



Automating Biomedical Research Through Machine Learning

Robert F Murphy

Ray & Stephanie Lane Professor of Computational Biology and
Professor of Biological Sciences, Biomedical Engineering and Machine Learning
Affiliated Senior Fellow, Freiburg Institute for Advanced Studies
Honorary Professor, Faculty of Biology, University of Freiburg, Germany

RAY AND STEPHANIE LANE
Center for Computational Biology

Carnegie Mellon

Why Comp Bio?

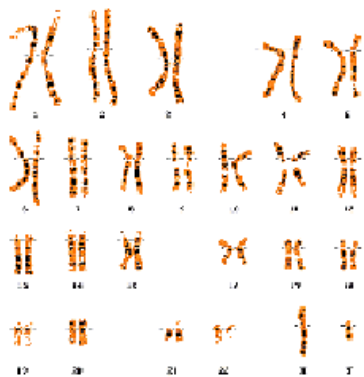
Personal Genomics Revolution

Personal Genome Company 23andMe Receives Google Investment

by Dr. Hsien-Hsien Lei

Posted May 22, 2007 in [DNA Products](#), [DNA in General](#)

[Email this](#) • [Stumble It!](#) • [Digg This!](#) (10 Diggs, 1 comment) • [Discuss on Newsvine](#)



Google has invested \$3.9 million in 23andMe along with Genentech and two venture capital firms, New Enterprise Associates and Mohr Davidov Ventures. 23andMe is a personal genome company that wants to help consumers get in tune with their genes. Check out the blurb on the newly revamped front page:

Even though your body contains

Wednesday, January 11, 2012

Ion Torrent claims to be first with \$1K genome sequencer

By Lori Valigra



Jonathan Rothberg, founder, Ion Torrent Systems

Other matches for "":

[More Search Results](#)

Ion Torrent Systems Inc. of Guilford, Conn., showed off its [new DNA sequencing machine and chip](#) that it claims can map a human genome in 24 hours for a cost of \$1,000 at the J.P. Morgan Healthcare Conference in San Francisco Tuesday, according to a news release from parent company Life Technologies Corp.

The same day, Illumina Inc. of San Diego also announced a [24-day sequencer](#), but at a much higher initial machine price tag of \$740,000. The price to sequence a genome was not disclosed.

Ion Torrent last summer wrote in a Nature paper that it had developed a DNA sequencing technique to [target the \\$1,000 genome industry goal](#). At the time the paper was published, Ion Torrent CEO and founder Jonathan Rothberg, told Mass High Tech, "Sequencing on an ion semiconductor chip makes the \$1,000 genome both inevitable and predictable. Extrapolating from our current progress we will break the \$1,000 genome barrier in 2013." The company appears to be ahead of schedule.

<http://www.eyendna.com/2007/05/22/personal-genome-company-23andme-receives-google-investment>

<http://www.bizjournals.com/boston/blog/mass-high-tech/2012/01/ion-torrent-claims-to-be-first-with-1k-genome.html>

Why Comp Bio?

Cancer Diagnosis and Treatment

Agendia Gets FDA Clearance for MammaPrint on Agilent Systems

February 23, 2011



By a [GenomeWeb staff reporter](#)

NEW YORK (GenomeWeb News) – Agendia has received another clearance for the firm's MammaPrint breast cancer recurrence assay.

The new clearance is for running the test on two additional Agilent microarray scanners and two Agilent Bioanalyzers. Agendia said the additional clearance will expand laboratory capacity to handle the increasing number of MammaPrint, TargetPrint, and Blueprint test results.

Previously, the MammaPrint test was only allowed to be performed at Agendia's lab in Irvine, Calif. The new clearance enables the test to be run in a central laboratory and allows Agendia to also perform the test at its lab in Amsterdam, where the firm is based.

MammaPrint was the first *in vitro* diagnostic multivariate index assay device [approved by the FDA](#) in early 2007.

<http://www.genomeweb.com/arrays/agendia-gets-fda-clearance-mammaprint-agilent-systems>

Type size:

Email

Printer-friendly version

RSS Feed



RSS-Feeds | [About us](#) | [Advisory Board](#) | [Sponsorship and Donations](#)

[Home](#) | [Science News](#) | [Events](#) | [Groups](#) | [Jobs](#) | [Videos](#) | [Forum](#)

Published on 2 February 2010, 07:02



Categories: [Cancer Stem Cells](#) [Computational Biology](#) [Computer Simulations](#) [Oncology](#)

Computer Simulations Reveal How Tumour Growth is Controlled

Computational scientists and oncologists working for the University of Amsterdam (UvA) have discovered that cancer stem cells control how tumours grow. The researchers also determined the underlying mechanism. The study, which used computer simulations and laboratory experiments, may help improve the treatment of cancer patients. The results were published earlier this month in *Cancer Research*, a leading journal published by the American Association for Cancer Research.



Professor of Computational Science Peter Slood, of the Faculty of Science, and Louis Vermeulen, a PhD student in the field of Experimental Oncology at the UvA's Academic Medical Centre, headed up a multidisciplinary group of researchers. The discovery that cancer stem cells control invasive growth and the identification of the underlying mechanism are particularly significant because that growth is one of the first stages in the manifestation of local metastases.

Why Comp Bio?

Develop better drugs

Science, Technology and Innovation

Projects

Subscribe Enews Alert

Search...

HOME MAGAZINES R&D PROJECTS FEATURES **OPINION** NEWS EVENTS WORLD CITIES SU

Categories

- Agriculture/Food
- Biology/Medicine
- Energy
- Environment/Climate
- ICT
- Industry/Technology
- Society/Economy
- Transport/Construction

Ahead in the cloud - computational drug discovery reaches for the next level

Cloud computing offers the promise of a flexible computing platform for drug discovery researchers, enabling access to technologies previously unavailable, unaffordable or at a scale previously unobtainable.

Cloud computing – the general term used for delivering computational services over the [internet](#) – resembles in some respects the longer established grid computing paradigm. Grid computing divides a task into many smaller tasks which are distributed among a large number of, usually, low power computers (for a life sciences example see the [Screensaver Lifesaver project](#)). However, setting up a grid computing infrastructure requires significant



twitter

ProjectsZine: New plant extracts could help in the fight against #obesity

http://www.projectsmagazine.eu.com/opinion/ahead_in_the_cloud_computational_drug_discovery_reaches_for_the_next_level

Why Comp Bio?

Solve the energy crisis with microorganisms

WSJ BLOGS

Environmental Capital

Daily analysis of the business of the environment by The Wall Street Journal.

JULY 14, 2009, 11:14 AM ET

Biofuels Bonanza: Exxon, Venter to Team up on Algae

Article

Comments (15)



Email



Print



Like



Send



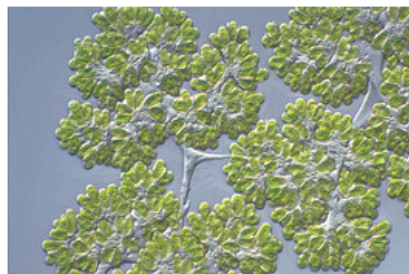
+ More



Text

By Keith Johnson

Exxon's [\\$600 million deal](#) with Synthetic Genomics to brew fuel from algae could mark a coming of age for alternative fuels.



National Institute of Environmental Studies

Put some of this pond scum in your tank.

As human genome mapper Craig Venter, now the chief executive of SGI said, turning algae into biofuel simply won't happen without Big Oil's deep pockets. He said it was the biggest program he knows of worldwide to produce biofuels.

At the same time, both Exxon and SGI were at pains to stress that the "collaborative research project" is in its early phases. "There's no guarantee of success," said Exxon's vice-president for R&D, Emil Jacobs. "We're giving a new reality to the timelines" of algae biofuel

development, Mr. Venter said.

Why Comp Bio?

Get a great job:

Future Forecast: 10 Hot Job Markets in 2012



In our information-rich society there is an ever increasing demand for workers in the fields of computers, health care, science and space technology—much of it driven by the demands of the retiring baby boomers. If you like to plan ahead, here is sampling of some of the jobs that will be hot in the next several years and beyond.

1) Organic food Industry

By 2010, organic food and beverage will represent about 10 percent of the total market — a tenfold increase from 1998. Bob Scowcroft, executive director of the Organic Farming Research Foundation says the industry will soon need more organic food producers, certification experts, retailers and scientists as organic becomes mainstream.

Qualifications: Organic food expertise in farming, business or science.

Salary range: \$50,000 to \$80,000

2) Computational Biology

There is a growing need to combine computer science, biology and math to make sense of research data in massive quantities, says Leroy Hood, co-founder of the Institute for Systems Biology. This field may eventually allow physicians to test for a patient's unique genetic markers and tailor the best treatments and medicine for that patient.

Qualifications: A bachelor's degree or higher in bioinformatics, computer science, mathematics,

http://www.greatnewsnetwork.org/index.php/news/article/future_forecast_10_hot_job_markets_in_2012/

Jobs in Computational Biology: Career Options and Requirements

The study of biology using computer software and mathematics is called computational biology. Jobs in the field typically require a high level of integrated education in biology, mathematics and computer science. A variety of organizations employ people in this application-oriented field.



Career Options

This multidisciplinary specialty has a number of practical and research applications. Computational biologists analyze large

volumes of information and devise computer modeling simulations for academic research and health applications for businesses such as biotechnology and pharmaceutical companies, as well as for government health and research institutions.

Career Options in Academia

Computational biologists in academia can work with other theoretical and applied researchers to devise models, simulations and predictions for molecular biological systems and interactions. They can also analyze large quantities of data related to genetics and genomics. Academic researchers typically teach courses in their field and supervise student research projects in addition to conducting their own research.

Academic Career Requirements

Careers in academia commonly require a graduate degree in computational biology. To obtain a faculty position at a university, a Ph.D. is recommended, though a master's degree is typically sufficient to teach at a 2-year college. In addition, working in a postdoctoral position for several years after graduation is usually necessary before a permanent faculty or research position can be obtained.

Career Options in Commercial Industries

Pharmaceutical companies, scientific software companies and biotechnology companies all employ computational biologists in research and development. Modeling drug interactions, developing analytical software for use by biologists and analyzing large data sets for biotech companies that provide genetic

http://education-portal.com/articles/Jobs_in_Computational_Biology_Career_Options_and_Requirements.html



Examples of Computational Biology Problems

- Drug Development
- Cell Modeling

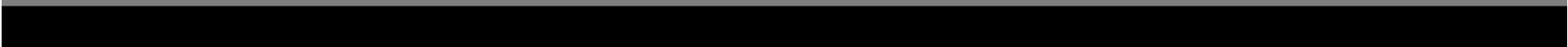


Observation, Modeling and Computation

- Traditional Approach (Scientific Method):
Make observations, construct model/
theory, make new predictions, test them
 - Good news
 - Computers can help build models
 - Bad news
 - Difficult/impossible to prove biological theories directly
 - Lots of possible theories and predictions to test!

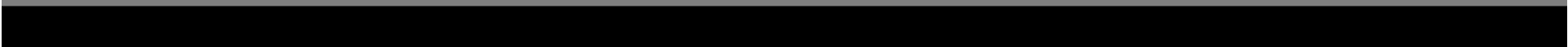


Problem – Cell/Developmental Biology

- Tens of thousands of molecules within cells can change on time scales from below seconds to months during differentiation or disease processes
 - Can undergo changes in expression, interaction, localization
 - Too many combinations to measure all...
- 

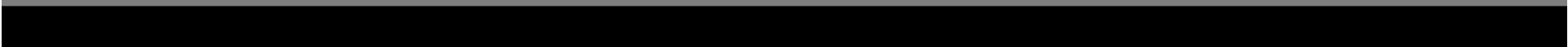


Problem – Drug Development

- Drugs fail late in development because of unanticipated side effects – only real solution is to choose drugs with desired effect on target and no undesired effect
 - This requires determining the effects of millions of potential drugs on tens of thousands of potential targets
 - Exhaustive experimentation too costly...
- 



Solution?

- Represent problems as a matrix of possible experiments
 - Build model to predict full matrix from whatever data we have
 - Use active learning to choose new experiments and iteratively improve the model
- 

Traditional drug development

A controlled experiment that measures the effect of a compound on a target's activity



Typically, a protein whose activity is believed to cause a disease



Negative



Positive



Measured



Unmeasured

Traditional drug development

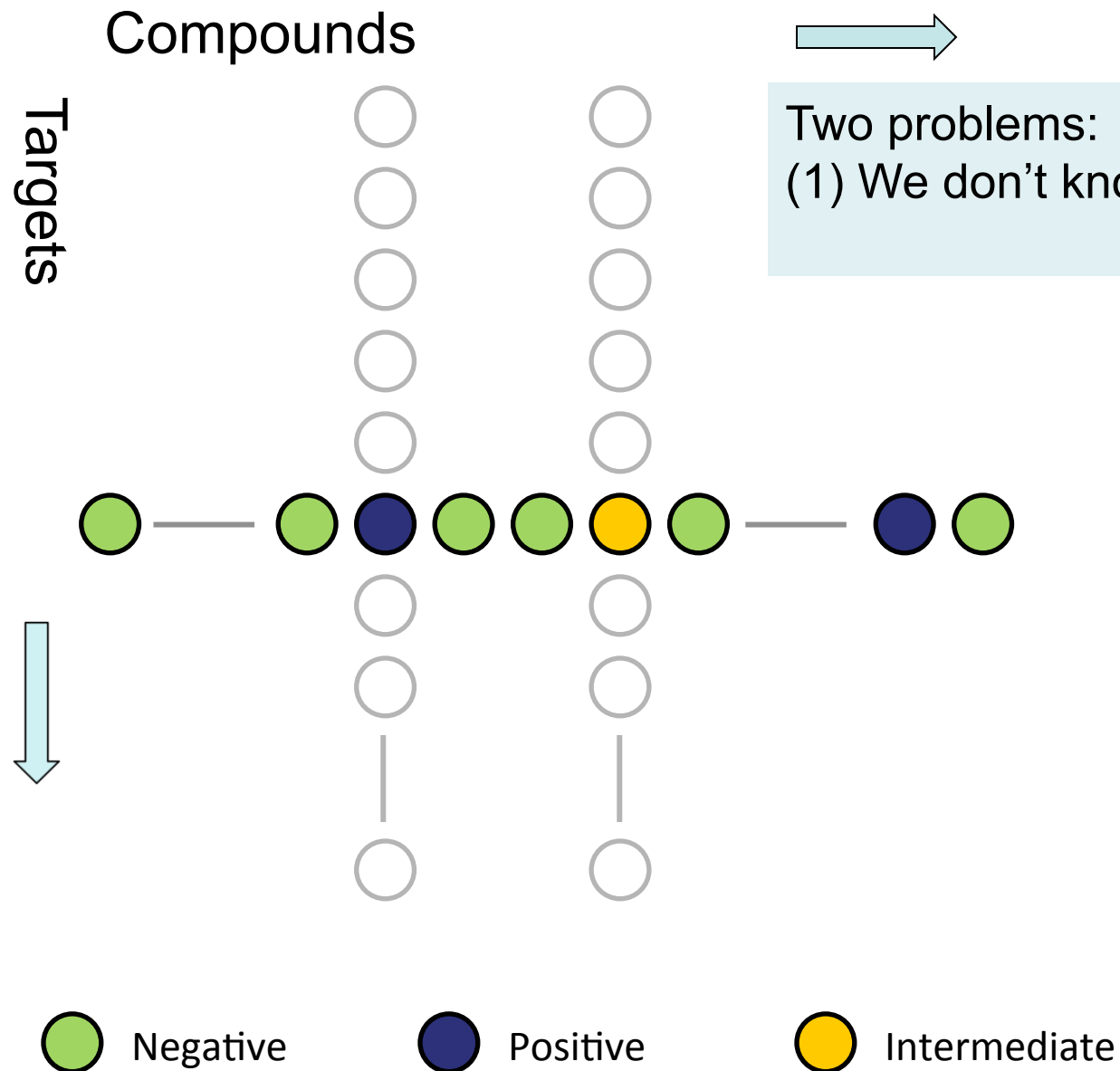
Chemicals that are potential drugs or might be modified to become drugs

