## Deliverable 3

For this task, we used LIME (Local Interpretable Model-agnostic Explanations) to visualize and analyse the decision regions of a pre-trained ResNet-50 classifier across 10 ImageNet images. The aim was to understand which parts of each image contributed most to the model's prediction, and to optimize LIME's settings for both interpretability and computational efficiency.

### Approach and Best Parameters

We automated a grid search over LIME hyperparameters, evaluating each configuration based on the average Intersection over Union (IoU) and runtime per explanation. The best configuration, striking a balance between accuracy and speed, used these settings:

- **num_features:** 12 (number of superpixels in the explanation)
- **num_samples:** 600 (perturbed samples generated per explanation)
- **segmentation_fn:** quickshift with kernel_size=4, max_dist=150, ratio=0.3
- **batch_size:** 10
- **distance_metric:** 'cosine'
- **model_regressor:** Ridge regression (alpha=0.5)
- **random_seed:** 42

With these parameters, we achieved an average IoU of **0.3305** and an average explanation time of **2.8 seconds** per image.

The LIME visualizations revealed strong and consistent patterns depending on the type of object and the image background. For images where the main object was large and centrally located—such as **goldfish**, **orange**, **tiger shark**, or **West Highland white terrier**—the explanations were sharply focused. The "Explained Only" masks in these cases matched the main subject very closely, confirming that both LIME and the underlying model rely primarily on the target object for their decision. Background regions, even when complex or visually rich, were largely ignored in the explanations.

In cases featuring multiple similar objects, such as **flamingo** and **American coot**, LIME distributed the mask across all the primary instances, successfully capturing each relevant subject even in the presence of reflections or cluttered environments.

For man-made objects like **racer** (car), LIME's attention centered on the most distinctive parts—such as the car's body and wheels—while avoiding less important areas like the racetrack or scenery.

However, in some images, the object boundaries were thin, fragmented, or overlapped with backgrounds that shared similar visual cues. Notably, in the **kite** and **vulture** images, some superpixels selected by LIME extended into branches, sky, or rooftop, rather than being restricted to just the bird. This was especially noticeable in the **kite** image, where both object and background have fine, intertwined details, and in **vulture**, where part of the rooftop was also included in the explanation. These cases highlight the challenges faced by LIME (and the model itself) in segmenting thin or overlapping objects and show that interpretability can be more difficult when backgrounds are visually like the target.

Overall, the clearest, most interpretable LIME masks appeared in images with distinct, well-separated subjects, while scenes with visual ambiguity or fine structures led to broader, less precise explanations. While tuning, we also observed an important trade-off between the average IoU and computation time. Some parameter settings gave even better IoU values, but the explanation time per image increased significantly—sometimes several times higher than the best-balanced configuration. For practical use, it was important to find a setup that gave both strong IoU and reasonable speed. Lowering the number of samples helped speed up computation without sacrificing explanation quality, and increasing the granularity of superpixels (with higher kernel size and max_dist) produced clearer, more relevant explanations.