



ChemFlow, chemometrics using Galaxy

Virginie Rossard, Jean Claude Boulet, Fabien Gogé, Eric Latrille, Jean-Michel Roger

► To cite this version:

Virginie Rossard, Jean Claude Boulet, Fabien Gogé, Eric Latrille, Jean-Michel Roger. ChemFlow, chemometrics using Galaxy. Galaxy Community Conference - GCC2016, Indiana University. USA., Jun 2016, Bloomington, United States. 10.7490/f1000research.1112573.1 . hal-01837798

HAL Id: hal-01837798

<https://hal.science/hal-01837798>

Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ChemFlow, chemometrics using Galaxy

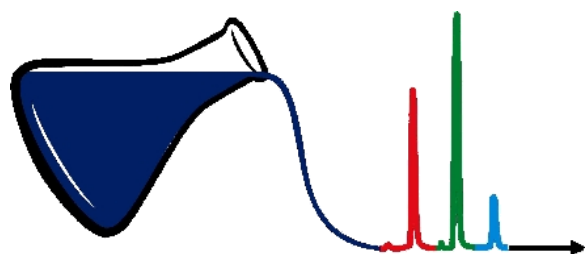
V. Rossard^{*}, J.C. Boulet^{*}, F. Gogé^{**}, E. Latrille^{*}, J.M. Roger^{**}

^{*}INRA, Institut National de la Recherche Agronomique

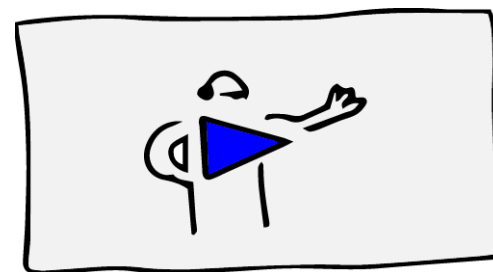
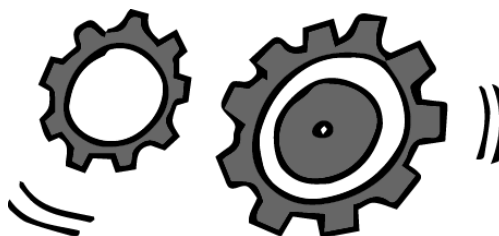
^{**}IRSTEA, Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture



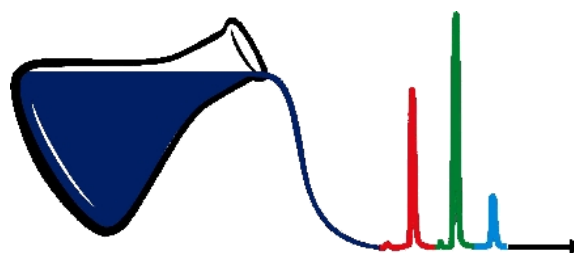
1. Chemometrics
2. The ChemFlow Tool
3. The MOOC



By courtesy of the Division of chemometrics,
Norwegian Chemical Society



1. Chemometrics
2. The ChemFlow Tool
3. The MOOC



By courtesy of the Division of chemometrics,
Norwegian Chemical Society

Definition (International Chemometrics Society):

Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.

Chemical system /
process

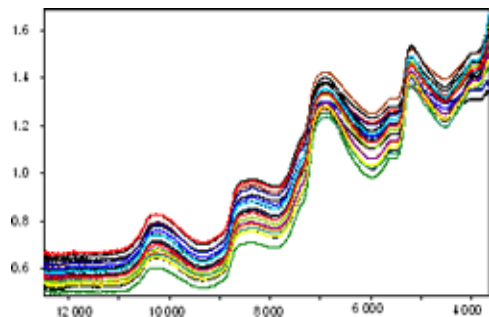
Measurements

Chemometrics

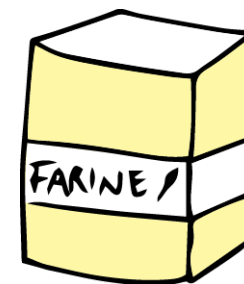
State

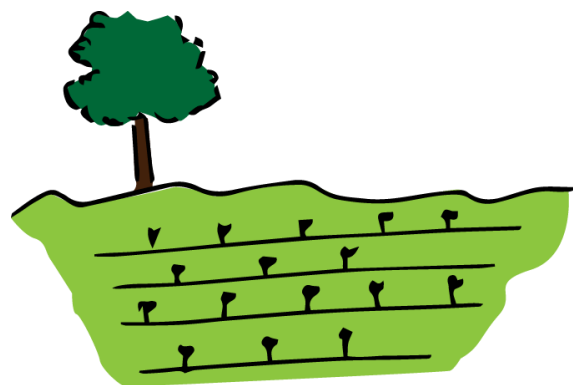
- Multivariate model
- Mathematics
- Statistics

- Gluten content of flour
- Year of production



Infrared spectra





Application domain



Product



Sample



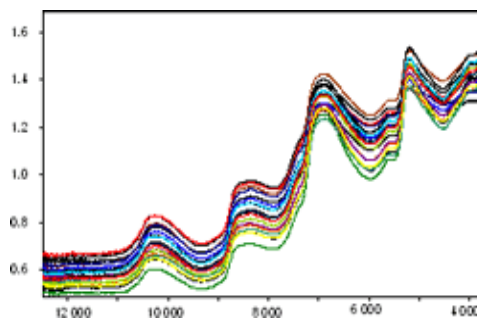
Operator

Chemometrics

```
LOAD('MY DATA.CSV');  
PLOT(X');  
X=DETREND(X)  
RESPLS=PLS(X,Y,10,20,1)  
PLOT(RESPLS...
```

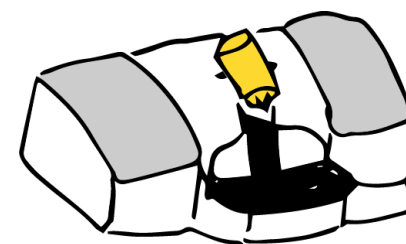


Measurements

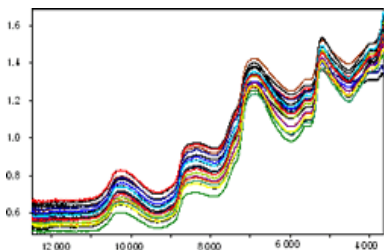
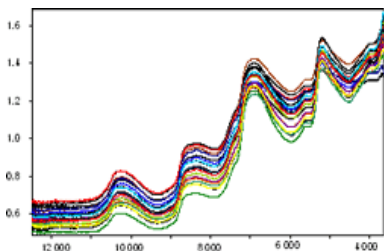


