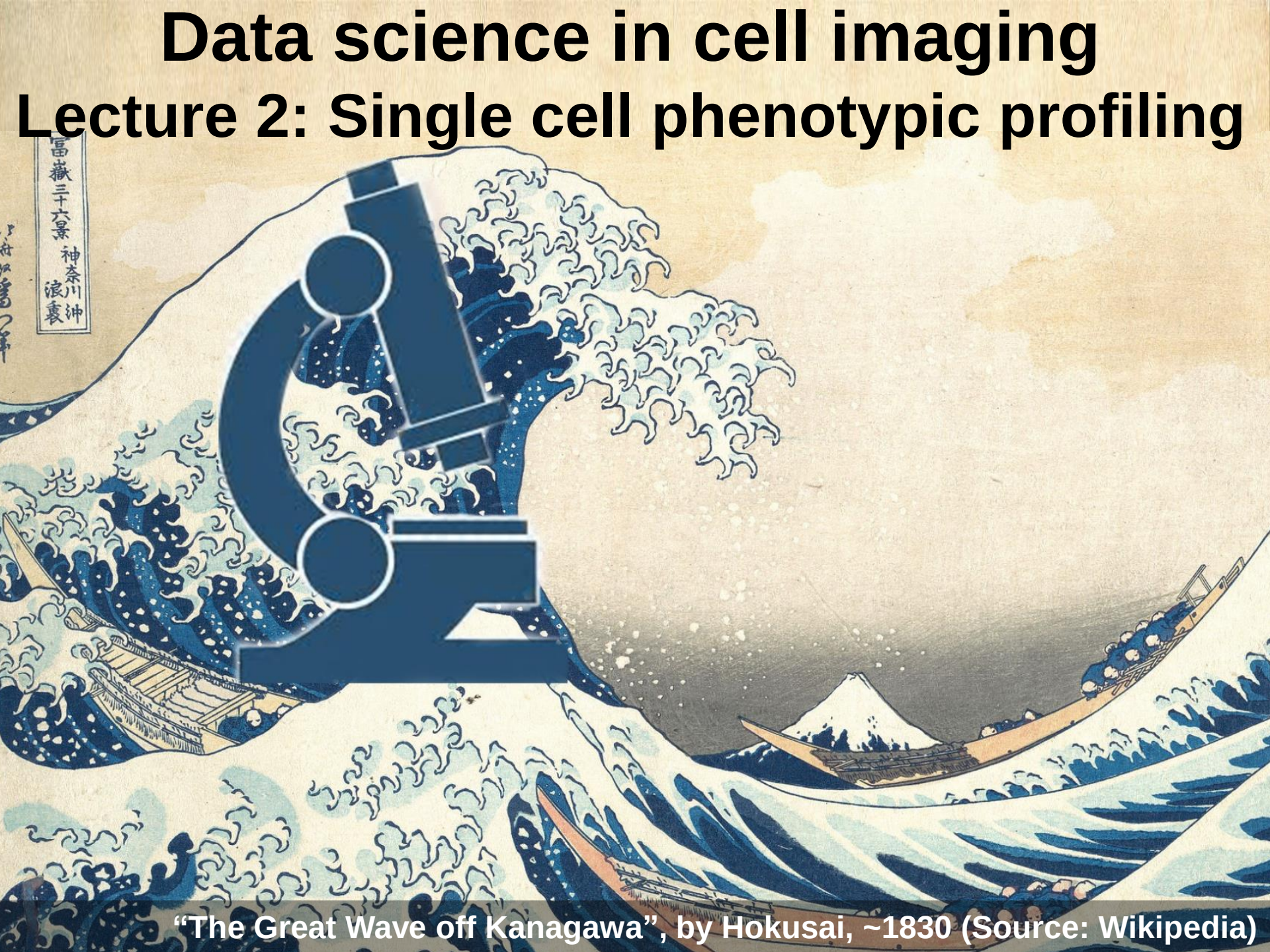
Data science in cell imaging
Lecture 2: Single cell phenotypic profiling



"The Great Wave off Kanagawa", by Hokusai, ~1830 (Source: Wikipedia)

All slides are open under the cc-by license

You are free to share and adapt any content from this presentation provided that you attribute the work to its author and respect the rights and licenses associated with its components

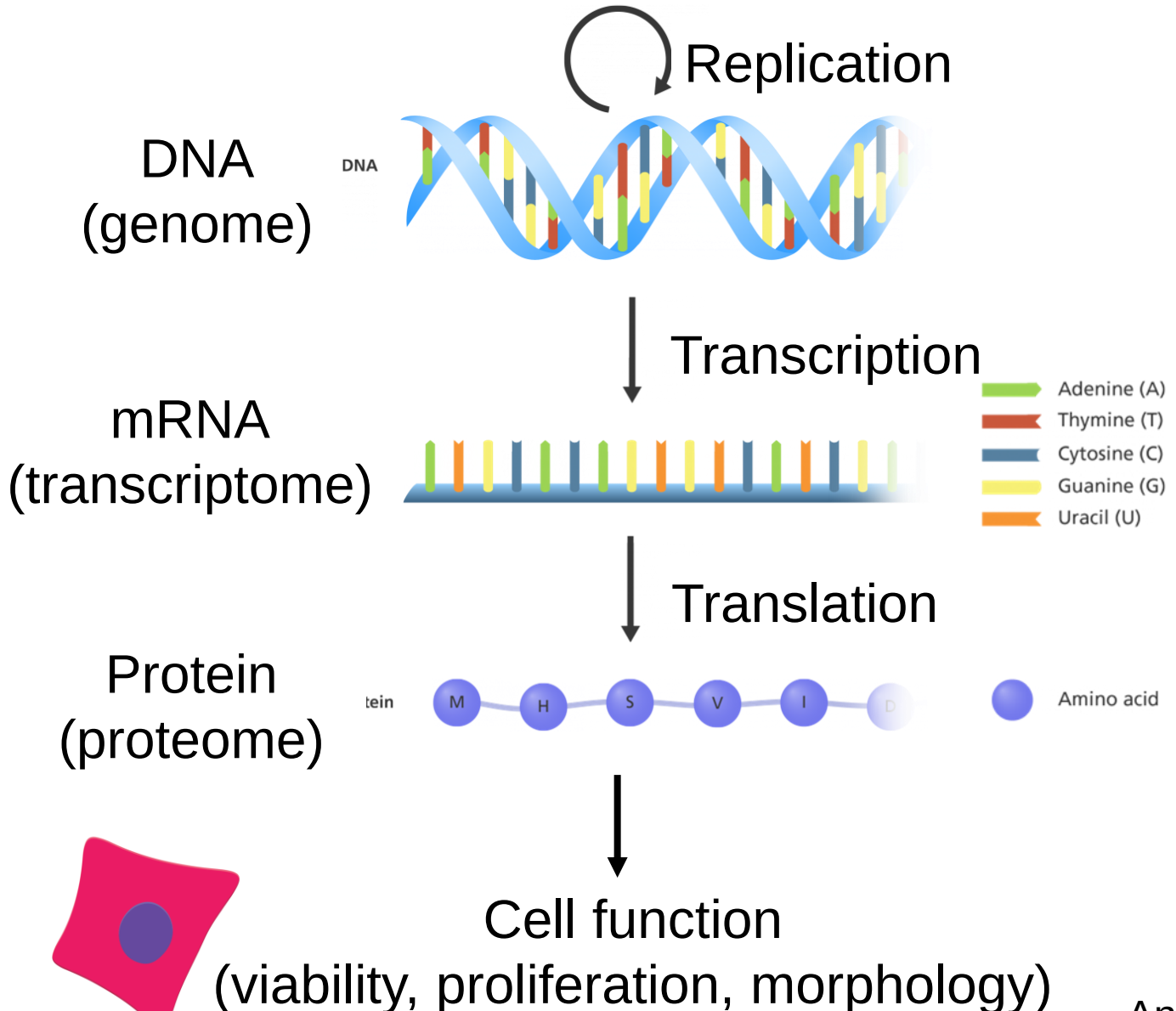
PPTX slides available here

Thank you and caveat toward today's lecture

Most of today's slides were adapted from Anne Carpenter (Broad Institute)!

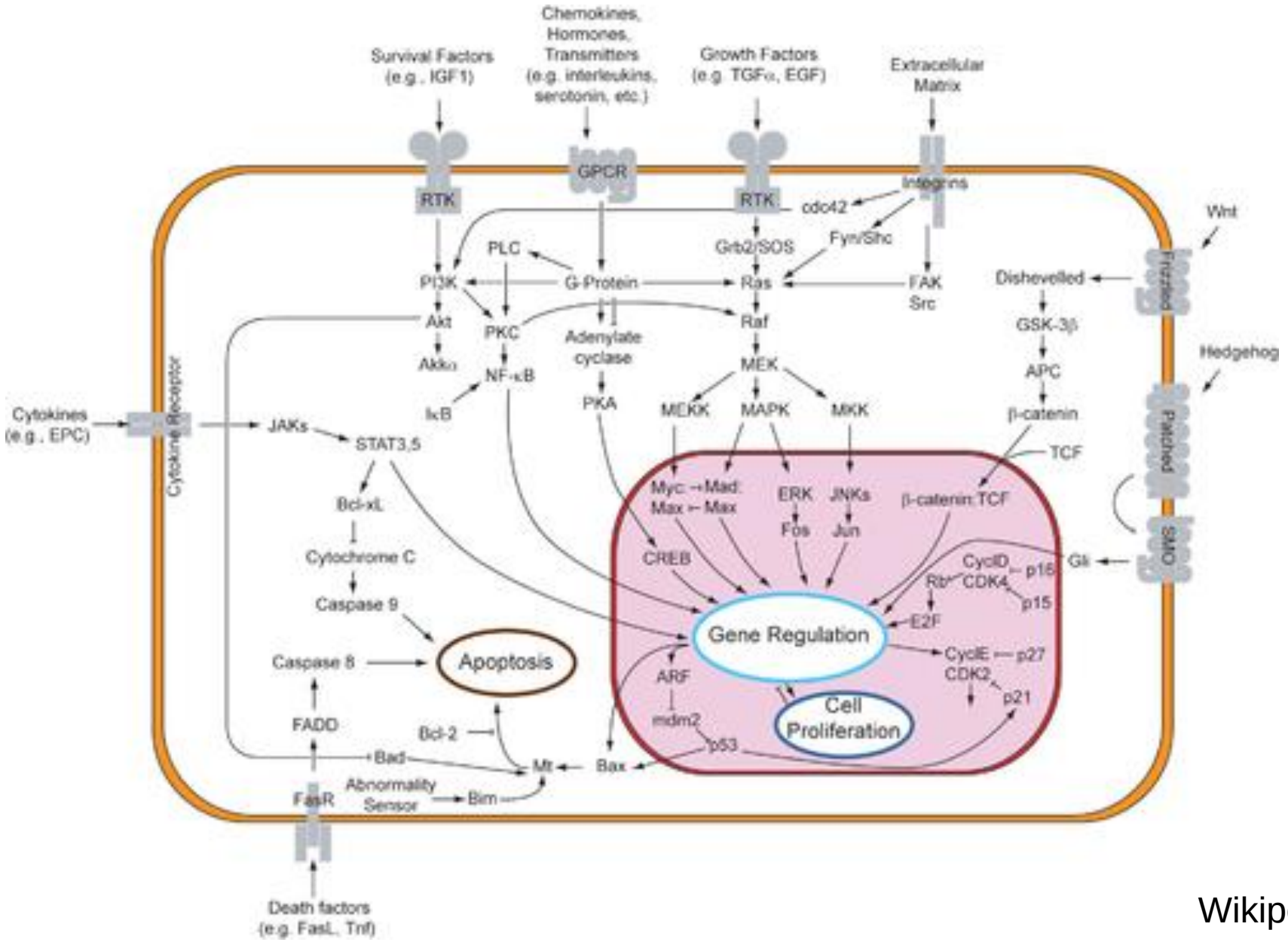
Heavily biased toward her lab's work
(to be improved next year…)

The central dogma of molecular biology



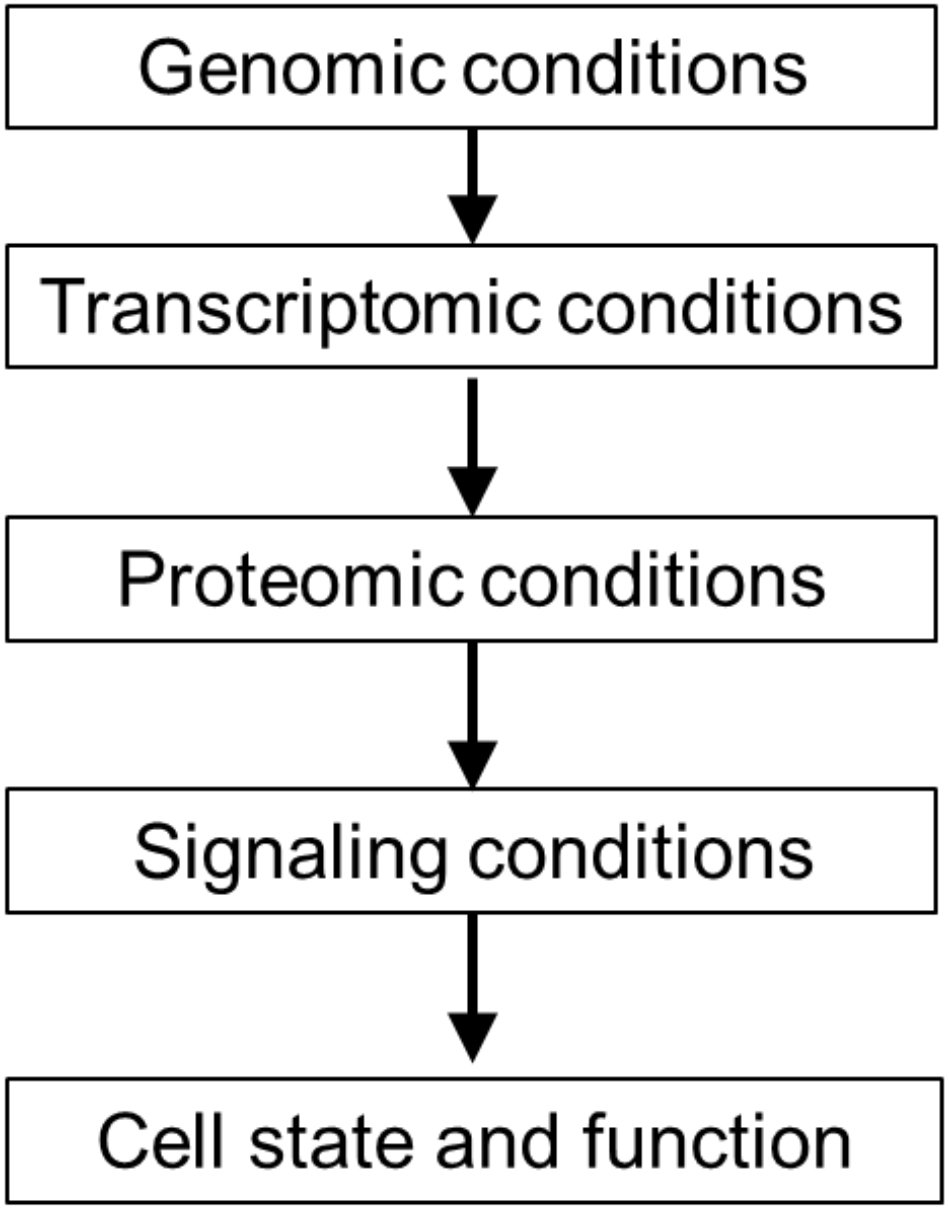
Anne Carpenter

Cell signaling

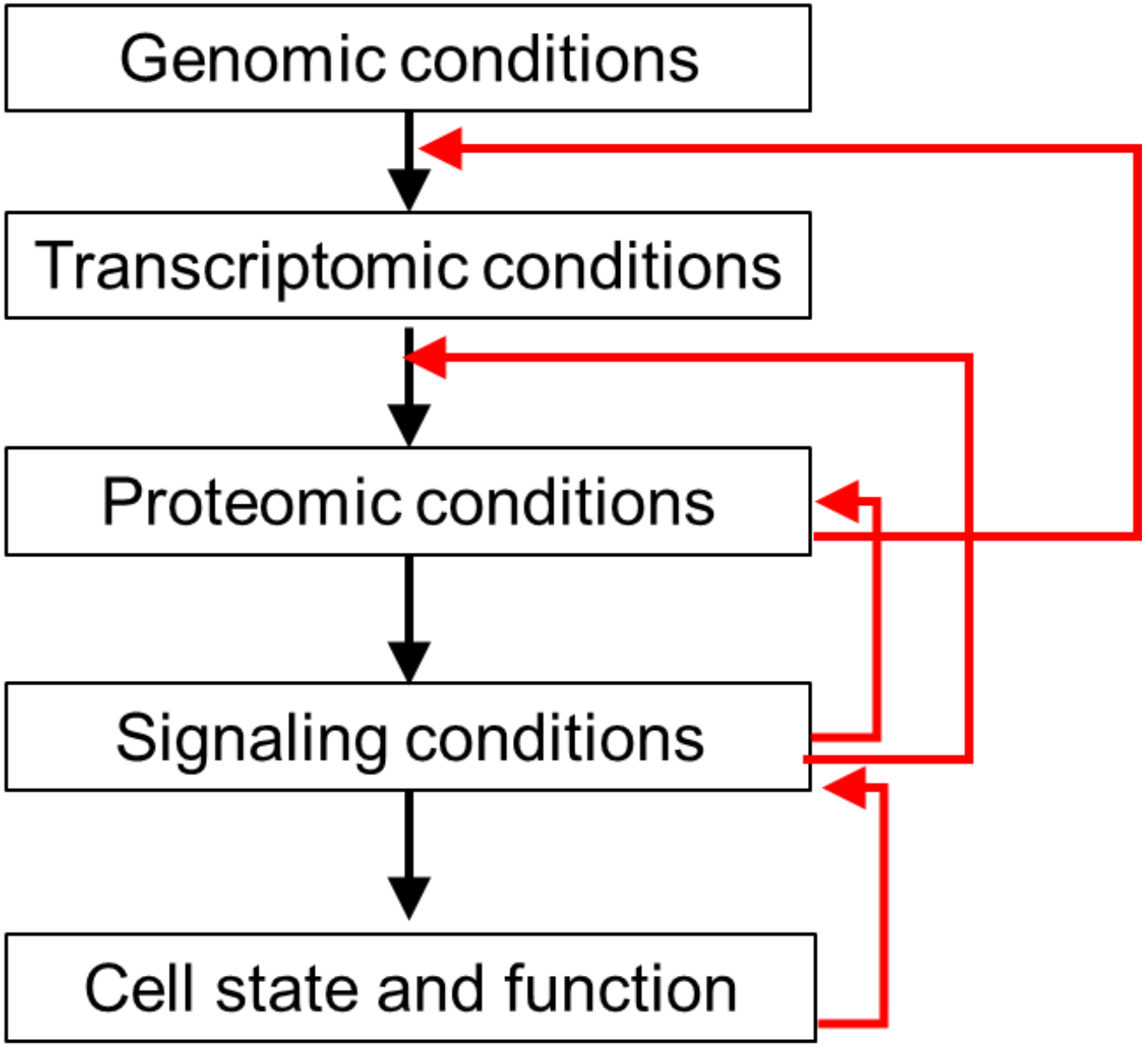


Wikipedia

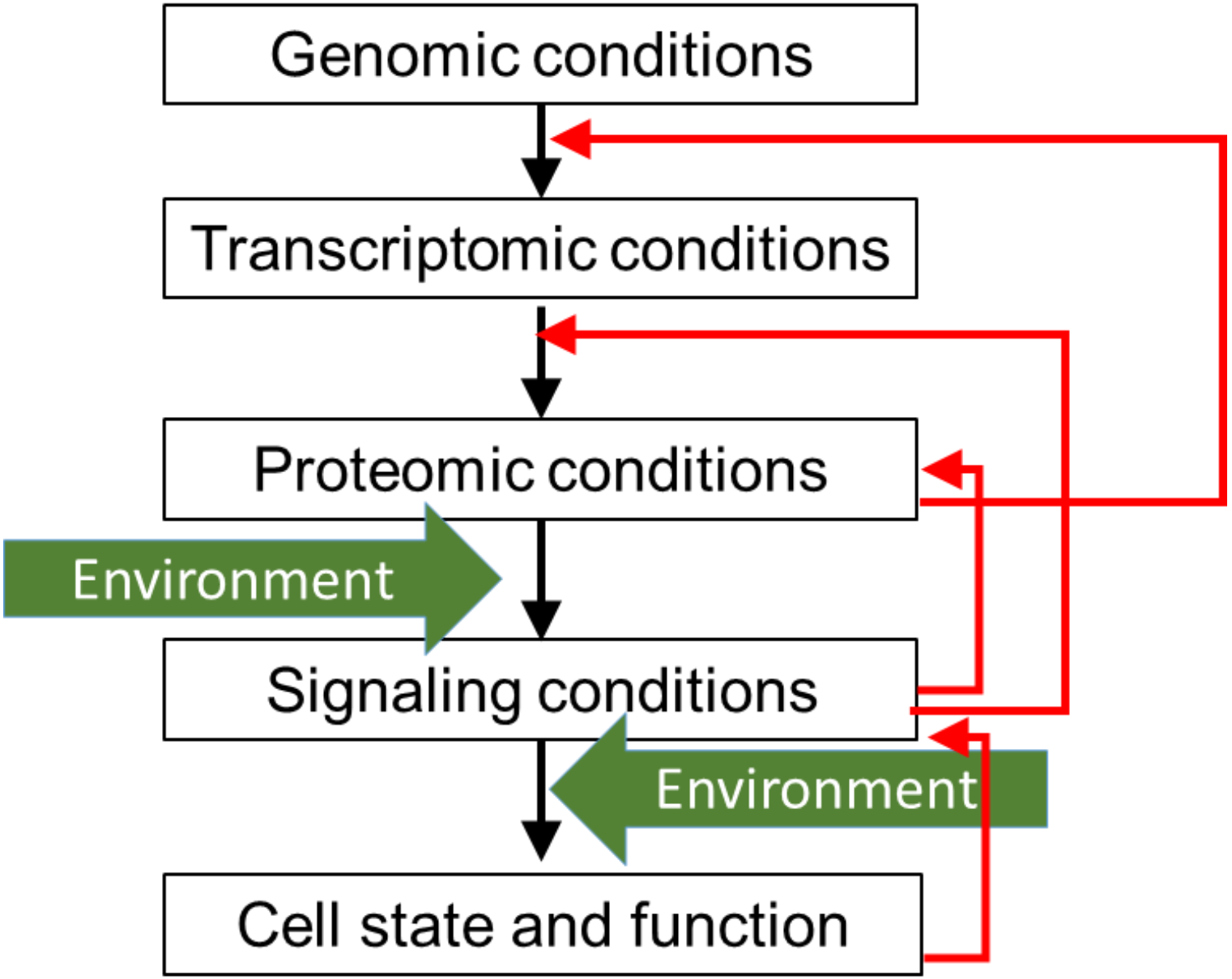
The central dogma of biology



But in reality…



But in reality…



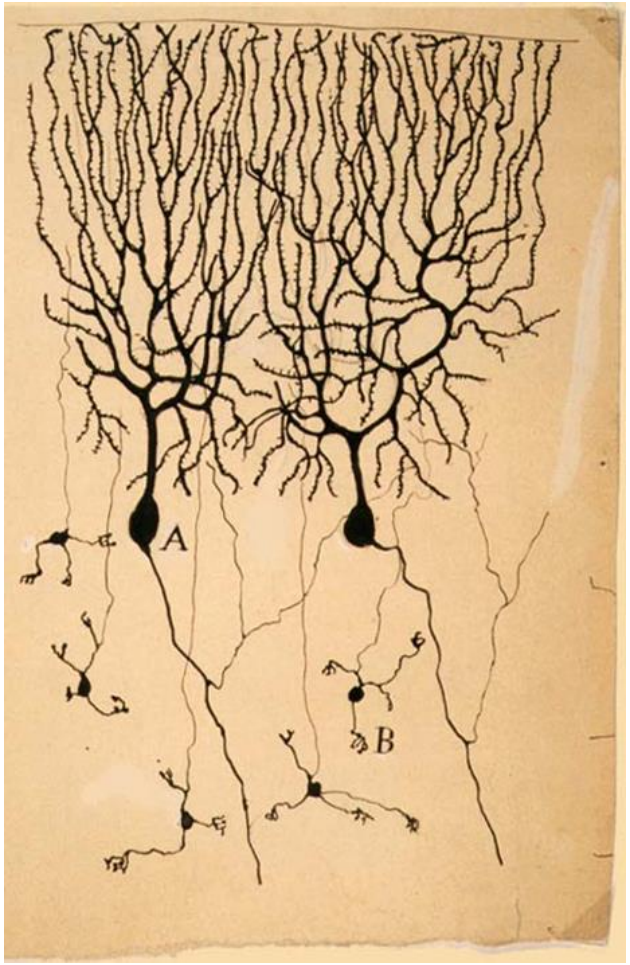
We need functional readouts!

Why microscopy?

