Indicator Frameworks

Joshua Tan, Christine Kendrick, Abhishek Dubey, and Sokwoo Rhee

March 7, 2018

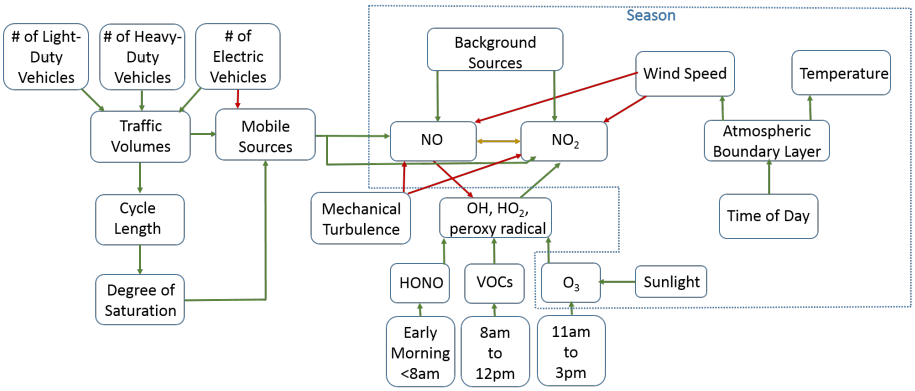
Abstract

We consider the mathematical semantics of operational (key performance) indicator frameworks, especially as they are used in city governance, and give a rigorous, statistically-motivated process for constructing such indicator frameworks.

1 Introduction

"You shall know an indicator by the company it keeps."

We take as our starting point a diagram of simple correlations, as below:



This diagram correlates the measurement variables of an air pollution monitoring system with other, partially-observable variables like traffic composition, mechanical turbulence, and the presence of sunlight. It presents an intuitive and apparently useful description of a system at large. We would like to clarify the meaning of this diagram, and of others like it, by giving its interconnections a precise mathematical meaning. Clarifying the meaning of the diagram will not only make it more useful; it will allow us to connect this local, correlation-based picture of a system with other local pictures, as well as with more sophisticated scientific models of the world.

In this paper, we develop *abstract indicator frameworks*, a mathematical tool for interpreting and combining systems of random variables and their correlations. Abstract indicator frameworks are modeled off operational (key performance) indicator frameworks, especially as they are used in city planning and project governance. Such operational indicator frameworks have three main uses: (1) to communicate quantitative information and strategic priorities to a wide audience, (2) to enable policy reactions to data, especially in the optimization of processes, and (3) to restrict attention to a set of 'relevant' indicators—thus discarding the information from many other, 'non-relevant' indicators.

In city planning, there are several strategy-setting frameworks for constructing operational indicator frameworks, from balanced scorecards [3] to SMART [2] to more specialized urban planning frameworks; in such frameworks, the indicators are often designed by mayors, chief strategy officers, and sizable expert committees in tandem with new projects, new policies, and new processes. Even assuming that the participants adhere to a framework, the process of choosing indicators is often ad hoc, the choice of indicators does not account for pre-existing statistical relationships between the indicators, and integrated data for all such indicators can be hard to capture. Once built, the indicator frameworks for cities also tend to be large and unwieldy: there are 175 indicators in the CITYkeys standard [**?**] and 212 in the Boston Indicators Project [**?**]. (And there are already more than 43 indicator frameworks built for "smart and sustainable cities" alone!) The point: indicator frameworks can be useful locally, but they are hard to build at scale, hard to sustain at scale, and hard to use at scale.

We propose an alternative: "start with the data you have, not the data you want". Instead of constructing operational indicator frameworks expensively and internally, meaning indicator-by-indicator, we can specify them abstractly and externally, by means of their causal and statistical relationships to other, already-extant sets of indicators. Our approach is especially suited to situations where heterogeneous measurement data is distributed across many projects, many subject domains, and many localities, as in a city.

However, such an approach faces several practical data-engineering challenges:

Problem 1.1 (Data integration)**.** It is not always possible to directly compare local sets of indicators—e.g. to construct a correlation matrix—especially when the indicator sets are disjoint. The combined data over all the indicators may not exist; or if the data does exist, it may not be sufficient to support the kinds of arguments the user wants to make; or it may be difficult (and prohibitively expensive) to integrate the data.

Problem 1.2 (Model integration)**.** The approach fundamentally limits our ability to choose the right indicators. We are given observational data that is presumed sufficient to manage some local systems of interest, like a bus line or a classroom, but this data cannot usually capture all the possible causal interactions between the variables of a larger system, e.g. transportation *and* education in the City of Boston. Managing—not solving—this problem requires that we

supplement the missing data with background knowledge about the larger system under study. But such knowledge, like the data, tends to arrive modularly, over the local systems in question. So we need not only the background knowledge, but also a way to stitch different models together.

Problem 1.3 (Compositional semantics)**.** An operational account of indicator frameworks needs to tell us not just what they are, but how they are *used*. To us, this is a question of semantics: the way we use an indicator—the kind of arguments we make with it—depends on how we interpret its meaning, and this meaning is governed by installing a mathematical model over the data.

These three problems motivate the definition of abstract indicator frameworks. An operational indicator framework is a list of indicators along with an account of how these indicators are used. An abstract indicator framework is, roughly, a formula for turning "scientific models" into spaces of random variables. (Such frameworks may be compared with the usual algorithms for using data to construct a scientific model.) But an abstract indicator framework can also be understood as an object on its own, rather than as a formula; in particular, it is possible to define maps and gluings between different abstract indicator frameworks. We will apply *category theory*, originally developed to relate and analyze topological spaces, as an efficient language for relating and analyzing the causal and statistical aspects of abstract indicator frameworks. To illustrate the mathematics, we will develop a running example of how a mayor can implement the approach.

