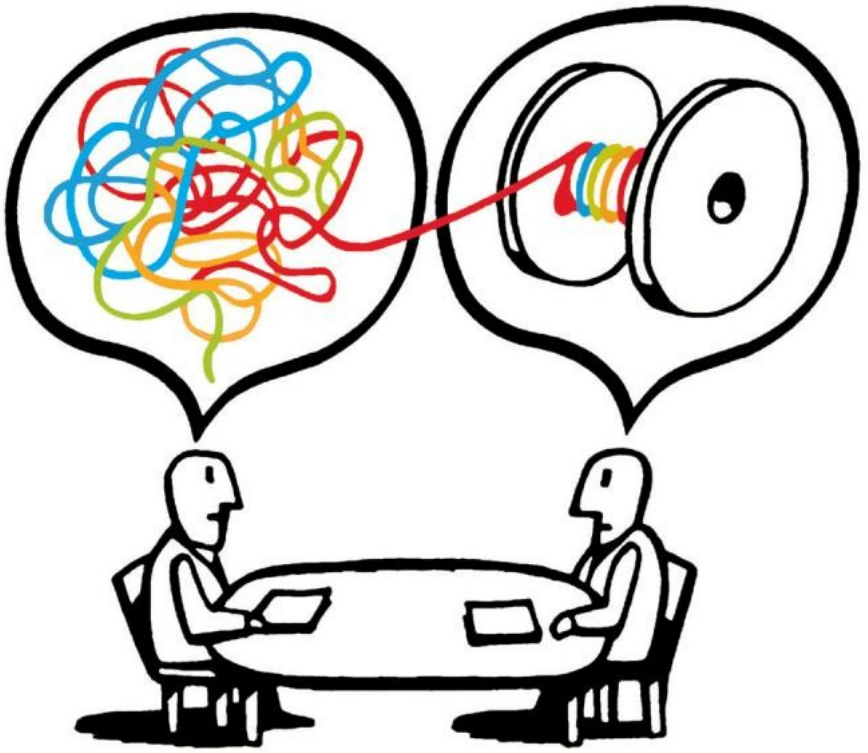
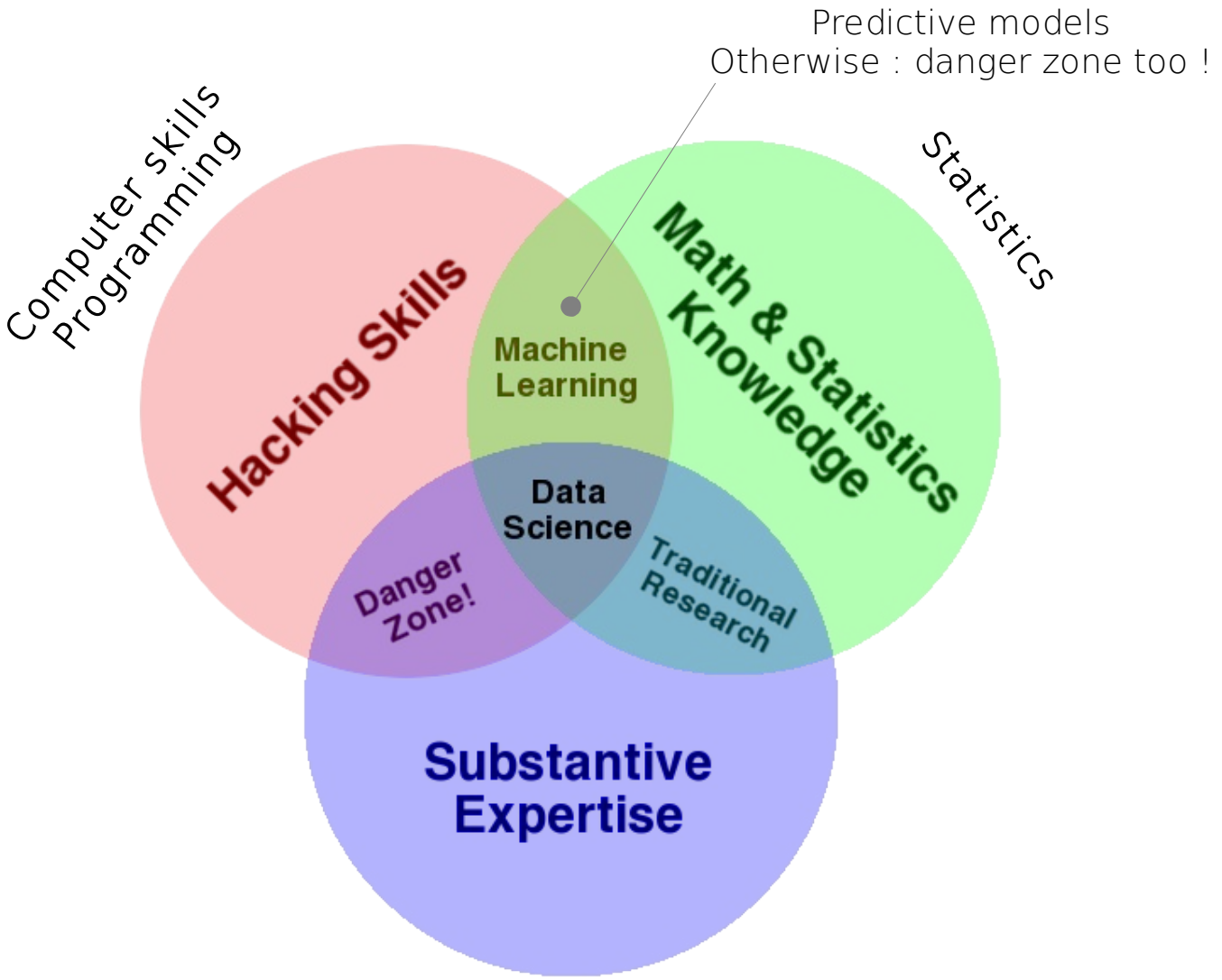
Gilles San Martin
Centre Wallon de Recherches Agronomiques

Data Science

Domaine interdisciplinaire visant à analyser des quantités massives de données brutes pour en extraire des nouvelles connaissances





http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science : pour qui ?

1 - Chercheur : de plus en plus

2 - Biologiste/agronome "de terrain"
(ONG, administration, bureau d'étude)
plus manipulation de données que statistiques

Statistiques : tendances de populations, habitat d'une espèce,...

Récolte de données automatique
enregistreurs météo, sonores (chauves-souris) - metabarcoding - animaux avec balises GPS,...

GIS : carte de qualité habitat / risque

Télédétection : repérage plantes invasives, état sanitaire des forêts, comptages mammifères, anciens lits de rivière drainée, aires de faulde, ...

Automatisation du rapportage
ex : évaluations annuelles : qualité des eaux, comptages oiseaux,...

Data Science : pour qui ?

3 - Étudiant(e) en biologie/écologie

Comment crée-t-on de nouvelles connaissances ?
--> lecture critique de la littérature
--> mise en pratique (TFE)

4 - Citoyenne/journaliste

Grâce à l'Open Data

Data Science : pour qui ?