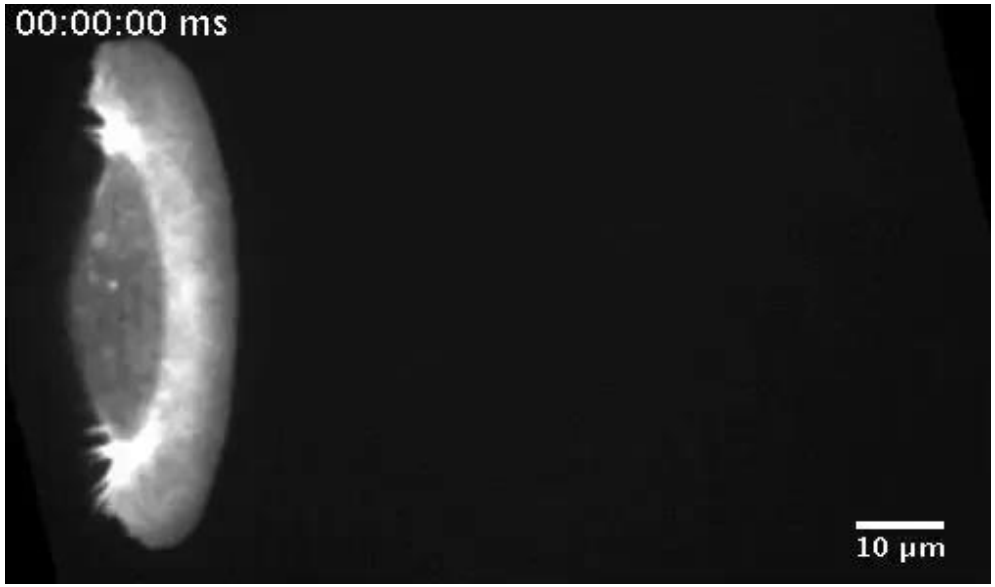
The only technology that allows us to see live cell behavior and to correlate cell function to intracellular (protein quantity and location) and extracellular (environment) factors

Morphology is a marker of the cell’s functional state

Diverse cell morphologies enable diverse functions

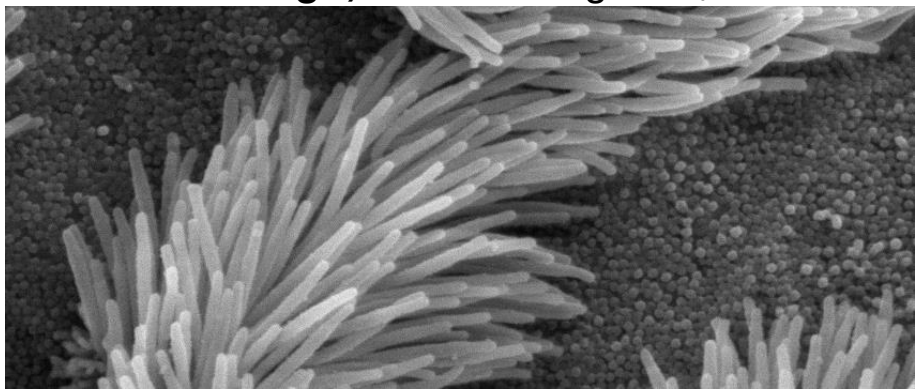


Purkinje cells
Santiago Ramón y Cajal



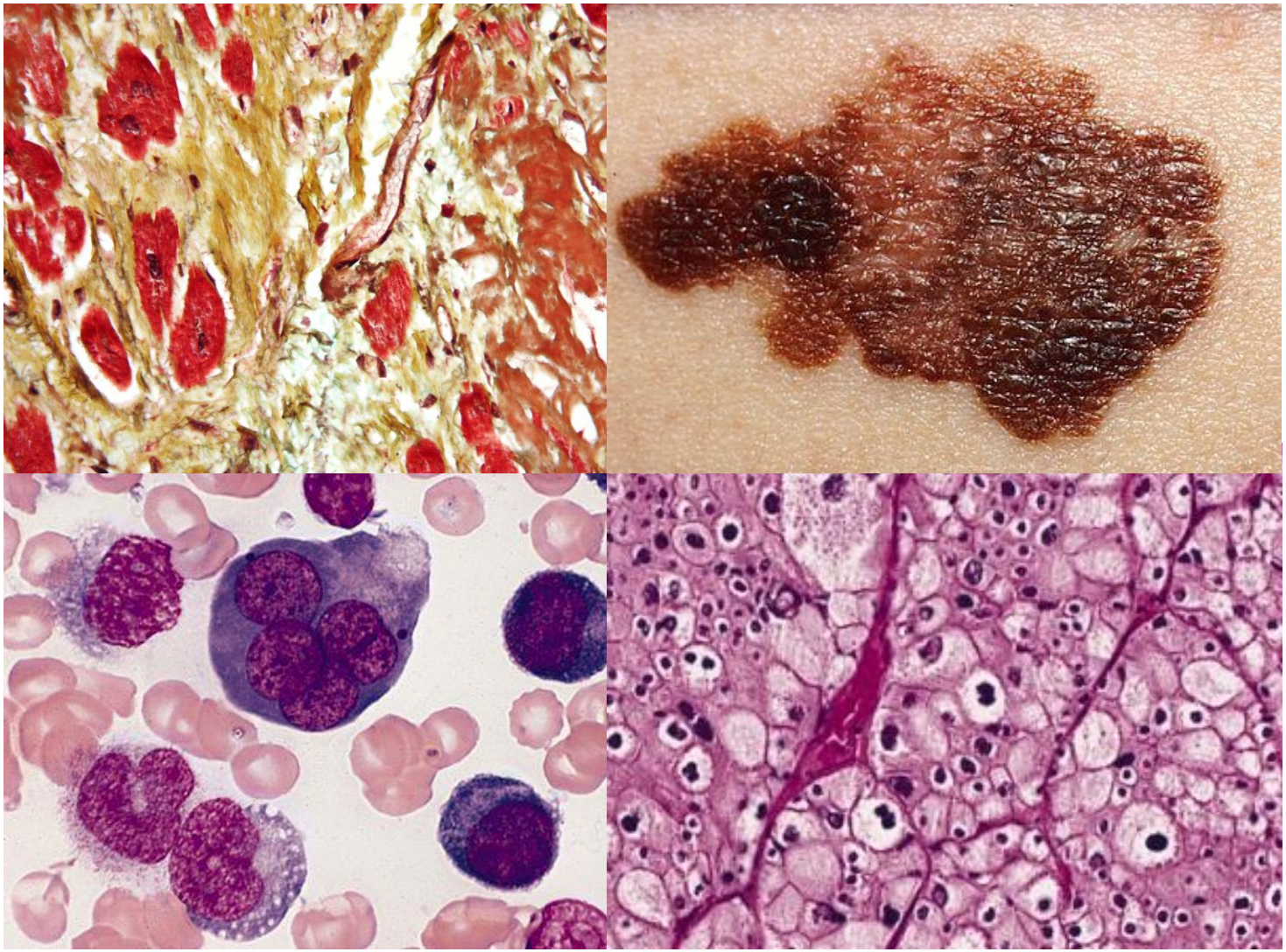
Zebrafish keratocyte; Mueller *et al* (2017)

Cilia in the lungs; Charles Daghlian, Dartmouth



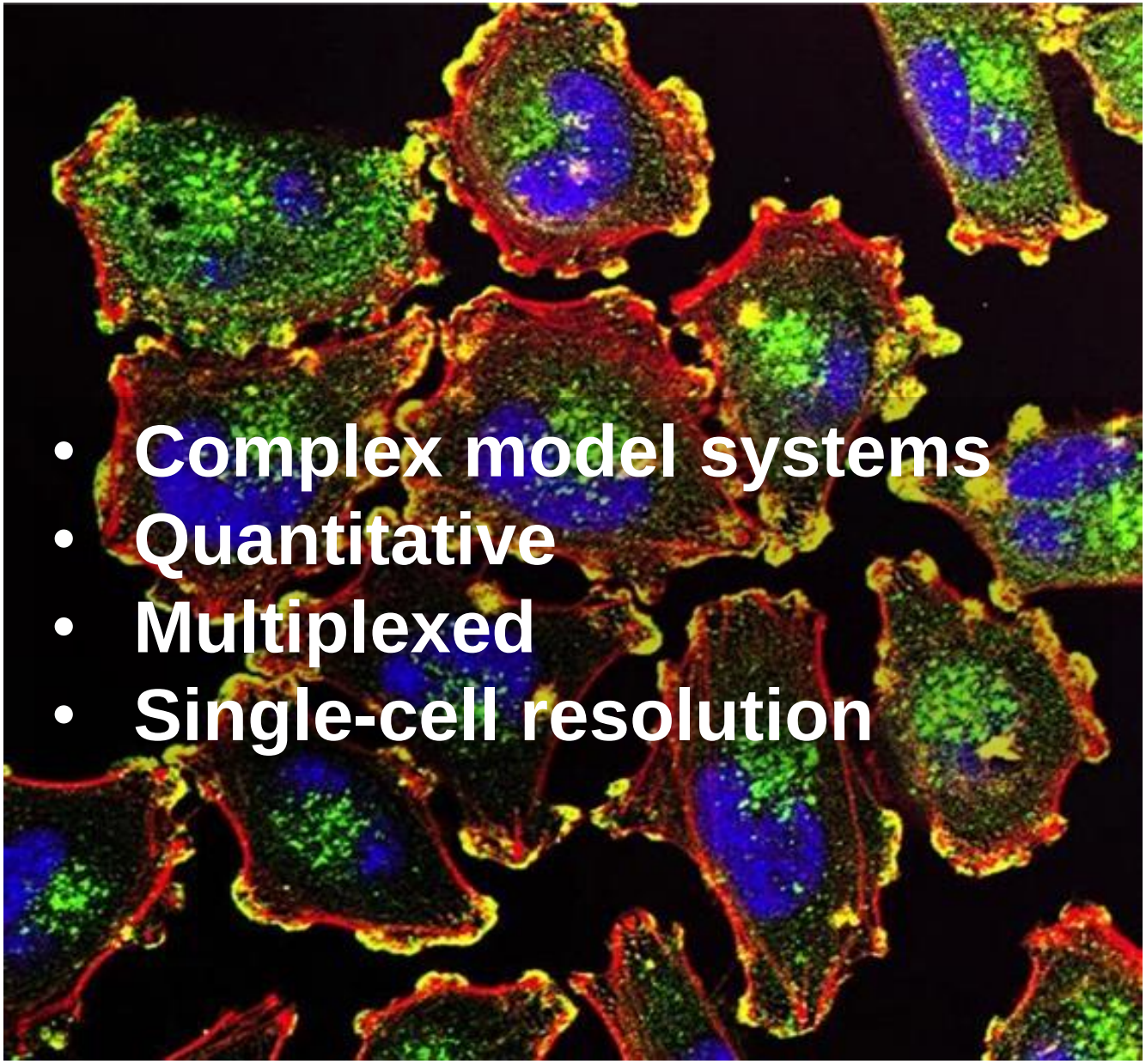
Meghan Driscoll

Visual appearance indicates cell (and disease) state



Anne Carpenter, images from Wikipedia

Images contain a wealth of information

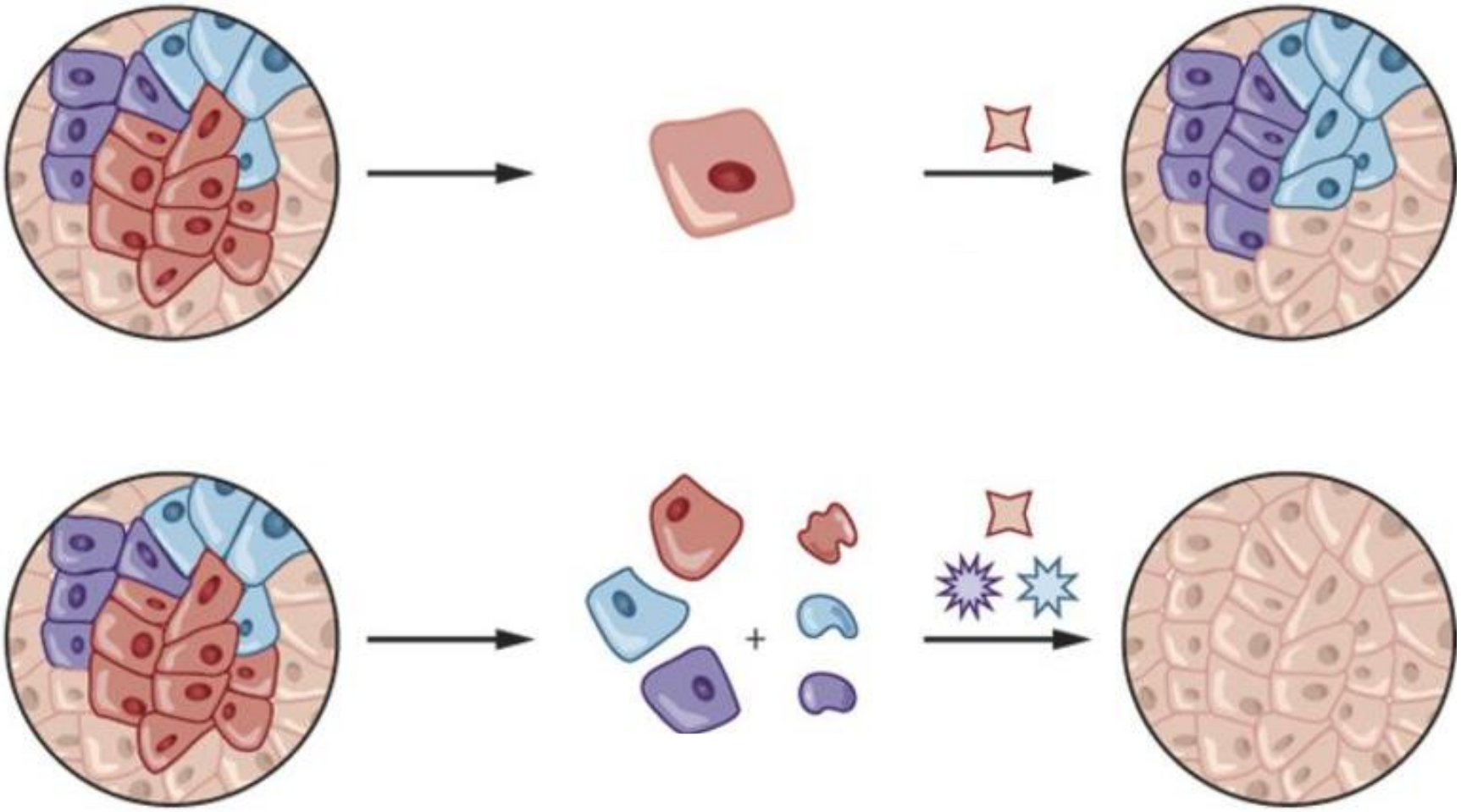


Anne Carpenter, from Wikipedia

Why do we care about single cell resolution?

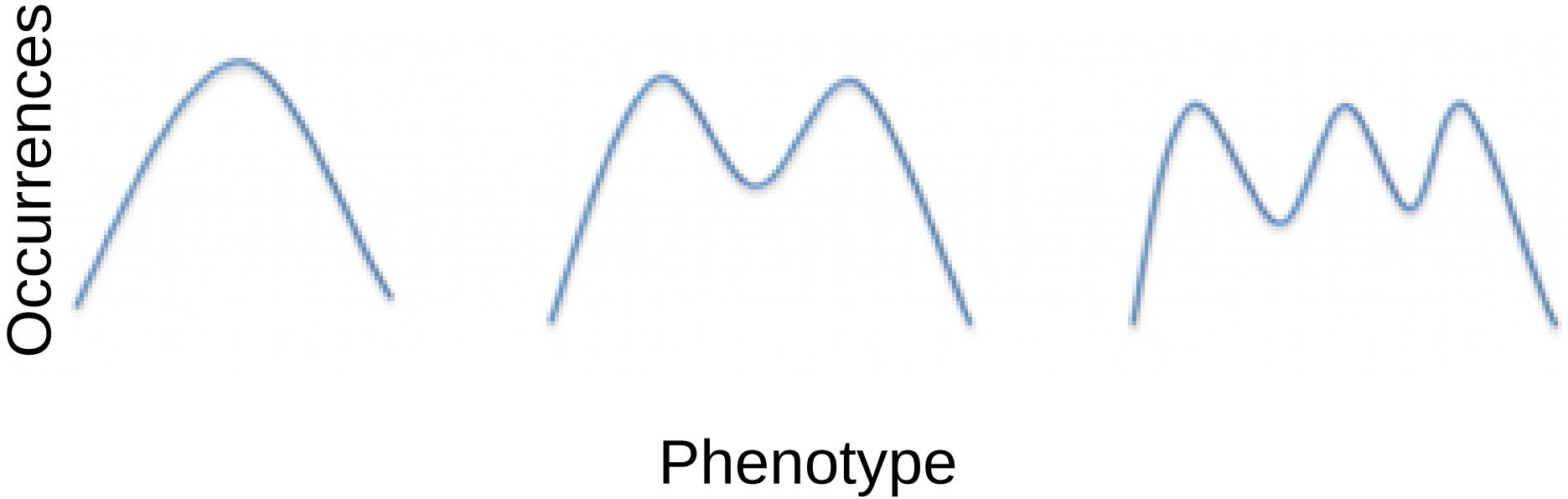
Source: https://pt.slideshare.net/QIAGENscience/why-scientist-analyze-single-cells

Genetic heterogeneity



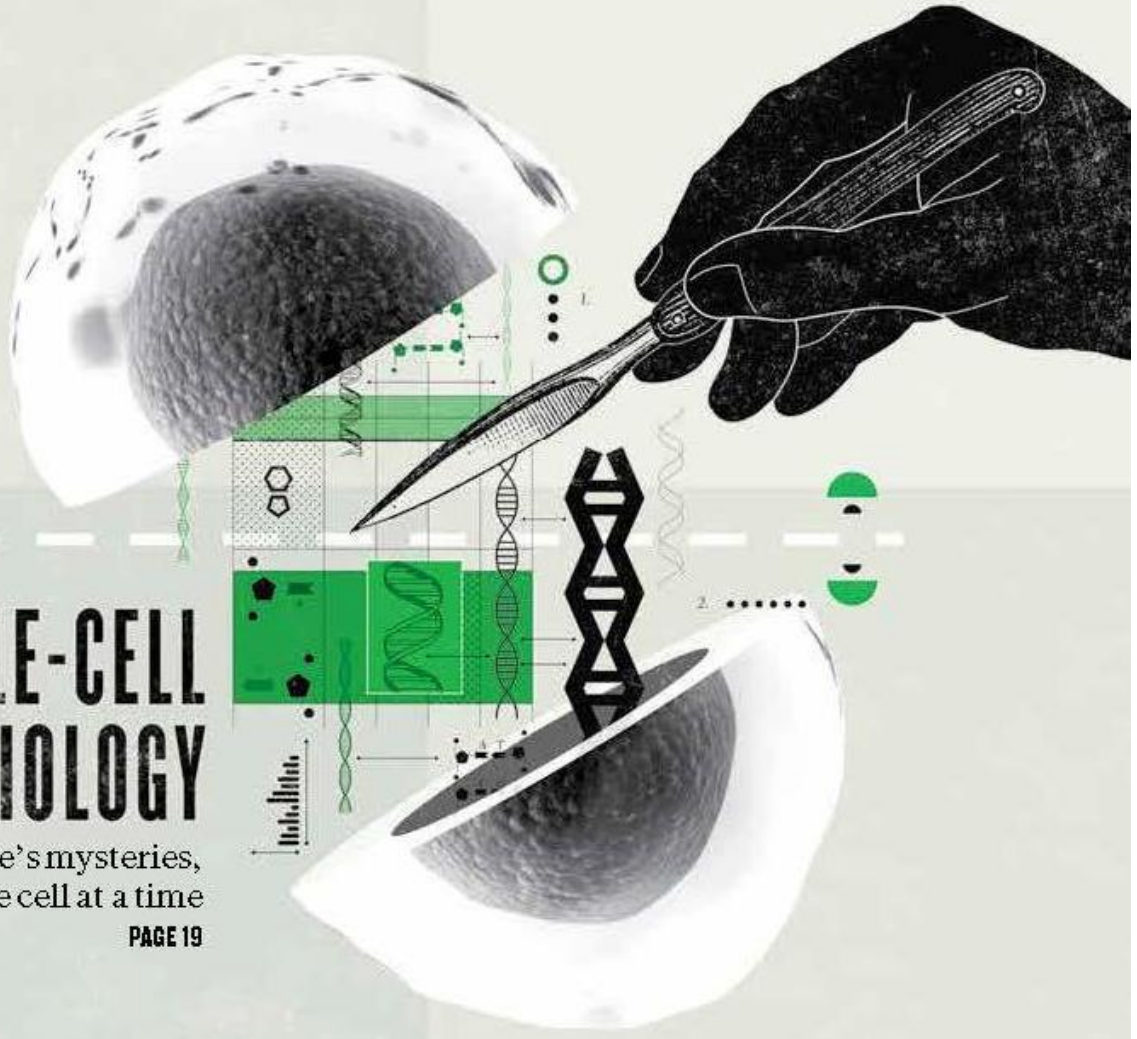
Shalek and Benson (2017)

Phenotypic heterogeneity



nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

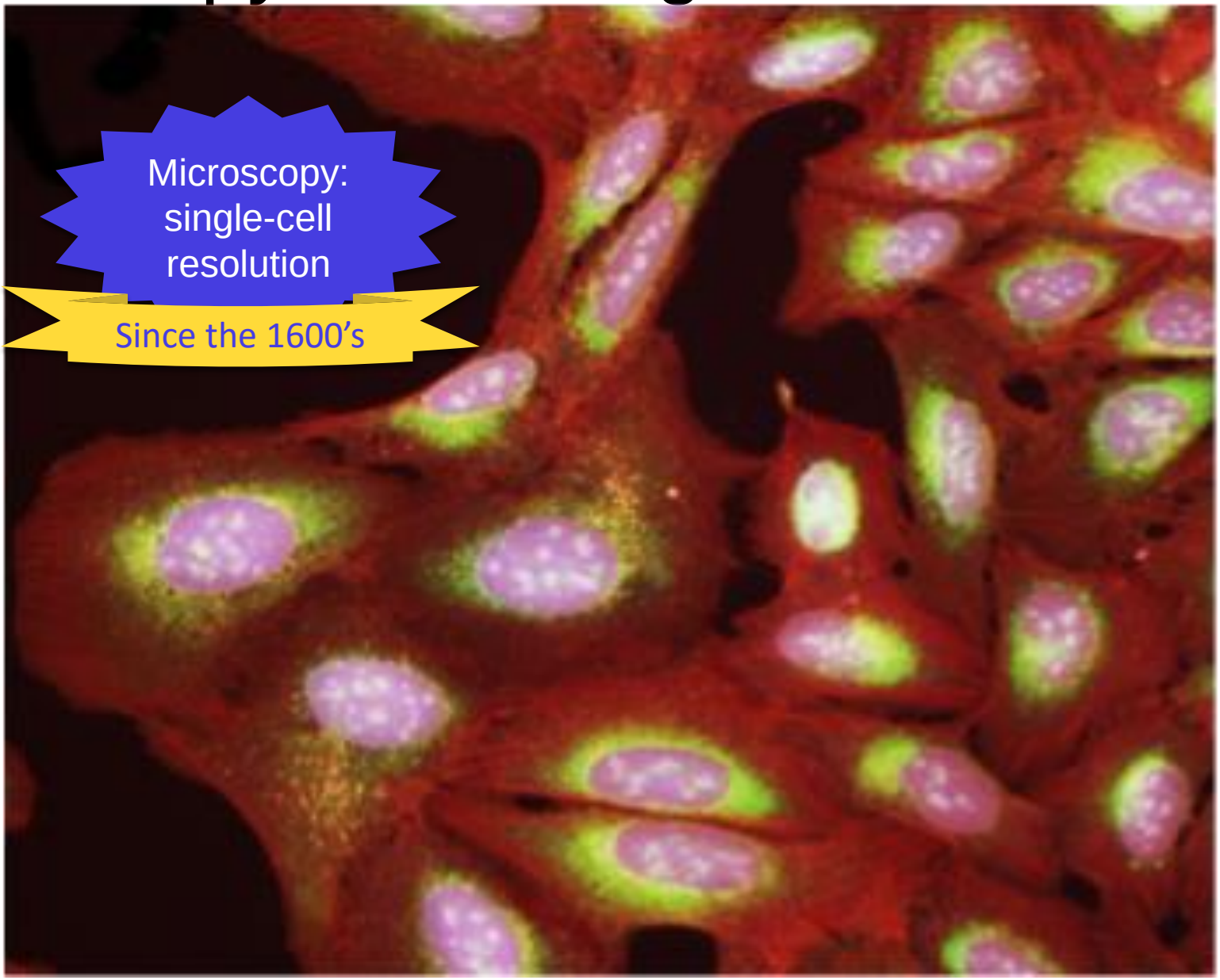


SINGLE-CELL BIOLOGY

Dissecting life's mysteries, one cell at a time

PAGE 19

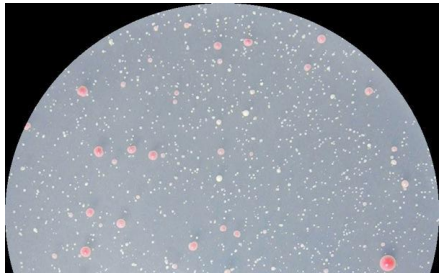
Microscopy offers single cell resolution



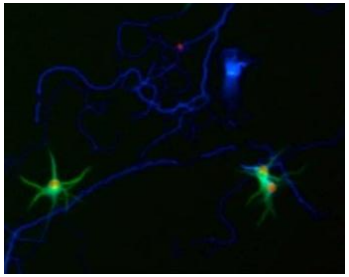
Anne Carpenter

Complex cell models can be quantified by imaging

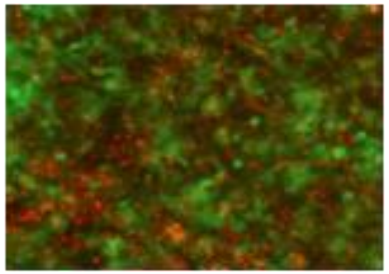
Yeast



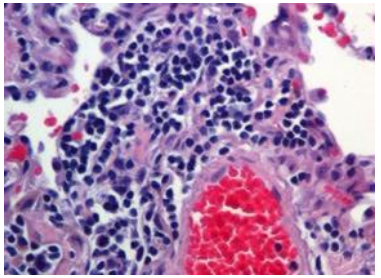
Neurons



Co-cultures

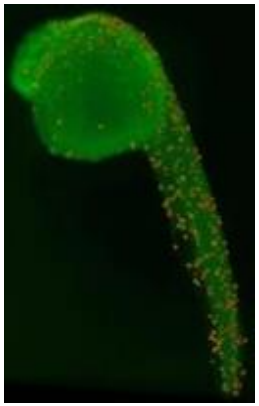


Tissue

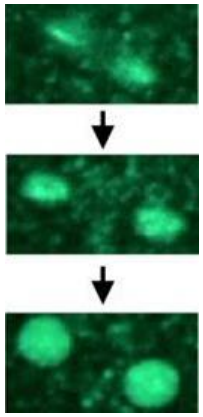


C. elegans

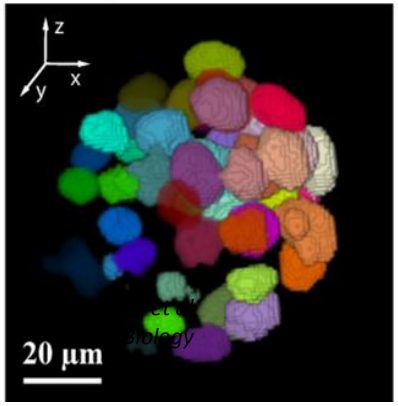
Zebrafish



Time-lapse



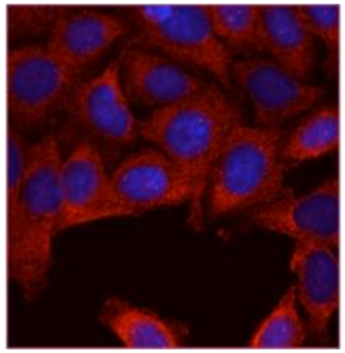
3D



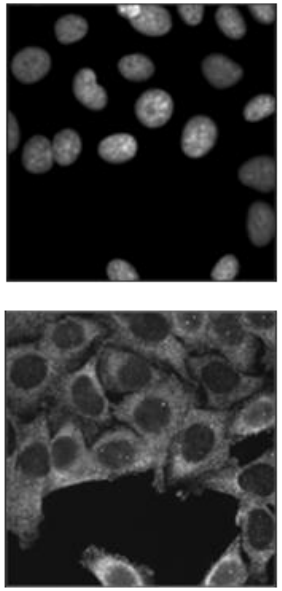
Anne Carpenter

Software (e.g., CellProfiler) can quantify cells

Original image

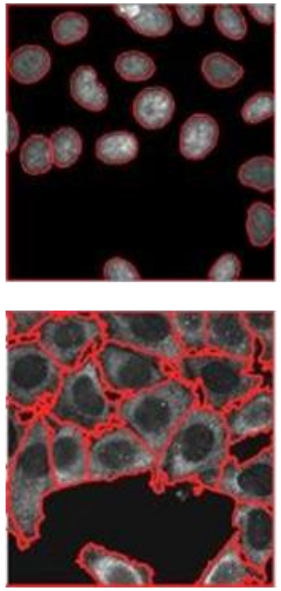


Split colors



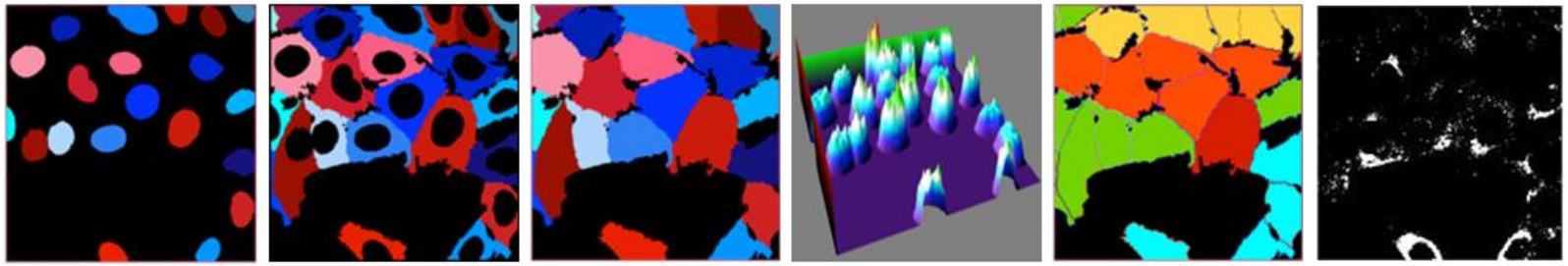
Correct illumination

Identify cells/ compartments



Measure everything

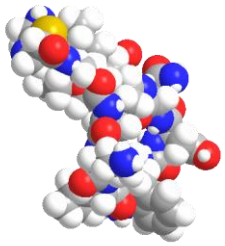
Counts, Shapes, Sizes, Intensities, Textures, Correlations, Relationships



