



HAL
open science

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

ChemFlow, chemometrics using Galaxy

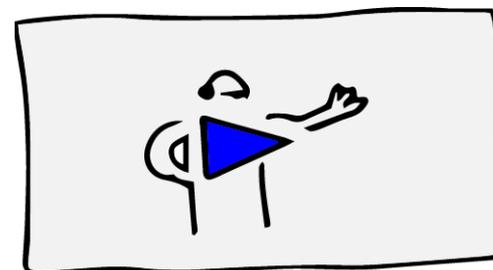
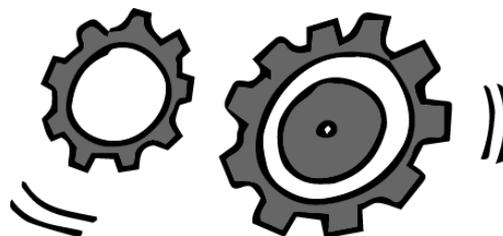
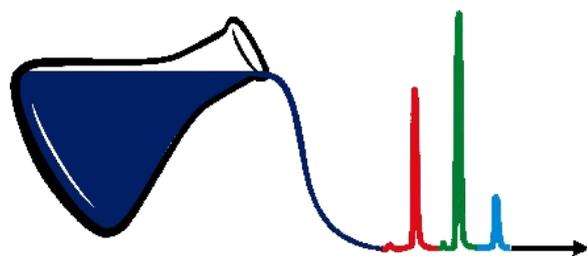
V. Rossard*, J.C. Boulet*, F. Gogé**, E. Latrille*, J.M. Roger**

*INRA, Institut National de la Recherche Agronomique

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

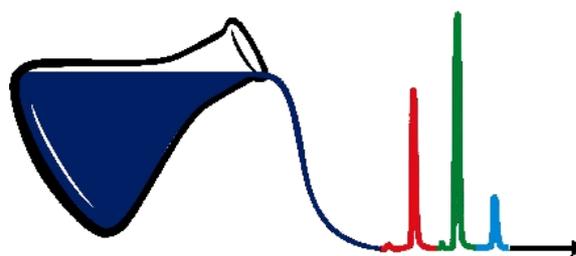


1. Chemometrics
2. The ChemFlow Tool
3. The MOOC



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

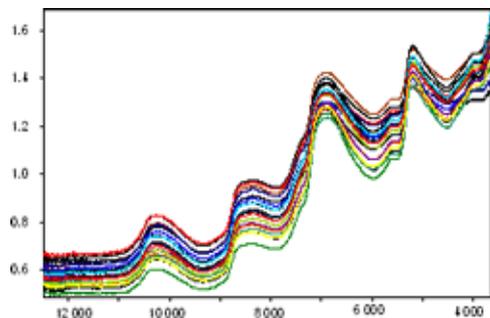
Measurements

Chemometrics

State

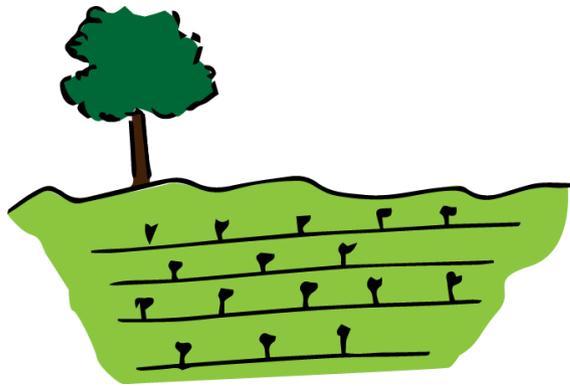
- Multivariate model
- Mathematics
- Statistics

- Gluten content of flour
- Year of production



Infrared spectra





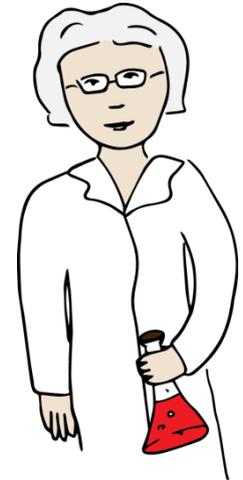
Application domain



Product



Sample



Operator



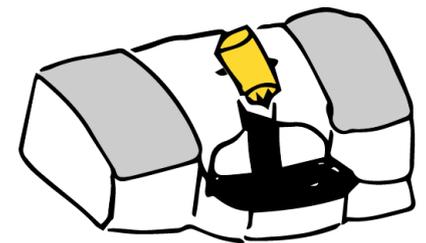
Chemometrics



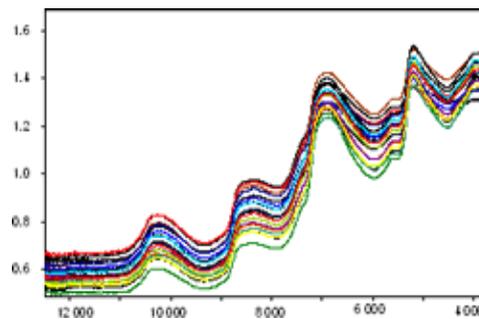
Measurements



Spectrometer



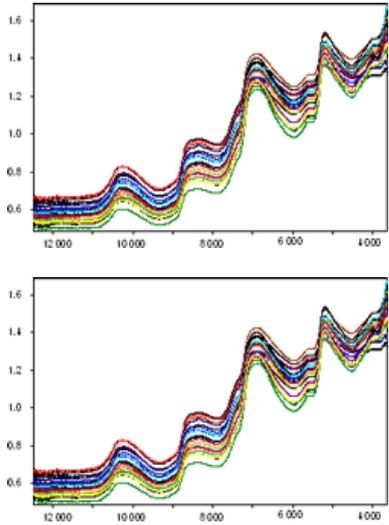
- Gluten content of flour
- Year of production



