

Measure-Relational Data: A Practical Foundation for Correctness

Michael P. Laing
Independent
ORCID 0009-0004-5593-5759
mlaing@post.harvard.edu

ABSTRACT

Codd's 1981 Turing lecture did not argue the relational model; it argued what the model bought. The productivity case rested on three failures of the systems that preceded it, three objectives that answered them, and a sharp line separating systems that had the relational processing capability from systems that merely had tables. This article makes the same kind of case for a different enlargement of the data model. Where Codd replaced positional addressing by associative addressing, we replace the exact-valued cell by a probability measure, with dependence among cells carried as an explicit coupling and operations derived by pushforward. The mathematics is elementary and long known. The argument is structural: correctness under measurement uncertainty currently resides in the analyst's vigilance, and the model should make that vigilance unnecessary. One theorem gives the argument its edge—collapsing cells to points before a derivation gives the same answer as collapsing after it, for every joint input law, only when the derivation is affine—and one recent *Nature* result gives it its evidence: a saturation in Earth's geomagnetic response, long taken for physics, is what a linear response looks like when the driver is stored as a value and its measurement error is stored nowhere. We state the minimal capability a system must have to be called measure-relational, answer the two skeptical questions Codd answered in 1981, and say honestly what we do not have: installations. What is claimed is a specification—what a conforming system must preserve, and the instruments to gauge what an implementation loses; not a system, not a benchmark result, and not a claim about how often the discrepancy is large.

1. THE QUESTION

Codd's 1970 paper [3] gave the relational model. His 1981 Turing Award lecture [4] gave the reason to adopt it, and the reason was not mathematical. It was a productivity crisis: demand for applications was outrunning the capacity to write them, and the database systems of the sixties, though successful as

instruments of data control, had failed to relieve it. He gave three reasons for the failure, three objectives of the relational model that answered them, and then spent most of the lecture on skepticism—range of services, performance—and on drawing a line between systems that possessed what he called the relational processing capability and systems that merely stored tables.

A specification of a measure-relational algebra now exists [8]: cells as probability measures on metrized domains, dependence as an explicit coupling, operations as pushforward, collapse as an operation rather than a default. It has a closure result, a theorem, a worked mechanism, and eight open problems. What it does not have, because a specification cannot supply it, is the answer to the question Codd's lecture answered. Why do it?

This article is the attempt. Its structure is Codd's, and the debt is acknowledged at each step, including the step where the parallel breaks.

2. WHY THE CORRECTIONS ARE NOT APPLIED

Modern metrology starts from the position that what a measurement yields is not a value but a state of knowledge about one [5]. The consequences of pretending otherwise are textbook: regression dilution, regression to the mean, and the corrections for them—errors-in-variables models, regression calibration [2]—have been in the statistics curriculum for generations. Nothing in this article is new mathematics.

The question is therefore not whether the corrections exist but why they are so rarely applied, and the answer parallels Codd's three reasons.

The data model has no place for the uncertainty. Codd's first complaint was that programmers were forced to think at a needlessly low level of structural detail. Ours is that analysts are forced to think about uncertainty outside the structure entirely. A

relational cell holds a value. The instrument’s error model, the calibration certificate, the fact that two columns share a clock: none of it has a home in the schema. It lives in a methods section, a notebook, or the head of the person who took the measurement, and it does not travel with the number when the number is joined to something else. The low-level detail that Codd removed was navigation; the detail that remains is the analyst’s private bookkeeping of what each number actually knows.

There is no propagation, so every analysis does its own. Codd’s second complaint was the absence of set processing: with no way to say “all of these,” programmers wrote loops. The analogue is the absence of propagation. With no way to say “carry the law through this derivation,” analysts who know they must propagate uncertainty write the delta method by hand, once per analysis, for the derivations they happen to notice are nonlinear. The loop that Codd eliminated was mechanical and visible. The propagation that is missing is mechanical and invisible, and invisibility is the worse property: a missing loop produces no output, while a missing propagation produces a plausible one.

The case that matters is the case nobody suspected. Codd’s third complaint was that ad hoc query had been treated as an afterthought—something to add later, for the unanticipated question. The analogue is that uncertainty propagation is treated as something to add later, for the analysis where someone suspects the error structure matters. The systems that do propagate uncertainty [6, 7, 10] do so where an analyst has declared a generator, a variable, an error model. Where no one suspects one is needed, none is declared. Section 3 is a case in which, for as long as the saturation was modelled, no one suspected one was needed.

3. THE EXHIBIT

The case is set out in full in [8, §4.2] and in its source [11]; we give only its shape. Earth’s magnetosphere is driven by the solar wind. The driver is measured far upstream, by spacecraft, and carried forward to where it acts through a transit whose duration is not known exactly. As the record of extreme events accumulated, the response appeared to saturate at high drive, and the saturation was taken for physics: mechanisms were proposed to explain it, and a ceiling on the response entered the models by which the consequences of extreme space weather are estimated.

Sivadas et al. [11] show that the observations carry

no statistical evidence for the saturation. Because the recorded driver is a temporally displaced observation of the true one, the regression of response on the recorded value bends whenever the displacement has positive dispersion, even for a response that is exactly linear in the true driver; the observed curve is reproduced over twenty-five years of data with no mechanism invoked; and regression calibration—replacing the recorded value by its conditional expectation given the observation—straightens it. Without the ceiling, the estimated impact of the largest storms roughly doubles.

In the vocabulary of this article: the driver was collapsed to a point at observation and then joined to the response. The loss was taken at the observation, not in the join: what the recorded value says about the true driver is concave in the recorded value, so a response that is linear in the true driver is concave in the recorded one. The output was not noisy. It had shape, and the shape was believed, because nothing in the recorded data distinguished it from physics. The corrective mathematics existed throughout and was not applied, because no column, no schema, and no query knew that it was needed.

This is the difference between Codd’s evidence and ours, and it is worth stating plainly. Codd, in 1981, had System R, INGRES, and seven hundred Tandem installations behind him; the productivity case was made with running systems in evidence. We have no installations. What we have is one instance, priced, of what point-valued practice cost, where someone happened to look; the same reduction documented, and repaired, under other names in disciplines that do not cite one another—distances inverted from recorded parallaxes, with a posterior per star recommended instead [1]; associations fitted to once-recorded blood pressure, corrected upward by about sixty percent [9]; calendar ages read off a point radiocarbon age, with the full calibrated distribution recommended in their place [12]; and a theorem saying that the mechanism is general—collapsing before a non-affine derivation is a different computation from collapsing after it, not an approximation of it—though not how often the difference is large, or noticed.

4. WHAT THE MODEL ASSERTS

The specification [8] is short and we do not reproduce it. Four commitments carry the argument.

A *cell* over a metrized domain (D, d) —a complete separable metric space with its Borel sets—is a probability measure $\mu \in \mathcal{P}(D)$. A classical value v is the Dirac mass δ_v , so the classical model is the image of $D \hookrightarrow \mathcal{P}(D)$ and is not displaced but located.

Dependence among cells is itself data—a coupling stored with the cells, one joint law per database state—from which each operation takes the marginal it needs. Because the marginals do not fix the joint law, the joint law has to be stored, not rebuilt from them; two measurements that share an instrument, a calibration event, or a clock are coupled, and that coupling is knowledge the record should hold.

Derivation by a measurable f is the pushforward $f_*\Delta$ of the joint law. Selection, projection, join, and aggregation are defined on states and are closed; in the Dirac, discrete-metric limit they reduce to Codd’s operations under bag semantics.

Collapse π —canonically the mean—takes the general model back to the classical one, and it is an operation with a place in the plan. Point-valued practice performs it at observation and combines afterwards; the model combines first and collapses at the end, if a point is wanted at all.

The theorem is what makes the four commitments more than a change of representation. For real-valued f measurable on a product of convex domains and π the mean,

$$\pi(f_*\Delta) = f(\pi\mu_1, \dots, \pi\mu_k) \quad \text{for every joint law } \Delta$$

(over every joint law on which both sides are defined) if and only if f is affine; and for a fixed family of cells, collapse commutes with every measurable f only when the cells are Dirac to begin with. The proof is Jensen’s functional equation. So computing on points is not a cheaper version of computing on laws; it is a different computation, and what separates the two is a bias rather than noise. The bias has a definite sign when the derivation is convex or concave; its size to second order is $\frac{1}{2} \text{tr}(\nabla^2 f \Sigma)$, the derivation’s curvature weighted by the covariance the collapse threw away; and no quantity of additional samples reduces it while the errors are generated the same way. There is a second loss, distinct from this one, taken at observation: a noisy proxy written down as if it were the quantity. Take the smallest case that closes in one formula: a linear response, a lognormal prior on the driver, multiplicative observation error. There the observation-time loss on its own produces a concave power law, with the derivation’s curvature entering only as a constant factor [8, §4.1]. That is the mechanism of the exhibit of Section 3 in three lines.

5. THREE OBJECTIVES, TRANSPOSED

Codd stated three objectives for the relational model; met, they were his answer to the three failures. The measure-relational model has three analogues.

Error-model independence. Codd’s data independence separated the logical view from physical storage, so that a change in layout did not force a change in programs. The analogue separates the recorded quantity from its error model, so that a recalibration does not force a re-analysis. When the coupling is data, keyed by the cause the cells share—the instrument, the calibration event, the clock—a new calibration certificate updates the coupling in one place and every consumer of the affected cells sees the change. When the coupling is not data, the same certificate requires every analysis that ever touched the instrument to be found and redone, and in practice is not. This is the normal-form question of the specification (Problem 2 of [8]): an update anomaly of exactly the kind normalization was invented to remove, one level up.

Propagation by default. Codd’s set-processing objective eliminated the loop by letting the system do the navigation. The analogue eliminates the hand-written delta method by letting the system do the propagation. The analyst declares the cell—which instrument, which error model—once, at the point of observation, and every derivation downstream carries the law. Nothing has to be noticed. The nonlinear derivation three joins away, which no one wrote down as nonlinear, is propagated through exactly as the linear one is. Where propagation is unnecessary the theorem says so: above the affine frontier collapse is exact and execution is classical, at no cost.

Auditable derivation. Codd’s communicability objective was structural simplicity: a table that programmers and end users could both read. The analogue is that the table stays a table, a cell renders as a distribution when presentation is wanted, and a presented result carries the derivation that produced it—a term of the algebra in which the collapse, when a decision finally requires one, is an explicit node recording which cell was collapsed, at what stage of the plan, by which loss, and how much dispersion was given up. The number in the report is no longer the only thing the report knows, and the report can be checked: a small reference evaluator that implements the algebra and nothing else re-derives the result from its term, so that an implementation may be as large and as opaque as it likes while what it presents stays checkable by something small, for its handling of uncertainty if not for the maps it applied. It is decision, not reporting, that forces the collapse, and the model makes that moment a recorded operation rather than an unexamined

default.

6. DRAWING THE LINE

Codd’s lecture drew a sharp line so that the label would not be used in misleading ways, and he drew it with a minimal capability: tables without user-visible navigation links, and a sublanguage in which select, project, and unrestricted join could be expressed without iteration. A system with the tables but not the sublanguage was tabular, not relational.

The analogue is needed for the same reason. “Uncertainty-aware” is already a label, and most of what it labels does not meet the bar. A system is *measure-relational* if it supports:

- (1) *The measure as the cell.* The law of the quantity is the datum, not metadata attached to a point. A column of values with a companion column of standard errors does not qualify: the pair is not closed under nonlinear pushforward, cannot express the lognormal cell of the exhibit at all, and carries no coupling.
- (2) *Coupling as data.* Dependence among cells is declared and carried, not assumed independent by default. An implementation that must default a coupling records that the coupling was constructed rather than elicited.
- (3) *Pushforward as the default derivation.* A derivation propagates the law unless the system can show, by the affine criterion, that propagation is unnecessary. Propagation is not an opt-in attached to a declared generator.
- (4) *Derivation as a checkable record.* A result carries the term that produced it; a point is produced only where requested, the request is a node of that term, and a reference evaluator can re-derive the result from the term.

A system with (1) but not (3) stores distributions and computes on their means; it is the tabular case. A system with the propagation of (3) but neither its default nor (1)—MCDB [6] is the mature example—propagates faithfully wherever someone has declared a generator, and the failure of Section 3 happened exactly where no one had. The line is drawn at the conjunction, and the theorem is what makes the line enforceable: a conforming implementation may omit propagation exactly above the affine frontier, and a non-conforming one pays, below it, a bias whose leading term is the second-order correction.

Codd added, in a postscript to the lecture’s 1987 reprint, that within four years most vendors were claiming the label, many claiming more than the

minimal capability they delivered and a few not delivering even that, and that he had published twelve rules to score them. We expect the same and offer the four conditions above as the score sheet.

7. SKEPTICISM

Codd posed two questions on behalf of the skeptic: can a relational system provide the range of services expected of a database system, and can it perform. He answered both from installations. We answer both from the open problems, which is weaker, and say so.

