

---

# Shining light on data

---

**Akshat Kumar**  
Department of Mathematics  
Clarkson University  
Potsdam, NY 13699  
akumar@clarkson.edu

**Mohan Sarovar**  
Sandia National Labs  
Livermore, CA 94550  
mnsarov@sandia.gov

## Abstract

Experimental sciences have come to depend heavily on our ability to organize, interpret and analyze high-dimensional datasets produced from observations of a large number of variables governed by natural processes. Natural laws, conservation principles, and dynamical structure introduce intricate inter-dependencies among these observed variables, which in turn yield geometric structure, with fewer degrees of freedom, on the dataset. We show how fine-scale features of this structure in data can be extracted from *discrete* approximations to quantum mechanical processes given by data-driven graph Laplacians and localized wavepackets. This leads to a novel, yet natural uncertainty principle for data analysis induced by limited data. We illustrate some applications to learning with algorithms on several model examples and real-world datasets.

Nature is complex, yet organized – this basic facet has become a cornerstone for how we analyze and interpret vast amounts of data ranging from the biological, physical, geological, meteorological, all the way to the astronomical. In fact, we owe to this, nearly all of the significant advances in data analysis disciplines in the past two decades, including signals processing [8, 22] and machine learning [3, 5, 11], where tractable models capture, interpret and reproduce complex natural data. We can turn this observation on its head and ask: are there certain natural processes that are fundamental to understanding the structure in complex, yet organized data?

We propose that, much like how quantum mechanics models nature at fine scales, the fine-scaled resolution of the organization and structure of data is also best characterized using quantum mechanical processes. Consider the following general scenario: an experimentalist makes a sequence of measurements, each consisting of a value for  $D$  variables. Each measurement is thus a point in  $\mathbb{R}^D$ . Of course, measurements coming from nature are bound by physical laws, so in fact one must imagine that the *true* ambient space is some other, nonlinear structure  $\mathcal{M}$ , residing in  $\mathbb{R}^D$ . Moreover, if the experiment is governed by a number of parameters, then these give  $\mathcal{M}$  its local degrees of freedom around any given measurement. This principle is in fact often studied and known as the *manifold hypothesis* (MH) [9]: measurements of arbitrarily high dimensions (residing in  $\mathbb{R}^D$ ) arising from natural processes are confined to low-dimensional manifolds (*i.e.*,  $\mathcal{M} \subset \mathbb{R}^D$  with  $\dim \mathcal{M} \ll D$ ). Due to the MH, the experimentalist’s aim of connecting model parameters to observations through the relationships among measurements is just the analysis of the organizational structure, or *geometry* of data. In the natural sciences, such analyses are commonly performed through dimensionality reduction, classification, *etc.* using techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), or variants of Laplacian eigenmaps, that indirectly probe the geometry of data, *e.g.*, [7, 19, 14]. In addition, it has recently been appreciated that understanding and exploiting the structure of data, and sometimes reorganizing/reparameterizing it, has advantages in learning frameworks such as convolutional neural networks, *e.g.*, [24, 4]. While the state-of-the-art in learning the structure and organization of data is founded on principles of Markovian dynamics [2, 6] – primarily because of their *local* nature that forms an accessible link between random walks and Markov processes – in practice, these are suitable only for accessing coarse features of the data due to

the slowly varying nature of the steady-states that form the basis of these techniques. We propose that to perform a fine-scaled data analysis, it is advantageous to shift the paradigm to quantum mechanics, which provides a natural way to formulate dynamics on the data that respects limits to resolution of its geometric structure set by data sampling, through the introduction of *phase-space uncertainty*.

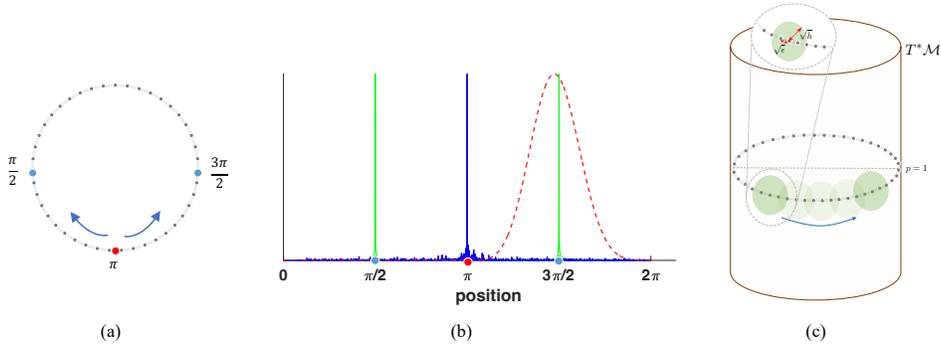
```

1: Inputs:  $X_N = \{v_1, \dots, v_N\}, v^*, \epsilon > 0, \alpha \geq 1, t > 0$ 
2: Output: Propagated state  $[\psi_h^\zeta](t)$ 
3: procedure PROPAGATE
4:   Compute  $[T_\epsilon]_{i,j} = k(\|v_i - v_j\|^2/\epsilon)$ 
5:   Compute diagonal matrix  $[D_\epsilon]_{i,i} = \sum_{j=1}^N [T_\epsilon]_{i,j}$ 
6:   Compute  $\Delta_{\epsilon,N} = \frac{4(I_N - D_\epsilon^{-1}T_\epsilon)}{\epsilon}$ 
7:    $U_{\epsilon,N}^t \leftarrow \exp(-it\sqrt{\Delta_{\epsilon,N}})$ 
8:   Set  $h = \epsilon^{\frac{1}{2+\alpha}}$ 
9:   Set  $p_0 = v_j - v^*$  for  $v_j$  closest to point  $v^* \Rightarrow \zeta = (v^*, p_0/\|p_0\|)$ 
10:  while  $1 \leq \ell \leq N$  do
11:     $[\psi_h^\zeta]_\ell \leftarrow e^{-\|v_\ell - v^*\|^2/2h} e^{\frac{i}{h}(v_\ell - v^*)^\top p_0/\|p_0\|}$ 
12:  return  $[\psi_h^\zeta](t) = U_{\epsilon,N}^t[\psi_h^\zeta]$ 

```

**Algorithm 1:** Pseudocode for algorithm that performs data-driven propagation of coherent states on a data graph. The inputs are the dataset  $X_N \subset \mathbb{R}^D$ , an initial data point to propagate from,  $v^*$ , parameters  $\epsilon > 0$ ,  $\alpha \geq 1$ , and a time to propagate for,  $t > 0$ . Lines 4-7 construct a data-driven quantum propagator,  $U_{\epsilon,N}^t$  using a graph Laplacian  $\Delta_{\epsilon,N}$  computed from the data.  $k(\cdot)$  (line 4) is an exponentially decaying function of the argument and  $I_N$  (line 6) is the  $N \times N$  identity matrix. Lines 10-11 form the  $N \times 1$  vector that approximates a coherent state at phase space point  $\zeta := (v^*, p_0/\|p_0\|)$ . A more detailed description of this algorithm, and extensions of it, are presented in [17, 16].

We realize this proposal through the simulation of quantum dynamics on data as sketched in Algorithm 1. The fundamental relationship connecting this dynamics to the structure of data is given by a *discrete* quantum-classical correspondence principle (QCC), which is established in Ref. [15] and summarized in [17, 16] and applies when the MH holds for the data in  $\mathbb{R}^D$ , confined to  $\mathcal{M}$  a smooth, compact, boundaryless submanifold. This relationship is supported by the traditional form of the QCC, which physically in our geometric setting connects the propagation of wavefunctions of photons (quantized excitations) on curved space ( $\mathcal{M}$ ) with geodesic propagation of light rays [12]. Such a propagation is illustrated in green in Figure 2. More formally, it is a well-known result in microlocal analysis [23, 25] that the trajectory of an impulse  $\delta_{x^*}$  at  $x^* \in \mathcal{M}$  with respect to the propagation  $U^t[\delta_{x^*}] := e^{-it\sqrt{\Delta}}[\delta_{x^*}]$ , where  $\Delta$  is the Laplace-Beltrami operator of  $\mathcal{M}$ , has singular support along geodesics in all directions at distance  $|t| < T$  from  $x^*$ , for a bounded time  $T$ . This is the continuum QCC in action: it connects the quantum mechanical propagation  $U^t[\delta_{x^*}]$  generated by the quantization  $\sqrt{\Delta}$  of the classical Hamiltonian (purely kinetic energy), to the geodesic flow, by moving the energy concentration of the singular state  $\delta_{x^*}$  to time  $t$  along the classical flow in directions of the state's asymptotically large momenta (*i.e.*, the frequencies of its wavefunction) [23].



**Figure 2:** (a) Data is  $N = 2500$  regularly spaced samples from the unit circle:  $x_j = (\cos \theta_j, \sin \theta_j)$  for  $\theta_j \in [0, 2\pi)$ . (b) We observe the squared magnitude of: (green) an undirected optical ray from  $x^* = \pi$  at time  $t = \frac{\pi}{2}$ , *i.e.*,  $|U^t[\delta_\pi]|^2$ , and (blue) the corresponding data-driven quantum propagation, *i.e.*,  $|U_{\epsilon,N}^t[\delta_\pi]|^2$ . While the former is concentrated at  $x_- = \pi - t = -\pi/2$ ,  $x_+ = \pi + t = 3\pi/2$ , which are points that minimize  $|d_g(x^*, x) - t|$  ( $d_g$  is geodesic distance), the latter state bears no resemblance to geodesic propagation, with significant attenuation over the green signal, and an undiminished peak at the source,  $x^*$ . We also show (dashed red) the data-driven propagation of a coherent state centered at  $x^*$  with uncertainty parameter  $h$ , *i.e.*,  $|U_{\epsilon,N}^t[\psi_h^\zeta]|^2$ , with  $\zeta = (\pi, 1)$ . This state is approximately centered at the point  $x_+$ , and in fact, its expected position,  $\langle x \rangle$ , satisfies  $|\langle x \rangle - x_+| \leq h$ . Note that all curves are normalized to have the same maximum value in order to plot them on the same scale. (c) Depiction of coherent state propagation in *phase space*,  $T^*\mathcal{M}$ . The zoom-in schematically shows  $\sqrt{\epsilon}$  and  $\sqrt{h}$ , the data-determined scale and uncertainty parameters, respectively.

Coming back to quantum dynamics on the data, we couple the above continuum QCC with the now well-known result that, assuming the MH, we can approximate from  $N$  measurement samples

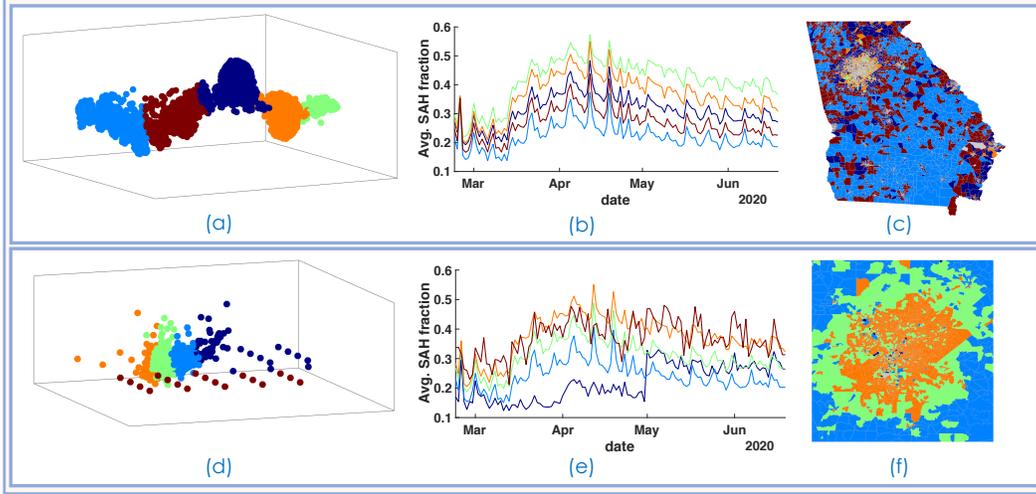
$X_N := \{v_1, \dots, v_N\} \subset \mathbb{R}^D$  the Laplace-Beltrami operator of their underlying manifold through the graph Laplacian  $\Delta_{\epsilon, N}$  of a  $\sqrt{\epsilon}$ -nearest-neighbour ( $\sqrt{\epsilon}$ -n.n.) graph on  $X_N$  with high probability (*w.h.p.*) [2, 13, 6]. Our recent work shows that a data-driven approximation to the quantum mechanical propagator  $U^t$  is given *w.h.p.*, by the  $N \times N$  matrix  $U_{\epsilon, N}^t := e^{-it\sqrt{\Delta_{\epsilon, N}}}$ . Given this construction, it is tempting to reproduce the geodesic propagation of light rays given through  $U^t[\delta_{x^*}]$  with  $U_{\epsilon, N}^t[\delta_{x^*}]$ . However, as shown in the blue curve in Figure 2(b), this data-driven propagation bears no resemblance to the continuum signal  $U^t[\delta_{x^*}]$  (green curve). This is because  $\Delta_{\epsilon, N}$  is defined through a Markov process discretized to  $\sqrt{\epsilon}$ -balls on the manifold and based on the uncertainty principle, our recent work shows that this Markov process acts like a low-pass filter by attenuating spatial frequencies with magnitude  $\gtrsim 1/\sqrt{\epsilon}$ . Therefore,  $\Delta_{\epsilon, N}$  approximates a scalar operator at bandwidth  $\gtrsim 1/\sqrt{\epsilon}$  and by virtue of being its spectral function, so does  $U_{\epsilon, N}^t$ . This explains the concentration of the blue curve in Figure 2(b) about  $x^*$  and predicts that this behaviour will persist even in the limit  $N \rightarrow \infty$  while  $\epsilon > 0$  remains bounded away from zero.

Understanding this issue is the fulcrum of our discrete QCC, which recovers light rays from data-driven simulation of quantum dynamics. As we've seen, the primary obstruction to approximating the continuum QCC with  $U_{\epsilon, N}^t[\delta_{x^*}]$  is the over-concentration of the initial state  $\delta_{x^*}$ , which gives frequency content above the data-determined threshold  $\sim 1/\sqrt{\epsilon}$ . To control this, we introduce a state  $\psi_h \in C^\infty(\mathcal{M} \times (0, 1]_h)$  surrogate to  $\delta_{x^*}$ , whose bandwidth we can control to scale as  $\sim 1/h$  and whose continuum propagation  $U^t[\psi_h]$  follows the light ray emanating from  $x^*$  until time  $t$ , to within a  $\sqrt{h}$ -ball throughout the propagation. This is satisfied by *coherent states*, which in the continuum have position-space representation:  $\psi_h^\zeta(v) := e^{-\frac{i}{2h}\|v-x^*\|^2} e^{-\frac{i}{h}\langle x^*-v, p \rangle}$  with position  $x^* \in \mathcal{M}$  and momentum  $p \in T_{x^*}^*\mathcal{M}$ . Rooted in the above arguments, it is established in Ref. [15] and further discussed in [17, 16] that in order for the data-driven propagation  $U_{\epsilon, N}^t[\psi_h^\zeta]$  to approximate  $U^t[\psi_h^\zeta]$  (*w.h.p.*) and therefore, to follow the light rays on data, the bandwidth  $1/h$  must be kept  $\ll 1/\sqrt{\epsilon}$ . Thus, the *uncertainty parameter*  $0 < h \leq 1$  governing the quantum mechanical properties of coherent states gives us the desired control to establish our discrete QCC.

In Figure 2(b) the dashed red line shows the result of propagating a coherent state using a data-driven propagator for the circle example. By incorporating the intrinsic uncertainty induced by finite sampling into the formulation of quantum dynamics, we recover accurate propagation. In Figure 2(c) we also depict coherent state propagation in phase space for this example, and graphically show the relationship between  $\epsilon$  and  $h$ .

In [15, 17, 16], convergence results have been derived to establish the following: given  $N$  data samples,  $X_N$ , from a smooth density on  $\mathcal{M}$ , the data-driven finite-dimensional matrix propagation  $[\psi_h^\zeta](t) := U_{\epsilon, N}^t[\psi_h^\zeta]$  returned by Algorithm 1, with  $h \propto \epsilon^{\frac{1}{2+\alpha}}$  for  $\alpha \geq 1$  agrees with  $U^t[\psi_h^\zeta]$  up to uniform error  $O(h)$  *w.h.p.*, provided  $h \gtrsim N^{-\frac{1}{\gamma}}$  for  $\gamma > 0$  a constant depending only on  $\alpha$  and  $\dim \mathcal{M}$ . Thus,  $[\psi_h^\zeta](t)$  traverses within an  $O(\sqrt{h})$  radius of the geodesic beam emanating from  $v^*$  in the direction  $p$  to time  $t$ . Moreover, the point  $\bar{x}_t := \|\psi_h^\zeta(t)\|^{-2} \sum_{j=1}^N (v_{j,1} \dots v_{j,D}) |\psi_h^\zeta(t)|^2$ , which is the expected position of the propagated state, is *w.h.p.* within geodesic distance  $O(h)$  to the point  $x_t^*$ , that is geodesic distance  $t$  from the initial position  $v^*$  in the direction  $p$ . While this expected value of the position coordinate,  $\bar{x}_t$ , is the best estimate of position along the geodesic path, due to the localization of the propagated state the maximum of the wavepacket distribution, we show that  $\hat{x}_t := \arg \max |\psi_h^\zeta(t)|^2$  is *w.h.p.* within geodesic distance  $O(\sqrt{h})$  from  $x_t^*$ .

The data-driven quantum dynamics formulated above enables estimation of intrinsic distances between points in a dataset: if  $\bar{x}_t(j; p_j) \in X_N$  denotes the data point closest to  $\bar{x}_t$  (or  $\hat{x}_t$ ) computed as above with respect to an initial state  $\psi_h^\zeta$  localized at  $\zeta := (v_j, p_j)$ , then  $t$  gives the propagation time of a coherent state following approximately a ray emanating in direction  $p_j$  from  $v_j$ . When the data is sampled from a smooth density on a smooth, compact, boundaryless manifold, we have established that  $t$  is *w.h.p.*, within  $O(h)$  of the geodesic distance. In fact, by using local PCA to define  $\psi_h^\zeta$  and hence,  $p_j$ , it is shown in [17, 16] that the dataset  $X_N$  can be *charted* with *geodesic polar coordinates* (GPC), or equivalently, normal coordinates. We pause to emphasize that the procedure we have described gives access to geodesics and GPC on a manifold, which are inherently described by non-linear dynamical equations, through linear, matrix computations. Computing such quantities, even when much more is known about the manifold, is generally computationally difficult since typical approximation methods are insufficient: polyhedral approximations are not a faithful model



**Figure 3:** Analysis of adherence to social distancing measures during the COVID-19 pandemic using the Social Distancing Metric dataset from SafeGraph Inc. [1]. This dataset is a collection of geolocation information from mobile devices in the United States, aggregated at the census block group (CBG) level and recorded daily for a period of over a year. As previously studied in Ref. [18], we compute the stay-at-home (SAH) fraction as a simple metric of adherence to social distancing from this data: for a given date, this is a measure of how curtailed mobility was within a CBG. To compare with Ref. [18] we consider data for the state of Georgia (GA) and limit it to the 117-day time period from February 23, 2020 to June 19, 2020, which provides a snapshot of mobility patterns during the first three months of the pandemic. After removing 17 CBGs with poor quality data, there are 5509 CBGs within the state. Therefore, our dataset  $X_N$  has  $N = 5509$  samples, each of dimension 117. **Top (a,b,c):** Embedding of  $X_N$  and clustering when the expected position  $\bar{x}_t$  is used to define points a geodesic distance  $t$  away. **Bottom (d,e,f):** Embedding and clustering when the maximum  $\bar{x}_t$  is used to define points a geodesic distance  $t$  away. (a) and (d) show 3-dimensional embeddings of  $X_N$  and clustering of the resulting points into five clusters using  $k$ -means clustering. (b) and (e) show the average SAH fraction time series for each cluster. In (b) we see a clear separation of clusters by their SAH behavior, and in (e) we have identified an anomalous SAH pattern (in purple) in one of the clusters. (c) and (f) color code the CBGs in GA according to the cluster they belong to. (c) shows a clear rural-urban divide in degree of social distancing behavior, and (f) is a magnification of the Atlanta metropolitan region, because many of the outlier CBGs identified are located in this area. We refer to [17, 16] for a detailed analysis, including parameter values and choice heuristics.

of manifolds with curvature restrictions [20] and even forward marching type approximations are known to be prone to failures [21]. Furthermore, our result establishes the first general convergence result for geodesics and GPC from data.

