Towards Compositional Interpretability for XAI

Sean Tull

QUANTINUUM

Naturalistic Approaches to Artificial Intelligence
IPAM, UCLA, 6 Nov 2024

Motivation

Most AI models lack **interpretability**, a major concern in high-stakes areas e.g. health sector.

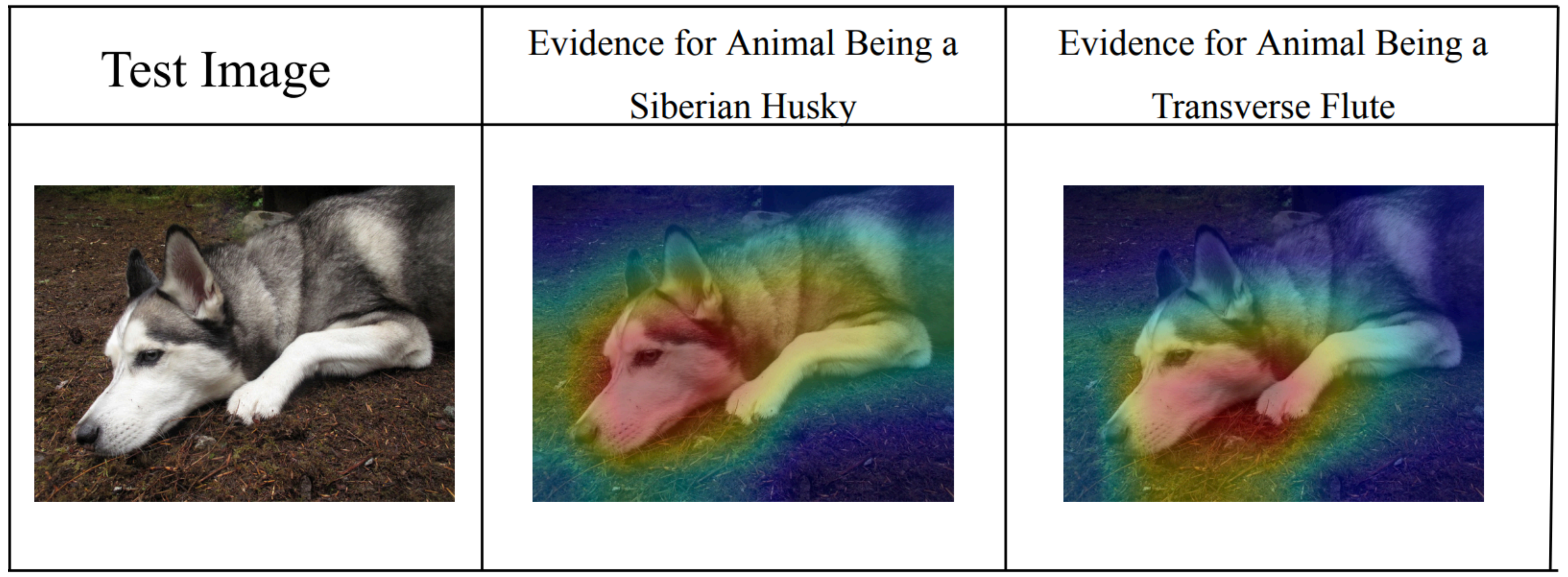
How does the model work?

Is it biased?

Why was the output X and not Y?

eXplainable (X)AI hopes to solve this, often via **post-hoc** explanations for outputs, but more formal work is needed.

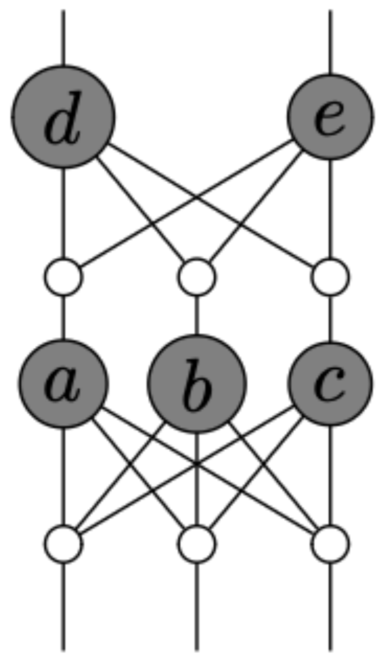
- These often only provide **limited** explanations (Rudin 2019).
- No standard **definition of interpretability.**



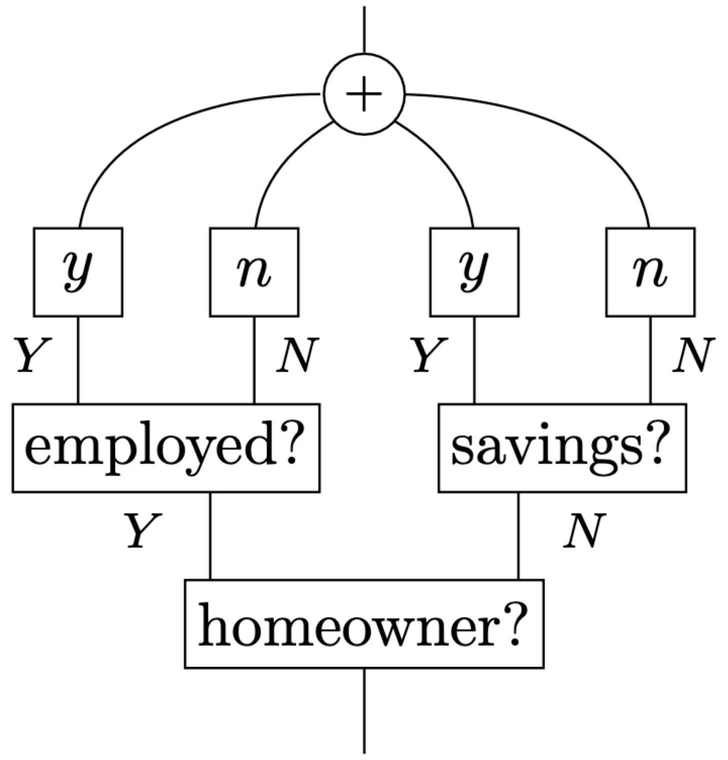
Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,* 2019.

Intuition: A model is **interpretable** when it has **meaningful compositional structure**.

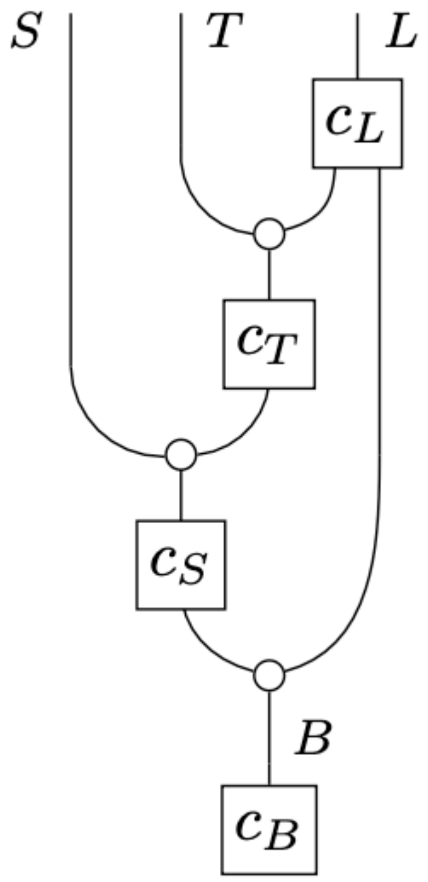
The mathematics of structure and **composition** is that of **category theory** and **string diagrams**.



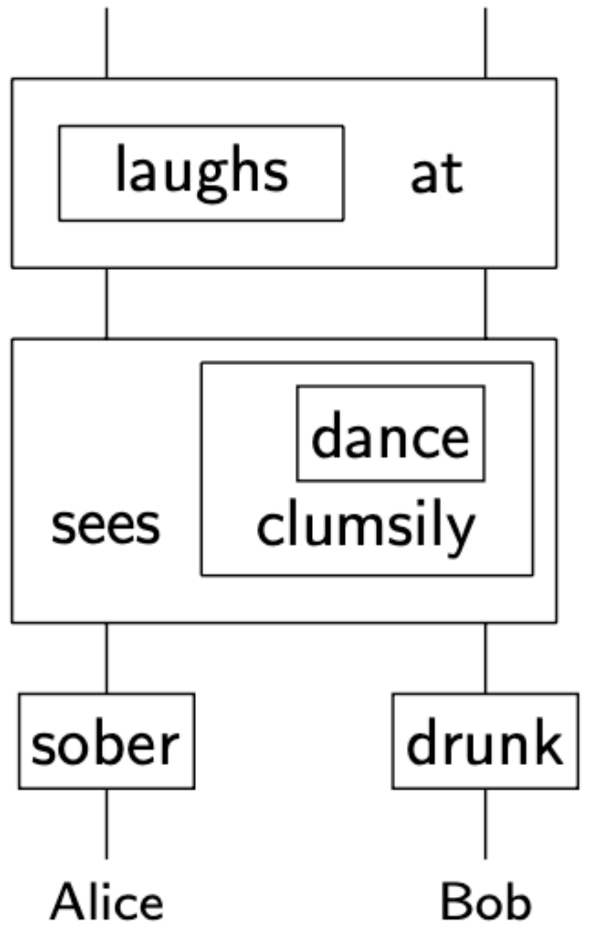
Neural network



Decision tree



Causal model



DisCoCirc model

Towards Compositional Interpretability for XAI

Sean Tull, Robin Lorenz, Stephen Clark, Ilyas Khan, Bob Coecke
{sean.tull,robin.lorenz,steve.clark,ilyas,bob.coecke}@quantinuum.com
Quantinuum, 17 Beaumont Street, Oxford, UK

- Categorical formalism for **defining AI models and interpretability**
- Make precise how **compositional structure can give explainable models**

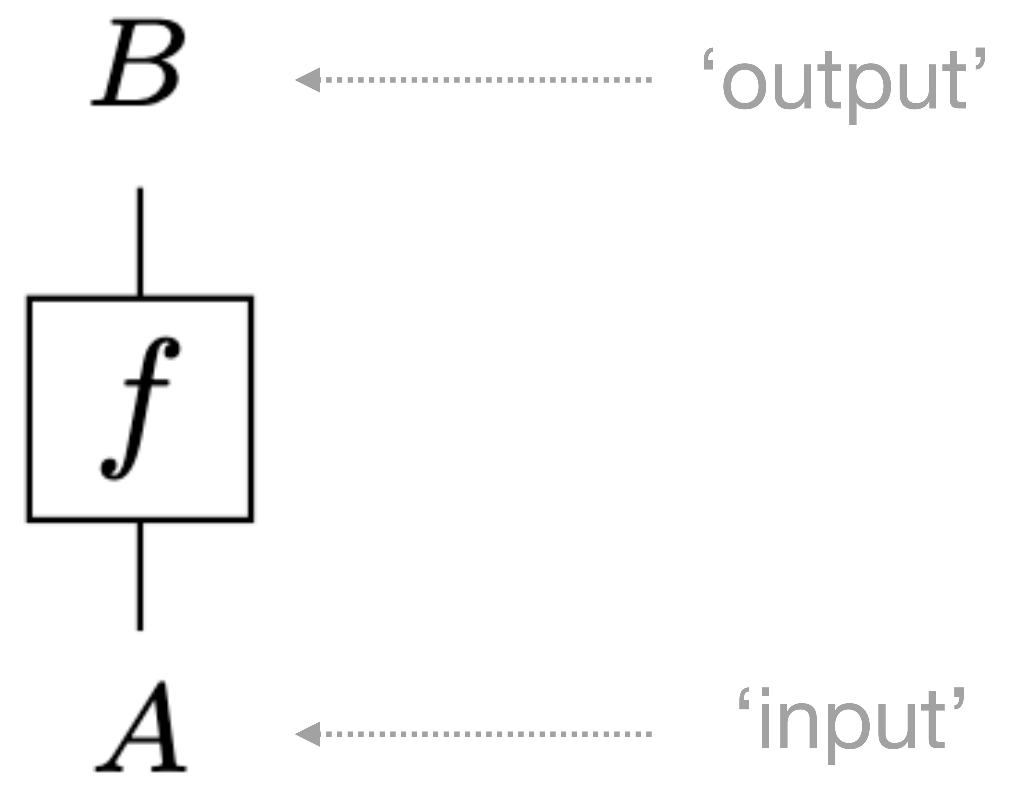
Applies to deterministic, probabilistic, and even **quantum** models

Category Theory

and String Diagrams

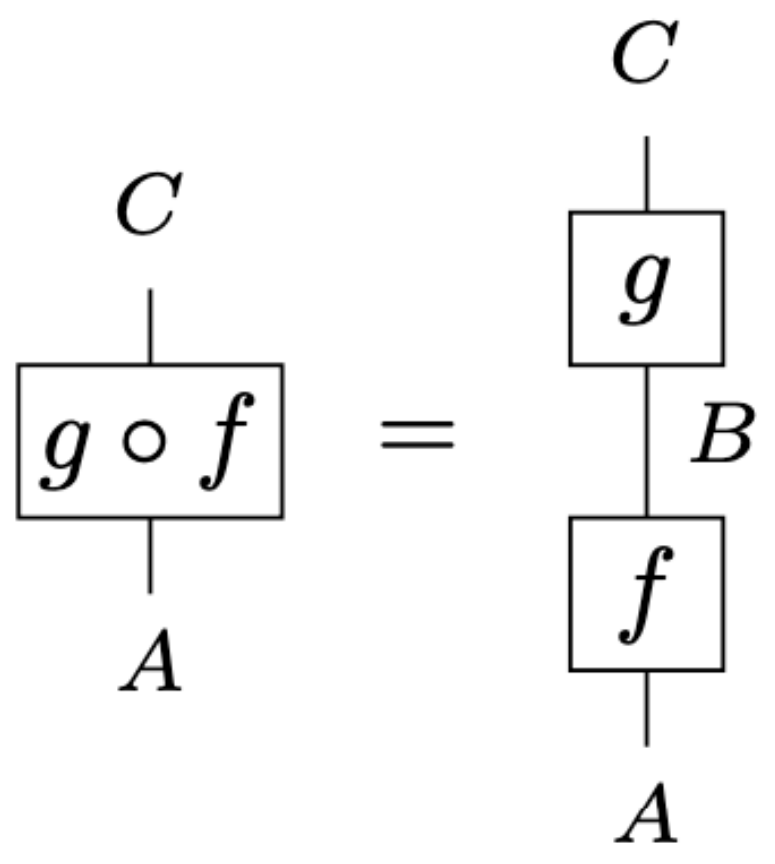
Categories

A **symmetric monoidal category** $\mathbf{C}$ consists of a collection of **objects** $A, B, C \ldots$ and **morphisms** or **processes** $f\colon A \to B$ between them, depicted in **string diagrams**:

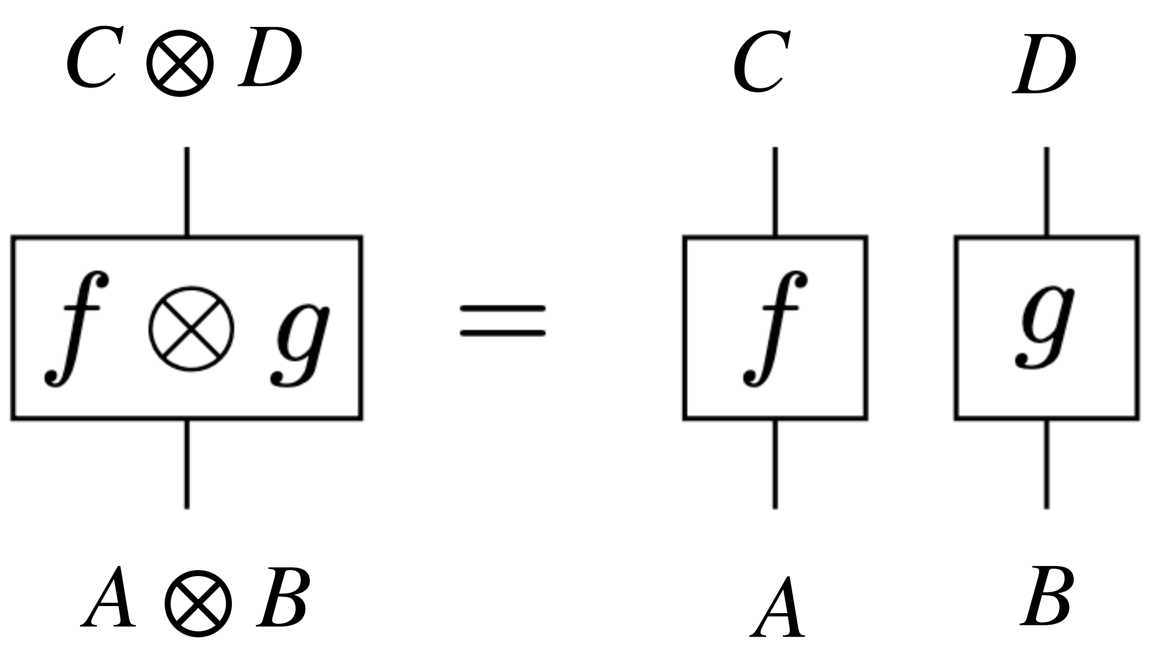


Categories

We can **compose** processes ‘in sequence’:



