AI Applications in Bioinformatics

University of Naples “Federico II”
Corso di Intelligenza Artificiale
A.A. 2009/2010

Bioinformatics

Bioinformatics is a scientific discipline dedicated to the resolution of biological problems at a molecular level with informatics methods

Schedule

Image Analysis
Location Proteomics
Gene Networks

Biologic Background

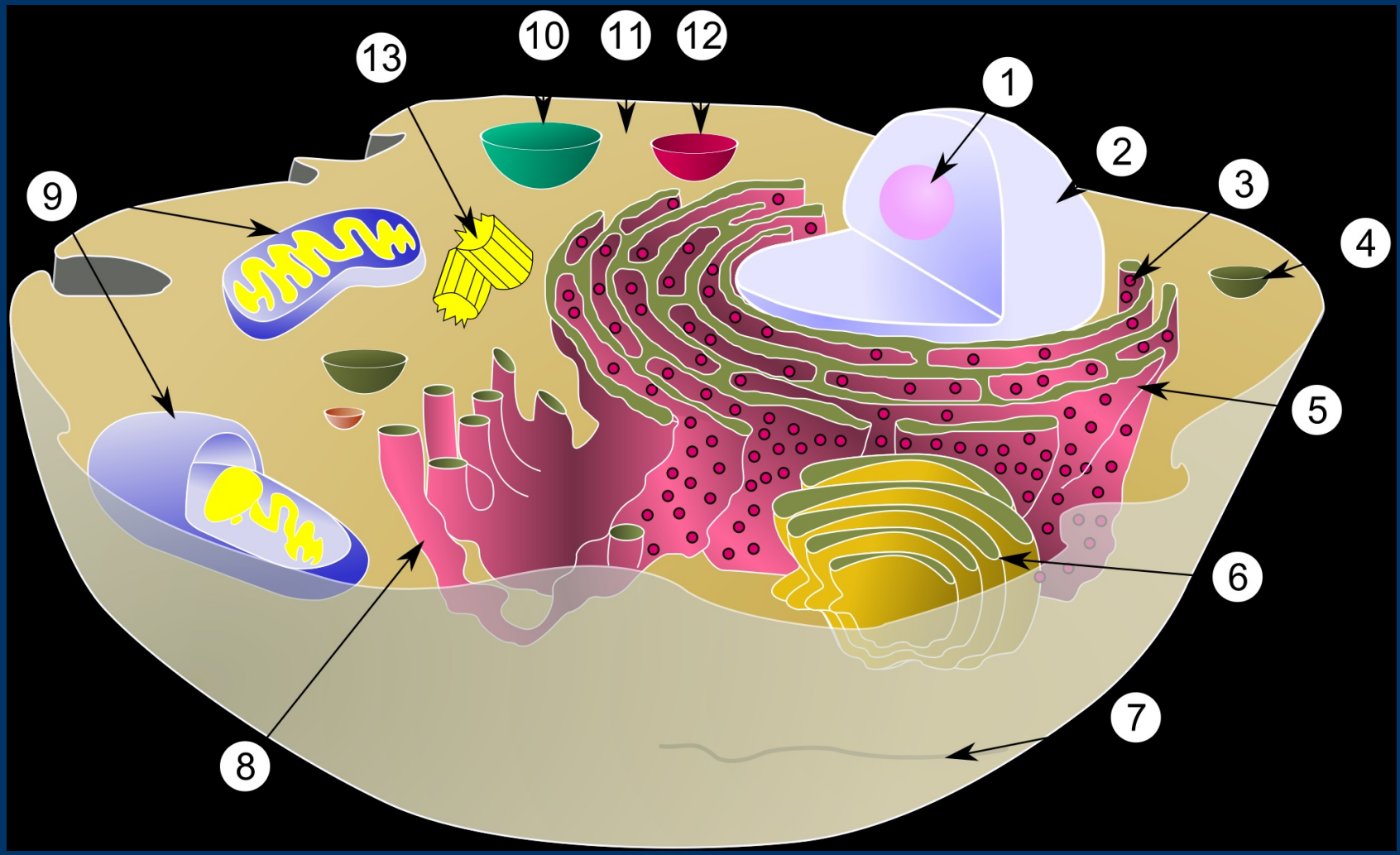
Cells
Organelles
DNA
Genes
Proteins

Cell

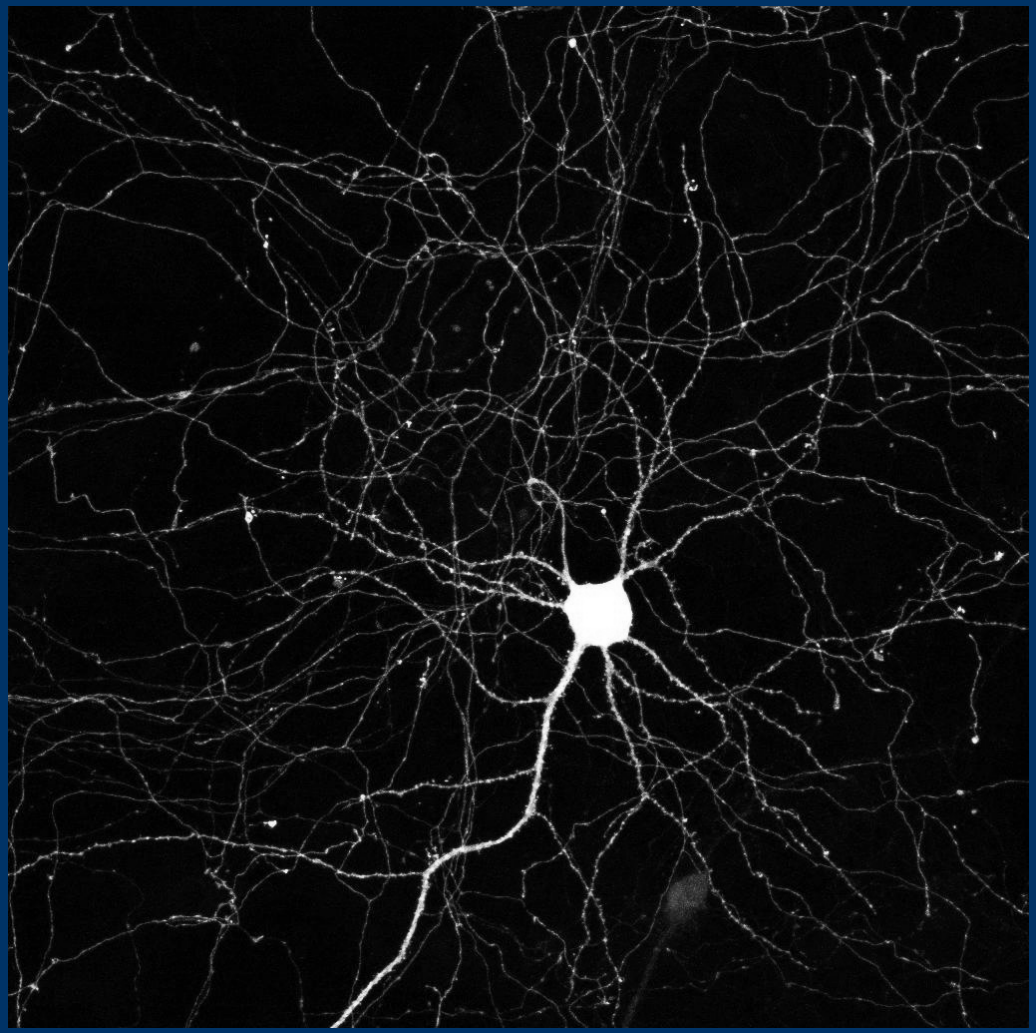
"The functional basic unit of life"

Eukaryotic cells are divided into functional compartments

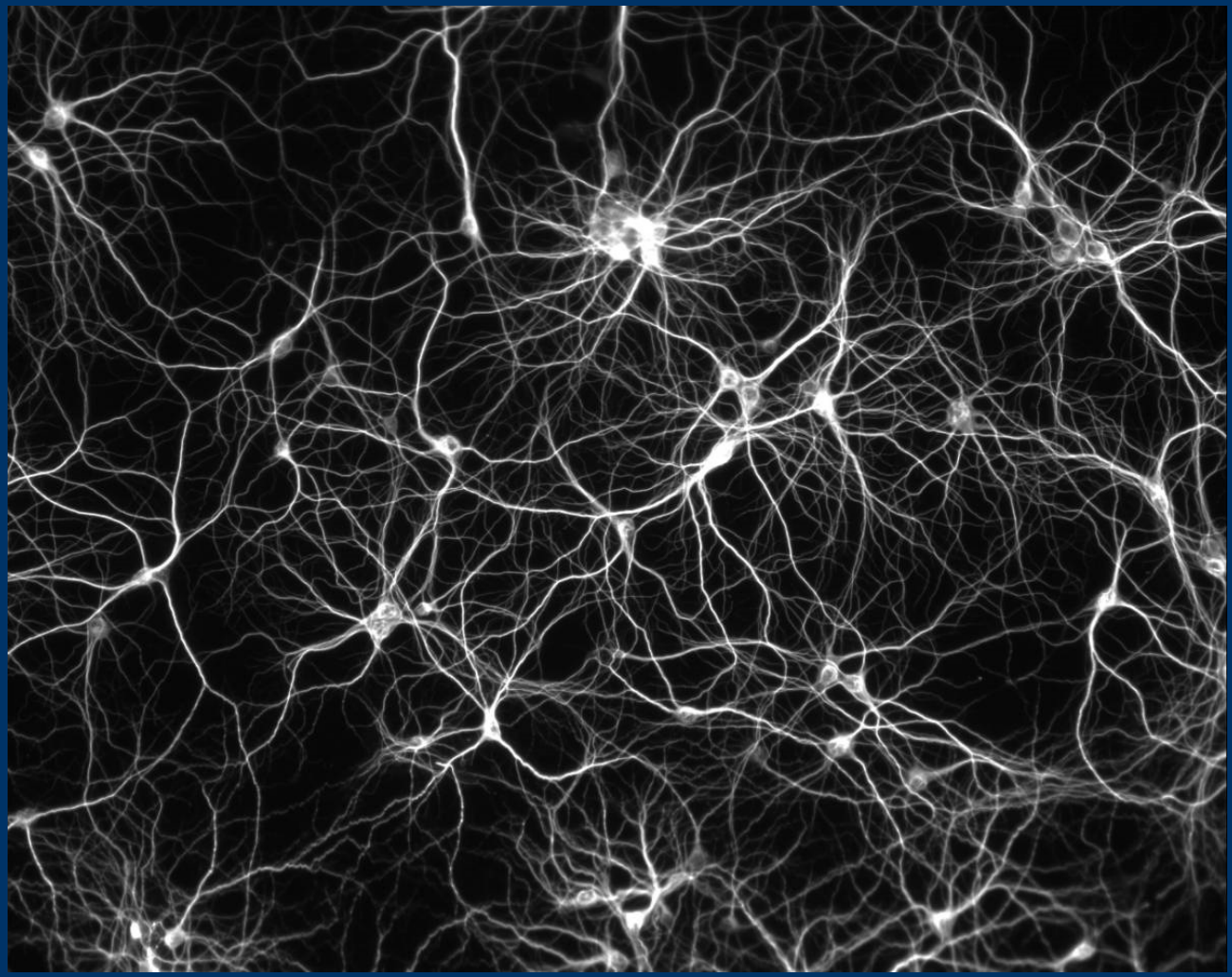
Cell Compartments Diagrams



Cell Example: A Neuron



Cell Example: A Neural Network

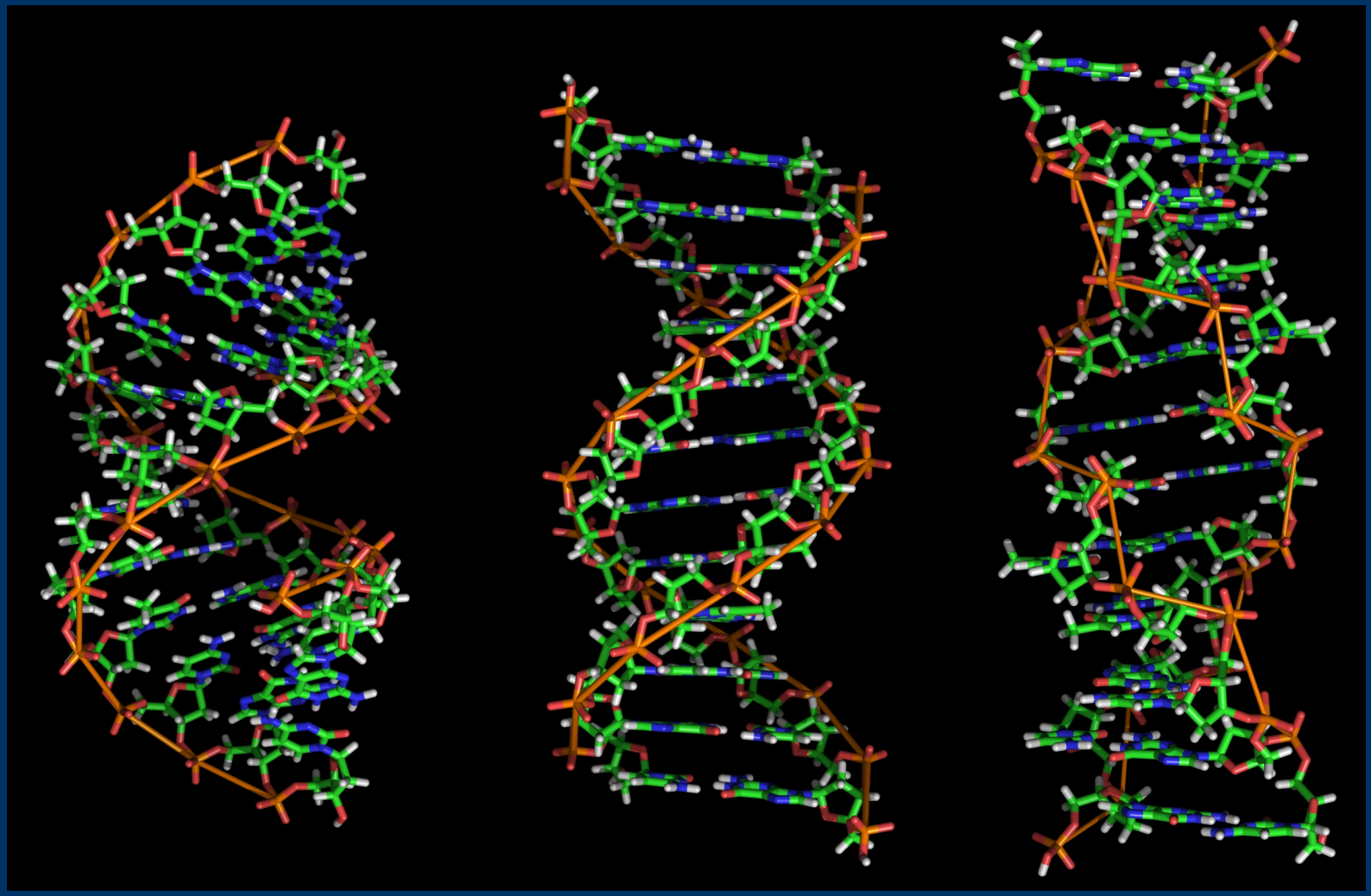


DNA

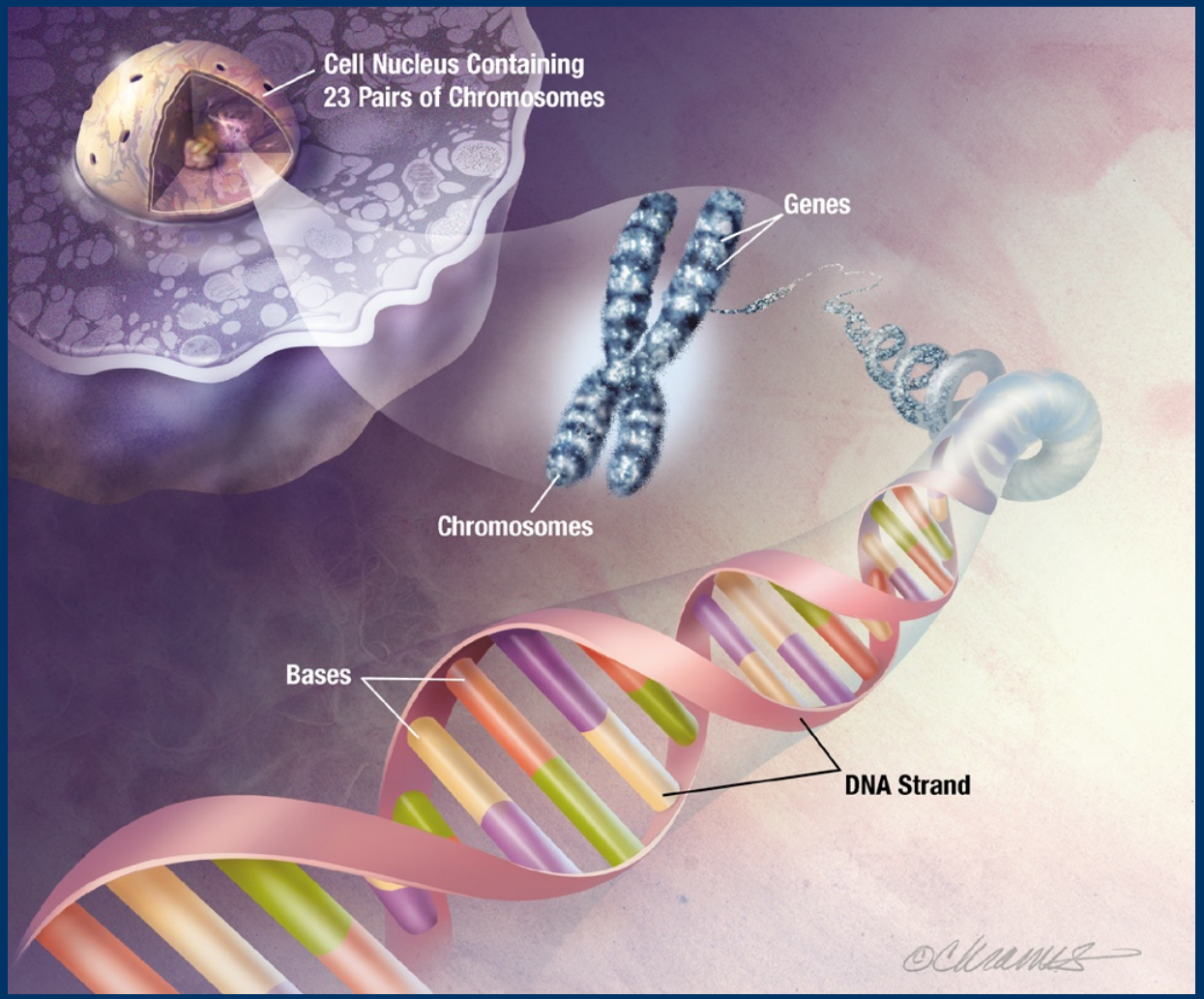
DNA is a long macromolecule that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses

The main role of DNA is the long-term storage of information in the form of a code

Double Helix Structure



DNA Subcellular Location



DNA Is A Code

DNA is a 'programming language' that 'encodes' living organisms

It has a quaternary alphabet: A T G C

Binary code: 01101010111010110010101101
Genetic code: GTCAGATCGTAAGATCA

DNA Memory Capability

DNA can be considered as an extremely miniaturized and powerful memory storage device

Each letter (*nucleotide*) has a value of 2 bits
Example: A→00 T→01 G→10 C→11

3×10^{9} nucleotides in human genome

6 Gigabits stored in 10^{-6} meters

DNA Memory Capability

Examples:

750 GB in 1 mm length and 10^{-3} mm height

or

75 PB (i.e. 75,000 TB) in a device 1×1 cm

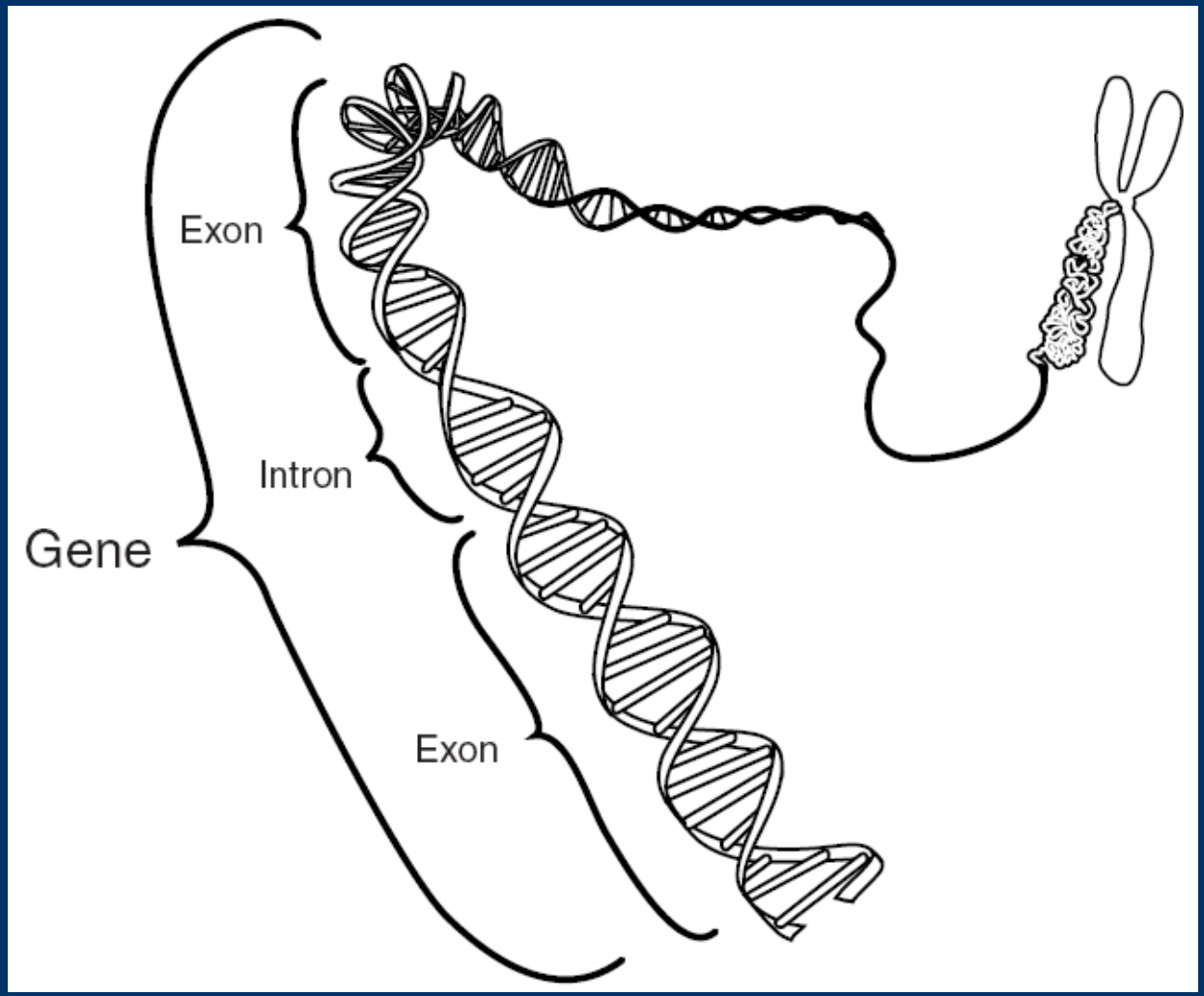
Genes

"A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and\or other functional sequence regions"

