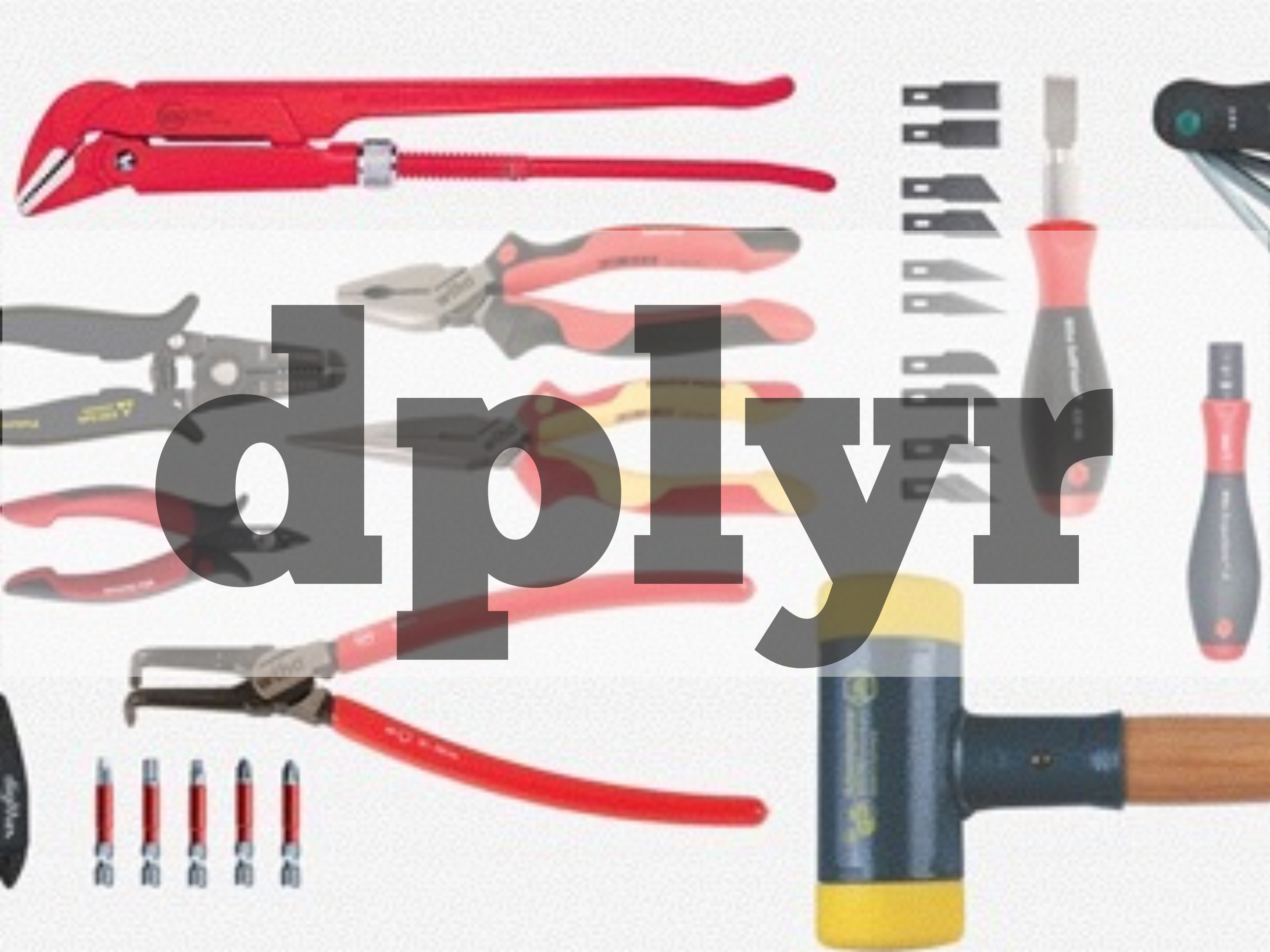
