

Towards Reliable Testing of Machine Unlearning

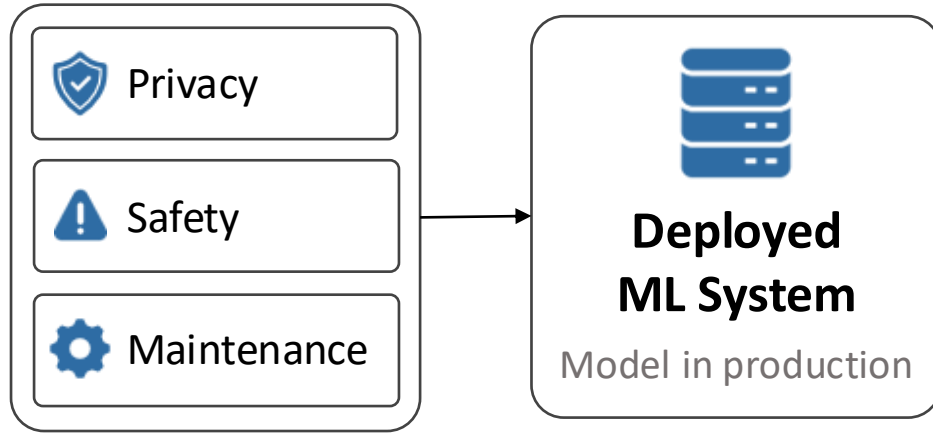
Anna Mazhar • Sainyam Galhotra

annam@cs.cornell.edu, sg@cs.cornell.edu

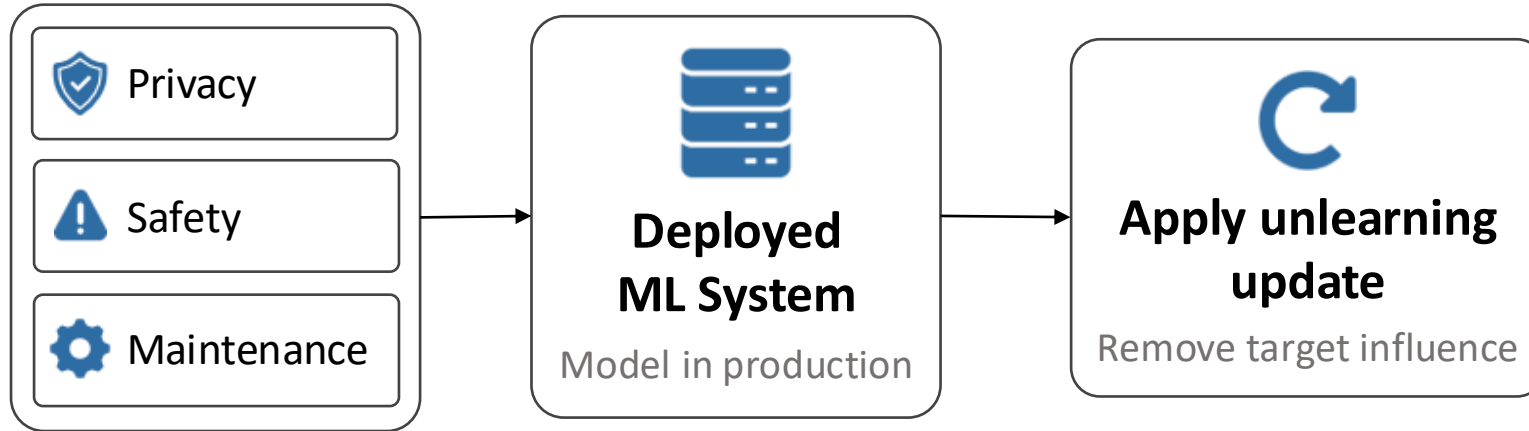
Cornell University



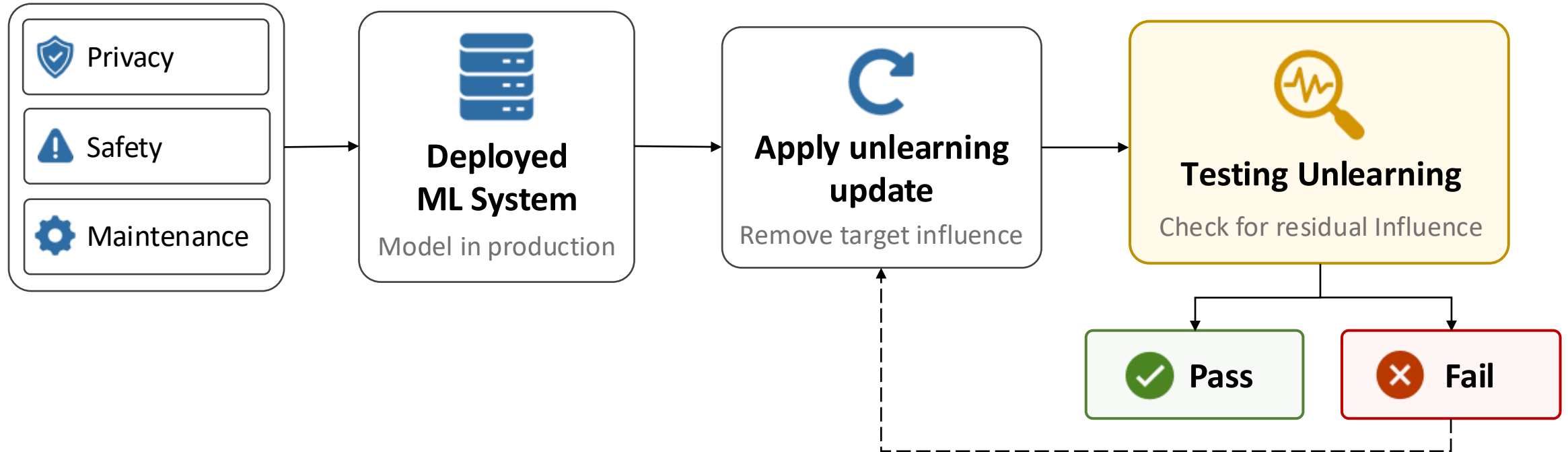
Machine unlearning is a software patch



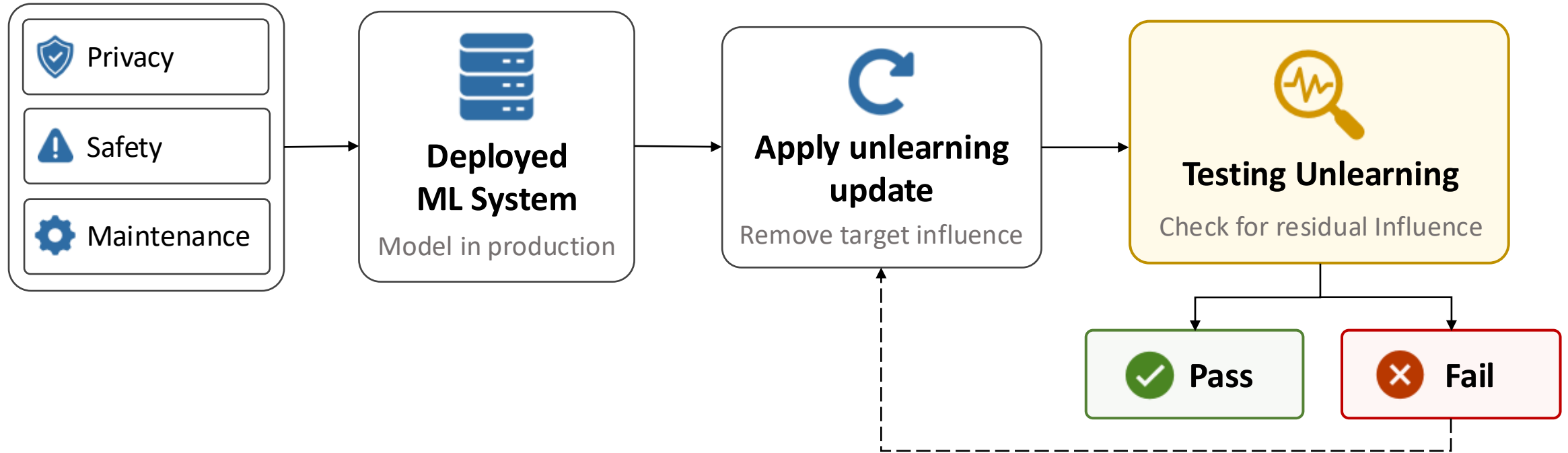
Machine unlearning is a software patch



Machine unlearning is a software patch



Machine unlearning is a software patch



Unlearning is a software change, and software changes need tests.

Why testing unlearning is hard



Problem: AI model for lung-disease (using X-ray & metadata) learns to associate **Hospital A** with **higher risk**.

