# Benchmarks and Challenges in Bioimage Analysis

## NEUBIAS 2017 Training School for Bioimage Analysts, February 14, 2017

**Michal Kozubek**

Centre for Biomedical Image Analysis
Faculty of Informatics, Masaryk University, Brno, Czech Republic

http://cbia.fi.muni.cz/

UNIVERSITAS MASARYKIANA BRUNENSIS

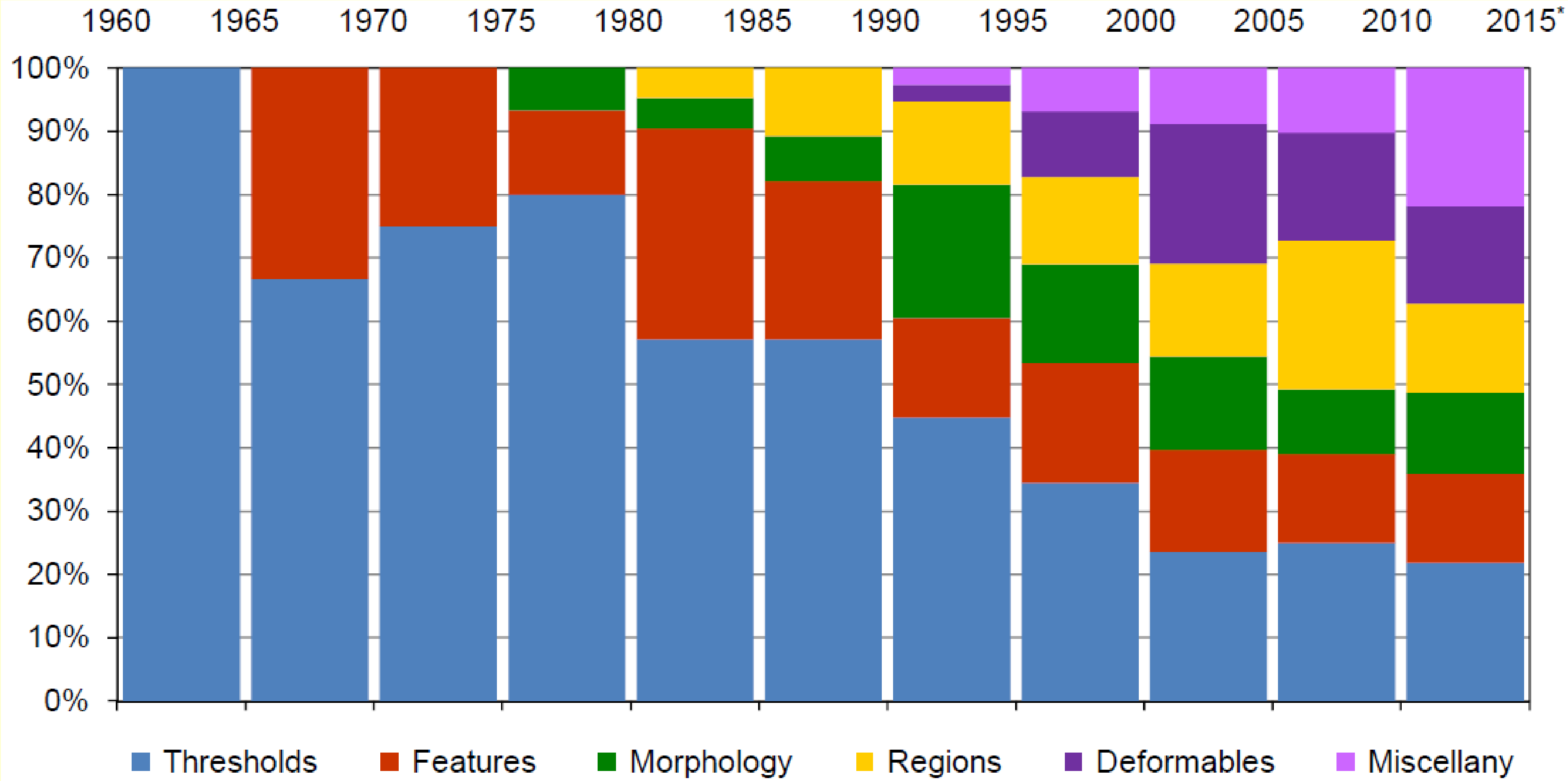
CBIA CENTRE FOR BIOMEDICAL IMAGE ANALYSIS

# Outline

- **Motivation**
  - **Why benchmarking?**
- Design of a benchmark or a challenge
  - Dataset selection
  - Algorithm evaluation / Metrics
- Summary and discussion
  - Future directions
- Existing benchmarks and challenges
  - Overview
  - Particle Tracking Challenge
  - Cell Tracking Challenge
  - Localization Microscopy Challenge (Daniel Sage)

# Computerized Cell Imaging: Observations

- Capabilities of computers have always been worse than capabilities of microscopes
- *Any reasonable solution* was welcome to cope with too complex microscope data
- Cell image analysis have thus frequently used ad hoc “cheap” methods



[Meijering E: Cell Segmentation: 50 Years Down the Road, IEEE SPM 29, 2012]

# From "Any Solution" to "Best Solution"

- The most important problem facing bioimaging
  - **Lack of standards** (reference datasets, algorithm performance measures)
  - Published algorithms often evaluated using ad hoc selected data and metrics
  - Performance of algorithms and software often can not be properly compared!

- History of tackling the problem in image processing

1970s: Stolen controversial image of Lenna becomes a benchmark image for demonstrating the performance of image enhancement algorithms

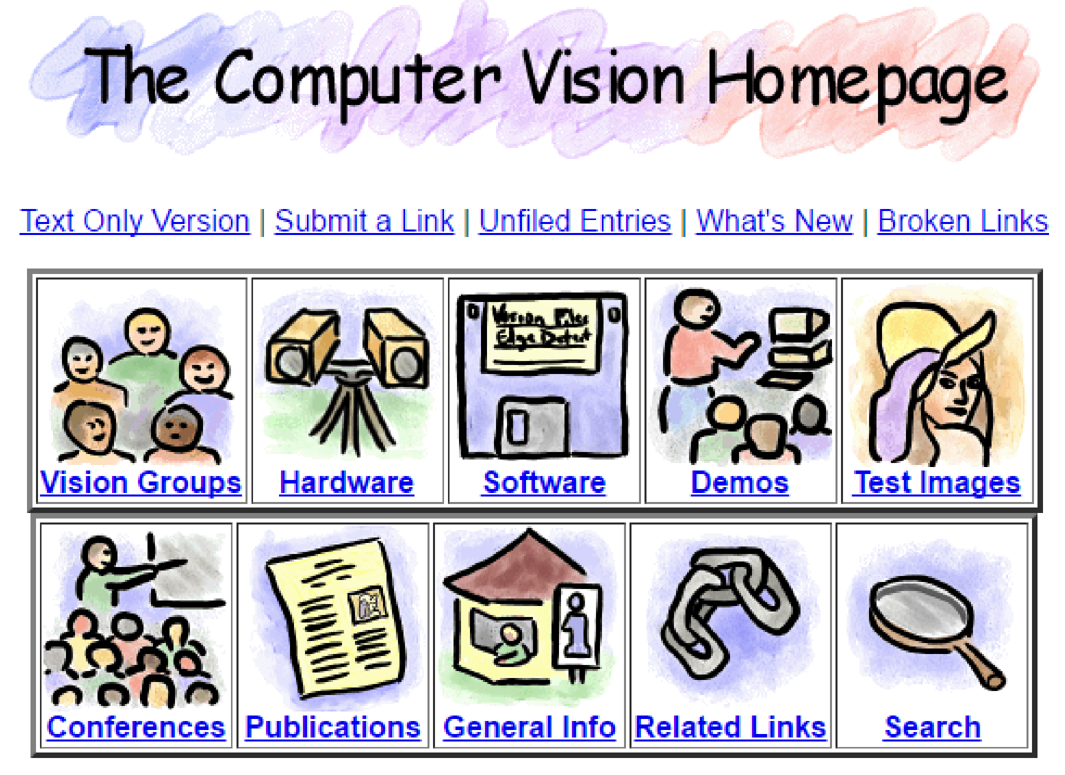
1980s: FTP sites appear with collections of images of specific type (e.g., Iris dataset repository at University of California, Irvine)

1980s: Growing awareness of poor building on previous work of others, Keith Price (University of Southern California): "Anything you can do, I can do better (No you can't)". *Comput Vision Graph* 36:387-391

# From "Any Solution" to "Best Solution"

- History of tackling the problem in the web age

1990s: Web home pages with reference data appear in computer vision community (with known correct solutions – so called **ground truth**)



The Computer Vision Homepage

Text Only Version | Submit a Link | Unfiled Entries | What's New | Broken Links

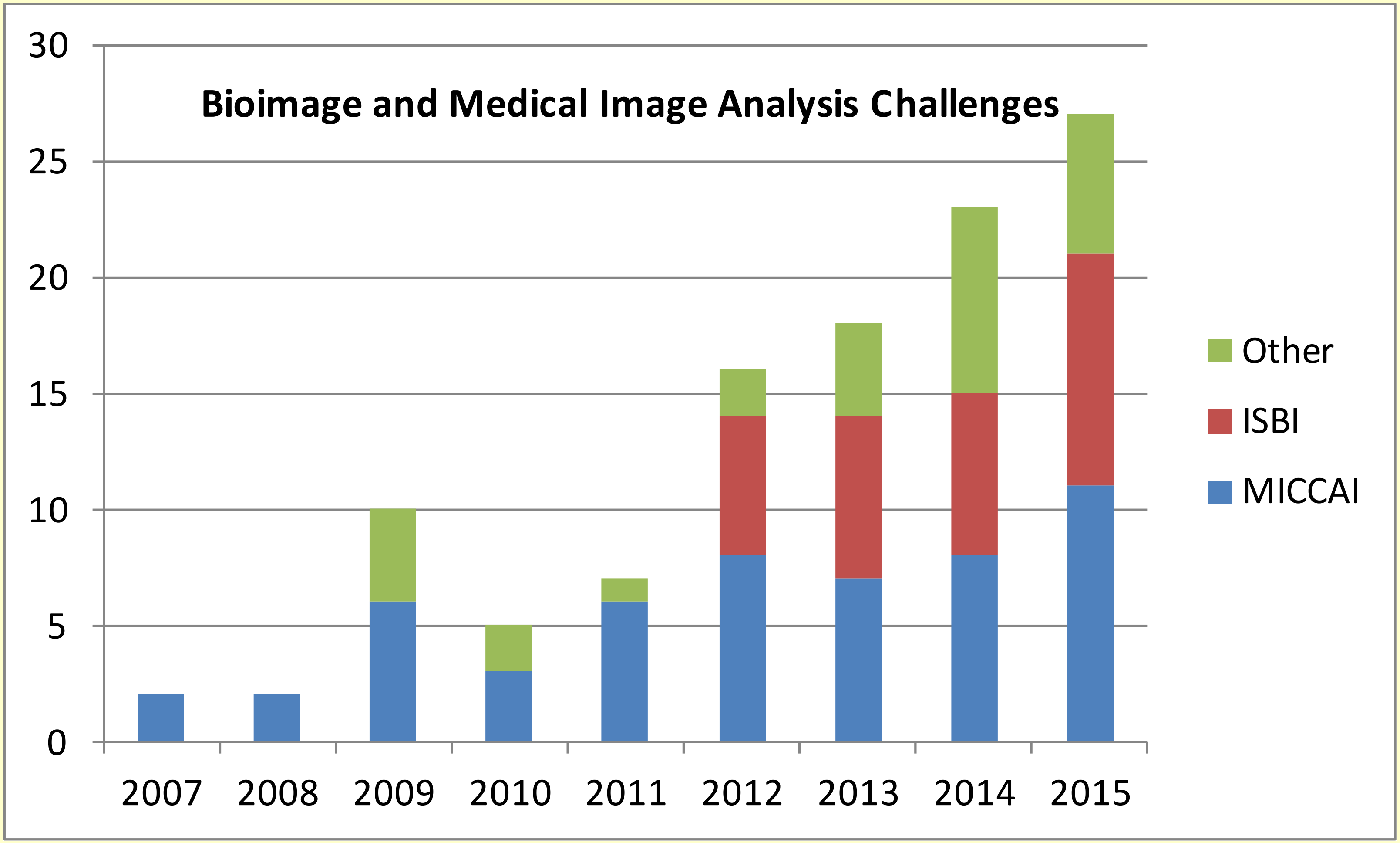
1990s: First benchmarking initiatives also in biomedical imaging – Retrospective Image Registration Experiment (RIRE) at Vanderbilt Univ

2000s: Medical imaging reference datasets appear (real as well as synthetic) with known ground truth (annotated by experts or computer generated)

2010s: Bioimaging (in the sense of biological imaging) reference datasets and **challenges** (benchmarks associated with competitions) appear

# Challenges Organized So Far

## Growing awareness of the need for benchmarking

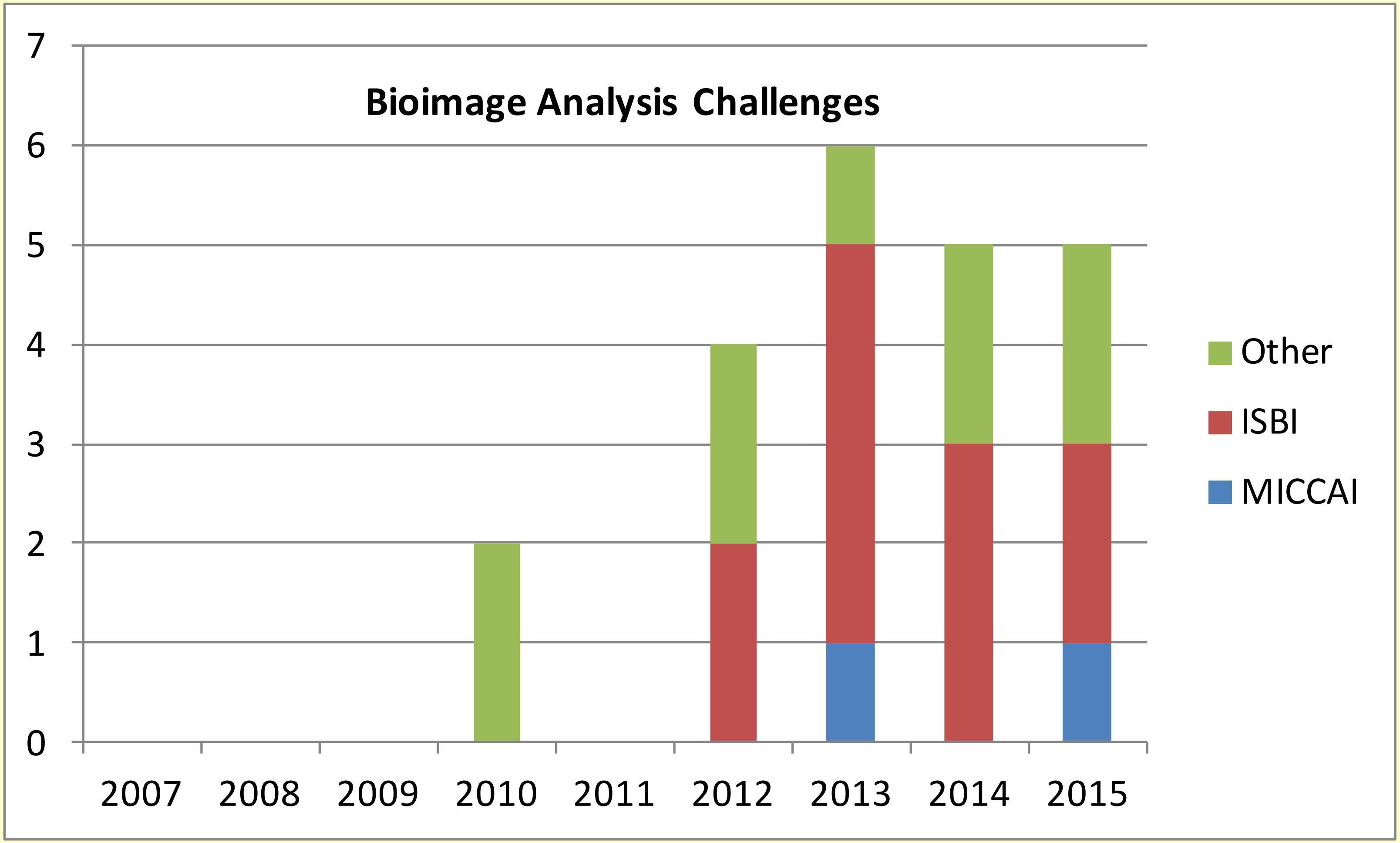


ISBI: International Symposium on Biomedical Imaging
MICCAI: Medical Image Computing and Computer Assisted Intervention

