Adaptive Resonance Theory and Diffusion Maps for Clustering Applications

Donald C. Wunsch, Steven Damelin and Rui Xu
Applied Computational Intelligence Laboratory
Missouri University of Science and Technology

Mathematical Reviews & Univ. Michigan

Acknowledgements

Mary K Finley Endowment
Intelligent Systems Center
Center for Infrastructure Science & Engineering

This talk is in honor of our prescient friend Alexander Gorban

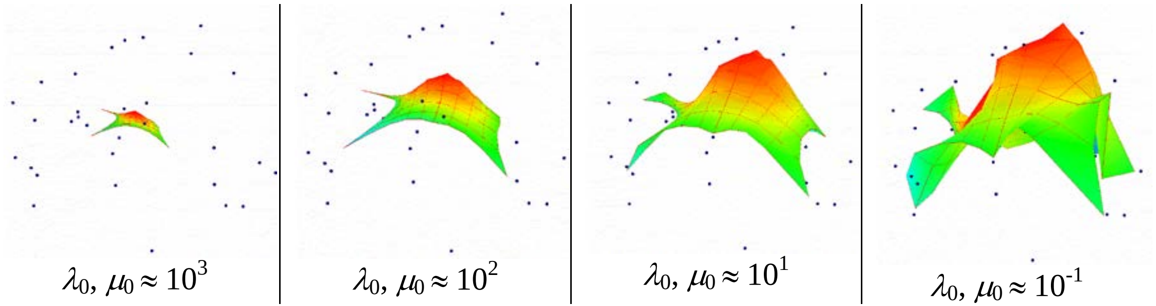


Fig.4: Training elastic net in several epochs

D.C. Wunsch, “ART properties of interest in engineering applications,” in *Proc. IEEE / INNS International Joint Conference on Neural Networks*, Atlanta, GA, 2009.

A. Gorban, B. Kégl, D.C. Wunsch, and A. Zinovyev, Eds., *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer, 2007.

J. Sieffertt and D.C. Wunsch, *Unified Computational Intelligence for Complex Systems: Studies in Neural, Economic and Social Dynamics*. Springer-Verlag, 2010.

R. Xu and D.C. Wunsch II, *Clustering*. IEEE Press / Wiley, 2009.

R. Xu, J. Xu, and D.C. Wunsch, “A Comparison Study of Validity Indices on Swarm Intelligence-Based Clustering,” *IEEE Trans. on Systems, Man and Cybernetics, part B*, Vol. 42, No. 4, pp. 1243 – 1256, 2012.

R. Xu and D.C. Wunsch, “BARTMAP: A viable structure for biclustering,” *Neural Networks*, vol. 24, no. 7, pp. 709-716, 2011.

L. du Plessis, R. Xu, S. Damelin, M. Sears, and D.C. Wunsch II, “Reducing dimensionality of hyperspectral data with diffusion maps and clustering with K-means and fuzzy ART,” *International Journal of Systems, Control and Communications*, Vol. 3, No. 3, pp. 232-251, 2011.

R. Xu and D.C. Wunsch, “Clustering algorithms in biomedical research: A review,” *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 120–154, 2010.

R. Xu, S. Damelin, B. Nadler, and D.C. Wunsch II, “Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps,” *Artificial Intelligence in Medicine*, vol. 48, no. 2-3, pp. 91-98, 2010.

R. Xu, L. du Plessis, S. Damelin, M. Sears, and D.C. Wunsch, “Analysis of hyperspectral data with diffusion maps and fuzzy ART,” in *Proc. IEEE / INNS International Joint Conference on Neural Networks*, Atlanta, GA, 2009.

R. Xu, S. Damelin, and D.C. Wunsch II, “Clustering of cancer tissues using diffusion maps and fuzzy ART with gene expression data,” in *Proc. IEEE / INNS International Joint Conference on Neural Networks*, Hong Kong, China, June 2008, pp. 183-188.

R. Xu, S. Damelin, B. Nadler, and D.C. Wunsch II, “Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps,” in *Proc. of the IEEE International Conference on Biomedical Engineering and Informatics*, Sanya, China, May 2008.

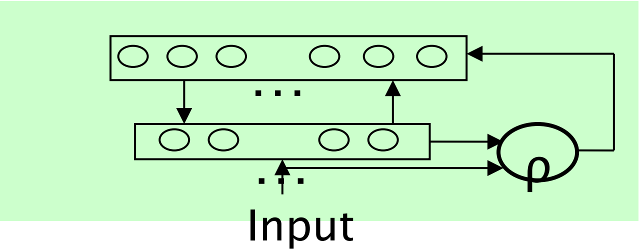
R. Xu, S. Damelin, and D.C. Wunsch, “Applications of diffusion maps in gene expression data-based cancer diagnosis analysis,” in *Proc. Of IEEE 29^{th} Annual Engineering in Medicine and Biology Society International Conference*, Aug. 22-26, 2007, pp. 4613–4616.

Charles Fefferman, Steven. B. Damelin and William Glover, BMO Theorems for ε distorted diffeomorphisms on R^D and an application to comparing manifolds of speech and sound, Involve, a Journal of Mathematics 5-2 (2012), pp 159—172.

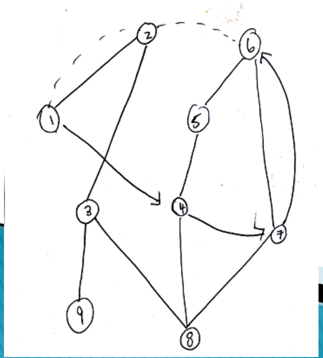
L du Plessis, S.B Damelin and M. Sears,“ Reducing the dimensionality of hyperpectral data using diffusion maps”, *Proceedings of the 2009 IEEE Geosciences and Remote Sensing Symposium*, Cape Town, pp 105-132.

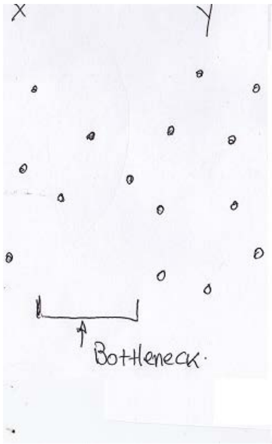
R.R. Coifman, S. Lafon, “Diffusion maps", Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets, Vol 21, July 2006, pp 5-30.

- Adaptive Resonance Theory: Learning switched on/off by resonant feedback loops in neural circuit



- Diffusion maps: Kernel–based, from edge–weighted graphs to smooth manifolds, we use for dimensionality reduction





Diffusion Maps

- Interpret eigenfunctions of Markov matrices as systems of coordinates on the original data set used in order to obtain efficient representation of data geometric descriptions (Coifman and Lafon, 2006)
- Given a set of d-dimensional data points, $\mathbf{x}_1,\ldots,\mathbf{x}_N$,
  - Construct affinity matrix **W** based on the Gaussian Kernel

$$w(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right)$$

  - Calculate the degree of $\mathbf{x}_i$,

$$d(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathbf{X}} w(\mathbf{x}_i,\mathbf{x}_j)$$

  - Derive the Markov or transition matrix $\mathbf{P} = \{p(\mathbf{x}_i, \mathbf{x}_j)\}$,

$$p(\mathbf{x}_i,\mathbf{x}_j) = \frac{w(\mathbf{x}_i,\mathbf{x}_j)}{d(\mathbf{x}_i)}$$

Diffusion Maps

- Given a set of d-dimensional data points, $\mathbf{x}_1, \ldots, \mathbf{x}_N$,
  - Obtain eigenvalues and eigenvectors of P,

$$\mathbf{P}^t \boldsymbol{\varphi}_j = \lambda^t \boldsymbol{\varphi}_j$$

    - Where larger *t* means fewer clusters
  - Map data objects to the new L-dimensional (L « d) Euclidean space by using the eigenvectors as a new set of coordinates on the data set,

$$\boldsymbol{\Psi}_t : \mathbf{x}_i \rightarrow \left(\lambda_1^t \boldsymbol{\varphi}_1(\mathbf{x}_i), \ldots, \lambda_L^t \boldsymbol{\varphi}_L(\mathbf{x}_i) \right)^T$$

  - Calculate the diffusion distance

$$D_t(\mathbf{x}_i, \mathbf{x}_j) = \left\| p^t(\mathbf{x}_i, \cdot) - p^t(\mathbf{x}_j, \cdot) \right\|_{1/\varphi j}$$

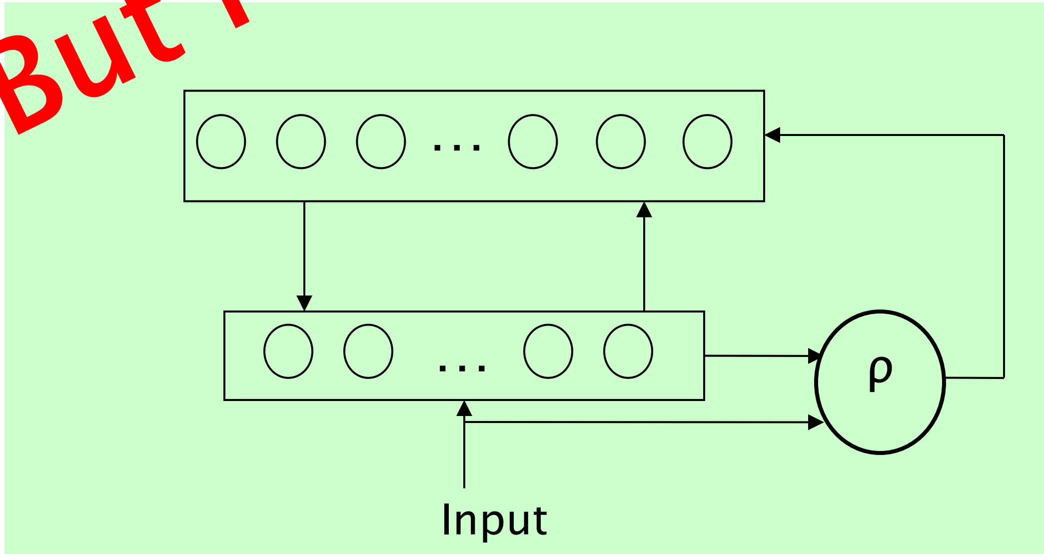
Adaptive Resonance Advantages in Engineering

- Scalability
- Speed
- Configurability
- Parallelization
- Results Interpretation

