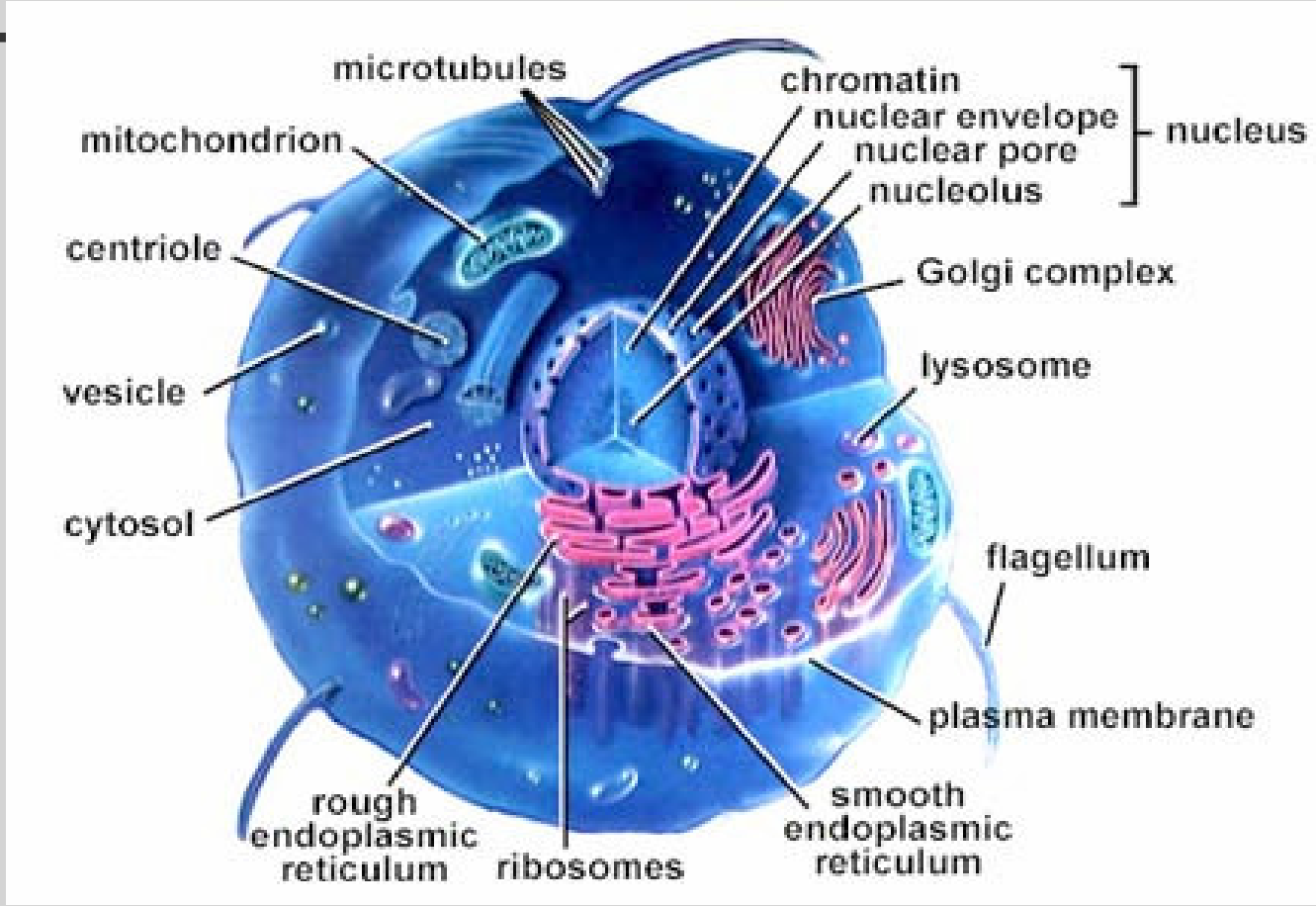
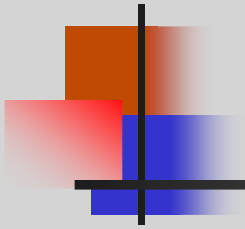
Automated Interpretation of Subcellular Patterns in Microscope Images: Bioimage Informatics for Systems Biology

Estelle Glory (eglory@cmu.edu)
Murphy Group - Center for Bioimage Informatics
Carnegie Mellon University

Carnegie Mellon

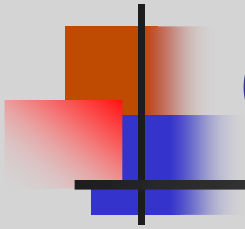
Eukaryotic cells have many parts





Protein localization

- The sequence of each protein determines where it is localized in cells
- Subsequences (“motifs”) within a protein’s sequence are responsible for targeting it to one (or more) locations (structures/organelles)

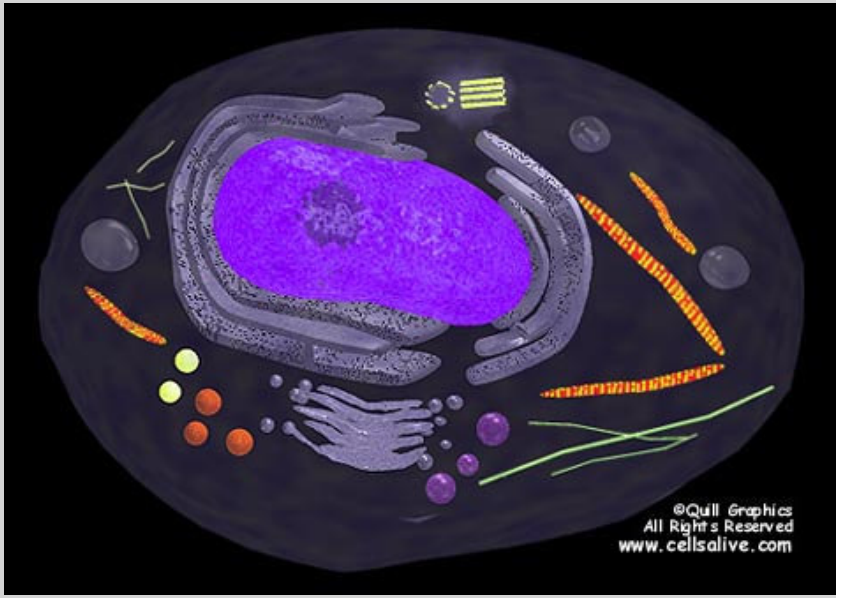


Open questions

- How many distinct locations can proteins be found in? What are they?
- How many distinct motifs direct proteins to those locations? What are they?

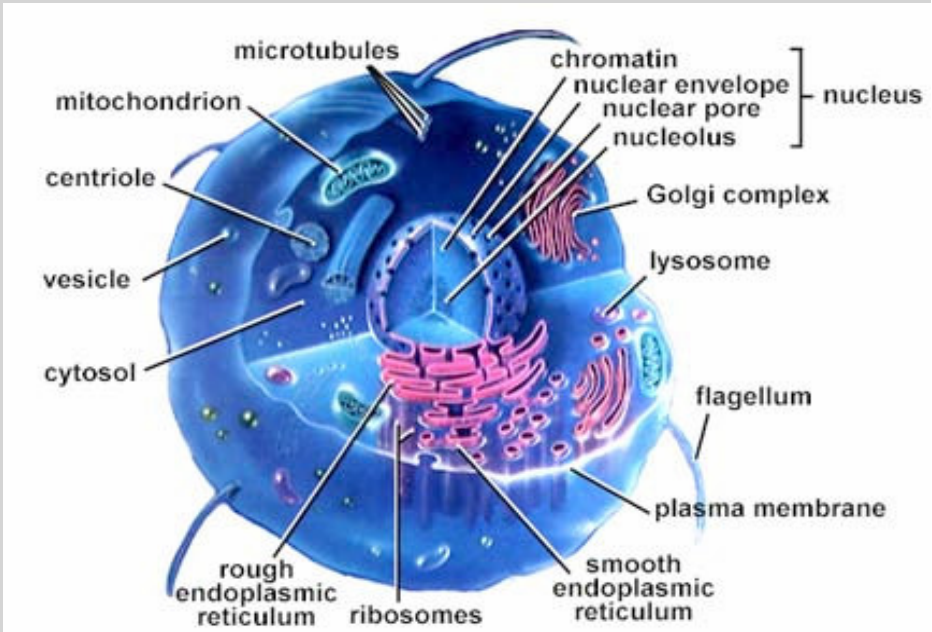
Proteomics

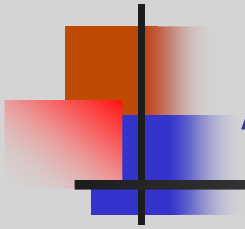
- The set of proteins expressed in a given cell type or tissue is called its *proteome*
- Proteomics projects
  - sequence
  - structure
  - activity
  - partners
  - **location**



Systems Biology and Location Proteomics

- All systems biology must be data driven
- Key to progress
  - identification of aspect that needs to be analyzed “ome-wide”
  - development of assays and automated analysis approaches
- Systems biology needs systematic information on high-resolution subcellular location
  - Eventually, for every expressed protein for all cell types under all conditions
- Providing this information is the goal of Location Proteomics





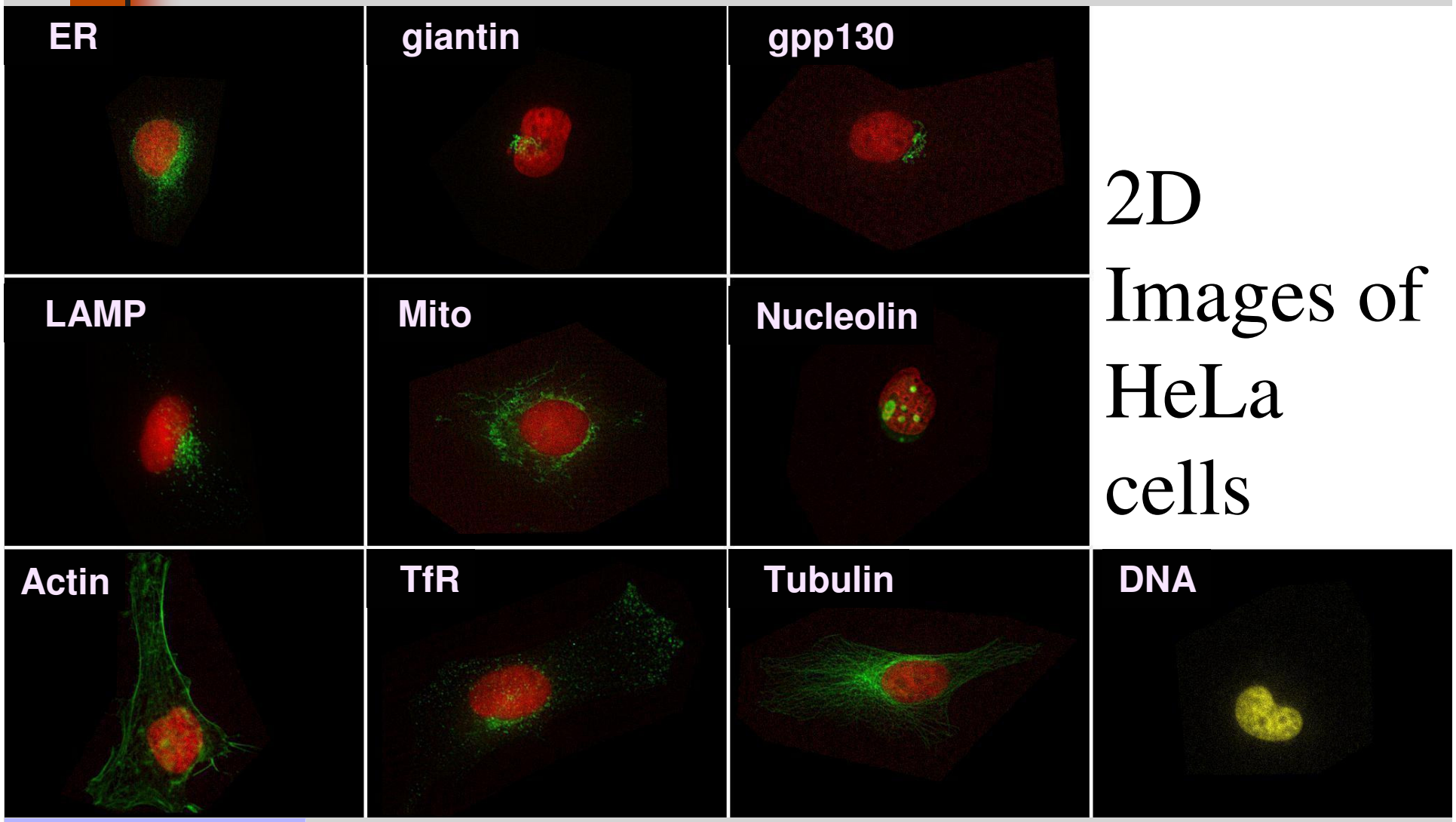
Automated Interpretation

- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past ten years has to been automate the interpretation, to yield better
  - Objectivity
  - Sensitivity
  - Reproducibility

Supervised Learning of High-Resolution Subcellular Location Patterns

Carnegie Mellon

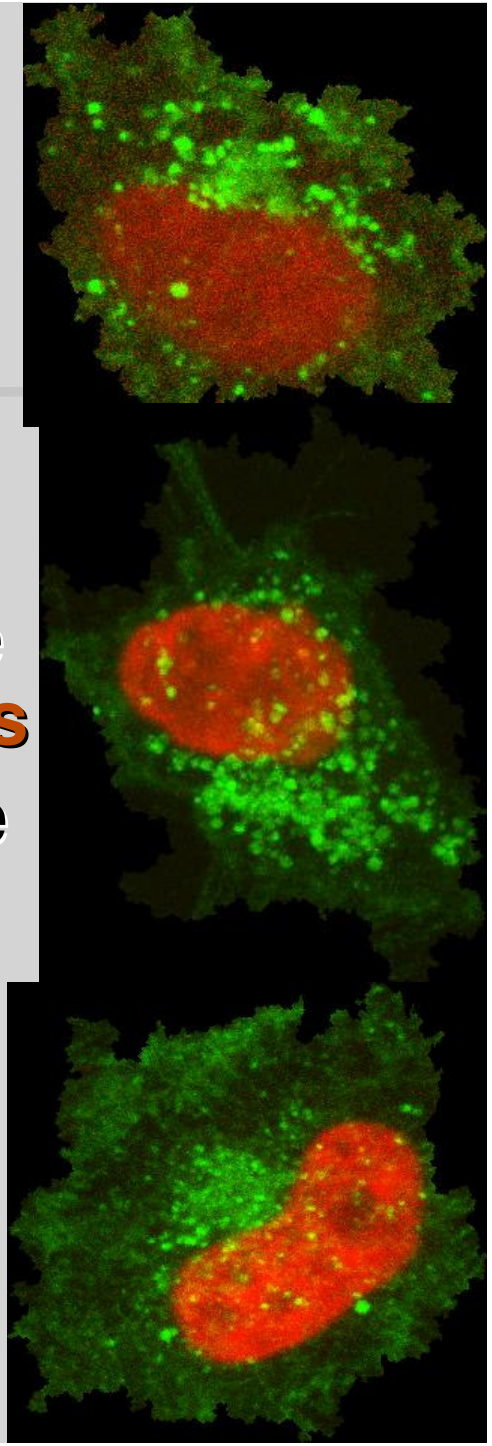
The goal: Learn to recognize all major subcellular patterns



2D Images of HeLa cells

The Challenge

- Pixel-by-pixel or region-by-region matching will not work for cell patterns because different cells have different **shapes, sizes, orientations**
- Organelles/structures within cells are **not found in fixed locations**
- ***Instead, describe each image numerically and compare the descriptors***

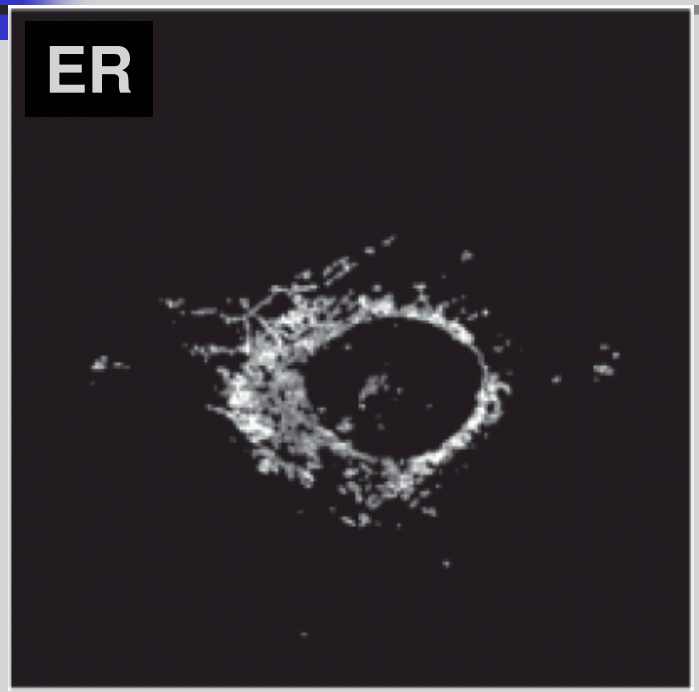


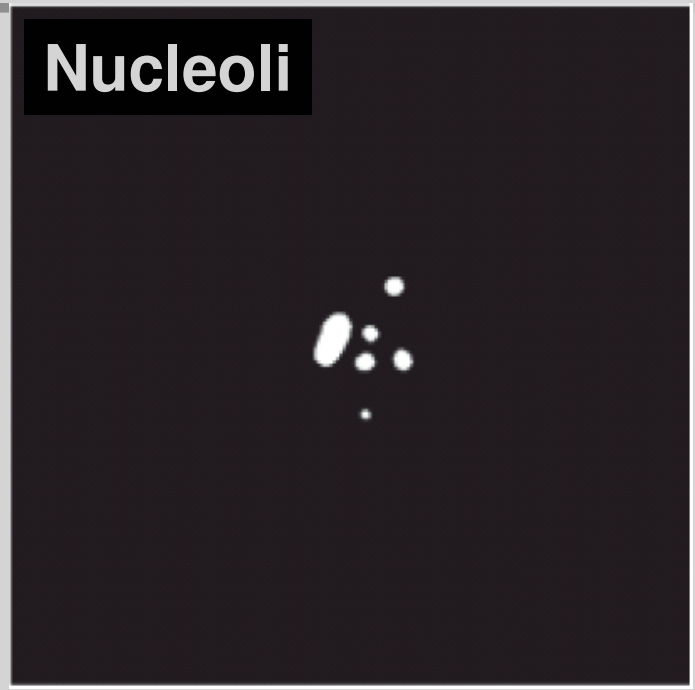
Feature-based, Supervised learning approach

1. Create sets of images showing the location of many different proteins (each set defines one **class** of pattern)
2. Reduce each image to a set of numerical values ("**features**") that are insensitive to position and rotation of the cell
3. Use **machine learning methods** to "learn" how to distinguish each class using the features

Boland et al 1996; 1997; 1998; Boland & Murphy 2001; Huang & Murphy 2004

Example of classification using Morphological Features





108	# of objects	6 ←
83	Average size of objects	232 ←
31	Average distance to COF	4 ←

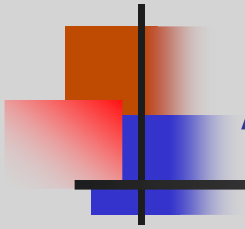
Any of these features could be used to distinguish these two classes

Acquisition considerations

- Resolution defined as ability to distinguish two “point-sources”
- Maximal resolution in x-y plane given by Rayleigh (or Abbe) limit

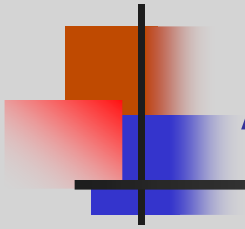
$$1.22\lambda/2NA$$

- where λ is wavelength of emitted light and NA is the numerical aperture of the objective; 244 nm for 520 nm light and 1.3 NA
- Sampling theorem (Nyquist) says maximum information can be obtained if we sample at twice the maximum frequency present in a sample
- Try to achieve Nyquist Sampling at Rayleigh limit



Acquisition considerations

- Maintain low cell density if single cell measurements desired
- Control acquisition variables
  - Select (initial) focal plane consistently
  - Select fields consistently (at least one full cell per field)
  - Maintain constant camera gain, exposure time, number of slices
  - Select interphase cells or ensure sampling of cell cycle

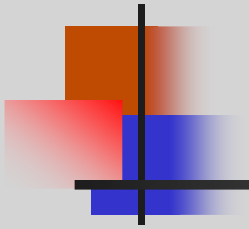


Acquisition considerations

- Collect sufficient images per condition
  - For classifier training or set comparison, more than number of features
  - For classification or clustering, based on confidence level desired
- Collect reference images if possible (DNA, membrane)

Annotation considerations

- Maintain adequate records of all experimental settings
- Organize images by cell type/probe/condition

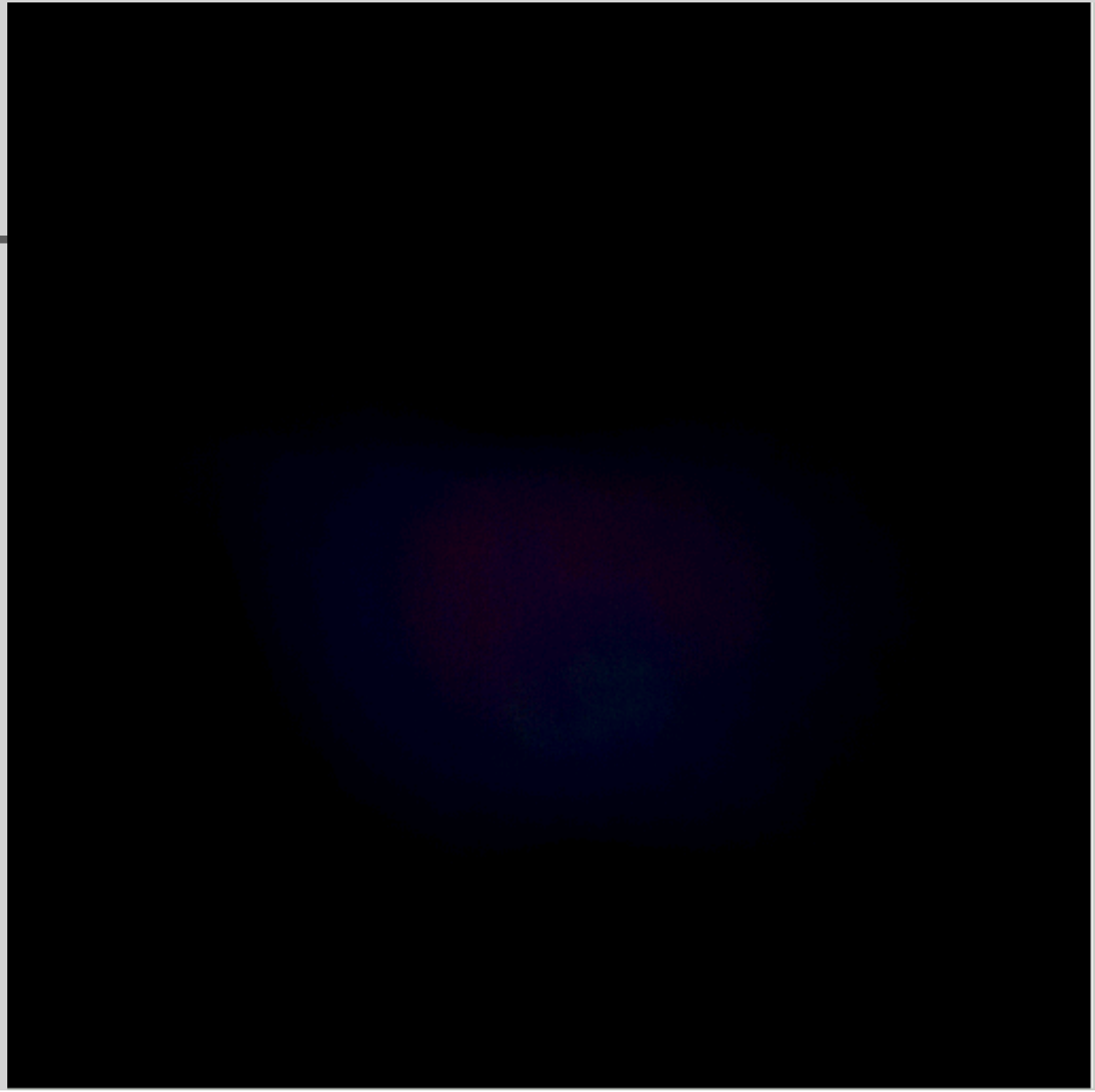


Preprocessing

- Correction for/Removal of camera defects
- Background correction
- Autofluorescence correction
- Illumination correction
- Deconvolution