Compounds

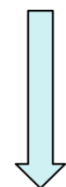


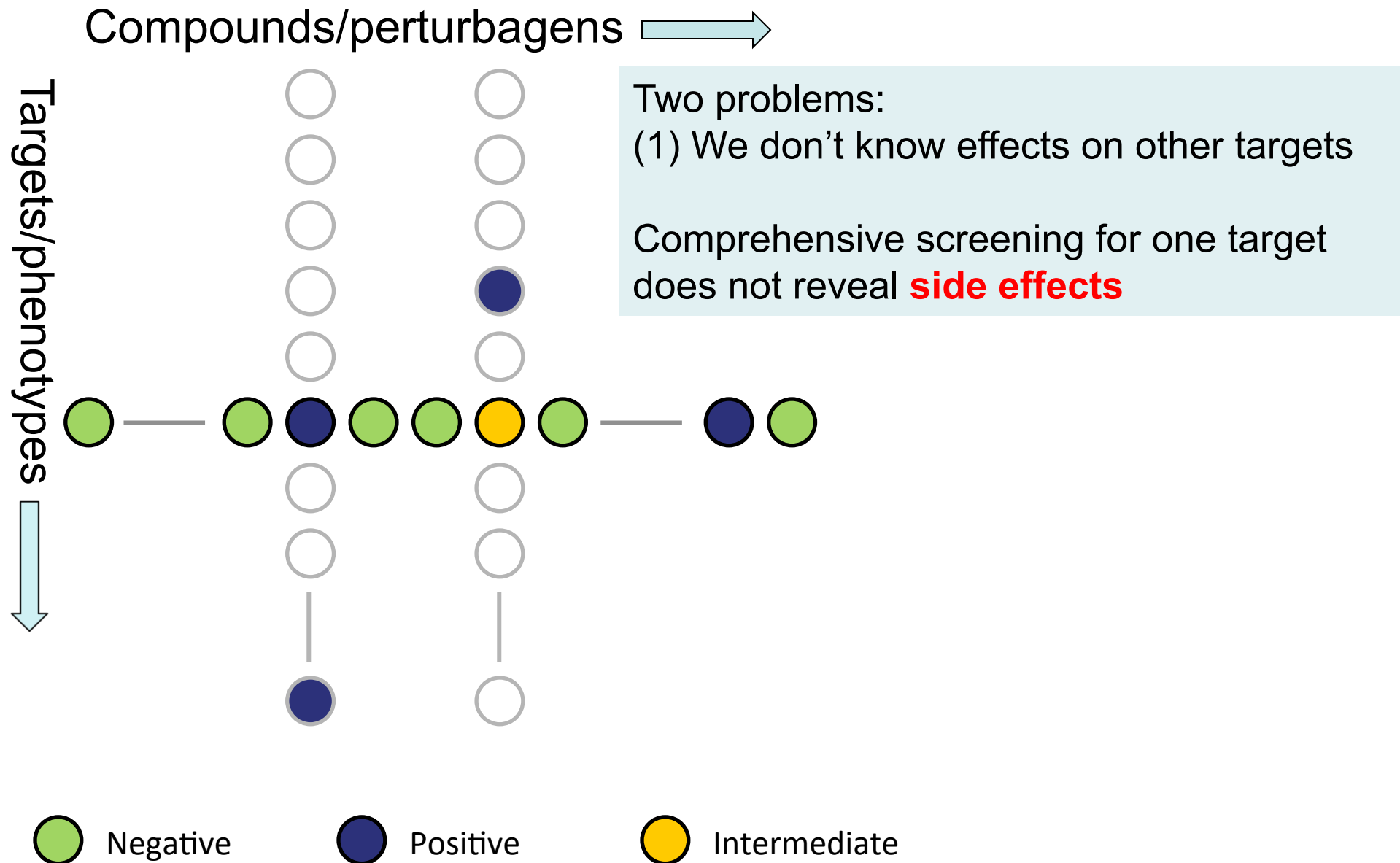
 Negative

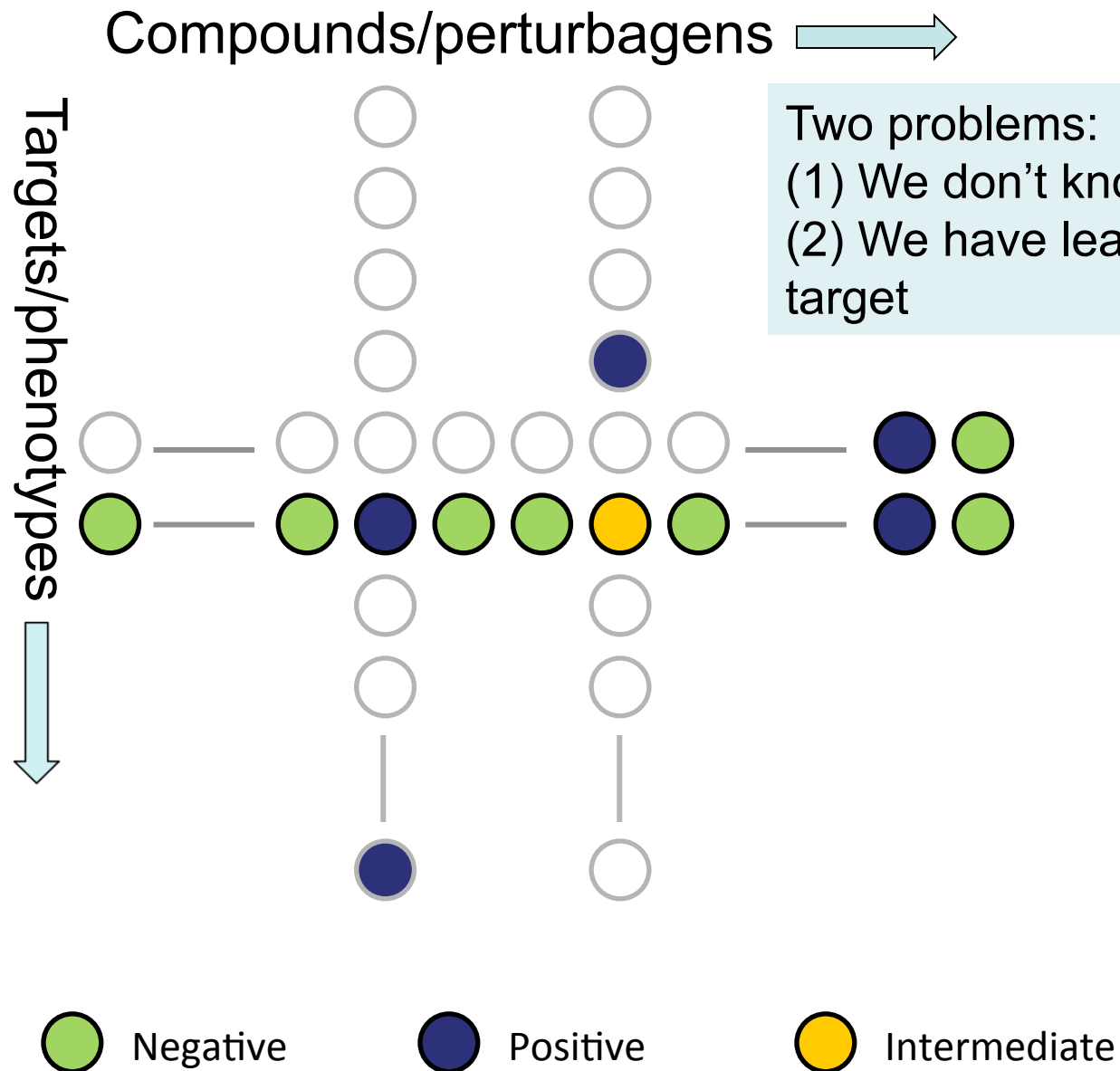
 Positive



Two problems:
(1) We don't know effects on other targets



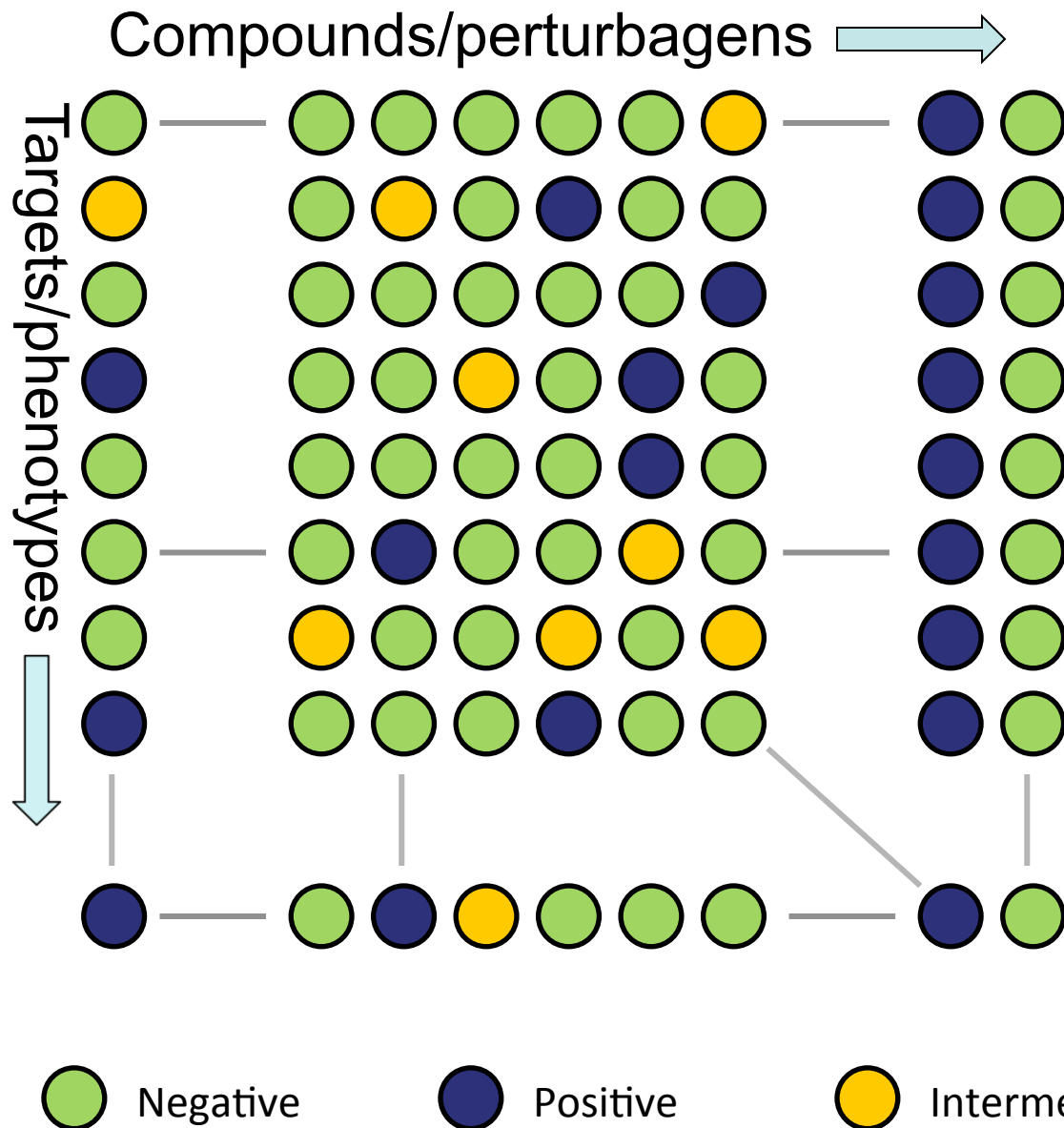




Two problems:

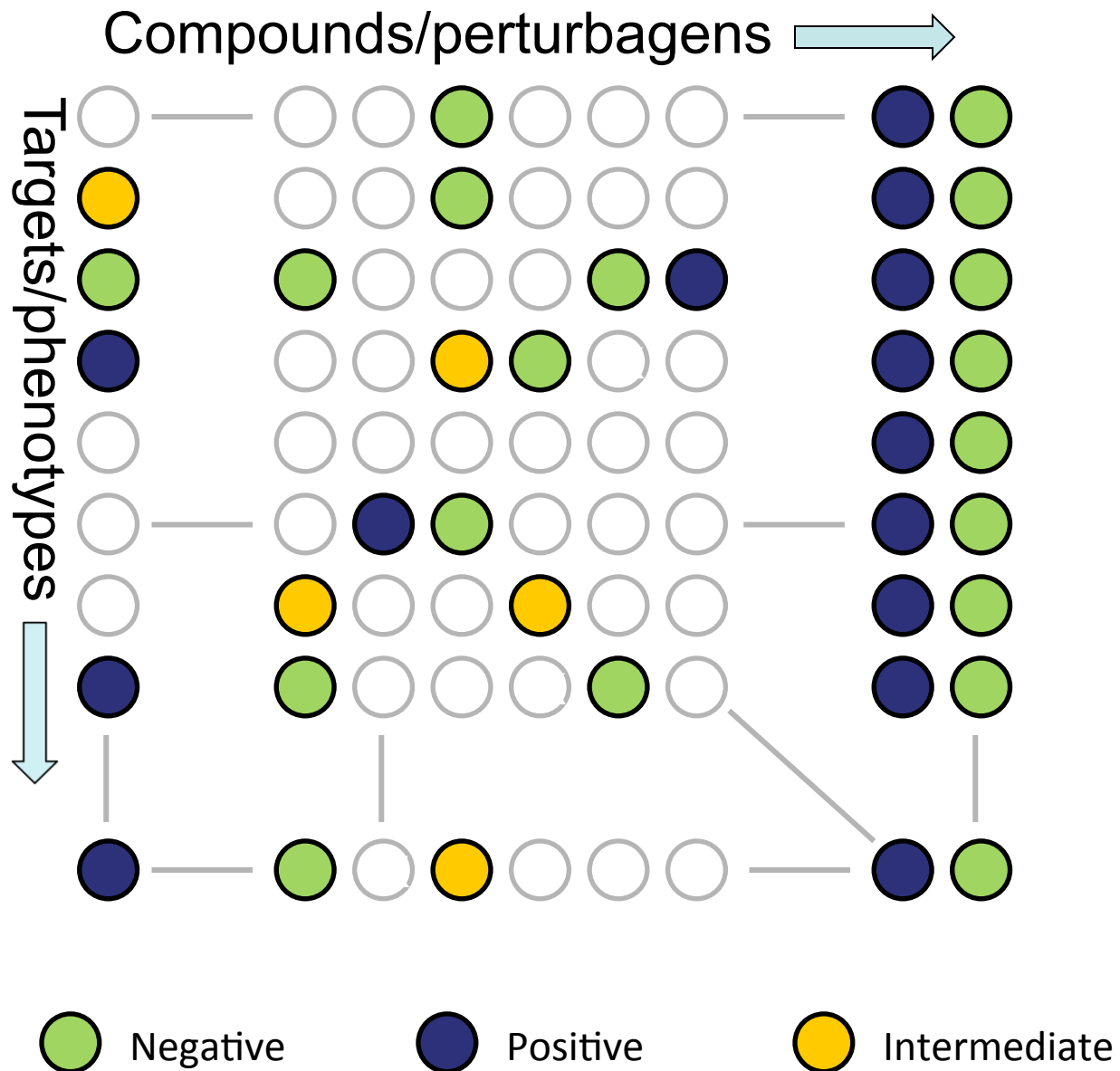
- (1) We don't know effects on other targets
- (2) We have learned nothing for the next target



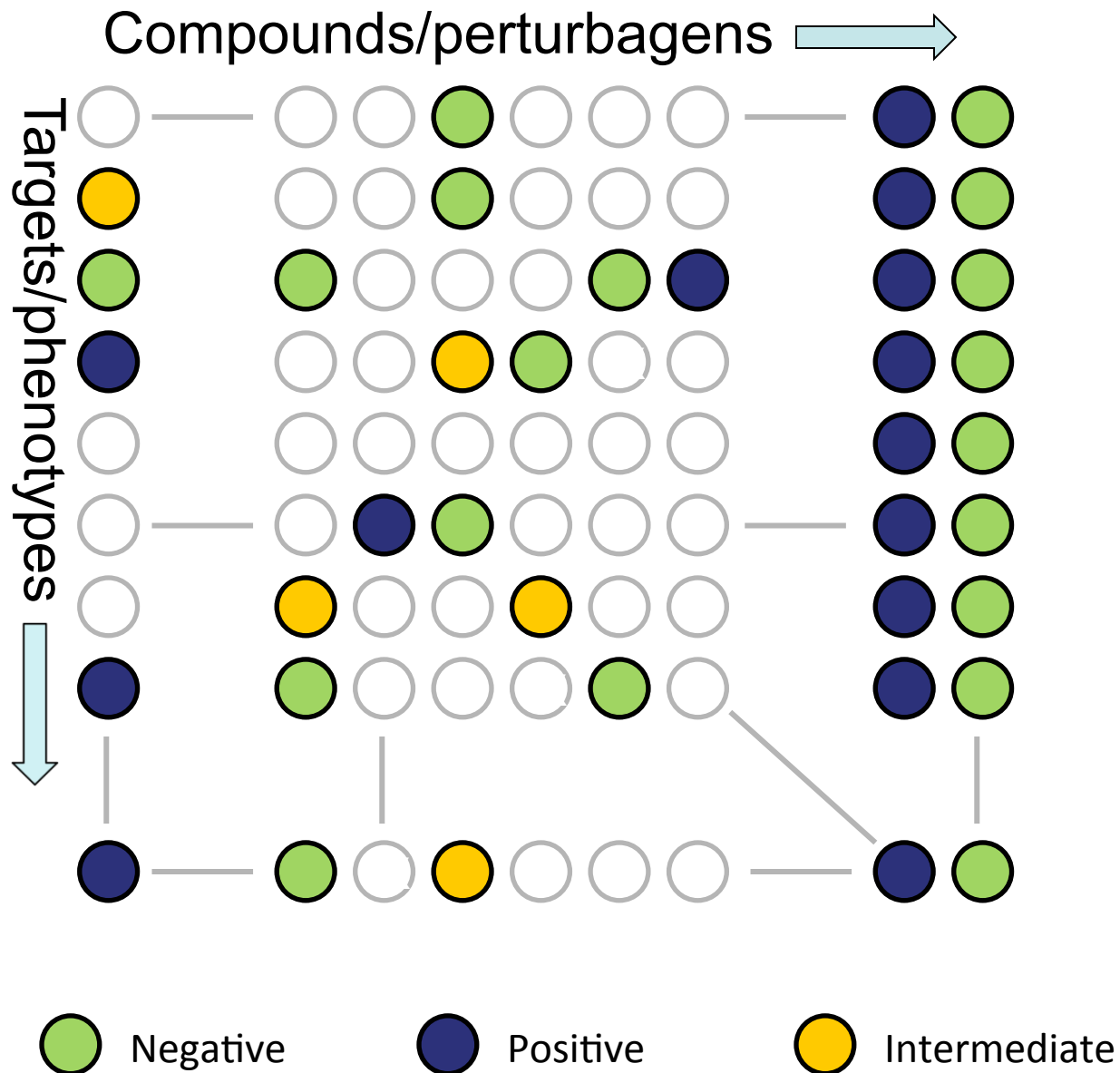


Much better if we could learn the entire matrix...

