The conditionality principle in dimensionality reduction

Carey E. Priebe

cep@jhu.edu

Department of Applied Mathematics & Statistics
Department of Computer Science
Center for Imaging Science

Johns Hopkins University

Multiscale Geometry and Analysis in High Dimensions

Multiscale Structures in the Analysis of High-Dimensional Data

October 25-29, 2004

IPAM, UCLA

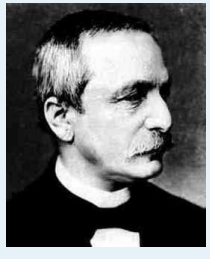
Kronecker Quote

"The wealth of your practical experience

with sane and interesting problems

will give to mathematics

a new direction and a new impetus."



– Leopold Kronecker to Hermann von Helmholtz –

< > - +

- **sensing —**

$$s : \mathcal{R} \rightarrow \mathbb{R}^p, \; p >> 1$$

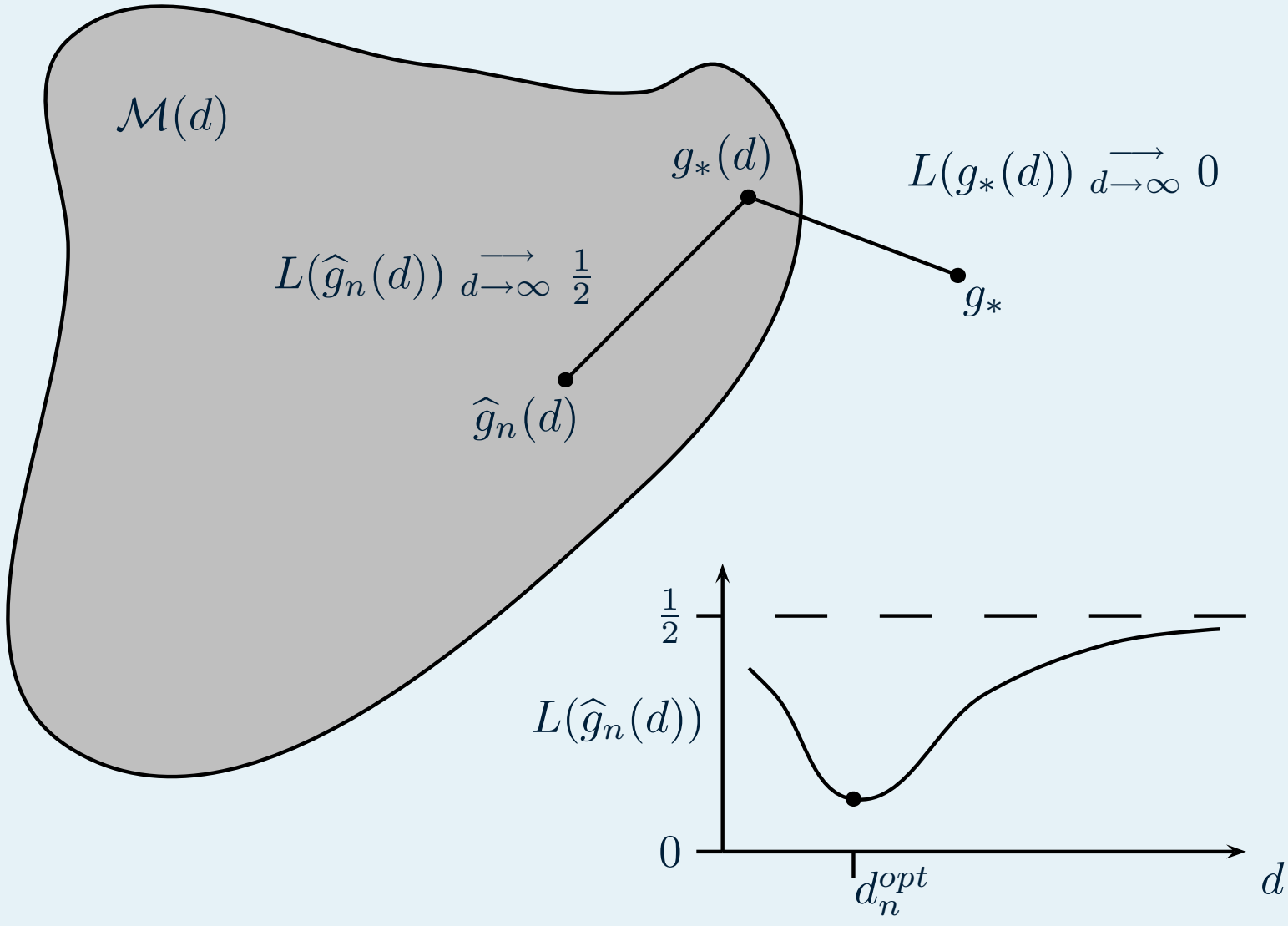
- **dimension reduction —**

$$r : \mathbb{R}^p \rightarrow \mathbb{R}^q, \; q << p$$

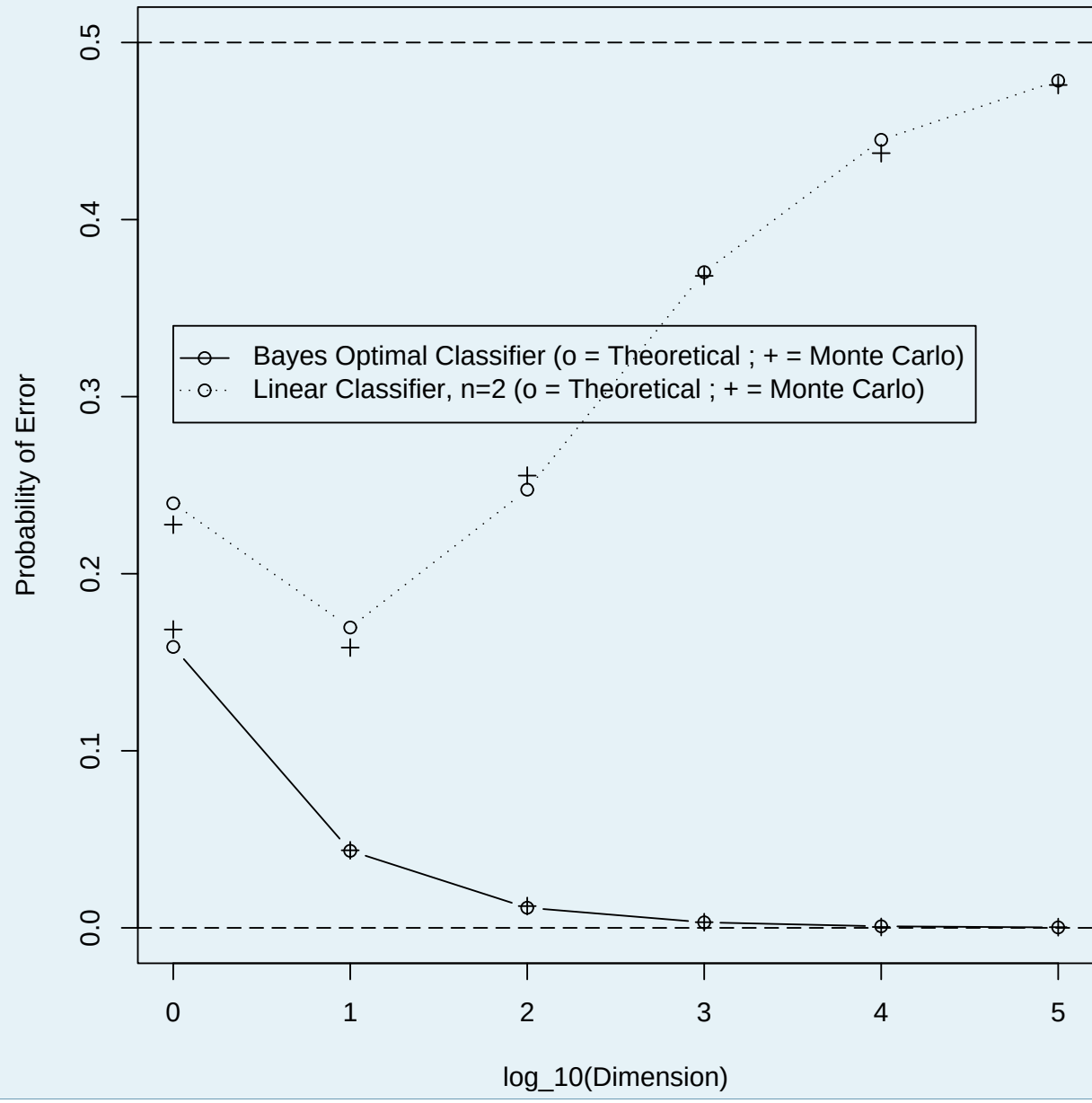
- **classification/clustering —**

$$g : \mathbb{R}^q \rightarrow [0,1]^{\mathcal{C}}$$

cep@jhu.edu

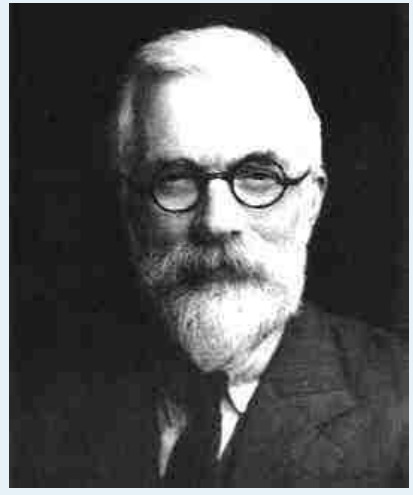

$\mathcal{M}(d)$
$g_*(d)$
$L(g_*(d)) \xrightarrow[d\to\infty]{} 0$
$L(\widehat{g}_n(d)) \xrightarrow[d\to\infty]{} \frac{1}{2}$
g_*
$\widehat{g}_n(d)$
$\frac{1}{2}$
$L(\widehat{g}_n(d))$
0
d_n^{opt}
d

Trunk (1979)


Probability of Error
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
log_10(Dimension)
Bayes Optimal Classifier (o = Theoretical ; + = Monte Carlo)
Linear Classifier, n=2 (o = Theoretical ; + = Monte Carlo)

Fisher's Conditionality Principle

Foundations of Statistical Inference:
Likelihood Principle, Sufficiency Principle, Conditionality Principle



Fisher's Conditionality Principle (1950,1956)

But: Welch (1939)?

cep@jhu.edu

Amari (1985)

Consider Amari's 1985 statement of the Conditionality Principle:

"When there exists an exact ancillary statistic r,
the conditionality principle requires that
the statistical inference should be performed
by conditioning on r. ..."

Shun-ichi Amari,
Differential Geometric Methods in Statistics,
Lecture Notes in Statistics, Vol 28, 1985,
page 217.

Amari continues ...

(inference about u is the goal)

"... A statistical problem then is decomposed into subproblems in each of which r is fixed at its observed value, thus dividing the whole set of the possible data points into subclasses. It is expected that each subclass consists of relatively homogeneous points with respect to the informativeness about u. We can then evaluate our conclusion about u based on r, and it gives a better evaluation than the overall average one. This is a way of utilizing information which ancillary r conditionally carries."

The X-Property

Most of the basic work in classifier construction
(Random Forests, Support Vector Machines, etc.)
focuses single-mindedly on classification at all times.
That is, a feature's value is in direct proportion
with its utility in estimating the true but unknown class label.
Little or no use is made of features ancillary to this task.

See *The X-property* (D-G-L, 1996, p. 319) for a general exception:

"[T]he form of the tree is determined by the [feature vectors] only, that is, the labels . . . do not play a role in constructing the partition, but they are used in voting."

cep@jhu.edu

ISP Decision Trees & Iterative Denoising

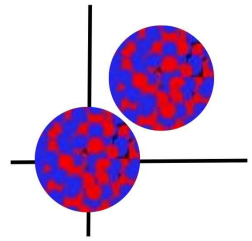
Priebe, C.E., Marchette, D.J., and Healy, D.M. (2004).
Integrated Sensing and Processing Decision Trees.
IEEE Trans. PAMI, Vol. 26, No. 6, pp. 699–708.

