The Sky Is Not the Limit*

Erzsébet Merényi

Rice University, Department of Statistics and Department of Electrical & Computer Engineering, 6100 Main Street MS-138, Houston, TX 77005

We live in the era of Big Data, or at least our awareness of Big Data's presence and impact has sharpened in the past ten years. Compared to data characteristics decades ago, Big Data not only means a deluge of unfiltered bytes, but even more importantly it represents a dramatic increase in data dimensionality (the number of variables) and complexity (the relationships among the often interdependent variables, intricacy of cluster structure). Along with the opportunities for nuanced understanding of processes and for decision making, these data created new demands for information extraction methods in terms of the detail that is expected to be identified in analysis tasks such as clustering, classification, regression, and parameter inference. Many traditionally favored techniques do not meet these challenges if one's aim is to fully exploit the rich information captured by sophisticated sensors and other automated data collection techniques, to ensure discovery of surprising small anomalies, discriminate important, subtle differences, and more. A flurry of technique developments has been spawned, many augmenting existing algorithms with increasingly complex features.

Self-Organizing Maps [1] have shown their staying power in the face of these changes and stood out with their simplicity and elegance in capturing detailed knowledge of manifold structures. In our research we have not yet encountered a limit in terms of data complexity. SOMs learn astonishingly well. They are extremely good "listeners" to what the data has to say. An outstanding challenge rather seems to be in equally sharp interpretation of what an SOM has learned.

I will present methods and tools we have developed for deciphering SOMs and for using their knowledge in various ways [2–7]. They are aimed at "precision mining" of large and high-dimensional, complex data, separating important from unimportant details of data characteristics in the presence of noise and some quantifiable degree of topology violation. Components of these tools build on seminal works by several colleagues in the SOM community (e.g., [8–13]), further developing or engineering the original ideas.

I will highlight applications and effectiveness through three types of Big Data: remote sensing hyperspectral imagery for characterizing planetary surface materials, functional Magnetic Resonance Images for brain mapping, and astro-

* This paper uses ALMA data ADS/JAO.ALMA#2011.0.00465.S. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada) and NSC and ASIAA (Taiwan), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

nomical imagery obtained with the world's most advanced radiointerferometric array, ALMA (the Atacama Large Millimeter / Submillimeter Array, in Chile) for answering astrophysical questions ranging from star and planet formation to the formation of the universe.

In this abstract I briefly describe the challenges associated with these representative Big Data and I give a preview of some results we obtained with SOMs and related knowledge extraction approaches.

Hyperspectral images (spectral signatures acquired in hundreds of narrow, contiguous band passes on a regular spatial grid over a target area) have long been utilized for remote geochemical analyses of terrestrial and planetary surfaces. Typical hyperspectral imagery spans the visible to near- and thermal-infrared wavelengths with 5-20 nm band width, sufficient to resolve the discriminating spectral features of (near-)surface compounds. For example, hyperspectral imagery affords identification of individual plant species, soil constituents, the paints of specific cars, and a large variety of roof and building materials, creating a need to extract as many as a hundred different clusters from a single image. These clusters can be extremely variable in size, shape, density, proximities and other properties. Another demand arising from such sophisticated data is to differentiate among clusters that have subtle differences, as the ability to do so can enable important discoveries or increased customization in decision making.

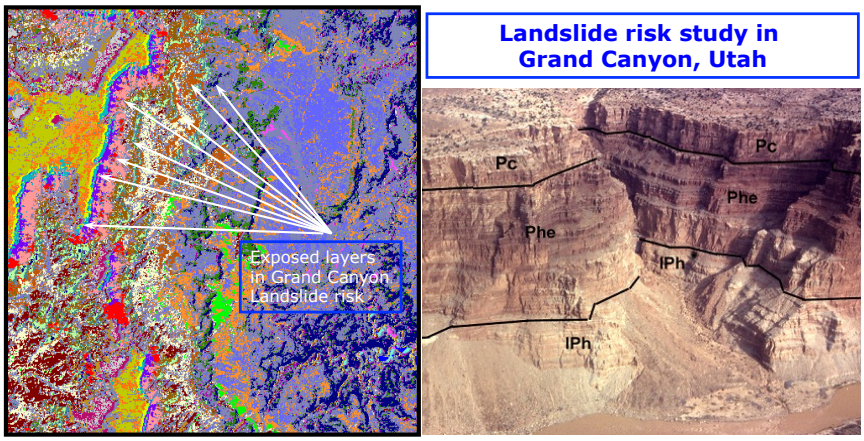

Fig. 1. Mapping clay distribution in soils for landslide risk assessment in Cataract Canyon, Grand Canyon, Utah, U.S.A. **Left:** Classification map produced from a remote sensing hyperspectral image. 15 of the 28 classes (each indicated by a different color) are exposed soil layers, several of which are indicated by the white arrows. **Right:** Photograph of some of the soil layers, in part of the imaged site. Figure adapted from [14].

