Computational Biology: A Measurement Perspective

Alden Dima

Information Technology Laboratory

alden.dima@nist.gov

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Problem

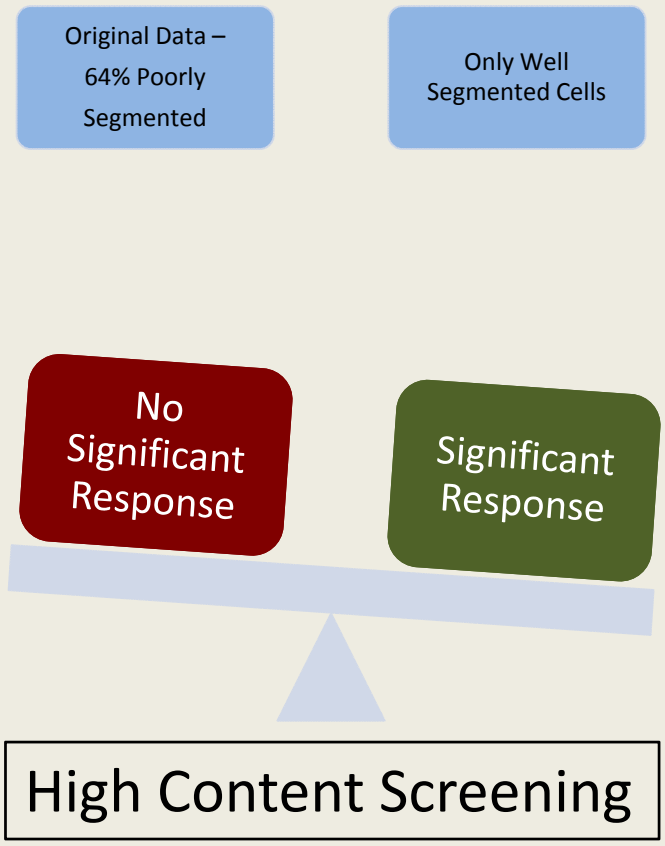
- High-throughput technologies are generating large amounts of complex data that is difficult to process and convert into knowledge
- Issues exist throughout data lifecycle:
  - Acquisition
  - Analysis
  - Archiving
  - Interchange

Imaging Technologies

- Imaging technologies are increasingly being used both as diagnostic tools and as research tools in the biosciences
  - Novel methods are needed for automated analysis and comparisons
  - Correlation among studies is difficult at best
  - Off-the-shelf methods are not well characterized and can contribute significantly measurement uncertainty
  - Need to combine features from images with other biological or medical sources

Example from Literature

- Relies heavily on cell imaging
- Gigabytes of images collected
- Algorithms treated as “black boxes”



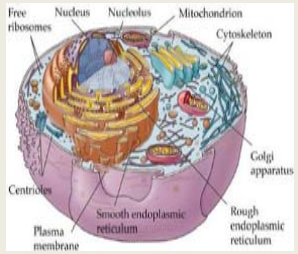
- Many algorithms published but rarely compared
- Poor segmentation significantly effects conclusions

Source:

Hill, LaPan, Li, Haney (Wyeth Research)
Impact of image segmentation on high content screening data quality for SK-BR-3 cells
BioMed Central Bioinformatics 2007

Computational Biology: *Single Cell Analyses*

Intracellular molecular reactions and interactions that control the response and fate of cells and organisms cannot be unambiguously compared and combined due to a lack of standards and validated protocols

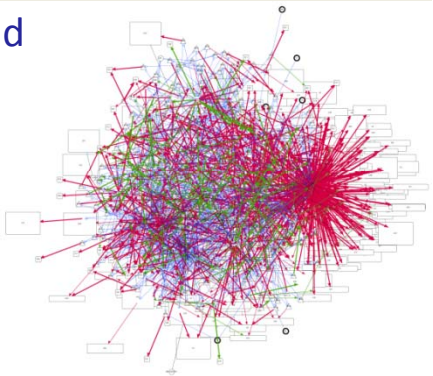


NIST Role

- Provide the measurement tools and standards that enable quantifiable and reproducible measurements of cells and their interactions through:
  - Standard data/metadata format for image capture, storage, retrieval, analysis
  - Software to enable high throughput cell image analysis and interoperability
  - Standards and validation required to ensure reproducible image analysis
  - Multi-site experiments to test software and validation protocols

Technical Approach

- Create and evaluate an integrated data collection, organization and analysis infrastructure for cellular imaging
- Experimentalists and computational scientists - focus on the physical standards and protocols for data collection, image processing and analysis, storage of data and metadata, and evaluation of results

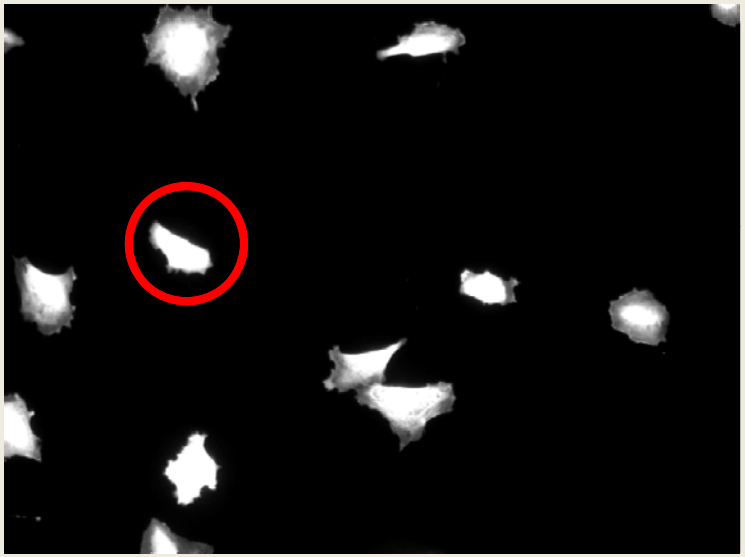


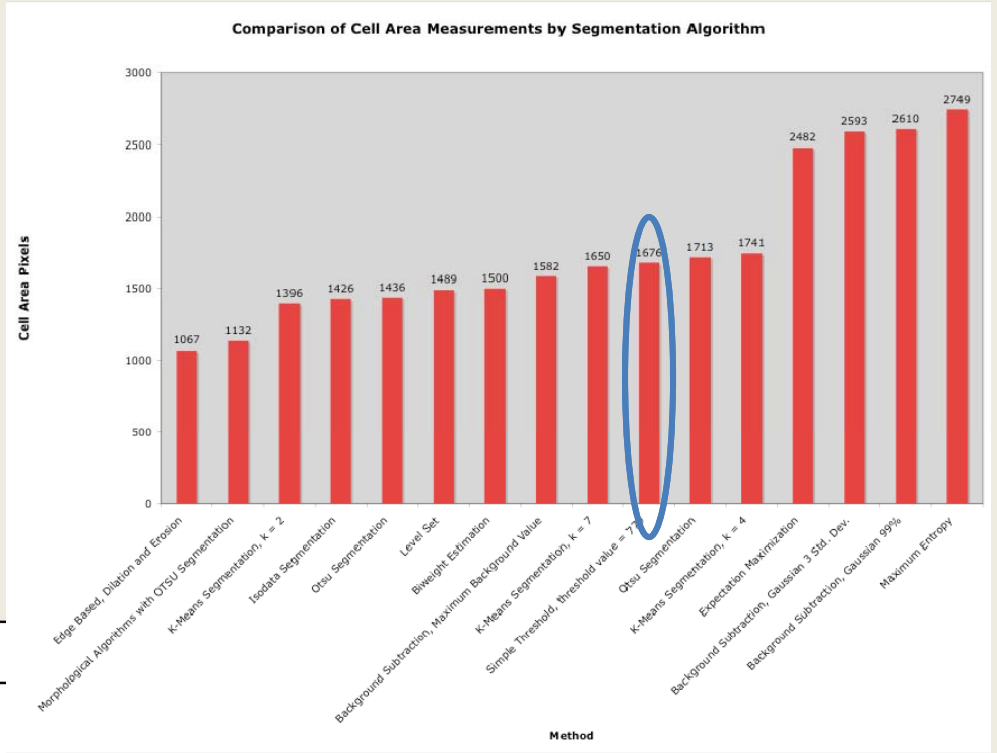
Segmentation Evaluation

- Determine which segmentation technique and associated parameters can be used to reliably determine the morphology of cells for the purposes of comparing cell lines as part of a new standard procedure under development

