Image content-based searching and tools for image-derived modeling

Robert F Murphy

Ray & Stephanie Lane Professor of Computational Biology and
Professor of Biological Sciences, Biomedical Engineering and Machine Learning
External Senior Fellow, Albert Ludwig University of Freiburg

RAY AND STEPHANIE LANE
Center for Computational Biology
Carnegie Mellon

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

FRIAS
FREIBURG INSTITUTE
FOR ADVANCED STUDIES

LIFE SCIENCES - LIFENET

Content-based image search

- Also called Query-by-image-content (QBIC)
- Find images whose content, as reflected by image features, is similar to one or more query images
- Can use positive and/or negative examples
- Can be iterative (relevance feedback)

Why content-based search?

- In compound screening, to see whether anyone else in the organization (or the world) has seen a particular pattern before
- In proteomics, to find other proteins that show a subcellular location that is difficult to assign

nature methods
Techniques for life scientists and chemists
Login
Search Go Advanced search
nature.com ▸ journal home ▸ archive ▸ issue ▸ correspondence ▸ abstract

NATURE METHODS | CORRESPONDENCE

OMERO.searcher: content-based image search for microscope images

Baek Hwan Cho, Ivan Cao-Berg, Jennifer Ann Bakal & Robert F Murphy

Affiliations | Corresponding author

Nature Methods **9**, 633–634 (2012) | doi:10.1038/nmeth.2086
Published online 28 June 2012

Full text | Download PDF | Citation | Reprints | Rights & permissions | Metrics

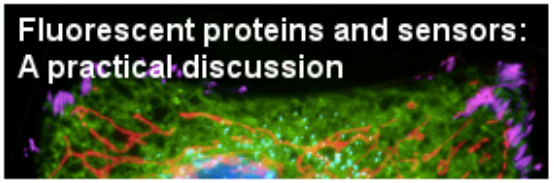
To the Editor:

Fluorescence microscopy is growing dramatically both in terms of technical capabilities and the volume of images generated. Online repositories have been created to provide public access to images and opportunities for joint research for many scientists[1]. This has reintroduced challenges faced when sequence and structure databases were being established: developing fast and

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

nature
الطبعة العربية
free Arabic language version of *Nature* magazine & regularly updated website.

Selected feature

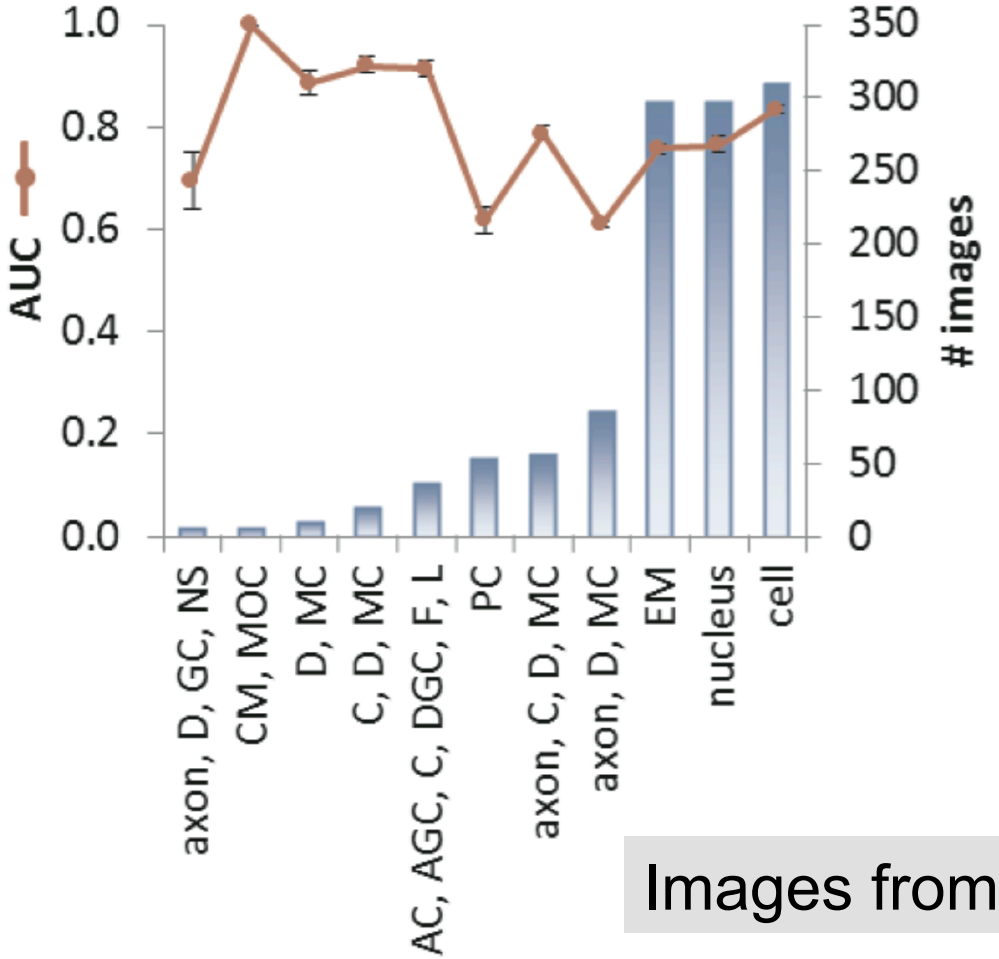


Fluorescent proteins and sensors webinar
Robert Campbell, David Piston and Thomas Knopfel
Fluorescent proteins are invaluable tools for fluorescent microscopy in the life sciences but researchers still have practical questions about

Testing search effectiveness

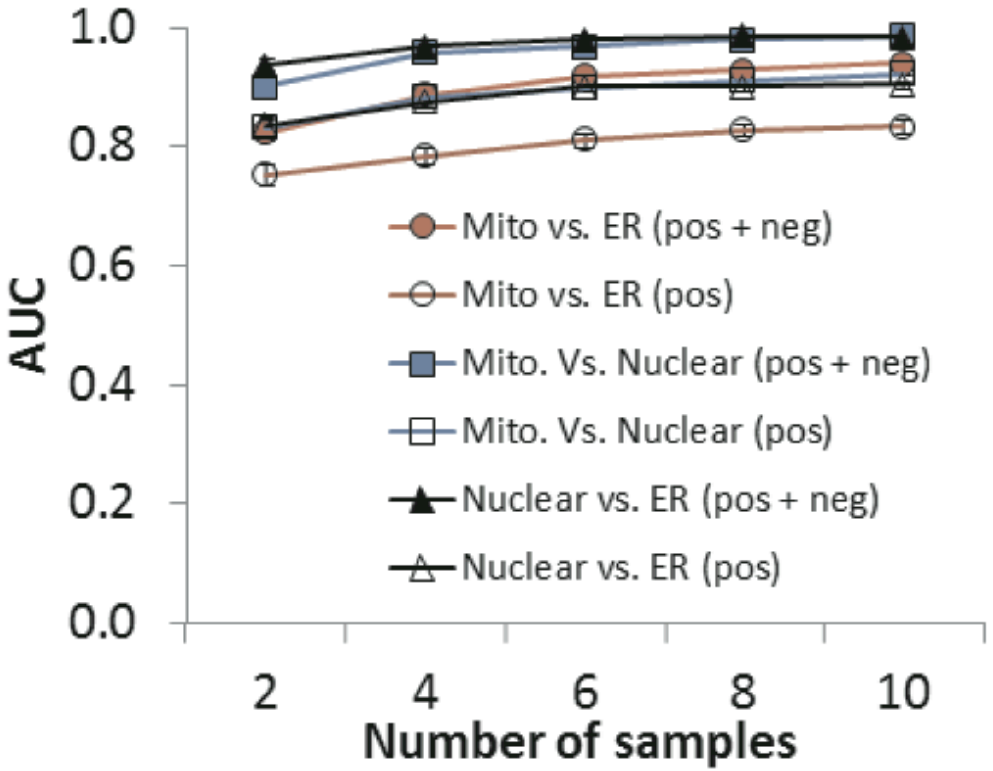
- Combined sets of images of different patterns into database
- Searched with one or more images of one of the patterns
- Observed how highly ranked images of that pattern are relative to other patterns (measure area under ROC curve)

Results for images from diverse image sources



Images from The Cell library

Results for different patterns from same source



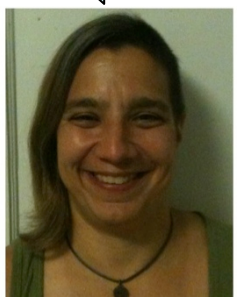
Images from RandTag

OMERO.searcher

- Open source content-based image database search
- Three ways to use:
  1. Server – install on top of your OMERO db
  2. External Search – search via web page
  3. Local Client – search via Python script

