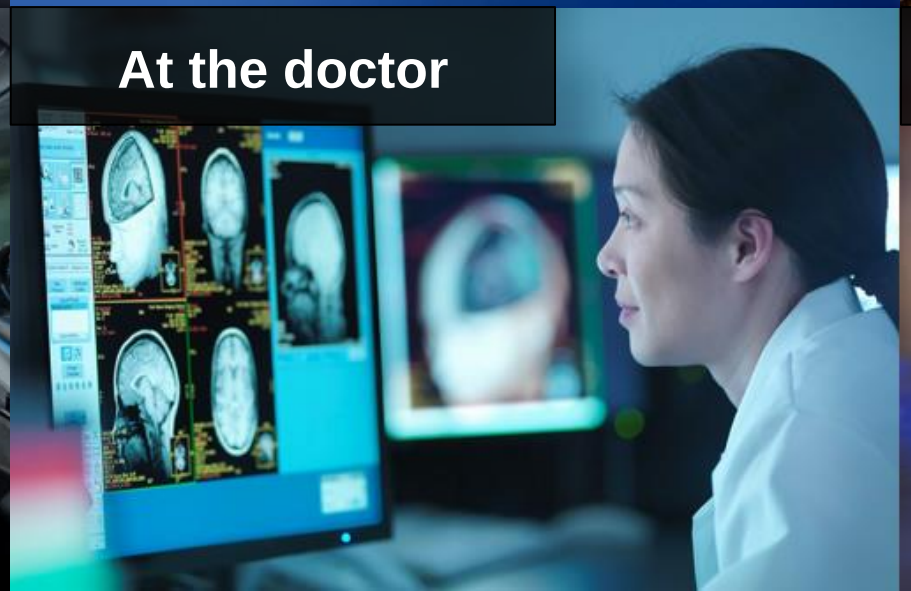
MATLAB EXPO
2021

인공지능 해부하기 :
설명 가능한 인공지능(eXplainable AI)

