
Rethinking Statistics and Causality: Why Mechanisms Cannot Be Inferred from Data Distributions

Egil Diau

Department of Computer Science
National Taiwan University
Taiwan, Taipei
egil158@gmail.com

Abstract

Statistical and causal inference have become universal currencies of explanation across the sciences, particularly in domains where underlying mechanisms remain opaque. Their apparent rigor—spanning psychology, economics and biomedicine—rests on the assumption that patterns within data can reveal the processes that generate them. Yet persistent mismatches between empirical predictions and real-world behaviour expose a deeper limitation: mechanisms cannot be inferred from data distributions alone. To address this limitation, we revisit the foundations of both paradigms, showing how statistical inference reduces explanation to geometric alignment, while causal inference, evolved from Bayes’ theorem and graphical models, extends this misstep by conflating probabilistic structure with causal truth. Both expose the same epistemic gap: data encode a lower-dimensional projection of structure, not the mechanism that generates it. We argue that understanding the world follows two routes: one is data-driven, expanding models toward richer function classes to achieve high-precision prediction, as exemplified by modern deep learning; the other is mechanism-driven, proposing and testing structural hypotheses as in the physical sciences. A robust framework requires both: data-driven models for high-precision prediction, and mechanistic models for reconstructing how the world produces the data we observe.

1 Introduction

Across many scientific domains, statistical and causal inference have become default tools for turning complex observations into interpretable results. Psychology Collaboration [2015], economics Camerer et al. [2016] and biomedicine Ioannidis [2005], in particular, depend on these methods whenever direct access to underlying mechanisms is out of reach. Over time, this dependence has given rise to an illusion of methodological certainty: as long as the analysis is significant, the claim appears sound. Yet the growing number of failed replications—from behavioural studies to clinical trials—suggests that what these methods capture is often a surface regularity rather than a genuine principle of how the world works.

Most researchers acknowledge that correlation does not imply causation, yet the same procedures remain deeply entrenched. The reason is pragmatic rather than epistemic. When mechanisms cannot be measured and sample sizes are limited, statistics offers a way to project uncertainty into mathematical form. Patterns that would otherwise look noisy can be fitted into curves, slopes or conditional dependencies, each assigned a numeric level of confidence. In this sense, statistics does not uncover meaning—it produces geometric regularity in places where the underlying structure remains unknown.

From its very beginning, statistical reasoning has relied on symbols that conceal its geometric nature. Correlation is the cosine of an angle, variance the length of a vector, regression the alignment of a



Figure 1: **Data as lower-dimensional projections of structure.** A single underlying world model gives rise to multiple observational projections, each encoding a different low-dimensional representation of the same structure. Once projected and compressed, the original properties become irrecoverable without prior knowledge of the generating process—illustrating why statistical and causal inference cannot recover mechanisms from data distributions alone.

plane Rencher and Schaalje [2008]. These are operations on geometry, not mechanisms. Modern causal inference inherits the same logic, with an added layer of semantic drift Pearl [2009]. Bayes’ theorem—originally a measure-an identity built from the formal definition of conditioning—was later reinterpreted as a calculus of belief through priors and posteriors, and eventually recast in graphical form as if conditional dependence entailed causal direction. Entire causal frameworks were built on this conflation. What unites both traditions is the same epistemic limit: data capture only the projection of a structure, never the process that generates it.

Understanding the world, therefore, follows two complementary yet incomplete paths. One—mirroring modern machine learning—accumulates ever more parameters and basis functions to perfect prediction, yet remains blind to the mechanisms that generate the data Hornik et al. [1989], LeCun et al. [2015]. The other follows the logic of the physical sciences: starting from structured prior knowledge, it proposes candidate mechanisms and tests whether they can produce the observed phenomena. But this approach falters whenever the underlying structure is too complex or only partially observable Hempel and Oppenheim [1948]. A genuine theory of inference must unify these two modes: models that predict reliably from data, and mechanisms that explain how the phenomena are generated.