but we cannot afford to exhaustively perform every experiment



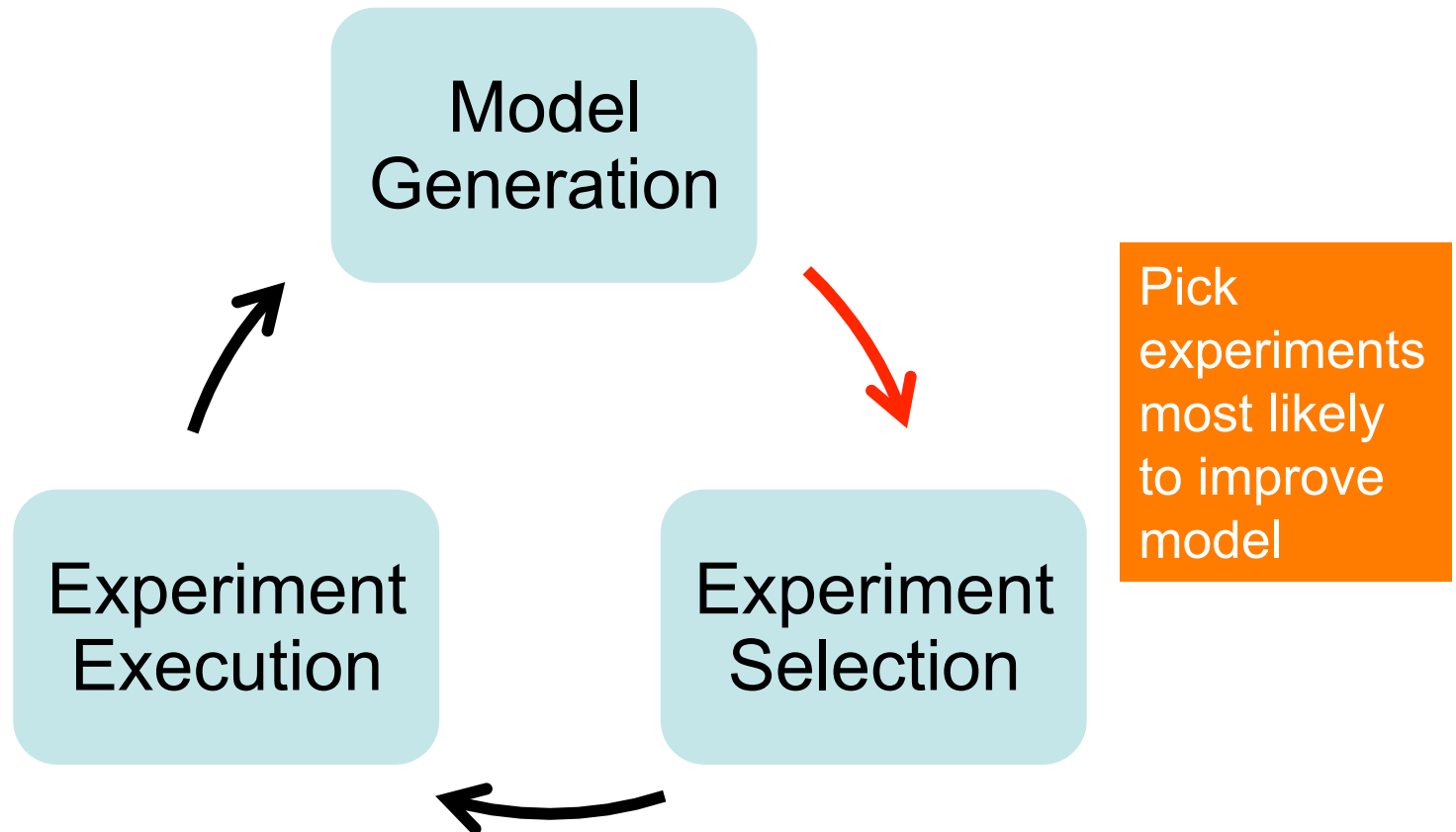
Solution: just do some experiments...



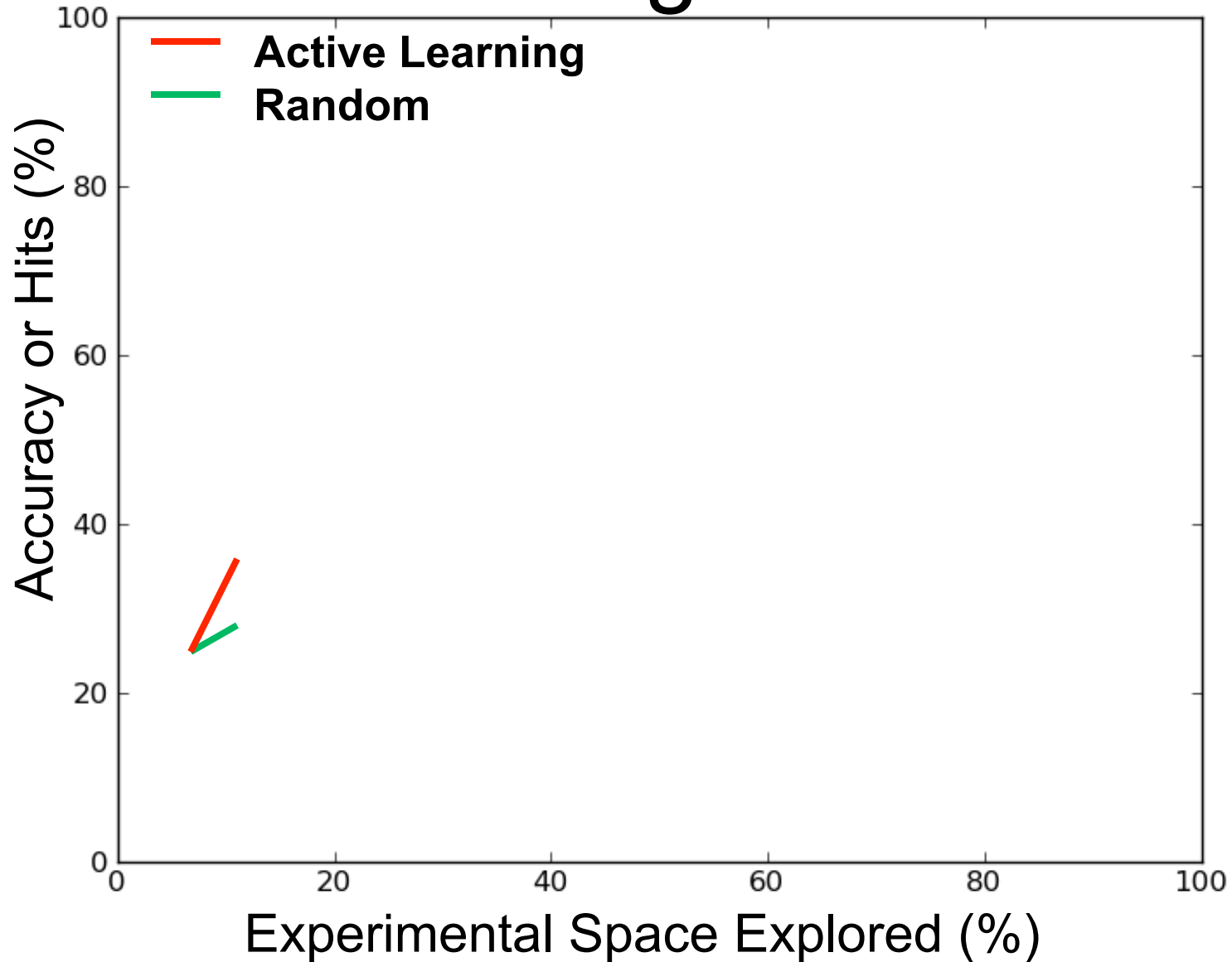
Solution: just do some experiments... and predict the rest

Active Learning Pipeline

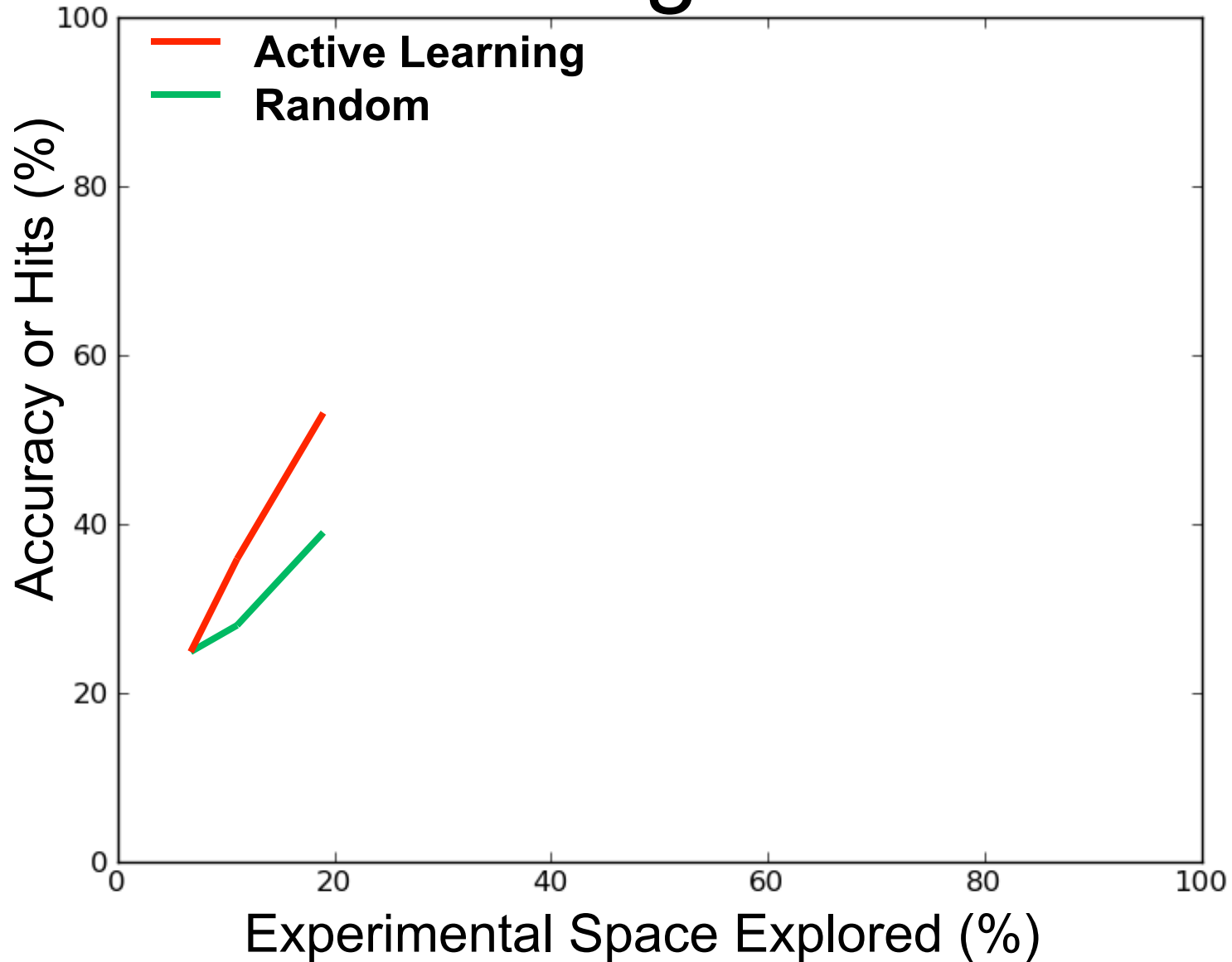
- Efficiently learn accurate predictive model



