# Final Project: Cdiscount product image data mining
# Phase 1

Haijin He

**Introduction:**

From Kaggle competition: https://www.kaggle.com/c/cdiscount-image-classification-challenge

The data is from Cdiscount, France's largest online non-food retailer.

As the amount of products Cdiscount sell is millions and still rapidly growing, it's a very difficult task to assign them to correct categories manually. Right now, they've already applying machine learning algorithms from text descriptions of the product to predict its category. But it seems the current method is reaching its limits. And in this challenge, they want to try learning from the images associated with the product.

The data contains information of it's products, with product id, category, and image information.

**Data description:**

Training data: training.bson size: 59.2 GB

Contains a list of 7,069,896 dictionaries, one per product. Each dictionary contains a product id (key: _id), the category id of the product (key: category_id), and between 1-4 images.

Testing data: test.bson size: 14.5 GB

Contains 1,768,182 products. Same format as train.bson, but without the category id.

Category_names.csv:

Contains the 3 level name for each category id.

**Data storage:**

Bson format

BSON(Binary JSON) format.

To read it, use bson module included in pymongo module. Image are then transformed into numpy.ndarray format.

**Preliminary analysis:**

1, Image per product: 1.34

Most product have only 1 image.

2, images:

Size: 180 X180 X 3

Most are color images.

3, Categories:

Each category id corresponds to a 3 level category name.

Total number of categories is 5271

Number of level1 category is about 50.

Number of level2 category is about 350

**Objectives & Execution:**

1: As stated in the Kaggle competition, object 1 is to predict the category id for the test data set.

Planned methods:

Deep learning is the state of art method for image classification. Since I plan to use Python environment, I plan to use TensorFlow library with Python, and related libraries like numpy, pandas.

Image annotation method will also be used to enhance prediction accuracy. Image annotation automatically assigns key words to images, and by comparing the key words to category names, it should help classification accuracy.

2: Image clustering

Here the data set contains 3 level of categories. For some categories, I can try to use clustering algorithm to separate them into sub-categories. The result can be compared with the given category label to see if anything interesting is given.

To implement the clustering, the method I plan to try several methods include K-means and hierarchical clustering. How to represent distance between images will carefully considered.

**Conclusion**

The classification task is a challenging one for it's data size, and for it's big amount of categories. Never the less, it is one that is manageable. The current best score on Kaggle is approaching 0.77 and surely it'll go up with still 2 month to go. I would be very interested to see the final score.