Our contribution. This work re-examines the foundations of statistical and causal inference. We reveal how their core constructs—correlation, significance, regression, and conditional dependence—originated as geometric or probabilistic operations but were later misinterpreted as explanations of mechanism. Building on this diagnosis, we outline two complementary yet incomplete paradigms for understanding the world: one that, like deep learning, expands toward infinite bases for prediction while remaining mechanistically blind, and another that generates systems from prior semantic bases but falters when those priors fail to capture complexity. Finally, we explain why such methods persist across psychology, economics, and biomedicine: they offer numerical certainty where mechanistic understanding is absent. Together, these arguments recast inference as a problem of meaning rather than measurement.

2 The Root of the Error: Projection, Sampling, and Estimation

Modern inference traditions emerged from two pressures:

- **Generative mechanisms were not specified.** Many fields operated without explicit accounts of how observations are produced. The underlying processes remained undefined, conceptually diffuse, or absent from the modeling framework.
- **Conclusions were drawn from minimal observations.** Researchers routinely attempted to make claims about complex systems using small, sparse or convenience datasets. Limited data were treated as sufficient to characterize the structure of the world.

These pressures encouraged a substitution: data were treated not merely as evidence *about* the world, but as stand-ins *for* the world’s generative structure. Samples were assumed to come from an underlying distribution; the distribution was assumed to reflect stable properties of the world; and stability was taken to imply mechanism.

Once this substitution was accepted, inference inverted the logic of scientific explanation. Rather than defining how the world produces observations, researchers attempted to extract generative structure from the observations themselves. The projection, sampling, estimation and semantic fallacies that follow arise directly from this initial move: asking patterns in the observation space to substitute for mechanisms that were never specified.

2.1 The Projection Fallacy: When Observations Are Mistaken for the World

For more than a century, statistical inference has treated data as if they were direct observations of the world. In practice, data are not observations but *projections*—compressed, transformed, and lossy encodings of far richer semantic states. A distribution therefore captures only the geometry of these encodings, not the mechanisms that generate them.

Formally, statistical analysis operates on a reduced observation space:

$$x = f(s), \quad x \in \mathbb{R}^n, \quad s \in \mathcal{S},$$

where the mapping f collapses high-dimensional semantic structure into a finite set of measurable quantities. The original state s contains the generative variables—interactions, processes, transformations—that explain *why* events occur. After projection, this semantic layer is irretrievably lost. Neither $P(x)$ nor $P(x, y)$ encodes the causal grammar that produced them.

Once meaning has been removed, the resulting geometry can be endlessly rearranged. Comparing distributions becomes an exercise in *alignment without understanding*: curves can be made to match, variables can be regressed, and patterns can be interpreted as structure, even when no mechanism survives the projection that created them.

Data therefore cannot be read as windows onto the world. They are shadows cast by a generative process, and the shape of a shadow cannot reveal the object that made it. Mechanism does not reside in the distribution; it must be specified in the semantic space where generation occurs.

2.2 The Sampling Fallacy: When Distribution Is Mistaken for World

Statistical and machine-learning theory typically assumes that observed data are “samples” drawn from an underlying probability distribution. This assumption contains two distinct errors.

The first error is semantic. The generative world does not operate as a probability distribution; it operates as a mechanism. States in the semantic space \mathcal{S} evolve according to physical, biological, cognitive, or social processes—not according to draws from a mathematical object. There is no world-defined distribution from which the true states are sampled.

The second error is geometric. Even if such a distribution existed, the analyst never observes it. What is recorded is a compressed projection:

$$x = f(s), \quad x \in \mathbb{R}^n, \quad s \in \mathcal{S},$$

where the map f is many-to-one, lossy, and often opaque. The projected data x do not themselves form a mathematical distribution in any principled sense; they merely trace the geometry induced by f . Nothing guarantees that the resulting dataset conforms to any parametric or nonparametric family assumed by statistical theory—not Gaussian, not exponential family, not i.i.d., not stationary, and not smooth.

Thus the sampling assumption commits a double fallacy. What is treated as sampling is, in fact, the output of a mechanism; and what is treated as a distribution in \mathbb{R}^n is, in fact, the image of a projection that destroys the very semantics needed to justify distributional reasoning.

