*Astroinformatics*
*Proceedings IAU Symposium No. 325, 2016*
*M. Brescia, S.G. Djorgovski, E. Feigelson,*
*G. Longo & S. Cavuoti, eds.*


doi:10.1017/S1743921317000175

# Deep data: discovery and visualization Application to hyperspectral ALMA imagery

**Erzsébet Merényi[1,2], Joshua Taylor[1] and Andrea Isella[3]**

[1]Dept. of Statistics
[2]Dept. of Electrical and Computer Engineering
[3]Dept. of Physics and Astronomy
Rice University, 6100 Main Street, Houston, Texas 77005, U.S.A.
email: {erzsebet,jtay,isella}@rice.edu

**Abstract.** Leading-edge telescopes such as the Atacama Large Millimeter and sub-millimeter Array (ALMA), and near-future ones, are capable of imaging the same sky area at hundreds-to-thousands of frequencies with both high spectral and spatial resolution. This provides unprecedented opportunities for discovery about the spatial, kinematical and compositional structure of sources such as molecular clouds or protoplanetary disks, and more. However, in addition to enormous volume, the data also exhibit unprecedented complexity, mandating new approaches for extracting and summarizing relevant information. Traditional techniques such as examining images at selected frequencies become intractable while tools that integrate data across frequencies or pixels (like moment maps) can no longer fully exploit and visualize the rich information. We present a neural map-based machine learning approach that can handle all spectral channels simultaneously, utilizing the full depth of these data for discovery and visualization of spectrally homogeneous spatial regions (spectral clusters) that characterize distinct kinematic behaviors. We demonstrate the effectiveness on an ALMA image cube of the protoplanetary disk HD142527. The tools we collectively name "NeuroScope" are efficient for "Big Data" due to intelligent data summarization that results in significant sparsity and noise reduction. We also demonstrate a new approach to automate our clustering for fast distillation of large data cubes.

**Keywords.** neural machine learning, clustering, ALMA, protoplanetary disks

## 1. Deep data and challenges

The currently most advanced radio telescope, the Atacama Large Millimeter and sub-millimeter Array (ALMA), and next-generation telescopes produce data not only with unprecedented volume but also with unprecedented complexity. ALMA opened an era where hyperspectral data cubes are becoming the norm in radio and millimeter observations. Spatially resolved images of a source are simultaneously recorded in many different molecular lines, each line resolved by dozens to hundreds of spectral (velocity) channels. This offers a magnifying lens for our understanding of the physical conditions (temperature, density, and kinematics) of atomic and molecular gas, as well as of the distribution of solid particles, in objects such as protoplanetary disks, molecular clouds, interstellar medium, nearby galaxies, and more. Current data analysis approaches, however, are underutilizing this rich information. In particular, the "depth" of the data — the detailed spectral information — often forces either prior dimensionality reduction or integration across spectral channels causing loss of potentially critical detail for discovery.

Traditionally, two approaches have been used for the extraction of physical (velocity) structure from 3-D (3-dimensional) data cubes generated by radio and millimeter telescopes like ALMA. One is to visually inspect simultaneously displayed images of each