Simply, a gene is a unit of heredity in a living organism

It is a stretch of DNA that codes for a type of protein that has a function in the organism

Genes



Proteins

Genes encodes the genetic code to create molecules that will have active role in the organism, called proteins

Structural roles

Functional roles

Application

Image Analysis

Image Analysis

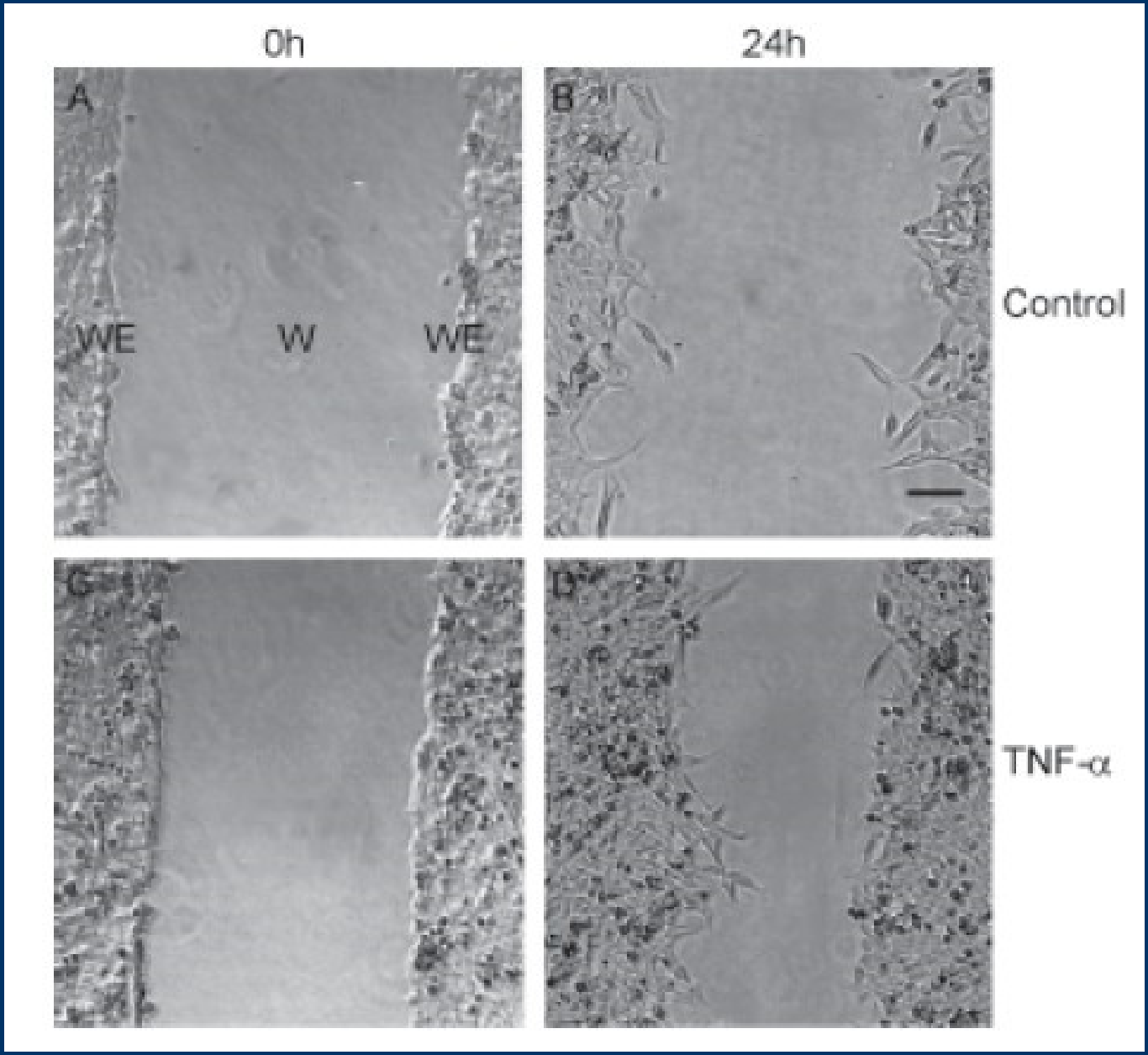
Artificial Intelligence can be used to analyze images and extract informations about objects in them and their behaviors

Cell Tracking

Wound healing
Parasite spreading
Immune cells migration
Stem cells migration
Embryo development
Tissue engineering

Normal movements vs. movements on drug treatment

Wound Healing



High-Throughput

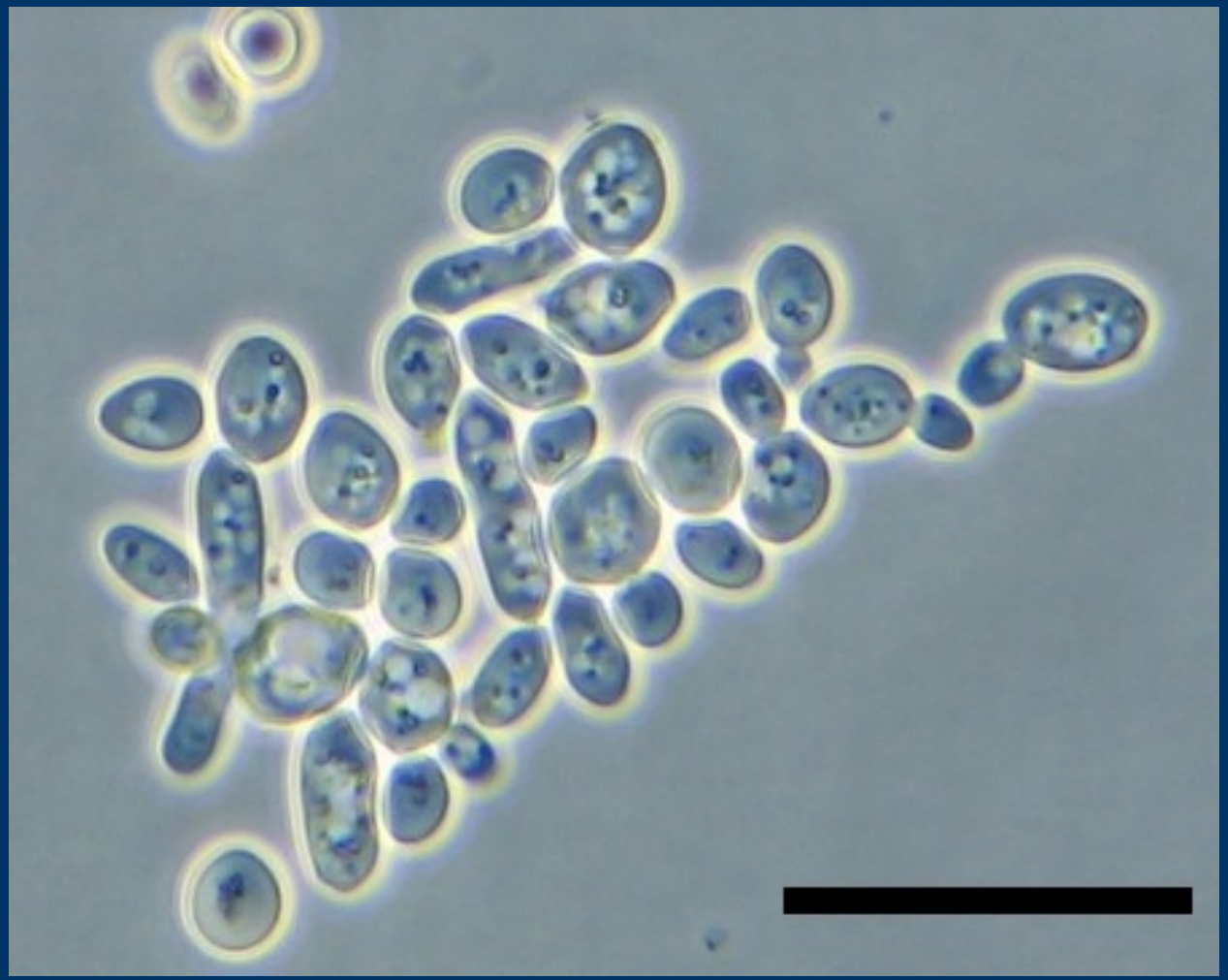
High-throughput

Example: 400,000 to 1,000,000 images with hundreds of cells in each for a single study

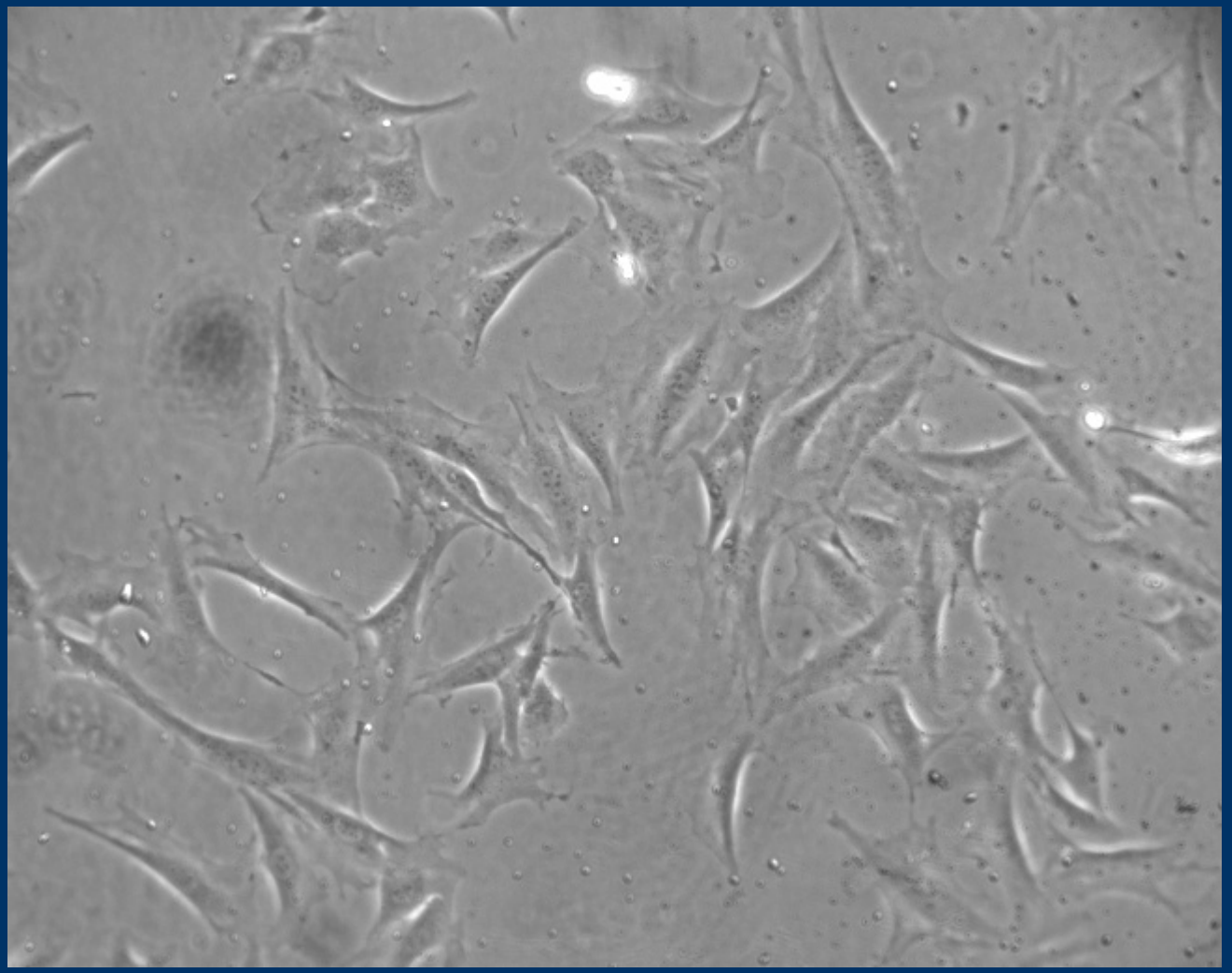
Need of automatic systems!

Spores of Diatrype Disciformis

Saccharomices Cerevisiae *(Yeast)*



Mouse Embryonal Fibroblasts



Case: Identification and Tracking

Objective: identification and tracking of time-lapse phase-contrast microscopy video of cells

Identification: Approaches

Manual identification

Pixel intensity

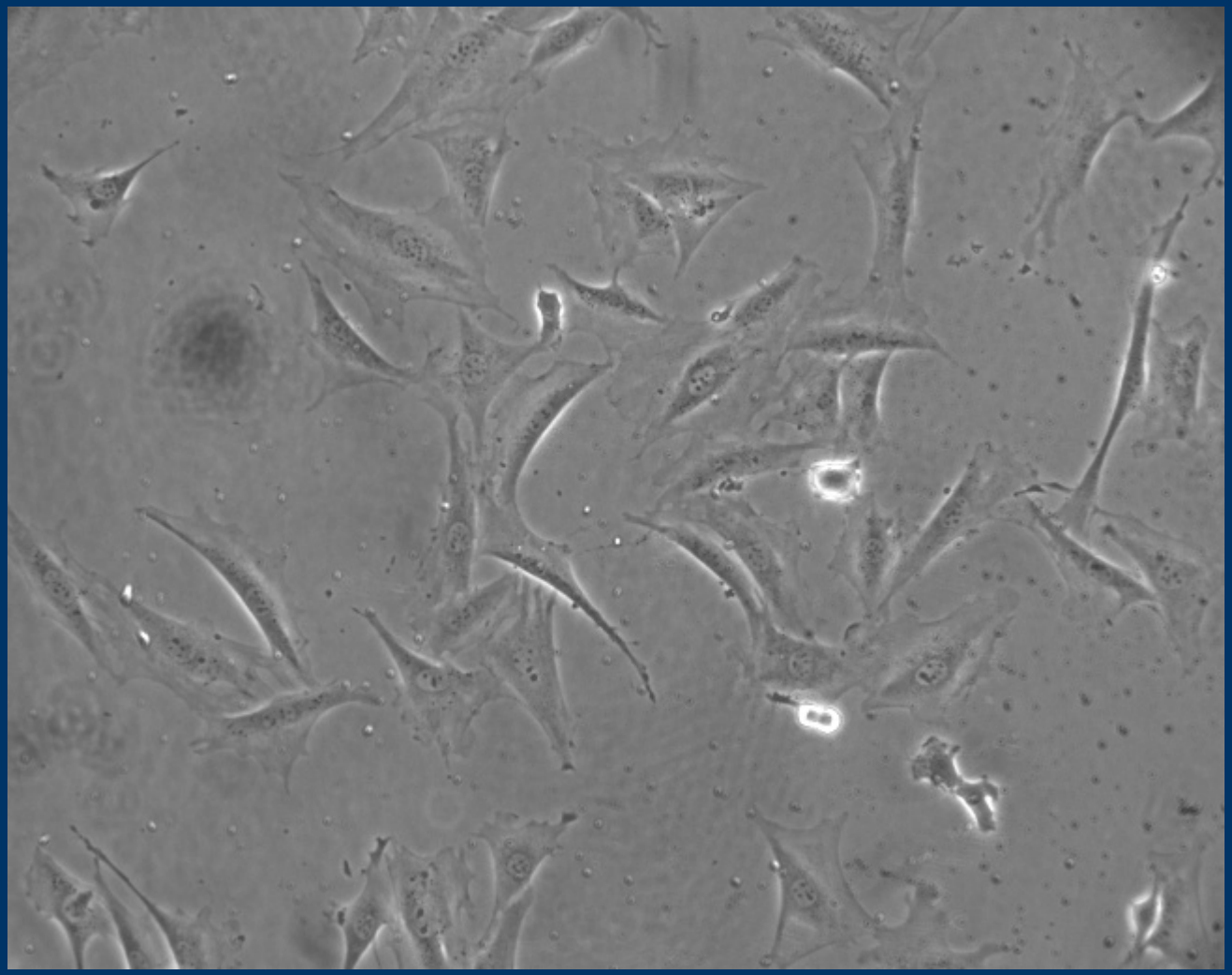
Image gradient

Active contours [Zimmer *et al.*, 2002]

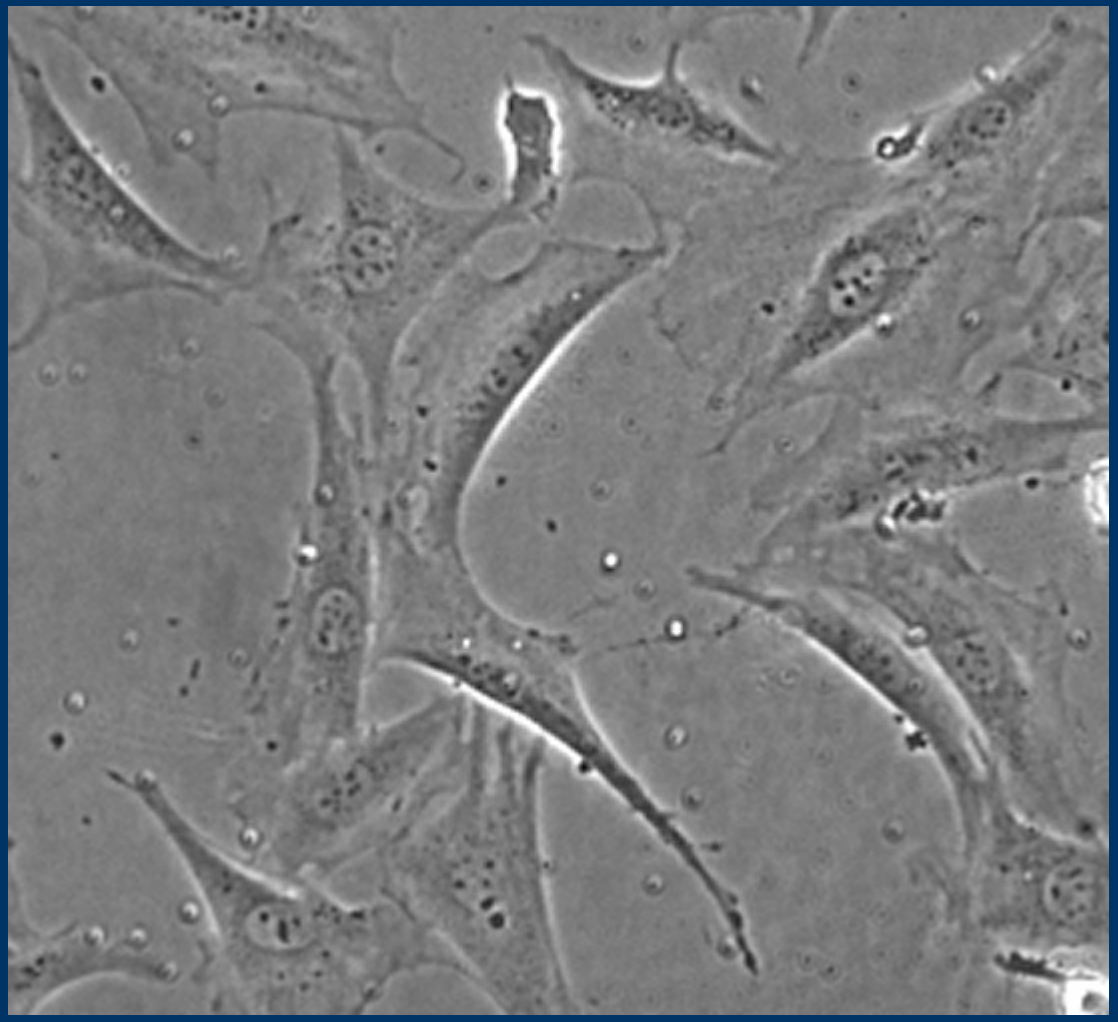
Watershed [Wählby *et al.*, 2002, Zhou *et al.*, 2005]

Energy functional minimization [Yan *et al.*, 2008]

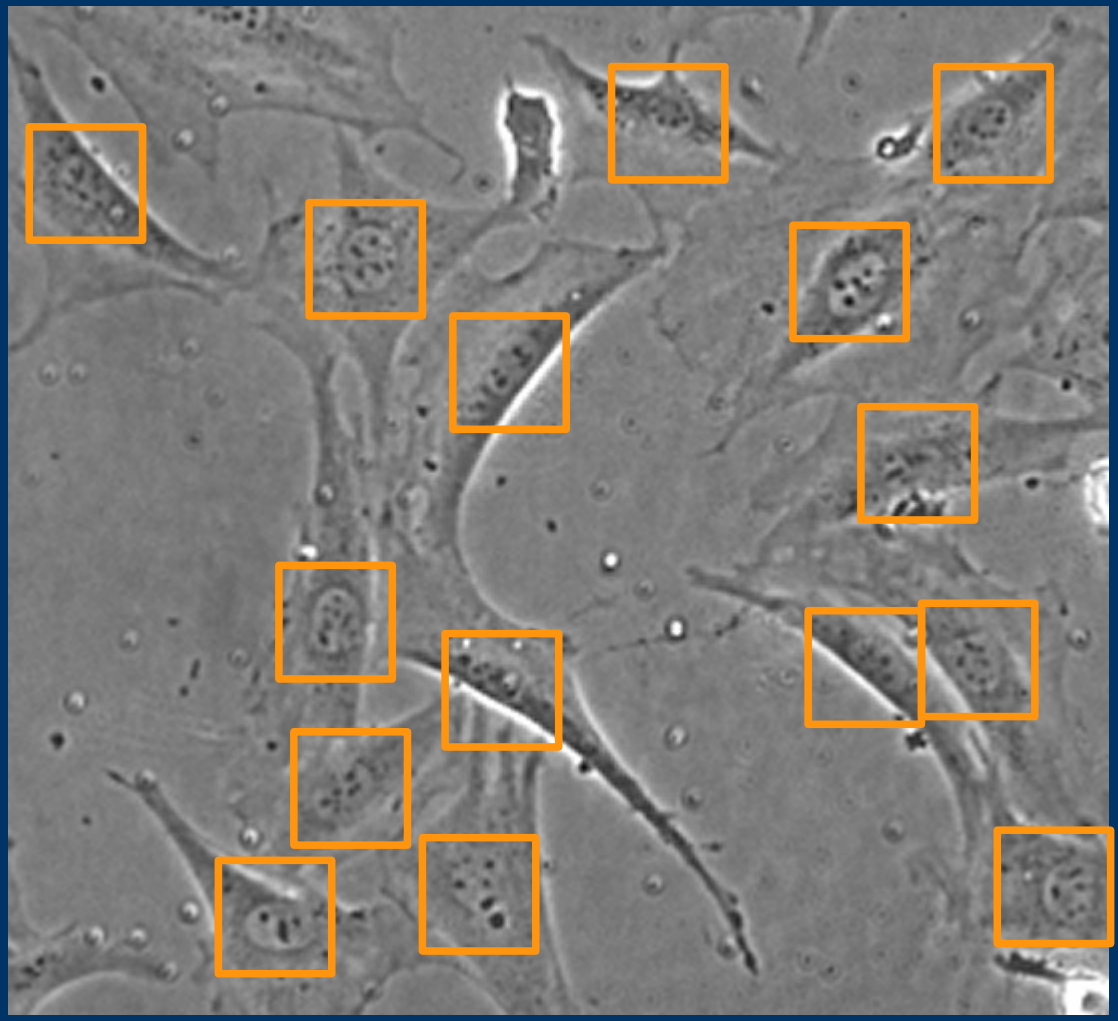
Cell Identification: MEF Example



What To Use?



