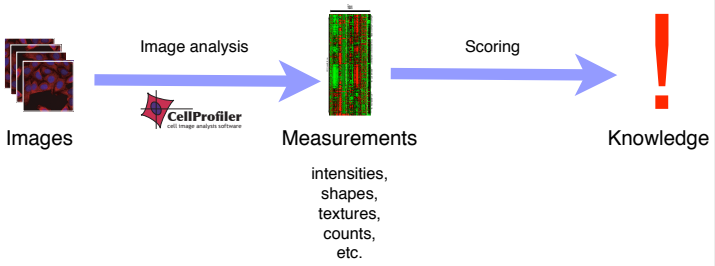
Large-scale learning of cellular phenotypes from images

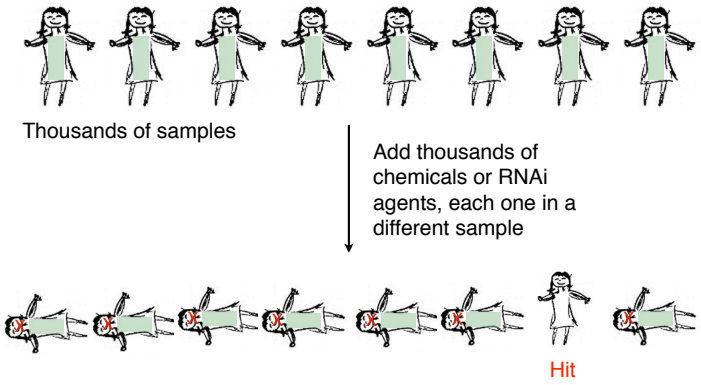
Vebjorn Ljosa, Piyush B. Gupta, Thouis R. Jones,
Eric S. Lander, and Anne E. Carpenter

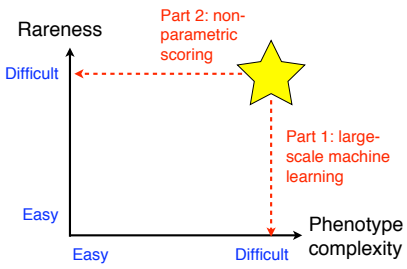
RECOMB Systems Biology
Cambridge, MA – 2009-12-06

From images to knowledge in high throughput



Screens





Simple methods

Few assumptions

Little modeling

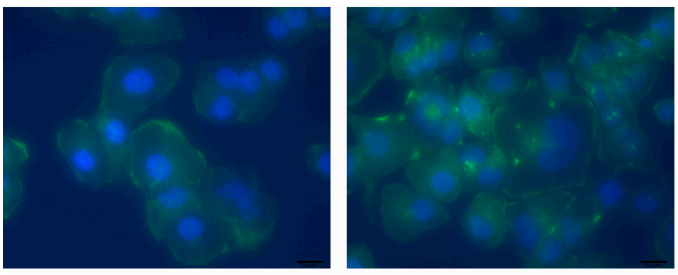
:
:

Lots of data

The phenotype of motile T47D cells

Normal T47D cells

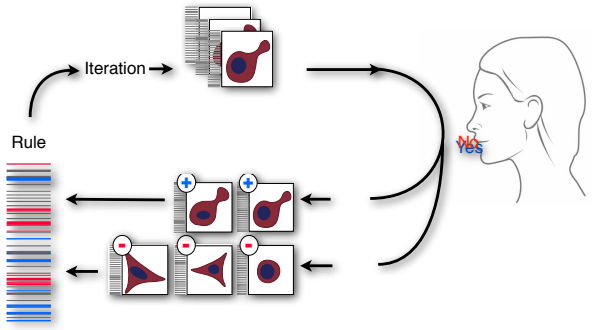
Features associated with cell motility: lamellipodia, filopodia, polarized cell shape, F-actin nucleation at filopodia, less clumping



