The 38th Annual AAAI Conference on Artificial Intelligence

# Using Stratified Sampling to Improve LIME Image Explanations

## Authors:

UNIVERSITÀ DI TORINO

Elvio G. Amparore
elviogilberto.amparore@unito.it

Muhammad Rashid
muhammad.rashid@unito.it

RuleX

Enrico Ferrari
enrico.ferrari@rulex.ai

Damiano Verda
damiano.verda@rulex.ai

## Fundings:

EU Horizon-2020 ECSEL-JU project

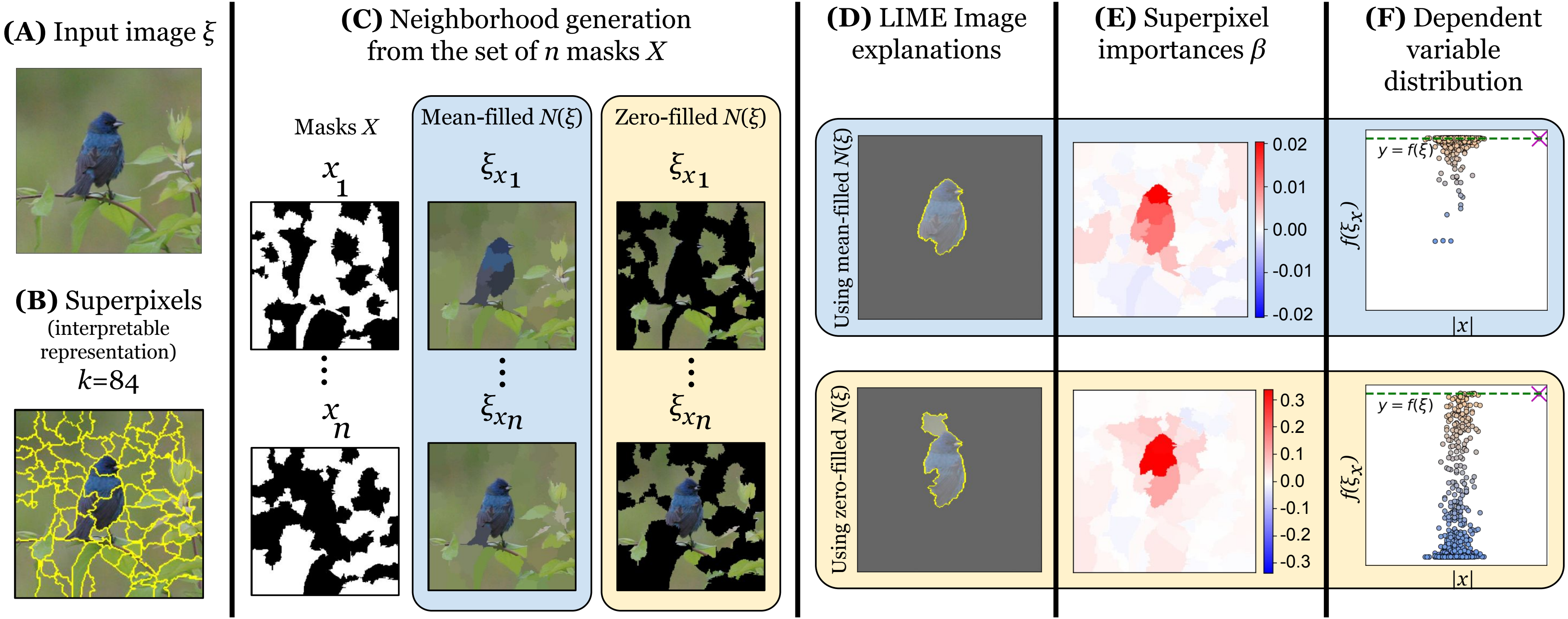
NEXT PERCEPTION

https://nextperception.eu/

# Topic of the presentation

- Consider linear explanations of **image data**
- Many popular techniques: Grad-CAM, SHAP, LIME, etc…
- We focus on **model-agnostic** systems that generate linear explanations by probing a black-box model using perturbations of the image input.
- Focus on **LIME Image** and its **sampling strategy**.
- LIME Image employs **Monte Carlo sampling** to generate synthetic neighborhoods

