

Discovering Causal Signals in Images

David Lopez-Paz
Facebook AI Research
dlp@fb.com

Robert Nishihara
UC Berkeley
rkn@eecs.berkeley.edu

Soumith Chintala
Facebook AI Research
soumith@fb.com

Bernhard Schölkopf
MPI for Intelligent Systems
bs@tue.mpg.de

Léon Bottou
Facebook AI Research
leon@bottou.org

Nasim Rahaman (HCI / IWR)
Seminar on: *Explainable Machine Learning*

Q & A

To Build Truly Intelligent Machines, Teach Them Cause and Effect

 51 | 

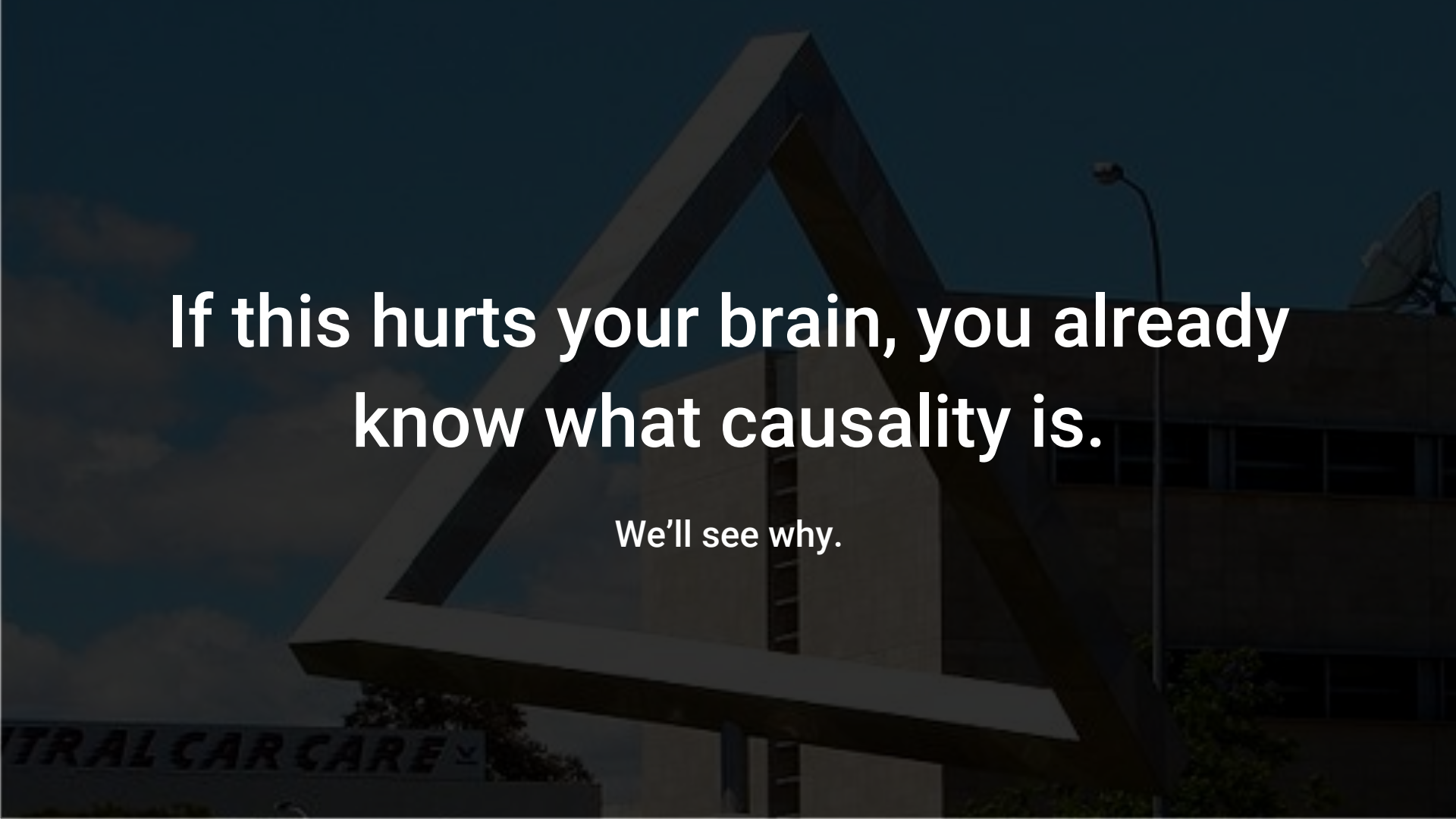
Judea Pearl, a pioneering figure in artificial intelligence, argues that AI has been stuck in a decades-long rut. His prescription for progress? Teach machines to understand the question why.