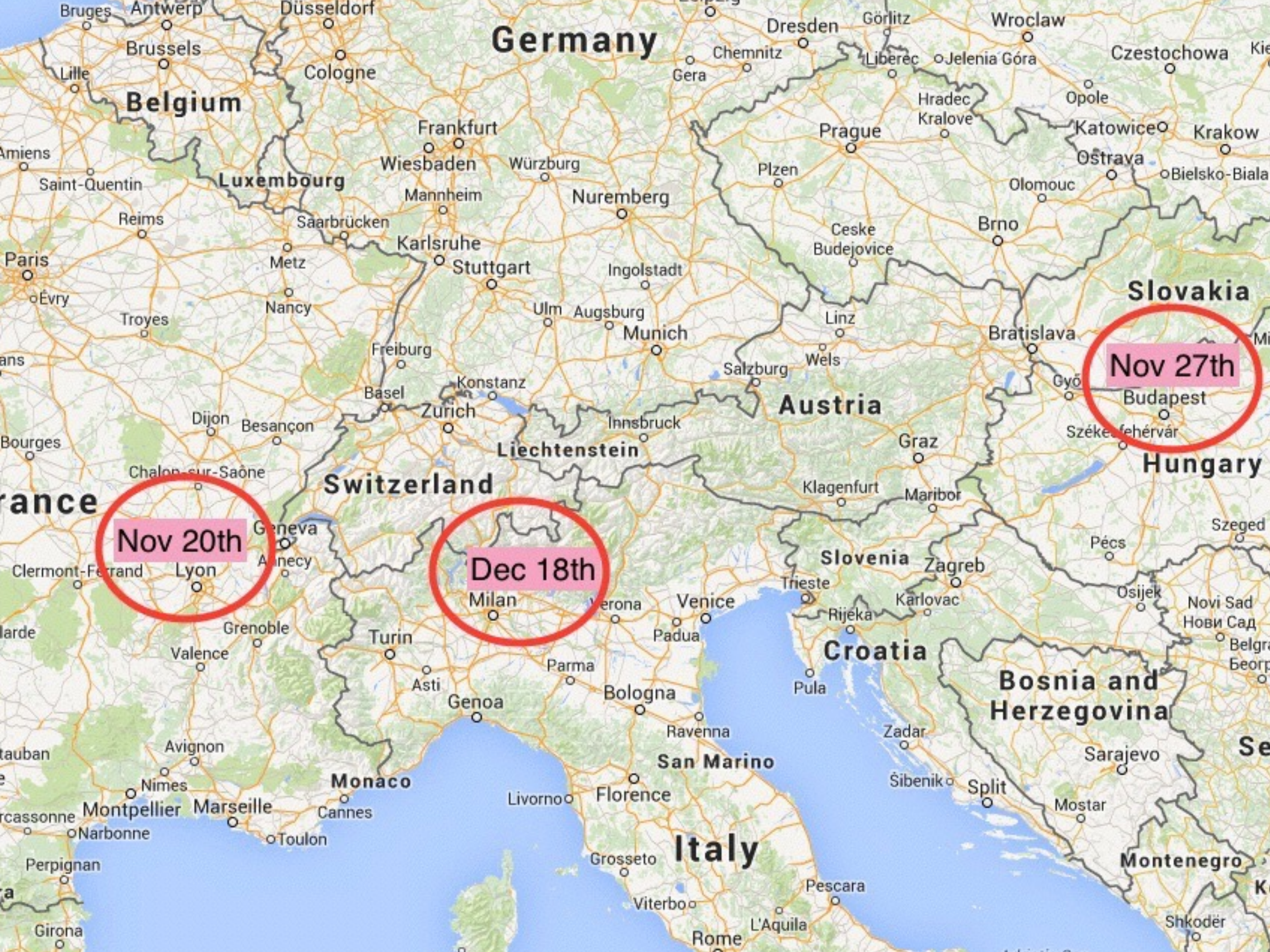