- New Metrics
- Distributed Representation
- Match-based vs. Error-based

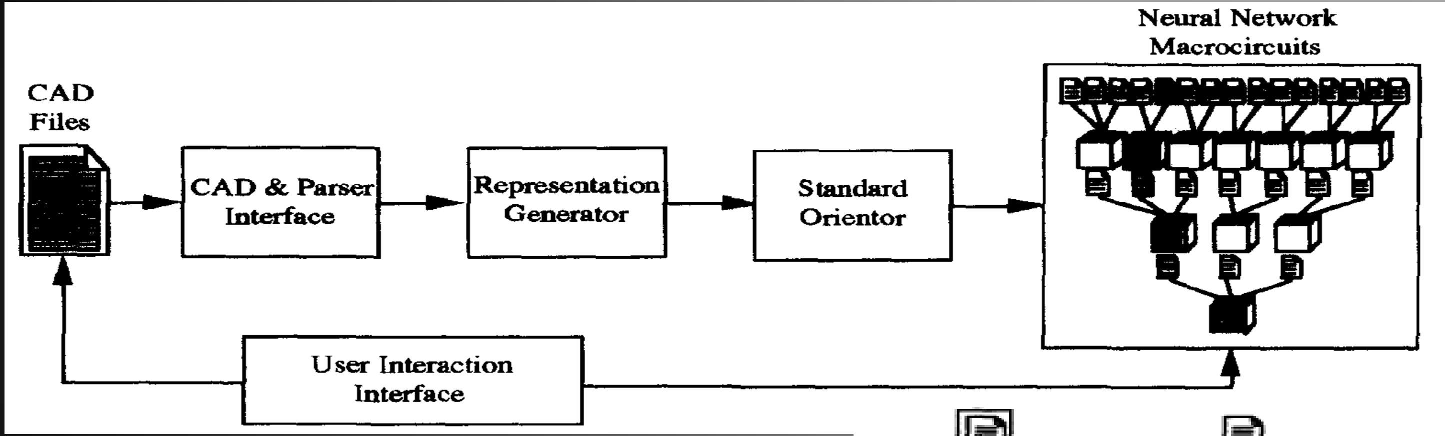
Theory, Not Architecture

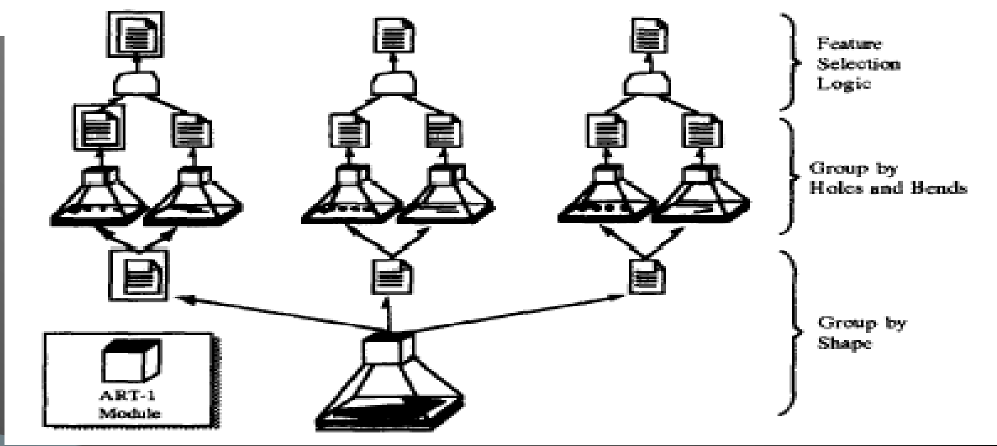
But First...



Resonance mediates learning

Scalability

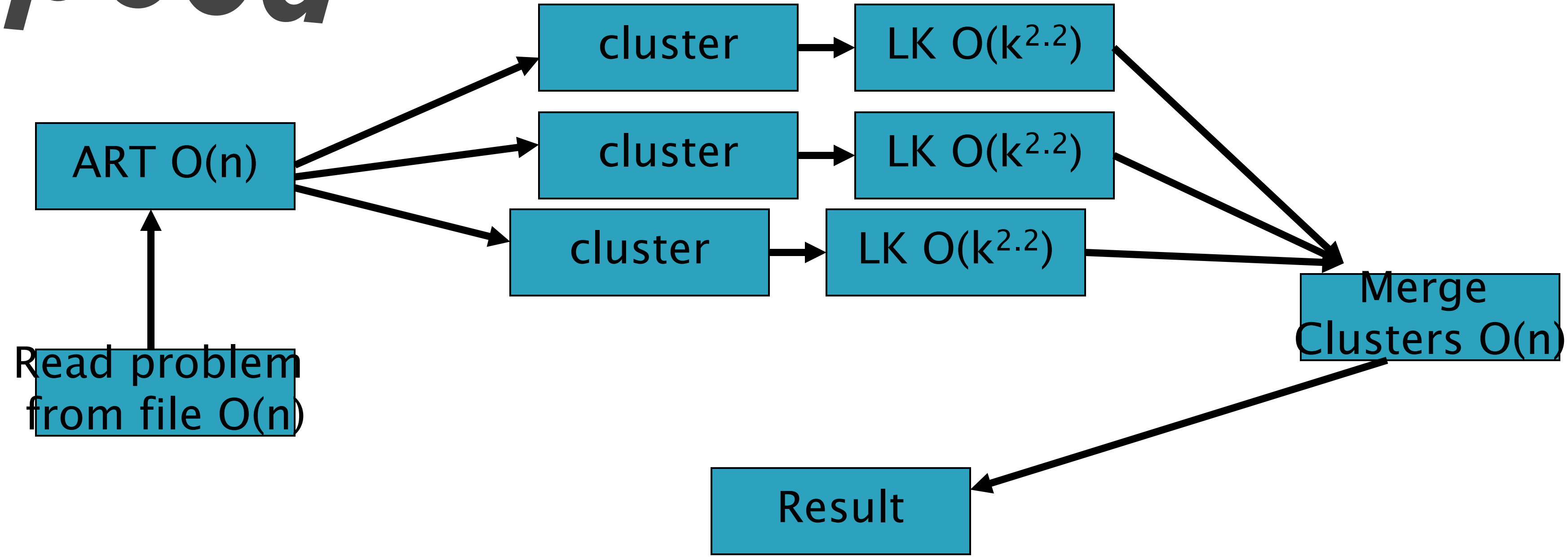




T. Caudell, S. Smith, R. Escobedo, and M. Anderson, “NIRS: Large scale ART-1 neural architectures for engineering design retrieval,” *Neural Networks*, vol. 7, pp. 1339–1350, 1994

Traveling Salesman Problem

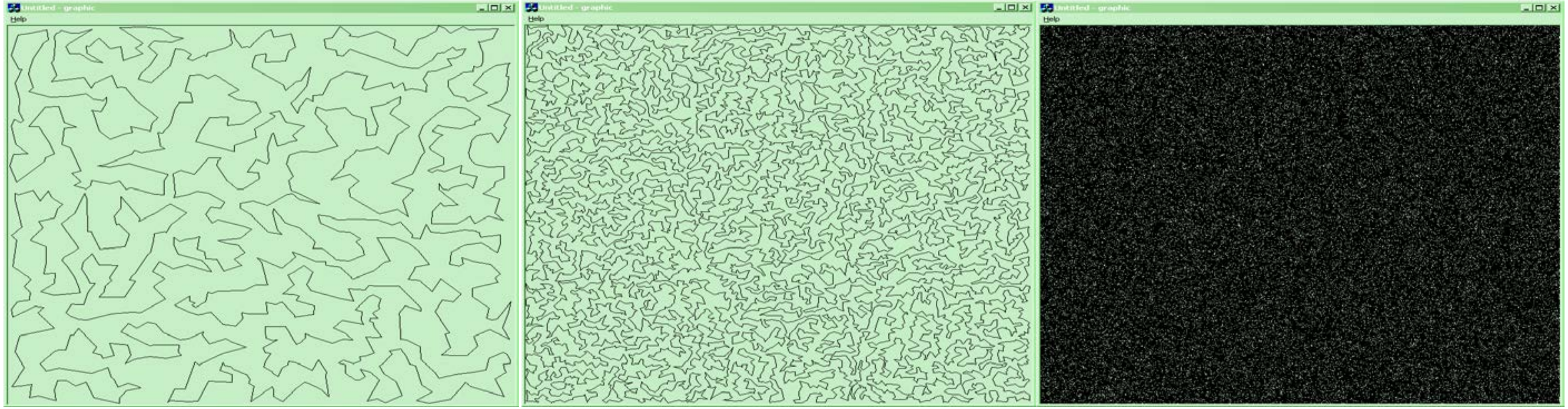
Speed



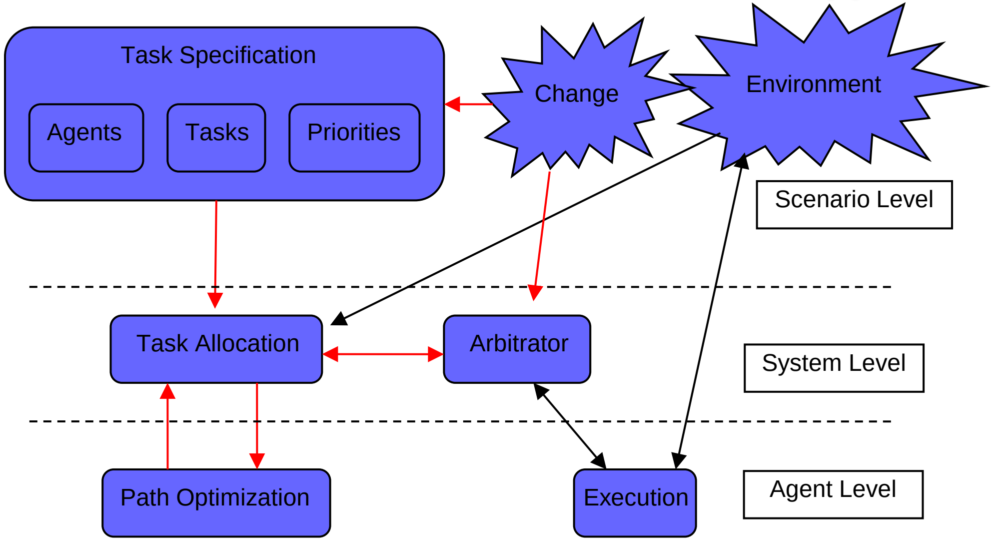
Ref: S. Mulder and D. Wunsch, “Million city traveling salesman problem solution by divide and conquer clustering with adaptive resonance neural networks,” *Neural Networks*, vol. 16, pp. 827–832, 2003.

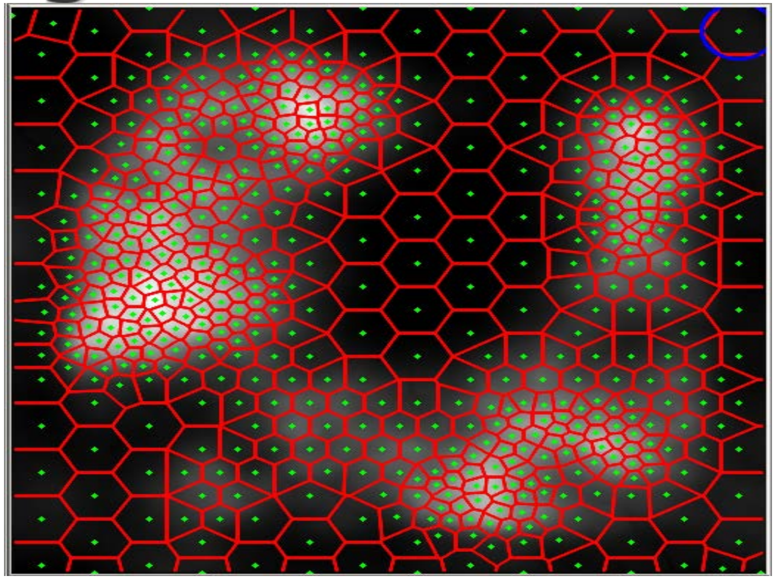
- Divide and Conquer Algorithm
  - #cities Tour Length Time % off
  - 10002.58E+07 0.422 10.40%
  - 20003.61E+07 1.031 10.64%
  - 80007.14E+07 8.328 10.97%
  - 10000 7.97E+07 11.359 10.57%
  - 20000 1.12E+08 24.641 10.53%
  - 250000 4.00E+08 315.078 11.64%
  - 1000000 7.94E+08 1468.165 11.03%
  - 10000000 2.52E+0910528.7 1.27%
- CONCORDE
  - 10002.34E+07 1.670
  - 20003.26E+07 3.500
  - 80006.43E+07 26.570
  - 10000 7.20E+07 37.620
  - 20000 1.01E+08 84.830
  - 250000 3.58E+08 1379.540
  - 1000000 7.15E+08 9013.53
  - 10000000 2.495E+09 43630.7