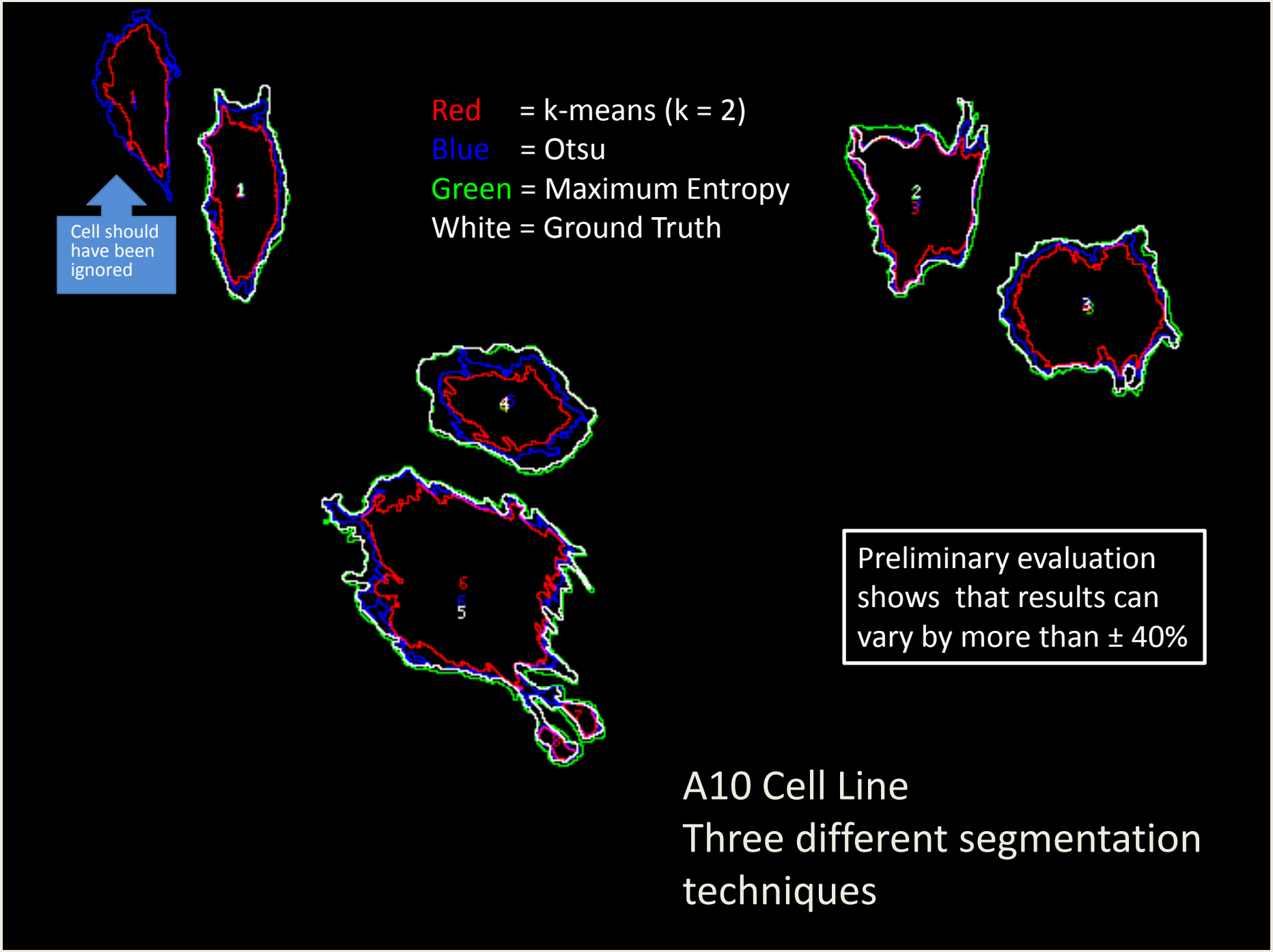
Variability Across Methods

Different segmentation techniques can change results



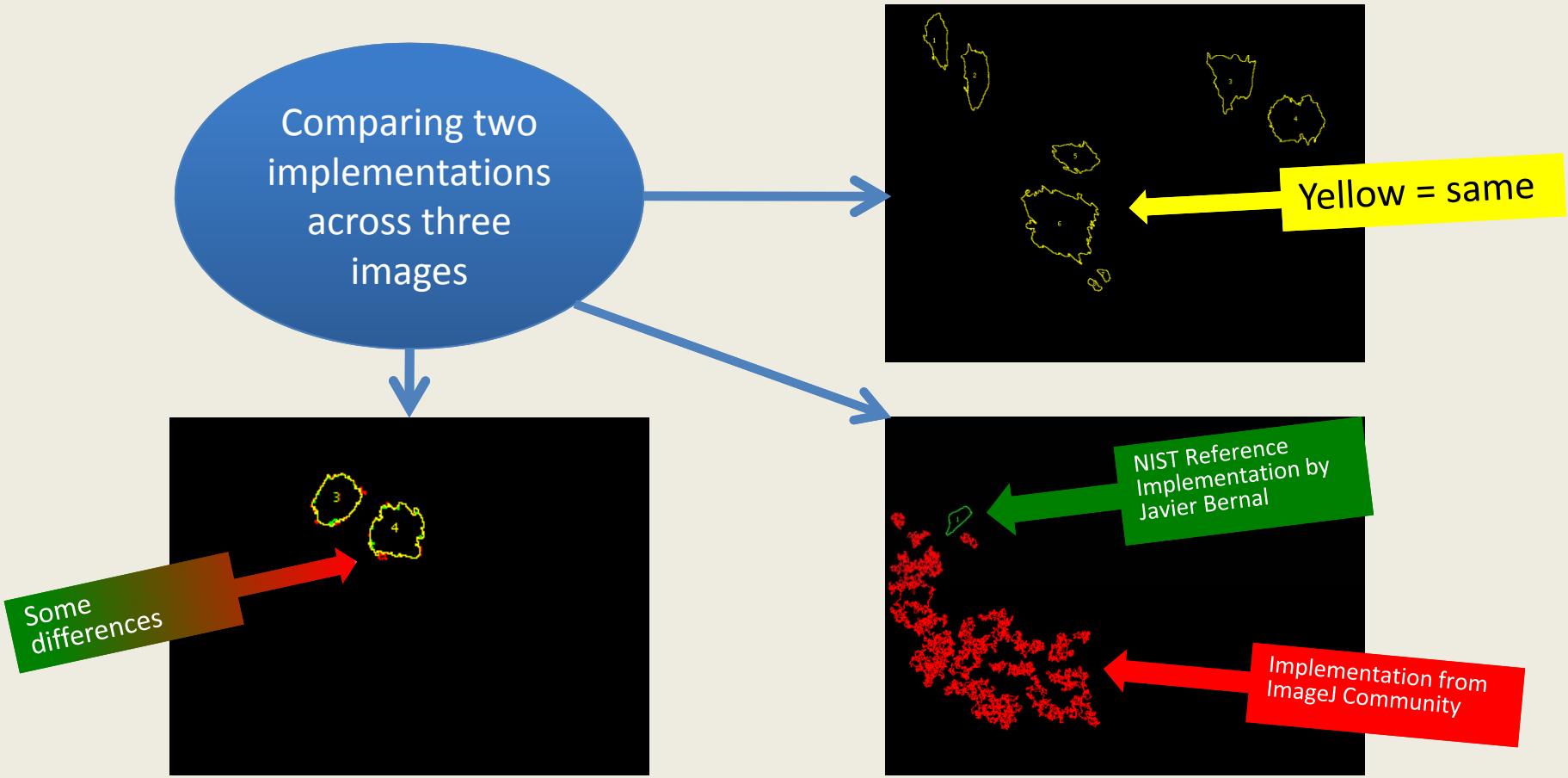


Method ID	Person	Reference Cell Area Value (pixels)	Tool	Method
1	John Elliot	1676	ImageJ	Simple Threshold, threshold value = 779
2	Asim Wagan	1436	Matlab	Otsu Segmentation
3	Asim Wagan	1426	Matlab	Isodata Segmentation
4	Asim Wagan	1741	Matlab	K-Means Segmentation, k = 4
5	Asim Wagan	2482	Matlab	Expectation Maximization
6	Marcin Kociolek	2593	C++	Background Subtraction, Gaussian 3 Std. Dev.
7	Marcin Kociolek	2610	C++	Background Subtraction, Gaussian 99%
8	Marcin Kociolek	1582	C++	Background Subtraction, Maximum Background Value
9	Xiao Lan Li	1489	Matlab	Level Set
10	Xiao Lan Li	1067	Matlab	Edge Based, Dilation and Erosion
11	Rui Fang	1132	Matlab	Morphological Algorithms with OTSU Segmentation
12	Jim Filliben	1396	Dataplot	K-Means Segmentation, k = 2
13	Adele Peskin	1500	C++	Biweight Estimation