Infrared spectra

Spectrometer



- Gluten content of flour
- Year of production

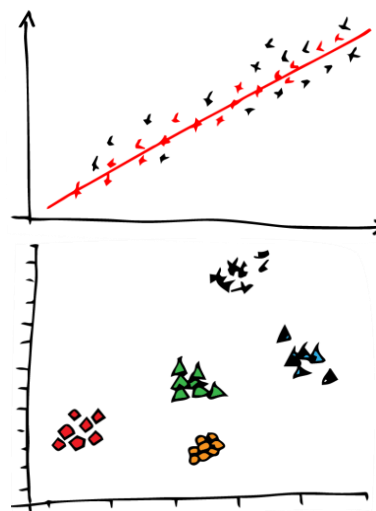


Infrared spectra

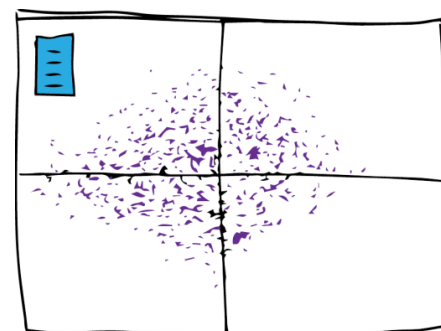


To

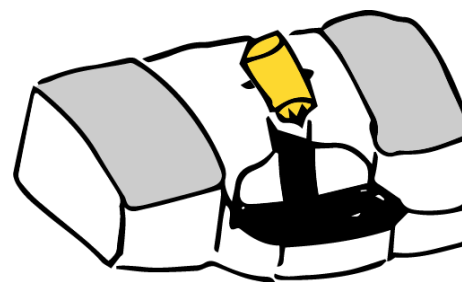
```
LOAD('MY DATA.CSV');  
PLOT(X);  
X=DETREND(X)  
RESPLS=PLS(X,Y,10,20,1)  
PLOT(RESPLS...
```



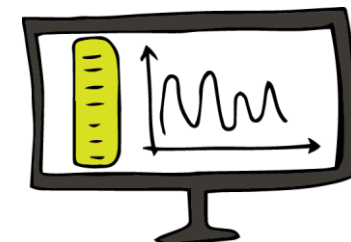
Predict biochemical content using a model



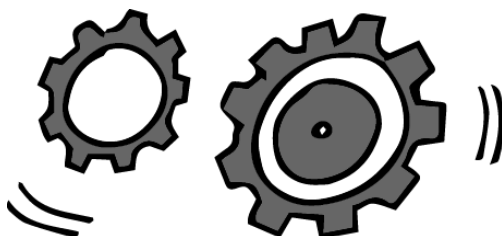
Laboratory



Inexpensive, faster,
preserve sample



1. Chemometrics
2. **The ChemFlow Tool**
3. The MOOC



Common practice

Specificities and added values of ChemFlow

One software for each type of spectrometer

Accessibility, web service

The collaborators develop tools in different languages

Reinforce the chemometrics community by **collaborative, multi-users and shared** tools

Developed in various languages

Interoperability



Difficulties to reproduce data treatment chains

Reproducibility and traceability: **workflow**

- User friendly software: all copyright.
- Free software: programming.