Nuclei Are Always Recognizable



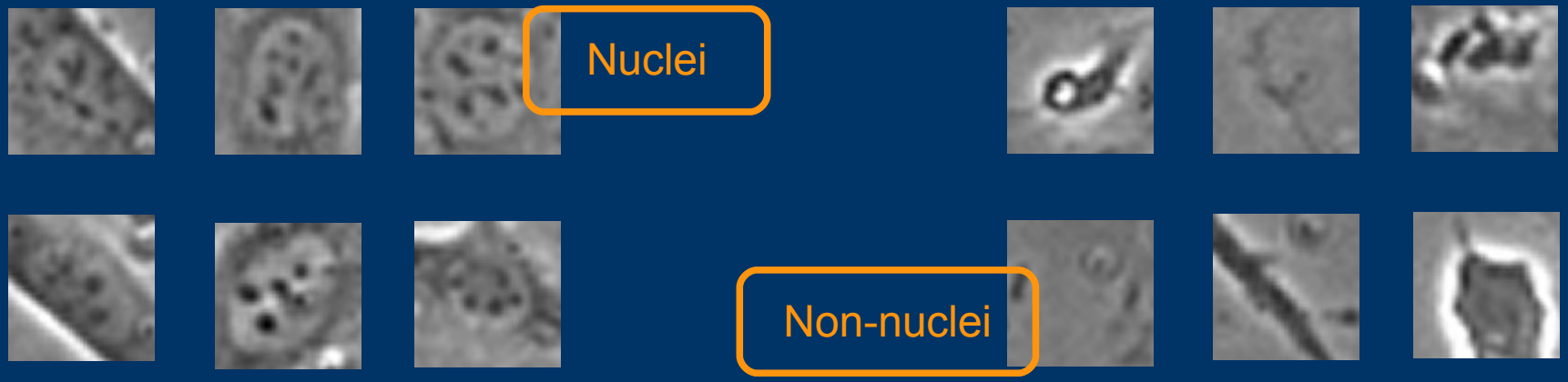
Question

How about if we train an SVM to make it able to discriminate between nuclei and non-nuclei structures?

Training Samples

A set of 100 nuclei portions were extracted
200 non-nuclei portions were extracted too, including very nuclei-like areas

20×20 pixels (a big nucleus size)



Training Samples

All the images were rotated 90, 180 and 270 degrees to have more samples

400 nuclei and 800 non-nuclei

1200 samples to train the SVM

Features

How can a numeric SVM work with images?

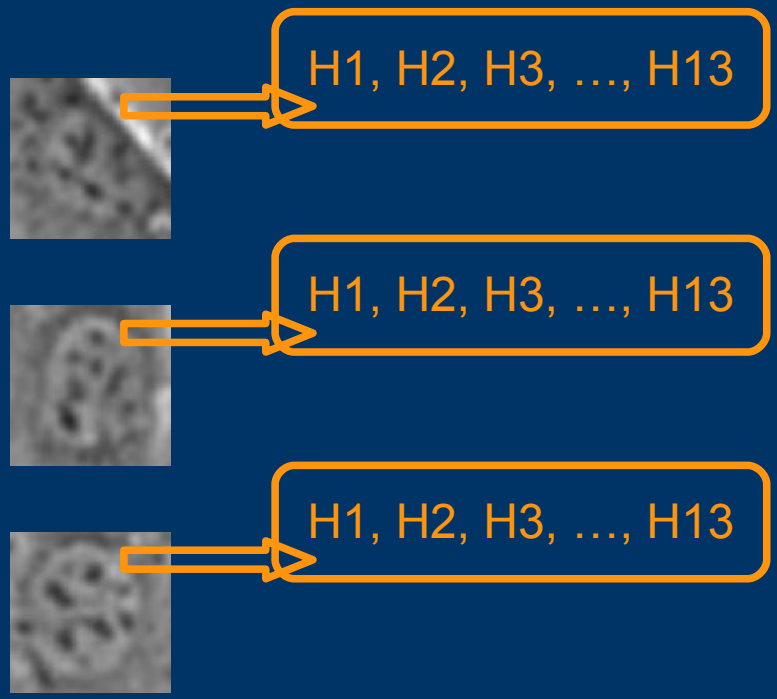
Need for numeric descriptors

Haralick Textural Features

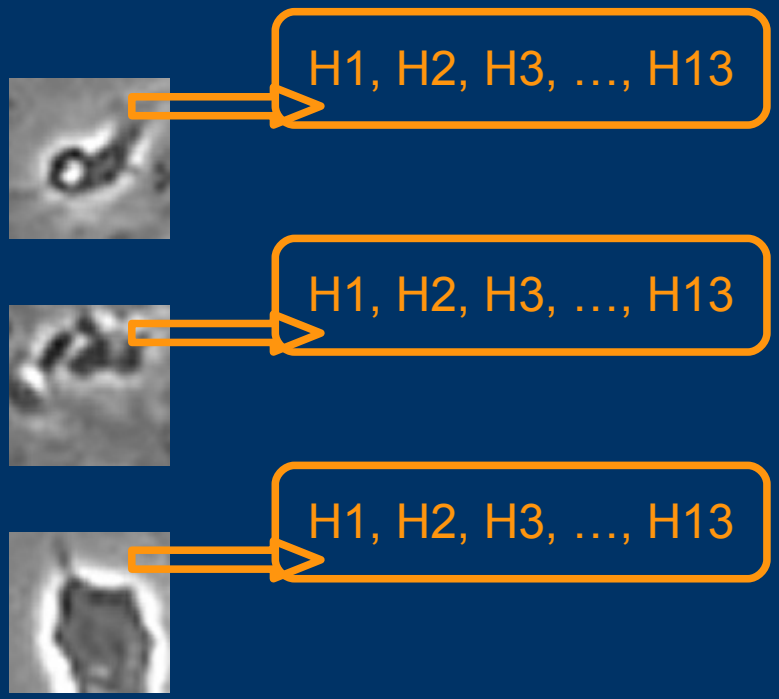
Very important features for image classification are the Haralick textural features

Those are 13 features that can describe numerically textural characteristics thus making able the SVM to discriminate them