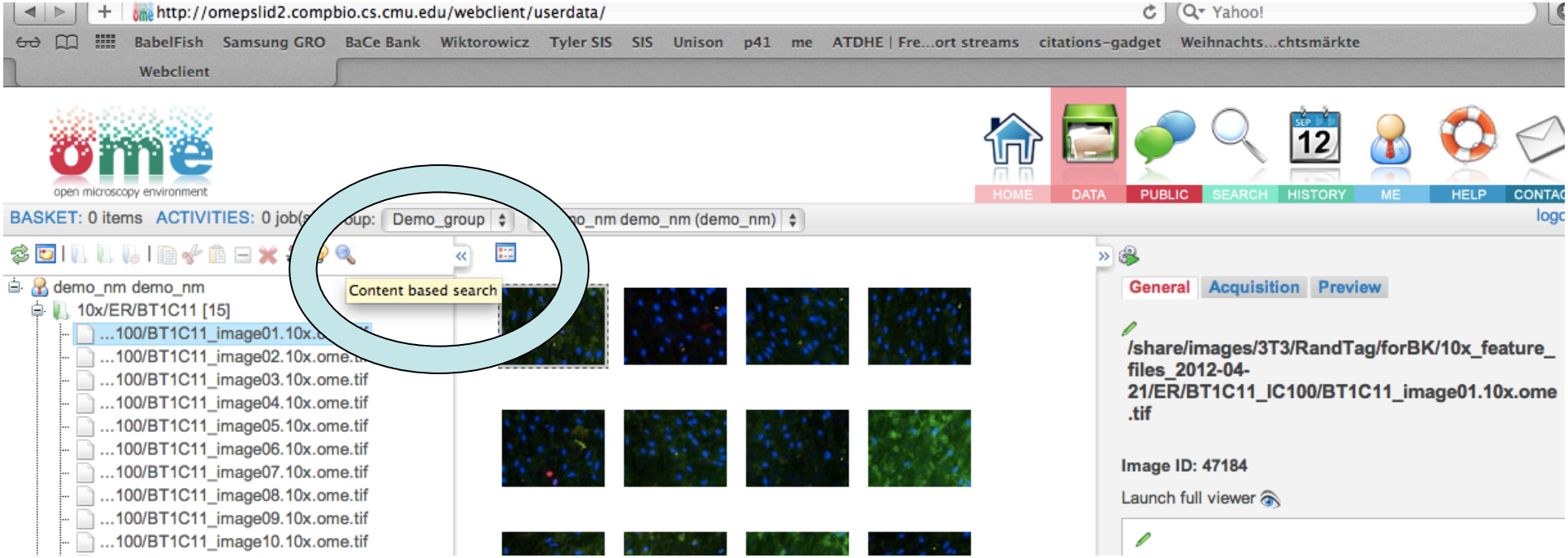
Can search with local images on any computer



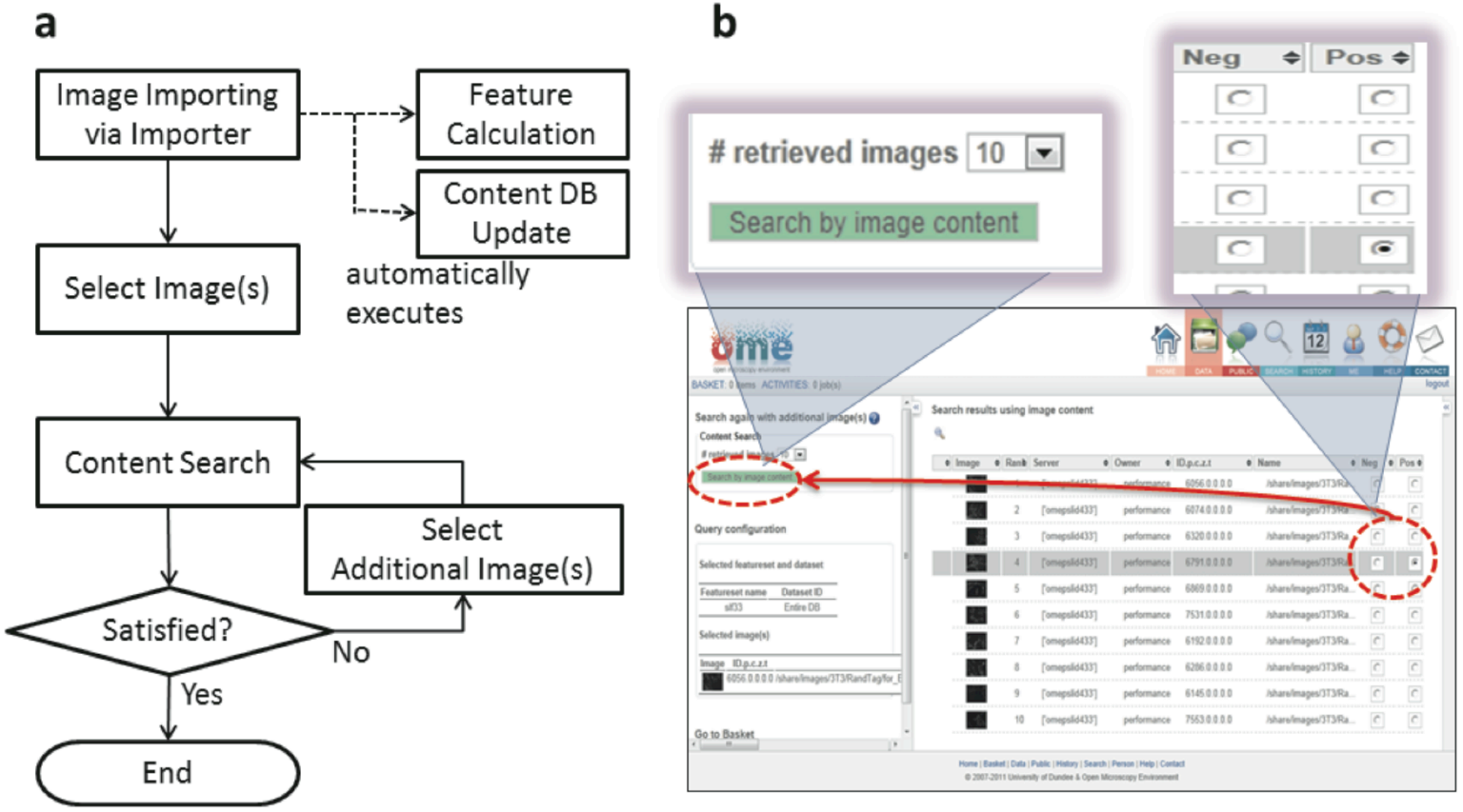


1. OMERO.searcher Server

- Install on top of OMERO 4.3
- Installs feature calculation, content search and changes to OMERO.web