Yet much of statistical and machine-learning theory proceeds as if both assumptions were correct. Learning bounds rely on i.i.d. sampling from a fixed distribution; Bayesian inference relies on a likelihood that corresponds to the true generative model; asymptotic theory presumes that larger “samples” converge to an underlying truth. None of these guarantees survive the recognition that data are projections rather than samples.

A projected dataset cannot reveal the distribution of the world because neither the world nor the projection produces such a distribution. Once semantics have been stripped away, no mathematical refinement—no resampling, no bootstrapping, no asymptotics, no density estimation—can recover what was never present.

The real inferential problem is therefore not to estimate a “true” distribution behind the data. It is the inverse problem: given a projected dataset and a set of semantic priors, *how can one reconstruct the generative structure that produced it?* This is a problem of mechanism recovery, not density estimation. And any such recovery requires priors about the projection itself—about how semantic structure was compressed, erased, or entangled on its way into \mathbb{R}^n . Without priors on the projection, the inverse problem is formally underdetermined: no amount of mathematics can reconstruct generative structure from geometry alone.

2.3 The Estimation Fallacy: When Precision Is Mistaken for Understanding

Statistical and machine-learning methods often seek ever more precise estimates—lower variance, tighter confidence intervals, smaller error bars. But precision in the projected space does not imply insight into the generative world. One can estimate the wrong object with arbitrary accuracy.

Formally, estimation procedures optimize with respect to the geometry of the projected data $x = f(s)$, not the mechanism that produced s . Even perfect estimation of $P(x)$, its parameters, or its latent factors does not recover the erased semantic variables or the transformations that govern them. An estimator can converge, asymptotically or exactly, to a value that has no interpretation in the semantic domain.

This is the estimation fallacy: the belief that reduced uncertainty in the data space reflects reduced uncertainty about the world. In reality, tighter estimates merely refine the geometry induced by the projection f ; they do not reconstruct the mechanism g that generated the states. Precision in the space of shadows does not illuminate the object that cast them.

2.4 The Semantic Fallacy: When Meaning Disappears Under Projection

The deepest error in modern inference is not geometric but semantic. Once the generative world is compressed into an observation space \mathbb{R}^n , the semantic content of the original states—their roles, relations, and causal functions—no longer exists within the data. Semantics are erased by projection, yet are reintroduced by interpretation.

The projected variables x do not correspond to the entities that produced them; they correspond only to the coordinates assigned after compression. Distances, similarities, and conditional relations among the x do not preserve the meaning that governed the transformations of the underlying states $s \in \mathcal{S}$. But statistical methods routinely treat these geometric relations as if they reflected the logic of the generative process.

This is the semantic fallacy: the belief that patterns in the projected space retain the meanings that existed before projection. Correlations are interpreted as relationships, regressions as influences, likelihoods as mechanisms, and posterior distributions as knowledge—even though none of these quantities contain the semantic properties they are taken to represent.

Formally, a projection $f : \mathcal{S} \rightarrow \mathbb{R}^n$ is not a semantic map but an information-destroying transformation. Different semantic states collapse to the same observation; distinct processes produce indistinguishable geometries; and meaningful distinctions in \mathcal{S} become conflated in \mathbb{R}^n . No statistical procedure operating solely in the projected space can recover the roles, intentions, or causal capacities that structured the original system.

Yet much of scientific interpretation assumes the opposite. Machine-learning representations are treated as concepts; statistical contrasts as explanations; graphical factorizations as mechanisms. Semantics are not contained in the data—they are supplied by the analyst and mistaken for properties of the distribution.

3 The Illusion of Inference: When Statistics Confuse Geometry with Mechanism

The confusion between geometry and meaning did not arise by accident—it was institutionalized. Across the twentieth century, each reform in statistical reasoning sought to formalize uncertainty and discipline scientific judgment. Fisher quantified experimental claims Fisher [1970], Neyman codified decision rules Neyman and Pearson [1933], and Pearl expressed causality in graphical syntax Pearl [2009]. Each advance increased mathematical rigor but displaced questions of *how* with questions of *how much*. Inference became a substitute for understanding.

As experimental inquiry gave way to modeling, explanation was redefined as prediction, and mechanism as estimation. The tools grew sharper, but their targets became thinner. By the time data-driven inference matured, the very notion of “knowing why” had been replaced by “fitting what.”

