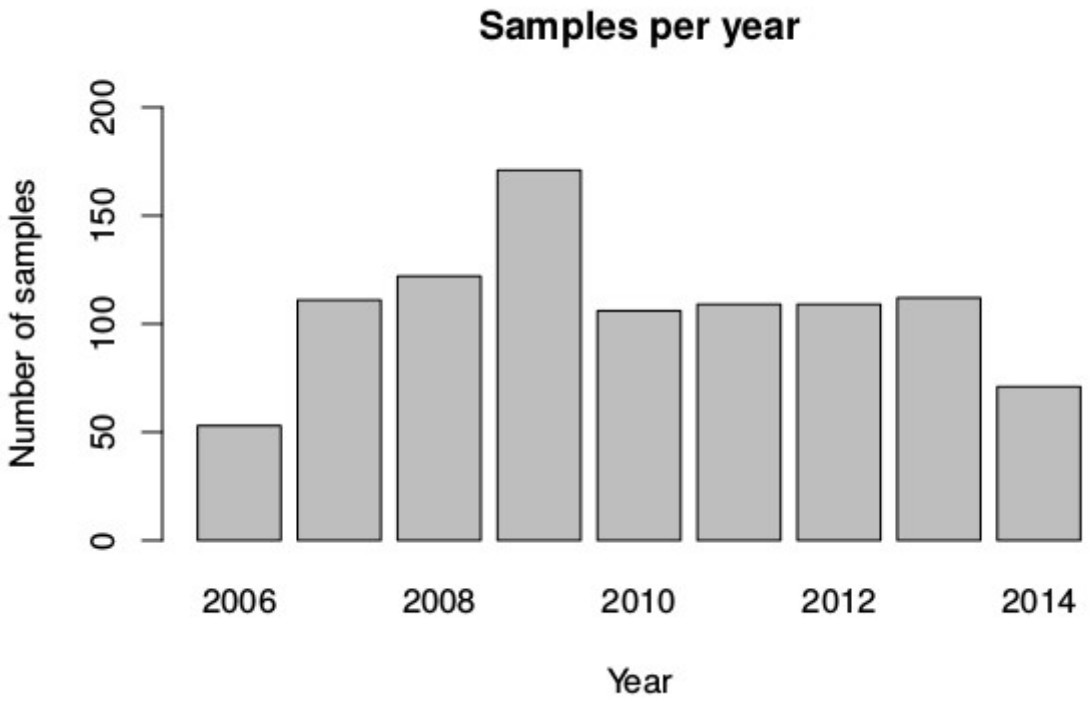
CONJUNTO DE DATOS IFCB

Datos básicos

- 3.457.819 ejemplos.
- Originalmente distribuidos en **115 clases**, que se han combinado (por recomendaciones del WHOI), en **51 clases**. Todavía es un número alto. Quizás se podrían combinar más clases pero en este apartado yo me pierdo un poco.
- 964 muestras completas.

