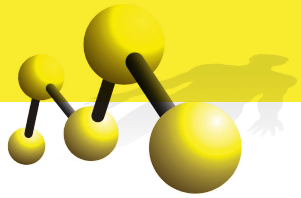
Bioimage Informatics: Computer Vision for Biology

Luis Pedro Coelho

Institute for Molecular Medicine, Lisbon
Mhlanga Lab

November 2011

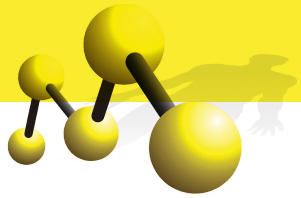
High Throughput Science



"The real measure of success is the number of experiments that can be crowded into twenty-four hours."

— Thomas Edison

High Throughput High Content Biology

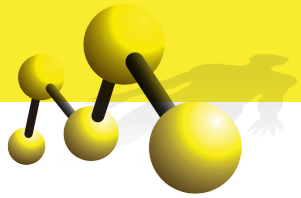


Lab Technologies

- Liquid handling robots
- Multi-well plates
- Automated microscopes

One can generate thousands of images per hour.

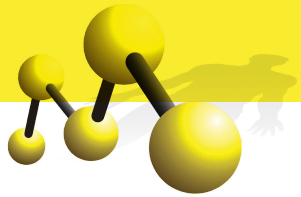
Images



8	2	2	1	1	1	2	2
8	8	2	2	2	2	2	8
21	8	8	2	2	2	8	8
21	8	8	8	2	8	8	8
21	8	8	8	8	8	8	8
21	8	8	8	2	8	8	8
21	8	8	2	2	2	8	8
8	8	2	2	2	2	2	8

This is the raw data.

Image Processing

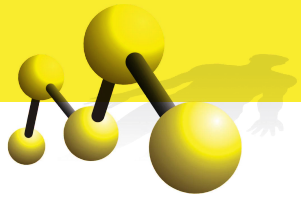


Typical Tasks

- Denoising
- Particle detection
- Segmentation
- ...

At the end of these steps, you still have an image which must be interpreted by computer or human.

Image Processing



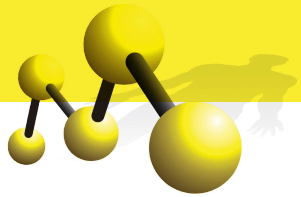
Typical Tasks

- Denoising
- Particle detection
- Segmentation
- ...

At the end of these steps, you still have an image which must be interpreted by computer or human.

I am not discussing any of this today.
See Alexandre's talk.

Image Processing



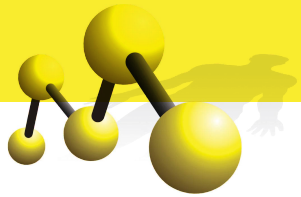
Typical Tasks

- Denoising
- Particle detection
- Segmentation
- ...

At the end of these steps, you still have an image which must be interpreted by computer or human.

I am not discussing any of this today.
See Alexandre's talk.

First Task



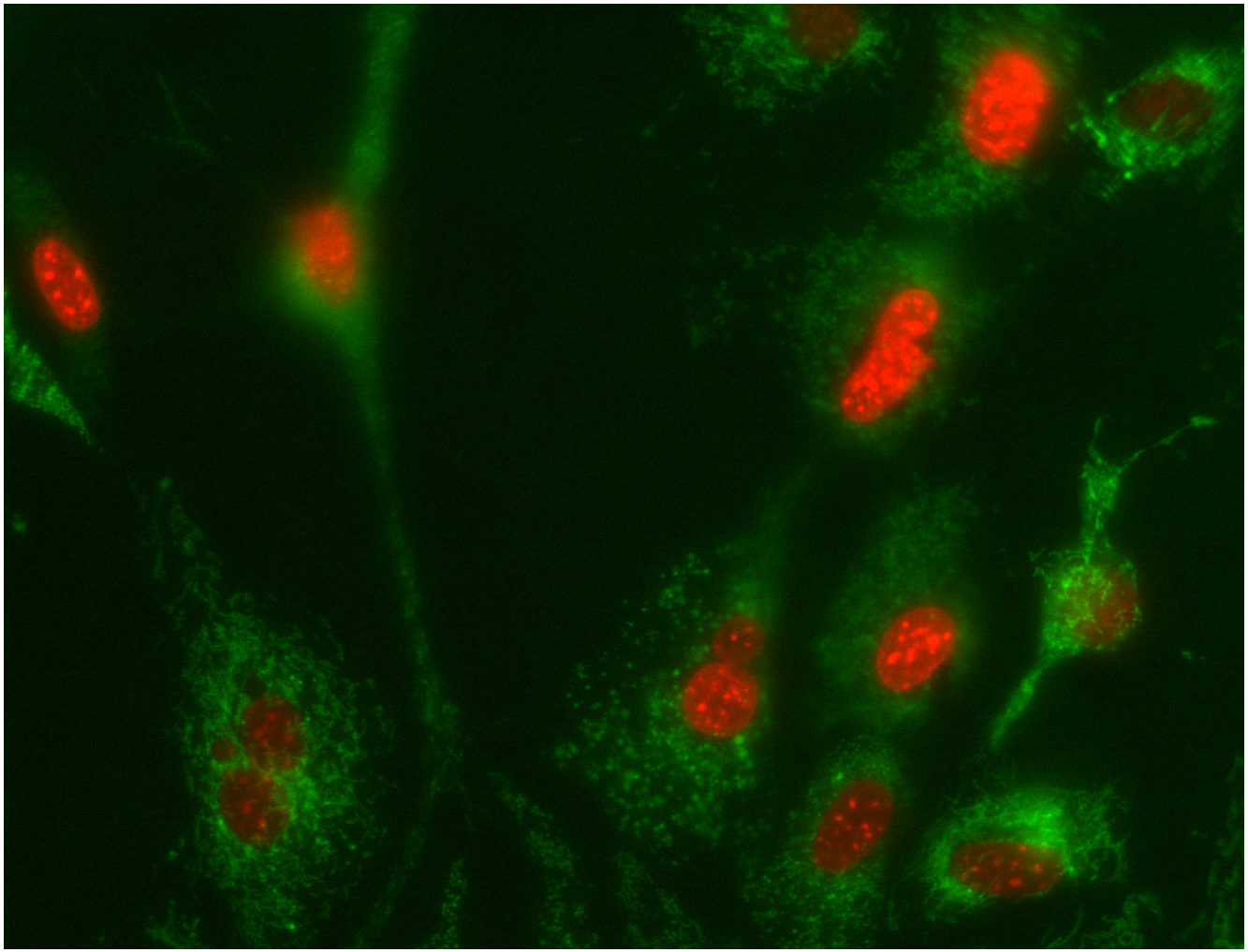
Classification

Given **labeled data**, can we learn a classification model?

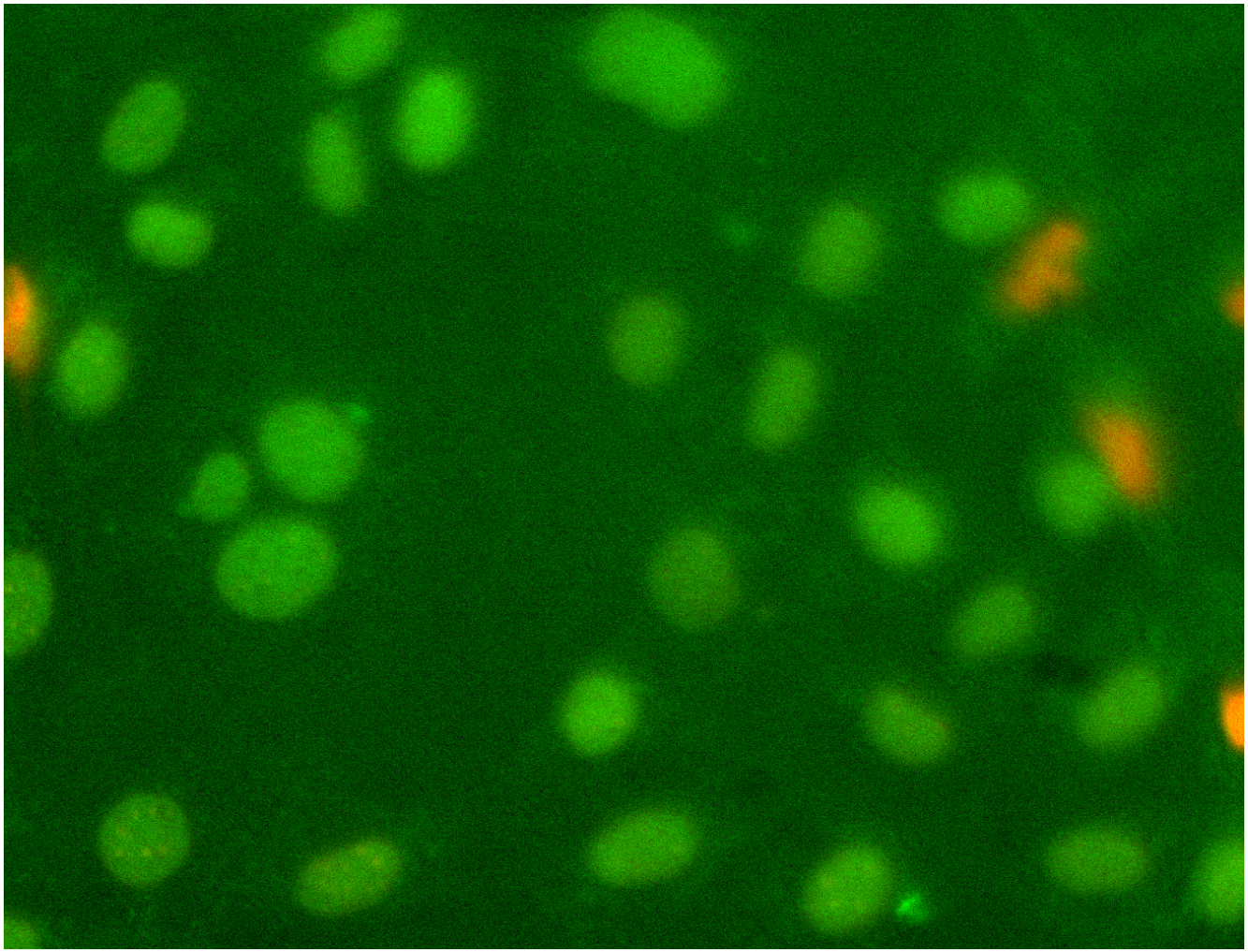
Labeled Data

A small dataset of images with **labels**.
The goal is to then **assign labels** to other images.

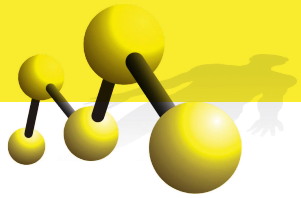
Example



Example



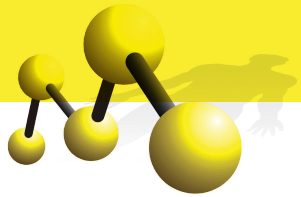
Features



Feature Based Approach

- Represent the image by a small number of features.
- Proposed by Boland and Murphy (1998) for subcellular location.
- Very successful for many applications.

Features

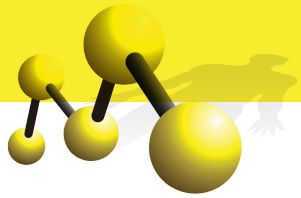


- A feature is any number you can compute from the image.

- For a good features, you wish to simmultaneously
  1. Capture the important variations.
  2. Disregard the unimportant variations.

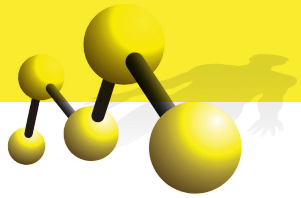
- These are naturally problem dependent,
- but machine learning helps.

Example Feature



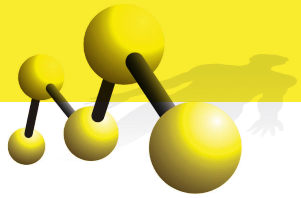
12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

Example Feature



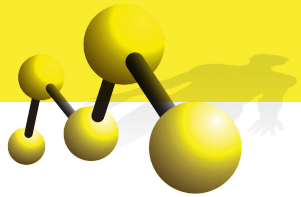
12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

Example Feature



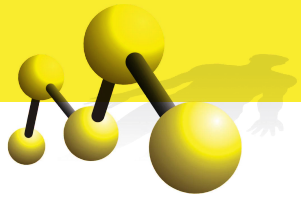
12	6	5	4	3	5
11	10	4	6	7	4
4	5	3	10	8	9
3	4	12	9	8	14
7	12	10	8	11	13

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).

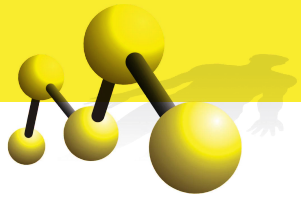
Algorithm



- 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

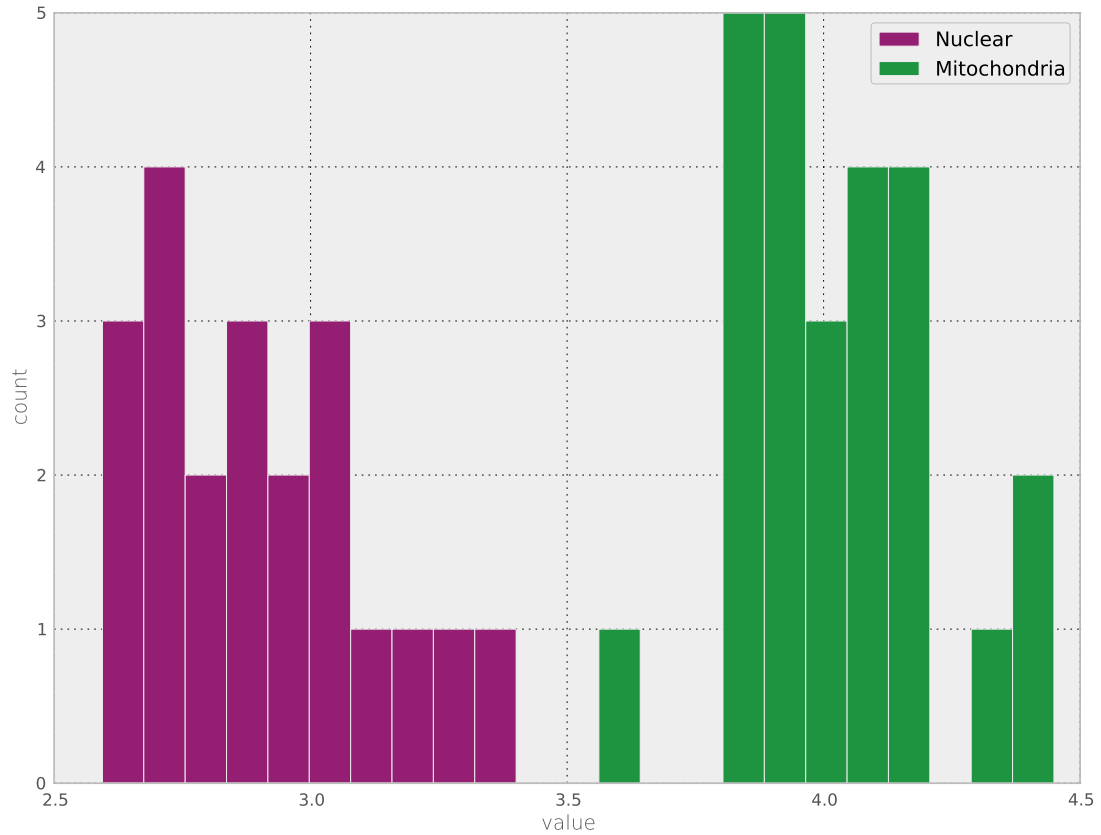


- For each 3×3 region:
- Find the maximum and the minimum.
- Subtract the minimum from the maximum.
- You end up with a number per region (per pixel).

