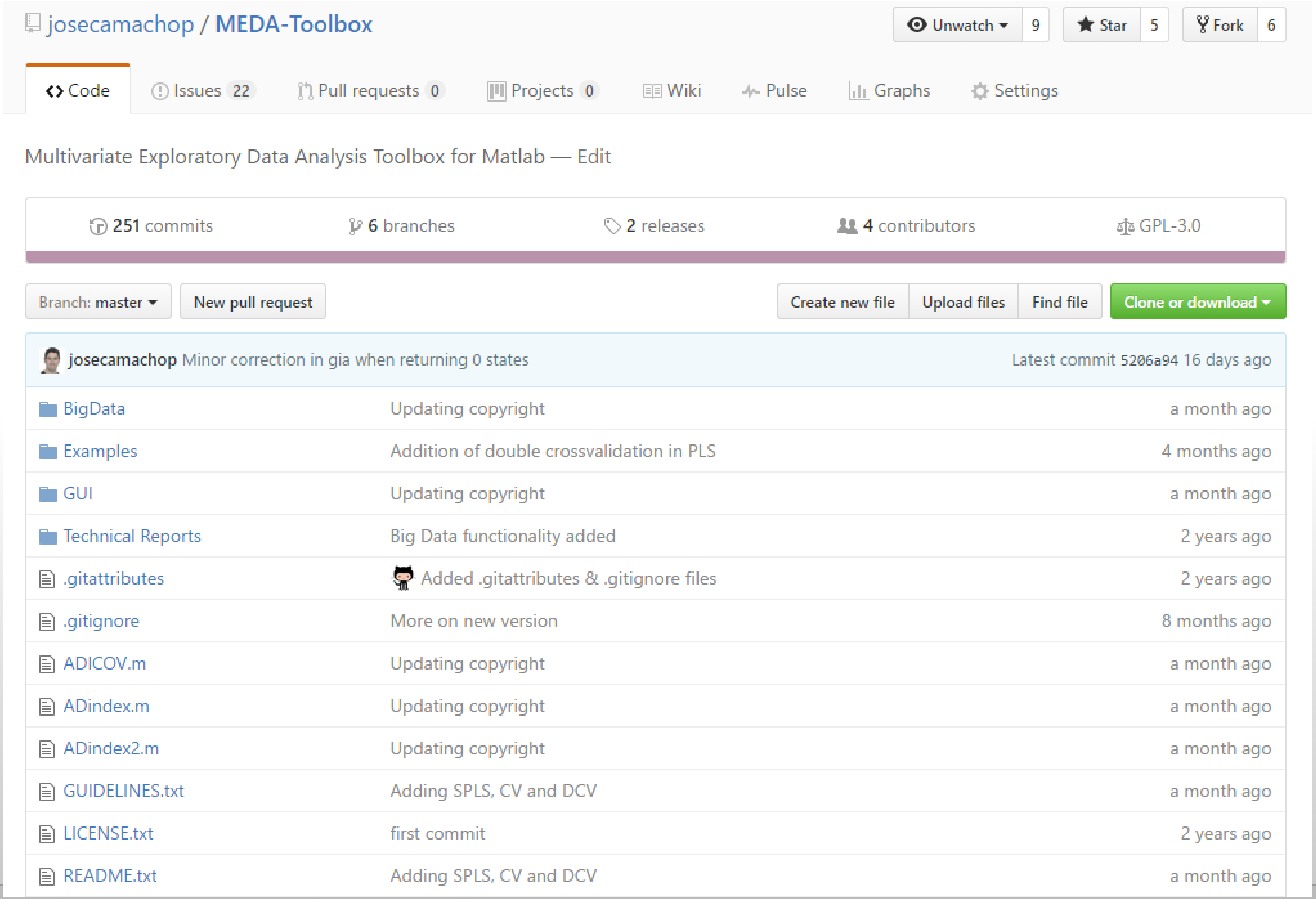
MathWorks



ChemoLab, (2015) 143: 49

# THE MEDA TOOLBOX

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



Multivariate Exploratory Data Analysis Toolbox for Matlab — Edit

josecamachop Minor correction in gia when returning 0 states — Latest commit 5206a94 16 days ago

| | | |
|---|---|---|
| BigData | Updating copyright | a month ago |
| Examples | Addition of double crossvalidation in PLS | 4 months ago |
| GUI | Updating copyright | a month ago |
| Technical Reports | Big Data functionality added | 2 years ago |
| .gitattributes | Added .gitattributes & .gitignore files | 2 years ago |
| .gitignore | More on new version | 8 months ago |
| ADICOV.m | Updating copyright | a month ago |
| ADindex.m | Updating copyright | a month ago |
| ADindex2.m | Updating copyright | a month ago |
| GUIDELINES.txt | Adding SPLS, CV and DCV | a month ago |
| LICENSE.txt | first commit | 2 years ago |
| README.txt | Adding SPLS, CV and DCV | a month ago |

# MEDA FOR BIG DATA

## Extensions for Big Data

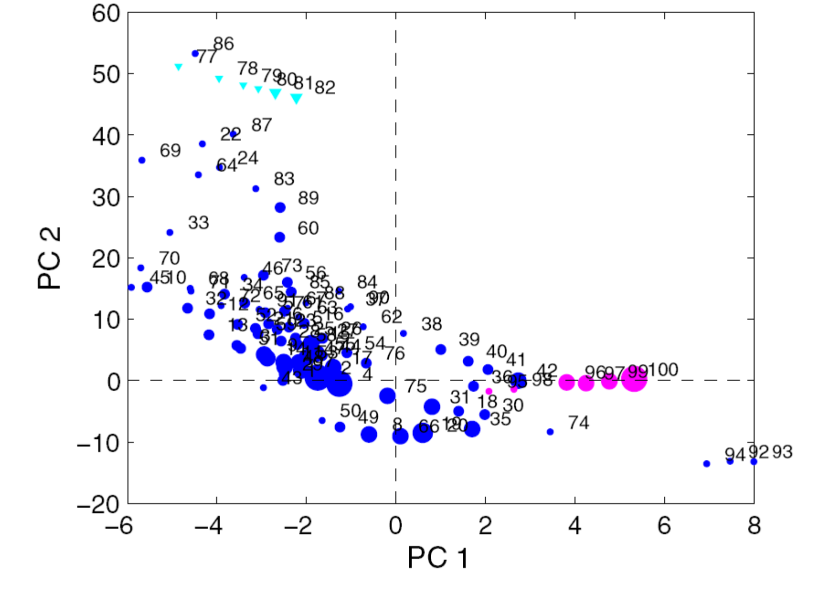
- For variables ➔ (linear) kernel calibration with EWMA update

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

  - <u>Scalable to any size</u>
  - PCA/PLS, MEDA, oMEDA
  - GPCA, GPLS
  - ADICOV MSPC

