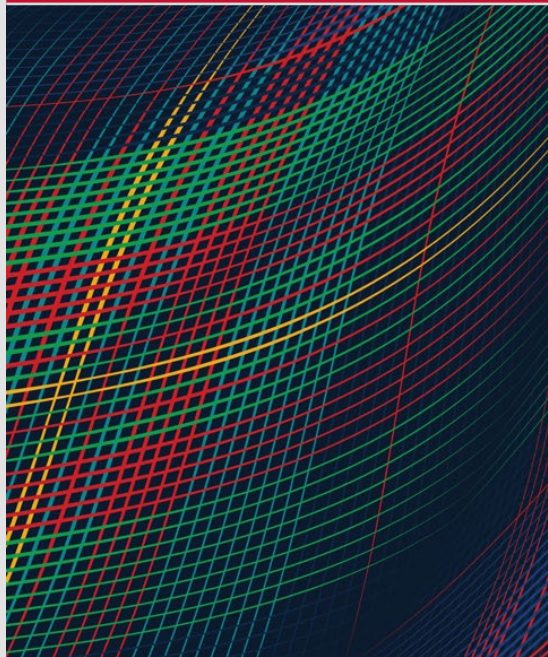
Carnegie Mellon University
Software Engineering Institute

Using end-to-end Causal Inference to Assess AI/ML Classifier Health

SEPTEMBER 17, 2024

Dr. Nicholas Testa
Senior Data Scientist



Document Markings

The following markings MUST be included in work product when attached to this form and when it is published.

For purposes of double anonymous peer review, markings may be temporarily omitted to ensure anonymity of the author(s).

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

This work product was created in part using generative AI.

DM24-1119

Who we are:

The Software Engineering Institute (SEI) was founded in 1984 as an independent entity that operates within Carnegie Mellon University (CMU) as a Federally Funded Research and Development Center (FFRDC) specializing in software engineering, artificial intelligence, and cybersecurity.

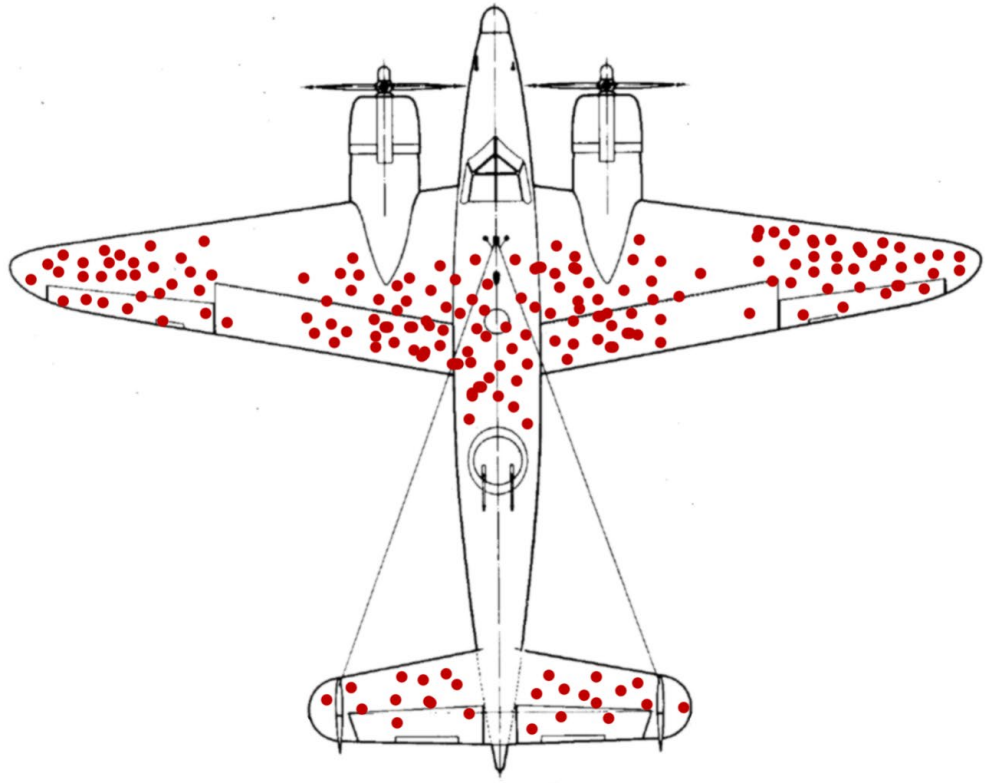
Our mission is to establish and advance software as a strategic advantage for national security. We lead and direct research and transition of software engineering and related disciplines at the intersection of academia, industry, and government.

Can you rely on your AI?

Carnegie Mellon University Software Engineering Institute

Every year, the DoD is **increasing** its **use** of AI/ML classifiers and predictors. However, AI classifiers are subject to a **lack of robustness** leading to a lack of trust.

Current test and evaluation methods are inadequate for ongoing evaluation of AI as they **heavily rely on correlations** within the data, which are undermined **by data/concept drift**, **evolving edge cases**, and **emerging phenomena**.



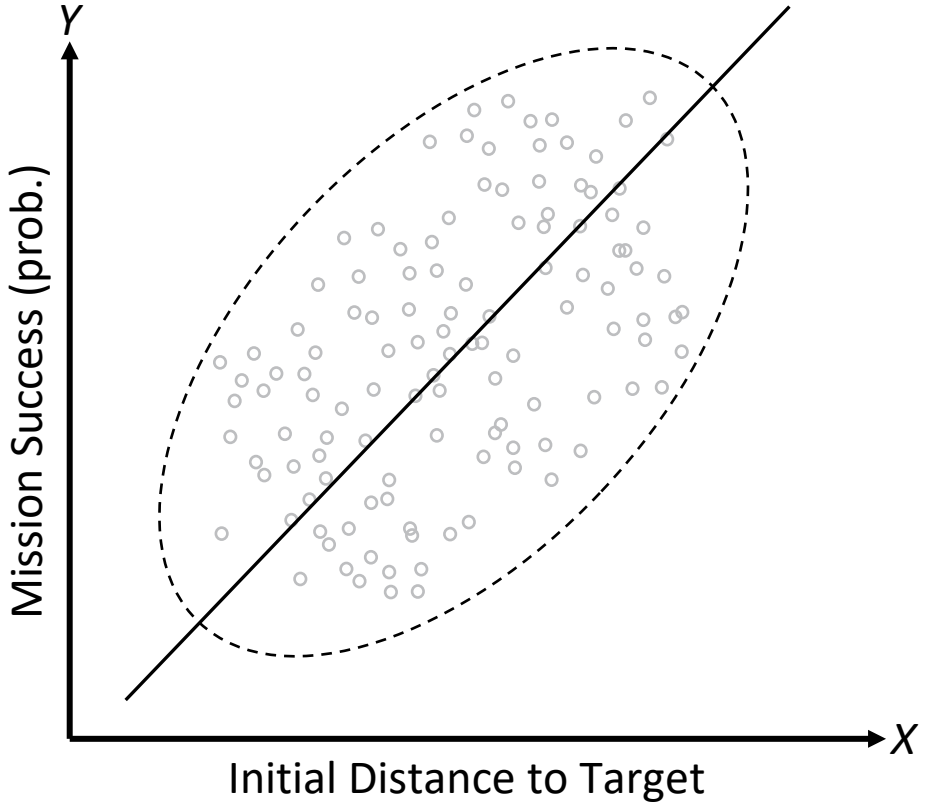
What's wrong with a little correlation?

AI/ML tools work by learning associations among attributes of objects to be classified and (typically) don't account for causation.

