plots_datamining

ukke

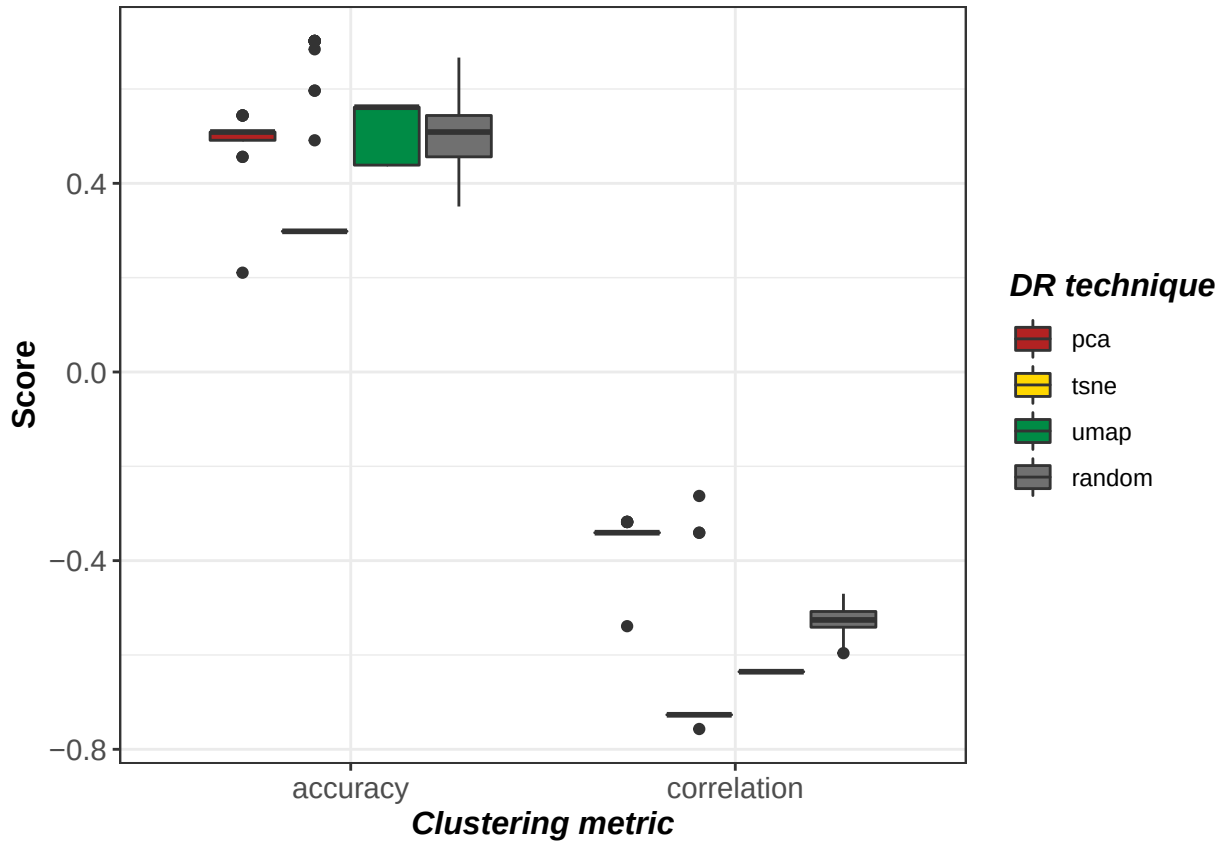
24 december 2021

Data exploration clustering data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
data <- read.csv("cluster_results.csv")
data$DR.technique <- ordered(data$DR.technique, levels = c("pca", "tsne", "umap", "random"))
ggplot(data) +
  geom_boxplot(aes(x = metric, y = value, fill = DR.technique))+
  theme_bw()+
  scale_fill_manual(values=c("firebrick", "gold", "springgreen4", "grey44"))+
  labs(fill = "DR technique")+
  xlab("Clustering metric")+
  ylab("Score")+
  theme(axis.title.x = element_text(size = 12, face = "bold.italic"))+
  theme(axis.title.y = element_text(size = 12, face = "bold"))+
  theme(legend.title = element_text(size = 12, face = "bold.italic"))+
  theme(axis.text = element_text(size = 11))
```




