# Math Foundations of Multiscale Graph Representations and Interactive Learning

Mauro Maggioni, Rachael Brady, Eric Monson

Mathematics and Computer Science
Duke University

FODAVA Kickoff meeting, 09/16/08

Funding: NSF/DHS-FODAVA, DMS, IIS, CCF; ONR.

# Research Goal

- Develop mathematical framework for multiscale analysis of graphs and of functions on graphs
- Dynamics of and on graphs
- Active learning within a multiscale framework
- Incorporation of uncertainty

# Basic tools

Main motivation: let the geometry of the data dictate what the relevant patters are.

- Random walk (r.w.) as a starting point
- Large time analysis of r.w.'s leads to spectral embeddings, cuts, clustering - and Fourier-like analysis on graphs. Classical, widely used, many open problems, some recent results.
- Multiscale time analysis of r.w.'s leads to graph hierarchical representations - and wavelet-like analysis on graphs. New and in development.

# Basic tools

Main motivation: let the geometry of the data dictate what the relevant patters are.

- Random walk (r.w.) as a starting point
- Large time analysis of r.w.'s leads to spectral embeddings, cuts, clustering - and Fourier-like analysis on graphs. Classical, widely used, many open problems, some recent results.
- Multiscale time analysis of r.w.'s leads to graph hierarchical representations - and wavelet-like analysis on graphs. New and in development.

# Basic tools

Main motivation: let the geometry of the data dictate what the relevant patters are.

- Random walk (r.w.) as a starting point
- Large time analysis of r.w.'s leads to spectral embeddings, cuts, clustering - and Fourier-like analysis on graphs. Classical, widely used, many open problems, some recent results.
- Multiscale time analysis of r.w.'s leads to graph hierarchical representations - and wavelet-like analysis on graphs. New and in development.

# From data to graphs and networks

- Often data sets are modeled as graphs: similar data points connected by an edge
- Important classes of statistical models are graphs (graphical models)
- The topological properties of these graphs help analysis the data and functions on the data
- Continuous/infinite structures underlying graphs in several instances shed light on existing techniques

# From data to graphs and networks

- Often data sets are modeled as graphs: similar data points connected by an edge
- Important classes of statistical models are graphs (graphical models)
- The topological properties of these graphs help analysis the data and functions on the data
- Continuous/infinite structures underlying graphs in several instances shed light on existing techniques

# From data to graphs and networks

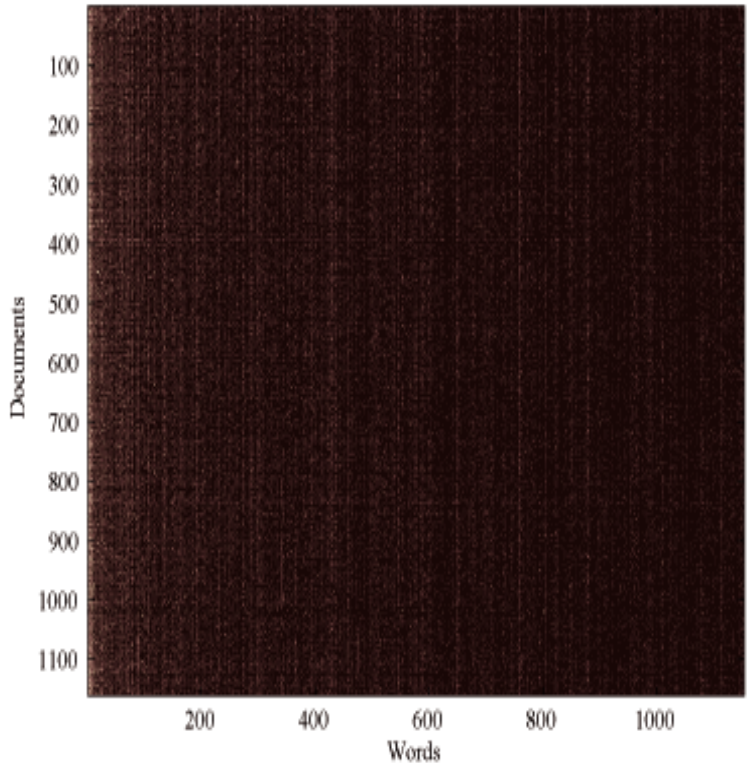
- Often data sets are modeled as graphs: similar data points connected by an edge
- Important classes of statistical models are graphs (graphical models)
- The topological properties of these graphs help analysis the data and functions on the data
- Continuous/infinite structures underlying graphs in several instances shed light on existing techniques

