

|Q>COGNITIVE

The Shape of Your Data

LUCA CANDELORI & NAG SANTHANAM



"I hold it equally impossible to know the parts without knowing the whole and to know the whole without knowing the parts in detail."

— Blaise Pascal, 17th century mathematician, physicist and philosopher

Beyond Repair: Learning the Whole

Data quality problems have long been treated as engineering challenges. They have been addressed through pre-processing, managed with heuristics, and accepted as a cost of doing business. This was a rational approach given the available tools. However, the real nature of the problem is one of geometry, not engineering. Real data do not fill the entire space spanned by input features. Instead, they cluster on lower-dimensional structures within that space. This is the *manifold hypothesis* and is the basis of every machine learning model in use today. Understanding the global shape of the data therefore provides the foundation for faithful data repair. Geometry holds the key, and until now, no practical method has captured it.

Firms invest resources in data engineering because the business and financial consequences of incomplete data can be significant. For example, in finance, forecasting prices and volatilities, detecting fraud, estimating credit risk, and extracting signals from alternative data sets are among the highest-value analytical tasks. Overwhelmingly, these are *tabular inference* tasks operating on structured data to forecast variables that drive investment, risk, lending, and operational decisions. The datasets involved are large and heterogeneous, spanning from traditional market and fundamental data to a rapidly expanding universe of alternative data. However, every prediction is limited by the quality of the data the model is trained on.

Our Quantum Cognition Machine Learning (QCML) based platform for imputation and anomaly detection tackles these problems at an architectural level. Unlike current methods that estimate missing values locally, our approach learns a **quantum geometric** model of the data, and performs imputation by choosing values that best fit the geometrical constraints of the data manifold.

Why quantum geometry? Simply put, it is the ideal framework for modeling data manifolds. In physics, quantum geometry describes manifolds as emergent properties arising from fundamental, discrete matrix dynamics, reconciling the quantized nature of space-time found at the microscopic scale with the smooth structures and properties that emerge at the macroscopic scale [8]. This is exactly how we think about data. Out of a finite collection of individual data points, larger patterns and smooth features emerge at scale: the shape of your data.

The QCML architecture mimics Right Brain Intelligence (RBI), embodying **Pascal's observation that we cannot know the parts without the whole**. Projecting incomplete records onto a learned quantum manifold improves imputation accuracy. A missing feature is not treated as a local gap to be patched; it is a coordinate to be inferred from the data's geometric structure. Data points that stray significantly from the manifold are identified as anomalies. In this sense, our methodology is a close relative of **energy-based models** (EBMs), which have recently been identified as the key ingredient to building true AI reasoning agents [10]. Like an EBM, we carve a model for the data manifold out of the energy landscape, where the actual data lies within low energy valleys while anomalies or other incoherent inputs lie on high energy mountain tops. Unlike EBMs, our model provides a latent representation of data into quantum states, able to model uncertainty and ambiguity about the exact position of an input within the energy landscape. This latent, quantum representation allows us to seamlessly handle continuous and categorical data within a unified architecture: the quantum state captures the *gestalt* of the individual data point, perhaps unable to reconstruct every single detail of the original, but able to capture the salient patterns that are relevant for imputation, forecasting, or any other downstream task.

Our framework provides reliable imputations even as data gaps and dimensionality increase. Additional features help define the manifold more precisely, increasing reliability in high-dimensional regimes where conventional methods degrade. Resource requirements scale linearly rather than exponentially, making QCML scalable to enterprise datasets and infrastructure. The result is not an incremental improvement over existing methods. It is a geometric foundation to tackle data problems that have so far been addressed with the wrong tools.

"Geometry is not true, it is advantageous."