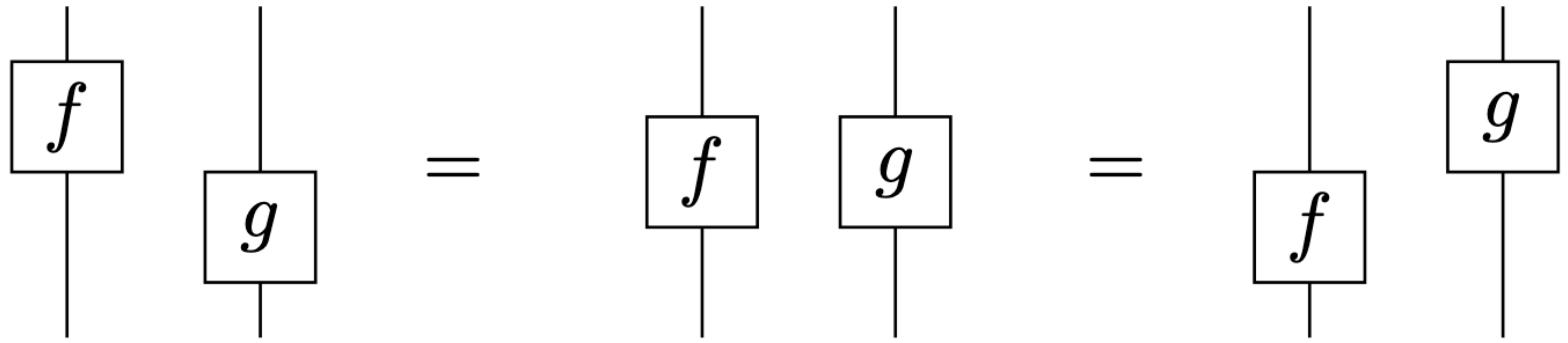
and ‘in parallel’:



‘tensor’ of A, B

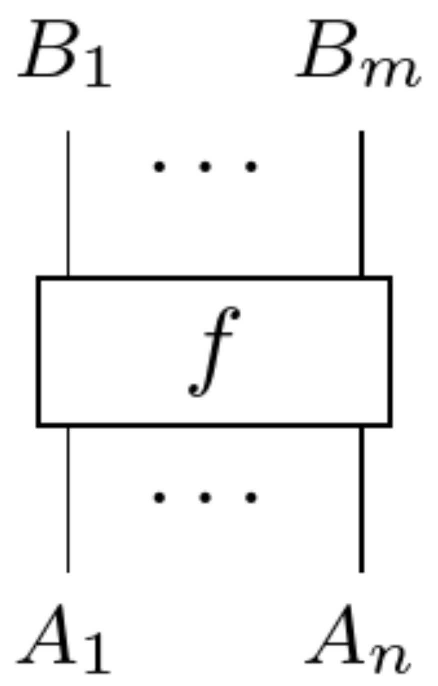
Categories

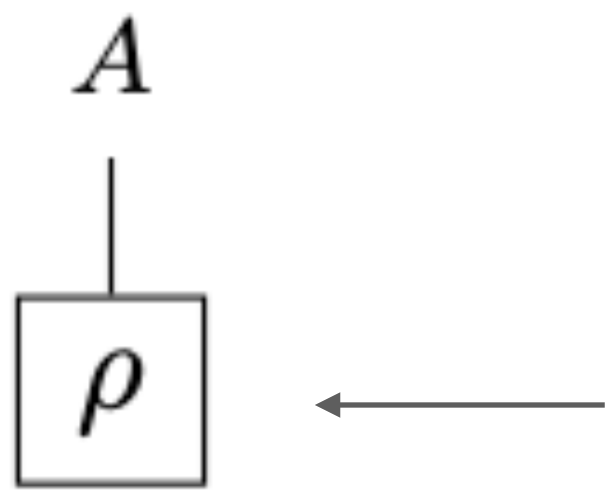
Categories satisfies various equations that come ‘for free’ in the diagrams:



Categories

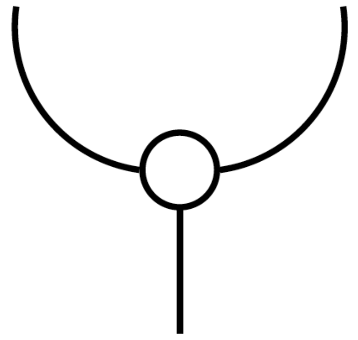
Processes can have multiple (or zero) inputs or outputs:



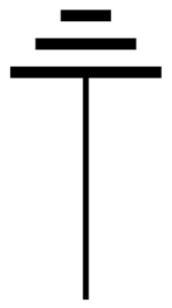


state of A

Many categories also come with processes for **copying** and **discarding**:



copy



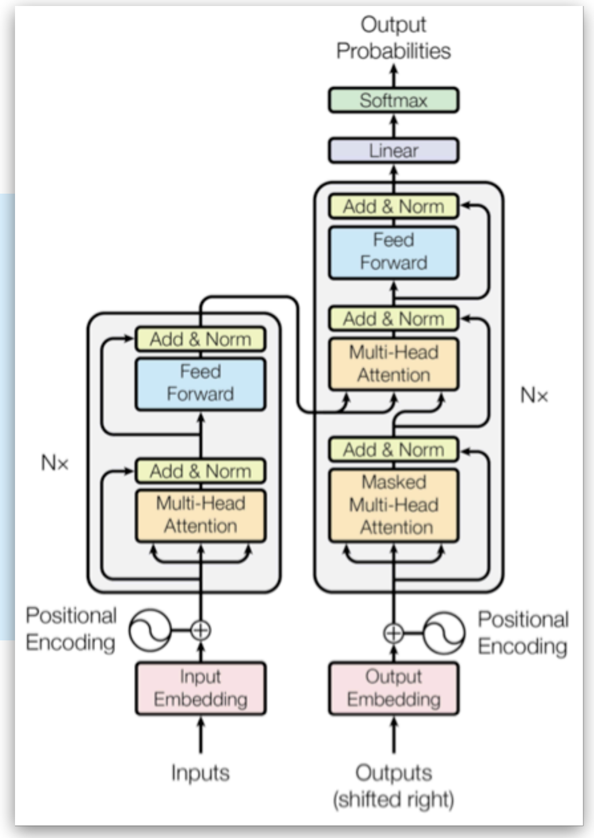
discard

Formally the input to a state is the *unit object,* drawn as ‘empty space’

Examples

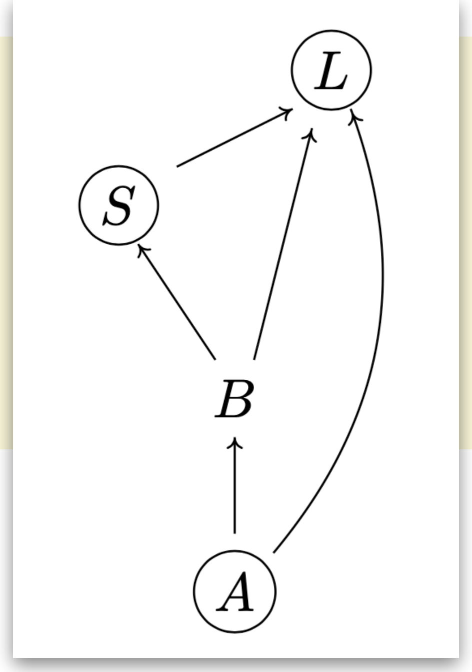
NN : Objects are spaces $\mathbb{R}^n$, morphisms are functions $f\colon \mathbb{R}^n \to \mathbb{R}^m$.

- Diagrams capture **neural networks.**



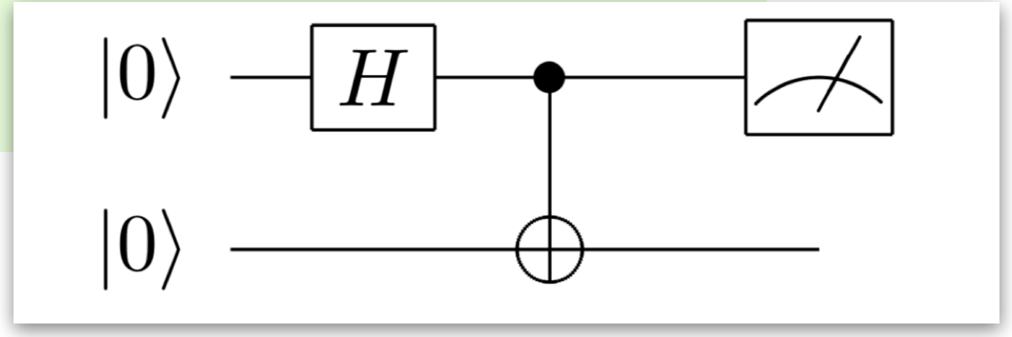
Stoch : Objects are finite sets X, morphisms are probability channels $P(Y \mid X)$.

- Diagrams capture **Bayesian networks.**



Quant : Objects are finite-dimensional Hilbert spaces $\mathscr{H}$, morphisms are CP maps $f\colon L(\mathscr{H}) \to L(\mathscr{K})$.

- Diagrams capture **Quantum Circuits.**



Compositional Models

Compositional Models

A **signature** G consists of sets of abstract ‘objects’ (**variables**) and ‘morphisms’ (**generators**) between them*.

A **compositional model** $\mathbb{M}$ is then given by:

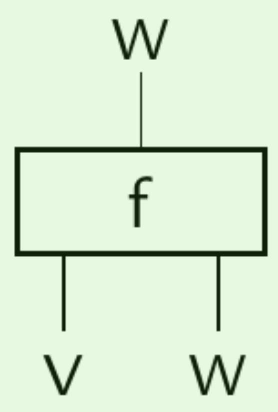
$$\mathbb{S} = \mathbf{Free}(G) \xrightarrow{[\![-]\!]} \mathbf{C}$$

structure category of diagrams built from G — **representation** functor — **semantics** category

Example

G

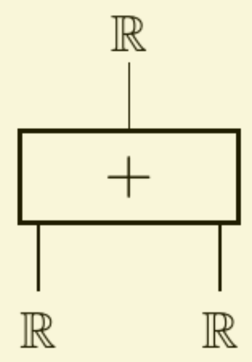
V, W,



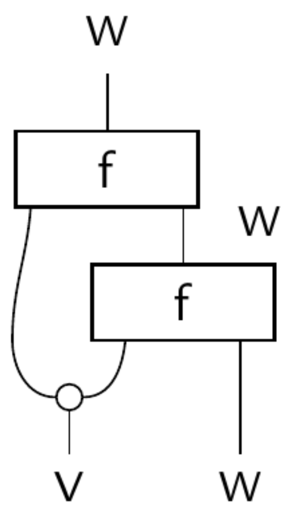
$\longmapsto$

$\mathbf{C} = \mathbf{NN}$

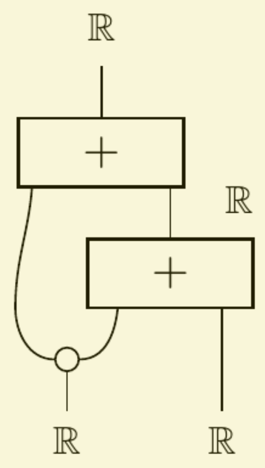
$\mathbb{R}$,



$:: (x, y) \mapsto x + y$



$\longmapsto$



$:: (x, y) \mapsto 2x + y$

*Along with optional equations between morphisms.

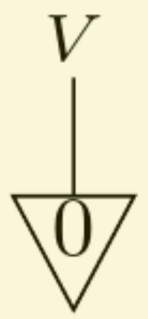
Interpretations

An **interpretation** consists of a signature $\mathbb{H}$ of '**human-friendly**' terms, along with two partial maps:

- an **abstract interpretation** $\mathcal{I}^{\mathsf{A}}$ of variables and generators in G.

  e.g. $\mathsf{V} \mapsto$ *'Brightness'*

- a **concrete interpretation** $\mathcal{I}^{\mathsf{C}}$ of morphisms in $\mathbf{C}$, such as states.

  e.g.  $\mapsto$ *'dark'*  $\mapsto$ *'bright'*

$$\begin{tikzcd} G \arrow[r, "[\![-]\!]"] \arrow[dr, "\mathcal{I}^{\mathsf{A}}"'] & \mathbf{C}_G \arrow[d, "\mathcal{I}^{\mathsf{C}}"] \\ & \mathbb{H} \end{tikzcd}$$

Say variable V has…

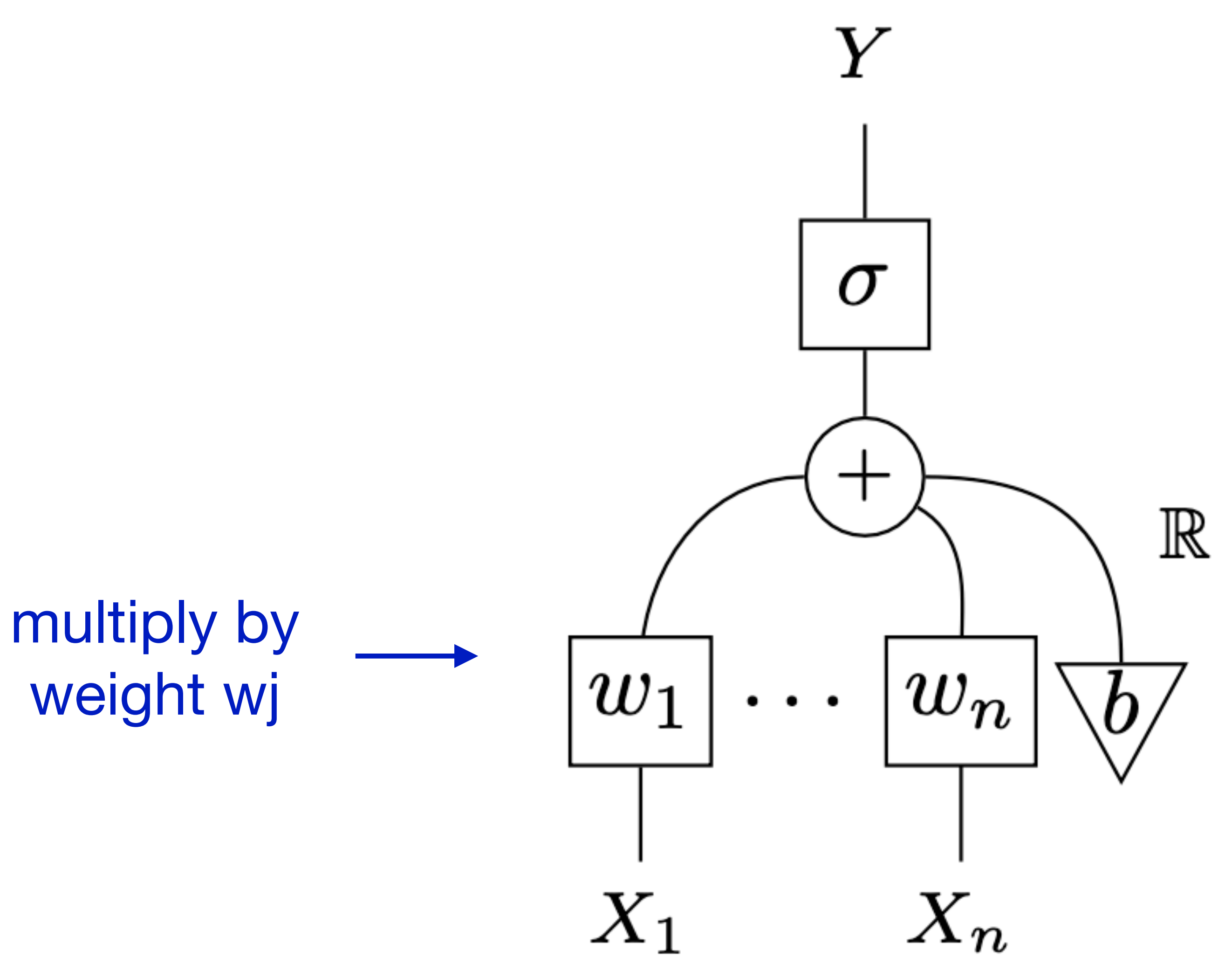
an abstract interpretation when $I^A(\mathsf{V})$ is defined.
a concrete interpretation when $I^C(\mathsf{v})$ is defined for every state v of $[\![V]\!]$in $\mathbf{C}$.

Formally $\mathcal{I}^{\mathsf{A}}$ $\mathcal{I}^{\mathsf{C}}$ are partial maps of signatures, and in $\mathbf{C}_G$ objects are lists of variables and morphisms $(\mathsf{A}_i)_{i=1}^n \to (\mathsf{B}_j)_{j=1}^m$ are $f : \otimes_{i=1}^n A_i \to \otimes_{j=1}^m B_j$ in $\mathbf{C}$.

