DEEL

DEpendable & Explainable Learning

Presentation given by Lucas Hervier and Antonin Poché

first_name dot last_name at irt-saintexupery dot com

IVADO

IRT SAINT EXUPÉRY

CRIAQ
Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI
Porté par l'Université de Toulouse

Institut intelligence et données

© DEEL- All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Table of content

- Introduction to explainability and Xplique
- What’s new?
- Two demonstations
- Feed-back from an user

© DEEL - All rights reserved to IVADO, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

DEEL

DEpendable & Explainable Learning

Explainability and Xplique

IVADO

IRT SAINT EXUPÉRY

CRIAQ

Consortium de recherche et d'innovation en aérospatiale au Québec

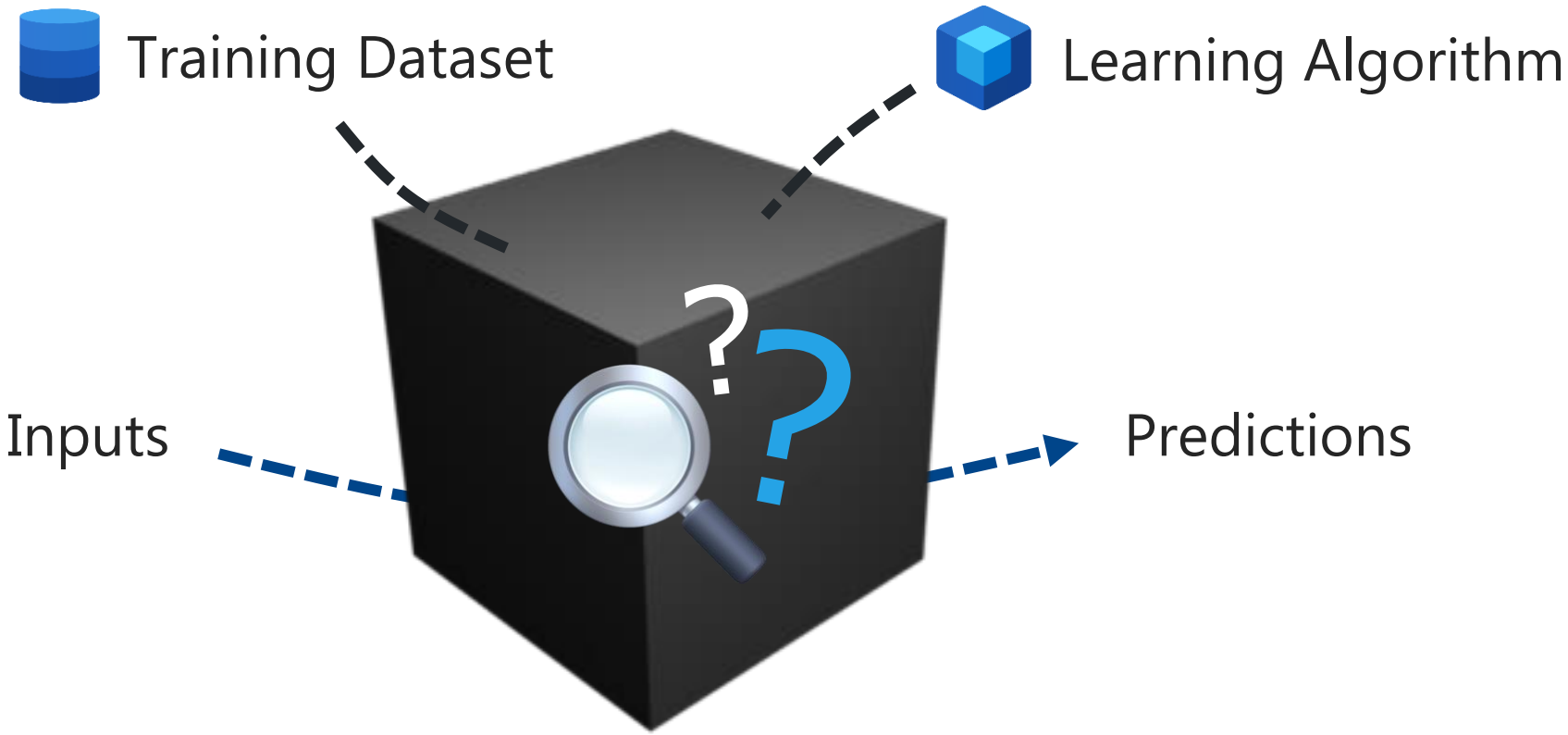
ANITI

Porté par l'Université de Toulouse

Institut intelligence et données

© DEEL- All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

The black-box problem



© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Why do we need explainability?