Motivation

What is
Causality?

A large, white, three-dimensional Penrose triangle sculpture stands in the foreground. The sculpture is a classic impossible object, appearing as a solid triangular frame. Behind it is a modern, multi-story building with a light-colored, textured facade and several windows. A tall, thin street lamp is positioned to the right of the sculpture. The sky is blue with scattered white clouds. In the bottom left corner, a sign for 'CENTRAL CAR CARE' is partially visible.

CENTRAL CAR CARE



**If this hurts your brain, you already
know what causality is.**

We'll see why.

Why Causality?

Consider two random variables: the **Altitude** (A) and the **Temperature** (T) of a city in Austria. The corresponding joint distribution can be expressed as:

$$P(\mathbf{A}, \mathbf{T}) = P(\mathbf{A} \mid \mathbf{T}) P(\mathbf{T})$$

-- or equivalently --

$$P(\mathbf{T}, \mathbf{A}) = P(\mathbf{T} \mid \mathbf{A}) P(\mathbf{A})$$

But are they *intuitively* equivalent? Intuitively, does $P(\mathbf{T} \mid \mathbf{A})$ feel more or less *fundamental* than $P(\mathbf{A} \mid \mathbf{T})$?

Why Causality?

Somehow, $P(\mathbf{T} \mid \mathbf{A})$ *feels right* (hopefully), in the sense that:

* if we manage to **magically elevate** a city while keeping all other *laws of physics* constant, we would expect the **temperature** of that city to drop.

* if we manage to **magically cool** the entire city, the remaining laws of physics do not imply that the city **elevates** itself.

There is this asymmetry that we intuitively latch on to - in doing so, we infer causally that **altitude** causes **temperature**, or $\mathbf{A} \rightarrow \mathbf{T}$. In other words, **A** is the cause and **T** is the effect.

Fine-print: The asymmetry need not be temporal.

Magical Interventions

All we have to do now is to replace the word “magically” (we don’t do that around here) with *interventionally* and we’re on to something. In math:

If $A \rightarrow T$:

$$P(A \mid do(T)) = P(A)$$

but not vice versa.

$do(X)$ is the act of performing a *localized intervention* on the random variable X (i.e. magically changing its value without affecting any other *laws of physics*).

But what makes a *law of physics*?

The Generating Mechanism and Structured Causal Models

When we talked about *the laws of physics*, what we meant more generally was the **mechanism** that generated the effect **E** (**temperature**) from the cause **C** (**altitude**). The big idea here is that the **mechanism** that we use to generate effect from cause itself does not depend on the cause.

In math,

$$\begin{aligned} \mathbf{C} &\sim P(\mathbf{C}) \\ \mathbf{E} &= \mathbf{f}_{Z \sim P(Z)}(\mathbf{C}) \end{aligned}$$

where \mathbf{f}_Z is a deterministic mechanism function of the cause **C** and **Z** models stochasticity in the mechanism independent of the cause, i.e $P(\mathbf{C}) \perp P(\mathbf{Z})$.

A close-up shot of Morpheus from the movie The Matrix. He is wearing his signature black sunglasses, which reflect the faces of Neo and Trinity. The background is a blurred green, suggesting an outdoor setting. The overall tone is serious and contemplative.

