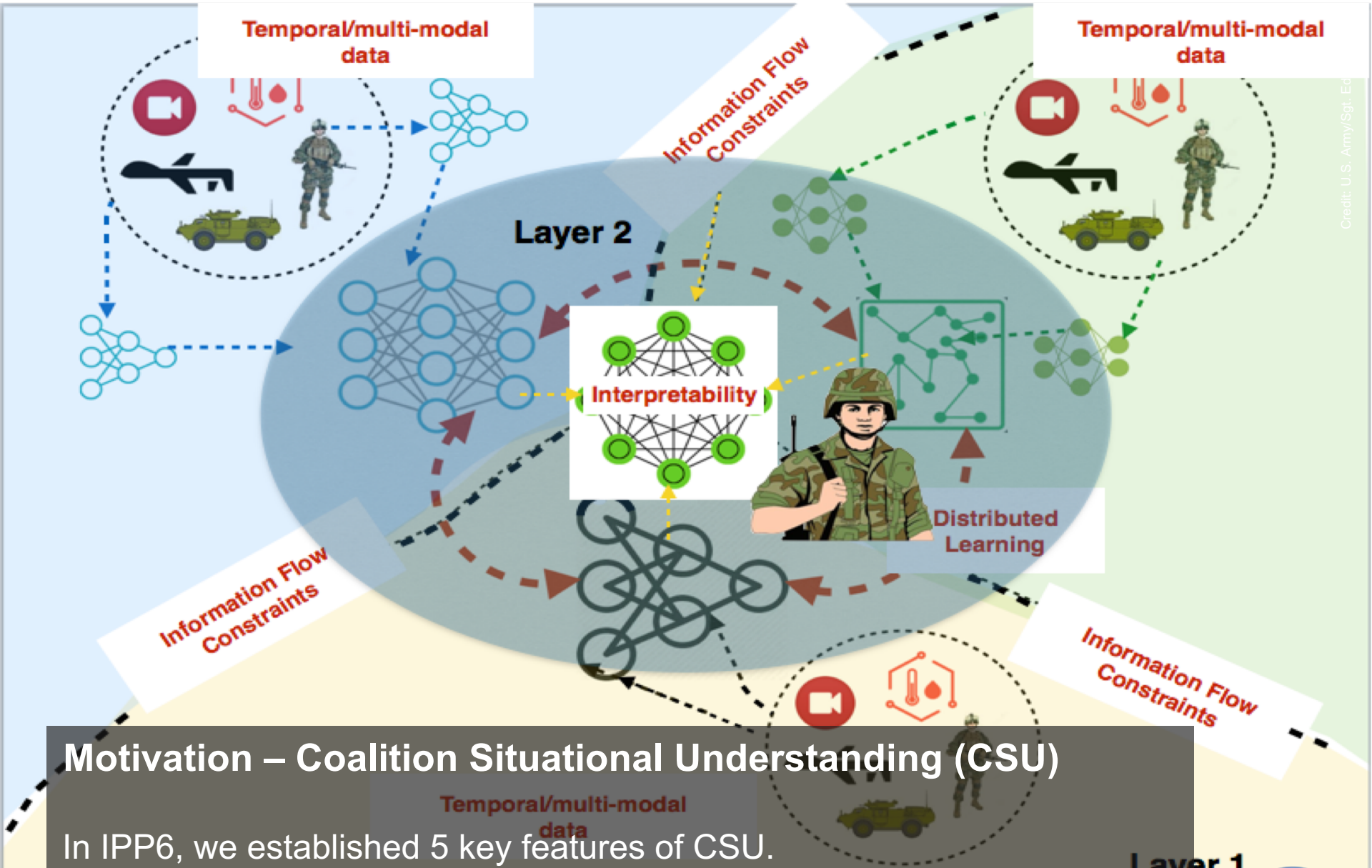
P5.2 Demo: Exploration of the Effect of Task and User Role on the Evaluation of Interpretability Techniques