# From data to graphs and networks

- Often data sets are modeled as graphs: similar data points connected by an edge
- Important classes of statistical models are graphs (graphical models)
- The topological properties of these graphs help analysis the data and functions on the data
- Continuous/infinite structures underlying graphs in several instances shed light on existing techniques

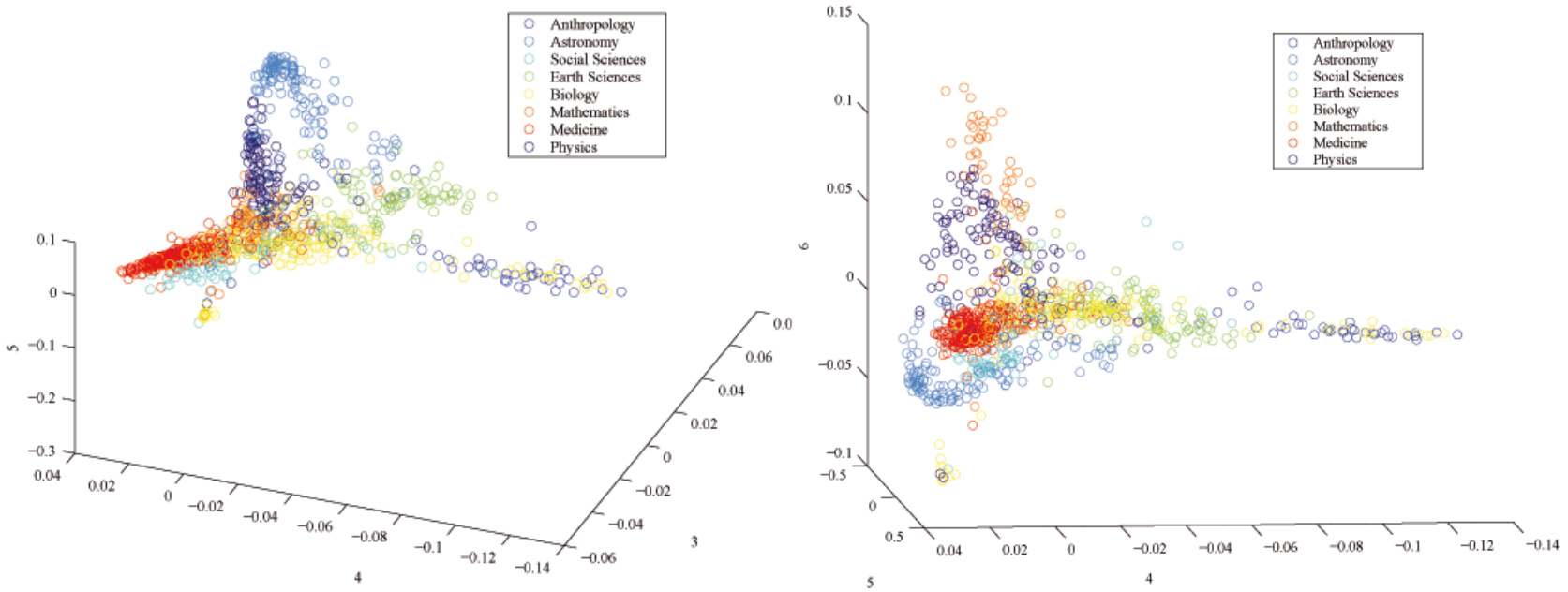
# Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, *i*-th coordinate of document *d* represents frequency in document *d* of the *i*-th word in a fixed dictionary.

