INTRODUCTION TO R

PIETER PROVOOST
OBIS DATA MANAGER

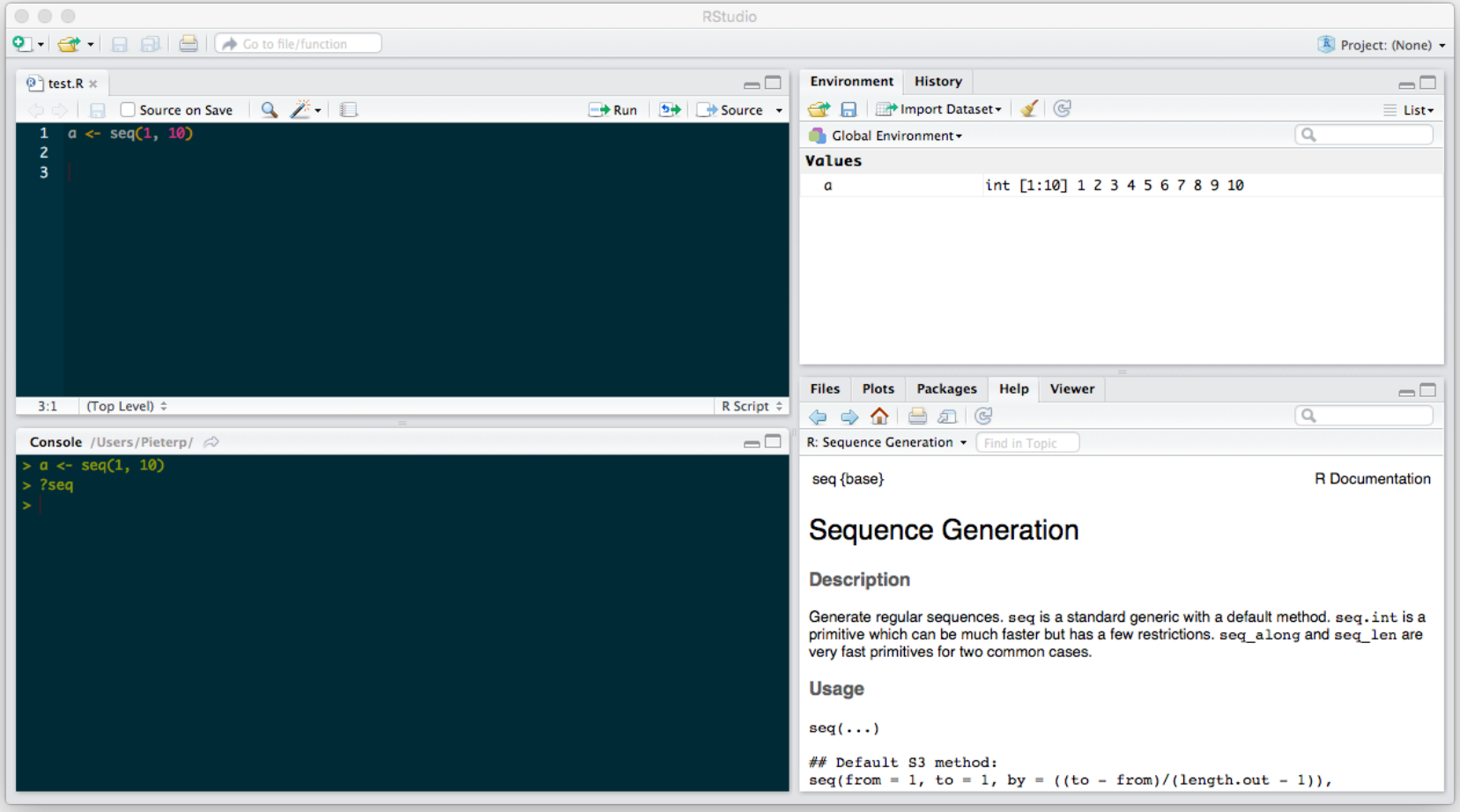
OBIS
OCEAN BIOGEOGRAPHIC
INFORMATION SYSTEM

RSTUDIO

- Install R from https://www.r-project.org
- Install RStudio from https://www.rstudio.com