14	Javier Bernal	1650	FORTRAN	K-Means Segmentation, k = 7
15	Javier Bernal	2749	FORTRAN	Maximum Entropy
16	Javier Bernal	1713	FORTRAN	Otsu Segmentation


Cell should have been ignored
Red = k-means (k = 2)
Blue = Otsu
Green = Maximum Entropy
White = Ground Truth
Preliminary evaluation shows that results can vary by more than ± 40%
A10 Cell Line
Three different segmentation techniques

Variability Across Implementations

Different implementations of the **same technique** can change results as well

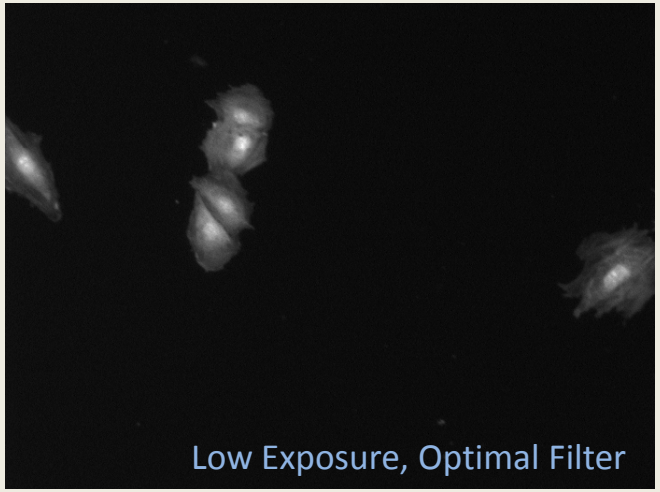


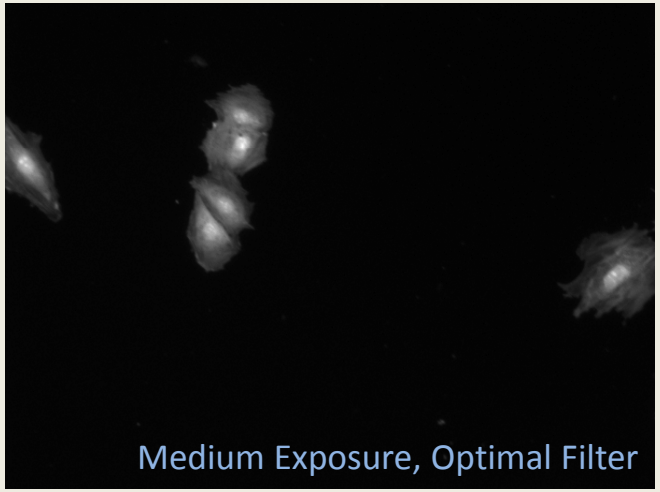
"Software as measurement"

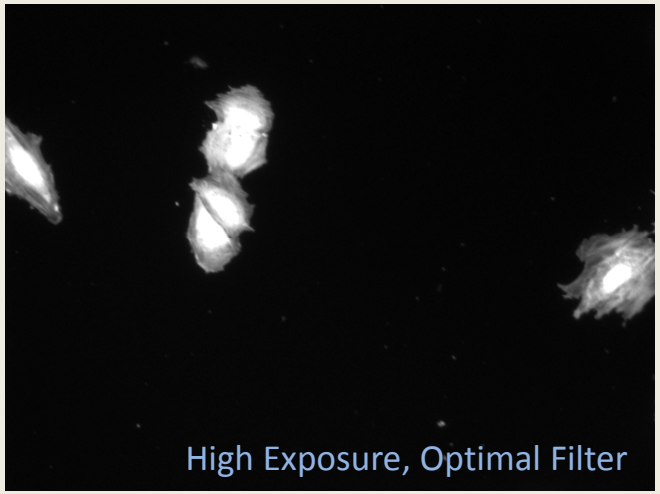
Experimental Design

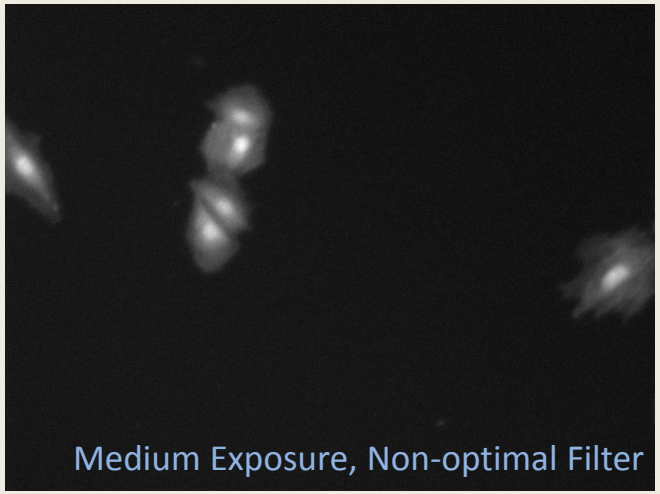
- Two cell lines – A10 & 3T3
- Cell preparation follows ASTM document
- Different image exposure and filter levels
- No flat-field correction
- Multiple sampling to capture noise
- Ground truth: expert manual segmentation
- Result: approximately 8000 images – 80 of which are used for initial evaluation work

