K-Means clustering

Please read the comments in each code block. The comments provide instructions and there are places that you are expected to fill in your own code. In order to get familiar with scikit learn's library you are expected to read the documentation. In the comments for the code links have been provided.

If you have followed my instructions on setting up virtual environments, then you might not have skimiage or tqdm installed. You need to install the following packages

```
pip install scikit-image
pip install tqdm
```

In [ ]:
```
# import stuff that we need
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import skimage
import skimage.io as skio
import tqdm as tq
from tqdm import tqdm_notebook as tqdm
```

We will work a particular image taken from Wikipedia. It is included in the repository as `talos.jpg`. The orginal can be downloaded from wikimeda be sure to save it as `talos.jpg`.

Part 1 - K-Means clustering

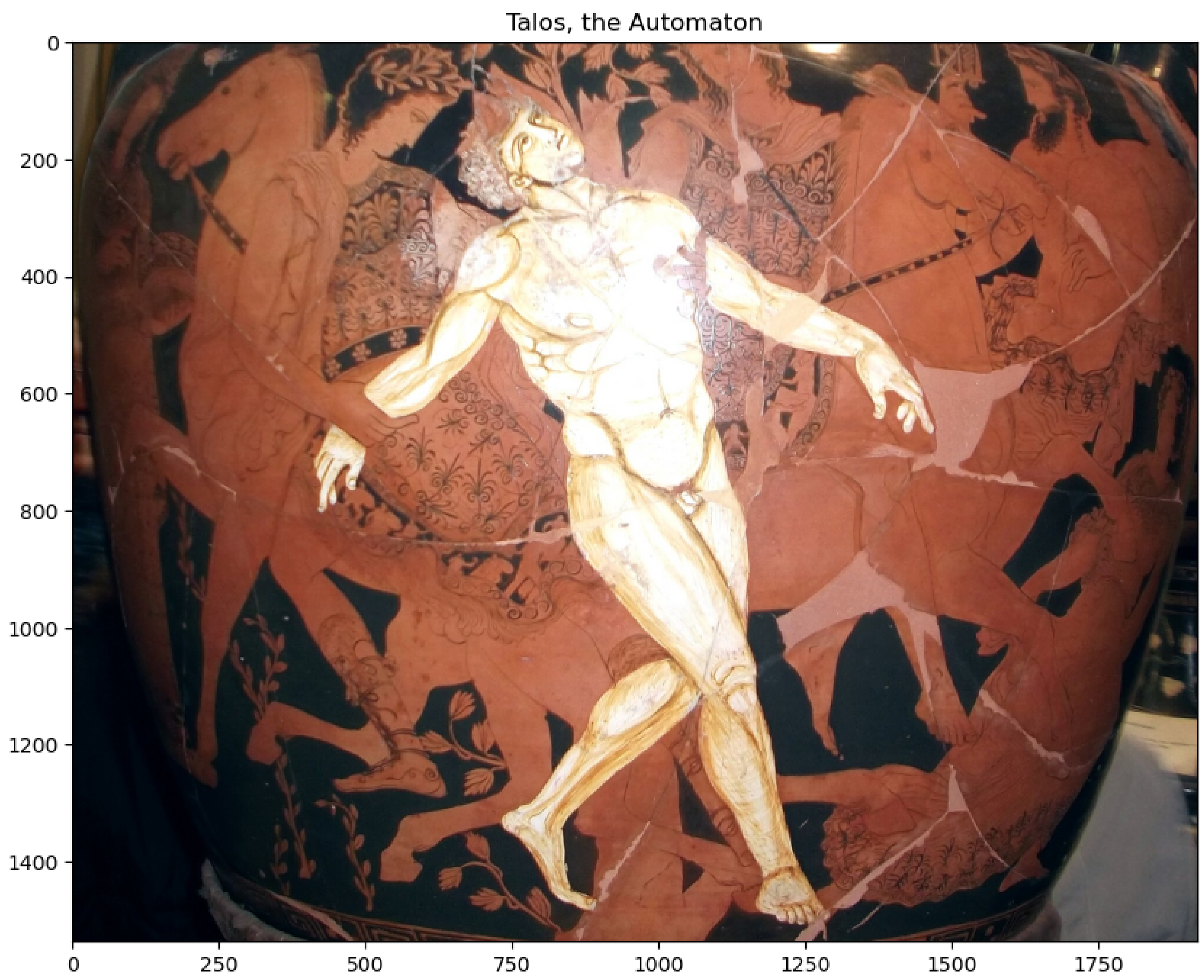
K-means clustering is an unsupervised method for finding clusters in data. There can be any amount of clusters and there can by any dimensions. Let's name the number of clusters K, and the number of dimensions D. The algorithm works as:

1. Specify the number of clusters *k*
2. Randomly initialize *k* centroids in the data.
3. Assign each point to its closest centroid
4. Compute the new centroids (mean) of each cluster
5. Repeat 3 and 4 until cluster centers does not change or until a pre-defined number of iterations

One special case of K-means is image compression, this will be the topic of this notebook.

Let's start with an image. Load the image and standardize it, i.e scale the values to a range 0-1.

In [ ]:
```python
# Loading the image
# Documentation: https://scikit-image.org/docs/dev/api/skimage.io.html#skima
# [CODE HERE] Load the image in a variable called "image"
image = skimage.io.imread("talos.jpg")
# [/CODE HERE]

# Standardization of the image (values go from the range 0-255 to 0-1)
# Documentation: https://scikit-image.org/docs/dev/api/skimage.html#skimage.
# [CODE HERE] Standardise the image
image = skimage.img_as_float32(image)
# [/CODE HERE]
assert image.max() - 1.0 < 1e-7, "The image must be standardized."

# Plotting the image
plt.figure(figsize=(10, 10))
plt.title("Talos, the Automaton")
plt.imshow(image)
plt.show()
```



In [ ]:
```python
print(f"Image width   : {image.shape[0]}")
print(f"Image height  : {image.shape[1]}")
```

```
print(f"Image channels: {image.shape[2]}")
print(f"Image size    : {np.prod(image.shape)}")
```

```
Image width   : 1537
Image height  : 1920
Image channels: 3
Image size    : 8853120
```

Image Information:

- **Width**: 1537 pixels
- **Height**: 1920 pixels
- **Channels**: RGB (Red, Green, Blue)
- **Uncompressed Size Calculation**:
  - Total pixels = Width × Height = 1537 × 1920 = 2,956,160 pixels
  - Each pixel has 3 channels (RGB), so the total number of values = 2,956,160 × 3 = 8,868,480 values
  - Each value is coded as a byte (8 bits), so the full uncompressed image size = 8,868,480 bytes or approximately 8.9 MB (megabytes).

Color Depth and Compression:

- **Current Color Depth**: Each channel uses 8 bits, resulting in 24 bits per pixel (8 bits for Red, 8 for Green, and 8 for Blue). This allows for 2^24(around 16 milion) different colors.
- **Reducing Colors for Compression**:
  - By choosing a palette with K colors (e.g., 16), the image would be restricted to only 16 different colors.
  - With a palette of 16 colors, each pixel can be represented using 4 bits (as 2^4 = 16). This is significantly smaller compared to the original 24 bits per pixel.
- **Compression Factor**:
  - Reducing from 24 bits (RGB) to 4 bits (with a palette of 16 colors) results in a compression factor of 83%.
  - The reduction in bits per pixel leads to a smaller file size while sacrificing some color fidelity.

But first, let's reshape the data to be able to display the RGB data in 3-dimensions and discard some data for faster computations.