Feature Extraction



Positive Examples



Negative Examples

SVM

The features were passed to an SVM with Polynomial kernel

The best parameters were found with KNIME

Then they were used with the LibSVM implementation for MATLAB

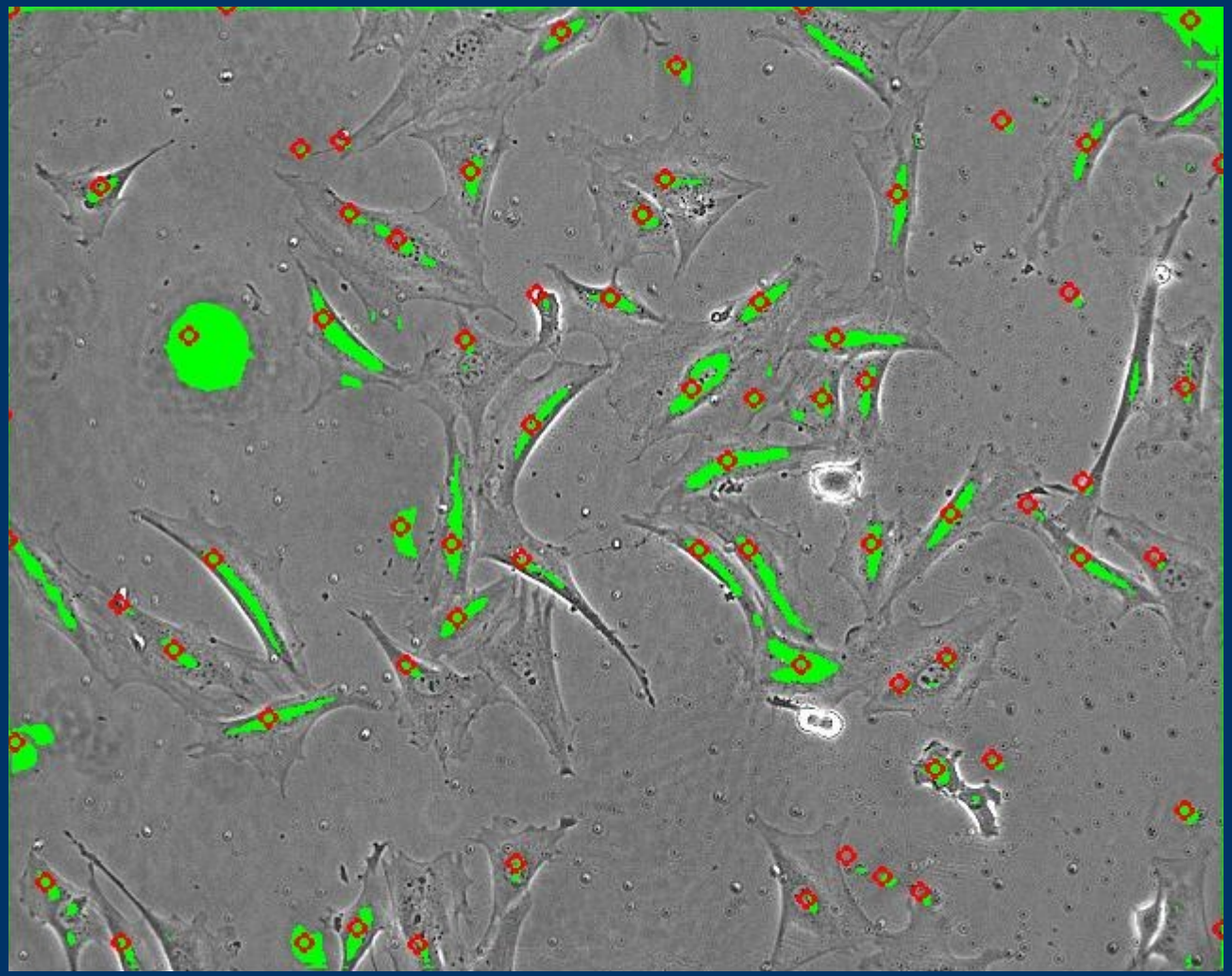
Identification: Preprocessing

First preprocessing

Zonal thresholding

The center of each zone identified a possible cell

Preprocessing Phase



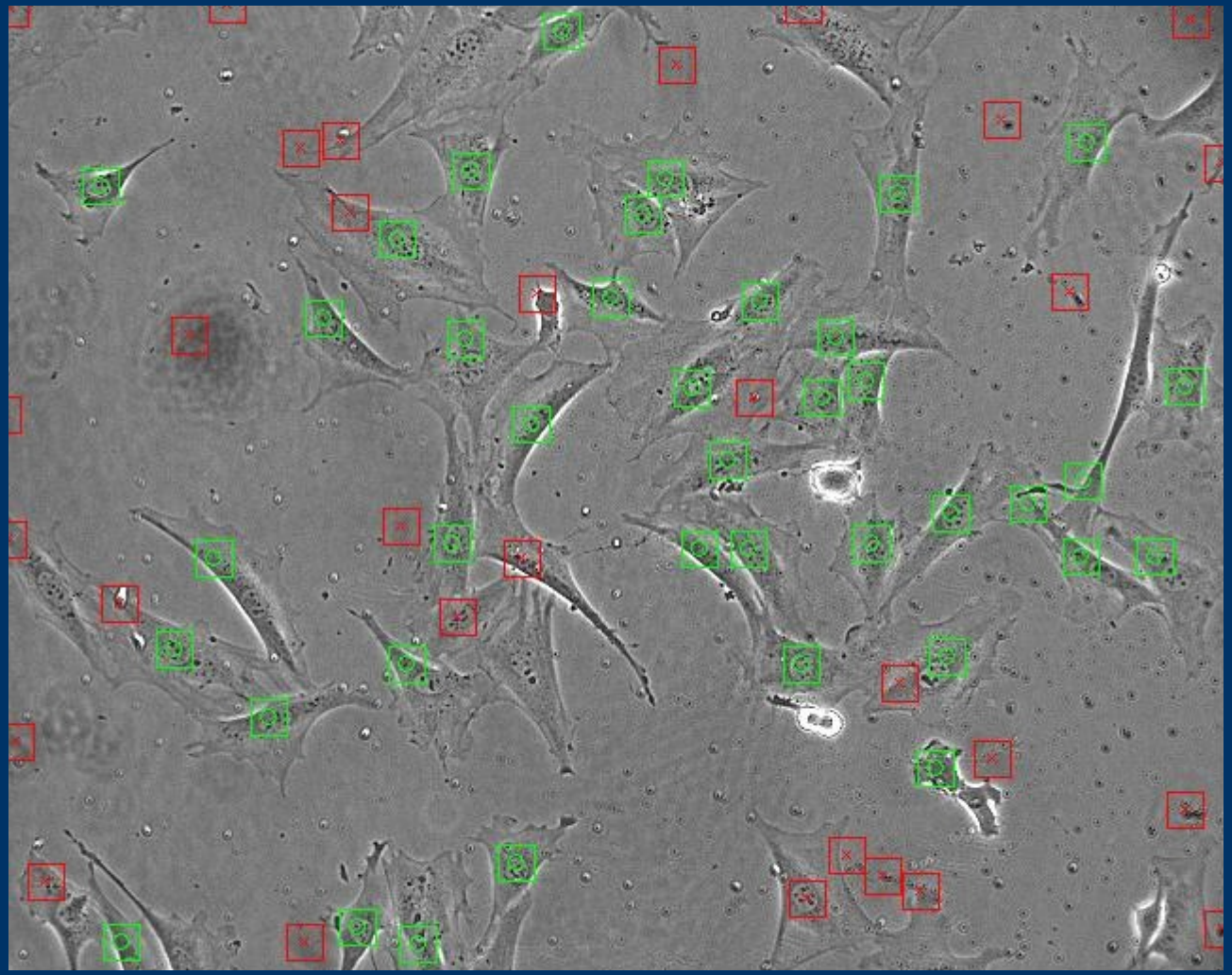
Identification: Supervised Phase

20×20 pixels area centered in each point

Features extraction

SVM classification

Supervised Phase

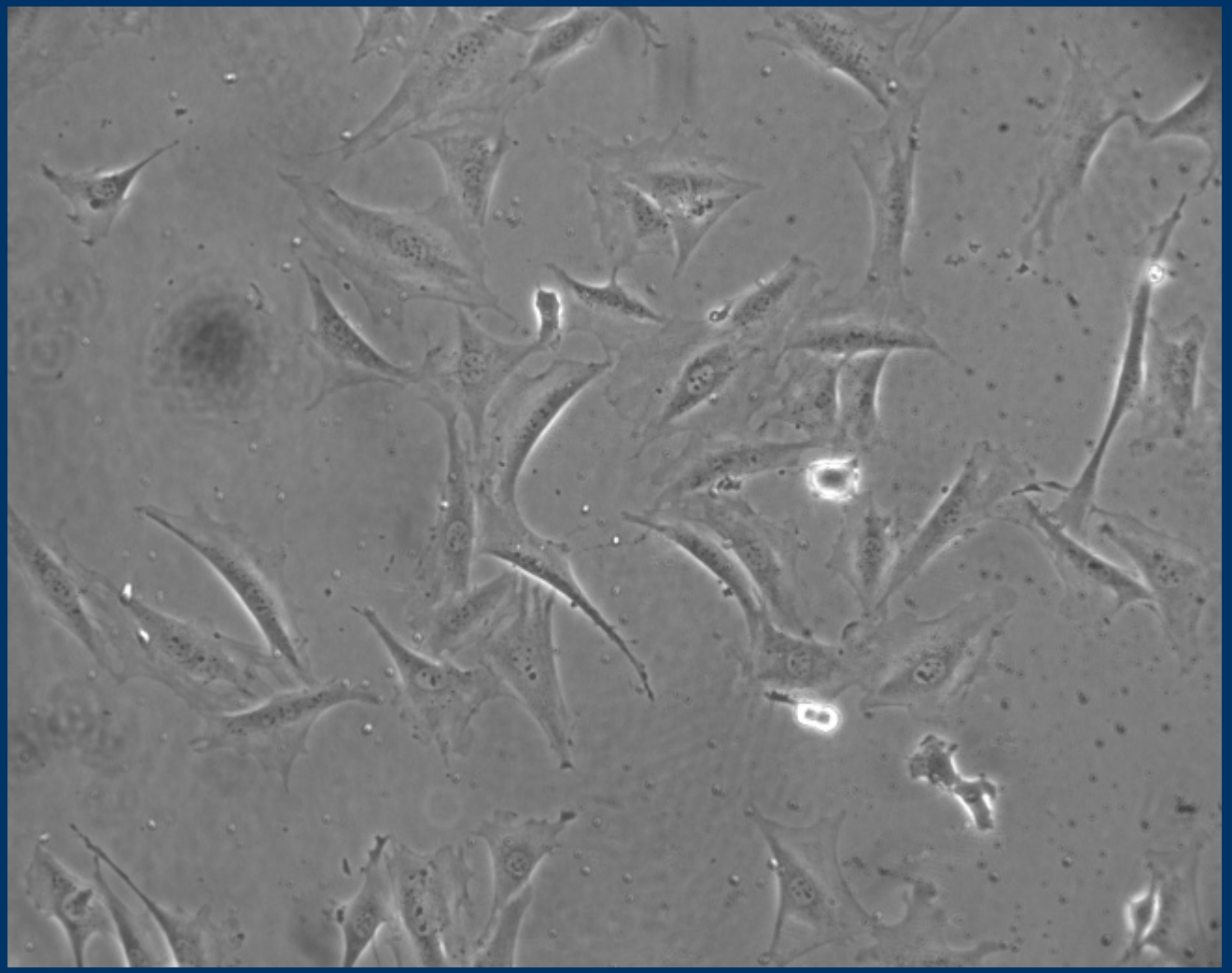


Time-Lapse Microscopy

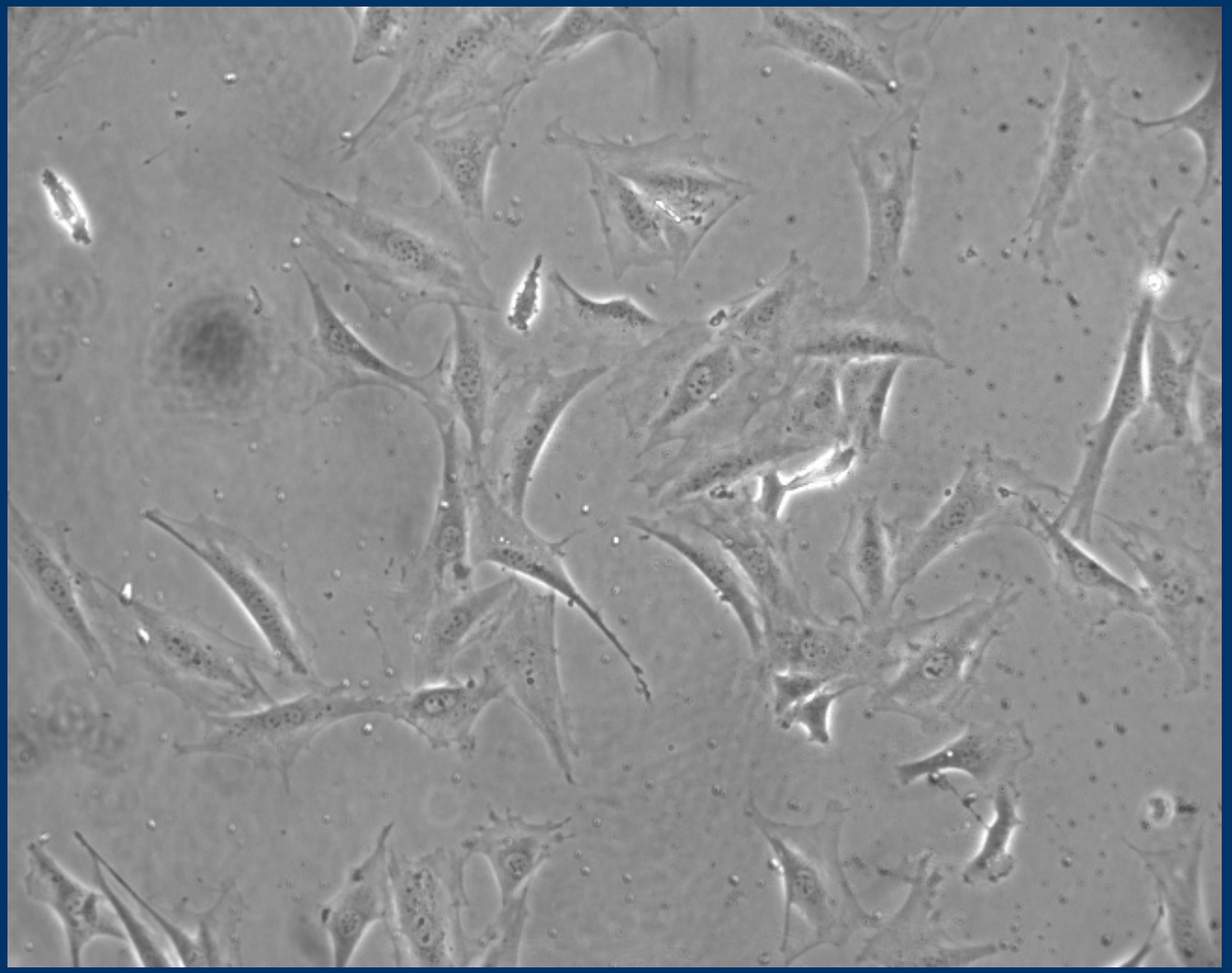
Tracking

Time-lapse phase-contrast microscopy movie

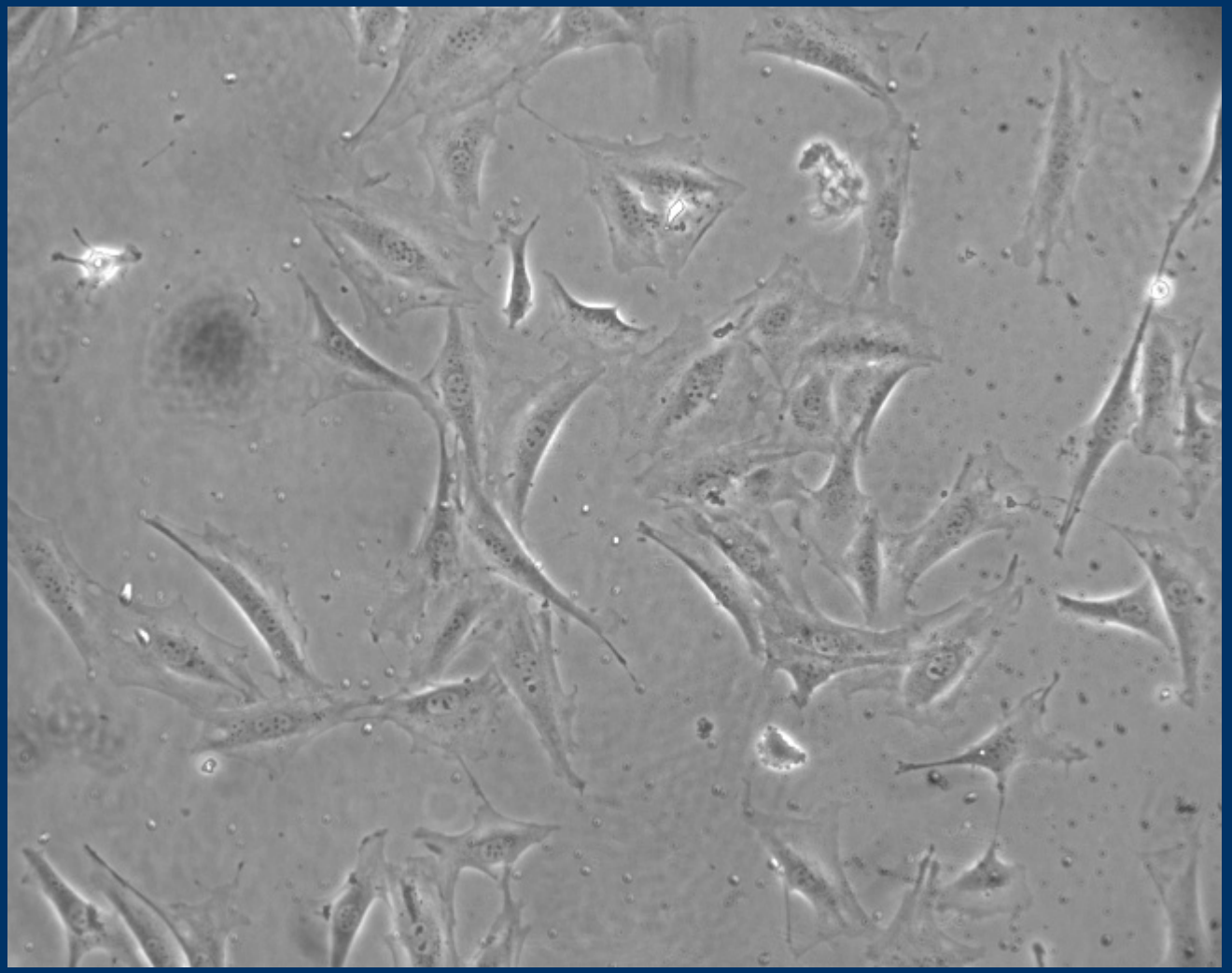
Frame 1



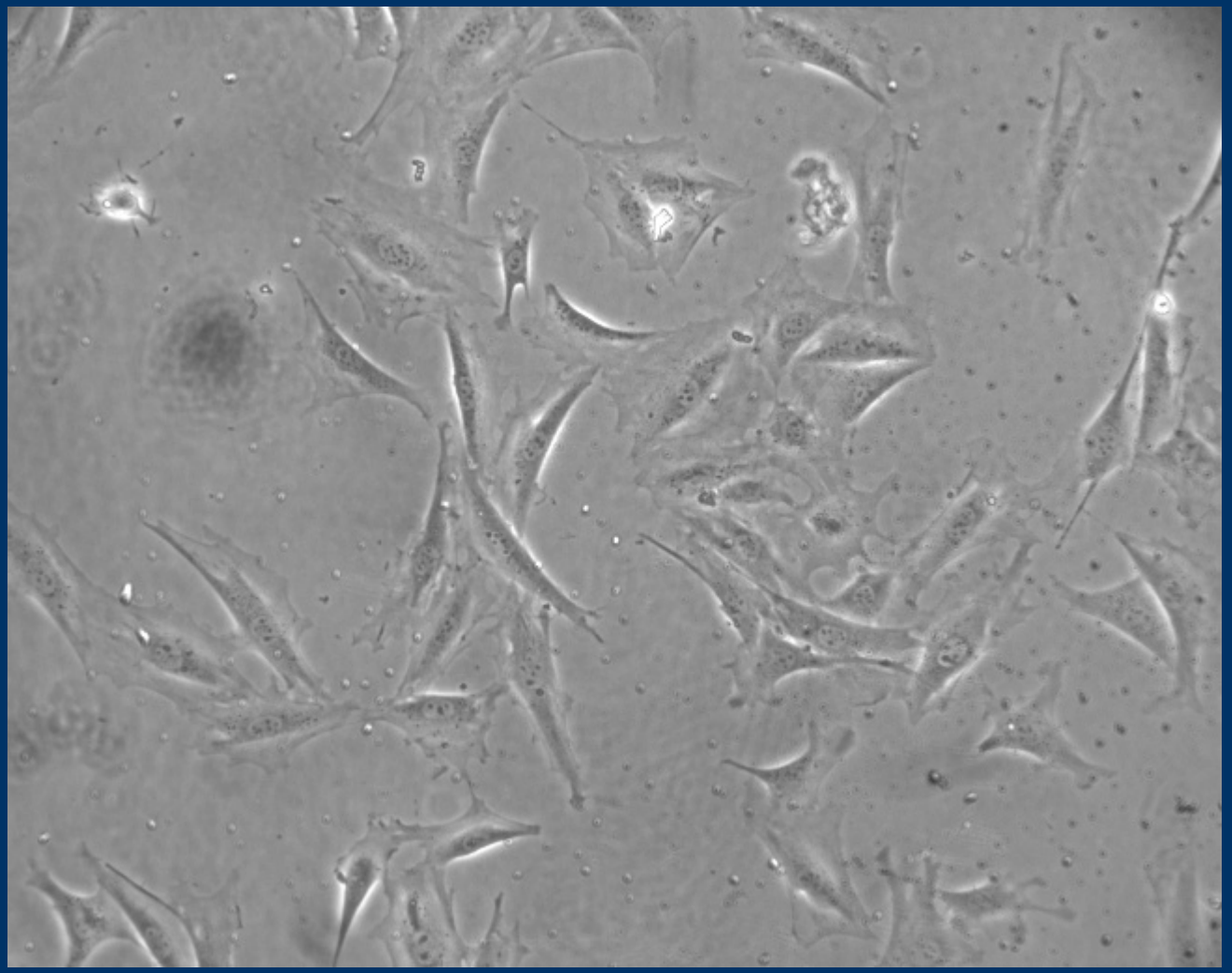
Frame 4



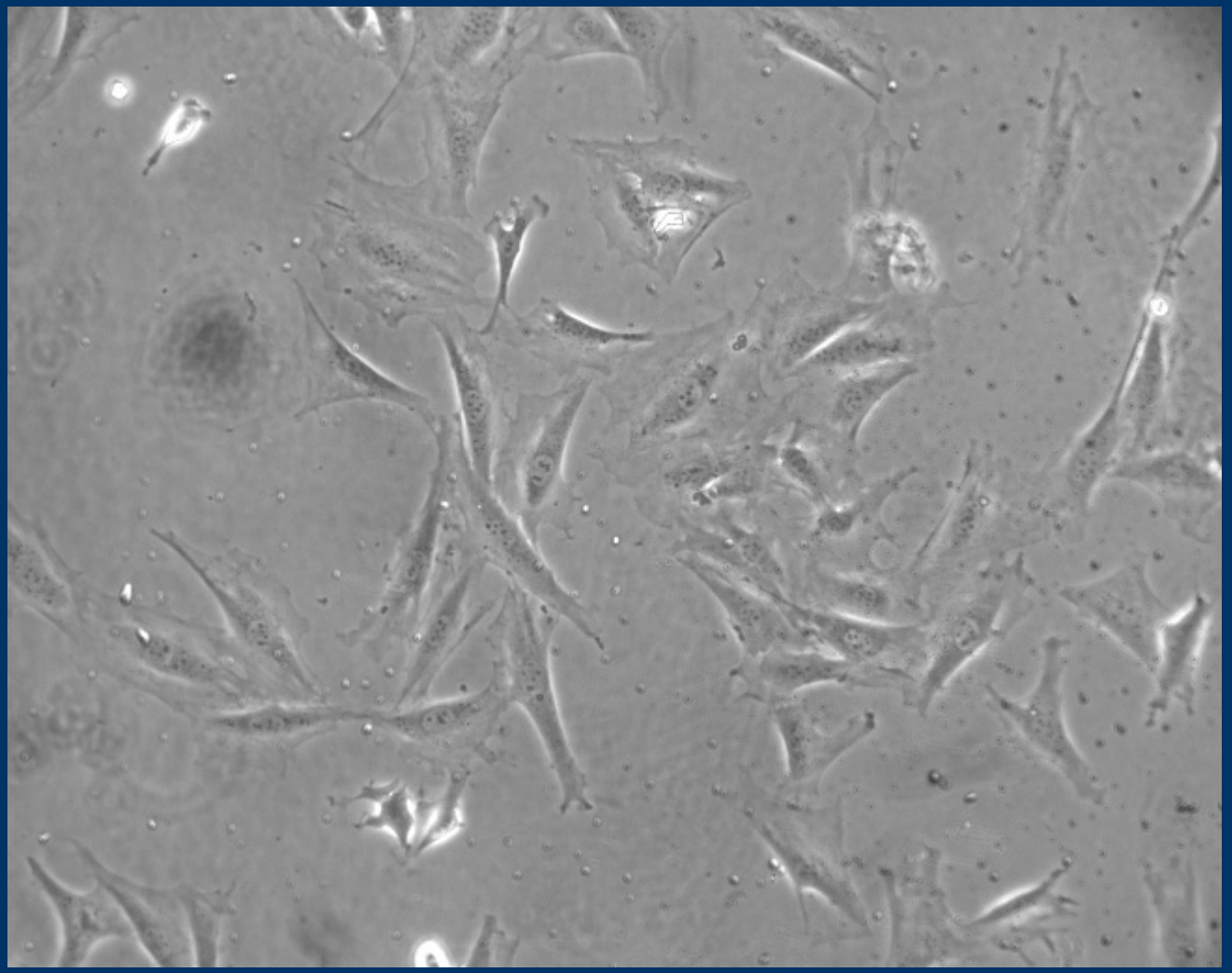
Frame 8



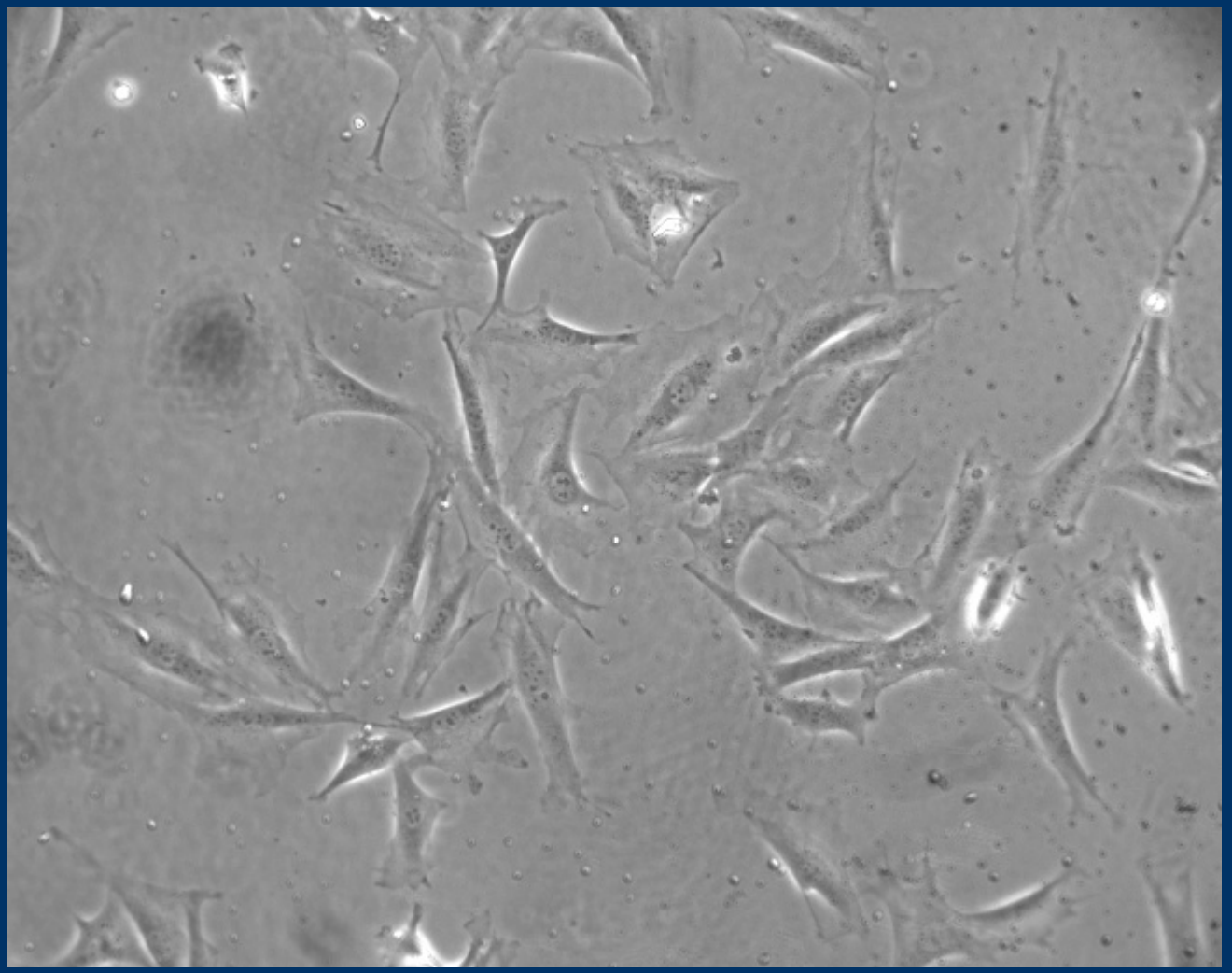
Frame 12



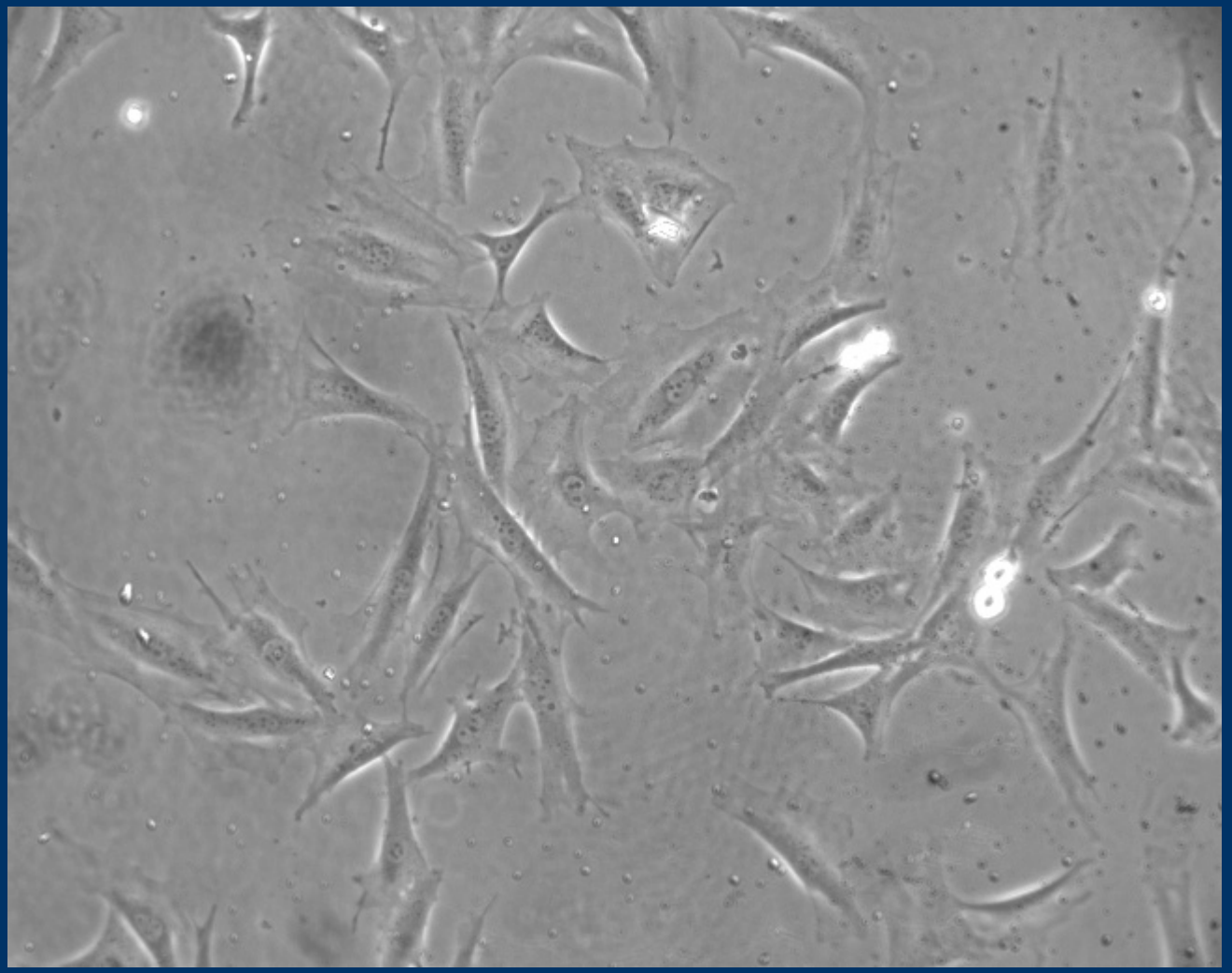
Frame 16



Frame 20



Frame 24



Tracking: Approaches

Manual tracking [Cantarella *et al.*, 2009]

Multiple-models dynamics filters [Li *et al.*, 2008]

Cross-correlation [Perez-Careta *et al.*, 2008]

Seeded Watershed [Pinidiyaarachchi, 2005]

Time as extra spatial dimension [Yang *et al.*, 2005]