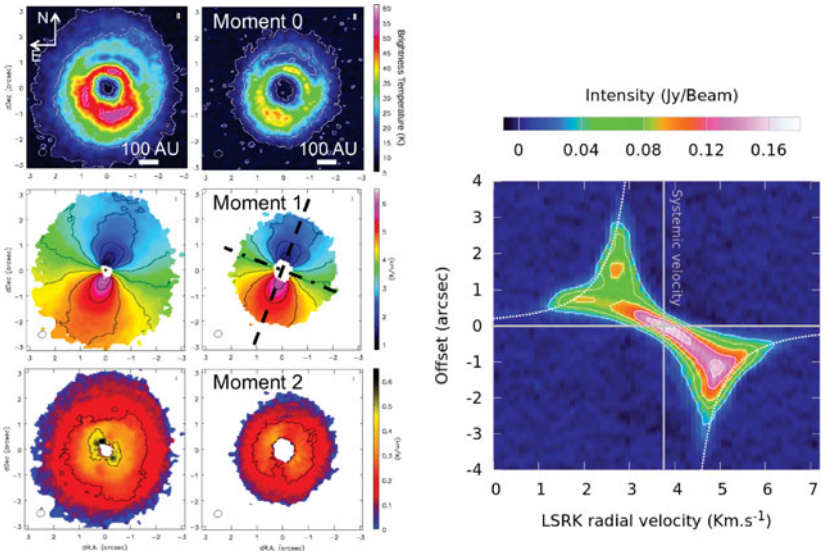


**Figure 1.** From (Boehler *et al.* 2016). **Left:** Traditional moment maps generated separately from two molecular lines, $^{13}$CO J=3-2 at left, and C$^{18}$O J=3-2 at right, from high-resolution observations of the protoplanetary disk HD142527. The Moment 0 map shows the intensity of the line emission, expressed in units of brightness temperature which corresponds to the gas kinetic temperature if the line emission is optically thick, as in the case of HD142527. The moment 1 map shows the velocity of the emitting gas relative to the observer. The dot-dashed line indicates the rotation axis of the disk, and the dashed line shows the apparent major axis of the disk. The moment 2 map shows the width of the line emission. **Right:** Position-Velocity diagram for the $^{13}$CO J=3-2 line emission measured toward HD142527. The x-axis shows the offset with respect to the center of the disk measured along the major axis of the disk. The y-axis shows the velocity along the line of sight of the emitting gas relative to the observer. The white dotted curves indicate the expected velocity for gas rotating at Keplerian velocity around the central star. The vertical line shows the systemic velocity, i.e., the velocity along the line of sight of the star+disk system relative to the observer.

spectral channel (a "channel map") within a spectral line. A second technique is to integrate the data along one dimension of a single line to calculate the so-called "intensity moments", at all pixel locations. The first three intensity moments correspond to the spectrally integrated intensity (moment 0), the velocity corresponding to the center of the line (moment 1), and the width of the line emission assumed to have Gaussian shape (moment 2). The moment 0 image maps the spatial distribution of the emitting gas, the moment 1 and moment 2 images, respectively, inform about the motion of the gas on spatial scales larger and smaller than the spatial resolution of the observations. Alternatively, 3-D data cubes can be integrated along one spatial dimension. This leads to the so-called "position-velocity" diagrams, which are sensitive to kinematical properties of the gas such as inflows or outflows.

Figure 1, left, shows the moment maps of the $^{13}$CO J=3-2 and C$^{18}$O J=3-2 line emission and the position-velocity diagram for the $^{13}$CO line of the protoplanetary disk HD142527. This object is particularly interesting because the bright red and yellow ring in the moment 0 maps is thought to have been formed by a newborn planetary system inside the dark blue center of about 100 AU diameter (Isella *et al.* 2013). In addition, the moment 1 and velocity-position maps reveal that the general motion of the gas is consistent with Keplerian rotation around a star twice as massive as the Sun. The moment 2 map indicates a decrease in line width with the distance from the star. This is due to a decrease in both the gas velocity and temperature with radial distance.

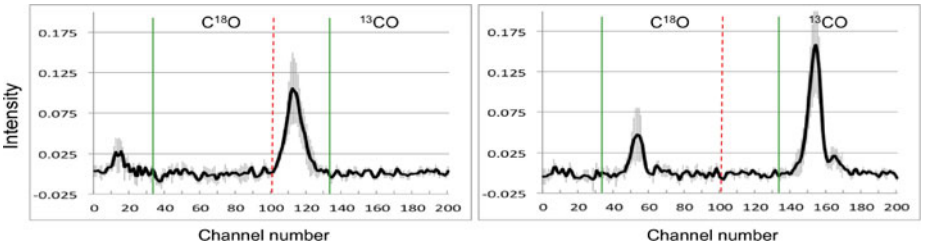


**Figure 2.** Sample spectral signatures (means of small areas) from an ALMA data cube at two different spatial locations of the protoplanetary disk HD142527. The cube comprises 100 channels from each of two molecular lines (a total of 200 channels) stacked at the red dotted line in the order of increasing frequency. The green vertical lines show the channel of the rest frequency for each of the molecular species. The shift of the peaks relative to the rest frequency (velocity difference or Doppler shift, between observer and source) is different in the two samples. The radiation intensity (height of the peaks), as well as the relative difference in the peak intensities between the two species is also different at the two locations. Notably, while the $^{13}$CO line at left has a simple shape, the $^{13}$CO line of about the same width at right has a double peak. These variations exemplify some of the complexity of the structure of the gas across the disk. Data from (Boehler *et al.* 2016). The 200-element stacked spectra from individual pixels are the 200-D input vectors to our clusterings, straight after the standard ALMA reduction.

For precursors of ALMA, moment maps and position-velocity diagrams were suitable to visualize the main characteristics of the molecular line emission, at least for simple objects such as protoplanetary disks. However, moment maps do not scale up well to the much more complex ALMA cubes. In particular, they might lead to erroneous conclusions if the gas kinematics along a line of sight cannot be represented as a Gaussian line, as in the case of gas moving at different velocities. At the same time, with potentially dozens of spectral lines consisting of several thousand channels, visual analysis of channel maps becomes infeasible, especially across multiple spectral lines.

