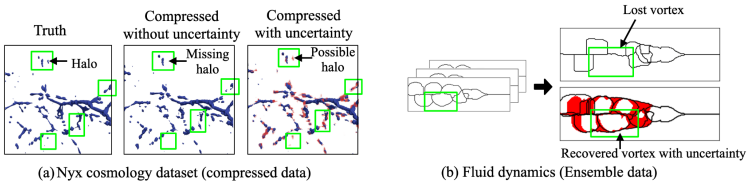
Uncertainty-Aware Inverse Feature Mapping for Reliable Scientific Discovery

Tushar M. Athawale[1] (athawaletm@ornl.gov), David Pugmire[1] (pugmire@ornl.gov), Chris R. Johnson[2] (crj@sci.utah.edu), Kenneth Moreland[1] (morelandkd@ornl.gov), Paul Rosen[2] (paul.rosen@utah.edu), Alireza Entezari[3] (entezari@ufl.edu), Antigoni Georgiadou[1] (georgiadoua@ornl.gov), Tom Beck[1] (becktl@ornl.gov)
[1]Oak Ridge National Laboratory, USA, [2]University of Utah, USA, [3]University of Florida, USA

Topic: Inverse problems using incomplete, noisy, or multi-modal data.

Challenge: Large scientific datasets produced by modern supercomputers and experimental facilities are inherently uncertain due to a number of factors. These factors include fixed-bit representations, sensor inaccuracies, approximations in simulation models, uncertain initial and boundary conditions for simulations, and the effects of spatiotemporal discretization. Extreme-scale data often need to be reduced and compressed to meet storage bottlenecks and for post-hoc analysis, which further amplifies uncertainty in the data. Given accumulation of uncertainty in data from noisy acquisition and preprocessing workflows, data-driven inverse analysis faces a major challenge of loss or truncation of important features in the data, thereby leading to either incorrect scientific conclusions or spending significant time on assessing and resolving uncertainty. Thus, the propagation of uncertainty through data-driven feature extraction algorithms is crucial to mitigating misinformation, a challenge that must be addressed and prioritized to ensure reliability of scientific discovery under uncertainty.

As shown in Fig. 1, ignoring uncertainty in the data can lead to significant errors in inverse analysis. For a common data compression use case in Fig. 1a, ignoring error bounds used for data compression can lead to a loss of halo features (indicative of dark energy matter) that are important to cosmologists. Propagation of error bounds into feature detection algorithms, however, can help to recover the missing halo features (depicted in red), thereby enabling robust scientific analysis and discovery. As illustrated in Fig. 1b, for ensemble datasets, feature extraction without uncertainty propagation (e.g., features of the statistical mean of the ensemble) can lead to distortion or loss of vortex features. However, incorporating uncertainty into the feature extraction process can recover vortex features, which can have a significant positive impact on vortex shedding studies in aerodynamics and fluid dynamics domains.



(a) Nyx cosmology dataset (compressed data) (b) Fluid dynamics (Ensemble data)

Figure 1: Enhanced feature recovery through uncertainty-aware inverse mapping (uncertainty depicted in red) indicating significant positive impacts on scientific conclusions and discovery.

Incorporating uncertainty into inverse feature mapping, however, faces significant barriers. (1) *There is no standard theory on how to best model uncertainty in data with different modalities for propagation into feature extraction*. Uncertainty can be represented with parametric or nonparametric probability distributions for ensemble data [1] or with error bounds used in data compression [2]. However, there is no widespread understanding of how uncertainty can be optimally modeled for data with different modalities (e.g., data compressed with a wide range of data compressors, data reduced with machine learning (ML)/ artificial intelligence (AI) surrogates, multivariate and multidimensional data). (2) Even though uncertainty can be effectively modeled for data with different modalities, most data-driven feature extraction algorithms (e.g., critical points, topological segmentation, streamline, and vortex core tracking) ignore uncertainty in the data due to complex nonlinear interaction between uncertain data and algorithmic models. Such a *lack of understanding of nonlinear interactions between uncertain data and complex algorithmic models prevents analysis of uncertainty propagated into extracted features*, thereby impeding reliable scientific analysis. To circumvent complexity of nonlinear interactions, most existing algorithms resort to Monte Carlo sampling of input uncertain data to estimate uncertainty in extracted features. The Monte Carlo models, however, cannot be used in practice due to their prohibitive costs and limited scalability with an