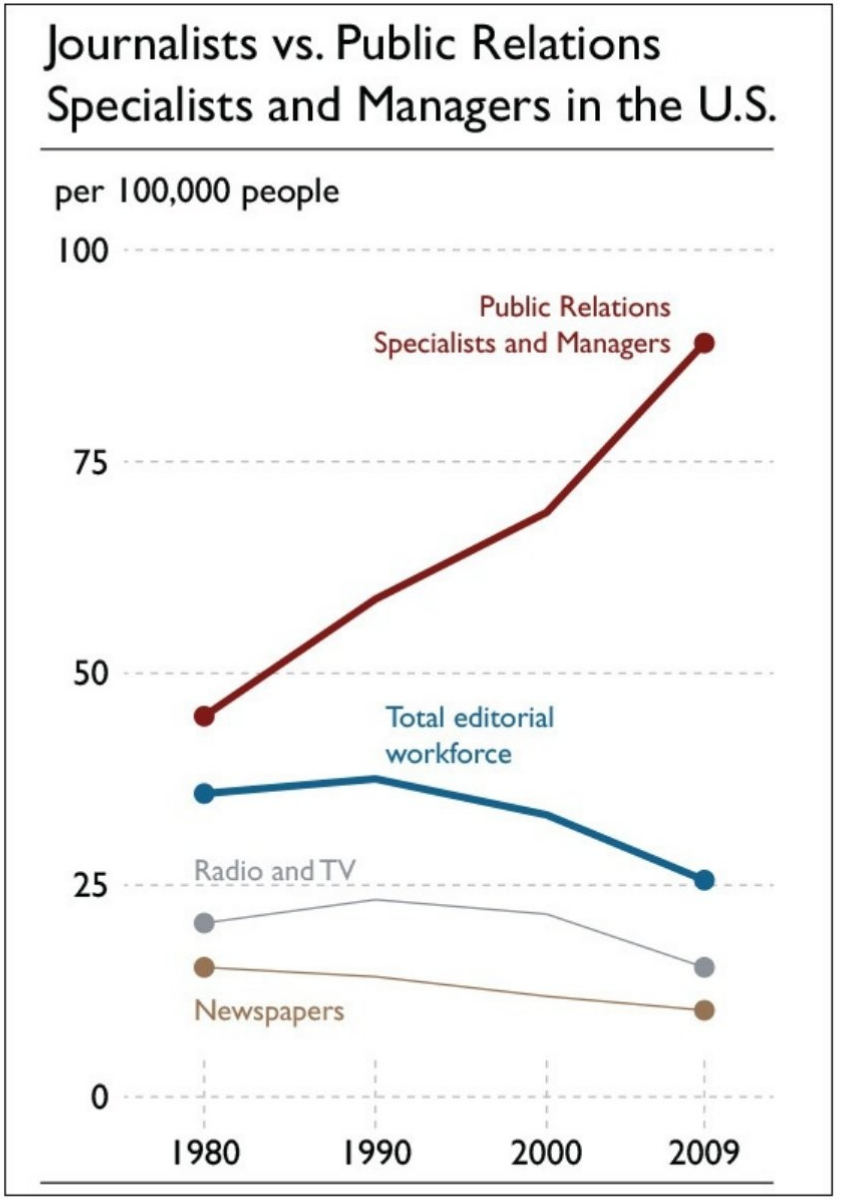
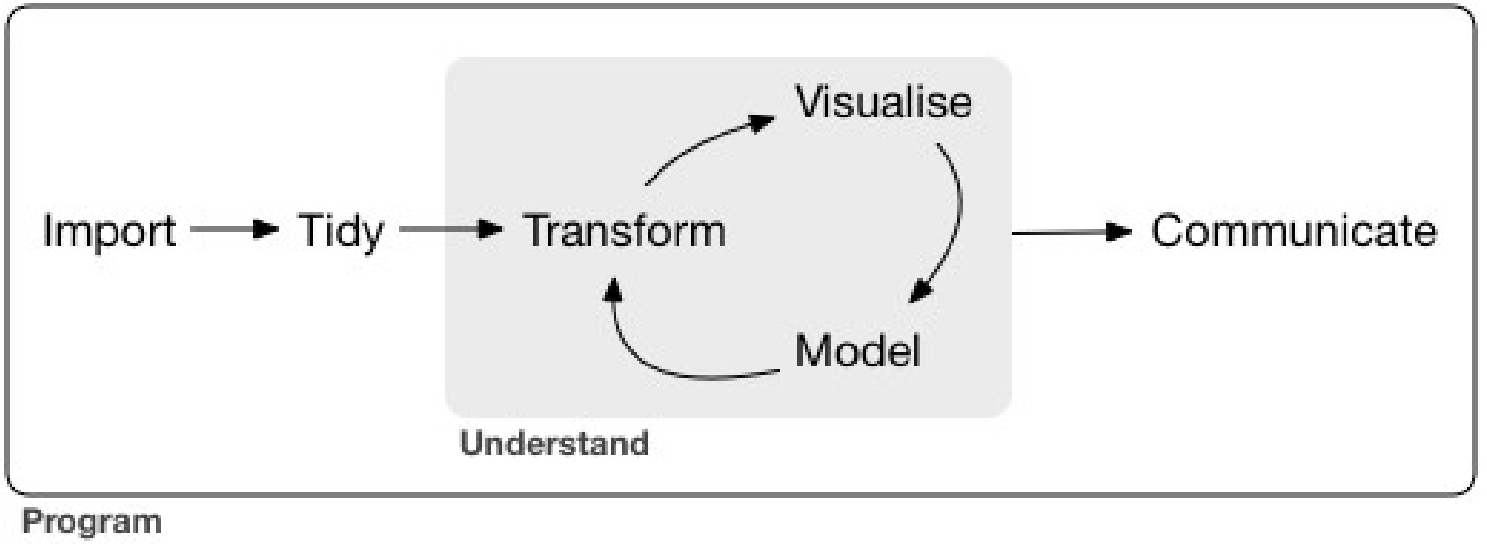
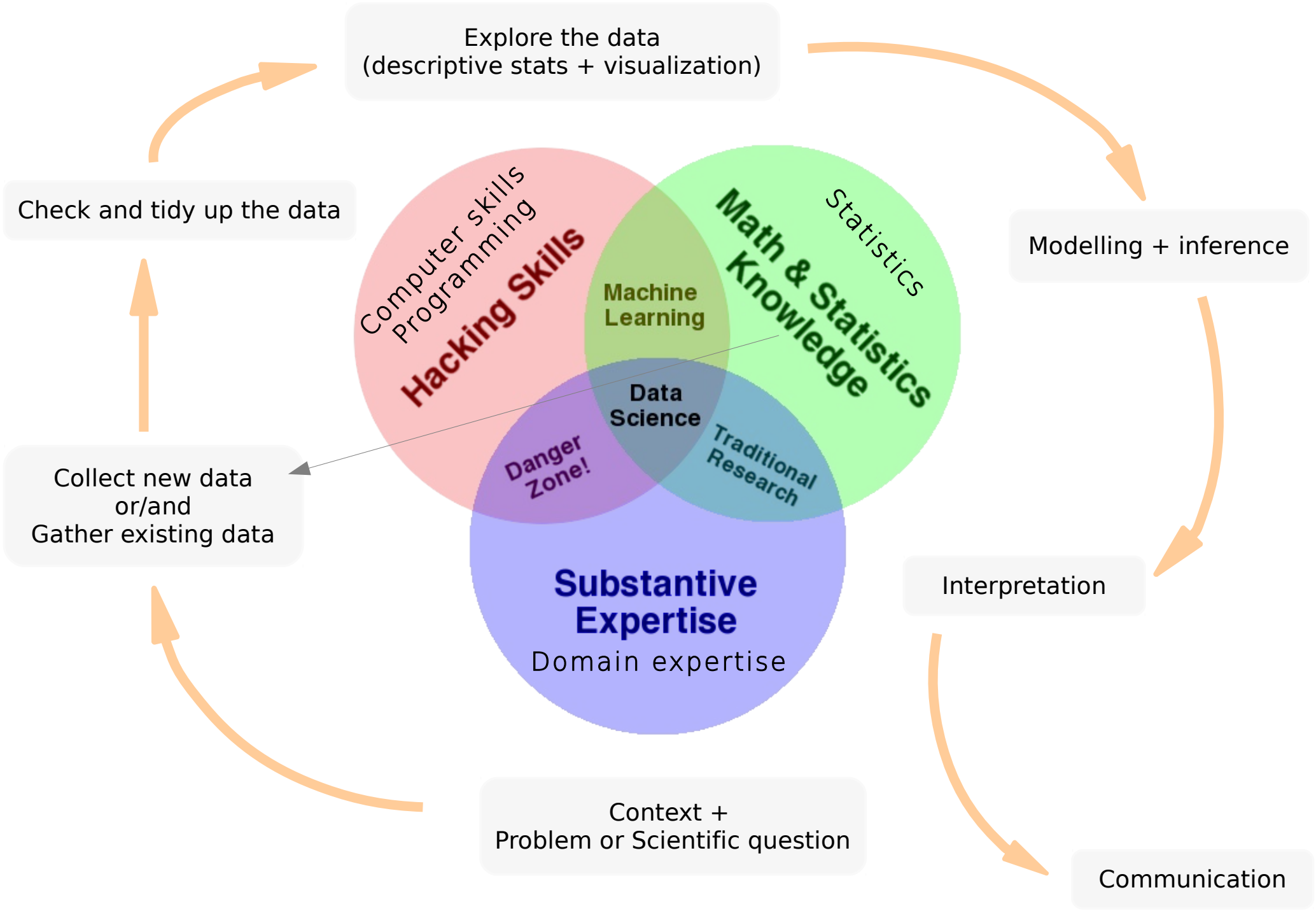


Figure I.10 The number of professionals working in public relations has expanded greatly in the past three decades, while the number of journalists has dropped. (Graph based on McChesney and Nichols, 2011.)

Data Science Process



Context + Problem at hand



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

1 - Contexte et question scientifique

Définir clairement :

le **contexte de l'étude**
problème à résoudre ou la **question scientifique**

--> conséquences majeures sur
1) la récolte des données
2) la gravité ou non de certains problèmes

