Explainable AI (Overview)

Trevor Bonjour

What will we cover today?

- What do we mean by interpretability?
- Motivation
- Interpretable Models
- Model Agnostic Methods
- Explainable Reinforcement Learning

What is interpretability

Interpretability is the degree to which a human can understand the cause of a decision.

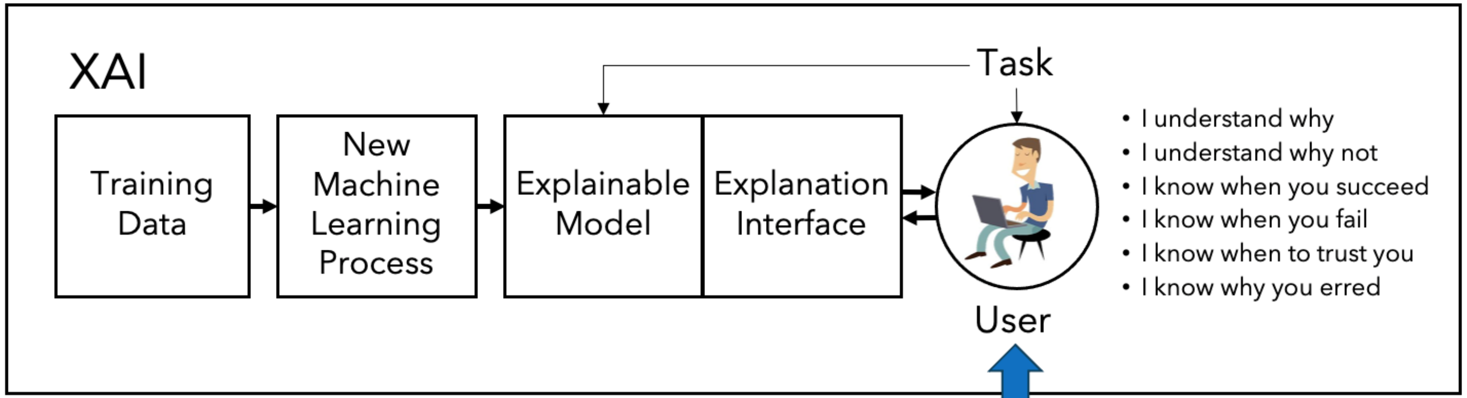
Why interpretability?

Decisions are critical in high-risk environments. Often machine learning algorithms are opaque.

Why interpretability?