Infrared spectra



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

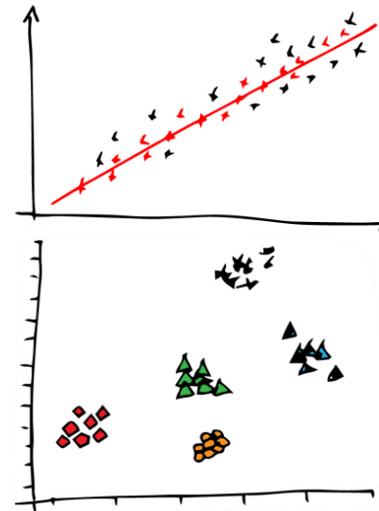


Infrared spectra

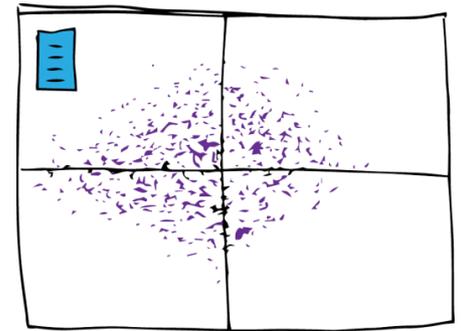


To

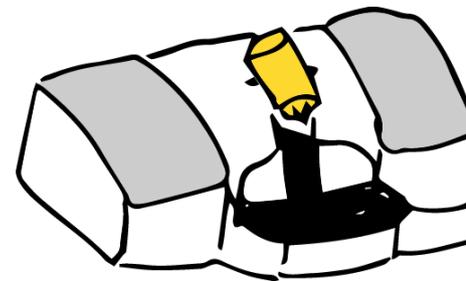
```
LOAD('MY DATA.CSV');  
PLOT(X);  
X=DETREND(X)  
RESPLS=PLS(X,Y,10,20,1)  
PLOT(RESPLS...
```



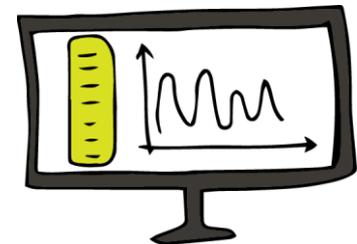
Predict biochemical content using a model



