



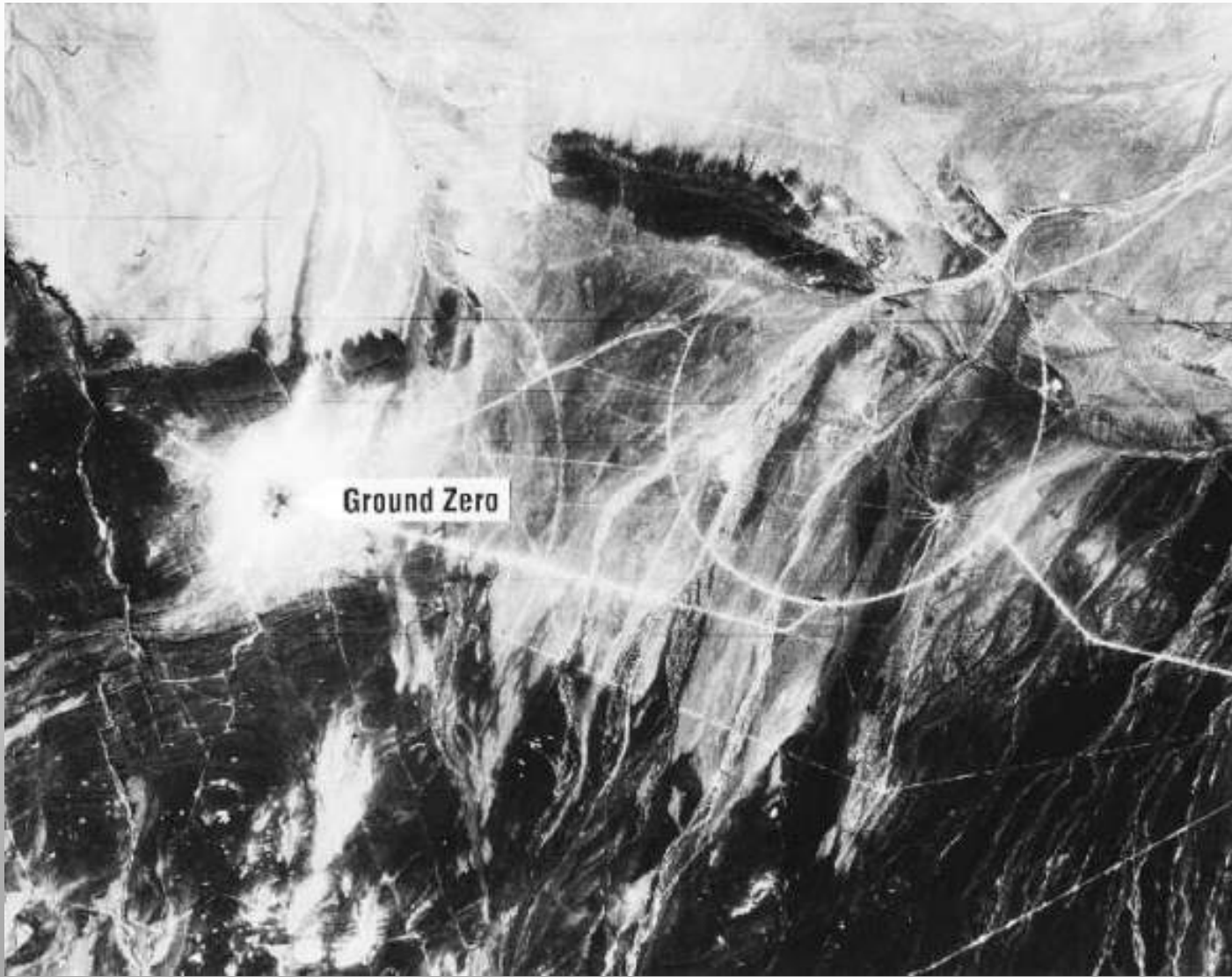
Technologies and Tools to Search Images with Images

Ulysses. J. Balis, MD

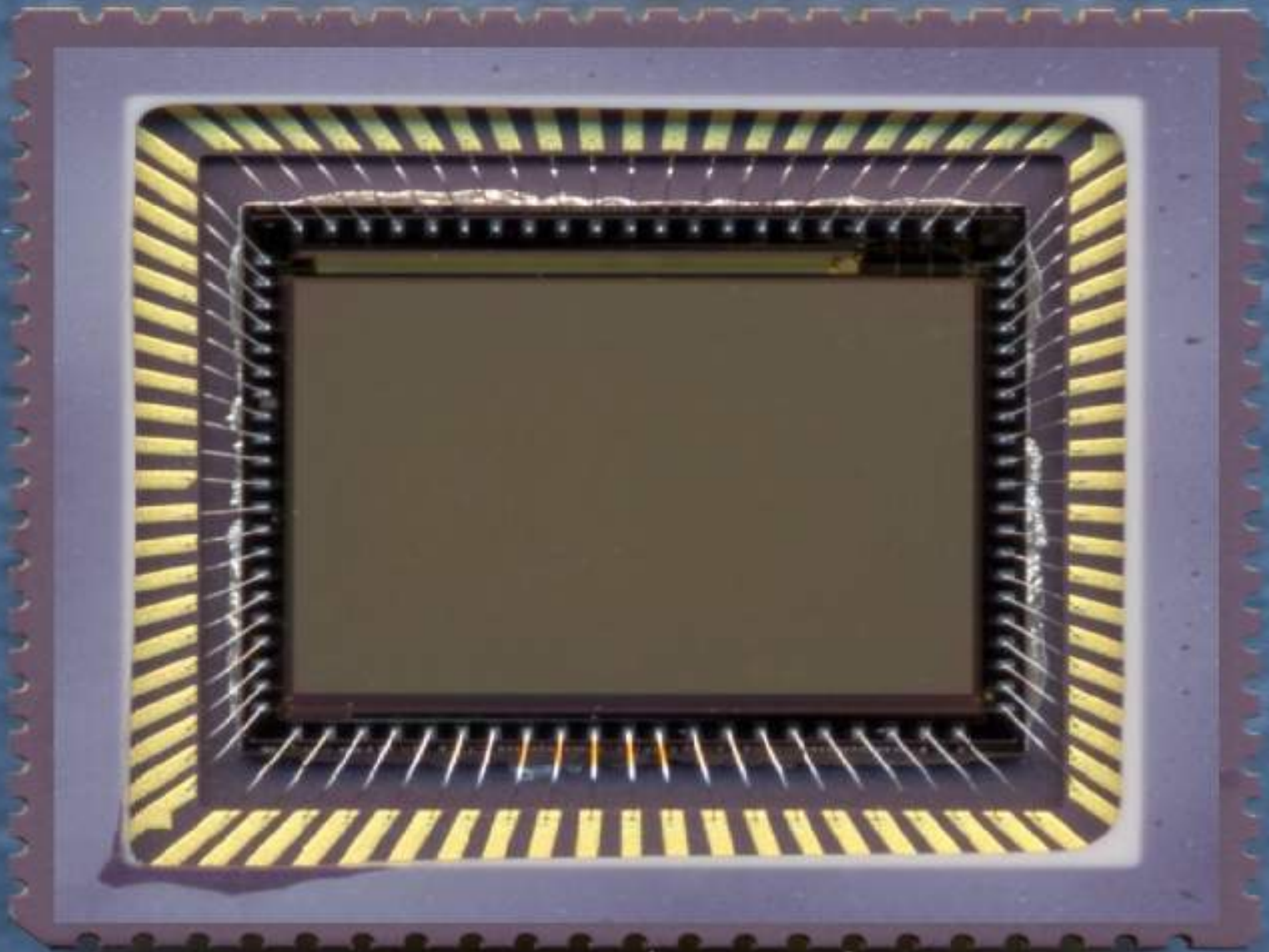
Director of Clinical Informatics
Co-Director, Division of Informatics
Department of Pathology
University of Michigan Health System

ulysses@umich.edu



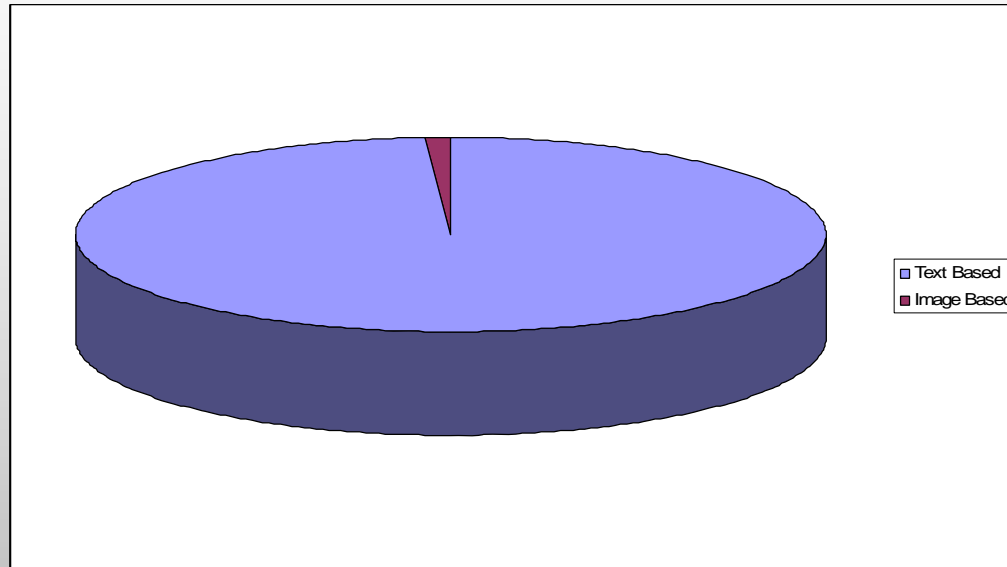


Lop Nor



The CCD – the fundamental transformative technology enabling creation of wide-field datasets