+

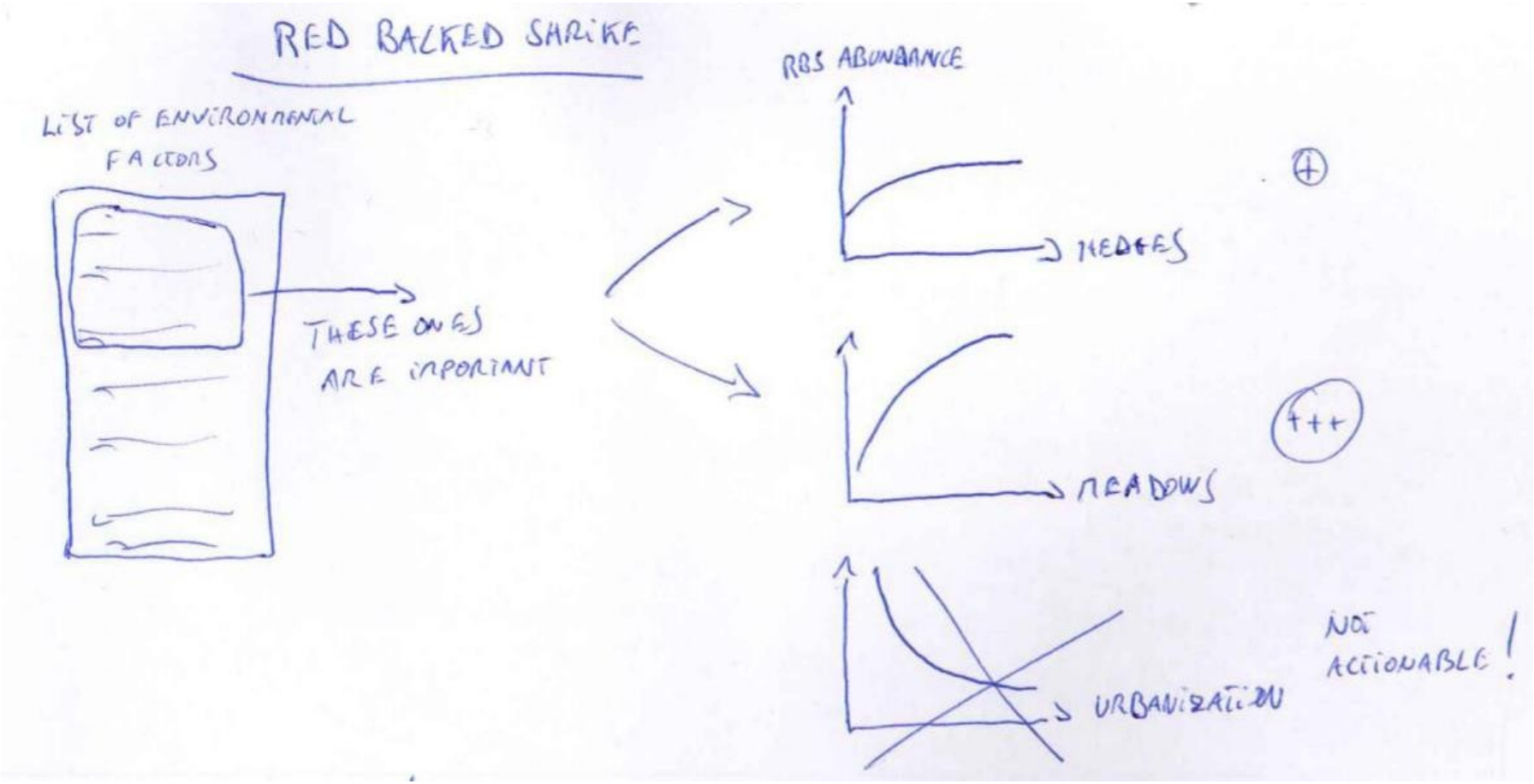
Qui va utiliser le résultat et **pourquoi** ?
Quelle sera la **forme idéale de l'output** ?

Idée approximative du type d'analyses statistiques

1 - Contexte et question scientifique

Ex : Croquis imaginant le résultat final d'une étude sur l'habitat de la pie grièche écorcheur

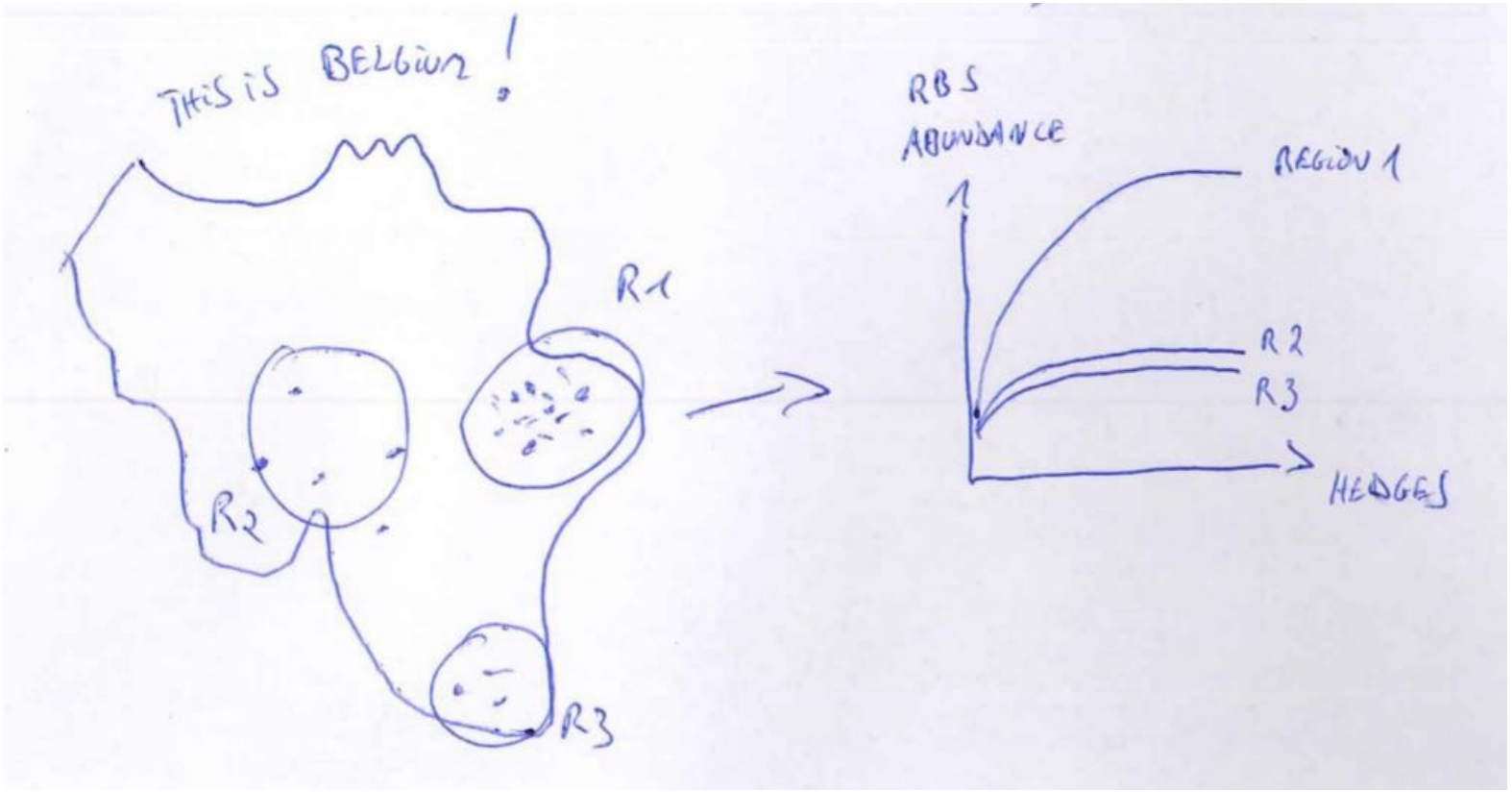
+ description sous forme de phrases



1 - Contexte et question scientifique

Les premiers résultats permettent parfois d'affiner les questions d'origine

Attention la fouille approfondie des données permet de générer des hypothèses mais pas toujours de les tester formellement



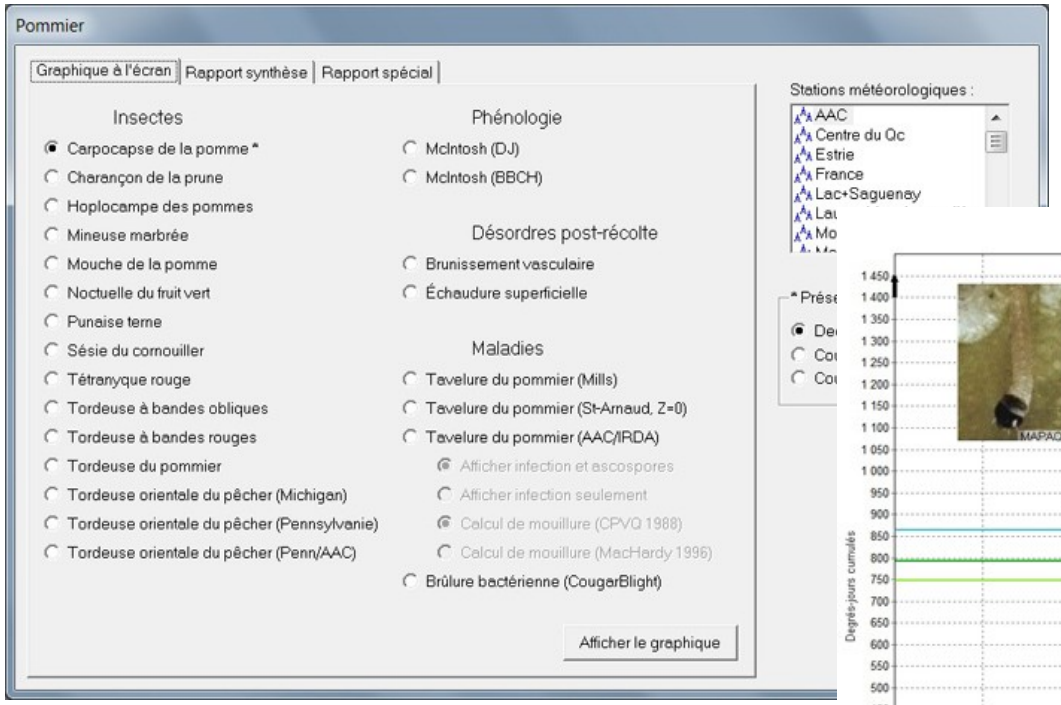
1 - Contexte et question scientifique

Conservation d'un papillon dans un parc naturel.
Les utilisateurs sont les gestionnaires du parc
--> On veut localiser les zones d'action prioritaires