```
ggsave("clustering_metrics.png", dpi = 750)
```

```
## Saving 6.5 x 4.5 in image
```

Statistical testing clustering data

```
acc_data <- data[data$metric == "accuracy",]
cor_data <- data[data$metric == "correlation",]
kruskal.test(value ~ DR.technique, data = acc_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by DR.technique
## Kruskal-Wallis chi-squared = 137.63, df = 3, p-value < 2.2e-16
```

```
kruskal.test(value ~ DR.technique, data = cor_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by DR.technique
## Kruskal-Wallis chi-squared = 371.01, df = 3, p-value < 2.2e-16
```

p-value < 0.05 => significant difference between clustering techniques in both correlation and accuracy.

```
pairwise.wilcox.test(acc_data$value, acc_data$DR.technique,
                     p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  acc_data$value and acc_data$DR.technique
##
##        pca    tsne   umap
## tsne   <2e-16 -      -
## umap   0.313  <2e-16 -
## random 0.082  <2e-16 0.536
##
## P value adjustment method: BH
```

only tsne's mean accuracy differs significantly from the accuracy of clustering on random data

```
pairwise.wilcox.test(cor_data$value, cor_data$DR.technique,
                 p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  cor_data$value and cor_data$DR.technique
##
##        pca    tsne   umap
## tsne   <2e-16 -      -
## umap   <2e-16 <2e-16 -
## random <2e-16 <2e-16 <2e-16
##
## P value adjustment method: BH
```

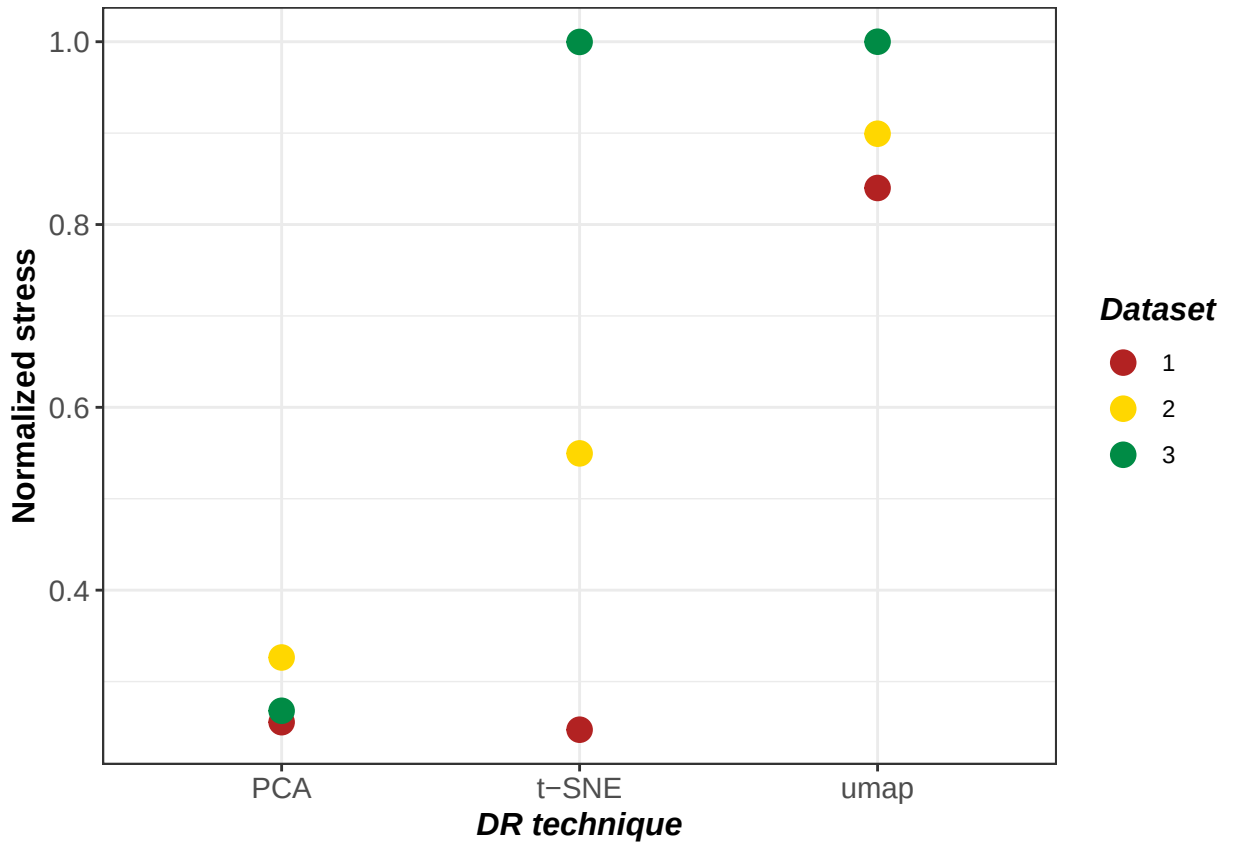
normalized stress

