**Problem statement**

The current dataset represent only watched movies without compromise. This is not the case for unwatched movies.

Unwatched movies can represent the following: movies that I would not watch, movies I haven’t seen yet but would watch or anywhere in between. This requires some way to compare availability versus watchability. For this reason there are a couple of ways to rate the availability:

1. Movies released longer than 6 months ago start losing availability at a rate of 1 every year on a scale of 10. Also the watched movies will be judged for their availability at the time of their viewing. The resulting value will be used to select a subset of the unwatched that have an availability lower than the average availability that the watched movies have.
2. Using unsupervised clustering we can determine the most watched clusters and assume all movies in these clusters still have a chance of being watched. All other movies will be considered unwatchable. This requires a lot of experimentation to build an appropriate size of clusters.
3. A random sample is chosen from the unwatched movies that is the same size as the watched movies. This will be used to train a predictor of the Boolean label. This will the prediction algorithm will be chosen to allow for qualitative analysis on the prediction, like SVM etc.