[Kozubek M: Challenges and Benchmarks in Bioimage Analysis, *Adv Anat Embryol Cell Biol*, 219, 2016]

# Challenges Organized So Far

## Growing awareness of the need for benchmarking



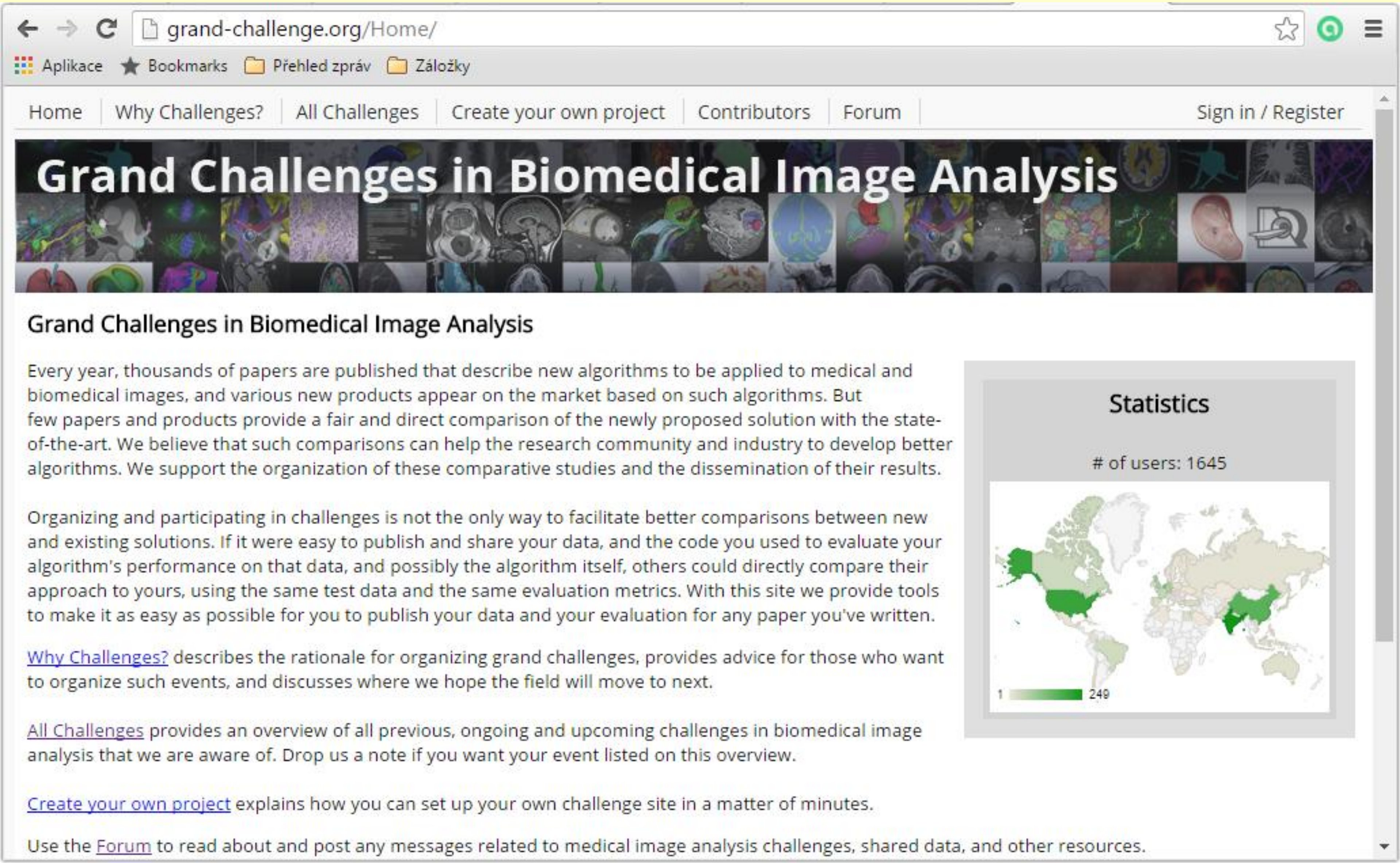
ISBI: International Symposium on Biomedical Imaging

MICCAI: Medical Image Computing and Computer Assisted Intervention

[Kozubek M: Challenges and Benchmarks in Bioimage Analysis, *Adv Anat Embryol Cell Biol*, 219, 2016]

# Database of challenges in biomedical image analysis

http://grand-challenge.org/ (Bram van Ginneken *et al.*)



## Grand Challenges in Biomedical Image Analysis

Every year, thousands of papers are published that describe new algorithms to be applied to medical and biomedical images, and various new products appear on the market based on such algorithms. But few papers and products provide a fair and direct comparison of the newly proposed solution with the state-of-the-art. We believe that such comparisons can help the research community and industry to develop better algorithms. We support the organization of these comparative studies and the dissemination of their results.

Organizing and participating in challenges is not the only way to facilitate better comparisons between new and existing solutions. If it were easy to publish and share your data, and the code you used to evaluate your algorithm's performance on that data, and possibly the algorithm itself, others could directly compare their approach to yours, using the same test data and the same evaluation metrics. With this site we provide tools to make it as easy as possible for you to publish your data and your evaluation for any paper you've written.

Why Challenges? describes the rationale for organizing grand challenges, provides advice for those who want to organize such events, and discusses where we hope the field will move to next.

All Challenges provides an overview of all previous, ongoing and upcoming challenges in biomedical image analysis that we are aware of. Drop us a note if you want your event listed on this overview.

Create your own project explains how you can set up your own challenge site in a matter of minutes.

Use the Forum to read about and post any messages related to medical image analysis challenges, shared data, and other resources.

### Statistics

# of users: 1645

# Outline

- Motivation
  - Why benchmarking?
- **Design of a benchmark or a challenge**
  - **Dataset selection**
  - **Algorithm evaluation / Metrics**
- Summary and discussion
  - Future directions
- Existing benchmarks and challenges
  - Overview
  - Particle Tracking Challenge
  - Cell Tracking Challenge
  - Localization Microscopy Challenge (Daniel Sage)

# Design of a Benchmark or a Challenge

- Dataset selection
  - Representative dataset selection: Covering variability of imaged objects
  - Real versus synthetic data: Advantages and disadvantages
  - Annotation of real data: Combining ground truth from several experts
  - Training versus test data: Splitting principles
- Algorithm evaluation
  - Evaluation metrics: Measuring performance of classification, segmentation, tracking, restoration and other methods
  - Merging multiple metrics: Normalization and weighting
  - Creating rankings: Coping with variable method performance across datasets
- Benchmark or challenge maintenance
  - Benchmark or challenge lifecycle: Updates, repetitions and open submission modes

# Representative Dataset Selection

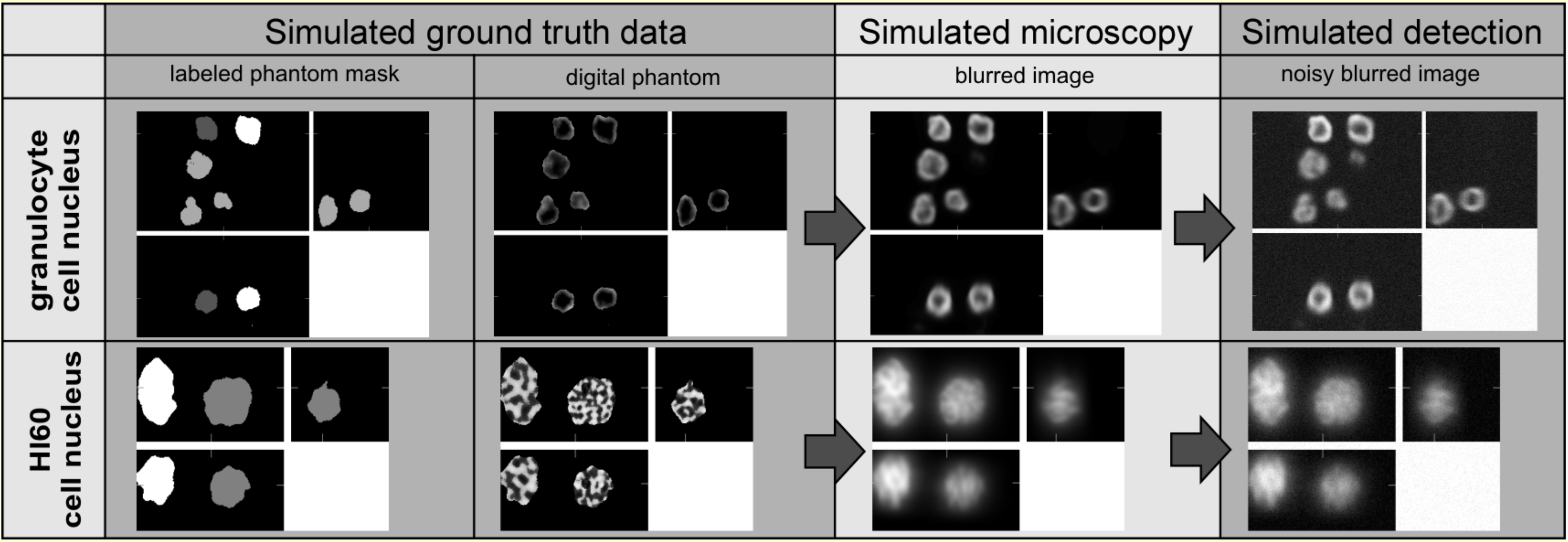
- Covering variability of imaged objects
  - Size, shape, texture, density, speed, etc.
- Covering various events or processes
  - Mitotic or apoptotic events, especially in time-lapse imaging
- Covering artefacts even if they occur rarely
  - Fluorescence bleaching, dust, uneven illumination, various types of noise, etc.
- Balanced occurrence of various types of objects, events or artefacts
  - In a way that corresponds to natural proportions
  - Otherwise, the developed algorithms will adjust to the most frequent types
  - If benchmark datasets for the given application already exist but are not comprehensive enough, it may be worth releasing additional datasets

# Real versus Synthetic Data

- Real datasets
  - Advantages:
    - The best available representation of imaged objects
  - Disadvantages:
    - Due to blur and noise, the correct answer to the biological question (ground truth) is often ambiguous (various experts give different answers)
    - Available only in limited quantities
- Synthetic datasets
  (also called simulated data or digital phantom images)
  - Advantages:
    - Can be easily generated at low cost in large quantities at different settings of noise level, cell density, cell speed, etc.
    - The ground truth is known precisely
  - Disadvantages:
    - Often questionable how similar simulated and real data are; they should be similar not only visually but also their characteristics / features should match

# Synthetic Data Generation

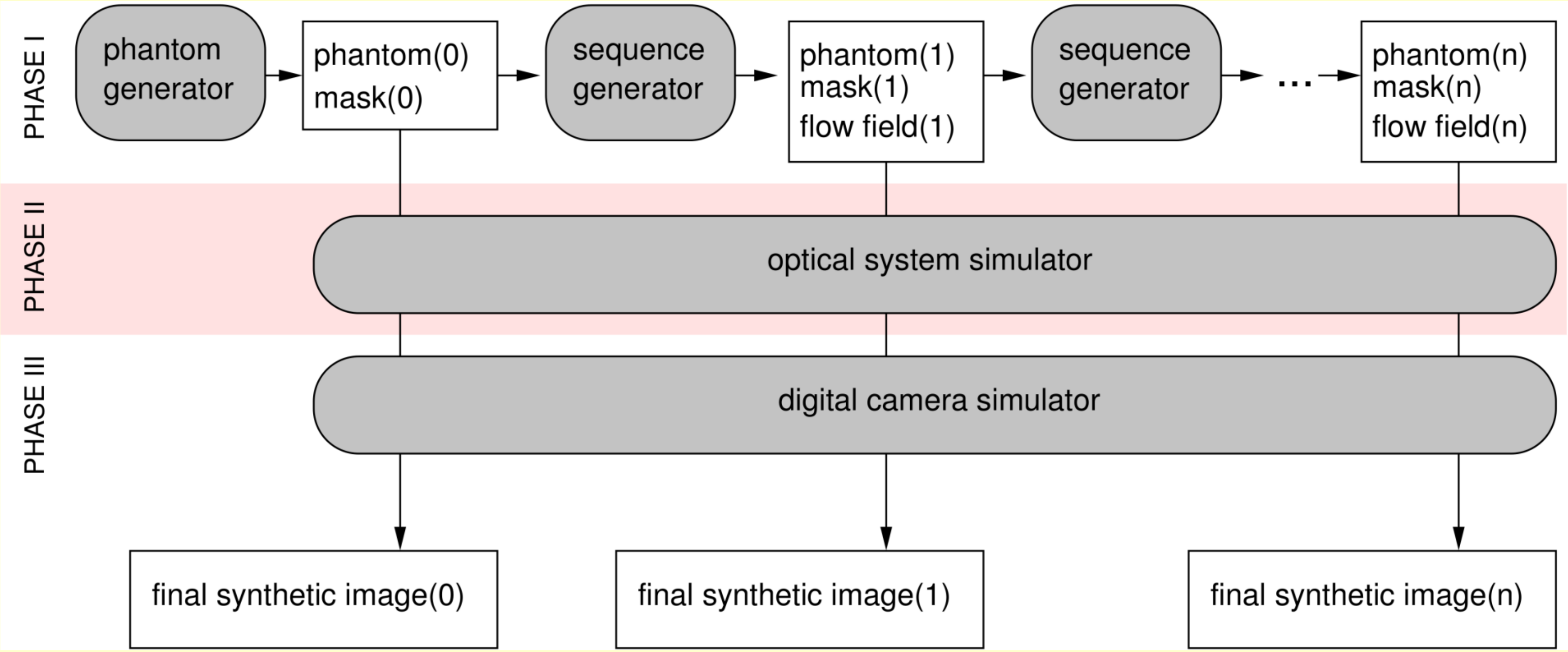
## Three phases of single frame simulation



[Ortiz-de-Solórzano et al. Toward Morphodynamic Model of the Cell. IEEE SPM 32(1), 20-29, 2015]