Red Channel Images - A10



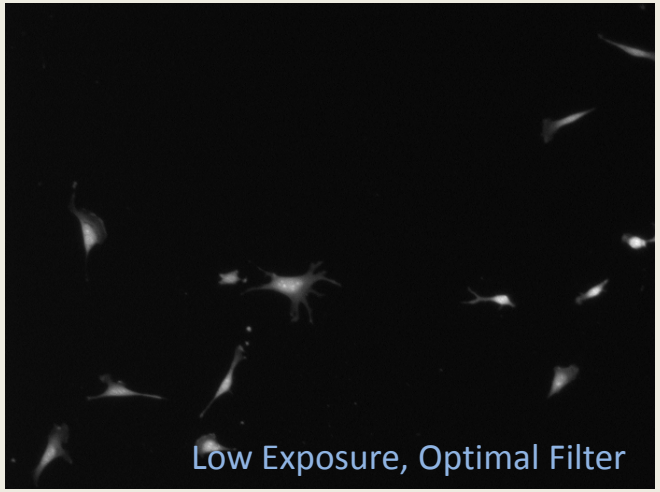


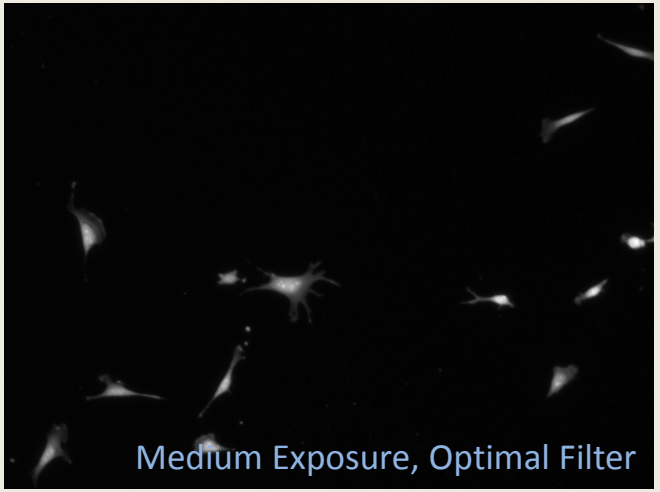


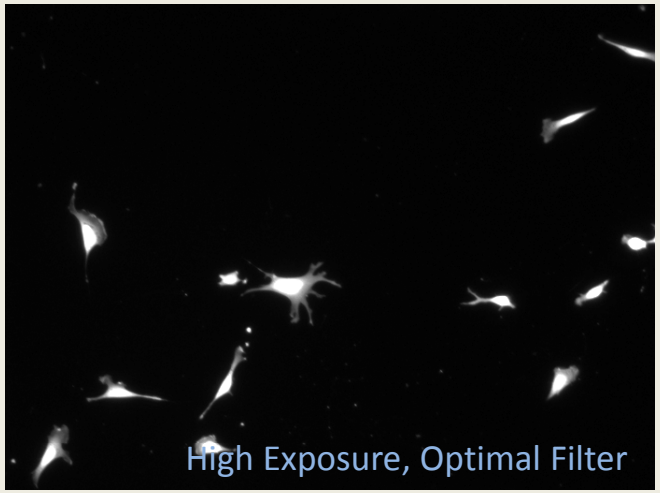


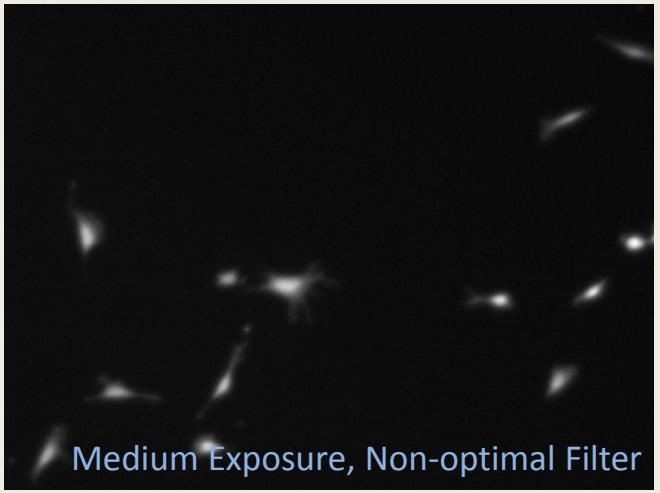
16-bit Gray-scale Images

Red Channel Images - 3T3



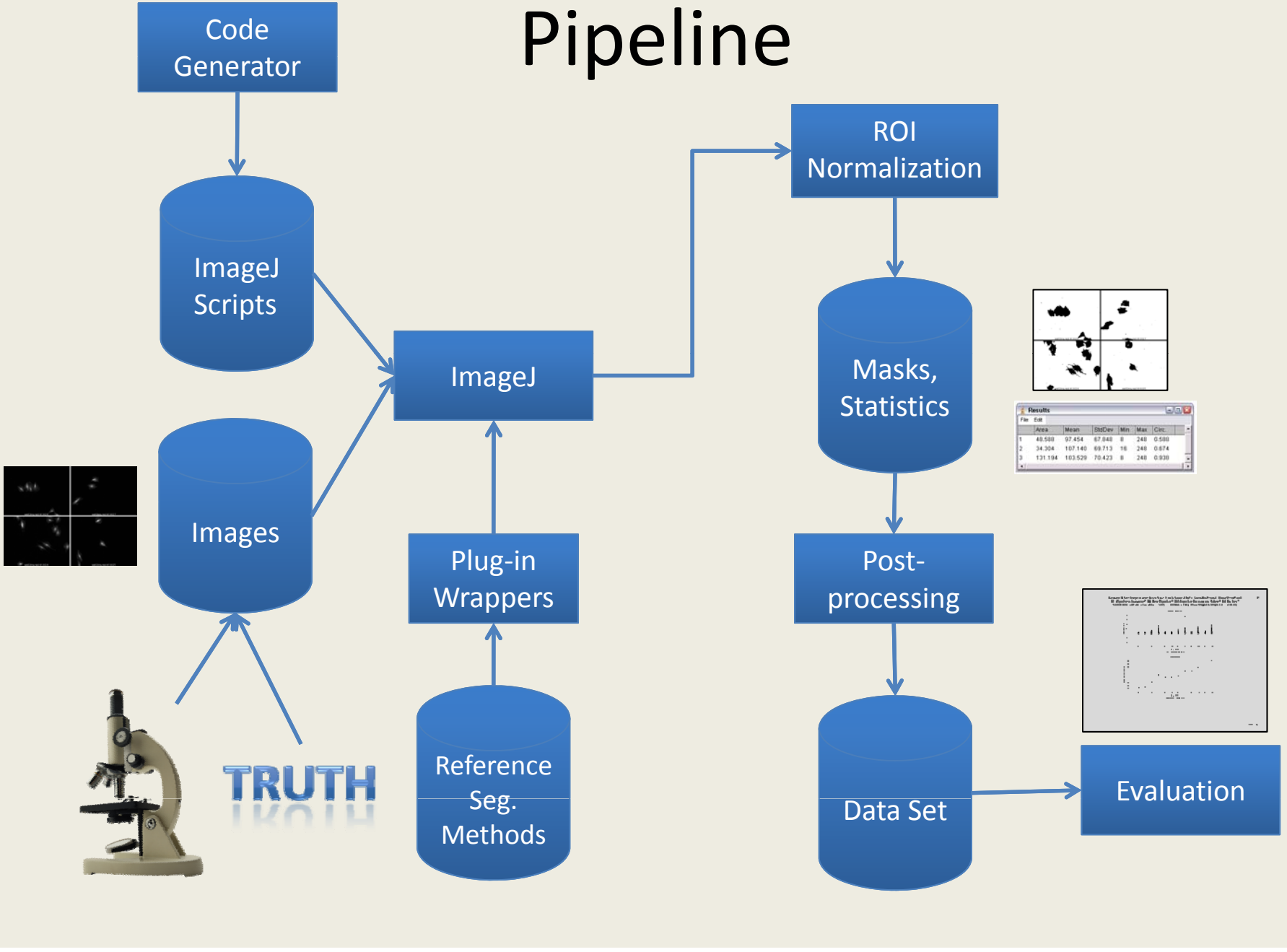




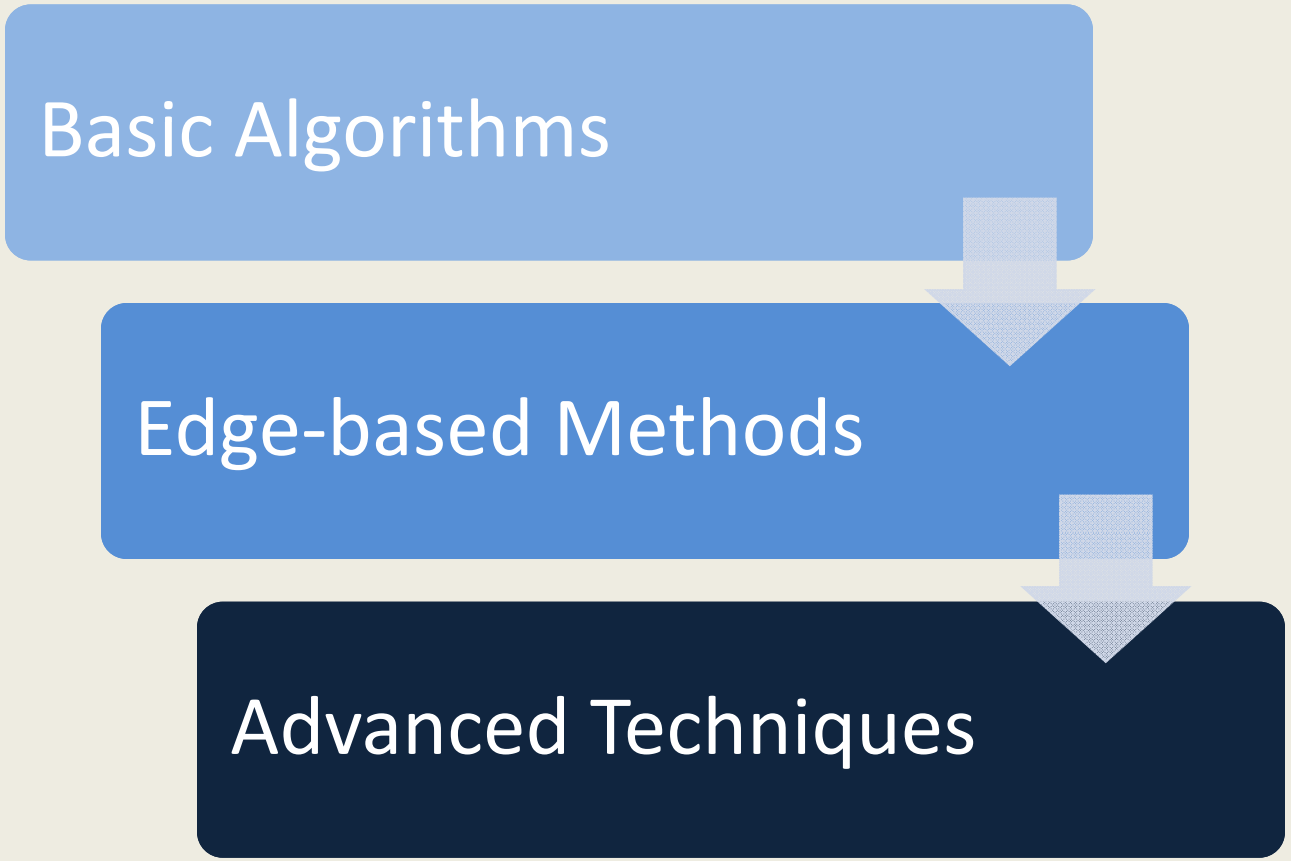


16-bit Gray-scale Images