1.0.1 Notation

Let $\mathcal{X}$ stand for an indicator set. Abstract indicator frameworks have a notion of *process* that transforms one indicator set into another, a notion of *state* that represents the process of picking a specific indicator in $\mathcal{X}$, and a notion of *effect* that represents the process of "measuring" or computing the correlation with respect to a specific indicator in $\mathcal{X}$. Processes, states, and effects are represented, respectively:

These are needed to capture the operations of composing processes in sequence, called *composition*, and combining them in parallel, called *tensoring*. The composition $g \circ f$ (first f, then g) and tensor $f \otimes g$ are represented as:

We call any formalism with a notion of composition and tensoring a *process theory*. The semantics of process theories and their diagrams are governed by the

theory of monoidal categories, which is surveyed in [11]. The goal of the paper is to specify an appropriate symmetric monoidal category, Rand, representing the appropriate operations on random variables, after which we can define a causal model as a strong monoidal functor from a causal theory into Rand. These causal models—essentially, diagrams in Rand—will be the promised abstract indicator frameworks. In **??**, we will consider a preliminary version of Rand along with some of the possible alternatives. In **??**, we will give the statistical justification for our choice of Rand, review the notion of a causal model from [4], and then give the full definition of abstract indicator frameworks.

2 Background

As mentioned above, there are a variety of approaches to choosing indicator frameworks as part of the process of strategic priorities. Of the many specialized approaches to choosing indicator frameworks in various fields, Niemejer and de Groot [9] have suggested a similar methodology for choosing environmental indicator sets based on explicit causal networks of environmental forces and societal response; while their methodology is still largely qualitative rather than formal or statistical, their paper handily illustrates how (diagrams of) causal models can facilitate the selection of relevant indicator sets.

Relate to "conceptual spaces" idea in Bob's work as well as in the original Gardenfohrs, since we are depending very much on Hilbert spaces to store the geometry of our data.

Even within the constraints of a process theory, there are still a number of approaches to formalizing probability. In this section, we will go over three examples: the Hilbert space interpretation of random variables, the original category of probabilistic mappings suggested by Lawvere [8], and the diagrammatic approach of Coecke and Spekkens [1] to Lawvere's work. Also discuss the approach, in quantum physics, of mapping various models into categories of C^*-algebras. We also briefly discuss graphical models such as those surveyed in [7], which are the most obvious applications of [8] and [1], e.g. see [4].

The Hilbert space interpretation, which we call Rand, uses the fact that real-valued random variables over some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ form a Hilbert space H where the inner product $\langle X, Y \rangle$ is just the covariance $\mathbb{E}(XY)$. Assuming that we restrict ourselves to standard variables with zero mean and unit variance, the covariance equals the correlation, and we can represent both by the inner product in Rand. This inner product can be represented by a process diagram, namely as the composition of a state and an effect in Rand:

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y) = \langle X, Y \rangle =$$

As we will see in the next section, Rand is actually already very close to what we want; the problem is that the obvious categorical interpretation of Rand does

not give a natural way of analyzing the data of "intermediate" correlations.

Rand is closely related to older work on the categorical foundations of probability initiated by Lawvere in [8] and developed in Giry [5]:

Definition 2.1. The category Stoch of stochastic processes is defined by the following data:

1. objects are measurable spaces (A, Σ_A) of sets A with a σ-algebra Σ_A
2. morphisms $P : (A, \Sigma_A) \to (B, \Sigma_B)$ are stochastic kernels, i.e. functions $P : A \times \Sigma_B \to [0, 1]$ that assign to (a, σ_B) the probability of σ_B given a, denoted $P(\sigma_B|a)$
3. composition $Q \circ P : A \times \Sigma_C \to [0, 1]$ of $P : (A, \Sigma_A) \to (B, \Sigma_B)$ and $Q : (B, \Sigma_B) \to (C, \Sigma_C)$ is defined by

$$(Q \circ P)(\sigma_C|a) = \int_{b \in B} Q(\sigma_C|b) dP_a,$$

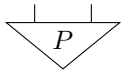
   i.e. marginalization over B

As suggested by the notation, morphisms in Stoch represent probability measures on an event/outcome space (A, Σ_A). If we restrict to the subcategory FinStoch whose objects are *finite* measurable spaces—since the outcomes are finite, one can imagine these spaces as sets of natural numbers $\{1, ..., n\}$—then we can think of stochastic kernels are stochastic matrices, i.e. matrices whose column entries sum to 1. Taking $1 = (\{*\}, \Sigma_*)$ as the monoidal unit, we can see that a probability distribution in FinStoch is just a vector $P : 1 \to (A, \Sigma_A)$ whose entries are the probabilities of all the possible atomic outcomes in A. The usual tensor product described in Stoch is the functor $\otimes : \mathsf{Stoch} \times \mathsf{Stoch} \to \mathsf{Stoch}$ that assigns to two probability distributions $P : 1 \to (A, \Sigma_A)$, $Q : 1 \to (B, \Sigma_B)$ their product measure, i.e. the map $PQ : 1 \to (A \times B, \Sigma_A \otimes \Sigma_B)$ s.t. $PQ(*, (a, b)) = P(*, a)Q(*, b)$.