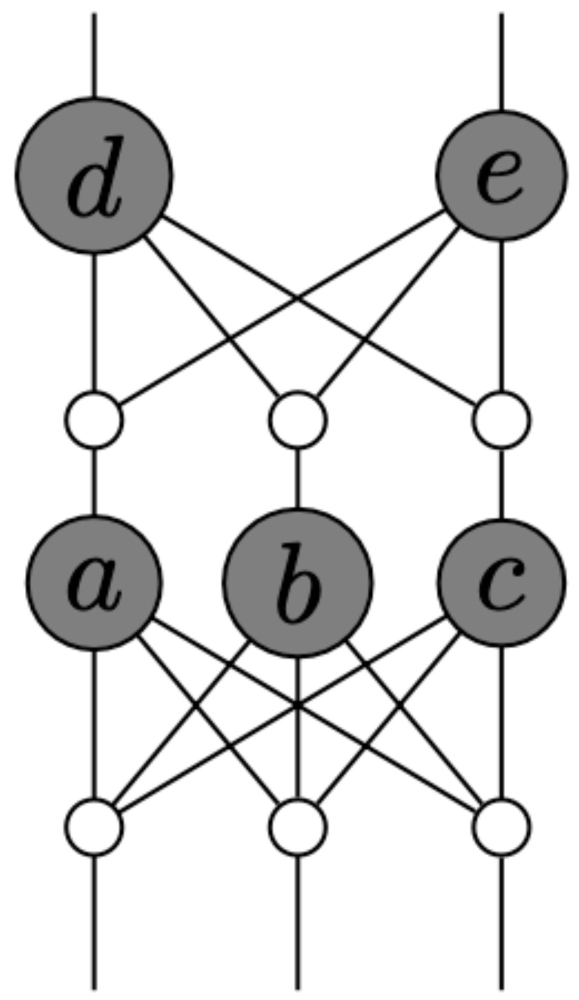
Analysing AI Models

Neural Networks

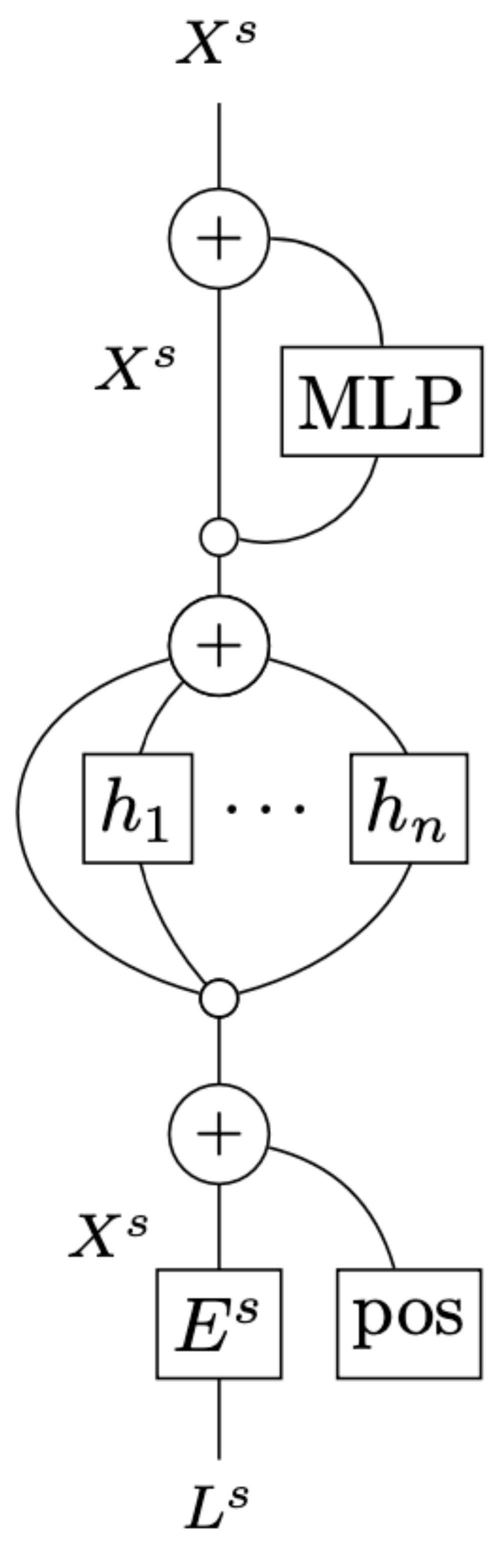
Neuron



Network



Transformer



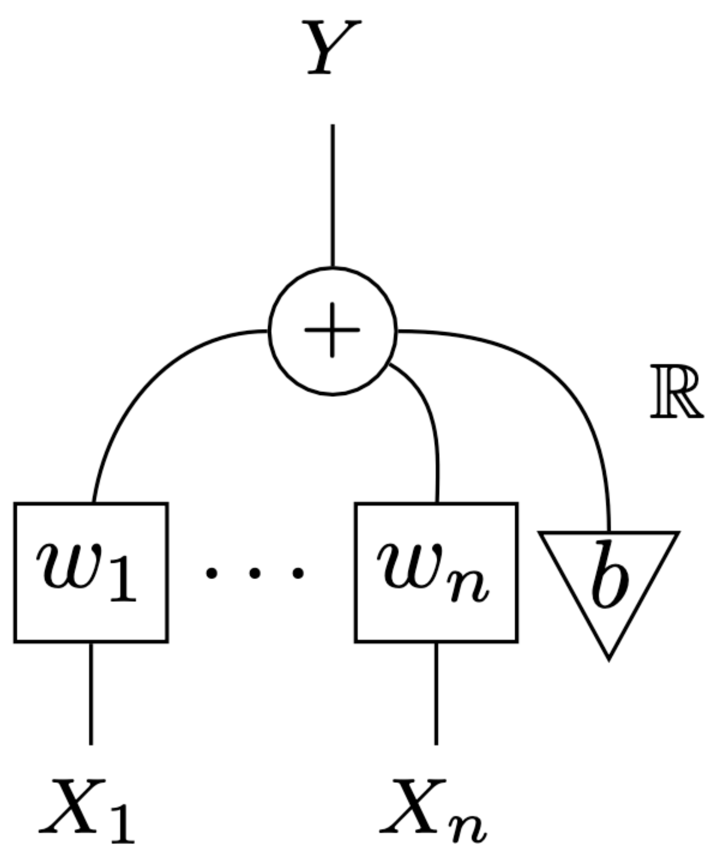
In the category $\mathbf{NN}$ of functions $f\colon \mathbb{R}^n \to \mathbb{R}^m$.

Observations

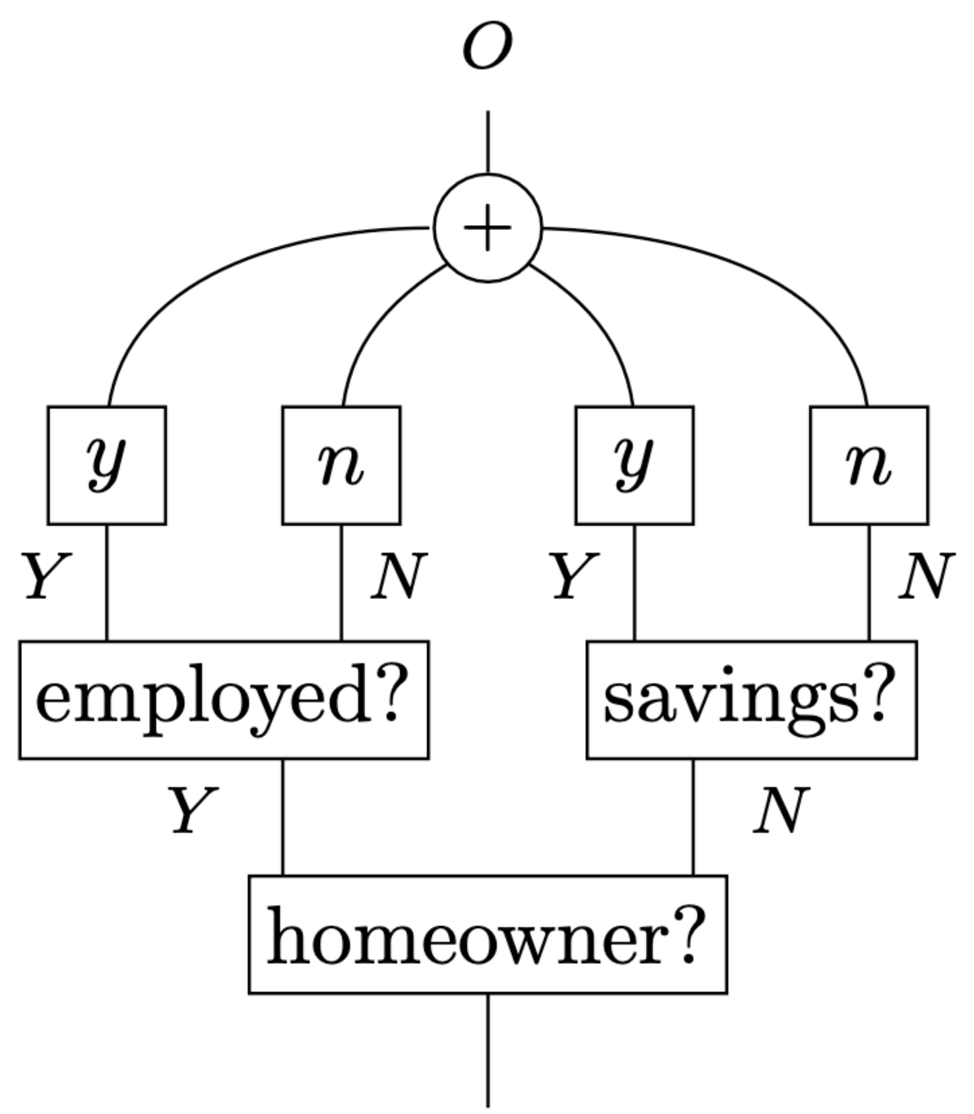
- Some forms of composition are common in ML.
- Compositional structure $\not\Rightarrow$ interpretability.
- Only inputs and outputs typically interpretable, so this is where XAI focuses.

Intrinsically Interpretable Models

Linear



Decision tree



Observation

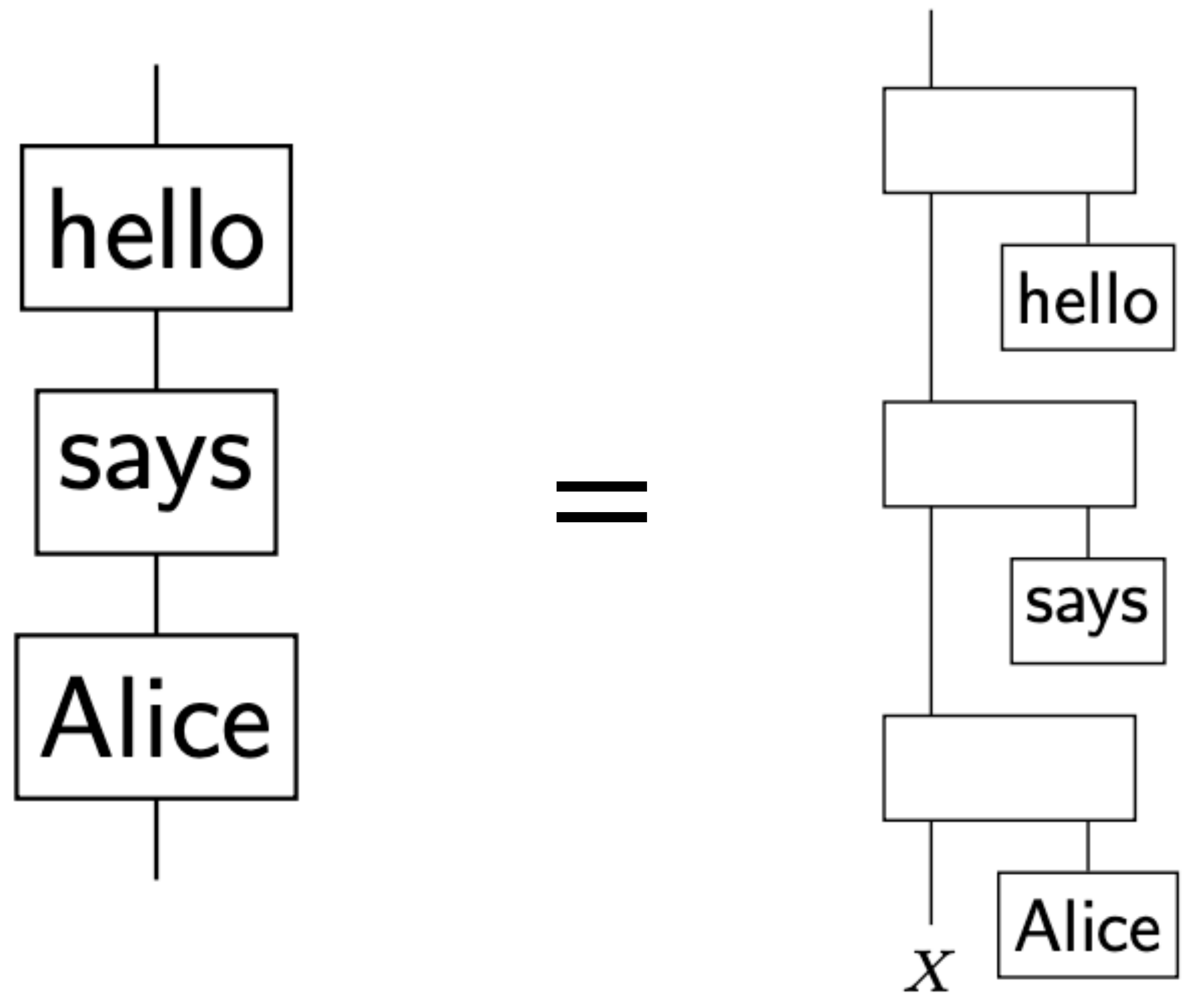
Intrinsic interpretability of models is manifest diagrammatically, and fits our definition.

Compositionally Interpretable Models

We call a model $\mathbb{M}$ **compositionally interpretable (CI)** when it has a complete abstract interpretation.

Every intrinsically interpretable model is CI, but the following models provide further examples.

Recurrent Neural Networks (RNNs)

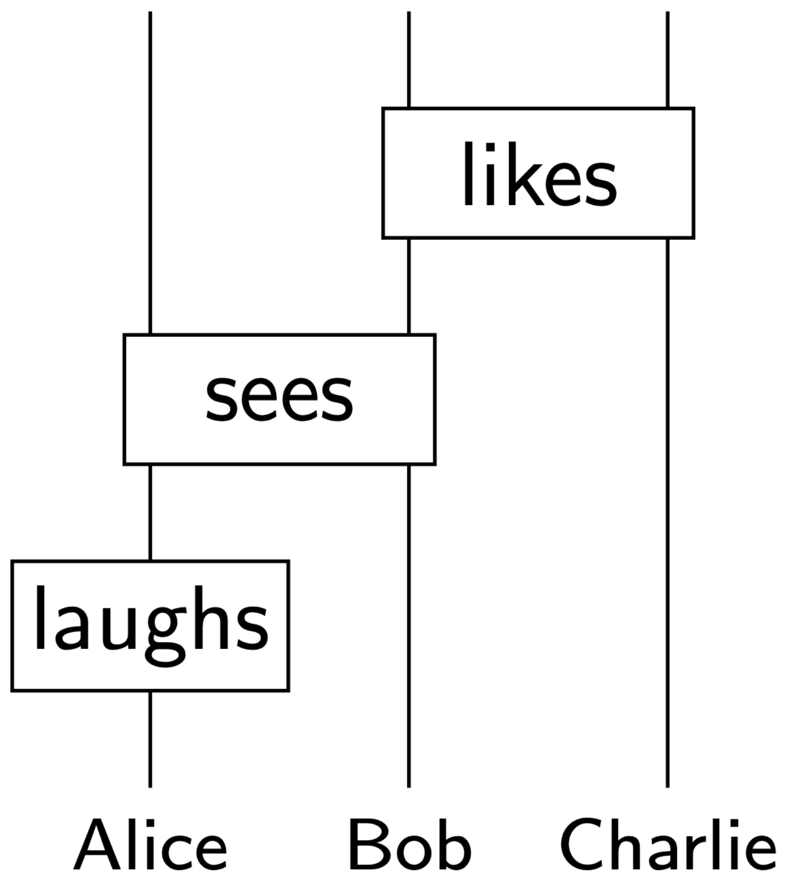


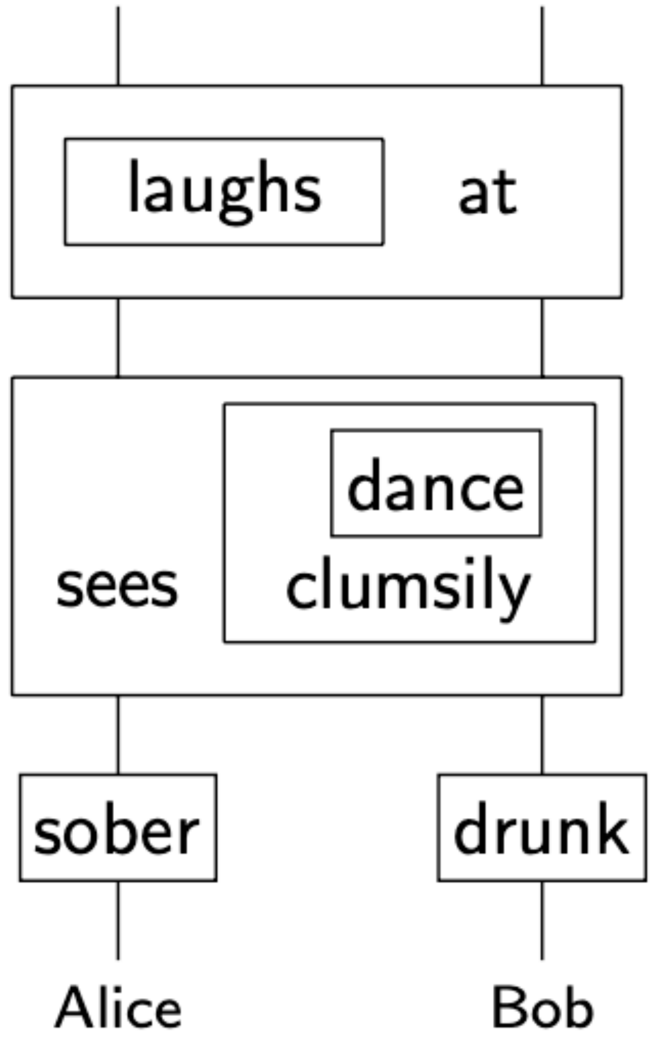
Any RNN forms a CI model in $\mathbf{C} = \mathbf{NN}$ with one variable and a generator for each word, represented by a NN.

