A Touchstone Information Criterion for Federal Evidence Building

Sealed-Corpus Learned Algorithm Length, and What It Would Take to Identify the Dynamics of a Social System

James A. Bowery

September 2026

Abstract

Federal agencies rank competing accounts of the evidence by the standing of their authors, because the Foundations for Evidence-Based Policymaking Act of 2018 (P.L. 115–435) supplies access to data and a process for using them and no touchstone information criterion. This paper proposes one and rests it on two claims. First, causal inference from observational data should begin from a law of motion, the rule by which a system moves from one time point to the next, and treat statistical information criteria over the same data as ranking kinematic descriptions. The literature, invariant causal prediction among its best instances, does the reverse: it identifies which statistic survives a change of environment, where the object to identify is the recurrence that generated the data. No causation without dynamics. Second, the literature cites the many-analysts design, one corpus and one question given to many analysts, and does not adopt it, because the design has no criterion and yields only the spread of its analysts' conclusions. The remedy is to charge the analysts' discretion. The gold ideal of causal induction is Solomonoff's, the shortest algorithm possible that generates a sealed corpus at the precision of the instruments that measured it. The computable test is the touchstone information criterion, the length in bits of a learned algorithm that generates the corpus: it scores the shortest algorithm attained, never the shortest possible, so it ranks any two accounts of the evidence side by side with no gold in hand. A custodian runs each submitted learning algorithm inside the restricted-access environment the Act provides, measures that length without judgement, publishes only the length, and discloses no measurement event. The spread of conclusions becomes an ordering, and any party contests a finding by submitting a better learning algorithm. The criterion may also prove useful within the privileged access a tier of custody enjoys over its own data selection before it seals the corpus it publishes: that tier then pays for every cleaning step it takes in the bits of its own internal learned algorithm. We demonstrate the format on 3,146 United States counties observed at eight time points from 1970 to 2024 on 777 items, with an intervention operator, cause(C), defined by the fitted recurrence alone. We asked one question of it: what follows for family structure when women's labour-force participation rises with men's employment held fixed. Of seven outcomes read, three keep their sign under all five of the paper's tests: female-headed families with children rise, never-married women rise, and married couples with own children fall, and the reverse contrast reverses all three. Of the other four, two crossed zero as the model improved across panel versions and two change sign between base years, and we assert no direction for them.

Part I
The methodology

1 Two claims

The first claim is about the order of foundation and addition. The literature on causal inference from observational data, invariant causal prediction [Peters et al., 2016] among its best instances, stands on a statistical foundation, a distribution over a cross-section, and treats the dynamics of the system observed as an addition to it, admitted where a time axis happens to be available. This paper inverts the order. Newton's laws were not an addition to the kinematics that preceded them: they replaced the description of trajectories with a law that generates them, and the description became the degenerate case, correct where the law holds and silent on what happens when someone acts on the system. The literature has no name for the same inversion in inference, and this paper does not coin one; it sets out its content. The object to identify is the recurrence that generates the corpus, a law of motion of the whole system state. A statistic of the cross-section, however invariant across environments, describes trajectories and may not define an intervention, its geometry, its cost or its recurrence reach. A description that stops at statistics is itself a learned algorithm with no recurrence, which the one criterion of §3 ranks with the rest on one scale, and which loses, on a corpus with a time axis, wherever a law shorter than any closed form is there for the criterion to find; which mode wins is a measurement the criterion makes and not a presumption it carries. Invariant causal prediction asks which conditional law survives a change of environment, which is the kinematic question; §3.4 charges the same invariance and asks it of the recurrence. The rule that follows is no causation without dynamics, and §3.4 says what it admits.

The first claim has a consequence for the width of the record that the statistical foundation hides. Under a statistical criterion every item added to the cross-section is a variable the analyst must estimate and the model grows with the record, so a wide record is a cost, and the practice is to choose one outcome and take the rest of the record as given. Taking an item as given denies that it is an output of the model at each time point, and with that denies the model to everyone whose question is about that item. Under the criterion of §3 the width of the record is on the other side of the count: every longitudinal series added is more of the corpus the recurrence must generate, and a law that generates more series from the same lines is shorter per bit coded, not longer. An item taken as given is the part of the system state the law does not generate, and a forecast departs from the observed motion at whatever rate that item in fact moves, from the first step; an item generated carries only the growth of the error in its initial condition and in the arithmetic. The wider the set of items the recurrence generates, the further any one of them projects under the same initial condition and the same arithmetic. The three-body problem is the plain case. Hold two of the bodies fixed and the third moves in a potential that does not change, Euler's problem of two fixed centres, which is integrable, solvable by quadrature, and wrong: the two bodies move, and the solution departs from their motion from the first step. Let all three move and no solution in closed form exists, but a numerical integration tracks the motion until the error in the initial condition and in the arithmetic has grown, which is as far as any integration can track it. The converse world would decide the other way. Were the motions a closed-form function of time to every resolution the instruments report, as finitely many epicycles would give them, the kinematic description would be the shorter learned algorithm, since the

receiver holds the time points as it holds the item indices and the closed form pays only its parameters and its residual, and the criterion would rank it first; no closed form exists for three bodies that move, which is why the law wins on every record we know of, and that is a fact about the records and not about the criterion. The bodies held fixed are the exogenous variables of a model written for one favoured outcome, and the integrable solution is the conditional law, tractable and wrong wherever what it holds fixed in fact moves. Which added items the law generates and which are noise that shortens nothing is what the criterion decides, and without a criterion no one can tell the two apart, which is why the consequence is not plain from the statistical foundation.

The second claim is about why the field cites and does not adopt the one design it has tried for making analytic discretion visible. The many-analysts design gives one corpus and one question to many analysts, compares their conclusions, and reports the spread it finds (§2) as its result. Reporting a spread settles nothing and costs no one anything, so the field cites the design as a caution while the practice it measured continues; it lacks a criterion, and without one the spread is all it can produce. The mechanism of §4 is that design with the discretion charged: every analyst's choices are a submitted learning algorithm, the custodian publishes the bit count its learned algorithm achieves on the sealed corpus, and the spread of conclusions becomes an ordering of accounts of the evidence, in which the shortest attained stands until a rival beats it.

2 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 advisory committee it created affirmed the need for the national secure data service that the Commission on Evidence-Based Policymaking had recommended, which the CHIPS and Science Act of 2022 (§10375) authorized as a five-year demonstration, which the National Science Foundation is now building. All of this is machinery for getting data to analysts. The Act is not silent on quality: it charges each agency's Evaluation Officer to "*continually assess the coverage, quality, methods, consistency, effectiveness, independence, and balance of the portfolio of evaluations, policy research, and ongoing evaluation activities of the agency*" (5 U.S.C. §313(d)(1)), and directs the Director of OMB to "*issue guidance for program evaluation for agencies consistent with widely accepted standards for evaluation*" (section 101(e)(1)(A)). What none of it supplies is a measure: the assessment is qualitative, and the standards are the profession's own. The Act therefore delegates the judgment of better evidence from worse to peer review and agency expertise—both exercised inside privileged access, and neither contestable from outside it.

The consequence is structural rather than venal. No one lacking the same clearance can check a finding produced from restricted data, so it rests on the standing of its author. The consequence holds inside the enclave the Act's structure demands as much as outside it: two analysts with the same clearance, reading the same restricted measurement events, have no measure between their accounts of the evidence either, so inside the enclave a finding rests on standing too, and a secure data service widens access to the data and leaves the ranking of accounts of the evidence where it was. 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 measurement events that matter most. We should stop treating this as a tension to manage and recognize it as a missing component: the Act provides

access to the data and no criterion for ranking accounts of the evidence.

The omission is visible in the Act's own definitions [Congress, 2019]. Title II defines statutory data: "*the term 'data' means recorded information, regardless of form or the media on which the data is recorded*" (44 U.S.C. §3502(16)). Title I defines statutory evaluation: "*the term 'evaluation' means an assessment using systematic data collection and analysis of one or more programs, policies, and organizations intended to assess their effectiveness and efficiency*" (5 U.S.C. §311(3)). Title III defines statutory evidence: "*the term 'evidence' means information produced as a result of statistical activities conducted for a statistical purpose*" (44 U.S.C. §3561(6)), and Title I adopts that definition by cross-reference (5 U.S.C. §311(4)). The same section defines the two terms inside it. Statistical activities "*means the collection, compilation, processing, or analysis of data for the purpose of describing or making estimates concerning the whole, or relevant groups or components within, the economy, society, or the natural environment*" and "*includes the development of methods or resources that support those activities, such as measurement methods, models, statistical classifications, or sampling frames*" (§3561(10)). Statistical purpose "*means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups*" (§3561(12)). Substituting the second definition into the first, information produced by collecting or compiling data for a statistical purpose is evidence, and so is information produced by analysing it with a model; the definition separates statutory evidence from statutory data by who produced the information and for what purpose, and by nothing else. Models appear once, in §3561(10)(B), as a resource that supports statistical activities, beside sampling frames and classifications, and no provision says how anyone is to rank one model against another. The implementing guidance does not supply the ranking either. The program evaluation standards issued under section 101(e) [OMB, 2020] note the statutory definition and take, in their words, "*a more expansive view of evidence*" (p. 9, n. 19), under which evidence "*can be quantitative or qualitative*", "*has varying degrees of credibility*", and "*the strongest evidence generally comes from a portfolio of high-quality, credible sources rather than a single source*" (p. 9); impact evaluation, which "*assesses the causal impact of a program, policy, or organization, or aspect thereof, on outcomes relative to those of a counterfactual*" (p. 10), is one such source. The guidance names degrees of credibility and gives no measure of them. A tabulation and the causal account built on it are therefore evidence on equal terms, which is the reading its commentators usually give the Act's gap: a question about which body of evidence is the better one. This paper uses the words differently, and says where: we set language quoted from the Act or from the guidance in quotation marks and italics, and roman text is this paper's own usage. Data, in this paper, are the sealed corpus at measurement instrument precision, with its public documentation, fixed once by data selection (§5). A model is a learned algorithm: whatever a receiver runs to decode the corpus from its description, with whatever cleaning the learning algorithm that produced it settled on, scored by the sequential code of §3.2. The general principle asks nothing more of it than that the receiver can run it, that its dynamics be a recurrence the receiver can step, and that its every bit enter the count (§3.2). We distinguish the two throughout: the learning algorithm is what a contestant submits and the custodian runs inside the enclave; the learned algorithm is what it produces and what has a length. Evidence is what the custodian publishes: the bit count the learned algorithm achieved on the sealed corpus, and the predictions read off the shortest learned algorithm yet produced, its accounting-based estimates (§5). On that reading evidence differs from data by a number, and a rival learning algorithm can overturn it without clearance.

Openness does not supply the criterion either, and the measurement of that is now in hand.

Silberzahn et al. [2018] gave one public data set and one question—whether soccer referees give dark-skinned players more red cards—to 29 teams of 61 analysts, let each choose its own method, and had the teams critique one another's plans before comparing the results. The teams used 29 distinct techniques and 21 distinct covariate sets; their estimates ran from an odds ratio of 0.89 to one of 2.93, with 20 of 29 finding a significant effect and 9 not. Neither the analysts' expertise nor peer ratings of their analyses' quality explained the spread, and the authors are careful to distinguish it from p-hacking and from the garden of forking paths: there was no incentive to find anything, and preregistration would not have removed it, because the analysts made the choices in good faith and each was defensible. Pre-specification is the safeguard the implementing guidance relies on: "*planned analyses produce stronger evidence than unplanned analyses; the latter are more likely to produce 'false positive' results, which appear to be reliable but in fact arise from selective adoption of analytic methods*" [OMB, 2020], p. 18. It fixes which of the defensible choices an analyst makes before the analyst sees the data, and 29 pre-specified analyses would have produced the same 29 estimates. Their remedies are to crowdsource the analysis where the budget allows and to report the whole specification curve where it cannot, and they leave open the question those remedies cannot answer: how much variability is too much. That question is what a touchstone information criterion is for. Each of their 29 analyses is a learned algorithm, and a number that ranks learned algorithms ranks them. Nothing in the design makes its adoption attractive: with no criterion the spread is its only product, so the field cites it as a caution and does not adopt it as a practice, and the mechanism of §4 is what adoption would require.

The stakes make the omission expensive. Federal transfer, tax, health and education programs turn on beliefs about how the social system responds when policy acts on it, 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. The mechanism of §4 answers that incentive with another: it pays for every bit a better model removes from the standing score, and a prize large enough to move a serious contestant to spend the compute a serious submission consumes must be commensurate with the sums at stake. A prize of that size is an appropriation, not an administrative act. Exploratory trials of the method need none; the improved models the method exists to elicit do. 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 a rival beats, and it protects one whose model no rival beats.

We use five terms throughout in one sense. A *corpus* is the whole body of observations the criterion counts over: every datum the sealed instruments recorded, at measurement instrument precision, with its documentation. An *observed object* is the thing a measurement instrument observes; inside an enclave, a respondent. A *measurement event* is one instrument's record of an observed object's attribute values at one time; inside an enclave, one respondent with every attribute held on them. An *item* is one quantity an instrument measures, and a *datum* is one item's value in one measurement event. In the statistician's triad of observation, variable and datum, the measurement event is the observation, an item is the variable, and a datum is the datum; row, column and cell describe the storage layout, and this paper does not use them for the data.

2.1 What this paper supplies

1. **The order of foundation and addition inverted.** A law of motion, the recurrence that generates the corpus, is the object to identify, and a statistic of the cross-section is its degenerate case, a learned algorithm with no recurrence ranked by the same criterion; the criterion asks invariance across environments of the recurrence and not of a conditional law, and no causation without dynamics is the rule (§1, §3.4).
2. **The many-analysts design with its discretion charged.** The design that measured the spread of defensible analyses of one corpus has no criterion and so no product but the spread; the mechanism makes each analysis a submitted learning algorithm and the spread an ordering (§1, §4).
3. **A touchstone information criterion.** The learned algorithm length of a corpus at instrument precision [Solomonoff, 1960, 1964]: a measurement in bits, with no tuning constant, no discretion in the scoring, and an unambiguous ordering. Over a corpus observed across environments it charges the absence of autonomy rather than asserting its presence (§3, §8).
4. **A mechanism whose whole disclosure is a budgeted number of bits.** Contestants submit learning algorithms; the custodian runs them inside the enclave and publishes only the bit count their learned algorithms achieve, at a stated resolution. The custodian discloses no measurement event, the published count's own length bounds what it can carry before any learning algorithm runs, and yet any outsider can contest any claim by beating its number (§4).
5. **The tiers of custody as a hierarchy of measurement instruments.** Every tier of custody is a measurement instrument: it takes the measurements the measurement instrument below it produced and abstracts them for the measurement instrument above it in accord with its purpose, a purpose that includes compliance with the statutes and regulations that bind it, and that abstraction is model selection at its own tier, charged when the mechanism runs there and uncharged data selection to the tier above (§5). The Oracle of Evidence, the learning algorithm that reads the raw measurement instrument's output with nothing hidden from it and nothing cleaned before it, is the limit from which the tiers count (§5.2).
6. **An intervention operator defined by the identified dynamics alone.** cause(C) is the smallest forcing of the recurrence that realizes a condition (§3.4); nothing is severed and there is no graph. The rule it keeps is no causation without dynamics, and §10 reports what happened when its first version did not keep it.
7. **A system closed against exogenous input.** The one component supplied to the fitted system from outside it is the one that does not extrapolate, and posing interventions as contrasts does not remove the dependence; we build and score the remedy, generating the aggregate from the cross-section (§9).
8. **A worked example**, at a scale small enough to audit, with its failures reported at the same resolution as its results, and, of the seven outcomes read for one contrast of family structure, the three that keep their sign under every test (§12).
9. **Three corrections to the neural coding of a longitudinal corpus**: a sequential code the receiver can decode (§3.2), a consistency closure on the recurrence (§7), and a

point mass in the decoder (§B.2), the last the largest effect measured on any fitted model's learned algorithm length.

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 one illustration, chosen as a fair specimen of the class the Act exists to inform. We built nothing in the learned algorithm around it: the operator poses any condition C on any items of the cross-section to the same model and reads it off in the same way (§10), and the results of §12 are one such reading among the many the worked example admits.

The paper is in two parts. Part I sets out the two claims and the methodology, the criterion, the mechanism and the boundary between data selection and model selection, without reference to the worked example; it is what the abstract says before the words "we demonstrate". Part II demonstrates the methodology on one worked example, a submission built on public releases at a scale small enough to audit, and everything in it, from the panel to the model's architecture to the operator's geometry, is one submission's choice, marked as such. The companion glossary, a separate document (`docs/glossary.tex`, rendered as `docs/glossary.html`) whose definitions also appear as hovercards on the HTML version of this paper, follows the same division: each entry defines its term in the methodology's terms first and in the worked example's second.

3 Why learned algorithm length, and why not a statistical criterion

The criterion has to be a number a custodian can compute without judgment, that no one can game by choosing a flattering test, and that measures understanding rather than fit. The learned algorithm length, the length in bits of a learned algorithm that generates the corpus, its dynamics included, is the only candidate we know of with all three properties, and in the limit it is not merely a proxy for identification: the shortest learned algorithm that generates a corpus converges, in what it predicts, on the law that generated it [Solomonoff, 1964, 1978], and there is no further fact of mechanism to be had beyond that law, selected over the widest corpus available. §3.4 is the same selection carried across environments—the pooled map against one map per environment, each paying its own parameter code. What it adds is not a second criterion but the evidentiary support over which the criterion selected the law: the environments and the region of state space the corpus spans. Within that evidentiary support the corpus has scored the law's answers to cause(C); outside it they are the selected law's accounting-based estimates, hypothetical interventions and historical counterfactuals alike, which is what a causal model is for, and the next corpus scores them.

We name the quantity for what it measures. A learned algorithm generates the corpus, so its length counts its dynamics, the recurrence a receiver steps from one period to the next, together with its parameters and the residuals coded against it. A description that stops at statistics is itself a learned algorithm, one with no recurrence, and the criterion ranks it with the rest; on a corpus with a time axis it is the longer one wherever a law is there for the criterion to find. This paper therefore says learned algorithm length wherever it means the criterion.

The criterion is Solomonoff's, and the two objections critics have repeated against it for sixty years he raised and answered himself at the outset [Solomonoff, 1960, 1964]. The first is that the length of a learned algorithm depends on the machine. His reply is that any universal machine runs any other after a fixed prefix of translation instructions, so the corpus pays the

prefix once and it cancels from every prediction conditional on it; what the choice of machine can still move the prefix bounds, and on a corpus long against it that bound is a rounding error in the ranking. He called the full form of this a conjecture, and a benchmark does not need it, because it publishes its machine: it scores every submission on the same one, and the conjecture only says that the ranking would not have changed much on another. The objection also assumes what it denies, since a statistical criterion itself runs on a machine whose primitives, counting and arithmetic, nobody pays for; Solomonoff's own example puts an arithmetic routine among the primitives and recovers frequency estimation, and with it Laplace's rule, as the shortest code once the corpus is long enough to pay for the routine, while on a shorter corpus the sequence written out is the cheaper description, which is the two-part code refusing a model that has not earned its bits. The second objection is that the quantity is uncomputable, because no procedure can tell whether an arbitrary algorithm halts. His reply was to bound the operations an algorithm may take, after which every term is computable, and that is the form a benchmark has anyway: a custodian runs submitted learning algorithms under a budget of operations, fixed by the custodian or paid for by the submitter (§4). What is uncomputable is the certificate that a given learned algorithm is the shortest possible. The search for a shorter one is computable, it is what every submission to the benchmark of §4 performs, and the benchmark scores it by what it attains rather than by what it could have attained. Reading the second objection as a verdict is an error as old as the dismissal of Solomonoff induction at its outset, and critics have repeated it across the whole span over which the compute to run such searches has grown by orders of magnitude. This paper therefore needs no bridge from an ideal to an approximation: the number published is the length a learned algorithm achieved, and the only approximation in it is the one every learned algorithm makes to the law.

The first objection has a second form, and it is the one a critic of a benchmark will make: that the submitter chooses the family of models before the count and it never enters the count, so that whatever complexity a submitter puts into the architecture rather than into the parameters goes uncharged, and the published machine is as arbitrary as the family of models it leaves free. The reply is the same reply. A submitter writes a learned algorithm for the published machine, and its length counts every line of it: the parameters, the residual code, and the lines that say what the parameters mean, which is the family of models. A family of models can hide any amount of complexity, from a table of the corpus's own values whose parameters are few because the table does the work to an architecture that is itself a fitted object, and a statistical criterion, which fixes its family of models before the data, never charges it; that is the free primitive again. On the published machine every submission writes its family of models in the same instruction set as its parameters, so the corpus pays for the family of models once in the same bits as everything else, and Solomonoff's arithmetic routine is the case in point: the routine that turns counting into Laplace's rule is lines of the learned algorithm, admitted once the corpus pays for them and not before, and a family of models is a routine of the same standing. This is also why the published machine must be exact, every intermediate value a defined bit pattern. Its instruction set is integer arithmetic at a stated width; a submission that wants floating-point arithmetic writes the standard it wants as a library in that instruction set and pays the library's length in the criterion, so that the library's lines, and not the hardware that happens to run it, fix the order of every reduction. A learned algorithm length is a fact about one learned algorithm on one machine, and a count that would differ from one copy of the machine to the next is not a measurement; floating-point hardware that reproduces the charged library's results bit for bit may execute the learned algorithm, and its count stays commensurable with every other

submission's.

A third objection is more often assumed than argued, and it is the one that matters to policy: that learned algorithm length scores prediction and not causation, so that the shortest learned algorithm attained may rest on a regularity that fails the moment someone acts on the system, and a causal claim therefore needs assumptions the corpus cannot score [Pearl, 2009]. The objection is correct about a static object. On a single cross-section with no time axis, compression cannot tell an autonomous relation from an accident of the sample, and nothing else can either, and an operator that reads its geometry off the cross-section inherits the accident. But the custodian applies the criterion to a corpus, and a corpus with a time axis that spans environments charges the absence of autonomy: a relation that holds in one period or one jurisdiction and not in another pays its parameter bits once per environment in which it differs, while a relation that holds across them pays them once (§3.4), which is invariant causal prediction's test [Peters et al., 2016] in the criterion's own currency, with no graph asserted by the analyst. Autonomy is then not a property the analyst asserts of the model but one whose absence the corpus charges, and the shortest learned algorithm attained is the most autonomous one the corpus selects. What remains is to ask of the selected learned algorithm only what its dynamics supply, which is the rule of §3.4: no causation without dynamics. Taken as a verdict, the objection is what leaves identification to discretion exercised inside privileged access (§2); over a corpus with a time axis, it is what the criterion measures.

3.1 The learned algorithm length in two parts

A learned algorithm that generates a corpus has two parts, the model and the corpus encoded with it, and the learned algorithm length is the sum of their lengths. This is the form the example in Solomonoff's first report already has, a method, its parameters, and the index of the data among the sequences the method admits [Solomonoff, 1960]. Writing θ for the parameters and δy for the resolution of datum y,

$$L \;=\; L(\theta) \;+\; \sum_{\text{observed data}} -\log_2\big(p(y \mid \theta)\,\delta y\big). \tag{1}$$

Both terms are bits, so they are commensurable and there is no free parameter trading them off. The model in $L(\theta)$ is the model as written for the published machine, the family of models with its parameters, for the reason given at the opening of this section. The sum runs over observed data 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 custodian publishes the data's own resolution. The corpus carries δy per datum: the standard error a published margin of error implies for a survey estimate, the tabulation's rounding step for a complete count, a propagated combination for a derived quantity. 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 the entropy code of the parameters at the precision at which the learned algorithm transmits them, because the criterion must score a model at the precision the receiver decodes it at; how a learning algorithm makes that cost felt while it fits is its own affair. The parameter code is not a regularizer standing in for simplicity. A norm penalty is not a proxy for it, because a weight of small magnitude can carry much information, and on a formal-language task where the analyst can write the correct network down by hand, that network is an optimum of the

two-part objective and is not an optimum of the residual code length with an L_1 or L_2 penalty, with early stopping or without [Lan et al., 2024].

3.2 What the receiver holds at each step

Equation (1) is a code only if a receiver who holds the learned algorithm and the data decoded so far, and nothing else, can form every predictive distribution in it. The residual bits of a datum are the bits that correct the model's imputation of that datum to its observed value, so each imputation must be computable before the datum arrives. We therefore write the description in the order in which the receiver decodes it: the code writes each datum against the imputation a receiver forms from the learned algorithm and the data decoded so far, whatever the learned algorithm is, and it is an observed-datum code: it codes every observed datum exactly once. A quantity the receiver recomputes for itself the learned algorithm never transmits, and it is not a term of the code, however it entered the fitting.

That is all the criterion and the mechanism require of a learned algorithm: the receiver can run it, its dynamics are a recurrence the receiver can step from the data already decoded, and the count includes every bit it uses. The recurrence need not be a smooth map of a continuous system state. A finite automaton, a pushdown automaton with a finite stack or a context-sensitive automaton with a finite tape is a recurrence in this sense, and the rule of §3.4, that the identified recurrence alone define an intervention, applies to whichever the criterion selects.

3.3 Held-out environments are an audit, not a second criterion

It is tempting to score submissions on a withheld environment—a period, a jurisdiction, a policy discontinuity—on the grounds that this is what tests whether a relation survives action on the system. In a sealed benchmark the distinction collapses: the shortest learned algorithm attained over the corpus necessarily generates the part of the corpus the contestant has not seen, so the learned algorithm 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.

The habit of selecting by held-out performance rests on a premise its users rarely make explicit: that data are plentiful and cheap, so that withholding a part of the corpus costs the model fitted on the rest nothing worth counting. Where that premise holds, cross-validation and the machine-learning benchmark are harmless, and harmless even for models capable of dynamical extrapolation to data the literature calls "out of distribution". A federal corpus is the opposite case. Its measurement events are few and costly, the instrument sees an observed object at a handful of time points, and the criterion did not select the law over any measurement event the audit withholds; the learned algorithm length over the whole corpus is the form of the same test that withholds nothing, and it asks of the whole corpus at once the question a held-out split asks only of the part the split happened to withhold.

What model selection is for decides the matter. Its purpose is decision support: one selects a model in order to predict the outcomes of the alternatives one is deciding among, and then decides under a utility, which in sequential decision theory is a function different in kind from the selection criterion; the formal statement of the pairing is Hutter [2005]'s AIXI, Solomonoff's predictor joined to a utility over actions with the two kept distinct by construction. The criterion ranks accounts of the evidence; the utility ranks actions; a held-out score is neither. The decision

maker needs the account of the evidence whose predictions of every alternative's outcome are the best the corpus supports, which are those of the shortest learned algorithm attained (§3), and not the one that did best on the one alternative a withheld split happened to resemble. The scientific process has the same shape and suffers the same conflation: practitioners routinely take the observation of an experimental result for model selection, when the experimenter could only have designed the experiment under a model already selected from prior observations, and its result is a widening of the corpus over which they select again. A held-out environment is that widening in miniature, and the next corpus is the audit.

3.4 Autonomy charged across environments, and what the dynamics supply

An environment is a grouping of the corpus across which a relation may or may not hold: a period, a jurisdiction, a policy discontinuity. The test of whether the selected learned algorithm's recurrence holds across environments is a code-length comparison, not a significance test: one set of coefficients for every environment against one set per environment, each coefficient paid for at $\frac{1}{2}\log_2 N_{\text{eff}}$ bits with N_{eff} an effective sample size allowing for the correlation among measurement events, and the recurrence is invariant when the per-environment coefficients cost more than they buy back. A relation that holds unchanged across environments is a candidate causal relation. That is the criterion of invariant causal prediction [Peters et al., 2016], which takes the predictors whose conditional law is the same in every environment as the direct causes of the target, asks the analyst for no graph, takes its environments from the data as it finds them without requiring anyone to know the shifts that produced them, and decides invariance by a hypothesis test at a chosen level. The test here differs in two respects, and the second is the one that matters: it charges invariance in bits instead of testing it, and its invariant object is the recurrence where theirs is a conditional law, which under the rule below is a statistic of the cross-section. It is the property Frisch [1938] and Haavelmo [1944] named autonomy, and the property Lucas [1976] observed a relation must have to survive action on the system. The test adds no second criterion: it is the learned algorithm length over a corpus of wider evidentiary support, and what it records is the evidentiary support over which the criterion selected the law.

Exogeneity and autonomy are one property seen from two sides. To take a quantity as determined outside the system is to assert that the relation conditioned on it stays fixed whatever generates it, which is autonomy relative to changes in that generating process, the property Engle et al. [1983] named super exogeneity; and because the generating process stands outside the model, the corpus holds no variation against which the criterion can charge that autonomy. An exogenous input is therefore a relation whose autonomy its author asserts and the record cannot judge, and that is the reason the method admits none. Closing the system reverses the assertion: a quantity the recurrence generates carries its process inside, where the test above charges it like everything else. The relationship runs the other way as well. The test spends exogenous variation: it can run only where the environments differ in what generated the recurrence's inputs, and an environment is exogenous to the relation tested in the one sense the method requires, as recorded variation the relation did not produce, and in no sense that lets it enter the relation as an input. The intervention the next paragraph defines is the same relationship at one step: the forcing overrides the recurrence's own generation of the forced items at one transition, an exogeneity of one step, and the warrant for reading the response is the recurrence's autonomy relative to that override, which the record has charged only for displacements within the recurrence reach and of the residual's own shape. The Lucas critique is this property failing for a policy quantity taken

as exogenous, and Pearl's operator is this property assumed for whatever the graph severs; the criterion charges what both assume.

A policymaker who asks what a choice will do a generation on is asking for autonomy relative to a change of rule, the case Lucas raised, and the record is the only evidence on offer, since no experiment runs a generation inside a term of office. The methodology's justification in that setting lies in the warrant it gives and the warrant it refuses. A corpus that spans generations records the policy choices of its own past, made across jurisdictions and periods, with everything that followed them; a recurrence invariant across those environments is autonomous relative to changes of the kinds past policymakers made, and a contemplated change of such a kind inherits that warrant at the invariance margin and within the recurrence reach, the answer travelling with both. A change of a kind never made has no warrant from the record, and the method says so instead of supplying one by assumption, which is what a model written in parameters declared invariant does. Closing the system brings the policy quantities themselves into the system state with their own dynamics, so that how policy has answered the system is part of what the recurrence learns, and a forcing of a policy item is an innovation in a process the recurrence carries rather than an input from outside. What the response misses is the announcement, the change in behaviour that precedes a change of rule; it is a transient of the first steps, and it is in the record wherever a past rule's announcement preceded its effect. Over a generation it is the autonomy of the law along the path that decides the answer, and that is the property the test charges.

Two kinds of intervention are worth putting to the recurrence, and a policymaker usually wants the second. The first is an impulse: a forcing applied at one transition and then left to the recurrence, whose response over the following steps is the system's own. The impulse is the diagnostic of the dynamics, what the system amplifies, damps and reverts, and cause(C) as the paragraph above defines it is an impulse. The second is a sustained intervention: a policy applied at every step, the forcing re-solved at each transition to hold the condition against whatever the recurrence has done to it, so that the intervened system is the recurrence followed by a correction for as long as the policy stands. A sustained intervention answers two questions an impulse cannot, and the recurrence answers both in its own currency. Whether the system can sustain the intervention is whether the forcing the condition needs at each step shrinks, holds or grows. A forcing that shrinks to nothing marks a system state the recurrence has come to hold on its own, a basin of the identified dynamics, and the policy that carried the observed object there is a capital investment, paid once in the sum of its forcings and maintained for nothing; a forcing that holds constant is a policy the system reverts at every step and the policymaker must reapply for as long as the outcome is wanted; a forcing that grows is a policy the system resists, and the recurrence's own residual code says it is unsustainable. Whether the intervention is desirable the policymaker then reads on the maintained trajectory and not on the impulse response, and the two readings can differ in sign, since a transient the impulse produces says nothing about what a standing policy does. Each step of a sustained intervention is a one-step exogeneity of the kind described above and inherits its warrant, within the recurrence reach and at the invariance margin, and the method assumes of no step that it persists on its own: the warrant for the maintained system state is that observed objects stood there in the record, and the warrant for the path is the sequence of its steps and nothing more. The interventions of most value are the ones the first kind finds and the second kind confirms: a sequence of small forcings along directions the recurrence amplifies, cheap in its residual code, that carries an observed object into a basin where the outcomes are better and the forcing falls to zero.

The institutions that stand between the record and the policy have an interest in that search, and the methodology has to name it rather than assume it away. A tier of custody whose own continuation depends on intermediating interventions, on being the channel through which an investment of the second kind passes, has an interest against the record showing a cheaper channel, and the record is of a system observed under that tier: a channel no observed object ever bypassed is a direction along which the recurrence has no residual, a forcing along it leaves the recurrence reach, and the method asserts nothing, which is correct and is also convenient for the intermediary. The silence is a property of the width of the record and not of the criterion. Its remedy is the one §5.2 gives, to widen the record toward the Oracle of Evidence by charging a tier nearer the raw measurement instrument, and its condition is that the custodian of the record and the beneficiary of its curation not be the same body, since a curator that intermediates the interventions the record would show cheap has reason to leave the record as it is. A methodology that can report a cheap, self-sustaining intervention serves the public only insofar as it is out of the hands of those the intervention would bypass, and that is an ethical property of the mechanism of §4 and not a courtesy of its operators.

What one may then ask of the selected learned algorithm is only what its dynamics supply. The identified recurrence alone defines an intervention: the forcing of the recurrence that realizes a condition, measured in the recurrence's own residual code and within the recurrence reach, with nothing severed and no graph. Any object unchanged for a corpus with no time axis is a statistic of the cross-section; it may describe or diagnose, but it may not define an intervention, its geometry, its cost or its recurrence reach. The rule is no causation without dynamics, and it applies to whichever recurrence the criterion selects.

The rule places a second literature beside invariant causal prediction. Single-cell biology recovers a developmental trajectory from a population in which no instrument observes a cell twice, because measurement destroys it: trajectory inference orders the cells along a coordinate recovered from the cross-section alone, called pseudotime [Trapnell et al., 2014, Haghverdi et al., 2016], and the orderings have proved useful enough that a benchmark compares dozens of methods for recovering them [Saelens et al., 2019]. Pseudotime is a dynamical axis in intent, and it stands where invariant causal prediction stands. Invariant causal prediction substitutes environments for interventions; trajectory inference substitutes the cross-section for time. Each recovers a statistic of the cross-section, the invariant conditional law in the one case and the ordering in the other, and under the rule each may describe and diagnose and neither may define an intervention. The ordering is a dynamical axis only under the model that the cross-section samples one trajectory at many stages, a model selection the ordering never charges; and the biology's own next step, estimating a velocity from the measurement event itself [La Manno et al., 2018], is the step from a statistic toward a recurrence, taken because the time axis was missing. Where a time axis exists, the recurrence is fitted on it, and the substitution of the cross-section for time becomes a hypothesis the corpus can charge.

4 The mechanism: learning algorithms in, one number out

The literature usually treats confidentiality and contestability as opposed, and they are not. Submit *learning algorithms* rather than findings; have the custodian run them inside the enclave the Act's confidentiality title already contemplates; publish only the bit count the learned algorithm achieves. The custodian discloses no measurement event; it discloses the count, and §4.2 charges it. The learning algorithm 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 learning algorithm 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.
- **The criterion scores cleaning at its tier.** The mechanism runs at one tier of the hierarchy of tiers of custody (§5.2), and whatever cleaning that tier does, it does inside the learned algorithm, where the criterion charges it; a submission that began from someone's pre-cleaned extract would be importing unscored judgement from a tier the mechanism does not run at. That includes dropping: a submission may decline to model any datum and pay its null code for it, and that, charged in bits and judged by the contestant, is the only sense in which anyone ever drops data here. Which data exist to be coded at this tier is data selection, in the sense the wider literature calls curation: the model selection of the tier of custody below, hidden from this one, which the mechanism charges only when it runs at that tier, and each tier it runs at moves the boundary one tier nearer the raw measurement instrument. Conflating data selection with model selection at one tier crosses the boundary between them, and the boundary matters for the same reason the boundary between statistics and dynamics does (§3.4): on one side of it the criterion can reach, on the other it cannot.
- **Symmetry.** The mechanism exposes an analyst whose model a rival beats and protects one whose model no rival beats. An agency holding the best-scoring submission holds a public fact, not a claim of standing.

4.1 Rules of play: the standing submission, the prize, and three obstacles

Three rules of play follow from the same properties. Contestants develop against what is already public, and only the sealed portion scores. What is public need not stop at the sealed corpus's metadata and the aggregated tables that statistical disclosure limitation permits: an agency may release a learning algorithm and the learned algorithm it produced, with its score. Agencies already build models as a statistical activity under the Act's own definition (44 U.S.C. §3561(10)(B)), and released in this form the agency's model is the standing submission, contestable by anyone. For work done under restricted access the release is a precondition rather than an option. A model the custodian's own agents built on a linked corpus, the Opportunity Atlas's among them, has no standing until the custodian runs its learning algorithm over the sealed corpus and publishes the count, because the prize pays for bits removed from a standing score and for nothing else, and no one but the custodian can produce that score. Until it does, the public holds the model's statistics as a tier above the raw measurement instrument, data selection it cannot charge, and a competing learning algorithm has nothing it can be paid for beating. Once it does, a challenger develops against the released learning algorithm and the public releases, and tests inside the enclave at the disclosure budget's price per submission, so the incumbent's years inside the data are not matched but charged. The benchmark pays a submission for its

increment over the standing score, as the Hutter Prize pays, so a published learning algorithm's cleaning step belongs to everyone from the next cycle on and the order of submission settles credit for it. And the budget of operations a learning algorithm runs under (§3) the custodian can set in either of two ways: as one allowance the custodian fixes for every submission, or as whatever compute the submitter pays for at its financial cost. The second rations the enclave's compute by willingness to pay rather than by the custodian's choice or a queue, and so opens the competition to contestants a fixed allowance would exclude. It also makes the contestants' incentives explicit: a contestant spends its own compute and earns payment only for the bits it removes from the standing score, so the prize per bit against the price of compute decides who submits, and the sums of §2 that turn on the answer say what size of prize is reasonable. A prize of that size is an appropriation, and without it the benchmark draws exploratory trials and not the submissions that would improve the standing model. A hard-coded enumeration of the corpus's quirks costs the bits of the list in the learned algorithm and asserts knowledge that only a prior submission, the published metadata, leaderboard leakage or privileged access could have supplied, whereas a learning algorithm that induces the same list from the corpus it reads inside the enclave asserts nothing; the first three leave a public trail, and the learning algorithm is the audit trail for the fourth.

Three honest obstacles. The first is **linkage**: what makes restricted access valuable is joining a person's tax, survey and program measurement events, and several statutes with incompatible rules govern that joining. The mechanism narrows this obstacle without removing it. The joining need occur only inside one execution, of the learning algorithm or of the learned algorithm, and nothing joined need leave that execution: what leaves is the count, or the answers of cause(C) queries put to the learned algorithm, each counted against the disclosure budget of §4.2. At the limit the Oracle of Evidence (§5.2) has complete linkage for its learning algorithm and admits to its learned algorithm nothing but cause(C) queries. Whether the statutes as written permit a joining that never leaves one execution is a question of law the mechanism does not answer. The second is **data selection**: which instruments enter at all is a decision no criterion over the sealed corpus can make, because the tier of custody that made it hid the data it made it on (§5.2); the mechanism charges it only by running at that tier, and what no raw measurement instrument recorded no tier can charge. The third is that **some agency processing is undocumented**. Here a trend the literature usually reports 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.

4.2 The attack the mechanism must answer

The conclusion first. A published score carries at most a few bits about the sealed corpus, and obtaining it costs a full run of a learning algorithm over that corpus, so an adversary mining the leaderboard for measurement events pays orders of magnitude more compute per bit than any exploit could return; the bound below says how many bits leave, and the three answers that follow it cap the number. The larger risk is denial of service: an adversary with deep pockets submitting learning algorithms in bad faith to tie up the enclave's compute, a strategic resource. Rate limits on submissions answer it as they answer every denial of service, and the submission budget below is one. What follows gives the leakage its cost in bits.

A contestant can overfit a sealed corpus 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.

The leak is not a side effect that care removes; it is the output itself. A published score is a function of the sealed data, and a submitted learning algorithm can make it any function it likes: pad the code by one bit so that the parity of the count answers a yes-or-no question about one observed object, and a custodian who publishes the exact count has answered it. One score discloses one bit. Confidentiality "preserved exactly" is therefore not available to any mechanism that publishes a number, and this paper does not claim it. What is available is a bound in the mechanism's own currency. A count published at resolution δ over a range of R bits is one of R/δ values, so it carries at most $\log_2(R/\delta)$ bits about the corpus whatever the learning algorithm did, and n submissions carry at most $n \log_2(R/\delta)$. That is the whole of the disclosure, and we call it the disclosure budget: the custodian counts it before running any submission, and it asks nothing of the learning algorithm, which is what distinguishes it from a differential-privacy guarantee. The latter needs a bound on how much one measurement event can move the score, and the parity learning algorithm moves it by the whole range. The budget says how many bits leave, not which observed object they are about: a learning algorithm can spend a cycle's entire allowance on one observed object, and the custodian has to set the allowance with that in mind. Three answers compose.

- **Budget the submissions.** The number of scores the custodian releases bounds the leakage, so a budget of n submissions per cycle fixes the cycle's disclosure budget at $n \log_2(R/\delta)$ bits before the cycle opens, a quantity to compare with what the corpus already discloses by publication rather than with zero.
- **Publish at the paid-for resolution and no finer.** The custodian publishes whether a submission beats the standing score by at least δ, and the new standing score rounded to δ. The benchmark pays a submission for its increment, so it does not pay for an improvement below δ either, and the rounding costs the leaderboard nothing it was using: it cannot order two submissions closer than δ, and it should not. The reusable-holdout construction [Dwork et al., 2015] is the same move with noise in place of rounding; its privacy-composition accounting needs the sensitivity bound an arbitrary learning algorithm cannot supply, which is why the bound given here is on the bits the published score can carry.
- **Let the corpus renew.** This is the advantage a federal statistical corpus has over a static machine-learning benchmark: 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. Renewal does not retract what earlier scores disclosed about observed objects that recur; it bounds what the leakage can earn.

There is precedent for accepting a bound of this kind in place of a worst-case guarantee, and for what carries the residual. The Opportunity Atlas [Chetty et al., 2018] published an estimate of social mobility for every census tract from confidential tax and census measurement events, with noise added in proportion to each statistic's maximum observed sensitivity, the largest change one measurement event made to it across all the tracts; its authors write that the method "does not offer a formal privacy guarantee" [Chetty and Friedman, 2019], because the

worst-case sensitivity of a regression estimate is unbounded and the local sensitivity is itself a statistic of the sample. The estimates are public. What carried the residual risk was the law: the researchers worked as temporary staff of the Census Bureau under what the Bureau calls special sworn status, which the statute permits "*only if such temporary staff is sworn to observe the limitations imposed by section 9*" (13 U.S.C. §23(c)), and wrongful disclosure by anyone so sworn is a crime, "*fined not more than $5,000 or imprisoned not more than 5 years, or both*" (13 U.S.C. §214), a fine 18 U.S.C. §3571(b) raises to $250,000 for a felony. The mechanism asks for less on both sides. Its bound holds whatever the learning algorithm does, which the Opportunity Atlas's noise could not claim, and the channel it opens is one count per submission against an estimate for every tract and subgroup; and no submitter sees a measurement event, so the only people the sworn-staff provision need cover are the custodian's own, whom it covers already. A custodian that wants the Opportunity Atlas's human backstop as well may swear submitters as temporary staff under the same section, bringing them under §214 at no new cost.

Whether any of this requires new legislation is a question of law we do not settle. The Act provides the enclave, the confidentiality regime and the mandate to build evidence; the published number, the submission budget and the open queue are administrative in form. The prize is not: a payment per bit removed from the standing score, of the size the sums of §2 warrant, is an appropriation, and nothing less than the full value of getting those dynamics right justifies the compute a serious submission consumes. Exploratory trials of the mechanism may run on what the Act already provides; the improved models it exists to elicit will not. Its evaluation provisions—the Evaluation Officer's duty to "*continually assess the coverage, quality, methods, consistency, effectiveness, independence, and balance*" of the agency's evaluation portfolio (5 U.S.C. §313(d)(1)), and the Director's guidance "*consistent with widely accepted standards for evaluation*" (section 101(e)(1)(A) of the Act)—are a mandate to judge quality without a measure of it, which is the gap of §2 and not a bar to filling it.

5 The boundary between data selection and model selection

We use *data selection* for what the wider literature calls curation: deciding which instruments, sources and data exist to be read, and whatever anyone does to them before the custodian seals the corpus. The term that locates the boundary is the *tier of custody*. A measurement instrument is the act that produces a measurement event, and the party that performs it. The raw measurement instrument records the observed object itself, a thermometer's reading or an interviewer's record of a respondent's answer, and its output is the raw measurement. Every tier of custody is a measurement instrument too: it takes the measurements the measurement instrument below it produced and abstracts them for the measurement instrument above it in accord with its purpose, a purpose that includes compliance with every statute and regulation that binds it, so the hierarchy of tiers of custody is a hierarchy of measurement instruments and begins at the raw measurement instrument. Every party that holds data on the way from the raw measurement instrument to the finding is one: the agency that collects the raw measurements, the agency that publishes tabulations of them, the analyst who builds a panel from the tabulations and seals it. A tier of custody holds its own tier's data in full, abstracts them for the tier above, and hides from the tier above what law requires it to hide and whatever it does not pass up (§5.2). What a tier receives from its measurement instruments is its evidence, in the Act's sense at that tier, presented as those instruments' own estimates at their stated precision. What it produces is an account of the evidence: a learned algorithm from which every bit of the evidence

decodes at that precision, so that nothing is lost in the sense of compression and every bit is accounted for, and whose length the criterion measures. What the account produces beyond the evidence—extrapolations, interpolations, imputations, forecasts and the answers to cause(C) queries—are accounting-based estimates: the account's and not any instrument's, warranted within the account's evidentiary support and extrapolation outside it. And what the tier presents to the tier above is its own evidence: its count, its account of the evidence or an interface to it, and its accounting-based estimates, which the tier above receives as its measurement instrument's estimates. What a tier passes up need not be data in the form it holds them. It may pass up a model of its data, the learned algorithm its tier selected, or a query interface to that model that presents the tier above its permitted view of the data and answers questions put to the model, intervention contrasts among them; what it passes up is then its model selection made legible, charged at its own tier, and the disclosure budget of §4 counts each answer as it counts each published score. The boundary between data selection and model selection is the boundary between two tiers of custody. Whatever the tier below did to the data is data selection at this tier: unscored, a property of the data that no criterion over the data can reach. Whatever this tier does to the data inside its learning algorithm is model selection at this tier: a line in the learned algorithm, charged in bits like every other line. Cleaning can sit on either side, and which side depends on nothing but the tier of custody that performs it. The boundary between data selection and model selection is as consequential for this method as the one between statistics and dynamics (§3.4), and for the same reason: it separates what the criterion can adjudicate at a tier of custody from what it cannot, and §5.2 names the limit from which the tiers of custody count.

What follows is where the criterion charges and where it cannot. A cleaning step applied at a tier of custody below the corpus is invisible to any criterion computed over the corpus, because the criterion scores accounts of the evidence it receives: a defect there changes published results and leaves goodness-of-fit untouched, and no optimization over the corpus will ever surface it. The mechanism charges such a step only by running at the tier that performs it, and each tier abstracts for its own purpose, the law that binds it included (§5.2).

5.1 Why the boundary is the whole question

Consider a per-capita normalization keyed on the source's unit indicator. Dollars serve both an aggregate payroll and a median income, so the normalization divides a median by population. Ask what an information criterion makes of this. If population is in the corpus, the item the normalization corrupted, the median divided by population, is an invertible function of the median as the instrument reported it, given population, and the two corpora are the same object: the criterion is *exactly* indifferent, and correctly so.

That indifference is the argument for charging the step at the tier that performs it, not an argument against the criterion. As a step at the tier of custody the criterion charges, inside the learned algorithm, the division is free, and the choice is legible in the code. As a step at the tier of custody below, applied before the corpus, it is unfalsifiable from the tier that receives it, and the item silently denotes something other than what its name says. Where someone builds a corpus the second way, with the denominator consumed by the normalization and never appearing as an item, the defect is not information-preserving but merely negligible, a fraction of a bit per datum on one item among hundreds, and no optimization over that corpus would ever surface it.

Two lessons for a federal benchmark. A defect of this kind is invisible to the objective because a model fits whatever it receives; it surfaces only when someone reads a quantity in its own units of measurement and asks whether the number is *possible*, or forces a derivation 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' description of them. The many-analysts project of Silberzahn et al. [2018] met one such defect: the data collection had recorded the club and league of each player once, over careers spanning many seasons, and ten teams had conditioned on them. Debate and a coordinators' ruling settled it, a ruling that the teams could rerun or retain, and one coauthor's dissent is a footnote. Inside the scored object the same question is a cleaning rule that repays its cost or does not.

5.2 The Oracle of Evidence, and the tiers of custody beneath it

The limit needs a name. Call the Oracle of Evidence the learning algorithm that reads every raw measurement every raw measurement instrument produced, at that instrument's own precision, with nothing hidden from it and nothing cleaned before it reads. The word is computability theory's, where an oracle supplies what no machine can compute; here it supplies what no custodian can hold. The Evidence Act defines evidence as "*information produced as a result of statistical activities conducted for a statistical purpose*" and statistical activities as "*the collection, compilation, processing, or analysis of data*" (44 U.S.C. §3561(6), (10)), so under the Act's own definition every step from the raw measurement instrument to the finding is one class of activity, and the Oracle of Evidence is the learning algorithm for which that whole class is its own: it has no curator, every cleaning step is a charged line of its learned algorithm, and the only residual is the one its learned algorithm pays for in bits. No one can build it. The confidentiality title requires every custodian to hide measurement events in identifiable form, raw measurement instruments do not keep what they measure, and no custodian holds every raw measurement instrument's output. Its linkage is complete: its learning algorithm joins every measurement event of every observed object, and what it admits to its learned algorithm is cause(C) queries. It is the reference against which this section measures every actual arrangement of custody: by the measurement instruments that stand between the raw measurement and the learning algorithm.

What separates any analyst from the Oracle of Evidence is a hierarchy of tiers of custody, each of them a measurement instrument, and the structure the section opened with is the same at every one of them. The custodian of a tier of custody holds that tier's data in full. It abstracts them for the tier above, in accord with its purpose and the law that binds it: the agency edits and imputes the measurement events it collected, removes or obscures identifying information to produce "*a less sensitive and more accessible version of a data asset*" (44 U.S.C. §3582(b)(1)(C)), aggregates them to tabulations and publishes those, or publishes in their place the learned algorithm its tier selected, or a query interface to it; the analyst who reads the tabulations builds a panel, normalizes, drops, fills and seals it. It hides from the tier above what law requires it to hide and whatever it did not pass up. And the cleaning it does is model selection given full access to its own tier: an edit rule is a model of which measurement events are errors, an imputation is a model of the missing datum, a disclosure-avoidance mechanism is a model of what an attacker could infer, and a per-capita normalization is a model of which items are extensive. Each decides what is residual at its tier, and each decides it with a model. Seen from the tier above, the same step is data selection and unscored, because the tier above cannot see the data on which the tier below chose it. Seen from its own tier it is model selection, and that tier could have charged it.

The curator's claim, that what it dropped or corrected was residual, is a model's claim made without counting the residual, and that is the whole difference between a curator and a learning algorithm. The structure is the same at every tier of custody, which is what makes the process recursively self-similar: the agency is the Oracle of Evidence relative to the analyst and a curator relative to the respondent, and the analyst's sealed panel is the Oracle of Evidence relative to the fitted model and a curator relative to the agency.

That is what the boundary of this section is. Data selection is not a different kind of act from model selection; it is the model selection of the tier of custody below, hidden from the tier that receives its output. What no learning algorithm at any tier can reach is what no raw measurement instrument recorded and what law forbids any one custodian to hold, which is the distance from the Oracle of Evidence; everything else is a residual judgement that some tier could have charged and did not. The mechanism of §4 charges one tier of custody. The custodian holds the tier's data in full, the submitted learning algorithm has the Oracle of Evidence's access to that tier, the analysts above it have only the count, or the standing submission's learned algorithm and the answers of an interface to it, and the disclosure budget is what the hiding costs in bits. Run at the agency's own tier it would charge the agency's editing, imputation and disclosure avoidance; run at the analyst's tier it charges the construction of the panel; and each tier it runs at moves the boundary one tier nearer the raw measurement instrument. The criterion is honest where the curator is not, at every tier and for the same reason: it counts the residual against the model that called it residual.

Part II
The worked example

Part I stated the methodology without reference to this part. What follows is one worked example of it: a submission built on public releases, at a scale small enough to audit, with its failures reported at the same resolution as its results. Everything in it is the worked example's choice, made so that we could run and audit the worked example, and the general principle carries no prior in favour of any of it. We developed the worked example on public releases and release it whole (§14), which is the form §4 allows an agency's standing submission to take.

This matters for how a reader should take the worked example. Everything it reports 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, the we name the violation that caused it (§9); none of the failures is evidence against the underlying thesis, and we labour the point because objectors have used the claim that there is "no empirical evidence" for that thesis to decline precisely the instrument that would produce the evidence.

6 The worked example's panel: construction and its defects

The observed object 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. A *panel version* is one of the three successive assemblies of the data, of 525, 770 and 777

items; the counts here are the third's. Of the $25{,}168 \times 777$ data, the sources supply 9,280,572 (47.5%). The chart stores values as $\log(1+x)/s_{95}$ with precision propagated accordingly. These normalized coordinates, with the per-capita and per-item normalizations below, are the *chart*, in the sense of a coordinate chart: the coordinates in which we fit the manifold of §7 and in which the code writes every datum. The six transitions the recurrence fits step from 1980 to 2024; we call each a decade, although the last three span five, five and four years, and the step length Δt carries the actual interval wherever it matters, a step-size-aware transition. The per-datum δy of (1) is a published margin of error divided by 1.645 for an ACS estimate, the tabulation's rounding step for a complete count, and a propagated combination for a derived quantity.

Three defects in the worked example's own cleaning, every one applied as data selection, upstream of the corpus, 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. We had applied every one *upstream* of the corpus, where a criterion cannot reach it.

The first defect, measured. We keyed per-capita normalization on the source's unit indicator, and dollars serve both aggregate payroll and median income, so the normalization divided four dollar medians by population, the median per-capita defect. With population in the corpus each divided median is an invertible function of the median as published, and we measure the difference at -0.0000 bits per datum: the criterion is exactly indifferent, as §5 says it must be. We built our panel the other way: the normalization consumes the population denominator, which never appears as an item, and in that configuration the defect is not information-preserving but merely negligible. It costs 1.86 bits per datum on that item in 1980 and 1.49 in 1990, which is 0.04% of a 6.6 bit-per-datum total spread over 777 items. No optimization over that corpus would ever have surfaced it.

Reference-year items. 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 items 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 §12: an economic variable the corpus did not contain had confounded the earlier estimates.

Non-additive quantities aggregated. Connecticut replaced its counties with planning regions in 2022, a planning region gap. 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. The rule for admitting a reconstructed item is verify-then-emit: the same reconstruction must reproduce a *published* value in a period where one exists. 635 of 676 items cleared, and Connecticut's observed density at 2024 rose from 0.004 to 0.290 against 0.319 nationally. We then found that procedure itself defective, through a list misalignment that paired each item with another item's error—which is what had let the $3.77 million through.

We wrote the cleaning rules of this section by hand, and the corpus can settle most of them itself: whether an item is extensive, whether a near-constant item is a measurement, whether an item-year is on one scale. Inducing those rules reads the corpus, which under the sealed

mechanism the analysts cannot do but the learning algorithms they submit can: a rule the learning algorithm induces inside the enclave is a line of the learned algorithm, charged like every other, and the access the analysts lack is exactly the access the criterion charges for. Appendix B reports the induction as we ran it, by hand on the corpus we hold, which is a stage of data selection; the same computation placed inside a learning algorithm is model selection. We keep the appendix out of the main text so that a procedure that sat, as we ran it, on the far side of the boundary this section draws does not blur it.

7 The worked example's model: a manifold autoencoder with a closed recurrence

§3 and §3.2 say what the criterion and the mechanism require of a learned algorithm, and it is all they require: the receiver can run it, its dynamics are a recurrence the receiver can step from the data already decoded, of any computable kind, and the count includes every bit it uses. Everything else in this section is a choice of the worked example, made so that we could run and audit it, and the general principle carries no prior in favour of any of it: the manifold autoencoder, the mixture head, the latent width, the function library and the closure are one family of models (§13, Appendix A), and a submission to the benchmark inherits none of them.

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's output must land on the encoder's own embedding of the next period, not merely decode into that period's observations. Without it, one-step prediction looks fine and multi-step propagation is meaningless—which matters here because the operator propagates intervention results for decades.

An encoder pushes a county's measurement event, with unobserved inputs entering at the per-item mean, forward onto the manifold, to a latent state $z \in \mathbb{R}^{32}$: the *pushforward*. A decoder pulls the latent state back to the observables, the *pullback*, returning, for each item, a three-component Gaussian mixture. The latent state is, in static terms, the county's position on the manifold at a time point; in dynamical terms it is that point with its motion, the rate (2) assigns it there, as Newton's laws took position and its derivatives together as the dynamical state of a body. The mixture matters: it lets the residual term be an honest code length for asymmetric and multi-modal item distributions without transformation tricks. Null initialization sets the weights so that the untrained output is exactly the per-item null, so epoch zero scores the null and the model earns every subsequent bit from a baseline it itself reproduces. During training a denoising mask hides a fraction of observed inputs while still coding them, which prevents the encoder from copying inputs to outputs.

We compute $L(\theta)$ of (1) two ways: for the code-length 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. The forward pass rounds the weights, the practice called quantization-aware training, so that the model we fit is the model we score. The limitation Lan et al. [2024] state, that the parameter code is not differentiable, is what the differentiable code at the quantization step is for.

Coding a county's own measurement event from its own pushforward, at each datum's precision, we call the *pullback*, after the operation whose residual it measures. Its residual measures how much structure the manifold has captured, in the data's own units of measurement.

It is a measurement and not a part of the code: the encoder computed the latent it decodes from out of the measurement event under coding, which is the circularity §3.2 excludes. The pullback trains the autoencoder and we report it as a diagnostic; the learned algorithm length is the sequential one of (4).

7.1 The recurrence, and why we close it

A recurrence carries z from one period to the next,

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

with $c_{t,s}$ a displacement shared by all counties 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 county. The decoder is far from injective, and with half the target data 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 land on the encoder's own embedding of the next period, penalized by

$$\text{cons} \;=\; \tfrac{n}{2}\sum_k \log_2\!\big(2\pi e\, s_k^2\big), \qquad s_k^2 = \big\langle (z_{t+1}^{\text{pred}} - z_{t+1}^{\text{enc}})_k^2 \big\rangle, \tag{3}$$

a Gaussian code on the discrepancy at its own fitted scale per coordinate. It is a constraint on the fit and not a term of the code: the learned algorithm never transmits the embedding, since the receiver recomputes it from data already decoded, and as a differential entropy the expression turns negative once s_k^2 falls below $(2\pi e)^{-1}$, which at the fitted scale it does (§3.2). It enters the training objective and nothing else.

The closure buys the multi-step results: the consistency-closed model predicts *better* than the unconstrained one (8.43 against 8.76 bits per datum in the panel version where we fitted both) 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.

7.2 The sequential code of the worked example

A manifold autoencoder invites a violation of the rule of §3.2 that is easy to miss: coding a measurement event from the latent state the encoder computed from that same measurement event. The residual so measured is a genuine measurement of how much of the measurement event the manifold reproduces—§7 calls it the pullback and uses it for exactly that—but no receiver can form that imputation without the measurement event, and in the limit an identity encoder with a decoder concentrated on its input would code every measurement event in zero bits. We therefore write the description in the order in which the receiver decodes it. Writing S for the parameters of the chart of §6 (the normalization mask, s_{95}, the per-item mean and variance, the per-capita denominator), $y_{c,t}$ for county c's observed data in period t, E for the encoder and Φ for the recurrence (2),

$$L \;=\; L(\theta) + L(S) \;+\; \sum_{\text{start periods}} -\log_2\!\big(p_0(y_{c,t})\,\delta y\big) \;+\; \sum_{\text{transitions}} -\log_2\!\Big(p\big(y_{c,t+1} \,\big|\, \Phi(E(y_{c,t}))\big)\,\delta y\Big). \tag{4}$$

A county's start periods are the periods no transition targets—1970, which the dynamics exclude because only 46 items have values there, and 1980, the first period the recurrence can step from—and p_0 is whichever of two imputations the receiver can form before it has seen the county is cheaper: the decoder run from the mean latent (thirty-two numbers, transmitted) or the per-item null, with one bit to say which. The code writes every later period against the recurrence's step of the encoder's embedding of the period already decoded, which the receiver recomputes for itself. It codes every observed datum exactly once, and nothing against information the receiver lacks. The consistency term (3), which holds the recurrence's output to the encoder's embedding, is a Gaussian differential entropy on a quantity the learned algorithm never transmits; it belongs to the fitting and not to the code, and so does the pullback. We write equation (4) for the worked example's decomposition into chart, encoder and recurrence; the general form is the rule of §3.2, whatever the learned algorithm is.

For the worked example's model, scored at its transmitted checkpoint (§11):

term	data	bits	bits per datum
$L(\theta)$, entropy code over quantized weights		4,421,084	
atom locations (§B.2)		282	
$L(S)$: per-item mean and variance		24,864	
start periods, 1970 and 1980, against the null	1,029,855	11,311,901	10.98
transitions, against the stepped embedding	8,250,717	54,686,184	6.63
learned algorithm length	9,280,572	70,444,312	7.59
pullback (diagnostic, not in the code)	9,280,572	60,541,596	6.52
consistency term (diagnostic, not in the code)		−2,748,044	

The table shows two things the code-length accounting we used while fitting these models hid. That code-length accounting, pullback plus prediction plus consistency plus parameters, totalled 1.169×10^8 bits for this checkpoint: it coded every period after 1970 twice, once against an imputation the receiver cannot form, and it carried a differential entropy that is negative at the fitted scale and took 2.7×10^6 bits off the total. Every learned algorithm length quoted in this paper is now the sequential one; where we made a section's comparisons under the previous code-length accounting it says so and reports what the sequential code makes of them. The per-capita denominator of §B.3 remains outside $L(S)$: it is the one object in S this submission does not yet transmit, and that section gives its cost.

8 Identification and invariance

Decade motion separates into a component shared across counties and a county-specific remainder, and the two behave oppositely. The developmental map, the part of the recurrence that acts on the county-specific remainder, is invariant, in quadratic form across the six decade environments and in linear form across the 41 US-state environments tested. The shared component does not extrapolate at all. That asymmetry is what licenses posing interventions as contrasts, which remove the shared displacement's explicit term from the difference equation—though not, as §10.1 shows, entirely from the result—and, as §9 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 county-specific, the national share inflated by nominal-dollar items moving with the price level.

The shared displacement'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 datum against 14.0 for assuming no shared displacement at all. Worse than nothing.

The developmental map is a different object. Under the test of §3.4, with N_{eff} allowing for spatial and temporal correlation among counties, fitted in quadratic form it is invariant across all six decade environments, the split costing 4.0×10^5 bits more than it saves; in linear form it is invariant across all 41 US-state environments, by 3.7×10^4 bits, and the quadratic form is untestable there because most US states hold fewer transitions than it has terms. Linear-in-decades fits are *not* invariant, by 9×10^4 bits, which we read as curvature mistaken for time-variation when we force the form to be linear.

The name of the map is the biology's, and we ran the biology's method as a check. We took *developmental map* from the trajectory-inference literature of single-cell biology (§3.4) and computed a diffusion pseudotime [Haghverdi et al., 2016] on the fitted latents with the time axis withheld, taking of the leading diffusion coordinates the one most aligned with the county-specific displacement, in the script `develop.py` of the released repository. The check was informative in the negative. Even that coordinate correlates with calendar decade at 0.93 in magnitude, the sign of a diffusion coordinate being arbitrary, and has a median cosine of 0.005 in magnitude with the county-specific displacement: it reads the shared displacement and nothing of the developmental map. Shared national conditions synchronize counties as no shared condition synchronizes the cells of a dissociated sample, which is the difference that stops the biological method from transferring, and it is the decomposition this section tests. What pseudotime recovers is the component that fails invariance; what passes invariance is the component pseudotime cannot see.

Autonomy in the sense of §3.4 we have for the developmental map and demonstrably lack for the shared displacement.

9 The exogeneity we smuggled back in, and the system closed against it

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.

This is the relationship of §3.4 met in the data. The shared displacement we imported was an autonomy claim, that the county recurrence conditioned on the national series stays fixed whatever generates those series, and nothing in the panel could charge it; the component that failed to extrapolate is that claim failing.

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: one conditioned on an input whose generating process changed between the fitted decades and the withheld one.

The remedy is to close the system, and it is concrete. Of the twelve national series we used, eight —unemployment, three labour-force participation measures, median home value, total fertility, adolescent fertility—are aggregates of the county cross-section, which the model should *compute from the system state* rather than receive as inputs. 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 model generates the aggregate rather than assuming it.

We report this prominently because of how readers are liable to take a study like this. 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 corpus we forbade it to explain, and it declined to explain it.

9.1 We built the closed system: split from the open model it neither costs nor buys anything we can resolve; fitted apart, it codes shorter

Closing the system does not, on this panel, change the learned algorithm length by an amount four paired fits could distinguish from zero under the code-length accounting in force when we made them; forecasts no better in a comparison that hands the open model the one input a forecast cannot have; and recovers the aggregate we denied it in two fits of six. We built the remedy of §9 and it runs. This section first reports the comparisons as we made them, and says why four seeds could not resolve them, then rescores every checkpoint under the sequential code of §3.2, which removes from both arms the pullback and the consistency term. Under that code the four from-scratch pairs all favour closure, by 2 to 3×10^5 bits; the pairs split from one fitted model still resolve it to about 4×10^4 bits of nothing. The pullback had been hiding the first result, because the closed fits spend more on it and it is not part of the code.

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

$$u_{t+1} \;=\; u_t \;+\; \Delta t \,\big[\,1,\; u_t,\; \bar{z}_t\,\big]\,A, \tag{5}$$

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: no effect at the resolution available. We recompute every learned algorithm length in this section from a saved checkpoint with a clean encoder input, weights quantized on both the pullback and the prediction branch, and the model in evaluation mode. Paired by seed, closed minus open, at the two checkpoints each run keeps:

	seed 0	seed 1	seed 2	seed 3	mean	p
checkpoint the trainer marks best	−846,590	−52,881	+488,929	+332,921	−19,405	0.95
last checkpoint	−398,415	−112,115	+510,042	−54,022	−13,628	0.95

Under the previous code-length accounting (pullback, prediction, consistency and parameters), the mean is a hundredth of a per cent of the total and its sign depends on nothing. The spread is the finding. Across four seeds of one architecture a total has a standard deviation of 1.9 to 4.9×10^5 bits, and a single run differs from *itself* by up to 5.4×10^5 between its two checkpoints. Four pairs therefore cannot see an effect smaller than about 5×10^5 bits, and there is none that large. The differences that do appear are trades rather than gains: in three pairs of four the

closed fit settled on coarser quantization, spending 4.4 to 8.6×10^5 fewer bits on parameters and 4 to 14×10^5 more on residuals.

Supplying the point mass of §B.2 to both arms moves every total down by more than a million bits and leaves the comparison where it was: a mean of -1.2×10^5 at either checkpoint with the rest of the model frozen, $p = 0.63$ and 0.35, and $+0.5 \times 10^5$ when we refit both arms with it, $p = 0.67$. The two arms gain from the atom alike, 1.48 against 1.39×10^6 bits, a difference well inside the scatter of the pairs.

Rescored under the sequential code, the same four pairs all favour closure. The sequential code of §3.2 drops the pullback and the consistency term from both arms. At the checkpoints the trainer marked best, closed minus open is then -4.1, -2.7, -0.9 and -3.0×10^5 bits, a mean of -2.7×10^5 with $p = 0.03$; at the last checkpoints -2.2×10^5, $p = 0.10$; with the frozen-fit point mass in both arms -3.0×10^5, $p = 0.004$, four pairs of four in every case. The decomposition says where the previous code-length accounting lost this: the closed fits code the transitions 2.6×10^5 bits worse but spend 5.3×10^5 fewer on parameters, and they reproduce their own measurement events 2.6×10^5 bits worse in the pullback. The pullback entered the old total and cancelled the parameter saving; it does not enter the code. The seed-to-seed standard deviation of a learned algorithm length is also smaller under the sequential code, 2.0 to 2.5×10^5 bits in either arm, because the pullback carried much of the scatter.

Paired by model rather than by seed, the bound tightens to 10^5 and the answer does not move. Two seeds differ by the model, so the pairing that cuts the scatter is one in which both arms are the same model until the moment they differ. We continue each open fit twice from its own checkpoint with the same optimizer moments, learning rate and mask stream: once as it stands, and once as a closed model whose generated national system state starts on the very path the open model imports—u_0 its first point, A the minimum-norm solution that reproduces the rest of the path from the checkpoint's own cross-county means. The two arms begin $1{,}728 \pm 71$ bits apart, which is the cost of transmitting the seventy-two numbers we hand the open model, and otherwise as one model. Scored on the total averaged over the final 500 epochs, closed minus open:

continued from	without the point mass	with it
seed 0 (`fq_all`)	+32,550	+46,437
seed 1 (`fq_seed1`)	+36,065	−8,350
seed 2 (`fq_open_s2`)	+157,817	−117,658
seed 3 (`fq_open_s3`)	+185,043	+17,061
mean (p)	+102,869 (0.08)	−15,627 (0.69)

The scatter of a pair is now 7 to 8×10^4 bits, three to eight times tighter than pairing by seed allowed. Over all eight pairs closure changes the code by $+4.4 \times 10^4$ with a standard error of 3.4×10^4, $p = 0.23$: within 0.1% of the total, generating the aggregate from the county cross-section costs nothing and buys nothing. The question this design asks is narrower than two architectures fitted from scratch—whether closing a fitted open model changes its code—and under the previous code-length accounting the two agreed.

Rescored at each run's minimum checkpoint under the sequential code, the paired branch gives -6.8×10^4 bits with a standard error of 6.9×10^4 without the point mass ($p = 0.40$) and

$+3.7 \times 10^4$ with a standard error of 1.0×10^4 with it ($p = 0.035$, closure costing a twentieth of a per cent in all four pairs); over all eight pairs -1.6×10^4 with a standard error of 3.8×10^4, $p = 0.69$. Split from one fitted model, closure is still worth nothing that four pairs resolve, now to 4×10^4 bits. The from-scratch pairs above say that two architectures fitted apart settle in different places, the closed one with a cheaper parameterization; the branch says that a fitted open model does not get there by being closed. Both are measurements of the same remedy, and the second is the one a custodian scoring a submitted learning algorithm would make.

The same instrument charges the per-US-state share of the shared displacement, G_{state}, removed from each open fit and refitted against the fit as it stands. Taken away cold it costs 4.1×10^5 bits with a scatter of 2×10^4; after the rest of the model adapts, three quarters of that returns. The residual ends about 10^5 bits longer without it and the parameter code 4×10^4 shorter, a net of $+6.3 \times 10^4$ with a standard error of 3.4×10^4, the same sign in three pairs of four ($p = 0.17$; $+1.1 \times 10^5$, $p = 0.11$, on the minimum). Three thousand numbers that let each US state carry its own share of the shared displacement are worth something of the order of 10^5 bits, at the edge of what the design resolves. Under the sequential code the result crosses the edge: at the minimum checkpoints removing G_{state} lengthens the code by $+1.0 \times 10^5$ bits with a standard error of 1.9×10^4, $p = 0.013$, in all four pairs, almost all of it on the transitions.

The trainer's own log says otherwise, and is not a learned algorithm length. The number the trainer prints as its total is the forward pass it differentiates: it masks the encoder input at a rate it redraws every epoch, and decodes the prediction branch at unquantized weights. It moves by a median of 1.3×10^6 bits between diagnostics ten epochs apart, and its minimum over a run sits 0.18 to 1.46×10^6 bits below the previous code-length accounting's total by a different amount in each run. Ranked by that minimum the same eight fits give closure -1.8, $+0.5$, -5.2 and -3.4×10^5 bits—three wins of four, and a result.

This is §4 in miniature. A contestant optimizes whatever they like; the custodian runs the submitted learning algorithm and computes the learned algorithm length itself, from the sealed corpus and the learned algorithm it outputs, and never sees the objective the contestant descended. Under that arrangement a training objective cannot reach the published number, because that number is not the contestant's to compute. In a study one writes alone there is no custodian: the hand that trains the model reports its score, the log is the nearest number to hand, and it is a number that flatters. The discrepancy is findable by anyone holding a checkpoint and the corpus, which is what "contestable without privileged access" means operationally, and it is about 1% of the total—enough to manufacture a finding, which is why someone other than the optimizer has to compute any criterion worth sealing a corpus for.

Out of sample: no difference either. Holding out the 2020–2024 transition, the open model codes it at 4.138 bits per datum at its best-marked checkpoint and 4.127 at its last; six closed runs average 4.159 and 4.154, 0.5 to 0.7 per cent worse, with two of the six ahead of the open model at one checkpoint and none at the other. A closed run moves by up to 0.08 bits per datum between its own two checkpoints, three times the gap under comparison. The comparison is also unfair to the closed model in a way that matters more than its outcome: at prediction time the we hand the open model the 2020 national system state, which is the one input a genuine forecast cannot have. Whatever closure costs here, it buys a system that needs nothing from outside itself.

And the model recovers the aggregate only sometimes. In a held-out run the model produces the generated u_{2020} 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 as multiples 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.

Six fits do not decide whether the criterion ranks its own fits. Among fits of one architecture, where nothing but the optimizer's path distinguishes them, a shorter code should go with better held-out prediction if the learned algorithm length is tracking what transfers. Across the six held-out closed runs the rank correlation between learned algorithm length and held-out bits per datum is -0.03 at the best-marked checkpoint and -0.26 at the last under the sequential code ($+0.43$ and -0.26 under the previous code-length accounting). With six points neither is distinguishable from zero. Against how well a run reconstructs the published national components the correlation is $+0.43$ ($+0.14$ before), which six points cannot distinguish from zero either.

So §9's remedy stands where the evidence leaves it: built; 2 to 3×10^5 bits shorter than the open model at each of four seeds when we fit both from scratch and score them by the sequential code; within 4×10^4 bits of it when we make the two one model and split them; indistinguishable from it in held-out prediction at six seeds; and recovering the aggregate we denied it in two runs of six. What this section also establishes is the size of the instrument's noise. A fit's learned algorithm length depends on where the optimizer stopped by about as much as on which of the two architectures we fitted. With from-scratch paired differences scattered by 1.3×10^5 bits under the sequential code (4 to 6×10^5 under the previous code-length accounting, most of it the pullback's), four pairs resolve an effect of 3×10^5 and an effect of 10^5 would take about fifteen.

The scatter is not an artifact of how we score a fit. For the fits continued in §B.2, averaging the total over a run's final 500 epochs instead of reading one checkpoint leaves the paired differences scattered by 3.3×10^5 bits. Re-searching every quantization step at fixed weights, on a grid four times finer than the one training uses, shortens a total by a median of 1.3×10^4 bits across 48 checkpoints and leaves the scatter where it was; fits of one architecture are still 1.4×10^6 parameter bits apart afterwards, because the fit writes the trade between parameter and residual bits into the weights. What differs between two seeds is the model, which is why the paired branch above, in which both arms share one, is the comparison that resolves anything—and it resolves closure to about 10^5 bits of nothing.

10 Intervening: cause(C), and no causation without dynamics

The rule of this section is no causation without dynamics, stated in §3.4 and tested here on an operator whose first version broke it.

cause(C) is the smallest forcing of the identified recurrence that realizes condition C in the decoded observables: of the displacements d of an observed object's latent system state whose

decoded reading satisfies C, the one the recurrence codes cheapest as a transition residual,

$$d^* \;=\; \arg \min_{d:\, C(\mathrm{dec}(z+d))} \; d^{\mathsf{T}} \Sigma_e^{-1} d, \tag{6}$$

with Σ_e the covariance of the recurrence's own residuals over its fitted transitions. Here d is a tangent vector at z and the quadratic form is the Hessian of the residual code length: the bits the residual code adds when d displaces the residual are $d^{\mathsf{T}} \Sigma_e^{-1} d/(2 \ln 2)$, exactly for a Gaussian residual at fixed covariance, so the smallest displacement is the cheapest forcing and not only to second order. Information geometry names this metric the Fisher information metric of the residual distribution. The quadratic form is the worked example's: its recurrence codes its residuals as Gaussian at a fitted covariance, so the cheapest transition residual is the smallest in that metric; for a recurrence of another kind, an automaton among them, the definition is the sentence before the equation, the forcing the recurrence's own residual code charges cheapest. We apply it at one transition, the recurrence propagates it, and the decoder reads it with intervals; the decoder reads every other observable off the system state the recurrence reaches. It severs nothing and there is no graph. Every result carries two diagnostics—how much of the requested change survives re-encoding, and what fraction of observed objects end in system states within the recurrence reach—and we read every result as a difference from the baseline trajectory, which removes the non-extrapolating shared displacement from the difference equation, though §10.1 bounds what survives it; we pose interventions as contrasts, a condition that moves one item and holds another.

Everything in that definition is an object of the dynamics, as the rule requires: the encoder's coordinates and their Euclidean norm, the covariance of the latent population, the decoder's cross-sectional correlations and the density of observed system states would each be unchanged for a corpus with no time axis, so each is a statistic of the cross-section and none may define the intervention, its geometry, its cost or its recurrence reach. The warrant the paper claims is the autonomy of the recurrence, so the recurrence has to supply all four. §10.2 reports what the operator answered before it did. The same rule is what separates this operator from the reading of cause as the shorter factorization of a static joint: that asymmetry exists for a corpus with no time axis, so here it may describe but not warrant.

Because the system state is the whole measurement event, a condition on one item necessarily moves others; we report those co-moves rather than assuming them away. They are the honest content of refusing exogeneity: in a closed system the co-moves are not contamination to purge, they are what intervening consists of. (6) decides which co-moves accompany the condition—the direction of the push is the one the recurrence's residual code finds cheapest—and then by the recurrence carrying the system state forward. Minimum norm remains a choice—it encodes an assumption that an intervention disturbs the system by the least forcing the system could have produced itself—and the data do not establish it; what the data do supply is the metric, and §10.2 measures how much that mattered.

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 §8. One asymmetry favours the recurrence: its autonomy is testable out of sample, and we test it.

The reachability test is not decorative, and it too is the recurrence's. A system state is

reachable if it lies within the recurrence's residual radius—the 99th percentile of its fitted residuals, in the metric of (6)—of the one-step image of some observed system state: the system could have produced it in one step from a system state the instrument observed it in. A result whose intervened system states are largely unreachable describes observed objects the system does not produce; we report it flagged and do not quote it, and we also read every effect among the reachable observed objects. In this application reachability disqualified the two income contrasts, as the density test it replaces had (§13); the intervened branch of the primary contrast is reachable for 76 to 88 per cent of counties across base years and horizons, and the effect among the reachable counties agrees in sign with the median over all counties for every effect and in magnitude by a median factor of 0.95. The residual radius's percentile is the one quantity in this test we set by hand, so we sweep it (`support_audit.py`, which re-reads every threshold-dependent figure from the captured per-county distances): from the 90th percentile of the fitted residuals to the largest of them, the effect among reachable counties agrees in sign with the all-county median in all 84 effects, with magnitudes a median 0.81 to 1.00 of it; at the 80th percentile, where fewer than half the counties are reachable, 82 of 84. The two income contrasts stay reachable for under half the counties at every horizon up to the 99.9th percentile, and one of them at the largest residual itself.

We pose interventions as *contrasts*: cause(C) with a condition C that moves 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, the pairing adds nothing to the reading: every cause(C) result is a difference between two system states propagated by the same recurrence, the intervened and the baseline, and the explicit shared term cancels from that difference whatever the condition. Writing the baseline as z_t and the intervened system state as $z_t + d_t$,

$$d_{t+1} \;=\; d_t \;+\; \Delta t \,\big[\, f(z_t + d_t) - f(z_t) \,\big], \tag{7}$$

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

10.1 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 (7) 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 the report divides an effect reported as a percentage by a baseline that the shared path moves.

The first route is small, and it has a bound. 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.

Direct audit bounds the second route, and it is the only way to reach it: it runs through a non-linear 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. The script

	from 1980		from 1990	
outcome at +30 years	fitted path	range	fitted path	range
female-headed families with children	+13.5	[+12.0, +13.5]	+17.1	[+11.4, +17.1]
never-married women	+3.4	[+2.5, +3.4]	+6.0	[+5.2, +6.2]
next generation's bachelor's share	−0.6	[−0.7, −0.6]	+2.1	[+1.6, +3.6]
median home value	+0.8	[**−0.4**, **+0.8**]	−3.4	[−4.7, −1.0]
births per 1,000	−1.8	[−2.0, −1.8]	−2.0	[−2.4, −1.6]
children under 5	−2.7	[−2.7, −2.6]	−3.9	[−3.9, −2.4]
married couples with own children	−2.5	[−2.7, −2.5]	−4.3	[−4.3, −3.8]

Table 1: The primary contrast re-read under the shared paths that keep both trajectories in region—two from the 1980 base, four from 1990—per cent change at the median county; the first column of each pair is the fitted path, computed under the recurrence's residual metric. Signs hold except for median home value from a 1980 base.

applies each path identically to both branches and 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.

Nothing in the audit is a scenario within the recurrence reach: it is a sensitivity to an exogenous input, and we read it under the density test of the earlier convention. Under the recurrence's own reachability test (§10) none of the alternative paths is reachable: the one-step images of observed system states carry the fitted shared displacement, so a doubled, reversed, zero or frozen path lands outside the residual radius of every one of them by construction. The contrast itself is the recurrence's, computed under its residual metric; only the region test is the older one.

We drop most of the paths unread, and the most recently fitted shared displacement, sustained, is among them: it takes most counties out of the data region within three decades. Six of the eight paths from the 1980 base, and four from 1990, carry the trajectories out of the data region and we drop them before reading anything from them. Doubling the shared displacement, reversing it, and holding the 1980 rate leave the region within one or two steps from either base; holding the 2020 rate for the whole run drifts out of it over three decades, to 0.23 of counties in region from 1980 and 0.53 from 1990; the zero path and the mean historical rate leave it from the 1980 base only. That is the non-extrapolability of §9 appearing as a failure of the region test rather than as a confident wrong number, which is the behaviour the region diagnostic exists to produce.

The magnitudes are conditional on the shared path; the signs are not, with one exception. On the paths that stay in region, magnitudes move by a median of 0.16 times the effect the fitted path gives over the 84 effects—0.08 from the 1980 base, where only two paths stay in region, and 0.36 from 1990, where four do; at thirty years the two medians are 0.13 and 0.38, and an effect that is itself small on the fitted path can move by more than its own size, which happens for four effects, three of them median home value. The reader should therefore take every magnitude in §12 as conditional on the fitted path. Across two base years, two contrast directions, seven outcomes and three horizons—84 effects—83 hold the sign the fitted path gives. The one that does not is median home value from the 1980 base at thirty years, and from the 1990 base the same outcome moves across its whole fitted effect without quite crossing zero.

The signs survive the two thresholds set by hand, every loss falling on median home value or

the next generation's bachelor's share. Which paths stay in region depends on the percentile of the nearest-neighbour radius and the in-region share a path must keep at every horizon, and the two trade off against each other. The pair used here, the 99th percentile and 0.90, sits on a line of settings that admit the same two paths from 1980 and four from 1990 and return the same 83 of 84 and 0.16. Admitting one path more from the 1980 base, the zero path, returns 80 of 84 and a median of 0.35; admitting five and seven paths at the largest observed density radius, the doubled and reversed paths among them, returns 77 of 84 and 0.63. Every sign lost at any setting is in the next generation's bachelor's share, one effect, or in median home value, up to six; married couples with own children, female-headed families with children, never-married women, births and children under five hold their sign at every setting of both thresholds.

Median home value is a systematic exception rather than a scattered one, and the reason is legible: no one can separate a nominal contrast from a shared path that is mostly the price level, a fragility of the worked example's nominal items and not of the methodology. This is the same outcome §12.1 finds changing sign between base years; it was already outside the set we state without qualification, and what the audit adds is the reason. 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 composes it: at thirty years the primary contrast on median home value goes from −\$3,818 to −\$451 at the median county from the 1990 base, the one base from which the zero path stays in region. No real-quantity outcome behaves this way.