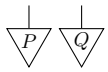
Using the language of symmetric monoidal categories, Coecke and Spekkens [1] give a graphical calculus for FinStoch and use it to elaborate Bayesian reasoning (in particular, a diagrammatic representation of Bayes' rule). As above, objects of FinStoch are natural numbers, morphisms from m to n are $n \times m$ stochastic matrices, composition is matrix product, and the monoidal product is the matrix tensor product. States are probability distributions over the set $1, ..., n$:

$: 1 \to A = (p_1, ..., p_n)$ such that $\sum_{j=1}^{n} p_j = 1$

A *joint state* is a state over the composite object $1, ..., mn$, of the form



A joint state is "uncorrelated"—one should be careful, since this is correlation between probability distributions, not between random variables per se—when it can be decomposed into a tensor product, and perfectly correlated when it can be represented as a delta function. Uncorrelated and perfectly correlated joint states are depicted, respectively:



$: 1 \to A \otimes B = (p_1 q_1, ..., p_n q_m)$

$: I \to A \otimes A = (\delta_{i,i'} \in \{1, ..., n\}.$

A perfectly anti-correlated joint state in FinStoch is just a cup with a NOT-gate attached to one end. More generally, any correlation can be obtained by attaching a suitable box to one of the ends of the cup.

There are two major problems with this graphical formalism, and with Stoch and FinStoch in general. First and foremost, there is not a very convenient way of talking about *random variables*. Technically, a real-valued random variable is given by the diagram below, of a measurable function $X : (\Omega, \Sigma_\Omega) \to (\mathbb{R}, \Sigma_\mathbb{R})$ which takes possible outcomes in Ω to their numerical representations in $\mathbb{R}$ (technically, X is not a function but a stochastic matrix whose columns represent point probabilities), a probability measure $P : 1 \to (\Omega, \Sigma_\Omega)$ on the outcomes in Ω, and finally the pushforward $X(P)$ of P along X, which represents the probability distribution of the random variable X.

$$\begin{array}{ccc} 1 & \xrightarrow{P} & (\Omega, \Sigma_\Omega) \\ & \searrow^{X(P)} & \downarrow X \\ & & (\mathbb{R}, \Sigma_\mathbb{R}) \end{array}$$

Besides being difficult to work with, the point of view of this paper is that probability measures are not random variables nor are they sufficient replacements; a probability measure is something probability theorists invented in order to talk about random variables. While useful in the context of Bayesian networks, which can be articulated primarily in terms of stochastic processes, probability measures are less visible in cases driven by data and by correlational arguments.

The second problem is that there is not a good interpretation of *effect*, i.e. of "measuring" or computing something with respect to a specific state. An effect in FinStoch

is defined to be a morphism $X^\dagger : (A, \Sigma_A) \to 1$, i.e. a function $X^\dagger : A \times \Sigma_* \to [0, 1]$. The problem is that 1 is terminal in Stoch: there is only one possible morphism from any object to 1 due to the constraint on morphisms of being a probability measure. In particular, for any (A, Σ_A) and for all $a \in A$, the unique map $X^\dagger : (A, \Sigma_A) \to 1$ is given by $X^\dagger(*|a) = 1$ and $X^\dagger(\varnothing|a) = 0$. In

other words, any 'measurement' of a state (i.e. a probability distribution) in FinStoch and Stoch simply kills the state.

In Bayesian statistics and machine learning, graphical models have been developed to model the conditional (in)dependence of systems of random variables; the joint distribution over all the random variables in a graphical model is the product of their conditional distributions. Among the most familiar examples of graphical models are (directed) Bayesian networks and (undirected) Markov networks. The upshot is that complex questions about joint distributions of many interrelated variables can be answered in terms of the topology of the graph. Graphical models, Bayesian networks, stochastic matrices ground several existing mathematical approaches to integrating causality with probability, including that of Stoch and FinStoch [**?**]. The study of causality in probability has deeper roots in structural equation modeling and Pearl's work on the do-calculus [10].

3 Method

In this section, we describe how to construct, use, and compare abstract indicator frameworks, at each step giving both the operational and the mathematical specification.

3.1 Constructing indicator frameworks

Here is a step-by-step guide to constructing an abstract indicator framework.

1. Take as input a list of $\mathbb{R}$-valued random variables, representing indicators, and a correlation matrix between them.

2. Construct a finite-dimensional vector space N with the original random variables as basis elements. The inner product between basis elements i, j should match the value of the i, j-th entry in the correlation matrix.

3. Separately, define a scientific model over those random variables. For example, $X \to Y \leftarrow Z$ is a causal model of (causal) variables X, Y, Z.

   (a) Divide the basic components s_i of your scientific model into types or "grammatical roles". For example, a causal model has only two types: causal variables and arrows.

   (b) Define the grammatical rules for how types can be composed. For example, an arrow in a causal model is only grammatical when it appears between two causal variables.

4. Assign each type in your scientific model to a vector space. In a causal model, variable types go to the vector space N and causal arrows go to $N \otimes S \otimes N$, where S is the one-dimensional $\mathbb{R}$-vector space with $\vec{0}$ as origin and $\vec{1}$ as basis element. (In general, S can be any vector space.) For example, the causal model $A \to B$ has type $N \otimes (N \otimes S \otimes N) \otimes N$.

We will see an example of this in Section **??**.

The types in the scientific model correspond to types of sentences. So, really, any shape of causal diagram should be an object in Caus!

The sentence space could be $is_surprising \times is_positive$.

3.2 Using indicator frameworks

Having constructed an abstract indicator framework, F, we can now use it.

