Deep Data

Discovery and Visualization

Erzsébet Merényi

Department of Statistics, and
Department of Electrical and Computer Engineering
Rice University, Houston, Texas

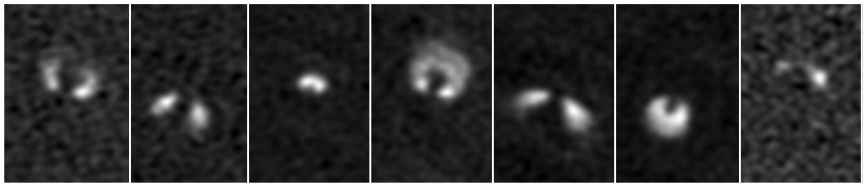
Joint work with
Josh Taylor, Dept. of Statistics
Andrea Isella, Dept. of Physics and Astronomy
Rice University

Deep Data: co-registered stacked data

- Many spectral channels in an image cube – becoming the norm
  - Or spectral data without spatial context
- fMRI (high-D time series at each spatial voxel)
- Images stacked from different heterogeneous observations
  - Image cubes concatenated from UV, VIS-NIR, radio
- The richness can hold the key to discovery of the unknown, and subtle relations

ALMA hyperspectral image

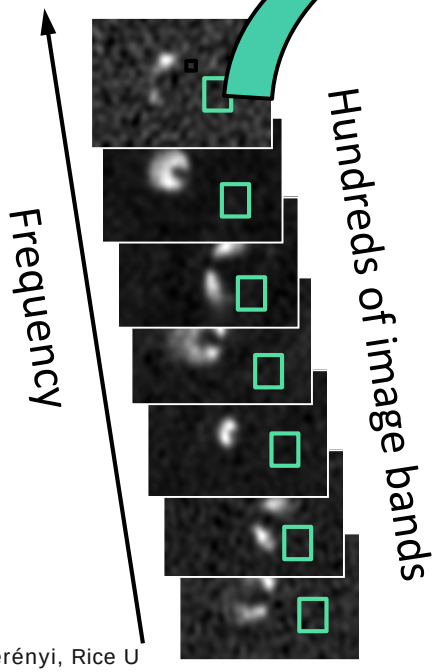
Image planes from ALMA Band 7, protoplanetary disk HD 142527



GHz 329.299-329.305 330.555 – 330.564 342.850-342.856

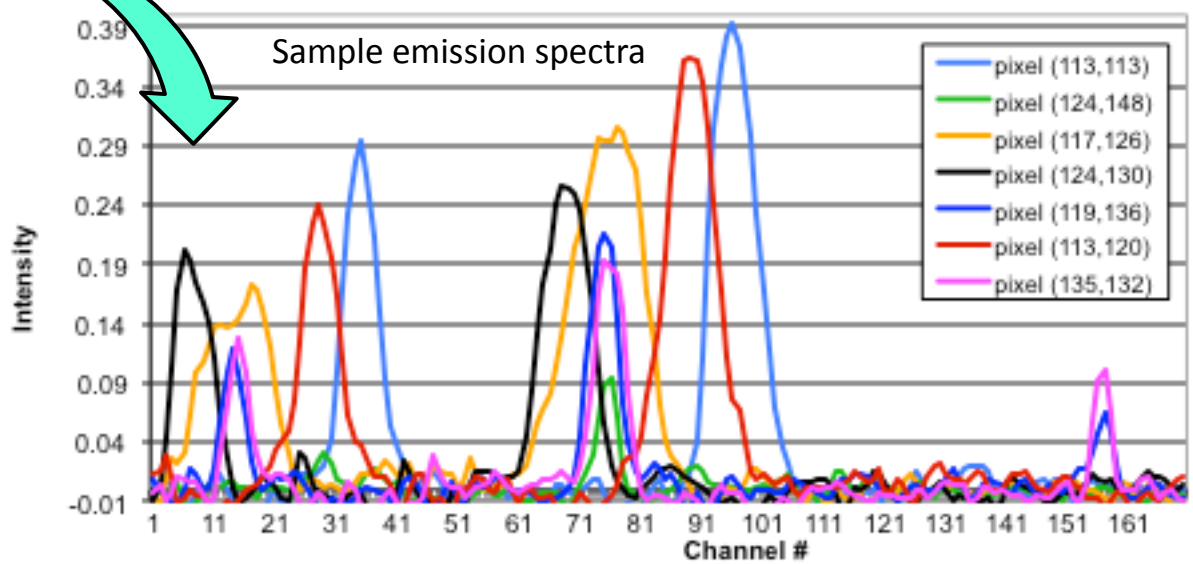
Ch 1 50 51 120 121 170

170 channels: $C^{18}O$, ^{13}CO, CS lines stacked
Spectral resolution: 0.122 MHz



Astronomical images can have thousands of channels!

ALMA has receiver Bands 1 – 10.
This sample is only from one receiver (Band 7).



ALMA spectra from combined $C^{18}O$, ^{13}CO, CS lines, showing differences in composition, Doppler shift, temperature

(Data credit: JVO, project 2011.0.00318.5)