Today, the legacy of this history persists: significance, likelihood, and model fit are often mistaken for understanding. But inference, however precise, only maps the geometry of what has been observed—it cannot reveal the process that made it so. To trace how this drift became normalized, we now revisit three canonical paradigms of statistical inference—correlation, significance, and regression—each a historical attempt to extract mechanism from geometry.

3.1 The Geometry of Correlation: What Pearson Actually Measured

Long before computers and data science, researchers sought a numerical way to capture *how things might be related*. In the late nineteenth century, scientists in biology and social science were collecting measurements—height, intelligence, income, behavior—without access to the mechanisms that produced them. When causation was unobservable, they turned to co-variation: if two quantities rose and fell together, perhaps they were connected.

KARL PEARSON formalized this intuition by defining a single measure of linear association between two observed variables Pearson [1896]. The resulting coefficient, later known as the Pearson correlation, quantifies how strongly two datasets align in shape rather than in cause:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|x - \bar{x}\| \|y - \bar{y}\|} = \cos(\theta_{xy}).$$

Formally, this is a normalized inner product between two centered data vectors. Geometrically, r_{xy} equals the cosine of the angle θ between $(x - \bar{x})$ and $(y - \bar{y})$ in a Euclidean data space. Variance represents vector length, covariance their dot product, and correlation the cosine of their relative orientation.

Table 1: Geometric interpretation of the Pearson correlation.

	Statistical expression	Geometric meaning
Variance (σ_x^2)	$\frac{1}{n} \sum_i (x_i - \bar{x})^2$	Squared vector length of x
Covariance ($\text{cov}(x, y)$)	$\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$	Inner product between x and y
Correlation (r_{xy})	$\frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$	Cosine of angle between x and y

This geometric framing is elegant yet deceptively simple. Real-world data rarely inhabit a flat Euclidean plane: they lie on high-dimensional, curved manifolds. Projecting such complexity into a single pairwise angle inevitably discards structural information and exaggerates apparent relationships.

Such large angular separations expose how little alignment typical empirical correlations imply. In psychology, mean reported correlations near $r = 0.3$ correspond to vectors separated by roughly

Table 2: **Approximate angular separation for common correlation values.**

r_{xy}	Cosine interpretation	Angle θ_{xy} (°)
1.0	Perfect alignment	0
0.9	Very strong	26
0.8	Strong	37
0.7	Moderate–strong	46
0.5	Moderate	60
0.3	Weak	73
0.2	Very weak	79
0.0	None	90

73°—barely pointing in the same quadrant. Economics fares no better: correlations that appear robust in-sample often collapse out-of-sample, underscoring that correlation captures orientation, not mechanism.

To appreciate this semantic limitation, consider its modern analogue—*cosine similarity* in embedding spaces Mikolov et al. [2013]:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}.$$

Mathematically identical, yet conceptually distinct. Cosine similarity compares two representations within a shared latent manifold, where dimensions carry learned meaning. Pearson correlation, by contrast, relates two distinct variables extracted through independent measurement pipelines; their coordinate systems are not semantically aligned. It is, in effect, the angle between two unrelated shadows—formally defined, yet devoid of shared reference.

Even within embedding spaces, cosine similarity can mislead: it often retrieves items sharing superficial geometric or frequency patterns rather than genuine semantic relations. If similarity fails even in co-trained spaces, correlation between semantically isolated variables is exponentially more fragile. It reproduces the geometry of relation, but not the meaning behind it.

Correlation preserves shape, not meaning—and in doing so, it mistakes orientation for explanation.

3.2 From Significance to Ritual: The p-Value Illusion

Before mechanistic models were available, researchers needed a numerical rule to decide whether an observed pattern looked unusual under a simple baseline. Fisher’s *p*-value supplied such a rule—a geometric index of deviation, never a statement about underlying process Fisher [1970].

Formally, a *p*-value is a tail probability in the space of a summary statistic:

$$p = P(T(X) \geq t_{\text{obs}} \mid H_0).$$

This quantity lives entirely inside a one-dimensional projection. It reflects the geometry induced by the null model, not the structure of the generative system that produced the data.