Priebe, C.E., et al. (2004).
Iterative Denoising for Cross-Corpus Discovery.
COMPSTAT: Proceedings in Computational Statistics,
16th Symposium Held in Prague, Czech Republic, 2004.
Edited by Jaromir Antoch.
Physica-Verlag, Springer. pp. 381–392.

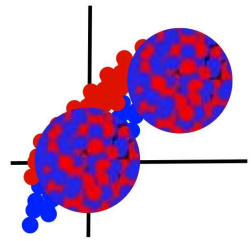
Example

(from Priebe *et al*., PAMI '04)

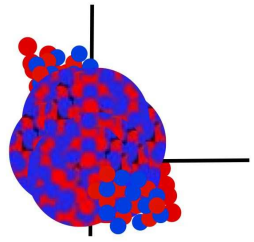
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim$$



$$\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} \sim$$

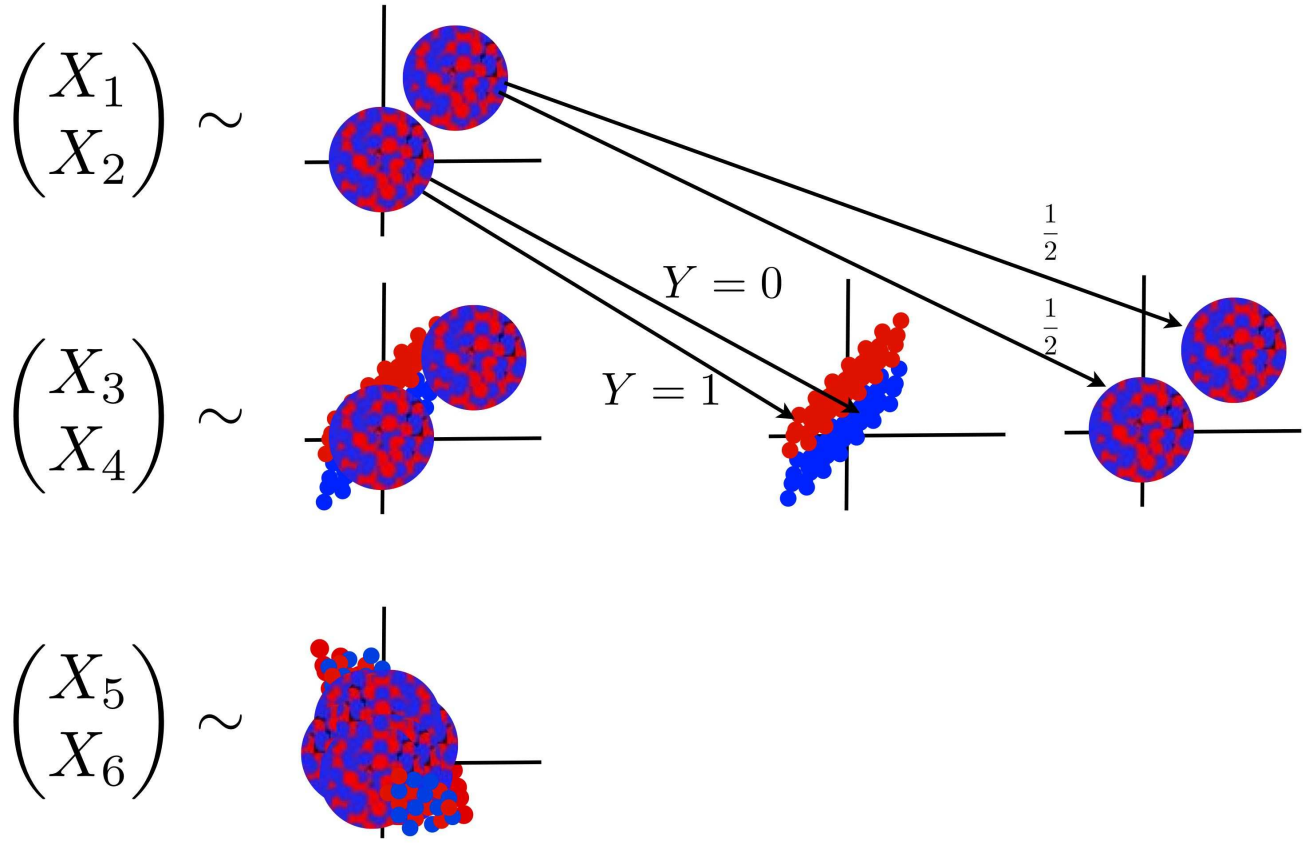


$$\begin{pmatrix} X_5 \\ X_6 \end{pmatrix} \sim$$



Example

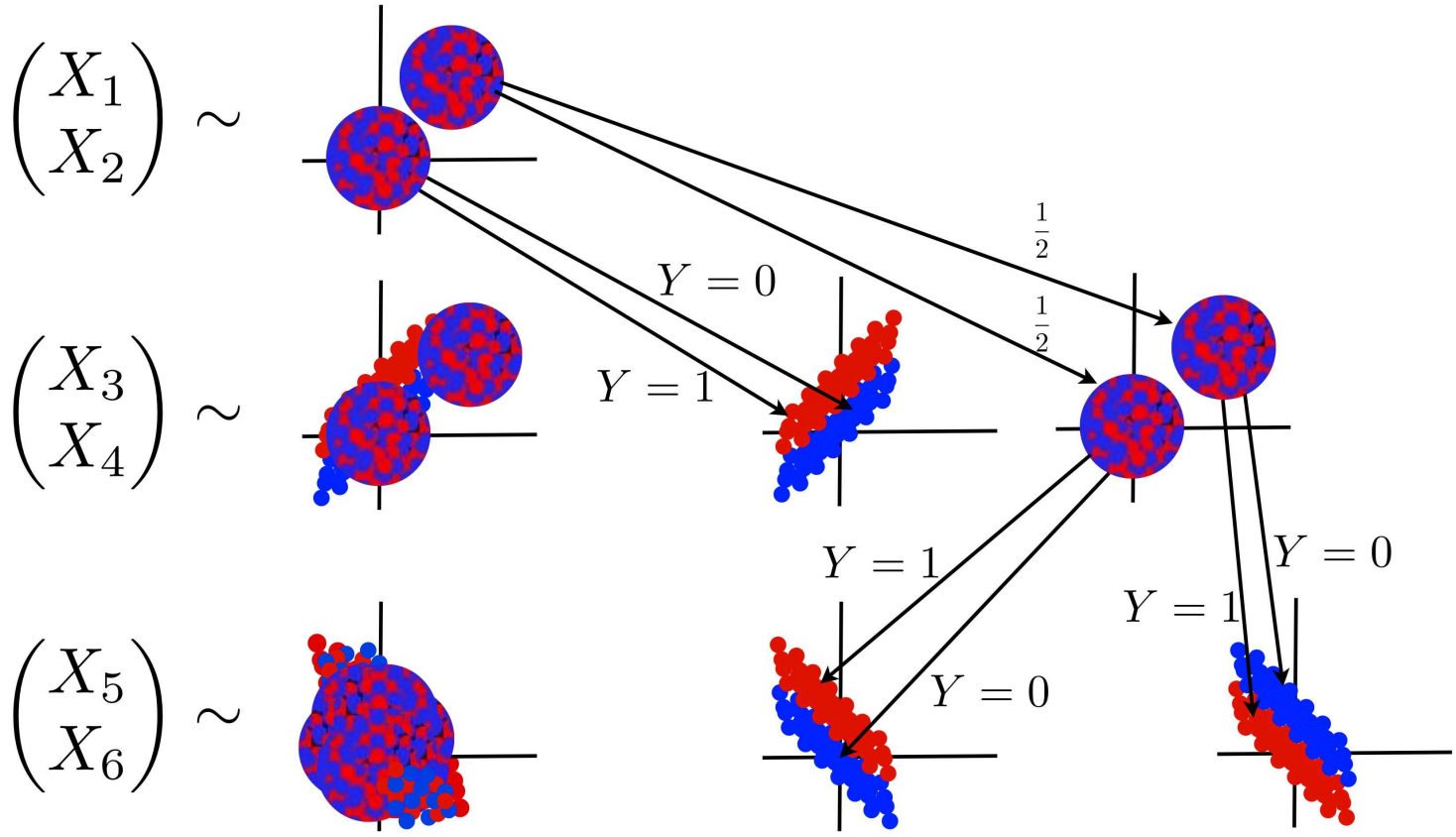
(from Priebe *et al.*, PAMI '04)



cep@jhu.edu

Example

(from Priebe *et al.*, PAMI '04)



cep@jhu.edu

Pedagogical Example

In this pedagogical example,
a willingness to consider partitioning
based on *ancillary features*
when constructing the tree
yields a superior solution:
the iterative denoising tree.

P-M-H, IEEE PAMI, 2004.

Theorem

Given $\epsilon > 0$, we construct F_{XY} with feature vectors $X = [X_1, \cdots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$\min_{g:\mathbb{R}^d \to \{0,1\}} \min_{i,j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon$$

while

$$\exists g \quad \text{with} \quad P[g(X_1, X_{X_1}) \neq Y] = 0,$$

and X_1 is ancillary for the classification task at hand.

That is, there is no pair of features X_i, X_j which work, while conditioning on the ancillary X_1 —using X_1, X_{X_1} —works.

P-M-H, IEEE PAMI, 2004.

Theorem

Given $\epsilon > 0$, we construct F_{XY} with feature vectors $X = [X_1, \cdots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$\min_{g:\mathbb{R}^d \to \{0,1\}} \min_{i,j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon$$

while

$$\exists g \quad \text{with} \quad P[g(X_1, X_{X_1}) \neq Y] = 0,$$

and X_1 is ancillary for the classification task at hand.

That is, there is no pair of features X_i, X_j which work, while conditioning on the ancillary X_1 —using X_1, X_{X_1} —works.

P-M-H, IEEE PAMI, 2004.

Theorem

This theorem illustrates the potential benefit
of a willingness to consider
conditionality / partitioning / X-property methods.

cep@jhu.edu

But there is an added benefit ...

(huge bags of cash, only $2.98 each)

Corpus-Dependent Feature Extraction

CDFE:

The features —

the features sensed, as with an adaptive sensor, or

the features extracted, as with dimensionality reduction —

depend on the collection of entities under consideration.

Appropriate denoising can allow the new “local” features,
extracted at internal nodes of the tree,
to be considerably more valuable than their “global” brethren.

