Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Presenter:
Md Mahfuzur Rahman
Postdoctoral Research Associate
TReNDS Center
Georgia State University

Georgia State University

TReNDS
Center for Translational Research in Neuroimaging & Data Science

About the paper

Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond

Anna Hedström[1,†] ANNA.HEDSTROEM@TU-BERLIN.DE
Leander Weber[3] LEANDER.WEBER@HHI.FRAUNHOFER.DE
Dilyara Bareeva[1] DILYARA.BAREEVA@CAMPUS.TU-BERLIN.DE
Daniel Krakowczyk[4] DANIEL.KRAKOWCZYK@UNI-POTSDAM.DE
Franz Motzkus[3] FRANZ.MOTZKUS@HHI.FRAUNHOFER.DE
Wojciech Samek[2,3,5] WOJCIECH.SAMEK@HHI.FRAUNHOFER.DE
Sebastian Lapuschkin[3,†] SEBASTIAN.LAPUSCHKIN@HHI.FRAUNHOFER.DE
Marina M.-C. Höhne[1,5,†] MARINA.HOEHNE@TU-BERLIN.DE

Journal of Machine Learning Research 24 (2023) 1-11

Link to the paper

XAI Evaluation Challenges

- No “ground truth”
- No “universally accepted” correctness
- Unknown properties to fulfill

- Conceive experimental ways, BUT:
  - Parameterizations
  - preprocessing
  - Normalizations

- ❑ Contrasting results
- ❑ One-sided conclusions
- ❑ Questionable procedure

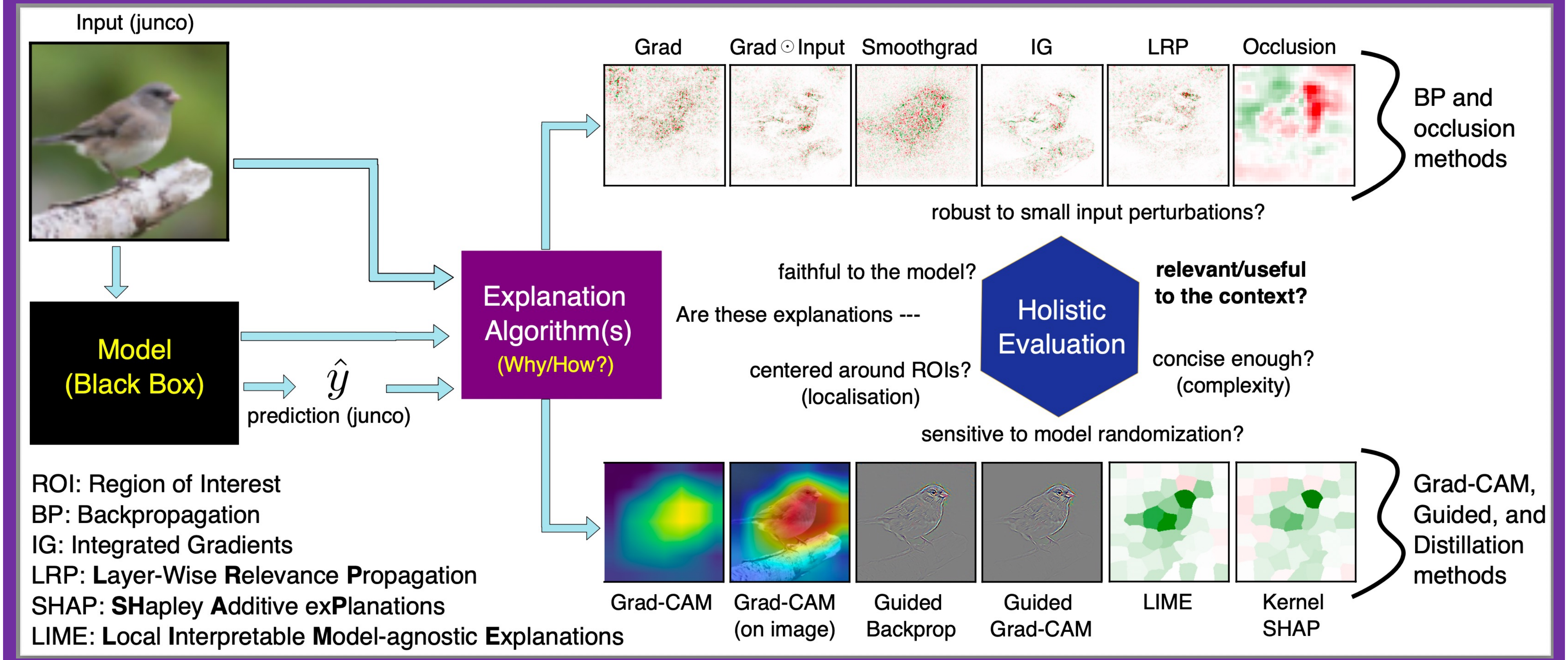
Post hoc explanations and quality evaluation