송완빈 과장

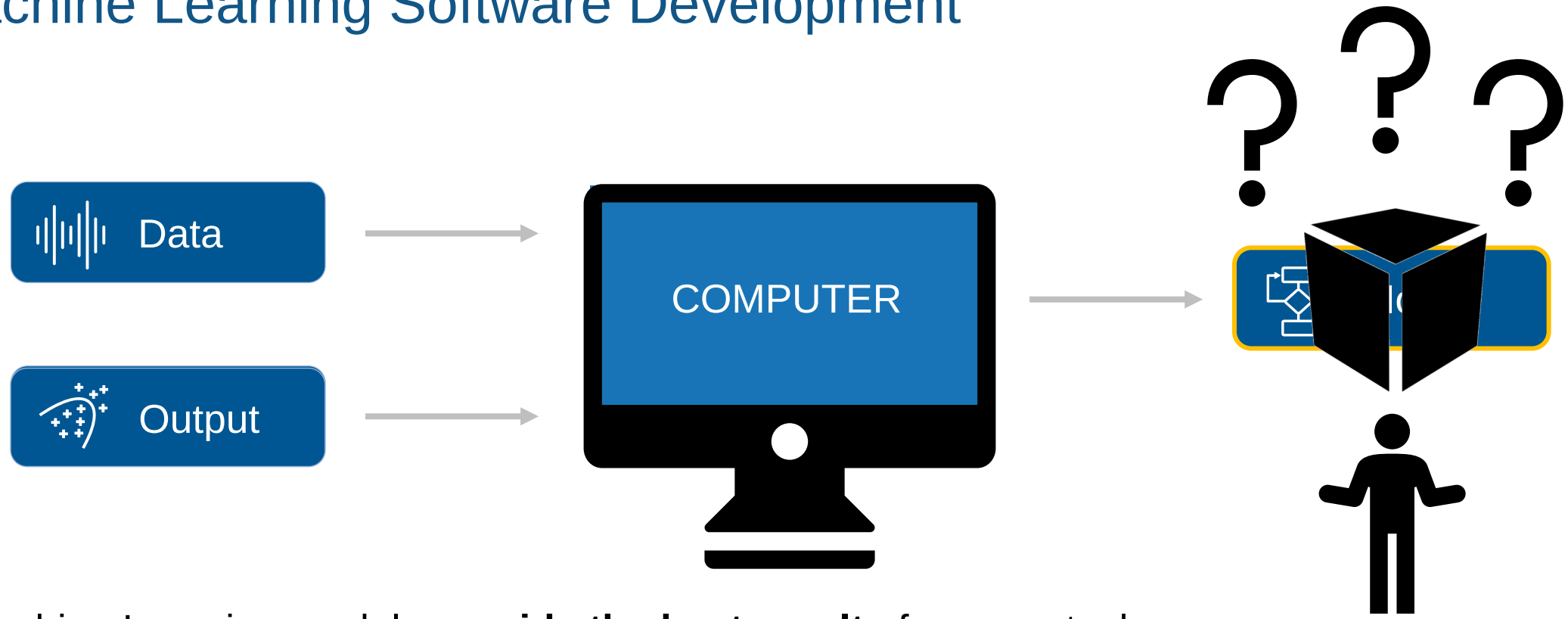
MathWorks

AI models you interact with regularly



Traditional Software Development

Machine Learning Software Development

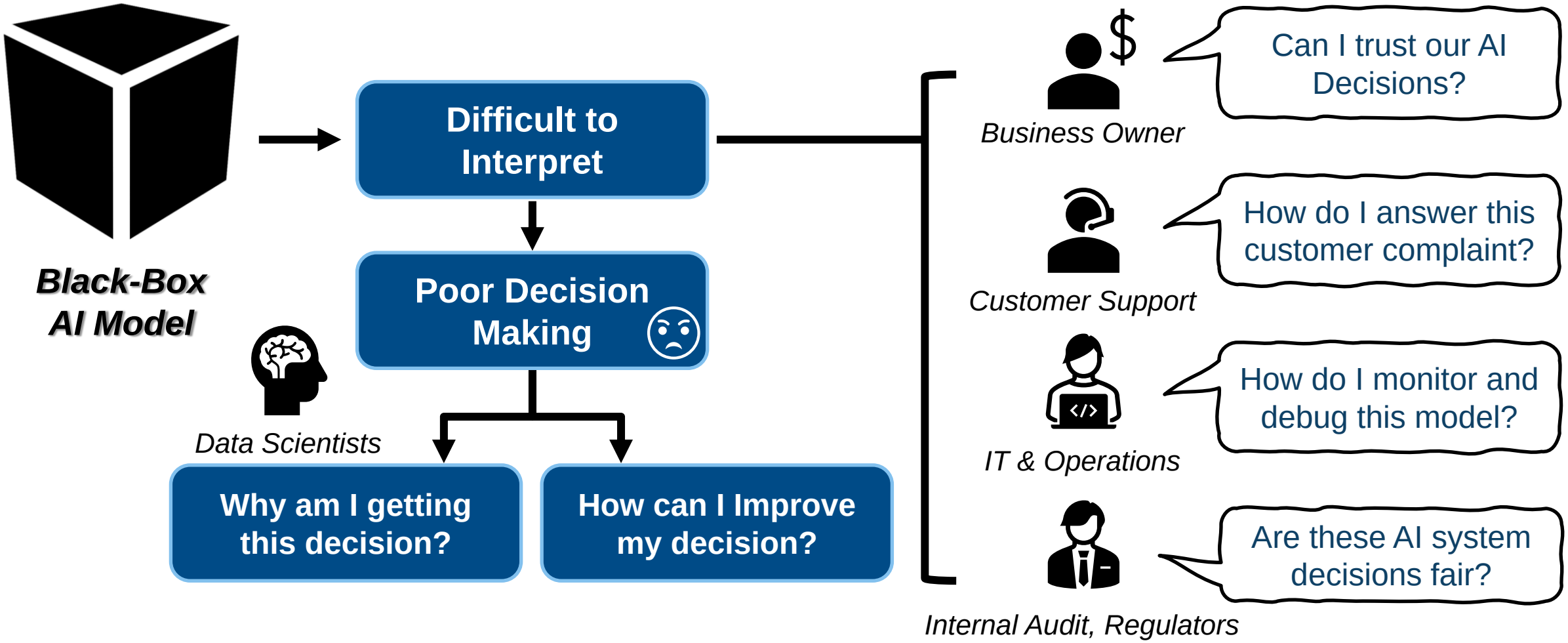


✓ Machine Learning models **provide the best results** for many tasks




✓ Desire to put ML models in production in safety critical situations.

Problems with Black-Box AI Models

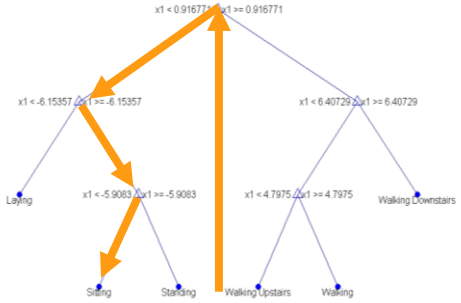


What is “eXplainable AI”?

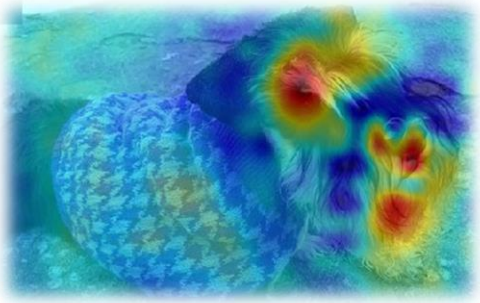
- Methods or techniques used to explain machine learning reasoning to humans