A Hyperspectral Imaging Example

HYMAP airborne hyperspectral scanner

Characteristics

Number of Bands: 126

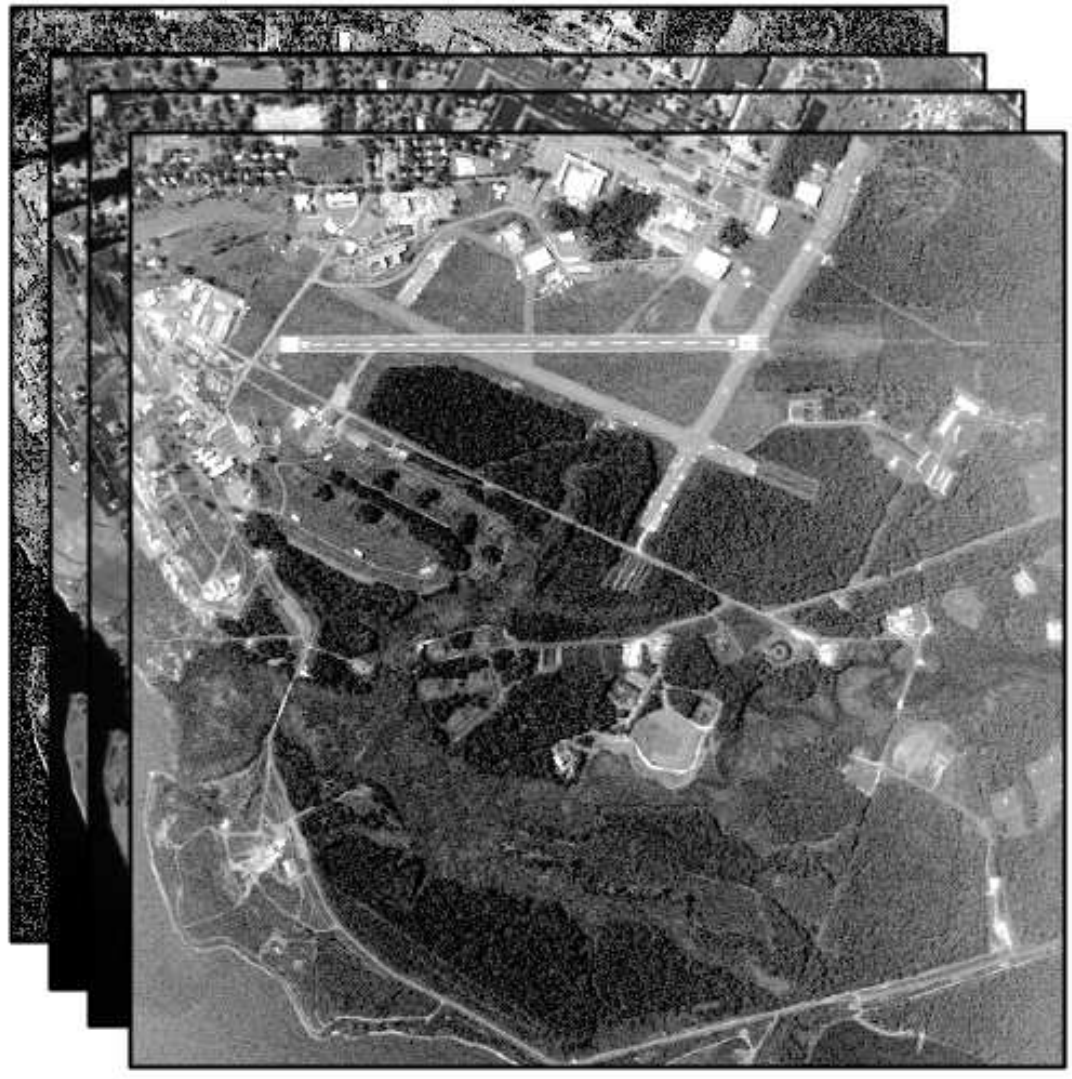
Wavelength: 450 - 2500 nm

Spatial resolution: varies

Spectral resolution: 15 - 20 nm

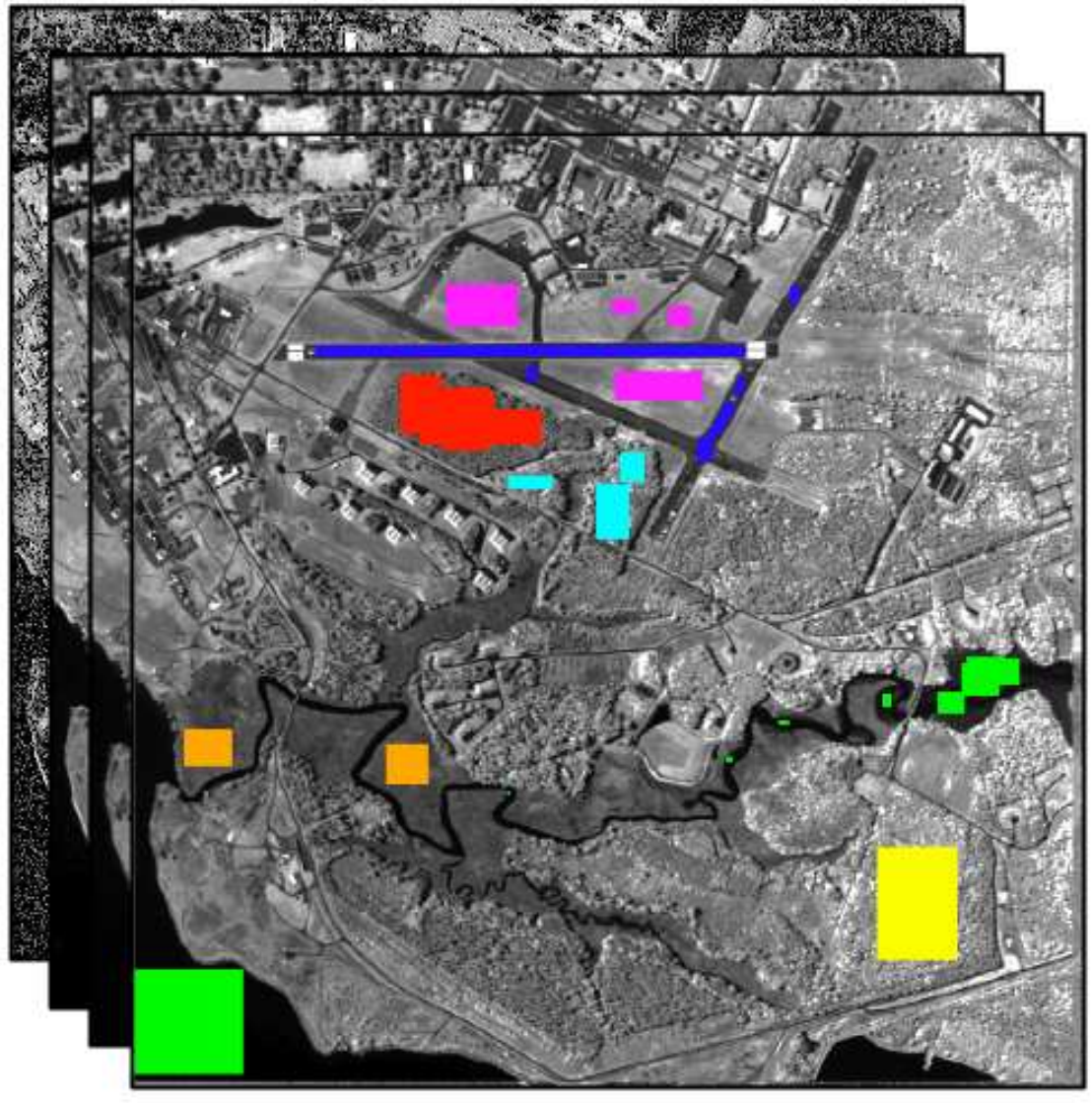
cep@jhu.edu

a HYMAP hyperspectral image

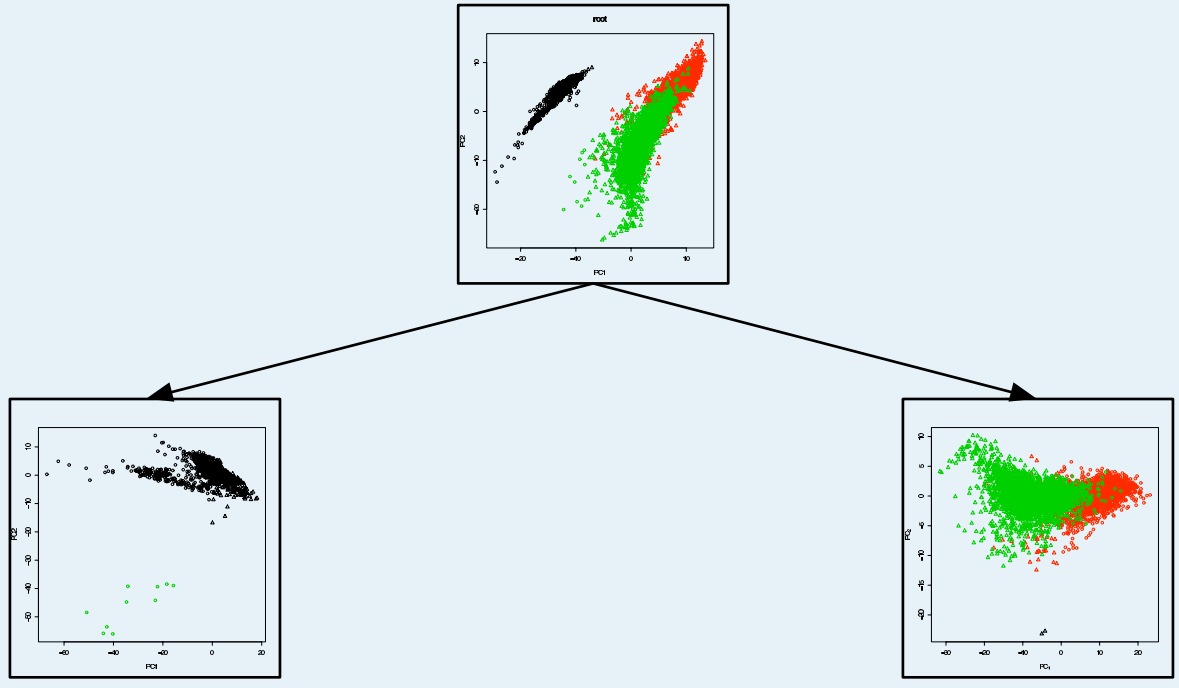


cep@jhu.edu

a HYMAP hyperspectral image



HYMAP hyperspectral imaging example I



black: runway, red: pine, green: scrub

< > - +

HYMAP hyperspectral imaging example II

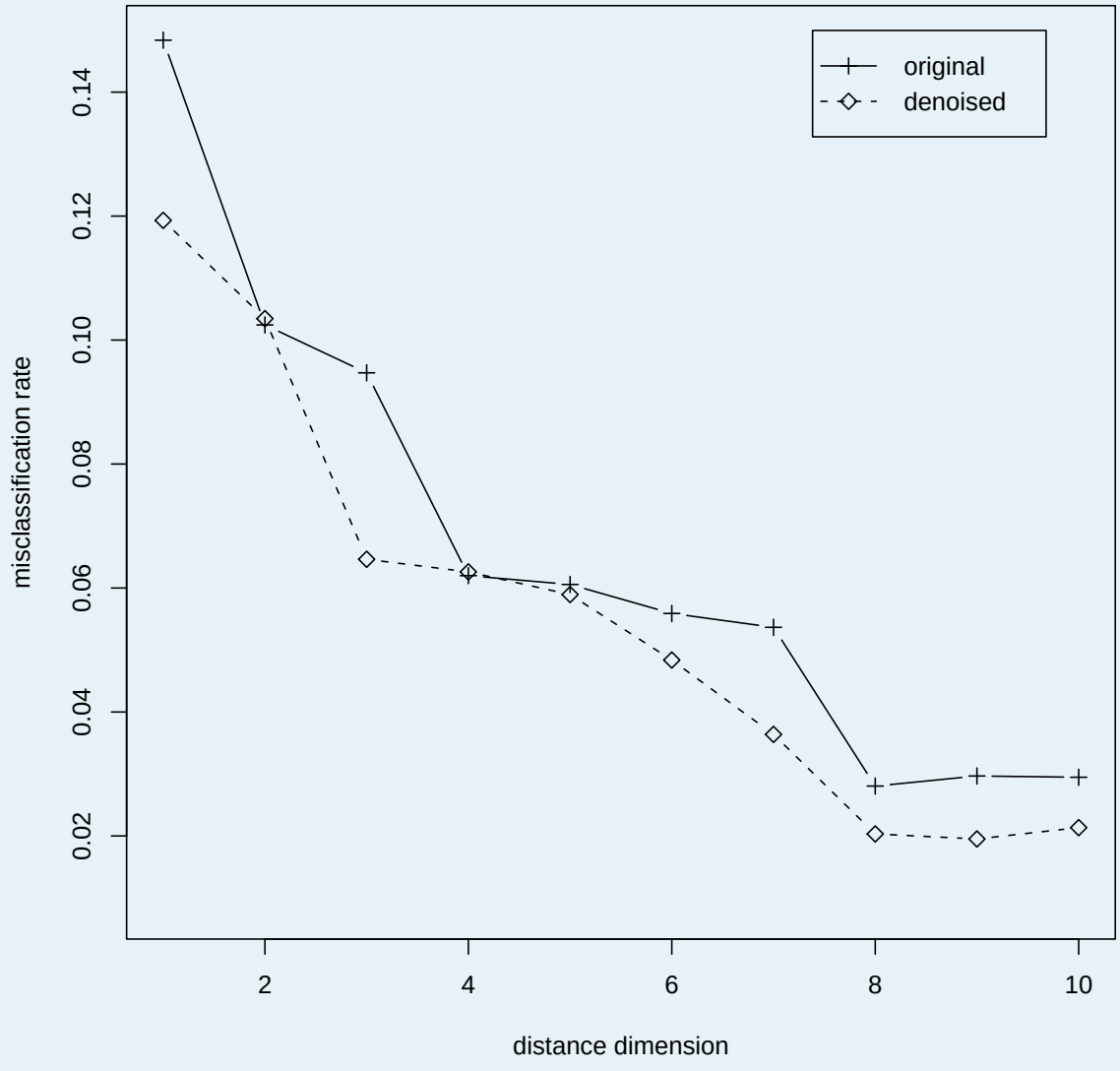
	original	denoised
misclassification of scrub as pine	367	284
misclassification of pine as scrub	363	293
sum	730	577 (+ 10)