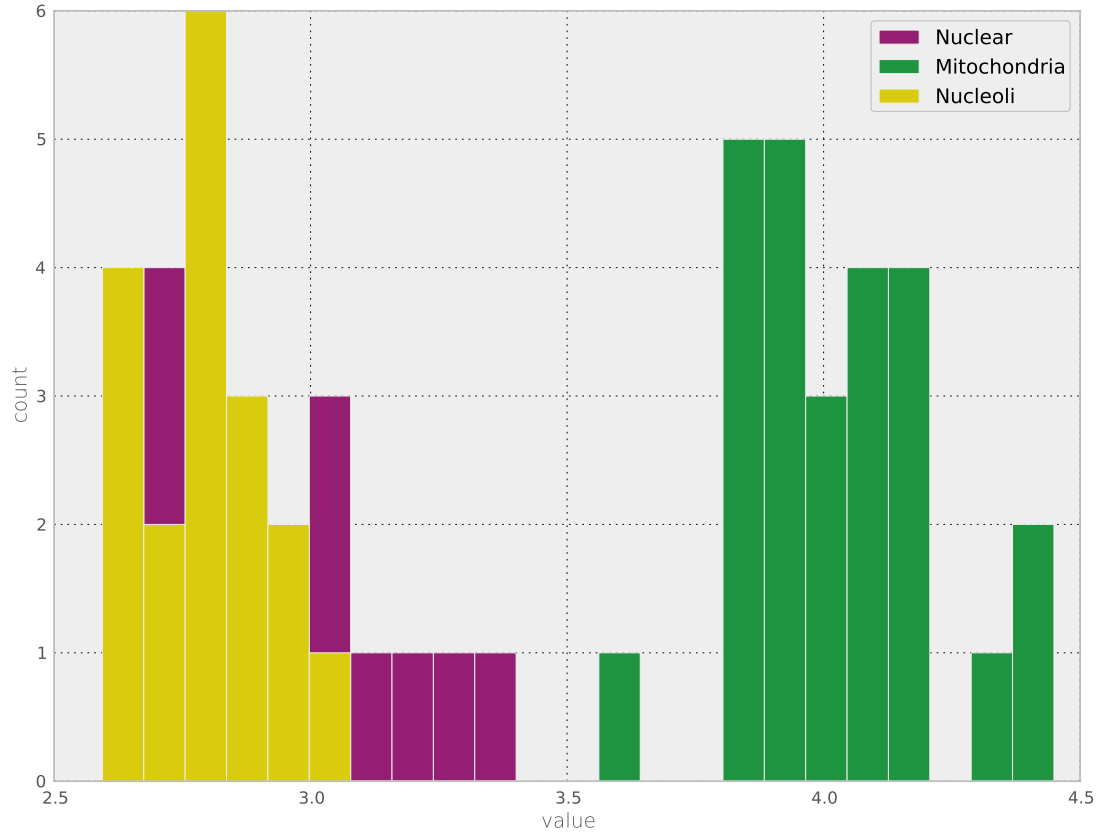
For an image level feature, average this number

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

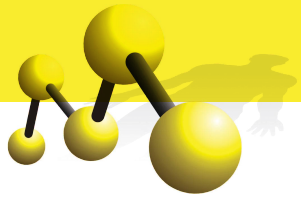
Example



Example


Nuclear
Mitochondria
Nucleoli
count
value
0
1
2
3
4
5
6
2.5
3.0
3.5
4.0
4.5

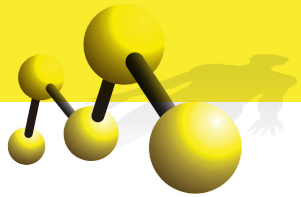
Complex Examples



Alternatives

- Manually design features by trial and error
- Machine learning approach

Complex Examples



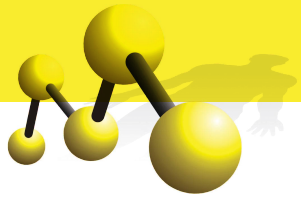
Alternatives

- Manually design features by trial and error
- Machine learning approach

Machine Learning

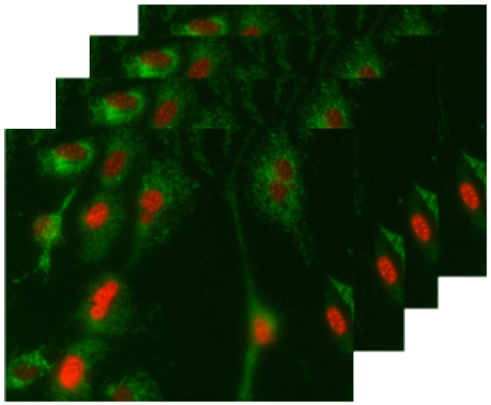
1. Use many generic features (tens to hundreds)
2. Automatically learn which features are important

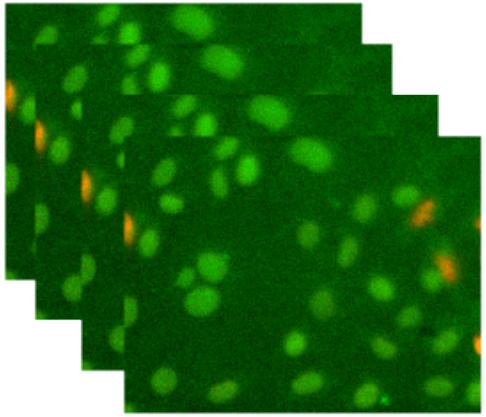
Typical Features

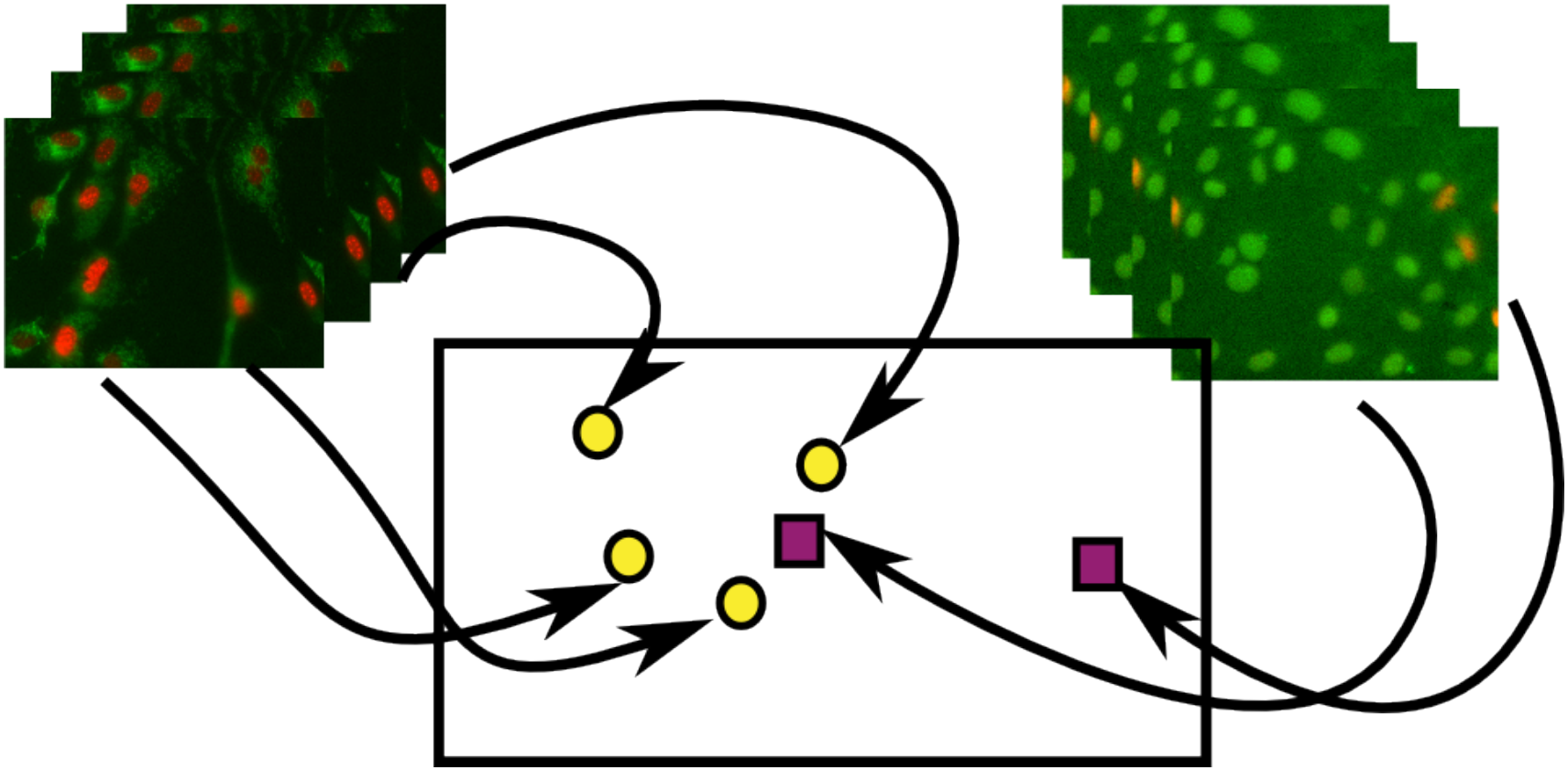


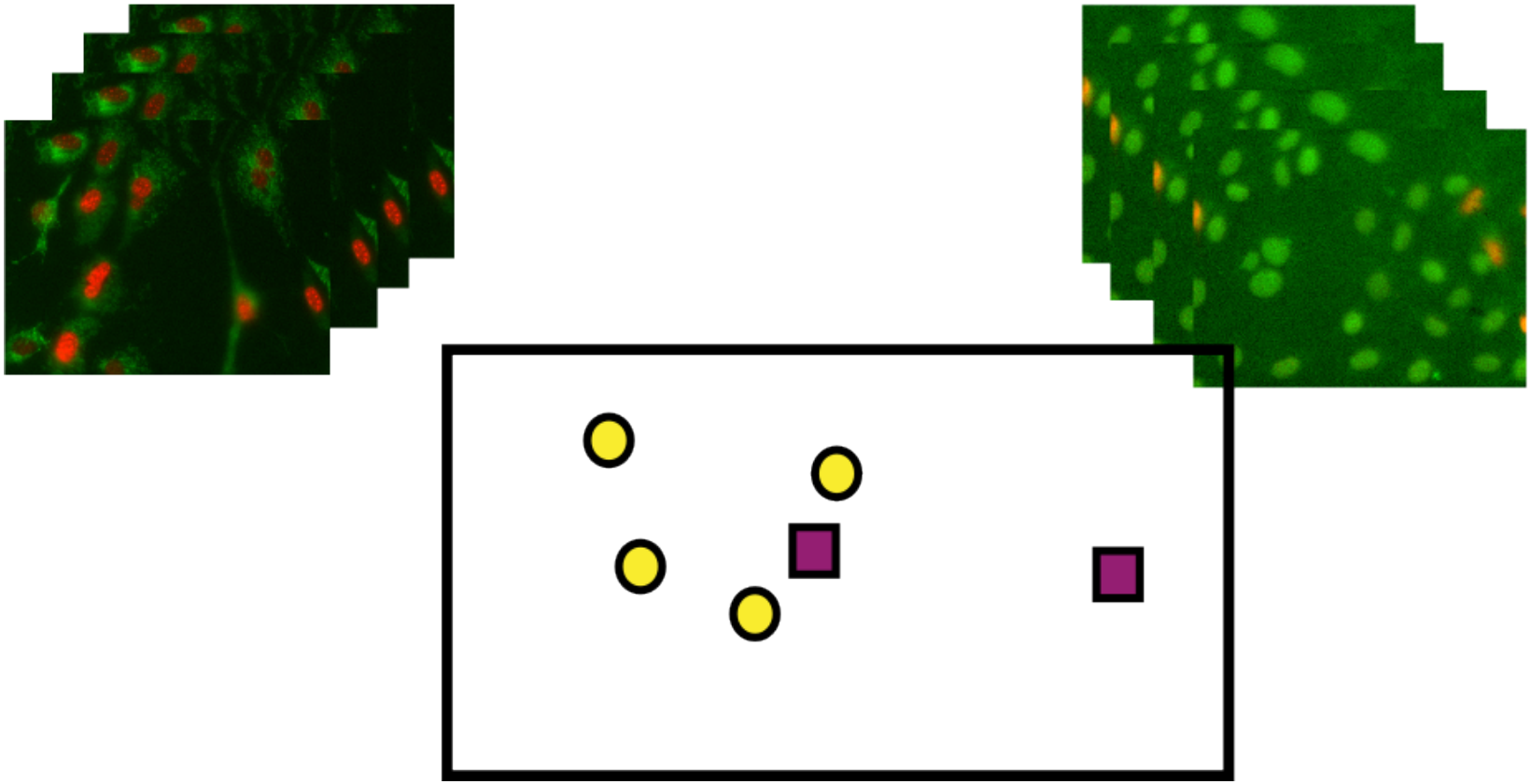
- Texture (Haralick, Gabor, ...)
- Edginess, smoothness, ...
- Local features, ...
- ...

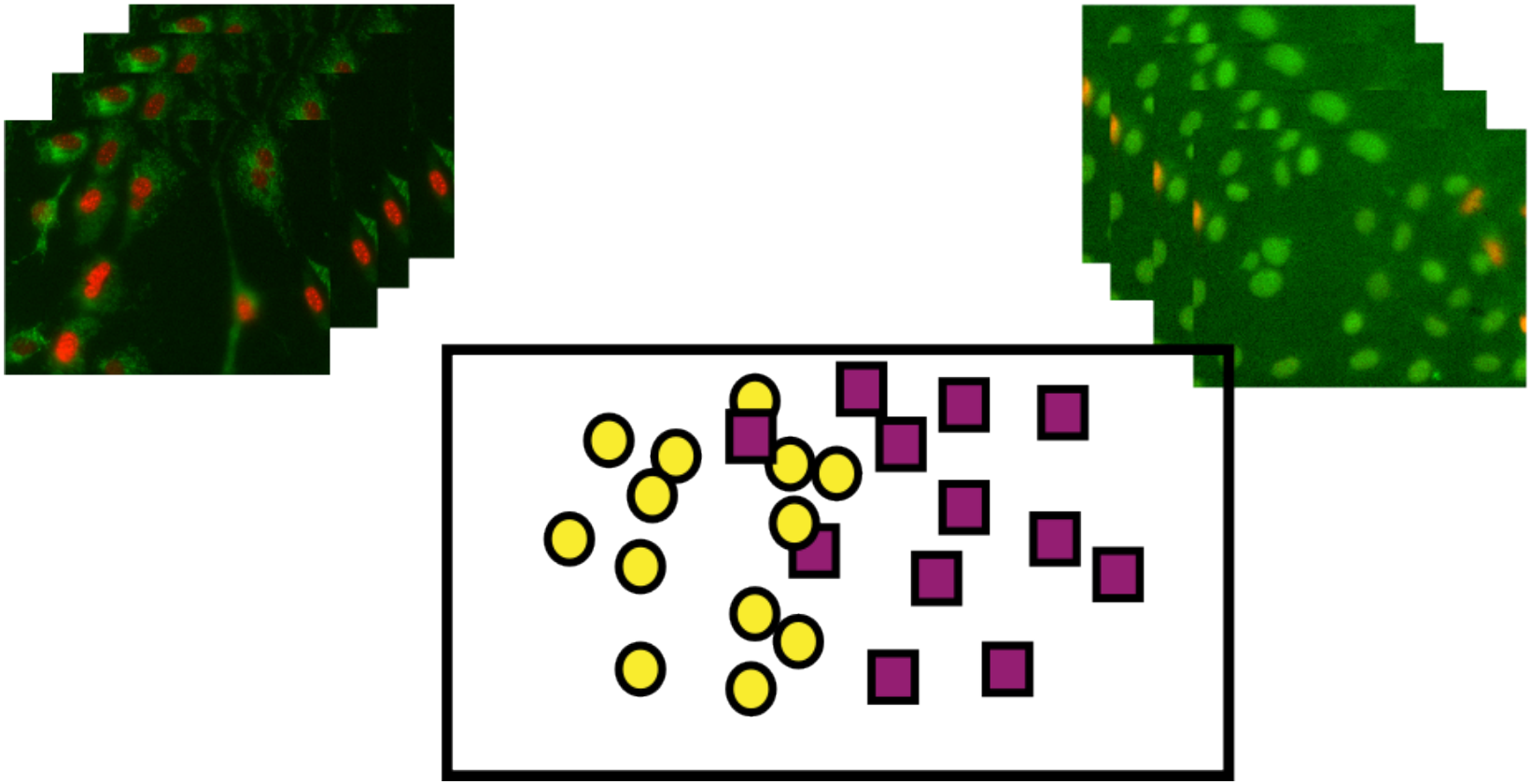
The literature is very vast.

