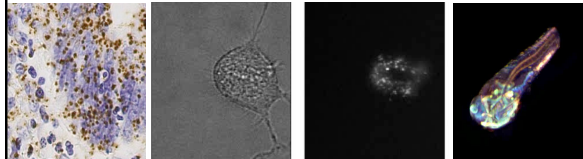


**Experimental design and image based screening**  
Image Analysis II, spring 2015.

Carolina Wahlby  
Professor in Quantitative Microscopy at SciLifeLab/the Centre for Image Analysis, Uppsala University  
And  
the Imaging Platform of the Broad Institute of Harvard and MIT, Cambridge, MA



## Outline

- Experimental design
  - What to think about when approaching a new problem
  - How do you know that your method produces correct results?
  - How to quantify a method's performance?
    - Dice score
    - Accuracy, precision, recall, and F-score
    - Hausdorff distance
    - ROC and AUC
    - Feature measurements and classification
  - Quality control
    - How can this be used in relation to the course projects?
- Inventing the wheel, or using available resources?
  - Short introduction to some different software tools
- Examples from High Throughput Screening in biomedicine

## A future scenario:



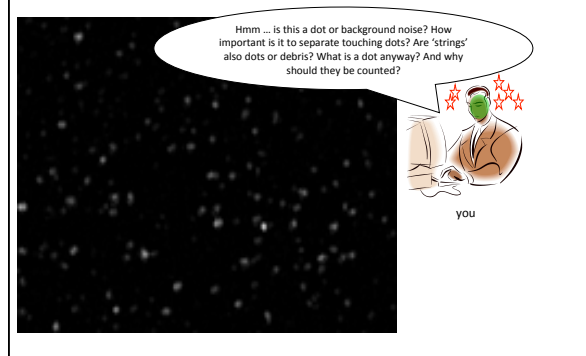
Can you write a program that automatically counts all the dots in these images?

Sure. Just give me the images. Easy thing, will deliver tomorrow.

An expert (and very important collaborator)

you

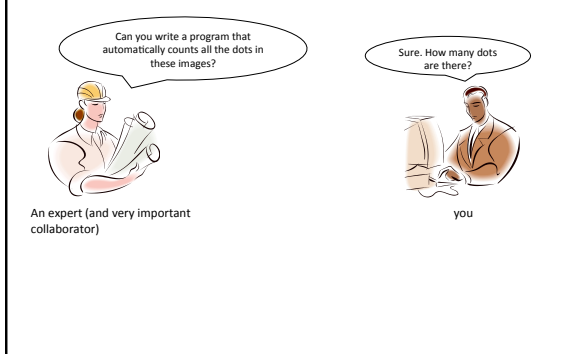
## Many hours later:



Hmm ... is this a dot or background noise? How important is it to separate touching dots? Are 'strings' also dots or debris? What is a dot anyway? And why should they be counted?

you

## A better future scenario:



Can you write a program that automatically counts all the dots in these images?

Sure. How many dots are there?

An expert (and very important collaborator)

you

## Method Evaluation

- You only know whether your method produces correct results if you know what results you expect.
- Thus: first try to solve a solved problem!
- "Ground truth" or "gold standard"
  - the correct solution to the problem
- Compare method's results to ground truth
  - Segmentation:
    - Dice similarity
    - Hausdorff distance
  - Detection/decision:
    - recall & precision
    - ROC curve and AUC
    - F-measure

### Scenarios of different hypotheses



- Hypothesis: The number of dots changes at a given treatment.
  - Do you have positive (treated) and negative (untreated) control samples?
  - Do you have more than one positive and negative control image so that I can measure the variance within the controls?
- Hypothesis: Dots appear when adding X.
  - Do you have any images of samples where X has not been added (only background)?
- Hypothesis: Dots cluster and increase in size when adding X.
  - Measurements such as shape, size and intensity may better quantify the change than a simple count?
- Hypothesis: Image based measurements of Y (hypothesis above) is more powerful than the standard approach Z.
  - How can we benchmark against Z?



### A clear hypothesis or goal will be helpful for experimental design



- What is the trade-off when it comes to improvements/consistency of image acquisition and/or sample preparation?
- Higher resolution vs number of spots per field?
  - using auto-focusing
  - imaging all samples using the same illumination source
  - avoiding glares and shadows etc
- Is this collection of images representative of the images you want to analyze?
- What do the worst images look like?



