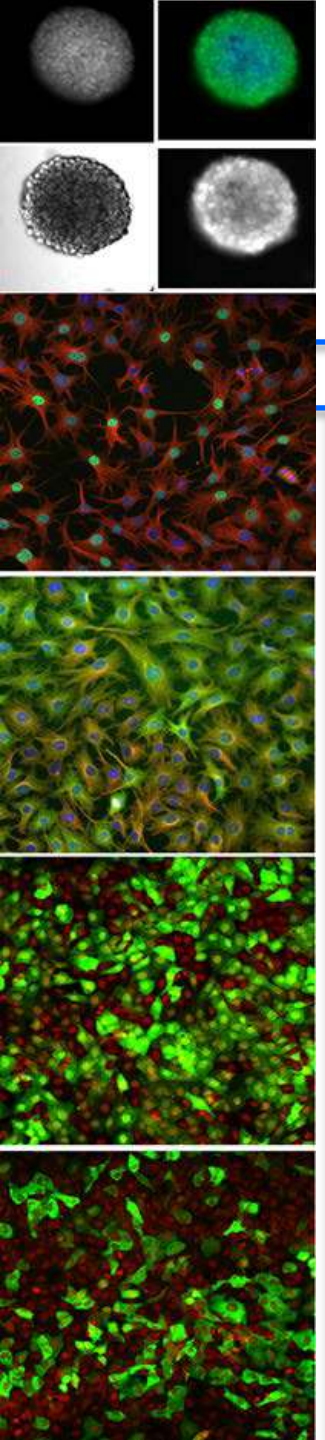


SOCIETY OF BIOMOLECULAR IMAGING AND INFORMATICS

High Content 2016

September 12th-14th

3rd Annual Conference

Joseph B. Martin Conference Center at Harvard Medical School, Boston, MA

Advanced Tools for Data Analysis: CellProfiler

Mark-Anthony Bray, Ph.D
Novartis Institutes of Biomedical Research
Cambridge, Massachusetts, USA
mark.bray@novartis.com

START HERE: https://github.com/braymp/sbi2

Overview and Requisites

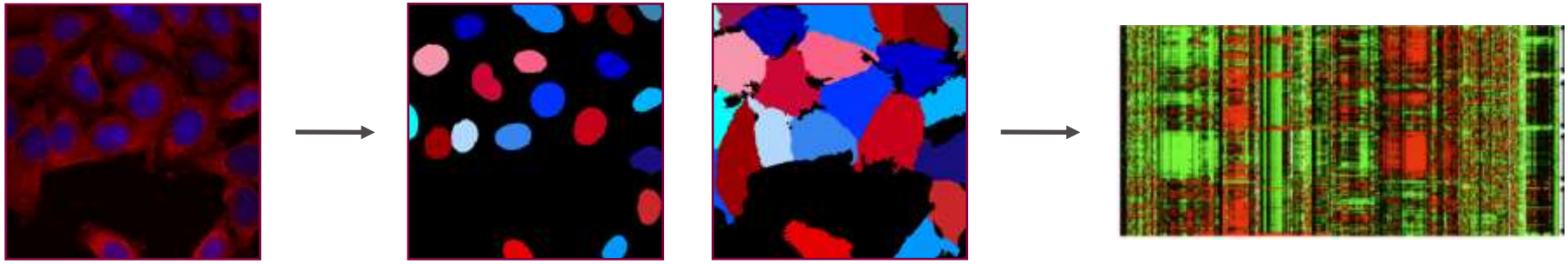
- Agenda: Hands-on demo using **CellProfiler** for high-content analysis

- This session assumes that you either:
  - Attended the Introductory HCS image analysis session
  - Have a good working knowledge of the basics of image analysis

CellProfiler: Overview

CellProfiler™
cell image analysis software

- **Process** large sets of images
- **Identifies and measures** objects
- **Export** data for further analysis

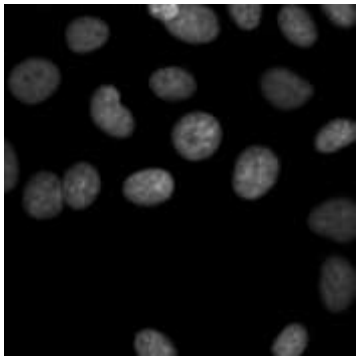
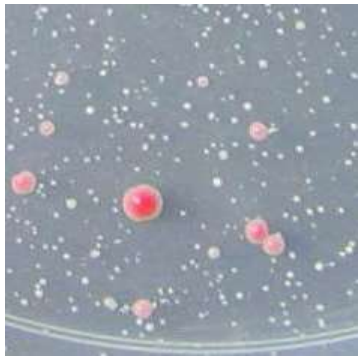


Goal: Provide powerful image analysis methods with a user-friendly interface

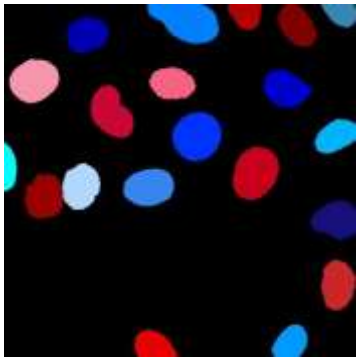
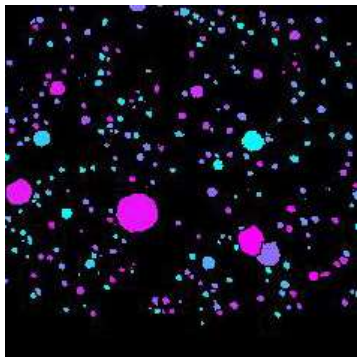
Philosophy: Measure everything, ask questions later

Typical CellProfiler Pipeline Workflow

- For image-based assays, the basic objective is always to
  - Identify cells/organisms
  - Measure feature(s) of interest

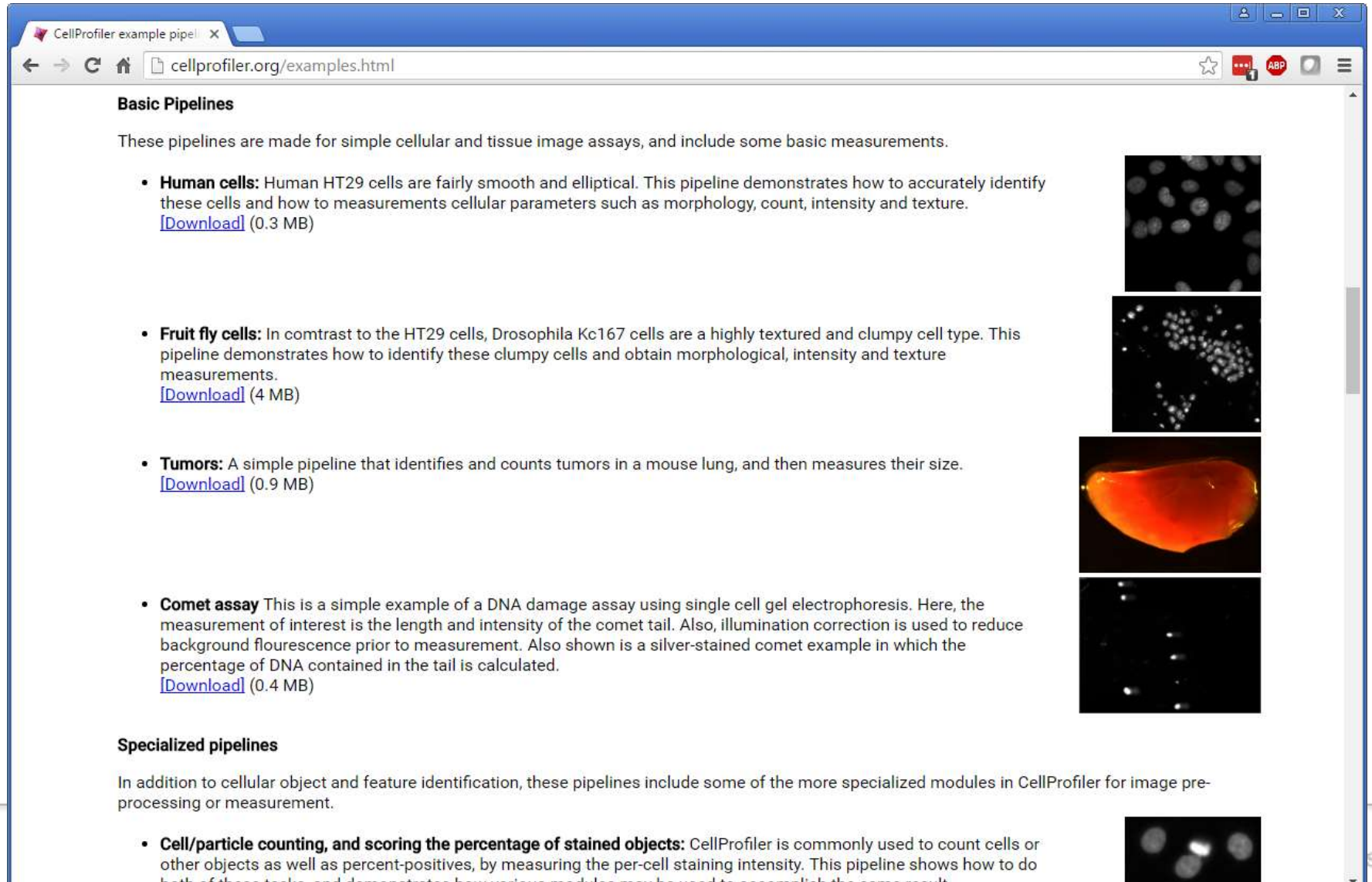



- The uniqueness of each assay comes in
  - Deciding what compartments to identify and how to identify them
  - Determining which measure(s) are most useful to identify interesting samples

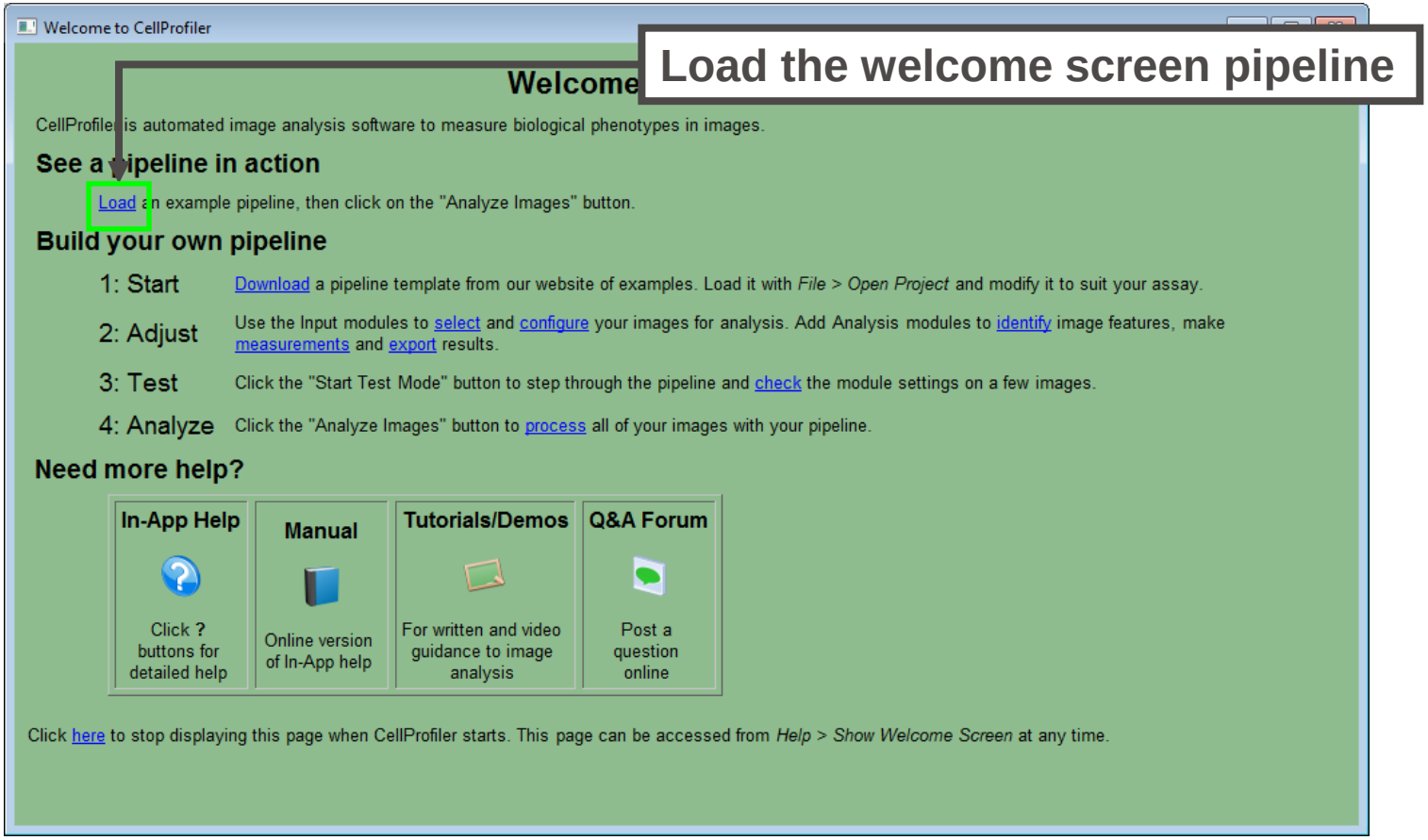
Typical Starting Point: An Existing Pipeline

Go to www.cellprofiler.org/examples.html for many examples

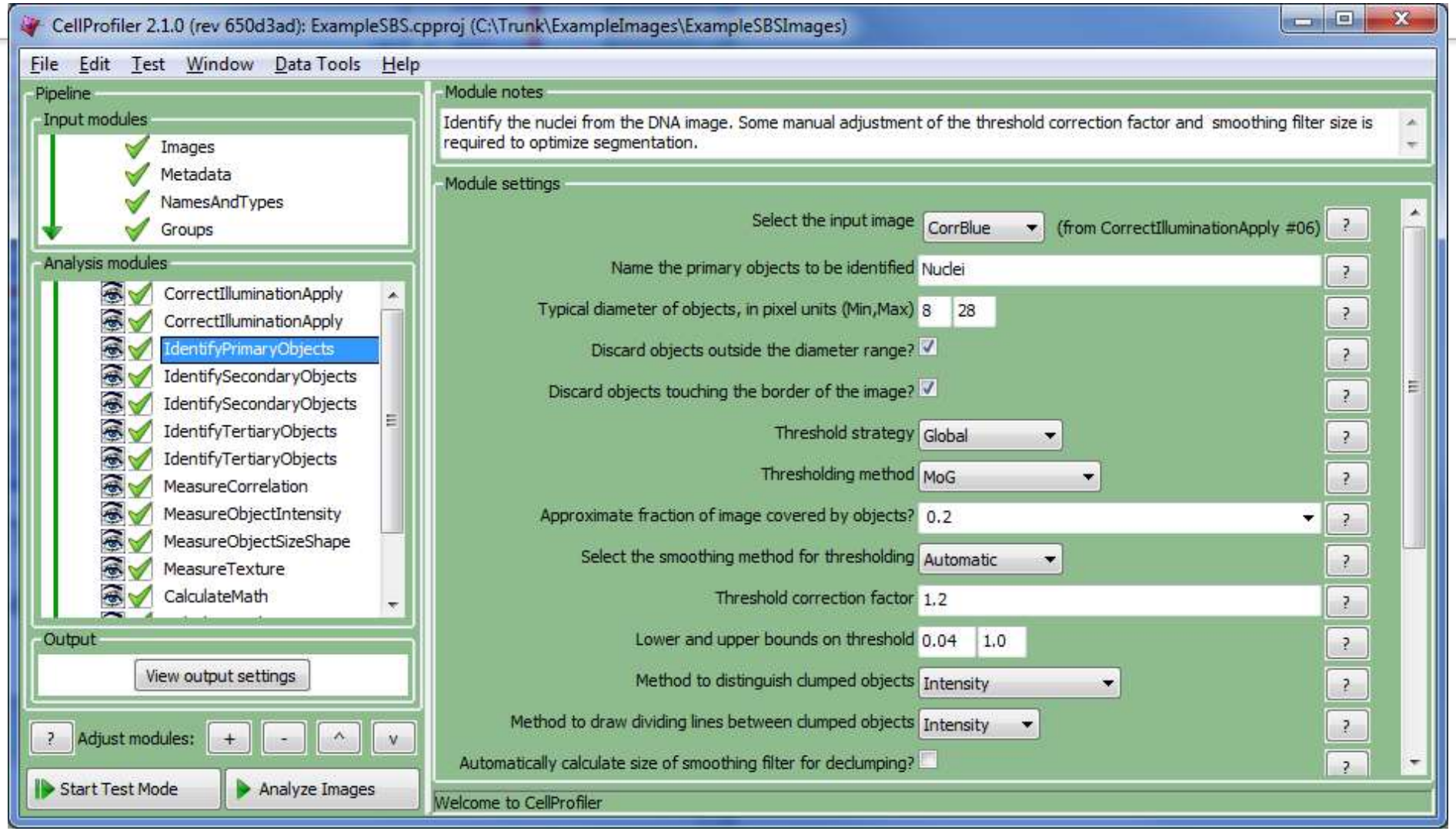


Typical Starting Point: An Existing Pipeline

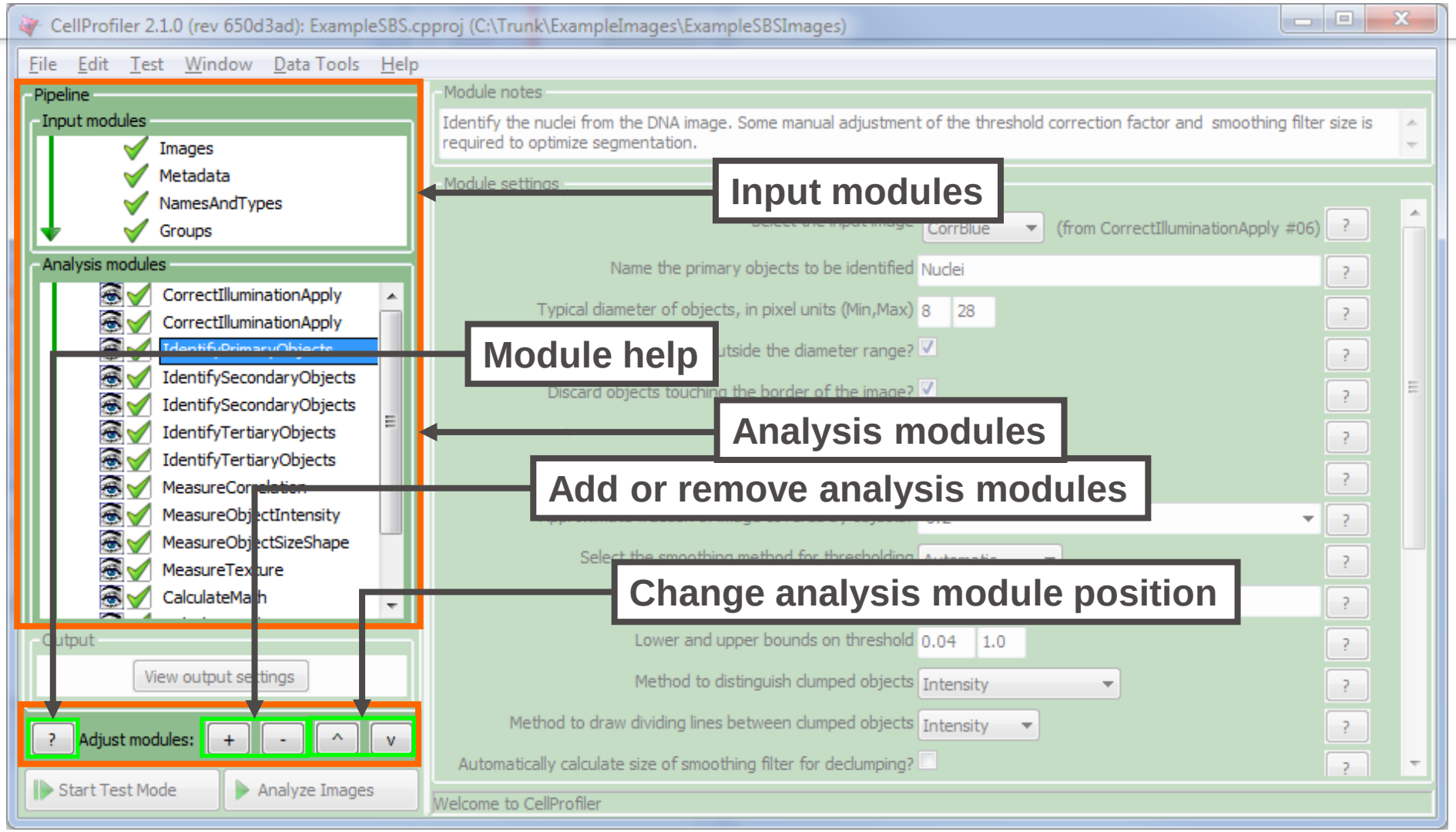
- Open CellProfiler



The CellProfiler Interface

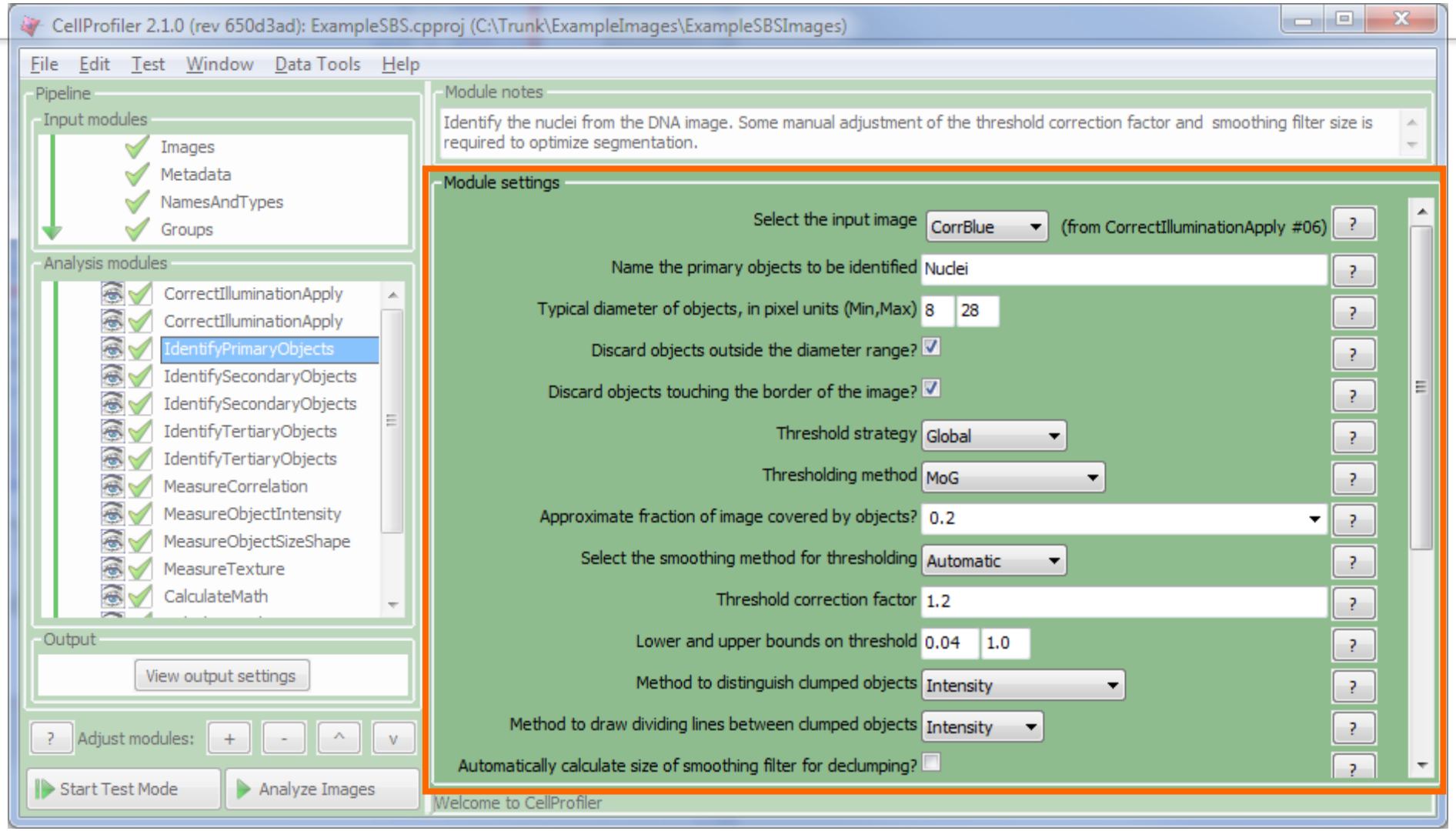


The CellProfiler Interface



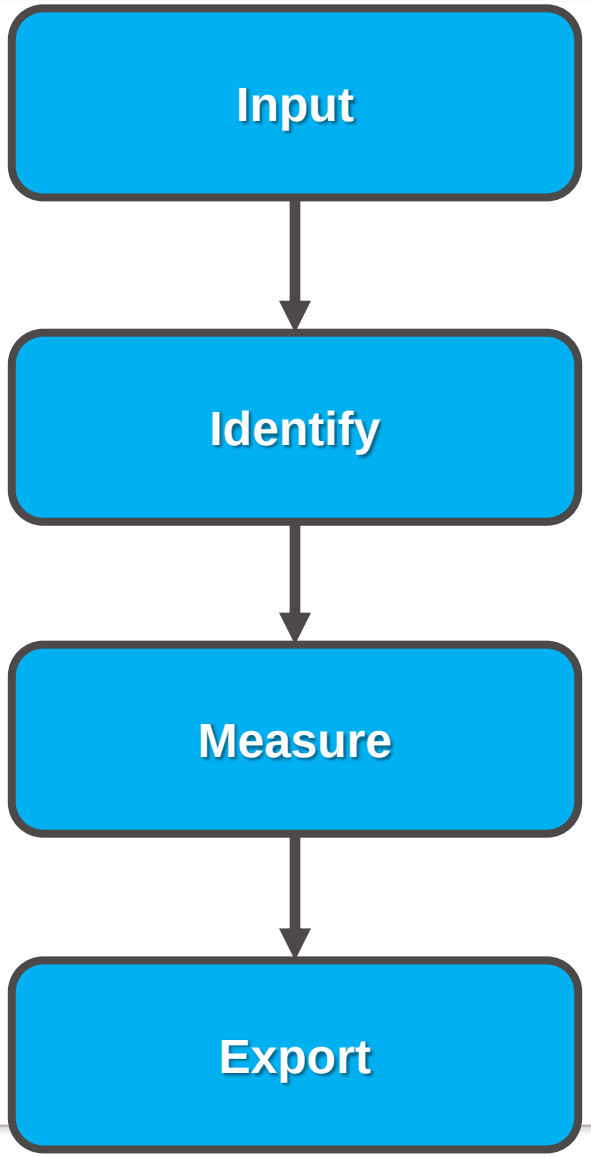
- **Pipeline panel:** Displays modules in pipeline
  - Modules executed in order from top to bottom, extensive help for each

The CellProfiler Interface



- **Settings panel:** View and change settings for each module
  - Clicking on a different module updates the settings view

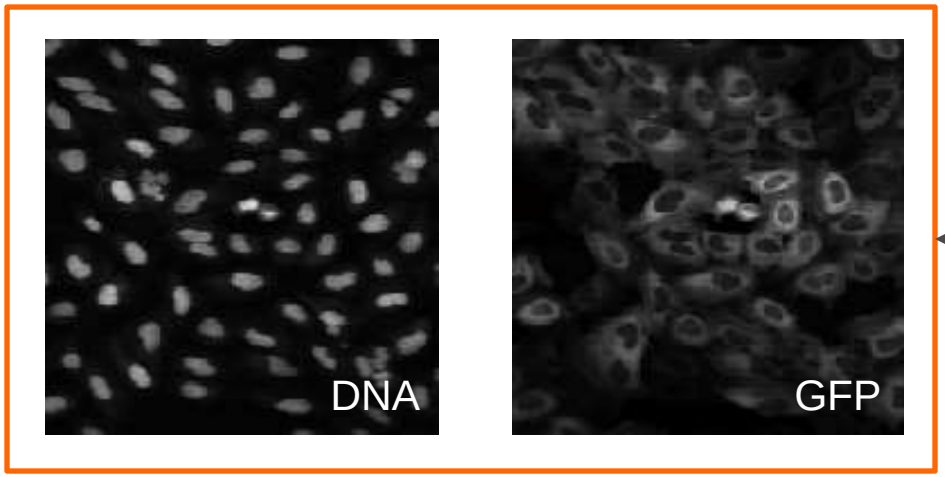
Typical CellProfiler Workflow



Configure:	Based on:
Images	1. File locations 2. Subsets?
Metadata	1. Plate-based?
Channels	1. Mnemonic 2. Matching
Grouping	1. Movies? 2. 3D stacks?