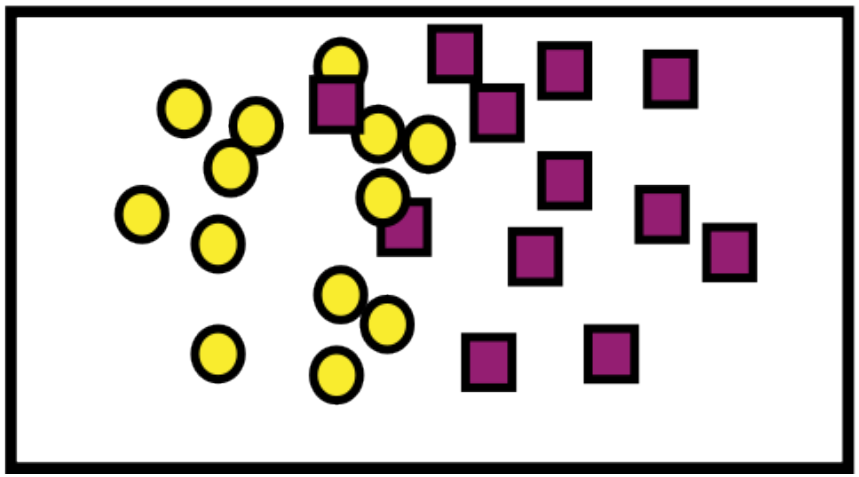


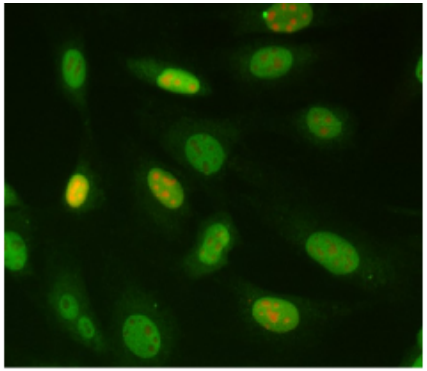


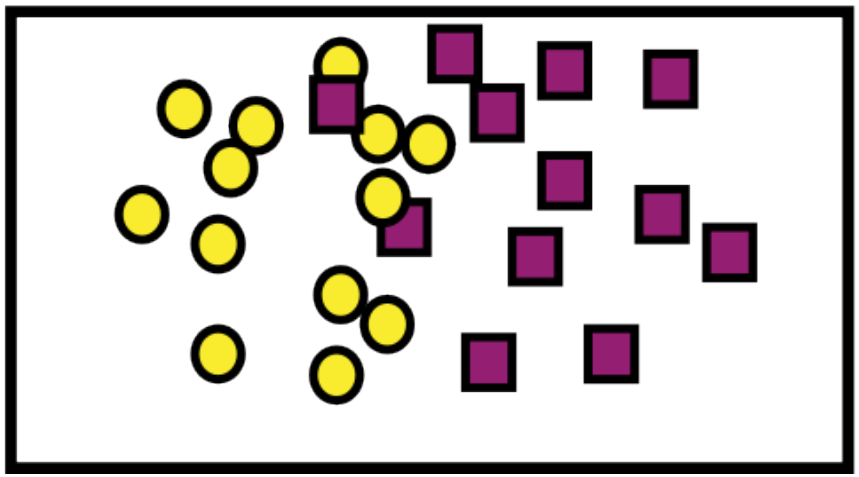


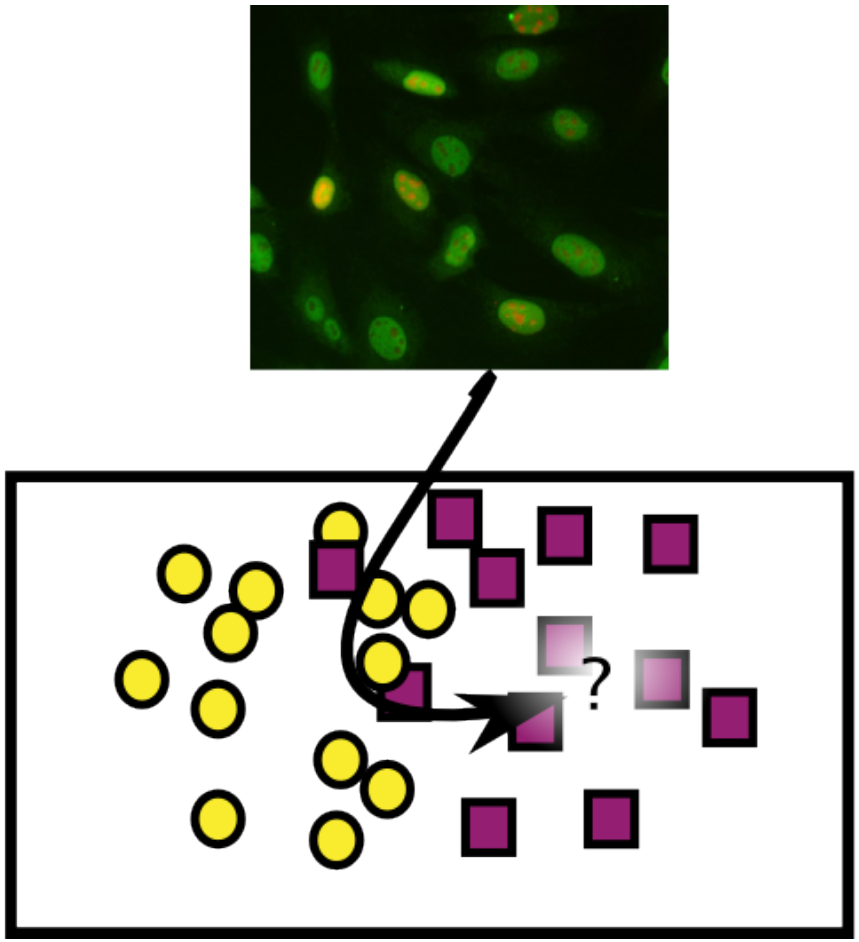




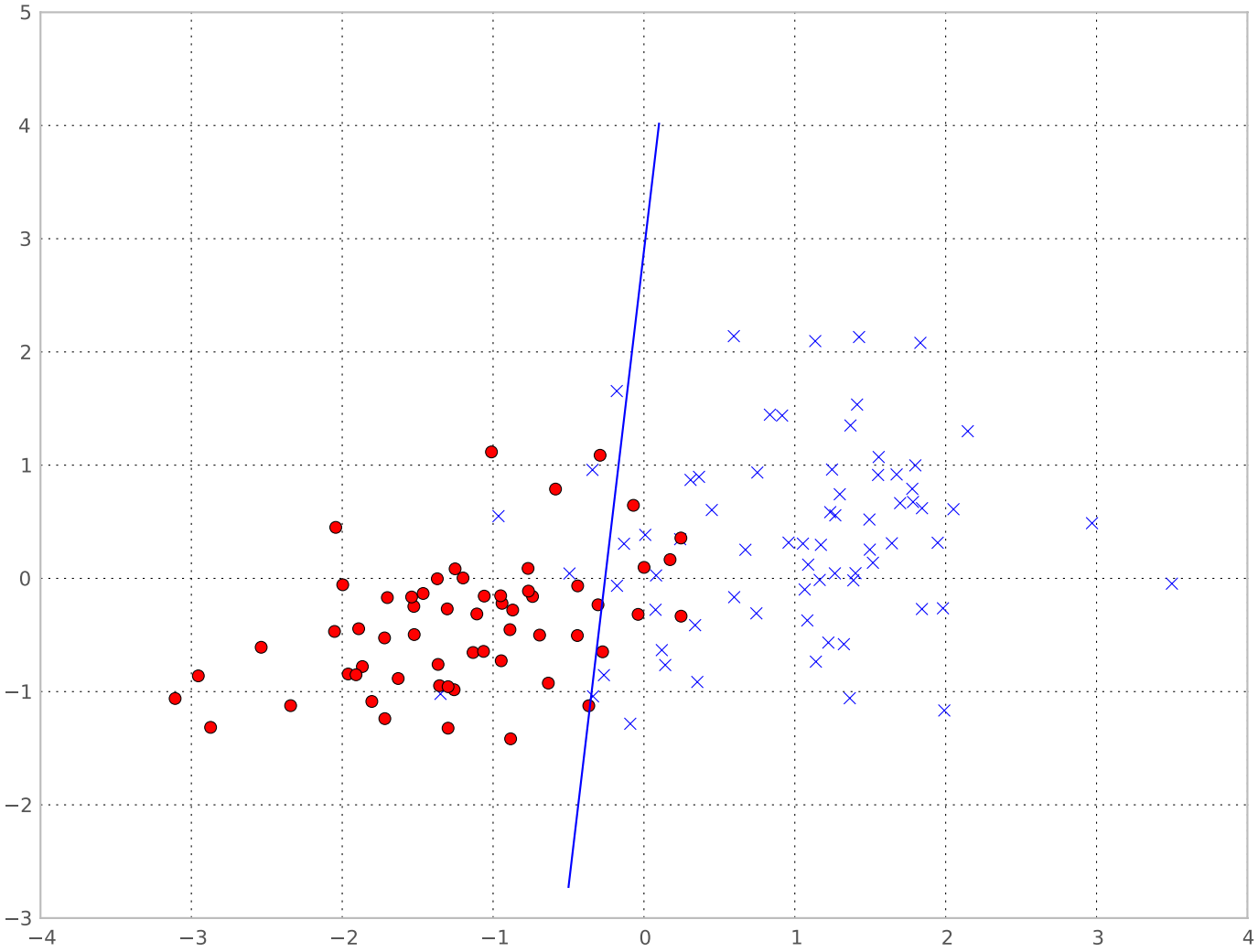




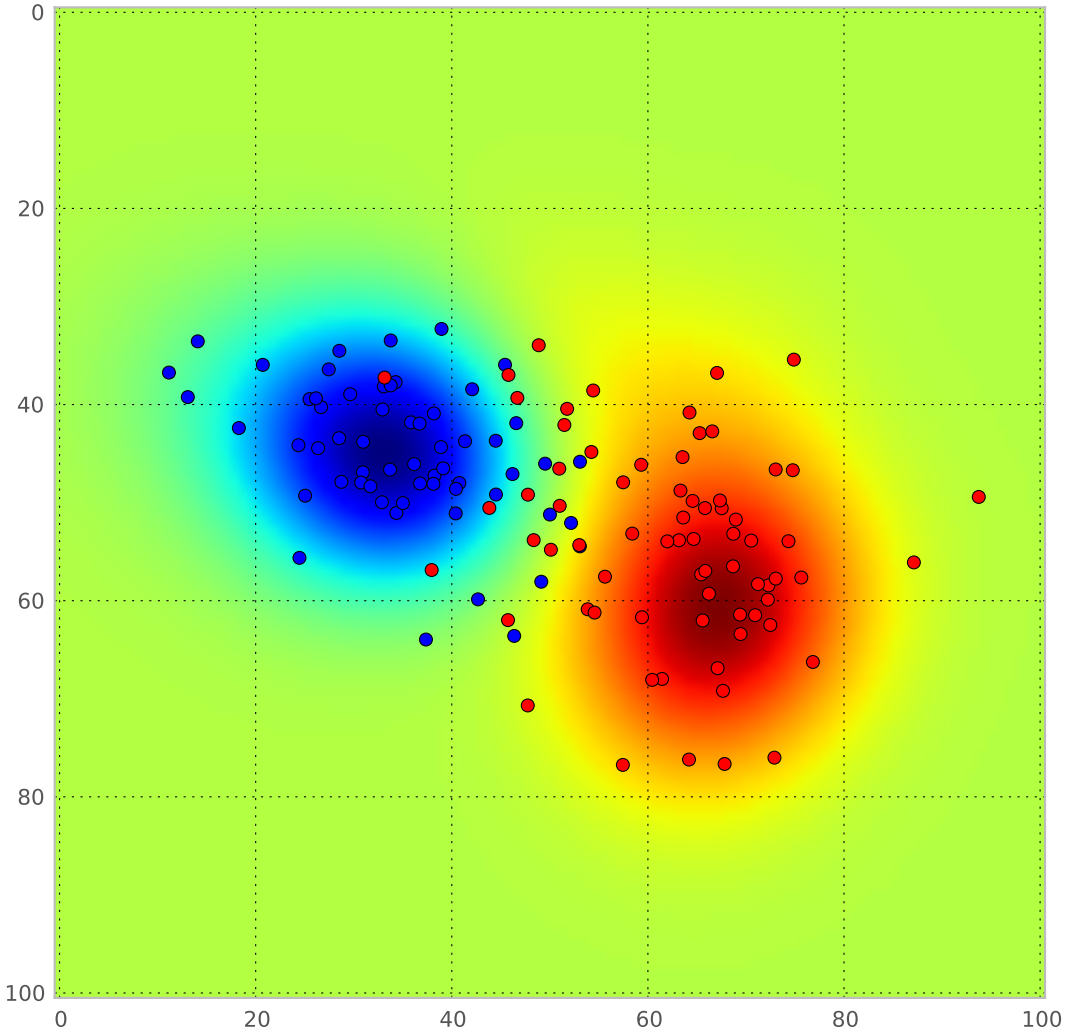




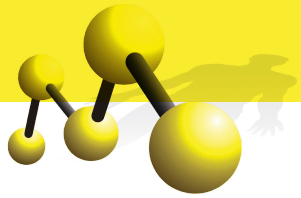
Classifiers



Classifiers



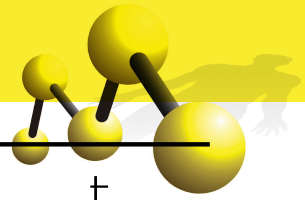
Results



	Cyto	Cytosk	Lyso	PM	Mito	N	NO
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
NO	1	0	0	0	1	16	95

Average: 72%

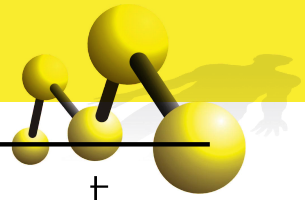
HeLa Dataset



	dna	er	gi	gii	l	m	n	a	e	t
dna	86	0	1	0	0	0	0	0	0	0
er	0	84	0	0	0	1	0	0	0	1
gi	0	0	84	2	0	1	0	0	0	0
gii	0	0	4	79	0	1	0	0	1	0
l	0	0	1	0	72	0	1	0	10	0
m	0	3	1	0	1	64	0	0	3	1
n	0	0	1	1	0	0	78	0	0	0
a	0	0	0	0	0	0	0	98	0	0
e	0	2	3	0	5	1	0	0	79	1
t	0	1	0	0	0	1	0	0	1	88

Average: 94%

HeLa Dataset



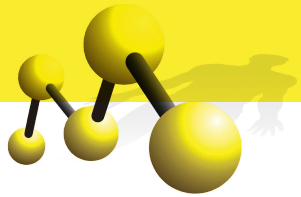
	dna	er	gi	gii	l	m	n	a	e	t
dna	86	0	1	0	0	0	0	0	0	0
er	0	84	0	0	0	1	0	0	0	1
gi	0	0	84	2	0	1	0	0	0	0
gii	0	0	4	79	0	1	0	0	1	0
l	0	0	1	0	72	0	1	0	10	0
m	0	3	1	0	1	64	0	0	3	1
n	0	0	1	1	0	0	78	0	0	0
a	0	0	0	0	0	0	0	98	0	0
e	0	2	3	0	5	1	0	0	79	1
t	0	1	0	0	0	1	0	0	1	88

Average: 94%

Human performance: 83%

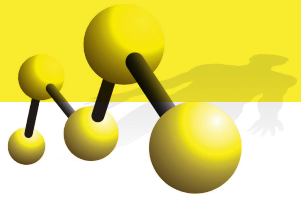
(Murphy et al., 2003)

Typical Results



- Comparable to or better than human!
- Better with multiple replicates.
- Classification times: a few seconds per image.

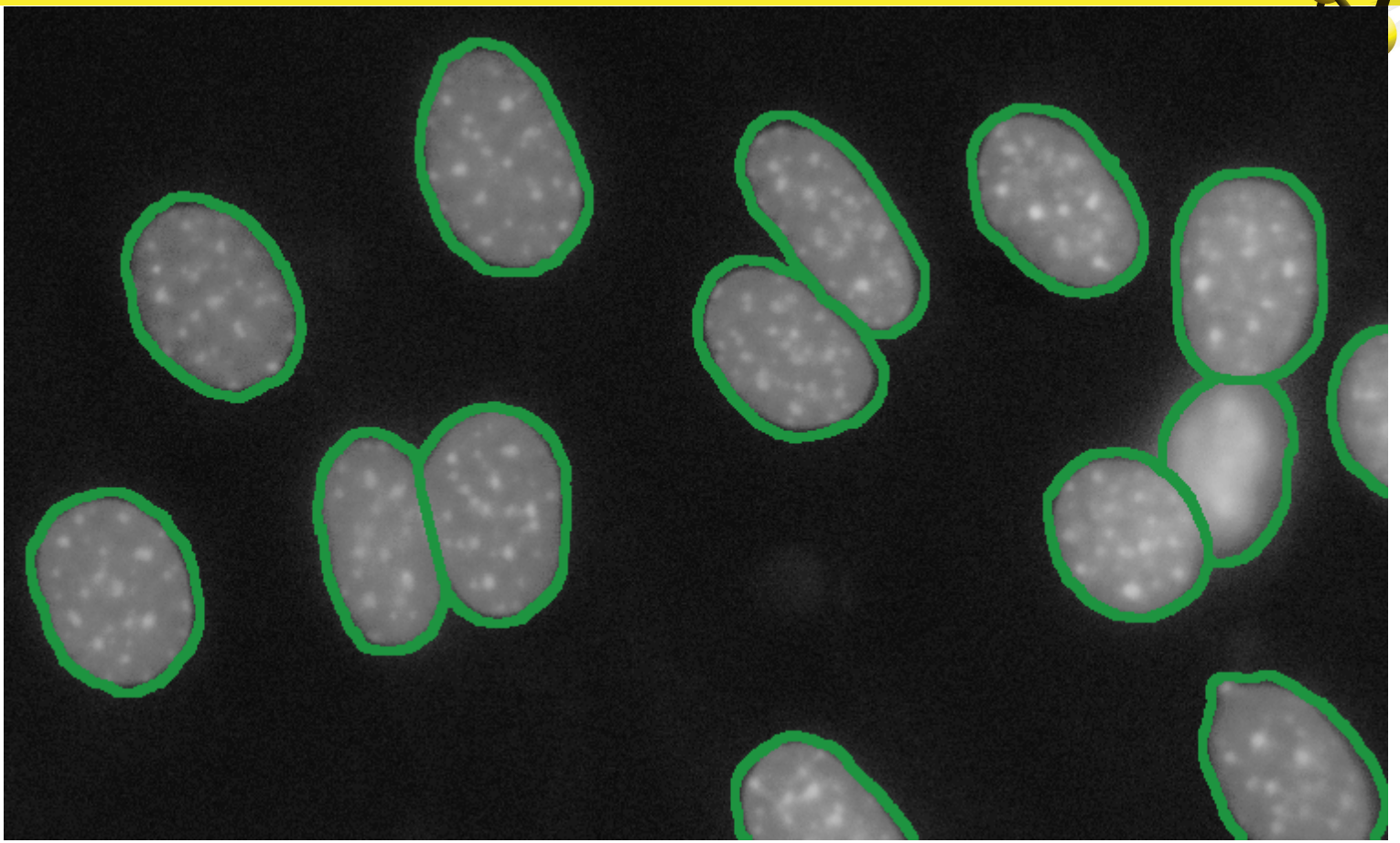
Other Problems



Other Typical Classification Problems

- Phenotype in a screen
- Stem cell differentiation
- ...

