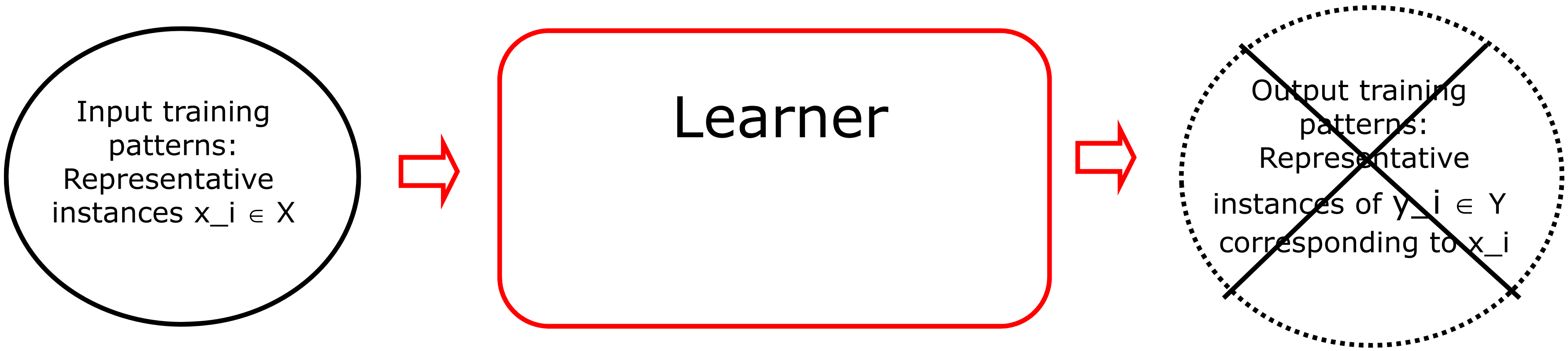
Magnifying (unknown) rare clusters to increase the chance of detection, using unsupervised learning

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

Learning Without a Teacher
(unsupervised learning)



Learning Without a Teacher
(unsupervised learning)

Input training patterns: Representative instances $x_i \in X$ ⇨ **Learner** ⇨ Model of the input space

An unsupervised learner captures some internal characteristics of the input data: structure, mixing components / latent variables, ...

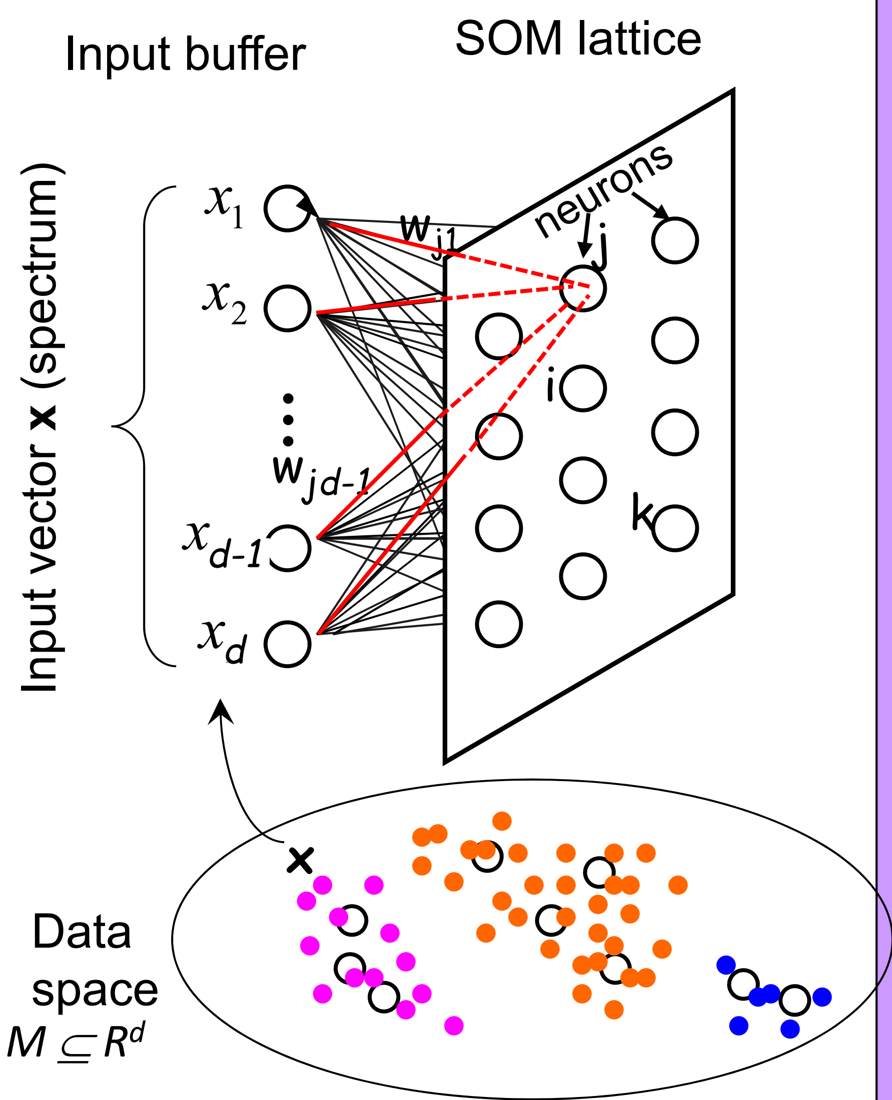
- Ex: clusters
- Ex: principal components
- Ex: independent components

- No (explicit) cost function

- Best for discovery: model-free
- Some “model-free” have implicit assumptions

E. Merényi, Rice U
erzsebet@rice.edu

Self-Organizing Map: model-free structure learner
Machine learning analog of biological neural maps in the brain



Formation of basic (Kohonen) SOM:

$\mathbf{x} = (x_1, x_2, \ldots, x_d) \in M \subseteq R^d$ **input pattern**

$\mathbf{w}_j = (w_{j1}, w_{j2}, \ldots, w_{jd})$ $j=1, \ldots, P$ **weight vector of neuron j (prototype j)**

Learning: cycle through steps 1. and 2. many times

1. <u>Competition</u>

Select a pattern x randomly.

Find winning neuron c as

$$c(\mathbf{x}) = \arg\min_j \|\mathbf{x} - \mathbf{w}_j\|, \; j=1, \ldots, P$$

Euclidean dist. in data space

2. <u>Synaptic weight adaptation / cooperation</u>

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + a(t)\, h_{j,c(\mathbf{x})}(t)\, (\mathbf{x} - \mathbf{w}_j(t))$$

for all $\mathbf{w}_j$ in *influence region* of node c in the SOM lattice, prescribed by $h_{j,c(\mathbf{x})}(t)$

h(t): most often Gaussian centered on node c

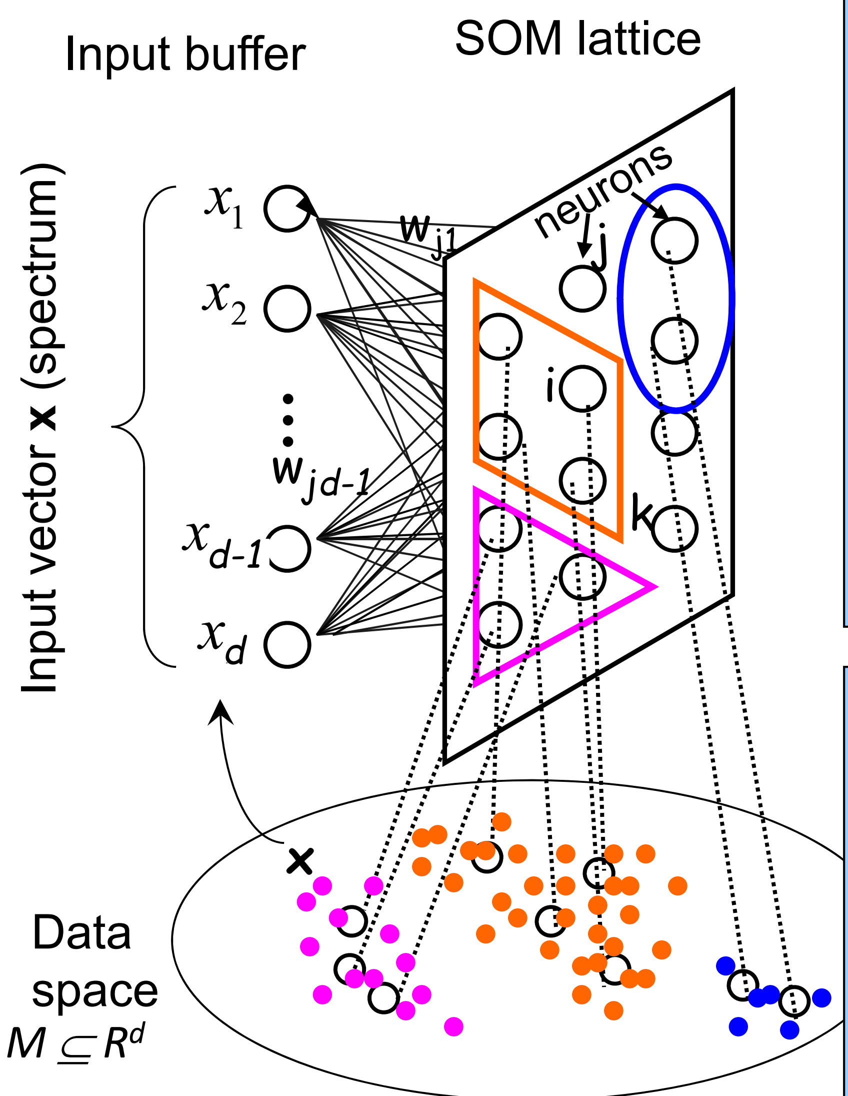
$$h_{j,c(\mathbf{x})}(t) = \exp(-(c-j)^2/\sigma(t)^2)$$

Manhattan dist. In SOM lattice

E. Merényi, Rice U
erzsebet@rice.edu

Self-Organizing Map: model-free structure learner

Machine learning analog of biological neural maps in the brain



Two simultaneous actions:

- ***Adaptive*** Vector Quantization (n-D binning): puts the prototypes in the "right" locations, encoding salient properties of data distribution
- Ordering the prototypes on the SOM grid according to similarities: expresses the topology on a low-dimensional lattice

Finding the prototype groups: post-processing – segmentation of the SOM based on the SOM's knowledge (both the summarized distribution and topology relations)

Summarization of N data vectors by ***O(sqrt(N)) prototypes***;

E. Merényi, Rice U
erzsebet@rice.edu

Map magnification in SOMs (Magnification of Vector Quantizers, in general)

*pdf*s of SOM weight vectors (VQ prototypes) and inputs related by

$$Q(\mathbf{w}) = \text{const} \cdot P(\mathbf{w})^{\alpha}$$

where

Q(**w**) is *pdf* of prototype vectors

P(**w**) is *pdf* of input vectors

and

α is the Magnification Exponent – an inherent property of a given Vector Quantizer

(Zador, 1982; Bauer, Der, and Hermann, 1996)

What does α mean?