# How LIME Image works



# Synthetic neighborhood generation

Initial image ξ $\longrightarrow$ divided into $k$ superpixels

Superpixel masking: $x \in \{0,1\}^k$ generates a perturbed image $\xi_x$

In LIME Image, masks are sampled using an **unbiased Monte Carlo** strategy:

$$x[i] \sim B(0.5), \qquad 1 \leq i \leq k$$

where $B$(p) is a Bernoulli-distributed random variable having probability p = 0.5

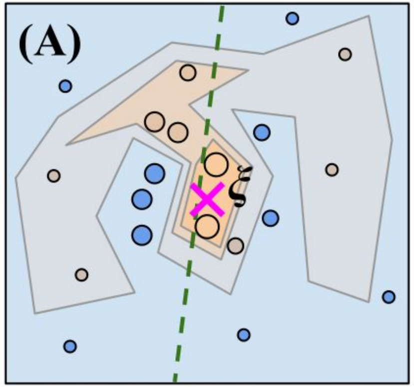
$N$(ξ) = synthetic neighborhood of n perturbed images

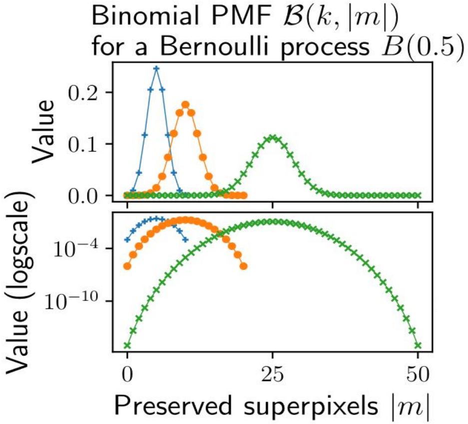
Dependent variables: $Y = \{ f(\xi_x) \mid \xi_x \in N(\xi) \}$

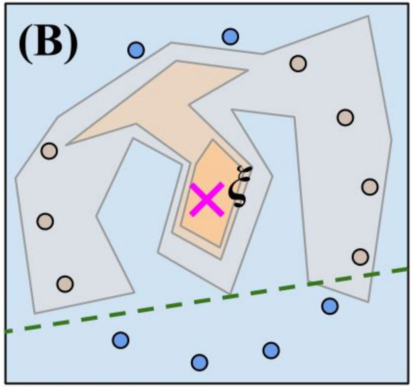
**black box model being explained**

# The problem with the unbiased Bernoulli distribution

- Ideally, the synthetic neighborhood $N(\xi)$ should provide a "good enough" coverage of the variations around $x$.