E. Merényi, Rice U
erzsebet@rice.edu

Deep Data: co-registered stacked data

- We collect all these variables to discover more details, more unknowns

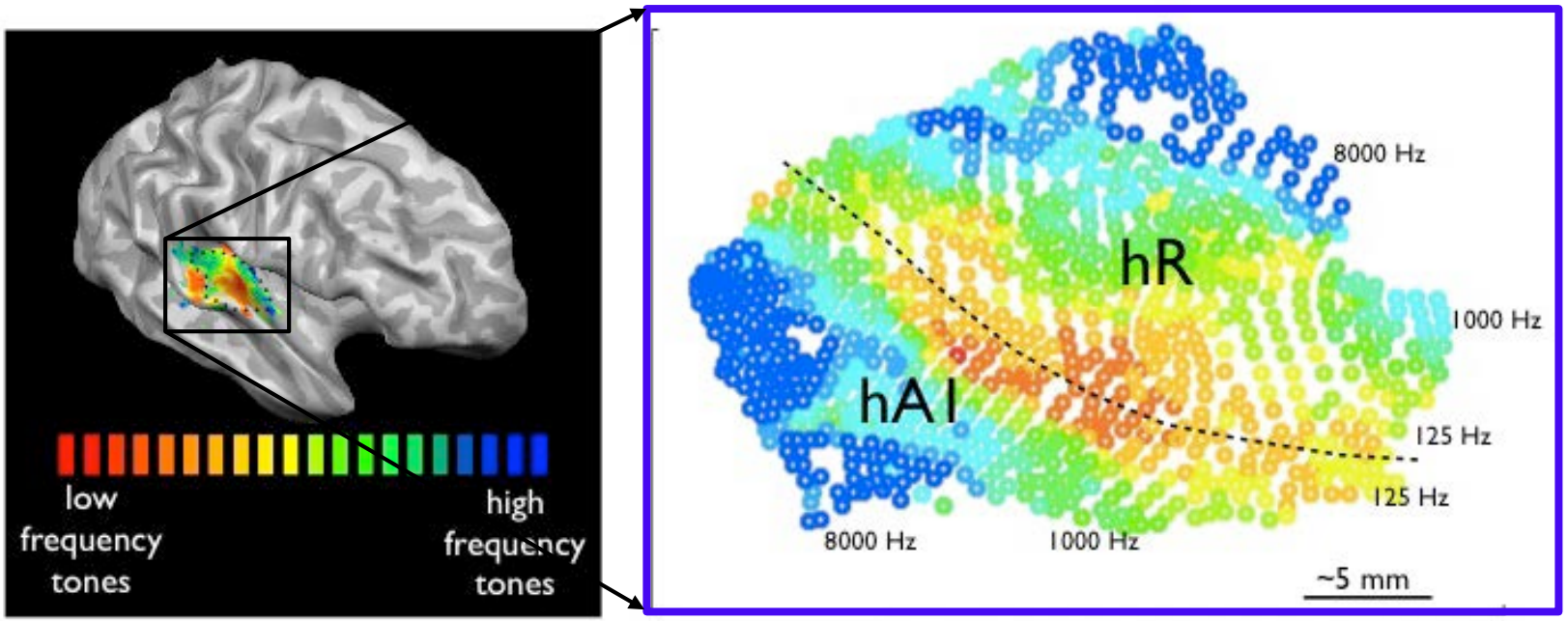
- Main challenge: complexity, high-D feature vectors (mathematically difficult)
  - often forces dimension reduction -> losing discovery potential
  - cannot be fully solved by tailoring the data to tools
  - cannot be solved by increased computing power alone

- Our brain deals with it …

Natural neural maps in the brain

Example of biological neural map: tonotopic map

In the auditory cortex (2-D surface) *tonotopic maps* are formed where the spatial *order* of cell responses corresponds to the similarity of the acoustic frequency of tones perceived.



Topology preserving mapping of acoustic frequencies in the auditory cortex

Image source: http://wp.unil.ch/neuroaudio/resaerch/auditory-cortex/

Ordering and quantization emerge by *self-organization through iterative learning*.

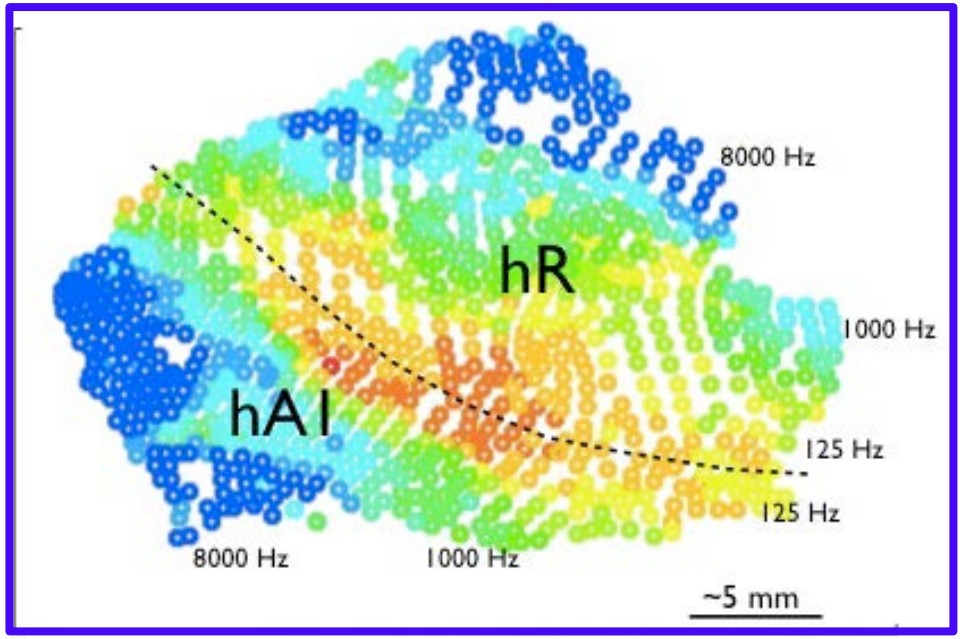
Natural neural maps in the brain

Example of biological neural map: tonotopic map

In the auditory cortex (2-D surface) *tonotopic maps* are formed where the spatial *order* of cell responses corresponds to the similarity of the acoustic frequency of tones perceived.

Neural maps in brain

- Summarize intelligently: more neurons where more details
- Express the topological *ordering* of high-D inputs in 2-D



Topology preserving mapping of acoustic frequencies in the auditory cortex

Image source: http://wp.unil.ch/neuroaudio/resaerch/auditory-cortex/

Ordering and quantization emerge by *self-organization through iterative learning*.