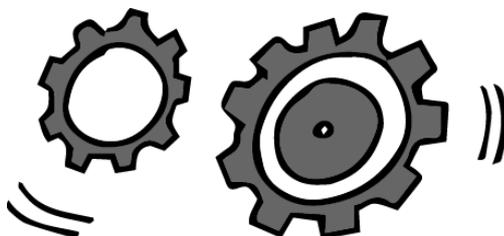
Laboratory



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
langua

Develo

Difficul
chains



Galaxy

PROJECT

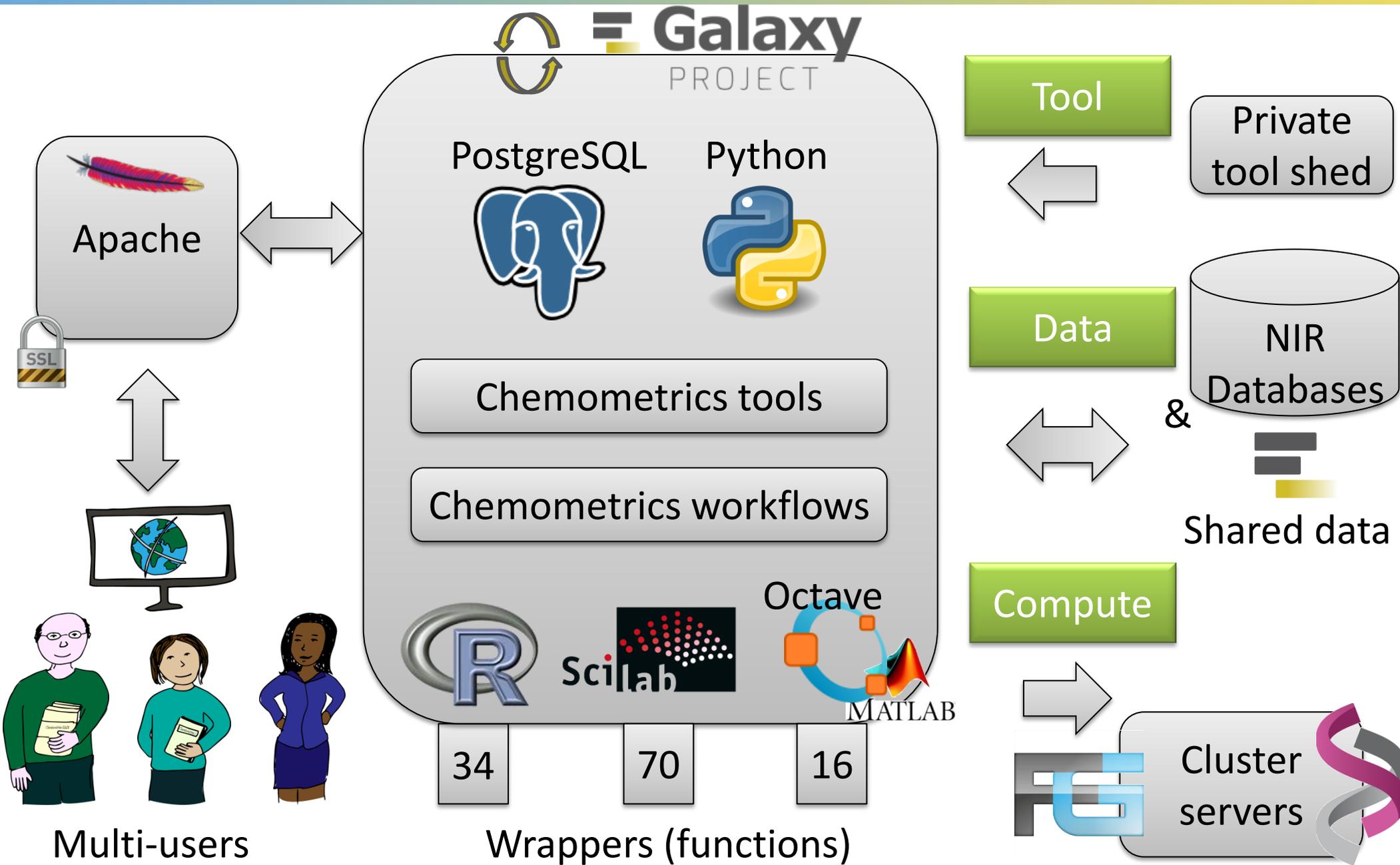
by
tools

traceability, workflow

Application to spectral analysis

- 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.

Chain tools into workflows

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

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

Galaxy 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 programing 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**

Galaxy <https://chemserver.supagro.inra.fr>

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 35.4 MB

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	-0.048245	-0.048092	-0.047762	-0.047316	-0.046829	-0.046383	-0.045937	-0.045491	-0.045045	-0.044599
AP02	-0.039741	-0.039554	-0.039367	-0.039180	-0.038993	-0.038806	-0.038619	-0.038432	-0.038245	-0.038058	-0.037871	-0.037684
AP03	-0.044596	-0.044463	-0.044330	-0.044197	-0.044064	-0.043931	-0.043798	-0.043665	-0.043532	-0.043399	-0.043266	-0.043133
AP04	-0.046453	-0.046489	-0.046525	-0.046561	-0.046597	-0.046633	-0.046669	-0.046705	-0.046741	-0.046777	-0.046813	-0.046849
AP05	-0.04702	-0.046511	-0.046002	-0.045493	-0.044984	-0.044475	-0.043966	-0.043457	-0.042948	-0.042439	-0.041930	-0.041421
AP06	-0.049078	-0.049004	-0.048930	-0.048856	-0.048782	-0.048708	-0.048634	-0.048560	-0.048486	-0.048412	-0.048338	-0.048264
AP07	-0.049906	-0.049996	-0.050086	-0.050176	-0.050266	-0.050356	-0.050446	-0.050536	-0.050626	-0.050716	-0.050806	-0.050896
AP08	-0.050282	-0.050152	-0.050022	-0.049892	-0.049762	-0.049632	-0.049502	-0.049372	-0.049242	-0.049112	-0.048982	-0.048852
AP09	-0.048378	-0.048315	-0.048252	-0.048189	-0.048126	-0.048063	-0.047999	-0.047936	-0.047873	-0.047810	-0.047747	-0.047684
AP10	-0.046083	-0.045663	-0.045243	-0.044823	-0.044403	-0.043983	-0.043563	-0.043143	-0.042723	-0.042303	-0.041883	-0.041463
AP11	-0.050291	-0.050357	-0.050423	-0.050489	-0.050555	-0.050621	-0.050687	-0.050753	-0.050819	-0.050885	-0.050951	-0.051017
AP12	-0.04934	-0.049129	-0.048918	-0.048707	-0.048496	-0.048285	-0.048074	-0.047863	-0.047652	-0.047441	-0.047230	-0.047019
AP13	-0.048376	-0.048506	-0.048636	-0.048766	-0.048896	-0.049026	-0.049156	-0.049286	-0.049416	-0.049546	-0.049676	-0.049806
AP14	-0.046614	-0.046385	-0.046156	-0.045927	-0.045698	-0.045469	-0.045240	-0.045011	-0.044782	-0.044553	-0.044324	-0.044095
AP15	-0.047614	-0.047479	-0.047344	-0.047209	-0.047074	-0.046939	-0.046804	-0.046669	-0.046534	-0.046399	-0.046264	-0.046129
AP16	-0.050376	-0.050324	-0.050272	-0.050220	-0.050168	-0.050116	-0.050064	-0.050012	-0.049960	-0.049908	-0.049856	-0.049804
AP17	-0.052726	-0.052701	-0.052676	-0.052651	-0.052626	-0.052601	-0.052576	-0.052551	-0.052526	-0.052501	-0.052476	-0.052451
AP18	-0.050036	-0.049736	-0.049436	-0.049136	-0.048836	-0.048536	-0.048236	-0.047936	-0.047636	-0.047336	-0.047036	-0.046736
AP19	-0.053388	-0.053846	-0.054304	-0.054762	-0.055220	-0.055678	-0.056136	-0.056594	-0.057052	-0.057510	-0.057968	-0.058426
AP21	-0.046617	-0.046224	-0.045831	-0.045438	-0.045045	-0.044652	-0.044259	-0.043866	-0.043473	-0.043080	-0.042687	-0.042294
AP22	-0.050909	-0.050995	-0.051081	-0.051167	-0.051253	-0.051339	-0.051425	-0.051511	-0.051597	-0.051683	-0.051769	-0.051855
AP23	-0.052071	-0.05187	-0.05167	-0.05147	-0.05127	-0.05107	-0.05087	-0.05067	-0.05047	-0.05027	-0.05007	-0.04987
AP24	-0.05252	-0.052267	-0.052014	-0.051761	-0.051508	-0.051255	-0.051002	-0.050749	-0.050496	-0.050243	-0.050090	-0.049837
		-0.047917	-0.047714	-0.047511	-0.047308	-0.047105	-0.046902	-0.046699	-0.046496	-0.046293	-0.046090	-0.045887

