

Travail journalistique sur les labos de recherche CNRS avec R et le package data.table



Le jeu de données

Code Unité au 31/12/2018	Intitulé 2018	ETPT CNRS 31/12/2018	Masse salariale CNRS 31/12/2018	Crédits FEI CNRS 2018	Etablissement	Type partenariat
FR2488	Institut de recherche en s	0	0	21 000	AGROCAMPUS OUE	Tutelle
UMR6590	ESPACES ET SOCIETES	13,83	1 159 623	29 194	AGROCAMPUS OUE	Tutelle
UMR6625	INSTITUT DE RECHERC	21,87	1 952 542	122 000	AGROCAMPUS OUE	Partenaire institutionnel
UMS3343	Observatoire des Sciences	6,72	435 423	192 030	AGROCAMPUS OUE	Tutelle
ERL3559	Du gène à la graine	6,81	744 426	33 000	AGROPARISTECH	Partenaire institutionnel
FR3020	Fédération Ile de France	2,25	163 378	20 000	AGROPARISTECH	Partenaire institutionnel

- **Données partielles : CNRS uniquement. 7 variables de départ :**
- code unité, intitulé du labo,
- effectifs CNRS, masse salariale CNRS, crédits CNRS,
- établissement tutelle (principale ou secondaire)

Le mille-feuille des UMR

- En France, les labos de recherche ont souvent plusieurs tutelles : => unités « **mixtes** » de recherche
- **tutelle** = source principale des sous et du personnel du labo :
 - cosigne les publications scientifiques,
 - co-propriétaire des brevets.
- **partenaire institutionnel** = source secondaire de personnels et de sous. Cosigne les publis et empoche les brevets uniquement si son personnel est impliqué.

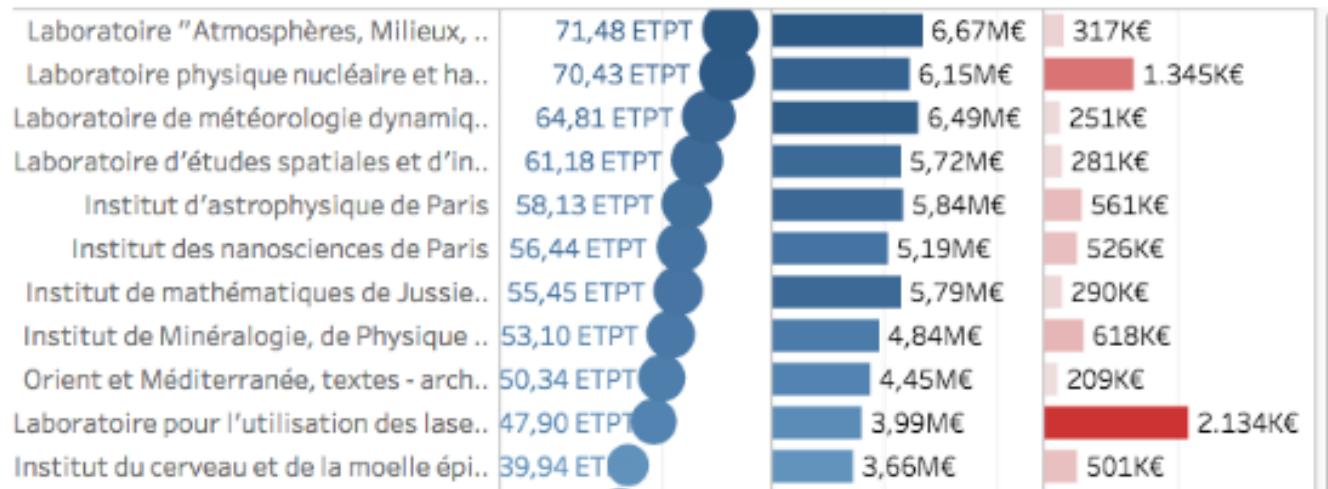


Unités de recherche : quels moyens financiers et humains le CNRS a-t-il octroyés en 2018 à chacune d'elles ?

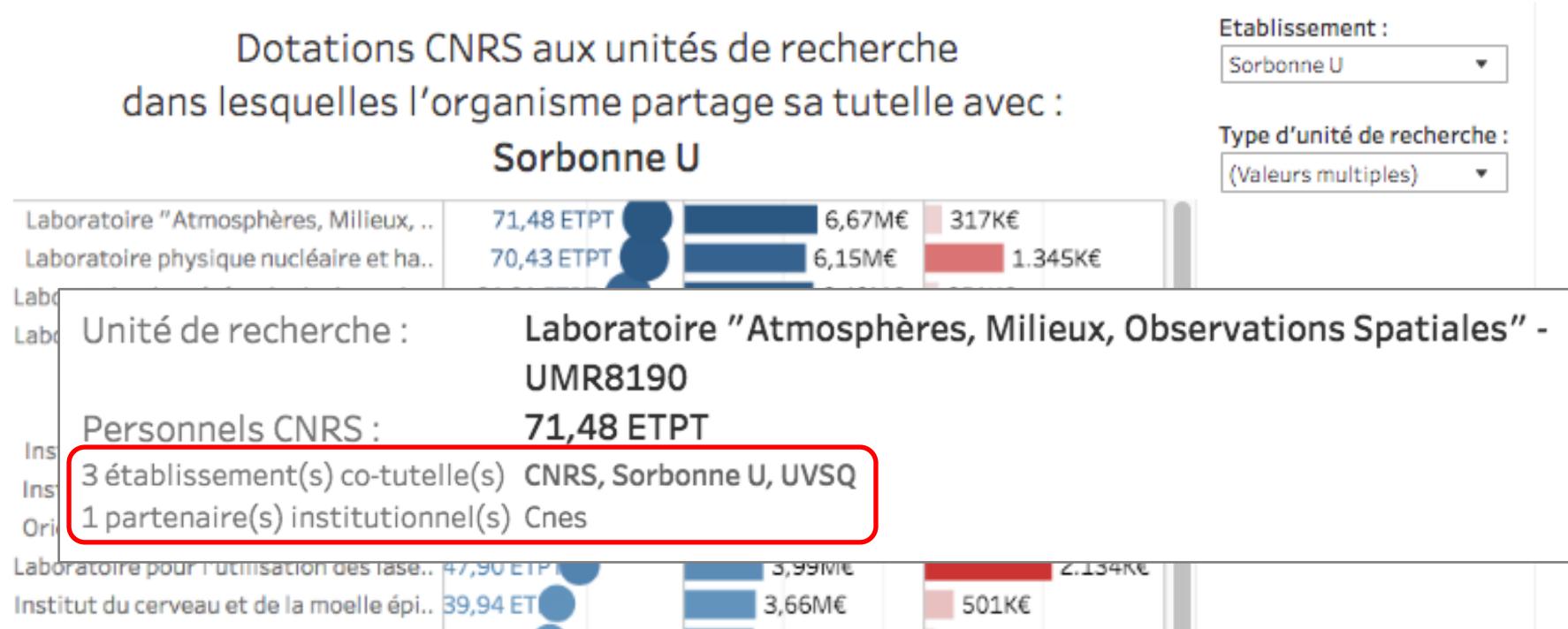
Dotations CNRS aux unités de recherche
dans lesquelles l'organisme partage sa tutelle avec :
Sorbonne U

Etablissement :
▼

Type d'unité de recherche :
▼



Unités de recherche : quels moyens financiers et humains le CNRS a-t-il octroyés en 2018 à chacune d'elles ?