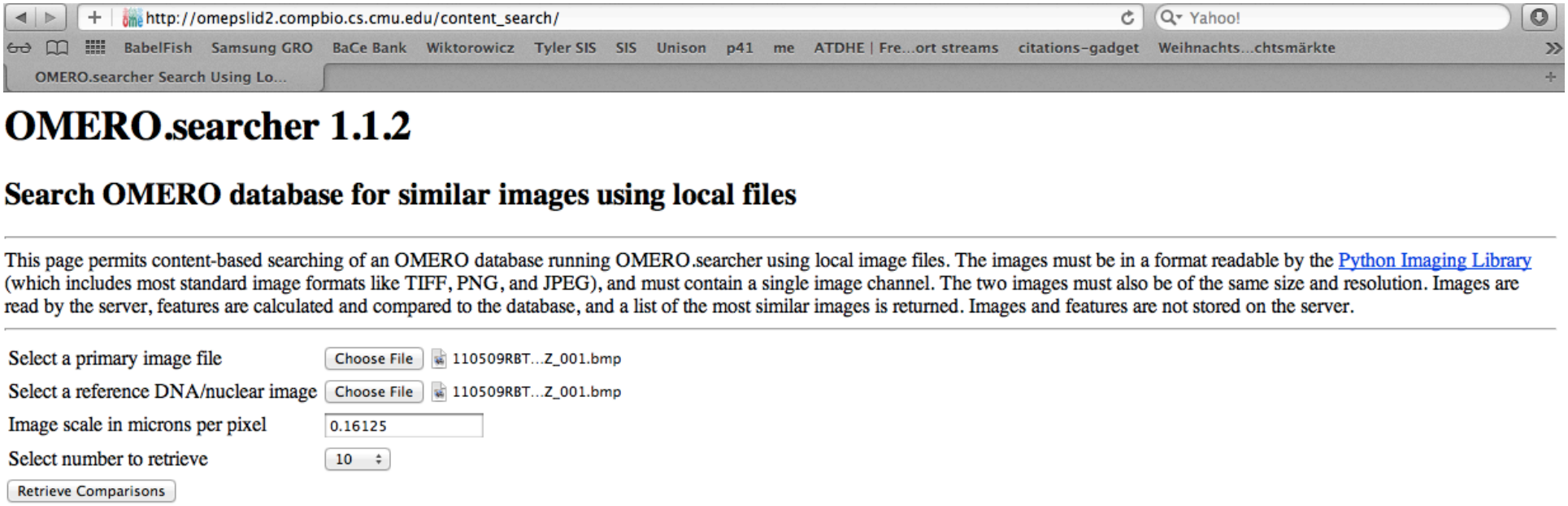


OMERO.searcher workflow

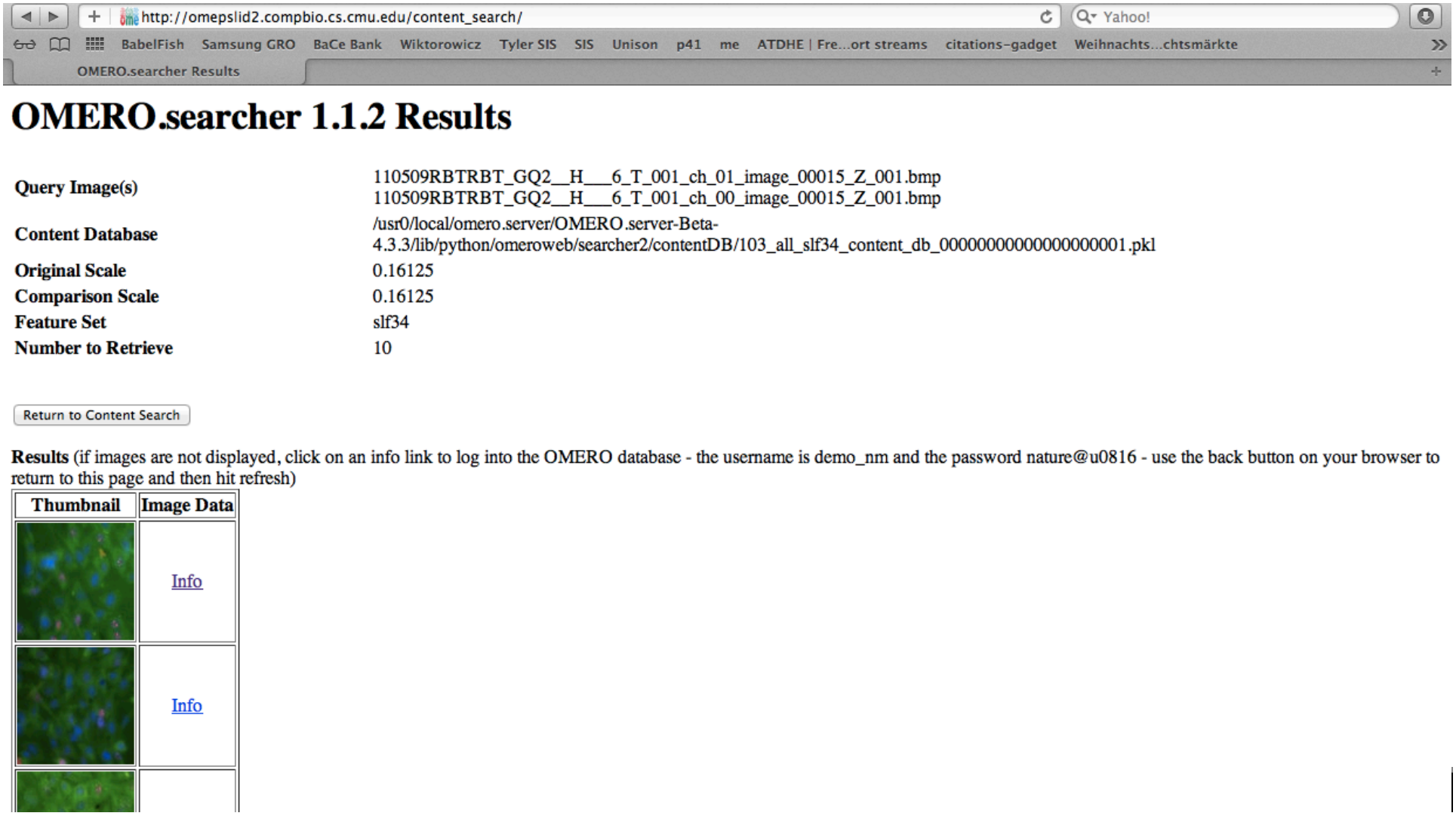


2. External Search page

- Django app installed with Server
- Allows search using external images

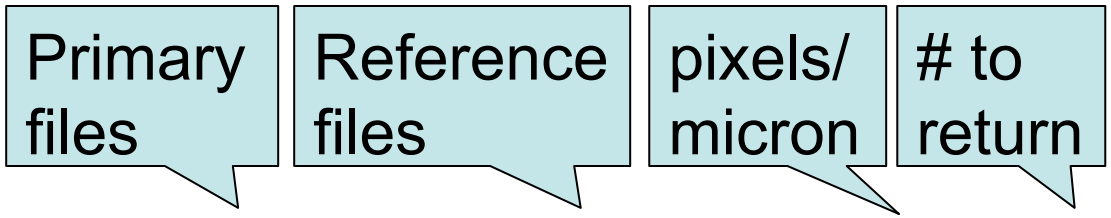


2. External Search results



3. Local Client

- Runs on local computer
- Requires python and packages
- Returns HTML page




```
python omero.searcher.py *c1*.bmp *c0*.bmp 0.2 10
```

Local Client version 1.2 released January 8, 2013

- Supports user-specified feature sets
- Supports non-OMERO databases

User-defined feature sets

- Define a feature set as an ordered set of individual features
- Use python template to produce a function that calculates desired features

Portable content database

- Contains feature values (for one feature set) for each image in database, with URLs linking to that image
- Native format is python .pkl file
- Support for converting to/from .hd5 files
- Allows embedding of user-defined feature calculation code

Making a database available for search

- If OMERO database, install OMERO.searcher
- If non-OMERO database, create content DB file containing features and URLs
- In either case, register the content DB so that other users can access, or let people know URL for database in some other manner

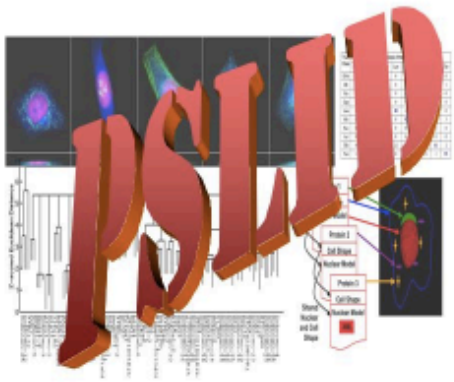
Local Client workflow

- User specifies database URL, feature set name, query images
- Program
  - looks for content DB for specified DB (and checks if it is latest version)
  - runs feature calculation code on the query images (if features not already calculated)
  - produces HTML file showing results of search

Publicly searchable databases

- Released
  - Nature Methods demo1
  - **PSLID RandTag Release 2 (new)**
    - http://pslid.org
    - 6099 CD-tagged clones, 99955 HTM images, 2777 confocal images, 2937 genomic sequences
    - Identified 116 tagged proteins in 189 clones with high confidence
    - Identified additional 70 tagged proteins in 133 clones with medium confidence

http://pslid.org



Protein Subcellular Location Image Database

PSLID - Protein Subcellular Location Image Database

PSLID is both a web service providing access to collections of images showing subcellular location patterns and an open-source image database software system. There are two main collections: the original Public database containing various small image sets useful for training and testing various automated analysis approaches, and the RandTag database containing images of NIH 3T3 clones expressing randomly-tagged proteins.

Go to PSLID RandTag Release 2 (**released January 9, 2013**)

Go to the PSLID Public Database Release 4 (released August 1, 2008)

Download datasets contained within PSLID

Download SLIC feature calculation and pattern analysis software used in PSLID and full PSLID software

Go to PSLID RandTag Release 1 (released July 1, 2009)

Supported by the National Institute of General Medical Sciences, Grant GM075205

Carnegie Mellon

RAY AND STEPHANIE LANE
Center for Computational Biology

Center for Bioimage Informatics
from image to knowledge

Department of Biological Sciences

actin related protein 2/3 complex, subunit 2

- **Gene:** ARPC2_MOUSE Ensembl
- **Location from external databases:**

Uniprot	MGI	Locate
cell leading edge;focal adhesion;cell projection;arp2/3 protein complex	golgi apparatus;cytoplasm;cell leading edge;focal adhesion;arp2/3 protein complex;cytoskeleton;cell projection	none found

- **Location from image classification:**

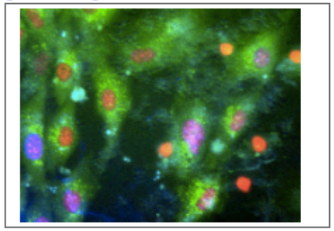
Major(minor)
cytoplasm (membrane, no_apparent_fluorescence); (cytoplasm, membrane, no_apparent_fluorescence)

- **Current Curated Location:** none listed
  Current Proposed Location(s): none listed
  Proposed Location:
  email address:
  Submit proposed location

show:
most typical
brightest by dna
brightest by protein
image

EM1D8_IC100

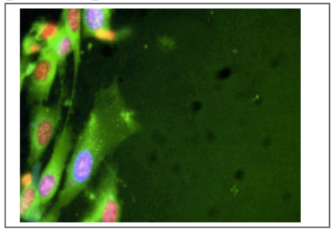
set preview set details
sequence information
gene assignment confidence: Low