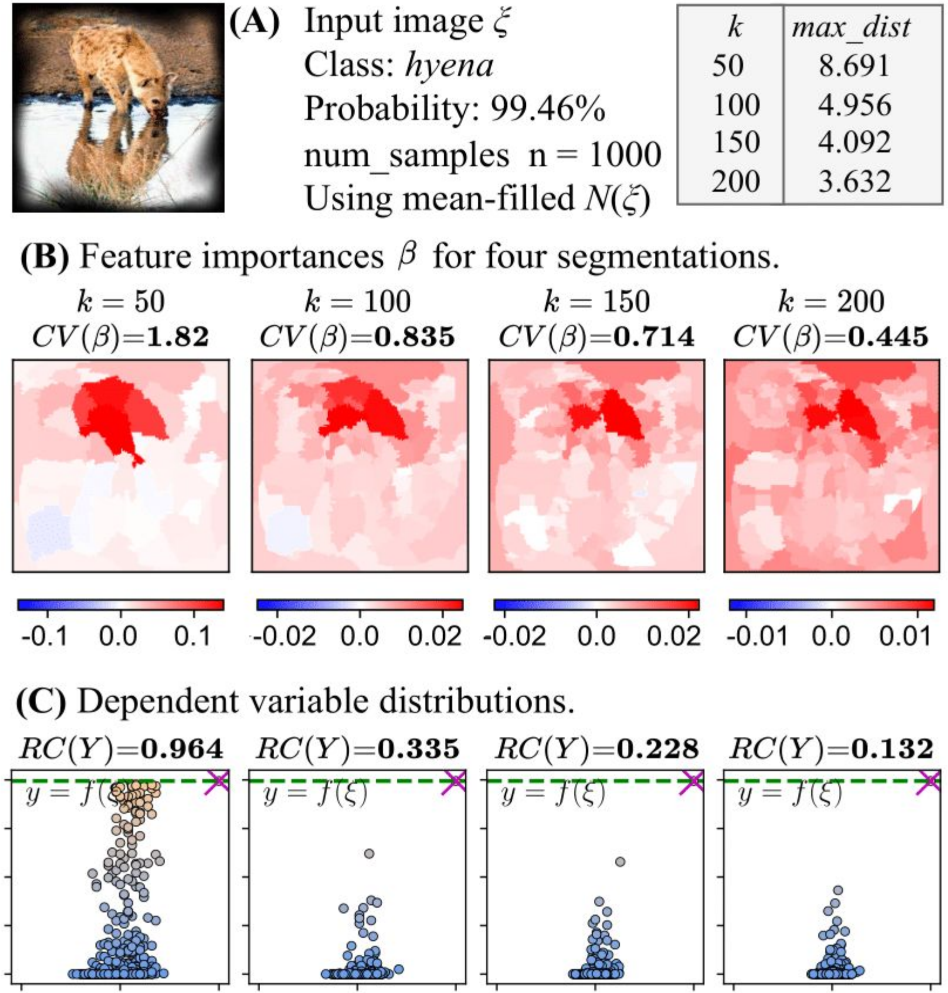



- May result in **under-representation** of the neighborhood

* Image freely inspired by: Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD int. Conf. 1135–1144.




# Dependent variable undersampling results in confused explanations

**C**oefficient of **V**ariation of the explanation $\beta$

$$CV(\beta) = \frac{\sigma_\beta}{\mu_\beta}$$

**R**ange **C**overage of the values of the dependent variable Y in the synthetic neighborhood.

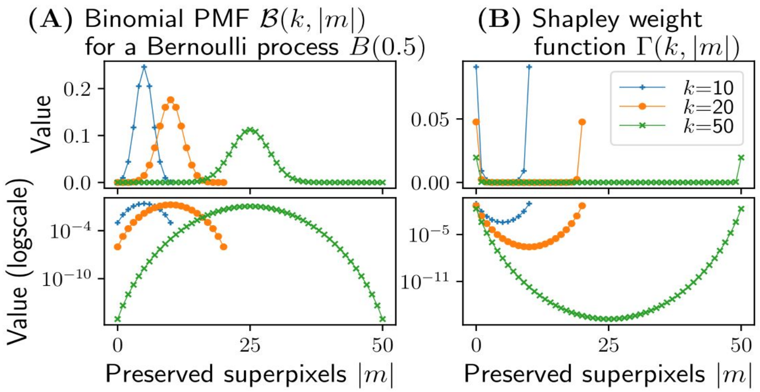
$$RC(Y) = \frac{IQR_{1-99}(Y)}{f(\xi)}$$

- Almost no sample is close to ξ.
- Samples drawn from $B$(p) have all about 50% of the superpixels masked.
- Under-representation of the local behaviour of the black-box model $f$.

# Sampling relevance

Bernoulli distribution is not the only option.

