Measure-Relational Algebra: The Core

Michael P. Laing
mlaing@post.harvard.edu
Prepared with drafting assistance from Anthropic Claude Fable 5

August 24, 2026

Abstract

Codd's relational model takes a data cell to be an exact value and equality to be the primitive comparison. We replace the cell by a probability measure on a metrized domain, carry dependence between cells as an explicit coupling, and derive operations by pushforward of the joint measure. The five defined operations reduce exactly to their classical counterparts in the Dirac, discrete-metric limit. The content of the generalization is one theorem: mean collapse commutes with a numerical derivation for every joint input law if and only if the derivation is affine, and commutes with every measurable derivation if and only if the cells are already Dirac. Point-valued practice collapses first and combines afterwards, and is therefore not an approximation of the correct computation but a different one; the discrepancy is systematic — of definite sign for convex or concave derivations, with second-order magnitude set by curvature and covariance — and it does not vanish merely by increasing the number of sampled cases while the measurement-error mechanism is unchanged. We exhibit in closed form a minimal model — linear response, multiplicative measurement error, lognormal prior — in which this premature collapse manufactures a *concave power law*: an apparent saturation with no counterpart in the underlying system. A 2026 *Nature* result exhibits this at scale: an apparent saturation in geomagnetic response, theorized for forty years, is reproduced by observational uncertainty alone — the data providing no statistical evidence that the saturation is physical.

Keywords: relational model; probability measure; pushforward; coupling; measurement uncertainty; errors-in-variables; regression dilution; probabilistic databases.

MSC 2020: 68P15 (primary); 60A10, 62J05 (secondary).

ACM CCS: Theory of computation → Incomplete, inconsistent, and uncertain databases; Information systems → Relational database model; Mathematics of computing → Probability and statistics.

1 The classical model

Codd's relational model [3] fixes for each attribute a domain D of admissible values. A tuple assigns to each attribute an element of its domain; a relation is a finite set of tuples over a common schema. Operations — selection, projection, join, aggregation, and derivation of new attributes by application of a function f — are defined on exact values, and the primitive comparison is equality.

The model's power is its closure: every operation returns a relation, so operations compose without limit. Its assumption is that a cell *is* a value.

For quantities obtained by measurement, that assumption is false. A measurement does not deliver a value; it delivers a state of knowledge about a value [10]. Recording the number and discarding the rest is a choice, and the purpose of this paper is to say precisely what that choice costs.

2 The core

2.1 Cells

Definition 2.1 (Metrized domain)**.** A *metrized domain* is a pair (D, d) with (D, d) a complete separable (Polish) metric space, equipped with the Borel σ-algebra $\mathcal{B}(D)$ generated by the open d-balls. Write $\mathcal{P}(D)$ for the probability measures on $(D, \mathcal{B}(D))$.

Completeness and separability exclude nothing arising in measurement practice and guarantee well-behaved products, disintegrations, and regular conditional probabilities, used from §3.1 onward. The structures divide the labour: the metric supplies comparison — proximity and similarity — the probability structure carries uncertainty, and affine structure on the domain, where present, is what licenses mean collapse (§3).

Definition 2.2 (Cell)**.** A *cell* over (D, d) is an element $\mu \in \mathcal{P}(D)$.

The interpretation is epistemic: the unknown true value is represented by a random variable with law μ, conditional on the available information; μ is what is known. A classical value v is the cell δ_v, the Dirac mass at v; thus $D \hookrightarrow \mathcal{P}(D)$, and the classical model occupies the image of that embedding.

2.2 Dependence

Definition 2.3 (Dependence structure)**.** For cells $\mu_1, \ldots, \mu_n$ over $D_1, \ldots, D_n$, a *dependence structure* is a measure $\Delta \in \mathcal{P}(D_1 \times \cdots \times D_n)$ whose i-th marginal is μ_i. The cells are *independent* when $\Delta = \bigotimes_i \mu_i$.

Δ is a coupling. It is carried as data — not derived, not assumed away, and not optimized over. Formally there is one such law per database state (§2.4); the couplings named in what follows are its marginals. Measurements sharing an instrument, a calibration, or a clock are coupled, and the coupling is part of what is known. Marginals do not determine it, which is why it must be represented explicitly rather than reconstructed. Nothing restricts Δ to the cells of one tuple: dependence runs within a tuple (the joint error of one observation event), across tuples within an attribute (a shared instrument, calibration, or clock), and across relations after a join; the tabular position of a cell is where it lives, not where its dependence stops.

2.3 Operations

Definition 2.4 (Proximity)**.** For $\mu_1, \mu_2 \in \mathcal{P}(D)$ with coupling Δ_{12} — taken to be the product $\mu_1 \otimes \mu_2$ when no dependence is specified — and tolerance $\theta \geq 0$,

$$S_\theta(\mu_1, \mu_2) \;=\; \mathbb{P}\big[d(X_1, X_2) \leq \theta\big], \qquad (X_1, X_2) \sim \Delta_{12},$$

and for a confidence level p, write $\mu_1 \approx_{\theta,p} \mu_2$ when $S_\theta(\mu_1, \mu_2) \geq p$.