Regulator's request: *"Unlearn references to the hospital identifier"*

After unlearning, hospital IDs are gone. Are we done?

Why testing unlearning is hard



Problem: AI model for lung-disease (using X-ray & metadata) learns to associate **Hospital A** with **higher risk**.

Regulator's request: *"Unlearn references to the hospital identifier"*

After unlearning, hospital IDs are gone. Are we done?

No -- influence reroutes!

Why testing unlearning is hard



Problem: AI model for lung-disease (using X-ray & metadata) learns to associate **Hospital A** with **higher risk**.

Regulator's request: *"Unlearn references to the hospital identifier"*

After unlearning, hospital IDs are gone. Are we done?

No -- influence reroutes!

Structured proxies

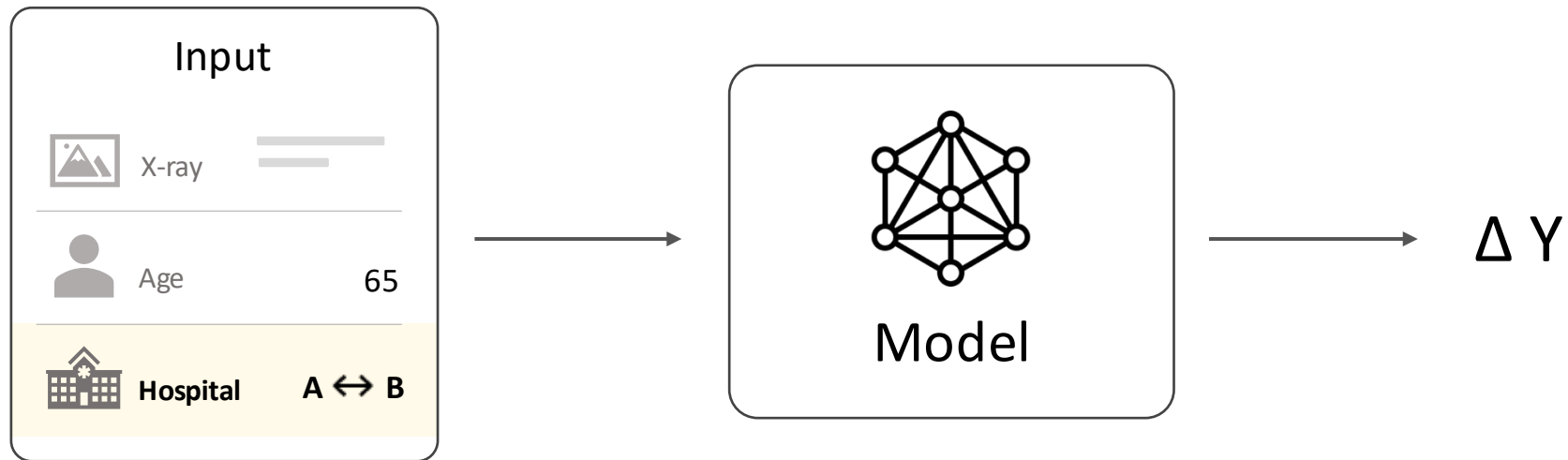
Zip code, insurance type, referral pattern still encode where a patient is treated.

Unstructured proxies

Scanner settings, contrast curves and reconstruction artifacts leave hospital specific signatures