# Synthetic Data Generation

## Time-lapse image sequence simulation

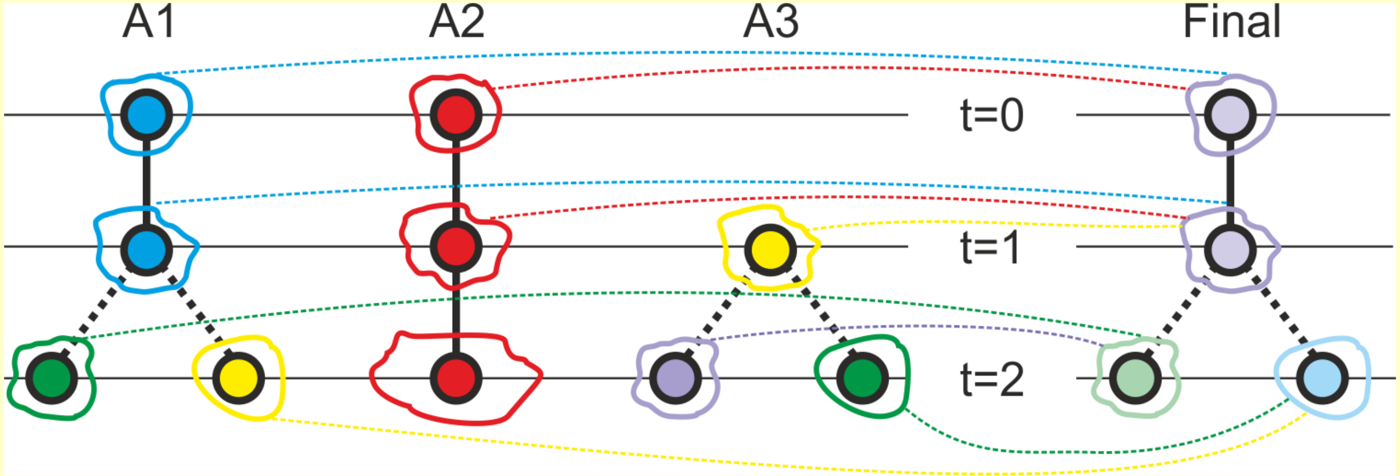


# Real versus Synthetic Data

- Physical phantom datasets
  - Advantages:
    - Acquired using real instrument with real noise and blur
    - Thanks to known properties, can be used for instrument calibration
  - Disadvantages:
    - Hard and expensive to manufacture
    - Available only for large objects
- Which type should be used for benchmarking?
  - Ideally all because their properties are complementary
  - In some applications, however, specific type might be not available, e.g.:
    - Real data in deconvolution: ground truth cannot be determined even by an expert
    - Synthetic data in electron microscopy: too complex to simulate currently
    - Physical phantom data in cell imaging: too small and complex to manufacture

# Annotation of Real Data

- Combining ground truths from several experts
  - Due to the subjectivity of expert answers, it is beneficial and more reliable to ask several experts and combine their answers
  - For binary or multiple choice decisions (especially classification), reasonable number of experts is three because **majority voting** scheme can then be applied:



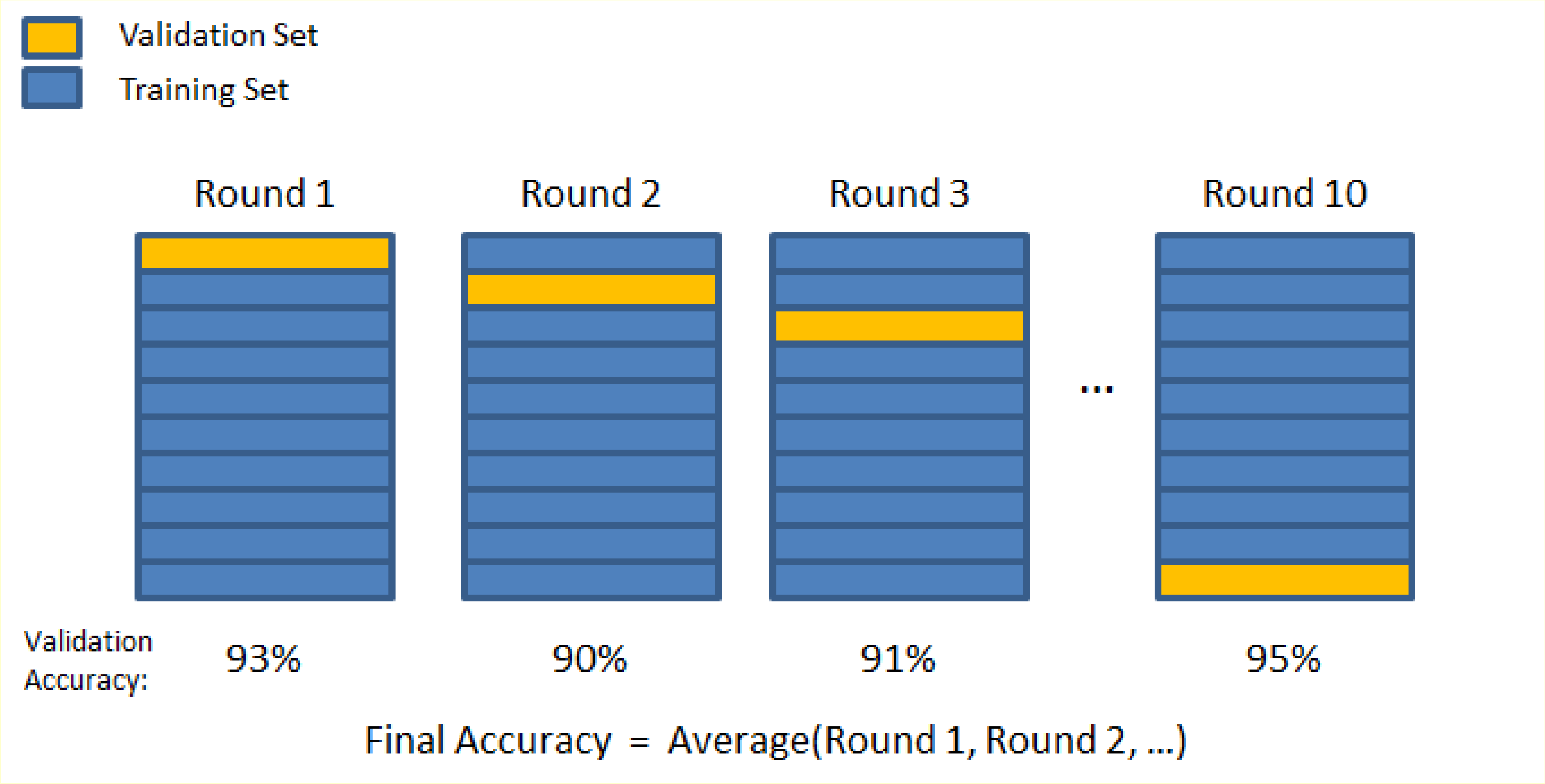
  - For answers in the form of a number (especially position determination or boundary delineation), two or three expert answers can be **averaged** while checking that their variability does not exceed a certain threshold:
    - *Annotators: 10, 123, 118, 259, 131* *Final: (123 + 118 + 131) / 3 = 124*

# Training versus Test Data

- Why to split datasets?
  - Training phase: the developers fine-tune their methods to work well on the particular data type using training datasets with **supplied ground truth**
  - Test (competition) phase: the developers apply developed software to previously unseen test data, whose **ground truth is kept secret**
  - In statistics, such an approach is called **holdout method** because test data are held out while training
- How to split datasets?
  - It is suggested to use about two-thirds of the data for training and one-third for testing (in practice, however, often 50:50 is used)
  - The split should be done **in a balanced way**; i.e., both training and test data should be representative and have similar properties

# Training versus Test Data

- Splitting possibilities
  - Single holdout method: Repeated once (most common)
  - Multiple holdout method: Repeated K times ⇒ **K-fold cross-validation**
  - (often used in machine learning, especially for classification tasks)

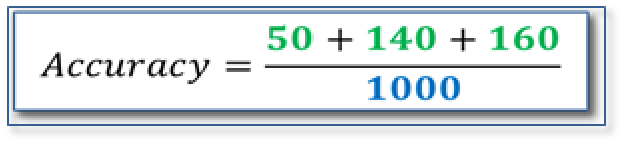


Courtesy of Chris McCormick

# Most Common Measures/Metrics

- Pixel or object classification : General case
  - Confusion matrix and accuracy

| | Predicted class A | Predicted class B | Predicted class C | |
|---|---|---|---|---|
| Actual class A | **50** | **80** | **70** | 200 |
| Actual class B | **40** | **140** | **120** | 300 |
| Actual class C | **120** | **220** | **160** | 500 |
| | 210 | 440 | 350 | **1000** |

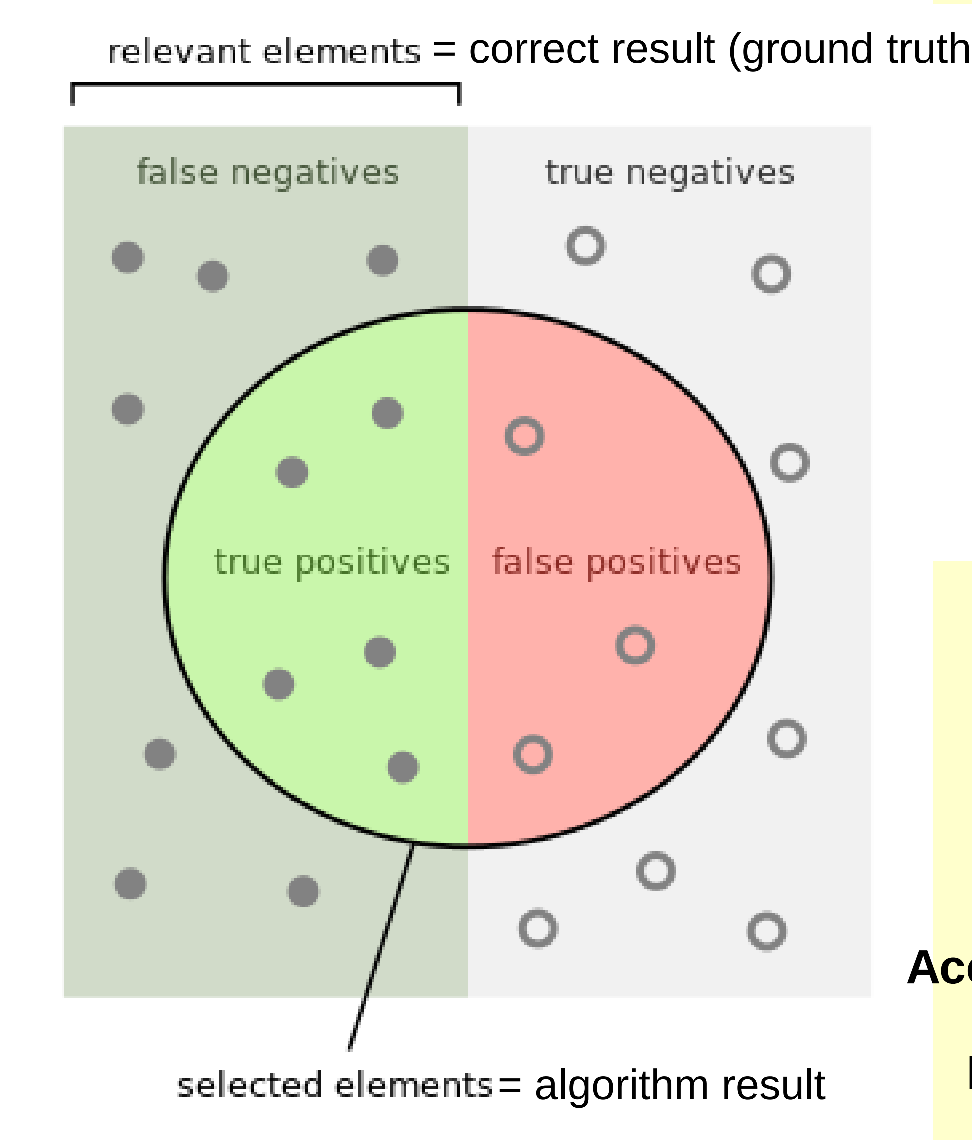


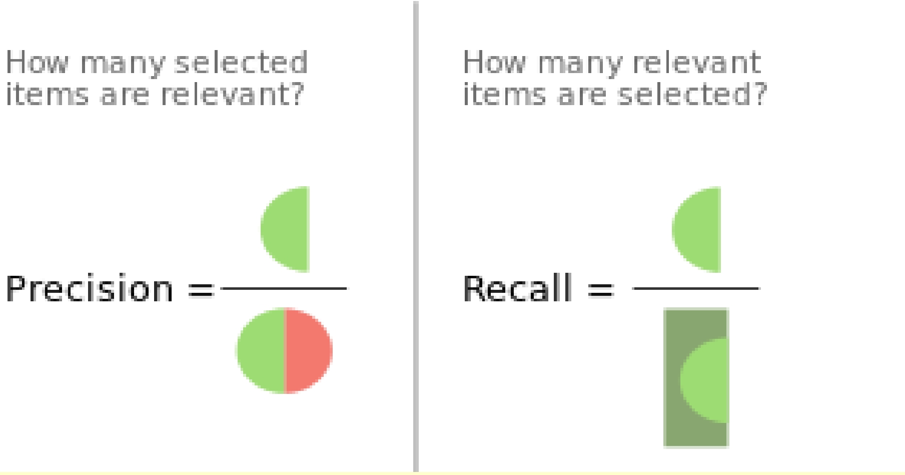
$$Accuracy = \frac{50 + 140 + 160}{1000}$$

Courtesy of Nicolas Nicolov

# Most Common Measures/Metrics

- Pixel or object classification: Binary Case


relevant elements = correct result (ground truth)
false negatives
true negatives
true positives
false positives
selected elements = algorithm result


How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =

**Precision = TP / (TP + FP)**

**Sensitivity = Recall = TP / (TP + FN)**

**Specificity = TN / (TN + FP)**

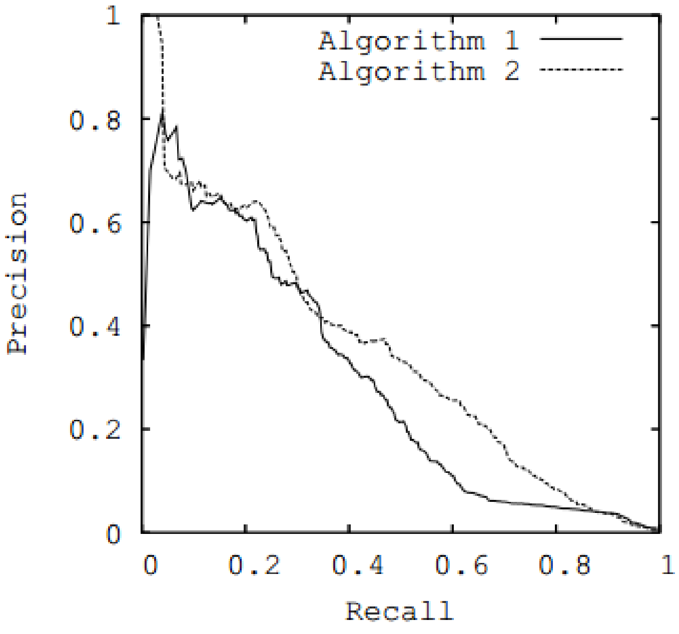
**Accuracy = (TP + TN) / (TP + TN + FP + FN)**

**F-score = 2 (Prec · Rec) / (Prec + Rec)**

# Most Common Measures/Metrics

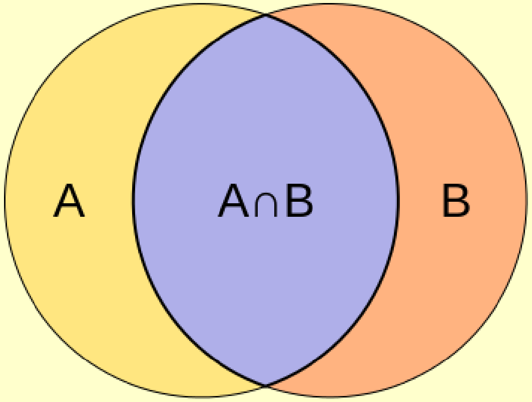
## Pixel or object classification: Typical values of precision, recall, accuracy and F-score

- All these characteristics are normalized
  - Their value ranges from 0 (worst) to 1 (best)
  - Values above 0.9 are considered excellent for most applications
- Precision and recall
  - Mutually dependent, a plot can be computed to show their mutual dependence
  - Usually a balance is kept, i.e. they are usually of similar value
  - Sometimes precision is favored
- Accuracy and F-score
  - Each combines algorithm performance into a single number
  - For the best methods typically reach the value of 0.6-0.9 depending on how hard the application is, rarely above 0.9 for easy tasks



# Most Common Measures/Metrics

- Segmentation results (binary masks comparisons) in 2D and 3D
  - One binary mask corresponds to ground truth, the other to algorithm result