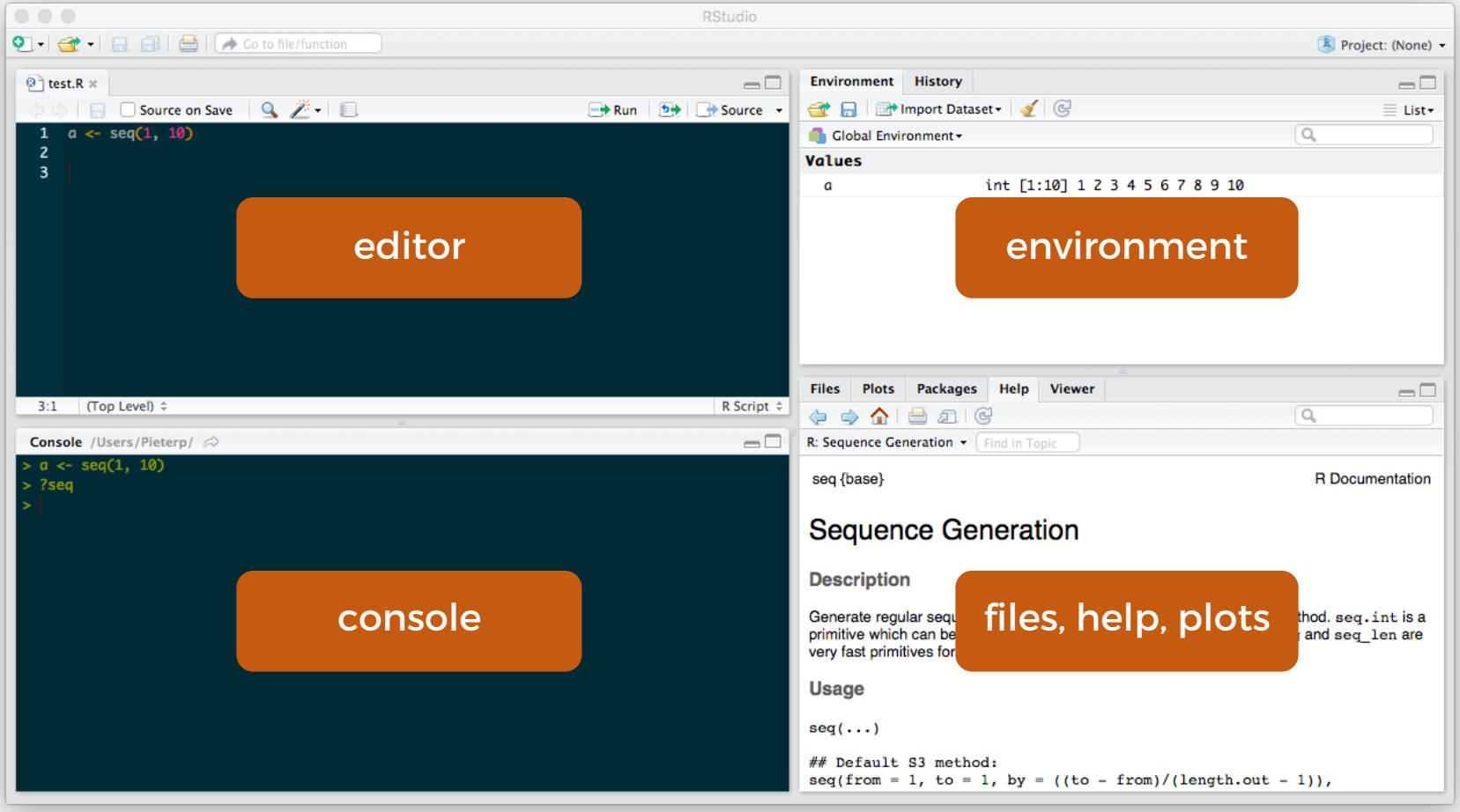
OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

RSTUDIO



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

RSTUDIO



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

VECTORS

- Most basic data structure
- Values of different classes such as numeric, character or logical

```
> a <- 1
> a
[1] 1
> class(a)
[1] "numeric"
> length(a)
[1] 1

> b <- "banana"
> b
[1] "banana"
> class(b)
[1] "character"
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

VECTORS

```
> a <- c(1, 2)
> a
[1] 1 2

> b <- seq(1, 10)
> b
[1]  1  2  3  4  5  6  7  8  9 10
> length(b)
[1] 10
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

DATA FRAMES

- Tabular data structures with columns of different classes

```
> d <- data.frame(a = c(1, 2, 3), b = c("x", "y", "z"))
> d
  a b
1 1 x
2 2 y
3 3 z
> d$a
[1] 1 2 3
> d[1]
  a
1 1
2 2
3 3
> d[1,]
  a b
1 1 x
> d[,1]
[1] 1 2 3
```

LISTS

- Collections of objects

```
> a <- data.frame(a = c(1, 2, 3), b = c("x", "y", "z"))
> l <- list(a = a, b = 1)
> l
$a
  a b
1 1 x
2 2 y
3 3 z

$b
[1] 1
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

LISTS

- Collections of objects

```
> l$a
  a b
1 1 x
2 2 y
3 3 z
> l[[1]]
  a b
1 1 x
2 2 y
3 3 z
> l[["a"]]
  a b
1 1 x
2 2 y
3 3 z
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

READING DATA

- Delimited text files

```
data <- read.table("data.txt", header = TRUE, sep = "\t",
                           dec = ".", stringsAsFactors = FALSE)
data <- read.csv("data.csv")
```

- Excel files

```
require(xlsx)

data <- read.xlsx("data.xlsx", 1)
data <- read.xlsx("data.xlsx", sheetName = "somesheet")
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

INSTALLING PACKAGES

- From CRAN