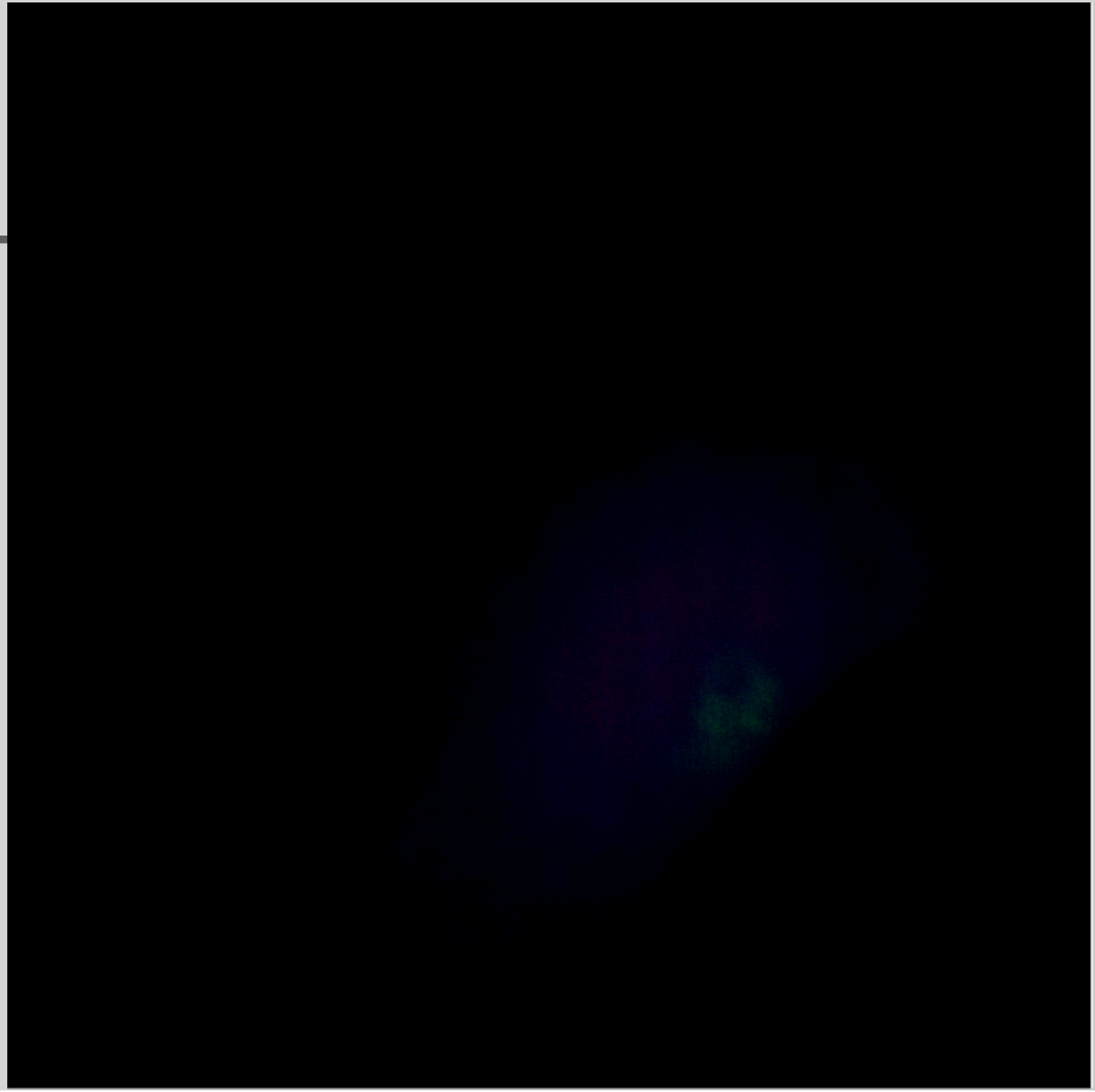
3D HeLa

- 2D slices (from bottom to top) for cell labeled for **transferrin receptor** (primarily in endosomes)



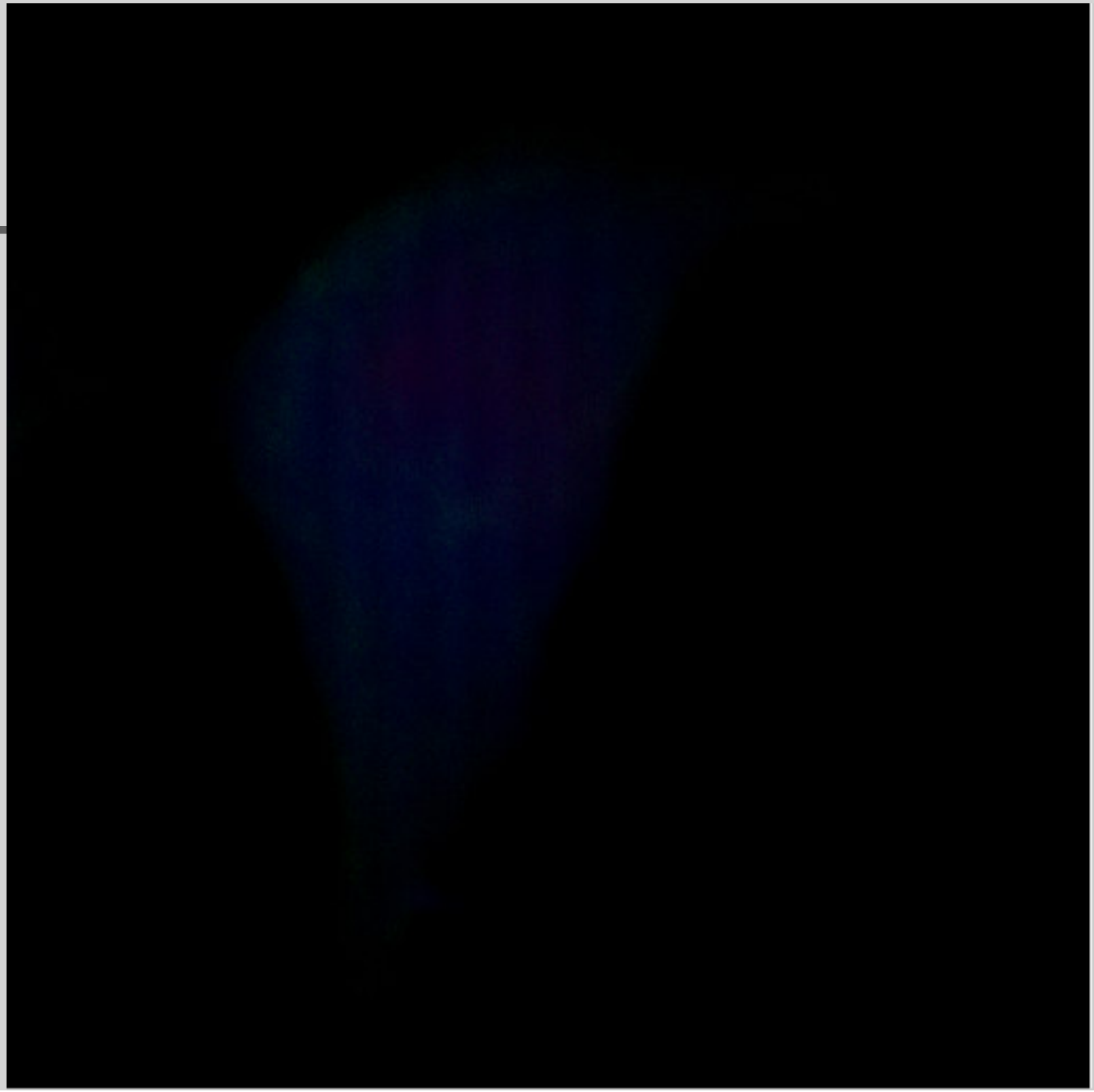
3D HeLa

- 2D slices (from bottom to top) for cell labeled for **giantin** (primarily in Golgi)



3D HeLa

- 2D slices (from bottom to top) for cell labeled for **tubulin** (major constituent of microtubules)

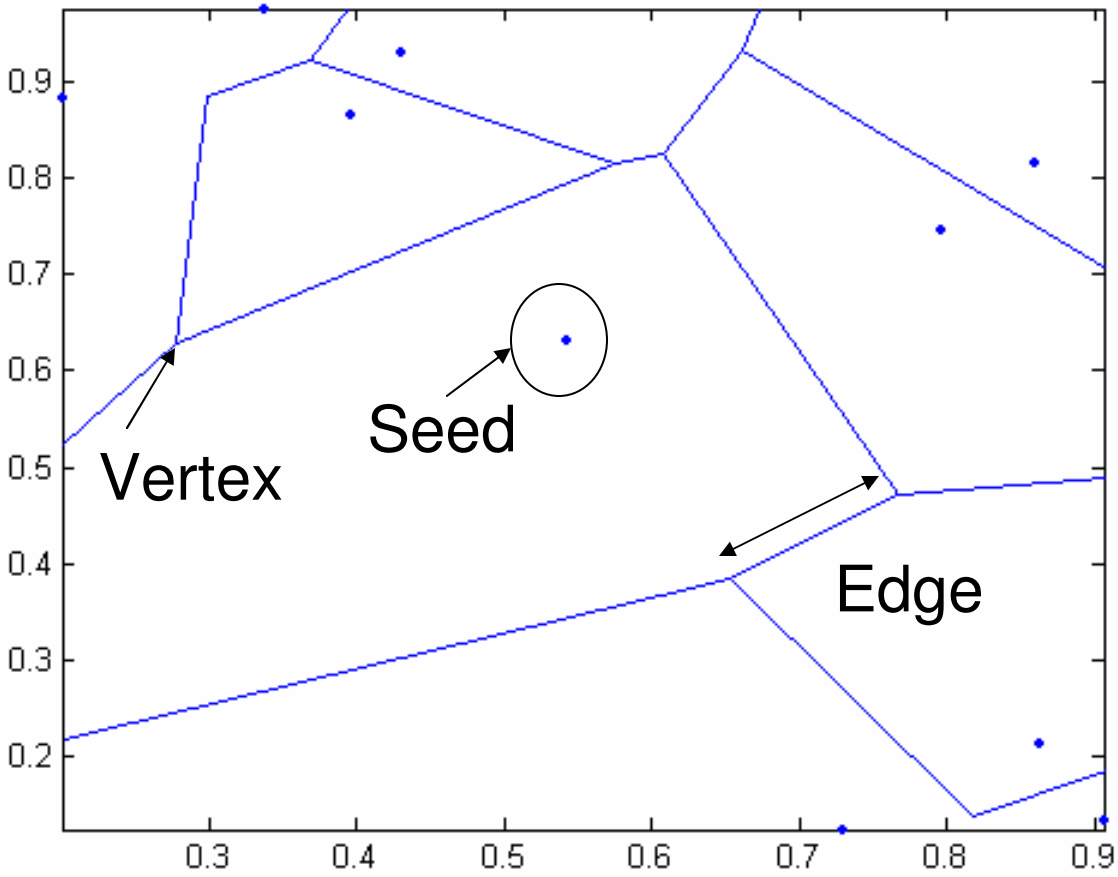


Single cell segmentation approaches

- Voronoi
- Watershed
- Seeded Watershed
- Level Set Methods
- Graphical Models

Voronoi diagram

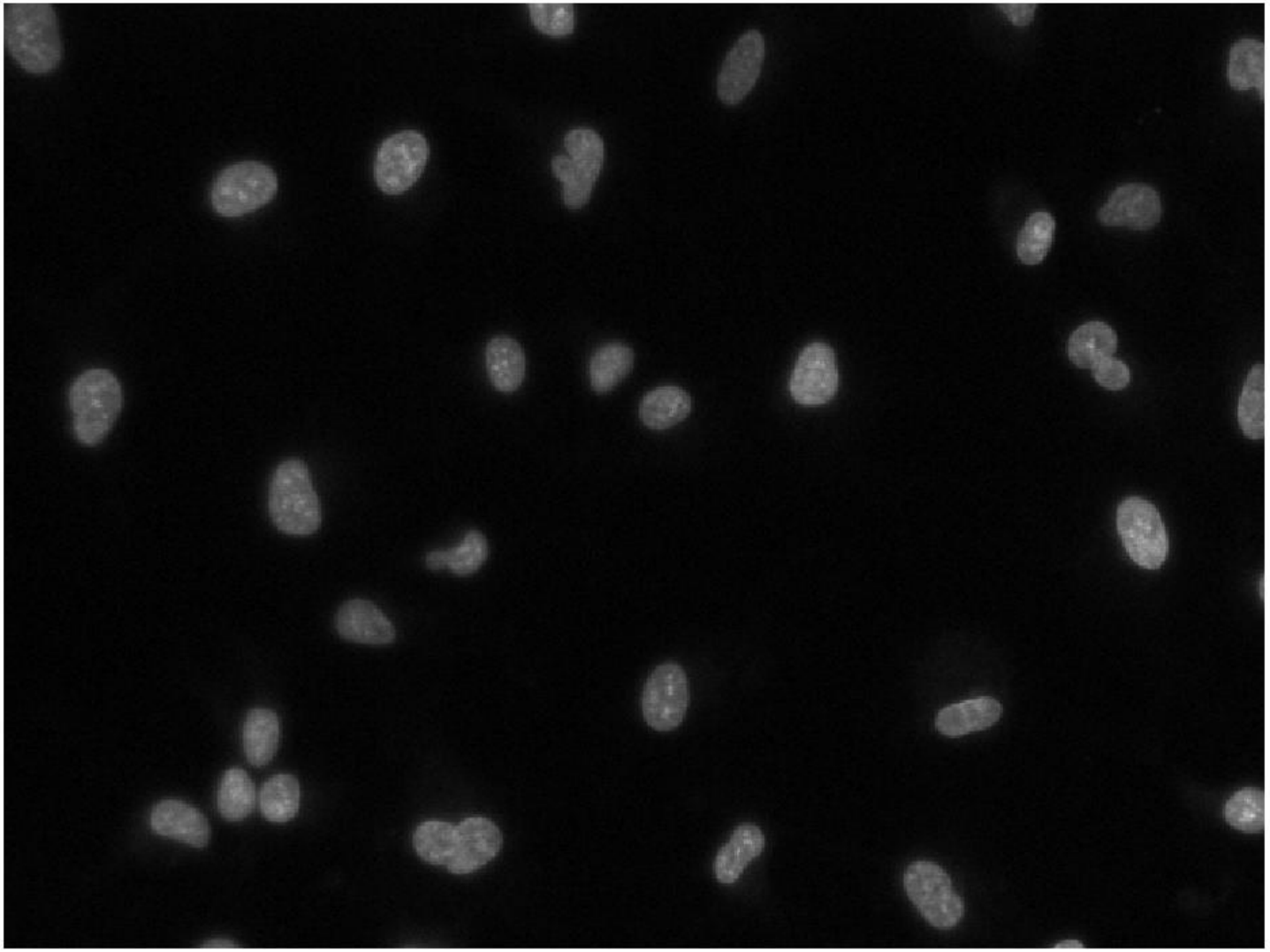
Given a set of seeds, draw vertices and edges such that each seed is enclosed in a single polygon where each edge is equidistant from the seeds on either side.



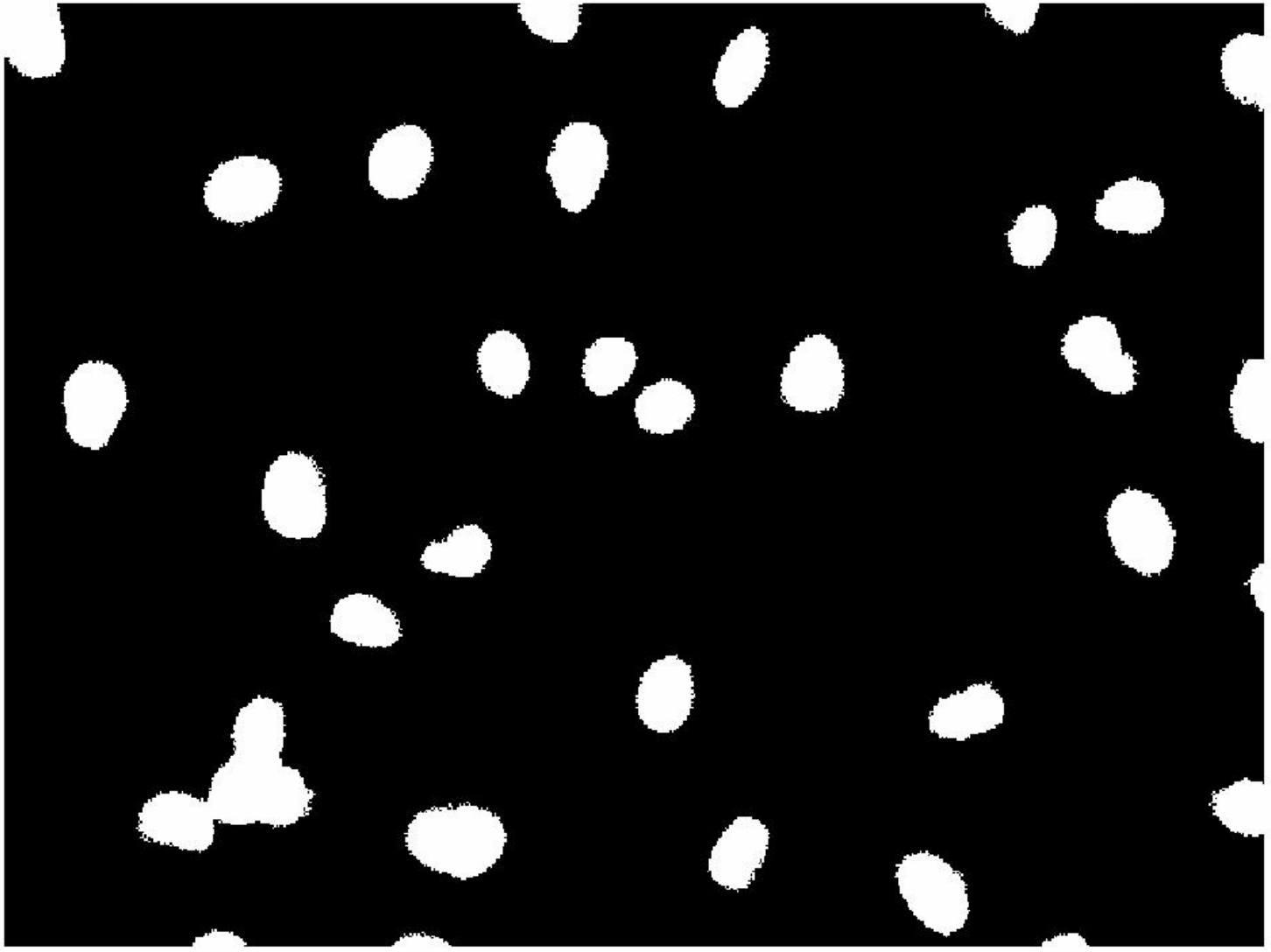
Voronoi Segmentation Process

- Threshold DNA image (downsample?)
- Find the objects in the image
- Find the centers of the objects
- Use as seeds to generate Voronoi diagram
- Create a mask for each region in the Voronoi diagram
- Remove regions whose object that does not have intensity/size/shape of nucleus

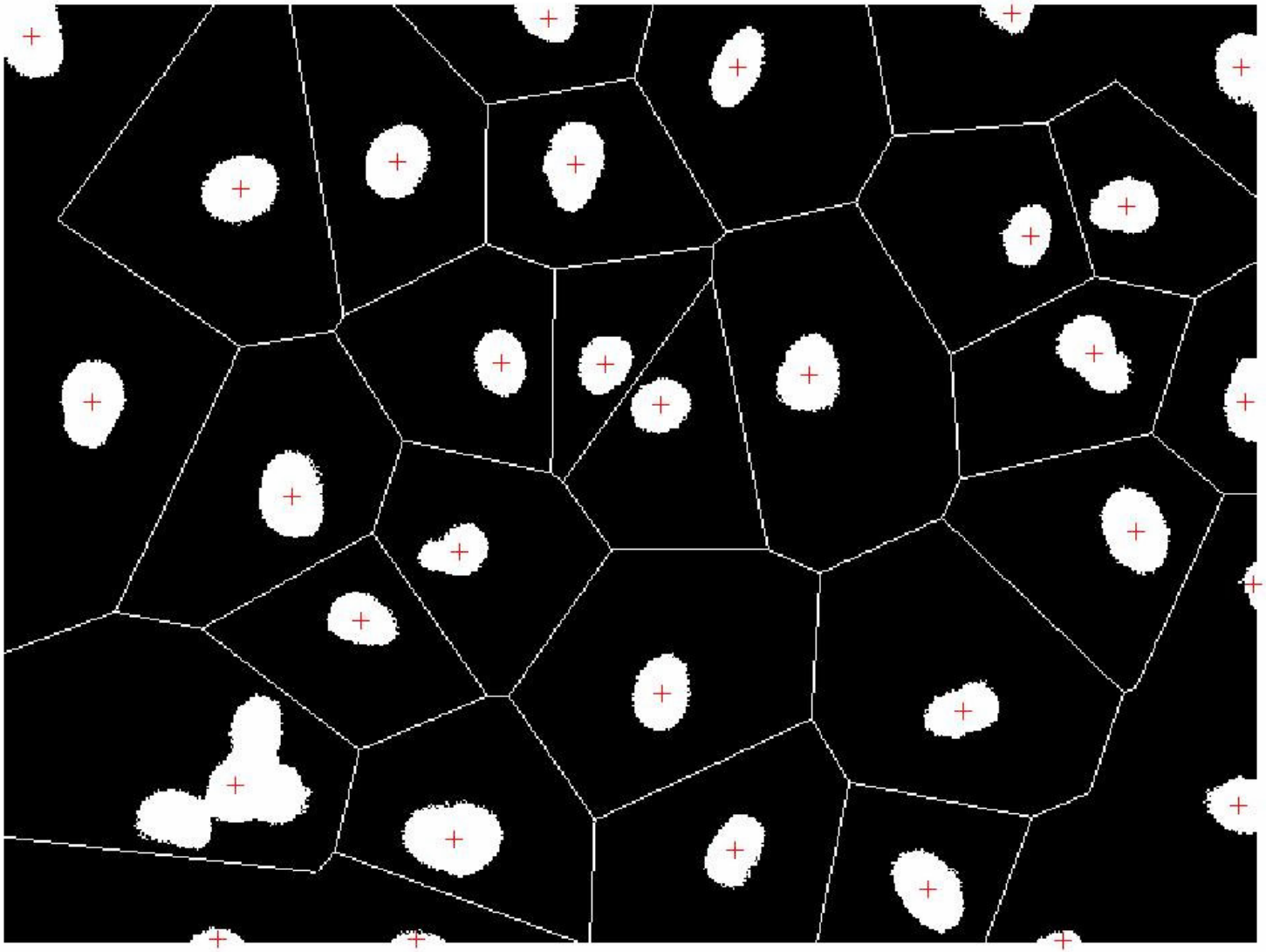
Original DNA image



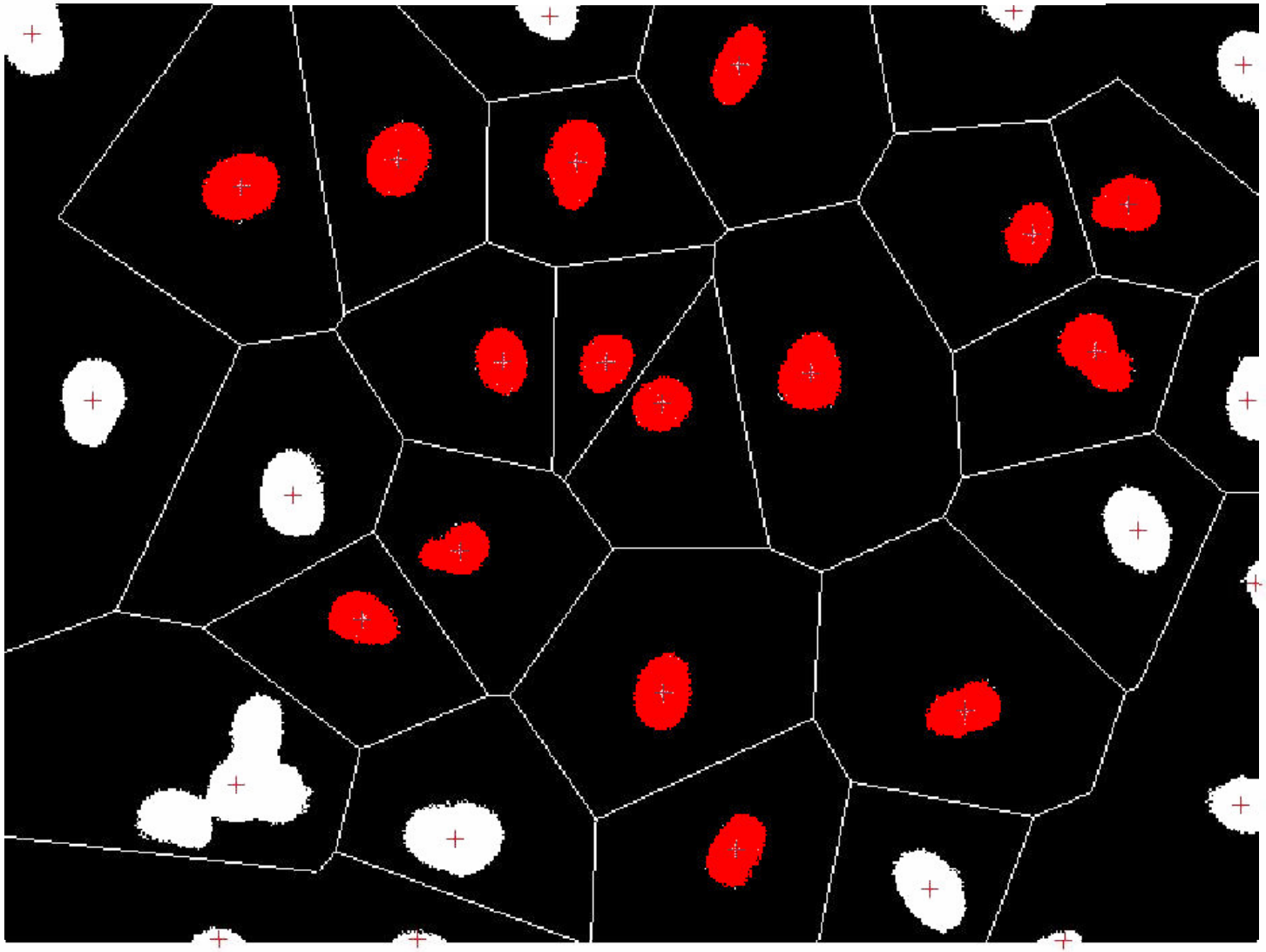
After thresholding and removing small objects