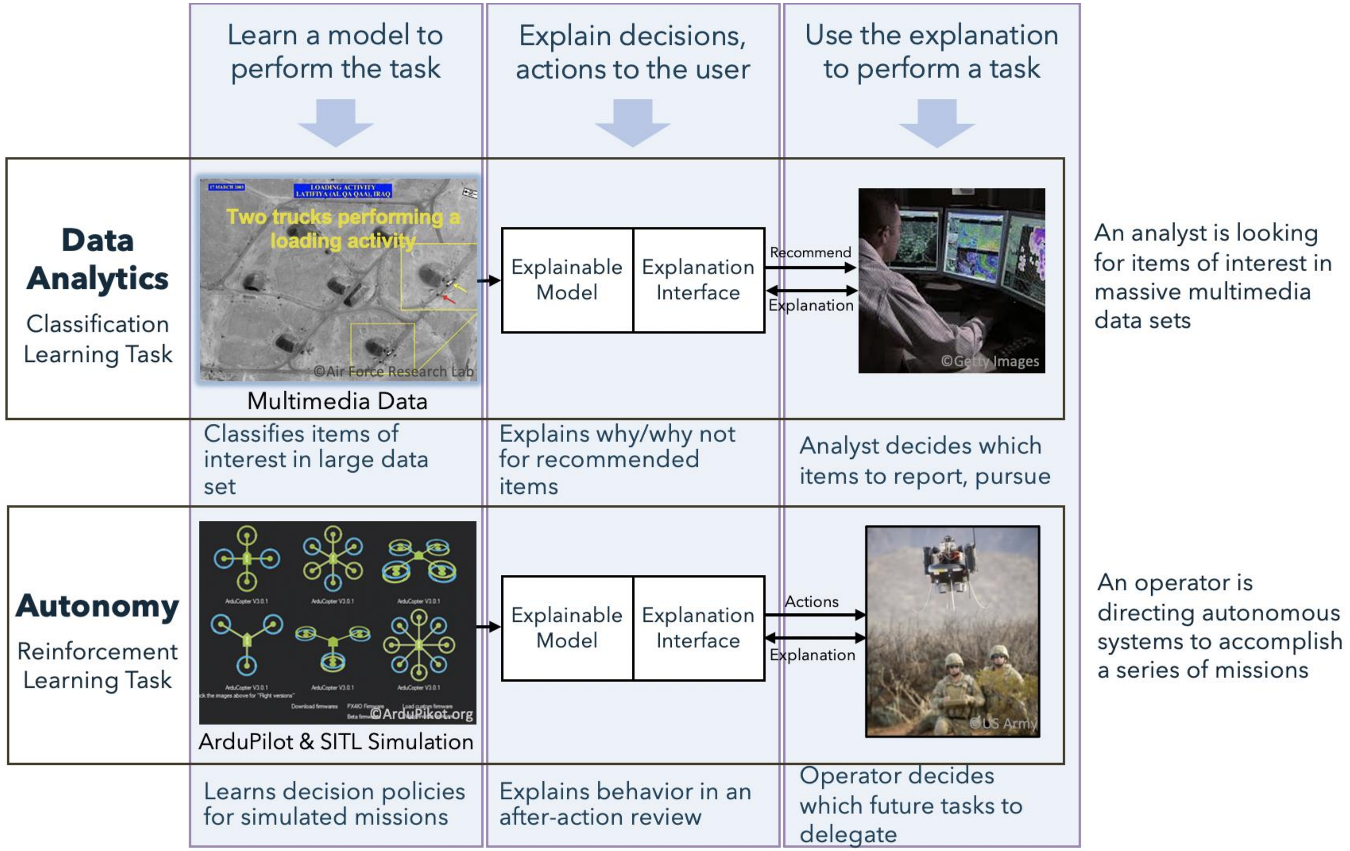
- To verify the model works as expected. Wrong decisions can be costly and dangerous.
- Learn new insights.
- Ensure fairness
- Enable trust in the system
- Ensure reliability: Small changes in the input should not lead to large changes in the output/prediction
- To check only causal relationships are picked up.
- To be able to debug mis-classifications by the model.

Need for interpretability



- The target of XAI is an end user who:
  - depends on decisions, recommendations, or actions of the system
  - needs to understand the rationale for the system's decisions to understand, appropriately trust, and effectively manage the system
- The XAI concept is to:
  - provide an explanation of individual decisions
  - enable understanding of overall strengths & weaknesses
  - convey an understanding of how the system will behave in the future
  - convey how to correct the system's mistakes (perhaps)

Problem Areas



Interpretable Models in Classification Tasks

- Linear/Logistic Regression
- Decision Trees

Linear/Logistic Regression

- Pros:
  - Predicts the target as a **weighted sum** of the feature inputs making the mechanism somewhat transparent.
  - Widely used – high level of collective experience and expertise
  - Guaranteed to find optimal weights(provided assumptions are met)
- Cons:
  - Can only represent linear relationships (non-linearity must be hand-crafted)
  - Often not that good regarding predictive performance
  - Interpretation of weights unintuitive

Decision Trees

- Pros:
  - Ideal for capturing interactions
  - Has a natural visualization
  - Creates good explanations
- Cons:
  - Does not deal with linear relationships
  - Slight changes in the input feature can have a big impact on the predicted outcome
  - Unstable - few changes in the training dataset can create a completely different tree
  - Decision trees are very interpretable -- as long as they are short

Model Agnostic Methods

- Permutation Feature Importance
- Global Surrogate
- Local Surrogate(LIME)
- Shapley Values