Anne Carpenter

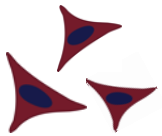
Cell based biomedical research in a nutshell (over simplified)

chemical or gene



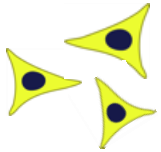
+

cells



+

any change? (phenotype)



=

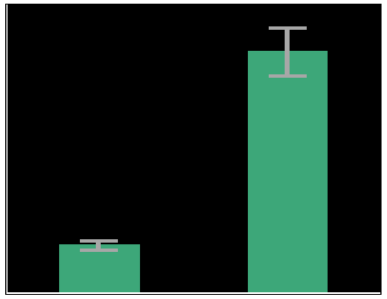
insight

Anne Carpenter

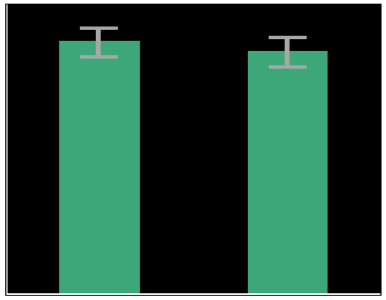
Visual appearance indicate cell state and can be quantified

GFP labeled protein which undergoes a translocation from the cytoplasm to the nucleus in response to perturbations

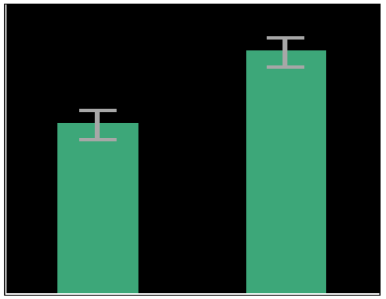
localization



protein levels



morphology



… + hundreds of other features

Anne Carpenter

Automated image analysis is

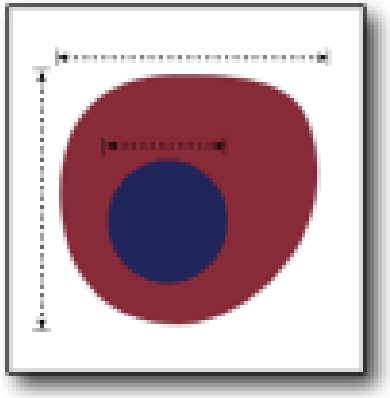
- Objective
- Quantitative, with statistics
- Measure multiple properties at once
- Distinguishes subtle changes, even those undetectable by eye
- Faster, less tedious

Over automation? Advise for scientists who are not tool builders

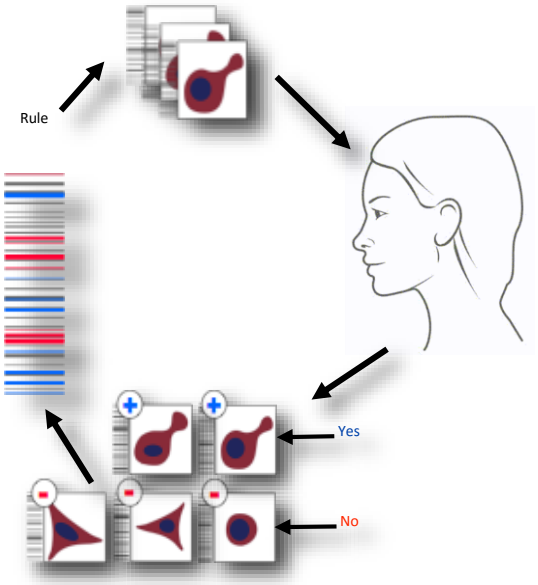
- Your goal is answering a biological question?
- Fully-automated analysis is not always worth the time/effort. Sometimes manual is better!
  - Quantity of data?
  - Accuracy needed?
- Semi-automation can save development time, and improve accuracy
- Application specificity: use specific application knowledge (at the cost of reduced generalization)
- We are not all tool builders!

Three waves of quantitative image analysis

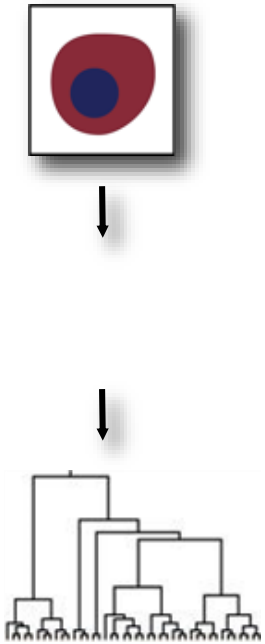
Measure known phenotypes



Train for known phenotypes



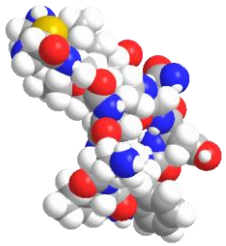
Discover new phenotypes



Anne Carpenter

Drug discovery (we wish!)

chemical



+

diseased human

=

healthy human

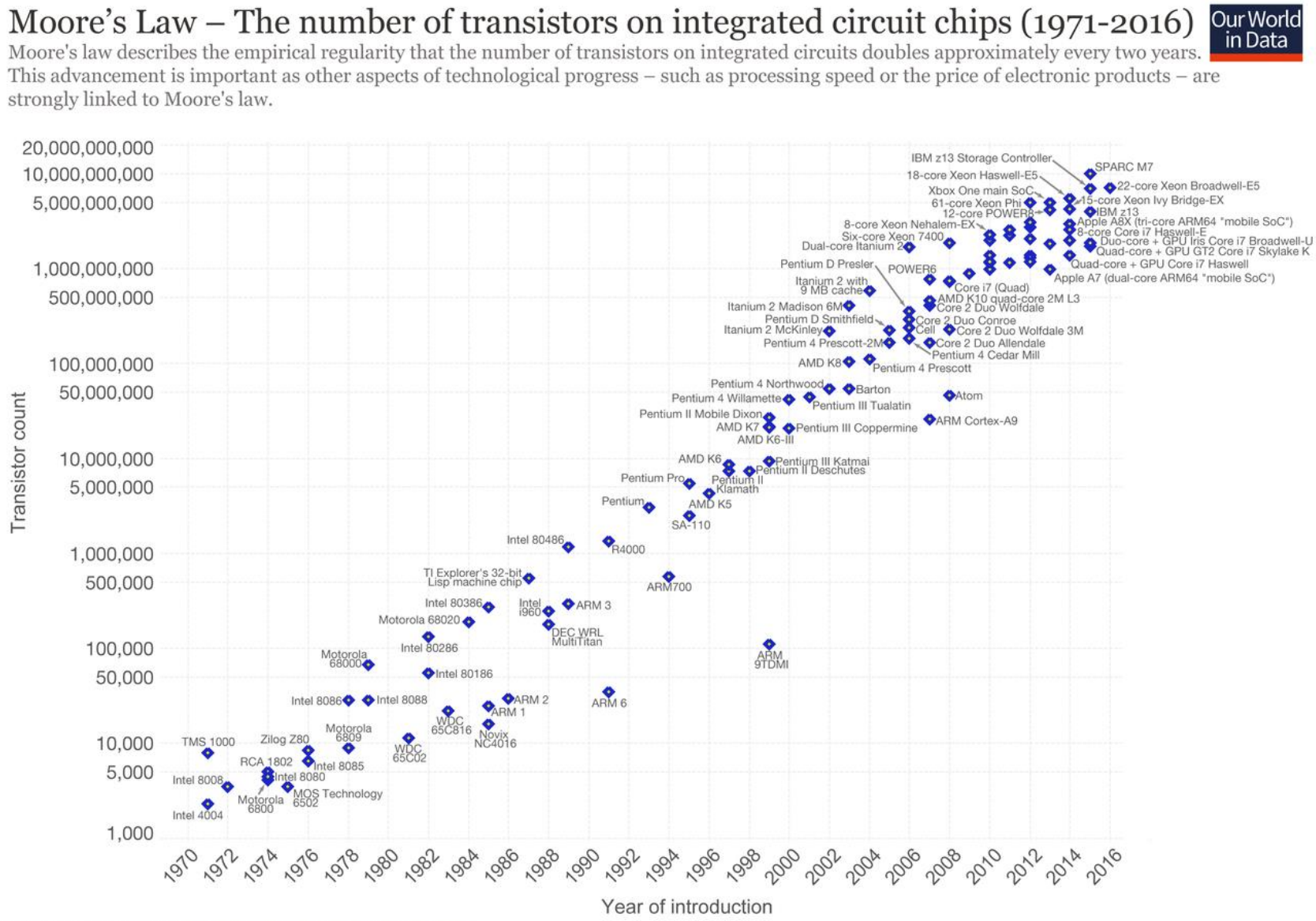
Anne Carpenter

Tradeoffs in cell biology (& drug discovery)

Physiological relevance vs. experimental complexity

- Physiological relevance:
  - In vivo
  - Resolution (spatial and temporal)
  - 3D
  - In context of microenvironment
- Experimental complexity:
  - Technology
  - Complexity of experiment
  - Costs
  - Amount of data collected
  - Complexity of data analysis

The Moore's law

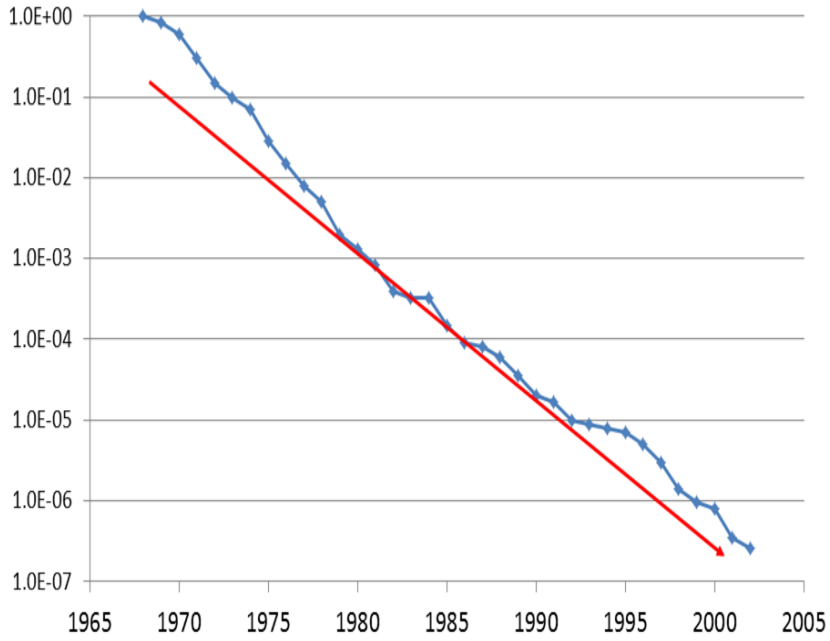


Anne Carpenter

Tale of two industries

Moore's Law

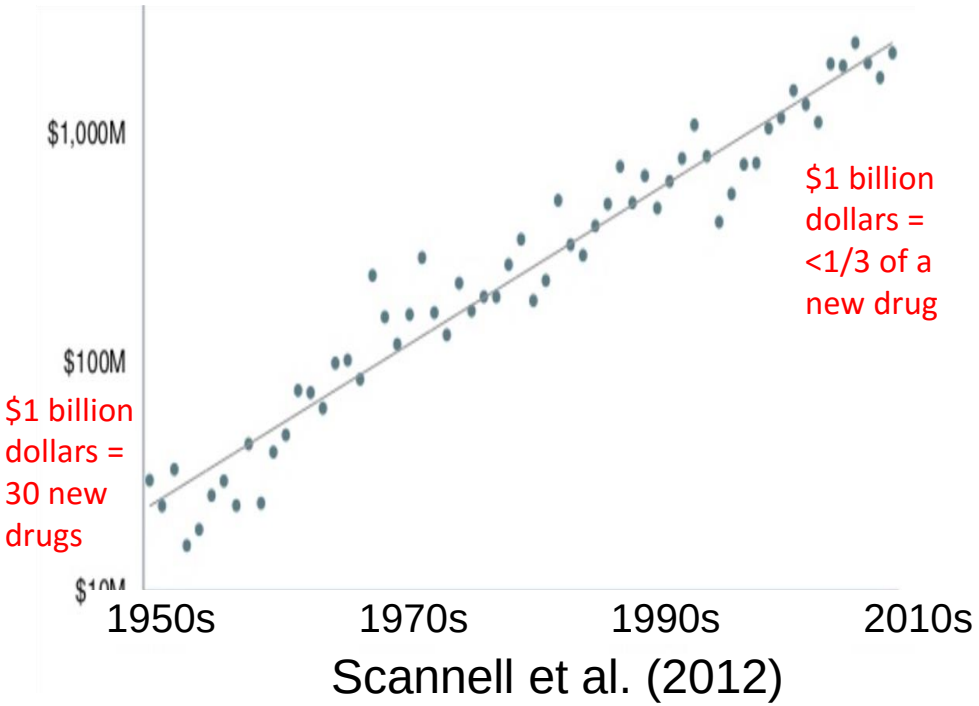
Compute gets cheaper and cheaper



SweptLaser.com

Eroom's Law

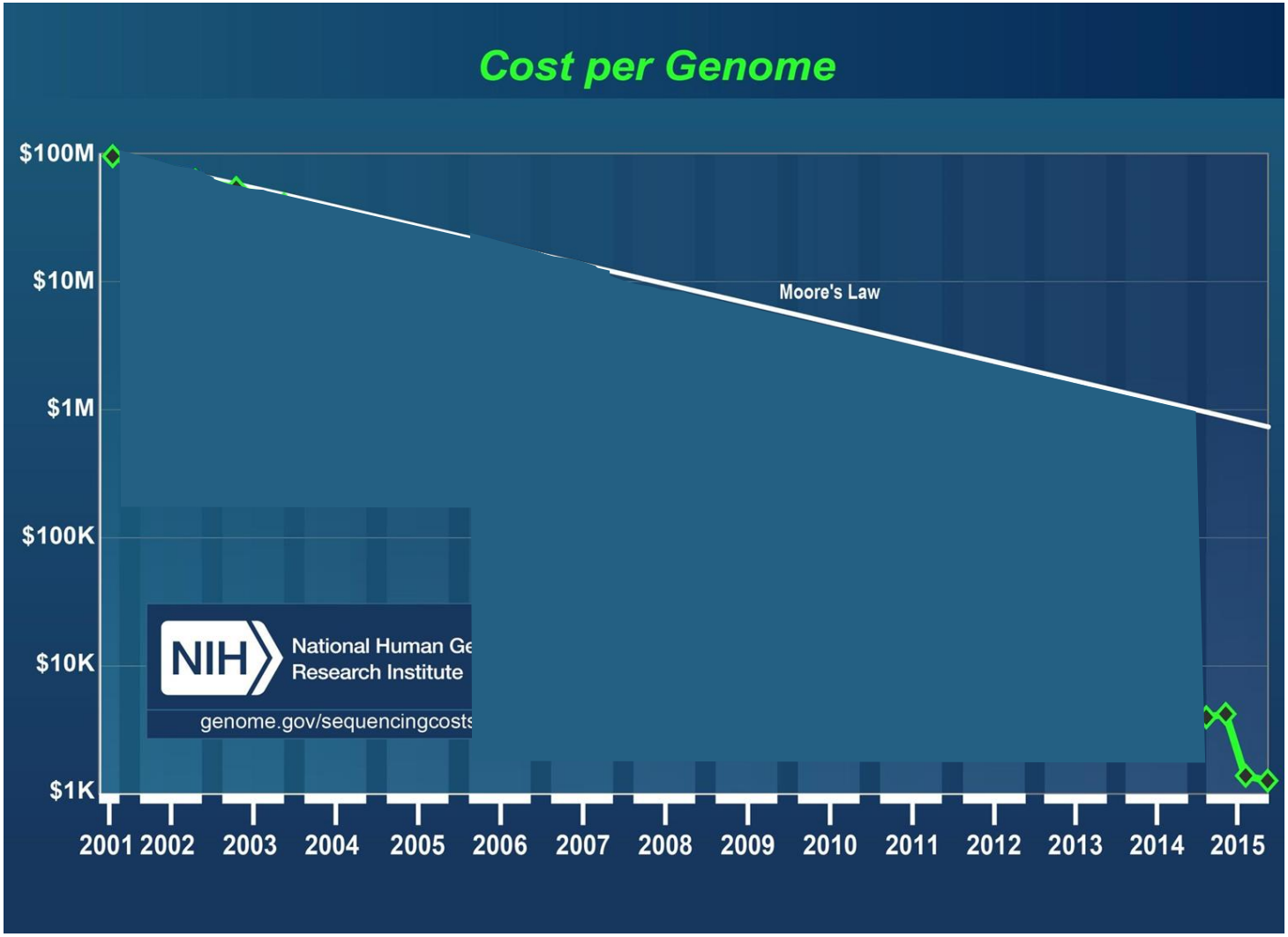
Discovering new medicines gets more and more expensive



Scannell et al. (2012)

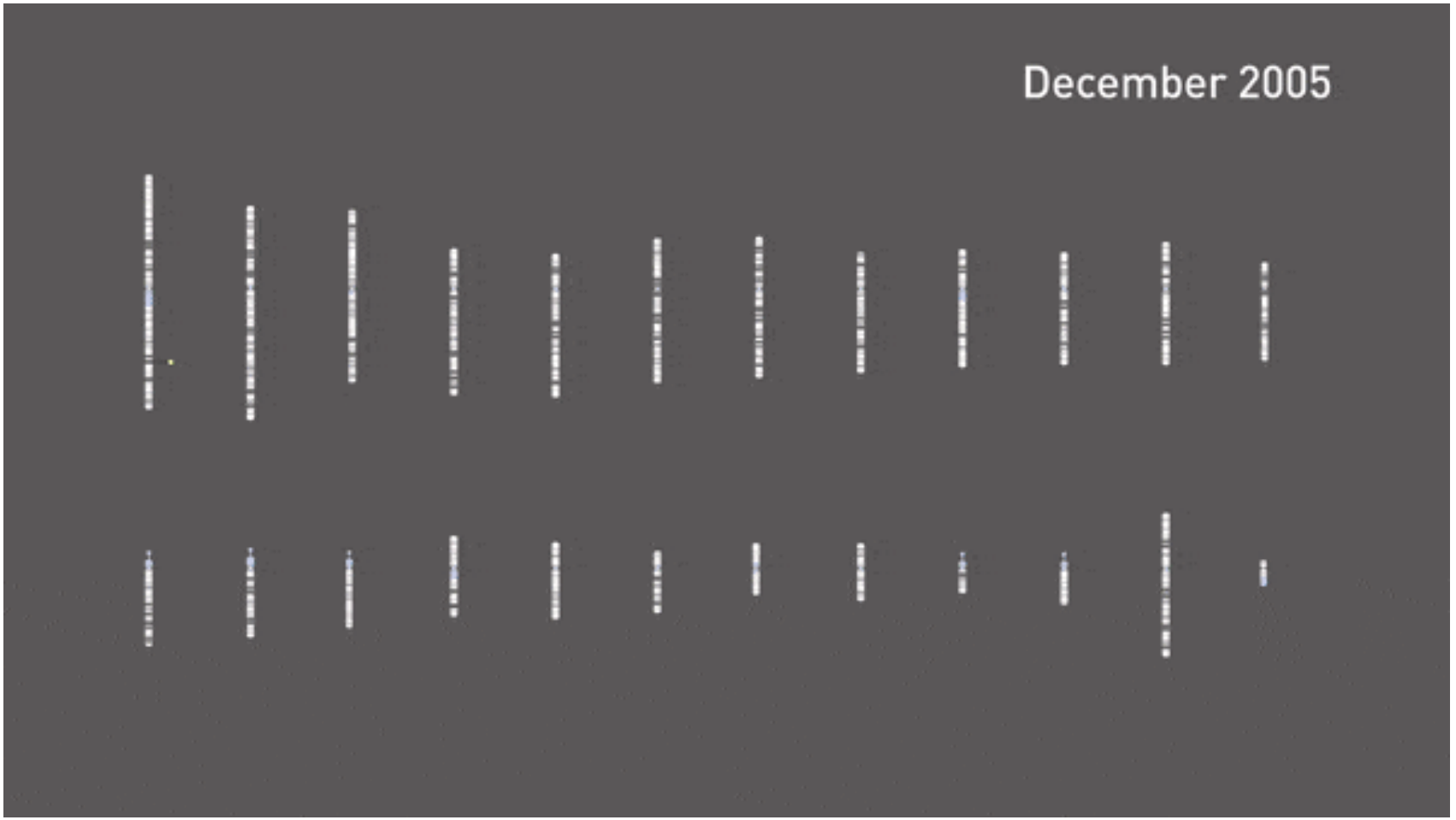
Anne Carpenter

Sequencing costs have dramatically declined



Anne Carpenter

Sequencing is uncovering the cause of many diseases

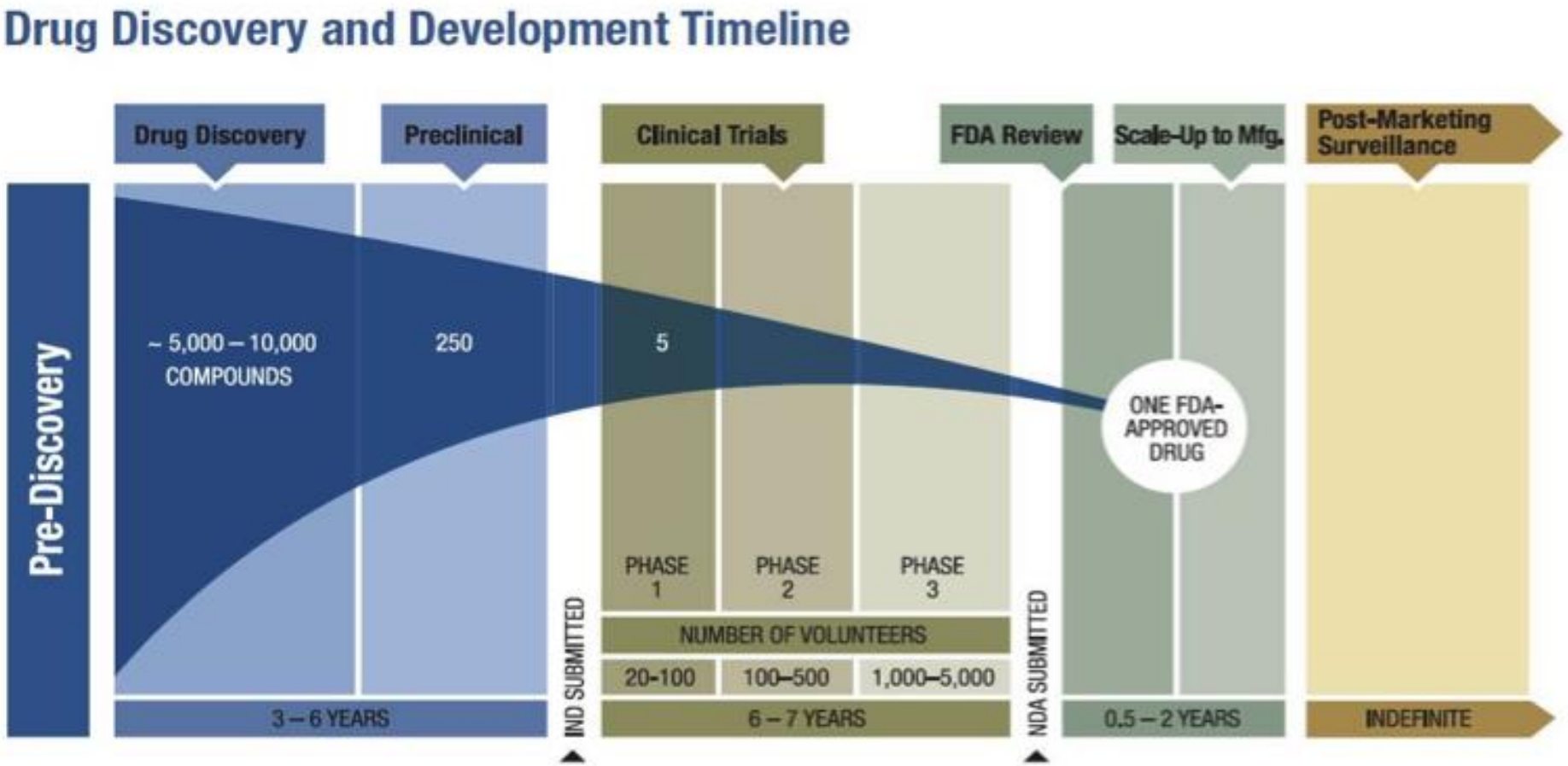


Genome regions associated with traits or disease through genome-wide association studies

MacArthu et al. (2017)

