

A Criterion of Merit for Federal Evidence Building Sealed-Corpus Description Length, and What It Would Take to Identify the Dynamics of a Social System

James A. Bowery

September 2026

Abstract

The Foundations for Evidence-Based Policymaking Act of 2018 (P.L. 115–435) created access and governance for evidence building—an openness presumption for public data, a confidentiality framework for restricted data, and the secure-enclave machinery that became the National Secure Data Service—but it defined no criterion of merit. The gap is usually read as a question about data, about which body of evidence is the better one, and that reading is part of why it has not been closed. What the Act leaves undefined is a criterion for *model* selection: a rule for saying which account of the same record is better, computable without discretion. Data selection cannot settle itself. Admitting or excluding a source is defensible only against a statement of what the resulting account buys, so in the absence of such a statement both questions fall to expert judgment exercised inside privileged access, and a member of the public cannot contest a finding without obtaining the same access. This is not an abuse of the Act’s access provisions; it is those provisions working as written. We propose the missing piece: a single hard number, the total description length of a sealed corpus at instrument precision, scored by a custodian who runs submitted *programs* and publishes only the resulting bit count. The corpus is held fixed; what competes is the program. Curation therefore falls *inside* the object being ranked rather than beside it, which is how a model selection criterion settles the data question that cannot settle itself: a cleaning step, a source, an imputation is admitted when the account it makes possible costs fewer bits than the account without it. The data never move, so confidentiality is preserved exactly; the program is public, so the data-cleaning algorithm is scored rather than presumed; and replication becomes the act of submitting a rival program and beating the number. We then demonstrate the submission format at small scale on 3,146 United States counties observed at eight time points from 1970 to 2024 on 777 items, reporting the identified dynamics, an intervention operator warranted by measured invariance, the curation defects the exercise exposed, and two failures that instruct. The sole component of the fitted system that does not extrapolate is the sole component we supplied from outside it; and the contrasts a difference was supposed to make independent of that component turn out to be conditional on it—stable in sign across three panel versions, but not across two base years, and, by direct audit rather than by assertion, conditional in magnitude on which shared path is propagated, by a median of a third of the effect itself. Both were located by the method’s own diagnostics rather than by argument, which is the property a criterion of merit has to have. Both bound what the study may claim, to the four of seven reported outcomes whose direction survives. And both have the same remedy: let the aggregate be generated by the system rather than supplied to it—which we then built and tested, and which shortens the code on three paired seeds of four while recovering the aggregate it was denied in only two runs of six, though among those runs the shortest code does reliably mark the best held-out forecast.

1 The gap P.L. 115–435 leaves

The Evidence Act supplies access and process and no measure of merit. Title II establishes a presumption that government data are open and machine readable; Title III establishes a confidentiality framework and a presumption of accessibility for statistical purposes, and the committee it created recommended the secure data service now being built. All of this is machinery for getting data to analysts. None of it says what distinguishes better evidence from worse. The Act therefore delegates that judgment to peer review and agency expertise—both exercised inside privileged access, and neither contestable from outside it.

The consequence is structural rather than venal. A finding produced from restricted data cannot be checked by anyone lacking the same clearance, so it rests on the standing of its author. The most consequential empirical social science of the last two decades has this shape, and the customary remedy—release the data—is precisely what the confidentiality title forbids for the records that matter most. We should stop treating this as a tension to be managed and recognize it as a missing component: the Act built the library and omitted the scoreboard.

The stakes make the omission expensive. Federal transfer, tax, health and education programs turn on beliefs about how the social system responds when acted upon, and the sums involved run to the tens of trillions of dollars over a planning horizon. Where a quantity that large depends on an unfalsifiable claim, the incentive to keep the claim unfalsifiable is of the same order. A hard public number is the only instrument we know of that removes the appearance of interest as effectively as it removes the substance, and it does so symmetrically: it exposes an analyst whose model is beaten, and it protects one whose model is not.

A note on three overloaded words. Three words carry a methodological sense and a substantive sense at once. This is not bad luck: system identification and demography have grown their vocabularies independently, so a paper that applies the first to the second will keep colliding. We always qualify and never use any of them bare.

A *system state* is the dynamical object—a point in the state space whose motion the recurrence describes; a *US state* is the jurisdiction counties nest within, and which supplies one of the regimes across which invariance is tested. Watch for this in “state variable” and “state regime”, which belong to opposite senses.

A *family of models* is the class the description length is minimized within; a *family household* is a census unit, and family structure—married couples with own children, female-headed families with children—supplies several of the outcomes reported below. The paper’s most robust substantive finding and its frankest methodological limitation are both about a “family”, and they are not about the same thing. A related reservation, made once here: *regime* always means an environment across which invariance is tested—a decade, a US state—and never a policy or access arrangement.

A *panel version* is one of the three successive assemblies of the data, of 525, 770 and 777 items; a *generation* is people. This paper is about intergenerational causation and reports an outcome named the next generation’s bachelor’s share, so “generation” is reserved for that sense throughout and the versions are never called generations.

1.1 What this paper supplies

1. **A criterion of merit.** Total description length of a corpus at instrument precision [Rissanen, 1978]: a measurement in bits, with no tuning constant, no discretion in the scoring, and an unambiguous ordering.
2. **A mechanism that preserves confidentiality exactly.** Contestants submit programs; the custodian runs them inside the enclave and publishes only the bit count. No record is disclosed, and yet any outsider can contest any claim by beating its number.
3. **A rule that puts curation inside the scored object.** The corpus is held as close to the measurement instrument as law allows, so that cleaning is part of the submitted algorithm rather than an unexamined step upstream of it.
4. **A worked submission,** at a scale small enough to audit, with its failures reported at the same resolution as its results.

The substantive question we use to exercise the method—whether young men can outbid the labour market for the fertile years of the most economically valuable women, and with what consequence for the next generation—is a fair specimen of the class the Act exists to inform, and its results are reported in §11.

2 Why description length, and why not a statistical criterion

The criterion has to be a number a custodian can compute without judgment, that cannot be gamed by choosing a flattering test, and that measures understanding rather than fit. Total description length is the only candidate we know of with all three properties, and in the limit it is not merely a proxy for identification—it *is* identification, because the shortest program that generates a record is the mechanism that generated it [Solomonoff, 1964, Kolmogorov, 1965].

2.1 Two-part description length

Following Rissanen [1978], the best description of a body of data minimizes the total length of a description comprising the model and the data encoded with it. Writing θ for the parameters and δy for the resolution of cell y ,

$$L = L(\theta) + \sum_{\text{observed cells}} -\log_2(p(y \mid \theta) \delta y). \quad (1)$$

Both terms are bits, so they are commensurable and there is no free parameter trading them off. The sum runs over observed cells only: coding imputed values would let a submission earn bits for predicting its own guesses.

The factor δy is not decoration. Coding a real number to unlimited precision costs unlimited bits, so the criterion is well posed only once the data’s own resolution is stated. We carry a per-cell δy : a published margin of error divided by 1.645 for a survey estimate, the tabulation’s rounding step for a complete count, a propagated combination for derived quantities. For a federal benchmark this has a second virtue—it forces the custodian to publish the precision of its own instruments, which is itself evidence-building.

$L(\theta)$ is computed two ways: for accounting, an entropy code over quantized weight levels; inside the gradient, a differentiable code at the same quantization step, so the optimizer feels the cost of its parameters. Weights are rounded in the forward pass, because a model must be scored at the precision at which it would be transmitted.

2.2 Held-out regimes are an audit, not a second criterion

It is tempting to score submissions on a withheld regime—a period, a US state, a policy discontinuity—on the grounds that this is what tests whether a relation survives being acted upon. In a sealed benchmark the distinction collapses: the shortest program generating the record necessarily generates the part of the record the contestant has not seen, so code length over the full corpus *is* the held-out score. Reported separately, held-out performance is a finite-sample audit of whether a submission’s parameter code was honest within its chosen family of models, which is worth publishing but is not a second criterion.

This matters for how the present study should be read. Everything we report about the gap between compression and causal identification is a statement about a *bounded approximation inside a chosen family of models*, not about the criterion. Where the approximation failed, we can name the violation that caused it (§8); none of the failures is evidence against the underlying thesis, and we are at pains to say so because the objection that there is “no empirical evidence” for that thesis has been used to decline precisely the instrument that would produce the evidence.

3 The mechanism: programs in, one number out

Confidentiality and contestability are usually treated as opposed, and they are not. Submit *programs* rather than findings; have the custodian run them inside the enclave the Act’s confidentiality title already contemplates; publish only the resulting bit count. No record is disclosed, so the statutory protection is preserved exactly. The program is public, so every cleaning decision is legible. And replication—which for restricted-data findings is currently unavailable at any price—becomes the act of submitting a rival program and beating the number.

The properties that matter for a federal benchmark follow directly.

- **No discretion in scoring.** The custodian executes and counts bits. There is no protocol to negotiate, no specification to choose, no test set to select.
- **Adversarial verification without access.** An outsider who believes an official model is wrong does not need the data to demonstrate it. This is the specific failure of the present arrangement that the mechanism repairs: today such an outsider has no move at all.
- **Curation is scored.** Under the rule of §4 the corpus sits as close to the instrument as law allows, so a submission that began from someone’s pre-cleaned extract would be importing unscored judgment. The cleaning is in the program or it is nowhere.
- **Symmetry.** The mechanism exposes an analyst whose model is beaten and protects one whose model is not. An agency holding the best-scoring submission holds a public fact, not a claim of standing.

Three honest obstacles. The first is **linkage**: what makes restricted access valuable is joining tax to survey to program records, and that joining is governed by several statutes with

incompatible rules—so the benchmark needs the linked corpus to exist, which is a political question the mechanism does not answer. The second is **corpus selection**: which instruments enter at all is a decision no criterion over a fixed corpus can make, and it is the part of the consensus problem that remains genuinely open. The third is that **some agency processing is undocumented**. Here a trend that is usually reported as a loss is a gain: a release protected by a formally private mechanism publishes its noise parameters [Abowd, 2018], whereas one protected by swapping does not, so formal privacy makes agency processing modellable and therefore scorable.

3.1 The attack the mechanism must answer

A sealed corpus can be overfitted through its own leaderboard. Each published score is a measurement of the withheld data, so a contestant submitting repeatedly learns about the corpus without ever seeing it, and a sequence of adaptively chosen submissions can reach a good score by exploiting that leakage rather than by understanding anything. This is the adaptive data analysis problem [Dwork et al., 2015], and any benchmark proposed for policy consequence has to say what it does about it. Three answers compose.

- **Budget the submissions.** The leakage is bounded by the number of scores released, so a submission budget per entrant converts an open-ended attack into a quantified one.
- **Report the score through a privacy mechanism.** Publishing the bit count with calibrated noise preserves the ordering that matters while bounding what a sequence of scores reveals—the reusable-holdout construction, applied to the leaderboard rather than to the data.
- **Let the corpus renew.** This is the advantage a federal statistical corpus has over a static machine-learning benchmark and it is decisive: the agencies publish again every year, so the sealed portion regenerates on its own. A submission tuned to leakage from last cycle’s scores must face next cycle’s releases, and a method that only ever fitted the leaderboard will show it.

None of this requires new legislation. The Act already provides the enclave, the confidentiality regime and the mandate to build evidence; what is missing is a published number, a submission budget and an open queue, all of which are administrative.

4 Curation inside the scored object

The corpus should be held as close to the measurement instrument—mechanical or institutional—as law allows, so that every cleaning step is a line in the submitted program rather than a property of the data. This is not fastidiousness. Three defects in our own curation each changed published results and none was visible in goodness-of-fit: one reversed the sign of the estimated housing response; one had silently removed the entire income and poverty domain, 245 items, from a panel used to study economic pressure on family formation; one produced a county median household income of \$3.77 million. Every one had been applied *upstream* of the corpus, where a criterion cannot reach it.

4.1 Why the boundary is the whole question

Consider the first defect. Per-capita normalization was keyed on the source’s unit code, and dollars serve both aggregate payroll and median income, so four dollar medians were divided by population. Ask what an information criterion makes of this. If population is in the corpus, the corrupted column is an invertible function of the clean one given population, and the two corpora are the same object: we measure the difference at -0.0000 bits per cell. The criterion is *exactly* indifferent, and correctly so.

That indifference is the argument for holding the boundary at the instrument, not an argument against the criterion. As a step inside a submitted program the division is free, and the choice is legible in the code. As a step applied before the corpus it is unfalsifiable, and the column silently denotes something other than what its name says. Our panel was built the second way—the population denominator is consumed by the normalization and never appears as an item—and in that configuration the defect is not information-preserving but merely negligible: it costs 1.86 bits per cell on that column in 1980 and 1.49 in 1990, which is 0.04% of a 6.6 bit-per-cell total spread over 777 items. No optimization over that corpus would ever have surfaced it.

4.2 Construction and the two remaining defects

The unit is the United States county, keyed by US-state-plus-county FIPS: 3,146 counties at eight time points—1970, 1980, 1990, 2000 from the decennial census and the American Community Survey five-year windows 2006–2010, 2011–2015, 2016–2020 and 2020–2024—on 777 items. Of the $25,168 \times 777$ cells, 9,280,572 (47.5%) are observed. Values are stored as $\log(1 + x)/s_{95}$ with precision propagated accordingly.

Reference-year columns. The source records income and poverty by the year the quantity refers to: income collected in the 1980 census is income in 1979. A grid admitting only exact decade years silently dropped 245 items—the whole income and poverty domain. Admitting columns within one year of a grid point recovered them, taking the panel from 525 to 770 items, and roughly halved the estimated fertility effects of §11: the earlier estimates had been confounded by an economic variable the corpus did not contain.

Non-additive quantities aggregated. Connecticut replaced its counties with planning regions in 2022. Both geographies are exact unions of the same 169 towns, so counties rebuild by summation for counts and for shares of counts—but not for medians, which is how the \$3.77 million arose. We admit a reconstructed column only if the same reconstruction reproduces a *published* value in a period where one exists: 635 of 676 columns cleared, and Connecticut’s observed density at 2024 rose from 0.004 to 0.290 against 0.319 nationally. That procedure was then itself found defective, through a list misalignment that paired each column with another column’s error—which is what had let the \$3.77 million through.

Two lessons for a federal benchmark. These defects are invisible to the objective because a model fits whatever it is given; they surface only when a quantity is read in its own units by someone asking whether the number is *possible*, or when a derivation is forced to reproduce an answer already known. And a verification step is code and fails like code, so the benchmark must publish submissions in full rather than accept their authors’ account of them.

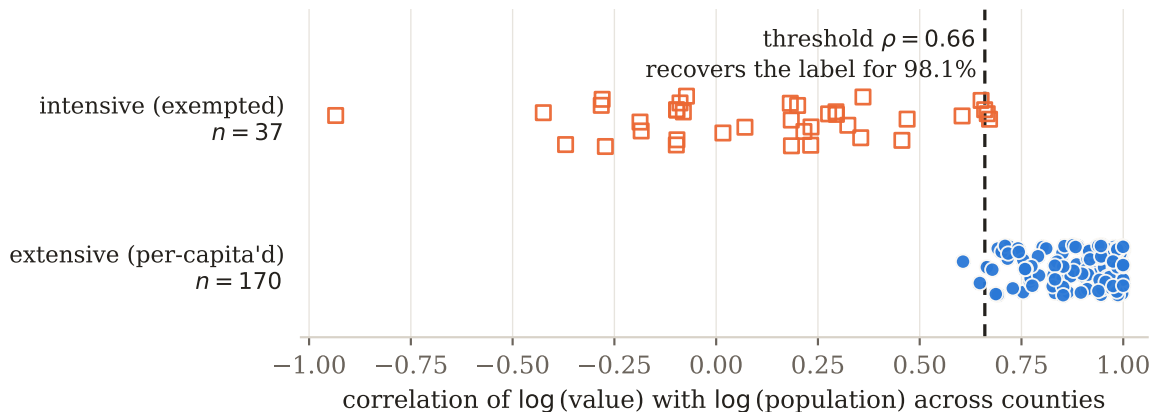


Figure 1: Whether a column is extensive is a property of the column, and the corpus can settle it. Each mark is one of 207 items measured at the 2000 point. Identity is carried by row, marker shape and direct label as well as hue.

5 Inducing the cleaning algorithm

The cleaning rules of §4 were hand-written, and two of them were wrong. That is an argument for deriving them rather than asserting them, and the derivation is available: whether a column is extensive is not a fact about the analyst’s intentions but a property of the column, and the corpus can settle it. We measure that a single threshold on how a column scales with population recovers the hand-written classification for 98.1% of items.

5.1 A cleaning rule is a hypothesis, not a convention

Take the first defect. Deciding whether to divide a column by population is deciding whether the quantity is extensive, and an extensive quantity is one that grows with the size of the unit. That is testable. Across 207 items measured at the 2000 point, the correlation of log value with log population separates the two classes cleanly: median $\rho = +0.946$ for items the hand-written rule treats as extensive against $+0.185$ for those it exempts. A single threshold, $\rho > 0.66$, reproduces the hand label for 98.1% of items.

Figure 1 shows the separation. The four disagreements are instructive rather than noisy. Two are counts of small, geographically concentrated minority populations, which grow with unit size more weakly than a general count; two are percentages that happen to rise with unit size because the composition they measure is urban. Both failures are visible, nameable, and would have been surfaced by the test—which is more than can be said for a regular expression that omitted the word “mean”.

5.2 Why the metadata belongs inside the corpus

The stronger move is to give the compressor the *documentation* as well as the numbers. The agencies publish, for every column, a unit indicator, a universe statement, a decimal resolution and a description in English. Each of the three defects was a failure to use information that this metadata already contained:

- whether a quantity is extensive is largely predictable from its unit indicator and the words “median”, “mean”, “average” or “per capita” in its description;
- the reference year of an income series is stated in its own description—“median household income in 1979”—so the mapping onto a survey year need not have been guessed at all;
- whether a quantity may be summed across geographies follows from its universe and unit: a count of households is additive, a median of them is not.

There is an information-theoretic reason to prefer this to an enumeration, and it is not the obvious one. On a *fixed* corpus, enumerating the exceptions is cheap: marking 10 intensive columns among 698 costs about 73 bits, less than the character count of a rule that finds them. Minimum description length over the data alone would therefore prefer the list. The rule wins on two other grounds. It generalizes to columns the corpus does not yet contain, which is the situation of any benchmark whose corpus grows. And once the metadata is itself part of what must be transmitted, the rule stops being an extra cost: the description strings are already in the corpus, so conditioning the normalization on them is free, and the rule becomes a statement about a regularity the corpus already exhibits rather than an assumption imposed on it.

For a benchmark this settles what the corpus is: not a table of numbers but the agency’s release as published, documentation included. A submission’s cleaning is then a conditional model of data given metadata, priced in the same bits as everything else, and testable in the way §4 could not test a hand-written regular expression—which is how the same class of error escaped twice.

6 Model

The model is a manifold autoencoder with a mixture output head, fitted by (1), together with a recurrence on its latent state. The one design choice that is not conventional is the closure: the recurrence must code the encoder’s own embedding of the next period, not merely that period’s observations. Without it, one-step prediction looks fine and multi-step propagation is meaningless—which matters here because intervention results are propagated for decades.

An encoder maps a unit’s cross-section, with unobserved inputs entering at the feature mean, to a latent state $z \in \mathbb{R}^{32}$; a decoder returns, for each item, a three-component Gaussian mixture. The mixture matters: it lets the residual term be an honest code length for asymmetric and multi-modal item distributions without transformation tricks. Weights are initialized so that the untrained output is exactly the per-item null—each item an independent Gaussian at its own mean and variance—so epoch zero scores the null and every subsequent bit is earned from a baseline the model itself reproduces. A fraction of observed inputs is masked during training while still being coded, which prevents the encoder from copying inputs to outputs.

Coding a unit’s own cross-section through the latent and back, at each cell’s precision, we call the *pullback*. Its residual measures how much structure the manifold has captured, in the data’s own units.

6.1 The recurrence, and why it must be closed

A recurrence carries z from one period to the next,

$$z_{t+1} = z_t + \Delta t (c_{t,s} + f(z_t)), \quad (2)$$

with $c_{t,s}$ a displacement shared by all units in period t and US state s , f built from a library of constants, linear terms and pairwise products—sparse selection from a function library follows Brunton et al. [2016]—and Δt the step length so that both are rates.

Fitted only against next-period observations, such a recurrence can score well one step ahead while its output lands nowhere the encoder ever puts a unit. The decoder is far from injective, and with half the target cells unobserved the set of latent points decoding acceptably is a wide region. Iterating then evaluates f outside its training distribution. We therefore require the recurrence’s output to code the encoder’s own embedding of the next period,

$$\text{cons} = \frac{n}{2} \sum_k \log_2(2\pi e s_k^2), \quad s_k^2 = \langle (z_{t+1}^{\text{pred}} - z_{t+1}^{\text{enc}})^2 \rangle, \quad (3)$$

the bits to transmit the encoder’s embedding given the recurrence’s guess, in the same currency as the rest.

The closure costs nothing and buys the multi-step results: the closed model predicts *better* than the unconstrained one (8.43 against 8.76 bits per cell in the panel version where both were fitted) at equal pullback, and two-step error grows as $\sqrt{2}$ times one-step error rather than compounding. Requiring the recurrence’s output to be a system state, and not merely to decode like one, turns out to regularize it.

7 Identification and invariance

Decade motion separates into a component shared across units and a unit-specific remainder, and the two behave oppositely. The unit-specific map is invariant across every decade and US-state regime tested. The shared component does not extrapolate at all. That asymmetry is what licenses interventions to be posed as contrasts, which remove the shared component’s explicit term from the difference equation—though not, as §9.1 shows, entirely from the result—and, as §8 argues, it is also a self-inflicted wound rather than a property of the world.

In the present panel the shares of squared motion are 0.79 national, 0.05 US state and 0.16 unit-specific, the national share inflated by nominal-dollar items moving with the price level.

The shared component’s failure is quantitative: parameterized as a function of national time series and projected onto a withheld decade, it coded that decade at 27.9 bits per cell against 14.0 for assuming no shared motion at all. Worse than nothing.

The unit-specific map is a different object. Fitted in quadratic form it is invariant across all six decade regimes and, in linear form, across all 41 US-state regimes, scored against an effective sample size accounting for spatial and temporal correlation. Linear-in-decades fits are *not* invariant, which we read as curvature mistaken for time-variation when the form is forced to be linear.

A relation that holds unchanged across regimes is a candidate causal relation [Peters et al., 2016]; it is the property Frisch [1938] and Haavelmo [1944] named autonomy, and the property Lucas [1976] observed a relation must have to survive being acted upon. We have it for the unit-specific map and demonstrably lack it for the shared displacement.

8 The exogeneity we smuggled back in

The single component of this system that fails to extrapolate is the single component we supplied from outside it. That is not a limitation of the method; it is the method’s central claim being confirmed by our own violation of it.

The shared displacement is either a free parameter per period—an unexplained constant, which is exogeneity in its purest form—or a linear map from national principal components that we *import* from published macroeconomic series. Having refused exogenous variables at the county level, we reintroduced them at the aggregate, and then measured exactly the failure the autonomy literature predicts for a relation that is not autonomous.

The remedy is to close the system, and it is concrete. Of the twelve national series used, eight—unemployment, three labour-force participation measures, median house price, total fertility, adolescent fertility—are aggregates of the county cross-section and should be *computed from the system state* rather than supplied to it. The remaining four—consumer prices, real output per capita, the long mortgage rate, federal debt relative to output—are monetary, fiscal and price-level quantities genuinely outside a county panel, and the consistent treatment is not to import them but to promote them to system-state variables carrying their own dynamics. A closed system has no shared displacement to extrapolate, because the aggregate is generated rather than assumed.

We report this prominently because of how a study like this is liable to be read. Taken at face value, “the component that compresses well does not extrapolate” looks like evidence that compression is causally insufficient. It is nothing of the kind. It is evidence that we gave a compressor a record it was forbidden to explain, and it declined to explain it.

8.1 We built the closed system: a reliable in-sample win, an unreliable out-of-sample one

Closing the system shortens the code every time we try it and does not, on this panel, produce a model that forecasts better or that recovers the aggregate it was denied. Both halves of that sentence are measured, and the second is the reason the remedy stays a proposal.

The implementation replaces the imported basis with a generated one. A national system state u_t starts from a free parameter and evolves under

$$u_{t+1} = u_t + \Delta t [1, u_t, \bar{z}_t] A, \quad (4)$$

where $\bar{z}_t$ is the cross-county mean latent—the system’s own aggregate—and the shared displacement reads u exactly as it previously read the imported components. Seventy-two parameters replace twelve numbers the open model receives as a fixed input and never pays for in the code.

In sample it wins, three times out of four. Paired by seed against the open model, closing the aggregate changes the total code by -1.8 , $+0.5$, -5.2 and -3.4×10^5 bits. Run-to-run spread at fixed architecture is 5.2×10^4 bits, so the three wins clear the floor and the one loss sits at it. By the criterion this paper proposes the closed system is the better model of the corpus, and it is better while also paying for an aggregate the open model was given—but not on every seed, which is the first sign of what §8.1 ends on.

These are the code lengths the diagnostics evaluate: clean encoder input, quantized weights throughout, in evaluation mode. They are not the quantity the optimizer minimizes, which charges the prediction branch at unquantized precision and corrupts the encoder input with the denoising mask. That training objective is the right thing to descend and the wrong thing to report—it is not a transmittable code length—and the two can disagree by more than 10^5 bits at the same checkpoint. Every figure in this section is the transmittable one.

Out of sample it does not. Holding out the 2020–2024 transition, the open model codes it at 4.1316 bits per cell and six closed runs average 4.1726—one per cent worse, with *one* of the six beating the open model. That comparison is unfair to the closed model, and the unfairness is the finding: at prediction time the open model is handed the 2020 national system state, which is the one input a genuine forecast cannot have. Read as a trade rather than a contest, closure costs about 0.05 bits per cell and buys a system that needs nothing from outside itself.

And the aggregate is recovered only sometimes. In a held-out run the generated u_{2020} is produced from u_{2015} and the 2015 aggregate and enters no training loss, so where it lands is a genuine forecast of the national system state. Mapping the generated coordinates onto the published components on the fitted decades and applying that map at 2020, two of six runs beat both naïve baselines—persistence and linear extrapolation of 2010–2015, at errors of 1.155 and 1.134 in units of the components’ own spread—at 0.395 and 0.579. The other four are worse than both. The county cross-section evidently contains its own aggregate in some fits and not in others, and the difference is the seed.

The criterion does rank the runs, once it is a code length that does the ranking. Across the six held-out runs, rank correlation between the transmittable code length at a checkpoint and that checkpoint’s held-out bits per cell is +0.83. The shortest-coding run has the best held-out prediction and the only win over the open model; the longest-coding run has the worst. Description length earns its keep here, and it earns it on the comparison it is supposed to lose—among fits of one architecture, where nothing but the optimizer’s path distinguishes them.

We first measured this wrong, and we report the error because it is a demonstration of §3 rather than an aside about carefulness. Ranking the same six runs by the quantity the optimizer *descends* gives -0.31 against held-out bits and -0.77 against aggregate error—an apparent refutation of the criterion, by the criterion’s own proponents. That quantity is not a code length: it corrupts the encoder input with the denoising mask and charges the prediction branch at unquantized precision. Both are correct for an objective and disqualifying for a score.

The two roles the mechanism separates are exactly the two we conflated. A contestant optimizes whatever they like; the custodian runs the submitted program and computes the description length itself, from the sealed corpus and the program’s output, and never sees the objective that was descended. Under that arrangement this error cannot reach a scoreboard, because the number on the scoreboard is not the contestant’s to compute. We made it because in a study one writes alone there is no custodian—the same hand that trains the model reports its score, which is the configuration the proposal exists to abolish, and which is the configuration nearly all published quantitative social science is in.

Two further properties of the criterion are on display, and both are claims §3 makes in the abstract and can now make concretely. The discrepancy was *findable*: it was found by

recomputing bits from a checkpoint, which anyone holding the program and the corpus can do, and which is what “contestable without privileged access” means operationally. And the criterion is *sharp* in the way an ordering has to be if it is to decide anything—the two quantities differ by at most 0.13% of the total, and that was enough to reverse the sign of a reported correlation and to move one architectural comparison from four wins to three. A criterion loose enough that a tenth of a per cent could not overturn a finding would not be worth sealing a corpus for.

What the code length does *not* track is the other test. Against how well a run reconstructs the published national components the correlation is +0.14, which is nothing. That asymmetry is not mysterious: held-out bits score the decoded county cross-section, which the code charges for directly, while u_{2020} is an extrapolation of the national recurrence that no term in the code ever scores. The criterion ranks what it is paid to rank.

So §8’s remedy stands where the evidence leaves it: built, shortening the code on three seeds of four, selecting reliably among its own fits, and recovering the aggregate it was denied in two runs of six. The right thing to build, and not yet delivering what it was built for.

9 Intervening: $\text{cause}(C)$

$\text{cause}(C)$ is the minimum-norm displacement of a unit’s latent system state that realizes condition C in the decoded observables, with every other observable co-moving as the manifold requires, propagated by the identified recurrence and decoded with intervals. Nothing is severed and there is no graph. Every result carries two diagnostics—how much of the requested change survives re-encoding, and what fraction of units end where data exist—and interventions are posed as contrasts, which remove the non-extrapolating shared displacement from the difference equation, though §9.1 bounds what survives it.

Because the system state is the whole cross-section, a condition on one item necessarily moves others; those co-moves are reported rather than assumed away. They are the honest content of refusing exogeneity: in a closed system the co-moves are not contamination to be purged, they are what intervening consists of. Minimum norm is a stated choice—it encodes an assumption that an intervention disturbs the system state as little as it can—and the data do not establish it.

We write *cause* rather than *do not* in opposition to Pearl [2009] but because the object assumed autonomous differs. Pearl’s operator formalizes autonomy for the case where the invariant structure is a graph, and obtains its warrant from the graph’s correctness; here the invariant structure is a recurrence, there is no edge to delete, and the warrant is the measured invariance of §7. One asymmetry favours the recurrence: its autonomy is testable out of sample, and we test it.

The region diagnostic is not decorative. A result whose intervened system state is largely outside the region the data occupy describes units that do not exist; it is reported flagged and not quoted. In this application it disqualified two of the eight contrasts attempted (§12).

Interventions are posed as *contrasts*: move one item while holding another at its factual value. Substantively, a single-item move on this manifold drags the whole labour market with it. Technically, a contrast is a difference between two system states propagated by the same recurrence, and the explicit shared term cancels from that difference. Writing the baseline as z_t and the intervened system state as $z_t + d_t$,

$$d_{t+1} = d_t + \Delta t [f(z_t + d_t) - f(z_t)], \quad (5)$$

outcome at +30 years	from 1980		from 1990	
	fitted path	range	fitted path	range
female-headed families with children	+10.7	[+4.2, +11.9]	+14.4	[+12.7, +14.5]
never-married women	+5.8	[+3.9, +6.0]	+8.3	[+7.9, +8.6]
next generation’s bachelor’s share	+3.2	[+2.7, +3.5]	+4.1	[+3.4, +4.3]
median home value	+4.3	[−0.5, +4.3]	−1.4	[−4.9, −1.4]
births per 1,000	+1.9	[+1.7, +2.7]	−1.3	[−1.4, −1.1]
children under 5	+1.4	[+0.5, +2.0]	−1.8	[−1.8, −1.3]
married couples with own children	−1.2	[−2.5, −0.8]	−5.7	[−5.8, −5.2]

Table 1: The primary contrast re-read under the four shared paths that keep both trajectories in region, per cent change at the median unit; the first column of each pair is the fitted path. Signs hold except for median home value from a 1980 base.

in which $c_{t,s}$ does not appear.