For multiple objects per image
(or per whole dataset):

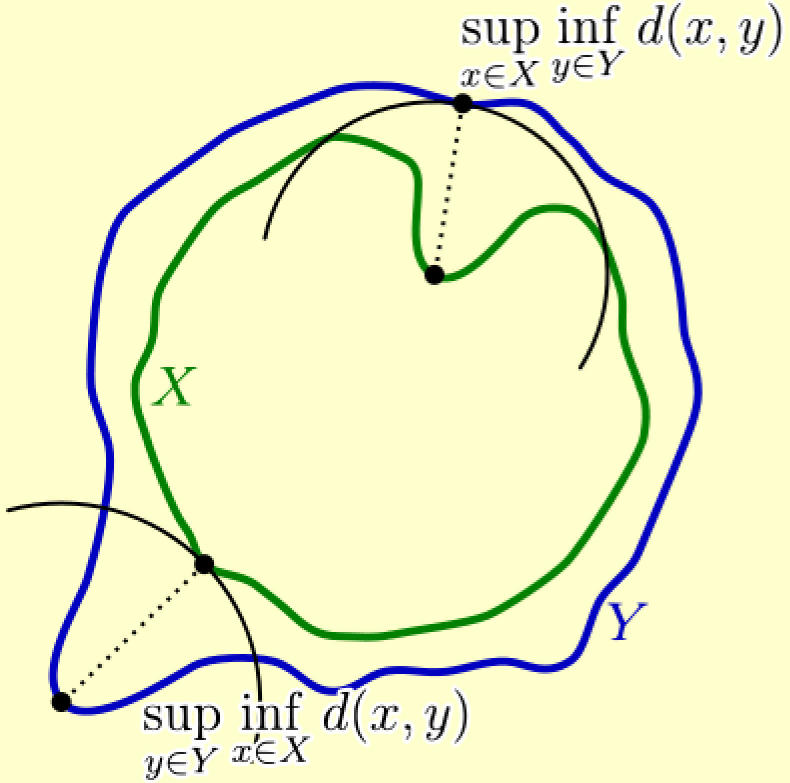
1) For each reference GT object A find segmented object B so that |A∩B| > 0.5 |A|, for no match: set B=∅, hence |A∩B|=∅
2) Compute Jaccard or Dice measure for each such pair
3) Compute average of all such measures from previous step

- **Jaccard similarity index** $= \dfrac{|A \cap B|}{|A \cup B|}$
- Dice coefficient $= \dfrac{2|A \cap B|}{|A| + |B|}$
- Both are normalized to reach values from 0 (worst) to 1 (best)
- Typically <0.5 for hard tasks, 0.5-0.8 for intermediate, >0.8 for easy ones

# Most Common Measures/Metrics

- Shape similarity in 2D and 3D
  - **Hausdorff distance**: the longest out of the shortest distances

$$d_{\mathrm{H}}(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y),\, \sup_{y\in Y}\inf_{x\in X} d(x,y)\}$$


$\sup_{x\in X}\inf_{y\in Y} d(x,y)$
$X$
$Y$
$\sup_{y\in Y}\inf_{x\in X} d(x,y)$

Source: Wikipedia

# Most Common Measures/Metrics

- Measurement results (position, length, size/area/volume)
  - Ground truth = correct value of the measurement
  - Error = difference between measured value and GT value
  - Errors of multiple measurements are typically averaged as follows:
    - **Root-mean-square error (RMSE)** = Root-mean-square distance (RMSD)

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

    - Not only scalar but also vector values (e.g., 3D positions) can be compared

$$\begin{aligned}\mathrm{RMSD}(\mathbf{v},\mathbf{w}) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}\end{aligned}$$

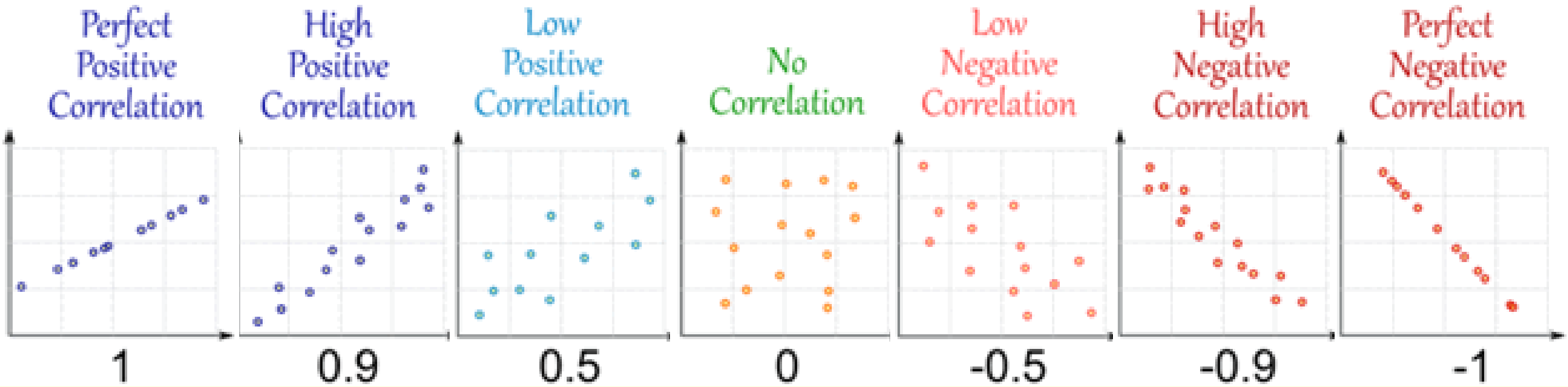
  - Plotting error distribution histogram may also be helpful

# Most Common Measures/Metrics

- Comparing patterns, textures, distributions
  - In order to compare two 1D signals (or 2D/3D signals transformed into 1D signals), most frequently correlation is used, especially **normalized cross correlation**:

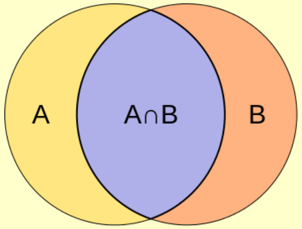
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - To save computation time, normalization (denominator) may be omitted to yield just **cross correlation** (numerator)
  - Plotting the two 1D signals against each other tells the story:

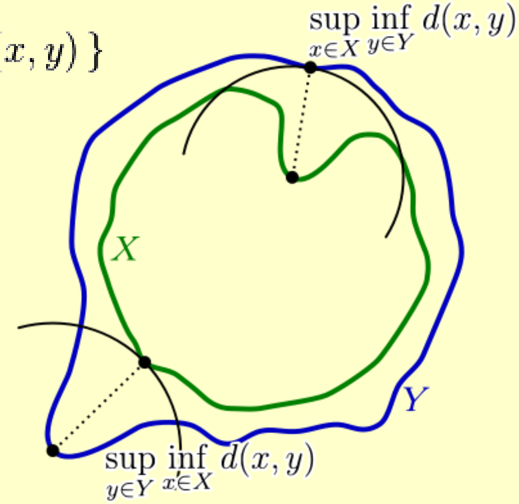


# Most Common Measures/Metrics

- Pixel or object classification
  - Confusion matrix and accuracy
  - In binary case at least (even if true negatives are not defined): true positives, false positives, false negatives, precision, recall and F-score
- Segmentation results (binary masks comparisons)
  - Jaccard similarity index $= \frac{|A \cap B|}{|A \cup B|}$ or Dice coefficient $= \frac{2|A \cap B|}{|A| + |B|}$



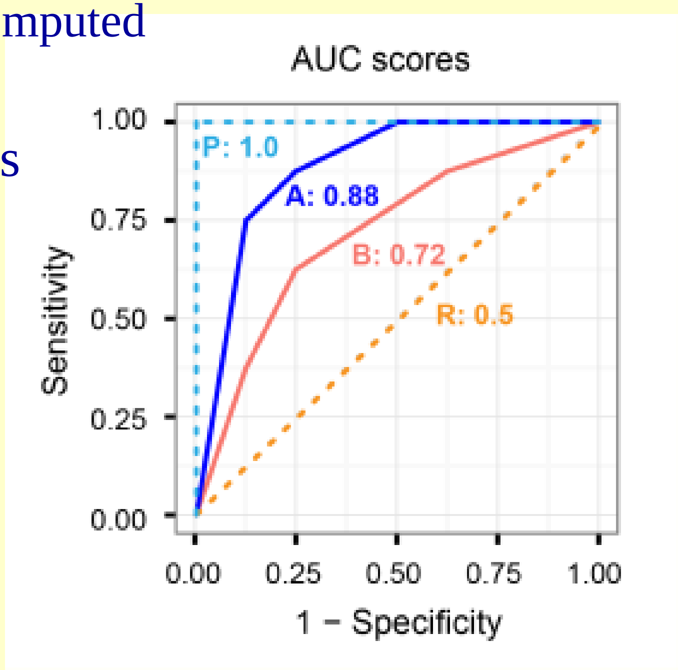
- Shape similarity
  - Hausdorff distance $d_{\mathrm{H}}(X,Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y)\}$


$\sup_{x \in X} \inf_{y \in Y} d(x,y)$
$X$
$Y$
$\sup_{y \in Y} \inf_{x \in X} d(x,y)$

- Position (localization) error
  - Root-mean-square distance (RMSD)
- Comparing distributions
  - Correlation
- Speed and memory consumption

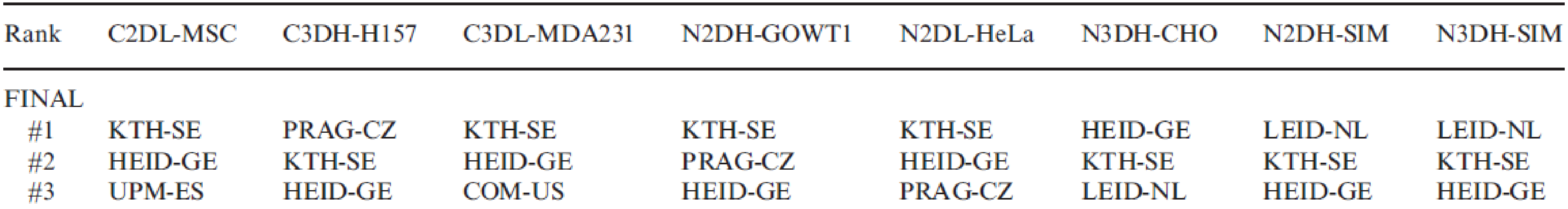
# Merging Multiple Metrics

- Treating multiple metrics on the same dataset
  - Their values should be published separately anyway (to assess strong and weak sides of each algorithm)
  - If single final score for each method is required, then:
    - Weights should be assigned to each metric (how important it is), some metrics can be put aside as supplementary (e.g., time, memory)
    - Each metric should be normalized before merging (e.g., to [0-1] interval where 0 means worst and 1 means best)
    - Weighted sum of selected metrics is typically computed
  - Special attention should be paid to combining mutually (typically inversely) dependent metrics
    - Curve / scatter-plot analysis can help (e.g., area under the curve – AUC – can be measured as in ROC curve analysis)
    - Minimal values for each can be defined
    - Alternatively can be replaced with single measure (e.g., F-score)



# Creating Rankings

- Merging performance measures on multiple types of datasets
  - **There is no algorithm that performs best on all types of data**
  - The *individual rankings* for each type should be published separately (to assess for which type of data the method works well or not so well)
  - Some methods have no individual rank for certain types of data (just can not cope with certain data types) – so called *missing scores*
  - If single *global ranking* from individual rankings is required, then:
    - Averaging could be used but missing scores cause troubles – how to penalize them?
    - Alternatively, one can count the occurrences of a particular algorithm at winning position or among the top three best performing methods across various data types – this naturally copes with missing scores



| Rank | C2DL-MSC | C3DH-H157 | C3DL-MDA231 | N2DH-GOWT1 | N2DL-HeLa | N3DH-CHO | N2DH-SIM | N3DH-SIM |
|---|---|---|---|---|---|---|---|---|
| FINAL | | | | | | | | |
| #1 | KTH-SE | PRAG-CZ | KTH-SE | KTH-SE | KTH-SE | HEID-GE | LEID-NL | LEID-NL |
| #2 | HEID-GE | KTH-SE | HEID-GE | PRAG-CZ | HEID-GE | KTH-SE | KTH-SE | KTH-SE |
| #3 | UPM-ES | HEID-GE | COM-US | HEID-GE | PRAG-CZ | LEID-NL | HEID-GE | HEID-GE |

[Maška et al. A benchmark for comparison of cell tracking algorithms, *Bioinformatics* 30, 2014]

# Benchmark or Challenge Maintenance

- Benchmark lifecycle

  Creation → Release (web) → Use → Publication (paper) → Use / Updates

- Challenge lifecycle

  Creation → Release (web) → Contest → Publication (paper) → Use / Updates

- Approaches to maintenance (from worst/cheap to best/expensive)
  - *One time event*, kept on-line but no further maintenance
  - *Repeated event*, kept on-line and updated once a year or every other year
  - *Open submission mode*, kept on-line, updated on-demand with reasonable response time

- Open access to
  - Datasets + ground truth (except for test data ground truth)
  - Annotation/evaluation tools, performance results + rankings
  - Image analysis methods ideally as open source (not all groups agree)
  - **Problem: open access to most items often delayed (1-2 years) until paper is out! This hampers the progress in the field and building on previous work!**

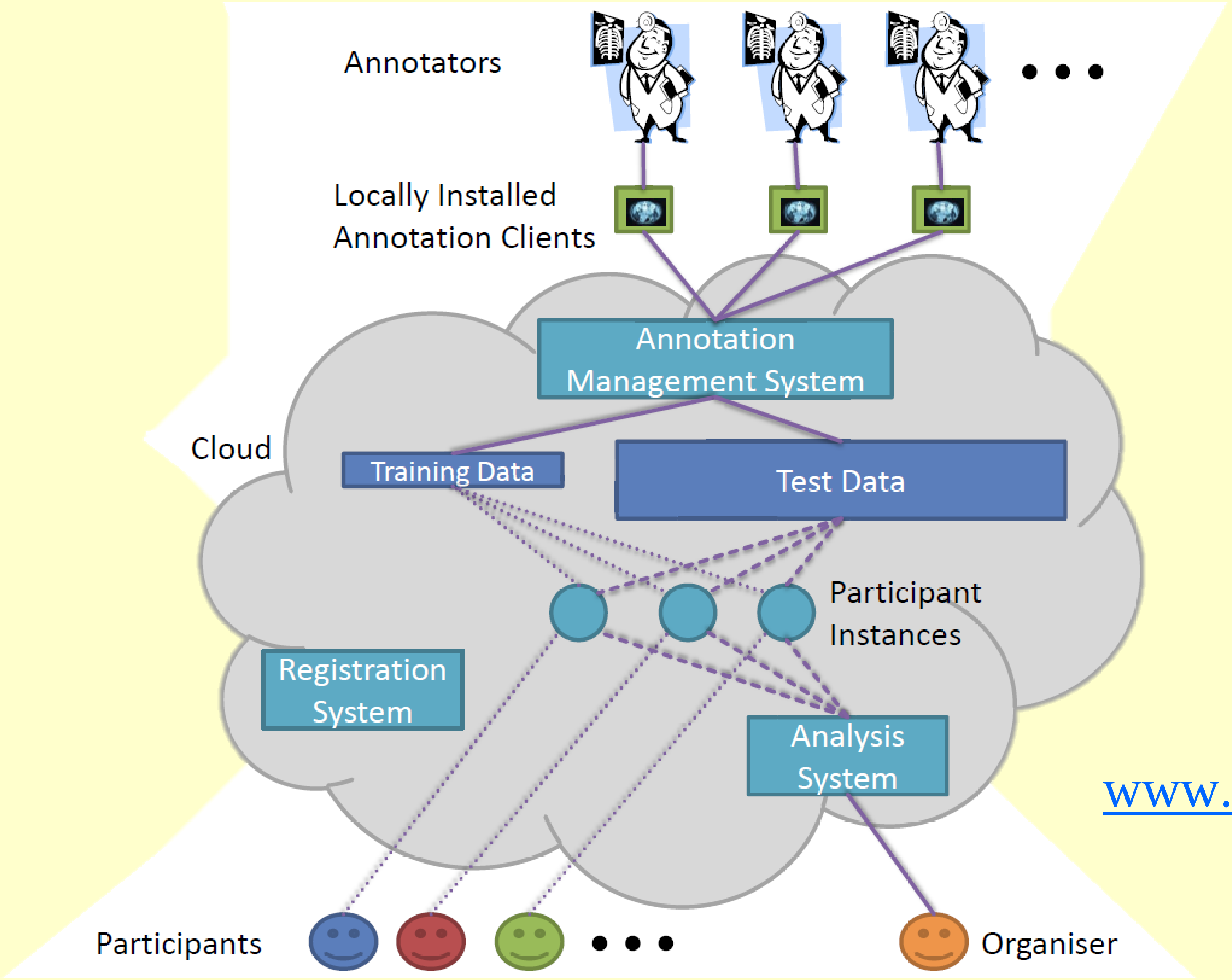
# Outline

- Motivation
  - Why benchmarking?
- Design of a benchmark or a challenge
  - Dataset selection
  - Algorithm evaluation / Metrics
- **Summary and discussion**
  - **Future directions**
- Existing benchmarks and challenges
  - Overview
  - Particle Tracking Challenge
  - Cell Tracking Challenge
  - Localization Microscopy Challenge (Daniel Sage)