Plus 25 M city results paper on IEEE Explore

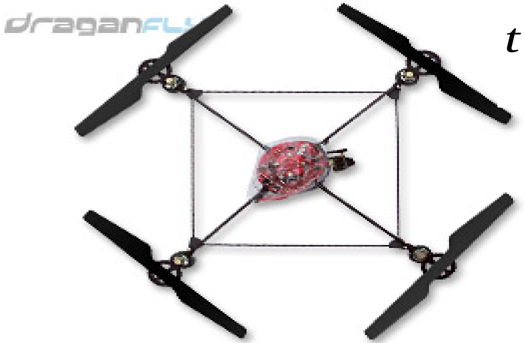

Untitled - graphic
Help
Untitled - graphic
Help
Untitled - graphic
Help

Heterogeneous Vehicle Swarm Path Planning



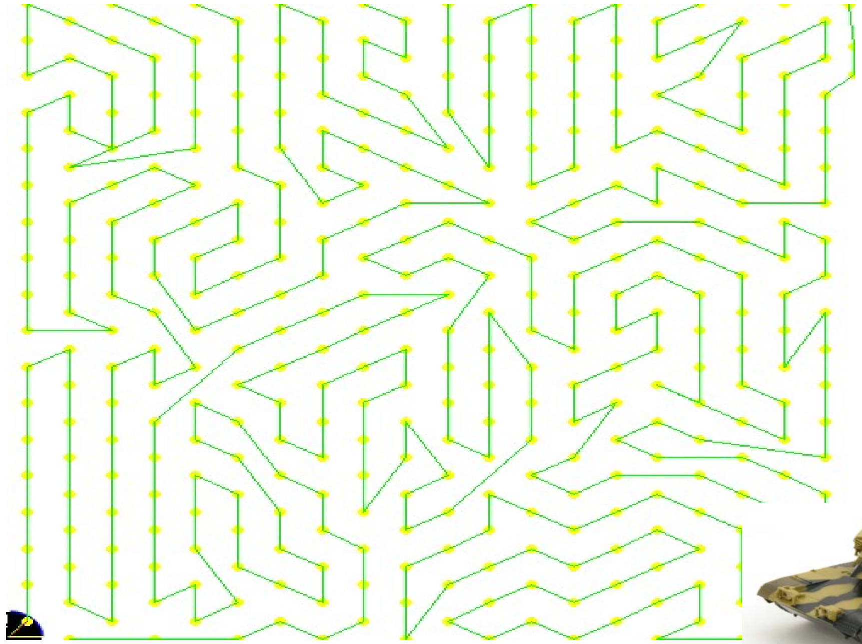


Heterogeneous Vehicle Heuristic Performance



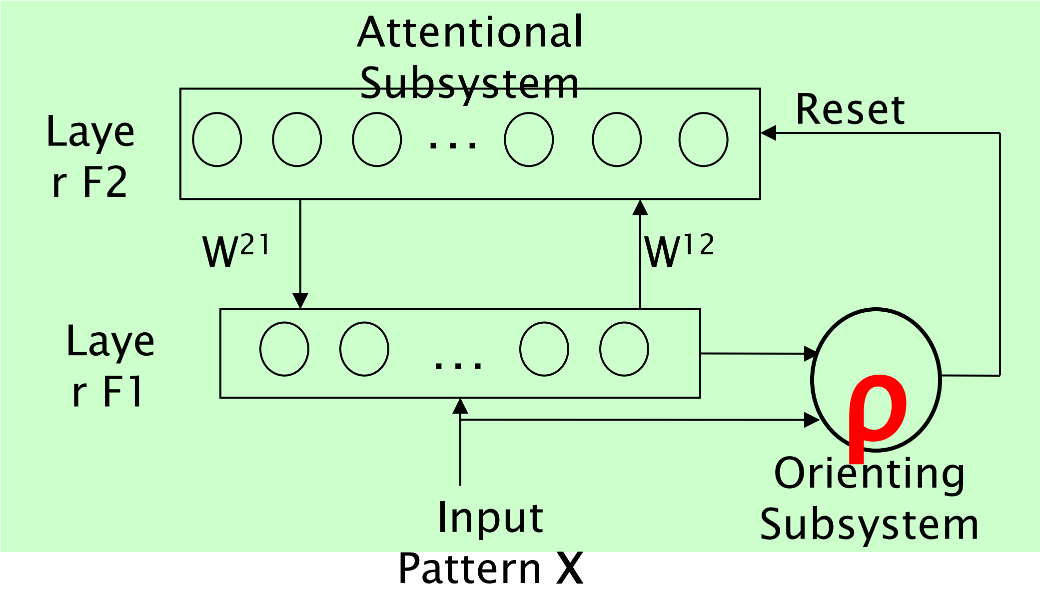
$$t(e_i) = \frac{d(a_{ei}, b_{ei})}{c_g v_{\max}} + c_a \left(\frac{c_g v_{\max}}{a} \right) \left(1 - \cos(\theta_{a_{ei}, b_{ei}, c_{ei}})\right)$$

Non-Euclidean
TSP in real-time



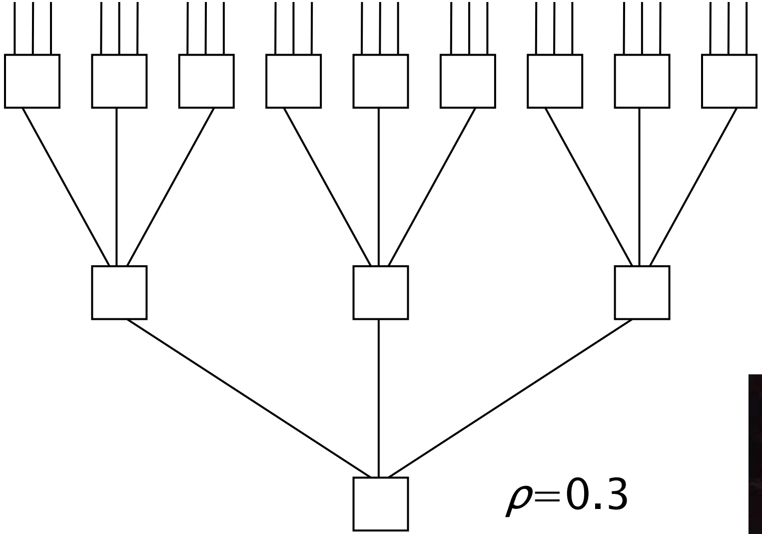
Configurability

- VIGILANCE ρ



- ART1
- ARTMAP
- LAPART
- Fuzzy ART
- Ellipsoid ART
- GramART

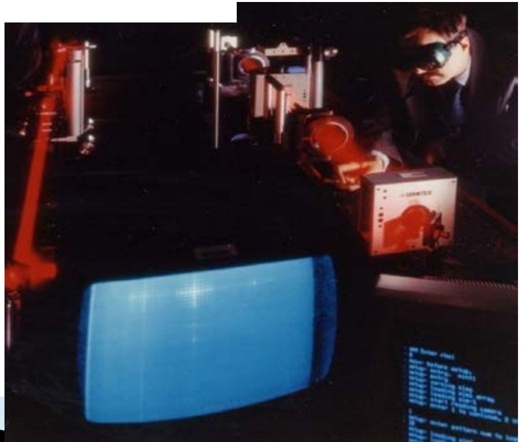
Hierarchical -- Parallelizable



$\rho = 0.9$

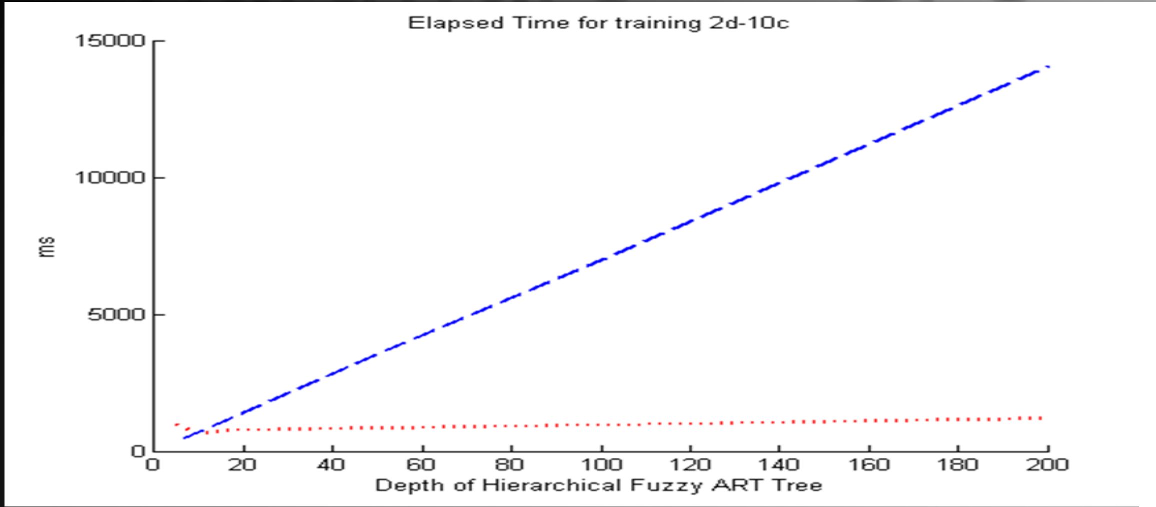
$\rho = 0.6$

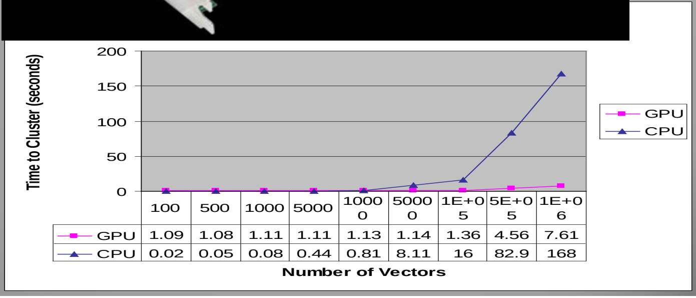
$\rho = 0.3$



#	Operation	Electronics	Optics	%
1†	I		✓	-
2†	T_k		✓	-
3	$T_k \bullet I$		✓	80
4	$\lvert I \rvert$		✓	5
5	$\lvert T_{k_m} \rvert$	✓		5
6	$T_{k_m} = T_{k_m} \cap I$	✓		5
7	$T_{n_c} = I$	✓		1
8††	$\mathrm{maxi}\{\ \}$	✓		3
9	$\{A_{0i}=1;\ i=1, n_t\}$	✓		<1
10	$p_i \rightarrow C_{n_c}$	✓		<1
11	$\geq$	✓		<1
12	$*$	✓		<1
13	$/$	✓		<1
14	$+$	✓		<1