9.1 What does not follow

It does not follow that the contrast is independent of the shared displacement. The base trajectory z_t depends on every shared displacement applied before t , and for a quadratic f the bracket in (5) is $J(z_t)d_t + Q(d_t)$, which depends on z_t through the Jacobian. The shared path therefore reaches the contrast by two indirect routes: through the latent difference, and through the decoder, where an effect reported as a percentage is divided by a baseline that the shared path moves.

We bound both. Doubling the shared displacement at every step for thirty years—far outside any plausible path—moves the propagated latent difference by 2.8%, and by the same 2.8% at intervention magnitudes of 0.05, 0.10 and 0.20, so this route is linear and scale-free. One decade of the fitted displacement changes the growth rate $\|J(z)d\|$ by 9.1% (10th to 90th percentile 7.2–12.6%), which over three steps accumulates to the 2.8% above.

The second route is bounded by direct audit, which is the only way to reach it: it runs through a nonlinear decoder and a percentage baseline, so there is no inequality to write down. `shared_path_audit.py` freezes the checkpoint, the base-year cohort, the intervention and the reporting definitions, and varies only the shared path, re-reading the contrast under eight paths: as fitted; the national component with the US-state component removed; zero; doubled; reversed; the mean of the six fitted per-decade rates; and the 1980 and the 2020 rate each held for the whole run. Each path is applied identically to both branches, and the script asserts that the realized intervention at $t = 0$ is unchanged across paths, so a difference in the answer is the indirect dependence and nothing else. These are sensitivity paths, not forecasts.

Two paths carry the trajectories out of the data region and are dropped before anything is read from them: reversing the shared displacement, and holding the 2020 rate for a decade. The second is worth stating plainly—the most recently fitted shared rate, sustained, takes every unit off the manifold within one step. That is the non-extrapolability of §8 appearing as a support failure rather than as a confident wrong number, which is the behaviour the region diagnostic exists to produce.

On the paths that stay in region, magnitudes move by a median of 0.32 times the effect the fitted path gives. The contrasts are therefore conditional on the shared path in size, and every

magnitude in §11 is to be read as conditional on the fitted one. The signs are another matter: across two base years, two contrast directions, seven outcomes and three horizons—84 cells—79 hold the sign the fitted path gives. Four of the five that do not are median home value; the fifth is a births effect of -0.26% .

Median home value is a systematic exception rather than a scattered one, and the reason is legible. Decoding the fitted shared displacement at the mean system state and reading which items it moves, dollar-denominated items carry 37.6% of its motion while being 17.4% of the panel—over-represented by a factor of 2.2, and rising from 27–39% before 2010 to 41–51% after it. Eight of the thirteen largest movers are dollar items: median home value twice, per capita income twice, median household income twice, gross rent, owner costs. Setting the shared path to zero therefore does not perturb the housing contrast so much as remove a large part of what it is made of: the primary contrast on median home value goes from $+\$4,228$ to $-\$605$ at the median unit. No real-quantity outcome behaves this way.

The same decoding says what else the shared displacement is made of, and the answer bears directly on §8. Its five largest movers are population change, population density, women’s and men’s unemployment rates, and families below poverty: population growth and the business cycle, with the price level behind them. Those are three components, and the remedy already sorts all three—growth and unemployment are aggregates of the county cross-section and should be computed from the system state; the price level is genuinely exogenous to a county panel and should be promoted to a system-state variable carrying its own dynamics. We did not choose that split to fit this measurement; the measurement is what the split predicts. This is the same outcome §11.1 finds changing sign between base years. It was already outside the set we state without qualification; what the audit adds is the reason, which is not a general fragility of the method but a specific one—a nominal contrast cannot be separated from a shared path that is mostly the price level.

The cancellation of the explicit term is why the method yields contrasts rather than forecasts: the component that would have to be extrapolated for a level prediction never enters (5) at all. What remains is a weaker and measurable dependence, which the contrasts below are conditional on, and which §8’s remedy would remove entirely—a closed system has no shared displacement to choose.

10 Validation

On transitions withheld entirely from fitting, the model sits 0.43 to 0.52 bits per cell above the attainable floor, and that margin is flat whether the step is four years or ten. We report this as the audit described in §2, and we caution against the summary statistic most natural here, because a federal benchmark that adopted it would be ranking the wrong thing.

Skill is not comparable across transitions: its denominator varies more than fourfold, because assuming no change over a decade fails catastrophically while assuming no change over four years does not. The 0.965 reports a collapsed baseline, not an easy problem. Excess over the attainable floor survives the comparison, and its flatness across step lengths is the strongest evidence here that the recurrence captured something of the process rather than of the sampling: nothing in the fitting arranges for a four-year and a ten-year step to be equally well predicted.

Against the per-item null the model codes the panel at 6.64 bits per cell versus 9.29 and predicts the next period at 6.42 versus 9.08. Permuting the correspondence between units and

held out	step	oracle	model	persistence	skill	model – oracle
2000–2010	1.0 dec	6.404	6.871	19.730	0.965	0.467
2010–2015	0.5	3.554	3.983	4.607	0.592	0.429
2015–2020	0.5	3.528	3.964	4.860	0.673	0.436
2020–2024	0.4	3.606	4.128	5.006	0.627	0.522

Table 2: Held-out transitions, bits per cell. *Oracle* codes the target period from the encoder’s own embedding of it and is the attainable floor. *Persistence* decodes the base period’s system state. Skill is $(b_{\text{pers}} - b_{\text{model}})/(b_{\text{pers}} - b_{\text{oracle}})$.

outcome at +30 years	525-item	770-item	777-item
female-headed families with children	+15.6	+9.4	+10.7
never-married women	+8.6	+6.8	+5.8
next generation’s bachelor’s share	+8.0	+4.9	+3.2
median home value	+4.0	+5.0	+4.3
births per 1,000	+4.4	+3.0	+1.9
children under 5	+2.0	+2.2	+1.4
married couples with own children	−5.6	−1.6	−1.2

Table 3: cause(women’s labour force participation up, men’s employment held), one cross-unit standard deviation, per cent change at the median unit, across three successive panel versions.

their latent codes costs 46 bits per cell. Nominal 90% predictive intervals cover 92% of observed cells.