Population growth, the business cycle and the price level compose the shared displacement, which is what the remedy of §9 predicts; we did not choose that split to fit this measurement. The same decoding says so: the displacement's five largest movers are population change, population density, women's and men's unemployment rates, and families below poverty, with the price level behind them. Those are three components, and the remedy of §9 already sorts all three: growth and unemployment are aggregates of the county cross-section, and the price level is genuinely exogenous to a county panel.

The worked example reports directions and withholds magnitudes: it yields qualitative comparisons outright and quantitative ones only conditionally. The cancellation of the explicit term is what separates the two kinds of comparison. A forecast is a quantitative comparison, a predicted quantity set against the quantity later observed, and it needs the explicit term: the component one would have to extrapolate to predict a quantity never enters (7) at all. A contrast, the difference d_t of (7), yields both kinds. The cancellation makes the qualitative comparison, which way an outcome moves under the condition, independent of the explicit term, and the audit above finds that comparison nearly indifferent to the trajectory behind the term; the quantitative comparison, by how much, moves with that trajectory by a measured fraction, the weaker and measurable dependence which the contrasts of §12 are conditional on. The remedy of §9 does not remove that dependence but changes its standing. A closed system has no shared displacement to choose, but its contrasts still run along the trajectory it generates, so the dependence becomes a property of the fitted system rather than of an input, and we have not yet run the propagator of §10 on one.

10.2 Which displacement is smallest

The first version of this operator took the minimum-norm displacement in the Euclidean geometry of the coordinates the encoder happens to emit. That is a statistic of the fit, not a property of the dynamics, and nothing selects it: write the latent in new coordinates $z' = Az + b$, re-express the decoder and the recurrence, and every observable prediction is unchanged, but the Euclidean-smallest displacement realizing C is unchanged only if A is orthogonal. With two decoded changes $X = \Delta z_1 + \Delta z_2$ and $Y = \Delta z_1 - \Delta z_2$, the condition $\Delta X = 1$ gives $\Delta Y = 0$ in the original coordinates and $\Delta Y = (1 - a^2)/(1 + a^2)$ after rescaling z_1 by a, so $a = \frac{1}{2}$ and $a = 2$ co-move Y in opposite directions from the same model and the same condition. The example is a reviewer's, and so is the experiment: reparameterize the frozen model, confirm its predictions are unchanged, recompute cause(C). The encoder's coordinates as learned have scales spanning a factor of 5; the receiver recomputes the latent, so only the quantized parameter code sees the parameterization at all.

`coordinate_audit.py` freezes the checkpoint, the base-year cohort, the intervention, the propagator and the reporting, and varies only the metric M in which the displacement is smallest: the Gauss–Newton step becomes $M^{-1}J^{\mathsf{T}}(JM^{-1}J^{\mathsf{T}})^{-1}$ applied to the remaining gap, which for $M = I$ is the convention and for $M = \Sigma_e^{-1}$ is (6). We run the literal experiment first. With the encoder's last layer scaled by a diagonal D and the decoder's first layer by D^{-1}, the frozen model's decoded means agree with the original to 2×10^{-6} and its residual code to the bit, and its cause(C) co-moves agree with the metric formulation at $M = D^2$ to 3×10^{-6} against a median co-move of 5×10^{-4}; a random rotation, under which the Euclidean norm is invariant, reproduces the convention's numbers to 2×10^{-6}.

We then read what the convention's directions were worth off six random diagonal rescalings, log-uniform per coordinate by factors up to 2 and up to 4, and three further metrics that are also statistics of the cross-section—the latent population's covariance, and the decoder's Jacobian pulled back in chart coordinates and as multiples of its residual variance per measurement event. None of these is a candidate geometry; under the rule of §10 none could be. They are the measurement of how much the convention's answers depended on the convention: across those that keep the trajectories in region, the sign held in 31 of the 84 effects—two base years, two contrast directions, seven outcomes, three horizons—and the range of the effect across geometries was a median 3.9 times the convention's effect. Only female-headed families with children kept its direction in every one of its effects. The directions the earlier tests of §12 had found stable across panel versions, decoders and shared paths were, to that extent, facts about the encoder's coordinates. This is the many-analysts finding of Silberzahn et al. [2018] reproduced inside one team: the same corpus, the same model, and a defensible choice no one had examined moving the answer. What differs is what the rule of §10 then does with the choice. It is not one more specification to report alongside the others; the rule disqualifies it, because nothing in the dynamics selects it, and the recurrence's own residual metric is what remains.

The recurrence's residual metric answers differently, and we test its answers against the two things that could undermine it. It agrees with the convention in sign in 72 of the 84 effects, with magnitudes a median 1.41 times larger; the disagreements are births and children under five, which under the convention rose when women's participation rose with men's employment held and under the recurrence's metric fall, and the nominal home value. The residual scale is not itself invariant across decades: each decade's own residual variance repays its cost under the two-part test, the largest per-decade to pooled standard-deviation ratio being 1.14. Reading

	from 1980		from 1990	
outcome at +30 years	convention	recurrence	convention	recurrence
female-headed families with children	+10.5	+13.5	+11.6	+17.1
never-married women	+3.0	+3.4	+6.1	+6.0
next generation's bachelor's share	+0.7	**−0.6**	+2.9	+2.1
median home value	+5.0	+0.8	−0.6	−3.4
births per 1,000	+3.2	**−1.8**	+1.4	**−2.0**
children under 5	+2.0	**−2.7**	−0.5	−3.9
married couples with own children	−0.5	−2.5	−2.4	−4.3

Table 2: The primary contrast at thirty years from two base years, under the encoder-norm convention the operator first used and under the recurrence's residual metric (6); per cent change at the median county. Bold marks a direction the recurrence reverses relative to the convention. Under the recurrence's metric twelve of the fourteen outcome-by-direction patterns keep one sign across both base years and all three horizons; the exceptions are the bachelor's share and the nominal home value under the primary contrast.

every contrast with the base decade's own residual metric instead of the pooled one changes the sign of 4 of the 84 effects, all of them small, on births and home value, and the magnitudes by a median factor of 0.96. And the reachability test, carried into the same geometry, leaves the intervened branch reachable for 76 to 88 per cent of counties, with the effect among reachable counties agreeing in sign with the all-county median for every effect. Table 2 reads the primary contrast at thirty years under the convention and under the recurrence.

The rule paid for itself twice here. It removed a dependence on a convention that the earlier stability tests could not see, because they were all run inside it; and it moved two directions. The dependence that remains is the one the data genuinely leave open: whether the forcing an actual policy applies is the one the recurrence finds cheapest. A cost model of a real intervention would settle that, and this corpus contains none; until one does, the contrasts of §12 are the recurrence's own answers to its own cheapest forcing, and the directions we report are those that survive the base year, the horizon, the decade's residual scale and the panel version.

11 Validation on withheld transitions

On transitions withheld entirely from fitting, the model sits 0.40 to 0.54 bits per datum above the oracle, and that margin is flat whether the step is four years or ten. We report this as the audit described in §3, 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 tenfold, from 0.99 to 10.59 bits per datum, because assuming no change over a decade fails catastrophically while assuming no change over four years does not. The 0.955 reports a collapsed baseline, not an easy problem. Excess over the oracle 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 the model to predict a four-year and a ten-year step equally well. One caveat on the last row: the 2016–2020 and 2020–2024 windows share the 2020 sample year, so the base and target of that transition are not independent samples, which

held out	step	oracle	model	persistence	skill	model − oracle
2000–2010	1.0 dec	6.362	6.840	16.948	0.955	0.479
2010–2015	0.5	3.528	3.934	4.518	0.590	0.405
2015–2020	0.5	3.542	3.974	4.873	0.675	0.432
2020–2024	0.4	3.597	4.140	4.850	0.567	0.543

Table 3: Held-out transitions, bits per datum, from the fits with the point mass, each scored at its shortest-coding checkpoint with the withheld measurement events never in the loss. *Oracle* codes the target period from the encoder's own embedding of it: a reference rather than a bound, since the encoder is a fitted map and not the minimizer of the code over latents, so a recurrence could in principle land below it; none here does. *Persistence* decodes the base period's system state. We compute the chart's per-item mean, variance and s_{95} over every measurement event, withheld ones included, which is all the withheld measurement events contribute to the fit. Skill is $(b_{\text{pers}} - b_{\text{model}})/(b_{\text{pers}} - b_{\text{oracle}})$.

flatters persistence and model alike; the two sub-decade rows above it have no such overlap.

The learned algorithm length under (4) is 7.59 bits per observed datum: the transitions, 89% of the data, at 6.63 against a null of 9.08, and the 1970 and 1980 start periods at the null's 10.98. As a diagnostic, the pullback reproduces the panel at 6.52 bits per datum against 9.29; permuting the correspondence between counties and their latent codes costs 46 bits per datum. Nominal 90% predictive intervals cover 92% of observed data.

12 Application: three directions of family structure, and what stability looks like

Every number in this section is the recurrence's answer under (6), read off the model with the point mass of §B.2, 0.87×10^6 bits shorter under the sequential code than the same checkpoint with the point mass switched off. The contrast reported here is one question put to the fitted recurrence, chosen to illustrate the reading; the operator poses any other condition on any items of the cross-section to the same learned algorithm and reads it off in the same way, and we built nothing in the model for this one. We test the motivating contrast and its mirror five ways: across the two earlier panel versions that we can still run, which differ from the final one by seven items and a time point; across the change of decoder; across two base years and three horizons; in §10.2, across the residual scale of the decade the contrast starts from; and across six seeds of the fit, each continued with the point mass. Three directions survive all five, and they are one thing: when women's labour-force participation rises with men's employment held, female-headed families with children rise, never-married women rise, and married couples with own children fall, and the mirror reverses all three. Children under five fall under the final model from either base year but crossed zero as identification improved across panel versions, and births did the same; we assert a direction for neither. The next generation's bachelor's share and the nominal home value change sign between base years and we assert no direction for them.

Across six seeds of the open fit, each continued with the point mass, the three directions and their mirror keep their signs in all 36 of their comparisons, a comparison being one outcome under one direction from one base year at one horizon. Over all seven outcomes, 79 of 84 comparisons keep the sign the paper's model gives under every seed; the five that do not are births from

outcome at +30 years	770-item	777-item	777-item, point mass
female-headed families with children	+15.1	+14.5	+13.5
never-married women	+9.9	+7.6	+3.4
next generation's bachelor's share	+2.8	+1.6	−0.6
median home value	+0.9	−0.0	+0.8
births per 1,000	+2.1	+0.5	−1.8
children under 5	+1.2	−0.2	−2.7
married couples with own children	−2.5	−3.3	−2.5

Table 4: cause(women's labour-force participation up, men's employment held), one cross-county standard deviation, per cent change at the median county under (6), across the two earlier panel versions that the current panel builder can still run (the 525-item checkpoint no longer loads) and, in the last column, the same panel with a point mass in the decoder. The outcomes are items of the county panel: female-headed families with children and married couples with own children are family households with own children under 18, female householder with no spouse present (`HSD580`) and married couple (`HSD520`); never-married women are females 15 and over never married (`POP540`); the next generation's bachelor's share is the per cent of persons 25 and over with a bachelor's degree or higher (`EDU685`), read thirty years on; births are per 1,000 population (`VST020`); children under 5 are the resident population under 5 (`AGE110`); median home value is the median value of specified owner-occupied housing units in nominal dollars (`HSG490`). The counts enter the chart per capita.

the 1980 base at ten and twenty years, the bachelor's share from 1980 at thirty years and the nominal home value from 1990 at ten and twenty years, outcomes for which we assert no direction. Children under five keeps its sign in all twelve of its comparisons, and we still assert no direction for it, since it crossed zero across panel versions. Over the 36 comparisons of the three directions, the coefficient of variation of the effect across seeds, its standard deviation over its absolute mean, has median 0.06, lies at or below 0.12 in 33 comparisons and reaches 0.31 in the most variable one, married couples with own children from the 1980 base at thirty years, where the largest magnitude among the six seeds is 3.1 times the smallest. The first four seeds stop their continuation at the checkpoint the earlier pullback code-length accounting chose and the last two at the checkpoint the sequential code chose, so the comparison varies the stopping epoch with the initialization and does not separate the two.

The change of decoder moves the magnitudes by a median factor of 0.98 from the 1980 base and 0.90 from 1990, with an interquartile range of 0.7 to 1.2, and keeps the sign of 369 of the 375 contrasts whose own interquartile band excluded zero before it. Among the seven reported outcomes it changes no sign from the 1990 base; from 1980 it changes births at every horizon, children under five at twenty years, the bachelor's share at thirty and the nominal home value, which are the crossings Table 4 shows.

Table 4 reports the contrast. The mirror—men's employment up, women's participation held—reverses family structure: married couples with own children +8.5%, female-headed families with children −14.2%, never-married women −5.5%. That the same three variables move oppositely at comparable magnitude under the mirrored contrast is the strongest internal check the method affords, and under the recurrence's metric the mirror reverses all seven outcomes from both base years.

The movement across panel versions is not a rescaling. The operator realizes the intervention

identically in every version, and the per-outcome ratios from the 770-item version to the final model are heterogeneous—0.9 for female-headed families with children, 0.3 for never-married women, 1.0 for married couples with own children—where a scaling artifact would be uniform, and births and children under five cross zero. The outcomes that moved most are the fertility ones.

12.1 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	−2.5	−4.3	1.72
female-headed families with children	+13.5	+17.1	1.27
never-married women	+3.4	+6.0	1.76
next generation's bachelor's share	−0.6	**+2.1**	−3.50
births per 1,000	−1.8	−2.0	1.11
children under 5	−2.7	−3.9	1.44
median home value	+0.8	**−3.4**	−4.25

Table 5: The same contrast propagated from two base years, same model, same reporting. Two of seven outcomes change sign.

Two of seven change sign: the next generation's bachelor's share and median home value. Both are the smallest effects from the 1980 base, −0.6 and +0.8 per cent, so here the casualties are the effects nearest zero, and one of them is a price. The five that hold are four composition measures—married couples with own children, female-headed families with children, never-married women, children under five—and one flow rate, births, at −1.8 and −2.0 per cent. We record the pattern without a mechanism for it.

This is the trajectory dependence of §10.1 made visible: the base trajectory differs between the two runs, so the Jacobian through which the recurrence propagates the intervention differs, and so does what the nonlinear decoder makes of the latent difference. The baseline that divides each percentage differs too, but a positive denominator changes a magnitude, not a sign.

Two conclusions, and the second is the one that should govern how anyone uses these numbers. The movement across panel versions is most likely confounding removed as identification improves, and the fertility effects are where it shows. But the contrasts are *conditional on a base year and a shared path*, and only the three directions that also hold across panel versions are the ones this paper is entitled to report without that qualification. Producing them with a band over plausible base years and shared paths is unfinished work, not a refinement.

12.2 The same contrast held for three decades

Every contrast above is an impulse, the diagnostic kind of §3.4: one forcing at the base year and the recurrence left to itself. We also put the sustained kind to the same recurrence, with the same operator applied at every step of the intervened branch. After each decade the recurrence

contrast	base	move (sd)	forcing base, +10, +20, +30	retained +10, +20, +30	reachable (%) +10, +20, +30	released +40, +50
primary	1980	1	0.98, 0.50, 0.43, 0.17	0.58, 0.62, 0.76	40, 18, 32	0.69, 0.57
primary	1980	1/2	0.46, 0.21, 0.18, 0.06	0.61, 0.63, 0.77	100, 91, 89	0.70, 0.59
primary	1980	1/4	0.23, 0.10, 0.09, 0.03	0.63, 0.64, 0.77	100, 99, 98	0.70, 0.59
primary	1990	1	1.01, 0.41, 0.15, 0.17	0.61, 0.78, 0.72	38, 48, 53	0.79, 0.64
primary	1990	1/2	0.47, 0.18, 0.06, 0.07	0.62, 0.78, 0.72	100, 99, 97	0.79, 0.64
primary	1990	1/4	0.23, 0.09, 0.03, 0.03	0.63, 0.78, 0.72	100, 100, 100	0.79, 0.63
mirror	1980	1	1.05, 0.28, 0.20, 0.06	0.72, 0.76, 0.96	49, 41, 53	0.81, 0.64
mirror	1980	1/2	0.51, 0.14, 0.10, 0.03	0.72, 0.77, 0.95	100, 95, 93	0.83, 0.69
mirror	1980	1/4	0.25, 0.07, 0.05, 0.02	0.72, 0.77, 0.95	100, 99, 98	0.84, 0.71
mirror	1990	1	0.98, 0.21, 0.06, 0.11	0.74, 0.90, 0.76	62, 69, 74	0.76, 0.61
mirror	1990	1/2	0.47, 0.10, 0.03, 0.06	0.73, 0.89, 0.77	100, 99, 98	0.78, 0.64
mirror	1990	1/4	0.23, 0.05, 0.02, 0.03	0.73, 0.88, 0.77	100, 100, 100	0.78, 0.65

Table 6: The primary contrast (women's labour-force participation up, men's employment held) and its mirror held at every step for three decades and released for two, from both base years, at one cross-county standard deviation and at half and a quarter of it. The forcing is the minimum-norm displacement in the recurrence's residual metric that restores the condition against the factual branch, as a multiple of the residual radius; at the base year it is the impulse. Retained is the share of the held condition the free run left after a decade, before the correction. Reachable is the share of counties whose maintained system state passes the reachability test of §10. Released is the share of the condition remaining one and two decades after the policy ends, with no correction.

advances the factual branch and the intervened branch, and we re-solve the minimum-norm displacement in the recurrence's residual metric that restores the condition, women's participation one cross-county standard deviation above the factual branch with men's employment held at it, and apply it to the intervened branch; we hold the condition for three decades and release it for two. We did this for the primary contrast and its mirror from both base years, at one standard deviation and at half and a quarter of it, and we judge every step by the recurrence reach as we judge the impulse. Three numbers describe each step: the forcing the condition needs, as a multiple of the residual radius; the share of the condition the free run retained over the decade before the correction; and the share of counties whose maintained system state is reachable. Table 6 reports them.