4 colonnes avec le nombre et la liste des tutelles par labo

Code Unité au 31/12/2018	Intitulé 2018	ETPT Masse CNR salarial S e CNRS			CréditsF EICNRS2 018	Etablisseme	NbTutell	NbParte naireInsti		ListePartenaireIn stitutionel
		31/12/ 2018	31/12/ 2018	018				tutionel	ListeTutelles	
UMR5320	Acides r	6,05	501 415		58 000	INSERM	3	0	Inserm, université de Bordeaux	
UMR5320	Acides r	6,05	501 415		58 000	UNIV BOF	3	0	Inserm, université de Bordeaux	
UMR5281	Acteurs	11,3	893 921		21 952	CIRAD	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV MOI	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV PAL	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV PER	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR8256	Adaptati	16,1	#####		220 000	INSERM	2	1	Sorbonne U	Inserm
UMR8256	Adaptati	16,1	#####		220 000	UNIV SOF	2	1	Sorbonne U	Inserm

R et la librairie `data.table` à la rescouisse

Lire/écrire un fichier

```
Table = fread('monfichier.csv')
fwrite(Table, file='monfichier.csv')
```

- Rapide (utilise tous les processeurs)
- Devine les types d'objet, les séparateurs, s'il y a des noms de lignes / colonnes

R et la librairie `data.table` à la rescousse

`dt[i, j, by]`

Take `data.table dt`,
subset rows using `i`
and manipulate columns with `j`,
grouped according to `by`.

Opérations sur les lignes

dt[i, j, by]

Take data.table **dt**,
subset rows using **i**
and manipulate columns with **j**,
grouped according to **by**.

dt[a > 5,]

The diagram illustrates the operation `dt[a > 5,]`. It shows a 4x4 grid representing a data.table. The first column is labeled 'a' and contains values 2, 6, 6, and 5 respectively. The second column is yellow, the third is grey, and the fourth is grey. An arrow points from the original grid to a modified version where the row for 'a=5' has been removed, leaving only the row for 'a=6'.

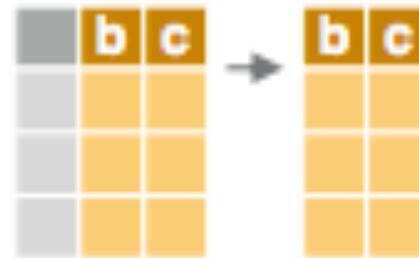
a			
2			
6			
5			

→

a			
6			

Opérations sur les colonnes

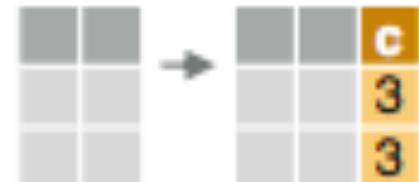
`dt[, .(b,c)]`



`dt[i, j, by]`

Take `data.table dt`,
subset rows using `i`
and manipulate columns with `j`,
grouped according to `by`.

`dt[, c := a + b]`

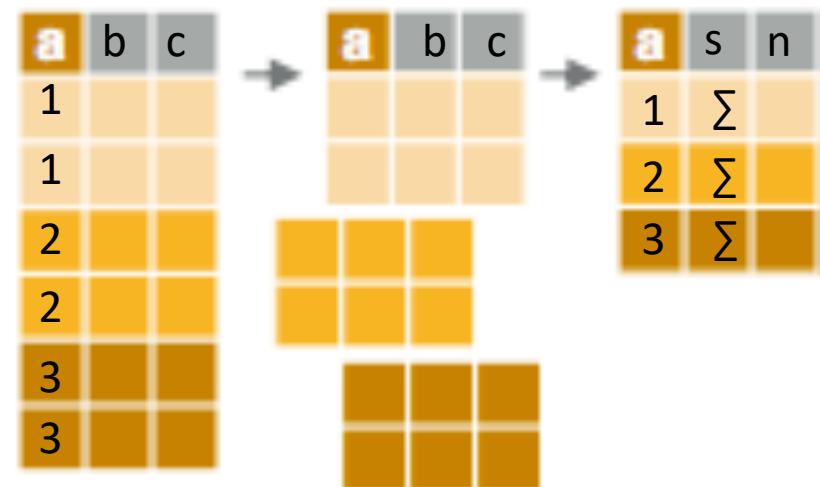


Calculer des valeurs de lignes en regroupant « par »

dt[, j, by = a]

dt[i, j, by]

Take data.table **dt**,
subset rows using **i**
and manipulate columns with **j**,
grouped according to **by**.



dt[, .(somme = sum(b), nombre = length(b)), by= a]

dt[, .(somme = sum(c)), by=(a, b)]

Mettre des colonnes en lignes

RESHAPE TO WIDE FORMAT

id	y	a	b
A	x	1	3
A	z	2	4
B	x	1	3
B	z	2	4

→

id	a	x	a	z	b	x	b	z
A	1		2		3		4	
B	1		2		3		4	

```
dcast(dt,  
      id ~ y,  
      value.var = c("a", "b"))
```

=> « spread » dans tidyverse

Mettre des lignes en colonnes

The diagram illustrates the transformation of a wide data frame into a long data frame. On the left, a wide data frame is shown with columns labeled 'id', 'a_x', 'a_z', 'b_x', and 'b_z'. The rows are labeled 'A' and 'B'. The values in the 'a_x' and 'a_z' columns are 1 and 2 respectively, while in 'b_x' and 'b_z', they are 3 and 4. An arrow points from this wide frame to a long data frame on the right. The long data frame has columns 'id', 'y', 'a', and 'b'. It contains four rows for each original row, with 'y' values 1 and 2, and 'a' and 'b' values corresponding to the original 'a' and 'b' columns.

id	a_x	a_z	b_x	b_z
A	1	2	3	4
B	1	2	3	4

id	y	a	b
A	1	1	3
B	1	1	3
A	2	2	4
B	2	2	4

```
melt(dt,  
      id.vars = c("id"),  
      measure.vars = patterns("^a|  
      variable.name = "y",  
      value.name = c("a", "b"))
```

=> « gather » dans tidyverse

Compter le nombre de tutelles et de partenaires institutionnels par unité de recherche

Créer une colonne où serait renseigné le **nombre de tutelles par labo**, auquel on rajoute 1 pour compter le CNRS

```
Table_UMR[ ,  
NbTutelle:=length(etab>Type.partenariat=='Tutelle')+ 1,  
by = Code.unité]
```

