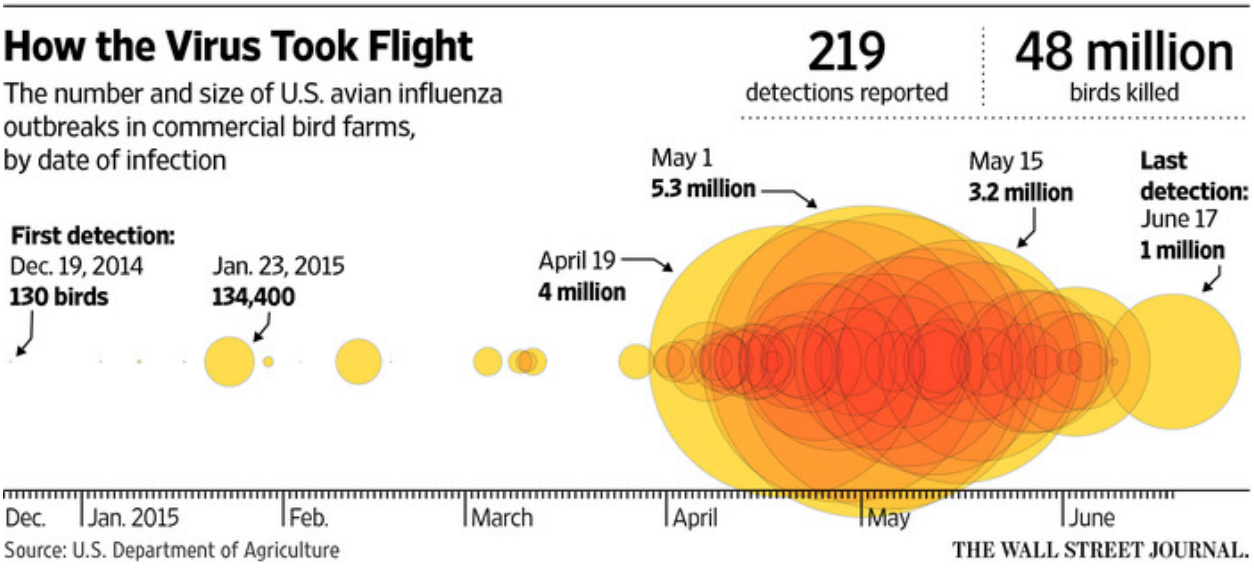
Bird Flu

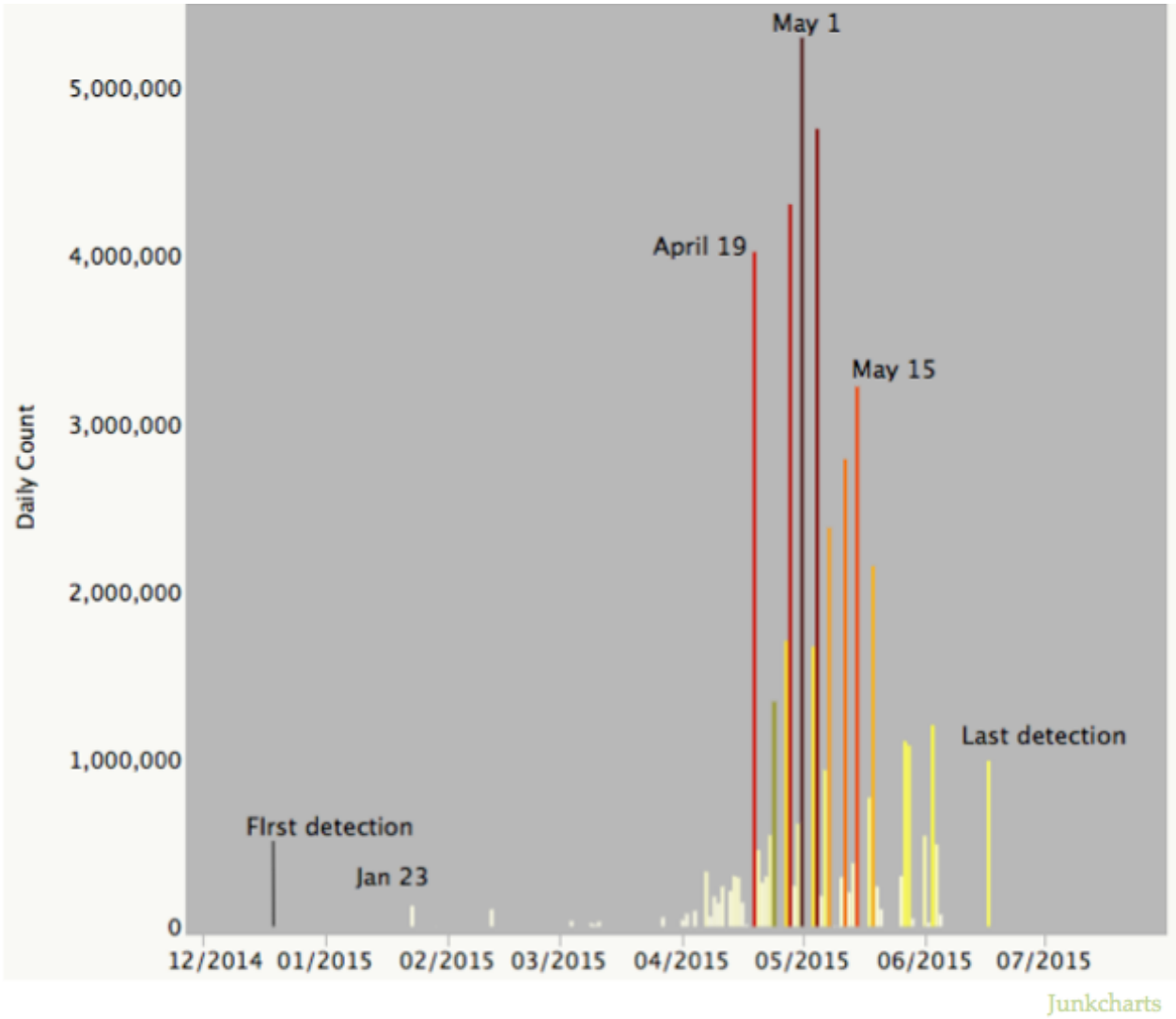
JcB

25/10/2015

Ce travail est inspiré de l'article hrbrmstr, lui même inspiré par un article du Wall Street Journal concernant la circulation de **virus aviaire pathogènes** dans les élevages US et illustré par un graphique accrocheur (Quite the eye-catching chart)



Le graphique est spectaculaire mais difficile à appréhender. L'objectif de l'auteur est de le transformer en barre graphe, moins tape à l'oeil mais plus informatif.



Reprendre ce travail m'a permis de:

- apprendre à extraire un tableau de données à partir d'une page HTML
- découvrir la librairy **viridis** (palette de couleur)
- de tansformer des formats de dates en modifiant la LOCALE de mon ordinateur
- de manipuler la librairie dplyr

Library nécessaires

```
library(xml2)
library(rvest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(ggplot2)
library(scales)
library(viridis)
library(ggthemes)
```

Récupérer les données source

Les données de base (raw data) sont issues de l'USDA (United States Department of Agriculture) à la page *Update on Avian Influenza Findings - Poultry Findings Confirmed by USDA's National Veterinary Services Laboratories* entre le 19/12/2014 (date de la première épidémie) et le 17/6/15.

```