Black-Box AI Model

- **Interpretability**
  - Discern the mechanics without necessarily knowing why



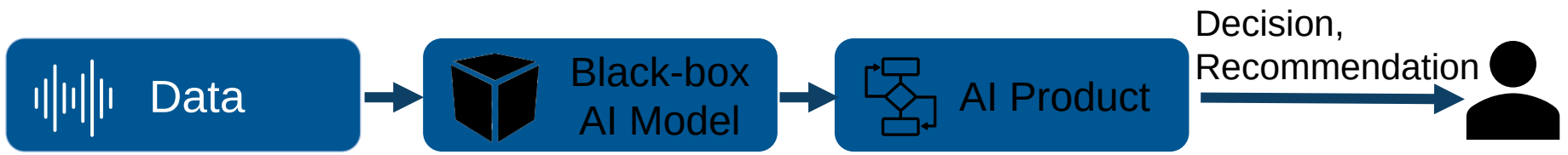
- **Explainability**
  - Being able to quite literally explain what is happening



The goal of eXplainable AI

Today

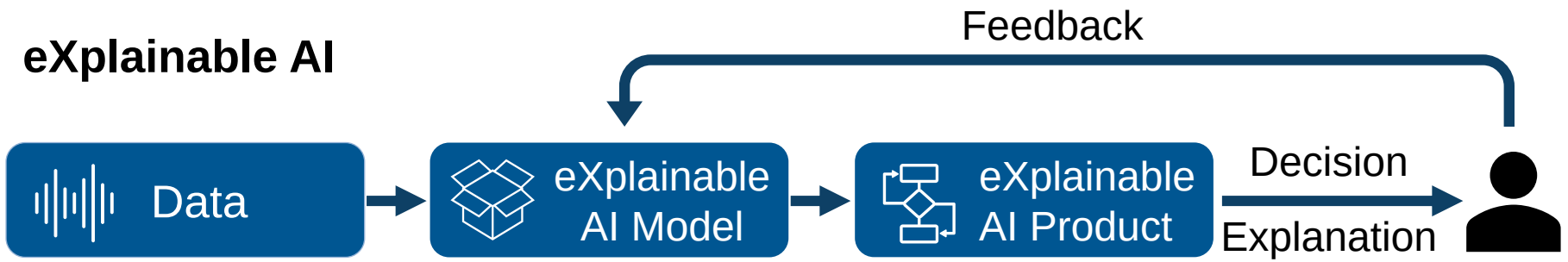
Black Box AI



- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Tomorrow

eXplainable AI



- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Categorize with respect to the viewpoint of eXplainable AI

