Interactive session - LIME

In LIME, what is the order of the high-level steps?

- **Generate neighborhood → Get predictions → Weight by proximity → Train interpretable model → Explain**
- Train model → Generate neighborhood → Weight by proximity → Get predictions → Explain
- Get predictions → Generate neighborhood → Train interpretable model → Weight by proximity → Explain
- Generate neighborhood → Train interpretable model on original labels → Weight by proximity → Explain

What is an "interpretable representation" in LIME for images?

- **Superpixel/patch segments encoded as a binary vector**
- The gradient map of the image
- The raw pixel matrix (WxHxC)
- A learned embedding from the neural network

In the LIME objective — explanation(x) = argmin L(f, g, πx) + Ω(g) — what does Ω(g) represent and why is it minimized?

- The proximity between x and perturbed samples — minimized to focus on nearby points
- The prediction error of the original model f — minimized to improve accuracy
- **The complexity of the surrogate model g — minimized to keep explanations interpretable**
- The number of perturbed samples generated — minimized to reduce computation time

Why might LIME produce unrealistic neighbor samples, and what is the main reason for this?

- Because the linear surrogate model is too simple to capture complex boundaries
- **Because perturbations are generated independently per feature, ignoring correlations between features**
- Because proximity is measured using cosine similarity instead of Euclidean distance
- Because the number of neighbors sampled is too small

A data scientist runs LIME twice on the same instance and gets different explanations. Which of the following actions would most directly reduce this problem?

- Switch from a linear surrogate to a decision tree
- **Increase the number of perturbed samples generated for the neighborhood**
- Use a smaller value of K (number of interpretable features)
- Replace cosine similarity with Euclidean distance as the proximity measure

Which of the following best describes the trade-off captured by the full LIME objective L(f, g, πx) + Ω(g)?

- The trade-off between model accuracy on training data and generalization to unseen data

- **The trade-off between faithfully approximating the black-box locally and keeping the surrogate model simple enough to interpret**
- The trade-off between the size of the neighborhood and the speed of computation
- The trade-off between using interpretable features for explanation and raw features for training