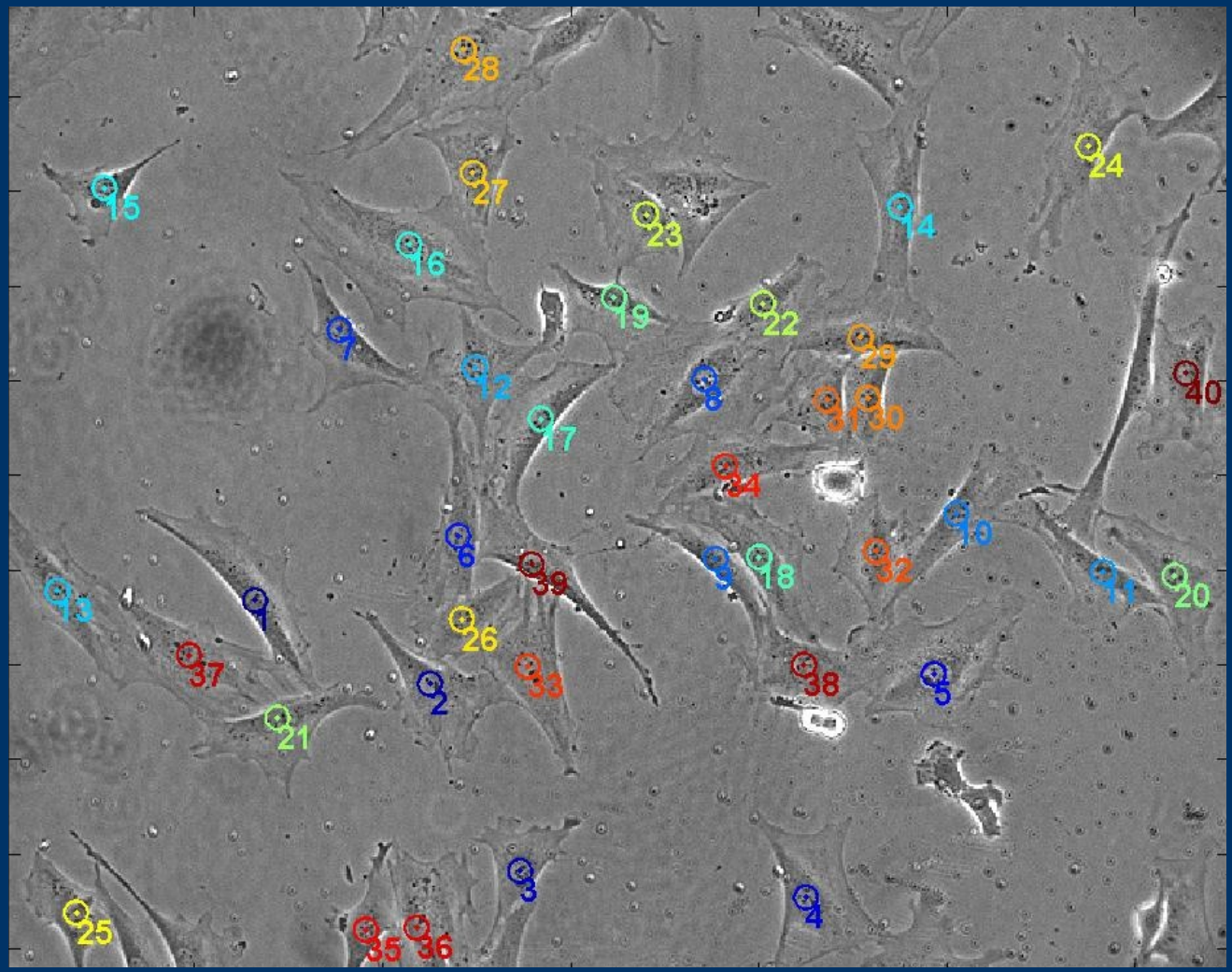
Tracking: Example Approach

Motion Estimation

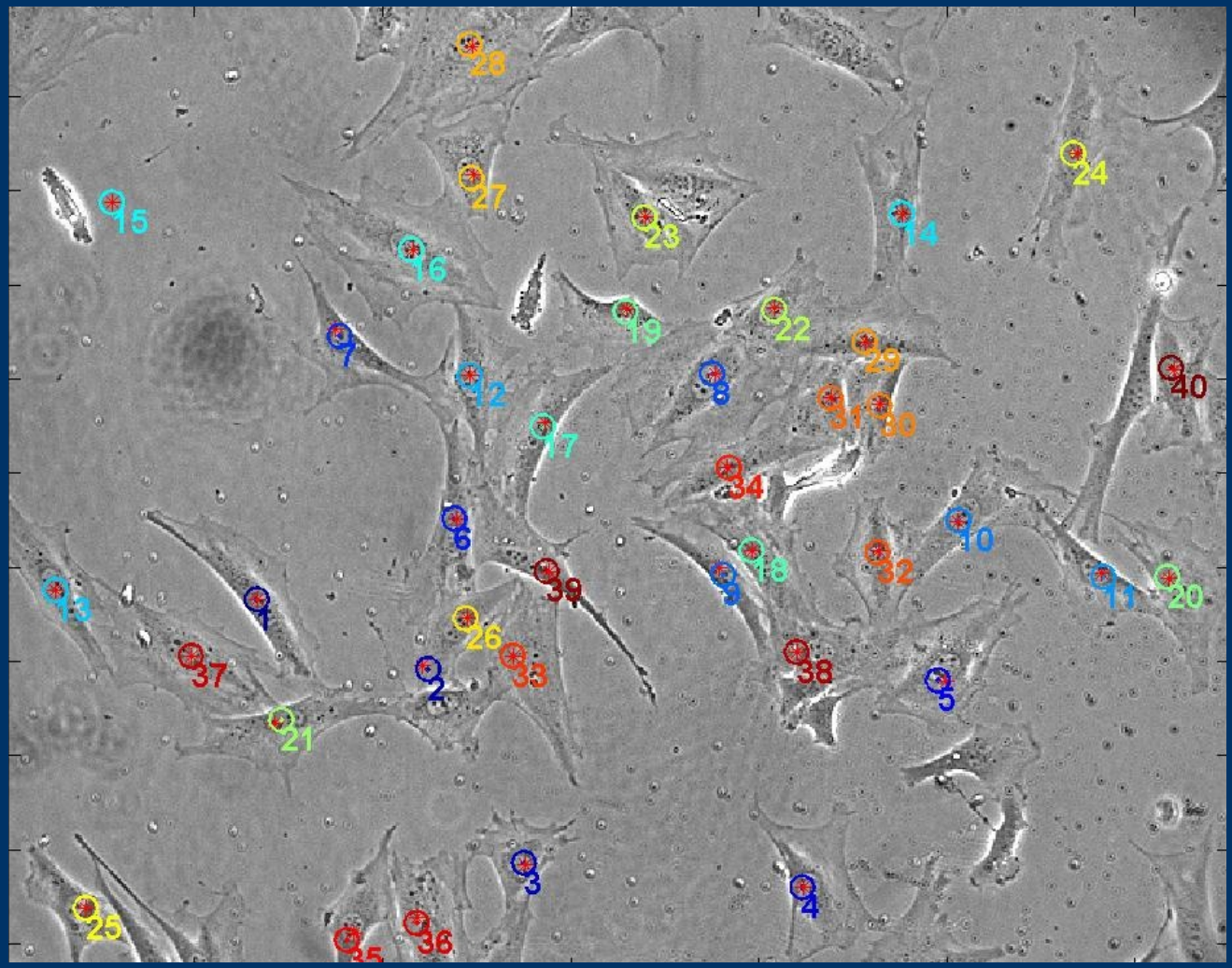
Kalman filter

SVM classification

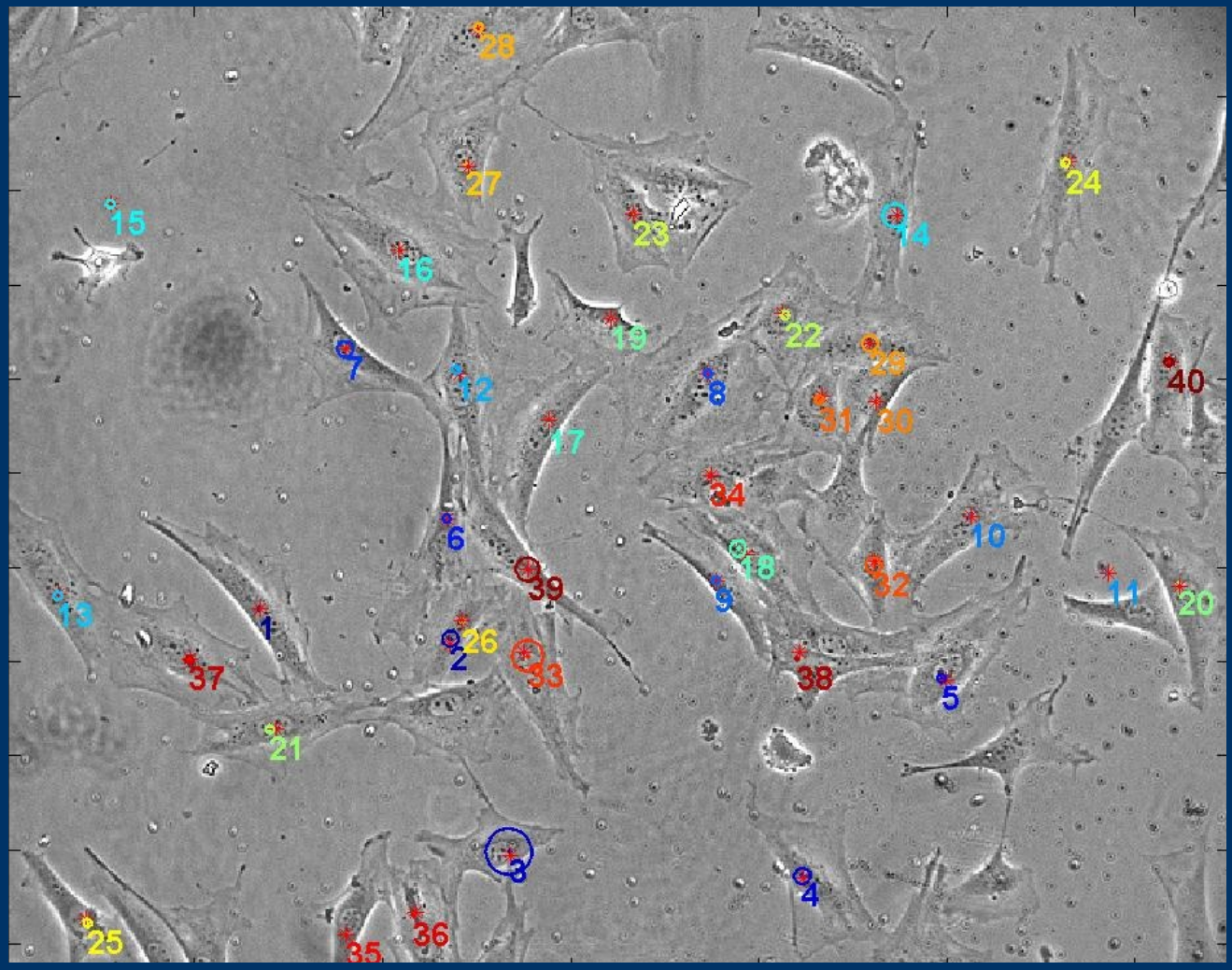
Frame 1


28
15
27
24
23
14
16
19
22
7
29
12
8
40
31
30
17
34
6
10
39
18
32
11
20
13
9
26
37
33
38
5
2
21
3
4
25
35
36

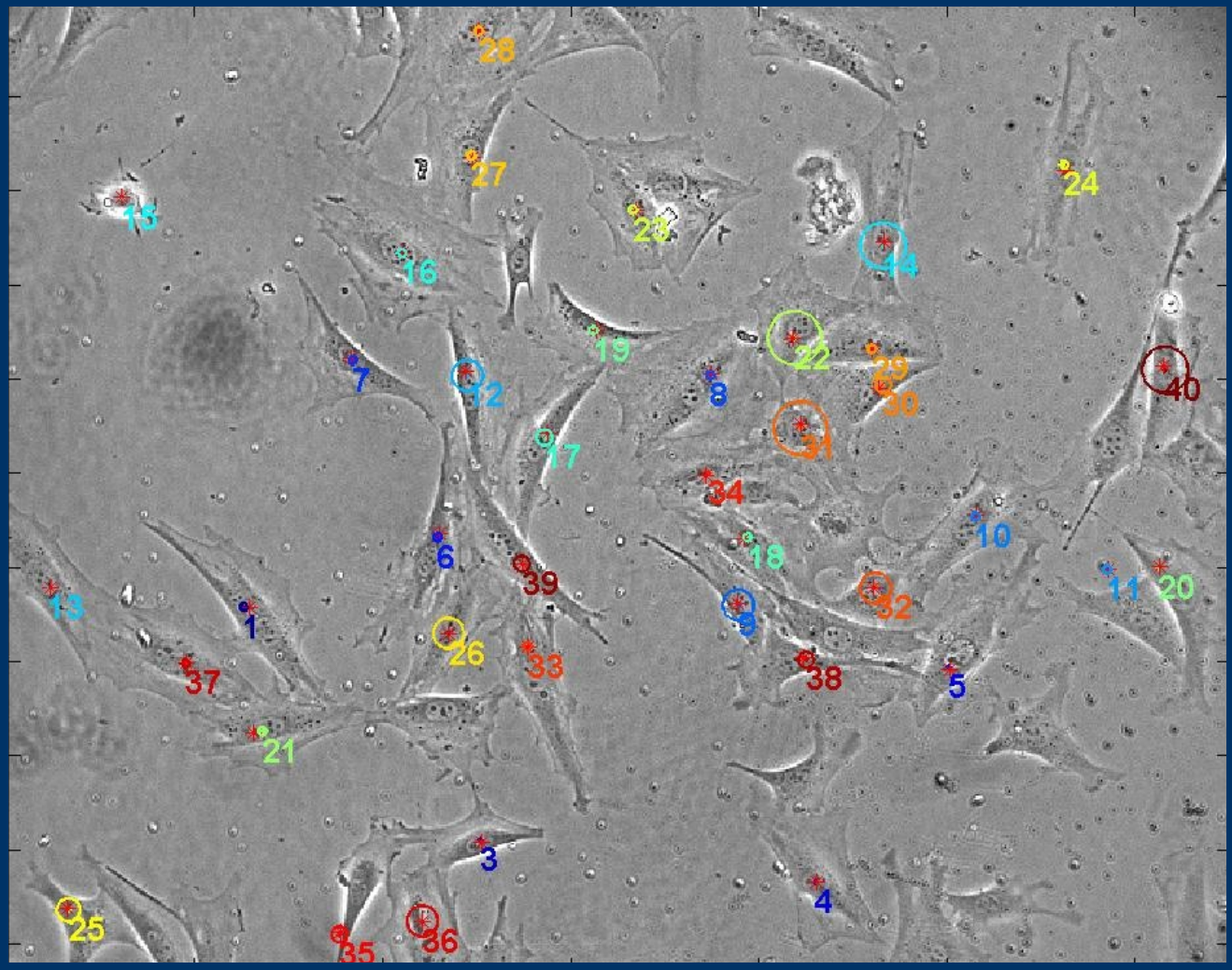
Frame 4


28
24
27
15
23
14
16
22
19
29
7
40
12
8
31
30
17
34
6
10
18
32
39
3
11
20
13
1
26
37
33
38
2
5
21
3
4
25
36
35

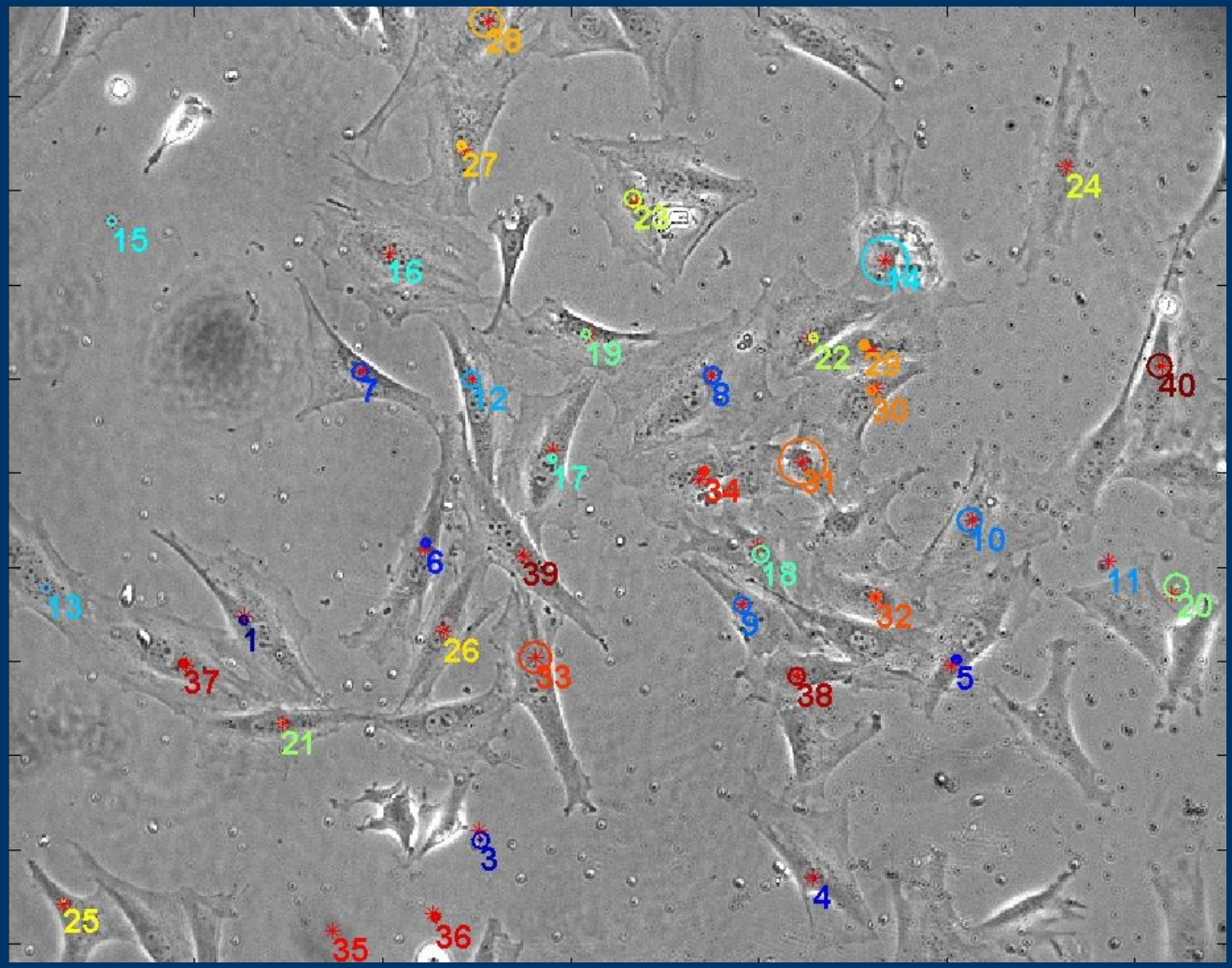
Frame 8


28
27
24
15
23
14
16
19
22
29
40
7
12
8
31
30
17
34
6
10
18
39
32
11
20
13
1
9
26
2
33
38
5
37
21
3
4
25
35
36

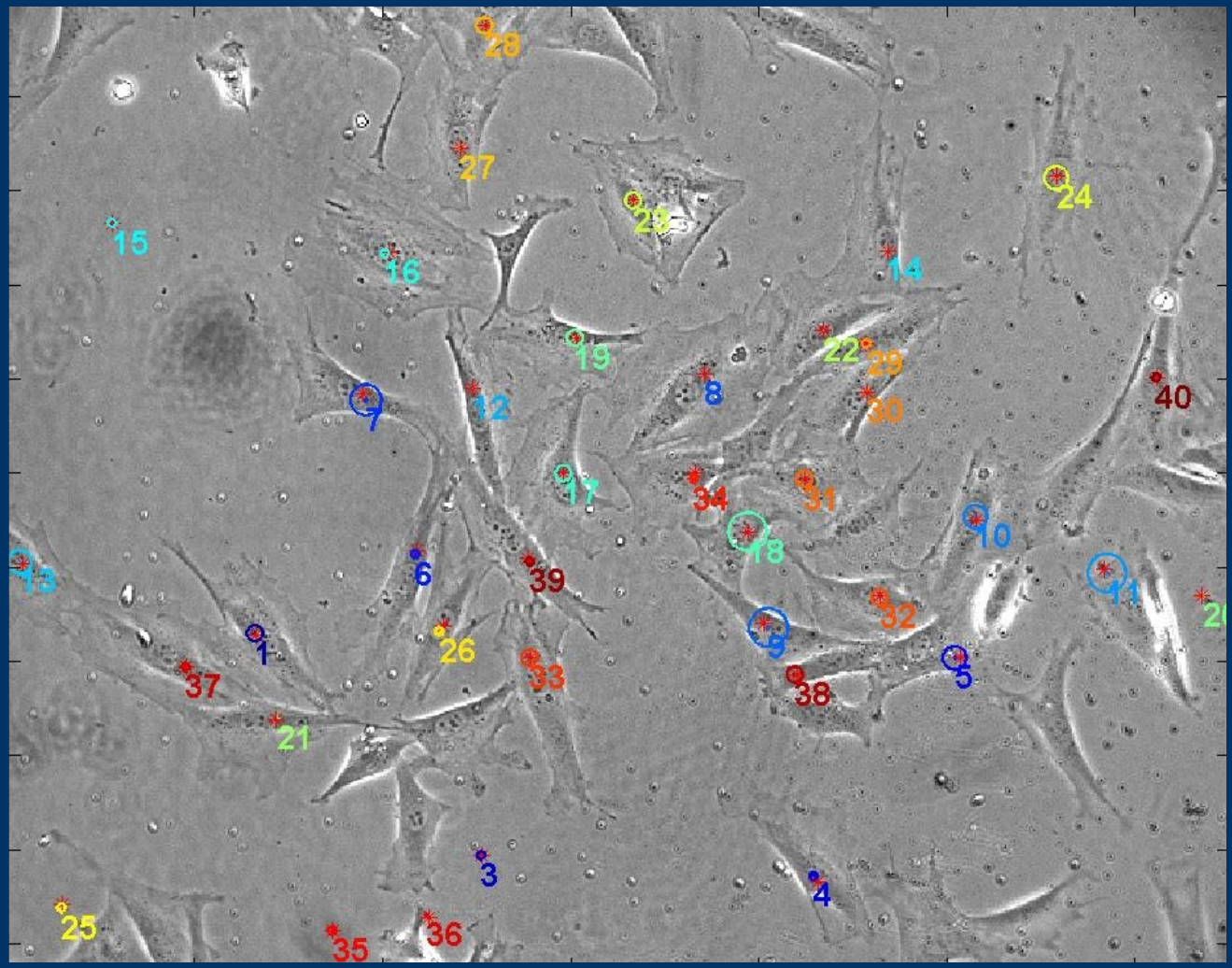
Frame 12


28
27
24
13
23
16
14
19
22
29
7
12
8
40
30
31
17
34
10
6
18
39
13
1
32
11
20
26
33
38
37
5
21
3
4
25
35
36

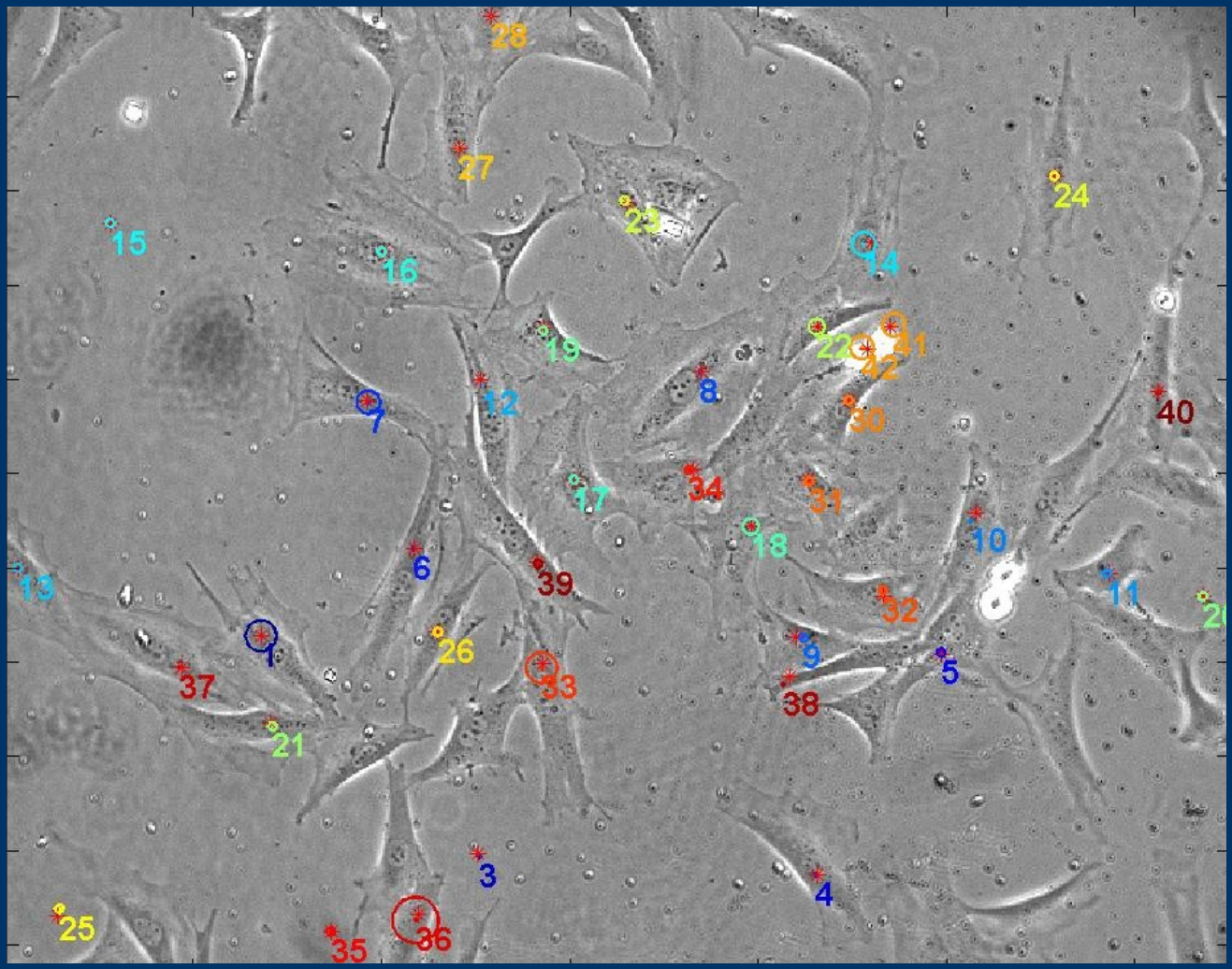
Frame 16


28
27
24
15
23
16
14
19
22
29
40
7
12
8
30
17
34
31
10
6
39
18
11
20
13
1
32
9
26
33
38
5
37
21
3
4
25
35
36

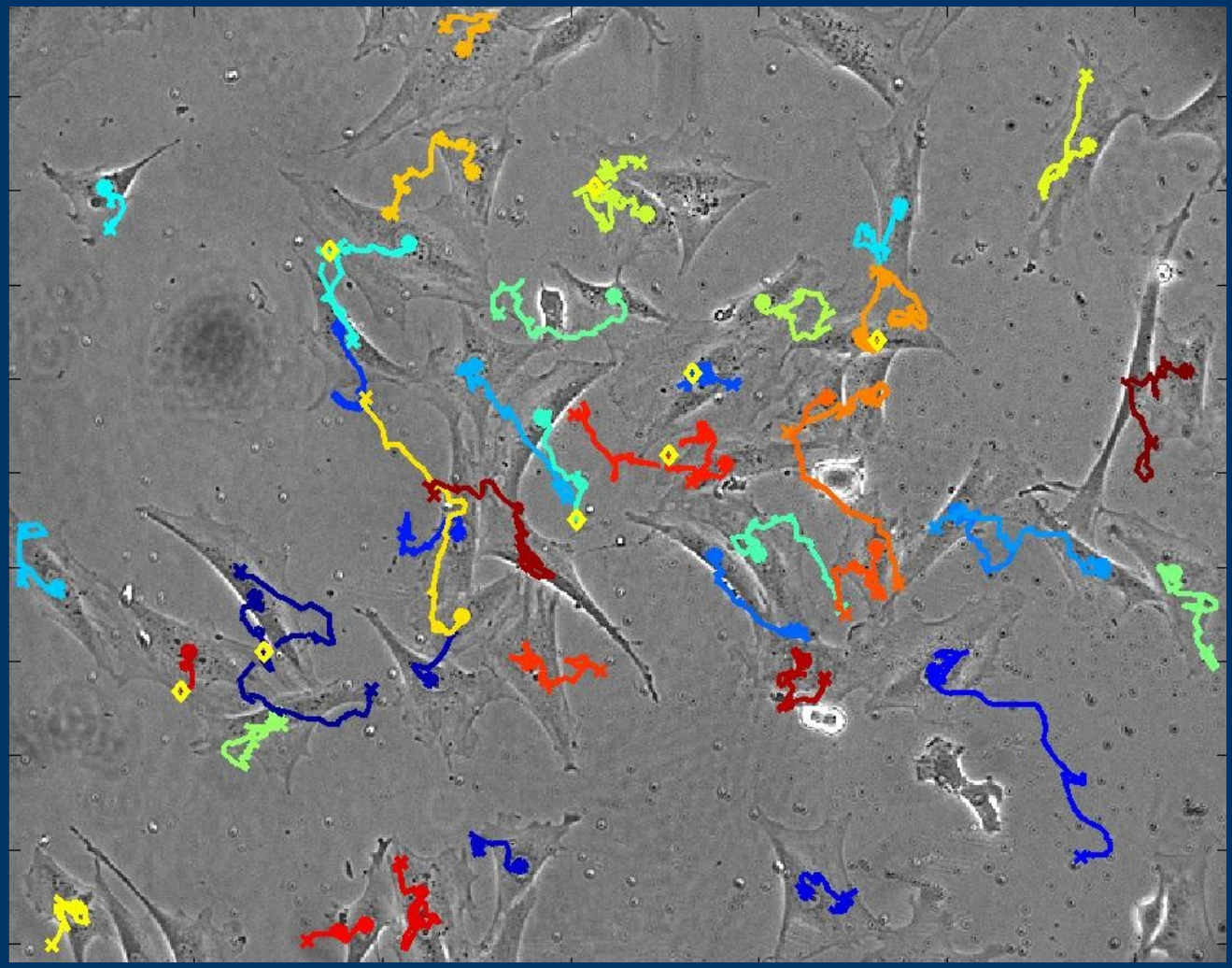
Frame 20


28
27
24
15
23
16
14
19
22
29
8
12
7
40
30
17
34
31
18
10
13
6
39
11
32
1
26
9
33
5
37
38
21
3
4
25
35
36