```
dat_normstress <- data.frame(c(0.8399765761963699, 0.8993401734587501, 0.9999928274887576, 0.2473133762
colnames(dat_normstress) <- c("norm_stress")
dat_normstress$dr_technique <- c(replicate(3, "umap"), replicate(3, "t-SNE"), replicate(3, "PCA"))
dat_normstress$dataset <- c(replicate(3,c("1", "2", "3")))
dat_normstress
```

```
##   norm_stress dr_technique dataset
## 1   0.8399766         umap       1
## 2   0.8993402         umap       2
## 3   0.9999928         umap       3
## 4   0.2473134        t-SNE       1
## 5   0.5496279        t-SNE       2
## 6   0.9997497        t-SNE       3
## 7   0.2555434          PCA       1
## 8   0.3263308          PCA       2
## 9   0.2680290          PCA       3
```

```
ggplot(dat_normstress)+
  geom_point(aes(x = dr_technique, y = norm_stress, color = dataset), size = 4)+
  theme_bw()+
  scale_color_manual(values = c("firebrick", "gold", "springgreen4"))+
  ylab("Normalized stress")+
  xlab("DR technique")+
  labs(color = "Dataset")+
  theme(axis.title.x = element_text(size = 12, face = "bold.italic"))+
  theme(axis.title.y = element_text(size = 12, face = "bold"))+
```

```
  theme(legend.title = element_text(size = 12, face = "bold.italic"))+
  theme(axis.text = element_text(size = 11))
```




