**Background:**

You have data on passive mutual funds and ETFs from the Morningstar database for a specific date (snapshot as of October 31, 2017).

**Requirements:**

1. Start by inspecting and cleaning the data. Produce a brief “Data Quality Report” outlining what you find (maximum 1 page in Word).
2. Next, make a database using the cleaned data (any type of DB you want), justify your choice of DB and schema, and write a query of some kind to answer the following question:
   * For funds with more than 1 share class, what is the average 5-year information ratio?
3. Use the dataand your favorite visualization techniques to describe the risk/return characteristics of the different asset classes. Create a presentation or webpage (or any other medium you choose) to present the results.
4. Finally, do ONE of the following:

* **Machine learning exercise:** How well can “machine learning” (automated algorithm of some kind) categorize funds into asset classes (e.g. Fixed Income, Equity, Alternatives, Allocation)? Write code to answer this question, and write a summary or make a presentation of your methodology & results (no more than 2 pages in Word, or 4 slides in PowerPoint). Assume your audience is technical but not expert in machine learning. Include at least 1 paragraph discussing potential problems/issues with the analysis.
* **Portfolio optimization exercise**: Given the data you have, define an ‘optimal portfolio’ and build an algorithm to choose an optimal portfolio. The definition of ‘optimal’ and the constraints are up to you. E.g. you might come up with some algorithm to pick the minimum number of investments needed to achieve a given risk/return profile, defined somehow. Assume that having more securities has some cost, e.g. so you don’t end up with 7 European stock ETFs, etc. Write a summary or make a presentation of your methodology & results (no more than 2 pages in Word, or 4 slides in PowerPoint).

Feel free to use any programming language(s).

To summarize, in 1 week you need to send us the following in a zipped file via email:

1. **Part 1:**
   1. Your code (if applicable)
   2. The 1-page “Data Quality Report”
2. **Part 2:**
   1. Your database schema (e.g. a diagram)
   2. Your query (e.g. a SQL query)
   3. 1 paragraph explaining your choice of database and schema
3. **Part 3:**
   1. Your code (if applicable)
   2. Presentation, webpage, etc.
4. **Part 4**
   1. Your code
   2. Presentation, Word Doc, etc.

**Evaluation:**

You will be evaluated based on:

* Clear articulation, understanding and further refinement of the problem(s) – they are intentionally somewhat vague
* Thoroughness of data quality checks and cleaning
* Explanation of database choice and explanation of schema
* Quality of your visualization, in terms of both clarity and aesthetics (i.e. displays information efficiently; tells a coherent story; not confusing; etc.)
* One of the following:
  + Machine learning techniques (choice of algorithm, proper use of k-fold cross-validation, training/test split, etc.)
  + Optimization assumptions and methodology
* Quality and documentation of code used to produce any of the above results (e.g. formatting, comments)
* Clarity of writing (e.g. in Word Doc, if you choose to present that way) and/or clarity of presentation (PowerPoint)