Shapley theory uses a different distribution:



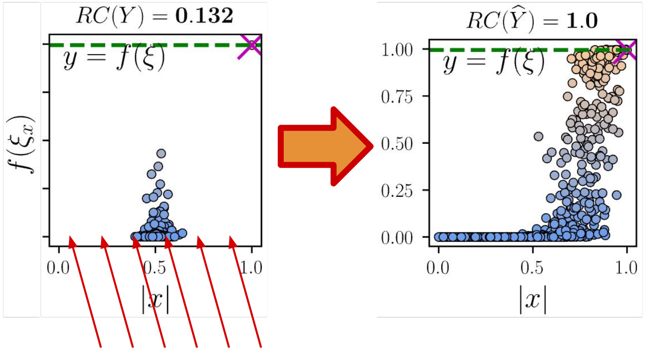

Shapley Importance function

$$\Gamma(k, |x|) = \frac{1}{(k+1)\binom{k}{|x|}}$$

Shapley Importance is the reciprocal of the Bernoulli distribution

$$\mathcal{B}(k, |x|) \cdot \Gamma(k, |x|) = \frac{\binom{k}{|x|} p^{|x|}(1-p)^{k-|x|}}{(k+1)\binom{k}{|x|}} = \frac{0.5^k}{k+1}$$

# Proposed methodology: **stratified sampling**




The goal is to draw samples to cover the entire |x| sampling space

- Consider a stratified partitioning
- $\mathcal{X}^{(i)}$ = set of possible masks having $|x| = i$
- Stratum i size is known a priori:

$$|\mathcal{X}^{(i)}| = \binom{k}{i}, \qquad 0 \leq i \leq k$$

- Oversampled probability

$$Prob\{x \in \mathcal{X}^{(i)} \mid x \in \hat{X}\} = \frac{1}{k+1}$$

- Adjustment factors

$$adj(i) = \frac{Prob\{x \in \mathcal{X}^{(i)} \mid x \in X\}}{Prob\{x \in \mathcal{X}^{(i)} \mid x \in \hat{X}\}} = \frac{(k+1)\binom{k}{i}}{2^k}$$

# Monte Carlo vs. Stratified Sampling

LIME Image uses a simple linear homoscedastic model

$$Y = X \cdot \beta + \epsilon$$

The explanation coefficients $\beta$ results from

$$\beta = (X^\mathsf{T} W X)^{-1} X^\mathsf{T} W Y$$

$\beta$ coefficients may vary by stratum.
We adopt instead a mixture model

$$\widehat{Y}^{(i)} = \widehat{X}^{(i)} \cdot \widehat{\beta}^{(i)} + \widehat{\epsilon}^{(i)}$$

The explanation coefficients $\beta$ results from

$$\widehat{\beta} = (\widehat{X}^\mathsf{T} \widehat{W} \widehat{X})^{-1} \widehat{X}^\mathsf{T} \widehat{W} \widehat{Y}$$

where $\widehat{W}$ accounts for the adjustment factors that correct the bias introduced by oversampling the distribution tails.

# Impact of stratified sampling in LIME Image

**Case (A):** The mean and variance of $\widehat{\beta}^{(i)}$ are independent from the strata.

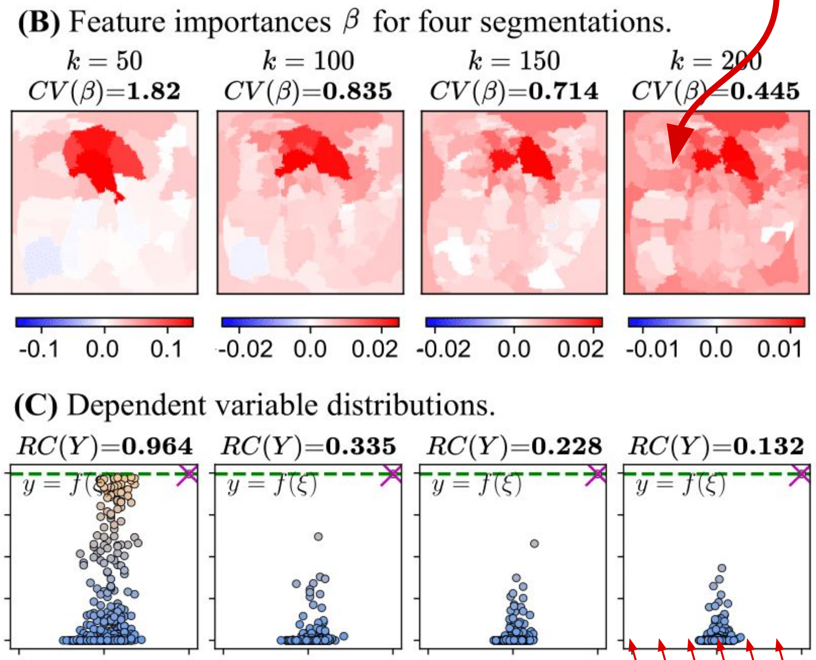
⇒ weighted regression model is not needed.
Monte Carlo and stratified sampling should behave similarly.
Unlikely to happen using complex black-box machine learning models.

