

## Multiscale Information Science

Multiscale Information Science is the application of classical and novel multiscale techniques to this largely *man-made* phenomenon: the deluge of information threatening to overwhelm scientists in a wide variety of application areas. While multiscale methods in carrying out computational simulations of physical process are critical to the success of these simulations, a second, higher level of multiscale problems arise as a team of scientists seeks to apply all the information available for insight into a scientific problem, not just that approximate solution to a set of equations. Images and text contribute as well, and the text may actually be in the form of a vast network of publications and clustered at multiple scales by topic. Such text, for example, may include reports of physical experiments and theoretical results relevant to the simulation.

For example the last five years have seen a revolution in our understanding of the topology of large networks relevant to the DOE mission. These networks include the power grid and other infrastructural networks, communication networks, biochemical pathways, networks underlying cellular signaling, protein-protein networks, social networks, as well as many networks underlying models of various complex systems and models of large non-numerical datasets. An explosion in experimental data show that most of these networks display a small-world and scale-free topology that is very different from the classic networks studied by mathematicians and computer scientists. The current challenge in this rapidly-growing field is understanding how topology at multiple scales impacts structure and function, how topology and dynamics ensure system-wide robustness with respect to ubiquitous fluctuations, how to address uncertainty in this framework, and how to build hierarchical models of the dynamics on such networks. A major roadblock to researchers in this area is the absence of software and fast algorithms that will allow multidisciplinary teams to address these issues on modern computing platforms.

Another example that requires analysis on multiple scales is the extraction of new knowledge from a disparate collection of scientific papers. At the lowest level is syntactic and semantic analysis of the raw text which is transformed into mathematical representations in high dimensional space from which relationships among documents can be inferred. These relationships can be aggregated into networks. Analysis of these networks reveals clusters of related scientific results which, when combined with models, simulation, and other information forms, can give insight into the discovery of unexpected results.

Outside of support for simulation and text analysis, multiscale analysis techniques are needed wherever high-level conclusions are required from low-level details. Logs of internet traffic and logs of events in parallel programs both provide masses of data that cannot be usefully interpreted without changes in scale. The mathematical representations, the resulting high dimensional relationships, and visualization all are critical issues of multiscale information science.

### *Multiscale Information Visualization*

Visualization is a critical technology because it provides the broadest bandwidth to the brain for understanding and discovery. Multiscale scientific simulations pose significant visualization challenges due to the differing representations of information on different scales. In addition many informatics problems arise from non-physical sources and have no natural geometry for visualization (e.g. text mining, network analysis, etc.). These challenges are particularly acute with the information science problems have multiscale structure. Visualization methods must be developed for these problems that support interactivity and discovery, while supporting understanding across scales. Without the cooperating multi-scale mathematical and informational representations founding the corresponding visual and interaction representation one cannot calibrate or validate the high dimensional interactions between scales. The core technical challenges that must be addressed within mathematics and information sciences working together include:

- 1) Mathematical representations, data structures & algorithms that enable multiscale simultaneous and linked visualizations and interaction paradigms for discovery and validation,
- 2) Visual and interaction paradigms that represent the user's conceptual model
- 3) Steering analytics for multiscale analysis

### *Dimensionality or Model Reduction*

High dimensionality will be the norm within multiscale science. With the broad number of information sources, including referenced literature, we expect most problems to start at 200 dimensions. However most solutions are discovered as a function of a few of these dimensions that themselves are functions or influenced but further dimensions.

Therefore the core technical challenges are

- 1) The mathematical and informational transformations of information from higher to lower dimensional spaces
- 2) The preservation of high dimensional characteristics and reverse mappings from lower into high dimensional spaces

Such approaches a nonlinear mappings and non-Gaussian statistics are a couple to consider. In many problems the dimensional reduction methodology is the enabling or restricting methods for scientific discovery.

### *Automatic Detection of Coherent Structure in Multiscale Data*

In a solution space we see multiple models, simulations, experiments, and related data spaces being developed by different groups with different study goals. As these become integrated for multiscale analysis a fundamental challenge will be the cross scale mapping/relationships between information structures. These relationships are problem domain specific. Discovering the within-scale problems may involve clustering or pattern analysis. Discovering and representing mathematically relationships cross scale may be manual, semi-manual, or automatic. The end goal is the:

- 1) The automatic discovery of the mathematical and informational relationships and structures between scales

Typical challenges are discovery of dislocation in materials at different scales and different motifs in biochemical networks.