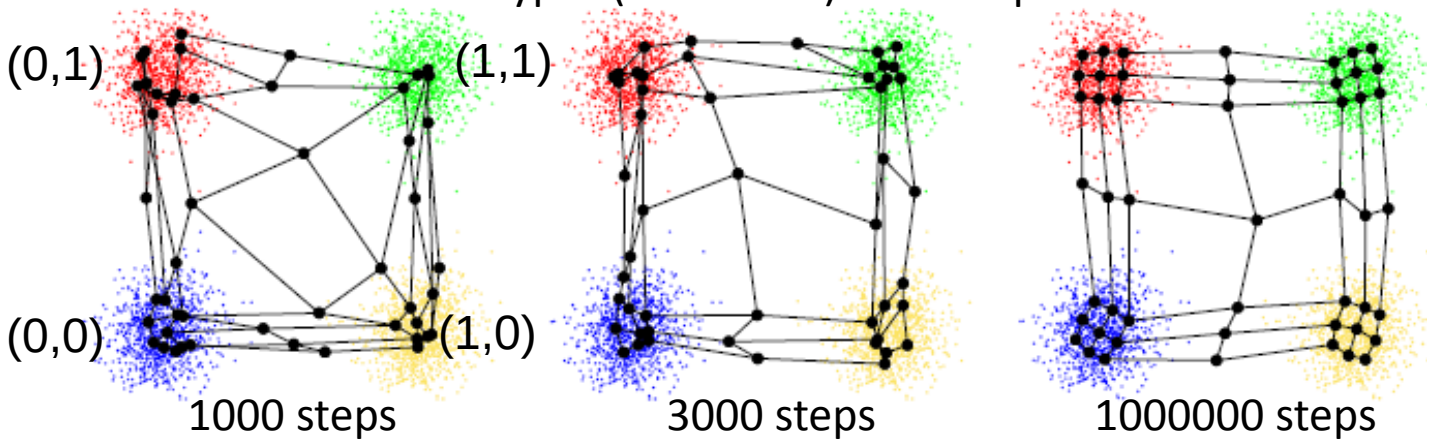
Deep Data

- We can mimic the information processing of natural neural maps – with *artificial neural maps*

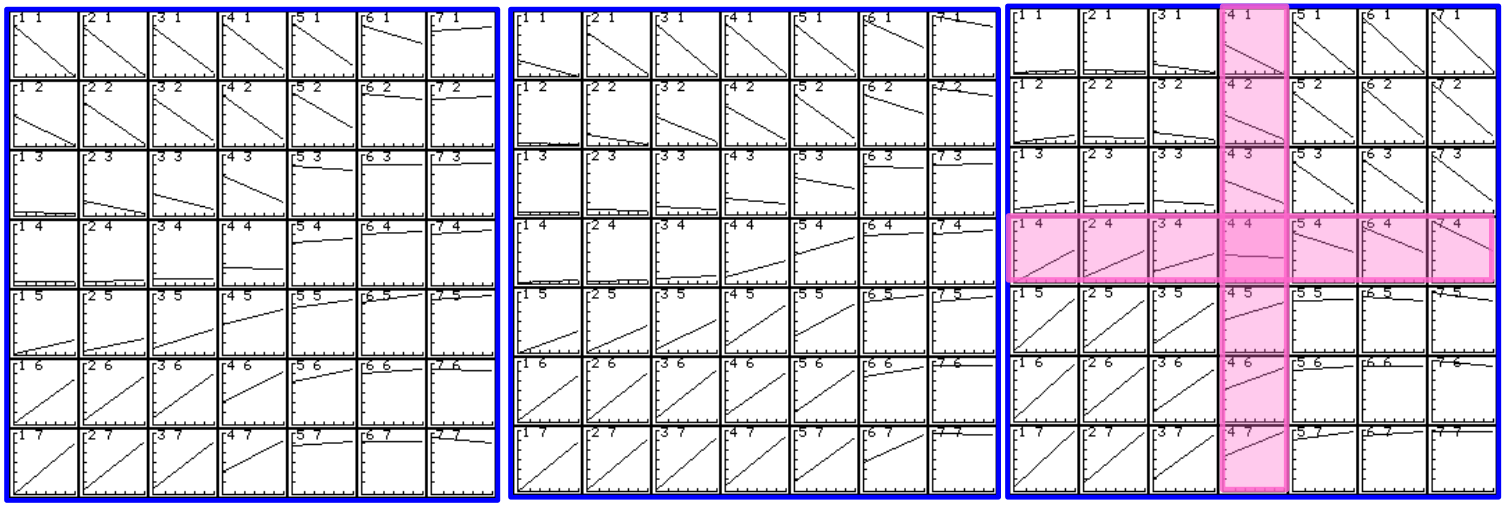
- The Kohonen Self-Organizing Map (KSOM) is best known. More variety exists with different powerful capabilities. All are *prototype-based learning* algorithms.
  - Conscious SOM (CSOM), Neural Gas, LVQ "supervised" variants, SOMs with magnification control, …

CSOM learning four 2-D Gaussian clusters: evolution of prototype vectors, unsupervised