# Summary

- Creating a benchmark dataset
  - Selection of **representative** image data (containing variability, events, etc.)
  - Choice of **real versus synthetic** datasets (or both), rarely physical phantoms
  - **Annotating** real datasets (merging several annotations) to get GT
- Creating a challenge based on a benchmark dataset
  - Splitting datasets into **training data** (public GT) / **competition data** (secret GT)
  - Defining **measure(s)** to compare an analysis result with GT (+ speed/memory)
  - Defining **weights** in the case of multiple measures
  - Creating **tools/software** that can compute the measure(s)
- Organizing a challenge
  - **Releasing** datasets, training data GTs, tools and deadlines on a web page
  - In the frame of a **known event** more participants can be attracted
  - Evaluating / verifying the submitted results, creation of **ranking(s)**
  - **Publishing** results (web page, journal)

# What About Future?

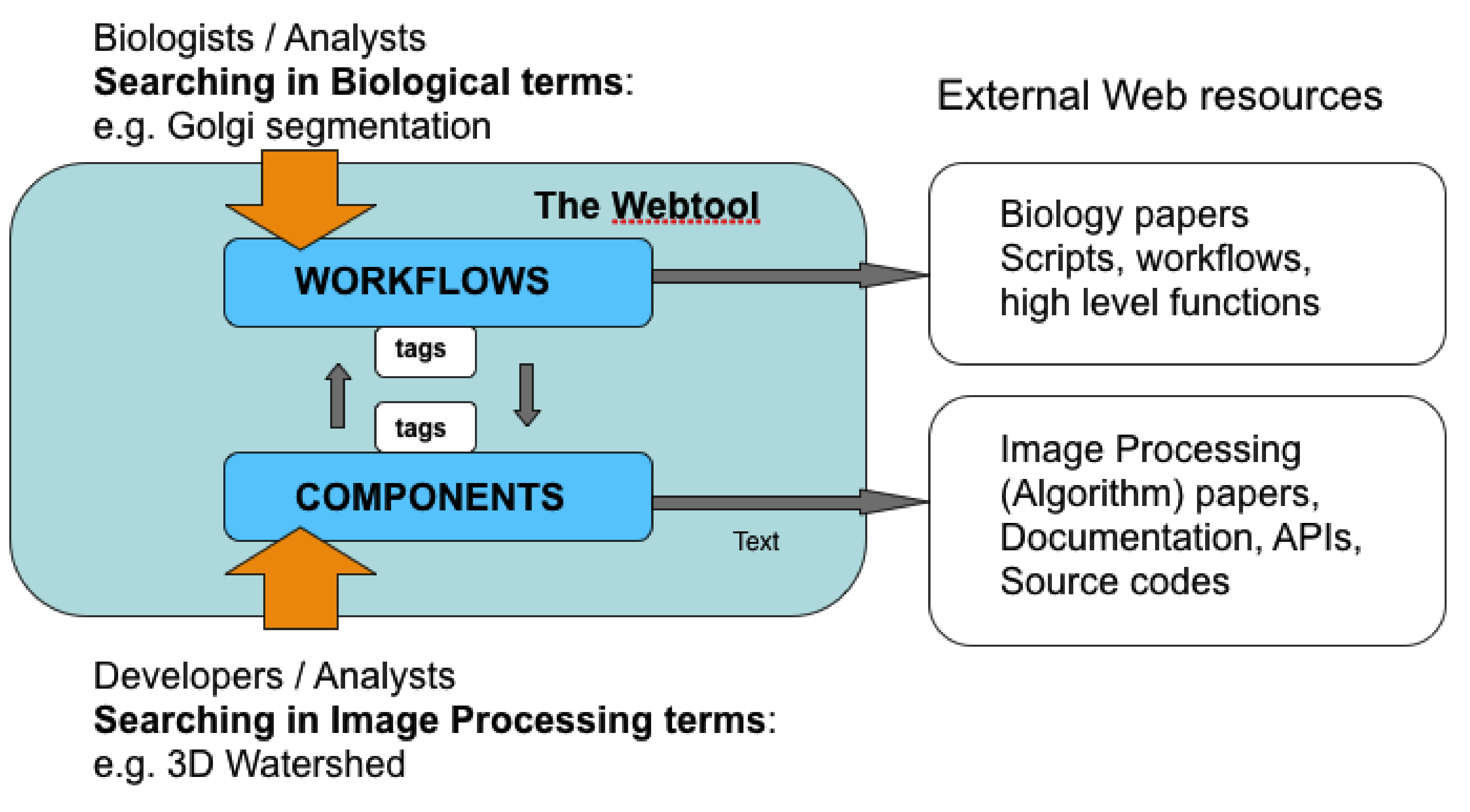


www.visceral.eu

[Hanbury et al. Cloud-Based Evaluation Framework for Big Data. LNCS 7858, 104–114, 2013]

# COST NEUBIAS
# (2016-2020)

neubias
network of european
bioimage analysts



Biologists / Analysts
**Searching in Biological terms**:
e.g. Golgi segmentation

External Web resources

The Webtool

WORKFLOWS

tags

tags

COMPONENTS

Text

Biology papers
Scripts, workflows,
high level functions

Image Processing
(Algorithm) papers,
Documentation, APIs,
Source codes

Developers / Analysts
**Searching in Image Processing terms**:
e.g. 3D Watershed

# Outline

- Motivation
  - Why benchmarking?
- Design of a benchmark or a challenge
  - Dataset selection
  - Algorithm evaluation / Metrics
- Summary and discussion
  - Future directions
- **Existing benchmarks and challenges**
  - **Overview**
  - **Particle Tracking Challenge**
  - **Cell Tracking Challenge**
  - **Localization Microscopy Challenge (Daniel Sage)**

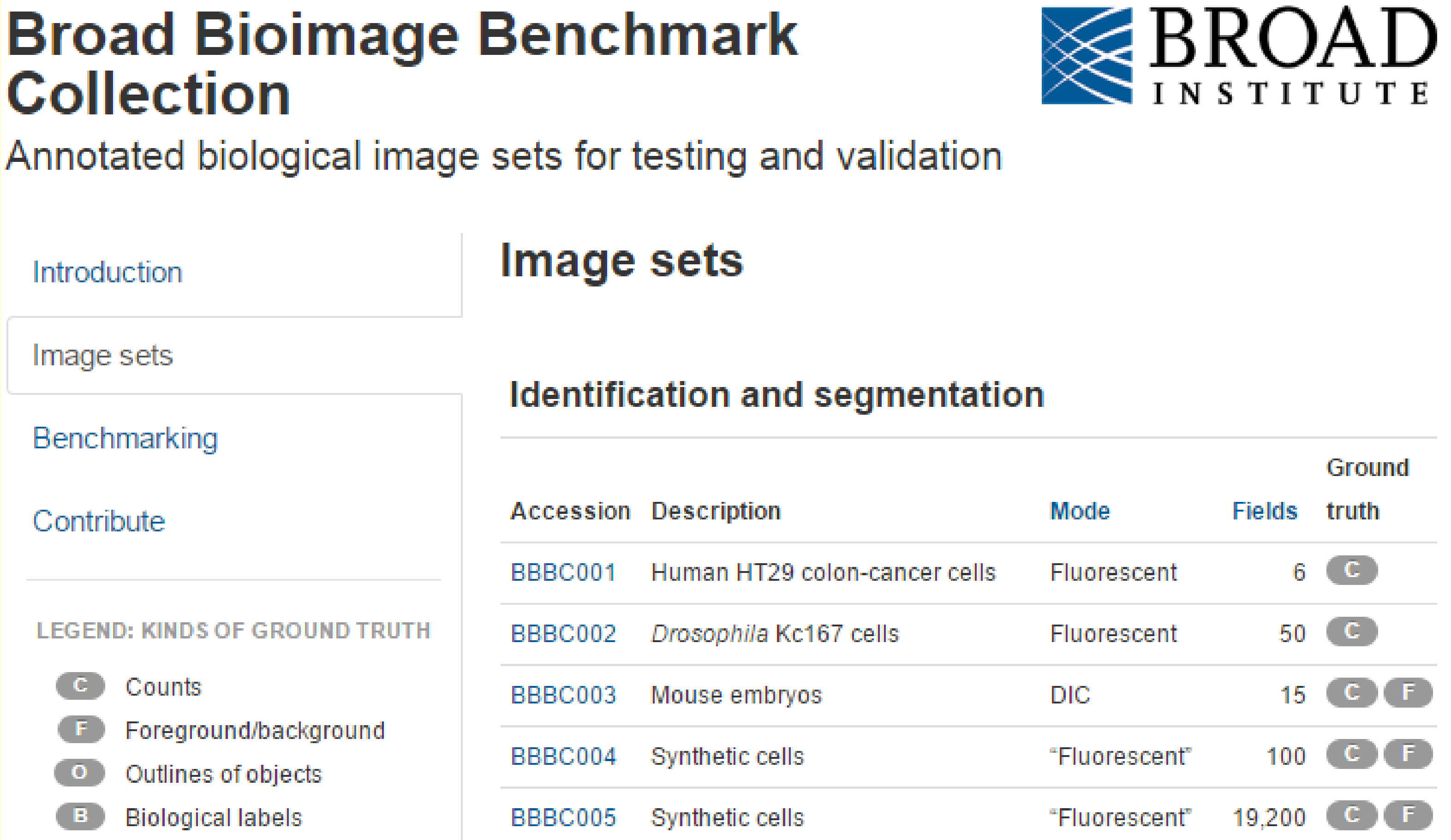
# Benchmarks in Bioimage Analysis

## (not associated with challenges)

- Broad Bioimage Benchmark Collection (BBBC)
  - Broad Institute, since 2008, real (fluorescence, brightfield, DIC) + simulated
- Cell Centered Database (CCDB)
  - UCSD, 2002, 2D - 4D real datasets from light and electron microscopy, lack of GT
- CellOrganizer and Murphy Lab Data
  - CMU, since 1999, mainly data from papers, CellOrganizer since 2012 – synthetic data
- Masaryk University Cell Image Collection (MUCIC)
  - Masaryk Univ, since 2008, synthetic cell microscopy data up to 3D+time
- Deconvolution in Microscopy
  - EPFL, 2010, synthetic hollow bars, C. elegans embryo, fluorescent beads
- JCB DataViewer
  - Journal of Cell Biology, since 2008, sharing of image data from papers, lack of GT
- SimuCell
  - UCSF, since 2012, synthetic data but only 2D
- The Cell: An Image Library (CIL)
  - ASCB, since 2010, database of cell-related image data, lack of GT, joined with CCDB in 2012
- UCSB Biosegmentation Benchmark
  - UCSB, since 2008, presented at ICIP 2008, from subcellular to tissue level, 2D and 3D

# Benchmarks in Bioimage Analysis
## (not associated with challenges)

- Broad Bioimage Benchmark Collection (BBBC)
  - https://data.broadinstitute.org/bbbc/
  - **The oldest + most famous collection** of bioimaging data with ground truth
  - Broad Institute, since 2008, real (fluorescence, brightfield, DIC) + simulated



Broad Bioimage Benchmark Collection

Annotated biological image sets for testing and validation

Introduction

Image sets

Benchmarking

Contribute

LEGEND: KINDS OF GROUND TRUTH

| | |
|---|---|
| C | Counts |
| F | Foreground/background |
| O | Outlines of objects |
| B | Biological labels |

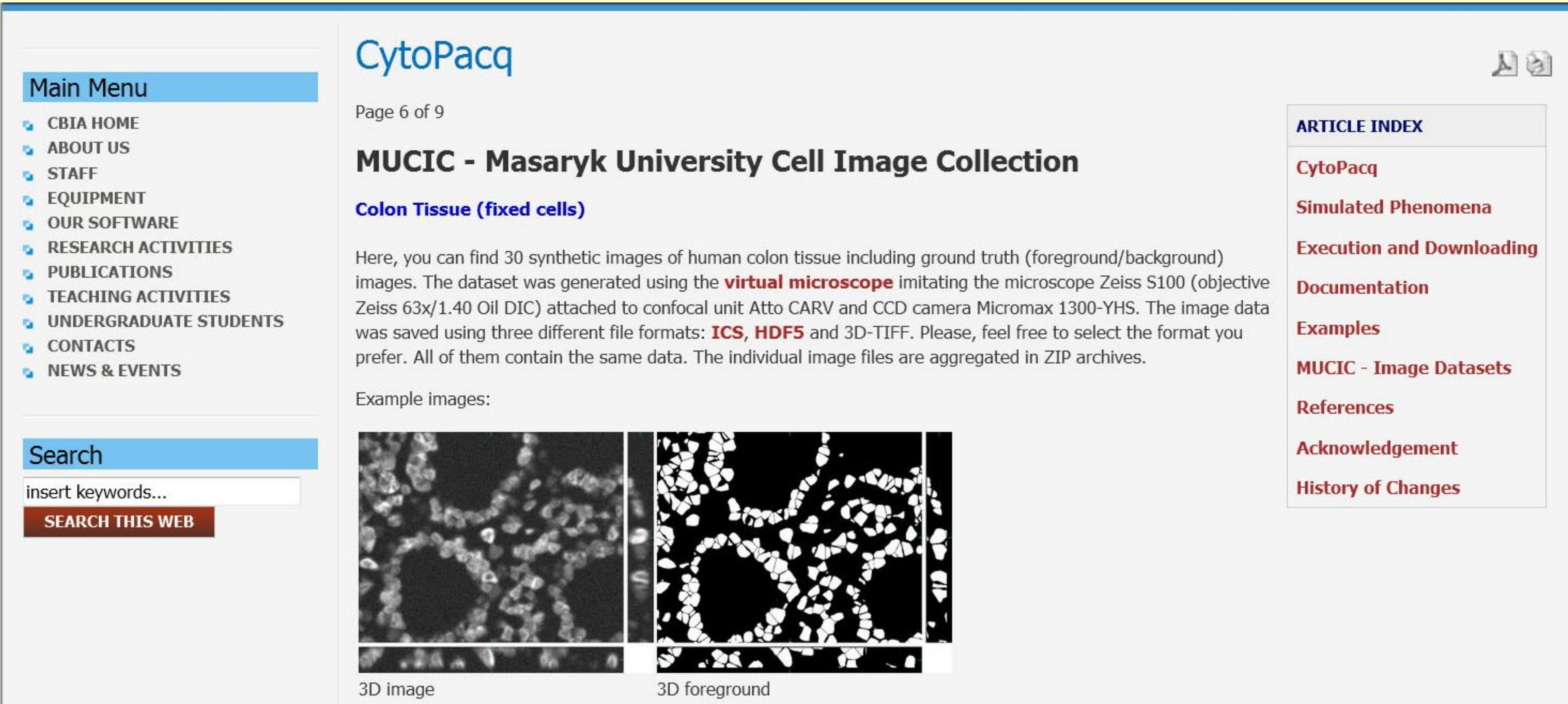
Image sets

Identification and segmentation

| Accession | Description | Mode | Fields | Ground truth |
|---|---|---|---|---|
| BBBC001 | Human HT29 colon-cancer cells | Fluorescent | 6 | C |
| BBBC002 | *Drosophila* Kc167 cells | Fluorescent | 50 | C |
| BBBC003 | Mouse embryos | DIC | 15 | C F |
| BBBC004 | Synthetic cells | "Fluorescent" | 100 | C F |
| BBBC005 | Synthetic cells | "Fluorescent" | 19,200 | C F |

# Benchmarks in Bioimage Analysis
(not associated with challenges)

- Masaryk University Cell Image Collection (MUCIC)
  - http://cbia.fi.muni.cz/datasets/
  - **The largest collection of synthetic** cell microscopy data up to 3D+time
  - Based on Brno CytoPacq simulation package, Masaryk University, since 2008



CytoPacq

Page 6 of 9

**MUCIC - Masaryk University Cell Image Collection**

**Colon Tissue (fixed cells)**

Here, you can find 30 synthetic images of human colon tissue including ground truth (foreground/background) images. The dataset was generated using the **virtual microscope** imitating the microscope Zeiss S100 (objective Zeiss 63x/1.40 Oil DIC) attached to confocal unit Atto CARV and CCD camera Micromax 1300-YHS. The image data was saved using three different file formats: **ICS**, **HDF5** and 3D-TIFF. Please, feel free to select the format you prefer. All of them contain the same data. The individual image files are aggregated in ZIP archives.

