Knowledge Discovery from the Hyperspectral Sky

Erzsébet Merényi
Research Professor

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

erzsebet@rice.edu

Past support for projects by

Science Mission Directorate

- *Applied Information Systems Research Program*
- *Solid Earth and Natural Hazards Program*
- *Mars Data Analysis Program*
- *Outer Planets Research Program*

Focus on Complexity and Discovery

- *Complexity* is a a challenge in the analytics of Big Data, new algorithms are needed

- Big Data have various complexities, not only "more" or "less", but "different"
  - Even among different *hyperspectral data*

- *Discovery* is finding what we do not know ... can't characterize in advance (no models) -> more / unknown complexity makes it more difficult

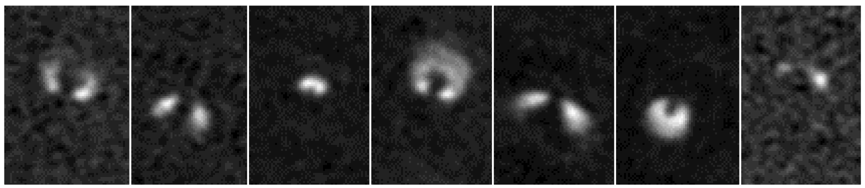
- *Neural maps* as tools: may be the closest analog to how the brain makes sense of big / complex data

Hyperspectral data: fused "wide data" – in this talk all channels are used together.

Hyperspectral imaging
of terrestrial (planetary) and astronomical objects

Astronomy example

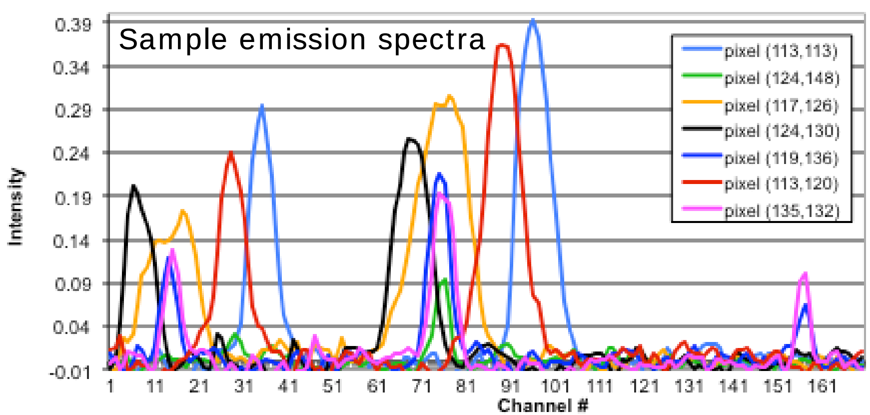
Sample image planes from ALMA Band 7, HD 142527



GHz 329.299-329.305 330.555 – 330.564 342.850-342.856

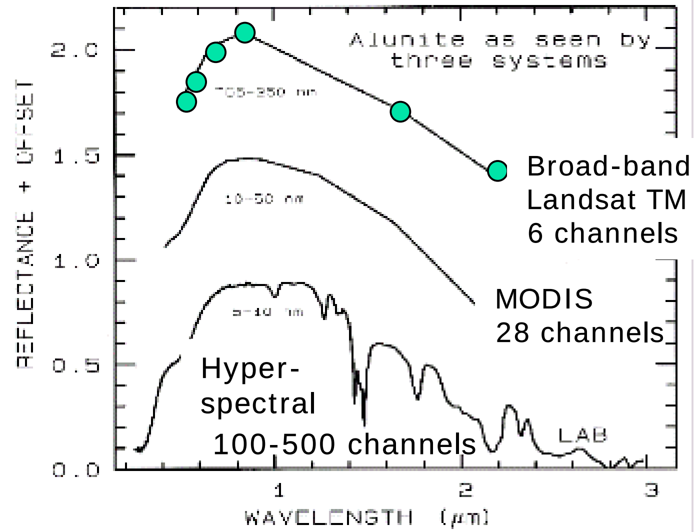
Ch 1 50 51 120 121 170

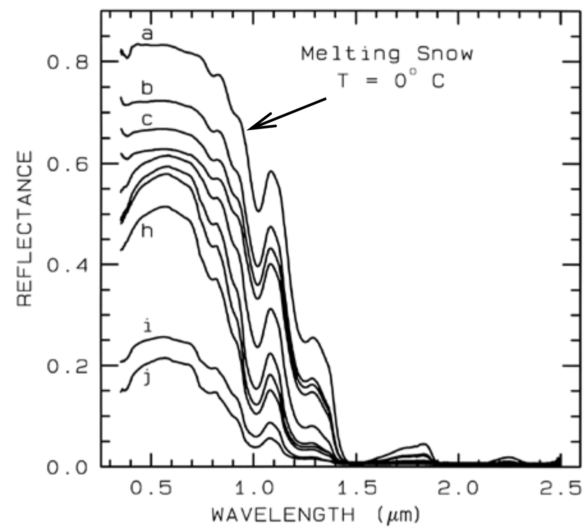
170 channels: $C^{18}O$, ^{13}CO, CS lines stacked
Spectral resolution: 0.122 MHz



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)*

Evolution of terrestrial imaging





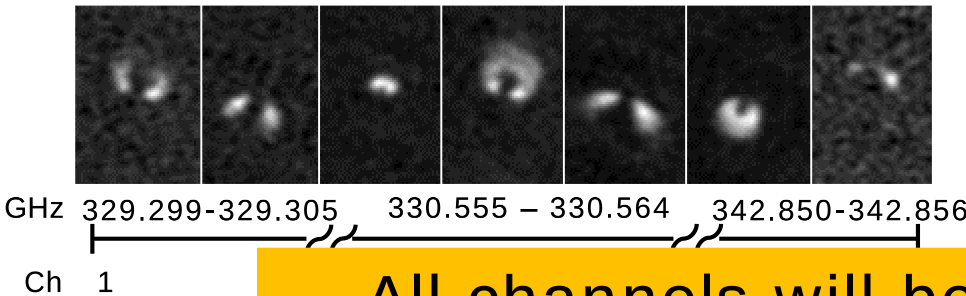
Variations of absorptions in spectra of melting snow in response to temperature changes *(speclab.cr.usgs.gov)*

E. Merényi, Rice U
erzsebet@rice.edu

Hyperspectral imaging of terrestrial (planetary) and astronomical objects

Astronomy example

Image planes from ALMA Band 7, HD 142527

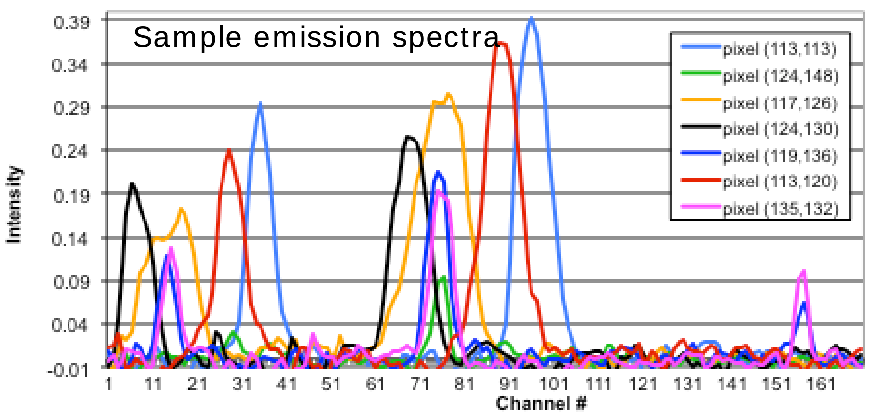


Astronomical images can have thousands of channels!

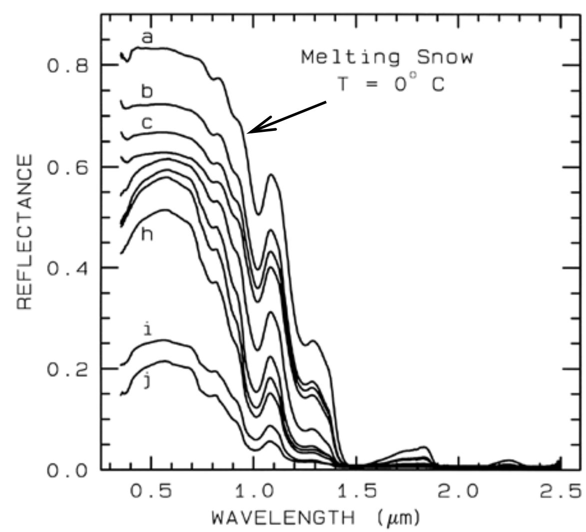
ALMA has receiver Bands 1 – 10.

one

All channels will be used together (as n-dim pattern vectors) for analysis



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)*



Variations of absorptions in spectra of melting snow in response to temperature changes *(speclab.cr.usgs.gov)*

Complex (complicated) data space

Imagine in 100 dimensions!

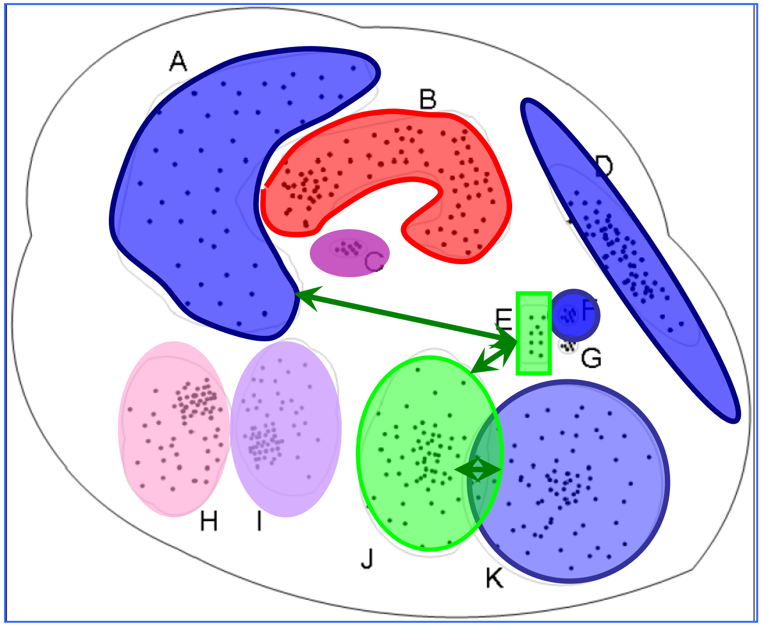
- High-dimensional
- Large (number of data points)
- Multi-modal (has clusters)
- <u>Highly structured</u>
  - Not linearly separable
  - Widely varying shapes and sizes
  - … densities (vary within and across clusters)
  - … proximities
  - … local dimensionalities

No statistical models

Hyperspectral data have many clusters with widely varying shapes, sizes, densities, proximities, local dimensionalities …

Small hyperspectral data can also be complex and resist discovery with many methods.