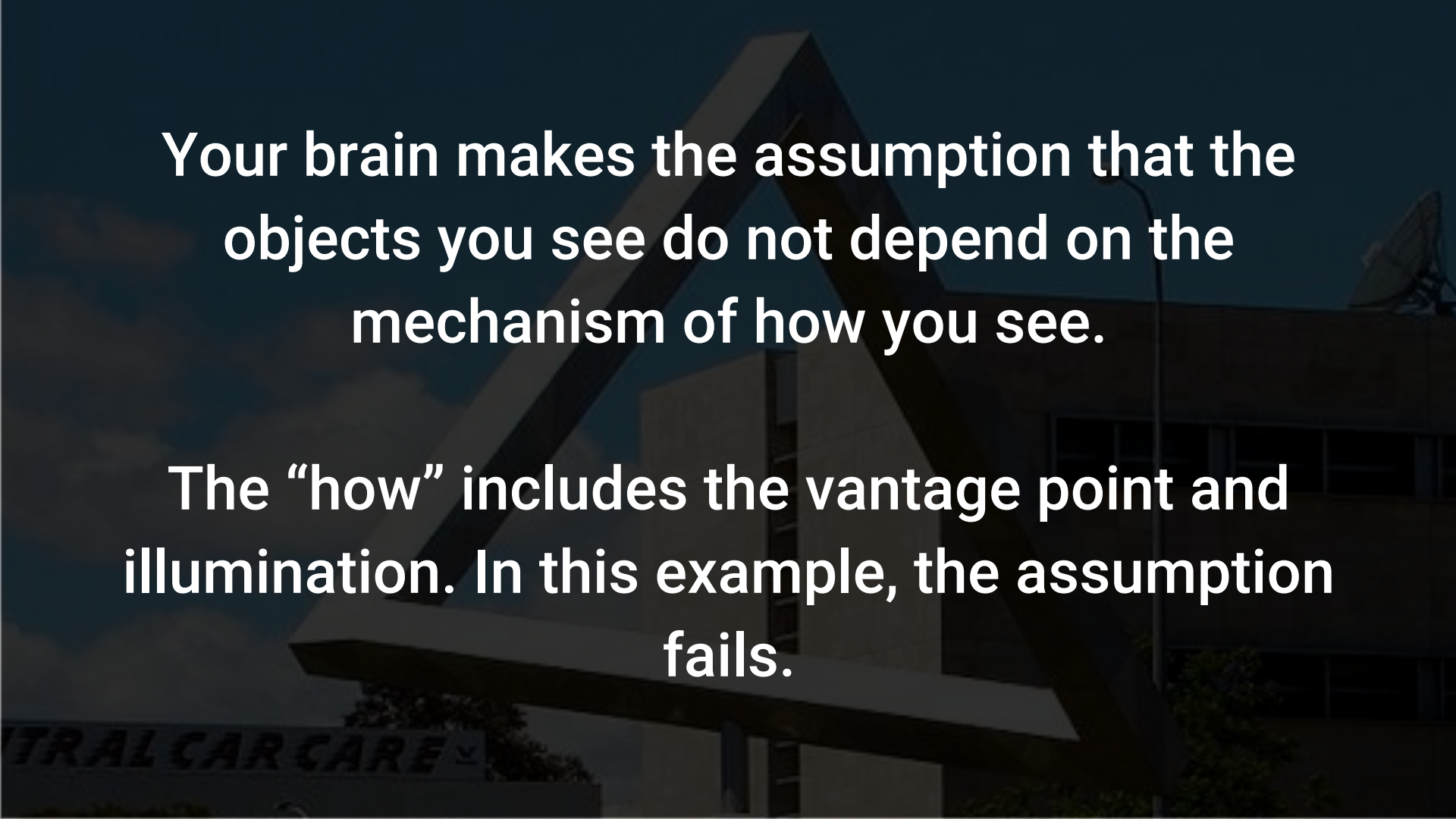
WHAT IF I TOLD YOU

THAT YOU KNEW THIS ALL ALONG?

A large, white, three-dimensional Penrose triangle sculpture stands in the foreground. The sculpture is a classic impossible object, appearing as a solid triangular frame. Behind it is a modern, multi-story building with a light-colored, textured facade and several windows. A tall, thin street lamp is visible to the right of the sculpture. The sky is blue with scattered white clouds. In the bottom left corner, a sign for 'CENTRAL CAR CARE' is partially visible.

CENTRAL CAR CARE





Your brain makes the assumption that the objects you see do not depend on the mechanism of how you see.

The “how” includes the vantage point and illumination. In this example, the assumption fails.