HYMAP Misclassification Errors

|pine|=2236 ; |scrub|=2284

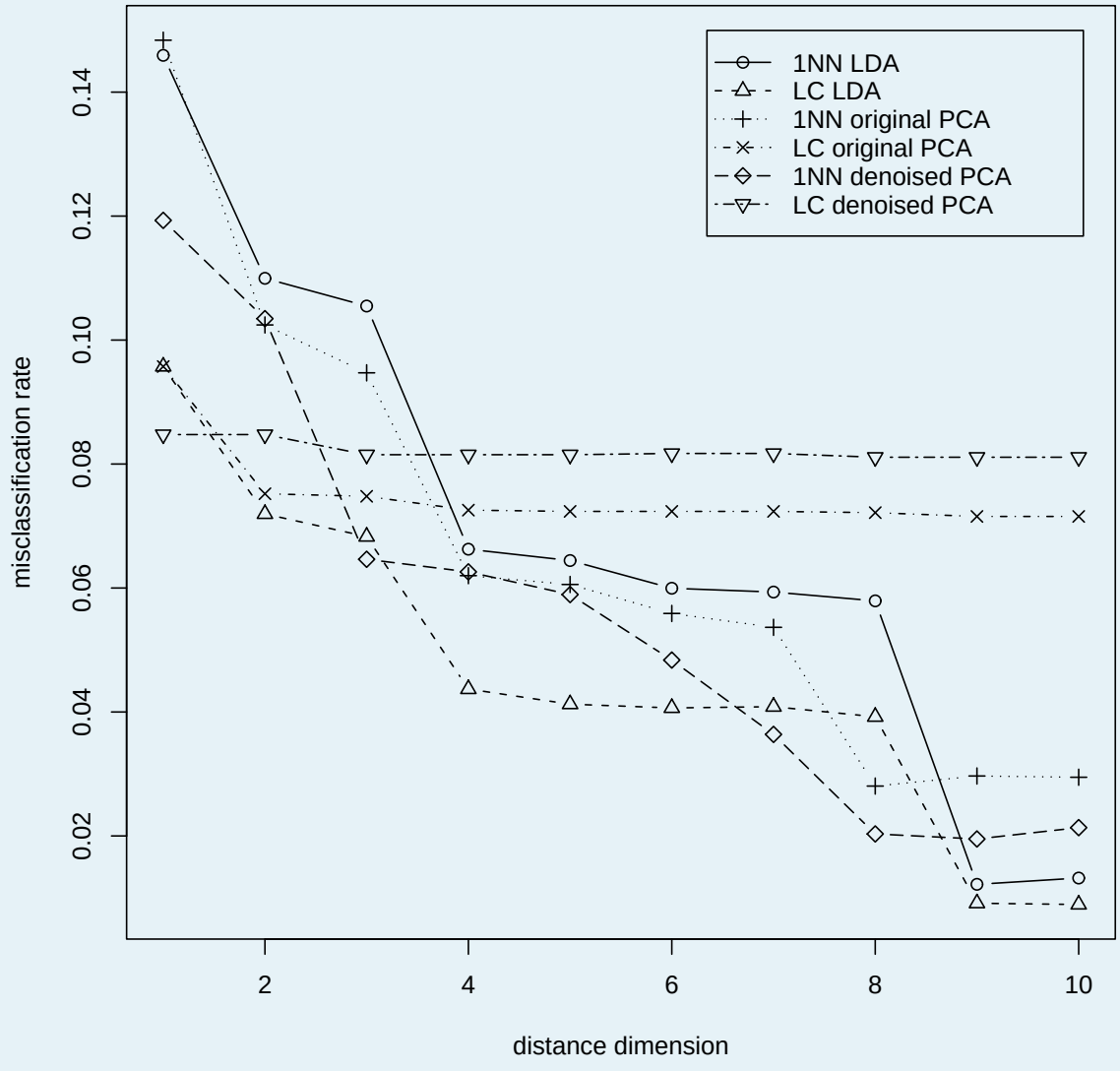
Note: 10 scrub observations are misclassified (as runway) via the denoising partition

HYMAP hyperspectral imaging example III



< > - +

HYMAP hyperspectral imaging example IIIb



cep@jhu.edu

A Text Document Example

Glimpses of Genius

Mathematicians and historians piece together a puzzle that Archimedes pondered

Erica Klarreich

At the start of the 20th century, a Danish mathematical historian named Johan Ludvig Heiberg made a once-in-a-lifetime find. Tucked away in the library of a monastery in Istanbul was a medieval parchment containing copies of the works of the ancient Greek mathematician Archimedes, including two never-before-seen essays. To mathematicians' astonishment, one of the new essays contained many of the key ideas of calculus, a subject supposedly invented two millennia after Archimedes' time. The essay caused a sensation and landed Heiberg's discovery on the front page of a 1907 New York Times.

RECYCLING PROGRAM. In the 13th century, monks reused a copy of Archimedes' work by writing prayer texts (horizontal lines) over the underlying mathematical text (vertical lines).

The other new essay, by contrast, mystified mathematicians. A fragment of a treatise called the Stomachion, it appeared to be nothing more than a description of a puzzle that might have been a children's toy. Mathematicians wondered why Archimedes, whose other works were so monumental, should have spent his time on something so frivolous.

The Stomachion fragment offered only tantalizing glimpses into Archimedes thinking. The parchment, probably first written on in the 10th century in Constantinople, is a palimpsest—a document whose surface has been scraped and reused. In 1204, the Fourth Crusade sacked Constantinople, and shortly thereafter monks unbound the Archimedes's parchment, did their best to erase the mathematical text, and recycled the volume as a Christian prayer book. Only the beginning pages of the Stomachion made it into the new book, and on those pages, the underlying math text was faint and hard to read.

< > - +

Armed with only a magnifying glass, Heiberg managed to make out a large portion of the palimpsest. What he read, however, offered few clues to Archimedes' interest in the puzzle. And before other scholars could

Example: Lin & Pantel Mutual Information Feature

$$\mathcal{L}_C(\cdot) : \mathcal{D}\text{ocument}\mathcal{S}\text{pace} \rightarrow [\mathcal{M}\text{utual}\mathcal{I}\text{nformation}\mathcal{F}\text{eature}]^{d_{\mathcal{L}}(C)}.$$

Both the features themselves and the number of features $d_{\mathcal{L}}(C)$ depend on the corpus C. Thus $\mathcal{L}_C(C)$ is a $|C| \times d_{\mathcal{L}}(C)$ *mutual information feature matrix*. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document x in corpus C, and associated word w, the mutual information between x and w is given by

$$m_{x,w} = \log \left(\frac{f_{x,w}}{\sum_{\xi \in C} f_{\xi,w} \sum_{\omega} f_{x,\omega}} \right).$$

Here $f_{x,w} = c_{x,w}/N$ where $c_{x,w}$ is the number of times word w appears in document x and N is the total number of words in the corpus C.

cep@jhu.edu

Example: Lin & Pantel Mutual Information Feature

$$\mathcal{L}_C(\cdot) : \mathcal{D}\text{ocument}\mathcal{S}\text{pace} \to [\mathcal{M}\text{utual}\mathcal{I}\text{nformation}\mathcal{F}\text{eature}]^{d_{\mathcal{L}}(C)}.$$

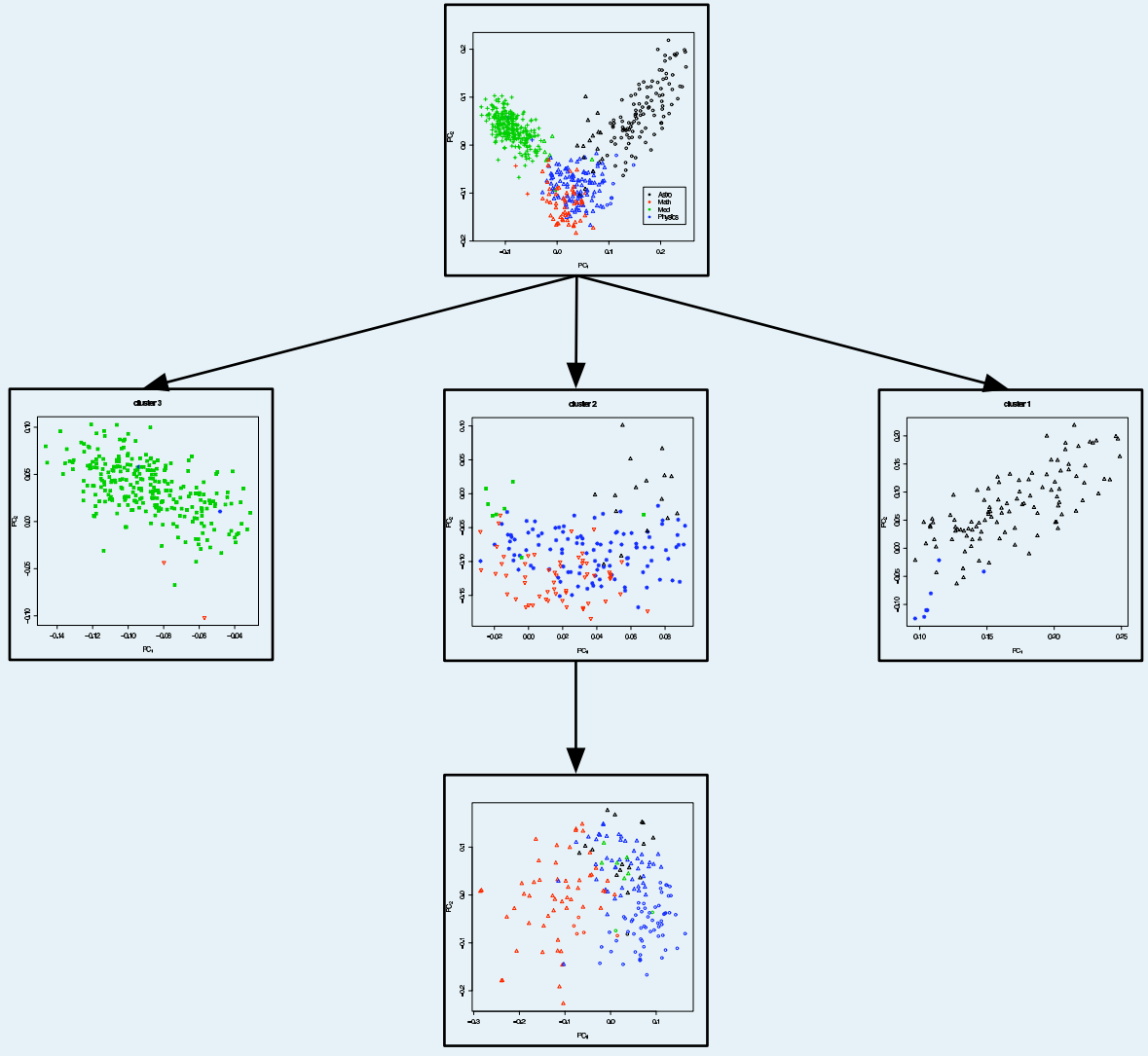
Both the features themselves and the number of features $d_{\mathcal{L}}(C)$ depend on the corpus C. Thus $\mathcal{L}_C(C)$ is a $|C| \times d_{\mathcal{L}}(C)$ *mutual information feature matrix*. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document x in corpus C, and associated word w, the mutual information between x and w is given by

$$m_{x,w} = \log\left(\frac{f_{x,w}}{\sum_{\xi \in C} f_{\xi,w} \sum_{\omega} f_{x,\omega}}\right).$$