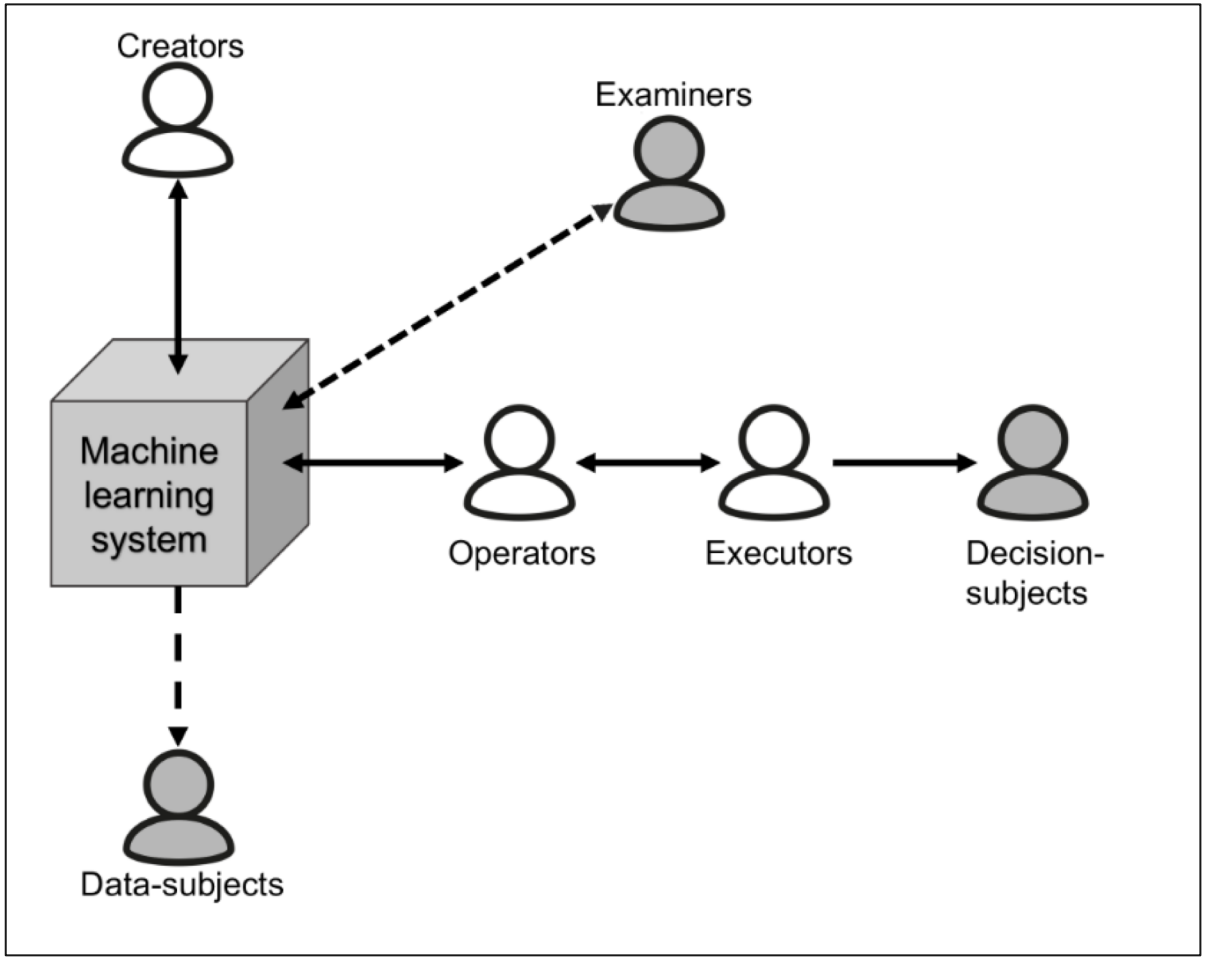
Annual Fall Meeting, New York, September 2018



Motivation – Coalition Situational Understanding (CSU)