Traditional approaches to ML evaluation fail to account for underlying causal structure:

- Leaving alternative explanations for the impacts of a scenario unexplored
- Failing to account for key drivers
- Attributing causes to the wrong factors
- Failing to properly cross-validate their evaluation results

Resulting in a failure to identify where and when ML predictions can't be trusted



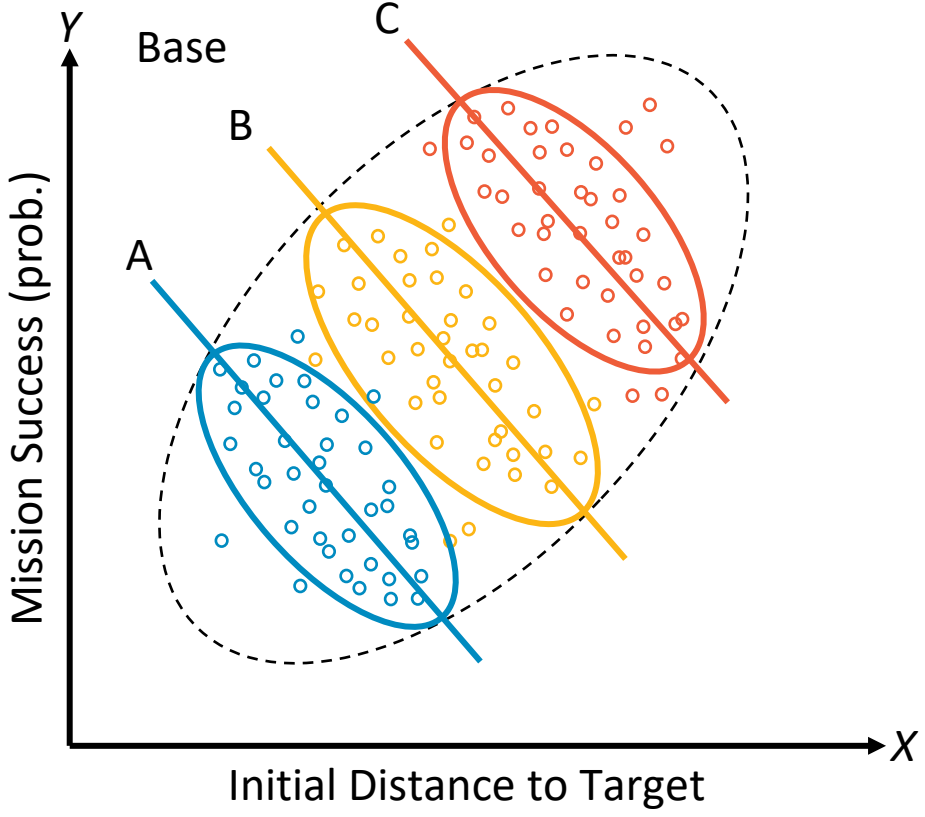
What's wrong with a little correlation?

AI/ML tools work by learning associations among attributes of objects to be classified and (typically) don't account for causation.

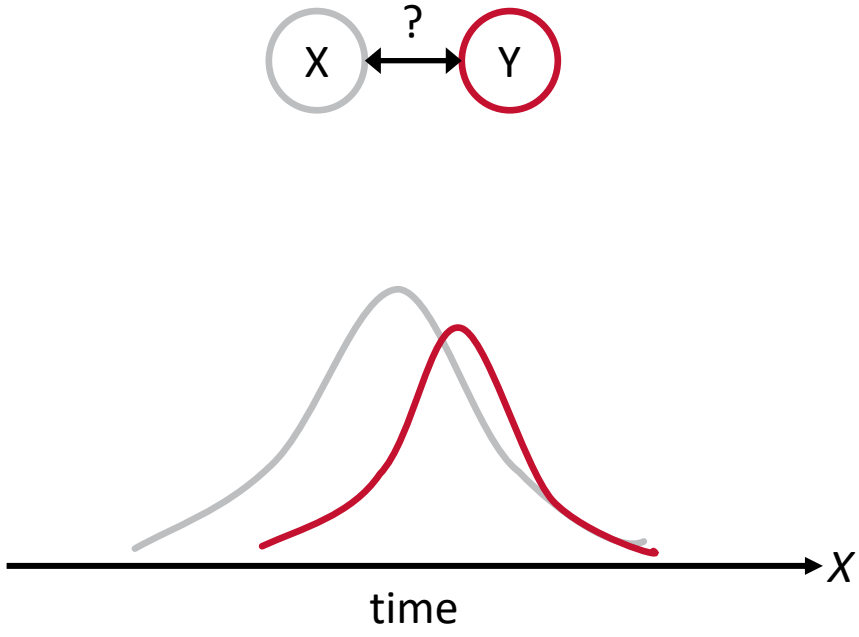
Traditional approaches to ML evaluation fail to account for underlying causal structure:

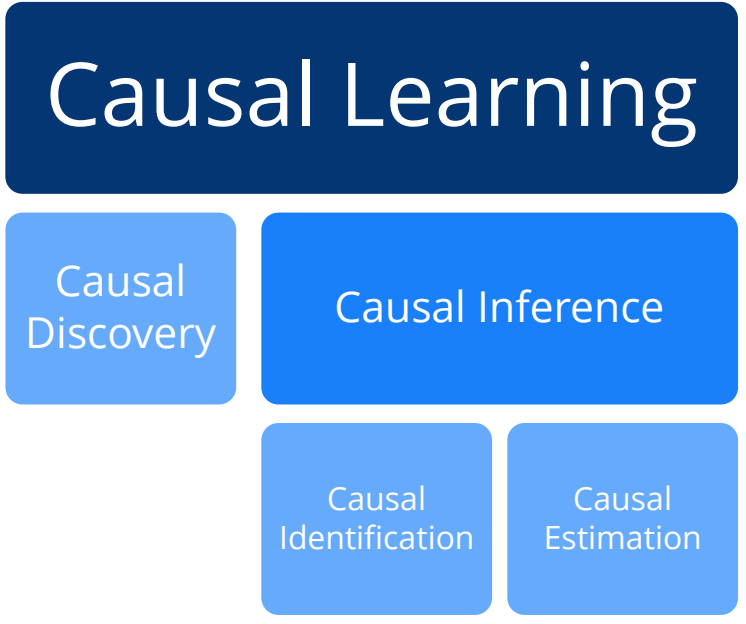
- Leaving alternative explanations for the impacts of a scenario unexplored
- Failing to account for key drivers
- Attributing causes to the wrong factors
- Failing to properly cross-validate their evaluation results

Resulting in a failure to identify where and when ML predictions can't be trusted

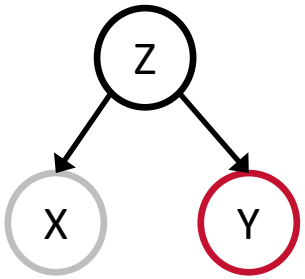


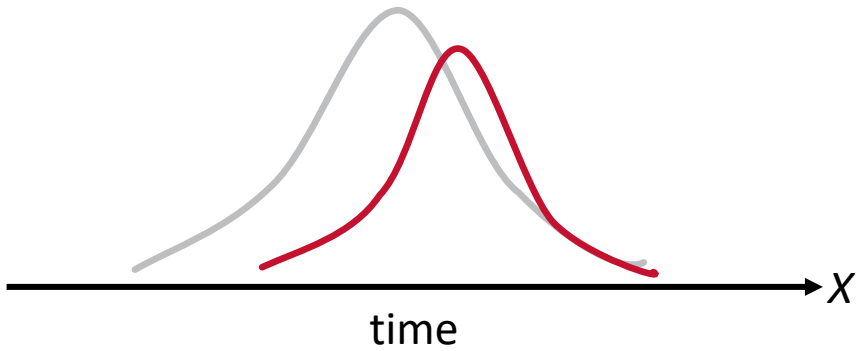
What is Causal Learning and how does it help?

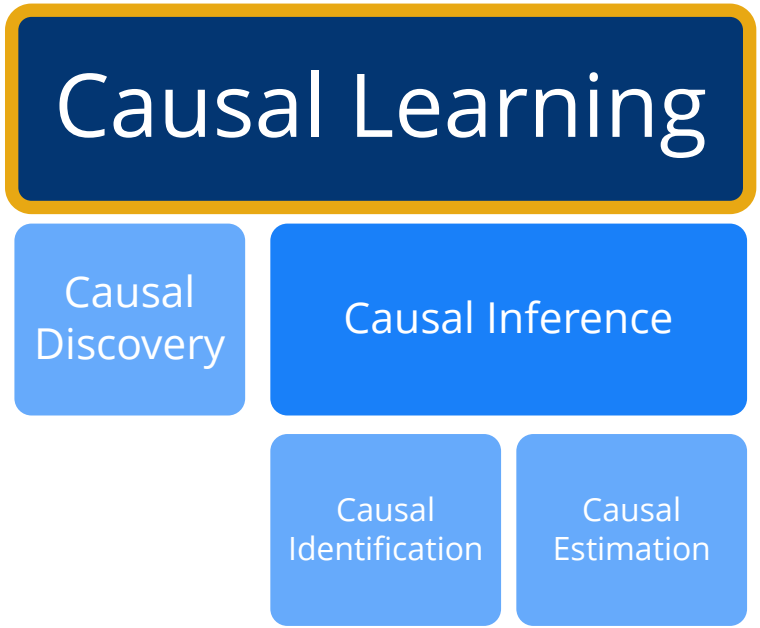




What is Causal Learning and how does it help?