Remark 2.5 (The product default). Taking $\Delta_{12} = \mu_1 \otimes \mu_2$ when no dependence is specified installs independence as a silent default — the same species of unexamined choice this model exists to make explicit. It is retained as operationally necessary and flagged as a modeling decision: independence is conservative in some directions and anticonservative in others, and the schema-level treatment of unspecified couplings is deferred to §6. The default is a provisional convention, not an inference; a conforming implementation should record that a defaulted coupling was constructed rather than elicited.

Definition 2.6 (Derivation)**.** For measurable $f\colon D_1 \times \cdots \times D_k \to E$ and cells with coupling Δ, the derived cell is the pushforward $f_*\Delta \in \mathcal{P}(E)$, where $(f_*\Delta)(A) = \Delta\big(f^{-1}(A)\big)$.

Selection retains tuples satisfying a predicate, which may be a proximity predicate. Projection marginalizes. Join pairs tuples whose cells satisfy $\approx_{\theta,p}$ and carries the induced coupling into the result. Aggregation is the pushforward of a group's joint measure through the aggregate function. These descriptions are made precise, and closure of the algebra is proved, in §2.4.

Proposition 2.7 (Codd limit)**.** If every cell is a Dirac mass, every metric is discrete, the tolerance satisfies $0 \leq \theta < 1$ (equivalently, θ below the minimum positive distance), and the confidence level satisfies $p > 0$, then each operation reduces to its classical counterpart: $\approx_{\theta,p}$ becomes equality, couplings are products of Diracs, and derivation is exact function application, $f_*\delta_v = \delta_{f(v)}$.

The corresponding classical fragment is thus not displaced but located: it is the algebra of $\mathcal{P}(D)$ restricted to Dirac masses.

Remark 2.8 (Where this sits)*.* A derivation is a deterministic Markov kernel between the associated measure spaces, and the algebra is a fragment of the category of such kernels. We note this only to locate the construction; nothing below depends on it.

2.4 Relations and closure

The definitions so far concern cells and their couplings; the relational object is now fixed, in the least structure sufficient for closure.

Definition 2.9 (Database state; measure-relation)**.** Fix a schema: finitely many relation names, each with attributes A carrying metrized domains (D_A, d_A). A *database state* is a pair (T, Δ), where T assigns to each relation name a finite set of tuple identifiers — determining a finite index set S of *slots* (t, A) — and, writing $D_s = D_A$ for $s = (t, A)$, $\Delta \in \mathcal{P}\big(\prod_{s\in S} D_s\big)$ is a single joint law over all slots. The *cell* of slot s is the marginal of Δ at s; the coupling of any finite family of slots is the corresponding marginal of Δ. A *measure-relation* is the restriction of the state to the slots of one relation name.

Coherence of dependence information is thereby definitional rather than assumed: every local coupling an operation consumes is a marginal of the one law Δ, so the marginals are automatically mutually compatible. (Pairwise couplings specified independently do not have this property: three pairwise laws with consistent one-dimensional marginals need not be the marginals of any joint law. The state is the joint law; pairwise materializations are views of it.) The finite *representation* of Δ is Problem 1, unchanged.

The operations act on states as follows; all are deterministic maps of states.

- *Derivation* by measurable f on slots $s_1, \ldots, s_k$ creates a new slot s' (over the schema extended by the new attribute) and replaces Δ by the pushforward of Δ under the graph map $x \mapsto \big(x, f(x_{s_1}, \ldots, x_{s_k})\big)$.
- *Projection* deletes slots and marginalizes Δ onto those remaining.

- *Selection* by a predicate — in particular a proximity predicate $\approx_{\theta,p}$ — retains the tuples that satisfy it and marginalizes Δ onto the surviving slots. The predicate is a *functional of the state*: each score S_θ is a number computed from a marginal of Δ, so selection is deterministic filtering, not conditioning; no zero-probability event arises. Conditioning on the value of a cell is a different operation — observation, §3.1 — with its own regularity questions (Problem 4).
- *Join* pairs tuples across two relations and retains the pairs satisfying the join predicate as evaluated on Δ. Since one tuple may participate in several retained pairs, its slots must appear once per pair: the output state is the pushforward of Δ under a coordinate-duplicating map $x \mapsto (\dots, x_s, \dots, x_s, \dots)$ giving each retained pair its own copies of the participating slots — the copies emerging perfectly coupled, as befits the same physical quantity — followed by marginalization onto the output slots. Dependence across the two relations is the corresponding marginal of Δ; the product default of Remark 2.5 is, in this formulation, a rule for *constructing* Δ when no cross-relation dependence has been elicited, not a property a state can lack.
- *Aggregation* of a group through an aggregate g is derivation by g on the group's slots followed by projection.

Proposition 2.10 (Closure)**.** Each operation above maps a database state to a database state.

Proof. Each operation is a composition of pushforwards of Δ under measurable maps — the graph map for derivation, coordinate projections for projection and marginalization, coordinate-duplicating maps for join — together with deterministic re-indexing of slots governed by predicates that are functions of Δ. The pushforward of a probability measure is a probability measure, so the result is again a single joint law over a finite slot set. □

Duplicate elimination, union, difference, and renaming are deliberately not defined here: renaming is trivial, and the set operations require a semantics for identity between measure-valued tuples, deferred with the completeness question (Problem 3). "Closed algebra" in this paper means closure of the five operations above.

3 What the generalization asserts

Enlarging the datum from D to $\mathcal{P}(D)$ is by itself only a change of representation. Its content lies in the relationship between the two, governed by the map that discards the enlargement.

Definition 3.1 (Collapse)**.** Write $\mathcal{P}_1(D) \subseteq \mathcal{P}(D)$ for the cells with finite first moment. A *collapse* is a map $\pi\colon \mathcal{P}_1(D) \to D$ assigning to a cell a point summary — canonically the mean, $\pi\mu = \mathbb{E}_{X\sim\mu}[X]$ on convex $D \subseteq \mathbb{R}^n$, so that the mean lies in D. It satisfies $\pi\delta_v = v$, and is the passage from the general model to the classical one.

Collapse acts on whatever cell is stored. Applied to a posterior cell $\mathcal{L}(X \mid W = w)$ it returns the conditional mean $\mathbb{E}[X \mid W = w]$; recording the observation itself — the cell δ_w — is a different point reduction, and the two coincide only in the exact-observation limit (§3.1).

Point-valued practice computes with collapsed values: given measured quantities and a function f, it forms $f(\pi\mu_1, \dots, \pi\mu_k)$ — collapse each input, then combine. The general model forms $f_*\Delta$ and collapses only if and when a point answer is required — combine, then collapse. These do not agree. Throughout, "the correct computation" means correctness under the declared cells and coupling; a misspecified error model is propagated with the same formal confidence as a right one (Problem 4).

Theorem 3.2 (Non-commutation). *Let $f\colon D_1\times\cdots\times D_k\to\mathbb{R}$ be measurable, with $D_i\subseteq\mathbb{R}^{n_i}$ convex, and let π be the mean. Then*

$$\pi\big(f_*\Delta\big) \;=\; f\big(\pi\mu_1,\dots,\pi\mu_k\big) \quad \textit{for every coupling } \Delta \textit{ for which both sides are defined}$$

if and only if f is affine.

Proof. ($\Leftarrow$) If $f(x)=a\cdot x+b$ then $\mathbb{E}[f(X)]=a\cdot\mathbb{E}[X]+b=f(\mathbb{E}X)$ by linearity of expectation, for any joint law Δ.

($\Rightarrow$) Suppose the identity holds whenever both sides are defined; for finitely supported couplings they always are. Take any $x,y\in D_1\times\cdots\times D_k$ and $\lambda\in(0,1)$, and let Δ put mass λ at x and $1-\lambda$ at y. The mean of the i-th marginal is the i-th block of $\lambda x+(1-\lambda)y$, so the right-hand side is $f(\lambda x+(1-\lambda)y)$, while the left-hand side is $\lambda f(x)+(1-\lambda)f(y)$. The identity therefore forces

$$f(\lambda x+(1-\lambda)y)=\lambda f(x)+(1-\lambda)f(y) \qquad \text{for all } x,y,\ \lambda\in(0,1),$$

i.e. f satisfies Jensen's functional equation for every $\lambda\in(0,1)$. It is therefore both convex and concave on the convex domain $D_1\times\cdots\times D_k$, hence affine; no regularity argument is required. □

Remark 3.3 (The quantifier matters). For $k\geq 2$ some non-affine f satisfy the identity for every *product* coupling: if $f(x,y)=xy$ and X, Y are independent with finite means, then $\mathbb{E}|XY|=\mathbb{E}|X|\,\mathbb{E}|Y|<\infty$, so both sides are defined, and $\mathbb{E}[XY]=\mathbb{E}[X]\,\mathbb{E}[Y]$. It is the dependent couplings that detect such f; Theorem 3.2 quantifies over all of them.

Corollary 3.4 (Why the classical model is self-consistent). If every cell is Dirac, both sides equal $f(v_1,\dots,v_k)$ for every measurable f. Conversely, if collapse commutes with every measurable derivation on a fixed family of cells, the cells are Dirac: taking f to be the indicator of $\{\pi\mu_i\}$ in the i-th coordinate makes the right-hand side 1 and the left-hand side $\mu_i(\{\pi\mu_i\})$, so $\mu_i=\delta_{\pi\mu_i}$. Collapse thus commutes with every measurable derivation exactly on the classical sub-model, where the remaining operations reduce to their classical forms outright (Proposition 2.7) — which is why no inconsistency is ever observed from within it.

Remark 3.5 (The discrepancy is systematic, and does not average away). For convex f, Jensen's inequality gives $\pi(f_*\Delta)-f(\pi\mu_1,\dots,\pi\mu_k)\geq 0$, with equality iff f coincides Δ-almost surely with an affine minorant of f; for concave f the inequality reverses. For convex or concave operations the discrepancy therefore has a definite sign; for general non-affine f it is a systematic transformation error whose sign is determined by the operation and the joint law. For a family concentrating at the mean — Δ_ε the joint law of $m+\varepsilon Z$ with $m=(\pi\mu_1,\dots,\pi\mu_k)$ fixed, Z centered with $\mathbb{E}\|Z\|^2<\infty$, and f twice continuously differentiable with bounded second derivative — Taylor expansion with dominated convergence gives $\pi(f_*\Delta_\varepsilon)-f(m)=\frac{1}{2}\operatorname{tr}\big(\nabla^2 f(m)\,\Sigma_\varepsilon\big)+o(\varepsilon^2)$, with $\Sigma_\varepsilon=\varepsilon^2\operatorname{Cov}(Z)$ the covariance of Δ_ε: a bias, not noise, whose second-order magnitude is controlled by the curvature of the operation and the covariance that was discarded.

Remark 3.6 (Not an artifact of choosing the mean). The mean is a natural collapse: it fixes Dirac masses, commutes with every affine map — including multi-argument maps such as the sum — and has a canonical definition on convex domains in every dimension. Alternative summaries do not share this. The one-dimensional median commutes with monotone maps, but already fails the two-argument sum: independent cells μ and ν with $\mu(\{0\})=0.6$, $\mu(\{10\})=0.4$ and $\nu(\{0\})=0.6$, $\nu(\{1\})=0.4$ each have median 0 while their sum has median 1. The mode commutes with little beyond bijections. Choosing the mean as π is therefore the most favourable case for the classical model; no uniqueness claim is made or needed, and Theorem 3.2 is stated for the collapse that gives the classical model its best chance.

3.1 The second failure mode: observation

Theorem 3.2 concerns the first failure mode: point reduction before a nonlinear derivation. A second and distinct loss occurs at *observation*, when a noisy proxy is recorded as though it were the latent quantity. Let X be a quantity of interest and W an observation of it under a known error model. What is known after observing $W = w$ is the conditional law $\mathcal{L}(X \mid W = w)$ — a cell. Point-valued practice instead records the number w, i.e. the cell δ_w. In the nondegenerate measurement-error setting these do not coincide; they agree in the exact-observation limit.

Proposition 3.7 (Attenuation)**.** Let $X \sim N(m, \tau^2)$ with $\tau^2 > 0$, and $W = X + \eta$ with $\eta \sim N(0, \sigma^2)$, $\sigma^2 > 0$, independent of X. Then

$$\mathbb{E}[X \mid W = w] \;=\; m + \lambda\,(w - m), \qquad \lambda = \frac{\tau^2}{\tau^2 + \sigma^2} < 1.$$

Consequently, if the true response is $Y = \beta X$, the population regression slope of Y on the recorded W is not β but $\lambda\beta$, since $\operatorname{Cov}(Y, W)/\operatorname{Var}(W) = \beta\tau^2/(\tau^2 + \sigma^2)$.

So recording w as though it were X biases subsequent analyses conditionally on the observation, and the bias is largest exactly where observations are most extreme. This is classical regression dilution. Correctly modelled repeated measurement of the same quantity does shrink the cell, and with it the gap — but shrinking the cell is an operation of this model (§3.1), not of the collapsed one. The next section shows that a slight change of prior turns this linear attenuation into something qualitatively worse.

4 Premature collapse can manufacture a law

4.1 A minimal model, in closed form

Attenuation rescales a slope but preserves linearity. If the quantity is positive and multiplicatively uncertain — the common case for physical intensities — the same mechanism produces curvature.

Proposition 4.1 (Spurious saturation)**.** Let $\log X \sim N(m, s^2)$ with $s^2 > 0$, and let the observation be multiplicatively perturbed, $\log W = \log X + \eta$ with $\eta \sim N(0, \sigma^2)$, $\sigma^2 > 0$, independent of X. Suppose the true response is exactly linear, $Y = \beta X$. Then

$$\mathbb{E}[\,Y \mid W = w\,] \;=\; C\,w^{\lambda}, \qquad \lambda = \frac{s^2}{s^2 + \sigma^2} \in (0, 1),$$

with $C = \beta \exp\big((1-\lambda)m + \tfrac{1}{2}\lambda\sigma^2\big)$ independent of w.

Proof. Write $L = \log X$ and $\ell = \log w$. Since $(L, \log W)$ is jointly Gaussian with $\log W = L + \eta$, the conditional law of L given $\log W = \ell$ is Gaussian with mean $m + \lambda(\ell - m)$ and variance $\lambda\sigma^2$, where $\lambda = s^2/(s^2 + \sigma^2)$. Hence, using $\mathbb{E}[e^Z] = e^{\mu_Z + \sigma_Z^2/2}$ for Z Gaussian,

$$\mathbb{E}[X \mid W = w] = \mathbb{E}\big[e^L \mid \log W = \ell\big] = \exp\Big(m + \lambda(\ell - m) + \tfrac{1}{2}\lambda\sigma^2\Big) = C'\,w^{\lambda},$$

with $C' = \exp\big((1-\lambda)m + \tfrac{1}{2}\lambda\sigma^2\big)$. Multiply by β. □

The $\frac{1}{2}\lambda\sigma^2$ in the exponent is the exact analogue of the correction of Remark 3.5: $\lambda\sigma^2$ is the conditional variance $\operatorname{Var}(L \mid W = w)$, and for $f = \exp$ with Gaussian input the curvature correction resums exactly *in the exponent*, $\log \mathbb{E}[e^Z] - \mathbb{E}[Z] = \frac{1}{2}\operatorname{Var}(Z)$, agreeing with the additive second-order term $\frac{1}{2}\operatorname{tr}(\nabla^2 f\,\Sigma)$ of Remark 3.5 to leading order in the variance. The two failure modes are thus both present in Proposition 4.1 and remain separable in the formula: conditioning regresses the mean of L toward m, producing the exponent $\lambda < 1$ and with it the concavity, while the curvature of exp contributes the factor $e^{\lambda\sigma^2/2}$ to the constant C. One is paid at observation, the other at transformation; they are mathematically distinct effects of the single decision to replace a measure by a point.