- Build **trust** in the model predictions

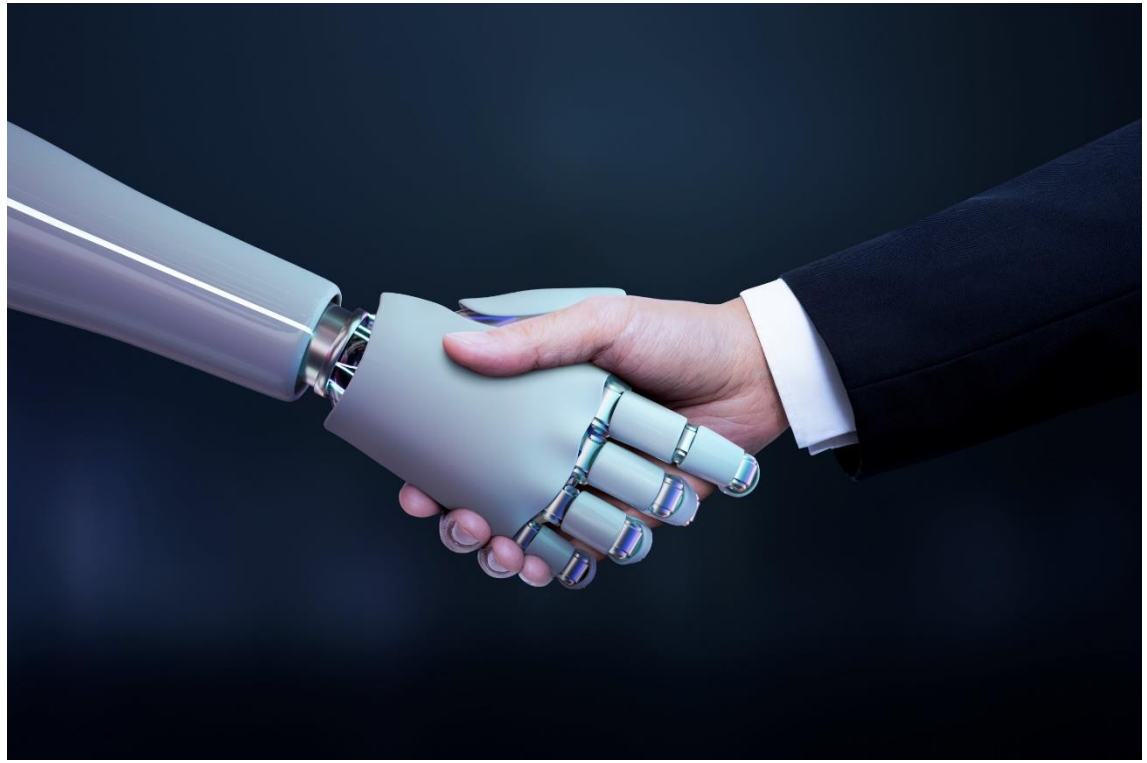


Image by rawpixel.com on Freepik

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Why do we need explainability?

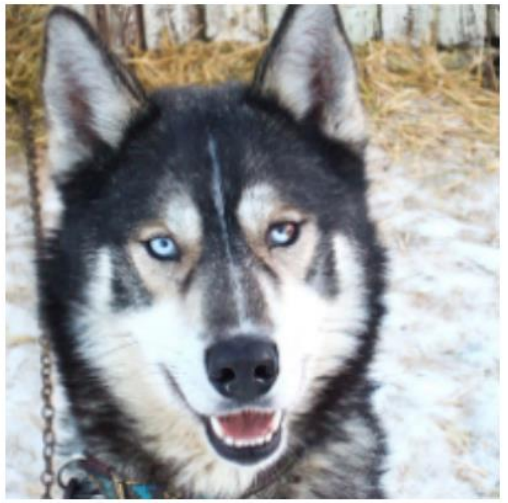
- Build **trust** in the model predictions
- **Satisfy regulatory requirements** and Certification process

Image by rawpixel.com on Freepik

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Why do we need explainability?