1 - Contexte et question scientifique

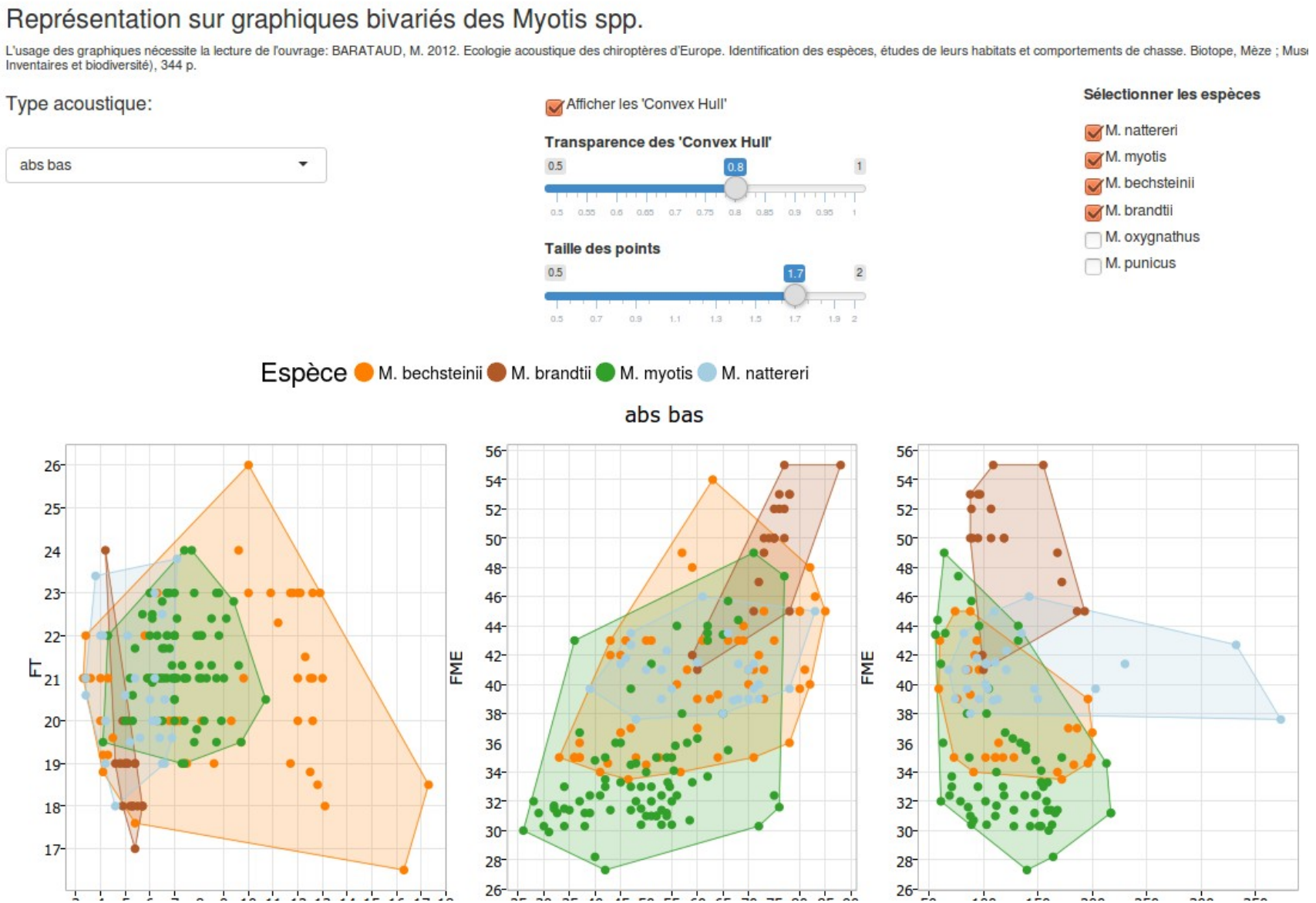
Ex : résultat de l'analyse sous la forme d'un logiciel (CIPRA)
Prédiction des ravageurs en fonction de la météo locale

Autre possibilité : avertissements par e-mail



1 - Contexte et question scientifique

Ex : Application interactive en ligne pour aider à l'identification des sons de chauves-souris



https://jeff37.shinyapps.io/Shiny1fileBarataud2016/

2 - Récolte des données

Récolter de nouvelles données

--> Expertise de domaine + notions de design expérimental

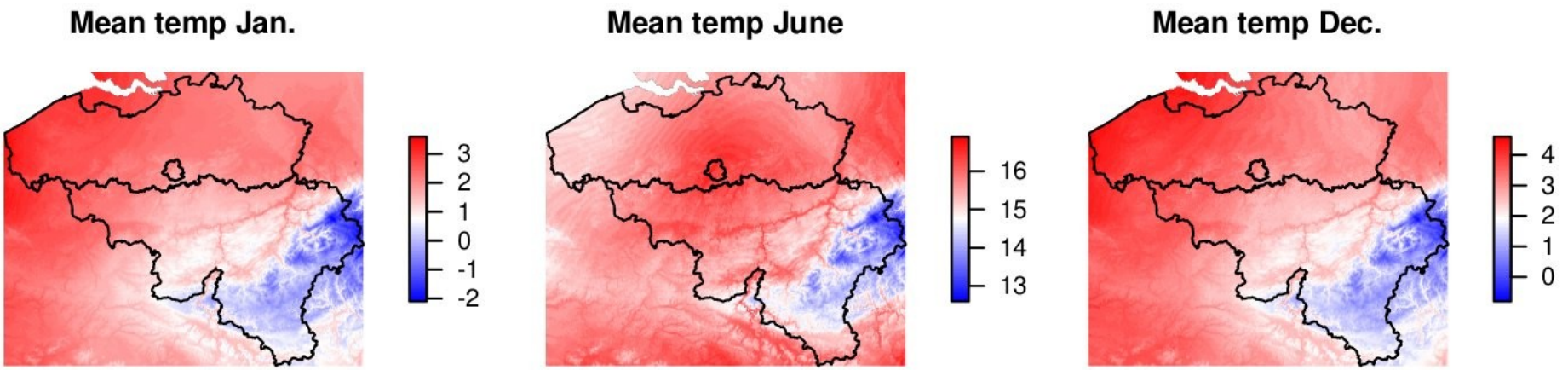
ET/OU

Rassembler des données existantes

--> Compétences informatiques

2 - Récolte des données

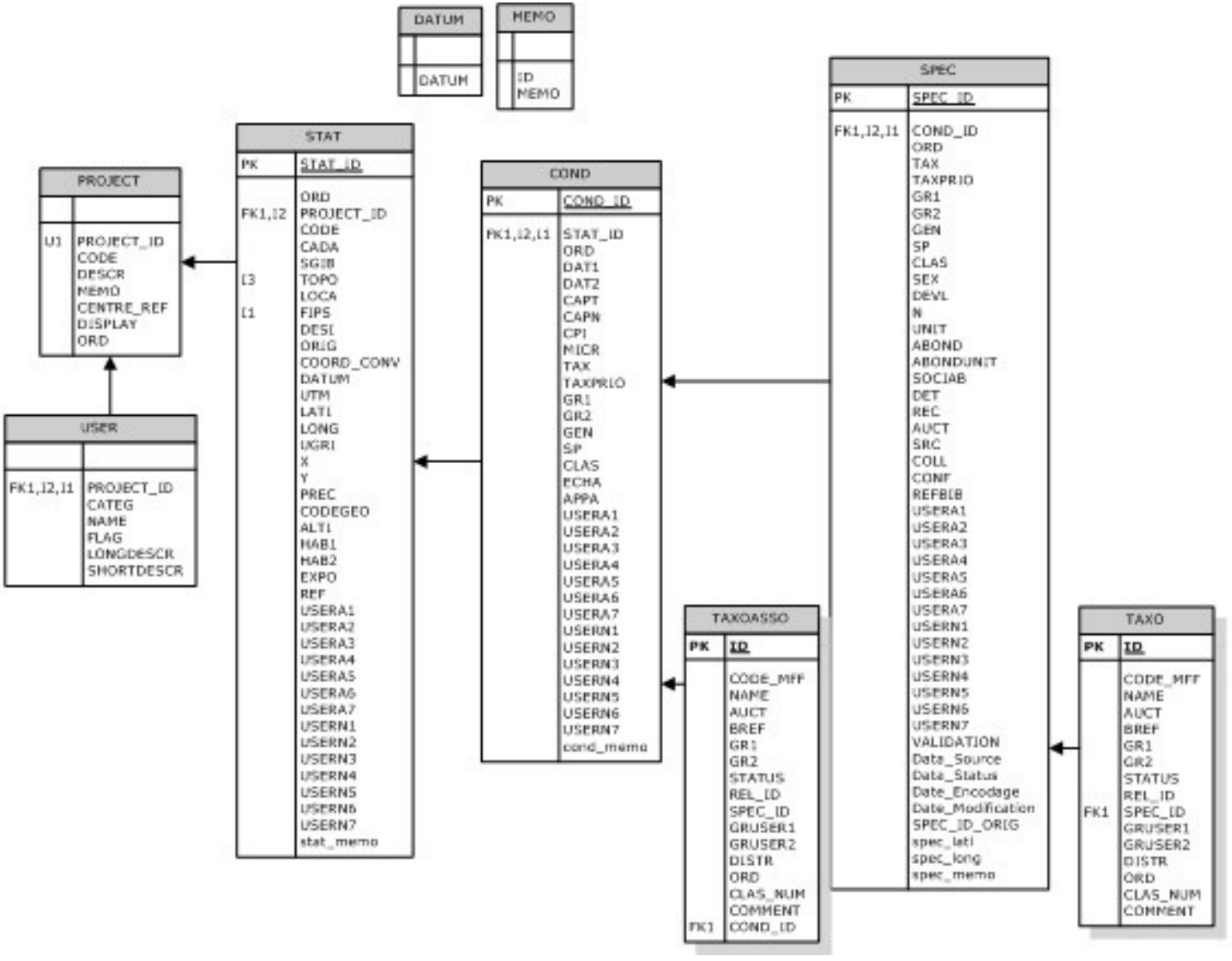
Ex : Données sous forme cartographique très fréquentes
Il faut pouvoir les extraire