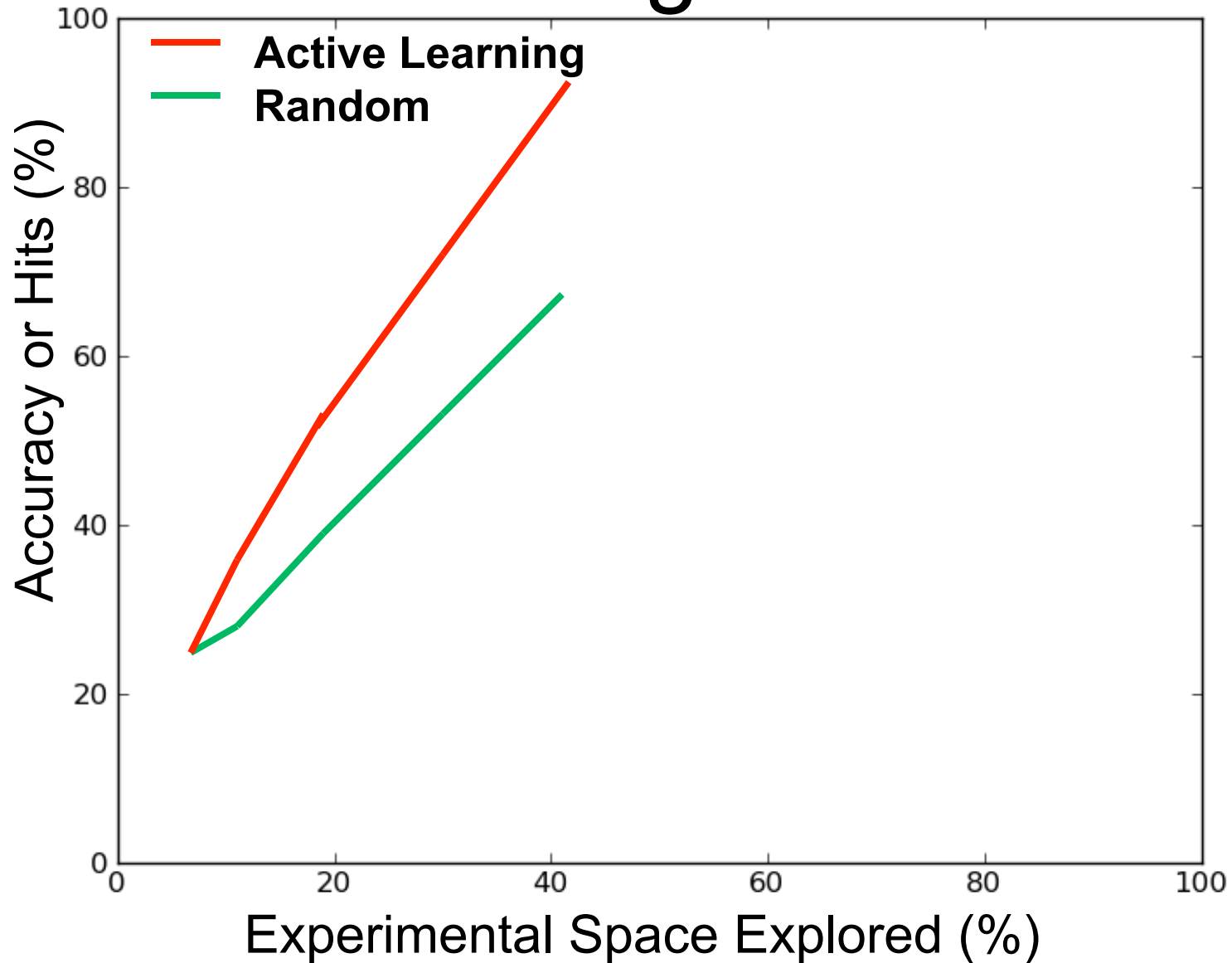
Active Learning Assessment



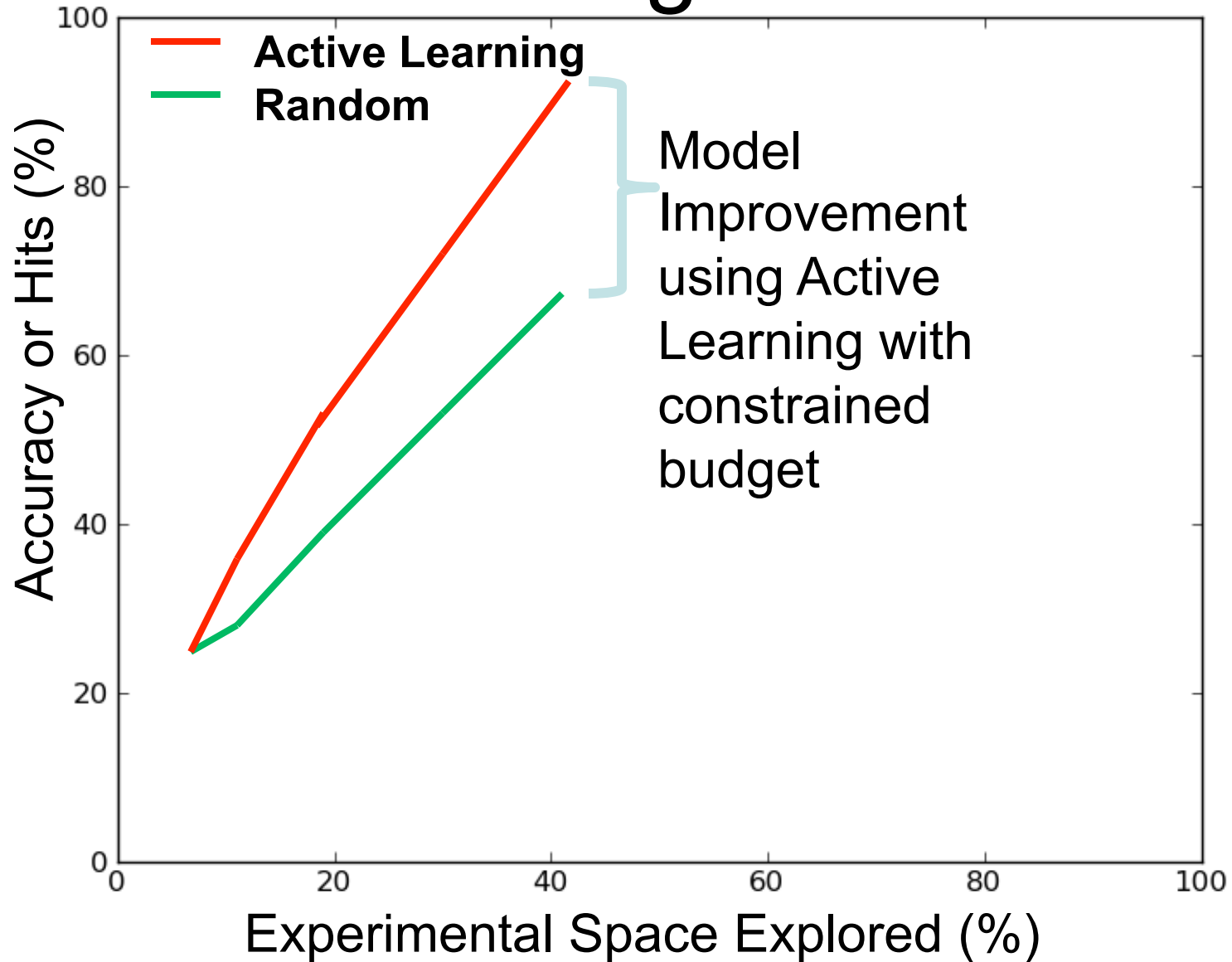
Active Learning Assessment



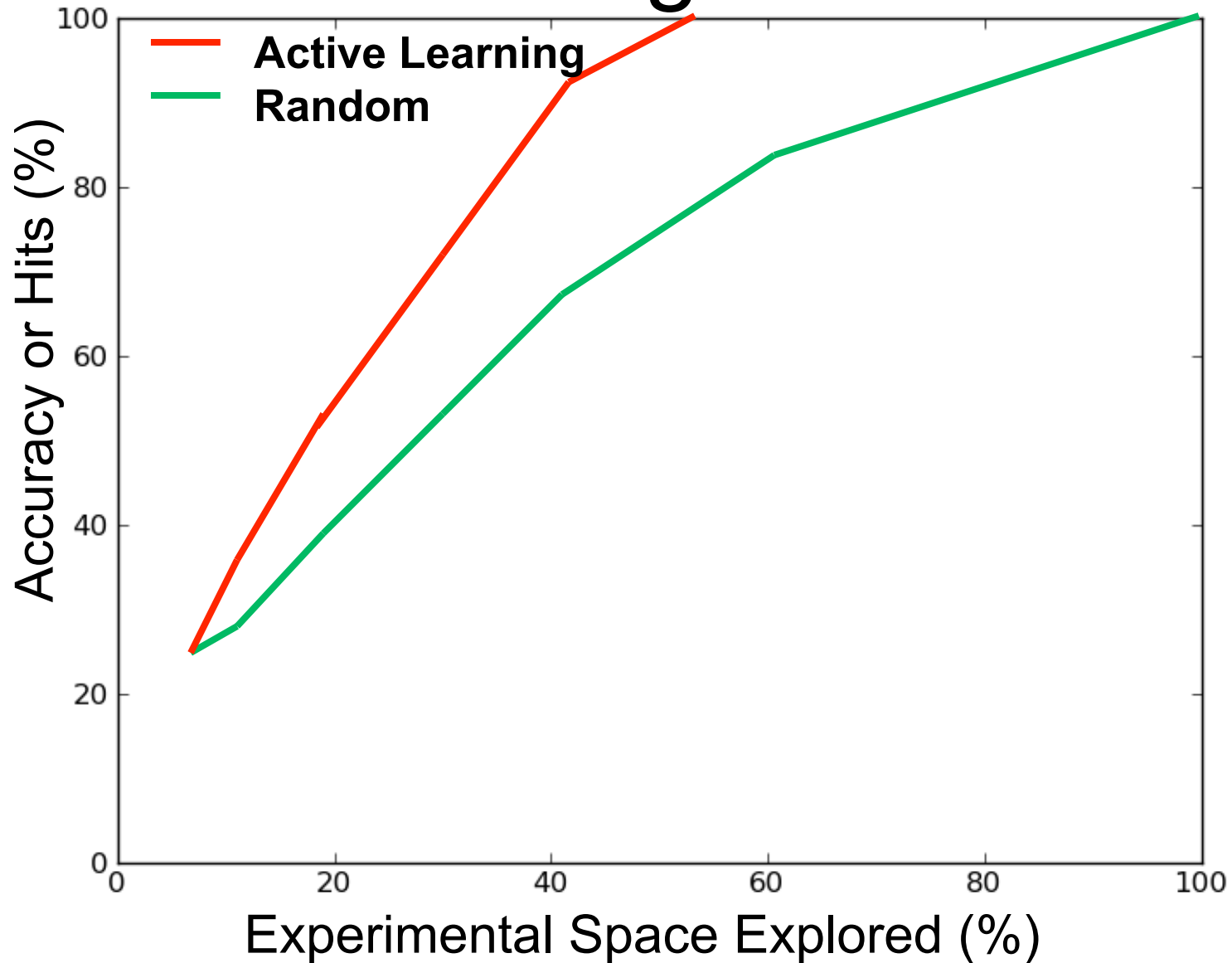
Active Learning Assessment



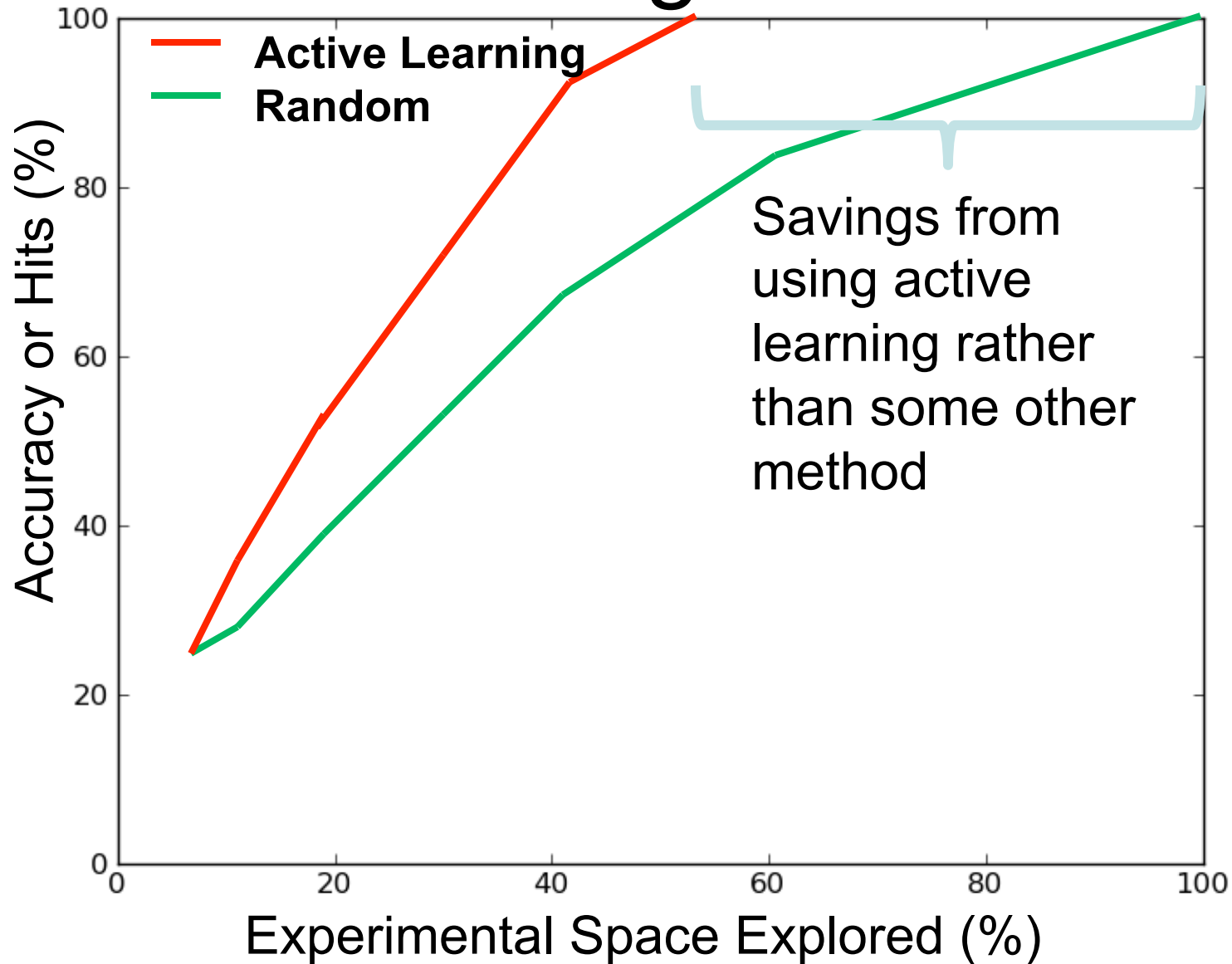
Active Learning Assessment



Active Learning Assessment

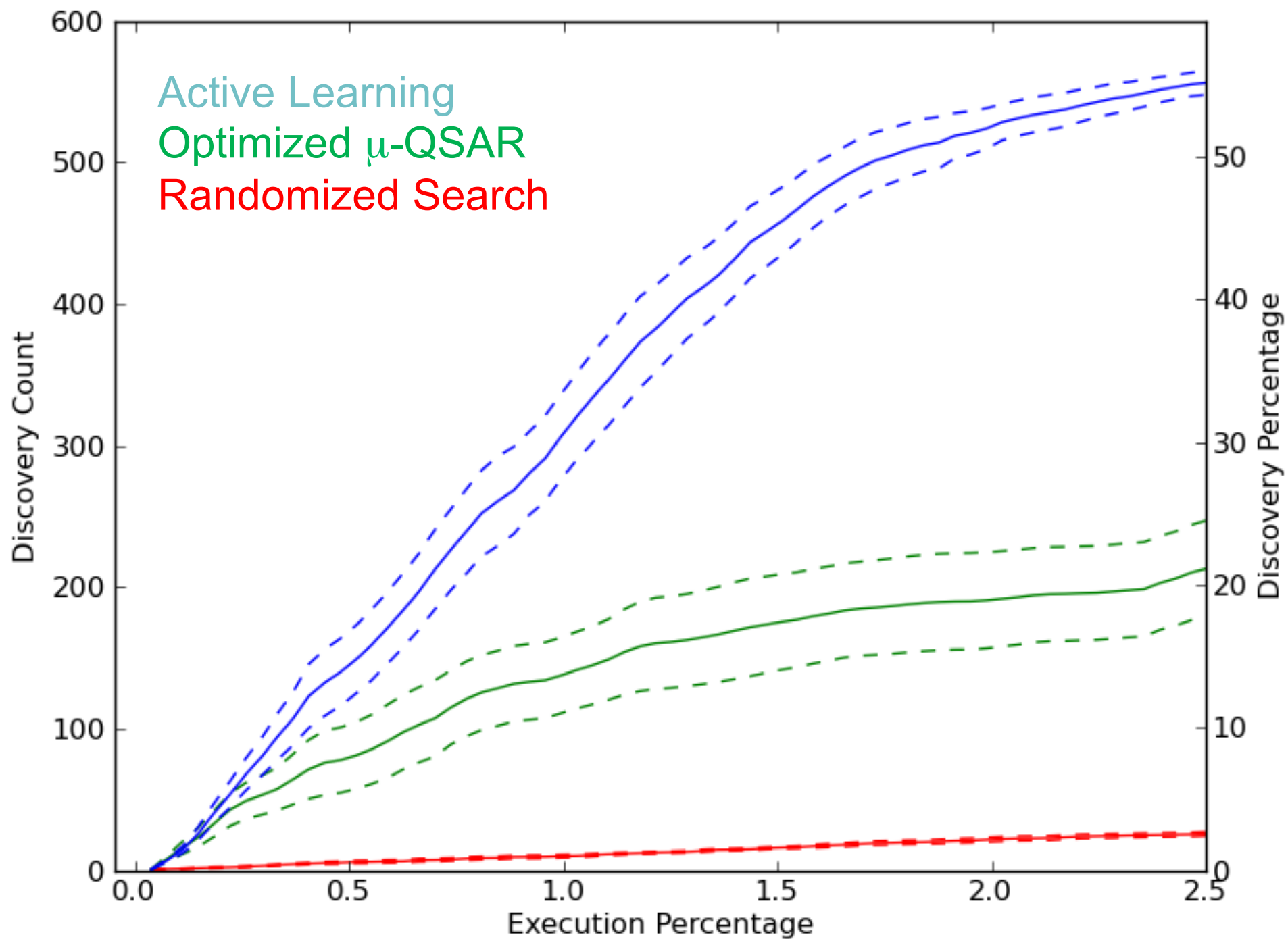


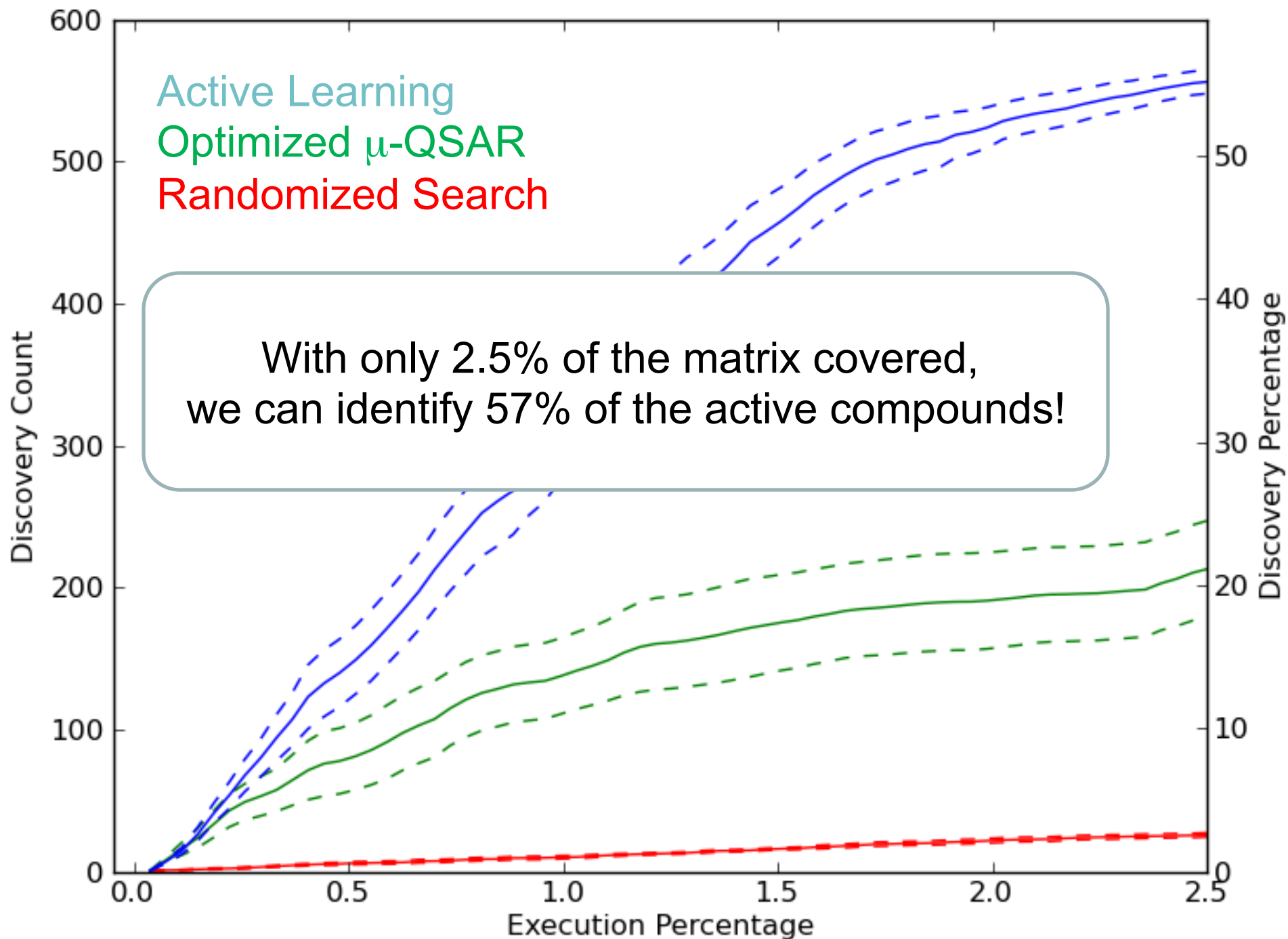
Active Learning Assessment



Test with PubChem Data

- Assays: 177
 - 108 *in vitro*
 - 69 *in vivo*
 - Sign of score modified to reflect type of assay (inhibition or activation)
- Unique Protein Targets: 133
- Compounds: 20,000
- Experiments: ~1,000,000 (30% coverage)
- Compare discovery rate across different methods
 - Discovery: a drug-protein pair whose $|\text{rank score}| > 80$





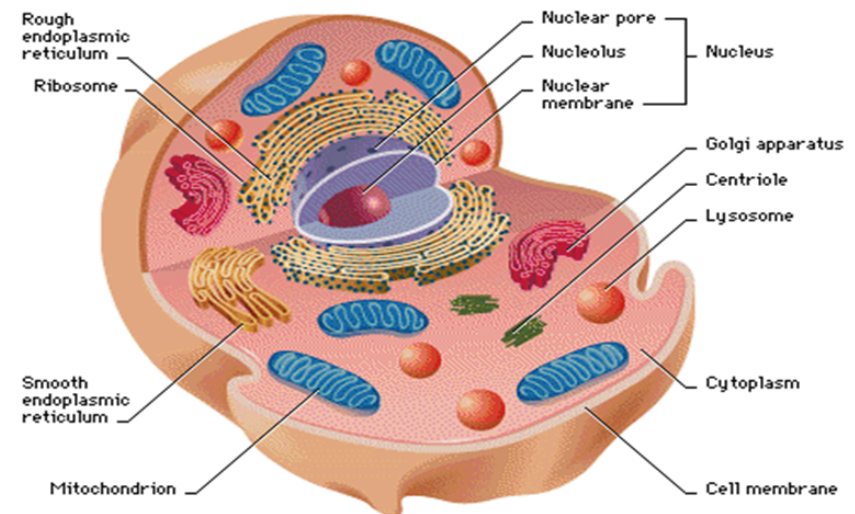


BUILDING CELL MODELS FROM IMAGES



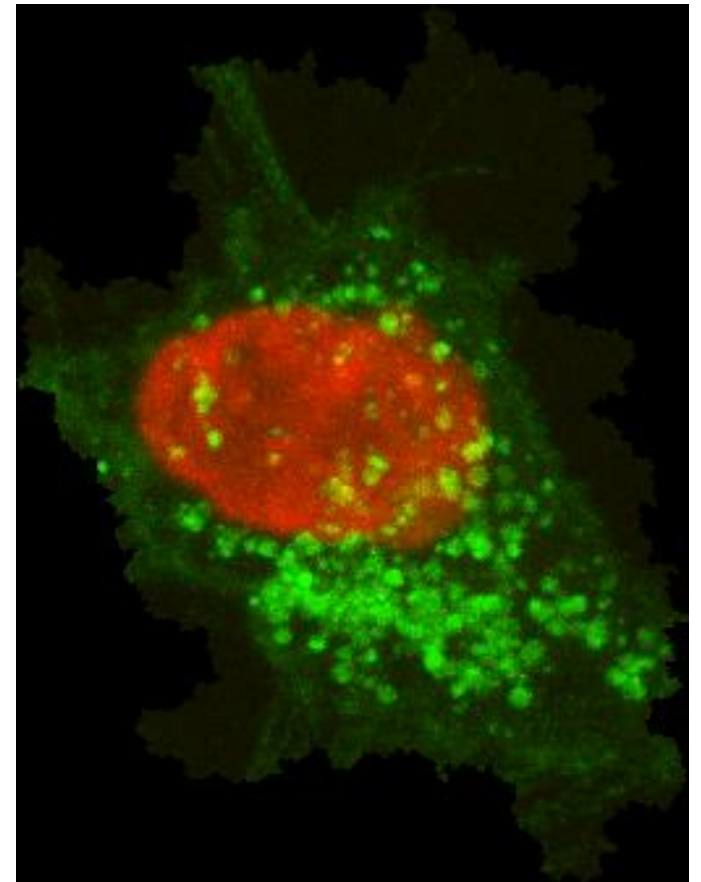
Cell Organization

- How do we learn and represent
 - the sizes and shapes of different cell types
 - the number, sizes, shapes, positions of subcellular structures
 - the distribution of proteins across those structures?
 - how these change in the presence of perturbagens?