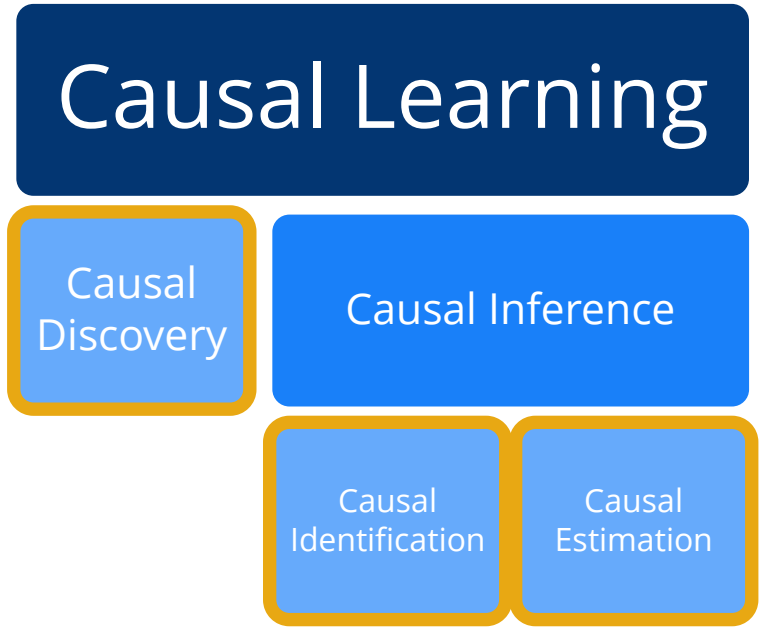
Calling in AIR support!

The SEI's new AI Robustness (AIR) helps answer the question of "*what effect **should** I be seeing on the outcome when I change the scenario and is this different from what my classifier is predicting?*"

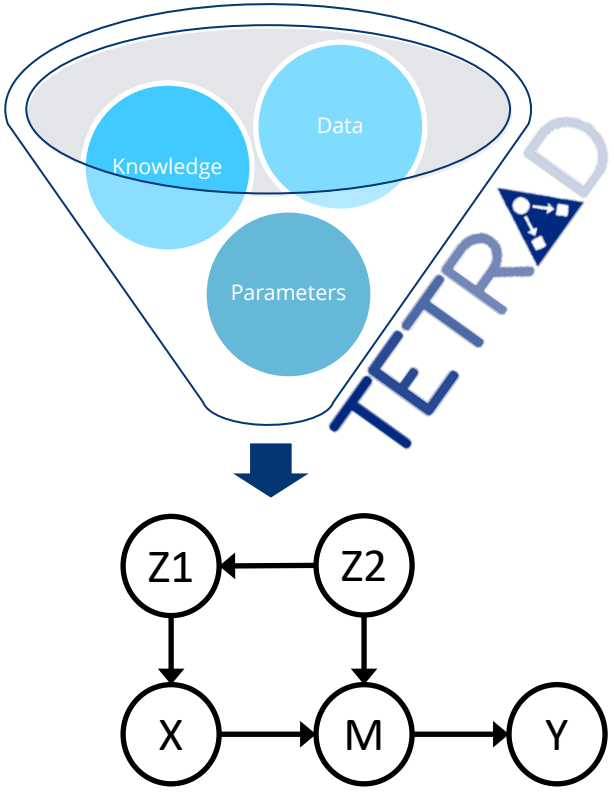
AIR applies an innovative mix of *causal discovery and inference* methods for:

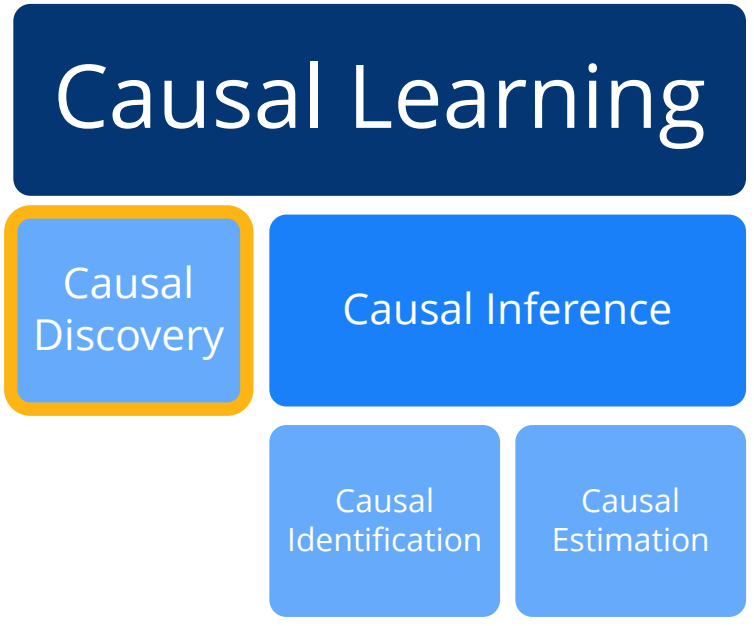
- evaluating the *robustness* of ML classifiers in dynamic contexts
- addressing *explainability* and *why*

So that decision makers can develop appropriate confidence in their model's predictions.



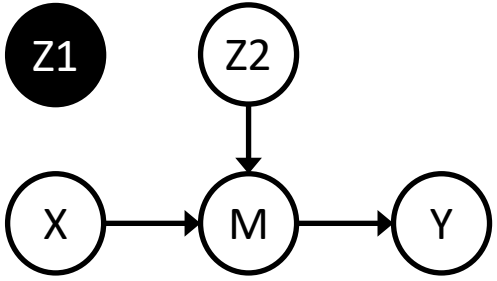
Step 1: Causal Discovery
Characterizing Key Causal Dependencies