### Optimization and Evaluation

- Crucial for methods development
  - Does my proposed approach support the goal/hypothesis?
  - Is this approach 'better' than a previous existing approach?
- Requires 'ground truth'

### What is 'ground truth'

- The word comes from observations made on the ground when validating measurements from remote sensing (satellite or airplane)
- Often the output of visual assessment; manually drawn outlines, counts and/or visual classification (e.g. visual assessment of license plate numbers to optimize and validate an automated car toll system)
- Also referred to as 'benchmarking data'

### Ground-truth is needed for optimization/ training *and* testing/validation

- Training set
  - this is the data (images) you use to develop your method
- Test set
  - this is the data (images) you use to evaluate your method
- Do not mix!!!
- The training set will let you:
  - tweak your method until it does as you need it to
  - optimize the parameters of your method
- The test set will let you:
  - see how well your method works (accuracy, precision)
  - compare different methods

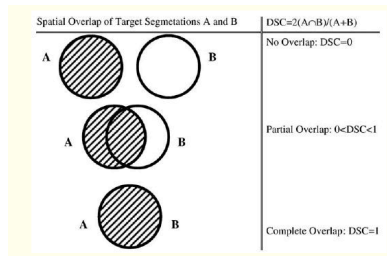
### k-fold cross-validation

- Used when you don't have enough data with ground truth
- Repeated training of parameters and method testing with different subsets of the data
  - Divide data into k subsets
  - Pick k-1 subsets for training, last subset for testing
  - Repeat so that each subset has been used for testing once
  - Average performance measure over all k repetitions
- Yields unbiased estimate of method performance
- Finally, train method parameters with all data
- When  $k = N$  (the number of images/cells/...)
  - leave-one-out cross-validation

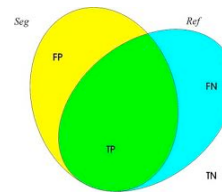
## Dice similarity coefficient

- A: ground truth segmentation
- B: your method's segmentation

The Dice similarity coefficient (DSC) measures spatial overlap  
 $DSC = 2(A \cap B) / (A + B)$



## Recall, precision, accuracy, and the F-measure



Recall (=Sensitivity) =  $TP / (TP + FN)$  will ignore FP  
 Precision =  $TP / (TP + FP)$  will ignore FN  
 Accuracy =  $TP / (TP + FN + FP)$   
 Specificity =  $TN / (TN + FP)$

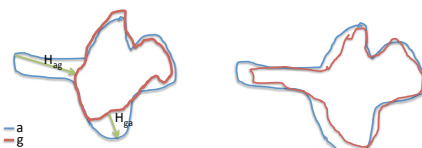
F-factor =  $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$   
 harmonic mean of recall and precision

DSC =  $2 * TP / (FP + FN + 2 * TP)$

Recall and Precision provide more detailed information than just the Dice score.

## Hausdorff distance

These two pairs of objects have similar Dice scores, but different Hausdorff distances.



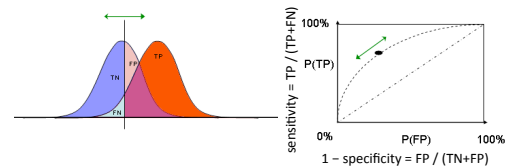
$$H_{ag} = \max(d_{\min}(a_i, g))$$

$$H = \max(H_{ag}, H_{ga})$$

## The ROC curve

### 'Receiver Operator Characteristics'

Plots true positive rate vs false positive rate as one parameter in the method is changed (e.g. changing the classifier boundary (threshold) in a binary classifier)  
 Often used for cost/benefit analysis.

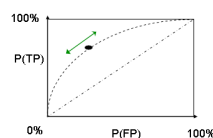


If the two populations completely overlap, the ROC-curve will be a diagonal line.

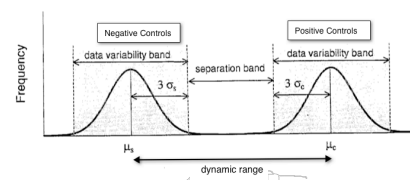
The better the separation of the classes, the larger the 'Area Under Curve' or AUC.