Creating A CellProfiler Project

- Use the **Input** modules to create an *image set*



An *image set*

- Most commonly defined as 1 or more channels at one imaging location
- Other definitions depend on the assay

- Add **Analysis** modules to pre-process the images, identify objects, make measurements

- Add **Export** modules to write out images/measurements to disk or database

Loading and Configuring Your Images

- Browse to the files extracted from the ZIP file you downloaded from GitHub

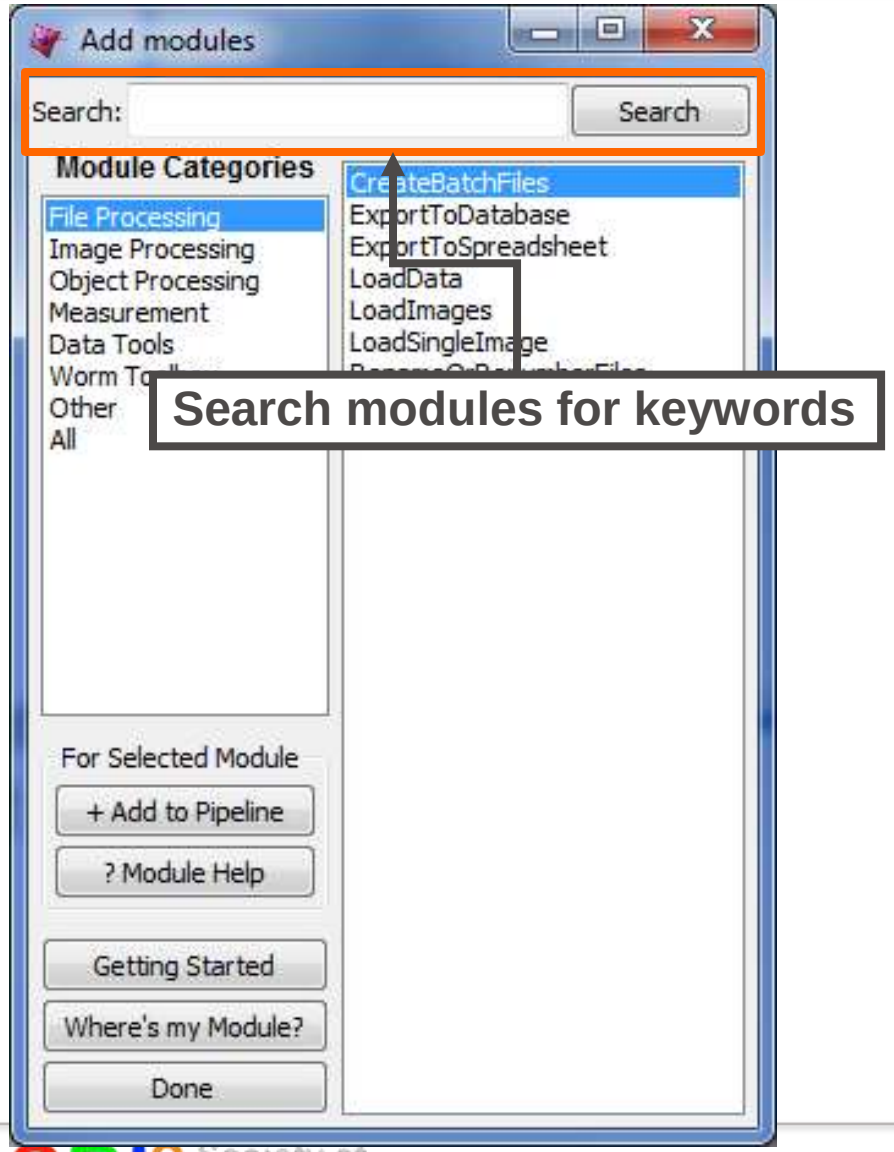
- **Images** module
  - Right-click on the file list panel; select “Clear file list” from context menu
  - Drag-and-drop the “images” folder into the file list panel module

- **NamesAndTypes** module: Make the following changes
  - **OrigBlue**: [File] [Does] [Contain] ”**d0.png**”
  - **OrigGreen**: [File] [Does] [Contain] ”**d1.png**”
  - **OrigRed**: [File] [Does] [Contain] ”**d2.png**”

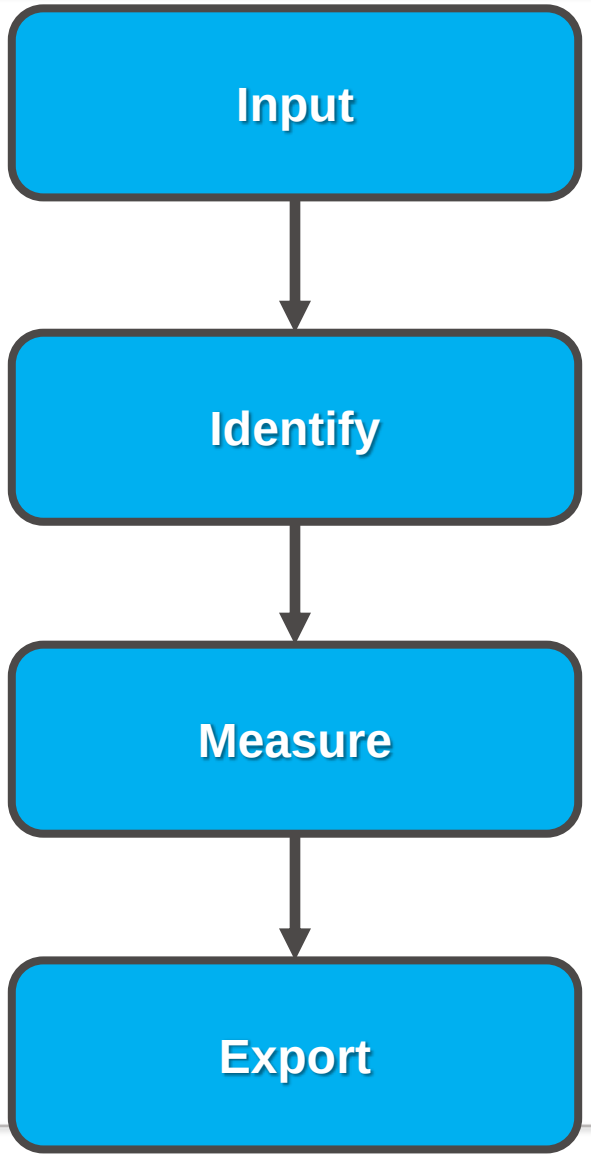
- **Crop** module: Adjust the settings
  - Left and right rectangle positions:
  - Top and bottom rectangle positions:

Module Categories

- **File processing:** Image input, file output
- **Image processing:** Often used for pre-processing prior to object identification
- **Object processing:** Identification, modification of objects of interest
- **Measurement:** Collection of measurements from objects of interest
- **Data Tools:** Measurement exploration, measurement output
- **Worm Toolbox:** *C. elegans*-specific operations

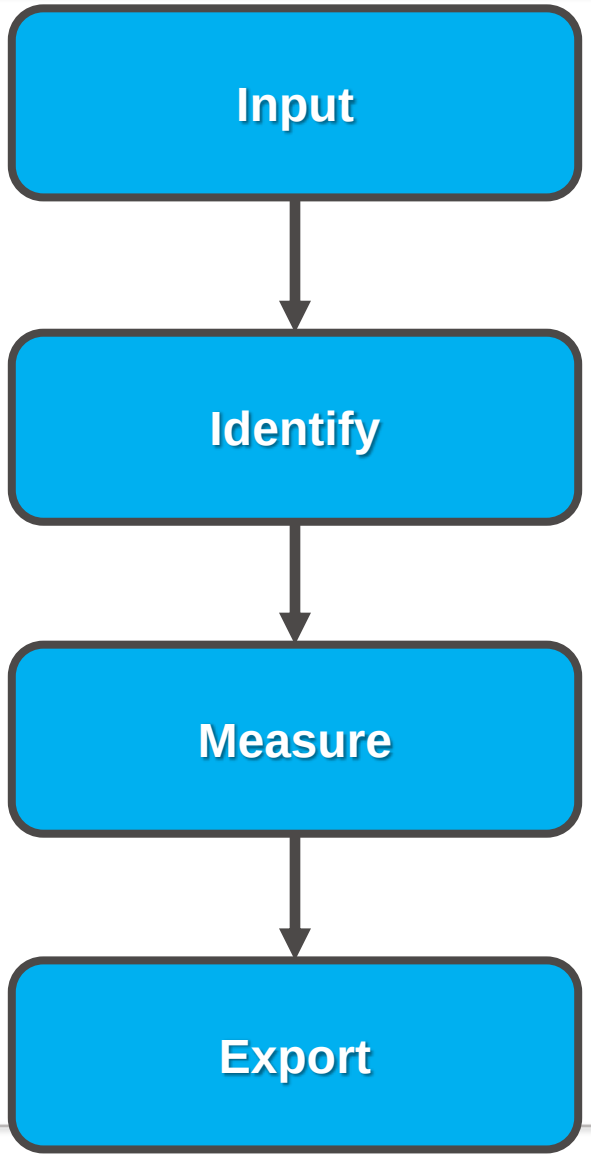


Typical CellProfiler Workflow



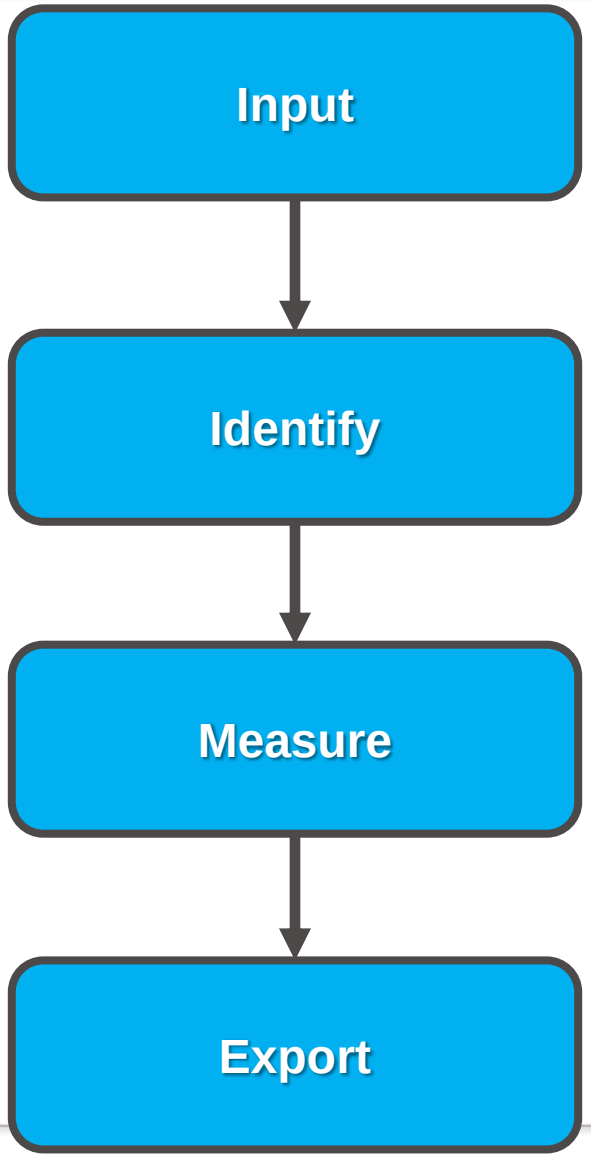
Configure:	Based on:
Images	1. File locations 2. Subsets?
Metadata	1. Plate-based?
Channels	1. Mnemonic 2. Matching
Grouping	1. Movies? 2. 3D stacks?

Typical CellProfiler Workflow



Configure:	Based on:
Images	1. File locations 2. Subsets?
Metadata	1. Plate-based?
Channels	1. Mnemonic 2. Matching
Grouping	1. Movies? 2. 3D stacks?

Typical CellProfiler Workflow

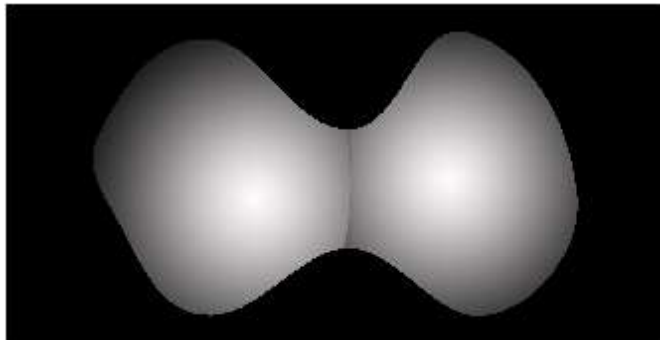


Configure:	Based on:
Images	1. File locations 2. Subsets?
Metadata	1. Plate-based?
Channels	1. Mnemonic 2. Matching
Grouping	1. Movies? 2. 3D stacks?

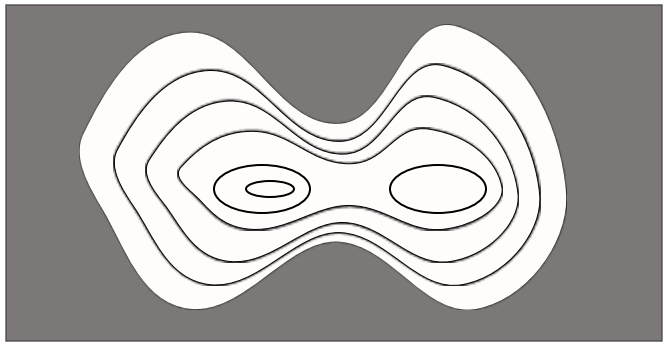
Object Identification

- Once the images are loaded, how do you find objects of interest?

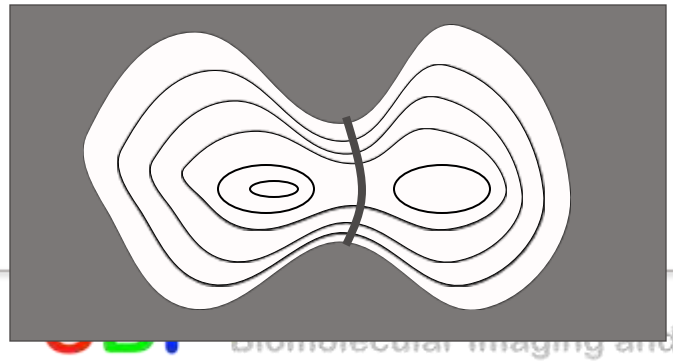
Step 1: Distinguish the foreground from the background by picking a good threshold



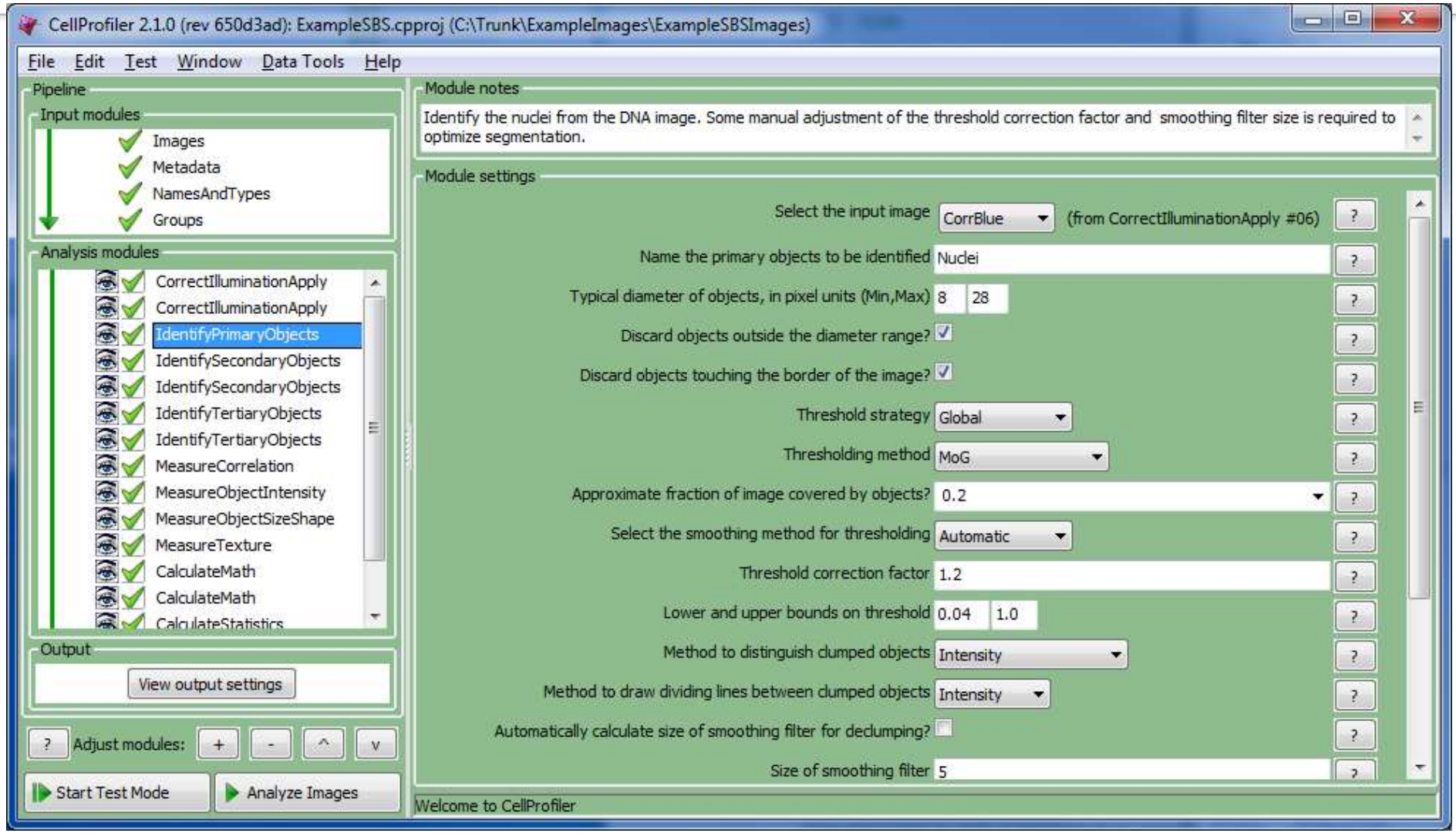
Step 2: Identify objects as regions brighter than the threshold