```
install.packages("dplyr")
```

- From GitHub

```
install.packages("devtools")
devtools::install_github("iobis/robis")
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

DATA EXPLORATION

```
> data <- occurrence("Abra")
Retrieved 49319 records of 49319 (100%)
> head(data)
      id decimalLongitude decimalLatitude depth           eventDate
1  56021        -80.65300        32.35300   6.3 2000-06-20 10:00:00
2  77027        -80.86700        32.16200   2.5 2001-07-10 10:00:00
3 118357        -90.16278        29.04722  12.0 1978-05-27 10:00:00
4 121952        -88.81600        30.31400   2.0 1992-08-23 10:00:00
5 136799        -81.45900        31.00000   1.2 2002-07-17 10:00:00
6 141960        -84.23700        30.01000   5.0 1993-07-23 10:00:00
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

DATA EXPLORATION

```
> dim(data)
[1] 49319    67
> names(data)
[1] "id"                    "decimalLongitude"
[3] "decimalLatitude"       "depth"
[5] "eventDate"             "institutionCode"
[7] "collectionCode"        "catalogNumber"
[9] "datasetName"           "phylum"
[11] "order"                 "family"
[13] "genus"                 "scientificName"
[15] "originalScientificName"  "scientificNameAuthorship"
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

DATA EXPLORATION

```
> summary(data)
       id              decimalLongitude   decimalLatitude       depth
Min.   :    56021   Min.   :-178.950   Min.   :-45.28   Min.   :  -1.36
1st Qu.:344350030   1st Qu.:   2.552   1st Qu.: 54.15   1st Qu.:  11.79
Median :345705350   Median :   9.912   Median : 55.67   Median :  15.00
Mean   :336837348   Mean   :   4.862   Mean   : 54.08   Mean   :  44.14
3rd Qu.:346089380   3rd Qu.:  10.534   3rd Qu.: 56.37   3rd Qu.:  21.90
Max.   :364573132   Max.   : 178.633   Max.   : 74.65   Max.   :5413.00
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

DATA EXPLORATION

data

Filter

	id	decimalLongitude	decimalLatitude	depth	eventDate	institutionCode
1	56021	-80.65300	32.35300	6.30	2000-06-20 10:00:00	EMAP_NCA
2	77027	-80.86700	32.16200	2.50	2001-07-10 10:00:00	EMAP_NCA
3	118357	-90.16278	29.04722	12.00	1978-05-27 10:00:00	USNM
4	121952	-88.81600	30.31400	2.00	1992-08-23 10:00:00	EMAP_NCA
5	136799	-81.45900	31.00000	1.20	2002-07-17 10:00:00	EMAP_NCA
6	141960	-84.23700	30.01000	5.00	1993-07-23 10:00:00	EMAP_NCA
7	235193	109.00000	21.40000	*NA*	*NA*	CASMBM
8	239473	-72.80000	39.23333	90.00	1977-03-24 11:00:00	USNM
9	242801	-87.99400	30.36000	3.00	1992-07-29 10:00:00	EMAP_NCA
10	259690	-73.40500	38.71000	75.00	1976-08-18 11:00:00	USNM
11	278804	-80.82900	32.13200	14.80	2002-08-06 10:00:00	EMAP_NCA

Showing 1 to 12 of 49,319 entries

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MANIPULATING DATA

- **dplyr** package
  - http://genomicsclass.github.io/book/pages/dplyr_tutorial.html
- Filtering (**filter**)

```
require(robis)
require(dplyr)

data <- occurrence("Abra")
data %>% filter(scientificName == "Abra alba" & yearcollected > 2005)
```

- Reordering (**arrange**)

```
data %>% arrange(datasetName, desc(eventDate))
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MANIPULATING DATA

- Selecting and renaming columns (**`select`**)

```
data %>% select(scientificName, eventDate,
                    lon = decimalLongitude, lat = decimalLatitude)
```

- Finding distinct combinations (**`distinct`**)

```
data %>% select(scientificName, locality) %>% distinct()
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MANIPULATING DATA

- Adding columns (**`mutate`**)

```
data %>% mutate(
             zone = .bincode(
                minimumDepthInMeters,
                breaks = c(0, 10, 100)
             )
         )
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