```
ggsave("normalized_stress.png", dpi = 750, width = 6, height = 4)
```

```
# umap
reductions <- read.csv("reductions.csv")
reductions
```

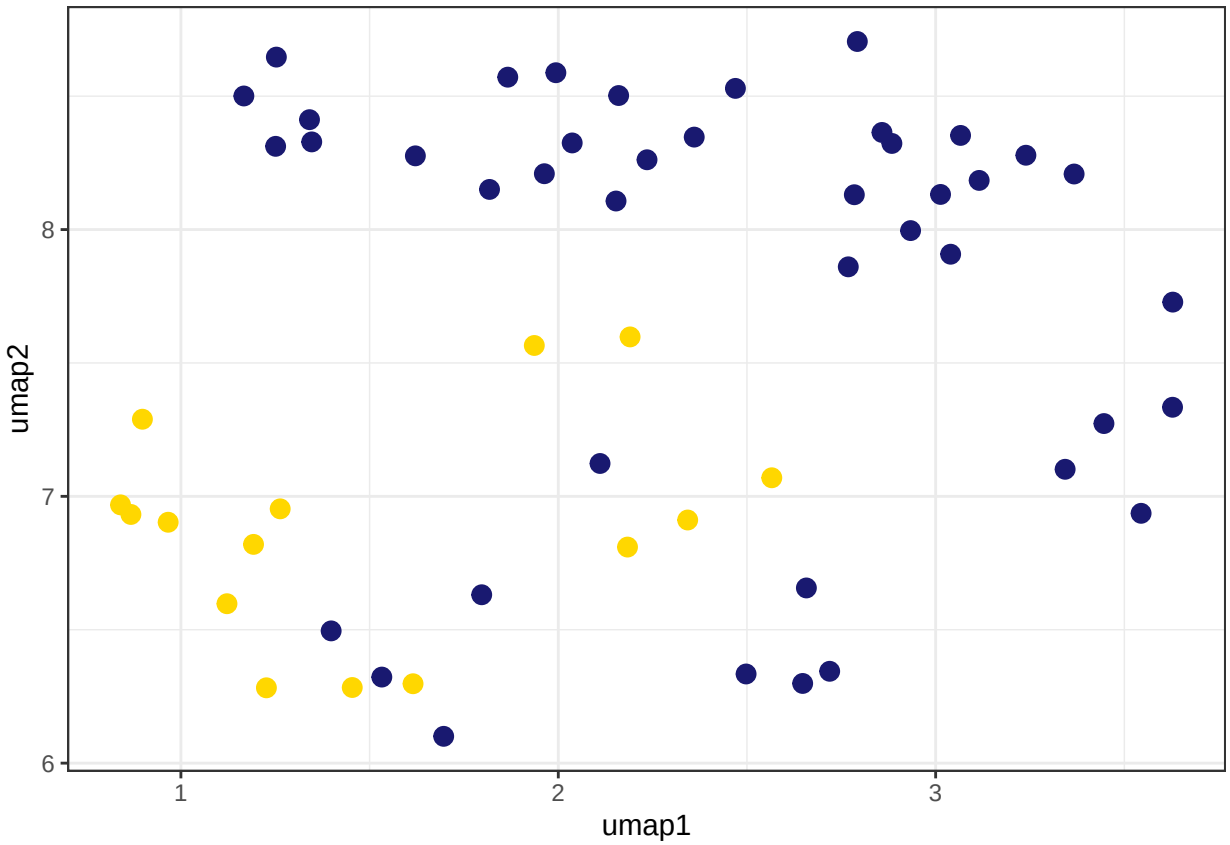
```
##      X                            label     umap1    umap2      umap3
## 1    0     inflammatory breast carcinoma 2.1600063 8.502431  0.9097819
## 2    1     inflammatory breast carcinoma 2.2351987 8.261634  0.5250331
## 3    2     inflammatory breast carcinoma 1.9630872 8.209410  0.7058328
## 4    3     inflammatory breast carcinoma 2.9335866 7.996118  1.6326293
## 5    4     inflammatory breast carcinoma 2.8840632 8.322863  1.4231348
## 6    5     inflammatory breast carcinoma 3.2391524 8.278604  1.6374718
## 7    6     inflammatory breast carcinoma 2.7924411 8.704933 -1.0961978
## 8    7     inflammatory breast carcinoma 2.1530509 8.107027 -1.0657251
## 9    8     inflammatory breast carcinoma 2.4694164 8.529235 -0.9993476
## 10   9     inflammatory breast carcinoma 1.9936638 8.587930  0.1940951
## 11  10     inflammatory breast carcinoma 1.8661669 8.571263 -0.1600585
## 12  11     inflammatory breast carcinoma 2.3602703 8.346621 -1.2797695
## 13  12     inflammatory breast carcinoma 2.0368590 8.324702 -0.8774019
## 14  13     inflammatory breast carcinoma 1.8177962 8.150499 -0.1014046
## 15  14     inflammatory breast carcinoma 1.6213261 8.276342  0.1619748
## 16  15     inflammatory breast carcinoma 1.3467263 8.328655  0.7812057
## 17  16     inflammatory breast carcinoma 2.6476181 6.298695  1.7030498
```