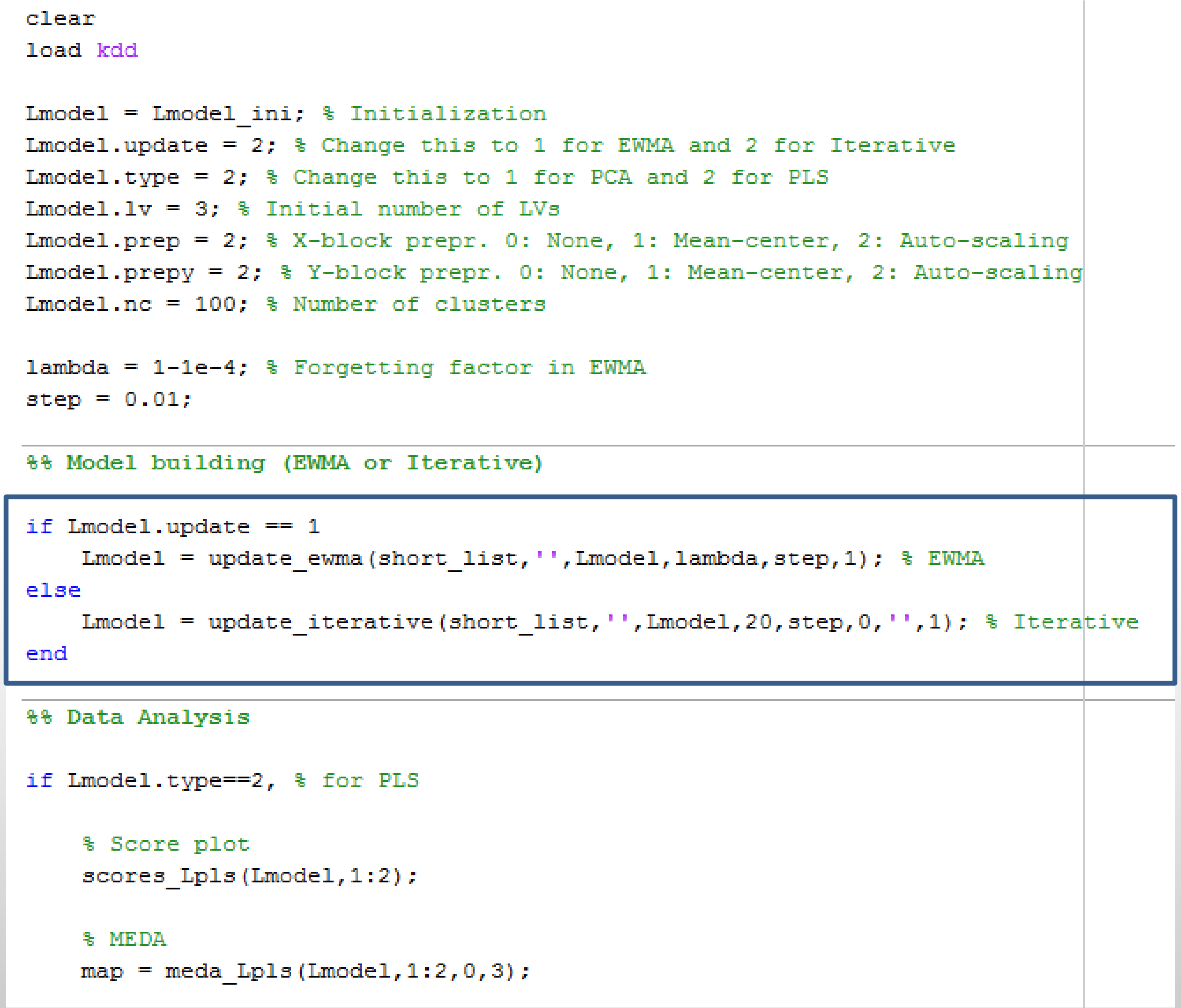
- For observations ➔ Clustering

  - <u>Scalable to any size</u>
  - Compressed Score Plots
  - Compressed MSPC



ChemoLab, (2014) 135: 110

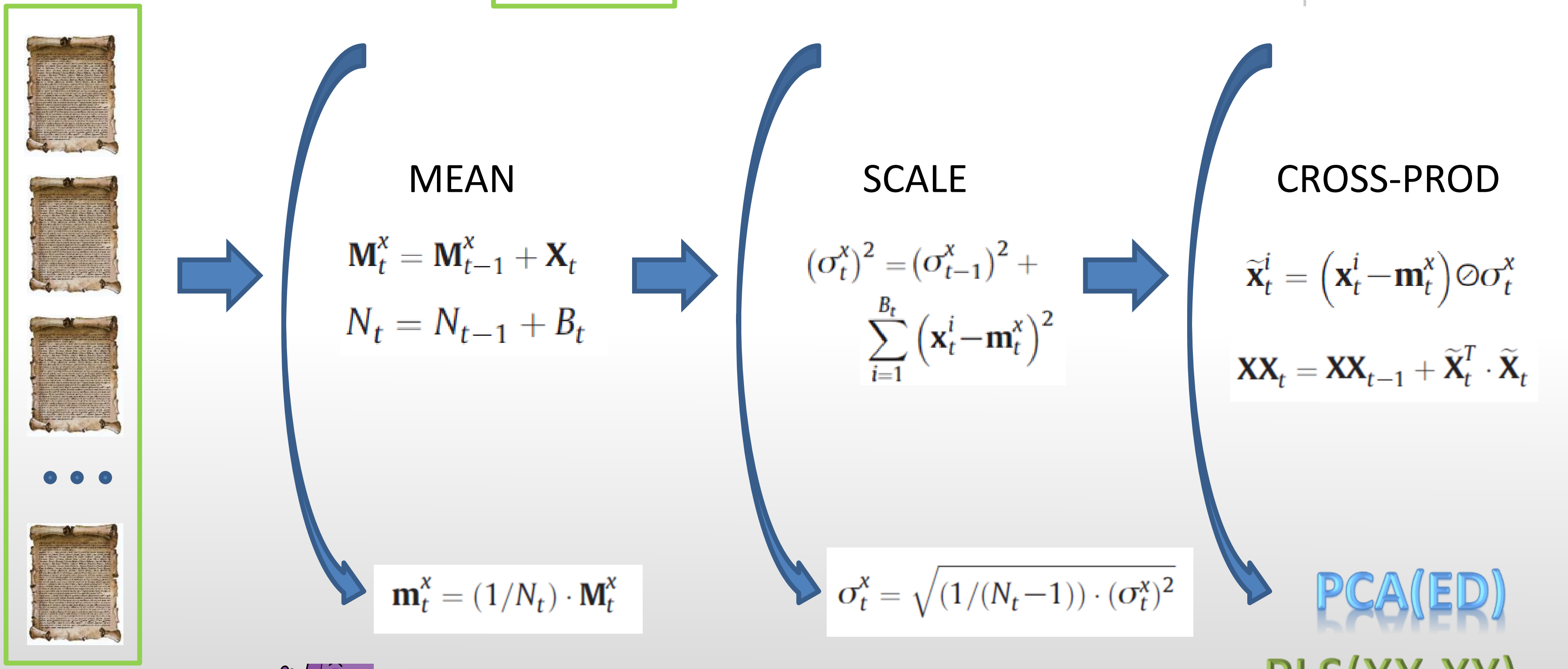
# MEDA FOR BIG DATA



```
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);
```

# BIG DATA SETS

```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```



MEAN

$$\mathbf{M}_t^x = \mathbf{M}_{t-1}^x + \mathbf{X}_t$$

$$N_t = N_{t-1} + B_t$$

$$\mathbf{m}_t^x = (1/N_t) \cdot \mathbf{M}_t^x$$

SCALE

$$(\sigma_t^x)^2 = (\sigma_{t-1}^x)^2 + \sum_{i=1}^{B_t} \left(\mathbf{x}_t^i - \mathbf{m}_t^x\right)^2$$

$$\sigma_t^x = \sqrt{(1/(N_t - 1)) \cdot (\sigma_t^x)^2}$$

CROSS-PROD

$$\tilde{\mathbf{x}}_t^i = \left(\mathbf{x}_t^i - \mathbf{m}_t^x\right) \oslash \sigma_t^x$$

$$\mathbf{XX}_t = \mathbf{XX}_{t-1} + \tilde{\mathbf{X}}_t^T \cdot \tilde{\mathbf{X}}_t$$

PCA(ED)

PLS(XX,XY)

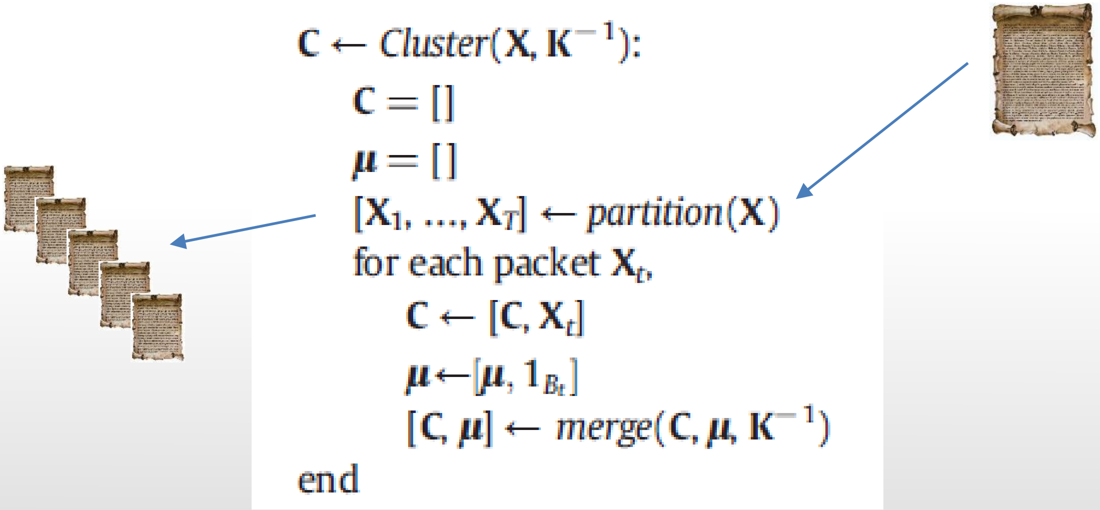
MEDA, Loading plots, CV, ...

ChemoLab, (2014) 135:110

# BIG DATA SETS