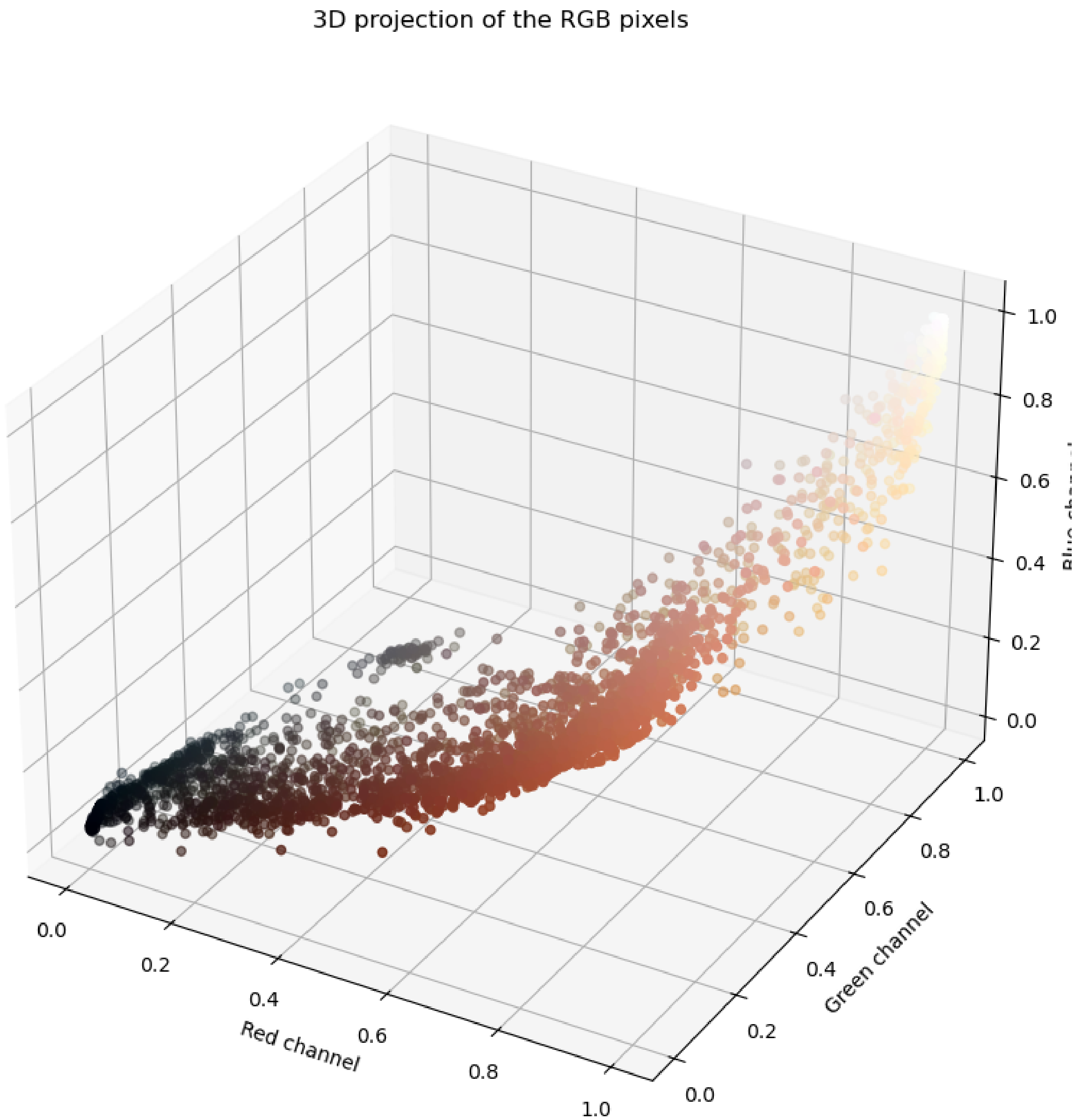
In [ ]:
```
# This steps reshape the image in the format (N, D) for N points in D dimens
# D will always be 3 since we will only deal with RGB images today.
pixels = image.reshape(-1, 3)
print("Pixels array shape                   :", pixels.shape)

# There are 2 951 040 pixels in 3 dimensions, which is A LOT!
# If you have too much data, algorithms will be slow, and displaying the dat
# So let's keep 0.1% of them (in other words randomly discard 99.9%)
# Documentation: https://docs.scipy.org/doc/numpy-1.14.0/reference/generated
```

```
# [CODE HERE] Create a variable "keep" of size pixels.shape[0] that contains
#             Make sure it is a numpy array
keep = np.random.rand(pixels.shape[0]) < 0.001

# [/CODE HERE]
assert keep.dtype == bool, "keep must be containing booleans"
assert len(keep) == pixels.shape[0], "keep has the wrong shape, it should be
assert (np.unique(keep) == [False, True]).all(), "keep must only contain Tru
# Now the smaller dataset is named pixels_small
pixels_small = pixels[keep]
print("Pixels array shape (after discarding):", pixels_small.shape)
print(f"Reduction in size                    : {1-pixels_small.shape[0]/pixe
```

```
Pixels array shape                   : (2951040, 3)
Pixels array shape (after discarding): (2963, 3)
Reduction in size                    : 99.9%
```

