Can machine intelligence help?

Quantitative analysis of large-scale single cell-based screens

Peter Horvath

RISC (**R**NAi **I**mage-based **S**creening **C**enter), LMC

High-Content Analysis – San Francisco 2010 Jan.

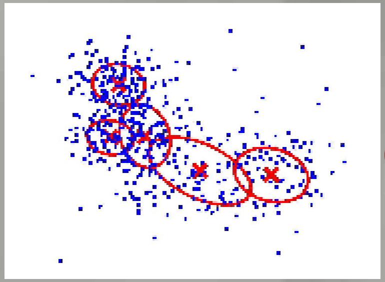
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

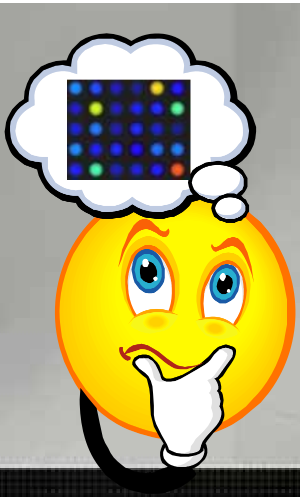
Roadmap

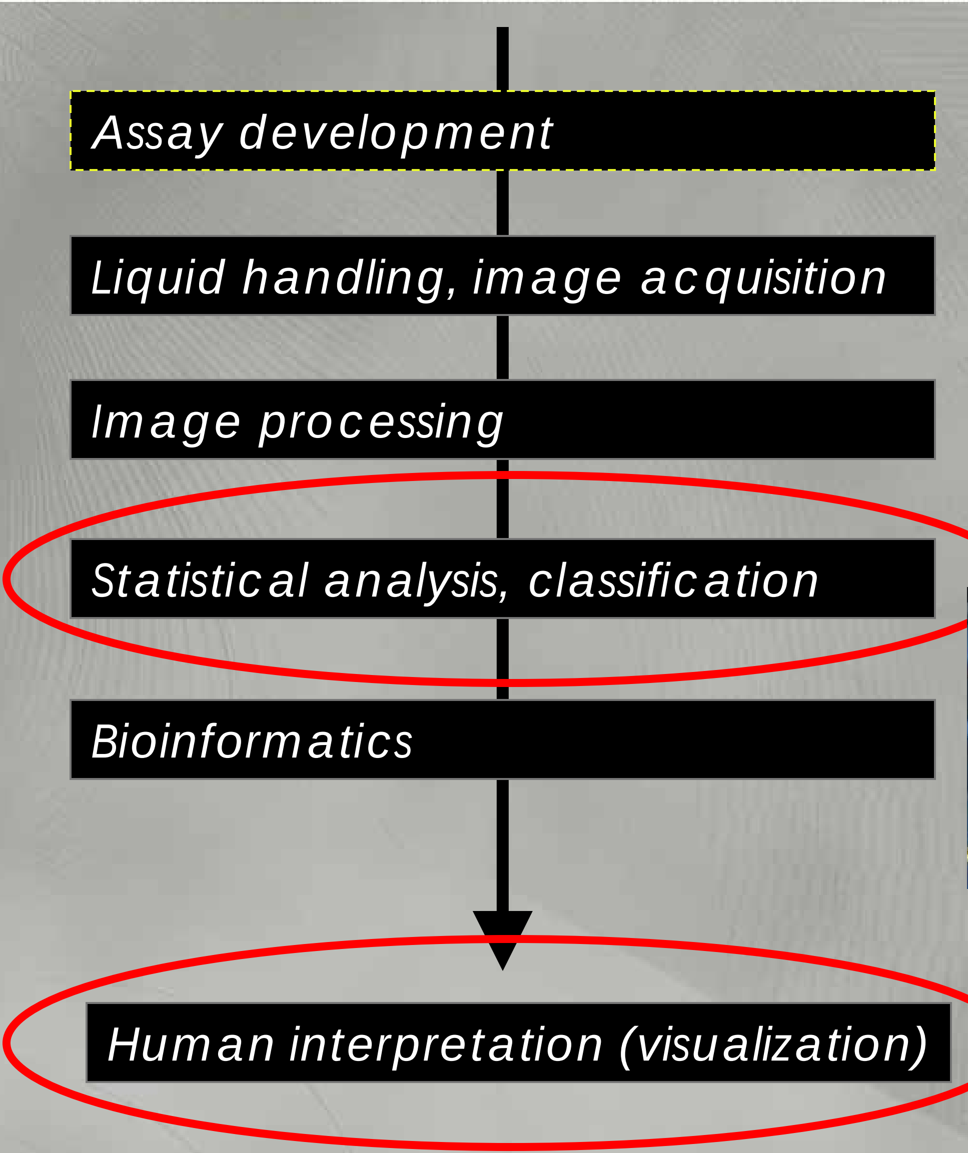
- Data processing pipeline
- From biology to image processing

- Advanced Cell Classifier
- Screen quality – method selection
- The human factor