User friendly and **free** software

Common practice

Specificities and added values of ChemFlow

One software for each type of spectrometer

Accessibility, web service

The co
language

Develo



Galaxy

PROJECT

by
tools

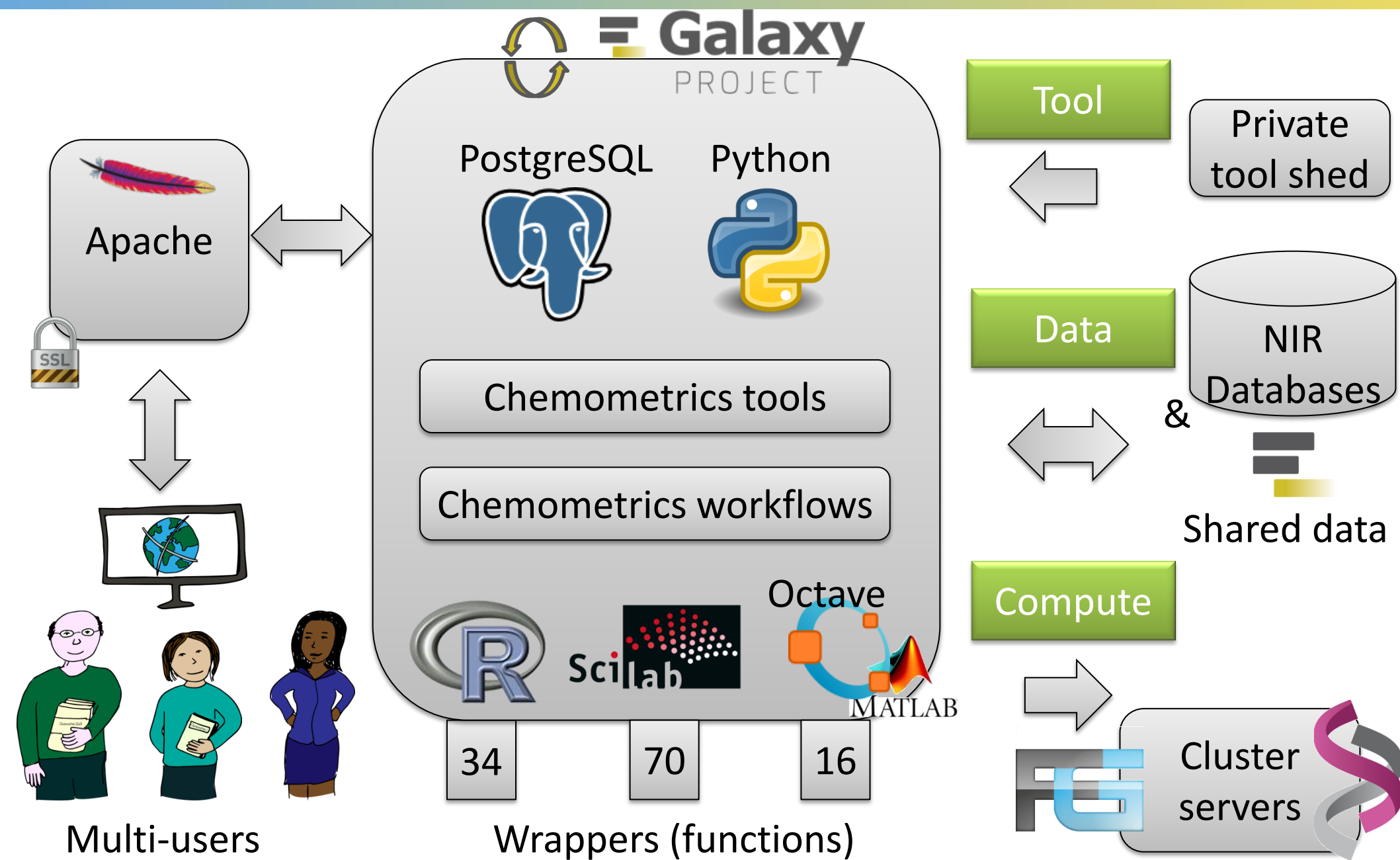
Difficul
chains

Application to spectral analysis

traceability. workflow

- User friendly software: all copyright.
- Free software: programming.

User friendly and free software





Step 1

Register an
account
automatically

Step 2

Connect to
ChemFlow

Step 3

Upload your
data OR use
shared data

Step 4

Play
chemometrics

A tool with a configuration immediately useful
CONNECT & PLAY



&



Galaxy

https://chemserver.supagro.inra.fr/user/login?use_panels=True

Galaxy Analyze Data Workflow Shared Data Visualization Help User

This Galaxy instance has been configured such that only users who are logged in may use it. If you don't already have an account, you can create one here.

Login

Username / Email Address:

Password:

[Forgot password? Reset here](#)

Login

https://chemserver.supagro.inra.fr/workflow

In Google Chrome, the tool



is automatically translate

In Google Chrome, the tool



is automatically translate

Galaxie x

← → ↻ https://chemserver.supagro.inra.fr/user/login?use_panels=True

Galaxie Analyser les données flux de travail données partagées Visualisation aider Utilisateur

Cette instance Galaxy a été configuré de telle sorte que seuls les utilisateurs qui sont connectés peuvent l'utiliser. Si vous ne possédez pas déjà un compte, [vous pouvez en](#)

S'identifier

Nom d'utilisateur / Adresse e-mail:

Mot de passe:

[mot de passe oublié? réinitialiser ici](#)

S'identifier

Galaxy <https://chemserver.supagro.inra.fr> Using 73.7 MB

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Capture Fenêtre

search tools

Import Data

Convert data format

[Convert](#) delimiters to TAB

[Convert MAT or RDATA to TAB](#)
convert mat or rdata file in tab file

[Convert MAT or RDATA to xml](#)
convert mat or rdata file in xml file

Utils

[Merge files](#)

[Edit files](#) delete or extract row(s) or column(s) from a file

[Sort files](#) sort files

[Transpose](#) transpose matrix file

[Command](#) execute line command

Statistics

Plot

Calibration/Validation

Pretreatment

ChemFlow

Chemometrics without programming Welcome !

More informations : [ChemProject](#)
If you have a problem : Virginie.Rossard@supagro.inra.fr

History

search datasets

pretreat
17 shown, 1 deleted
26.1 MB

17: PCA model:SG(Xnir)

16: PCA explained variance(%):SG(Xnir)

15: PCA eigenvectors:SG(Xnir)

14: PCA scores:SG(Xnir)

13: Spectra plot SG (Xnir) degPol1

12: SG(Xnir)

11: Spectra plot on data 6

10: Spectra plot on data 5

ChemFlow: plot interface

Galaxy <https://chemserver.supagro.inra.fr> Using 35.4 MB

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

- Import Data
- Convert data format
- Utils
- Statistics
- Plot**
 - Histogram
 - Boxplot
 - Barplot
 - Correlation plot
 - Spectra plot
 - Heatmap
 - Scatter plot for multiple series and graph types
 - Interactive graph create interactive plot with Rshiny in a new tab
 - Interactive graph create interactive plot with ggvis and Rshiny in a new tab
- Calibration/Validation

History

- 58: Xnir (187 lines, format: tabular, database: ?)
- 57: origin e huile
- 56: SG(x.c sv)
- 55: NIPAL S-PLS on Y(2):model
- 54: NIPAL

Warning: This dataset is large and only the first megabyte is shown below. [Show all](#) | [Save](#)