For example, landslide risk models can be greatly improved by including (in addition to the traditional factors of mountain slope and rain fall) the types and amounts of clay minerals contained in the exposed soil layers of mountains. For assessment of large areas remote sensing is used, which detects the clay minerals highly diluted in the soil matrix, resulting in weakened signatures. To distinguish

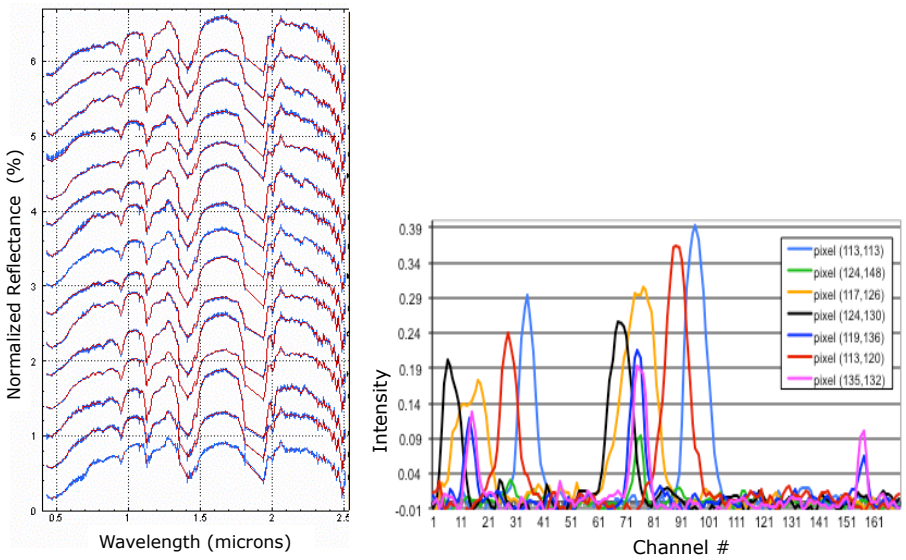

Fig. 2. Left: Visible-Near-Infrared hyperspectral signatures of clay-bearing soil classes, mapped in Fig. 1. Spectra are vertically offset for viewing convenience. The variations in the spectral window most discriminating for clay species (0.5 – 0.7 and 2.0 – 2.3 microns) are very subtle. Blue and red curves are the means of training and test samples for the classification experiment in [14], with standard deviations shown by the vertical bars. **Right:** Sample emission spectra, from combined C18O, 13CO, CS lines of ALMA receiver band 7, showing differences in composition, Doppler shift, depth and temperature. 170 channels were stacked from the C18O, 13CO, CS lines. Data credit: JVO, project 2011.0.00318.5.

and map the 15 or so layers of different soils around a landslide area in the Grand Canyon (Fig. 1) hyperspectral signatures with such slight variations as in Fig. 2 must be discriminated precisely by a classifier and produce maps showing the spatial distribution of the various soils, as in Fig. 1. An SOM was instrumental in accomplishing the delicate task [14].

In stellar astronomy, where Ångström resolution is typical, the data complexity can grow even higher. 21st century observatories such as ALMA achieve, for the first time, data sets that begin to approach, and in some dimensions exceed, the richness of data from terrestrial and planetary remote sensing. High spatial and spectral resolution image cubes with thousands of frequency channels are extending into new and wider wavelength domains, and at the same time capturing several different physical quantities that characterize 3-dimensional plasma structures. The "spectra" are no longer vectors of homogeneous variables. Effects of spatial depth, Doppler shift, temperature and densities are influencing the signatures in addition to chemical composition. Fig. 2, right, gives an illustrative sample of ALMA data, combined from three different emission lines. The left image in Fig. 3 shows structural details of a protostar produced (to my knowledge) by the first SOM clustering of a complex ALMA image cube [15], in comparison to details extracted from a single doppler line by [16]. This protostar