The central illusion is straightforward: a low-density region in this one-dimensional geometry is treated as evidence about a high-dimensional world. The tail area is a property of a mathematical construction, yet it is routinely interpreted as a property of the mechanism that produced the observations.

A p-value does not test a mechanism; it tests the position of a statistic in a reference curve.

Thresholding converts this geometric position into a semantic claim—“real,” “significant,” “true.” These categories derive from the assumptions of the null distribution, not from the generative structure of the world.

The fragility of significance follows directly from this mismatch. A one-dimensional projection cannot reveal the structure of a generative system, regardless of the threshold applied. Signals in the projected space cannot recover the mechanisms erased by the projection itself.

Significance is a one-dimensional shadow interpreted as a generative explanation.

3.3 Regression as Projection: The Mirage of Explanation

Regression began as a purely geometric tool. In Galton’s original formulation, the fitted line was simply the orthogonal projection of one centered data vector onto another—a descriptive summary of co-variation, not a statement about generative influence Galton [1886], Rencher and Schaalje [2008].

Over the twentieth century, this geometric construct was gradually recast as a mechanistic one. The equation

$$y = \beta x + \varepsilon$$

was no longer read as a projection in data space but as a causal pathway in the world. The coefficient β , which is mathematically defined as

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|x - \bar{x}\|^2},$$

became interpreted as the “effect” of x on y . Yet this quantity is merely the optimal scalar in a least-squares projection: it measures the angle and relative lengths of two vectors after centering. Nothing in its definition refers to mechanisms, transformations, or influence.

Formally, the fitted value $\hat{y} = \beta x$ is the orthogonal projection of y onto the span of x . Geometrically, this means that regression finds the point in a one-dimensional subspace that minimizes Euclidean reconstruction error. But reconstruction in data space is not explanation in semantic space. A projection reveals how well one shadow aligns with another—not how the underlying objects interact.

The distinction is not cosmetic but fundamental. Regression coefficients are invariant under rotations, rescalings, and reparameterizations of the projected space \mathbb{R}^n . Mechanistic relations are not. Two distinct generative processes can produce identical regression estimates, and identical generative processes can yield wildly different coefficients depending on the projection f that maps semantic states $s \in \mathcal{S}$ into observed values $x = f(s)$. Thus the geometry of fit contains neither the direction nor the logic of the underlying causal transformation.

A regression coefficient is the angle between two shadows—not the force that moves the objects casting them.

This slippage from geometry to meaning underlies the widespread misuse of regression as explanation. In psychology, regression coefficients are treated as cognitive parameters; in economics, as behavioral laws. In each case, models fit historical data with impressive precision yet fail to generalize, because precision in the projected space does not translate into truth about the generative world.

Regression offers precision without understanding: geometry mistaken for mechanism, alignment mistaken for influence.

4 The Illusion of Causality: When Models Confuse Syntax with Semantics

The geometric barrier outlined above would seem to make one conclusion inevitable: if data distributions cannot contain generative structure, then no algebra defined on those distributions can produce it. Yet much of modern statistics, Bayesian reasoning, and causal inference was built on precisely the opposite hope.

Faced with the impossibility of recovering mechanism from projection, researchers redirected the problem into the symbolic domain—attempting to read semantics out of syntax, treating probabilistic expressions as if they could encode information, relevance, belief, or causal influence. This was not a later misapplication but a foundational assumption: that manipulating expressions within a probability space could reveal the processes that gave rise to it.

Conditioning was formalized as a ratio of measures, yet was implicitly treated as if it expressed knowledge or relevance Kolmogorov [2018]. Bayes’ identity was an algebraic equality, yet was reframed as a rule of learning or belief revision de Laplace [1820]. Graphical factorizations were notational devices for decomposing joint distributions, yet were interpreted as diagrams of mechanism.

And modern causal inference built an entire philosophy on the premise that statistical regularities can recover generative structure—an assumption that is false in principle Pearl [2009].