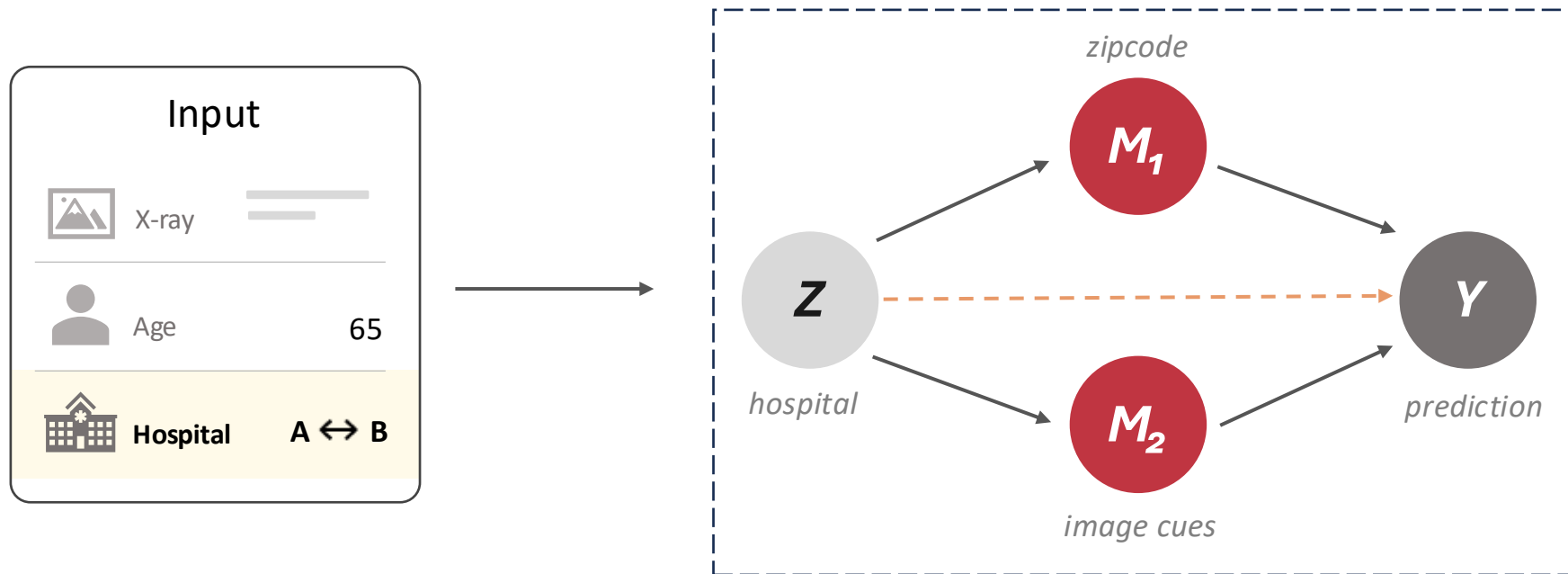
Forgetting “Hospital”

After unlearning Z, changing Z in the input should not change Y.



