Global Clustering Approach Example

Let's use 15 images to create 3 clusters, showing how the global approach differs from the sequential approach.

Initial Data

Let's say we have images of:

```
A1: Dog (German Shepherd)
A2: Dog (Poodle)
A3: Dog (Bulldog)
B1: Cat (Siamese)
B2: Cat (Persian)
B3: Cat (Tabby)
C1: Bird (Eagle)
C2: Bird (Parrot)
C3: Bird (Penguin)
D1: Fish (Goldfish)
D2: Fish (Shark)
D3: Fish (Tuna)
E1: Horse
E2: Cow
E3: Pig
```

Step 1: Global Hierarchical Clustering

First, AgglomerativeClustering looks at ALL images at once and groups them by feature similarity:

Cluster Alpha: [A1, A2, A3, E1, E2] (Large four-legged animals) Cluster Beta: [B1, B2, B3, E3] (Small four-legged animals) Cluster Gamma: [C1, C2, C3, D1, D2, D3] (Flying/Swimming animals)

Step 2: Ensure Internal Diversity

For each cluster, select most diverse members if > 5 images:

Cluster Alpha Processing:

Has 5 images, calculate internal diversity:

```
A1 ↔ A2: 0.7 similarity (similar breed features)
A1 ↔ E1: 0.3 similarity (different species)
```

Keep all 5 as is: [A1, A2, A3, E1, E2]

Cluster Beta Processing:

Has 4 images, keep all: [B1, B2, B3, E3]

Cluster Gamma Processing:

Has 6 images, select most diverse 5:

1. Start with C1 (Eagle)
2. Add D2 (most different - Shark)
3. Add C3 (most different from both - Penguin)
4. Add D1 (Goldfish)
5. Add C2 (Parrot) Final: [C1, D2, C3, D1, C2]

Key Differences from Sequential Approach

1. Global View First:
    - Sequential: Might group all dogs together
    - Global: Recognizes larger patterns (land/sea/air animals)
2. Natural Groupings:
    - Sequential: Arbitrary first choices affect all clusters
    - Global: Groups form based on feature similarity
3. Better Distribution:
    - Sequential: Later clusters get leftover images
    - Global: All clusters formed considering full dataset

Example Output Structure

```
Cluster 1/
    - German_Shepherd.jpg
    - Poodle.jpg
    - Bulldog.jpg
    - Horse.jpg
    - Cow.jpg
    cluster_dendrogram.png
    distance_matrix.csv

Cluster 2/
    - Siamese.jpg
    - Persian.jpg
    - Tabby.jpg
    - Pig.jpg
    cluster_dendrogram.png
    distance_matrix.csv

Cluster 3/
    - Eagle.jpg
    - Shark.jpg
    - Penguin.jpg
```

```
    - Goldfish.jpg
    - Parrot.jpg
    cluster_dendrogram.png
    distance_matrix.csv
```

This approach ensures:

1. Naturally related images tend to cluster together
2. Each cluster still maintains internal diversity
3. Clusters are meaningfully different from each other
4. No "leftover" effect in later clusters