Example images:

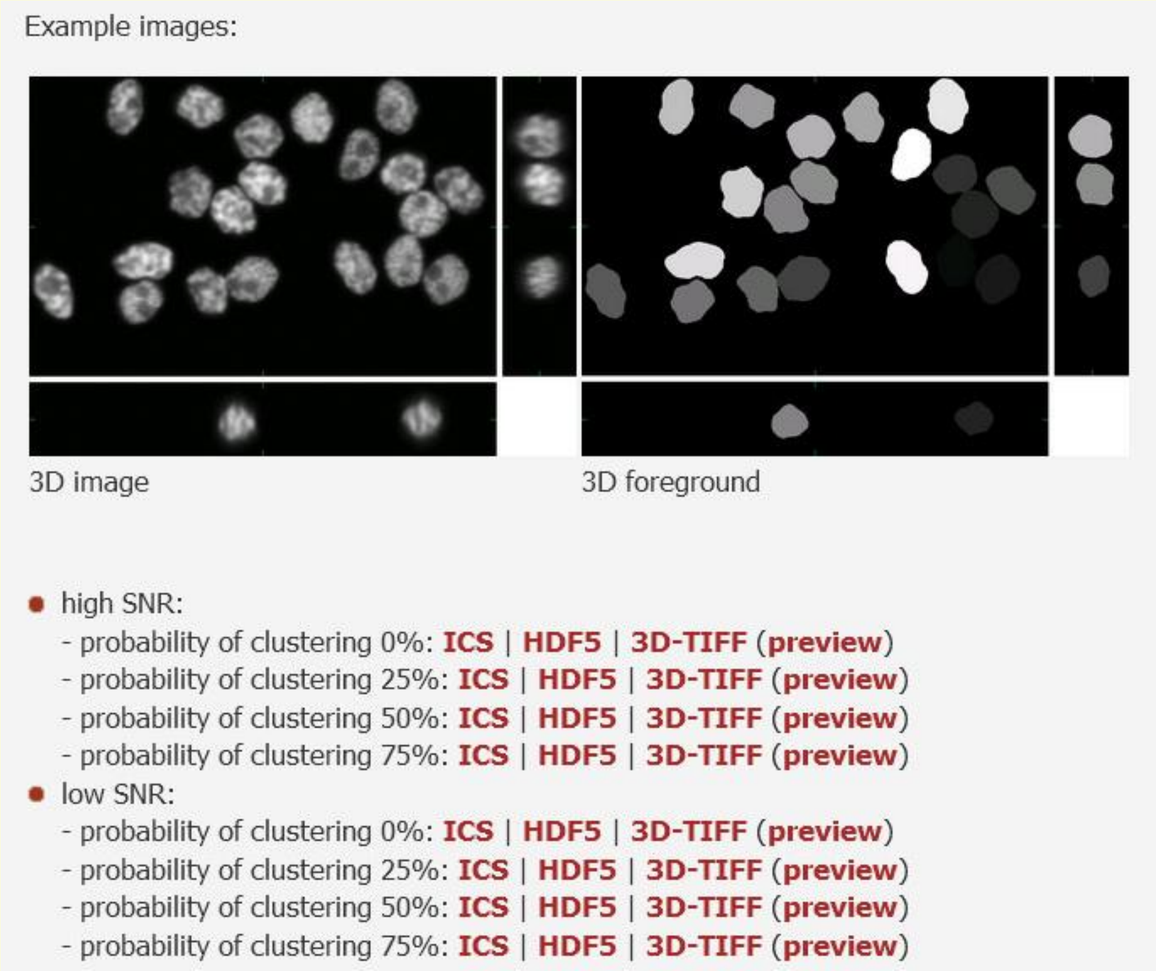
3D image

3D foreground

# Benchmarks in Bioimage Analysis

## (not associated with challenges)

- Masaryk University Cell Image Collection (MUCIC)
  - http://cbia.fi.muni.cz/datasets/
  - Synthetic clustered 3D cell nuclei (4 clustering densities and 2 SNR levels):



Example images:

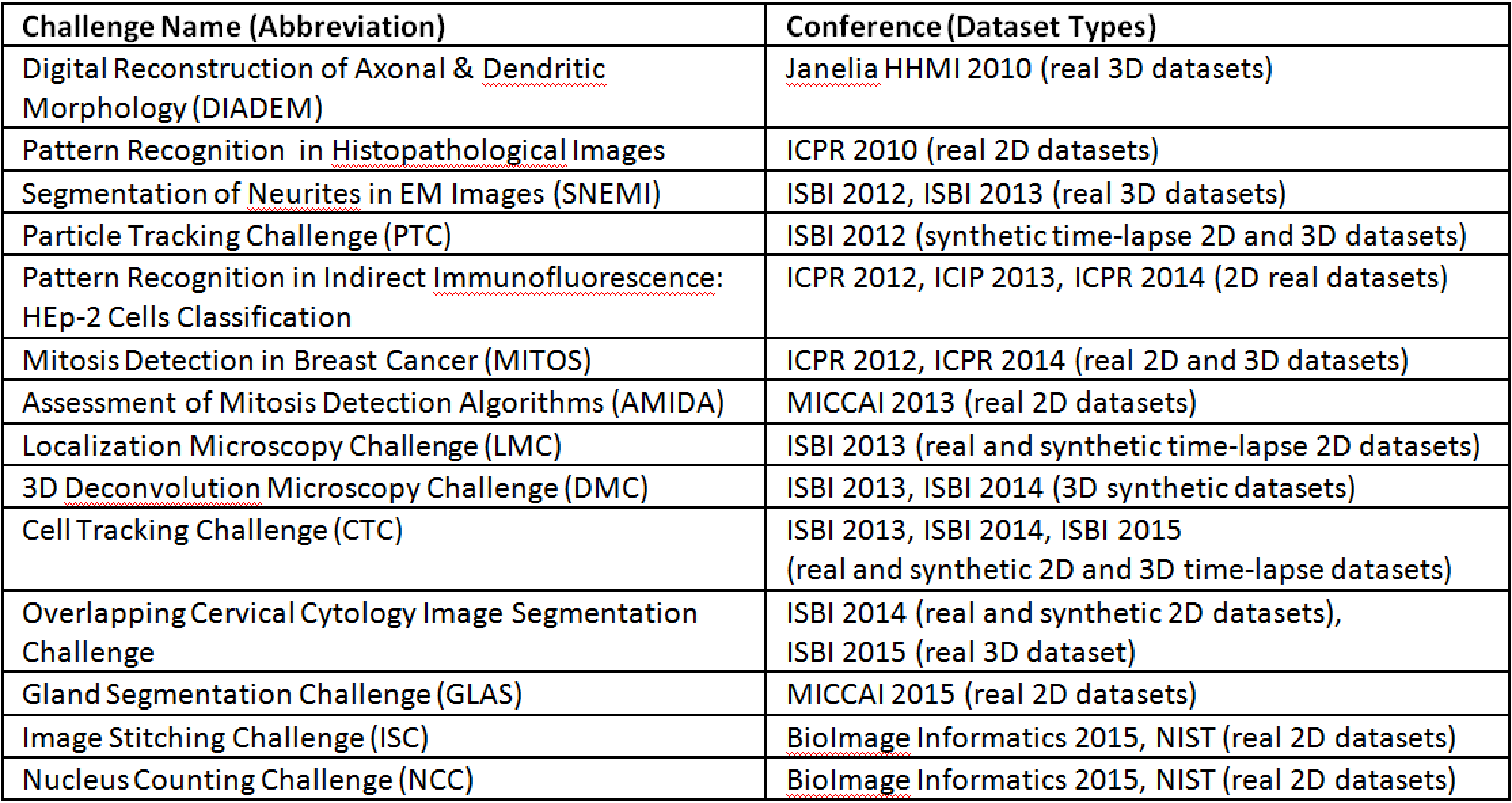
3D image

3D foreground

- high SNR:
  - probability of clustering 0%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 25%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 50%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 75%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
- low SNR:
  - probability of clustering 0%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 25%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 50%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)
  - probability of clustering 75%: **ICS** | **HDF5** | **3D-TIFF** (**preview**)

# Challenges in BioImage Analysis

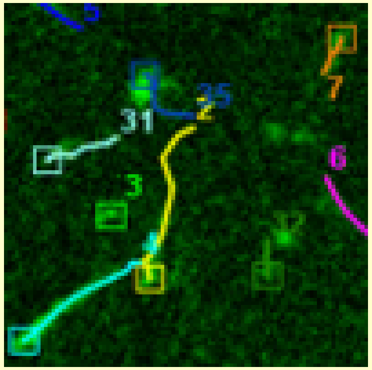
## Challenges organized until 2015



| Challenge Name (Abbreviation) | Conference (Dataset Types) |
|---|---|
| Digital Reconstruction of Axonal & Dendritic Morphology (DIADEM) | Janelia HHMI 2010 (real 3D datasets) |
| Pattern Recognition  in Histopathological Images | ICPR 2010 (real 2D datasets) |
| Segmentation of Neurites in EM Images (SNEMI) | ISBI 2012, ISBI 2013 (real 3D datasets) |
| Particle Tracking Challenge (PTC) | ISBI 2012 (synthetic time-lapse 2D and 3D datasets) |
| Pattern Recognition in Indirect Immunofluorescence: HEp-2 Cells Classification | ICPR 2012, ICIP 2013, ICPR 2014 (2D real datasets) |
| Mitosis Detection in Breast Cancer (MITOS) | ICPR 2012, ICPR 2014 (real 2D and 3D datasets) |
| Assessment of Mitosis Detection Algorithms (AMIDA) | MICCAI 2013 (real 2D datasets) |
| Localization Microscopy Challenge (LMC) | ISBI 2013 (real and synthetic time-lapse 2D datasets) |
| 3D Deconvolution Microscopy Challenge (DMC) | ISBI 2013, ISBI 2014 (3D synthetic datasets) |
| Cell Tracking Challenge (CTC) | ISBI 2013, ISBI 2014, ISBI 2015<br>(real and synthetic 2D and 3D time-lapse datasets) |
| Overlapping Cervical Cytology Image Segmentation Challenge | ISBI 2014 (real and synthetic 2D datasets),<br>ISBI 2015 (real 3D dataset) |
| Gland Segmentation Challenge (GLAS) | MICCAI 2015 (real 2D datasets) |
| Image Stitching Challenge (ISC) | BioImage Informatics 2015, NIST (real 2D datasets) |
| Nucleus Counting Challenge (NCC) | BioImage Informatics 2015, NIST (real 2D datasets) |

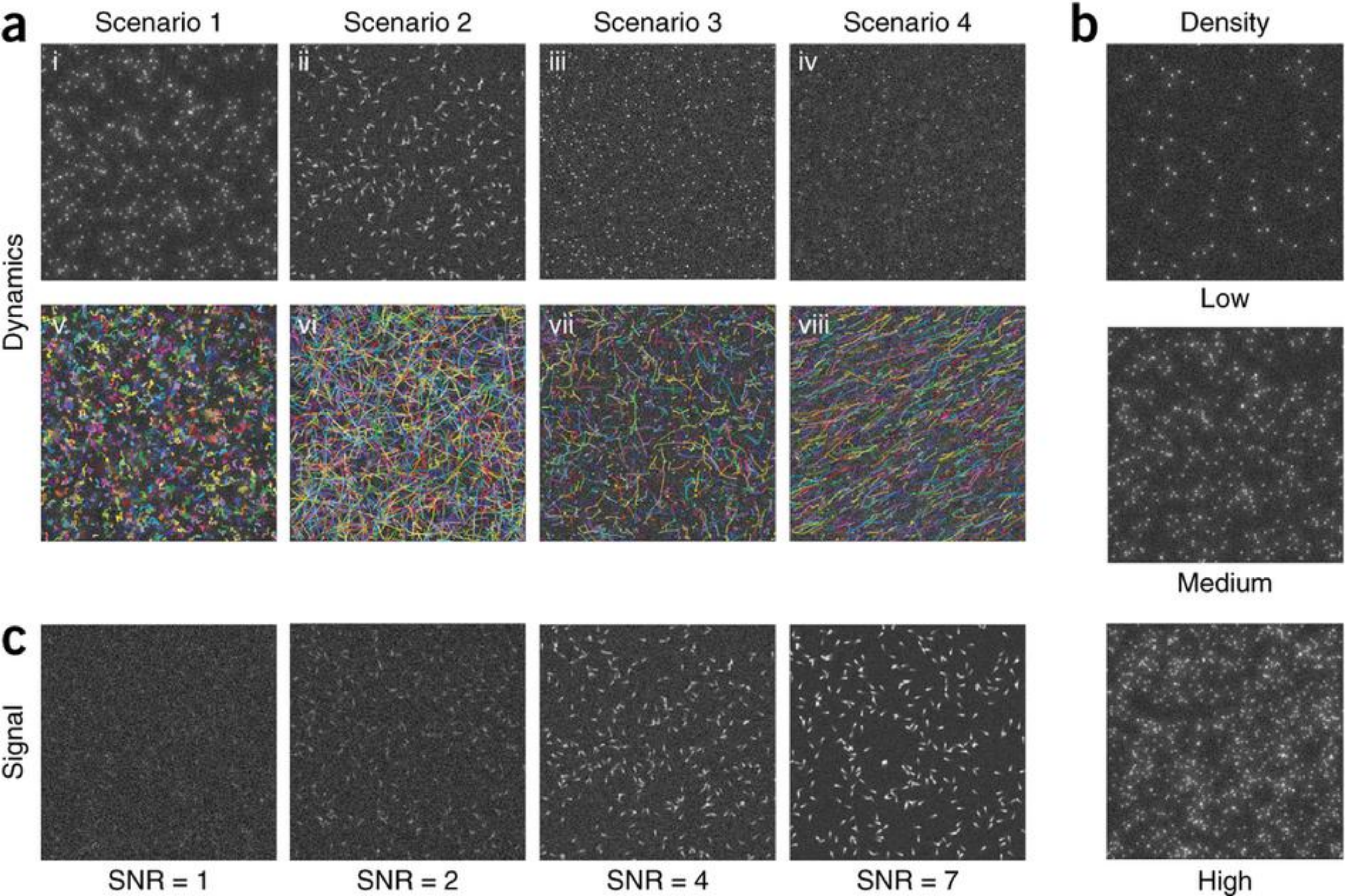
[Kozubek M: Challenges and Benchmarks in Bioimage Analysis, *Adv Anat Embryol Cell Biol*, 219, 2016]