```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```

## Compressed Scores



```
$\mathbf{C} \leftarrow Cluster(\mathbf{X}, \mathrm{K}^{-1})$:
    $\mathbf{C} = []$
    $\boldsymbol{\mu} = []$
    $[\mathbf{X}_1, \ldots, \mathbf{X}_T] \leftarrow partition(\mathbf{X})$
    for each packet $\mathbf{X}_t$,
        $\mathbf{C} \leftarrow [\mathbf{C}, \mathbf{X}_t]$
        $\boldsymbol{\mu} \leftarrow [\boldsymbol{\mu}, 1_{B_t}]$
        $[\mathbf{C}, \boldsymbol{\mu}] \leftarrow merge(\mathbf{C}, \boldsymbol{\mu}, \mathrm{K}^{-1})$
end
```

ChemoLab, (2014) 135:110

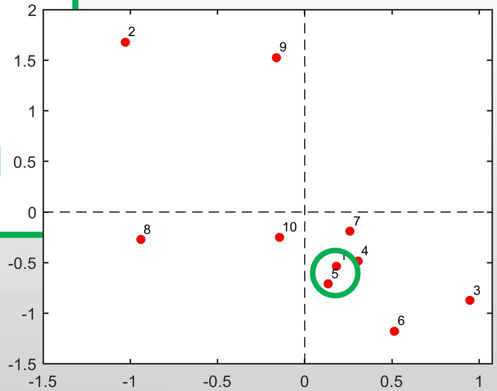
# BIG DATA SETS

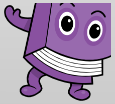
```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```

## Compressed Scores

```
$C \leftarrow Cluster(X, K^{-1})$:
    $C = []$
    $\mu = []$
    $[X_1, \ldots, X_T] \leftarrow partition(X)$
    for each packet $X_t$,
        $C \leftarrow [C, X_t]$
        $\mu \leftarrow [\mu, 1_{B_t}]$
        $[C, \mu] \leftarrow merge(C, \mu, K^{-1})$
end
```

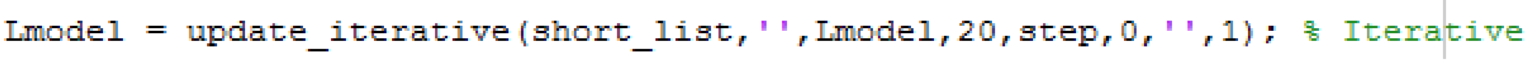
```
$[C, \mu] \leftarrow merge(C, \mu, K^{-1})$:
    $L := \#(C)$
    $C := [c_1, \ldots, c_L]$
    $\mu := [\mu_1, \ldots, \mu_L]$
    while $(\#(C) > L_{end})$,
        $[c_i, c_j] \leftarrow min\_dist(C, K^{-1})$
        $c_i \leftarrow centroid(\mu_i \cdot c_i, \mu_j \cdot c_j)$
        $C \leftarrow [c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_L]$
        $\mu_i \leftarrow \mu_i + \mu_j$
        $\mu \leftarrow [\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_L]$
end
```





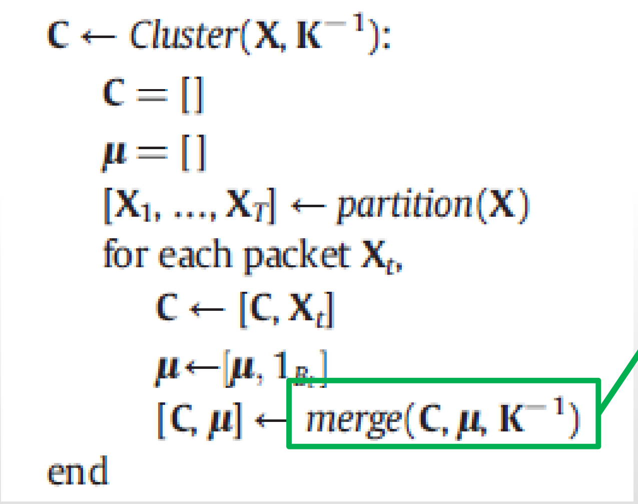
ChemoLab, (2014) 135:110

# BIG DATA SETS

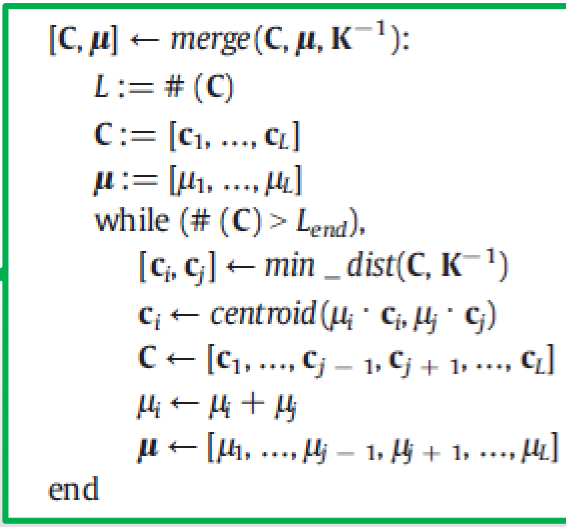


```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```

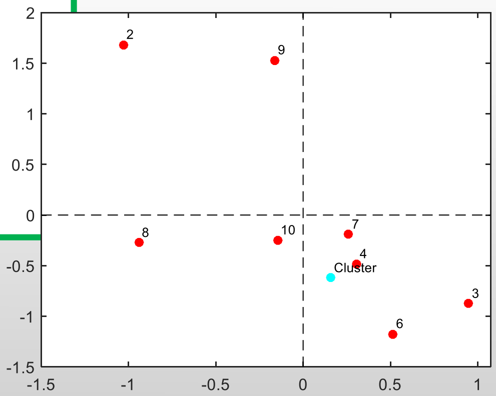
## Compressed Scores

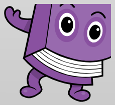


```
$C \leftarrow Cluster(X, K^{-1})$:
    $C = []$
    $\mu = []$
    $[X_1, \ldots, X_T] \leftarrow partition(X)$
    for each packet $X_t$,
        $C \leftarrow [C, X_t]$
        $\mu \leftarrow [\mu, 1_{B_t}]$
        $[C, \mu] \leftarrow \boxed{merge(C, \mu, K^{-1})}$
end
```



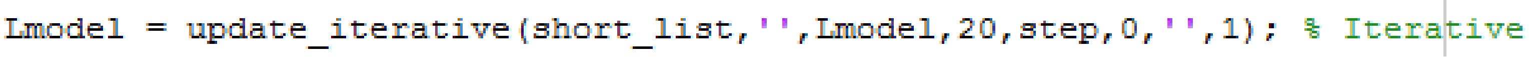
```
$[C, \mu] \leftarrow merge(C, \mu, K^{-1})$:
    $L := \#(C)$
    $C := [c_1, \ldots, c_L]$
    $\mu := [\mu_1, \ldots, \mu_L]$
    while $(\#(C) > L_{end})$,
        $[c_i, c_j] \leftarrow min\_dist(C, K^{-1})$
        $c_i \leftarrow centroid(\mu_i \cdot c_i, \mu_j \cdot c_j)$
        $C \leftarrow [c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_L]$
        $\mu_i \leftarrow \mu_i + \mu_j$
        $\mu \leftarrow [\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_L]$
end
```





ChemoLab, (2014) 135:110

# BIG DATA SETS



```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```

## Compressed Scores

```
$\mathbf{C} \leftarrow Cluster(\mathbf{X}, \mathbf{K}^{-1})$:
    $\mathbf{C} = []$
    $\boldsymbol{\mu} = []$
    $[\mathbf{X}_1, \ldots, \mathbf{X}_T] \leftarrow partition(\mathbf{X})$
    for each packet $\mathbf{X}_t$,
        $\mathbf{C} \leftarrow [\mathbf{C}, \mathbf{X}_t]$
        $\boldsymbol{\mu} \leftarrow [\boldsymbol{\mu}, 1_{B_t}]$
        $[\mathbf{C}, \boldsymbol{\mu}] \leftarrow merge(\mathbf{C}, \boldsymbol{\mu}, \mathbf{K}^{-1})$
end
```