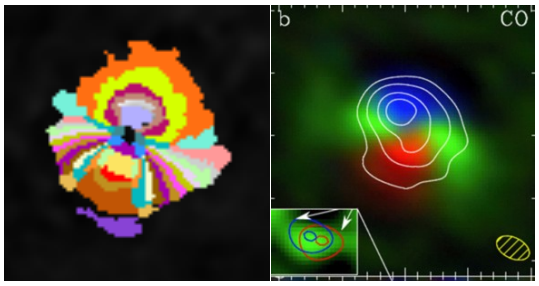

Fig. 3. Structure found in protostar HD 142527 from ALMA data. **Left:** SOM clustering from hyperspectral ALMA data cube by the author. **Right:** From single doppler line, by [16]. Figure reproduced with permission. Data credit: JVO, project 2011.0.00318.5.

has stirred great interest recently because of a planet formation process that has been detected deep in its interior.

Functional Magnetic Resonance Imagery (fMRI) poses many similar challenges as hyperspectral data, with typically higher-dimensional data vectors and potentially more clusters. The time courses — vectors of measurements of blood-oxygen-level dependece (BOLD) signals at hundreds of time points recorded during the observation of a subject at each of several hundred thousands of voxels in the brain volume — can be clustered to find brain areas with similar activation patterns. Correlation analysis of the characteristic time courses of the identified clusters can further reveal temporal relationships of various sub-networks in the brain. With SOM tools we can glean detailed maps of the entire brain with more complete coverage than seen in many published results.

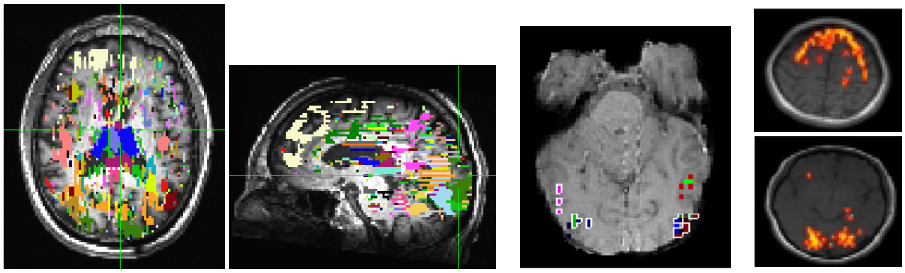

Fig. 4. Clusterings of fMRI images, based on the BOLD time courses. **Left pair of images:** SOM clusters obtained with our tools, in two selected brain slices, showing good coincidence with and coverage of known functional regions such as the thalamus (dark blue), insula (mauve symmetrically placed spots on either side of the thalamus in the left image), visual cortex (light blue, dark green and orange), and superior frontal gyrus (light yellow, at top of the left image, at front left in the right image) [17]. Data credit: The Methodist Research Institute, Houston, Texas. **Center:** Clusters generated by statistical hypothesis testing, from [18]. **Right:** Clusters found in the motor cortex (top) and visual cortex (bottom) by [19] using SOMs. Figures from [18] and [19] reproduced with permission.

Fig. 4, left, shows a pair of representative brain slices with our recent SOM clustering [17], for which all available voxels were used from the entire brain volume. The clusters coincide well with several known functional areas throughout all slices (not shown here). In comparison, clustering with statistical methods in [18] (center) was applied only to two selected slices, and the clusters identified are highly segmented, with very sparse coverage. The brain maps on the right, from [19] were obtained by SOM clustering, and have good coverage of selected functional areas. An important property in the face of Big Data, SOMs are not nearly as limited by large data volumes as many other methods (for example, graph-based clustering, where the number of vertices grows quadratically with the number of data vectors). The ability of learning well from large volumes of data allows precise identification of a large variety of functional regions, which in turn enables more nuanced investigation of such fundamental questions as — in our study — the generation of the conscious movement in healthy and impaired brains.

SOMs arguably provide a key to accurate learning of diverse types of highly structured data. However, with this power come new puzzles. While sharpening knowledge extraction methods to match the richness of the data, we must also recognize that interpretation of the increasing detail emerging from data like these may be the next challenge in the Big Data picture. Sophisticated tools that allow penetration of previously unidentified relationships in the data may return "distilled" results that look complicated, hard-to-digest, and not straighforward to interpret or verify.