1. Construct an embedding σ from the components s_i of your scientific model into the vector space corresponding to the type of that component. In a causal model, σ is fixed by the definition of N, while the causal arrow $\rightarrow$ maps to the vector $\vec{\Psi}$ in $N \otimes S \otimes N$ given by

$$\vec{\Psi} = \sum_{i,j,k} c_{ijk}(n_i \otimes s_j \otimes n_k) \in N \otimes S \otimes N$$

   where $c_{ijk} = 1$ when n_i causes n_k, and 0 otherwise. So we have directly defined the embedding by encoding each component s_i within the basis of its target vector space. More generally, one can apply machine learning methods to a large corpus of data in order to construct such an embedding σ.

2. Using the embedding above, embed each component s_i of your scientific model in order to obtain a "linearized" version of that model, $v_1 \otimes ... \otimes v_m \in V_1 \otimes ... \otimes V_m$.

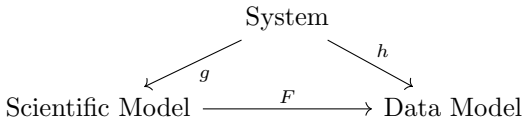
3. Choose a *type reduction* $F(f) : V_1 \otimes ... \otimes V_m \rightarrow U_1 \otimes ... \otimes U_k$. For example, given a causal model $A \rightarrow B$, we can define a reduction $F(f) : N \otimes (N \otimes S \otimes N) \otimes N \rightarrow S$ by $F(f) = \epsilon_N \otimes 1_S \otimes \epsilon_N$ where $\epsilon_N : N \times N \rightarrow \mathbb{R}$ is the inner product map $\langle -|- \rangle$ (a.k.a. the co-unit).

4. Apply $F(f)$ to your vector $v_1 \otimes ... \otimes v_n$. The result, in the reduced vector space, is the "meaning" of the scientific model in the form of a single random variable.

term	semantics
scientific model	an object in an autonomous category
causal model	an object $C \in \mathsf{Caus}$
indicator framework	functor $F : \mathsf{Caus} \rightarrow \mathsf{Rand}$

Table 1: A glossary of basic mechanisms for describing the behavior of a learning algorithm A with hypothesis h on a given (labeled) data set S.

Roughly:



An operational account of indicator frameworks should describe not only the data model but how we *use* that data model. Our contention: we *use* that data model in ways that are approximated by the scientific models inside our heads.

Any meaningful scientific model is composed of parts, each of which has a meaning.

In general, it is not possible to identify a causal model uniquely. But neither is this the problem that we want to solve. Instead: how do we "value" the data of one system with respect to the data of another system? Alternately: how do we compare systems of variables with other systems of variables?

Mathematically: Put into 2-column format.

1. Using indicator data, specify an object A in the data category Rand.
2. This is a rather subtle point. E.g. see `http://math.ucr.edu/home/baez/quantum/node3.html`.
3. Pick an object in your grammar category, Caus.
4. Define the objects of the grammar category of your scientific model, e.g. the category of causal models Caus.
5. Define a strict monoidal functor $F : \mathsf{Caus} \to \mathsf{Rand}$.
6. Define a morphism σ in Rand from the unit object to A.
7. Consider the morphism σ as a generalized element.
8. Choose a morphism f in Caus.
9. Consider $F(f)$.
10. Consider $F(f) \circ \sigma$.

Recall that real-valued, square-integrable random variables over a given probability space form a Hilbert space $L^2(\Omega, \Sigma, \mathbb{P})$ whose inner product is just the covariance.

Definition 3.1. The category of random variables, Rand, is defined by the following data:

1. objects are finite-dimensional Hilbert spaces

$$\mathcal{X} = L^2(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$$

of square-integrable random variables (under the equivalence relation $X_1 \sim X_2$ if $\mathbb{P}_{\mathcal{X}}(X_1 = X_2) = 1$) with inner product $\langle X, Y \rangle = E(XY)$, defined over probability spaces $(\Omega_{\mathcal{X}}, \Sigma_{\mathcal{X}}, \mathbb{P}_{\mathcal{X}})$, with an associated basis $\mathcal{B}_{\mathcal{X}} = \{X_1, X_2, ..., X_n\} \cup \mathbf{1}$, where $\mathbf{1}$ is the random variable with constant value 1.

2. morphisms $F : \mathcal{X} \to \mathcal{Y}$ are bounded linear operators
3. identity $1 : \mathcal{X} \to \mathcal{X}$ is the identity matrix.
4. the composition is the usual composition of bounded linear operators
5. the tensor product of $\mathcal{X}$ and $\mathcal{Y}$ is the pushout over their joint support in $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$, with monoidal unit $\mathbf{1}$.

Lemma 3.2. Rand is a symmetric monoidal category with the above tensor.

We defer the proof of the lemma to the appendix.

The fact that Rand is a symmetric monoidal category is a strong constraint on the semantics of the data model. It also allows us to write down equations in Rand in the language of string diagrams [**?**]. [Placeholder here for statistical interpretation of Rand.]