Anticipated Evolution of Data Content of Typical APLIS Systems



Compelling Use Cases for Image Query

- Diagnostic decision support
- Longitudinal evaluation
- Differential diagnosis generation
- Detection of rare events
- Teaching
- Discovery

LETTERS

A network-based analysis of systemic inflammation in humans

Steve E. Calvano^{1*}, Wenzhong Xiao^{2*}, Daniel R. Richards³, Ramon M. Felciano³, Henry V. Baker^{4,5}, Raymond J. Cho³, Richard O. Chen³, Bernard H. Brownstein⁶, J. Perren Cobb⁶, S. Kevin Tschoeke⁵, Carol Miller-Graziano⁷, Lyle L. Moldawer⁵, Michael N. Mindrinos², Ronald W. Davis², Ronald G. Tompkins⁸, Stephen F. Lowry¹ & the Inflammation and Host Response to Injury Large Scale Collaborative Research Program[†]

Oligonucleotide and complementary DNA microarrays are being used to subclassify histologically similar tumours, monitor disease progress, and individualize treatment regimens^{1–5}. However, extracting new biological insight from high-throughput genomic studies of human diseases is a challenge, limited by difficulties in recognizing and evaluating relevant biological processes from huge quantities of experimental data. Here we present a structured network knowledge-base approach to analyse genome-wide transcriptional responses in the context of known functional interrelationships among proteins, small molecules and phenotypes. This approach was used to analyse changes in blood leukocyte gene expression patterns in human subjects receiving an inflammatory stimulus (bacterial endotoxin). We explore the known genome-wide interaction network to identify significant functional modules perturbed in response to this stimulus. Our analysis reveals that the human blood leukocyte response to

endotoxin activates innate immune responses and presents with physiological responses of brief duration¹⁰. Notably, there is an initial proinflammatory phase and a subsequent counterregulatory phase, with resolution of virtually all clinical perturbations within 24 h.

K-means cluster and principal component analyses were first used to visualize the overall response to endotoxin administration. Figure 1a reveals probe sets clustered by *K*-mean analysis, where each bin has a distinct endotoxin-induced temporal pattern. The signal intensity of 5,093 probe sets—representing 3,714 unique genes—out of a total of >44,000 probe sets changed significantly in response to endotoxin, whereas no significant changes were observed in control subjects (estimated false discovery rate <0.1%). Of the 5,093 probe sets identified, over half showed reduced abundance at 2, 4, 6 and 9 h, returning to baseline by 24 h (see bins 0–4). In contrast, a smaller number of probe sets were induced by 2 h (bins 5, 6), and the remaining probe sets showed a delayed response, peaking at 4–9 h but

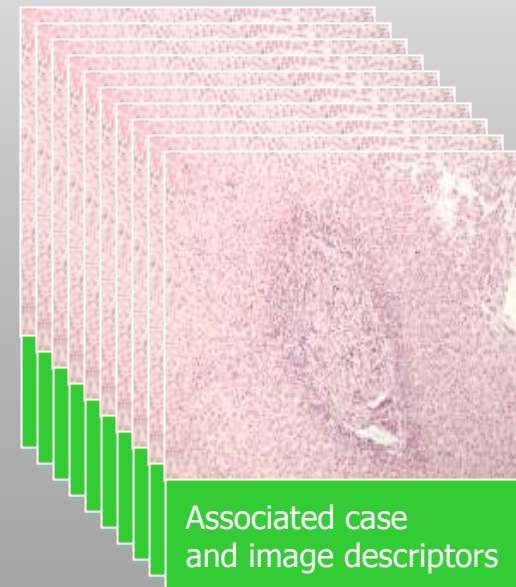
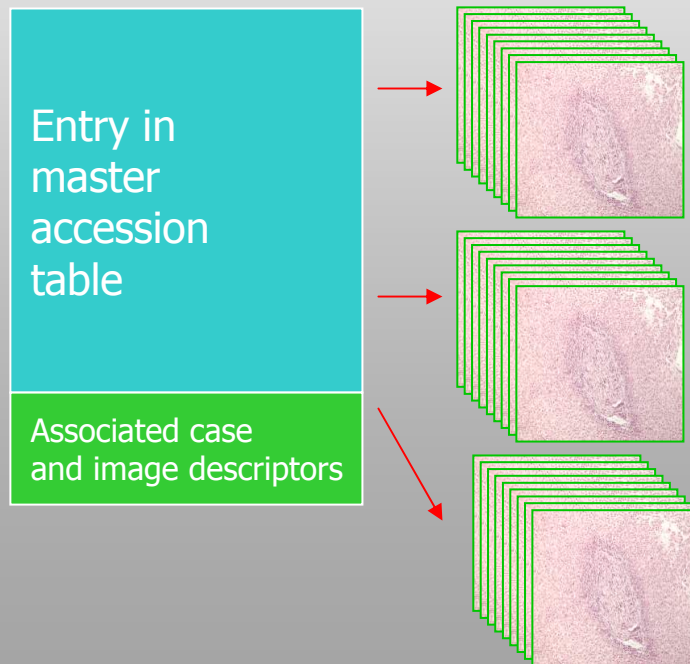
Current World View of Pathology Imagery Repositories

- Model 1: Relational Database

- Image Metadata associated with case-level data
- Entire Schema required to carry out discovery
- Text-based
- Image data is a passive component of the query

- Model 2: Metadata-tagged Images

- Image Metadata associated with each image
- Image becomes a self-contained dataset available for discovery
- Text-based
- Image data is a passive component of the query

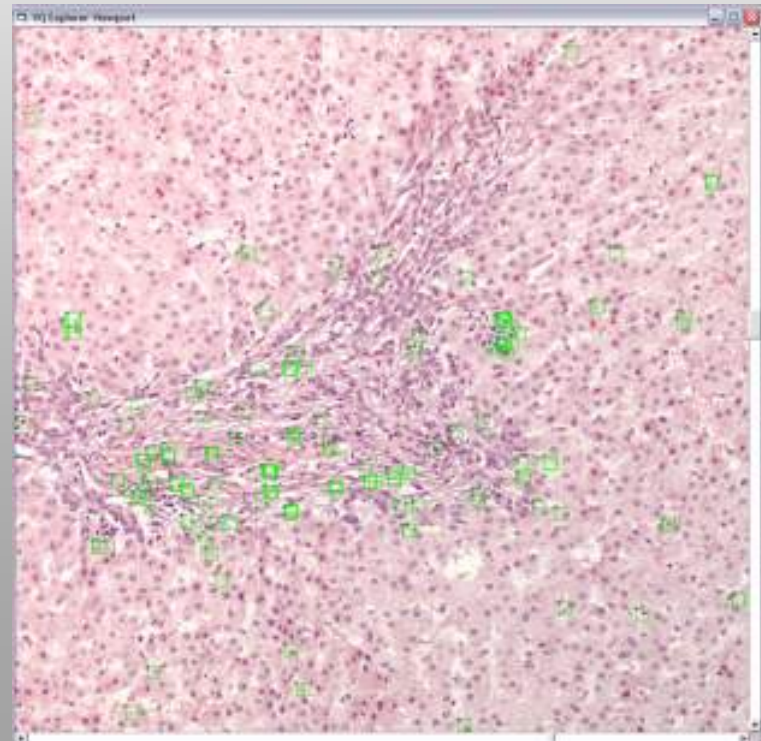


Highly Desirable World View of Pathology Imagery Repositories

(Future State)

- Model 3: Metadata-tagged surface map
 - Image Metadata exists at the image level and is spatially coupled to underlying digital imagery
 - Discovery can be carried out on the image-space itself, with retrieved metadata classifiers available for generating search result sets (e.g. differential diagnosis generation)
 - Image-based
- Model 4: Surface discovery
 - Non-metadata-associated digital imagery is spatially probed for statistical convergence with an image-based query set
 - Imagery becomes a self-contained dataset available for discovery
 - Image-based

?  Region-of-interest based predicate ∈



Synthesis of Disparate Vectorized Data sets

- Increased size of global composite vectors
- Added analysis complexity
- Enhanced opportunity for discovery
- No commercial software
- Paucity of synthetic algorithms
- Few domain-specific publications



On the prospect of
analyzing 1000's of
Gigabytes of data in
real-time...

*“...the difference between
myself and a madman is
that, quite obviously, I am
not mad...”*

-Salvador Dali

Some Observations Concerning Slide data Density



- Characteristics:
 - ~2.5 by ~7.5 cm
 - 1/3 used for label
 - 2.5 x 5.0 cm for tissue display
 - Typical light microscopy is diffraction-limited to 0.25 microns
 - Yields an effective required pixel count of 100K by 200k pixels (2.3 Gb) or a 20k MPixel Image
 - This is the same things as saying that one would need to capture 20,000 images with a 1 MPixel camera to obtain a single slide
 - Herein lies the essence of why telepathology has been so long in approaching an operational reality.