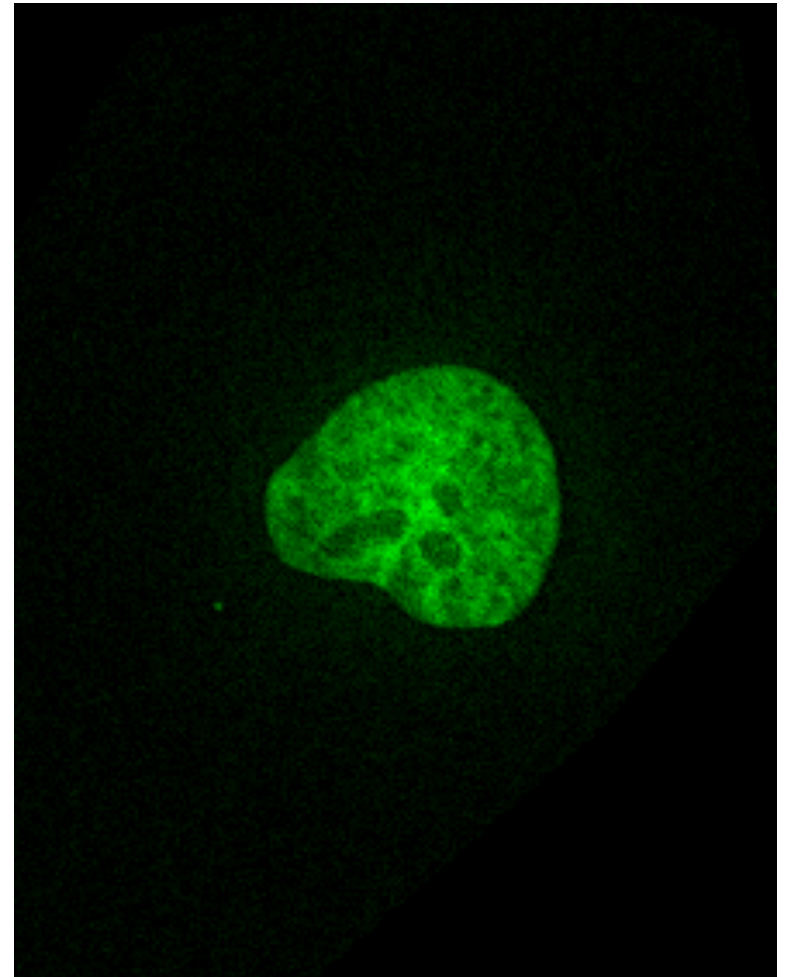
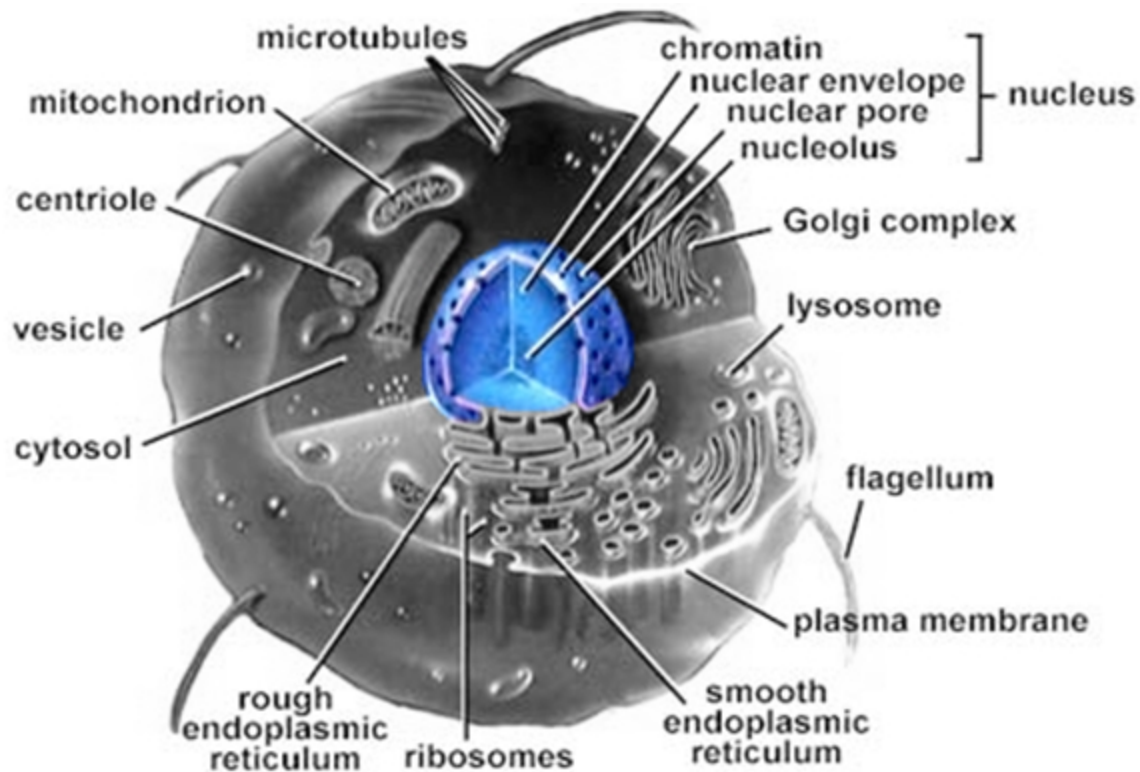


Fluorescence microscopy

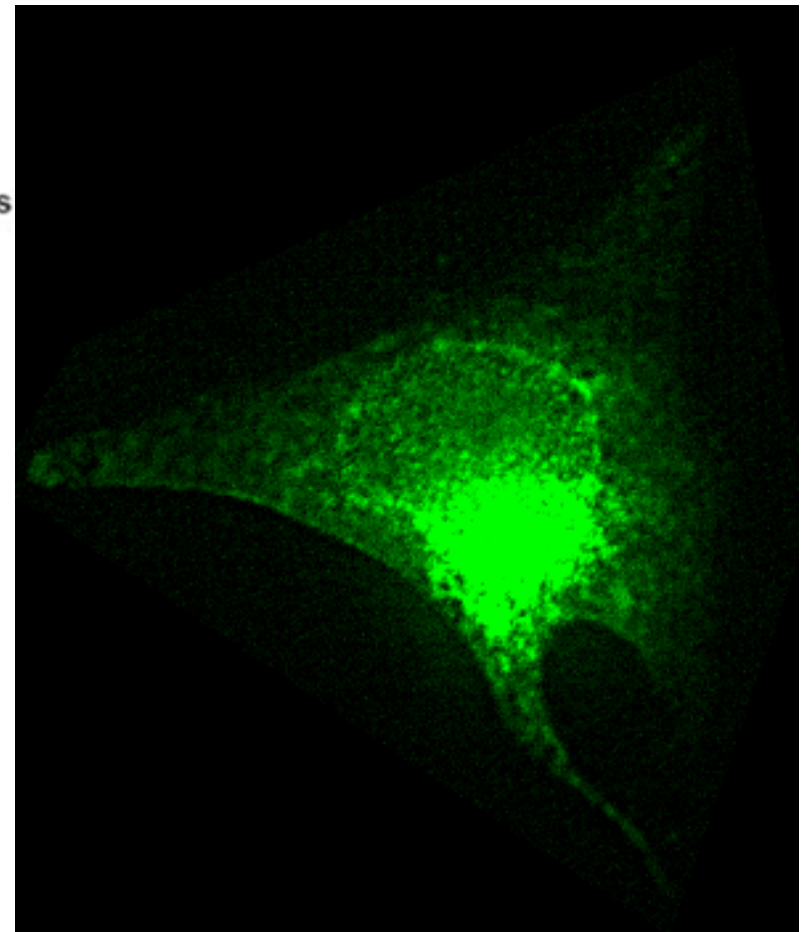
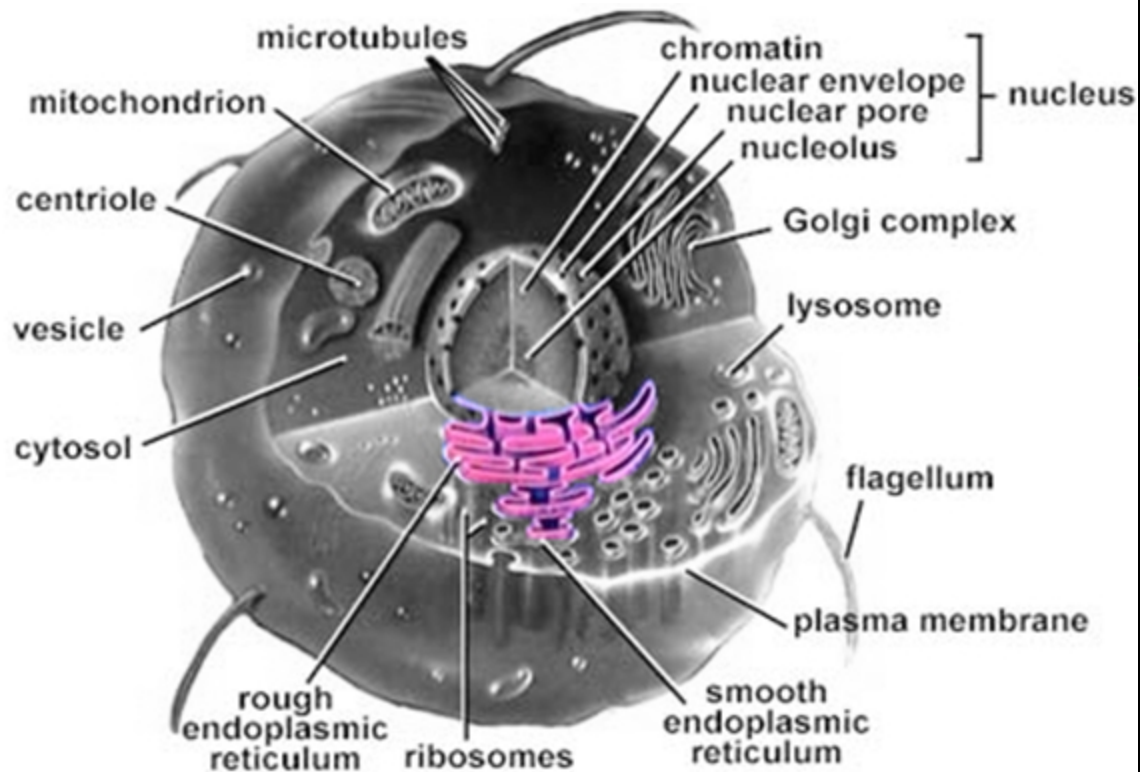
- Primary method used to **determine** the subcellular location of a protein is to “tag” it with a fluorescent probe



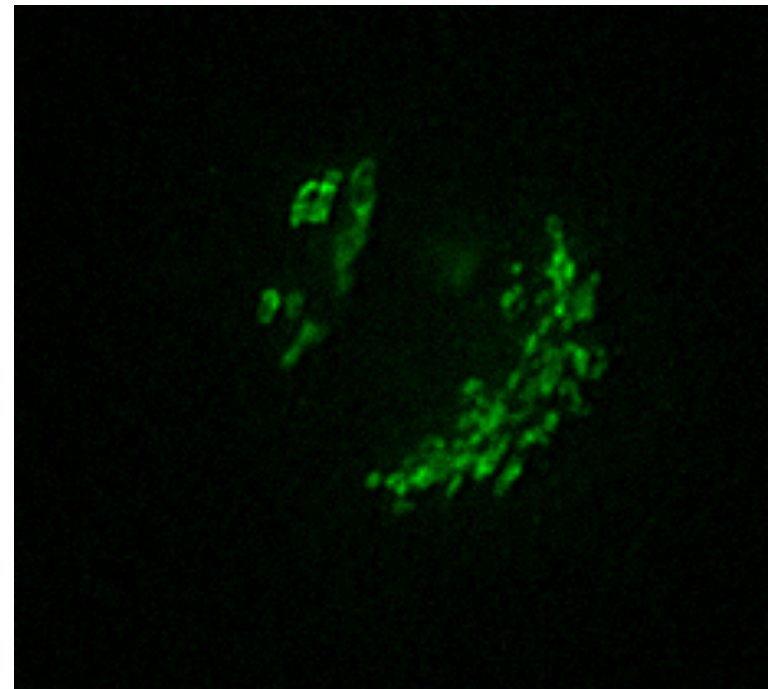
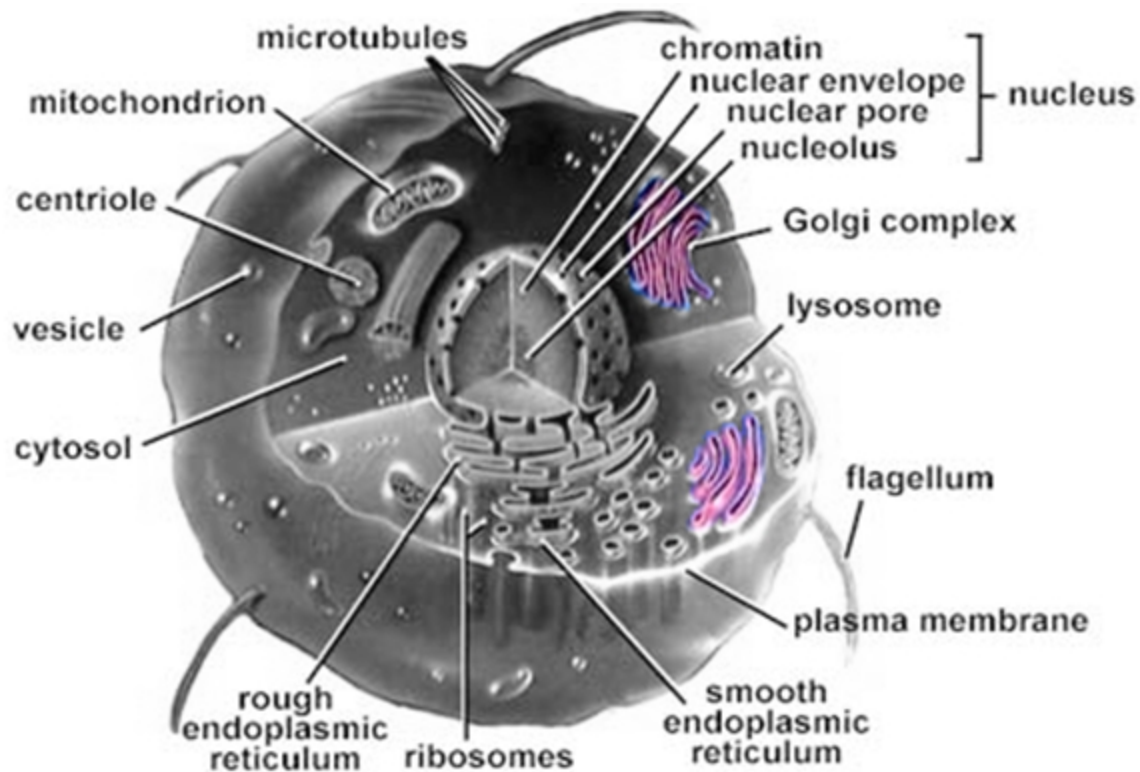
Subcellular Location