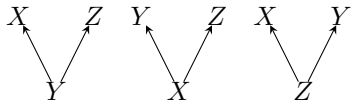
Definition 3.3. An *abstract indicator framework* F is a strong monoidal functor

$$F : \mathsf{Caus} \to \mathsf{Rand}.$$

Example 3.4.

$$F(A \to B \leftarrow C) = N \otimes N \otimes S \otimes N \otimes N \otimes N \otimes S \otimes N \otimes N$$

Rand and the notion of mediating framework supply the basic statistical foundation for a theory of indicator frameworks. We would now like to incorporate a causal foundation. There are several reasons for doing so. First: many statistical arguments, e.g. partial correlation, actually rely on an implicit choice of causal model—see discussions related to confounding and mediating variables by Pearl [10], among others. For example, given three random variables X, Y, Z; the partial correlations $\rho_{XZ \cdot Y}$, $\rho_{XY \cdot Z}$, or $\rho_{YZ \cdot X}$ are all equally valid; which one we take as 'true' depends on which of the following causal structures we believe is true:

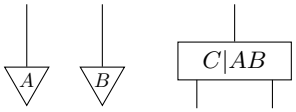


Second, many operational indicator frameworks are constructed based on experts' causal models of the indicators, e.g. as in [9] or as in the air pollution diagram shown at the very beginning of this paper. Any story of indicator frameworks would be incomplete without mentioning causation. And third, the

pattern developed here for causation will be useful later, when we want to incorporate not only causal models but arbitrary scientific models (such as those in Bayesian networks) into our indicator frameworks.

We recall the definition of causal theory from Fong [4], as a certain symmetric monoidal category induced from a directed acyclic graph (i.e. the causal structure), such as any of the three graphs above. Without going into the details, the idea is that given such a causal structure, we can specify a symmetric monoidal category whose objects are collections of the letters $\{X, Y, Z\}$, and whose morphisms are generated by the counit (representing 'deletion') and comultiplication (representing 'copying'), depicted respectively by

and by a set of causal mechanisms generated from the causal structure, $[A] : \varnothing \to A$, $[B] : \varnothing \to B$, and $[C|AB] : AB \to C$, depicted as



Given a causal theory, i.e. a symmetric monoidal category, we can define a model of that causal theory $\mathcal{C}$ in Rand as a strong monoidal functor $F : \mathcal{C} \to \mathsf{Rand}$. To specify such a functor, it suffices to define its behavior on every atomic variable and every generating map in $\mathcal{C}$, i.e. the counit, comultiplication, and causal mechanisms of $\mathcal{C}$, since the values of the functor on the rest of $\mathcal{C}$ is specified up to isomorphism by the definition of a strong monoidal functor. For example, if A is an atomic causal variable of the causal theory $\mathcal{C}$, then F sends A to a one-dimensional Hilbert space, e.g. one with basis set $\{X\}$. On tensor products of atomic causal variables, $F(A \otimes B)$ gives the tensor product $F(A) \otimes F(B)$, i.e. the space of random variables with basis set $\{F(A), F(B)\}$ and probabilities inherited from the product measure. On morphisms, $F([A]) : F(*) \to F(A)$ is just the single random variable in $F(A)$, and a causal mechanism $[C|AB]$ becomes a linear operator $F([C|AB]) : F(A \otimes B) \to F(C)$.

Note that diagrams in the causal theory do not, typically, give rise to diagrams of the same shape in Rand. For example, a confounding variable Y with causal structure $X \leftarrow Y \to Z$ will typically generate the diagram corresponding to $\rho_{XZ \cdot Y}$. In general, each strong monoidal functor $\mathcal{C} \to \mathsf{Rand}$ converts a causal theory into a certain "package" of related indicator sets, where the operational indicator framework is represented by the terminal leaves of the causal theory. Picking the appropriate functor constitutes an optimization problem.

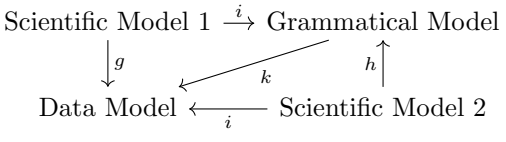
[Move to where?] So we lack data. But to take just one example, in a database setting there are ways to reason about "missing data", e.g. database nulls, especially when that data is the subject of a data migration or integration, as described in [12]. In particular, one can impose a set of algebraic equations that each null value must satisfy, where the equations are given by a diagram of

database schema mappings. More generally, almost every diagram in a category articulates a set of constraints on the objects of that category.

3.3 Comparing indicator frameworks

Now, given two indicator frameworks $F : \mathsf{Caus} \to \mathsf{Rand}$ and $G : \mathsf{Caus} \to \mathsf{Rand}$, we can compare them in the following way:

Roughly:



An analogy

Suppose that we are given a set of n nouns, along with a word embedding of those nouns to a vector space spanning those words $\mathbb{R}^n$. Separately, someone has arranged (some of) those words into a sentence.

Suppose, now, that we have many such sets of nouns, and a sentence for each set of nouns. There may or may not be overlaps between the noun spaces, but we assume that the inner product (i.e. relative distance) between nouns is preserved across different embeddings. We assume that the sentences all belong to the same language (of, if you will, all the models are of the same "type"), though really the only thing that matters is that the functor F is defined on the verbs; i.e. the verbs have meaning.

The key question: how do I best arrange these sentences? Well... I cannot calculate correlations between words coming from different sentences directly; in other words, the embeddings are supposed to agree on their overlaps

We can now state the definition of Ind.

Definition 3.5. The category Ind of abstract indicator frameworks is defined by the following data:

1. an object I of Ind is a strong symmetric monoidal functor $\mathcal{C} \to \mathsf{Rand}$ from a causal theory $\mathcal{C}$ to the category of random variables.

2. a morphism η between abstract indicator frameworks is a natural transformation of strong symmetric monoidal functors

In other words, an object of Ind represents a diagram in Rand, whose nodes are indicator sets and whose edges have been organized to represent the various relationships between indicator sets. One may directly compare Ind with the category of stochastic causal models in [4], which are generalizations of Bayesian networks.