**Case (B):** The mean and variance of $\widehat{\beta}^{(i)}$ varies by stratum.

⇒ weighted regression model is highly advisable (DuMouchel and Duncan 1983).
Monte Carlo will perform badly, stratified sampling is relevant.
Common scenario for complex models and/or large number of superpixels.

# Monte Carlo sampling

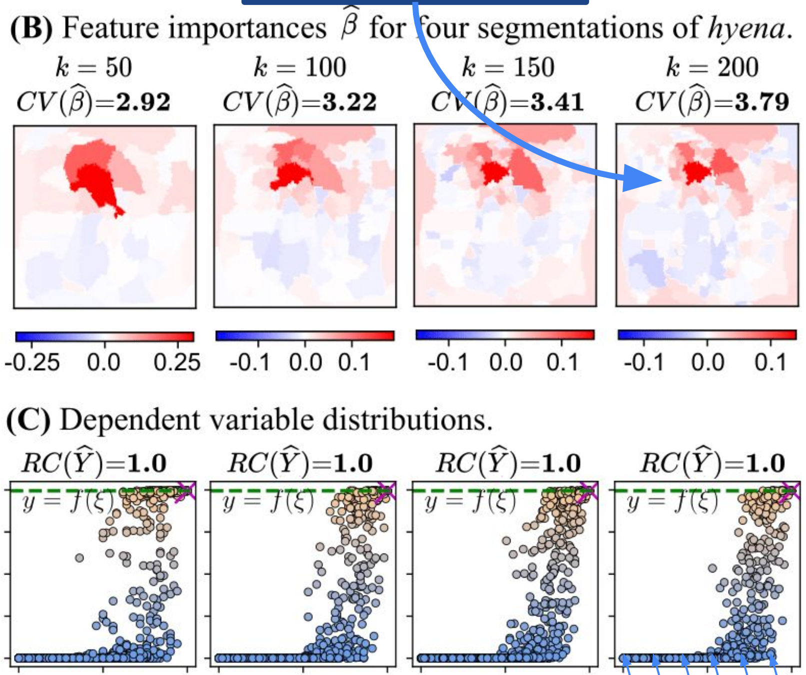
confused explanation