DisCoCirc Models

In **DisCoCirc** models, a text is represented as a **text circuit** acting on its relevant nouns, where each word forms either a process (e.g. verbs) or **higher-order** process (e.g. adverbs).

‘Alice laughs and sees Bob, who likes Charlie’

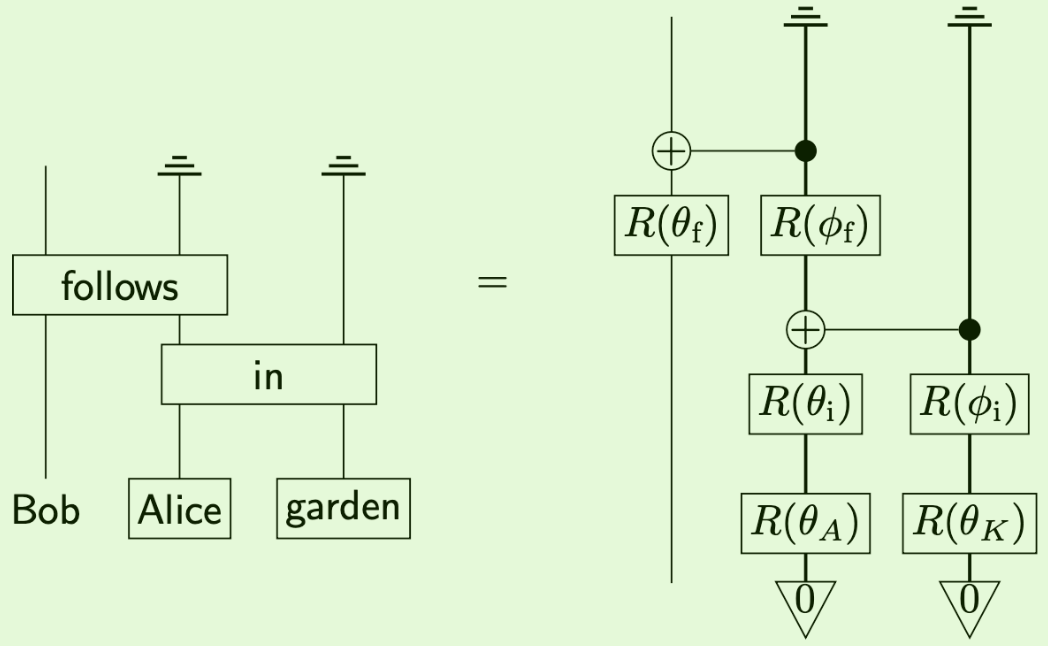




‘Alice, who is sober, sees Bob, who is drunk, clumsily dance and laughs at him’.

Implementable either as:

- neural networks, $\mathbf{C} = \mathbf{NN}$
- quantum circuits, $\mathbf{C} = \mathbf{Quant}$.



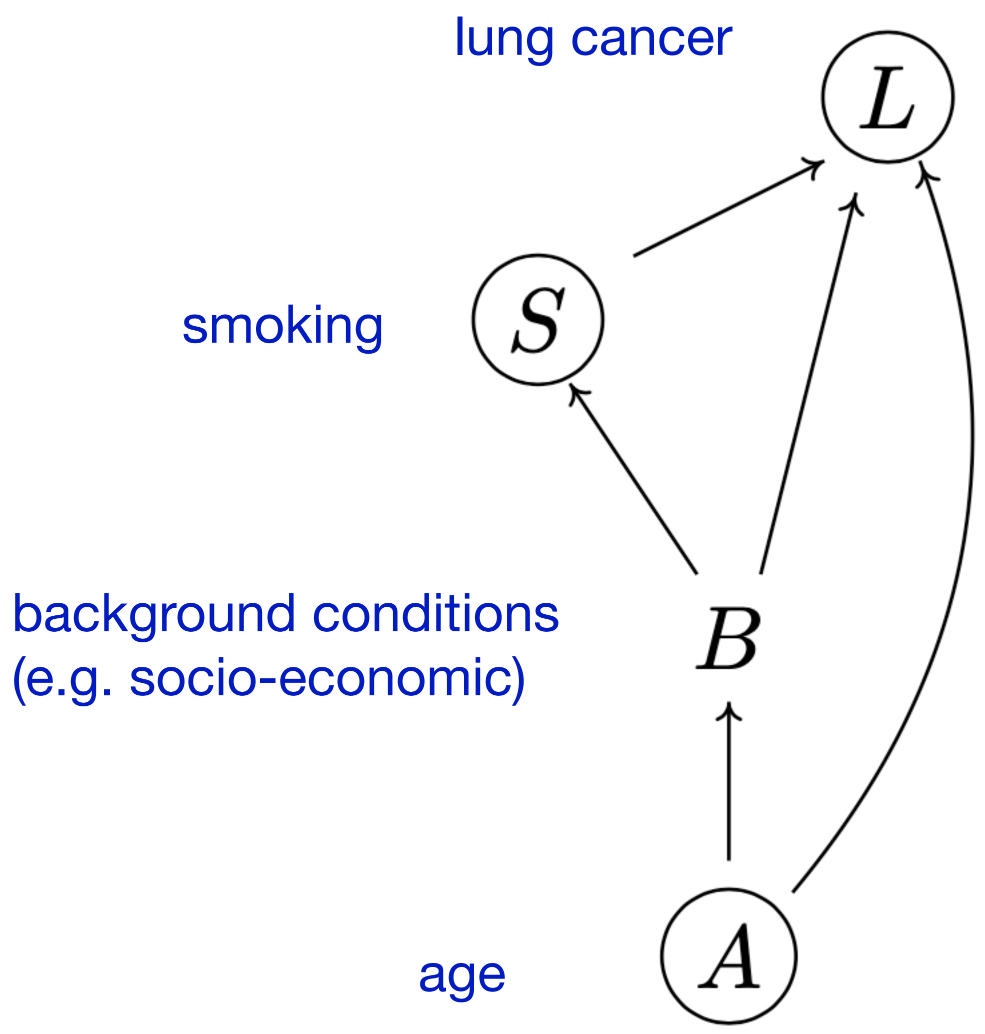
Scalable and interpretable quantum natural language processing: an implementation on trapped ions

Tiffany Duneau[1,2], Saskia Bruhn[1,*], Gabriel Matos[1,*], Tuomas Laakkonen[1], Katerina Saiti[3], Anna Pearson[1,*], Konstantinos Meichanetzidis[1], Bob Coecke[1]

[1]Quantinuum, [2]University of Oxford, [3]Leiden University

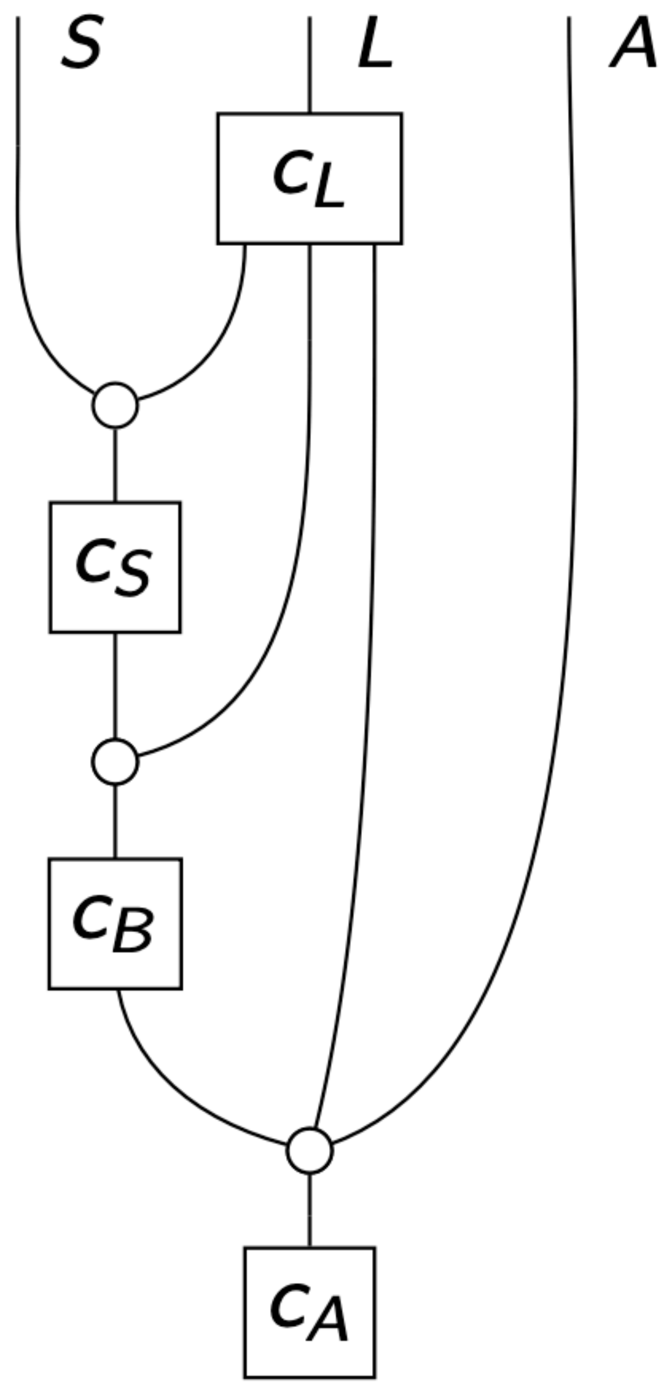
Causal Models

Causal models (causal Bayesian networks) form a well-known class of CI models, widely studied in **Causal ML.**



$$P(L|SBA)$$
$$P(S|B)$$
$$P(B|A)$$
$$P(A)$$

$$\Longleftrightarrow$$



Causal Inference by String Diagram Surgery

Bart Jacobs[1], Aleks Kissinger[1], and Fabio Zanasi[2]

Causal models in string diagrams

Robin Lorenz*, Sean Tull†

Quantinuum, 17 Beaumont Street, Oxford, UK

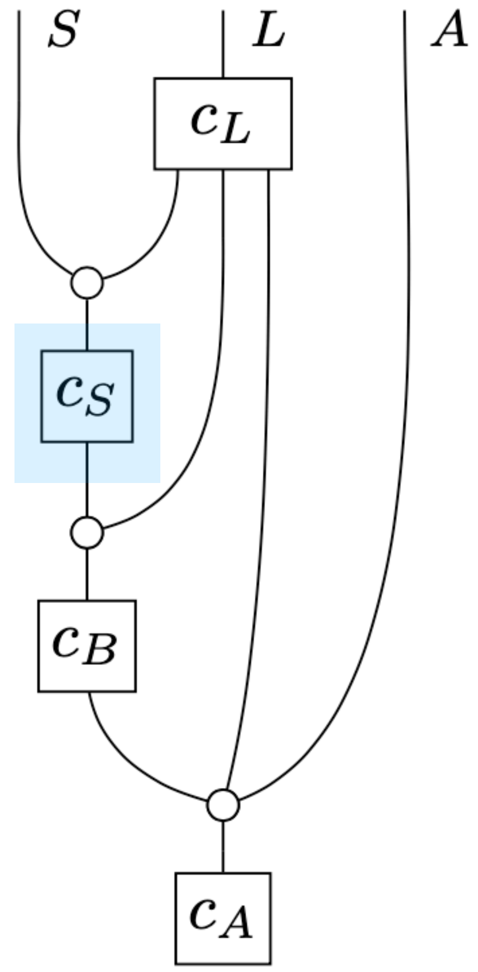
Usual description

Network Diagram in $\mathbf{Stoch}$.

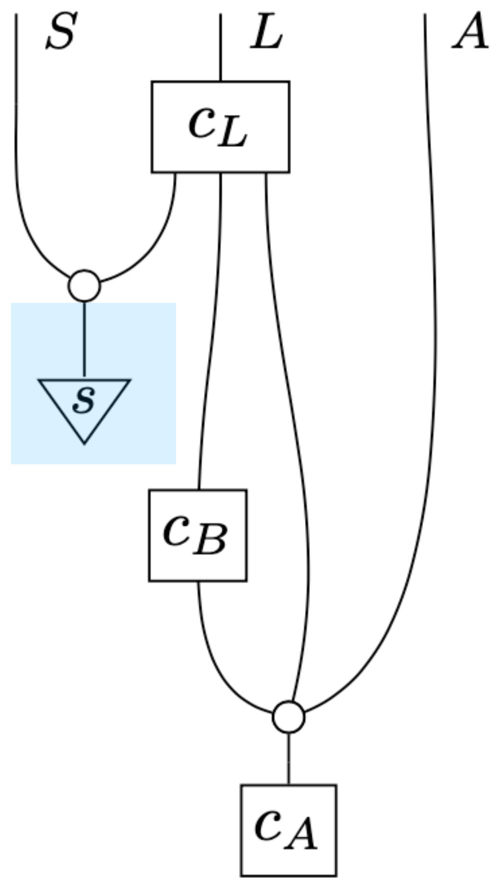
Causal Models

The causal model **framework** provides further interpretability benefits:

Interventions



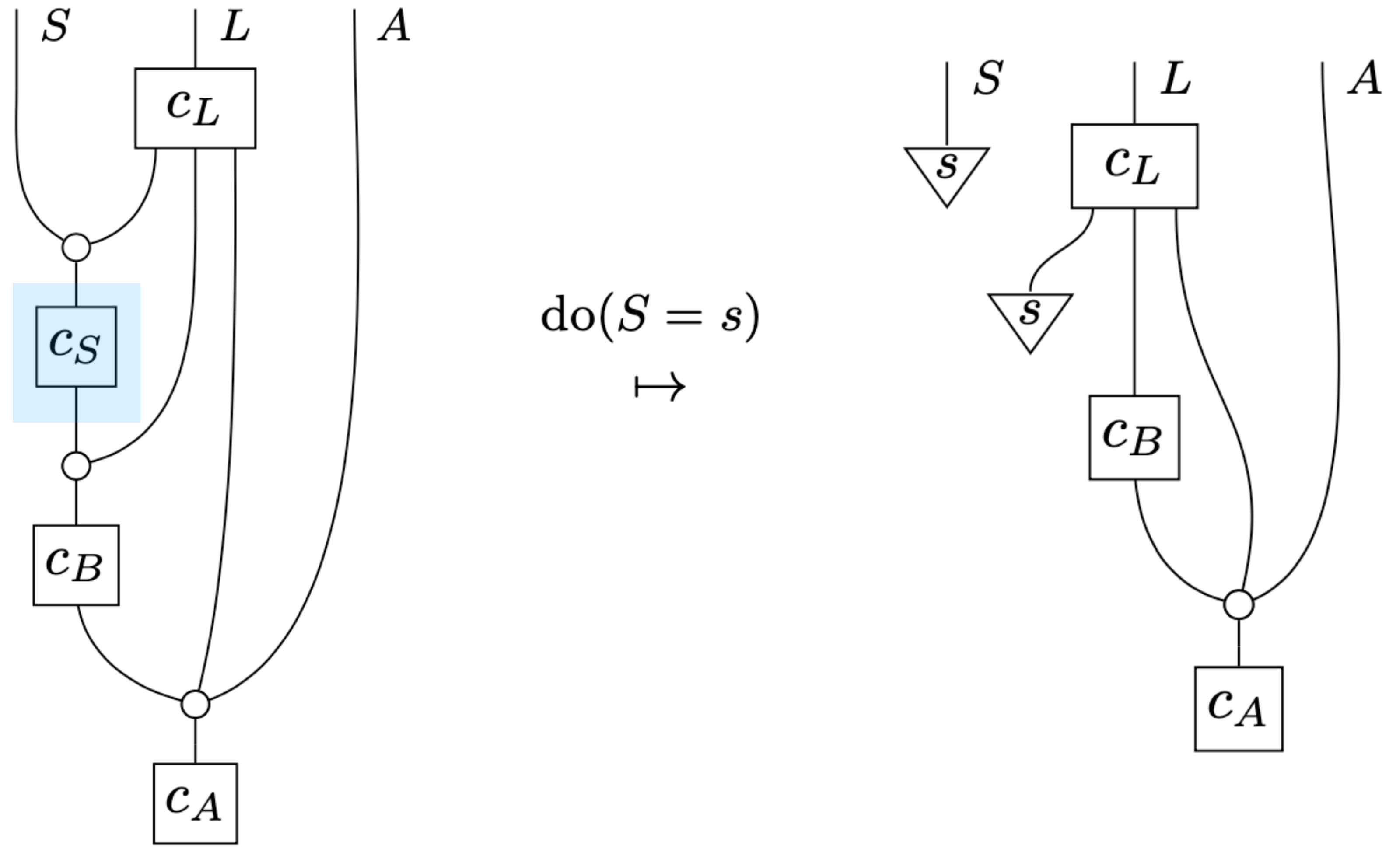
$$\mathrm{do}(S = s) \mapsto$$



Causal Models

The causal model **framework** provides further interpretability benefits:

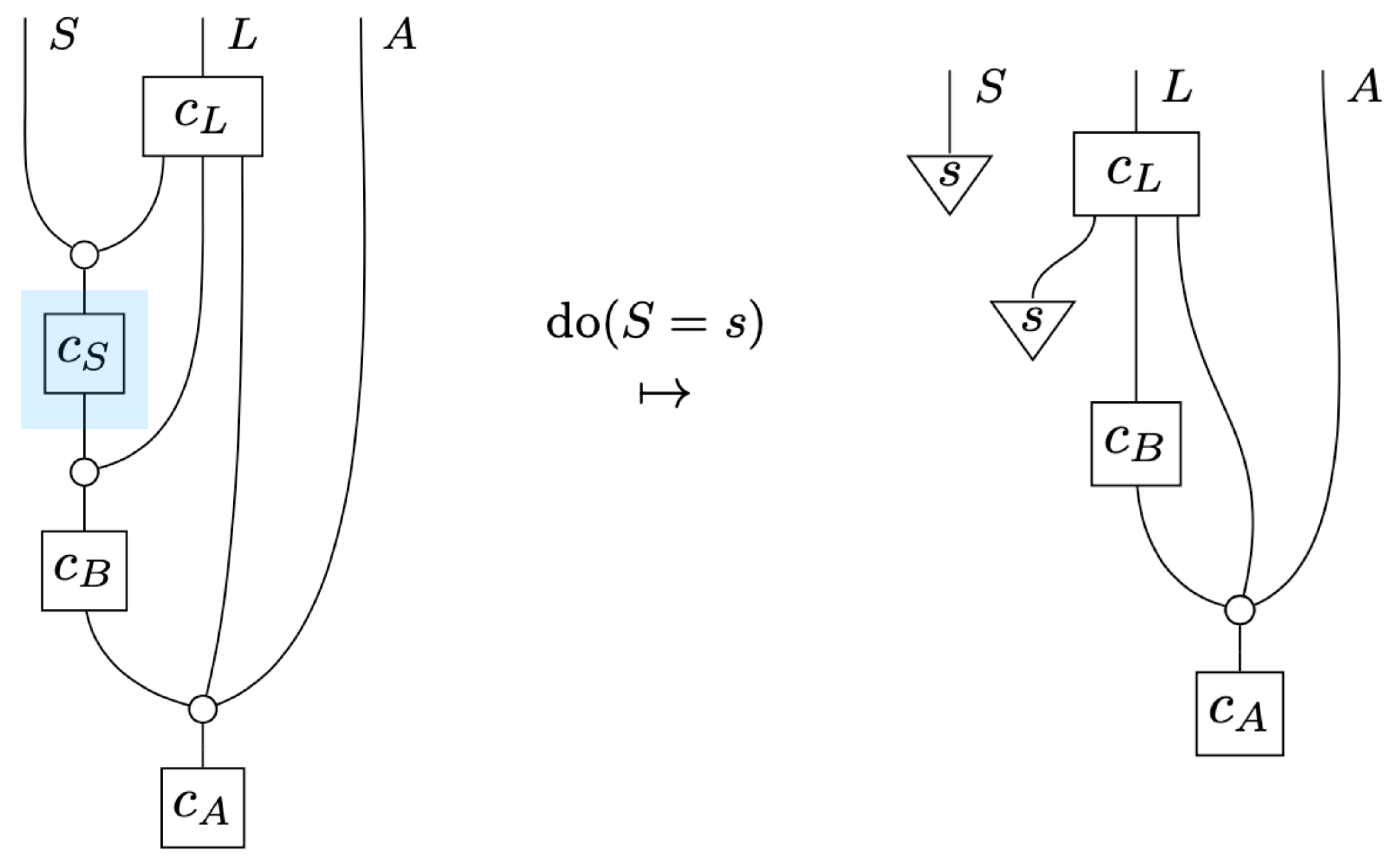
Interventions



Causal Models

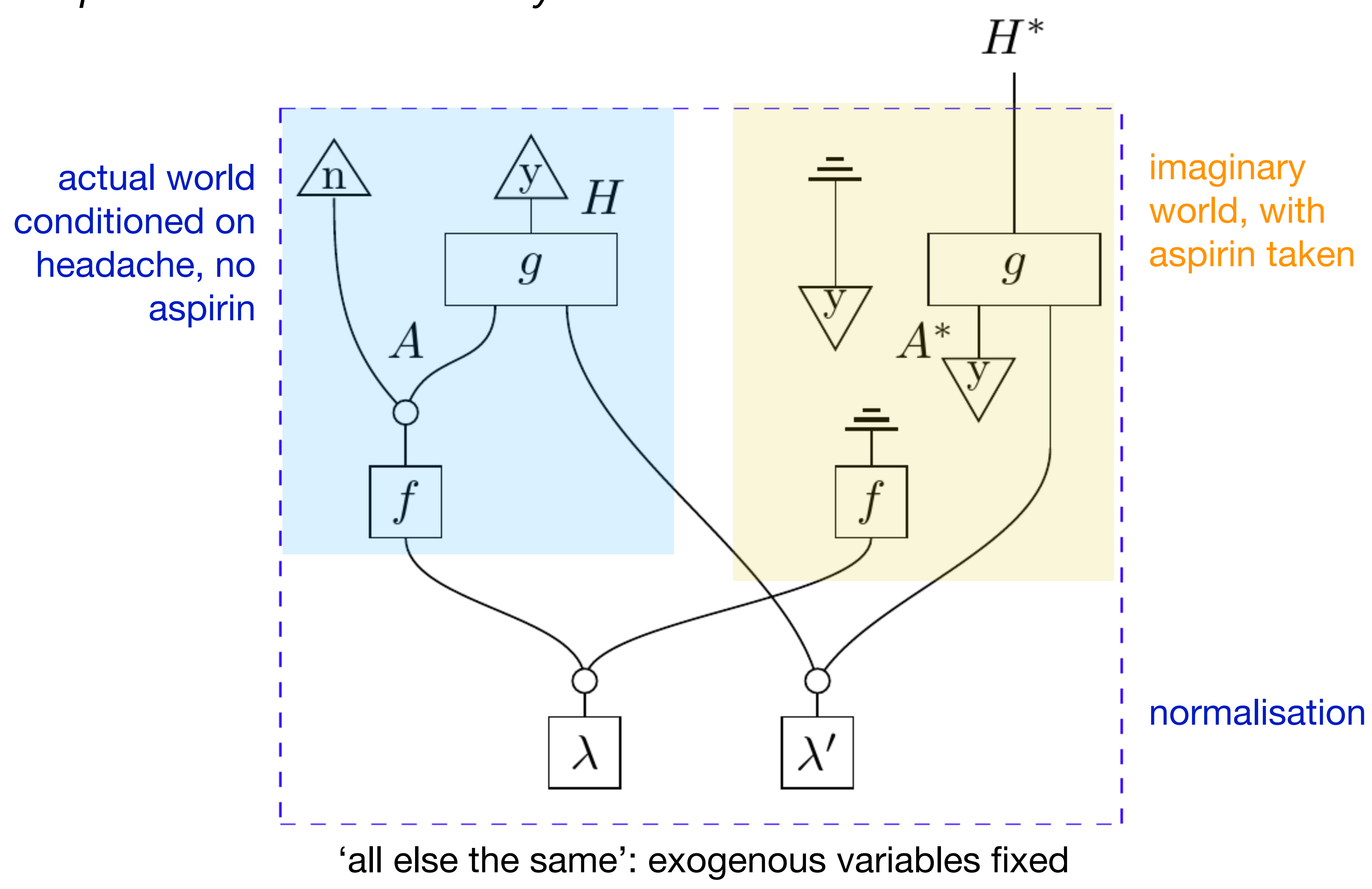
The causal model **framework** provides further interpretability benefits:

Interventions



Counterfactuals

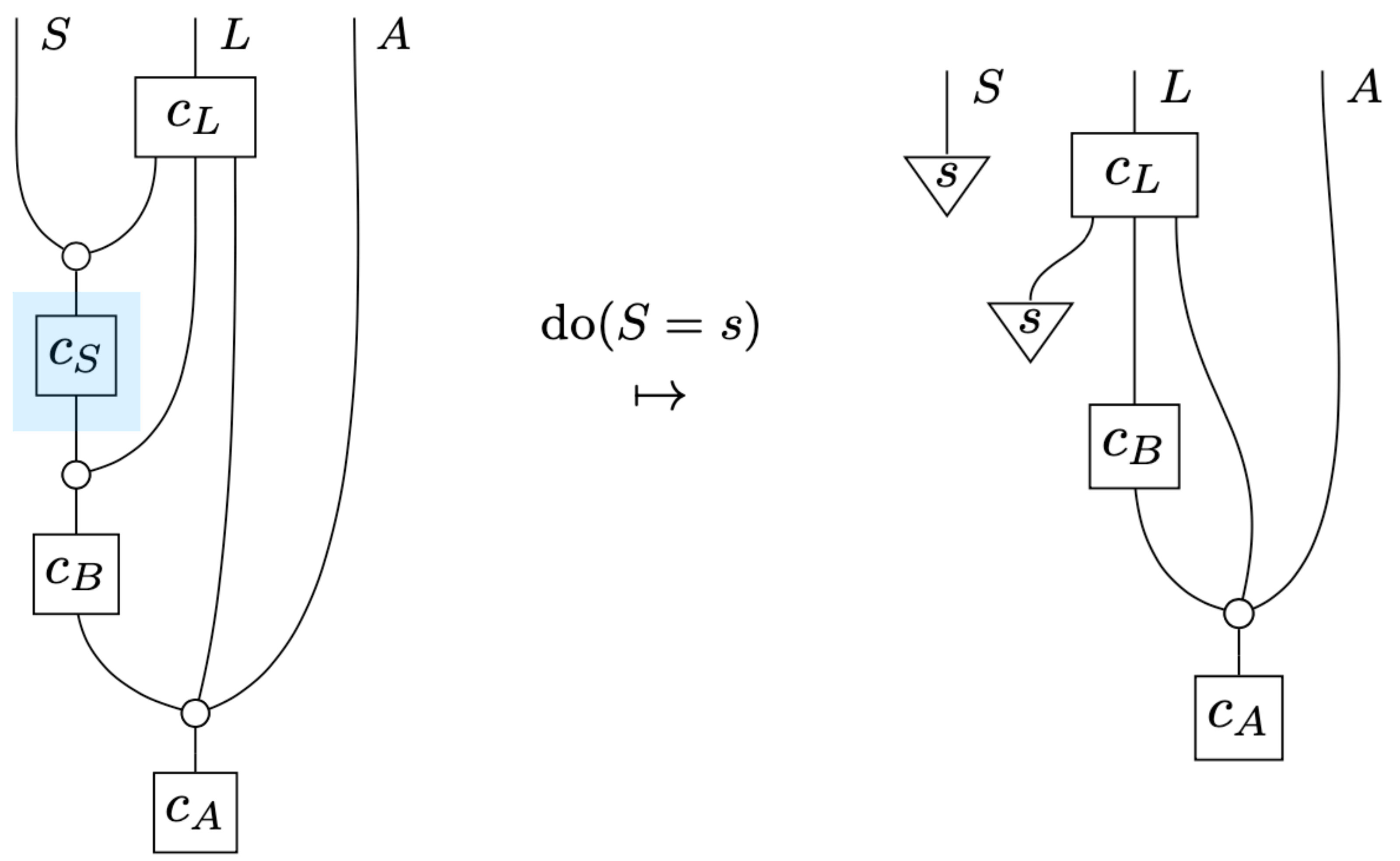
"Had Mary taken an aspirin last night, would she still have woken up with a headache today?"



Causal Models

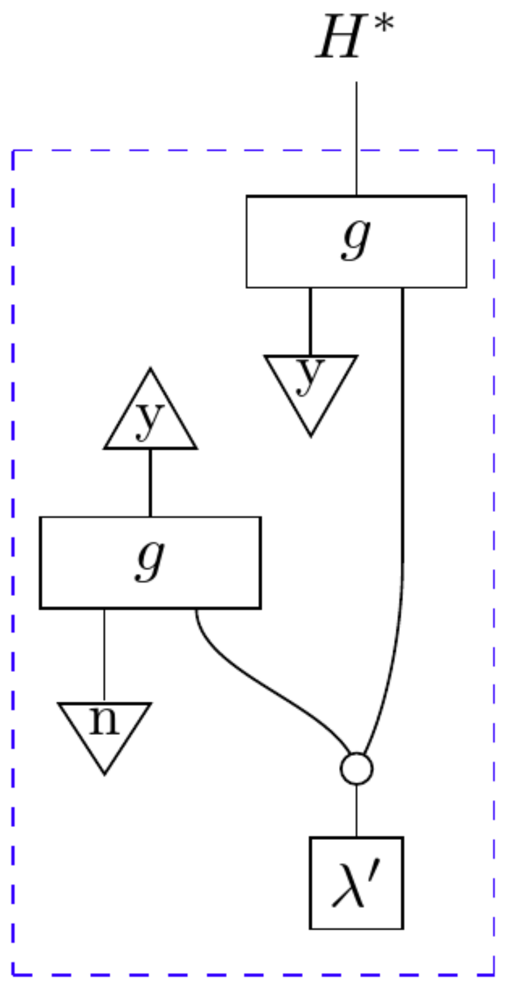
The causal model **framework** provides further interpretability benefits:

Interventions



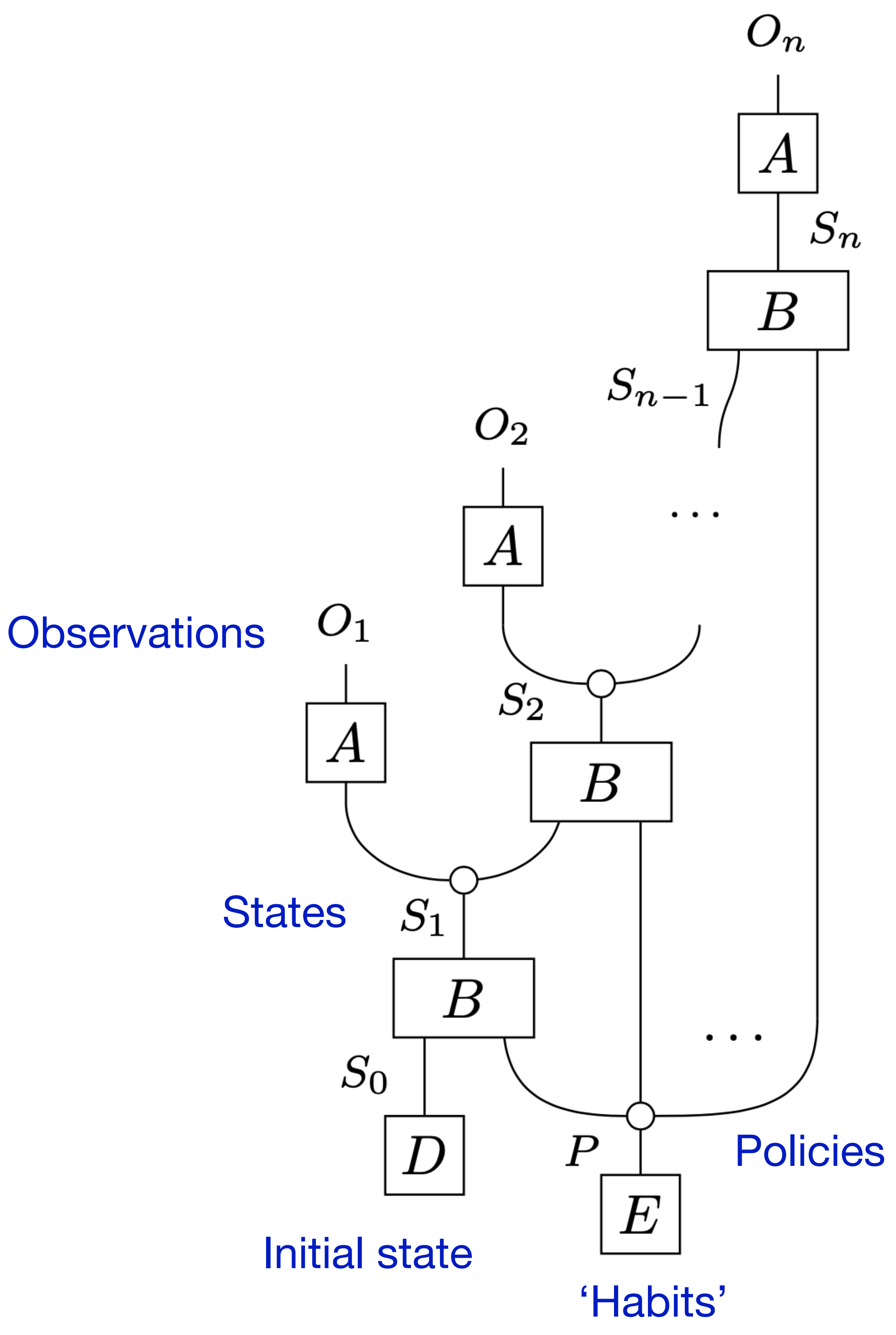
Counterfactuals

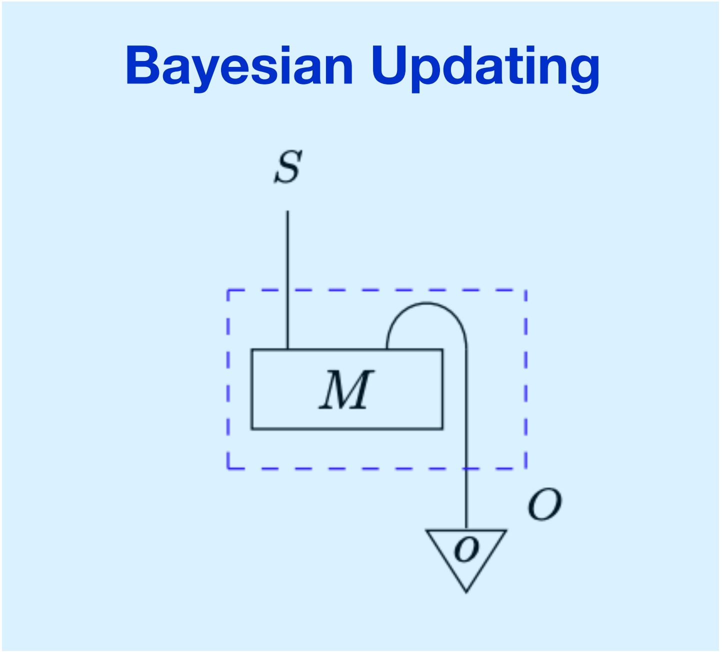
"Had Mary taken an aspirin last night, would she still have woken up with a headache today?"