Range of services. The specification defines five operations—selection, projection, join, derivation, aggregation—and proves them closed. It does not define union, difference, or duplicate elimination, because each of those needs a rule for when two measure-valued tuples count as the same tuple (Problem 3 of [8]: completeness). It does not define grouping by an uncertain key or conditioning on an uncertain cell (Problem 4: the conditional cell). It takes the regular conditional kernel of the observation model as given and asks the schema to commit to a version. These are gaps, not oversights; each is stated precisely enough to be worked on, and none is required for the exhibit.

Performance. Codd’s performance argument had two parts: performance-oriented physical structures, and a compiler at least as good as the average programmer. The analogues are Problem 1 and Problem 5. Problem 1 asks for a class of finitely representable measures closed under the algebra with cost bounded output-sensitively through a k -way join; no candidate class is known to satisfy every condition, and the obvious candidates—point-plus-tolerance, parametric families, sample bundles, moment closures—each fail one leg. A stratified representation, exact where affine and approximate elsewhere, is the expected shape of the answer. Problem 5 is the optimizer. Mark each derivation node whose path to the root passes only through affine nodes; the marks form a frontier. Above it, execution is classical and exact. Below it, the affine nodes still carry mean and covariance through without loss and hand them to the second-order correction wherever a curved node sits. Read constructively, the theorem tells the optimizer where collapsing is free and what it costs at each node where it is not.

We do not claim performance. We claim that the specification says where performance is free and what it costs elsewhere, which is the input an optimizer needs and which point-valued practice cannot supply

because it has no frontier to find.

8. WHAT IS ASKED

Codd closed by saying why his title promised only a foundation: relational database dealt with the shared-data component of applications, and the rest needed complementary technologies. The measure-relational model is a foundation in the same sense and for the same reason, with one addition: it is a foundation without a building.

The specification is offered as work. Problem 7 of [8]—replicate the geomagnetic case end to end from public data, classically and measure-relationally, and specify a benchmark suite with known ground truth—requires no new theory and can begin now. Problem 1 gates the systems questions; Problems 2 through 6 are separable and of thesis scale; Problem 8, retrieval over embedding stores, is the nearest to current practice, since an embedding store already treats its domain as metric and its predicate as proximity; what remains is for it to treat its cells as measures.

The parallel with 1981 breaks at the point where it would be most convenient to hold. Codd could say that the practicality of the relational approach had been proven by installations already in operation. We can say only that the cost of the alternative has been priced, once, by a theorem and an exhibit, and that a data model which had carried the law would have kept the artifact's cause in the record rather than in a methods section—provided the observation's error model had been elicited, which the model asks for (Problem 4 of [8]) and cannot itself supply. Whether that is worth a system is the community's question. It is the reason this article exists.

Acknowledgments

Anthropic Claude Fable 5.1 was used to draft and revise the text of this article under the author's direction; the author is responsible for its content. The specification, theorem, and open problems are set out in [8]; the canonical reference for the measure-relational model is <https://realql.org>.

9. REFERENCES

- [1] C. A. L. Bailer-Jones. Estimating distances from parallaxes. *Publ. Astron. Soc. Pacific* 127(956):994–1009, 2015.
- [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] E. F. Codd. Relational database: a practical foundation for productivity (1981 Turing Award lecture). *Comm. ACM* 25(2):109–117, 1982. Reprinted with a postscript in *ACM Turing Award Lectures: The First Twenty Years 1966–1985*, ACM Press, 1987, pp. 391–410.
- [5] JCGM 100:2008. *Evaluation of measurement data—Guide to the expression of uncertainty in measurement* (GUM).
- [6] 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.
- [7] O. Kennedy, C. Koch. PIP: a database system for great and small expectations. *IEEE ICDE*, 157–168, 2010.
- [8] M. P. Laing. Measure-relational algebra: the core. <https://doi.org/10.5281/zenodo.22999911> (all versions); <https://realql.org>, 2026.
- [9] S. MacMahon, R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin, A. Dyer, J. Stamler. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 335(8692):765–774, 1990.
- [10] S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, R. Shah. Orion 2.0: native support for uncertain data. *ACM SIGMOD*, 1239–1242, 2008.
- [11] 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.
- [12] R. J. Telford, E. Heegaard, H. J. B. Birks. The intercept is a poor estimate of a calibrated radiocarbon age. *The Holocene* 14(2):296–298, 2004.