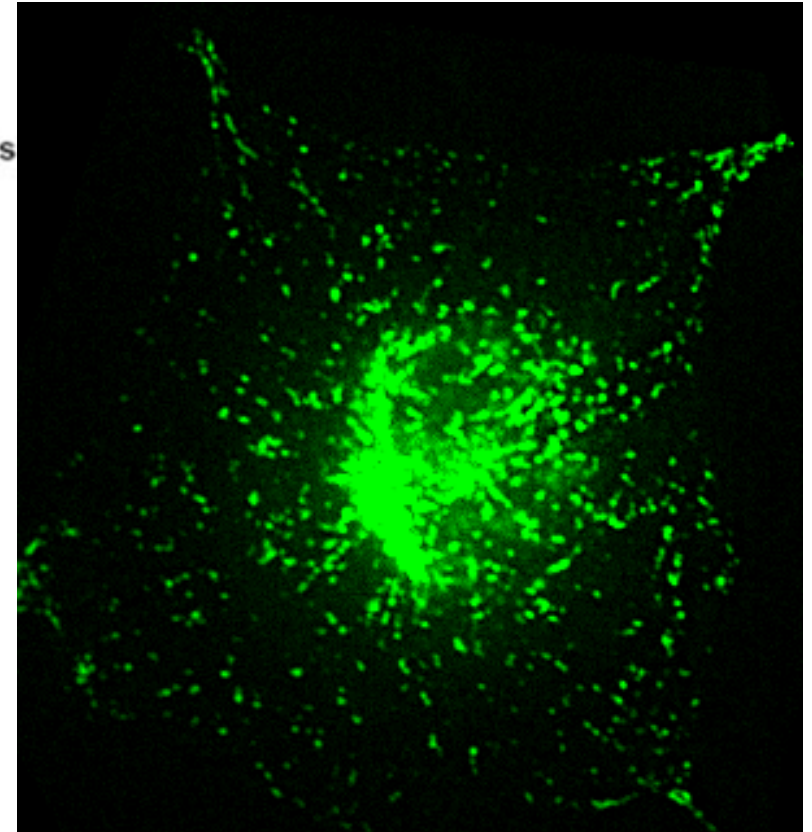
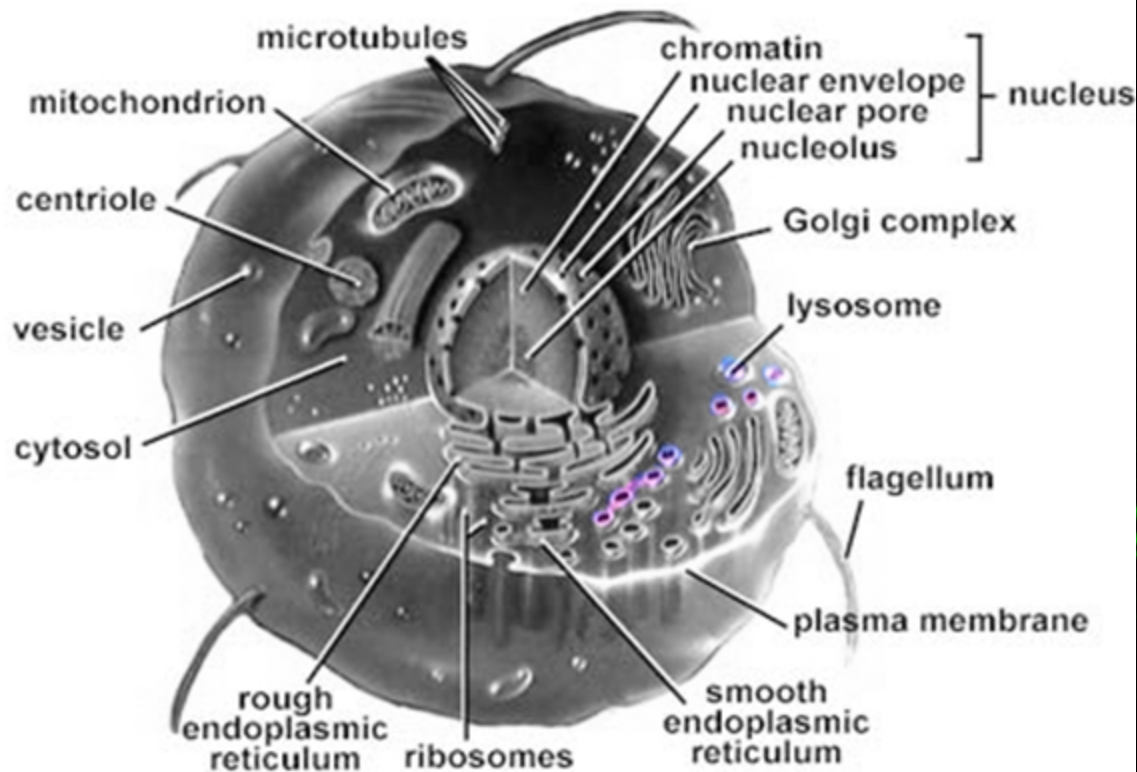
Subcellular Location



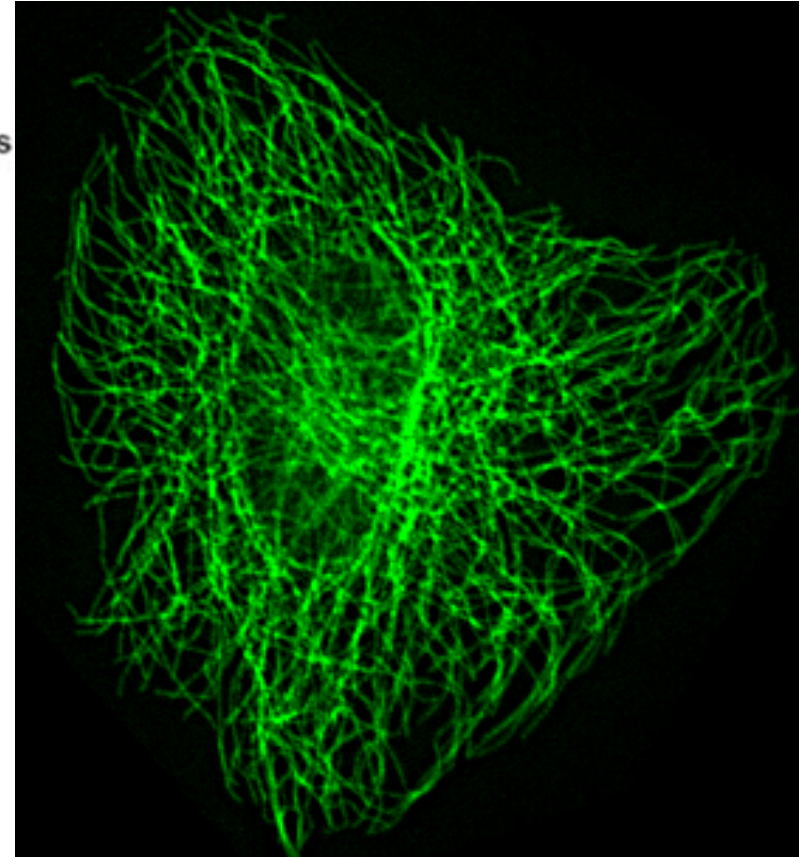
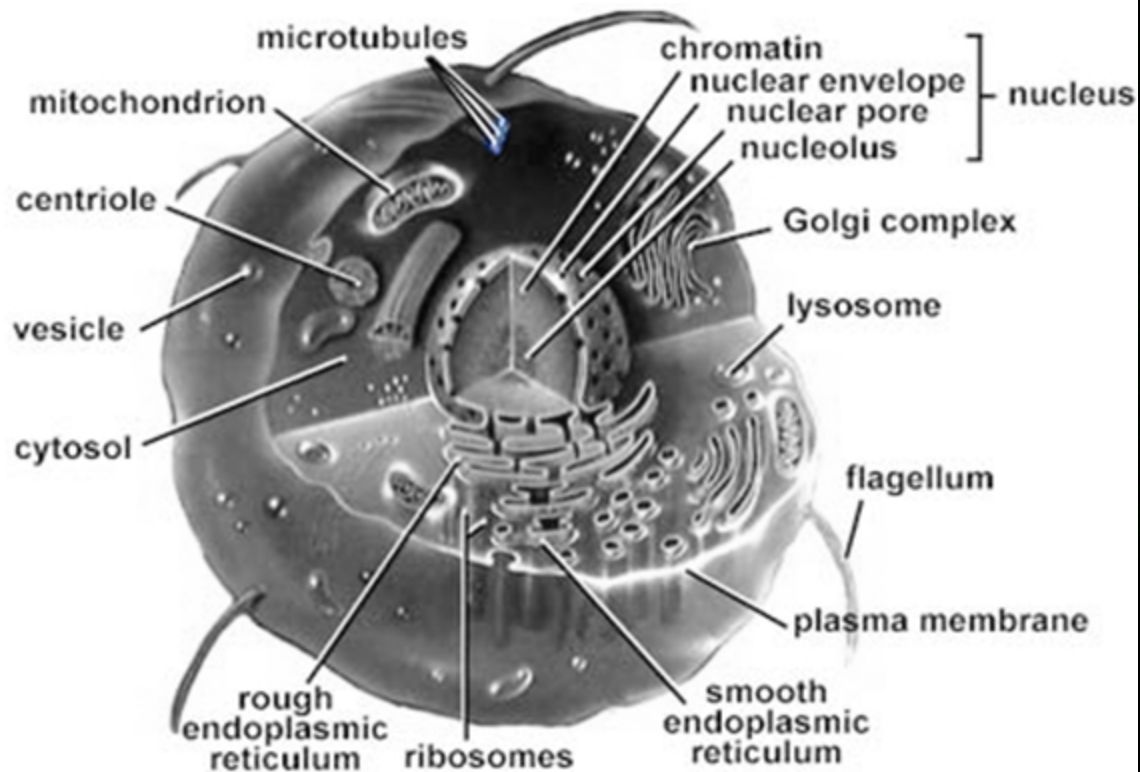
Subcellular Location



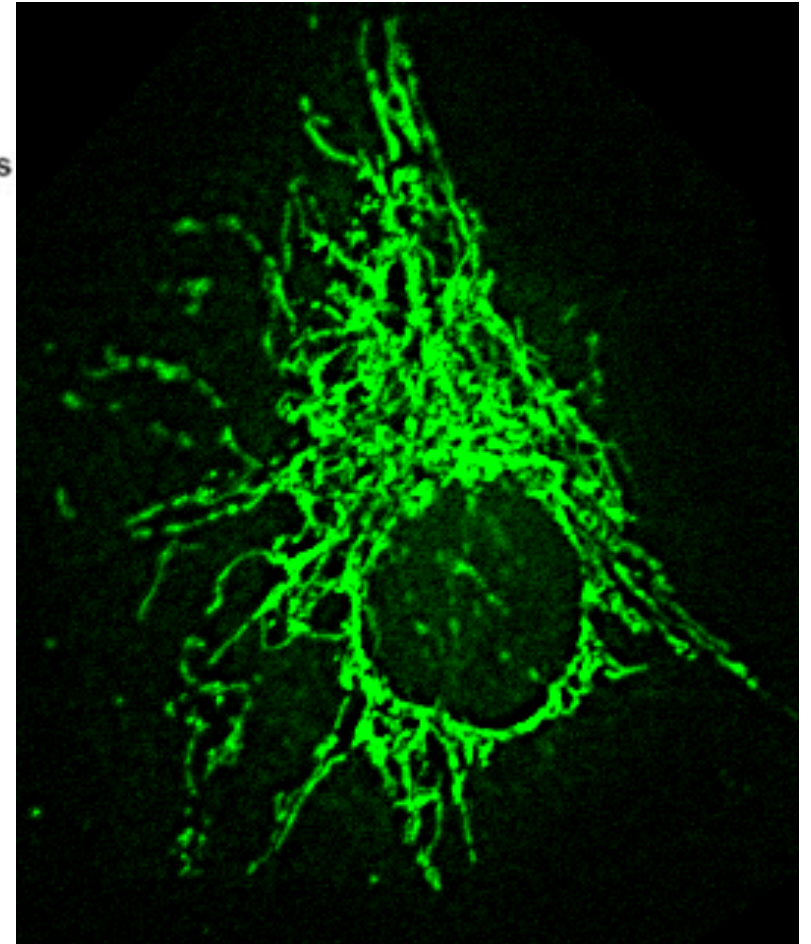
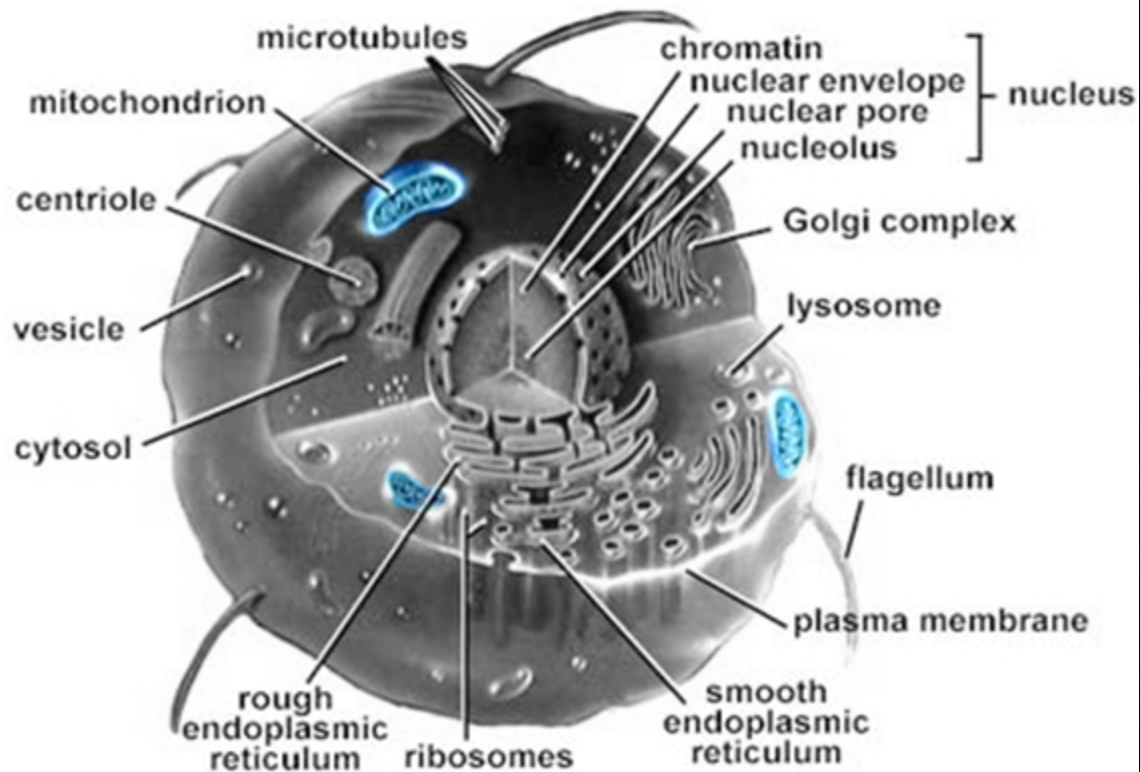
Subcellular Location



Subcellular Location

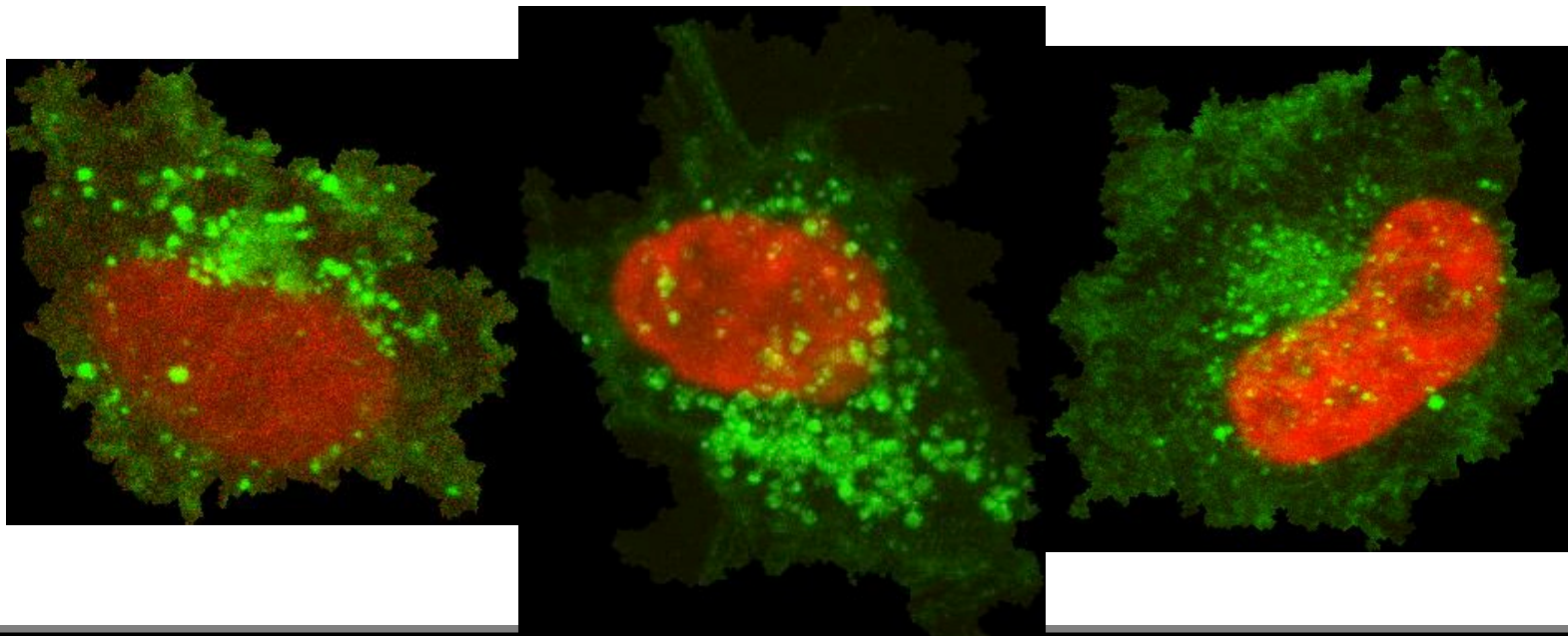


Subcellular Location



Can recognition of subcellular location be automated?