Anne Carpenter

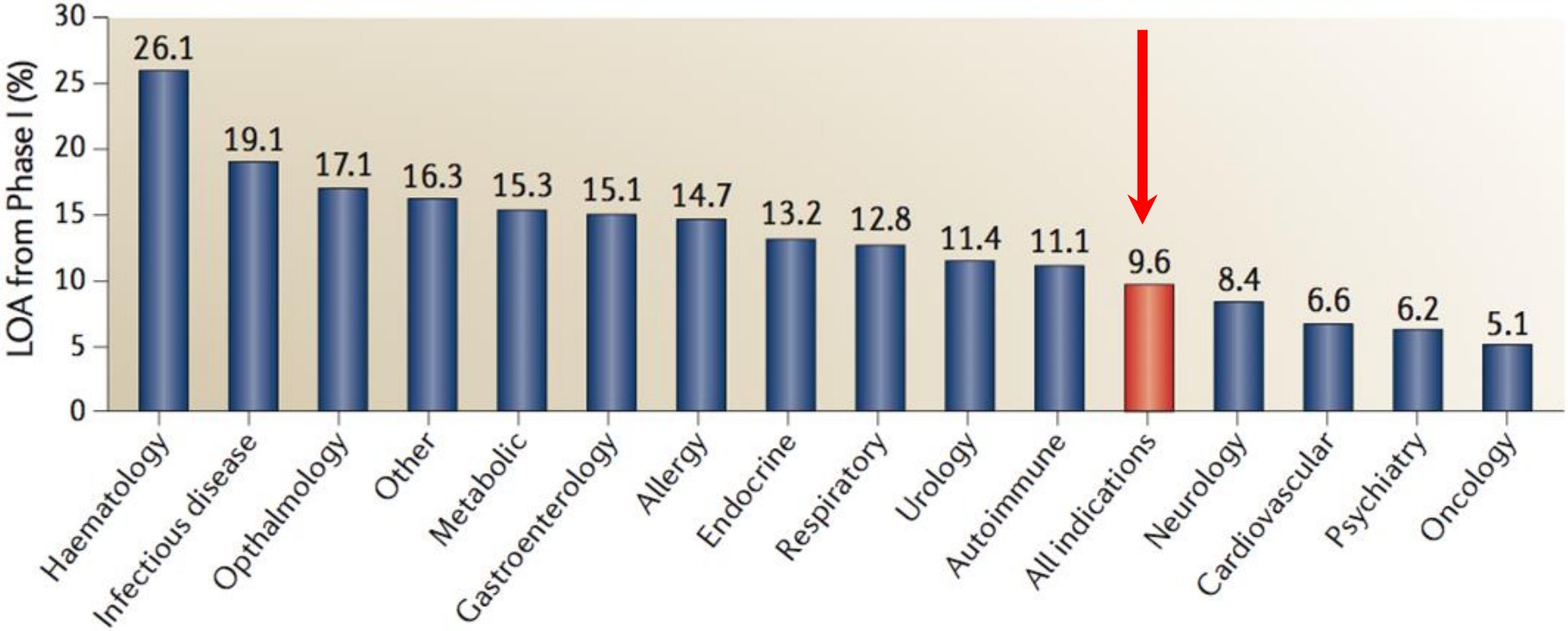
Drug discovery is time consuming



Anne Carpenter, from PhRMA/Applied Clinical Trials

Drug discovery fails most of the time

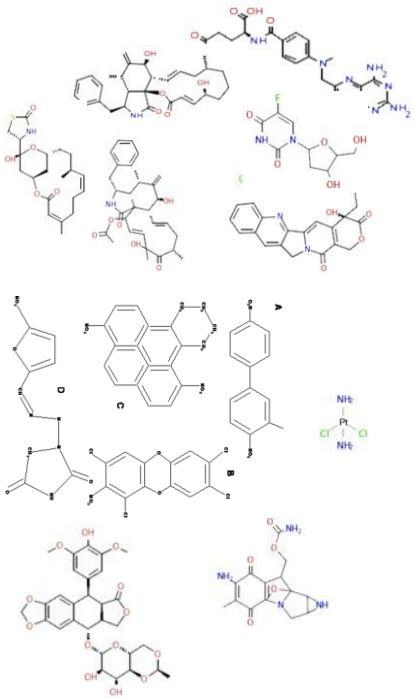
Success rate of a drug entering Phase I clinical trials
Average = 9.6% make it to market



Anne Carpenter, from Nat Rev Drug Discov July 2016

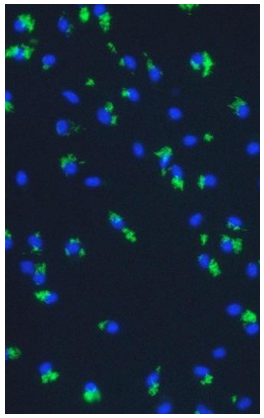
Discovering drugs in high throughput?

millions of chemicals



+

cells



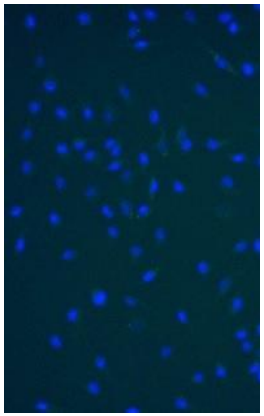
cells

+

tuberculosis bacteria

→

readout (phenotype)



cells are alive!

+

tuberculosis is gone!

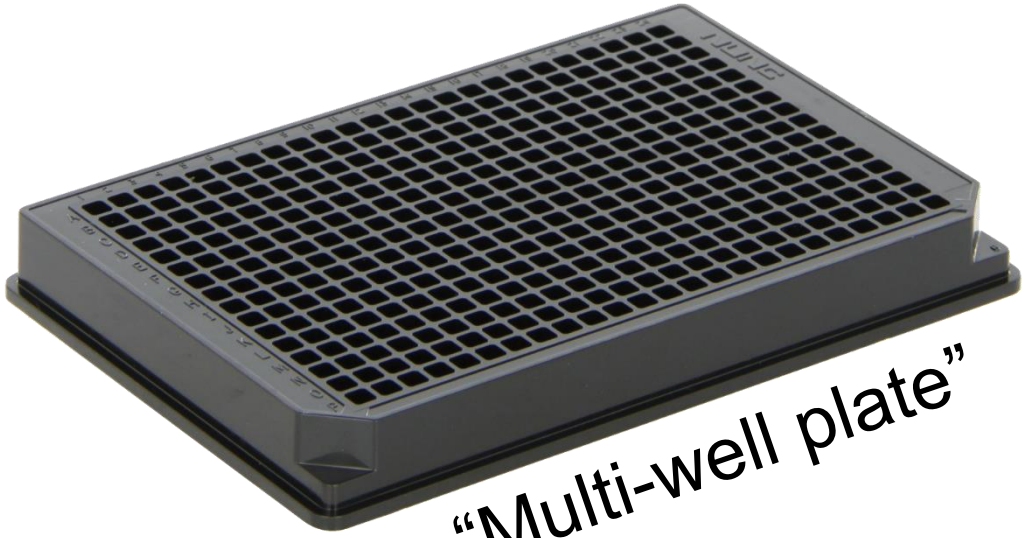
=

drug!

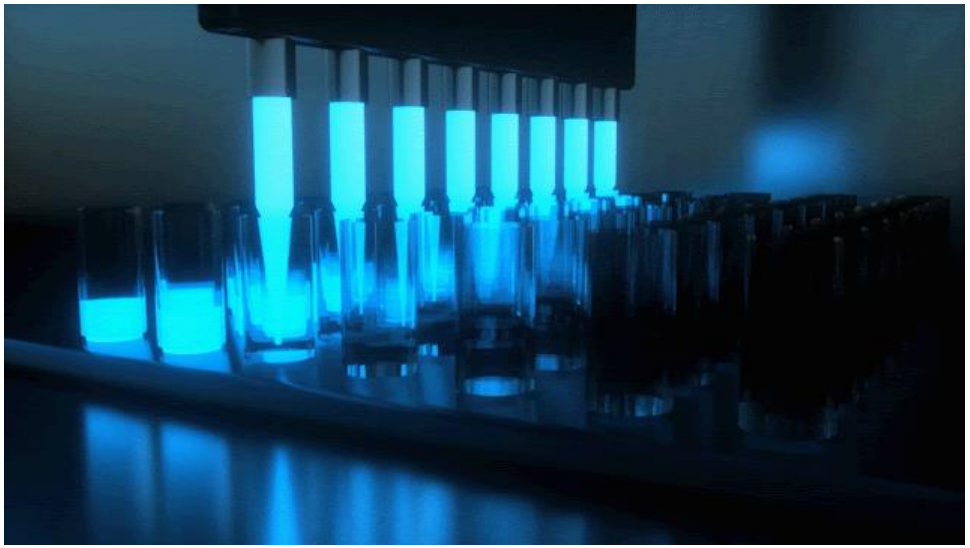
Anne Carpenter

Discovering drugs in high throughput

X 384 =



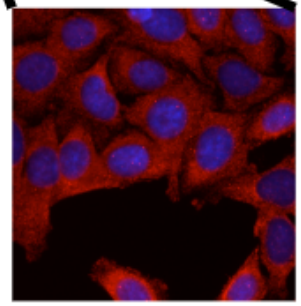
"Multi-well plate"



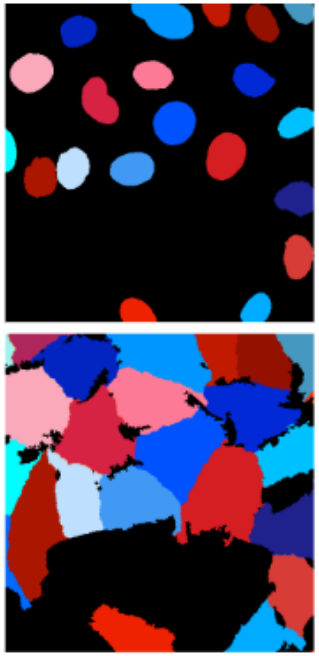
Anne Carpenter, Image: Nalgene; video: Chemistry World

Large scale imaging experiments

Cells or organisms in multiwell plates, each well treated with a gene or chemical perturbant

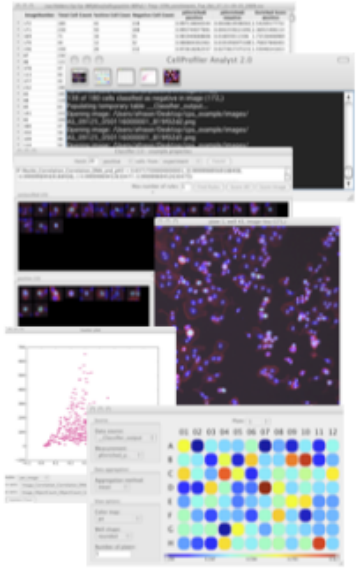


automated
microscopy
(any manufacturer)



Cell
measurements

(size, shape,
intensity, texture,
etc.)



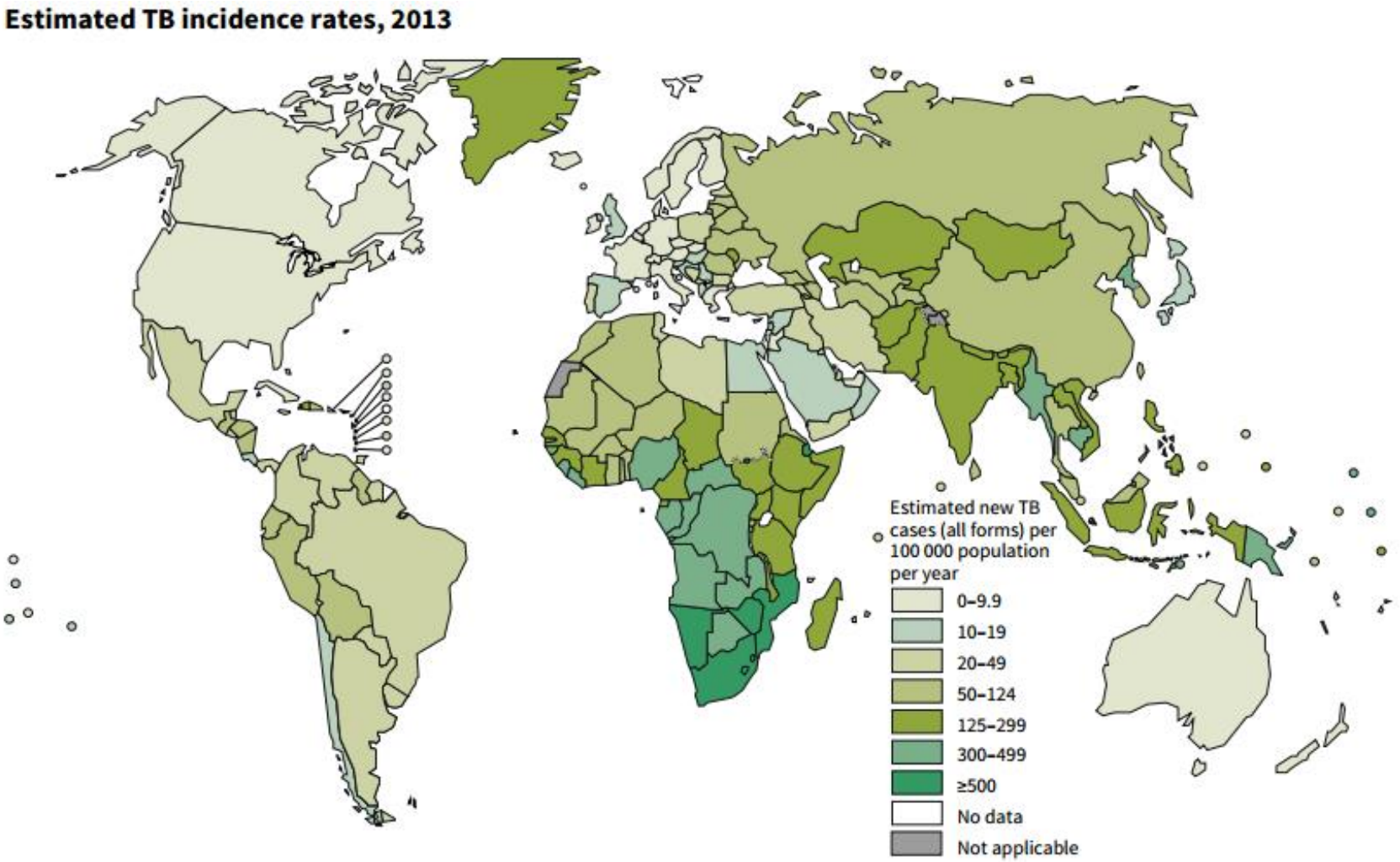
Data exploration
& machine learning

Anne Carpenter

Case study: Tuberculosis

Remains a leading cause of mortality globally

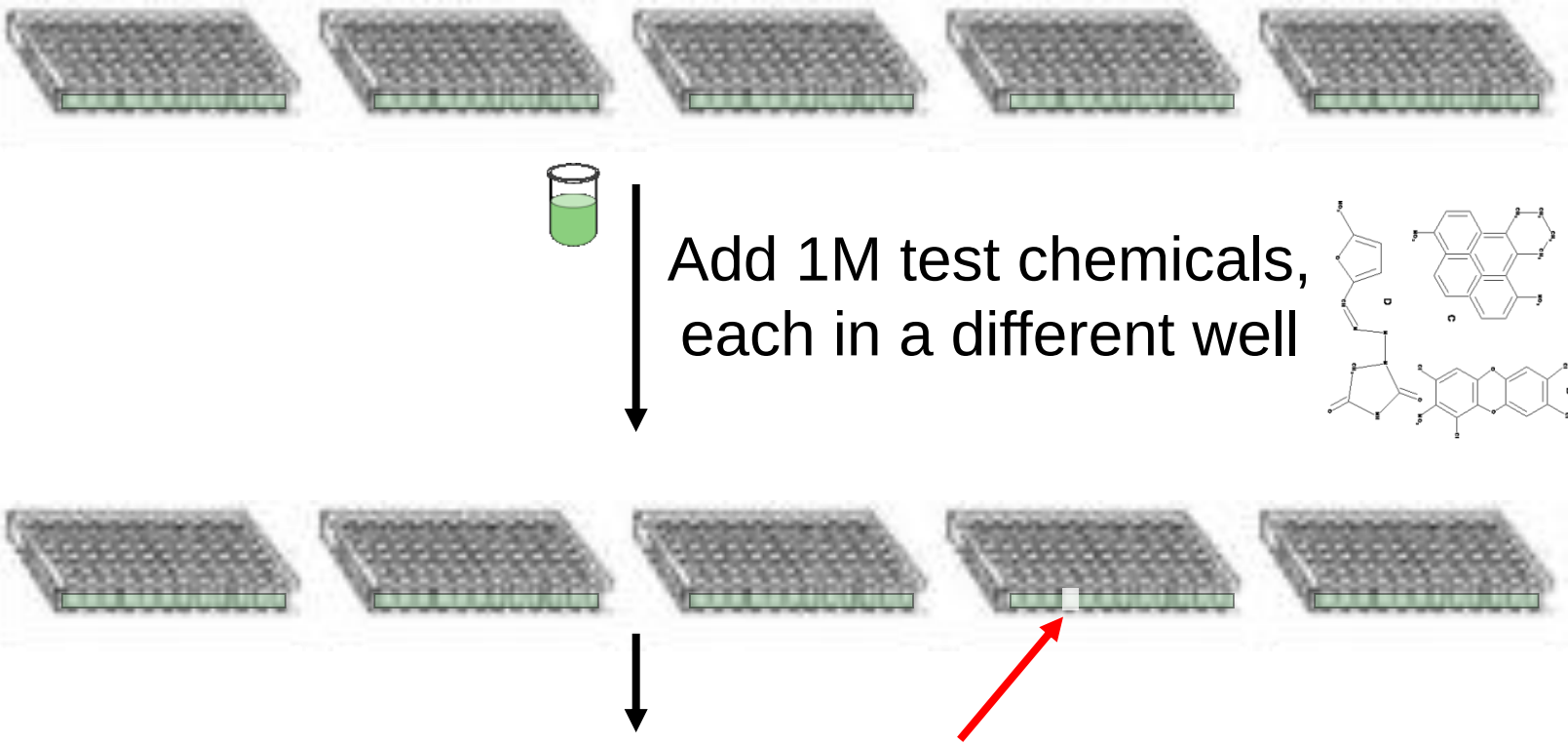
1/3 of the world is latently infected



Anne Carpenter, from WHO report, Global Tuberculosis Control 2013

Traditional approach to discover new antibiotics

Try to kill bacteria in individual wells of multi-well plates

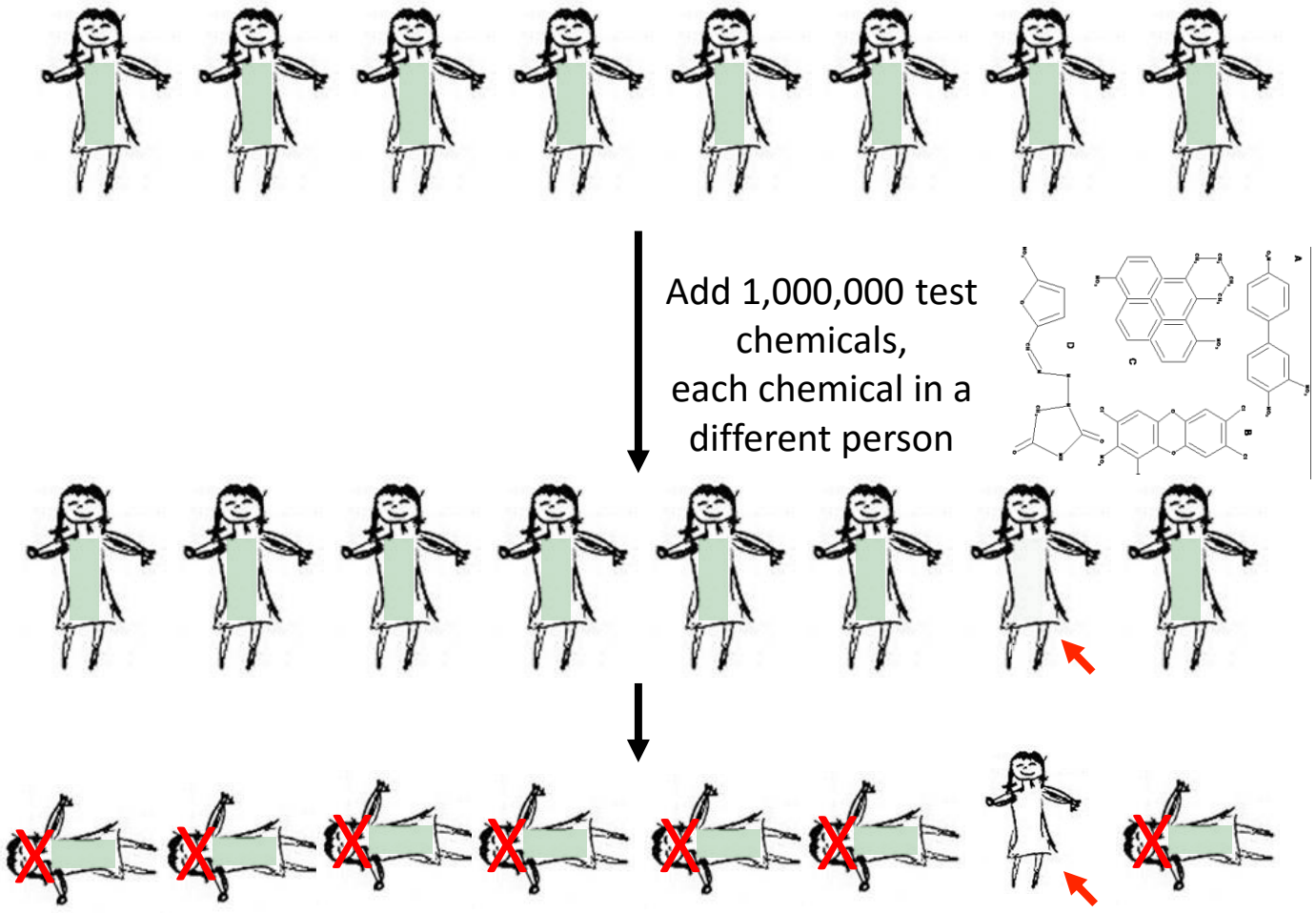


Add 1M test chemicals, each in a different well

Measure amount of bacteria (fluorescence plate reader) and look for wells where bacteria died

Anne Carpenter

Alternative approach to discover new antibiotics (effective, not ideal)



Anne Carpenter

Search for tuberculosis treatments

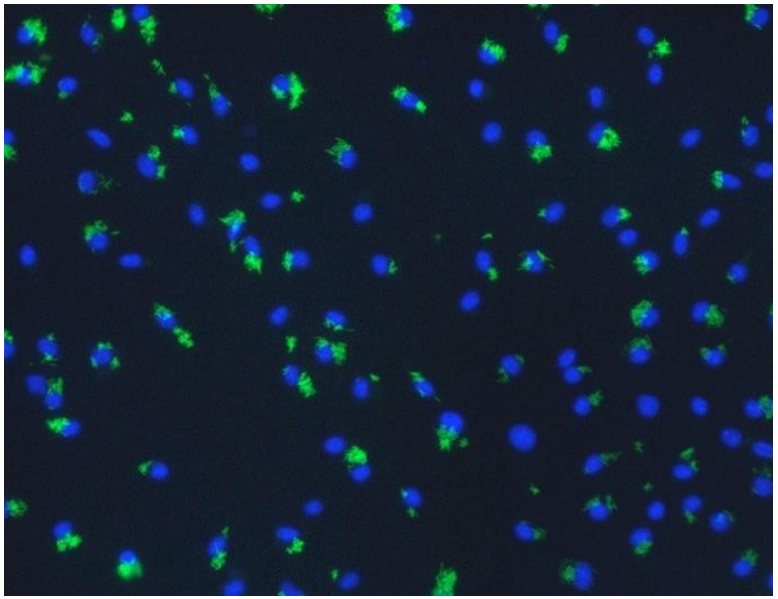
that prevent infection but do NOT kill the bacterium directly

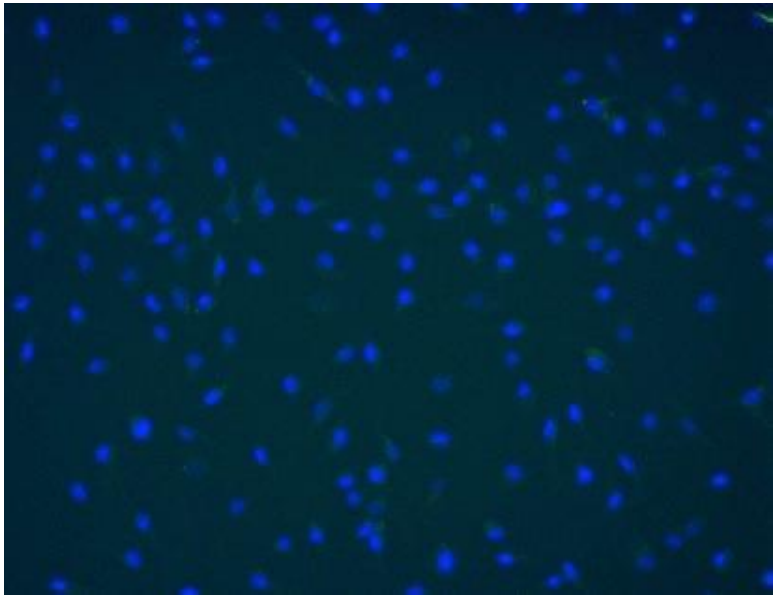
Without drug

With drug

mouse nuclei

tuberculosis bacteria

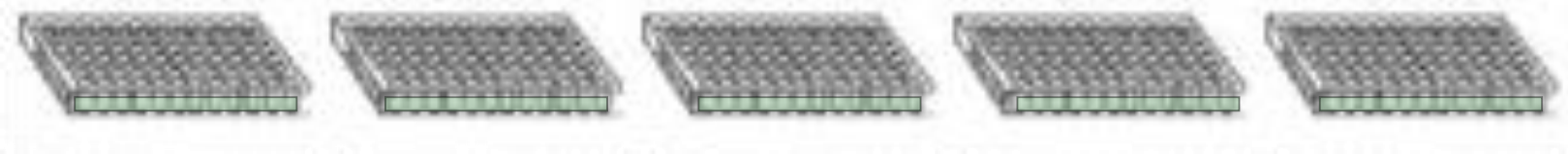




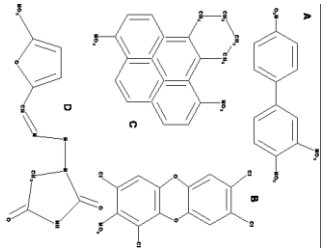
Anne Carpenter, from Stanley et al. (2014)