	1000	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011
AP01	-0.04844		-0.048245		-0.048092		-0.047762		-0.047316		-0.046829	
AP02	-0.039741		-0.039554		-0.039632		-0.039297		-0.039018		-0.03855	
AP03	-0.044596		-0.044463		-0.044075		-0.043648		-0.043265		-0.042964	
AP04	-0.046453		-0.046489		-0.046504		-0.046383		-0.045753		-0.044689	
AP05	-0.04702		-0.046511		-0.046133		-0.045755		-0.044943		-0.044656	
AP06	-0.049078		-0.049004		-0.048913		-0.048488		-0.048018		-0.04748	
AP07	-0.049906		-0.049996		-0.049674		-0.049195		-0.048522		-0.048037	
AP08	-0.050282		-0.050152		-0.049891		-0.049495		-0.049185		-0.048511	
AP09	-0.048378		-0.048315		-0.048596		-0.048738		-0.048497		-0.047873	
AP10	-0.046083		-0.045663		-0.04555		-0.045343		-0.045024		-0.044771	
AP11	-0.050291		-0.050357		-0.050325		-0.050178		-0.049881		-0.049257	
AP12	-0.04934		-0.049129		-0.048792		-0.048361		-0.047875		-0.047513	
AP13	-0.048376		-0.048506		-0.048463		-0.047818		-0.047171		-0.046934	
AP14	-0.046614		-0.046385		-0.045998		-0.045632		-0.045188		-0.044723	
AP15	-0.047614		-0.047479		-0.047122		-0.046992		-0.046829		-0.046101	
AP16	-0.050376		-0.050324		-0.050235		-0.049835		-0.049539		-0.048966	
AP17	-0.052726		-0.052701		-0.052205		-0.051963		-0.051723		-0.050938	
AP18	-0.050036		-0.049736		-0.049392		-0.049131		-0.048745		-0.048277	
AP19	-0.053388		-0.053846		-0.053827		-0.053454		-0.052955		-0.052495	
AP21	-0.046617		-0.046224		-0.045481		-0.04534		-0.045222		-0.044811	
AP22	-0.050909		-0.050995		-0.051167		-0.050922		-0.049724		-0.048774	
AP23	-0.052071		-0.05187		-0.05169		-0.051759		-0.051451		-0.050704	
AP24	-0.05252		-0.052267		-0.051943		-0.051569		-0.051111		-0.050512	
			-0.047917		-0.047714		-0.04746		-0.04711		-0.046675	

Galaxy <https://chemserver.supagro.inra.fr> Using 38.8 MB

Tools **Statistics** **Plot** **Calibration/Validation** **Pretreatment** **Exploration** **Regressions** **Clustering**

Plot

- [Histogram](#)
- [Boxplot](#)
- [Barplot](#)
- [Correlation plot](#)
- [Spectra plot](#)**
- [Heatmap](#)
- [Scatter plot](#) for multiple series and graph types
- [Interactive graph](#) create interactive plot with Rshiny in an new tab
- [Interactive graph](#) create interactive plot with ggvis and Rshiny in an new tab

Spectra Plot

History

search datasets

ChemFlow
50 shown, 26 [deleted](#)
29.36 MB

76: Spectra plot on data 58
971.4 KB
format: **pdf**, database: ?
Image in pdf format

72: Trans(Xnir)

66: NIPALS-PLS on Y (None (value not yet validated)):model

65: NIPALS-PLS on Y (None (value not yet validated)):ypredcv

64: NIPALS-PLS on Y (None (value not yet validated)):ypredcv

ChemFlow: chemometrics tools interface

Galaxy <https://chemserver.supagro.inra.fr> Using 35.4 MB

Tools Capture Fenêtre

Calibration/Validation

Pretreatment

Exploration

Regressions

PLS Regression Partial Least Square

PCR Regression Principal component Regression

MLR Regression Multiple Linear Regression

SLR Simple Linear Regression

Applies a regression model to a new set of spectra

Clustering

Discrimination

Variable selection

Orthogonal projection

Calibration transfert

Unmixing/ Curve Resolution

Multitable/Multiway Analysis

58: Xnir
Dataset (n x p) containing the n spectra of p variables.

Select y data
60: y_triglycerides
Dataset (n x q) containing the reference values.

