Studying the subcellular location space with bioimages and other data modalities

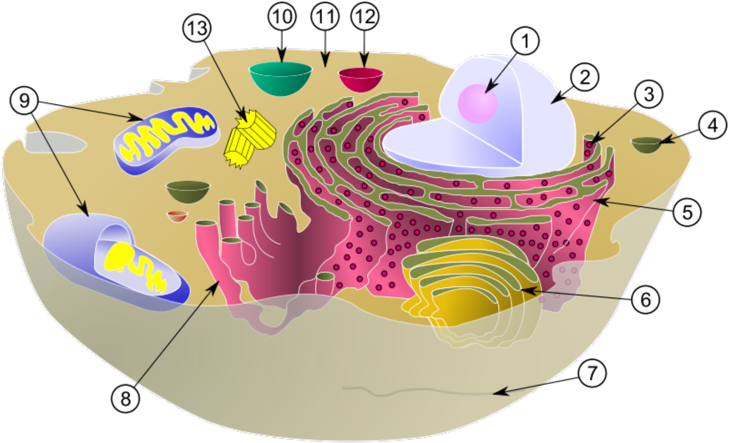
Luis Pedro Coelho

Carnegie Mellon University

September 22 2011

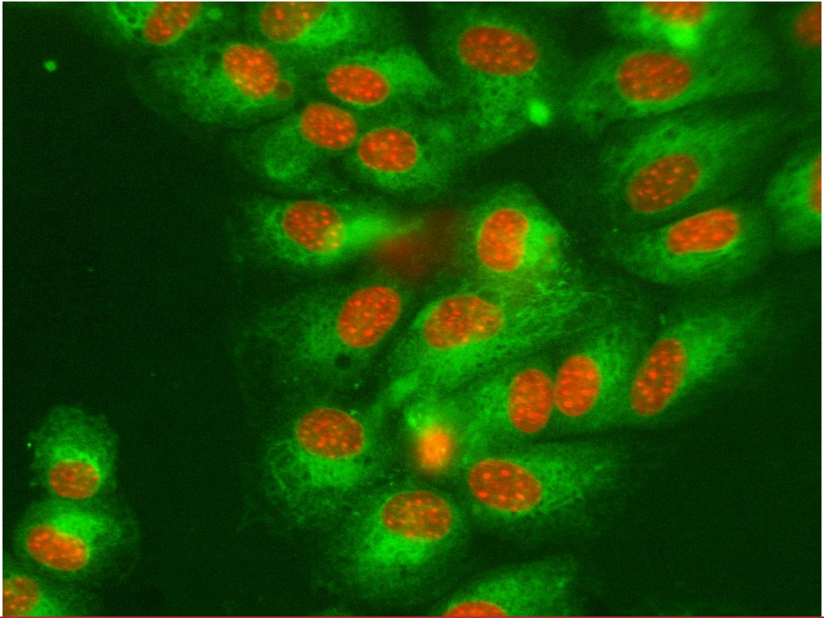
Note: All material related to unpublished work has been removed from this online version.

The Cell

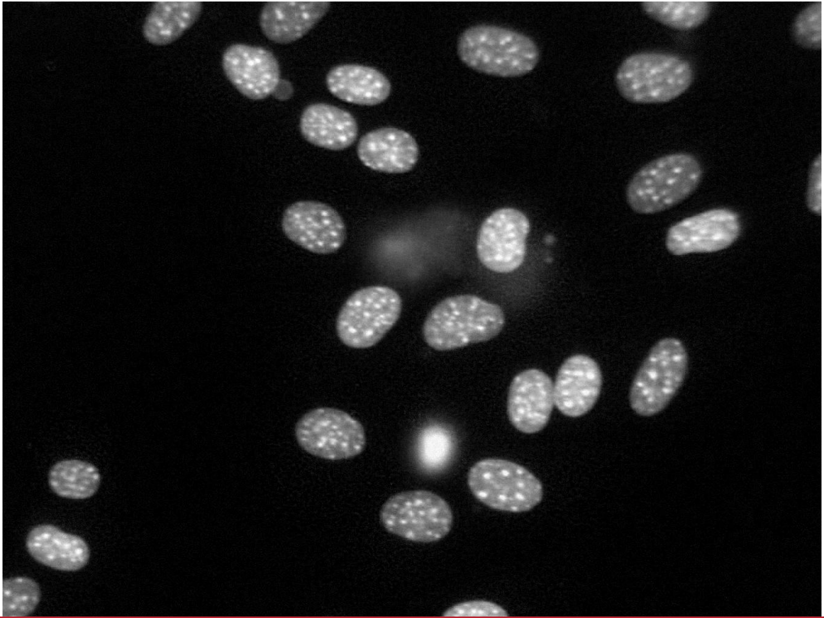


(Wikipedia)