```
$[\mathbf{C}, \boldsymbol{\mu}] \leftarrow merge(\mathbf{C}, \boldsymbol{\mu}, \mathbf{K}^{-1})$:
    $L := \#(\mathbf{C})$
    $\mathbf{C} := [\mathbf{c}_1, \ldots, \mathbf{c}_L]$
    $\boldsymbol{\mu} := [\mu_1, \ldots, \mu_L]$
    while $(\#(\mathbf{C}) > L_{end})$,
        $[\mathbf{c}_i, \mathbf{c}_j] \leftarrow min\_dist(\mathbf{C}, \mathbf{K}^{-1})$
        $\mathbf{c}_i \leftarrow centroid(\mu_i \cdot \mathbf{c}_i, \mu_j \cdot \mathbf{c}_j)$
        $\mathbf{C} \leftarrow [\mathbf{c}_1, \ldots, \mathbf{c}_{j-1}, \mathbf{c}_{j+1}, \ldots, \mathbf{c}_L]$
        $\mu_i \leftarrow \mu_i + \mu_j$
        $\boldsymbol{\mu} \leftarrow [\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_L]$
end
```

$$d_{\mathbf{K}}\left(x_i, x_j\right) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_{\mathbf{K}} = \\ = \left(\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \mathbf{K}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)\right)^{1/2}$$

$$\mathbf{K}^{-1} = \mathbf{P} \cdot \mathbf{K}_{PCA}^{-1} \cdot \mathbf{P}^T,$$

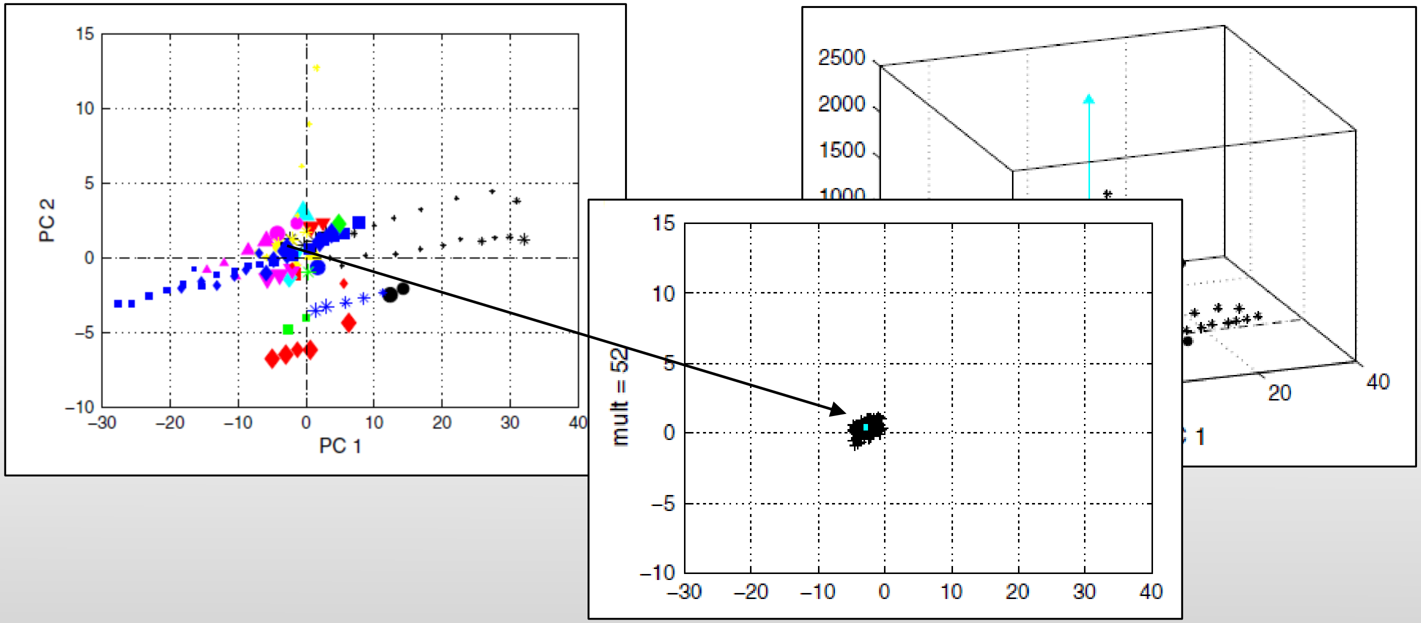
$$\mathbf{K}^{-1} = \mathbf{R} \cdot \mathbf{K}_{PLS}^{-1} \cdot \mathbf{R}^T$$

ChemoLab, (2014) 135:110

# BIG DATA SETS

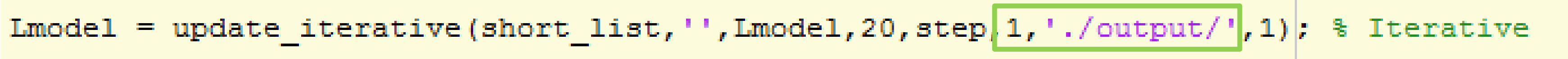
```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,0,'',1); % Iterative
```

## Compressed Score Plot (CSP)



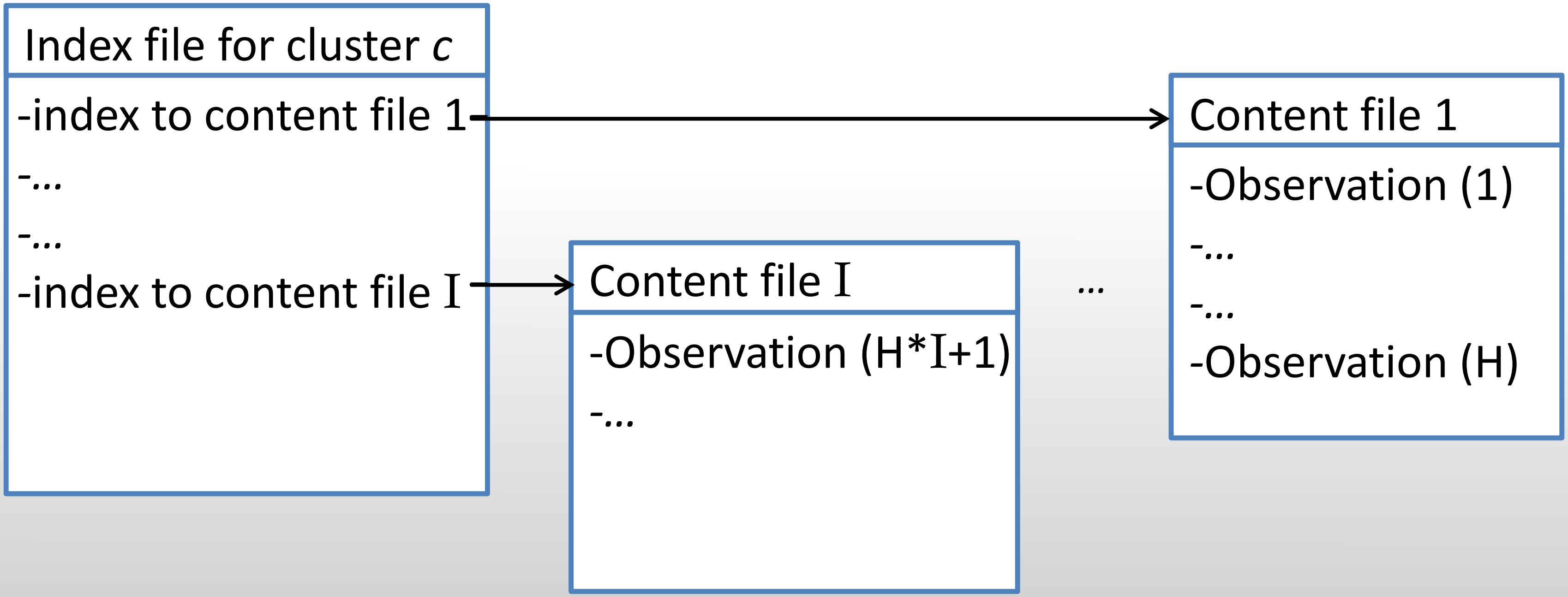
ChemoLab, (2014) 135:110

# BIG DATA SETS