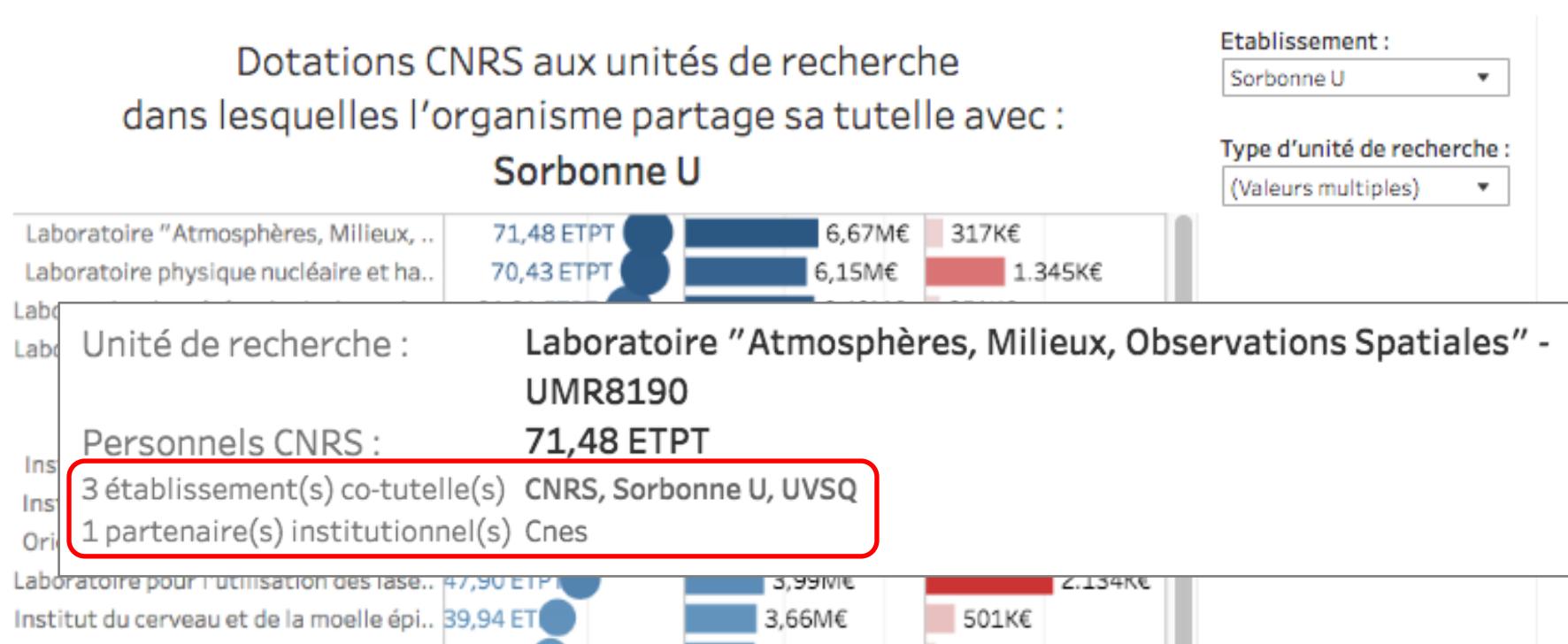
Créer une colonne avec **la liste des noms des tutelles**, séparés par des virgules

```
Table_UMR[ , ListeTutelles:=paste(etab>Type.partenariat=='Tutelle',  
collapse=', '), by = Code.unité]
```

4 colonnes avec le nombre et la liste des tutelles par labo

Code Unité au 31/12/2018	Intitulé 2018	ETPT Masse CNR salarial S e CNRS			CréditsF EICNRS2 018	Etablisseme	NbTutell	NbParte naireInsti		ListePartenaireIn stitutionel
		31/12/ 2018	31/12/ 2018	018				tutionel	ListeTutelles	
UMR5320	Acides r	6,05	501 415		58 000	INSERM	3	0	Inserm, université de Bordeaux	
UMR5320	Acides r	6,05	501 415		58 000	UNIV BOF	3	0	Inserm, université de Bordeaux	
UMR5281	Acteurs	11,3	893 921		21 952	CIRAD	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV MOI	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV PAL	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR5281	Acteurs	11,3	893 921		21 952	UNIV PER	5	0	Cirad, Montpellier, Montpellier-III, Pe	
UMR8256	Adaptati	16,1	#####		220 000	INSERM	2	1	Sorbonne U	Inserm
UMR8256	Adaptati	16,1	#####		220 000	UNIV SOF	2	1	Sorbonne U	Inserm

Unités de recherche : quels moyens financiers et humains le CNRS a-t-il octroyés en 2018 à chacune d'elles ?

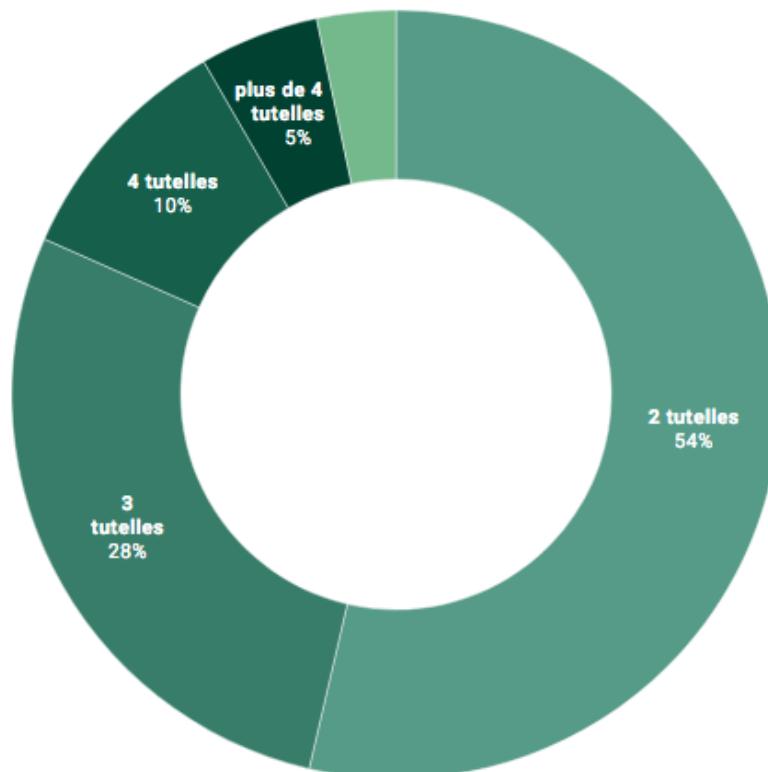


43% des unités de recherche CNRS ont plus de 2 tutelles en 2018

43% des unités de recherche ont plus de deux tutelles

Nombre de tutelles des unités de recherche dont le CNRS est tutelle au 31 décembre 2018, comprenant les UMR, UPR, USR, FRE et ERL.