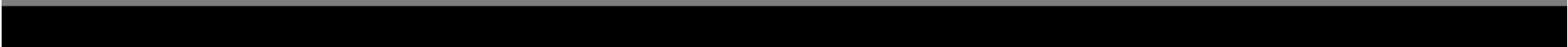
- Problem is hard because different cells have different **shapes, sizes, orientations**
- Organelles **not found in fixed locations**

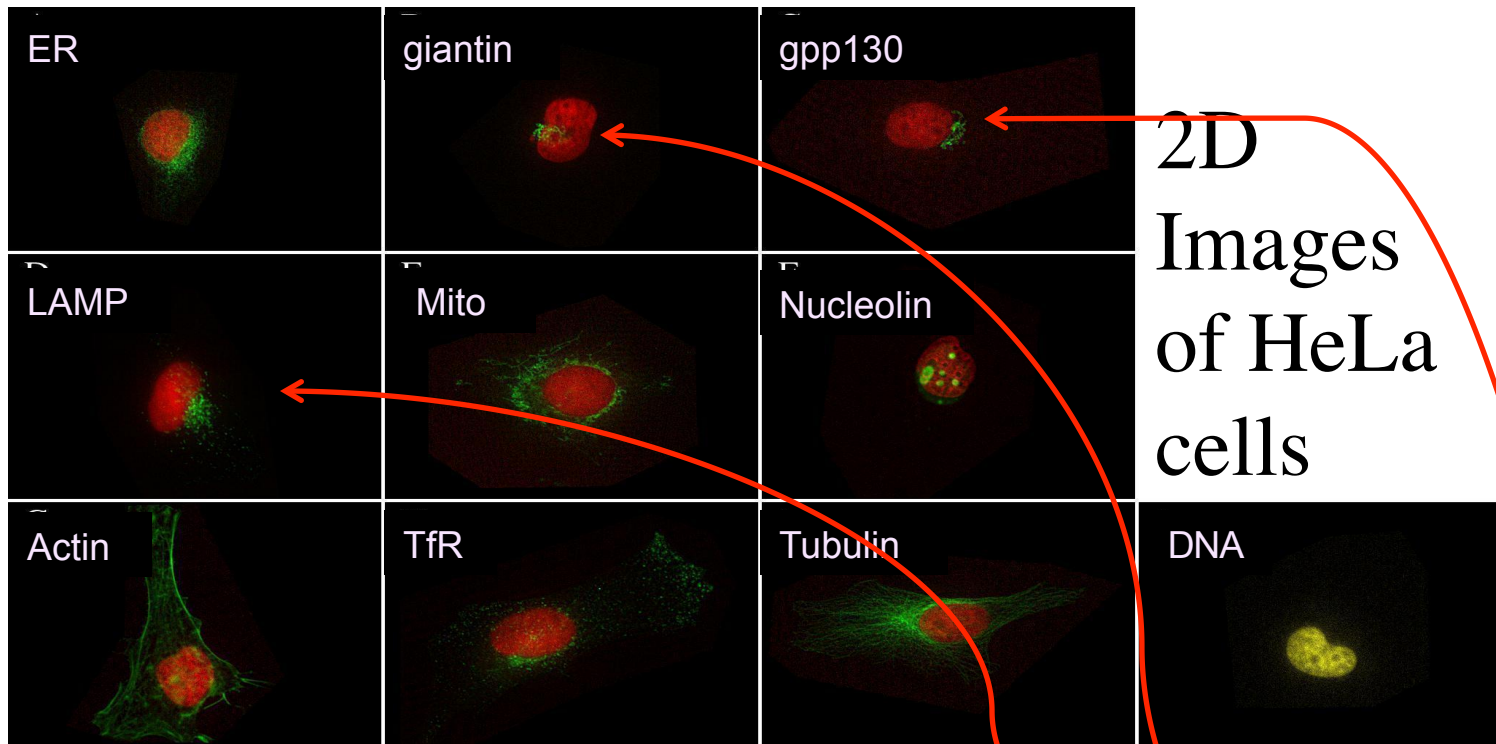




Supervised Machine Learning Approach

- Design features to describe subcellular patterns
- Use examples of images of different subcellular patterns to train classifier

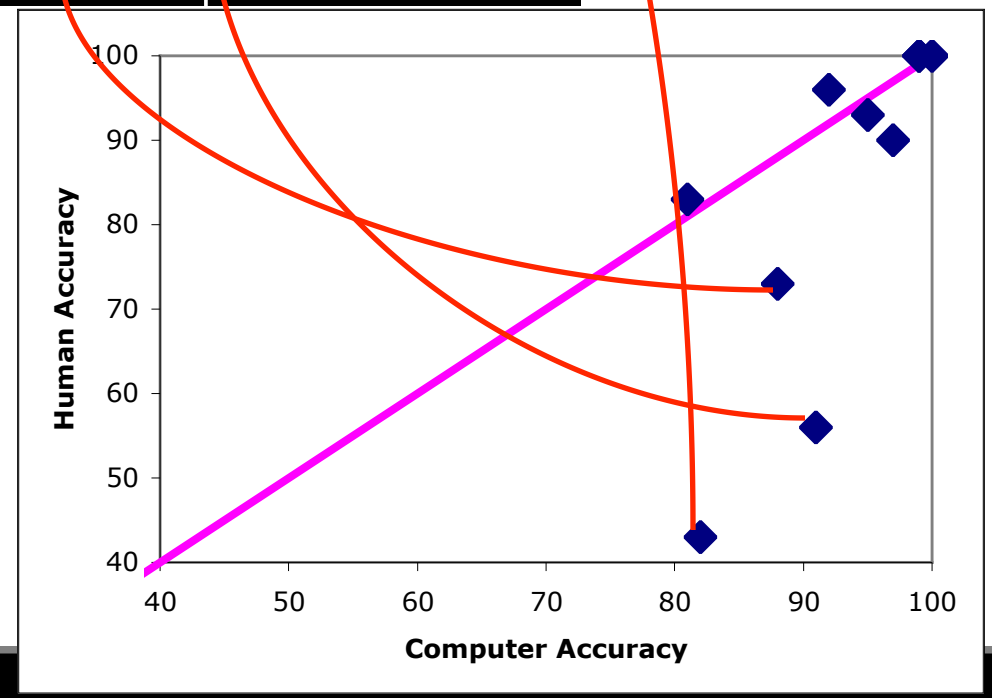


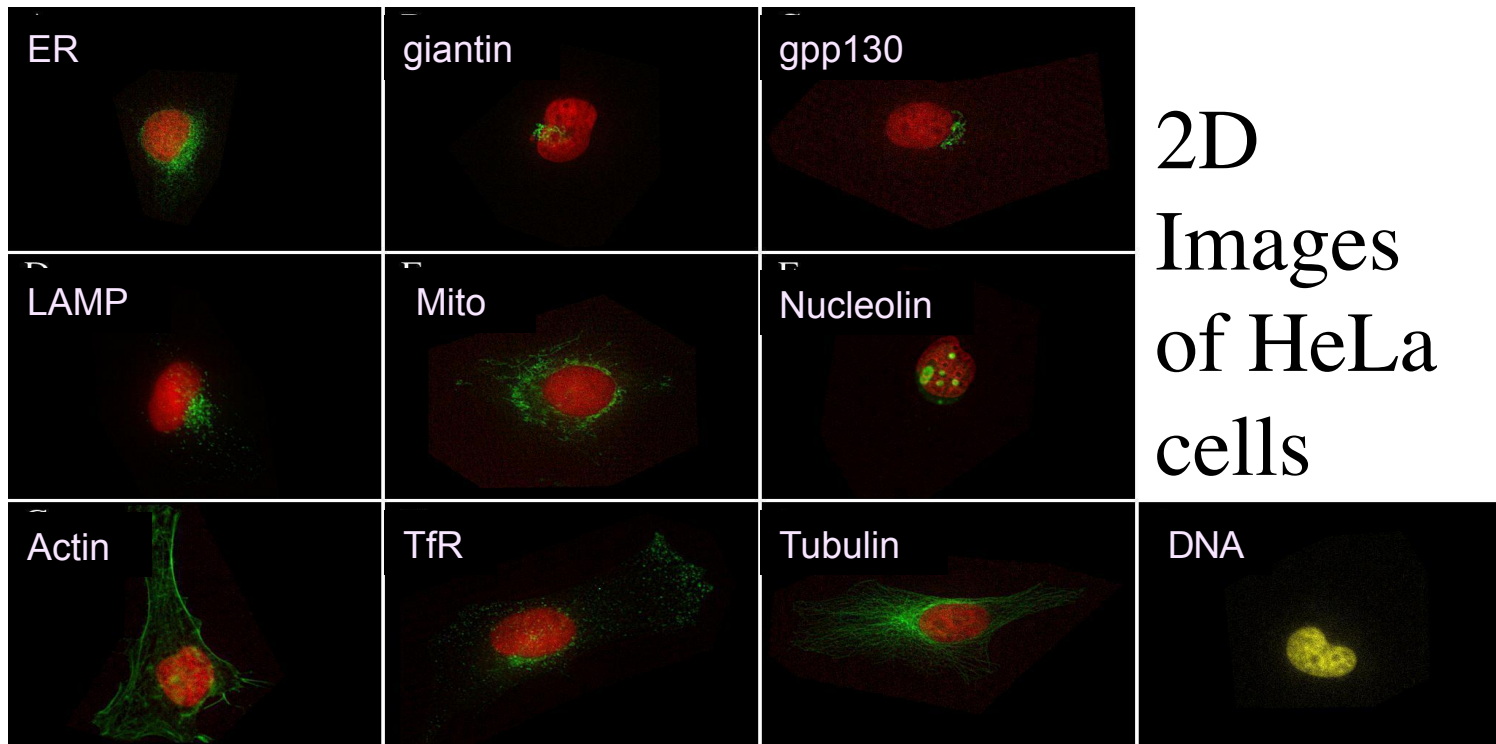


2D
Images
of HeLa
cells

Boland et al 1997;
Murphy et al 2000;
Boland & Murphy
2001; Murphy et al
2003; Huang &
Murphy 2004

Subcellular Pattern Classification: Computer vs. Human

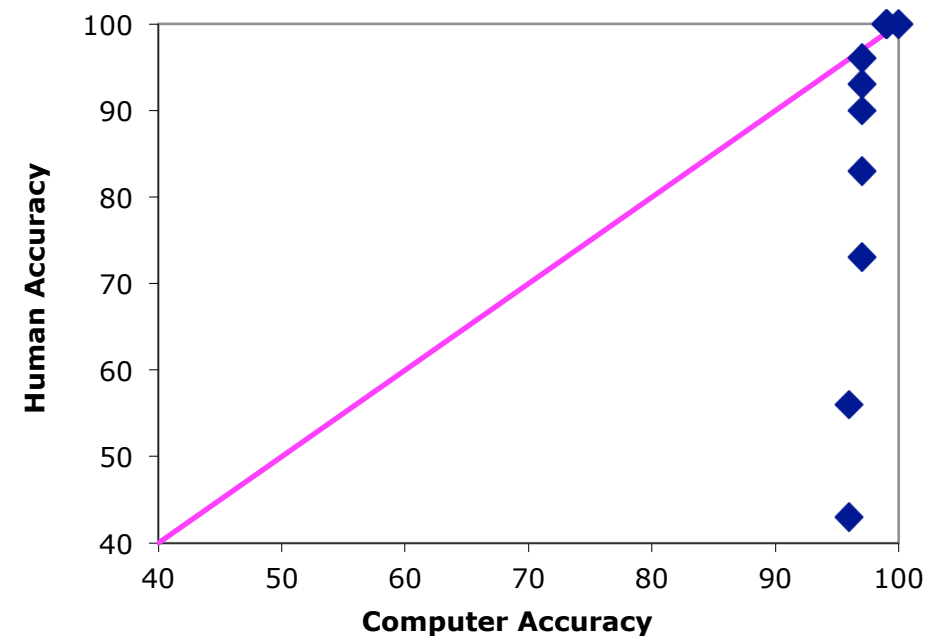




Chebira et al 2007;
Nanni et al 2010

2D Images of HeLa cells

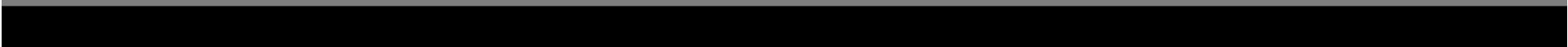
Subcellular Pattern Classification: Computer vs. Human





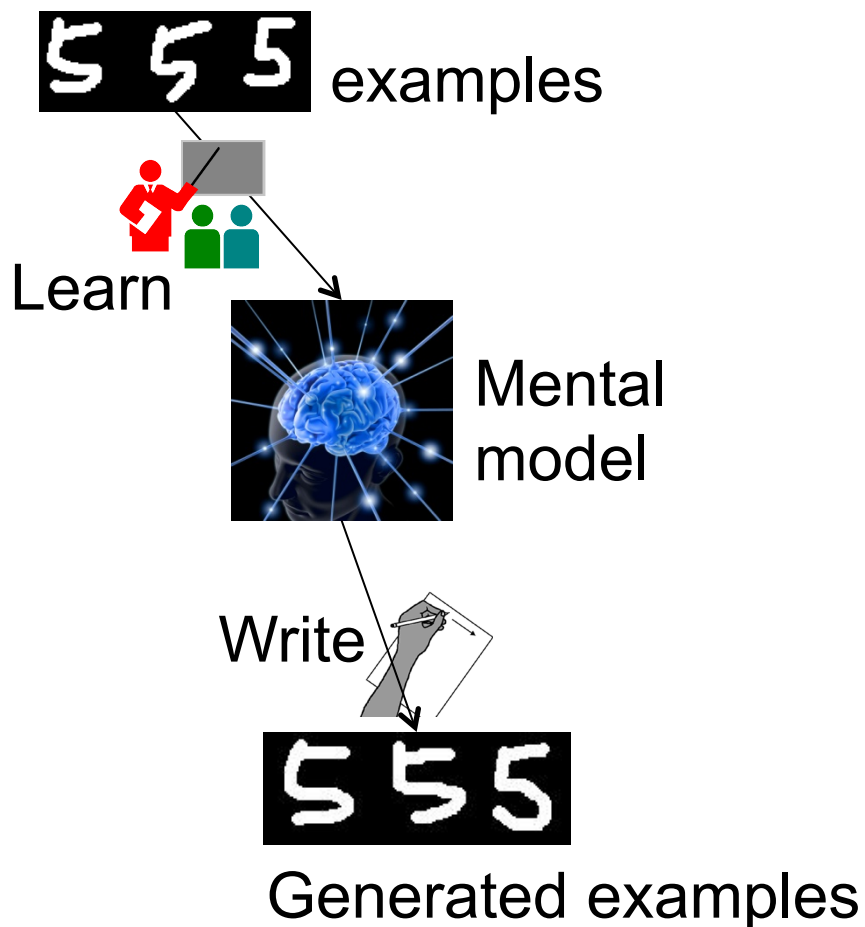
However...

- Assigning words not sufficient
 - Knowing that apples and oranges can be distinguished by their color does not allow you to understand how either are formed

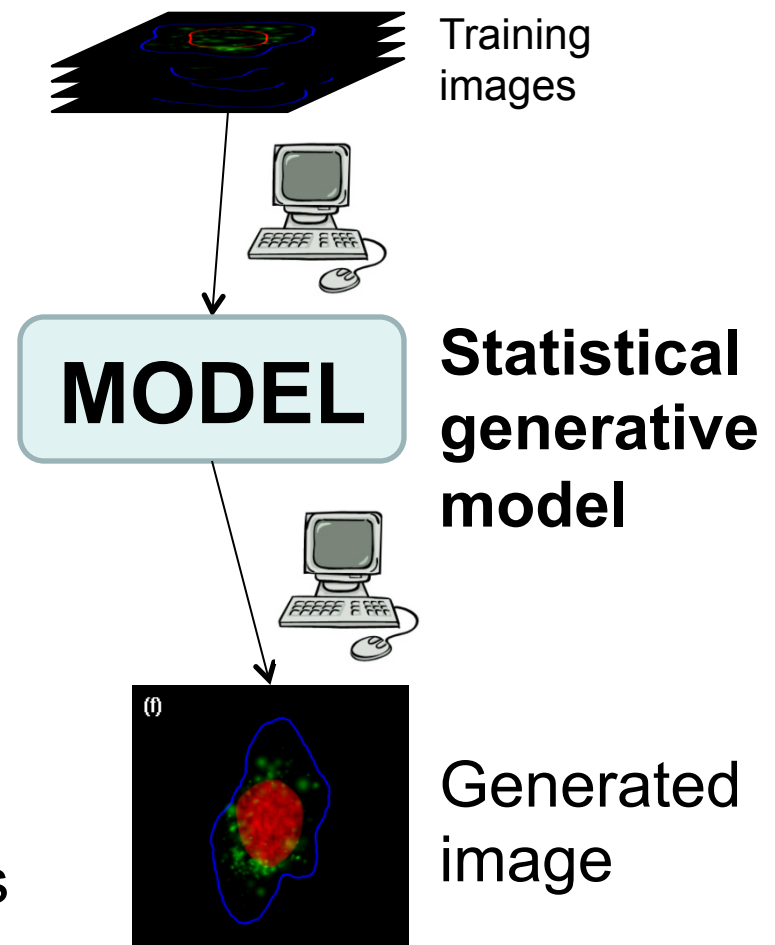


Solution: Generative Modeling

- Human cognition

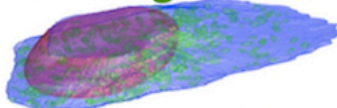


- Generative model



http://CellOrganizer.org

CellOrganizer



Images ↔ Models

RAY AND STEPHANIE LANE
Center for Computational Biology

Carnegie
Mellon
University



[Home](#)
[People](#)
[Publications](#)
[Downloads](#)

May 17, 2013: **Version 1.9.0 released!**

New: Now allows synthesis of cell and nuclear shape instances for HeLa cells using a diffeomorphic model.

The **CellOrganizer** project provides tools for

- learning generative models of cell organization directly from images
- storing and retrieving those models in XML files
- synthesizing cell images (or other representations) from one or more models

Model learning captures variation among cells in a collection of images. Images used for model learning and instances synthesized from models can be two- or three-dimensional static images or movies.

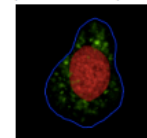
CellOrganizer can learn models of

- cell shape
- nuclear shape
- chromatin texture
- vesicular organelle size, shape and position
- microtubule distribution.

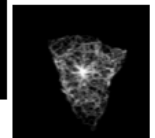
These models can be *conditional* upon each other. For example, for a given synthesized cell instance, organelle position is dependent upon the cell and nuclear shape of that instance.

Cell types for which generative models for at least some organelles have been built include human HeLa cells, mouse NIH 3T3 cells, and Arabidopsis protoplasts. Planned projects include mouse T lymphocytes and rat PC12 cells.

Synthesized Cell Images
(click to view)



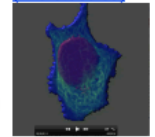
[2D HeLa
\(endosomes\)](#)



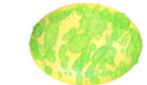
[3D HeLa
\(microtubules\)](#)



[3D HeLa
\(mitochondria\)](#)



[3D HeLa movie](#)



[3D protoplast
\(chloroplasts\)](#)

Support for **CellOrganizer** has been provided by grants GM075205 and GM090033 from the [National Institute of General Medical Sciences](#), grants MCB1121919 and MCB1121793 from the [U.S. National Science Foundation](#), by a [Forschungspreis from the Alexander von Humboldt Foundation](#), and by the [School of Life Sciences of the Freiburg Institute for Advanced Studies](#).