Here $f_{x,w} = c_{x,w}/N$ where $c_{x,w}$ is the number of times word w appears in document x and N is the total number of words in the corpus C.

cep@jhu.edu

Science News text document example I



black: Astronomy, red: Mathematics, green: Medicine, blue: Physics

cep@jhu.edu

Science News text document example II

branch	Astronomy	Math	Medicine	Physics
left branch	0	2	272	2
center branch	16	58	8	109
right branch	105	0	0	7

Initial Science News Denoising Partition

|Math|=60 ; |Physics|=118

Note: 7 Physics documents are misclassified as Astronomy;
2 Physics and 2 Math documents are misclassified as Medicine

Science News text document example III

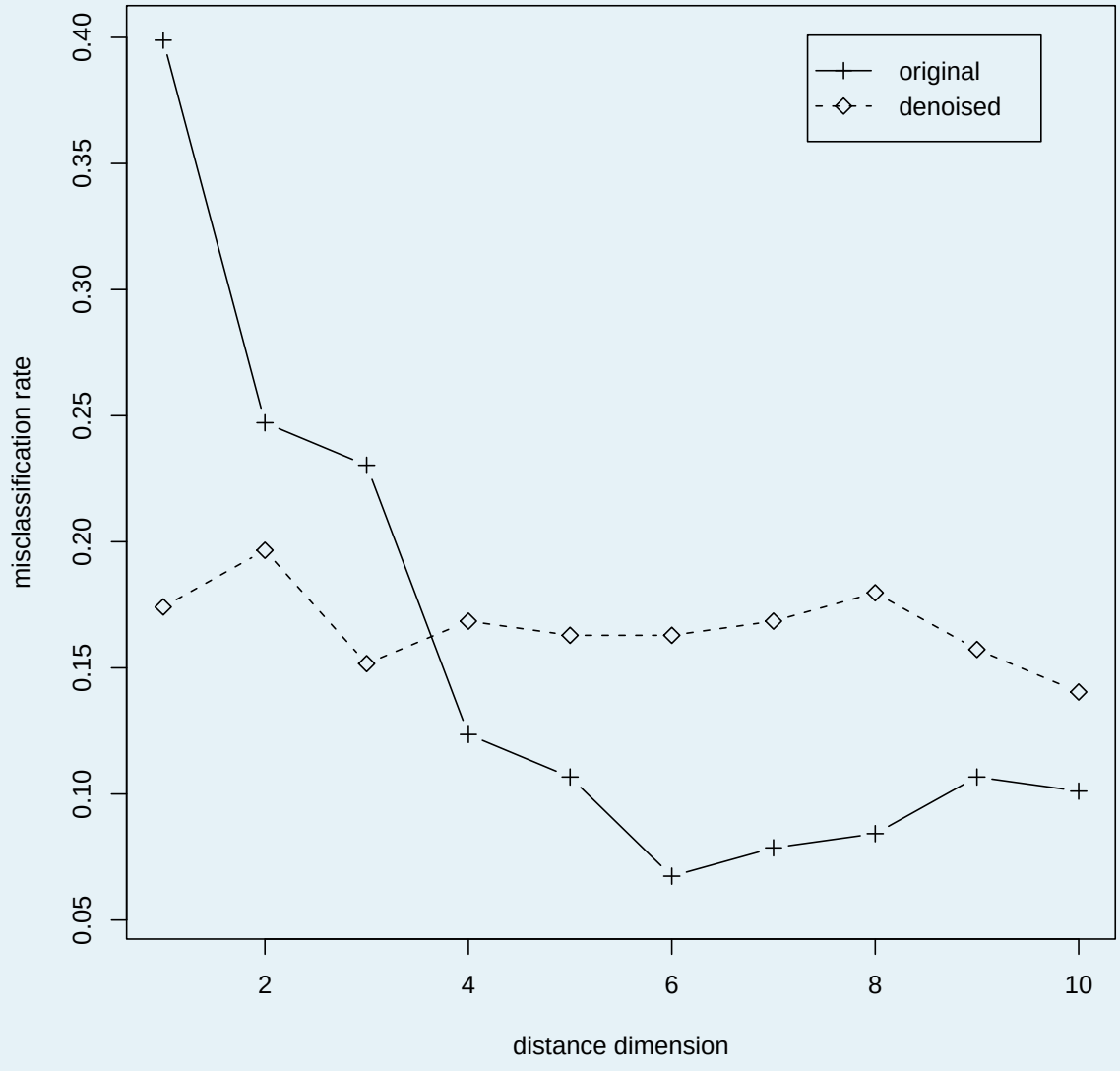
	original	denoised
misclassification of Physics as Math	42	9
misclassification of Math as Physics	35	11
sum	77	20 (+ 11)

Science News Misclassification Errors

|Math|=60 ; |Physics|=118

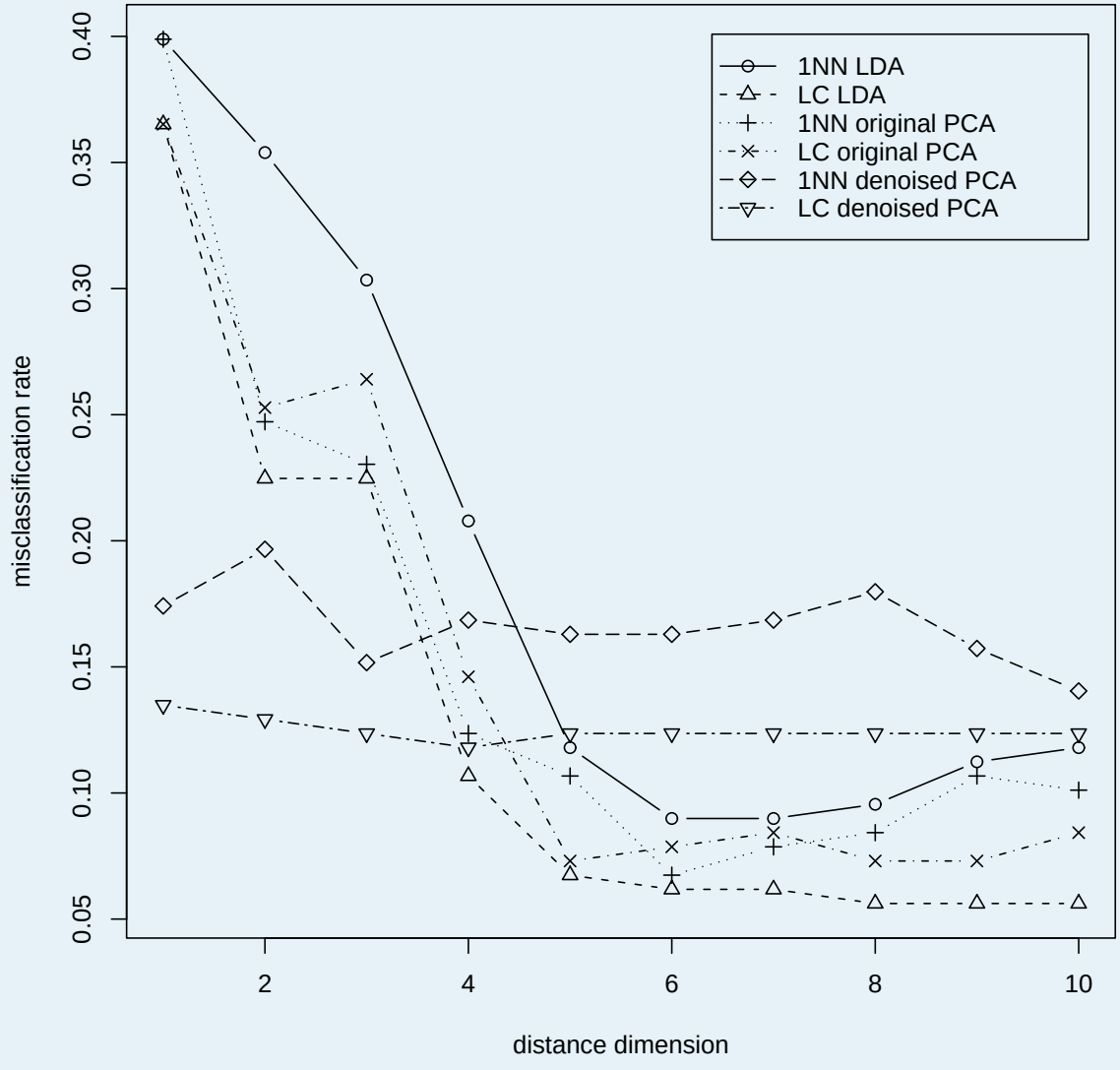
cep@jhu.edu

Science News text document example IV



cep@jhu.edu

Science News text document example IVb



cep@jhu.edu

Information-Theoretic Iterative Denoising

On the minimization of concave information functionals
for unsupervised classification via decision trees