$(1000 \times 25) / 0.25 \text{ microns} = 100,000 \text{ linear pixels}$

$(1000 \times 50) / 0.25 \text{ microns} = 200,000 \text{ linear pixels}$

This is a 20 GPixel image

vs. a relatively insignificant
4 MPixel Image



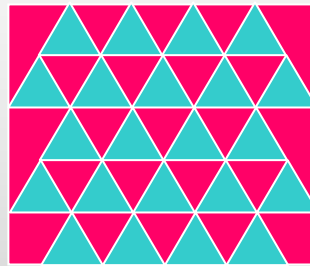
Project Objectives

- Develop a self-training, domain independent image segmentation / classification tool.
- Utilize this tool to create two novel image search modalities:
 - Region of interest Query by example (image space search; not text based)
 - Retrieve diagnostic information associated with prior classified fields, enabling the generation of dynamically generated differential diagnosis
- Explore the stochastics of multi-dimensional image space data as it applies to other emerging massively parallel data collection approaches (genomics, proteomics, etc.)
 - i.e. *Morphogenomics*

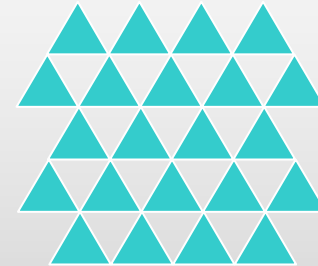
Vector Quantization



Original Image

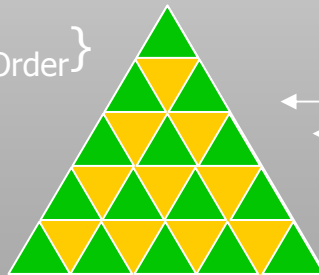


Division of
image into local
domains

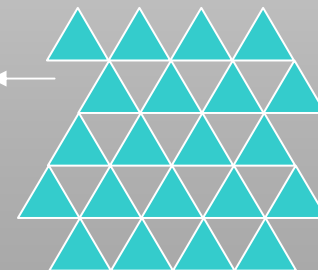


Extraction of
Local Domain
Composite
Vectors

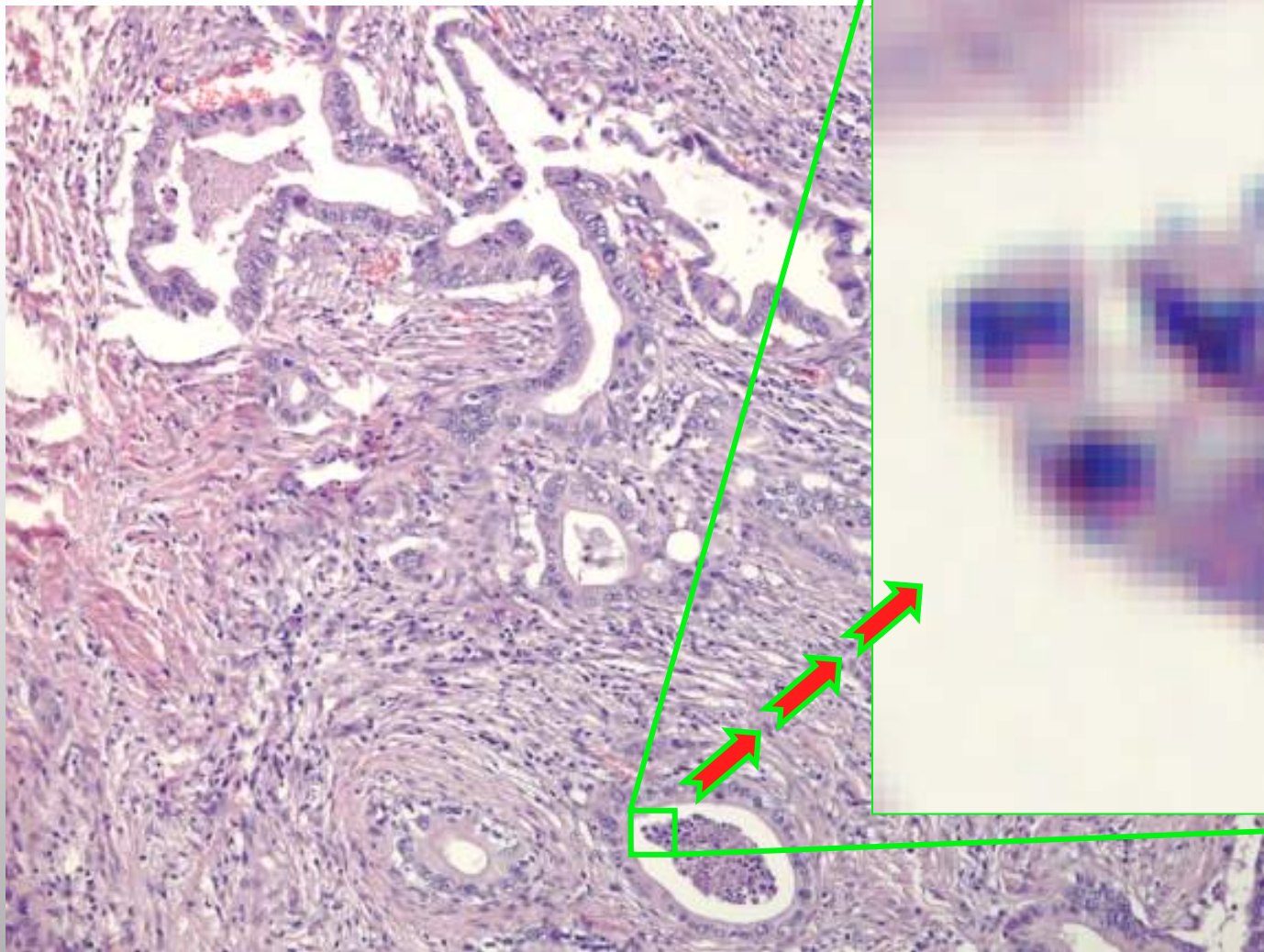
$$V^K = \sum \{ [L \cdot x_0 y_0]_{\text{Order } r} \dots [L \cdot x_n y_m]_{\text{Order } s} \}$$

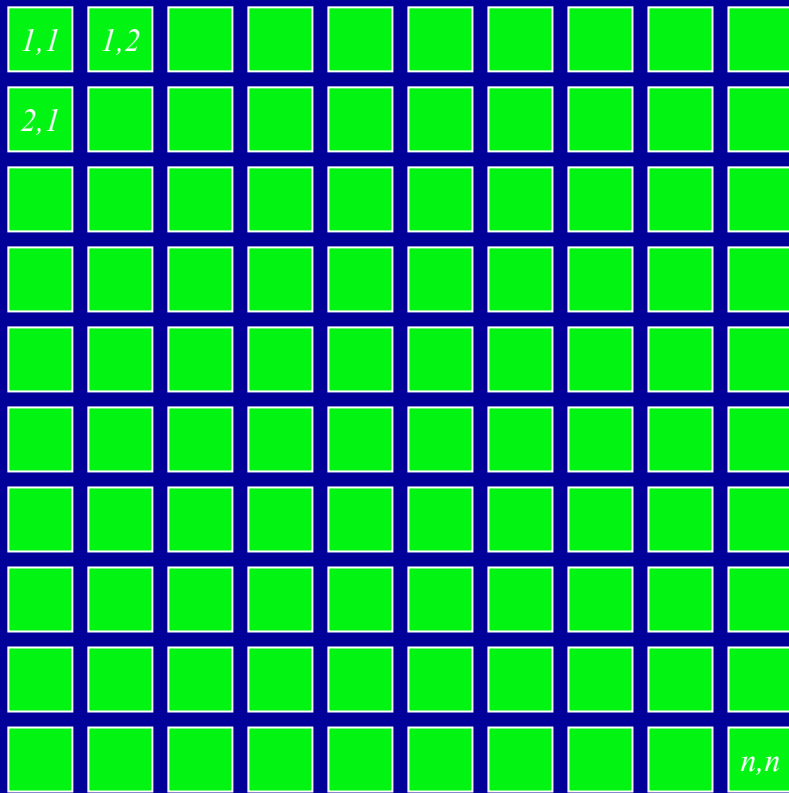


Vectorization of
each local kernel

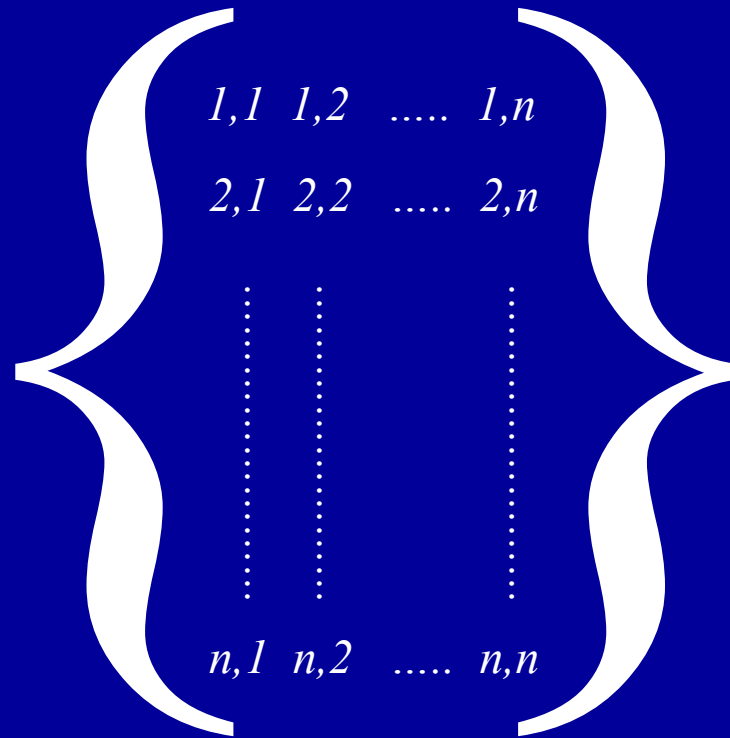
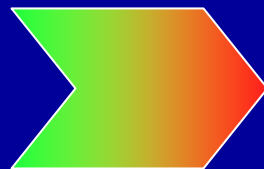


Individual
assessment of each
composite vector





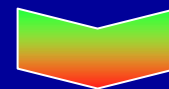
Initial n by n sub-region of image



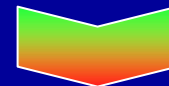
*Resultant Input Vector
Kernel of $n \bullet n \bullet 3$
dimensionality*

For every location 

Each location is an RGB triplet; hence, each vector component is itself a triplet sub-vector.