# page internet où se trouve le tableau de données
pg <- read_html("https://www.aphis.usda.gov/wps/portal/aphis/ourfocus/animalhealth/sa_animal_disease_in

# pg est une liste de 2 éléments
names(pg)

# lecture de la table principale
dat <- html_table(html_nodes(pg, "table"))[[1]]

# sauvegarde
write.csv(dat, file = "dat.csv")

# dat est le dataframe correspondant au tableau 1
names(dat)
```

Les données brutes sont sauvegardées dans **dat.csv** pour économiser du temps

```
dat <- read.csv("dat.csv")
head(dat)
```

```
##   X     State    County      Flyway Flock.type         Species
## 1 1      Iowa    Wright Mississippi Commercial  Layer Chickens
## 2 2      Iowa     Sioux Mississippi   Backyard Mixed Game Fowl
## 3 3 Minnesota Kandiyohi Mississippi Commercial         Turkeys
## 4 4 Minnesota     Brown Mississippi Commercial         Turkeys
## 5 5      Iowa       Sac Mississippi Commercial         Turkeys
## 6 6 Minnesota  Renville Mississippi Commercial Pullet Chickens
##   Avian.influenza.subtype. Confirmation.date Flock.size
## 1               EA/AM-H5N2      Jun 17, 2015  1,000,000
## 2               EA/AM-H5N2       Jun 9, 2015      2,500
## 3               EA/AM-H5N2       Jun 5, 2015     44,000
## 4               EA/AM-H5N2       Jun 5, 2015     39,000
```

```
## 5                EA/AM-H5N2       Jun 4, 2015     42,200
## 6                EA/AM-H5N2       Jun 4, 2015    415,000
```