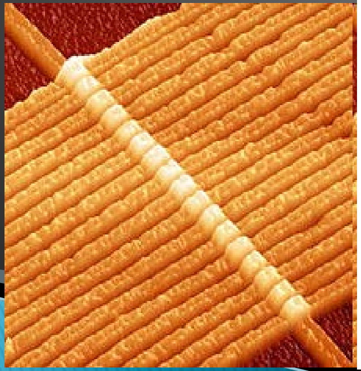
Hardware -- GPU

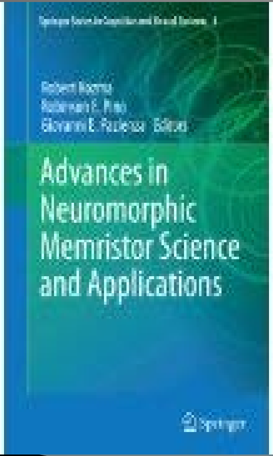


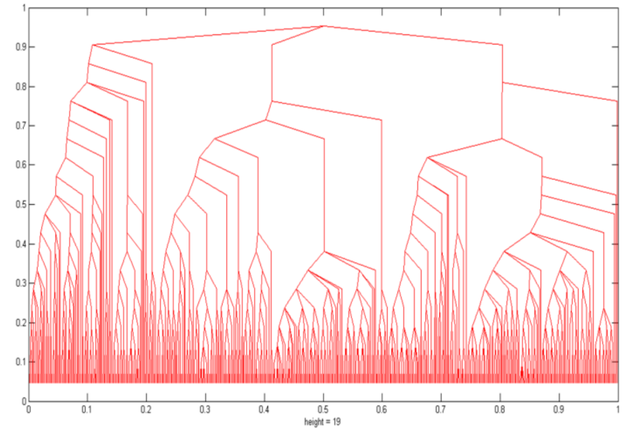


Memristor

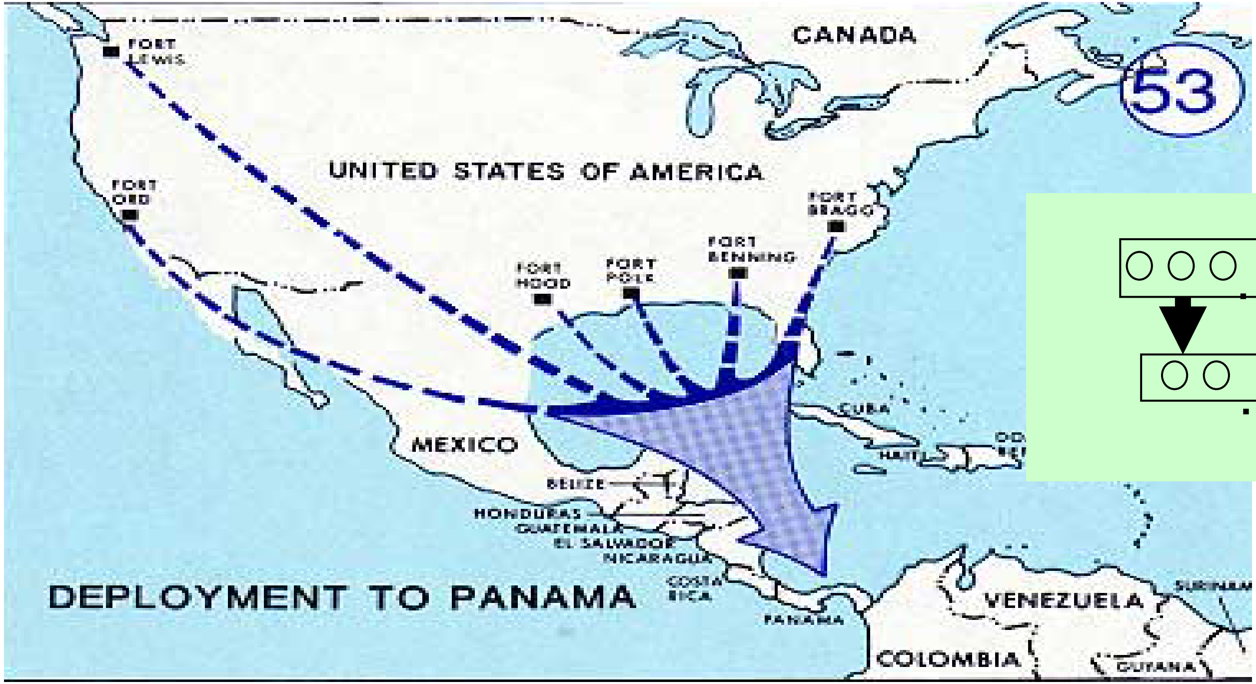
ART

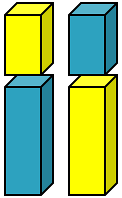


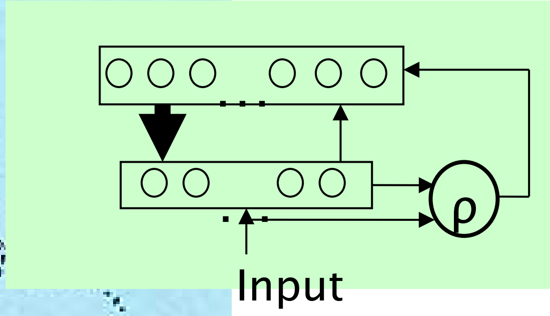




Results Interpretation – ART Templates as Chokepoint Estimators

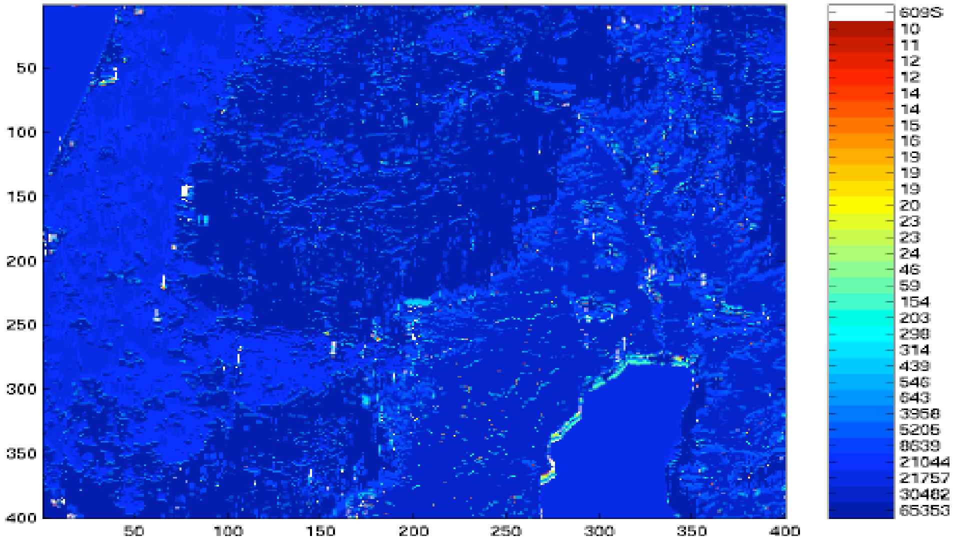






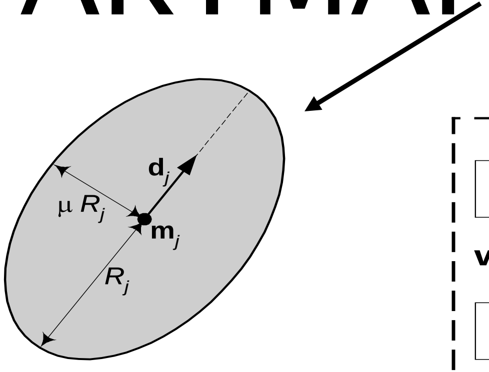
Turner & Wunsch

Knowledge Representation / Template Interpretation via Category Theory

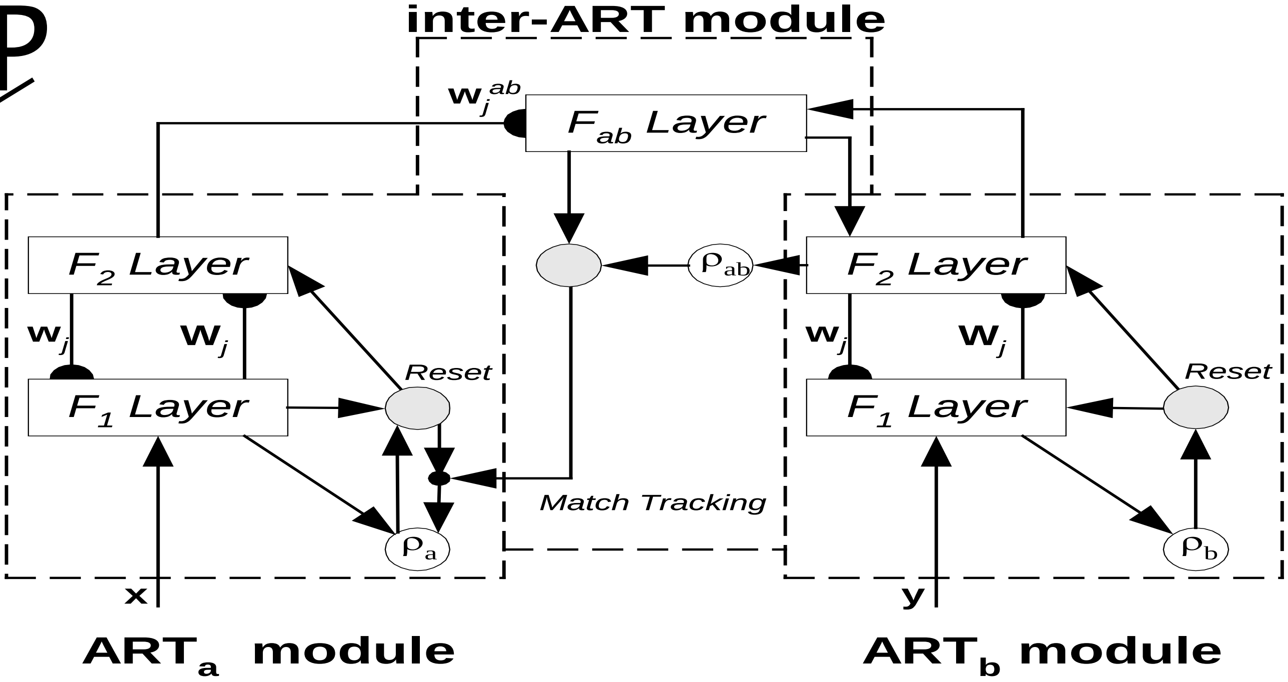


Healy, Olinger, Young, Caudell, Larson

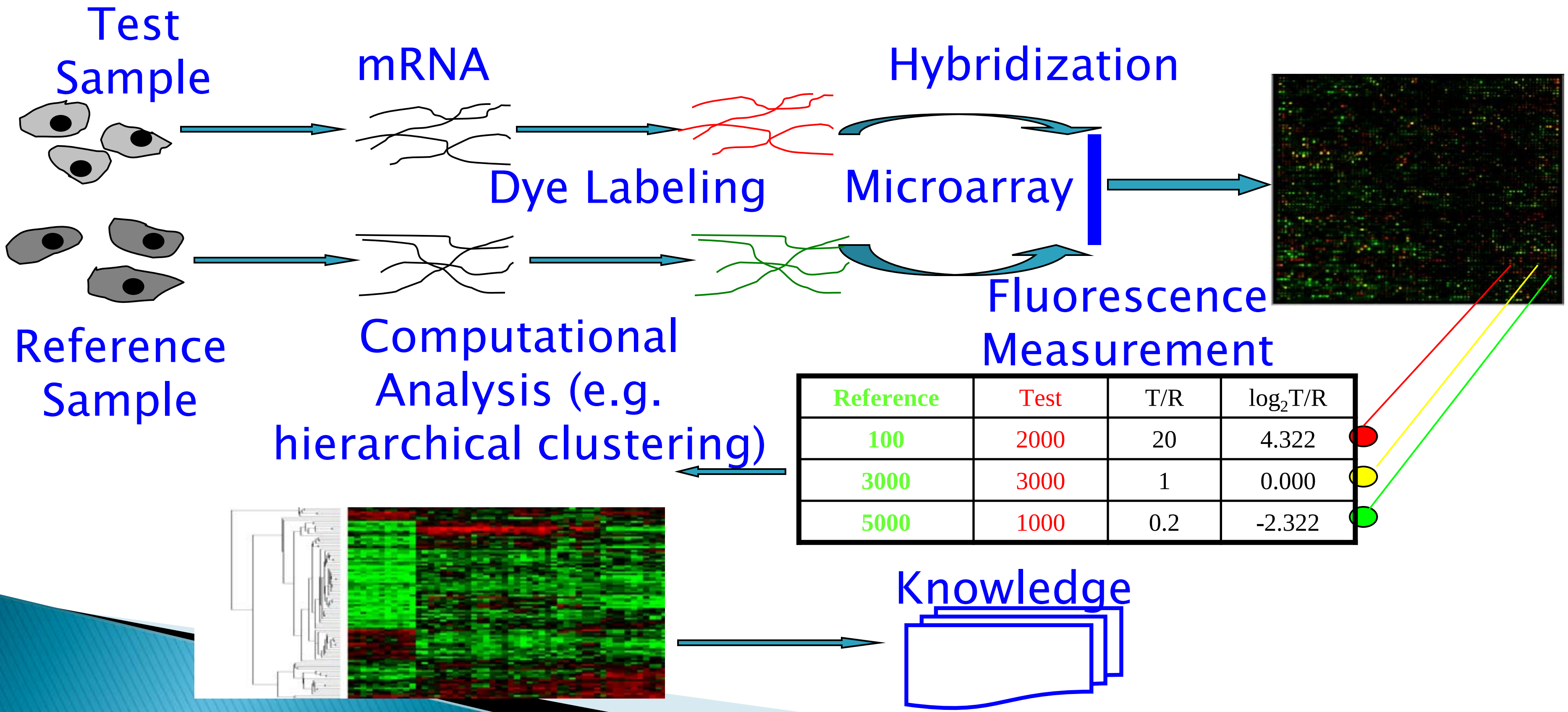
Metrics, e.g., Ellipsoidal ARTMAP



An Ellipsoid ART category

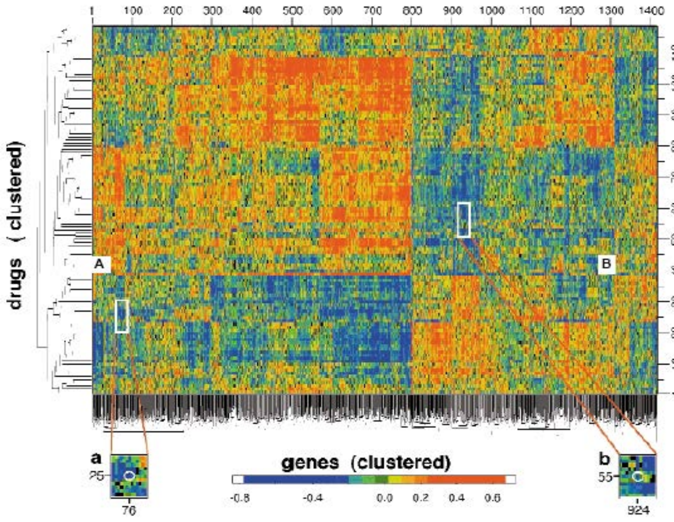


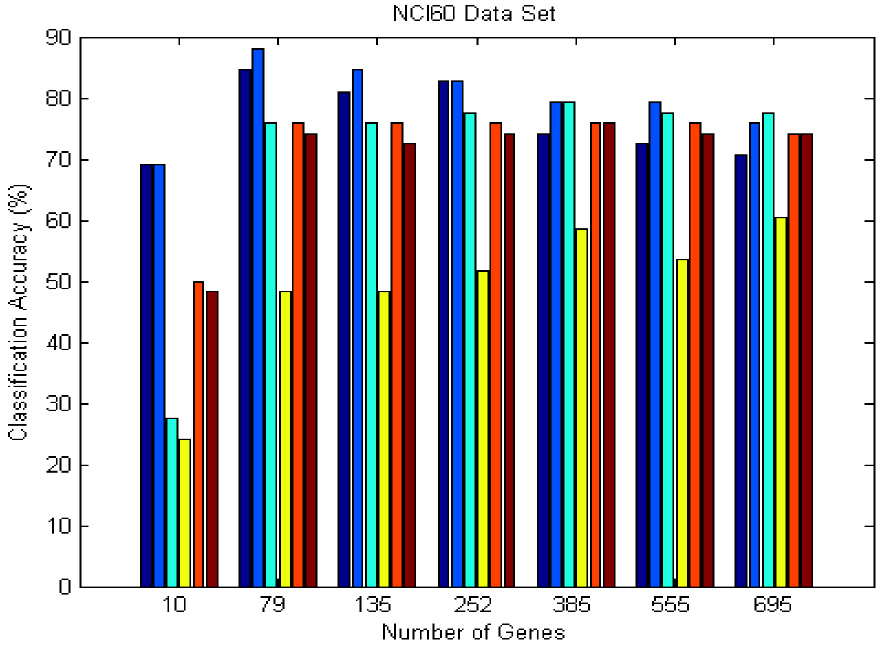
cDNA Microarray Technology



Reference	Test	T/R	$\log_2 T/R$
100	2000	20	4.322
3000	3000	1	0.000
5000	1000	0.2	-2.322

NCI60 Cancer Identification

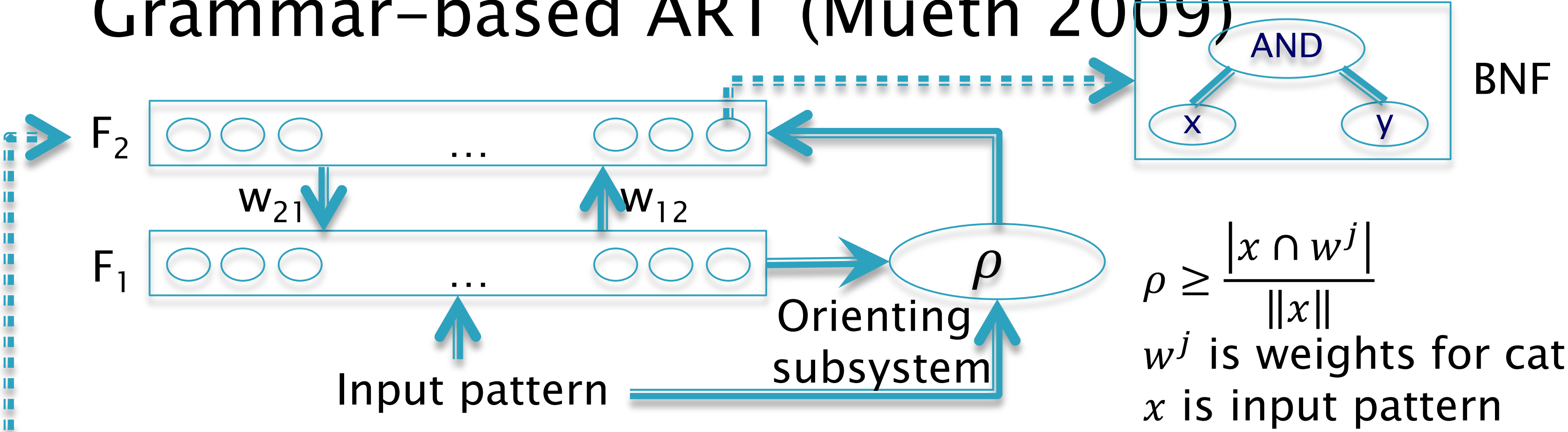




Classification rate comparison: EAM, ssEAM, PNN, ANN, LVQ1, and kNN

Gram-ART

Grammar-based ART (Mueth 2009)