Figure 2 shows sample spectra from a high-resolution ALMA data cube of HD142527 at two different spatial locations and in two molecular lines, C$^{18}$O and $^{13}$CO, each line comprising 100 channels concatenated along the spectral axis. The lines have an approximate width of 6 MHz and a spectral resolution of about 65 KHz per channel (0.11 km/sec velocity resolution). Within each line, the apparent shift of the peaks relative to the rest frequency at which the observed source emits radiation at rest in the observer's reference frame, indicates that the relative velocity of the gas varies across spatial locations due to the disk rotation. The structure of the line (width, height and number of peaks) can vary by spatial location and by species. For example, while the $^{13}$CO line at left has a simple Gaussian shape, the one at right shows a main peak around channel 152 and a second faint peak around channel 165, which would be missed by fitting a Gaussian. The stacked full spectral signatures capture the complex variations of the combined compositional and kinematic properties which may not be expressed by reduced-dimensionality data or by single molecular lines. Our objective is to extract and map these variations within a spatially resolved source, and to produce results efficiently for on-board analyses or for discovery in large archives.

*Recent methods to improve on the moment analysis* include Clumpfind (Williams *et al.* 1994), Cloudprops (Rosolowsky & Leroy 2006), dendogram object identification (Houlahan & Scalo 1992; Rosolowsky *et al.* 2008), and Discrete Persistent Structures Extractor (DisPerSE) by Sousbie (2013).

Some rely on fitting Gaussian distributions (e.g., Clumpfind), limiting discovery to simple structures. Some work well for finding particular structures (e.g., filaments in the

case of DisPerSE) but do not generalize to different ones. Others (as the dendrogram analysis) assume uniform and well-characterized noise.

## 2. NeuroScope for structure discovery

We demonstrate a clustering / pattern recognition approach with *NeuroScope*, a set of neural machine learning tools, for identification of spatial regions that exhibit distinctly coherent kinematic and compositional behavior based on the full spectral signatures of (possibly) combined emission lines. This approach is insensitive to most limitations discussed above therefore more robust and flexible for analysis of ALMA data. The emerging details indicate fuller exploitation of the rich ALMA data than by the traditional methods. In this study the input vectors to all analyses are the 200-D stacked spectra as in Figure 2. The data have undergone the standard ALMA data reduction to correct for atmospheric and instrumental effects (Boehler *et al.* 2016) but no other preprocessing.

*Cluster discovery with NeuroScope tools* involves learning the n-D data manifold with advanced variants of Self-Organizing Maps (SOMs), and a recent similarity measure that facilitates interpretation of the SOM's knowledge based on *connectivity* properties of the data manifold. This allows deeper exploitation of relevant details than customary uses of SOMs and common similarity metrics, resulting in sensitive distinction of clusters.

The SOM (Kohonen 1988) is an unsupervised neural network which mimics the information summarization and organization of cortical areas in natural brains. The SOM consists of a rigid (usually 2-D) lattice of artificial neurons each of which is connected to an input layer of n neurons by an n-D weight vector, also called a *prototype* vector. During SOM learning the prototypes are moved in the data space to reflect the distribution (the *pdf*) of the data. At the same time they are ordered on the SOM lattice according to their similarity relations in data space. This *topology-preserving* mapping expresses an intelligent summarization of both the statistics (the n-D density distribution) and the topology of the data manifold. It facilitates discovery of clusters (groups of similar patterns, e.g., similar spectra) from a learned SOM by evaluating the similarity relationships of prototypes neighboring in the SOM grid, and segmenting the SOM into groups of similar prototypes (data points mapped to a prototype cluster make up a data cluster). This can be very challenging for data with complex structure, and success depends on the quality of manifold learning and on the expressiveness and sensitivity of the representation of prototype relationships.

For learning the data manifold we use the Conscience SOM (CSOM) by DeSieno (1988) which produces more faithful *pdf* matching than the Kohonen SOM (KSOM) by inducing equiprobabilistic (maximum entropy) mapping. For high-dimensional input data this was verified by Merényi *et al.* (2007). The CSOM introduces a bias to achieve equal winning probabilities across all neural units. Briefly, the prototype (weight) vector $\mathbf{w}_i$ of neural unit $i$ in the SOM lattice $A$ of $NP$ neural units, is updated iteratively at every time step $t$ as follows. First, a winner neuron (or best matching unit, BMU) $i$ is selected for a random input vector $\mathbf{x} \in \mathcal{R}^{\mathbf{n}}$ such that with bias $b_j$ for neuron $j$

$$\| \mathbf{w}_i - \mathbf{x} \|^2 - b_i \leqslant \| \mathbf{w}_j - \mathbf{x} \|^2 - b_j, \forall j \in A. \tag{2.1}$$