Unstimulated

Stimulated by heregulin

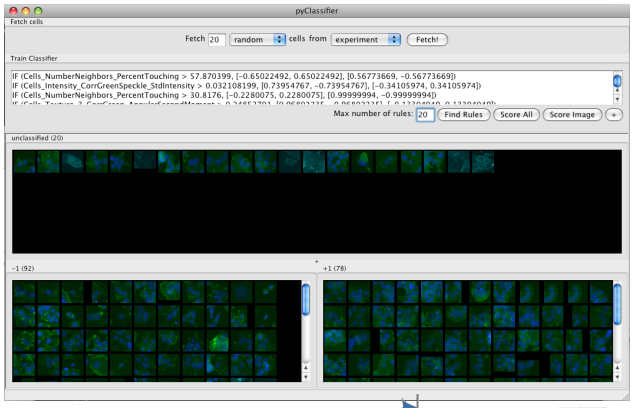
Iterative machine learning



CellProfiler Analyst
data exploration software

[Jones et al., PNAS, 2009]
Using gentle boosting [Friedman et al., 1998]

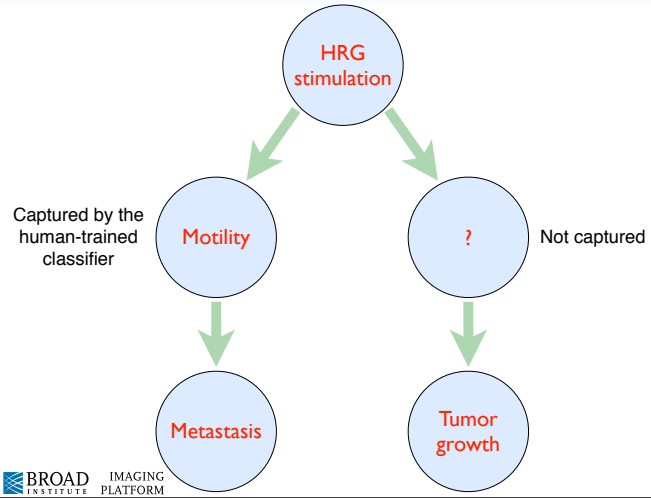
Built training set of ~300 cells



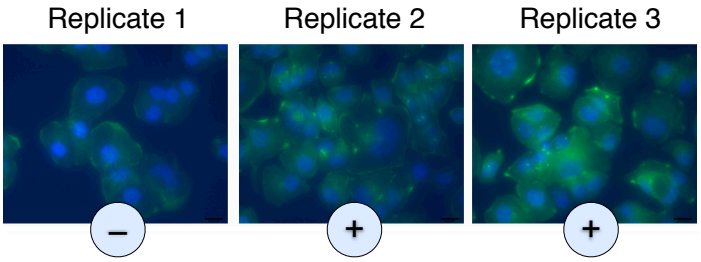
CellProfiler Analyst
data exploration software

Adam Fraser

Why cut out the human?



Labeling for automatic training set



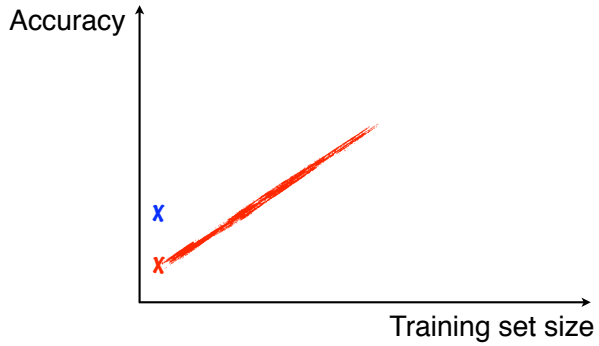
Unstimulated

45 % motile cells

Stimulated by heregulin

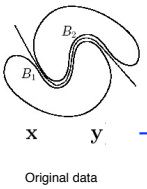
55 % motile cells

Two ways to improve the classifier



See [Banko & Brill, 2001]

Random features



$\mathbf{x}$ $\mathbf{y}$

Original data

The kernel trick

$k(\mathbf{x}, \mathbf{y})$

Kernel function

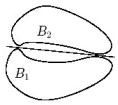
$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \approx \mathbf{z}(\mathbf{x})' \mathbf{z}(\mathbf{y})$

Inner product used, e.g., by SVM.