$$\rho \geq \frac{|x \cap w^j|}{\|x\|}$$

w^j is weights for category j
x is input pattern

$$|x \cap w^j| = \sum_{i=0}^{r} w^j_{i,x_i}$$

r = # nodes in tree
$\|x\|$ = # of inputs in x

Category selection

$$T(j) = \frac{|x \cap w^j|}{\|w^j\|}$$

w^j is wgts for category j
x is input pattern

Resonance: weight update

$$w^j_i = \frac{w^j_i * N + \delta_j}{N+1}$$

$$\delta_j = \begin{cases} 1 \; if \; x_i = j \\ 0 \; otherwise \end{cases}$$

prior updates

Ontologies

Sir Tim Berners-Lee

- *WWW*
- *HTML*
- *Semantic Web*

Graph Theory

- *Arbitrary DAG*

W3C

- *OWL*
- *RDF*
- *Other standards*

Computer/Information Science

Formal representation of a set of concepts within a domain and the relationships between those concepts

(Wiki)

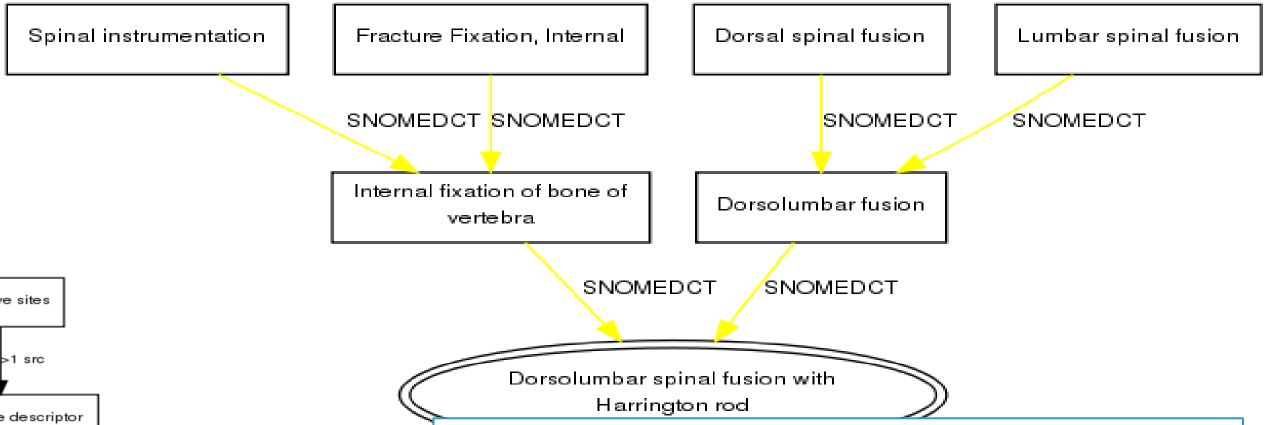
Metaphysics

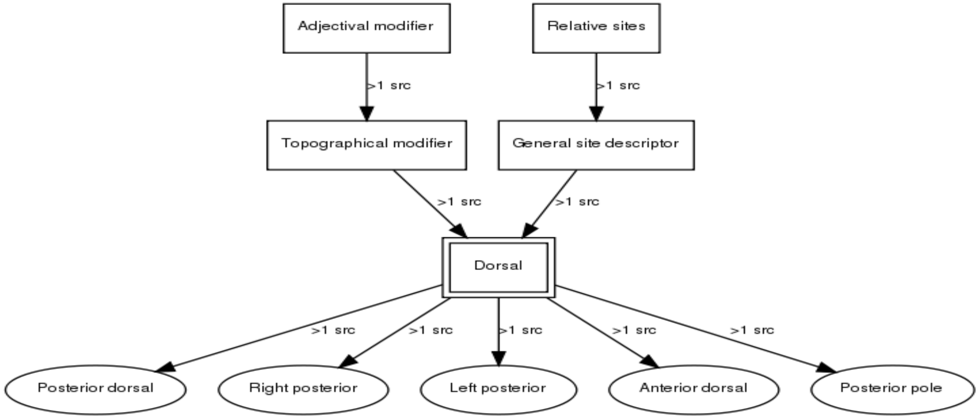
Theory on existence of beings

(Webster dictionary)

ONTOLOGY is_a KNOWLEDGE ARTIFACT

SNOMED Example





Simple schema may be sufficient

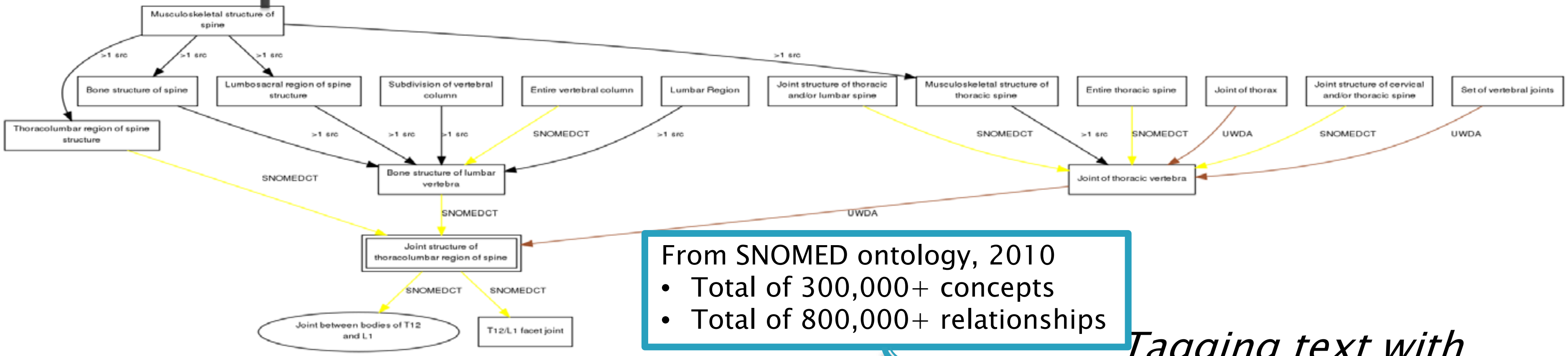
CONCEPT
ID | Name

* *

RELATIONSHIP
From| To | Type

LEXICON
Base | Variant (i.e., dorsal = dorso)