Les colonnes retenues sont:

- confirmation date: date de début de l'épidémie
- flock size : taille de l'élevage

Préparation des données

La colonne date est au format US avec un nom de mois abrégé. Les mois abrégés US ne sont pas reconnus par le Locale français. Il faut donc temporairement mettre locale au format US. Pour la colonne *flock size*, il faut supprimer le séparateur de milliers. Par ailleurs ertains chiffres sont remplacés par *pending*.

```
# sauvegarde des constantes locales
local_time <- Sys.getlocale(category = "LC_TIME")
local_time
```

```
## [1] "fr_FR.UTF-8"
```

```
# mise en place du système US
Sys.setlocale(category = "LC_TIME", locale = 'en_GB.UTF-8')
```

```
## [1] "en_GB.UTF-8"
```

```
Sys.getlocale(category = "LC_TIME")
```

```
## [1] "en_GB.UTF-8"
```

```
# dat$"Confirmation date" <- as.Date(dat$"Confirmation date", "%b %d, %Y")

# transformation des données en une passe
dat %>%
  mutate(`Confirmation.date` = as.Date(`Confirmation.date`, "%b %d, %Y"),
    week = format(`Confirmation.date`, "%Y-%U"),
    week_start = as.Date(sprintf("%s-1", week), "%Y-%U-%u") ,
    `Flock.size` = as.numeric(str_replace_all(`Flock.size`, ",", ""))) %>%
  select(week, week_start, `Flock.size`) %>%
  filter(!is.na(`Flock.size`)) %>%
  group_by(week_start) %>%
  summarize(outbreaks=n(),
    flock_total=sum(`Flock.size`)) -> dat
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduits lors de la
## conversion automatique
```

```
dat
```

```
## Source: local data frame [23 x 3]
##
##    week_start outbreaks flock_total
##        (date)     (int)       (dbl)
## 1  2014-12-15         1         130
## 2  2015-01-05         2         730
## 3  2015-01-12         2         140
## 4  2015-01-19         1      134400
## 5  2015-01-26         1        5830
## 6  2015-02-02         1          40
## 7  2015-02-09         1      112900
## 8  2015-02-16         1          70
## 9  2015-03-02         1       44000
## 10 2015-03-09         4       93130
## ..        ...       ...         ...
```