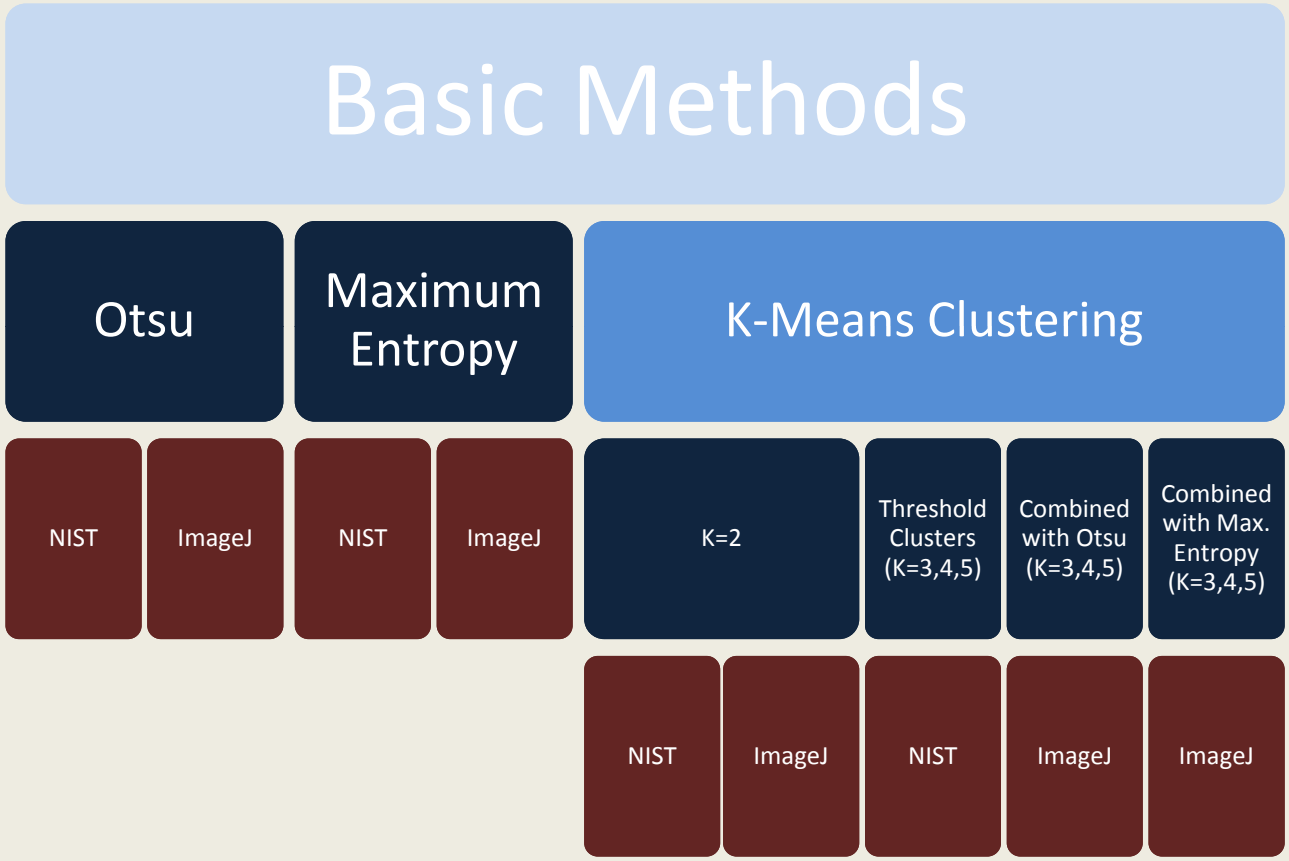
Pipeline


Code Generator
ImageJ Scripts
Images
ImageJ
Plug-in Wrappers
Reference Seg. Methods
TRUTH
ROI Normalization
Masks, Statistics
Post-processing
Data Set
Evaluation

Planned Progression of Evaluations



Algorithms Evaluated



01/07/09
1

Bio-Image Sensitivity Analysis (Campfire Project) (Brady/Dims/Plant)

Q1. What is the best algorithm?