Ontologies Can Be Large, Complicated



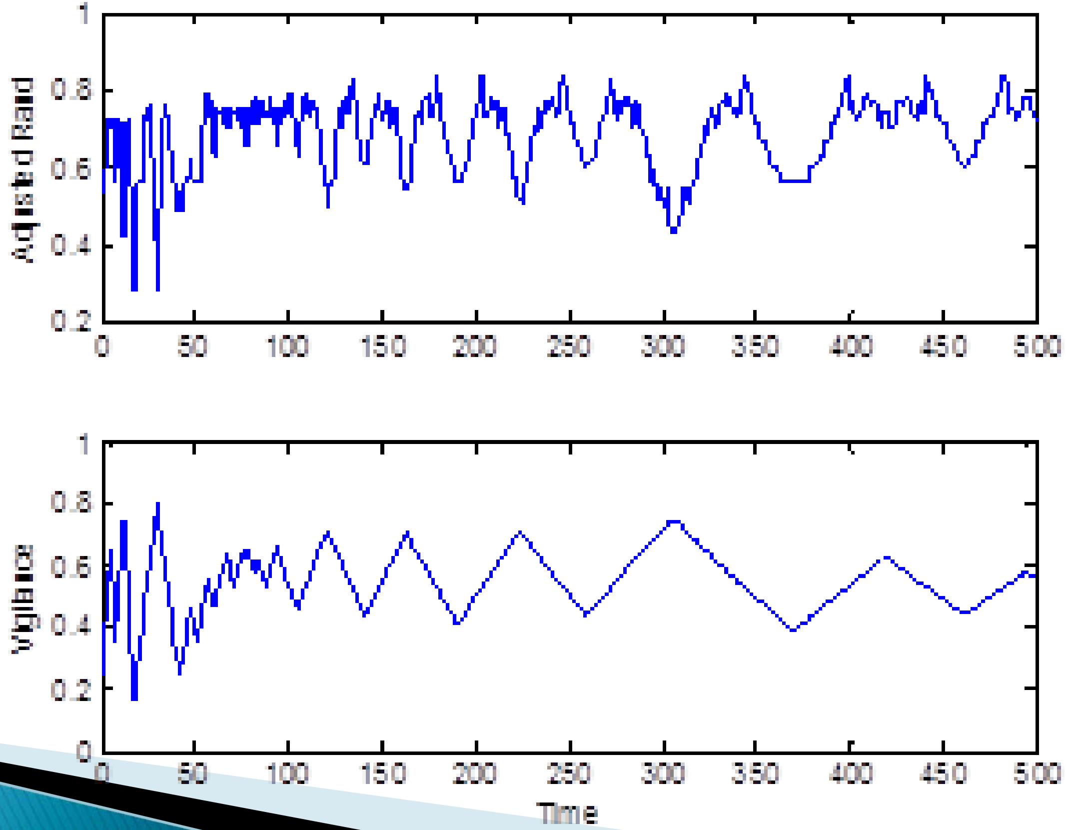
From SNOMED ontology, 2010
- Total of 300,000+ concepts
- Total of 800,000+ relationships

Tagging text with concepts can be NP-complete

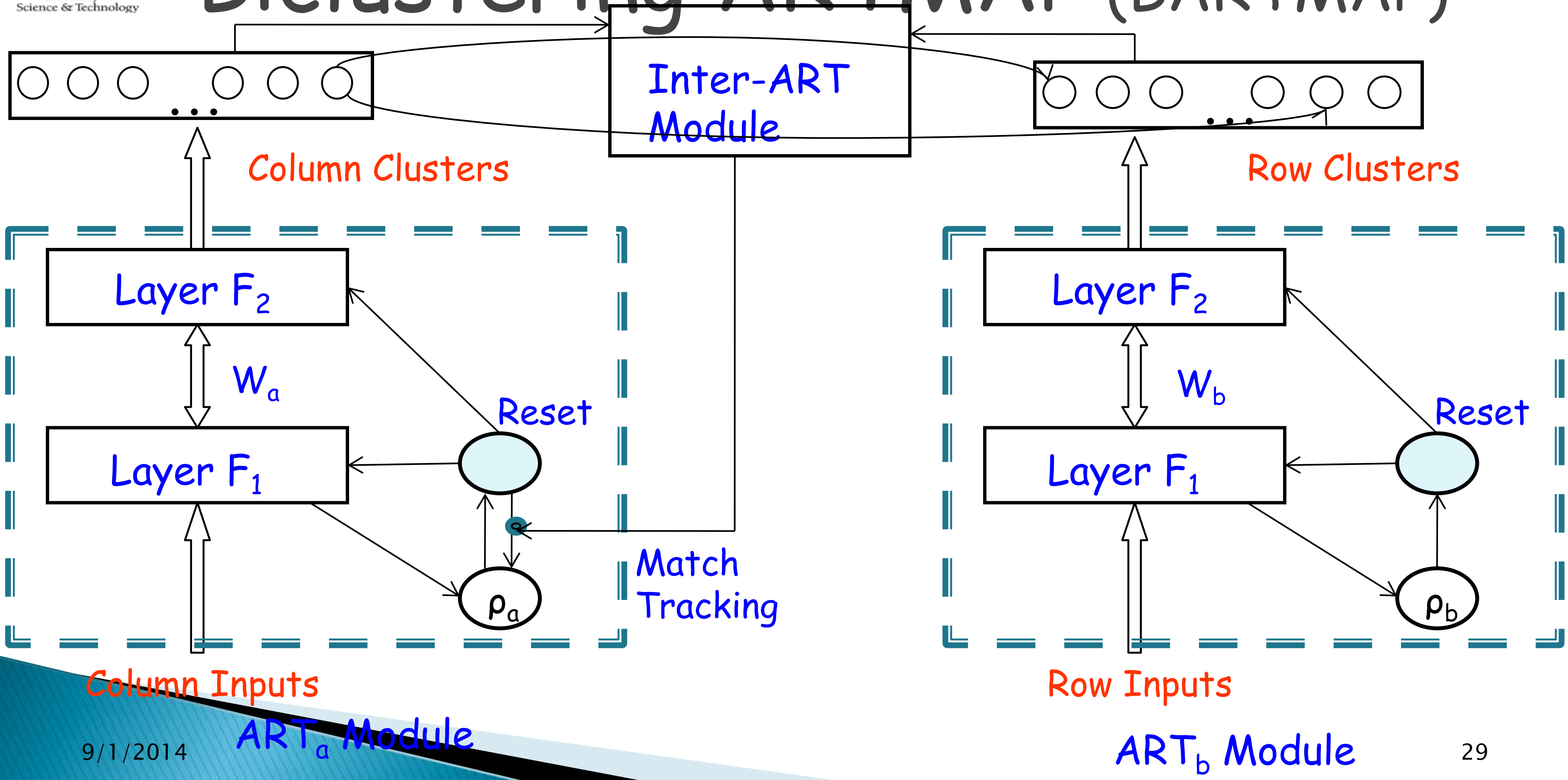
<u>Search criteria can be large and complicated also (real-world example HPI):</u>

A 64-year-old women presents with a 3 cm mass in her left upper lobe, which was not present 18 months previously. Computed tomography confirms the presence of the mass without evidence of mediastinal adenopathy. Transthoracic fine needle aspiration reveals non-small cell lung cancer. The surgeon reviews the patient's medical record, x-ray findings, pulmonary function studies, laboratory results, and bronchoscopy report. A mediastinoscopy has been

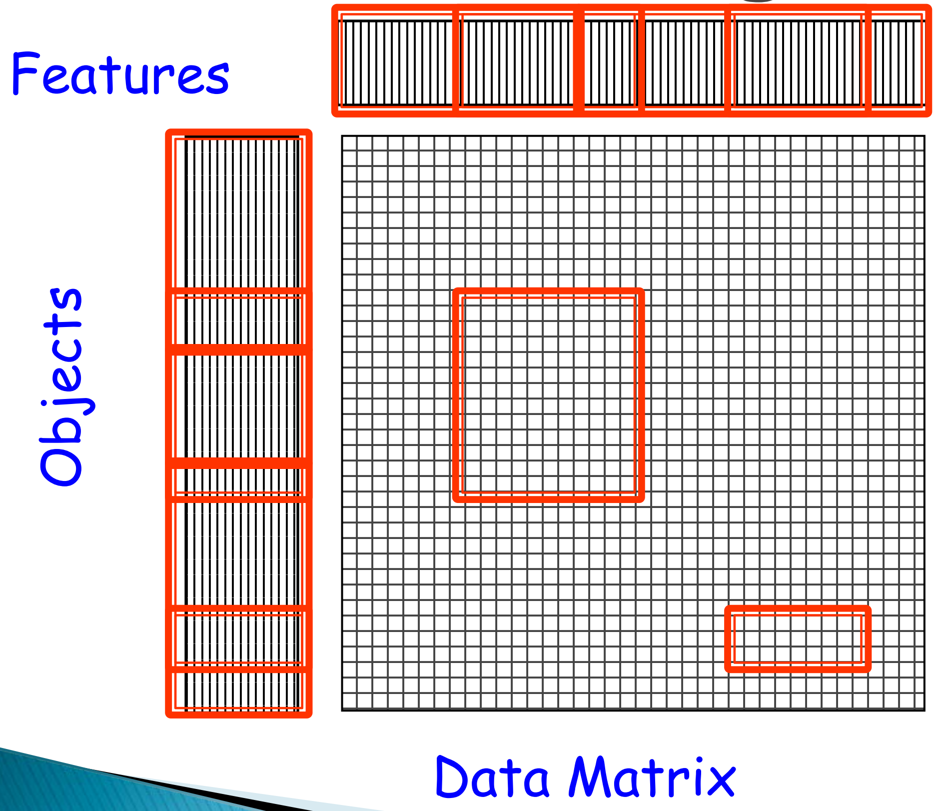
Adaptive Dynamic Programming for Optimizing Clustering



Biclustering ARTMAP (BARTMAP)



Biclustering



- Clustering separately – Global model
  - Rows
  - Columns
- Biclustering (subspace clustering, coclustering, bidimensional clustering) – Clustering of two dimensions simultaneously (clustering + feature selection) – Local model
- How hard? – NP complete
  - Iterative row and column clustering combination
  - Greedy iterative search
  - Distributed parameter identification
  - Divide-and-conquer
  - Exhaustive bicluster enumeration

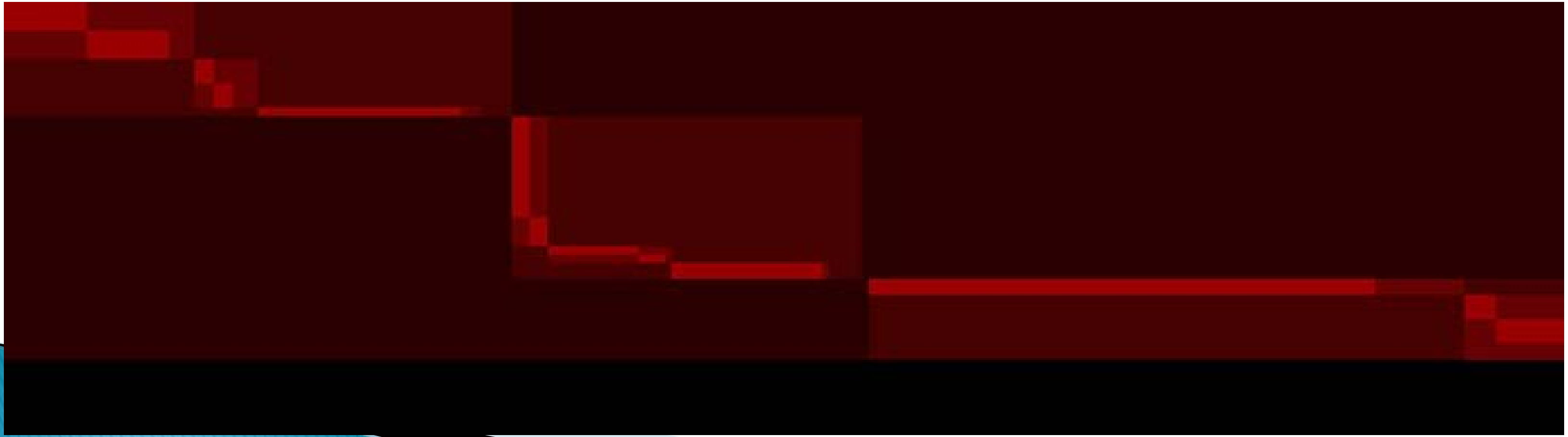
Hierarchical BiFAM

- BARTMAP
  - State of the art biclustering algorithm
  - Significantly outperforms other approaches
- HBiFAM
  - Hierarchical Biclustering Fuzzy ARTMAP algorithm
  - Provides deeper / more precise biclustering