In [ ]:

```
"""
Now that we have the image collapsed in a list of pixels, it is possible
to display each individual pixel in 3D, just by connecting the RGB intensiti
to the axis X, Y, Z.
"""
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(10, 10))
ax = fig.add_subplot(111, projection="3d")
ax.scatter(*pixels_small.T, color=pixels_small)
ax.set_xlabel("Red channel")
ax.set_ylabel("Green channel")
ax.set_zlabel("Blue channel")
plt.title("3D projection of the RGB pixels")
plt.show()
```



Step 1: Computing distances

The first step consist of computing the euclidean distance between two sets of points which we will later use in the k-means algorithm. The eucledian distance between two points p and q is given by

$$d = \sqrt{(q-p)^2}$$

which for multiple dimensions extends to

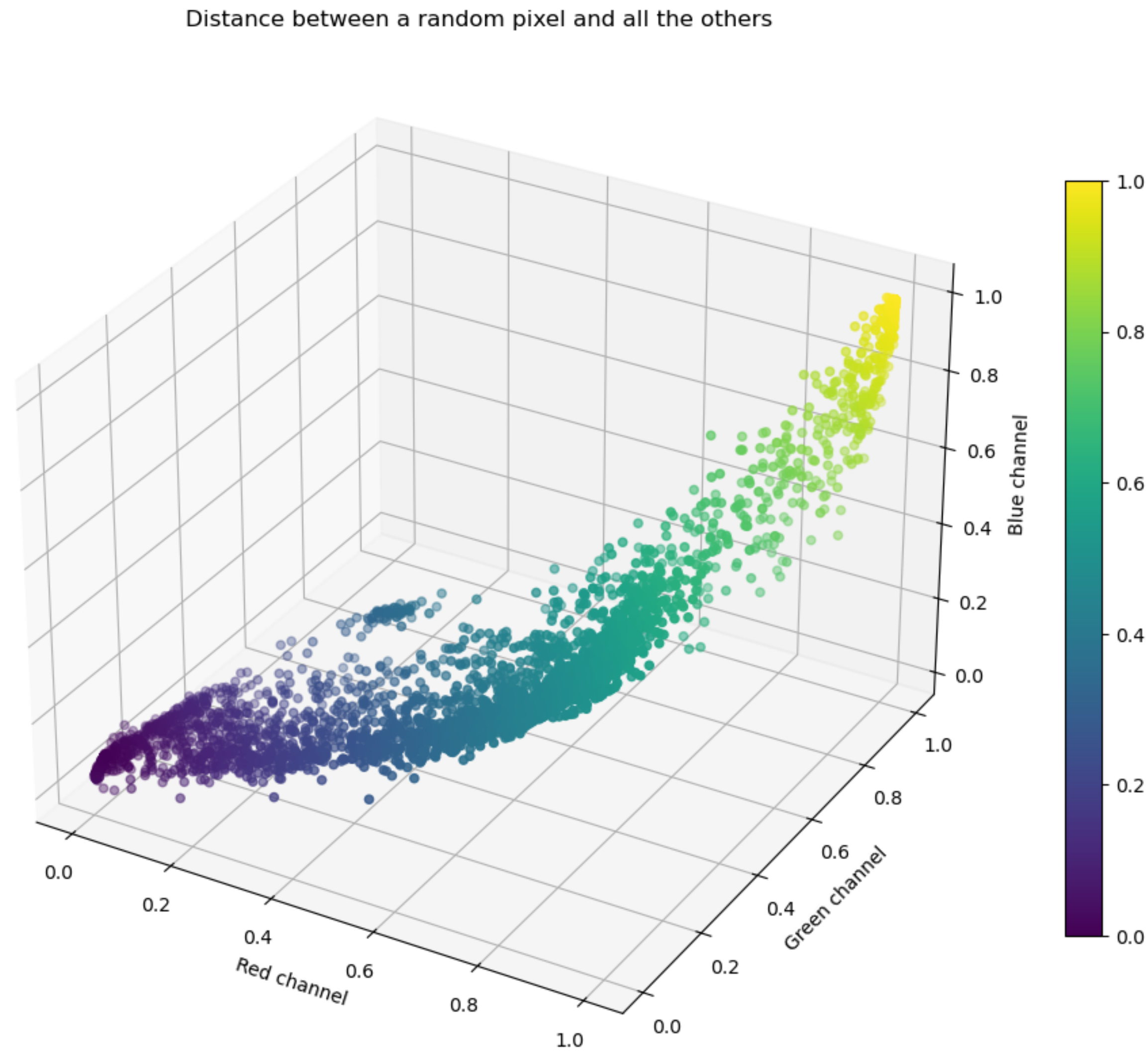
$$d = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + (q_3-p_3)^2 + \ldots + (q_n-p_n)^2}$$