The bias $b_j$ is computed from the winning frequency $p_j$, of neural unit $j$, as

$$b_j = \gamma(t) \times ((NP \times p_j) - 1), \tag{2.2}$$

where $\gamma$ is a parameter. Second, all prototypes $\mathbf{w}_j$ are updated:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t) h_{i,j}(t)(\mathbf{x} - \mathbf{w}_j(t)) \tag{2.3}$$

Here, $h_{i,j}(t)$ is a neighborhood function, $\alpha(t)$ is the learning rate. For the CSOM $h_{i,j}(t)$

can be fixed and of small size (e.g., the immediate SOM neighbors) instead of a large neighborhood (e.g., Gaussian) that has to decrease with time in a KSOM. This provides significant computational savings. With $\gamma = 0$ and large time-decreasing $h_{i,j}(t)$ the above algorithm reverts to KSOM. The learning (as SOM learning in general) does not require a pre-specified number of clusters, and needs very little parameter tuning. For successful and correct identification of clusters from a learned SOM, the correctness of the manifold learning and the quality of cluster extraction are critical. Many of these issues have been addressed by Merényi *et al.* (2009) and Tasdemir & Merényi (2011), and references therein. These include measures of topology preservation, to separate serious topology violations from those that are inconsequential for cluster detection and can be ignored.

To delineate complex cluster structure we use the CONN measure by Tasdemir & Merényi (2009) which expresses connectivity (hence the name CONN), rather than data space distances, of the prototypes. The pair wise connectivities — the CONN *graph* — can be visualized over the SOM lattice regardless of the data dimensionality, providing a view of the salient properties of the manifold structure as in Figure 3, top of left panel for the protoplanetary disk HD142527. On the lattice, black dots mark the positions of neurons. A cell with no dot has no data vectors mapped to it (has an empty prototype). The thickness of the line segments between two prototypes signifies the *absolute strength* of their connection. The connection strength, $CONN(i, j)$, between prototypes $\mathbf{w}_i$ and $\mathbf{w}_j$, is measured as the number of data vectors that fall closest to $\mathbf{w}_i$ or $\mathbf{w}_j$ and second closest to the other one, after the SOM has converged. Colors indicate the *relative importance* of the connections to other prototypes. Red is most-connected, followed by blue, green, yellow, and grey shades (not present for this data). The combination of global connection strengths and their local ranking provides rich information about where the manifold is strongly woven and where it is disconnected or thin. This representation and visualization is done automatically and it is the input to clustering the SOM prototypes.

Cluster boundaries are found between regions that are strongly connected inside and have thin or no connections to other regions. Default parameters for filtering unimportant connections (inconsequential for clustering), and for a non-linear binning of the connection strengths to aid the human eye, are automatically computed from CONN statistics and applied. Details on this and a cluster extraction procedure are given in Tasdemir & Merényi (2009) and Merényi *et al.* (2009). The clusters of similar prototypes extracted interactively from the CONN graph representation are shown in Figure 3, bottom of left panel, while the corresponding clusters of pixels in the disk of HD142527 are at top right.

## 3. Comparison with the state-of-the-art

Figure 3, top of right panel, shows the cluster structure found in the protoplanetary disk HD142527, using the combined (stacked) molecular lines $C^{18}O$ and $^{13}CO$ (as in Figure 2) as input data vectors. The first thing to notice is that the cluster map — similarly to the moment 1 map — reveals regions of the disk moving at similar velocities. Since the gas kinematics are dominated by rotation around the star, the cluster map is roughly symmetric with respect to the axis of rotation of the disk (the dot-dashed line on the moment 1 image). Looking more carefully, clusters at two symmetric positions with respect to the minor (NW-SE) axis show an asymmetry in intensities, in agreement with the moment 0 map. See, e.g., those indicated by the arrows (labeled W and N, two different brown shades in the map) at the top of the right panel of Figure 3. The cluster means additionally reveal that the relative intensity change for the two gas species is also different at the two symmetric locations.