If data dimensionality = d,

- $\alpha = 1$ equiprobabilistic mapping
  (max entropy mapping, information theoretical optimum)
- $\alpha = d/(d+2)$ minimum MSE distortion quantization
- $\alpha = d/(d+p)$ minimum distortion in p norm
- $\alpha < 0$ enlarges representation of low-frequency inputs

- Kohonen's SOM (KSOM) attains $\alpha = 2/3$ (under certain conditions) *(Ritter and Schulten, 1986).* Not ideal by any of the above measures.

- Conscience SOM (CSOM) attains $\alpha = 1$ *(D. DeSieno, 1988)*

- α of KSOM or CSOM cannot be changed (not a parameter of the algorithm);

BDH: Modification of KSOM to allow control of α

(Bauer, Der and Hermann, 1996)

KSOM learning rule: $w_j(t+1) = w_j(t)+\varepsilon(t)\, h_{j,r(v)}(t)\,(v - w_j(t))$

winner index

Time-decreasing learning rate

Idea: Modify the learning rate ε(t) in KSOM to

force the local adaptabilities to depend on the input density P at the lattice position, r, of prototype w_r. Require

$$\varepsilon_r = \varepsilon_0\, P(w_r)^m,$$

where m is a free parameter that will allow control of α.

How to do this when $P(w_r)$ is unknown?

Use the information already acquired by the SOM and exploit

$$P(w_r) \propto Q(w_r)P'(r)$$

where P'(r) is the winning probability of the neuron at r.

Approximate $Q(w_r)$ and $P'(r)$ by quantities the SOM has learnt so far

Compute $P(w_r) \propto Q(w_r)P'(r)$:

$Q(w_r) \propto 1/vol$ vol = Volume of the Voronoi polyhedron of w_r

$vol \propto |v - w_r|^d$

$P'(r) \propto 1/(\Delta t_r)$,

$\Delta t_r \propto$ (present t value – last time neuron r won)

Substitute into $P(w_r) \propto Q(w_r)P'(r)$ to get

$$\varepsilon_r(t) = \varepsilon_0(t)\left[\frac{1}{\Delta t_r}\left(\frac{1}{|v - w_r|^d}\right)\right]^m \qquad (1)$$

Update weight vectors (prototypes) of ALL SOM lattice neighbors by using ε_r of the winning neuron.

E. Merényi, Rice U
erzsebet@rice.edu

Controlling α through m in the learning rate formula

- Given α = 2/3 for KSOM, it can be shown that a "desired" SOM magnification with exponent α' is related to m as

  $$Q(\mathbf{w}) = \text{const} \cdot P(\mathbf{w})^{\alpha'} = \text{const}\ P(\mathbf{w})^{(2/3)*(m+1)}$$

- Now we have a free parameter to control α

- EXAMPLE: to achieve max entropy mapping, we want α' = 1.

  α' = 2/3 (m+1) = 1 -> set m = 3/2-1 = 0.5 in eq. (1)

- EXAMPLE: to achieve α' = -1 negative magnification, set

  m = -3/2 -1 = -2.5

Limitations of the BDH algorithm

Theory guarantees success only for

1. 1-D input data
2. n-D data, if and only if $P(\mathbf{v}) = P(v_1)P(v_2)\ldots P(v_n)$

(i.e., the data are independent in the different dimensions)

1 and 2 $\rightarrow$ "Allowed" data

Rest $\rightarrow$ "Forbidden" data

Central question:

Can BDH be used for "forbidden" data?

Carefully designed controlled experiments suggest YES.

(Merényi, Jain, Villmann, IEEE TNN 2007).

Magnification control for higher-dimensional data

I. Noiseless, 6-D 5-class synthetic data cube

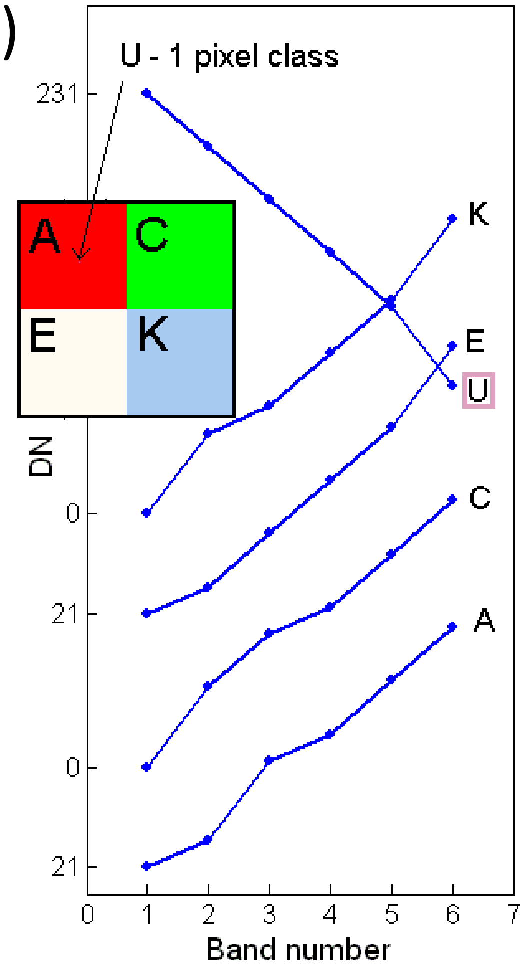
128×128 pixel image where a 6-D vector is associated with each pixel (16,384 6-D patterns)

5 classes:

Class	No. of inputs
A	4095
U	**1** (rare class)
C	4096
E	4096
K	4096

$0.004 \leq$ Pairwise correlation coefficients ≤ 0.9924

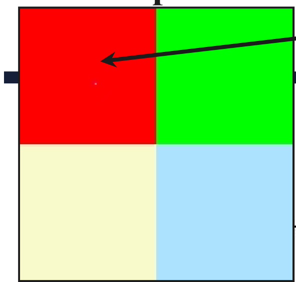
$\Rightarrow$ "Forbidden" data



(Merényi et al. IEEE TNN 2007)

E. Merényi, Rice U
erzsebet@rice.edu

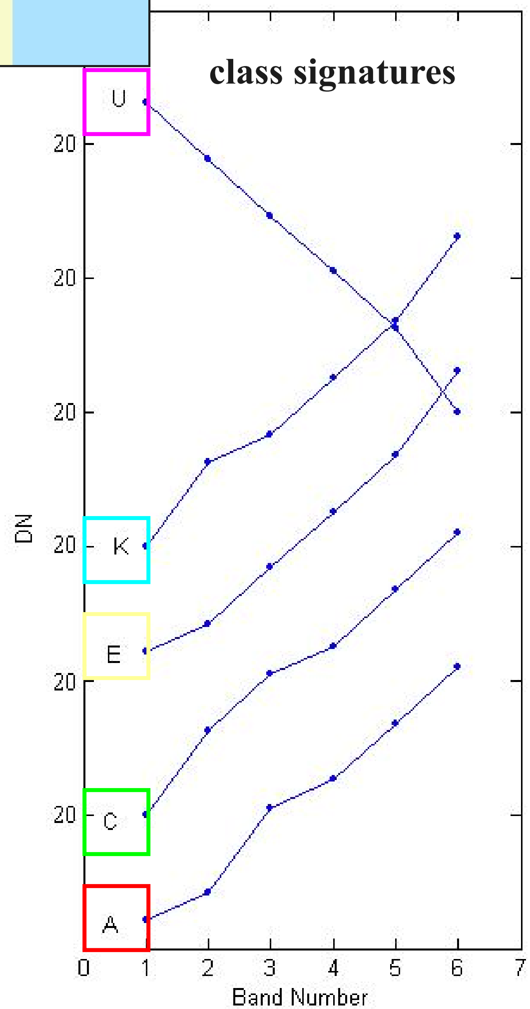
128 x 128 px image data cube
6-D spectrum (feature vector)
at each pixel location



1-px class U

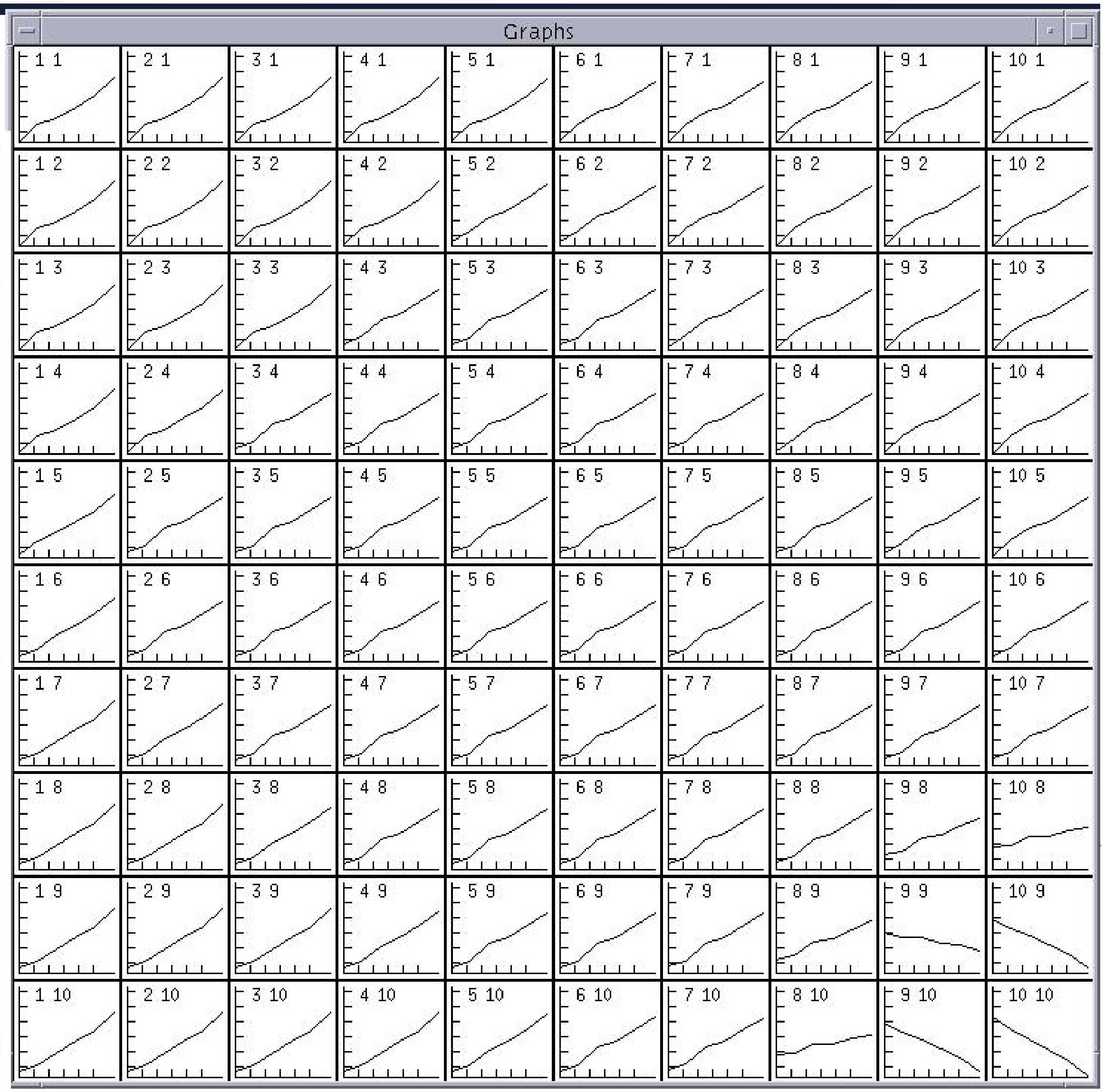
5 spectral classes
synthetic, noiseless

class signatures



SOM Visualization for >3-D Data

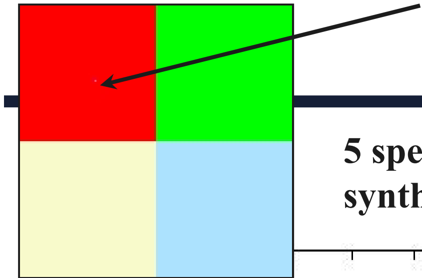
Weight vectors of 10 x 10 KSOM, after learning