— Henri Poincaré, mathematician and theoretical physicist

Geometry & Energy: The Key Ingredients

Our point of departure for solving data imputation are energy-based models (EBMs), which have been gaining popularity with proponents like Yann LeCun [9]. An EBM assigns a scalar (the energy) to every possible input, representing a fitness score. Lower energy indicates that the input is consistent with the data, while higher energy indicates that the input is likely to lie outside the data manifold. Crucially, energy is not normalized like a probability distribution. The basic idea then is to turn missing value imputation into an energy optimization problem: from the given data vector, **we fill the missing coordinates with the values that result in the lowest possible energy configuration.**

Current state-of-the-art EBMs are implemented as neural networks that take a data point x as input and output a scalar energy value $E(x)$. This neural network can be a simple multilayer perceptron, a convolutional neural network for image data, or a transformer for sequential data. The loss function for an EBM is based on the principle of maximum likelihood estimation. The energy function $E(x)$ is used to define a probability distribution over the input space, via the Boltzmann distribution:

$$p(x) = \frac{e^{-E(x)}}{Z}$$

where the normalization constant $Z = \int e^{-E(x)} dx$ is typically called the 'partition function', by analogy with physics. The ideal loss function for an EBM is therefore the negative log-likelihood of the data under this distribution

$$-\log p(x) = E(x) + \log Z.$$

Minimization of this loss would ensure that the EBM assigns high-probability (low energy) to data points and low probability (high energy) to all the other possible input points. However, here it lies one of the most serious flaws of EBMs: the partition function Z requires integration of a more-or-less arbitrary function over the entire input space. Since the input space is infinite, and no particular constraint is imposed on the neural network to ensure convergence of the integral, this becomes numerically intractable.

For this reason, EBMs have to abandon direct maximum likelihood estimation and instead rely on alternative training techniques, such as contrastive divergence or score matching, which are notoriously difficult to implement and scale effectively. Tabular data in particular represents unique challenges for EBMs, due to the mix of continuous and categorical variables, which require different approaches for generation and sampling. To this day, only a handful of experimental EBMs (such as TabEBM [11]) have been proposed for tabular data, and are mostly confined to the academic literature.

Quantum Geometry holds the keys to unleash the true potential of energy-based modeling at scale, especially in the realm of tabular data. Quantum Geometry was initially developed in theoretical physics to provide a quantum theory of gravity [8], and more recently it has been re-discovered at Qognitive, Inc. within the framework of QCML [6]. In this approach, each point on the data manifold is represented as a quantum state, a vector in a finite-dimensional Hilbert space. Each feature or variable is an observable, an operator in that space. Both the quantum state encoding and the observables are learned jointly through a training process, producing a new representation of the entire dataset in which every data point, every feature, and every relationship between them is encoded in a single coherent structure. The result is not a patchwork of neighborhood approximations. It is one global representation that captures the full shape of the data manifold all at once [2, 3].

But quantum states do not behave like abstract points. They are not confined to an infinitesimal, dimensionless location in space, but rather represent a dense cloud of possible locations that can all be held concurrently in quantum superposition. This allows for flexibility in data modeling [1, 5], with built-in capabilities for uncertainty quantification and better generalization [4, 7], much like a variational auto-encoder can achieve by encoding points as probability density functions. But this cloud of uncertainty needs to be controlled, or otherwise the states will spread

out uniformly over input space. Quantum Geometry achieves this by forcing the states into an energy well that is carved around the data manifold, and requiring the energy of a state to be proportional to the squared distance from the data manifold.

This physically principled architecture has two fundamental advantages over generic EBMs:

- The partition function integral $Z = \int_x \exp(-E(x)) dx$ **converges** to an actual finite value, since $\exp(-E(x))$ approaches $\exp(-\|x\|^2/2)$ for large $\|x\|$, the unnormalized density function of an isotropic Gaussian. This opens up numerical approaches to maximum likelihood estimation in low dimension, and effective sampling strategies in high dimension, that are not available to EBMs based on generic neural network architectures.
- Each data point x has a **latent representation** $\psi(x)$ consisting of its associated quantum state. Measurement of this state allows reconstruction and sampling without expensive and brittle Markov-Chain Monte-Carlo techniques, incorporating the key property of variational autoencoders and flow matching models into the EBM framework.

We illustrate the power of this new approach in Figure 1, where we plot the Boltzmann probability density function $p(x) = \exp(-E(x))/Z$ learned by our model on some common two-dimensional benchmarks for EBMs.