Germany
Belgium
Luxembourg
Slovakia
Austria
Hungary
Switzerland
Liechtenstein
Slovenia
Croatia
Bosnia and Herzegovina
Montenegro
Italy
San Marino
Monaco
Nov 27th
Nov 20th
Dec 18th
Bruges
Antwerp
Düsseldorf
Brussels
Cologne
Lille
Dresden
Görlitz
Wroclaw
Chemnitz
Gera
Liberec
Jelenia Góra
Czestochowa
Hradec Kralove
Opole
Prague
Katowice
Krakow
Ostrava
Plzen
Olomouc
Bielsko-Biala
Frankfurt
Wiesbaden
Würzburg
Mannheim
Nuremberg
Saarbrücken
Karlsruhe
Stuttgart
Ceske Budejovice
Brno
Ingolstadt
Ulm
Augsburg
Munich
Linz
Bratislava
Wels
Salzburg
Freiburg
Konstanz
Basel
Zurich
Innsbruck
Graz
Budapest
Székesfehérvár
Klagenfurt
Maribor
Zagreb
Trieste
Rijeka
Karlovac
Osijek
Novi Sad
Pécs
Szeged
Venice
Padua
Verona
Milan
Turin
Asti
Genoa
Parma
Bologna
Ravenna
Pula
Zadar
Sarajevo
Šibenik
Split
Mostar
Florence
Livorno
Grosseto
Viterbo
Pescara
L'Aquila
Rome
Shkodër
Amiens
Saint-Quentin
Reims
Paris
Évry
Metz
Nancy
Troyes
Dijon
Besançon
Bourges
Chalon-sur-Saône
Geneva
Annecy
Clermont-Ferrand
Lyon
Grenoble
Valence
Avignon
Nimes
Montpellier
Marseille
Cannes
Toulon
Narbonne
Perpignan
Girona