4 Applications

Simple lists of indicators are used widely, and correlations between these indicators are one of the basic tools of elementary statistics. However, we are most interested in cases where the systems under study are large and complex, and thus necessitate the "gluing together" of many such frameworks. Abstract indicator frameworks are designed to make possible the gluing together of such frameworks, even in the absence of data to support that gluing.

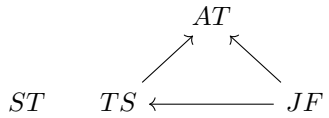
Don't just think single variables: think sets and systems of variables thereof—like a big tensor of these things! Each abstract indicator framework can be thought of as a summary of the local system. We can use them to explore higher-order phenomena like "cooperation" in the local system.

We present an example based on bus data from Nashville, TN [**?**]. Start with a table of correlations, as below.

	ST	AT	JF	TS
ST	1	.81	**.15**	-.04
AT		1	-.05	-.02
JF			1	-.48
TS				1

Table 2: Correlation table for bus data collected in 2016 from Nashville, TN. ST represents the scheduled travel time of the bus, AT represents the actual travel time, JF represents the jam factor (intuitively, the higher, the greater the traffic), and TS represents the average traffic speed.

Why is there a mild positive correlation between the jam factor and the actual travel time of the bus, when we would expect a mild-to-strong *negative* correlation? To resolve this question, the analyst comes up with a causal diagram:



The idea is to the turn both the table and the diagram into morphisms in Rand.

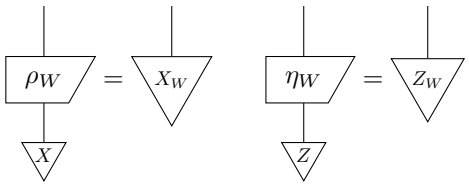
The objects in the causal theory C of the diagram above are { AT, ST, JF, TS, the tensor products, and the unit object **1**. }

The morphisms are need to draw a big list of diagrams.

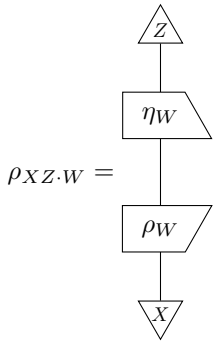
Example 4.1. Suppose that the transportation department buys a new bus and designates an indicator, X, that counts the number of riders on the bus per day. Elsewhere, the education department tracks an indicator, Z, that counts the number of students per day who are absent from class across the whole city. Assume that X and Z live in indicator sets $\mathcal{X}$ and $\mathcal{Z}$.

First, we "integrate" the data by computing $\mathcal{X} \otimes \mathcal{Z}$, so that the correlation is computed only on days for which X, Z both have data. We compute the

correlation: then the correlation may be very small, or conversely it may be absurdly high, especially if there is some confounding variable correlated with both X and Z, e.g. an economic boom. Suppose the federal government tracks a separate variable, W, on aggregate economic performance per quarter. The first step is to get rid of the influence of W, i.e. compute the residuals X_W, Z_W of X, Z resulting from their linear regression with W. Assuming that X and Z live in indicator sets $\mathcal{X}$ and $\mathcal{Z}$ respectively, and that W lives in both $\mathcal{X}$ and $\mathcal{Z}$, we can represent computing the residuals as applying transformations ρ_W, η_W on $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$, respectively:



For variables X, Y, let us denote the linear regression of X with respect to Y by $X|Y$. ρ_W and η_W are indeed morphisms in Rand, i.e. bounded linear operators, since the residual of a linear regression can be written in the form $X_W = X - X|W = X - (a\mathbf{1} + bW)$, where a, b are constants. (Technically, everything above happens in the "larger" Hilbert space $\mathcal{X} \otimes \mathcal{Z}$; ρ_W, η_W are projections from this larger space.) Then by definition, we have



Example 4.2. Recall our earlier example, where X stands for the number of riders on a particular bus in a city, and Z stands for the number of absent students across a city. Suppose that we have already controlled X and Z for economic performance (i.e. W) along with any number of other confounding variables, and that we have found (or suspect) a small but significant correlation between X and Z. We are now interested in understanding *how* X correlates with Z.

There may be a variety of possible explanations for why this correlation exists: maybe dropping the price of a ticket (thus promoting more bus ridership) allows more students to go to school, or perhaps additional bus ridership decreases traffic, which gives harried parents more time to track their truant

children. Without choosing any one explanation, we can represent the statistical properties of a set of mediating, "explanatory" variables $\mathcal{Y}$ by a 'sum' of the possible explanations:

$$\mathrm{Cor}(X, Z) = \sum_{Y \in \mathcal{Y}} \left(\begin{array}{c} Z \\ \eta_Y \\ \mathcal{Y} \\ \rho_Y \\ X \end{array} \right) \tag{1}$$

The equation above succinctly represents a set of constraints that we can impose on the intermediate framework $\mathcal{Y}$, and motivates the following definition.

5 Discussion

In this paper, we sought to give a rigorous mathematical alternative to the traditional, indicator-by-indicator process of constructing indicator frameworks, especially in city planning and project governance. We proposed that indicator frameworks could be defined (and optimized) by means of their relationships to other indicator frameworks. These relationships were probabilistic as well as causal. Therefore, we sought to develop a semantics for the problem of constructing indicator frameworks that could accommodate both probabilistic and causal modes of reasoning.