- **Location from image classification:** Major class(es): cytoplasm Minor class(es): membrane, no_apparent_fluorescence
- **Cell name:** 3T3-CDtag-EM1D8
- **Experiment title:** 3T3-CDtag-EM1
- **Download image files:**

FX1B4_IC100

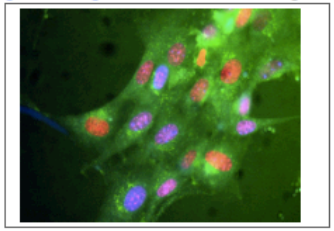
set preview set details
sequence information
gene assignment confidence: Low



- **Location from image classification:** Major class(es): cytoplasm Minor class(es): membrane, no_apparent_fluorescence
- **Cell name:** 3T3-CDtag-FX1B4
- **Experiment title:** 3T3-CDtag-FX1
- **Download image files:**

FX1E9_IC100

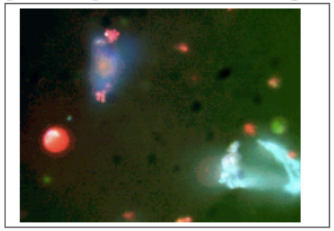
set preview set details
sequence information
gene assignment confidence: High



- **Location from image classification:** Major class(es): cytoplasm Minor class(es): membrane, no_apparent_fluorescence
- **Cell name:** 3T3-CDtag-FX1E9
- **Experiment title:** 3T3-CDtag-FX1
- **Download image files:**

FX2B3_IC100

set preview set details
sequence information
gene assignment confidence: High



- **Location from image classification:** Major class(es): Minor class(es): cytoplasm, membrane, no_apparent_fluorescence
- **Cell name:** 3T3-CDtag-FX2B3
- **Experiment title:** 3T3-CDtag-FX2
- **Download image files:**

URL-based searching of PSLID

- Search by target or clone
- Return HTML or XML
- Can download images matching a query as a ZIP file of OME-TIFF images

Planned publicly searchable databases

- The Cell library FM Subset1 (should be released next week)
- Human Protein Atlas confocal images

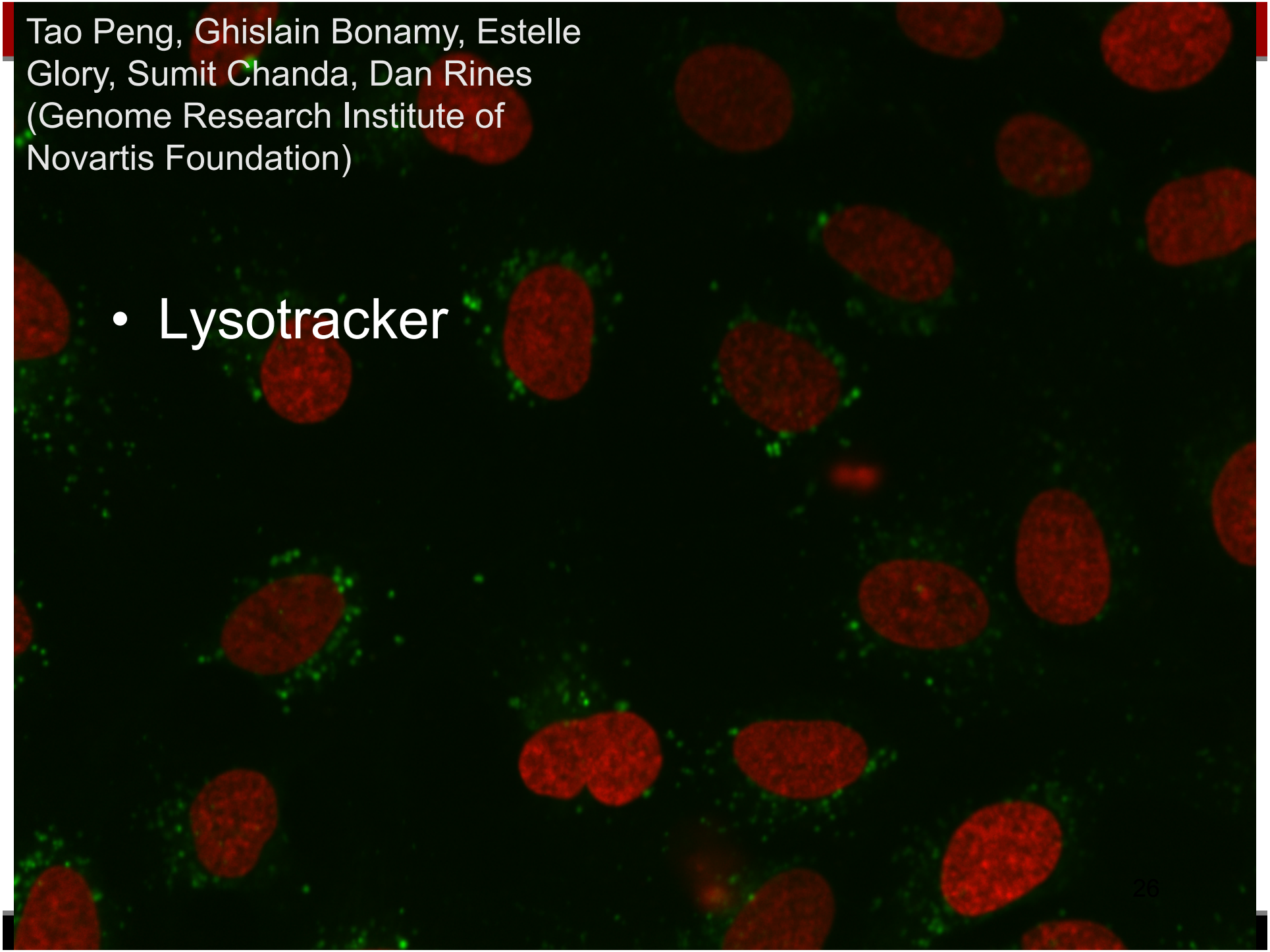
OMERO.searcher plans

- Support for
  - OMERO 4.4 (done?)
  - Search across multiple servers
  - BioFormats for reading images
  - Other feature sets
  - Combining content and context for search
  - Integration into base OMERO system
- Additional OMERO databases available for public search