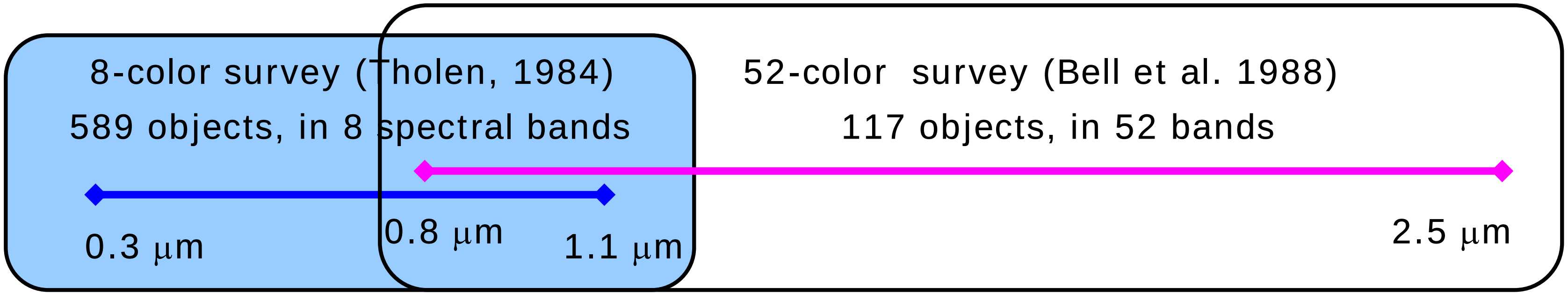
Highly structured data space



Merényi, Taşdemir, Zhang, Springer,LNAI 5400. 2009

Motivation - The First Case, from Tucson ☺

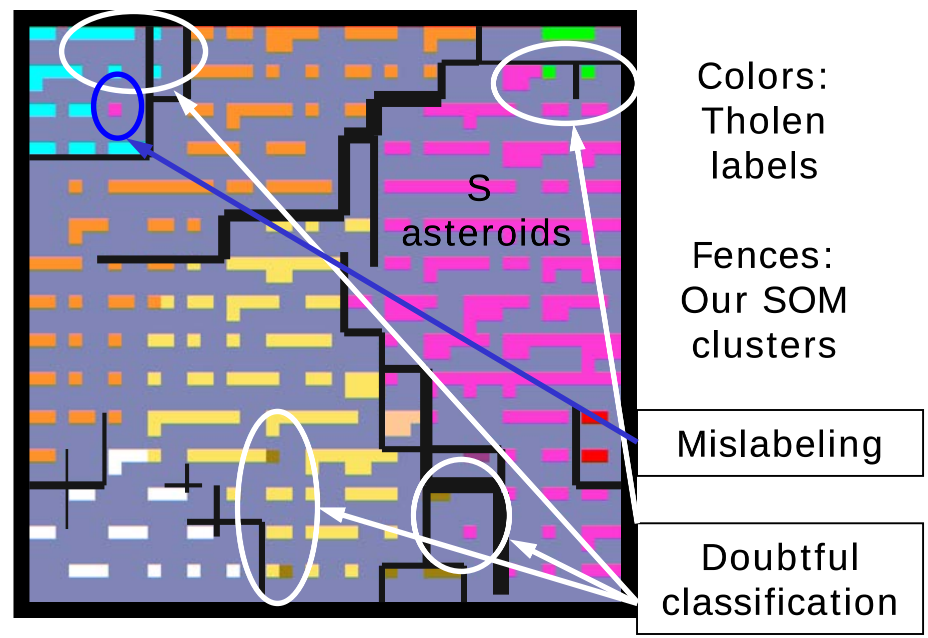
Finding olivine and pyroxene subgroups of S asteroids with Self-Organizing Maps



Previous work

- Tholen taxonomy of asteroid compositions established based on spectral shapes in 8-color survey
- Techniques used for clustering: PCA, Minimum Spanning Tree, band ratios, G-mode analysis
- Bell's 52-color survey: extended spectral range and (hyperspectral) resolution (albeit less objects)
- Discovery of more structure was expected – but not found
  - Specifically, end groups of silicate (S) asteroids

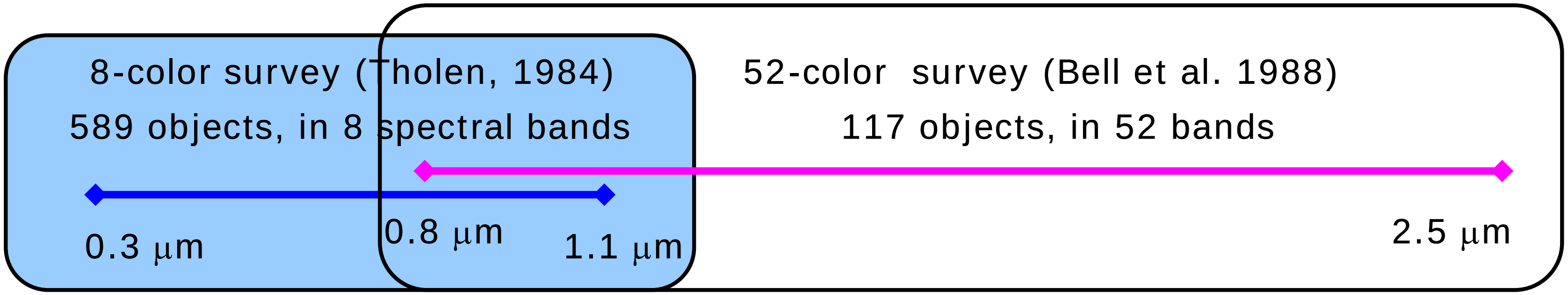
Our SOM portrait of ***8-color*** objects: Matches Tholen's taxonomy



Howell, Merényi, Lebofsky, JGR 99, 1994

Motivation - The First Case, from Tucson ☺

Finding olivine and pyroxene subgroups of S asteroids with Self-Organizing Maps



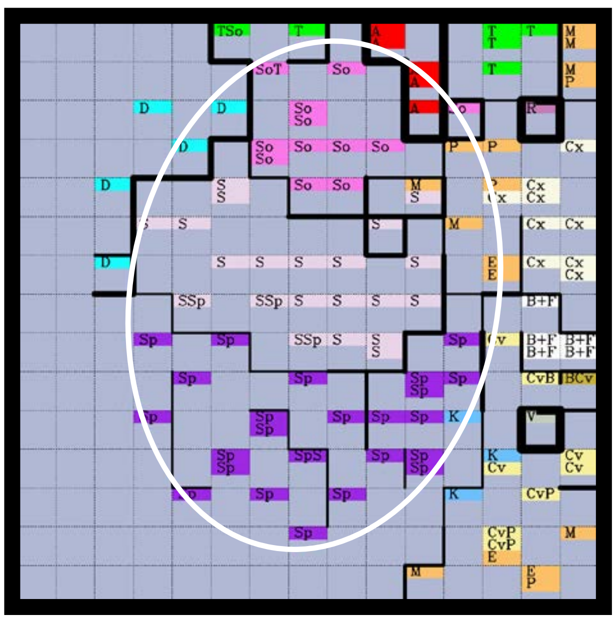
SOM portrait of ***60-color*** objects:
Does not match Tholen's taxonomy

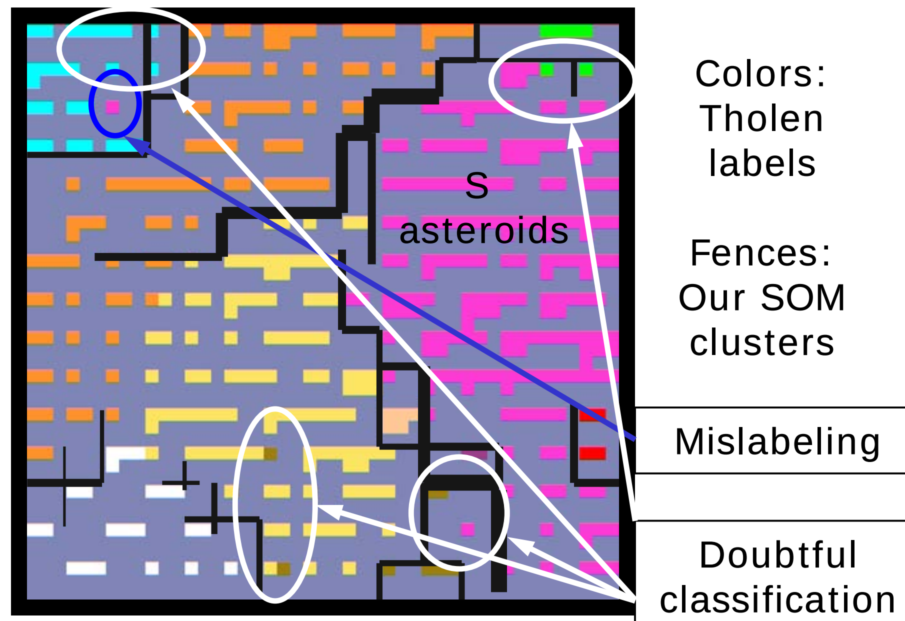
Our SOM portrait of ***8-color*** objects:
Matches Tholen's taxonomy

Pink & Purple: S types

So: Olivine rich

Sp: Pyroxene rich





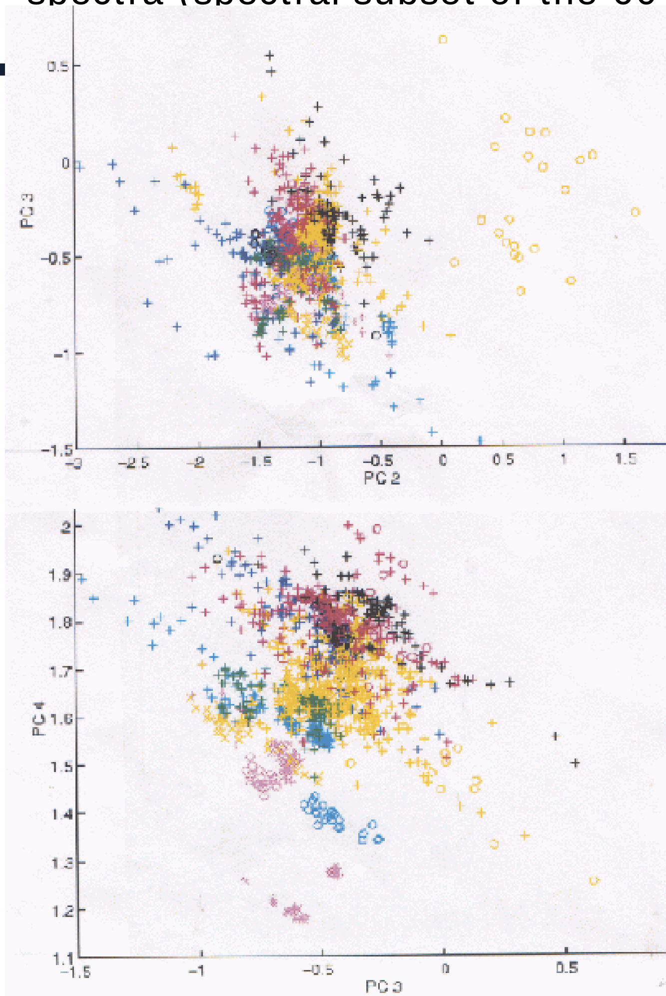
Colors: Tholen labels

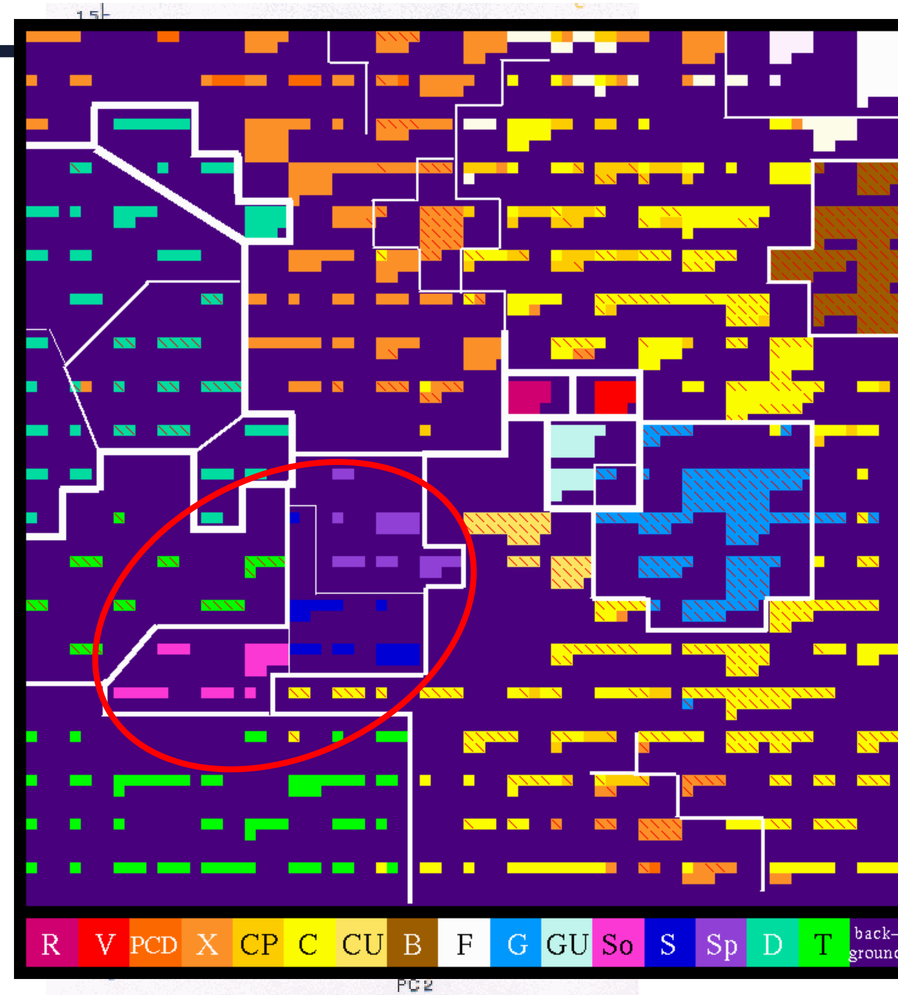
Fences: Our SOM clusters

Howell, Merényi, Lebofsky, JGR 99, 1994

Principle Component views of 13-color asteroid spectra (spectral subset of the 60-color data)