Suite

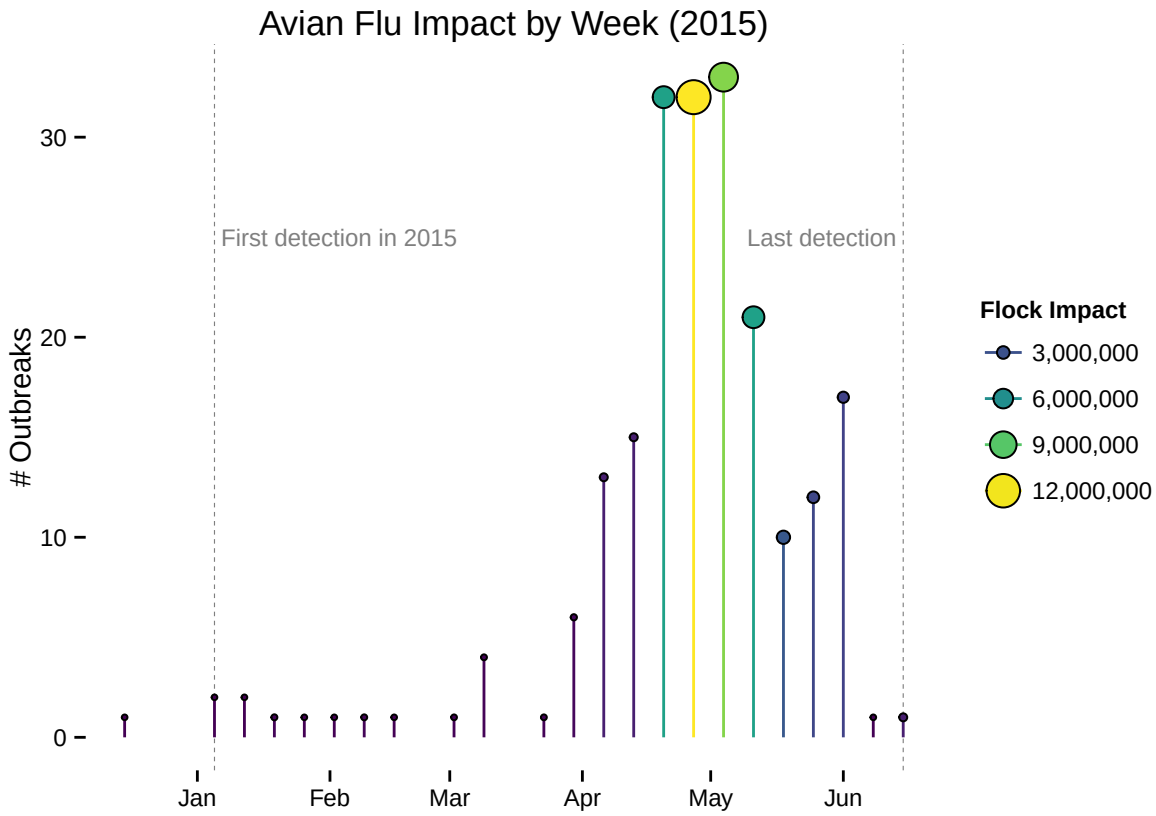
```
first <- dat[2,]
last <- tail(dat, 1)

first
```

```
## Source: local data frame [1 x 3]
##
##   week_start outbreaks flock_total
##       (date)     (int)       (dbl)
## 1 2015-01-05         2         730
```

```
last
```

```
## Source: local data frame [1 x 3]
##
##   week_start outbreaks flock_total
##       (date)     (int)       (dbl)
## 1 2015-06-15         1       1e+06
```

```
gg <- ggplot(dat, aes(x=week_start, y=outbreaks))
gg <- gg + geom_vline(xintercept=as.numeric(first$week_start), linetype="dashed", size=0.2, color="#7f7
gg <- gg + geom_text(data=first, aes(x=week_start, y=25), label=" First detection in 2015", hjust=0, si
gg <- gg + geom_vline(xintercept=as.numeric(last$week_start), linetype="dashed", size=0.2, color="#7f7f
gg <- gg + geom_text(data=last, aes(x=week_start, y=25), label="Last detection ", hjust=1, size=3, colo
gg <- gg + geom_segment(aes(x=week_start, xend=week_start, y=0, yend=outbreaks, color=flock_total), siz
gg <- gg + geom_point(aes(size=flock_total, fill=flock_total), shape=21)
gg <- gg + scale_size_continuous(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_color_viridis(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_fill_viridis(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_x_date(label=date_format("%b"))
gg <- gg + guides(color=guide_legend(), fill=guide_legend(), size=guide_legend())
gg <- gg + labs(x=NULL, y="# Outbreaks", title="Avian Flu Impact by Week (2015)")
gg <- gg + theme_tufte(base_family="Helvetica")
gg <- gg + theme(legend.key=element_rect(color=rgb(0,0,0,0)))
gg
```



If we really want to see the discrete events, we can do that with our less-ZOMGOSH color scheme, too:

```
# dat <- html_table(html_nodes(pg, "table"))[[1]]
dat <- read.csv("dat.csv")

dat %>%
  mutate(`Confirmation.date` = as.Date(`Confirmation.date`, "%b %d, %Y"),
         `Flock.size` = as.numeric(str_replace_all(`Flock.size`, ",", ""))) %>%
  filter(!is.na(`Flock.size`)) %>%
  rename(date=`Confirmation.date`) %>%
  arrange(date) -> dat
```