■ 2 tutelles ■ 3 tutelles ■ 4 tutelles ■ plus de 4 tutelles ■ une tutelle (le CNRS)

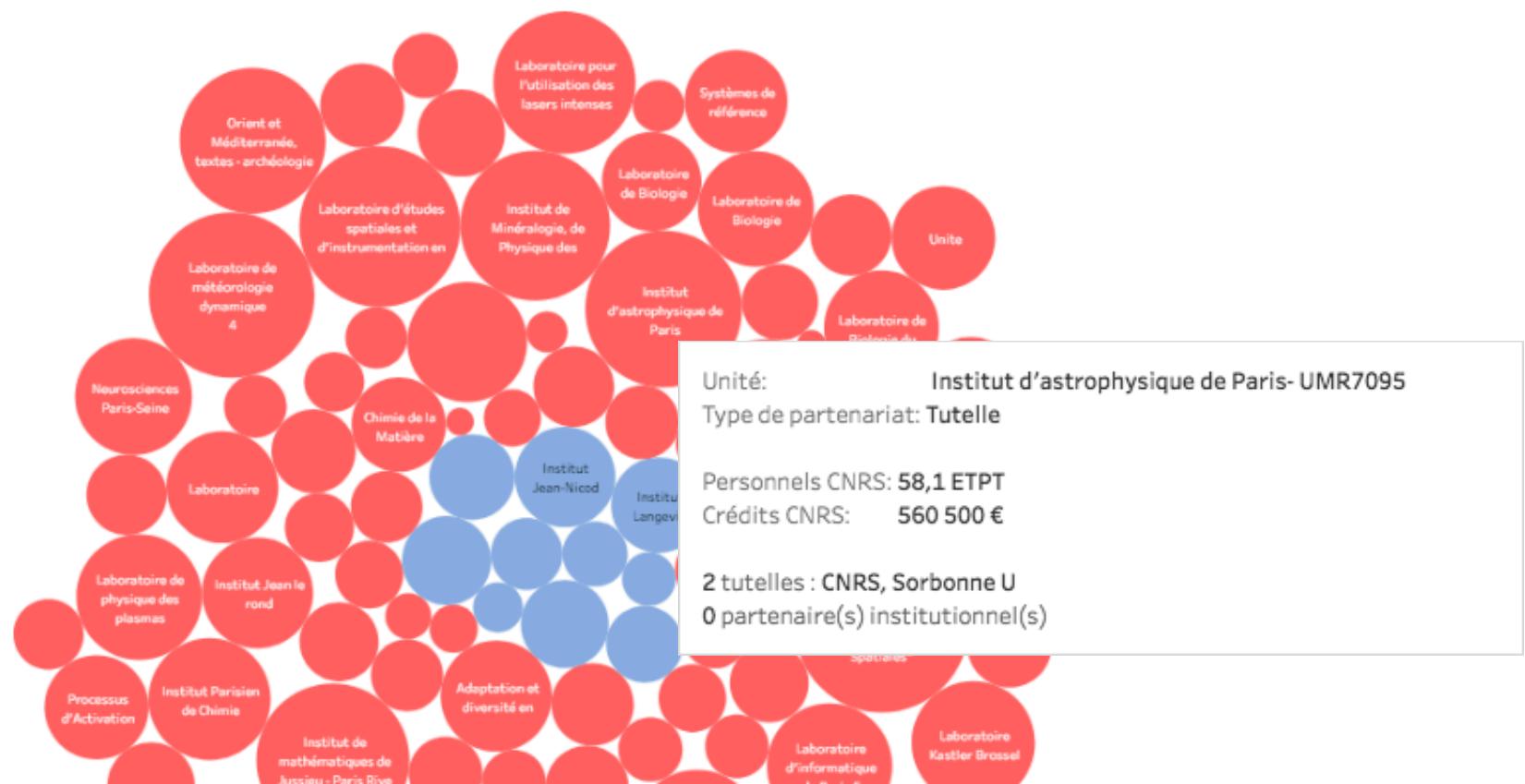


Universités, écoles, organismes : qui sont les tutelles des unités mixtes du CNRS ?

Etablissement

Sorbonne U

Unités de recherche dont Sorbonne U est tutelle principale ou secondaire



Universités, écoles, organismes : qui sont les tutelles des unités mixtes du CNRS ?

nom court	Type partenariat	Regroupement	type d'établissement
Inserm	Tutelle	Bordeaux	organisme de recherche
université de Bordeaux	Tutelle	Bordeaux	université
Cirad	Tutelle	Languedoc-Roussillon	organisme de recherche
Montpellier	Tutelle	Languedoc-Roussillon	université
Montpellier-III	Tutelle	Languedoc-Roussillon	université
Perpignan	Tutelle	Languedoc-Roussillon	université
Inserm	Partenaire inst	Sorbonne U	organisme de recherche