Galois Field Transform



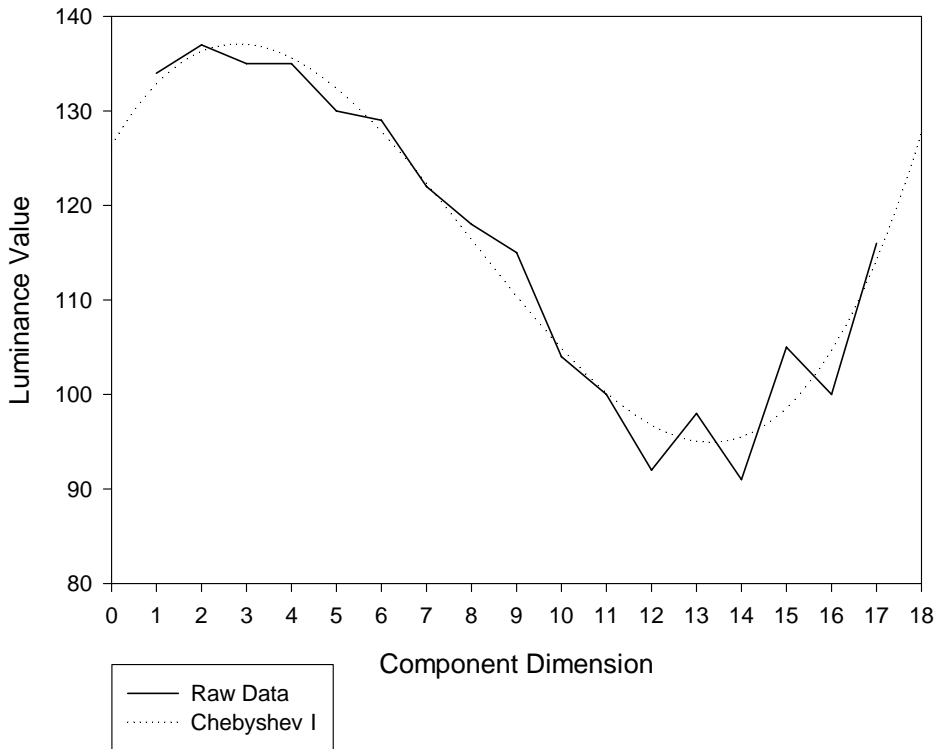
Canonical V.Q. Tensor

What about higher order data, which may also constitute complete vector sets?

- Multi-planar (cytology)
- Synthetic data sets
 - Image-genome
 - Image-proteome
 - Image-physiome, etc.
- Hyperspectral

From a vector analysis perspective, added vectors simply add robustness to a system, independent of their phenomenological derivation.

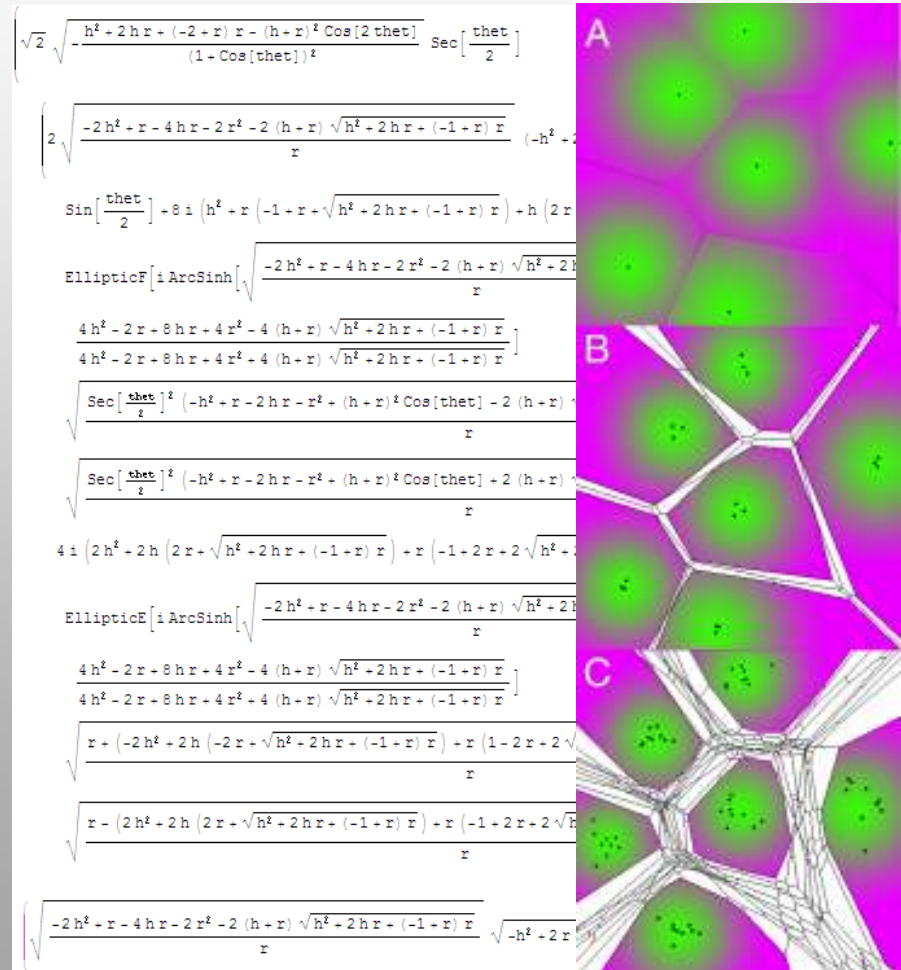
Polynomial Model Stringency



- Polynomial Model Considerations
 - Vector data need not be exactly like source data
 - Provides for concurrent compression and opportunity to search in a greatly reduced search space.
 - Very useful for hyperspectral imaging search
 - Minimal exploration in the life sciences and specifically, histopathology

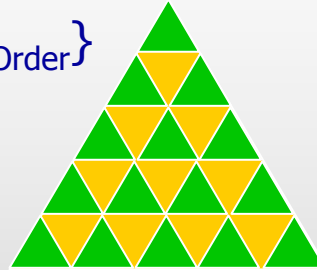
Typical Galois Field mapped to the even Jacobian/Chebyshev tensor polynomials manifested on the edge of the complexity transition

- On Galois Fields...
 - Not merely a clustering algorithm
 - The resulting field is a non-linear N-space manifold selected for its distinctiveness from all other modular functions in the Galois set space
 - Fields may have local minima and local extrema
 - Any Galois manifold is exclusive of any other Galois set
 - Non-trivial to calculate; trivial to query

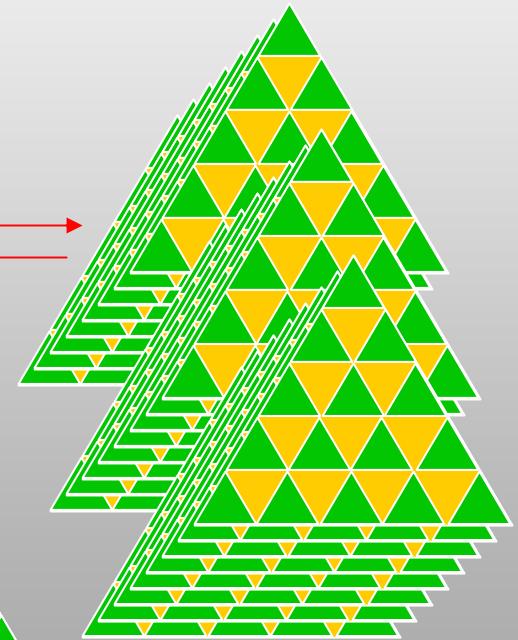


Vector Quantization

$$V^K = \Sigma \{ [L \cdot x_0 y_0]_{\text{Order}}, \dots, [L \cdot x_n y_m]_{\text{Order}} \}$$



Established Vocabulary



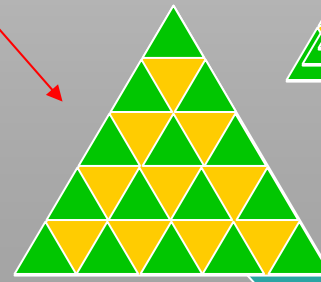
Query Against library (Vocabulary) of established Galois Vectors