Prototypes (black dots) in data space

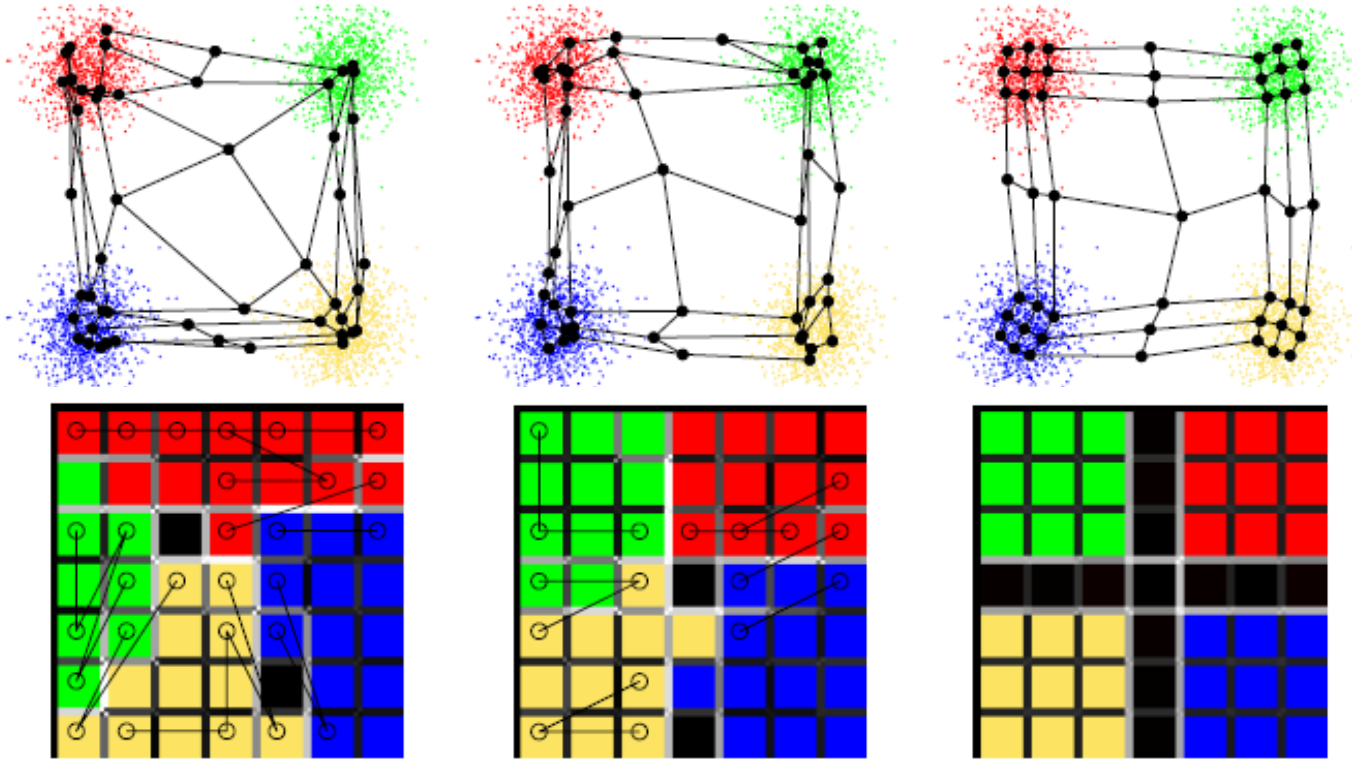


4,000 data vectors, 49 SOM prototypes



Prototype vectors in SOM lattice (drawn in parallel coordinates)

SOM learning 4 Gaussian clusters



Prototypes in data space

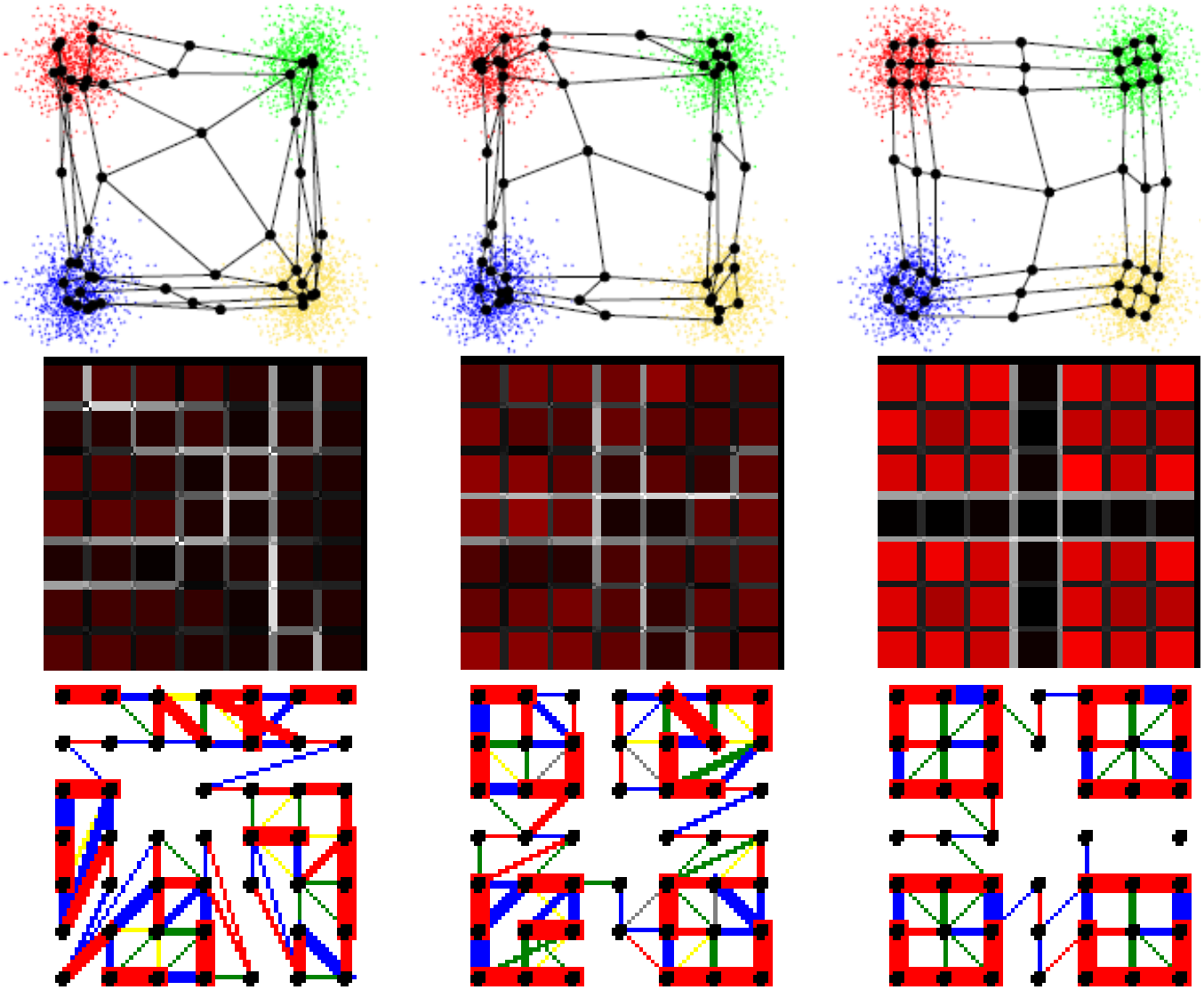
Prototypes in SOM with majority labels and topology violations shown

From Merényi, Taşdemir, Zhang, Springer LNAI 540.0, 2009

But, we don't know the labels. We need to find the clusters from the learned SOM.

Step 1: SOM learns. This is only needs 2-3 parameters, and is easy.
Step 2: We interpret the SOM's knowledge. This can be hard for complex data.