AGGREGATING DATA

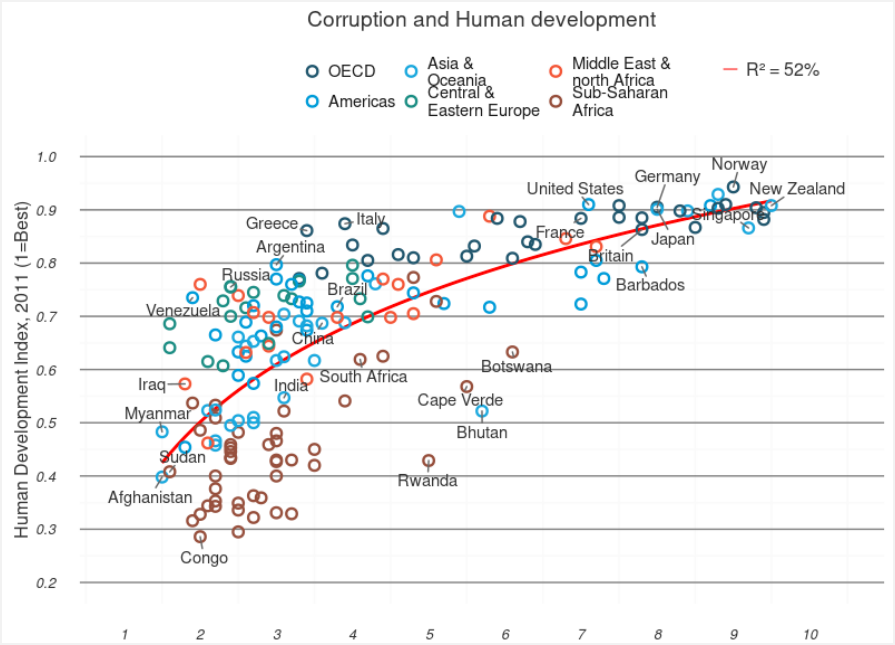
- **group_by, summarise**

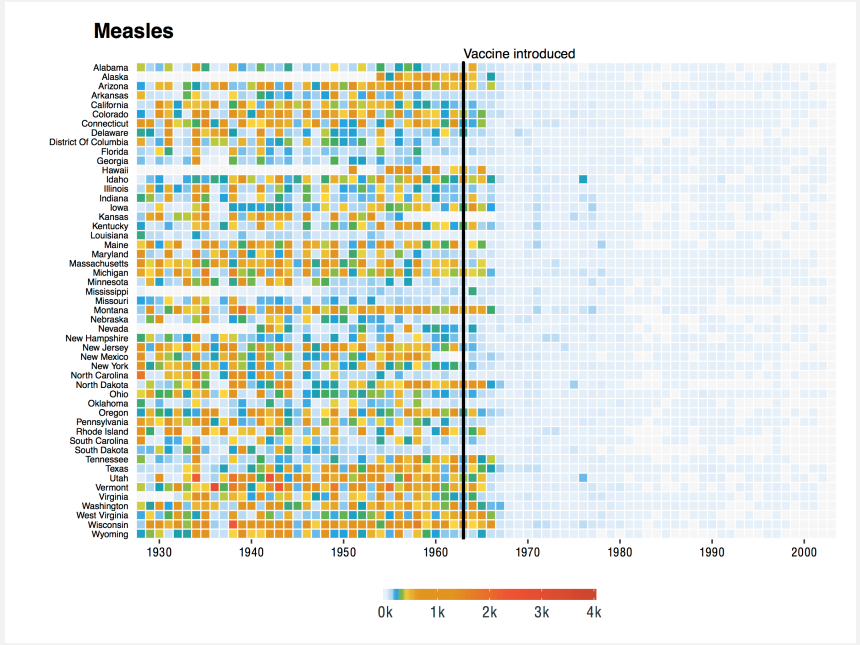
```
> data %>% summarise(lat_mean = mean(decimalLatitude),
                     lat_sd = sd(decimalLatitude))
  lat_mean  lat_sd
1 54.07561 7.00586

> data %>%
    group_by(scientificName) %>%
    summarise(records = n(),
              datasets = n_distinct(datasetName))
      scientificName records datasets
                <chr>   <int>    <int>
1              Abra    1544       40
2      Abra aequalis     693       15
3          Abra alba   33188       85
```

OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

PLOTTING

- **`ggplot2`** package
  - http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html
  - http://www.r-graph-gallery.com/all-graphs

