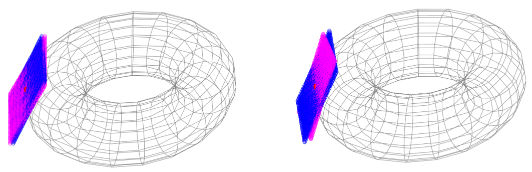
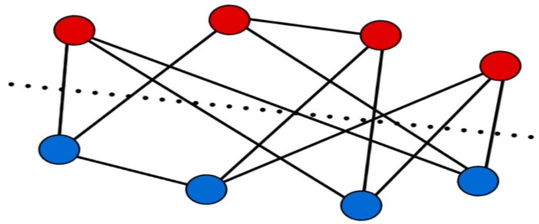
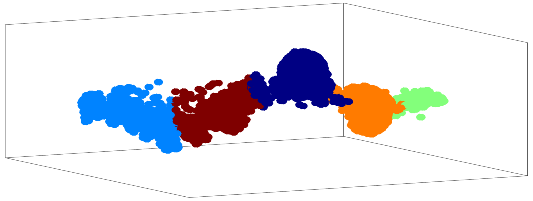
SAND2022-12946C

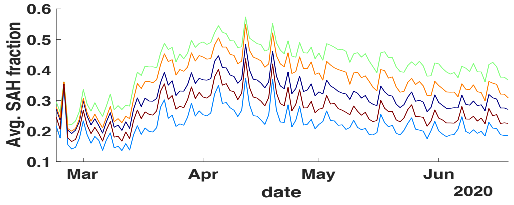
Sandia National Laboratories

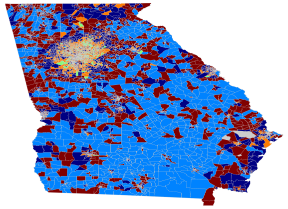
Geometric data analysis through quantum dynamics











Mohan Sarovar

Sandia National Laboratories, Livermore, CA

SIAM Mathematics of Data Science
September 2022

U.S. DEPARTMENT OF ENERGY NNSA

Manifold learning and its applications

- Manifold learning, or resolving geometric structure of data enables many tasks:
  - Visualization
  - Representation of data in reduced order coordinates
  - Classification, anomaly detection, image segmentation, autonomous driving, augmented reality, …

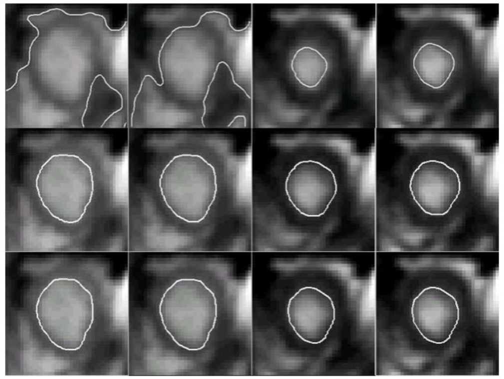
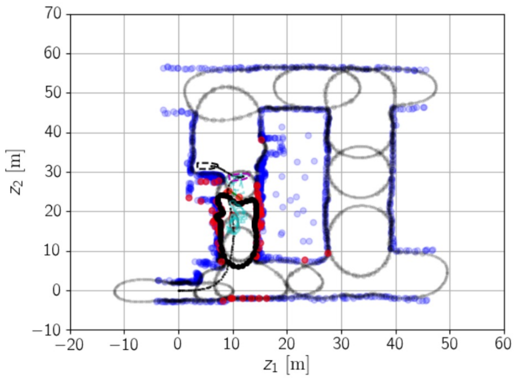
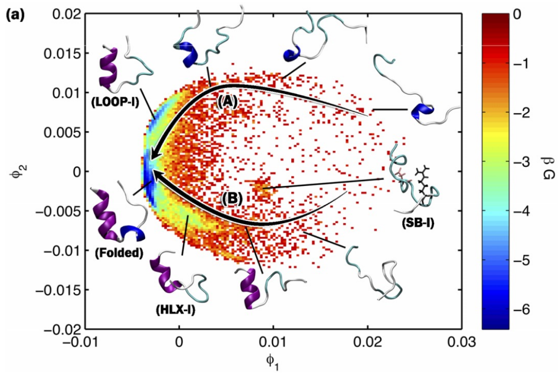


Image segmentation for medical imaging
Qilong Zhang, et al., *2006 IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, p. 1092.



Obstacle avoidance in autonomous driving
Diwale et al., https://infoscience.epfl.ch/record/265381?ln=en



Identification of protein folding pathways from molecular dynamics simulations
Kim et al., J. Chem. Phys. 142, 085101 (2015)

Manifold learning and its applications

- Manifold learning, or resolving geometric structure of data enables many tasks:
  - Visualization
  - Representation of data in reduced order coordinates
  - Classification, anomaly detection, image segmentation, autonomous driving, augmented reality, …

- **The manifold hypothesis**: "high dimensional data tend to lie in the vicinity of a low dimensional manifold".
  - *e.g.*, images, randomly generated image of *NxN* pixels will almost surely not correspond to a real world scene.
  - *e.g.*, data generated by a dynamical system will follow some equation of motion

Fefferman, Mitter, Narayanan. J. Am. Math. Soc., **29**, 983 (2016)

Continuum quantum dynamics on manifolds & geodesic flow

Consider: $\hat{H} = \sqrt{\Delta_g}$ ← Laplace-Beltrami operator for manifold

$|\psi_0\rangle = |\delta_x\rangle$

$|\psi_t\rangle = e^{it\sqrt{\Delta_g}}|\delta_x\rangle$ is a state that has singular support along **geodesics** in all directions

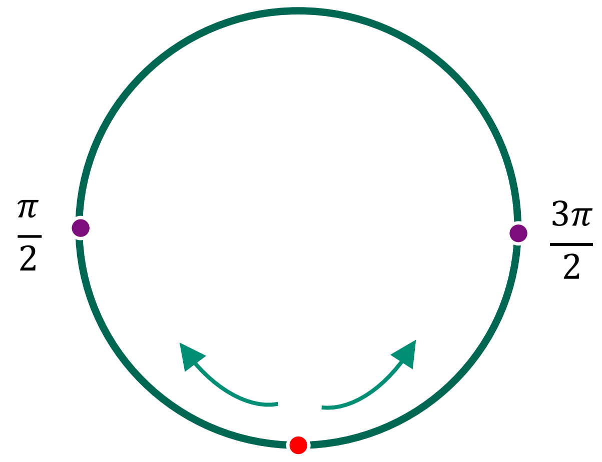
Continuum quantum dynamics on manifolds & geodesic flow

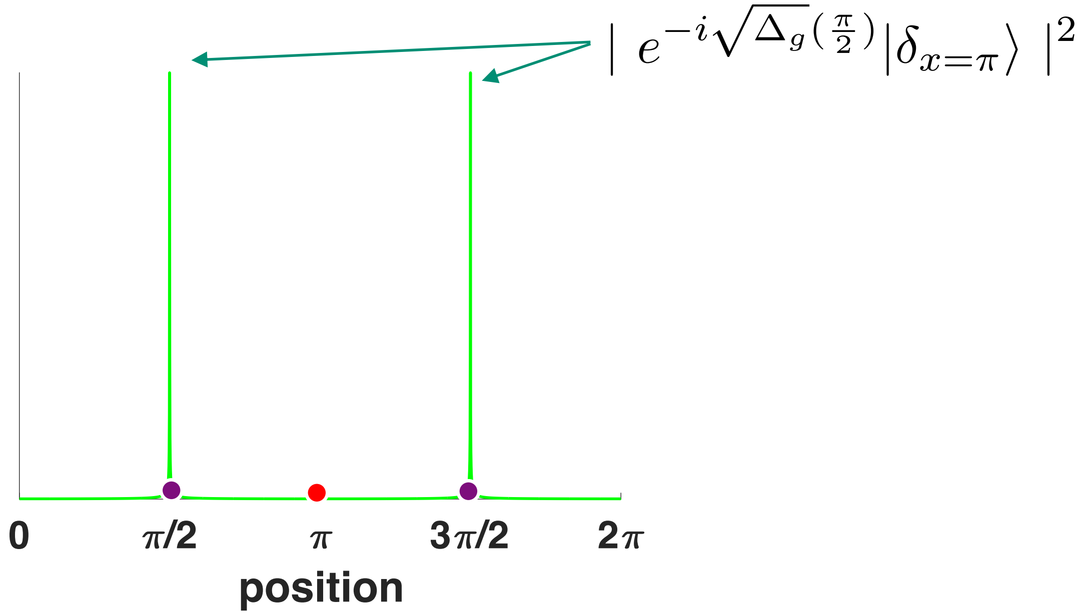
Consider: $\hat{H} = \sqrt{\Delta_g}$ ← Laplace-Beltrami operator for manifold

$|\psi_0\rangle = |\delta_x\rangle$

$|\psi_t\rangle = e^{it\sqrt{\Delta_g}}|\delta_x\rangle$ is a state that has singular support along **geodesics** in all directions

Example: circle





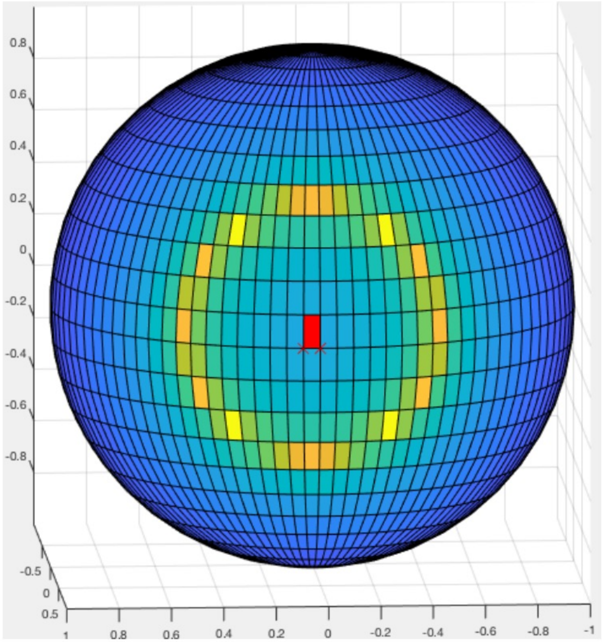
Continuum quantum dynamics on manifolds & geodesic flow

Consider: $\hat{H} = \sqrt{\Delta_g}$ ← Laplace-Beltrami operator for manifold

$|\psi_0\rangle = |\delta_x\rangle$

$|\psi_t\rangle = e^{it\sqrt{\Delta_g}}|\delta_x\rangle$ is a state that has singular support along **geodesics** in all directions

Example: sphere



Continuum quantum dynamics on manifolds & geodesic flow

Consider: $\hat{H} = \sqrt{\Delta_g}$ ← Laplace-Beltrami operator for manifold $\qquad |\psi_0\rangle = |\delta_x\rangle$

$|\psi_t\rangle = e^{it\sqrt{\Delta_g}}|\delta_x\rangle$ is a state that has singular support along **geodesics** in all directions

This statement can be understood in various ways

1. **Math**: Microlocal analysis – wavefront set associated to hyperbolic dynamics
2. **Physics**: Free motion of localized particle (photon) – light travels in straight lines
3. **Mathematical physics**:

$\hat{H} = \sqrt{\Delta_g}$ is quantization of kinetic energy/free motion $\qquad \sqrt{\sum_{i,j} g^{ij}(x) p_i p_j} = |\mathbf{p}|_g$

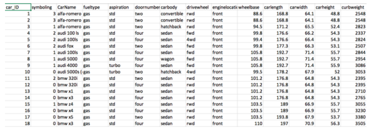
$|\psi_0\rangle = |\delta_x\rangle$ is a localized wavepacket with momentum in all directions

Can we exploit this observation to extract geodesic distances and information about geometry from quantum dynamics **on data**?