11 Application, and what stability looks like

The motivating contrast is stable in sign across three panel versions differing by 252 items and a time point—twenty-one sign agreements out of twenty-one—and its mirror reverses family structure cleanly. Two things qualify that. Its magnitudes are not converged: five of seven attenuate monotonically as identification improves. And stability across panel versions is not stability across base years: propagated from 1990 rather than 1980, three of the seven outcomes change sign. Four of the seven survive both tests—married couples with own children, female-headed families with children, never-married women, and the next generation’s bachelor’s share—and they are the only ones we state without qualification.

Table 3 reports the contrast. The mirror—men’s employment up, women’s participation held—reverses family structure: married couples with children +5.6%, female-headed families with children −9.9%, never-married women −3.9%. That the same three variables move oppositely at comparable magnitude under the mirrored contrast is the strongest internal check the method affords.

The attenuation is not mechanical. The intervention is identically realized and fully in region in all three panel versions, and the per-outcome ratios are strongly heterogeneous—0.21 for married-with-children, 1.08 for home value—where a scaling artifact would be uniform. The outcomes that shrank most are the fertility and marriage ones, and the panel version that shrank them is the one that admitted the income and poverty domain.

11.1 Stable across panel versions, not across base years

Sign stability across panel versions is a statement about identification improving while the answer stays put. It is not a statement that the contrast is independent of the trajectory it starts from, and propagating the same intervention from a 1990 base rather than a 1980 base shows that it is not.

outcome at +30 years	from 1980	from 1990	ratio
married couples with own children	−1.2	−5.7	4.75
female-headed families with children	+10.7	+14.4	1.35
never-married women	+5.8	+8.3	1.43
next generation’s bachelor’s share	+3.2	+4.1	1.28
births per 1,000	+1.9	−1.3	−0.68
children under 5	+1.4	−1.8	−1.29
median home value	+4.3	−1.4	−0.33

Table 4: The same contrast propagated from two base years, same model, same reporting. Three of seven outcomes change sign.

Three of seven change sign: births, children under five and median home value. It is tempting to explain this by magnitude—that the survivors are the large effects and the casualties the small ones—and the table does not support it. Married couples with own children survives at −1.2%, while median home value reverses from +4.3%, the third largest effect reported. What the four survivors have in common is that they are composition measures, shares of a population or of a household stock; what reverses is a flow rate, an age-specific count and a price. We record the pattern without a mechanism for it.

This is the second route of §9.1 made visible: the base trajectory differs between the two runs, and with it the baseline each percentage is divided by.

Two conclusions, and the second is the one that should govern how these numbers are used. The attenuation across panel versions is most likely confounding removed—the fertility effects shrank in the version that admitted the income domain, which the panel did not contain until the defect of §4 was repaired. But the contrasts are *conditional on a base year and a shared path*, and the four surviving directions are the only ones this paper is entitled to state without that qualification. Producing them with a band over plausible base years and shared paths is unfinished work, not a refinement.

12 What the method does not establish

The method establishes directions, not magnitudes; relations among counties, not among people; and contrasts, not forecasts. Two of the eight contrasts attempted are disqualified by its own region diagnostic. We set these out as a numbered part of the argument rather than as a closing hedge, because each is a measured property of the results rather than a caveat of manners.

Ecological inference. Every effect is an effect on a county, not on a person [Robinson, 1950]. A county where women’s education and childlessness both rise says nothing directly about whether the educated women are the childless ones. Stating this does not dispose of it.

Level forecasts. Propagating levels requires the synchronized displacement, which §7 shows does not extrapolate. Across panel versions a decade-ahead forecast of the same quantity moved by a factor of five for births and changed sign for female-headed families. A contrast removes that component from the difference equation; it does not make the result independent of it (§9.1).

Interventions outside the data region. A one-standard-deviation move in correctly scaled income leaves the manifold: the intervened system state is in region for 14% of units. Such results are reported flagged and are not interpreted. The diagnostic disqualified two of the eight contrasts attempted.

Residual dependence on the shared path. A contrast removes the shared displacement’s explicit term from the difference equation but not the trajectory dependence behind it (§9.1). The latent route is bounded at 2.8% under a doubling stress test; the decoding route is bounded by audit at a median of 0.32 times the fitted-path effect in magnitude, with 79 of 84 sign cells holding across the shared paths that stay in region. Magnitudes are conditional on the shared path. Separately, §11.1 shows three of seven outcomes changing sign between two base years; four of the seven survive both tests, and median home value fails both for the same reason—it is a nominal item and the shared path is mostly the price level.

Magnitude. See §11. The direction of an effect is established as firmly as this method can establish it; its size is not.

Choice of the family of models. The two-part code is minimized within a chosen family of models, and within that, a chosen architecture. Neither choice is free, and the criterion cannot tell us what a different family of models would have found.

Three of these five are themselves measurements—the region fraction, the forecast divergence, the attenuation sequence—which is the property a benchmark should demand of every submission: that its boundaries be located by the same instrument as its results. A submission that reports where it fails, in the units it reports its successes, is one a custodian can rank; one that does not is a claim of standing.

13 Conclusion

P.L. 115–435 built the library and omitted the scoreboard. The access and confidentiality machinery it created is sound and is in place; what it lacks is a rule for saying which of two accounts of the same record is better—a model selection criterion, from which the data question follows—and in that vacuum both are settled by authority exercised where the public cannot follow. A criterion exists that does not require the public to follow: total description length of a sealed corpus, with programs submitted and only the bit count returned. It preserves confidentiality exactly, it scores the data cleaning rather than presuming it, and it converts replication from a privilege into an act anyone can perform.

We have demonstrated the submission format at a scale small enough to audit, and reported its failures at the resolution of its results: three curation defects invisible to the objective, one

verification step that was itself defective, and two failures of the dynamics that are worth more than the successes around them.