Step 3: Cut and join objects to “improve” their shape

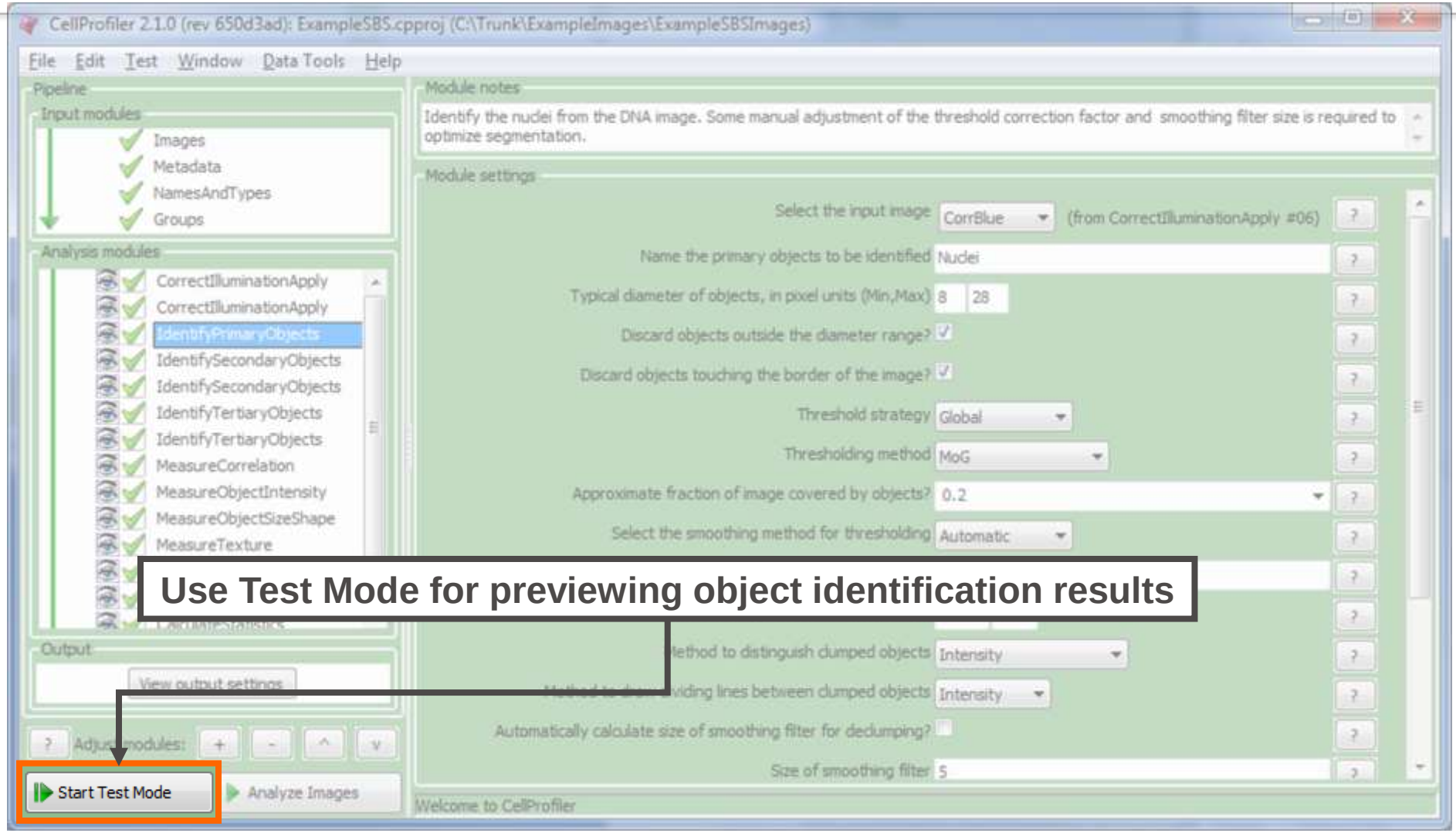


Primary Object Identification

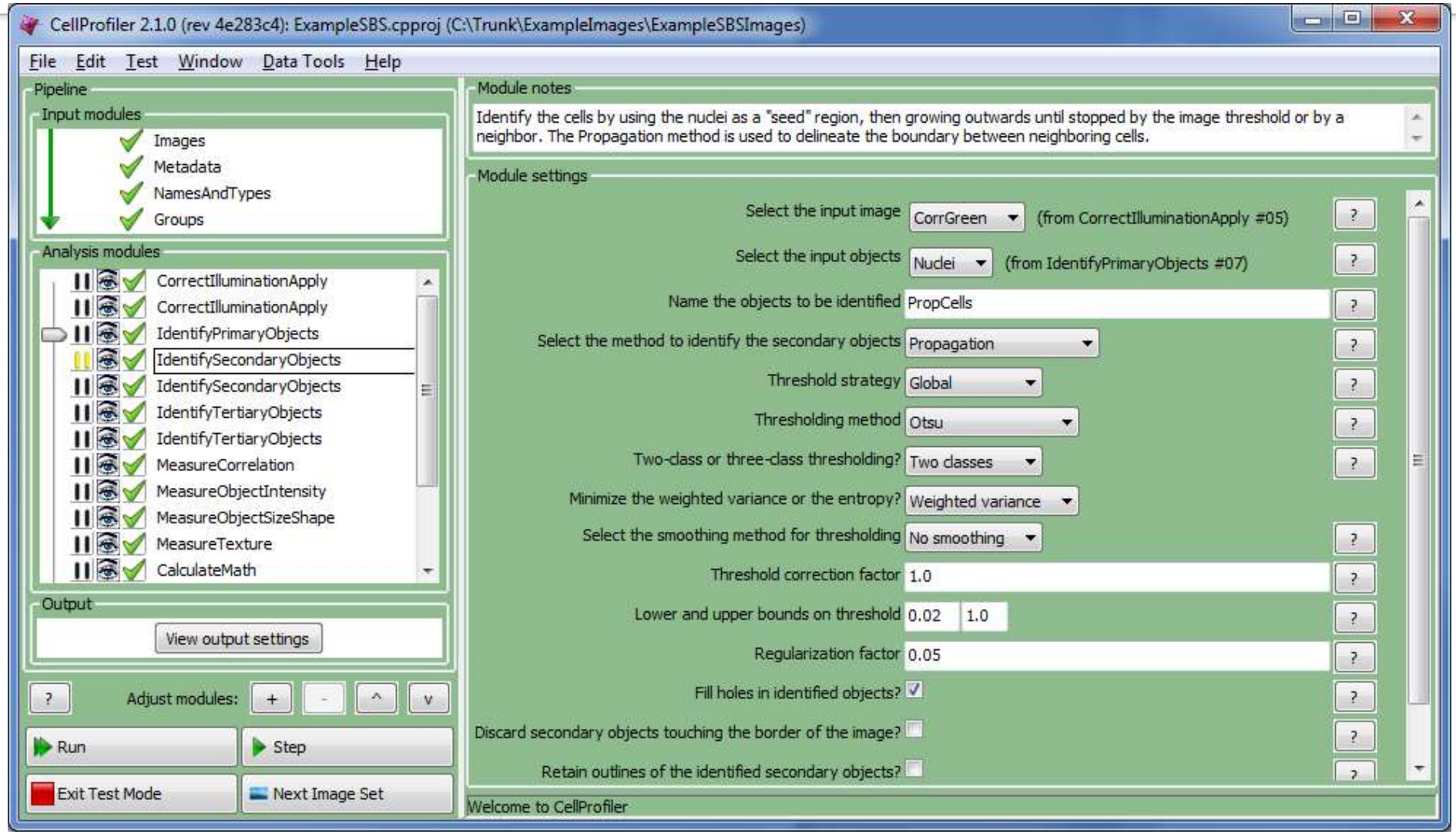


- Many options for thresholding, cut and join methods, etc.

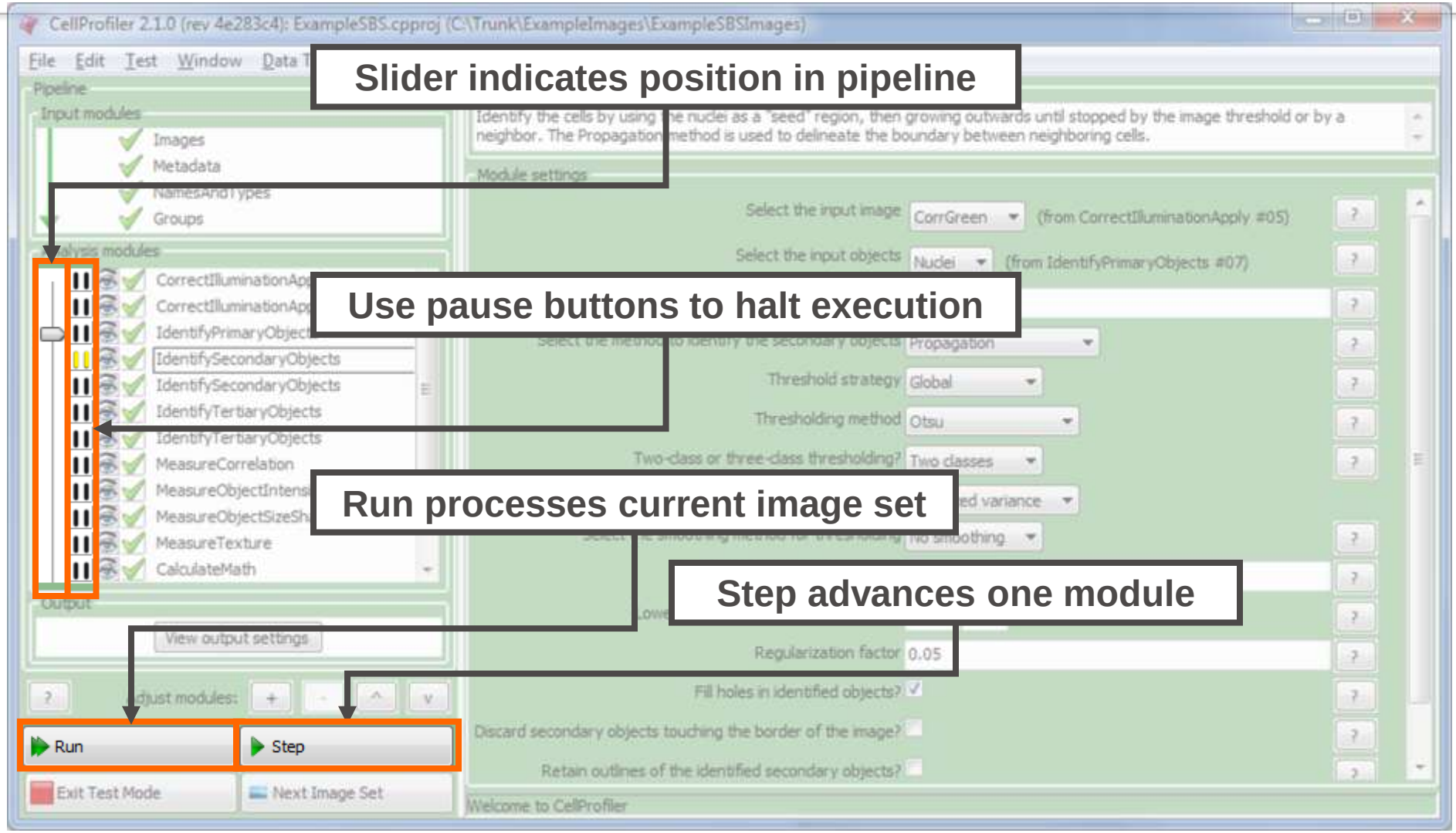
Primary Object Identification



Test Mode

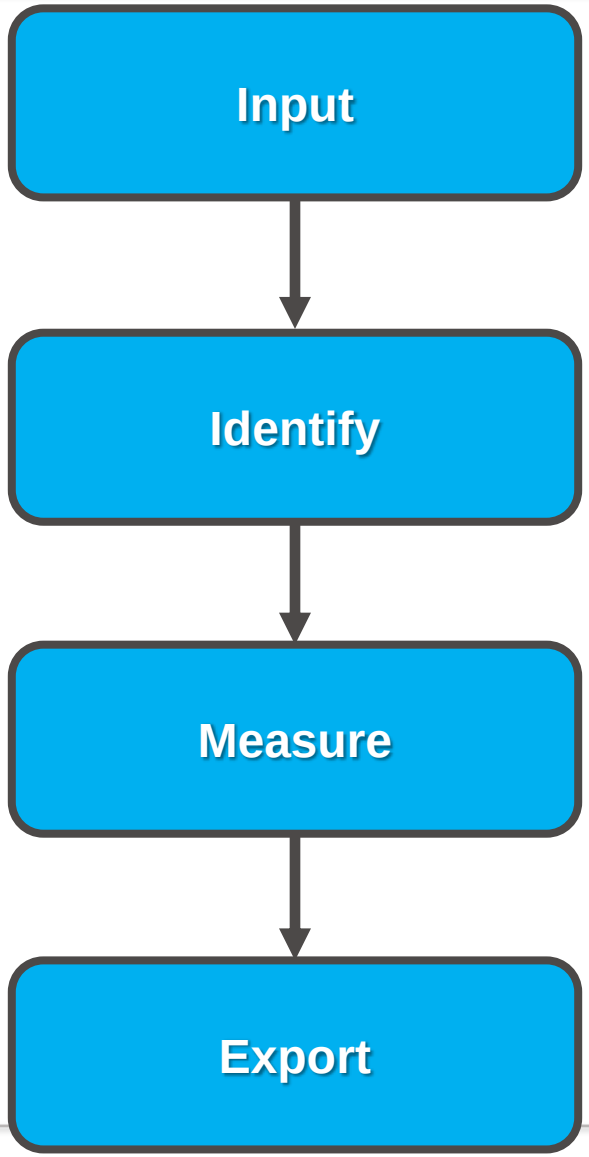

CellProfiler 2.1.0 (rev 4e283c4): ExampleSBS.cpproj (C:\Trunk\ExampleImages\ExampleSBSImages)
File Edit Test Window Data Tools Help
Pipeline
Input modules
Images
Metadata
NamesAndTypes
Groups
Analysis modules
CorrectIlluminationApply
CorrectIlluminationApply
IdentifyPrimaryObjects
IdentifySecondaryObjects
IdentifySecondaryObjects
IdentifyTertiaryObjects
IdentifyTertiaryObjects
MeasureCorrelation
MeasureObjectIntensity
MeasureObjectSizeShape
MeasureTexture
CalculateMath
Output
View output settings
Adjust modules:
Run
Step
Exit Test Mode
Next Image Set
Module notes
Identify the cells by using the nuclei as a "seed" region, then growing outwards until stopped by the image threshold or by a neighbor. The Propagation method is used to delineate the boundary between neighboring cells.
Module settings
Select the input image CorrGreen (from CorrectIlluminationApply #05)
Select the input objects Nuclei (from IdentifyPrimaryObjects #07)
Name the objects to be identified PropCells
Select the method to identify the secondary objects Propagation
Threshold strategy Global
Thresholding method Otsu
Two-class or three-class thresholding? Two classes
Minimize the weighted variance or the entropy? Weighted variance
Select the smoothing method for thresholding No smoothing
Threshold correction factor 1.0
Lower and upper bounds on threshold 0.02 1.0
Regularization factor 0.05
Fill holes in identified objects?
Discard secondary objects touching the border of the image?
Retain outlines of the identified secondary objects?
Welcome to CellProfiler

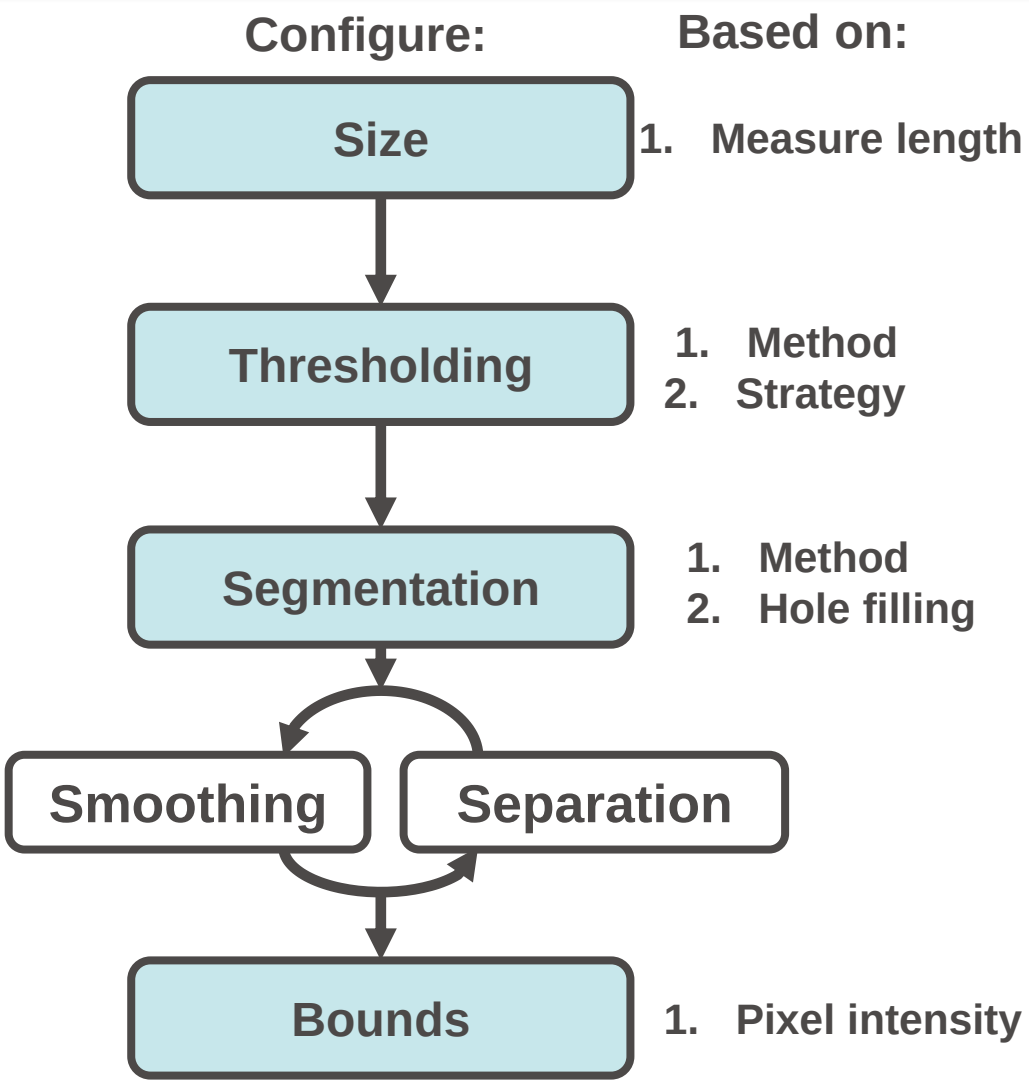
Test Mode



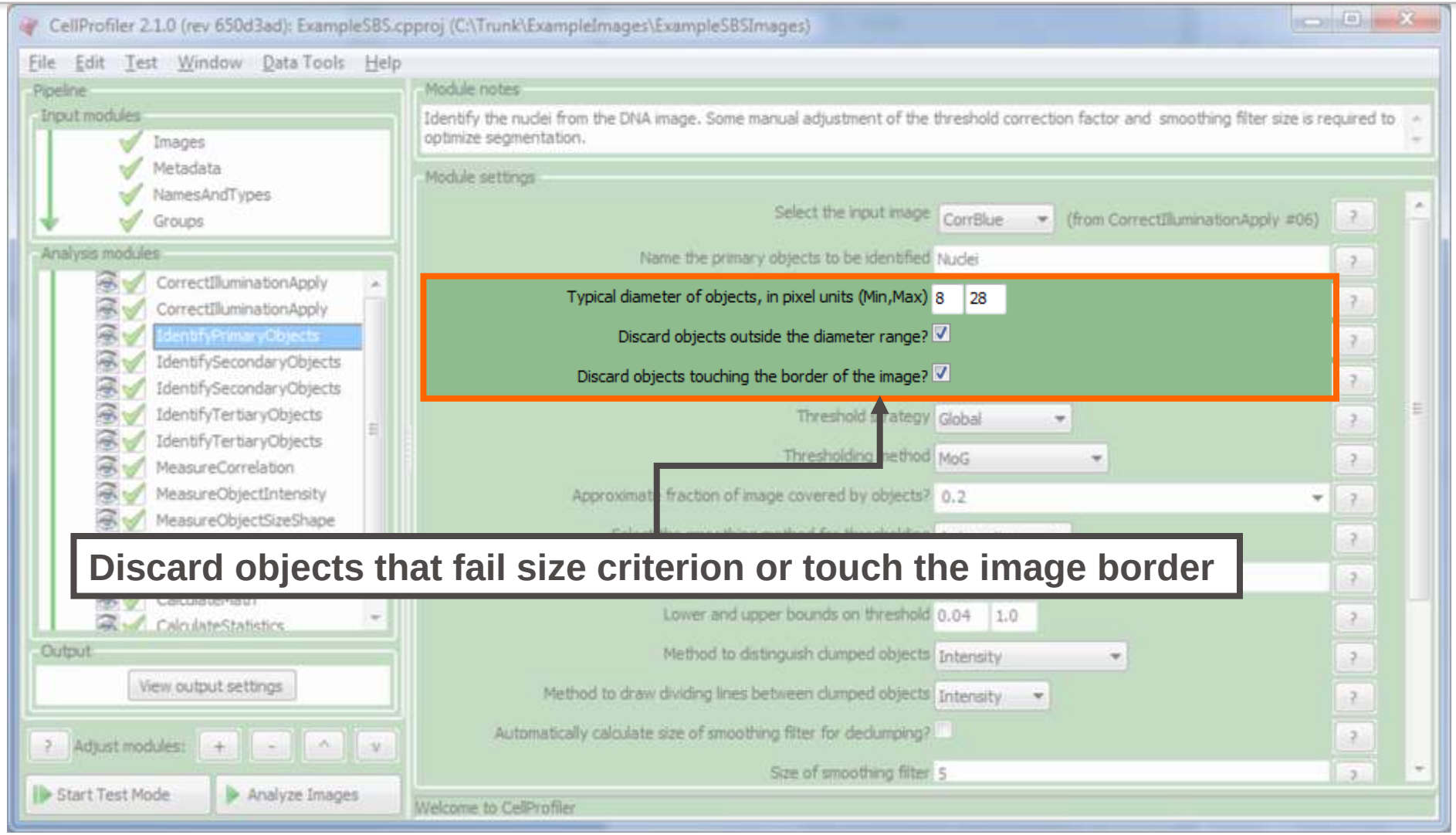
- Use the Test menu item for more options

Typical CellProfiler Workflow





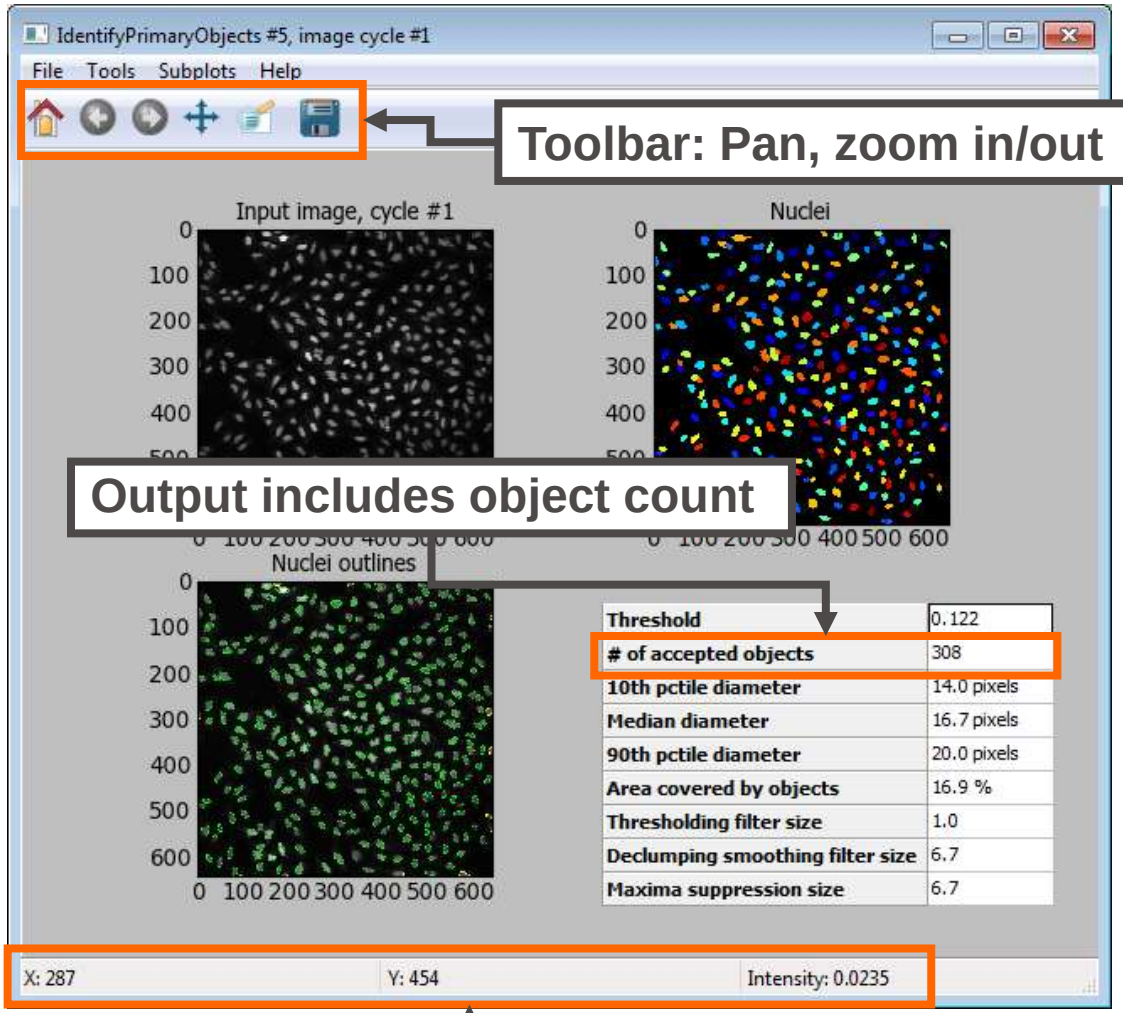
Filtering Invalid Objects



- See FilterObjects module for more advanced filtering options

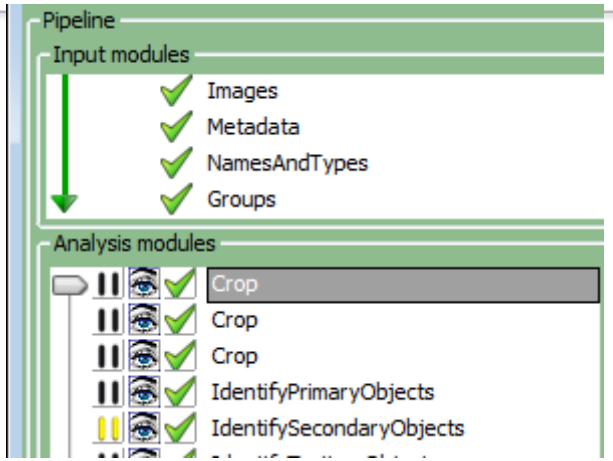
Primary Object Identification

- Segmented objects are colored
  - Shows if each object has been identified and separated properly
- Outlines: Valid objects
  - **Green:** Valid
  - **Yellow:** Invalid – Touching border
  - **Red:** Invalid – Size criterion



Primary Object Identification

- Press the Test mode button
- Adjust **Crop** module settings
  - Left and right rectangle positions: 250, end
  - Top and bottom rectangle positions: 1, 250
- Set a pause after **IdentifyPrimaryObjects**, run the pipeline



- Experiment with the exclusion settings
  - Measure a few nuclei with the measuring tool
  - Toggle the border touching criteria, re-run the module
  - Toggle the size exclusion criteria, re-run the module

- What min/max diameter setting would be most appropriate in this case?