In [ ]:
```python
def euclidean_dist(a, b):
    """Computes the euclidean distance between two sets of D-dimensional poi

    Args:
        a (array): A list of points of shape (N, D).
        b (array): A list of points of shape (M, D).

    Returns:
        (array): An array of shape (N, M) containing all the pairwise distar
    """
    # If a or b are python lists, they are transformed into numpy arrays.
    if isinstance(a, list): a = np.array(a)
    if isinstance(b, list): b = np.array(b)
    assert a.ndim == 2 and b.ndim == 2, "a and b must be 2-dimensional array
    assert a.shape[1] == b.shape[1], "a and b must have the same dimension D

    N = a.shape[0]
    M = b.shape[0]

    distances = np.ones((N, M))

    # [CODE HERE] Fill in the matrix "distances" so that it contains the pai
    # Advice: try using only two nested for-loops, for speed's sake.
    '''
    # My slow solution, which does not terminate when doing the full image i
    # the section "Clustering the image", likely due to a stack overflow.
    # It does however pass the asserts in the cell below...
    for i in range(N):
        for j in range(M):
            distances[i, j] = np.sqrt(np.sum((a[i] - b[j])**2))
    '''

    # Faster solution which terminates. It is not written by me, I found it
    a_squared = np.sum(a**2, axis=1).reshape(-1, 1)  # (N, 1)
    b_squared = np.sum(b**2, axis=1).reshape(1, -1)  # (1, M)
    distances = np.sqrt(a_squared + b_squared - 2 * np.dot(a, b.T))

    # [/CODE HERE]

    return distances
```

In [ ]:
```python
# Testing the euclidean distance
assert euclidean_dist([[0]], [[1]]) == 1, "Unit test 1 failed."
assert euclidean_dist([[0, 0, 0]], [[1, 1, 1]]) == np.sqrt(3), "Unit test 2
np.random.seed(0)
random_array1 = np.random.rand(10, 4)
random_array2 = np.random.rand(10, 4)
assert np.abs(euclidean_dist(random_array1, random_array2).mean() - 0.8897)
```