Special Case: The Additive Noise Model

Take the structured causal model, and require that the mechanism is a deterministic function of the cause plus an additive noise, ergo:

$$\begin{aligned} \mathbf{C} &\sim P(\mathbf{C}) \\ \mathbf{E} &= \mathbf{f}_{Z \sim P(Z)}(\mathbf{C}) = \mathbf{f}(\mathbf{C}) + \mathbf{Z} \end{aligned}$$

A causal relationship following this model would leave a statistical signature on the joint probability distribution $P(\mathbf{C}, \mathbf{E})$. In other words, it's possible to tell cause from effect just by looking at observations, or samples from $P(\mathbf{C}, \mathbf{E})$, i.e. without actually having to perform interventions. In causality jargon, *the problem of causal discovery is identifiable*.

Example of a Causal Signature on the Joint Distribution

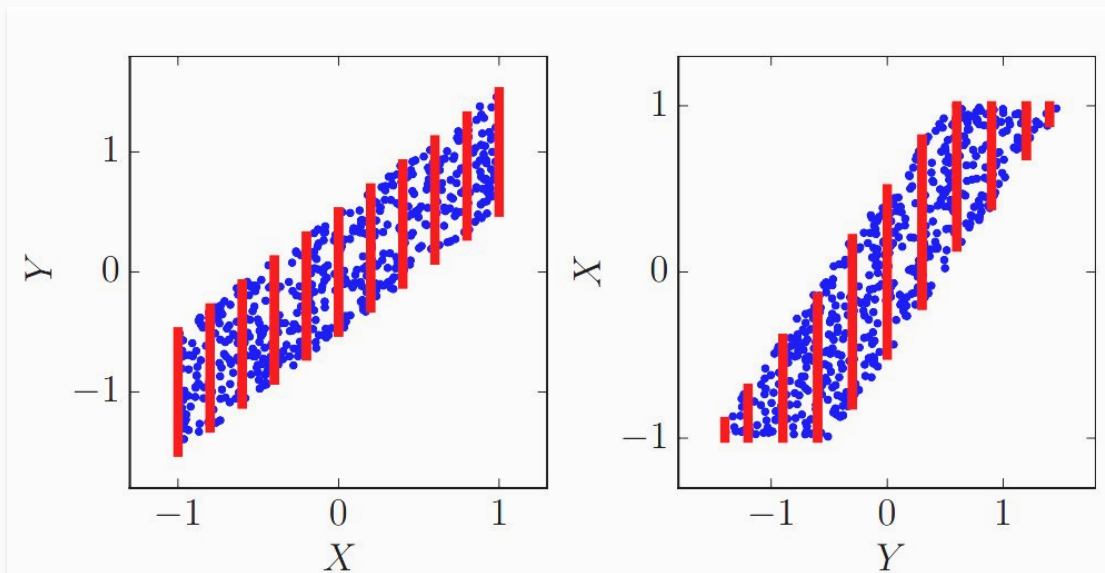
Say X is the cause and Y is the effect, and we have the SCM:

$$Y = X + Z$$

where $P(X) \perp P(Z)$. It's impossible to construct

$$X = f(Y) + Z'$$

with $P(Y) \perp P(Z')$.



(a) ANM $X \rightarrow Y$. (b) ANM $Y \rightarrow X$

If there's a statistical
signature of causal influence,
can we learn to find it?

Given a model powerful enough, apparently.