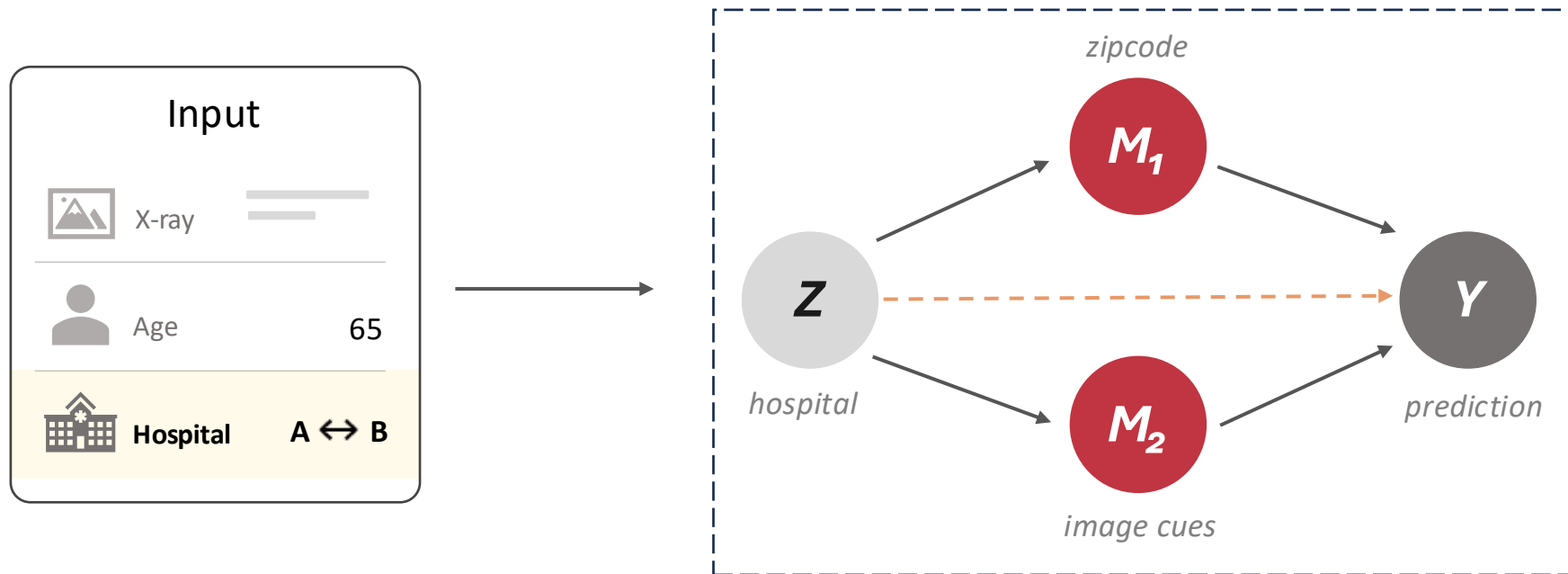
Forgetting “Hospital” as a causal claim

Representing the predictor as a **causal model**



Forgetting “Hospital” as a causal claim

Representing the predictor as a **causal model**



✓ Direct path: $Z \rightarrow Y \approx 0$

✓ Indirect path: $Z \rightarrow M \rightarrow Y \approx 0$

Causal fuzzing: a four-step workflow