dplyr dplyr
dplyr
dplyr
dplyr
dplyr
dplyr dplyr
dplyr dplyr
dplyr dplyr
dplyr
dplyr dplyr dplyr dplyr
dplyr dplyr dplyr
dplyr dplyr
dplyr
dplyr
dplyr
dplyr dplyr dplyr
dplyr
dplyr
dplyr
dplyr
dplyr
dplyr
dplyr dplyr
dplyr
dplyr
dplyr
dplyr
dplyr
dplyr dplyr
dplyr
dplyr
dplyr
dplyr dplyr
dplyr #rcatladies
dplyr
dplyr
dplyr dplyr
dplyr
dplyr
dplyr
dplyr dplyr
dplyr dplyr
dplyr dplyr
dplyr
dplyr dplyr dplyr dplyr dplyr
dplyr
dplyr dplyr
dplyr
dplyr dplyr
dplyr dplyr

verb

function that takes a data frame as its first argument

Examples of R verbs

head, tail, …

verb subject …

```
> head( iris, n = 4 )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
```

%>%

from `magrittr`

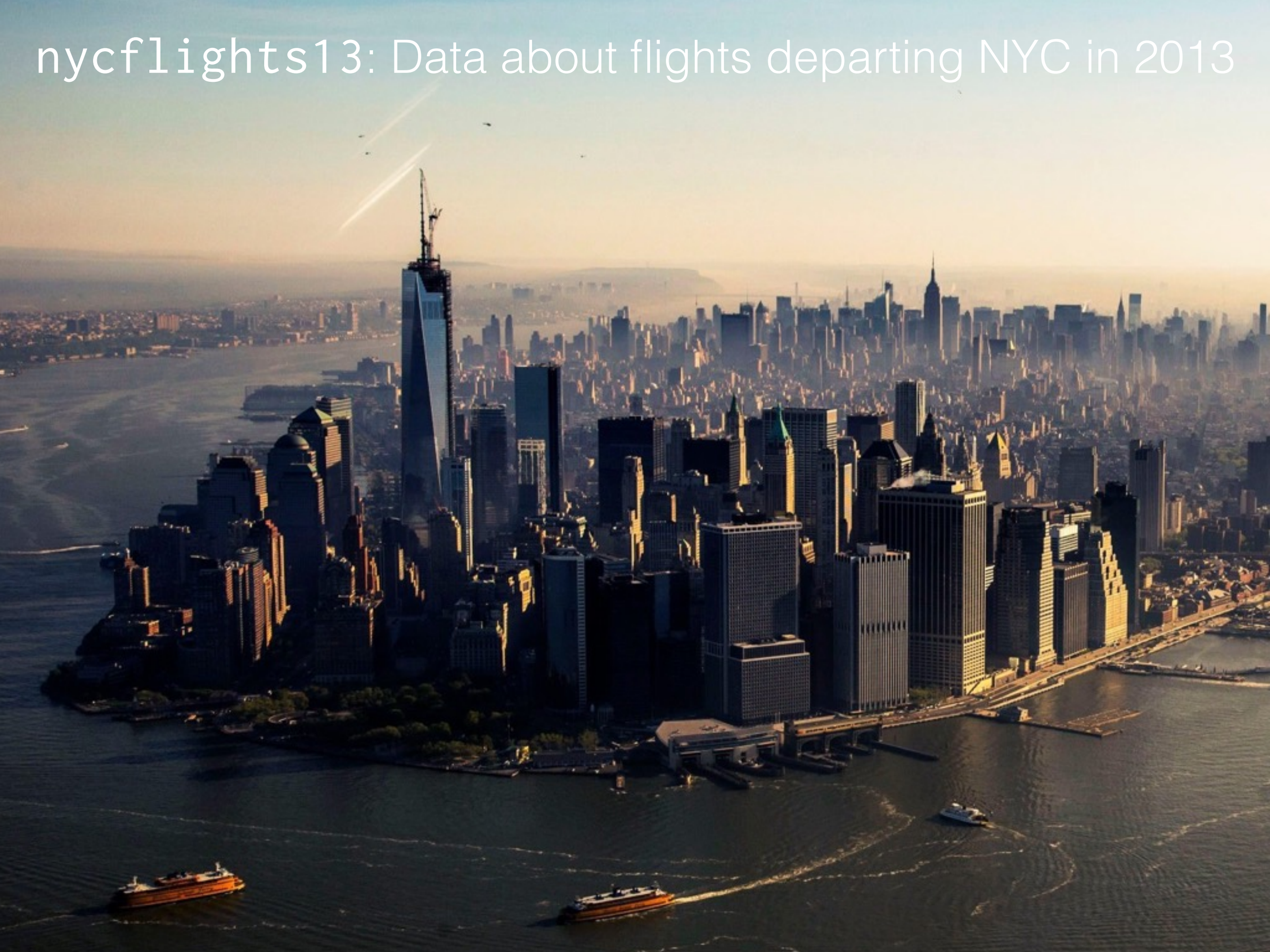
Classic R code

```
mean( rnorm( 100, mean = 4, sd = 4), trim = .1 )
```

Pipeline R code with %>%

```
100 %>%
  rnorm( mean = 4, sd = 4) %>%
  mean( trim = .1 )
```