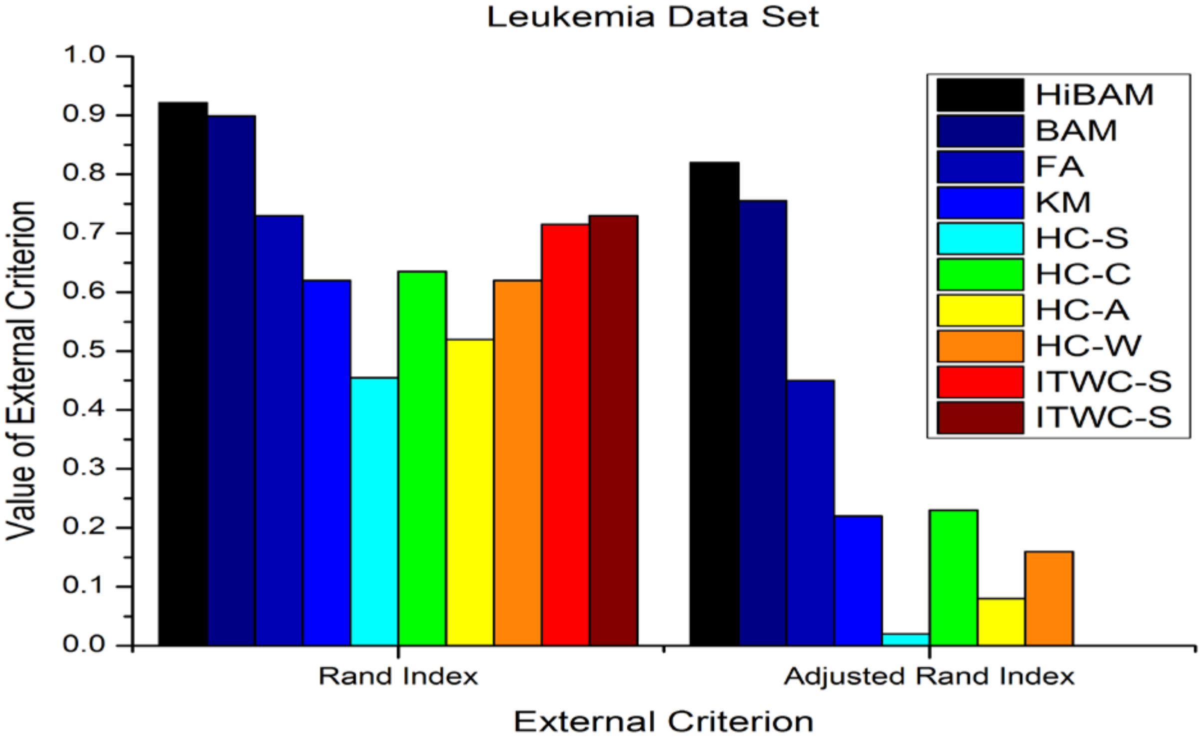
Data source: M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 14863–14868, Dec 1998.

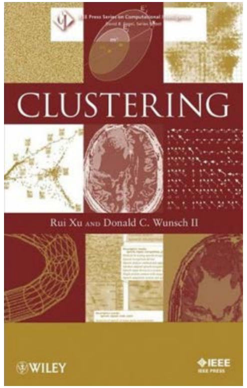
Hierarchical BiFAM

- Figure. Heat map of the correlation between gene and sample of leukemia sample (prototype) presented contrast wise (brighter = more correlated)

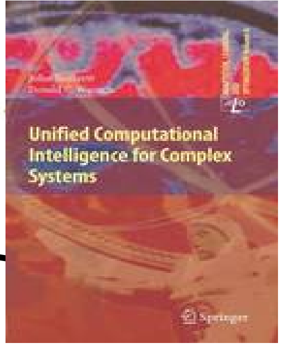


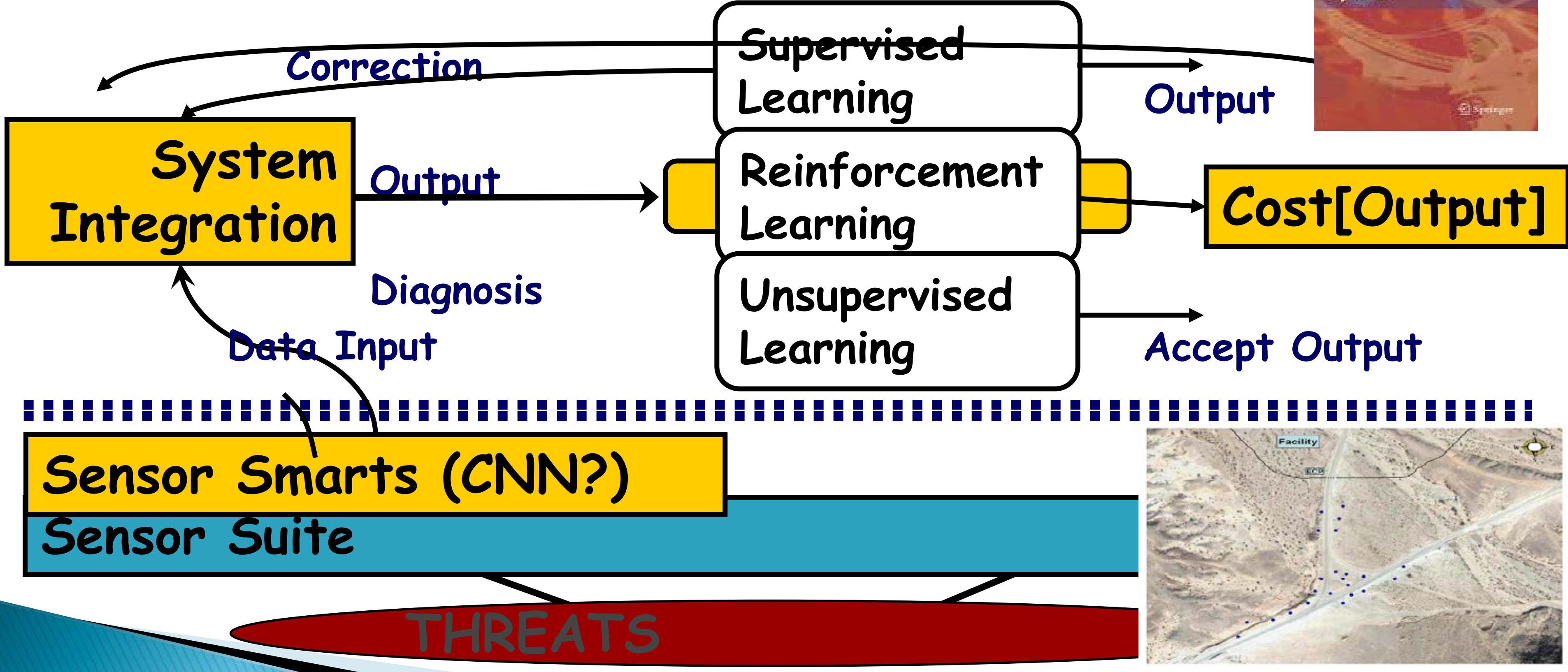
BARTMAP & Hierarchical BARTMAP





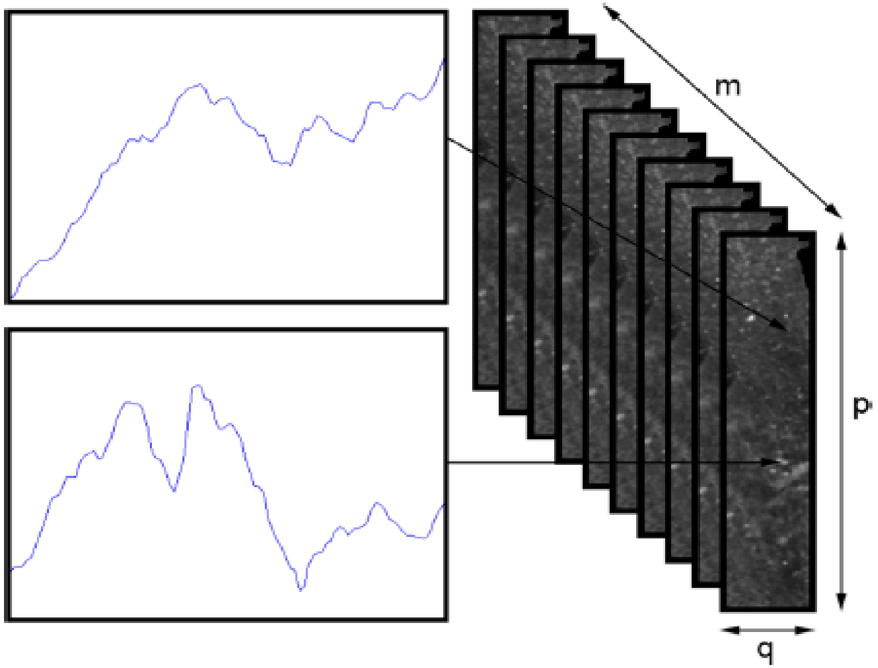
Unified Learning Modalities



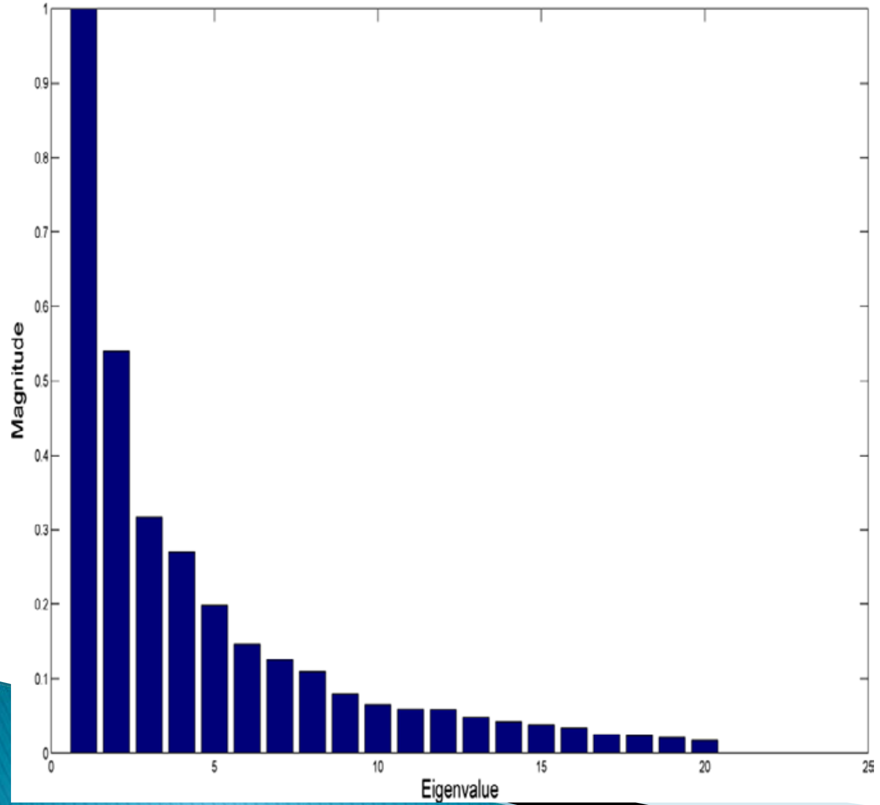


DM & ART for Hyperspectral Imaging

- Every pixel generates a continuous spectrum
- Image -> hypercube
- Agriculture, environment, mining, military
- Particularly challenging at high resolution
- E.g mining samples: over 200 spectral bands
- 250 k pixels / meter
- 5 meters / hour