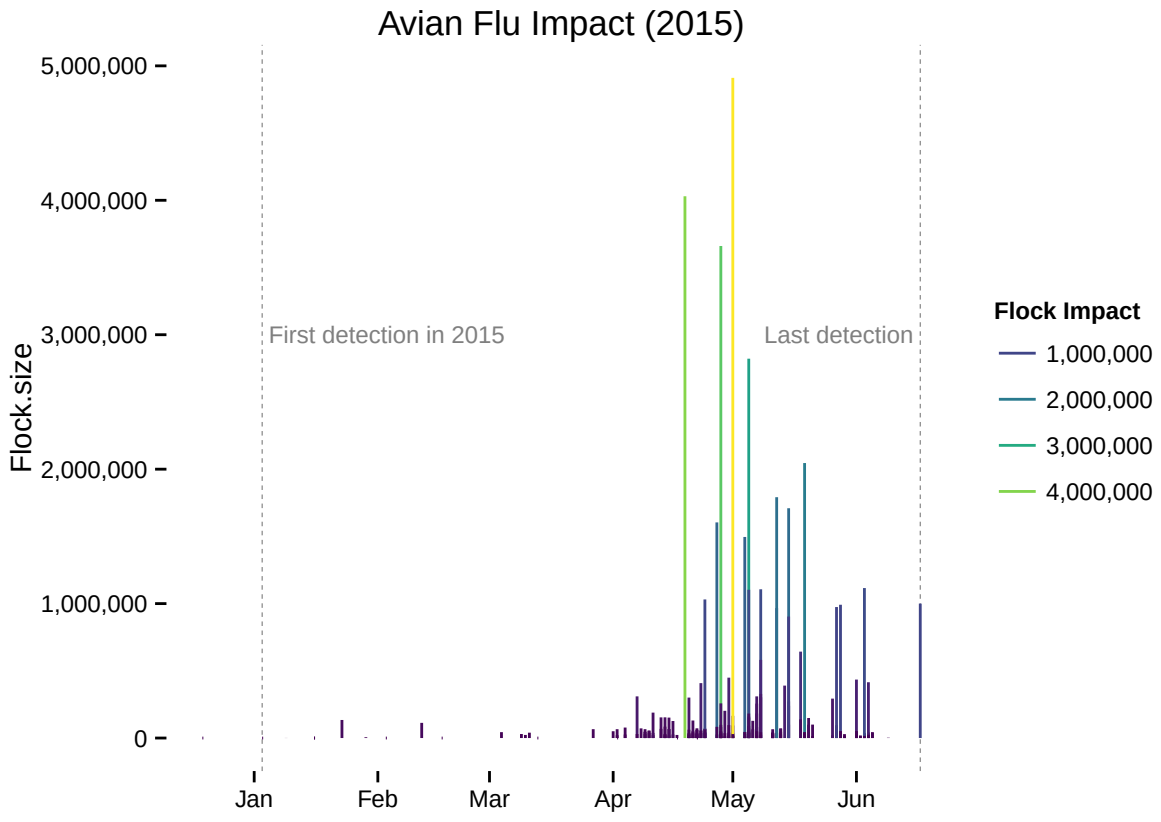
```
## Warning in eval(substitute(expr), envir, enclos): NAs introduits lors de la
## conversion automatique
```

```
head(dat)
```

```
##     X      State     County  Flyway Flock.type       Species
## 1 219     Oregon    Douglas Pacific   Backyard Mixed poultry
## 2 218 Washington     Benton Pacific   Backyard Mixed poultry
## 3 217 Washington     Benton Pacific   Backyard Mixed Poultry
## 4 215 Washington    Clallam Pacific   Backyard Mixed Poultry
## 5 216      Idaho     Canyon Pacific   Backyard Mixed poultry
## 6 214 California Stanislaus Pacific Commercial       Turkeys
##   Avian.influenza.subtype.       date Flock.size
## 1                EA -H5N8 2014-12-19        130
## 2              EA/AM-H5N2 2015-01-03        140
```

```
## 3               EA/AM-H5N2 2015-01-09        590
## 4               EA/AM-H5N2 2015-01-16        110
## 5               EA/AM-H5N2 2015-01-16         30
## 6                  EA-H5N8 2015-01-23     134400
```

```
first <- dat[2,]
last <- tail(dat, 1)

gg <- ggplot(dat, aes(x=date, y=`Flock.size`))
gg <- gg + geom_vline(xintercept=as.numeric(first$date), linetype="dashed", size=0.2, color="#7f7f7f")
gg <- gg + geom_text(data=first, aes(x=date, y=3000000), label=" First detection in 2015", hjust=0, siz
gg <- gg + geom_vline(xintercept=as.numeric(last$date), linetype="dashed", size=0.2, color="#7f7f7f")
gg <- gg + geom_text(data=last, aes(x=date, y=3000000), label="Last detection ", hjust=1, size=3, color=
gg <- gg + geom_segment(aes(x=date, xend=date, y=0, yend=`Flock.size`, color=`Flock.size`), size=0.5, al
gg <- gg + scale_size_continuous(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_color_viridis(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_fill_viridis(name="Flock Impact", label=comma, guide="legend")
gg <- gg + scale_x_date(label=date_format("%b"))
gg <- gg + scale_y_continuous(label=comma)
gg <- gg + guides(color=guide_legend(), fill=guide_legend(), size=guide_legend())
gg <- gg + labs(x=NULL, y="Flock.size", title="Avian Flu Impact (2015)")
gg <- gg + theme_tufte(base_family="Helvetica")
gg <- gg + theme(legend.key=element_rect(color=rgb(0,0,0,0)))
gg
```



Restauration

```r
# restauration du système français
Sys.setlocale(category = "LC_TIME", locale = local_time)
```

```
## [1] "fr_FR.UTF-8"
```