These were not neutral mathematical tools later assigned the wrong meaning. Their meanings were imported from the start through linguistic metaphors, intuitive readings of notation, and a natural tendency to treat grammar as semantics. The result is a family of formalisms that appear to explain how systems work, even though their mathematics never contained any account of mechanism.

To clarify this structural error, we examine conditioning, Bayes’ theorem, graphical models, and causal inference in their original formulation. In each case, the problem is not misuse but ontology: syntax was taken for semantics, and algebra was taken for transformation. This confusion forms the foundation of the modern causal illusion.

4.1 Conditioning: A Linguistic Metaphor Mistaken for a Mathematical Operation

Conditioning is often treated as a semantically meaningful transformation, as if the notation “ $A \mid B$ ” described a shift of context, information, or assumption. Formally, however, conditioning in probability theory is defined only as a ratio of measures Kolmogorov [2018]:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

This expression performs a purely algebraic task: it normalizes the measure of the intersection $A \cap B$ by the total measure of B , producing a dimensionless scalar. The operator “ \mid ” therefore encodes no semantic content; it introduces neither information restriction nor belief refinement. It merely rescales a subset of measure space.

A deeper problem is rarely acknowledged: *probability multiplication can acquire set-theoretic meaning, but probability division cannot*. When two events are independent, the probability of their joint occurrence factorizes:

$$P(A \cap B) = P(A) P(B) \quad (\text{iff independence}).$$

In this special case, scalar multiplication corresponds to a legitimate event-level operation: the intersection of A and B . Outside such structural assumptions, however, the product $P(A)P(B)$ is merely the product of two scalars, with no guaranteed interpretation in the underlying sample space.

In contrast, probability division has no event-level interpretation. The ratio

$$\frac{P(A \cap B)}{P(B)}$$

does not denote an event, a transformation of the sample space, or any operation in measure theory; it merely rescales one scalar by another. Nothing in the mathematics of division guarantees closure within the unit interval—indeed, division can yield values exceeding 1, a clear indication that no underlying event could correspond to the operation. Thus conditioning is not a lawful transformation on events but a notational convention that disguises a purely algebraic normalization.

The confusion persists because the notation “ $A \mid B$ ” mimics the linguistic phrase “given B ,” encouraging readers to treat an algebraic ratio as if it encoded information restriction, relevance, or causal constraint. None of these appear in the definition. Conditioning performs a geometric rescaling in measure space; the semantics are imported by the analyst, not supplied by the mathematics.

4.2 Bayes’ Theorem: An Algebraic Identity Misread as Epistemology

Bayes’ theorem is often introduced as the mathematical foundation of inference, learning, and belief updating. Yet the theorem itself possesses none of these properties. It is nothing more than a symbolic rearrangement of the definition of conditioning:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}.$$

This identity introduces no new information, mechanism, or process. It merely restates a relationship among numerical quantities already defined.

Conditioning, however, is not a lawful operation on events. It is only a ratio of scalars with no set-theoretic or mechanistic meaning. Bayes' theorem, being an algebraic rearrangement of this ratio, cannot acquire epistemic content from an operation that has none.

The first semantic reinterpretation came from statistics. Early Bayesian analyses renamed the three numerical terms in the algebraic identity underlying Bayes' theorem as "prior," "likelihood," and "posterior." A static equality was presented as an updating procedure, even though nothing in the identity specifies change, learning, or cognition. This interpretation did not arise from the theorem itself. It was created by assigning epistemic roles to the three positions of an algebraic relationship that is, in essence, equivalent to the structure $a \times b = c$. The meaning resides in the naming convention, not in the mathematics.

Later, cognitive science adopted this statistical metaphor and elevated it into a model of reasoning. Hypotheses and evidence replaced sets and events; beliefs and updates replaced ratios and products. But again, none of these concepts appear in the theorem. They are cognitive narratives layered onto an algebraic identity that is indifferent to interpretation.

Bayes' theorem never contained epistemic meaning. It appeared to gain meaning only because successive communities—first statistics, then cognitive science—misread algebraic syntax as a model of reasoning.