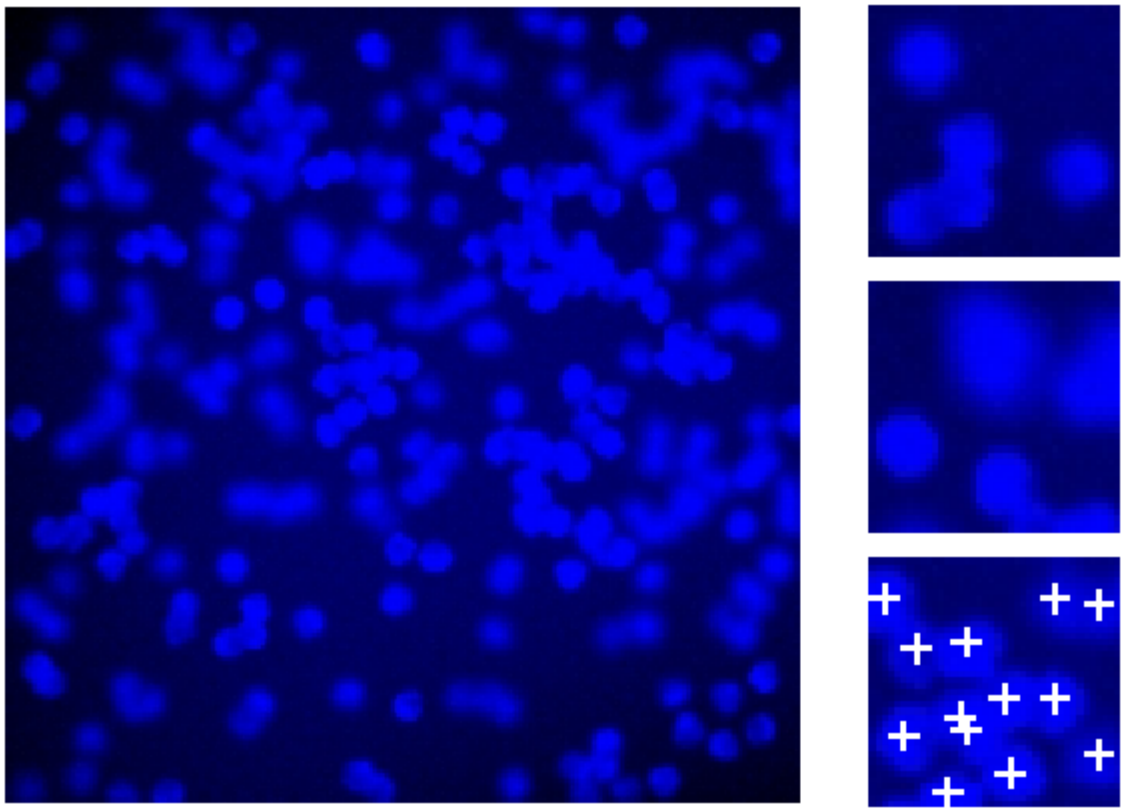
Segmentation as Classification



(Coelho et al., 2009)

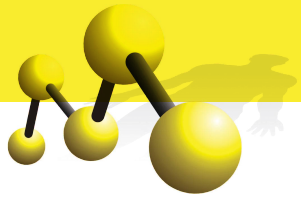
(Chen et al., 2011)

Learning to Count



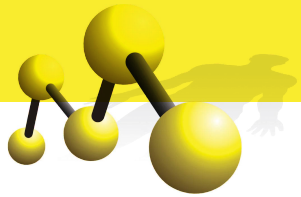
(Lempitsky & Zisserman, 2010)

Conclusions



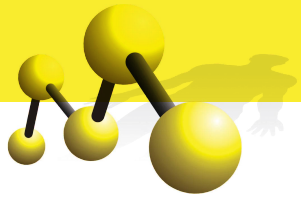
- Computers can do very well at classification.
- Flexible tool if you have the training data.

Mixture Patterns Classification



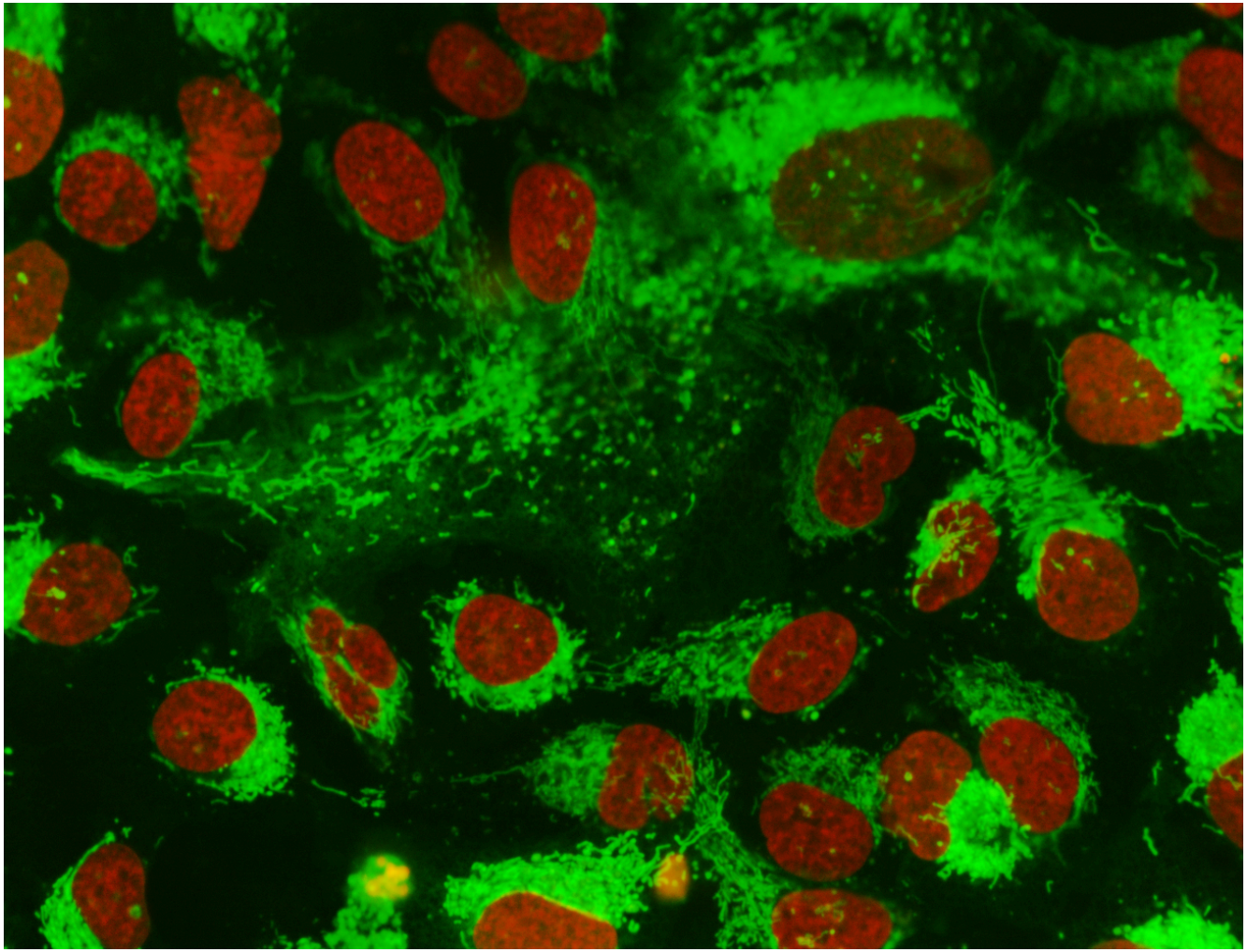
Previously reported methods work well for simple classes, like “endosomes” or “mitochondria.”

Mixture Patterns Classification

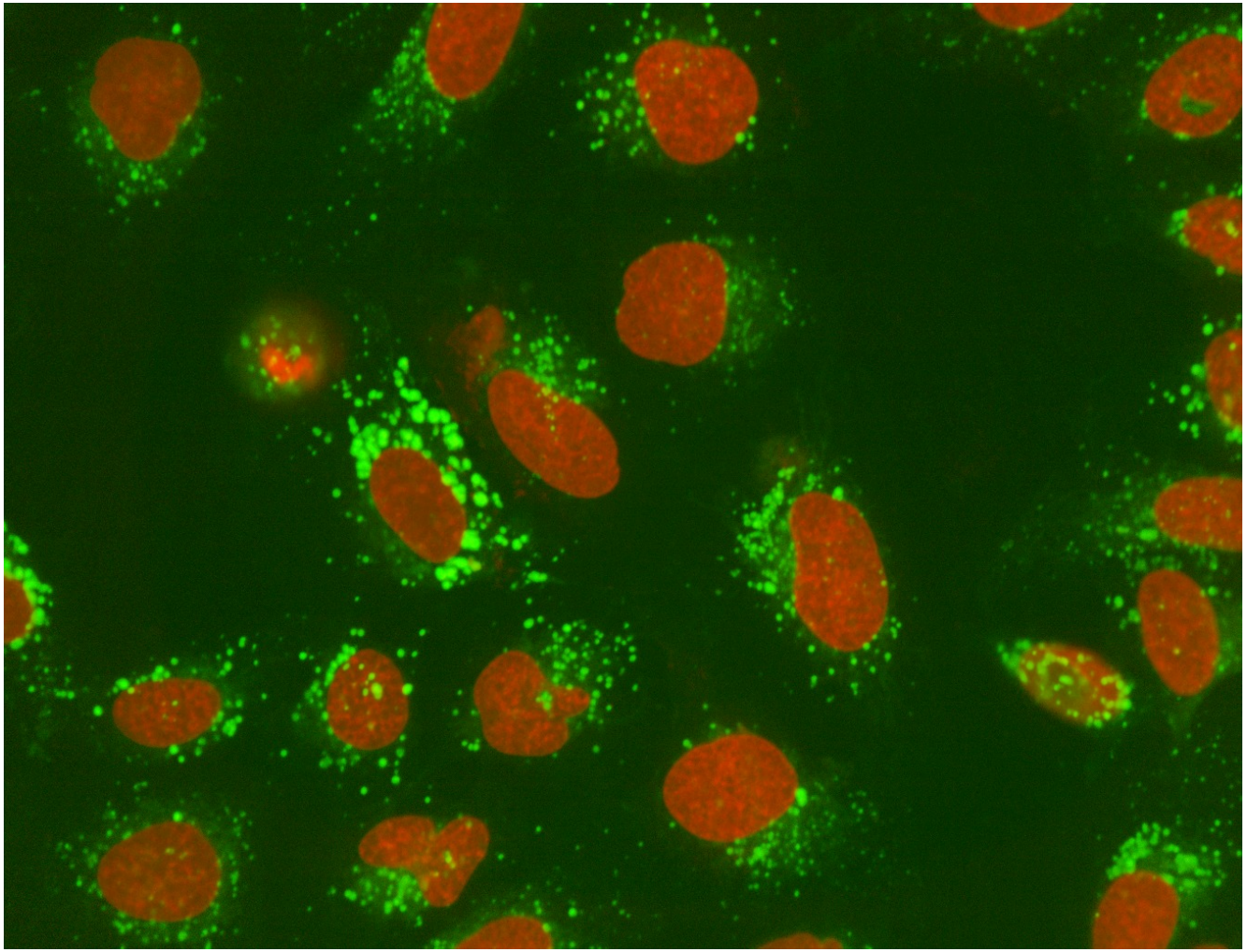


Previously reported methods work well for simple classes, like “endosomes” or “mitochondria.”
What if a protein is present in both endosomes and mitochondria?

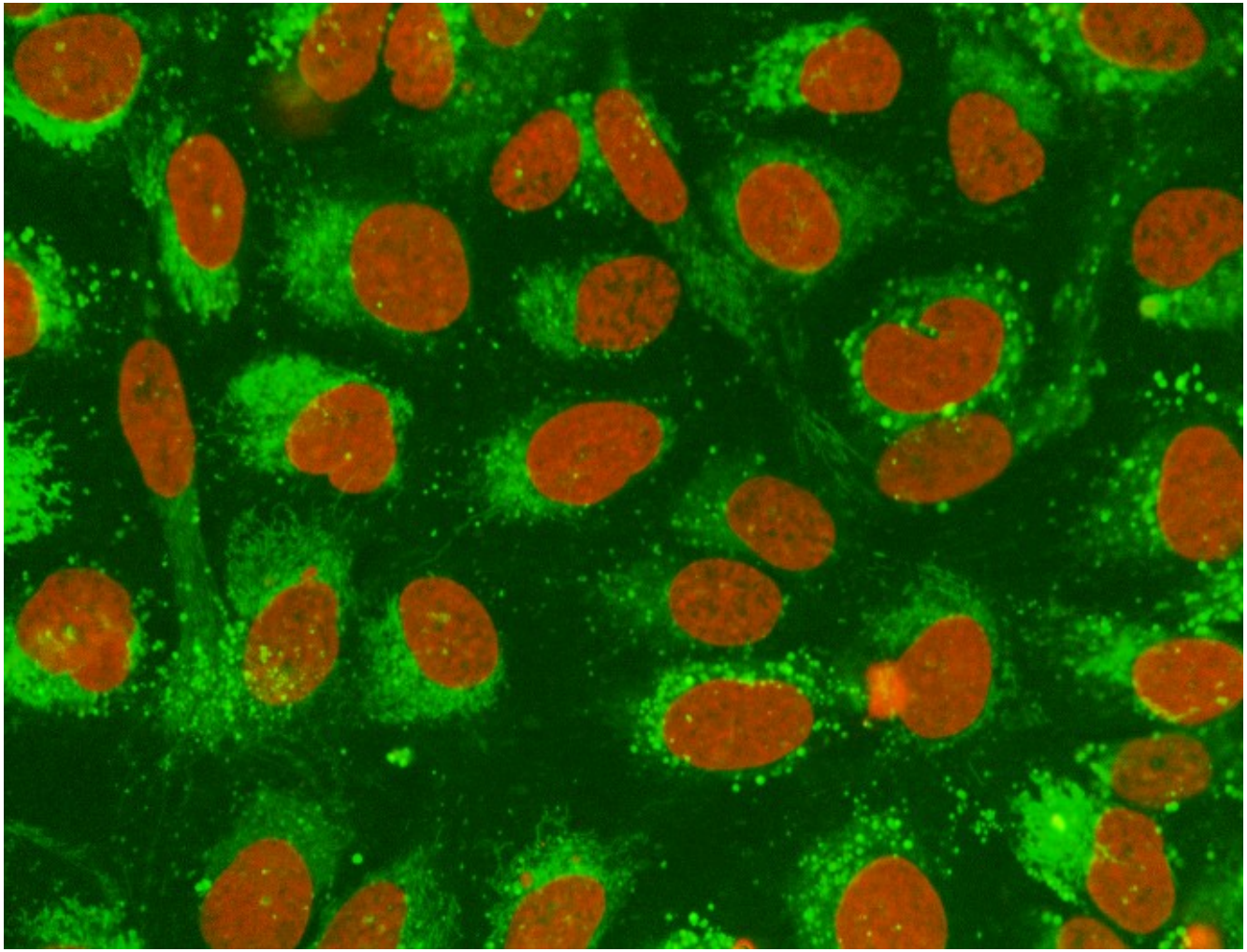
Mixture Pattern Example



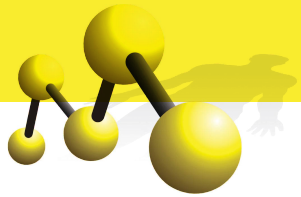
Mixture Pattern Example



Mixture Pattern Example

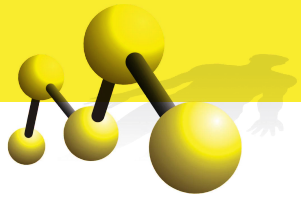


Supervised Unmixing Problem



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

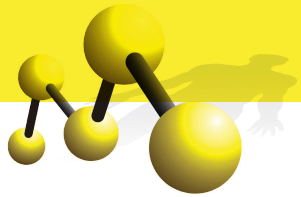
Supervised Unmixing Problem



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.