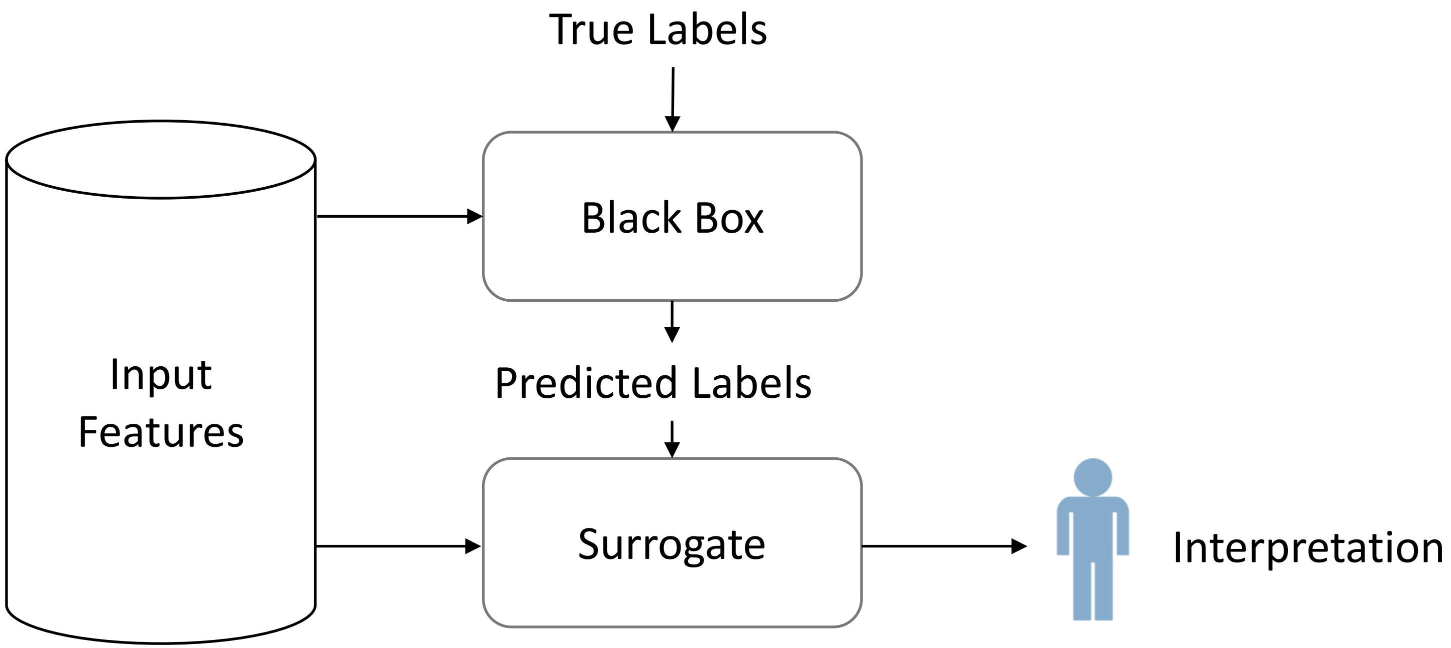
Permutation Feature Importance

Measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature.

- Introduced for Random Forests by Breiman (2001)
- Model agnostic method proposed by Fisher, Rudin, and Dominici (2018)
- **Important Feature:** If shuffling the values increases the error
- **Unimportant Feature:** If shuffling the values leaves the model error unchanged

Global Surrogate

An interpretable model that is trained to approximate the predictions of a black box model.