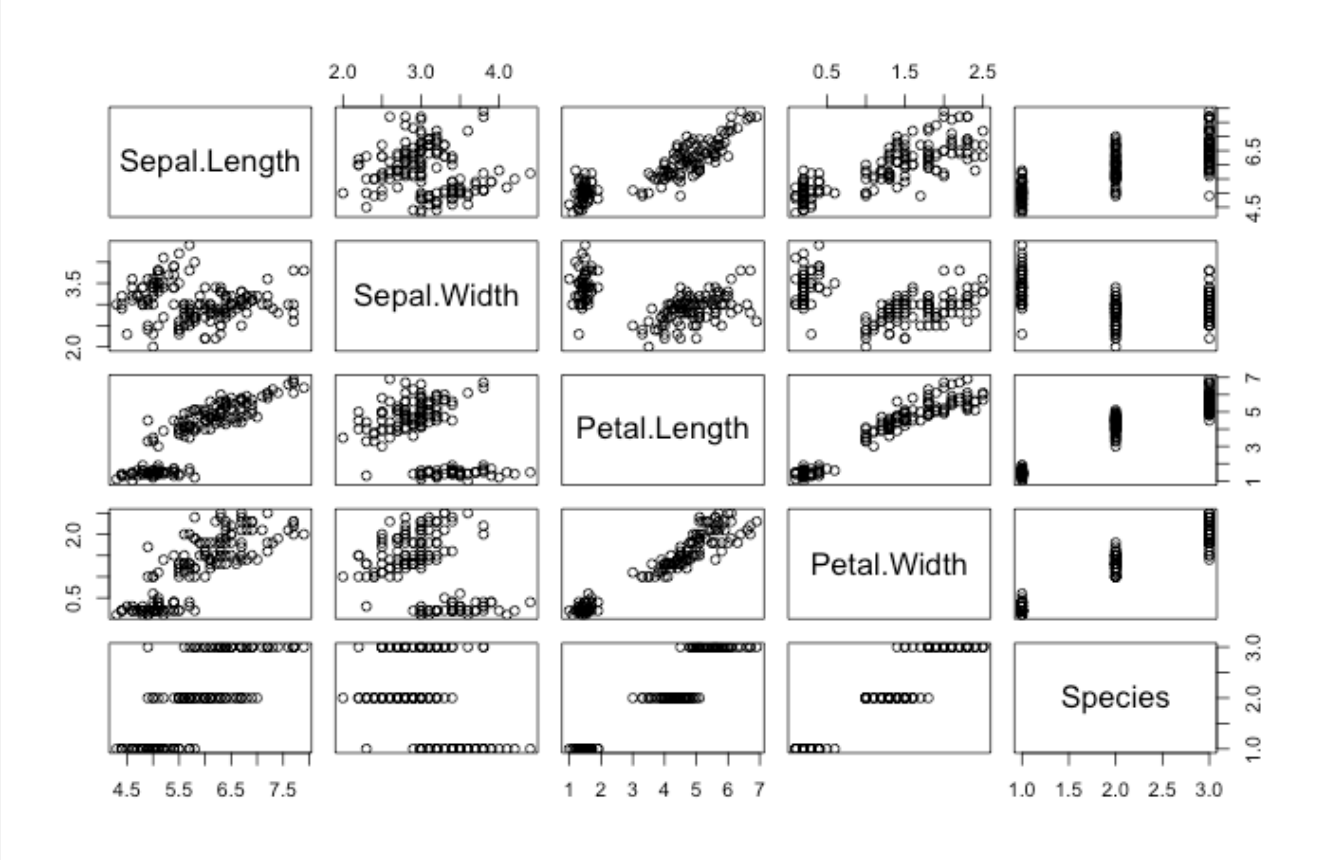




OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

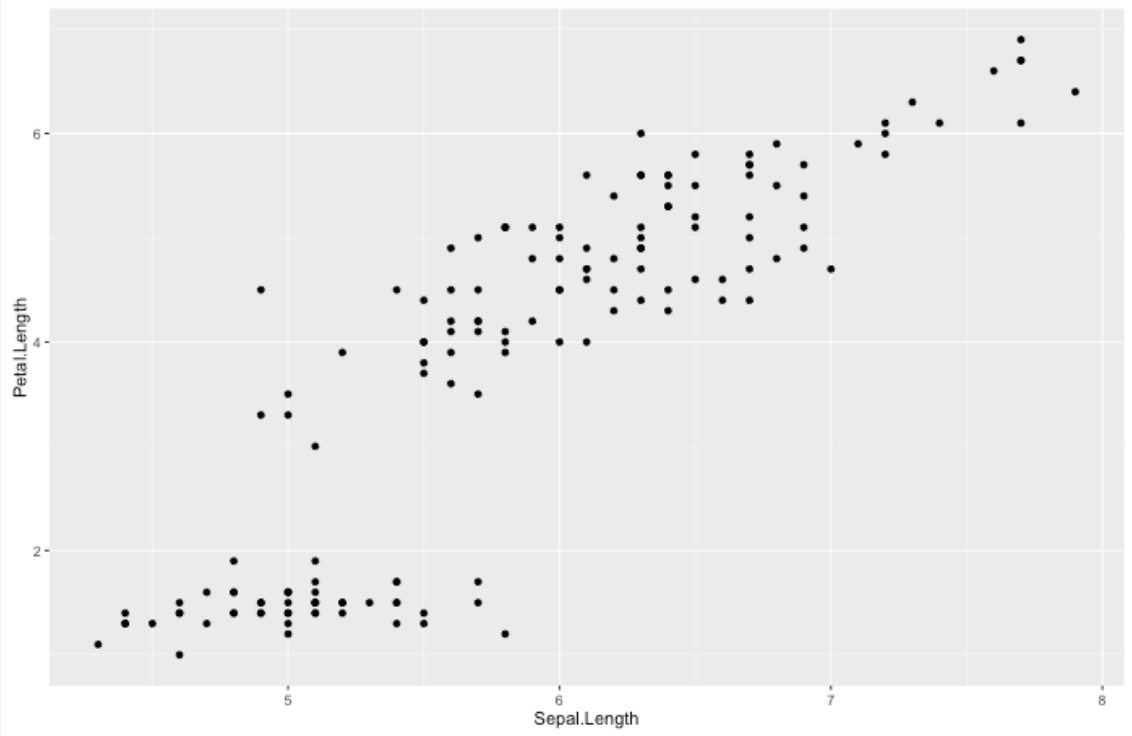
PLOTTING

```
plot(iris)
```



PLOTTING

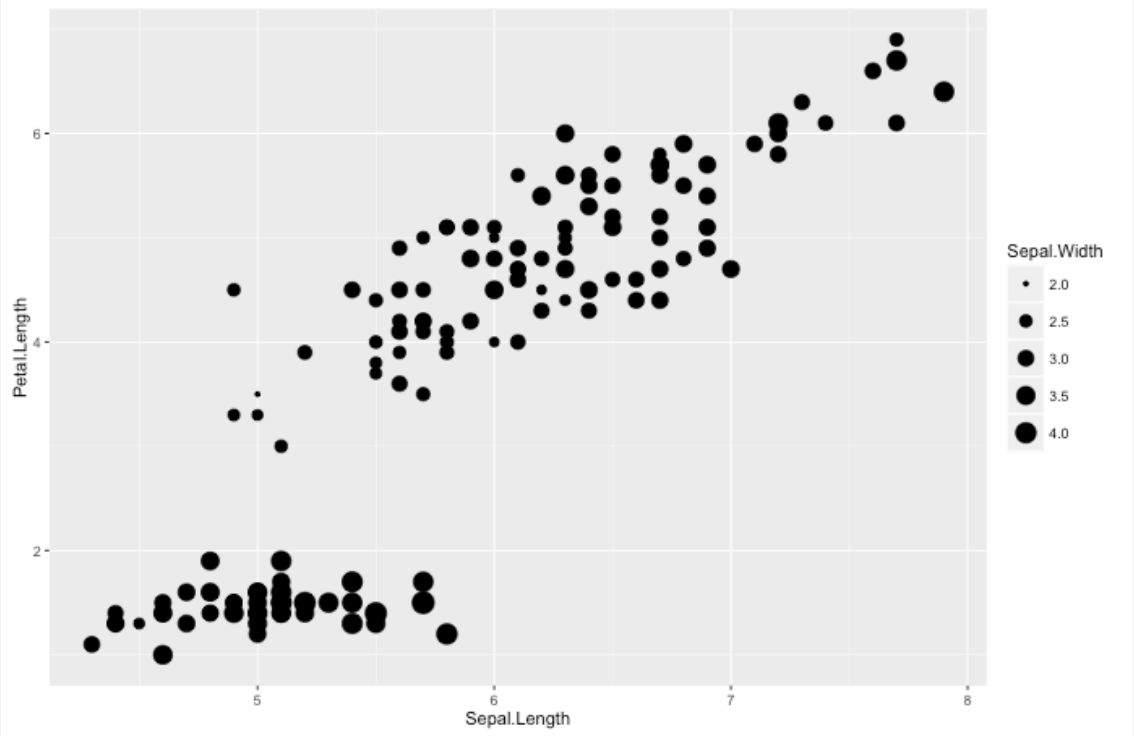
```
ggplot(data = iris) +
    geom_point(aes(x = Sepal.Length,