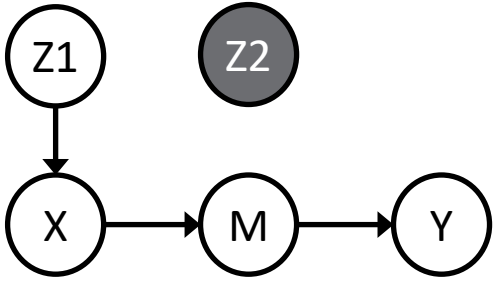


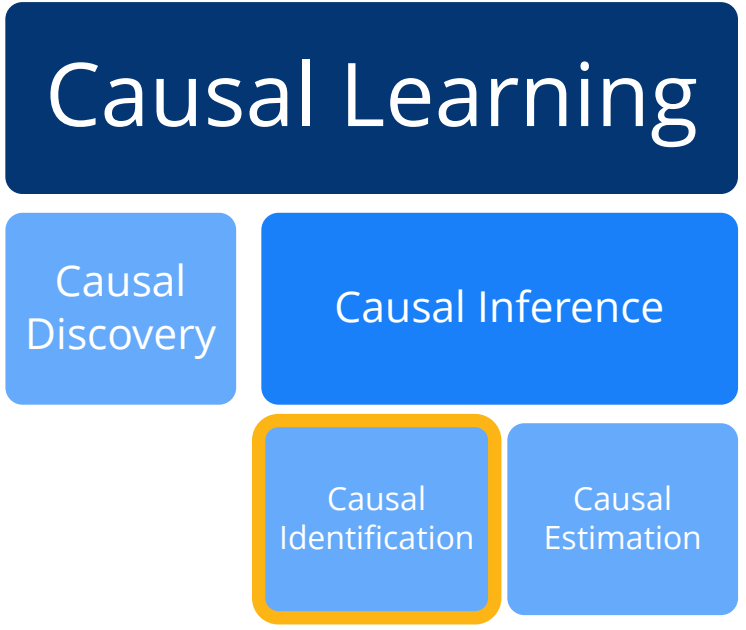


Step 2: Causal Identification

Providing Alternate Ways to Control Spurious Correlations

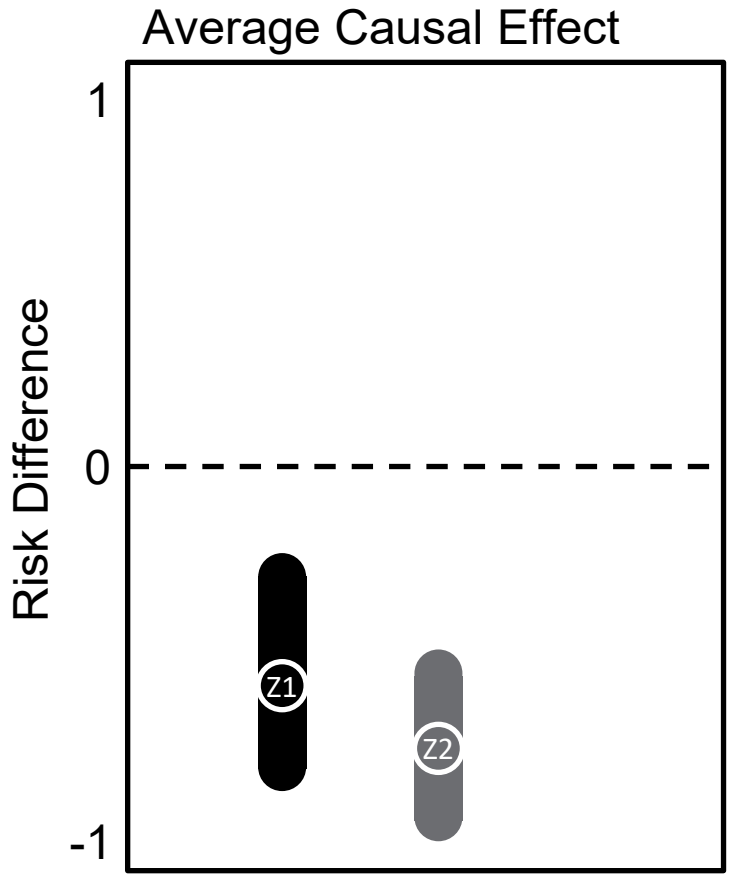


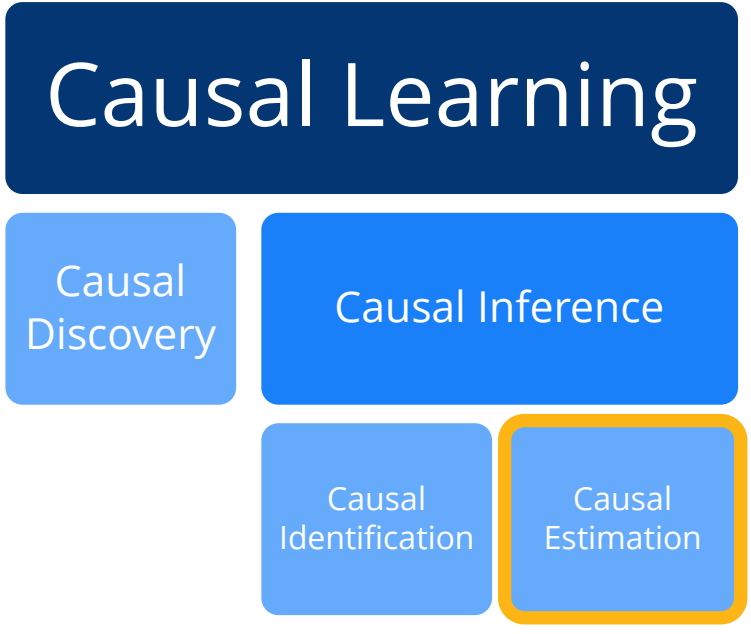




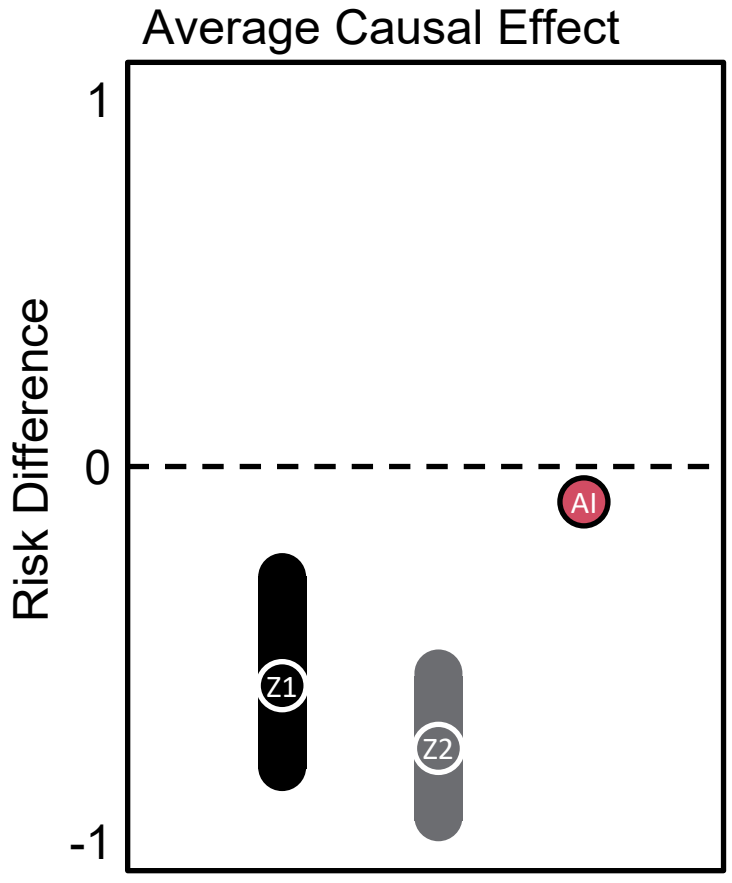
Step 3: Causal Estimation

Visualizing Potential Treatment Effects





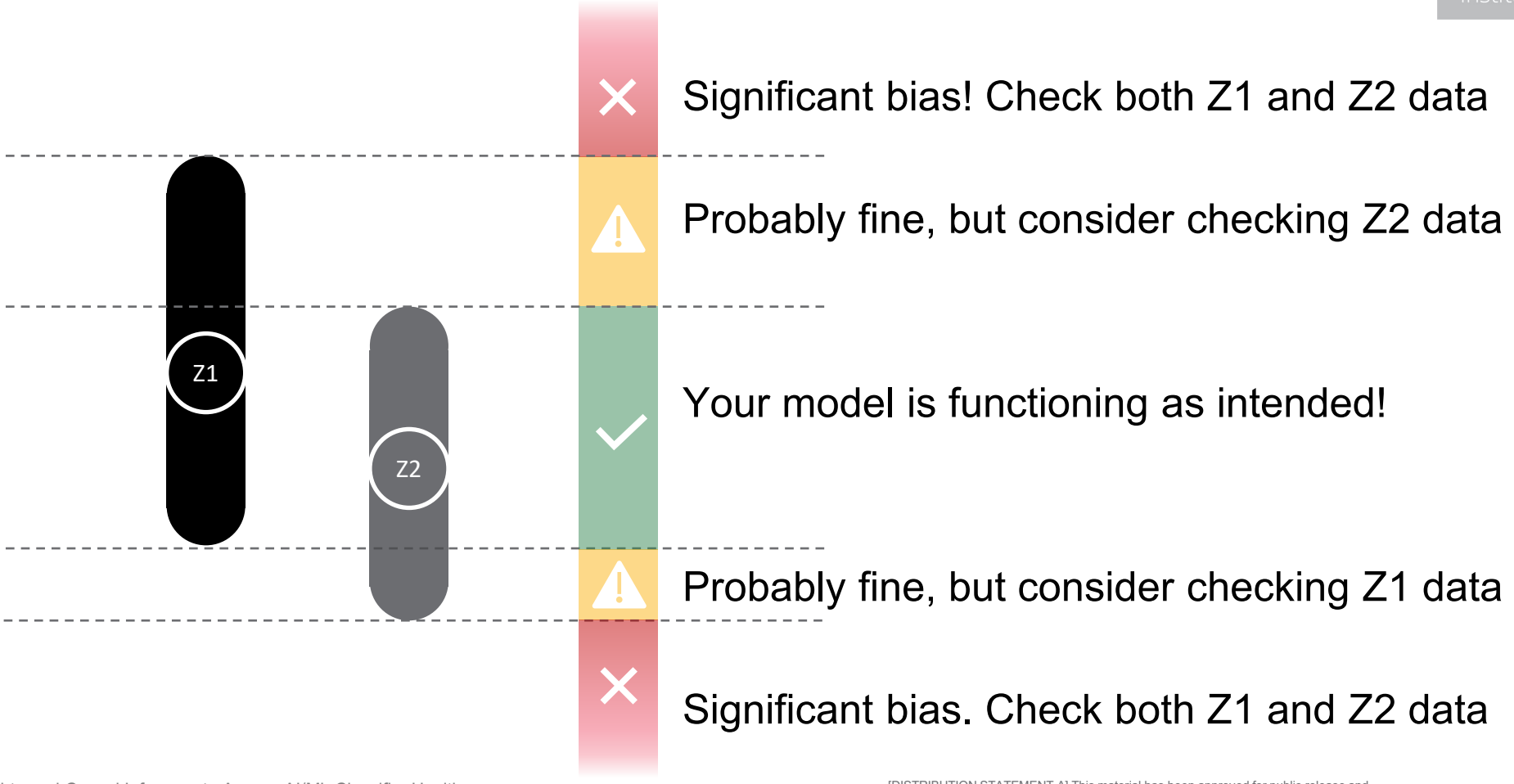
Applying Results of AIR



Assembling the final output requires that the original classifier's predicted risk difference be calculated and plotted alongside the causally-derived confidence intervals.

The position of the AI-classifier value relative to the causally-derived ACE estimates is what we're after.

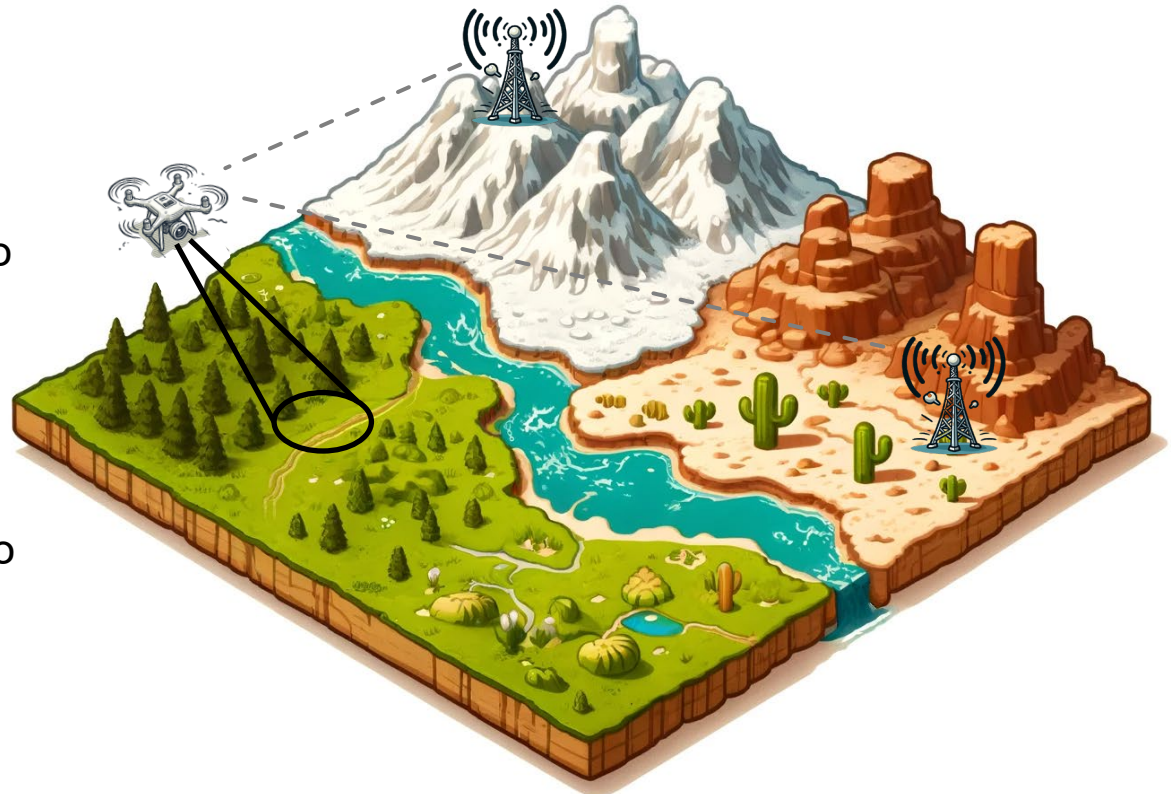
