Each user can be represented by a row vector of ratings of length equal to the total number of movies where each entry is a tuple of rating and timestamp:

The actual ratings data is in the form of a 20m x 4 matrix:

The complete ratings matrix can be represented as a matrix with each column representing a movie and each row representing a column.

The complete ratings matrix is much too large to hold in local machine memory (~5x typical local machine memory).

The complete ratings matrix is extremely sparse; only .53% of all entries are non-zero. Therefore, we should try not to create the complete ratings matrix.

Collaborative Filter (Pearson’s Correlation Coefficient):

Not the right formula to use yet.