Previously Identified Vector

Novel Vector

Assembly of compressed dataset

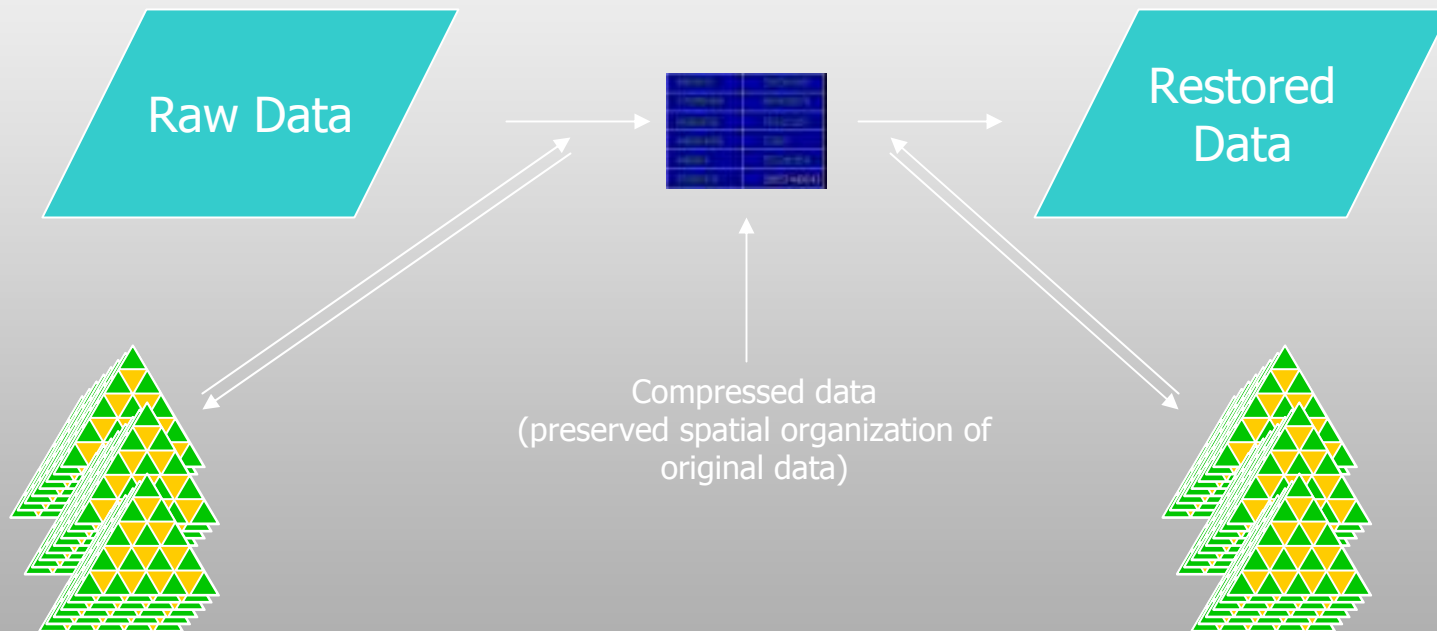
Assignment of a unique serial number and inclusion into global vocabulary



38857448643

| | |
|-----------|-------------|
| 8865433 | 354554343 |
| 776956468 | 865438676 |
| 66963658 | 554323267 |
| 446854456 | 53887 |
| 446854 | 553246564 |
| 55565435 | 38857448643 |

VQ - Based Image Compression



Depending on the selected compression ratio, restored loss-compression imagery may or may not be of diagnostic quality.

Information Theory pertaining to Galois Mapping Systems

- Ludwig von Boltzman
 - What is an efficient manner to model processes that have essentially infinite discrete elements (gas kinetics)?
 - $** \Rightarrow$ Boltzman distribution
 - Model many discrete elements with a continuous function
 - computationally feasible
 - conceptually palatable
 - Phenomenologically correct

The Mean-free-path problem

- In Astrophysics: What is the incidence of two stars colliding for a given tensor volumetric distribution?
- In Histology: What is the likelihood of two comparable Galois tensors sharing a common region in N-space for a given homomorphic stringency?

The Mean-free-path problem

- $\lambda = 1/(n\sigma)$ and $\rho = \lambda/v$
 - Mean free path of λ and collision interval of ρ
 - Where n is the number density, σ is the cross section and is the random velocity
 - For our galaxy, $\rho = 10^{19}$ years
 - $\sigma = \pi (2R_{\odot})^2$; $R_{\odot} = 6.96 \times 10^{10}$ cm
 - For Vector quantization of histologic data, with use of 64-dimensional vectors or higher orders, the incidence of overlap of non-homomorphic regions is greater than 1 in 256^{30} (1.766×10^{72}) which allows for unique identification of structural components.
 - When combined with multivariate Bayesian analysis, the identification profile effectively becomes a fingerprint for underlying unique histomorphic status of a region of interest.

Combining these equations, we obtain

$$1 + \frac{4\pi G}{k^2} \int \frac{\mathbf{k} \cdot \partial f_0 / \partial \mathbf{v}}{\mathbf{k} \cdot \mathbf{v} - \omega} d^3 \mathbf{v} = 0. \quad (5-27)$$

This equation is the required dispersion relation since it relates \mathbf{k} and ω .

For illustrative purposes assume that f_0 is Maxwellian,

$$f_0(\mathbf{v}) = \frac{\rho_0}{(2\pi\sigma^2)^{3/2}} e^{-\frac{1}{2}v^2/\sigma^2}, \quad (5-28)$$

where ρ_0 is the density. When f_0 is of this form, the integral over all velocities in equation (5-27) can be done in rectangular coordinates (v_x, v_y, v_z), where the v_x -axis is chosen to lie in the direction of \mathbf{k} . The integrals over v_y and v_z are simple, using $\int_{-\infty}^{\infty} \exp(-\frac{1}{2}v^2/\sigma^2) dv = \sqrt{2\pi\sigma^2}$, and equation (5-27) becomes

$$1 - \frac{2\sqrt{2\pi}G\rho_0}{k\sigma^3} \int_{-\infty}^{\infty} \frac{v_x e^{-\frac{1}{2}v_x^2/\sigma^2}}{kv_x - \omega} dv_x = 0. \quad (5-29)$$

By analogy with the fluid case, we expect the boundary between stable and unstable solutions to occur at $\omega = 0$. At $\omega = 0$ the integral in (5-29) is evaluated easily, and we have

$$k^2(\omega = 0) \equiv k_J^2 = \frac{4\pi G\rho_0}{\sigma^2}, \quad (5-30)$$

where k_J is the Jeans wavenumber for the stellar system. Thus the formula for the Jeans length of a collisionless system is the same that obtained for fluids, equation (5-22), except that the velocity dispersion σ is substituted for the sound speed v_s .

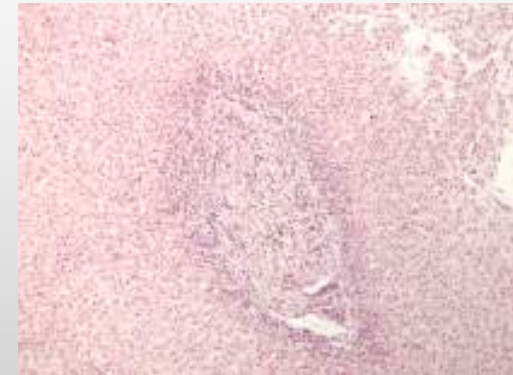
N-Space systems exhibit Maxwellian energy distributions, regardless of length-scale, making them available for modeling in reverse-discretized form.

Thus, the cluster of homomorphs created by any histologic architecture can be modeled by a family of continuous functions, simplifying computational complexity and search-space size.

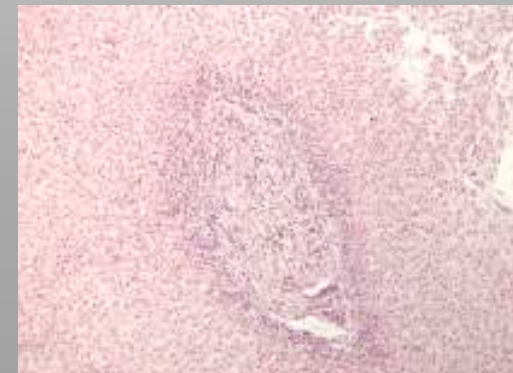
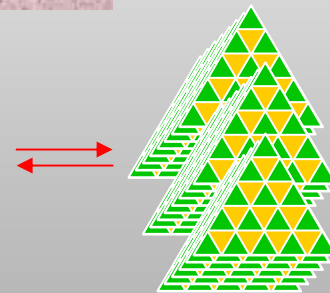
From: Galactic Dynamics, Binney J and Tremaine S. Princeton University Press, 1987