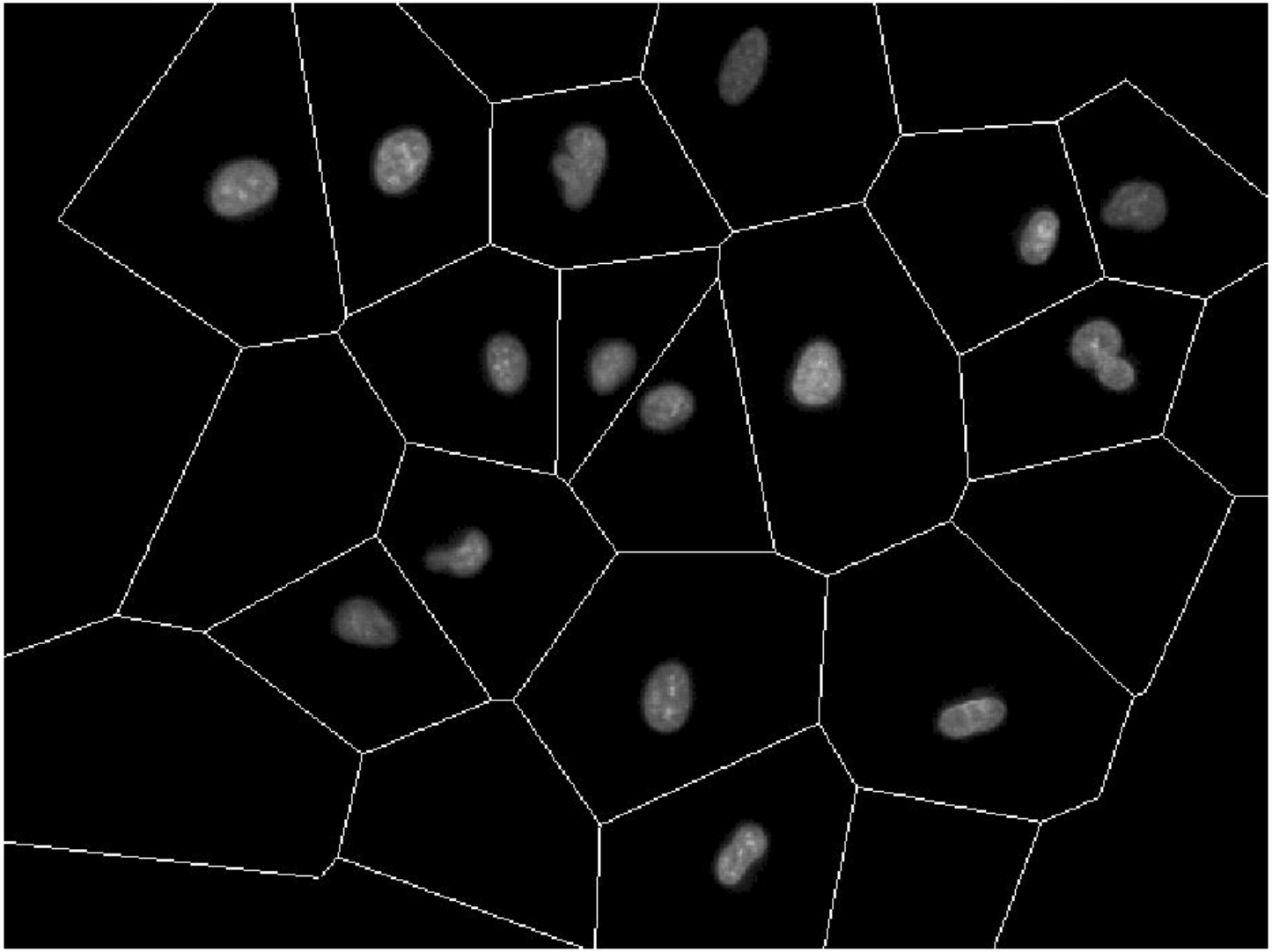
After triangulation



After removing edge cells and filtering

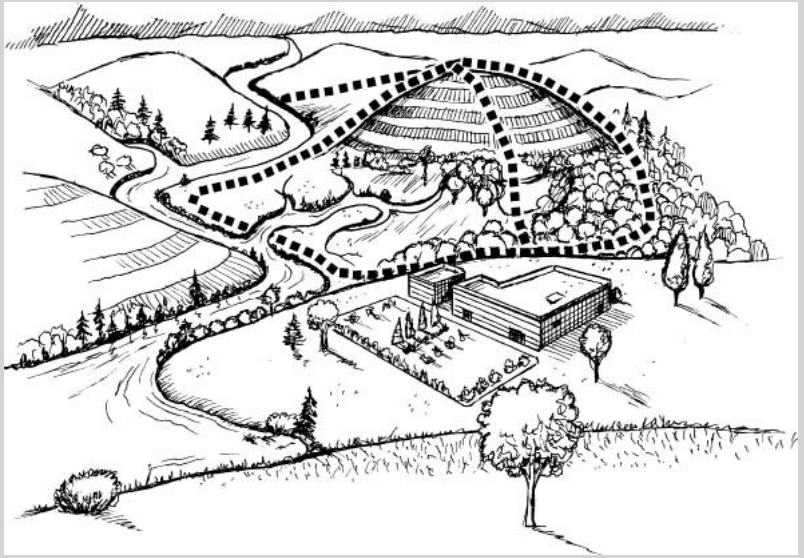


Final regions masked onto original image

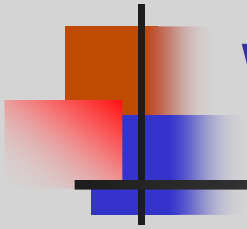


Watershed Segmentation

- Intensity of an image ~ elevation in a landscape
  - Flood from minima
  - Prevent merging of “catchment basins”
  - Watershed borders built at contacts between basins



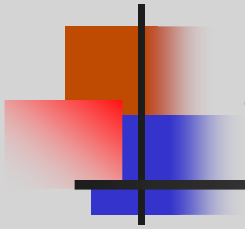
http://www.ctic.purdue.edu/KYW/glossary/whatisaws.html



Watershed Segmentation

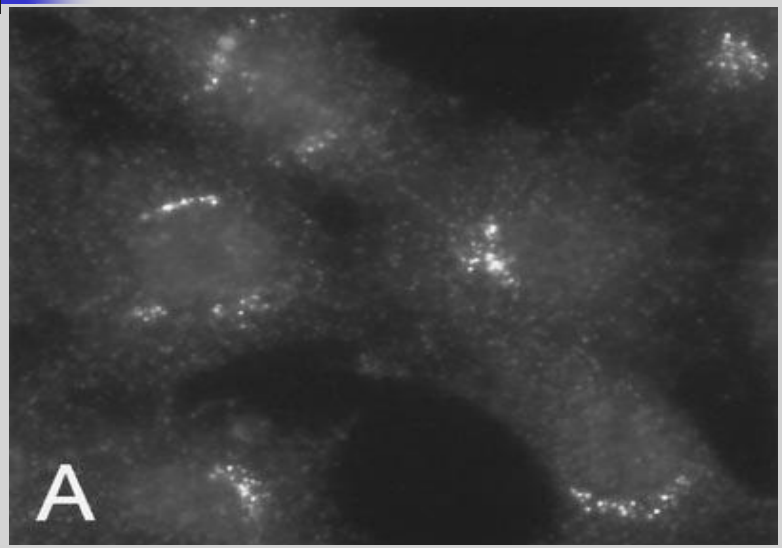
- If starting image has intensity centered on the cells (e.g., DNA) that you want to segment, invert image so that bright objects are the sources
- If starting image has intensity centered on the boundary between the cells (e.g., plasma membrane protein), don't invert so that boundary runs along high intensity

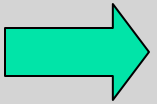
Seeded Watershed Segmentation

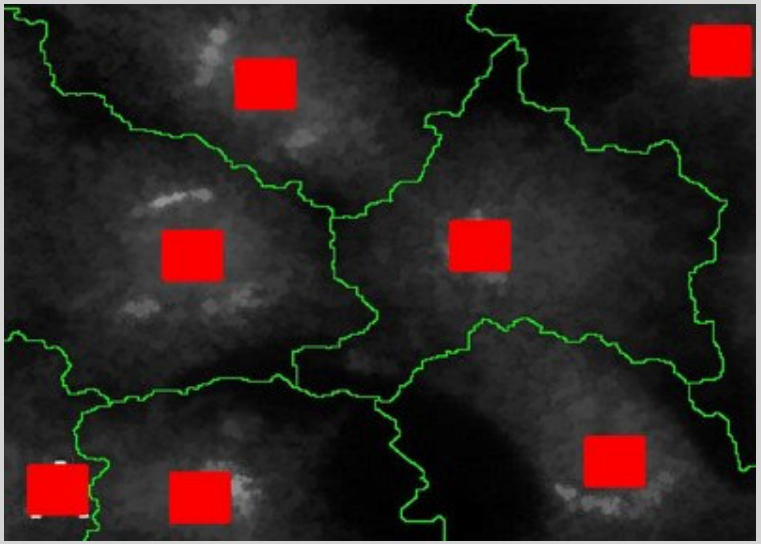


- Drawback is that the number of regions may not correspond to the number of cells
- Seeded watershed allows water to rise only from predefined sources (seeds)
- If DNA image available, can use same approach to generate these seeds as for Voronoi segmentation
- Can use seeds from DNA image but use total protein image for watershed segmentation

Seeded Watershed Segmentation







Original image

Seeds and boundary

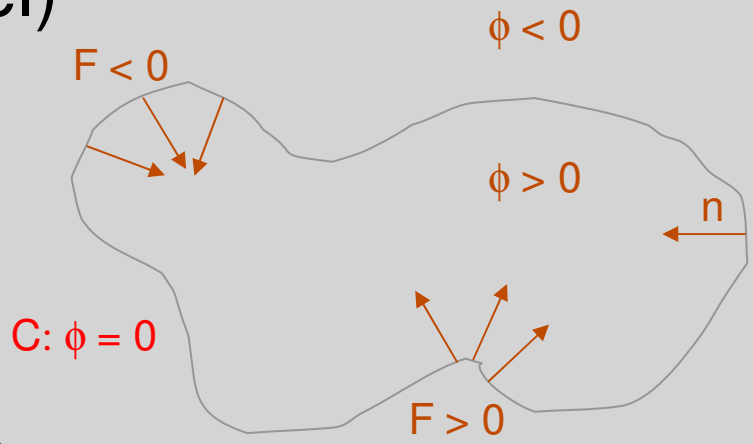
Applied directly to protein image (no DNA image)

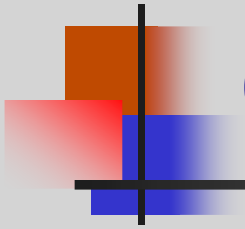
Note non-linear boundaries

Level Set Methods

- Level set function $\phi(x,y,t)$
  - Positive inside the contour (mountain)
  - Negative outside the contour (valley)
  - Zero on the contour, C embedded at its zero level (sea level)

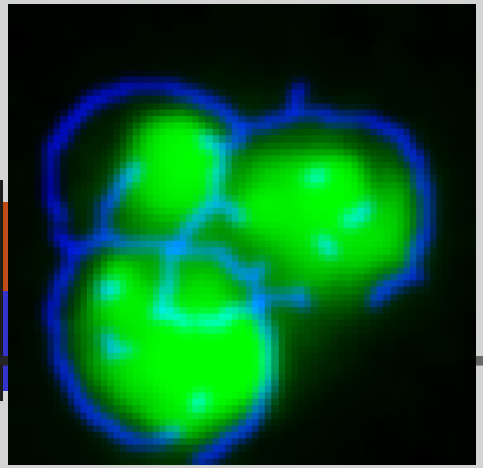
http://ranger.uta.edu/~alp/personal/travelImageGallery.htm



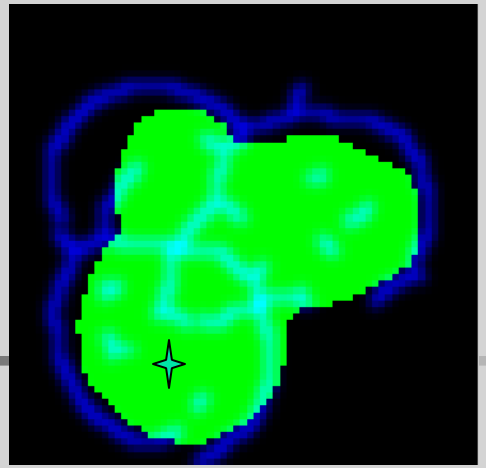


Graphical Model Methods

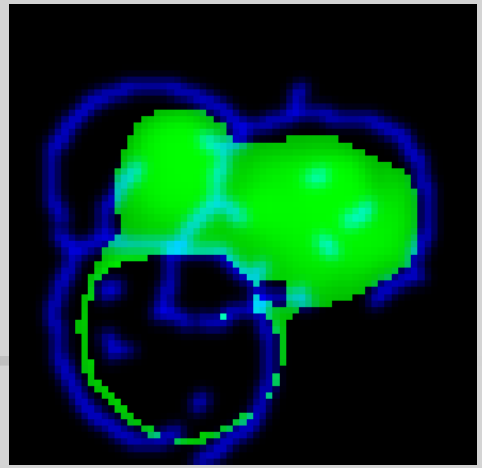
- Assumptions
  - Two classes of pixels: those part of a cell or part of the background
  - Each pixel is likely to be the same class as its neighbors
  - Have information about where cells are likely to be and where boundaries (edges) are likely to be
  - Probability that two pixels are same class related to probability that there is an edge between them



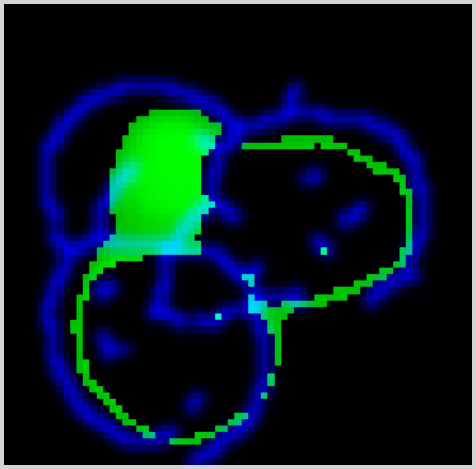
1. Start with initial DNA and edge potential



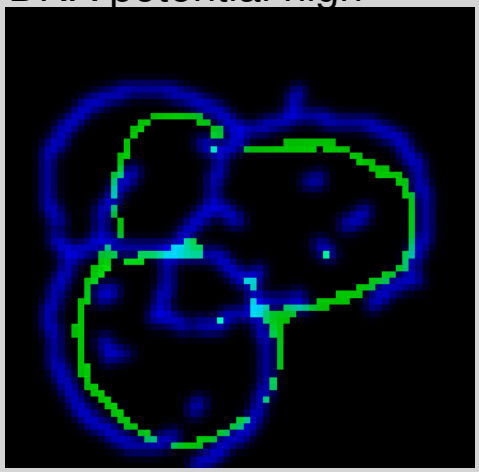
2. Run 1st believe propagation (BP), separate foreground and background. Pick the most confidence foreground pixel p, set its DNA potential high



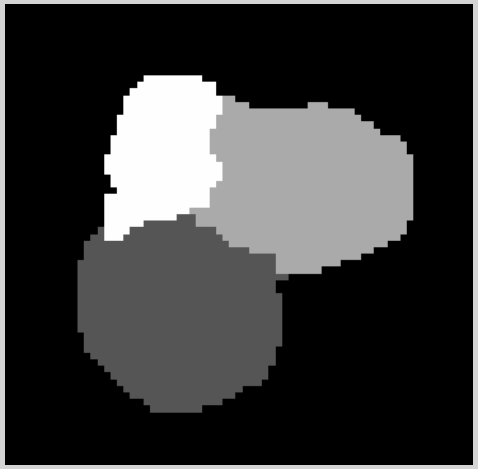
3. Run 2nd BP, assign the pixels with the same class of p to be ***segmented_cell1***, then set these pixels to be background



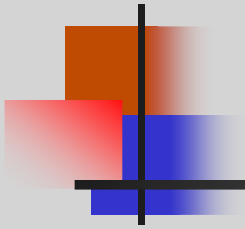
4. Pick the most confident foreground pixel , Run BP, find another cell, and iterate....



5. Iteration stops when the segmented cell is too small

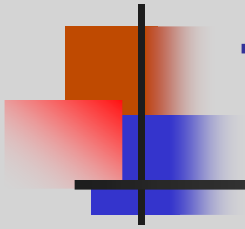


6. The resulting masks

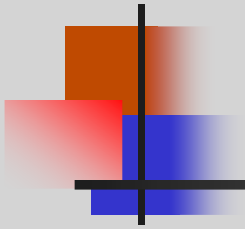


Feature extraction

Morphological Features - Thresholding



- Morphological features require some method for defining objects
- Most common approach is global thresholding
- Methods exist for automatically choosing a global threshold (e.g., Riddler-Calvard method)

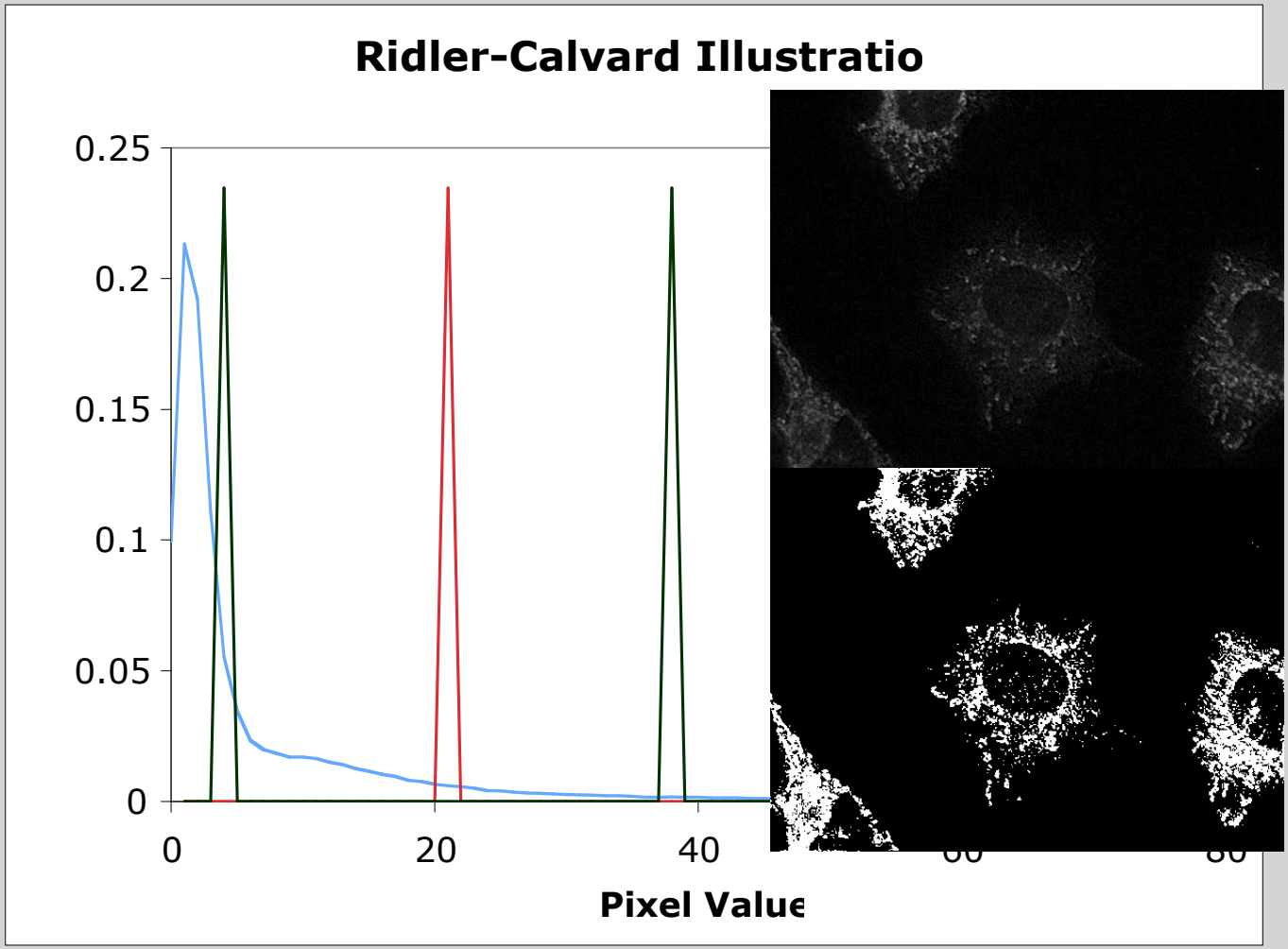


Ridler-Calvard Method

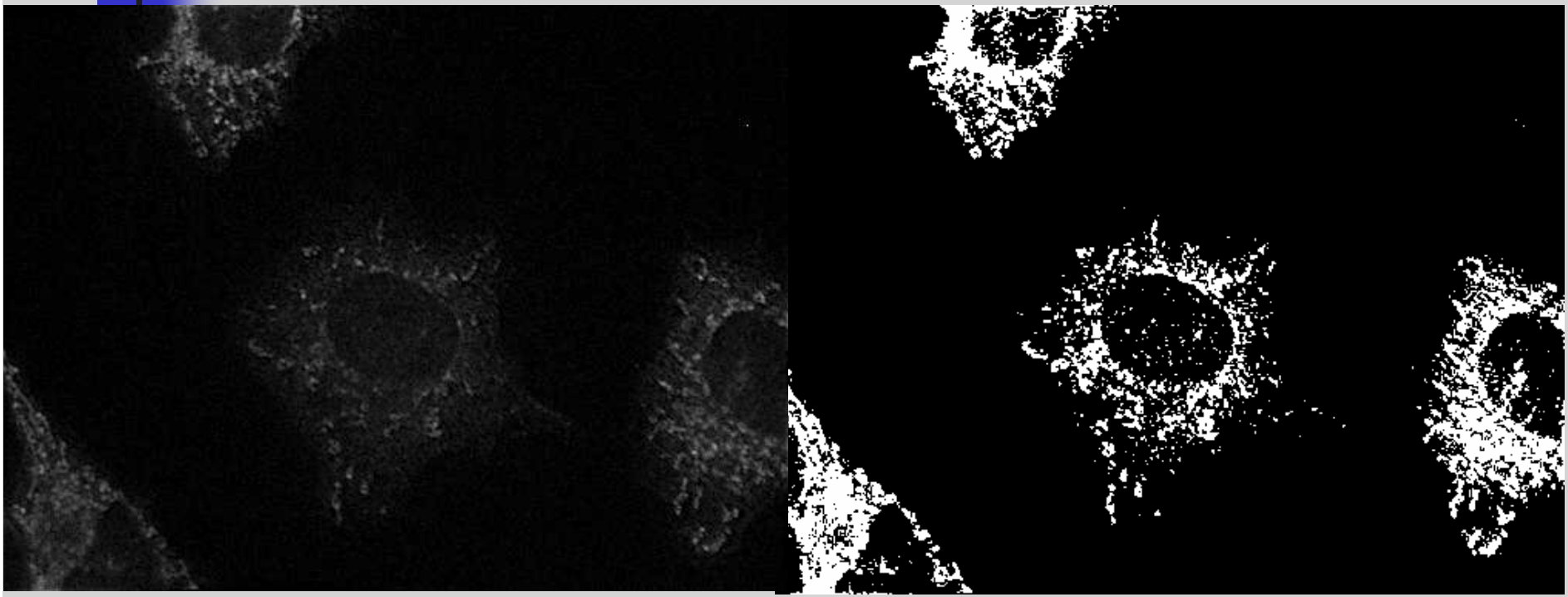
- Find threshold that is equidistant from the average intensity of pixels below and above it
- Ridler, T.W. and Calvard, S. (1978) Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics* 8:630-632.