SOM view of the 13-color asteroid spectra





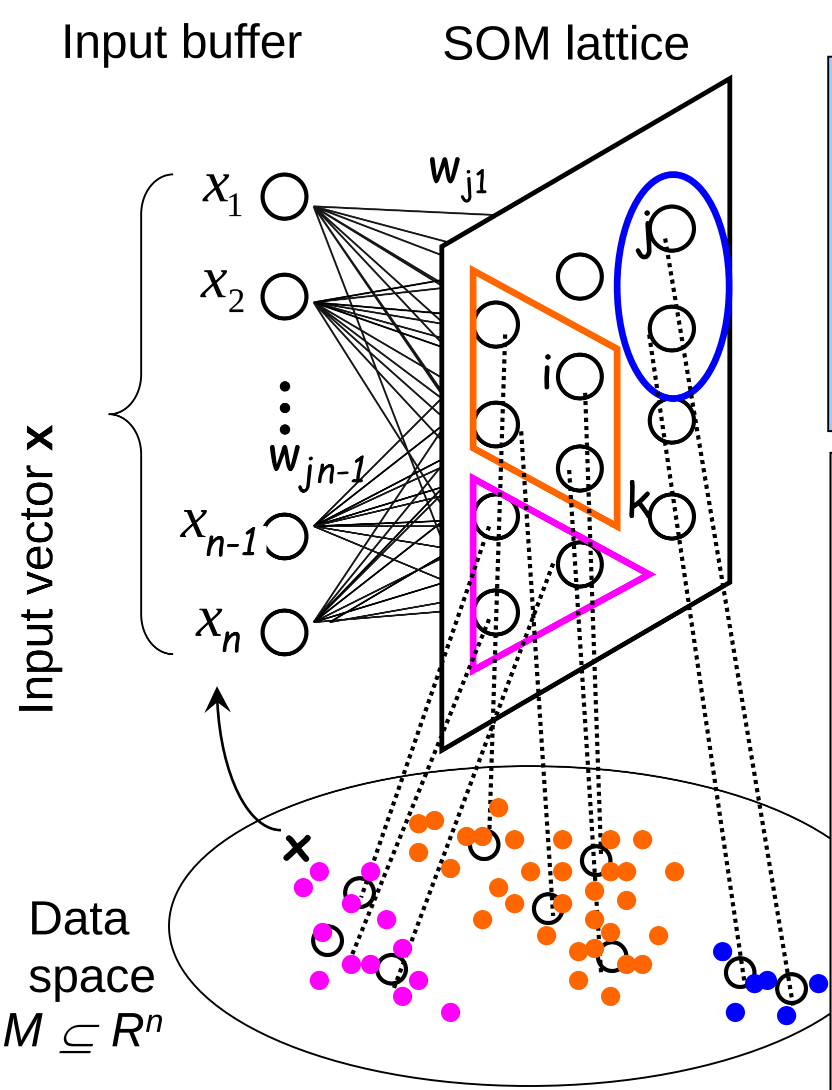
None of the 76 pair-wise PC plots resolve all 16 known clusters in this data set. Colors and symbols indicate the known labels.

All clusters were found from an SOM.

Merényi, Howell, Rivkin, Lebofsky, Icarus 129 (1997)

Prototype-based Learning With Self-Organizing Map

Most widely used machine learning model of biological neural maps



Formation of basic SOM (Kohonen, early 80's)

Simultaneous
- Adaptive Vector Quantization (VQ), and
- Ordering (indexing) of the prototypes in the SOM grid according to their similarities

SOM learns the structure of the data and represents it on a low-dimensional lattice, in a *__topology preserving__* fashion.

(If learning goes correctly …)

The SOM learns very well. Extraction of the prototype groups from the learned SOM is the challenge.

E. Merényi, Rice U
erzsebet@rice.edu

Structure discovery in *<u>complex</u>* data with Self-Organizing Maps

- Effective post-processing of the SOM is key to the extraction of clusters
  - The information learned by SOMs is generally underutilized for interpretation of data structure (cluster extraction)
  - Advanced / information theoretical variants are underutilized, <u>metrics untapped</u>.
- Exploitation of the SOM makes a difference for complex data

Side note on SOM efficiency:

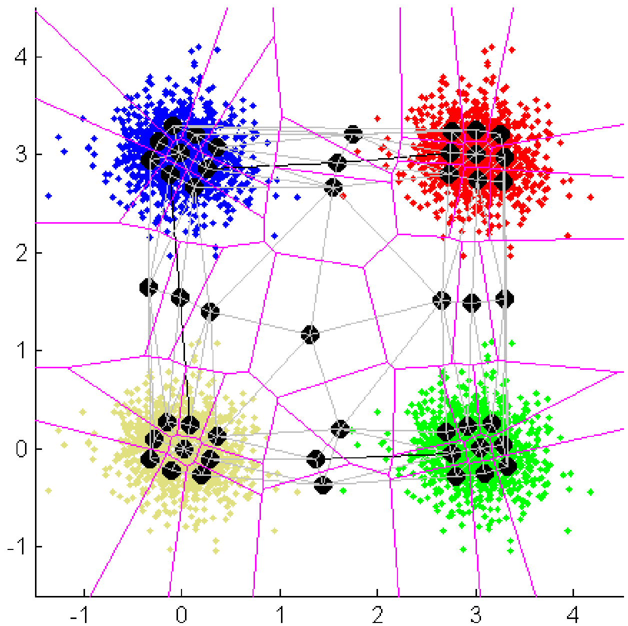
Prototype-based learning produces sparse representation of data, reduces volume *<u>during learning</u>* – advantage over graphical methods for Big Data

N = 10^6 data need ~ 10^12 graph edges; N -> N^2

N = 10^6 data can be expressed by ~ 10^3 SOM prototypes; N -> sqrt(N)

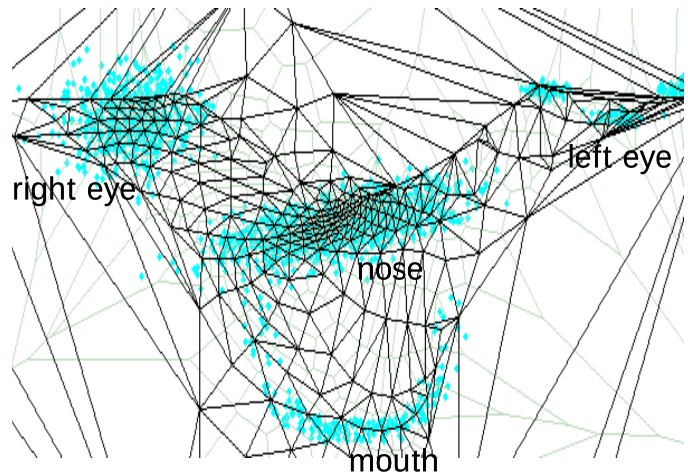
Structure / complexity of data as expressed by Voronoi tessellation and Delaunay graph

Artificial (noiseless) 2-d data, with learned SOM prototypes shown in the data space



Simple

V-cell: pink, D-graph: gray



2-d “clown data”
(Vesanto and Alhoniemi, IEEE TNN, 2000)

Somewhat complicated

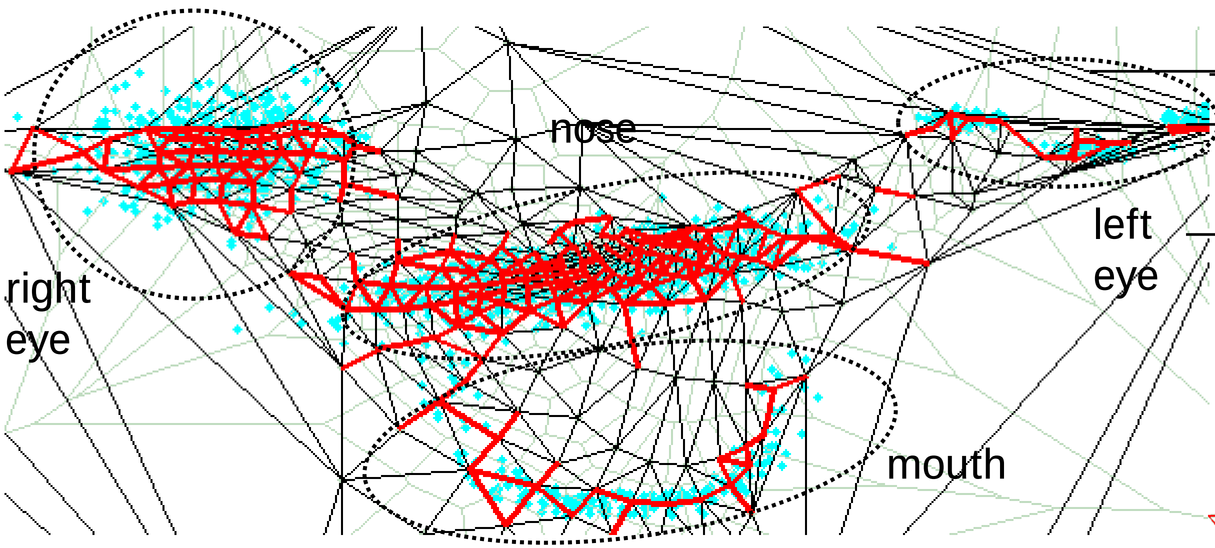
V-cell: green, D-graph: black

The V-cell and D-graph structure increases from left to right.

Cannot show the V-cells / D-graph of higher-d data in data space!

D-graph and masked D-graph of the Clown
wrt 17 x 17 SOM prototypes

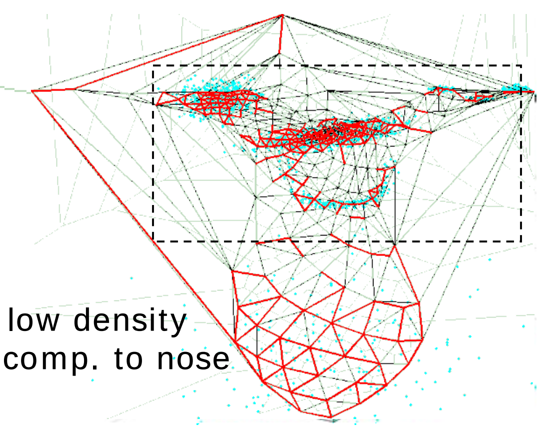
- The prototypes, learned by an SOM, nicely follow the data distribution
  - The prototypes are at the vertices of the D-graph



Emerging: Curve of mouth

Mass/shape of nose, eyes

Gaps between eyes and nose



Body: low density comp. to nose

E. Merényi, Rice U
erzsebet@rice.edu

The importance of SOM learning: builds masked D-graph

- *Martinetz and Schulten, Topology Representing Networks, IEEE TNN 1994:*

  Competitive Hebbian learning – as in neural maps - guarantees the construction of the masked D-graph of the (learned) prototypes (under one condition).