- **Complexity**
  - Intrinsic
  - Post-hoc
- **Scope**
  - Global
  - Local
- **Dependency**
  - Model-specific
  - Model-agnostic

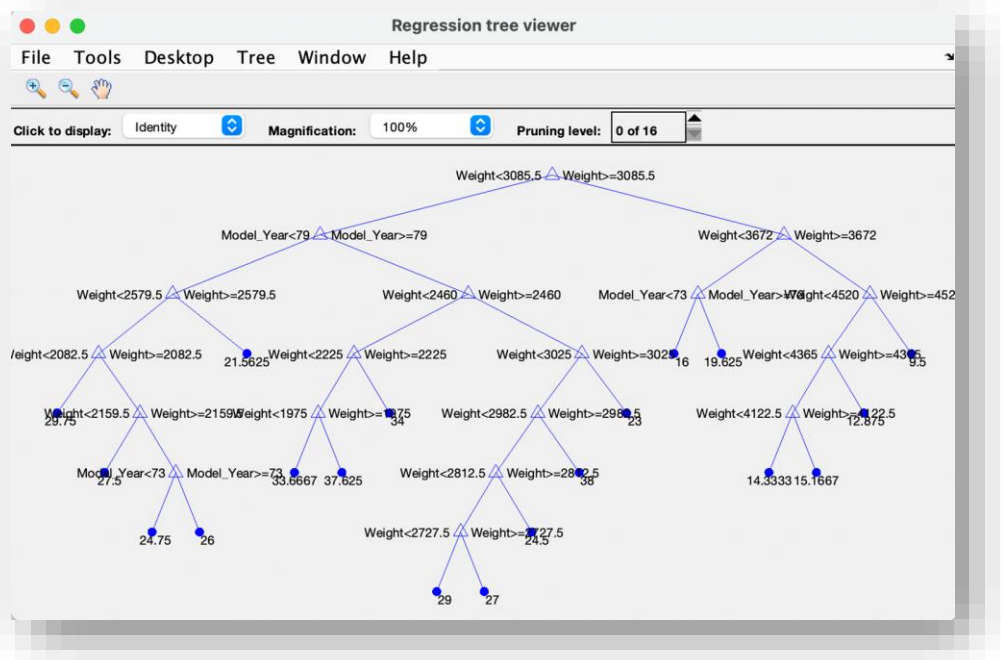
Viewpoint of eXplainable AI

Complexity

Intrinsic

Post-hoc

- Intrinsic / Transparency
  - Model itself already have transparency and interpretable




```
Linear regression model:
    Model_Year ~ 1 + Weight + MPG

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	49.758	4.1453	12.003	1.8646e-20
Weight	0.0032738	0.00080383	4.0728	9.9032e-05
MPG	0.70279	0.080191	8.7639	9.8456e-14

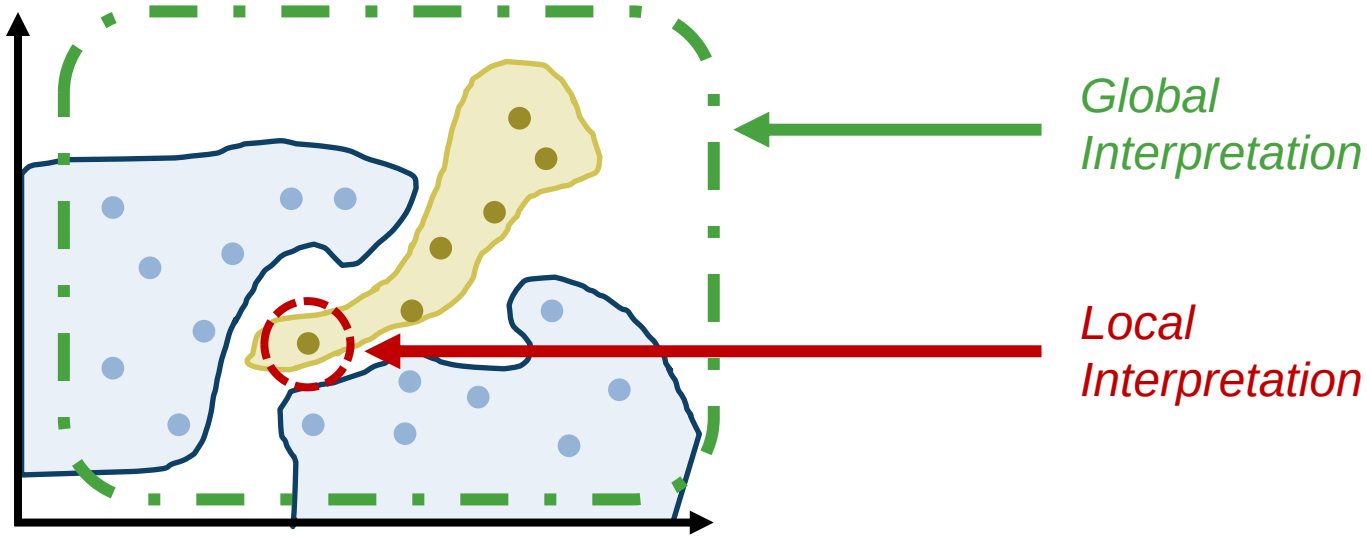
- Post-hoc
  - Normally use complex model which has large predictive power and interpret later

Viewpoint of eXplainable AI

Scope

Global

Local



- Global
  - Provides an overview of the most influential variables in the model, based on the data input and the predicted variable.
- Local
  - Explains conditional interaction between input/output with respect to single prediction result

Viewpoint of eXplainable AI