Ridler-Calvard Method

Blue line shows histogram of intensities, green lines show average to left and right of red line, red line shows midpoint between them or the RC threshold

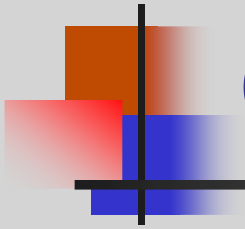


Ridler-Calvard Method



original

thresholded

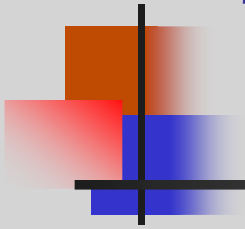


Object finding

- After choice of threshold, define objects as sets of touching pixels that are above threshold

2D Features
Morphological Features

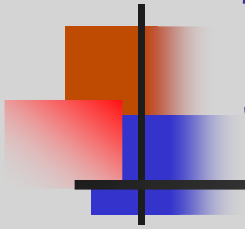
SLF No.	Description
SLF1.1	The number of fluorescent objects in the image
SLF1.2	The Euler number of the image
SLF1.3	The average number of above-threshold pixels per object
SLF1.4	The variance of the number of above-threshold pixels per object
SLF1.5	The ratio of the size of the largest object to the smallest
SLF1.6	The average object distance to the cellular center of fluorescence(COF)
SLF1.7	The variance of object distances from the COF
SLF1.8	The ratio of the largest to the smallest object to COF distance



2D Features
DNA Features

DNA features (objects relative to DNA reference)

SLF No.	Description
SLF2.17	The average object distance from the COF of the DNA image
SLF2.18	The variance of object distances from the DNA COF
SLF2.19	The ratio of the largest to the smallest object to DNA COF distance
SLF2.20	The distance between the protein COF and the DNA COF
SLF2.21	The ratio of the area occupied by protein to that occupied by DNA
SLF2.22	The fraction of the protein fluorescence that co-localizes with DNA

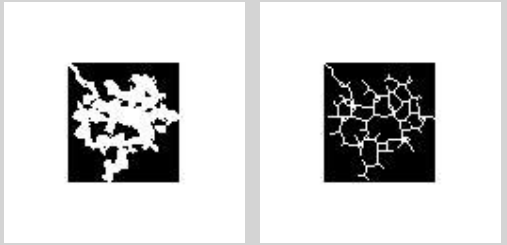


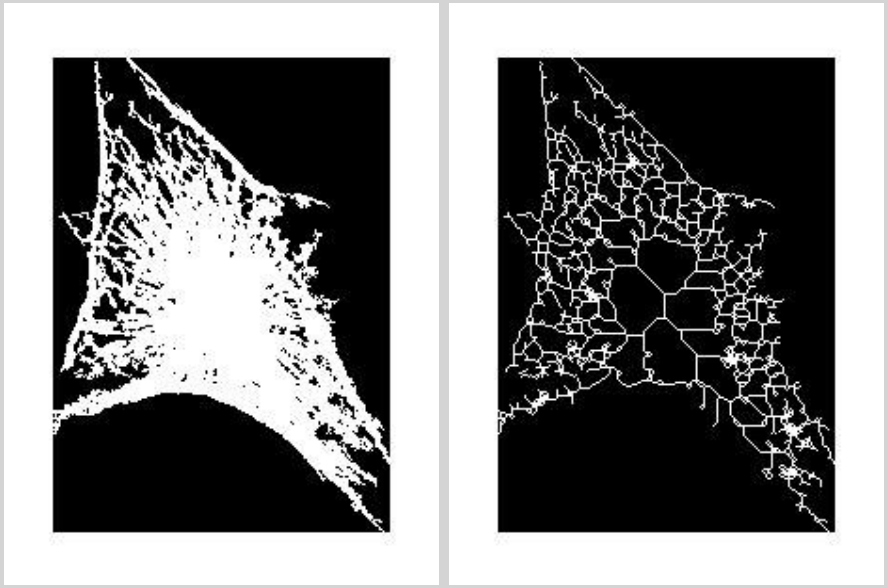
2D Features
Skeleton Features

Skeleton features

SLF No.	Description
SLF7.80	The average length of the morphological skeleton of objects
SLF7.81	The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects
SLF7.82	The fraction of object pixels contained within the skeleton
SLF7.83	The fraction of object fluorescence contained within the skeleton
SLF7.84	The ratio of the number of branch points in the skeleton to the length of skeleton

Illustration – Skeleton





2D Features
Edge Features

Edge features

SLF No.	Description
SLF1.9	The fraction of the non-zero pixels that are along an edge
SLF1.10	Measure of edge gradient intensity homogeneity
SLF1.11	Measure of edge direction homogeneity 1
SLF1.12	Measure of edge direction homogeneity 2
SLF1.13	Measure of edge direction difference

2D Features
Haralick Texture Features
(SLF7.66-7.78)

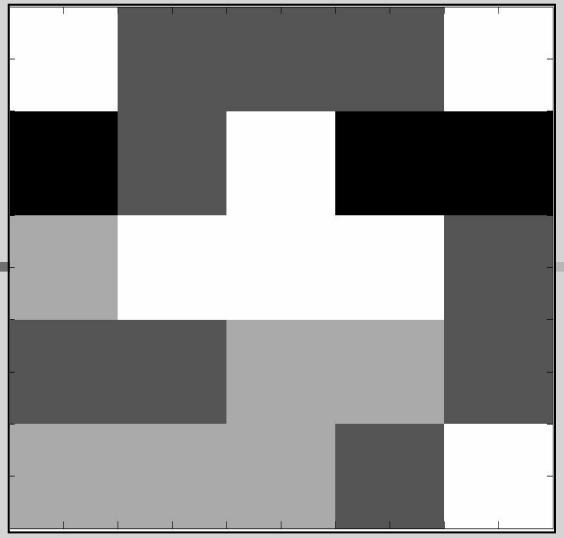
- Correlations of adjacent pixels in gray level images
- Start by calculating co-occurrence matrix P:

  N by N matrix, N=number of gray level.

  Element P(i,j) is the probability of a pixel with value i being adjacent to a pixel with value j
- Four directions in which a pixel can be adjacent
- Each direction considered separately and then features averaged across all directions

Co-occurrence Matrices

Example image with 4 gray levels



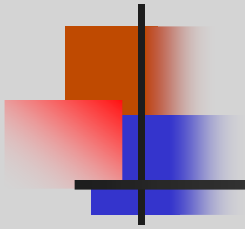
4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

	1	2	3	4
1	0	2	1	3
2	2	4	4	3
3	1	4	2	2
4	3	3	2	2

	1	2	3	4
1	2	1	0	1
2	1	6	3	4
3	0	3	6	2
4	1	4	2	4

	1	2	3	4
1	0	1	0	3
2	1	4	3	3
3	0	3	4	1
4	3	3	1	2

	1	2	3	4
1	0	3	0	1
2	3	0	4	4
3	0	4	0	3
4	1	4	3	2



Pixel Resolution and Gray Levels

- Texture features are influenced by the number of gray levels and pixel resolution of the image
- Optimization for each image dataset required
- Alternatively, features can be calculated for many resolutions

2Dt or 3Dt Features
Temporal Texture Features

- **Haralick texture features** describe the correlation in intensity of pixels that are next to each other in **space**.
  - These have been valuable for classifying static patterns.
- **Temporal texture features** describe the correlation in intensity of pixels in the same position in images next to each other over **time**.

Temporal Textures based on Co-occurrence Matrix

- Temporal co-occurrence matrix P:

  N_{level} by N_{level} matrix, Element P[i, j] is the probability that a pixel with value i has value j in the next image (time point).
- Thirteen statistics calculated on P are used as features

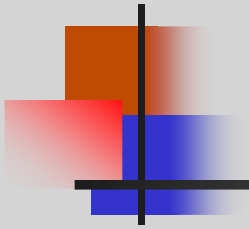


Image at t0

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Image at t1

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Temporal co-occurrence matrix (for image that does not change)

	1	2	3	4
1	3	0	0	0
2	0	9	0	0
3	0	0	6	0
4	0	0	0	7

Image at t0

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Image at t1

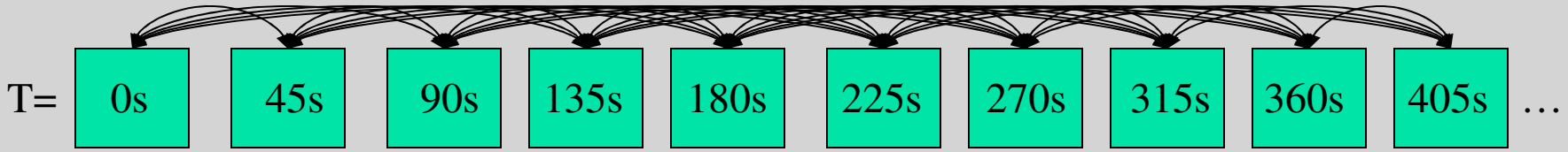
2	1	4	4	3
1	4	2	3	3
2	3	3	2	2
4	4	2	2	3
2	4	2	1	4

Temporal co-occurrence matrix (for image that changes)

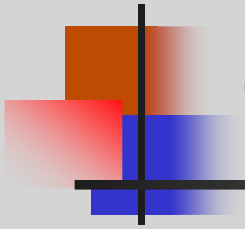
	1	2	3	4
1	1	0	2	0
2	2	1	1	5
3	0	5	0	1
4	0	3	3	1

Implementation of Temporal Texture Features

- Compare image pairs with different time interval, compute 13 temporal texture features for each pair.

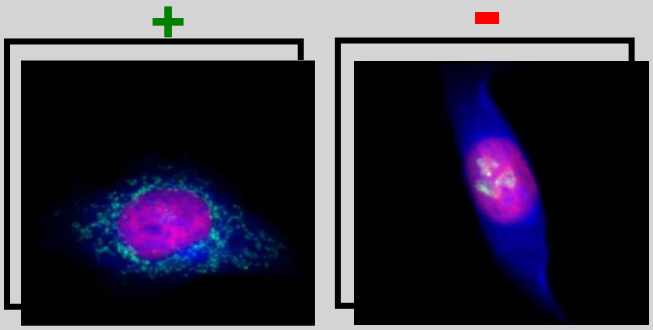


- Use the average and variance of features in each kind of time interval, yields 13*5*2=130 features

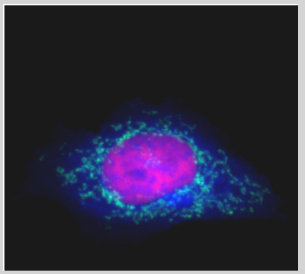


Machine Learning - Classification Methods

Simple two class problem

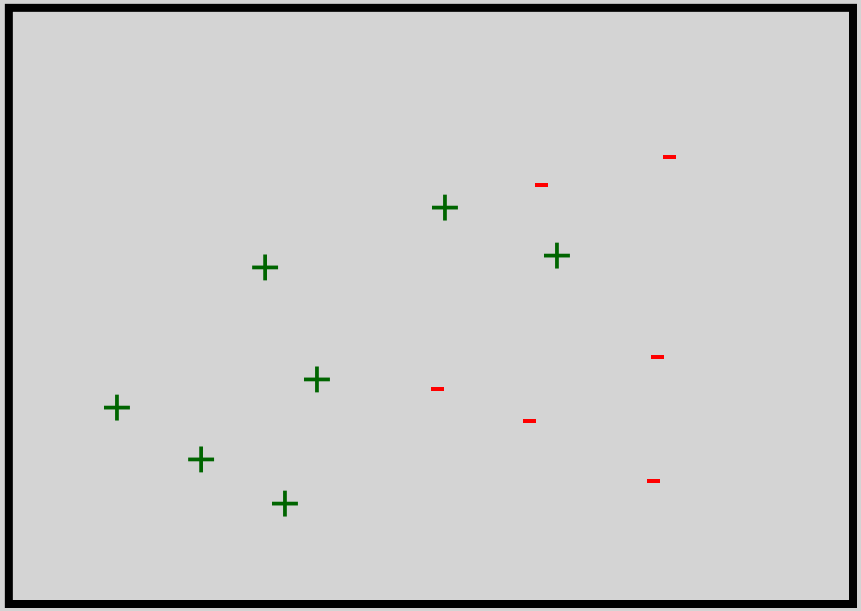


???



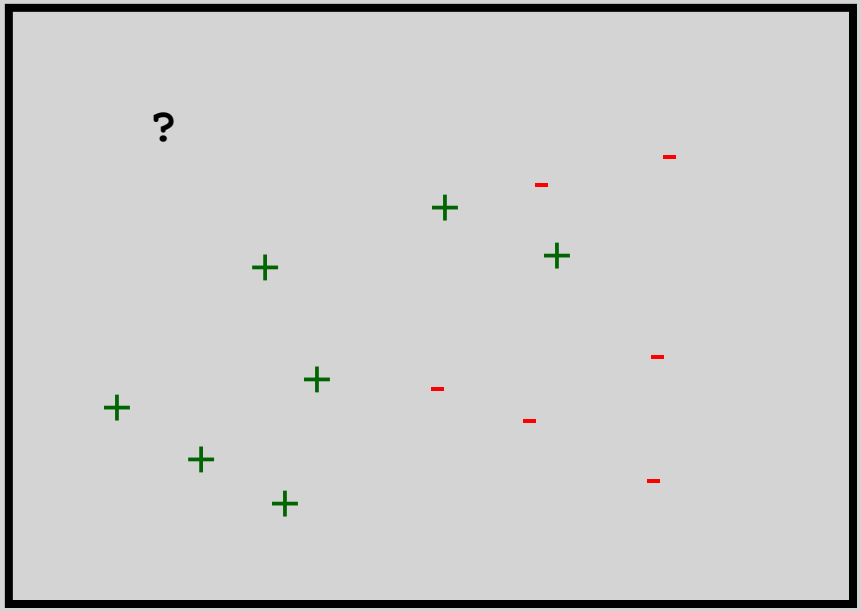
k-Nearest Neighbor (kNN)

- In feature space, training examples are



k-Nearest Neighbor (kNN)

- We want to label '**?**'

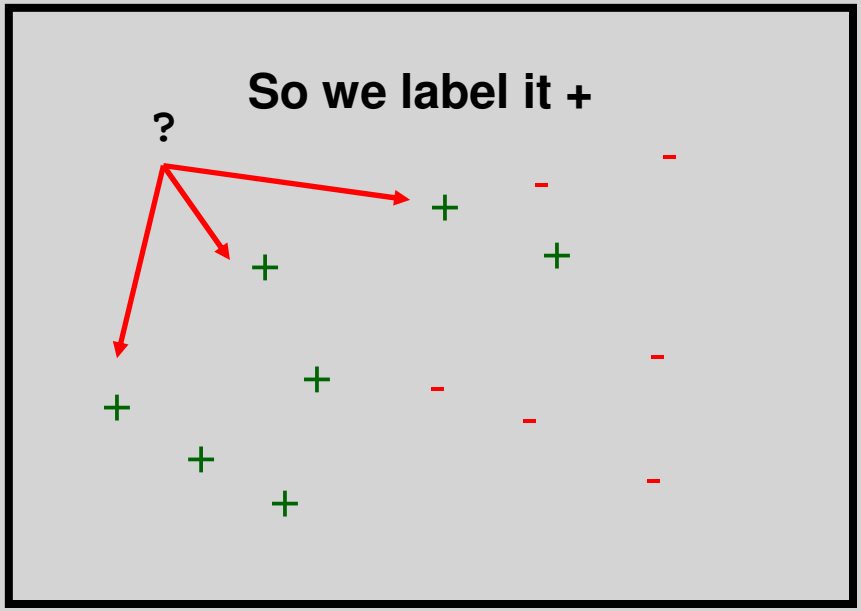


Feature #2
(e.g.., roundness)

Feature #1 (e.g.., 'area')

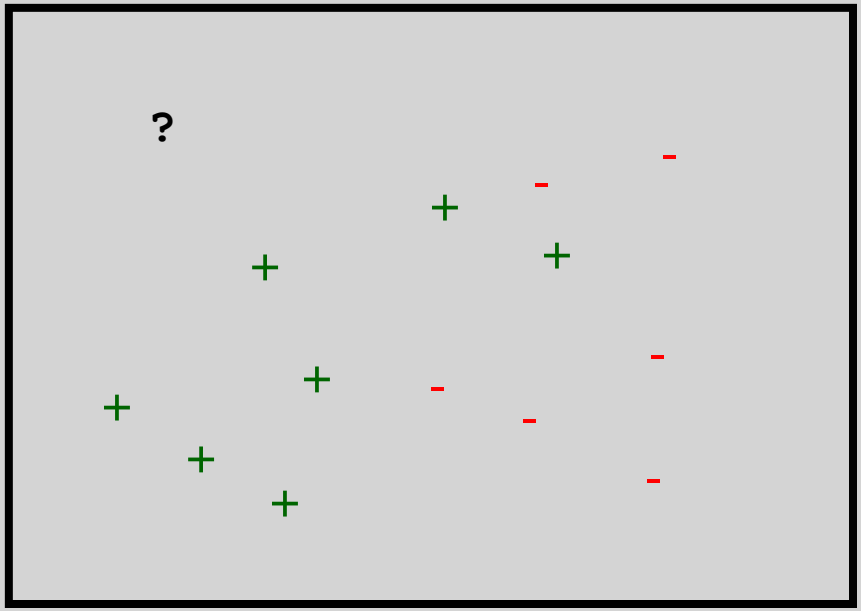
k-Nearest Neighbor (kNN)

- Find k nearest neighbors and vote



Decision trees

- Again we want to label '**?**'



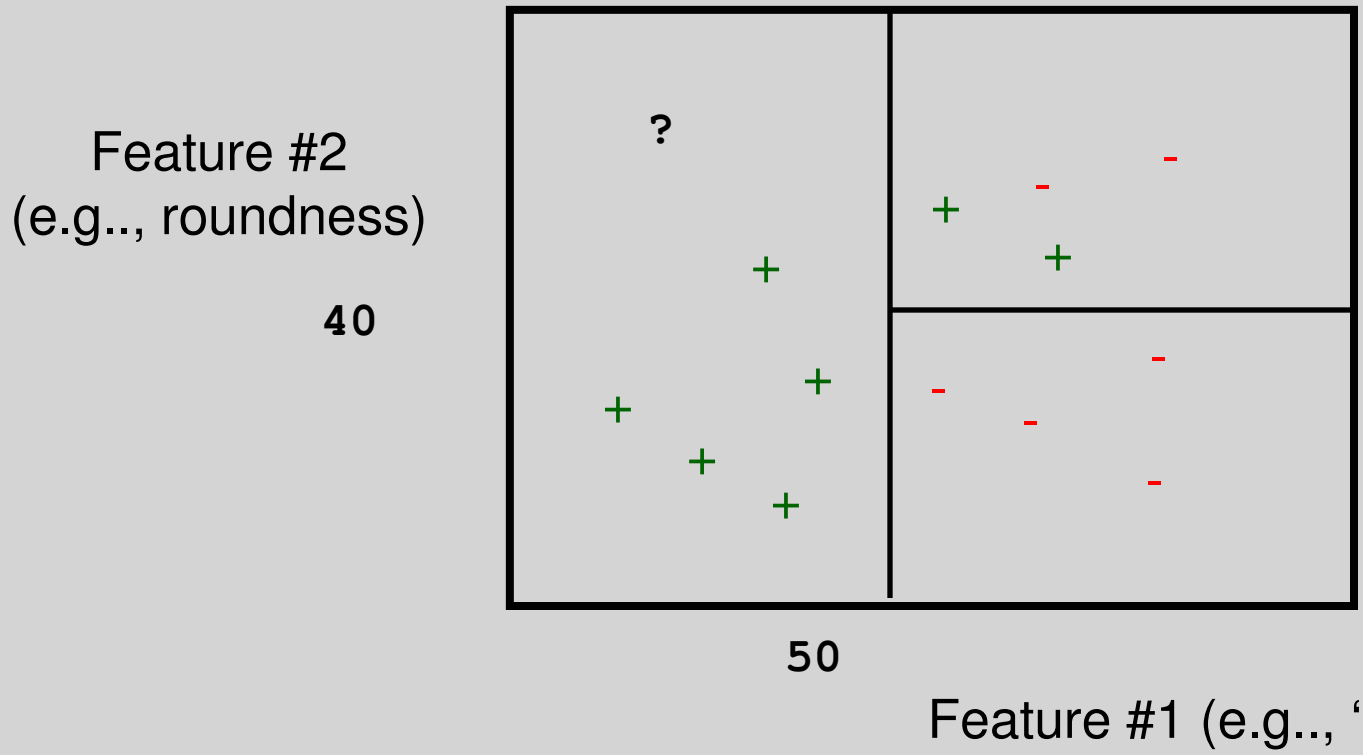
Feature #2
(e.g.., roundness)

Feature #1 (e.g.., 'area')

Slide courtesy of Christos Faloutsos

Decision trees

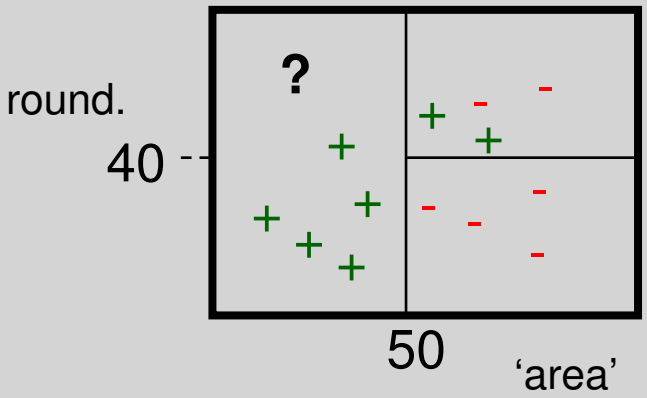
- so we build a decision tree:

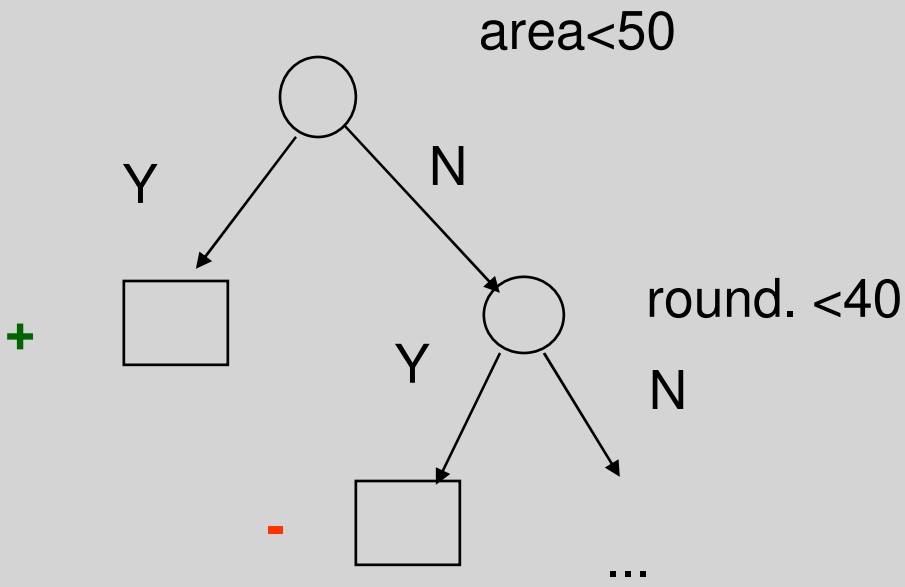


Slide courtesy of Christos Faloutsos

Decision trees

- so we build a decision tree:

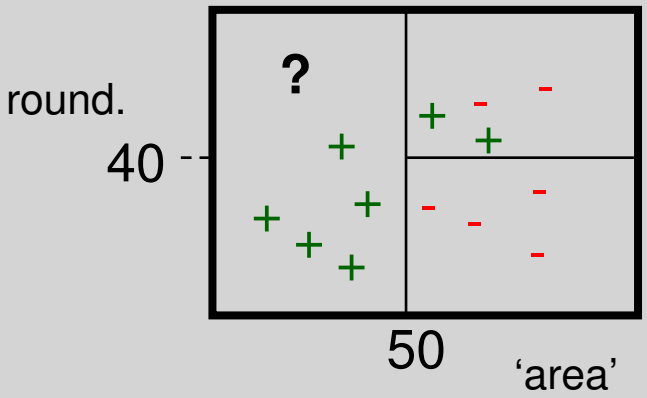


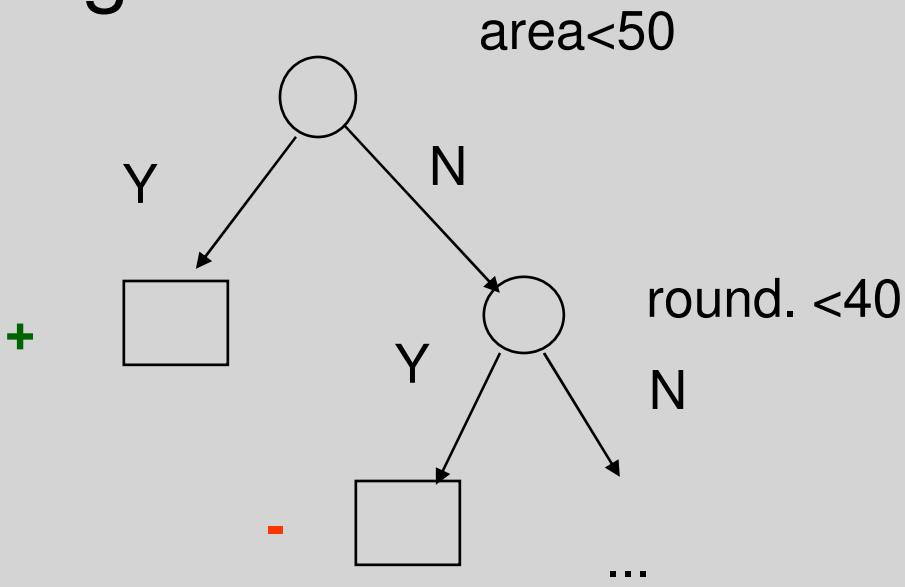


Slide courtesy of Christos Faloutsos

Decision trees

- Goal: split address space in (almost) homogeneous regions

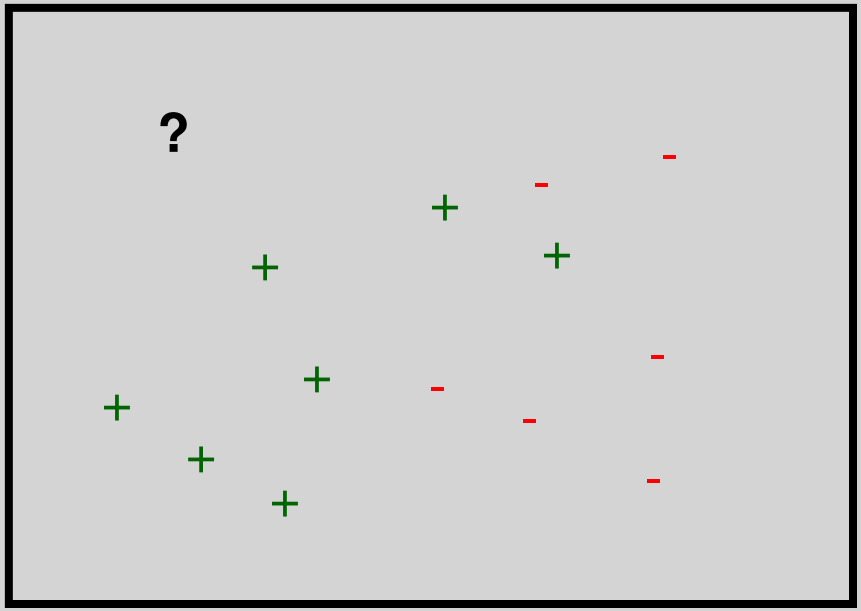




Slide courtesy of Christos Faloutsos

Support vector machines

- Again we want to label '**?**'



Feature #2
(e.g.., roundness)

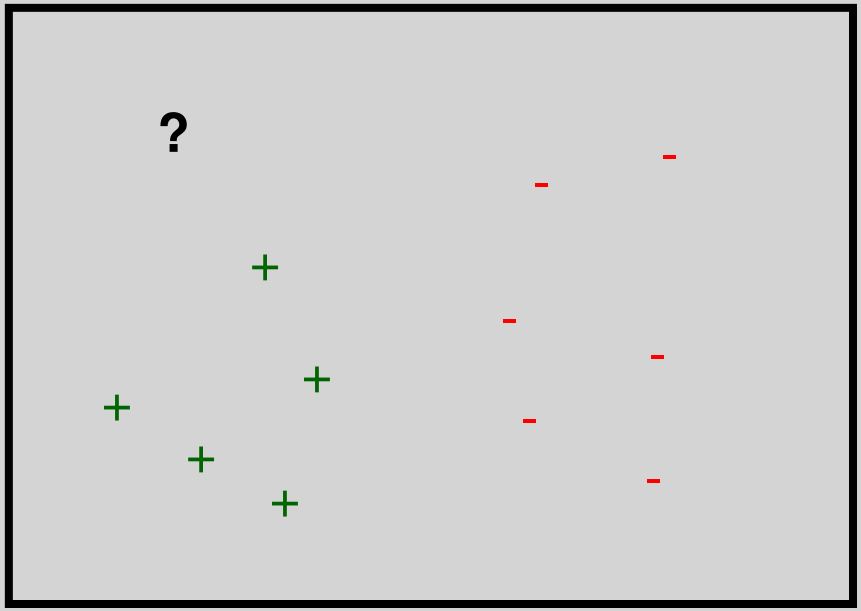
Feature #1 (e.g.., 'area')

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- Use single linear separator??

round.



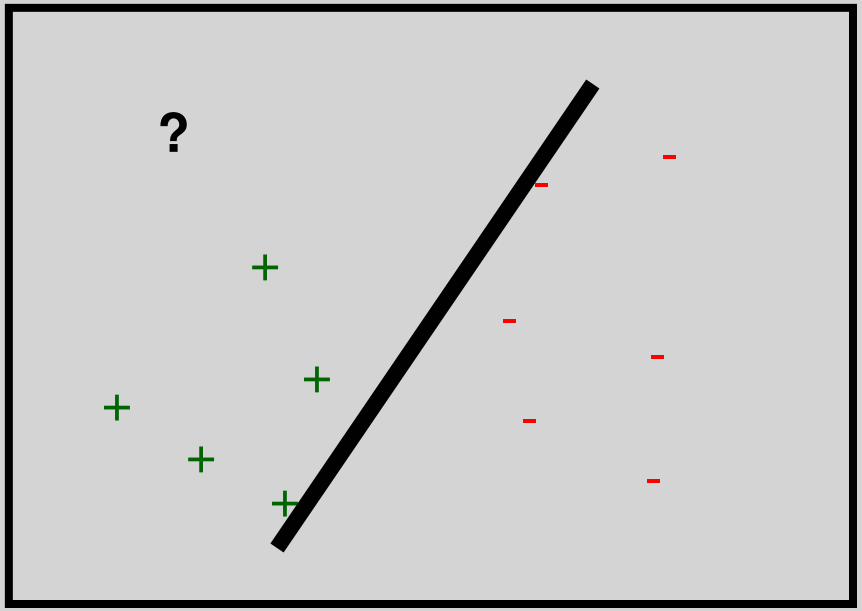
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- Use single linear separator??

round.



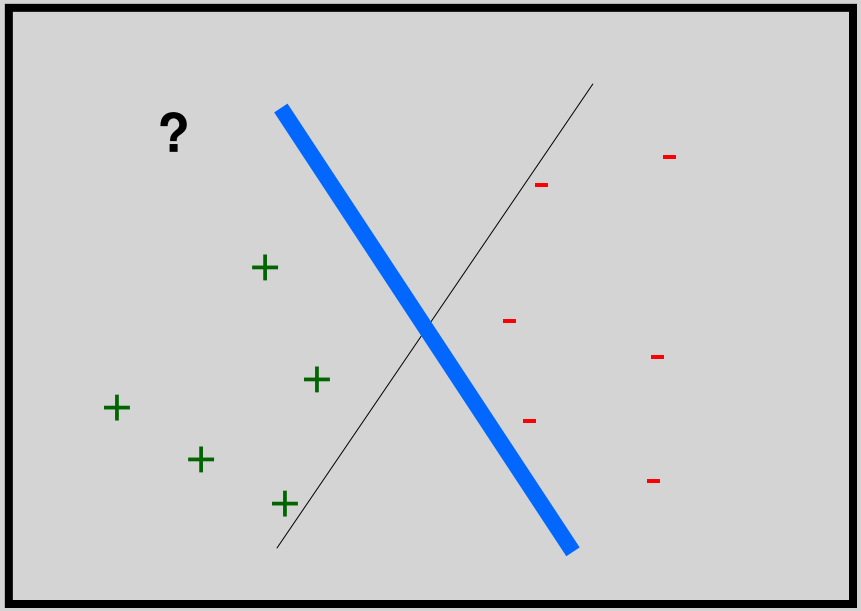
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- Use single linear separator??

round.



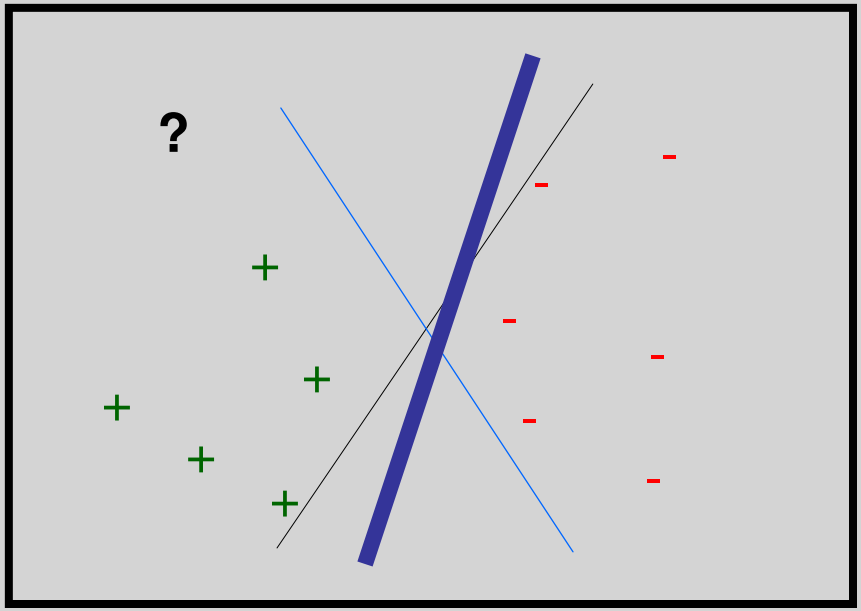
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- Use single linear separator??

round.



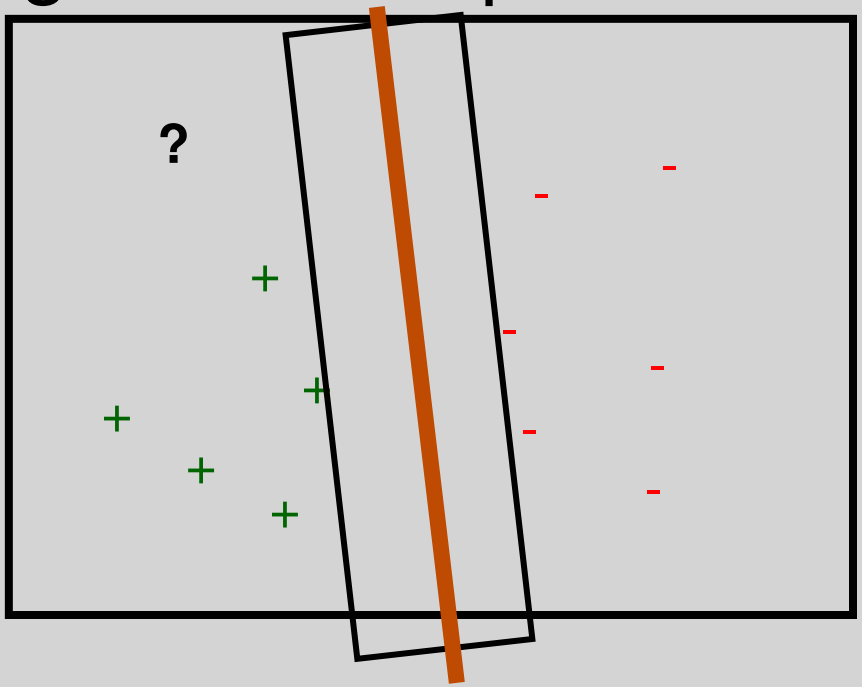
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- Use single linear separator??

round.



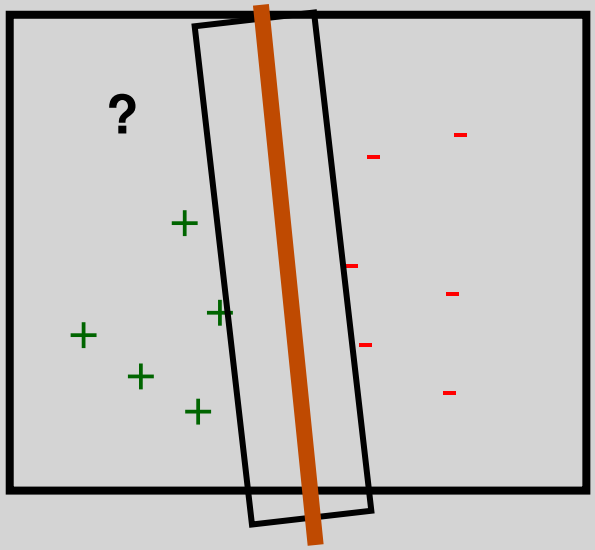
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??
- A: the one with the widest corridor!

round.



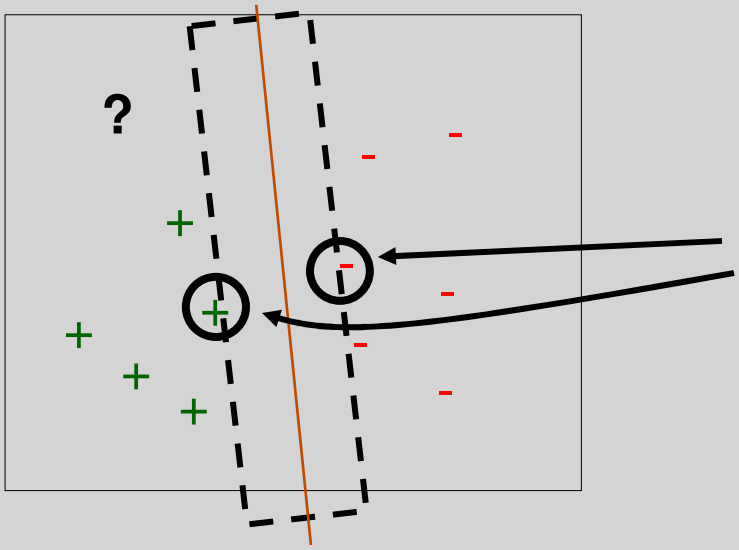
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??
- A: the one with the widest corridor!

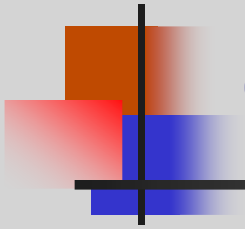
round.



'support vectors'

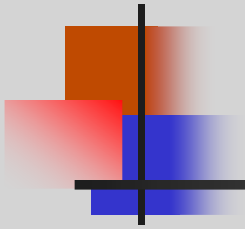
area

Slide courtesy of Christos Faloutsos



Cross-Validation

- If we train a classifier to minimize error on a set of data, have no ability to generalize error that will be seen on new dataset
- To calculate *generalizable* accuracy, we use *n*-fold cross-validation
- Divide images into *n* sets, train using n-1 of them and test on the remaining set
- Repeat until each set is used as test set and average results across all trials



Describing classifier errors

- For multi-class classifiers, typically report
  - Accuracy = $\frac{\text{\# test images correctly classified}}{\text{\# test images}}$
- For binary classifiers (positive or negative), define
  - TP = true positives, FP = false positives
  - TN = true negatives, FN = false negatives
  - Recall = TP / (TP + FN)
  - Precision = TP / (TP + FP)
  - F-measure= 2*Recall*Precision/(Recall + Precision)

Murphy et al 2000;
Boland & Murphy 2001;
Huang & Murphy 2004

2D Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	**99**	1	0	0	0	0	0	0	0	0
ER	0	**97**	0	0	0	2	0	0	0	1
Gia	0	0	**91**	7	0	0	0	0	2	0
Gpp	0	0	14	**82**	0	0	2	0	1	0
Lam	0	0	1	0	**88**	1	0	0	10	0
Mit	0	3	0	0	0	**92**	0	0	3	3
Nuc	0	0	0	0	0	0	**99**	0	1	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	1	0	0	12	2	0	1	**81**	2
Tub	1	2	0	0	0	1	0	0	1	**95**

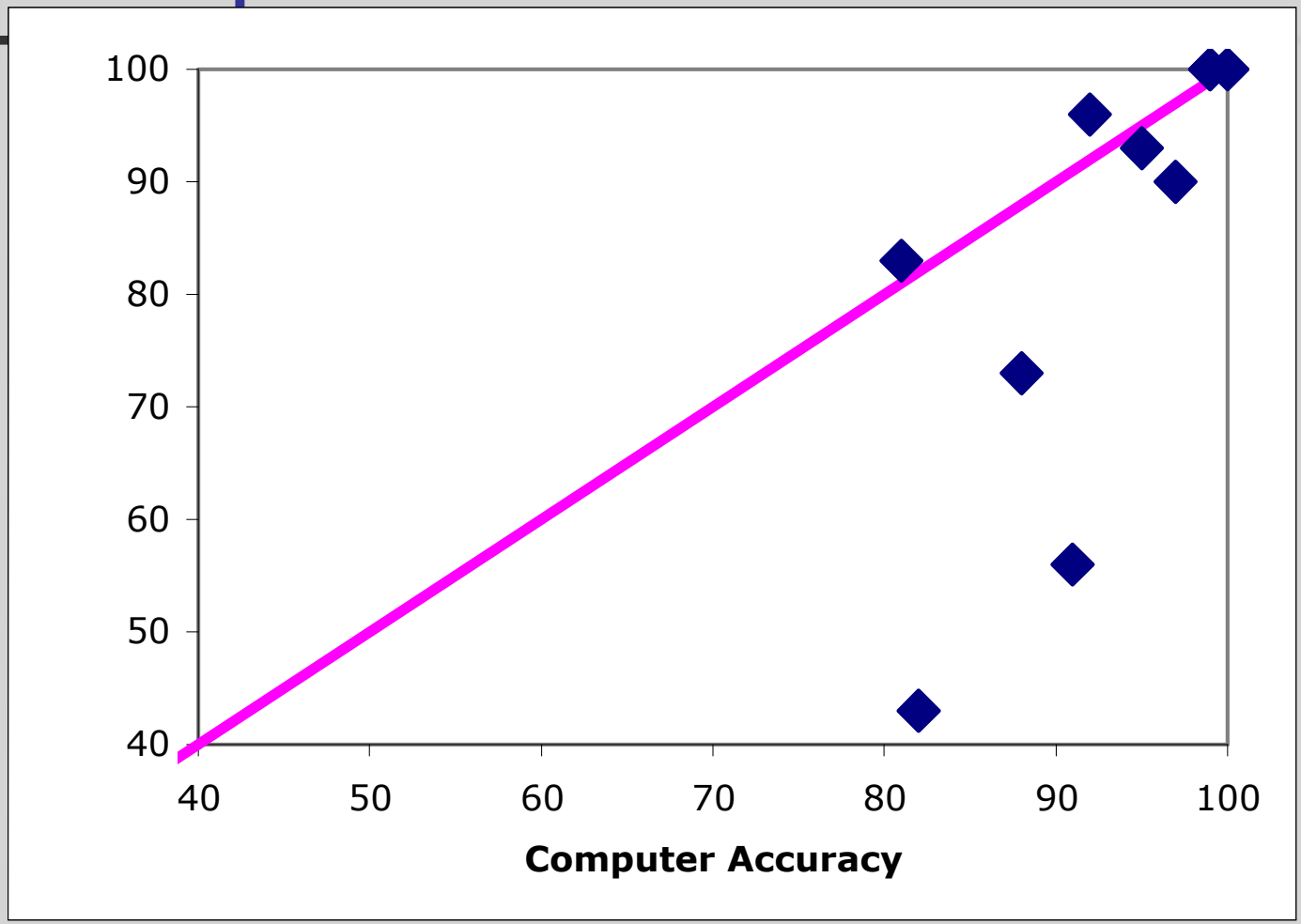
Overall accuracy = 92%

Human Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	**100**	0	0	0	0	0	0	0	0	0
ER	0	**90**	0	0	3	6	0	0	0	0
Gia	0	0	**56**	36	3	3	0	0	0	0
Gpp	0	0	54	**33**	0	0	0	0	3	0
Lam	0	0	6	0	**73**	0	0	0	20	0
Mit	0	3	0	0	0	**96**	0	0	0	3
Nuc	0	0	0	0	0	0	**100**	0	0	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	13	0	0	3	0	0	0	**83**	0
Tub	0	3	0	0	0	0	0	3	0	**93**