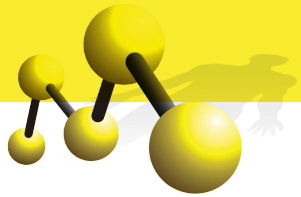
(T. Zhao et al., 2005)
(T. Peng, G. Bonami et al., 2010)

Unsupervised Unmixing Problem



What if we don't know the pure patterns?

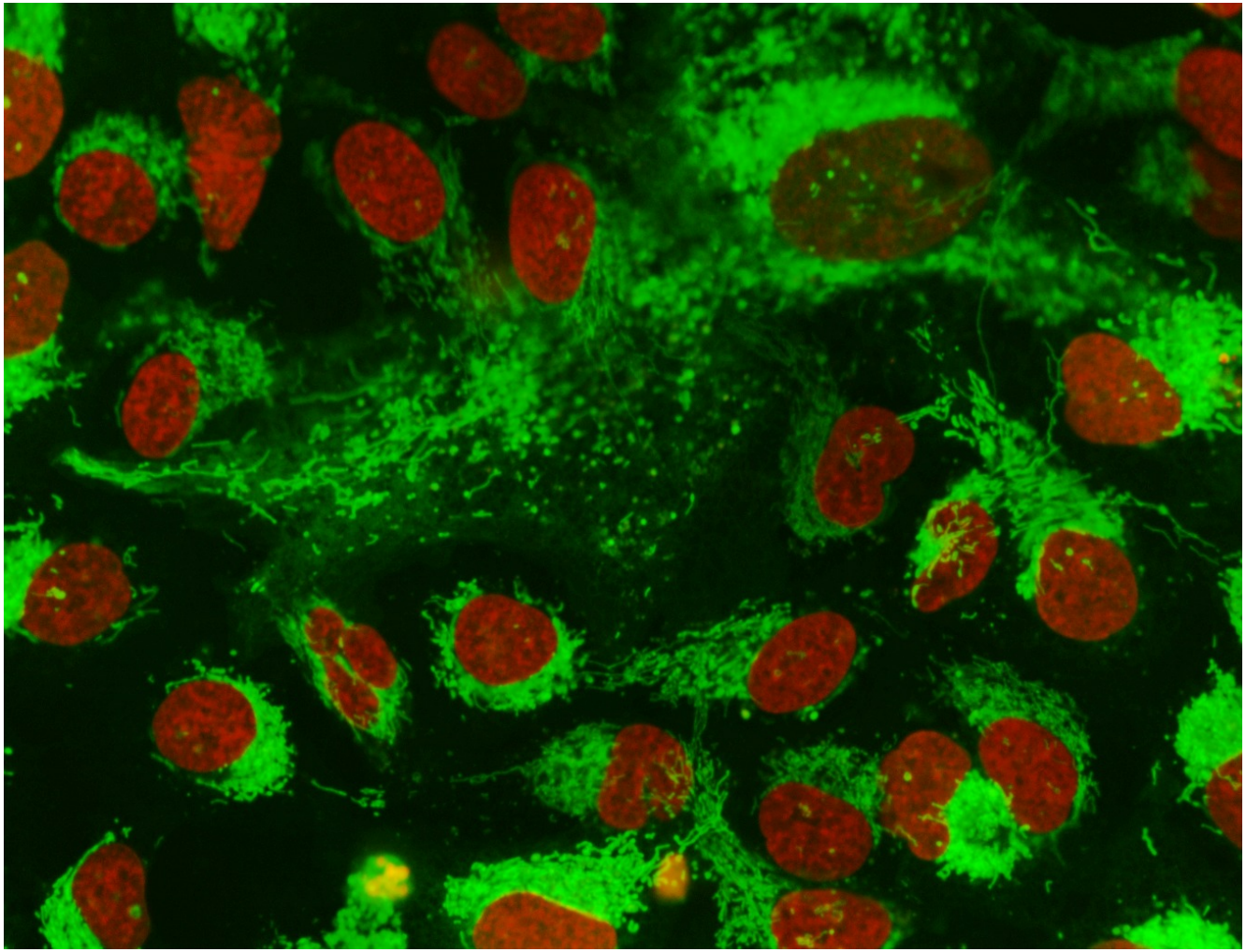
Unsupervised Unmixing Problem



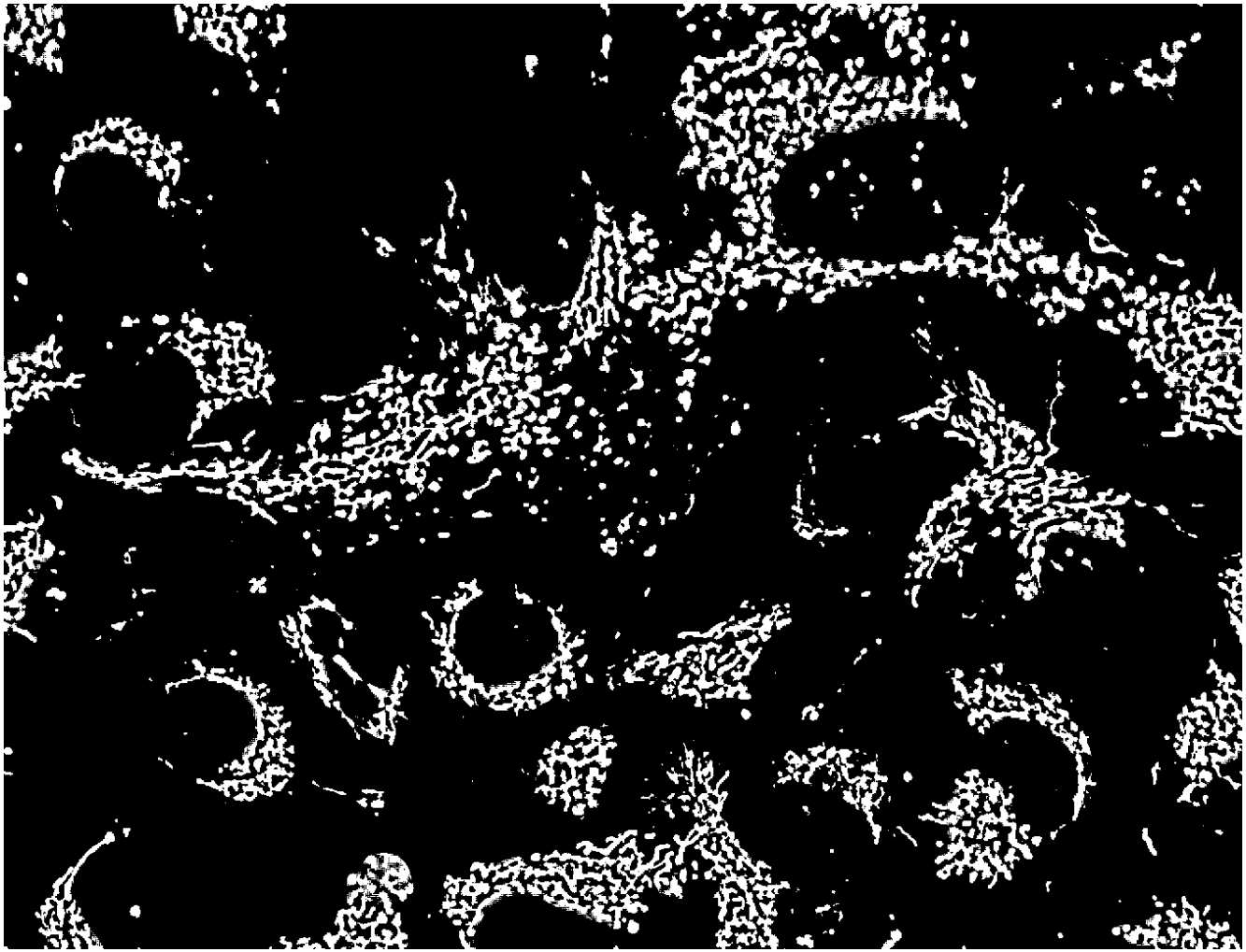
What if we don't know the pure patterns?