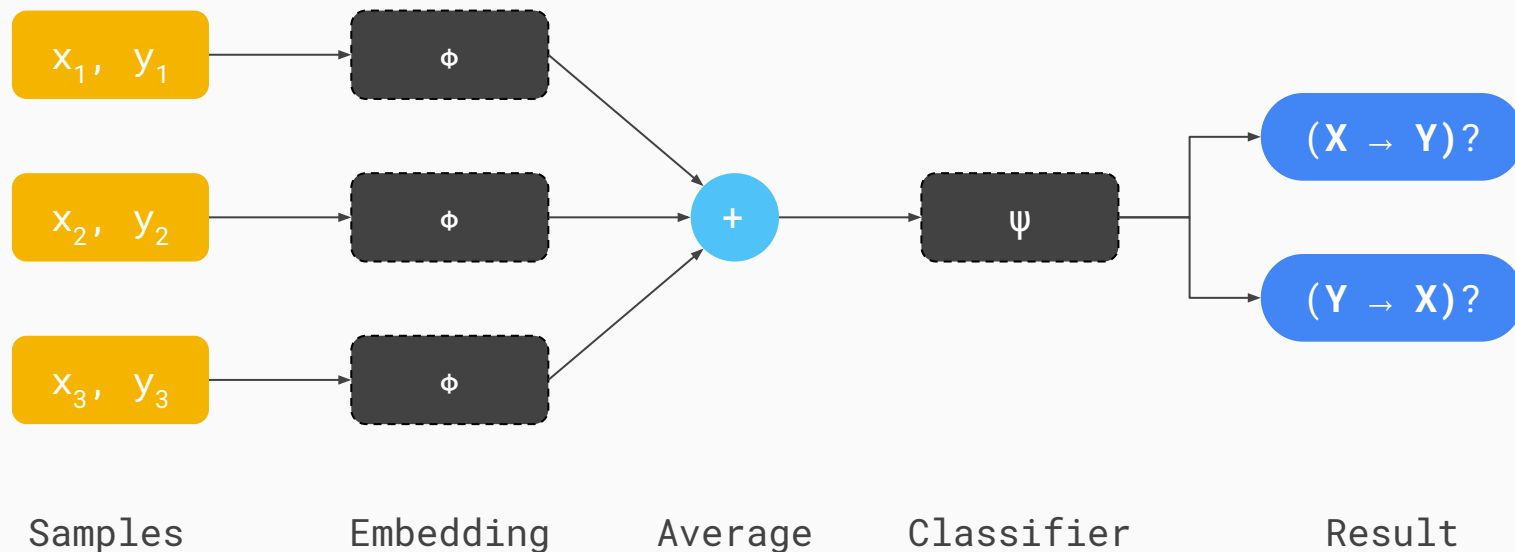
Enter Neural Networks.



The joint *distribution* $P(X, Y)$ is fed to a network which is tasked with predicting the causal direction of the variables.

How do we feed a distribution to a Neural Network?

Feed it Samples!



This architecture has been reinvented over and over again.

A simple neural network module for relational reasoning

Adam Santoro*, David Raposo*, David G.T. Barrett, Mateusz Malinowski,
Razvan Pascanu, Peter Battaglia, Timothy Lillicrap

adamsantoro@, draposo@, barrettdavid@, mateuszm@,
razp@, peterbattaglia@, countzero@google.com

DeepMind
London, United Kingdom

Deep Sets

**Manzil Zaheer^{1,2}, Satwik Kottur¹, Siamak Ravanbakhsh¹,
Barnabás Póczos¹, Ruslan Salakhutdinov¹, Alexander J Smola^{1,2}**

¹ Carnegie Mellon University ² Amazon Web Services
{manzilz,skottur,mravanba,bapoczos,rsalakhu,smola}@cs.cmu.edu

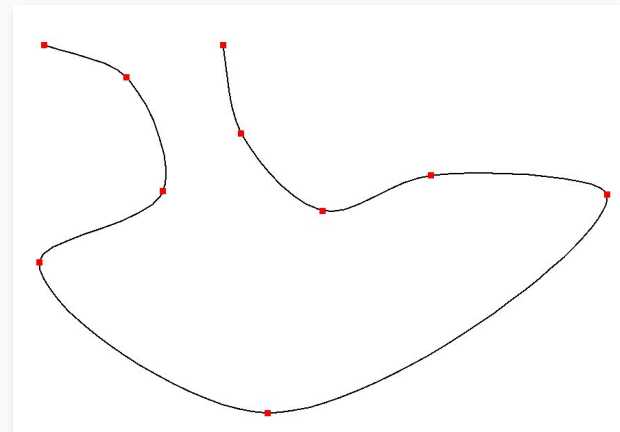
Where do we get the training data?