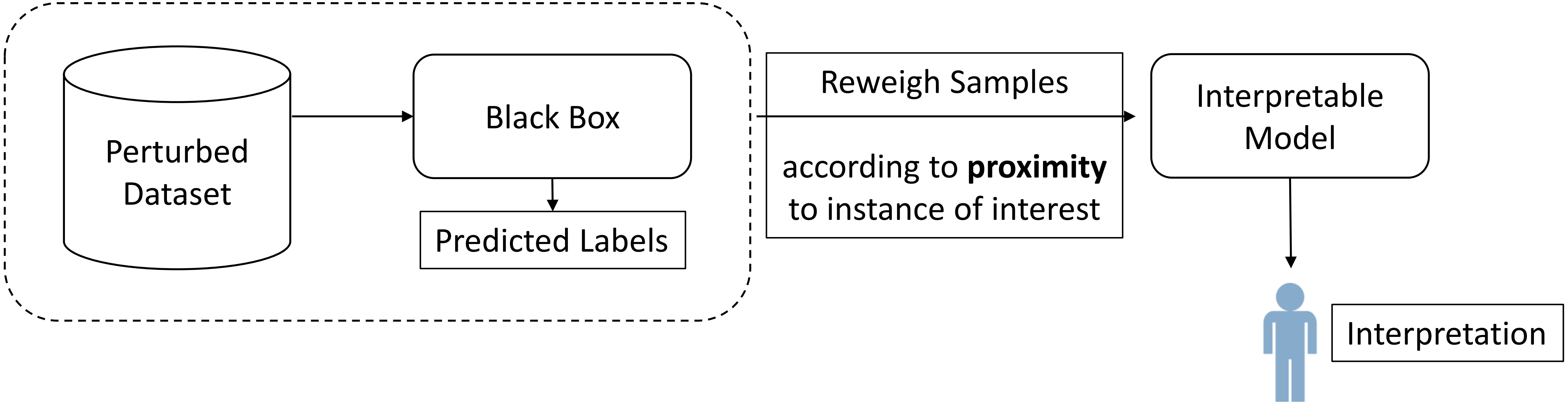


Local Surrogate

Train local surrogate models to explain individual predictions.

Concrete Implementation: Local interpretable model-agnostic explanations (LIME) by Ribeiro et al.(2016)

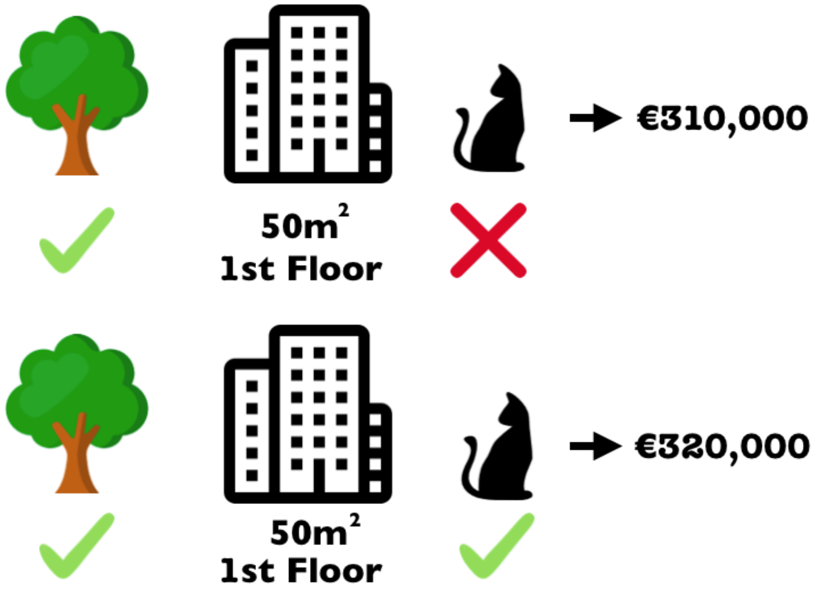
- LIME uses exponential smoothing kernel for calculating proximity
- Good approximation of predictions locally, not globally