```
Lmodel = update_iterative(short_list,'',Lmodel,20,step,1,'./output/',1); % Iterative
```

## Compressed Score Plot (CSP)



Index file for cluster *c*

-index to content file 1

-…

-…

-index to content file I

Content file I

-Observation (H*I+1)

-…

…

Content file 1

-Observation (1)

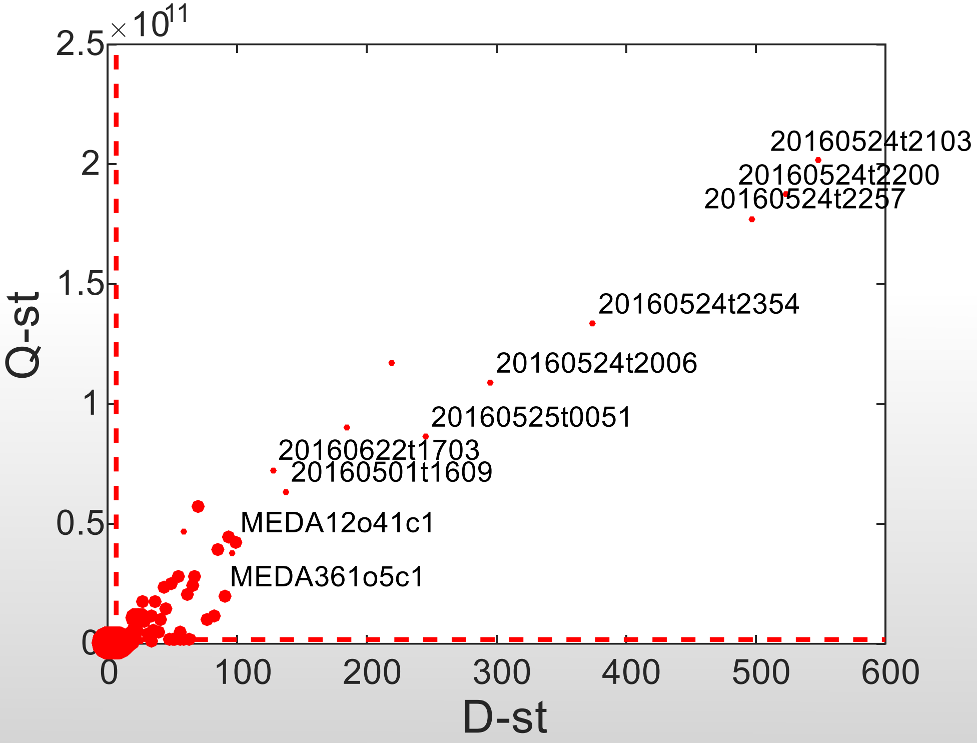
-…

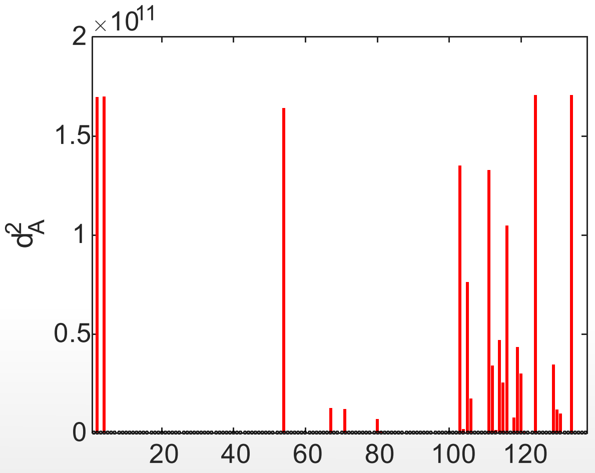
-…

-Observation (H)

# BIG DATA SETS

## Example: vw PCA-MSPC in Big Data (Networkmetrics)

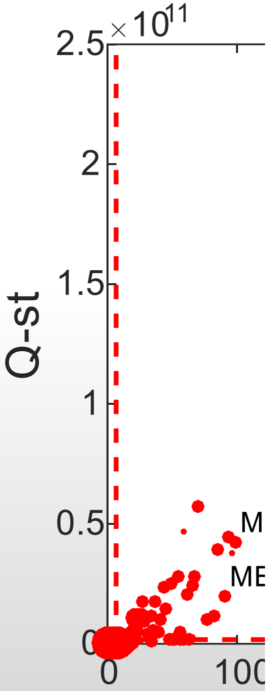


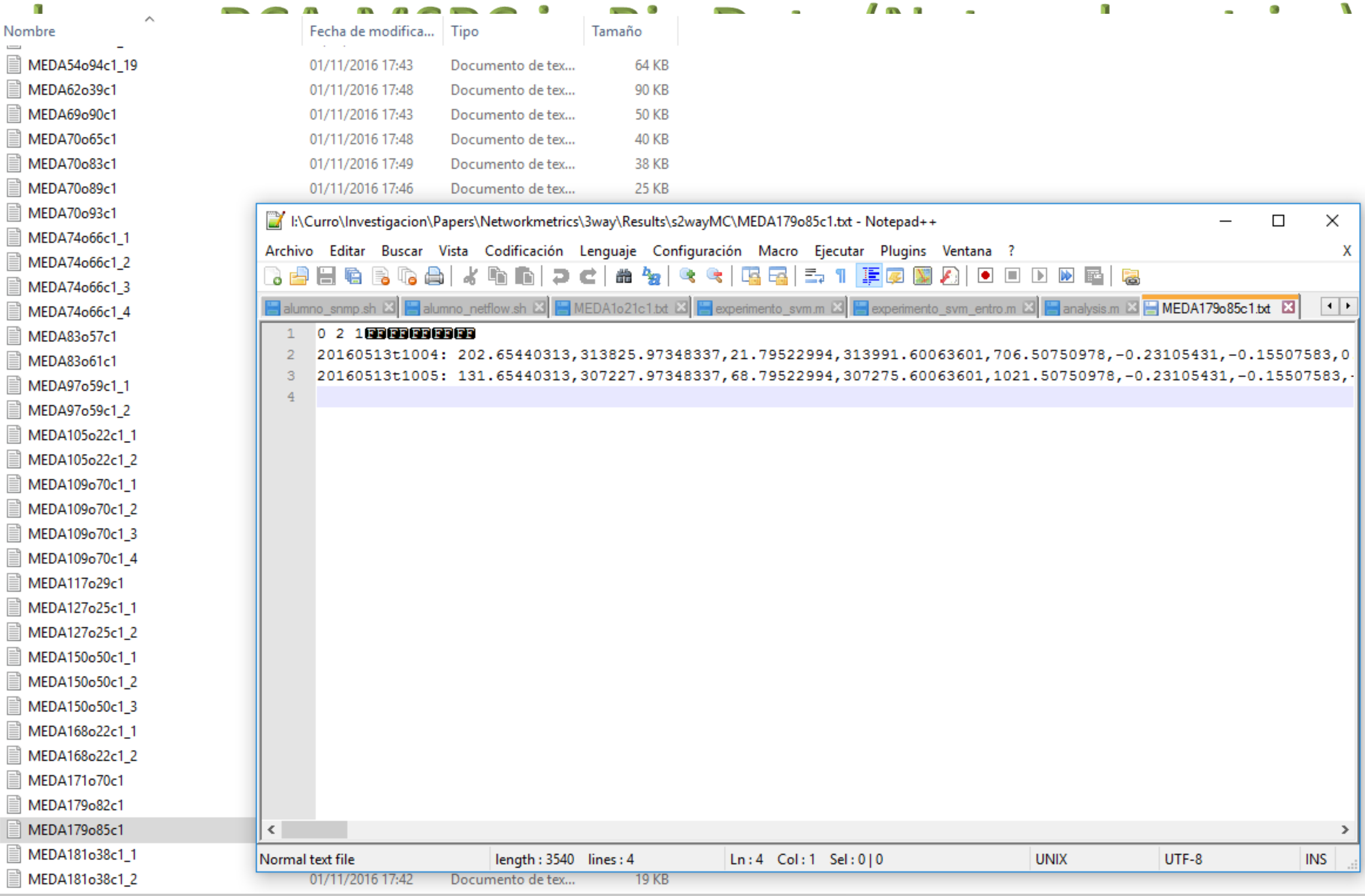


SPAM ATTACK

# BIG DATA SETS

Exam





| Nombre | Fecha de modifica... | Tipo | Tamaño |
|---|---|---|---|
| MEDA54o94c1_19 | 01/11/2016 17:43 | Documento de tex... | 64 KB |
| MEDA62o39c1 | 01/11/2016 17:48 | Documento de tex... | 90 KB |
| MEDA69o90c1 | 01/11/2016 17:43 | Documento de tex... | 50 KB |
| MEDA70o65c1 | 01/11/2016 17:48 | Documento de tex... | 40 KB |
| MEDA70o83c1 | 01/11/2016 17:49 | Documento de tex... | 38 KB |
| MEDA70o89c1 | 01/11/2016 17:46 | Documento de tex... | 25 KB |
| MEDA70o93c1 | | | |
| MEDA74o66c1_1 | | | |
| MEDA74o66c1_2 | | | |
| MEDA74o66c1_3 | | | |
| MEDA74o66c1_4 | | | |
| MEDA83o57c1 | | | |
| MEDA83o61c1 | | | |
| MEDA97o59c1_1 | | | |
| MEDA97o59c1_2 | | | |
| MEDA105o22c1_1 | | | |
| MEDA105o22c1_2 | | | |
| MEDA109o70c1_1 | | | |
| MEDA109o70c1_2 | | | |
| MEDA109o70c1_3 | | | |
| MEDA109o70c1_4 | | | |
| MEDA117o29c1 | | | |
| MEDA127o25c1_1 | | | |
| MEDA127o25c1_2 | | | |
| MEDA150o50c1_1 | | | |
| MEDA150o50c1_2 | | | |
| MEDA150o50c1_3 | | | |
| MEDA168o22c1_1 | | | |
| MEDA168o22c1_2 | | | |
| MEDA171o70c1 | | | |
| MEDA179o82c1 | | | |
| MEDA179o85c1 | | | |
| MEDA181o38c1_1 | | | |
| MEDA181o38c1_2 | 01/11/2016 17:42 | Documento de tex... | 19 KB |

I:\Curro\Investigacion\Papers\Networkmetrics\3way\Results\s2wayMC\MEDA179o85c1.txt - Notepad++