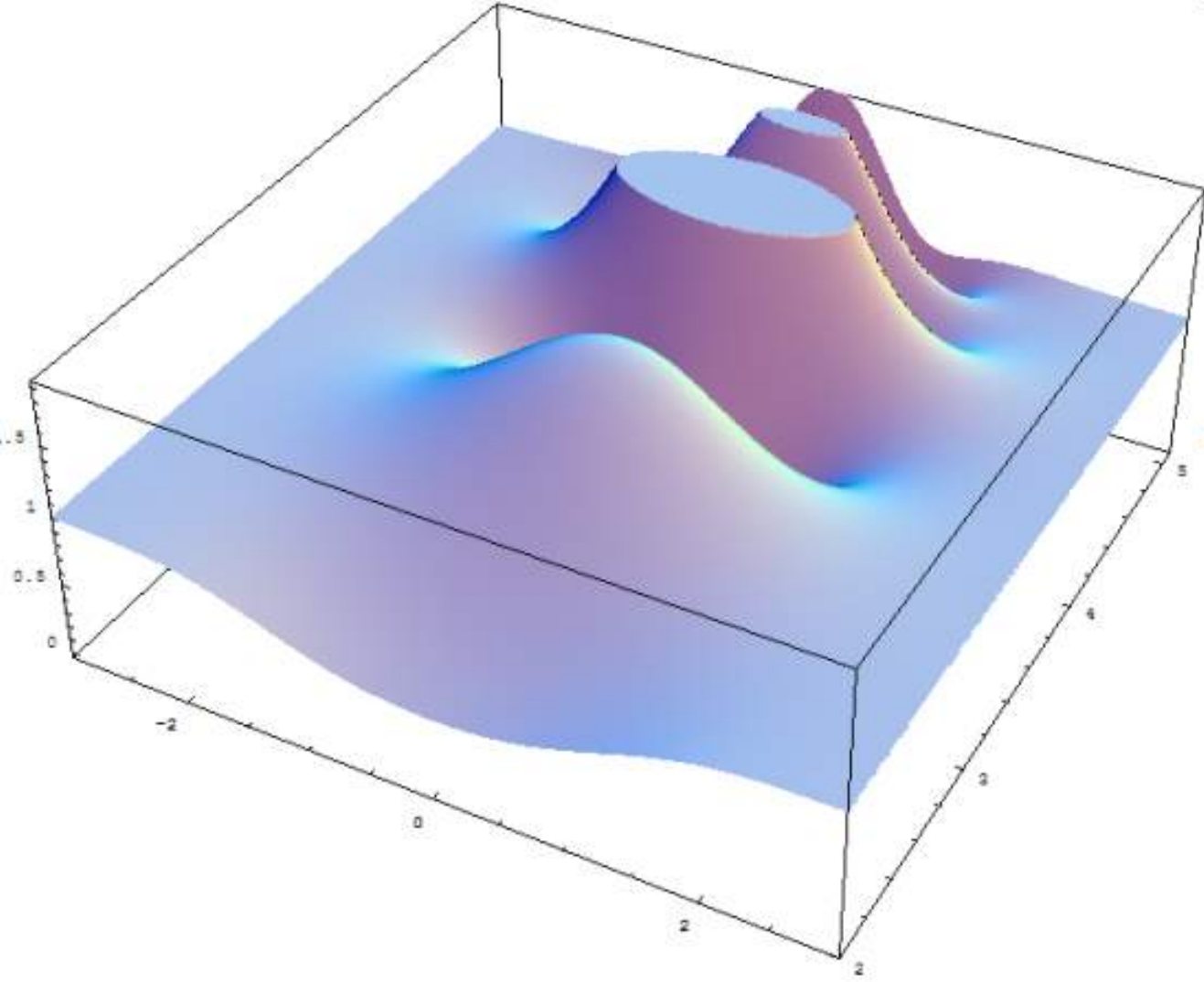
Consequences of VQ representation, in light of Maxwellian complexity

- If an image can be compressed by six log, and subsequently restored with minimal degradation of diagnostic clarity, is it not the case that the sum total of “knowledge” is similarly contained in the compressed data set as it is obviously present in the primary and restored data.
- Searches carried out upon the compressed data set represent an enormous computation opportunity for simplified query.
- As VQ vectors are structural homologs of repeating histologic elements, the query can be carried out by searching for a set of recurring vectors in the image set space, using a region-of-interest source template.



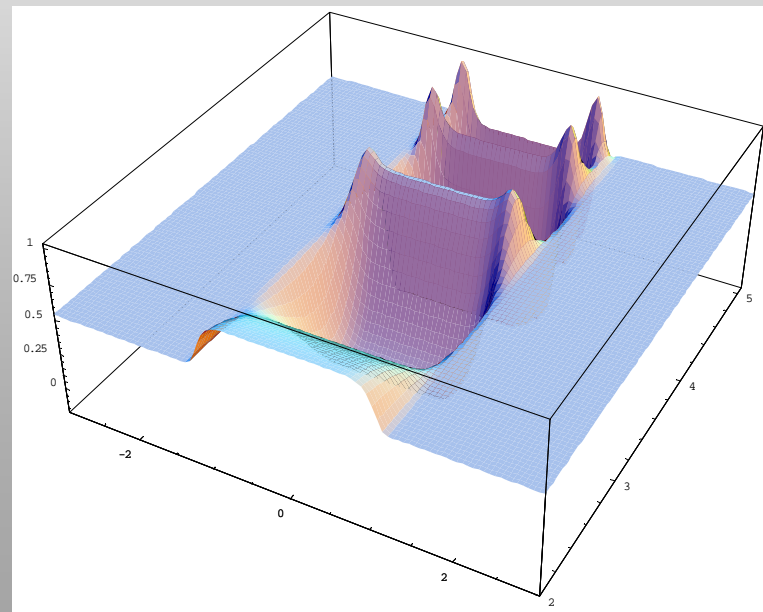
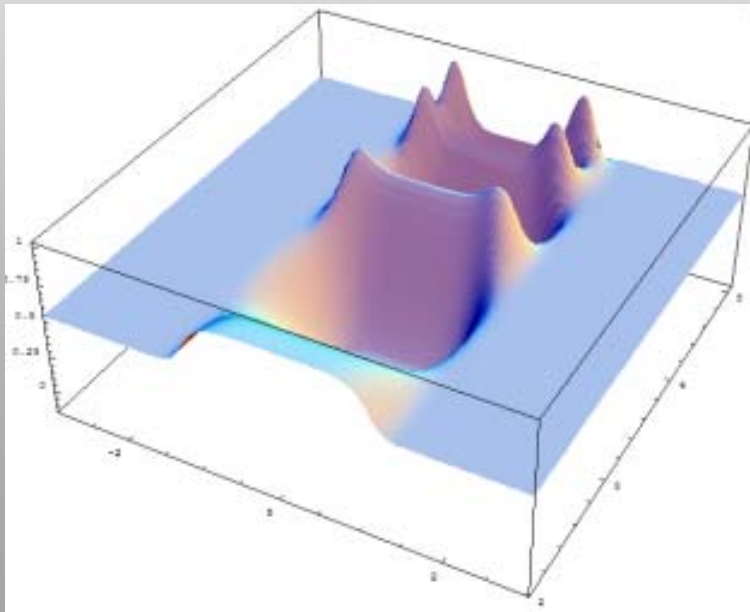
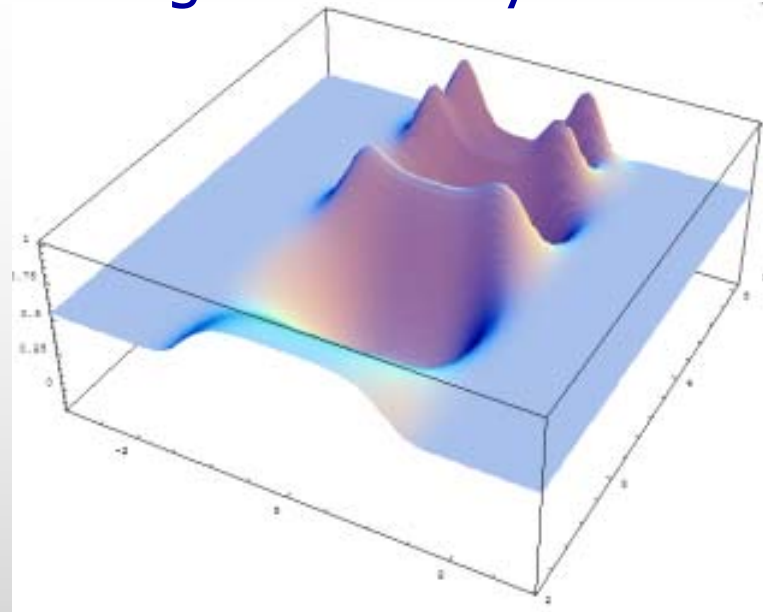
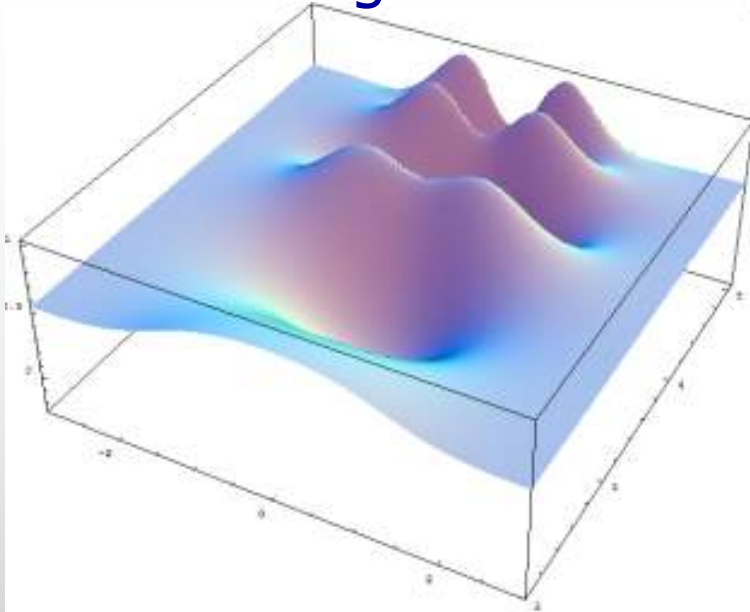
| | |
|-----------|-------------|
| 8805433 | 354054343 |
| 776956408 | 865438676 |
| 66963658 | 554323267 |
| 446854456 | 53887 |
| 446854 | 553246564 |
| 55565435 | 38857448643 |