Shapley Values

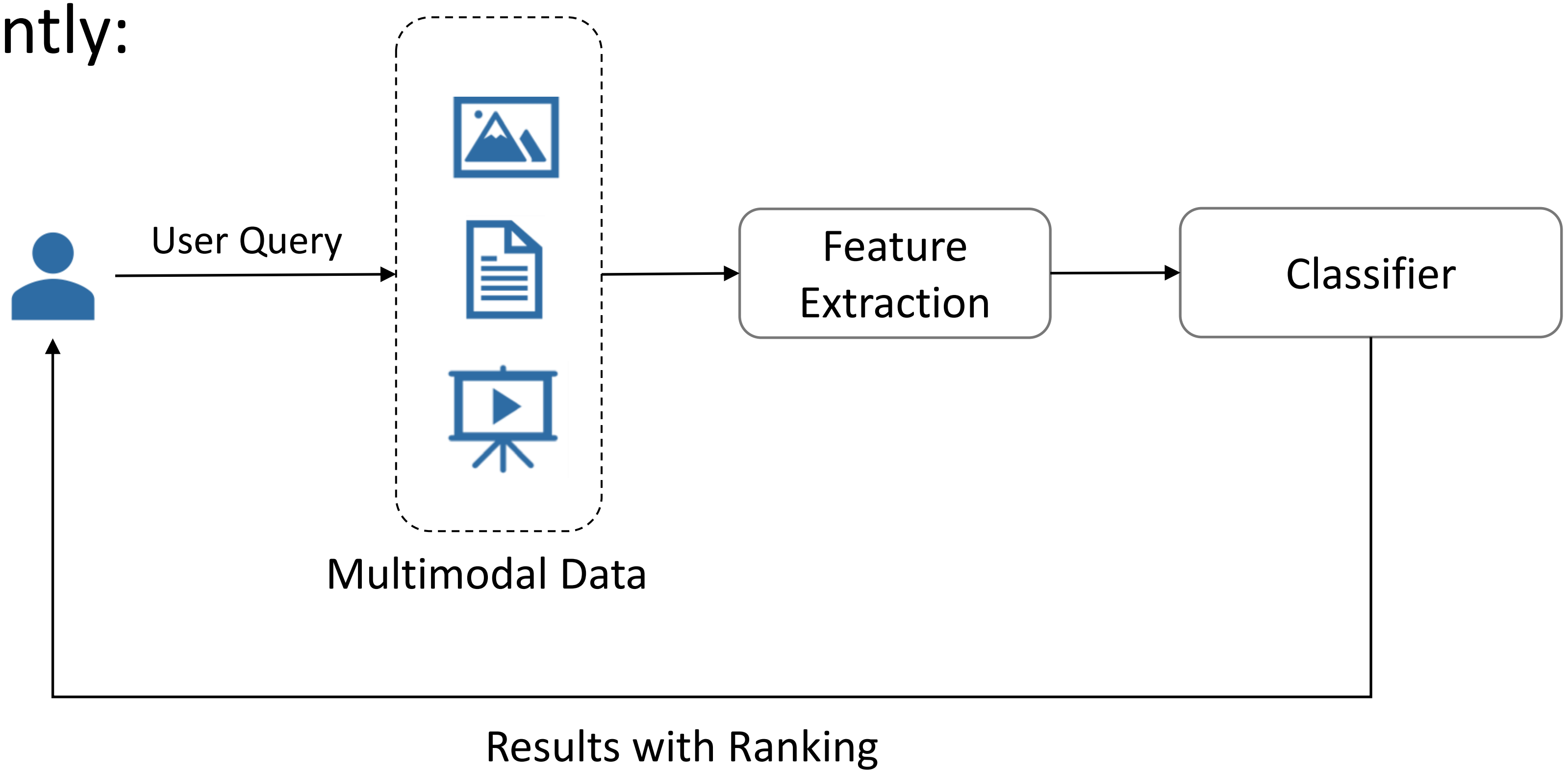
Explain the prediction of an instance by computing the contribution of each feature to the prediction.

- The Shapley value is the average marginal contribution of a feature value across all possible combinations.
- Assign values to features depending on their contribution to the prediction – e.g. buying an apartment with *pets allowed* adds 10k to the cost
- SHAP (SHapley Additive exPlanations) by Lundberg and Lee (2017) connects Shapley Values to LIME