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

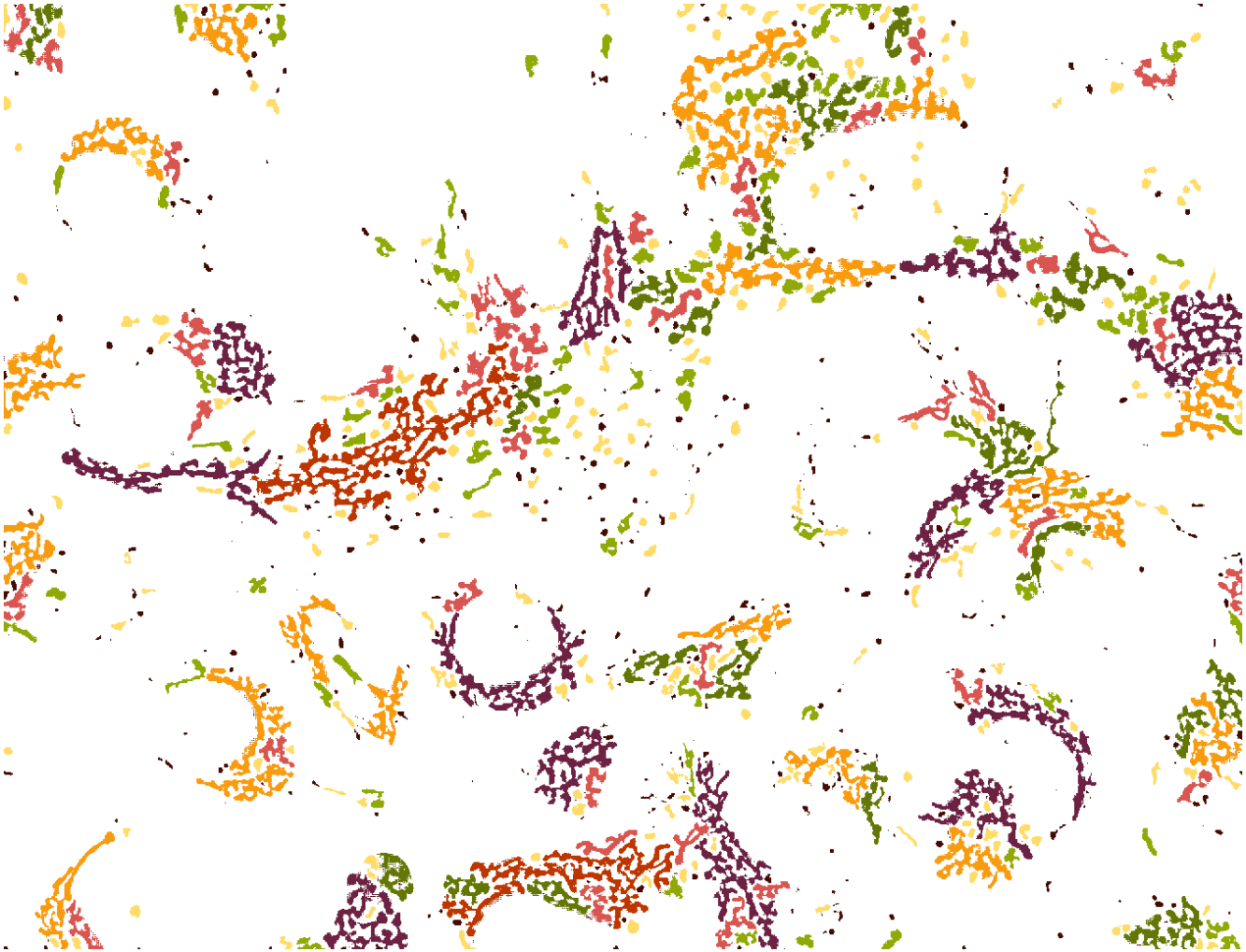
Process



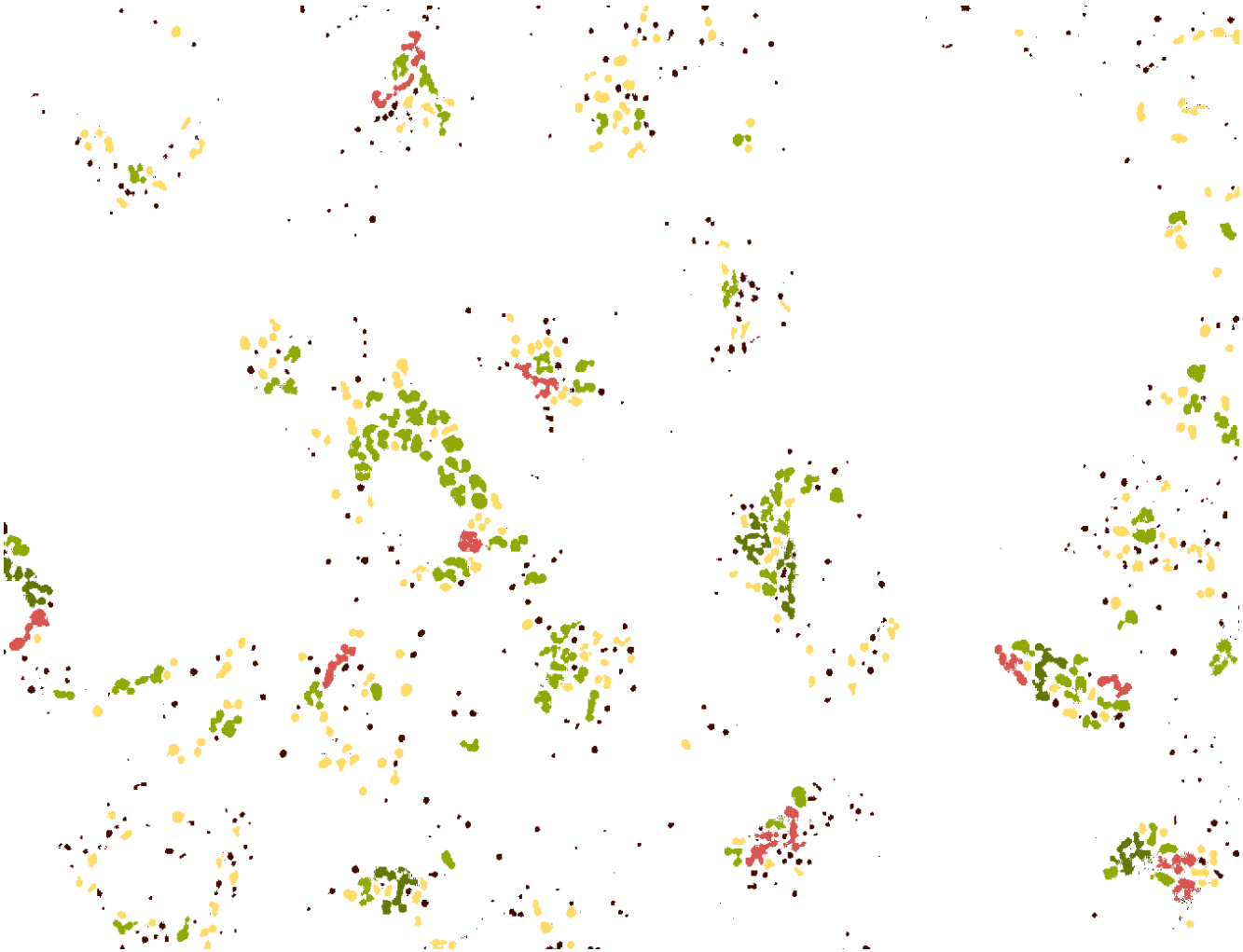
Process



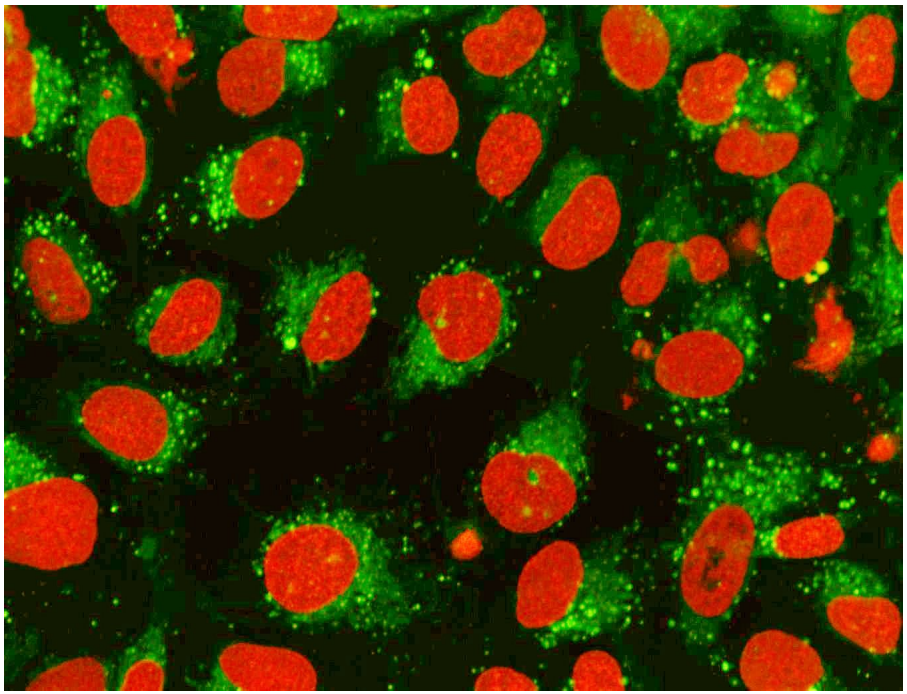
Process

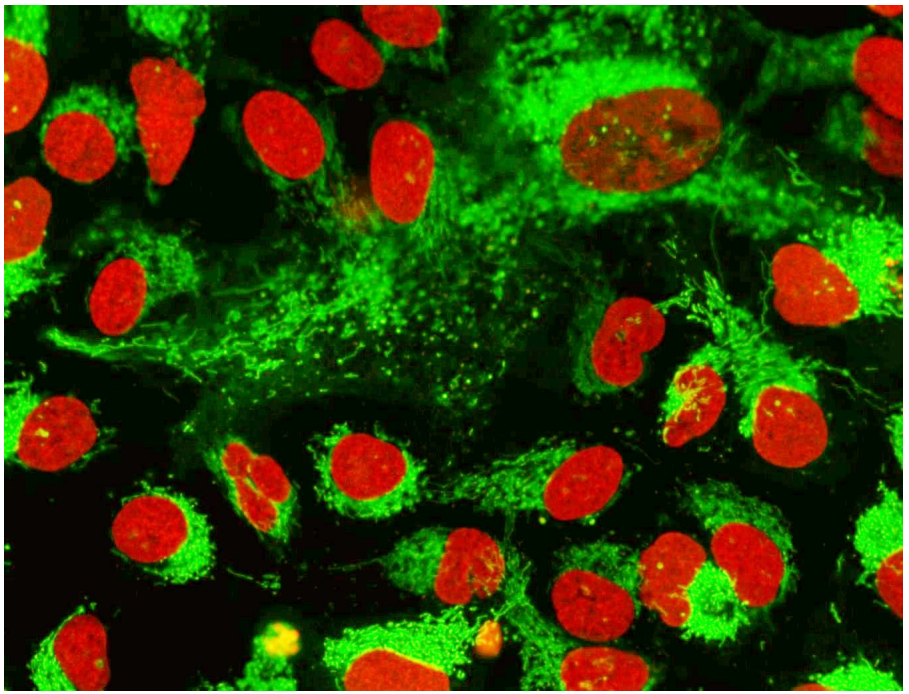


Process



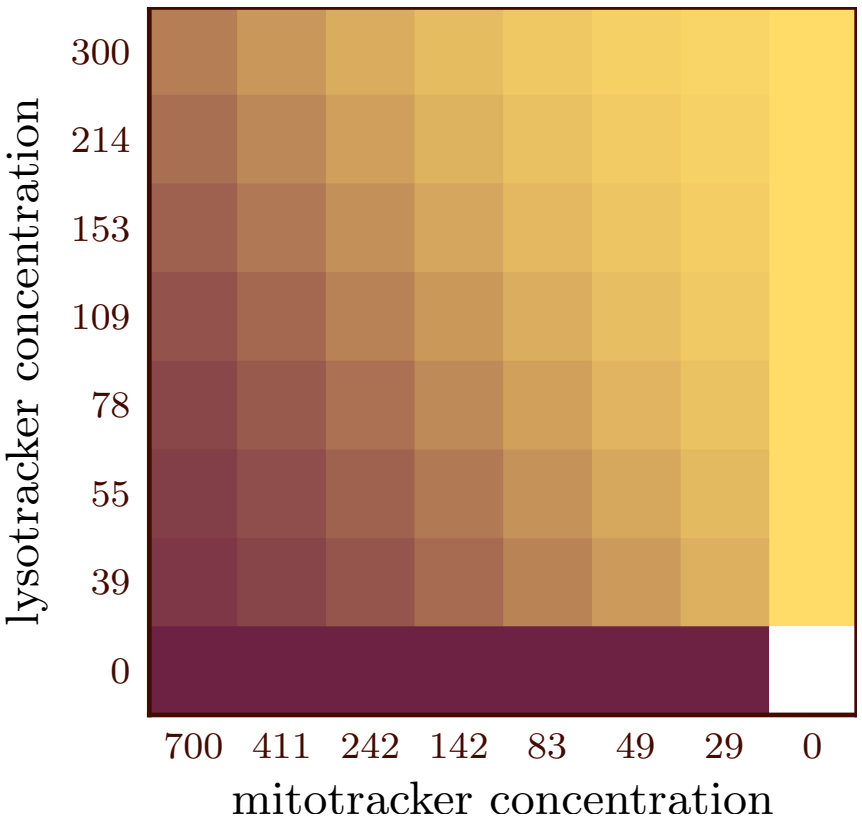
Results: Mixing Bases



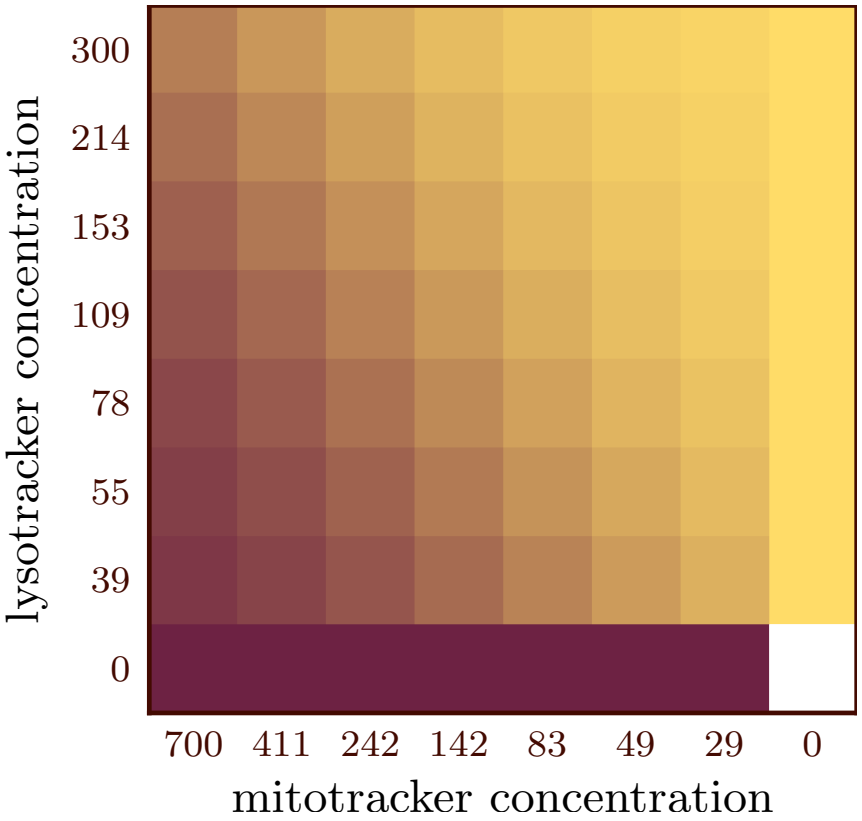


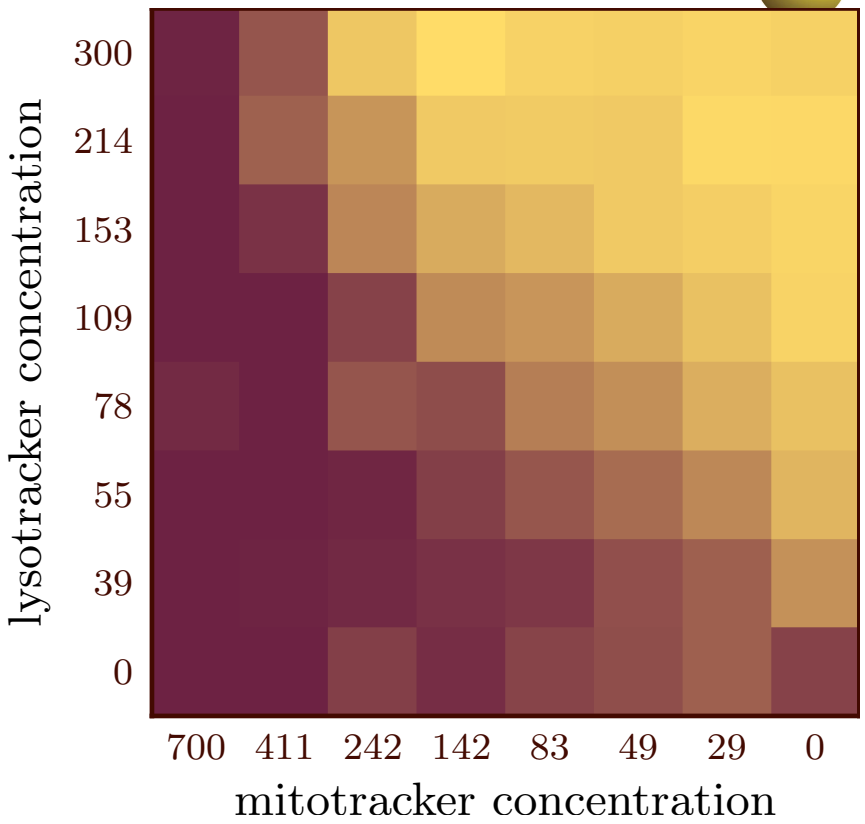
(Coelho et al., 2010)

Results: Mixing Fractions



Results: Mixing Fractions



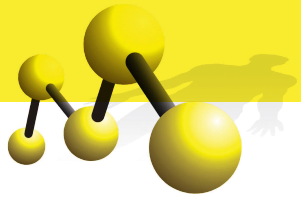


Correlation: 91%

(Coelho et al., 2010)

- Pattern unmixing works both in supervised and unsupervised modes.

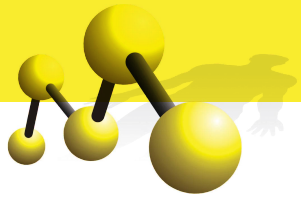
Other Heterogeneous Problems



Problems

- Multiple cells in a field
- Multiple cells in a tissue
- ...

Multiple Heterogeneous Cells

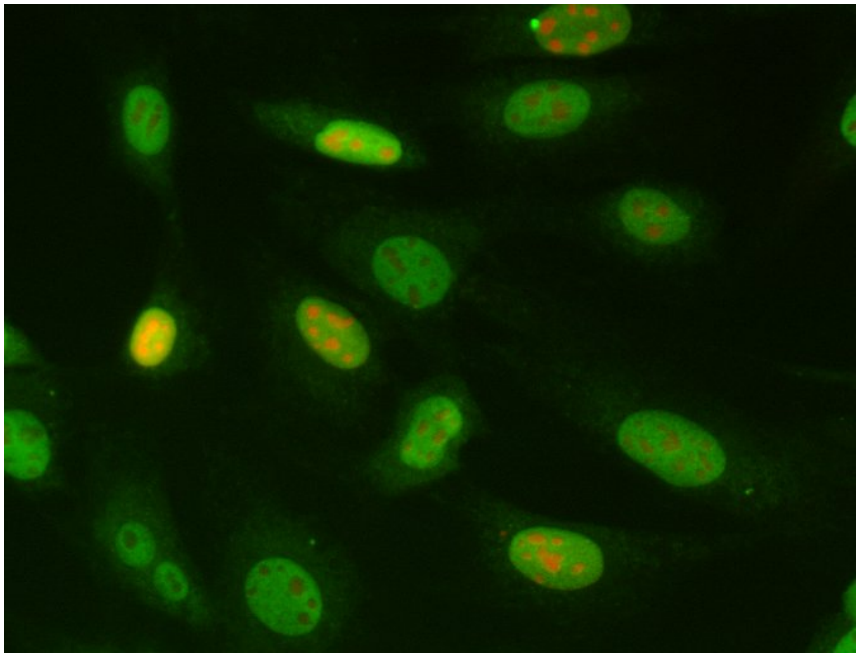


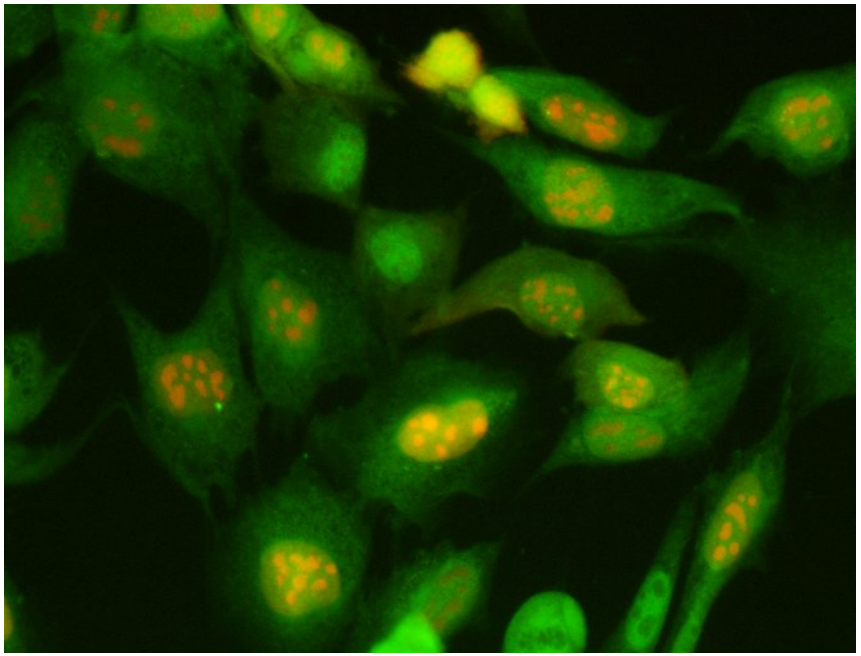
Approach

1. Segment cells
2. Classify cells independently
3. Group classifications

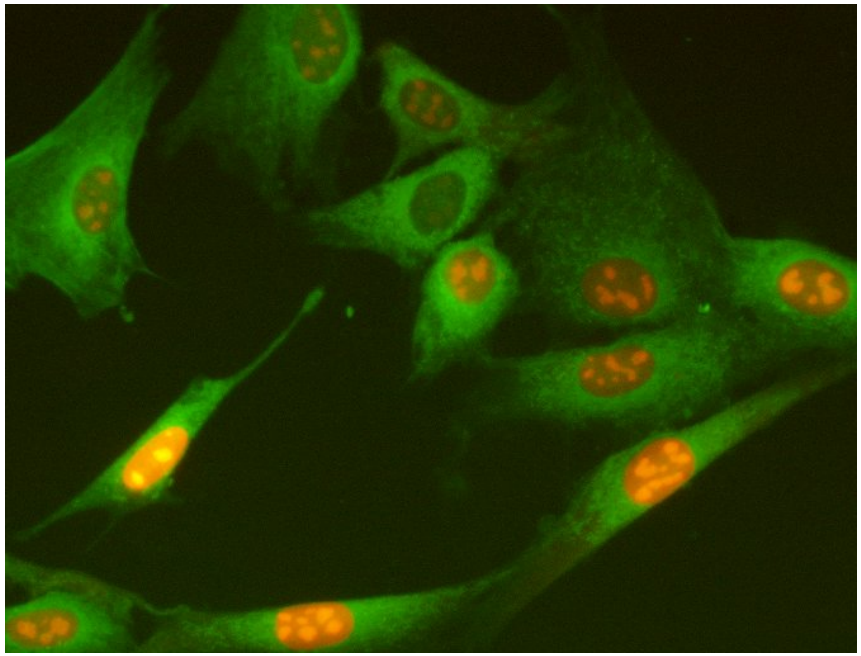
(Altschuler & Wu, 2010)