Unités dont Polytechnique est tutelle principale



Part des unités mixtes CNRS ayant plus de deux tutelles

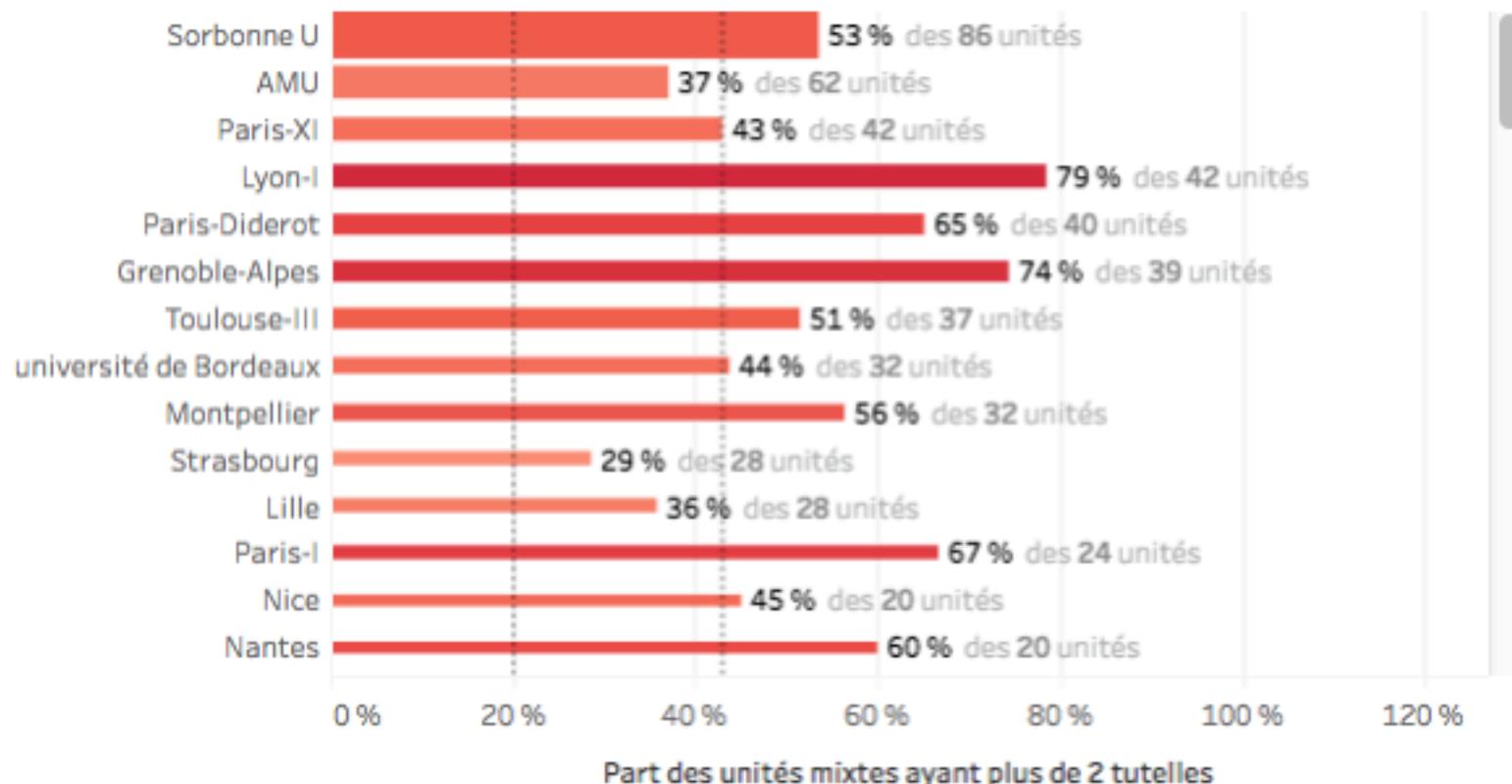
Type d'établissement

université



Etablissement

Surligner Nom.Court



Quel est le 2ème meilleur partenaire d'un établissement ?

Code.Unité.au.31.12.2018	nom.court	associated_etab	liste_tutellies
UMR6590	AgroCampus Ouest	Angers	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Caen	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Le Mans	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Nantes	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Rennes-II	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR8172	AgroParisTech	Cirad	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Inra	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Antilles	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Guyane	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8079	AgroParisTech	Paris-XI	AgroParisTech,Paris-XI
UMR8120	AgroParisTech	Inra	AgroParisTech,Inra,Paris-XI
UMR8120	AgroParisTech	Paris-XI	AgroParisTech,Inra,Paris-XI
UMR8586	AgroParisTech	IRD	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR8586	AgroParisTech	Paris-I	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR8586	AgroParisTech	Paris-Diderot	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR9000	AgroParisTech	Cirad	AgroParisTech,Cirad,Irstea
UMR9000	AgroParisTech	Irstea	AgroParisTech,Cirad,Irstea
UMR6265	AgroSup Dijon	Inra	AgroSup Dijon,Inra,Dijon
UMR6265	AgroSup Dijon	Dijon	AgroSup Dijon,Inra,Dijon
UMR7025	Amiens	UTC	Amiens,UTC

Trouver le 2^{ème} meilleur partenaire

Sortir les paires d'établissements qui partagent la tutelle principale d'un labo

```
Table_UMR[, .(associated_etab=strsplit(ListeTutelles, ','[[1]] ,  
by=.(Code.unità, etab)]
```

Quel est le 2ème meilleur partenaire d'un établissement ?

Code.Unité.au.31.12.2018	nom.court	associated_etab	liste_tutellies
UMR6590	AgroCampus Ouest	Angers	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Caen	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Le Mans	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Nantes	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR6590	AgroCampus Ouest	Rennes-II	AgroCampus Ouest,Angers,Caen,Le Mans,Nantes,Rennes-II
UMR8172	AgroParisTech	Cirad	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Inra	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Antilles	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8172	AgroParisTech	Guyane	AgroParisTech,Cirad,Inra,Antilles,Guyane
UMR8079	AgroParisTech	Paris-XI	AgroParisTech,Paris-XI
UMR8120	AgroParisTech	Inra	AgroParisTech,Inra,Paris-XI
UMR8120	AgroParisTech	Paris-XI	AgroParisTech,Inra,Paris-XI
UMR8586	AgroParisTech	IRD	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR8586	AgroParisTech	Paris-I	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR8586	AgroParisTech	Paris-Diderot	AgroParisTech,IRD,Paris-I,Paris-Diderot
UMR9000	AgroParisTech	Cirad	AgroParisTech,Cirad,Irstea
UMR9000	AgroParisTech	Irstea	AgroParisTech,Cirad,Irstea
UMR6265	AgroSup Dijon	Inra	AgroSup Dijon,Inra,Dijon
UMR6265	AgroSup Dijon	Dijon	AgroSup Dijon,Inra,Dijon
UMR7025	Amiens	UTC	Amiens,UTC

Trouver le 2^{ème} meilleur partenaire

Sortir les paires d'établissements qui partagent la tutelle principale d'un labo

```
Table_UMR[, .(associated_etab=strsplit(ListeTutelles, ','[[1]]),  
by=.(Code.unità, etab)]
```

Retirer les paires avec 2 fois le même établissement

```
Table_2MP[associated_etab!=etab,]
```