4.3 From Conditional Distribution to Directed Graph: The Syntax Illusion

Once conditioning is recognized as an algebraic ratio rather than a lawful operation on events, the status of graphical models becomes clearer. Bayesian networks were built on this ratio by treating conditional terms as meaningful structural components. Formally, they provide a compact syntactic factorization of a joint distribution Pearl [2009]:

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i \mid \text{Parents}(X_i)).$$

This factorization expresses nothing about mechanism, generation, or direction. It is simply one of many ways to rewrite a joint probability as products of scalar ratios.

The illusion arises when this syntactic decomposition is mistaken for structure. A directed edge—introduced as a bookkeeping device—was reinterpreted as a causal arrow, even though conditional terms contain no directional or generative semantics. The expressions

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}, \quad P(Y \mid X) = \frac{P(X \cap Y)}{P(X)},$$

are algebraic normalizations, not statements about influence. Nothing in the distribution distinguishes $X \rightarrow Y$ from $Y \rightarrow X$, and no factorization order corresponds to a mechanism.

Graphical models therefore encode only syntactic relationships among algebraic ratios. They describe permissible rearrangements of a joint distribution, not the processes that produce it. Interpreting their arrows as mechanisms extends the same category error that treats conditioning as meaningful. It is the syntax illusion: the belief that grammatical form can reveal generative process.

4.4 When Syntax Becomes Philosophy: The Expansion of the Causal Illusion

Once conditioning is understood not as a meaningful operation but as a ratio of scalars, and Bayes' theorem as a restatement of that ratio, the status of causal inference becomes clearer. What later appeared as a philosophical framework began as a sequence of syntactic devices—none of which carried semantics. Directed graphs, conditional terms, and Bayesian factorizations encode permissible algebraic rearrangements, not mechanisms, dependencies, or processes.

Causal inference marks the point where syntactic constructs were mistaken for claims about the world. Algebraic devices with no semantic content were reinterpreted as descriptions of mechanism, giving rise to the belief that causality can be recovered or manipulated through operations defined entirely within a probability space.

This rests on a categorical mistake. Observational data are projections that preserve patterns while discarding generative structure. Once a process is collapsed into data space, no algebraic manipu-

lation—conditioning, factorization, or intervention—can reconstruct the mechanism that produced it.

Yet modern causal inference is built precisely on the hope that it does. It inherits the initial confusion: that conditional relations can stand in for causal ones, and that rewriting those relations yields insight into interventions. Correlation and causation became separated only by notation, not by ontology.

The field’s history reflects successive attempts to strengthen this substitution. Pearl’s graphical models and *do*-calculus Pearl [2009] treat intervention as algebraic reassignment of conditionals. Rubin’s potential-outcome model Rubin [1974] replaces processes with contrasts between unobservable counterfactual states. Econometric traditions—from Haavelmo to Angrist and Imbens Haavelmo [1943], Angrist and Imbens [1995]—simulate exogenous variation through instruments and natural experiments. Each framework adds formal apparatus, but none touches generative mechanism. All operate solely on observational quantities, performing algebraic rearrangements that cannot, even in principle, recover the processes that produce the data.

Even recent machine-learning approaches, which embed causal reasoning within pattern recognition, inherit the same limitation. By treating structure in data as structure in reality, they reproduce the fallacy at scale. The result is a closed epistemic loop in which causality is defined by syntax rather than semantics, and computation substitutes for understanding.

5 Regrounding Scientific Inference: From Correlation to Mechanism

Scientific inference follows two fundamentally different logics. Both seek structure in data, yet they diverge on what “structure” means.

- **1. Data-driven prediction.**

Mechanism. The data-driven view treats the world as a mapping problem:

$$y \approx f(x) = \sum_i w_i \phi_i(x).$$

By enlarging the feature space—adding more basis functions ϕ_i —any observed relation can, in principle, be interpolated Hornik et al. [1989]. Deep learning LeCun et al. [2015] pushes this logic to its limit: high-capacity models construct extremely expressive representations that reproduce outcomes across diverse contexts. The model succeeds by matching the geometry of the projected data, not the process that produced it.

Limitation. Once the basis is sufficiently large, infinitely many distinct functions fit the same observations. Because nothing in the mapping identifies which function corresponds to the underlying process, prediction remains decoupled from explanation. Data-driven models can match every observed case, but they cannot recover—or prefer—the mechanism that generates the world.

