sparseDataSimulation

```
library(dplyr)
```

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

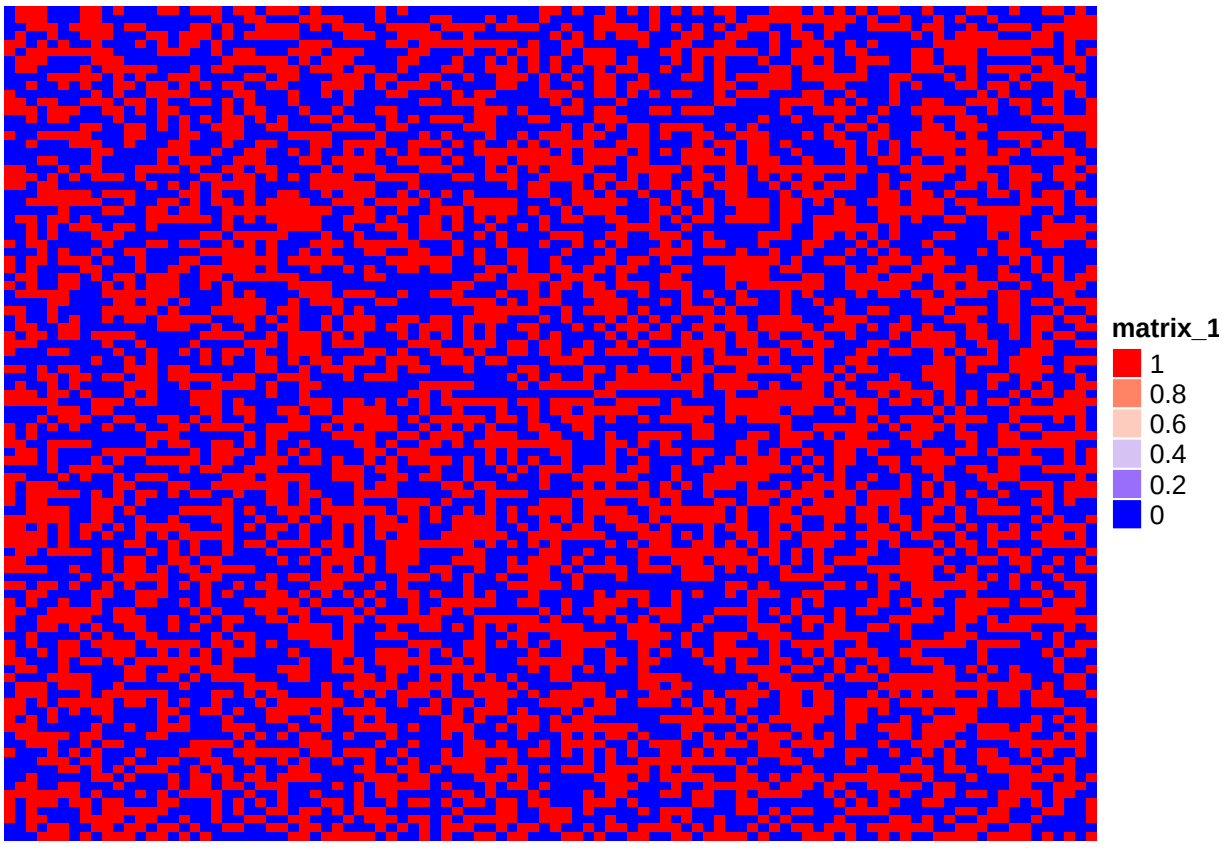
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ComplexHeatmap)
```