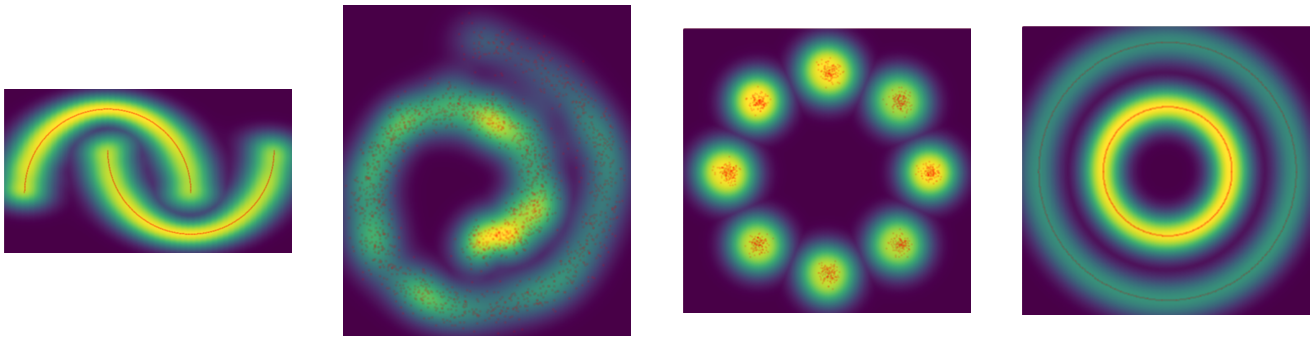


Figure 1: Probability densities generated by our model for some common 2D benchmarks. The training data is shown in red (and solid green for the concentric circles)

All these examples only required a Hilbert space dimension of sixteen, equivalent to 4 qubits in the language of quantum computing. Every example only took a few seconds of training on a laptop, using standard auto differentiation libraries such as PyTorch. In internal testing, we were able to train on genetics datasets with up to 20 thousand features with minimal resources. Our approach is effective and scalable, overcoming much of the training instability of current state-of-the-art EBMs.

A note on the Hilbert space dimension

A natural question is how a high-dimensional dataset can be faithfully represented by a small Hilbert space. The answer lies in the concentration of measure phenomenon discussed in the previous section. Because real financial data concentrates on a low-dimensional manifold, the intrinsic dimensionality of the dataset is far lower than the number of features. The Hilbert space dimension N acts as a magnification factor — larger values resolve finer details of the manifold, while smaller values produce a coarser but still faithful representation. Qognitive’s published research demonstrates that for a wide range of financial and other real-world datasets, N between 8 and 32 is sufficient to capture the structure that matters for downstream analytical tasks.

Armed with an accurate quantum geometric model for the data manifold, **missing data imputation becomes a simple energy optimization problem**: given an input x with possibly missing coordinates, we find the quantum state $\psi(x)$ that minimizes the energy of x ,

$$\psi(x) = \operatorname{argmin}_{\psi \in \mathbb{CP}^{N-1}} E(x),$$

and then measure the expected position of the state $\psi(x)$ (the ‘center’ of the quantum point cloud) to obtain the imputation $\hat{x}$. We demonstrate this method in Figure 2 and compare it to other popular imputation methods.