In [ ]:
```python
# We will test the euclidean function by picking a random pixel
# and checking its distance with all the other pixels in the image.
# you can choose to modify "random_pixel" to a specific pixel or
# position in the 3D space.
N = pixels_small.shape[0]
random_index = np.random.randint(N)
```

```
random_pixel = np.array([pixels_small[random_index]])

# Computing all the distances
distances = euclidean_dist(random_pixel, pixels_small)[0]
# The distances are normalised to be between 0 and 1.
distances /= distances.max()
import mpl_toolkits.mplot3d.art3d as art3d

fig = plt.figure(figsize=(12, 10))
ax = fig.add_subplot(111, projection="3d")
p = ax.scatter(*pixels_small.T, c=distances)
fig.colorbar(p, fraction=0.03)
ax.set_xlabel("Red channel")
ax.set_ylabel("Green channel")
ax.set_zlabel("Blue channel")
plt.title("Distance between a random pixel and all the others")
plt.show()
```



KMeans clustering

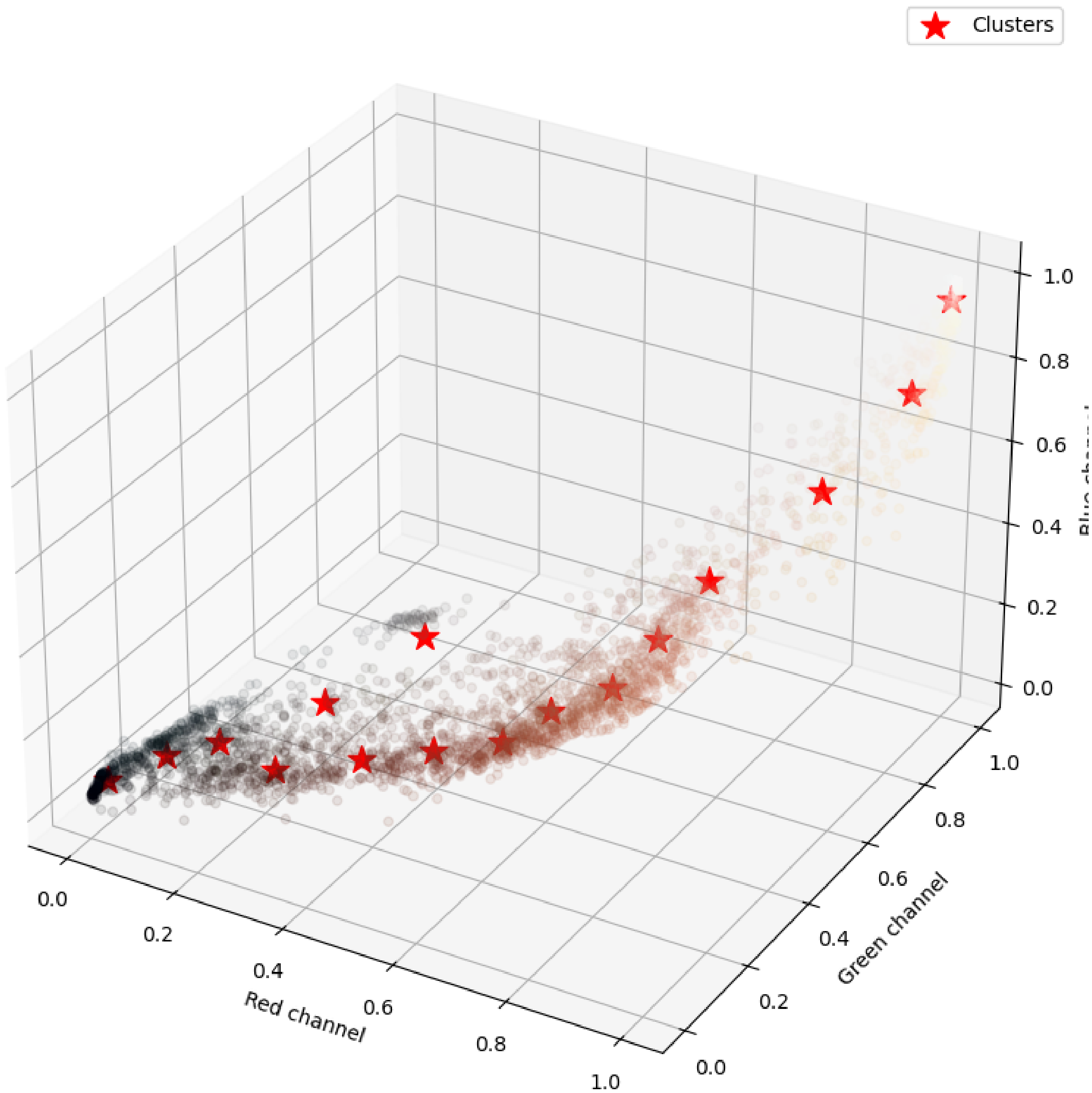
Now lets implement the k-means algorithm. We will construct a class that has two functions. One for fitting the data, i.e iterativly finds the optimal cluster centers and one for predicting data points into the respective clusters.