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

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User

Tools [Analysis home view](#)

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

Pretreatment

Exploration

Regressions

Clustering

Spectra Plot

Abs

wavelength

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

ChemFlow: chemometrics tools interface

The screenshot displays the Galaxy web interface for the ChemFlow tool. The browser address bar shows <https://chemserver.supagro.inra.fr>. The Galaxy navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The 'Tools' panel on the left lists various analysis categories, with 'PLS Regression Partial Least Square' circled in red. The main configuration area for the tool is set to '58: Xnir' as the dataset and '60: y_triglycerides' as the y data. The 'Algorithm choice' is 'NIPALS' and the 'cross-validation type' is 'k-bloc cross-validation'. The 'Number of blocs for cross-validation' is set to 4, and the 'Number of latent variables' is 20. The 'Centering option' is 'yes'. The 'compute outliers statistics' checkbox is checked. The 'History' panel on the right shows a list of previous runs, with the most recent ones (63-68) highlighted in green and red, indicating their status.

Tools

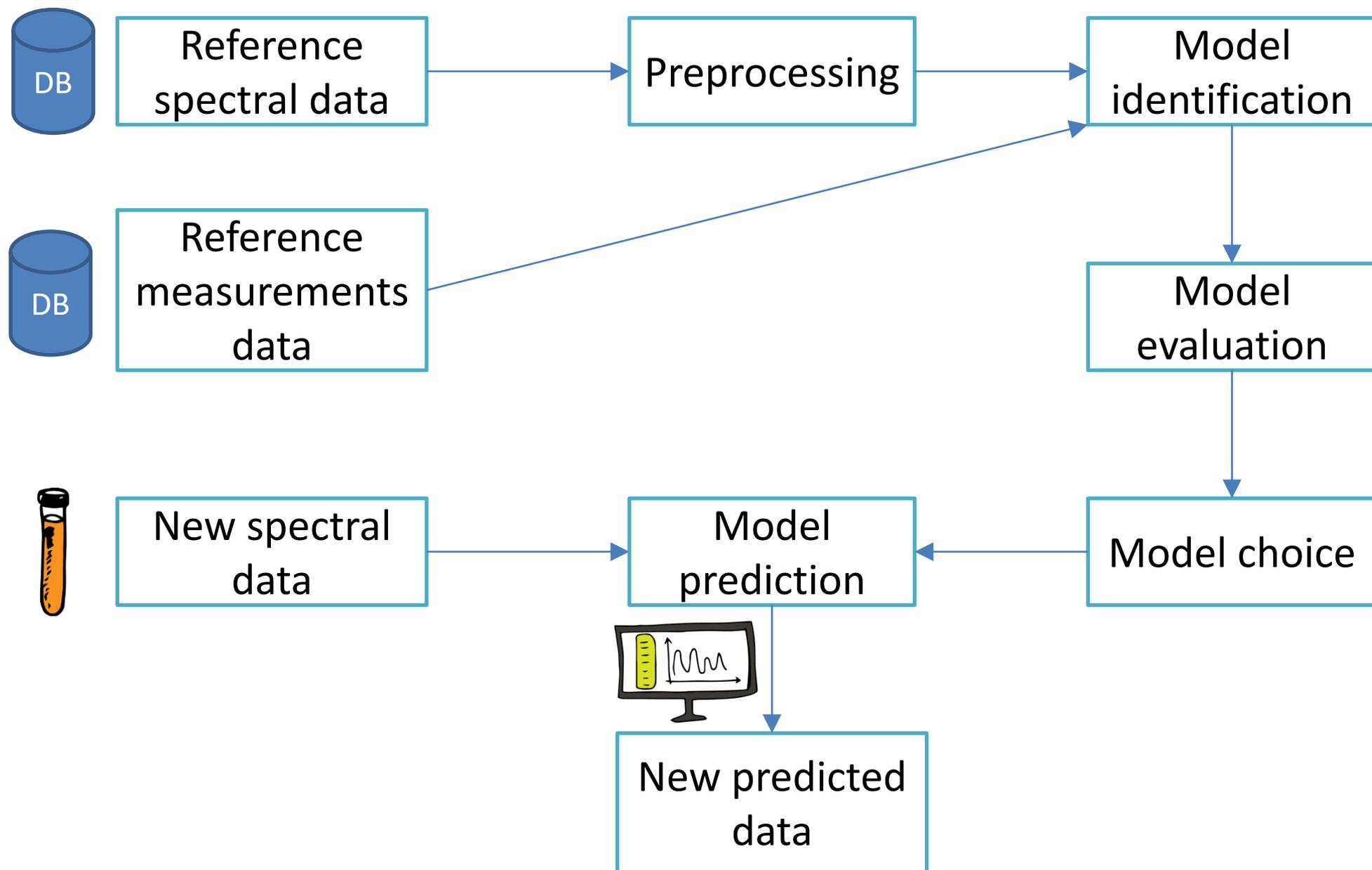
- Calibration/Validation**
- Pretreatment**
- Exploration**
- Regressions**
 - PLS Regression Partial Least Square**
 - PCR Regression Principal component Regression
 - MLR Regression Multiple Linear Regression
 - SLR Simple Linear Regression
 - Applies a regression model to a new set of spectra
- Clustering**
- Discrimination**
- Variable selection**
- Orthogonal projection**
- Calibration transfert**
- Unmixing/ Curve Resolution**
- Multitable/Multiway Analysis**