The data processing pipeline







Assay development

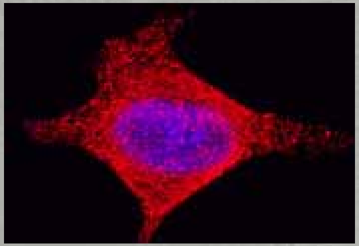
Liquid handling, image acquisition

Image processing

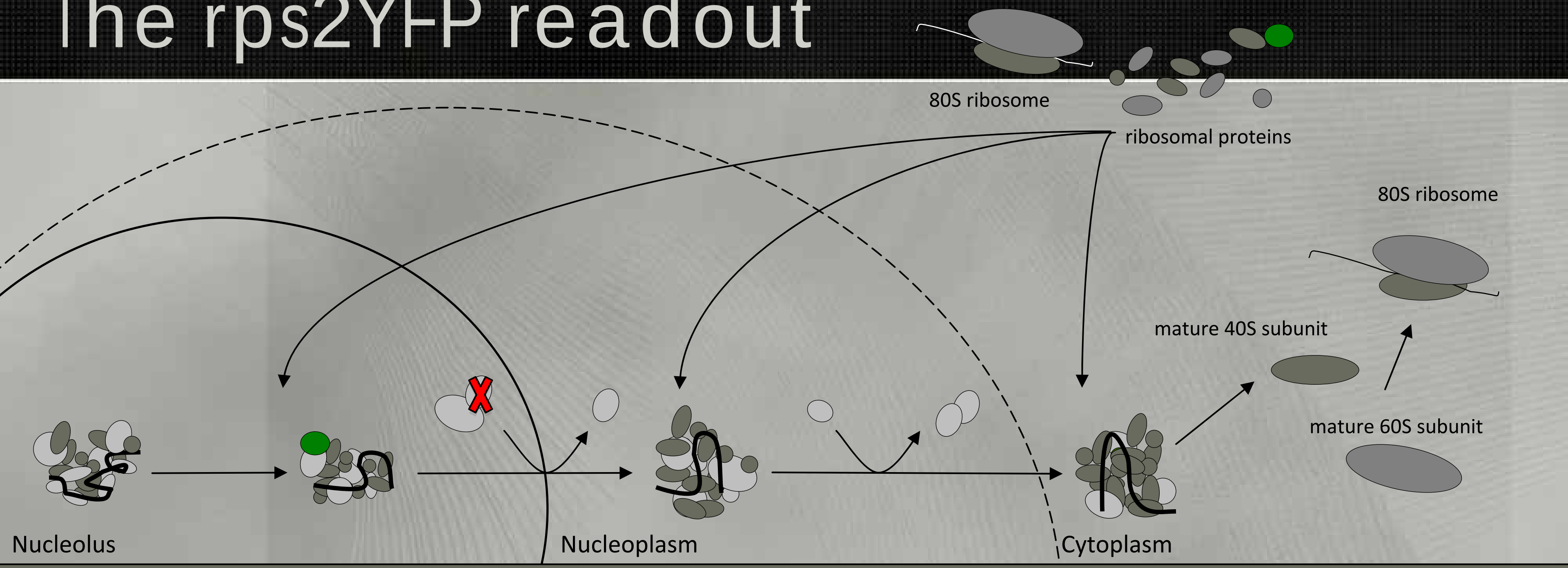
Statistical analysis, classification

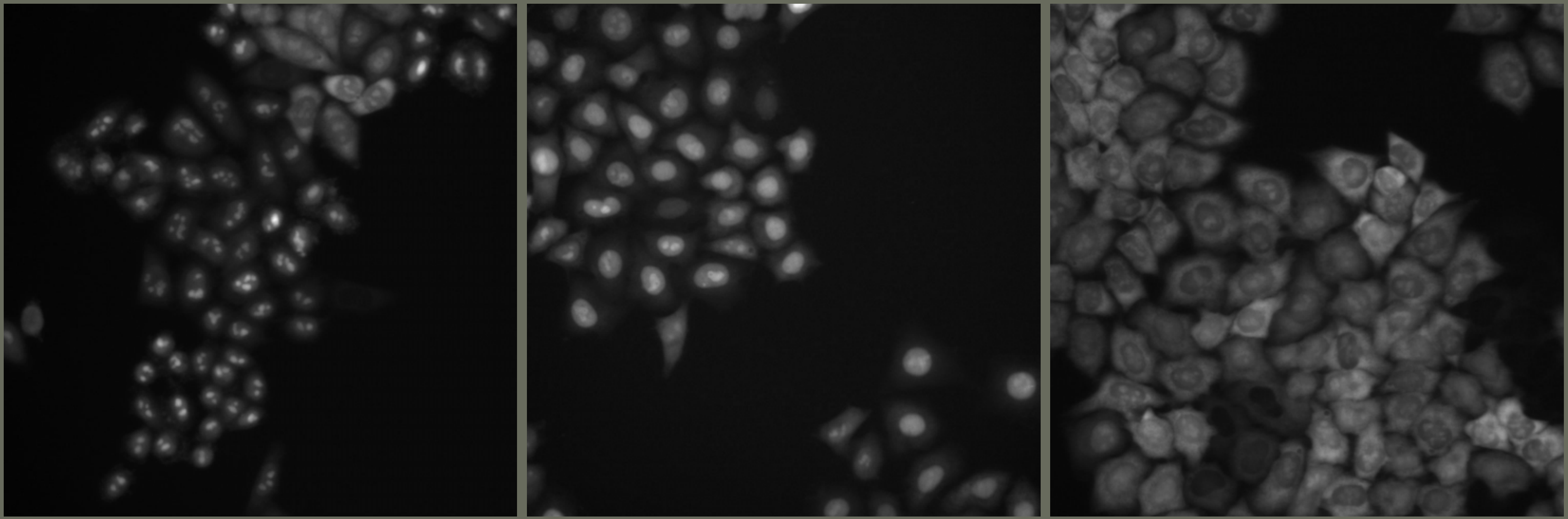
Bioinformatics

Human interpretation (visualization)



The rps2YFP readout





Genome-wide siRNA screen

- 22.000 genes 4 oligo/gene
  - Over 2.000.000 fluorescent images
  - Cell based analysis with ~80.000.000 cells
- Advanced Cell Classifier project