- Easy to do: For each data point $v \in M \subset R^n$ record the BMU and 2nd BMU pairs (in the learned SOM)

  -> these will be the connected edges of the masked D-graph (V-neighbors in data space);

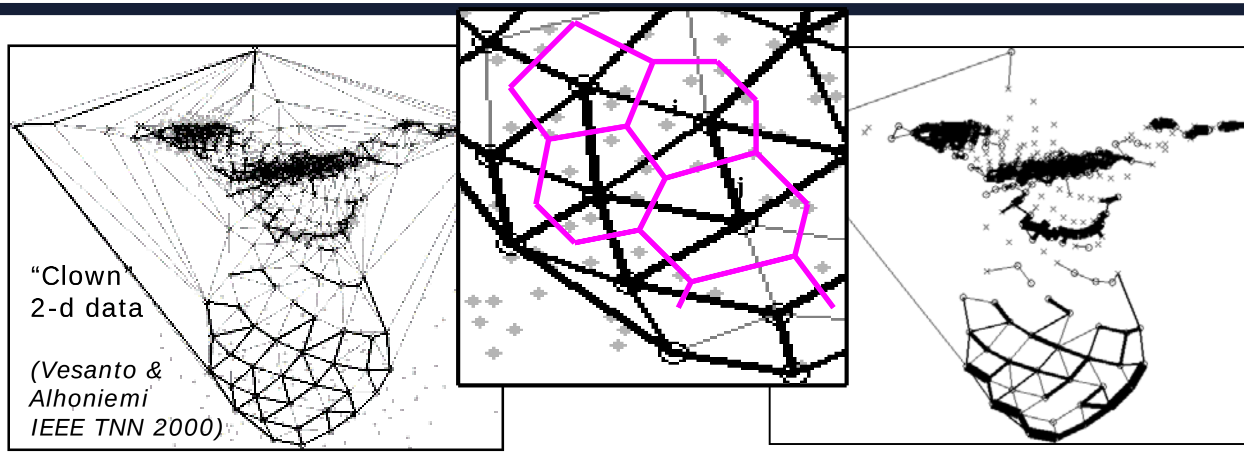
  -> pairs of prototypes that are not chosen together as BMU and 2nd BMU by any data point, will not be connected in the D-graph.

  - The generated masked D-graph can be stored as an *Adjacency matrix* A that has a 1 at A(i,j) if prototypes i and j are connected (selected together) by at least one data point.

Connectivity (CONN) graph representation

(Taşdemir & Merényi, IEEE TNN 2009)

Masked Delaunay graph - binary



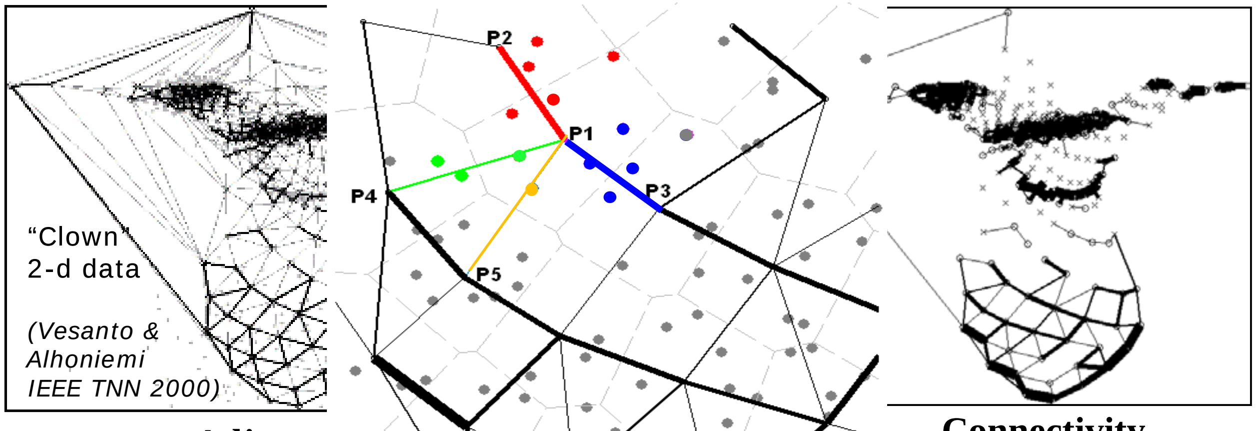
Adjacency **Connectivity**

Weighted masked Delaunay graph

Connectivity (CONN) graph representation

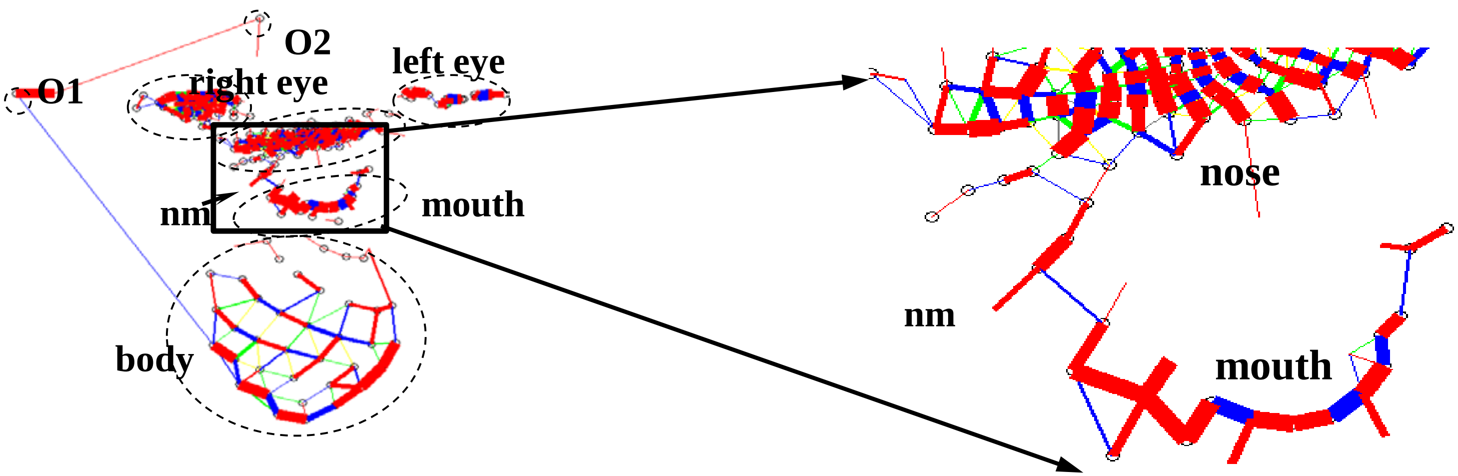
(Taşdemir & Merényi, IEEE TNN 2009)

Masked Delaunay graph - binary



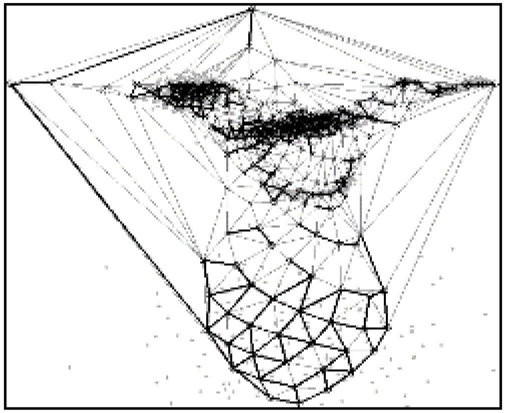
Adjacency **Connectivity**

Weighted masked Delaunay graph

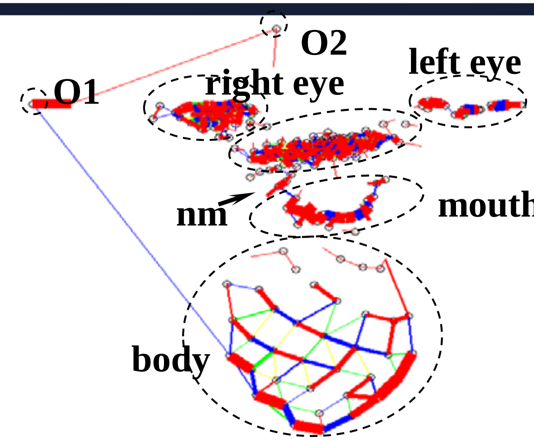


Connectivity

Connectivity (CONN) graph representation & visualization in data space vs on the SOM lattice *(Taşdemir & Merényi, IEEE TNN 2009)*

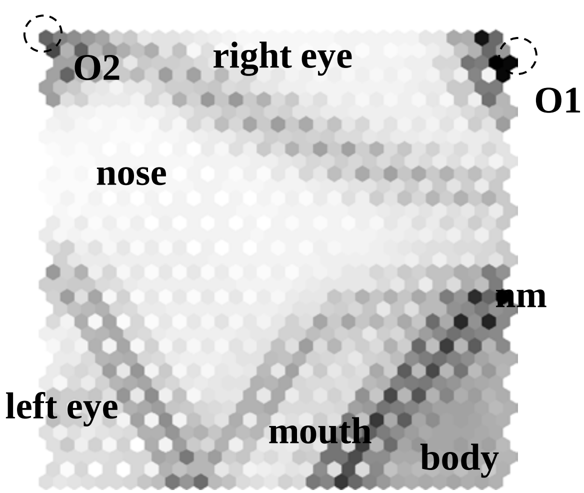


Adjacency

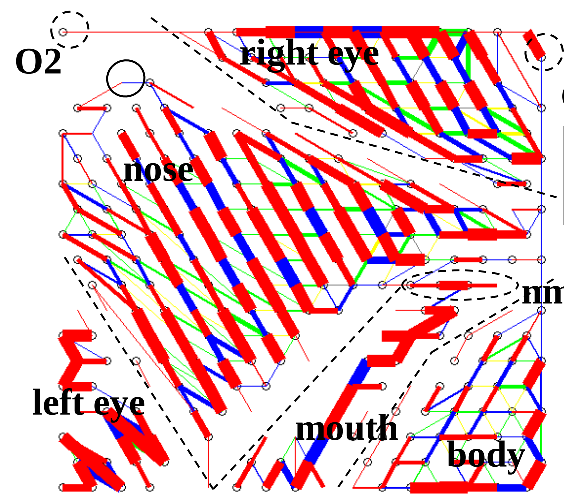


Connectivity

Cannot be shown for data dim > 2



A classic representation: U-matrix ($\Sigma|| w_i - w_j ||$) overlain the SOM grid



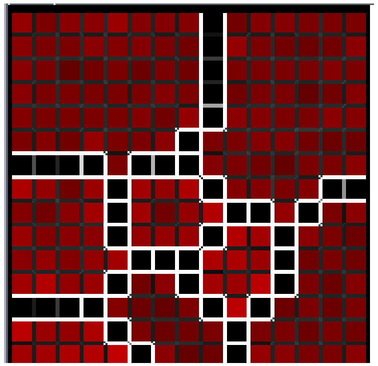
Can be shown for data dim >2

Bonus: CONN shows topology violations

CONNectivity matrix draped over SOM grid: The SOM / CONN portrait of the Clown (hexagonal SOM lattice)

Maximum entropy mapping with Conscience SOM *(De Sieno, 1988)*
Learning a 6-d synthetic data set with 8 known classes

