McDonald’s Canada

Data Scientist

Interview Assignment

# Million Song Challenge

The Million Song Dataset Challenge aims at being the best possible offline evaluation of a music recommendation system. You may propose any algorithm or approach or language or tool as part of your solution.

# Deliverable

Please create a 25 minute presentation to answer the questions below:

1. Based on the objective, how would you build a recommendation model on which songs a user will listen to? Describe your approach.
2. As data is never perfect, how would you approach cleaning the data? What considerations would you make when moving this into production?
3. How would you test your model? How would you evaluate success?
4. How would you explain how the model works? What are the limitations of the model? What stages could be taken to evolve the model?
5. If an organization like Spotify wanted to implement the model, what strategy could be taken productionalise the model?

Note: This assignment requires you to only create a presentation to describe your approach and answers to each of the questions above - this is not a test on building an actual model.

# Resources

Ref: <https://www.kaggle.com/c/msdchallenge/data>

Ref: <http://millionsongdataset.com/>

Key Data Sets - <https://www.kaggle.com/c/msdchallenge/overview/custom>

* [user data](http://labrosa.ee.columbia.edu/millionsong/tasteprofile)
* [list of tracks](http://labrosa.ee.columbia.edu/millionsong/sites/default/files/AdditionalFiles/unique_tracks.txt) (careful, large text file!)
* [features](http://labrosa.ee.columbia.edu/millionsong/pages/tutorial) (and more, tutorial)
* track-level [tags and similarity](http://labrosa.ee.columbia.edu/millionsong/lastfm)
* [lyrics](http://labrosa.ee.columbia.edu/millionsong/musixmatch)
* [cover songs](http://labrosa.ee.columbia.edu/millionsong/secondhand)