  - classification accuracy: 93% (10 fold c.v., ANN)
  - Z factor: 0.755;
- Computational time:
  - Segmentation: 1.5 hour/plate* ~ 300 plates
  - Classification: ~ 1 hour/GW screen* ~ 80.000.000 cells

Z factor	
0.5-1.0 –	excellent
0.0-0.5 –	marginal
< 0 –	overlap

**32 core server, 2.4 GHz, 80 GB RAM*

Image processing

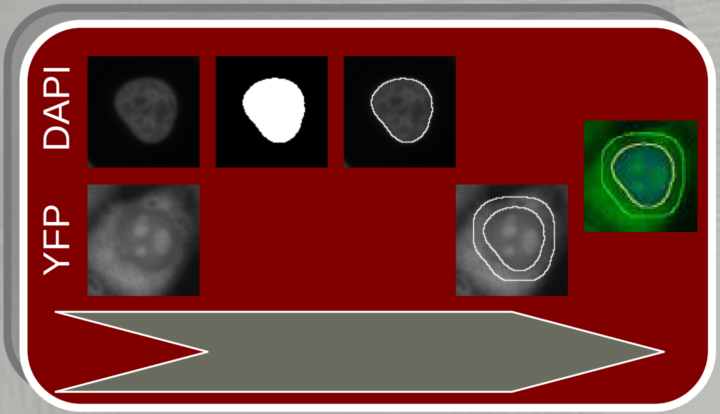
CellProfiler™
cell image analysis software

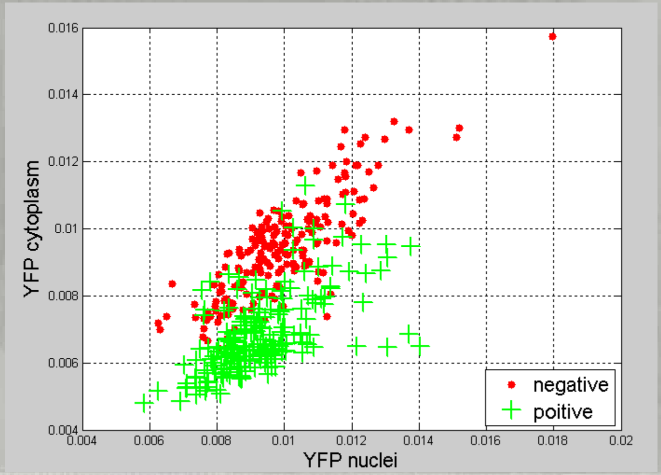
- CellProfiler – MatLab based software for high-throughput manner