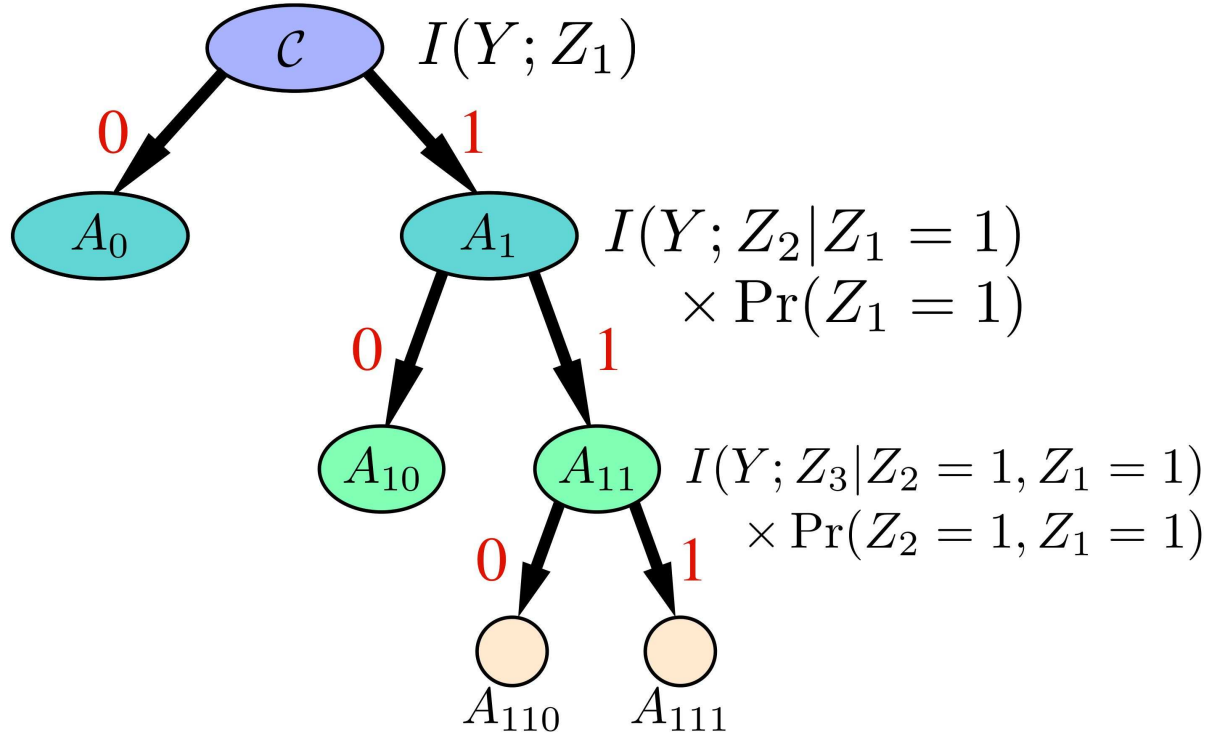
Joint work with

Damianos Karakos, Sanjeev Khudanpur, David Marchette.

cep@jhu.edu

Information-Theoretic Iterative Denoising

Total Score: $I(Y;\boldsymbol{Z}) = I(Y;Z_1) + I(Y;Z_2|Z_1) + I(Y;Z_3|Z_2,Z_1)$



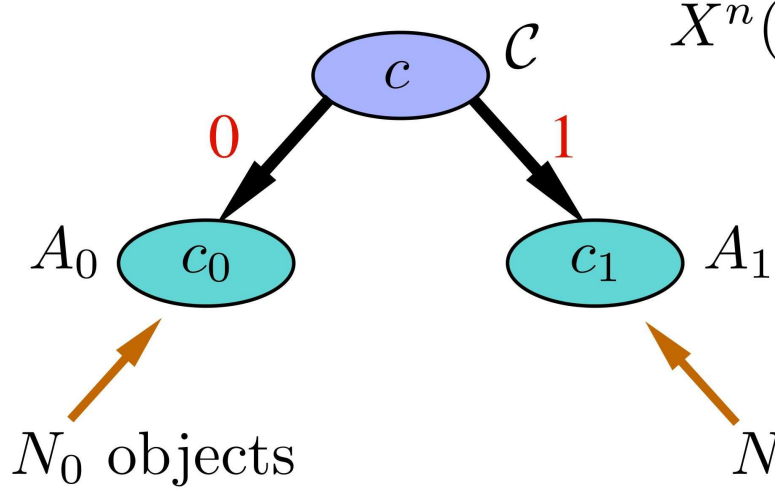
Information-Theoretic Iterative Denoising

Mutual Information as the Goodness Criterion for ISPDT Construction

N objects: $X^n(1), \ldots, X^n(N)$

Transform to $\tilde{X}^n(1), \ldots, \tilde{X}^n(N)$.

... and cluster.



Transformation & Clustering Score:

$$S(c, c_0, c_1) = \frac{N_0}{N} D(c_0||c) + \frac{N_1}{N} D(c_1||c)$$

Information-Theoretic Iterative Denoising

Optimality of Mutual Information as
the Goodness Criterion

Transformation & Clustering Score:

$$S(c, c_0, c_1) = \frac{N_0}{N} D(c_0||c) + \frac{N_1}{N} D(c_1||c)$$

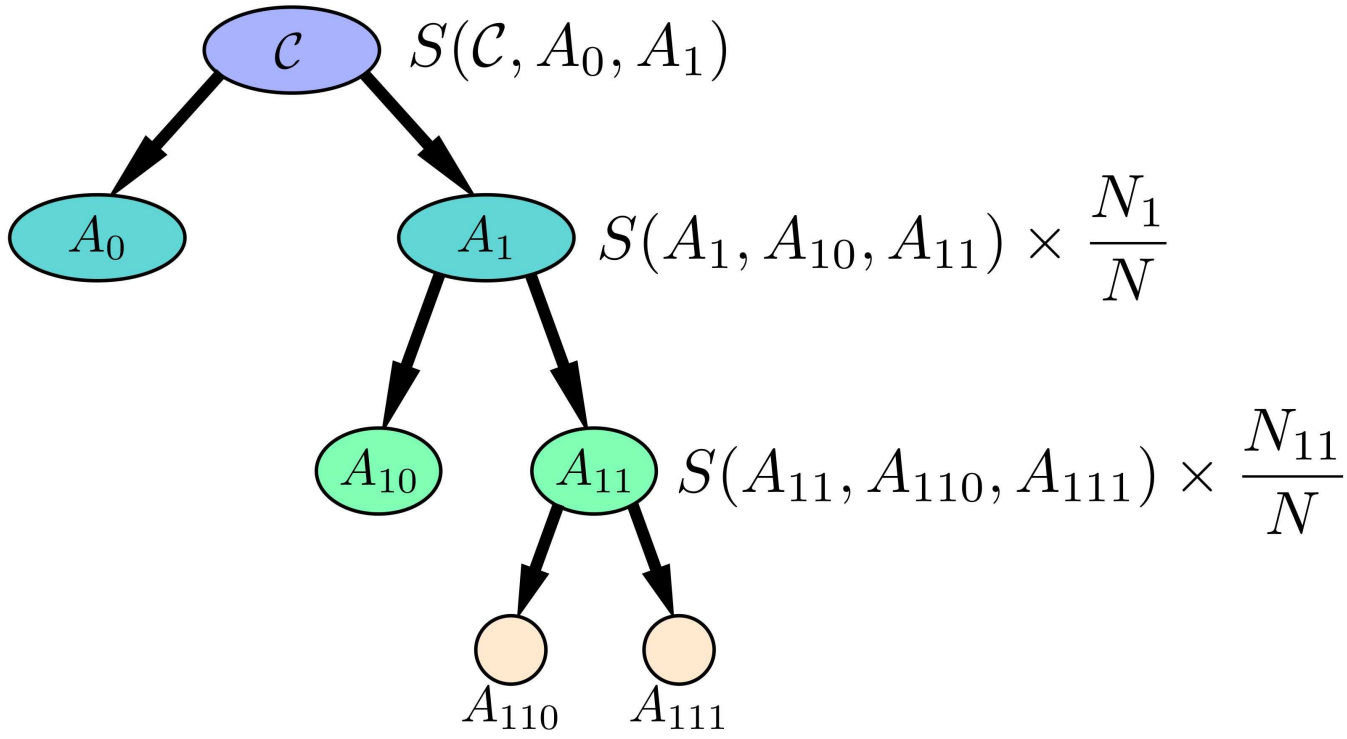
$$\stackrel{N,n\to\infty}{\longrightarrow} I(Y;Z)$$

(for stationary and ergodic data)

Information-Theoretic Iterative Denoising

Total Score:

$$S(\mathcal{C}, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N}$$

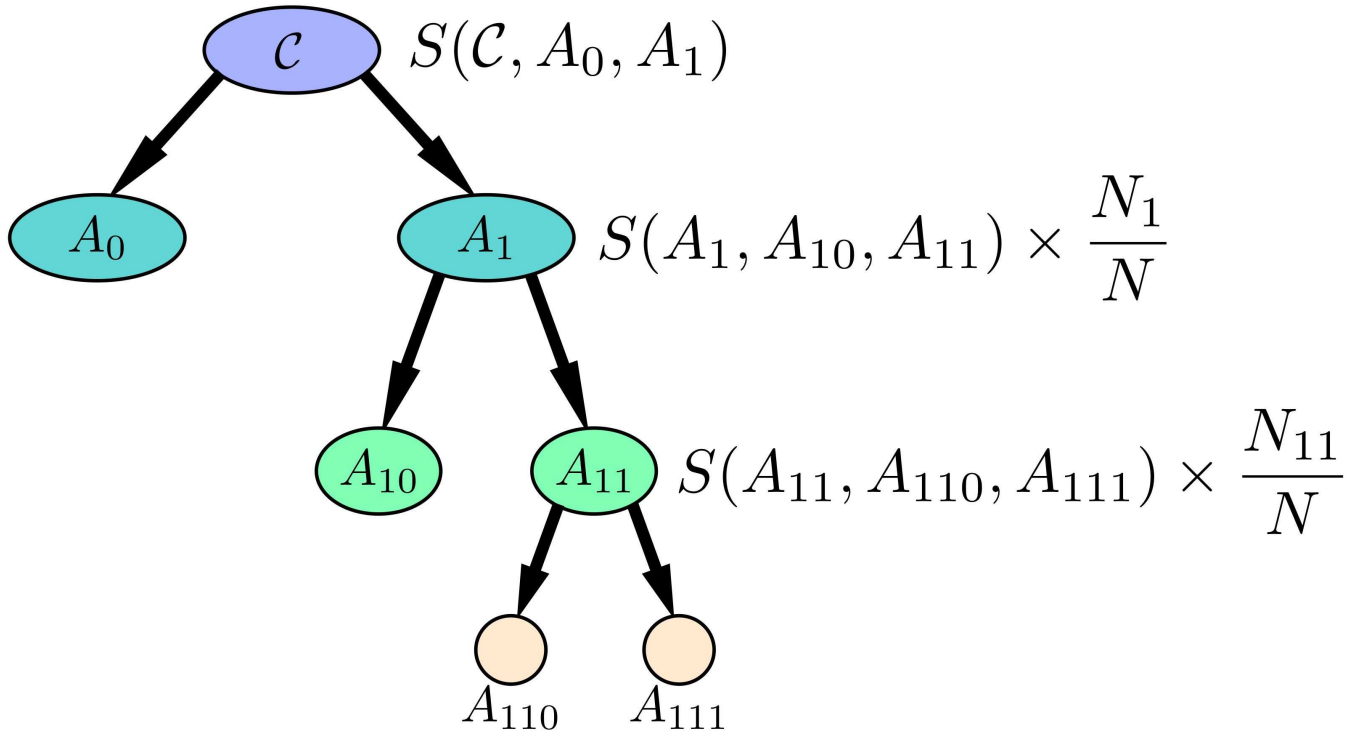


Goal: Maximize Total Score

Information-Theoretic Iterative Denoising

Total Score:

$$S(\mathcal{C}, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N}$$

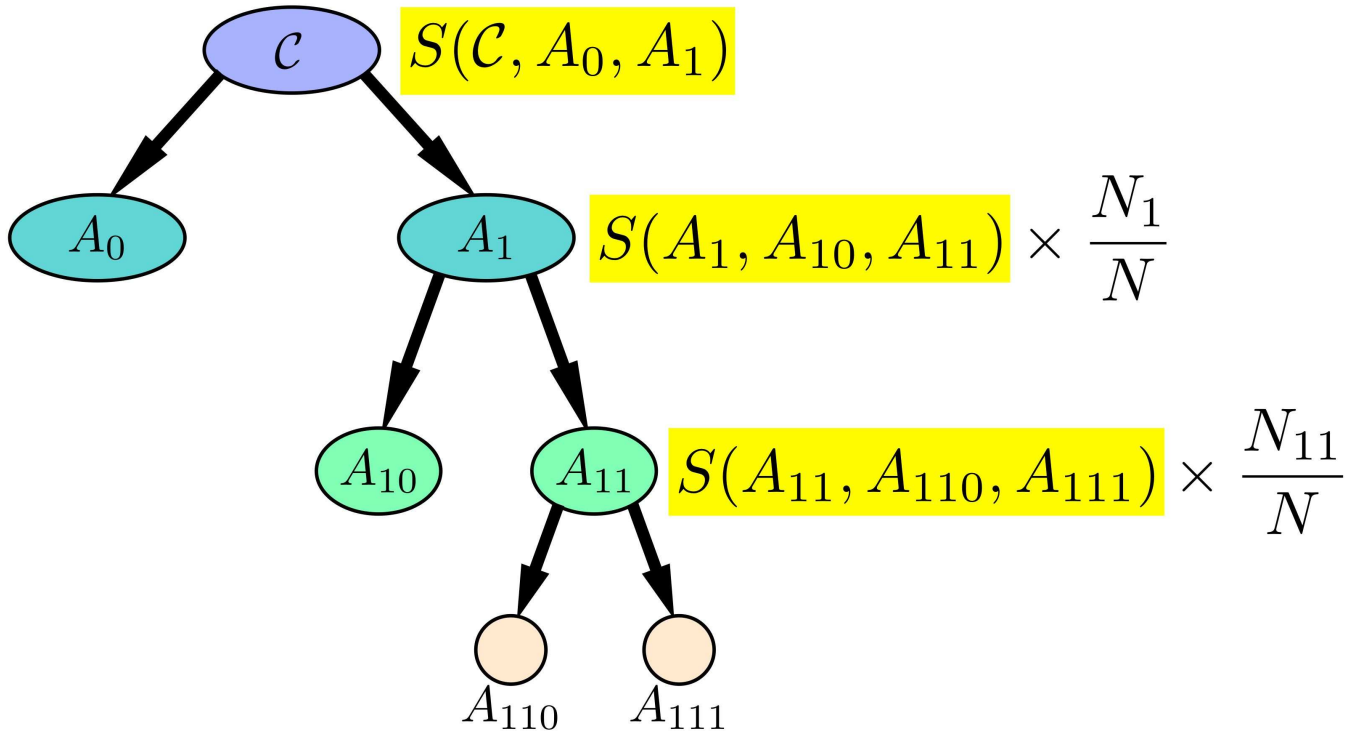


Nice Property: Iterative Computation

Information-Theoretic Iterative Denoising

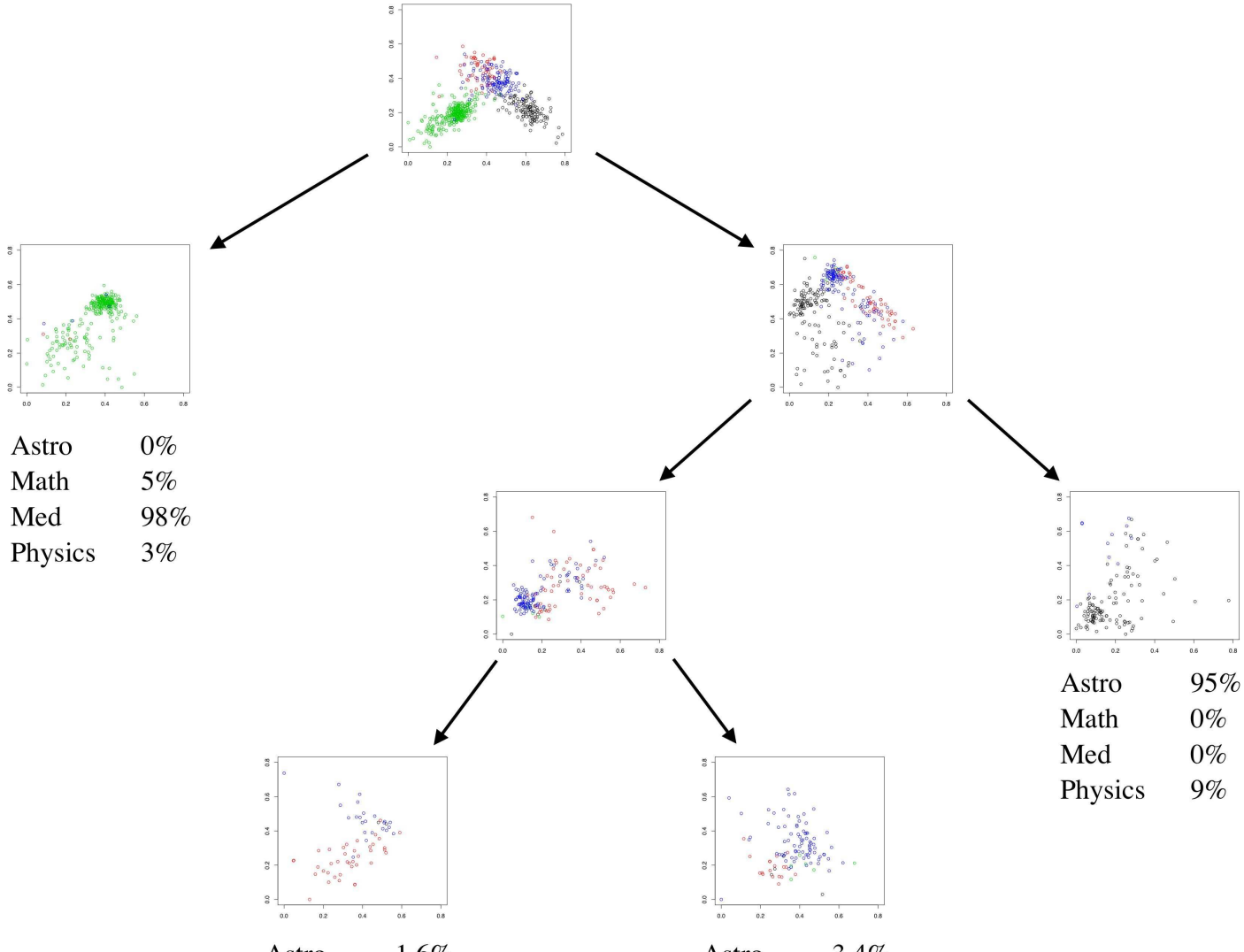
Total Score:

$$S(\mathcal{C}, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N}$$



Score depends only on local features: CDFE

Information-Theoretic Iterative Denoising Example



Astro	0%
Math	5%
Med	98%
Physics	3%

Astro	95%
Math	0%
Med	0%
Physics	9%

Astro	1.6%
Math	65%
Med	0%
Physics	21%

Astro	3.4%
Math	30%
Med	2%
Physics	67%

In the first example (hyperspectral imaging),
the gain from cdfe was (simply) due to dimension reduction.

In the text document examples,
new corpus-dependent features were indeed extracted.

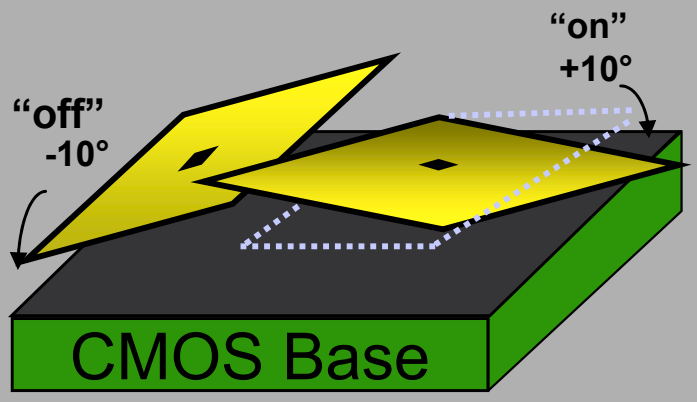
But there was no *re-sensing* involved.

An adaptive sensor example?

Real-time tunable hyperspectral imagers!

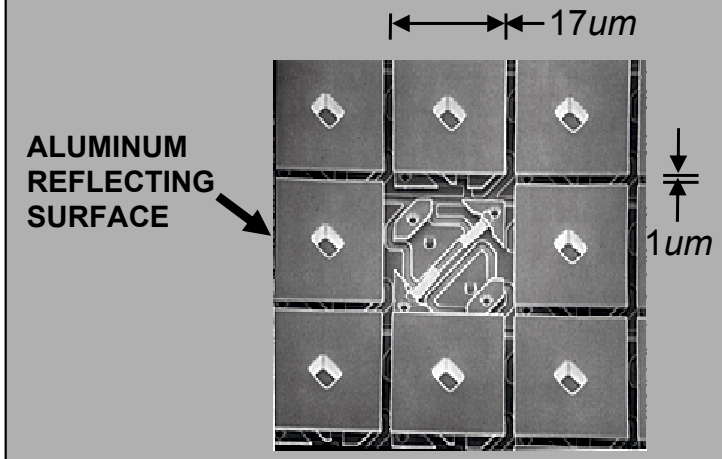
cep@jhu.edu

Architecture of the Digital Micro-Mirror Array (DMA) used as a switching device for combining spatio-spectral features



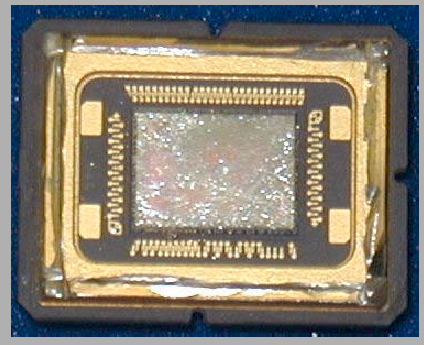
Mirrors rotate along their diagonal axis by exactly ± 10°, which is what makes the DMA a digital devise

- **0 = no light reflected (-10°)**
- **1 = all light reflected (+10°)**

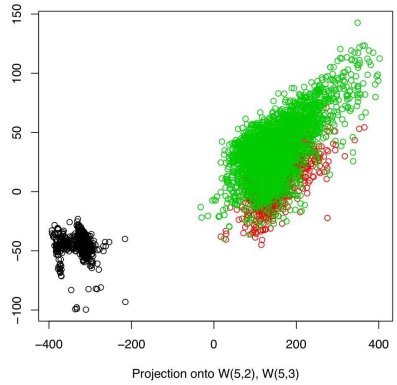


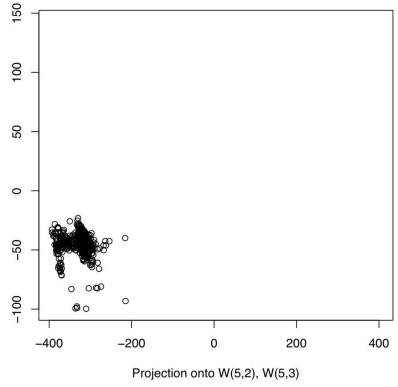
Mirrors are built on top of integrated circuits and assembled in a grid-like formation

848 X 600 DMA

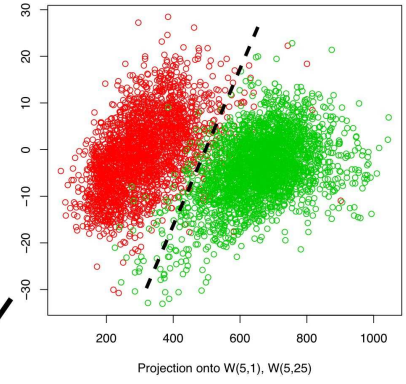


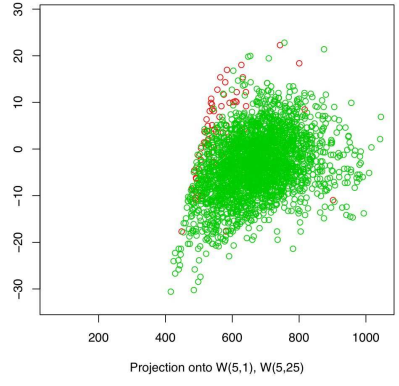
A complete DMA contains 848 columns and 600 rows of mirrors and measures 10.2 mm x 13.6 mm. Here, a DMA is shown with its glass cover removed.