                   y = Petal.Length))
```



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

PLOTTING

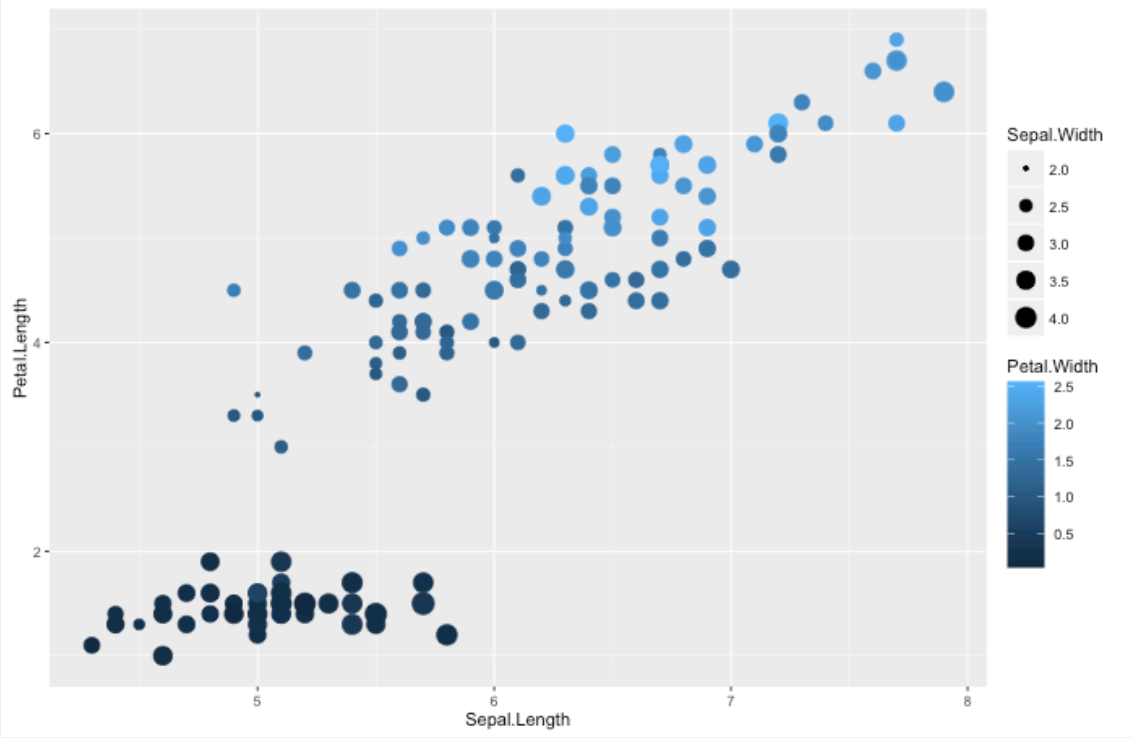
```
ggplot(data = iris) +
    geom_point(aes(x = Sepal.Length,
                   y = Petal.Length,
                   size = Sepal.Width))
```