Typical CellProfiler Workflow

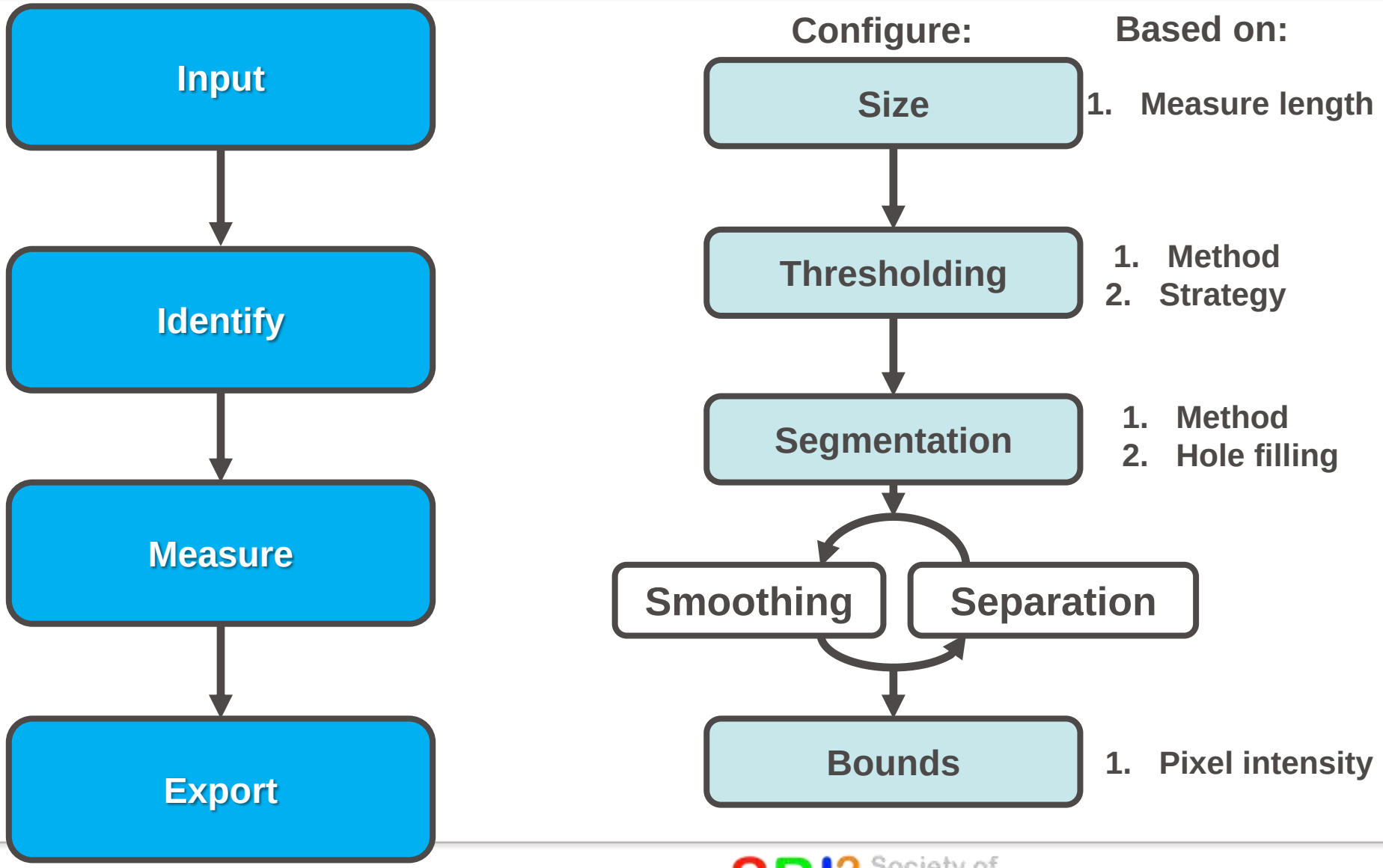


Image Thresholding

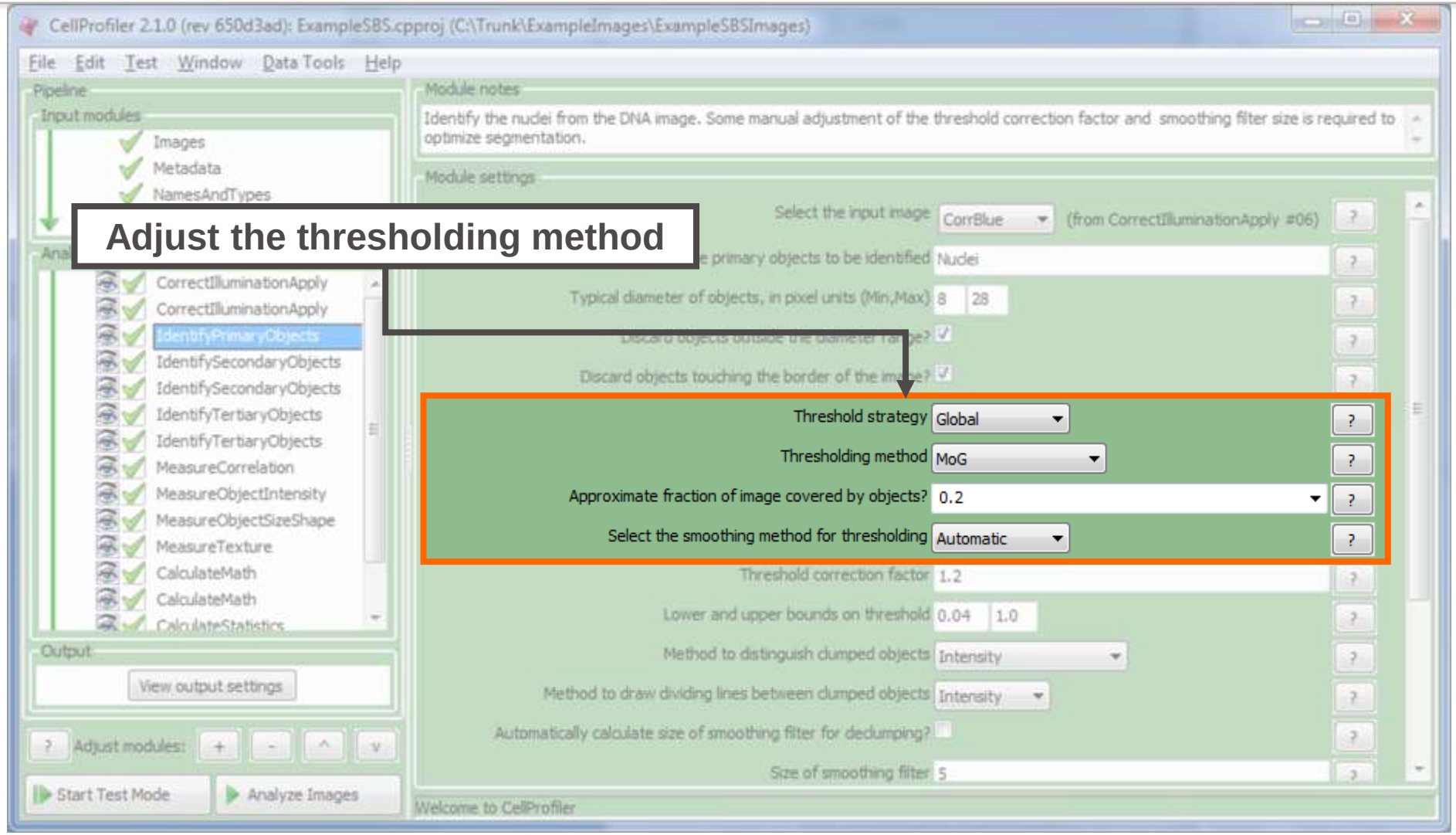
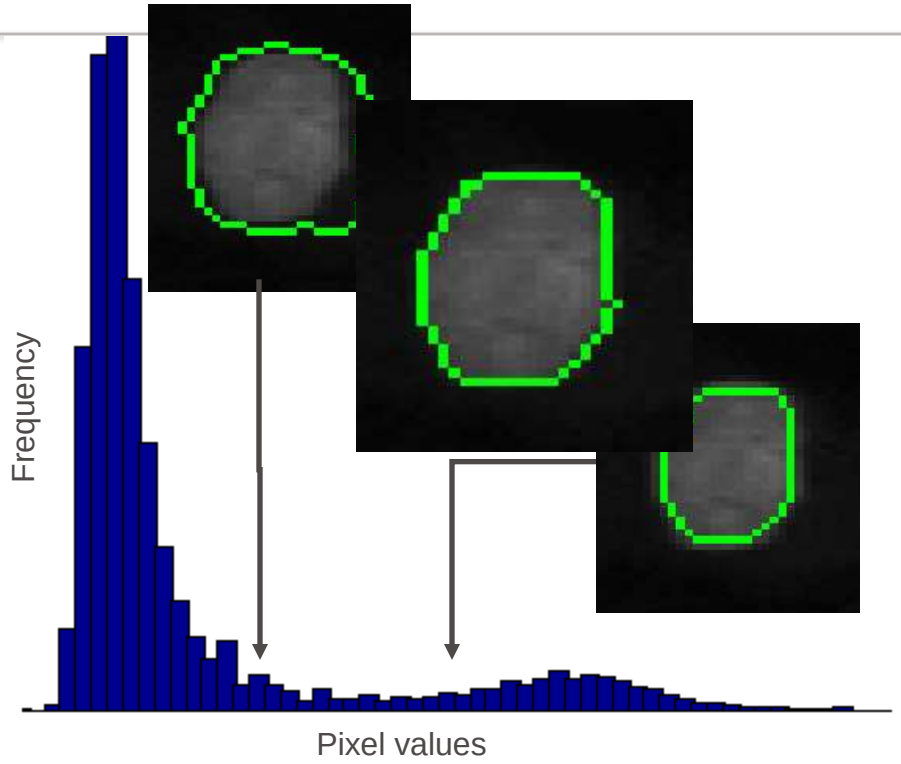


Image Thresholding

- **Definition:** Division of the image into background and foreground

What is the best threshold value for dividing the intensity into foreground and background pixels?

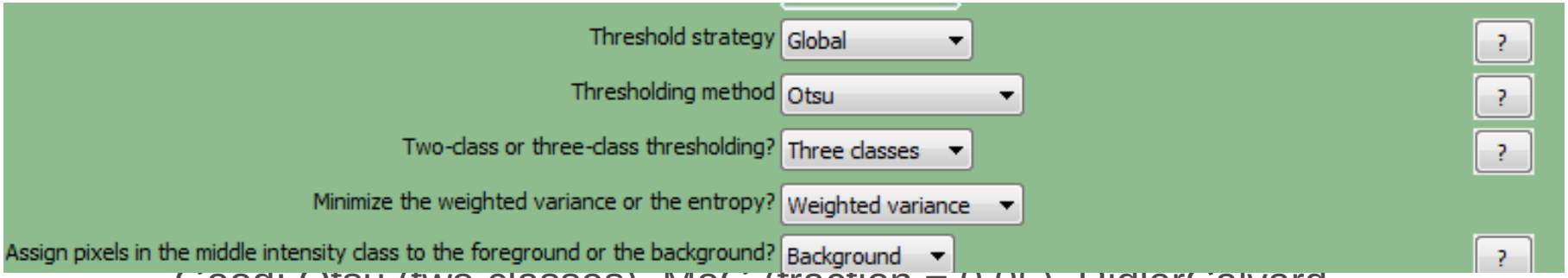


Method: Pick the method that provides the best results

- Automatic: Good for readily identifiable foreground / background
- Otsu: Choose between 2- or 3-class if mid-level intensities present
- Background, RobustBackground: Good for images in which most of the image is comprised of background

Image Thresholding

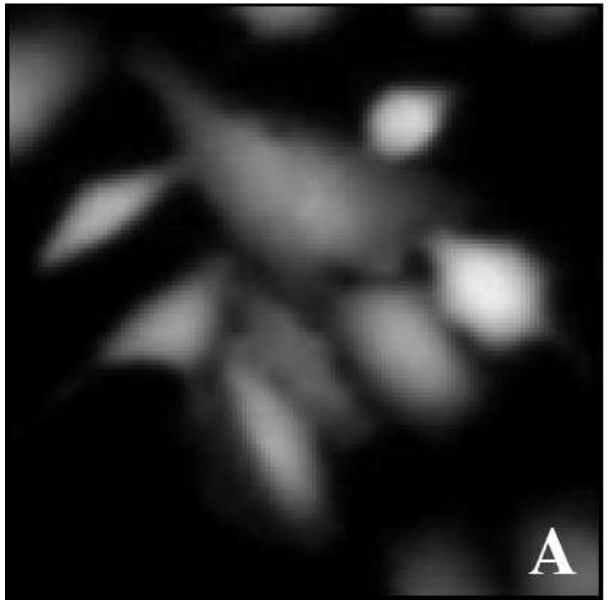
- Experiment with the threshold settings
  - Start with the “Thresholding method” setting
  - Try adjusting the “Thresholding strategy”, if time permits

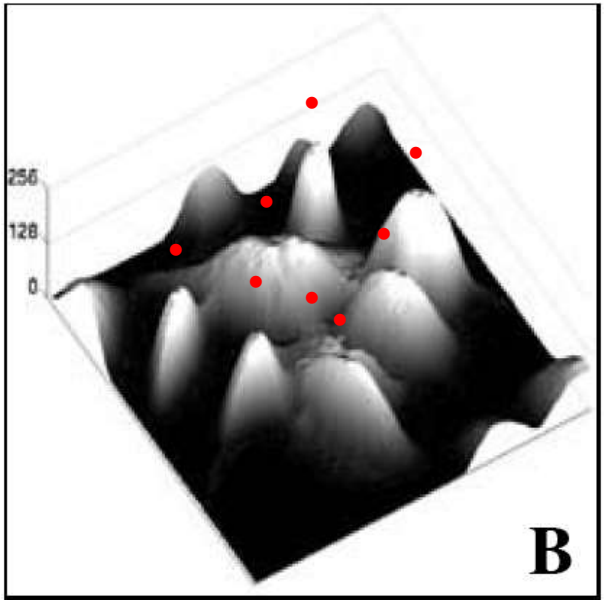


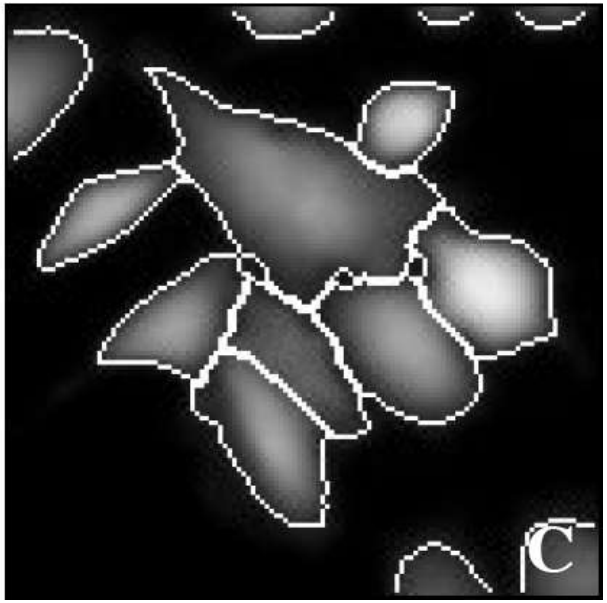
  - Good: Otsu (two classes), MoG (fraction = 0.05), RidlerCalvard
  - Too lenient: Background, RobustBackground, Kapur
  - Too strict: MCT

Separating Touching Objects

- Once the foreground objects have been identified, what next?



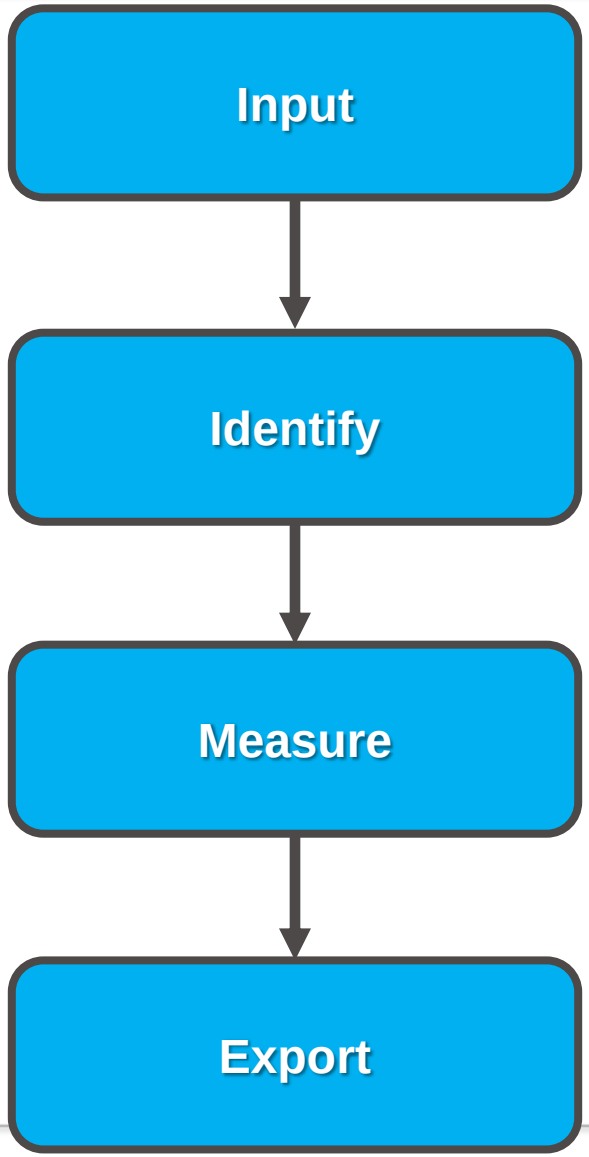


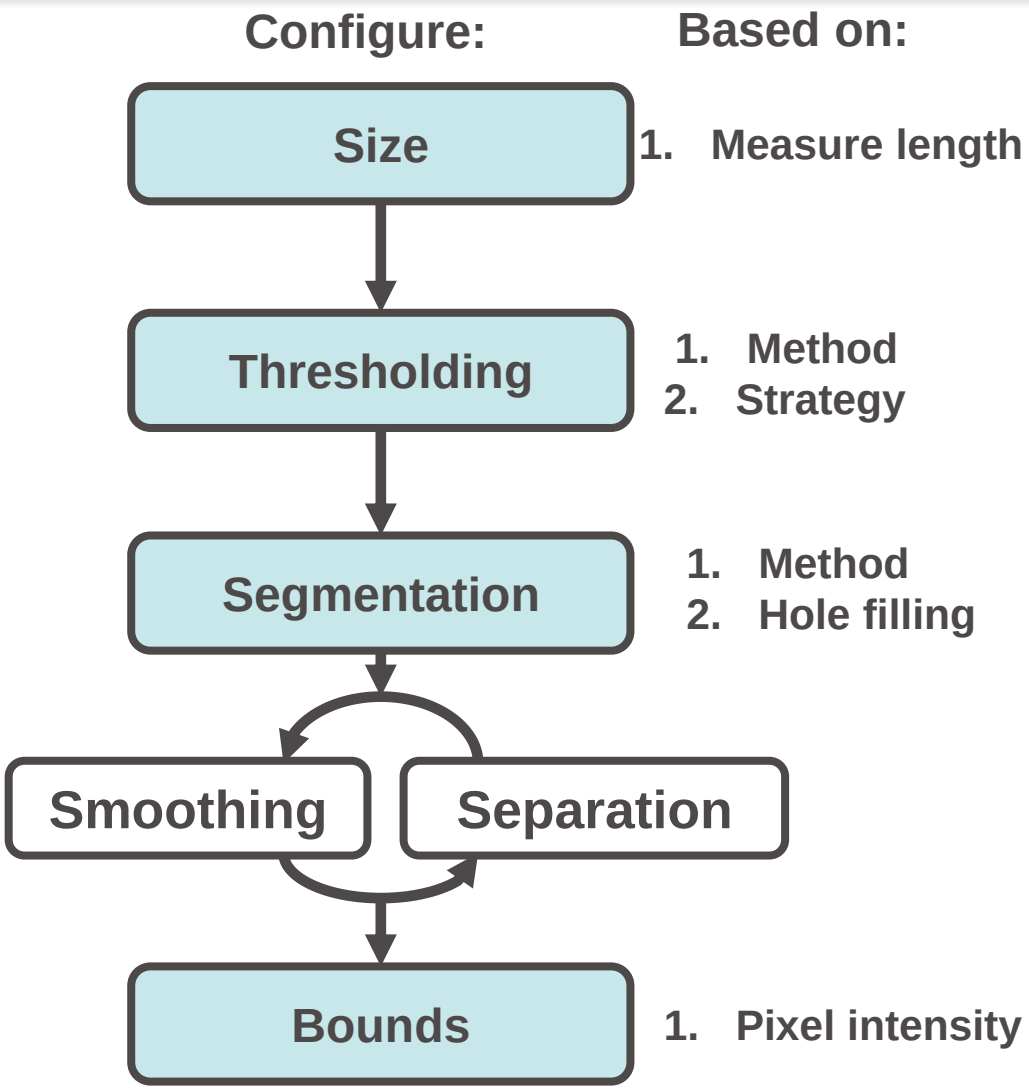


- We need to distinguish multiple objects contained in the same “clump”

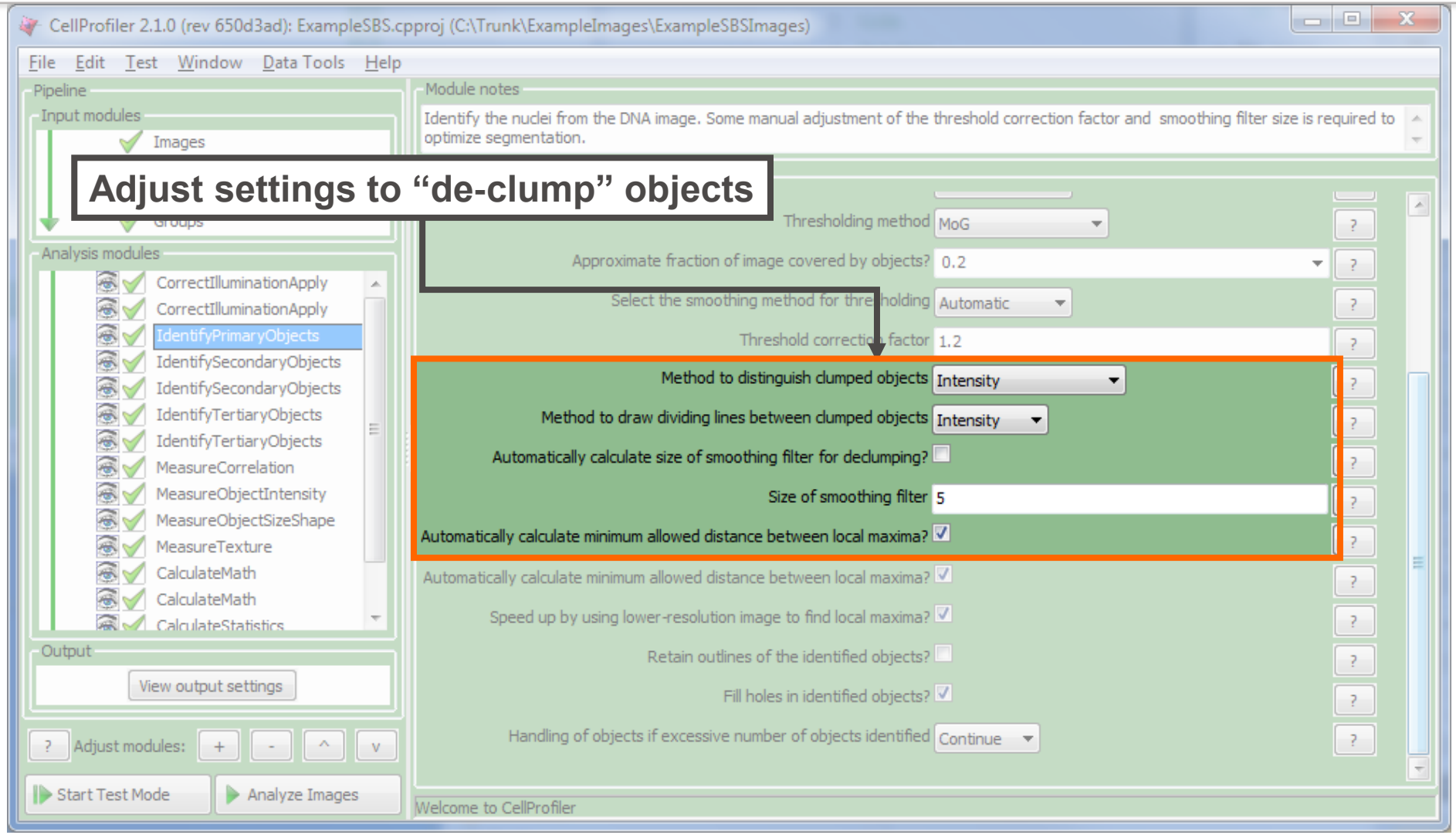
Images from Carolina Wahlby

Typical CellProfiler Workflow

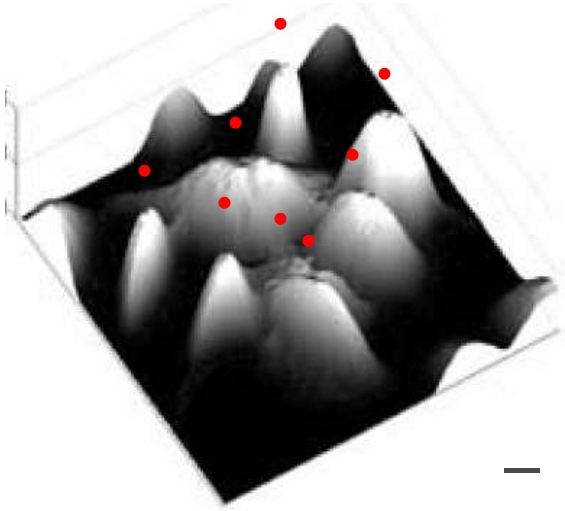




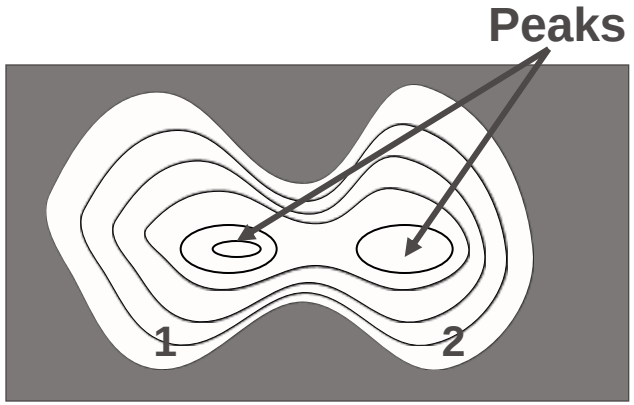
Separating Touching Objects



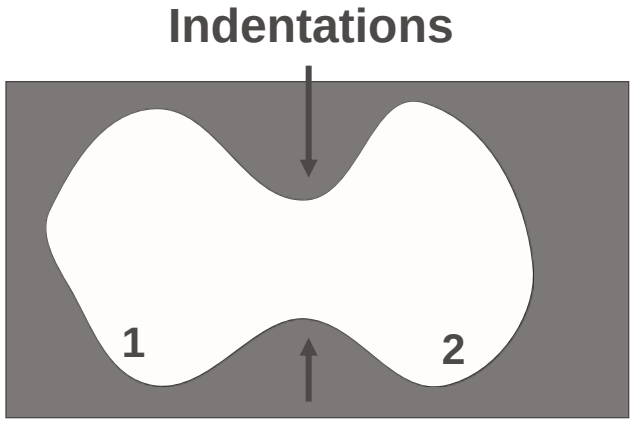
Clump identification: Two options



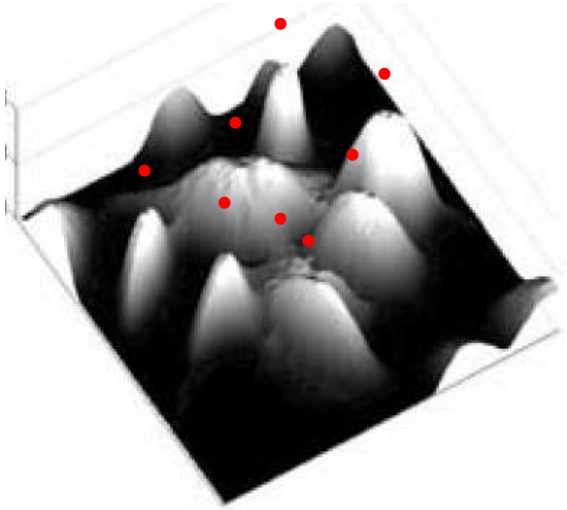
- **Intensity:** Works best if objects are brighter at center, dimmer at edges