The first is that the only component of our fitted system that does not extrapolate is the only component we supplied from outside it. The second is that a contrast, which removes that component’s explicit term from the difference equation, does not thereby become independent of it—three of seven reported outcomes change sign when the same intervention is propagated from a different base year. Read carelessly the pair looks like evidence against the premise. Read correctly it is the premise operating: a relation that is not autonomous fails exactly where autonomy is what was needed, and the failure is visible because the method measured for it.

Three things follow, and they are what make the failures useful rather than merely honest. They locate the boundary of the claim in the same units as the claim itself, so this paper states the four surviving directions and withholds the three that reverse. They were found by the diagnostics rather than by an objector, which is the behaviour a criterion of merit must produce if it is to be trusted where no objector has access. And they converge on one remedy—close the system, let the aggregate be generated rather than assumed—which removes both at once and is the obvious next piece of work.

The sums that turn on getting these dynamics right run to the tens of trillions of dollars. An incentive of that size attaches equally to getting the answer right and to ensuring that no hard number can ever show the answer was wrong. That is the strongest argument we know for adopting a criterion whose scoring admits no discretion, and for adopting it before rather than after the next occasion on which the public is asked to take an unfalsifiable claim on trust.

Reproducibility. The worked submission—code, panel construction, every model checkpoint and the diagnostic logs—is available at the project repository, and each reported figure prints the invariance verdict and held-out performance of the model that produced it.

A Architecture sensitivity

The criterion is minimized within a chosen family of models, so the obvious question is how much of the result is the class. Table 5 reports a sweep on a 90/10 split of the county cross-section, varying latent width, mixture components, hidden width and the reconstruction head.

run	latent	components	hidden	total (Mbit)	bits/cell
L8_K3	8	3	512,256	35.155	9.378
L16_K3	16	3	512,256	35.001	9.260
L32_K1	32	1	512,256	35.323	9.642
L32_K3_norec	32	3	512,256	34.992	9.344
L32_K3_wide	32	3	1024,512	35.165	9.389
L32_K5	32	5	512,256	34.797	9.308
L64_K3	64	3	512,256	34.921	9.281
L128_K3	128	3	512,256	35.079	9.393

Table 5: Architecture sweep. Total description length varies by 1.51% from best to worst.

Two things follow, one reassuring and one not. The spread across a sixteen-fold range of latent width is 1.51% of total bits, so the reported code lengths are not an artifact of a lucky

architecture; and a single mixture component is the worst configuration, which is the mixture head earning its place. What does *not* follow is that the panel model’s architecture is optimal—this sweep was run on the cross-sectional model and the panel model inherited its settings rather than repeating the search. A submission to the benchmark of §3 would have no reason to inherit anything, which is the point of scoring rather than asserting.

Restarts are search, and search belongs inside the program. Fitting is not convex, so where two architectures are compared we run each from several seeds and report the shortest. That raises an obvious objection—the number of restarts looks like a parameter nobody paid for, and a minimum over K tries is a biased estimate of what a model can achieve. The objection dissolves in the mechanism and survives in the paper, and it is worth separating the two.

In the mechanism of §3 there is nothing to correct, because nothing is being estimated. The custodian runs a submitted program and publishes the bits that program achieved. A program that restarts internally spends its submitter’s compute and returns one number; the criterion never asks what code length was *attainable*, only what this program attained. That is a quiet advantage of scoring programs over estimating an infimum, and it is the same property that lets curation sit inside the scored object.

Within this paper the objection stands, and we handle it the only honest way: K is matched across the architectures being compared, runs are paired by seed, and the spread is reported alongside the difference. Run-to-run spread at fixed architecture is 5×10^4 bits, about 0.005 bits per cell, which is the floor any architectural claim here has to clear.

That restarts are needed at all is a fact about the objective rather than a nuisance: the architectures whose aggregate is generated rather than supplied have a markedly rougher surface than those whose aggregate is handed to them, with five times the seed spread. Fixed restarts are the crudest possible response to that, and the adaptive-reannealing literature [Ingber, 1989] is the natural place to look for a better one. We note the connection and do not pursue it: changing the optimizer mid-study would confound the architectural comparison this section exists to make.

Data availability

Every source is public and no restricted data were used. *USA Counties* (U.S. Census Bureau) supplies the 1970–2010 backbone; American Community Survey five-year detailed tables for 2006–2010, 2011–2015, 2016–2020 and 2020–2024 are retrieved from data.census.gov; county-subdivision tables for Connecticut come from the same endpoint. Harmonized long-run series are IPUMS NHGIS extract `nhgis0001` (time series tables BX3, A34, BM4, A67, BS4, AG4, B79, AB2, BD5), which requires the citation: Schroeder, Van Riper, Manson, Cooper, Krause, Kugler, Zhu and Ruggles, *IPUMS National Historical Geographic Information System: Version 21.0*, IPUMS, 2026, doi:10.18128/D050.V21.0. County population estimates are from the Census Bureau’s Population Estimates Program (vintages as cited in code); single-year-of-age county populations from the National Cancer Institute’s SEER program; county student-debt measures from the Urban Institute. Code, the panel’s construction, every model checkpoint and the diagnostic logs are in the project repository; each figure prints the invariance verdict and held-out performance of the model that produced it.

Note on assisted authorship

The analysis, code and prose were developed in an extended collaboration with a large language model acting as a research assistant under the author’s direction. Every numerical result in this paper was produced by code in the repository and can be regenerated from it; the measurement defects reported in §4 were introduced, and found, in that process, which is part of why they are reported.

References

- Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. In *Proceedings of KDD '18*, 2867.
- Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113(15), 3932–3937.
- Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248), 636–638.
- Frisch, R. (1938). *Statistical versus Theoretical Relations in Economic Macrodynamics*. Memorandum, League of Nations.
- Gács, P., Tromp, J., and Vitányi, P. (2001). Algorithmic statistics. *IEEE Transactions on Information Theory*, 47(6), 2443–2463.
- Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.
- Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12 (Supplement), iii–115.
- Ingber, L. (1989). Very fast simulated re-annealing. *Mathematical and Computer Modelling* 12(8), 967–973.
- Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1–7.
- Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46.
- Ljung, L. (1999). *System Identification: Theory for the User*, 2nd ed. Prentice Hall.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press.
- Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction. *Journal of the Royal Statistical Society B*, 78(5), 947–1012.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.

- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.
- Solomonoff, R. J. (1964). A formal theory of inductive inference, Parts I and II. *Information and Control*, 7(1), 1–22; 7(2), 224–254.
- Wallace, C. S., and Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, 11(2), 185–194.