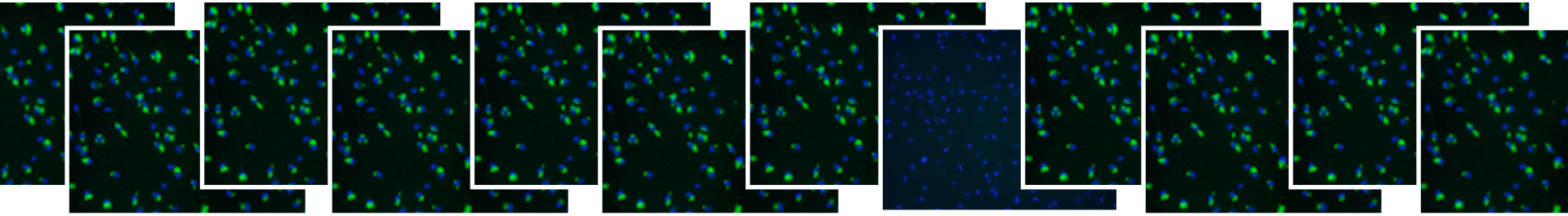
Search for tuberculosis treatments

Put bacteria and mouse cells in individual wells of multi-well plates



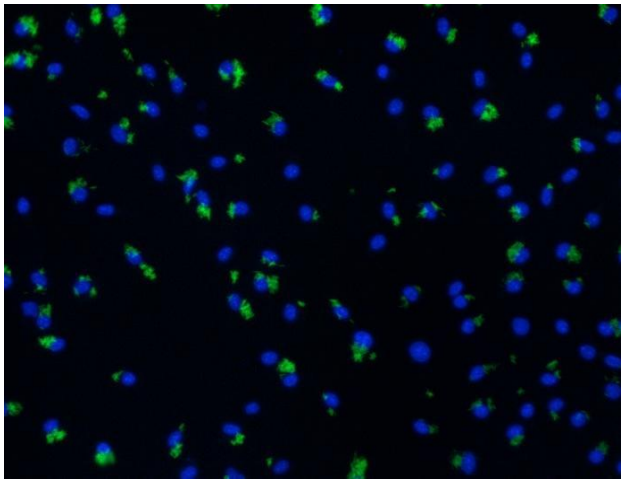
Add 10,000 bioactive chemicals, each chemical in a different well





Anne Carpenter

Automated image analysis pipeline



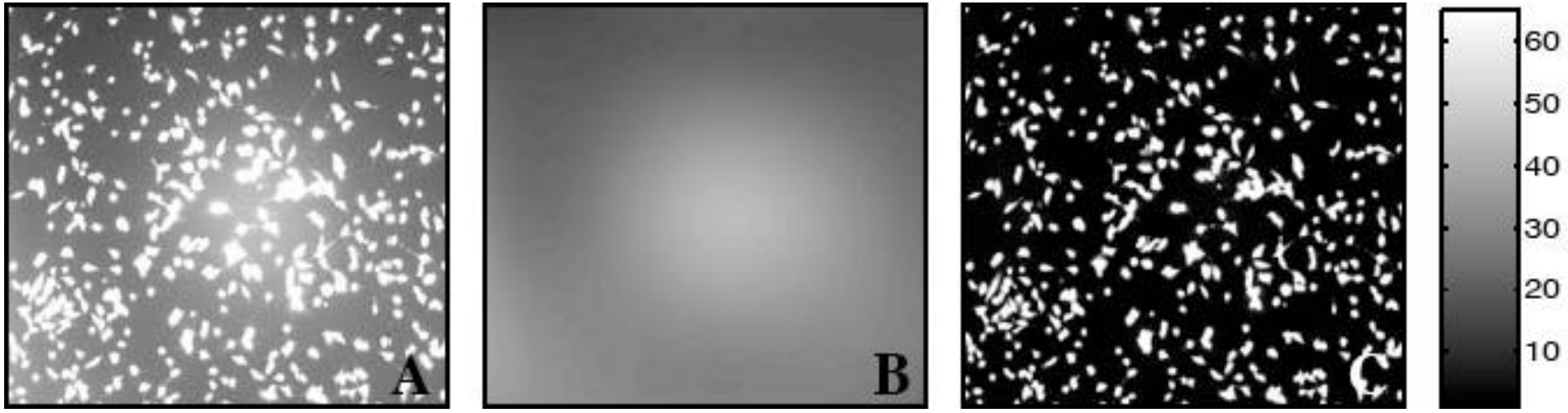
Find mouse nuclei

Find tuberculosis bacteria

Count the number of bacteria per nucleus

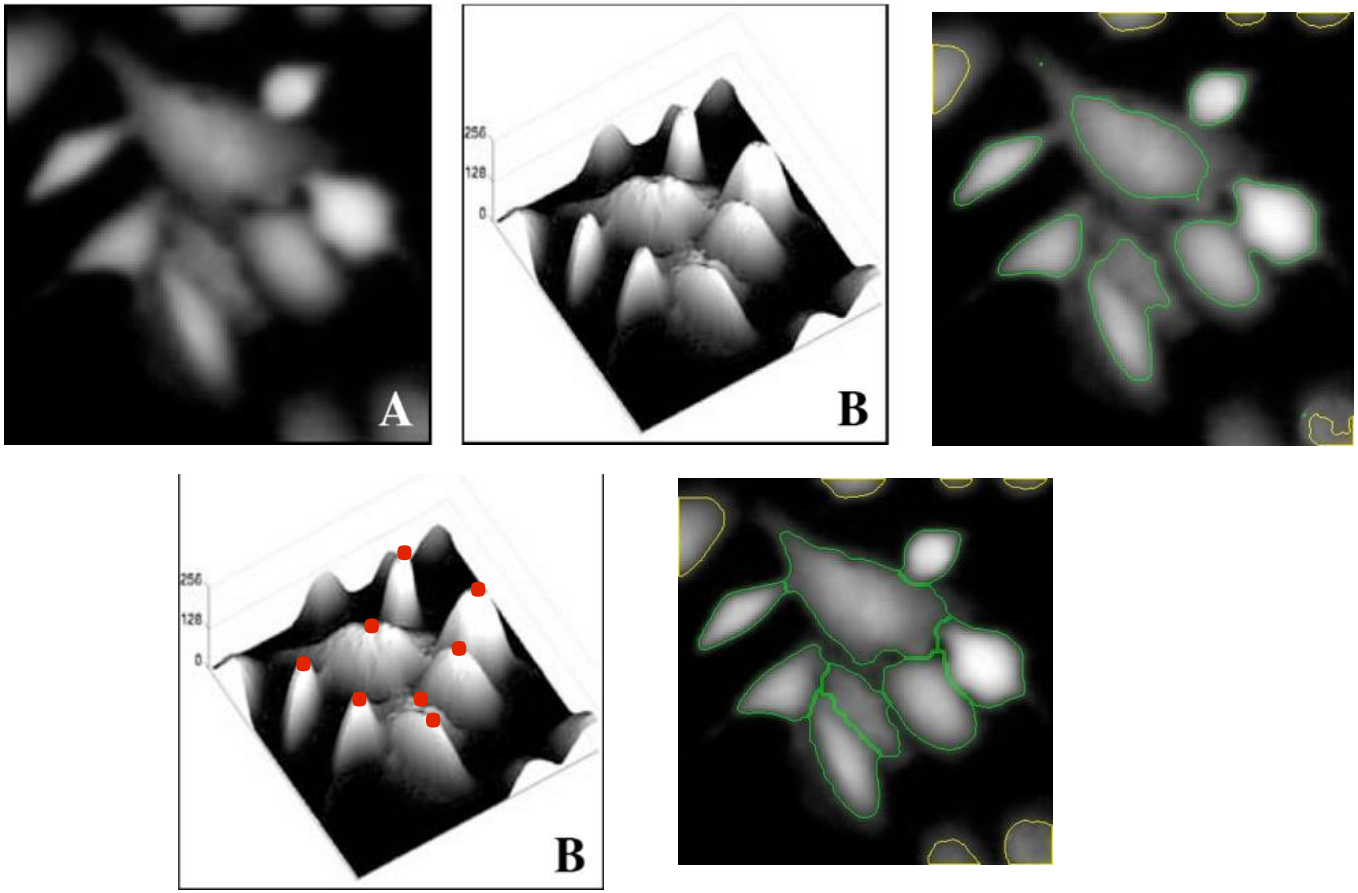
Anne Carpenter

Correct illumination



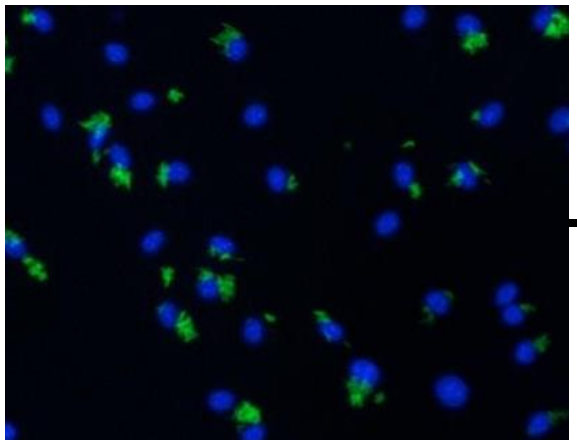
Anne Carpenter, images from Carolina Wahlby

Segment cells/nuclei

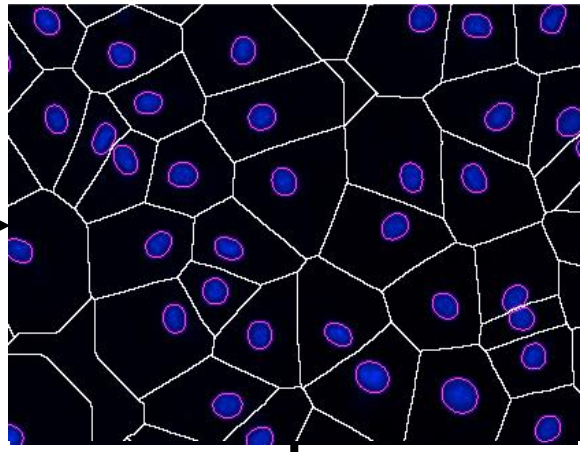


Anne Carpenter, images from Carolina Wahlby

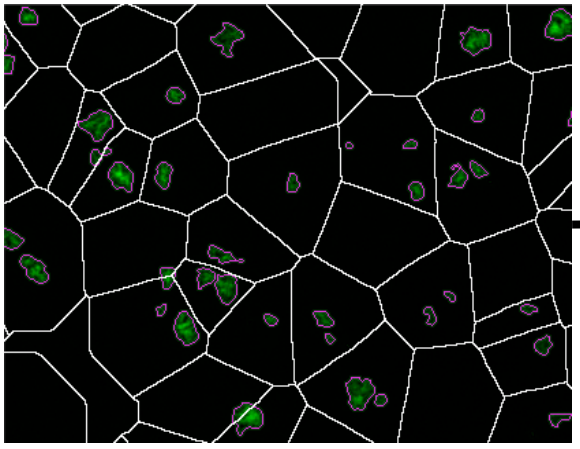
Automated image analysis



Find mouse nuclei



Find bacteria



Quantify the bacteria per mouse nucleus

Anne Carpenter

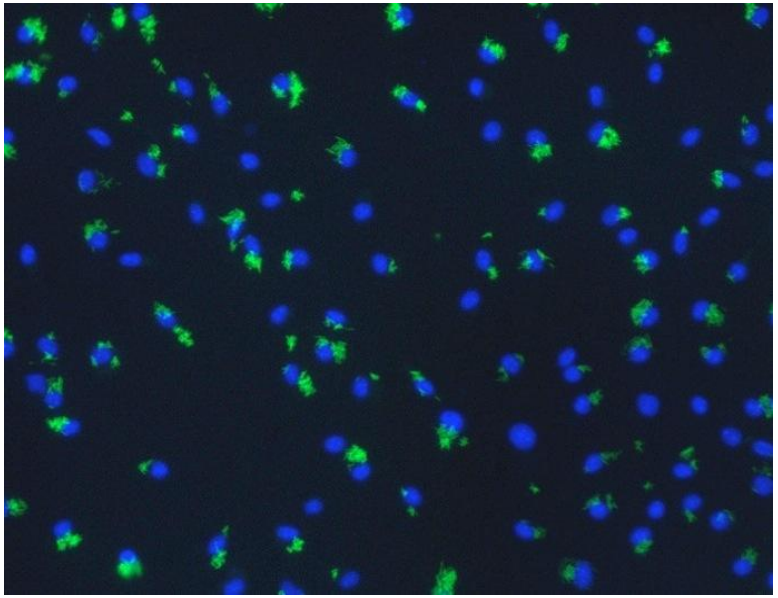
Search for tuberculosis treatments

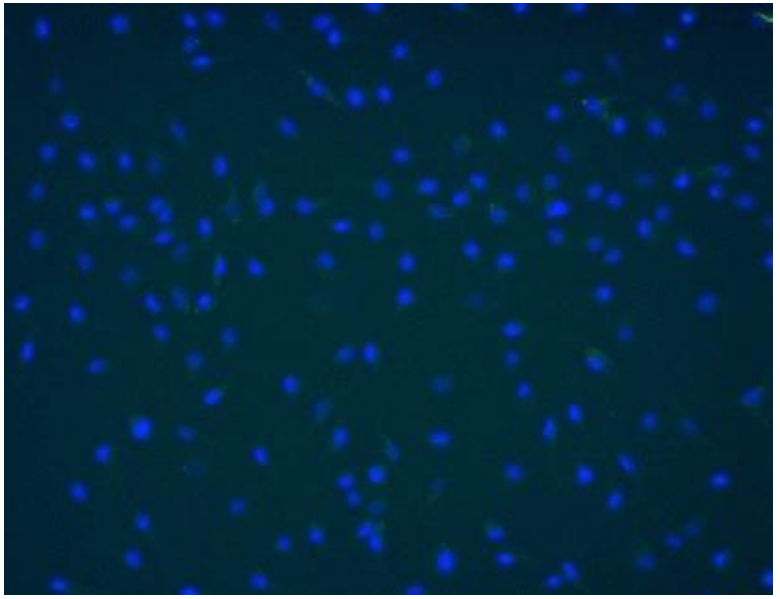
Without drug

With drug

mouse nuclei

tuberculosis bacteria





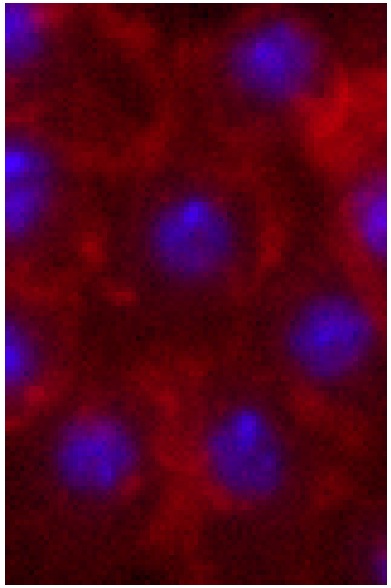
- Identified compounds that prevent bacterial infection/expansion but do NOT kill the bacterium directly
- Gefitinib reduces Mtb growth in the lungs of infected mice

Anne Carpenter, from Stanley et al. (2014)

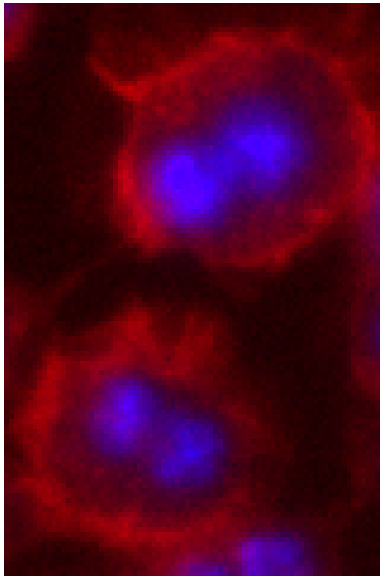
Regulators of cell division

DNA

Actin



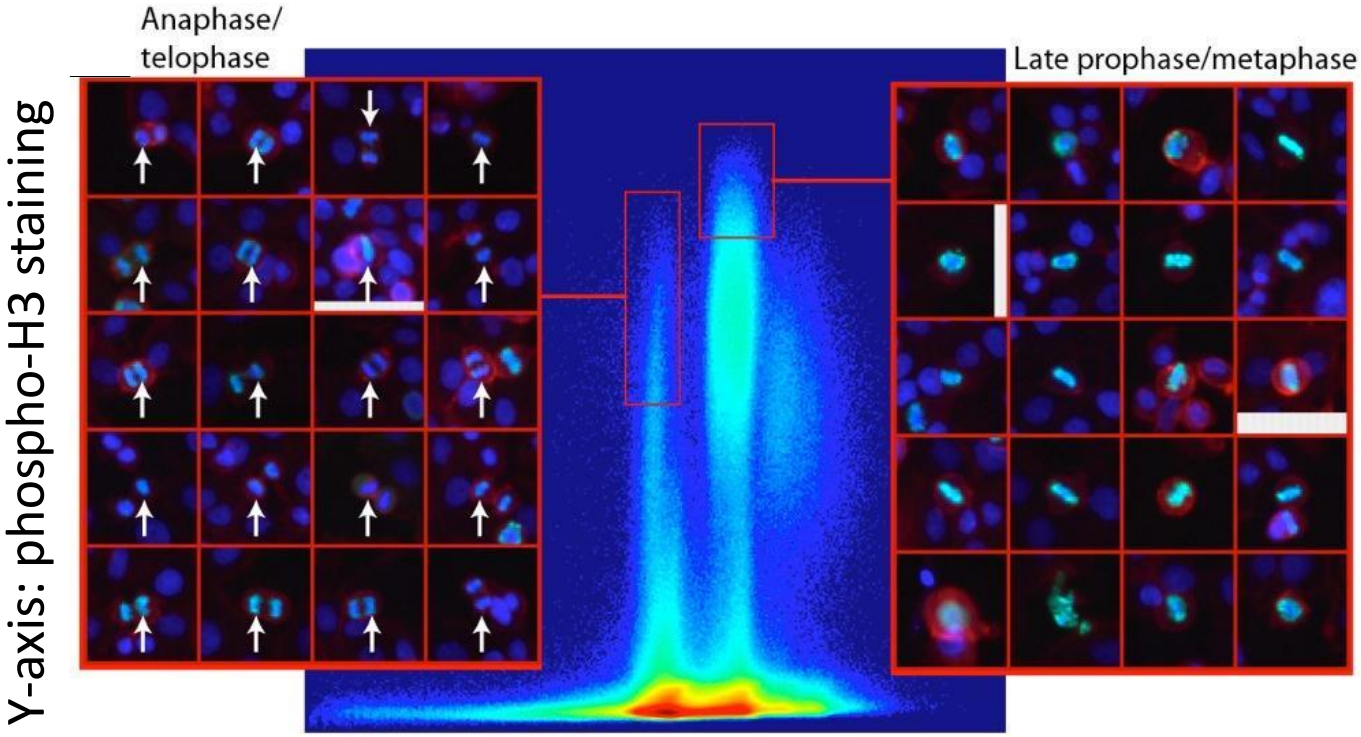
Normal:
one nucleus
per cell



Abnormal:
two nuclei
per cell

Anne Carpenter, from Castoreno et al. (2010)

Using > 1 features can identify interesting cells



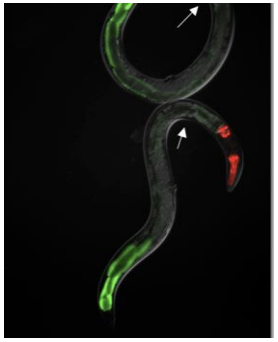
X-axis: DNA content

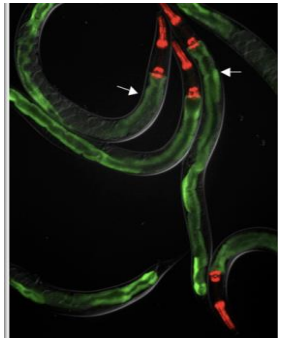
Anne Carpenter, from Castoreno et al. (2010)

Screen everything!

(Carpenter lab)

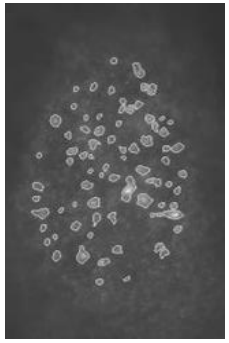
S. aureus infection

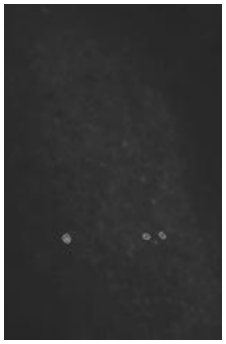




Ausubel/ Irazoqui labs

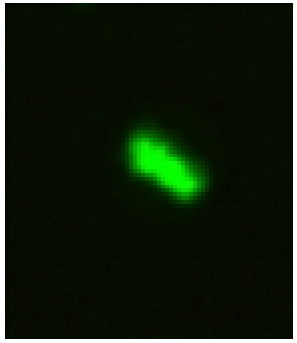
DNA damage

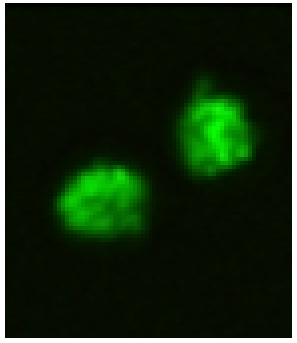




Yaffe lab

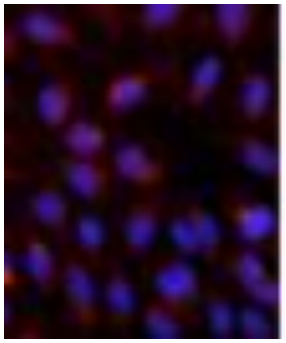
Mitosis

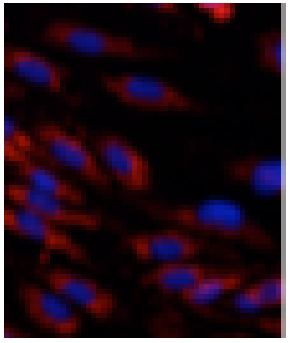




Mitchison lab

Mitochondrial abundance





Mootha lab

E. faecalis

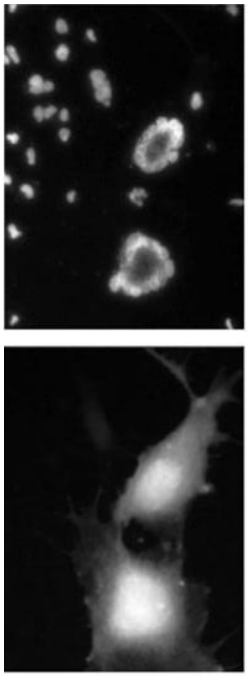
Ausubel lab

Anne Carpenter

Screen everything!

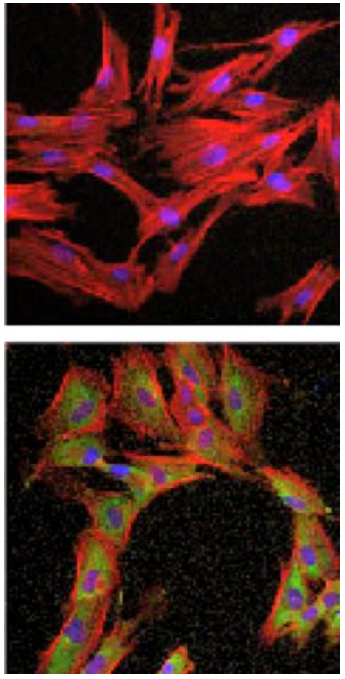
(Carpenter lab)

HIV neutralization



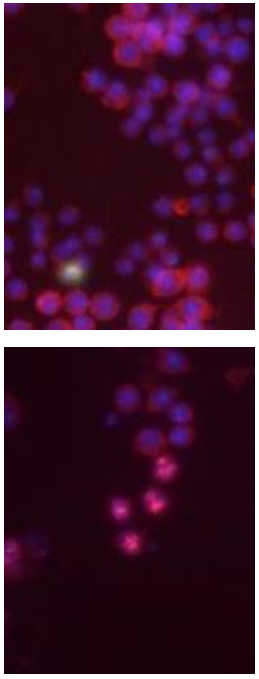
Fenyo lab

mTOR pathway activation



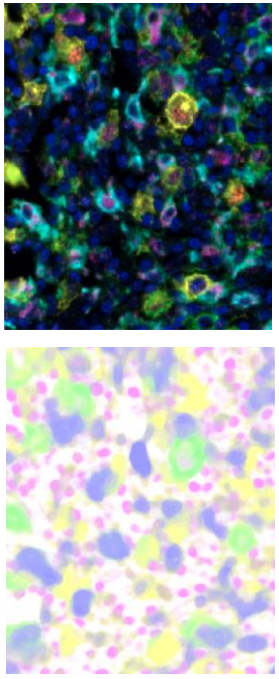
Sabatini lab

Meiosis



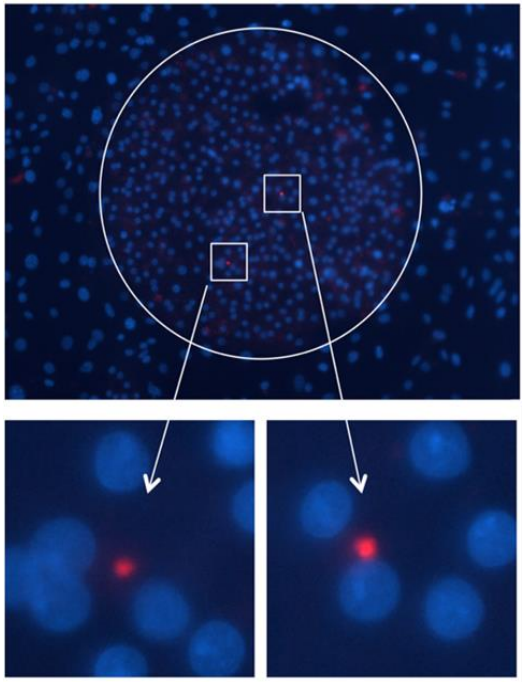
Orr-Weaver lab

Immune environment



Shipp/Rodig labs

Malaria



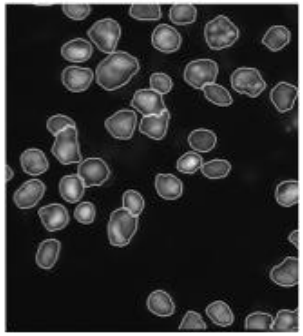
Bhatia lab

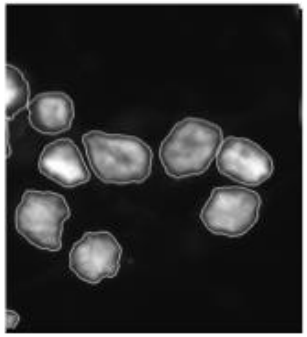
Anne Carpenter

Translational impact

(Carpenter lab, cellprofiler.org/impact)