Applying Results of AIR



Example: UAV Demo

A unit within the DoD is tasked with deploying Unmanned Aerial Vehicles (UAVs) to acquire images in various locations of interest. There are two bases available to the UAV team, ‘Home’ and ‘Away,’ and they want to be able to predict mission success (acquiring images) based on which location the UAV starts the mission.

Naturally, they built a series of ML classifiers to predict mission success and kept the best performing one. That classifier, however, sometimes seems to produce biased results suggesting main base location has little effect.



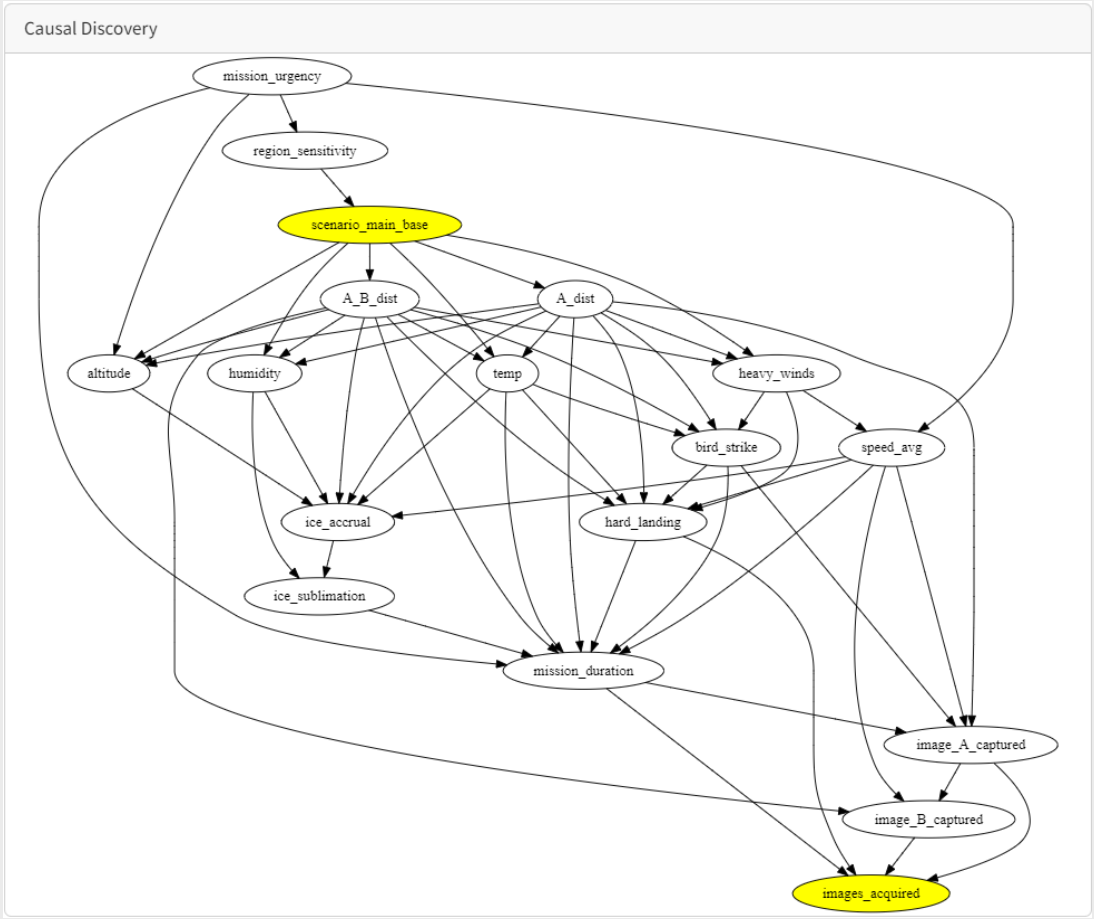
Graphic generated by combining several ChatGPT responses. Version 4o. OpenAI. June 2024. https://chat.openai.com/chat