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

PLOTTING

```
ggplot(data = iris) +
    geom_point(aes(x = Sepal.Length,
                   y = Petal.Length,
                   size = Sepal.Width,
                   colour = Petal.Width))
```



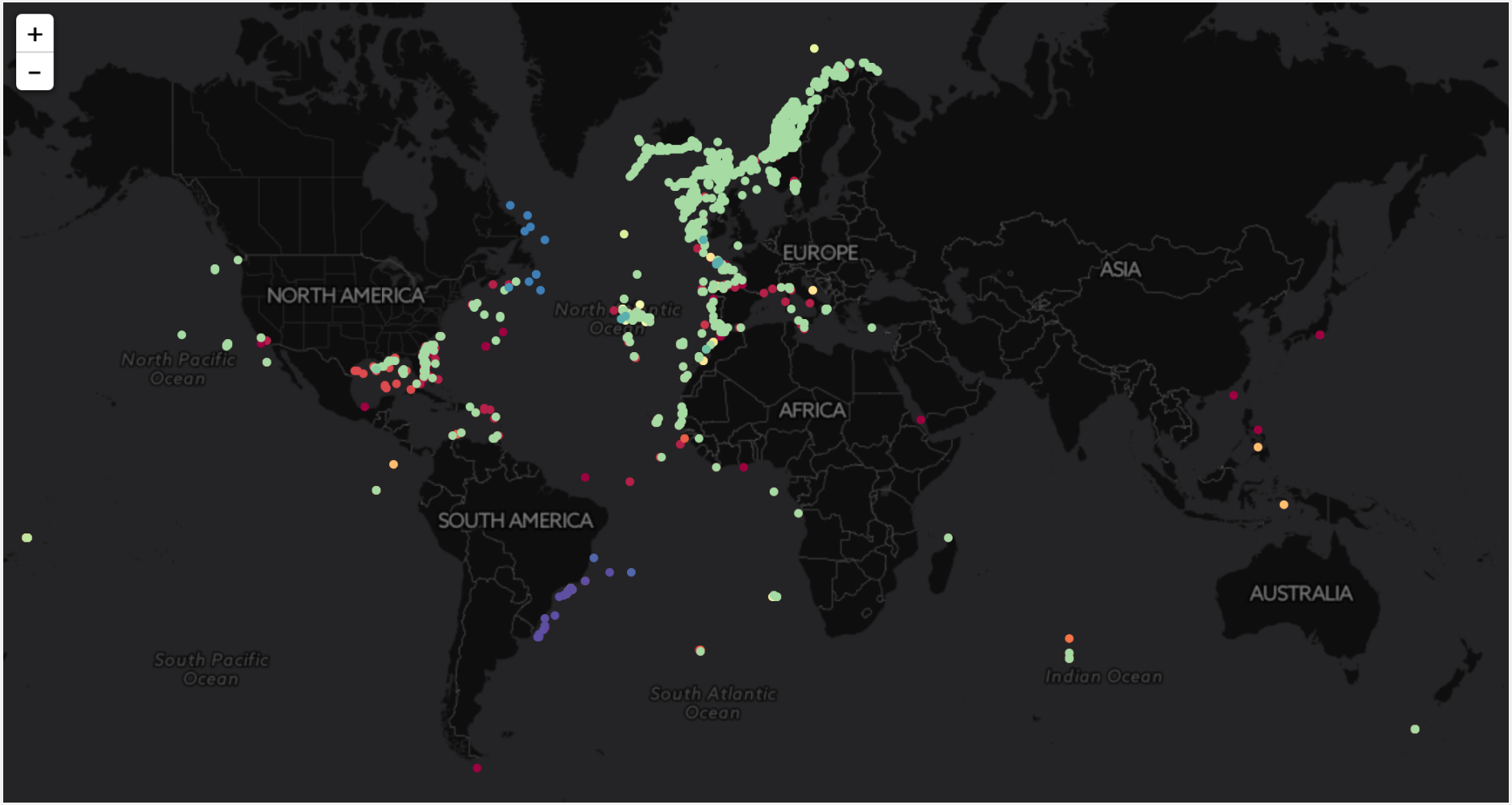
OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MAPPING

```
require(leaflet)
require(robis)
require(dplyr)

data <- occurrence("Lophelia")

pal <- colorFactor("Spectral", levels = unique(data$datasetName))

m <- leaflet() %>%
     addProviderTiles("CartoDB.DarkMatter") %>%
     addCircleMarkers(
         data = data %>% select(lat = decimalLatitude,
                                lng = decimalLongitude),
         radius = 3,
         weight = 0,
         color = pal(data$datasetName),
         fillOpacity = 1)
m
```

