Alla Zelenyuk, PhD

Senior Research Scientist
Chemical Physics & Analysis
Physical Sciences Division
Pacific Northwest National Laboratory

P.O. Box 999 (K8-88)
Tel.:   (509)-371-6155
Richland, WA 99352
Email: alla.zelenyuk-imre@pnnl.gov

Our project (”Chemistry and Microphysics of Small Particles”) is focused on the development and application of unique methods for single particle *multidimensional* characterization, to yield quantitative, comprehensive, real-time information on the properties and transformation of small particles that are ubiquitous in natural and human-made environments. These methods are applied to study particles of interest to basic and applied sciences, including nanoscience, nanotoxicology, mesoscience, atmospheric science, and combustion research.

The behavior and impact of small particles depends on a multitude of their properties, many of which are strongly coupled. The size, internal composition, phase, density, shape, morphology, optical properties, hygroscopicity, activity as cloud warm and ice nuclei, and others - all play a role. Unlike traditional approaches that rely on parallel measurements conducted on heterogeneous mixtures of particles, we simultaneously measure all relevant properties of *millions of individual* particles in real-time.

By necessity this *multidimensional* single particle characterization routinely produces vast amounts of high dimensionality data, on millions of particles, the classification, visualization, mining, and analysis of which calls for unconventional methods that must draw on statistical methods, while preserving the wealth and depth of information._ENREF_100 Moreover, analysis should be based on data generated by all relevant instruments, and include the relationships between them, their temporal evolution, and, for data acquired on aircraft, a way to visualize it all in a geo-spatial context.

The challenges we face are clearly common to other fields that generate massive multidimensional, complex datasets that require matching advances in the science of data organization, visualization, and mining. To that end, we have been developing and applying novel approaches for analysis of large multidimensional complex datasets, in collaboration with Prof. Klaus Mueller (State University of New York at Stony Brook), a specialist in data mining and visualization.

Our unique data visualization and mining program, SpectraMiner,[1] makes it possible to handle data on millions of particles organized into hundreds of clusters, limiting loss of information and thus overcoming the boundaries set by traditional data cluster analysis approaches. SpectraMiner organizes the data and generates an interactive circular hierarchical tree, or dendrogram, providing the user with a visually driven, intuitive interface to easily access and mine the data of millions of particles in *real-time*.

In addition, we developed ClusterSculptor,[2, 3] our expert-driven visual classification software, which uses a novel, expert-steered data classification approach to provide an intuitive visual framework to aid in data clustering. It overcomes the limitation of traditional statistical approaches by offering the scientist the ability to insert their expert knowledge into the data classification.

To visualize and analyze the relationships between multitude of observables and to do so in a geo-spatial context, we recently developed the interactive visual analytics software package, ND-Scope,[4, 5] that is designed to explore and visualize the vast amount of complex, multidimensional data acquired by our single particle mass spectrometers, along with all other aerosol and cloud characterization instruments on-board the aircraft. We demonstrate that the interactive and fully coupled Parallel Coordinates and Google Earth displays of ND-Scope make it possible to visualize the relationships between different observables and to view the data in a geo-spatial context.

Analysis of large data includes data pre-processing (data preparation, integration, reduction, and clustering) that often consumes, as much as 80-90% of the effort. Since data analysis can often be mission critical, pre-processing should be done effectively and fast. Recently we have developed a GPU-accelerated incremental correlation clustering of large data with visual feedback[6] that makes it possible to achieve significant speed-ups over the sequential clustering algorithms. It can be used to detect and eliminate redundant data points, offering a viable means for data reduction and data sampling, and find the outliers or a few “golden nuggets” in vast amounts of data.

Moreover, big data pre-processing, is rarely interactive, which stands in conflict with expert-users, who seek immediate feedback and answers. Our approach provides streaming visual feedback of the clustering process using Multi-Dimensional Scaling dynamic display, allowing interactive visualization and input by the user to fine-tune the clustering parameters and process. Note that visualizing results is also more intuitive than raw text or tabulated data and can increase the chance of spotting patterns that might otherwise go unnoticed.[3]

All software packages we are developing have applications far beyond single particle characterization and are suitable for other large and complex, multidimensional data. While the pre-processing of large data may require the multiple-GPU approach or use of supercomputers, it is highly beneficial to perform visualization and mining of the data in interactive and intuitive manner through visual interface on personal computer. This approach is central to all our software packages, as it puts the user at the center of information flow and decision making.

1. A. Zelenyuk, D. Imre, Y. Cai, K. Mueller, Y. P. Han and P. Imrich, *International Journal of Mass Spectrometry*, 2006, **258**, 58-73.
2. E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk and D. Imre, presented in part at the IEEE Symposium on Visual Analytics Science and Technology, VAST 2007. 2007.
3. A. Zelenyuk, D. Imre, E. J. Nam, Y. P. Han and K. Mueller, *International Journal of Mass Spectrometry*, 2008, **275**, 1-10.
4. Z. Zhang, X. Tong, K. McDonnell, A. Zelenyuk, D. Imre and K. Mueller, *Tsinghua Science and Technology*, 2013, **18**, 111-124.
5. A. Zelenyuk, D. Imre, J. Wilson, Z. Zhang, J. Wang and K. Mueller, *Journal of The American Society for Mass Spectrometry*, 2014, DOI: 10.1007/s13361-13014-11043-13364.
6. E. Papenhausen, B. Wang, S. Ha, A. Zelenyuk, D. Imre and K. Mueller, *IEEE Digital Library, 2013 IEEE International Conference on Big Data*, 2013, 63-70.