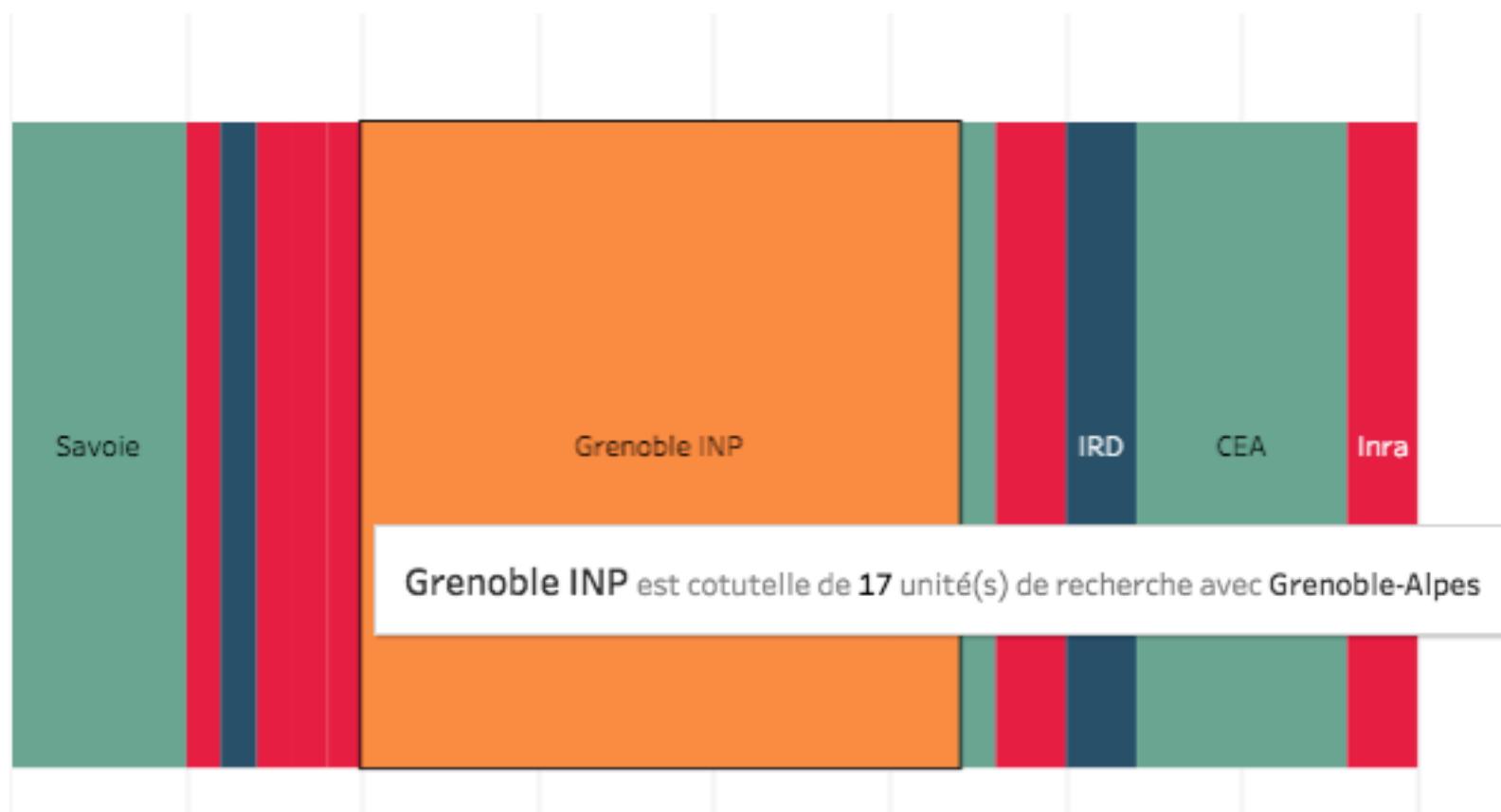
Quel est le plus important deuxième partenaire, pour chaque établissement ?

Etablissement

Grenoble-Alpes

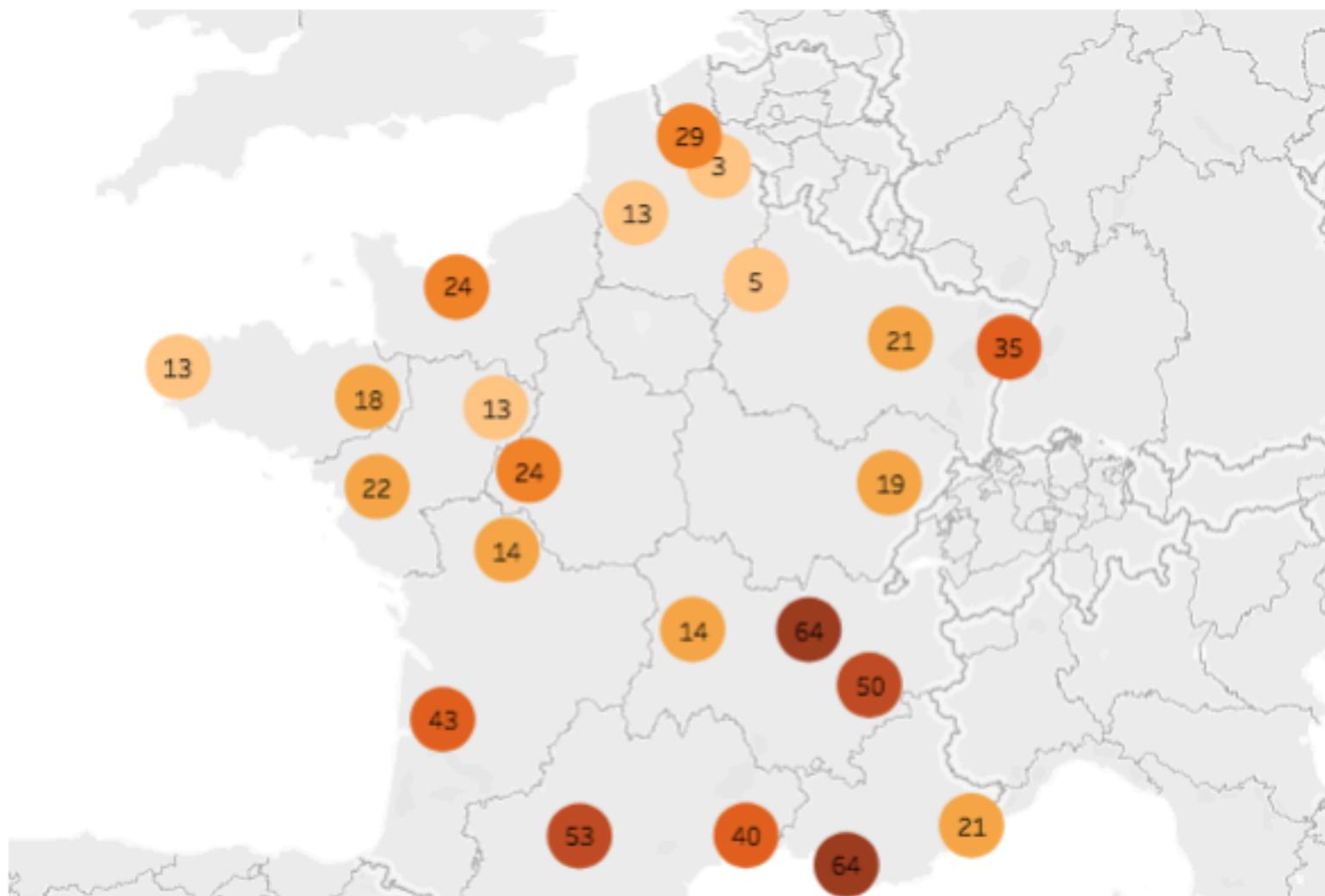


Grenoble-Alpes partage la tutelle de 1 à 17 unités mixtes CNRS avec chacun des établissements suivants :



Regroupements !

Unités mixtes CNRS dans les sites universitaires en région, et en Ile-de-France



Regroupements !