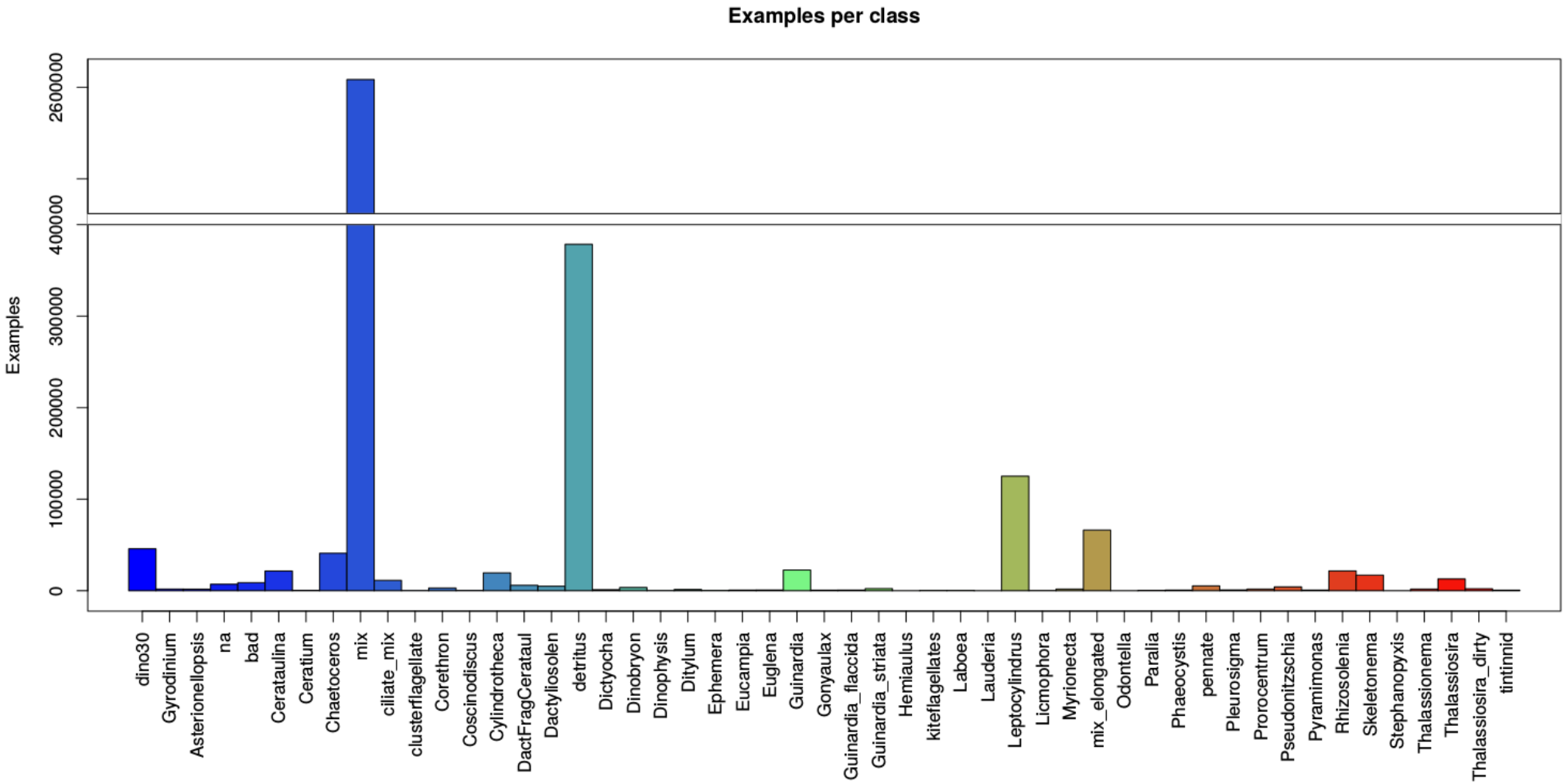
Distribución de las muestras

Las muestras van desde el año 2006 al 2014. La distribución de muestras por año se puede ver aquí:



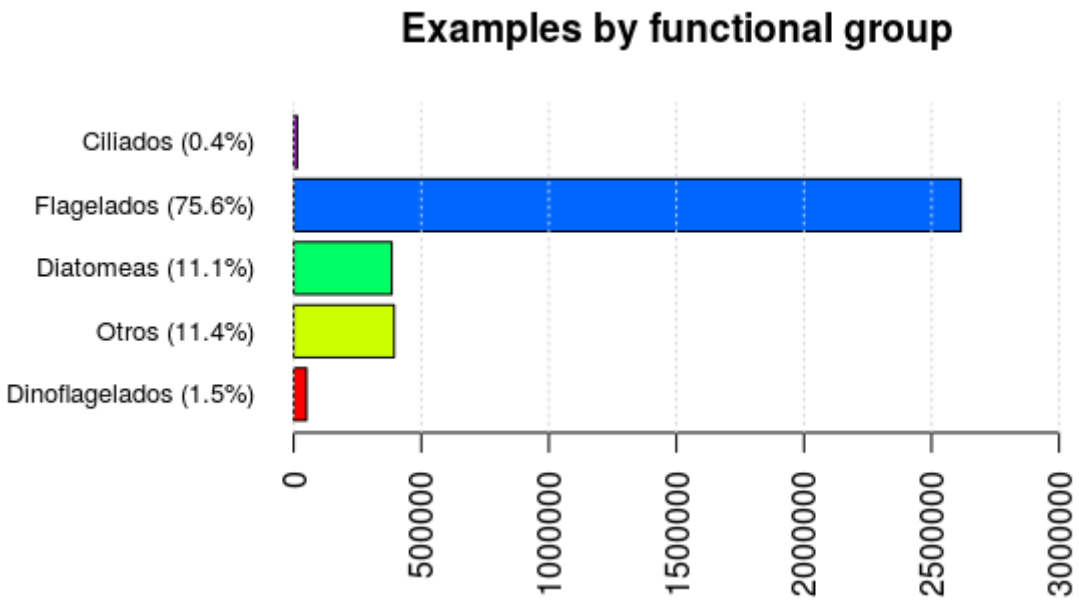
Distribución por clases

Como se puede ver en la gráfica estamos ante un conjunto de datos muy poco balanceado. La clase 'mix' tiene más de 2,5M ejemplos, que **suponen un 75% de los ejemplos totales**.



Distribución por grupos funcionales

Si agrupamos las 51 clases en grupos funcionales, obtenemos la siguiente distribución.



Descripción de los datos y CSV

El archivo CSV principal con los datos [IFCB.csv] tiene un tamaño de **7.8 Gb** [3.1 Gb comprimido]. El CSV tiene **239** columnas. Las tres primeras columnas son 'Sample', 'roi_number' y 'Class':

- Sample: Identificador de la muestra.
- roi_number: Identificador del ejemplo dentro de la muestra.
- Class: Clase (una de las 51 posibles).
- FunctionalGroup: Uno de los cinco grupos funcionales posibles.

Las 236 columnas restantes son las características de cada ejemplo. Estas características están descargadas del IFCB Dashboard. Hay siete características con valores nulos en 15.717 ejemplos. Estos ejemplos tienen los 7 valores nulos a continuación:

- "Area_over_PerimeterSquared"
- "Area_over_Perimeter"
- "H90_over_Hflip"
- "H90_over_H180"
- "Hflip_over_H180"

- "summedConvexPerimeter_over_Perimeter"
- "rotated_BoundingBox_solidity"

Descripción de las muestras

En el archivo IFCB_SAMPLES.csv se puede encontrar la información de las muestras. Para cada muestra tenemos:

- Sample: Id de la muestra (que coincide con el campo Sample del archivo IFCB.csv)
- Instrument: Instrumento con el que se ha sacado la muestra.
- Year: Año de la muestra.
- Day: Día del año de la muestra (1 a 365).
- Time: Hora de la muestra (formato hhmmss).