## AUC: area under curve

- Simply compute the area under the ROC curve; a larger area indicates better performance.
- Summarizes the ROC curve, but might not provide useful information.
- The F-factor is typically a better summary for a method's performance
  - or e.g. determine distance to the top-left corner



## discriminative power; the Z-factor



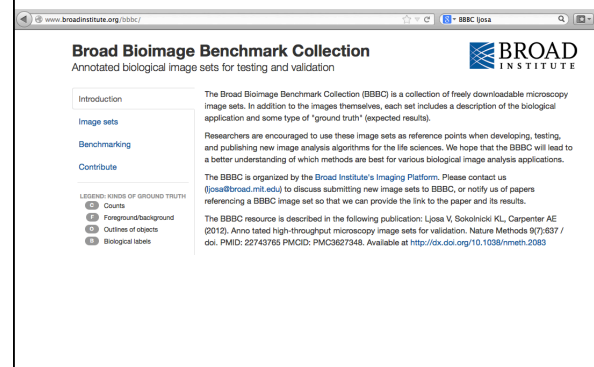
$$\text{Estimated Z-factor} = 1 - \frac{3(\hat{\sigma}_p + \hat{\sigma}_n)}{|\hat{\mu}_p - \hat{\mu}_n|}$$

A common metric when evaluating the 'power', or ability of an assay to discriminate between positive and negative control samples.  
 Note: What if distributions are not Gaussian?

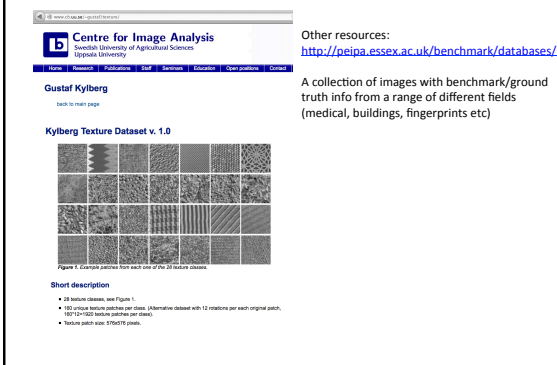
## Approaches for validation?



## Image collections with 'ground truth'



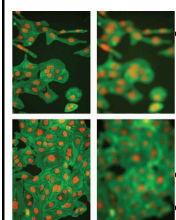
## Image collections with 'ground truth'



## Quality control

- Is the input what you expect?
  - i.e: similar enough to the training set
- Things that can go wrong:
  - staining sub-par
  - imaged region contains something unexpected
  - camera was out of focus
  - illumination not aligned (uneven illumination)
  - etc.
  - etc.
  - etc.
- Do you build a test for each possible issue?

## Data quality control



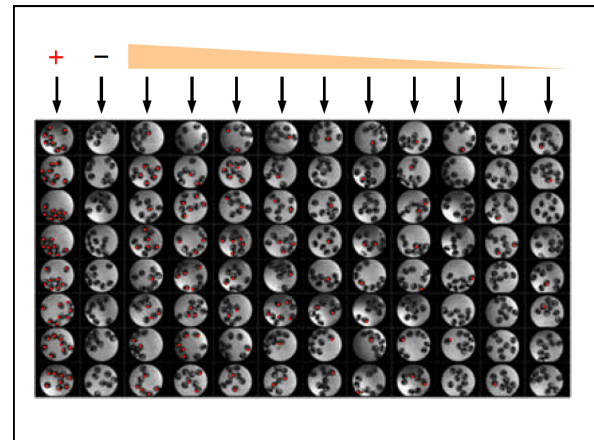
- If we have 100 000 images, and 10% of them are out of focus or contain debris that skews our measurements, there is a large risk of getting many false hits and also missing hits.
- How can we find these 'bad' images?
- What to do once we've found them?