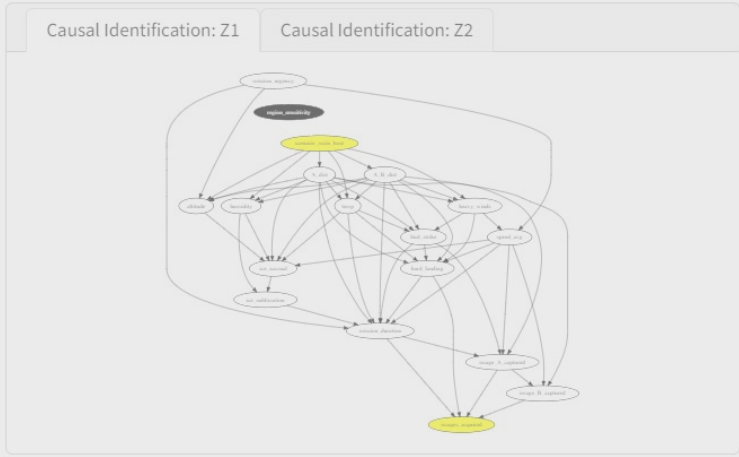
AIR Tool: Inputs

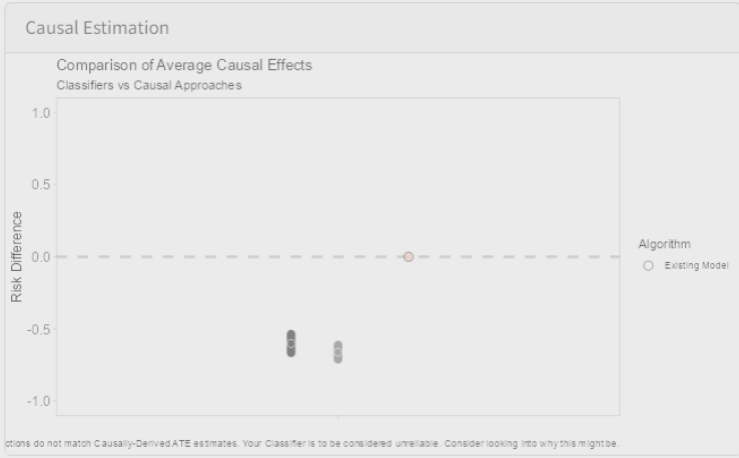
- **Data**: This is usually the processed data file used to train the original classifier.
- **Knowledge file**: This is a .csv file (right) representing a general hierarchy of known cause and effect relationships across all features. For example, temperature can affect ice accrual, but not the other way around.
- **Scenario Variable**: scenario_main_base
- **Outcome Variable**: images_acquired
- **Your model**: a saved model file or a pre-calculated average causal effect.

level	variable
0	mission_urgency
1	region_sensitivity
2	scenario_main_base
3	A_B_dist
3	A_dist
4	altitude
4	heavy_winds
4	humidity
4	temp
5	bird_strike
5	speed_avg
6	fuel_consumed
6	ice_accrual
7	hard_landing
7	ice_sublimation
8	image_A_captured
8	image_B_captured
8	mission_duration
8	images_acquired