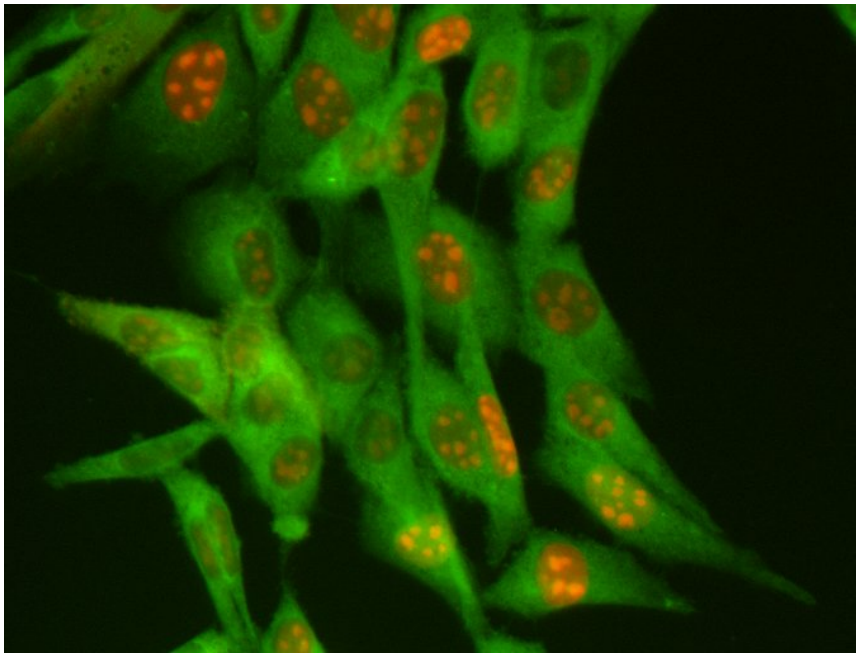
Positive Example

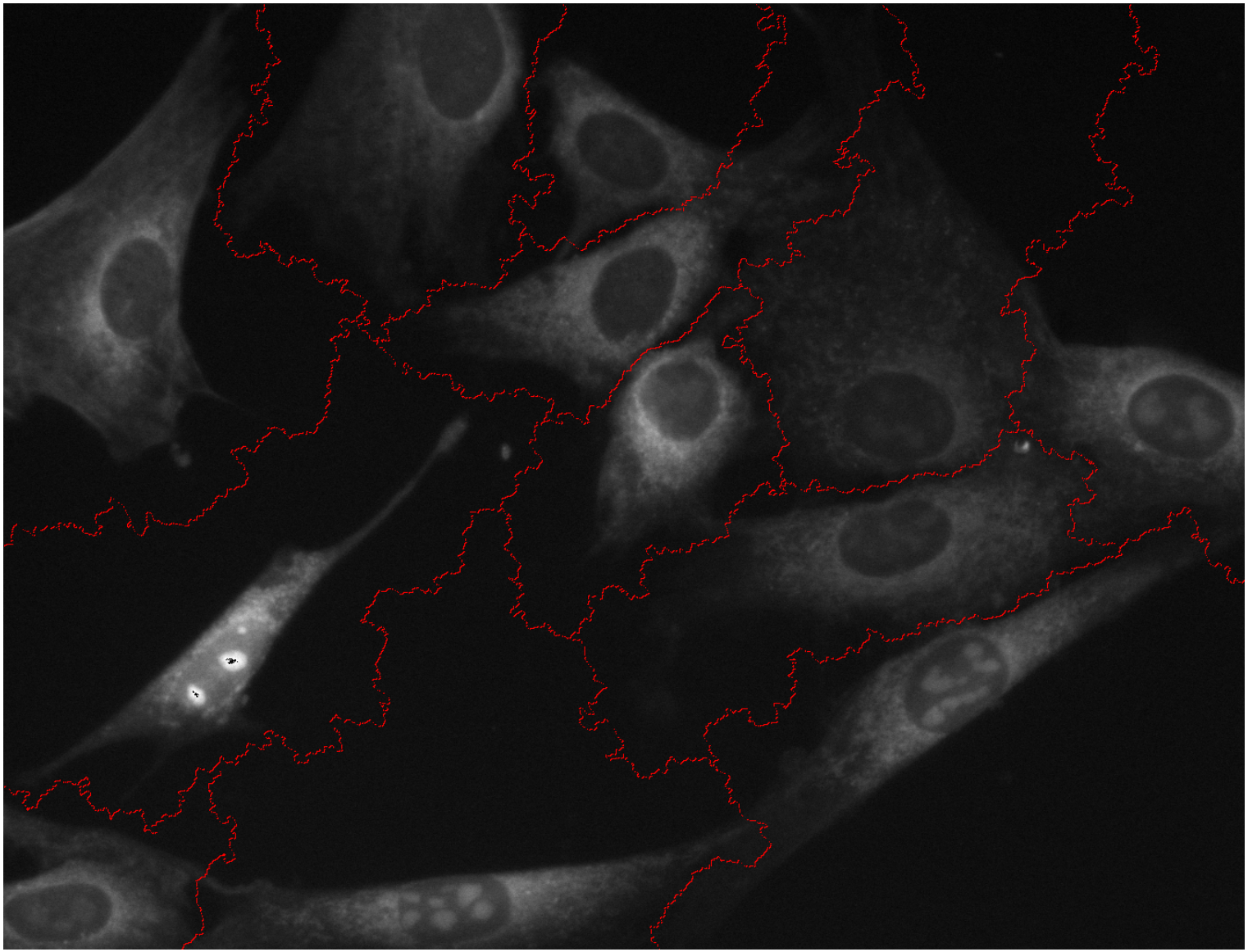




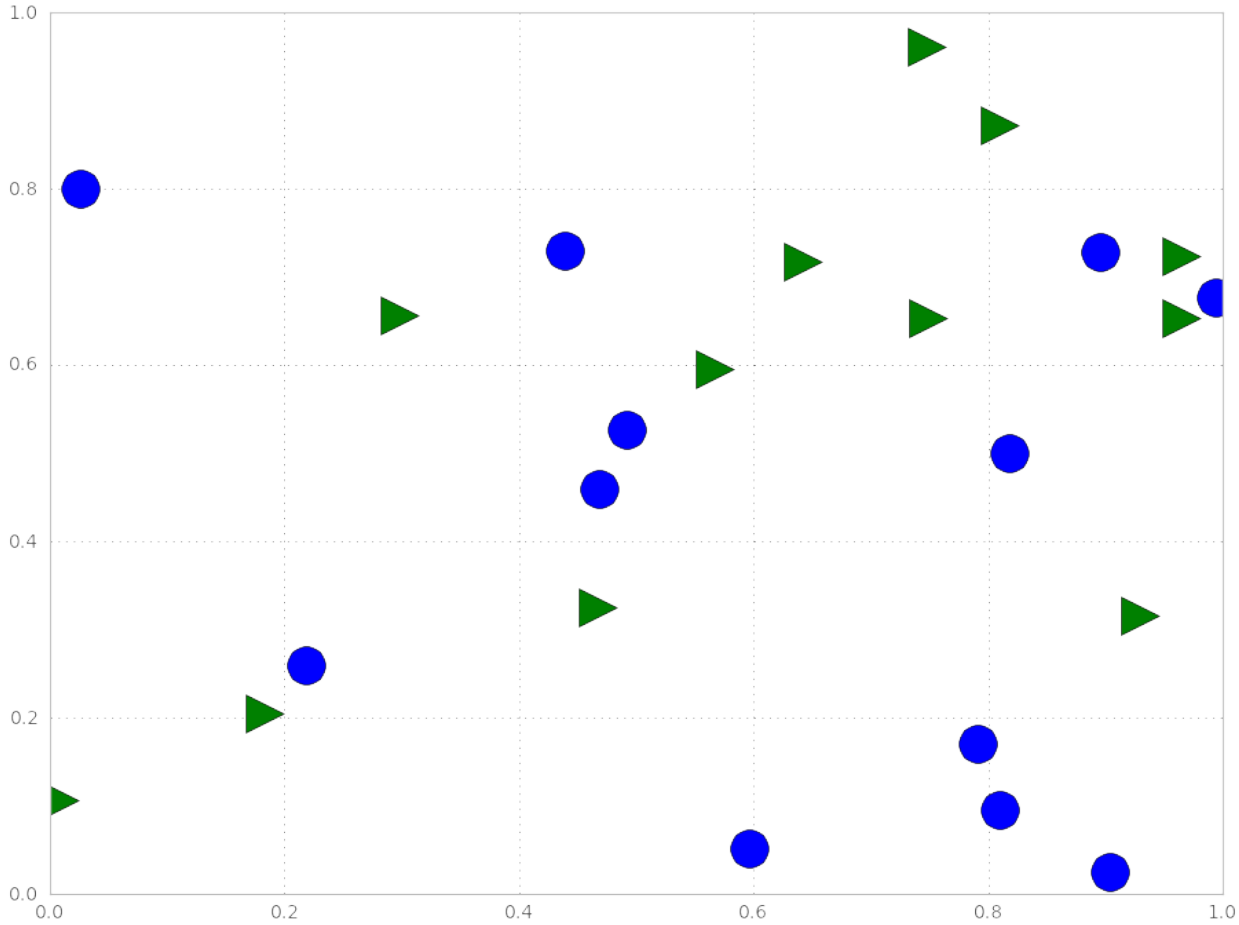
Negative Example







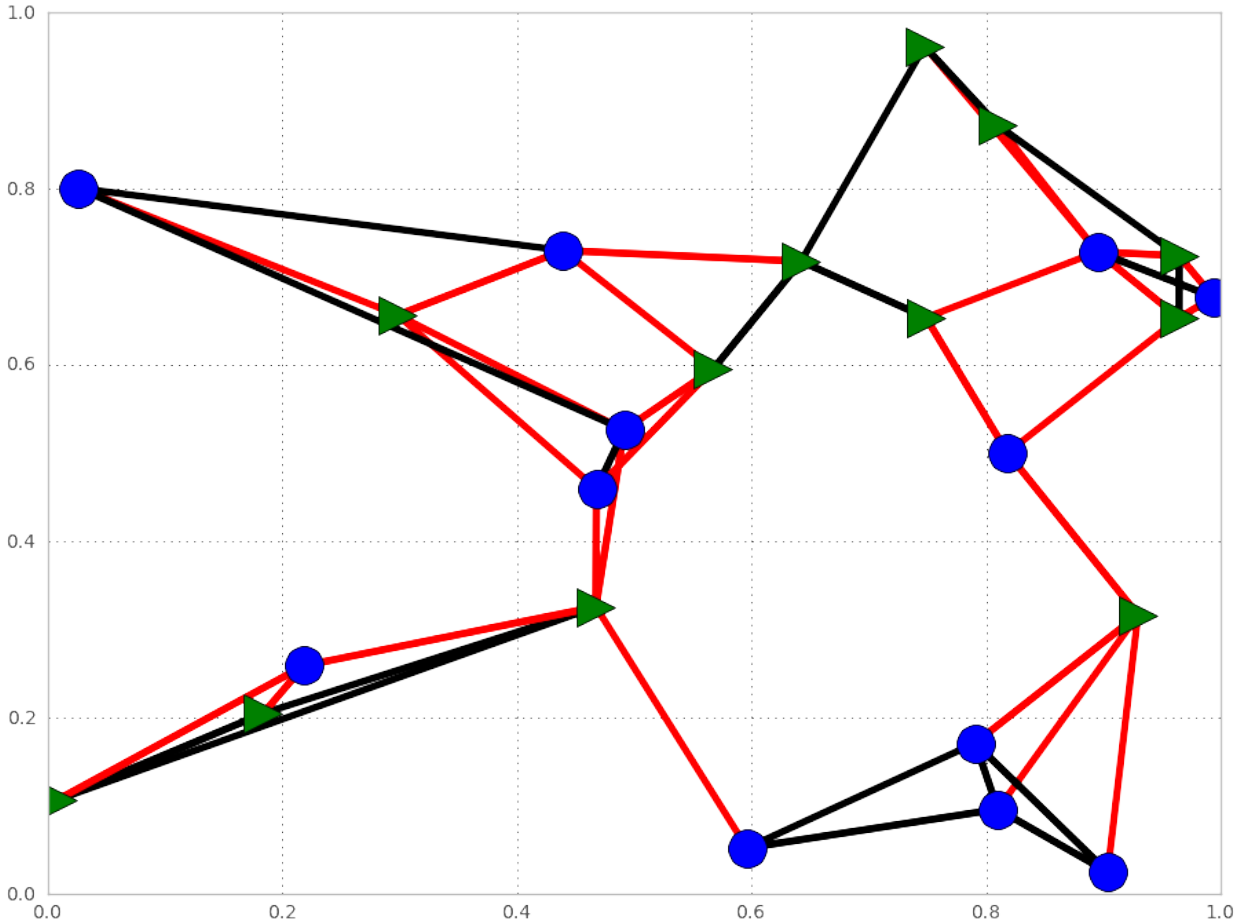
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

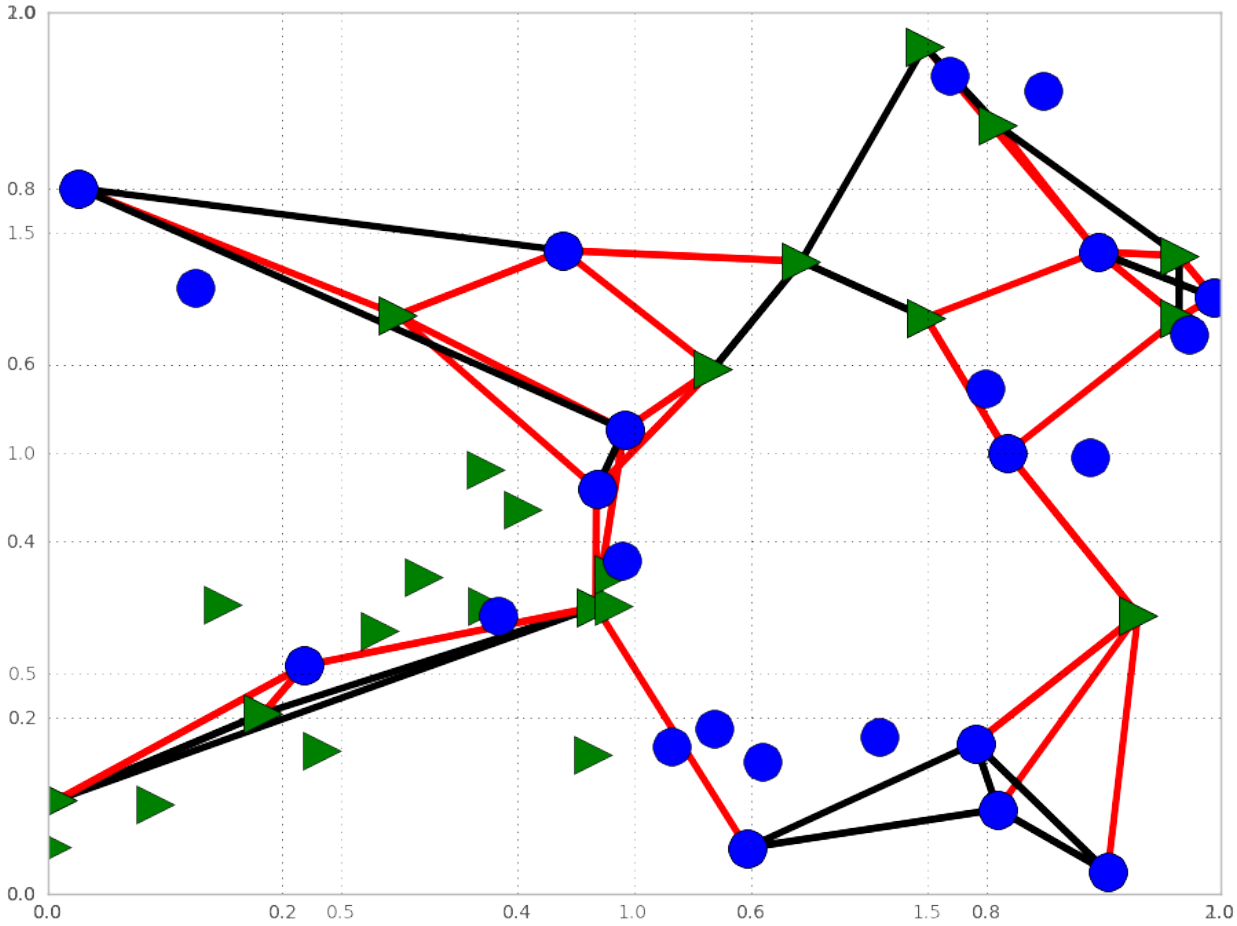
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