```
## Loading required package: grid
```

```
## Make a synthetic sparse data matrix
x <- sample(c(1, 0), 100 * 100, replace = T) %>% matrix(ncol = 100, nrow = 100)
Heatmap(x, cluster_rows = F, cluster_columns = F)
```



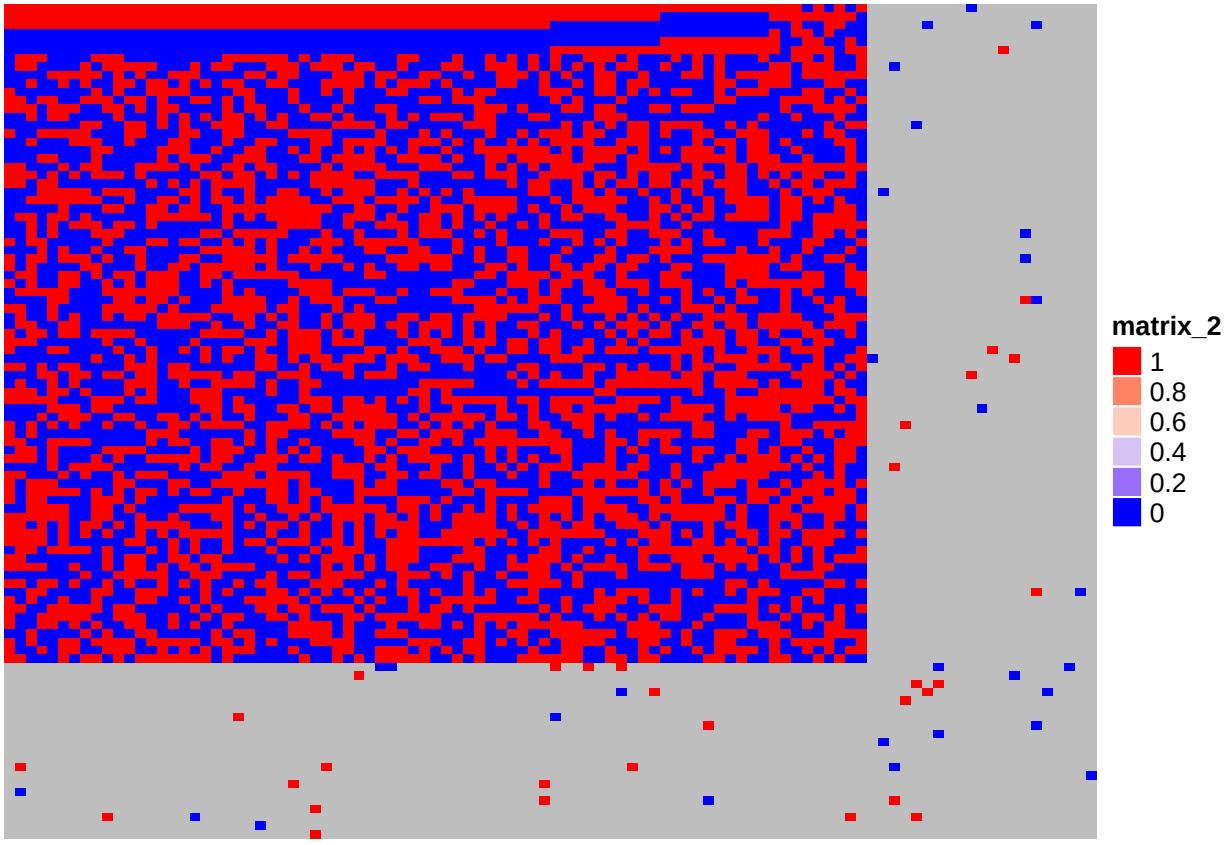

```
## 30*100+30*100-30*30=5100. Keep 60 values and make others NA => 5040 NA
x[80:100, ] <- NA
x[, 80:100] <- NA