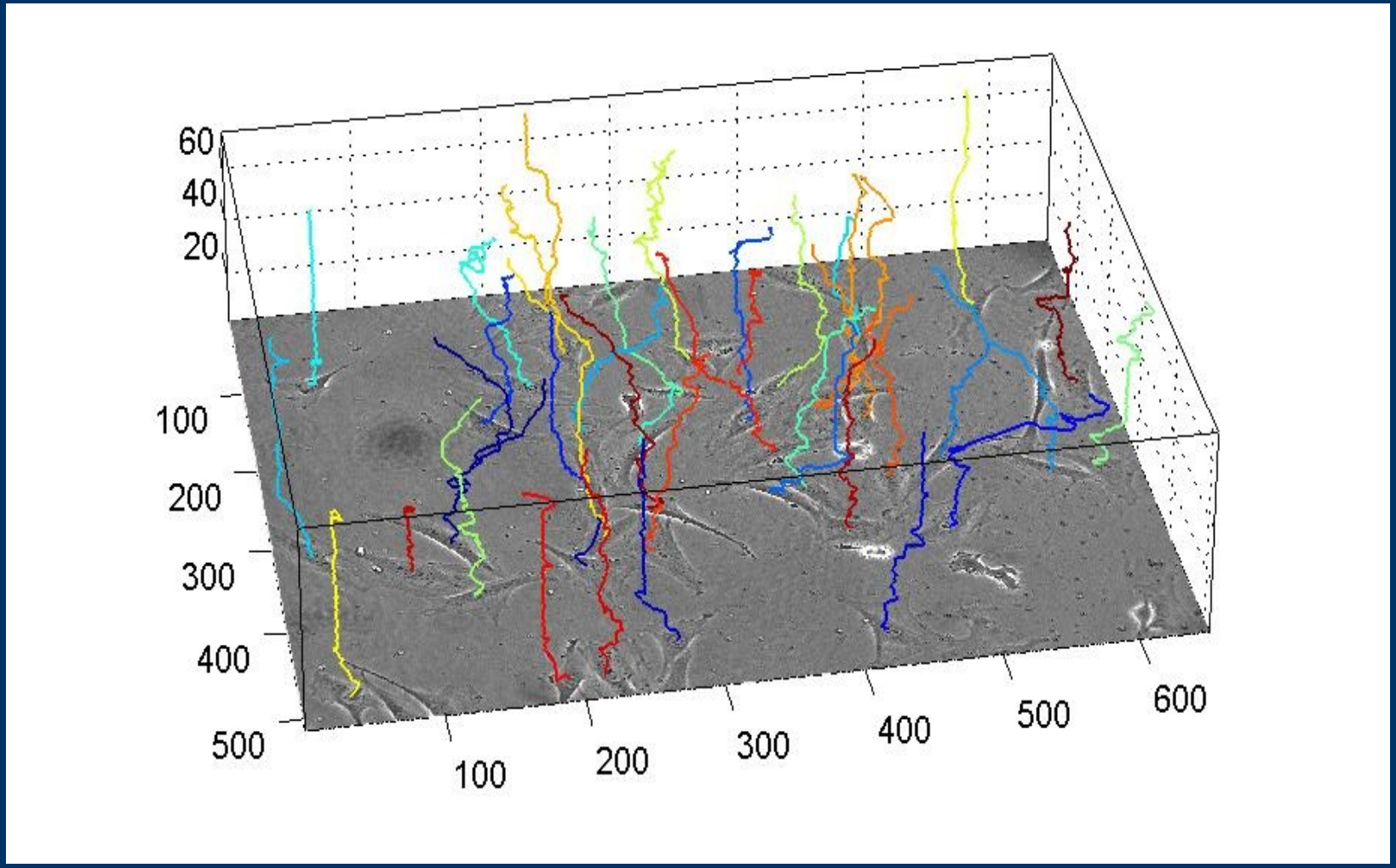
Frame 24


28
27
15
16
23
24
14
19
22
41
42
12
8
7
30
40
17
34
31
18
10
6
39
13
32
11
26
9
5
37
33
38
21
3
4
25
35
36

Trajectories



Spatio-Temporal Visualization



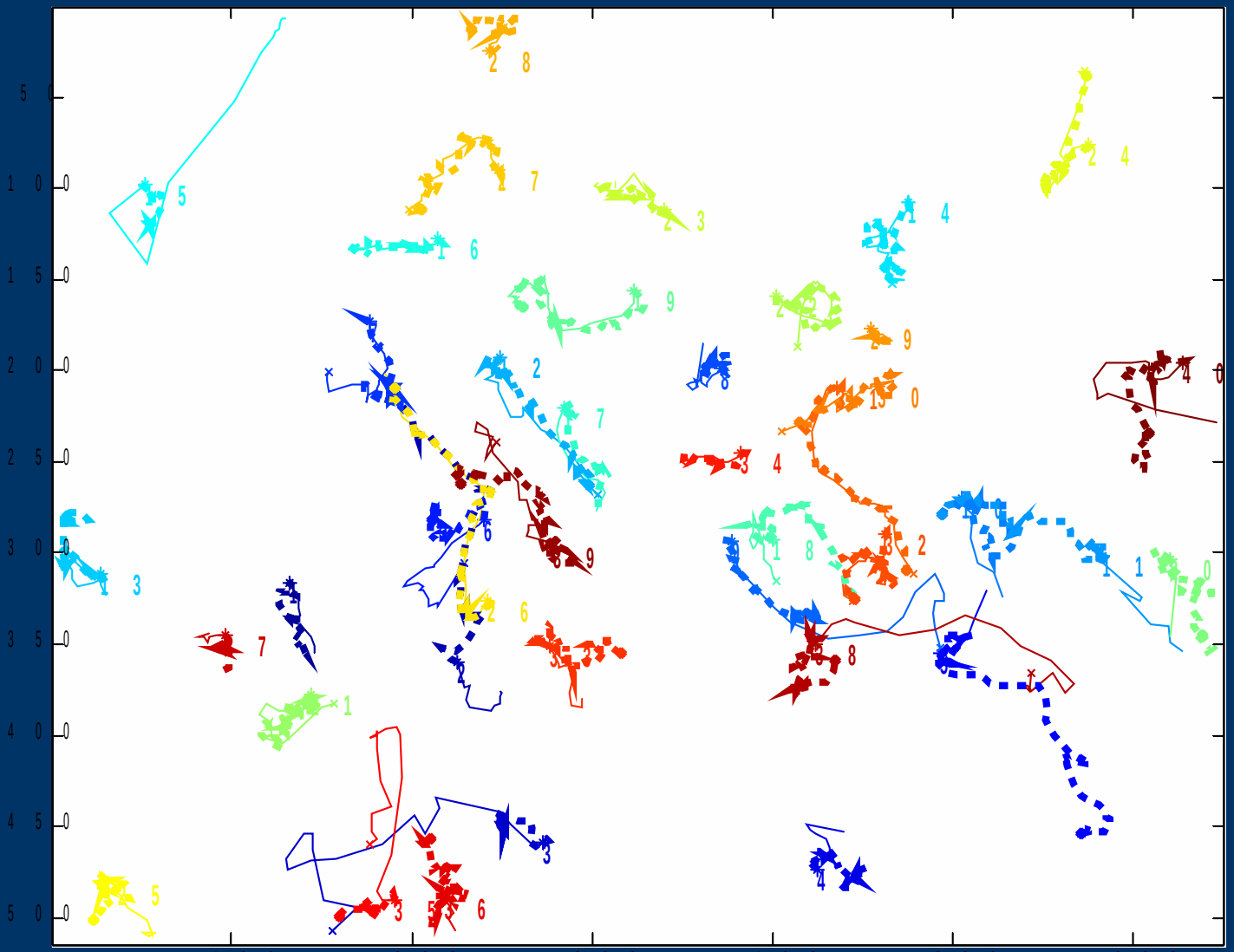
Validation

How to validate?

Best methods around

Ground-truth

Example: Ground-Truth Validation



Application

Location Proteomics

Location Proteomics

Genomic Era
1980 - 2001
Objective: to identify organism DNA sequences and gene mapping

Proteomics Era
1997 - in progress
Objective: to identify proteins, their structure and their functions

Location Proteomics

Genes may encode more than just one protein, and in fact they usually do

Proteins do interact with other proteins to do a specific job

Location Proteomics

Organelles in the cells are related to specific functions

→

proteins sharing the same purpose have a specific location or travel between particular structures (pathways)

Location Proteomics

Therefore knowing the protein localization within the cell may yield an essential role in the identification of their function

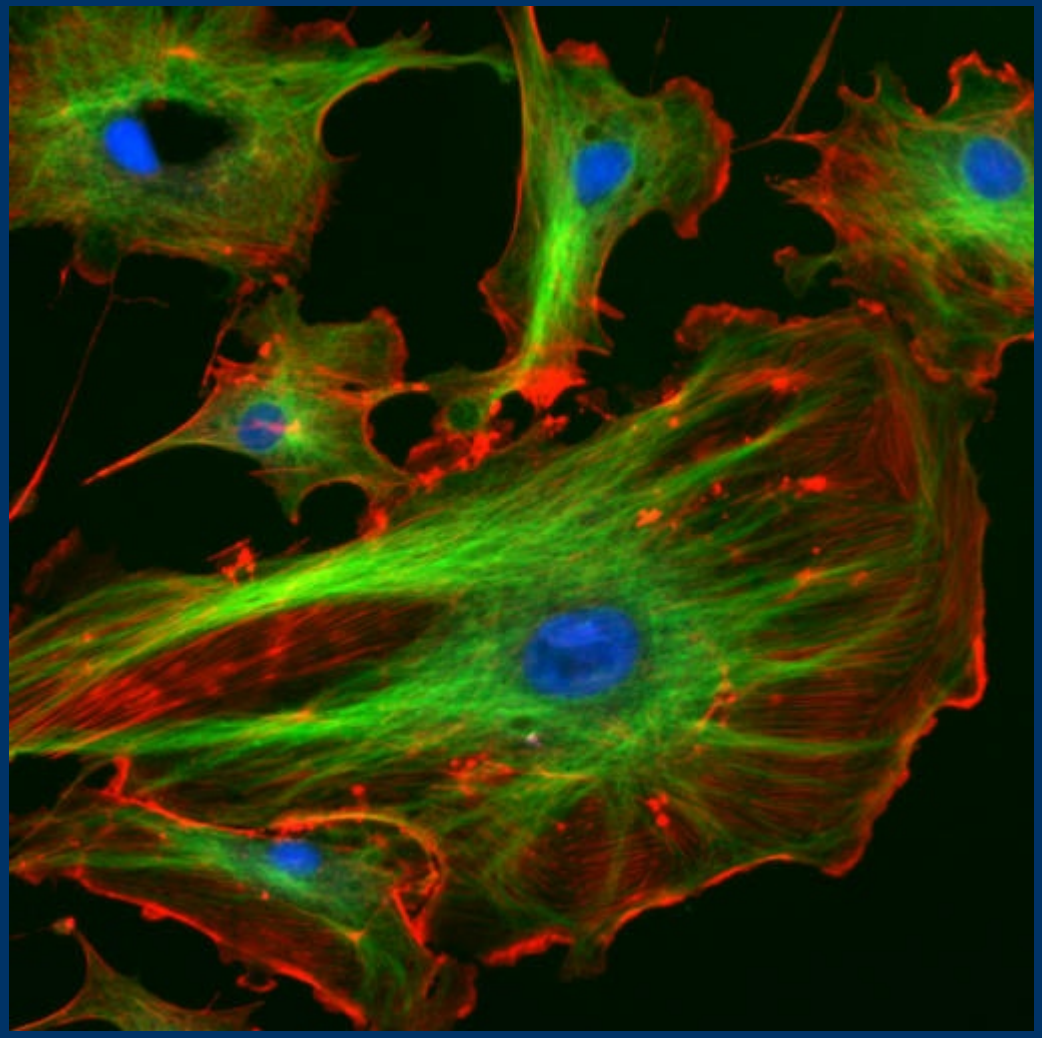
Location Proteomics

Identification of protein subcellular localization may be difficult even for a human expert

Fluorescence microscopy is used for this task

A target protein is marked with a fluorescent marker, then a picture is taken

Fluorescence Microscopy



AI Role

Artificial Intelligence may help in this task

Boland and Murphy [2001] were able to create a classifier system capable of protein localization detection even *better* than a human expert

Location Proteomics

Images of proteins whose location was known were used to train a Support Vector Machine

The SVM was then able to analyze images of new proteins, correctly revealing their subcellular location

They were better than human classification even for very close locations

Features Example

Haralick texture features [Haralick, 1979]

Hu shape descriptors [Hu, 1962]

Zernike moments [Zernike, 1934]

Areas, perimeters

Location Proteomics*: 3D*

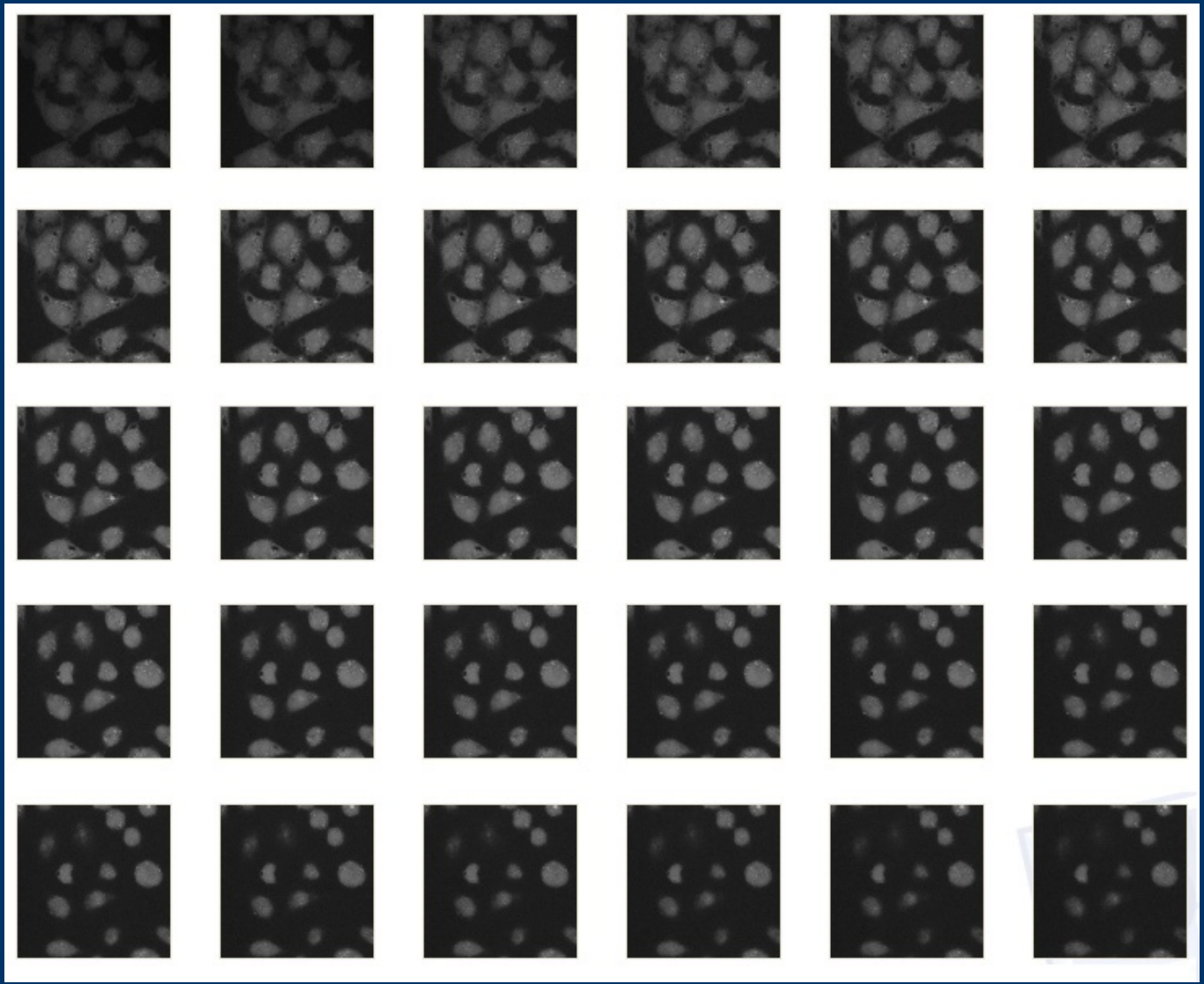
Some fluorescence microscopy techniques, such as confocal microscopy, are able to return three-dimensional volumes instead of two-dimensional images of the observed specimen

Confocal Microscopy

A microscopy technique that can extract slices of the observed specimen

Slices can be used to reconstruct a fully 3D image (*z-stack*) of the specimen

Slices



3D Reconstruction Methods

Image-Based reconstruction

Surface-Based reconstruction

Image-Based 3D Reconstruction

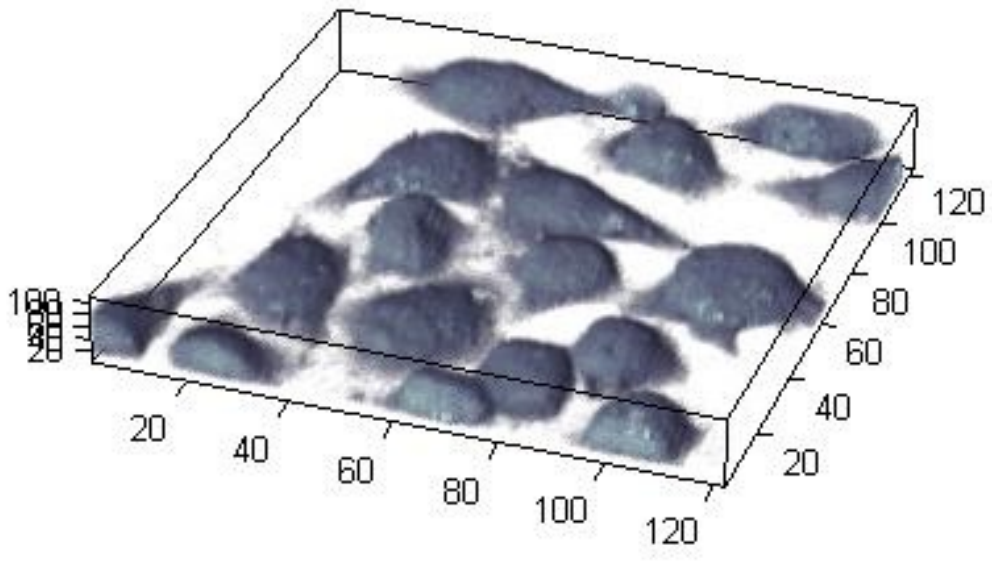


Image-Based 3D Reconstruction

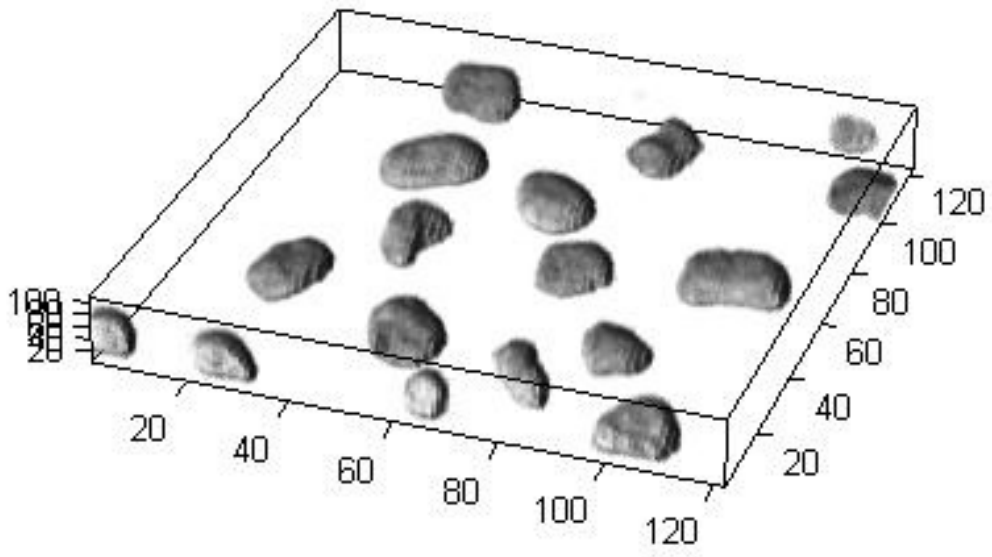
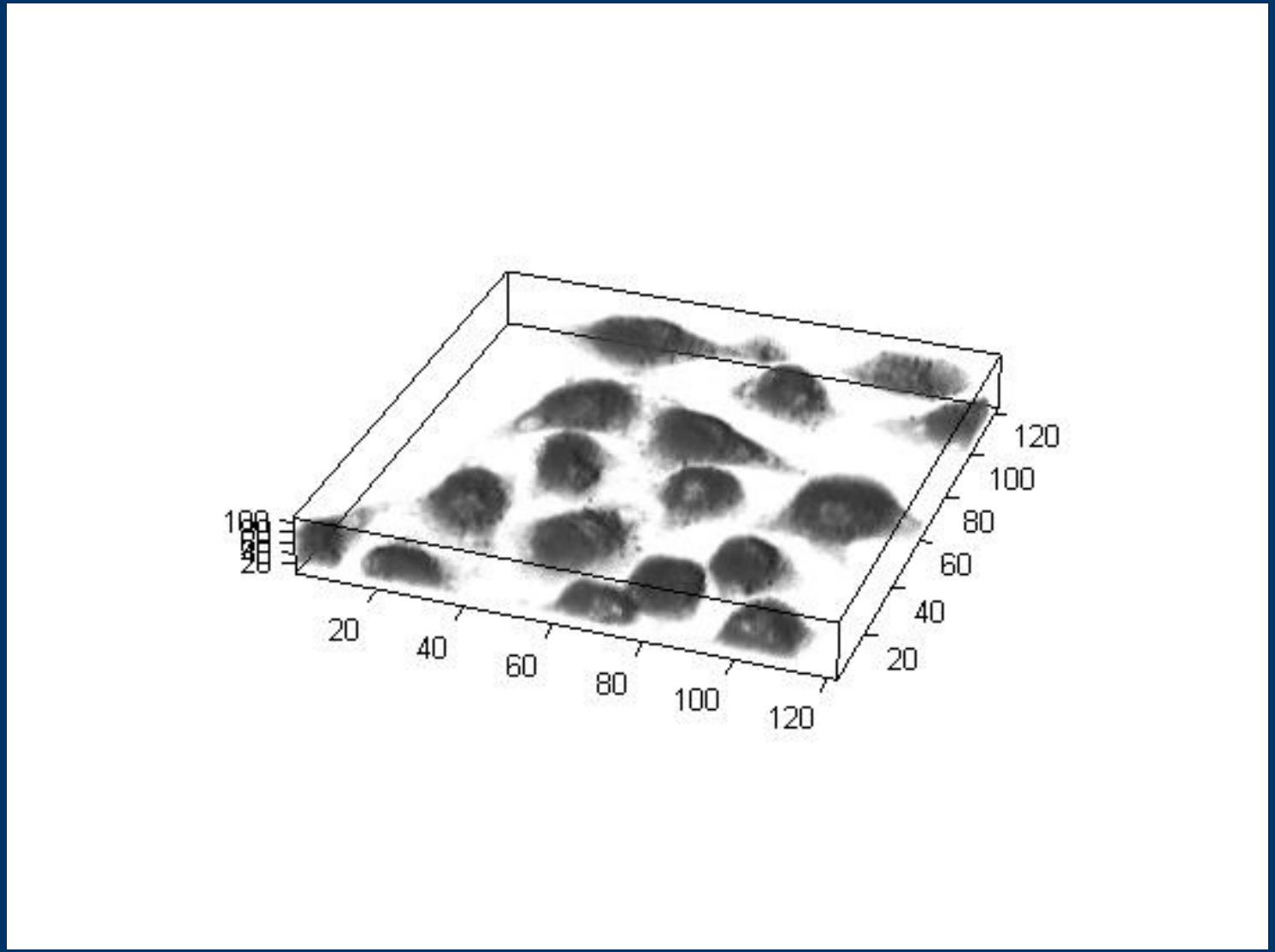
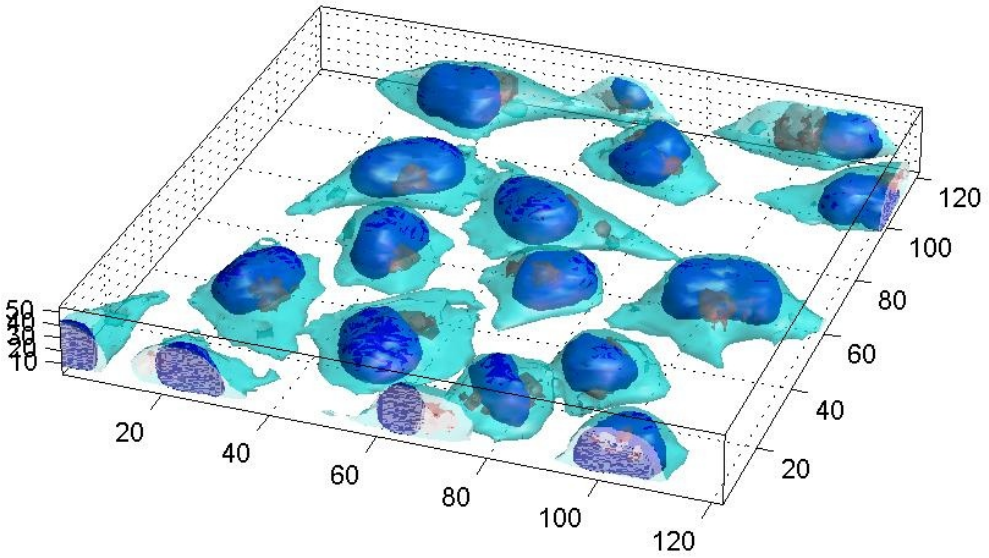