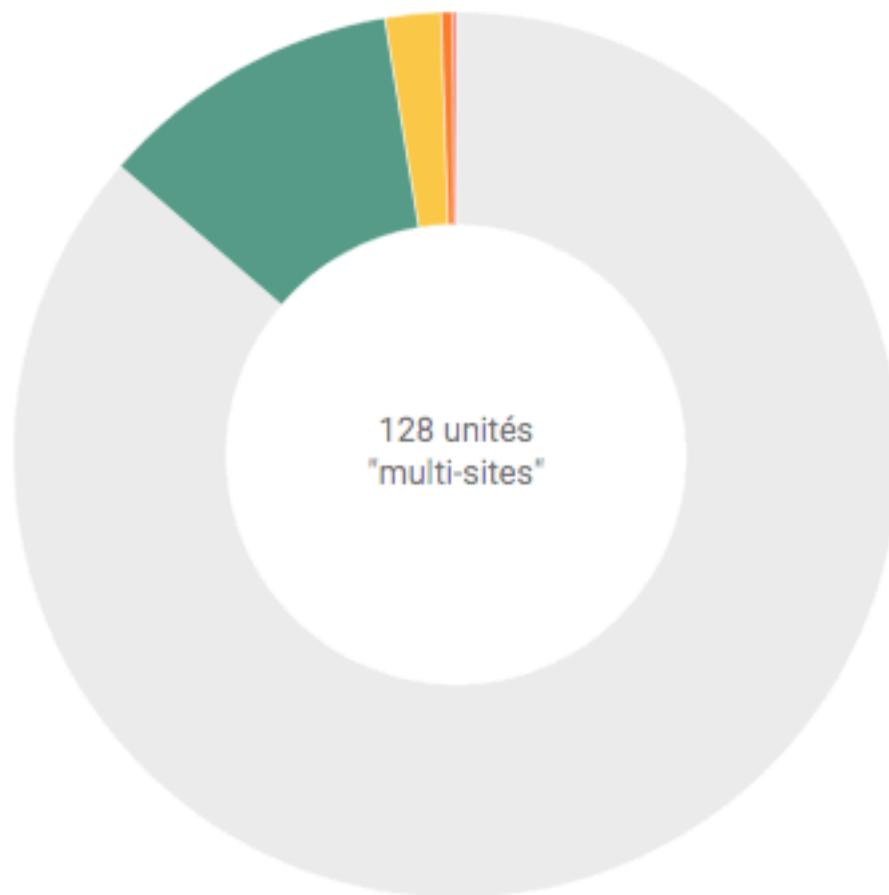
Etablissement	nom court	Type partenariat	Regroupement
INSERM	Inserm	Tutelle	Bordeaux
UNIV BORDEAUX	université de Bord	Tutelle	Bordeaux
CIRAD	Cirad	Tutelle	Languedoc-Rous
UNIV MONTPELLI	Montpellier	Tutelle	Languedoc-Rous
UNIV PAUL VALE	Montpellier-III	Tutelle	Languedoc-Rous
UNIV PERPIGNAN	Perpignan	Tutelle	Languedoc-Rous

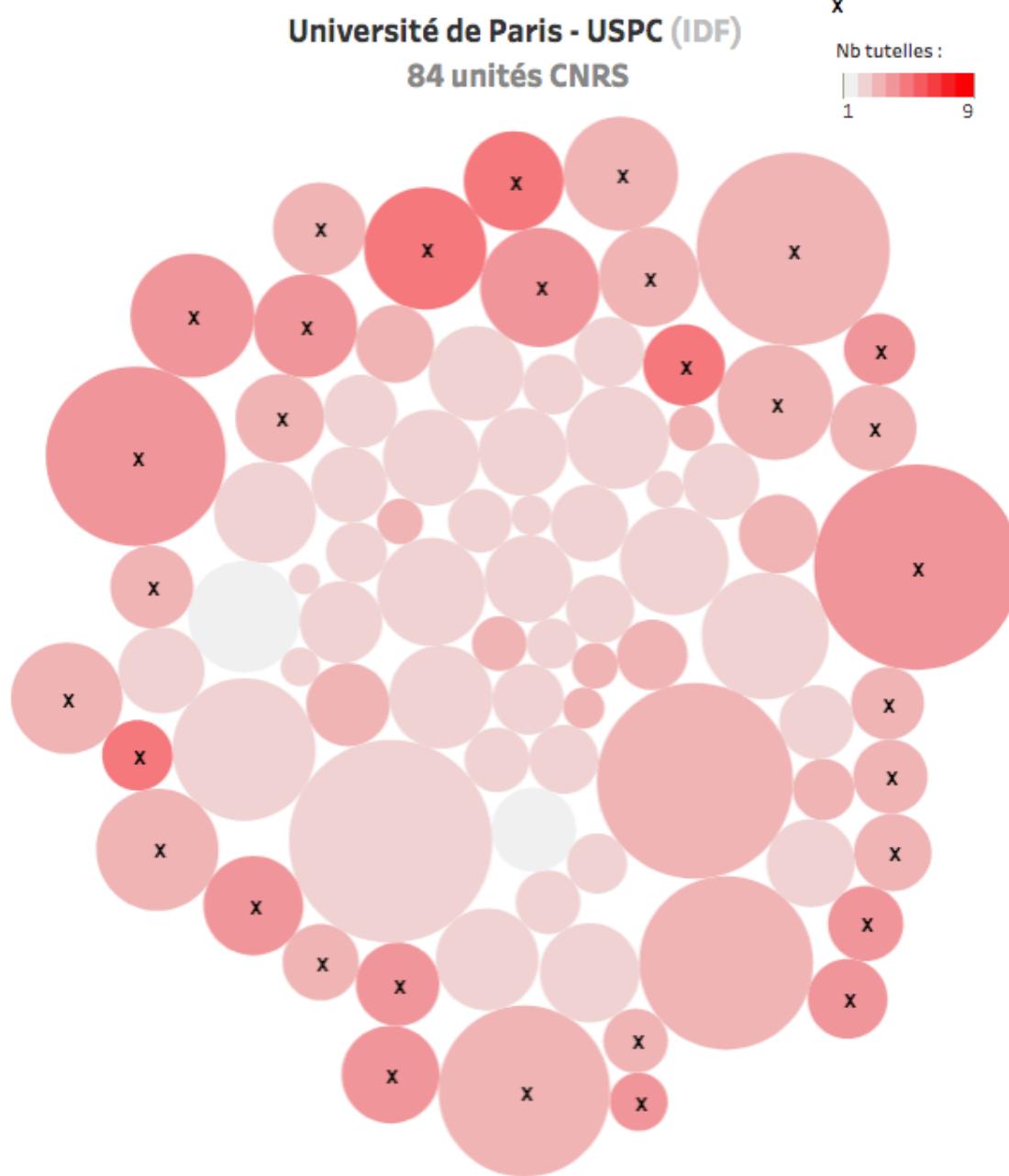
Regroupements !

14% des unités de recherche CNRS relèvent de plusieurs regroupements

Parmi les 934 unités de recherche CNRS qu'il est possible de rattacher à un site ou ensemble universitaire, 128 laboratoires sont "multi-regroupements".