Column of y data chosen for the calculation
c3: OLnL

Algorithm choice:
NIPALS

cross-validation type:
k-bloc cross-validation

Number of blocs for cross-validation
4

Number of latent variables
20

Centering option
yes

compute outliers statistics

History

search datasets

ChemFlow
50 shown, 20 deleted
25.99 MB

68: Scatter plot on data 65 and data 60

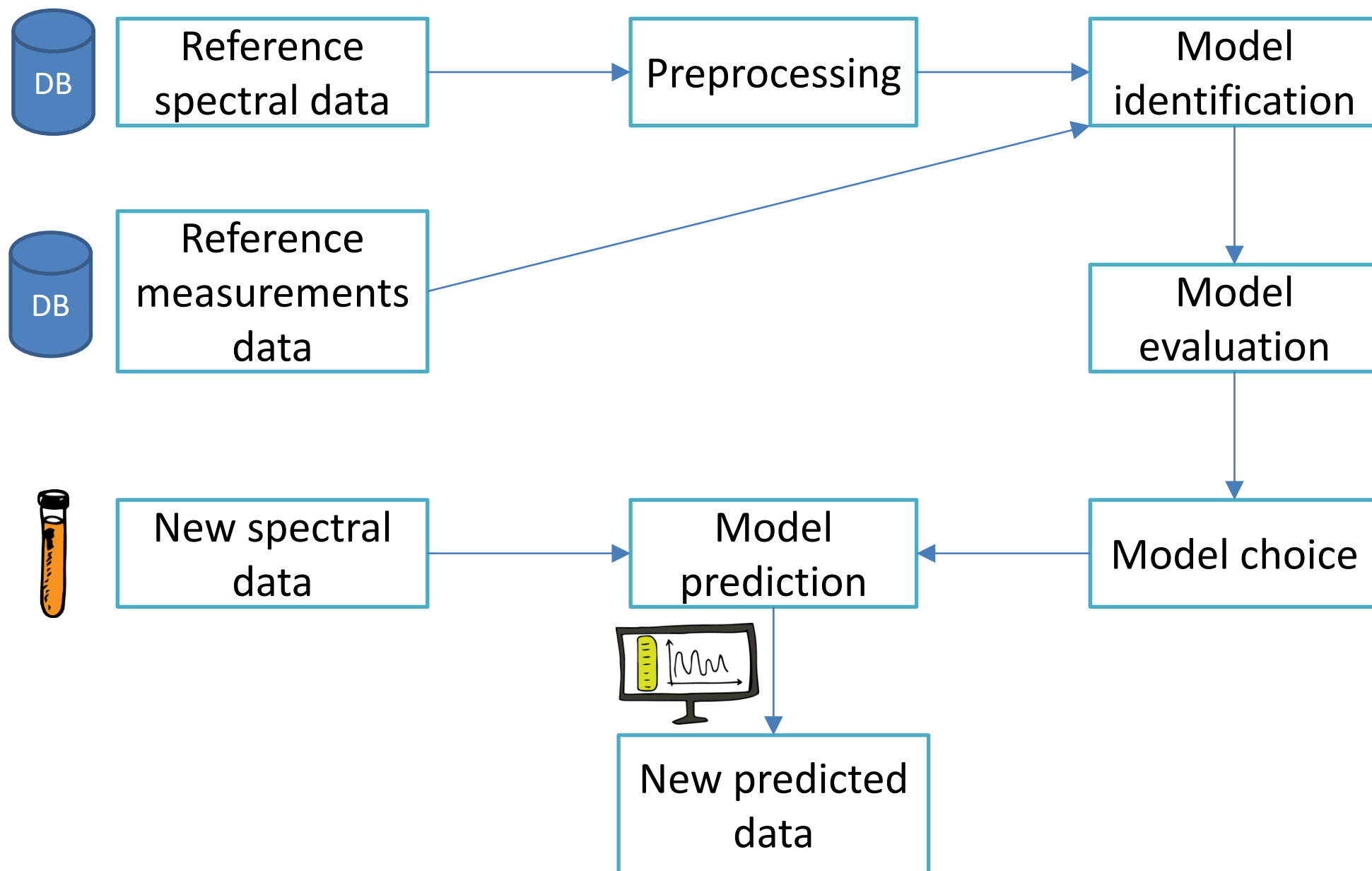
67: Scatter plot on data 63

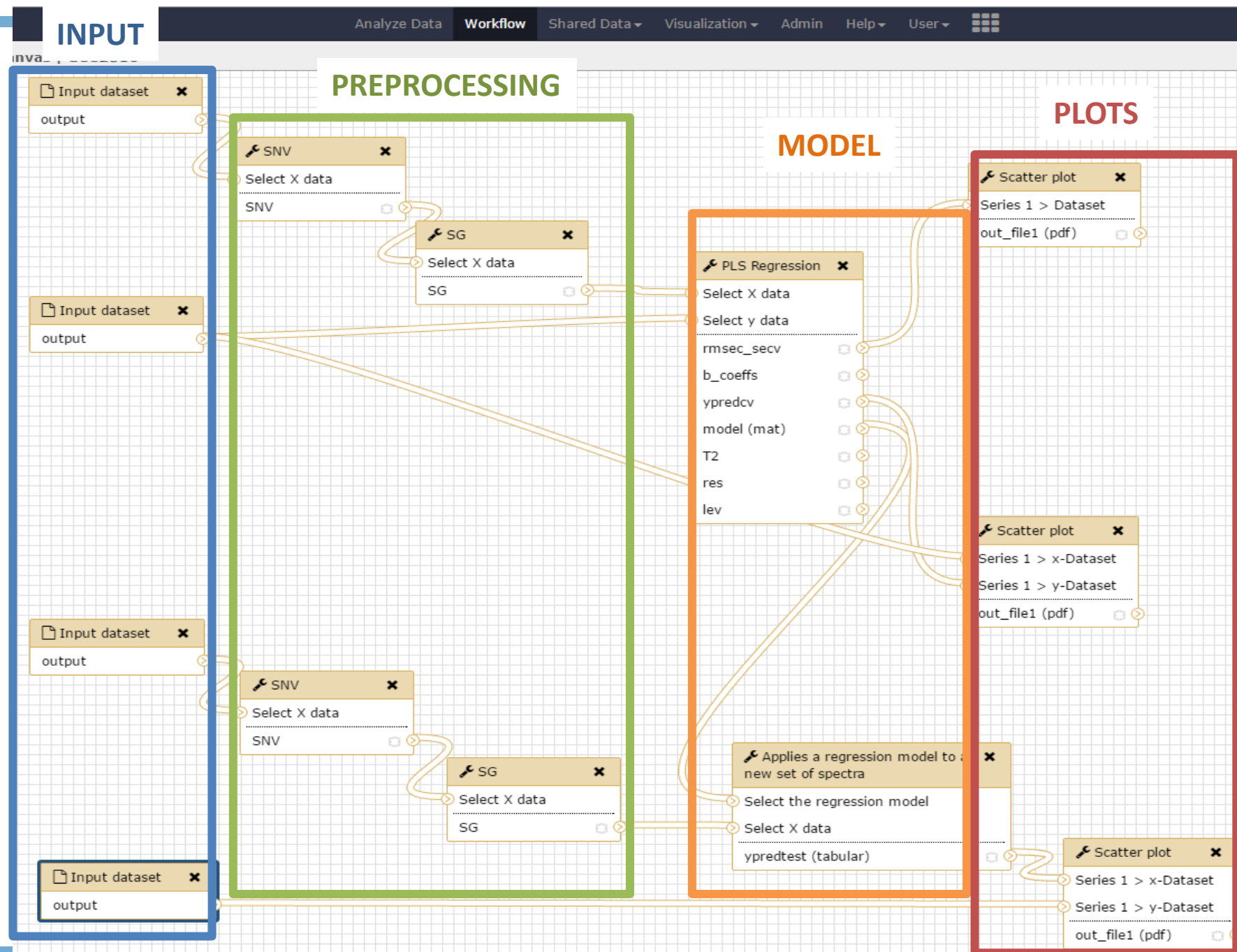
66: NIPALS-PLS on Y (None (value not yet validated)):model