2 - Récolte des données

Nombreux formats de données et de base de données

Ex structure BD SQL de données biogéographiques (DFF)



3 - Vérification et nettoyage des données

Vérification des erreurs, valeurs manquantes,...
Élimination de certaines données (ex GBIF)

```
> head(d)
  taille poids site nid   sexe            memo
1    9.4   5.3    3   2   male
2    9.8   5.3    1   1 female
3   11.6   6.3    2   2   <NA>       Mal formé
4   10.1   5.7    4   2 female
5   12.0   3.7    4   1   male peson déréglé ?
6   11.7   6.7    1   1 female

> summary(d)
     taille          poids           site           nid          sexe
 Min.   : 7.70   Min.   :3.700   Min.   :1.000  Min.   :1.000  female:36
 1st Qu.: 9.40   1st Qu.:4.900   1st Qu.:2.250  1st Qu.:2.000  femle : 1
 Median :10.05   Median :5.500   Median :4.000  Median :3.000  male  :38
 Mean   :10.29   Mean   :5.471   Mean   :3.987  Mean   :2.821  NA's  : 3
 3rd Qu.:10.78   3rd Qu.:6.000   3rd Qu.:6.000  3rd Qu.:4.000
 Max.   :29.00   Max.   :6.900   Max.   :7.000  Max.   :5.000
                 NA's   :2
```

4 - Exploration des données

Buts :

acquérir une connaissance approfondie des données
(cfr interprétation du résultat final)

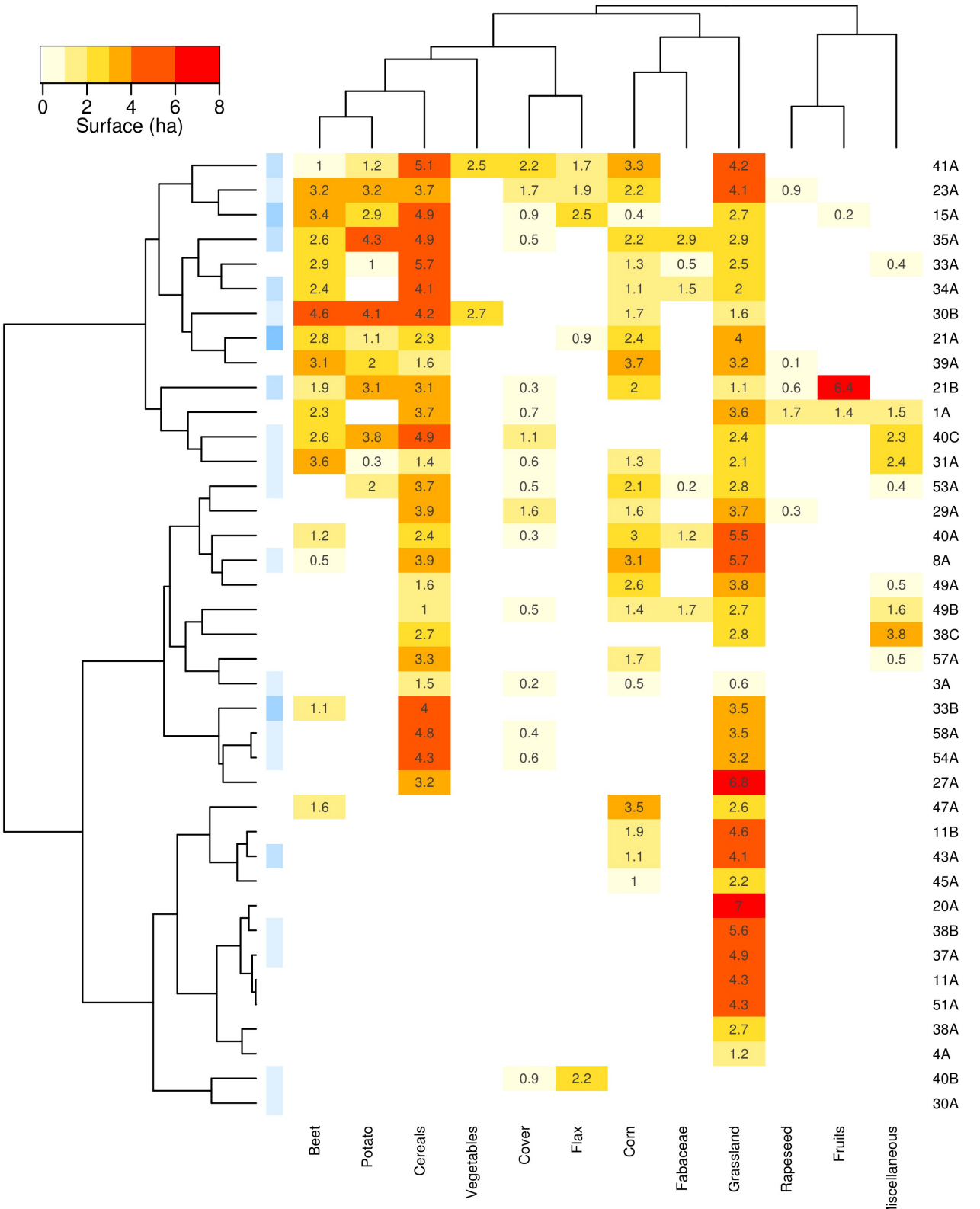
repérer les problèmes potentiels pour les analyses
statistiques subséquentes

DataViz
Visualisation des données particulièrement importante ici
(mais utile à toutes les étapes)

Visualisation d'une matrice complète : *surfaces de cultures 500m autour de ruchers*

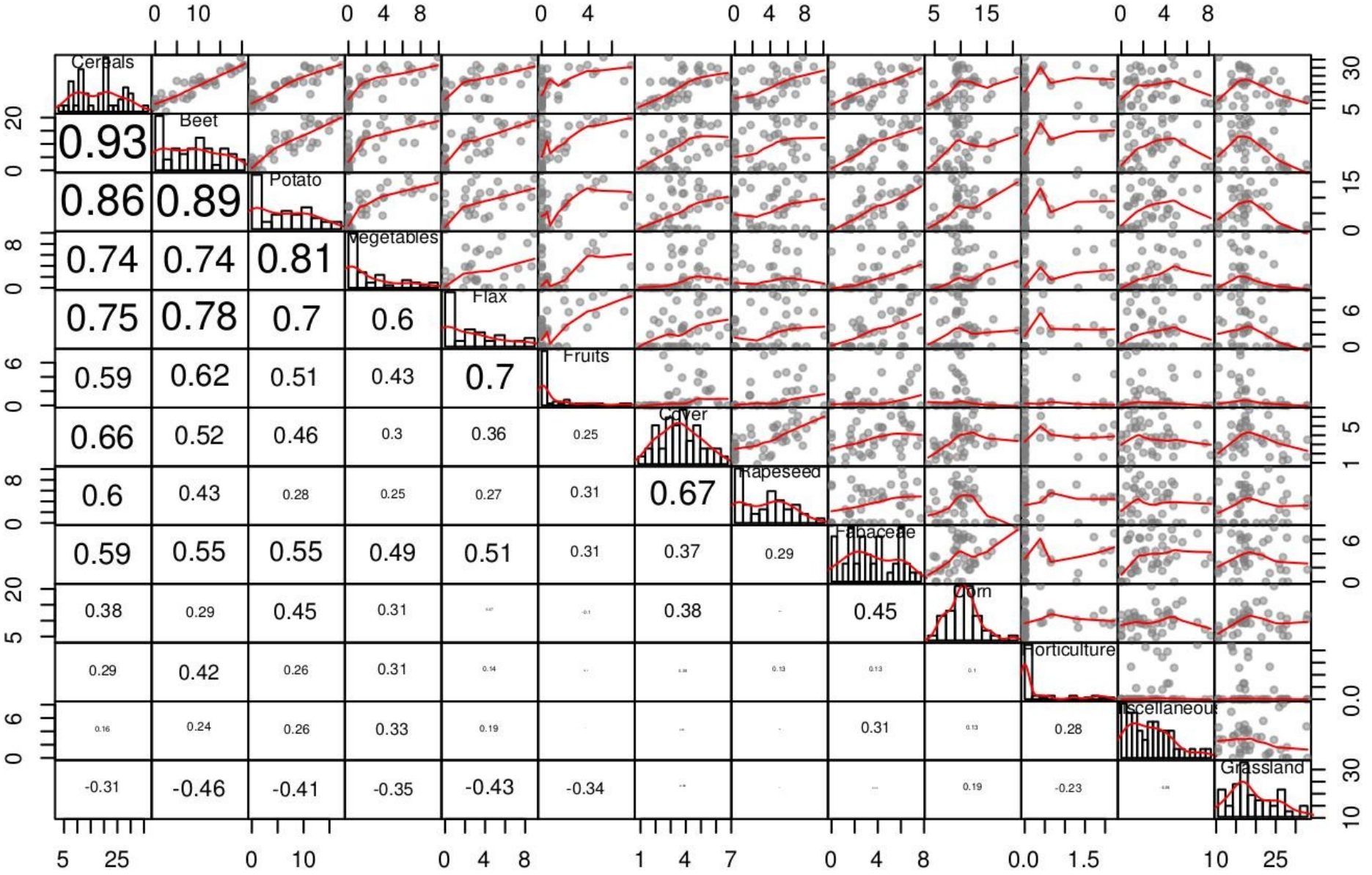
Statistiques :
distances Col + Row
clustering
réorganisation