- Speed: ~1 image/sec
- ~500.000 cell/hour

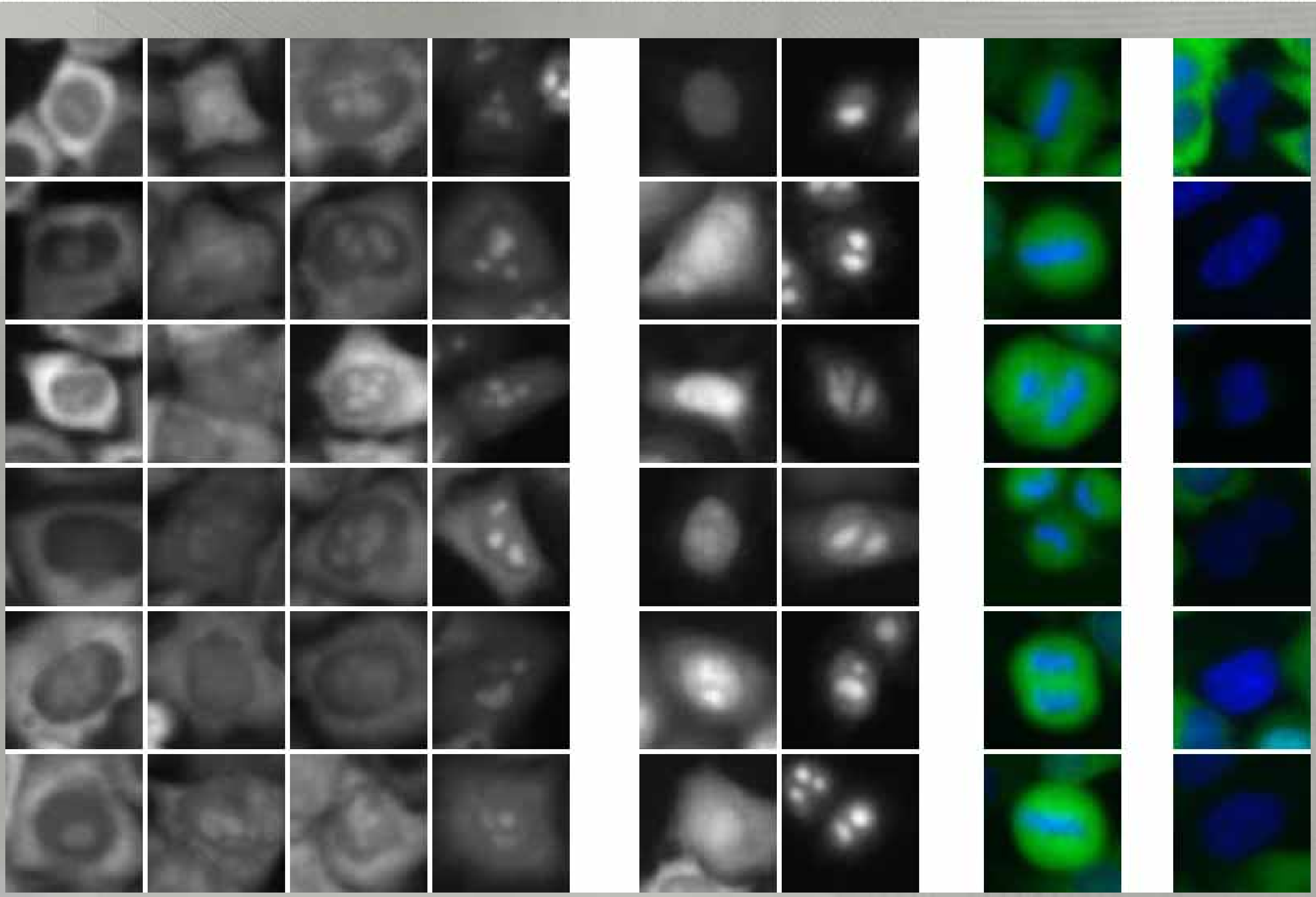
- Nuclei extraction on DAPI, extended ring on YFP

- 30 different features/cell
  - Intensity mean and std values
  - Morphological descriptors
  - Texture features





Advanced Cell Classifier I.



Mitotic cells

Normal cells (1-4)

Hit cells (1, 2)