```
## 18 17     inflammatory breast carcinoma 2.4976869 6.334167  1.3145161
## 19 18     inflammatory breast carcinoma 3.0660479 8.352873 -1.2064062
## 20 19     inflammatory breast carcinoma 2.7192056 6.344221  1.3242422
## 21 20     inflammatory breast carcinoma 2.6574500 6.656671  1.0424001
## 22 21     inflammatory breast carcinoma 2.1106741 7.123216  0.8423802
## 23 22     inflammatory breast carcinoma 1.1668473 8.500697  1.5774854
## 24 23     inflammatory breast carcinoma 1.2510432 8.312054  1.2893339
## 25 24     inflammatory breast carcinoma 1.3407848 8.411951  1.4214667
## 26 25     inflammatory breast carcinoma 1.2529628 8.646497  1.5303801
## 27 26 non-inflammatory breast carcinoma 2.5658126 7.069632  1.7360228
## 28 27 non-inflammatory breast carcinoma 2.1834617 6.809886  1.7339787
## 29 28 non-inflammatory breast carcinoma 1.9365003 7.565526  2.0864983
## 30 29 non-inflammatory breast carcinoma 2.1903188 7.597658  1.8772724
## 31 30 non-inflammatory breast carcinoma 2.3429244 6.911337  1.3896978
## 32 31     inflammatory breast carcinoma 2.7684908 7.860325  2.7905686
## 33 32     inflammatory breast carcinoma 3.0397277 7.907635  2.6233008
## 34 33     inflammatory breast carcinoma 2.7844927 8.130567 -0.7583109
## 35 34     inflammatory breast carcinoma 3.3669057 8.208245  2.2358370
## 36 35     inflammatory breast carcinoma 3.1153188 8.184312  2.3397655
## 37 36 non-inflammatory breast carcinoma 1.4540627 6.283473  2.1562130
## 38 37 non-inflammatory breast carcinoma 1.2265625 6.282343  1.8213420
## 39 38 non-inflammatory breast carcinoma 1.6151458 6.297604  1.7334732
## 40 39 non-inflammatory breast carcinoma 1.1924306 6.819987  1.7840652
## 41 40 non-inflammatory breast carcinoma 1.2631432 6.953164  2.1043706
## 42 41     inflammatory breast carcinoma 3.6283443 7.728055  1.7739753
## 43 42     inflammatory breast carcinoma 1.3981574 6.495658  1.2625128
## 44 43     inflammatory breast carcinoma 1.5324097 6.322556  0.9353400
## 45 44     inflammatory breast carcinoma 1.7970895 6.631147  1.1083804
## 46 45     inflammatory breast carcinoma 1.6963269 6.100503  1.4025795
## 47 46     inflammatory breast carcinoma 2.8577716 8.364118 -1.2938393
## 48 47 non-inflammatory breast carcinoma 0.9662817 6.902884  2.7552595
## 49 48 non-inflammatory breast carcinoma 1.1220905 6.597536  2.5998511
## 50 49 non-inflammatory breast carcinoma 0.8675160 6.932119  2.2410219
## 51 50 non-inflammatory breast carcinoma 0.8980108 7.289110  2.1211562
## 52 51 non-inflammatory breast carcinoma 0.8400339 6.968087  1.9229612
## 53 52     inflammatory breast carcinoma 3.6277285 7.333962  2.0474241
## 54 53     inflammatory breast carcinoma 3.5443943 6.936367  2.1326714
## 55 54     inflammatory breast carcinoma 3.3431232 7.101486  1.8934696
## 56 55     inflammatory breast carcinoma 3.4459717 7.272841  2.3184304
## 57 56     inflammatory breast carcinoma 3.0129538 8.131695 -0.8813540
##          tsne1        tsne2       tsne3        pca1        pca2        pca3
## 1   -179.60103   141.726776  100.280663  -83279.157  -57478.186  -35719.521
## 2     46.35705  -156.347641  184.606659 -184946.913  -52275.679  -25506.223
## 3     95.32909   -39.852020  173.602219 -112768.098  -57765.963  -38194.547
## 4    227.92278   -16.983332  105.087570  -55297.805  -79530.920  137561.270
## 5    -49.99157  -284.266418   84.393829  -85054.359  -78805.096  129197.047
## 6   -330.68076    37.746323  -86.415840    7733.696  -98570.619  169254.893
## 7    -45.08903   244.649323  113.561302 -153401.769  336071.880   -9961.763
## 8    -55.89178   124.381165    8.587130 -438994.603  135299.380  -61908.318
## 9   -167.74689    82.259224 -204.956177 -344118.726  197681.587  -55791.190
## 10    24.79041    25.820082 -221.275726  -77371.288   87596.664  -45419.967
## 11   141.58971   101.097412  179.276260 -251183.225    5377.500  -42344.218
## 12  -130.82703   231.047119  -62.551964 -266677.894  211092.293  -85449.378
## 13   -95.03152   -20.139072  247.204803 -360178.877  157851.766  -91837.016
```