The neural map-based clustering is superior to the moments visualization in identifying deviations from Keplerian rotation. For example, the bottom right panel in Figure 3 shows

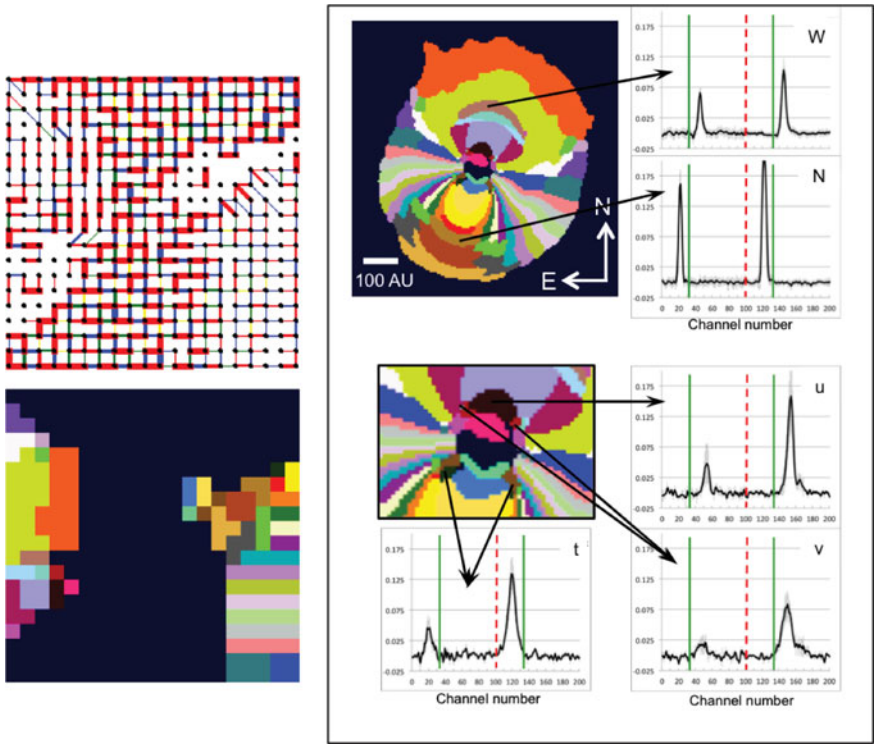


**Figure 3. Left panel, top:** SOM lattice of 20 x 20 neurons, with CONN graph representation of the learned manifold structure of HD142527. **Left panel, bottom:** SOM lattice with clusters of similar prototypes extracted interactively from the CONN graph representation. Each color represents a different similarity group of prototypes. The colors are chosen for contrast and do not represent relative similarities. The largest and also strongest-connected cluster (dark blue) comprises the sky background. While visually overwhelming in the CONN graph at left, it does not affect the extraction of smaller clusters with more subtle differences, by analysis of the local relations. **Right panel, top:** Visualization of the spectral clusters — showing coherent regions of distinct kinematic and compositional properties — in the disk of HD142527, where each pixel is colored as its prototype in the SOM lattice (at left, bottom); and cluster signatures indicating NW-SE asymmetry. Red and green lines are as in Figure 2. **Right panel, bottom:** The magnified center region surrounded by the mean signatures of three interesting clusters. Cluster u, and the tiny whisker-like clusters v and t flanking the arch-shaped clusters g (hot pink, North of the center) and j (turquoise, South of the center), respectively, exhibit double or widened peaks in the $^{13}$CO line. These can indicate deviation from the Keplerian motion.

clusters (u, and t) characterized by a double peak profile, or a widened peak (cluster v) in the $^{13}$CO line. The main peak in cluster u (at channel 153) arises from gas moving at 2.2 km/sec relative to the star in the direction of the observer, and traces gas orbiting the central star at Keplerian velocity. The second, minor peak (at channel 165) arises from gas moving at 3.5 km/sec relative to the star in the direction of the observer, and might arise from gas blown away by a stellar wind. Although low-intensity, the second peak emerges cleanly in the average cluster signature with very small standard deviations. Similarly for cluster t (the pair of small downward pointing whisker-like dark green features marked by arrows at either sides of the arch-shaped turquoise cluster), where the main and minor peaks (at channels 120 and 148, respectively) are more distinct and have larger velocity difference (−1.43 km/sec vs 1.7 km/sec) than in cluster u. Here the two velocities have opposite signs, which may be caused by two gas components moving

in opposite directions. While the moment 2 map indicates a widening of the line at the location of clusters u and v, it does not provide information of the shape of the line. Furthermore, the moment 2 fails to highlight the location of cluster t. This example shows how NeuroScope clustering can enlarge the discovery space by fully exploiting the transformational imaging capabilities provided by current telescopes.

There are important advantages of this clustering analysis compared to the traditional moment visualization. First, the capability of combining multiple lines at once allows one to identify correlation and anticorrelation between different gas tracers. Second, clustering analysis naturally combines signals from similar regions augmenting the capabilities to see faint structures. Third, this — data-driven — technique does not assume priors for the line emission and therefore delivers an unbiased interpretation of the observations. SOM learning is also robust because the summarization of data by the prototypes reduces noise, and because the topological ordering of the prototypes facilitates the preservation of subtle differences despite possibly small vector distances between spectra. A brief discussion of some complementarities and differences with leading clump finder and dendrogram methods are given in Merényi *et al.* (2016).

