MedCrave
Step into the World of Research

Sociology International Journal

Review Article

Open Access

CrossMark

The empirical gap: matching social science methods to contemporary data environments

Abstract

This article asks whether the empirical methods routinely used in the social sciences remain commensurate with the analytical possibilities opened by contemporary data environments. The central issue is methodological rather than technological. Established qualitative and quantitative approaches retain substantial explanatory value, yet the technical frontier has moved rapidly toward forms of evidence that are larger in scale and more difficult to represent through conventional tabular structures. Part of that pressure comes from digitally mediated socioeconomic interaction, which now generates traces that purpose-built research instruments observe only indirectly. A second source is independent of digital socialization itself: advances in statistical learning and computational representation have relaxed technical restrictions that previously forced researchers to simplify complex evidence before analysis. The article proposes a selective empirical update in which newer techniques enter the toolkit only when they solve an identified problem of representation or estimation. Three matrices organize the argument by linking changes in empirical access to methodological adequacy, then comparing established methodological families with their recent extensions, and finally mapping technical restrictions to concrete analytical responses. The broader claim is that a narrow methodological repertoire can influence which dimensions of social and economic processes become measurable enough to enter explanation.

Keywords: social science methodology, empirical methods, methodological updating, computational social science, statistical learning, causal inference, data environment, digital methods

Volume 10 Issue 4 - 2026

Rodrigo Kataishi
National University of Tierra del Fuego Ushuaia, Argentina

Correspondence: Rodrigo Kataishi, Ph.D., Institute of Economic Development and Innovation, National University of Tierra del Fuego, Argentina, Tel +5492901652878

Received: August 12, 2026 | **Published:** September 3, 2026

JEL Classification: C10, C50, C80, B40, O30

Introduction: why empirical updating has become a methodological question

As digitally mediated interaction becomes more common, the empirical environment faced by the social sciences changes in ways that cannot be reduced to the simple growth of available data. A growing share of interaction now unfolds through infrastructures that mediate communication while also recording activity. The relevant shift is not the disappearance of physical interaction, but the increasing coexistence of settings in which social processes extend across physical and digital domains. This widens the empirical field and weakens the assumption that the locus of interaction can always be reconstructed through instruments designed around physically delimited settings.[1,2]

Within that broader field, economic activity changes alongside social interaction rather than appearing as a separate dimension. Platform-mediated work offers a clear illustration because the infrastructure through which people coordinate activity may also organize access to income and shape the conditions under which participation is valued. Data generated through ordinary use can likewise become an input into accumulation. The economic dimension therefore forms part of the social process itself while also helping to configure the terms under which interaction develops.[3,4]

Seen from the standpoint of empirical research, these shifts matter because the location of the process and the location of observation no longer coincide as closely as before. A labor relation may remain physically situated while its monitoring occurs digitally. A territorially bounded community may sustain much of its ongoing interaction online. In such cases, studying actors outside the infrastructure can reveal important attributes and experiences while leaving part of the interactional structure unobserved.[5]

Against this background, the ordinary empirical repertoire of the social sciences retains its value but encounters new conditions of sufficiency. Surveys remain powerful for standardized measurement, while interviews continue to provide access to meaning and experience. Yet part of the evidence relevant to contemporary processes is now generated outside those research instruments. The mismatch begins when a method captures the participant but not the relational process in which that participant is embedded.[6,7]

Alongside this first mismatch, a second one has emerged from the evolution of analytical capacity itself. Statistical learning now offers procedures for settings in which manual specification becomes unstable because the covariate space is too large. Graph-based methods preserve forms of interdependence that conventional tabular representations often compress. Related advances have changed the analysis of documentary sources and fine-grained territorial evidence. These developments expand what can be analyzed, but they do not yet form a homogeneous part of the methodological repertoire used across the social sciences.[8]

As access to digital sources expands, the status of empirical evidence changes with it. Researchers can now work with large public repositories that existed before the research question was formulated, while institutional digitization makes extensive administrative and documentary records searchable at a scale that was previously impractical. Earlier work on scientific careers already showed how public web information and curriculum vitae could be crawled and transformed into a relational database for social-science analysis.[9]

Digital platforms add another layer because interaction itself can generate logs whose evidentiary value depends on understanding how the underlying system produced them.[1,10]

From this perspective, methodological updating is best treated as a problem of correspondence rather than replacement. The relevant question is whether the empirical structure of a research problem can be represented adequately with the methods that researchers routinely consider available. When that correspondence weakens, newer techniques may extend the range of observable dimensions without displacing the interpretive or inferential functions of established approaches. The discussion that follows therefore uses changes in socioeconomic interaction as context before turning to the paper's main concern: the expansion of technical capacity and its uneven incorporation into ordinary research practice.[11,1]

From changing socioeconomic interaction to the problem of empirical access

For much of the twentieth century, many central objects of sociological and economic inquiry could be studied through a relatively close relationship between the place where interaction occurred and the place where observation was organized. Workplaces and formal organizations offered settings that could be entered directly, while households could be reconstructed through purpose-built surveys. This proximity never implied complete access, but it provided a comparatively stable empirical architecture for much of the methodological repertoire that became standard.[12]

Within that architecture, qualitative and quantitative approaches developed complementary strengths. Fieldwork could reconstruct context because the relevant setting was accessible to the researcher, while survey research could estimate distributions because standardized observations were produced through a controlled instrument. The issue raised here is not that these methods cease to work once interaction becomes digitally mediated. Their adequacy depends more strongly on whether the chosen design still corresponds to the actual structure of the process.[7]

As social practices become distributed across physical and digital settings, the empirical field becomes less easily bounded. Interaction may begin in one setting and continue in another, while coordination can remain attached to a physical organization even when important decisions are mediated through software. The same hybrid structure increasingly appears in economic activity when labor remains materially situated but assignment or evaluation is handled through digital systems. Methodologically, observing only one environment can therefore produce a partial reconstruction of the process.[5,1]

The widening empirical field also changes what counts as a viable source. Administrative databases become richer as public institutions digitize their operations, while large documentary banks can be searched without first reducing them to small hand-selected samples. Digital traces provide a different kind of source when the interaction itself leaves a record. Each of these sources broadens empirical access, yet each also requires reconstruction of the conditions under which the record was produced.[12,1]

When a process depends on digital infrastructure for the form it takes, the methodological implications become stronger. Online communities offer an obvious case because the relational structure exists within the platform itself. Platform-mediated labor creates a similar problem when matching or evaluation is organized through digital systems. In such cases, the infrastructure cannot be treated as an external communication layer because it partly constitutes the process under study.[2,5]

This distinction becomes clearer when actor-centered and interaction-centered evidence are compared. A survey can describe the people who participate in a platform, and an interview can reconstruct how they experience that participation. Neither source alone can recover the topology of a large network nor the circulation pattern produced within it. The gap therefore concerns the structure of observation: methods may remain valid for one dimension of the phenomenon while leaving another dimension outside the empirical design.[6,10]

Table 1 summarizes this relationship through three ideal-typical configurations. They should not be read as a rigid chronology because each can coexist with the others, and a single research object may combine more than one. The purpose of the matrix is to show how methodological adequacy changes when social and economic interaction becomes more dependent on digitally mediated environments.

Table 1 Locus of interaction and methodological adequacy

Predominant locus	Qualitative methods	Quantitative methods
Predominantly physical interaction	High capacity to reconstruct meaning and situated practice because the context remains broadly accessible to direct observation or participant reconstruction.	High capacity to measure attributes and regularities through surveys, censuses and purpose-built administrative records.
Hybrid physical-digital interaction	Interpretive capacity remains strong, but field reconstruction must follow practices across more than one environment.	The empirical base expands through richer records, while part of the interaction begins to occur outside purpose-built research instruments.
Predominantly digital interaction	Meaning and experience remain central, but digital infrastructure becomes part of the context that must itself be interpreted.	Population-level regularities remain essential, yet a substantive part of the phenomenon may reside in interaction structure rather than in actor attributes alone.

Read in this way, the matrix suggests a cumulative rather than substitutive logic. The more constitutive the digital environment becomes for a given process, the greater the need for methods capable of accessing evidence generated within that environment. Yet the methodological case for updating the toolkit does not depend only on digital interaction. Even when the substantive object remains conventional, the contemporary data environment introduces technical constraints that older analytical workflows were not designed to absorb.[11,8]

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

The changing data environment

In classical research designs, data are often imagined as the outcome of a planned observation. The researcher defines a population and constructs an instrument in order to obtain comparable evidence. Digital records partly reverse this sequence because relevant evidence may already exist before the study begins. Administrative activity can generate data as a by-product of institutional functioning, while documentary repositories may accumulate material over long periods without reference to a later research question. The methodological task therefore moves partly from constructing the source toward reconstructing its provenance.[10,12]

Once evidence is generated in this way, control over measurement cannot be inferred from the size or precision of the database. A platform log may record an event accurately while hiding the rule that classified it, and a public database may offer extensive coverage while containing breaks created by administrative reform. These problems do not make the sources unusable. They require a more explicit account of how the record came to exist and how closely that process corresponds to the concept being measured.[1]

As data sources grow, computational constraints increasingly become part of empirical design. Large administrative files can exceed the memory limits of ordinary desktop workflows, while digitized documentary collections may contain volumes that make manual inspection impossible. At that point, storage architecture and parallel processing affect which operations can be performed without reducing the source in advance. Big data is therefore not only a statistical condition but also an infrastructural one.[1,13]

A related problem appears when evidence is generated continuously rather than as a closed dataset. Some digital records are better understood as streams whose state changes while observation is taking place. In these settings, a workflow based on periodic snapshots may discard temporal structure before analysis begins. Streaming architectures and online-learning procedures respond to this feature by allowing the analytical process to update as new observations arrive.[14]

Higher dimensionality introduces a different kind of restriction. When a dataset contains a large number of potential predictors, conventional estimation becomes increasingly sensitive to specification. Regularization addresses this problem by constraining model complexity rather than forcing the researcher to reduce the covariate space entirely through prior selection. This expands the range of empirical structures that can be analyzed without abandoning statistical discipline.[15,16]

Nonlinear relationships create another source of difficulty because they increase dependence on the functional form chosen by the researcher. Flexible statistical learning offers a way to relax that dependence while retaining explicit validation against data that were not used to fit the model. The analytical gain lies in representing relationships that would otherwise require extensive prior specification, not in a generic superiority of newer algorithms over conventional statistics.[17,18]

Some contemporary sources also challenge the unit-by-variable representation itself. Networks contain ties whose analytical meaning depends on structural position, while text contains contextual information that cannot be reduced fully to word counts. Trajectories create a related problem because ordered states may carry information that disappears when they are collapsed into period averages. In each case, the question concerns how much substantive structure is lost before estimation even begins.[7,19]

For this reason, methodological expansion also concerns representation. Graphs preserve relations instead of collapsing them into actor attributes, and embeddings allow similarity to be modeled in a continuous space. Work on semantic-statistical analysis in Latin American research has explicitly framed this transition as a methodological problem rather than as a purely technical innovation.[6,7] The same logic provides the bridge to the paper's central argument: contemporary analytical capacity has expanded because different methodological developments now relax restrictions that previously forced stronger simplification.

Expanding methodological capacity

Flexible statistical learning and causal inference

Parsimonious modeling remains a major strength of conventional statistics because it supports interpretation and keeps the relation between theory and estimation visible. The limitation appears when empirical structure contains more potentially relevant dimensions than can reasonably be handled through manual specification. In those settings, parsimony can shift from analytical virtue to a source of information loss if relevant structure is removed before the model is estimated. The tension between interpretability and flexibility has a long history in statistical learning, where the central challenge has always been to expand modeling capacity without losing the ability to understand what the model represents.[16,20]

Regularization provides one response to that problem. Penalized estimators control the effective complexity of a model, making broader covariate spaces empirically tractable without treating every available dimension as an unrestricted parameter. The methodological contribution is therefore tied to a specific restriction: high dimensionality can be incorporated without relying entirely on manual variable selection. The theoretical foundations for this approach were developed well before the current wave of interest in machine learning, as statistical learning theory had already formalized the trade-off between model complexity and generalization.[21,16]

Flexible function learning addresses a different restriction by reducing dependence on a single prespecified functional form. Tree-based methods can recover thresholds or interactions that would otherwise need to be anticipated in advance, while smoother approaches provide less rigid departures from linearity. Rather than collapsing these procedures into the label of machine learning, their value is better understood through the particular modeling constraint each one relaxes. Ensemble methods such as random forests and boosting illustrate this principle directly because they aggregate simple learners into representations that capture structure which a single parametric form would miss.[17,22] This perspective, in which the data themselves guide the choice of functional form, is a central theme in the data mining tradition and has been systematized in widely used methodological references.[23,24]

As flexibility increases, validation becomes more central to design. Out-of-sample assessment helps determine whether a recovered pattern generalizes beyond the observations used to construct it, while cross-validation provides a disciplined basis for comparing specifications. Theoretical discipline does not disappear in this setting. Part of it shifts from specifying every relationship ex ante toward justifying the learning procedure and interpreting the patterns it recovers. This concern with generalization is itself a classical topic

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

in both statistical learning[16] and data science methodology.[25,26] The emergence of data science as a distinct analytical field has further emphasized that technical skills in modeling must be combined with the capacity to frame problems in ways that make them empirically tractable, a point that applies with equal force to the social sciences.[27]

Within econometrics, the strongest dialogue with statistical learning begins from a limit rather than a substitution. Predictive accuracy does not identify a causal effect because the counterfactual problem remains unresolved.[28,29] The distinction between prediction and causal inference is foundational to modern econometrics and remains the organizing principle of its engagement with machine learning methods.[30] Recent causal machine-learning approaches are most useful precisely when they retain that distinction and use flexible estimation only where it supports a clearly defined inferential target.[8,31]

High-dimensional control selection offers one such case. When researchers face large covariate sets, manual specification can become unstable and sensitive to discretionary choices. Double-selection strategies use regularization to organize this part of the estimation problem while preserving an explicit causal parameter. This allows greater flexibility without treating prediction as an end in itself.

Double or debased machine learning develops the same principle through orthogonalization and cross-fitting. Flexible algorithms estimate nuisance functions, while the parameter of interest is constructed to be less sensitive to small errors in those estimates. The methodological gain lies in combining flexible prediction with a design that still separates the causal question from the auxiliary modeling task.[31,32]

Heterogeneous treatment effects extend the analysis further by allowing researchers to ask how an intervention varies across empirical contexts. Generalized random forests make it possible to estimate conditional parameters without restricting heterogeneity to a small set of predefined subgroups.[33] For sociology and development economics, this creates a more direct way to study variation that may depend on context rather than assuming that an average effect adequately summarizes the process.

Beyond unit-by-variable representations

Large documentary collections create another methodological frontier because exhaustive manual reading becomes impossible long before the source loses substantive relevance. Classical text-as-data methods responded by transforming documents into statistical representations that could be compared systematically. This already changed the scale at which documentary evidence could enter social research, while historical-methodological reviews show that the transition toward contemporary NLP has involved successive changes in how meaning itself is represented.[6]

Topic models expanded that repertoire by identifying latent co-occurrence structures, while supervised classification allowed coding schemes developed on smaller samples to be extended to much larger document collections. These procedures provide an intermediate methodological layer between close reading and contextual language models. Their continuing relevance is especially clear when topic structure carries substantive information that should not be discarded in favor of a purely contextual representation.[7]

Embeddings introduce a different representation by placing textual units in a continuous space where proximity reflects patterns of use. Sentence-level representations make it possible to compare documents without reducing each one to the same fixed vocabulary structure.[34,35] Recent work in social-science retrieval has explored hybrid representations that align latent topic structure with contextual semantics rather than treating them as competing alternatives.[11]

Contextual language models deepen that capacity because the representation of a word canvary with the sentence in which it appears. BERT-based transfer learning therefore provides a technical response to context sensitivity in supervised text analysis.[36,37] Topic-enriched embedding approaches extend the same methodological concern by integrating classical statistical signals with contextual representations, first in preprint form and later in a peer-reviewed journal version.[38,39] Large language models expand annotation possibilities further, although validation against expert coding remains necessary.[40]

Network analysis addresses a different representation problem because observations may be connected in ways that are substantively constitutive of the process. Centrality measures describe position within a relational structure, while community detection can identify groupings that emerge from patterns of connection. These techniques make interdependence part of the empirical object rather than treating it only as a statistical nuisance.

Graph representation learning extends this logic by encoding structural position numerically. Node2vec, for example, learns low-dimensional node representations from network neighborhoods and can then support downstream prediction.[41] More recent graph-learning architectures push the same principle further by combining relational structure with attributes, illustrating how methodological progress can occur through a change in representation rather than through a more elaborate regression on the same table.

Economic geography has made productive use of this relational turn because physical distance does not exhaust the meaning of proximity. Product spaces and relatedness measures infer productive or technological closeness from observed specialization patterns, allowing territorial capabilities to be represented relationally rather than as isolated attributes.[42,43] This is methodologically significant because the data structure itself becomes part of the theoretical argument about development trajectories.

High-resolution spatial evidence adds another extension. Remote sensing allows researchers to recover information below conventional administrative scales, while computational image representations make large visual sources usable for socioeconomic measurement.[44,45] These methods widen the empirical scale at which territorial processes can be studied, even though substantive interpretation still depends on whether the extracted feature is a valid representation of the social concept of interest.

From technical expansion to disciplinary incorporation

Tables 2 and 3 bring the technical argument together at two complementary levels. The first compares methodological families according to the limits that newer techniques extend. The second starts from the empirical restriction itself and identifies the response that makes the source analytically tractable. Reading them together avoids ranking methods by novelty and instead asks what specific limitation is being relaxed in each case.

Table 2 is organized around five analytical dimensions, each revealing a different aspect of methodological change. The first column names seven core methodological families, from qualitative methods through massive-data processing. The second column identifies what each family does best, which makes explicit the analytical function that should be preserved even when techniques are extended. The third column then isolates the characteristic limit that appears once a research question pushes beyond what the family's conventional

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

toolkit can represent comfortably. It is precisely this limit, and not a generic preference for computational tools, that motivates each extension listed in the fourth column. The final column specifies the additional analytical capacity that the extension makes available, so that the methodological gain remains tied to a concrete improvement in the researcher's ability to represent, estimate, or interpret evidence.

Table 2 Methodological families and extensions of analytical scope

Family analytical capacity	Core strength	Characteristic limit	Techniques extending the limit	Additional capacity	Key references
Qualitative methods	Meaning and contextual reconstruction	Limited scale for very large corpora or extensive relational structures	Assisted coding; topic modeling; semantic embeddings; digital ethnography supported by trace data	Extends interpretation to larger bodies of evidence and supports systematic case selection	Kataishi and Milia,[7] Hine,[46]
Conventional quantitative statistics	Population description and explicit inference	High dimensionality and rigid functional assumptions	Ridge; LASSO; elastic net; generalized additive models; decision trees; random forests; gradient boosting	Adds regularization and flexible functional modeling where manual specification becomes restrictive	Tibshirani,[15] Hastie et al.,[16] Zou and Hastie,[47]
Econometrics	Causal identification and interpretable parameters	High-dimensional controls and complex effect heterogeneity	Double/debiased machine learning; generalized random forests; causal forests; meta-learners	Permits flexible nuisance estimation and more granular analysis of heterogeneous causal effects	Chernozhukov et al.,[30] Belloni et al.,[48] Athey et al.,[33]
Classical quantitative text analysis	Reproducible coding of vocabulary or predefined categories	Loss of contextual meaning when text is reduced to frequencies	Topic models; word embeddings; sentence embeddings; transformers	Preserves more semantic context while extending analysis to larger corpora	Blei et al.,[49] Devlin et al.,[36] Mikolov et al.,[50]
Spatial analysis	Spatial dependence and territorial spillovers	Coarse administrative resolution or rigid proximity structures	Remote sensing; spatially flexible learning; image-based feature extraction; spatiotemporal models	Adds finer-grained observation and allows proximity to be represented beyond simple contiguity	Anselin,[51] Jean et al.,[52] Burke[45] et al.,
Network analysis	Interdependence and structural position	Limited treatment of dynamic or attribute-rich networks	Node embeddings; graph representation learning; temporal and multilayer networks	Integrates topology with attributes and permits structural information to enter predictive tasks	Grover and Leskovec,[41] Wasserman and Faust,[19] Barabási and Albert,[53]
Massive-data processing	Empirical access at large scale	Storage and computation become binding constraints	Columnar storage; distributed databases; cluster computing; parallel processing	Keeps large sources analytically usable without premature aggregation	Dean and Ghemawat,[13] Zaharia et al.,[54] Lazer et al.,[55]

Read transversally, the table reveals two structuring patterns. First, every family faces a limit, but the limits are substantively different across families: dimensionality restricts conventional statistics, context loss restricts text analysis, and fixed adjacency structures restrict spatial methods. Because the limits differ, the solutions differ, and no single technique addresses all of them. Regularization is not a response to semantic complexity, just as embeddings are not a solution for causal identification. Second, the direction of extension is cumulative rather than substitutive. The family retains its core analytical function while the extension adds capacity along the margin where the conventional toolkit loses traction. Interpreting the table in this way supports the paper's central claim: methodological updating should be selective and tied to identifiable analytical gaps, not driven by the assumption that newer techniques render established families obsolete.

Seen as a whole, the matrix also clarifies what is not at stake in the proposed update. It does not claim that every social scientist must master every technique, nor that computational methods are inherently superior across all research contexts. What it does suggest is that the boundary between a conventional toolkit and a computationally extended one is not a single frontier but a series of distinct analytical thresholds. Recognizing those thresholds as methodologically relevant rather than as exotic specializations is a disciplinary question as much as a technical one.

Table 3 adopts a diagnostic rather than a taxonomic structure. Each row isolates a specific empirical restriction that conventional analytical workflows were not designed to absorb easily. The first column names the restriction; the second describes why it becomes binding under standard research designs; the third identifies the methodological response that makes the source analytically tractable; and the fourth provides concrete techniques that implement that response. By organizing the argument around restrictions rather than around disciplinary traditions, the matrix avoids classifying techniques by their origin and instead groups them by the problem they solve.

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

Table 3 Technical restrictions and methods that expand the analytical frontier

Dimension	Restriction in conventional workflows	Methodological response	Representative techniques	Key references
Volume	Memory and computation time become binding before the substantive source has been fully processed.	Distribute storage or computation so the source can remain analytically available at scale.	MapReduce; Apache Spark; Dask; GPU processing; Parquet/Arrow	Shvachko et al.,[56] Dean and Ghemawat,[13] Zaharia et al.,[4]
Velocity	A closed-dataset workflow loses information when observations are generated continuously.	Process events incrementally and update analytical procedures as new observations arrive.	Kafka; Flink; Spark Streaming; online learning	Kreps et al.,[57] Bottou,[14]
High dimensionality	A large predictor space increases instability and overfitting under unrestricted estimation.	Control effective model complexity through regularization or sparse estimation.	Ridge; LASSO; elastic net; sparse models	Tibshirani,[15] Bühlmann and van de Geer,[58] Hastie et al.,[16]
Nonlinearity	Results depend strongly on a functional form that must be specified in advance.	Allow flexible approximation while evaluating generalization outside the fitted sample.	GAM; CART; random forests; gradient boosting; neural networks	Breiman,[17] Friedman, 2001;[18] Breiman et al.,[59]
Heterogeneous effects	Average effects can conceal systematic variation across contexts.	Estimate conditional effects without restricting heterogeneity to a few predefined groups.	Causal forests; generalized random forests; Bayesian causal forests; meta-learners	Athey and Imbens,[60] Wager and Athey,[61] Athey et al.,[33]
High-dimensional causal adjustment	Manual control selection becomes unstable when potential confounders are numerous.	Use flexible nuisance estimation while protecting the inferential target through orthogonalization.	Double/debiased ML; double selection; cross-fitting	Belloni et al.,[47,62] Chernozhukov et al.,[30]
Interdependence	Unit-by-variable data treat relations indirectly even when ties are constitutive of the process.	Represent the empirical object as a graph and model structural position directly.	Social network analysis; ERGM; stochastic block models; graph embeddings	Wasserman and Faust,[19] Robins et al.,[63]
Community structure	Institutional categories may not coincide with groupings generated by observed interaction.	Detect meso-level structure from patterns of connection.	Louvain; Leiden; Infomap; stochastic block models	Blondel et al.,[64] Traag et al.,[65] Newman,[66]
Large-scale text	Exhaustive reading becomes impossible while sampling may discard relevant documentary structure.	Transform large corpora into comparable statistical representations.	TF-IDF; topic models; supervised text classification	Salton and Buckley,[67] Blei,[68] Grimmer and Stewart,[69]
Contextual semantics	Frequency-based representations lose meaning that depends on linguistic context.	Represent textual units in learned semantic spaces that retain contextual information.	Word2Vec; BERT; sentence transformers; domain-specific embeddings	Devlin et al.,[35] Reimers and Gurevych,[33] Mikolov et al.,[49]
High-resolution space	Administrative units can conceal local heterogeneity that is visible in remote-sensing data.	Extract spatial features below conventional statistical scales.	Satellite imagery; CNN/ViT feature extraction; raster analysis	Jean et al.,[51] Burke et al.,[44]
Relational space	Physical distance can miss productive or technological proximity.	Construct similarity spaces from observed capabilities or connections.	Product space; relatedness; multilayer networks	Hidalgo and Hausmann,[70] Hidalgo et al.,[71] Boschma[72]
Complex temporality	Panels can compress trajectories when transitions or state sequences are substantively relevant.	Model ordered states or changing network structures directly.	Hidden Markov models; state-space models; sequence analysis; temporal networks	Abbott,[73] Rabiner,[74] Allison,[75]
Images	Manual visual coding cannot scale to large image repositories.	Transform visual structure into reproducible features for subsequent analysis.	CNN; vision transformers; object detection; image embeddings	Krizhevsky et al.,[76] Dosovitskiy et al.,[77]
Heterogeneous sources	Analyzing each modality separately can remove cross-source structure before inference.	Construct joint representations that retain information across source types.	Multimodal transformers; joint embeddings; graph-based multimodal models	The references for this final row were cut off in the text you provided.