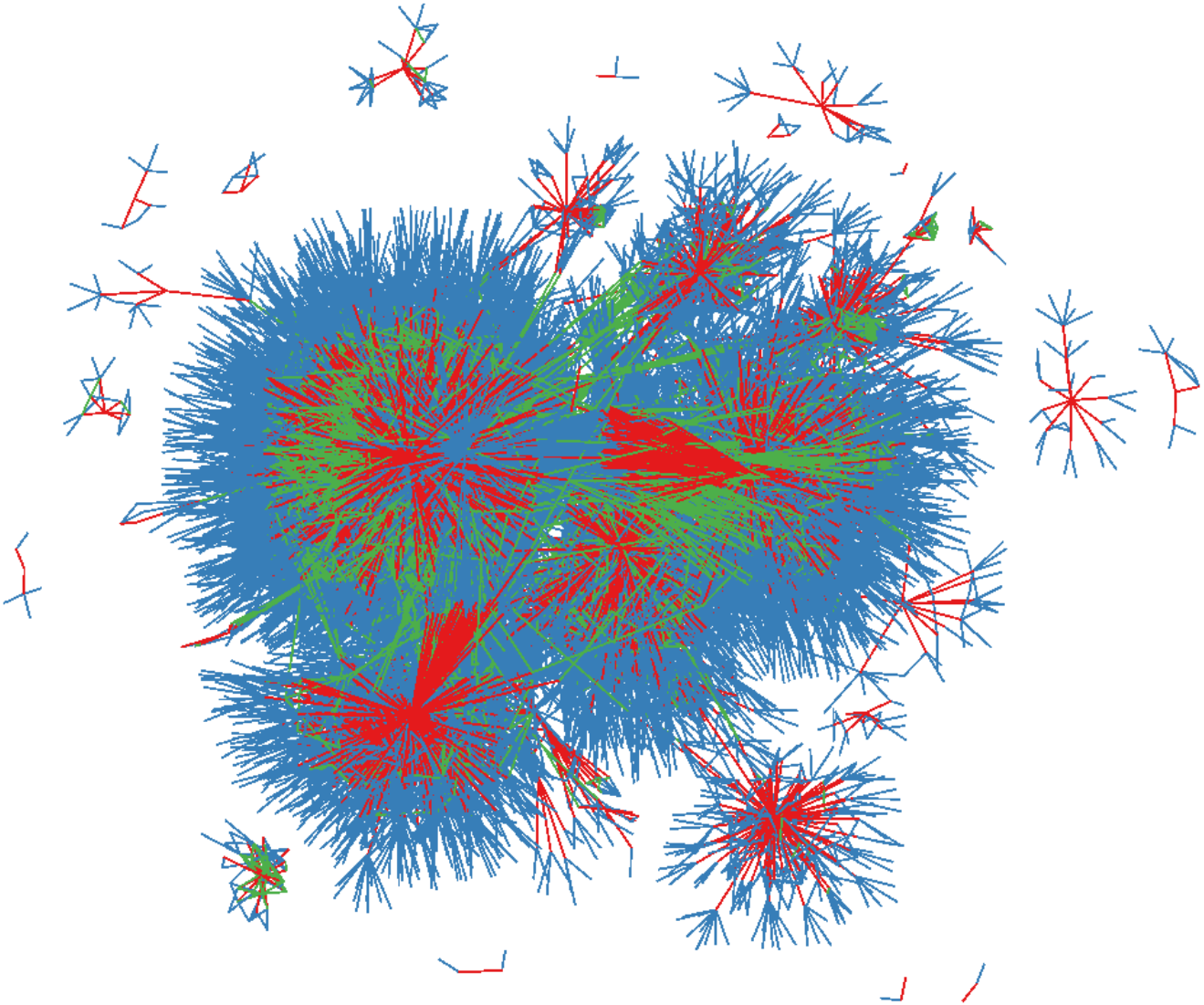


# Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, *i*-th coordinate of document *d* represents frequency in document *d* of the *i*-th word in a fixed dictionary.



# Visualization of large graphs



Example of a patent network of North Carolina biotechnology since 2002. Edges connect patents to people (blue), patents to each other (green), and patents to companies (red). This is a good example of a network which would be much easier to understand and explore using our proposed multiscale representations.