15 x 15 SOM lattice

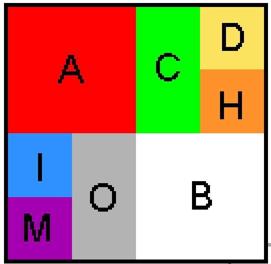


The knowledge of SOM
mU-matrix

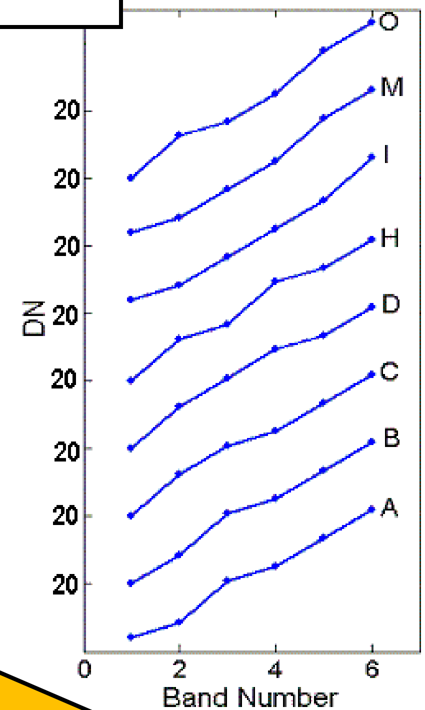
"Fences" between SOM cells show degree of dissimilarity of prototypes on gray scale (white is large difference)
-> outline cluster boundaries

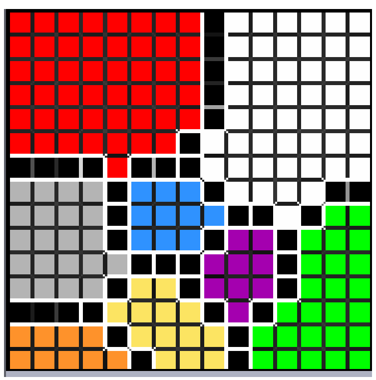
Intensity of monochrome red is proportional to # data points mapped to the prototype in each SOM cell
-> shows even distribution

Spatial map of classes in image cube



Spectral signatures, offset for clarity





The truth labels super-imposed on the SOM

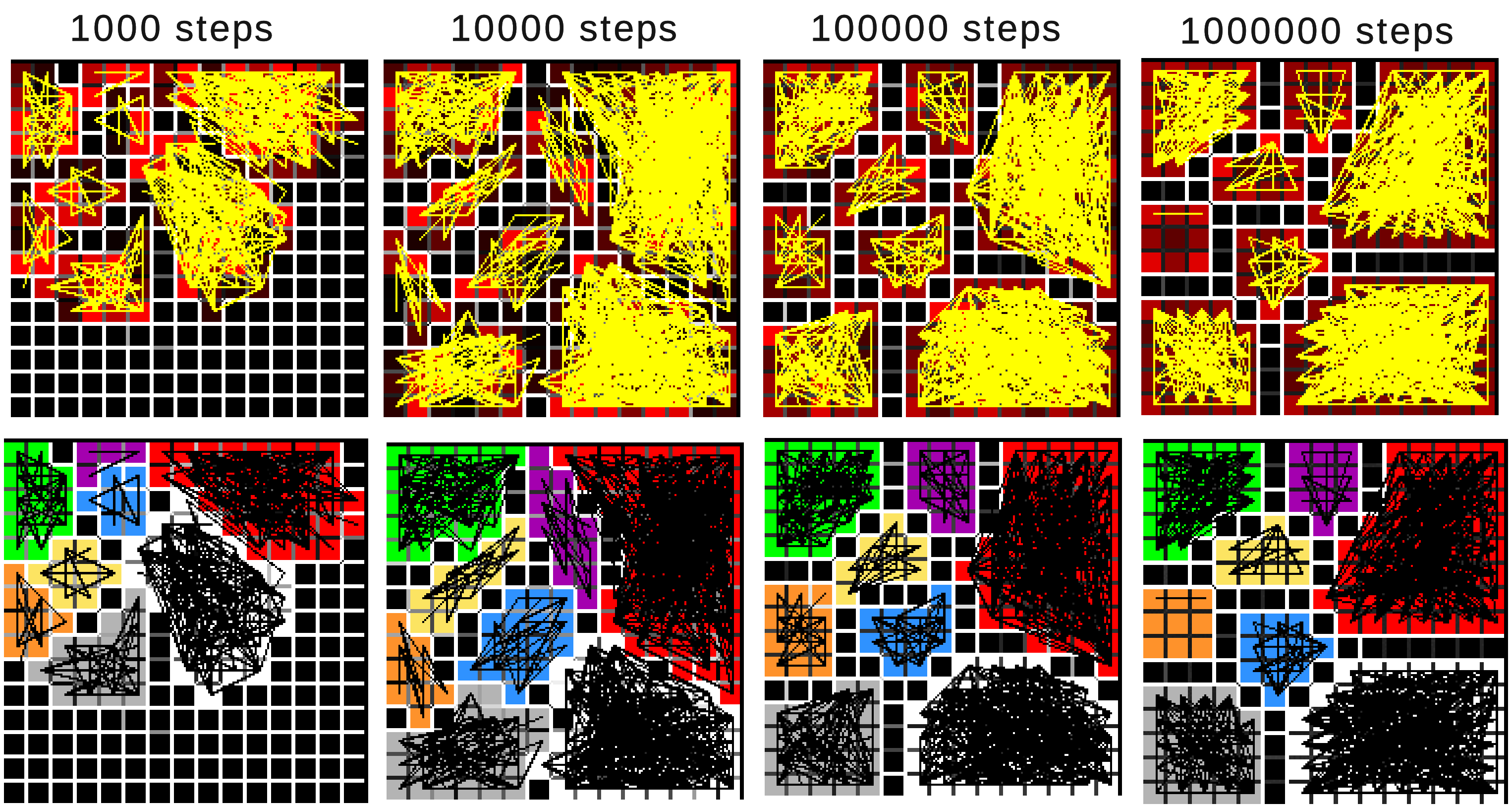
# Prototypes in clusters:	Linear ⟷	# Data points in clusters:
A:48		A: 4096
B:49		B: 4096
C:25		C: 2048
O:21		O: 2048
M:9		M: 1024

Discovery of small clusters with "SOM magnification"

Deviations from the exact 4:2:1 proportions are due to the small size of the SOM, integer arithmetic, and the formation of inter-cluster gaps

Merényi et al. IEEE TNN 2007

Monitoring the learning of the 8-class synthetic data with TopoView *(Merényi, Taşdemir, Zhang, Springer, LNAI 5400. 2009)*

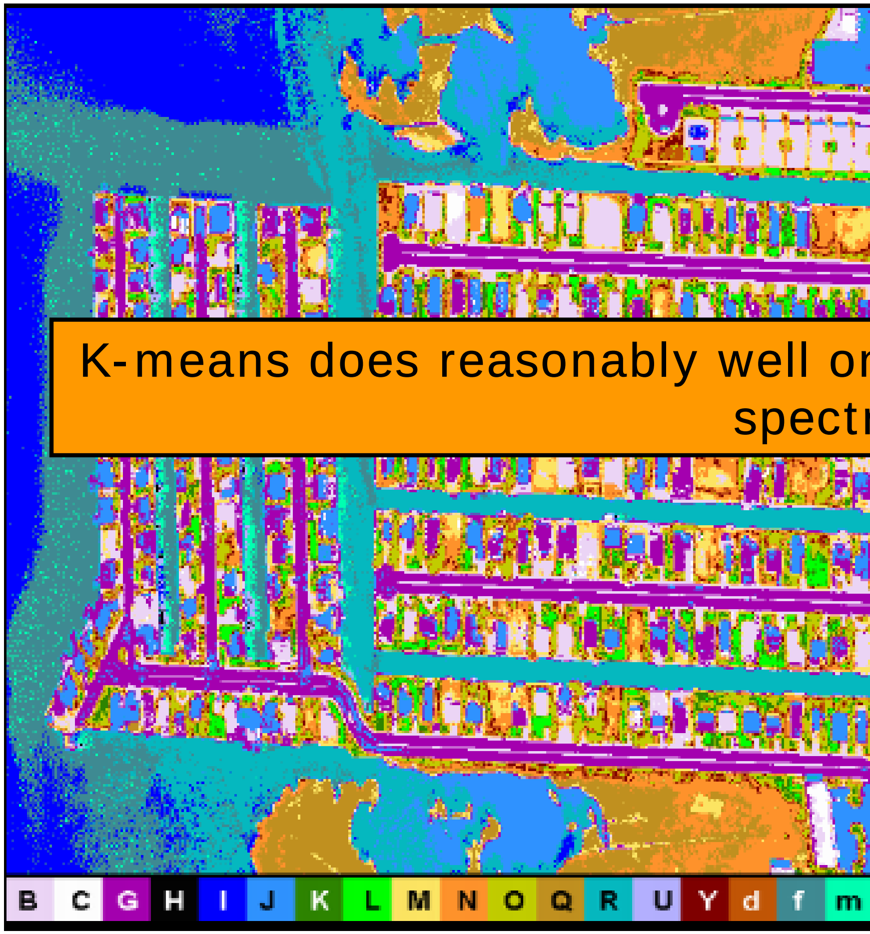


Top: All topology violating connections superimposed on mU-matrix
Bottom: Same with majority truth labels overlain.

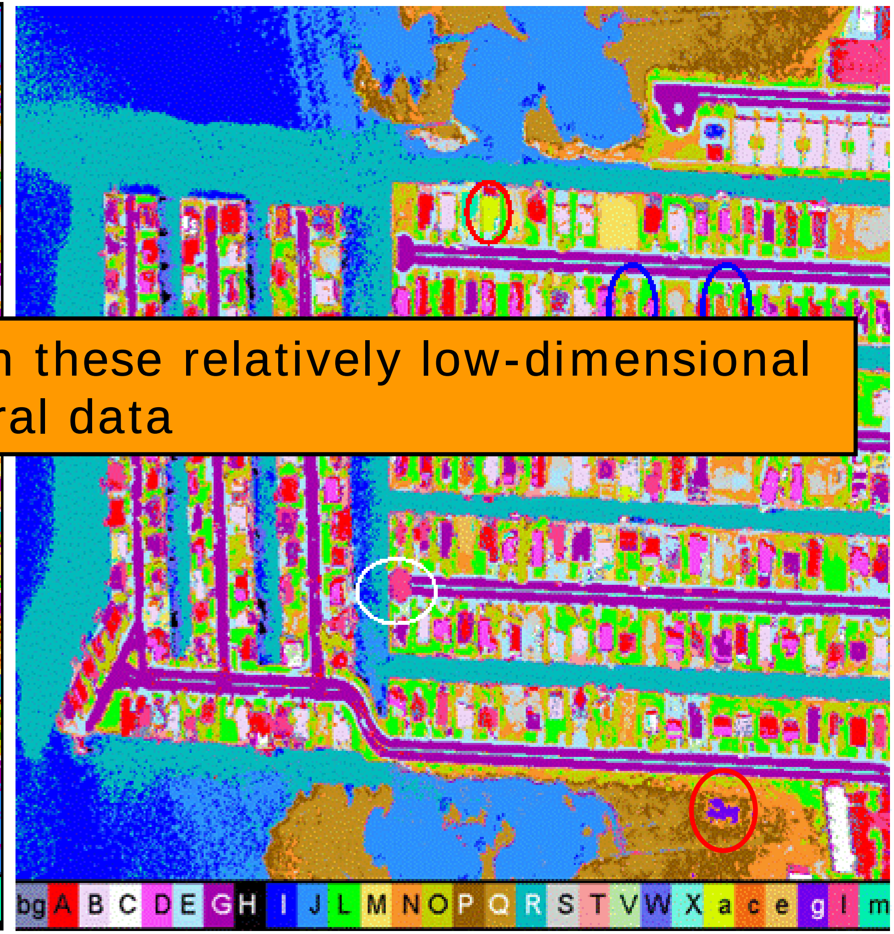
Learning of topography not yet complete but SOM state is perfect for cluster capture.