SOM learning 4 Gaussian clusters



SOM prototypes (black dots) in data space

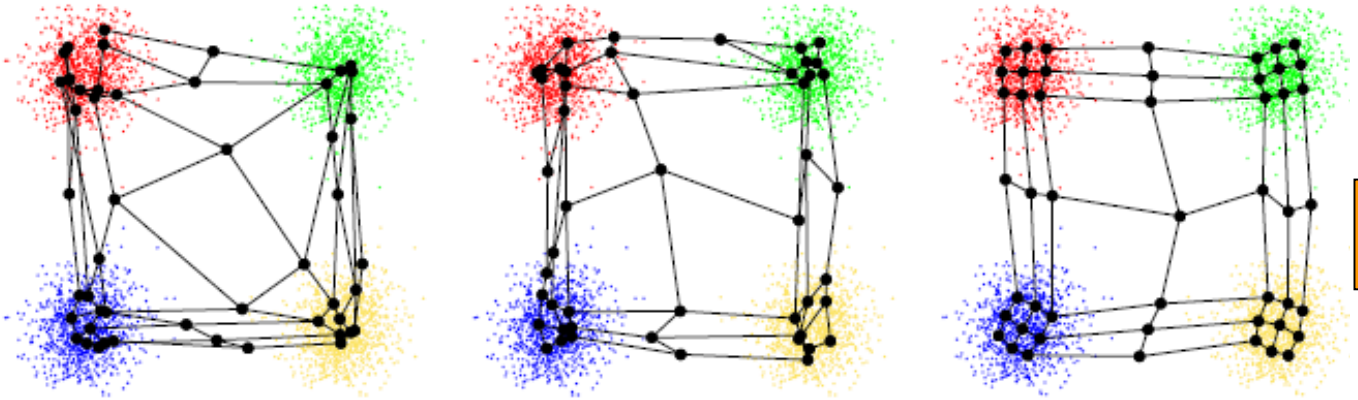
SOM knowledge visualized

mU-matrix representation

CONN graph representation,

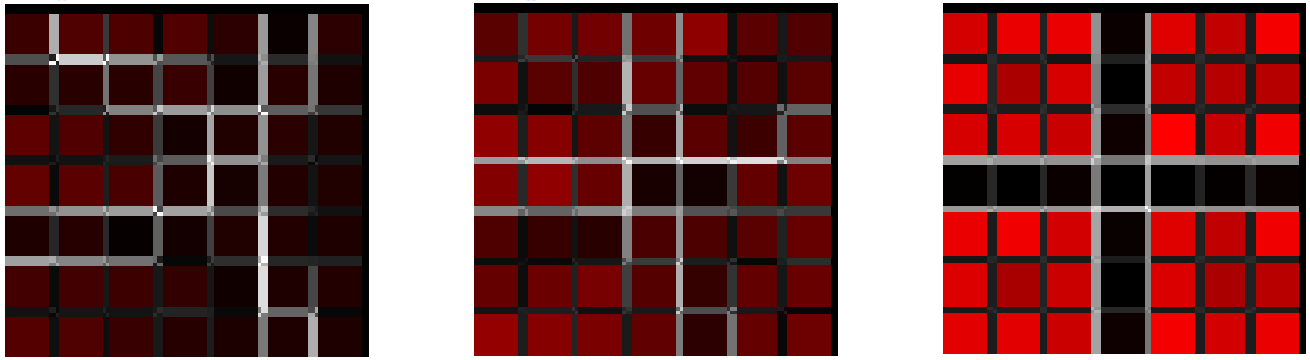
Merényi et al., 2009
Tasdemir & Merényi, 2009

SOM learning 4 Gaussian clusters
Visualizations to guide cluster identification



SOM prototypes (black dots) in data space

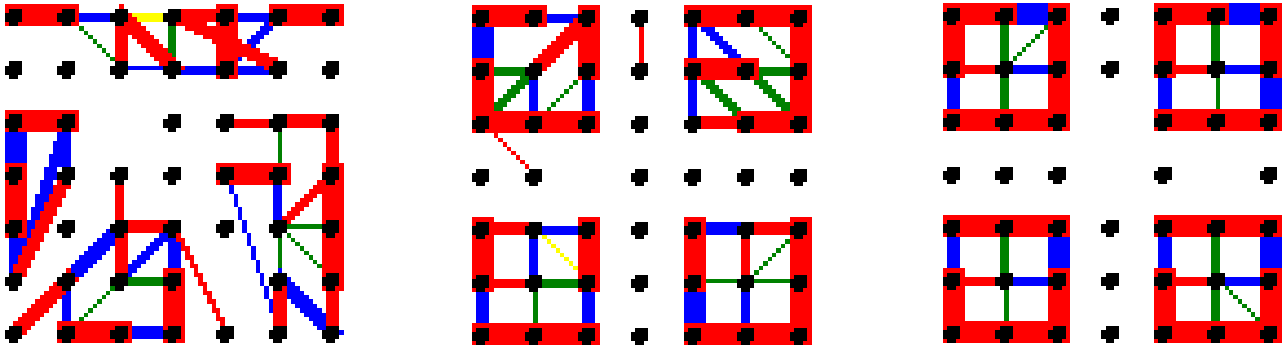
Cannot be shown for >2-D data



SOM knowledge visualized

Can be shown for >2-D data

mU-matrix representation

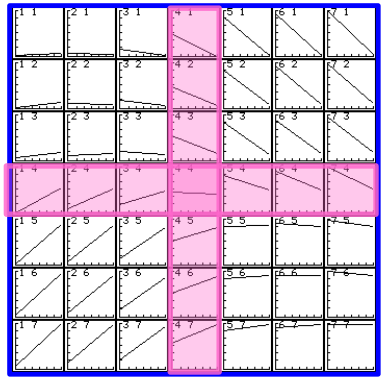


CONN graph representation, unimportant connections removed

NeuroScope approach

NeuroScope: collection of neural map based clustering and classification methods and related tools (visualization, similarity metric, evaluation tools) that we have been developing

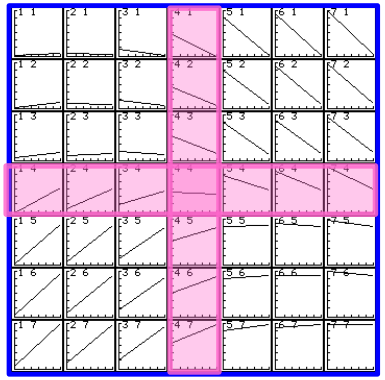
- Use all features – keep the discovery potential
- Summarize the data by unsupervised neural map learning
  - prototype-based learning: N -> O(sqrt(N))
  - preserve the relevant structure (match the data distribution), reduce noise
  - Does NOT reduce feature dimension
- Visualize in 2-D - interpretable