- 1 regroupement (806)
- 2 regroupements (104)
- 3 regroupements (19)
- 4 regroupements (4)
- 6 regroupements (1)





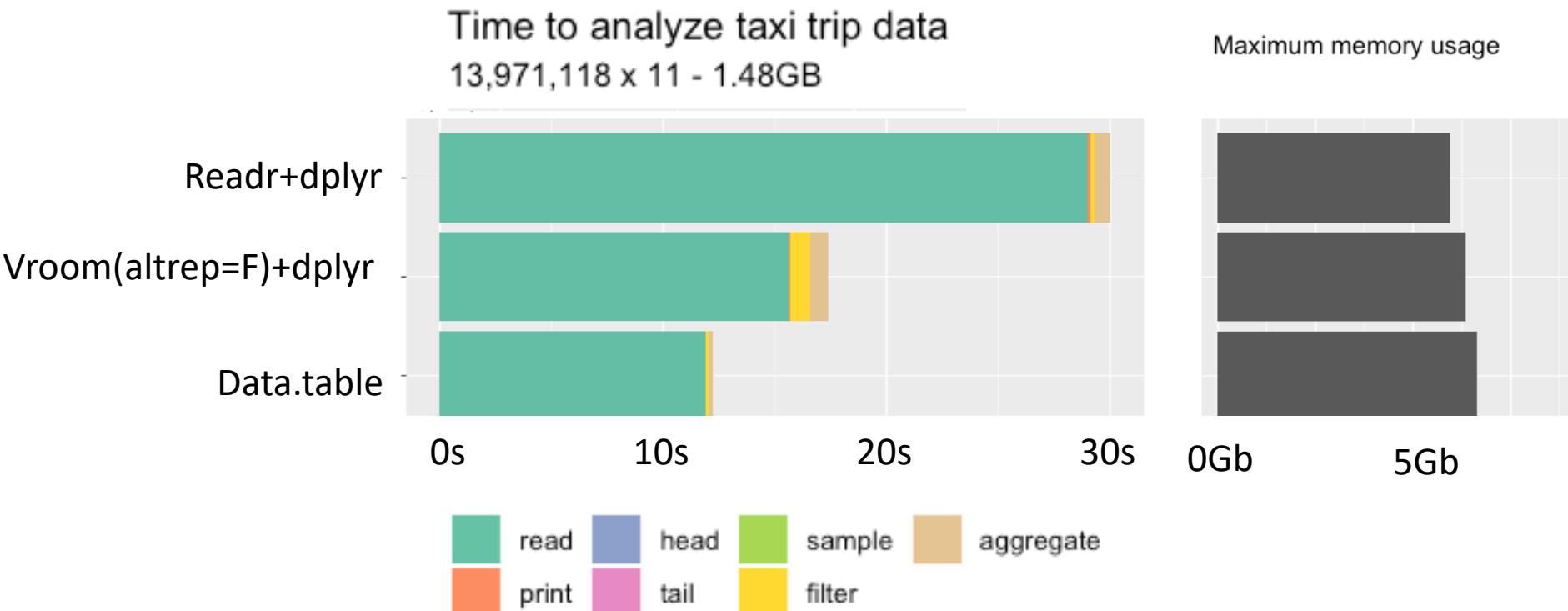
Les scripts sont sur github

- <https://github.com/anouchk/DistributionNbTutellesParEtab>
- https://github.com/anouchk/paste_CNRS
- https://github.com/anouchk/unique_gsub_asnumeric_CNRS
- <https://github.com/anouchk/Meilleur2emePartenaire>
- <https://github.com/anouchk/UMR dans les sites universitaires>



data.table ou tidyverse ?

Plus performant



data.table ou tidyverse ?

Moins facile d'injecter des variables/
factoriser le code

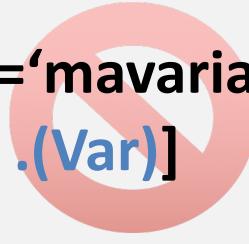
R de base

```
dt[, c('b', 'c')]
```

```
Var='mavariable'  
dt[, Var]
```

Data.table

```
dt[, .(b,c)]
```

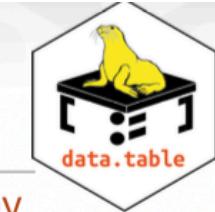
Var='mavariable'
dt[, .(Var)]

```
Var='mavariable'  
dt[, mget(Var)]
```

Y'a une cheatsheet

- <https://github.com/rstudio/cheatsheets/raw/master/datatable.pdf>

Data Transformation with data.table :: CHEAT SHEET



Basics

data.table is an extremely fast and memory efficient package for transforming data in R. It works by converting R's native data frame objects into data.tables with new and enhanced functionality. The basics of working with data.tables are:

dt[i, j, by]

Take data.table **dt**,
subset rows using **i**
and manipulate columns with **j**,
grouped according to **by**.

data.tables are also data frames – functions that work with data frames therefore also work with data.tables.

Create a data.table

data.table(a = c(1, 2), b = c("a", "b")) – create a data.table from scratch. Analogous to `data.frame()`.

setDT(df)* or as.data.table(df) – convert a data frame or a list to a data.table.

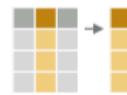
Subset rows using i



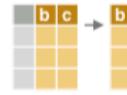
dt[1:2,] – subset rows based on row numbers.

Manipulate columns with j

EXTRACT



dt[, c(2)] – extract columns by number. Prefix column numbers with “-” to drop.



dt[, .(b, c)] – extract columns by name.

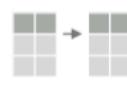
SUMMARIZE



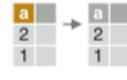
dt[, .(x = sum(a))] – create a data.table with new columns based on the summarized values of rows.

Summary functions like `mean()`, `median()`, `min()`, `max()`, etc. can be used to summarize rows.

COMPUTE COLUMNS*



dt[, c := 1 + 2] – compute a column based on an expression.

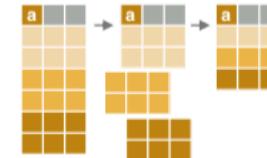


dt[a == 1, c := 1 + 2] – compute a column based on an expression but only for a subset of rows.



dt[, `:=` (c = 1, d = 2)] – compute multiple columns based on separate expressions.

Group according to by



dt[, j, by = .(a)] – group rows by values in specified columns.

dt[, j, keyby = .(a)] – group and simultaneously sort rows by values in specified columns.

COMMON GROUPED OPERATIONS

dt[, .(c = sum(b)), by = a] – summarize rows within groups.

dt[, c := sum(b), by = a] – create a new column and compute rows within groups.

dt[, .SD[1], by = a] – extract first row of groups.

dt[, .SD[N], by = a] – extract last row of groups.

Chaining

dt[...][...] – perform a sequence of data.table operations by chaining multiple “[]”.

Functions for data.tables

REORDER

MERCI !