increase in the data size, dimensions, and variables. (3) Furthermore, treating uncertainty as data in its own right can increase the storage and computational burden of systems. *Such a memory and computational cost overhead can prevent efficient analysis of uncertainty in derived data features.* Significant research is needed to understand how data uncertainty can be represented in a memory-efficient manner, how uncertainty computation can be scaled and accelerated with Department of Energy's (DOE's) leadership computing facilities, and how cost-accuracy trade-offs of different uncertainty models can be balanced to make propagation of uncertainty in inverse features efficient and practical.

Opportunity: The mentioned barriers to propagating uncertainty into data-driven feature extraction present several new opportunities for research in robust and trustable scientific discovery. (1) There are multiple new directions that can be explored for effective modeling of uncertainty in the data with diverse modalities. Widely used data compression techniques (e.g., MGARD, SZ, and ZFP) and data reduction techniques (e.g., adaptive mesh refinement [AMR] and multiresolution data) [2] must be studied to understand how uncertainty can be optimally modeled in these techniques. The use of ML/AI models is becoming pervasive to compactly represent large simulation data, however, representation of uncertainty for these models (e.g., ensemble and Monte Carlo dropout methods [3]) must be studied for reliable feature extraction. New techniques based on information-driven metrics (e.g., correlation, entropy, and mutual information) should be investigated for accurate uncertainty representation. For example, uncertainty with large entropy can be represented by a uniform distribution, and uncertainty with small entropy can be represented by histograms. (2) Extensive research is needed to improve theoretical understanding of how uncertainty in data is nonlinearly transformed by inverse feature mappings. There are a few theoretical developments on how uncertainty modeled with probability distributions is propagated into inverse mappings (e.g., critical points [1] and level-sets [2]). However, such work must be expanded to other data-driven algorithms fundamental to scientific discovery (e.g., local gradient and Hessian determination, topology extraction, and streamline tracking with Euler/Runge-Kutta integration). Such an investigation will be critical to overcoming the inefficiency and limited scalability of Monte Carlo methods and radically improving credibility of analysis. (3) New data structures (e.g., octrees, hash maps) and utilization of data compressors and ML/AI models must be investigated to reduce memory footprint of storing uncertainty. Utilization of low-/mixed-precision can be investigated for memory-efficient representation of uncertainty parameters (e.g., standard deviation, number of histogram bins) and to speed up uncertainty propagation. In addition, parallelization of algorithms with acceleration using DOE's parallel resources and balancing of cost-accuracy overheads of diverse uncertainty models (e.g., Gaussian vs. histogram) must be investigated to further speedup uncertainty propagation [1]. The above research thrusts will be key to efficiently storing and propagating uncertainty and reducing time to assess uncertain data, enabling reliable data analysis under uncertainty.

Innovation: The capability to efficiently derive uncertainty in inverse feature mappings for data with different modalities and complexities is presently lacking, which prevents scientists from trusting their analysis and, consequently, spending significant time to assess uncertainty. Innovating algorithms that enable efficient and accurate quantification of uncertainty in feature extraction will be key to overcoming the inefficiency of existing Monte Carlo methods, integrating uncertainty into data analysis pipelines and tools, and enabling timely and reliable analysis and discovery.

References: [1] T. M. Athawale et al., "Uncertainty Visualization of Critical Points of 2D Scalar Fields for Parametric and Nonparametric Probabilistic Models," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 108-118, Jan. 2025, doi: 10.1109/TVCG.2024.3456393. [2] D. Wang et al., "A High-Quality Workflow for Multi-Resolution Scientific Data Reduction and Visualization," in SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1-18, doi: 10.1109/SC41406.2024.00091. [3] S. Saklani et al., "Uncertainty-Informed Volume Visualization using Implicit Neural Representation," *2024 IEEE Workshop on Uncertainty Visualization: Applications, Techniques, Software, and Decision Frameworks*, FL, USA, 2024, pp. 62-72, doi: 10.1109/UncertaintyVisualization63963.2024.00013.