Input: $\boldsymbol{x} \in \mathbb{R}^d$

Model: $F : \mathbb{R}^d \to \mathbb{R}^C$

Class logit: $F_c(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$

Explanation: $E : \mathbb{R}^d \to \mathbb{R}^d$



Comparison of XAI Libraries

Library	Faithfulness	Robustness	Localisation	Complexity	Axiomatic	Random.
Captum (2)	1	1	0	0	0	0
AIX360 (2)	2	0	0	0	0	0
TorchRay (1)	0	0	1	0	0	0
Quantus (27)	9	4	6	3	3	2

Categories of Metrics

Faithfulness (↑): Do explanations follow the predictive behaviour of the model?

Robustness (↓): Are explanations stable when subject to slight perturbations in the input?

Localization (↑): Is the explainable evidence centered around a ROI?

Categories of Metrics

Complexity (↓): captures to what extent explanations are concise

Randomization (↑): tests to what extent explanations deteriorate

Axiomatic (↑): measures if explanations fulfill certain axiomatic properties

Faithfulness Metrics

Faithfulness Correlation (Bhatt et al., 2020): iteratively replaces a random subset of given attributions with a baseline value and then measuring the correlation between the sum of this attribution subset and the difference in function output.

Pixel Flipping (Bach et al., 2015): captures the impact of perturbing pixels in descending order according to the attributed value on the classification score.

Robustness Metrics

Local Lipschitz Estimate (Alvarez-Melis et al., 2018): tests the consistency in the explanation between adjacent examples (via perturbation or from test pool).

Max-Sensitivity (Yeh et al., 2019): measures the maximum sensitivity of an explanation using a Monte Carlo sampling-based approximation in a neighborhood of radius r.

Localization Metrics

Pointing Game (Zhang et al., 2018): checks whether attribution with the highest score is located within the targeted object

Relevance Rank Accuracy (Arras et al., 2021): measures the ratio of highly attributed pixels within a ground-truth mask towards the size of the ground-truth mask

Complexity Metrics

Sparseness (Chalasani et al., 2020): uses the Gini Index to measure, if only highly attributed features are truly predictive of the model output

Effective Complexity (Nguyen at el., 2020): measures how many attributions in absolute values are exceeding a certain threshold

Randomization Metrics

Model Parameter Randomisation Test (Adebayo et. al., 2018): randomize layers and measure similarity

Random Logit Test (Sixt et al., 2020): computes the distance between the original explanation and the explanation for a random other class

Axiomatic Metrics

Non-Sensitivity (Nguyen at el., 2020): ensures that a method assigns zero-importance only to features to which the model f is not functionally dependent on.

Input Invariance (Kindermans et al., 2017): adds a shift to input, and tests if attributions change in response.

Input Invariance – An Axiomatic Evaluation

As a result the first layer activations are the same for $f_1(x)$ and $f_2(x)$:

$$z = \boldsymbol{w}^T \boldsymbol{x}_2 + b_2 = \boldsymbol{w}^T \boldsymbol{x}_1 + \boldsymbol{w}^T \boldsymbol{m}_2 + b_1 - \boldsymbol{w}^T \boldsymbol{m}_2.$$

Note that the gradient with respect to the input remains unchanged as well:

$$\frac{\partial f_1(\boldsymbol{x}_1^i)}{\partial \boldsymbol{x}_1^i} = \frac{\partial f_2(\boldsymbol{x}_2^i)}{\partial \boldsymbol{x}_2^i}.$$

- Gradient and signal methods satisfy
- Gradient x INPUT is sensitive
- IG (all) and DTD with LRP reference do not satisfy

Kindermans, Pieter-Jan, et al. "The (un) reliability of saliency methods." Explainable AI: Interpreting, explaining and visualizing deep learning (2019): 267-280.

Metric Calculation Example

- Define metric

```
# Define XAI methods and metrics.
xai_methods = list(explanations.keys())
metrics = {
    "Robustness": quantus.AvgSensitivity(
        nr_samples=10,
        lower_bound=0.2,
        norm_numerator=quantus.norm_func.fro_norm,
        norm_denominator=quantus.norm_func.fro_norm,
        perturb_func=quantus.perturb_func.uniform_noise,
        similarity_func=quantus.similarity_func.difference,
        abs=False,
        normalise=False,
        aggregate_func=np.mean,
        return_aggregate=True,
        disable_warnings=True,
    ),
    "Faithfulness": quantus.FaithfulnessCorrelation(
        nr_runs=10,
        subset_size=224,
        perturb_baseline="black",
        perturb_func=quantus.perturb_func.baseline_replacement_by_indices,
        similarity_func=quantus.similarity_func.correlation_pearson,
        abs=False,
        normalise=False,
        aggregate_func=np.mean,
        return_aggregate=True,
        disable_warnings=True,
    ),
}
```

Metric Calculation Example (Contd.)

- Calculate score

```
# Or, run quantification analysis!
results = {method : {} for method in xai_methods}

for method in xai_methods:
    for metric, metric_func in metrics.items():

        print(f"Evaluating {metric} of {method} method.")
        gc.collect()
        torch.cuda.empty_cache()

        # Get scores and append results.
        scores = metric_func(
            model=torchvision.models.mobilenet_v3_small(weights=True).to(device),
            x_batch=x_batch,
            y_batch=y_batch,
            a_batch=None,
            s_batch=s_batch,
            device=device,
            explain_func=explainer_wrapper,
            explain_func_kwargs={
                "method": method,
                "posterior_mean": copy.deepcopy(
                    torchvision.models.mobilenet_v3_small(weights=True)
                    .to(device)
                    .state_dict()
                ),
                "mean": 1.0,
                "std": 0.5,
                "sg_mean": 0.0,
                "sg_std": 0.5,
                "n": 25,
                "m": 25,
                "noise_type": "multiplicative",
                "device": device,
            },
        )
        results[method][metric] = scores
```

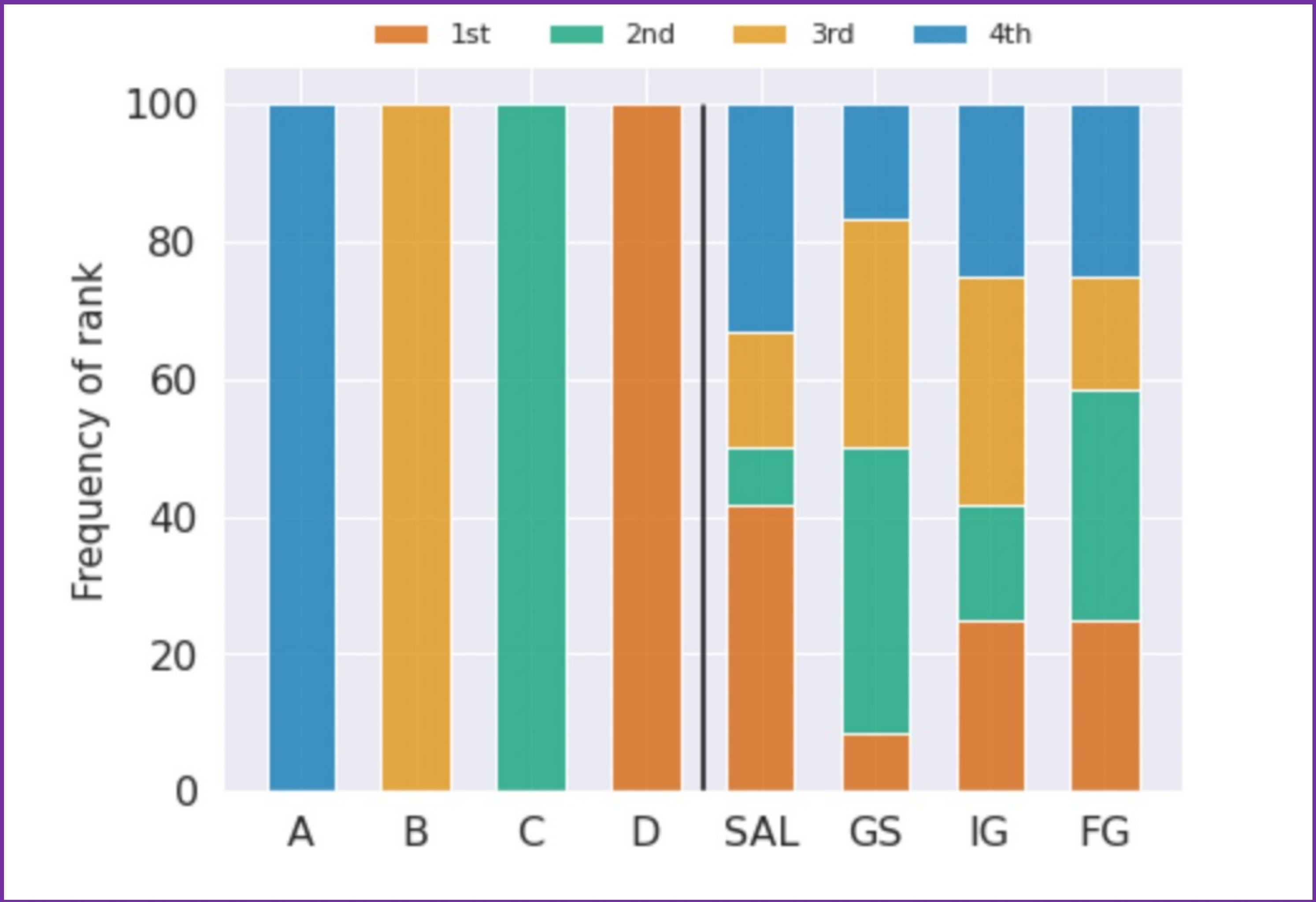
Sensitivity Analysis on Faithfulness Correlation

```python
# Define some parameter settings to evaluate.
baseline_strategies = ["mean", "uniform"]
subset_sizes = np.array([2,  52, 102])
sim_funcs = {"pearson": quantus.similarity_func.correlation_pearson, "spearman": 
quantus.similarity_func.correlation_spearman}

result = {
    "Faithfulness score": [],
    "Method": [],
    "Similarity function": [],
    "Baseline strategy": [],
    "Subset size": [],
}

# Score explanations!
for b in baseline_strategies:
    for s in subset_sizes:
        for method, attr in explanations.items():
            for sim, sim_func in sim_funcs.items():
                metric = quantus.FaithfulnessCorrelation(abs=True,
                                                        normalise=True,
                                                        return_aggregate=True,
                                                        disable_warnings=True,
                                                        aggregate_func=np.mean,
                                                        normalise_func=quantus.normalise_func.normalise_by_negative,
                                                        nr_runs=10,
                                                        perturb_baseline=b,
                                                        perturb_func=quantus.perturb_func.baseline_replacement_by_indices,
                                                        similarity_func=sim_func,
                                                        subset_size=s)
                score = metric(model=model.cuda(), x_batch=x_batch.cpu().numpy(), y_batch=y_batch.cpu().numpy(), 
a_batch=attr, device=device)
                result["Method"].append(method)
                result["Baseline strategy"].append(b.capitalize())
                result["Subset size"].append(s)
                result["Faithfulness score"].append(score[0])
                result["Similarity function"].append(sim)
```

Sensitivity Analysis on Faithfulness Correlation



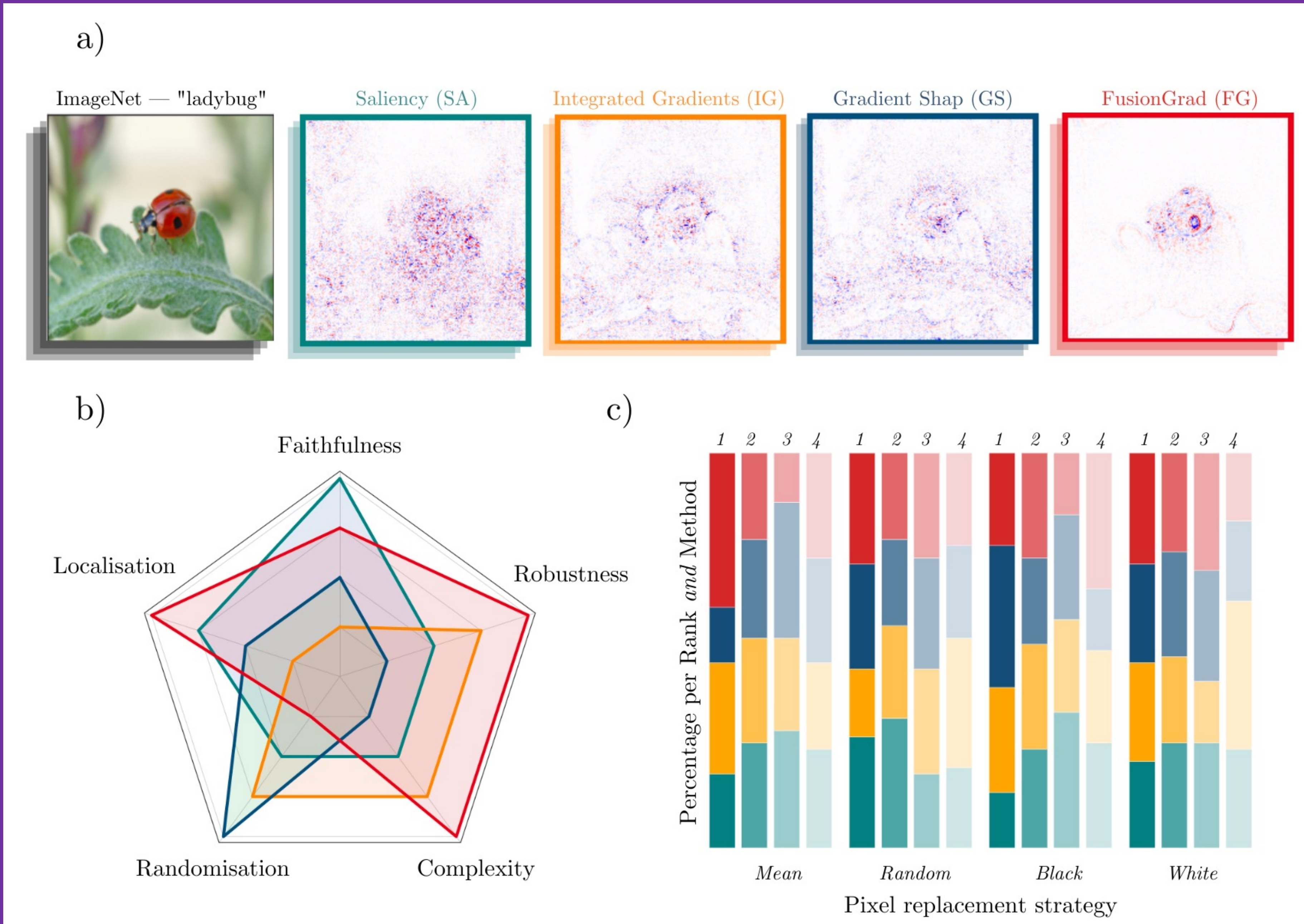
ranking significantly differs depending on parameters

Pixel Flipping Example

```
import quantus
pixelflipping = quantus.PixelFlipping(perturb_baseline="black", abs=True)
scores = pixelflipping(model, x_batch, y_batch, a_batch, **params)
pixelflipping.plot(y_batch=y_batch, scores=scores)
```

Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

Results on Pixel Flipping



Conveniences and Caveats

- Highly customizable and easily extendable
  - Supports new metrics
  - Customization of existing metrics
- API documentation available
- supporting functions replaceable by users
- Outcomes are highly sensitive to the parameters.
- Cautionary support is available.

- No one-size-fits-all metric
- Contextual calibration required: the application, data, model, and intended stakeholders

Human-centered evaluation – An Invitation

Explanations must make sense to humans

What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods

Julien Colin[1,3,5 * †] **Thomas Fel**[1,3,4 *] **Rémi Cadène**[1,2 ‡] **Thomas Serre**[1,3]
[1]Carney Institute for Brain Science, Brown University, USA [2]Sorbonne Université, CNRS, France
[3]Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
[4] Innovation & Research Division, SNCF
[5] ELLIS Alicante, Spain
{julien_colin,thomas_fel,remi_cadene}@brown.edu

Caution!! Humans are questionable judges

Questions?

Contact:
mrahman21@gsu.edu

Georgia State University

TReNDS
Center for Translational Research in Neuroimaging & Data Science