# Multiscale Analysis, a bit more precisely

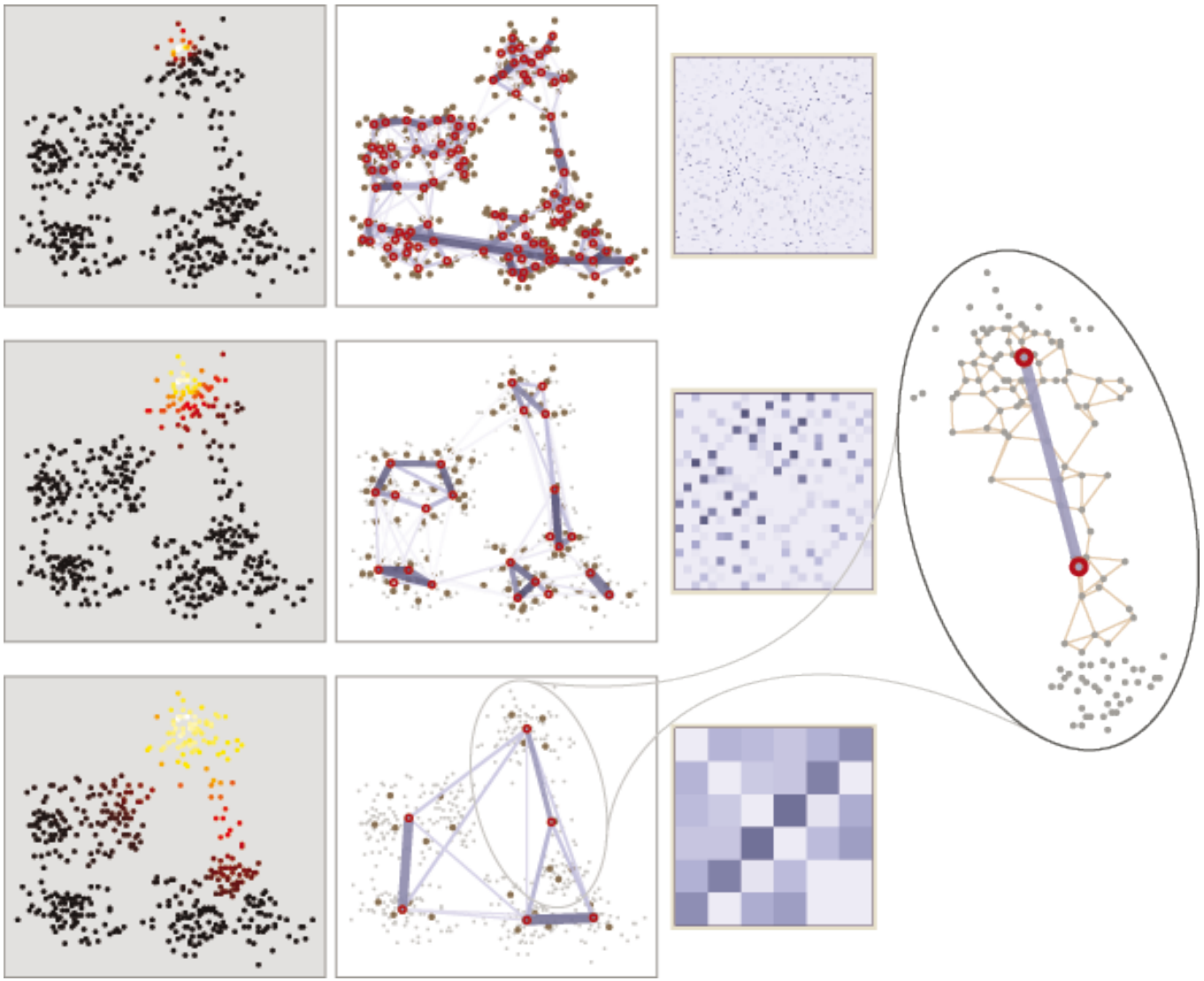
Global embeddings are troublesome - cannot control distortion, interpretability etc...
Recent results show how to localize and obtain good embeddings
We construct multiscale analyses associated with a diffusion-like process $T$ on a space $X$, be it a manifold, a graph, or a point cloud. This gives:

(i) A coarsening of $X$ at different "geometric" scales, in a chain $X \to X_1 \to X_2 \to \cdots \to X_j \ldots$;