STEP 1



Scope

Identify candidate
mediators M
e.g., zipcode, image cues

Causal fuzzing: a four-step workflow

STEP 1



Scope

Identify candidate
mediators M
e.g., zipcode, image cues



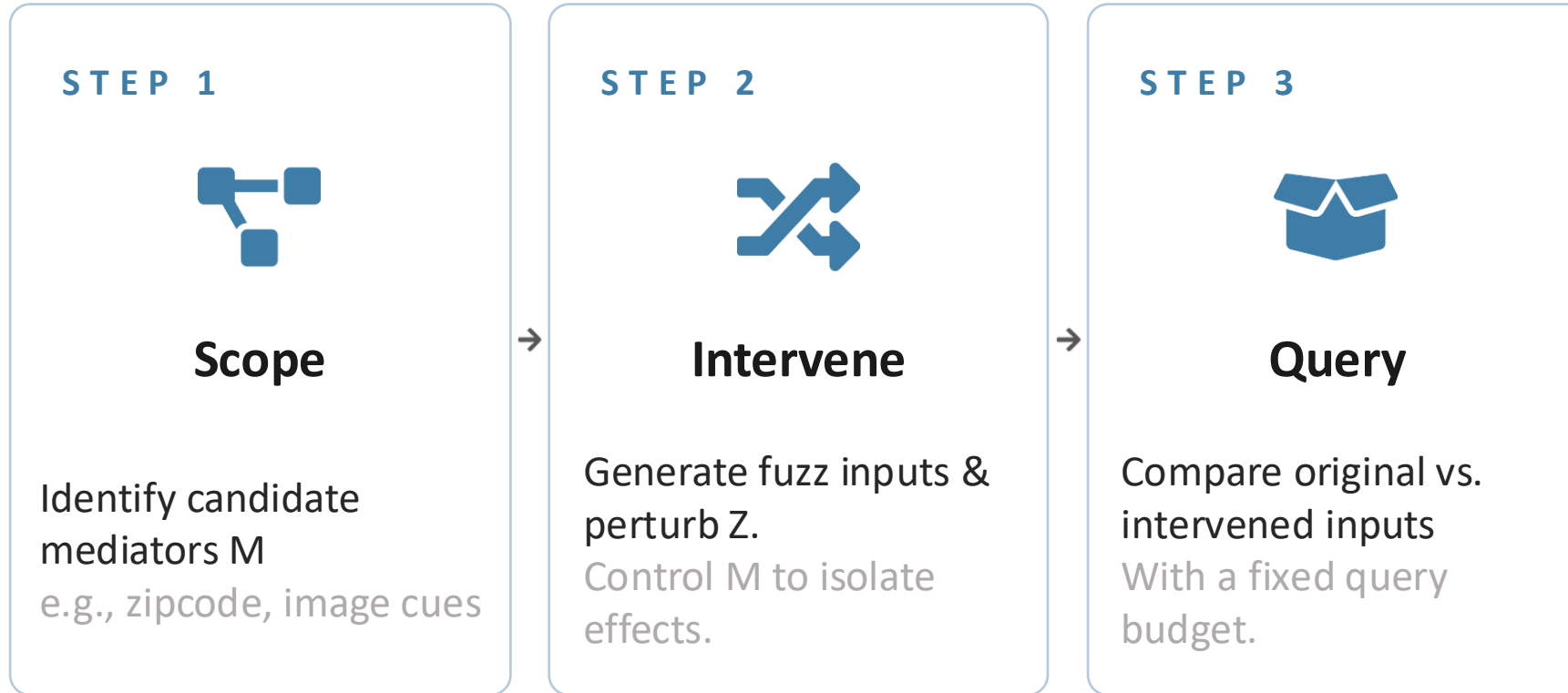
STEP 2



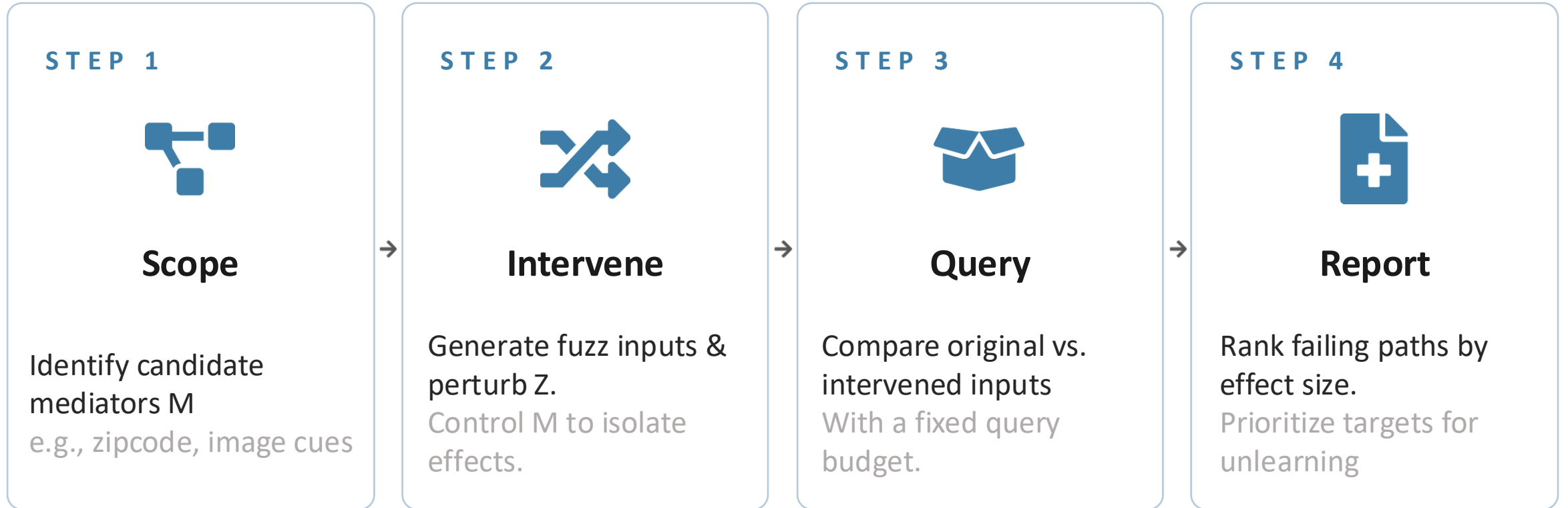
Intervene

Generate fuzz inputs &
perturb Z.
Control M to isolate
effects.