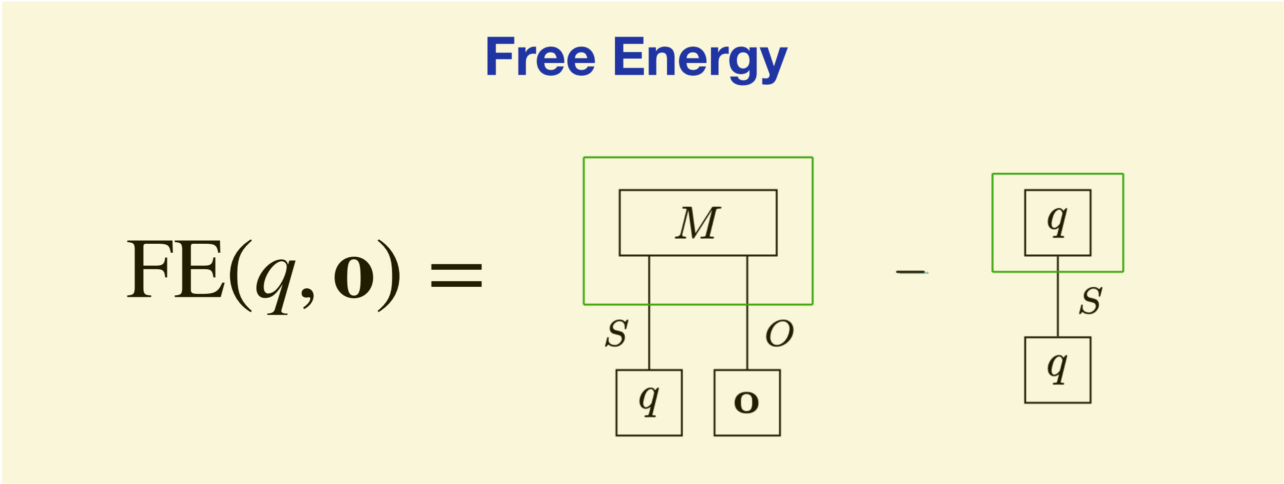


Models from Cognitive Science

Bayesian + Active Inference



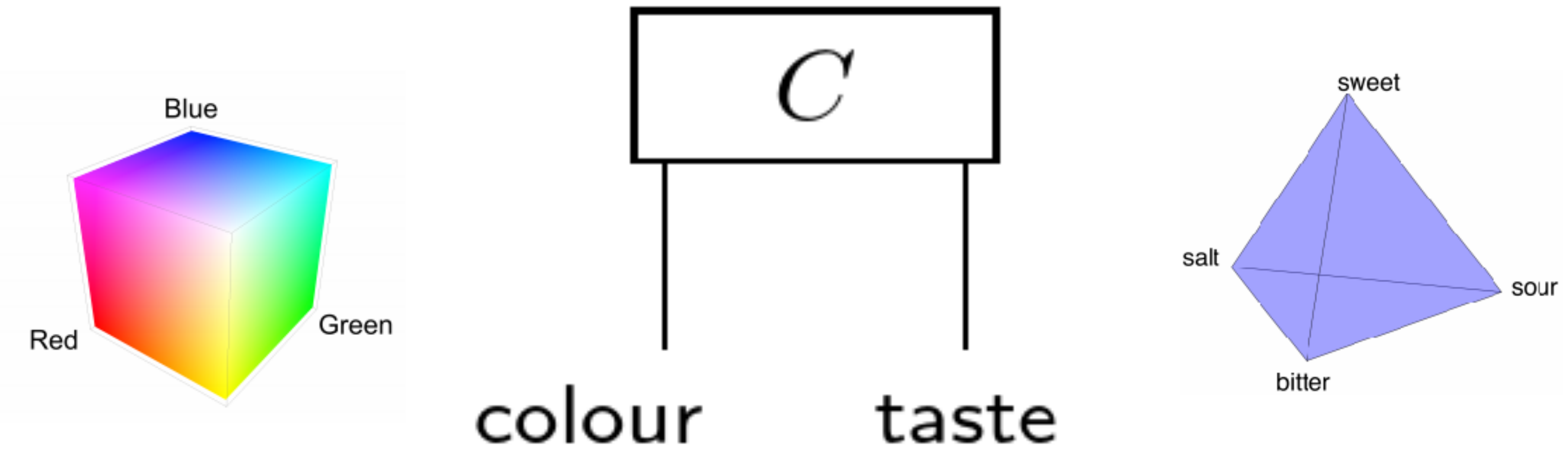


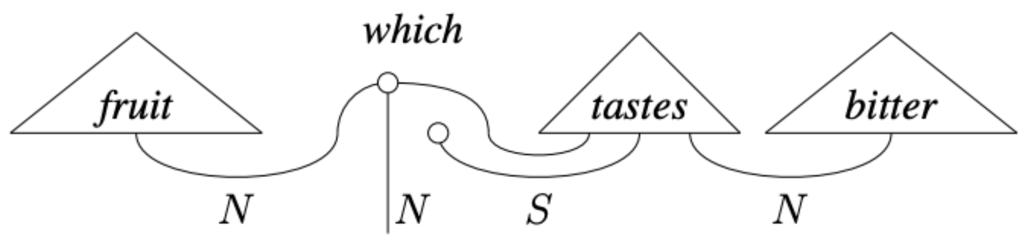


Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy

Sean Tull[1,2], Johannes Kleiner[2,3,4], and Toby St Clere Smithe[5,6]

Conceptual Spaces





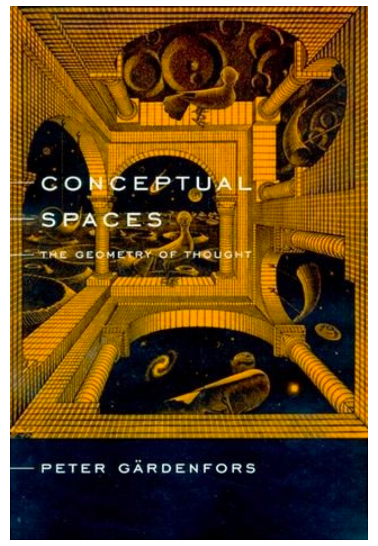
Interacting Conceptual Spaces I :

Grammatical Composition of Concepts *

Joe Bolt Bob Coecke Fabrizio Genovese Martha Lewis Dan Marsden Robin Piedeleu †

From Conceptual Spaces to Quantum Concepts: Formalising and Learning Structured Conceptual Models

Sean Tull, Razin A. Shaikh, Sara Sabrina Zemljič and Stephen Clark
Quantinuum



Explanations from Diagrams

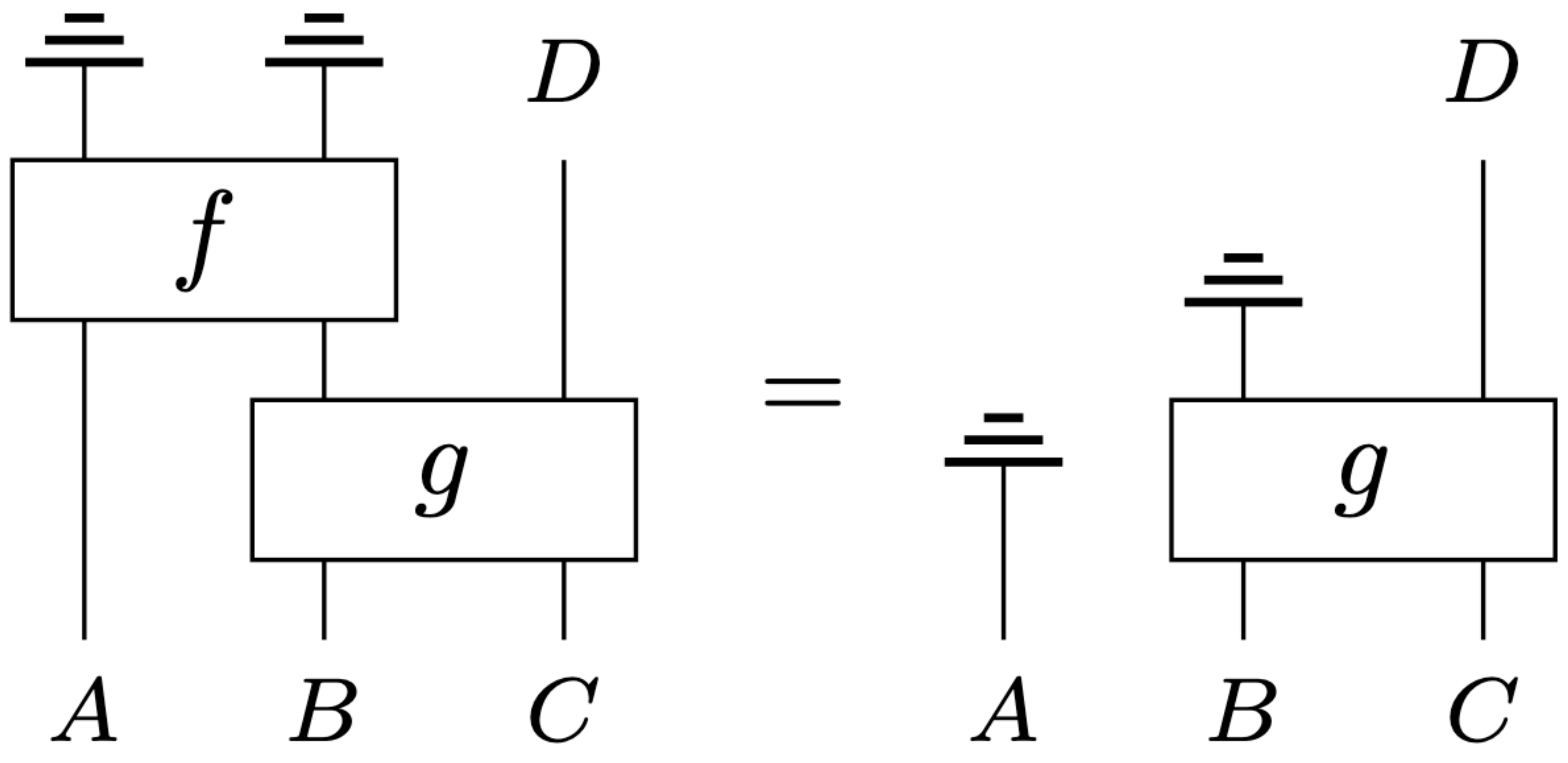
Explanations from Diagrams

How exactly does the compositional structure of a CI model yield **explanations** for its behaviour?

We propose three ways which are purely **diagrammatic**, and so in particular apply equally to e.g. classical or **quantum** models.

Influence Relations

For models based on (discard-preserving) **channels**, diagrams let us see which inputs can **influence** which outputs.



This is not possible for trivial compositional structure **e.g.** fully-connected NN layers.

Diagram Surgery

Each piece of an interpreted diagram forms a point where we may *intervene* by **diagram surgery**, to learn more about the process.

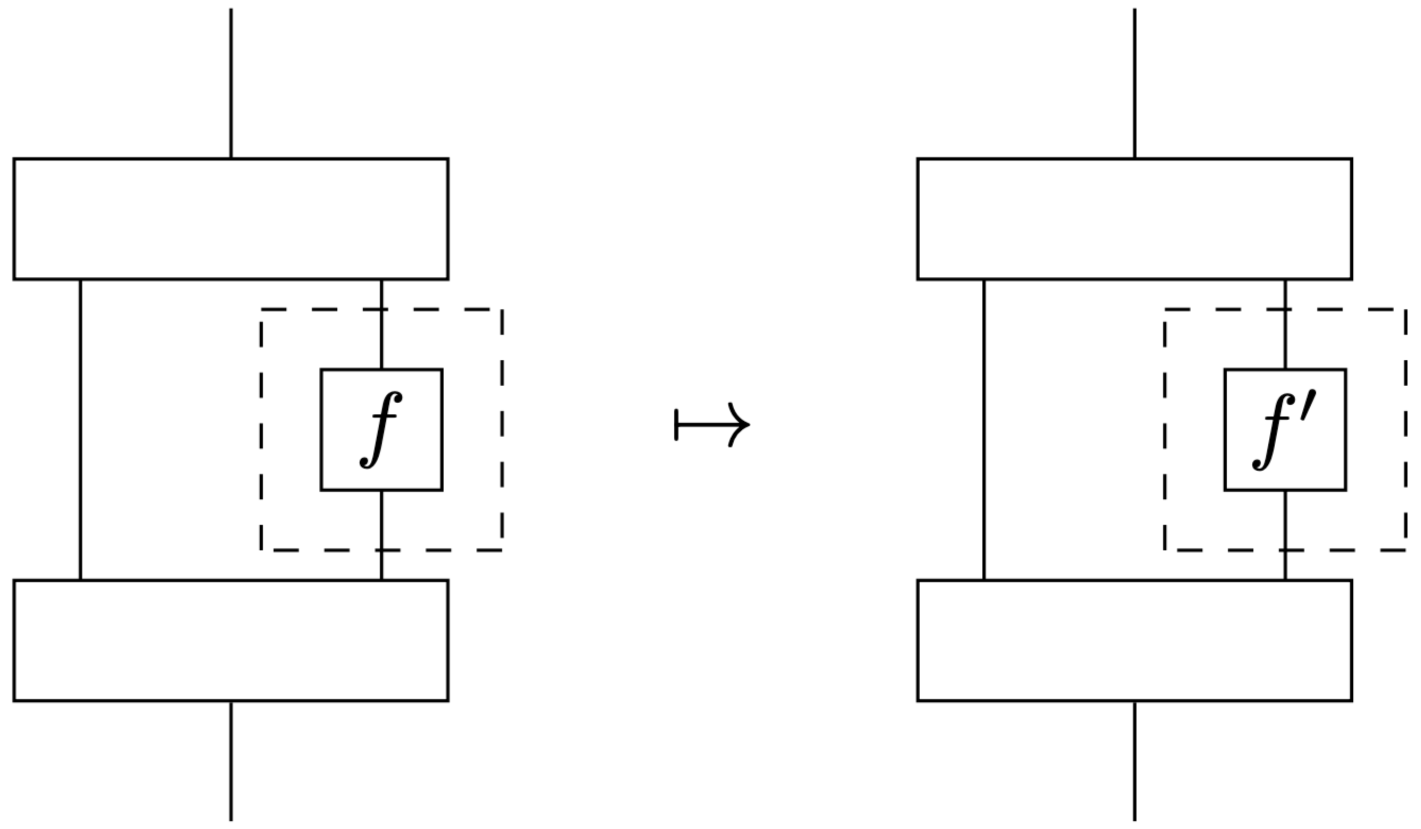
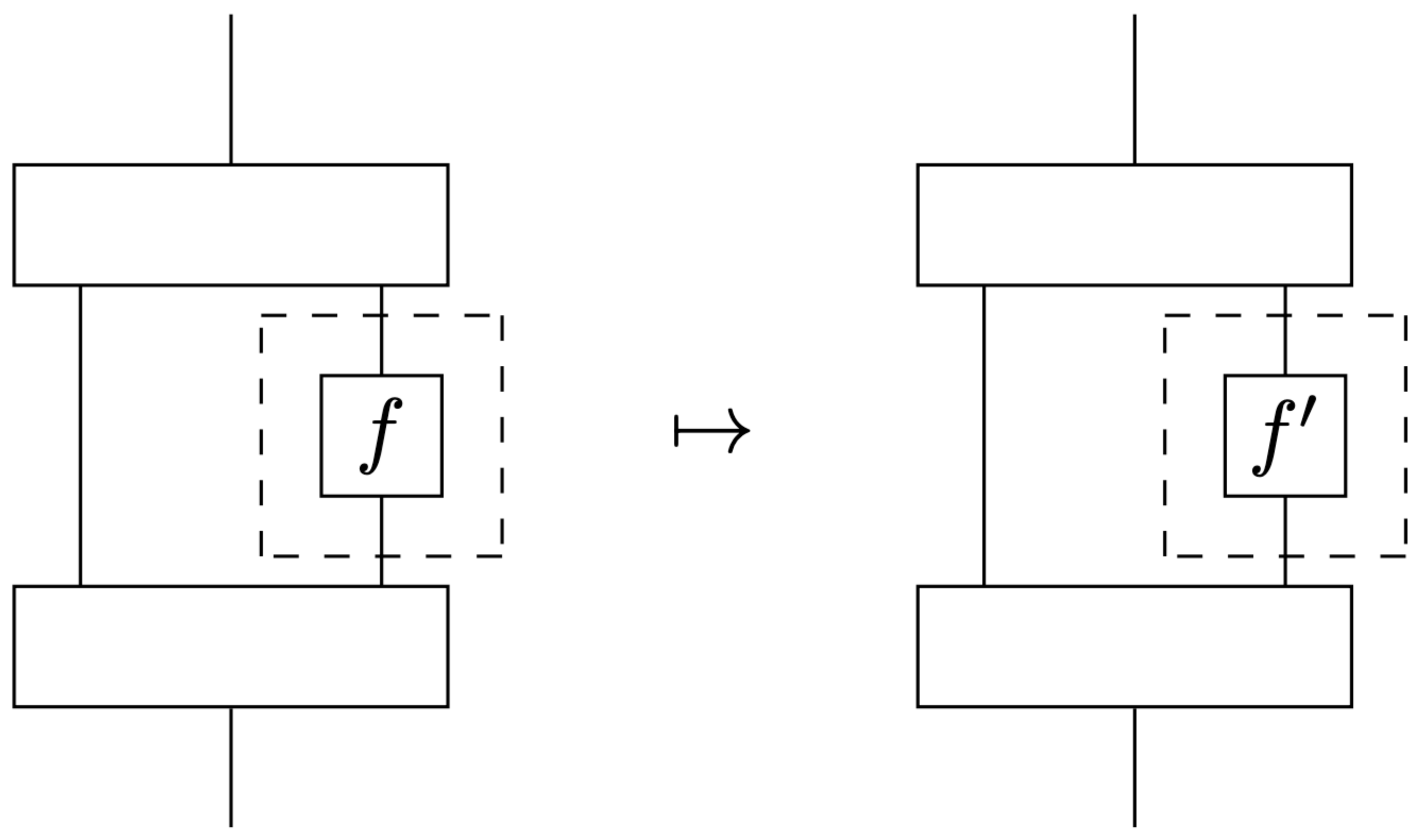
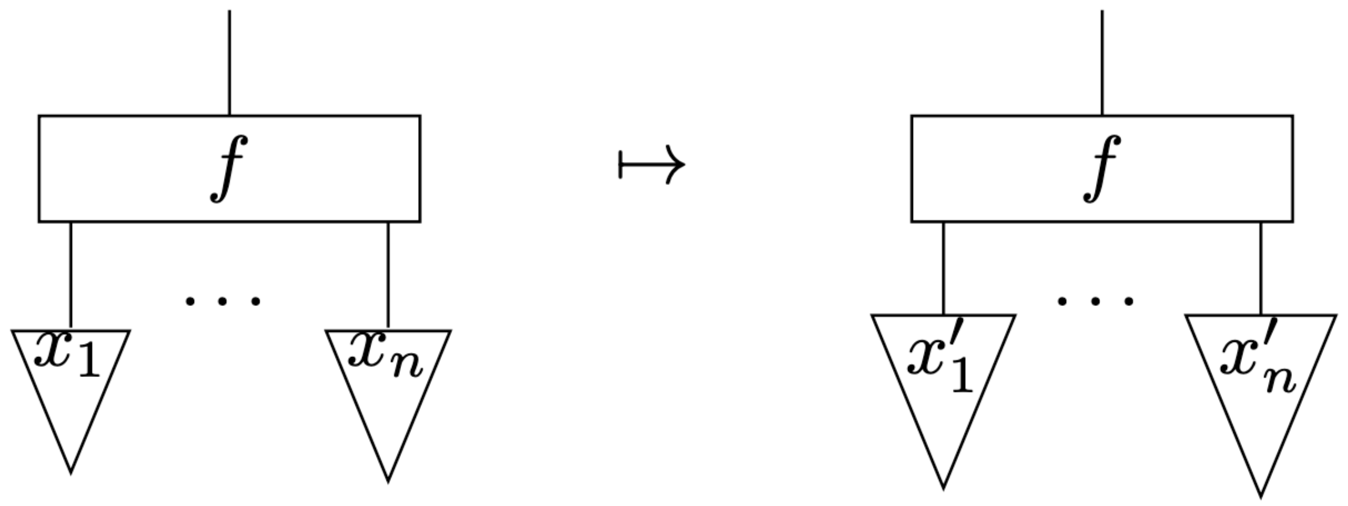