Alexander von Humboldt
Stiftung/Foundation

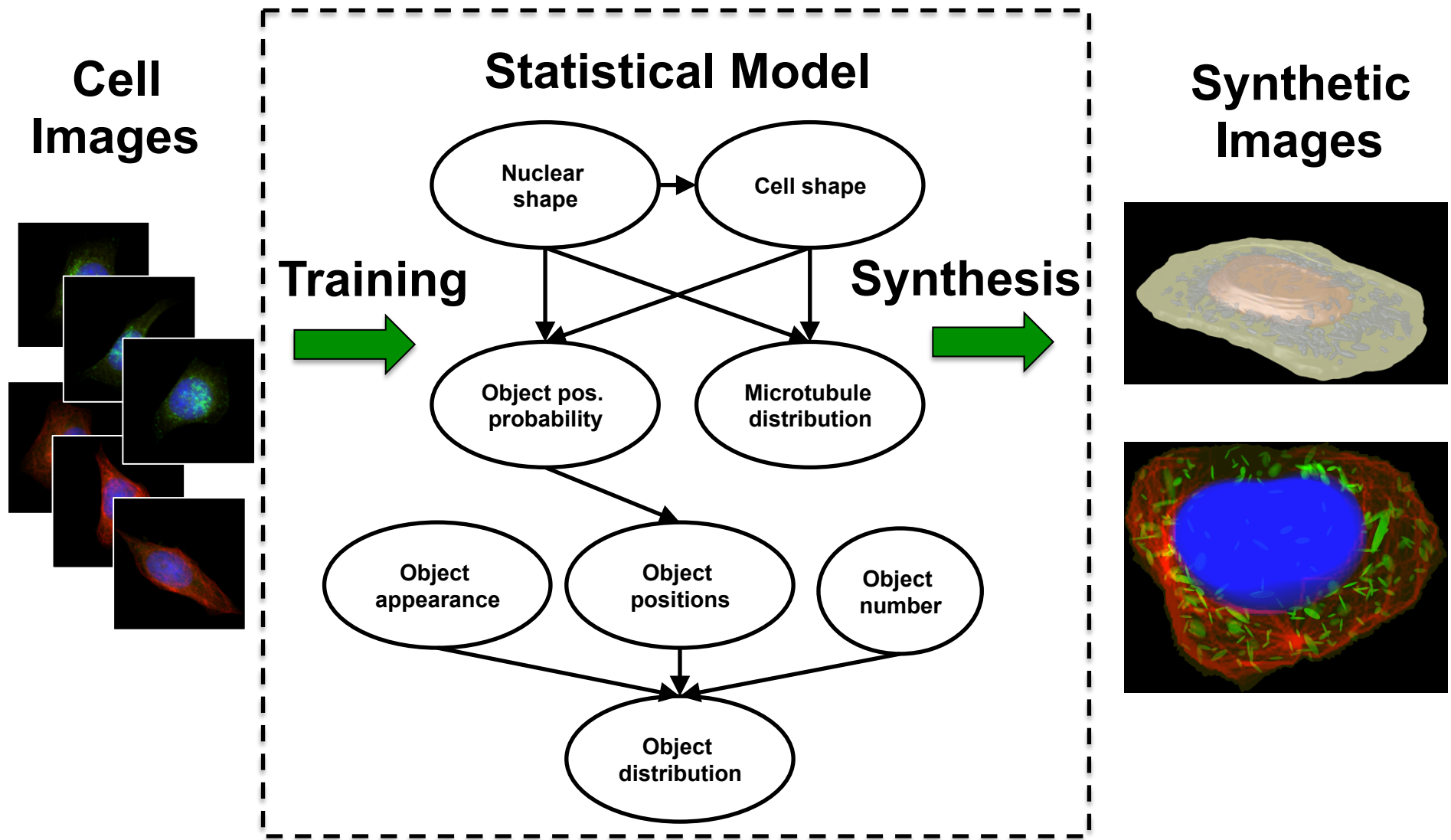


ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG



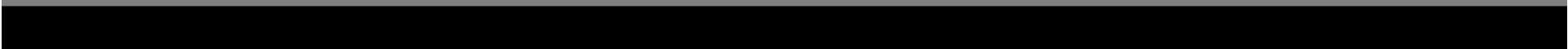
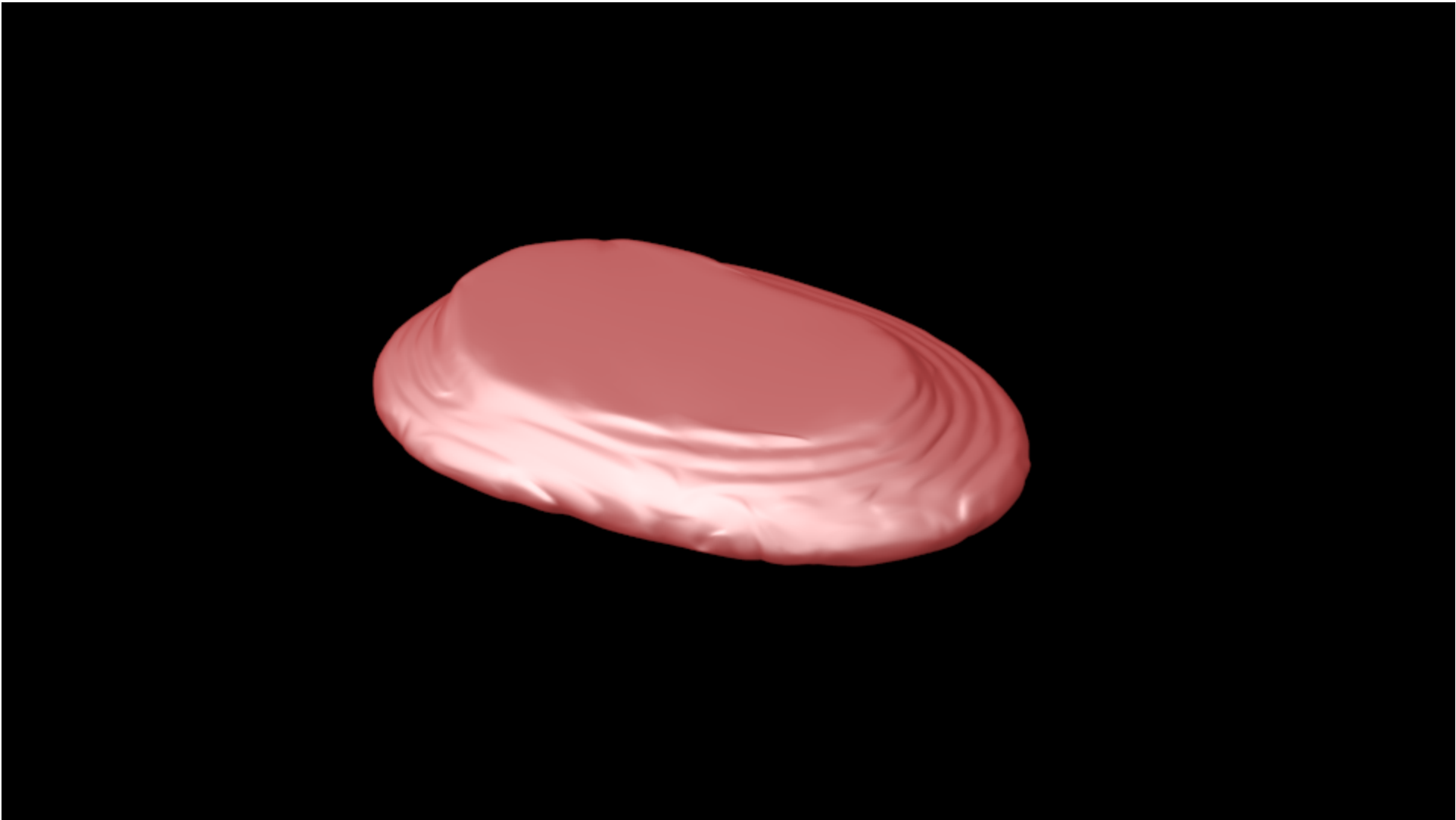
FRIAS
FREIBURG INSTITUTE
FOR ADVANCED STUDIES
LIFE SCIENCES - LIFE NET

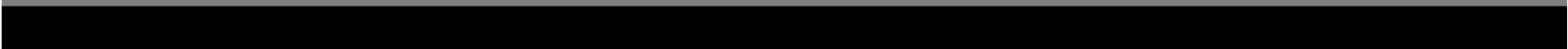
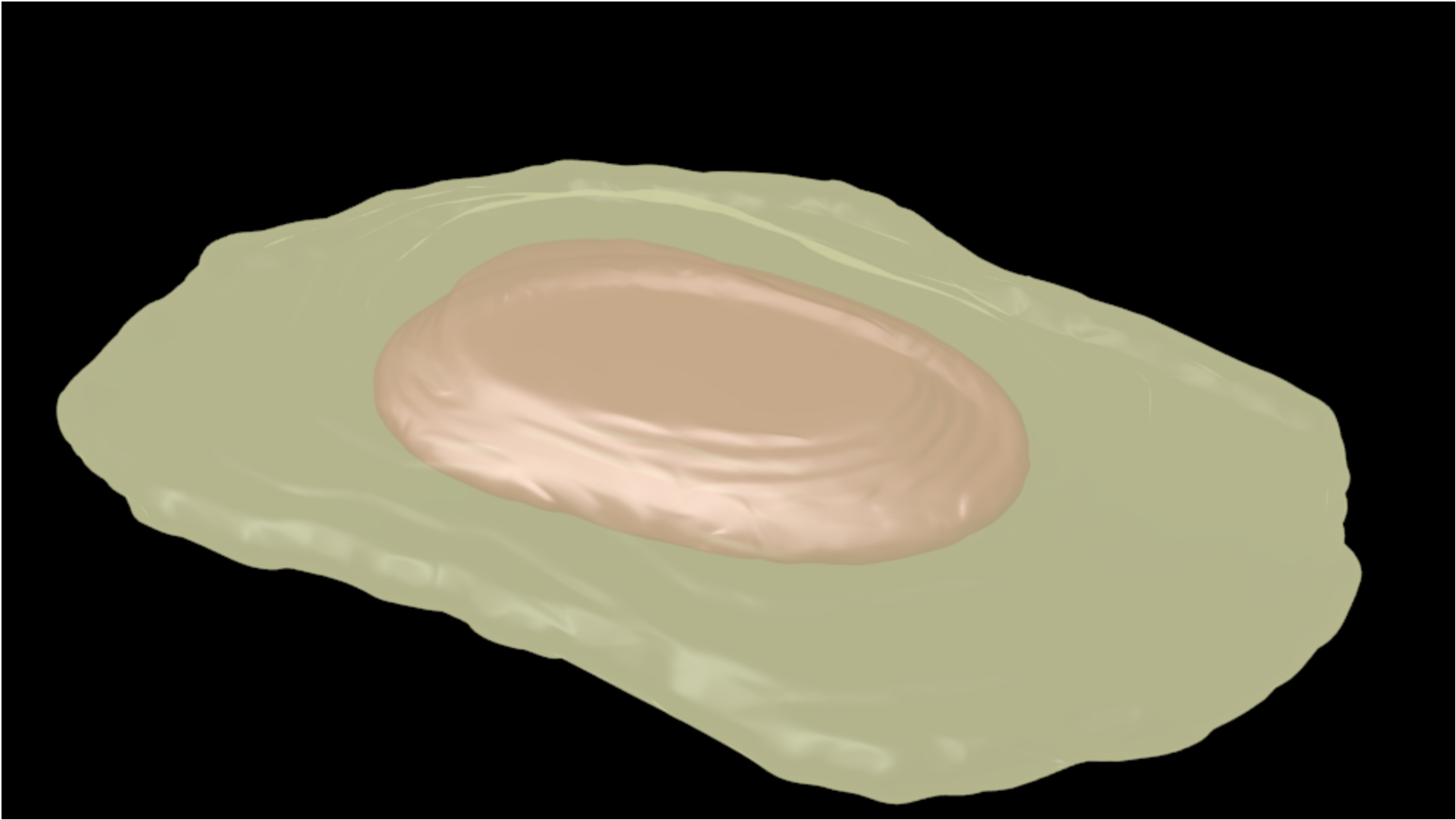
CellOrganizer

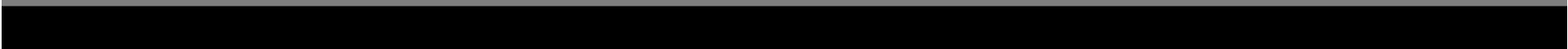
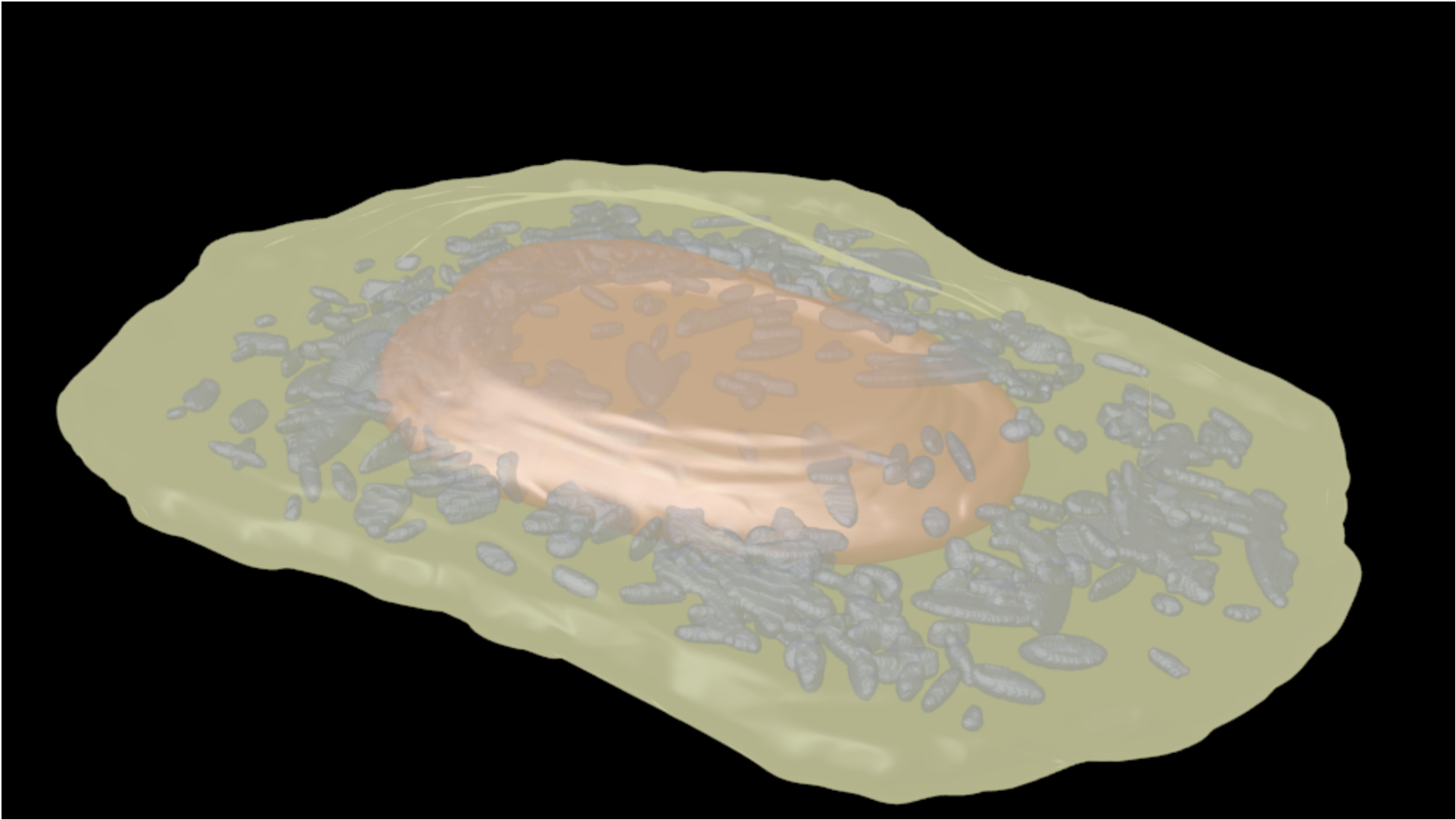


What does CellOrganizer Do?

- Given a population of cells we can model
 - Distribution of nuclear and cell shapes
 - Distribution of vesicular patterns
 - Number, shape, intensity, location of objects
 - Distribution of microtubules
 - Number, stiffness, location of centrosome
 - Relationships between these models
- Given models we can
 - Synthesize new images
 - Model populations that change over time/condition

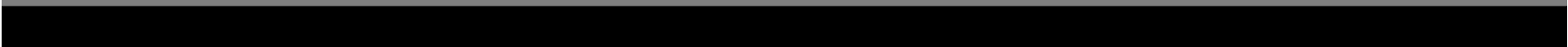








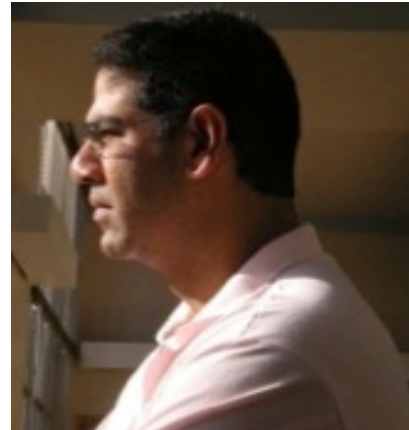
Conclusions

- Active learning provides solution to infeasibility of performing all possible perturbation and differentiation experiments
 - Tools beginning to be available to build image-derived generative models
 - Learn the underlying cell “model” from which individual cell images are drawn
- 

Active Learning



Josh Kangas



Armaghan Naik



Chris Langmead

CellOrganizer

Project Leaders



Robert F.
Murphy



Gustavo
Rohde

Major Collaborators



Klaus Palme



Christoph Wülfing



Jörn Dengjel



Hauke Busch

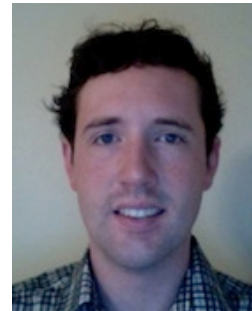


Melanie Boerries



Ivo Sbalzarini

Current Team Members



Devin Sullivan



Taraz Buck



Gregory R. Johnson



Ivan Cao-Berg

Past Contributors

Ting Zhao

Tao Peng

Wei Wang

Aabid Sharif

Joshua Kangas

Jianwei Zhang

Alexander Dovzhenko

Rüdiger Trojok

Jieyue Li

Baek Hwan Cho

Support:

NIH GM075205, GM090033, GM103712

NSF MCB1121919