```
0 2 1
20160513t1004: 202.65440313,313825.97348337,21.79522994,313991.60063601,706.50750978,-0.23105431,-0.15507583,0
20160513t1005: 131.65440313,307227.97348337,68.79522994,307275.60063601,1021.50750978,-0.23105431,-0.15507583,-
```

Normal text file | length : 3540 lines : 4 | Ln : 4 Col : 1 Sel : 0 | 0 | UNIX | UTF-8 | INS

# Data Streams

```
Lmodel = update_ewma(short_list,'',Lmodel,lambda,step,1); % EWMA
```

$$\mathbf{M}_t^x = \lambda \cdot \mathbf{M}_{t-1}^x + \mathbf{X}_t$$

$$\mathbf{m}_t^x = (1/N_t) \cdot \mathbf{M}_t^x$$

$$N_t = \lambda \cdot N_{t-1} + B_t$$

$$(\sigma_t^x)^2 = \lambda \cdot (\sigma_{t-1}^x)^2 + \sum_{i=1}^{B_t} \left(\mathbf{x}_t^i - \mathbf{m}_t^x\right)^2$$

$$\sigma_t^x = \sqrt{(1/(N_t - 1)) \cdot (\sigma_t^x)^2}$$

$$\tilde{\mathbf{x}}_t^i = \left(\mathbf{x}_t^i - \mathbf{m}_t^x\right) \oslash \sigma_t^x$$

$$\mathbf{XX}_t = \lambda \cdot \mathbf{XX}_{t-1} + \tilde{\mathbf{X}}_t^T \cdot \tilde{\mathbf{X}}_t$$

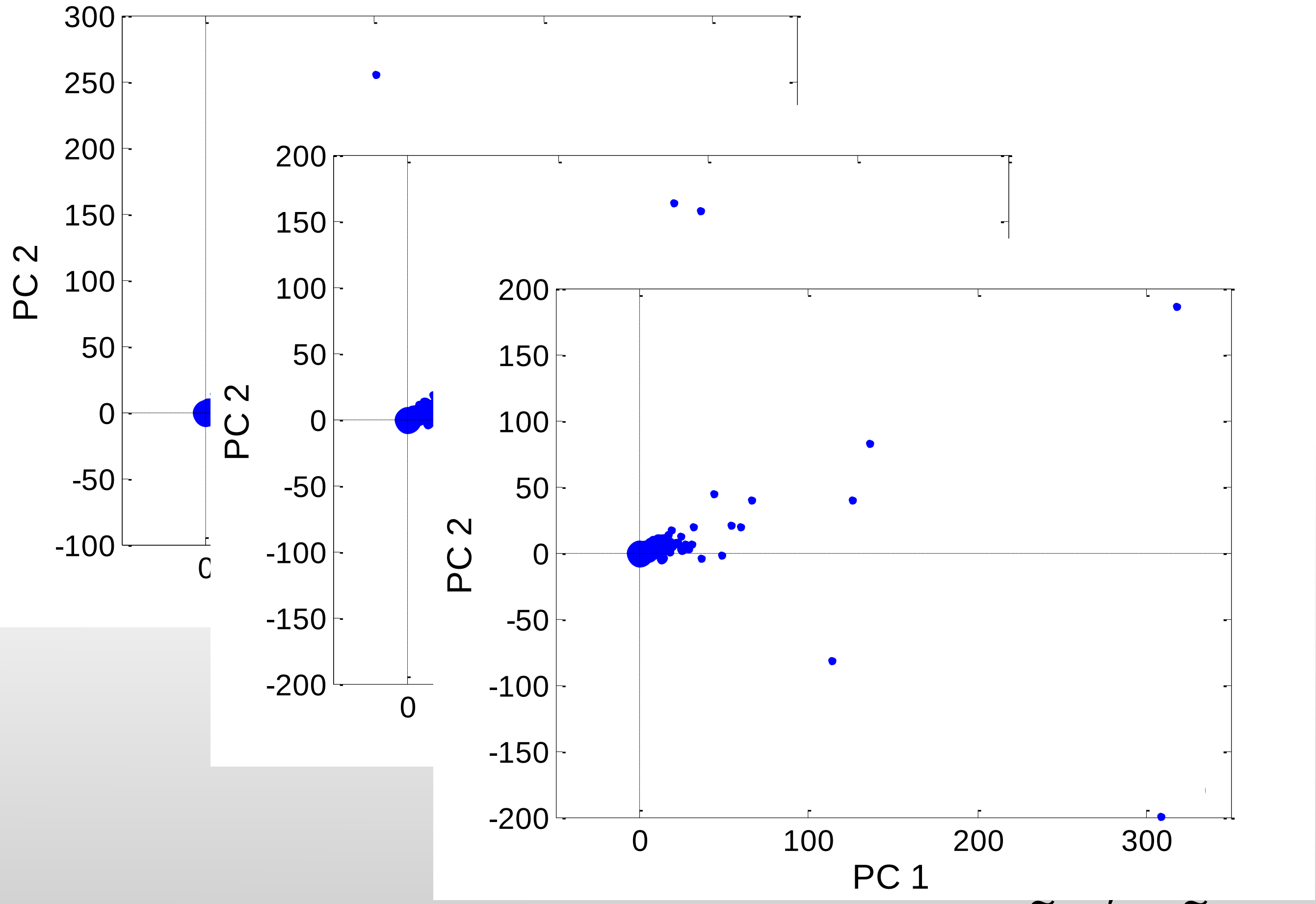
PCA(ED)

PLS (XX,XY)

MEDA,
Loading plots,
CV, ...

J.P.C., (1997) 7:169
