LabelScope: Context-Aware Discovery of Label Outliers in Radio Astronomy Signals

Rabeya Hossain
University of Utah
Salt Lake City, USA
rabeya.hossain@utah.edu

Brian Mason
National Radio Astronomy Observatory
Charlottesville, USA
bmason@nrao.edu

Ryan Loomis
National Radio Astronomy Observatory
Charlottesville, USA
rloomis@nrao.edu

Jeff M. Phillips
University of Utah
Salt Lake City, USA
jeff.m.phillips@utah.edu

El Kindi Rezig
University of Utah
Salt Lake City, USA
elkindi.rezig@utah.edu

ABSTRACT

The National Radio Astronomy Observatory (NRAO) operates one of the largest radio telescopes in the world. As part of its data processing pipeline, it produces thousands of bandpass calibration signals per observation cycle, each capturing how an antenna's frequency response varies across a spectral window—essentially a sequence of amplitude measurements over frequency channels. To ensure data quality, automated detectors assign anomaly scores to these signals, and human experts review flagged cases to assign labels. In practice, scores and labels can disagree, and expert review cannot scale to every mismatch. We present LabelScope, an interactive system developed in collaboration with NRAO to address this problem. LabelScope groups signals using contextual metadata and feature augmentation, identifies minority-label instances within locally coherent groups, and generates predicate-based explanations for why they are interesting disagreement cases. Through its interface, users explore candidate label outliers and determine whether score–label disagreements reflect labeling errors, meaningful anomalies, or limitations of the scoring method. We demonstrate LabelScope on real astronomy signals, enabling NRAO scientists to efficiently direct limited reviewer attention toward the most informative disagreements. Beyond astronomy, this demonstration exposes VLDB 2026 attendees to a class of real-world data quality problems, i.e., contextual disagreements between automated anomaly scoring and human labeling, that arises broadly in scientific data pipelines but remains underexplored in the data management literature.

PVLDB Reference Format:
Rabeya Hossain, Brian Mason, Ryan Loomis, Jeff M. Phillips, and El Kindi Rezig. LabelScope. PVLDB, 19(12): 4686 - 4689, 2026.
doi:10.14778/3827998.3828097

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828097

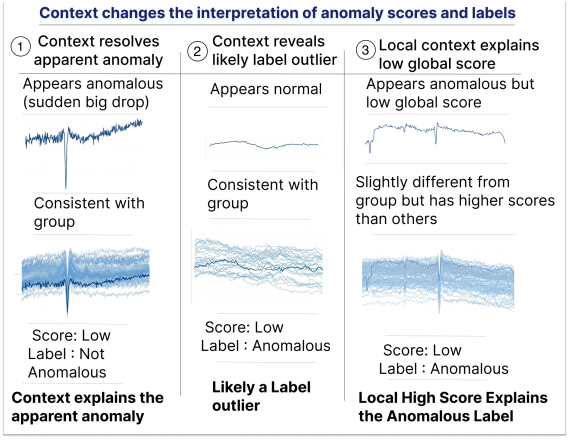

Figure 1: Motivating examples illustrating how context changes the interpretation of anomaly scores and labels.

1 INTRODUCTION

Radio telescopes are growing larger and capturing increasing volumes of signals from space, yielding rapidly growing and more complex datasets [3]. As system size and processing complexity grow, so does the likelihood of anomalies from instrument failures, miscalibrated observations, environmental interference, astronomical phenomena, or processing artifacts [7]. Automated anomaly detection is therefore essential for flagging signals that need attention, since unresolved anomalies may propagate into later processing stages, affecting downstream data quality and sometimes the reliability of scientific conclusions [1]. However, such detectors are often approximate or heuristic, and their outputs do not always align with downstream human judgment: a signal given a low anomaly score by an automated technique may later be labeled anomalous by a human expert.

The National Radio Astronomy Observatory (NRAO) operates the Atacama Large Millimeter/submillimeter Array (ALMA) [10], one of the most powerful radio telescopes in the world. ALMA's calibration pipeline processes thousands of bandpass signals per observation cycle, each representing the frequency response of

an antenna across a spectral window (a contiguous range of frequency channels that the telescope observes simultaneously). To ensure data quality, an automated detector assigns anomaly scores to these signals, and human experts at NRAO review those scores to assign labels. In practice, however, the scores and labels can disagree: signals deemed normal by the detector are sometimes labeled anomalous by reviewers, and vice versa. These disagreements are not simply noise. They arise because the detector relies on fixed heuristics that cannot anticipate every type of anomaly introduced by distant astronomical sources, instrument drift, or environmental interference, while expert review is performed under operational constraints, with varying levels of expertise and a cautious tendency to overflag ambiguous cases. Since expert review cannot scale to every mismatch, NRAO needs a principled way to surface the most informative disagreements and direct limited reviewer attention where it matters most.