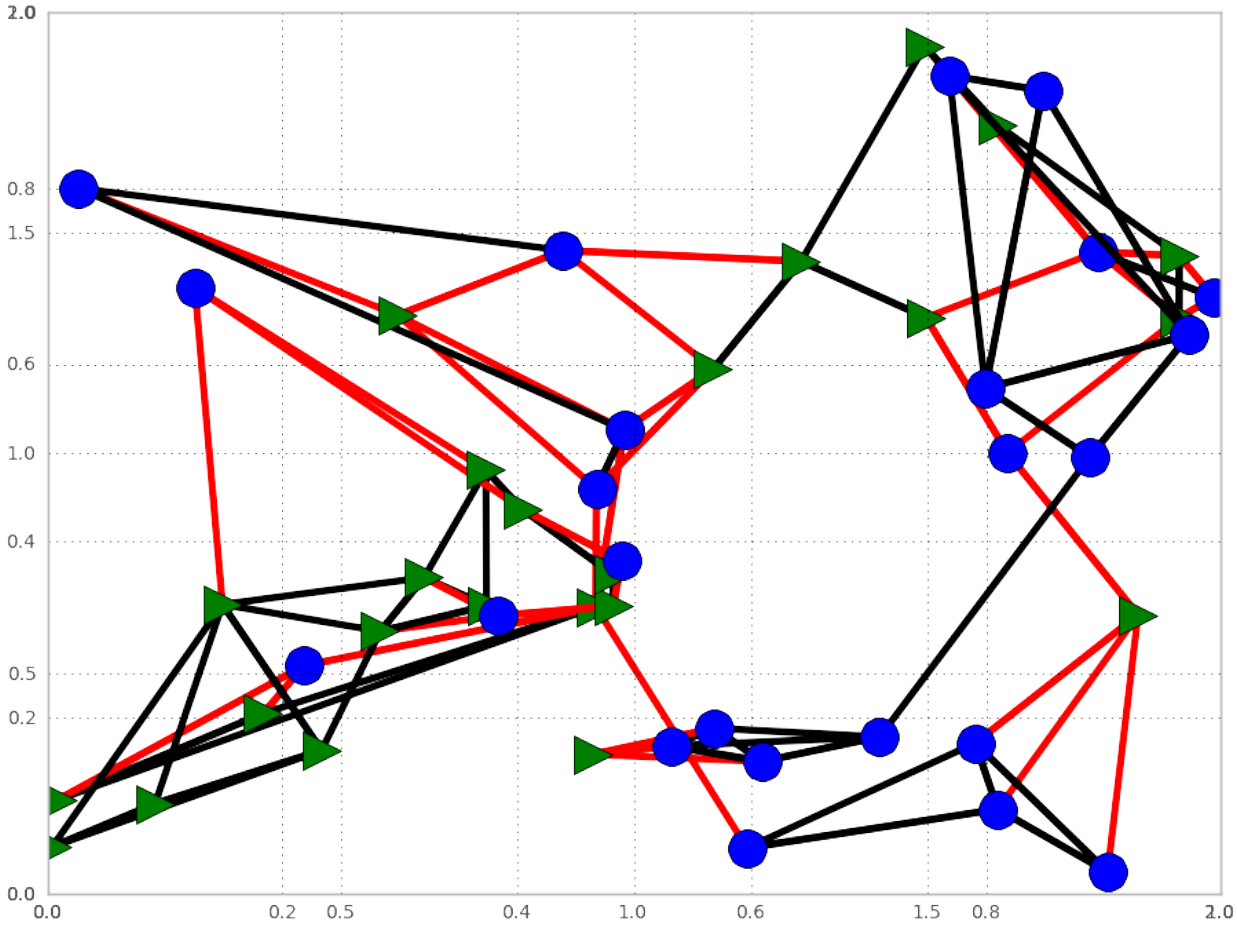
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

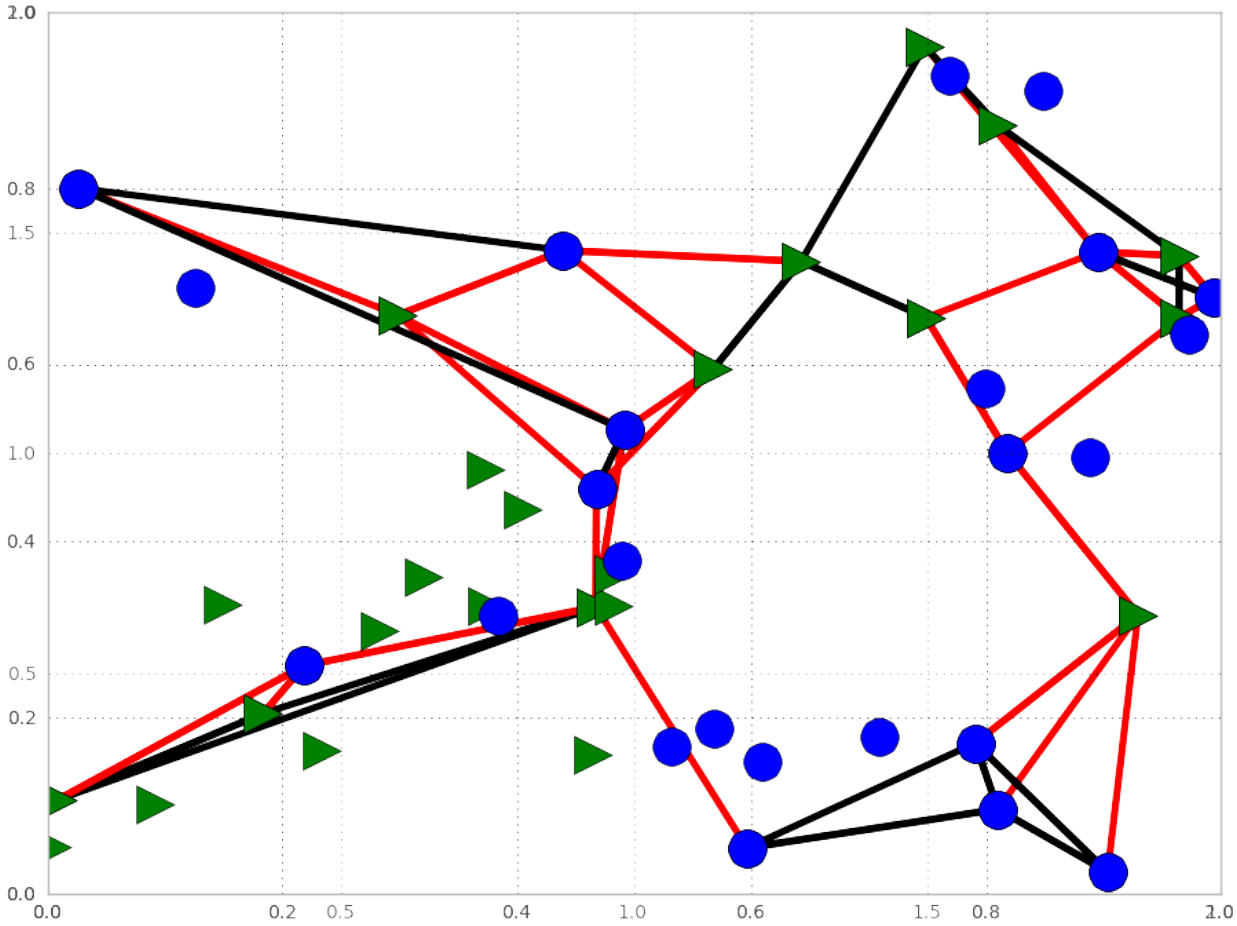
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

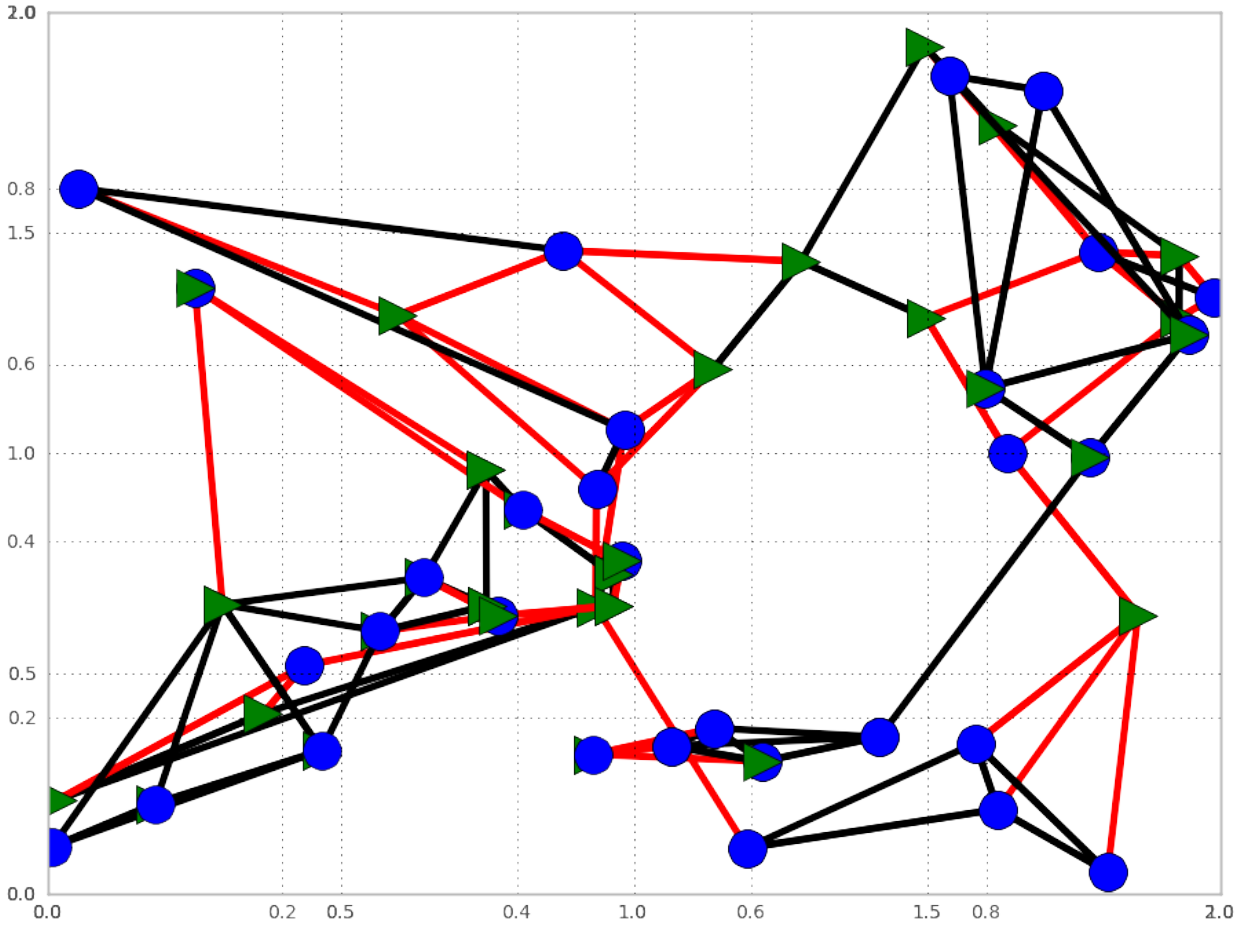
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

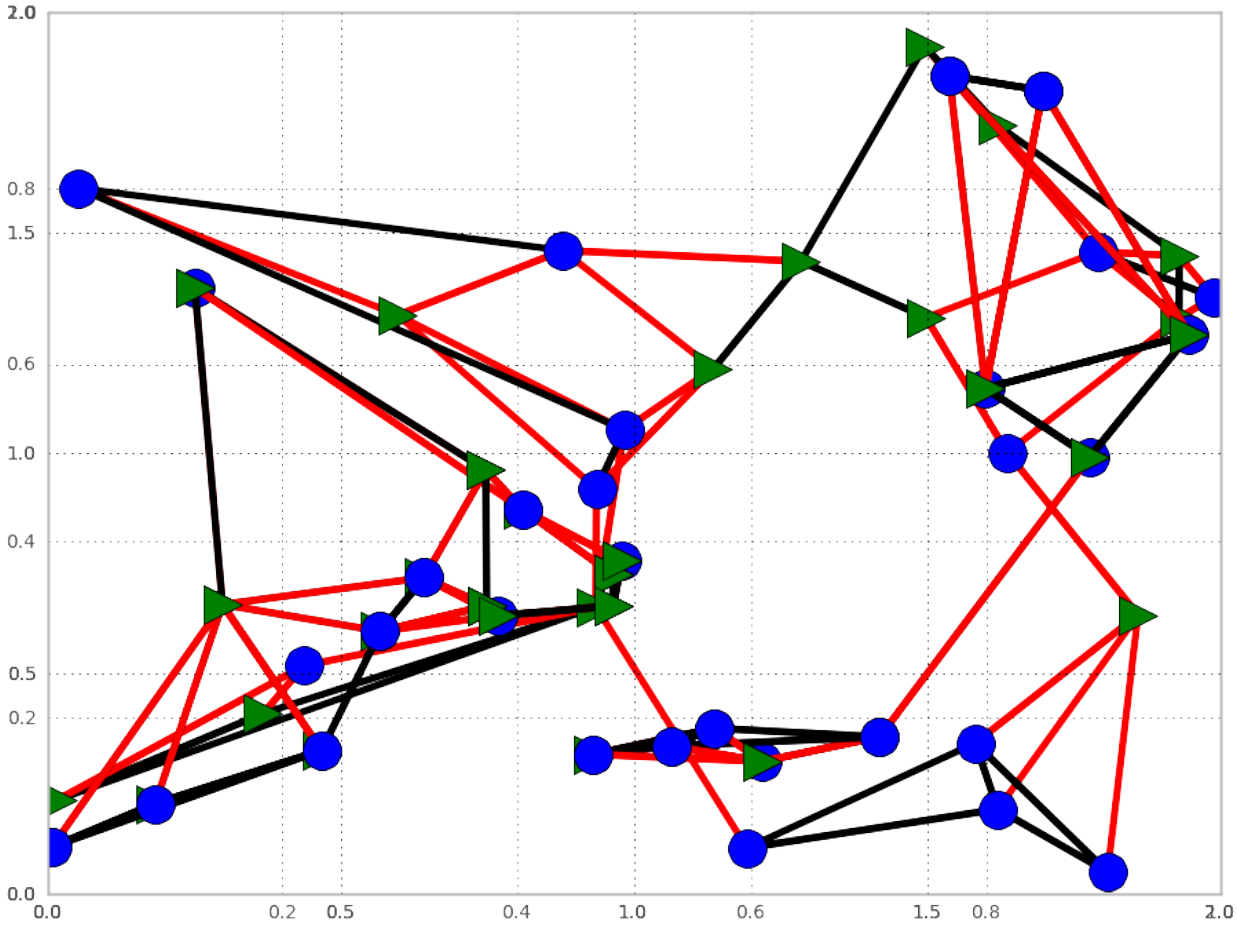
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

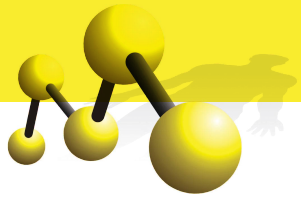
K-Nearest Neighbour Test



(Henze, 1988)

(T. Zhao et al., 2006)

Where we are going

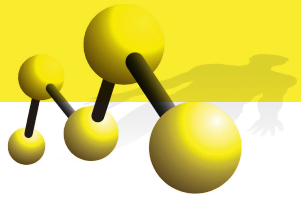


Data Integration

- Multiple image types
- Non-image data

(This was my PhD dissertation, but it is still unpublished)

Where we are going



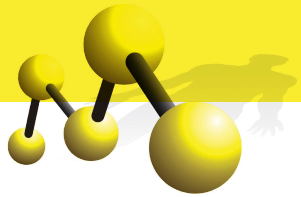
Active Learning

- Let the computer choose the experiment.
- Cut the human out of the loop.

(King et al., 2009)

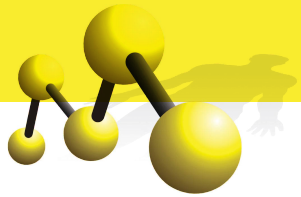
(Murphy, 2011)

Conclusions & Guidelines



- Automated methods can give better answers than humans
- (if the question is well defined)
- Interpretation need not be the bottleneck
  even in high-throughput settings
- Not so many user friendly tools available
- Collaboration can get you an expert
- Start your collaboration before you collect data

Acknowledgments



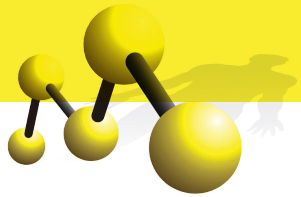
Prof. Robert F. Murphy
Dr. Tao Peng
Aabid Shariff
Dr. Estelle Glory-Afshar
Dr. Elvira Garcia-Osuna
Armaghan Naik
Joshua Kangas
...

Prof. Gustavo Rohde
Cheng Chen

Funding Agencies
Fulbright Program
National Institutes of Health
Fundação Para Ciência e Tecnologia
Siebel Scholars Foundation

thank you...

Slides



These slides (and complete references to all papers mentioned) are available at

http://luispedro.org/talks/2011/embo