## 4. Automation approach to SOM segmentation

The first step of clustering, SOM learning, requires little tuning, and it is efficient for large data sets. By intelligent summarization, it shrinks the data volume by magnitudes while retaining the manifold characteristics relevant for cluster discovery and preserving the spectral resolution in the prototypes. As an example, the protoplanetary disk in Figure 3 comprises approx. 56,000 pixels. These are characterized by 400 SOM prototypes, a factor of 140 down-sizing. While SOMs are slow on sequential machines, parallel implementation (e.g., in Field Programmable Gate Arrays) eliminates this problem (Lachmair *et al.* 2013). The second step, SOM segmentation, is currently more successful for complex data structure when done interactively from expressive visualizations such as in Figure 3, left. However, scalable processing requires automation. Graph segmentation algorithms have been proposed for automated clustering of "Big Data", but their use requires huge computational resources since they process a graph with $N$ vertices (representing data points) and $N^2$ edges (representing some pairwise point similarity). To remedy this, we apply these segmentation algorithms to a graph representing the SOM prototype structure (where approximately $\sqrt{N}$ prototype vectors generate $N$ edges), using the CONN values as point similarities (edge weights). As we show below, and summarize in Figure 4, this combination significantly enhances the graph segmentation algorithms' ability to find relevant clusters in minimal computing time.

*Graph segmentation overview.* Graph segmentation aims to identify a subgraph structure such that each subgraph is densely connected within itself and sparsely connected to other subgraphs. Decades of research in this area have resulted in many different algorithms (see the review from Fortunato 2010). While we experimented with a number of leading algorithms, the `Walktrap` (Pons & Latapy 2005) and `Infomap` (Rosvall & Bergstrom 2008) algorithms identified the most meaningful cluster structure of the ALMA data relative to the interactive clustering in Figure 3. Both methods are freely distributed in the `igraph` package (Csardi & Nepusz 2006) and require negligible runtimes ($\approx$ 1 sec on an ordinary MacBook Pro for the SOM-based graphs). Importantly, both require only one tuning parameter. For algorithm details we refer the reader to the original papers cited above. A short overview is given in Merényi *et al.* (2016).

*Positing the SOM as a graph* is straightforward but a distinction must be made between the SOM output space, which is traditionally visualized on a two-dimensional lattice, and the weighted graphs we derive as inputs to graph segmentation algorithms.

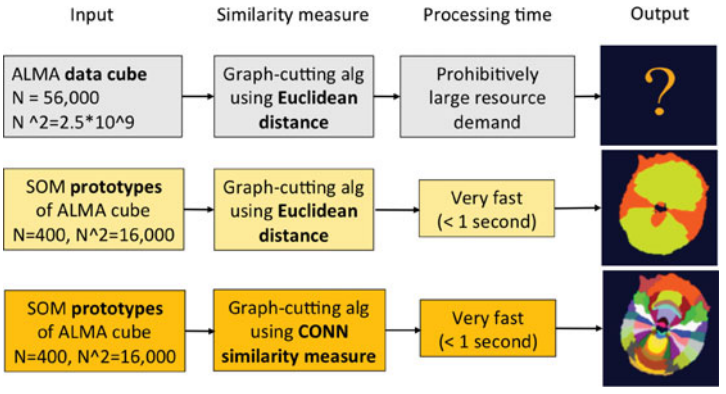


**Figure 4.** Schematics of three approaches to graph segmentation for clustering the ALMA cube of HD142527, and their results. The outputs in the last column are from the Walktrap algorithm, the best-performing of the graph-cutting methods we tested, but the relative differences in quality are representative of results from other algorithms as well. **Top row:** Using the graph segmentation algorithm with the individual 200-D spectra as inputs and Euclidean distance as similarity measure. No output was generated due to excessive resource demands. However, we do not expect better results than those shown in the middle row since SOM prototypes (if learned correctly) reflect the salient discriminative features of the data. **Middle row:** Using the learned (200-D) SOM prototypes as input vectors and Euclidean distance for similarity measure. **Bottom row:** Using SOM prototypes as input vectors and CONN similarity measure.

In both, vertices represent data prototype vectors but the latter is not a lattice; its connectivity structure is dictated by the function we choose to represent similarities between prototypes. For comparison, we consider both the CONN values and the (inverse) Euclidean distance between prototype vectors (IEDP) as similarity measures.