In the dataset used in this demonstration, each signal is associated with both a manually assigned label and an anomaly score computed by an external anomaly detector. Although the system can operate with scores produced by any external method, in our setting the anomaly scores are generated externally using scan statistics [5]. More generally, however, the broader problem is not restricted to scores and labels alone: it arises whenever the outputs of two stages in a processing pipeline disagree and those disagreements require contextual interpretation.

Motivating Example. Figure 1 illustrates why anomaly scores must be interpreted relative to context: in isolation each of these signals is ambiguous, and only its contextual group—signals sharing the same spectral window and execution block (a single uninterrupted observation session)—makes its score and label interpretable. In the first example (Figure 1①), the signal looks anomalous on its own because of a sharp drop, but the same drop is unremarkable across its group, so the low score and label are consistent with the local context. In the second example (Figure 1②), the signal looks unremarkable both on its own and within its group, so its anomalous label is puzzling; since neither the signal nor its context supports the label, it is a strong label-outlier candidate. In the third example (Figure 1③), the signal is labeled anomalous and, though its global score is low, it is still higher than its neighbors', whose scores are even lower—so the score is consistent with the local context and supports the label, which without the group could be mistaken for a label outlier. More importantly, contextual grouping enables label-quality assessment: many groups show strong internal agreement, and when both the scores and the context support the dominant pattern but a few labels differ, those minority cases become strong candidates for labeling errors [2].

Contributions. We present LabelScope, an interactive system for context-aware analysis of disagreements in scored and labeled signal datasets. LabelScope groups signals using contextual metadata and feature augmentation, surfaces groups with strong local agreement, and highlights signals whose labels disagree with both their anomaly scores and their contextual neighborhood, helping analysts identify likely label errors that would otherwise require costly manual inspection. Although demonstrated on ALMA data, it reflects a broader paradigm in which disagreements between automated scores, downstream labels, and contextual similarity are analyzed jointly.

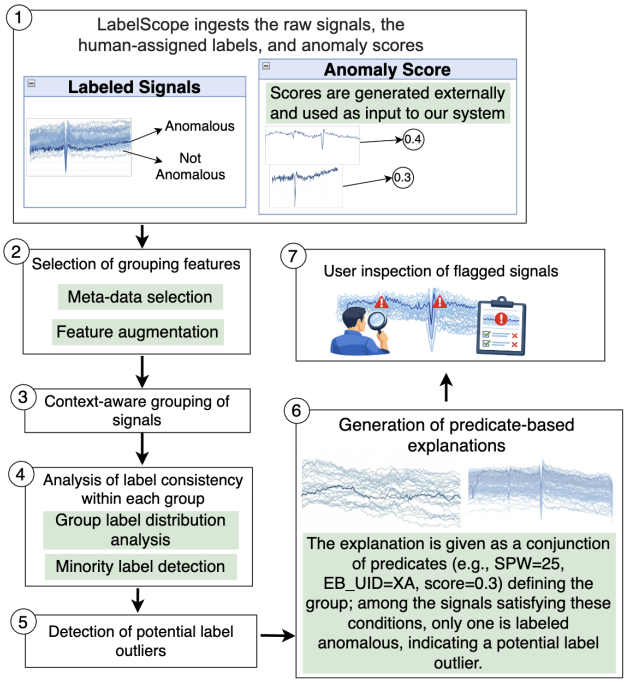

Figure 2: LabelScope's architecture. LabelScope ingests labeled signals and their anomaly scores, groups them into contextual regions based on selected features, analyzes them for label inconsistencies, and presents them to users with interpretable explanations for interactive inspection.

Related Work. Prior work on anomaly detection focuses on producing scores that rank unusual signals [4, 5]; contextual anomaly detection score–label disagreements.interprets such signals relative to similar instances [8], and label-quality methods identify potentially mislabeled data [2]—but these lines are studied separately. However, these lines of work are typically studied separately. They do not support joint analysis of disagreements between automated scores, contex- tual similarity, and downstream labels. LabelScope addresses this gap through context-aware grouping, minority-label analysis, and interactive inspection of candidate label outliers.