No YFP signal

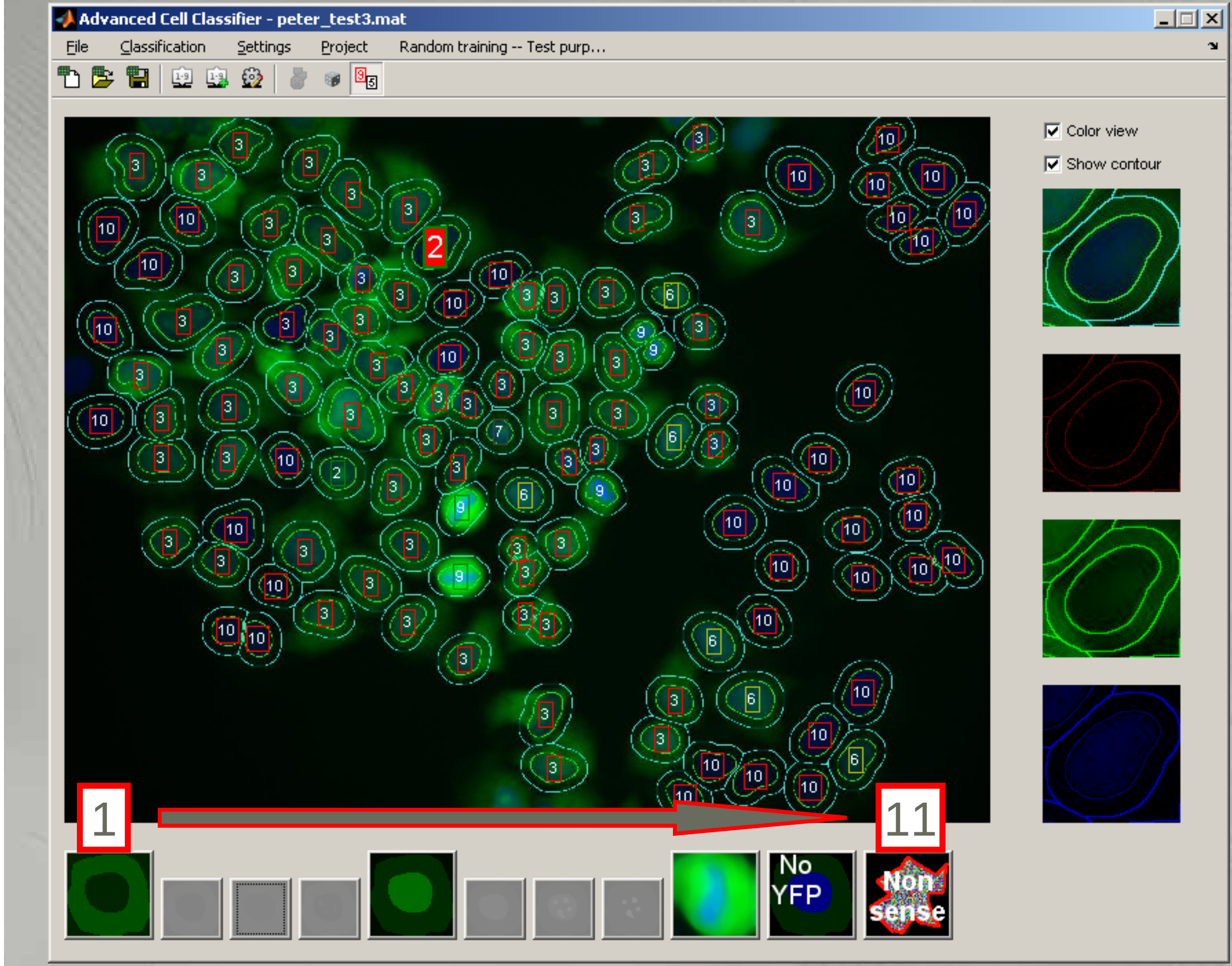
Training, classification

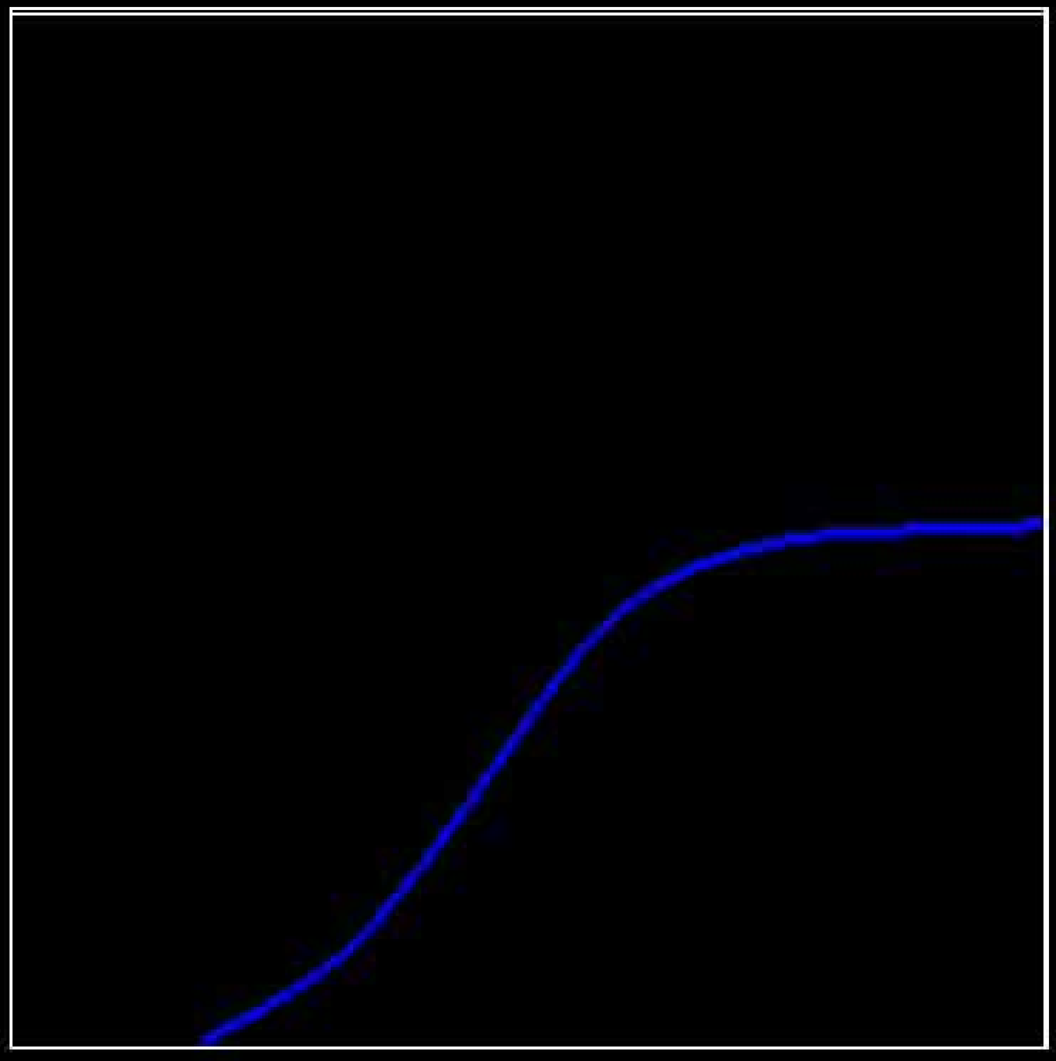
Advanced Cell Classifier

http://acc.ethz.ch

- Custom written software
- Easy training by clicking
- Predefined phenotypes and subtypes
- Quick prediction
- Machine learning and simple feature-based statistics
- Prediction of the entire screen and quick report (pdf, html, xml, csv)

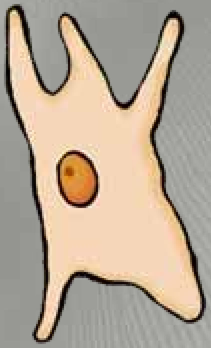
- Available learning methods:
  - Neural network
  - Support vector machine
  - Random forest
  - Logistic
  - 20+ more

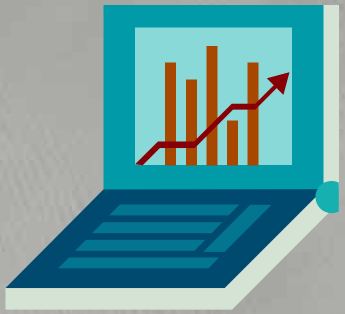




Screen and method quality

- Biology or method?





- What to maximize
  - Distance between the controls and std.
  - Accuracy of the analysis
  - Speed
- Best method

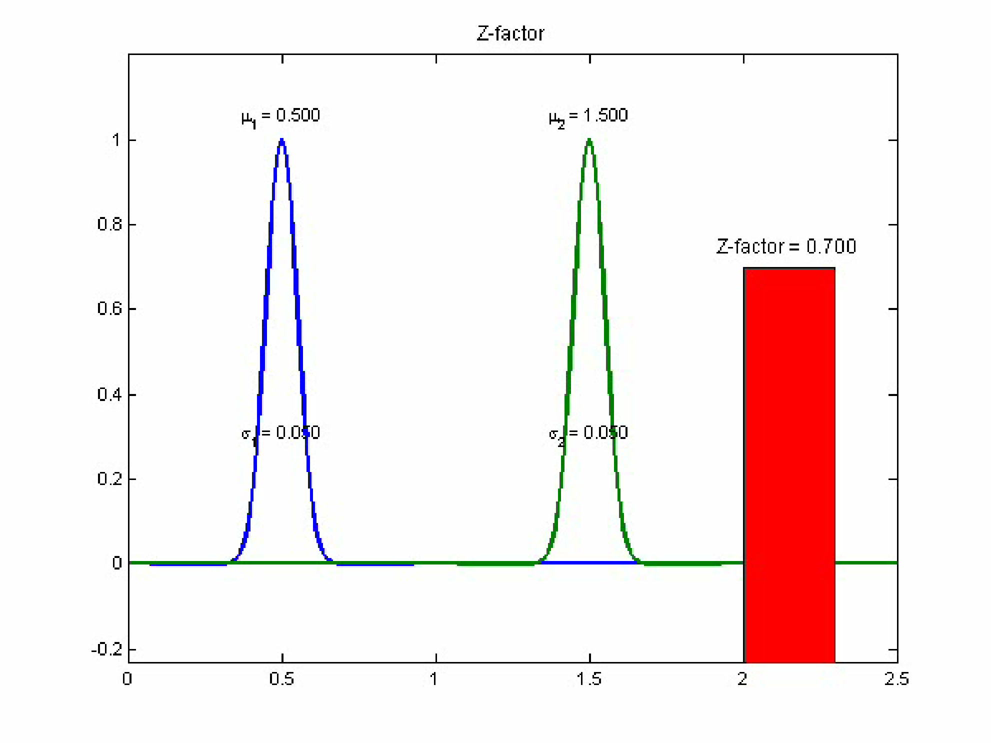
Z-factor and Cross Validation

- Z-factor
  - Metrics between two data point set

$$Z = 1 - \frac{3(\sigma(hr^{+}) + \sigma(hr^{-}))}{\left|\mu(hr^{+}) - \mu(hr^{-})\right|}$$