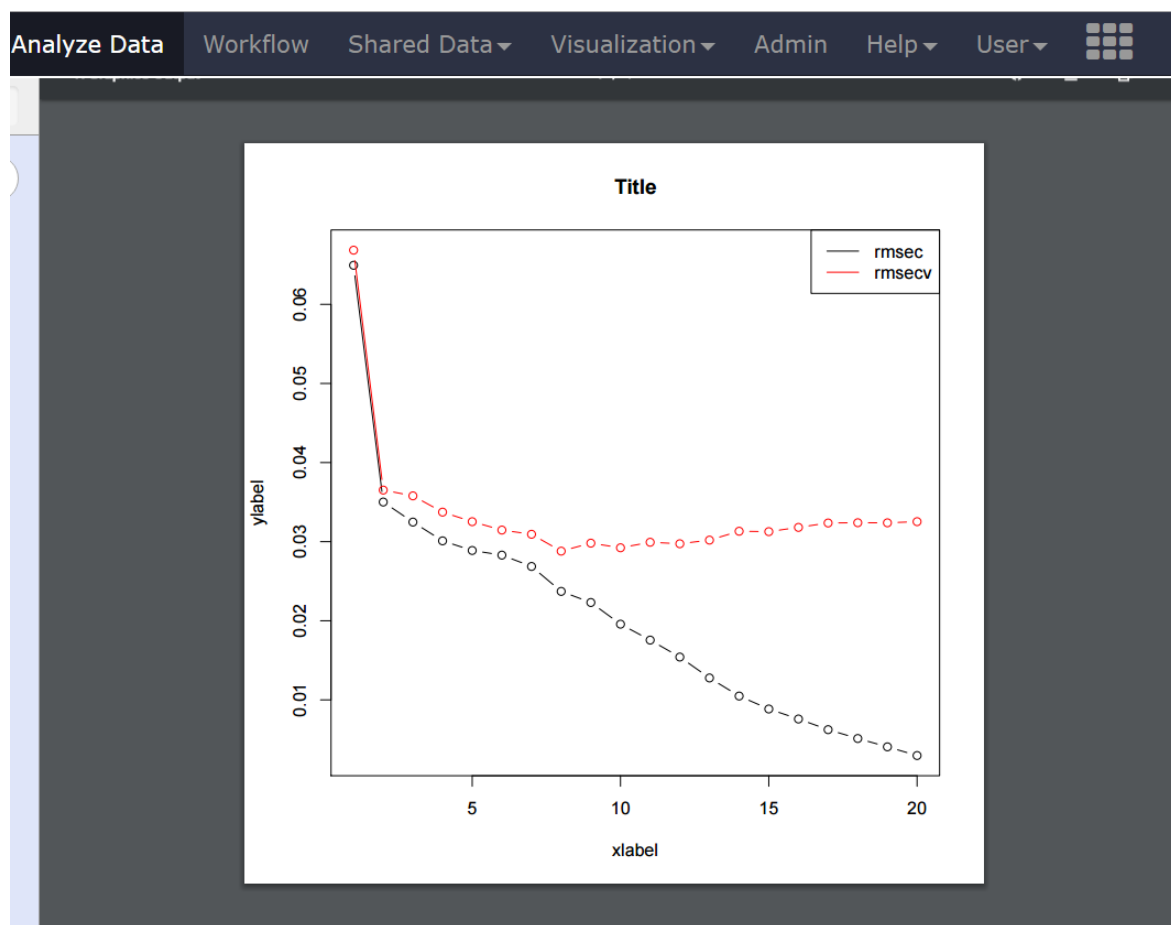
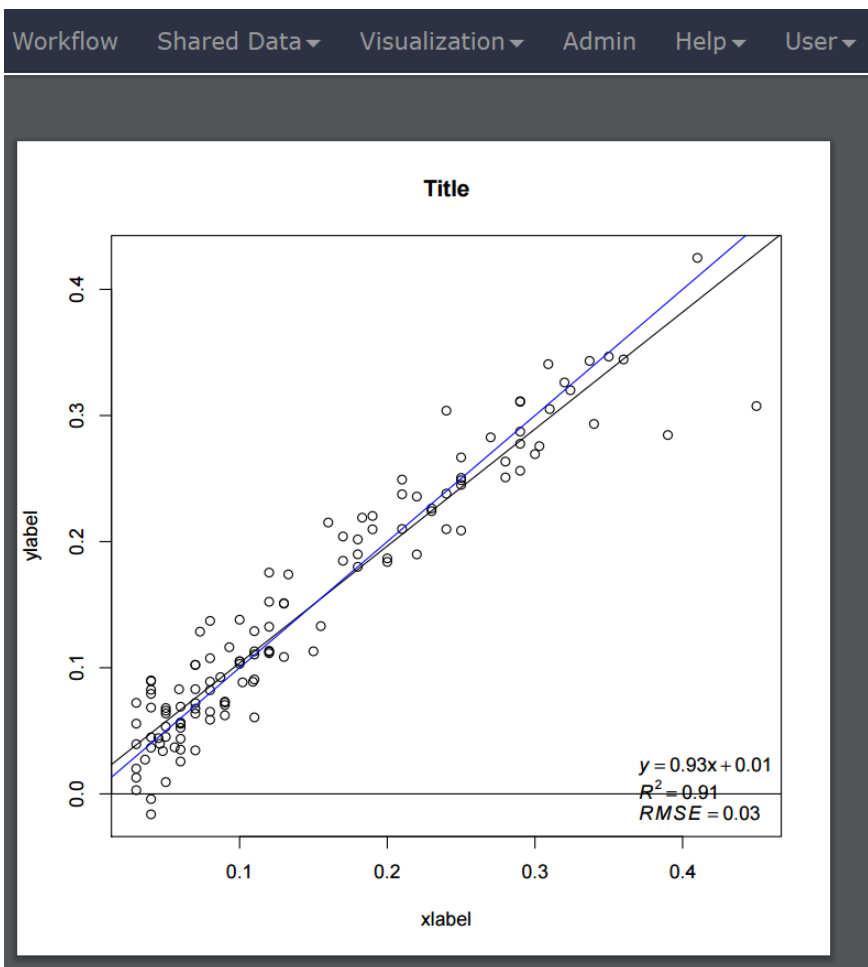
65: NIPALS-PLS on Y (None (value not yet validated)):ypredcv