`nycflights13`: Data about flights departing NYC in 2013



```
> library("nycflights13")
> flights
Source: local data frame [336,776 x 16]

   year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
1  2013     1   1      517         2      830        11      UA  N14228   1545
2  2013     1   1      533         4      850        20      UA  N24211   1714
3  2013     1   1      542         2      923        33      AA  N619AA   1141
4  2013     1   1      544        -1     1004       -18      B6  N804JB    725
5  2013     1   1      554        -6      812       -25      DL  N668DN    461
6  2013     1   1      554        -4      740        12      UA  N39463   1696
7  2013     1   1      555        -5      913        19      B6  N516JB    507
8  2013     1   1      557        -3      709       -14      EV  N829AS   5708
9  2013     1   1      557        -3      838        -8      B6  N593JB     79
10 2013     1   1      558        -2      753         8      AA  N3ALAA    301
..  ...   ... ...      ...       ...      ...       ...     ...     ...    ...
Variables not shown: origin (chr), dest (chr), air_time (dbl), distance (dbl),
  hour (dbl), minute (dbl)
```

tbl_df

A data frame that does print all of itself by default

```
> data <- tbl_df(mtcars)
> data
Source: local data frame [32 x 11]

    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
5  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
6  18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
7  14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
8  24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
9  22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
10 19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
..  ...  ...   ...  ...  ...   ...   ... .. ..  ...  ...
```

filter

A subset of the rows of the data frame

```
filter(flights, month == 1, day == 1)

flights %>%
  filter( dep_delay < 10 )


flights %>%
  filter( arr_delay < dep_delay )


flights %>%
     filter( hour < 12, arr_delay <= 0 )
```

arrange

reorder a data frame

```
flights %>%
  filter( hour < 8 ) %>%
  arrange( year, month, day )
```

```
flights %>%
  arrange( desc(dep_delay) )
```

select

select certain columns from the data frame

```r
# Select columns by name
select(flights, year, month, day)

# Select all columns between year and day
select(flights, year:day)

# Select all columns except those from year to
# day (inclusive)
select(flights, -(year:day))
```

mutate

modify or create columns based on others

```
d <- flights %>%
  mutate(
    gain = arr_delay - dep_delay,
    speed = distance / air_time * 60
  ) %>%
  filter( gain > 0 ) %>%
  arrange( desc(speed) )

d %>%
  select( year, month, day, dest, gain, speed )
```

summarise

collapse a data frame into one row …

```
summarise(flights,
  delay = mean(dep_delay, na.rm = TRUE))
```

```
flights %>%
  filter( dep_delay > 0 ) %>%
  summarise(arr_delay = mean(arr_delay, na.rm = TRUE))
```

group_by

Group observations by one or more variables

```
flights %>%
  group_by( tailnum ) %>%
  summarise(
    count = n(),
    dist = mean(distance, na.rm = TRUE),
    delay = mean(arr_delay, na.rm = TRUE)
  ) %>%
  filter( is.finite(delay) ) %>%
  arrange( desc(count) )
```

```
flights %>%
  group_by(dest) %>%
  summarise(
  planes = n_distinct(tailnum),
  flights = n()
) %>%
arrange( desc(flights) )
```

joins

joining two data frames

inner_join

all rows from x where there are matching values in y, and all columns from x and y. If there are multiple matches between x and y, all combination of the matches are returned.

```
destinations <- flights %>%
  group_by(dest) %>%
  summarise(
    planes = n_distinct(tailnum),
    flights = n()
  ) %>%
  arrange( desc(flights) ) %>%
  rename( faa = dest )

inner_join( destinations, airports, by = "faa")
```

inner_join

all rows from x where there are matching values in y, and all columns from x and y. If there are multiple matches between x and y, all combination of the matches are returned.

```
destinations <- flights %>%
  group_by(dest) %>%
  summarise(
    planes = n_distinct(tailnum),
    flights = n()
  ) %>%
  arrange( desc(flights) )

inner_join( destinations, airports,
  by = c( "dest" = "faa" ) )
```

other joins
See ?join

- `left_join, right_join`
- `inner_join, outer_join`
- `semi_join`
- `anti_join`

dplyr %>% summary

- Simple verbs: filter, mutate, select, summarise, arrange
- Grouping with group_by
- Joins with *_join
- Convenient with %>%
- F🛫ST

dplyr

Romain François
@romain_francois
romain@r-enthusiasts.com