SOM vs K-means clustering of multi-spectral image (8-d spectra as input data vectors)

18 clusters, found by ISODATA (K-means)



28 SOM clusters, extracted with CONN visualization

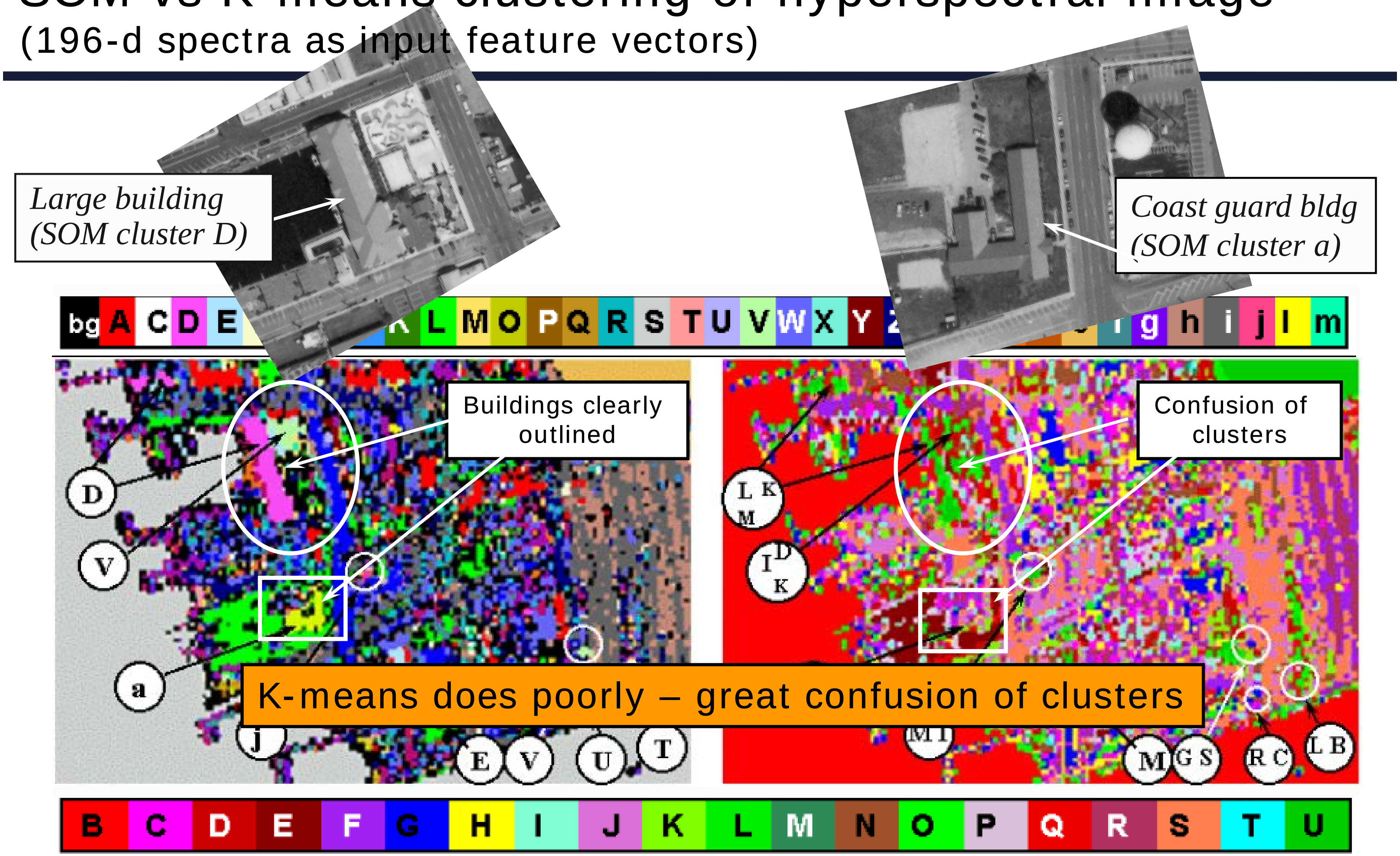


K-means does reasonably well on these relatively low-dimensional spectral data

Merényi et al., URBAN 2007

Data: Ocean City, Maryland,
Daedalus AADS 1260 scanner, bands 3 – 10
(Csathó, Krabill, Lucas and Schenk, 1998)

SOM vs K-means clustering of hyperspectral image
(196-d spectra as input feature vectors)

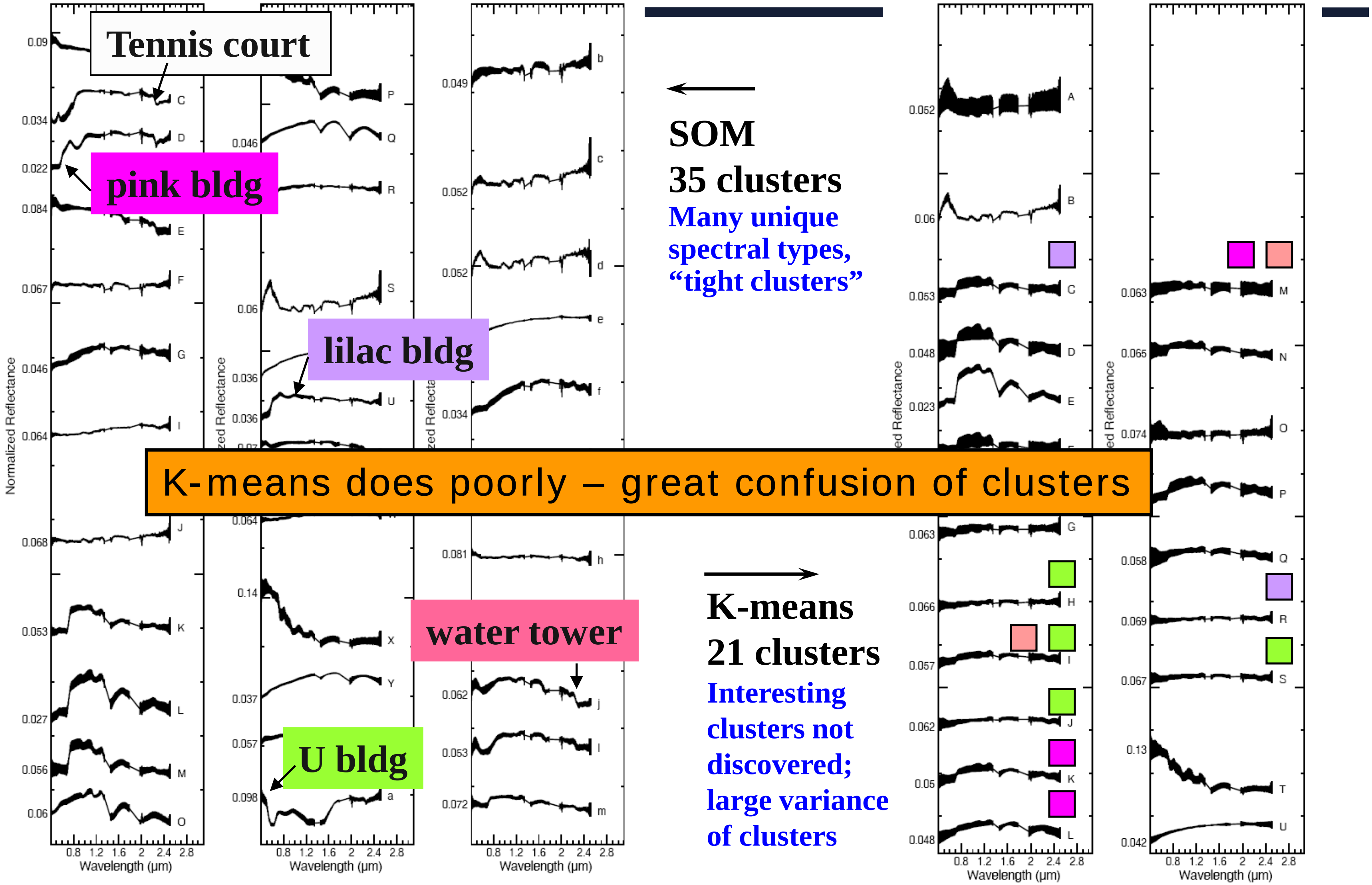


35 SOM clusters **21 K-means clusters**

Merényi et al., URBAN 2007

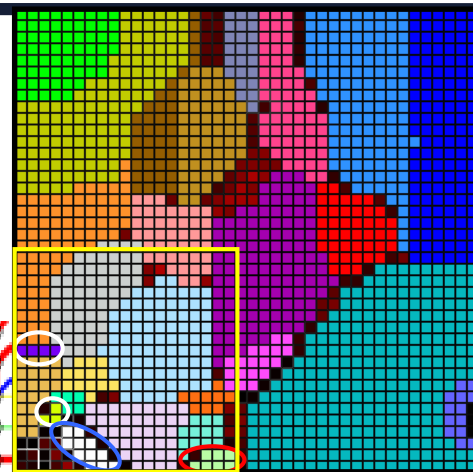
Spectral Statistics of Clusters, Ocean City

196-band Hyperspectral Image of Urban Area



CONN vs mU-matrix for identifying SOM clusters - effect on real data of moderate complexity

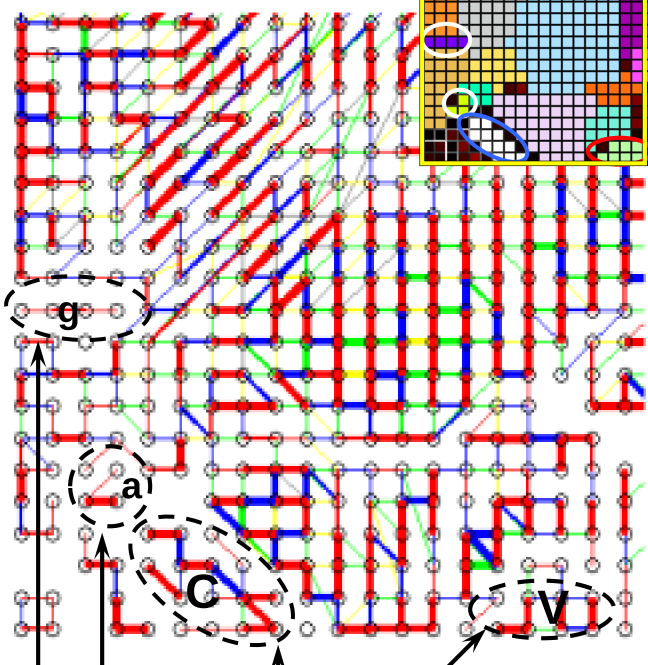
Clusters obtained from an 40 x 40 SOM



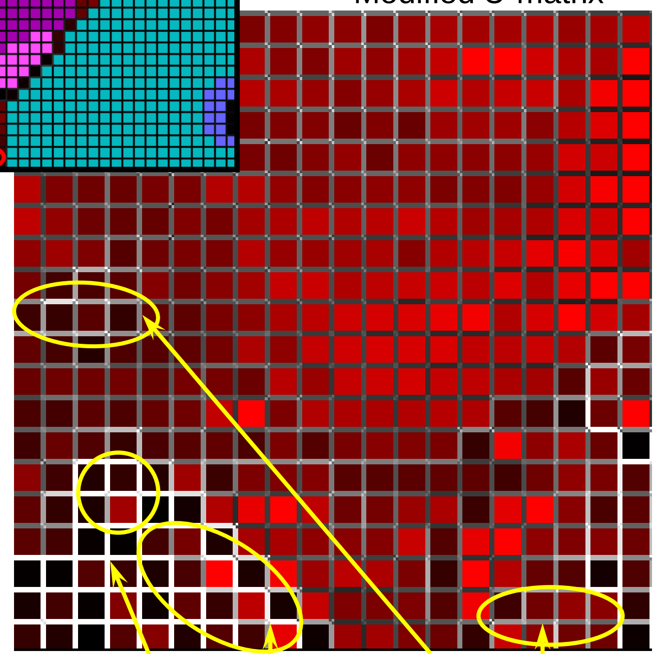
Data: Ocean City, Maryland

Daedalus AADS 1260 scanner
bands 3 – 10 used
(Csathó, Krabill, Lucas and Schenk, 1998)