- Build **trust** in the model predictions
- **Satisfy regulatory requirements** and Certification process
- **Reveal bias** or other unintented effects learned by a model
- Understand to **intervene on models**
- **…**



(a) Husky classified as wolf

(b) Explanation

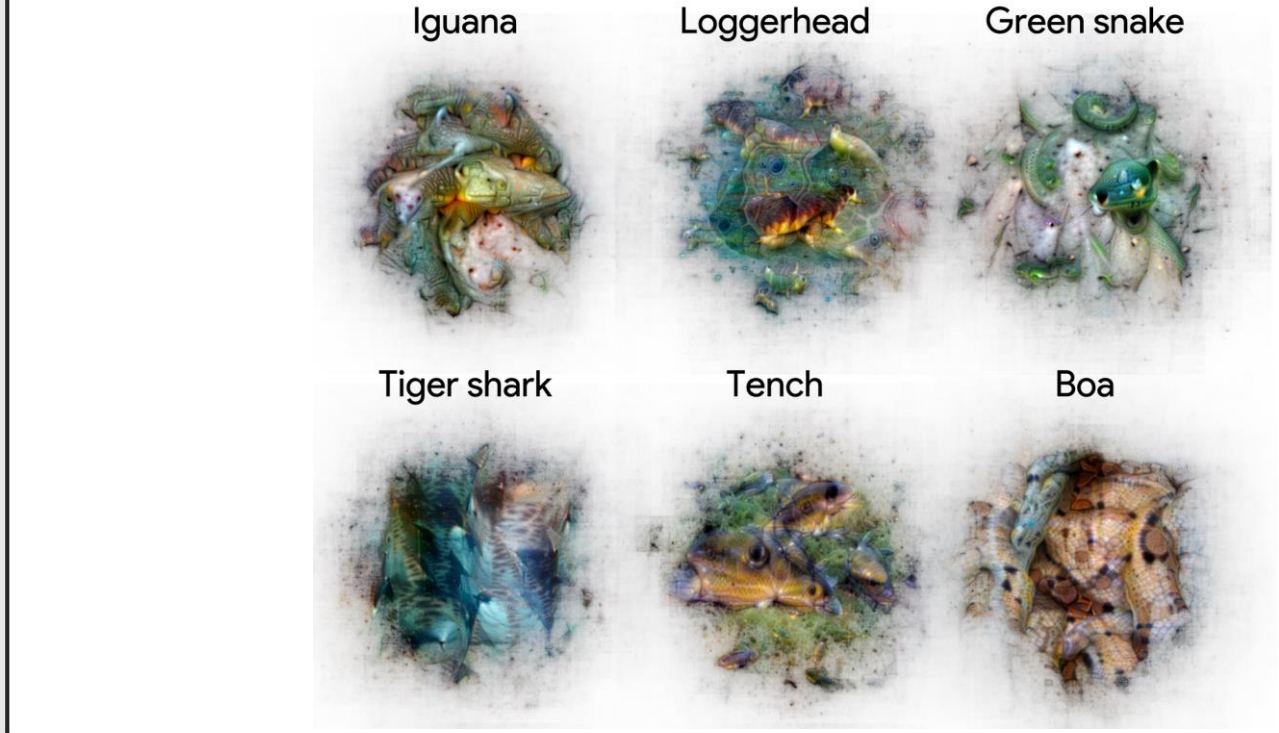
From Ribeiro et al.: "Why Should I Trust You?"