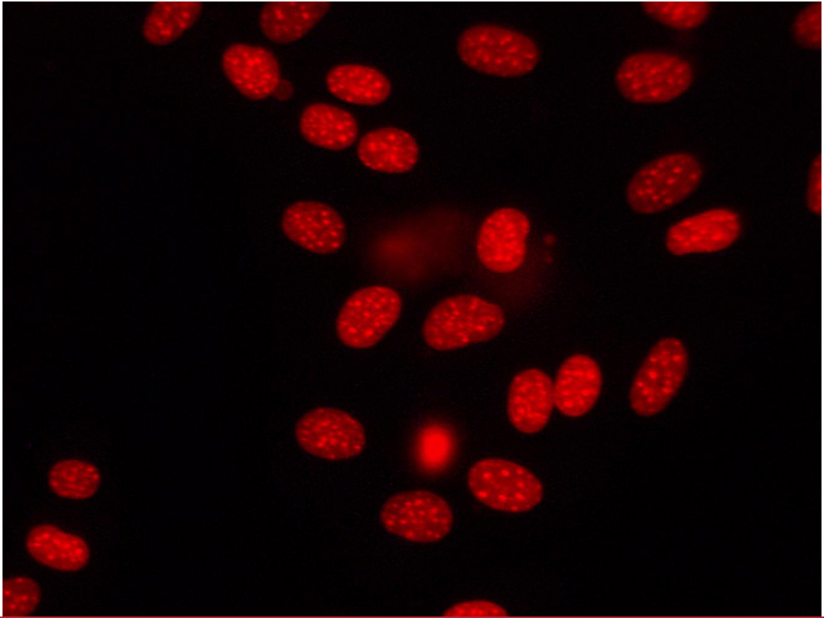
The Data



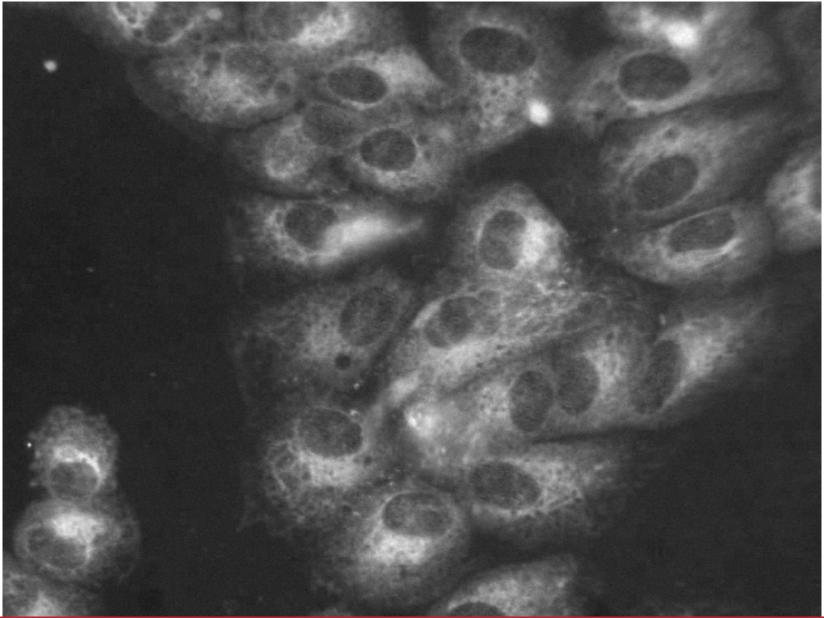
The Data



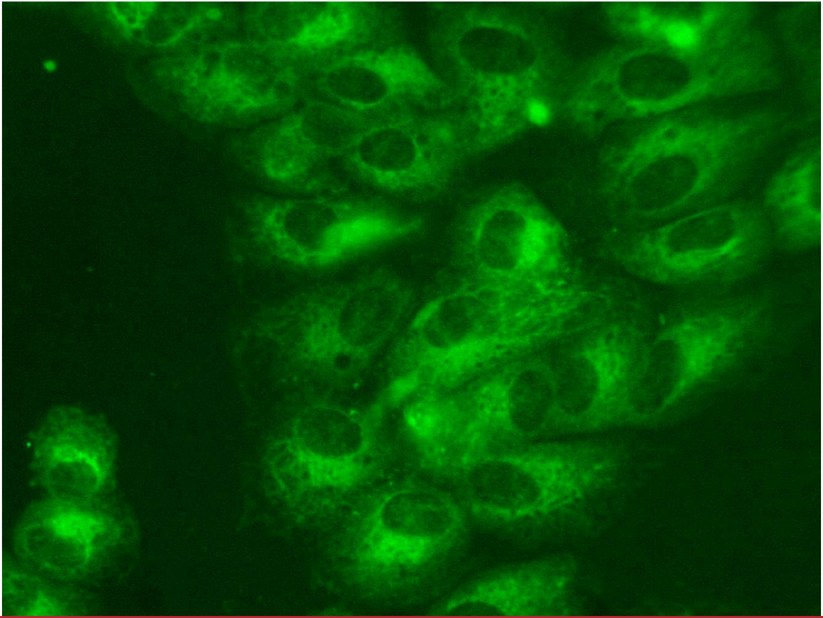
The Data



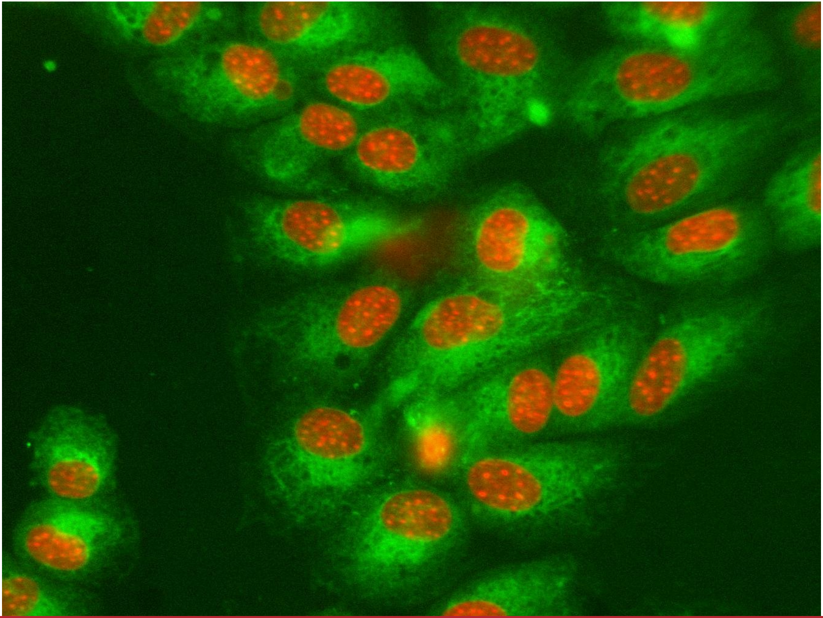
The Data



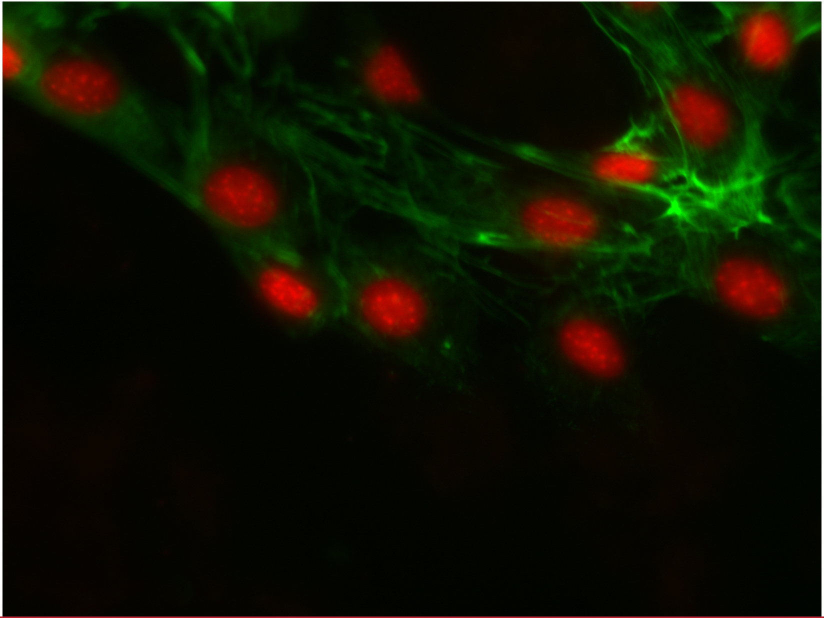
The Data



The Data



The Data



First Task

Classification

Given labeled data, can we learn a classification model?

Pre-processing

Preprocessing is denoising, background correction, . . .

I am going to ignore that step.

Features

Feature Based Approach

- Represent the image by a small number of features.
- Proposed by Boland and Murphy, in 1998.
- Very successful.

Features

- A feature is any number you can compute from the image.
- For a good features, you wish to simmultaneously
  1. Capture the important variations.
  2. Disregard the important variations.
- These are naturally problem dependent,
- but machine learning helps.

Typical Features

- Texture (Haralick, Gabor, . . .)
- Edginess, smoothness, . . .
- . . .

The literature is very vast.

Example Feature

12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

Example Feature

12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

Example Feature

12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

Algorithm

- For each 3×3 region:
- Find the maximum and the minimum.
- Subtract the minimum from the maximum.
- You end up with a number per region (per pixel).

Algorithm

- For each 3×3 region:
- Find the maximum and the minimum.
- Subtract the minimum from the maximum.
- You end up with a number per region (per pixel).

For an image level feature, average this number

Algorithm

- For each 3×3 region:
- Find the maximum and the minimum.
- Subtract the minimum from the maximum.
- You end up with a number per region (per pixel).