Merényi et al. IEEE TNN 2007

128 x 128 px image
6-D spectra

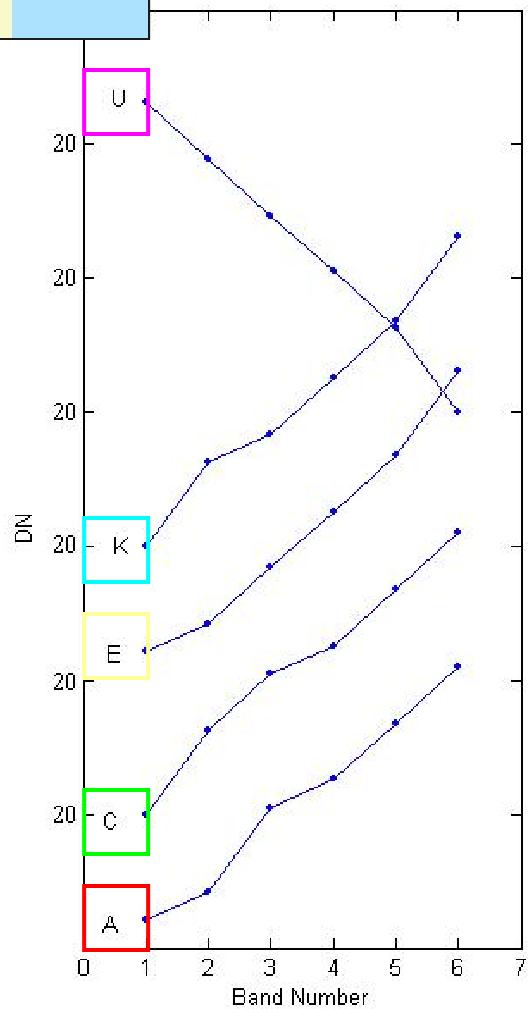
1-px class U

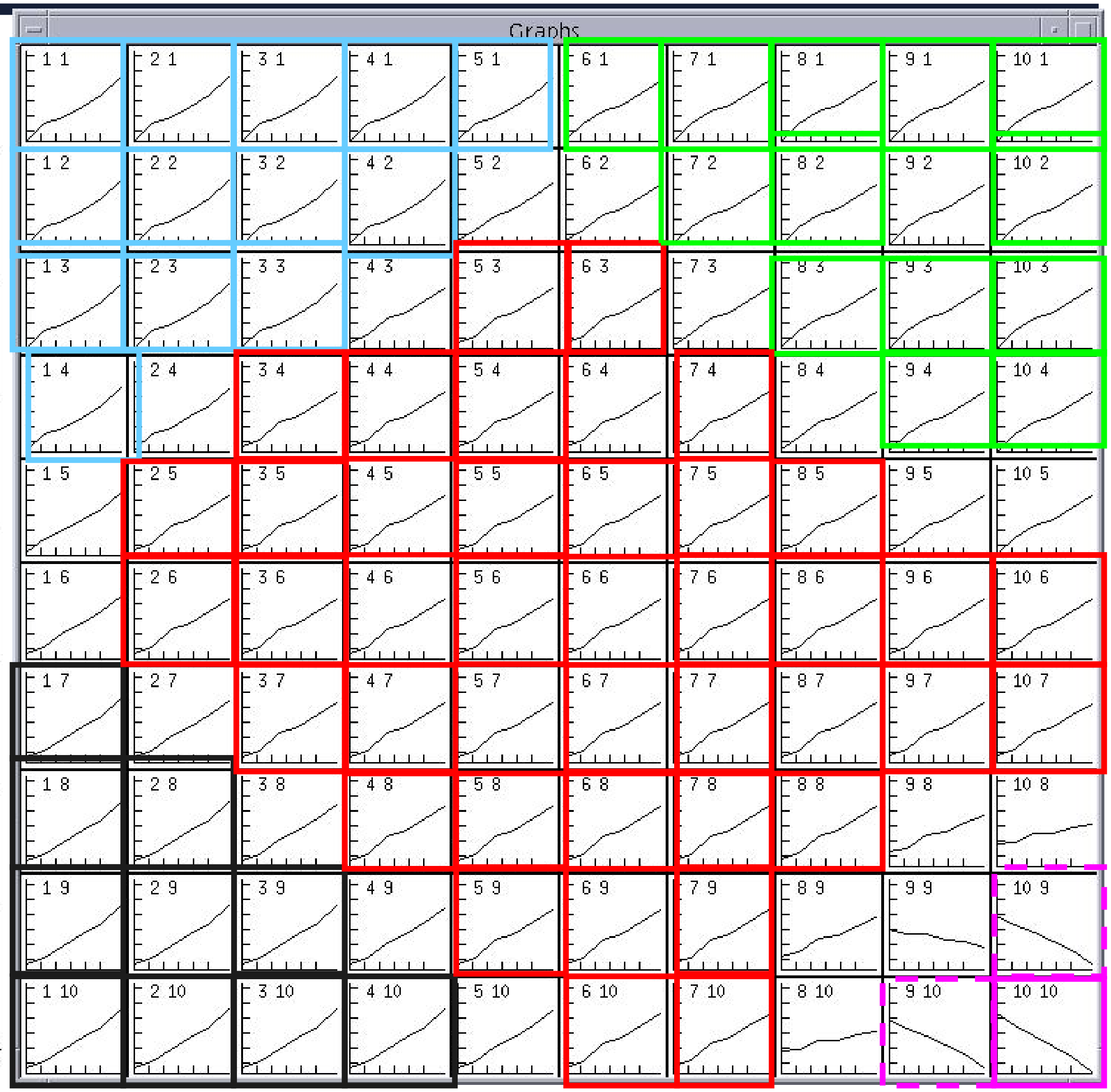


SOM Visualization, for >3-D Data

Weight vectors of 10 x 10 KSOM, after learning

5 spectral classes
synthetic, noiseless



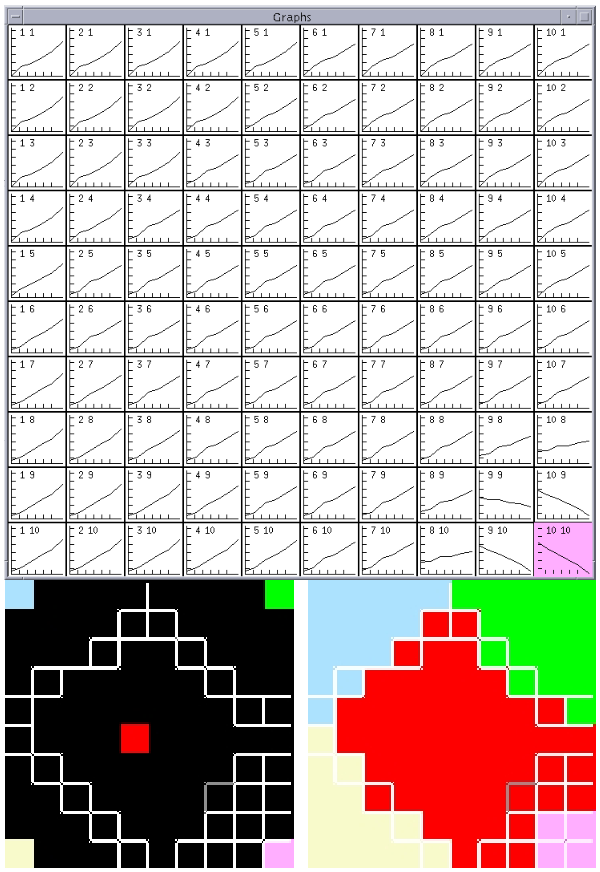


Merényi et al. IEEE TNN 2007

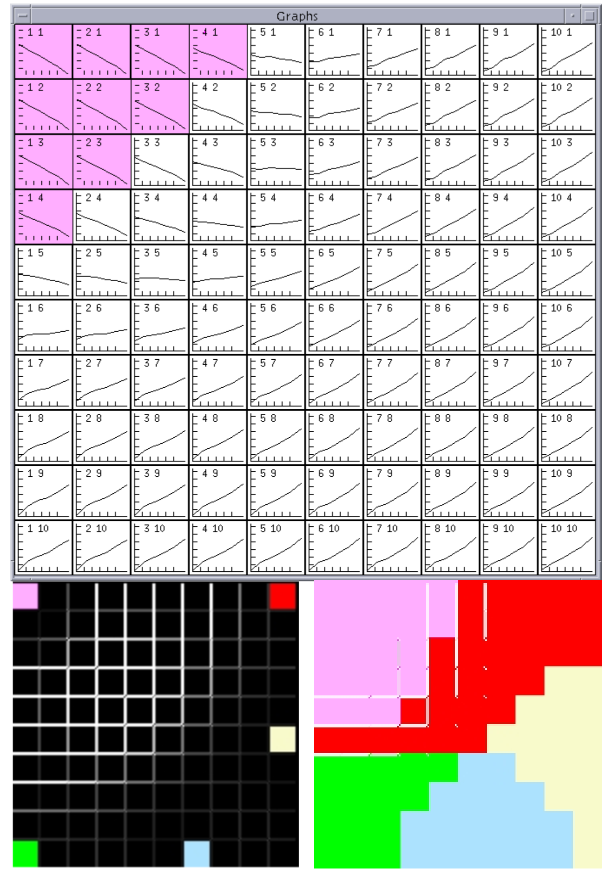
SOM learning without and with magnification

I: Noiseless, 6-D 5-class synthetic data cube

KSOM (no magnification)



BDH with $\alpha_{desired}$ = -0.8



Only 1 PE represents the rare class U

U now represented by 10 PEs!