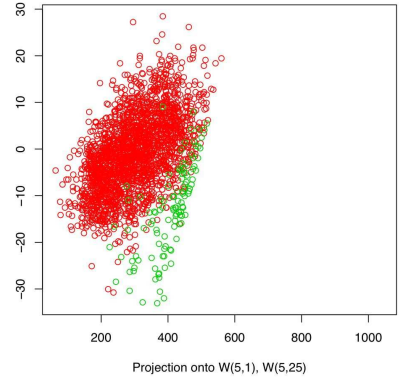


Runway: 100%

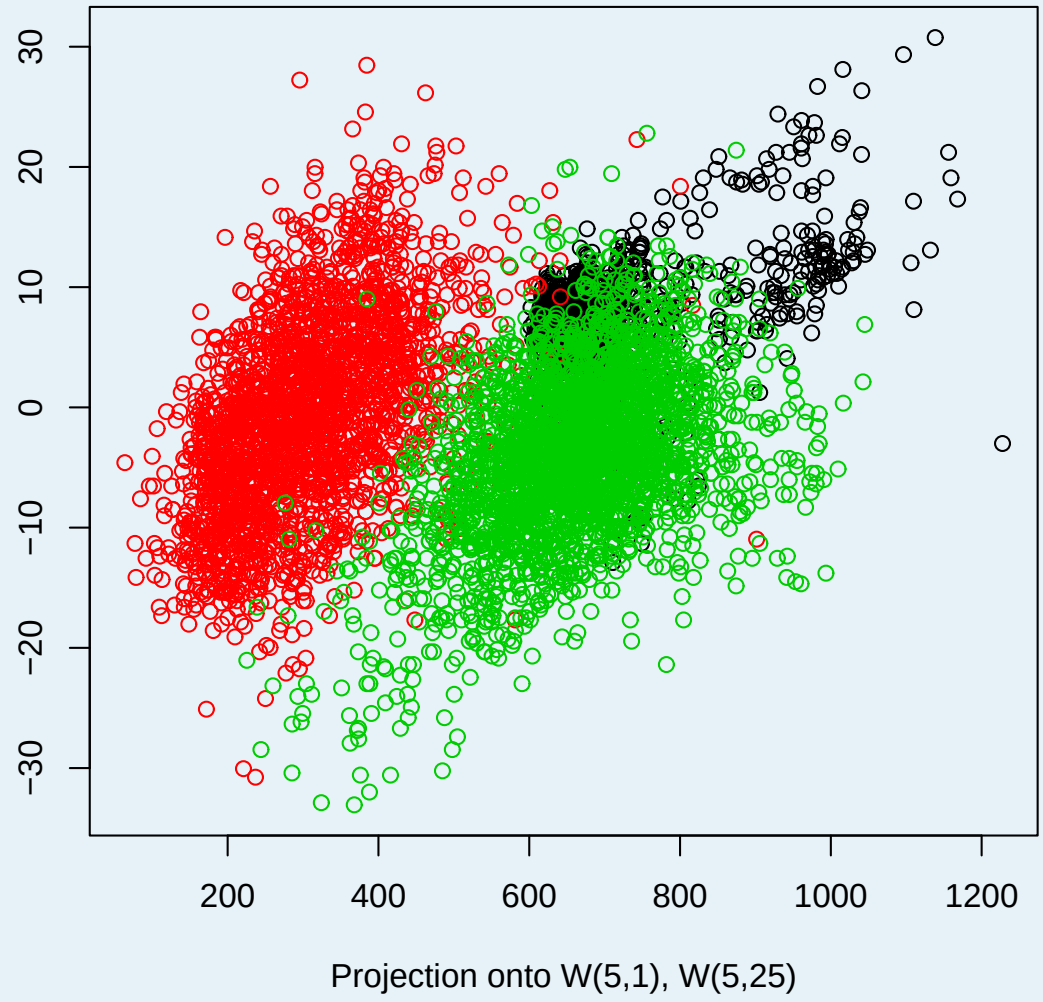




Pine: 3%
Scrub: 96%



Pine: 97%
Scrub: 4%


30
20
10
0
−10
−20
−30
200
400
600
800
1000
1200
Projection onto W(5,1), W(5,25)

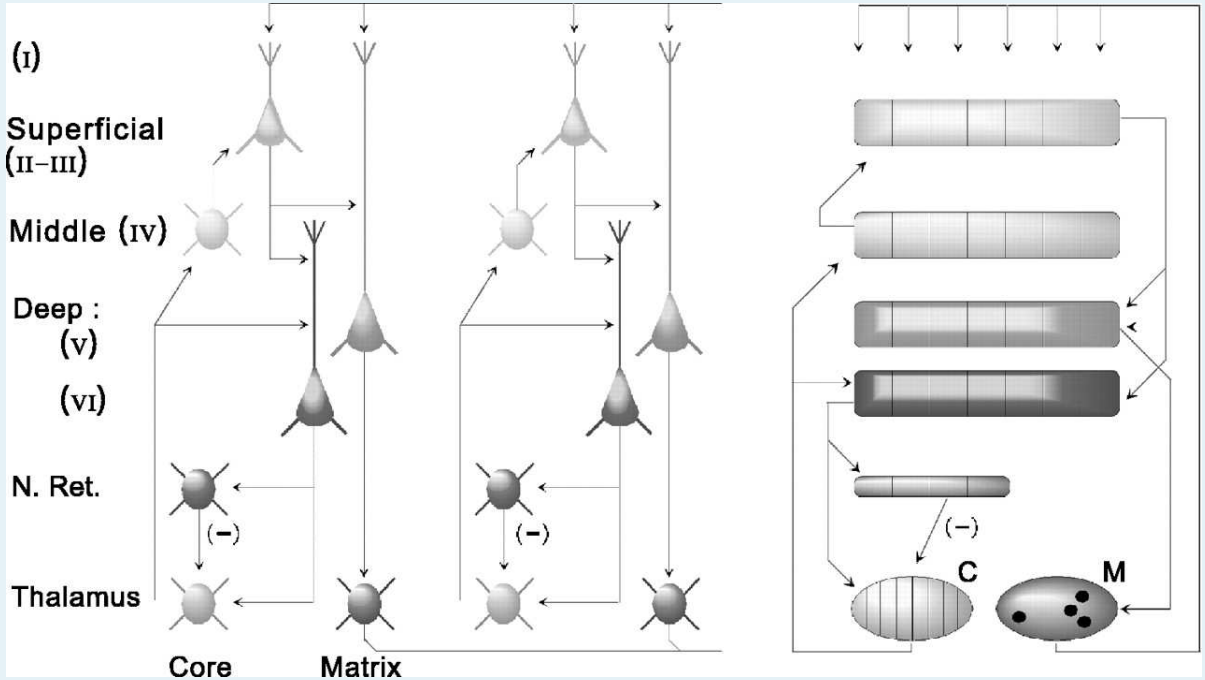
< > - +

a thalamocortical iterative denoising tree?

Rodriguez, A., Whitson, J., Granger, R. (2004).
Derivation and analysis of basic computational operations of thalamocortical circuits.
Journal of Cognitive Neuroscience 16:5, pp. 856-877.

Prof. Richard Granger
Director: Brain Engineering Laboratory
Computer Science and Cognitive Science
University of California, Irvine

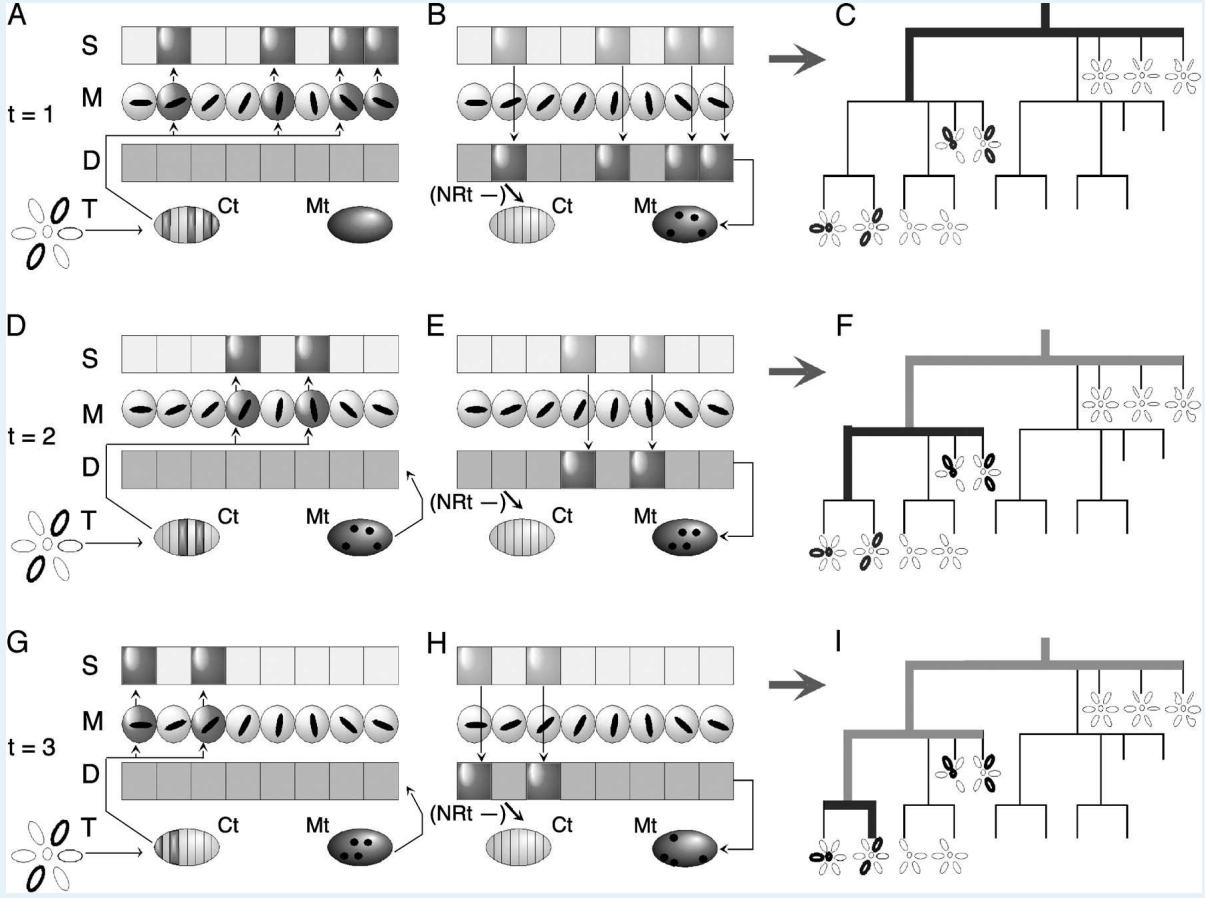
RWG 2004, Figure 3



Key features of anatomical organization of thalamocortical circuits to be modeled.

cep@jhu.edu

RWG 2004, Figure 6



Sequential operations at three successive time steps (t = 1, 2, 3) in middle (M), superficial (S), and deep (D) cortical layers, core and matrix thalamic projections (Ct & Mt), and inhibitory nucleus reticularis thalami (NRt-), in response to a static input.

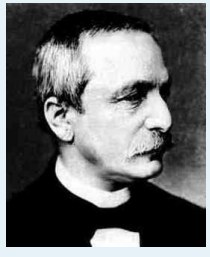
Kronecker Quote

"The wealth of your practical experience

with sane and interesting problems

will give to mathematics

a new direction and a new impetus."



– Leopold Kronecker to Hermann von Helmholtz –