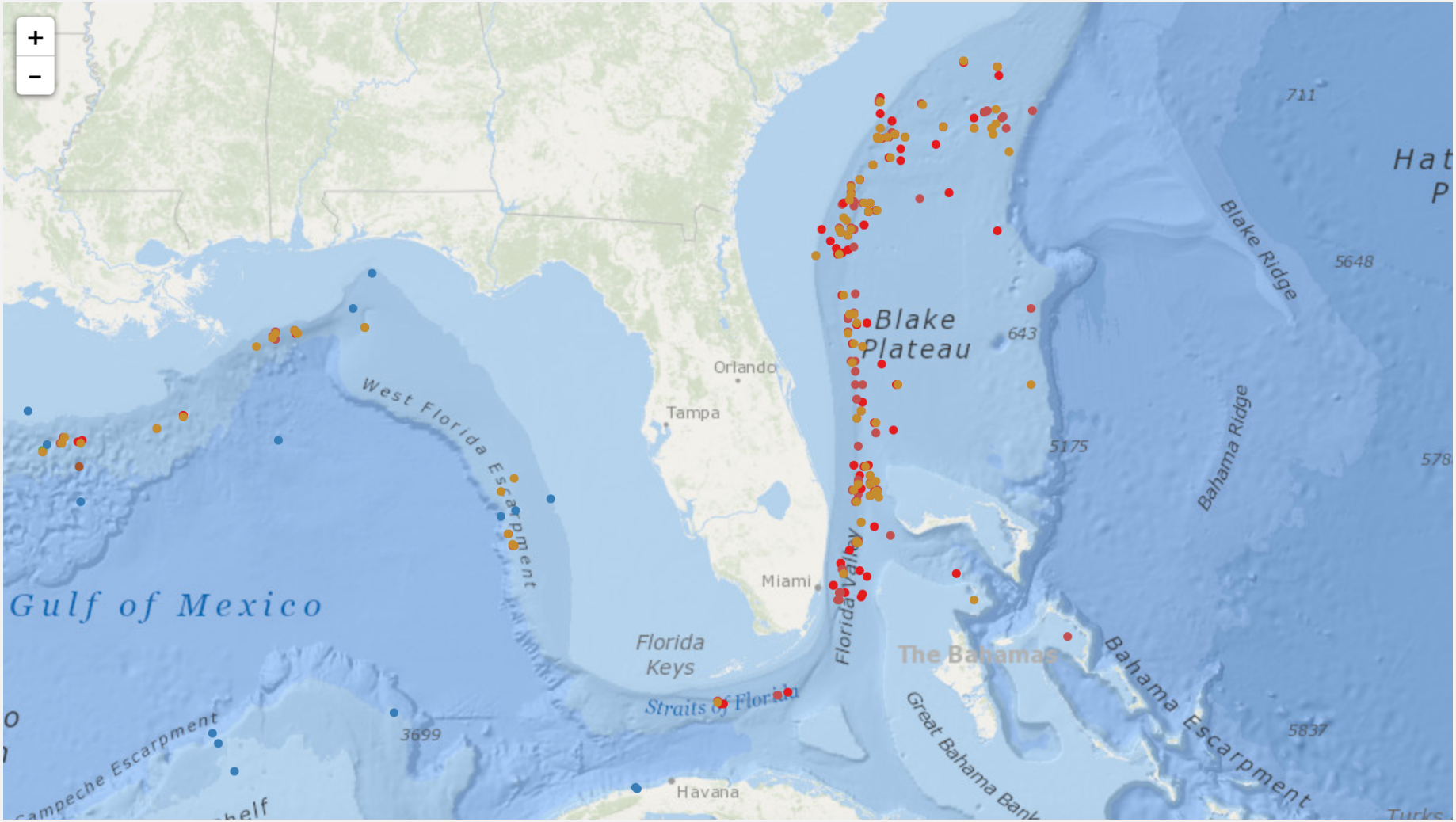
OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MAPPING



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

MAPPING



OBIS
OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM