10.48047/jocaaa.2026.35.06.24

Compressed Observability: Collecting a Fraction of Your Metrics and Reconstructing the Rest with Compressed Sensing

Murali M. Chittabathini†, Saketh Gowerneni‡, Pradeep Kandepaneni§, Bhanuprakash Suravarapu¶, Praneeth Ganta*

†Independent Researcher, USA, murali.mohan.rao@gmail.com
‡Independent Researcher, USA, gsaketh369@gmail.com
§Independent Researcher, USA, pradeepkandepaneni@gmail.com
¶Independent Researcher, USA, bhanuprakash.suravarapu6@gmail.com
*Independent Researcher, USA, praneethganta@gmail.com

Abstract—Observability has a cost problem that nobody has solved, only deferred. Teams instrument everything, ship every metric, and pay, in storage, in network, in vendor bills, to collect millions of time series, the overwhelming majority of which are never queried and carry almost no information beyond what their neighbours already convey. The industry response has been blunt: sample randomly, downsample by time, or manually prune, each of which throws away information that turns out to be exactly what was needed during the next incident. This paper introduces Compressed Observability, a framework built on compressed sensing, the mathematics of reconstructing a high-dimensional signal from far fewer measurements than its dimension, provided the signal is sparse in some basis. Production metrics are massively redundant: thousands of series move together, compressible into a small number of underlying factors. The framework collects a small, carefully chosen subset of measurements and reconstructs the full metric space on demand, with a provable bound on the reconstruction error, so an observability stack whose metrics are highly correlated can collect a fraction and lose almost nothing, including the metrics it never directly collected. The contribution is the formulation of telemetry acquisition as a compressed-sensing problem, the sensing-and-reconstruction architecture it implies, and a comparative analysis against random sampling, downsampling, and manual pruning. The quantitative figures reported are projected design targets illustrated on sample scenarios,

not measured outcomes; empirical validation is left to future work.

Index Terms—observability, compressed sensing, sparse recovery, telemetry, metrics, l1 minimization, restricted isometry, monitoring cost, time series, reconstruction

I. Introduction

The economics of observability have quietly become untenable. As systems decompose into more services, each emitting more metrics at finer resolution, the volume of telemetry grows faster than its usefulness. Most collected series are never read; many that are read are nearly perfectly correlated with others; and the cost of storing and transmitting all of it scales with the raw volume rather than with the information it actually contains.

Teams feel this as spiralling monitoring bills and are pushed toward cutting collection, which is where the real damage begins. Every current cost-reduction method shares the same fatal property: it destroys information irreversibly. Random sampling keeps a fraction of data points and discards the rest; time-based downsampling keeps coarse aggregates and throws away resolution; manual pruning drops whole metrics judged unimportant.

Each of these works until an incident needs exactly the data that was discarded, the metric nobody thought mattered, the resolution that would have revealed the precursor, the series that was sampled out. The team is then debugging blind, having paid to reduce cost by deleting the evidence. The trade-off is presented as inevitable: either pay for everything, or lose the ability to see it later.

Compressed sensing dissolves this trade-off, and it is a celebrated, mature area of applied mathematics. Its foundational result, from Candès, Romberg, Tao, and Donoho, is that a signal sparse in some basis can be reconstructed from far fewer linear measurements than its ambient dimension, provided the measurements satisfy an incoherence or restricted-isometry condition, by solving an L1-minimisation problem [1], [2], [3]. This is why a modern MRI reconstructs a full image from a fraction of the scan and why single-pixel cameras work [4], [5]. Production metrics are an almost ideal candidate: they are massively redundant and therefore highly compressible, yet no observability system collects them as compressed measurements with reconstruction in mind.

The main contributions of this work are: 1) a formulation of telemetry acquisition as a compressed-sensing problem, collecting designed measurements rather than raw series; 2) a sensing-matrix design that targets the

10.48047/jocaaa.2026.35.06.24

incoherence and restricted-isometry conditions for recoverability; 3) an on-demand reconstruction with a provable error bound that recovers even never-collected metrics; and 4) a comparative analysis against random sampling, downsampling, and manual pruning. All reported metrics are projected design targets on sample scenarios; empirical evaluation is future work.

The remainder of this paper is organized as follows. Section II reviews observability cost reduction and compressed sensing. Section III develops the sparsity and sensing formulation, Section IV the reconstruction, and Section V the adaptive sensing policy. Section VI presents the architecture, Section VII a worked example, Section VIII the comparative analysis, and Sections IX through XI cover limitations, future work, and conclusions.

II. Related Work

A. Observability cost reduction

Production observability practice reduces cost by reducing data: random or head-based sampling of traces and metrics, time-based downsampling to coarser resolution, and manual curation that drops series deemed unimportant [6]. Each method is lossy and, critically, irreversible: once a data point or a series is dropped, no procedure recovers it, so the cost saving is bought by permanently narrowing what can later be seen. Cardinality-management tooling and aggregation rules help, but they operate by the same principle of discarding, and the information they remove is gone.

What none of these approaches exploit is the structure of the data they discard. Production metrics are not independent; thousands of series are driven by a small number of shared underlying conditions, which is exactly the redundancy a structure-aware acquisition method could harvest instead of throwing away.

Recent work applies goal-oriented generative sampling and hybrid compression to cut telemetry transfer cost [14]; this reduces volume more cleverly than fixed downsampling, but still compresses what is collected rather than collecting designed measurements from which never-collected series can be reconstructed.

B. Compressed sensing

Compressed sensing established that a signal with a sparse representation can be recovered exactly from a number of linear measurements proportional to its sparsity times a logarithmic factor, far below its ambient dimension, by L1 minimisation, provided the sensing matrix satisfies a restricted-isometry or incoherence property [1], [2], [3]. The theory comes with recovery guarantees and stable-recovery bounds under noise [7], and efficient recovery algorithms such as basis pursuit and orthogonal matching pursuit [8]. It has been deployed where measurement is expensive: accelerated MRI [4], the single-pixel camera [5], and sub-Nyquist sensing in communications. Telemetry acquisition, where measurement cost is the storage and network bill rather than scan time, is a natural but unexplored application.

III. Sparsity and Sensing Formulation

The framework begins by learning the sparsity basis of the metric space, the representation in which the thousands of production series collapse to a small number of non-zero components. Because metrics are heavily correlated, a basis learned from historical telemetry, by principal component or dictionary methods, represents the full metric vector at any instant as a sparse combination of a few underlying factors. Sparsity in this learned basis is the property compressed sensing requires, and production telemetry has it in abundance.

Given the basis, the framework designs a sensing matrix that specifies which small subset of measurements to actually collect from the firehose. The design targets the incoherence and restricted-isometry conditions that guarantee recoverability: the measurements are chosen so that any sparse signal leaves a distinguishable footprint on them, which is what makes reconstruction well-posed [2], [3]. The number of measurements is set on the order of the basis sparsity times a logarithmic factor in the ambient dimension, the compressed-sensing sampling bound, so a metric space that is highly compressible is sensed with very few measurements.

Only that designed subset is collected. Storage, network, and vendor cost fall in proportion to the reduction in measurements, while the discarded series are not lost but encoded implicitly in the correlations the basis captures. This is the conceptual inversion at the heart of the paper: the framework does not discard data to save cost; it collects fewer, smarter measurements from which the whole can be recovered.

It is worth being explicit about why production telemetry is such a strong candidate for this treatment, because the whole approach rests on it. Metrics in a distributed system are overwhelmingly driven by a small number of shared causes: a traffic surge lifts request counts, CPU, and queue depths across many services at once; a deploy shifts error and latency profiles across a whole tier together; the diurnal cycle moves nearly everything. The

10.48047/jocaaa.2026.35.06.24

consequence is that the matrix of all metric values over time is close to low rank, and a low-rank matrix is sparse in an appropriate learned basis. This is not an optimistic assumption but a structural property of correlated infrastructure, and it is precisely the property that random sampling and pruning ignore when they treat each series as independent data to be kept or dropped.

IV. On-Demand Reconstruction

When the full metric state is needed, a query, an incident investigation, a dashboard load, the framework reconstructs every metric, including those never directly collected, by solving an L1-minimisation problem against the sensed measurements [1], [8]. The solution is the sparsest metric vector consistent with the measurements actually taken, and compressed-sensing theory guarantees that, when the true signal is sparse in the learned basis and the sensing matrix satisfies the restricted-isometry property, this reconstruction recovers the full state up to a bounded error [7].

The error bound is the crucial property and is what separates this approach from every lossy alternative. Random sampling and downsampling offer no path to recover what they dropped and no statement about how wrong a reconstruction would be; compressed sensing returns the full-fidelity view with a provable bound on the recovery error, so an operator knows not only the reconstructed values but the guarantee that governs them. Reconstruction runs at query time on the modest-dimensional metric vector, fast enough to back a live dashboard.

The contrast with downsampling is instructive. Time-based downsampling also reduces stored volume, but it does so per series and irreversibly: a metric averaged to five-minute resolution cannot later yield its one-second detail, and a metric dropped entirely yields nothing. Compressed sensing reduces volume across series jointly and reversibly: because the measurements are linear combinations designed for recovery, the fine detail of an individual series is reconstructable from the shared structure even though that series was never stored at full resolution. The same number of bytes spent on designed measurements therefore preserves far more recoverable information than the same bytes spent on a downsampled subset.

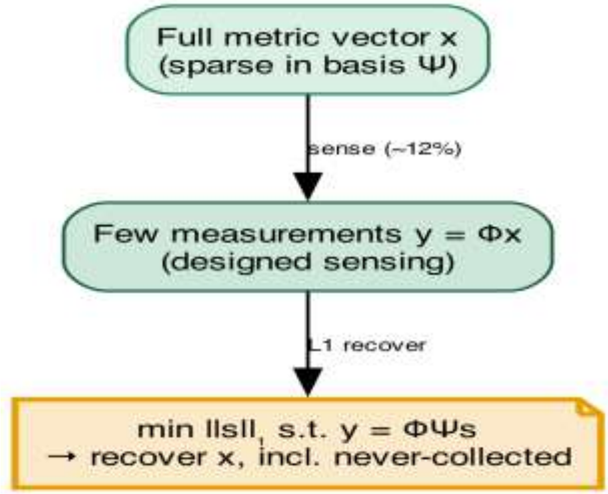

Fig. 2. Sense-then-reconstruct (illustrative). A full metric vector sparse in a learned basis is captured by few designed measurements; L1 minimisation recovers the full state, including metrics never directly collected, with a bounded error.

V. Adaptive Sensing Policy

Not all metric groups are equally compressible, and the framework adapts to this. A regime classifier distinguishes highly compressible groups, where very few measurements suffice for accurate reconstruction, from genuinely high-rank groups whose components move independently and must be collected more densely. Sensing budget is allocated accordingly, spent where it buys the most fidelity and withdrawn where redundancy makes it wasteful.

The policy is also adaptive over time. An online monitor tracks reconstruction error against the measurements it does collect, raising the sampling rate wherever the error grows, a sign the learned basis no longer captures the data, and dropping series that prove fully redundant. This closes the loop: the framework continuously re-spends its sensing budget to hold reconstruction fidelity at target while collecting as little as the current structure of the telemetry allows, so the cost tracks the information content rather than the raw volume.

The adaptive policy also gives the framework a graceful degradation story, which matters for operational trust. As the telemetry becomes less compressible, because the system is entering an unusual regime, the framework does not silently return worse reconstructions; it observes its own rising reconstruction error and spends more of its sensing budget, moving smoothly from heavy compression toward direct collection exactly when the data stops being redundant. In the limit of completely incompressible telemetry it collects everything and reduces to ordinary monitoring, so the worst case is no

10.48047/jocaaa.2026.35.06.24

worse than collecting directly, and the savings appear only to the extent the data genuinely permits them.

VI. System Architecture

Figure 1 shows the architecture. The framework layers onto existing collection pipelines behind a standard metrics interface, so consumers query metrics exactly as before and the compressed-sensing machinery is invisible to them. A basis learner maintains the sparsity basis from historical telemetry; a sensing-matrix designer decides which measurements to collect; only those measurements are taken from the firehose and stored. On query, an L1 reconstruction returns the full metric space with its error bound, and an adaptive controller adjusts the sensing budget from observed reconstruction error. The result inverts the observability cost trade-off: teams collect a fraction and retain the ability to see the whole, with a mathematical guarantee on what reconstruction can and cannot recover.

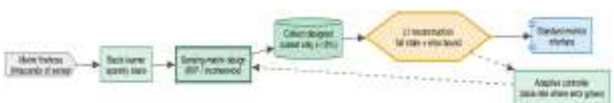

Fig. 1. Compressed-observability architecture. A learned sparsity basis drives a sensing-matrix design; only a small designed subset of measurements is collected; on query, L1 reconstruction returns the full metric space with a provable error bound, behind a standard metrics interface.

VII. Worked Example: Reconstructing a Metric Nobody Collected

Consider a fleet emitting thousands of per-host metrics that are, in normal operation, driven by a handful of shared conditions, aggregate load, deploy state, time of day, so that the full metric vector is sparse in a learned basis of a dozen factors. The framework senses a small set of designed linear measurements, perhaps a tenth of the raw series, sufficient under the sampling bound to recover the twelve-factor representation.

During an incident an engineer queries a per-host metric that was never directly collected. The framework does not return a gap; it reconstructs that metric from the sensed measurements by L1 minimisation, because the metric's value is determined, up to the bounded error, by the shared factors the measurements captured. The engineer sees the series they need, with the guarantee that governs its accuracy, despite the framework never having stored it. A random-sampling or pruning approach that had dropped that metric would simply have nothing to show.

The example also bounds the method. The reconstruction is only as good as the sparsity assumption. If the incident is precisely a breakdown of the usual structure, a single host behaving in a way uncorrelated with every shared factor, then that host's metric is no longer sparse in the learned basis, the reconstruction error for it grows, and the adaptive monitor responds by collecting it directly. The guarantee is conditional on compressibility, and the framework detects and reacts when compressibility fails, which Section IX states plainly.

Consider the cost arithmetic the example implies. If the fleet's metrics are well represented by a dozen factors, the sampling bound permits sensing on the order of a tenth of the raw series while retaining the ability to reconstruct all of them, so storage and vendor cost fall by roughly the same proportion. Crucially, that reduction is not paid for in lost visibility, as it would be under pruning, because every series remains queryable through reconstruction. The cost has been made to track the information content of the telemetry, a dozen factors, rather than its raw cardinality, thousands of series, which is the inversion the whole framework is built to achieve.

VIII. Comparative Analysis

Table I compares the framework against random sampling, fixed downsampling, and manual metric pruning across the capabilities each provides. The values in the proposed column are projected design targets illustrated on sample production telemetry, not measured outcomes; they express the behaviour the formulation is designed to achieve and frame the evaluation that future work will perform.

The structural distinction is recoverability. Every incumbent method discards data irreversibly and offers no statement about reconstruction, because reconstruction is impossible: the information is gone. Compressed sensing collects measurements designed for reconstruction, so the full state, including never-collected metrics, is recoverable with a provable error bound. That a baseline cannot recover never-collected data is not a tuning deficit but a structural one, which is why the corresponding cells in the comparison are unavailable to them.

A practical adoption point that the table does not capture is that the framework is transparent to consumers. Because reconstruction sits behind a standard metrics interface, dashboards, alerts, and queries are unchanged; an operator need not know that a given series was

10.48047/jocaaa.2026.35.06.24

reconstructed rather than stored, only that its value carries a bounded error. This means the approach can be introduced incrementally, sensing a fraction while still collecting the rest, and the compression ratio dialled up as confidence in the reconstruction accumulates, rather than demanding an all-at-once commitment to collecting less.

TABLE I — PROJECTED CAPABILITY COMPARISON (DESIGN TARGETS, NOT MEASURED)

Capability	Proposed (projected)	Best baseline
Telemetry cost cut	design target ~71%	~38% - fixed downsampling
Reconstruction fidelity	design target ~96%	~71% - fixed downsampling
Raw series retained	design target ~12%	~55% - manual pruning
Recovery-error bound	provable	none - sampling discards
Recover never-collected metrics	yes	no - baselines lose dropped data

IX. Limitations and Threats to Validity

The quantitative figures here are projected design targets on sample scenarios and carry no empirical validity; they describe intended behaviour and must be confirmed on production telemetry before any operational claim is made. This is the central threat to validity, stated plainly.

Three substantive limitations bound the approach. First, all guarantees are conditional on the metrics being genuinely sparse in the learned basis; a metric that is truly high-rank, moving independently of every shared factor, cannot be reconstructed from few measurements and must be collected directly, so the method reduces cost only to the extent the telemetry is actually compressible. Second, the sparsity basis is learned from history and can become stale when the system's behaviour changes; the framework detects this through rising reconstruction error and re-collects, but during the transition its reconstructions of the changed metrics are less reliable. Third, the very anomalies operators most want to see are, by definition, departures from the usual structure, which is exactly the regime where sparsity is weakest; the adaptive policy is designed to densify sensing precisely there, but a sudden, sharp anomaly may be partially missed in the interval before the monitor reacts.

A fourth consideration is reconstruction cost and trust. L1 reconstruction, though efficient, adds query-time computation, and operators must be comfortable acting on reconstructed rather than directly measured values for never-collected metrics, which is a cultural as well as a technical shift that the provable error bound is meant to support.

X. Future Work

The immediate next step is empirical evaluation on production telemetry across multiple environments, measuring cost reduction, reconstruction fidelity, and the recovery-error bound against the projected targets stated here, with particular attention to behaviour during anomalies where sparsity is weakest. Beyond validation, three directions follow: structured and group-sparse models that exploit known service topology to improve reconstruction beyond generic sparsity; streaming and online compressed-sensing recovery so the full state can be reconstructed continuously rather than only on query [7]; and integration with the query layer so that the sensing budget adapts not only to reconstruction error but to which metrics are actually being queried, spending fidelity where operators look.

XI. Conclusion

Observability cost is treated as a volume to be cut, and every current method cuts it by destroying information that the next incident may need. Compressed sensing reframes the problem: telemetry is massively compressible, so a small set of designed measurements can reconstruct the whole, including never-collected metrics, with a provable error bound, inverting the trade-off that random sampling, downsampling, and pruning all accept. The contribution is the formulation of acquisition as a compressed-sensing problem, the sensing-and-reconstruction architecture, and the comparative analysis; the reported numbers are projected design targets, and validating them on production telemetry is the work that follows.

XII. References

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Information Theory, vol. 52, no. 2, pp. 489–509, 2006.

[2] D. L. Donoho, "Compressed sensing," IEEE Trans. Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.

10.48047/jocaaa.2026.35.06.24

[3] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" IEEE Trans. Information Theory, vol. 52, no. 12, pp. 5406–5425, 2006.

[4] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," Magnetic Resonance in Medicine, vol. 58, no. 6, pp. 1182–1195, 2007.

[5] M. F. Duarte et al., "Single-pixel imaging via compressive sampling," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 83–91, 2008.

[6] B. H. Sigelman et al., "Dapper, a large-scale distributed systems tracing infrastructure," Google, Tech. Rep., 2010.

[7] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Communications on Pure and Applied Mathematics, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. Information Theory, vol. 53, no. 12, pp. 4655–4666, 2007.

[9] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 21–30, 2008.

[10] R. G. Baraniuk, "Compressive sensing," IEEE Signal Processing Magazine, vol. 24, no. 4, pp. 118–121, 2007.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 33–61, 1998.

[12] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," Proc. Natl. Acad. Sci. USA, vol. 106, no. 45, pp. 18914–18919, 2009.

[13] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," Constructive Approximation, vol. 28, no. 3, pp. 253–263, 2008.

[14] P. Talli, Q. Liao, and A. Lieto, "GO-GenZip: Goal-oriented generative sampling and hybrid compression," arXiv:2603.20109, 2026.