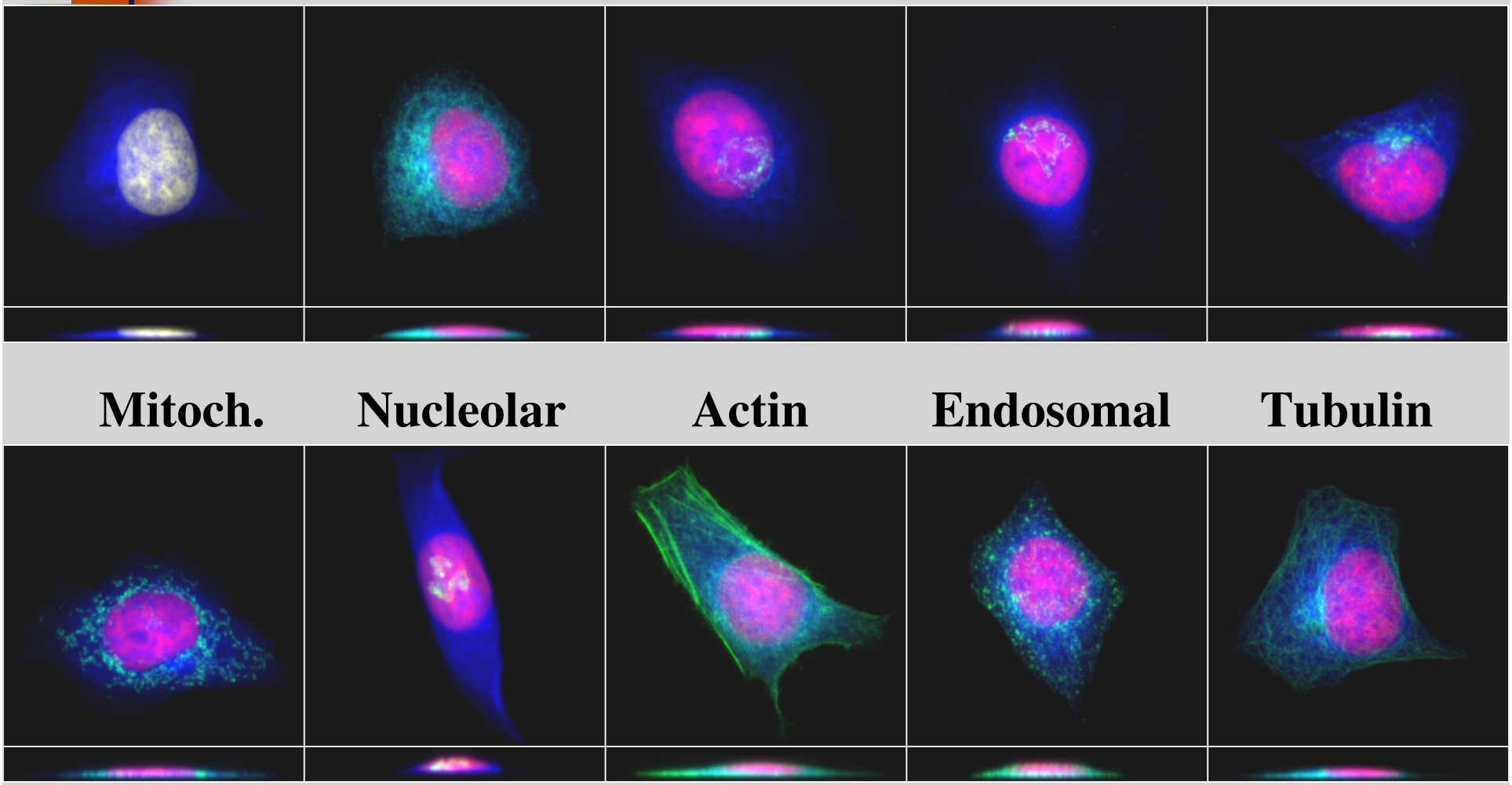
Overall accuracy = 83%

Computer vs. Human



Velliste & Murphy 2002

3D HeLa cell images



Images collected using facilities at the Center for Biologic Imaging courtesy of Simon Watkins

Carnegie Mellon

Velliste & Murphy 2002; Chen & Murphy 2004

3D Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	**98**	2	0	0	0	0	0	0	0	0
ER	0	**100**	0	0	0	0	0	0	0	0
Gia	0	0	**100**	0	0	0	0	0	0	0
Gpp	0	0	0	**96**	4	0	0	0	0	0
Lam	0	0	0	4	**95**	0	0	0	0	2
Mit	0	0	2	0	0	**96**	0	2	0	0
Nuc	0	0	0	0	0	0	**100**	0	0	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	0	0	0	2	0	0	0	**96**	2
Tub	0	2	0	0	0	0	0	0	0	**98**

Overall accuracy = 98%

Conclusions (1996-2004)

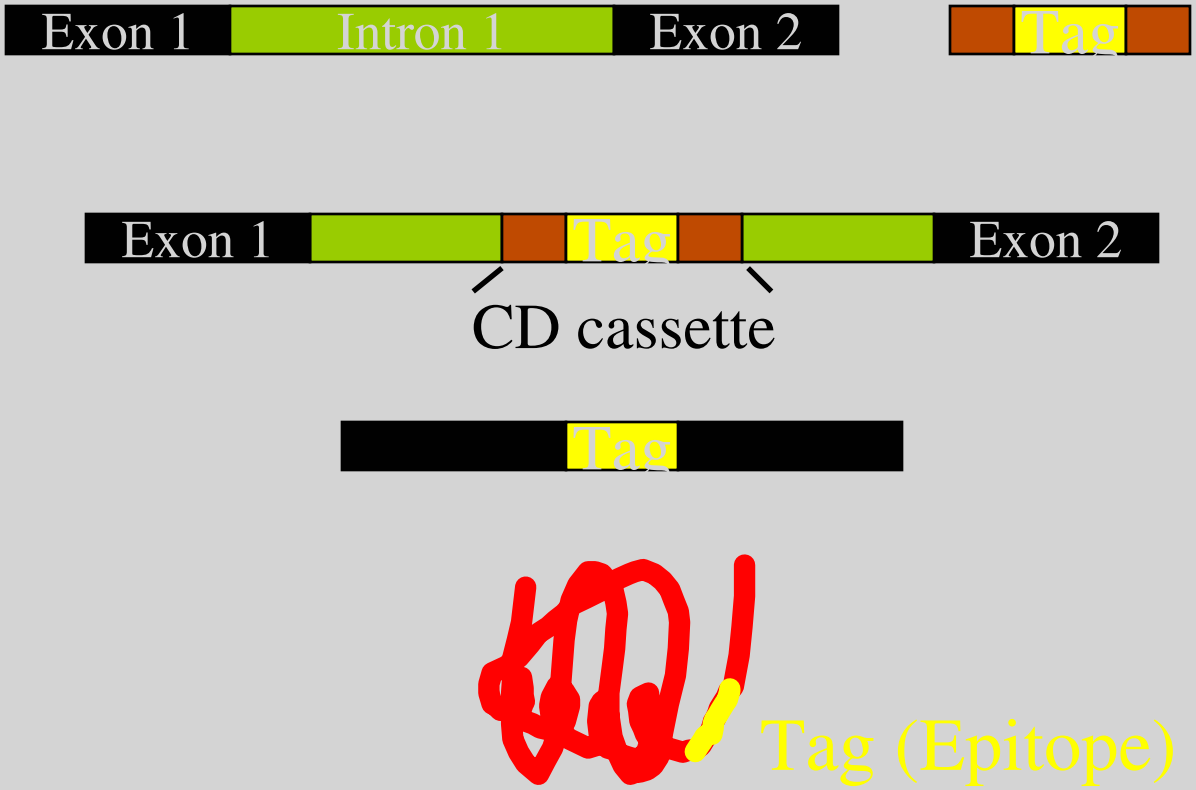
- Automated classification of subcellular patterns possible without colocalization
- Accuracy better than visual examination
  - Similar for basic patterns
  - Better for similar patterns
- 3D images give better accuracy than 2D
- **>> SLFs capture essence of patterns**

Unsupervised Learning to Identify High-Resolution Protein Patterns

Location Proteomics

- **Tag** many proteins
  - We have used **CD-tagging** (developed by Jonathan Jarvik and Peter Berget): Infect population of cells with a retrovirus carrying DNA sequence that will "tag" in a random gene

Principles of CD-Tagging (Jarvik & Berget) (CD = Central Dogma)



Genomic DNA + CD-cassette

Tagged DNA

Tagged mRNA

Tagged Protein

Location Proteomics

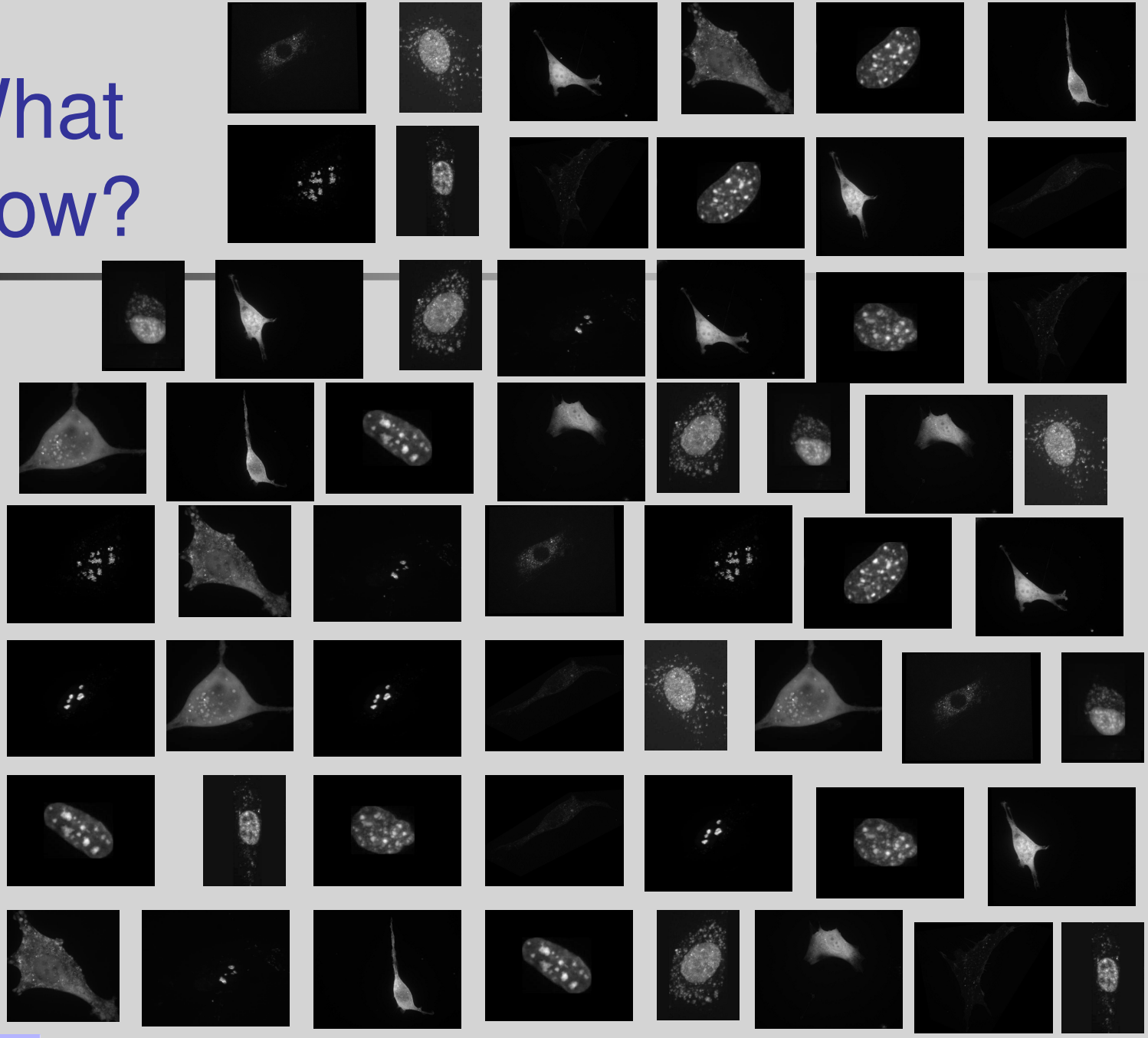
- **Tag** many proteins
  - We have used **CD-tagging** (developed by Jonathan Jarvik and Peter Berget): Infect population of cells with a retrovirus carrying DNA sequence that will "tag" in a random gene

Jarvik et al 2002

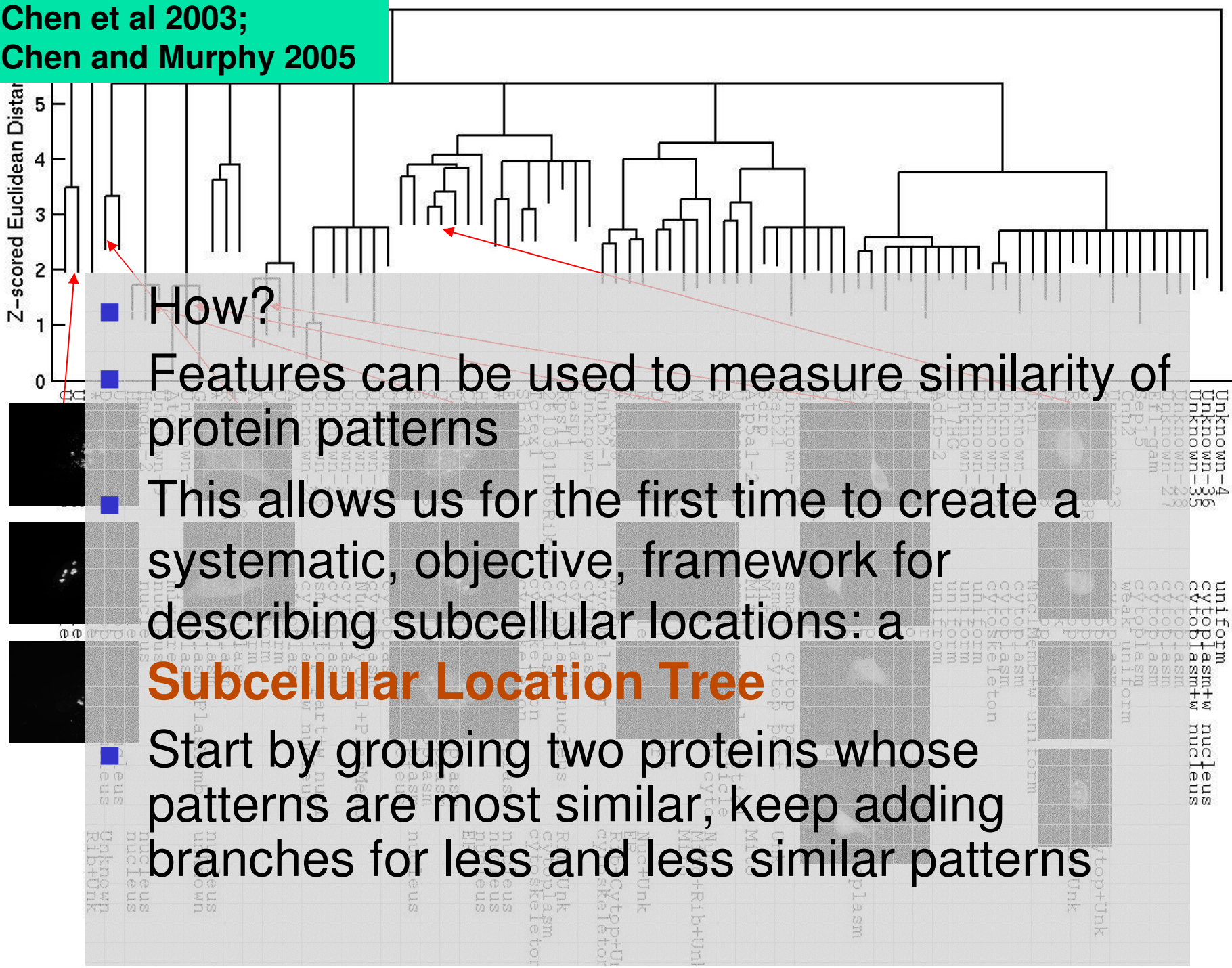
- Isolate separate **clones**, each of which produces express one tagged protein
- Use RT-PCR to **identify tagged gene** in each clone
- Collect **many live cell images** for each clone using spinning disk confocal fluorescence microscopy

What Now?

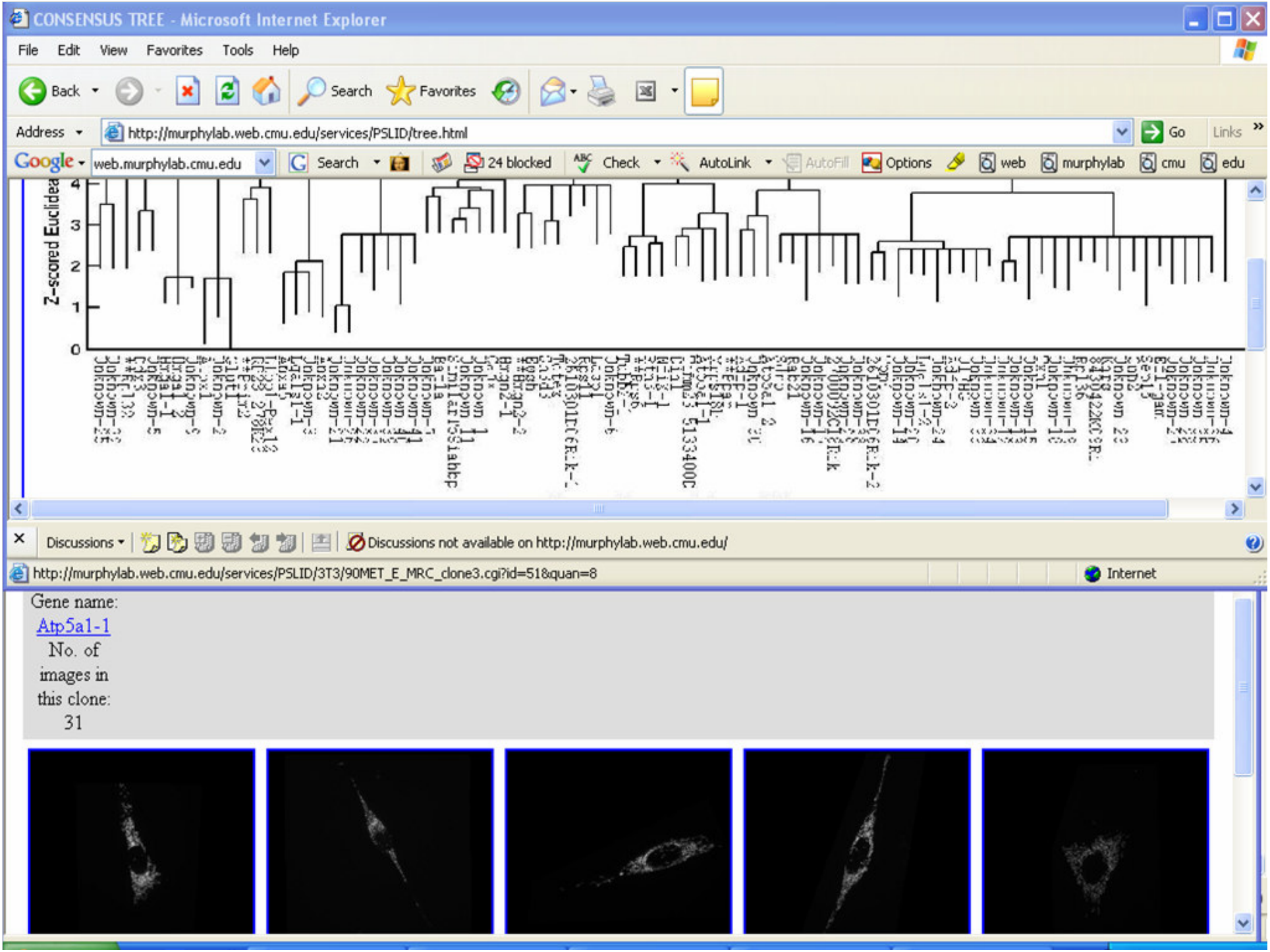
Group ~90 tagged clones by pattern



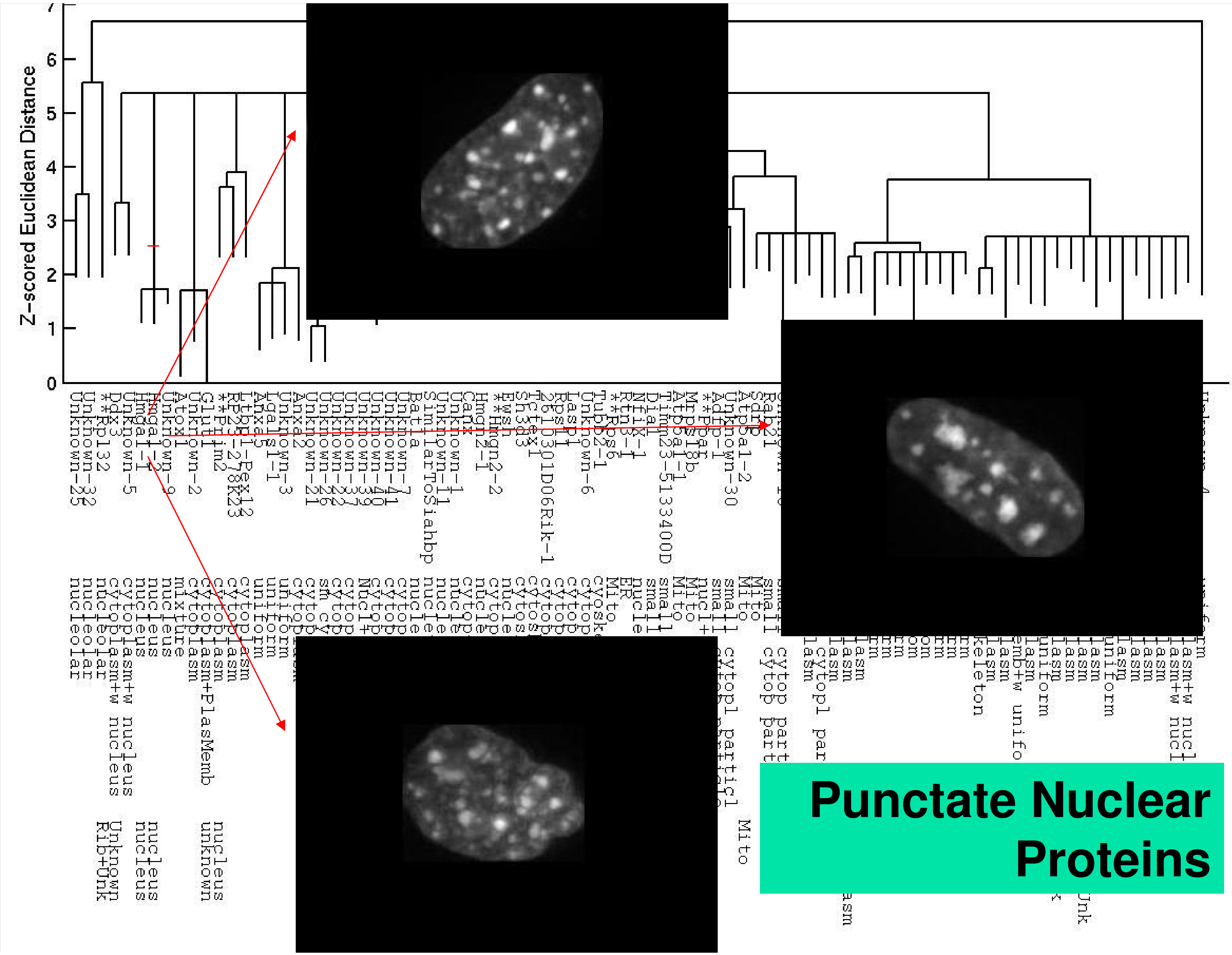
Chen et al 2003;
Chen and Murphy 2005



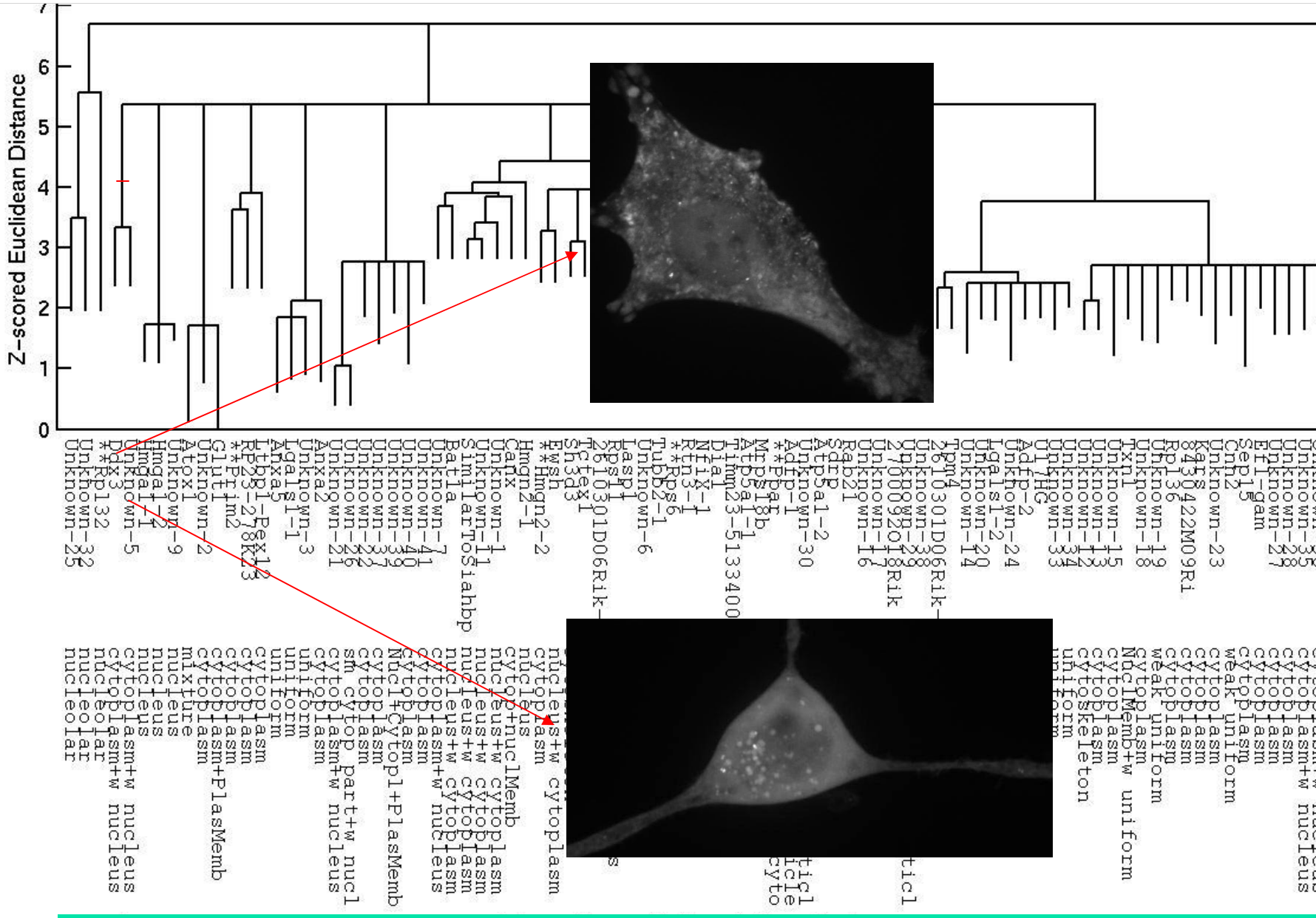
- How?
- Features can be used to measure similarity of protein patterns
- This allows us for the first time to create a systematic, objective, framework for describing subcellular locations: a **Subcellular Location Tree**
- Start by grouping two proteins whose patterns are most similar, keep adding branches for less and less similar patterns