64: NIPALS-PLS on Y (None (value not yet validated)):b-coeffs

63: NIPALS-PLS on Y (None (value not yet validated)):rmsec-secv







ChemFlow: fonctionnal test workflow

Galaxy <https://chemserver.supagro.inra.fr/workflow/editor?id=a0d84b45643a2678> Using 49.5 MB

Tools

- search tools
- Inputs**
- Import Data**
- Convert data format**
- Utils**
- Statistics**
- Plot**
- Calibration/Validation**
- Pretreatment**
- Exploration**
- Regressions**
- Clustering**
- Discrimination**
- Variable selection**
- Orthogonal projection**
- Calibration transfert**
- Unmixing/ Curve Resolution**
- Multitable/Multiway Analysis**
- Workflows**

Workflow Canvas | imported: pretreat (imported from uploaded file)

```
graph LR; Input[Input dataset] --> MSC[Select X data MSC]; Input --> SNV[Select X data SNV]; Input --> SG[Select X data SG]; Input --> SC[Select X data SC]; Input --> DT[Select X data DT]; MSC --> MSC_Plot[Spectra plot Spectra 1 > Dataset mat_plot pdf]; SNV --> SNV_Plot[Spectra plot Spectra 1 > Dataset mat_plot pdf]; SG --> SG_Plot[Spectra plot Spectra 1 > Dataset mat_plot pdf]; SC --> SC_Plot[Spectra plot Spectra 1 > Dataset mat_plot pdf]; DT --> DT_Plot[Spectra plot Spectra 1 > Dataset mat_plot pdf];
```

Details

SG Savitsky-Golay (Galaxy Version 0.0.1)

Select X data
Data input 'Xdata' (tabular)
Dataset (n x p) containing the n spectra of p variables.

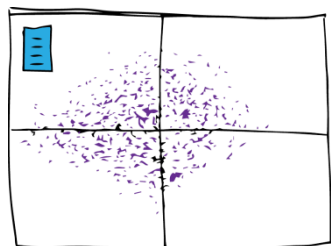
Derivative order
2

Size of the window
9

Degree of the polynomial
3

Annotation / Notes
Add an annotation or note for this step. It will be shown with the workflow.

Email notification
Yes No



–Test and validation tools (Planemo...)

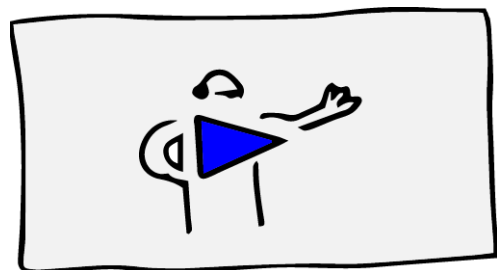


–Interactive graphics (R-Shiny)



–Connection with specific databases



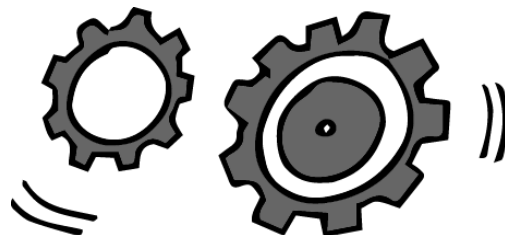


CheMOOCS

-MOOC, Massive Open
Online Course



Free and
available for
everybody



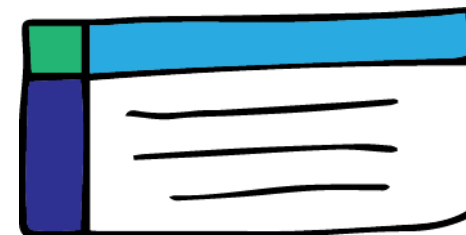
ChemFlow

- tool



ChemFlow

Using a
collaborative
conception



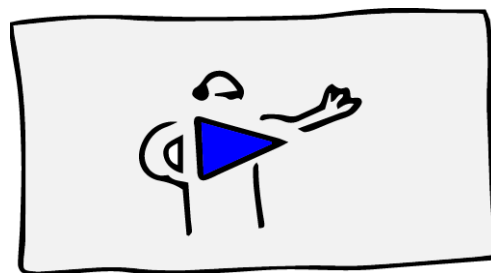
ChemData

- database



Data sharing

1. Chemometrics
2. The ChemFlow Tool
3. **The MOOC**



The MOOC: using FUN, a french course platform

FUN - Se former en | x

https://www.fun-mooc.fr

FUN Se former en liberté

Rechercher un cours Inscription Connexion

FUN : L'excellence de l'enseignement supérieur pour des cours en ligne, gratuits et ouverts à tous

FUN

Les cours à la une

Comprendre l'économie collaborative
NOUVEAU COURS
Comprendre l'économie collaborative
Institut Mines-Télécom
Débute 20 sep 2016 En savoir plus

Chirurgie de l'obésité
NOUVEAU COURS
Chirurgie de l'obésité
Université Fédérale Toulouse Midi-Pyrénées
Débute 15 sep 2016 En savoir plus

Initiation aux applications dynamiques
NOUVEAU COURS
Initiation aux applications dynamiques
Université Fédérale Toulouse Midi-Pyrénées
Débute 01 oct 2016 En savoir plus

CheMOOCs
PRINCIPES ET OUTILS DE LA CHIMISMETRIE POUR TOUS
NOUVEAU COURS
CheMOocs
Agreenium
Débute 12 sep 2016 En savoir plus