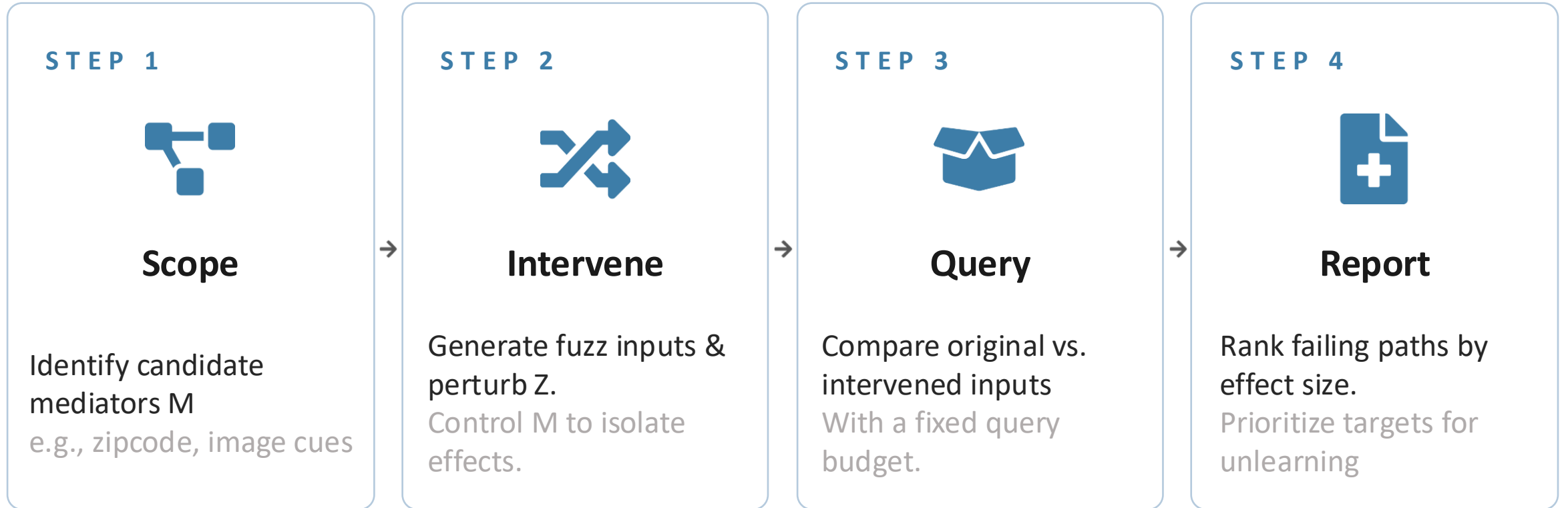
Causal fuzzing: a four-step workflow



Causal fuzzing: a four-step workflow



Causal fuzzing: a four-step workflow



Systematically testing for residual influence across all causal paths

Proof of concept: causal fuzzing insights



Datasets

Adult Income . Heart Disease . Drug Consumption



Baselines

SHAP . Permutation Imp. AIF 360



Unlearning

Remove Z + retrain

Proof of concept: causal fuzzing insights



Datasets

Adult Income . Heart Disease . Drug Consumption



Baselines

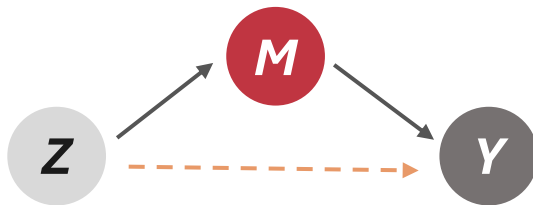
SHAP . Permutation Imp. AIF 360



Unlearning

Remove Z + retrain

Proxy pathways



Influence survives through
correlated mediators

Proof of concept: causal fuzzing insights



Datasets

Adult Income . Heart Disease . Drug Consumption



Baselines

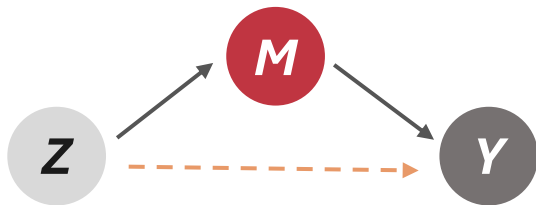
SHAP . Permutation Imp. AIF 360



Unlearning

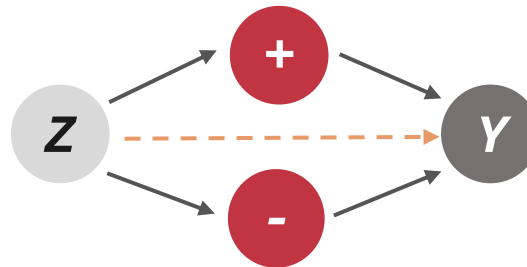
Remove Z + retrain

Proxy pathways



Influence survives through correlated mediators

Cancellation illusion



Strong and positive influence cancel out

Proof of concept: causal fuzzing insights



Datasets

Adult Income . Heart Disease . Drug Consumption



Baselines

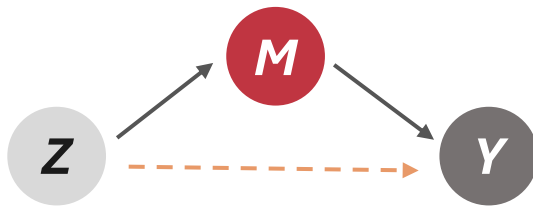
SHAP . Permutation Imp. AIF 360



Unlearning

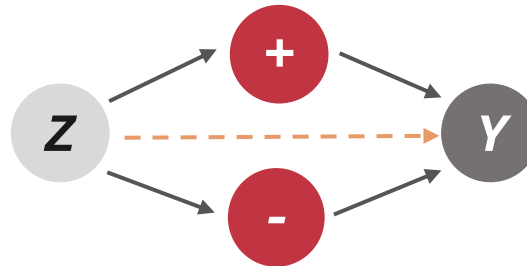
Remove Z + retrain

Proxy pathways



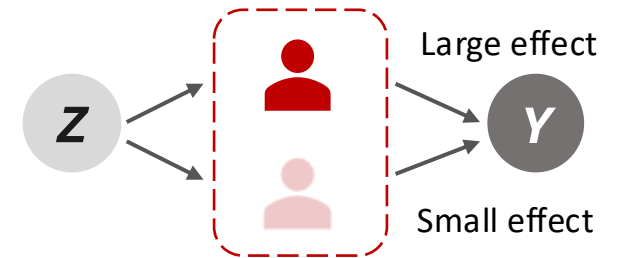
Influence survives through correlated mediators

Cancellation illusion



Strong and positive influence cancel out

Subgroup masking



Large residual influence remains in subgroup

Proof of concept: causal fuzzing insights



Datasets

Adult Income . Heart Disease . Drug Consumption



Baselines

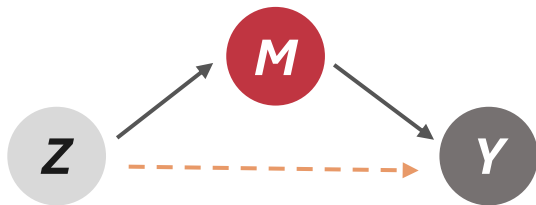
SHAP . Permutation Imp. AIF 360



Unlearning

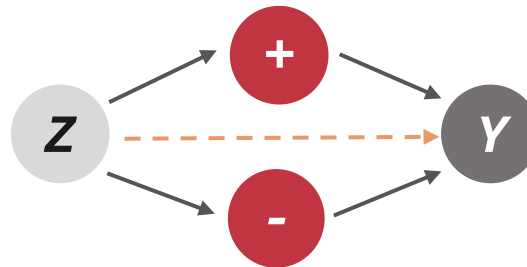
Remove Z + retrain

Proxy pathways



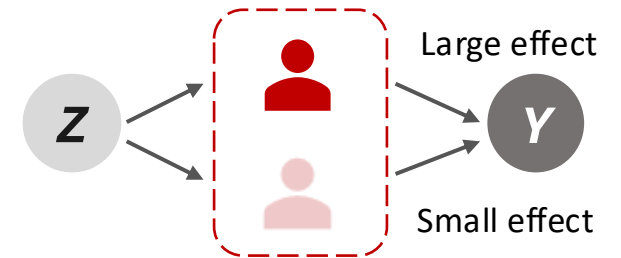
Influence survives through correlated mediators

Cancellation illusion



Strong and positive influence cancel out

Subgroup masking



Large residual influence remains in subgroup