Visualisation
Dendrogrames
Heatmap



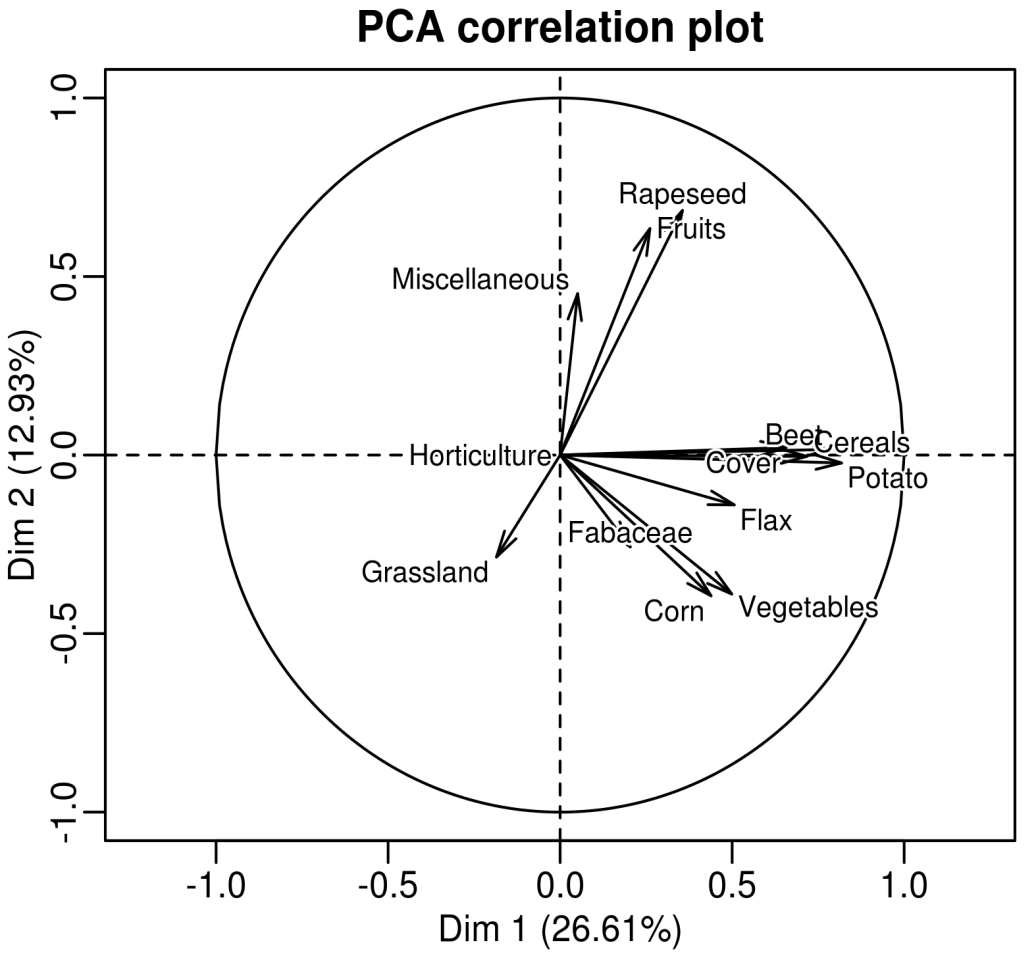
4 - Exploration des données

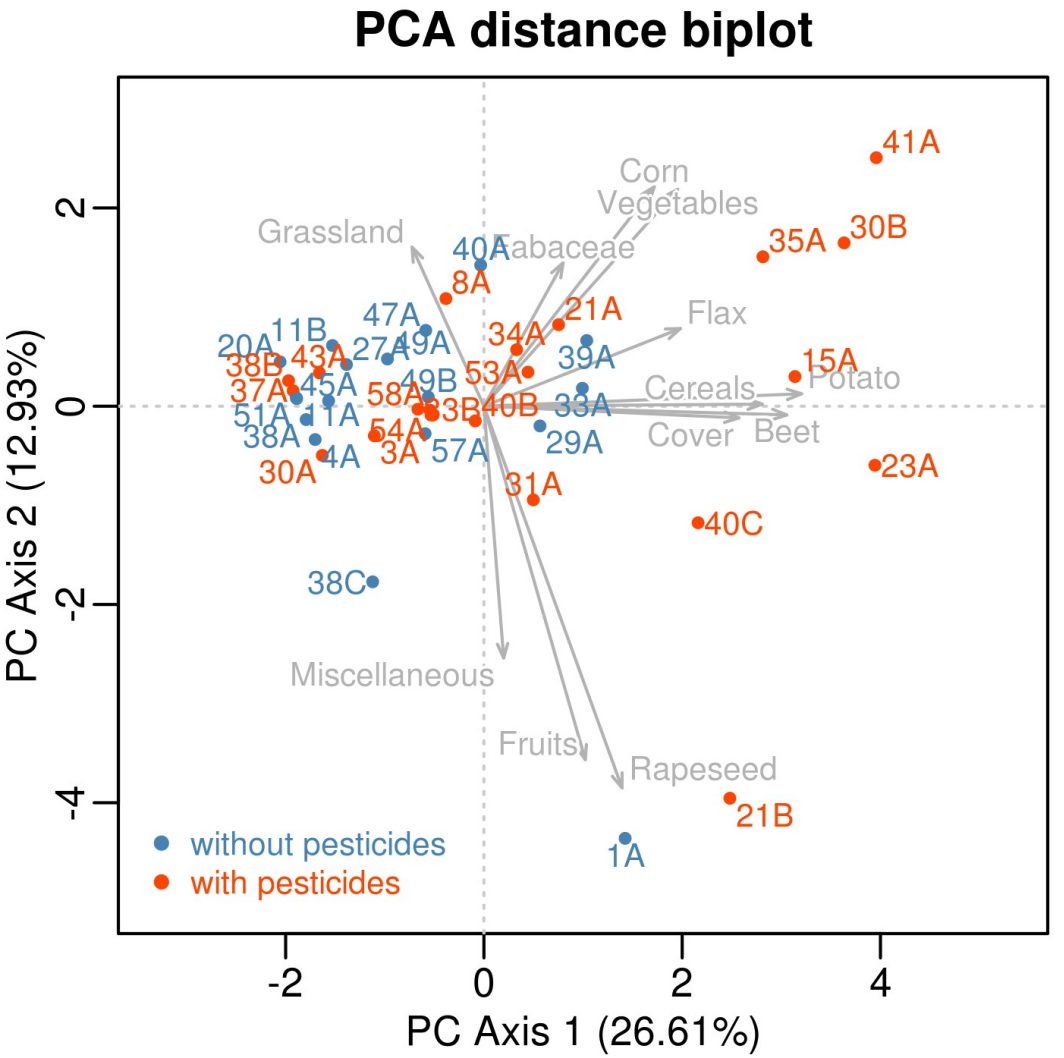
Autre visualisation de la même matrice de données