Random features
[Rahimi and Recht, NIPS, 2007]

$\mathbf{z}(\mathbf{x})$ $'\mathbf{z}(\mathbf{y})$

Datapoints in a fairly low-dimensional space (a few hundred dimensions) spanned by random Fourier bases

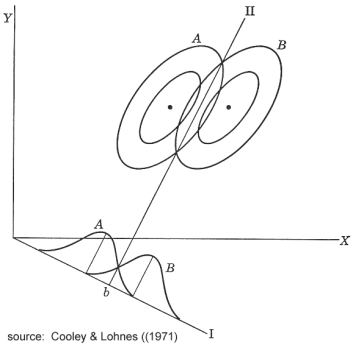


Linear discriminant on random features

7.6 million training cells, 130 measurements

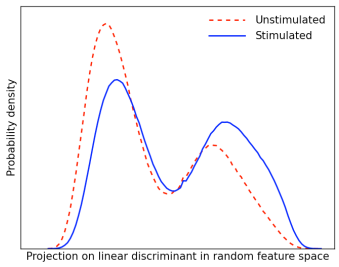
Mapped into 250-dimensional random feature space

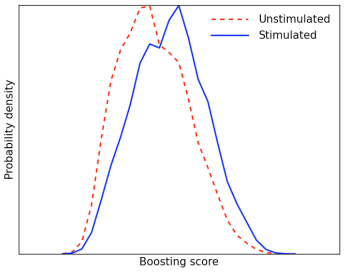
Trained Fisher's linear discriminant



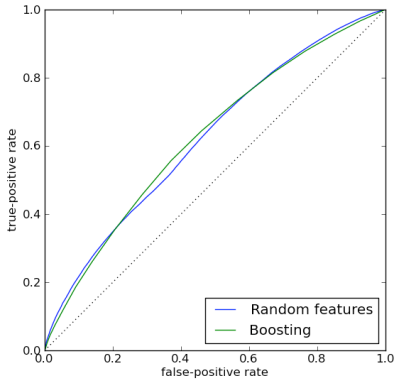
source: Cooley & Lohnes ((1971)

Automatic vs. hand training

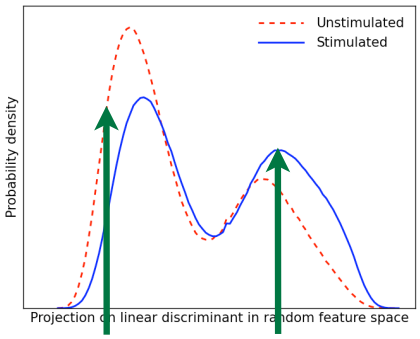




Automated classifier as good as human-trained



Soft labels for cells



Cell with this score: P(stimulated) = .85

Cell with this score: P(stimulated) = .65

Fuzzy counts

Each cell in a sample is assigned a probability of having the positive phenotype

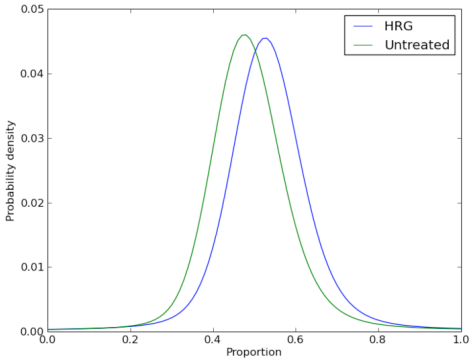
Compute the probability of the sample having

0 positive cells
1 positive cell
2 positive cells
…
6 positive cells

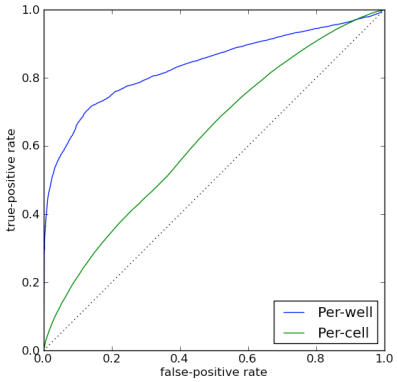
This pdf of counts can be turned into a pdf of proportions

Scoring samples by fuzzy counts

Mix the pdfs of proportions => empirical positive and negative control distributions

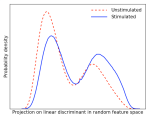


Per-well accuracy is quite good



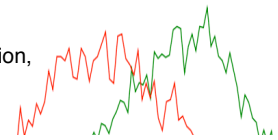
Summary

We can identify subtle, complex cellular phenotypes without human training



May enable screening for “invisible” phenotypes, as well as large-scale profiling experiments

Avoid premature thresholding, classification, and aggregation. Embrace populations, uncertain values, and fuzzy scores.



Acknowledgement: This work was supported in part by grant no. U54 HG005032 from the National Institutes of Health as part of the Broad Institute Comprehensive Screening Center of the MLPCN.