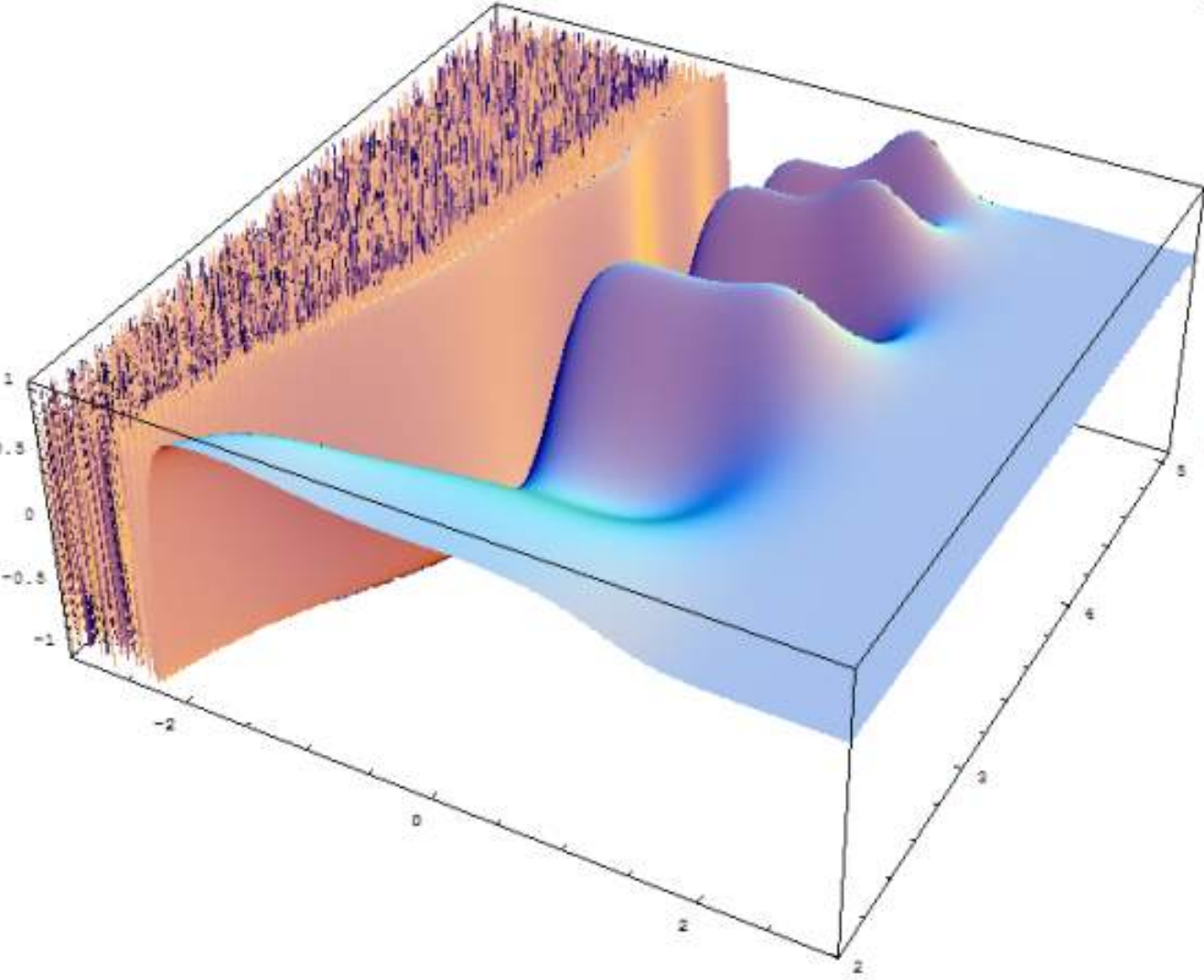




Local Islands in Galois Field Space of statistical convergence and near-convergence to high-probability feature matches using support vector analysis

Convergence with increasing Vocabulary Size





Regions of a typical Galois manifold with no correlation to established vocabulary tensors are easily recognized as exhibiting chaotic behavior and are therefore excluded.

How does this approach differ from traditional N-space cluster analysis?

- Conventional

- Algorithms are custom designed for a narrow recognition task
- Often requires customization with expert programming
- Low tolerance to variability in source format

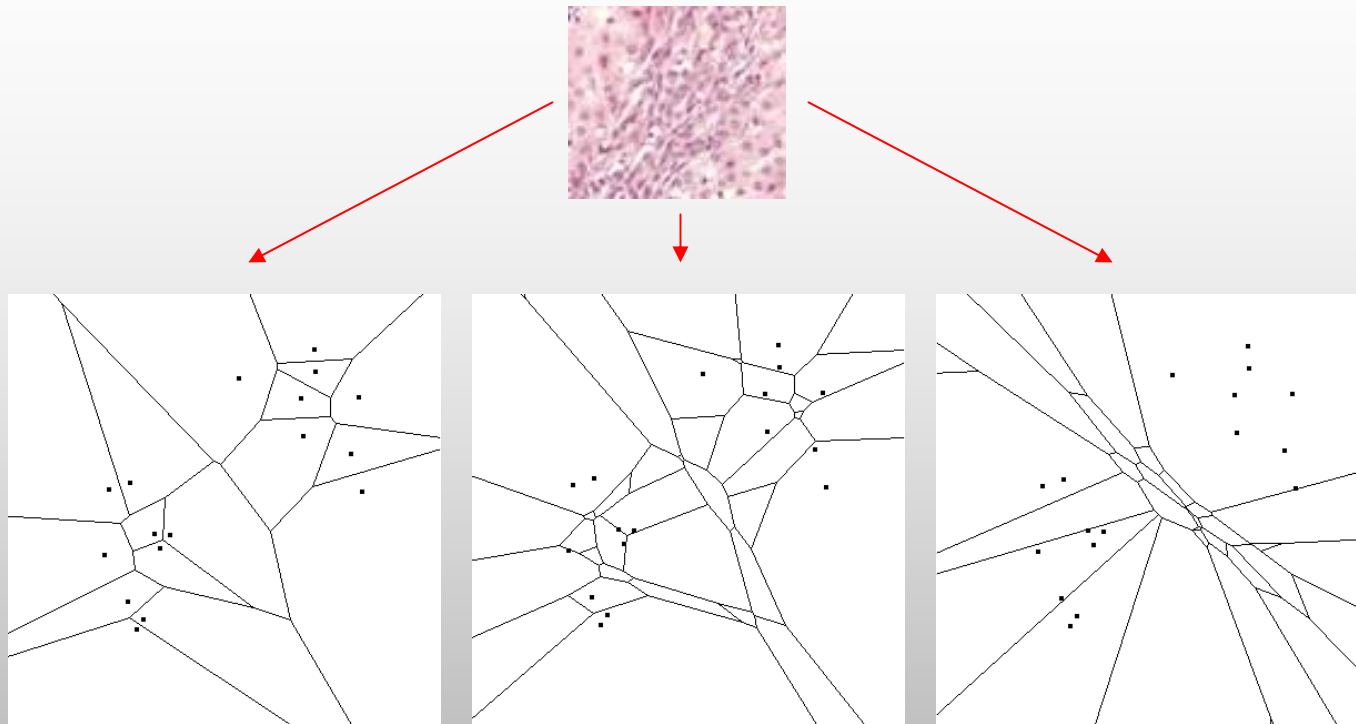
- VQ-Galois

- General matching algorithm agnostic to input data format
- No end-user customization required
- Designed to improve with increased data pool size (self-training)

Derivative Technology: Image-Based Query-by-Example

- New Class of Database
- User to select query by generating an image-based ROI (region of interest)
- ROI is vectorized for comparison with the highly compressed vocabulary library.
- Similar Images (with associated known diagnoses) are returned as a thumbnail gallery.
- A differential diagnosis tool is implicitly enabled

Typical Resultant Voronoi Class System Clusters as basis functions for Bayesian Belief Networks (BBNs)





System Console

File Operations

ENVI (.img format) file name

liver_3432

Load ENVI ImageCube

ENVI load is 100% Complete

Reload Composit VQ Libraries

Command15

ViewPort Mode

- Pseudo RGB
- Monochromatic Singlet
- VQ gated output

Load Viewport

Refresh Viewport

Current Wavelength: 0450.0

459

52

Vectors

1

Vector Creation Settings

X Dim

Y Dim

Vector Class Currently Selected

3

3

Early Fibrosis

Stringency

0.0015

Comment

Empty text area for comments.