(PE = Processing Element = neuron)

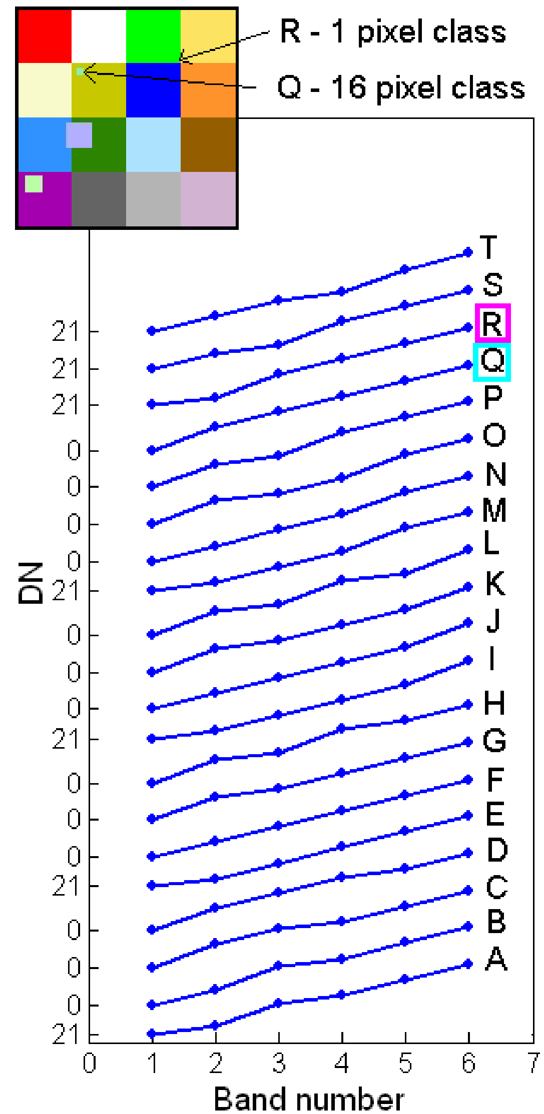
(Merényi et al. IEEE TNN 2007)

Magnification control for higher-dimensional data
II. Noiseless, 6-D 20-class synthetic data set

128 × 128 pixel image where each pixel is a 6-D vector (16,384 6-D patterns).
20 classes:

Class	No. of inputs
A,B,D,E,G,H,K,L,N,O,P	1024
C	1023
F	1008
I	979
J	844
M	924
Q	**16**
R	**1**
S	100
T	225

$0.008 \leq \rho \leq 0.6 \Rightarrow$ "Forbidden data"

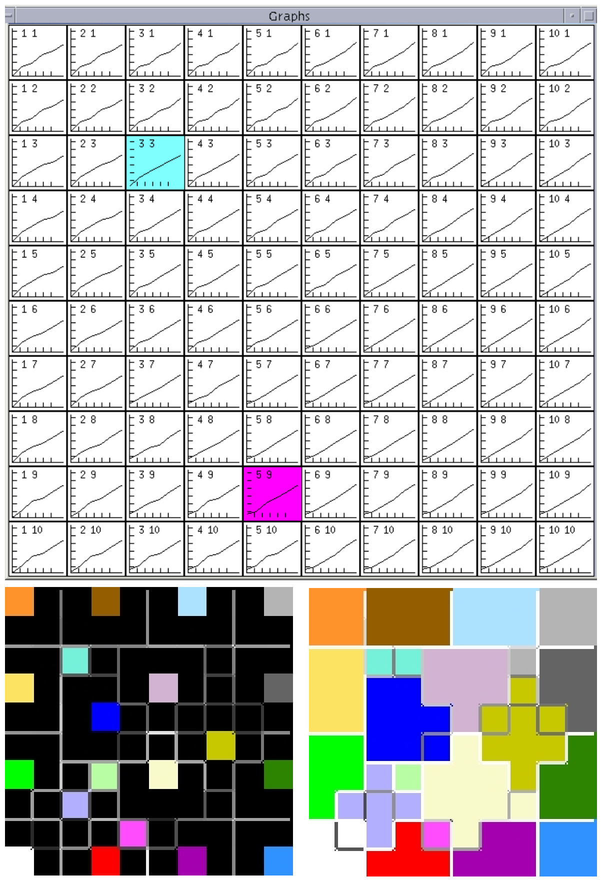


(Merényi et al. IEEE TNN 2007)

SOM learning without and with magnification

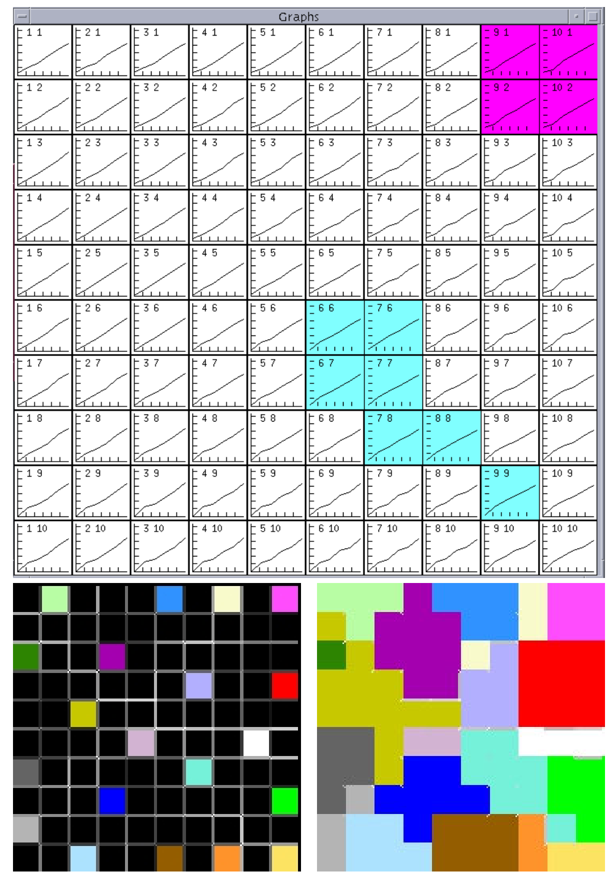
II: Noiseless, 6-D 20-class synthetic data set

KSOM (no magnification)



R: 1PE, Q: 1 PE

BDH with $\alpha_{desired}$ = -0.8



R: 4 PEs, Q: 7PEs

(Merényi et al. IEEE TNN 2007)

E. Merényi, Rice U
erzsebet@rice.edu

$\alpha < 0$ magnification for 8-D real data: discovery of rare clusters

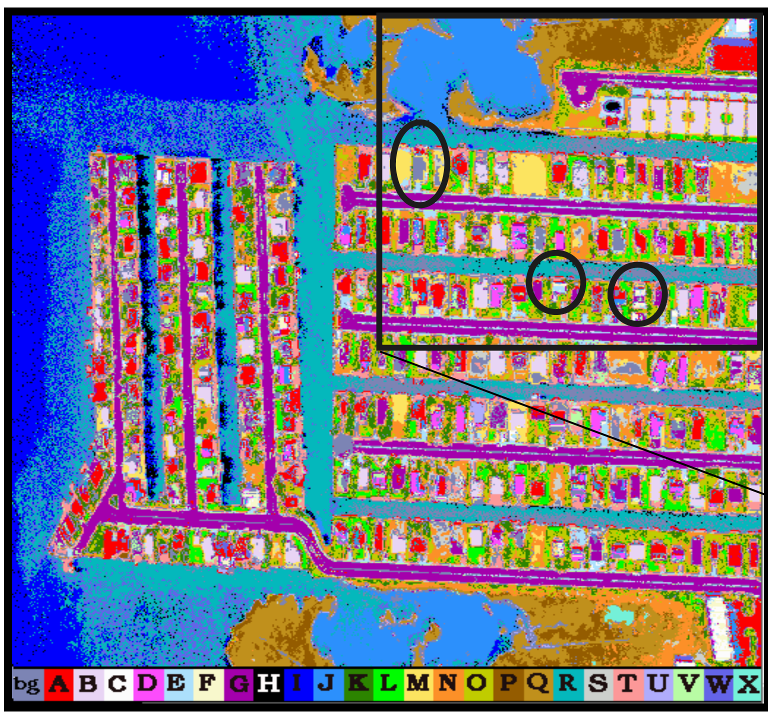
We assume now that the Conscience algorithm achieves a magnification of $\alpha_{achieved} = 1$.

We compare a BDH SOM with $\alpha_{desired} < 0$ to a Conscience SOM of the same data, to see if known small clusters have larger areal representation in the BDH SOM.

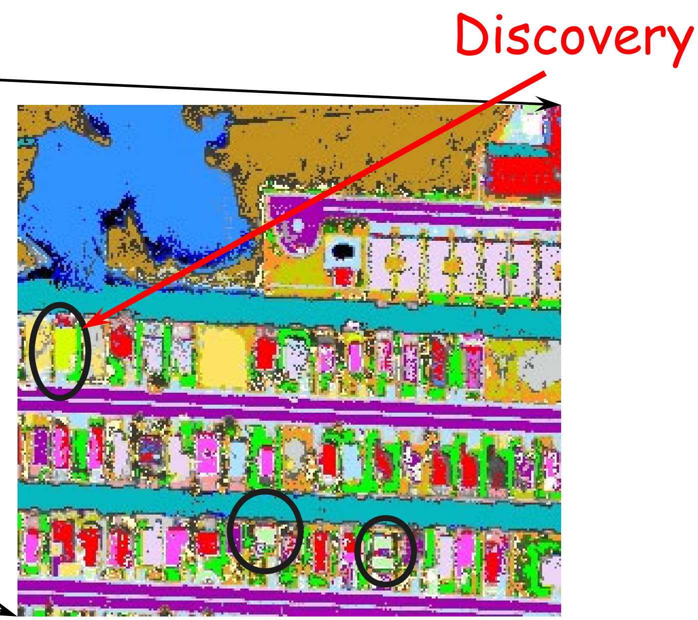
We also use a verified supervised class map to see if either Conscience or BDH SOM shows new discovery.