Cell Entry = Mean Bias (Over 3 Medium-Res Image Conditions)

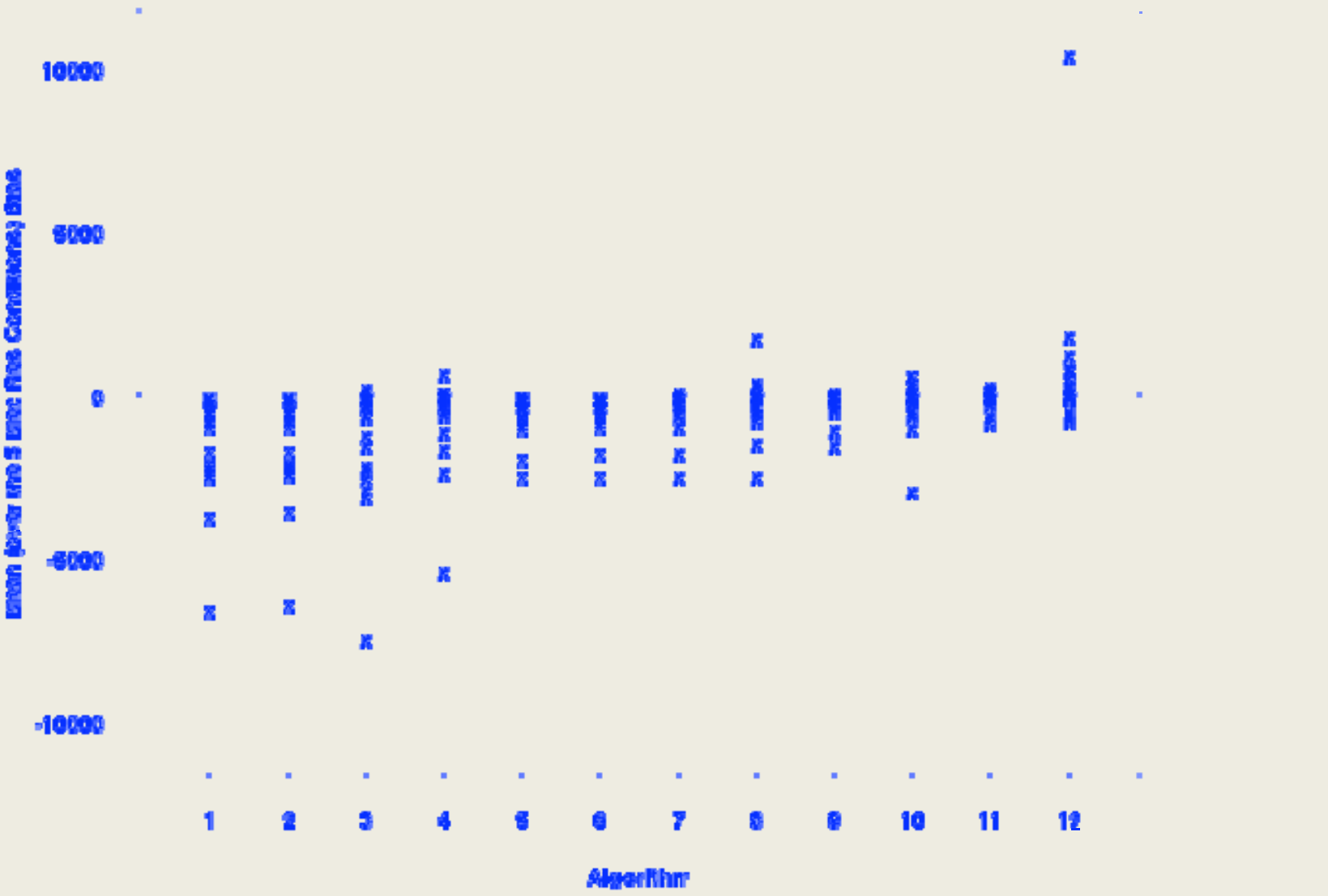
Algorithm

High Res Image ID

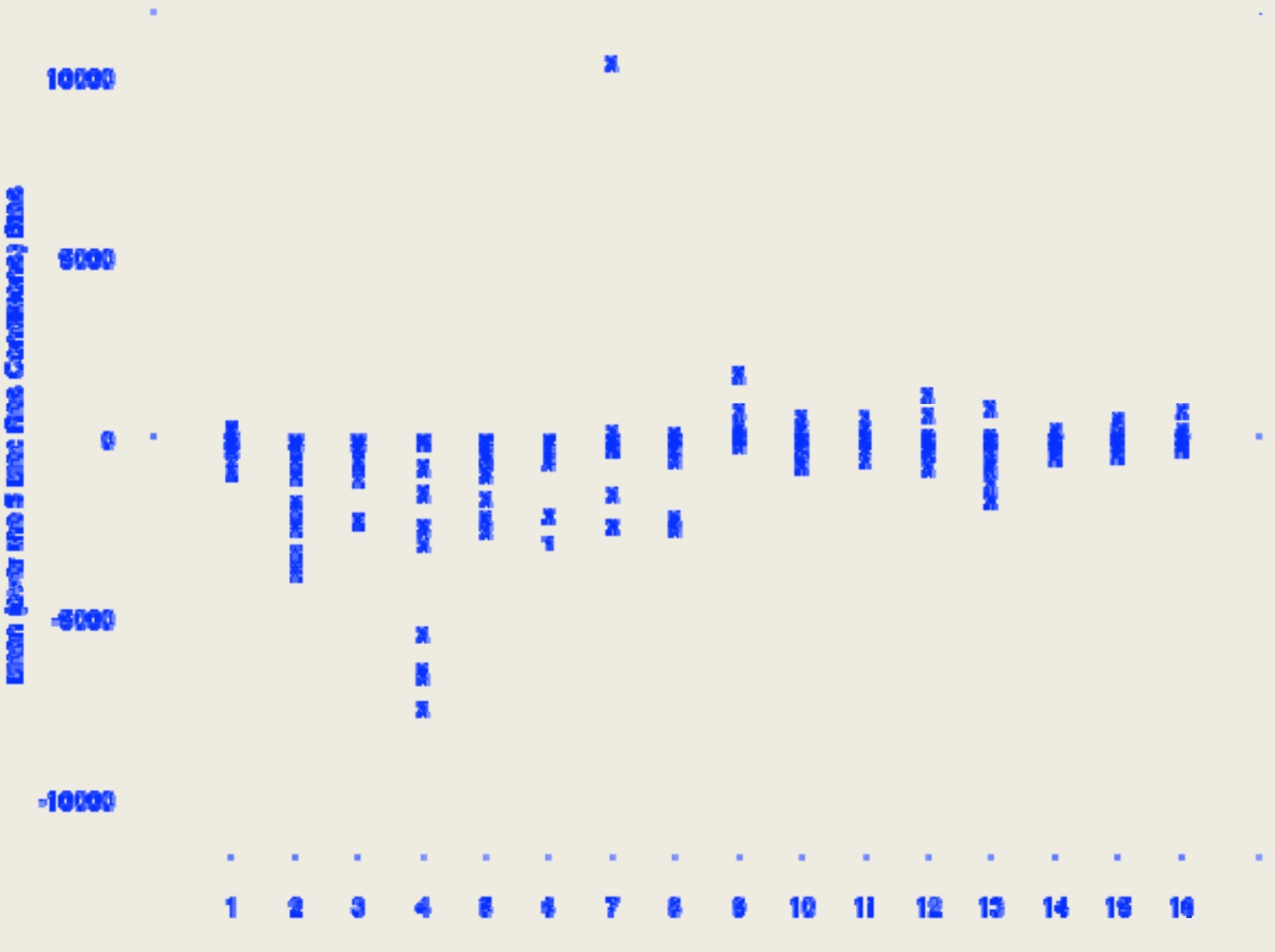
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Mean	Median	Rank
1	-907.	-809.	-378.	-164.	-244.	-223.	-229.	63	84	403	387	204	-189.	-193.	6
2	-3878.	-3319.	-3083.	-2271.	-1800.	-1728.	-1728.	-2398.	-1017.	-935.	-888.	-384.	-1880.	-1814.	15
3	-4258.	-4178.	-2231.	-1896.	-891.	-818.	-818.	-578.	-252.	-305.	-64.	-884.	-831.	-828.	12
4	-6851.	-6394.	-7483.	-3303.	-2487.	-2378.	-2378.	-1428.	-1481.	-2279.	-769.	-727.	-3348.	-2403.	16
5	-2208.	-2144.	-2448.	-1818.	-801.	-898.	-898.	-418.	-448.	-410.	-249.	-889.	-1108.	-888.	14
6	-2080.	-2040.	-2831.	-443.	-872.	-398.	-398.	-398.	-323.	-898.	-134.	-463.	-833.	-898.	11
7	-2331.	-2543.	-1488.	-461.	-258.	-233.	-233.	-252.	-131.	-160.	-6.	10808	234	-241.	8
8	-2387.	-2302.	-2143.	-872.	-348.	-340.	-340.	43	-169.	240	-64.	177	-681.	-340.	10
9	-106.	-88.	293	183	41	-6.	-6.	1738	140	184	228	1944	272	192	5
10	-721.	-700.	-531.	-219.	-251.	-242.	-242.	431	21	683	232	248	-109.	-231.	7