For an image level feature, average this number

1. What is this feature sensitive to?
2. What is this feature invariant to?

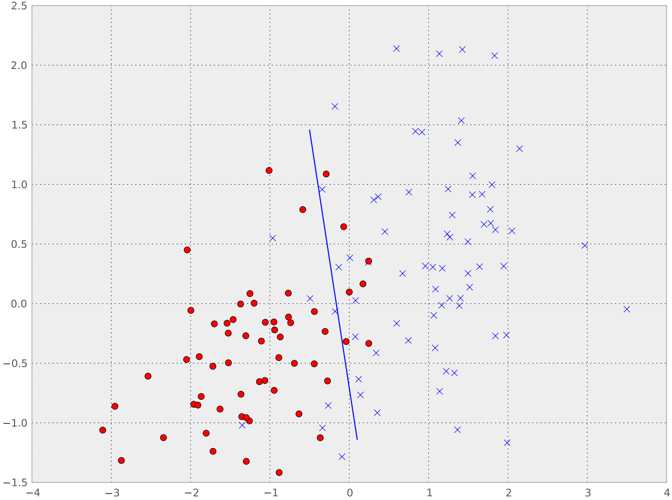
Classifiers

Simple classifier: if average of region is less than 10, call it sky.

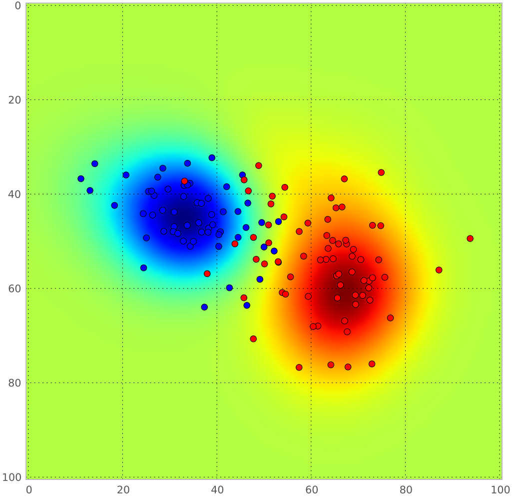
Classifiers

Simple classifier: if average of region is less than 10, call it sky.

Classifiers



Classifiers



Results

	Cyto	Cytosk	Lyso	PM	Mito	N	NN
Cyto	115	10	3	15	8	4	0
Cytosk	14	147	3	2	30	1	0
Lyso	3	1	14	0	50	0	1
PM	31	6	2	9	2	1	0
Mito	22	30	15	0	126	6	1
N	25	1	0	1	0	219	9
NN	1	0	0	0	1	16	95

Other Problems

Other Typical Classification Problems

- Phenotype in a Genome-wide Screen
- Stem cell differentiation
- . . .

Conclusions

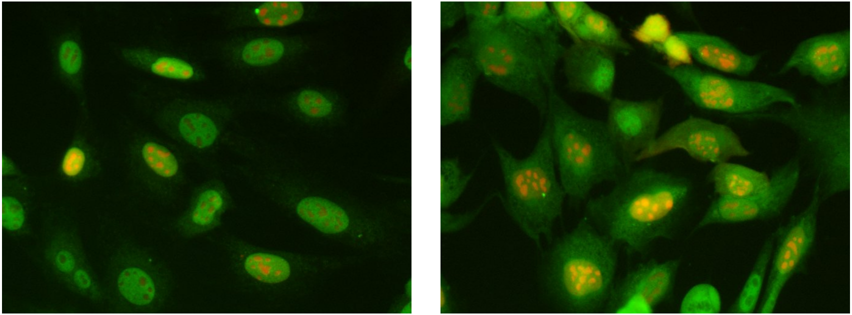
- Computers can do very well at classification.
- Flexible tool if you have the training data.

Detecting Changes in Pattern

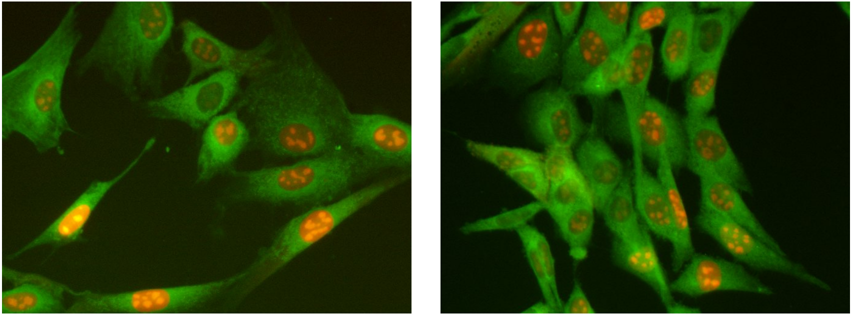
Which proteins change pattern when we stress the cells?

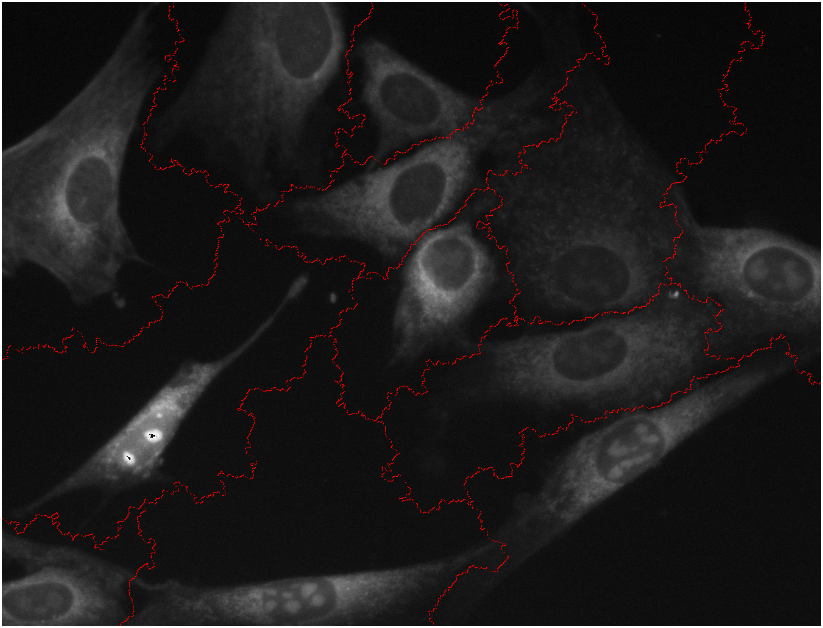
In collaboration with Marc Halterman @ Rochester