We then used Ind as the setting for an optimization problem: how to construct a mediating indicator framework that best explains the relationship between a given set of indicators, such as those of a specialized project in a city, and another set of indicators, such as headline indicators of broad interest to the public. Such a mediating framework can be used to answer the question, "what are the secondary impacts of my project?"

We examined several options for the semantics of probability, including Stoch [8], $\mathsf{FinStoch}$, and their corresponding diagrammatic representations [1]. After reflecting on the practical necessities of data analysis, we decided to base our construction on a category more directly in terms of random variables and correlations, and defined the symmetric monoidal category Rand of (spaces of) random variables. We then introduced the idea of a causal model from [4], and used this to motivate the definition of the category Ind of abstract indicator frameworks as models of a causal theory in Rand.

[Given a heterogeneous set of models over a common data model, what, in some sense, is the "total prediction" of that set of models on a single data set? Over many data sets? Given the total prediction, can we then define

a consistent, heterogeneous "total prediction error"? Draw inspiration from Friston 2009 (Trends in Cog Sci).

Further directions:

- Test more complicated sorts of scientific models used in city science, e.g. cyber-physical systems.
- Abstract over the particular scientific model and the particular data model entirely. In other words, discuss the general idea of strong monoidal functors as a mechanism for defining models over data.
- Implement the problem in a real-world example in city administration.
- Study possible obstacles gluings may introduce to producing a consistent global picture of the complex system.

Acknowledgements

We would like to thank Bob Coecke, Santi, Bilin Guvenc, Levent Guvenc, Derek Loftis, and Ed Griffor for helpful conversations in the writing of this paper.

Disclaimer

Certain commercial products may be identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

A Symmetric monoidal categories

Lemma A.1. Rand is a symmetric monoidal category with the above tensor.

Proof. ... □

B Statistical interpretation

Before giving the definition of the category Ind of abstract indicator frameworks, we will go through some of the statistical justification. Suppose that we have a correlation between random variables X and Y and another one between Y and Z. What can we say about the correlation between X and Z? One obvious guess would be

$$\mathrm{Cor}(X, Z) = \mathrm{Cor}(X, Y)\mathrm{Cor}(Y, Z). \tag{2}$$

Of course we know that Equation 2 is, in general, false.[1] But it is, under certain conditions, still the best guess.

The following result is a standard exercise in statistics.

Lemma B.1. If $a = \mathrm{Cor}(X, Y)$ and $b = \mathrm{Cor}(Y, Z)$, then

$$\mathrm{Cor}(X, Z) \geq ab - \sqrt{1 - a^2}\sqrt{1 - b^2} \tag{3}$$

$$\mathrm{Cor}(X, Z)\rangle \leq ab + \sqrt{1 - a^2}\sqrt{1 - b^2} \tag{4}$$

Proof. WLOG, assume that A, B, C are standard variables with zero mean and unit variance, since the correlation is invariant under changes to mean and variance. We can write $X = aY + E_{Y,X}$ and $Z = bY + E_{Y,Z}$ where, by construction, $E_{Y,X}, E_{Y,Z}$ are random variables uncorrelated with Y.

Then $\langle X, Z\rangle = \mathrm{Cor}(X, Z) = \langle aY + E_{Y,X}, bY + E_{Y,Z}\rangle = ab + \langle E_{Y,X}, E_{Y,Z}\rangle$, and we can use the Cauchy-Schwarz inequality to bound $\langle E_{Y,X}, E_{Y,Z}\rangle$ from above and from below, giving the lemma. □

The lemma tells us that there is a range of possible values, centered around $\mathrm{Cor}(X, Y)\mathrm{Cor}(Y, Z)$, for the composite correlation; unfortunately, in practice that range can be so large as to be useless. In such a situation, we may ask what is the obstruction, given $\mathrm{Cor}(X, Y)$ and $\mathrm{Cor}(Y, Z)$, to knowing the canonical or 'true' correlation of their composite, and whether we can reduce or get around that obstruction. Reading the proof of the lemma, we know the obstruction is just the correlation $\langle E_{Y,X}, E_{Y,Z}\rangle$; that is, if $\langle E_{Y,X}, E_{Y,Z}\rangle$ were 0, our guess would be valid.

One may also derive the above result from the definition of the partial correlation of X and Z, fixing Y. Recall that the partial correlation $\rho_{XZ\cdot Y}$ is defined as the correlation between the residuals of X and of Z, fixing Y. In terms of their component correlations,

$$\rho_{XZ\cdot Y} = \frac{\mathrm{Cor}(X, Z) - \mathrm{Cor}(X, Y)\mathrm{Cor}(Y, Z)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}\sqrt{1 - \mathrm{Cor}(Y, Z)^2}}.$$

Thus

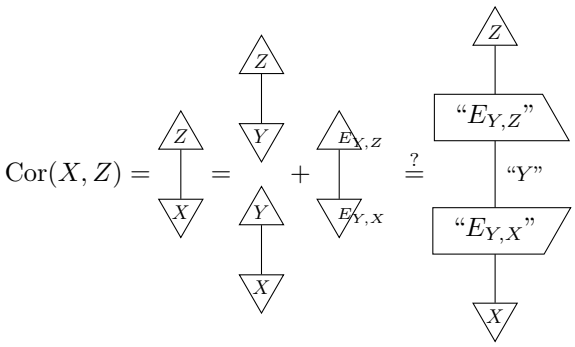
$$\langle E_{Y,X}, E_{Y,Z}\rangle = \rho_{XZ\cdot Y}\sqrt{1 - \mathrm{Cor}(X, Y)^2}\sqrt{1 - \mathrm{Cor}(Y, Z)^2}.$$

In other words, Equation 2 is correct just when the partial correlation $\rho_{XZ\cdot Y} = 0$, when $\mathrm{Cor}(X, Y) = 1$ or -1 (i.e. X and Y are linear functions of each other), or when $\mathrm{Cor}(Y, Z) = 1$ or -1. This allows us to produce another guess:

$$\text{“Cor}(X, Z)\text{”} = Y \text{ s.t. } \rho_{XZ\cdot Y} = 0 \tag{5}$$

[1] Correlations are rarely composed in practice because (1) the result is usually false and (2) because we can usually compute the composite correlation directly from the data. It is only when we lack the data that we use a causal model to infer the correlation. Unfortunately, causal models are often invoked in the process of imposing a learned model such as a Kalman filter or a dynamical Bayesian network, which will often conflate the statistical and causal contributions.