The recurrence reverts the condition at every step and nowhere comes to hold it. The free run retains 0.58 to 0.78 of the primary condition over a decade and 0.72 to 0.96 of the mirror's; the forcing re-solved after one decade is a fifth to a half of the impulse's and after three decades a twentieth to a sixth of it, shrinking at every step and never reaching zero; released, the condition decays to 0.57 to 0.64 of itself after two free decades under the primary and to 0.61 to 0.71 under the mirror. On the reading of §3.4 this is a policy the system reverts and the policymaker must reapply for as long as the outcome is wanted, not a capital investment, and the recurrence reports the same of the other six specifications of §10. In every run of the eight, from both base years at all three sizes of the move, the forcing at the third decade is between a twentieth and a quarter of the impulse's, except under the income contrast at one standard deviation, where it holds at the impulse's size; and the free run retains between 0.36 and 1.23 of the condition, carrying it past itself in some decade only under the three specifications that move or hold the nominal home value, income, or never-married men, where the forcing re-solved is the move back.

outcome	from 1980			from 1990		
	impulse +30	held +30	released +50	impulse +30	held +30	released +50
married couples with own children	−2.2	−8.6	−6.4	−3.6	−10.2	−9.1
female-headed families with children	+13.3	+43.7	+26.2	+17.4	+38.6	+18.6
never-married women	+3.3	+16.8	+11.1	+5.4	+17.4	+11.2
next generation's bachelor's share	−0.8	+4.6	+4.5	+1.7	+10.8	+10.2
births per 1,000	−1.8	−6.9	−3.7	−2.1	−6.8	−5.3
children under 5	−2.6	−8.7	−6.0	−3.8	−9.7	−6.6
median home value	+1.0	+0.3	−8.8	−3.9	−3.8	−10.8

Table 7: The primary contrast at one cross-county standard deviation: per cent change at the median county under the impulse at thirty years, on the maintained trajectory at thirty years with the condition held at every decade, and twenty years after release, each on the counties its own branch reaches at that horizon (2,559 and 2,707 of 3,146 counties under the impulse, 1,019 and 1,673 held). The released readings lie past the record's last cross-section and reuse its last shared displacement.

At one standard deviation the maintained system state is reachable for a minority of counties: the primary's third held decade for 32 per cent from 1980 and 53 per cent from 1990, where the impulse's branch is reachable for 81 and 86, and the mirror's for 53 and 74. At half a standard deviation the maintained branch is reachable for 89 to 100 per cent of counties and at a quarter for 98 to 100. A policy that holds the whole condition for thirty years carries most counties to system states no county occupied in the record; a policy half its size stays within it.

Table 7 reads the seven outcomes of Table 4 on the maintained trajectory at thirty years beside the impulse response, each on the counties its own branch reaches, and twenty years after release. The three directions keep their signs on the maintained trajectory from both base years at all three sizes of the move, and so does the mirror's reversal of them; their magnitudes at thirty years are two to five times the impulse's, female-headed families with children +44 per cent held against +13 under the impulse from 1980, never-married women +17 against +3, married couples with own children −9 against −2. Of the ten outcomes the operator reports, the two readings agree in sign for nine from 1980 and ten from 1990 under the primary, and for ten and ten under the mirror; the disagreement is the next generation's bachelor's share from 1980, −0.8 per cent under the impulse and +4.6 held, an outcome for which we assert no direction. On the maintained trajectory the mirror reverses the three directions, births and children under five from both base years, as it does under the impulse, but it does not reverse the bachelor's share, which rises under both held policies, nor the nominal home value from 1980. Across the eight specifications, three sizes of the move, two base years and three held horizons, the two readings agree in sign on all ten outcomes in 78 of 143 comparisons and on at least nine in 116; the eight comparisons that agree on seven or fewer all lie under the three specifications that move or hold income or the nominal home value, and the income contrast from 1990 at thirty years reaches no county at one standard deviation and gives no comparison. After release the three directions decay with the condition and keep their signs through both free decades, which lie past the record's last cross-section.

A sustained intervention is a capital investment only where the identified dynamics have a second basin to carry the county into, and we looked for one. At every observed system state of

1980, 1990, 2000 and 2010 we linearized the free developmental map in the recurrence's residual metric. At the median system state it expands in four, three, two and two of the thirty-two latent directions (90th percentile six, five, four and four), its largest eigenvalue modulus 1.09, 1.07, 1.05 and 1.04, and contracts the rest; at no more than five per cent of the system states of any year does it contract in every direction. We then stepped every observed system state three free decades. The cohort's spread fell to 0.65, 0.70, 0.73 and 0.77 of the observed, the median county stood at 0.64 to 0.75 of its observed distance from the cohort's centre, and of the 12,584 system states four ended farther from the centre than they began, by at most ten per cent. Within the record the free map has one basin, the drifting cohort, and the expanding directions carry no observed county out of it. That is the recurrence's reason, in its own currency, for a forcing that shrinks at every step and never reaches zero: the held condition is reverted toward the cohort from wherever it is held. The methodology admits the capital investment of §3.4; this recurrence, on this record, finds none.

The silence has the shape §3.4 gives it. A channel no county bypassed between 1980 and 2020 is a direction along which the recurrence has no residual, and a basin reached only through such a channel is one this recurrence cannot see. The record is of counties observed under the institutions that stood over them in those decades, and in it the method finds no cheap, self-sustaining displacement of family structure. Whether that is a property of the system or of the width of the record is what the worked example cannot tell, and what widening the record toward the Oracle of Evidence would.

The record holds one maintained branch of its own, and we can say where. For every county we took the residual of each observed decade transition against the free developmental map, 1980 to 1990 through 2010 to 2020, in the recurrence's residual metric, and the mean cosine between consecutive decades' residuals, which we call the residual's persistence. The free map is what the county cross-section generates on its own, so a residual it does not generate, re-applied in the same direction decade after decade, is the forcing of a maintained branch, the object §3.4 defines and Table 6 computes for a policy. At the median county the persistence is -0.04 and at the 90th percentile 0.14: the residual of one decade says nothing about the next. At the District of Columbia it is 0.69, the largest of the 3,146 counties, and at Loudoun, Anne Arundel, Fairfax and Prince George's counties 0.29, 0.27, 0.26 and 0.21, the 34th, 48th, 56th and 144th: the District and four of the counties around it stand among the most persistent five per cent. The residual's size at these counties is ordinary (Fairfax, Prince George's and Montgomery counties rank between 1,779th and 2,256th by its mean norm), and persistence is only weakly a property of outlying system states (Spearman 0.20 with the observed distance from the cohort's centre; of the hundred farthest counties, seven are among the hundred most persistent), so the capital's persistence is not its distance. Decoded, the forcing raises households with incomes over $150,000 and $200,000 at Fairfax and Loudoun above what the free map carries, and the foreign-born and Asian population at Loudoun with them; at the District it raises employment in public administration; at Prince George's it holds Defense expenditure, the Asian population and the highest rents below the map's free image; its leading items move by half to one cross-county standard deviation a decade. Its source is not an item of any county. The worked example admits no exogenous input, and the appropriation that pays the federal workforce is not a function of the county cross-section, so the channel that maintains the capital is silent in the sense of §3.4, and the silence lies in the counties that administer the policies the methodology would be imposed on. The method does not name the forcing. It shows that the counties with the most influence on macrosocial policy are, in the record of counties, the ones a forcing the county system does

not generate holds where they are, which is where §3.4 said the intermediary's interest would appear. The other reading, that a quadratic map fitted where no county has neighbours misfits the capital in a fixed direction, the weak tie between persistence and distance argues against and does not exclude; the District is also the most unusual system state in the record, and the second most persistent county, Ocean County, New Jersey, lies in no capital's metropolitan area.

13 What the worked example does not establish

The worked example establishes directions, not magnitudes, and directions only under the recurrence's own geometry; relations among counties, not among people, because its corpus is county tabulations; and contrasts, together with forecasts of what the recurrence generates and not of the one component we supplied from outside it. Its own reachability diagnostic disqualifies two of the eight contrasts we attempted. Each limit is the worked example's, and where the methodology removes it we say so. 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 reported here 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. This is a limit of the worked example's corpus and not of the methodology. We built the corpus from public tabulations, so the observed object is the county and nothing inside the data can dispel the fallacy. Inside the enclave the observed object is the respondent and the measurement events are the individuals' own, and the custodian does not merely show them to the learning algorithm: the learned algorithm it produces must generate them, and the criterion scores it on every bit of every one of them. A relation it asserts between education and childlessness is then a relation among the women it generates, coded against each woman's measurement event, and an account of the evidence that reproduced the county totals while generating the wrong individuals would pay, in bits, for every individual it got wrong. The methodology's answer to the ecological fallacy is the sealed individual corpus itself, which the mechanism of §4 exists to make scorable; the worked example, run outside the enclave, could not use it.

Forecasts of quantities that depend on the supplied component. Propagating a quantity forward requires the shared displacement, the one component we took as given, and §8 shows it is the one component that does not extrapolate, as the first claim says of any item a law does not generate. 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 with children. A contrast removes that component from the difference equation; it does not make the result independent of it (§10.1).

Interventions the system does not produce. A one-standard-deviation move in correctly scaled income leaves the recurrence reach: the intervened system state is reachable for 1 to 7 per cent of counties. We report such results flagged and do not interpret them. The diagnostic disqualified two of the eight contrasts attempted; the primary contrast's intervened branch is reachable for 76 to 88 per cent of counties, and the effect among those agrees with the all-county

median in sign for every effect. The residual radius's percentile is the one hand-set quantity in the test; swept from the 90th percentile to the largest fitted residual, the sign agreement holds for every effect and the income contrasts stay disqualified to the 99.9th (§10).

Sustained interventions, held for three decades in one record. Every contrast of §12 but those of §12.2 is an impulse, the diagnostic kind of §3.4, and every direction they report is the direction of a transient. §12.2 holds the primary contrast and its mirror for three decades and releases them for two, and what it reports is bounded three ways. The held horizon is three decades, so a forcing that shrinks at every step and never reaches zero within it is reported as reverted and not as converging; the maintained branch at one standard deviation is reachable for a minority of counties, so its outcomes are read on that minority; and the released decades lie past the record's last cross-section and reuse its last shared displacement. The free map carries no observed system state away from the cohort within three decades, so the worked example reports no basin other than the drifting cohort and no capital investment. The methodology admits both. Whether the record is silent because the system has no such basin or because no county in it was ever carried into one, the worked example cannot tell. The one maintained branch the record shows on its own, at the capital, the worked example reads from the persistence of a residual over four transitions, and it cannot name what maintains it.

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 (§10.1). A doubling stress test bounds the latent route at 2.8%; audit bounds the decoding route at a median of 0.16 times the fitted-path effect in magnitude (0.08 from the 1980 base, 0.36 from 1990), with 83 of 84 signs holding across the shared paths that stay in region; we read the paths under the density test because none of them is reachable under the recurrence's own. The density radius and the in-region share a path must keep we set both by hand; over settings that admit from two to seven paths, the median moves from 0.16 to 0.63 and the signs held from 83 to 77 of 84, every loss in bachelor's share or home value (§10.1). Magnitudes are conditional on the shared path. Separately, §12.1 shows two of seven outcomes changing sign between two base years; five 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.

Geometry of the displacement. The smallest displacement is the tangent vector of least length in the recurrence's residual metric, an object of the dynamics that is invariant to the latent's parameterization but whose scale varies by up to 14 per cent across decades; the base decade's own metric changes 4 of 84 signs, all near zero (§10.2). Under the encoder-norm convention the operator first used, a statistic nothing selects, only 31 of 84 signs survived reparameterization, and two of the directions reported under it—births and children under five—were the convention's rather than the recurrence's. What the criterion removes is discretion about the account of the evidence. The operator inherits that account's geometry, and the choice between the pooled residual metric and each decade's own is a fork the invariance test bounds, at 4 of 84 signs and a median magnitude factor of 0.96, but does not close. The crowd of Silberzahn et al. [2018] converged on a direction and stopped at the mechanism, because its corpus had no dynamics; the worked example converges on directions because its corpus has them, and it reports, in the same units of measurement, where the geometry those directions rest on stops being invariant.

Magnitude. See §12. The worked example establishes the direction of an effect as firmly as it can establish anything, under the recurrence's own geometry; its size it does not.

Choice of the family of models. The worked example minimized the learned algorithm length within a chosen family of models, and within that, a chosen architecture. Neither choice is free, and nothing in the worked example says what a different family of models would have found; under the mechanism that is a rival submission's to show, and the criterion ranks the two by their lengths. The worked example's count carries no line for its family of models: the manifold autoencoder's shape, the mixture head and the function library were primitives of the machine it ran on, and the table of §7.2 charges the quantized weights and not the lines that interpret them. Its arithmetic was the 32-bit floating point of a graphics processor, with no charged library in an integer instruction set fixing the order of its reductions, so the low-order bits of its counts depend on that hardware, a dependence the worked example did not measure. Under the mechanism both are lines on the custodian's published machine, counted with the rest (§3); here they are defects of the worked example and not of the methodology.

Five of these are themselves measurements—the reachable fraction, the forecast divergence, the attenuation sequence, the spread across geometries, the sweep of the reachability and density thresholds—which is the property a benchmark should demand of every submission: that the same instrument locate its boundaries as locates its results. A submission that reports where it fails, in the units of measurement in which it reports its successes, is one a custodian can rank; one that does not is a claim of standing.

14 Conclusion

The two claims of §1 are what the worked example exercised. The first, that the law of motion is the foundation and the statistic its degenerate case, is what each of the three failures of the dynamics reported below is a case of: a statistic stood where a recurrence should have, and the criterion found it because it measured for it. The second, that the many-analysts design wants a criterion and not more analysts, is what §10.2 reproduced inside one team, a defensible choice no one had examined moving the answer; what the rule of §10 then did with the choice, disqualifying it rather than reporting it alongside the others, is charging the incremental cost in bits of the defensible choice at the scale of one submission.

P.L. 115–435 provides access to the data and no criterion for ranking accounts of the evidence. 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 evidence is better—a model selection criterion, from which the data question follows—and in that vacuum authority exercised where the public cannot follow settles both. A criterion exists that does not require the public to follow: the learned algorithm length of a sealed corpus, with learning algorithms submitted and only the bit count returned. Its disclosure is the count and nothing else, bounded by the count's own length; 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 cleaning defects, each applied as data selection and invisible to the objective, one verification step that was itself defective, and three 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—two of seven reported outcomes change sign when the recurrence propagates the same intervention from a different base year. The third is that the displacement cause(C) first computed was smallest in a geometry the data do not select, the encoder's coordinates, and its signs survived reparameterization in 31 of 84 effects; the recurrence's own residual metric now defines the operator and the recurrence reach bounds it, under which twelve of fourteen directions hold across base years and horizons and two of the earlier directions reverse. Read carelessly the first two look like evidence against the premise. Read correctly they are the premise operating: a relation that is not autonomous fails exactly where autonomy is what it needed, and the failure is visible because the method measured for it. The third is the premise having been broken—a statistic of the cross-section standing where the dynamics should have stood—and then repaired; no causation without dynamics is the rule, and the paper's first operator did not keep 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 of measurement as the claim itself, so this paper reports, of the seven outcomes it read for one contrast of family structure, the three that keep their sign under every test, and their mirror, and reports the failures of the other four without asserting a direction for them. The diagnostics found two of them rather than an objector, and an objector whom the diagnostics then confirmed found the third, which between them is the behaviour a touchstone information criterion must produce if anyone is to trust it where no objector has access. And the first two converge on one remedy—close the system, generate the aggregate rather than assume it—which removes the first by construction and turns the second from a choice of input into a property of the fitted system; we have not yet read the contrasts of §12 off a closed model. Built and split from the open system it codes the corpus to within 4×10^4 bits of it; fitted from scratch it codes the corpus 2 to 3×10^5 bits shorter at four seeds of four; and it recovers the aggregate we denied it in two fits of six. The remedy costs nothing; what it buys is beginning to show, and measuring it is the 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. The same sums fix the size of the prize the mechanism must offer, and a prize of that size is an appropriation. 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 someone asks the public to take an unfalsifiable claim on trust.