Positive Example

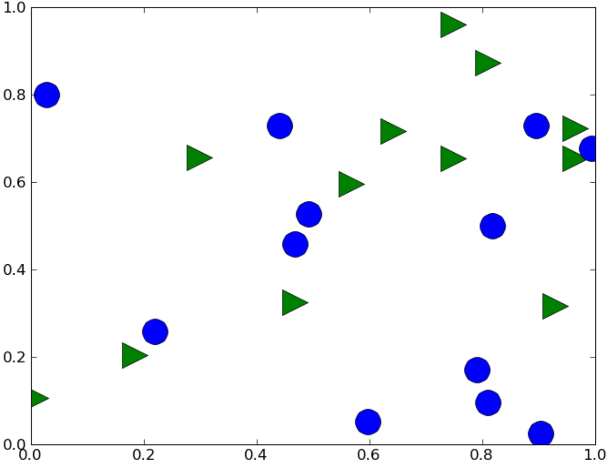


Negative Example





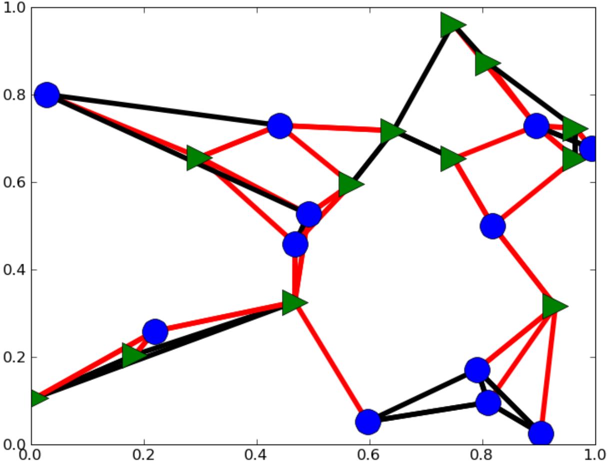
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

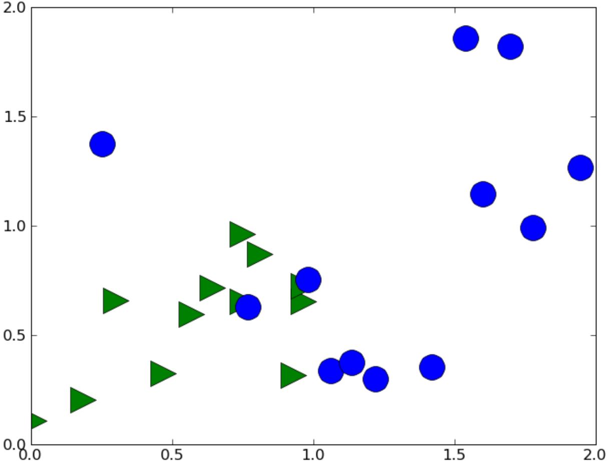
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

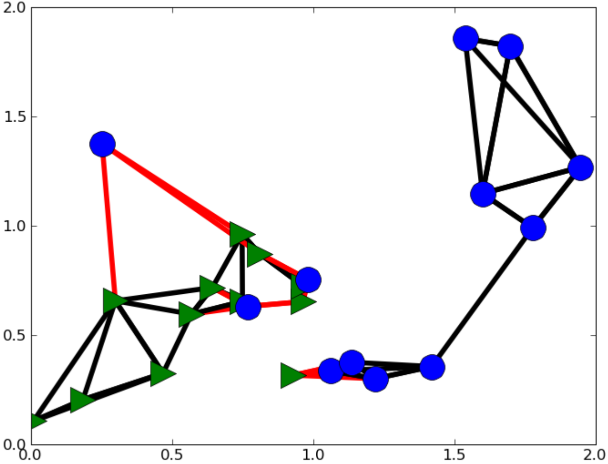
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

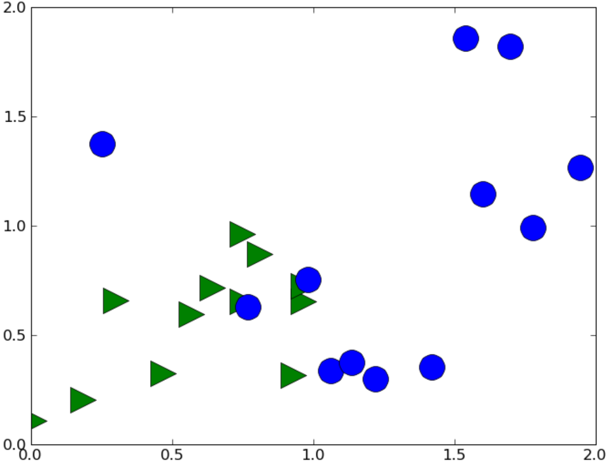
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

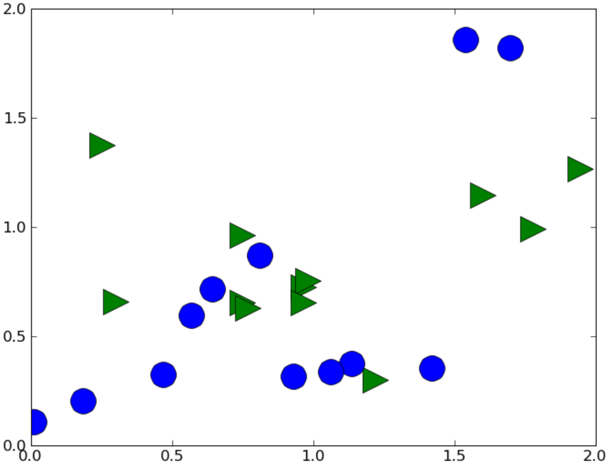
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

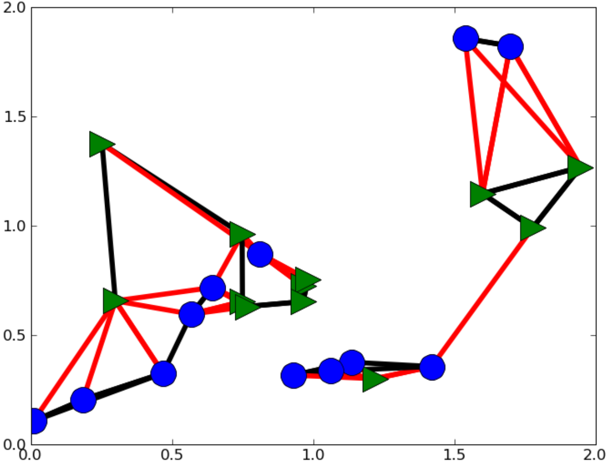
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

This is a simple example, but handling heterogeneity is an important area of research.

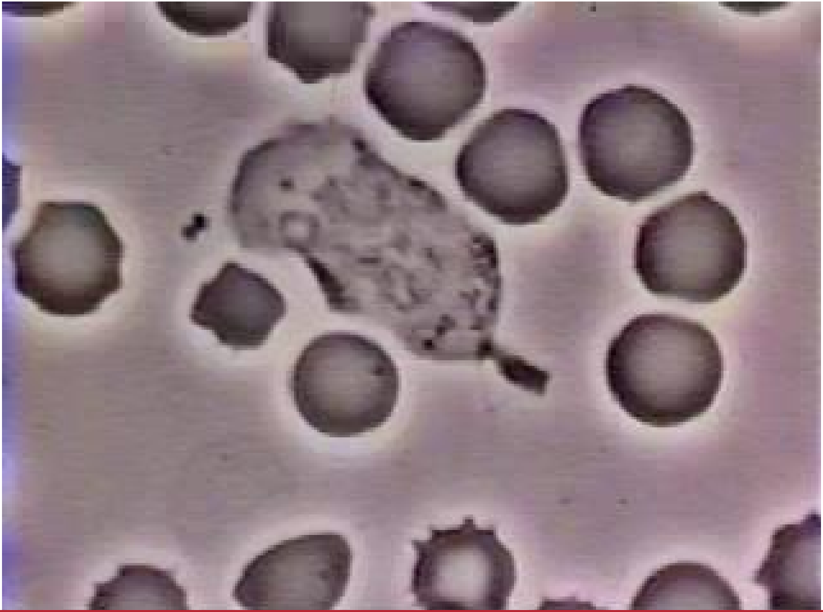
Tracking

- Can you follow the cells in a movie?