$\alpha < 0$ magnification for 8-D real data: discovery of rare clusters

Data: 8-D spectral image of Ocean City, Maryland.
512 x 512 pixels, very noisy



Supervised classification, 24 verified classes



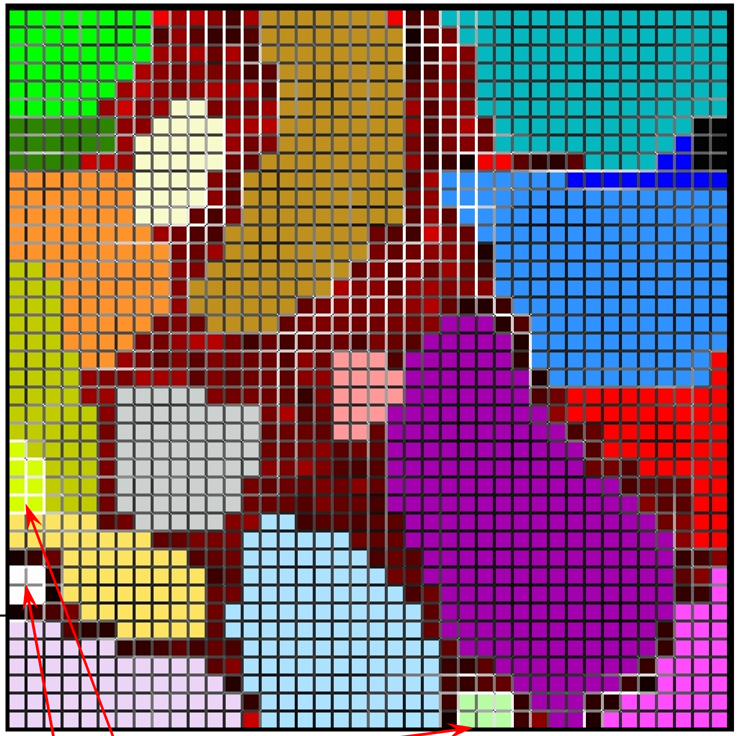
BDH clustering, with forced negative magnification,
$\alpha_{desired} = -0.8$

(Merényi et al. IEEE TNN 2007)

Comparison of BDH and Conscience SOM

Real Data: Ocean City, 8-D 512 x 512 pixel image

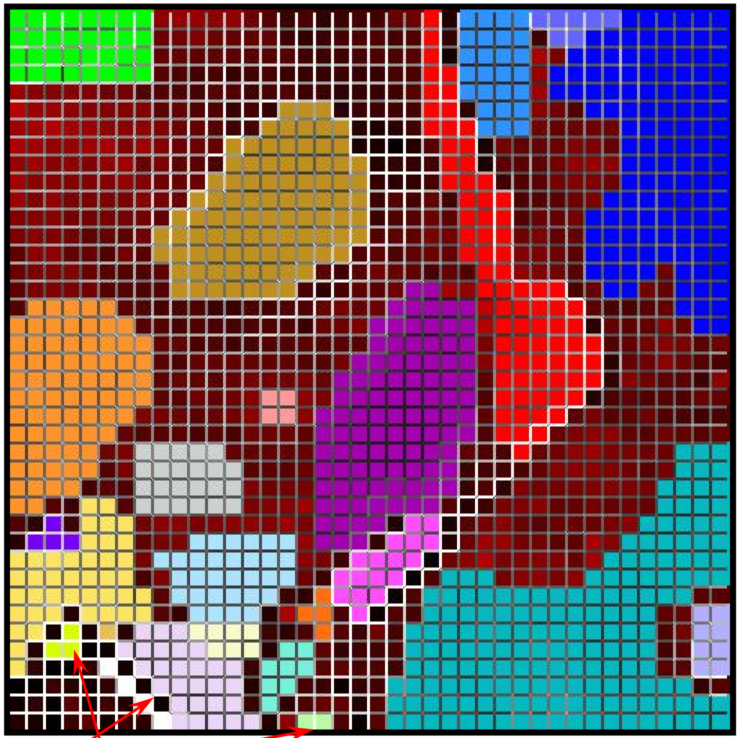
40 x 40 SOM, BDH, α -0.8
(used 128 x 128 pixel subset)



Rare classes

#PEs: 7, 4, 6

40 x 40 SOM, Conscience
(used entire 512 x 512 pixel image)



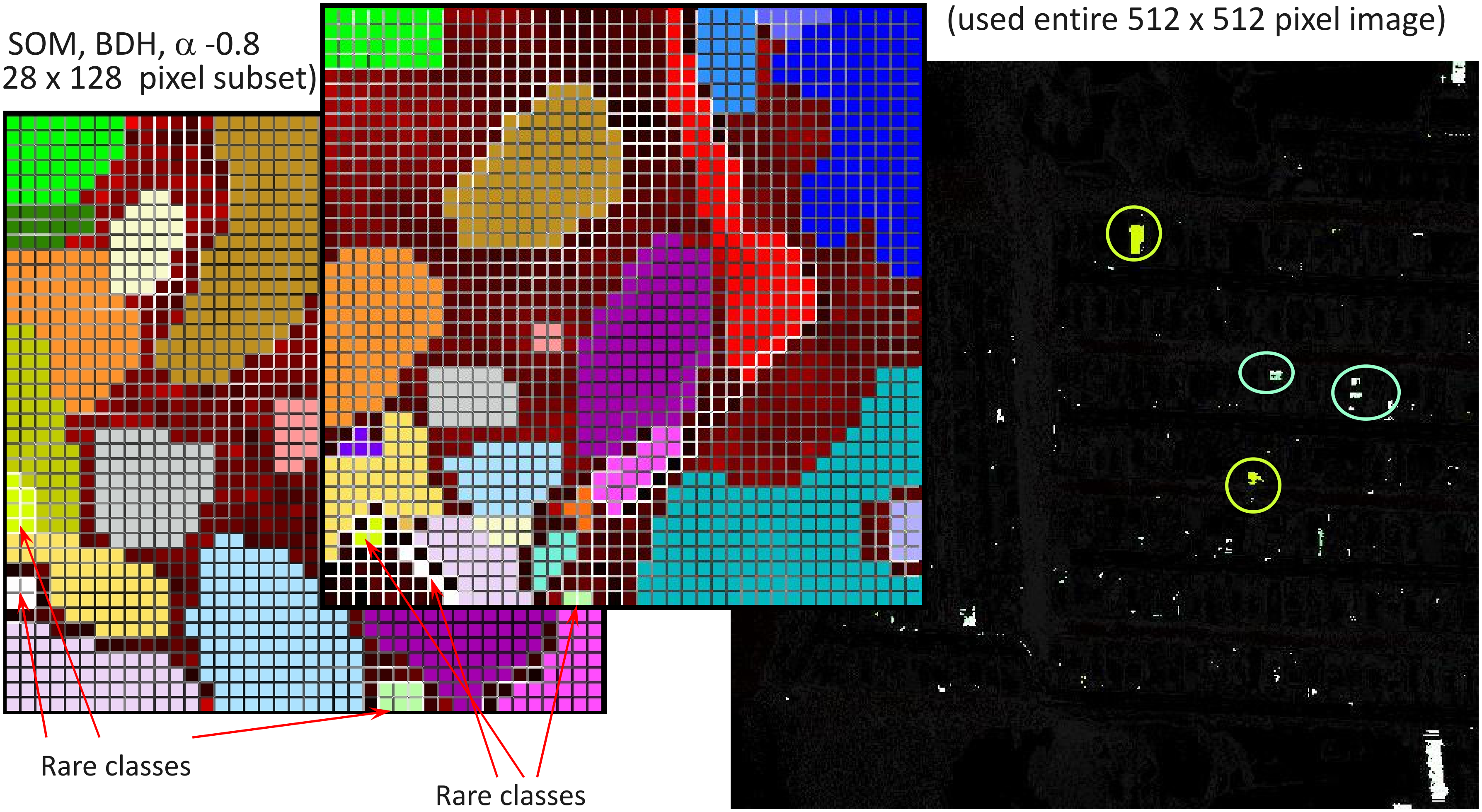
Rare classes

#PEs: 3, 4, 2

E. Merényi, Rice U
erzsebet@rice.edu

Rare clusters detected by Conscience SOM

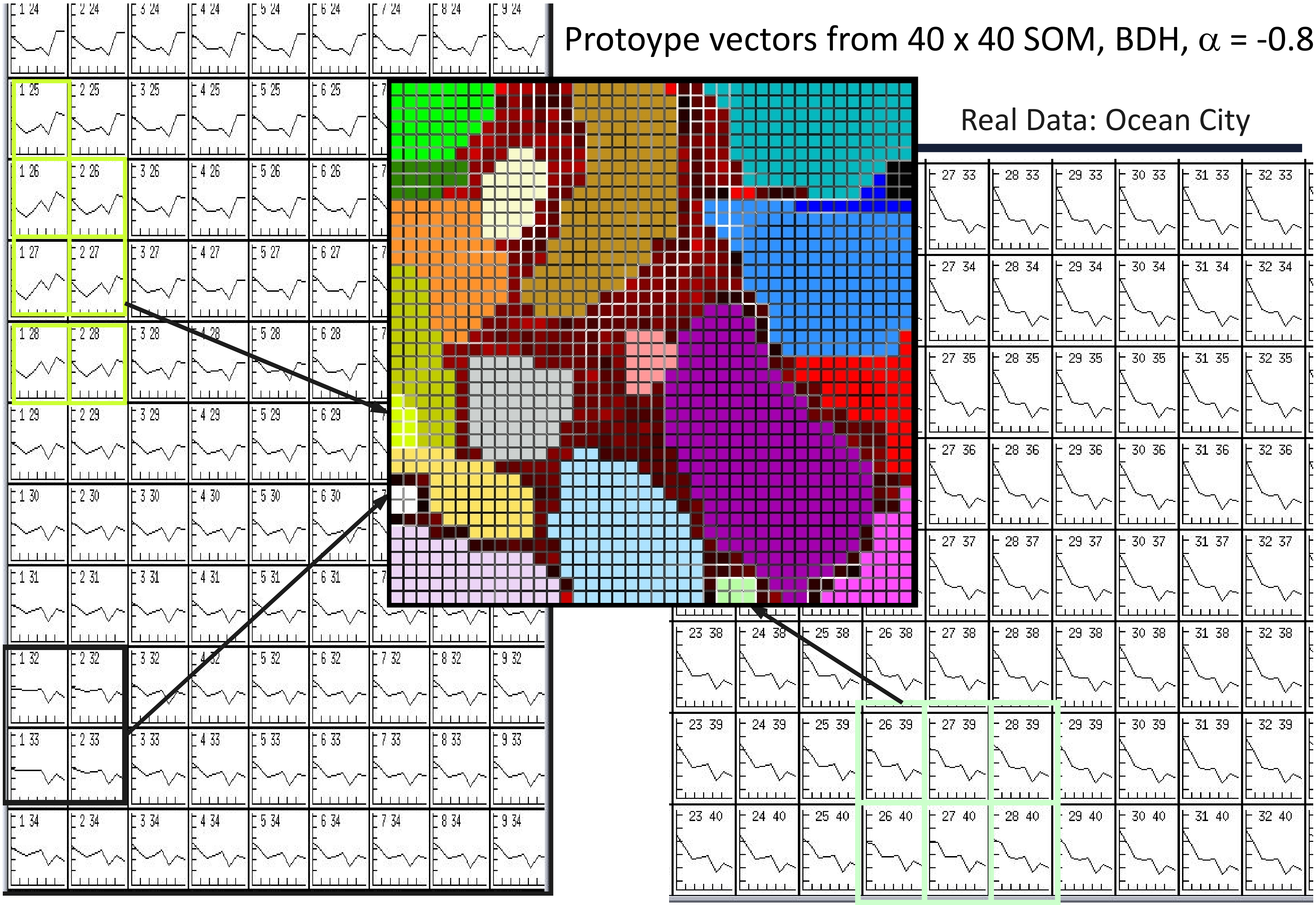
Real Data: Ocean City



Distribution of rare patterns in the 512 x 512 image

(Merényi et al. IEEE TNN 2007)