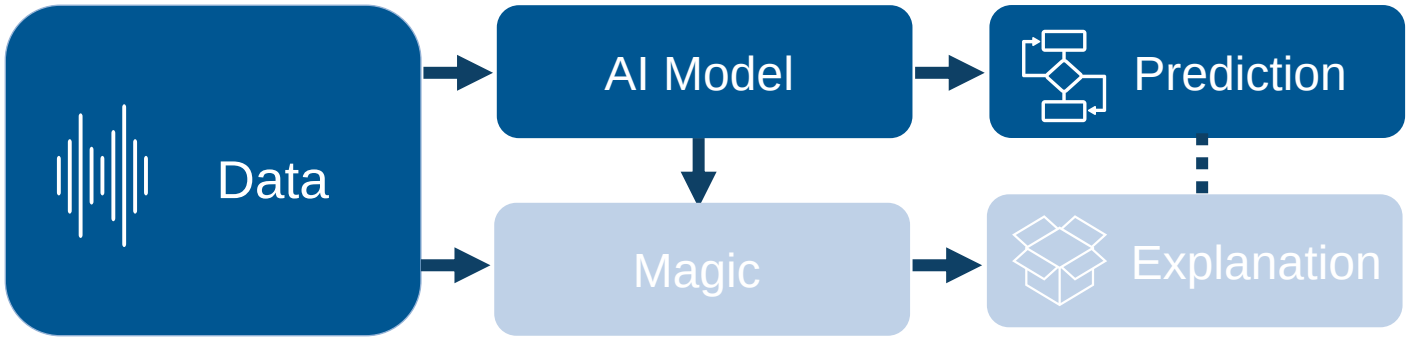
Dependency

Model-specific

Model-agnostic

- Model-specific
  - Deals with inner working of model to interpret its result

- Model-agnostic
  - Deals with analyzing the feature and its relationships with its output



Different Interpretability methods

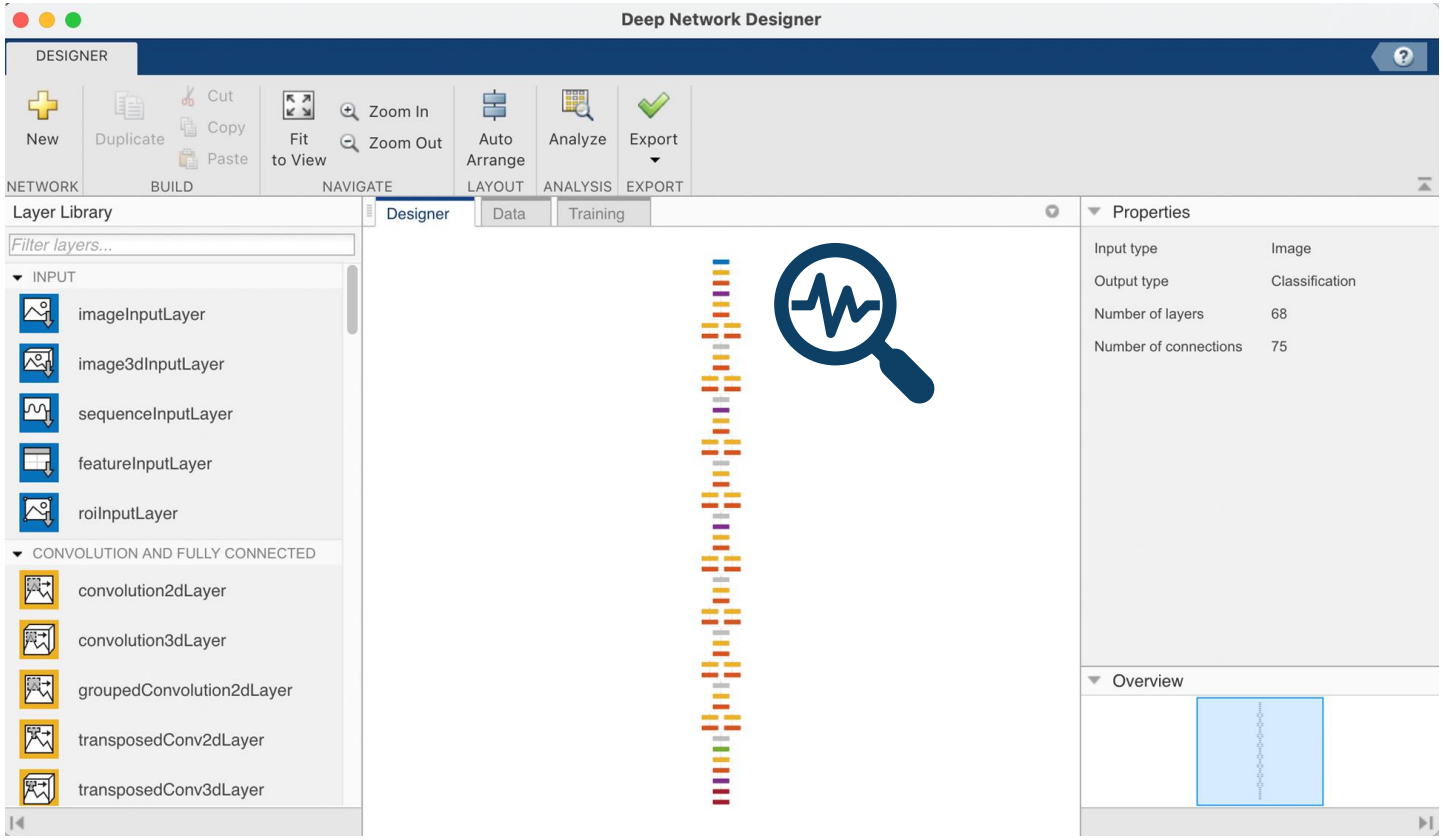
	Intrinsic	Post hoc
Model-Specific	Regression	Predictor Importance (trees & ensemble)
	Tree Models	Occlusion Sensitivity CAM, Grad-CAM Activation
	Correlation	
Model-Agnostic	ACP	Partial Dependence Plot Individual Conditional Exp Accumulated Local Effects
	tSNE	
	MRMR	LIME Shapley values *Local methods*

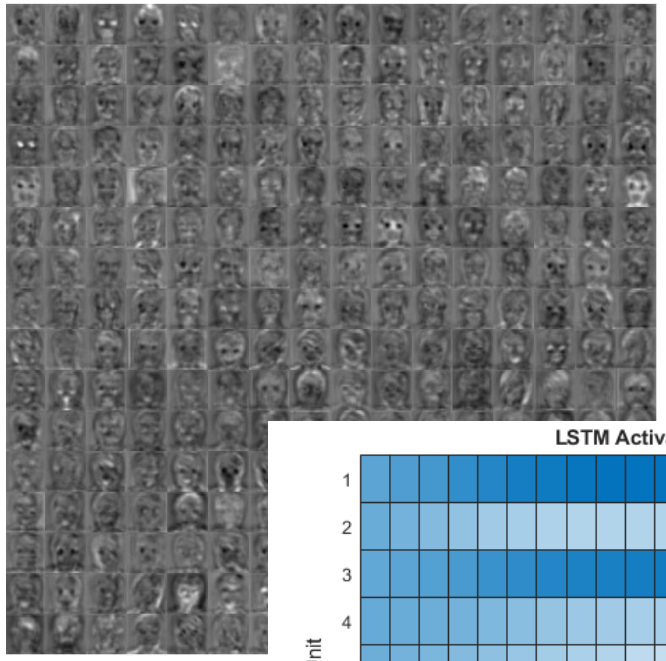
Visual Interpretation of Features Across Layers

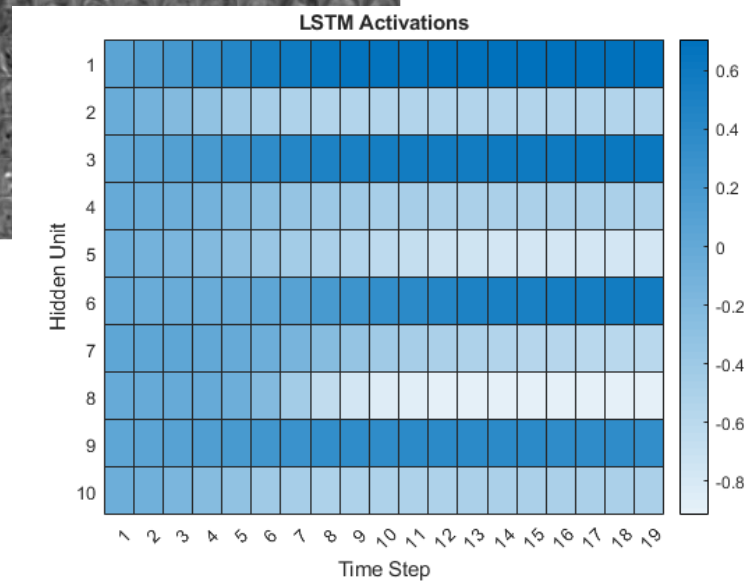
Activations

Post hoc

Model-Specific





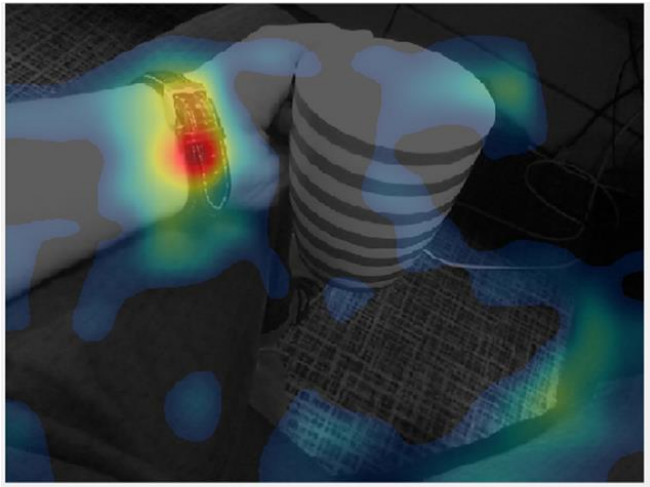


Deep Learning interpretability methods

Grad-CAM & Occlusion sensitivity

Post hoc

Model-Specific



Truth:	Coffee mug
AI:	Buckle (15%) ✗

AI classifies incorrectly as "buckle" due to the watch