CONSENSUS TREE - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://murphylab.web.cmu.edu/services/PSLID/tree.html
Go Links
Google web.murphylab.cmu.edu Search 24 blocked Check AutoLink AutoFill Options web murphylab cmu edu
Z-scored Euclidean
0
1
2
3
4
Discussions Discussions not available on http://murphylab.web.cmu.edu/
http://murphylab.web.cmu.edu/services/PSLID/3T3/90MET_E_MRC_clone3.cgi?id=51&quan=8
Internet
Gene name:
Atp5a1-1
No. of images in this clone:
31

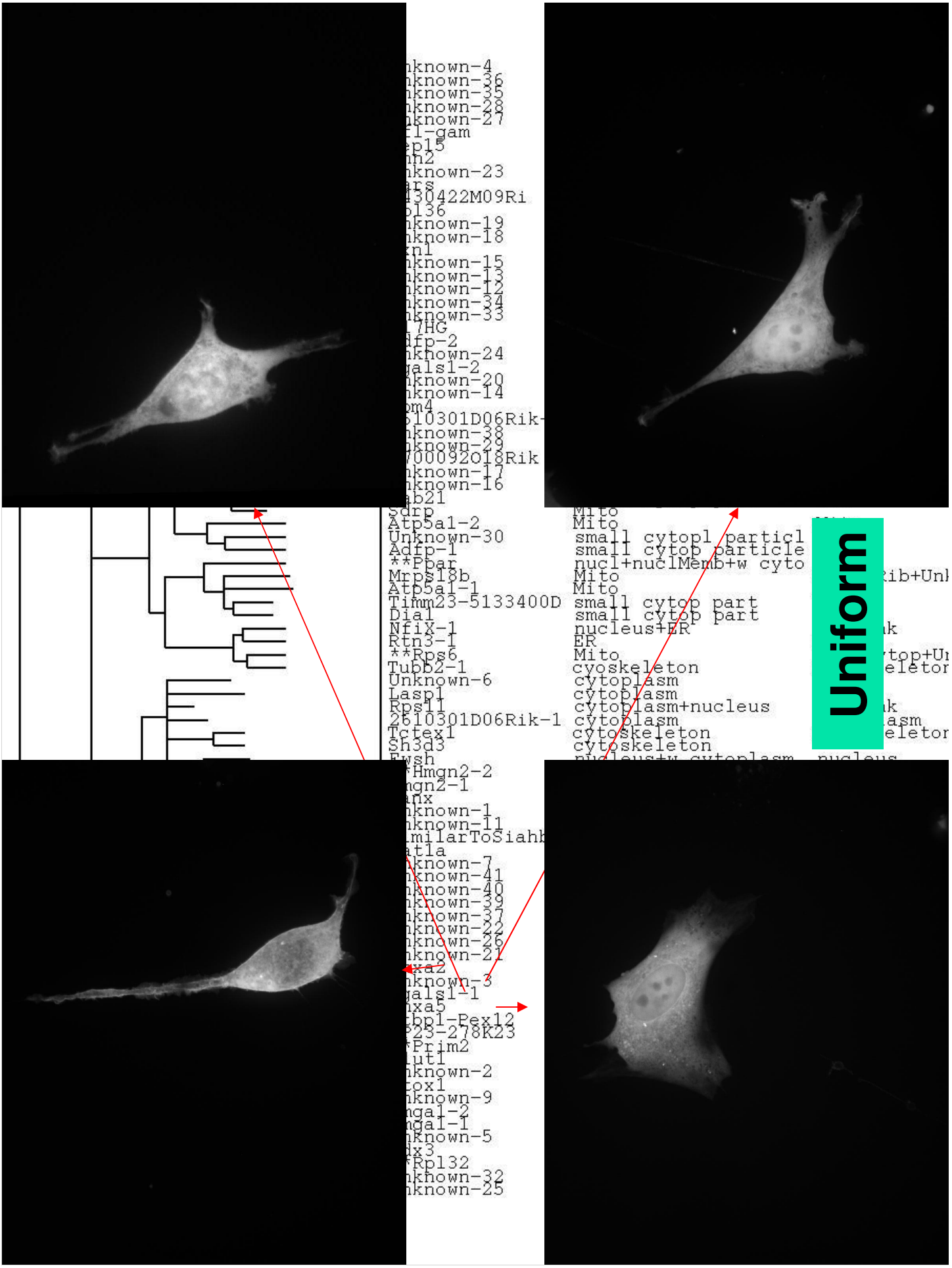
Z-scored Euclidean Distance
0
1
2
3
4
5
6
7
Unknown-25
Unknown-32
Unknown-5
Unknown-9
Unknown-2
Unknown-7
Unknown-41
Unknown-40
Unknown-39
Unknown-22
Unknown-26
Unknown-21
Unknown-3
Unknown-1
Unknown-11
Unknown-10
Unknown-6
Unknown-30
Unknown-24
Unknown-38
Unknown-20
Unknown-14
Unknown-19
Unknown-18
Unknown-15
Unknown-13
Unknown-12
Unknown-34
Unknown-33
Unknown-16
Unknown-4
Unknown-36
Unknown-35
Unknown-27
Unknown-23
Unknown-29
Unknown-28
Unknown-31
Unknown-17
Unknown-8
Unknown-37
Unknown-42
Unknown-43
Unk
Mito
Nuc
Mito+Rib+Unl
Mito
Nuc+Unk
ER
Rib+Cytopp+Ul
cytoskeleton
Rib+Unk
cytoplasm
cytoskeletor
Nucleolar Proteins


Z-scored Euclidean Distance
0
1
2
3
4
5
6
7
Punctate Nuclear Proteins

Predominantly
Nuclear
Proteins with
Some Punctate
Cytoplasmic
Staining

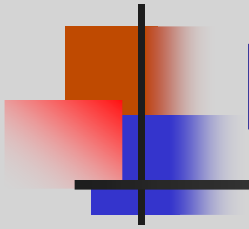


Nuclear and Cytoplasmic Proteins with Some Punctate Staining


Uniform

Generative Models for Subcellular Location Patterns

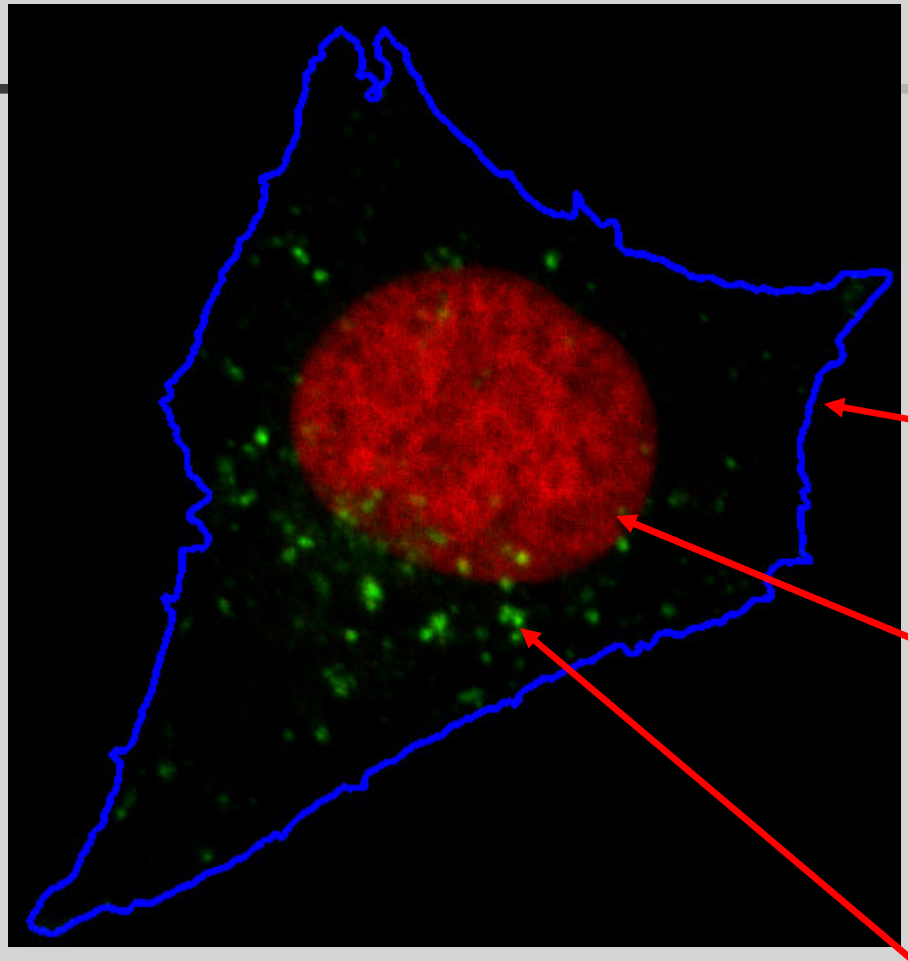
Carnegie Mellon



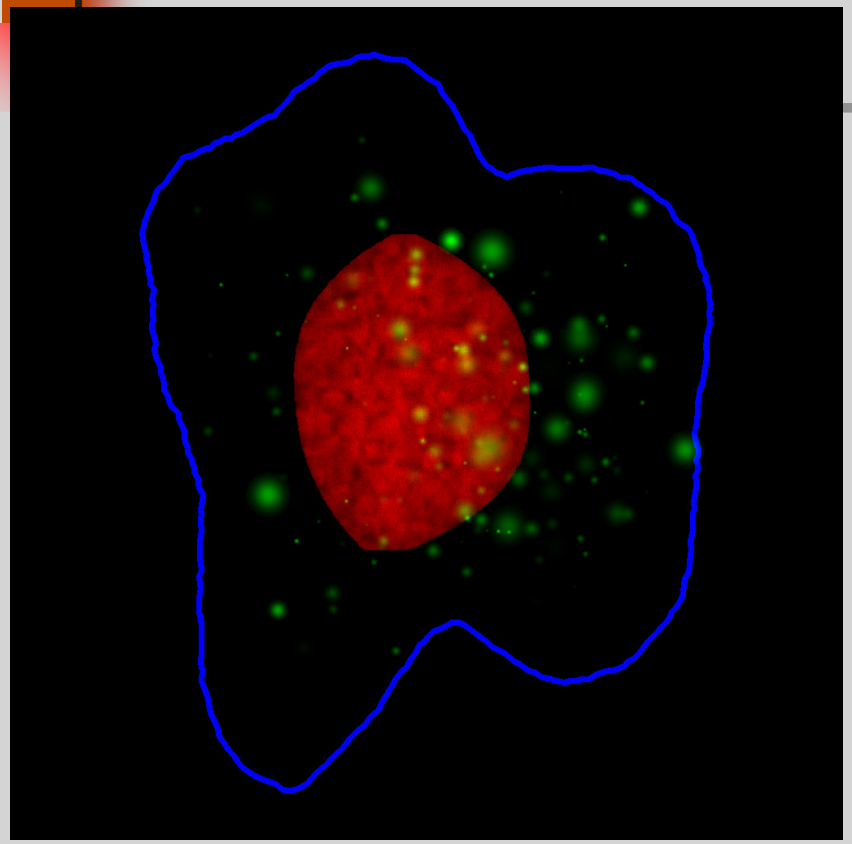
Need

- How do we communicate results of clustering patterns?
- Show all images from a given cluster?
  - Long download
  - No ability to generalize
- Proposal: Use generative models

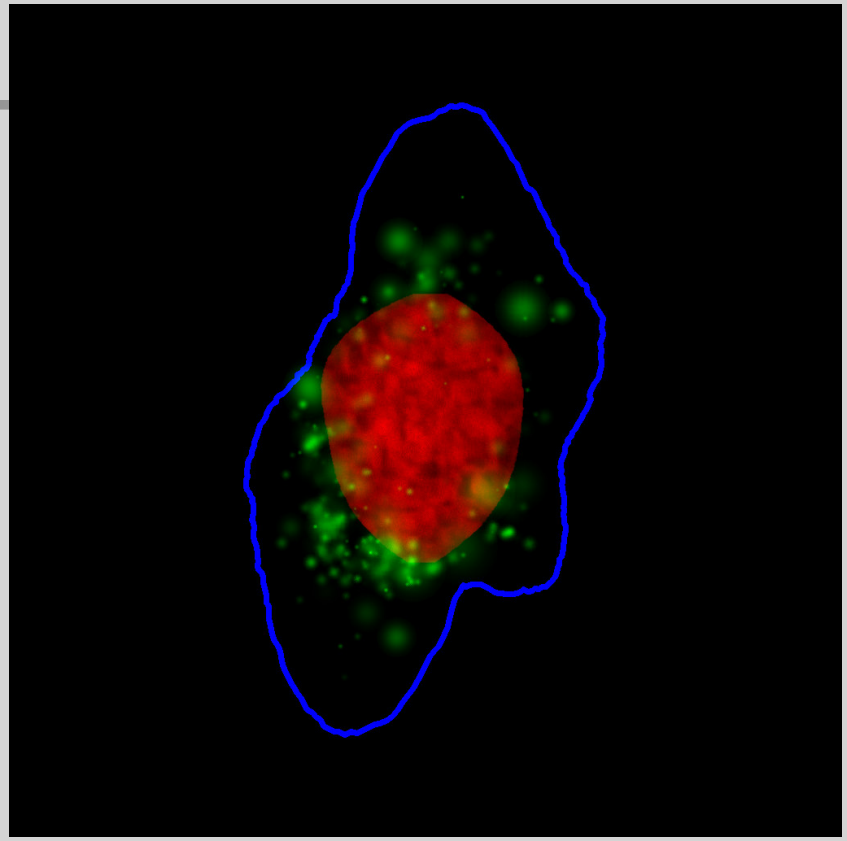
LAMP2 pattern



Synthesized Images

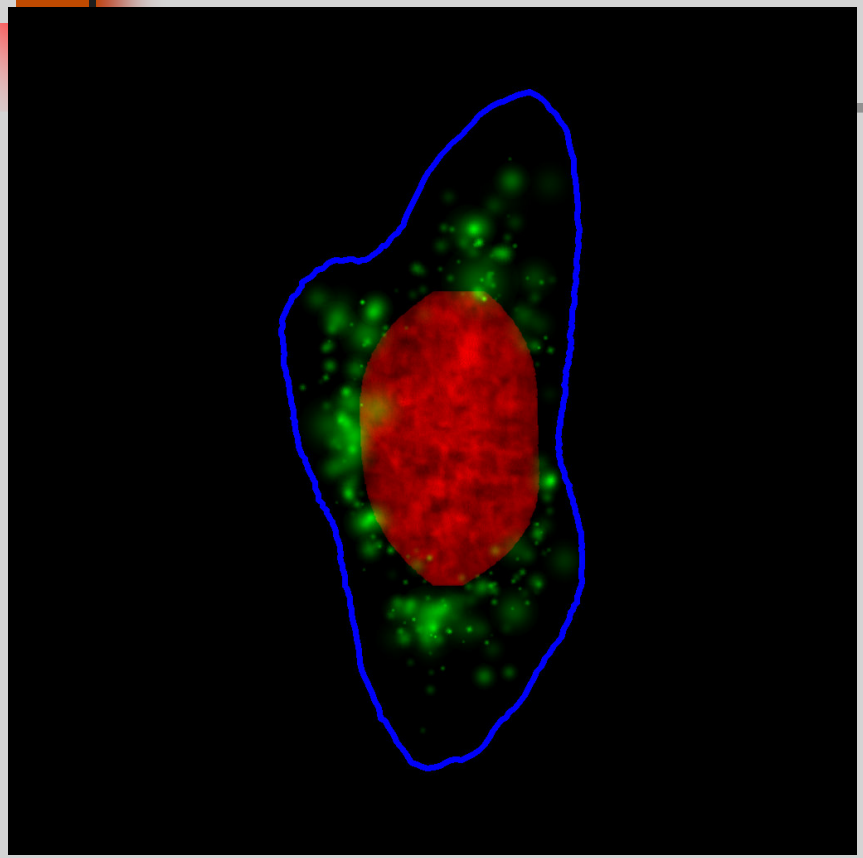


Lysosomes

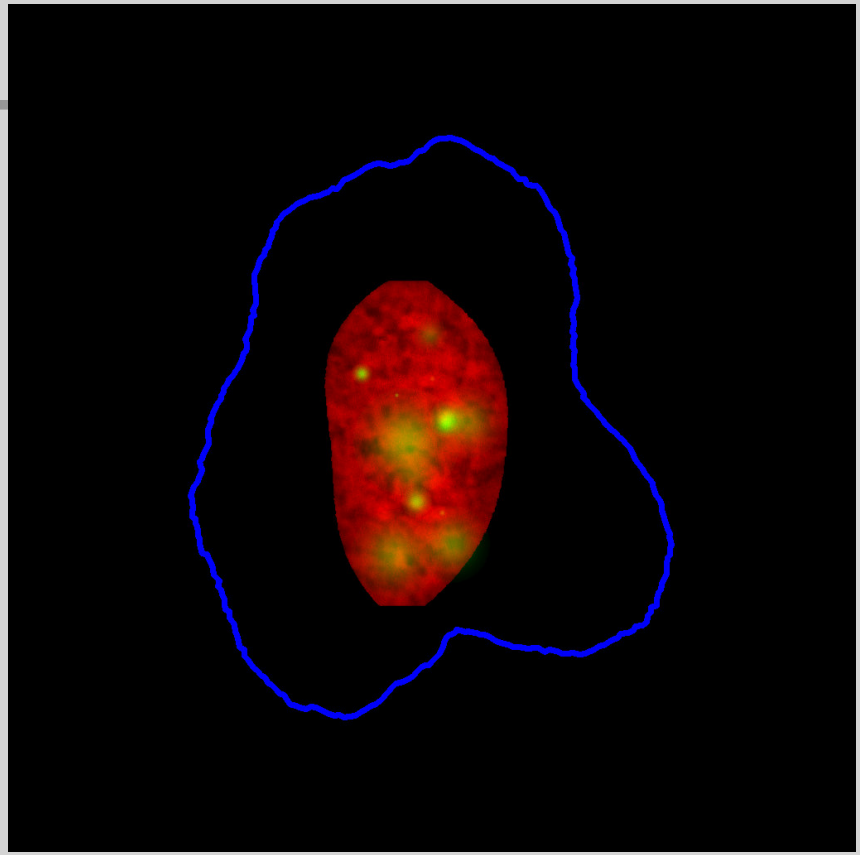


Endosomes

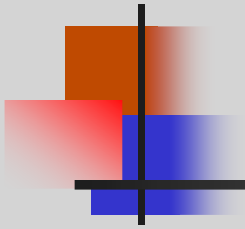
Synthesized Images



Mitochondria



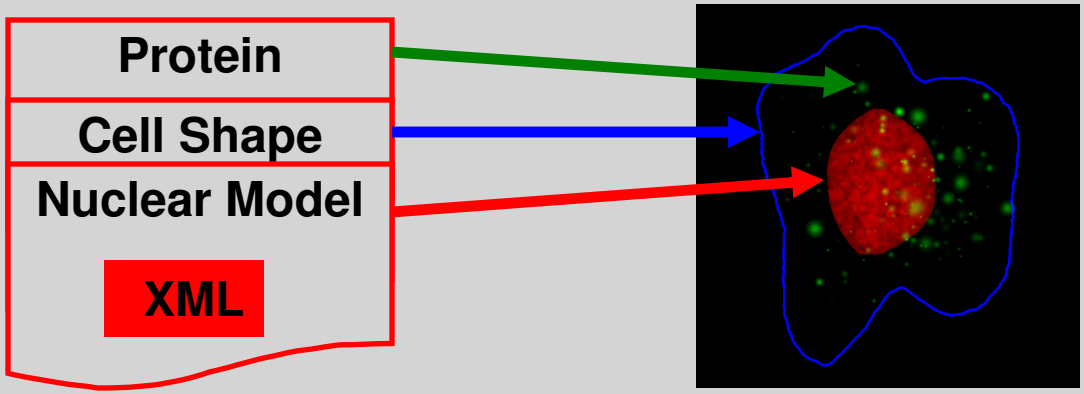
Nucleoli



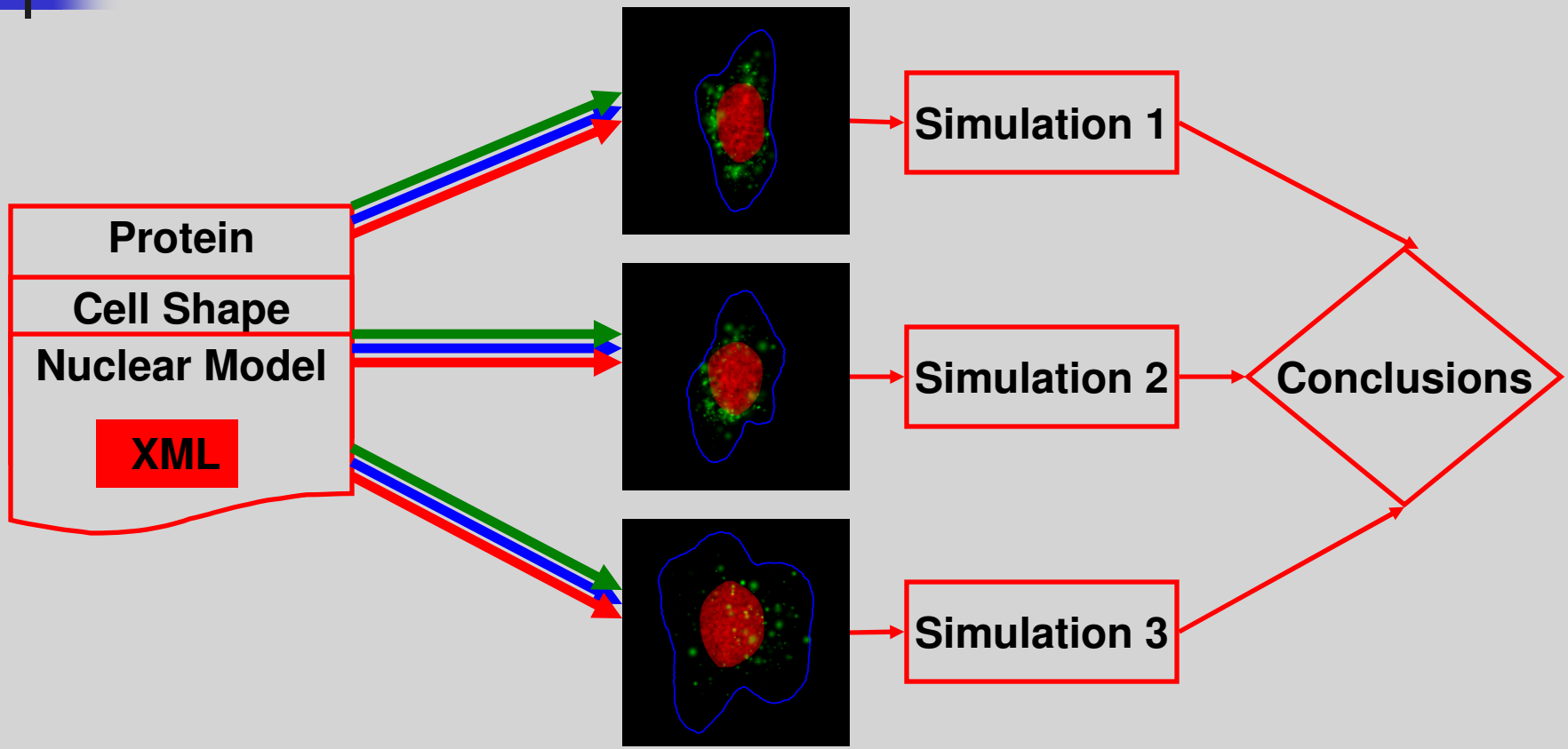
Model Distribution

- Generative models provide better way of distributing what is known about “subcellular location families” (or other imaging results, such as illustrating change due to drug addition)
- Have initial XML design for capturing the models for distribution
- Have portable tool for generating images from the model