NeuroScope approach to structure discovery

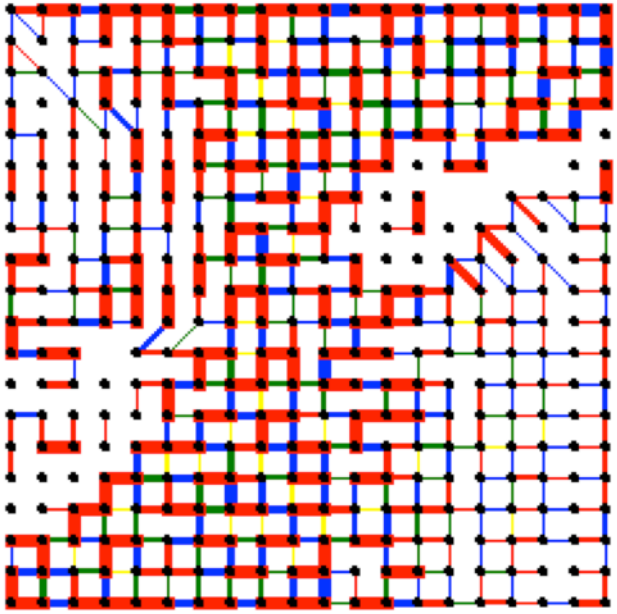
<u>Step 1:</u> Learn the data manifold

- Easy, reliable, little tuning needed, automatic.
- We use Conscience SOMs for maximum entropy mapping (best for information transfer), and other advanced SOM variants.



<u>Step 2:</u> cluster the SOM prototypes

- Hard for complicated data. Need good knowledge representation, and similarity measure, like the CONN graph.
- Interactive cluster extraction best so far.



SOM / CONN of the ALMA data

NeuroScope clustering - real data

- ALMA example of discovering subtle spectral feature consistent with non-Keplerian motion in a protoplanetary disk, and possibly indicating planet formation

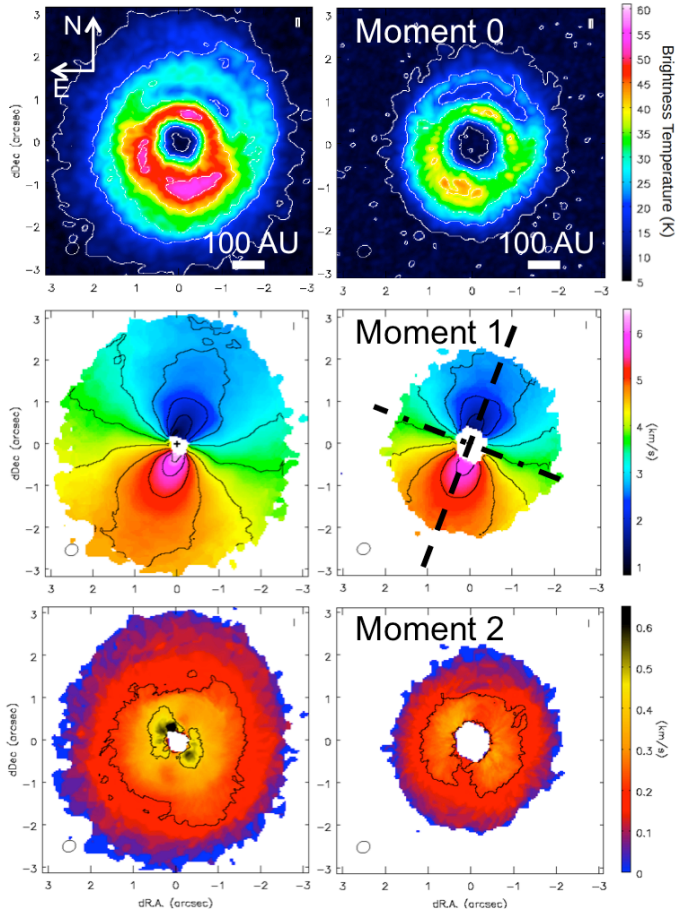
- Approach for automation and scalable processing

Discover more from deep data

Data: ALMA image cubes of HD142527 (Isella, 2015)

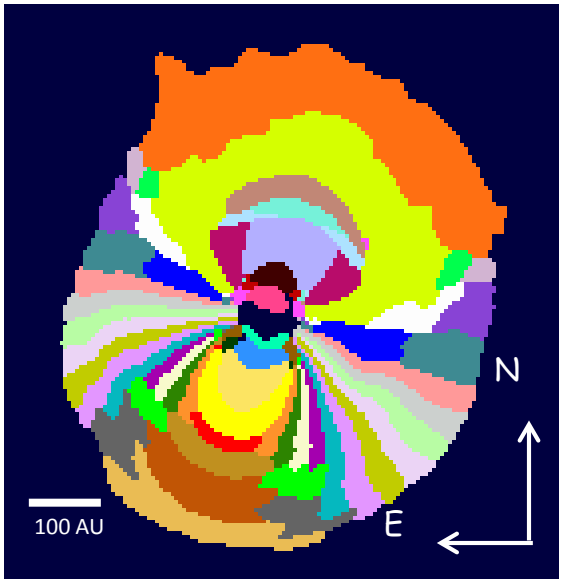
Moment maps
Each from 100 channels

^{13}CO J=3-2 $C^{18}O$ J=3-2



(Boehler et al., 2016)

SOM / CONN cluster map from stacked $C^{18}O$, ^{13}CO lines, 100 + 100 channels as input feature vectors