ChemoLab, (2014) 135:110

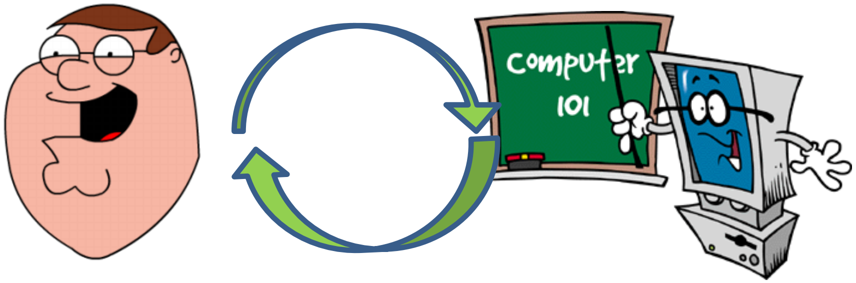
# Data Streams



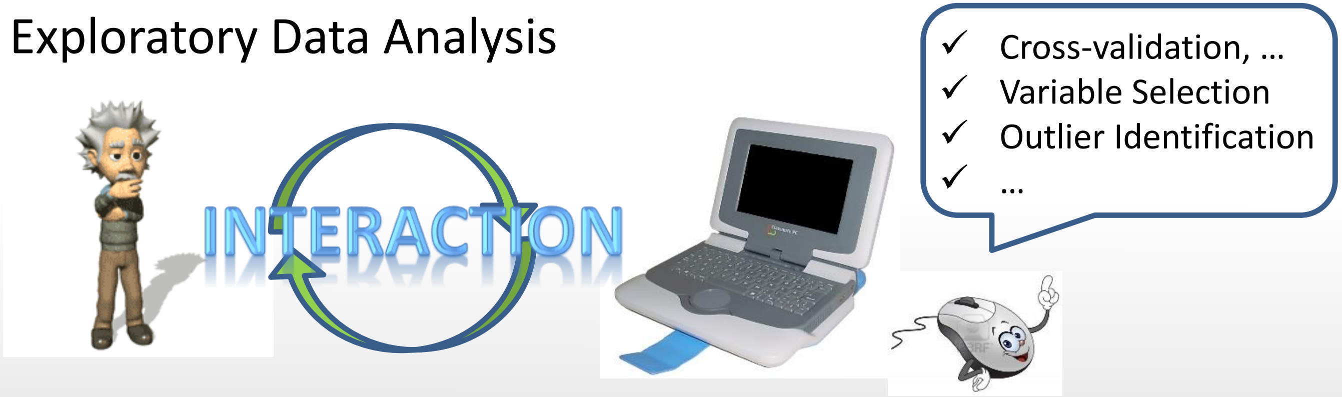
$$Lmodel\ at\ t\colon\ (X'X)_t = 0.9 \cdot (X'X)_{t-1} + \tilde{X}_t{}' \cdot \tilde{X}_t$$

# *i*MEDA: MEDA + VISUAL ANALYTICS

- Data mining / Machine learning



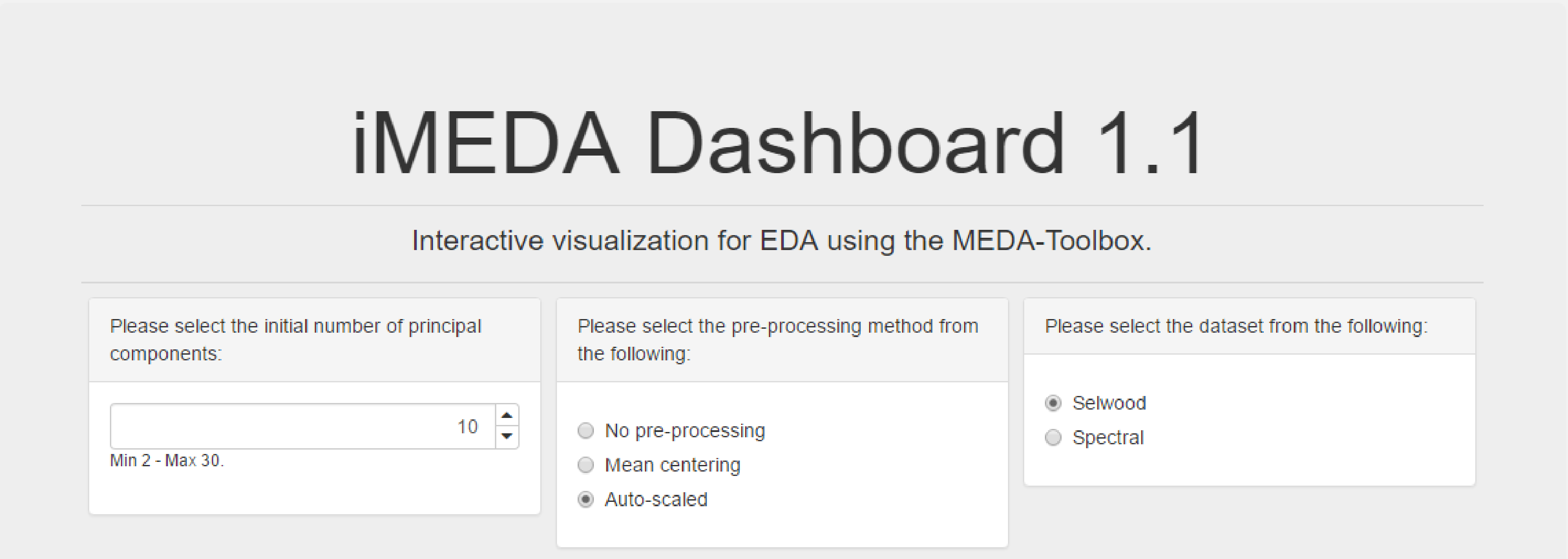
- Chemometrics / Exploratory Data Analysis



✓ Cross-validation, …
✓ Variable Selection
✓ Outlier Identification
✓ …

- EDA + Visual Analytics

# iMEDA: MEDA + VISUAL ANALYTICS



## iMEDA Dashboard 1.1

Interactive visualization for EDA using the MEDA-Toolbox.

Please select the initial number of principal components:

10

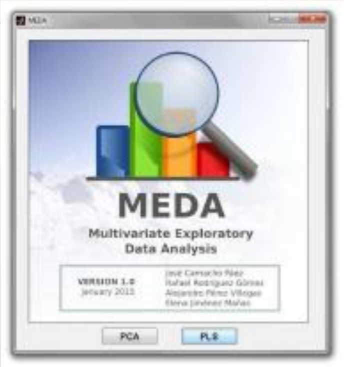
Min 2 - Max 30.

Please select the pre-processing method from the following:

○ No pre-processing
○ Mean centering
● Auto-scaled

Please select the dataset from the following:

● Selwood
○ Spectral



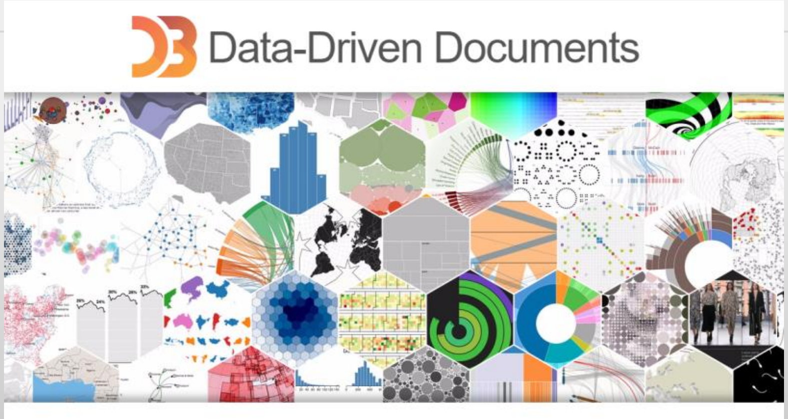
MEDA

Multivariate Exploratory Data Analysis

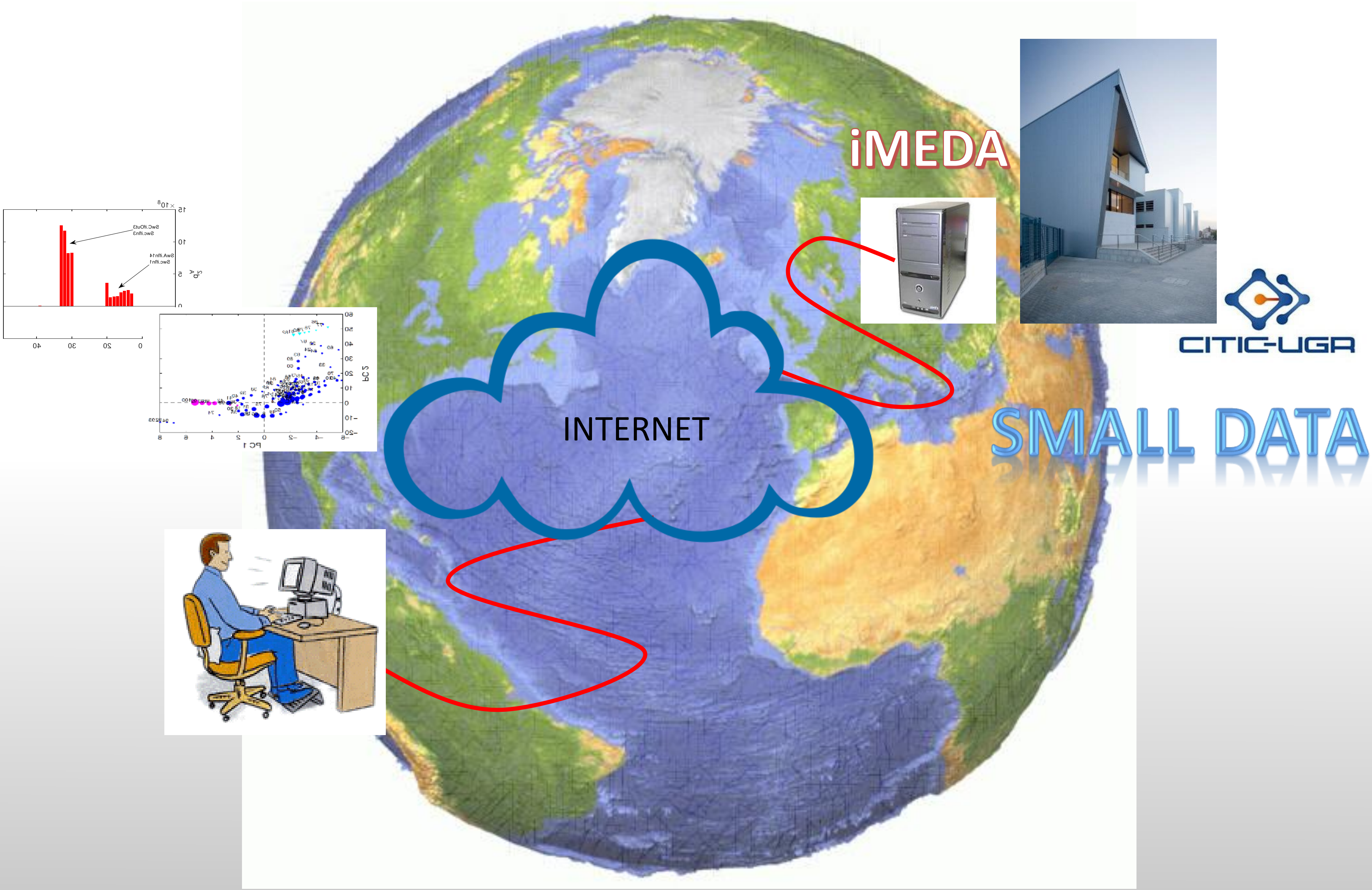
VERSION 1.0
January 2015

