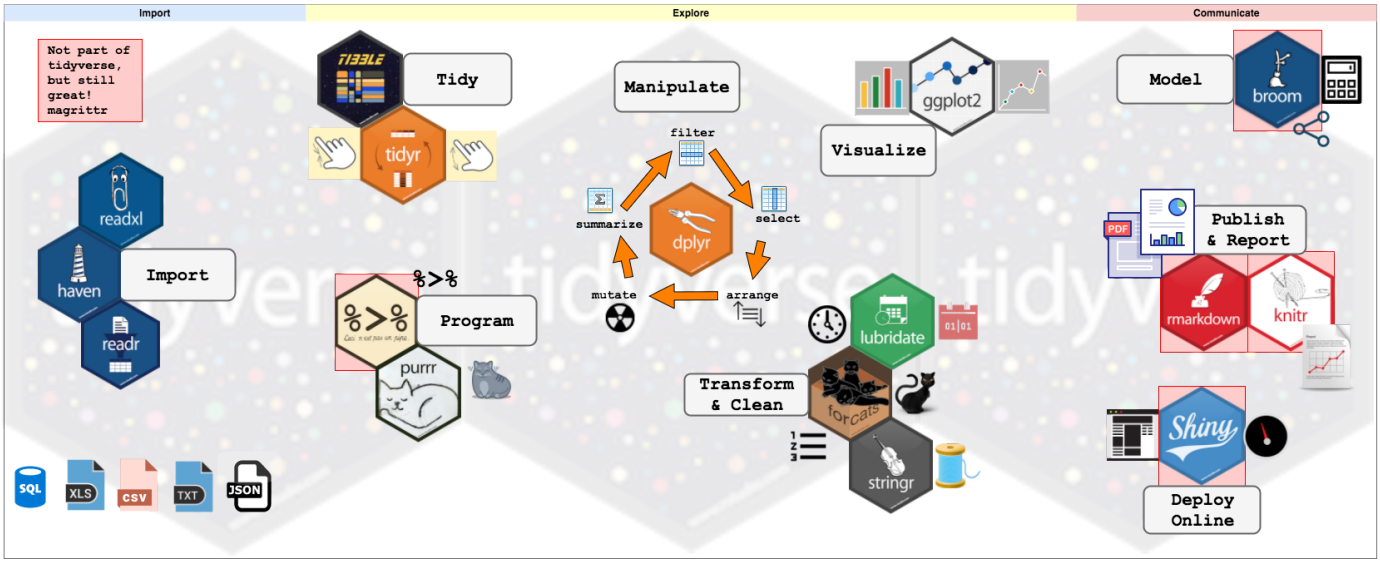
Chapitre 2 Le Tidyverse

Le tidyverse est un ensemble de packages proposant une syntaxe cohérente pour remplir l'essentiel des traitements propres à la science de la données, de la lecture des données à la valorisation en passant par la modélisation. Le manifeste du tidyverse comprend 4 principes clefs pour les packages du tidyverse :

- Utiliser les structures de données existantes : ne pas créer des objets ad hoc
- Utiliser l'opérateur pipe
- S'intègrer dans l'approche de programmation fonctionnelle de R
- Designé pour les être humains : favoriser la facilité d'usage à la performance machine



2.1 Présentation des packages

2.1.1 Des packages pour lire des données

2.1.1.1 tidyverse

- readr pour les fichiers plats
- readxl pour les fichiers tableur Excel
- haven pour les données stockées sous des formats propriétaires (SAS, SPSS, …)

2.1.1.2 Hors tidyverse

- odbc / Rposgresql pour accéder à des données stockées sous forme de base de données
- sf pour lire des données spatiales
- rsdmx pour lire des données sdmx

2.1.2 Des packages pour manipuler des données

2.1.2.1 tidyverse

- dplyr fonctions correspondant à des “verbes” pour manipuler ses données
- tidyr fonctions pour modifier l’agencement de nos tables entre les lignes et les colonnes

2.1.3 Des packages pour nettoyer des données

2.1.3.1 tidyverse

- forcats permet de manipuler les variables de type catégorielle (ou factor en R)
- stringr permet de manipuler des chaînes de caractères
- lubridate permet de manipuler des dates

2.1.3.2 Hors tidyverse

- RcppRoll qui regroupe des opérations fenêtrées ou glissantes

2.2 Activer les packages

```
library (dplyr)
library (tidyr)
library (forcats)
library (lubridate)
library (stringr)
library (RcppRoll)
library (DT)
library (readxl)
library (dbplyr)
library (RPostgreSQL)
library (rsdmx)
library (sf)
```

2.3 Les spécificités du tidyverse

Quelques spécificités des fonctions de ce package :

- Ces packages sont orientés manipulation de *dataframes* et non de *vecteurs*
- En conséquence, on utilise jamais l'indexation des colonnes de tables (le "$") pour appeler une variable
- Chaque fonction ne fait qu'une chose et une seule (c'est une opération élémentaire)
- L'ensemble des fonctions obéissent à la même logique, ce qui permet de simplifier l'apprentissage
- l'ensemble de ces opérations élémentaires peuvent s'enchaîner à la manière d'un ETL avec le pipe

2.4 D'autres approches possibles

Les fonctions que nous allons voir obéissent à une logique intégrée et simple, qui permet des manipulations complexes, à partir du moment ou l'on est capable d'identifier et de sérier chaque *opération élémentaire* à réaliser. D'autres packages permettent également de réaliser ce type de manipulations. La différence est qu'ils sont souvent dédiés à une tâche spécifique, ce qui rend la cohérence moins évidente lorsque l'on doit réaliser plusieurs opérations. Un autre package propose toutefois une vision intégrée de la sorte : data.table. Plusieurs différences sont à noter :

- *data.table* est plus rapide sur d'importants volumes de données, le code est très succinct.
- *dplyr* est plus simple à apprendre, le code est plus lisible, il peut s'appliquer à des formats de données multiples, il s'intègre dans un framework global qui va de la lecture des données (readr, readxl,haven…) à leur valorisation (ggplot2).