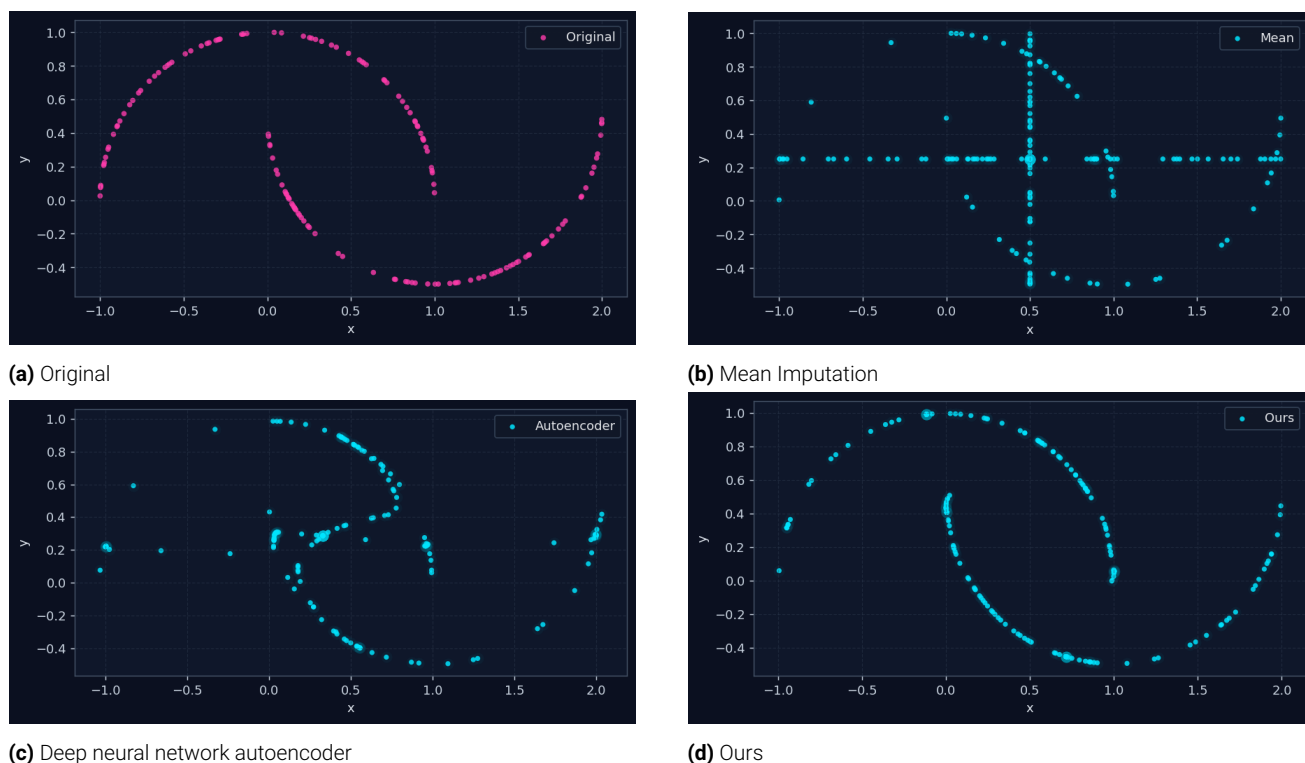


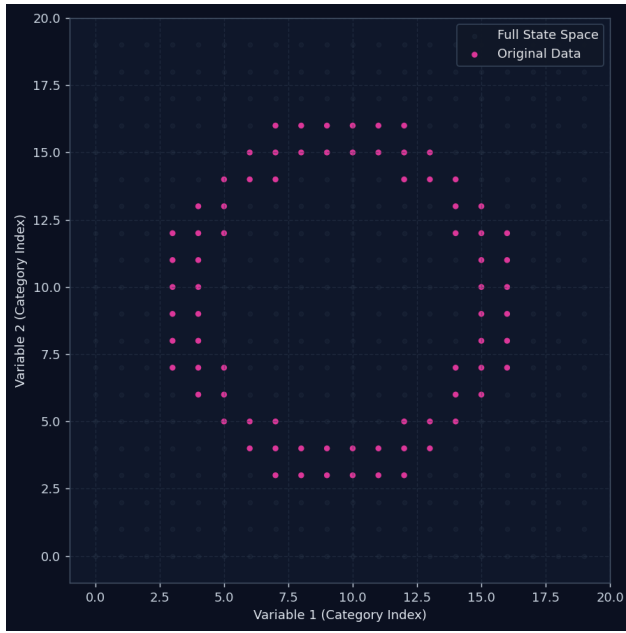
Figure 2: Comparison of different methods for missing value imputation. In this test, 50% of the coordinates were removed from the original half-moons (a), and then imputed using several methods.

The visual comparison between our imputation method and a deep neural network autoencoder illustrates the key difference between the energy-based and probabilistic approaches. A deep autoencoder will minimize the reconstruction mean-squared error, and when presented with uncertainty it will attempt to interpolate between the two arcs of the half-moons, completely disregarding the overall geometry of the data manifold. Our quantum geometry model, on the other hand, is able to handle multi-modal outputs and make an imputed value snap to either arc of the half-moon. The overall macro geometric features of the datasets are therefore preserved, ensuring higher fidelity and lower data-drift between the original and the imputed data.