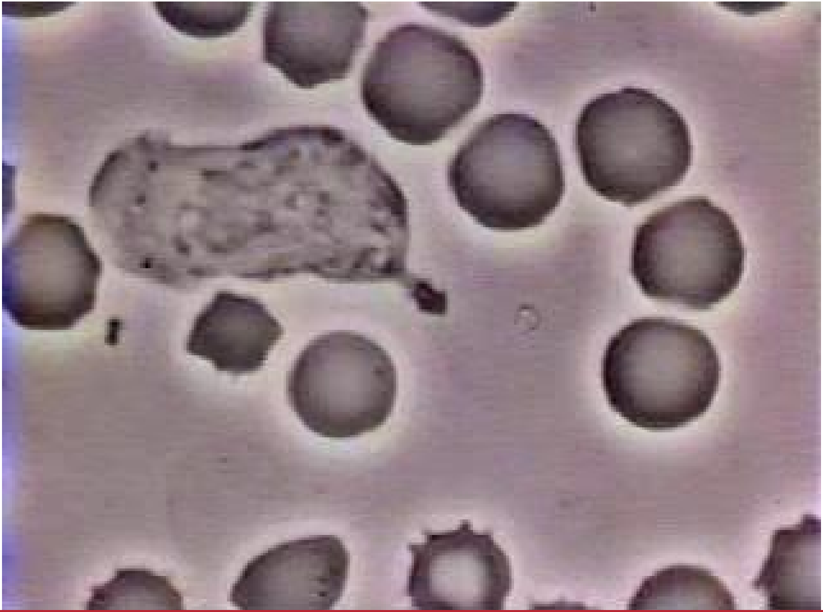
Example

Example

Example



Example

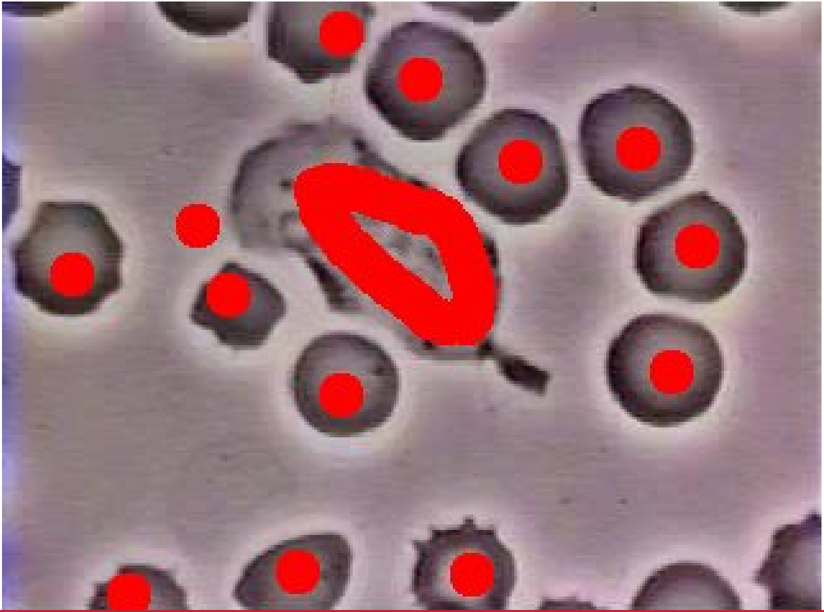


Techniques

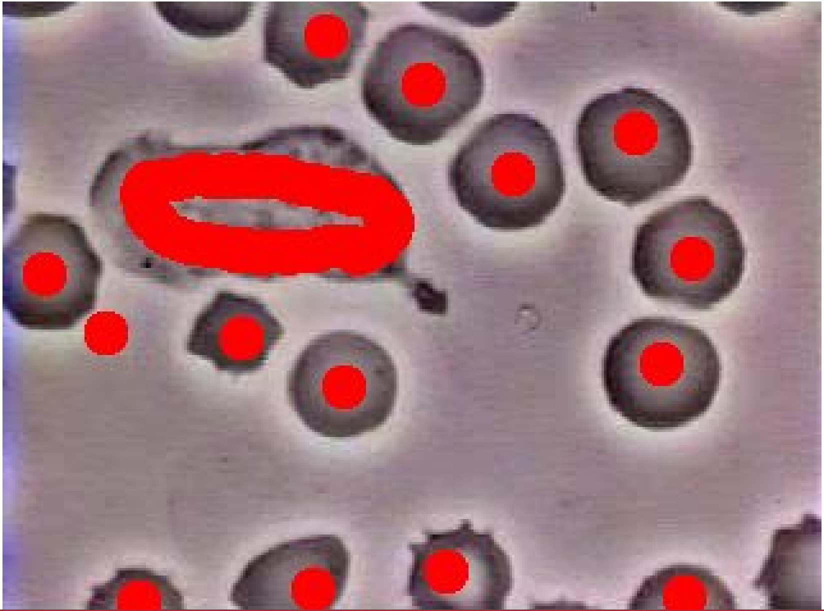
Pipeline Approach:

- (Pre-process)
- Detect/Segment
- Link

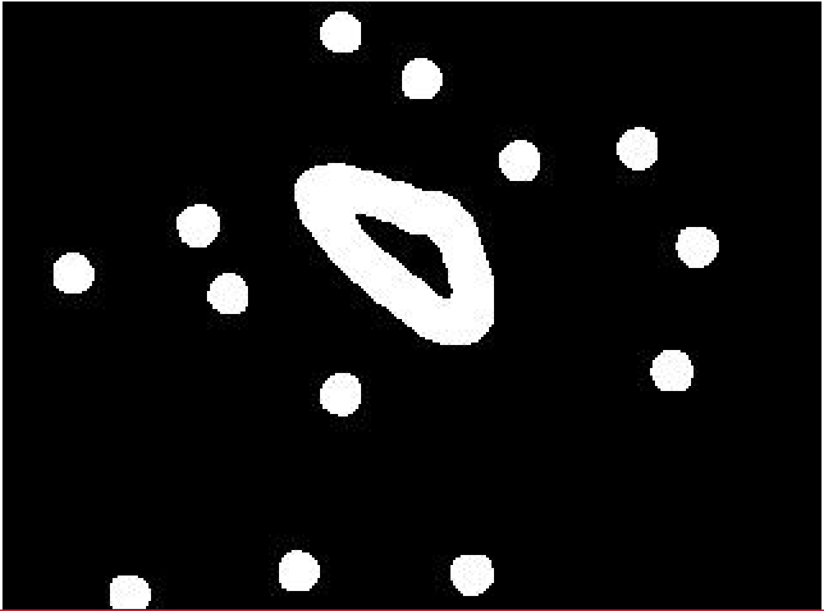
Example



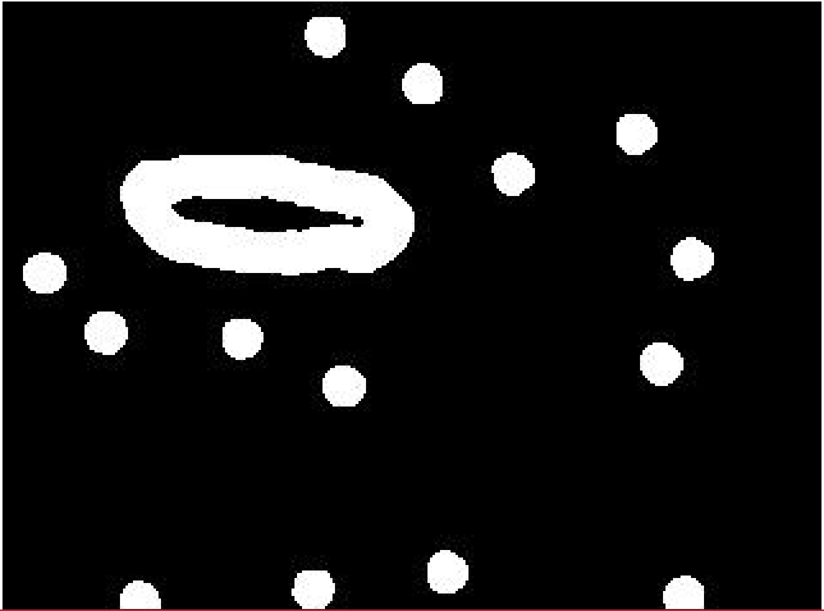
Example



Example



Example



Greedy Linking Approach

Linking Problem

Given a set of points in one frame $\{x_i\}_i$
and points in another frame $\{y_i\}_i$,
how do we link them together?

Greedy Linking Approach

Linking Problem

Given a set of points in one frame $\{x_i\}_i$
and points in another frame $\{y_i\}_i$,
how do we link them together?

Marriage Problem

Given a bipartate graph $M \cup F$ with weights w_{mw} between the members, can we find a matching that maximises $\sum w_{mf}$?

Yes, we can solve this in polynomial time using the Hungarian algorithm.

Better Approaches

Hungarian Algorithm

- Simple, off-the-shelf, solution.
- Computationally efficient

- Not very robust to noisy data

How to do better

- Take multiple time-steps into consideration.
- Active contours approaches combine segmentation and tracking.

Generative Models

Can we generate fake images that look like real images?

Generative Models

Why?

- Learn biology.
- Serve as input for simulation.
- Communication of image data.
- Test set generation.

Generative Models

How?

1. Define a model
2. Learn parameters from data
3. Generate images

(T. Zhao & Murphy, 2007)

(T. Peng et al., 2009)

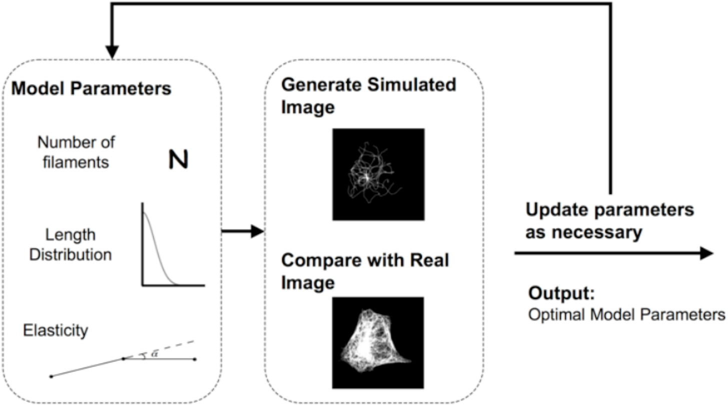
Generative Models

How?

1. Define a model
2. Learn parameters from data
3. Generate images

(T. Zhao & Murphy, 2007)

(T. Peng et al., 2009)



Model Parameters

Number of filaments N

Length Distribution

Elasticity

Generate Simulated Image

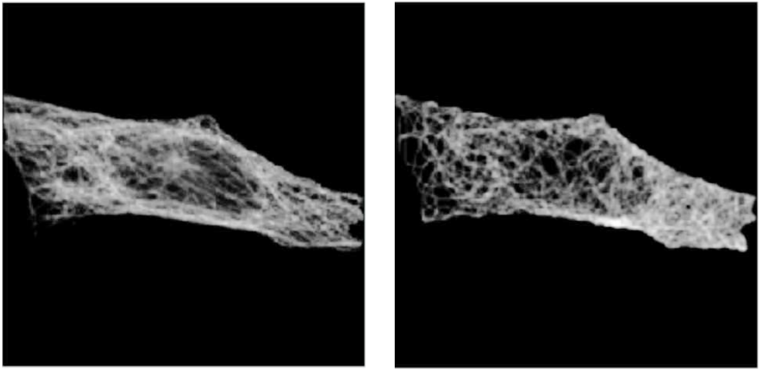
Compare with Real Image

Update parameters as necessary

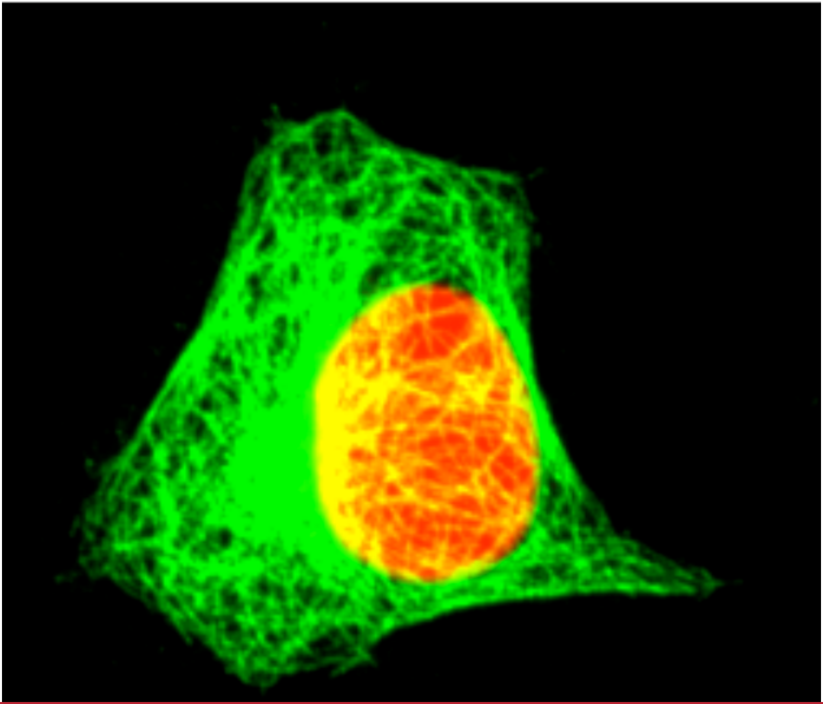
Output:
Optimal Model Parameters

(Shariff et al., 2009, 2010)

Two Images



(A. Shariff et al., 2009)