- **2. Mechanism-driven explanation.**

Mechanism. The mechanistic view begins from explicit assumptions about how observations are produced. Mechanisms are first *hypothesized* from prior knowledge and empirical regularities, then tested by deriving the consequences they should generate. This is the epistemology of the physical sciences Hempel and Oppenheim [1948]: mechanisms are proposed, used to derive predictions, and accepted only when experiment confirms that they can generate the phenomenon.

Formally:

$$m \sim \mathcal{M}_{\text{prior}}, \quad x = g(m).$$

Once a component of the mechanism is validated, it becomes a building block: a stable element that can be combined with others to reconstruct increasingly complex processes. Scientific inference proceeds by iterating this loop—observation, hypothesis, derivation, test—allowing mechanisms to accumulate combinatorially until the generative structure of the phenomenon becomes identifiable.

Limitation. Mechanistic reconstruction depends on how much of the true process is observable. When prior knowledge is weak—or when noise, entanglement, or limited access obscures the structure—only fragments of the mechanism can be recovered. Explanation loses resolution even if prediction remains viable. These limits reflect the information the world makes available; they bound the resolution at which mechanism can be inferred.

6 Discussion: The Narrow Window of Statistical Success

The analysis above does not imply that statistical methods never succeed. They succeed only in the restricted regimes where the underlying phenomena are so simple—linear, low-dimensional, and semantically shallow—that little mechanistic information is needed for adequate performance. In such settings, the task does not require recovering how the phenomenon unfolds. It requires only a numerically convenient rule that performs acceptably under the observed conditions—predicting the next value, stabilizing variation, or avoiding loss. When the world is sufficiently simple, coarse geometric approximations can meet these limited demands without representing the generative process.

These successes arise because the tasks can be solved without recovering the mechanism—not because inference from projected data is valid. The error arises when such narrow successes are generalized into a universal theory of reasoning. Methods that work only when semantics are minimal have been applied to domains in which semantics constitute the phenomenon—human behavior, social interaction, cognition, biological regulation, and economic dynamics. In these systems, the projection that produces data eliminates precisely the information that inference requires.

The challenge, therefore, is not to refine estimation or strengthen asymptotics. It is to replace the geometry-based ontology of inference with representations that preserve generative semantics. Only then can mechanisms be reconstructed rather than approximated from their shadows.

7 Conclusion

In domains where underlying mechanisms are opaque, statistics and causal inference have long served as instruments of certainty. Their appeal lies in translating complexity into geometry—fitting planes, aligning residuals, and manipulating conditional expressions—creating the appearance of explanation even when no mechanism is identified. Yet these operations construct structure within mathematics rather than recover structure from the world; projection can reveal alignment within data but never the process that produced them.

Physical science illustrates the opposite epistemology. Its progress has never come from fitting geometric patterns to observations, but from inferring the mechanisms that make those observations possible. Had physics relied on correlation, regression, or other forms of geometric alignment, it would have produced an ever-shifting landscape of plausible patterns rather than a stable science of causes. It is mechanism—not geometric fit—that allows explanation to move beyond the surface of data and reach the structure of the world.

Scientific progress ultimately follows only two valid paths. One infers mechanisms from structured priors and observation, as in physics. The other uses high-capacity models to achieve reliable prediction in the projected space without claiming to recover mechanism. Statistical inference offers neither, treating projected data as if it contained the structure that projection has removed. Understanding requires mechanism or prediction—not distributions mistaken for explanation.

Declaration of LLM Usage

The authors used OpenAI’s ChatGPT to assist in refining phrasing and improving clarity. All theoretical arguments and interpretations are original and authored by the researchers.

References

J. Angrist and G. Imbens. Identification and estimation of local average treatment effects, 1995.

- C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.
- O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
- P. S. de Laplace. *Théorie analytique des probabilités*, volume 7. Courcier, 1820.
- R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.
- F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.
- T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.
- C. G. Hempel and P. Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.
- K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.
- A. N. Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.
- Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- J. Pearl. *Causality*. Cambridge university press, 2009.
- K. Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.
- A. C. Rencher and G. B. Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.