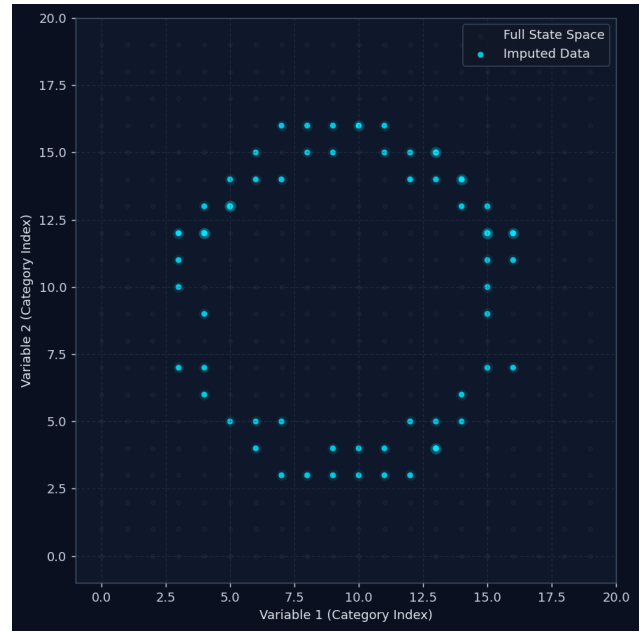
What this means in practice

QCML encodes your entire dataset as a learned geometric object. A data point with missing values is represented as a partially specified location in that geometry. Recovering the missing values is then a matter of finding the point on the learned manifold that is most consistent with the observed coordinates — a geometric projection rather than a statistical estimate.

In tabular data, categorical variables (such as industry category, country code, bond rating, etc.) are as prevalent and important as continuous variables. However, current EBM architectures based on deep neural networks cannot easily switch between a continuous input space and a discrete one. For example, sampling techniques like Langevin dynamics, which are used in contrastive divergence, cannot easily be adapted to handle categorical variables. Our quantum geometric approach, on the other hand, can be easily adapted to handle categorical data, simply by replacing the notion of ‘distance’ between data points from the Euclidean distance to an information-theoretic analog, such as cross-entropy. In this way, a data set consisting of a mix of categorical and continuous features can be handled by choosing the appropriate notion of distance for each feature. We illustrate how our missing value imputation method handles categorical data in Figure 3.



(a)



(b)

Figure 3: Benchmark dataset for discrete input space consisting of 2 categorical variables of 20 inputs each. (a) The dataset is in blue, and is arranged as an octagonal ring. 20% of the coordinates are removed for testing. (b) Imputed values using our method.

In addition to being able to capture the global geometry of the data manifold, there are other practical considerations that make our method appealing:

- *Dimensionality Scaling:* Classical imputation methods, such as MICE, require computational resources that grow exponentially with the number of features. QCML's resource requirements grow only linearly. This is a direct consequence of the quantum geometric representation — the global structure of even a very high-dimensional dataset can be captured in a finite-dimensional Hilbert space of modest size, typically between 8 and 32 dimensions. The curse of dimensionality that cripples local methods does not apply.
- *Natural Handling of Multicollinearity:* Highly correlated features are endemic in financial data, but do not represent a problem to be managed in QCML. They are simply reflected in the geometry: correlated features occupy nearby directions in the learned Hilbert space. There is no regression model to destabilize, no coefficient to inflate, no per-variable tuning required. The correlation structure of the data is encoded in the geometry automatically.
- *Multi-modal data:* Financial datasets are rarely uniform in their feature types — they combine continuous prices, categorical classifications, binary flags, ordinal ratings, and sparse alternative data signals. QCML handles mixed data types natively, without requiring specialized pre-processing or separate models for different feature types. The quantum geometric representation accommodates heterogeneous data within a single unified framework.

As Poincaré observed, **geometry is valuable not because it is literally true, but because of its advantages**. Our approach, grounded in QCML and Quantum Geometry, applies this insight directly to data. By modeling features using non-commutative matrix configurations, we are able to reconcile the discrete nature of data with the continuous geometrical abstraction of the manifold hypothesis. We do not attempt to model the data manifold explicitly, as other methods do, but rather let the manifold manifest itself as an emergent property of the quantum geometry.