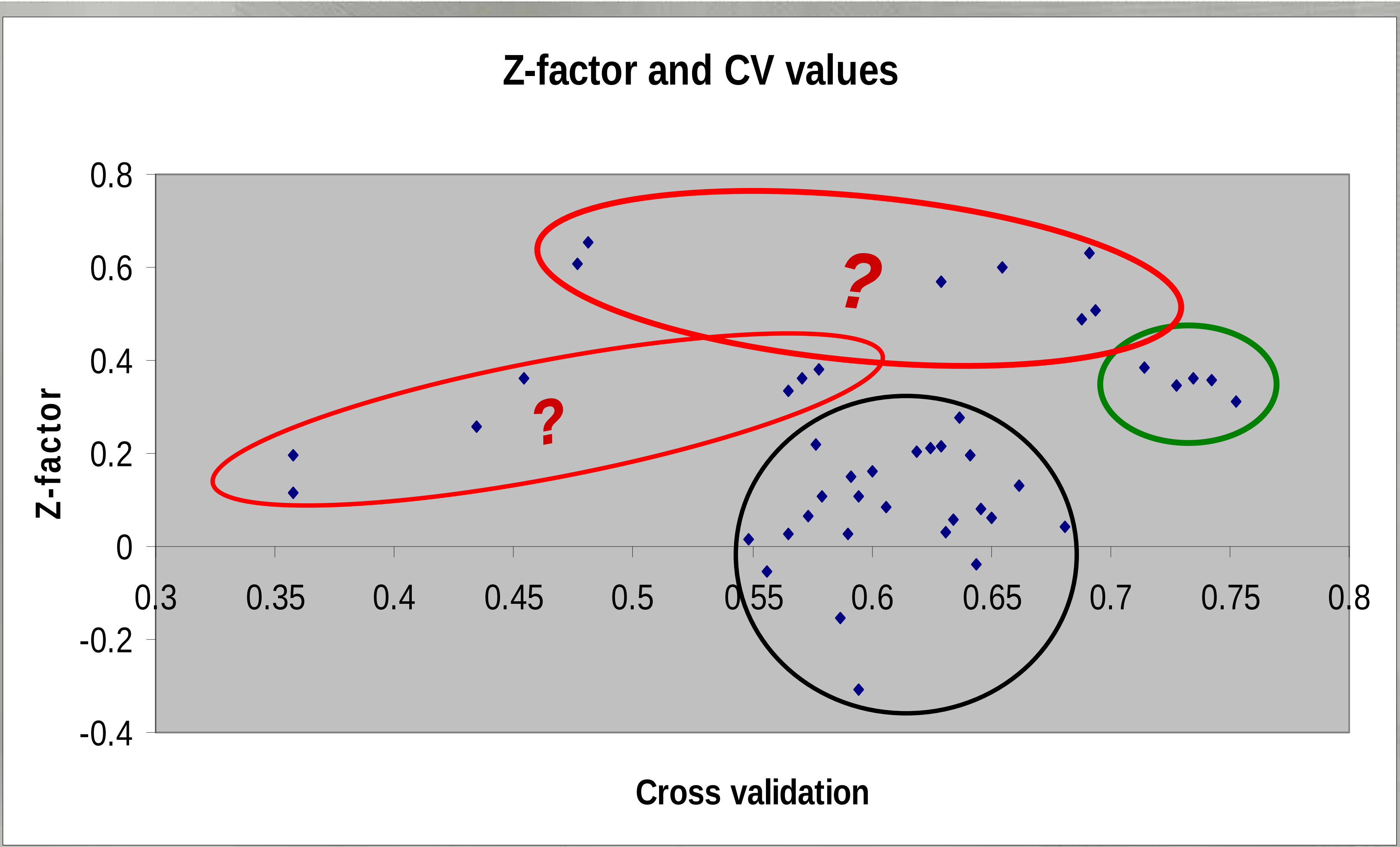
Z factor

0.5-1.0 –	**excellent**
0.0-0.5 –	**marginal**
< 0 –	**overlap**

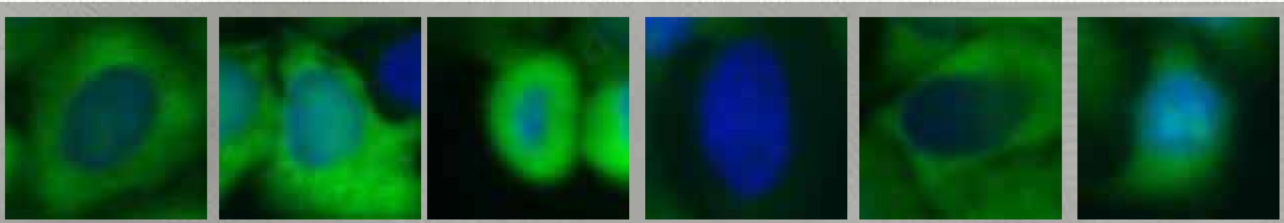


- Cross validation
  - How the results of a statistical analysis will generalize to an independent data set
    - K-fold
    - Leave-1-out

CV vs. Z-factor for supervised methods



The human factor



😎	N_1	N_2	M	D	N_3	H_1
😕	N_1	N_2	M	D	N_2	H_1
😘	N_2	N_3	M	D	N_3	H_1
😳	N_1	N_3	M	D	H_2	H_1

?	x	ok	ok	x	ok

😎 😕 😘 😳

N_2 N_3

The human factor II.