4 - Exploration des données

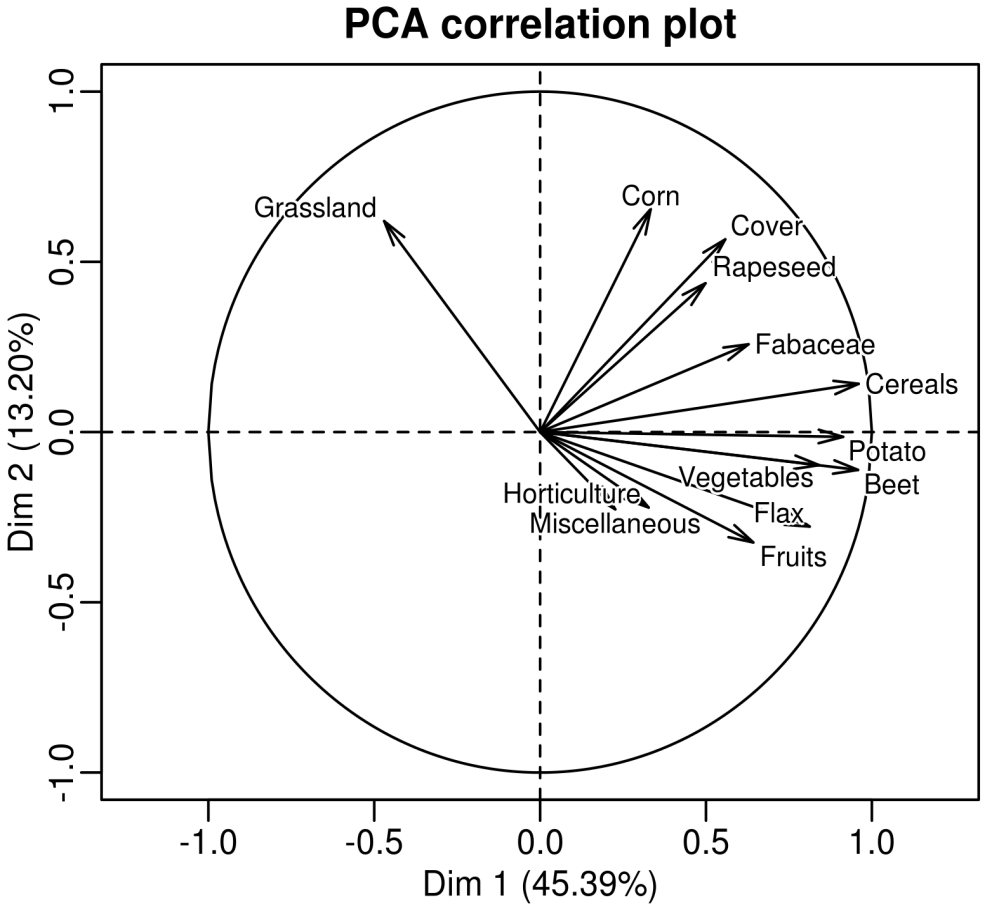
Encore une autre visualisation de la même matrice de données...

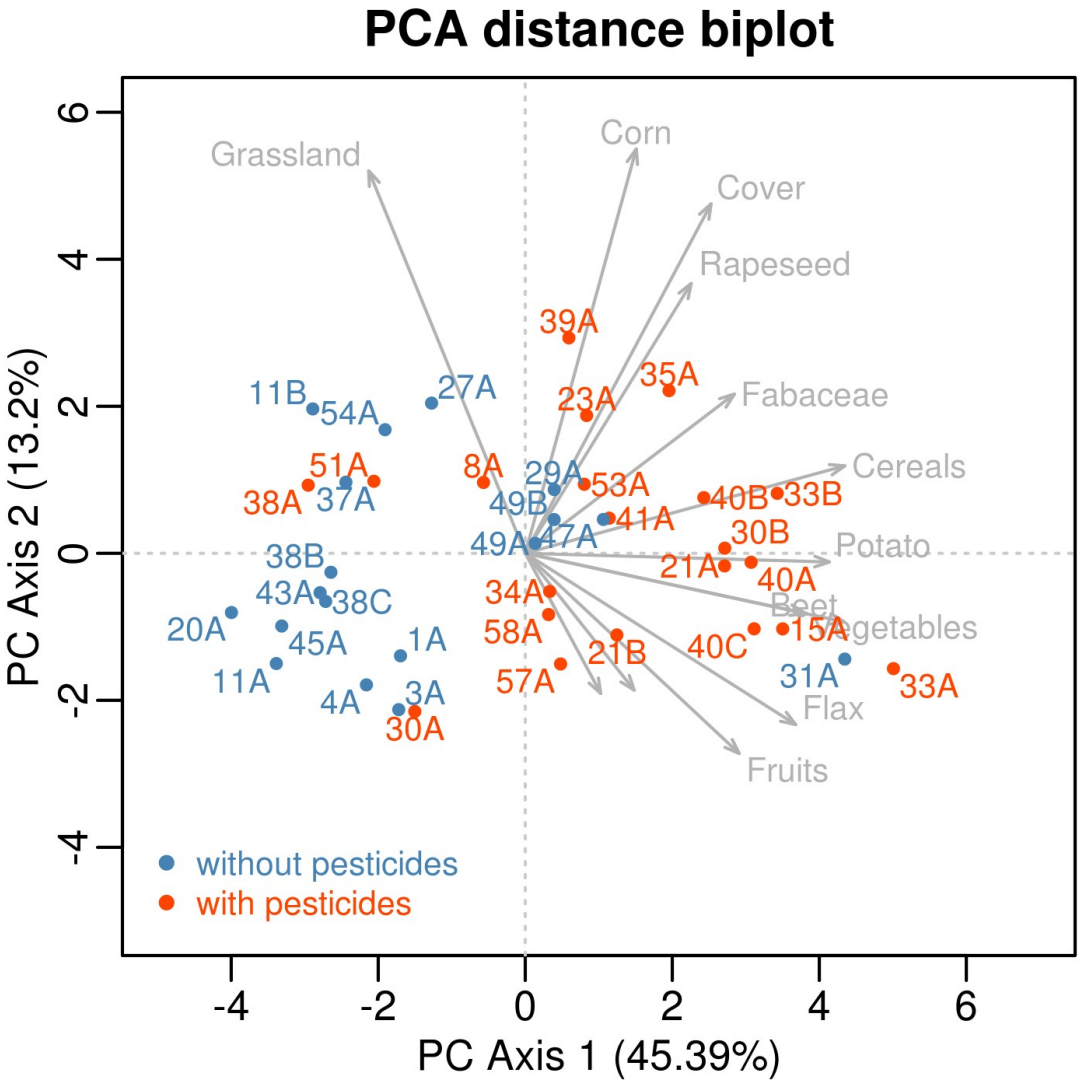




4 - Exploration des données

A une autre échelle (3km) d'autres patterns apparaissent

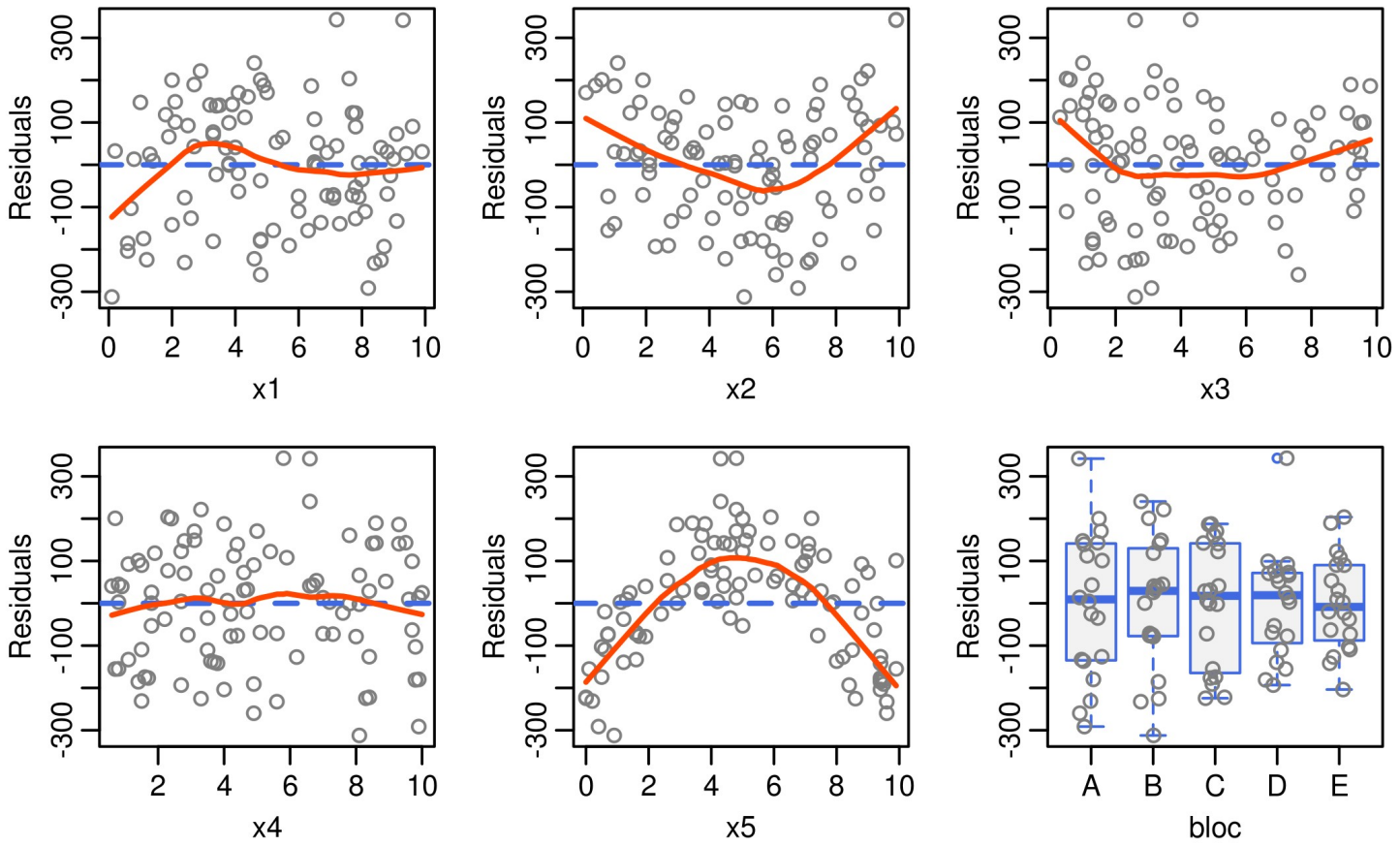




5 - Modélisation - Analyse statistique proprement dite

Etape A : Construction et vérification de modèle(s)

Examen critique pour repérer les problèmes pex pour GLM : linéarité, additivité, variance des résidus, multicolinéarité, valeurs extrêmes,...