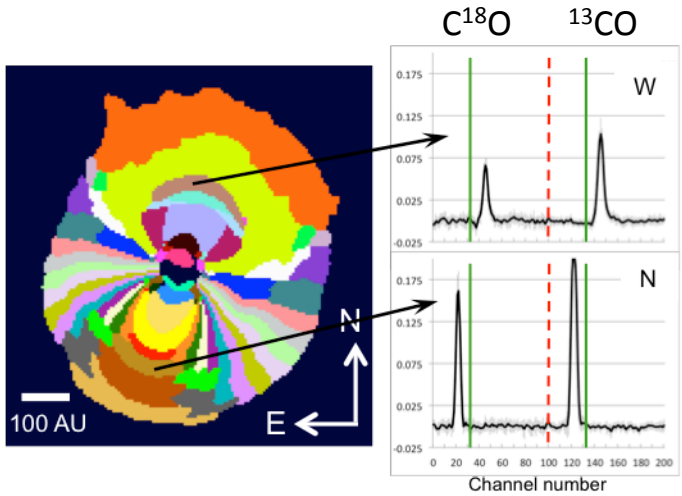


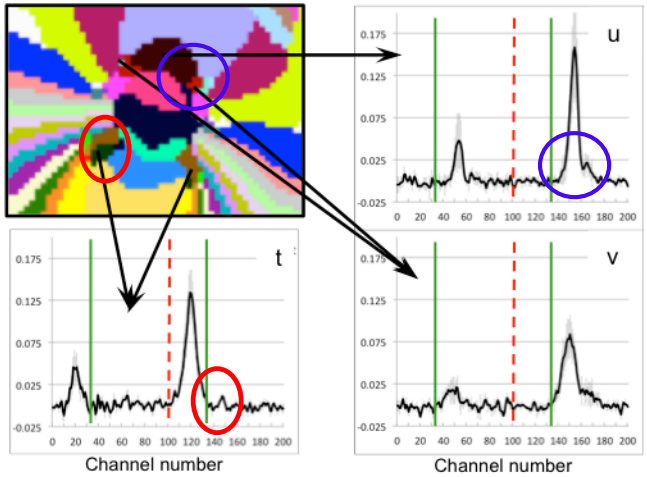
The emerging structure of the protoplanetary disk based on all channels of two molecular tracers, visualized in one 2-D view

Coloring of clusters is arbitrary (to provide contrast), not a heat map!

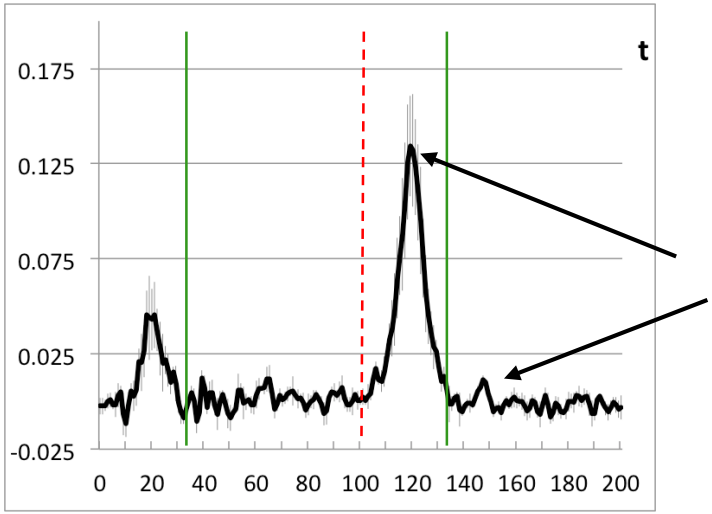
E. Merényi, Rice U
erzsebet@rice.edu

Discover more from deep data





Mean cluster signatures alert to interesting areas missed by the moment maps.



Two distinct peaks, shifted opposite from rest frequency. Two gas components moving in different directions.

(Merényi, Taylor, Isella, IEEE SSCI CIDM 2016)

More discovery within one molecular line

More discovery from the combination of lines

So far so good … but not enough

So far we have seen

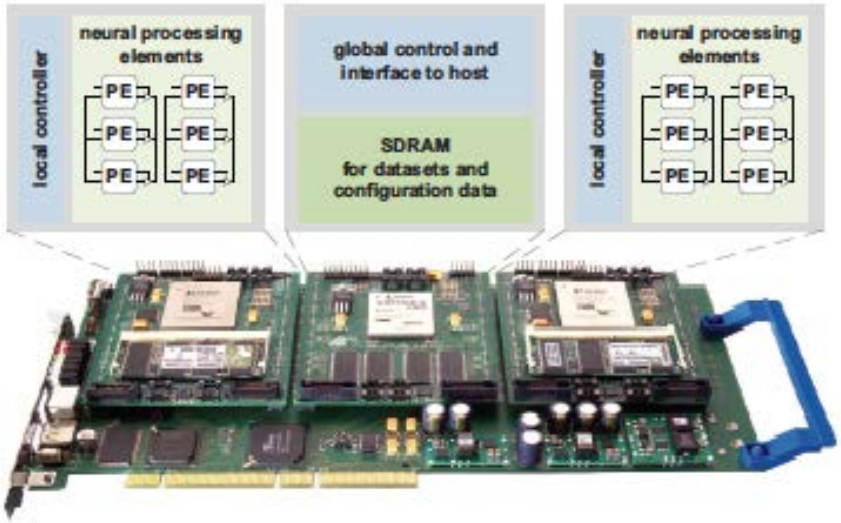
- We can make interesting discoveries, using the full input feature set with NeuroScope tools.
- We can visualize multi-dimensional relationships in high-D data

But does it scale?

- Step 1, Iterative SOM learning can take long on regular computers
- Step 2, Interactive cluster extraction from an SOM is slow and requires expert knowledge

Efficiency for large data
Step 1: SOM learning in parallel hardware

SOM accelerator gNBXe *(Lachmair, Merényi, Porrmann, Rückert, Neurocomputing, 2013)*



Designed for optimal algorithm mapping

- Three SOM variants implemented, <u>reconfigurable, on-chip learning</u>
- <u>Large-scale computation:</u> handles real hyperspectral imagery
- <u>2013:</u> FPGA-based prototype, ~ 12–25 x faster than Core-i7 PC, 4 threads, for large SOM / high-d data (consumes 80-90% less energy) . <u>2016:</u> > 100 x speed-up.
- <u>Future:</u> ASIC implementation is expected to gain another factor of 10 (or more, depending on the nano-scale technology)

Automation for Step 2, cluster extraction: Graph-segmentation informed by SOM and CONN

- ☺ Graph-cutting methods: automatic, only 1 or 2 parameters
- ☹ Can't deal with many data points. N vectors => N^2 edges. For this small image (56,000 vectors), over 10^9 edges !!!
- ☺ ☺ Use intelligently summarized data (SOM prototypes) as input.

ALMA **data cube**
N = 56,000
N ^2=2.5*10^9

→ Graph-cutting alg using **Euclidean distance**

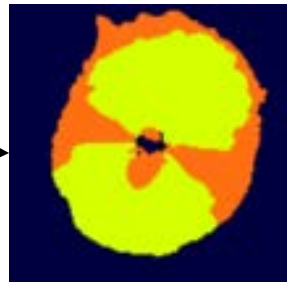
→ Takes forever
Burned down my student's Mac

→ ???