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134. DOI: 10.15406/sij.2026.10.00469

The transversal logic reveals that the restrictions are largely orthogonal to one another. Volume and velocity are infrastructure problems that appear when the data source is too large or too fast for a single-machine workflow. High dimensionality and nonlinearity are modeling problems that arise even in modestly sized datasets if the predictor space is rich. Interdependence, community structure, and contextual semantics are representational problems: the empirical object contains structure that the unit-by-variable format cannot capture without loss. Because these are different types of difficulty, the appropriate response depends on which restriction is active in a given research context. A dataset may be large but low-dimensional, in which case distributed computing suffices; another may fit in memory yet be so rich in predictors that regularization becomes essential; a third may be small but so deeply relational that graph methods are the only adequate starting point.

This diagnostic logic provides the operational core of the paper's selective updating proposal. Rather than asking whether a method is new, the matrix encourages researchers to ask whether a restriction prevents the phenomenon from being represented adequately at the current stage of analysis. When the answer is yes, the corresponding column identifies a concrete analytical response that relaxes the constraint without changing the inferential claim by default. When the answer is no, adding complexity offers no methodological gain. The matrix thus translates the general principle of problem-driven updating into a practical heuristic that can be applied across substantive domains.

A transversal reading of Table 2 shows why methodological change should not be described as a single computational revolution. Regularization extends regression under one type of constraint, while contextual embeddings address a different problem in textual representation. Distributed infrastructure solves yet another restriction at the level of computation. Keeping these distinctions visible prevents the paper from using machine learning as a residual category for every recent development.

Complementarity follows from that distinction. Qualitative methods remain indispensable when the question concerns meaning, and econometric identification remains necessary when the claim is causal.[29] Graph methods become relevant when relational structure is part of the object itself. The methodological toolkit therefore needs breadth not because every project should use every technique, but because method choice should follow the empirical architecture of the problem rather than the researcher's habitual repertoire.

Table 3 sharpens the same argument by moving from disciplinary traditions to technical diagnosis. A restriction becomes methodologically relevant only when it can be connected to the structure of the evidence and to a procedure that relaxes it without changing the inferential claim by default. This is the core of the paper's proposal for empirical updating: newer techniques enter the toolkit as responses to identifiable analytical problems rather than as a general technological layer.

The matrix also clarifies why different technical problems require different responses. A dataset may be extremely large but low dimensional, in which case distributed processing may be enough. Another may fit easily in memory yet contain so many predictors that regularization becomes the relevant issue. Text may present a semantic problem even when the corpus is small. Methodological precision therefore depends on diagnosing the restriction before selecting the tool.

Once that diagnosis is made, newer techniques can be understood as extensions of empirical observability. They allow researchers to retain structure that would otherwise need to be aggregated or discarded. This does not make the resulting pattern self-explanatory, but it changes which dimensions can be examined directly rather than inferred only through indirect proxies.

Despite these developments, methodological training in the social sciences often remains centered on a narrower repertoire. Advanced regularization may appear only after conventional regression, while graph representation learning is frequently treated as a specialized topic rather than as part of a broader relational toolbox. Contextual NLP can likewise remain separated from ordinary text analysis instead of being taught as the next step in a progression of representational choices. The result is a gap between the analytical frontier and the toolkit with which many projects are initially conceived.

That gap matters because methodological availability can shape object construction. If relational structure is not considered empirically accessible, researchers may formulate questions around actor attributes instead. If large documentary corpora appear technically unmanageable, they may be reduced to convenience samples before their internal structure can be explored. Methodological limitations can therefore influence what is treated as measurable and, by extension, what becomes visible enough to enter explanation.

Toward an updated empirical toolkit

Rather than opposing traditional and computational methods, a more productive update begins by distinguishing their functions. An interview can reconstruct meaning in a way that an embedding cannot, while a causal design can identify an effect that a purely predictive model cannot. A graph representation can preserve interdependence that a conventional table may suppress. These differences support expansion of the toolkit rather than replacement of one methodological family by another. This complementary logic has been a defining theme in computational social science, where the integration of digital-age data sources with established inferential frameworks is understood as a methodological enrichment rather than as a competition between paradigms.[78]

Method selection should therefore begin from the empirical restriction rather than from the novelty of the technique. A complex model adds little when a simpler specification already answers the research question. By contrast, an advanced method becomes justified when it relaxes a concrete limitation that prevents the object from being represented adequately. This problem-driven logic is the central criterion of the proposed update and echoes the broader turn in social science methodology toward tools selected by the structure of the evidence rather than by disciplinary habit.[79]

A first decision concerns the correspondence between object and evidence. When the process is fundamentally relational, data organized only around isolated attributes may be insufficient. When meaning depends on context, frequency-based representations can remove relevant information before analysis. Conceptualization must therefore precede representation so that the chosen data structure reflects the phenomenon rather than merely the convenience of the available method.

The next decision concerns the correspondence between evidence and technique. Scale problems call for infrastructure capable of handling large sources, whereas dimensionality problems require some form of complexity control. Relational problems require graph representations, while semantic problems require representations that preserve context. Distinguishing these cases prevents technically unrelated issues from being collapsed into a generic appeal to machine learning.

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J.* 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

A final decision concerns the relation between technique and inference. Predictive performance does not justify causal claims without identification, and semantic clustering does not produce sociological categories on its own. A detected network community likewise does not automatically imply shared identity. Methodological updating must therefore expand empirical capacity while preserving explicit limits on what each procedure allows researchers to infer.

Within statistics and econometrics, an updated core should include regularization and flexible modeling as ordinary extensions of conventional estimation rather than as exotic additions.[30] Causal machine learning belongs in the same progression because it shows how predictive tools can support inference without replacing identification. The aim is not to make every social scientist a specialist, but to ensure that high dimensionality or heterogeneous effects are recognized as methodological problems for which mature solutions already exist.

