Practical 7: Data Stream Mining (cont)

I. Evaluation of data stream clustering

Internal evaluation measures:

- "average.between" Average distance between clusters
- "average.within" Average distance within clusters
- "max.diameter" Maximum cluster diameter
- "entropy" entropy of the distribution of cluster memberships

External evaluation measures:

- "precision" and "recall":
  - Precision=TP/(TP+FP)
  - Recall=TP/(TP+FN)
- "purity": Average purity of clusters. The purity of each cluster is the proportion of the points of the majority true group assigned to it.
- "Euclidean": Euclidean dissimilarity of the memberships

See the stream package for more measures

```
library("stream")
stream <- DSD_Gaussians(k = 3, d = 2, noise = .05)
```

1. Use Reservoir sampling to generate 100 data points and use K-means to generate 4 clusters.

```
Reservoir_Kmeans =
  DSC_TwoStage(micro = DSC_Sample(k = 100), macro = DSC_Kmeans(k = 4))
update(Reservoir_Kmeans, stream, n=500)
Reservoir_Kmeans
plot(Reservoir_Kmeans, stream)
evaluate_static(Reservoir_Kmeans, stream
                , measure =
                  c("average.between", "precision", "recall"), n =500)
```

2. Use sliding window method rather than Reservoir sampling in the above example. Compare the precision and recall of the two methods.

   Hint: Window_Kmeans = DSC_TwoStage(micro = DSC_Window(horizon = 100), macro = DSC_Kmeans(k = 4)).

II. Concept Drift

Concept drift means the changes of the data generating process over time. It implies that the statistical properties of the data also change when time passes. A good data mining algorithm should be able to deal with concept drift. In the stream package, DSD_Benchmark(1) is an example data stream which contains concept drift. To show the concept drift we request four times 250 data points from the stream and plot them. To fast-forward in the stream we request 1400 points in between the plots and ignore them. The codes below will show 4 figures of the data at different time points.

```
stream <- DSD_Benchmark(1)
stream
for(i in 1:4) {
plot(stream, 250, xlim = c(0, 1), ylim = c(0, 1))
tmp <- get_points(stream, n = 1400)
}
```

We can use animation package to demonstrate this:

```
reset_stream(stream)
animate_data(stream, n = 10000, horizon = 100
             , xlim = c(0, 1), ylim = c(0, 1))
library("animation")
animation::ani.options(interval = .1)
ani.replay()
```

III. Evaluation of data stream clustering with concept drift

1. Using Reservoir sampling and K-means

```
stream = DSD_Benchmark(1)
Reservoir_Kmeans= DSC_TwoStage(micro = DSC_Sample(k = 100, biased = TRUE)
                               , macro = DSC_Kmeans(k = 2))
update(Reservoir_Kmeans, stream, n=500)
plot(Reservoir_Kmeans, stream)
evaluate_stream(Reservoir_Kmeans, stream
                , measure = c( "precision", "recall"), n =5000, horizon=100)
```

2. Evaluate the Sliding window + K-means clustering