## Three important means of QC

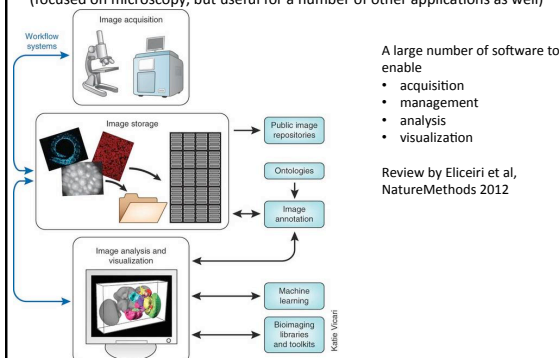
- good experimental design
  - helps to identify systematic errors (especially those linked for example with well position) and determine what normalization should be used to remove/reduce the impact of systematic errors
- the selection of effective positive and negative (chemical/biological) controls
- the development of effective QC metrics to measure the degree of differentiation so that images with inferior data quality can be identified and flagged/excluded.

## Controls

- Every experiment (data set) requires a positive and a negative control
  - controls = data with ground truth
  - every experiment can be tested for sensitivity and specificity
  - if these deviate from expectation, something went wrong!
- Every experiment should be independently replicated
  - in the ideal case
  - but it is expensive to do everything twice
- Controls:
  - known to be positive / negative
  - treated differently to look like positive / not treated
  - ...



## Free and open-source imaging software tools (focused on microscopy, but useful for a number of other applications as well)



Software name	Primary function	Website
µManager	Image acquisition	<a href="http://www.micro-manager.org/">http://www.micro-manager.org/</a>
ScanImage	Image acquisition	<a href="http://www.scanimage.org/">http://www.scanimage.org/</a>
OMERO	Image database	<a href="http://www.openmicroscopy.org/">http://www.openmicroscopy.org/</a>
Bisque	Image database	<a href="http://www.biomage.ucsb.edu/bisque/">http://www.biomage.ucsb.edu/bisque/</a>
OMERO searcher	Image content search	<a href="http://murphy-lab.web.cmu.edu/software/searcher/">http://murphy-lab.web.cmu.edu/software/searcher/</a>
Bio-Formats	Image format conversion	<a href="http://www.openmicroscopy.org/">http://www.openmicroscopy.org/</a>
ImageJ	Image analysis	<a href="http://rsbweb.nih.gov/ij/">http://rsbweb.nih.gov/ij/</a>
Fiji	Image analysis	<a href="http://www.fiji.sc/">http://www.fiji.sc/</a>
BioImageXD	Image analysis	<a href="http://www.biomedixd.net/">http://www.biomedixd.net/</a>
Icy	Image analysis	<a href="http://icy.biomageanalysis.org/">http://icy.biomageanalysis.org/</a>
CellProfiler	Image analysis	<a href="http://www.cellprofiler.org/">http://www.cellprofiler.org/</a>
Vaa3D	Visualization and image analysis	<a href="http://www.vaa3d.org/">http://www.vaa3d.org/</a>
FarSight	Visualization	<a href="http://www.farsight-toolkit.org/">http://www.farsight-toolkit.org/</a>
ITK	Bioimaging library	<a href="http://www.itk.org/">http://www.itk.org/</a>
OpenCV	Bioimaging library	<a href="http://opencv.willowgarage.com/wiki/">http://opencv.willowgarage.com/wiki/</a>
WIND-CHARM	Machine learning	<a href="http://code.google.com/p/wind-charm/">http://code.google.com/p/wind-charm/</a>
PSUID	Machine learning	<a href="http://psuid.org/">http://psuid.org/</a>
Ilastik	Machine learning	<a href="http://www.ilastik.org/">http://www.ilastik.org/</a>
CellProfiler Analyst	Machine learning and data analysis	<a href="http://www.cellprofiler.org/">http://www.cellprofiler.org/</a>
PatternUnmixer	Machine learning	<a href="http://murphy-lab.cbl.cmu.edu/software/PatternUnmixer2.0/">http://murphy-lab.cbl.cmu.edu/software/PatternUnmixer2.0/</a>
CellOrganizer	Machine learning, modeling and visualization	<a href="http://cellorganizer.org/">http://cellorganizer.org/</a>
KNIME	Workflow system	<a href="http://www.knime.org/">http://www.knime.org/</a>



Wayne Rasband

## Benefits of open-source software

- Educational value: anyone can go in and look at the source and learn.
- You often have the possibility to add your own algorithms.
- A user community for an open-source software will often be a more responsive and efficient source of help than what can be provided through an expensive service package for a commercial software.
  - Warning: Always validate the code before trusting it. Everyone makes mistakes...
- 'Reproducible research': with an open source solution, you can provide your analysis pipeline as part of the supplementary material of your published paper (along with data), and use the macro recorder of Image J to document exactly what you do with each image.
- You can easily share your analysis approach with colleagues.