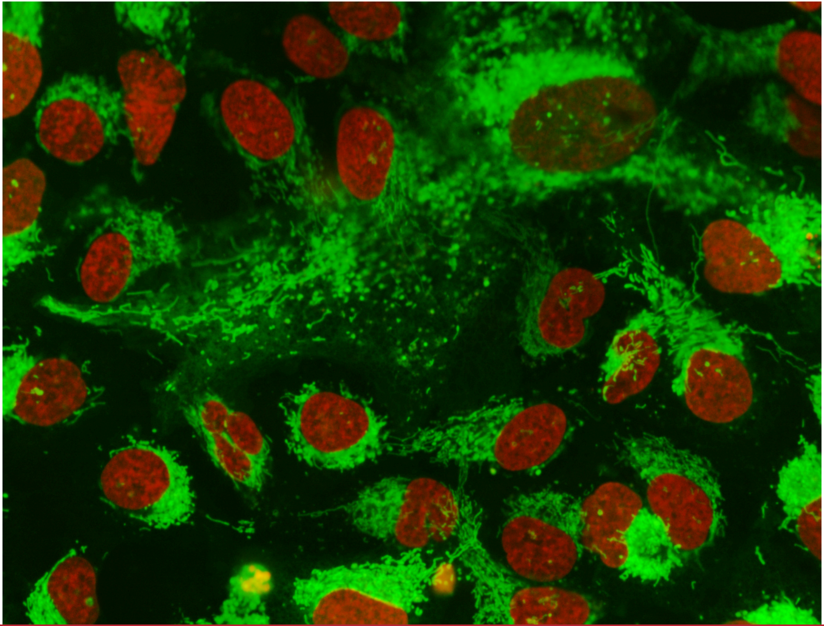
Mixture Patterns Classification

Previously reported methods work well for simple classes, like “endosomes” or “mitochondria.”

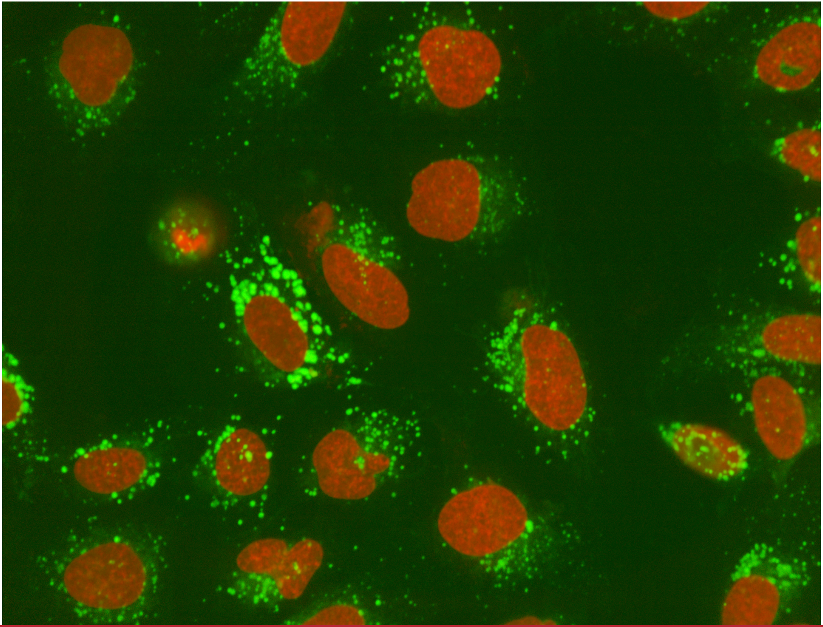
Mixture Patterns Classification

Previously reported methods work well for simple classes,
like “endosomes” or “mitochondria.”
What if a protein is present in both endosomes and mitochondria?

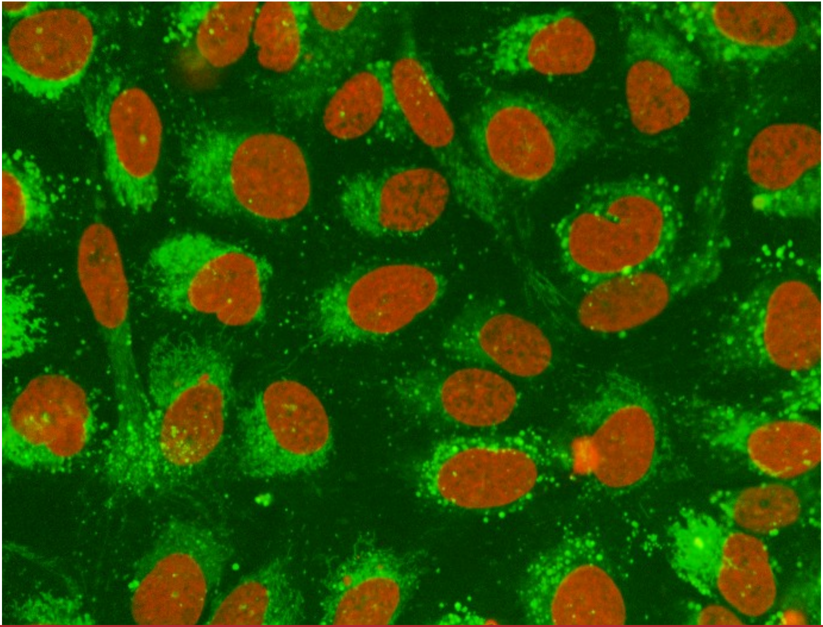
Mixture Pattern Example



Mixture Pattern Example



Mixture Pattern Example



Supervised Unmixing Problem

Given examples of pure patterns and a mixed pattern,
can we identify how much each pure pattern contributes to the mixture?

Supervised Unmixing Problem

Given examples of pure patterns and a mixed pattern,
can we identify how much each pure pattern contributes to the mixture?
Using an object-based approach, we can solve this.

(T. Zhao et al., 2005)
(T. Peng, G. Bonami et al., 2010)