Vector Tuning

Vector Number

0

Total in Class

1

Review Mode

- All
- Class-Specific

Vector Class

Early Fibrosis

X

Y

Z

Distance

% Complete

3

3

3

0.015

100

Test this vector only

Re-Render with below Metric

Render Results

0.07

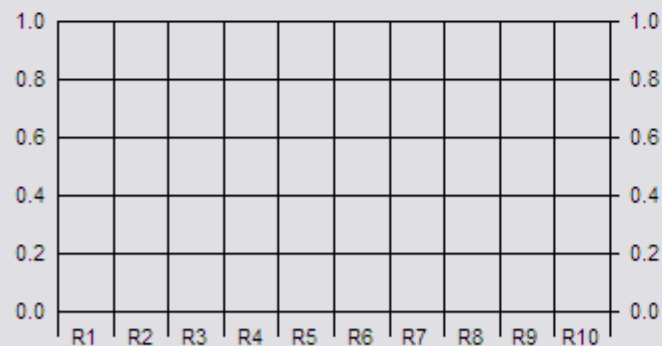
Save Modified Vector

Z Selector



Vertical slider for Z Selector.

Single-Point Spectrum



.1856019

Search Operations

Vector Class

OR

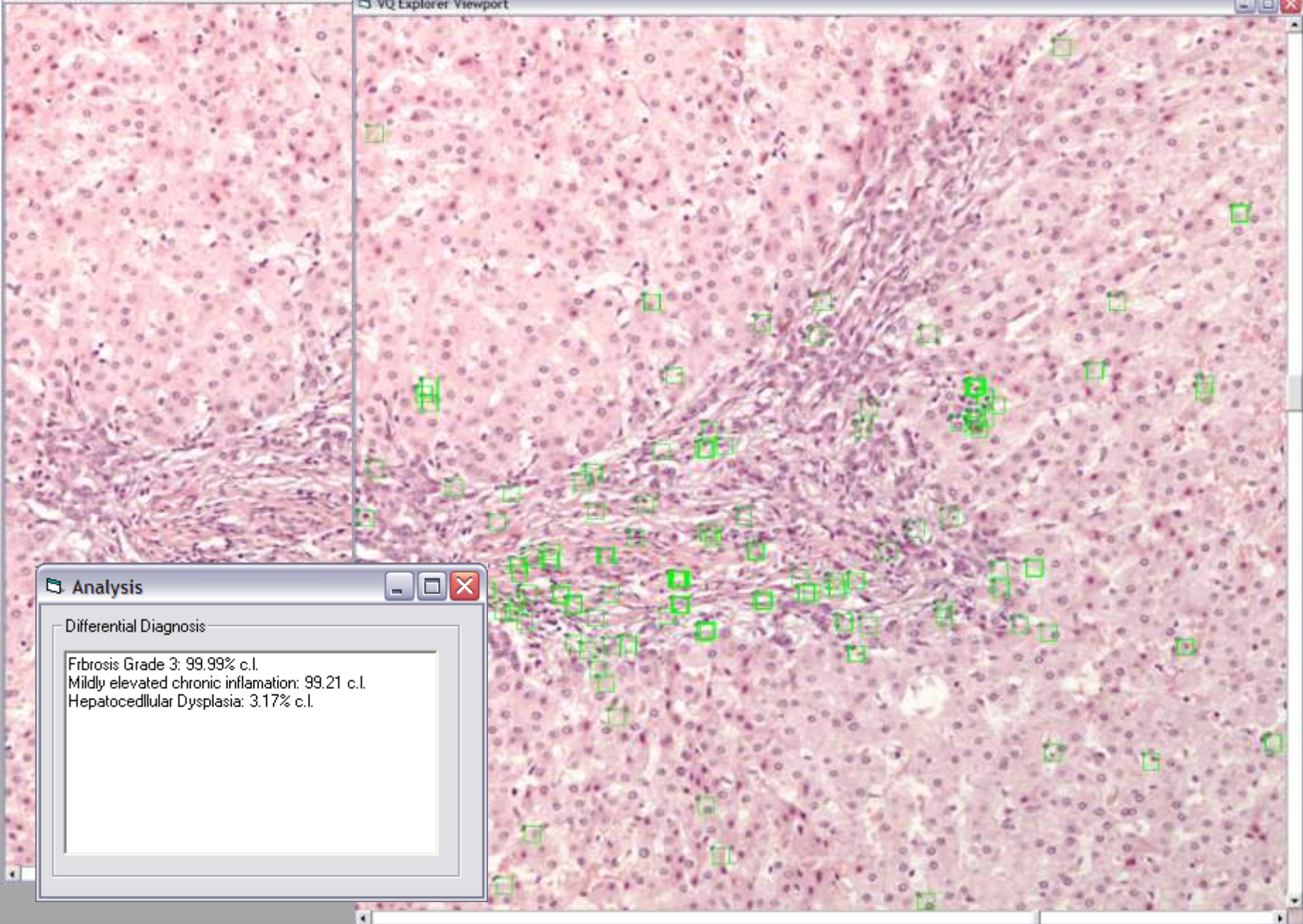
NOT

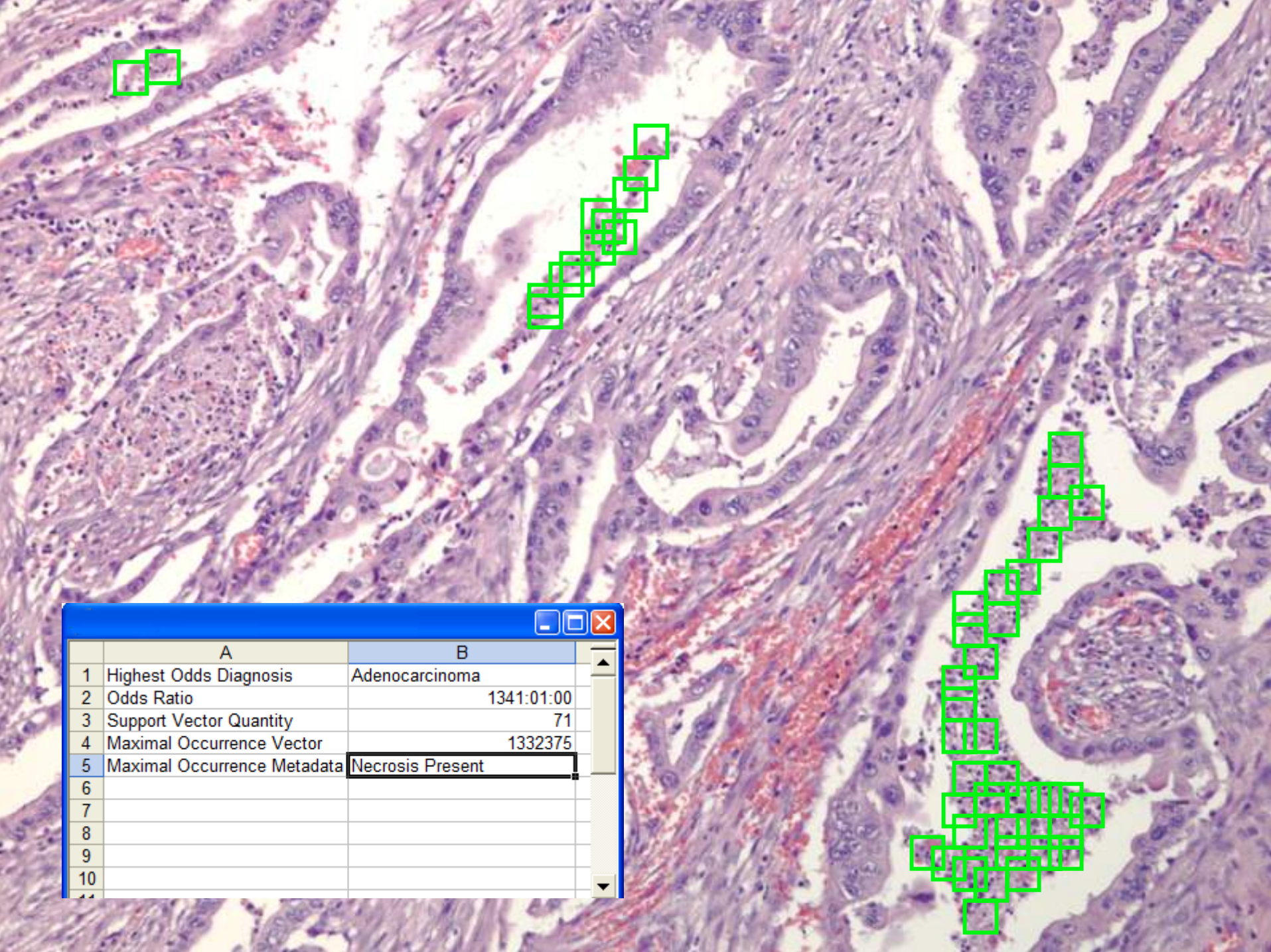
Early Fibrosis

VQ search based on above class settings

Terminate Application







| | A | B |
|----|-----------------------------|------------------|
| 1 | Highest Odds Diagnosis | Adenocarcinoma |
| 2 | Odds Ratio | 1341:01:00 |
| 3 | Support Vector Quantity | 71 |
| 4 | Maximal Occurrence Vector | 1332375 |
| 5 | Maximal Occurrence Metadata | Necrosis Present |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |



World Congress on Pathology Informatics

ANNOUNCEMENT AND CALL FOR PAPERS

1st World Congress on Pathology Informatics



Brisbane Convention Centre
16 & 17 August, 2007

- A global perspective on pathology informatics
- Analysis of leading regional issues
- Updates on key pathology informatics
 - Automation
 - Digital Microscopy
 - Omics
 - Shared Care
 - Disease Surveillance
 - Standards Development
 - Pathology Order Entry
 - Micro Electro-Mechanical Systems

www.wcpi07.org