2 SYSTEM OVERVIEW

Figure 2 presents the architecture of LabelScope: it ingests labeled signals with anomaly scores (Figure 2 ②), and LabelScope partitions signals into similar-feature groups (Figure 2 ③), analyzes each group's label distribution (Figure 2 ④), and surfaces minority-label signals as candidate label outliers (Figure 2 ⑤) with predicate-based explanations (Figure 2 ⑥) that the user inspects against contextual neighbors to decide whether each case is a labeling error, a meaningful anomaly, or a scoring limitation (Figure 2 ⑦). We detail each stage below.

2.1 Signal Representation and Anomaly Scoring

Signal Representation. The input to LabelScope is a collection of bandpass calibration signals from ALMA observations (Figure 2 ①).

A signal is $s = \{(f_1, a_1), \ldots, (f_m, a_m)\}$, where f_i is a frequency channel and a_i its amplitude; each instance also carries metadata, e.g., spectral window, execution block, and a manually assigned anomaly label.

Anomaly Scoring. Each signal is assigned an anomaly score $\sigma(s)$. In LabelScope, this score can come from any detector operating on the signal sequence; in our setting it is computed by a scan-statistics-based [5] method from domain experts that detects localized irregularities—large drops or peaks—by sliding a window across the amplitude profile.

2.2 Context Construction and Partitioning

How anomaly scores should be interpreted depends on which signals are treated as similar [8]. Since this notion varies across datasets and tasks, LabelScope constructs context in a user-guided manner and partitions the dataset accordingly.

2.2.1 User-Guided Feature Selection. Let $F = \{f_1, \ldots, f_k\}$ be the attributes available for each signal, including observational metadata and augmented features capturing signal shape (Figure 2 ②). Users select a subset $F_c \subseteq F$ that defines the similarity feature space, adapting the grouping to the dataset and task.

2.2.2 Context-Aware Partitioning. Given F_c, LabelScope partitions the dataset into groups of signals similar in those features (Figure 2 ③); each group is a local region of feature space within which scores and labels can be meaningfully compared.

Why naive grouping is insufficient. Explicitly enumerating combinations of feature values or threshold intervals scales combinatorially when multiple features are considered jointly, yielding many sparse and analytically weak groups. **Random Forest-based partitioning** To avoid this exhaustive search, LabelScope employs a Random Forest model trained on the selected features F_c together with the signal labels. Each decision tree recursively partitions the dataset by selecting feature splits that improve label separability. We choose a Random Forest over unsupervised clustering or manual binning because its supervised, its splits are simple per-feature conditions, so each group is defined by a readable rule that is also its explanation.

A root-to-leaf path defines a conjunction of predicates over the selected features, and signals routed to the same leaf node satisfy the same predicates. These signals are therefore assigned to the same contextual group. This way, the Random Forest efficiently discovers meaningful combinations of feature conditions without explicitly enumerating all possibilities.Each tree is constrained—maximum depth 5, at least 10 signals per leaf, and minimum impurity decrease 0.01—so a disagreeing label stays visible as a minority within a coherent group rather than isolated in its own leaf; because a human reviewer, not the model, adjudicates each flagged case, training on the audited labels does not form a closed loop.

2.3 Label Distribution Analysis

Each leaf node defines a contextual group $G = \{s_1, \ldots, s_n\}$ with labels $\{y_1, \ldots, y_n\}$, and LabelScope analyzes each group's label distribution to find potential inconsistencies (Figure 2 ④). A group is *pure* if all signals share one label, indicating strong agreement, and is excluded; *impure* groups contain multiple labels and represent potential labeling conflicts. Impure groups with a small minority relative to their size indicate strong contextual agreement with few deviations, so those minority signals are flagged as candidate label outliers (Figure 2 ⑤).

2.4 Predicate-Based Explanation Generation

For each candidate, LabelScope generates a predicate-based explanation (Figure 2 ⑥) capturing both the conditions under which the signal was grouped with its neighbors and the label disagreement that makes it suspicious. Traversing the decision path from the root to the leaf containing s yields predicates $P = \{p_1, \ldots, p_l\}$ whose conjunction is the group's contextual rule $E(s) = p_1 \wedge \cdots \wedge p_l$. Since a signal is flagged only when its label differs from the group's dominant label, the explanation reads: it satisfies the same contextual conditions as the others, but its label differs—e.g., *spectral window = 31 ∧ execution block = XAA2 ∧ anomaly score = 0.3*, where most signals are anomalous but this one is normal.

