# Lab 4. Data Preparation

Select the most relevant data:

```
ratings_movies <- MovieLense[rowCounts(MovieLense) > 50,
                             colCounts(MovieLense) > 100]
ratings_movies
```

```
## 560 x 332 rating matrix of class 'realRatingMatrix' with 55298 ratings.
```

Exploring the most relevant data:

```
min_movies <- quantile(rowCounts(ratings_movies), 0.98)
min_users <- quantile(colCounts(ratings_movies), 0.98)
image(ratings_movies[rowCounts(ratings_movies) > min_movies,
                     colCounts(ratings_movies) > min_users],
main = "Heatmap of the top users and movies")
```

```
average_ratings_per_user <- rowMeans(ratings_movies)
qplot(average_ratings_per_user) + stat_bin(binwidth = 0.1) +
  ggtitle("Distribution of the average rating per user")
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

# Normalising data:

```
ratings_movies_norm <- normalize(ratings_movies)
sum(rowMeans(ratings_movies_norm) > 0.00001)
```

```
## [1] 0
```

Visualize the normalized matrix (it's colored because the data is continuous):

```
image(ratings_movies_norm[rowCounts(ratings_movies_norm) > min_movies,
                          colCounts(ratings_movies_norm) > min_users],
main = "Heatmap of the top users and movies")
```

# Binarising data

## 1st option: define a matrix equal to 1 if the movie has been watched

```
ratings_movies_watched <- binarize(ratings_movies, minRating = 1)
min_movies_binary <- quantile(rowCounts(ratings_movies), 0.95)
min_users_binary <- quantile(colCounts(ratings_movies), 0.95)
image(ratings_movies_watched[rowCounts(ratings_movies) > min_movies_binary,
                             colCounts(ratings_movies) > min_users_binary],
main = "Heatmap of the top users and movies")
```

## 2nd option: define a matrix equal to 1 if the cell has a rating above the threshold

```
ratings_movies_good <- binarize(ratings_movies, minRating = 3)
image(ratings_movies_good[rowCounts(ratings_movies) > min_movies_binary,
colCounts(ratings_movies) > min_users_binary],
main = "Heatmap of the top users and movies")
```