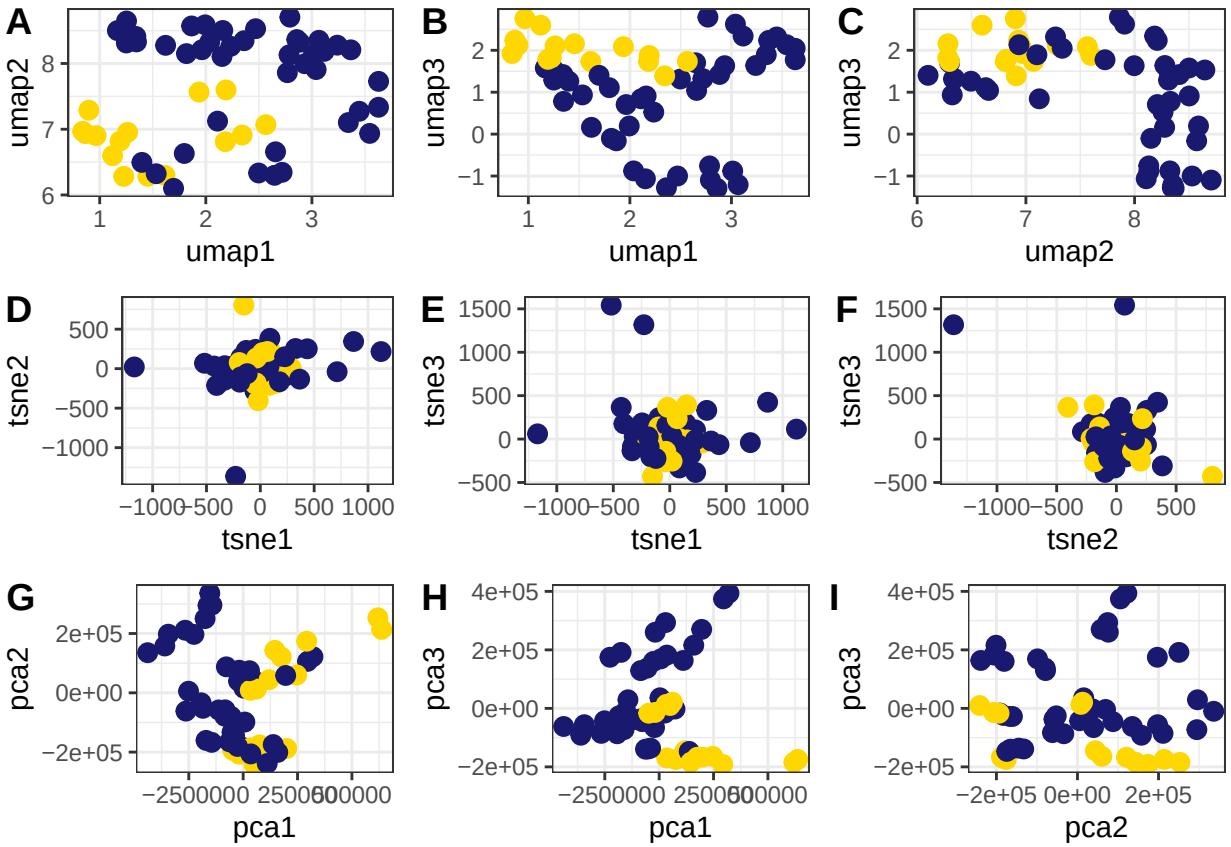
```
## 14  -429.60721    33.837788  364.950745 -263179.578  -61686.271  -83007.449
## 15   188.14606    32.794254 -175.401779 -193310.940  -33675.238  -86821.092
## 16  -139.90253   -13.610869   68.149445  -21349.769   39658.019  -65664.824
## 17    35.04933  -115.655602 -136.593079   26315.381 -191080.632  -15971.119
## 18  -314.88098   -91.444954   31.370518   -6367.775 -182742.906  -23962.670
## 19   327.27509   255.095062  330.469757 -143492.987  296035.235   29407.184
## 20  1121.02600   216.555435  114.376320    3642.762   15436.605   36135.017
## 21    85.82137   -12.721311 -332.706818 -144659.826 -167123.730  -27458.762
## 22    83.87376   133.243866   43.452385 -170777.930 -160460.495  -27621.032
## 23  -405.74686  -212.538101  175.736664   25644.234   23014.876   -9809.906
## 24  -244.72017    38.071243  186.273071   19421.768   69318.746   -3543.995
## 25   228.77365   -94.668922 -384.637268   54054.180   37367.108   -1954.243
## 26   435.92947   251.675339  -64.664520   72749.465   33746.415   -2323.611
## 27   148.98914  -186.095718  393.737671  -48055.902 -193586.627  -16559.724
## 28   -66.42537   -15.249841  -75.638168  -18605.266 -207050.266  -13899.335
## 29   -97.69703  -140.358429  139.330948   60961.913   12632.023   20985.005
## 30   102.00309   209.138596  -97.598526   32581.392    8451.347   13541.230
## 31  -152.48767   804.003479 -429.465912   49956.846 -241261.659    9729.369
## 32  -229.21451 -1363.166504 1316.963013  295825.195  105957.153  374267.601
## 33    89.09364   384.425079 -309.055145  320906.107  121830.669  393883.102
## 34   -25.36141    76.696098  157.218033 -226223.016  197051.699  174749.199
## 35 -1170.14636    22.043381   59.927895  -18125.025   76253.253  260391.959
## 36   713.20728   -38.418877  -41.092804   29191.557   74544.550  293351.581
## 37    82.45652  -211.658951    6.367661  200171.419 -187677.177 -166614.077
## 38   -50.38776  -193.510132  -48.765072   75715.735 -175605.435 -174876.528
## 39   152.63350   -96.174034  -10.435764   35137.907 -185252.108 -169638.970
## 40  -195.84482    77.937302  -48.345200  119253.437   44322.644 -144736.238
## 41   286.91614     8.639855  -46.387390  246243.271   60300.179 -164067.668
## 42   177.90875  -165.829559 -158.900391  159136.005 -200921.694  215885.953
## 43  -518.31836    68.381203 1542.420044  -39377.336 -144959.269 -135952.959
## 44    18.82808   -47.293377   36.309780  -51028.574 -132606.190 -138809.813
## 45   364.99210  -132.783661  -22.519257  -62826.936 -165405.960 -139489.260
## 46    80.60092    26.120234  -71.166809  138111.966 -174491.616 -147056.944
## 47  -333.96564  -146.750275 -134.968323 -161600.617  294640.825  -72750.219
## 48    20.09699   202.736465 -254.171158  619419.783  253009.530 -184627.566
## 49   -19.26101  -406.505463  366.562408  636806.544  214057.507 -174502.016
## 50   -34.96992   135.189590 -141.332794  292168.923  173953.328 -191350.316
## 51   -38.07642  -181.815521 -262.012573  175284.322  120846.190 -167144.531
## 52    65.58424   217.840195  234.741928  145631.238  143252.759 -184048.480
## 53  -182.76088  -104.982300  -88.003105   35264.021 -206088.454  182198.637
## 54   866.12299   344.064545  424.118622  110545.251 -238468.929  163833.390
## 55  -190.99696  -170.981995   25.609873  -26034.750 -180919.287  160608.624
## 56   227.33046   151.283295   -9.591022  195425.570   58490.163  269983.950
## 57  -121.90786   -63.217476 -223.521973 -175010.948  250348.513  191430.475
```