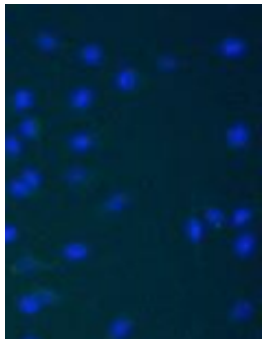
AMKL

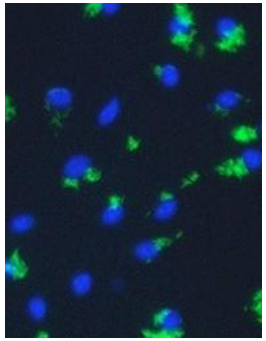




Crispino lab

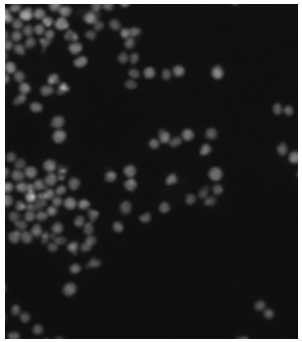
Tuberculosis

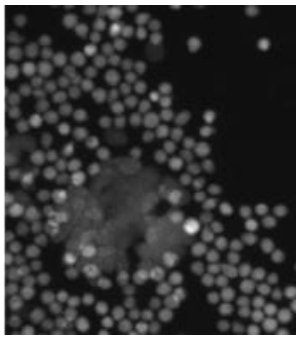




Hung lab

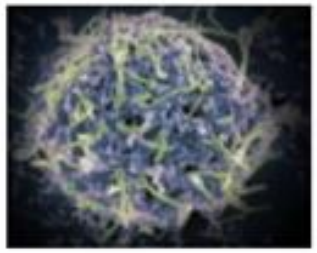
Leukemia





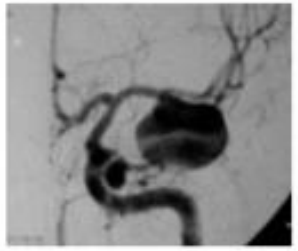
Gilliland/Scadden /Golub/Schreiber labs

Ebola



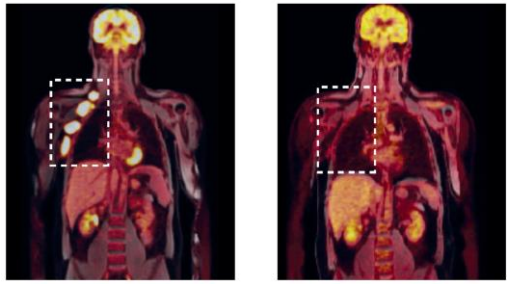
Texas Biomedical Research Institute

Cerebral cavernous malformation



University of Utah

Leukemias & Lymphomas

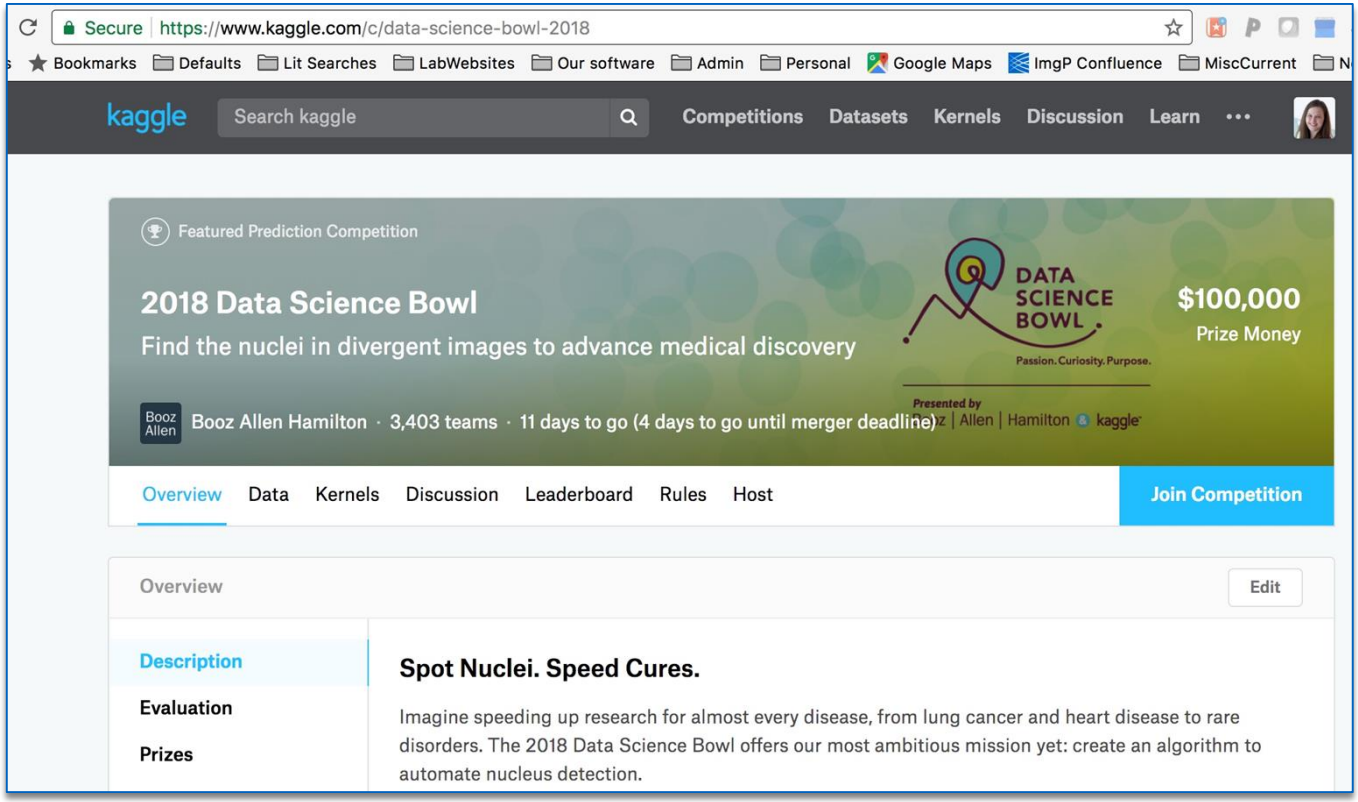


Vienna hospitals & medical institutes, ETH Zurich

Anne Carpenter

A key step is high throughput cell pheontyping is nuclei segmentation

2018 Data Science Bowl: Towards a Universal Nucleus Finder



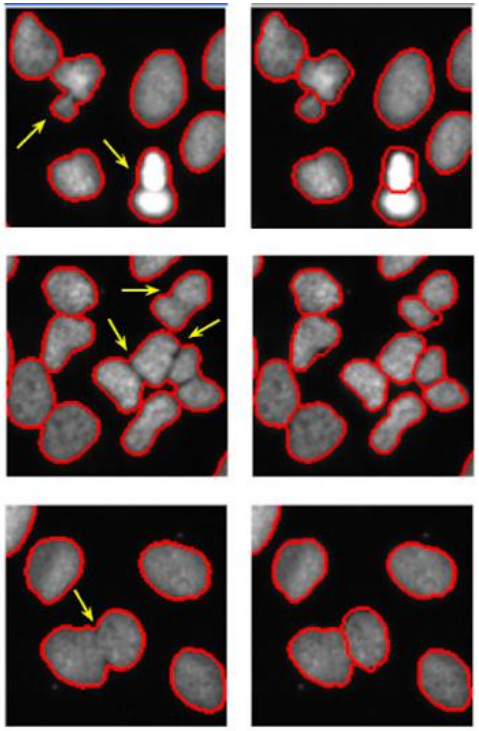
Anne Carpenter, from Caicedo et al. (2019)

A key step is high throughput cell pheontyping is nuclei segmentation

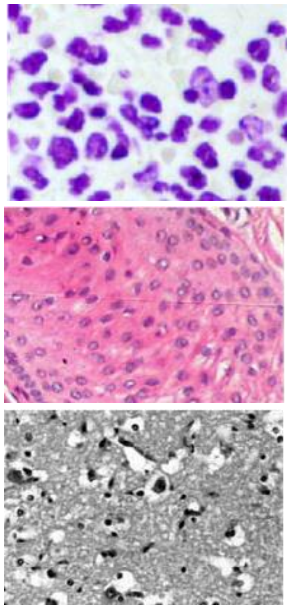
2018 Data Science Bowl: Towards a Universal Nucleus Finder

Split/merge errors

Ground truth



Goal: improve accuracy & generalizability



DATA SCIENCE BOWL

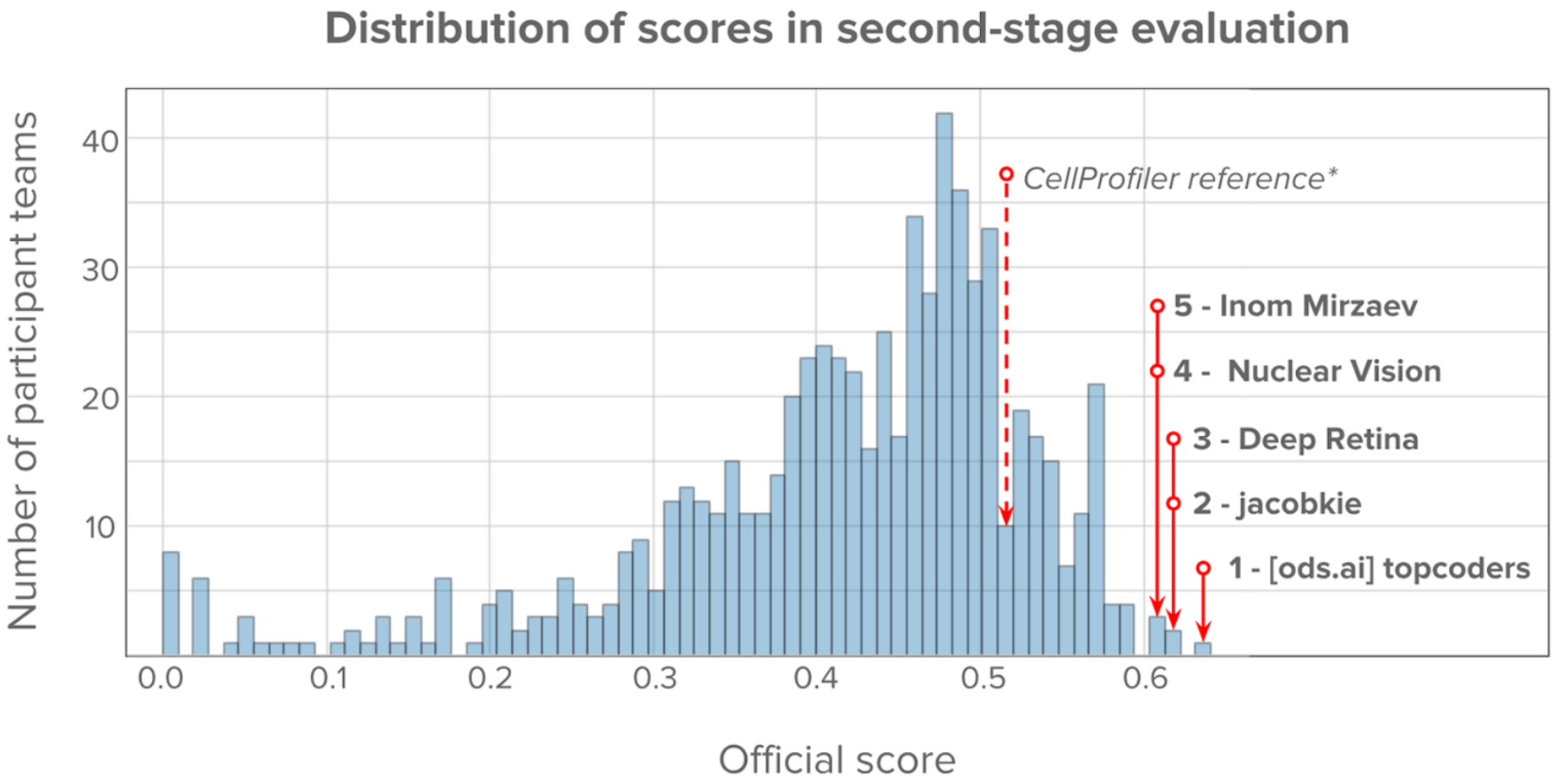
Passion. Curiosity. Purpose.

kaggle

Public dataset: 37,333 outlined nuclei
3,919 teams competed
65,333 submissions
Implemented web app:
www.NucleAIzer.org (Horvath)

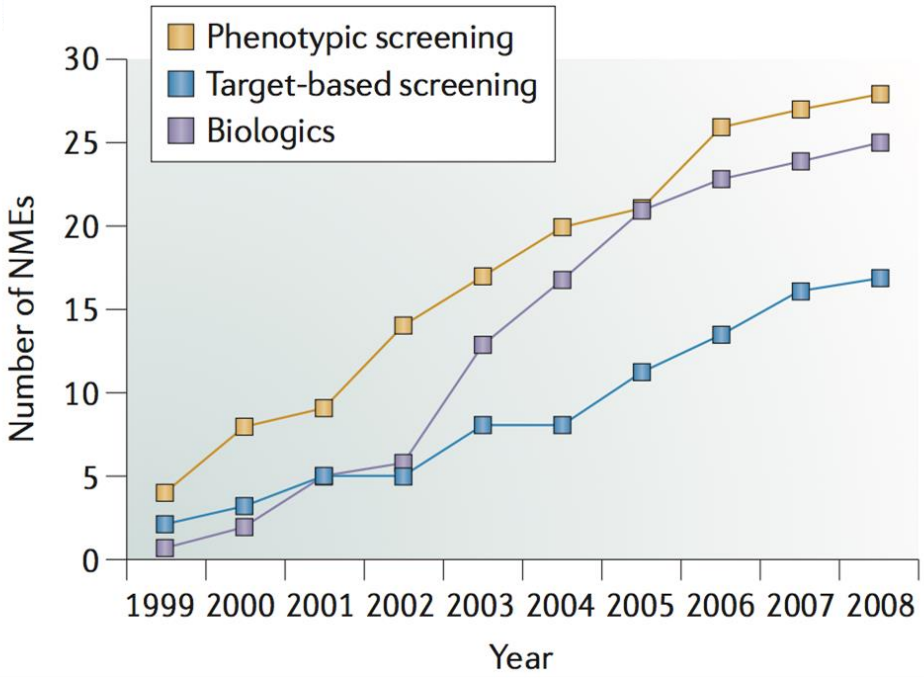
Anne Carpenter, from Caicedo et al. (2019)

Deep learning excels at segmentation

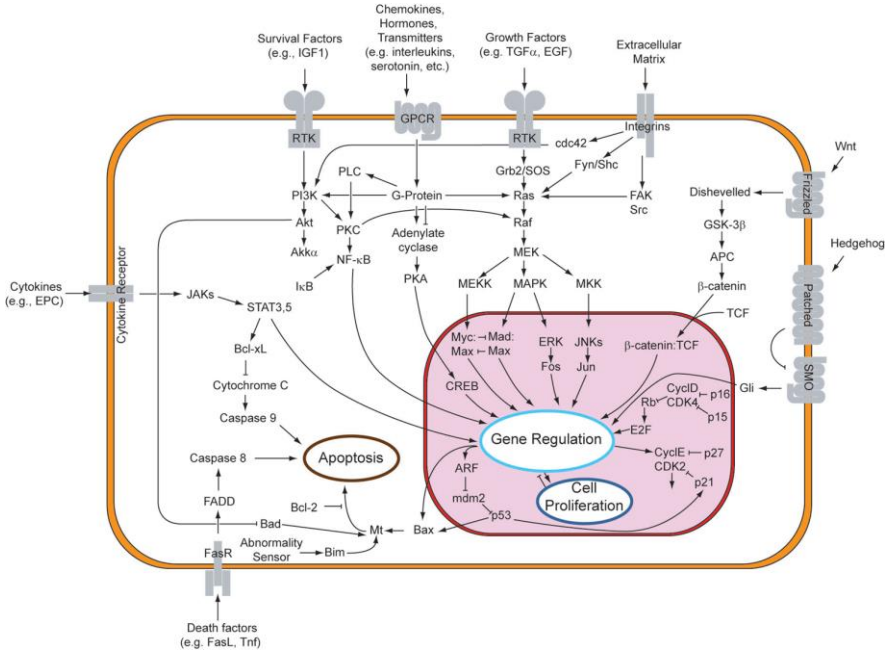


Anne Carpenter, from Caicedo et al. (2019)

Target-based vs. phenotypic drug discovery



Swinney and Anthony (2011),
Zheng et al. (2013),
Lee et al. (2013)



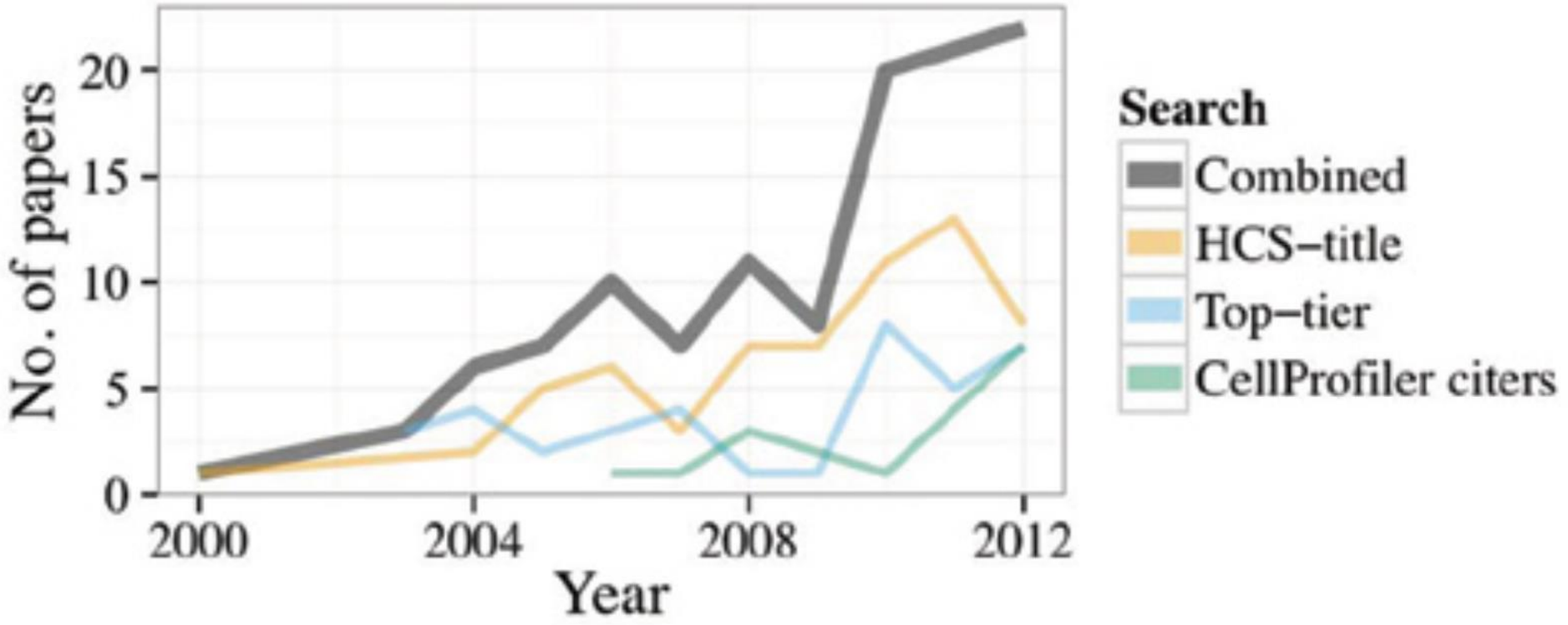
?

Genotype → phenotype

Anne Carpenter

Increased interest in high-content screens

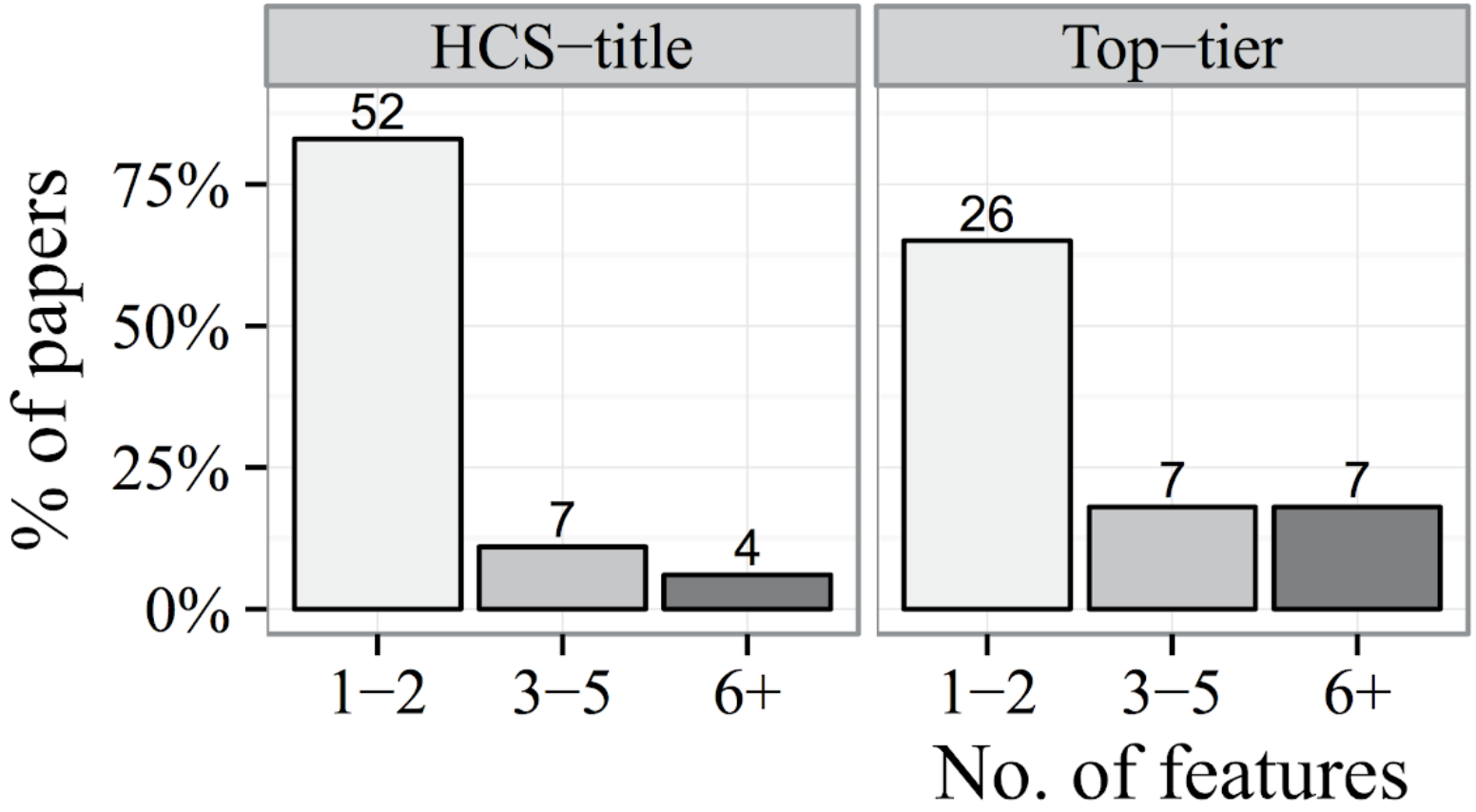
Number of papers in which a high-throughput, image-based experiment was used toward a discovery



Anne Carpenter, from Singh, Carpenter, Genovesio (2014)

Most high contact screens are not very high in information content…

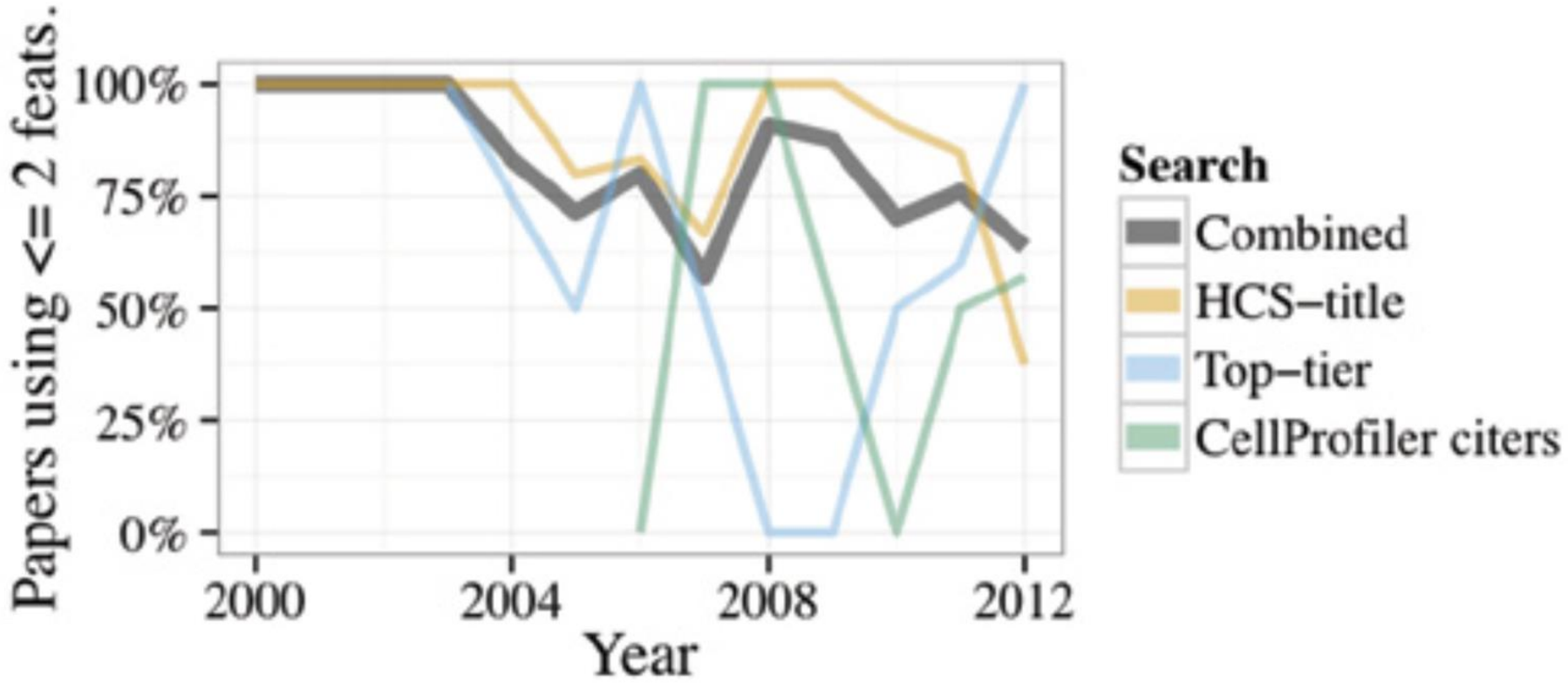
60-80% of “high-content” studies use only 1 or 2 cellular features



Singh, Carpenter, Genovesio (2014)

Most high contact screens are not very high in information content…

No dramatic improvement in information content in HCS over the past decade



Singh, Carpenter, Genovesio (2014)

The data science approach: measure everything, ask questions later…

"Cytological profile": collection of measurements describing the appearance of a cell

Perlman et al. (2004)

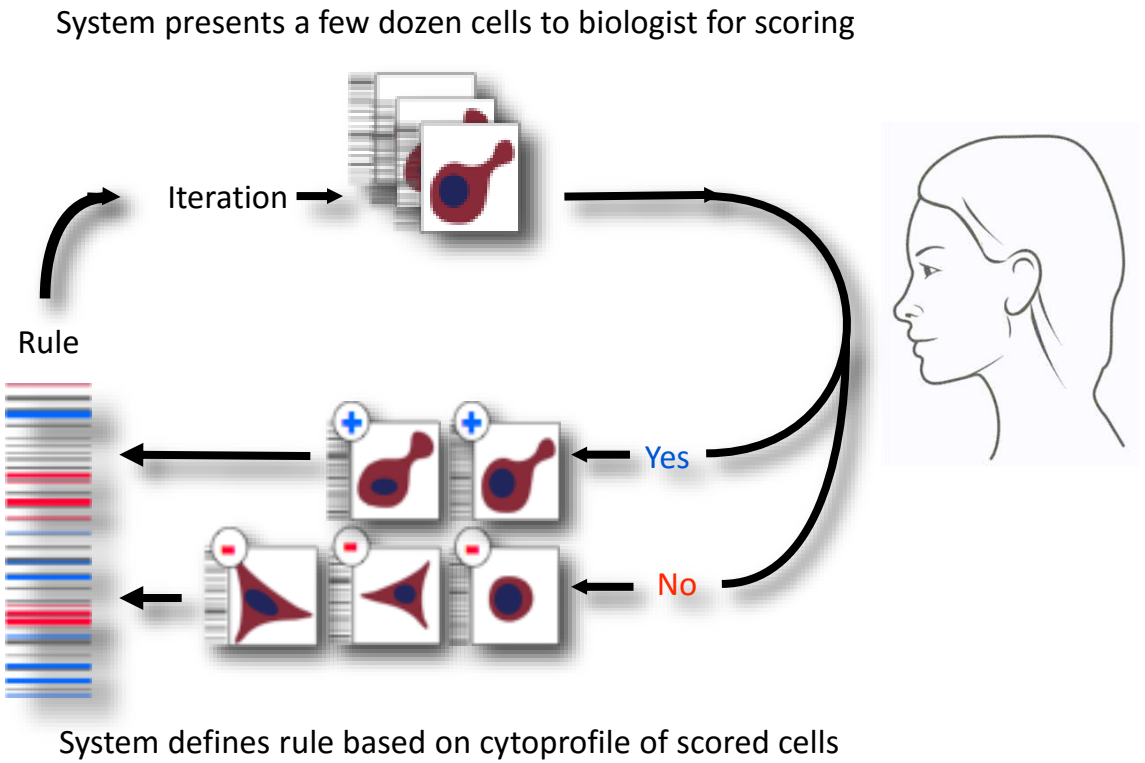
Anne Carpenter

Measuring everything, asking questions later

Hundreds of features: size, shape, staining intensity, texture, etc.