3 DEMONSTRATION PLAN

Demonstration outline. The user interacts with LabelScope to discover label outliers through context-aware grouping, rule exploration, visualization, and predicate-based explanations. The demonstration uses ALMA bandpass calibration solutions: amplitude values across frequency channels, with metadata describing the observational setting, augmented features (e.g., noise level, sharp amplitude variations), a human-assigned label, and an externally generated anomaly score. A companion video is available online.[1]

① **Uploading data and scoring module.** The user uploads the input dataset (Figure 3 ❶) and, optionally, a scoring module (Figure 3 ❷), showing that LabelScope is not tied to a specific detector.

② **Previewing data and selecting features.** The system previews the available records and attributes (Figure 3 ❸); the user selects the metadata and augmented features defining contextual similarity (Figure 3 ❹).

③ **Configuring analysis parameters.** The user adjusts the anomaly threshold (Figure 3 Ⓐ) and minority-ratio parameter (Figure 3 Ⓑ): signals above the threshold are anomalous, and the minority ratio controls how much label disagreement is allowed within a group.

④ **Exploring contextual rules.** LabelScope generates a ranked list of contextual rules (Figure 3 ❺); each summarizes a group through shared feature conditions and reports group size, label distribution, and minority-to-group ratio. Each rule is also rendered as a plain-language sentence with on-hover glosses for domain-specific fields (e.g., SPW = spectral window, EB_UID = execution-block identifier), so attendees without astronomy background can read it.

⑤ **Visual inspection of candidate outliers.** The user chooses how to visualize signals within a group (Figure 3 ❻); candidate outliers are highlighted relative to the other signals (Figure 3 ❼). For large groups, neighbor traces are drawn at low opacity as a density band while the candidate is highlighted in a contrasting color, and a single-index mode further declutters dense groups.

[1] https://drive.google.com/drive/folders/1Vsx6kMAcpNOxoDOChQy7r1PNOy50lzxK (last accessed June 24, 2026).

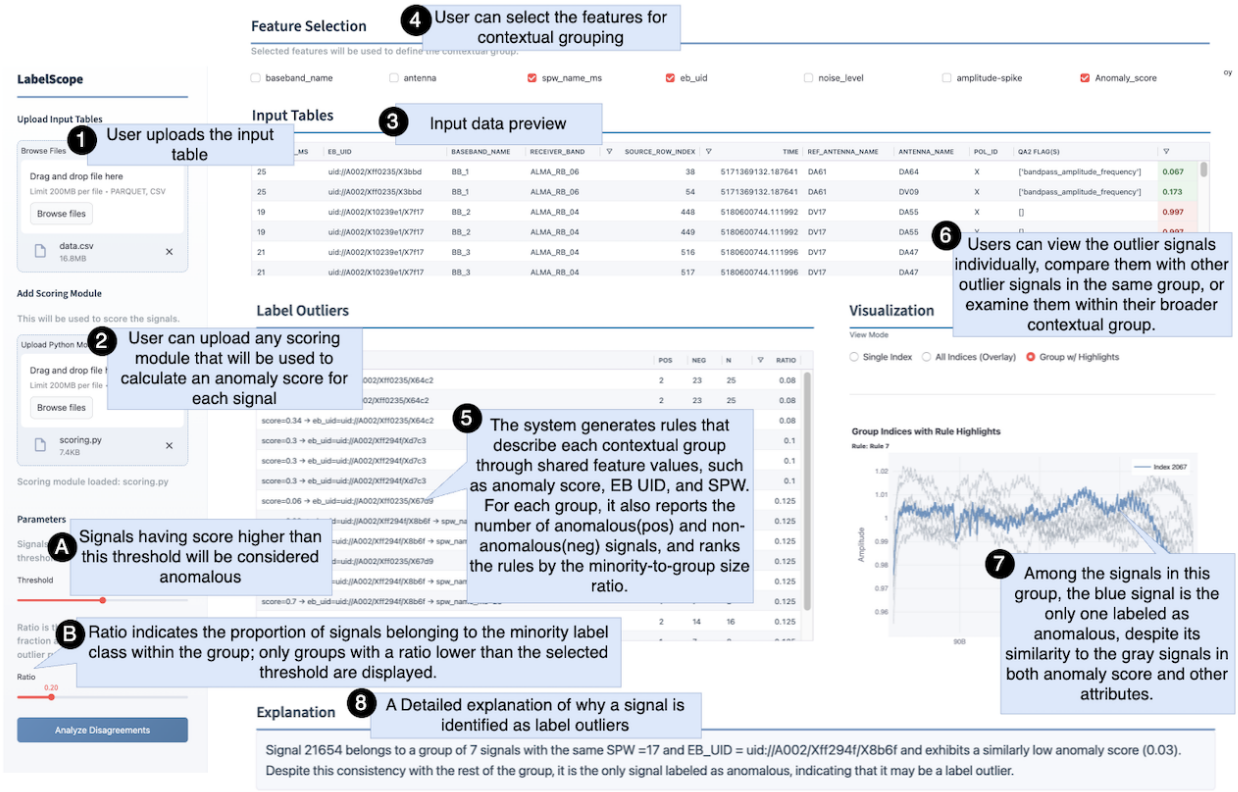

Figure 3: The LabelScope interface showing dataset upload, grouping configuration, ranked suspicious groups, and signal-level inspection panels.