ALMA, for example, represents such advanced observational capability that fully exploiting the information content will require — as much as anything else — new capabilities to synthetize, visualize, and interpret the extracted knowledge, the already summarized information! The cluster map on the left in Fig. 3, for example, can only show part of the protostar structure detected by the SOM. We yet have to devise a visualization to layer on and meaningfully convey the full information. In closing, I will illustrate some cases of this interesting problem.

Acknowledgments

Special thanks to ALMA project scientist Al Wootten, for sharing and helping with ALMA data. Collaboration with Drs. Robert Grossman and Christof Karmonik at the Methodist Research Institute, Houston, Texas, on the analysis of their fMRI data is gratefully acknowledged.

References

1. Kohonen, T.: Self-Organizing Maps. Second edn. Springer-Verlag Berlin Heidelberg (1997)
2. Zhang, L., Merényi, E., Grundy, W.M., Young, E.Y.: Inference of surface parameters from near-infrared spectra of crystaline h2o ice with neural learning. Publications of the Astronomical Society of the Pacific **122**(893) (February 2010) 839–852 DOI: 10.1086/655115.

3. Merényi, E., Tasdemir, K., Zhang, L.: Learning highly structured manifolds: harnessing the power of SOMs. In Biehl, M., Hammer, B., Verleysen, M., Villmann, T., eds.: Similarity based clustering. Lecture Notes in Computer Science, LNAI 5400. Springer-Verlag (2009) 138–168
4. Tasdemir, K., Merényi, E.: Exploiting data topology in visualization and clustering of Self-Organizing Maps. IEEE Trans. on Neural Networks **20**(4) (2009) 549–562
5. Mendenhall, M., Merényi, E.: Relevance-based feature extraction for hyperspectral images. IEEE Trans. on Neural Networks **19**(4) (April 2008) 658–672
6. Merényi, E., Jain, A., Villmann, T.: Explicit magnification control of self-organizing maps for "forbidden" data. IEEE Trans. on Neural Networks **18**(3) (May 2007) 786–797
7. Zhang, L., Merényi, E.: Weighted Differential Topographic Function: A Refinement of the Topographic Function. In: Proc. 14th European Symposium on Artificial Neural Networks (ESANN'2006), Brussels, Belgium, D facto publications (2006) 13–18
8. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Networks **15** (2002) 1059–1068
9. Villmann, T., Der, R., Herrmann, M., Martinetz, T.: Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. IEEE Transactions on Neural Networks **8**(2) (1997) 256–266
10. Bauer, H.U., Der, R., Herrmann, M.: Controlling the magnification factor of self–organizing feature maps. Neural Computation **8**(4) (1996) 757–771
11. Martinetz, T., Schulten, K.: Topology representing networks. Neural Networks **7**(3) (1994) 507–522
12. Ultsch, A., Simeon, H.P.: Kohonen's self organizing feature map for exploratory data analysis. In: Proc. INNC-90-PARIS **I**, Paris (1990) 305–308
13. DeSieno, D.: Adding a conscience to competitive learning. In: Neural Networks, 1988., IEEE International Conference on, IEEE (1988) 117–124
14. Rudd, L., Merényi, E.: Assessing debris-flow potential by using AVIRIS imagery to map surface materials and stratigraphy in cataract canyon, Utah. In Green, R., ed.: Proc. 14th AVIRIS Earth Science and Applications Workshop, Pasadena, CA (May 24–27 2005)
15. Merényi, E.: Hyperspectral image analysis in planetary science and astronomy. Presentation in Special Session "Building the Astronomical Information Sciences: From NASA's AISR Program to the New AAS Working Group on Astroinformatics and Astrostatistics. 223rd AAS meeting, Washington, D.C., January 7 2014.
16. Casassus, S., van der Pas, G., Perez M., S., et al.: Flowes of gas through a protoplanetary gap. Nature **493** (January 2013) 191
17. O'Driscoll, P.: Using Self-Organizing Maps to discover functinal relationships of brain areas from fMRI images. Master's thesis, Rice University (June 2014)
18. Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y.: Cluster-based analysis of FMRI data. NeuroImage **33**(2) (November 2006) 599–608 PMID: 16952467.
19. Liao, W., Chen, H., Yang, Q., Lei, X.: Analysis of fMRI data using improved self-organizing mapping and spatio-temporal metric hierarchical clustering. IEEE Transactions on Medical Imaging **27**(10) (2008) 1472–1483