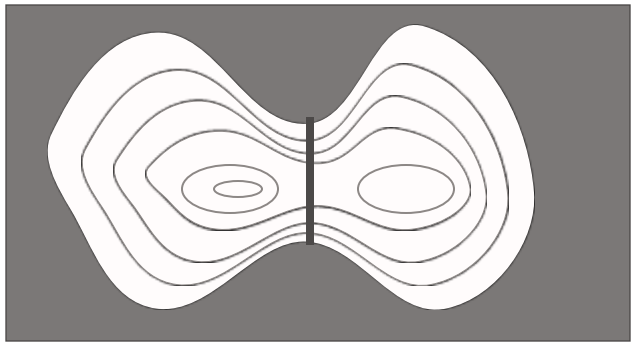
- **Shape:** Works best if objects have indentations where clumps touch (esp. if objects are round)



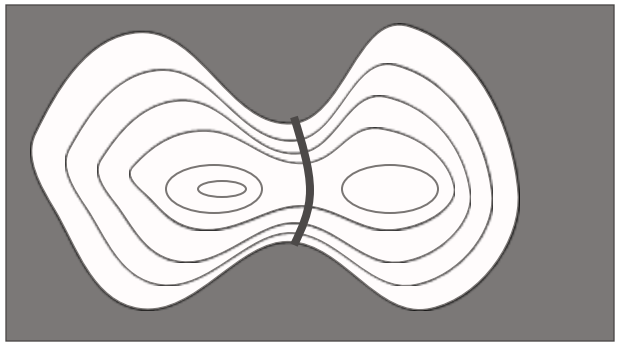
Drawing boundaries: Two options



- **Distance:** Draws boundary lines midway between object centers



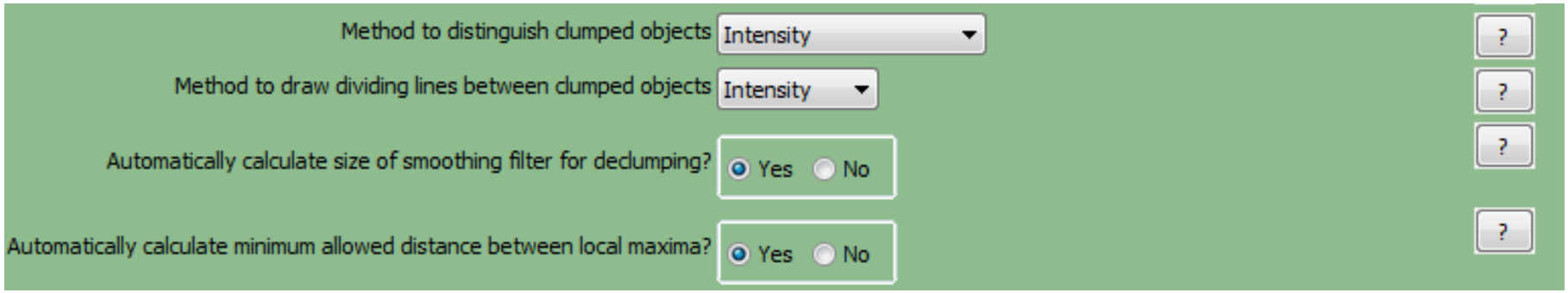
- **Intensity:** Draws boundary lines at dimmest line between objects



Remember to use *Test Mode* to view results of setting combinations

Separating Touching Objects

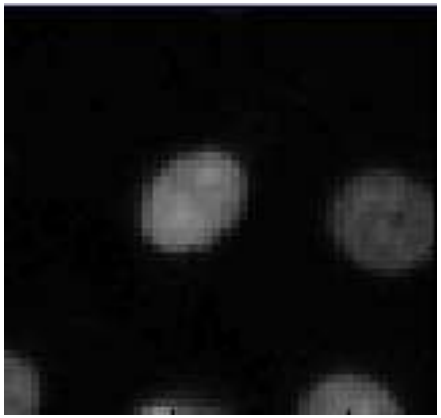
- Experiment with the declumping settings
  - Stick with the “Intensity” and “Shape” methods for now
  - Scan the whole image and look for differences
  - Try the others (“Laplacian of Gaussian”, “Propagate”) if time



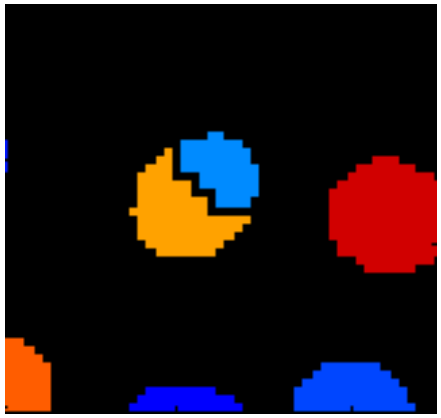
  - Decent: Shape/Shape, Shape/Intensity
  - Not good: Intensity/Shape

Separating Touching Objects

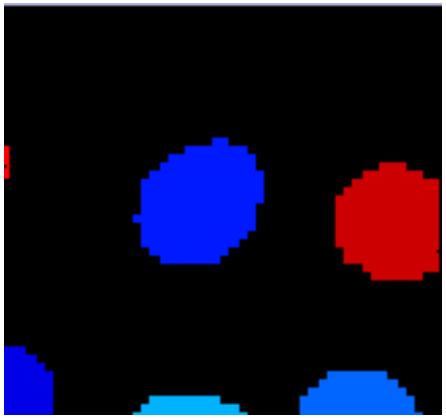
- **Additional separation settings:** Adjust these settings if objects are being incorrectly split into pieces or merged together



Original image



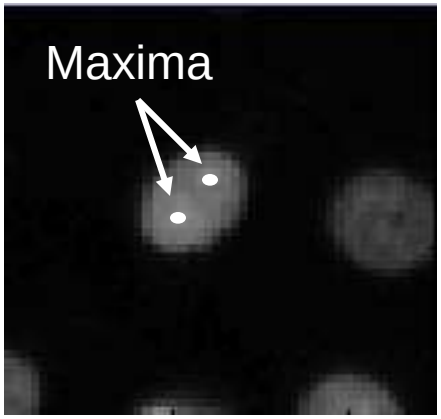
Smoothing filter size = 4



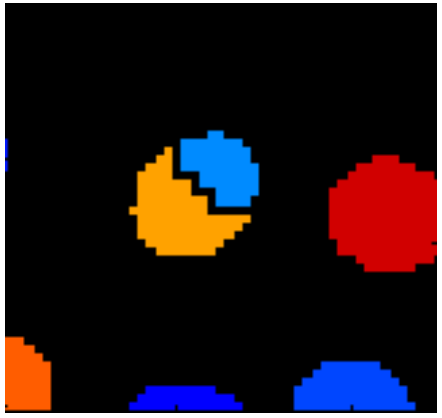
Smoothing filter size = 8

Smoothing: Increase to reduce intensity irregularities which produce over-segmentation of objects

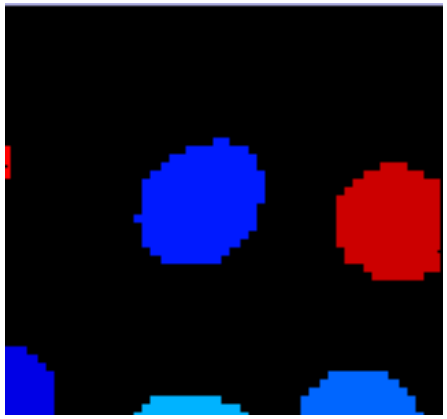
Separating Touching Objects



Original image



Maxima
distance = 4

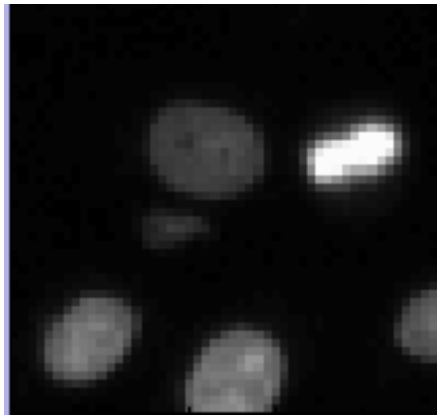


Maxima
distance = 8

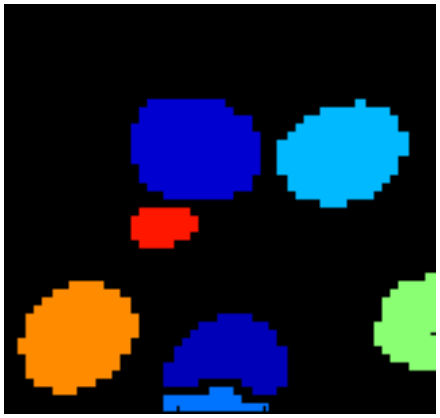
- **Suppress Local Maxima**
  - Smallest distance allowed between object intensity peaks to be considered one object rather than a clump
  - Decrease to reduce improper merging of objects in clumps

Separating Touching Objects

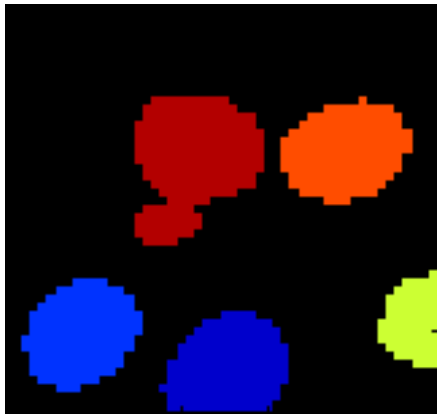
However….



Original image



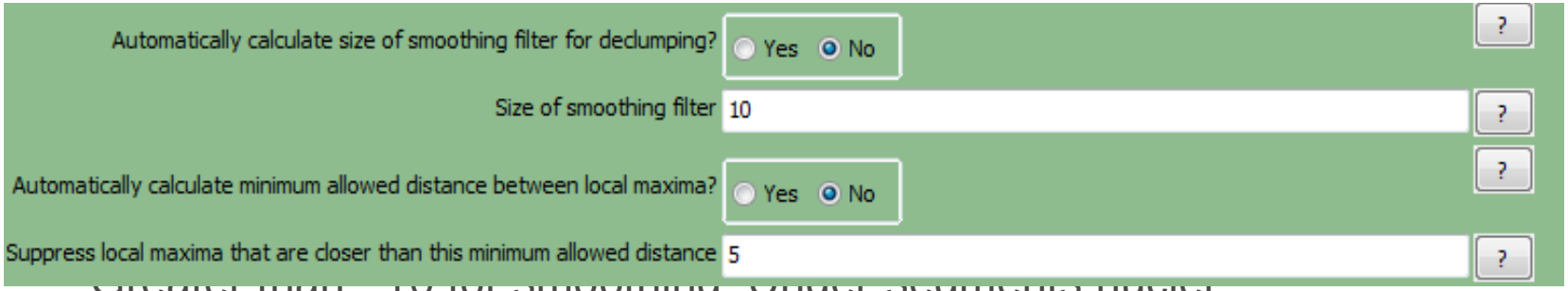
Smoothing filter size = 4



Smoothing filter size = 8

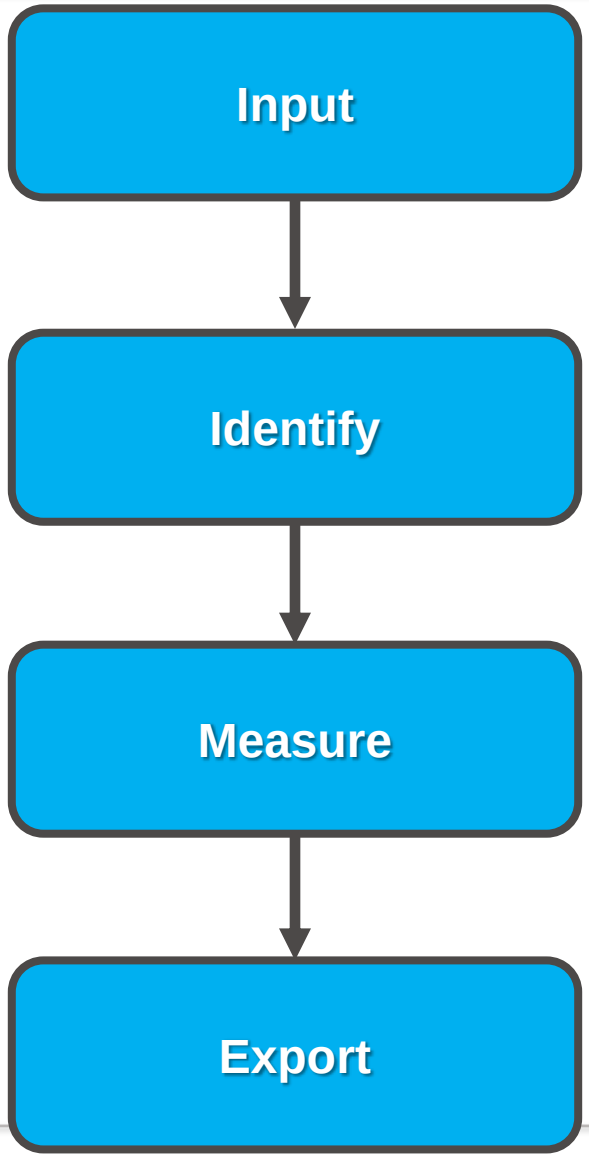
- Adjusting can produce more improper segmentation than it solves
- The proper settings are usually a matter of trial and error
  - The automatic settings are a good starting point, though

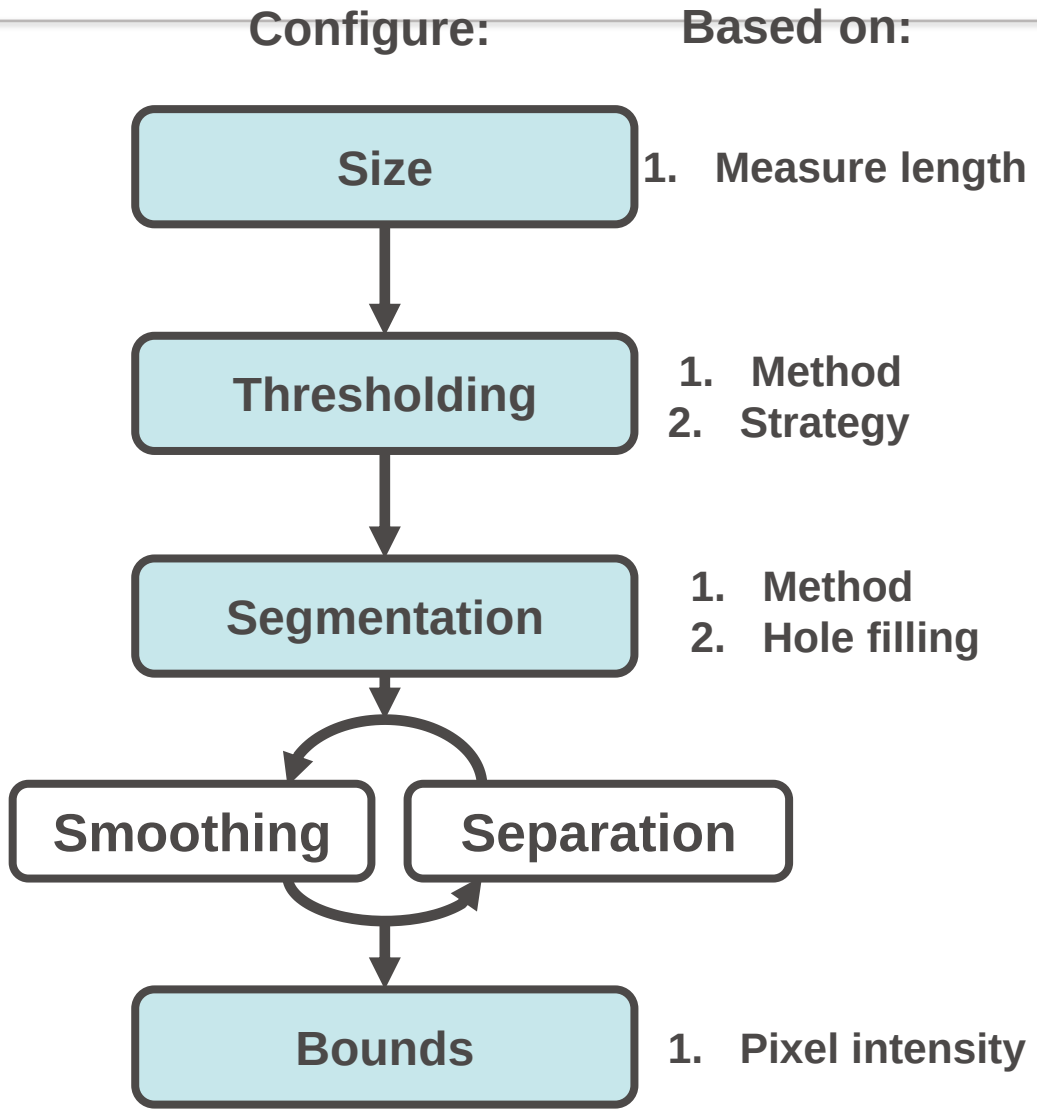
Separating Touching Objects

- Experiment with smoothing filter, minimum allowed distances
  - Note the current values
  - Suggestion: Step from 2 to 12 pixels for both settings
  - Use “Measure length” tool to ball-park minimum allowed distance
  - Scan the whole image and look for differences
- 

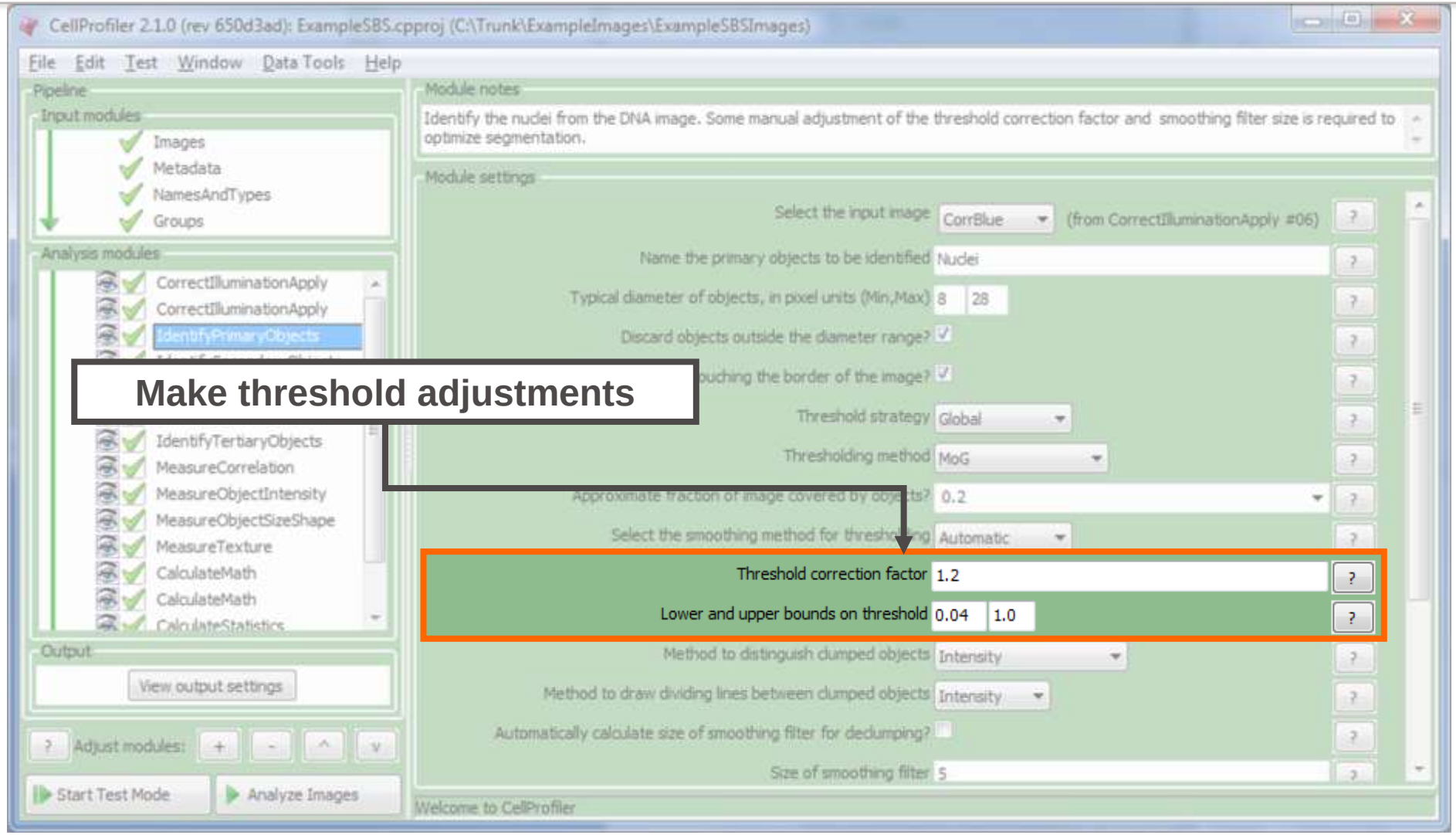
  - Greater than ~7 for minimum distance: Under-segments nuclei

Typical CellProfiler Workflow





Further Identification Adjustments

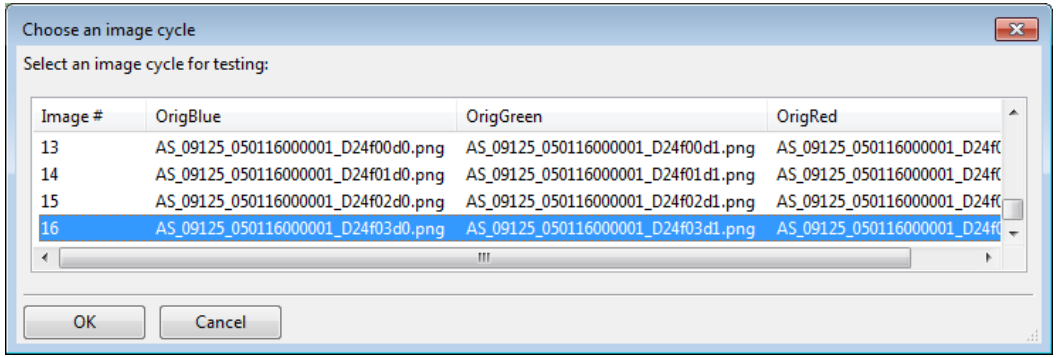


Further Identification Adjustments

- **Correction factor**
  - Multiplication factor applied to threshold
  - Adjusts threshold stringency/leniency
  - Setting this factor is empirical
- **Upper/lower bounds**
  - Set safety limits on automatic threshold to guards against false positives
  - Helpful for unexpected images: Empty wells, images with dramatic artifacts, etc