In [ ]:
```
# The following part needs to be implemented by you.
class KMeans:
    """ K-Means Algorithm. """

    def __init__(self, K, D):
        """Initialisation of the KMeans algorithm.
        Args:
            K (int): The number of clusters to use.
            D (int): The number of dimensions
        """
        self.K = K
        self.D = D
        # Initialise the clusters to zero
        self.clusters = np.zeros((K, D))

    def fit(self, data, iterations=5):
        """Trains the algorithm and iteratively refines the clusters' positi

        Args:
            data (array): The data points to cluster, shape must be (N, D)
            iterations (int): The number of iterations of the K-means algori

        Note:
            The algorithm updates the member variable "clusters".
        """
        assert data.ndim == 2 and data.shape[1] == self.D, "The data should
        assert iterations > 0, "The number of iterations should be positive.
        # Starting the algorithm
        N = data.shape[0] # Number of points in the data

        # [CODE HERE] Update the clusters in function of the data
        # 1. Pick K random points from the data and use them as starting pos
        points = np.random.choice(N, size=self.K, replace=False)
        self.clusters = data[points]
        for _ in range(iterations):
            # 2. Compute the distances between the data and the clusters
            distances = euclidean_dist(data, self.clusters)
            # 3. Associate each data point to the nearest cluster
            assoc = np.argmin(distances,axis=1)
            # 4. For each cluster
            for i in range(self.K):
                # 5. Gather all the points in the cluster
                clustered_points = data[assoc == i]
                # 6. Compute the mean value of the cluster
                if(len(clustered_points)> 0):
                    mean = np.mean(clustered_points, axis=0)
                    # 7. Update of the position of the cluster
                    self.clusters[i] = mean
                else:
                    self.clusters[i] = np.random.rand(self.D)
        # [/CODE HERE]
```

```
    def predict(self, data):
        """Predicts the cluster id for each of the

        Args:
            data (array): The data points to cluster, shape must be (N, D)

        Returns:
            (list of int): The id of the cluster of each of the data points,
        """
        assert data.ndim == 2 and data.shape[1] == self.D, "The data should

        # [CODE HERE] Update the clusters in function of the data
        # 1. Compute the distances between data and the clusters.
        distances = euclidean_dist(data, self.clusters)

        # 2. The datapoints are associated to each clusters.
        clustered_points = np.argmin(distances, axis = 1)
        # [/CODE HERE]

        return clustered_points
```

In [ ]:
```
# Choose a number of clusters
K = 16
kmeans = KMeans(K=K, D=3)
kmeans.fit(pixels_small, iterations=30)
```

In [ ]:
```
"""
Again, we show the 3D projection of the RGB pixels, along with the position
"""
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(10, 10))
ax = fig.add_subplot(111, projection="3d")
ax.scatter(*pixels_small.T, color=pixels_small, alpha=0.1)
ax.scatter(*kmeans.clusters.T, s=200, color="red", marker="*", depthshade=Fa
ax.set_xlabel("Red channel")
ax.set_ylabel("Green channel")
ax.set_zlabel("Blue channel")
plt.legend()
plt.show()
```



Clustering the image

Now let's test cluster the rest of the pixels in the image and assing new colors i.e compress the image. This might take a while