for (i in 1:30) {
```

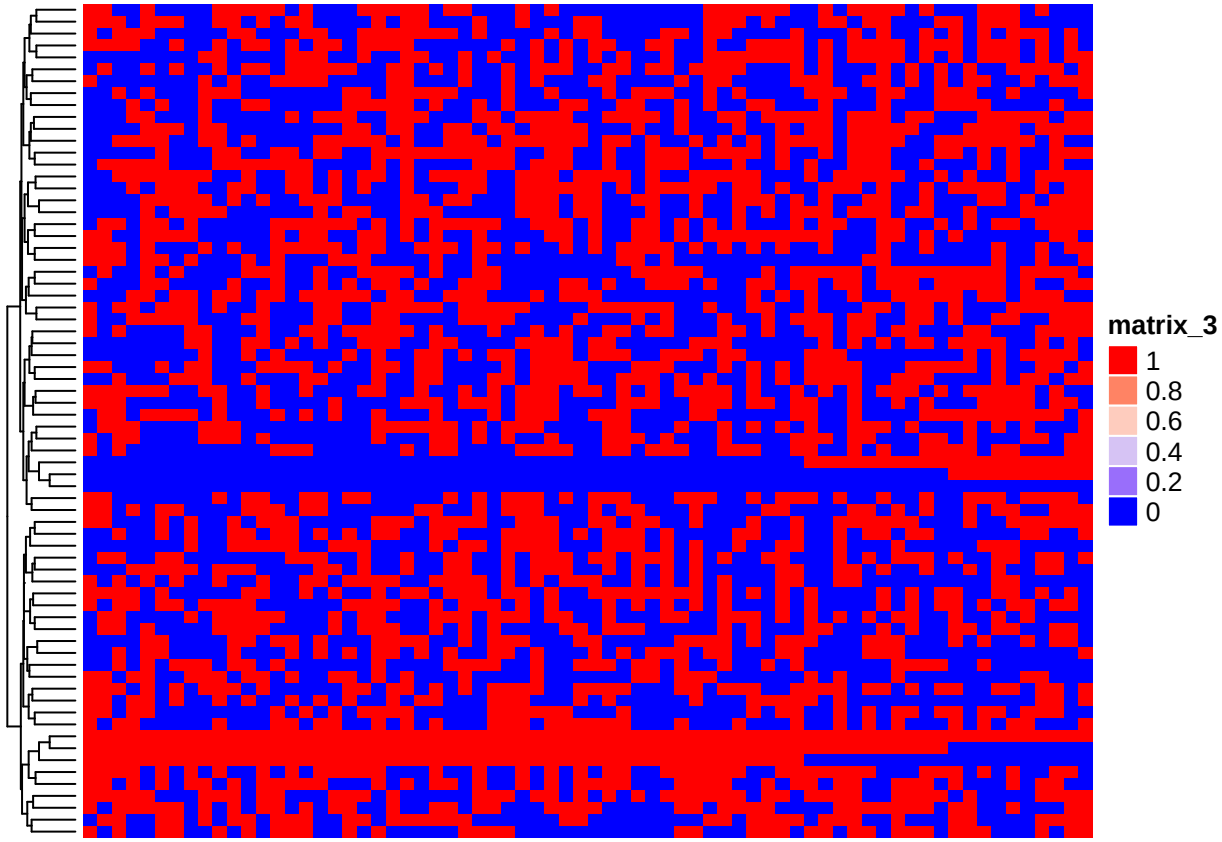
```
    x[sample(80:100, 1), sample(1:100, 1)] <- sample(c(1, 0), 1)
    x[sample(1:100, 1), sample(80:100, 1)] <- sample(c(1, 0), 1)
}

## Add 5 imaginary, highly correlated strains (Will use euclidean distance
## instead of pcc)
x[1, 1:70] <- rep(1, 70)
x[2, 1:70] <- c(rep(1, 60), rep(0, 10))
x[3, 1:70] <- c(rep(1, 50), rep(0, 20))
x[4, 1:70] <- rep(0, 70)
x[5, 1:70] <- c(rep(0, 60), rep(1, 10))
x[6, 1:70] <- c(rep(0, 50), rep(1, 20))