CONNvis



Modified U-matrix



Clear separation of rare clusters from other clusters

Rare clusters, hard to separate from this representation

Taşdemir & Merényi, IEEE TNN 2009

ALMA hyperspectral image of HD 142527

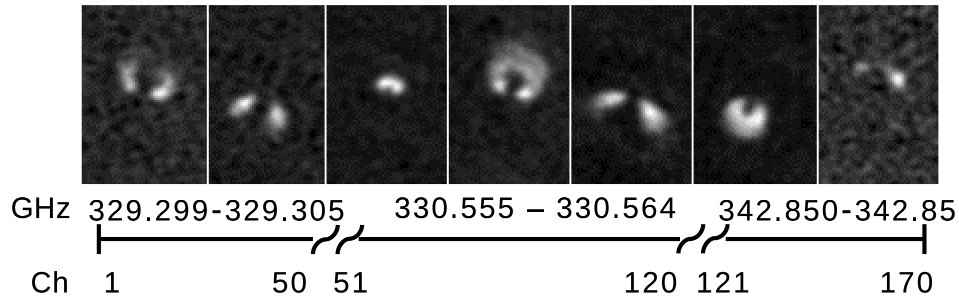
Data credit: JVO, project 2011.0.00318.5

ALMA: Atacama Large Millimeter Array

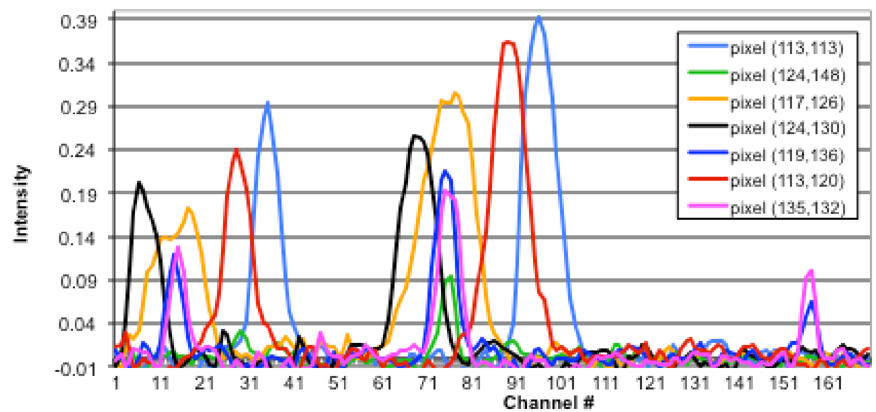
66 dishes at ~ 5,500 m

Artist's concept of planet formation in HD 142527

Sample image planes from ALMA Band 7, HD 142527



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



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)*

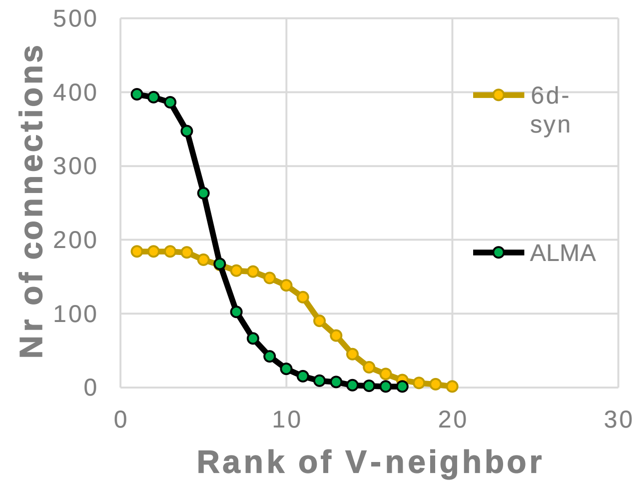
Data and connectivity statistics

<u>Passport, ALMA data</u>

# vectors	5,625
# dim	170
# clusters	? (many)
Noise	Moderate
Similarity	Variable
# V neighbors	17
# R-local	2

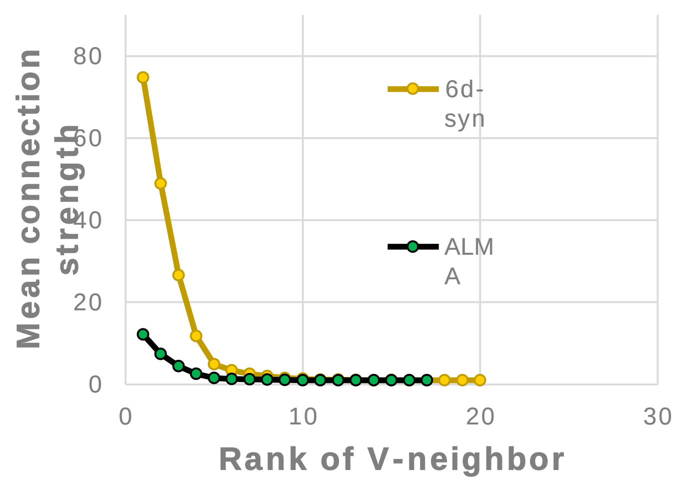
<u>Passport, 6d 8-class data</u>

# vectors	16,386
# dim	6
# clusters	8
Noise	Moderate
Similarity	High
# V neighbors	20
# R-local	2

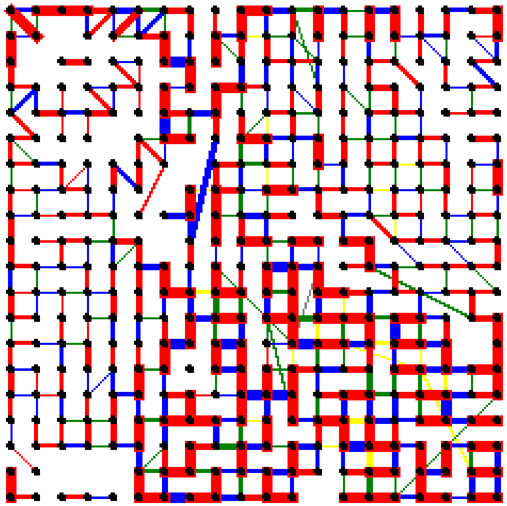


15 x 15 SOM

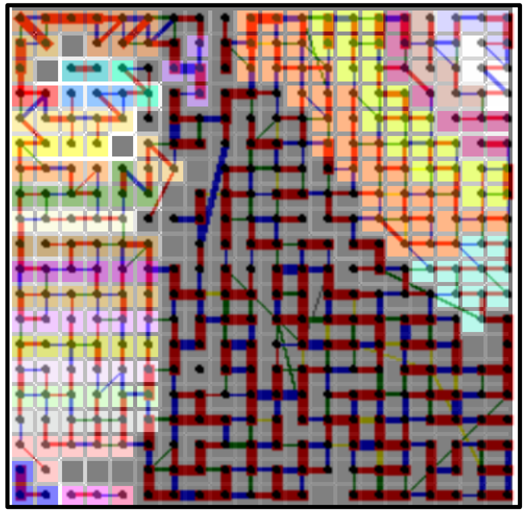
20 x 20 SOM

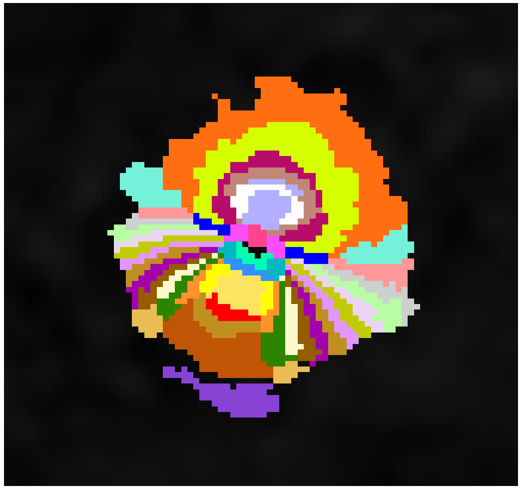


Clusters from 20 x 20 SOM of ALMA image

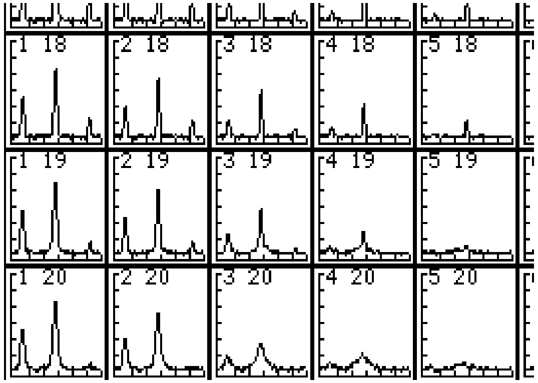


CONNvis





The clusters shown in the disk



Prototypes in the lower left corner

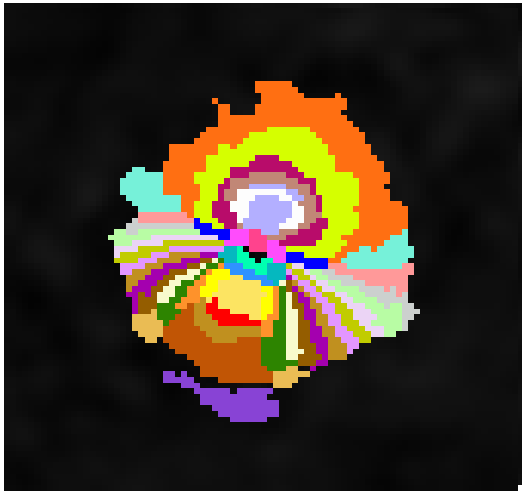
The extracted clusters in the SOM

Center detail

SOM clusters of HD 142527

First-cut hyperspectral analysis of ALMA data compared to Casassus et al. (2013)

Simultaneous CS, ^{13}CO, & $C^{18}O$



SOM clusters from 170-channel hyperspectral cube of protostar HD 142527

Coloring is arbitrary, not a heat map.

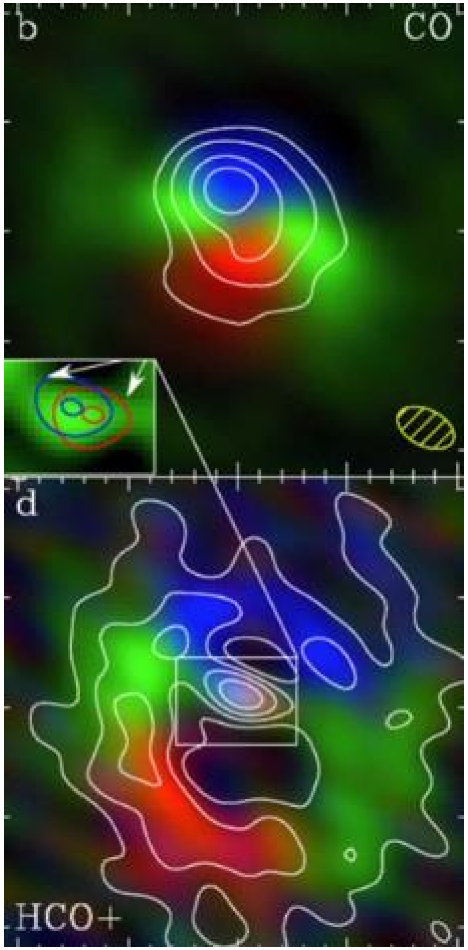
Thanks: Al Wootten
Data Credit: JVO Project Code 2011.0.00318.5

SOM clusters

- capture general Doppler structure found in single-species lines.
- incorporate line intensities, widths, shapes, et cetera, as well as Doppler.
- contain more structure than single-line analysis, and more than can be shown here.

.

