Problem: Feature Discovery

Terms:
- Dictionaries: Sets of characteristic functions or features. We use superpositions of these to build a "model", or data representation.
- Low-dimensional signals/functions: e.g. 10 sounds, 20 individual images.
- High-dimensional data: Point clouds in D-dimensional space, where each point is one piece of data and D=\#pixels/image, \#terms/corpus or \#samples/spoken vowel.

Solution: Geometric Wavelets

1. Cut the data into pieces.
2. Approximate each piece as a low-dimensional plane.
3. Use these approximations as an interpretable representation or feature set (dictionary) for the data that can be used for compression, filtering, outlier detection, etc.

More Details:
1. Use a similarity measure to create a graph from the data points. Construct a set of multi-scale partitions of \( \mathcal{M} \) by using recursive spectral cuts.
   (METB uses Eigendfunctions of the Laplacian over the graph – we are also implementing Cover Trees. This step often involves task-specific method variations.)
2. Compute the SVD of the data covariance for each piece. This gives Scaling Functions & \( \mathcal{M}_{j,k} \) – a manifold approximation at scale \( j \) for piece \( k \) – a projection onto that local approximate tangent space.
3. Higher-scale planes are small corrections to the previous parent. Efficiently encode the differences between \( \mathcal{M}_{j+1} \rightarrow \mathcal{M}_j \) by constructing Wavelet spanning space & "detail" operators analogous to Wavelet theory.

Experience: Interactive GUI

- Quickly see much more data so mathematicians can evaluate methods during development.
- Begin developing a platform onto which we can build more specialized applications for new tasks and data types.
- Help explain Geometric Wavelets and gain intuition about the representation.

Our Approach:

We do not try to solve this problem "in general", but exploit the fact that often real data has lower-dimensional geometric structure, such as lying near a manifold (\( \mathcal{M} \)) of dimensionality d<<D.

Geometric Wavelets is a novel construction which discovers features in high-dimensional data under these geometric assumptions.

It is explicit, which leads to interpretable features, and it comes with guarantees (as a function of an approximation error parameter) on computational cost, number of elements in the dictionary and sparsity of the representation. It is globally non-linear, but piecewise linear, so it is fast, but can adapt to arbitrary non-linear manifolds.

Observations & Future Directions:

- Coarser scales contain generalized approximations of the data, with readily interpretable node centers and wavelet directions.
- Finer scales reveal anomalous data through extreme wavelet coefficients or "odd" wavelet axis images.
- Coarser-scale wavelets contain information which could be ignored for classification tasks, but finer-scale wavelets encode more specific features which cluster and characterize individuals.
- Developers have changed their ideas about data encoding after viewing their results in the GUI.
- We are already working on variable dimensionality, group definition & labeling for classification and outlier detection, views for new data types.