Explicitly, $E_{Y,X}$ measures the nonlinear component of the relation between X and Y. But one may also think of it as a measure of the 'noise' or 'error' between X and Y, at least as it concerns the correlation. The idea of Equation 5 is that, if we are lucky in choosing Y, then the noise factors $E_{Y,X}$ and $E_{Y,Z}$ will "cancel out" to produce the true correlation $\mathrm{Cor}(X,Z) = \mathrm{Cor}(X,Y)\mathrm{Cor}(Y,Z)$. Heuristically, we can represent this process as below:



That is, the correlation between X and Z can be computed by applying a transformation "$E_{Y,X}$", representing some sort of structured noise factor, then applying a transformation "$E_{Y,Z}$" that cancels out the noise introduced by "$E_{Y,X}$".

As shorthand, we will sometimes refer to the space of random variables $\mathcal{X}$ as a set of variables or indicators; in such cases, we always mean the basis set of random variables, $B_{\mathcal{X}}$.

It will help to think of random variables as representing column vectors or "dimensions" of data in a table of such data, where row vectors in that table represent particular data points. The correlation between two column vectors is just their sample correlation. This has several benefits: it makes the inner product (correlation) and tensor product (entity resolution) very concrete, it is what a data analyst actually looks at, and it highlights the restrictions and challenges imposed by the presence and absence of data. In fact, we can define a category Data explicitly in such terms:

Definition B.2. The category of $\mathbb{R}$-valued data tables, Data, is defined by the following data:

1. objects $\mathcal{X} = (\mathcal{X}, \Omega_{\mathcal{X}}, \mathbb{I}_{\mathcal{X}})$ of Data are $m \times n$ tables of $\mathbb{R}$-valued data vectors whose rows are assigned an index key given by $\mathbb{I}_{\mathcal{X}} : \Omega_{\mathcal{X}} \to \mathbb{R}$ and whose columns, $B_{\mathcal{X}} = \{X_1, ..., X_n\}$, represent indicators
2. morphisms $f : \mathcal{X} \to \mathcal{Y}$ are linear transformations of the column values of $\mathcal{X}$ by vector addition (of other columns in $\mathcal{X}$) and scalar multiplication
3. the identity $1 : \mathcal{X} \to \mathcal{X}$ is the identity matrix
4. the composition is just the matrix product
5. the tensor product of $\mathcal{X} \otimes \mathcal{Y}$ is the integrated table of their data values over a table of linkages, $S \subset \Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$

Suppose we are working in a 3-dimensional Hilbert space $\mathcal{X}$ with a basis of random variables $B_{\mathcal{X}} = \{X, Y, Z\}$. In this basis, the random variable $E_{Y,X}$ is just the vector $X - \langle X, Y \rangle Y$ (and similarly with $E_{Z,X}$), but the problem with this space... is that there is no problem! In Rand, having a basis in X, Y, Z corresponds to the observation, in Data, that we already have the tabular data we need to compute $\langle X, Z \rangle$ directly. But in many situations of interest in a complex, open system, e.g. in computing the second-order impacts of local and/or technical projects, such broad-based data is difficult to obtain.

[REWORK? Introduce adjustment problem / Simpson's paradox first, and then define a confounding framework as "the set of confounding variables you should measure and marginalize over?"

Definition B.3. A *mediating framework* between two spaces of random variables $\mathcal{X}, \mathcal{Z}$ is a space of random variables $\mathcal{Y}$ such that Equation 1 is satisfied for all variables $X \in \mathcal{X}, Z \in \mathcal{Z}$.

References

[1] B. Coecke and R. Spekkens. Picturing classical and quantum bayesian inference. *Synthese*, 186(3):651–696, June 2012.

[2] G. T. Doran. *Management review*, 70(11):35–36, 1981.

[3] M. J. Epstein and J.-F. Manzoni. The balanced scorecard and tableau de bord: translating strategy into action. *Strategic Finance*, 79(2):28, 1997.

[4] B. Fong. Causal theories: A categorical perspective on bayesian networks. Preprint, April 2013.

[5] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, volume 915 of *Lecture Notes in Mathematics*, pages 68–85. Springer-Verlag, 1982.

[6] S. Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, 2011.

[7] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[8] F. W. Lawvere. The category of probabilistic mappings. Unpublished seminar notes.

[9] D. Niemeijer and R. S. de Groot. A conceptual framework for selecting environmental indicator sets. *Ecological Indicators*, 8:14–25, 2008.

[10] J. Pearl. *Causality*. Cambridge University Press, 2009.

[11] P. Selinger. A survey of graphical languages for monoidal categories. Preprint, August 2009.

[12] D. I. Spivak. Functorial Data Migration. *ArXiv e-prints*, Sept. 2010.