```
library(ggplot2)
plot <- function(dimension1, dimension2){
  dim1 <- which(colnames(reductions) == dimension1)
  dim2 <- which(colnames(reductions) == dimension2)
  ggplot(reductions)+
  geom_point(aes(x = reductions[,dim1], y = reductions[,dim2], color = label), size = 3)+
  scale_color_manual(labels = c("IBC", "non-IBC"), values = c("midnightblue", "gold"))+
  xlab(dimension1)+
  ylab(dimension2)+
```

```
  theme_bw()+
  theme(legend.position='none')
}
plot("umap1", "umap2")
```

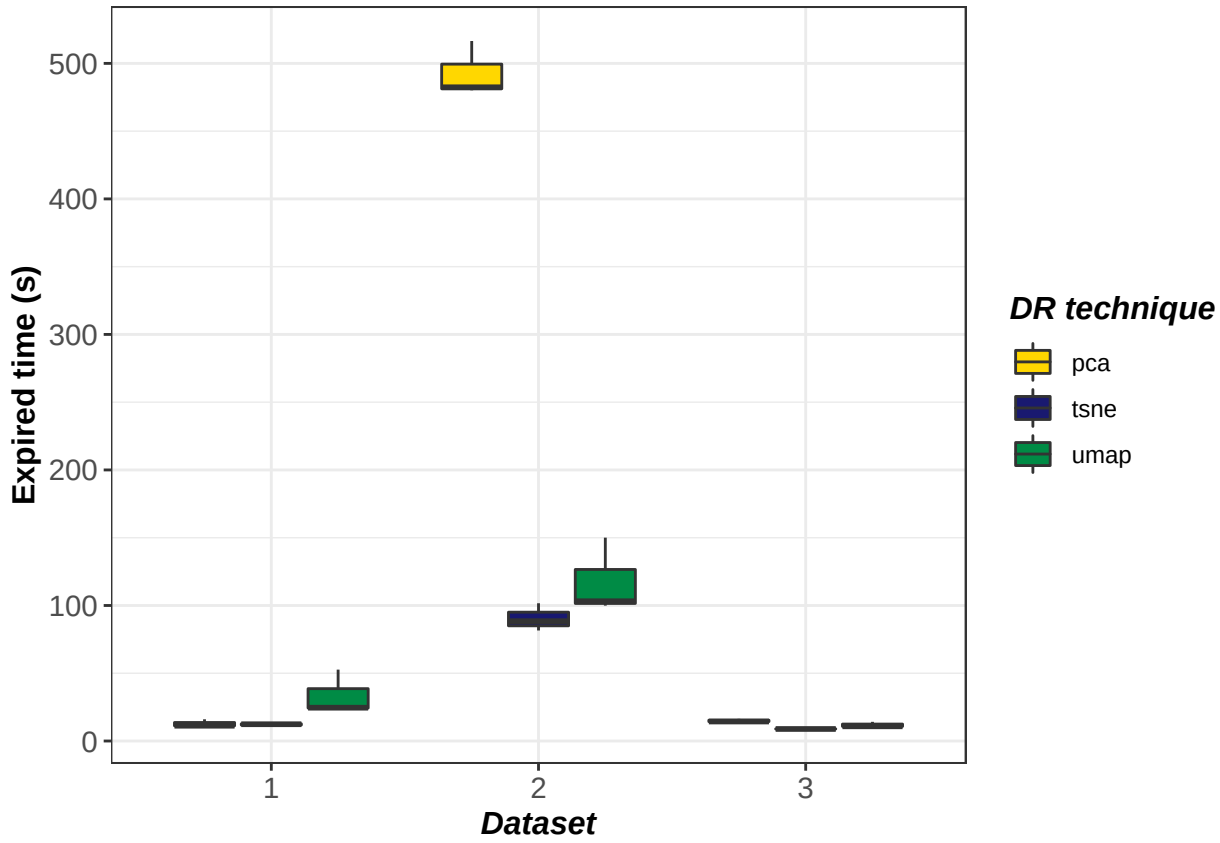