# Challenges in BioImage Analysis

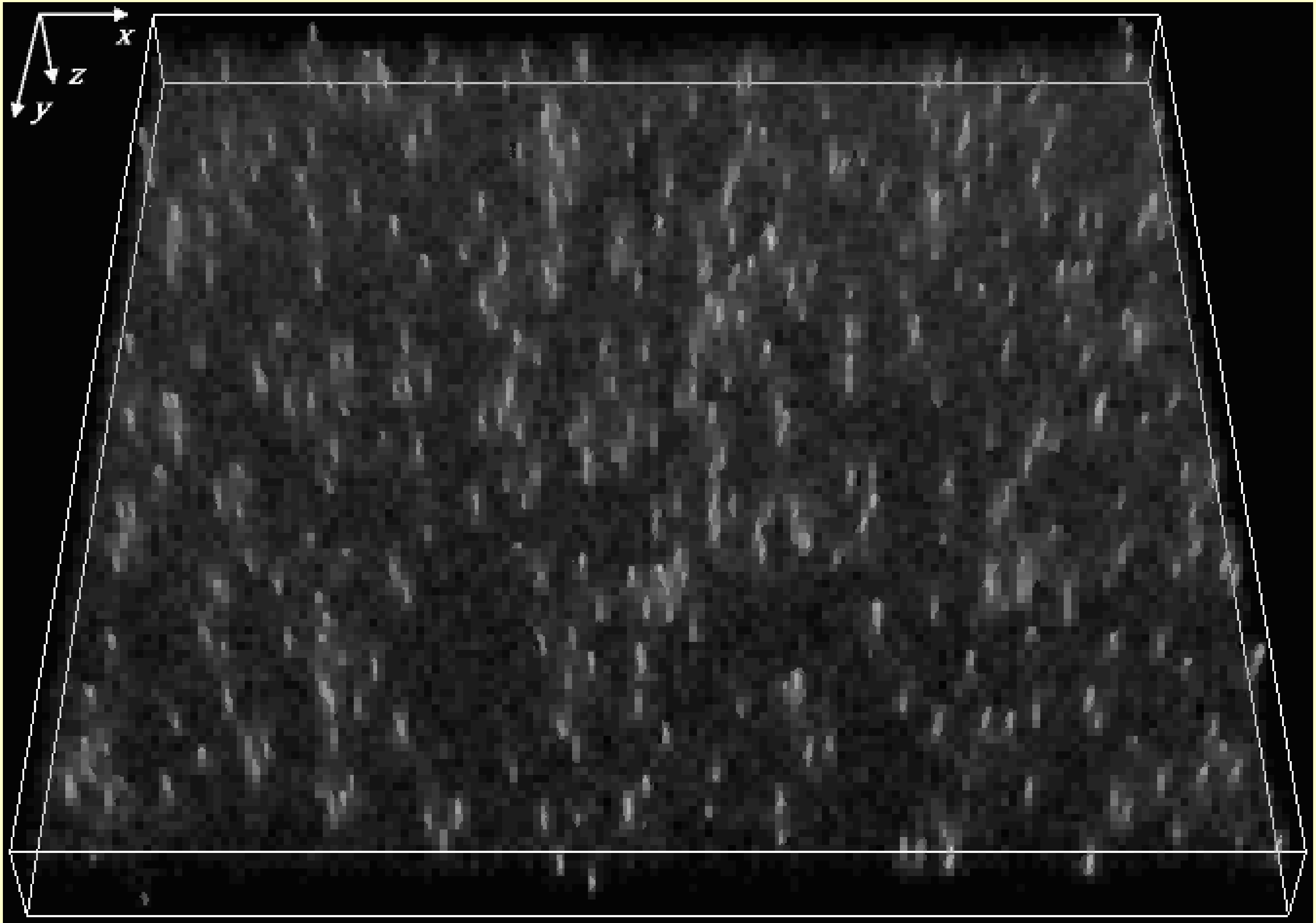


- Particle Tracking Challenge (PTC)
  - **The most cited and the first on synthetic datasets**
  - ISBI 2012, Erasmus MC + Institut Pasteur, 14 teams
  - Four different types of synthetic benchmark datasets:
    - Vesicles with Brownian (random-walk) motion (2D+time)
    - Microtubules (larger elongated particles) with directed motion (2D+time)
    - Receptors switching between Brownian and randomly-oriented directed motion (2D+time)
    - Viruses switching between Brownian directed motion with restricted orientation (3D+time)
  - Each of these four scenarios was simulated at 3 density levels and 4 SNR levels yielding 48 datasets in total
  - Optimal track pairing was found using Munkres algorithm and then five measures were computed to assess tracking accuracy: overall degree of matching not taking into account spurious (nonpaired estimated) tracks, penalization of spurious tracks, Jaccard similarity coefficient for track points as well as for entire tracks and RMSD between matching points
  - *Nature Methods* (Chenouard et al. 2014)

# Particle Tracking Challenge (PTC)



# Particle Tracking Challenge (PTC)



# PTC: Results and Open Access

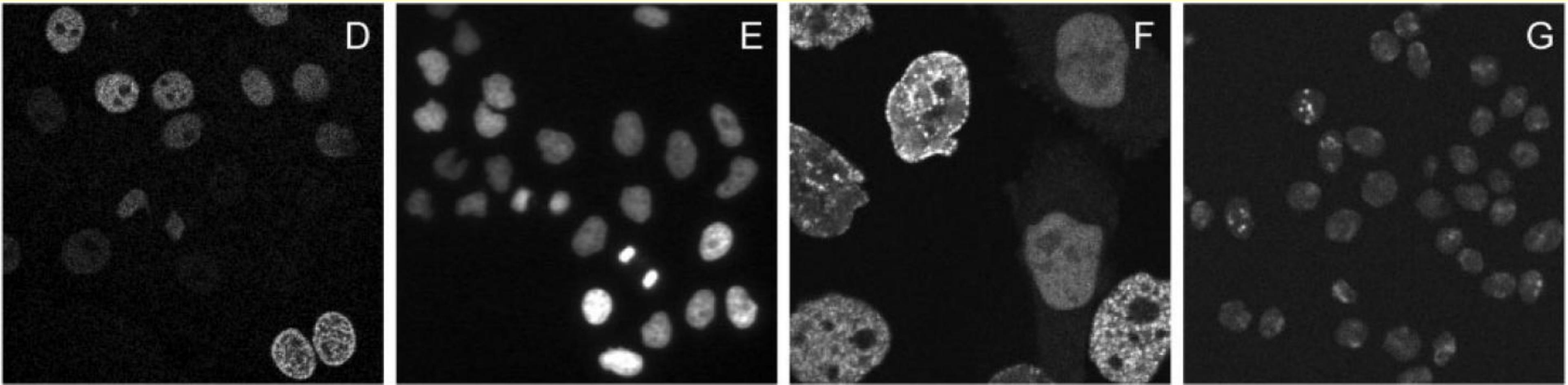
- Participants and submissions: 2012
  - Teams that submitted results: 14
- Rankings
  - Separate ranking for each dataset (no generally best method)
- Open access
  - All datasets, training data GTs, algorithms, results and evaluation tools available at PTC web: http://www.bioimageanalysis.org/track/

Login: anonymous@bioimageanalysis.org Password: erfg14d3

  - Open access paper about PTC and established benchmark:
    - *Nature Methods* 11 (3), 281-289, 2014 (WOS highly cited paper)

# Challenges in BioImage Analysis

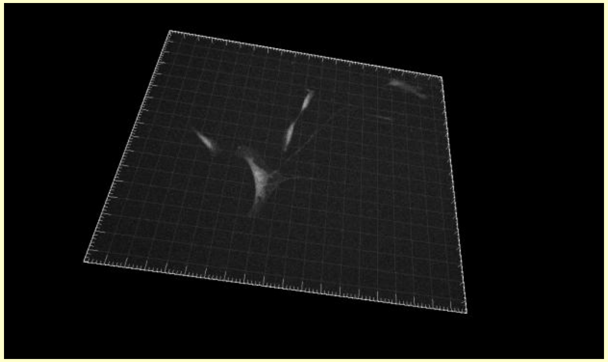
- Cell Tracking Challenge (CTC)
  - **The largest time-lapse dataset collection** (both 2D+time and 3D+time)
  - ISBI 2013, 2014 and 2015, Univ Navarra + Masaryk Univ + Erasmus MC
  - Real as well as simulated datasets of different types:
    - Mainly real fluorescence datasets from low density isolated cells or cell nuclei up to very complex developmental image series
    - Real phase contrast (PhC) and differential interference contrast (DIC) datasets
    - Simulated datasets with various settings of cell density, cell speed and SNR
  - Ground truth for real data created by 3 experts and major voting scheme
  - Segmentation quality is assessed using Jaccard similarity index, tracking quality using a specially developed measure
  - *Bioinformatics* (Maška et al. 2014)



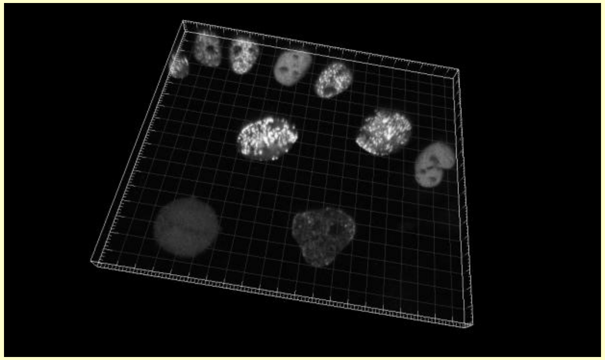
# CTC: Diverse Data Repository

- Diversity (of both real and synthetic data) with respect to
  - Cell types
  - Microscopy and experimental setup: 2D, semi 3D, full 3D
  - Nuclei and whole cells (for synthetic only nuclei)
  - Cell density, speed, frequency of mitotic events
  - Resolution: lateral, axial, temporal
  - Noise level / SNR
- New in the second edition
  - Brightfield modalities: Phase Contrast and DIC (only real data)
  - Developmental biology data – C. elegans (only real data)
  - Improved simulations
- New in the third edition
  - Drosophila development – highly challenging dataset (only real data)

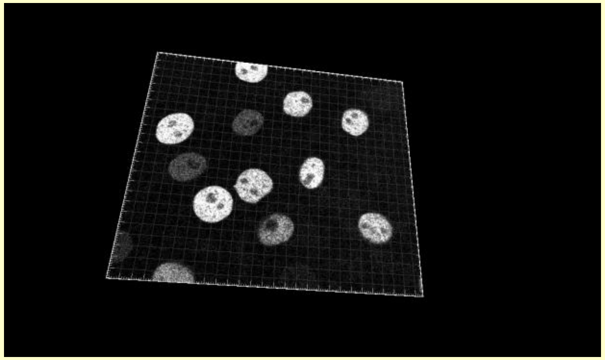
# CTC Real Fluorescence Datasets



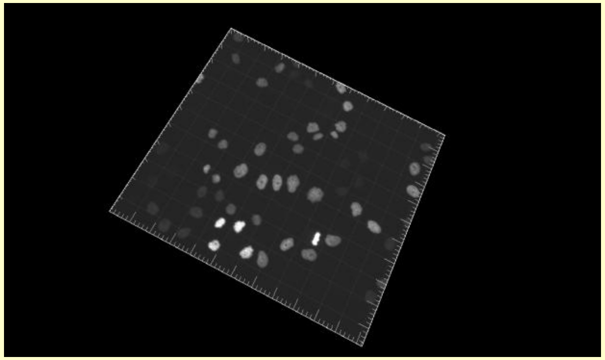
Fluo-C2DL-MSC



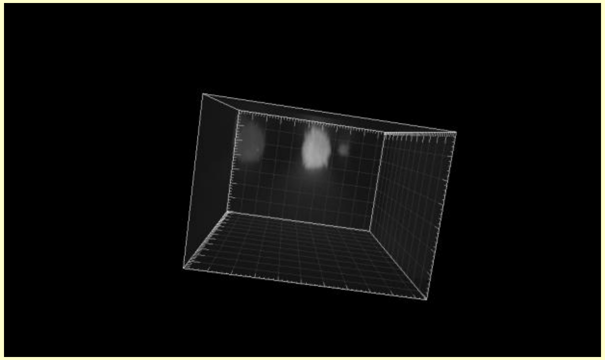
Fluo-N3DH-CHO



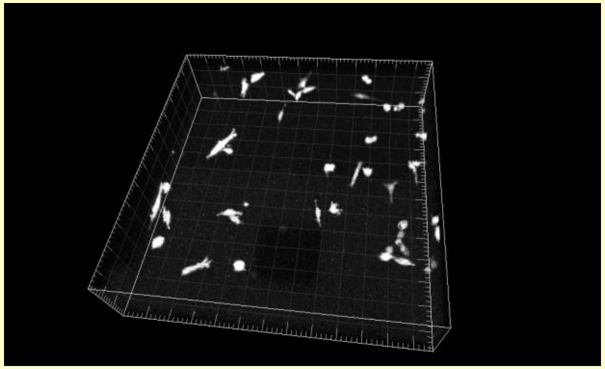
Fluo-N2DH-GOWT1



Fluo-N2DL-HeLa

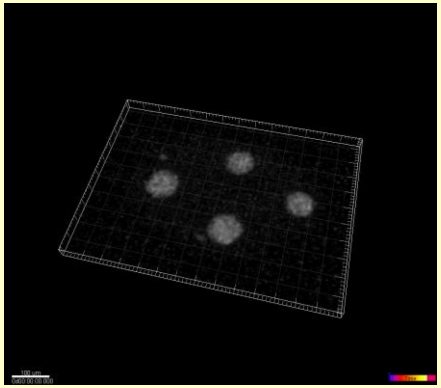


Fluo-C3DH-H157

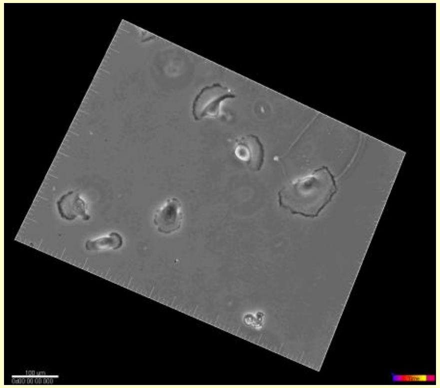


Fluo-C3DL-MDA231

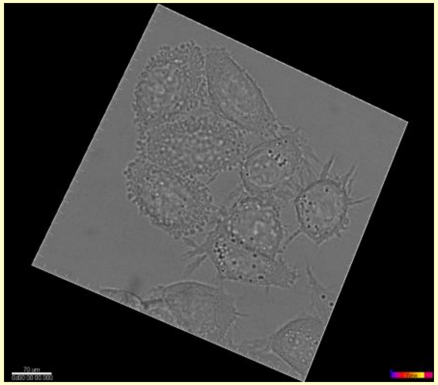
# CTC Real Fluorescence Datasets (added in the second edition)