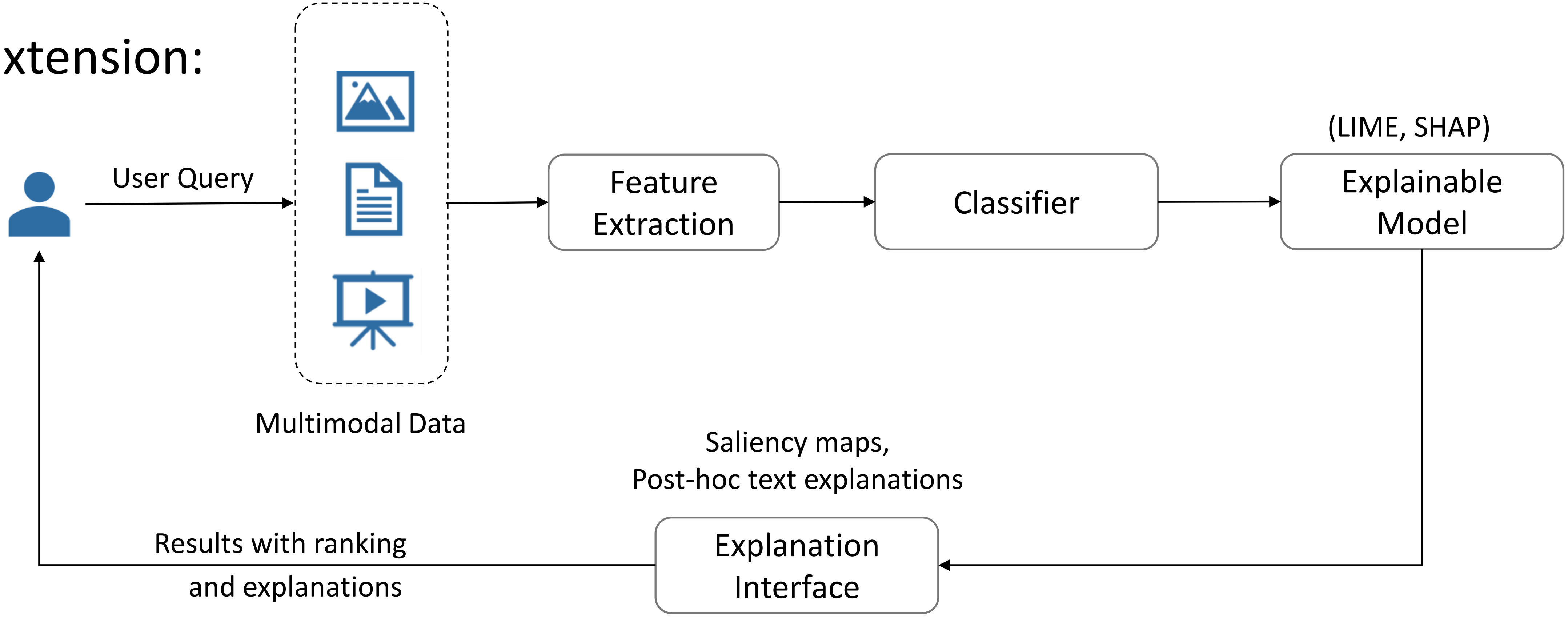
Application of XAI in Multimodal Predictions

- Currently:



Application of XAI in Multimodal Predictions

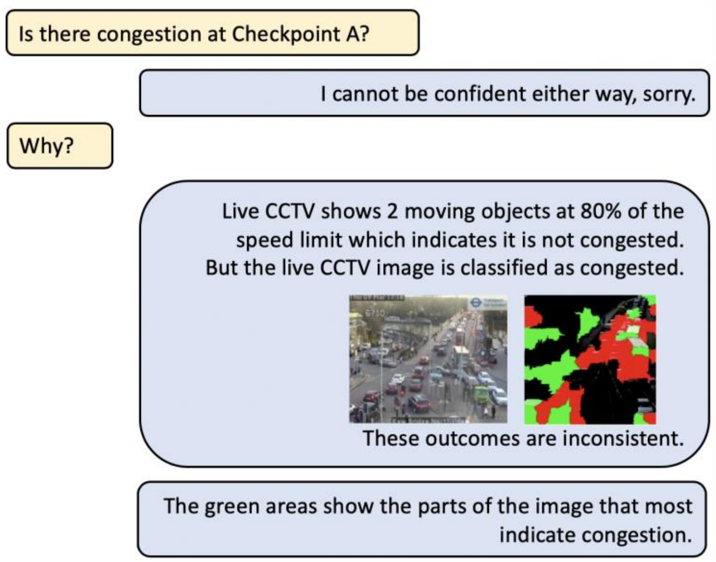
- Extension:



Multiple modalities provide complementary explanatory strengths.