## Visualize the result:
Heatmap(x, cluster_rows = F, cluster_columns = F)
```



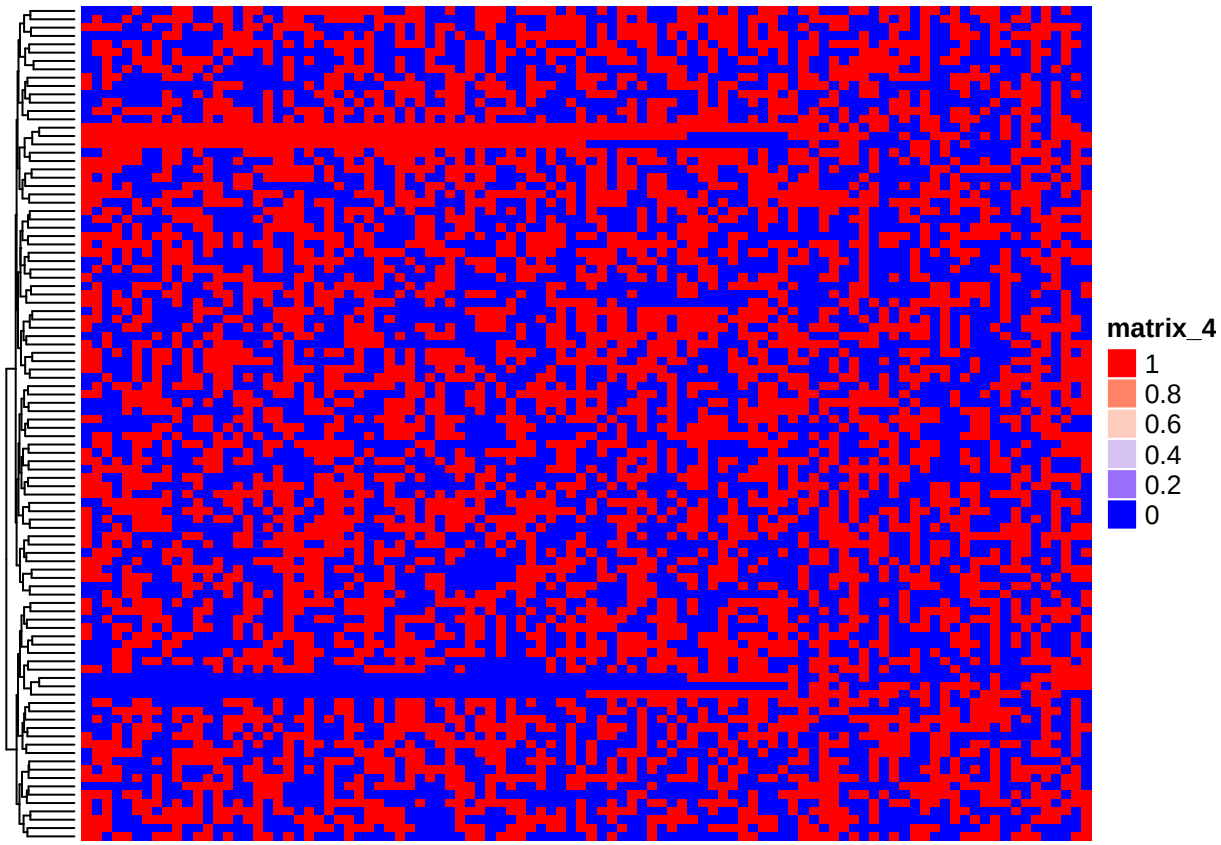

```
Heatmap(x[1:70, 1:70], cluster_columns = F)
```




```
## R would complain if this is done (because of NAs): x %>% dist %>% hclust

## Assuming that the distribution is normal (when the data is quantitative)
## => the imputed data would be 50% 0 and 50% 1
for (i in 1:dim(x)[1]) {
    for (j in 1:dim(x)[2]) {
        if (is.na(x[i, j]) == T) {
            x[i, j] <- sample(c(1, 0), 1)
        }
    }
}

## Visualize the result:
Heatmap(x, cluster_columns = F)
```




```
# The next step would be: By knowing more and more of the unknown data
# (which were imputed), what will be interesting? How can the fixed sparse
# data give us more info?
```