Remember our old friend, the additive noise model:

$$\begin{aligned} \mathbf{C} &\sim P(\mathbf{C}) \\ \mathbf{E} &= \mathbf{f}(\mathbf{C}) + \mathbf{Z} \end{aligned}$$

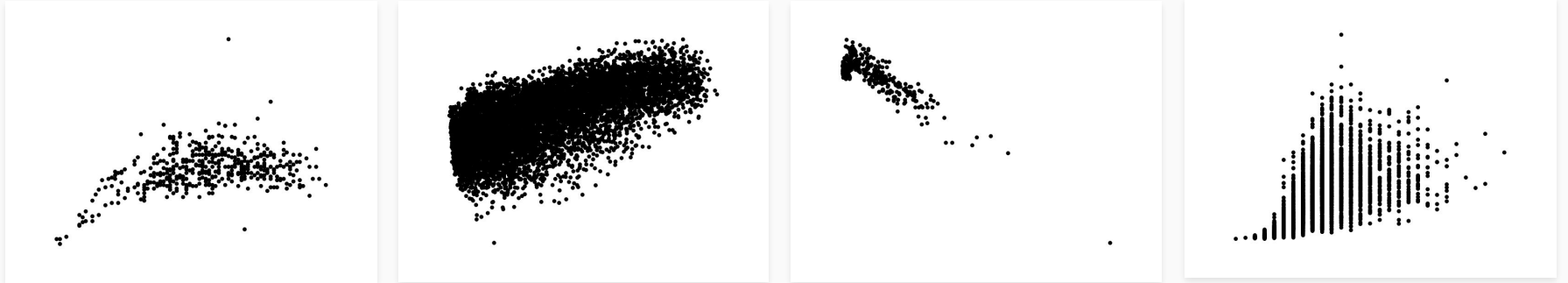
Now,

- * set $P(\mathbf{C})$ to a mixture of gaussians,
- * set $P(\mathbf{Z})$ to a random gaussian,
- * set \mathbf{f} to a cubic hermite spline and sample its parameters.
- * compute \mathbf{E} to obtain a sample from $P(\mathbf{C}, \mathbf{E})$.
- * assign label $\mathbf{1}$ to (\mathbf{C}, \mathbf{E}) and $\mathbf{0}$ to (\mathbf{E}, \mathbf{C}) .



A cubic hermite spline

Where do we get the validation data? The Tübingen Datasets



Samples from the Tübingen datasets: 107 $P(X, Y)$ real (non-synthetic) distributions with corresponding labels for causal directions.

Experiments & Results

Generalizing to the Tübingen Datasets

* Train the network (the “*Neural Causal Coefficient*”) on synthetic data generated by the Additive noise model.

* Validate on the Tübingen datasets.

* State of the Art on the Tübingen Datasets (with 79% accuracy)

* Previous state of the art was at 75% accuracy.