undersampling

# Stratified sampling

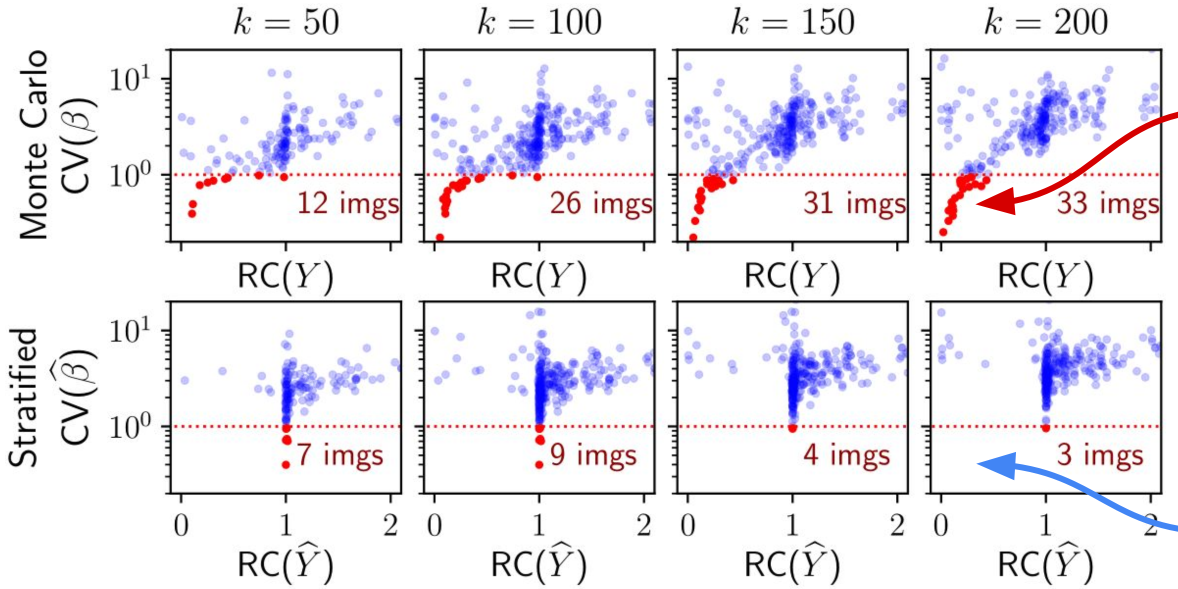
feature attribution is meaningful




samples from all strata

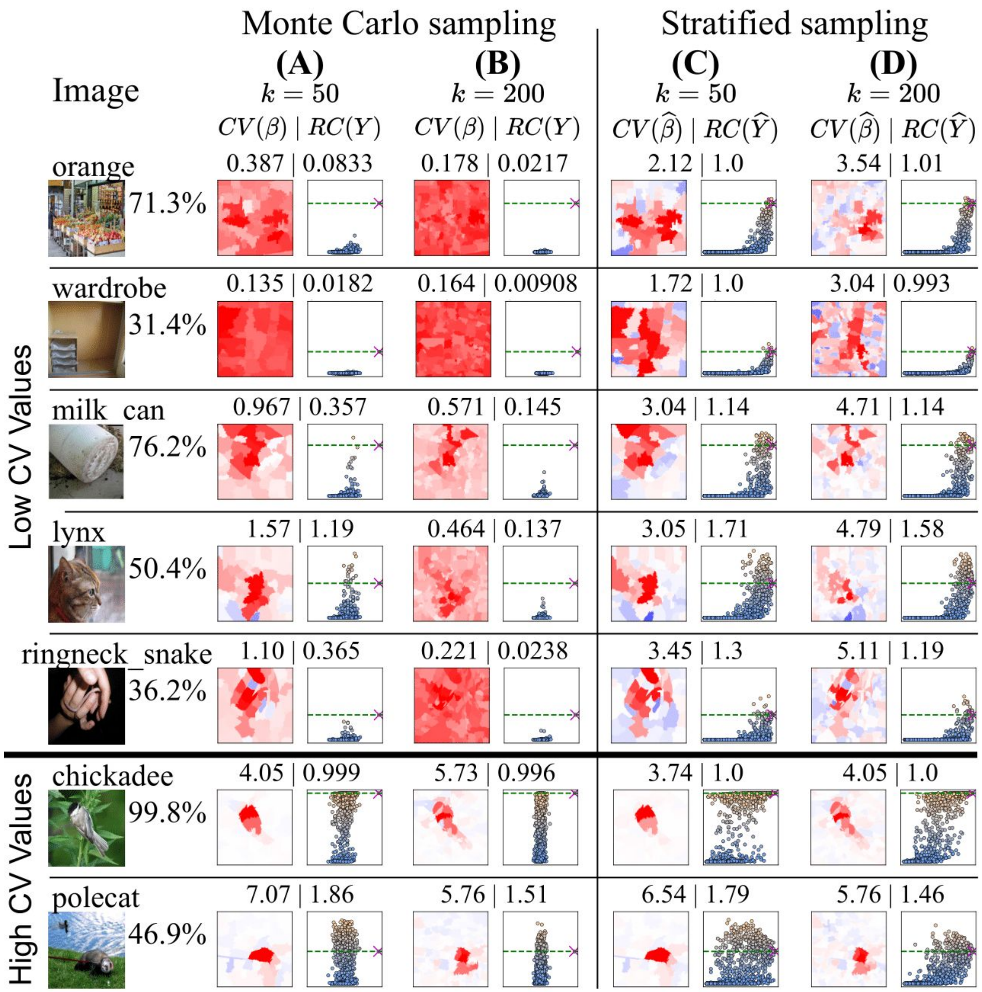
# Evaluation on the *ImageNet Object Localization Dataset*

150 images; four different choices of superpixels ($k$ = 50, 100, 150, 200)
n=1000 samples (average of 10 runs). Using *ResNet-50* model.