### *Multiresolution Algorithms for Multiscale Analysis*

The use of multiresolution concepts in scientific researches can substantially improve the effectiveness of many costly experiments as well as the quality of the experimental results. The information science community has existing multiresolution algorithms that have already applied in industrial wide applications. For example, direct cosine transforms (DCT) in a JPEG image and discrete wavelet transforms (DWT) in volume visualization. We need to extend this multiresolution concept to go beyond spatial and temporal datasets. These include non-numeric data such as text or DNA sequences:

- 1) New domain-related mathematical techniques to translate non-numeric data into descriptive signature vectors
- 2) New and customized multiresolution techniques to satisfy the needs of different scientific communities

### *Sampling Techniques for Multiscale Phenomena Analysis*

Since most of the resulting models in multiscale information science and the other multiscale application areas have a statistical nature, radical improvement in sampling methods is an essential part of progress in multiscale phenomena analysis and design. The dimension of the phase spaces that needs to be sampled ( $10^6$ - $10^9$ , for networks of man-made and biological origin) in this context is outside of the realm of what has been either practically demonstrated or theoretically inferred for sampling methods. Technical challenges are:

- 1) Effectively using the multiscale structure to define efficient sampling techniques.
- 2) Defining techniques that respect the constraints of the models at all scales (such as complex sampling domains)
- 3) Establishing convergence in distribution of key quantities (such as merit criteria and optimal values) that would allow for efficient quality control

### *Faithful and Consistent Representations Across Scales*

Multiscale models and data formats must ensure *relevant* information is propagated across scales. This information must include the information required to perform the analysis at the target scale, with appropriate error metrics, while presenting it at a level of resolution appropriate for that scale. Currently, simple aggregation or data projections are used, but error metrics are not well understood. Furthermore, transformations that work for small data sets can be misleading when the data sets are large. For example, the MatLab visualization of a large matrix will show most of the cells as occupied even when the matrix is sparse because it maps many matrix cells to each screen location. The core technologies that need to be developed by information scientists and mathematics working together included:

- 1) Development of algorithms to create a hierarchical and lossless data representation

- 2) Development of metrics appropriate for the quantification of errors introduced in both numeric and non-numeric information across scales
- 3) Development of efficient, lossy transformations that can preserve user required information across scales for both numeric and non-numeric information

### *Machine Learning within Multiscale Sciences*

A plethora of information sources exist in the multiscale community including text documents, large data sets from scientific simulations and scientific experiments, results of exercises that compare experimental data and simulations, etc. The magnitude of information available makes the task(s) of inferring and integrating knowledge from all available sources a daunting task(s) at best for human users. Automated machine learning techniques that will support development of decision support systems that help remove some of the burden of discovery from the user are necessary. Such systems will help avoid the tendency to gather the low-hanging fruit and avoid mundane details inferences that likely contain the revolutionary insights. Core technical challenges include:

- 1) Efficient parallel algorithms supporting discovery within and across scales of interest
- 2) Sampling techniques that accelerate convergence for appropriate cost functions relevant for discovery
- 3) Automated learning techniques that help avoid necessity for relearning via automatic creation of knowledge libraries

### *Uncertainty of Multiscale Information*

Uncertainty in working with information sciences data sets will be a major challenge. In addition to many of the problems associated with data derived from physical simulations and experiments, it is quite likely that the information sciences data will have missing, censored, and non-numeric data. Another major challenge will be to quantify the propagation of uncertainty between scales and from dimensionality reduction. We recommend the following core technical advances:

- 1) Uncertainty information representation within scale, mappings and transformations between scales complimenting the mathematics of uncertainty calculations
- 2) Visual analytics of uncertainty across scales and including data types such as experimental data, theoretical models, text, images, graphs,... for the critical needs of uncertainty refinement
- 3) Trust and reliability between scales is a domain and problem specific issue that must be quantified and represented within informational forms