Detecting Causal Signals in Images

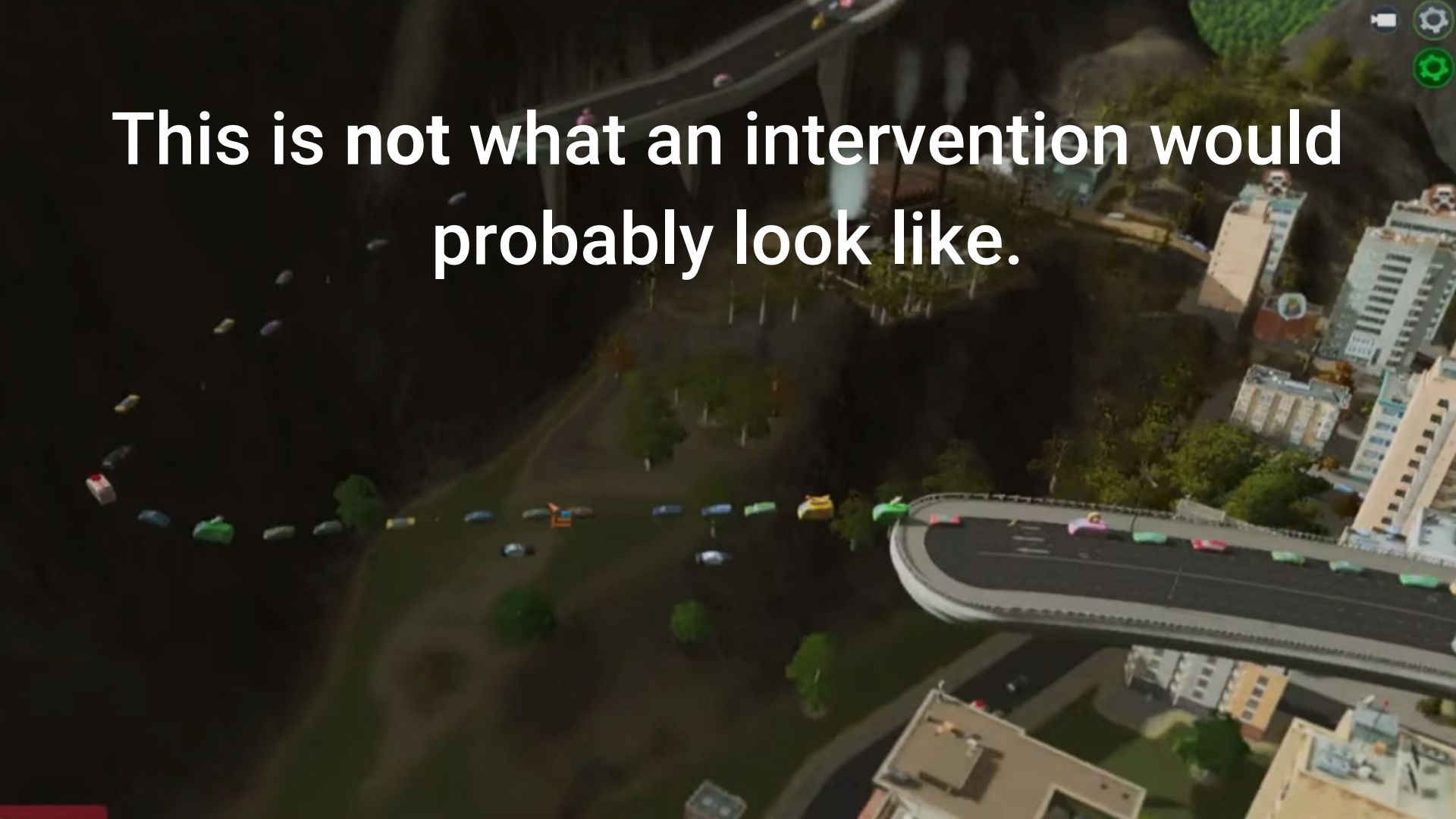


“Does the presence of the car cause the presence of the wheels?”

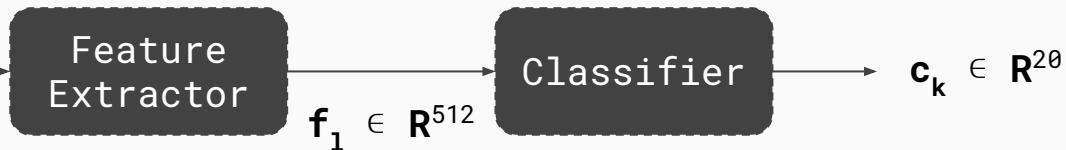
A photograph of a suspension bridge with several cars and bicycles on it. The bridge has tall towers and many vertical cables. The scene is dimly lit, possibly at dusk or dawn. The text is overlaid in the center of the image.

Does the presence of a bridge cause
the presence of cars on it?

This is not what an intervention would probably look like.



General Task



Given features $\mathbf{f}_1 \in \mathbf{R}^{512}$ from a feature extractor (e.g. the convolutional layers of an off-the-shelf network), use the NCC to predict the direction and strength of the causal relation between a given feature and the (pre-softmax) output from the classifier $\mathbf{c}_k \in \mathbf{R}^{20}$ corresponding to a given class.

Definitions

- * **Causal features:** features that cause the presence of an object in the scene.
- * **Anticausal features:** features that are caused by the presence of an object in the scene.
- * **Object features:** features that are most activated inside the bounding box around the object of interest.
- * **Context features:** features that are most activated outside the bounding box.

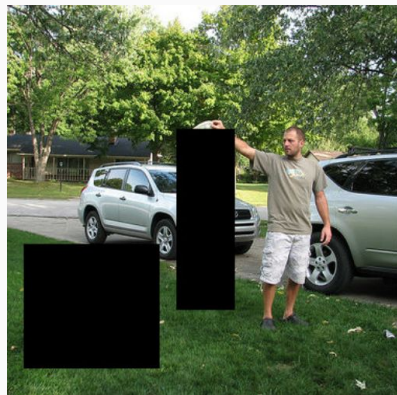
“There exists an observable statistical dependence between object features and anticausal features.”