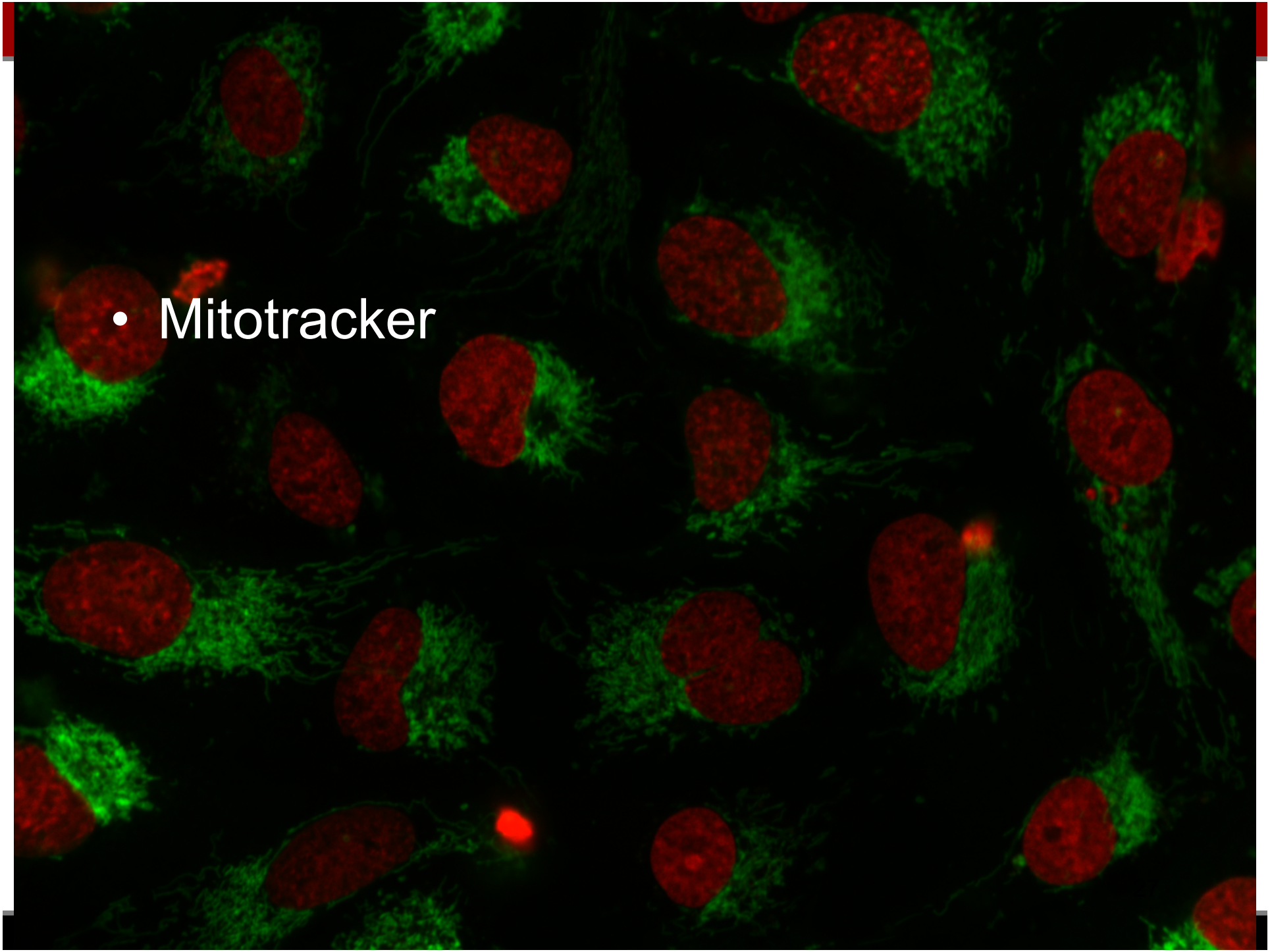
PatternUnmixer

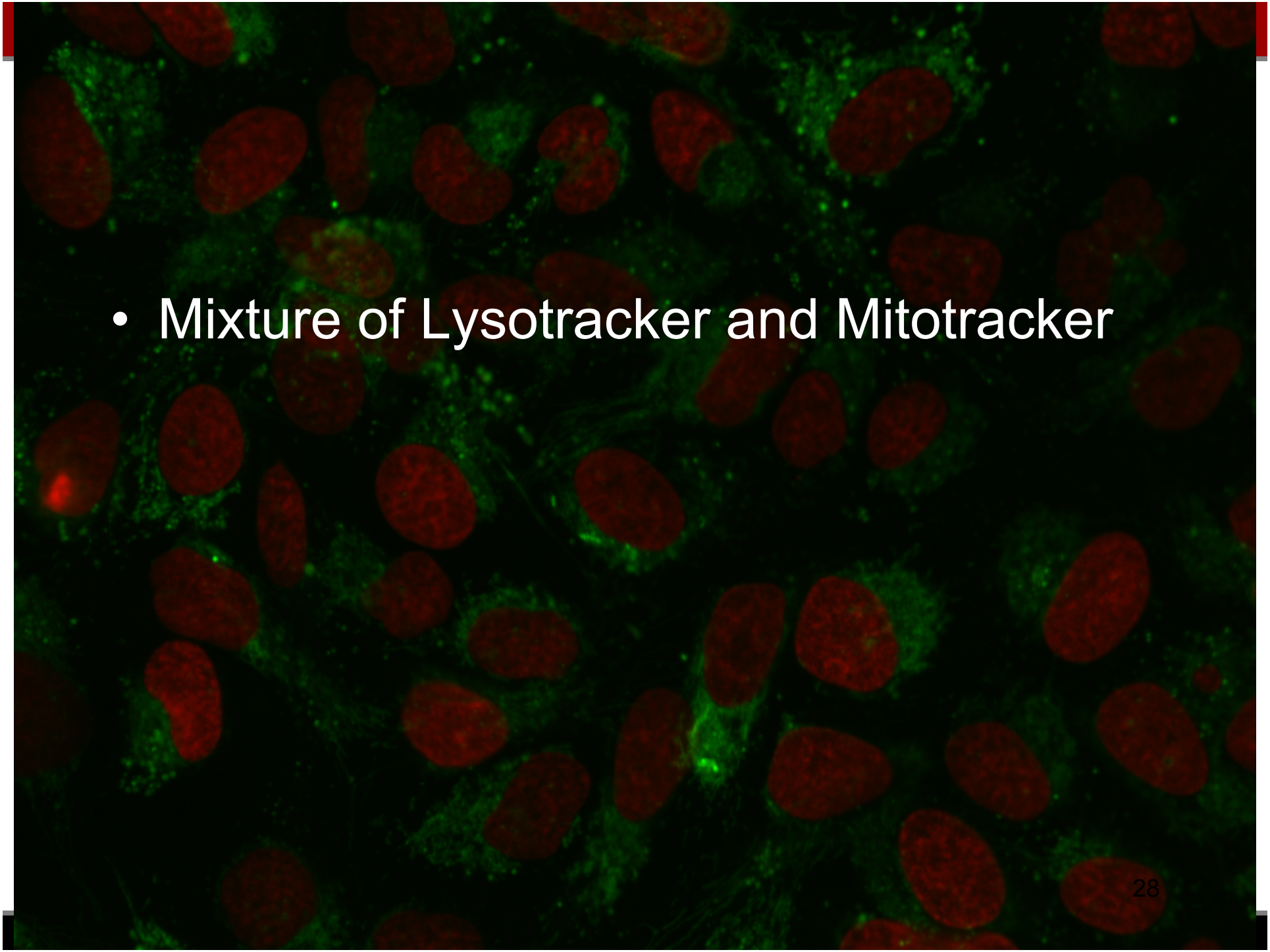
- Performs supervised unmixing of subcellular location
- Versions
  - Matlab source
- Supports training on or unmixing images from local disk or multiple OMERO servers

Tao Peng, Ghislain Bonamy, Estelle Glory, Sumit Chanda, Dan Rines (Genome Research Institute of Novartis Foundation)



- Lysotracker


• Mitotracker

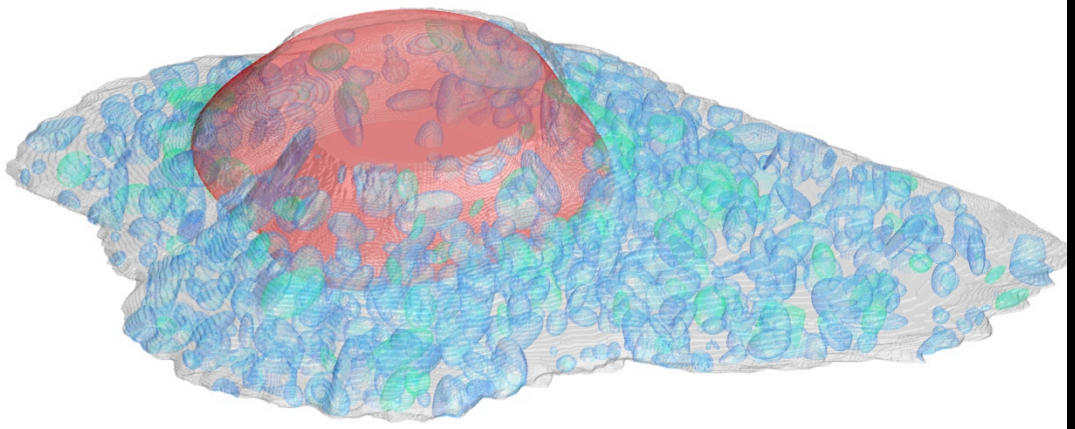

• Mixture of Lysotracker and Mitotracker

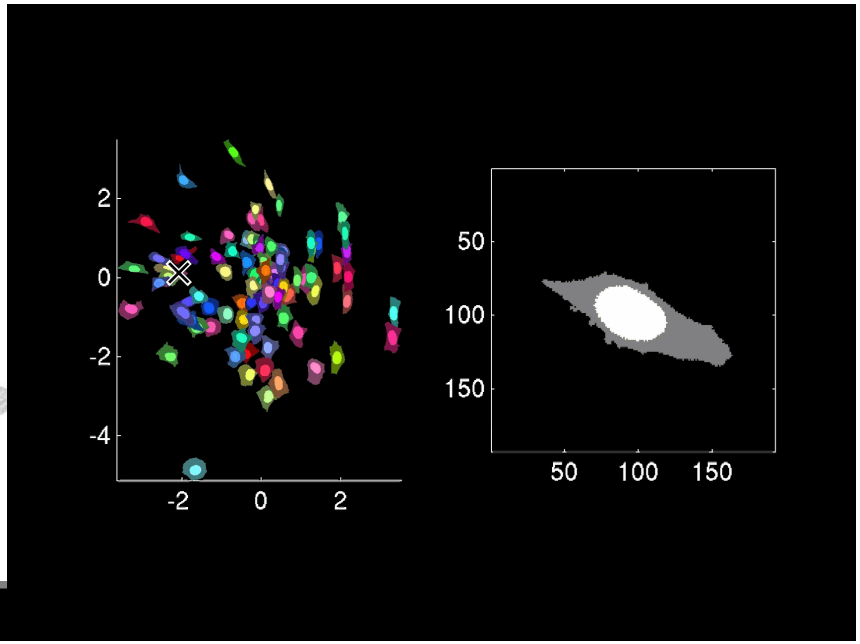
PatternUnmixer plans

- Django app running on server
  - No requirements on local computer
- Heatmap displays and hit finders
- Other local feature detectors
- Add unsupervised unmixing (have code already)

CellOrganizer

- Open-source system for learning generative models of cell organization
- Version 1.8 released tomorrow
  - Includes diffeomorphic models for generation of cell and nuclear shapes





CellOrganizer

Images ⇔ Models

RAY AND STEPHANIE LANE Center for Computational Biology

Carnegie Mellon University

Center for Bioimage Informatics

from image to knowledge

April 13, 2012: Integrated 2D/3D version 1.6 released!

