**Project Proposal**

Image processing for the apparel and fashion industry

**Problem:**

Be able to identify images of clothing items from targeted social media sites.

An apparel/fashion company wanting to start a new marketing campaign targeting users of social media such as Pinterest and Instagram.

People who post a lot of apparel pictures to **Pinterest or Instagram** may be planning a purchase or an event like a vacation. By identifying these images, the apparel/fashion company can develop:

* Models that can identify what makes a *blue blouse,* can use the information to recommend products that are visually similar.
* Understanding a person’s style in clothing can also help target them for related products.
* A following on these social media sites.

Also identify people who are following specific sites or people on Pinterest and Instagram that focus on apparel, can be potential customers

**Outline:**

* Data
* Process
  + Reduce the dimensionality of a data set to optimize performance of Machine-learning algorithms

**Data**

The data used for this project is from [Kaggle](https://www.kaggle.com/zalando-research/fashionmnist).

Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes:

0-Tshirt/top

1-Trousers

2-Pullover

3-Dress

4-Coat

5-Sandal

6-Shirt

7-Sneaker

8-Bag

9-Ankle boot

**Process**

**Reduce the dimensionality of a data set**

Principal component analysis (PCA) is an unsupervised method for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss. It does by creating new uncorrelated variables that successively maximize variance. PCA is mainly used for two purposes:

* Data Visualization
* Improve performance of Machine-learning algorithms

The main focus used in this project was for improving performance in benchmarking a number of Supervised learning models for classification

There are few stepsto reduce the dimensionality of a data set.

1. Standardize the Data

PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScaler to help standardize the dataset’s features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms.

1. Explained Variance

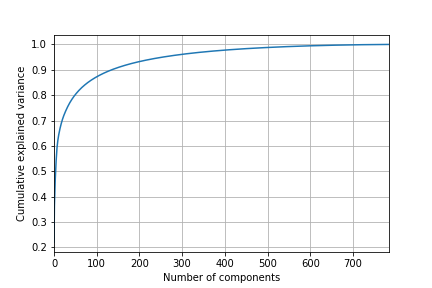
The explained variance tells how much information

(variance) can be attributed to each of the principal components.

1. Choose the number of components to use

Based on the acceptable accuracy

**PCA data based on Training Data**



The above plot shows:

~80% variance for 50 components

~88% variance for 100 components

~94% variance for 200 components

~96% variance for 300 components

We are able to observe that around 50 components the curve starts to level off and the variance is about ~80%. We will use 50 components for the PCA to reduce the dimensionality.

Was able to take advantage of the data visualization opportunities with PCA. Looking at the first two components, representing about 40% Cumulative explained variance

From the graph you can see the clustering, but not enough to set them apart.