A second core area concerns relational and semantic representation. Graph methods become essential when the object depends on interdependence, while contemporary NLP becomes relevant when the source is textual and context cannot be reduced safely to frequency. The methodological trajectory documented in social-science semantic analysis shows how older statistical techniques can remain useful inside hybrid architectures rather than being discarded by newer contextual models.[7,6,39]

A third core area concerns computational infrastructure. Large-scale empirical work may depend on storage formats that permit efficient access or on distributed computation that avoids premature aggregation. These tools are not methods of social explanation, yet they condition which methods remain feasible once the empirical source reaches big-data scale. Treating infrastructure as part of methodological literacy therefore becomes increasingly important, especially when public records or digital repositories are large enough that technical access determines the effective sample. Recent interdisciplinary guides have begun to bridge this gap by presenting computational infrastructure alongside research design, making the case that data engineering and methodological reasoning are no longer separable stages of the research process.[80]

Discussion and conclusions

Taken together, the preceding sections identify a first source of pressure on the empirical toolkit in the changing relation between socioeconomic interaction and observational design. As part of social and economic life becomes digitally mediated, evidence about the process may exist inside infrastructures that conventional research instruments observe only indirectly. This creates a stronger need to complement actor-centered evidence with methods capable of recovering interactional structure.

A second source of pressure appears independently from the transformation of socialization itself. Contemporary data environments can exceed conventional assumptions about scale or dimensionality, while newer methods provide increasingly specific responses to those problems. When those responses remain outside the ordinary toolkit, the limitation is not merely computational because potentially relevant dimensions may never enter the empirical design.

For this reason, the broader methodological issue concerns the relationship between technical capacity and substantive visibility. Theory continues to define what matters, but the available toolkit shapes how those dimensions can be operationalized. An updated empirical repertoire should expand the range of observable structures without allowing technical novelty to determine the research question.

Seen in this light, the contemporary challenge is best understood as an uneven expansion of empirical capacity. Established qualitative methods remain central where interpretation is required, while statistical and econometric approaches retain their role in measurement and inference. Newer computational techniques add the ability to work with empirical structures that previously required stronger simplification. Their contribution is therefore cumulative when they are selected according to the restriction they address.

The paper's central proposal is to organize methodological updating around the correspondence between the empirical object and the technical problem that prevents it from being represented adequately. Changes in socioeconomic interaction provide one context in which conventional observation becomes incomplete, while the changing data environment creates additional demands even for familiar substantive objects. The expansion of analytical methods offers concrete responses to those demands. The main risk for the social sciences lies less in failing to adopt every new technique than in allowing a narrow toolkit to determine what can be measured and, ultimately, what can be understood as a social or economic process.

Acknowledgments

None.

Conflicts of interest

The author declares that there is no conflicts of interest.

References

1. Lazer D, Hargittai E, Freelon D, et al. Meaningful measures of human society in the twenty-first century. *Nature*. 2021;595:189–196.
2. Edelmann A, Wolff T, Montagne D, et al. Computational social science and sociology. *Annu Rev Sociol*. 2020;46:61–81.
3. Andreoni A, Roberts S. Governing digital platform power for industrial development: towards an entrepreneurial-regulatory state. *Cambridge J Econ*. 2022;46(6):1431–1454.
4. Gawer A, Bonina C. Digital platforms and development: risks to competition and their regulatory implications in developing countries. *Inf Organ*. 2024;34(3):100525.
5. Vallas S, Schor JB. What do platforms do? Understanding the gig economy. *Annu Rev Sociol*. 2020;46:273–294.
6. Kataishi R. The technological trajectory of semantic analysis: a historical-methodological review of NLP in social sciences. *SSRN*. Published 2024.
7. Kataishi R, Milia M. El análisis semántico-estadístico como estrategia de abordaje metodológico: reflexiones sobre su pertinencia en el estudio de problemáticas latinoamericanas. In: Natera JM, Suárez D, eds. *Métodos para el análisis de los procesos de ciencia, tecnología e innovación: Herramientas para el estudio del desarrollo de América Latina. Volumen 3: Métodos mixtos y emergentes*. Universidad Nacional de General Sarmiento; Universidad Autónoma Metropolitana; 2024:265–308.
8. Brand JE, Zhou X, Xie Y. Recent developments in causal inference and machine learning. *Annu Rev Sociol*. 2023;49:81–110.
9. Geuna A, Kataishi R, Toselli M, et al. SiSOB data extraction and codification: a tool to analyze scientific careers. *Res Policy*. 2015;44(9):1645–1658.
10. Jungherr A, Schoen H, Posegga O, et al. Digital trace data in the study of public opinion: an indicator of attention toward politics rather than political support. *Soc Sci Comput Rev*. 2017;35(3):336–356.
11. Kataishi R. Hybrid retrieval for RAG in the social sciences: aligning latent topics and contextual semantics in enriched embedding spaces. *SSRN*. Published 2025.