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

OCTAVE



# THE FUTURE: DATA ANALYSIS AS A SERVICE (DAAS)

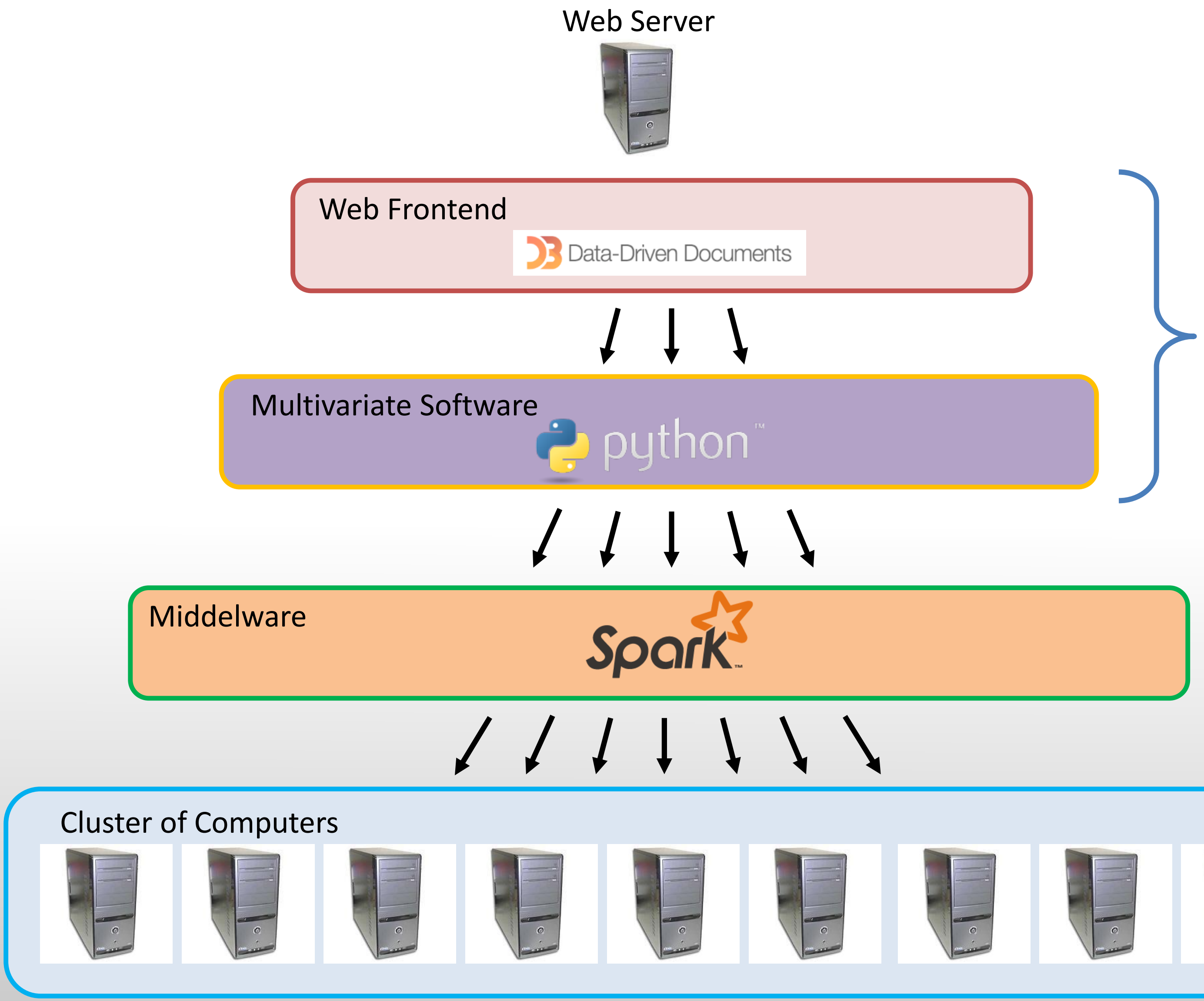


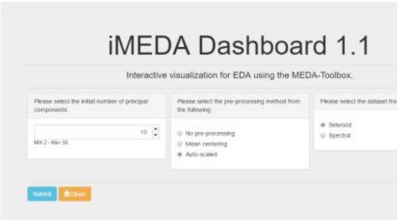
iMEDA

INTERNET

SMALL DATA

# THE FUTURE II: *i*MEDA + BIG DATA TOOLS





iMEDA Dashboard 1.1

Interactive visualization for EDA using the MEDA-Toolbox.

Please select the initial number of principal components:

10

Min 2 - Max 30.

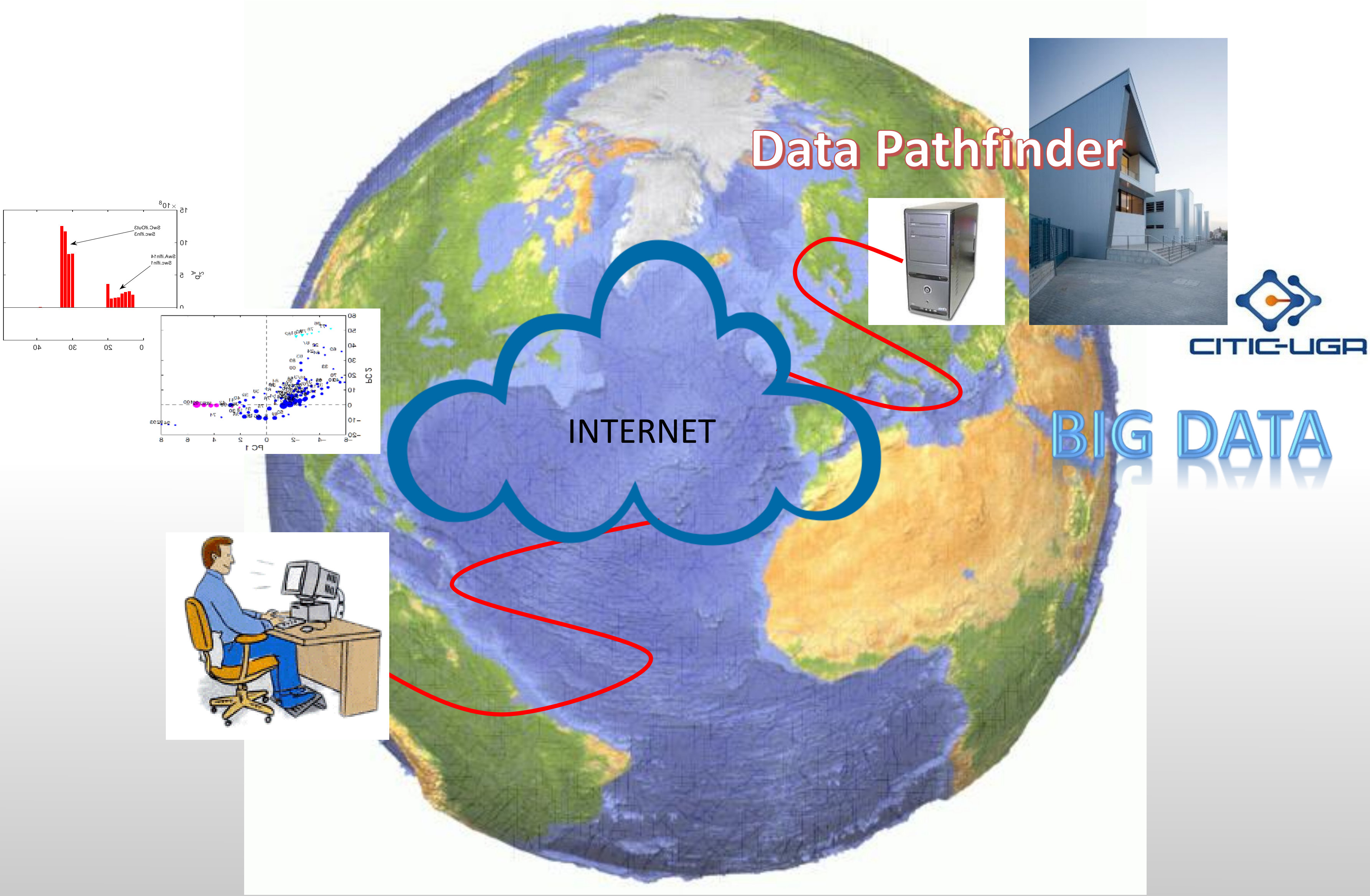
Please select the pre-processing method from the following:

○ No pre-processing
○ Mean centering
● Auto-scaled

Please select the dataset fro

● Belwood
○ Spectral

# THE FUTURE II: DAAS FOR BIG DATA



# QUESTIONS?

Is this the way for MA + BD?

Interested?

IS IT TO ACHIEVE
FULL DATA INTERACTIVITY?

# Exploratory Analysis on Big Data using the MEDA Toolbox

## Present and Future

**José Camacho,**

**Roberto Therón,**

**Roberto Magán**

*Departamento de Teoría de la Señal, Telemática y Comunicaciones*

*Universidad de Granada*

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

ugr Universidad de Granada