“The statistical dependence between context features and causal features is non-existent or much weaker.”

Features that are most activated in the bounding box around the object of interest are those that are often caused by the presence of the object in the scene.

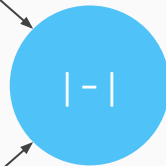
Features that are most activated outside the bounding box do not necessarily cause the presence of the object in the scene.

Proxy for Object and Context Features



Feature
Extractor

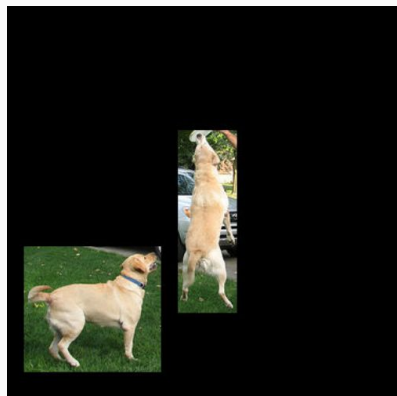
Feature
Extractor



Object Feature Ratio

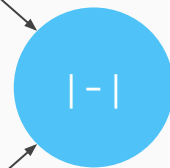
Relative difference

Proxy for Object and Context Features



Feature
Extractor

Feature
Extractor



Context Feature Ratio

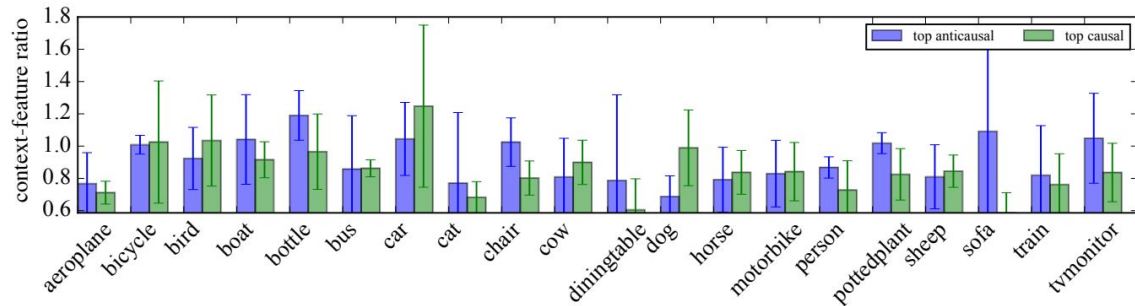
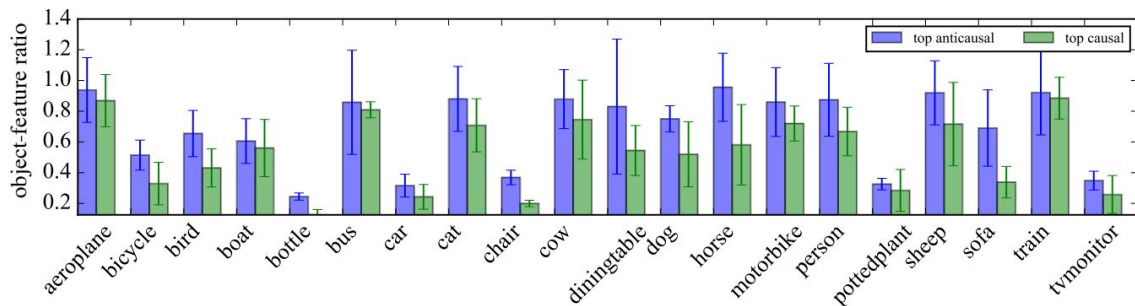
Relative difference

Results

The figure shows the object and context feature ratios of the top 1% of causal and anticausal features as predicted by the NCC model.

“The average object feature scores associated to the top 1% anticausal feature scores is always higher than the average object feature score associated to the top 1% causal features.”

“Such separation does not occur for context feature scores.”



The image features a central black circle surrounded by several concentric red rings, creating a tunnel-like effect. A white, stylized cursive letter 'J' is positioned on the left side, partially overlapping the red rings and the black center. The 'J' has a long, sweeping tail that curves downwards and to the left.

J