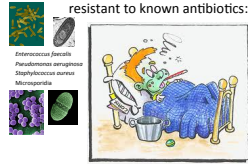
## Image based screening

- Screening: to have a large number of samples and identify those that deviate from the norm. The deviation could either be well defined or undefined.
- Many applications other than biomedical (control of product quality, damage detection, sorting etc.)
- Focus today: high throughput screening with biomedical applications.

## Image-based drug screening using model organisms

### Current problem:

People infected with bacteria resistant to known antibiotics:



### Possible solution:

Collections of thousands of different chemical compounds are available.

Perhaps one of them works as a drug?

### Now what?

Take a thousand sick patients and try a different chemical on each?

## Using *C.elegans* to search for novel anti-infectives

1. Infect worms with human pathogens (bacteria)



2. The worms get sick

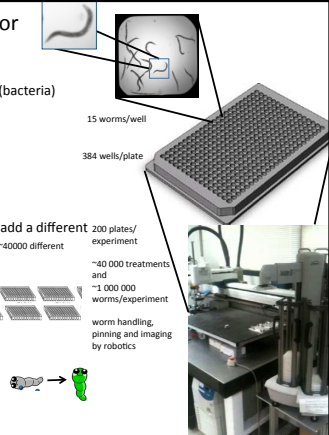


3. Place sick worms in wells, add a different compound to each well (~40000 different potential drugs)



4. Wait

5. Figure out if any treatment cured the worms: 'live/dead scoring'

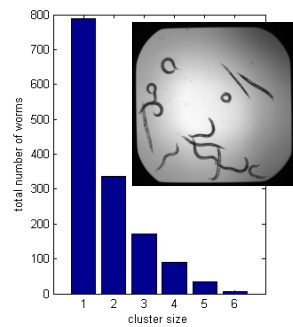


## Challenges with per-worm measurements:

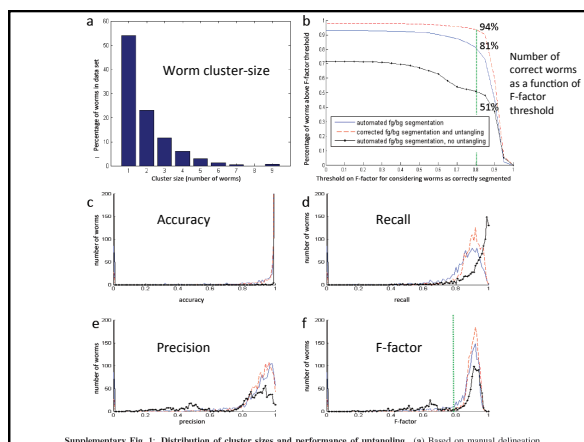
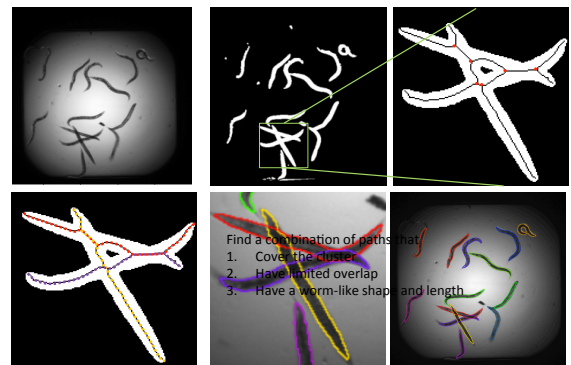
Worms touch, overlap and cluster.

In a typical screen on anti-infectives:

- 15 worms in each well to maximize statistics while minimizing the cost of the screen.
- ~50% of worms are clustered (based on visual examination of 100 wells, 1500 worms)



## How to 'untangle' worms



## Take home message

Knowledge about the image formation, possibilities and limitations, can greatly improve the scientific value of an experiment.

A better understanding of digital image processing, possibilities and limitations, can greatly improve the scientific value of an experiment.

A better communication between experts of from different fields can greatly improve the scientific value of an experiment.

Metrics and data for optimization/training and testing/validation should be considered already at the design of a project.

Carolina Wahlby, carolina@cb.uu.se