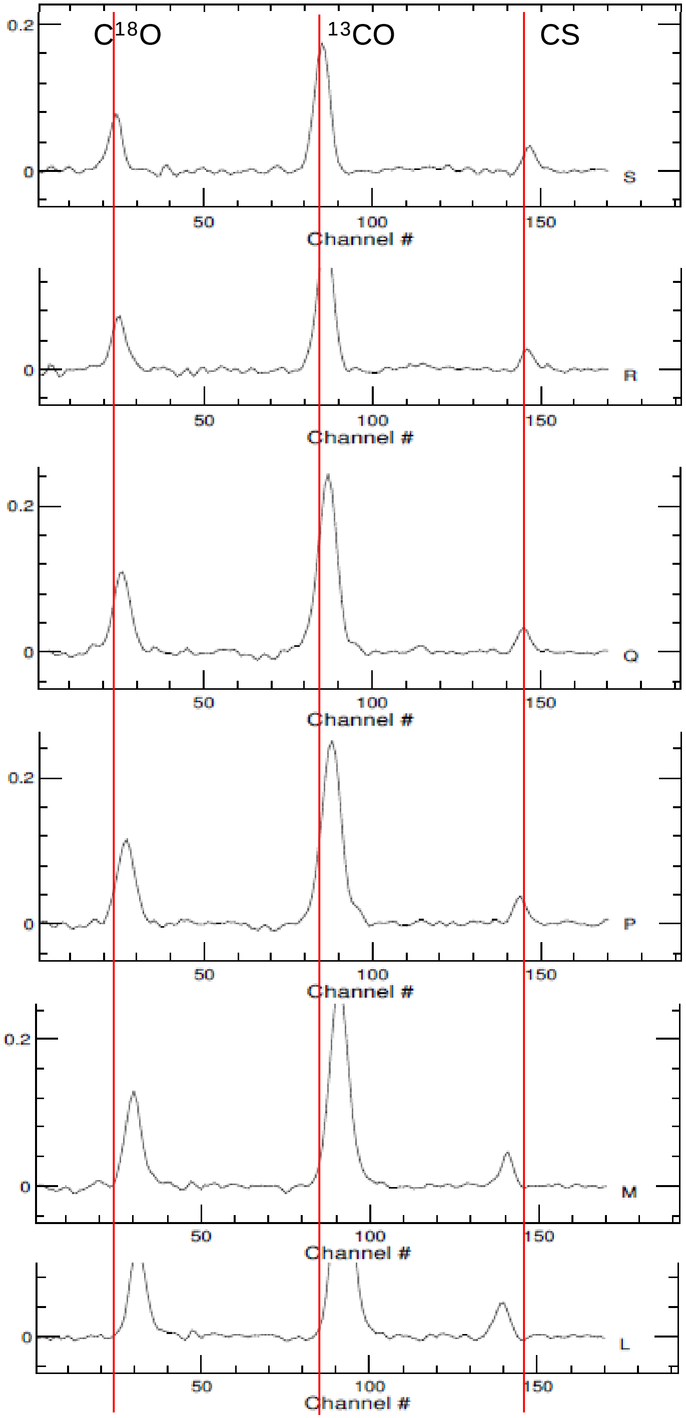
Single-line Doppler CO & HCO+

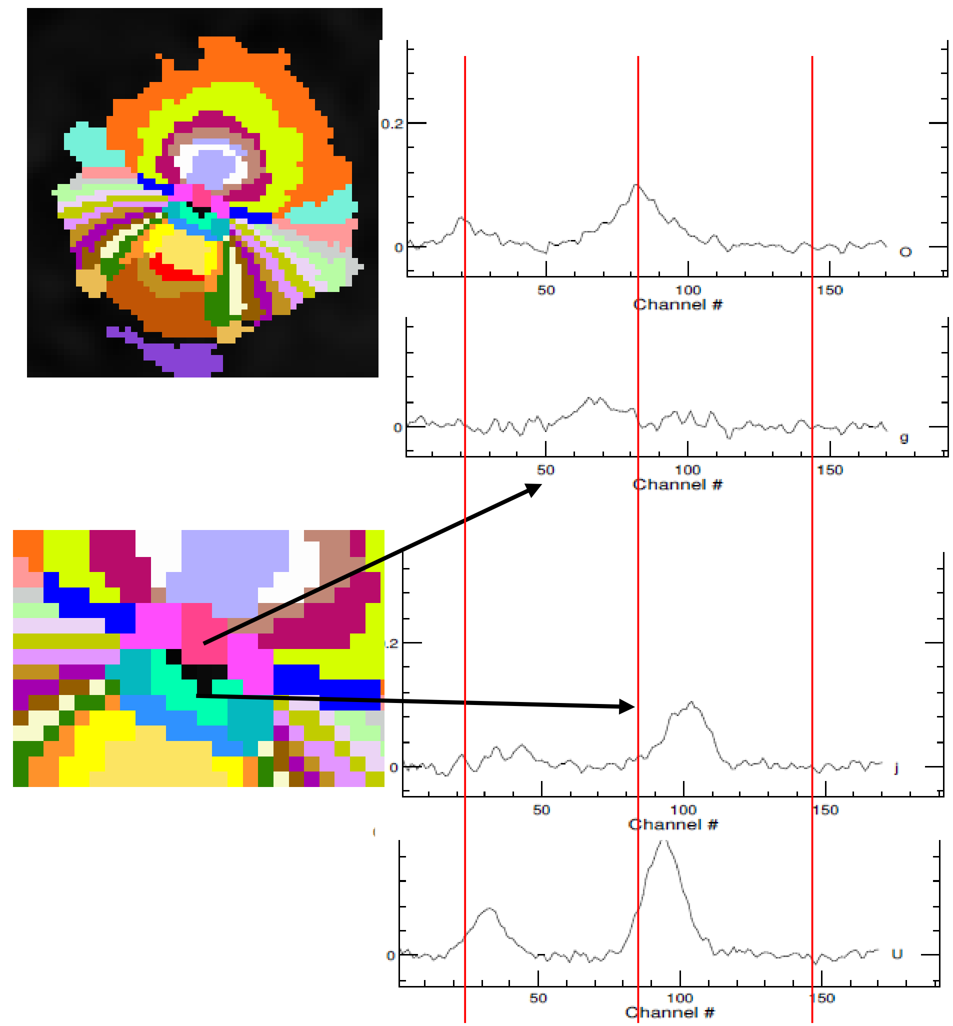


Extracted from: Casassus, et al., 2013, **Nature**, **493**,191.

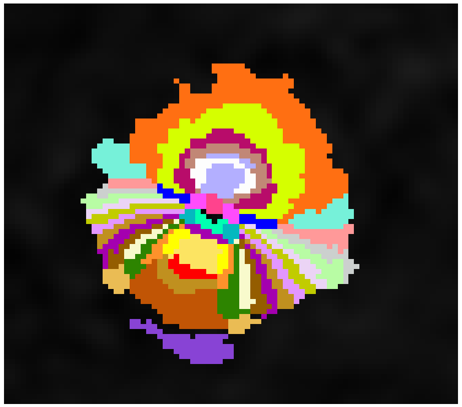
E. Merényi, Rice U
erzsebet@rice.edu

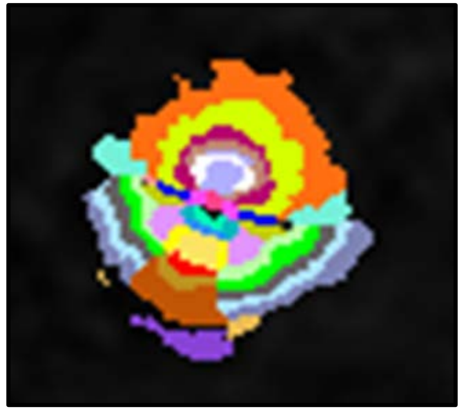
Mean cluster spectra

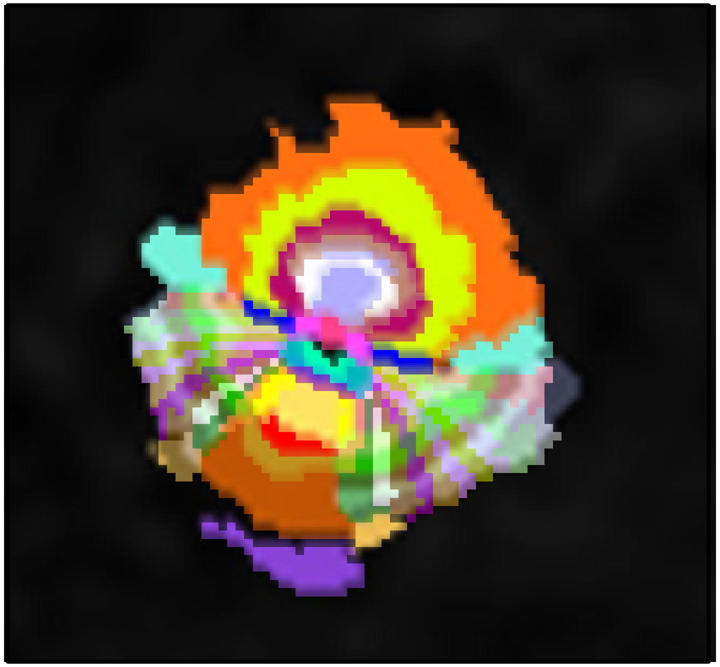




Layered Knowledge?







Superimposed cluster maps

Conclusions

- SOMs are powerful for structure discovery in complex data
  - The CONN(ectivity) similarity metric improves the segmentation of prototypes compared to distance-based metrics
- ALMA hyperspectral data cubes – a new type of complexity
- Showed intricate structure identified in ALMA data
- Emerging structure makes good sense, but it is also more complex than CONN seems to capture from the SOM
  - Motivates further development of metrics & visualization
- New types of astronomy data can present surprises we may not be ready for and will provide exciting opportunities for CI and ML research. ☺

Thank you

E. Merényi, Rice U
erzsebet@rice.edu

References

http://www.ece.rice.edu/~erzsebet/publications-EMerenyi.pdf

- E. Merényi, K. Taşdemir, L. Zhang (2009) Learning highly structured manifolds: harnessing the power of SOMs. 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.
- Howell, E. S., Merényi, E., L. A. Lebofsky (1994) Using Neural Networks to Classify Asteroid Spectra. *J. Geophys. Res*. 99 No. E5, pp. 10,847-10,865
- Merényi, E., E.S. Howell, L.A. Lebofsky, A.S. Rivkin (1997) Prediction of Water In Asteroids from Spectral Data Shortward of 3 Microns, *ICARUS* 129, pp 421- 439
- Teuvo Kohonen: Self-organizing Maps (Springer Series in Information Sciences S.). Springer-Verlag, 2001 (3rd Edition, ISBN: 3540679219)
- Martinetz, T. and Schulten, K. Topology Representing Networks, IEEE Trans. Neural Networks, 1994.
- 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.

References

http://www.ece.rice.edu/~erzsebet/publications-EMerenyi.pdf

- DeSieno, D., Adding a conscience to competitive learning. Proc. Int. Conf. Neural Networks, New Your, July 1988, vol. I., pp I-117-I-124.
- Merényi, E., Jain, A., Villmann, Th. (2007) Explicit Magnification Control of Self-Organizing Maps for "Forbidden Data". *IEEE Trans. Neural Networks 18(3)* May, 2007, pp 786-797.
- Merényi, E., B. Csathó, and Taşdemir, K. (2007) Knowledge discovery in urban environments from fused multi-dimensional imagery *Proc. 4th IEEE GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (URBAN 2007)*, Paris, France, April 11-13, 2007. (invited paper). pp 1-13. DOI: 10.1109/URS.2007.371860 , IEEE Catalog number 07EX1577.
- Zhang, L., Merényi, E., Grundy, W. M., Young, E. Y. (2010) Inference of Surface Parameters from Near-Infrared Spectra of Crystalline H_2O Ice with Neural Learning, *Publications of the Astronomical Society of the Pacific*. Vol. 122, No. 893: pp. 839-852.