3D nuclei Fluo-N3DH-

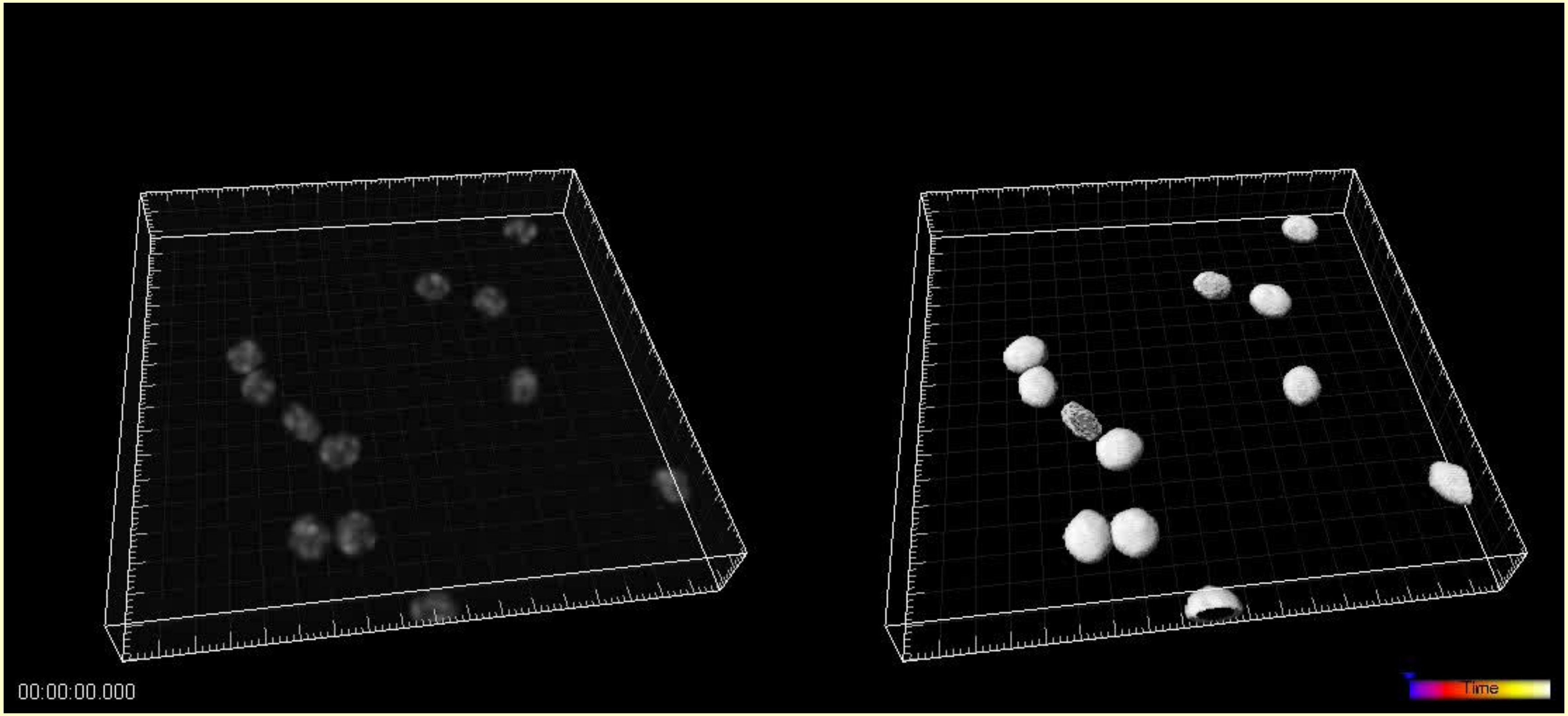


PhC-C2DH-U373



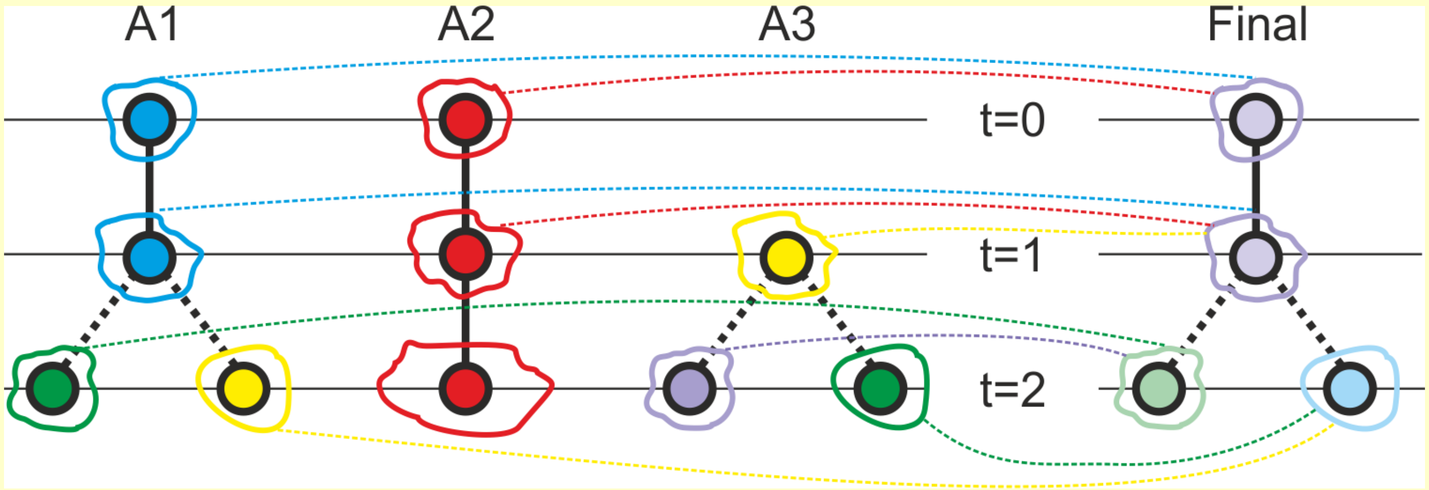
DIC-C2DH-

# CBIA Simulation Framework



# CTC: Ground Truth (GT)

- Synthetic datasets
  - GT provided inherently by the simulator
  - Complete tracking and segmentation information
- Real datasets
  - Manual annotations performed by three experts
  - Complete tracking information
  - Randomly selected slices manually segmented
  - Final GTs constructed based on majority voting:

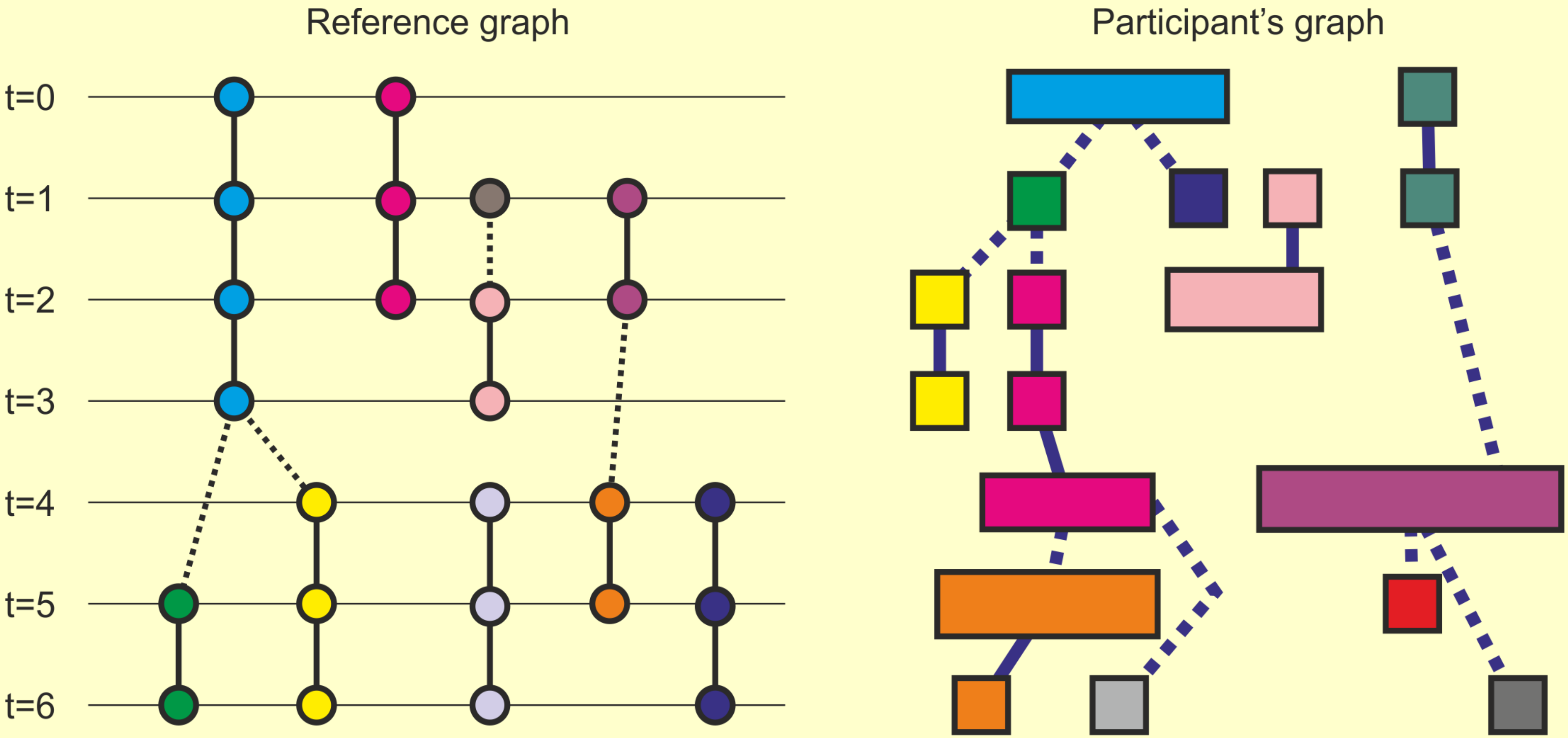


# CTC: Evaluation Measures

- Segmentation accuracy measure (SEG)
  - Based on Jaccard similarity index computed for each matched mask pair
  - Mean over all reference objects taken
- Tracking accuracy measure (TRA)
  - How difficult it is to change a computed graph to the reference one
  - Weighted sum of the number of basic operations that make both graphs identical
  - Operations and weights:
    - Delete node (1, one mouse click)
    - Split node (5, draw a divider)
    - Add node (10, draw a whole mask)
    - Delete edge (1, one mouse click)
    - Change edge (1, one mouse click)
    - Add edge (1.5, slightly more difficult than deleting an edge)

# CTC: Evaluation Measures

## Tracking accuracy measure (TRA)

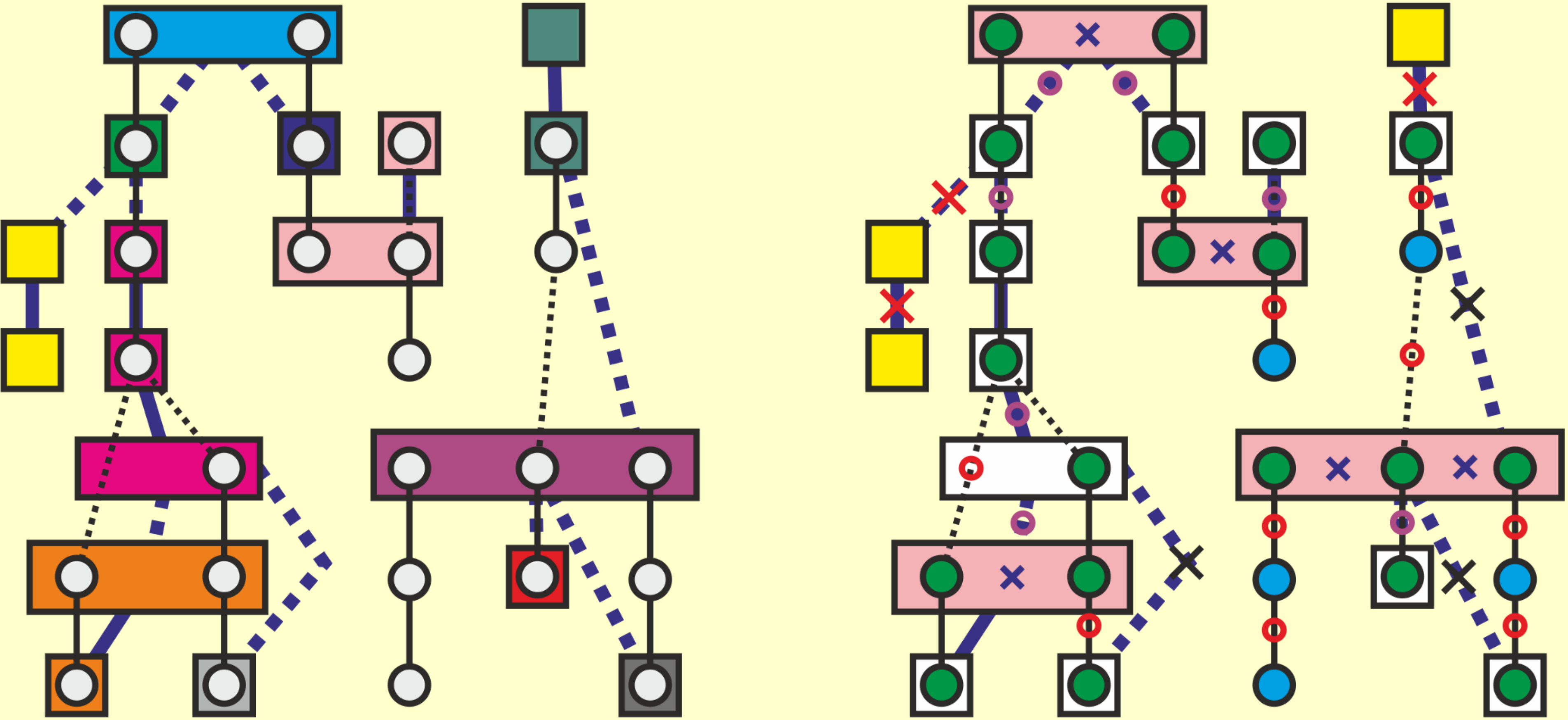


[Matula et al., Cell tracking accuracy measurement…, *PLOS ONE* 10(12), e0144959, 2015]

# CTC: Evaluation Measures

Participant’s graph overlaid with the reference one

Basic operations to be performed
TRA measure is 103



× Node splitting

🟨 Nodes to be deleted

× × Edges to be deleted

● Nodes to be added

▣ Matching nodes

○ Edges to be added

○ Edges to be changed

[Matula et al., Cell tracking accuracy measurement…, *PLOS ONE* 10(12), e0144959, 2015]

# CTC: Results and Open Access

- Participants and submissions: 2013 / 2014 / 2015
  - Registered participants: 67 / 59 / 58
  - Teams that submitted results: 8 /10 / 11
  - Out of them consistent results: 6 / 8 / 11 (2 the same)
- Rankings
  - Separate ranking for each dataset (no generally best method)
  - Normalized TRA and SEG measures added
  - Computation time used to distinguish equally performing methods
- Open access
  - All datasets, training data GTs, algorithms, results and evaluation tools available at CTC web: http://codesolorzano.com/celltrackingchallenge/
  - Open access paper about CTC and established benchmark:
    - *Bioinformatics* 30 (11), 1609-1617, 2014 (WOS highly cited paper)
    - To be continued: Second CTC paper being prepared

# Acknowledgement

- Benchmarking team (main activity – ISBI Cell Tracking Challenge)

UNIVERSITAS MASARYKIANA BRUNENSIS

Michal Kozubek

Martin Maška

Vladimír Ulman

David Svoboda

Pavel Matula

Petr Matula

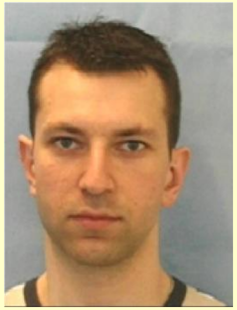
cima CENTER FOR APPLIED MEDICAL RESEARCH UNIVERSITY OF NAVARRA

Carlos Ortiz-de-Solórzano

Arrate Muñoz-Barrutia

Erasmus MC Universitair Medisch Centrum Rotterdam

Erik Meijering



Ihor Smal

- Further information
  - NEUBIAS 2017 Symposium – Talk on Thursday
  - [Kozubek M: Challenges and Benchmarks in Bioimage Analysis, *AAEC* 219, 231-262, 2016]