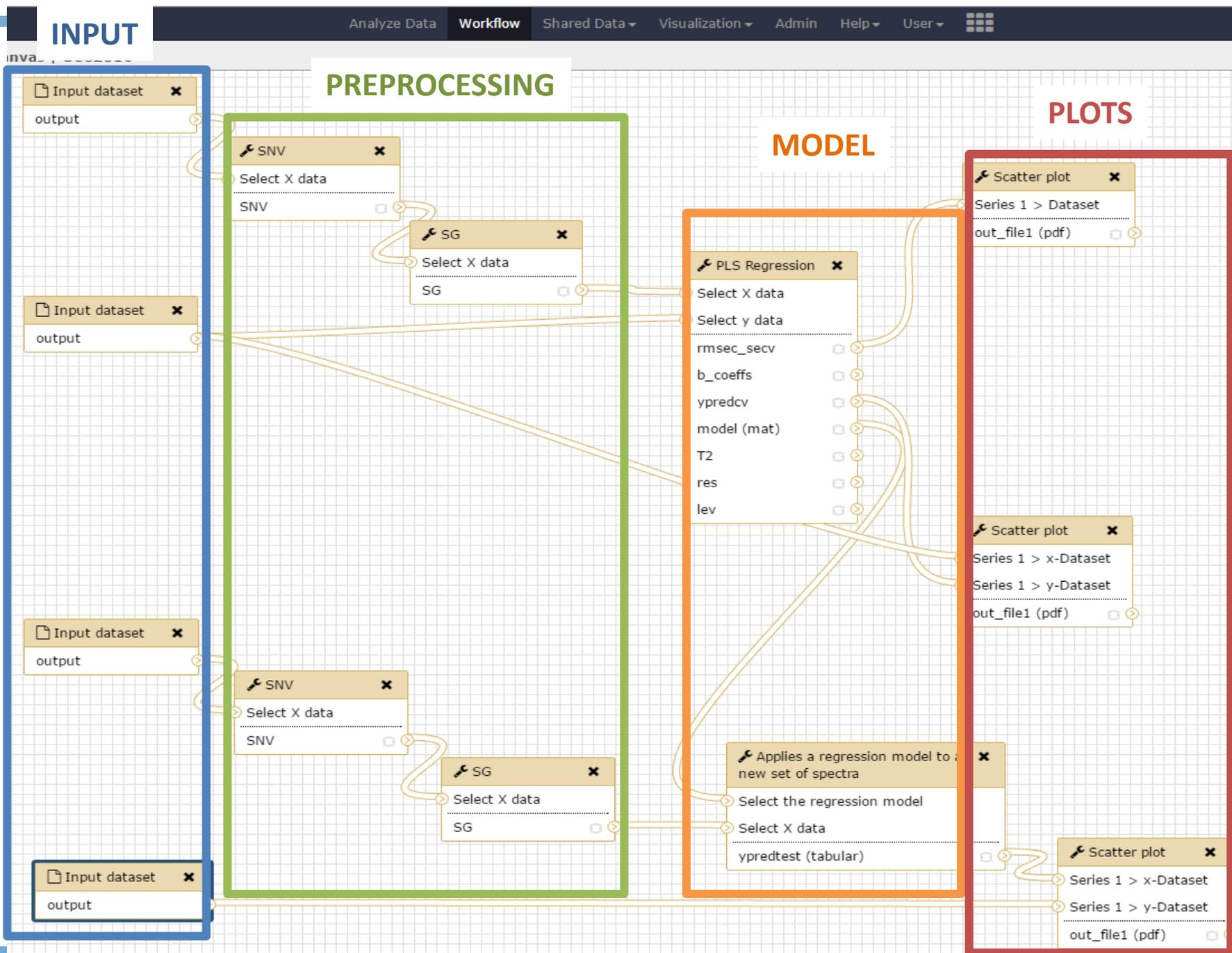
Configuration:

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

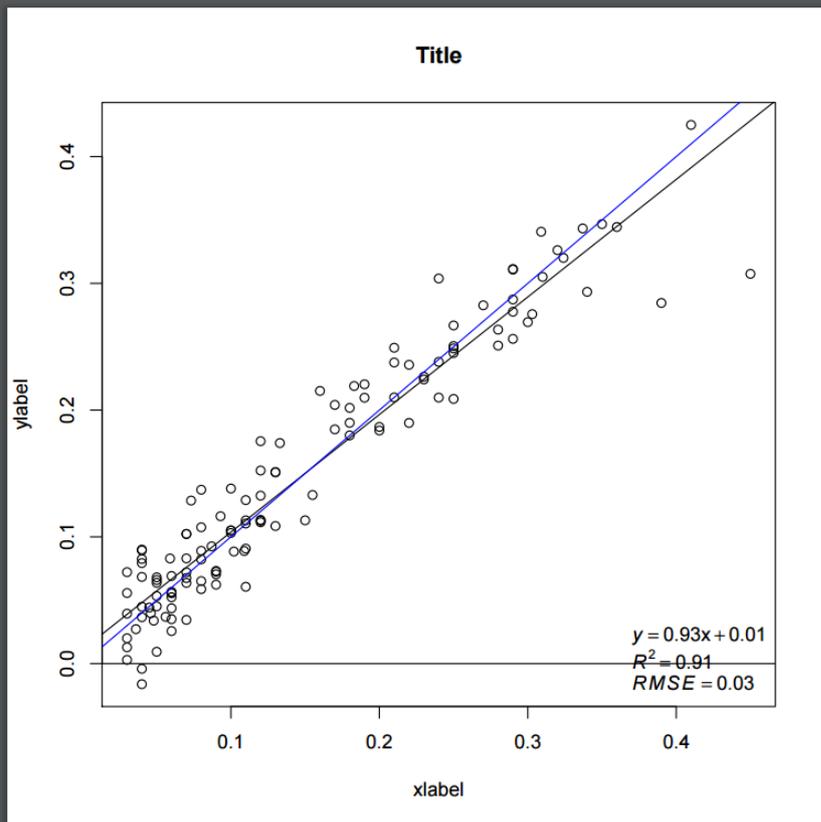
History:

- 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

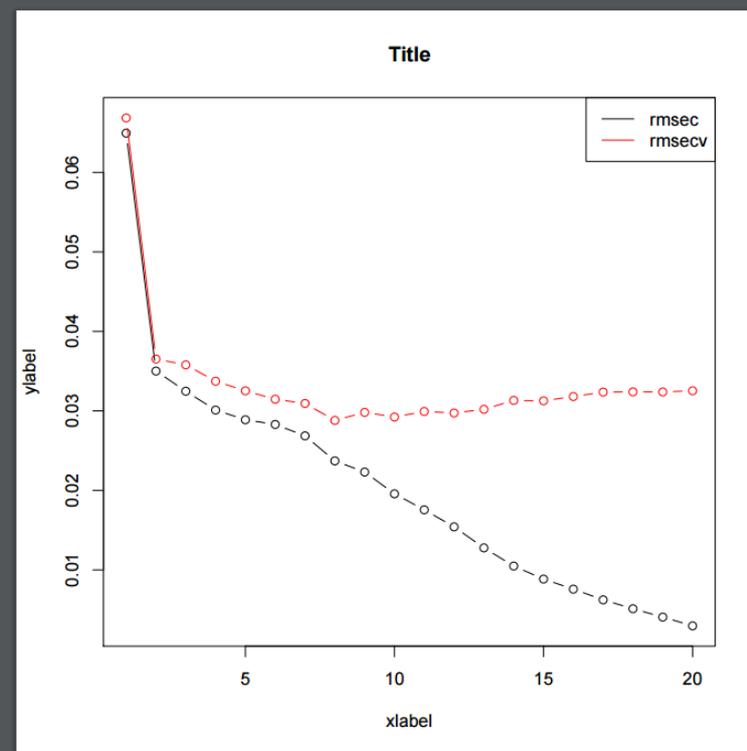




Workflow Shared Data Visualization Admin Help User



Analyze Data Workflow Shared Data Visualization Admin Help User



ChemFlow: fonctionnal test workflow

The screenshot displays the Galaxy workflow editor interface. The main canvas, titled "Workflow Canvas | imported: pretreat (imported from uploaded file)", shows a workflow starting with an "Input dataset" (output) that branches into five parallel paths. Each path consists of a "Select X data" tool followed by a "Spectra plot" tool. The "Spectra plot" tools are configured with "Spectra 1 > Dataset" and "mat_plot (pdf)". The central "SG" tool is highlighted with a blue border. On the right, the "Details" panel for the "SG Savitsky-Golay (Galaxy) Version 0.0.1" tool is visible, showing configuration options: "Derivative order" set to 2, "Size of the window" set to 9, and "Degree of the polynomial" set to 3. The "Annotation / Notes" section is empty, and the "Email notification" section is partially visible at the bottom.

Galaxy
https://chemserver.supagro.inra.fr/workflow/editor?id=a0d84b45643a2678

Analyze Data Workflow Shared Data Visualization Admin Help User 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)

Input dataset x
output

MSC x
Select X data
MSC

SNV x
Select X data
SNV

SG x
Select X data
SG

SC x
Select X data
CT

DT x
Select X data
DT

Spectra plot x
Spectra 1 > Dataset
mat_plot (pdf)

Spectra plot x
Spectra 1 > Dataset
mat_plot (pdf)

Spectra plot x
Spectra 1 > Dataset
mat_plot (pdf)

Spectra plot x
Spectra 1 > Dataset
mat_plot (pdf)

Spectra plot x
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.

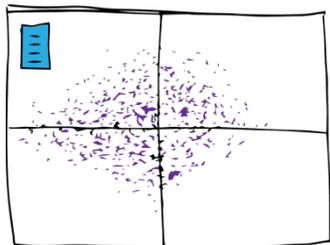
Derivative order
2

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.

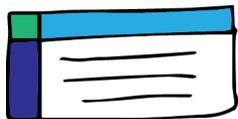
Email notification



–Test and validation tools (Planemo...)

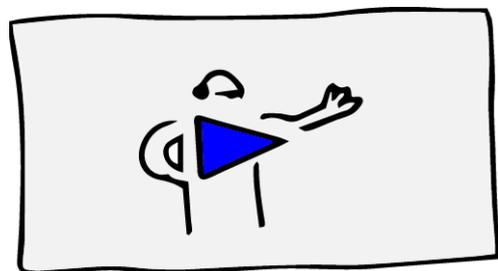


–Interactive graphics (R-Shiny)