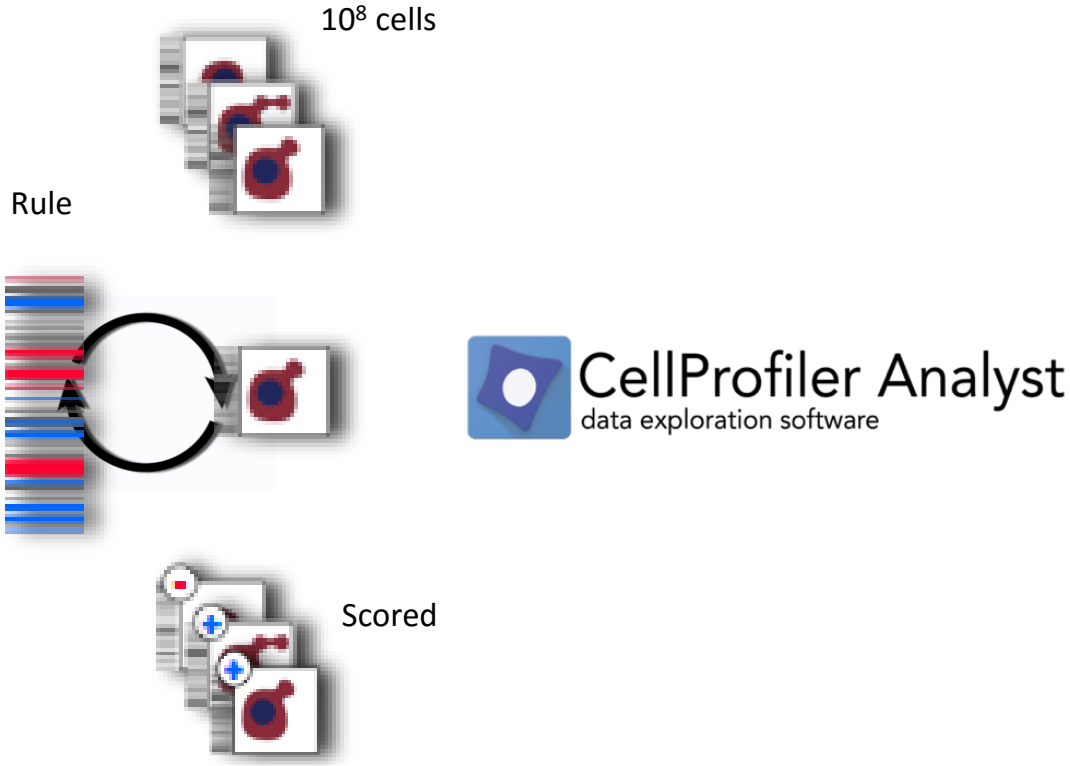
- Why?
  - Several features may be necessary to score the phenotype
  - Virtual secondary screens can help characterize hits
  - Later re-screening for new phenotypes
  - The measurements required to score the phenotype of interest may not be known a priori
  - The full spectrum of cellular responses to each treatment (even those not visible by eye) may be useful for data mining/machine learning/clustering…systems biology

Iterative machine learning



Anne Carpenter, from Jones et al. (2009), Dao et al. (2016)

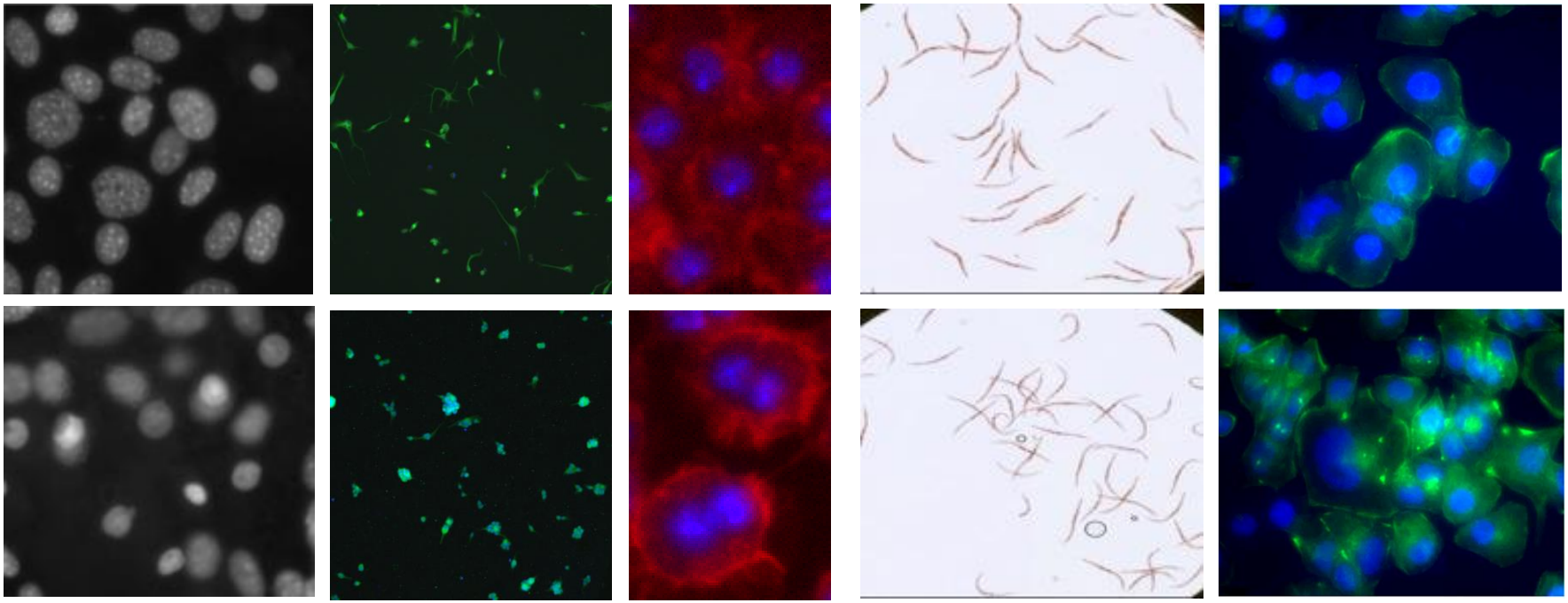
Iterative machine learning



Scored cells are sorted by well:
identify samples with a high proportion of positive cells

Anne Carpenter, from Jones et al. (2009), Dao et al. (2016)

Screen more complex phenotypes



CellProfiler Analyst
data exploration software

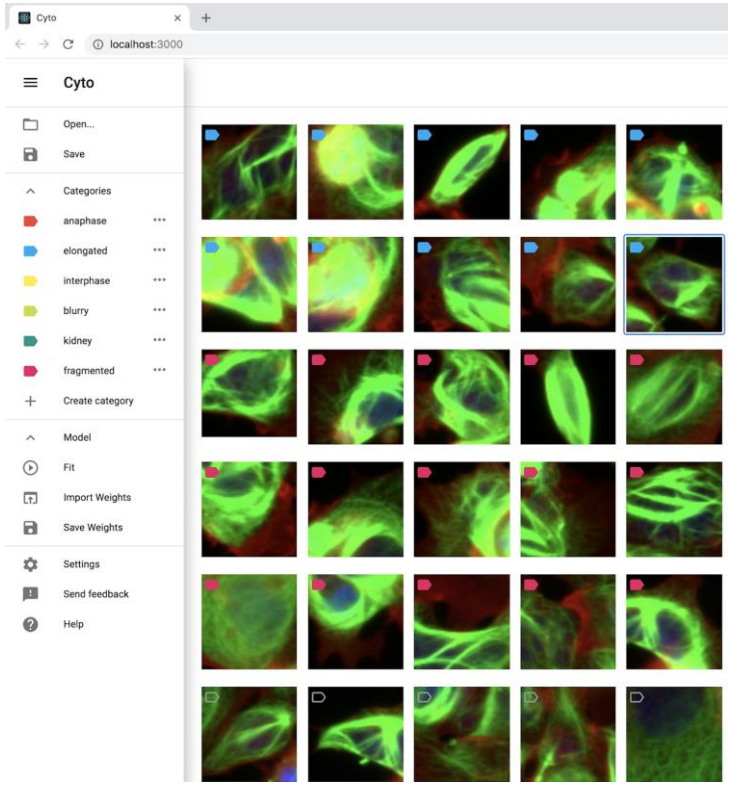
→ Piximi, https://github.com/piximi/application

Anne Carpenter

Piximi: deep learning cell classifier (under construction)

To replace CellProfiler Analyst (Carpenter lab)
+ Advanced Cell Classifier (Horvath lab)

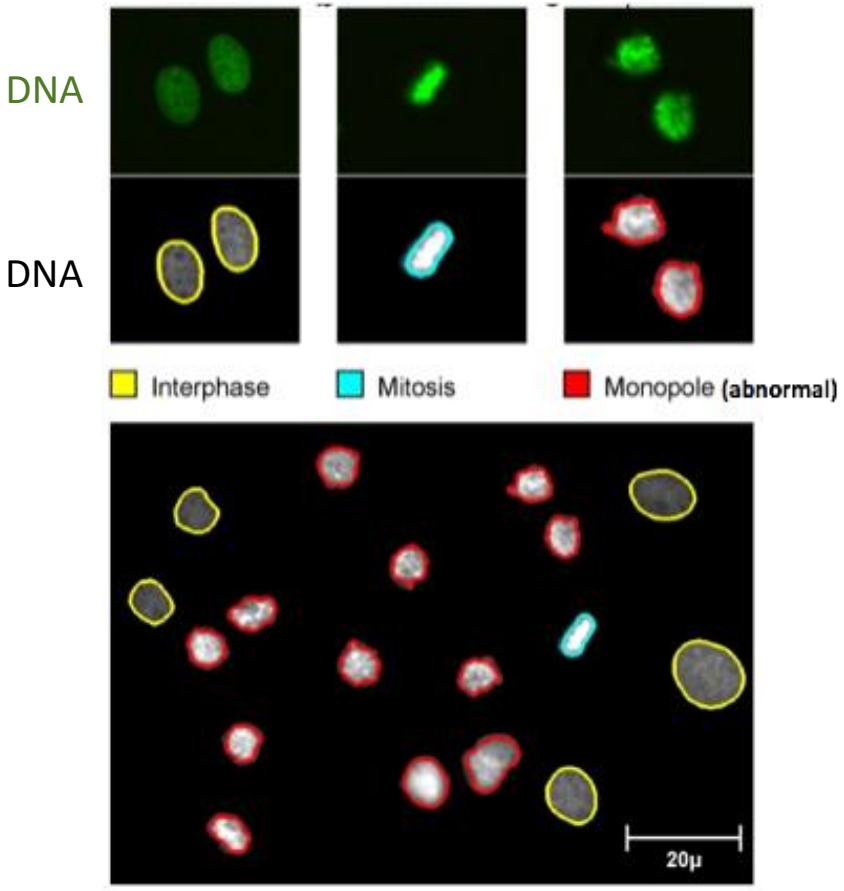
www.piximi.app
(under construction)



- Phenotype classifier for images
- **Web-based**
- Uses **deep learning** algorithms
- Interactive / intuitive user interface
- Source code: https://github.com/piximi
- Collaboratively developed!

Anne Carpenter

Regulators of cell division



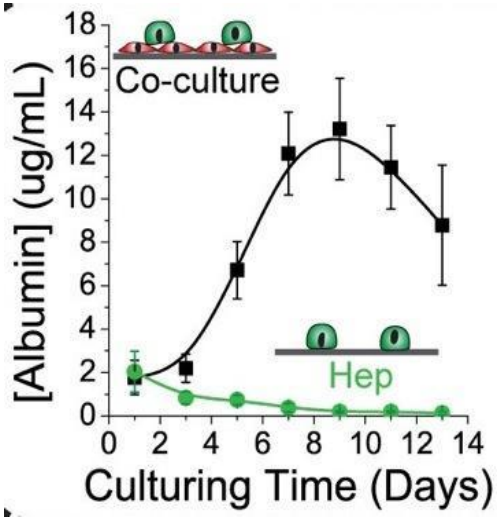
Anne Carpenter, from Tsue et al. (2009)

Co-cultured cell systems

Two or more cell types cultured together in order to maintain physiological conditions

Necessary

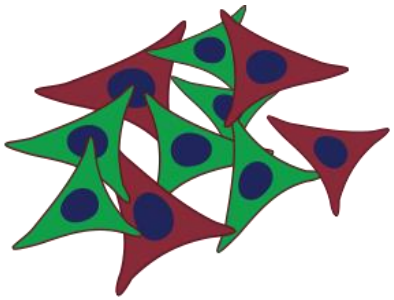
- Many primary cell types lose their physiological functions when grown in isolation



Challenging

- Culture conditions are difficult to optimize and less robust

- Need to distinguish the cell type of interest from the co-cultured cells, ideally without using additional cellular stains

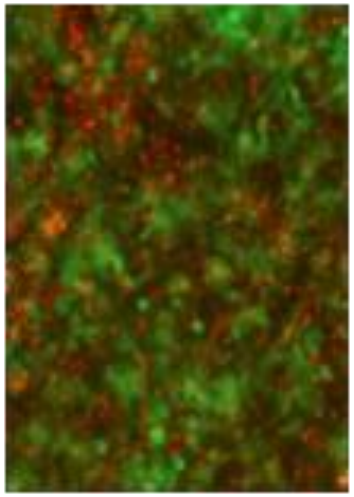


Anne Carpenter

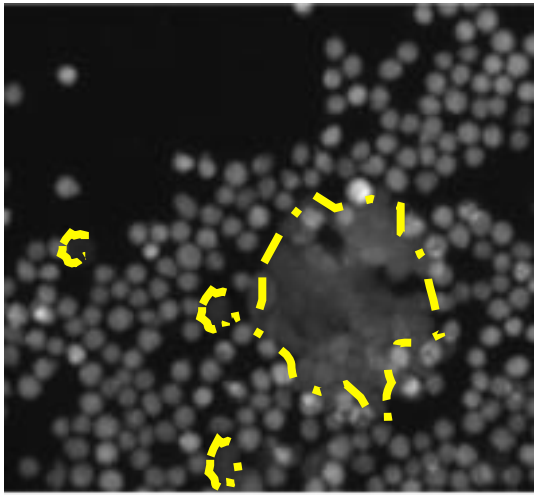
Leukemic & hematopoetic stem cells (HSCs/LSCs)

using mouse primary HSCs or LSCs co-cultured with stromal cells

Co-cultured LSCs and stroma



LSC channel only: live, no DNA stain



Cobblestones

Identified drugs that preferentially reduce leukemic cell growth; lovastatin extends lifespan of mice given leukemic bone marrow cells

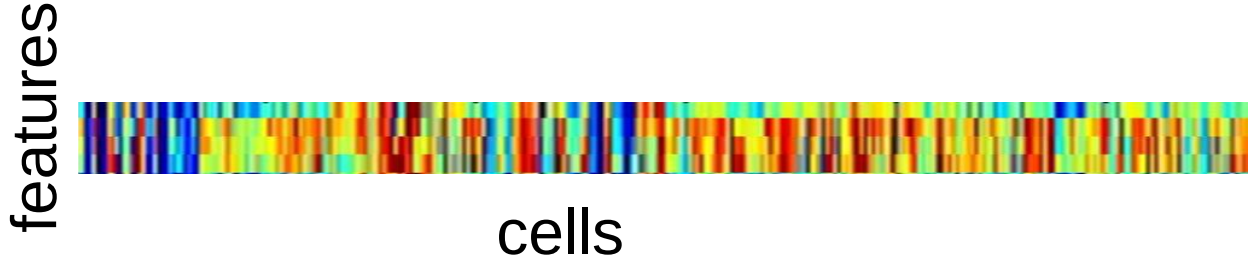
Anne Carpenter

Interpretability

- Check what features

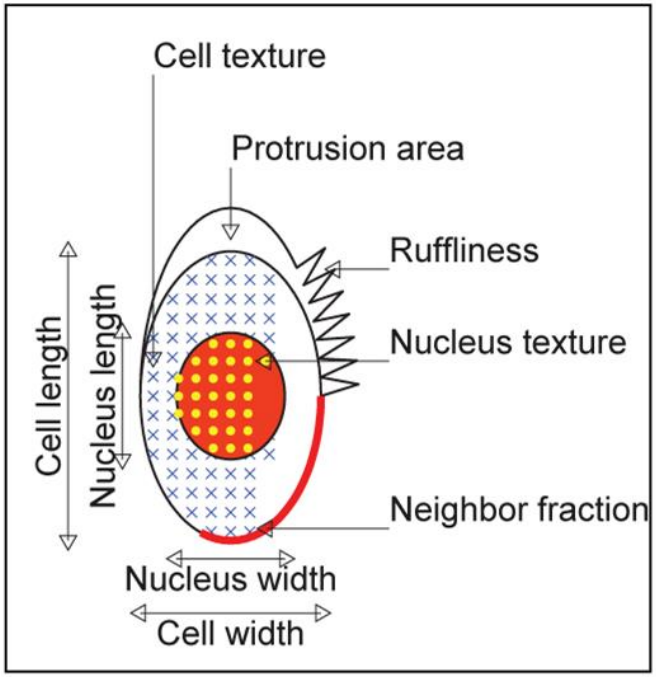
Visualization of high dimensional cellular data

- High dimensional cellular data
- Processing pipeline
- Why do we need visualization?
  - Data interpretation
  - Hypothesis formulation
  - Communication of results
- Visualization methods:
  - Bar charts, scatter plots (3D)
  - Heatmaps



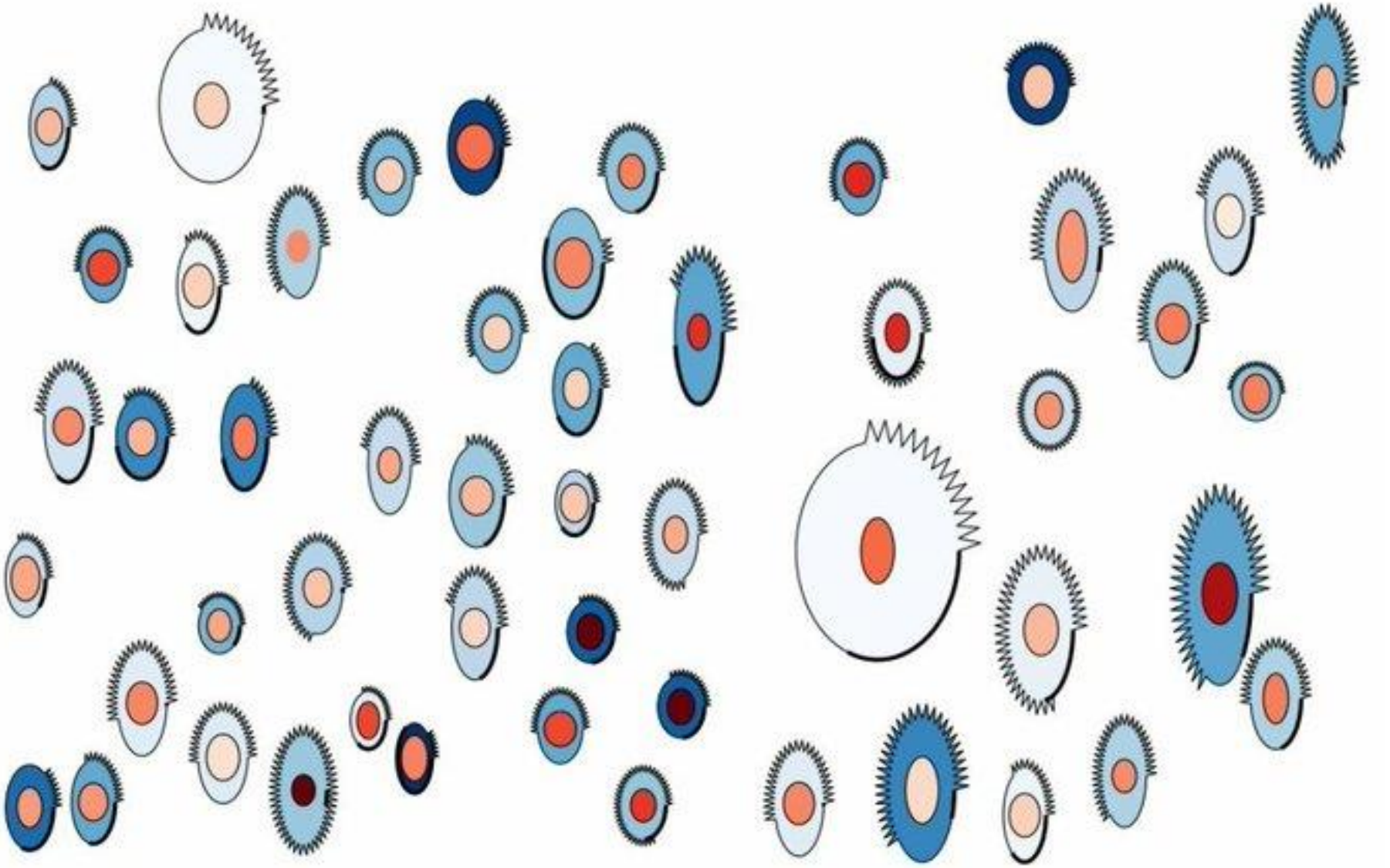
PhenoPlot

- A glyph-based method: “use a collection of visual elements such as size, colour, texture and/or orientation to depict multi-dimensional data”
- Free open source Matlab toolbox (+ GUI)
- 21 variables encoded
- Allows customization



Sailem et al. (2015)

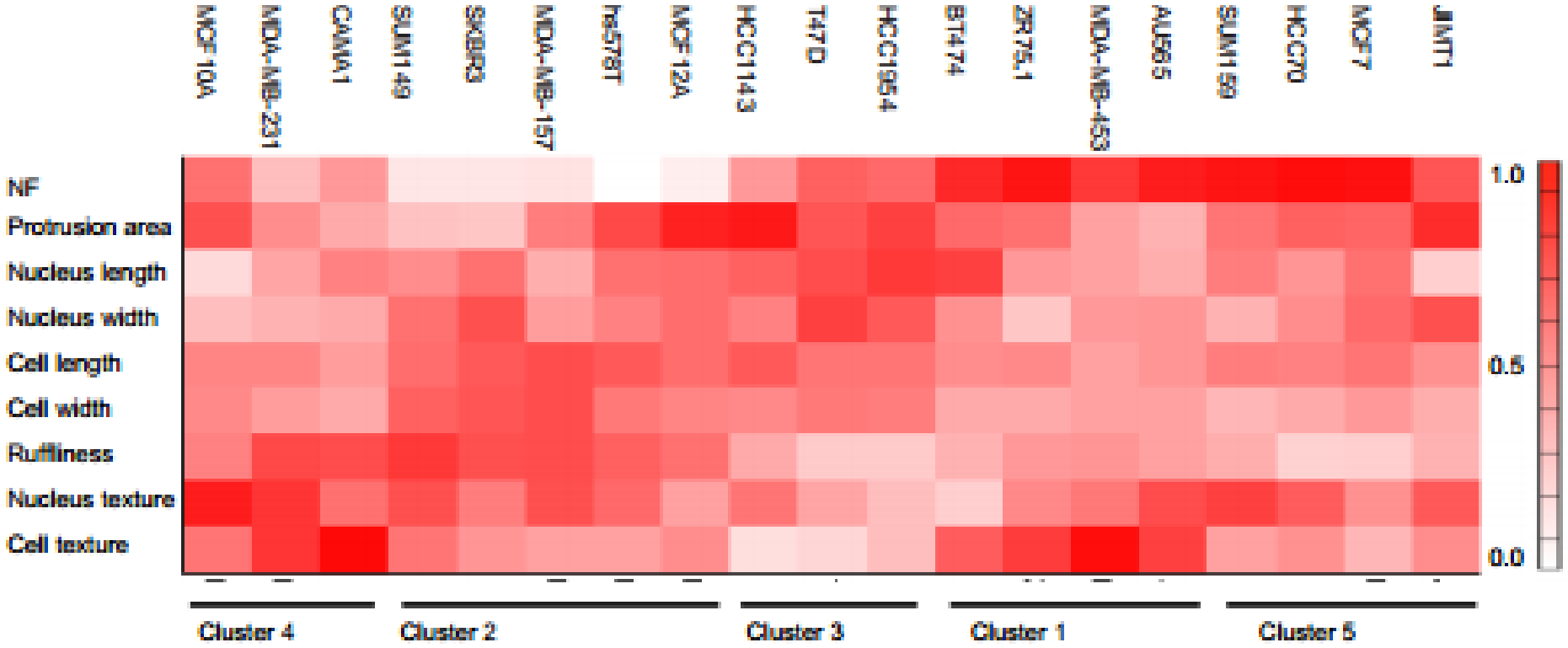
PhenoPlot



Sailem et al. (2015)