Protoype vectors from 40 x 40 SOM, BDH, α = -0.8

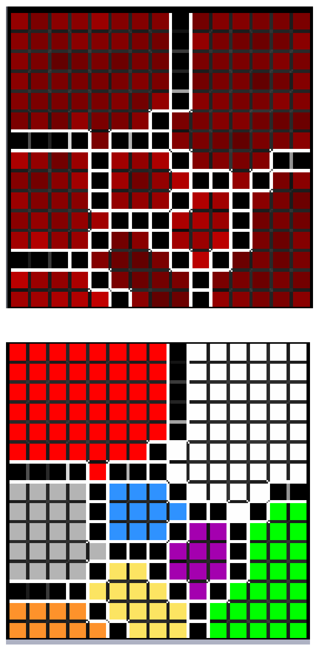


E. Merényi, Rice U
erzsebet@rice.edu

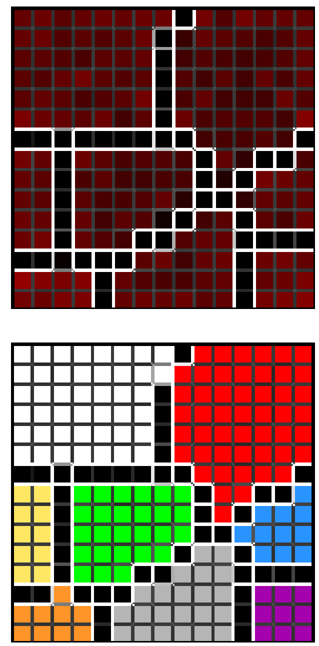
α =1 magnification: special case of max. entropy mapping

6-D synthetic data cube with 8 classes

Conscience SOM
$\alpha_{achieved}$ = 1?



BDH with $\alpha_{desired}$ = 0.7
to get $\alpha_{achieved}$ = 1



PEs allocated:

Conscience SOM:
A:48 B:49
C:25 O:21
D:13 H:9 I:10 M:9

BDH:
A:49 B:44
C:26 O:22
D:10 H:11 I:11 M:9

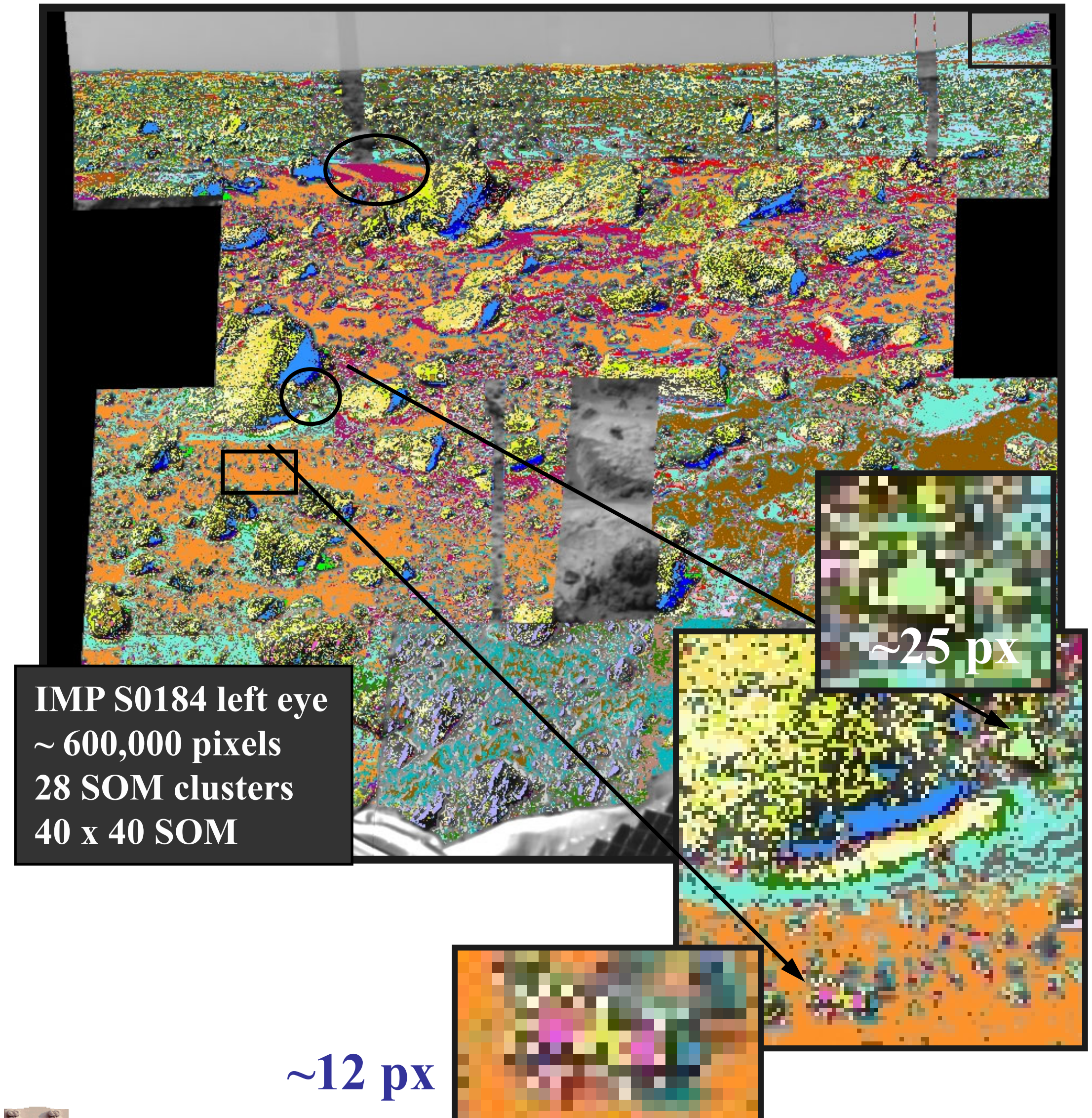
Class A: 4096 points
Class B: 4096 points
Class C: 2048 points
Class O: 2048 points
Class D: 1024 points
Class H: 1024 points
Class I: 1024 points
Class M: 1024 points

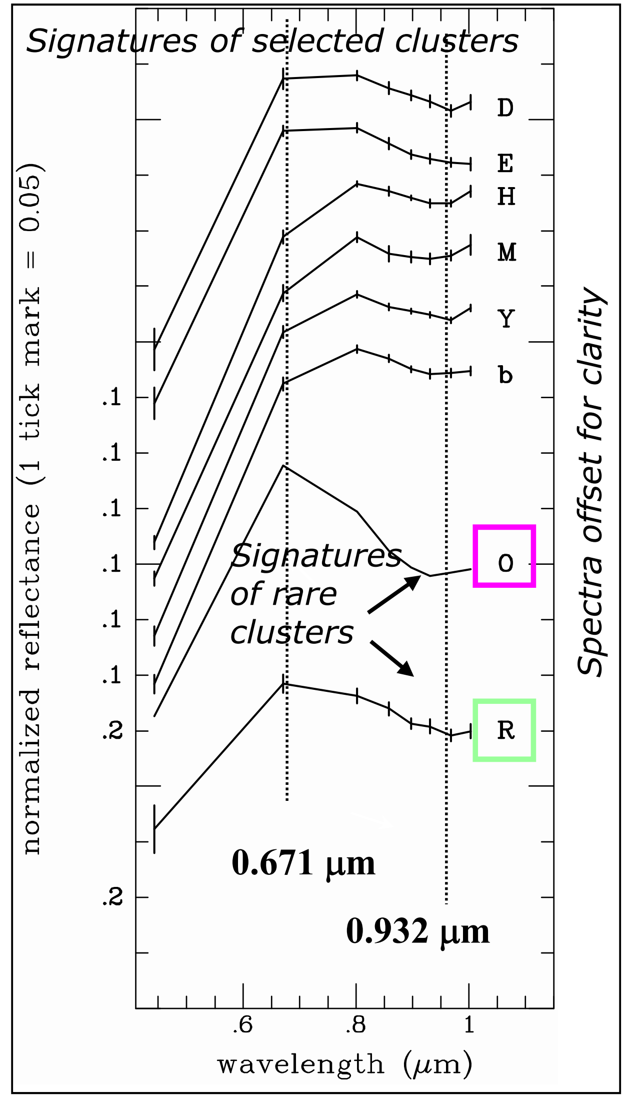
Deviations from the exact 4:2:1 proportions can be due to the small size of the SOM, integer arithmetic, and the formation of inter-cluster gaps