–Connection with specific databases



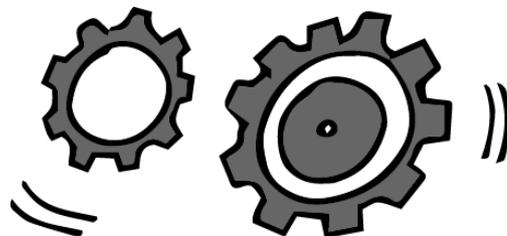


CheMOOCS

-MOOC, Massive Open
Online Course



Free and
available for
everybody



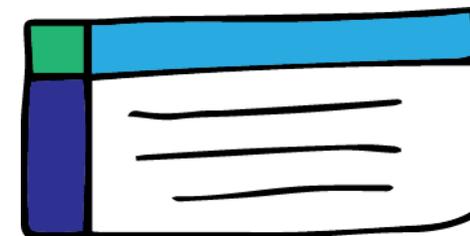
ChemFlow

- tool



ChemFlow

Using a
collaborative
conception



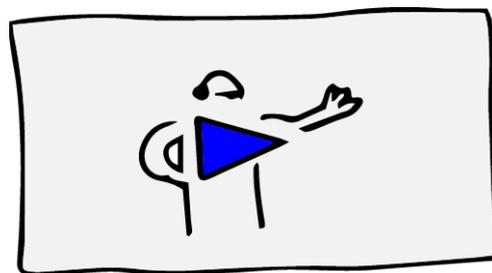
ChemData

- database

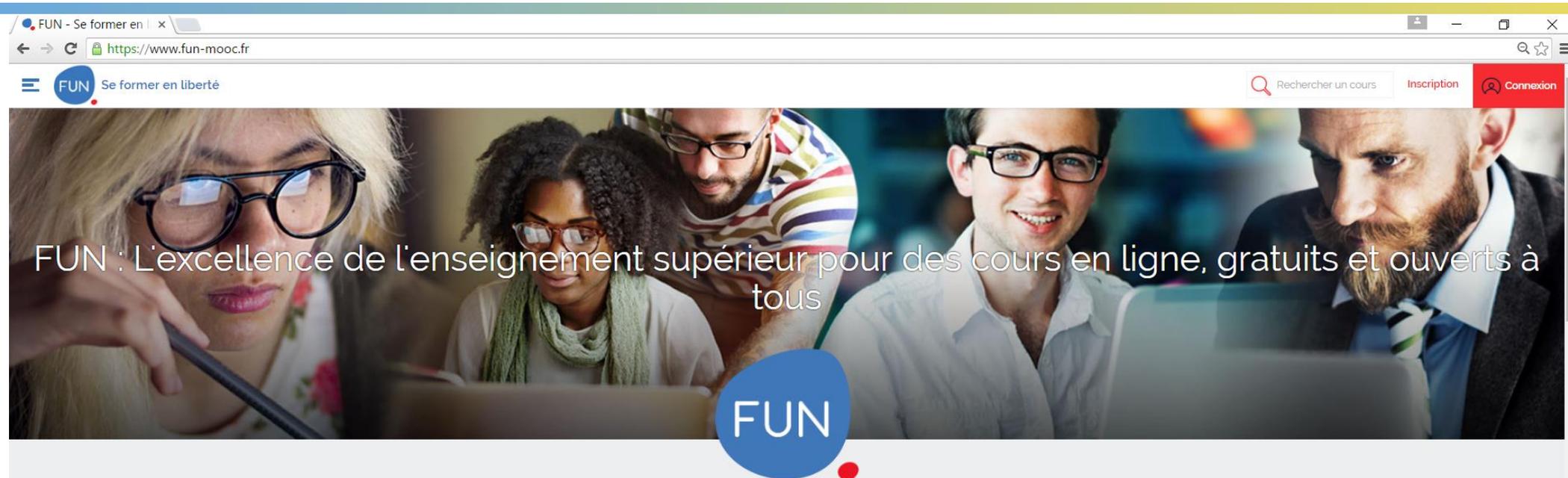


Data sharing

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



The MOOC: using FUN, a french course platform



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

NOUVEAU COURS

Chirurgie de l'obésité

Université Fédérale Toulouse Midi-Pyrénées

Débute 15 sep 2016

En savoir plus

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

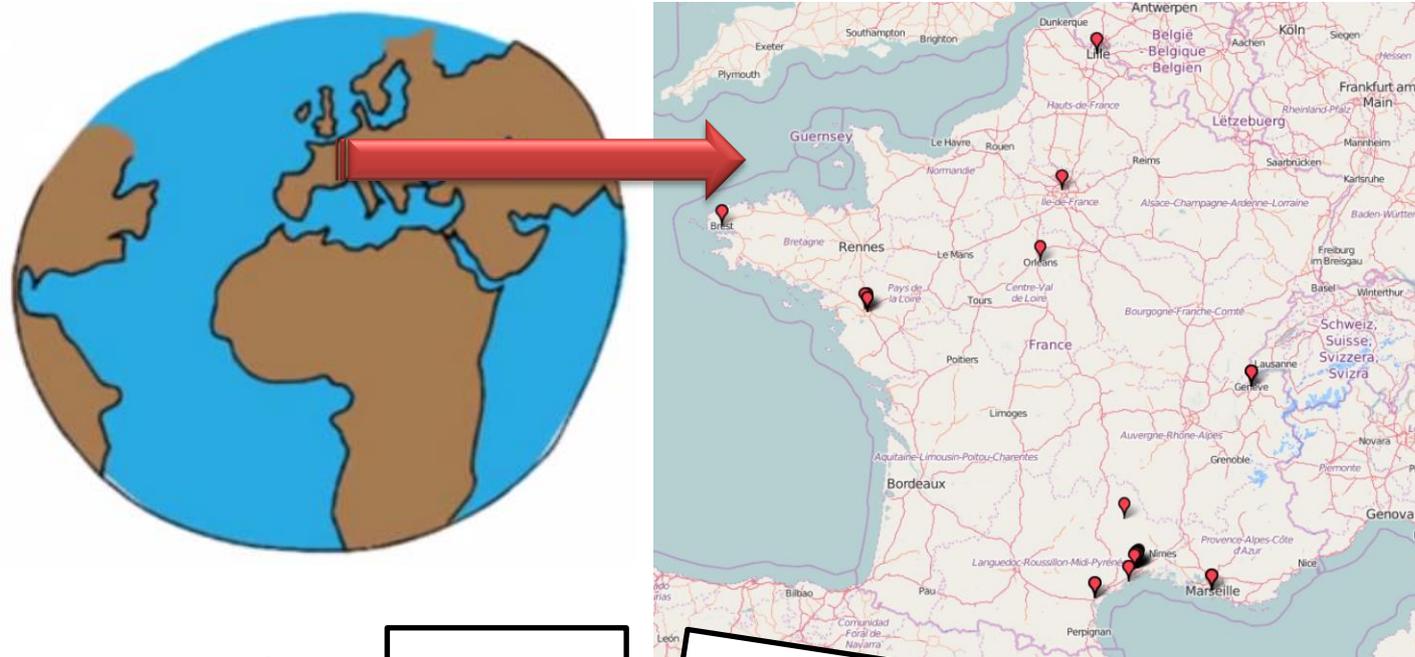
Agreenium