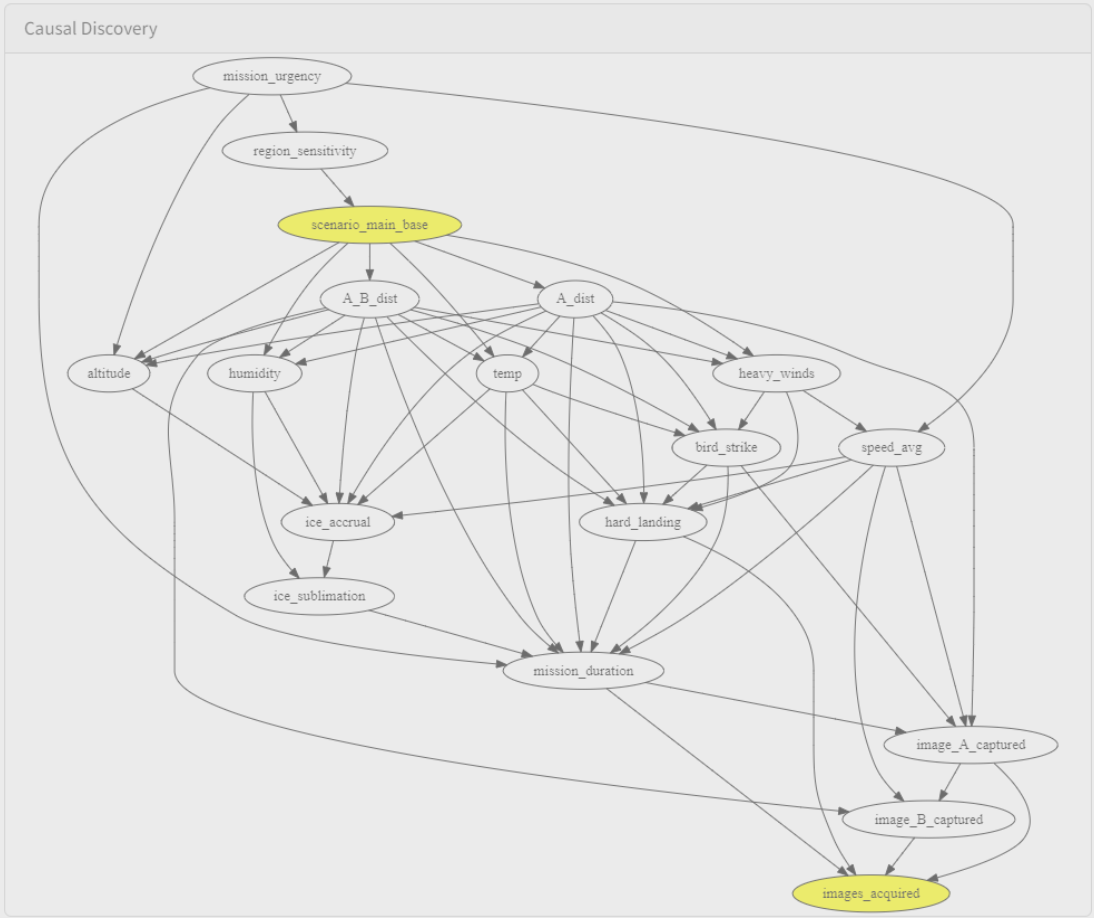
Example: UAV Demo

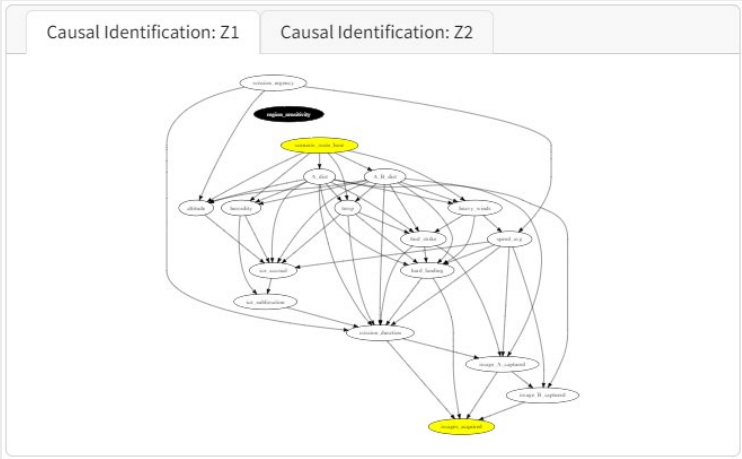




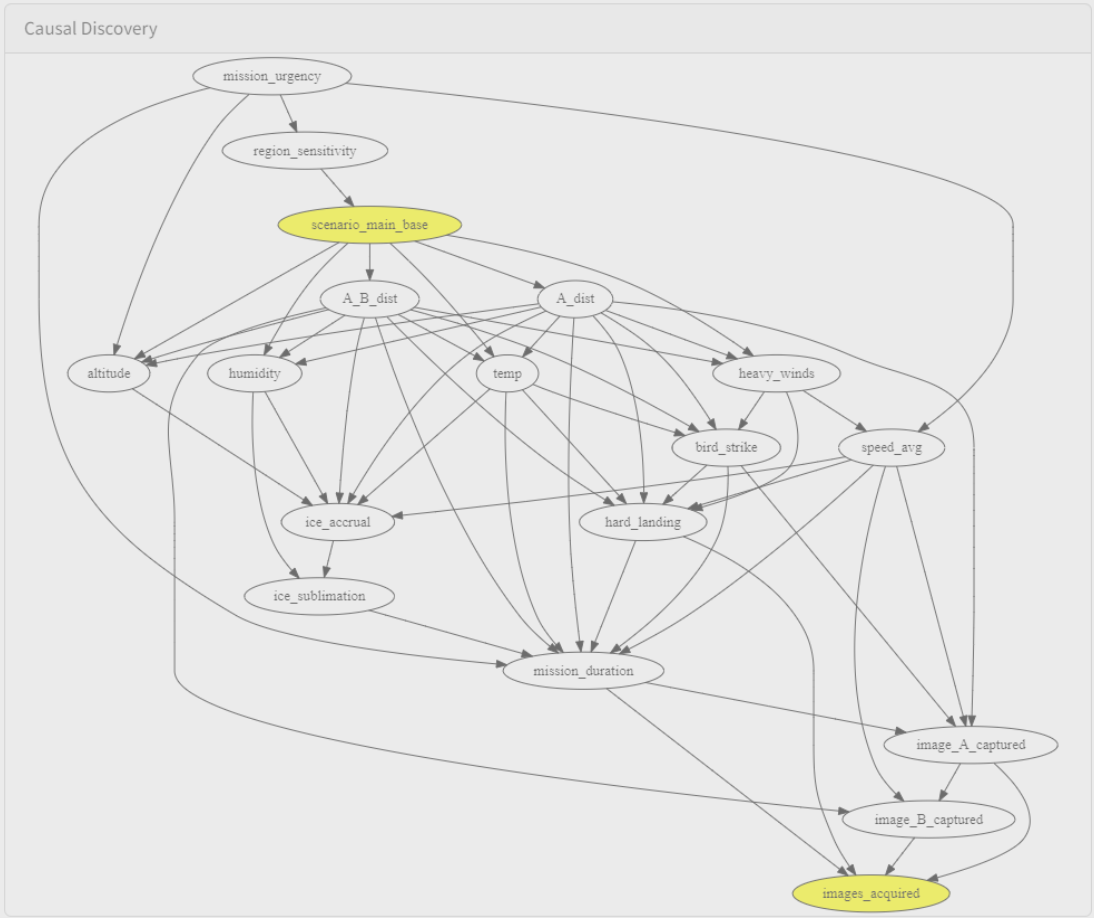


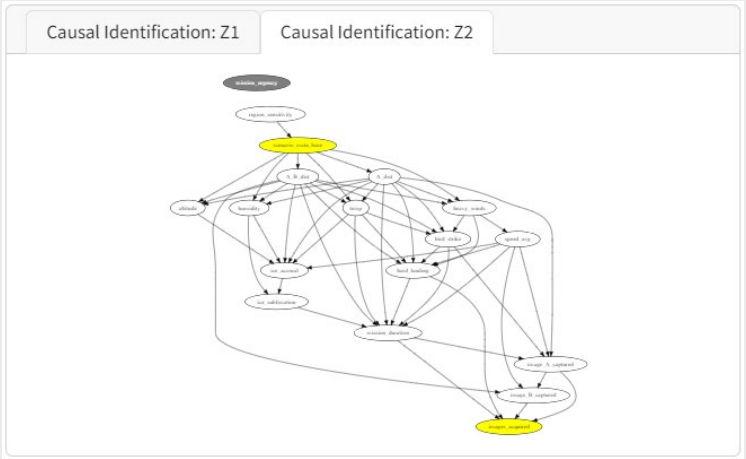
Example: UAV Demo



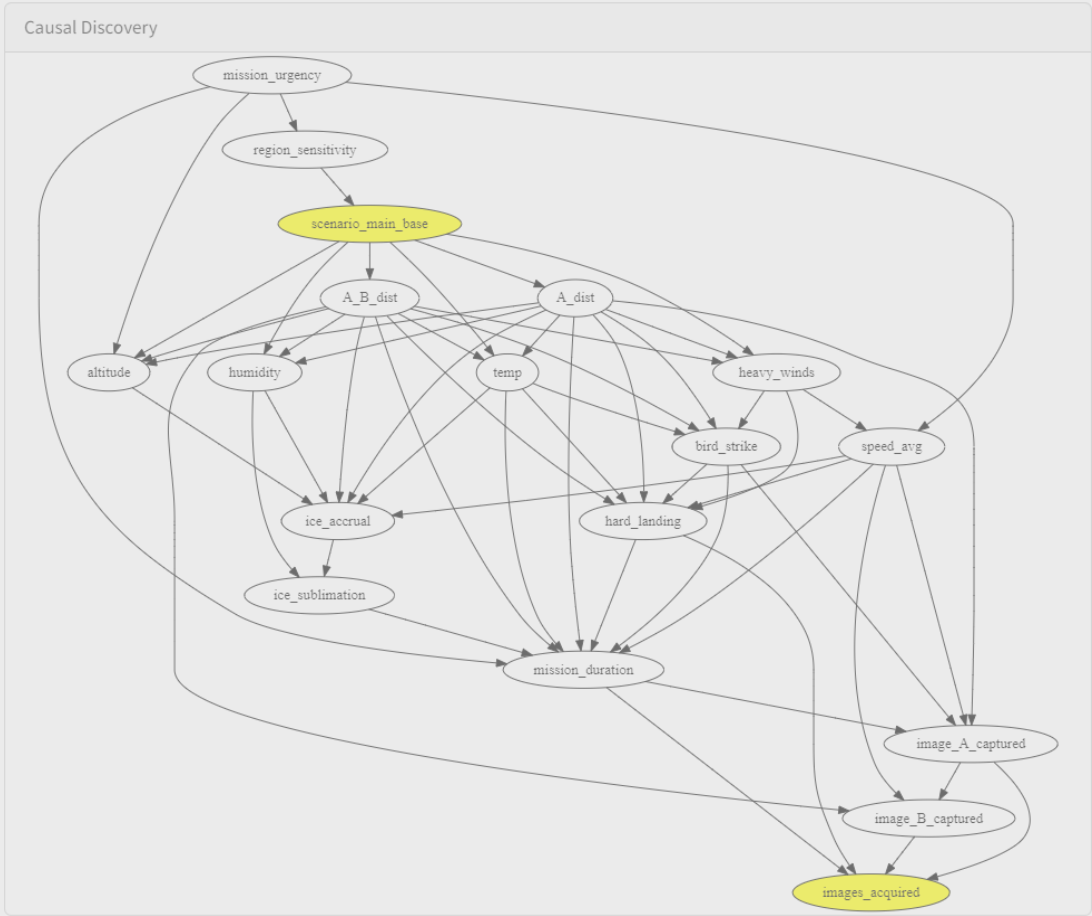


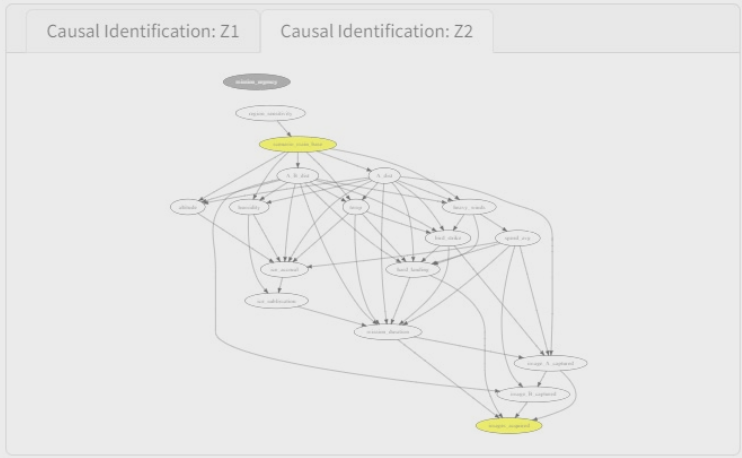
Example: UAV Demo

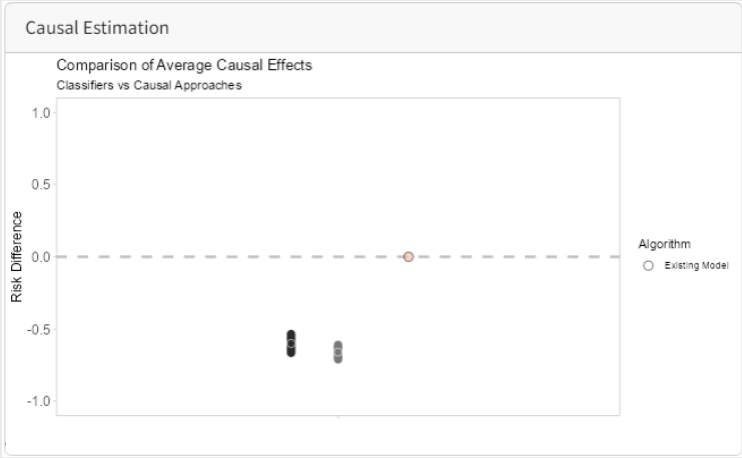




Example: UAV Demo



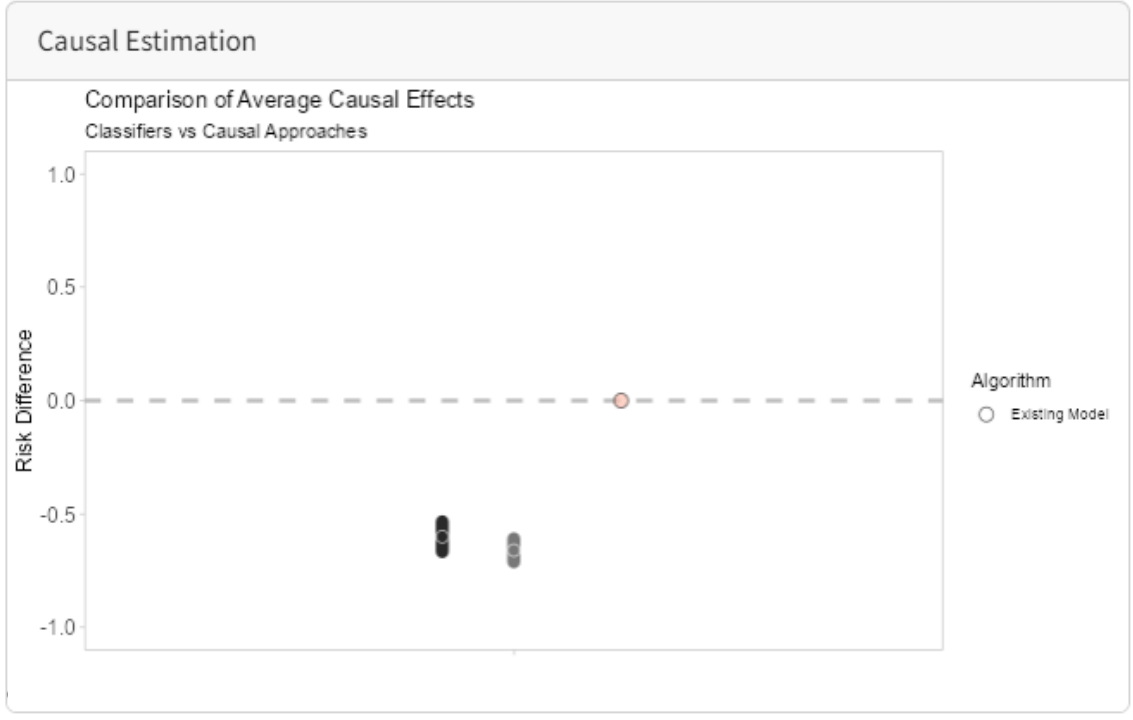




What has AIR taught us?

The ML classifier clearly falls outside the resulting causally-derived confidence intervals produced by the AIR Tool. Since both intervals are violated, the UAV mission planner should be wary of classifier predictions for this particular use case.

The two adjustment sets that are output of AIR (Step 2) provide recommendations of what variables/ features to focus on for subsequent classifier retraining, in this case region_sensitivity (Z1) and mission_urgency (Z2).



Next Steps