Reproducibility. The worked example—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

We minimized the criterion within a family of models we chose, so the obvious question is how much of the result is the family of models. We ran the sweep that Table 8 reports on a 90/10 split of the county cross-section, varying latent width, mixture components, hidden width and

the reconstruction head. The sweep is the many-analysts design of §1 run inside one team: eight rival architectures on one corpus, and the criterion ordering them, which is what the mechanism of §4 does across teams.

run	latent	components	hidden	total (Mbit)	bits per datum
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 8: Architecture sweep. The learned algorithm 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 learned algorithm 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—we ran this sweep on the cross-sectional model and gave the panel model its settings rather than repeating the search. A submission to the benchmark of §4 would have no reason to inherit anything, which is the point of scoring rather than asserting.

Restarts are search, and search belongs inside the learning algorithm. Fitting is not convex, so where we compare two architectures we run each from several seeds and report the shortest, each seed a rival submission and the criterion choosing among them. 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 §4 there is nothing to correct, because nothing is being estimated. The custodian runs a submitted learning algorithm and publishes the bits its learned algorithm achieved. A learning algorithm that restarts internally spends its submitter's compute and returns one learned algorithm; the criterion never asks what learned algorithm length was *attainable*, only what this one attained. That is a quiet advantage of scoring learned algorithms over estimating an infimum, and it is the same property that lets cleaning sit inside the scored object. It is also where the usual objection to algorithmic information theory, that it is uncomputable, misses its mark: the certificate of minimality is uncomputable, the search is not, and the benchmark runs the search (§3).

Within this paper the objection stands, and we handle it the only honest way: we match K across the architectures we compare, pair the runs by seed, and report the spread alongside the difference. Seed-to-seed standard deviation of a learned algorithm length at fixed architecture is 2.0 to 2.5×10^5 bits under the sequential code (2 to 5×10^5 under the previous code-length accounting), 0.02 bits per datum, and one run moves by as much between two of its own checkpoints. That is the floor any architectural claim here has to clear, and §9.1 is a report of one that clears it when we fit the arms apart and not when we split them from one model.

That the fit needs restarts at all is a fact about the objective rather than a nuisance: where

the optimizer stops moves the learned algorithm length more than the architectural choice under test, and the parameter code is jagged where a norm penalty is smooth [Lan et al., 2024]. 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.

B Inducing the cleaning algorithm

The criterion charges cleaning on the data the learned algorithm describes, which is the whole sealed corpus, whether or not the analysts who wrote the learning algorithm can see it. What the analysts cannot see they cannot hand-write a rule for; the learning algorithm, which reads the corpus inside the enclave, can induce the rule in place, and the learned algorithm pays for it at the lines that carry it. We ran the induction reported here by hand on the corpus we hold, so as we ran it it is a stage of data selection and sits on the far side of the boundary of §5; placed inside a learning algorithm, the same computation is model selection. We keep it out of the main text so that a procedure that, as we ran it, lives on the far side of the boundary does not blur it.

We wrote the cleaning rules of §6 by hand, and four of them were wrong. That is an argument for deriving them rather than asserting them, and the derivation is available: whether a item is extensive is not a fact about the analyst's intentions but a property of the item, and the corpus can settle it. We measure that a single threshold on how an item scales with population recovers the hand-written classification for 98.1% of items. We derive two further rules in §B.2 from the defects that exposed them, and each catches more than the case that prompted it—which is the only thing that distinguishes a rule from a patch.

B.1 A cleaning rule is a hypothesis, not a convention

Take the first defect. Deciding whether to divide an item by population is deciding whether the quantity is extensive, and an extensive quantity is one that grows with the size of the observed object. 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 county size more weakly than a general count; two are percentages that happen to rise with county size because the composition they measure is urban. Both failures are visible and nameable, and the test would have surfaced them—which is more than we can say for a regular expression that omitted the word "mean".

B.2 Two more rules, and what they catch

A cleaning defect is an argument for a rule, not an occasion for an edit. Patching the data by hand is the practice this paper exists to argue against: cleaning done as data selection, outside the scored object and justified by the analyst's say-so. So for each of two further defects we ask

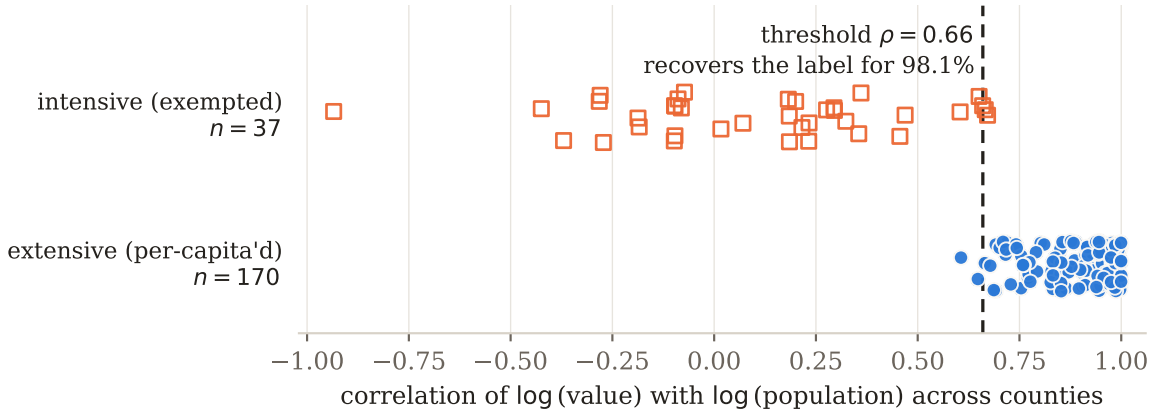

Figure 1: Whether an item is extensive is a property of the item, and the corpus can settle it. Each mark is one of 207 items measured at the 2000 point. Row, marker shape and direct label carry identity as well as hue.

what rule, derivable from the corpus, would have caught it, and then whether that rule catches anything else.

An item that says almost nothing is not a measurement. The item `AGE010`, resident population, is the per-capita denominator divided by itself: 1.0 at 99.76% of counties at 1980, 1990 and 2000. The extensiveness test of §B.1 gets it right—population scales with population—and the correct action is nonetheless exclusion rather than division.

The test has to be what the item *says*, which is its empirical entropy per datum, and our first attempt got this wrong in a way worth recording. Relative interquartile range reads zero whenever one value occupies both quartiles—which takes at least half the counties sharing it, and is certain once three quarters do—a situation that is common and is not degeneracy; it flagged thirteen items, eleven of which carry between 4,700 and 21,600 bits of real content, and acting on it would have deleted them. Entropy separates cleanly where dispersion did not. Two items of 777 fall below 0.1 bits per datum—`AGE010` at 0.028 and `HIS010` at 0.028—while the next lowest sits at 0.556 and the 0.5th percentile across all items is 0.726. Both are population counts divided by population.

Their content is 463 bits between them. The fitted model spends 131,689 bits transmitting them, 284 times what they say.

Within one item-year, values must be on one scale. The item `AGE040` holds three data above 10 and two above 1,000 at the 2000 point, in an item whose median is 1.0: raw counts sitting among ratios. What distinguishes a mixture from a heavy tail is a gap, so the statistic is the extreme over the 99.9th percentile of the same item-year, not over its median. The rule flags **59 items**, all at 2000, and the affected data fall in just **five observed objects**: Broomfield County, Colorado, created in 2001; Yellowstone National Park County, Montana, abolished in 1997; South Boston city, Virginia, which reverted to town status in 1995; and two reorganized Alaska census areas. These are county codes that did not exist, or barely existed, at the point in question, so the per-capita denominator was missing or degenerate and the normalization

silently passed raw counts through. That is the failure mode nominal integration is known to have, arriving by a route the hand-written rules did not anticipate.

The footprint is small—107 data of 9,280,572, about eleven per million—and the size is beside the point. We derived both rules rather than asserting them; both generalize past the case that exposed them, and both would have fired without anyone noticing the defect first.

A finding about the model, arrived at sideways. Charging the thirteen items the discarded dispersion test flagged turned up something that is not about cleaning at all. Their empirical entropy totals 121,718 bits; the fitted model spends 700,164 transmitting them, an overspend of 5.8×.

Two explanations suggest themselves and the measurement refutes both. The decoder is not failing to specialize: on these items it codes at 700,164 bits against the 1,303,516 of the per-item Gaussian we initialize it to, so the fit nearly halves a code we handed it for free. Nor does it charge them the median per-datum cost—they come in at 5.45 bits per datum against a panel-wide median of 6.95. A per-item scale is present already, since the mixture head carries a per-item bias over component weights, mean offsets and log-variance deltas.

What the decoder lacks is an *atom*. These items are a point mass with a few dozen exceptions: `AGE010` is exactly 1.0 at 99.76% of its data. The majority values are of two kinds. Six of the items are population totals—`AGE010`, `POP020`, `POP600`, `INC940` and the ancestry and Hispanic-origin totals—which the per-capita normalization of §6 divides by resident population, so they read 1.0 wherever the two counts agree: a redundancy of the chart, charged by the criterion before anyone read the item. The rest are zeros that most counties genuinely have—subway and ferryboat commuting, thermoelectric water withdrawals, solar heating, murders—for which a point mass is the model of the measure and not a cleaning step. A density with a point mass pays $H(0.9976) = 0.025$ bits per datum to say *the usual value*, which is where the 463-bit content figure comes from. A Gaussian mixture has no atoms. The nearest thing it can write is a component narrowed onto the modal value, and since the residual code adds the datum's own δy^2 to every component's variance, such a component codes a modal datum at about 1.3 bits at best—a density of width δy integrated over δy—fifty times the atom's cost. The fitted decoder does not find even that. Its components keep the item's own spread, and placing the modal value inside δy with a density of that width costs $\log_2(\sigma/\delta y) \approx 13$ bits per datum however well the decoder predicts the mean. The overspend we measure is the ratio of an item's spread to its precision *as fitted*; the residual code forces part of it and the optimizer the rest, and we have not separated the two.

We charge the remedy the same way as everything else here. Across the panel, 46 items have a majority value at some year, covering 237,111 data—2.6% of the corpus—on which the fitted code now spends 1,012,440 bits. A flag plus the exceptions costs 204,287, so a per-item point mass saves 808,152 bits for two numbers per item, about 2.5×10^4 bits of parameters: a return of 33 to 1.

That is a code-length accounting on one autoencoder. Put into the density for real—$P(\text{datum}) = \pi_p \mathbf{1}[\, y \text{ at the mode}\,] + (1 - \pi_p) f(y)\, \delta y$, one modal value per item, its mass set to the mode's share of the item's data and nothing else in the model moved—the point mass shortens the total under the previous code-length accounting of every one of the eight fitted models of §9.1, by 1.15 to 1.52×10^6 bits. The figure is about twice the code-length accounting because the model codes each datum twice, once from its own period's system state and once

from the previous period's, and both codes pay for the missing atom. Thirty-seven items carry one; its locations cost 282 bits to transmit. Fitted with the autoencoder held fixed, the mass lands on the empirical share to four decimals ($\pi = 0.9980$ against 0.9983 for `HIS010`), which is what a correctly specified component looks like rather than a fitted fudge.

Letting the fit respond changes little. We continued each of the eight models from its own checkpoint twice—once as it stood and once with the point mass, with the same optimizer moments, learning rate and seed, so that the pair differs in the atom and nothing else—and scored on its total under the previous code-length accounting averaged over the final 500 epochs. With the atom the continuation ends 1.24 to 1.51×10^6 bits shorter than without it: a mean of 1.44×10^6 with a standard deviation of 1.0×10^5 across the eight, against 1.32×10^6 at the frozen fit. Continuing without the atom gains nothing, which is the evidence that the optimizer does not reach the narrowed component on its own. The decoder does not spend against the saving the capacity it no longer needs for sharpening near-constant items, and the mass stays where the data put it, $\pi = 0.998$ for `HIS010` in every run. These figures are under the previous code-length accounting. Under the sequential code of §3.2 the pullback's share of the saving no longer counts: at the frozen fit of the transmitted model the atom saves 0.87×10^6 bits on the transitions, and the start periods, coded against the null, are unaffected.

At one per cent of the total this is invisible in a bits-per-datum figure. It is also the largest effect this paper measures on a fitted model's learned algorithm length, the only one that appears in every run, and several times the seed-to-seed spread that §9.1 finds swallowing the architectural comparison. The cheapest available improvement to this model is in what its decoder can write, not in its dynamics.

B.3 The cleaning charged in its own currency

Everything above argues from correctness: the hand-written rules were wrong four times, so derive them. The criterion we propose is denominated in bits, and on one piece of the cleaning bits can settle the argument instead.

The normalization mask is cheaper derived than asserted. Which items the transform divides by population is information a decoder needs, and §B.1 measured that a single threshold on how an item scales with population reproduces the hand-written answer for 98.1% of items. We charge it three ways over the 754 items carrying the metadata the hand-written rule reads, of which the rule classifies 678 as extensive:

the mask transmitted as	bits
data, one bit per item	754
data, entropy-coded at the observed rate	356
the rule of §B.1 plus its 14 exceptions	**133**

The derived form is 223 bits shorter, 63% of the asserted one. Against a learned algorithm length of 7.04×10^7 that is nothing, and the sign is the entire point: a submission carrying the rule beats one carrying the mask, so §B's thesis is not a methodological preference but a scoring fact. It also scales the right way—the mask costs $O(\text{items})$ and the rule $O(\text{exceptions})$, so the margin widens with the corpus rather than shrinking.

The denominator is a debt, and we state its size. The same honesty cuts against us elsewhere. The chart transform runs before we count any bit, so three things sit outside the reported learned algorithm length: the mask above, the scale s_{95} (one number), and the per-capita denominator—one population per county-year, which a decoder cannot invert the transform without. Charged as this paper charges everything else, from the previous decade at the instrument's own precision, that denominator is 14.26 bits per value over the 12,732 decennial county-years, or 1.8×10^5 bits. Imputing it first saves almost nothing: persistence predicts log population to about 13% while a complete count is exact to one person, three orders of magnitude apart, so nearly every bit lies in digits no model can reach. The leverage is in the precision required rather than in the prediction—a denominator need only be as sharp as the items it divides, and at the 10^{-3} relative precision the chart's data actually carry, the cost falls to 1.2×10^5 bits.

Either figure is the size of the seed-to-seed spread of a fitted model's total reported in §9.1. It is 0.2% of the learned algorithm length and changes no conclusion about this panel. What it does mean is that the demonstration codes a *derived* corpus rather than the sealed one, and that two submissions choosing different transforms would not yet be comparable—which is the thing §4 exists to prevent. The mechanism is sound and the worked example is one transform short of implementing it. We report the number rather than the intention.

B.4 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 item, 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;
- an income series carries its reference year in its own description—"median household income in 1979"—so we need not have guessed the mapping onto a survey year at all;
- whether a quantity may be summed across geographies follows from its universe and unit indicator: a count of households is additive, a median of them is not.

There is a reason, in the criterion's own terms, 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 items among 698 costs about 73 bits, less than the character count of a rule that finds them. The learned algorithm length over the fixed corpus alone would therefore prefer the list. The rule wins on two other grounds. It generalizes to items 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 the learned algorithm must transmit, 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, charged in the same bits as everything else, and testable in the way §6 could not test a hand-written regular expression—which is how the same class of error escaped twice.

Data availability

Every source is public, and we used no restricted data. *USA Counties* (U.S. Census Bureau) supplies the 1970–2010 backbone; we retrieved the American Community Survey five-year detailed tables for 2006–2010, 2011–2015, 2016–2020 and 2020–2024 from `data.census.gov`, and the county-subdivision tables for Connecticut 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. The release has the form §4 proposes for an agency's own: the public aggregated data, a learning algorithm and the learned algorithm it produced.

Note on assisted authorship

The author developed the analysis, code and prose in an extended collaboration with a large language model acting as a research assistant under the author's direction. Code in the repository produced every numerical result in this paper and can regenerate it; that process introduced, and found, the measurement defects §6 reports, which is part of why we report them.

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.

Chetty, R., and Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. *Journal of Privacy and Confidentiality*, 9(2). doi:10.29012/jpc.716.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., and Porter, S. R. (2018). The Opportunity Atlas: Mapping the childhood roots of social mobility. NBER Working Paper 25147.

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.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2), 277–304.

Frisch, R. (1938). *Statistical versus Theoretical Relations in Economic Macrodynamics*. Memorandum, League of Nations.

115th Congress (2019). *Foundations for Evidence-Based Policymaking Act of 2018*. Public Law 115–435, 132 Stat. 5529, January 14, 2019. Enrolled text from the Government Publishing Office, archived as `docs/PLAW-115publ435.htm`.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, 12 (Supplement), iii–115.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10), 845–848.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, Berlin.

Ingber, L. (1989). Very fast simulated re-annealing. *Mathematical and Computer Modelling* 12(8), 967–973.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719), 494–498.

Lan, N., Chemla, E., and Katzir, R. (2024). Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length. arXiv:2402.10013v2, June 6, 2024, archived as `docs/2402.10013v2.pdf`.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46.

Office of Management and Budget (2020). *Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices*. Memorandum M-20-12, March 10, 2020. Archived as `docs/M-20-12.pdf`.

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.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5), 547–554.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356.

Solomonoff, R. J. (1960). *A Preliminary Report on a General Theory of Inductive Inference*. Report ZTB-138, Zator Company, Cambridge, Mass. Revised 1962. `https://raysolomonoff.com/publications/z138.pdf`.

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.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24(4), 422–432.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4), 381–386.