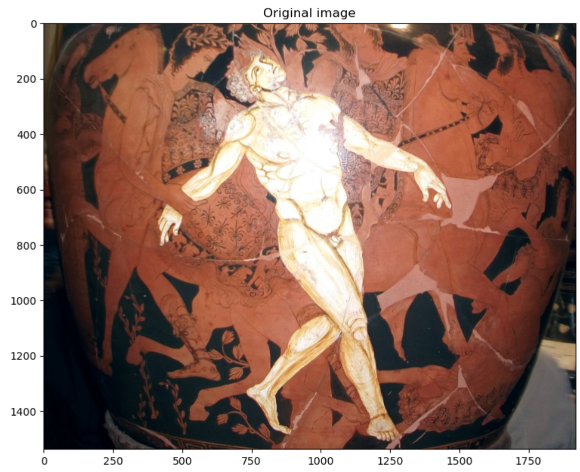
In [ ]:
```
# We take the original image (we dropped a lot of pixels in the beginning, r
# we will compress the full sized image.
# We already trained the kmeans algorithm by calling fit before.
# Each of the pixels of the big image is associated to a specific cluster.
clustered_image = kmeans.predict(pixels)

# We now need to determine which color is given to each cluster. In this cas
# we will take the mean.
# Constructing the palette
palette = np.empty((kmeans.K, kmeans.D))
# For each cluster
for i in range(kmeans.K):
```

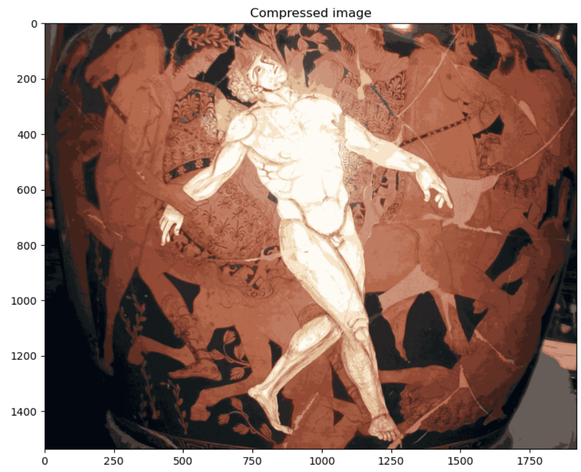
```
    # We take the pixels belonging to the ith cluster.
    cluster = pixels[clustered_image == i]
    # We compute the average color for the cluster
    color = cluster.mean(axis=0)
    # The ith color of the palette is set.
    palette[i] = color
```

In [ ]:
```
fig, ax = plt.subplots(1, kmeans.K, figsize=(2*kmeans.K, 3))
fig.suptitle("Palette")
for i in range(kmeans.K):
    ax[i].set_title(f"Cluster {i+1}")
    ax[i].imshow(palette[i].reshape(1, 1, kmeans.D), interpolation="None")
    ax[i].axis("off")
plt.show()
```

In [ ]:
```
def cluster2image(image, palette, imshape):
    """Constructs an RGB image from the clustered pixels and a palette.

    Args:
        image (list of int): a list of clustered pixels, shape (N).
        palette (array): a list of K different colors.
        imshape: the 2D shape of the image to create.
    """
    assert image.ndim == 1, "The image must have only one dimension."
    assert palette.ndim == 2, "The palette must have two dimensions."
    assert isinstance(imshape, tuple), "imshape must be a tuple."

    N = image.shape[0]
    K, D = palette.shape

    final_image = np.empty((N, D))
    for i in range(K):
        cluster = image == i
        for j in range(D):
            final_image[cluster, j] = palette[i, j]

    return final_image.reshape(imshape)
```

In [ ]:
```
final_image = cluster2image(clustered_image, palette, image.shape)

fig, ax = plt.subplots(1, 2, figsize=(20, 10))
ax[0].set_title("Original image")
ax[0].imshow(image)
ax[1].set_title("Compressed image")
ax[1].imshow(final_image)
plt.show()
```


Original image
0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750


Compressed image
0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750