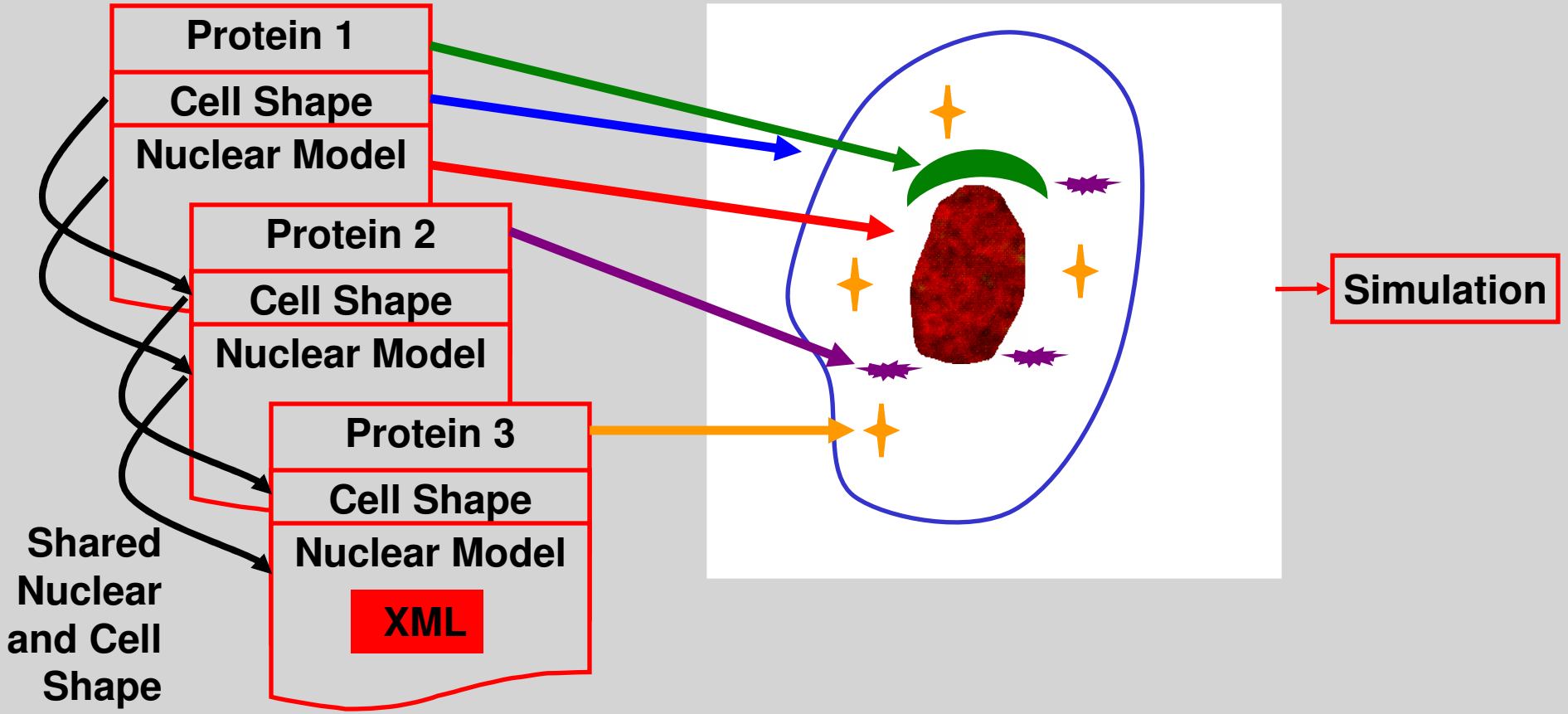
Generation Process



Generating Multiple Distributions for Simulations



Combining Models for Cell Simulations



The Protein Subcellular Location Image Database (PSLID)

Carnegie Mellon

Huang et al 2002; Huang et al 2007

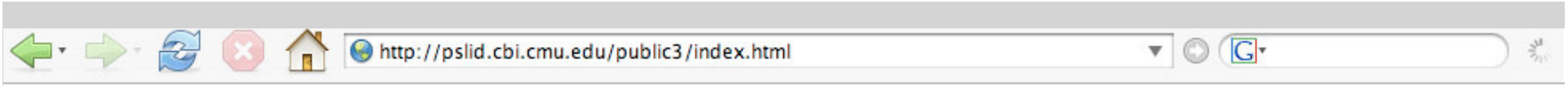
PSLID: Protein Subcellular Location Image Database

- A publicly accessible image database at http://pslid.cbi.cmu.edu
  - Version 3 released February 2, 2007
  - 2D and 3D images (single cell regions defined)
  - Two cell types, HeLa and 3T3
  - Over 120,000 images/3000 unique fields/14,000 cells
  - 111classes; 55 known proteins; 11 targeting mutants of a single protein
  - Programmatic search via URL

Huang et al 2002; Huang et al 2007

PSLID: Protein Subcellular Location Image Database

- A downloadable open source system for creating local databases
  - Version 3 of software released February 13, 2007
  - Focused on subcellular pattern analysis
  - SLF features integrated into database
  - Integrated comparison, classification, clustering tools
  - Designed for high-throughput microscopy
  - Interface to OME in the works
  - Large ITR project with UCSB for distributed system



Murphy Lab PSLID Service

PSLID stands for Protein Subcellular Localization Image Database. PSLID collects and structures 2-D through 5-D fluorescence microscope images, annotations, and derived features in a relational schema.

It is designed so that interpretations as well as annotations can be queried. The annotations in PSLID, composed of 44 linked tables with publicly available descriptions, provide a thorough description of sample preparation and fluorescence microscope imaging.

Image interpretation is achieved using Subcellular Location Features that have been shown to be capable of recognizing all major subcellular structures and of resolving patterns that cannot be distinguished by eye.

The fundamental unit of PSLID is an *image set*, which is simply a logical grouping of images. Image sets can be defined at the time of image loading, or they can be defined by searching for images that meet specified criteria (e.g., all images of "actin" or all images that are similar to a query image). They can also be created by analysis functions such as cluster analysis (e.g., the images in each cluster found by cluster analysis can be put into distinct sets).

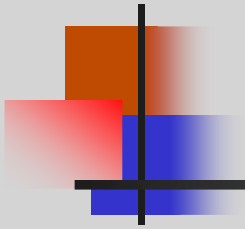
Analysis capabilities that are incorporated in PSLID include:

- *Searching* for images by context (annotations) or content
- *Ranking images by typicality* within a set
  - e.g., to choose an image for presentation or publication
- *Ranking images by similarity* to one or more query images
  - "searching by image content" or "relevance feedback"
- *Comparing* two sets of images (hypothesis testing)
  - e.g., to determine whether a drug alters the distribution of a tagged protein
- *Training a classifier* to recognize subcellular patterns
- *Using a trained classifier* to assign images to pattern classes
  - e.g., assigning images to "positive" or "negative"
- Clustering images by their subcellular patterns
  - e.g., finding "subcellular location families" within a large set of images

You can go to the Quick Start page to see instructions for PSLID installation, image loading, and image analysis using PSLID.

The public PSLID database currently contains a number of large image collections. It can be accessed interactively or via queries embedded in URLs. We encourage the submission to PSLID of other image collections documenting the subcellular location of proteins to facilitate "one-stop" searching for information on subcellular patterns.

Login Public Access Quick Start



External search

- http://pslid.cbi.cmu.edu/public3/search.jsp?protein=calponin-2

http://pslid.cbi.cmu.edu/public3/search.jsp?protein=calponin-2

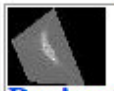
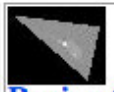
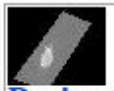
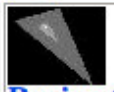
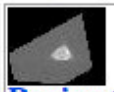
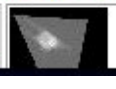
http://pslid.cbi...otein=calponin-2 | temp7_710B35DB64C10A8CF21...

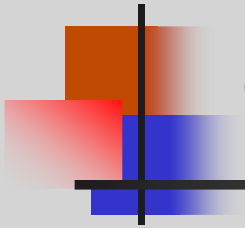
Search results for Image Type: 2D Static, Target: calponin-2

10 regions returned (30 regions shown) from the query.

View the summary of set **temp8_710B35DB64C10A8CF219992B3A193B57**.

Click besides a given image to retrieve similar images in the database.

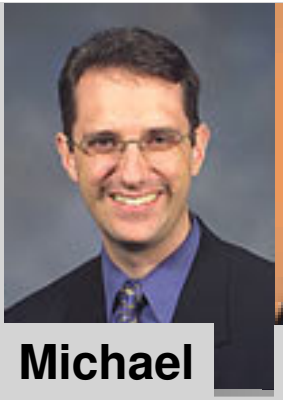
	Image	Cell Name	Organism	Segmenter	Experiment	Protocol	Target	Microscopy & Filter
	 Region 68249	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68280	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68311	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68342	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68373	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68404	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
	 Region 68435	3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500
		3T3	Mus musculus	External	Cyto039	GFP_Live	Calponin-2	Olympus IX500



Conclusions

- Methods well worked out for classifying and learning protein patterns - better than visual examination
- Temporal information improves discrimination
- Progress on decomposing complex patterns and synthesizing distributions
  - High-resolution, reliable data for bottom-up systems modeling
- Graphical models provide improved classification of single cells in fields (and potentially tissues)
  - New fast inference algorithm
- Image database integrated with interpretation tools (PSLID)
- Information extractor for online text and images (SLIF)

Acknowledgments


Michael


Mia


Greg

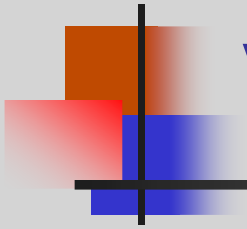

Meel

- Students
  - Dr. Michael Boland
  - Dr. Mia Markey (ugrad)
  - Gregory Porreca (ugrad)
  - Dr. Meel Velliste
  - Dr. Kai Huang
  - **Dr. Xiang Chen**
  - **Dr. Ting Zhao**
  - **Dr. Juchang Hua**
  - **Dr. Yanhua Hu**
  - **Shann-Ching Chen**
- Funding
  - NSF, NIH, Commonwealth of Pennsylvania
- Collaborators/Consultants
  - David Casasent, Simon Watkins, **Jon Jarvik, Peter Berget, Jack Rohrer**, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, William Cohen, **Geoff Gordon**
  - NSF ITR: B. S. Manjunath Ambuj Singh


Kai


Xiang

Carnegie Mellon



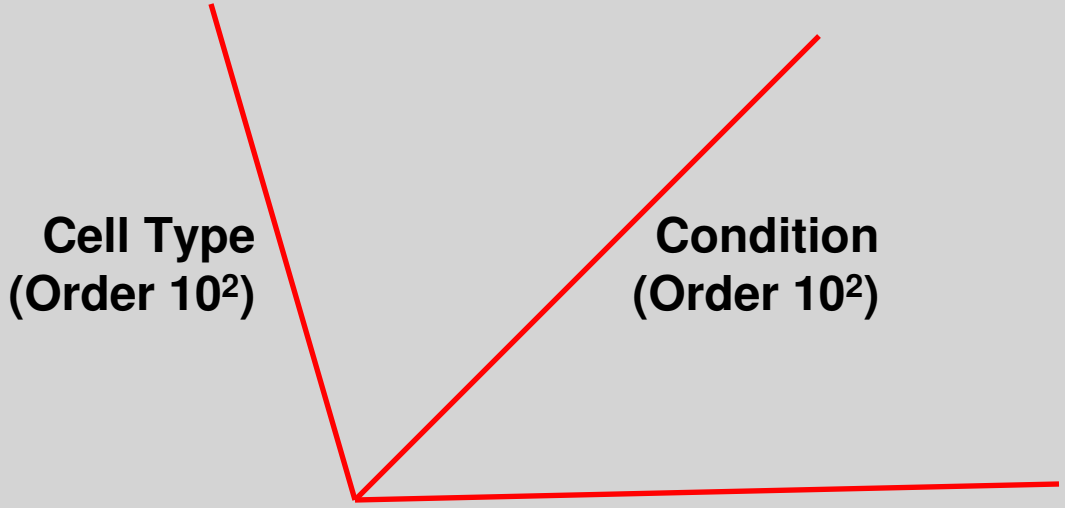
Vision

- Full automation of
  - experiment design
  - adaptive acquisition
  - model-based image interpretation
- to generate biological knowledge from images in a form suitable for systems modeling

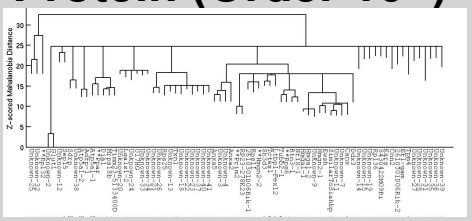
The Future of Subcellular Pattern Analysis

Carnegie Mellon

The problem



Protein (Order 10^4)



Plus: Time scale from subsecond to years

Other subcellular location projects

- O'Shea group - Yeast
  - GFP-tagged cDNAs
  - GFP and DNA images with some additional markers
- Pepperkok group - human (MCF7 cells)
  - GFP-tagged cDNAs
  - GFP and DNA images
- Uhlen group (Protein Atlas) - human
  - Immunohistochemistry with monospecific antibodies
  - DAB and hematoxylin images
  - Fixed tissues
- Schubert group (MELK technology)
  - Cycles of immunofluorescence, imaging and bleaching
  - Fixed tissues
- Teasdale group (Locate, Hela)
  - Immunofluorescence and GFP-tagged proteins
  - GFP and DNA images

How do we really analyze subcellular location?

- Classification and comparison good for focused questions but there are too many questions to ask
- Scope of problem argues for cooperation on grand scale: Human Cytome Project?
- Need intelligent (optimized) data collection: probabilistic methods to integrate available data, make predictions and suggest experiments

Carnegie Mellon

Molecular Biosensor and Imaging Center

Welcome to the Molecular Biosensor and Imaging Center website.

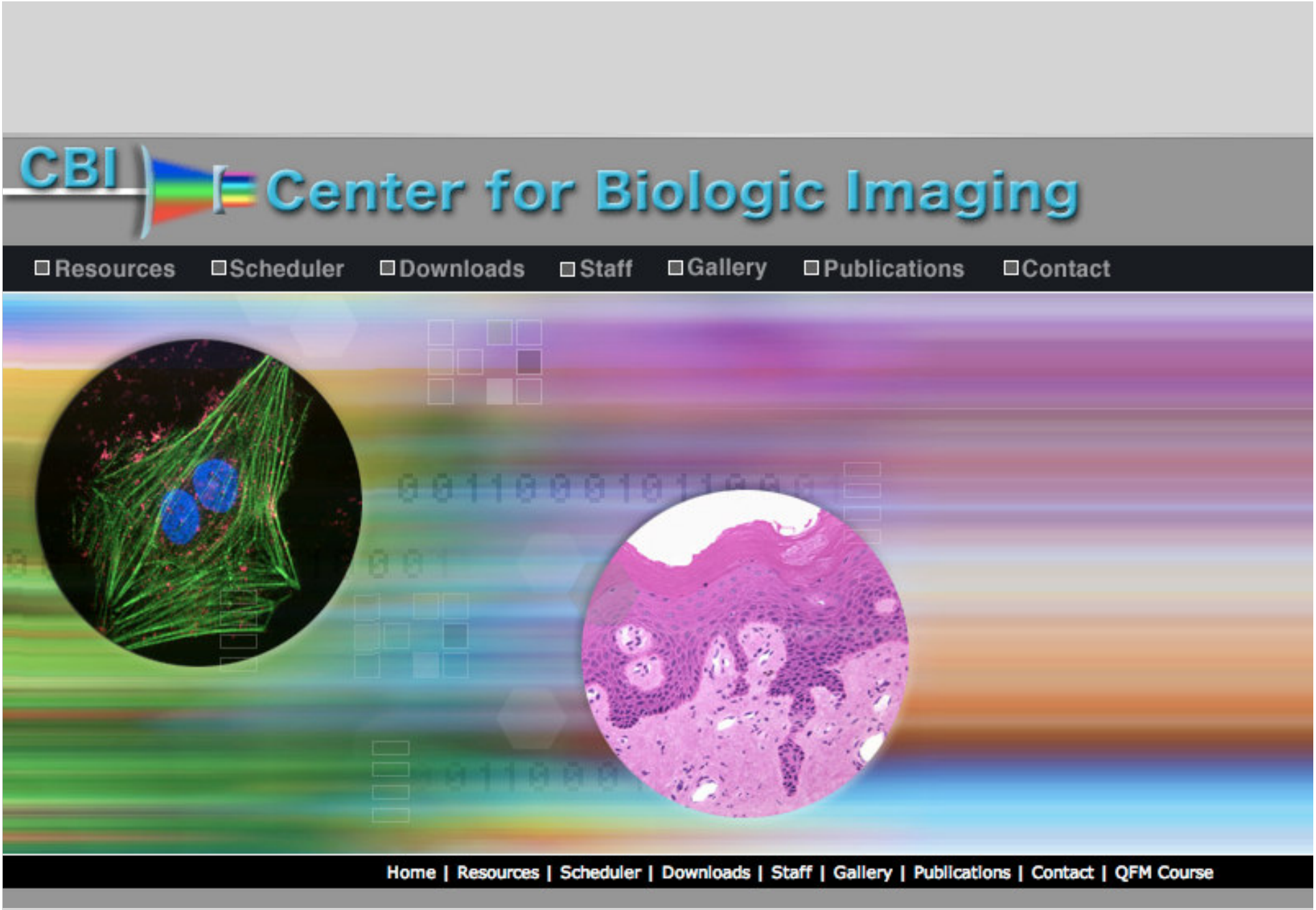
Mission

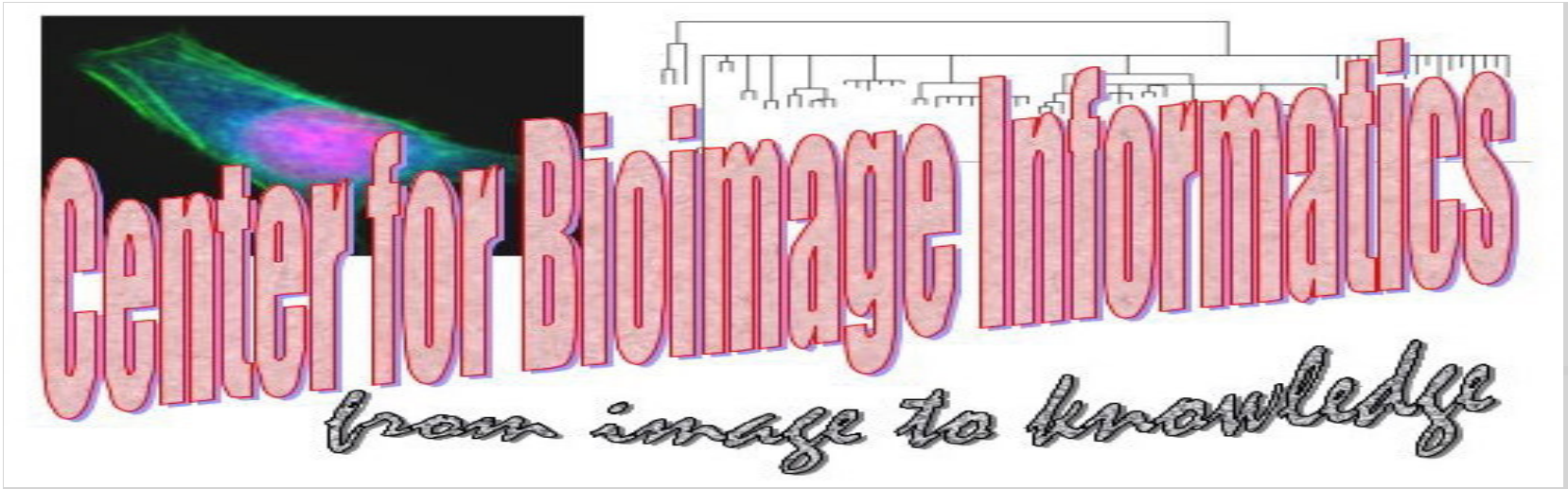
To develop fluorescence detection technologies for biomedical research and NASA space exploration.

NIH Technology Center for Networks and Pathways

Carnegie Mellon

Alan Waggoner


CBI
Center for Biologic Imaging
Resources
Scheduler
Downloads
Staff
Gallery
Publications
Contact
Home | Resources | Scheduler | Downloads | Staff | Gallery | Publications | Contact | QFM Course



What do we do

1. Biology: Pose a question about a biological system
2. Acquisition: Design strategy for collecting relevant information in the form of images of molecules, cells, organisms
3. Signal Processing/Computer Science: Find the answer through image processing and machine learning
4. Scientific Computing: Optimize computational performance for real-time applications and sharing

For more information: http://www.cbi.cmu.edu

NCIBI
NATIONAL CENTER FOR INTEGRATIVE BIOMEDICAL INFORMATICS

Search

home

you are here: home

navigation

- Home
- About NCIBI
- Computational Technology
- Driving Biological Problems
- Resources and Software
- Education and Training

National Center for Integrative Biomedical Inform

by plone — last modified 2005-09-29 09:29 AM

Mission

The mission of the NCIBI is to facilitate scientific exploration of complex di
currently feasible.

The Center develops and interactively integrates analytical and m
appropriate molecular biology information from emerging experim

UM Press Release

nation access and data analysis workflow
knowledge models of biological systems
roblems are prostate cancer progression,
ity of type 2 diabetes, and genetic susc

ach, training, and education programs.

borators

Collaborations with Bill Mohler, Ian Moraru, Les Loew, Paul Campagnola (U Conn)

Collaboration with Badri Roysam (RPI) and Sally Temple (Albany Med Coll), Stem Cell Patterning and FARSIGHT system

Collaboration with Dan Rines and Sumit Chanda (GNF San Diego) on high throughput location proteomics

Thank you !

Carnegie Mellon