Visual Question Answering

- Explanations can be integrated with a question answering system to provide justifications
- In case of conflict between different classifiers, explanations become critical



Credit: Braines et al.

Explainable Reinforcement Learning

- PIRL
- Hierarchical Policies
- Linear Model U-Trees

PIRL

Programmatically Interpretable Reinforcement Learning framework by Verma et al. (2018)

- A policy is represented using a high-level, domain-specific, human-readable programming language.
- Mimics Deep Reinforcement Learning model (DRL)
- *Neurally Directed Program Search(NDPS):* Uses DRL to compute a policy which is used as a neural ‘oracle’ to direct the policy search for a policy that is as close as possible to the neural oracle.

Hierarchical Policies

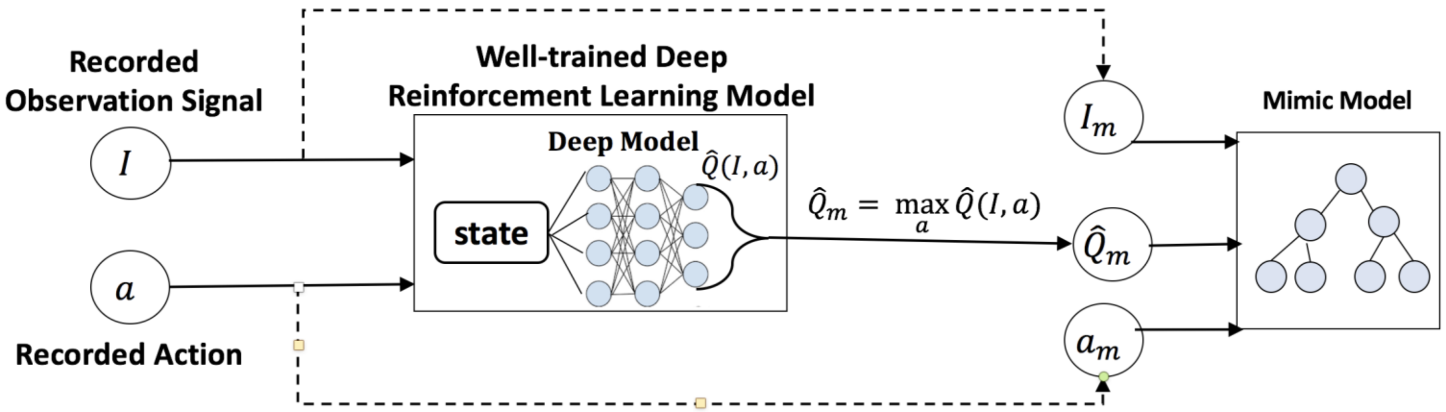
Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning by Shu et al.(2017)

- Complex task decomposed into several simpler subtasks.
- Each task is described by a human instruction
- Agents can only access learnt skills through these descriptions

Linear Model U-Trees

Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees by Liu et al.(2019)

- Approximates the predictions of an accurate, but complex model by mimicking the model's Q-function
- Records the state-action pairs and the resulting Q-values as 'soft supervision labels



References

- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017)
- Breiman, Leo.“Random Forests.” Machine Learning 45 (1). Springer: 5-32 (2001)
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. “Model Class Reliance: Variable importance measures for any machine learning model class, from the ‘Rashomon’ perspective.” http://arxiv.org/abs/1801.01489 (2018)
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)
- Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. (2017)
- Verma, A., Murali, V., Singh, R., Kohli, P., Chaudhuri, S.: Programmatically interpretable reinforcement learning. PMLR 80:5045-5054 (2018)
- Shu, T., Xiong, C., and Socher, R. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. (2017)
- Liu, G., Schulte, O., Zhu, W., Li, Q.: Toward interpretable deep reinforcement learning with linear model u-trees. In: Machine Learning and Knowledge Discovery in Databases, pp. 414–429. Springer International Publishing (2019)
- Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53 (https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf)
- Braines, Dave, Alun Preece, and Dan Harborne. "Multimodal explanations for ai-based multisensor fusion." NATO SET-262 RSM on artificial intelligence for military multisensor fusion engines. NATO (2018).

Thank you