The *CellOrganizer* project provides tools for

- learning generative models of cell organization directly from images
- storing and retrieving those models in XML files
- synthesizing cell images (or other representations) from one or more models

Model learning captures variation among cells in a collection of images. Images used for model learning and instances synthesized from models can be two- or three-dimensional static images or movies.

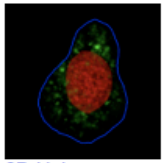
Version 1.6 of *CellOrganizer* can learn models of

- cell shape
- nuclear shape
- chromatin texture
- vesicular organelle size, shape and position
- microtubule distribution.

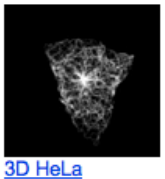
These models can be *conditional* upon each other. For example, for a given synthesized cell instance, organelle position is dependent upon the cell and nuclear shape of that instance.

Cell types for which generative models for at least some organelles have been built include human HeLa cells, mouse NIH 3T3 cells, and Arabidopsis protoplasts. Planned projects include mouse T lymphocytes and rat PC12 cells.

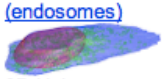
Synthesized Cell Images
(click to view)



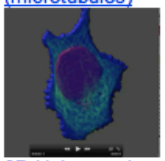
2D HeLa (endosomes)



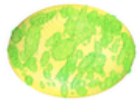
3D HeLa (microtubules)



3D HeLa (mitochondria)



3D HeLa movie



3D protoplast (chloroplasts)

Support for *CellOrganizer* has been provided by grants GM075205 and GM090033 from the National Institute of General Medical Sciences, by a Forschungspreis from the Alexander von Humboldt Foundation, and by the School of Life Sciences of the Freiburg Institute for Advanced Studies.

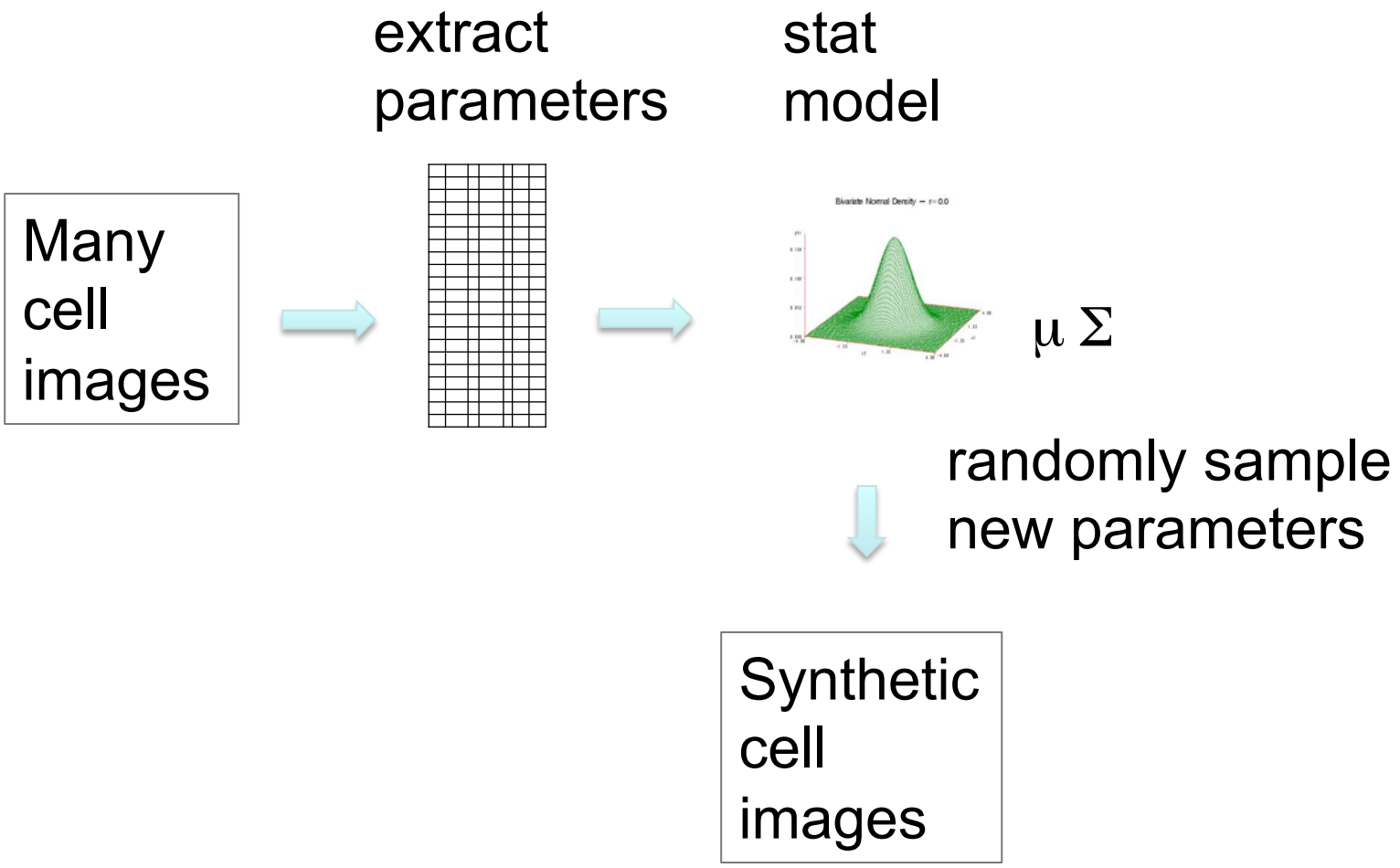
Alexander von Humboldt Stiftung/Foundation

ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

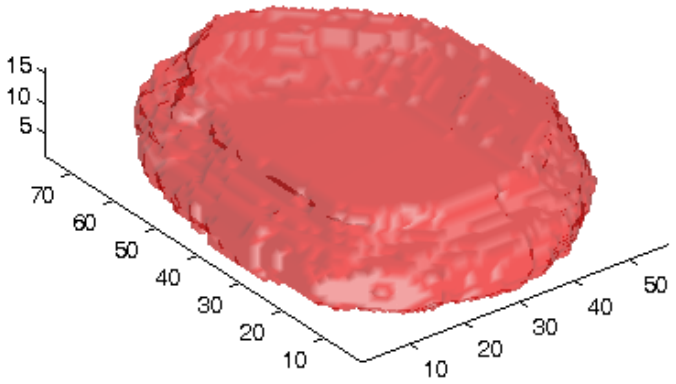
FRIAS FREIBURG INSTITUTE FOR ADVANCED STUDIES

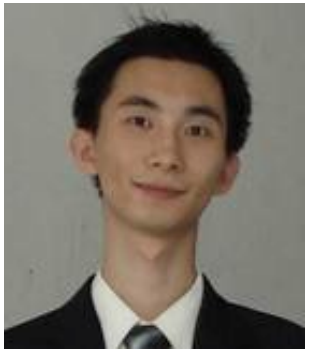
LIFE SCIENCES - LIFENET

Statistical modeling and generation

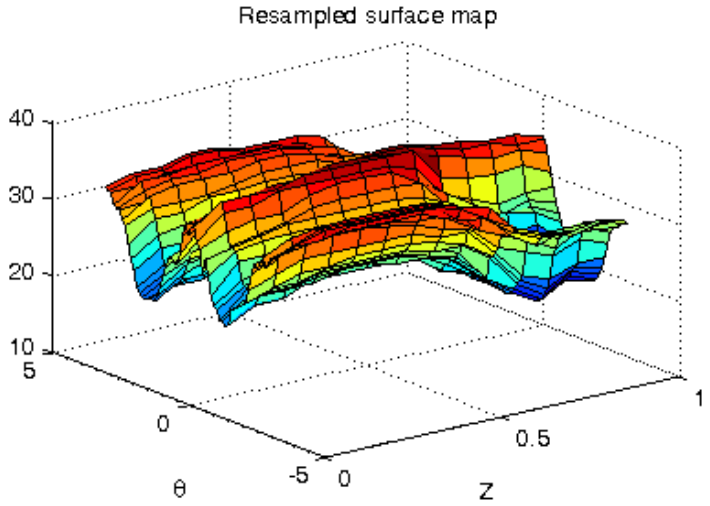


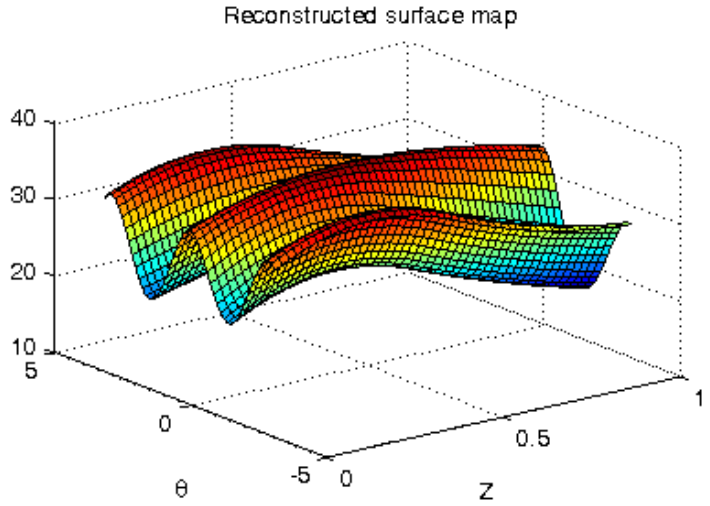
3D Nuclear Shape – Cylindrical Spline Surface





Tao Peng





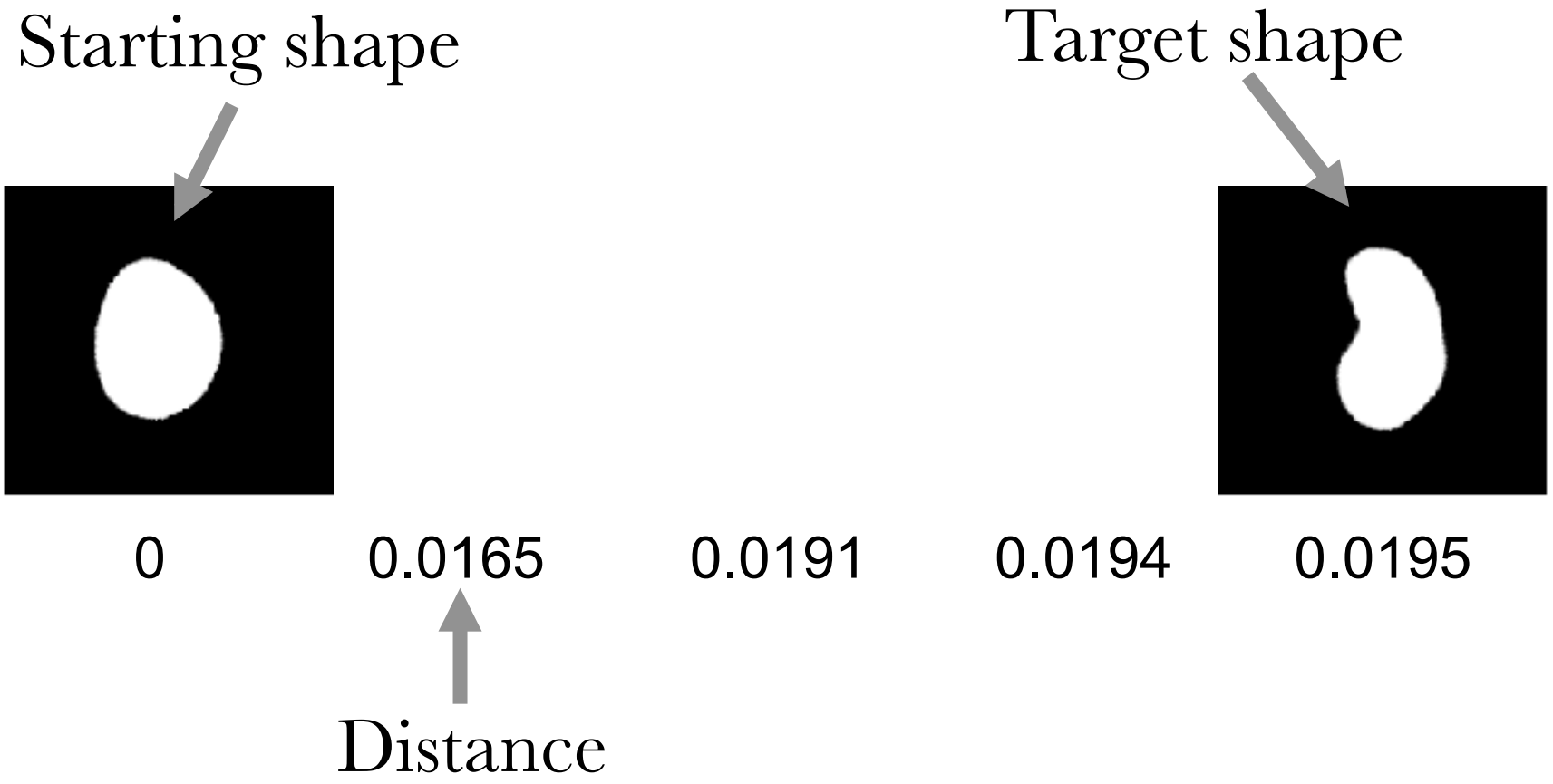
33 parameters (32 spline coefficients + height)

Diffeomorphic analysis of shape

- Sometimes cell or nuclear shapes are irregular
- Can use distance between shapes to characterize shape instead of parameters of model (based on work by Michael Miller and colleagues)

Gustavo Rohde

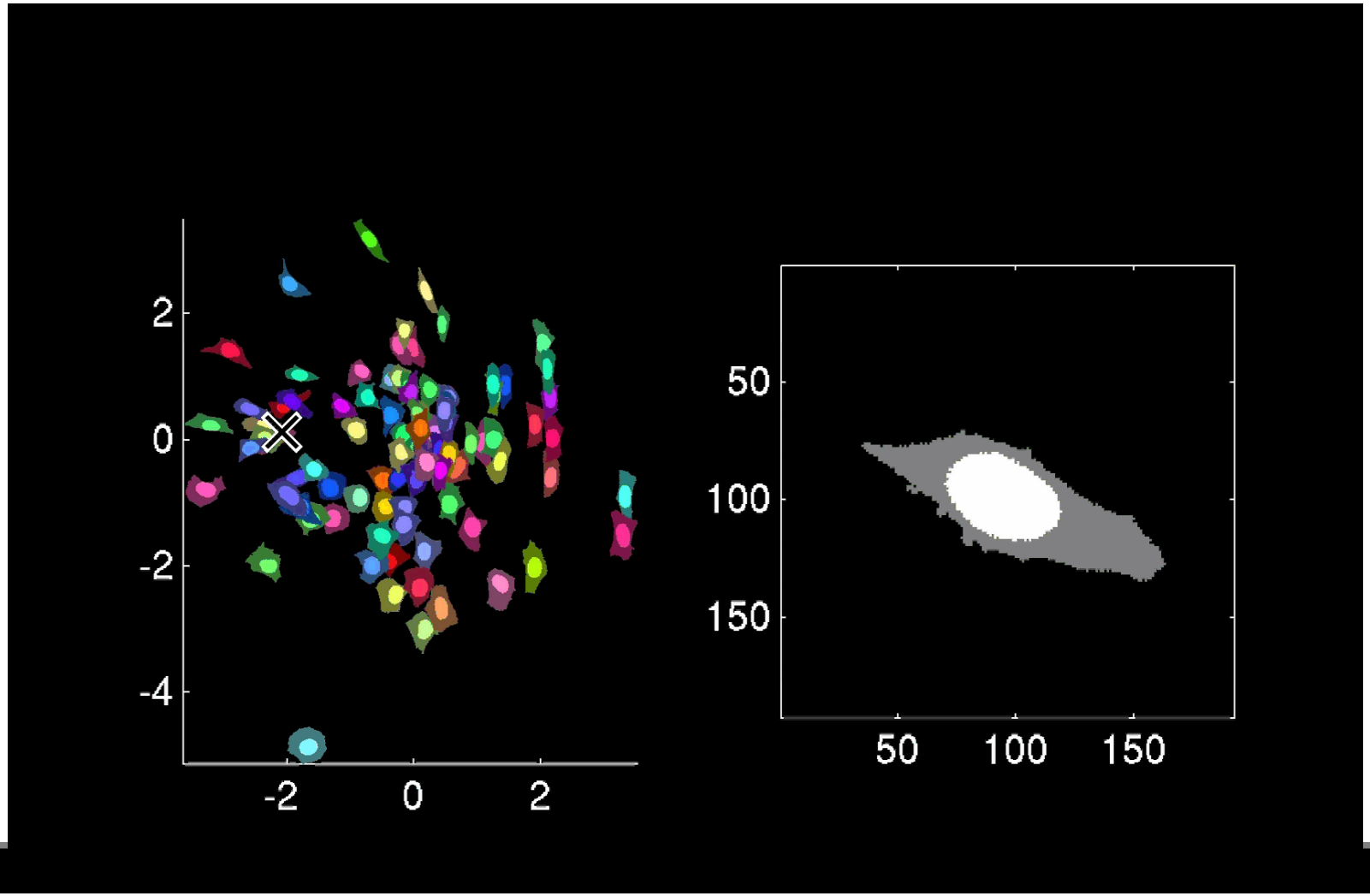
Morphing one shape into another



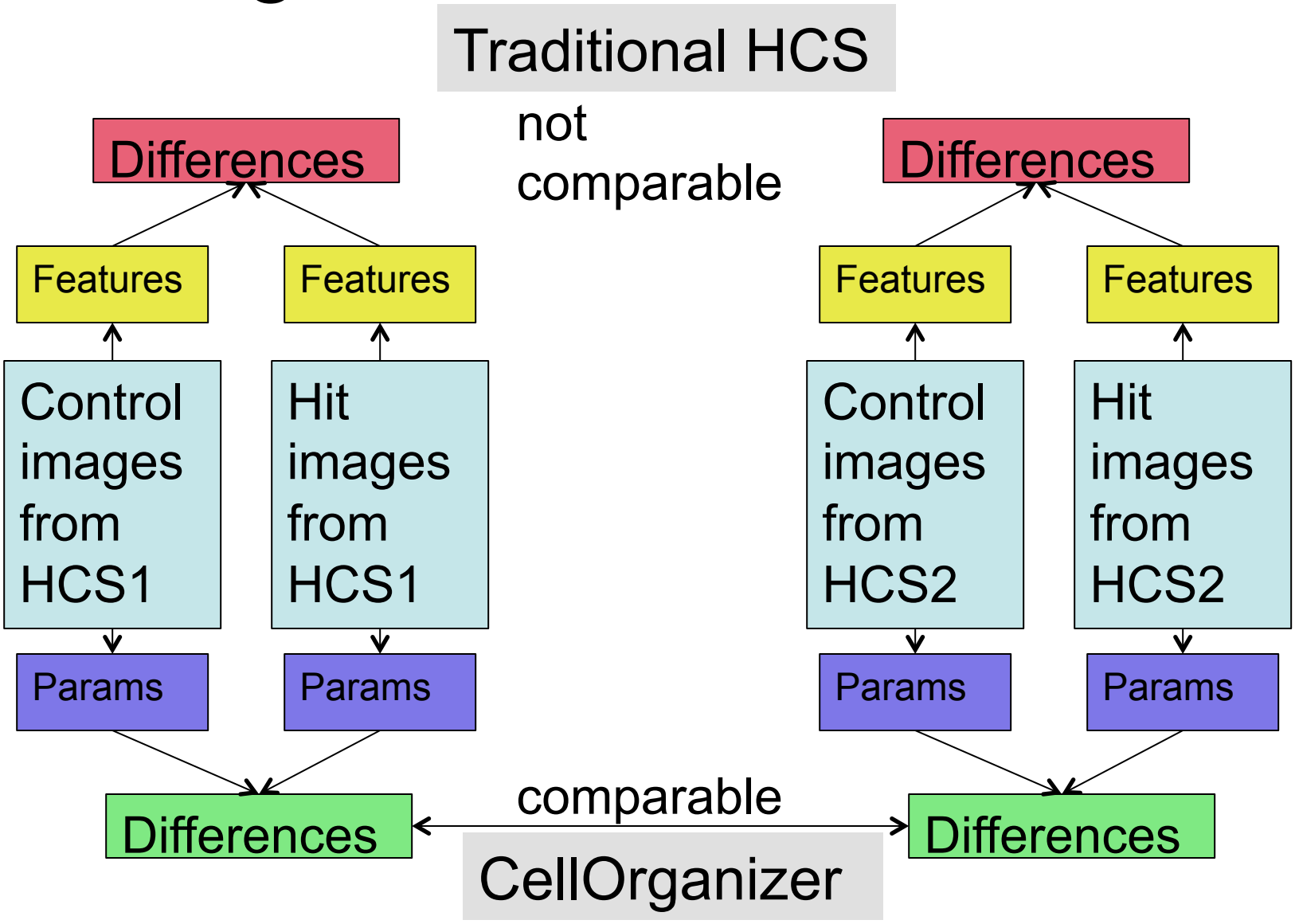
Shape space

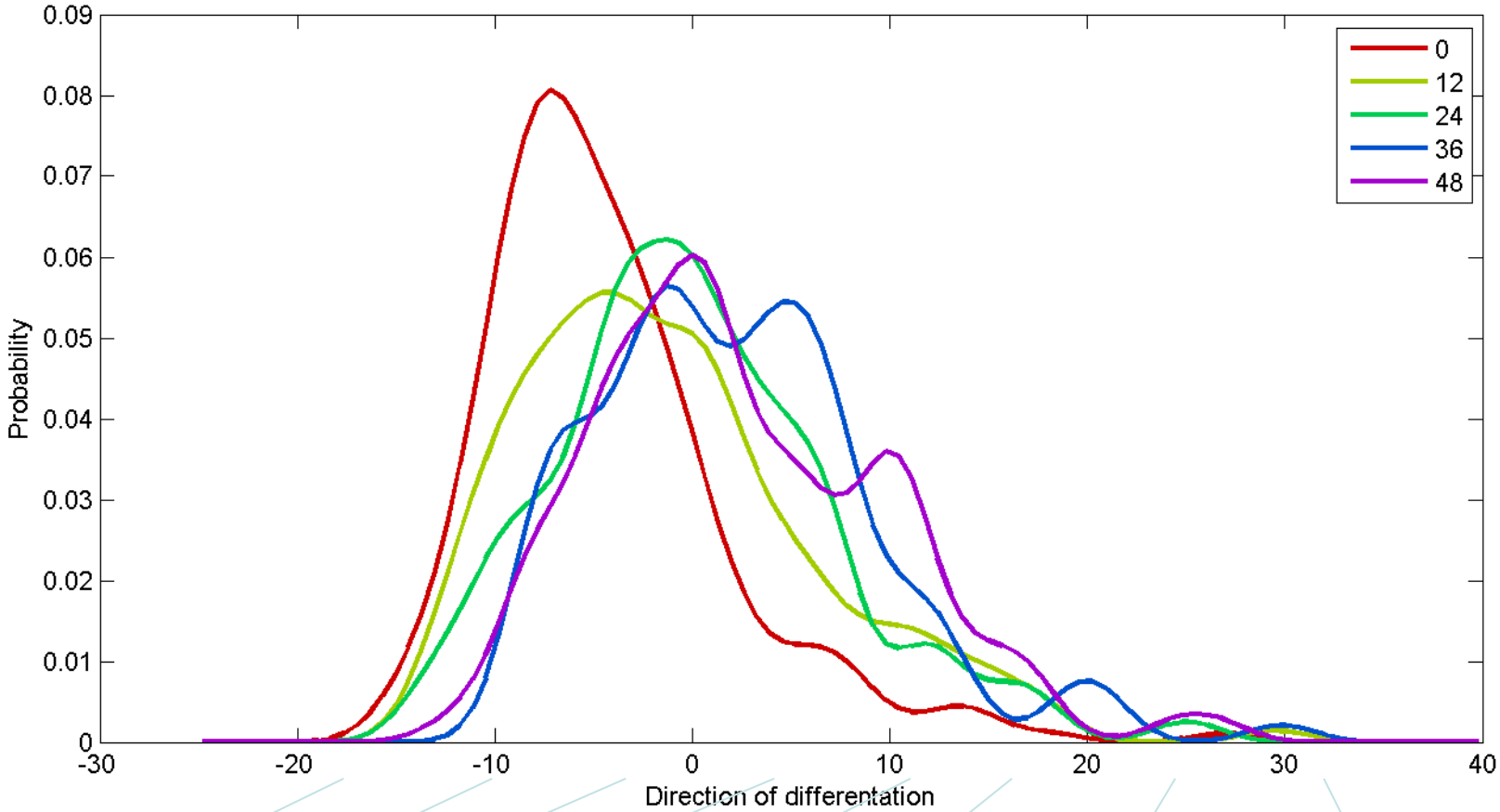
- Can measure distance between all pairs of shapes to construct a “shape space”
- Captures essential aspects of how shapes vary
- Can be applied to nuclei, cells, organelles, etc.
- Can be applied in 2D, 3D, 4D and to more than one component at a time

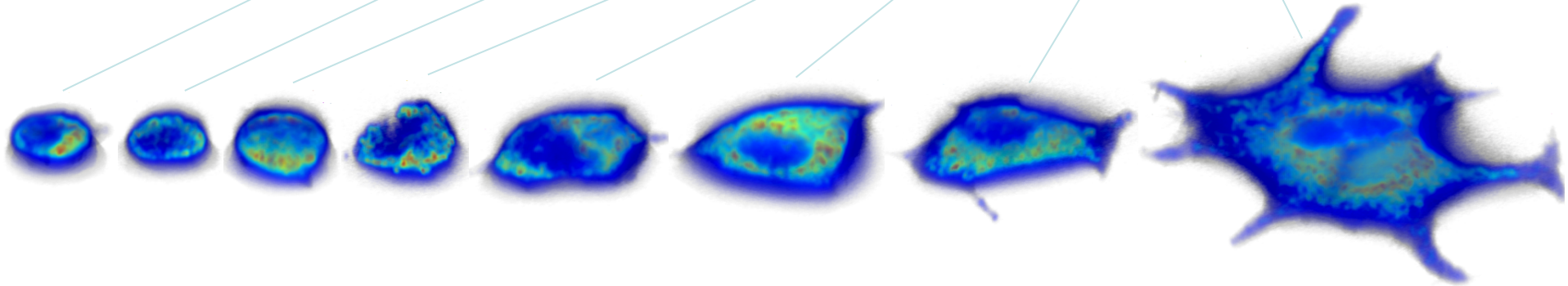
Cell and nuclear shape space



CellOrganizer vs. traditional HCS




0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Probability
-30
-20
-10
0
10
20
30
40
Direction of differentation
0
12
24
36
48



A Modest Proposal

- Register the large collection of cellular images within this consortium through generative models