Graphs based on prototype representations of the data, as opposed to graphs based on individual data points (in our case, the spectral signatures of the pixels of the HD142527 data cube), offer three benefits for automated segmentation algorithms. From a computational standpoint, it is intractable to compute and store the similarity measure between all pixel pairs of a large data cube. For example, the relatively small HD142527 cube requires on the order of $10^9$ edge weights. Edge sparsity, if available, can lower this demand somewhat but that typically requires some *a priori* knowledge or preprocessing of the data.

Prototype-based learning schemes have the added benefit of noise reduction, boosting the signal-to-noise ratio in their representation of the data distribution. Most importantly, such schemes permit the introduction of new similarity measures (i.e., CONN) which are unavailable in the data domain itself. Figure 4 summarizes the combined benefits of using SOM prototypes and CONN similarity measure over traditional choices, for both processing time and the quality of the segmentation.

*Results.* Figure 5 displays the clusterings by `Walktrap` (5a,5c) and `Infomap` (5b,5d) applied to the SOM prototypes using their default parameter. We note first that all algorithms we considered perform extremely poorly when inverse Euclidean distance (IEDP) is used as a similarity between graph vertices. This is evident for the two algorithms highlighted here when comparing 5c and 5d to Figure 3, where little more than the general shape of the protoplanetary disk is identified. The CONN similarity measure helps these algorithms discern clusterings (5a,5b) which are much more similar to Figure 3.

Visual comparison of the results in Figure 5 to those in Figure 3 favors the `Walktrap` algorithm. Three different kinds of statistical assessments also support this conclusion: a) the percentage of matching pixels; b)) cluster size distribution (via the Jensen-Shannon

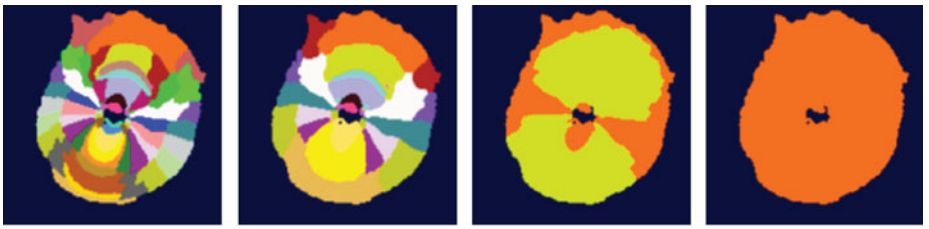


**Figure 5.** Automated SOM clusterings of the HD142527 data by the `Walktrap` (a, c) and `Infomap` (b, d) algorithms using CONN values (a, b) and IEDP (b, d) as a similarity measure.

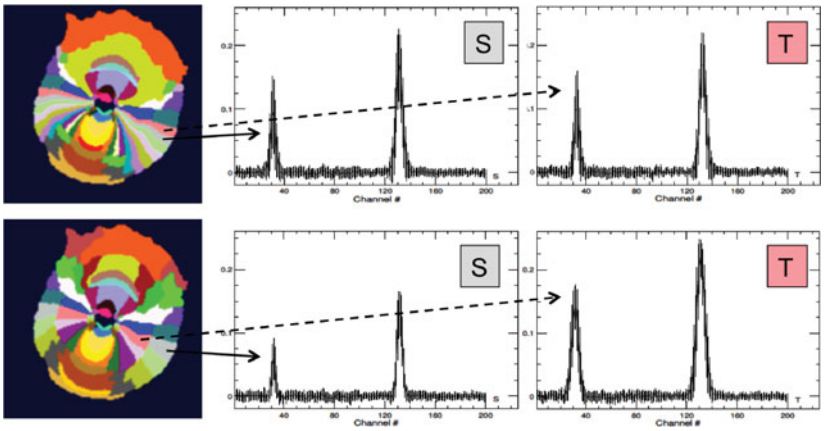


**Figure 6. Top row:** interactively produced cluster map from Figure 3, and mean spectra and standard deviations (vertical bars) of two "radial" clusters, labeled S and T. **Bottom row:** automatic cluster map by the `Walktrap`-CONN approach, and spectra of the two clusters, also labeled S and T, that cover approximately the same area as those from the interactive clustering.

(J-S) divergence); and c) cluster composition (via the Jaccard similarity coefficient ($JSC$)). These are described in Merényi *et al.* (2016). Despite their differences, the `Walktrap` and `Infomap` segmentations generally capture at least the high level structure in Figure 3, and without parameter tuning. However, this success is wholly dependent on the combination of a prototype-based representation of the data and the CONN similarity measure.