```
umap12 <- plot("umap1", "umap2")
umap13 <- plot("umap1", "umap3")
umap23 <- plot("umap2", "umap3")
umap <- cowplot::plot_grid(umap12, umap13, umap23, labels = c("A", "B", "C"), nrow = 1)

tsne12 <- plot("tsne1", "tsne2")
tsne13 <- plot("tsne1", "tsne3")
tsne23 <- plot("tsne2", "tsne3")
tsne <- cowplot::plot_grid(tsne12, tsne13, tsne23, labels = c("D", "E", "F"), nrow = 1)

pca12 <- plot("pca1", "pca2")
pca13 <- plot("pca1", "pca3")
pca23 <- plot("pca2", "pca3")
pca <- cowplot::plot_grid(pca12, pca13, pca23, labels = c("G", "H", "I"), nrow = 1)

cowplot::plot_grid(umap, tsne, pca, nrow = 3)
```




```
ggsave("reductions.png", dpi = 750, height = 12, width= 15)
```

```
library(ggplot2)
# umap
times <- read.csv("times.csv")
times$dataset <- as.character(times$dataset)

ggplot(times)+
  geom_boxplot(aes(x = dataset, y = value, fill = DR.technique))+
  theme_bw()+
  scale_fill_manual(values= c("gold", "midnightblue", "springgreen4"))+
  xlab("Dataset")+
  ylab("Expired time (s)")+
  labs(fill= "DR technique")+
  theme(axis.title.x = element_text(size = 12, face = "bold.italic"))+
  theme(axis.title.y = element_text(size = 12, face = "bold"))+
  theme(legend.title = element_text(size = 12, face = "bold.italic"))+
  theme(axis.text = element_text(size = 11))
```




```
ggsave("times.png", dpi = 750, height = 5, width = 8)
```