About 1 image out of 5 suffers from severe undersampling using the default Monte Carlo sampling of LIME-Image

Misbehaviours are corrected using stratified sampling



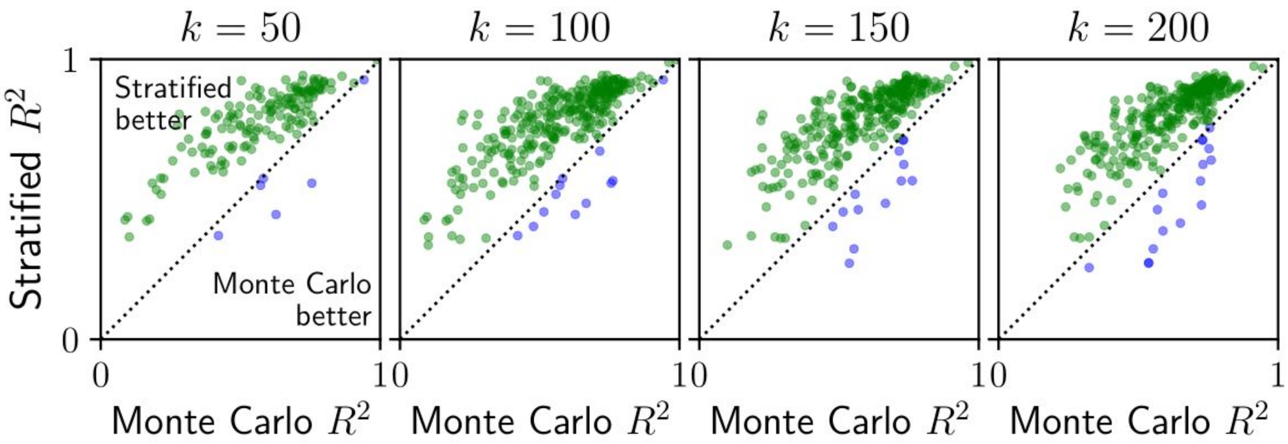

# Evaluation on the *ImageNet Object Localization Dataset*

Inspection shows that poor sampling of synthetic neighborhoods using unbiased Monte Carlo are not rare.

Unsurprisingly, explanations built using these (poor) neighborhoods do not identify the relevant portions of the image

# Evaluation on the *ImageNet Object Localization Dataset*

In general the dependent variable sampled using Stratified Sampling is a better explainer than the one sampled using Monte Carlo (tested using average $R^2$ coefficients of the linear regressors built by LIME).



# Conclusions

- Reformulation of LIME Image sampling strategy (not restricted to image data) for stratified sampling.
- Drawing lessons from the Shapley theory.
- Empirical evaluation shows that Monte Carlo undersampling is not rare, and stratified sampling provides practical improvements, at no additional cost.

**Possible improvements**

- Consider regularization factors for ridge regression.
- Mixed model could be improved using uniform weights for the strata.

**Code Availability**

- https://github.com/rashidrao-pk/lime_stratified
- https://github.com/rashidrao-pk/lime-stratified-examples

# Using Stratified Sampling to Improve LIME Image Explanations

# Thank you

## Authors:


UNIVERSITÀ DI TORINO

Elvio G. Amparore
elviogilberto.amparore@unito.it

Muhammad Rashid
muhammad.rashid@unito.it

RuleX

Enrico Ferrari
enrico.ferrari@rulex.ai

Damiano Verda
damiano.verda@rulex.ai


## Fundings:


EU Horizon-2020 ECSEL-JU project

NEXT PERCEPTION

https://nextperception.eu/