```
scoreMap = gradCAM(net,X,label)
```

```
scoreMap = occlusionSensitivity(net,X,label)
```

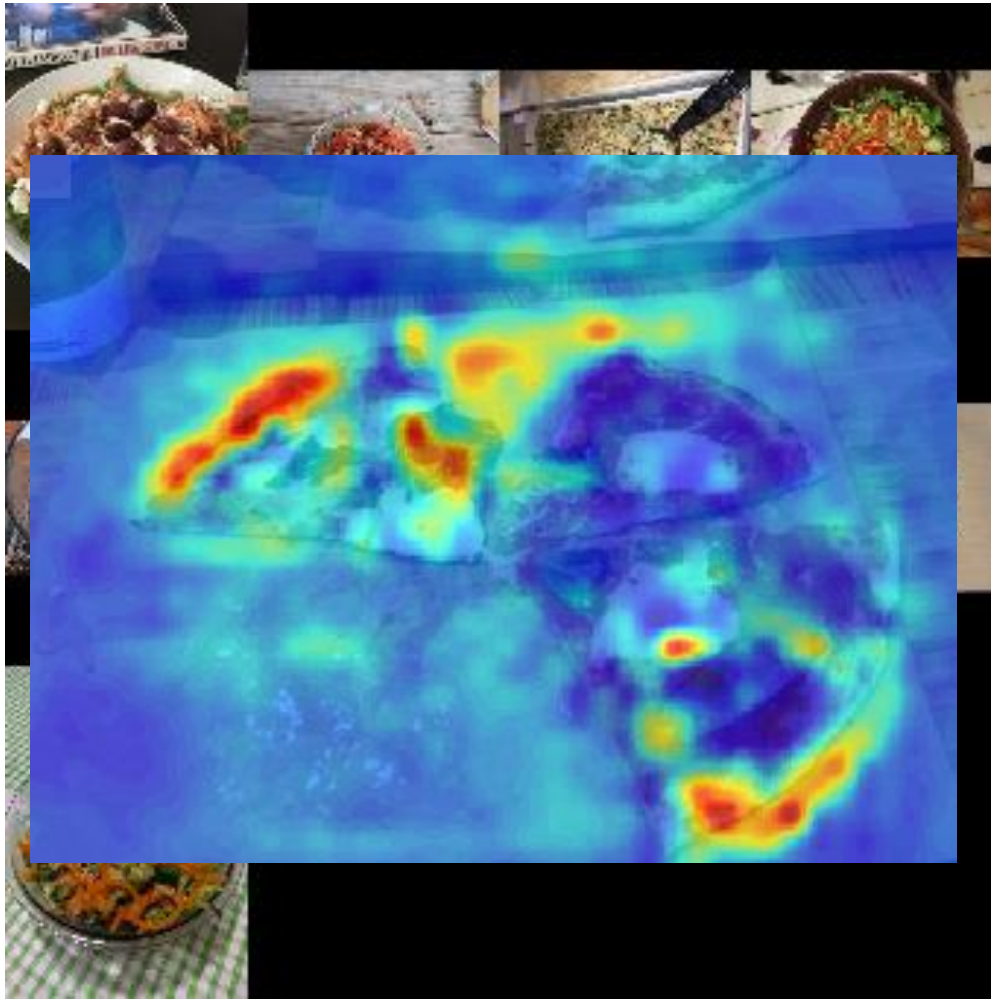
Deep Learning interpretability methods

Grad-CAM & Occlusion sensitivity

Post hoc

Model-Specific

- ✓ **Initial investigation:** why are my salad pictures misclassified as pizza?
- ✓ **Hypothesis:** the network is focused on the curving edges of pizzas
- ✓ **Test:** does this work for other data pizza images?
- ✓ **Fix:** add more pizza slice images and salads on plates with curved edges



Musashi Seimitsu Industry Uses Deep Learning for Visual Inspection of Automotive Parts

Challenge

Reduce the workload and cost for manually operated visual inspection of 1.3 million automotive parts per month, by implementing an anomaly detection system using deep learning.

Solution

Musashi Seimitsu built a camera connection setup, preprocess images, create a custom annotation tool, and improve the model accuracy. They generated code for the trained model using GPU Coder™, implemented it on NVIDIA® Jetson.

Benefits of using MATLAB and Simulink

- Enable a seamless development workflow from image capture to implementation on embedded GPU
- Estimate and visualize the defect area using **Class Activation Mapping**
- Create custom user interfaces with App Designer for improving labeling efficiency
- Leverage consulting services to maximize the benefits of using MATLAB

> *"Using camera connection, preprocessing, and various pretrained models in MATLAB enabled us to work on the entire workflow. Through discussions with consultants, our team gained many tips for solving problems, growing the skills of our engineers."*

Different Interpretability methods

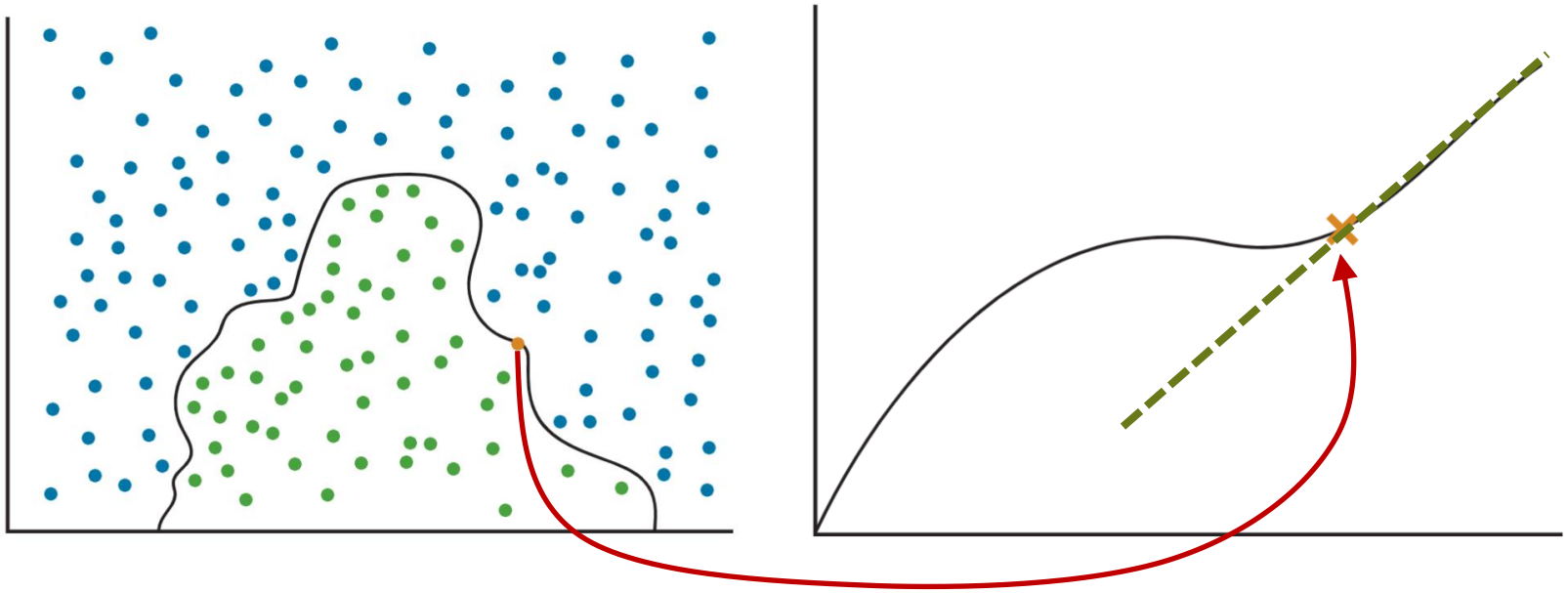
	Intrinsic	Post hoc
Model-Specific	Regression	Predictor Importance (trees & ensemble)
	Tree Models	Occlusion Sensitivity CAM, Grad-CAM Activation
	Correlation	
Model-Agnostic	ACP	Partial Dependence Plot Individual Conditional Exp Accumulated Local Effects
	tSNE	
	MRMR	LIME Shapley values *Local methods*

LIME(Local Interpretable Model-agnostic Explanation)

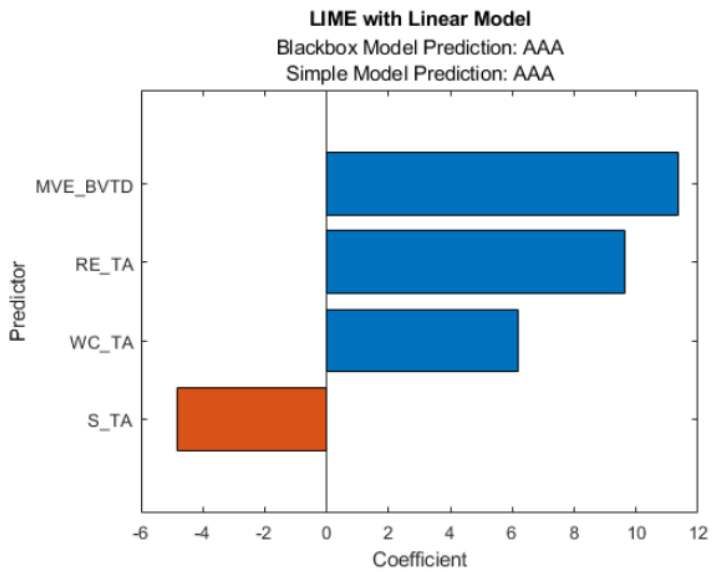
Post hoc

Model-Agnostic

- Fit a simple interpretable model for 1 query point