Image-Based 3D Reconstruction



Surface-Based 3D Reconstruction



Location Proteomics*: 3D*

Observing a cell in three dimensions gives more information about the localization of internal proteins

AI Role

The same Artificial Intelligence application just seen can also be applied to three-dimensional fluorescence microscopy

Velliste and Murphy [2002] in fact demonstrated that the extra dimension allows even better results

3D Features Example

Features are adapted to 3D:

3D Haralick texture features

3D Hu shape descriptors

Volumes and surfaces

Case: AI Application

Objective: classification of HeLa cells according to a mutant protein (UCE)

HeLa

A *cell line* is a permanently established cell culture that will proliferate indefinitely

HeLa is a famous cancer cell line

Obtained from a patient in 1952

Cancerous cells are immortal

Mutants

Identification of mutants of a particular protein in HeLa cells, called UCE

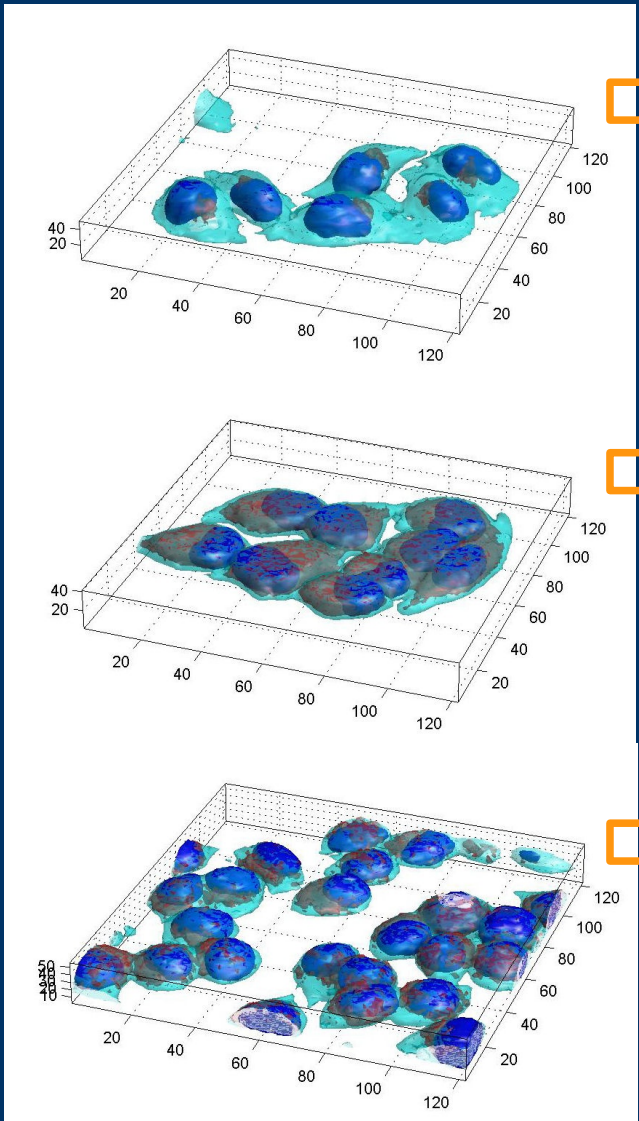
This protein in normal conditions travels in a particular way in the cell

Mutant cells show a different behaviour of the UCE protein

Question

Can we identify the mutants given their confocal microscopy images via an SVM?

Learning

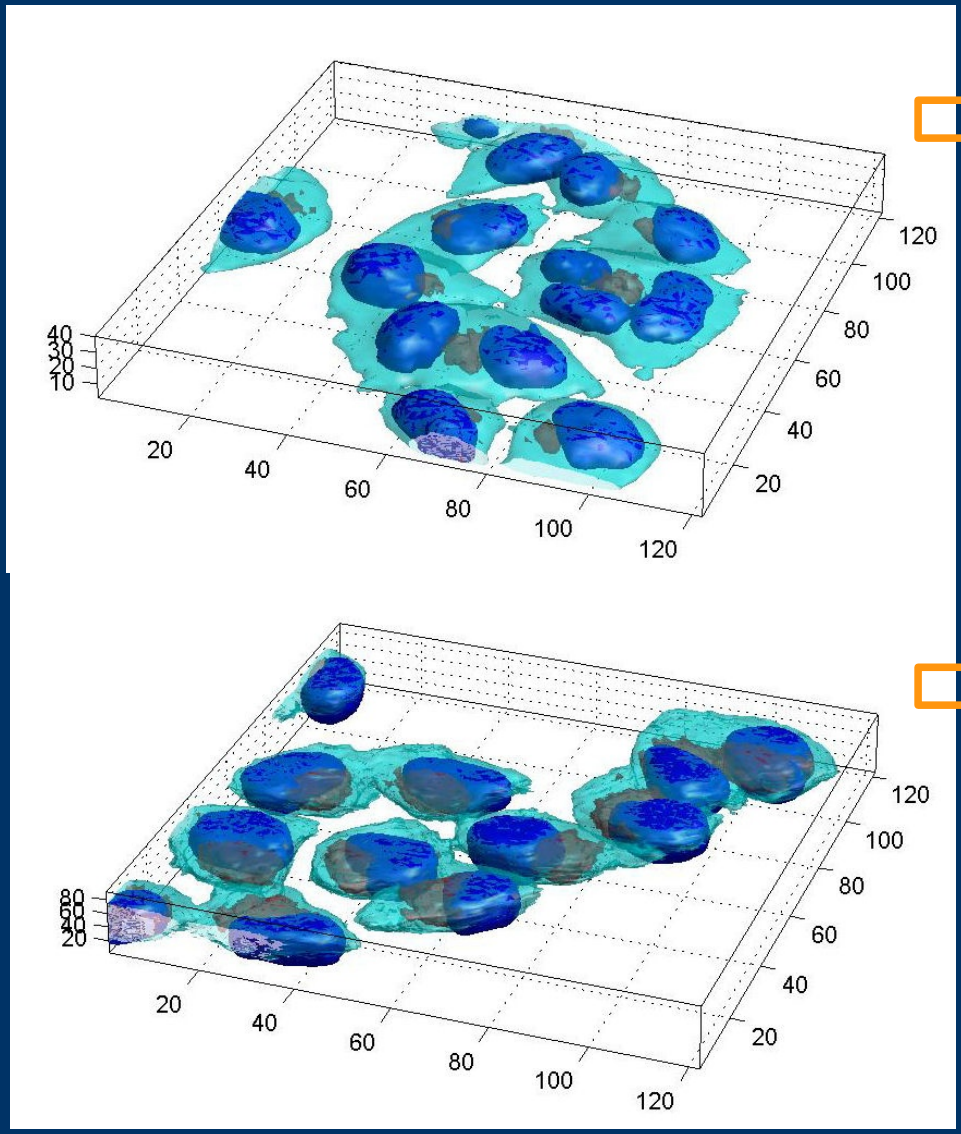


HeLa cells
mutant '502 Stop'

HeLa cells
mutant 'Y^{488}-A'

HeLa cells
mutant 'YQEMN-A'

Classify New Data



HeLa cells
which mutant?

HeLa cells
which mutant?

Features

We'll be using the same features above (Haralick, Hu, volumes)

2D or 3D?

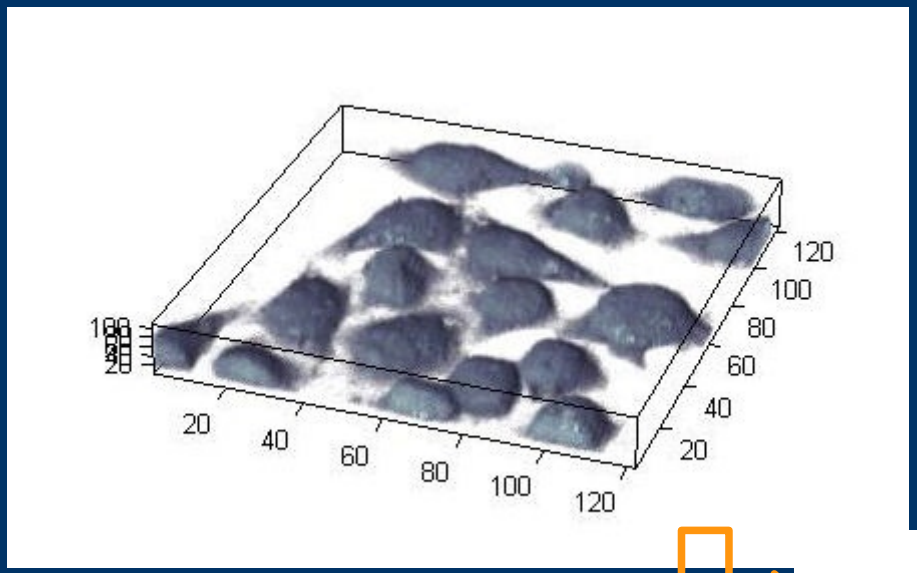
Identifying the mutants turns out to be a *Location Proteomics* problem

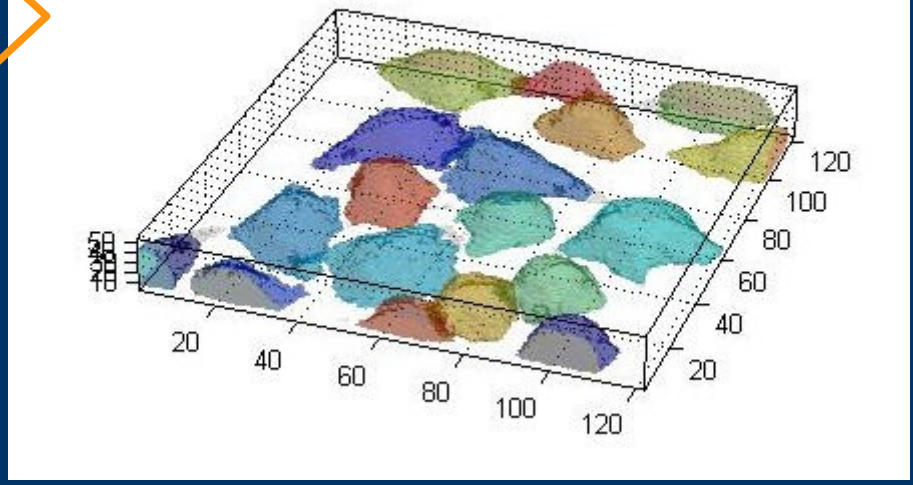
2D gives good results

3D gives better results

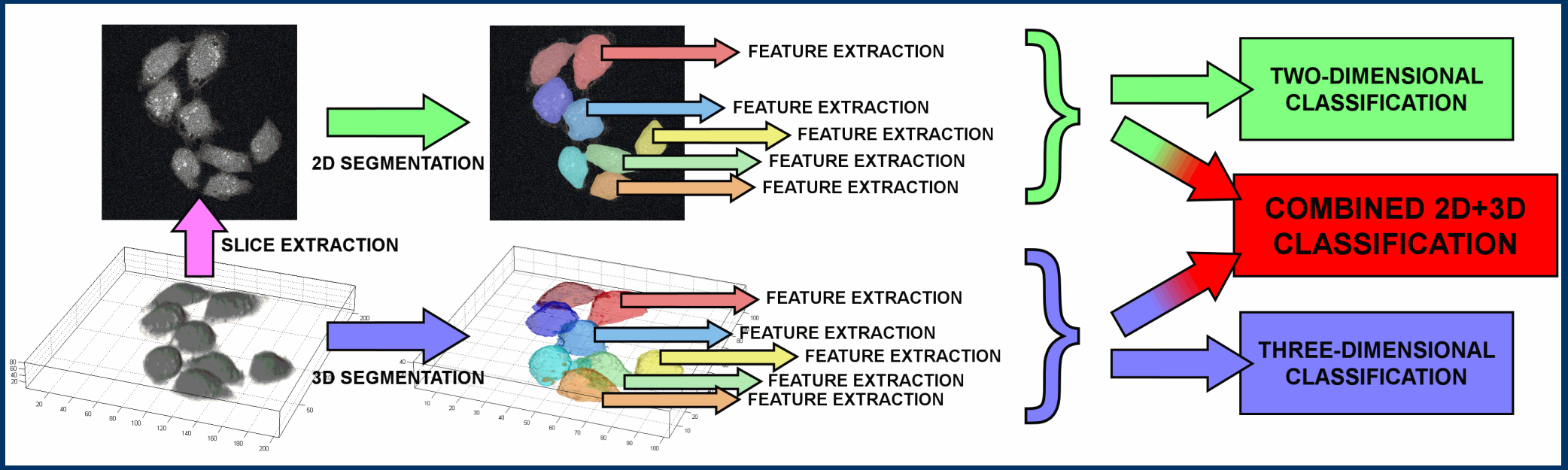
How about a combined 2D+3D classification?

Three-Dimensional Segmentation





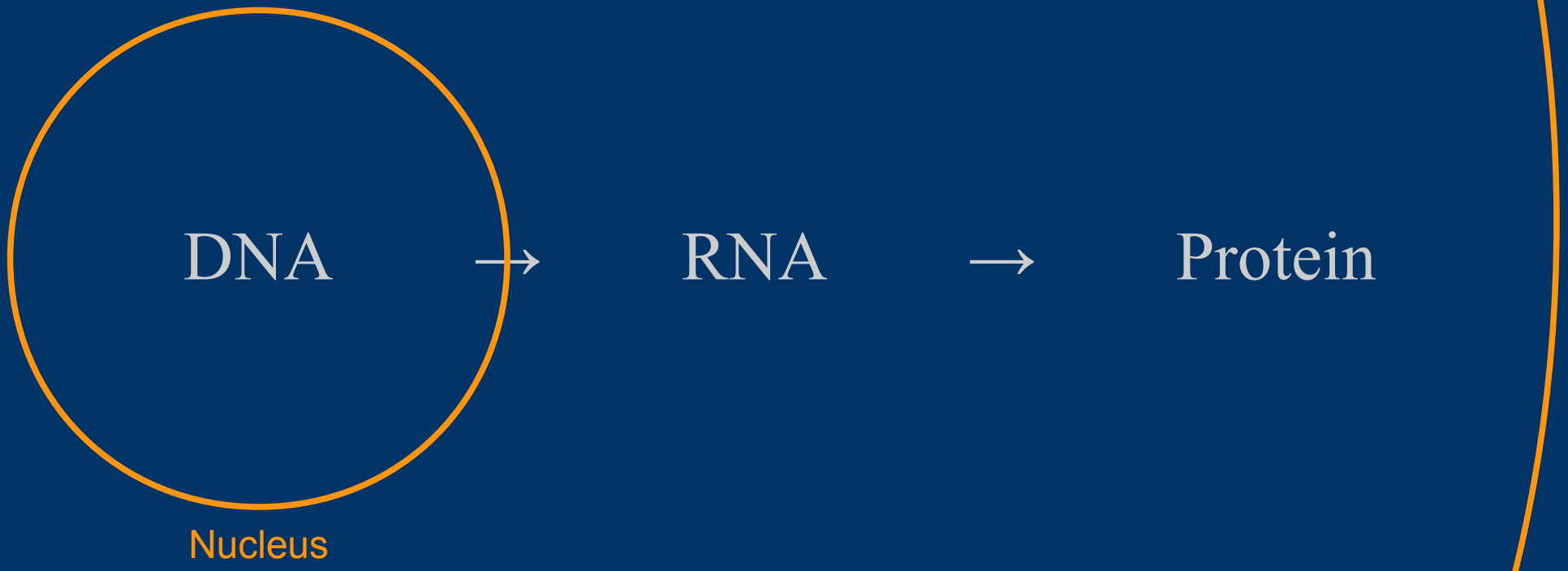
Combined Classification



Application

Gene Networks

From DNA to protein



“Gene Networks”

RNAs can interact between them

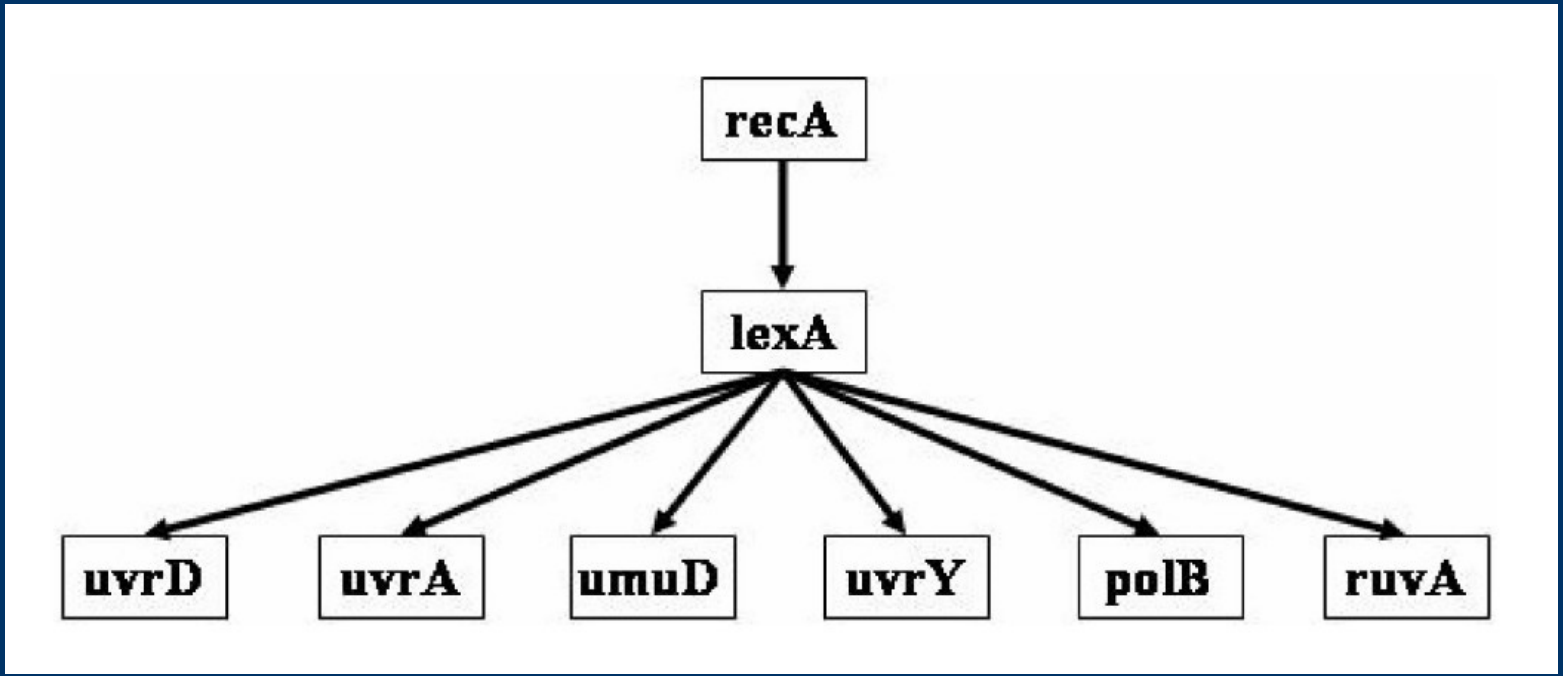
→

Genes interact between them

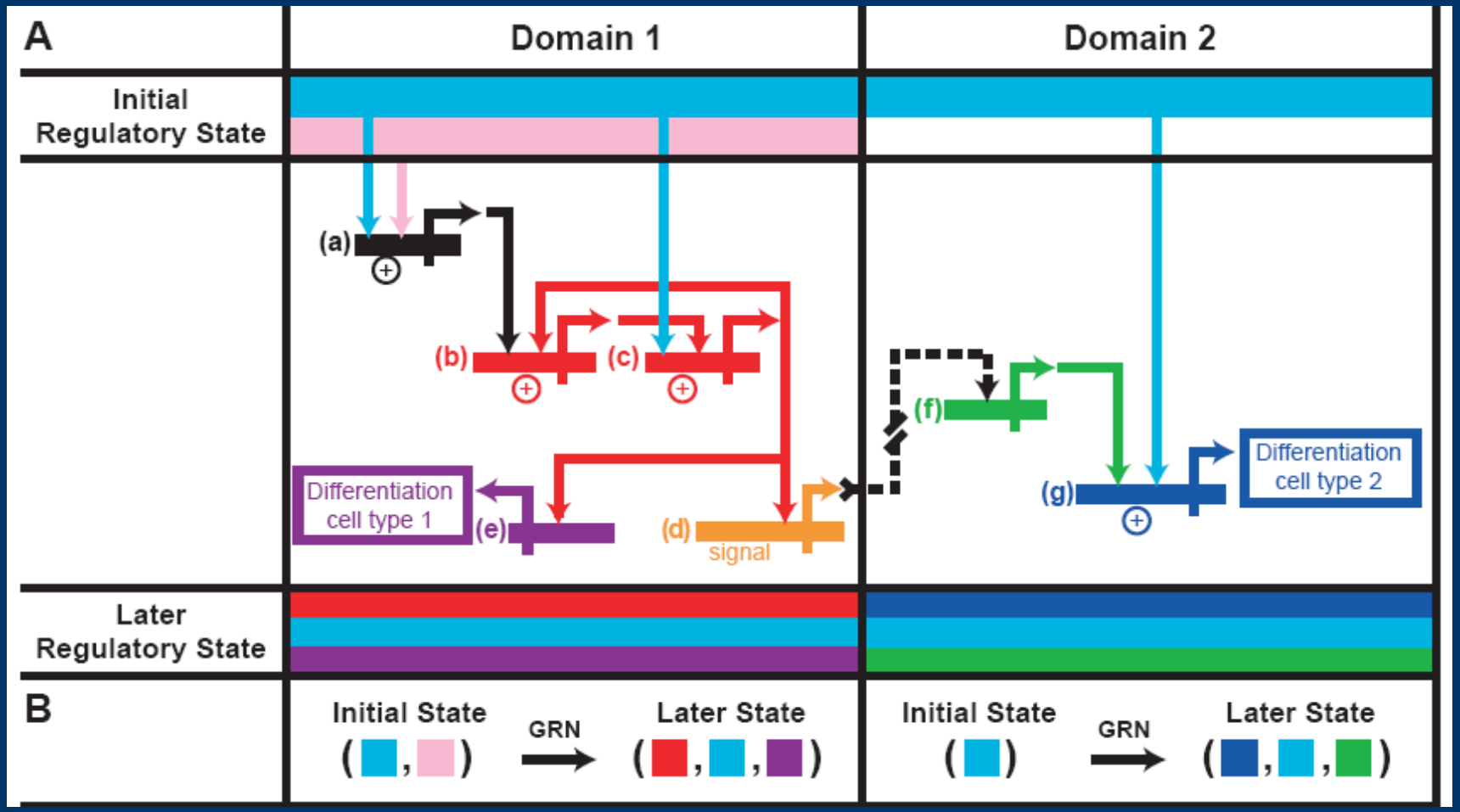
Interactions can be put in a graph

We talk about “Gene Networks”