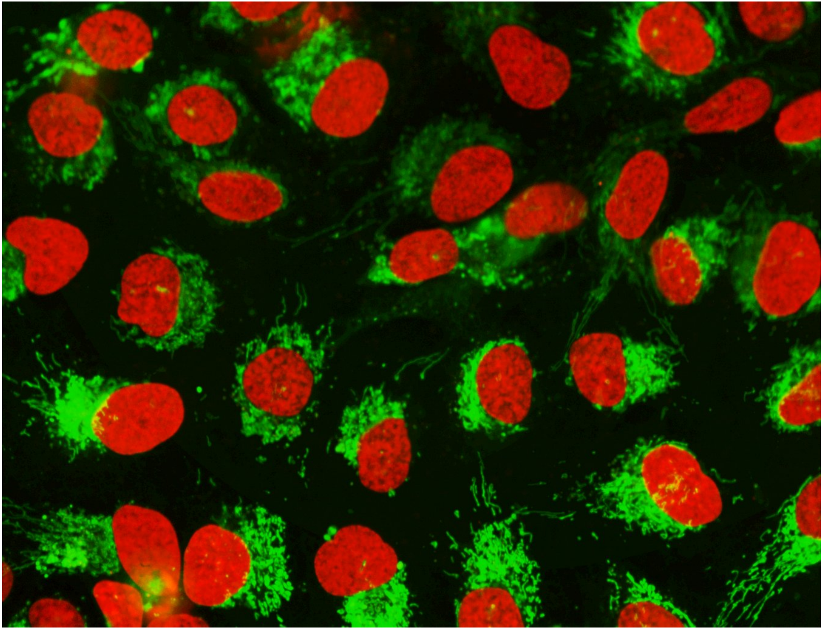
Unsupervised Unmixing Problem

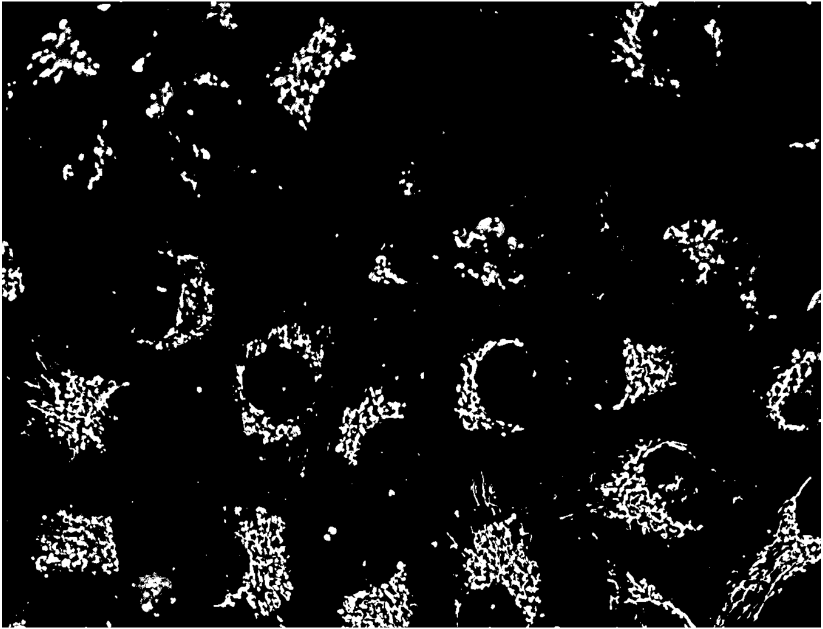
What if we don't know the pure patterns?

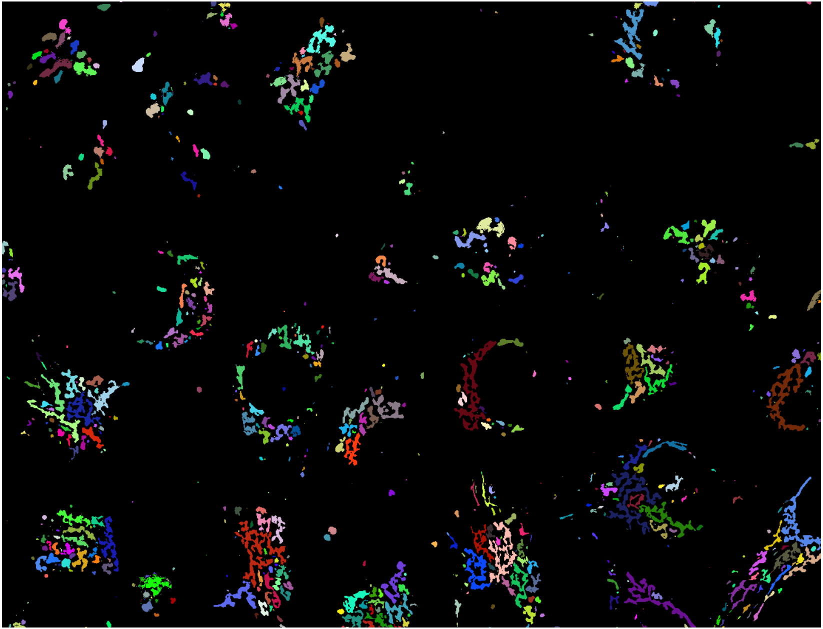
Unsupervised Unmixing Problem

What if we don't know the pure patterns?

Given a collection of untagged images,
can we identify the pure and mixed patterns?



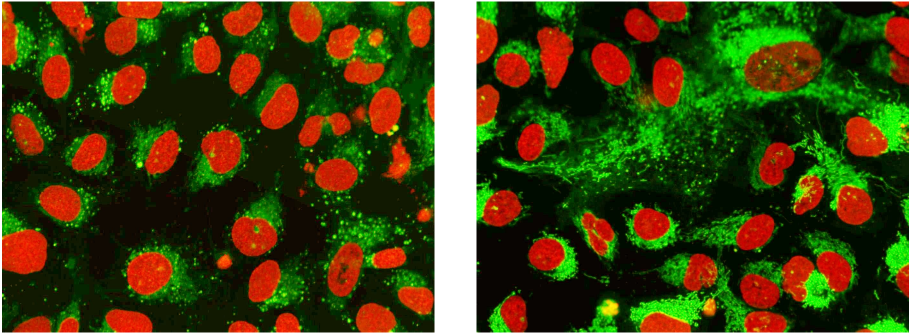




$$\min \sum_i \|X_i - Bc\|$$

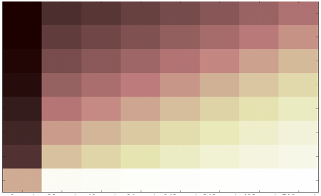
where B is the basis,
c are the coefficients,
and the problem subject to the constraint that the elements come from the dataset.

Results: Mixing Bases

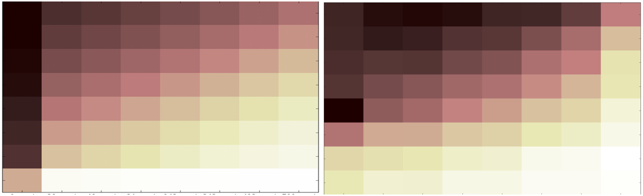


(Coelho et al., 2010)

Results: Mixing Fractions

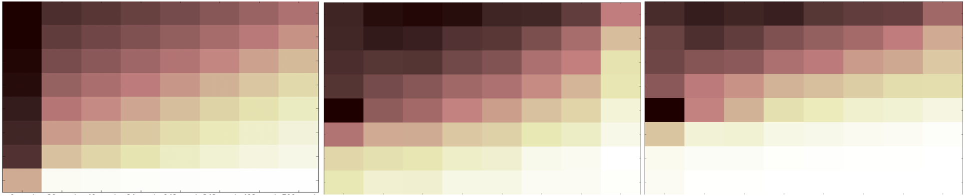


Results: Mixing Fractions



Correlation:
88%

Results: Mixing Fractions



Correlation:
88%

Correlation:
91%

(Coelho et al., 2010)

- Pattern Unmixing works both in supervised and unsupervised modes.

Conclusions

1. Intersection between computer vision and biology.
2. Basic problems are solved with machine learning.
3. Problems from image processing
4. Specific problems to this area:
    - generative (or indirect) modeling
    - pattern unmixing

Areas of Active and Future Research

1. New modalities.
2. Computational efficiency.
3. Robust algorithms (less parameter tuning).
4. Post-pipeline algorithms.
5. Integrating human and computer computation.
6. Integrating different data modalities.
7. Finding needles in haystacks.
8. Machine learning with very noisy data, but lots of it
   (web community is going similar things, what can we borrow?)
9. Active learning.
10. Unsupervised or semi-unsupervised methods.

Acknowledgements

- Robert F. Murphy
- Tao Peng
- Aabid Shariff
- Ting Zhao
- Estelle Glory-Afshar
- Murphy Lab

Thank you.