Diagram Surgery

Each piece of an interpreted diagram forms a point where we may *intervene* by **diagram surgery**, to learn more about the process.

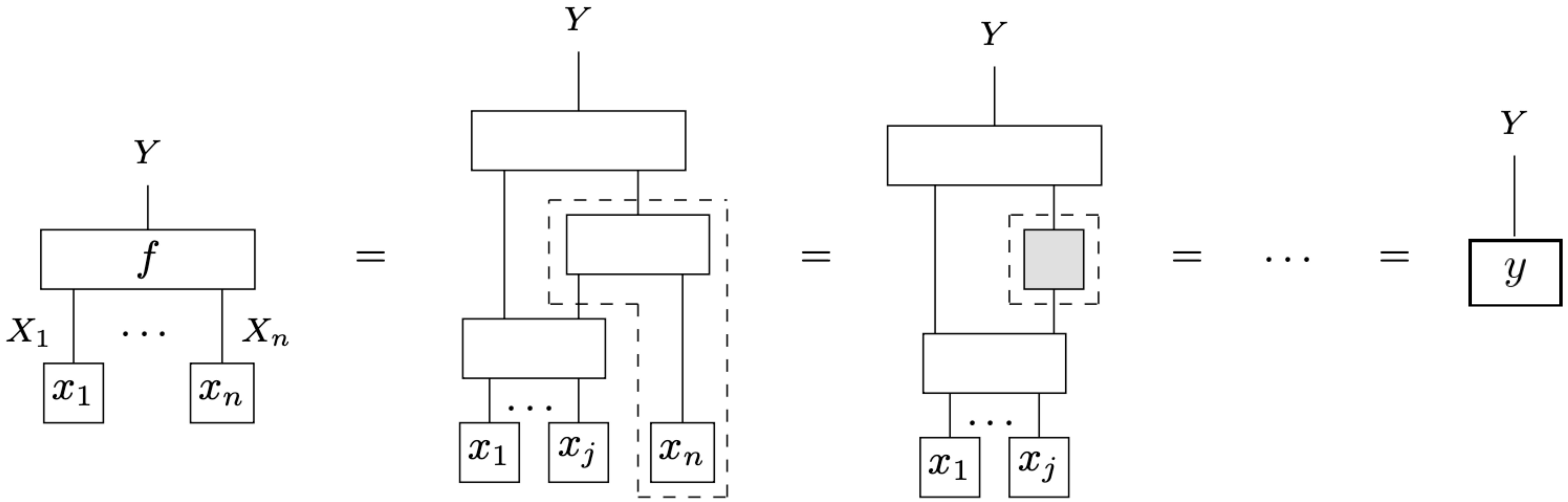


This generalises causal interventions, as well as **Counterfactual Explanations** in which one varies inputs to produce a given output.



Rewrite Explanations

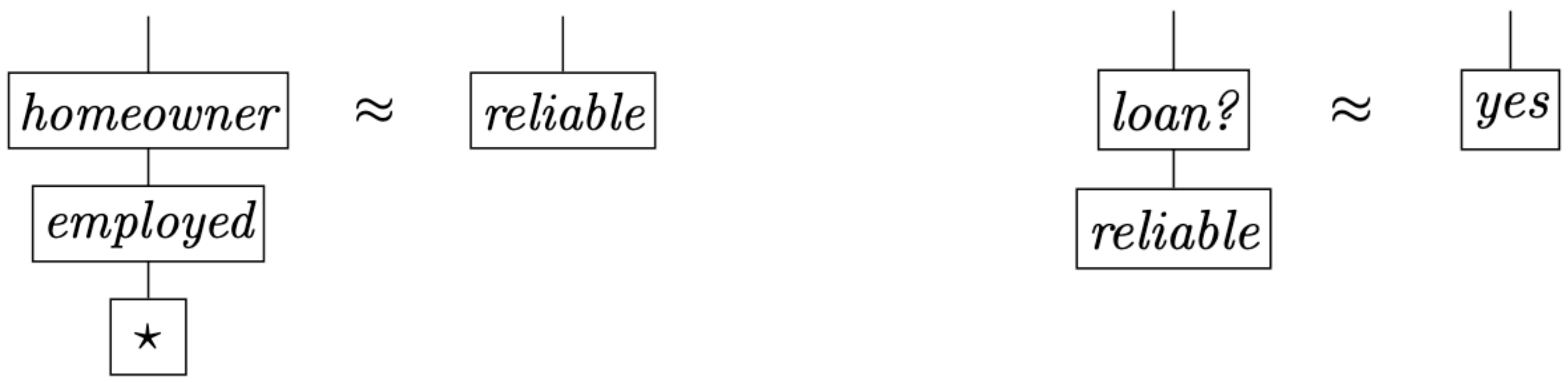
A **rewrite explanation** of an equality of **interpreted diagrams** $D = D'$ consists of a collection of further such equations $(D_i = D_i')_{i=1}^n$ and a proof that these imply $D = D'$.



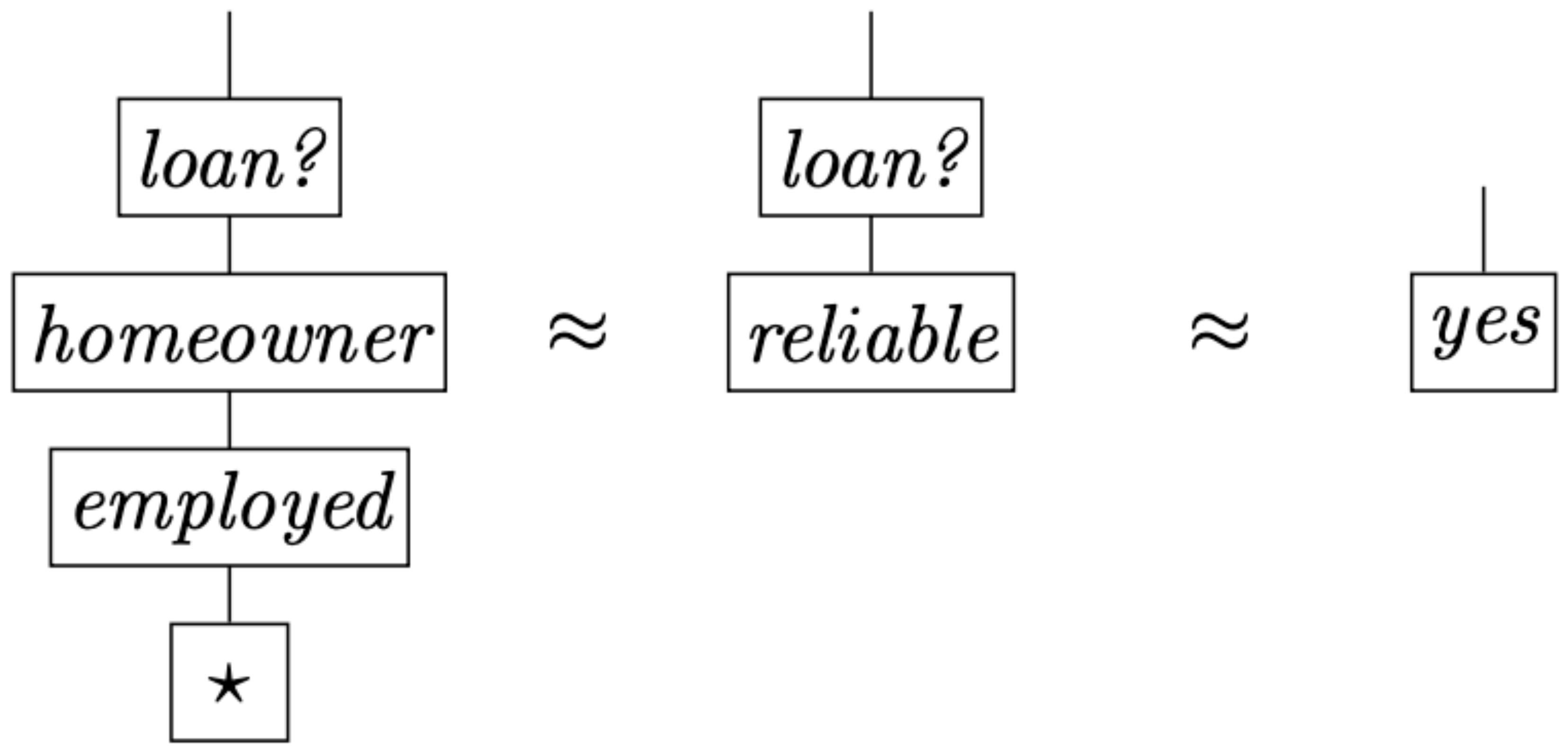
To count as an **explanation**, all diagrams involved must be interpreted.

Rewrite Explanations

Suppose a bank uses an RNN model, which (almost) always grants an employed homeowner a loan. An explanation is given by approximate equalities:

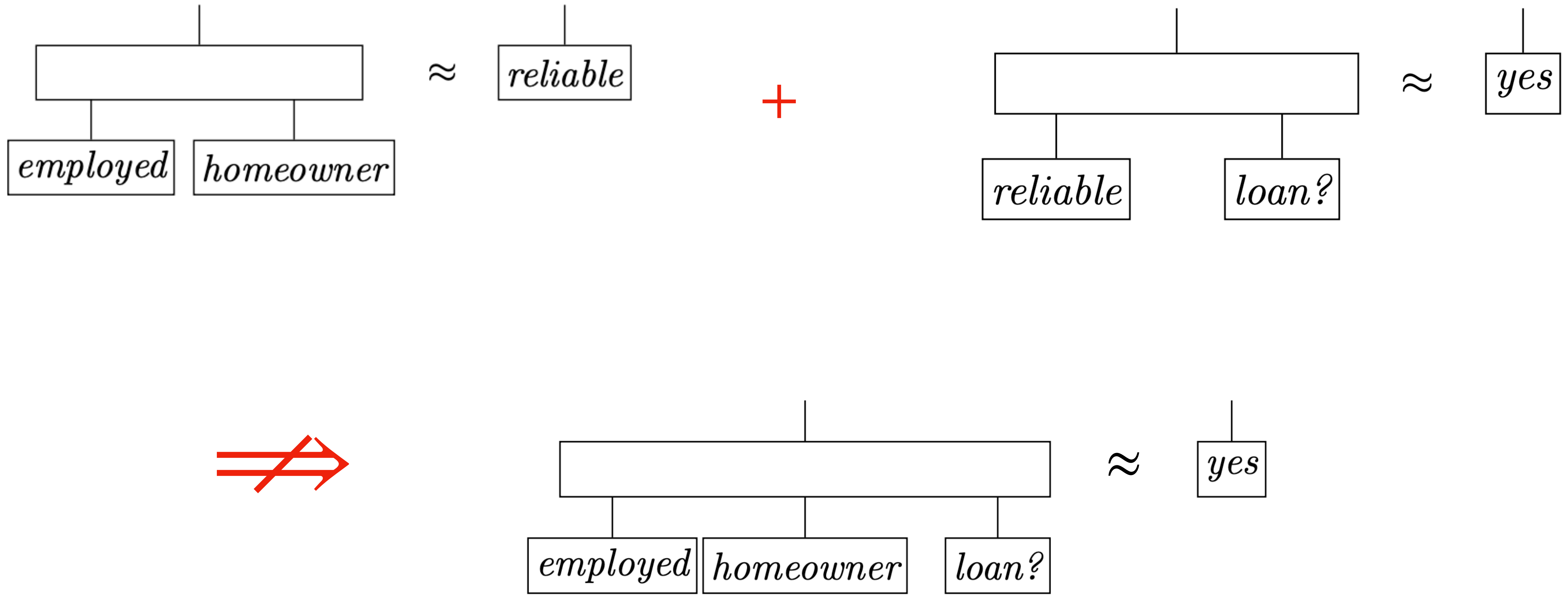


and the proof:



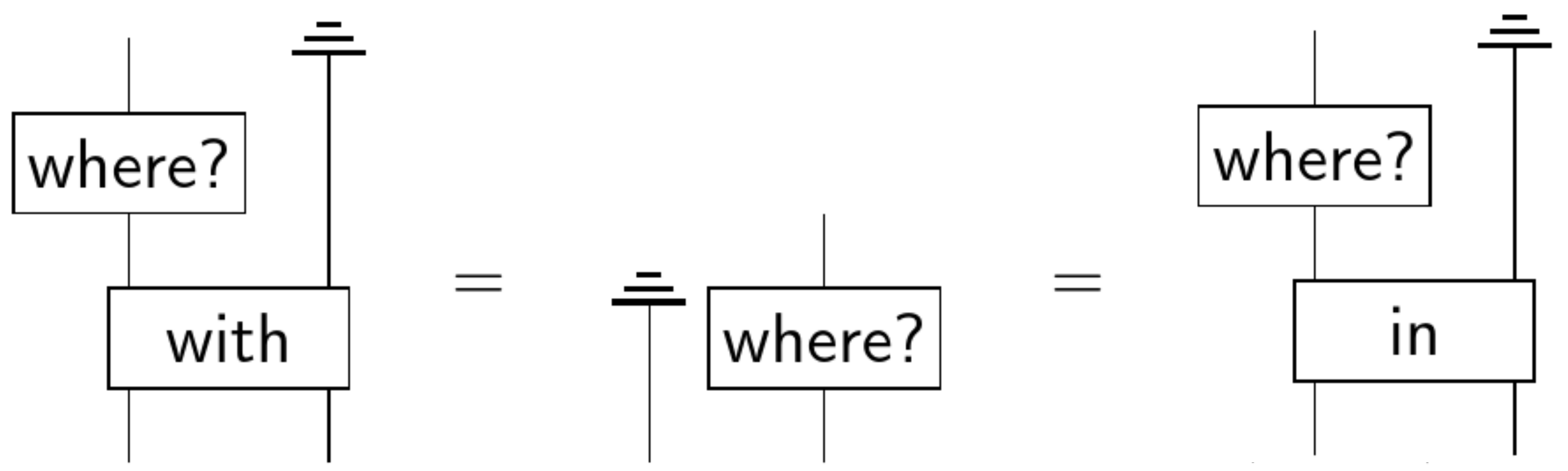
Rewrite Explanations

Such an argument is not possible for a black-box NLP model (e.g transformer):

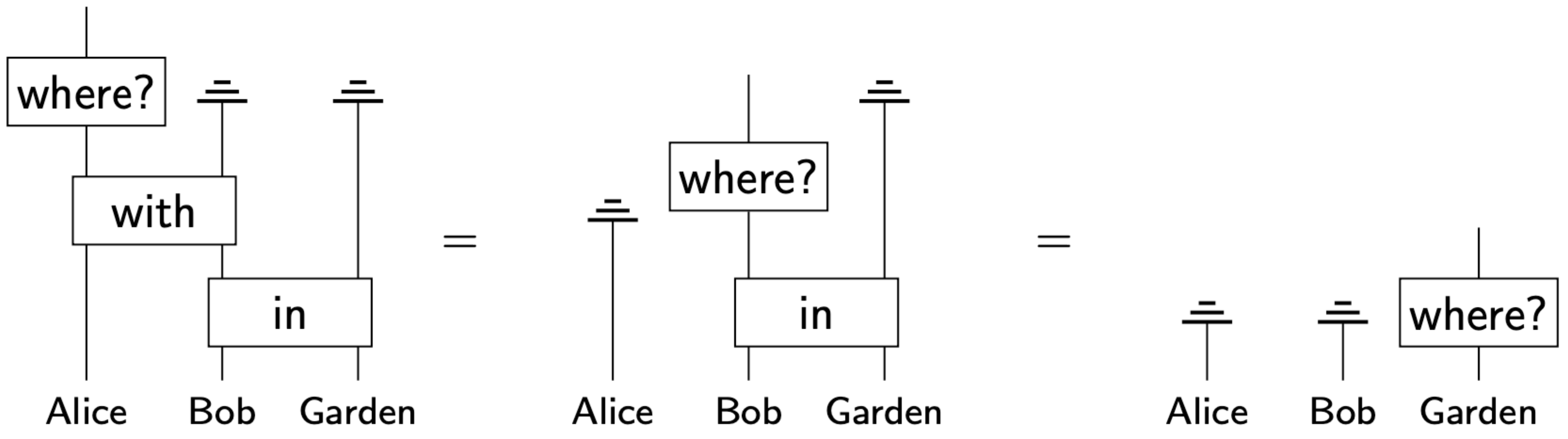


Rewrite Explanations

Consider a DisCoCirc model of text ‘*Alice is with Bob. Bob is in the garden. Where is Alice?*’.
An explanation for the answer *‘garden’* could consist of equations:



and proof:



Outlook

Summary

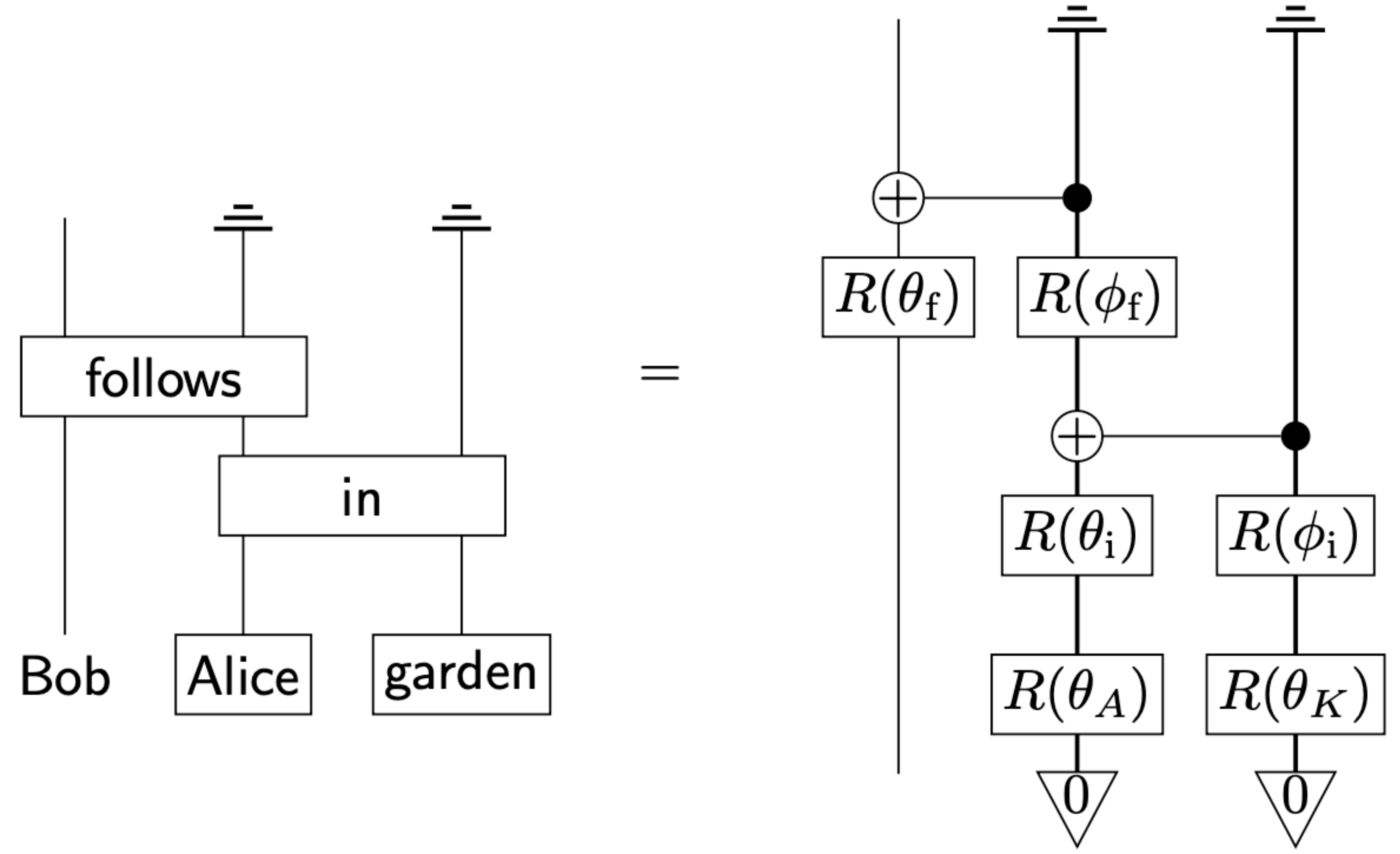
Categories provide a natural language for studying AI models and their **interpretability.**

Suggests broader class of interpretable models, those with **meaningful compositional structure** (CI).

Causal models are the CI models most widely studied in ML, but there are further examples e.g. DisCoCirc.

Quantum Models

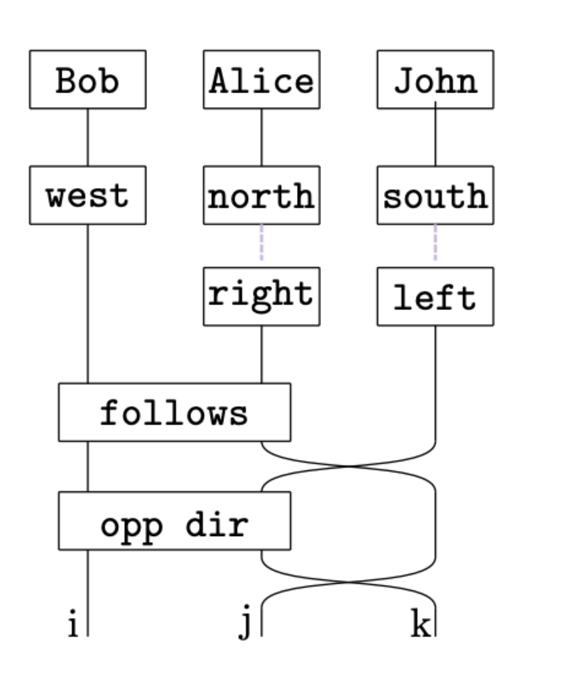
A categorical treatment is natural for **quantum AI models.**



- Quantum models are defined compositionally, as **parameterised quantum circuits** (PQCs)
- Our definitions, and explanation techniques, are independent of semantics so cover both classical and quantum
- Compositionally structured models (e.g. DisCoCirc) allow **'Train small, test big'**

Compositional Intelligence at Quantinuum

Quantum NLP



LAMBEQ

Natural Language Processing on Quantum Computers

`pip install lambeq`

GitHub | Discord

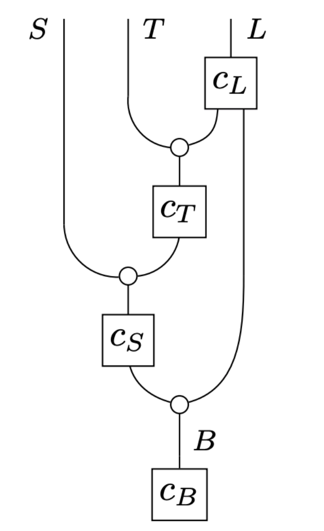
Scalable and interpretable quantum natural language processing: an implementation on trapped ions

Tiffany Duneau[1,2], Saskia Bruhn[1,*], Gabriel Matos[1,*], Tuomas Laakkonen[1], Katerina Saiti[3], Anna Pearson[1,*], Konstantinos Meichanetzidis[1], Bob Coecke[1]

Quantum Algorithms for Compositional Text Processing

Tuomas Laakkonen, Konstantinos Meichanetzidis, Bob Coecke
Quantinuum, 17 Beaumont Street, Oxford OX1 2NA, United Kingdom

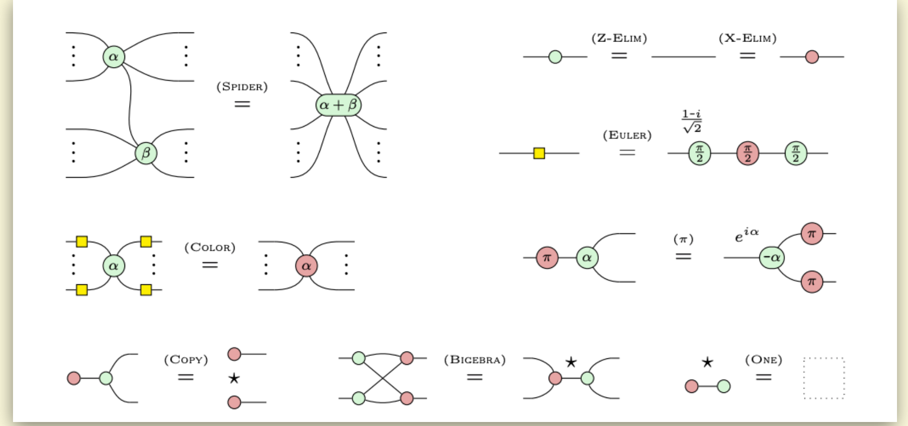
Causal Models



Causal models in string diagrams

Robin Lorenz*, Sean Tull†
Quantinuum, 17 Beaumont Street, Oxford, UK

ZX Calculus



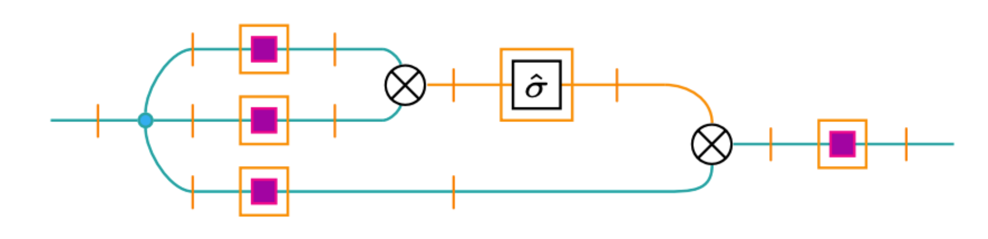
ZX-calculus is Complete for Finite-Dimensional Hilbert Spaces

Boldizsár Poór[1] Razin A. Shaikh[1,2] Quanlong Wang[1]
[1]Quantinuum, 17 Beaumont Street, Oxford, OX1 2NA, United Kingdom
[2]University of Oxford, United Kingdom

Fusion and flow: formal protocols to reliably build photonic graph states

Giovanni de Felice[1], Boldizsár Poór[1], Lia Yeh[1,2], and William Cashman[2]
[1]Quantinuum, 17 Beaumont Street, Oxford, OX1 2NA, United Kingdom
[2]University of Oxford, United Kingdom

Compositional ML



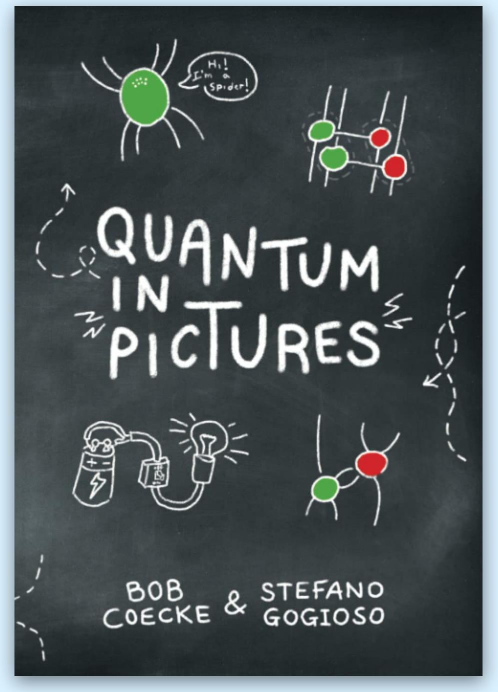
A Pattern Language for Machine Learning Tasks

Benjamin Rodatz*‡† Ian Fan*† Tuomas Laakkonen† Neil John Ortega† Thomas Hoffman† Vincent Wang-Maścianica*†‡#
†Compositional Intelligence, Quantinuum

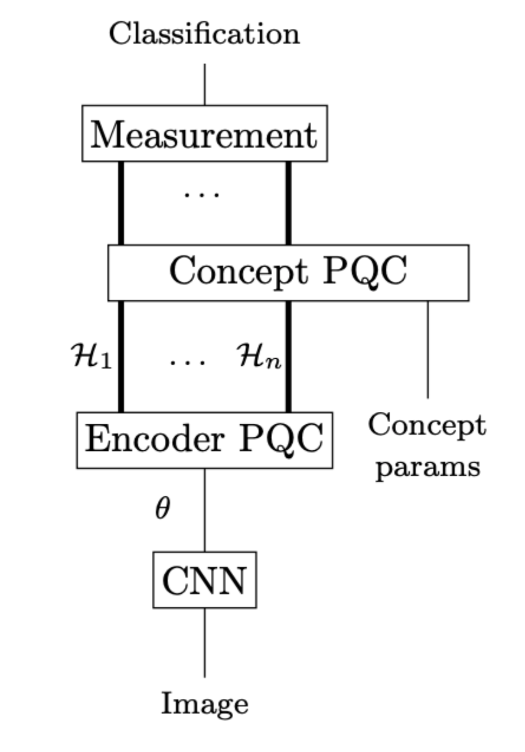
On the Anatomy of Attention

Nikhil Khatri* Tuomas Laakkonen* Jonathon Liu* Vincent Wang-Maścianica*‡
Compositional Intelligence, Quantinuum

Education

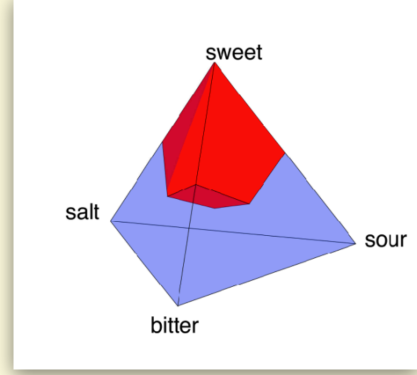


Conceptual Spaces



From Conceptual Spaces to Quantum Concepts: Formalising and Learning Structured Conceptual Models

Sean Tull, Razin A. Shaikh, Sara Sabrina Zemljič and Stephen Clark
Quantinuum



+ Much more!

Future Directions

How can we **learn** compositional structure from raw data? cf causal representation learning

How can we **relate** low-level neural networks to a high-level CI model? cf causal abstraction

What benefits do **quantum** compositional models bring?

Thanks!