11	-848.	-831.	-303.	161	-229.	-234.	77	219	-82.	889	129	294	-6.	27	1
12	-748.	-737.	-208.	728	-378.	-377.	-377.	-280.	24	-189.	187	1257	-81.	-244.	9
13	-1877.	-1898.	-1181.	-184.	-842.	-767.	198	-717.	-862.	-814.	-84.	819	-829.	-825.	13
14	-486.	-473.	-331.	-113.	-147.	-141.	-141.	-70.	16	20	137	314	-118.	-128.	4
15	-443.	-430.	-323.	-83.	-168.	-162.	-129.	89	98	478	149	898	-97.	-108.	3
16	-283.	-250.	-37.	143	-114.	-112.	-112.	-67.	16	333	139	848	41	-67.	2
Mean	-1713.	-1683.	-1647.	-683.	-567.	-583.	-486.	-241.	-241.	-188.	-27.	838			
Median	-1298.	-1254.	-483.	-207.	-238.	-231.	-238.	-174.	-77.	-89.	84	290			
Rank	12	11	10	6	9	8	5	4	3	2	1	7			

dims44.de

Bio-Image Sensitivity Analysis (Compile Project) [Brady/Dima/Plant]
Q. Best & Worst Algorithms?



Bio-Image Sensitivity Analysis (Compile Project) [Brady/Dima/Plant]
Q. Easiest and Hardest Images?

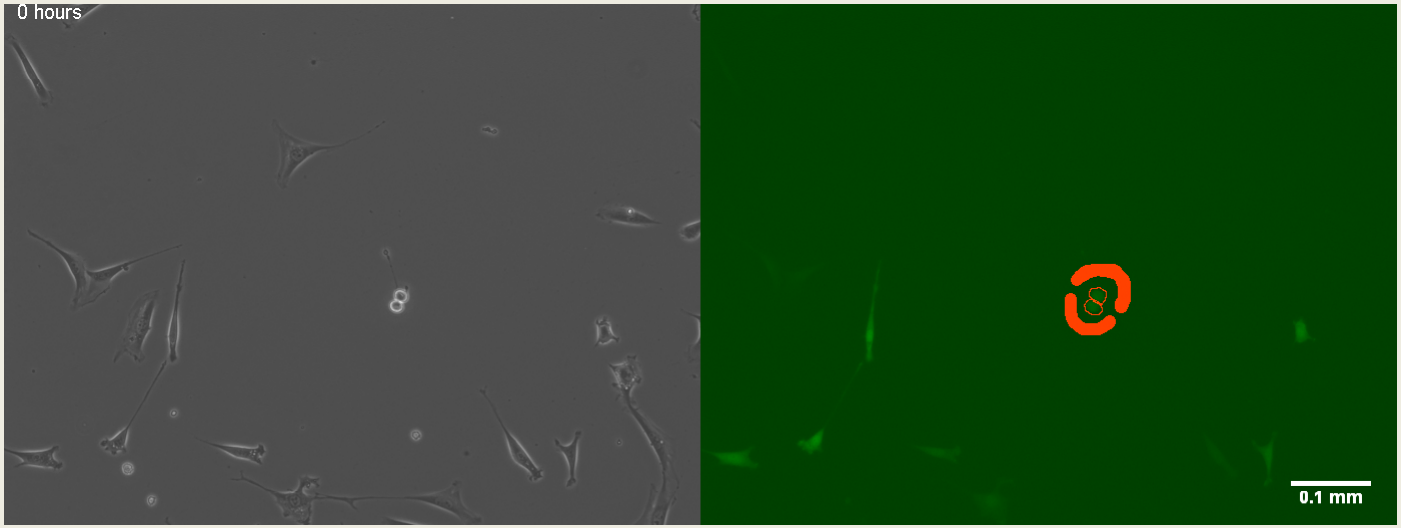


Mean	-199.	-1980.	-631.	-3348.	-1102.	-813.	234	-951.	372	-109.	-6.	-91.	-852.	-119.	-27.	41
Median	-183.	-1814.	-629.	-3403.	-888.	-816.	-841.	-340.	182	-231.	27	-244.	-926.	-199.	-106.	-67.
Rank	6	16	12	15	14	11	8	10	8	7	1	9	13	4	3	2