5 - Modélisation - Analyse statistique proprement dite

Etape B : Inférences - Sélection des variables importantes

Pex sélection de modèles par AICc :

The first 10 best models (Models with delta.AICc<2 are equally supported by the data):

	model	AICc	AICc.delta	AICc.w	sum.w
4	Period+ bra	42.27	0	0.099	0.099
132	Period+ bra+ tar	44.36	2.088	0.035	0.134
20	Period+ bra+ pha	44.44	2.168	0.034	0.168
516	Period+ bra+ api	44.45	2.179	0.033	0.202
68	Period+ bra+ ast	44.48	2.208	0.033	0.235
8	Period+ bra+ ivy	44.51	2.231	0.033	0.267
12	Period+ bra+ tri	44.52	2.244	0.032	0.299
260	Period+ bra+ vic	44.55	2.274	0.032	0.331
36	Period+ bra+ ros	44.57	2.294	0.032	0.363
148	Period+ bra+ pha+ tar	46.58	4.303	0.012	0.375

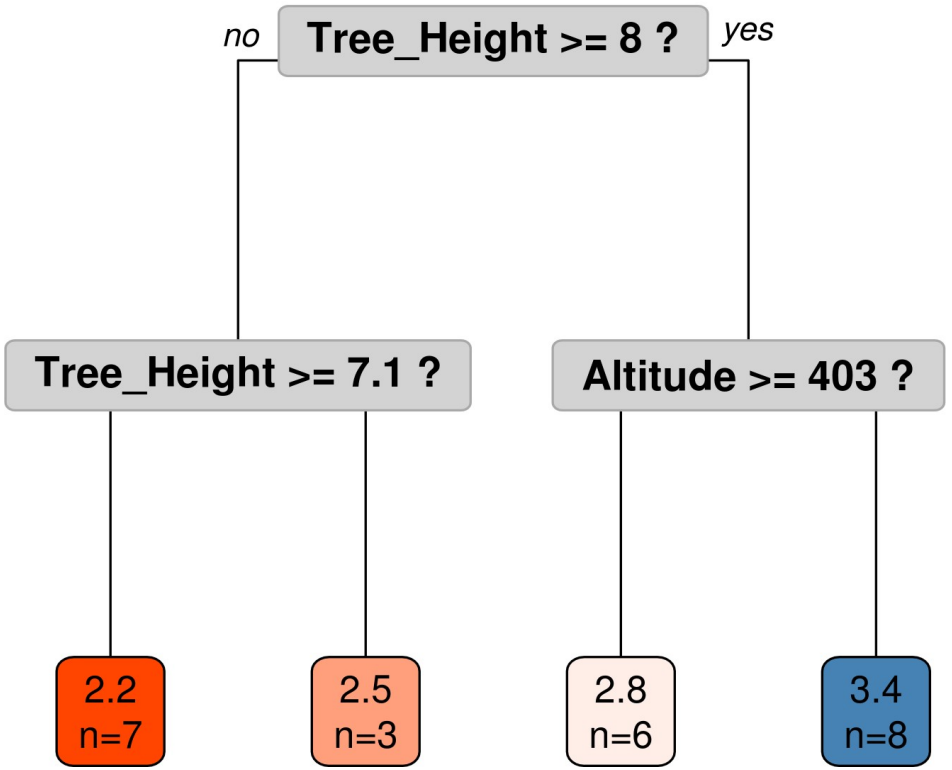
Model averaging results (variables with w>0.6 are supported by the data)

	freq	w	av.coef	av.se
(Intercept)	1	1	-4.179	2.274
PeriodSepOct	0.5	0.959	-3.282	1.288
bra	0.5	0.952	0.747	0.341
ivy	0.5	0.25	0.002	0.064
tar	0.5	0.248	0.028	0.071
api	0.5	0.244	-0.021	0.067

5 - Modélisation - Analyse statistique proprement dite

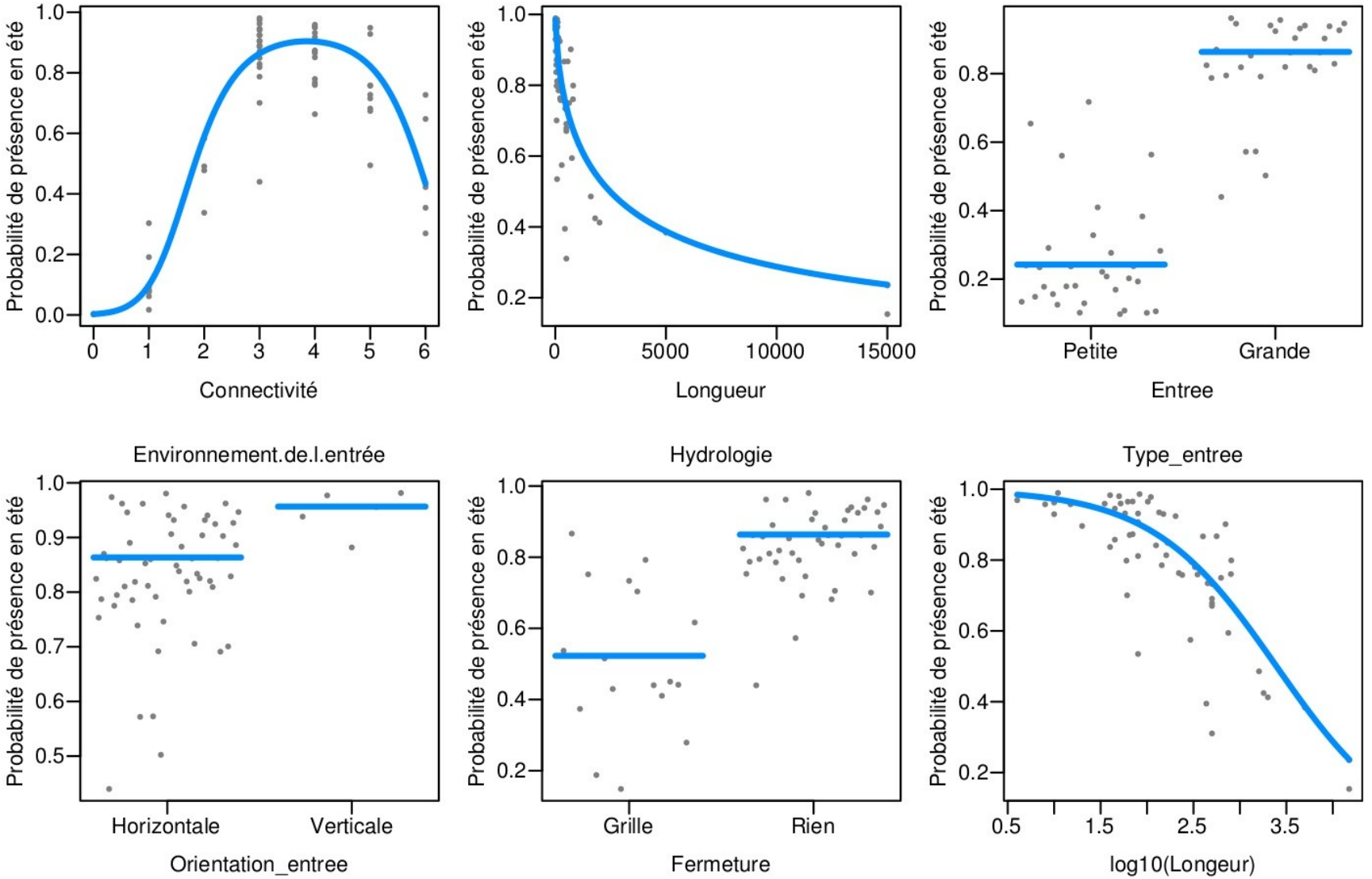
Éventuellement :
confirmation avec des approches complémentaires
--> robustesse des résultats ?

Exemple simple : Arbre de régression



6 - Interprétation biologique des résultats - Communication

Résultats chiffrés : souvent impossible à interpréter finement
--> représentation graphique !
Dépend de l'utilisation finale (cfr point 1)



6 - Interprétation biologique des résultats - Communication

Programmation + Communication
= "Reproducible research"

On doit pouvoir reproduire les résultats que vous avez obtenus depuis vos données brutes

Inclut votre futur vous-même !

Métadonnées
Litterate programming
(pex : knitr + R + markdown)