Further Identification Adjustments

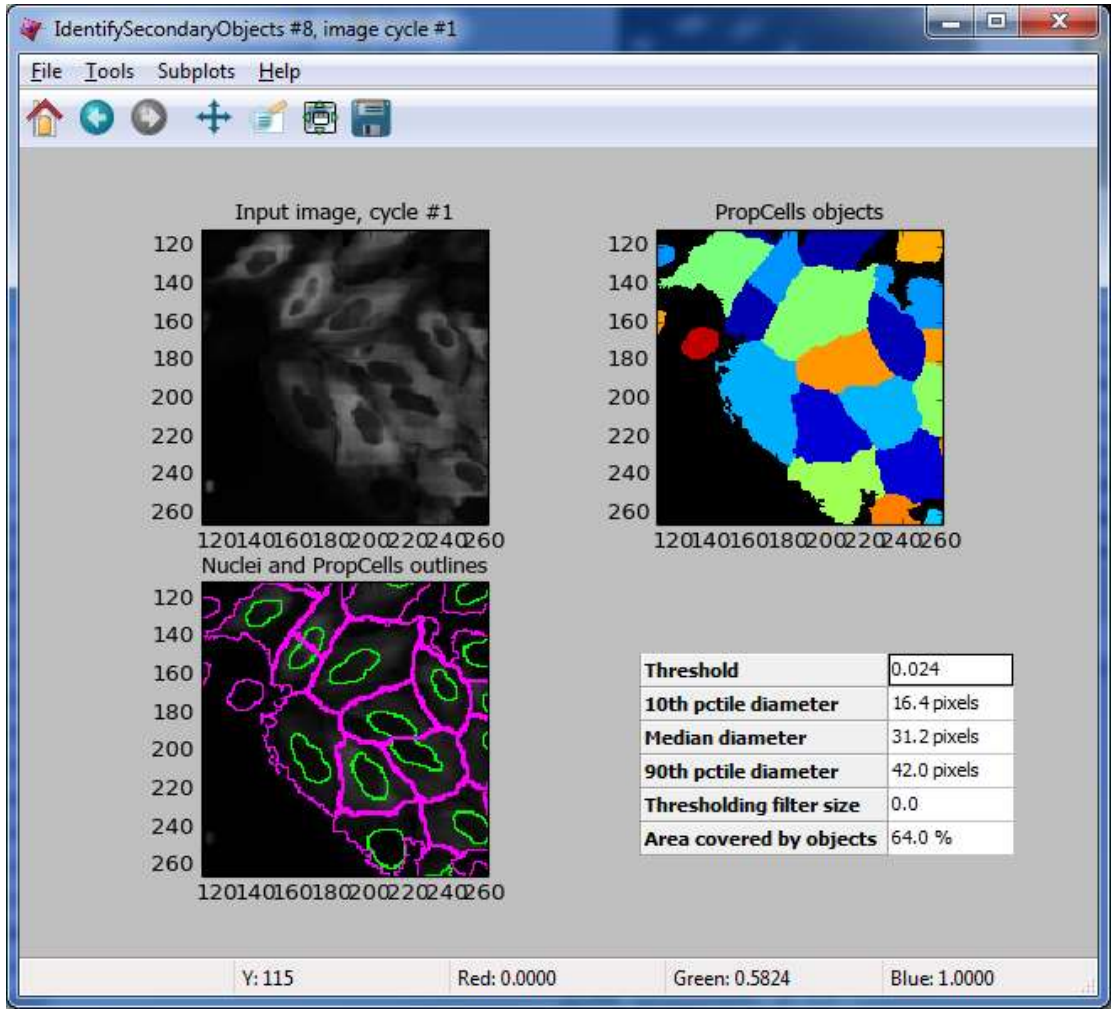
- From *Test* > *Choose Image set*, select the last image set



- Run the pipeline. How does the nuclei identification look? Why?
- What appears to be a good lower bound?
  - Using the intensity tool
  - Using the histogram
- Adjust the lower bound, re-run the module
- Confirm your settings: Use Test menu to go back to the 1st image, run the pipeline

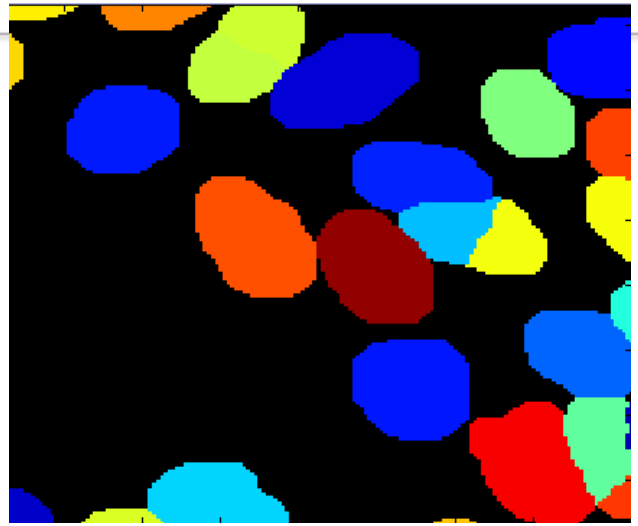
Secondary Object Identification

- **Goal:** Identify cell boundaries by "growing" primary objects
  - Nuclei typically more uniform in shape, more easily separated than cells
- **Approach:** Segment nuclei → Seeds for cell segmentation by using a cell stain channel



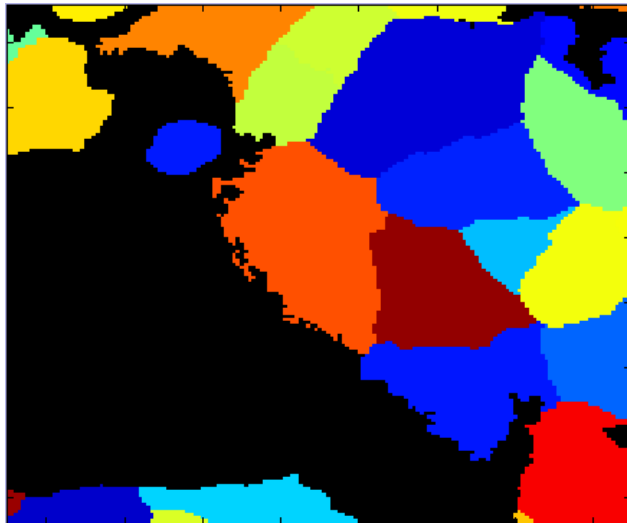
Secondary Object Identification: Methods

- Distance-N: Ignores image information
  - Useful when cell stain is absent



Distance-N

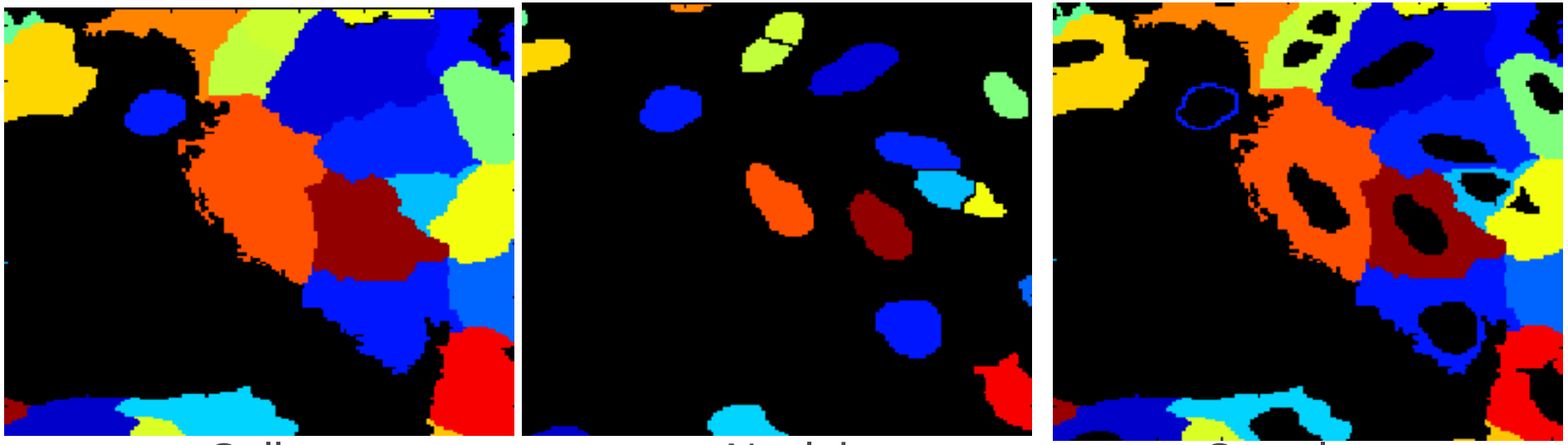
- Watershed, Propagate, Distance-B: Uses image information
  - Finds dividing lines between objects and background / neighbors



Propagation

Tertiary Object Identification

- **Goal:** Identify *tertiary objects* by removing the primary objects from secondary objects
  - "Subtract" the nuclei objects from cell objects to obtain cytoplasm



Cells — Nuclei = Cytoplasm

Secondary and Tertiary Object Identification

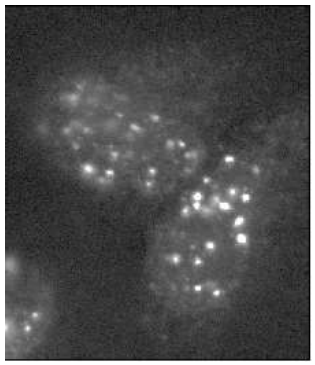
- In *IdentifySecondaryObjects*, change the input image to "CropRed"

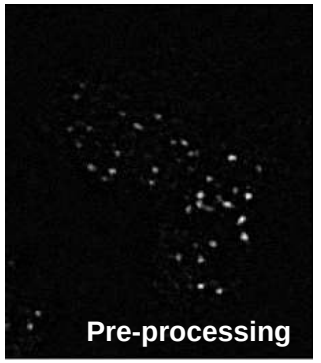
- Experiment with the secondary object identification methods
- Try different thresholding methods, if time permits
- Which method works best in this case?
  - Good: Propagation, Distance-B
  - Decent: Watershed-Image, Watershed-Gradient
  - Not good: Distance-N

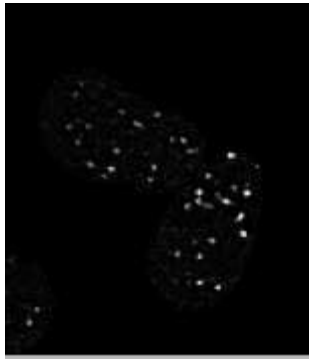
- Press Step button to execute *IdentifyTertiaryObjects*

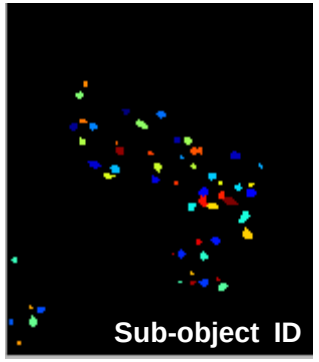
Identifying Subcellular Structures

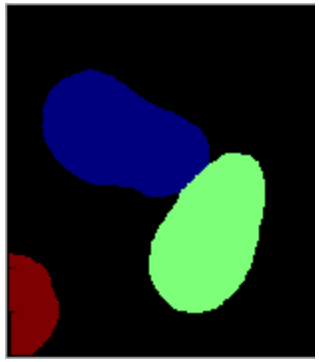
- With appropriate markers, other subcellular compartments can be labeled
- These can be identified using the same methods already mentioned
- Consider using enclosing object as mask for better pre-processing, thresholding
- Make sure to assign subfeatures to enclosing objects

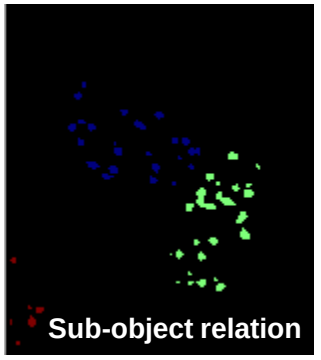












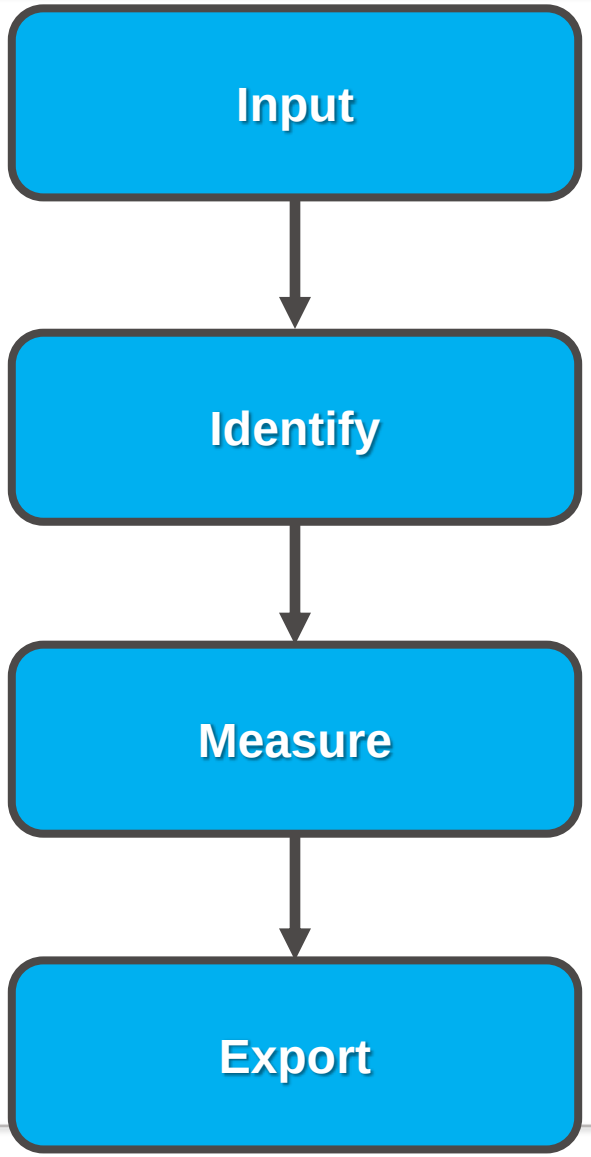
Identifying Subcellular Structures

- Add another *IdentifyPrimaryObjects* module
  - Position it after IdentifyTertiaryObjects
  - Re-enter Test mode if you need to
  - Adjust settings
    - Select “CropGreen” as the input image
    - Enter “pH3” as the primary object name

- Add *RelateObjects* module
  - Position it after *IdentifyPrimaryObjects*
  - Adjust settings
    - Set child objects as “pH3”
    - Set parent objects as “Nuclei”

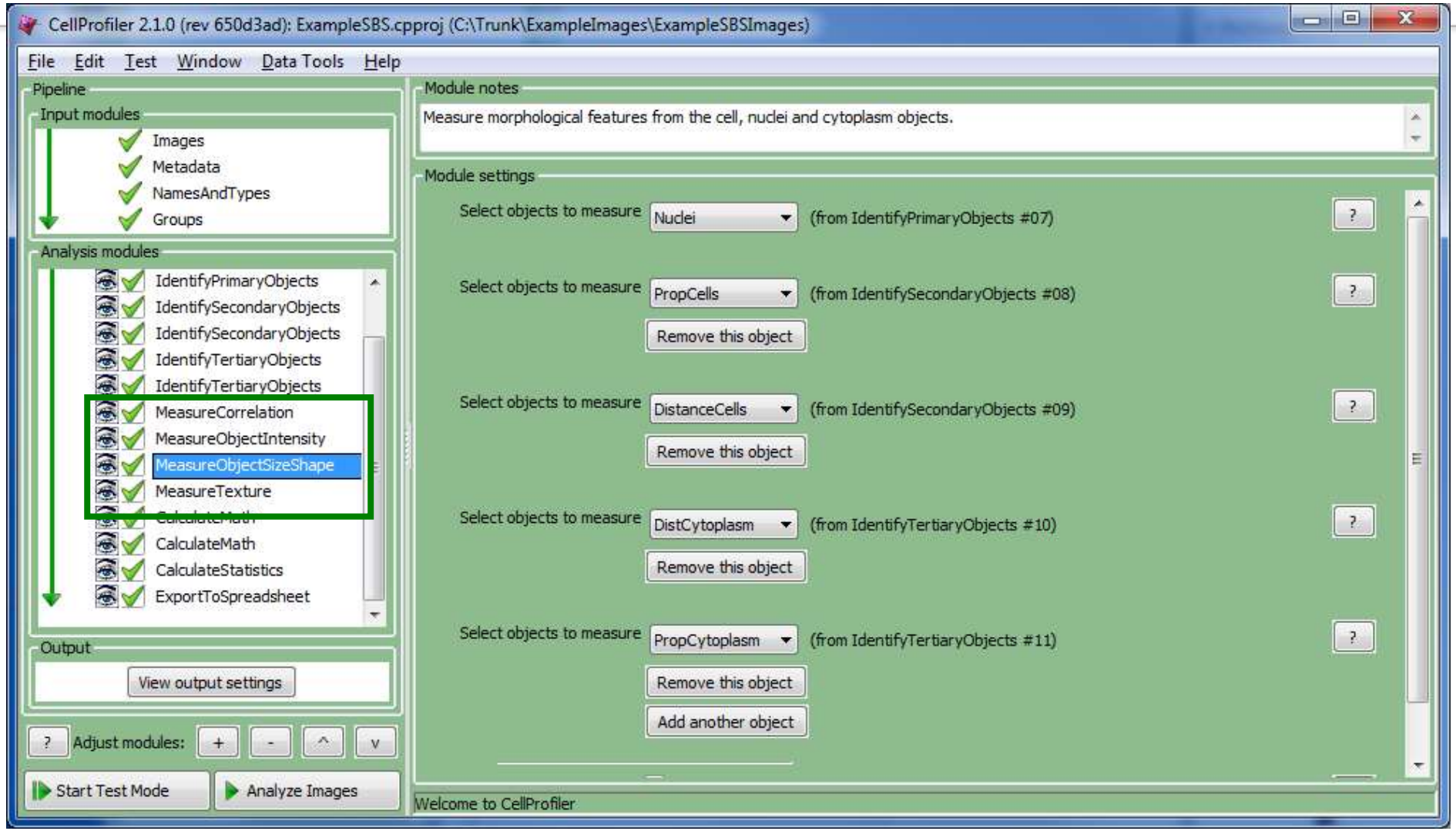
- Set new pause after *RelateObjects*, and run the pipeline

Typical CellProfiler Workflow

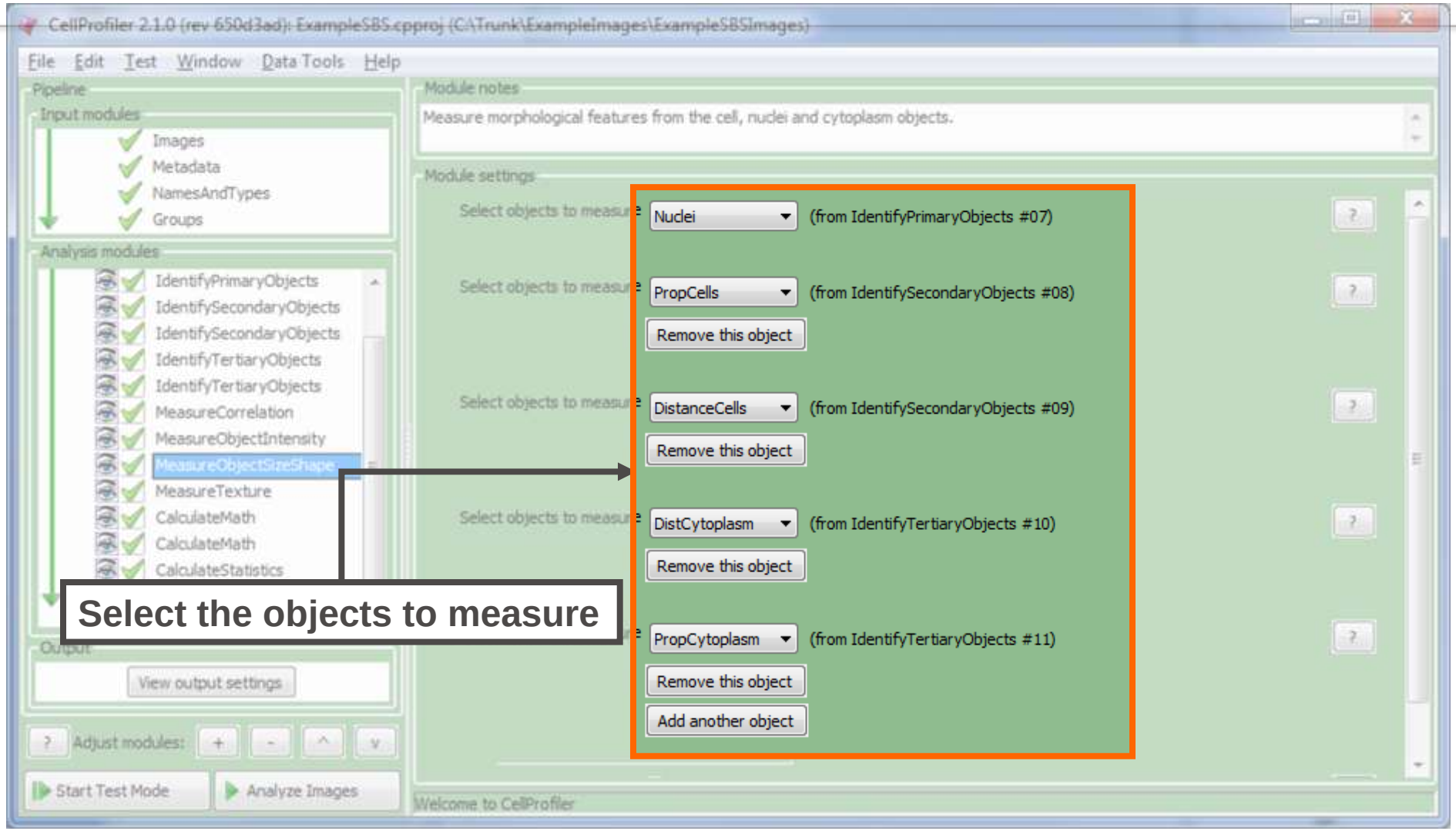


Configure:	Based on:
Morphology	Object
+	
Intensity	1. Object 2. Channel
+	
Texture	1. Object 2. Channel 3. Scale
+	
Location	1. Object 2. Time-lapse?
+	
Clustering	Object