Courses in french language

FUN, France Université numérique
France Digital University

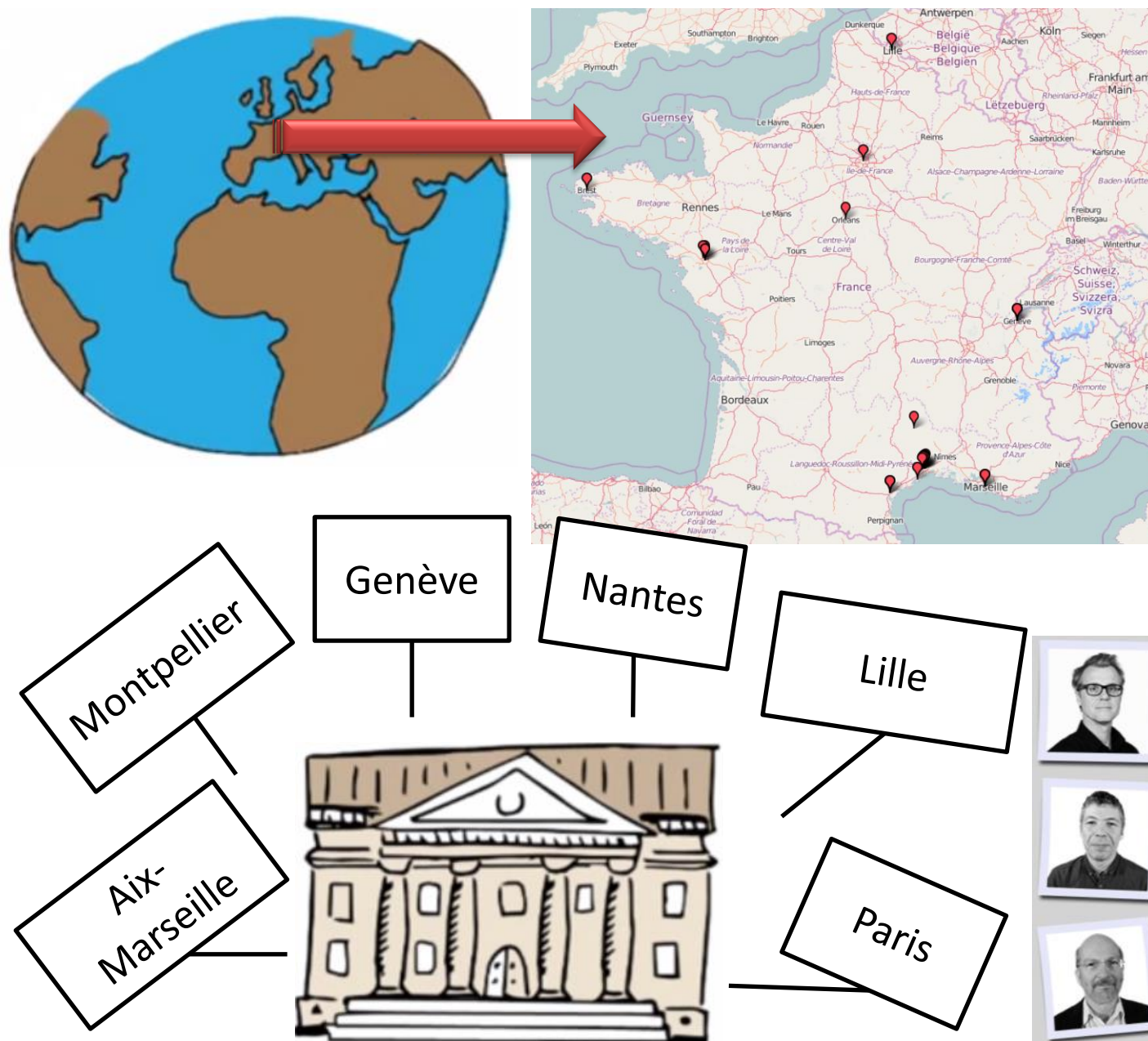
The MOOC: who?

IT team



24 collaborators

- Academic
- Industrial



–Targeted audience

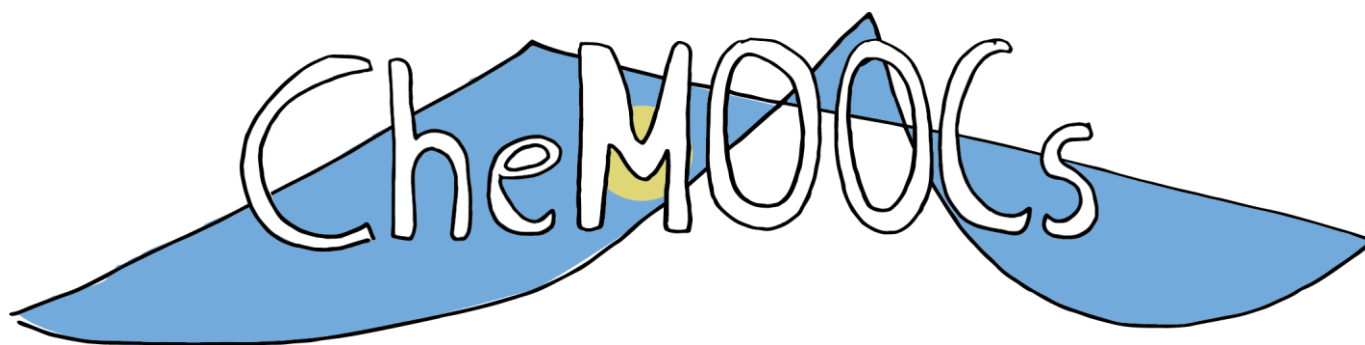
- Students: initial learning
- Professionals (industrial, teacher, researcher): training

– Aims:

- New competence
- Competence improvement



- Launch on FUN in French language: September 12th 2016
- Registration from June 15th to www.FUN-MOOC.fr
- MOOC from September 12th to November 10th



THANKS



data_frame



The EIC, *Equipe d'Informatique Collective*, team in Montpellier at Supagro/INRA

Illustration et Logo (sous licence CC BY SA) par Fanny Monod (fanzyl.net) et Marc Lanssens de SupAgro Montpellier



This project was funded under the reference ID-1401-005 from the program "Investissements d'avenir" (Labex-Agro : ANR610-LABEX-0001-01)