⑥ **Examining explanations.** For each flagged signal, LabelScope provides a predicate-based explanation (Figure 3 ⑧) describing the group's defining conditions and the observed label disagreement, supporting validation of whether the case is a labeling inconsistency, a meaningful anomaly, or a scoring limitation.

Demonstration engagement. VLDB 2026 attendees work hands-on with real ALMA data from NRAO, triaging score–label disagreements across thousands of signals. LabelScope responds immediately: changing features changes the anomaly context, the minority threshold surfaces different candidates, and selecting a flagged signal reveals its explanation beside its neighbors. The same score–label–context pattern recurs across pipelines—from Kepler exoplanet vetting [6, 9] to non-astronomy settings with reviewer-adjudicated scores—and LabelScope transfers by changing only the features and scoring module.

Acknowledgments This work was supported by the National Science Foundation under Grant No. AST-2421782 and Simons MPS-AI-00010515.

REFERENCES

[1] Alexander B. Akins, Sara Seager, Janusz J. Petkowski, Sukrit Ranjan, and Jane S. Greaves. 2021. Complications in the ALMA Detection of Phosphine at Venus. *The Astrophysical Journal Letters* 907, 2 (2021), L27. https://doi.org/10.3847/2041-8213/abd56a

[2] Carla E. Brodley and Mark A. Friedl. 1999. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research* 11 (1999), 131–167. https://doi.org/10.1613/jair.606

[3] Robert J. Brunner, S. George Djorgovski, Thomas A. Prince, and Alexander S. Szalay. 2001. Massive Datasets in Astronomy. *arXiv preprint astro-ph/0106481* (2001). arXiv:astro-ph/0106481

[4] Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. 2024. Deep Learning for Time Series Anomaly Detection: A Survey. *Comput. Surveys* 57, 1 (2024), 1–42. https://doi.org/10.1145/3691338

[5] Martin Kulldorff. 1997. A Spatial Scan Statistic. *Communications in Statistics – Theory and Methods* 26, 6 (1997), 1481–1496. https://doi.org/10.1080/03610929708831995

[6] Sean D. McCauliff, Jon M. Jenkins, Joseph Catanzarite, Christopher J. Burke, Jeffrey L. Coughlin, Joseph D. Twicken, Peter Tenenbaum, Shawn Seader, Jie Li, and Miles Cote. 2015. Automatic Classification of Kepler Planetary Transit Candidates. *The Astrophysical Journal* 806, 1 (2015), 6. https://doi.org/10.1088/0004-637X/806/1/6

[7] M. Mesarcik, A.-J. Boonstra, M. Iacobelli, E. Ranguelova, C. T. A. M. de Laat, and R. V. van Nieuwpoort. 2023. The ROAD to Discovery: Machine-Learning-Driven Anomaly Detection in Radio Astronomy Spectrograms. *Astronomy & Astrophysics* 680 (2023), A74. https://doi.org/10.1051/0004-6361/202347182

[8] Xiuyao Song, Mingxi Wu, Christopher M. Jermaine, and Sanjay Ranka. 2007. Conditional Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering* 19, 5 (2007), 631–645. https://doi.org/10.1109/TKDE.2007.1009

[9] Susan E. Thompson, Jeffrey L. Coughlin, Kelsey Hoffman, Fergal Mullally, Jessie L. Christiansen, Christopher J. Burke, Steve Bryson, Natalie Batalha, et al. 2018. Planetary Candidates Observed by Kepler. VIII. A Fully Automated Catalog with Measured Completeness and Reliability Based on Data Release 25. *The Astrophysical Journal Supplement Series* 235, 2 (2018), 38. https://doi.org/10.3847/1538-4365/aab4f9

[10] Alwyn Wootten and A. Richard Thompson. 2009. The Atacama Large Millimeter/Submillimeter Array. *Proc. IEEE* 97, 8 (2009), 1463–1471. https://doi.org/10.1109/JPROC.2009.2020572