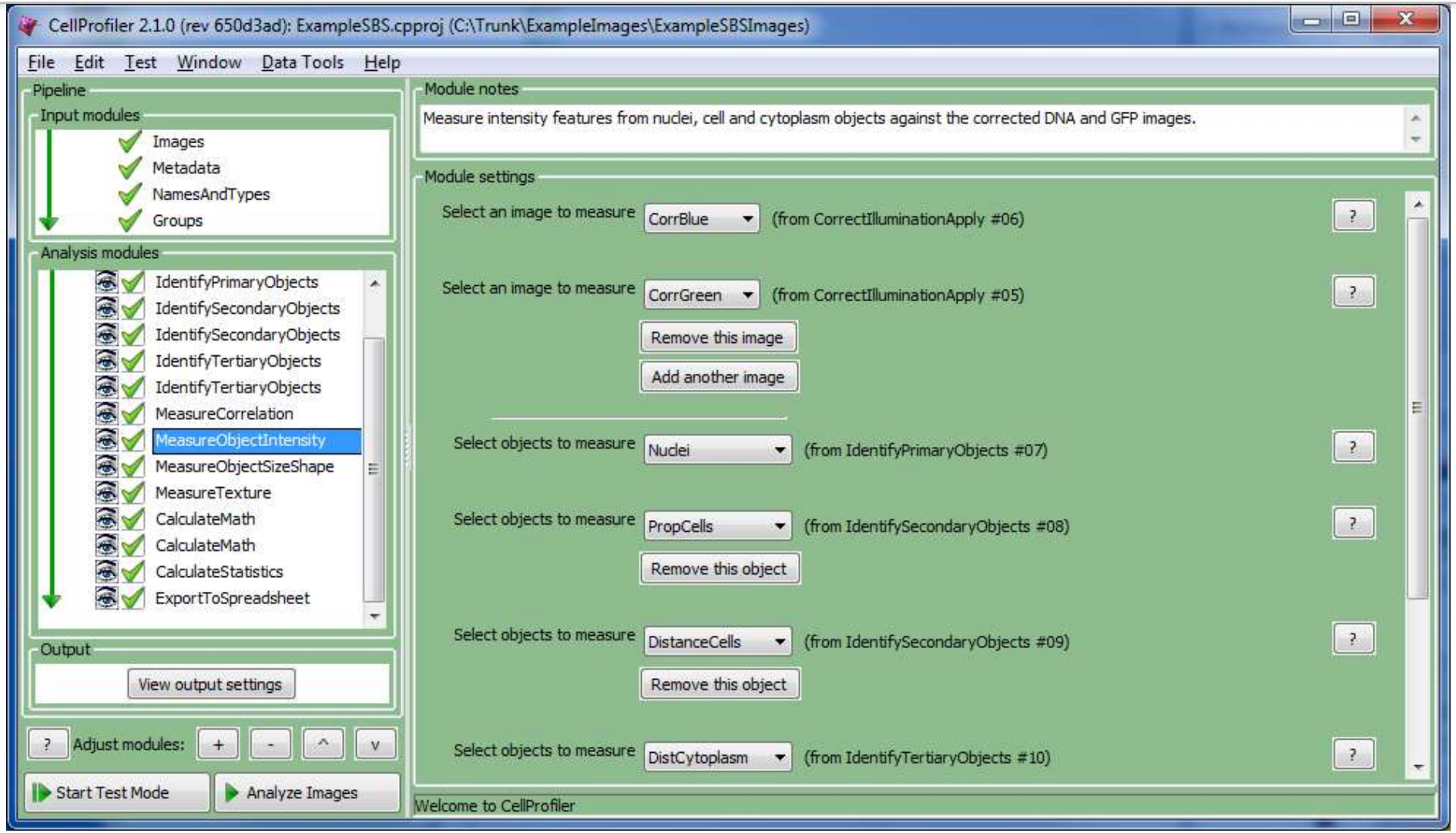
Measurement Modules: Morphology



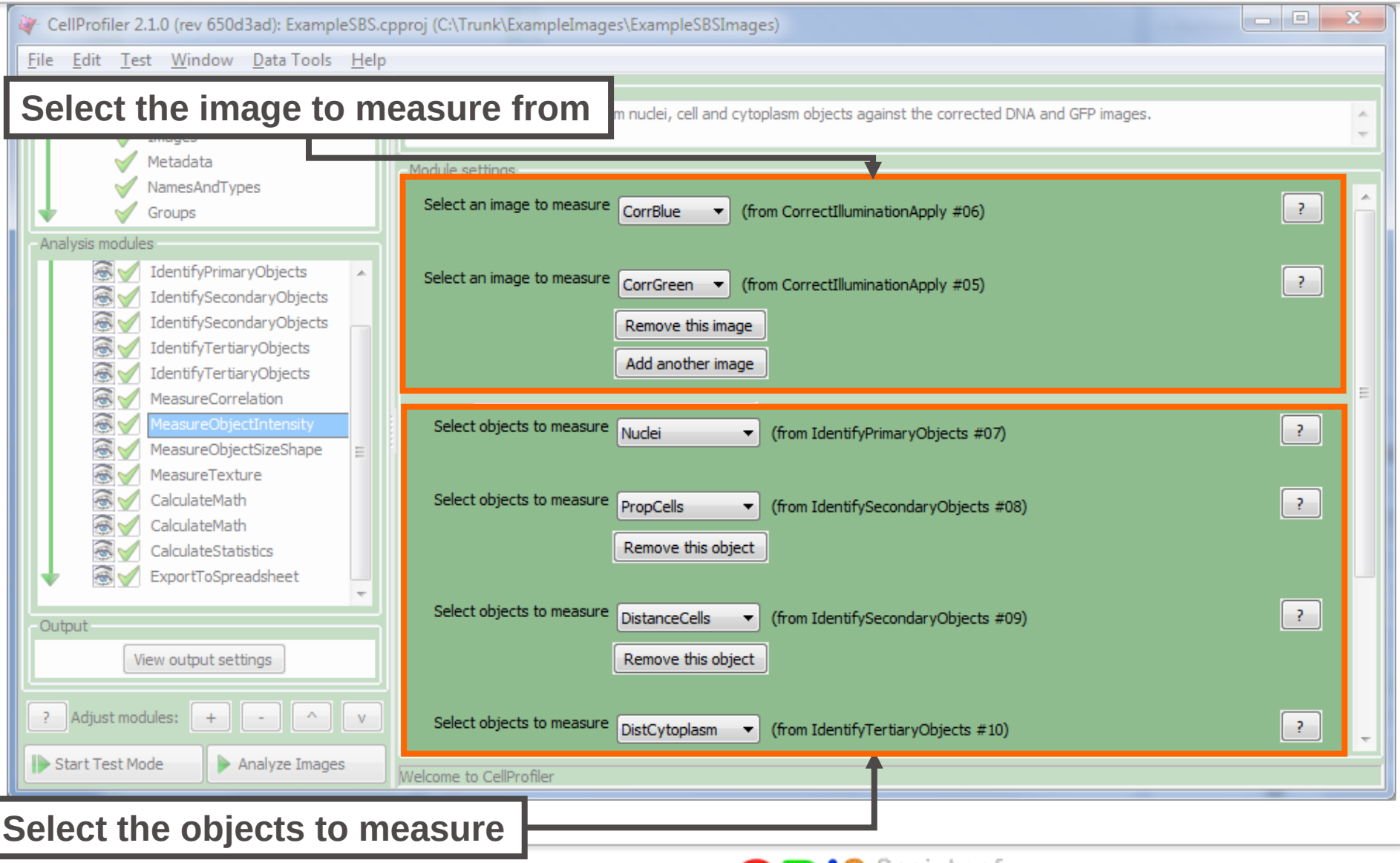
Measurement Modules: Morphology



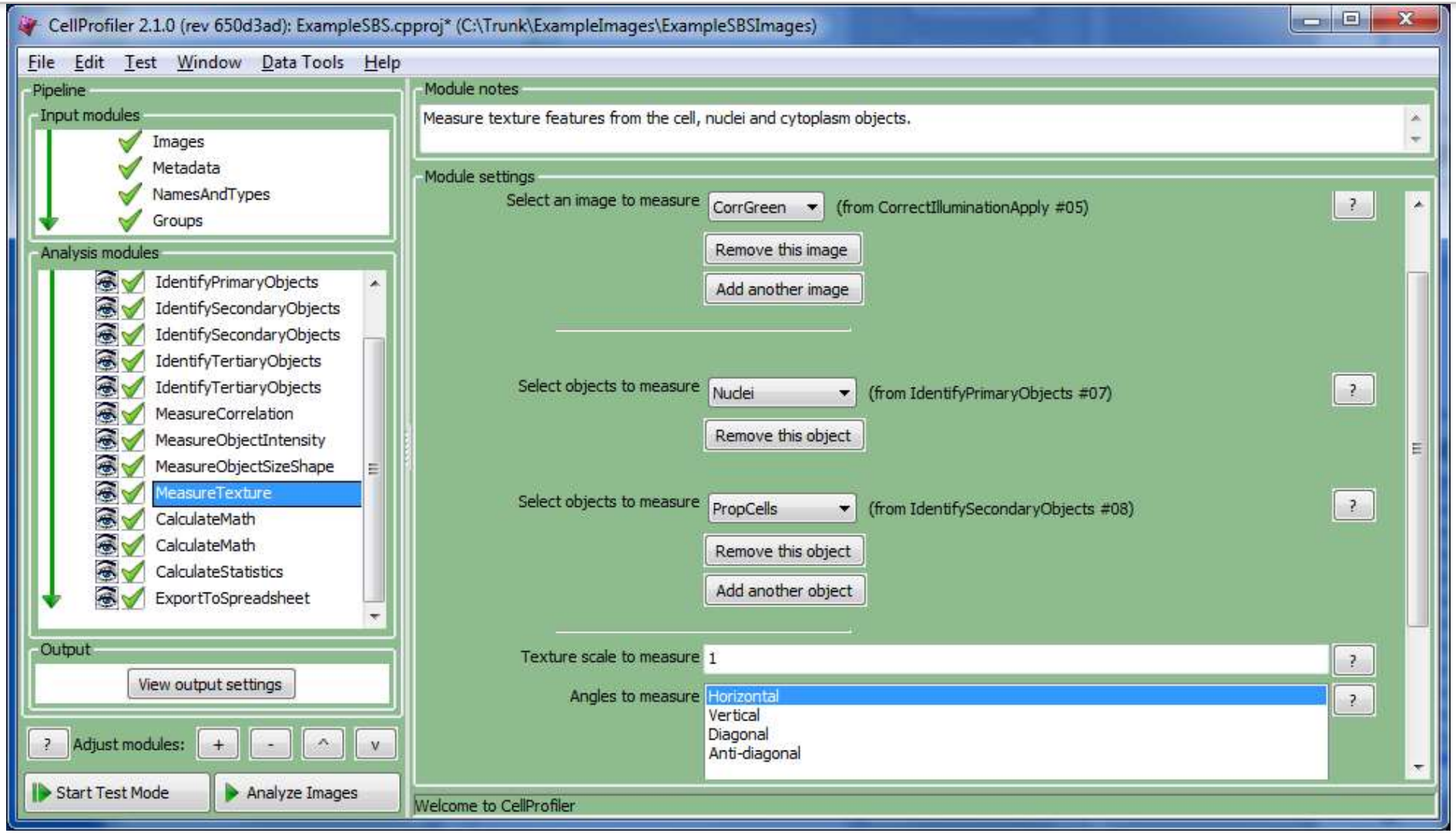
Measurement Modules: Intensity



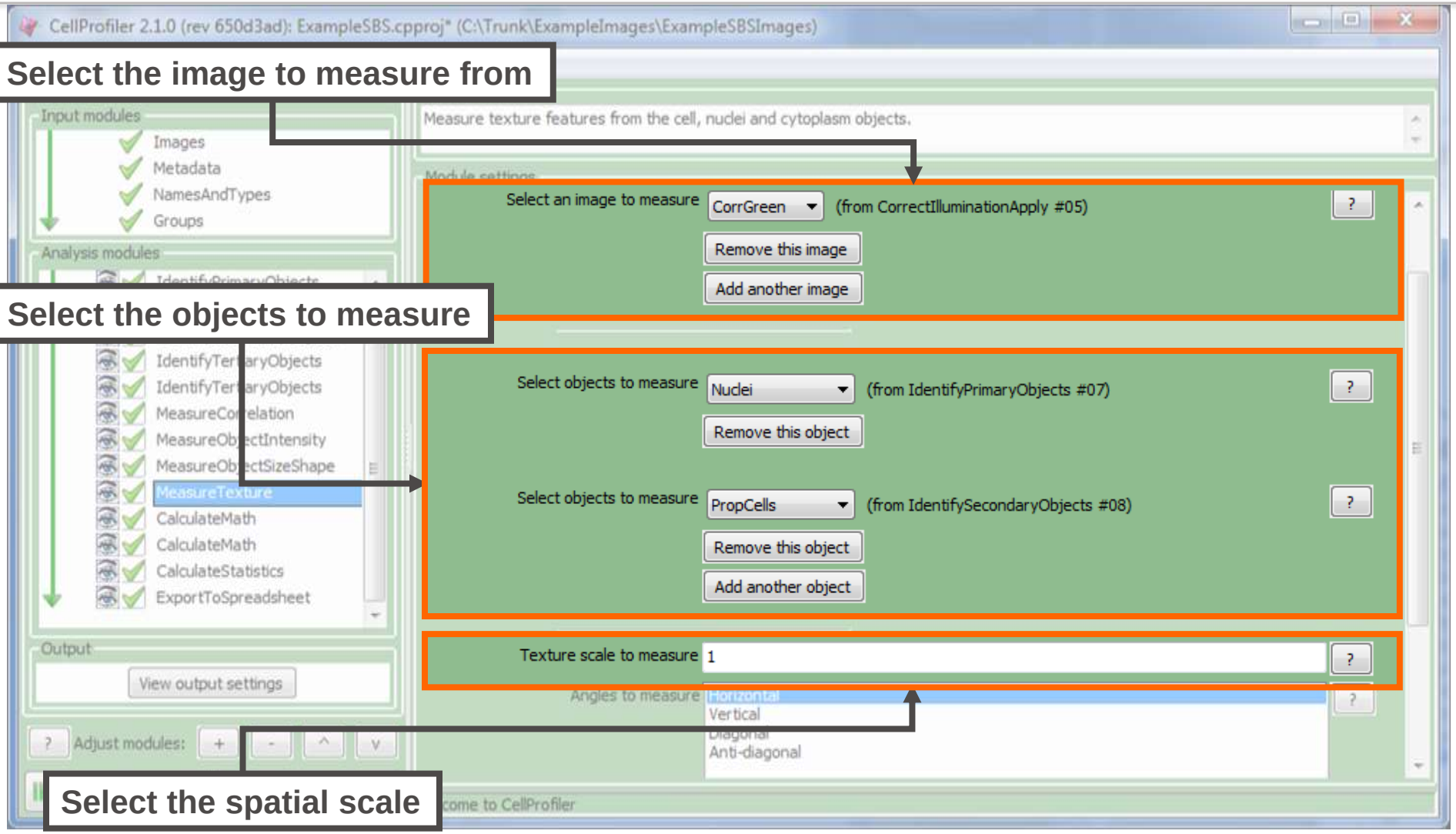
Measurement Modules: Intensity



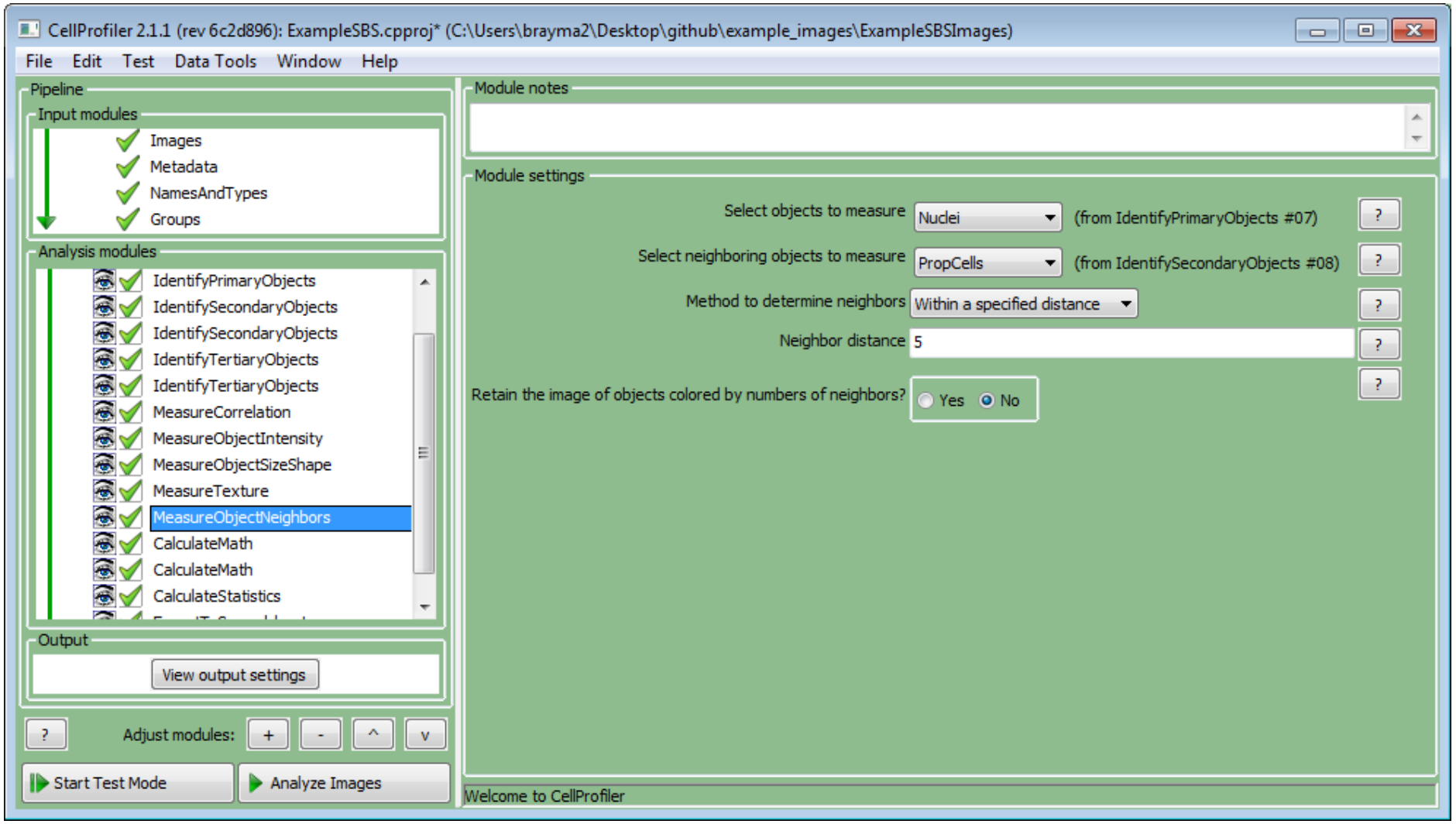
Measurement Modules: Texture



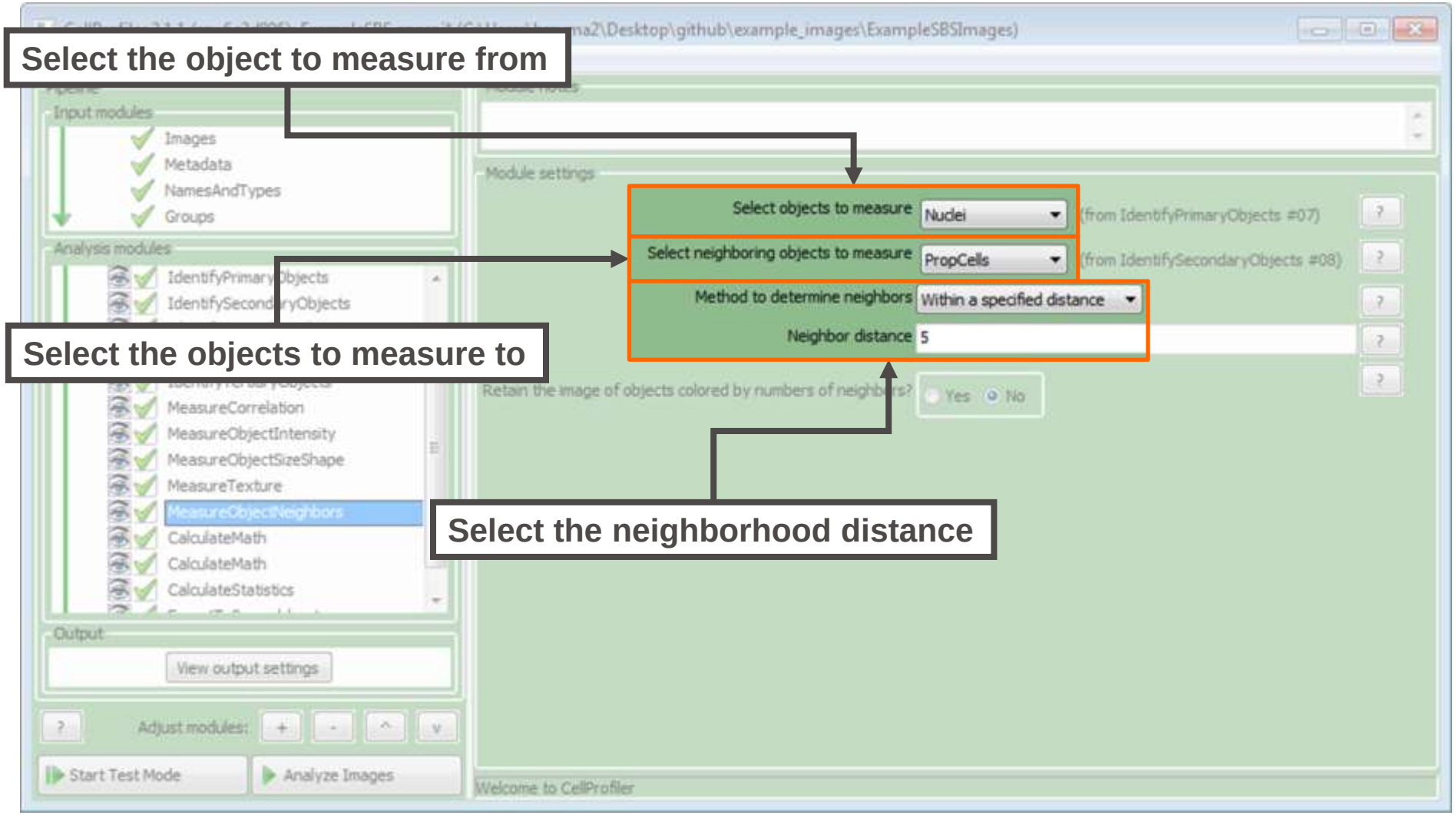
Measurement Modules: Texture



Measurement Modules: Clustering



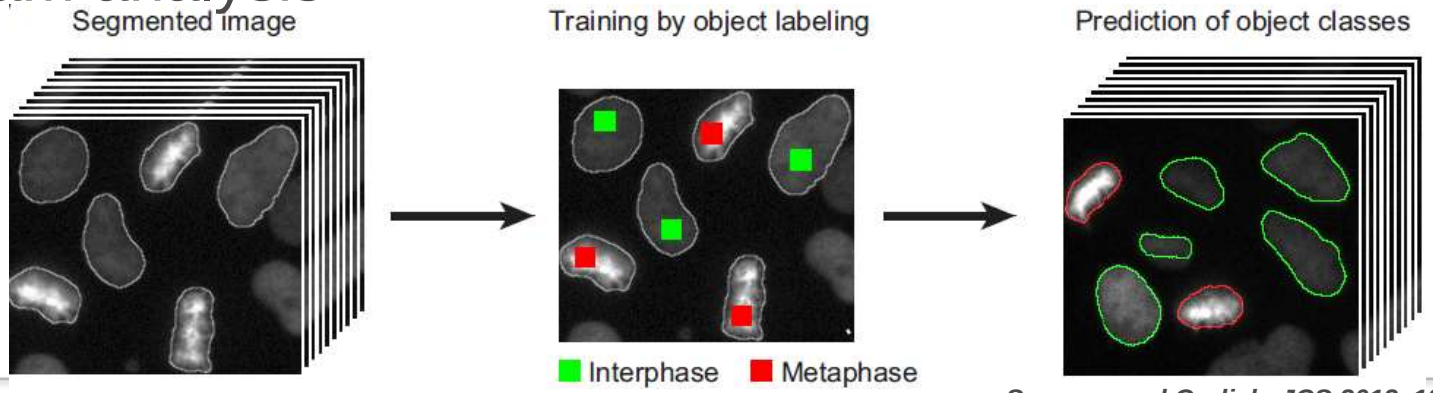
Measurement Modules: Clustering



Combinations of Measurements

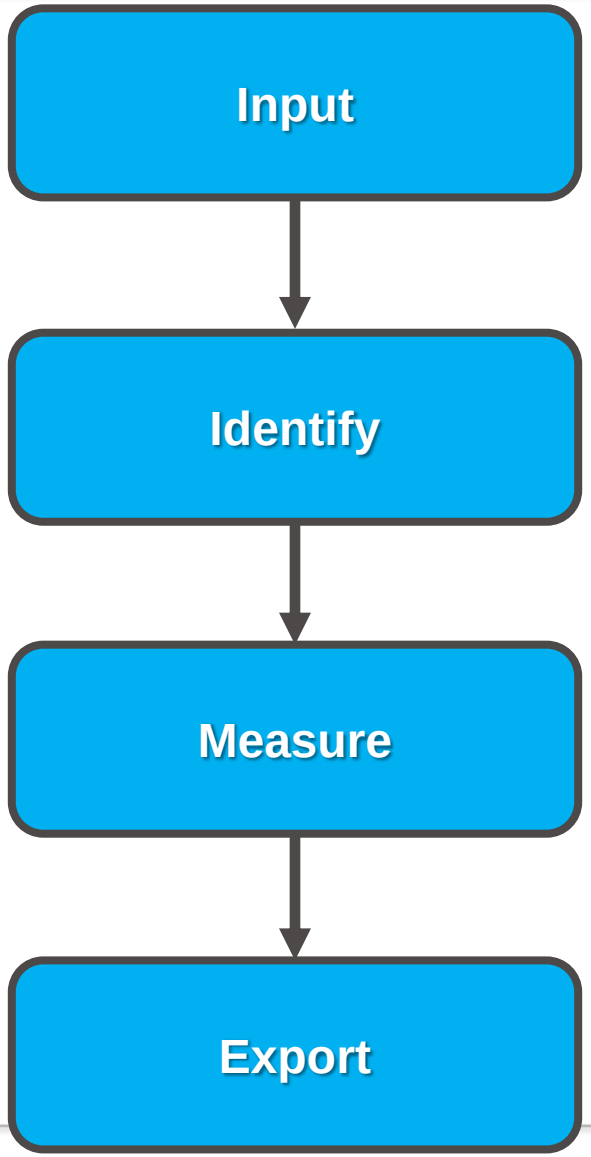
- Phenotype identification may be difficult if hand-selecting from a limited measurement set
- Machine learning (ML) approaches can identify phenotypes from a combination of measurements

- Some measurements (e.g., texture) are hard to interpret as readouts but are excellent fodder for ML approaches to downstream analysis



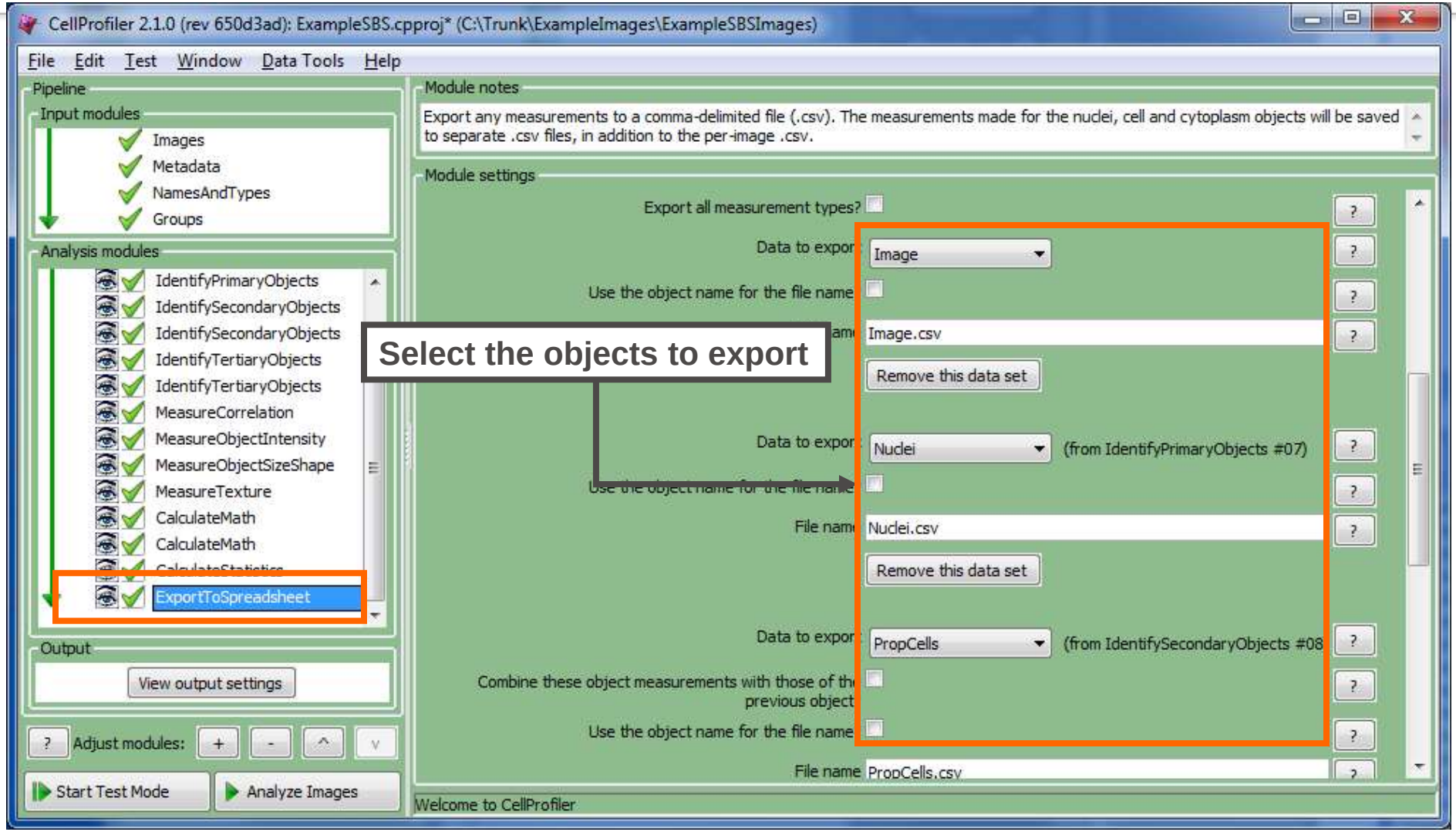
Sommer and Gerlich, JCS 2013, 126:1

Typical CellProfiler Workflow



Configure:	Based on:
Images	1. Merge channels 2. Outlines
+	
CSV	1. Object 2. Image
+	
Database	1. Object 2. Image

Data Export Modules



- User may output images or image measurements

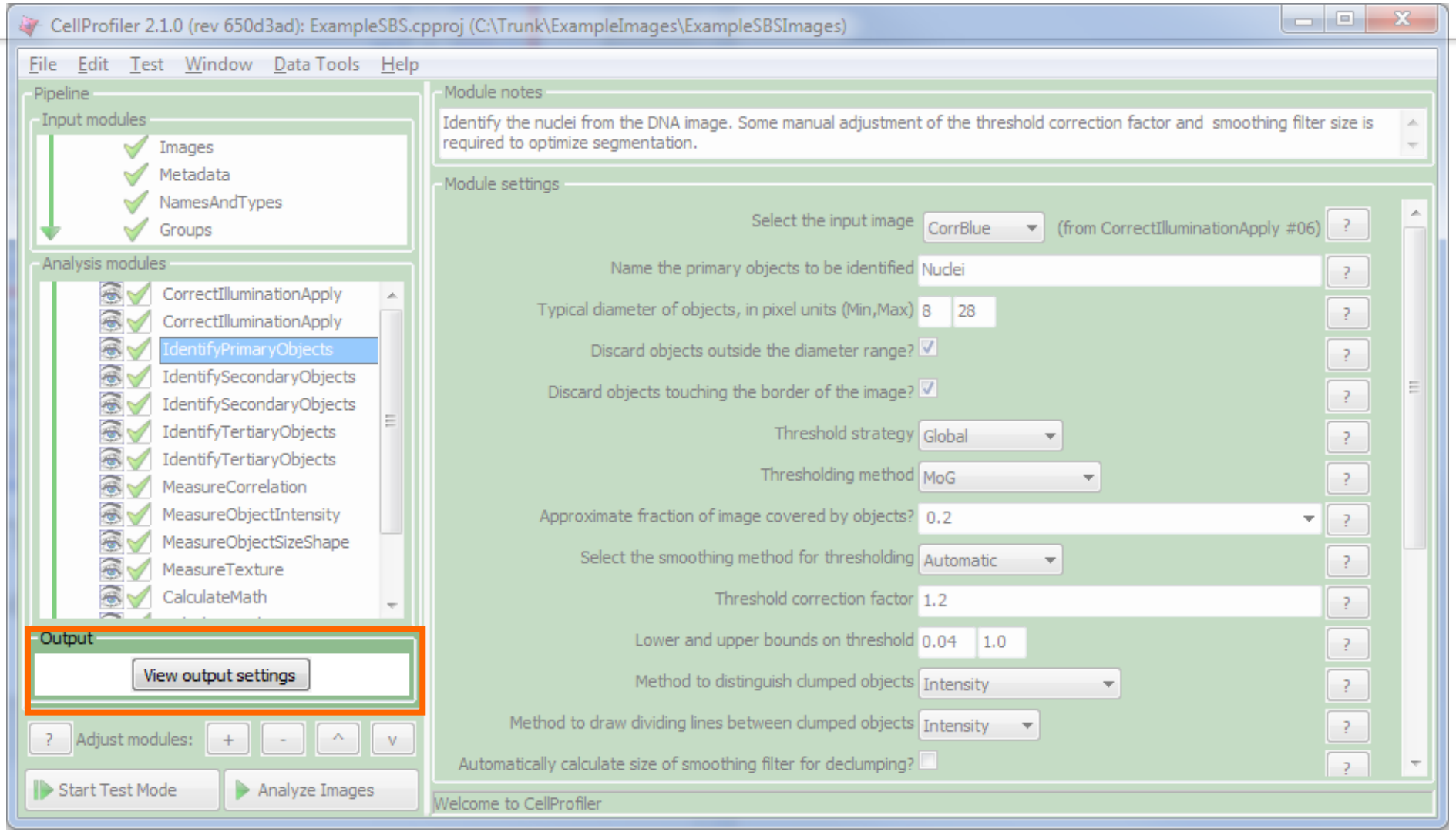
Data Export Modules

- **Goal:** Retain images of intermediate image processing steps for quality control *or* save measurements for later analysis and exploration
- **SaveImages:** Writes an image to a file
  - Intermediate images in the pipeline are not saved unless requested
  - Choice of many image formats to write → module can be used as an image format converter
- **ExportToSpreadsheet:** Export measurements as a comma-separated file readable by spreadsheet programs
- **ExportToDatabase:** Export measurements as a per-object and per-table plus configuration file for a MySQL or SQLite database

Data Export

- Remove the **Crop** modules, re-set identification module inputs
- Remove the **SaveImages** module
  - Not yet supported on HCSIA
- Leave settings on **ExportToSpreadsheet** as-is
  - Including this module is required if you want per-well results

Before The Analysis…



- **Output panel:** Specify the location to place outputs
  - Spreadsheets of measurements, saved images, etc

Before the Analysis…

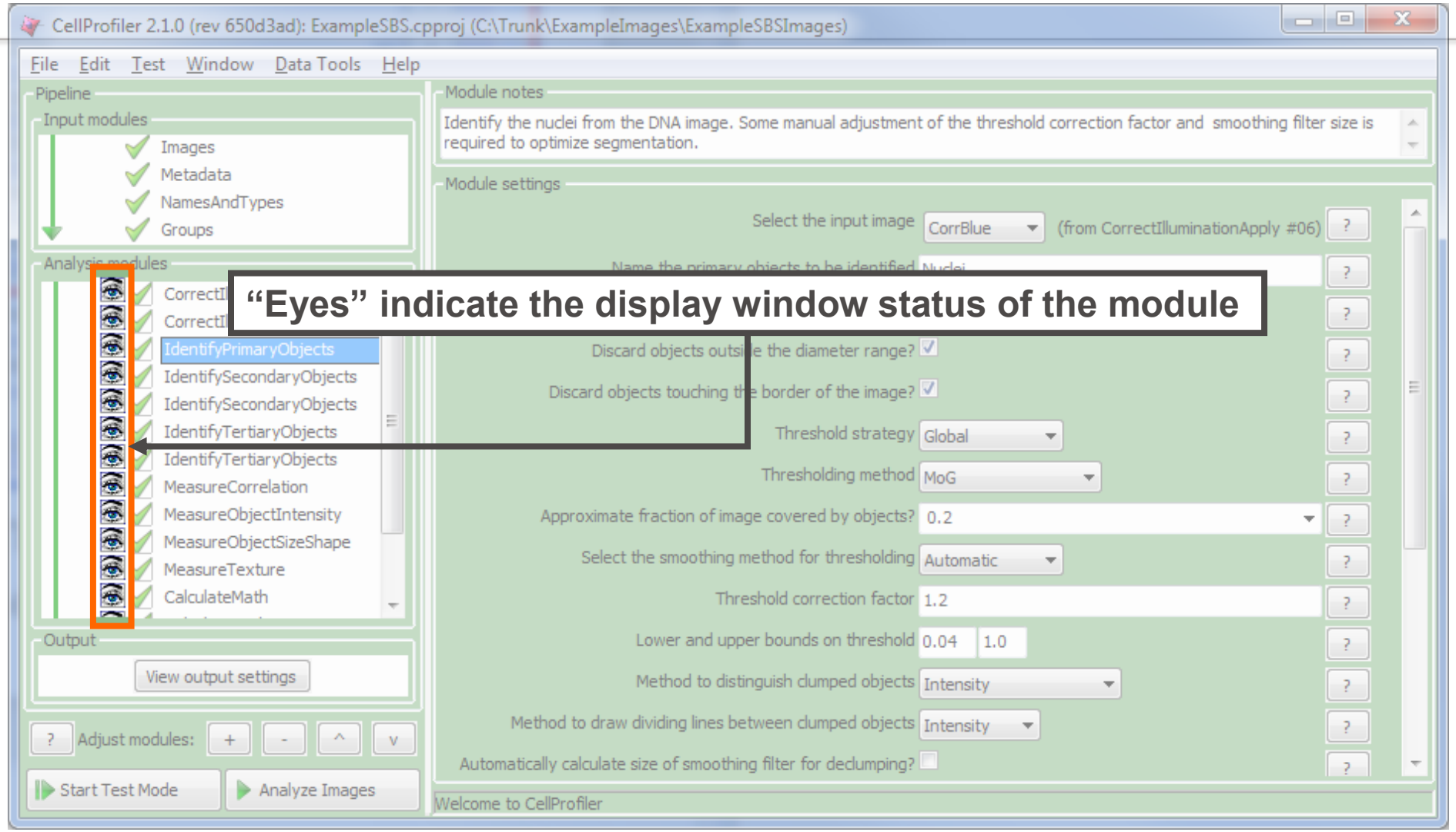
- Saving projects
  - Saves the pipeline, image locations, metadata, etc. (.cpproj)
  - Quick start-up, but not always portable

vs.

- Exporting pipelines
  - Saves just the pipeline (.cppipe)
  - More portable

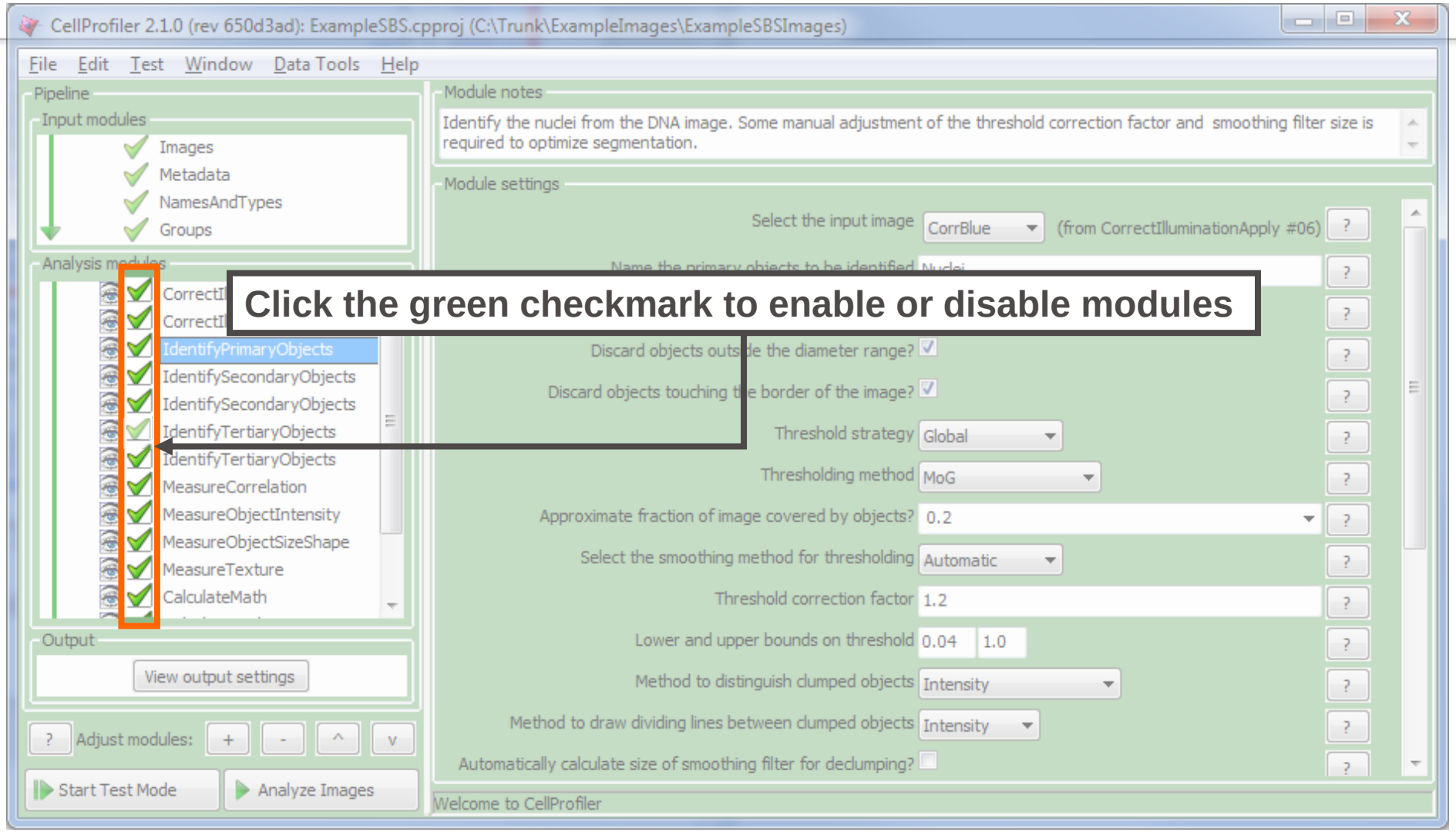
- If publishing, consider submitting your pipeline as **supplemental material**

Before The Analysis…



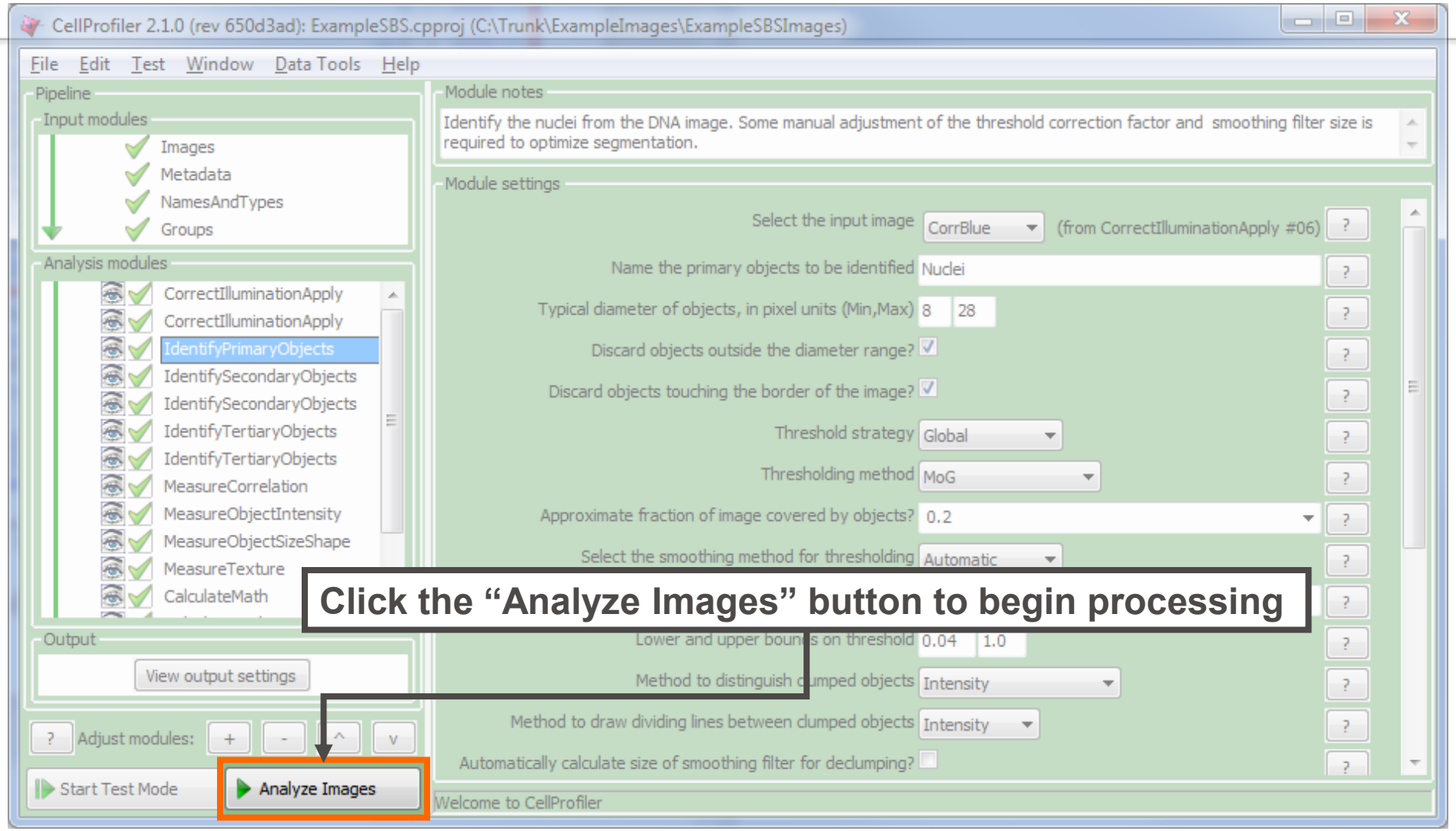
- Close the "eyes" using the Window menu (*Hide all windows on run*)
  - Saves run time and memory usage

Before The Analysis…



- Disabled module is grayed out, effectively removed from the run

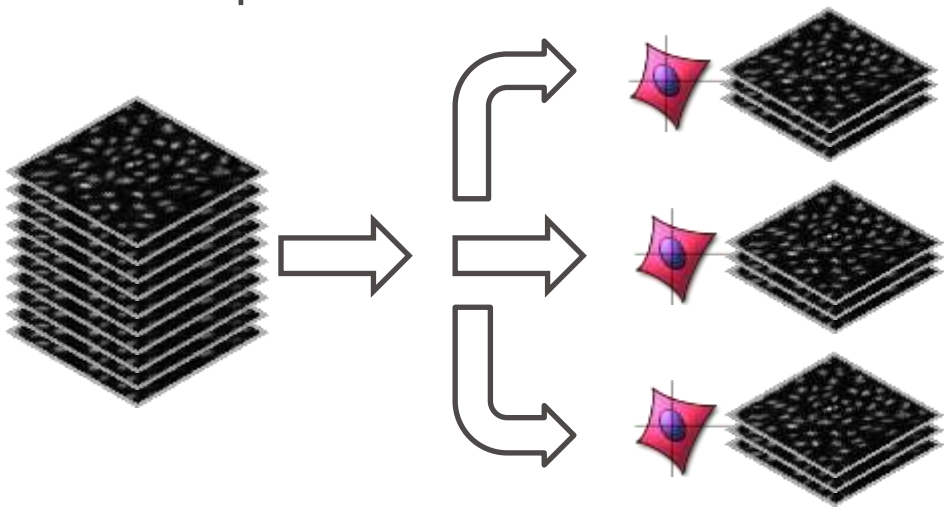
Before The Analysis…



- Analysis can be paused or halted with additional buttons

Multiprocessing in CellProfiler

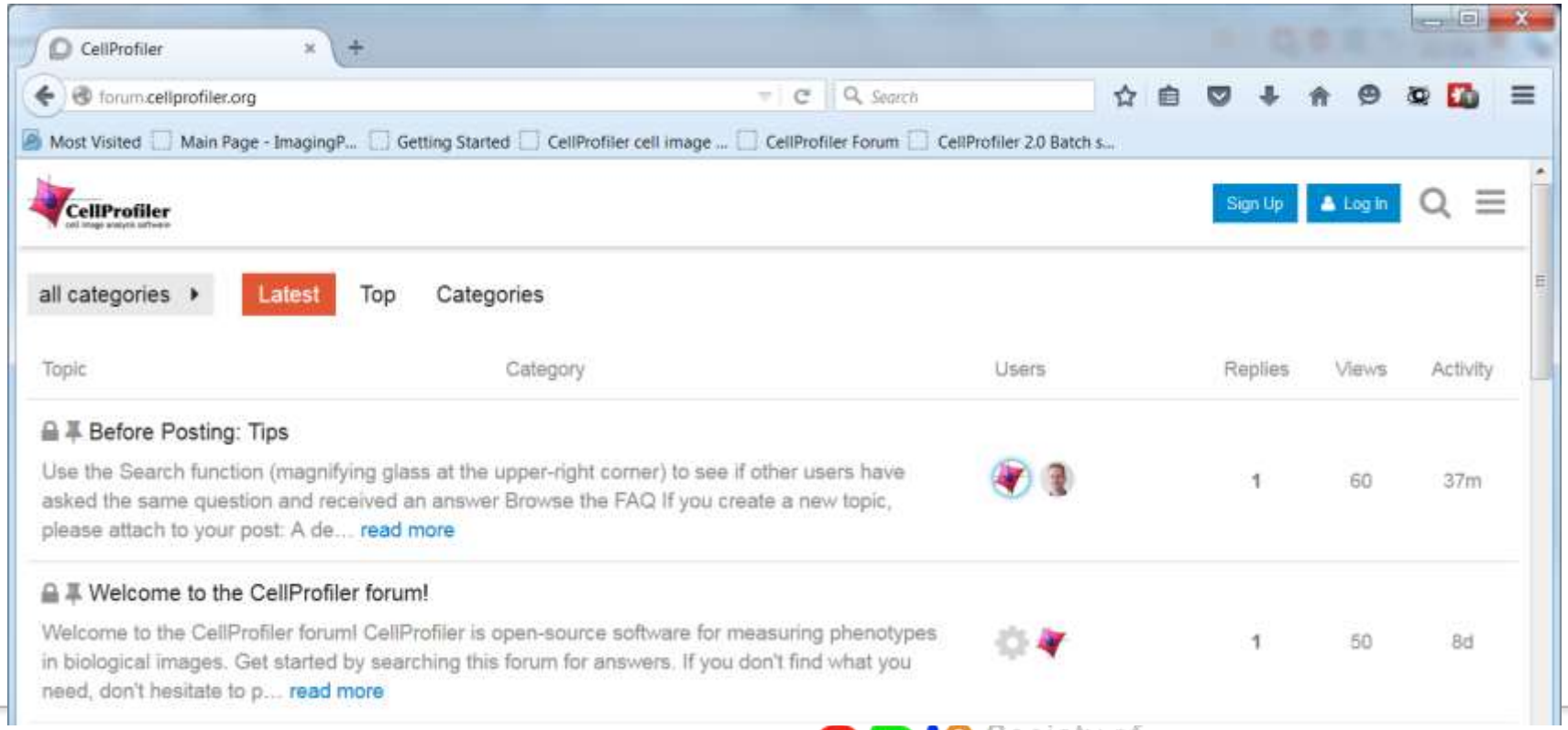
- The more CPUs your computer has, the better…
- Default behavior
  - Number of workers = Number of computing cores
  - Without the GUI (“headless”): One worker, use batch processing to distribute multiple workers



- Number of workers can be set under *File > Preferences*

Final Notes

- Where to get help
  - Access help from the CellProfiler main window
  - Ask for help on the CellProfiler.org forum



Acknowledgments

BROAD INSTITUTE IMAGING PLATFORM

Funding and support for the CellProfiler project has been provided by:

BROAD INSTITUTE, Lilly, National Institute of General Medical Sciences, AstraZeneca, Massachusetts Institute of Technology, Wellcome Trust, Society for Biomolecular Sciences SBS, MERCK, NOVARTIS, Goldman Partnerships, For Women in Science L'Oréal, U.S. Department of Defense

Free, at www.cellprofiler.org:

CellProfiler™ cell image analysis software

CellProfiler™ Analyst data exploration software

Contact:
imagingadmin@broadinstitute.org

Many thanks to our many biology collaborators who provide images

Recent funding for this work provided by:
NIH NIGMS (Carpenter: R01 GM089652 and Wahlby: R01 GM095672)
The Broad Institute of Harvard and MIT

*The Society of Biomolecular Imaging and Informatics (**SBI2**) is an international community of leaders, scientists, and students promoting technological advancement, discovery, and education to quantitatively interrogate biological models to provide high context information at the cellular level.*

SBI2 Society of Biomolecular Imaging and Informatics