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

A technical challenge

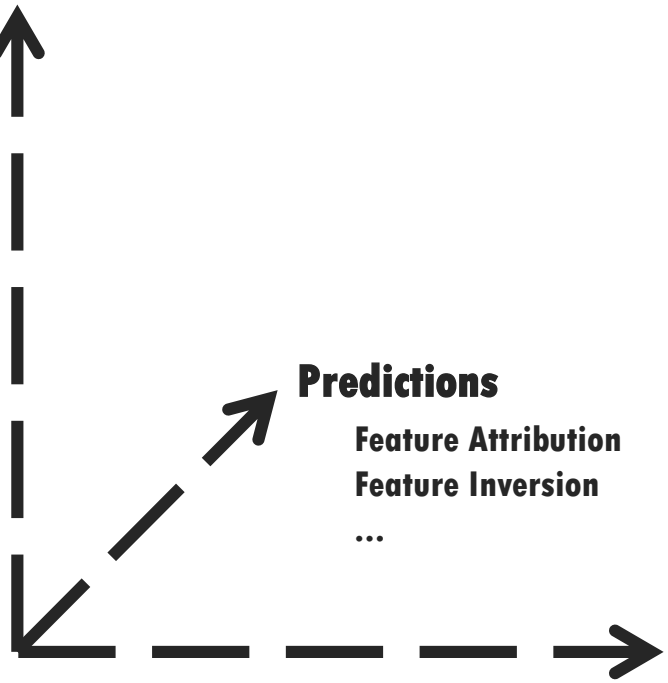
Model

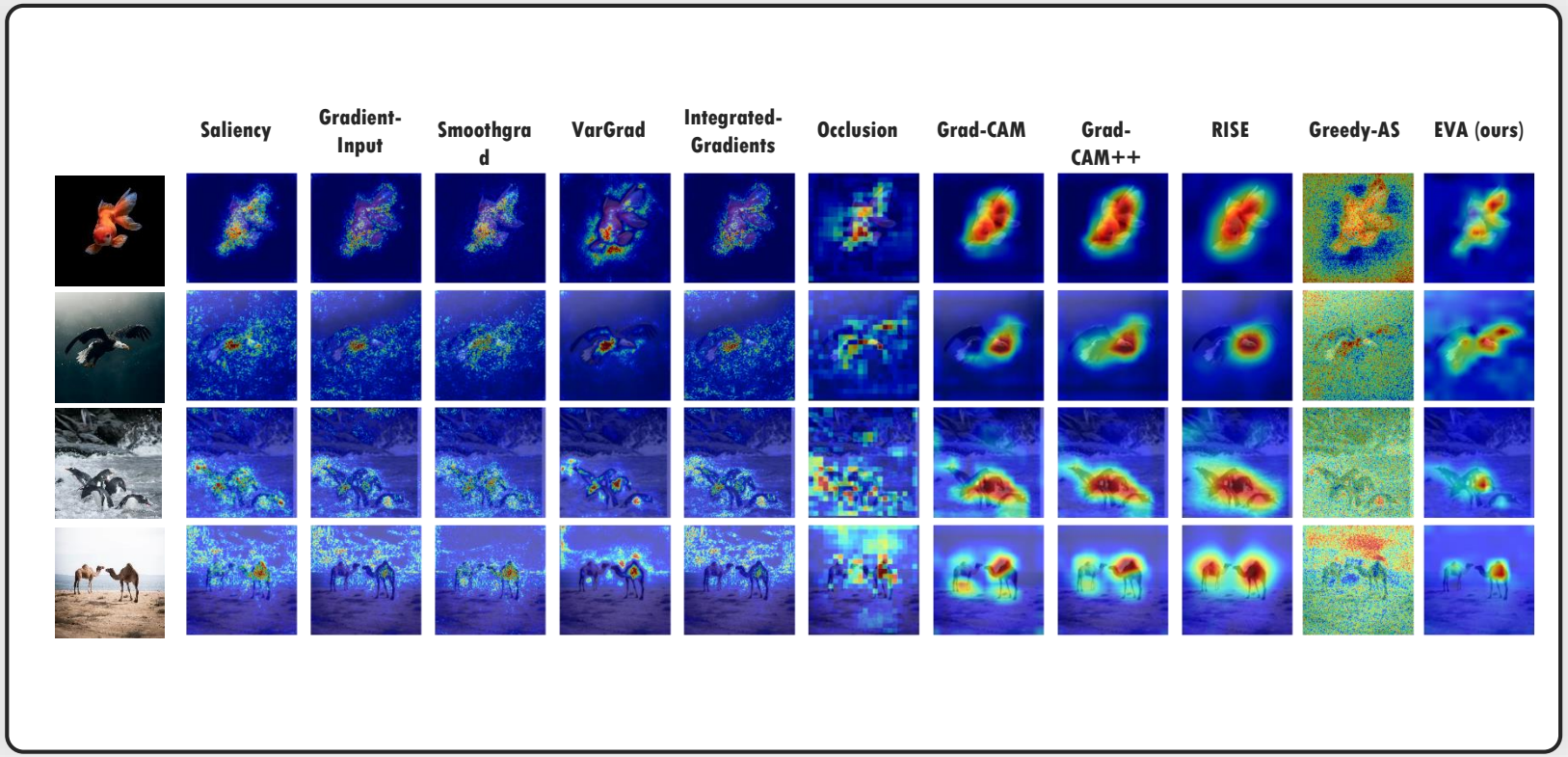
Feature Viz,
Concept Activation Vector
Explanation ‘by design’
…



© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

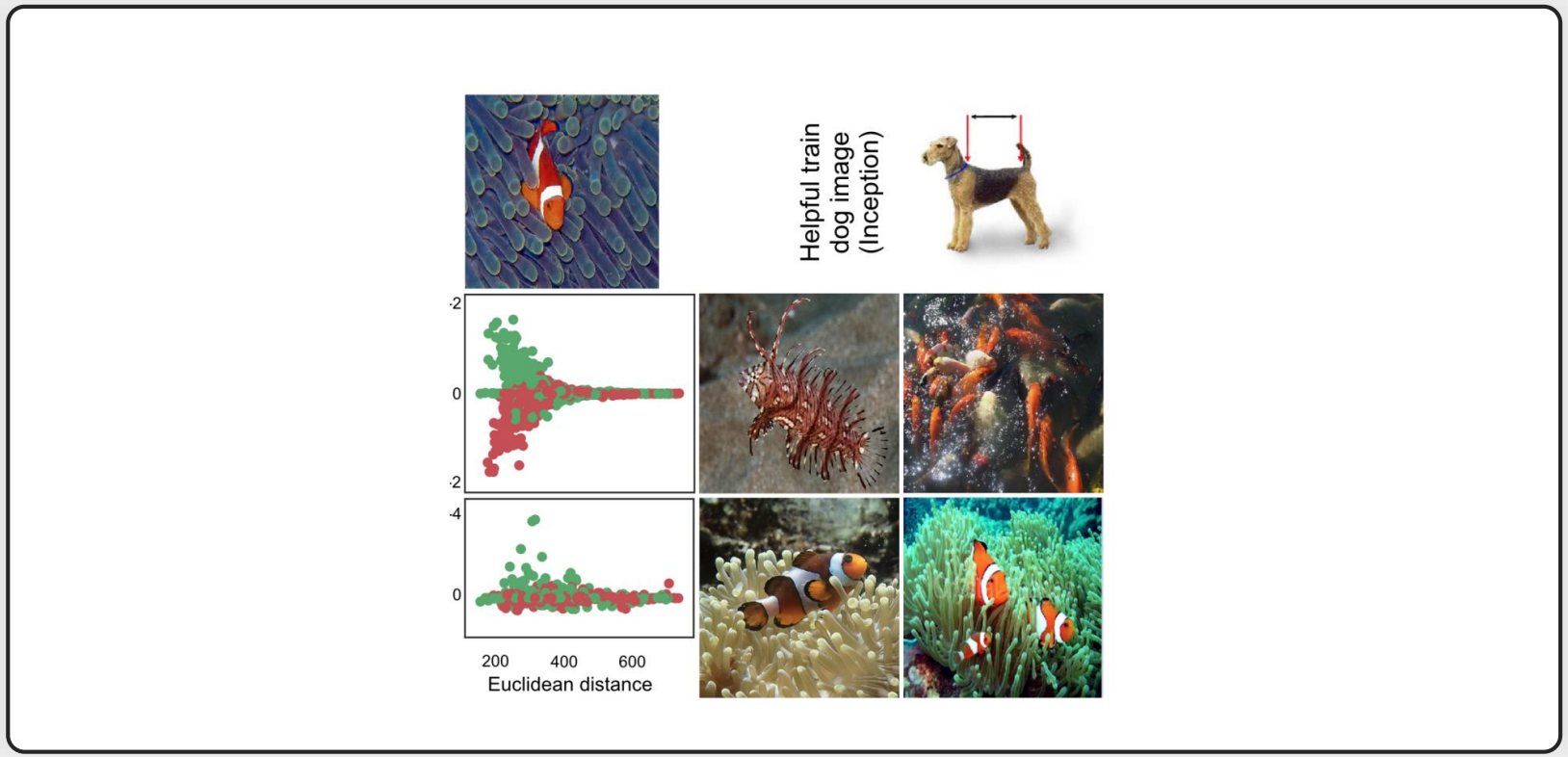
A technical challenge





© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

A technical challenge

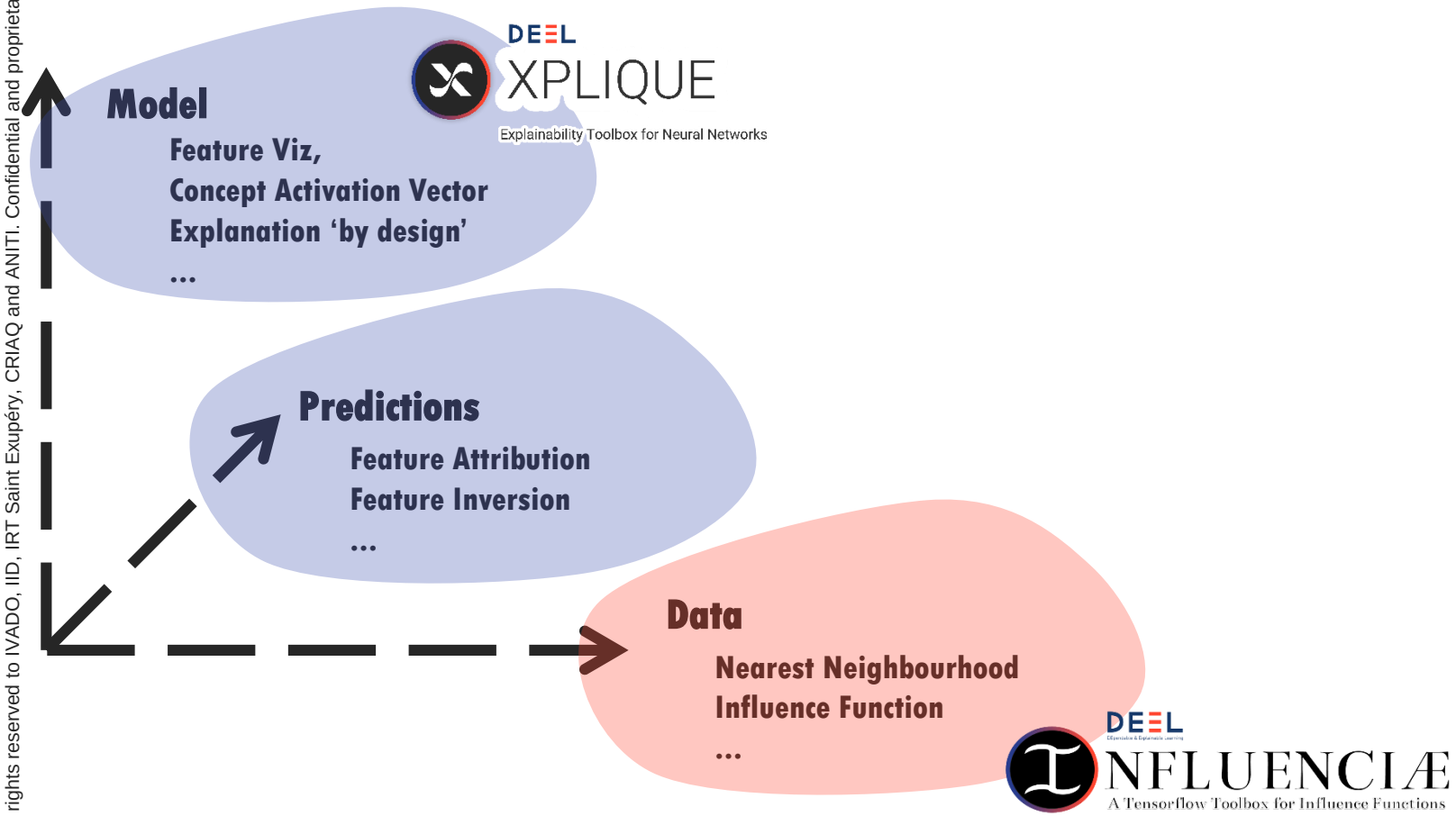


Data

Nearest Neighbourhood
Influence Function
...

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

A technical challenge

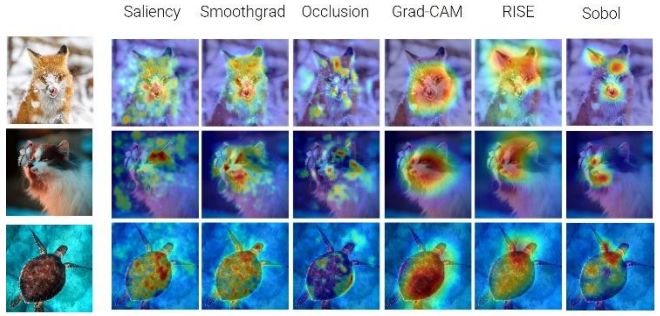


© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Xplique: A DL Explainability Toolbox

DEEL
DEpendable & Explainable Learning

Attribution Methods more than **14** black-box / white-box methods*

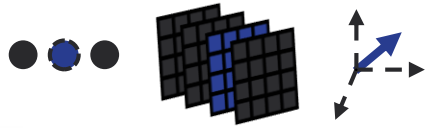



```
from xplique.attributions import GradCAM

explainer = GradCAM(model)
explanations = explainer(x, y)
```

*SOTA

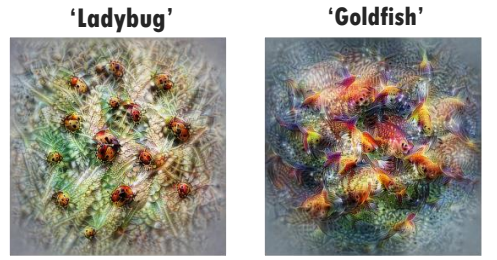
Feature Visualization

• Neurons • Channels • Directions

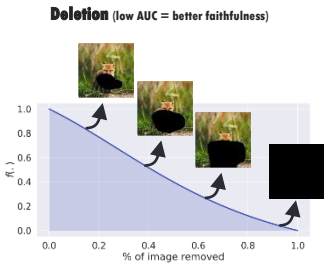


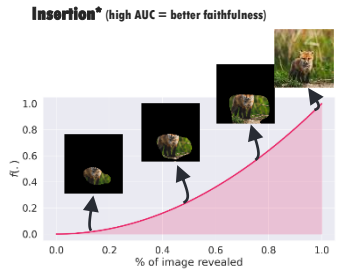
Visualize Neurons, Channels, Vectors in activation space (e.g. CAV) or a mix of them !

```
from xplique.feature_visualization import Objective,
optimize
obj = Objective.neuron(model, 'logits', 10)
images, obj_name = optimize(obj)
```



Metrics more than **6** attributions metrics each supporting multiple baselines

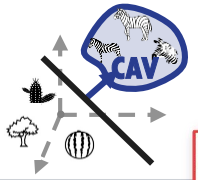


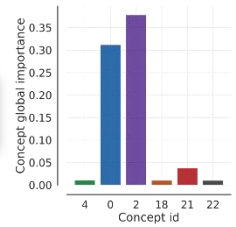



```
from xplique.metrics import Deletion
from xplique.attributions import GradCAM

metric = Deletion(model, x, y)
explanations = GradCAM(model)(x, y)
score = metric(explanations)
```

Concept based concept activation vector, CRAFT (new!)

Easily extract and test CAVs:






```
from xplique.concepts import Cav

extractor = Cav(model, 'mixed3')
concept_vector = extractor(striped_samples,
                           random_samples)
```

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

DEEL

DEpendable & Explainable Learning

What's new?

IVADO

IRT SAINT EXUPÉRY

CRIAQ
Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI
Porté par l'Université de Toulouse

Institut intelligence et données

© DEEL- All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

List of new features

- PyTorch wrapper
- Operators, they allow to treat new tasks:
  - Object detection
  - Semantic segmentation
- New state-of-the-art attribution methods:
  - Sobol, HSIC, and FORGrad
- Automatic concepts extraction
  - CRAFT
- Last feature visualization methods:
  - MaCO

© DEEL - All rights reserved to IVADO, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

PyTorch wrapper and operators API

DEEL
DEpendable & Explainable Learning

Initial API

```python
from xplique.attributions import Saliency
from xplique.metrics import Deletion

# load images, targets, and model
# ...

# initialize the explainer with the model and method parameters
explainer = Saliency(model)

# call the explainer on the sample to explain
explanations = explainer(inputs, targets)

# compute explanation metrics
metric = Deletion(model, inputs, targets)
deletion_score = metric(explanation)
```

PyTorch Wrapper

```python
import torch

from xplique.attributions import Saliency
from xplique.metrics import Deletion
from xplique.wrappers import TorchWrapper

# load images, targets, and torch_model
# ...

device = 'cuda' if torch.cuda.is_available() else 'cpu'
model = TorchWrapper(torch_model, device)

# initialize the explainer with the model and method parameters
explainer = Saliency(model)

# call the explainer on the sample to explain
explanations = explainer(inputs, targets)

# compute explanation metrics
metric = Deletion(model, inputs, targets)
deletion_score = metric(explanation)
```

Operators

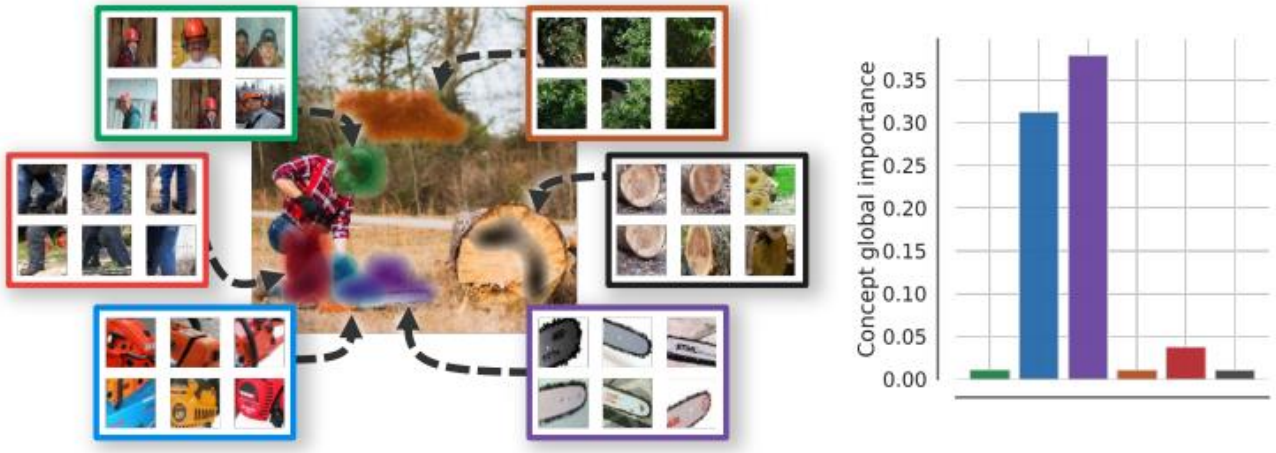
```python
from xplique.attributions import Saliency
from xplique.metrics import Deletion

# load images, targets, and model
# ...

# initialize the explainer with the model and method parameters
explainer = Saliency(model, operator=xplique.Tasks.SEMANTIC_SEGMENTATION)

# call the explainer on the sample to explain
explanations = explainer(inputs, targets)

# compute explanation metrics
metric = Deletion(model, inputs, targets, operator=xplique.Tasks.SEMANTIC_SEGMENTATION)
deletion_score = metric(explanation)
```

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

New methods

HSIC

Novello, Paul, Thomas Fel, and David Vigouroux. "Making sense of dependence: Efficient black-box explanations using dependence measure." *Advances in Neural Information Processing Systems* 35 (2022): 4344-4357.



CRAFT

Fel, Thomas, et al. "Craft: Concept recursive activation factorization for explainability." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

MaCO

Fel, Thomas, et al. "Unlocking feature visualization for deep network with MAgnitude constrained optimization." *Advances in Neural Information Processing Systems* 36 (2024).

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

DEEL

DEpendable & Explainable Learning

Demonstrations

IVADO

IRT SAINT EXUPÉRY

CRIAQ

Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI

Porté par l'Université de Toulouse

Institut intelligence et données

© DEEL- All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Demonstrations with Xplique tutorials

HSIC on a PyTorch model for semantic segmentation

- PyTorch wrapper
- Operators, they allow to treat new tasks:
  - Object detection
  - Semantic segmentation
- New state-of-the-art attribution methods:
  - Sobol and HSIC

CRAFT and MaCO

- Automatic concepts extraction
  - CRAFT
- Last feature visualization methods:
  - MaCO

© DEEL - All rights reserved to IVADO, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Future Works

Research

- Development of new methods (example-based)
- Extend to new Use Cases
- Development of new metrics
- ….

Tools

- Enhance our library efficiency
- Extend to new Use Cases
- CI/CD, include the newest methods
- ….

Benchmarking

- Provide a thorough User Guide
- Multi-criteria evaluation
- Challenge on more Use Cases
- ….

© DEEL - All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

DEEL

DEpendable & Explainable Learning

Feed-back from an user

IVADO

IRT SAINT EXUPÉRY

CRIAQ

Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI

Porté par l'Université de Toulouse

Institut intelligence et données

© DEEL- All rights reserved to IVADO, IID, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Want to know more ?

www.deel.ai

github/deel-ai

linkedin/showcase/deel-ai

Do not forget
to leave a star

Want to join DEEL ?

gregory dot flandin at irt-saintexupery dot com

Bâtiment B612
3 Rue Tarfaya, 31400 Toulouse