Focusing on the `Walktrap`-CONN results, closer examination reveals important differences with the interactive clustering. First, from the clusters which exhibit double or widened peaks in Figure 3 and possibly indicate non-Keplerian motion, the automated approach correctly captures cluster u. However, it misses v and t which include extremely small areas, suggesting that the `Walktrap` graph-cutting may be sensitive to small cluster size. Second, the `Walktrap`-CONN approach seems more tuned to an intensity drop at the apparent break line. The resulting clusters in Figure 6, bottom appear to have very similar mean velocities as their counterparts at top but the standard deviations are smaller, indicating possibly cleaner segmentation which may reveal interesting structure. The break line between clusters S and T in the `Walktrap`-CONN map appears to coincide with the transition between the bright yellow ring of $^{13}$CO emission and a more tenuous extended emission (Figure 1). Full understanding of the `Walktrap`-CONN map needs more analysis, but these preliminary results increase confidence in the automated approach and highlight an advantage over the more subjective interactive clustering.

## 5. Conclusions and next steps

We demonstrated on a relatively simple astronomical object the advantages of neural map-based clustering over traditional moment maps for finding structure. NeuroScope tools can highlight regions of distinct combinations of kinematics and gas densities for multiple molecular species in a single integrated map, alerting to potential discoveries. We also provided a mass processing perspective by using NeuroScope products as input to leading automatic graph segmentation algorithms. While some important details were missed by the automatic methods using default parameters, we expect to improve that by parameter tuning. This would allow replacement of the interactive clustering with automation and enable fast and reliable mapping of the structure of astronomical objects in large archives or in on-board processing.

Next we will target more complex astronomical objects (such as molecular clouds) with more chaotic kinematics and superimposed sources. With NeuroScope we can also naturally combine data from different wavelength regions or from different instruments. This opens up a path in answering many salient questions in broad astronomical investigations which will necessarily involve integrated study of physical structures, thermal conditions, and chemical processes along with kinematics. Clustering methods that can deal with the complexity and richness of the data involved and are capable of large-scale, automated processing will play an essential role in uncovering the intricate processes of the universe. We believe that NeuroScope offers tools toward this goal.

## Acknowledgment

A. I. acknowledges grant NNX15AB06G, NASA Origins of Solar Systems Program. J.T. thanks NIH/NCI T32 CA096520: Training program in Biostatistics for Cancer Research.

## References

Boehler, Y., Isella, A., Weaver, E., *et al.* 2016, submitted

Csardi, G. & Nepusz, T. 2006, *InterJournal, Complex Systems*, 1695

DeSieno, D. 1988, in *Proc. IEEE Int'l Conference on Neural Networks (ICNN)*, July 1988, Vol. I, New York, I–117–124

Fortunato, S. 2010, *Physics Reports*, 486, 75

Houlahan, P. & Scalo, J. 1992, *The Astrophysical Journal*, 393, 171

Isella, A., Pérez, L. M., Carpenter, J. M., *et al.* 2013, *The Astrophysical Journal*, 775, 30

Kohonen, T. 1988, *Self-Organization and Associative Memory* (New York: Springer-Verlag)

Lachmair, J., Merényi, E., Porrmann, M., & Rückert, U. 2013, *Neurocomputing*, 112, 189

Merényi, E., Jain, A., & Villmann, T. 2007, *IEEE Trans. on Neural Networks*, 18, 786

Merényi, E., Tasdemir, K., & Zhang, L. 2009, in *Lecture Notes in Computer Science*, Vol. 5400, Similarity-Based Clustering, ed. M. Biehl, B. Hammer, M. Verleysen, & T. Villmann (Berlin Heidelberg: Springer Verlag), 138–168

Merényi, E., Taylor, J., & Isella, A. 2016, in *Proc. of IEEE Symposium Series on Computational Intelligence (SSCI 2016)*, Athens, Greece, forthcoming

Pons, P. & Latapy, M. 2005, *ArXiv Physics* e-prints, physics/0512106

Rosolowsky, E. & Leroy, A. 2006, *Publications of the Astronomical Society of the Pacific*, 118, 590

Rosolowsky, E., Pineda, J., Kaufmann, J., & Goodman, A. 2008, *The Astrophysical Journal*, 678, 1338

Rosvall, M., & Bergstrom, C. 2008, *Proceedings of the National Academy of Science*, 105, 1118

Sousbie, T. 2013, arXiv

Tasdemir, K. & Merényi, E. 2009, *IEEE Trans. on Neural Networks*, 20, 549

—. 2011, *IEEE Trans. Systems, Man and Cybernetics*, Part B, 41, 1039, dOI: 10.1109/TSMCB.2010.2104319

Williams, J., de Geus, E., & Blitz, L. 1994, *The Astrophysical Journal*, 428, 693