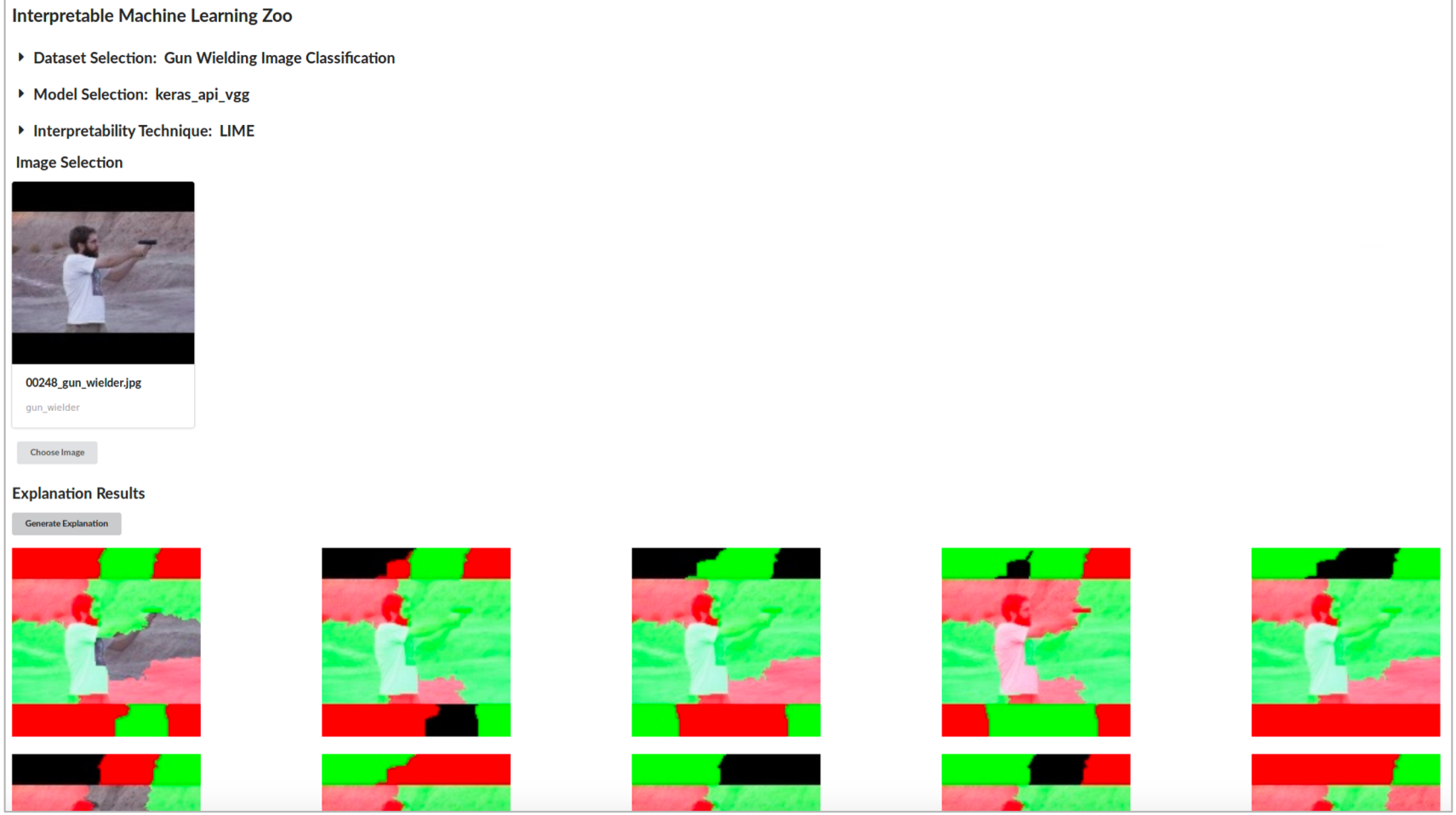
In IPP6, we established 5 key features of CSU.

In BPP5.2, ***interpretability of decision support models*** has been our focus.



An agent's role/relationship with a machine learning decision will effect the utility of interpretability techniques.

- **User Knowledge** may effect the details that can be provided within explanations.
- **Task** of the agent may effect what information is useable.



This demo showcases a framework developed that brings:

- **Easy Comparison** of existing (and new) interpretability techniques.
- **Generating Data** for experiments both analytical and human based testing.
- **Sharable Tool** that through open-sourcing can help engage the wider Machine Learning community.

Experimentation Framework - Datasets, Models and Explanations

Datasets

- Gun Wielder Classification
- Traffic Congestion
- CIFAR-10
- MNIST
- ImageNet

Models

Neural Networks:

- VGG16
- VGG19
- InceptionV3
- Inception ResNet V2
- MobileNet
- Xception

Other Models:

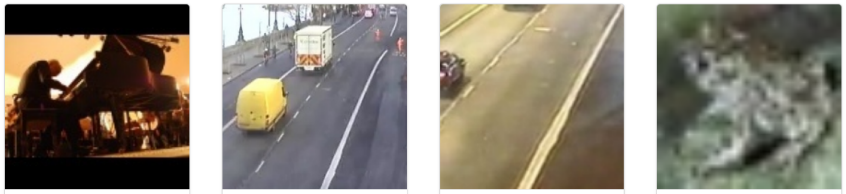
- Support Vector Machine

Explanation Techniques

- LIME
- Shapely
- Deep Taylor LRP
- Influence Functions

Experimentation Framework - Datasets, Models and Explanations

Dataset Selection: Gun Wielding Image Classification



Selected

Gun Wielding Image Classification

Image classification of people holding guns and

Traffic Congestion Image Classification

Image classification of traffic camera imagery collected from Transport for

Traffic Congestion Image Classification (Resized)

Resized version of the first traffic congestion image classification dataset Image

CIFAR-10

Dataset commonly used for benchmarking Machine Learning techniques.

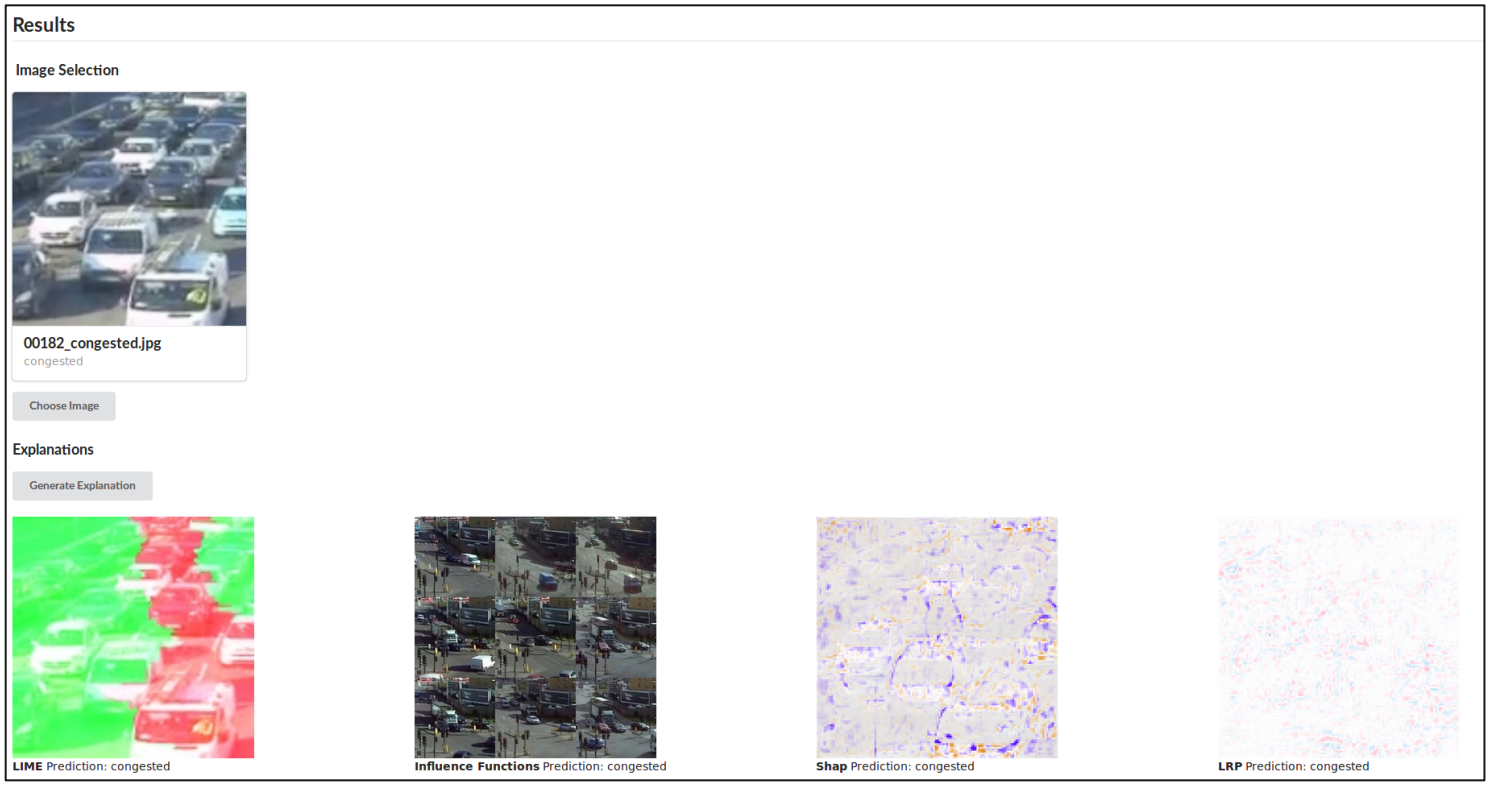
Model Selection: vgg16_imagenet

Model Name	Description	Performance Notes	
ConvSVM		Training Time: 228.53 Test Accuracy: 0.6015625	Use Model
VGG16Imagenet	A keras api VGG16 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model
VGG19Imagenet	A keras api VGG19 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model
InceptionV3Imagenet	A keras api InceptionV3 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model
InceptionResNetV2Imagenet	A keras api InceptionResNetV2 CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model
MobileNetImagenet	A keras api MobileNet CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model
XceptionImagenet	A keras api Xception CNN feature descriptor trained on Imagenet with newly trained fully connected layers.	Training Time: 0 Test Accuracy: 0	Use Model

Interpretability Technique: Influence Functions

Interpretability Technique	Description	
LIME	A local (example specific) decision-boundary explanation of evidence towards classes	Use Interpreter
Shap		Use Interpreter
Influence Functions	An explanation by example method that finds accurate approximations of the difference in loss at a test image due caused by retraining the model with the exclusion of a train image	Use Interpreter
LRP		Use Interpreter

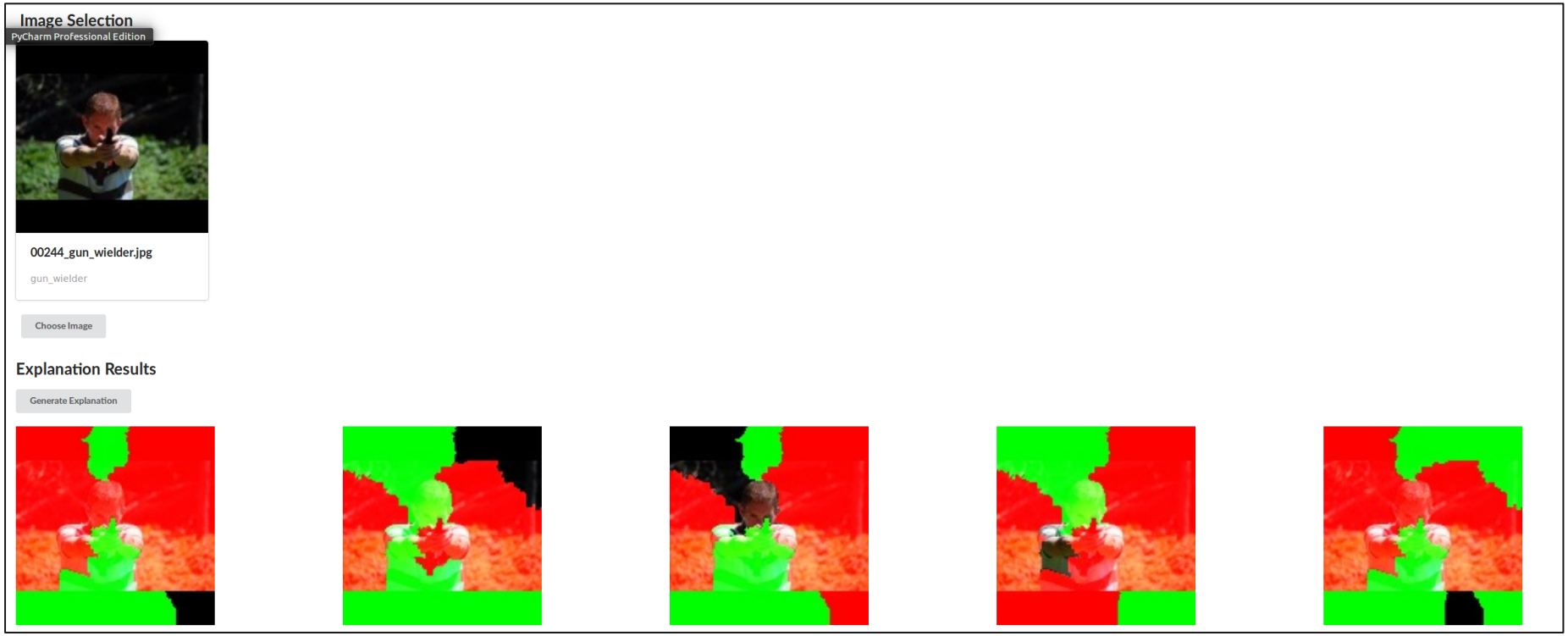
Experimentation Framework – Use Cases



Use Case 1 – Multiple Explanation Techniques on the Same Input:

- **Build Intuition** for different techniques.
- **Compare Utility** of different techniques – do different users (with different roles/knowledge) prefer different techniques.

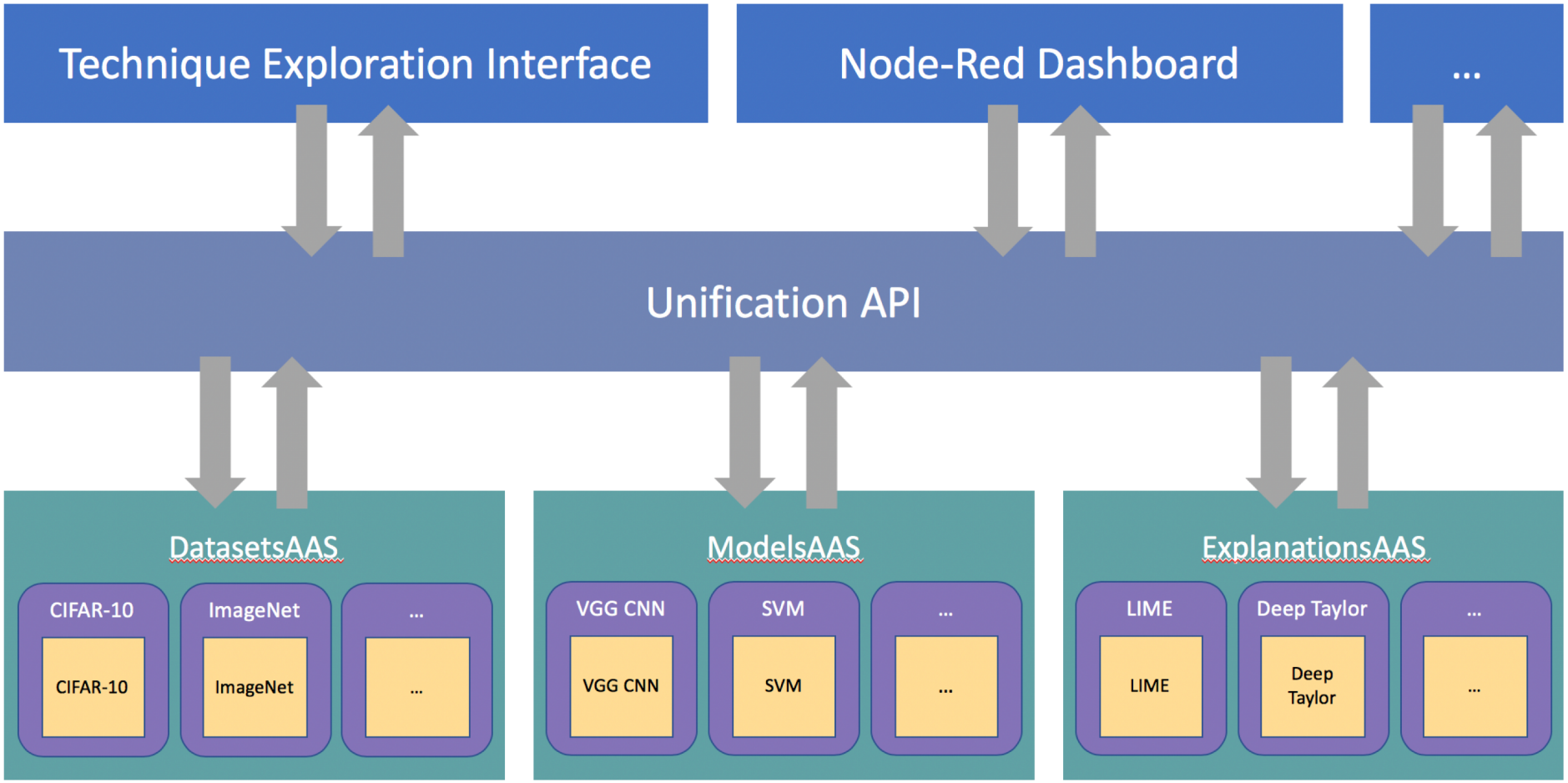
Experimentation Framework – Use Cases



Use Case 2 – Multiple Explanations from the Same Technique:

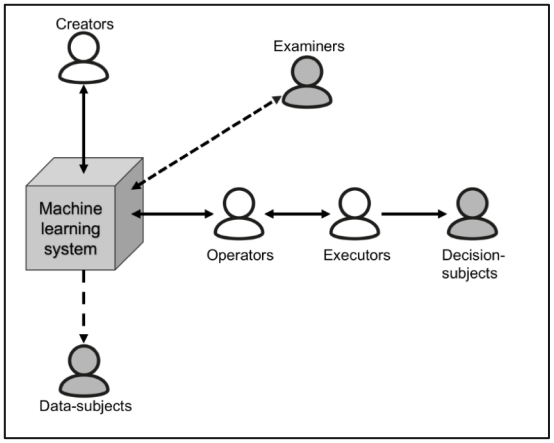
- **Build Intuition** for how stable the technique is.
- **Generating Data** for experiments.
- **Refine techniques** as seen in collaborative work with West Point cadets.

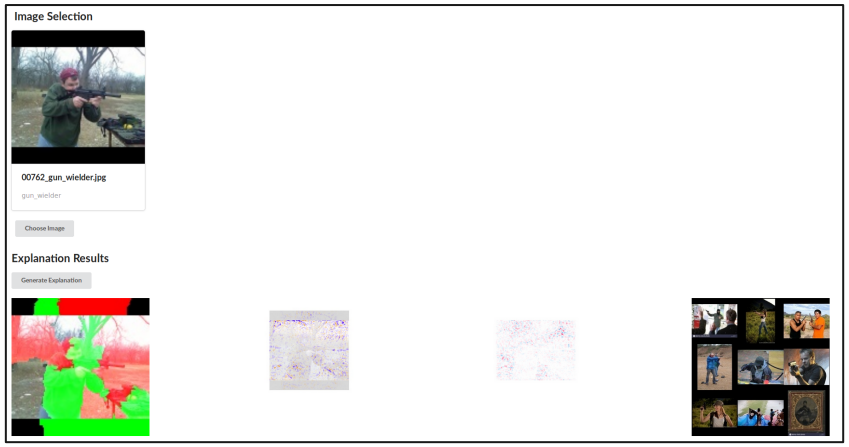
Experimentation Framework - Architecture



Framework - Summary

In response to the **need to compare and showcase** many **different existing and emerging interpretability techniques**, we have **produced a framework** that, from the ground up, is designed **to simplify** the **research process** of **our group** and the **wider community**.





Open-sourced at: github.com/dais-ita/interpretability_framework