First ingredient: the graph Laplacian

- Some of the most successful methods for manifold learning rely graph Laplacians constructed from data
  - Diffusion maps, Laplacian eigenmaps, Local linear embedding, …

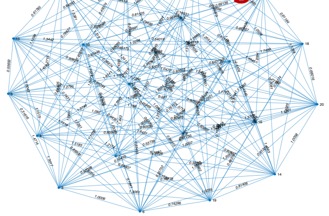
Data



$$V = \{v_1, v_2, \ldots, v_N\}$$

$$v_i \in \mathbb{R}^n$$

Graph/point cloud embedding

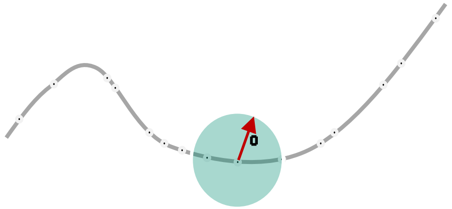


$$[T_{N,\epsilon}]_{i,j} = \exp(-\frac{\|v_i - v_j\|^2}{2\epsilon})$$

Adjacency matrix for symmetric weighted graph

Normalized graph Laplacian

$$L_{N,\epsilon} = \frac{4}{\epsilon}\left(I_N - D_{N,\epsilon}^{-1}\Sigma_{N,\epsilon}^{-1}T_{N,\epsilon}\Sigma_{N,\epsilon}^{-1}\right)$$



ϵ is a key "scale" parameter

Typically, $\epsilon \to 0$ as $N \to \infty$

Methods hinge on key convergence result

$$L_{N,\epsilon} \xrightarrow{N\to\infty} \mathcal{L}_\epsilon = \Delta_g + \mathcal{O}(\epsilon)$$

Coifman & Lafon, Appl. Comp. Harm. Anal., **21**, 5 (2006)
Hein, Audibert, von Luxburg, J. Mach. Learn. Res., 8, 1325 (2007)

"Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps.," Coifman et al. *Proc. Natl. Acad. Sci.*, **102**, 7426 (2005)
"Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Belkin, Niyogi, *Neural Computation*, **15**, 1373 (2003)

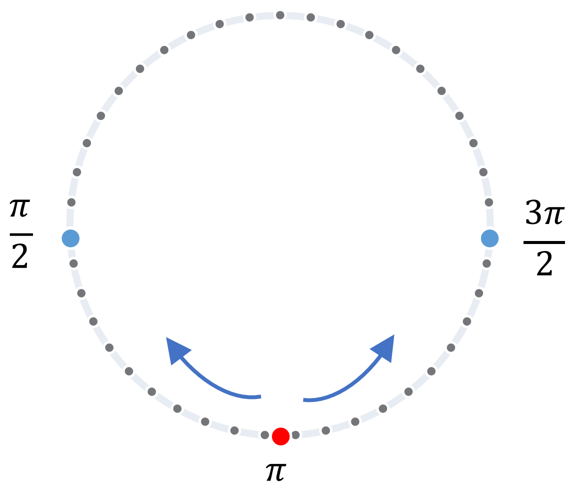
Quantum dynamics on data

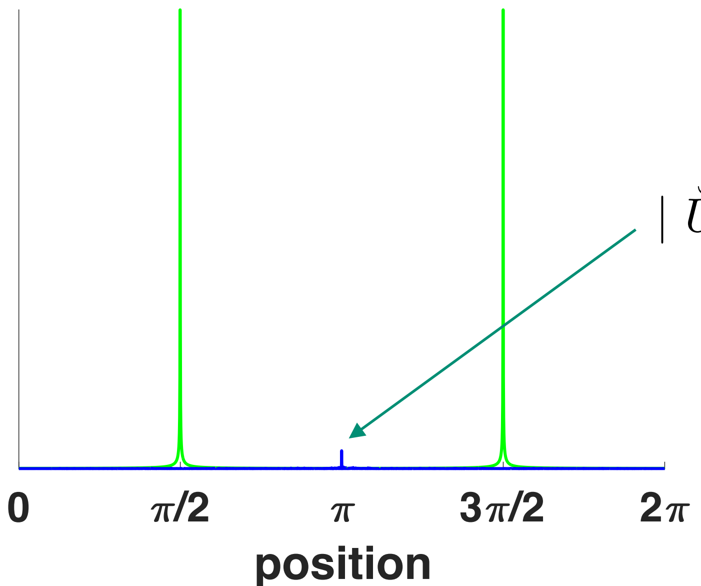
Normalized graph Laplacian

Recall: $L_{N,\epsilon} \xrightarrow{N\to\infty} \mathcal{L}_\epsilon = \Delta_g + \mathcal{O}(\epsilon)$

Build unitary propagator (NxN matrix) $\breve{U}(t) = e^{-i\sqrt{L_{N,\epsilon}}t}$

Example: Build data-driven propagator with $N = 2500$ samples from the circle





$|\ \breve{U}(t = \frac{\pi}{2})|\delta_{x=\pi}\rangle\ |^2$

Quantum dynamics on data

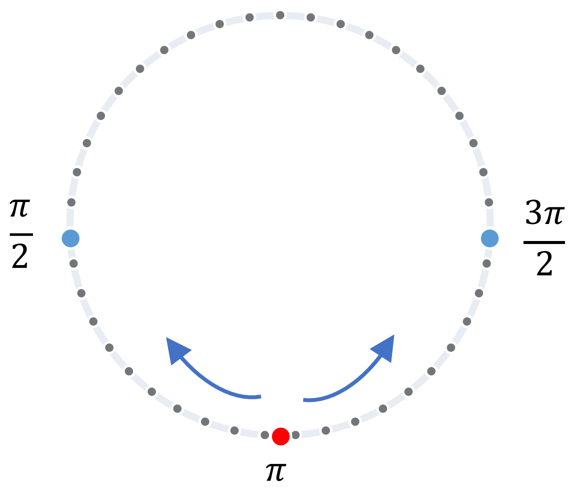
Normalized graph Laplacian

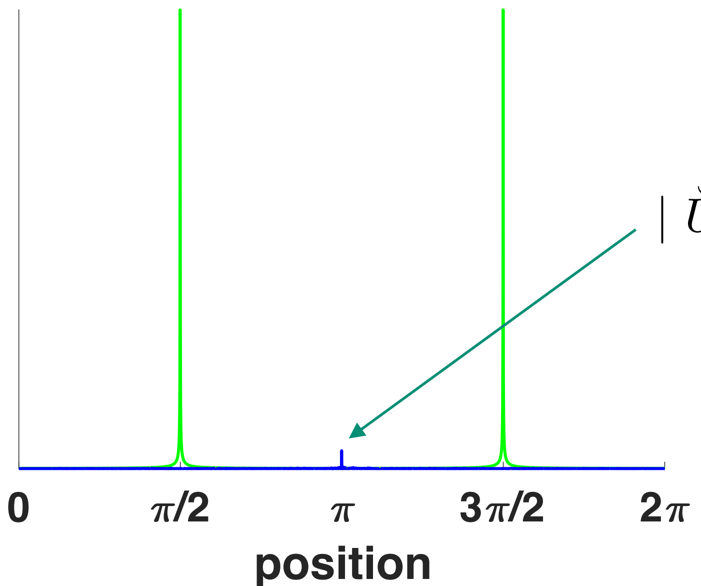
Recall: $L_{N,\epsilon} \xrightarrow{N\to\infty} \mathcal{L}_\epsilon = \Delta_g + \mathcal{O}(\epsilon)$

These error terms are the problem.

Build unitary propagator (NxN matrix) $\breve{U}(t) = e^{-i\sqrt{L_{N,\epsilon}}t}$

Example: Build data-driven propagator with $N = 2500$ samples from the circle





$|\ \breve{U}(t = \frac{\pi}{2})|\delta_{x=\pi}\rangle\ |^2$

Resolution limits of data-driven quantization

- The problem is the finite resolution of the manifold given by finite N, and quantified by ϵ.

- We are asking for too much to resolve delta function propagation when $\epsilon > 0$.

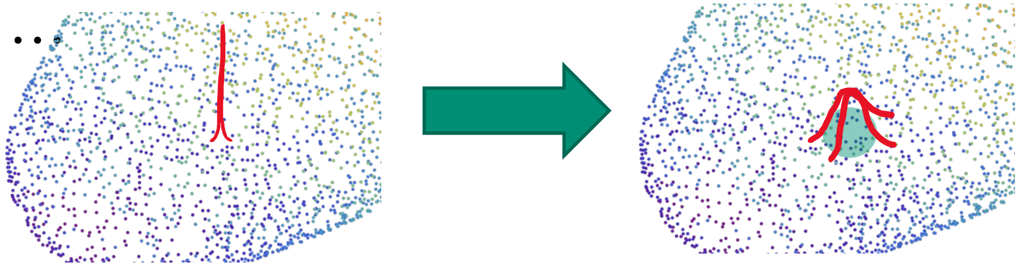
- Roughly: $\breve{U}(t) = e^{-i\sqrt{L_{N,\epsilon}}t}$ "resolves" position space at scale $\sqrt{\epsilon}$

  "resolves" momentum *up to* bandwidth $1/\sqrt{\epsilon}$

  Approximates the action of the operator $e^{-it\sqrt{\Delta_g}}$

- In contrast, initial states like $|\delta_x\rangle$ are infinitely localized *and* have unbounded momentum.

- So we need a way to coarsen the dynamics …

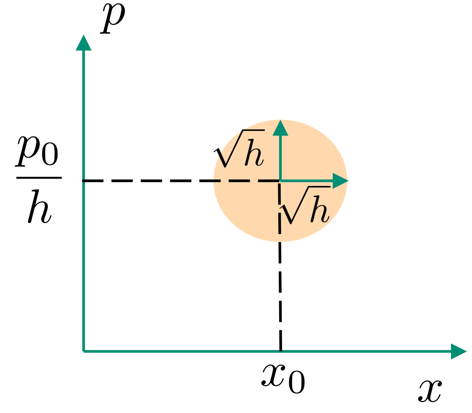


Coherent states to the rescue

- Coherent states are
  - The "most classical" of quantum states
  - Minimum uncertainty in phase space, maximum localization in phase space

- Can define coherent states on a general manifold. In position basis:

$$\psi_{\zeta_0}(\mathbf{x}) = \langle \mathbf{x} | \psi_{\zeta_0} \rangle = \frac{1}{(\pi h)^{\frac{\nu}{4}}} e^{\frac{i}{h}\langle \mathbf{x}-\mathbf{x}_0, \mathbf{p}_0 \rangle} e^{-\frac{\|\mathbf{x}-\mathbf{x}_0\|^2}{2h}} \qquad \zeta_0 = (\mathbf{x}_0, \mathbf{p}_0) \in T^*\mathcal{M}$$



Combescure & Robert. *Coherent States and Applications in Mathematical Physics*. Springer Netherlands, 2012.
Gazeau, *Coherent States in Quantum Physics*. Wiley-VCH, 2009.

- Localized to $\sqrt{h}$ in space and momentum bandwidth scales as $^1/_h$

- We can control coarse-ness of phase space resolution using h. Match to data resolution:

$$h > \sqrt{\epsilon} \quad \Rightarrow \quad \boxed{h = \epsilon^{\frac{1}{2+\alpha}}, \alpha > 1}$$

"Classical limit" $h \to 0$
is the large data limit $N \to \infty$

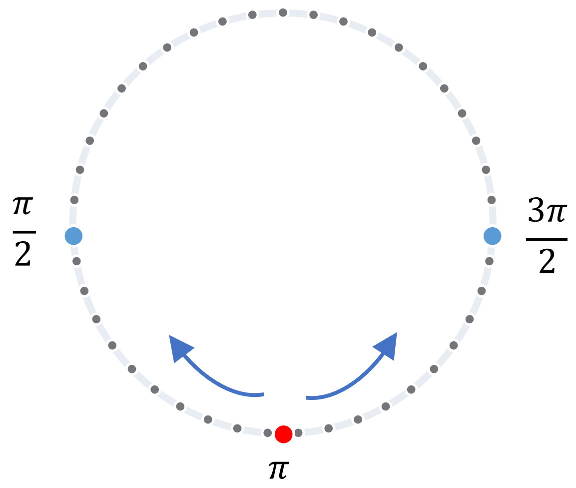
Coherent states to the rescue

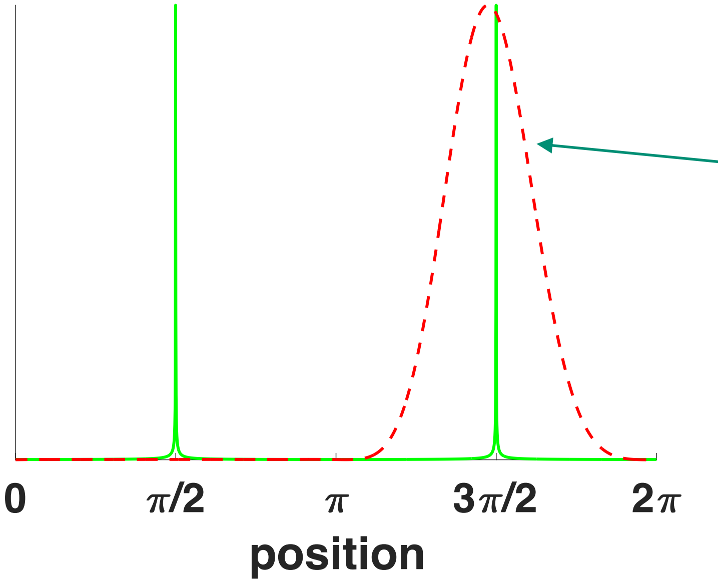
- Can approximate coherent state using the data. A coherent state centered at data point v_0 has elements

$$\left[|\psi_h^{\zeta_0}\rangle\right]_i \propto e^{\frac{i}{h}(v_i - v_0)^\top p_0} e^{-\frac{\|v_i - v_0\|^2}{2h}}, \quad 1 \le i \le N. \qquad |\psi_h^{\zeta_0}\rangle \in \mathbb{C}^N$$

- Initial momentum, p_0, approximated using a vector from v_0 to nearest point, or using local principal component analysis (LPCA)

Example: Build data-driven propagator with $N = 2500$ samples from the circle



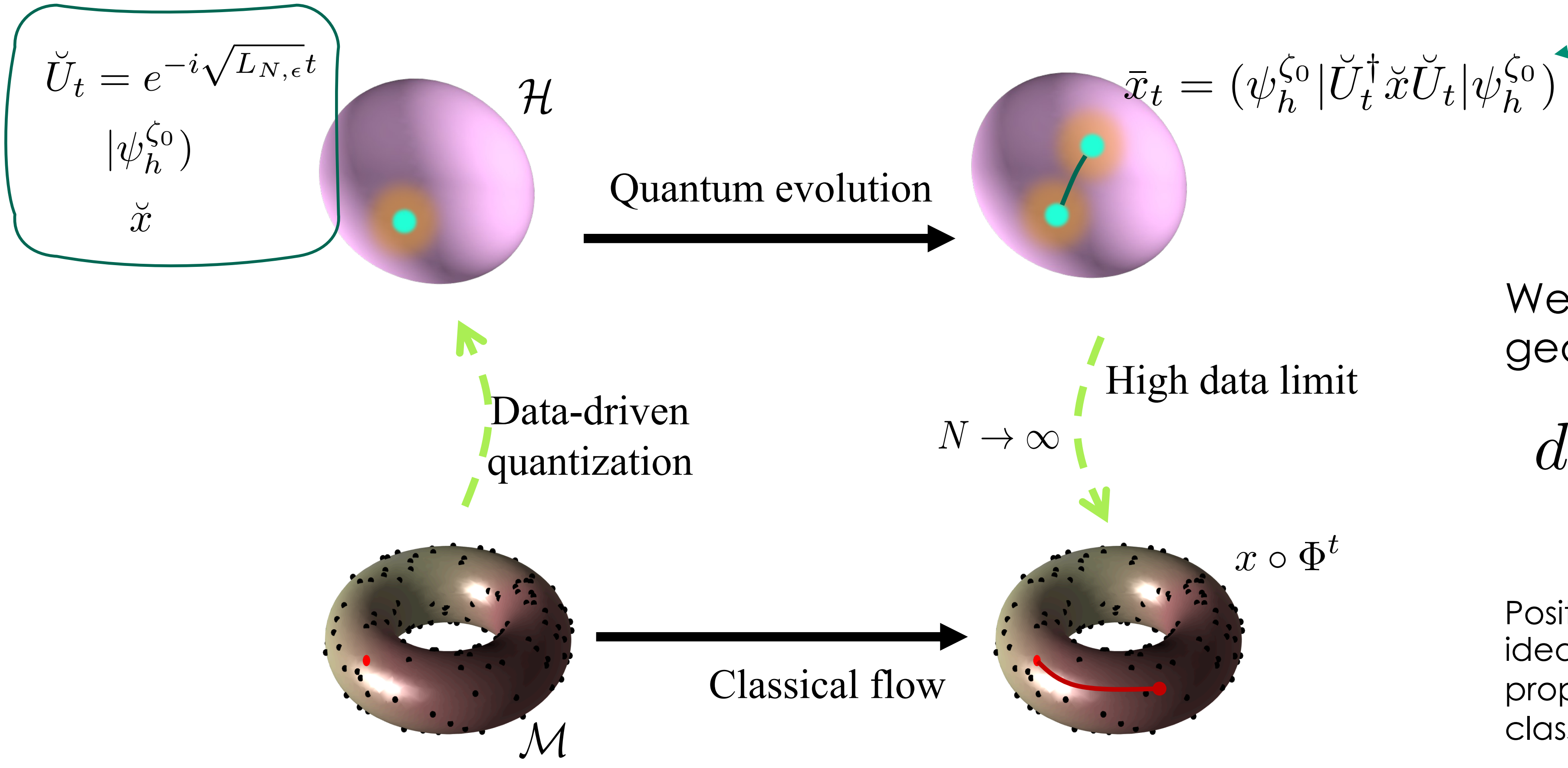


$$|\ \breve{U}(t = \frac{\pi}{2})|\psi_h^{\zeta_0}\rangle\ |^2$$

Discrete quantum-classical correspondence

"Manifold learning via quantum dynamics." A. Kumar & M. Sarovar. arXiv:2112.11161 (2021)

By approximating the correct quantum operators and quantum states from the data, and setting h appropraitely, we show that it is possible to approximate **geodesics on the data manifold**.



Expected position at time t for data-driven propagated quantum state

We prove bounded error in geodesic distance

$$d_g(\bar{x}_t, x \circ \Phi_t) < h$$

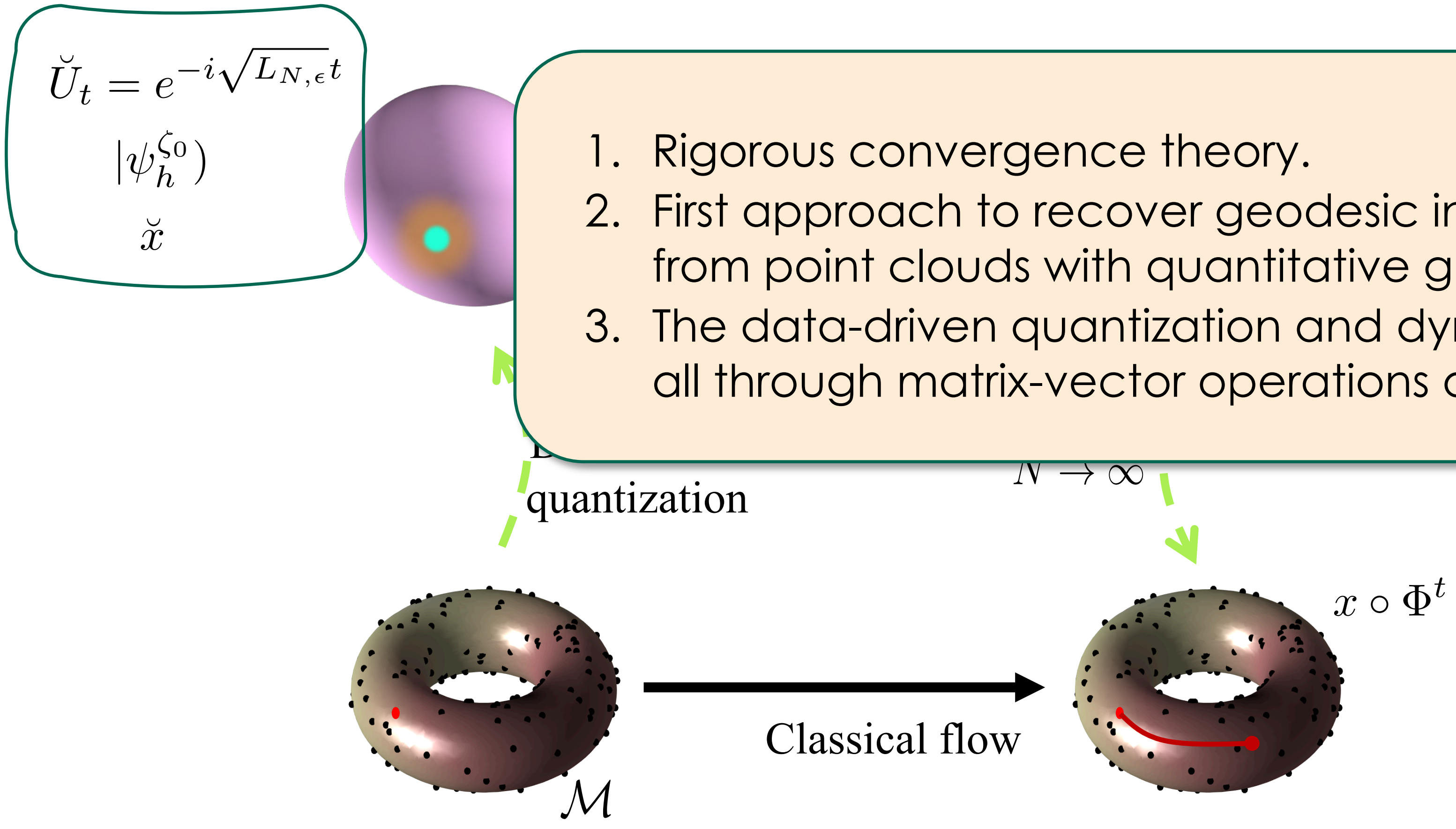
Position at time t for ideal free propagation on classical manifold

Data-determined "Planck's constant"

Discrete quantum-classical correspondence

"Manifold learning via quantum dynamics." A. Kumar & M. Sarovar. arXiv:2112.11161 (2021)

By approximating the correct quantum operators and quantum states from the data, and setting h appropraitely, we show that it is possible to approximate **geodesics on the data manifold**.



1. Rigorous convergence theory.
2. First approach to recover geodesic information from point clouds with quantitative guarantees.
3. The data-driven quantization and dynamics are all through matrix-vector operations on the data.

Expected position at time t for data-driven propagated quantum state

ove bounded error in
sic distance

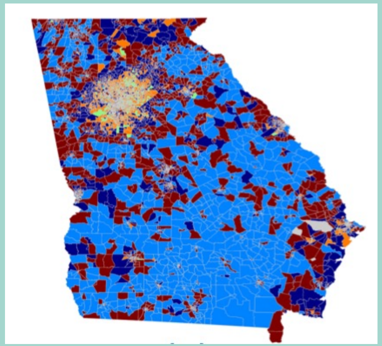
$$d_g(\bar{x}_t, x \circ \Phi_t) < h$$

Position at time t for ideal free propagation on classical manifold

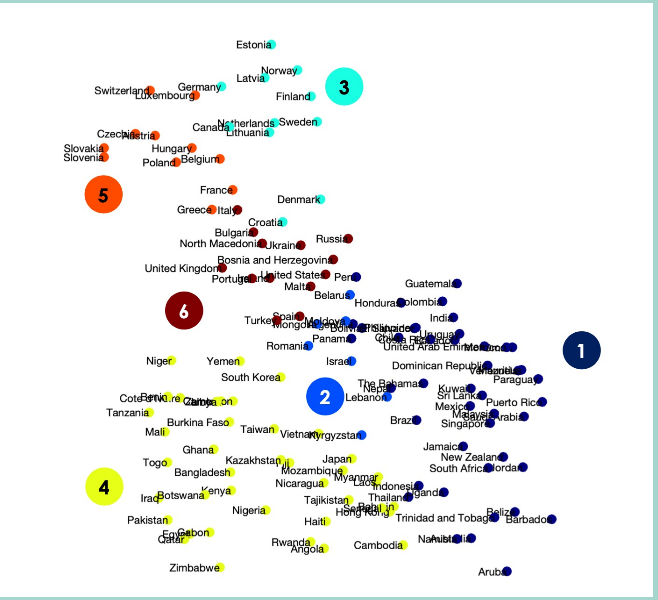
Data-determined "Planck's constant"

Applications to real-world data

Clustering, anomaly detection from data of COVID-19 social distancing behavior across geographic regions

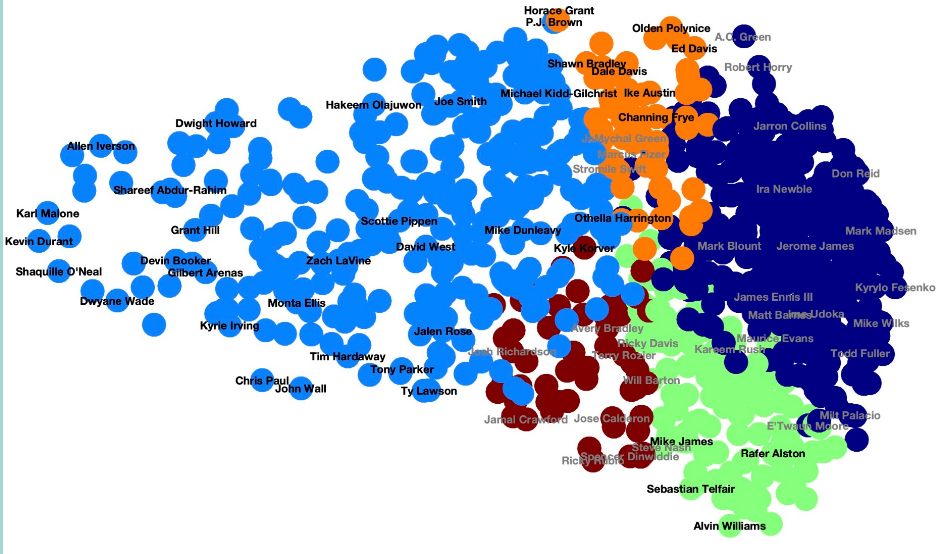


SafeGraph dataset for state of GA



Google dataset for worldwide mobility patterns

Clustering of NBA players based on performance

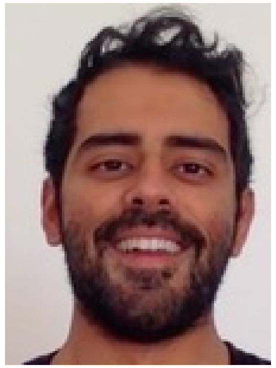


NOTES:

1. Manifold hypothesis not necessary for utility
2. Some of these datasets are small (e.g., N=126)

See "Manifold learning via quantum dynamics." A. Kumar & M. Sarovar. arXiv:2112.11161 (2021)

Acknowledgements



Akshat Kumar
Clarkson University &
Instituto de Telecommunicações, Portugal

QUANTUM SYSTEMS ACCELERATOR
Catalyzing the Quantum Ecosystem

For more details see:

Manifold learning via quantum dynamics

Akshat Kumar[1,2,a] and Mohan Sarovar[3,b]
[1] *Department of Mathematics, Clarkson University, Potsdam, NY 13699 USA*
[2] *Instituto de Telecomunicações, Lisbon, Portugal*
[3] *Sandia National Laboratories, Livermore, California 94550, USA*

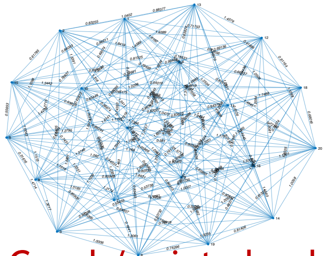
arXiv:2112.11161

Backup slides

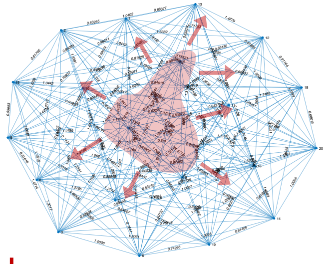
Markov methods for manifold learning

car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocatio	wheelbase	carlength	carwidth	carheight	curbweight
1	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
2	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
3	1	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823
4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337
5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824
6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507
7	1	audi 100ls	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844
8	1	audi 5000	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954
9	1	audi 4000	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086
10	0	audi 5000s (	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053
11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395
12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395
13	0	bmw x1	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710
14	0	bmw x3	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765
15	1	bmw z4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055
16	0	bmw x4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230
17	0	bmw x5	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380
18	0	bmw x3	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505

Data

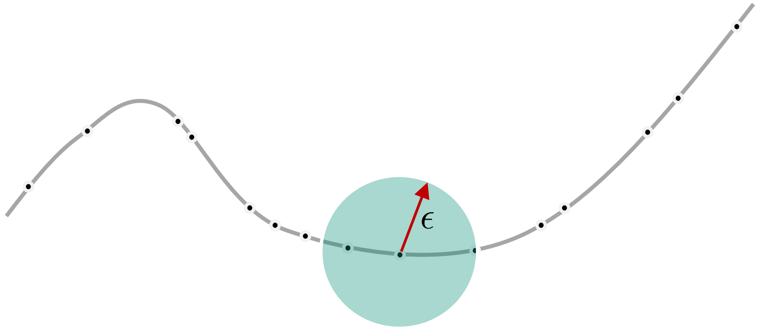


Graph/point cloud embedding



A Markov process connects local distances to global ones

$$V = \{v_1, v_2, ..., v_N\} \qquad v_i \in \mathbb{R}^n$$



Adjacency matrix for symmetric weighted graph

$$[T_{N,\epsilon}]_{i,j} = \exp(-\frac{\|v_i - v_j\|^2}{2\epsilon})$$

Normalized graph Laplacian

$$L_{N,\epsilon} = \frac{4}{\epsilon}\left(I_N - D_{N,\epsilon}^{-1}\Sigma_{N,\epsilon}^{-1}T_{N,\epsilon}\Sigma_{N,\epsilon}^{-1}\right)$$

Normalizations

$$\Sigma_{N,\epsilon} = \mathrm{diag}\left(\sum_{j=1}^{N}[T_{N,\epsilon}]_{i,j}\right)$$

$$D_{N,\epsilon} = \mathrm{diag}\left(\sum_{j=1}^{N}[\Sigma_{N,\epsilon}^{-1}T_{N,\epsilon}\Sigma_{N,\epsilon}^{-1}]_{i,j}\right)$$

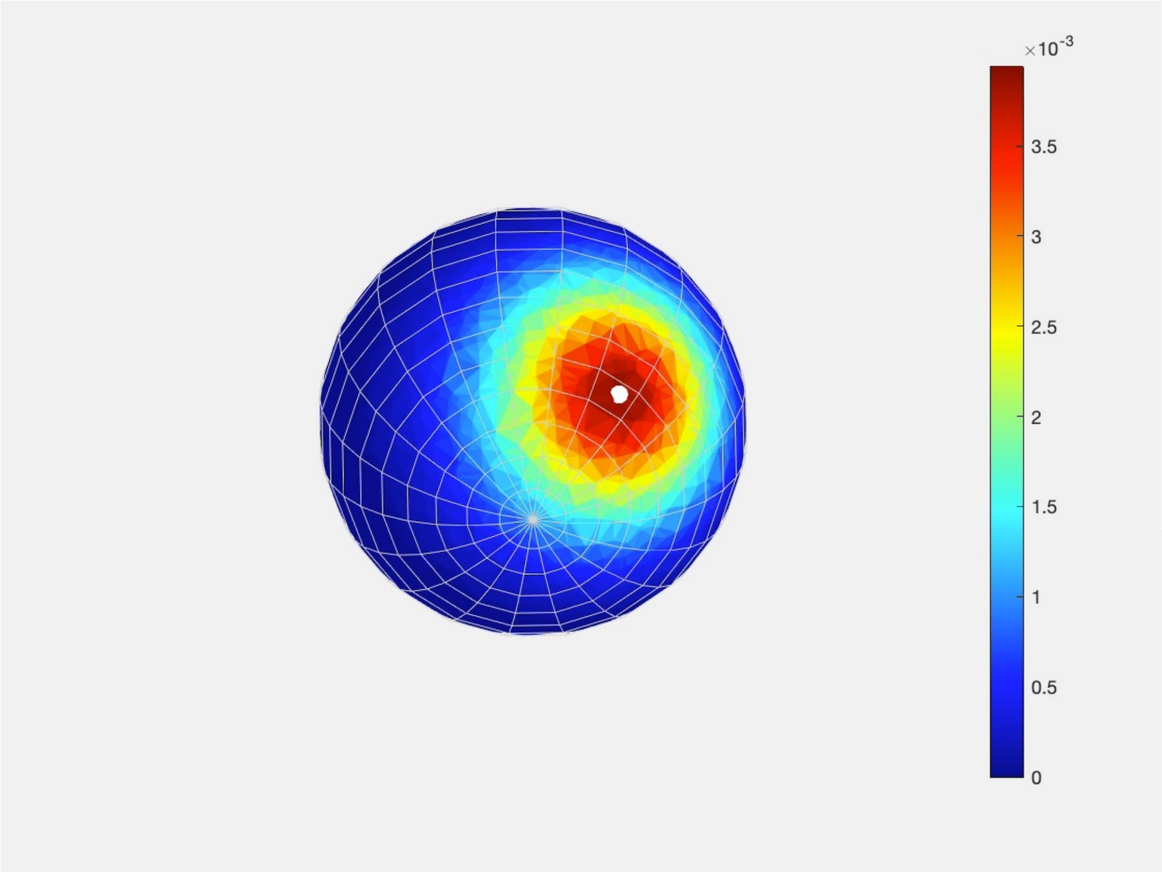
ϵ is a key "scale" parameter

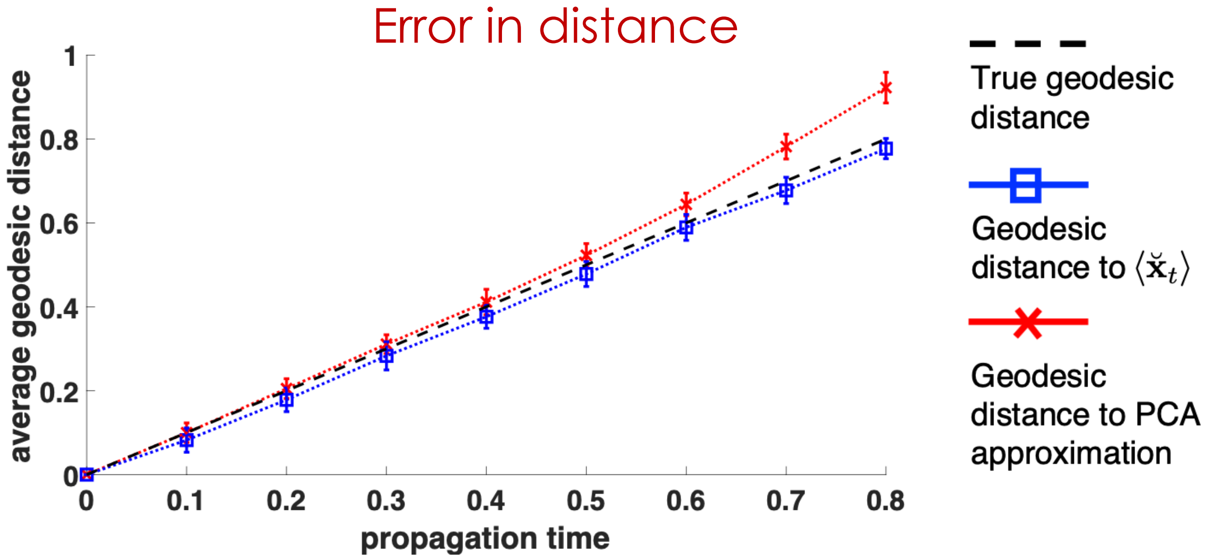
Typically, $\epsilon \to 0$ as $N \to \infty$

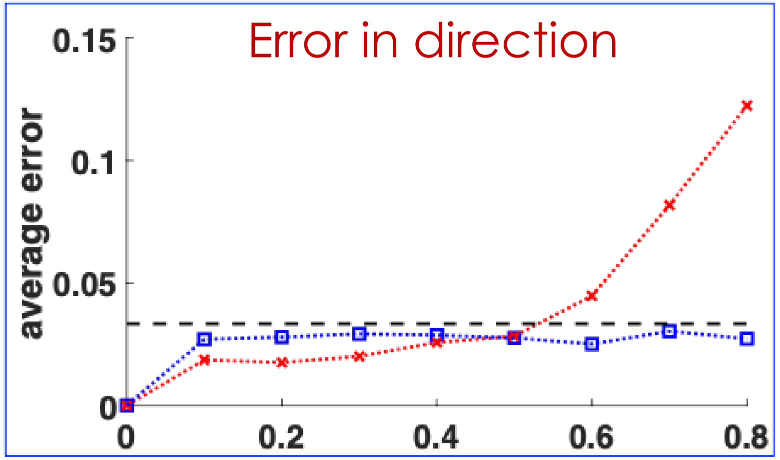
"Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps.," Coifman et al. *Proc. Natl. Acad. Sci.*, **102**, 7426 (2005)
"Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Belkin, Niyogi, *Neural Computation*, **15**, 1373 (2003)

Example: sphere

N=3000 points, uniformly sampled on unit sphere

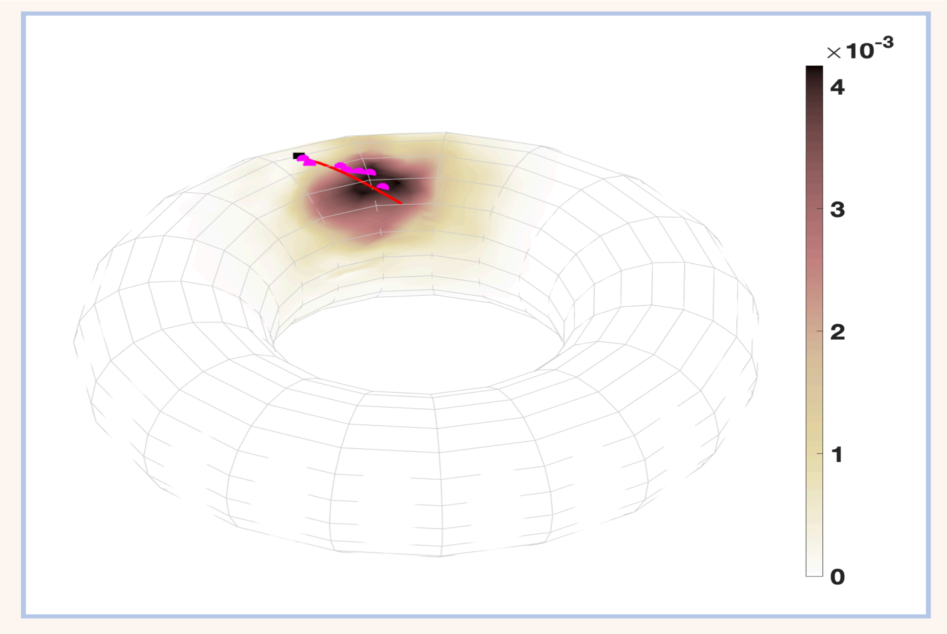


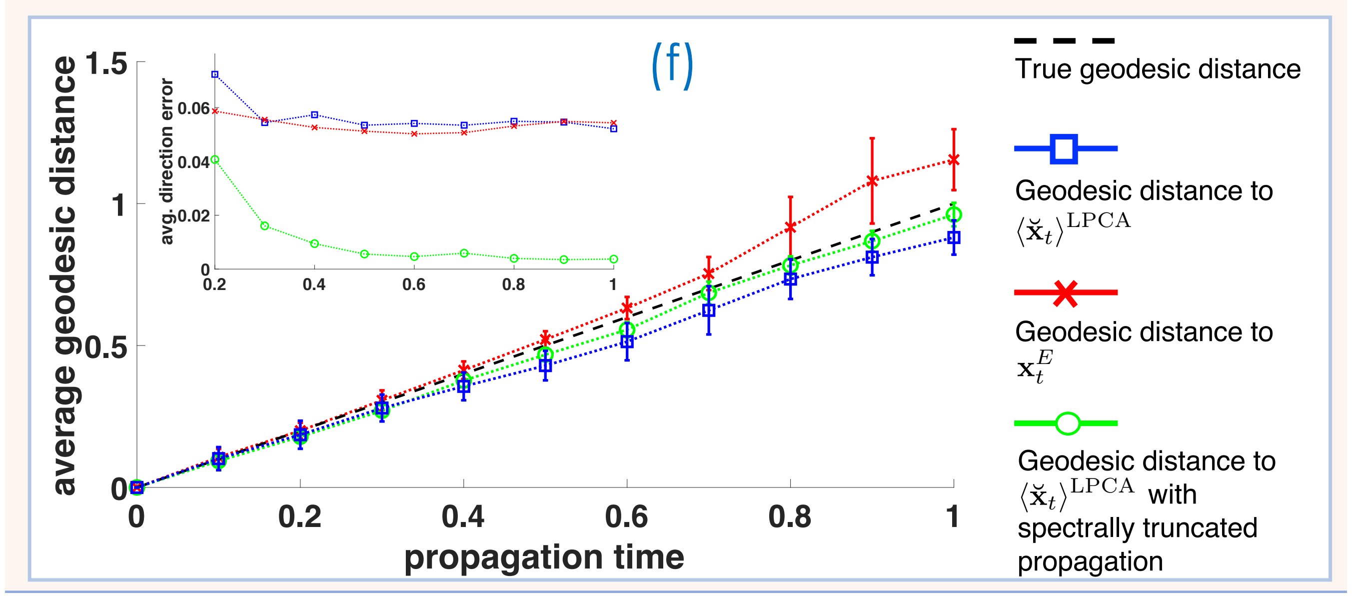




Example: torus

N=12000 points, uniformly sampled on 2-torus



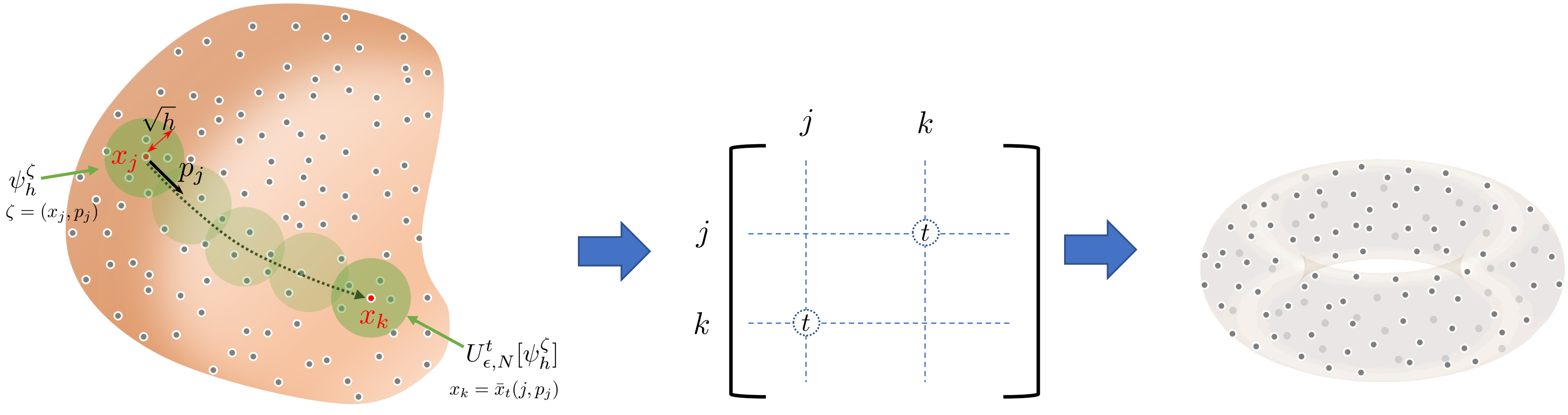


Using geodesic distances for practical tasks

We have shown that quantum propagation allows for extraction of geodesic distances between data points (under the manifold hypothesis). How is this useful?

Geodesic distances (or in the absence of a manifold, *quantum walk distances*) define a similarity metric between data points.

We can embed the data in a graph based on this similarity metric. This graph reflects the geometry of the dataset, and delivers new coordinates for the data => visualization, clustering, classification, …

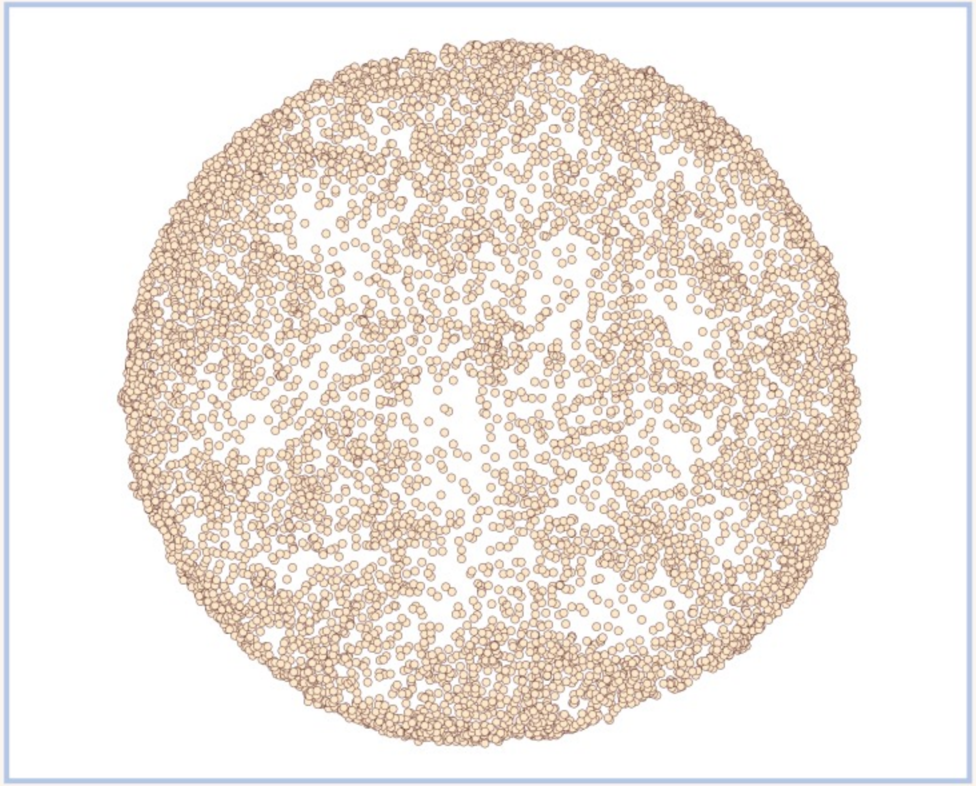


Cross-section of dataset of size N in $\mathbb{R}^n$

$\mathcal{G}$ ($N \times N$ matrix)

Graph embedding of dataset in $\mathbb{R}^3$ based on weighted adjacency matrix $\mathcal{G}$

Embeddings of sphere and torus



N=8000 points

N=12000 points

Example: COVID-19 mobility data

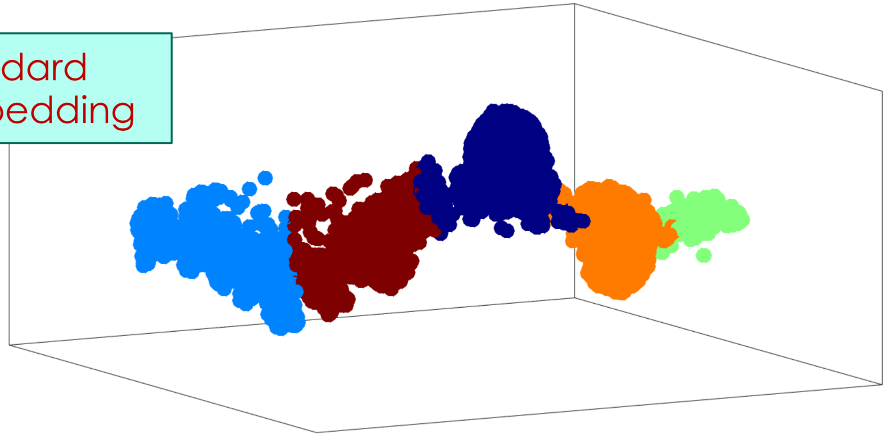
Social Distancing Metric dataset from SafeGraph Inc.
https://docs.safegraph.com/docs/social-distancing-metrics

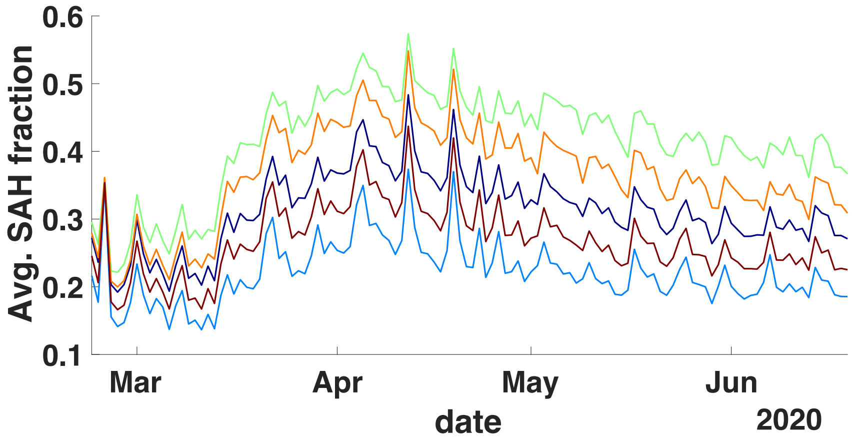
- Dataset collects user location information (from cellphone GPS data) over the course of the initial 3 months of the COVID-19 pandemic (Feb 23, 2020 – June 19, 2020: 117 days).
- Aggregated at the census block group (CBG) level.
- Understanding patterns in mobility behavior can help tune public health policy.
- We compute a "stay-at-home" fraction which represents the fraction of devices that stayed at their home location on a day.
- We concentrate on the data for Georgia (GA), which has 5509 CBGs.
- Dataset: 5509 x 117

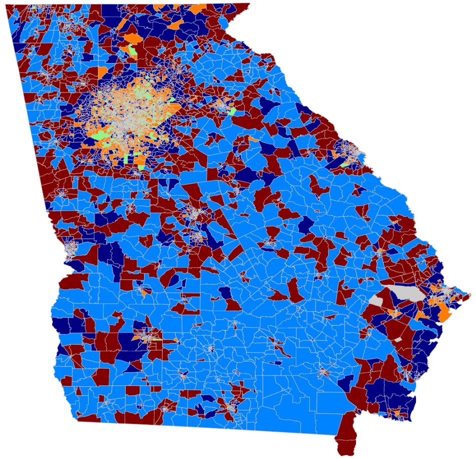
Apply manifold learning through geodesics and embed in 3 dimensions (**reduction from 117 dimensions**) and then perform clustering using K-means.

Example: COVID-19 mobility data

Standard embedding



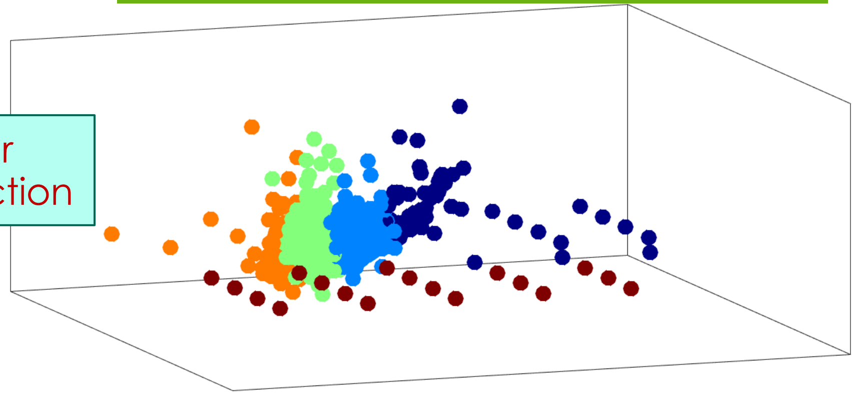


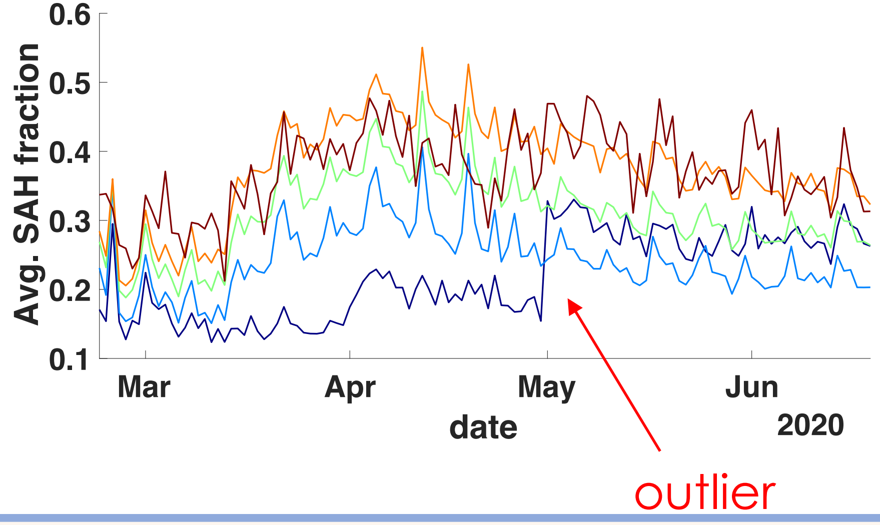


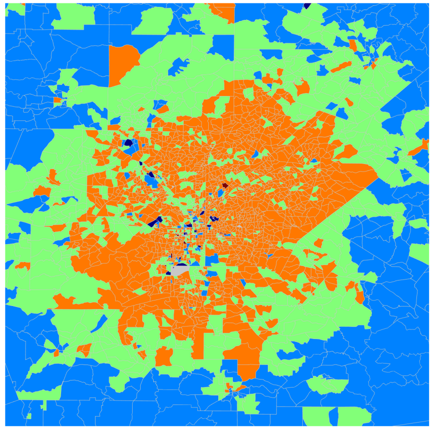
Embedding in 3D and k-means using these 3D coordinates

Average SAH fraction time series for each cluster

Outlier detection

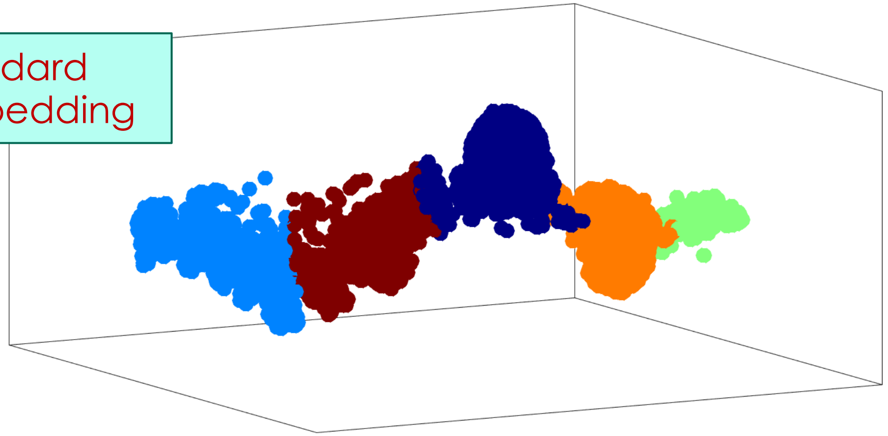


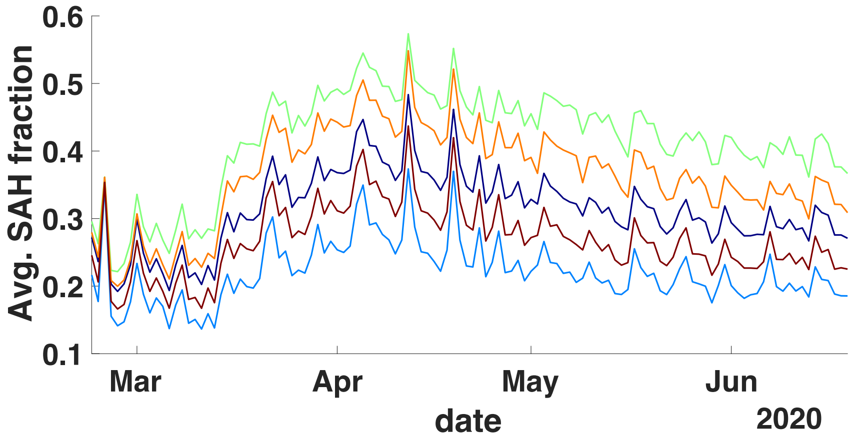


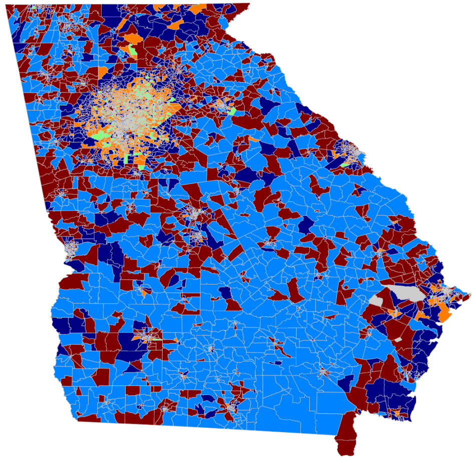


Example: COVID-19 mobility data

Standard embedding



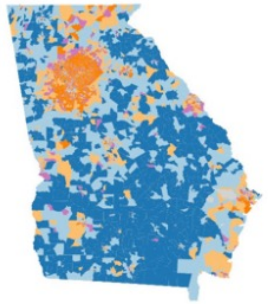




Embedding in 3D and k-means using

Average SAH fraction time series for

c.f. Levin et al., "Cell Phone Mobility Data and Manifold Learning." https://doi.org/10.1101/2020.10.31.20223776



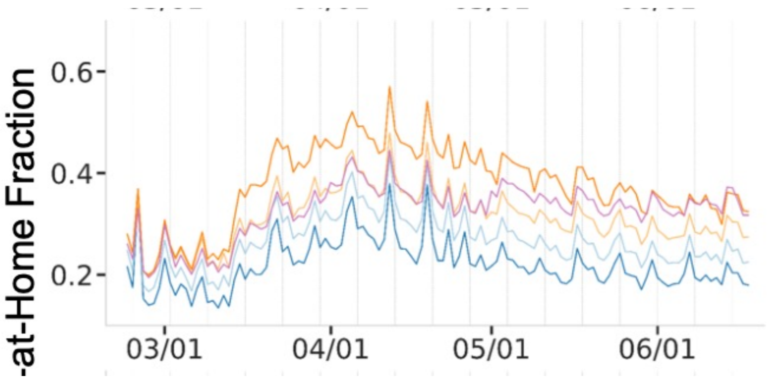


Figure 2: Laplacian eigenmaps, clustering done after embedding in **14 dimensions**

We achieve better quality clustering with QML, and are able to identify outliers, even with an embedding into fewer dimensions

Example: Global COVID-19 mobility data

Google COVID-19 Community Mobility Reports

https://www.google.com/covid19/mobility/

Dataset collects user mobility information (% change in mobility from baseline) over 6 categories for 132 countries and regions within these countries.

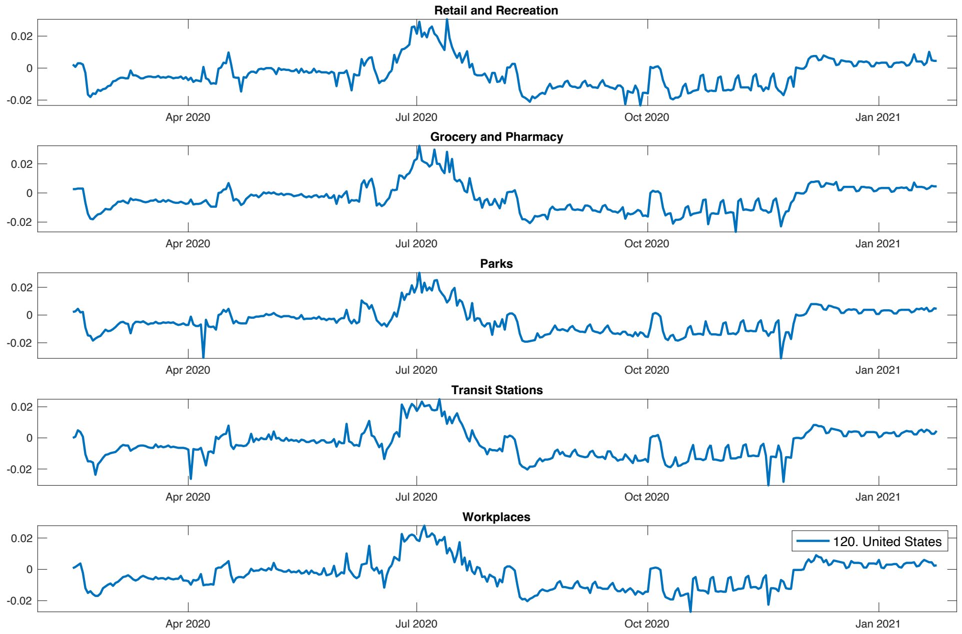
Timeframe: ~1 year (Feb 15, 2020 – Jan 24, 2021)
Baseline: Jan 3 – Feb 6, 2020
Categories: Retail and recreation, grocery and pharmacy, parks, transit stations, workplaces.

After pre-processing:

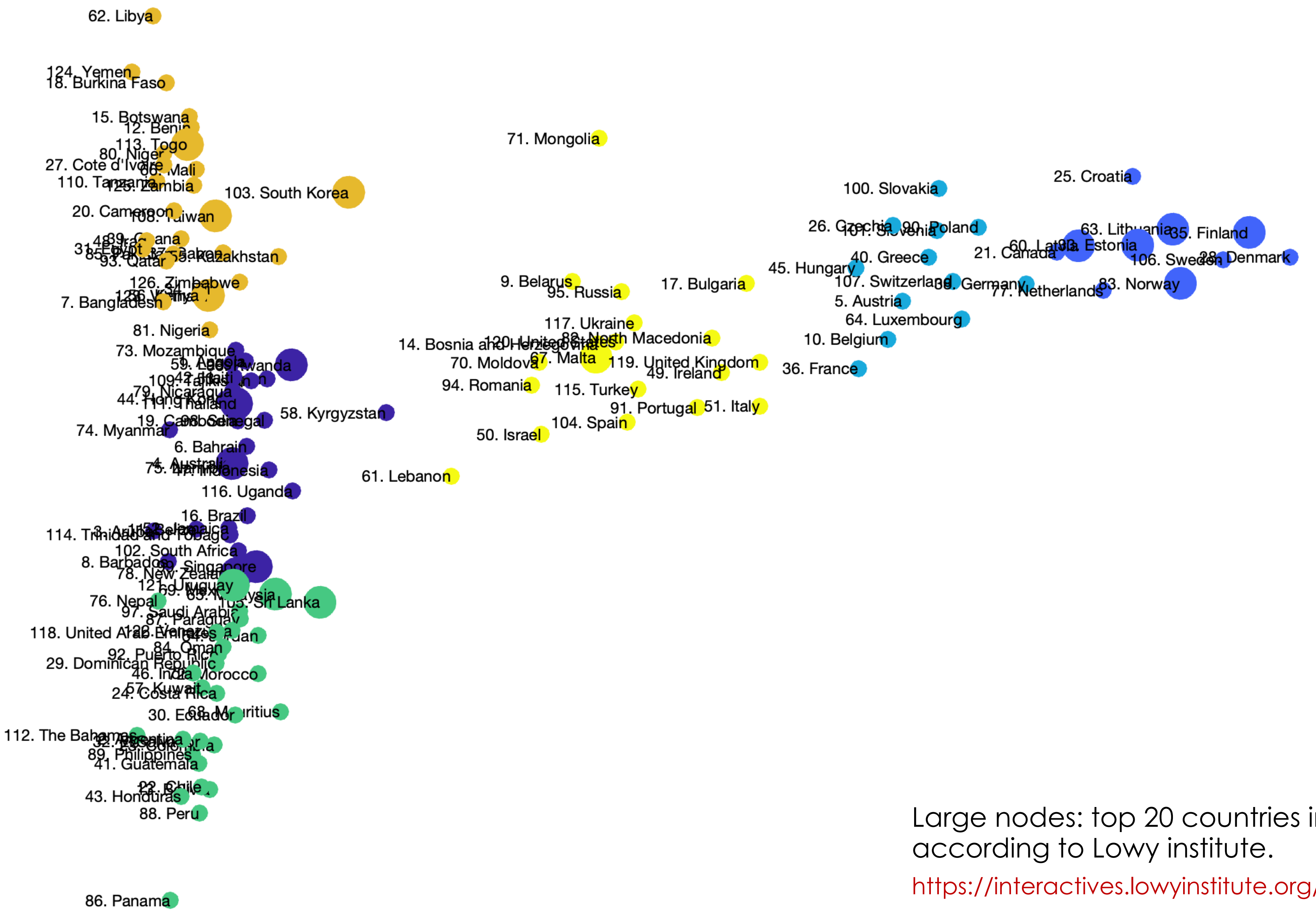
For each country, **345*5=1725 columns (features)** that represent a time series of mobility changes across 5 categories.

Apply manifold learning through geodesics and embed in 3 dimensions (**reduction from 1725 dimensions**)

e.g.,



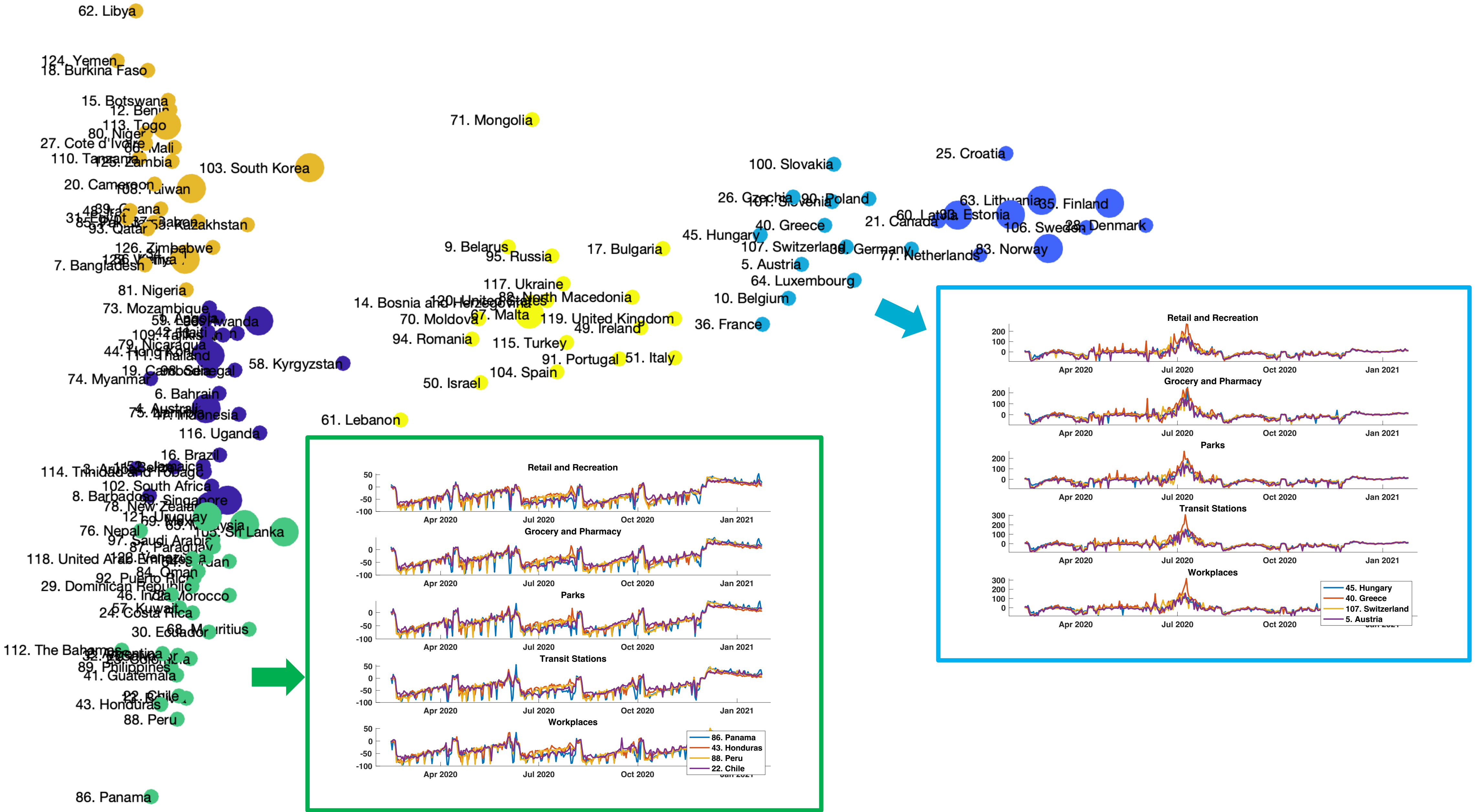
Example: Global COVID-19 mobility data



Large nodes: top 20 countries in COVID response according to Lowy institute.

https://interactives.lowyinstitute.org/features/covid-performance/

Example: Global COVID-19 mobility data



Signal processing interpretation

Propagate a coherent state and look at the result in the position basis

$$|\langle \delta_x | U^t | \psi_h^\zeta \rangle|^2 = |\langle \psi_h^\zeta | U^t | \delta_x \rangle|^2$$

Propagate a position eigenstate and look at the result in a coherent state basis

$|\langle \psi_h^\zeta | f \rangle|^2$ is actually a Gabor spectrogram of f (also a Husimi-Q function)

Gabor transform is a short-time Fourier transform, defined by integration against Gabor wavelets that are delocalized in time and frequency

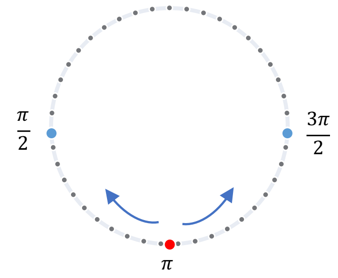
$$G_x(t,f) = \sqrt[4]{\sigma} \int_{-\infty}^{\infty} e^{-\sigma\pi(\tau - t)^2} e^{-j2\pi f\tau} x(\tau) d\tau$$

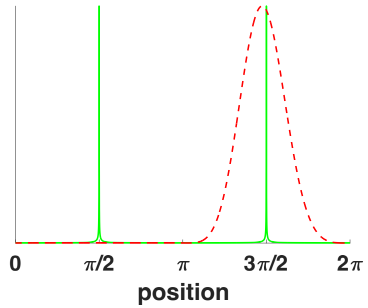
A coherent state is a Gabor wavelet (this time defined over coordinate and momentum, or phase, space)

By choosing h we are choosing the (phase space) scale at which to resolve the signal $U^t|\delta_x\rangle$

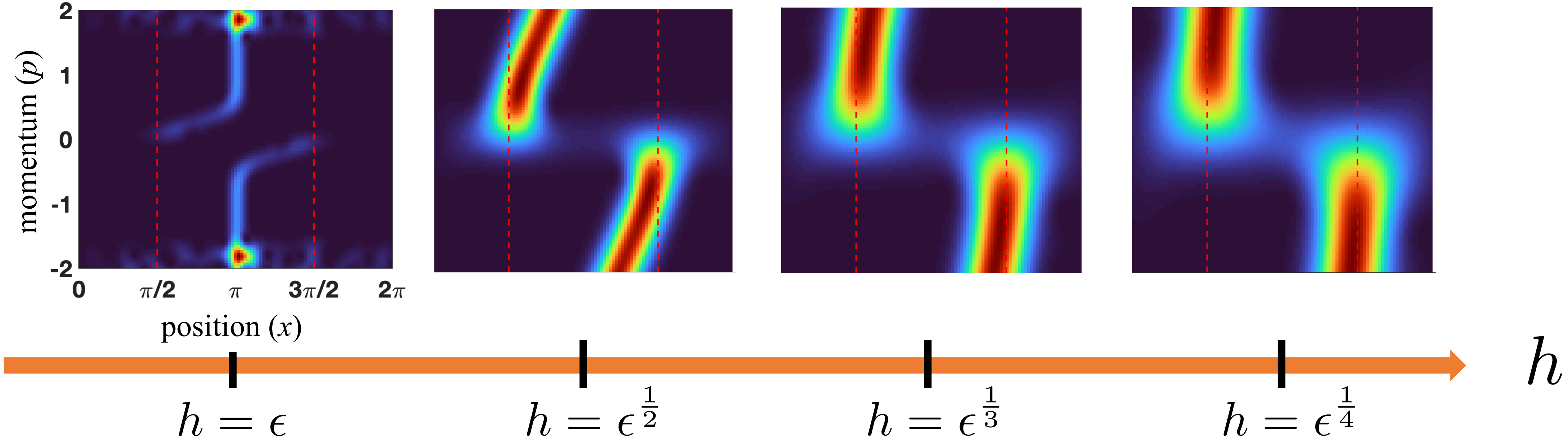
Signal processing interpretation

Example: $N = 2500$ samples from the circle



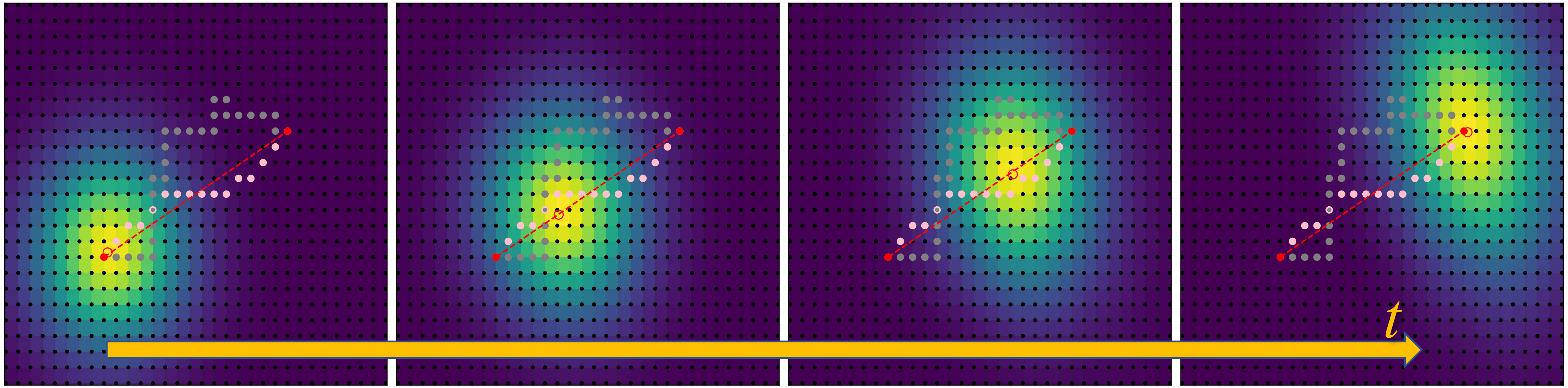


Gabor spectrogram at varying resolution h



Geodesic interpolation

Example of points sampled on a plane, what is shortest path between two points?



- Notice how coherent state propagates smoothly between known/sampled data points

- Coherent state charts a straight line (shortest) path, while other ”shortest-path” algorithms (e.g., Dijkstra’s algorithm) using the same data define longer paths.

- Our technique allows one to interpolate and extrapolate from the sampled data in a way that is consistent with the data geometry.