More broadly, the assignment of a data point  $\bar{x}_t(j, p_j) = v_k$  to a given data point  $v_j$  is, in itself, independent of further structural assumptions on the dataset. Even when the data  $X_N$  is not guaranteed to be sampled from a manifold, this defines a distance relationship between two points based on the *quantum walk*  $U_{\epsilon, N}^t[\psi_h^S]$ . Based on this distance relationship we can build an  $N \times N$  adjacency matrix,  $\mathcal{G}$ , for a graph  $\mathcal{X}_N$  on  $X_N$ , with elements  $\mathcal{G}_{j, k} = \mathcal{G}_{k, j} = t$ . Repeating this process for a collection of initial points  $v_j$  and time-steps,  $t_1, \dots, t_m$  populates this adjacency matrix, which captures a notion of distance between the data points in  $X_N$  given by the quantum propagation times of coherent states. We can perform embeddings of  $\mathcal{X}_N$  to achieve tasks such as recovering reduced-dimensional coordinates, clustering, classification, *etc.* We find that even the most classical embedding of  $\mathcal{X}_N$  in few dimensions, such as the Fruchterman-Reingold (FR) method of springs and electrostatic forces [10], recovers salient features of complex datasets. For example, Figure 3 shows examples of FR embedding into 3 dimensions followed by  $k$ -means clustering of a 117-dimensional dataset related to population mobility during the initial months of the COVID-19 pandemic. The embedding enabled by our discrete QCC and subsequent clustering is able to meaningfully separate geographic regions with differing levels of mobility as well as identify outlier patterns. On the same dataset, a state-of-the-art Markov-based approach was deployed in Ref.[18] and relied on analysis in 14 dimensions to identify poorer quality clusters (see Fig. 2 of Ref. [18]), and was also unable to identify the outlier patterns in mobility.

We have argued that, similar to the Fourier transform being fundamental for signal analysis, quantum propagation and sensing are fundamental for *structure* analysis in sampled data. The structure is carried by the dynamics of rays emitted on the dataset, which are *signals* generated by the quantum propagation  $U_{\epsilon, N}^t[\delta_{x^*}]$  that can be resolved with the aid of coherent states defined on the dataset. When the data satisfies the MH with a compact, smooth, boundaryless manifold, these signals reveal

the underlying geodesic flow that governs the data's organizational patterns. Our experiments indicate that these signals efficiently reveal underlying patterns and anomalies in even sparse datasets.

## References

- [1] Safegraph Inc. Social Distancing Metrics, 2021. URL <https://docs.safegraph.com/docs/social-distancing-metrics>.
- [2] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15:1373–1396, 2003.
- [3] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [4] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478 [cs, stat]*, 2021. URL <http://arxiv.org/abs/2104.13478>.
- [5] E. Candes, R. R. Coifman, A. Singer, and T. Strohmer. *Applied Harmonic Analysis, Massive Data Sets, Machine Learning, and Signal Processing*. 2019. Workshop Report from "Computational Harmonic Analysis and Data Science". Banff International Research Station.
- [6] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006.
- [7] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci.*, 102:7426–7431, 2005.
- [8] S. Damelin and W. Miller, Jr. *The Mathematics of Signal Processing*. Cambridge University Press, 2012.
- [9] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *J. Am. Math. Soc.*, 29:983–1049, 2016.
- [10] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21:1129–1164, 1991.
- [11] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio. Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, 121:9722–9758, 2021. doi: 10.1021/acs.chemrev.0c01195.
- [12] D. Gloge and D. Marcuse. Formal quantum theory of light rays. *JOSA*, 59:1629, 1969. doi: 10.1364/JOSA.59.001629.
- [13] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.
- [14] D. Kobak and P. Berens. The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 10(11):5416, 2019. doi: 10.1038/s41467-019-13056-x.
- [15] A. Kumar. On a quantum-classical correspondence: from graphs to manifolds, 2021. URL <https://arxiv.org/abs/2112.10748>. arXiv:2112.10748.
- [16] A. Kumar and M. Sarovar. Manifold learning via quantum dynamics, 2021. URL <https://arxiv.org/abs/2112.11161>. arXiv:2112.11161.
- [17] A. Kumar and M. Sarovar. Shining light on data: Geometric data analysis through quantum dynamics, In preparation.
- [18] R. Levin, D. L. Chao, E. A. Wenger, and J. L. Proctor. Insights into population behavior during the covid-19 pandemic from cell phone mobility data and manifold learning. *Nature Computational Science*, 1:588–597, 2021.
- [19] M. Maggioni. Geometry of data and biology. *Notices of the American Mathematical Society*, 62(10):1185–1188, 2015. doi: 10.1090/noti1289.

- [20] A. Petrunin. Polyhedral approximations of riemannian manifolds. *Turkish Journal of Mathematics*, 27(1):173–188, 2003.
- [21] G. Peyré, M. Péchaud, R. Keriven, and L. D. Cohen. Geodesic Methods in Computer Vision and Graphics. *Foundations and Trends in Computer Graphics and Vision*, 5:197, 2010.
- [22] J. Wright and Y. Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.
- [23] S. Zelditch. *Eigenfunctions of the Laplacian of Riemannian manifolds*. Am. Math. Soc., 2017.
- [24] W. Zhu, Q. Qiu, J. Huang, R. Calderbank, G. Sapiro, and I. Daubechies. Ldmnet: Low dimensional manifold regularized neural networks. *arXiv:1711.06246 [cs]*, 2017. URL <http://arxiv.org/abs/1711.06246>.
- [25] M. Zworski. *Semiclassical Analysis*. American Mathematical Society, 2012.

## Broad impact

In this work we develop a new framework for understanding structure within large datasets. These results will be useful for analyzing and extracting meaning from a broad range of datasets, with applications ranging from the analysis of statistical and experimental data to image analysis and to analysis of sensor and computer simulation data. The unique capability we have developed, to efficiently identify geodesic distances between data points, could be useful for performing heretofore infeasible analyses. As an example, we have shown in Figure 3 the capability of our methods to detect anomalous behavioral patterns in mobility during the COVID-19 pandemic, which surpasses the comparative state-of-the-art in literature; in the accompanying recent works, we have given further examples of pattern recognition on a variety of real-world datasets. As such, this could have tremendous societal benefit, since efficient analysis of large datasets is a ubiquitous task in almost every field of study. Negative societal impacts are also possible if the increased capabilities in data analysis are utilized for purposes such as surveillance. As with any fundamental data analysis method, the societal impact is eventually dictated by the way in which the method is applied.

## Checklist

1. For all authors...
  - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [\[Yes\]](#)
  - (b) Did you describe the limitations of your work? [\[No\]](#) We provide our assumptions and experiments on data that meet those assumptions and some that don’t. Longer discussion is in full-length works, while this is an extended abstract.
  - (c) Did you discuss any potential negative societal impacts of your work? [\[Yes\]](#)
  - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [\[Yes\]](#)
2. If you are including theoretical results...
  - (a) Did you state the full set of assumptions of all theoretical results? [\[Yes\]](#) We give only a flavor of our theoretical results and have included the necessary assumptions for them.
  - (b) Did you include complete proofs of all theoretical results? [\[No\]](#) This is an extended abstract; the complete proofs are given in the corresponding research papers.
3. If you ran experiments...
  - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [\[No\]](#) We provided a sample pseudocode and referred to the data we have used (in the real world example, we’ve provided a citation with the URL). We are happy to provide code and datasets on request.

- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No] These details are in the corresponding research papers and we are happy to provide them on request as well.
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- (a) If your work uses existing assets, did you cite the creators? [Yes]
  - (b) Did you mention the license of the assets? [No]
  - (c) Did you include any new assets either in the supplemental material or as a URL? [No]
  - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
- (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
  - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]