Corollary 4.2. Since $0 < \lambda < 1$, the function $w \mapsto Cw^\lambda$ is strictly concave with $d\,\mathbb{E}[Y \mid W = w]/dw \to 0$ as $w \to \infty$: the *marginal* response to increasing drive vanishes at the extremes, which is precisely what "saturation" denotes operationally. On log–log axes the observed relation is a line of slope $\lambda < 1$ against a true slope of 1 — a sublinear power law where the physics is linear. No ceiling exists in the model: the true response $Y = \beta X$ is linear and unbounded, and the apparent exponent λ measures nothing but the ratio of signal dispersion to total dispersion, with $\lambda \to 1$ (no saturation) exactly as $\sigma \to 0$.

Remark 4.3. The minimal model produces a pure power law; the empirical curves in the geophysical case are linear at low drive and bend only at high drive, a shape produced by the temporal-displacement error model of [15], which is state-dependent where ours is homogeneous. The mechanism — conditioning pulls extreme observations toward the bulk, and more strongly the larger the error — is the same; the minimal model trades fidelity of shape for a closed form.

This is the whole phenomenon in three lines. A linear law, observed through multiplicative error, and collapsed to a point at the moment of observation, presents itself as saturation: a concave power law whose marginal response vanishes at the extremes. The saturation is an artifact of the data model. It is smooth, reproducible, and physically plausible, and there is nothing in the recorded data to mark it as spurious.

4.2 The case in which this happened

The solar wind drives Earth's magnetosphere. The driver — a reconnection electric field $\widehat{E}_m$ — cannot be measured where it acts. It is measured by spacecraft at the L1 Lagrange point, some 1.5×10^6 km upstream, and propagated forward through an uncertain transit time and an uncertain change in magnitude en route. The available quantity is therefore an observation W of the true driver X — an instance of the structure of Propositions 3.7 and 4.1, with the error entering through temporal displacement rather than i.i.d. multiplicative noise.

Plotted against the geomagnetic response, the relationship is linear at low drive and bends over at high drive. Saturation was taken to be physics. At least nine mechanisms were proposed for it, and a response ceiling entered the models used to estimate the consequences of extreme space weather.

Sivadas et al. [15] show that the available observations provide no statistical evidence for the inferred saturation. Modelling the true driver as lognormal with autocorrelation time κ and the observation as displaced by a random temporal offset of dispersion σ_Δ, they prove the regression of response on W is nonlinear whenever $\sigma_\Delta/\kappa > 0$, with curvature increasing in that ratio, even when the true response is exactly linear in X. The apparent saturation curve is reproduced quantitatively over twenty-five years of observations with no physical mechanism invoked. Applying regression calibration — replacing the recorded W by $\mathbb{E}[X \mid W]$ — the curvature disappears and the response is linear throughout the observed range.

The consequence is not academic. Under the resulting non-saturating extrapolation, the geomagnetic impact of a Carrington-scale storm is roughly twice what the saturating models predicted.

In the vocabulary of §3: a cell was collapsed to a point at the moment of observation, and the collapsed value was then combined with another measured quantity. Because the observation enters nonlinearly — the conditional mean of the true driver is a concave function of the recorded value — and the collapse premature, the result was not merely imprecise. It exhibited structure for which the system itself provides no evidence: the apparent saturation that forty years of theory sought to explain is reproduced by the observation model alone, undermining the empirical motivation for a physical ceiling.

5 Relation to existing work

The mathematics used here is elementary and long known. What follows states precisely what is being claimed, by saying what is not.

Measurement science and errors-in-variables. That a measurement delivers a state of knowledge — a distribution, not a bare number — is the premise of modern metrology [10]; regression dilution, regression to the mean, and regression calibration are classical [2]. We introduce no estimator and no statistical result. Propositions 3.7 and 4.1 are textbook computations, included because they make the mechanism explicit, not because they are new. Indeed, for a single analysis whose error structure is already suspected, a statistician with an errors-in-variables package produces the same numbers at less cost. The claim is not mathematical but structural: correctness of this kind should reside in the data model rather than in the analyst's vigilance.

Probabilistic databases. A substantial literature equips the relational model with uncertainty. The classical line — conditional tables [12], probabilistic relational algebra [5], and the possible-worlds systems surveyed in [16] — annotates *tuples* with probabilities of existence or correctness and asks for the probability that a tuple belongs to the answer. Attribute-level uncertainty has also been studied: cells holding distributions over possible values, with cells taken to be independent and the emphasis on integrity constraints and most-probable instantiation [7]; and the possible-worlds semantics has been extended to continuous domains, with Polish-space foundations and measurability theorems for relational-algebra queries [9]. Mathematically, a database state in the sense of Definition 2.9 is close kin to these: a single law over instances is what a continuous possible-worlds semantics also maintains, and such a law can encode arbitrary dependence. The contribution claimed is therefore not the object but the discipline imposed on it: the measure *is* the cell rather than metadata about it; dependence is surfaced as a declared coupling rather than left implicit in an instance-level law; pushforward on metrized domains is the default derivation semantics rather than an opt-in; collapse is an explicit operation; and the collapse non-commutation theorem is the design criterion that makes the discipline enforceable. The question asked is also different: not the probability that a tuple belongs to the answer, but what happens to a value's uncertainty when the value is pushed through a nonlinear derivation.

Monte Carlo databases. Closest in spirit is MCDB [13], which propagates arbitrary uncertainty through arbitrary queries by generating realizations from user-declared variable-generation functions and executing the query over bundles of sampled tuples. MCDB establishes that such propagation is computationally practical — an important precedent. The structural difference is where the uncertainty lives: in MCDB it is external to the data model, supplied by the analyst as a generator, and query processing proceeds over the classical relational model; here the measure *is* the cell, propagation is the default, and collapse is the operation that must be requested. The externalization has a consequence: propagation is available only where a generator has been declared, and the failure of §4 occurred precisely where no one suspected one was needed — no generator for transit-time uncertainty was ever going to be declared for a saturation nobody doubted.

Similarity and fuzzy relational algebras. Relational algebras extended with graded matching — multi-similarity algebra [1], similarity relational calculus [14], and the De Morgan frame model of Hajdinjak and Bierman [11], which argues for attribute-level rather than tuple-level annotation — annotate data with *degrees* drawn from a semiring or lattice. These answer "how well does this match". They do not represent a distribution over the value, and have no analogue of pushforward, of dependence as a coupling, or of Theorem 3.2.

Approximate algebraic structures. Work on ε-relaxed algebraic axioms, including the recent framework of Ghojogh and Amanpour [6], is orthogonal: there the operations are exact maps and the *laws* hold only to within tolerance, collapsing to classical algebra as $\varepsilon \to 0$. Here the laws hold exactly and the *datum* is enlarged, collapsing to Codd in the Dirac limit.

Optimal transport. The objects coincide — cells are points of $\mathcal{P}(D)$ over a metric space, and dependence structures are couplings — but no optimization occurs. A coupling is carried because the physics specifies it, not solved for as the minimizer of a transport cost.

What is claimed

Only this: that taking the cell to be a measure on a metrized domain, with dependence as an explicit coupling and operations as pushforward, yields a closed algebra (Proposition 2.10) containing the corresponding fragment of Codd's exactly as its Dirac restriction (Proposition 2.7); and that collapse commutes with its numerical derivations only in the degenerate directions — for every coupling only when the derivation is affine (Theorem 3.2), and for every measurable derivation only when the cells are Dirac (Corollary 3.4) — so that point-valued practice is a *different* computation rather than an approximate one, differing systematically and capable of exhibiting structure the modelled system does not have (§4). The corrective mathematics is old; the claim is that the data model should make it unnecessary to know that one needs it.

A final statement of intent locates the paper among its neighbours: this is a specification, not a system. The literature above answers, in different ways, how a system may compute over uncertain data; the core presented here is meant, in the manner of Codd's paper [3], to be implementation-independent — to say what any conforming system must preserve, and to gauge what an implementation loses, with Theorem 3.2 marking where exactness is achievable and the second-order correction of Remark 3.5 estimating the cost where it is not. The problems of §6 measure the distance between this specification and a usable system.

6 Open problems

The preceding sections establish what the measure-relational model asserts and prove the one theorem on which it rests. Deliberately excluded were finite representation, query language, implementation, and cost — the questions that determine whether the model can be *used*. They are collected here, stated as precisely as their current state permits. The author, a retired engineer, does not expect to resolve them; they are offered as work.

Problem 1 (Finite representation under the algebra)**.** Exhibit a class $\mathcal{R}$ of finitely representable measures on metrized domains, and a class of finitely representable couplings over $\mathcal{R}$, such that (i) $\mathcal{R}$ contains the Dirac masses and the empirical, parametric, and mixture families arising in measurement practice; (ii) the pushforward of a represented coupling under each algebra operation is again representable in $\mathcal{R}$, exactly or with quantified error; and (iii) the cost of maintaining the representation through a k-way join is bounded polynomially in k and the representation sizes.

This is the load-bearing engineering problem, and it is genuinely open — I am not aware of a candidate class satisfying all three conditions. The candidates are the obvious ones, and each fails one leg as stated: point-plus-tolerance pairs are not closed under nonlinear pushforward and cannot express the lognormal cell of §4 at all; parametric families are closed only under maps preserving the family; sample bundles (the MCDB representation [13]) satisfy (i) and (ii) with statistical error but couple that error to sample size in ways that compound through joins; polynomial chaos and moment closures truncate with error that is quantifiable only under smoothness assumptions. A satisfactory answer may well be a *stratified* representation — exact where affine (propagation through affine maps is exact in closed form), approximate elsewhere — which would make the theorem an optimization criterion rather than only an impossibility result. I am not aware of results beyond folklore on the right stratification.

Problem 2 (Normal forms for carried couplings)**.** The model carries the coupling Δ as data (§2.2). A coupling among n cells that share a physical cause — one instrument, one calibration event, one clock — is then embedded in every stored Δ that includes any of the n cells ($\binom{n}{2}$ times in a pairwise materialization), and a recalibration must update every copy: precisely the update anomaly that normalization exists to prevent. Formulate a normal-form theory in which the *cause* of a coupling is a first-class entity, couplings are keyed by shared cause, and Δ is materialized from cause entities on demand. Characterize the schemas for which this factorization is lossless — i.e., for which the materialized coupling equals the carried one — and the class of dependence structures (non-pairwise, non-Gaussian, time-varying) for which no finite cause factorization exists.

This is the analogue, one level up, of the passage from unnormalized to normalized schemas: it moves a correctness burden (keeping ten thousand embedded couplings consistent with one calibration certificate) from vigilance into structure. Classical dependency theory (functional and multivalued dependencies, the chase) should be the model; what replaces a functional dependency when the dependent object is a joint law rather than a value is not known.

Problem 3 (Completeness of a measure-relational query language)**.** Codd's completeness theorem gives the relational algebra a calibration: a language is complete if it expresses everything the relational calculus does. State and prove the analogue. Specifically: define a measure-relational calculus over cells and couplings — presumably with proximity predicates $\Pr[X \in A] \geq p$ in the role that membership predicates play classically — and characterize the algebra (which operations beyond pushforward, marginalization, and conditioning are required?) that achieves exactly its expressive power. Establish whether collapse π must be a primitive of a complete language or is definable from the rest.

The Dirac restriction gives a lower bound — any complete language must contain Codd's — and the categorical literature on Markov categories and the Giry monad [8, 4] supplies composition machinery, but I am not aware of a completeness theorem in the database sense there. The answer disciplines language design: without it, every added construct is a taste judgment.

Problem 4 (Existence and elicitation of the conditional cell)**.** Section 3.1 takes as given that observing $W = w$ under a known error model yields the cell $\mathcal{L}(X \mid W = w)$. Two gaps separate that sentence from a data model. Mathematically: the conditional law is defined via disintegration, which exists under regularity (automatic for the Polish domains of Definition 2.1), is unique only up to W-null sets, and requires a prior law for X — state the weakest hypotheses under which the model's cells are well-defined objects rather than conveniences, and where the prior is to live in the schema. Practically: the cell is only as good as the elicited error model, and a wrong coupling is propagated with the same formal confidence as a right one. Characterize the sensitivity of downstream queries to misspecification of the observation kernel — a perturbation bound of the form: an error of ε in the kernel (in a named metric) moves π of any query output by at most $\delta(\varepsilon, f)$ — so that stored error models can carry quantified trust rather than implied exactness.

The second half is the model's honest exposure. Point-valued practice also guesses — Dirac cells and product couplings are guesses, invisible ones — but a system that prints distributions invites more trust than one that prints numbers, and that trust needs a warranted region.

Problem 5 (Cost models and the affine frontier)**.** For classical engines, selectivity estimation and access-path selection rest on statistics of stored values. Here the stored values are measures, and the frontier question is where propagation can be avoided. By Theorem 3.2, π may be pushed from the root of a plan through every affine node: the nodes whose entire path to the root is affine form a frontier above which execution is classical, with zero error. Below the frontier collapse is not licensed — an affine fragment's output cannot be replaced by its mean when a non-affine consumer awaits, since the collapse discards both its dispersion and its coupling to the fragment's siblings — but affine nodes still cost nothing there: they transport mean and covariance exactly ($m \mapsto Am + b$, $\Sigma \mapsto A\Sigma A^{\mathsf{T}}$, for any distribution), and the transported Σ is precisely the input to the second-order correction $\frac{1}{2}\operatorname{tr}(\nabla^2 f\, \Sigma)$ of Remark 3.5 at each curved node. Develop the optimizer theory this implies: plan rewriting that enlarges the region above the affine frontier; selectivity estimation for proximity predicates; and an error-budgeted mode in which the user names a tolerance on π of the answer and the optimizer chooses the cheapest plan whose accumulated second-order error estimate respects it.

This is the point at which the theorem stops being a warning and starts being a query optimization: the impossibility result, read constructively, is a licence to collapse everything above the affine frontier and a price list for everything below it.

Problem 6 (Decision operators and audited collapse)**.** An action is a point; collapse is therefore deferred to the moment of decision, not eliminated. Presentation is different: a result may be rendered as a distribution — a density, a quantile band, shaded probability mass — a terminal reduction that marginalizes the answer cell, preserving its law and discarding only the carried coupling, and strictly finer than the mean. It is decision, not reporting, that forces a point. The model makes collapse an explicit operation, which permits — but does not yet define — a decision layer: operators of the form $\arg\min_a \mathbb{E}_\Delta[\ell(a, X)]$ over a stored loss function ℓ, evaluated against the joint measure rather than against collapsed inputs. Specify the algebra of such operators (composition with pushforward; interaction with conditioning), the provenance a collapse should record (what was collapsed, when, under what loss, with what dispersion discarded), and the conditions under which the classical decision — threshold the collapsed value — coincides with the expected-loss decision. The divergence condition is the practically important object: it says exactly when carrying the measure changes what an organization does.

Problem 7 (Empirical replication and a benchmark). Reproduce the geomagnetic case of §4.2 end to end from public data (solar-wind measurements and ground-response indices, 1995–2019): execute the stimulus–response join classically, with Dirac cells, and re-derive the spurious saturation; execute it measure-relationally, with the observation kernel of the upstream instrument as the cell, and test whether the linear response is recovered under the declared observation model. Beyond the single case, specify a benchmark suite for uncertainty-propagating query processing — schema, queries, error models, and correctness criteria with known ground truth — so that representations proposed under Problem 1 and optimizers under Problem 5 can be compared on common ground. I am not aware of such a benchmark.

Problem 7 requires no new theory and can begin immediately; Problem 1 gates everything else; Problems 2–6 are separable and of thesis scale. The author's expectation is that the stratified-representation question of Problem 1 and the affine-frontier optimizer of Problem 5 are two halves of one dissertation, and that Problem 2 is another entire.

7 Statement of the core

C1. A measured cell is an element of $\mathcal{P}(D)$, not of D; dependence among cells is a coupling, carried as data.

C2. Operations are pushforwards of the joint measure; the five defined operations map database states to database states (Proposition 2.10), so they compose without leaving the model.

C3. Restricted to Dirac cells and discrete metrics, the defined operations are their classical counterparts, exactly.

C4. Collapse commutes with a numerical derivation for every joint law if and only if the derivation is affine, and commutes with every measurable derivation if and only if the cells are already Dirac. Hence point-valued practice differs from uncertainty-preserving computation by a systematic, curvature-determined error — of definite sign for convex or concave derivations — that does not vanish merely by increasing the number of sampled cases while the measurement-error mechanism is unchanged, and that can manufacture structure absent from the system.

C4 is what makes C1–C3 worth adopting; without it the enlargement from D to $\mathcal{P}(D)$ would be a matter of taste.

The questions excluded by design — finite representation under the algebra, completeness, query language, implementation, and cost — are posed precisely in §6. Consequences for machine reasoning over uncertain observations are treated separately.

References

[1] S. Adali, P. Bonatti, M. L. Sapino, V. S. Subrahmanian, *A multi-similarity algebra*, ACM SIGMOD, 402–413 (1998).

[2] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed., Chapman & Hall/CRC (2006).

[3] E. F. Codd, *A relational model of data for large shared data banks*, Comm. ACM **13**(6), 377–387 (1970).

[4] T. Fritz, *A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics*, Adv. Math. **370**, 107239 (2020).

[5] N. Fuhr, T. Rölleke, *A probabilistic relational algebra for the integration of information retrieval and database systems*, ACM Trans. Inf. Syst. **15**(1), 32–66 (1997).

[6] B. Ghojogh, G. Amanpour, *Similarity algebra: a framework for approximate algebraic and Lie structures with collapse to classical algebra*, arXiv:2602.14075 [math.RA] (2026).

[7] A. Gilad, A. Imber, B. Kimelfeld, *The consistency of probabilistic databases with independent cells*, Proc. 26th Int. Conf. on Database Theory (ICDT 2023), LIPIcs **255**, 22:1–22:19 (2023).

[8] M. Giry, *A categorical approach to probability theory*, in: Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics **915**, Springer, 68–85 (1982).

[9] M. Grohe, P. Lindner, *Infinite probabilistic databases*, Proc. 23rd Int. Conf. on Database Theory (ICDT 2020), LIPIcs **155**, 16:1–16:20 (2020); journal version, Log. Methods Comput. Sci. **18**(1), 34:1–34:43 (2022), doi:10.46298/lmcs-18(1:34)2022.

[10] JCGM 100:2008, *Evaluation of measurement data — Guide to the expression of uncertainty in measurement* (GUM).

[11] M. Hajdinjak, G. Bierman, *Extending relational algebra with similarities*, Math. Structures in Comp. Sci. **22**(4), 686–718 (2012).

[12] T. Imieliński, W. Lipski, *Incomplete information in relational databases*, J. ACM **31**(4), 761–791 (1984).

[13] R. Jampani, F. Xu, M. Wu, L. L. Perez, C. Jermaine, P. J. Haas, *MCDB: a Monte Carlo approach to managing uncertain data*, ACM SIGMOD, 687–700 (2008), doi:10.1145/1376616.1376686.

[14] I. Schmitt, N. Schulz, *Similarity relational calculus and its reduction to a similarity algebra*, FoIKS, Springer, 252–272 (2004).

[15] N. Sivadas, D. Sibeck, V. Subramanyan, M.-T. Walach, D. S. Ozturk, B. Ferdousi, B. Michotte de Welle, *Regression to the mean can explain saturation of geomagnetic storms*, Nature **655**, 1143–1147 (2026), doi:10.1038/s41586-026-10757-4.

[16] D. Suciu, D. Olteanu, C. Ré, C. Koch, *Probabilistic Databases*, Morgan & Claypool (2011).