```
results = lime(blackbox)
```



✓ Complex model and Point of Interest

✓ Approximate with Simple Model

✓ ‘Explain’ = drivers within area of interest

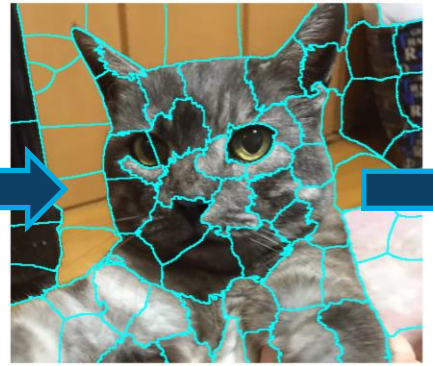
LIME(Local Interpretable Model-agnostic Explanation)

Post hoc

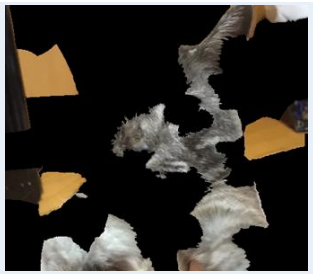
Model-Agnostic

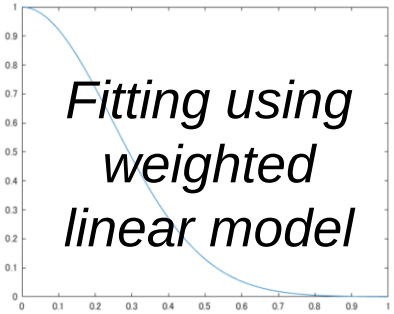
- Fit a simple interpretable model for perturbed instances

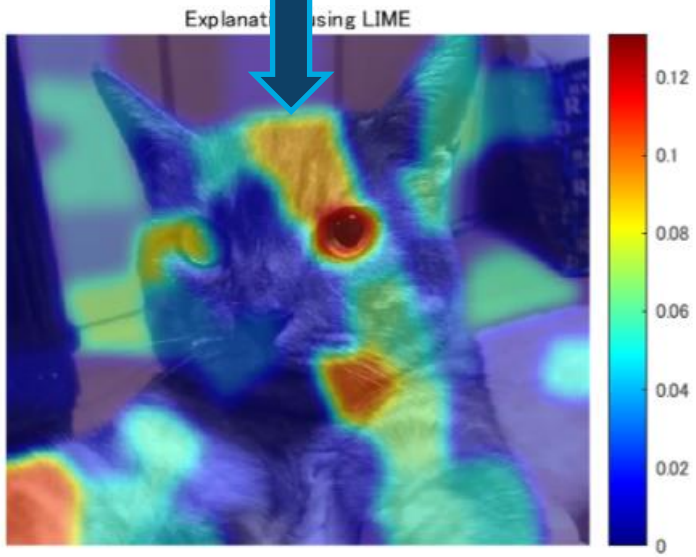
Original Image



Interpretable Components

Perturbed Instances	P(Egyptian Cat)
	0.8
	0.4
	0.2