E. Merényi, Rice U
erzsebet@rice.edu

Finding Clusters of Rare Materials on Mars

Data: VIS-NIR Spectral Imagery, Imager for Mars Pathfinder; Colors: clusters



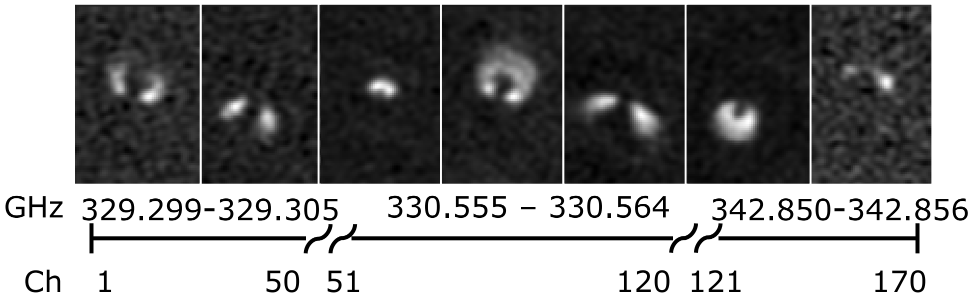


Farrand et al. Int'l Mars J. 2008

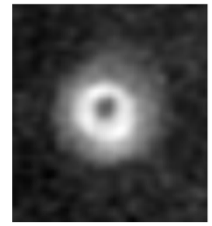
E. Merényi, Rice U
erzsebet@rice.edu

Example: ALMA hyperspectral image – spectral variations

Image planes from ALMA Band 7, protoplanetary disk HD 142527

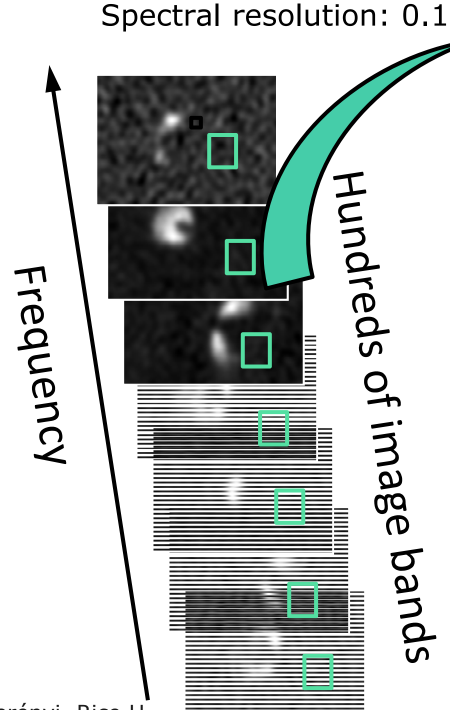


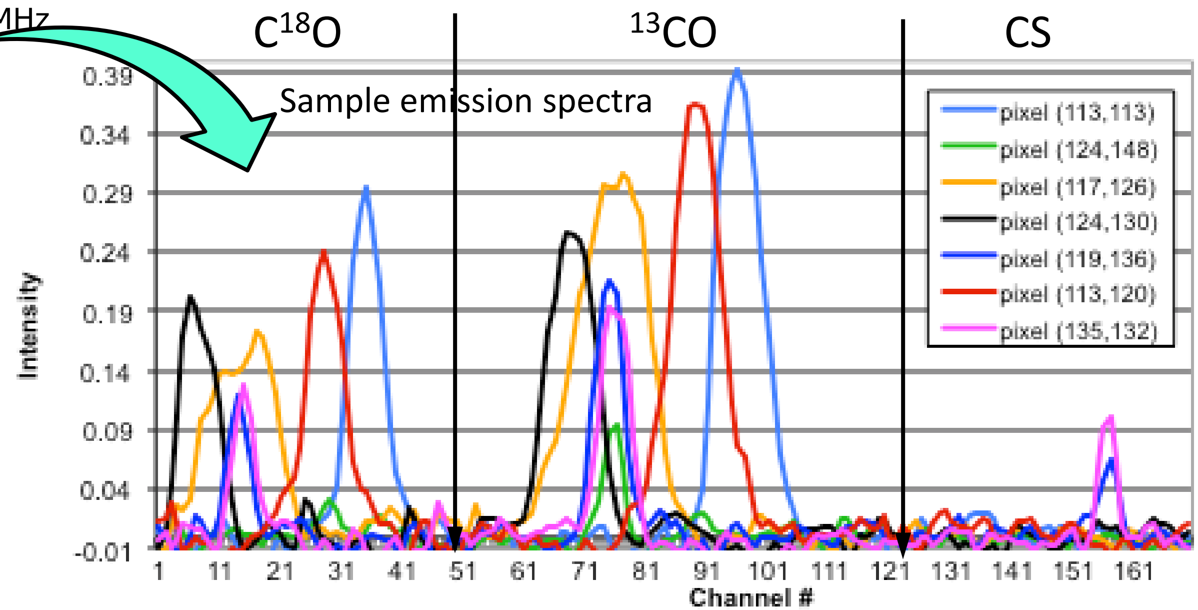
170 channels: $C^{18}O$, ^{13}CO, CS lines stacked
Spectral resolution: 0.122 MHz



Continuum image

Cluster the spectral signatures to map regions of distinct kinematic and compositional behavior.



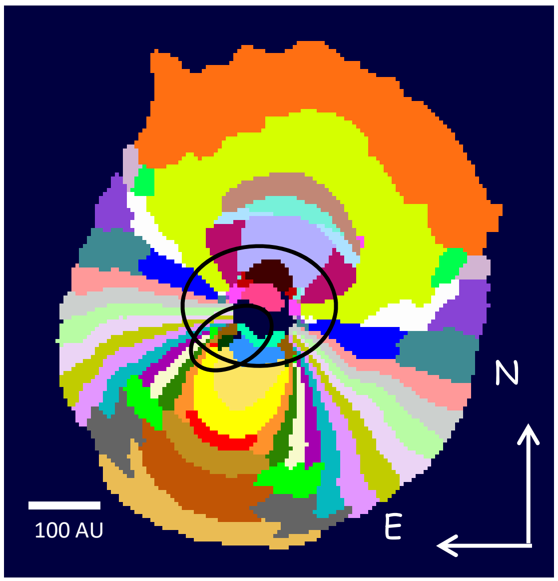


ALMA spectra from combined $C^{18}O$, ^{13}CO, CS lines, showing differences in composition, Doppler shift, temperature

(Data credit: JVO, project 2011.0.00318.5)

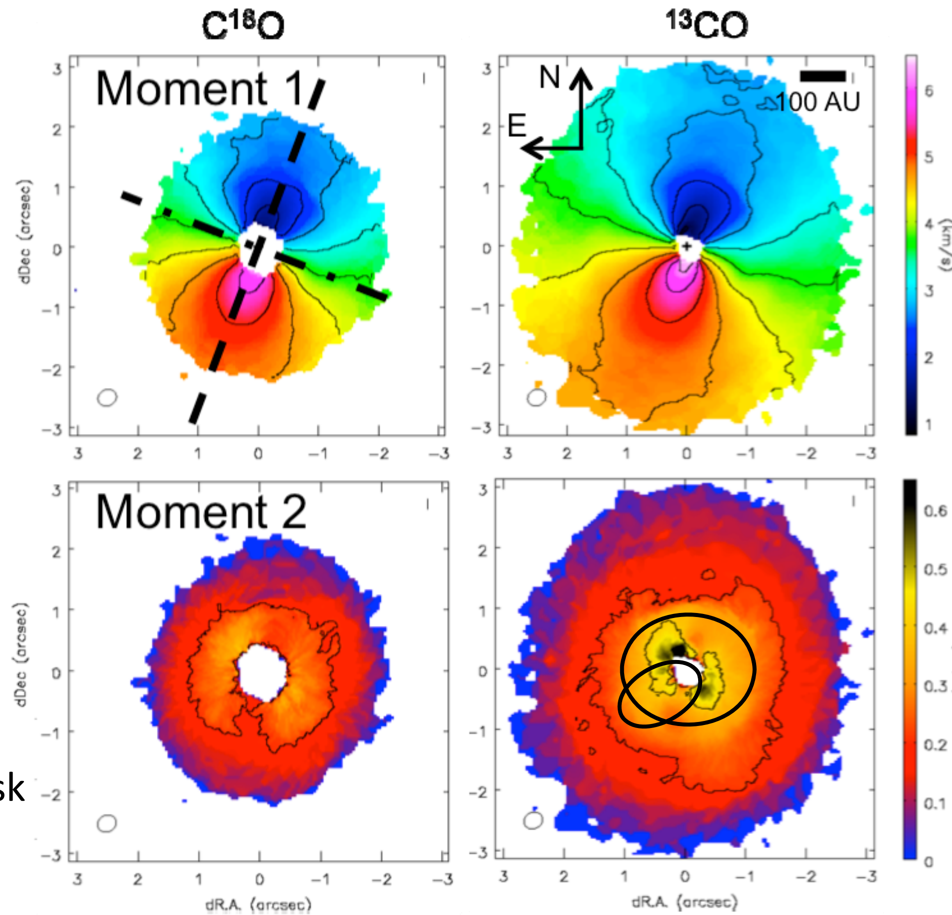
NeuroScope structure discovery from ALMA data
HD 142527 protoplanetary disk (data: Isella 2015)

NeuroScope 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, not a heat map!

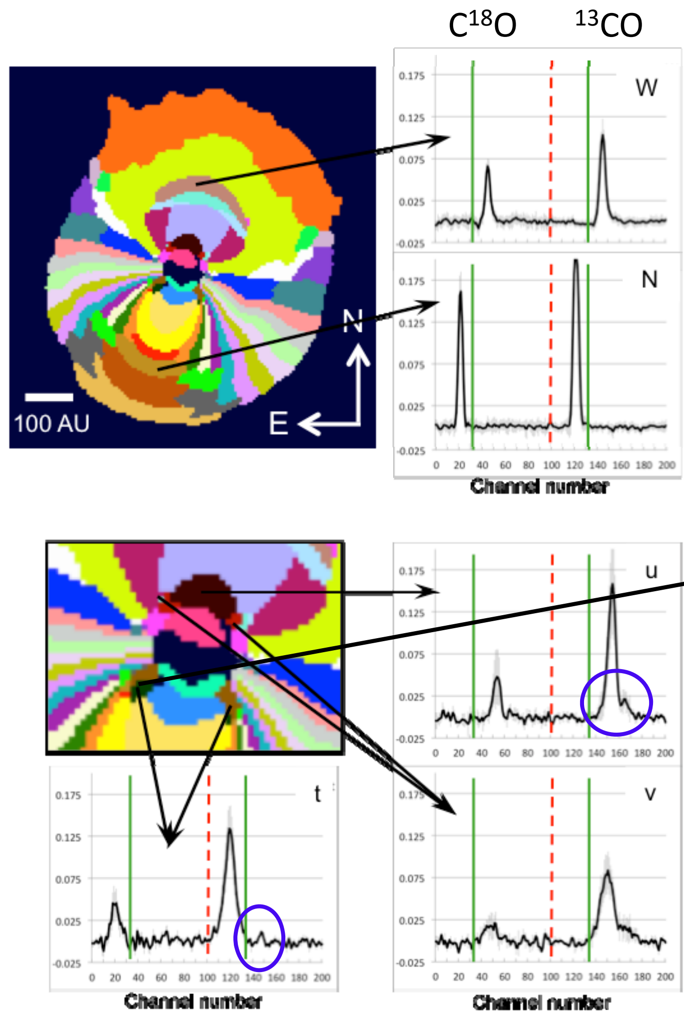


(Merényi, Taylor, Isella, Proc. IAU 325, 2016)

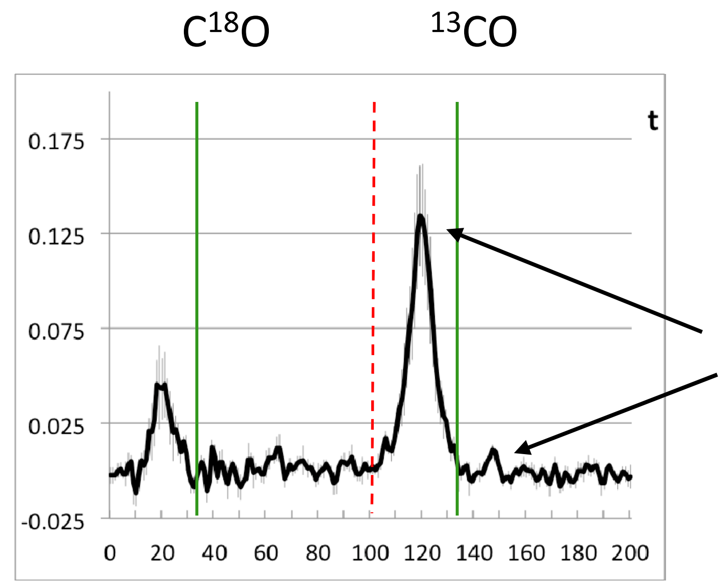
E. Merényi, Rice U
erzsebet@rice.edu

Clusters found in HD142527

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



Mean cluster signatures alert to interesting areas.