12. Boyd d, Crawford K. Critical questions for big data. *Inf Commun Soc*. 2012;15(5):662–679.
13. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM*. 2008;51(1):107–113.
14. Bottou L. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010*. Physica-Verlag; 2010:177–186.
15. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267–288.
16. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer; 2009.
17. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
18. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.
19. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press; 1994.
20. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199–231.
21. Vapnik VN. *The Nature of Statistical Learning Theory*. 2nd edition. Springer; 2000.
22. Bishop CM. *Pattern Recognition and Machine Learning*. Springer; 2006.
23. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Morgan Kaufmann; 2011.
24. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann; 2011.
25. Donoho D. 50 years of data science. *J Comput Graph Stat*. 2017;26(4):745–766.
26. Cleveland WS. Data science: an action plan for expanding the technical areas of the field of statistics. *Int Stat Rev*. 2001;69(1):21–26.
27. Provost F, Fawcett T. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media; 2013.
28. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(2):423–443.
29. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press; 2010.
30. Wooldridge JM. *Introductory Econometrics: A Modern Approach*. 7th ed. Cengage; 2020.
31. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21(1):C1–C68.
32. Knaus MC. Double machine learning-based programme evaluation under unconfoundedness. *Econom J*. 2022;25(3):602–627.
33. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat*. 2019;47(2):1148–1178.
34. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics; 2019:3982–3992.
35. Stöhr F. Advancing language models through domain knowledge integration: a comprehensive approach to training, evaluation, and optimization of social scientific neural word embeddings. *J Comput Soc Sci*. 2024;7:1753–1793.
36. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2019;1:4171–4186.
37. Wankmüller S. Introduction to neural transfer learning with transformers for social science text analysis. *Sociol Methods Res*. 2024;53(4):1676–1752.
38. Kataishi R. Enhancing retrieval-augmented generation: a hybrid approach integrating traditional NLP techniques. Preprint. Published 2025.
39. Kataishi R. Enhancing retrieval-augmented generation with topic-enriched embeddings: a hybrid approach integrating traditional NLP techniques. *Nat Lang Process J*. 2026;14:100200.
40. Heseltine M, Clemm von Hohenberg B. Large language models as a substitute for human experts in annotating political text. *Res Politics*. 2024;11(1).
41. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:855-864.
42. Balland PA, Broekel T, Diodato D, et al. The new paradigm of economic complexity. *Res Policy*. 2022;51(3):104450.
43. Balland PA, Boschma R. Complementary interregional linkages and smart specialisation: an empirical study on European regions. *Reg Stud*. 2021;55(6):1059-1070.
44. Rolf E, Proctor J, Carleton T, et al. A generalizable and accessible approach to machine learning with global satellite imagery. *Nat Commun*. 2021;12:4392.
45. Burke M, Driscoll A, Lobell DB, et al. Using satellite imagery to understand and promote sustainable development. *Science*. 2021;371(6535):eabe8628.
46. Hine C. *Ethnography for the internet: embedded, embodied and everyday*. Bloomsbury; 2015.
47. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320.
48. Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. *J Econ Perspect*. 2014;28(2):29-50.
49. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
50. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2013.
51. Anselin L. *Spatial econometrics: methods and models*. Kluwer Academic; 1988.
52. Jean N, Burke M, Xie M, et al. Combining satellite imagery and machine learning to predict poverty. *Science*. 2016;353(6301):790-794.
53. Barabási AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.
54. Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI)*. USENIX Association; 2012.
55. Lazer DMJ, Pentland A, Watts DJ, et al. Computational social science: obstacles and opportunities. *Science*. 2020;369(6507):1060-1062.

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J*. 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469

56. Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system. In: *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE; 2010.

57. Kreps J, Narkhede N, Rao J. Kafka: a distributed messaging system for log processing. In: *Proceedings of the NetDB Workshop*. 2011.

58. Bühlmann P, van de Geer S. *statistics for high-dimensional data: methods, theory and applications*. Springer; 2011.

59. Breiman L, Friedman JH, Olshen RA, et al. *Classification and regression trees*. Wadsworth; 1984.

60. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A*. 2016;113(27):7353-7360.

61. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.

62. Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud*. 2014;81(2):608-650.

63. Robins G, Pattison P, Kalish Y, et al. An introduction to exponential random graph (p*) models for social networks. *Soc Networks*. 2007;29(2):173-191.

64. Blondel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 2008;2008(10):P10008.

65. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9(1):5233.

66. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*. 2006;103(23):8577-8582.

67. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag*. 1988;24(5):513-523.

68. Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55(4):77-84.

69. Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Anal*. 2013;21(3):267-297.

70. Hidalgo CA, Hausmann R. The building blocks of economic complexity. *Proc Natl Acad Sci U S A*. 2009;106(26):10570-10575.

71. Hidalgo CA, Klinger B, Barabási AL, et al . The product space conditions the development of nations. *Science*. 2007;317(5837):482–487.

72. Boschma R. Proximity and innovation: a critical assessment. *Reg Stud*. 2005;39(1):61–74.

73. Abbott A. Sequence analysis: new methods for old ideas. *Annu Rev Sociol*. 1995;21:93–113.

74. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. 1989;77(2):257–286.

75. Allison PD. *Event History Analysis: Regression for Longitudinal Event Data*. Sage; 1984.

76. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84–90.

77. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*. 2021.

78. Salganik MJ. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press; 2018.

79. Foster I, Ghani R, Jarmin RS, et al. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman and Hall/CRC; 2016.

80. Snee H, Hine C, Morey Y, et al. *Digital Methods for Social Science: An Interdisciplinary Guide to Research Innovation*. Palgrave Macmillan; 2016.

Citation: Kataishi R. The empirical gap: matching social science methods to contemporary data environments. *Soc Int J*. 2026;10(4):125–134.
DOI: 10.15406/sij.2026.10.00469