Conclusion

QCML solves a longstanding problem by bringing together two ingredients that, separately, have each fallen short. Quantum Geometry provides a principled, tractable representation of the data manifold as a global structure rather than a patchwork of local approximations. Energy-based modeling provides the optimization framework that turns imputation into a coherent inference problem: find the configuration most consistent with the learned structure, not the value most probable given nearby observations. Together, they produce something neither achieves alone. The practical consequences follow directly from the architecture. Mixed data types, continuous and categorical alike, are handled within a single unified representation without requiring separate models or specialized pre-processing. High-dimensional datasets, far from posing a challenge, help define the manifold more precisely, turning the curse of dimensionality into an asset. Because the global structure of even a very high-dimensional dataset can be captured in a finite-dimensional Hilbert space of modest size, resource requirements scale linearly rather than exponentially, making QCML deployable at the scale of enterprise data infrastructure. In this white paper we focused on illustrating our basic principles, deliberately avoiding extensive benchmarking reporting. At the end of the day, what matters is your data, and we are ready to work with it. Our framework is a demonstration of Poincaré's idea applied to one of the most persistent practical problems in data science. The shape of your data was always there. We now have a method capable of representing it.

References

- [1] Ryan Samson, Adrian Banner, Luca Candelori, Sebastien Cottrell, Tiziana Di Matteo, Paul Duchnowski, Vahagn Kirakosyan, Jose Marques, Kharen Musaelian, Stefano Pasquali, Ryan Stever, and Dario Villani. Supervised similarity for firm linkages. October 2025.
- [2] Alexander G. Abanov, Luca Candelori, Harold C. Steinacker, Martin T. Wells, Jerome R. Busemeyer, Cameron J. Hogan, Vahagn Kirakosyan, Nicola Marzari, Sunil Pinnamaneni, Dario Villani, Mengjia Xu, and Kharen Musaelian. Quantum geometry of data. July 2025.
- [3] Luca Candelori, Alexander G. Abanov, Jeffrey Berger, Cameron J. Hogan, Vahagn Kirakosyan, Kharen Musaelian, Ryan Samson, James E. T. Smith, Dario Villani, Martin T. Wells, and Mengjia Xu. Robust estimation of the intrinsic dimension of data sets with quantum cognition machine learning. February 2025.
- [4] Joshua Rosaler, Luca Candelori, Vahagn Kirakosyan, Kharen Musaelian, Ryan Samson, Martin T. Wells, Dhagash Mehta, and Stefano Pasquali. Supervised similarity for high-yield corporate bonds with quantum cognition machine learning. February 2025.
- [5] Ryan Samson, Jeffrey Berger, Luca Candelori, Vahagn Kirakosyan, Kharen Musaelian, and Dario Villani. Quantum cognition machine learning: Financial forecasting. September 2024.
- [6] Kharen Musaelian, Alexander Abanov, Jeffrey Berger, Luca Candelori, Vahagn Kirakosyan, Ryan Samson, James Smith, and Dario Villani. Quantum cognition machine learning: AI needs quantum. March 2024.
- [7] Giuseppe Di Caro, Vahagn Kirakosyan, Alexander G. Abanov, Luca Candelori, Nadine Hartmann, Ernest T. Lam, Kharen Musaelian, Ryan Samson, Martin T. Wells, Richard J. Wenstrup, and Mengjia Xu. Quantum cognition machine learning for forecasting chromosomal instability.
- [8] Harold Steinacker. Quantum Geometry, Matrix Theory, and Gravity *Cambridge University Press*, 2024
- [9] Yann LeCun. A Tutorial on Energy-based Learning in "Predicting Structured Data", G. Bakir, T. Hofman, B. Scholkopf, A. Smola, B. Taskar (eds) MIT Press, 2006.
- [10] Eve Bodnia. The Art of Knowing
<https://logicalintelligence.com/blog/energy-based-ai-vision-the-art-of-knowing>
- [11] A. Margeloiu, et al. TabEBM: a tabular data augmentation method with distinct class-specific energy-based models, *NeurIPS* 2024