Aggregate scoring can hide away where the residual influence survives

Open challenges & future directions

? Imperfect causal knowledge

The causal graph may be incomplete or wrong. How does uncertainty in the graph affect coverage and confidence?

Open challenges & future directions



Imperfect causal knowledge

The causal graph may be incomplete or wrong. How does uncertainty in the graph affect coverage and confidence?



Budgeted regression suites

Exhaustive pathway testing is costly. How should testing balance coverage, risk, and computational budget?

Open challenges & future directions



Imperfect causal knowledge

The causal graph may be incomplete or wrong. How does uncertainty in the graph affect coverage and confidence?



Budgeted regression suites

Exhaustive pathway testing is costly. How should testing balance coverage, risk, and computational budget?



Unstructured data

Mediators are difficult to define in text and images. What constitutes a meaningful and valid intervention across modalities?

Open challenges & future directions



Imperfect causal knowledge

The causal graph may be incomplete or wrong. How does uncertainty in the graph affect coverage and confidence?



Budgeted regression suites

Exhaustive pathway testing is costly. How should testing balance coverage, risk, and computational budget?



Unstructured data

Mediators are difficult to define in text and images. What constitutes a meaningful and valid intervention across modalities?



Foundation models

Results may vary across prompts and contexts. How robust are pathway-level conclusions for LLMs and VLMs?

Conclusion



Scan to connect!

- **Unlearning testing is an SE problem.**
 - Not a downstream ML evaluation. Treat it like fault localization, coverage, and regression — under a query budget.
- **Attribution misses indirect leakage.**
 - Proxy paths, cancelling mediators, and subgroup-specific dependence survive unlearning and survive SHAP.
- **Causal fuzzing gives actionable evidence.**
 - Path-targeted, budgeted, black-box. Output: a leakage report a developer can act on.