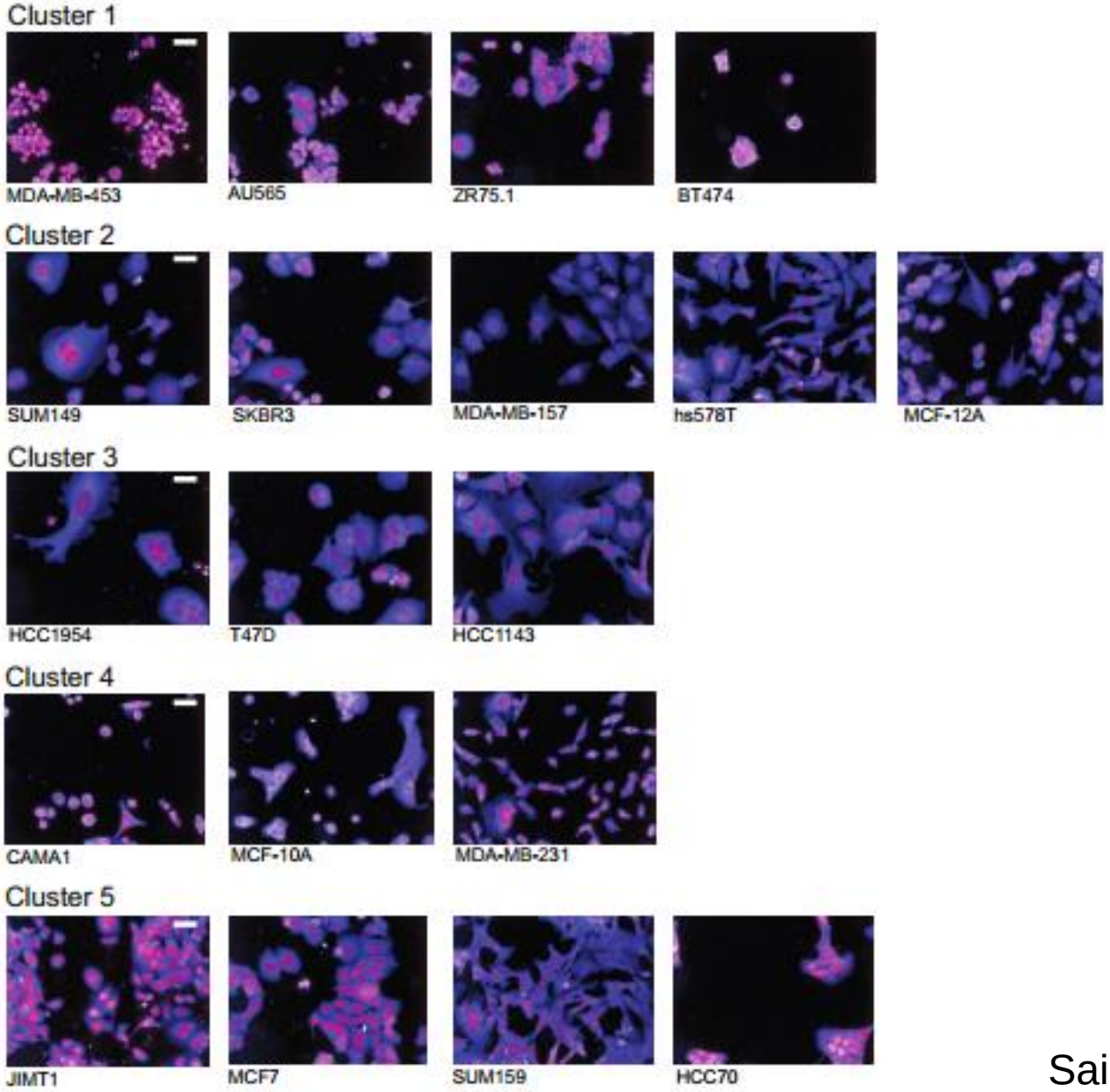
Hierarchical Clustering

N ~ 156,000 cells



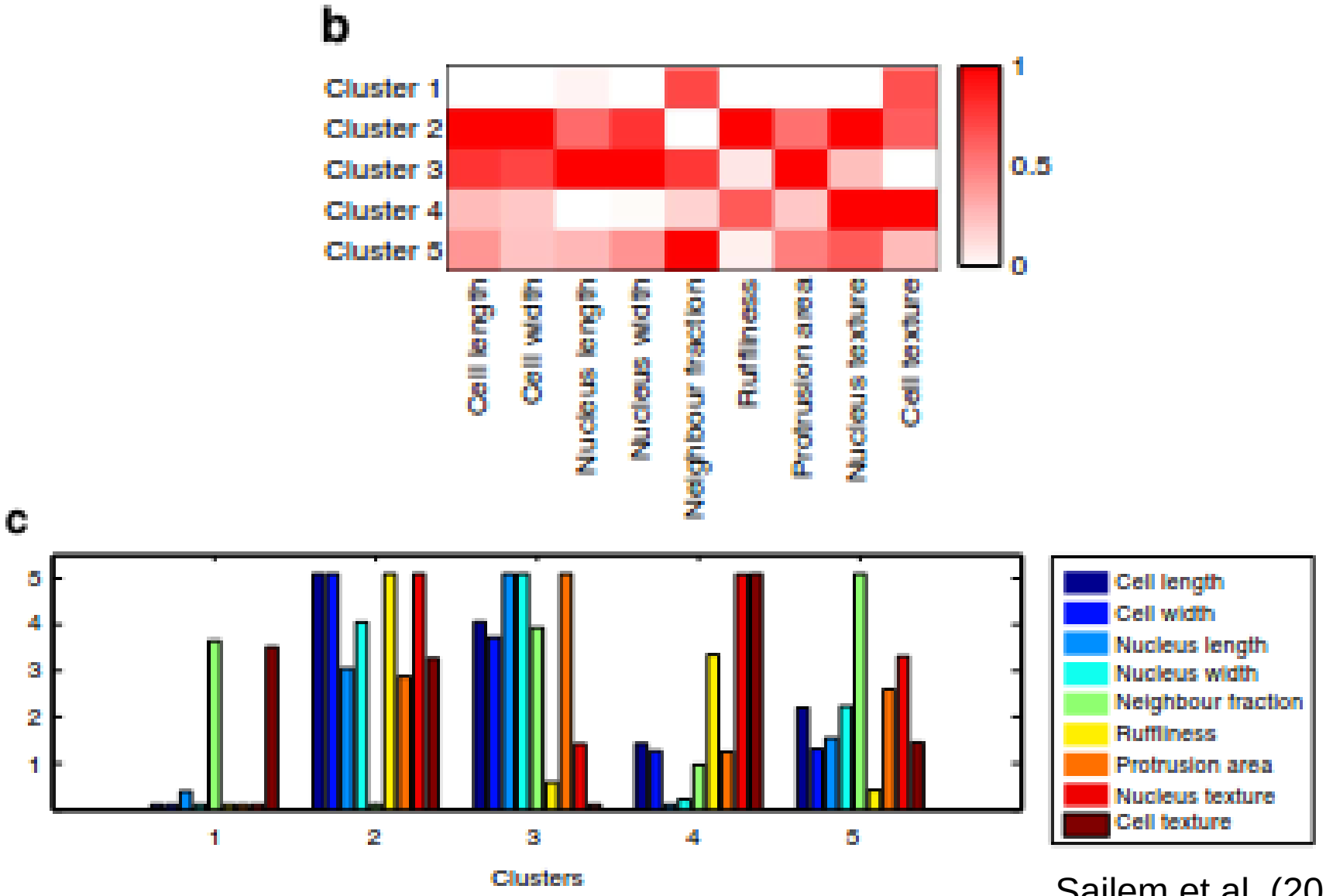
Sailem et al. (2015)

Clusters Visualization



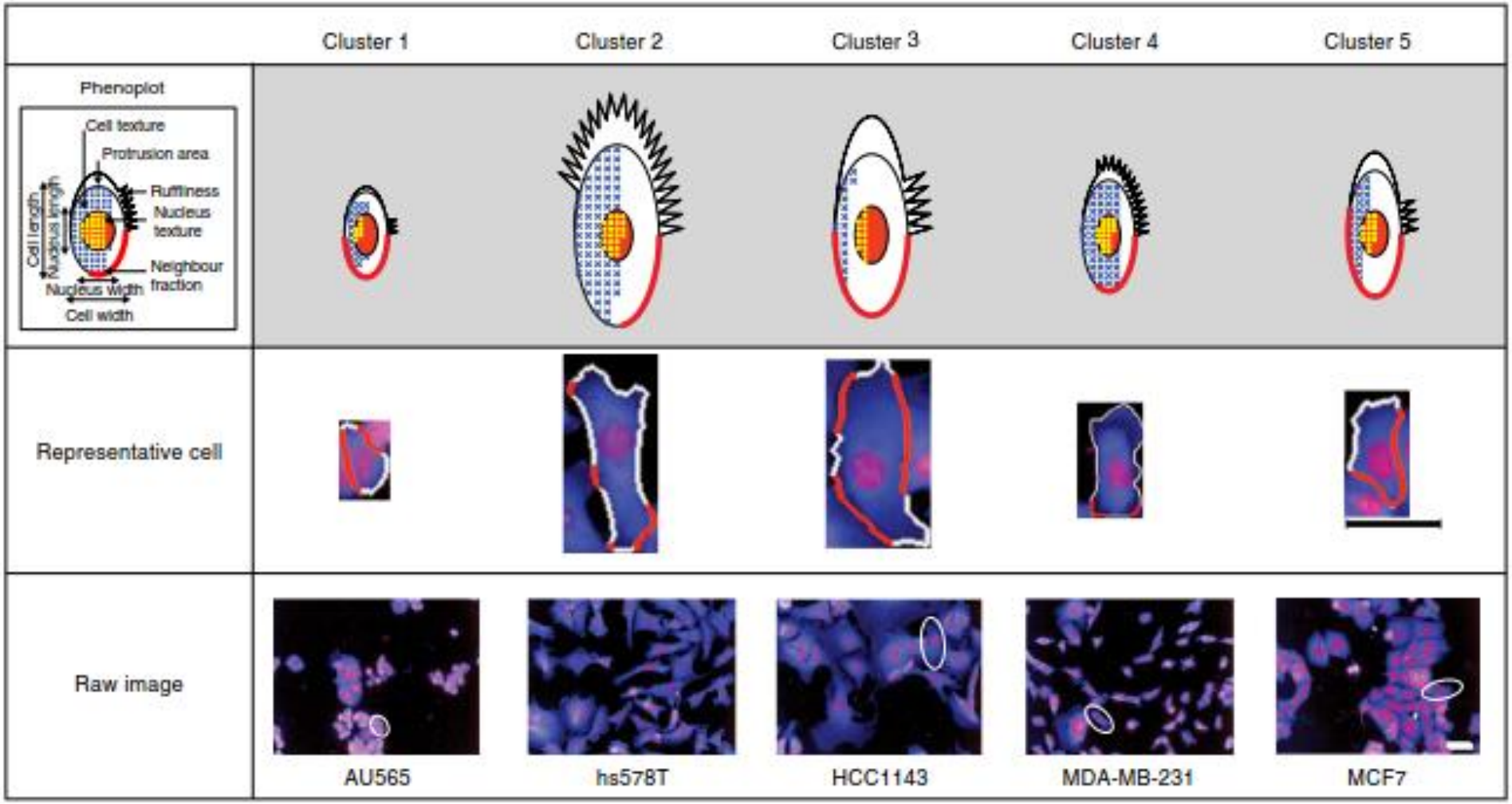
Sailem et al. (2015)

Clusters visualization



Sailem et al. (2015)

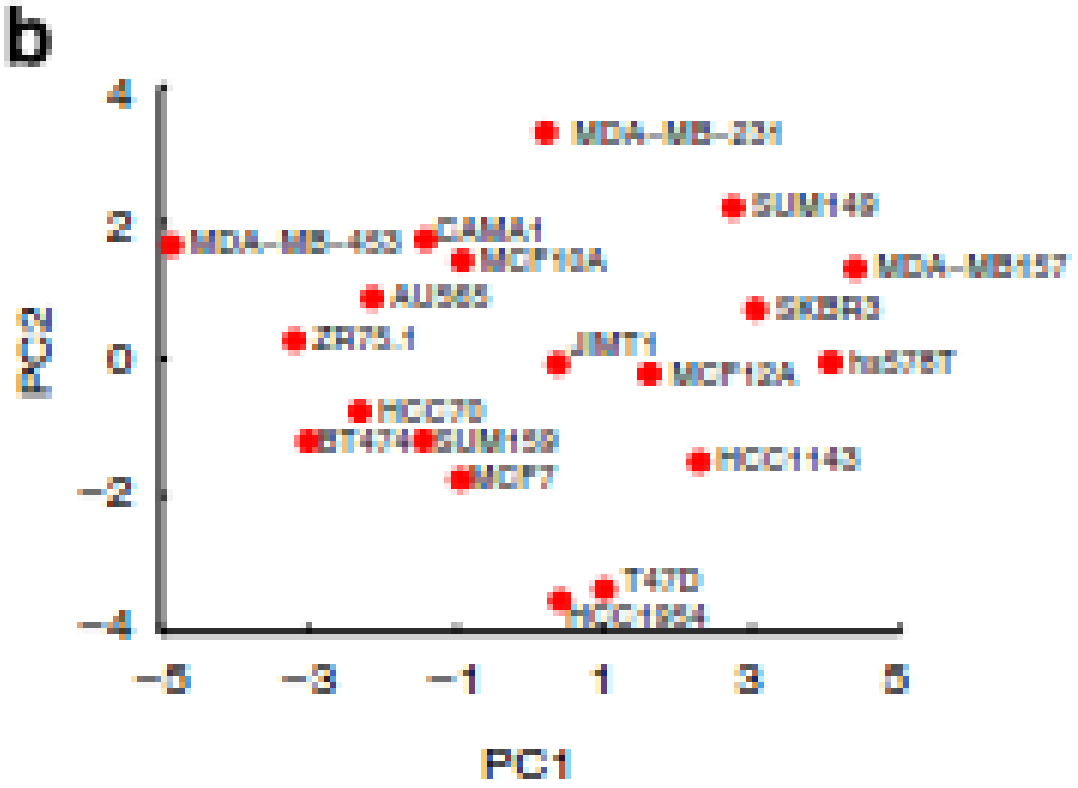
Clusters Visualization



Easier to interpret than “representative” cell

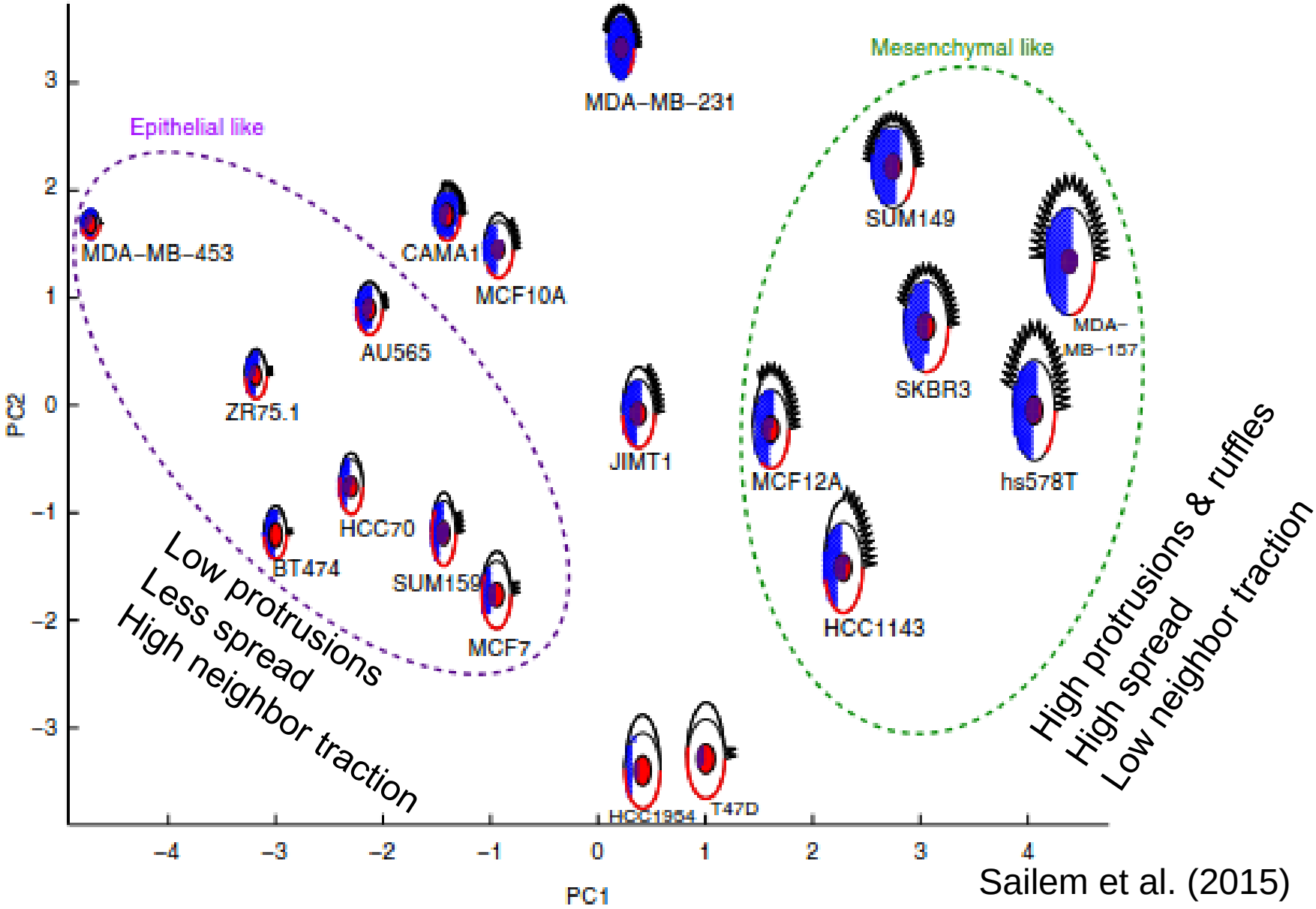
Sailem et al. (2015)

Scatter plot is not informative



Sailem et al. (2015)

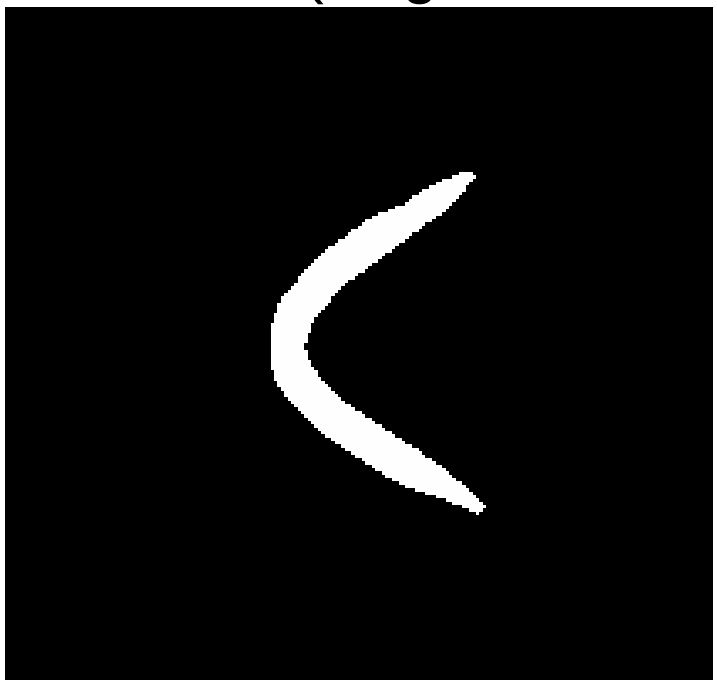
Clusters visualization



Sailem et al. (2015)

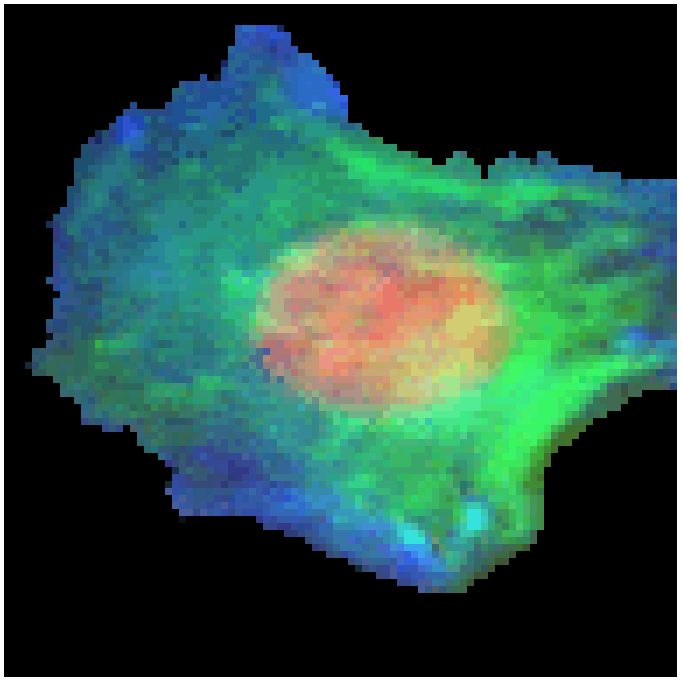
Visual interpretation of complex pheontypes

View examples along a phenotypic continuum (“Eigenworms”)



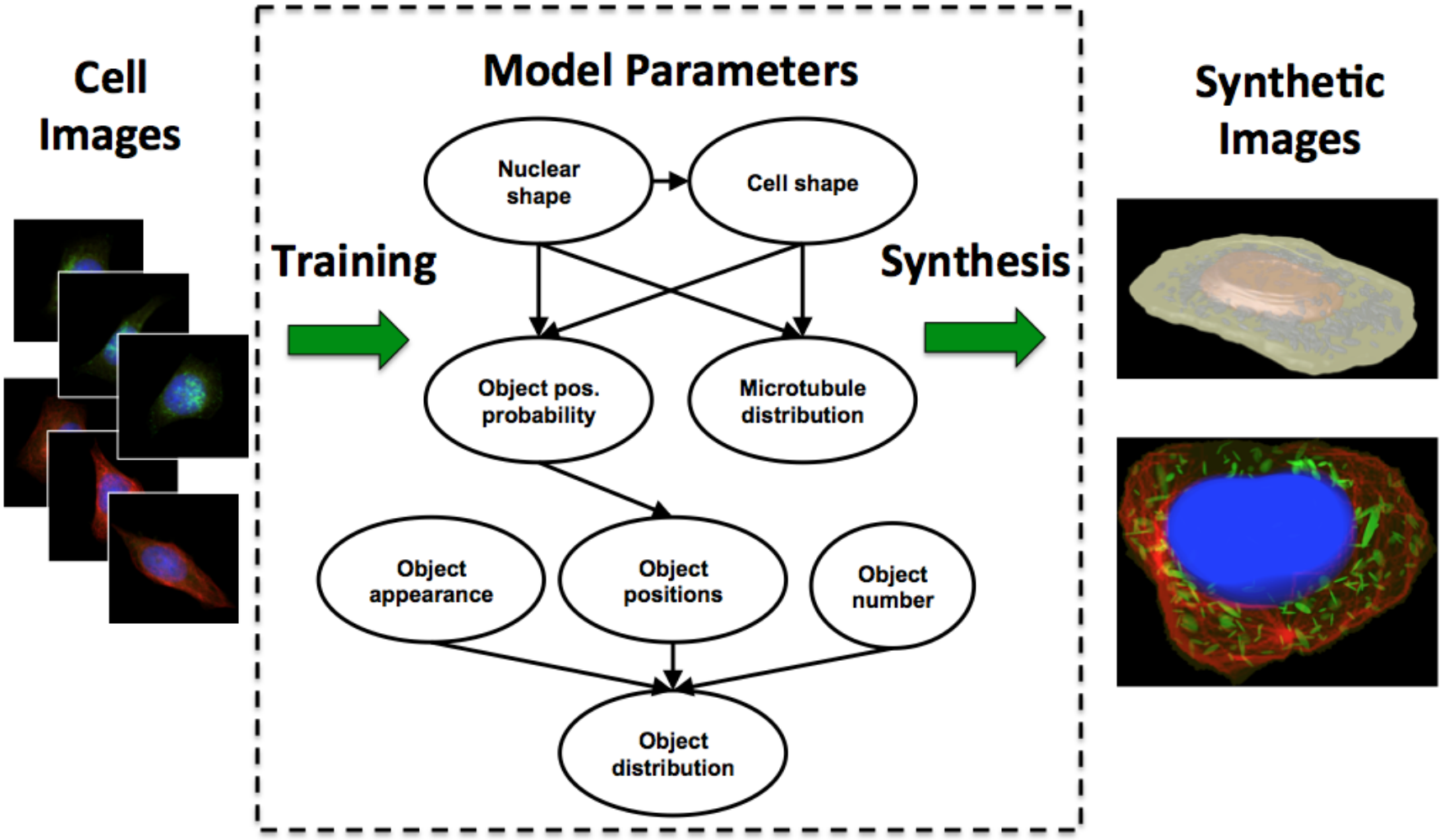
Anne Carpenter,
from Carolina Wahlby

Generative modeling



Goldsborough et al. (2017),
Lafarge et al. (2019)

Generative modeling



Cell Orginizer, Murphy lab

Adversarial Autoencoders

Research project types

- Data mining and integration in public repositories
- Collaborative bioimage analysis projects
- Tool building projects

Open resources for your research projects

- Image data resource (IDR), Williams et al. (2017), Image Data Resource: a bioimage data integration and publication platform https://idr.openmicroscopy.org/
- The Allen Institute of Cell Science
- The Human Protein Atlas
- Branda Andrews datasets http://sites.utoronto.ca/andrewslab/data.shtml

Open resources for your research projects

- Bray et al. (2017), A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. Data, https://github.com/gigascience/paper-bray2017
- Pascual-Vargas et al. (2017), RNAi screens for Rho GTPase regulators of cell shape and YAP/TAZ localisation in triple negative breast cancer Data via IDR
- Pizzagalli et al (2018), Leukocyte Tracking Database, a collection of immune cell tracks from intravital 2-photon microscopy videos (via figshare)