```
scoreMap = imageLIME(net,X,label)
```

By whitening your AI model, you can expect

Explainable AI

MATLAB

PDP, ICE
LIME, Shap
Occlusion Sensitivity, CAM, Grad-CAM, Predictor importance, etc.

Use MATLAB functions to **explain** your model

Data Scientist: I can understand my models & debug it easily

Manager: I trust & understand the data scientist's models

Regulator: I can validate model fairness and trustworthy

Where can I use Knowledge learned from XAI?

- Decision Critical Application

Fraud detection *Self driving* *Heath care*

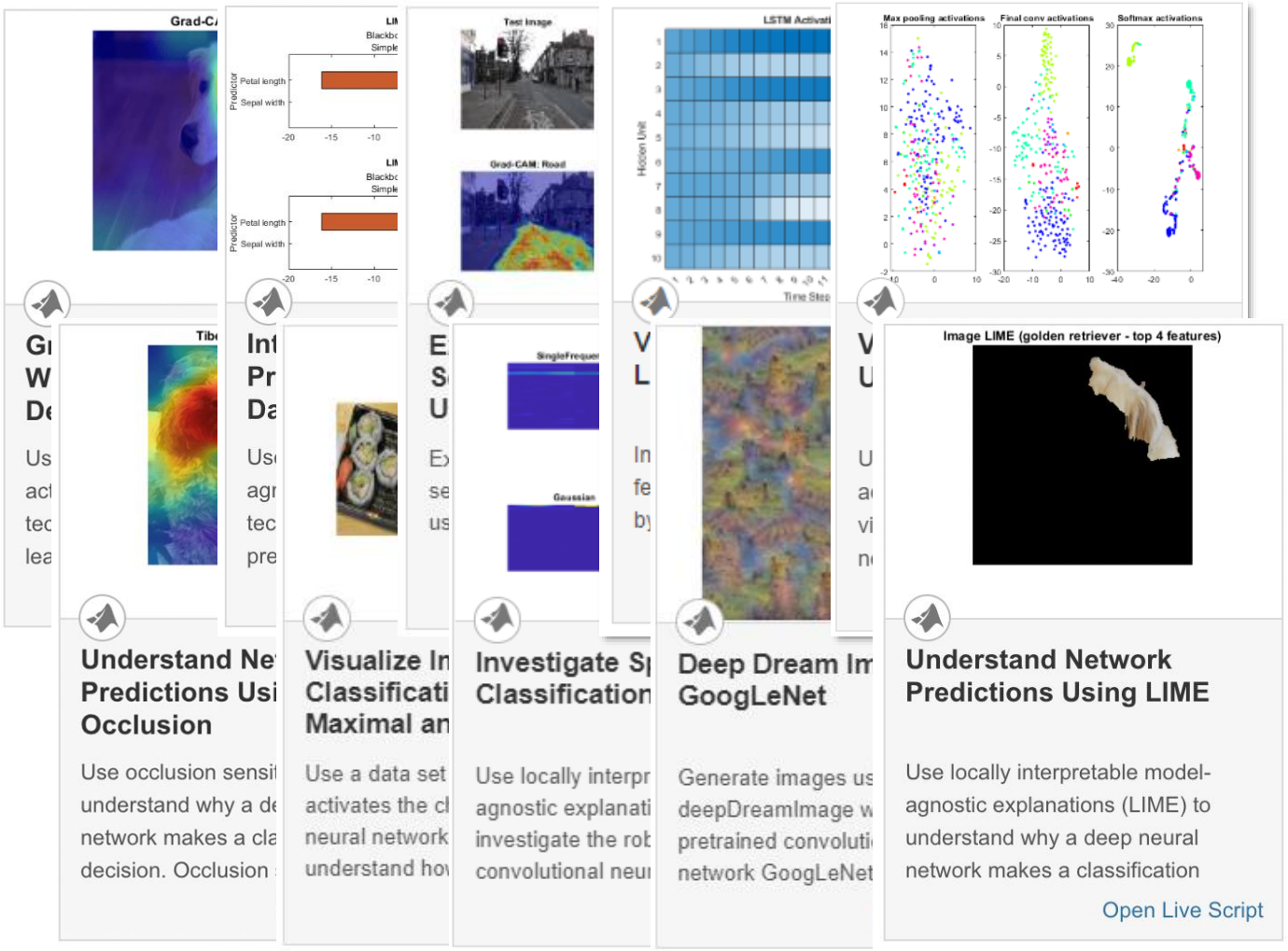
- AI Software = **Code** + **Data**
  (Model/Algorithm)

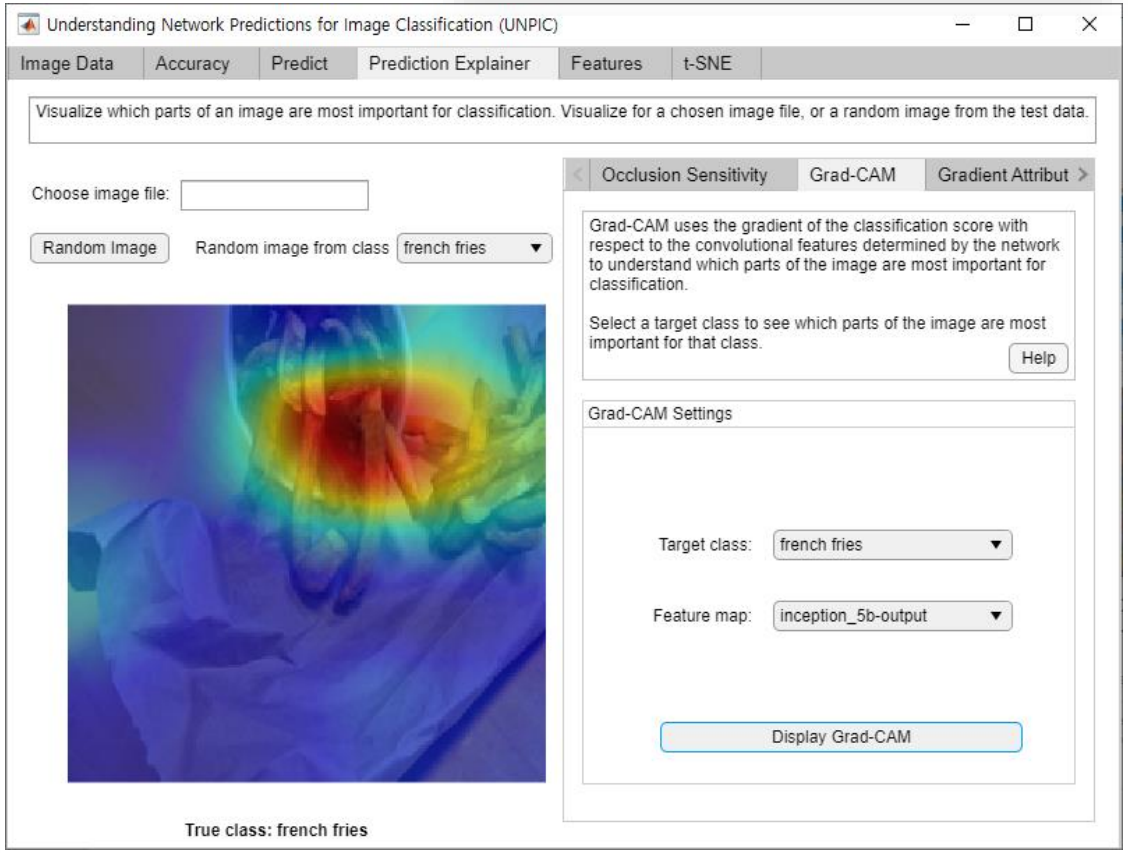
Scope Project → **Data Preparation** → **AI Modeling** → **Simulation & Test** → **Deployment**

Iterative Improvement

Big Data to Good Data

Start whitening your AI model with Golden References Today





MATLAB EXPO 2021

감사합니다

MathWorks®