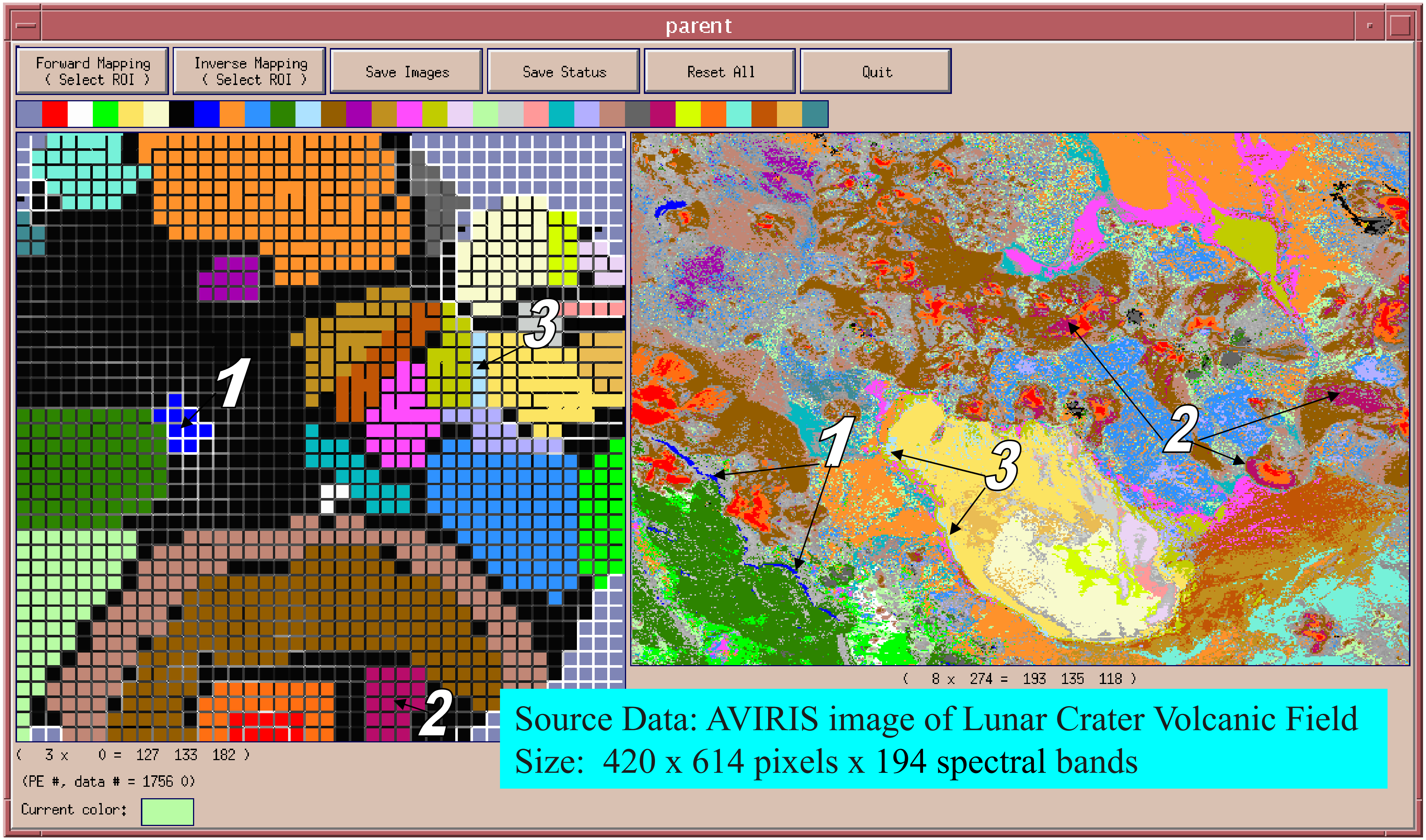
Two distinct peaks, shifted opposite from rest frequency. Two gas components moving in different directions.

(Merényi, Taylor, Isella, Proc. IAU 325, 2016)

More discovery within one molecular line

More discovery from the combination of lines

Discovery in large 194-D hyperspectral image with CSOM



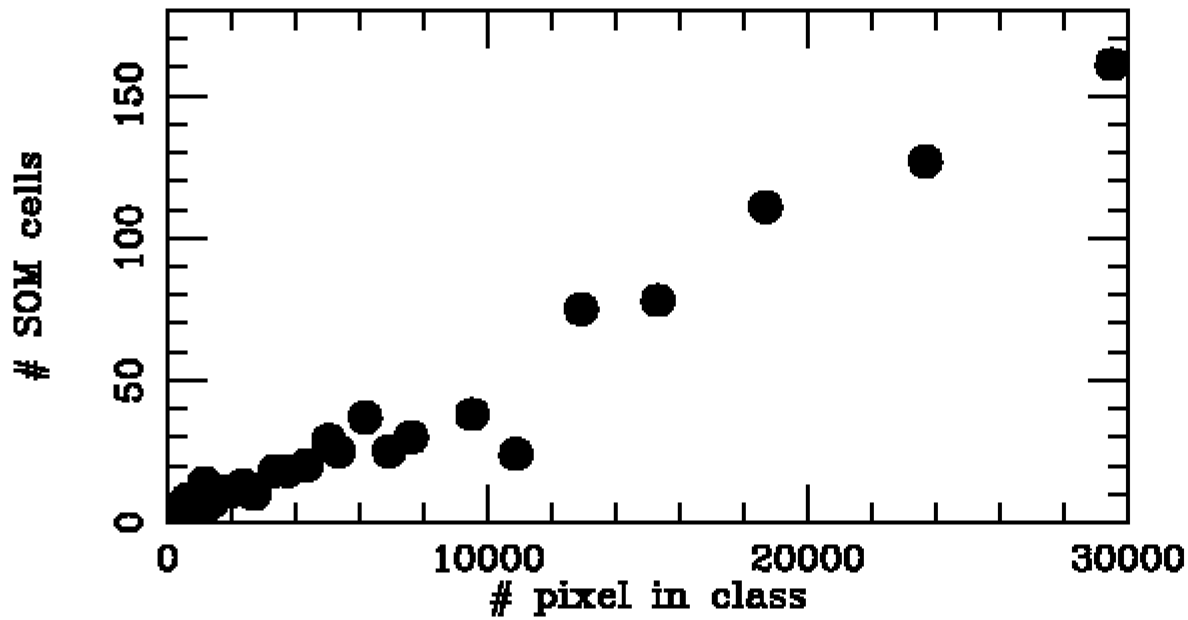
Source Data: AVIRIS image of Lunar Crater Volcanic Field
Size: 420 x 614 pixels x 194 spectral bands

Left: Clusters identified by a Conscience SOM. **Right:** Clusters shown in the spatial image.

(Merényi, 2000; Villmann and Merényi, 2001)

Density matching (max. entropy mapping) by Conscience SOM, 194-band hyperspectral data

Data: Lunar Crater Volcanic Field, 194-band AVIRIS image, segmented into 32 SOM clusters



SOM cells allocated to clusters is proportional to the # if pixels in the clusters.

(Merényi, ISCI 2000)

In Summary

- Predictability of the magnification exponent for “forbidden” data: $\alpha_{achieved}$ = 1 verified
- Negative magnification for “forbidden” data magnifies the rare classes in the BDH SOM
- Applicability of BDH may be justified for a broader range of data than the theory supports
- We used SOM magnification for rare clusters in data with
  - ~ 6 - 200-D feature vectors, some very noisy
  - ~ 2.5 - 6*10^5 patterns, some with subtle differences

  Promise for DM search?
- Behavior of BDH is worth (and needs!) more investigation to assess applicability for complex, high-D data with extremely rare clusters.

Note on mass-processing perspectives for pipelines

(Example numbers for the 6-D synthetic and 200-D hyperspectral image)

- Do SOM learning in parallel hardware : < 5 - 15 sec / 1M
  - Practically automatic
  - Dedicated mid-level FPGA implementation, could be much faster for more $$ *(Lachmair et al., Neurocomputing 2013)*
  - SOM size matters
- Cluster the SOM prototypes automatically with SOM-derived CONN graph as input to graph-segmentation algorithms: < 1 sec
  - Results comparable to interactive segmentation by expert. *(Merényi and Taylor, WSOM+ 2017)*
- Scales linearly with # of samples, and (within large range) with # of feature dimensions

References

- Bauer, H-U., Der, R., Herrmann, M. (1996) Controlling the Magnification of Self-Organizing Feature Maps. *Neural Computation* 8:4, pp 757-771.
- Zador, P. L. (1982) Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inf. Theory*, vol., IT-28, no. 2, pp. 139-149.
- Ritter, H. and Schulten, K. (1986) On the Stationary State of Kohonen's Self-Organizing Sensory Mapping. *Biol. Cybern*. 54, 99-106.
- DeSieno, D. (1988) Adding a Conscience to Competitive Learning, *Proc. Int. Conf. Neural Netw.* Vol I, pp I-117-I-127.
- Taşdemir, K., and Merényi, E. (2011) A Validity Index for Prototype Based Clustering of Data Sets with Complex Structures. *IEEE Trans. Sys. Man and Cyb., Part B.* 02/2011; Vol. 41, No. 4, pp 1039 - 1053. DOI: 10.1109/TSMCB.2010.2104319
- 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.
- Merényi, E., Taylor, J. and Isella, A. (2016), Deep data: discovery and visualization. Application to hyperspectral ALMA imagery. *Proc. Int'l Astronomical Union, 12*(S325), 281-290. doi:10.1017/S1743921317000175
- Farrand, W. H., Merényi. E.., Johnson, J., Bell, J. III (2008) Comprehensive mapping of spectral classes in the imager for Mars Pathfinder Super Pan, *The Int'l J. of Mars Science and Exploration*, Mars 4, 33-55, 2008; doi:10.1555/mars.2008.0004
- Lachmair, J., Merényi, E., Porrmann, M., Rückert, U. (2013) A Reconfigurable Neuroprocessor for Self-Organizing Feature Maps. *Neurocomputing* 112, pp 189-199.
- Merényi, E., Taylor, J. (2017) SOM-empowered Graph Segmentation for Fast Automatic Clustering of Large and Complex Data. *Proc. 12th WSOM+ 2017, Nancy, France, June 27-29, 2017. 9pp.* On-line: http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8019995