Débute 12 sep 2016

En savoir plus

Courses in french language

FUN, France Université numérique
France Digital University

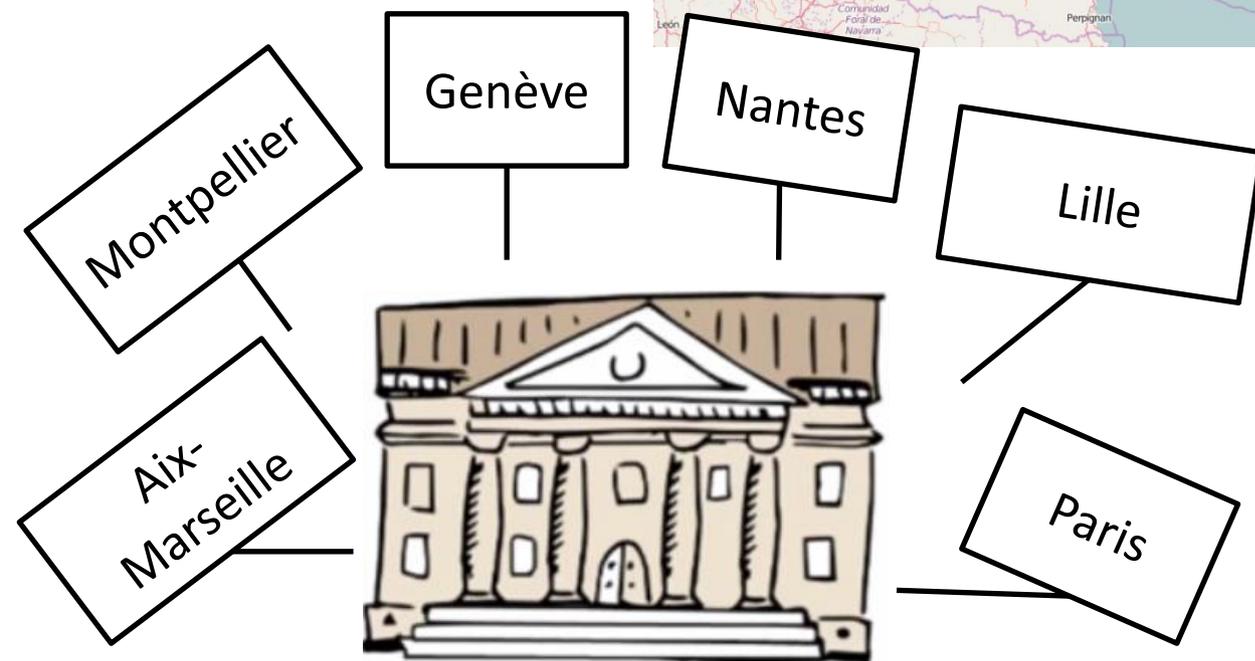


IT team



24 collaborators

- Academic
- Industrial



–Targeted audience

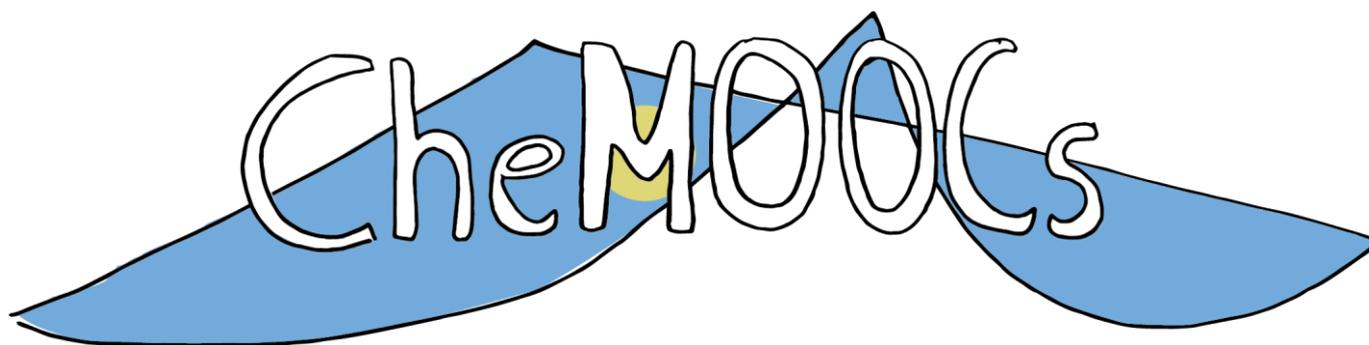
- Students: initial learning
- Professionals (industrial, teacher, researcher): training

– 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

The logo for CheMOOCs features the text 'CheMOOCs' in a white, rounded, sans-serif font. The letters are set against a blue, stylized background that resembles a mountain range or a series of overlapping peaks. The 'M' is notably larger and more prominent than the other letters.

26
registrations
per day

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)