Confusion between field experts

	😎	😕	💋	😳
😎	100	72.2	64.6	67.6
😕	72.2	100	71.2	71.4
💋	64.6	71.2	100	63.8
😳	67.6	71.4	63.8	100

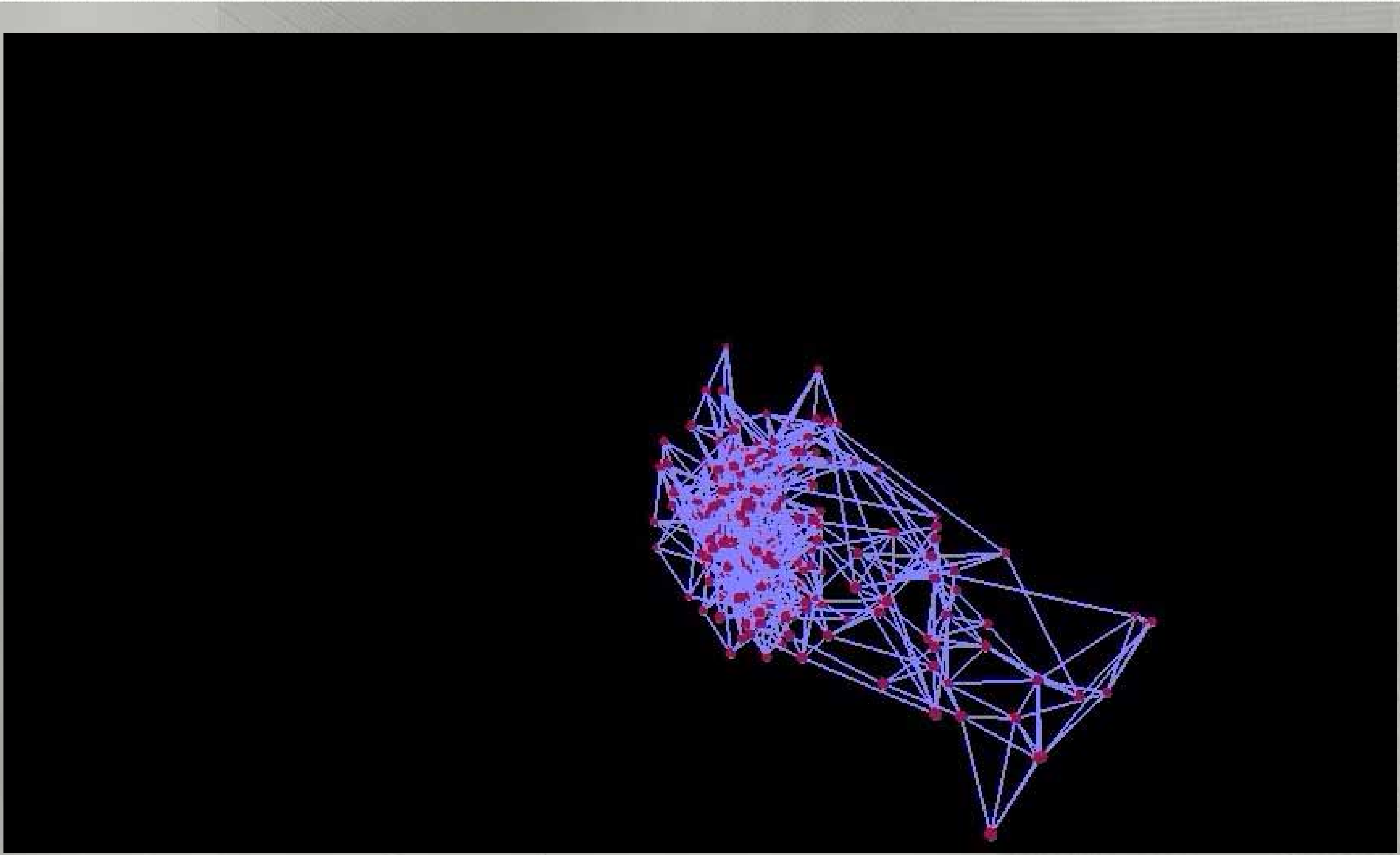
Accuracy
63.6% (85.2)
63.4% (88.2)
70.2% (87.0)
54.6% (86.2)

10-fold cross-validation using Logistic method

↓

😎 😕 💋 😳 ***After correction*** → **77% (93.8)**

Semi-supervised learning



Summary and future directions

- Classification
  - Semi-supervised learning (*“10 clicks from the hits”*)
  - Unsupervised classification
- Human factor
  - Worth to consider and reduce (with multiple independent labeling; 2-1, all-all)

Thank you for your attention!

http://acc.ethz.ch

http://www.lmc.ethz.ch/People/PeterHorvath

Peter.Horvath@lmc.biol.ethz.ch