SOM **prototypes** of ALMA cube
N=400, N^2=16,000

→ Graph-cutting alg using **Euclidean distance**

→ Very fast
(< 1 second)

→ 

SOM **prototypes** of ALMA cube
N=400, N^2=16,000

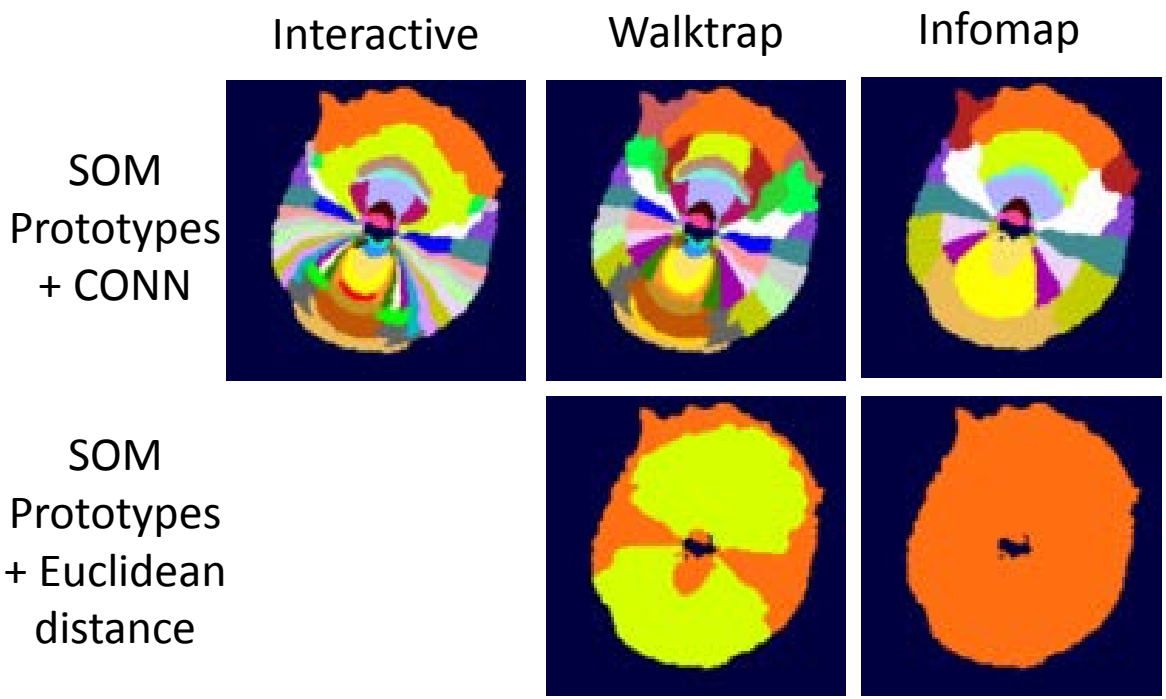
→ Graph-cutting alg using our **CONN similarity measure**

→ Very fast
(< 1 second)

→ 

Interactive vs automated results

- Walktrap (Pons & Latapy, 2005) and Infomap (Rosvall & Bergstrom) – two best results from graph segmenting algorithms with default setting, 1 or 2 parameters.
- Details don't quite match, but differences reasonable. Graph-segmentation with SOM + CONN finds structure, and FAST.



- Future work: explore non-default parameters
- Interpret differences

Interactive vs automated results

- Walktrap (Pons & Latapy, 2005) and Infomap (Rosvall & Bergstrom) – two best results from graph segmenting algorithms with default setting, 1 or 2 parameters.
- Details don't quite match, but differences reasonable. Graph-segmentation with SOM + CONN finds structure, and FAST.

Interactive Walktrap Infomap

SOM Prototypes + CONN

- Future work: explore non-default parameters
- Interpret differences

Mass-processing perspective for pipelines

- Achieve the quality of interactive structure discovery from SOMs with automated methods
  - by feeding SOM prototypes and CONN measure to suitable graph segmentation algorithms (work in progress, previous two slides)

- Perform SOM learning in parallel hardware -> 10^2 – 10^3 acceleration

- Segment the SOM prototypes automatically – a few seconds

We could map the structure of a protoplanetary disk (as in this example) and return the salient spectral properties within a few minutes

- Depending on the number of lines / channels combined

Heterogeneous data

- NeuroScope clustering is not limited to stacked data of the same kind (e.g., spectral bands; or spectral lines)

- Disparate data from different windows of the electromagnetic spectrum and from different instruments can be combined

Conclusions

- **Greatly increased feature space** (e.g., spectral resolution for ALMA) **offers a magnifying lens for understanding the physical processes that generate the features** (kinematics of atomic and molecular gas and the distribution of solid particles in the ALMA example).

- **If we can exploit the richness of features** (spectral details), let the data speak for itself in more articulate ways, **we can enlarge the discovery space**.
  - Especially important for discovery of the unknown and subtle

- The NeuroScope approach provides some tools to achieve this.

- It also shows promise for large-scale, automated processing.

References

- Merényi, E., Taşdemir, K., Zhang, L. (2009) Learning highly structured manifolds: harnessing the power of SOMs. Chapter in *"Similarity based clustering", Lecture Notes in Computer Science* (Eds. M. Biehl, B. Hammer, M. Verleysen, T. Villmann), Springer-Verlag. LNAI 5400, pp. 138 – 168.

- Taşdemir, K, and Merényi, E. (2009) Exploiting the Data Topology in Visualizing and Clustering of Self-Organizing Maps. *IEEE* Trans. *Neural Networks* 20(4) pp 549 – 562.

- Boehler, Y., Isella, A., Weaver, C., Grady, J., Perez, L., and Ricci, L, A close-up view of the horseshoe disk HD 142527. ApJ, submitted, 2016.

- Merényi, E., Taylor, J., Isella, A. (2016) Mining Complex Hyperspectral ALMA Cubes for Structure with Neural Machine Learning. Proc. IEEE *Symposium Series of Computational Intelligence and Data Mining,* Athens, Greece, December 6-9, 2016, in press.

- Lachmair, J., Merényi, E., Porrmann, M., Rückert, U. (2013) A Reconfigurable Neuroprocessor for Self-Organizing Feature Maps. *Neurocomputing* 112, pp 189-199.

- Pons, P. and Latapy, M. Computing communities in large networks using random walks (2005*) ArXiv Physics e-prints*, December 2005.

- Rosvall, M. and Bergstrom, C. (2008) Maps of random walks on complex networks reveal community structure*. Proc. National Academy of Science* 105, pp 1118-1123, Jan 2008.