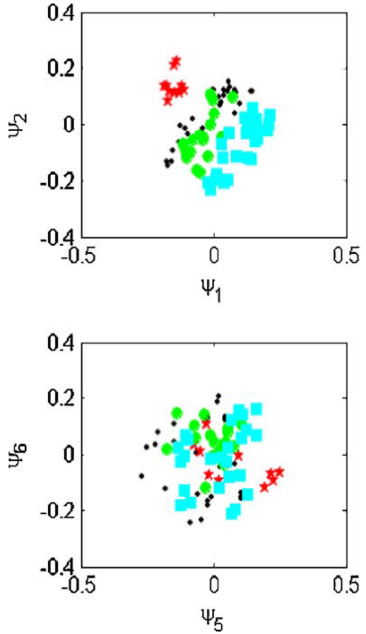


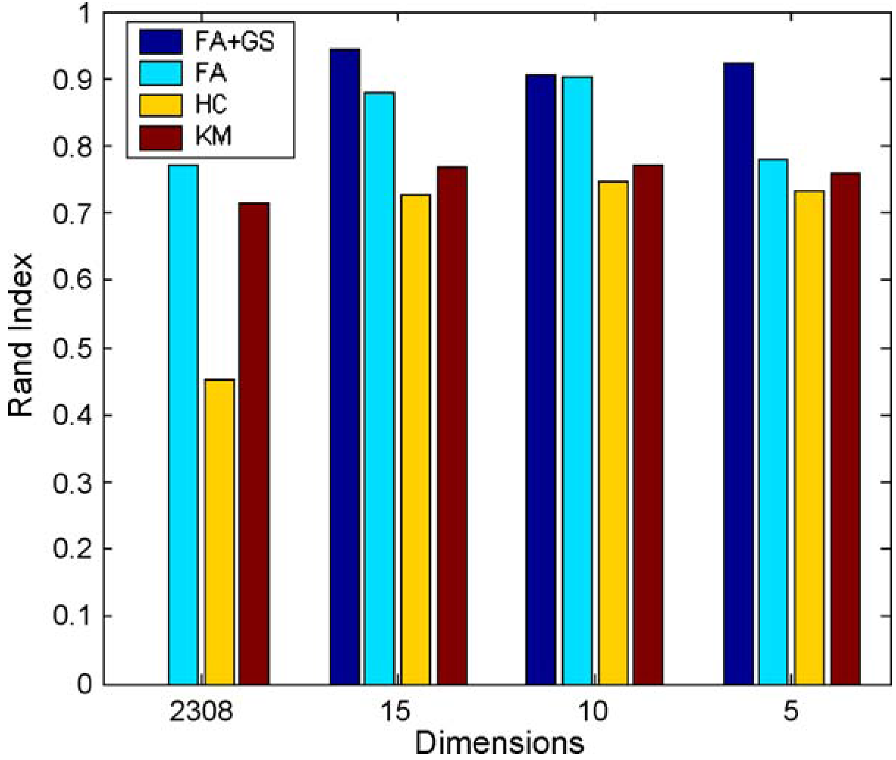
Can Achieve Several Orders of Magnitude Data Reduction



- Magnitude of largest eigenvalues (subset of many)
- Typically sparse matrix, only need the top few eigenvalues
- Amenable to parallelism

Cancer Gene Expression: Small Round Blue Cell Tumors





Using 30 genes instead of 2300. Cluster Visualization. Rand Index vs # used

Conclusions

- Plenty of opportunity in the space between approaches.
- Synergies can create unique capabilities
- No shortage of exciting applications
- The best is yet to come!

Thank You!

Question: Anything for Encore?
Integral Reinforcement Learning

Work of Draguna Vrabie

$$\dot{x} = f(x) + g(x)u$$

Can Avoid knowledge of drift term *f(x)*

Policy iteration requires repeated solution of the CT Bellman equation

$$0 = \dot{V} + r(x,u(x)) = \left(\frac{\partial V}{\partial x}\right)^T \dot{x} + r(x,u(x)) = \left(\frac{\partial V}{\partial x}\right)^T f(x,u(x)) + Q(x) + u^T R u \equiv H(x, \frac{\partial V}{\partial x}, u(x))$$

This can be done online without knowing *f(x)*

using measurements of *x(t), u(t)* along the system trajectories

D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, “Adaptive optimal control for continuous-time linear systems based on policy iteration,” Automatica, vol. 45, pp. 477-484, 2009.

Slide courtesy Frank Lewis

system $\dot{x} = f(x) + g(x)u$

value $V(x(t)) = \int_t^{\infty} r(x,u)\, d\tau$

Key Idea

Lemma 1 – Draguna Vrabie

$$0 = \left(\frac{\partial V}{\partial x}\right)^T f(x,u) + r(x,u) \equiv H(x, \frac{\partial V}{\partial x}, u), \quad V(0) = 0$$

Is equivalent to Integral reinf. form for the CT Bellman eq.

$$V(x(t)) = \int_t^{t+T} r(x,u)\, d\tau \quad + \quad V(x(t+T)), \quad V(0) = 0$$

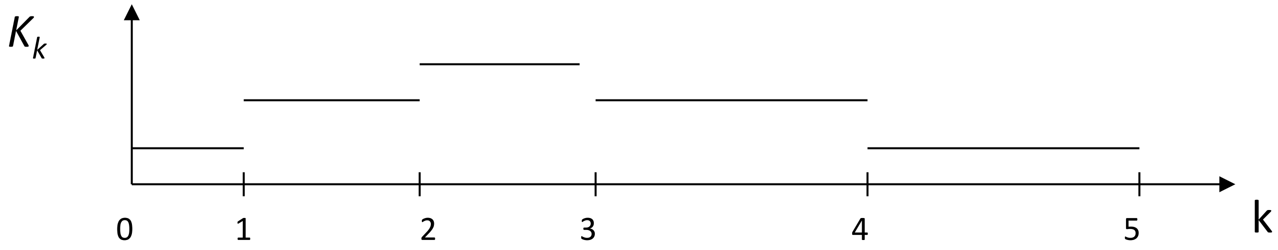
Solves Bellman equation without knowing *f(x), g(x)*

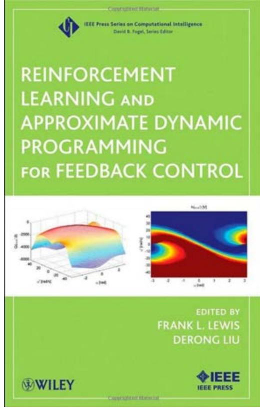
Allows definition of temporal difference error for CT systems

$$e(t) = -V(x(t)) + \int_t^{t+T} r(x,u)\, d\tau \quad + \quad V(x(t+T))$$

Slide courtesy Frank Lewis

Gain update (Policy)

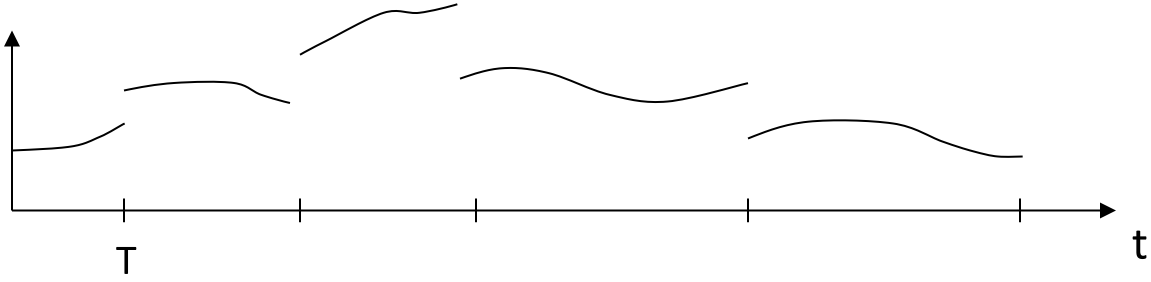




Ch. 15

Control

$$u_k(t) = -K_k x(t)$$



Reinforcement Intervals T need not be the same

They can be selected on-line in real time

Continuous-time control with discrete gain updates

Slide courtesy Frank Lewis

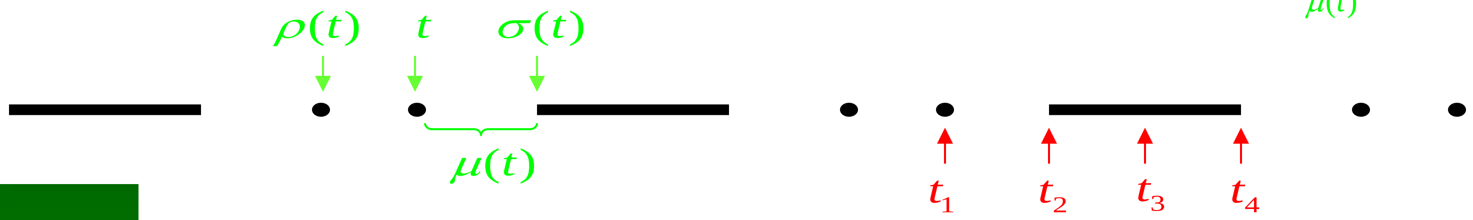
Time Scales Analysis Contributions

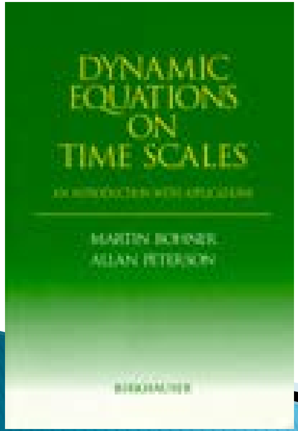
Forward Jump Operator: $\sigma(t) := \inf\{s \in T : s > t\}$

Backward Jump Operator: $\rho(t) := \sup\{s \in T : s < t\}$

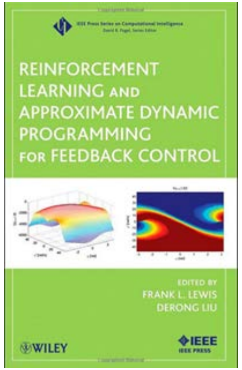
Graininess: $\mu(t) := \sigma(t) - t$

$$f^{\Delta}(t) := \frac{f(\sigma(t)) - f(t)}{\mu(t)} \quad \xrightarrow{\mu \equiv 0} \left(\frac{df}{dt}\right), \qquad \xrightarrow{\mu \equiv 1} \Delta f$$





t_1 is isolated	$\rho(t) < t < \sigma(t)$
t_2 is left-scattered (right-dense)	$\rho(t) < t = \sigma(t)$
t_3 is dense	$\rho(t) = t = \sigma(t)$
t_4 is right-scattered (left-dense)	$\rho(t) = t < \sigma(t)$



Seiffertt & Wunsch, Ch. 21

Let $x_1, \ldots, x_n$ be ordered variables such that $x_i \in T_i$ and $x_i = f_i(x_1, \ldots, x_{i-1})$

Define $F_n(x_1, \ldots, x_n) = x_n$ and $F_{i-1}(x_1, \ldots, x_{i-1}, f_i(x_{1,} \ldots, x_{i-1}))$

Define ordered delta derivative as $x_n^{\Delta^+_{x_i}} = F_i^{\Delta x_i}$

Theorem: $F_j^{\Delta_{x_i}} = \sum_{k=j+1}^{n} x_n(\sigma_1(x_1), x_2, \ldots, x_{n-1})^{\Delta^+_{x_i}} x_k^{\Delta_{x_i}}$

Hamilton-Jacobi-Bellman Equation:

$$0 = \min_u \left\{ r(t) + J^{\Delta_t}(x(t), t) + J^{\Delta_x}(x(t), \sigma(t)) f(x(t), t) \right\}$$

Theorem: Suppose $V(x(t),t)$ solves , $u^*(x(t),t)$ minimizes , $V(x(T),T) = r(x(T))$, $\hat{x}(t_0) = x(t_0)$, $x^*(t)$ is a state trajectory , and $x^*(t_0) = x(t_0)$. Then $V(x(t),t)$ and $u^*(x(t),t)$ are optimal.