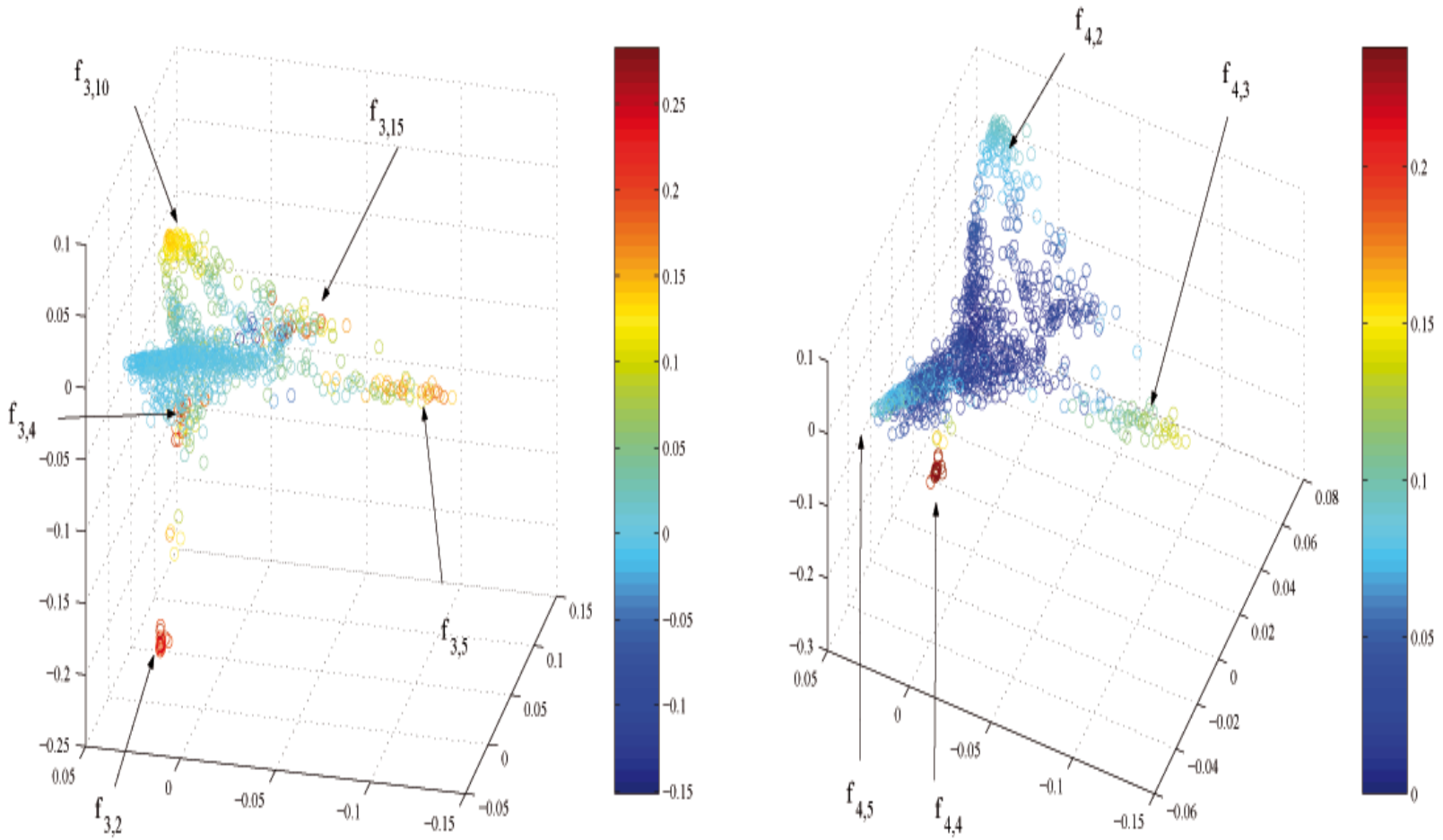
(ii) A coarsening (or compression) of the r.w. $T$ at all time scales $t_j = 2^j$, $\{T_j = [T^{2^j}]_{X_j}^{X_j}\}_j$, each acting on the corresponding $X_j$;

(iii) A set of wavelet-like basis functions for analysis of functions (observables) on the manifold/graph/point cloud/set of states of the system.

# Multiscale Analysis of and on graphs



# Example: Multiscale text document organization




Multiscale clusters on a set of 1100 documents (spectrally) embedded in $\mathrm{R}^3$. $\phi_{3,4}$: Mathematics - networks, encryption, number theory; $\phi_{3,10}$: Astronomy - X-ray cosmology, black holes, galaxies; $\phi_{3,15}$: Earth Sciences - earthquakes; $\phi_{3,5}$: Biology and Anthropology - dinosaurs; $\phi_{3,2}$: Science - talent awards, inventions and science competitions.

# Interactive/iterative learning

- Graph representations may be adapted to user input, e.g. labels or function values
- Recent work shows that even the most elementary implementations of the above ideas leads to state-of-art or better performance in semi-supervised learning tasks
- Relationships between the above models and Gaussian graphical models - uncertainty assessment

# Challenges

- Directed graphs
- Faster algorithms; randomized and online algorithms
- Fast updates and user interaction, making the multiscale representation amenable to easy browsing and labeling
- Sparse multiscale decompositions on graphs
- Time-frequency analysis of dynamic graphs

# Impact on FODAVA

- Mathematically sound tools for multiscale analysis on graphs
- Simplification of navigation of large graphs via multiscale representation
- Integration of active learning and multiscale visualization
- Visualizations adapted to user input in view of specific learning task
- Efficient monitoring of dynamic networks
- Incorporation of uncertainty highlights critical regions and speeds up learning

# Development of FODAVA

- FODAVA applications will be highlighted in publications outside usual fields targeted by FODAVA, in particular applied harmonic analysis as well as mathematically-oriented journals in machine learning, conferences such as VAST and IEEE Visualization.
- Interdisciplinary applications to data sets including patent databases, gene networks, machine learning testbeds, internet traffic
- Consider organizing a highly interdisciplinary workshop (e.g. at IPAM)