Carnegie Mellon University Software Engineering Institute

Hopefully, you now see the value in using an end-to-end Causal Learning solution like AIR for evaluating classifier performance. Drift and bias are significant threats to the longevity of ML/AI applications, and the AIR tool is a powerful resource for ensuring AI/ML classifier robustness.

This tool is still a work in progress. If you or someone you know might be interested in working with us to help make it better, we are looking for partners to test and collaborate with!

info@sei.cmu.edu

Want to learn more?

Causal Discovery:

- https://www.cmu.edu/dietrich/philosophy/tetrad/

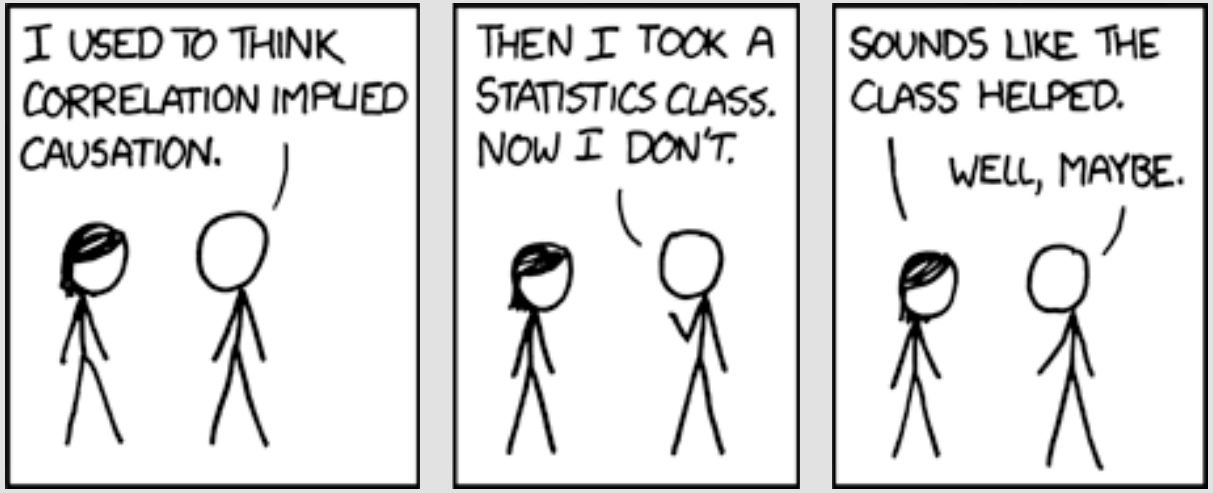
Causal Inference:

- Causal Inference in Statistics: A Primer, Judea Pearl, Madelyn Glymour and Nicholas P. Jewell, 2019

Causal Estimation:

- https://tlverse.org/tlverse-handbook/tlverse.html

Questions?



Source: Randall Munroe. “Correlation.” *XKCD*. CC BY-NC 2.5.