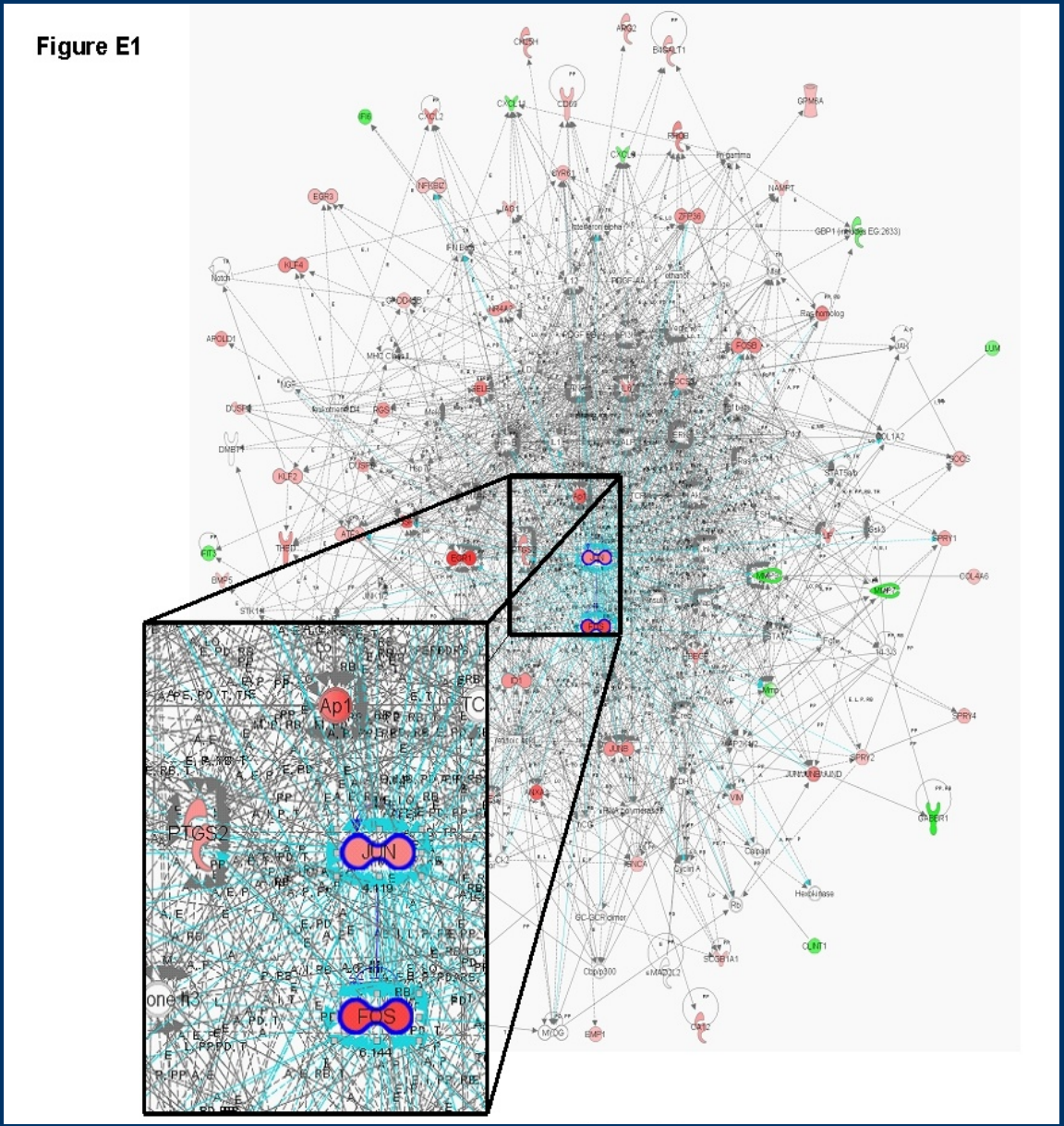
Example: E. Coli *SOS Network*



Engineer View: Network → Circuit



Things Can Get Very Complicated



Gene Networks

Not all gene interactions are known

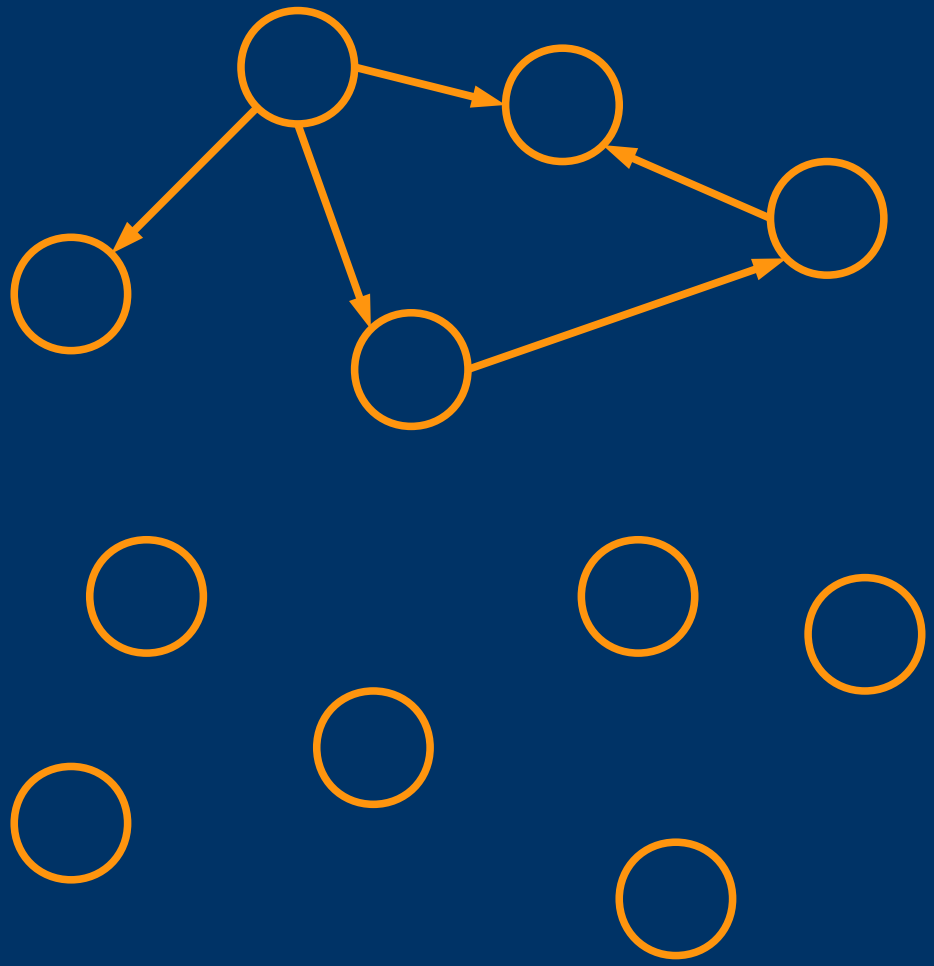
How can we retrieve the existing but unknown interactions?

Reverse engineering approaches on genetic code

Case: Gene Network Inference

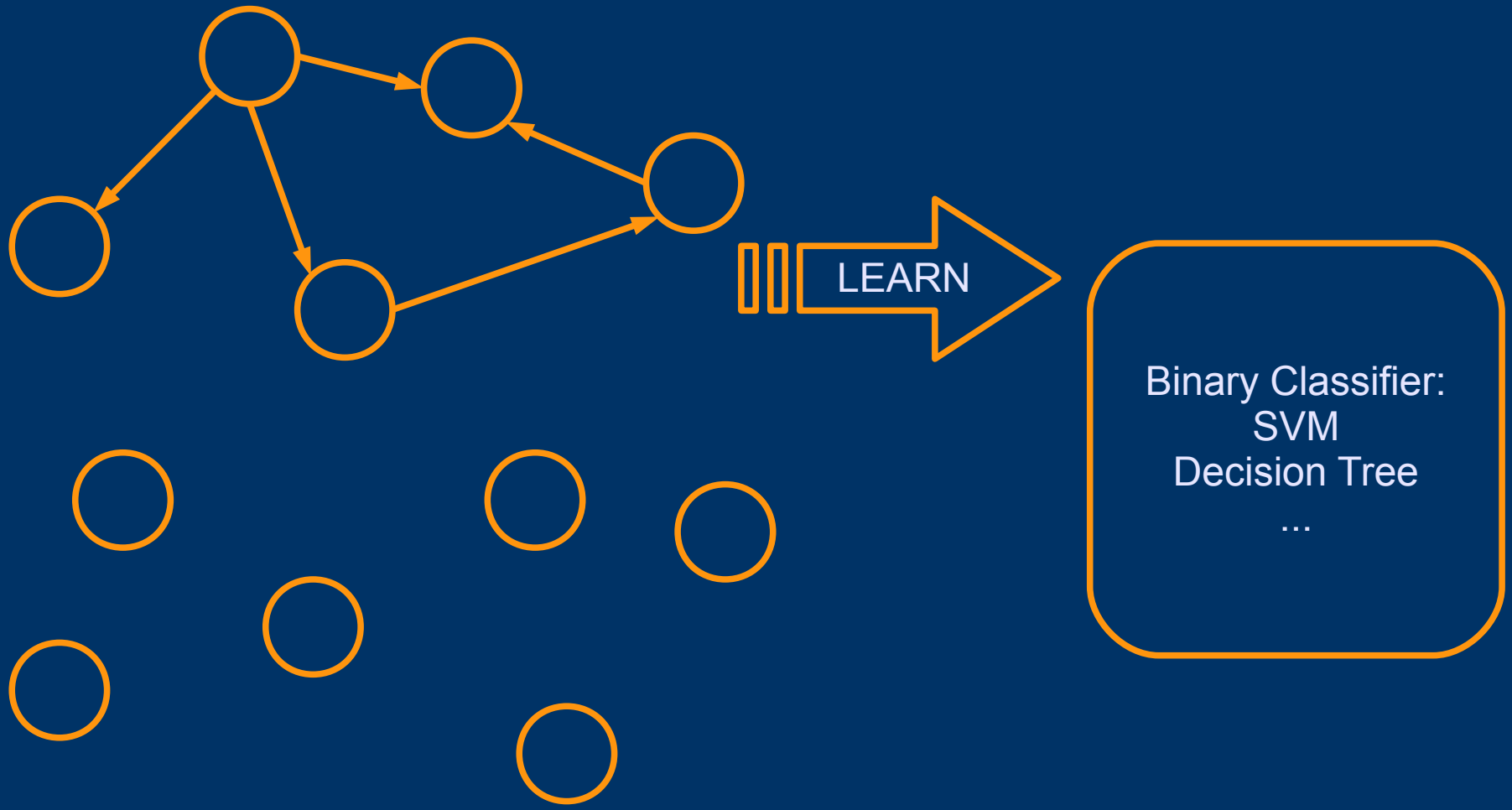
Objective: given genomic data and the known gene-gene interaction infer missing gene-gene interactions

Objective

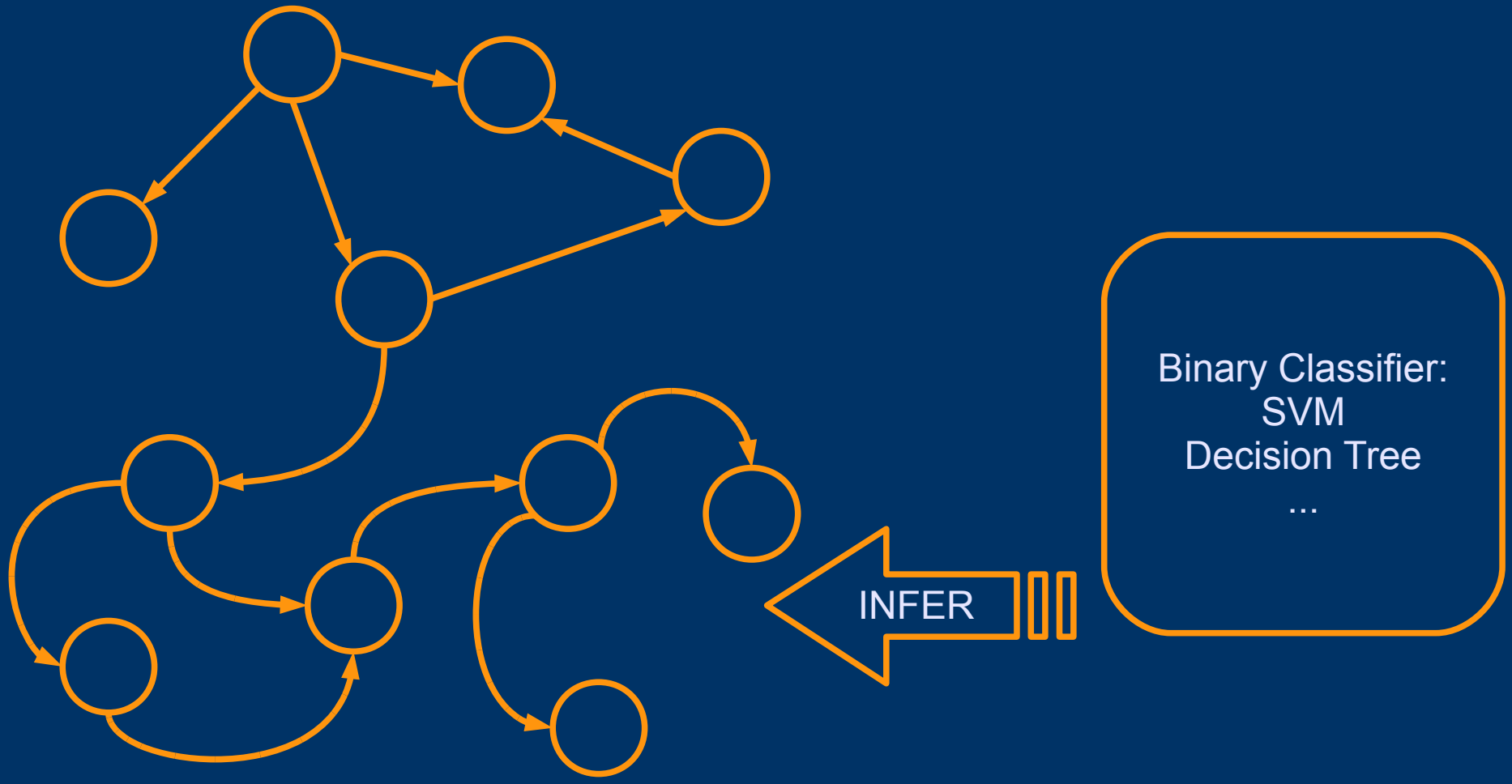


Part of the network is already known

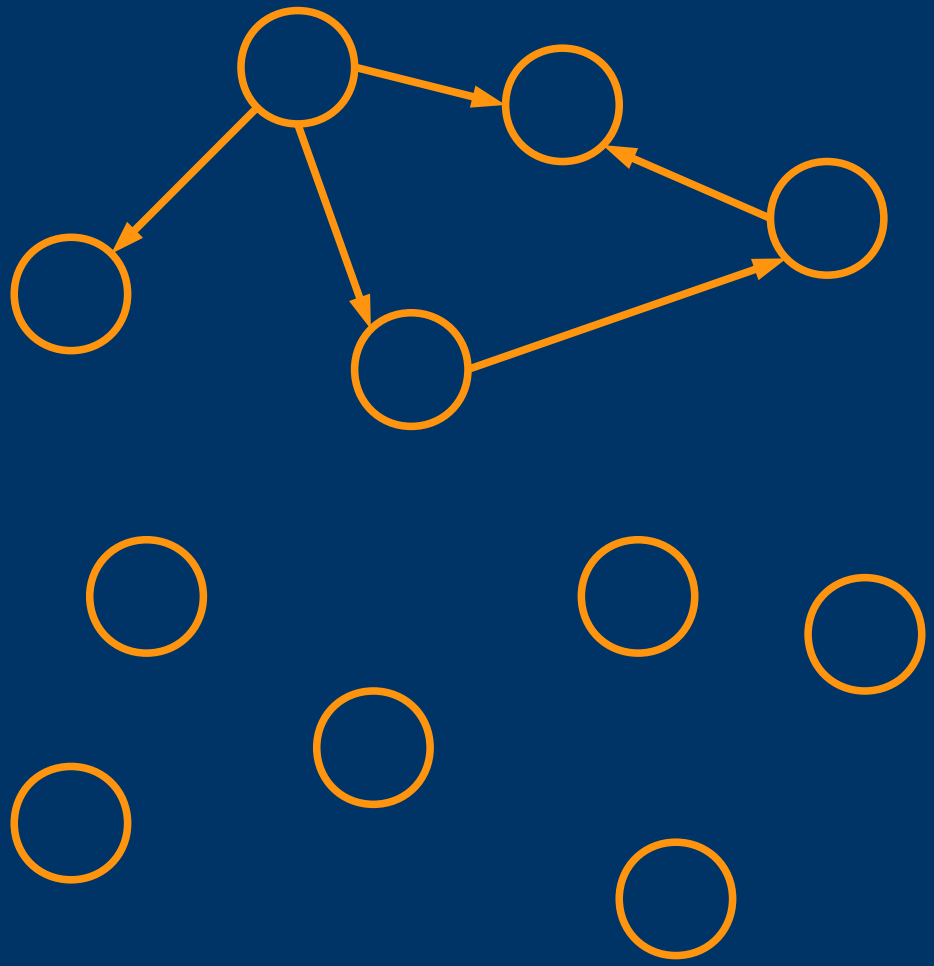
Objective



Objective



Objective



Problem: a binary classifiers learns from positive and negative examples

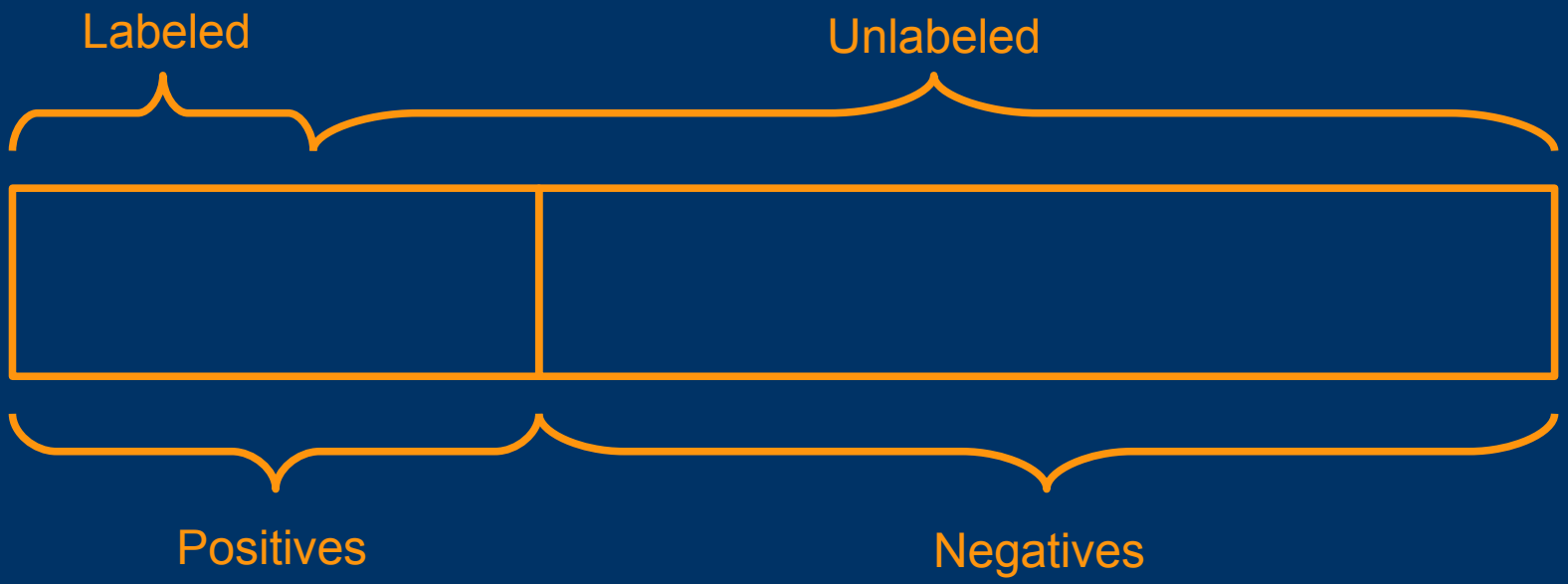
Which are the negatives?

Negatives?

Biologists are very confident on known positives but are not aware on non-interactions (negatives)

Current Approaches

Unlabeled → negatives [Mordelet, 2008]
Random selection [Grzegorczyk *et al.*, 2008]
Heuristics [Cerulo *et al.*, 2010]



Heuristics Example

The network has no or few cycles

The network has a tree like structure

Dataset

A network generated with the software GeneNetWeaver

Percentage of known positives going fro 100% to 10%

Dataset


Positives
Negatives
P=100%
P=90%
P=80%
Known positives
Unlabeled data
P=10%

Question

How should the precision/recall of positives and the precision/recall of negatives of the set selected by the heuristic vary with the percentage of known positives?

Answer

We want the precision/recall of negatives to be high

And the precision/recall of positives to be low

Question

What should be the performances of a classifier trained with the set selected by the heuristic and one trained with a random set?

Answer

We expect the heuristic set to outperform the random set in particular for small values of P

Thank You