Live Cell Tracking

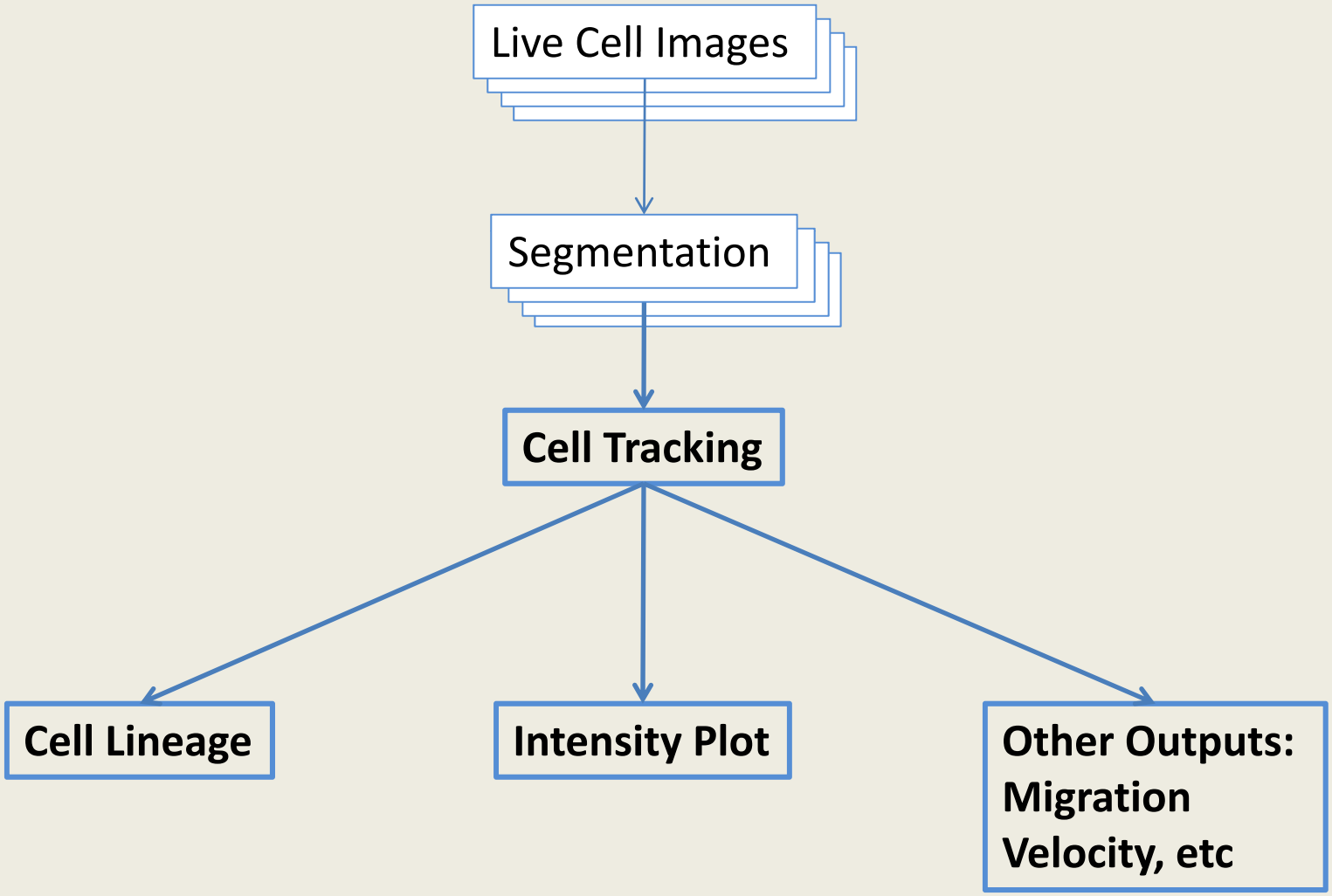
- There are few automated image analysis options available to the cell biologists to quantify live cell image data
- Segment and track cells in an image sequence to quantify the total fluorescence intensity of individual cells over time

Manually Tracking NIH 3T3 Cells

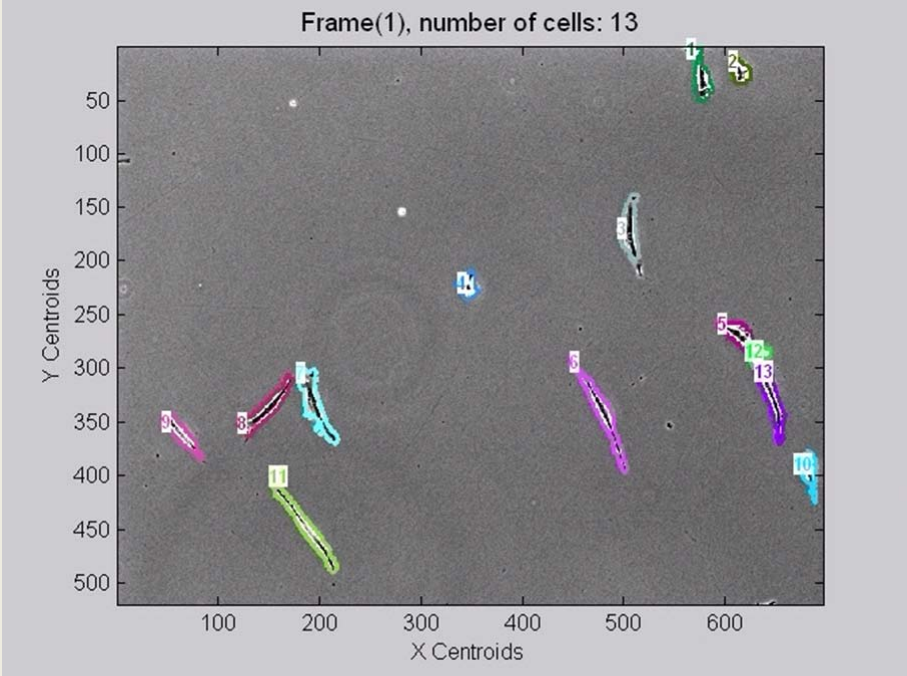


Phase contrast on left, GFP on right

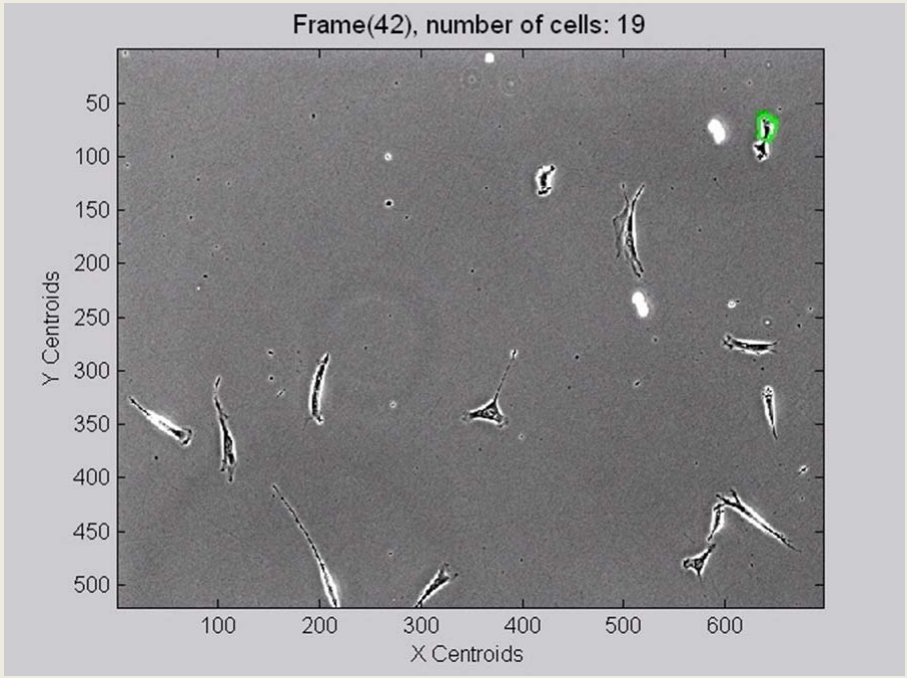
CompBio Cell Tracker



Automatic Tracking of NIH 3T3 Cells



Automatic Tracking of NIH 3T3 Cells



Thank You!

alden.dima@nist.gov

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce