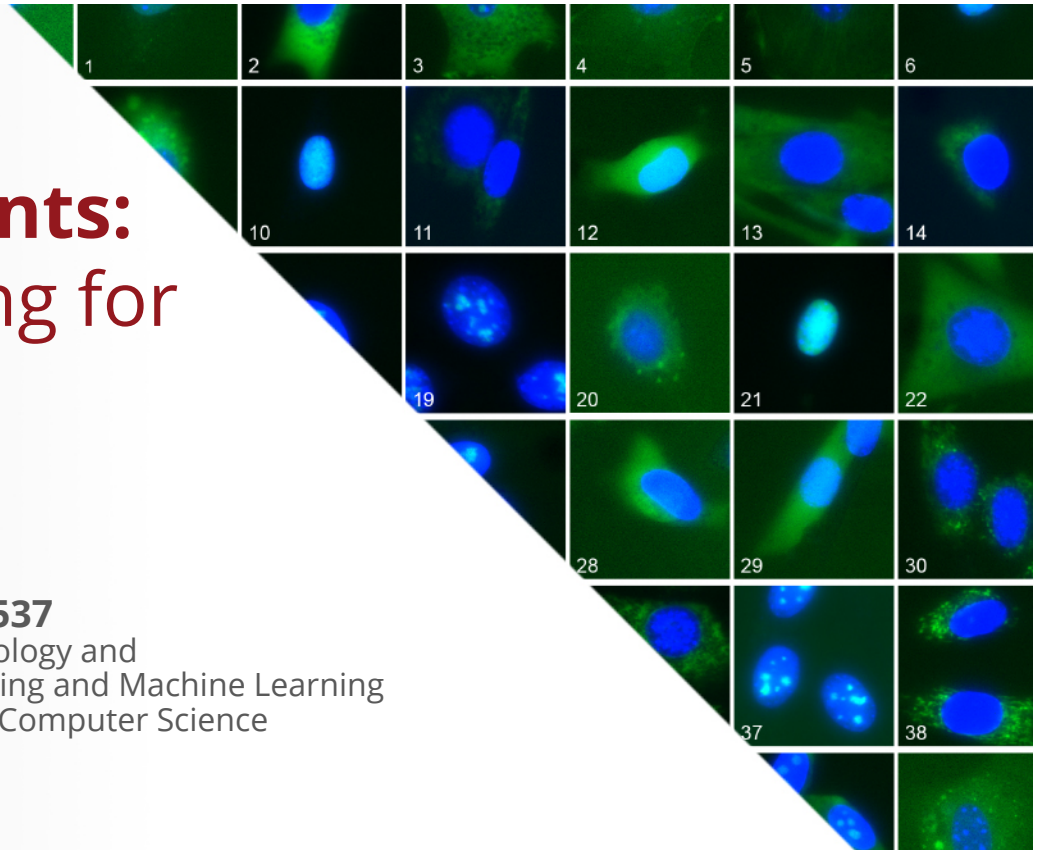
Computational Biology Department

Carnegie Mellon University
School of Computer Science

Self-driving instruments:
Active Machine Learning for Biological Discovery



Robert F. Murphy **@murphy2537**
Ray & Stephanie Lane Professor of Computational Biology and
Professor of Biological Sciences, Biomedical Engineering and Machine Learning
Head, Computational Biology Department, School of Computer Science

AAAS Annual Meeting 2019

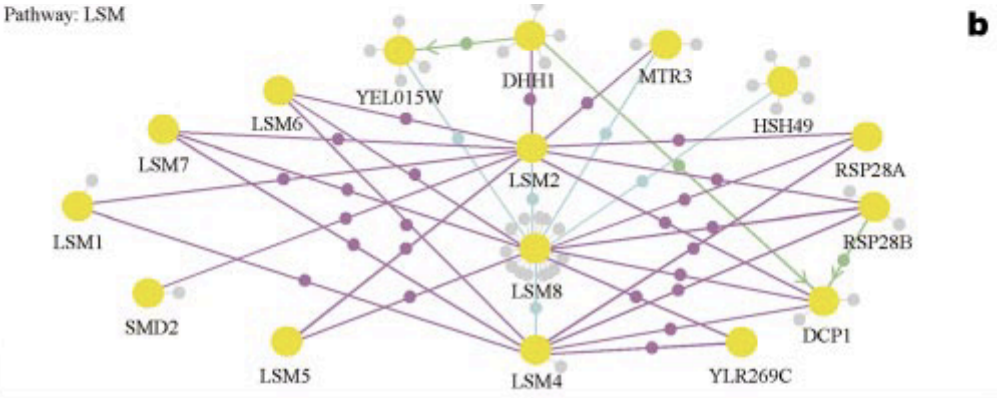
My goals

- Describe the future of self-driving instruments: how artificial intelligence/machine learning can do science without human intervention
- Review the background that makes self-driving instruments necessary
- Describe past results demonstrating feasibility

www.aarp.org/auto/trends-lifestyle/info-2018/self-driving-cars.html

The failure of Reductionism

- For many decades, biomedical research was based on ***reductionism***, the assumption that biological components could be understood in isolation
- By the 80's it was becoming clear that many, many components interacted
- Cells, Organs, Organisms are "complex systems" — "the whole is greater than the sum of the parts"



A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*

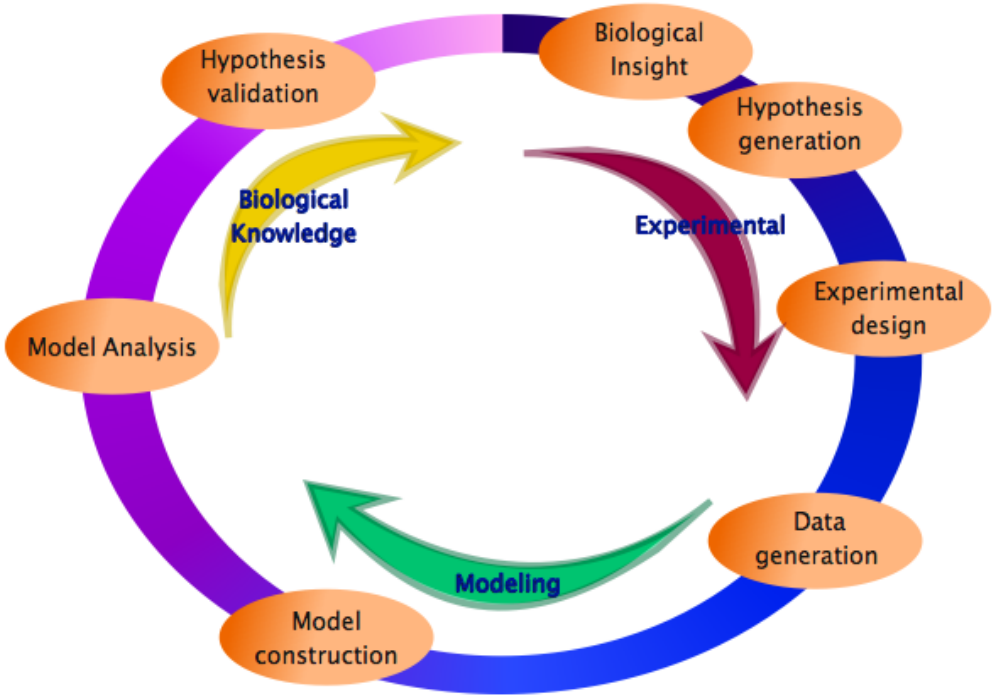
Peter Uetz*†, Loic Giot*‡, Gerard Cagney†, Traci A. Mansfield‡, Richard S. Judson‡, James R. Knight‡, Daniel Lockshon†, Vaibhav Narayan‡, Maithreyan Srinivasan‡, Pascale Pochart‡, Alia Qureshi-Emili†§, Ying Li‡, Brian Godwin‡, Diana Conover†§, Theodore Kalbfleisch‡, Govindan Vijayadamodar‡, Meijia Yang‡, Mark Johnston†‖, Stanley Fields†§ & Jonathan M. Rothberg‡

Complexity = combinatorics

- Assuming *n* genes, one gene=one function and reductionism, the number of experiments needed equals the number of genes, about 10,000
  - at (optimistically) one experiment per day, 28 years
- Given *m* average genes per function and *n* genes, the number of experiments is $n^m \sim 10^{4m} \sim 10^{20}$
  - at 10^9 experiments per day, 2 million centuries!

The rise of systems biology

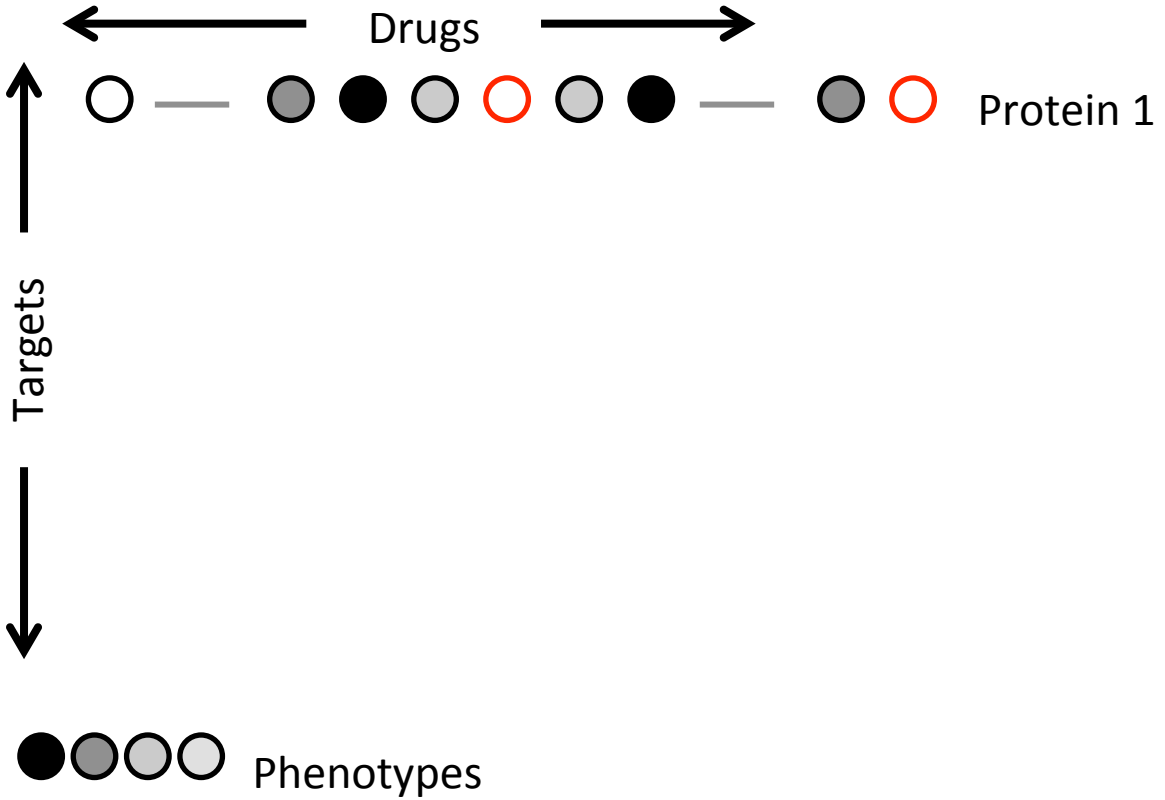
- Instead of doing all experiments build predictive models from a smaller number of experiments
- Emphasis on “validating” models by testing specific predictions
- But empirical models cannot be proven!



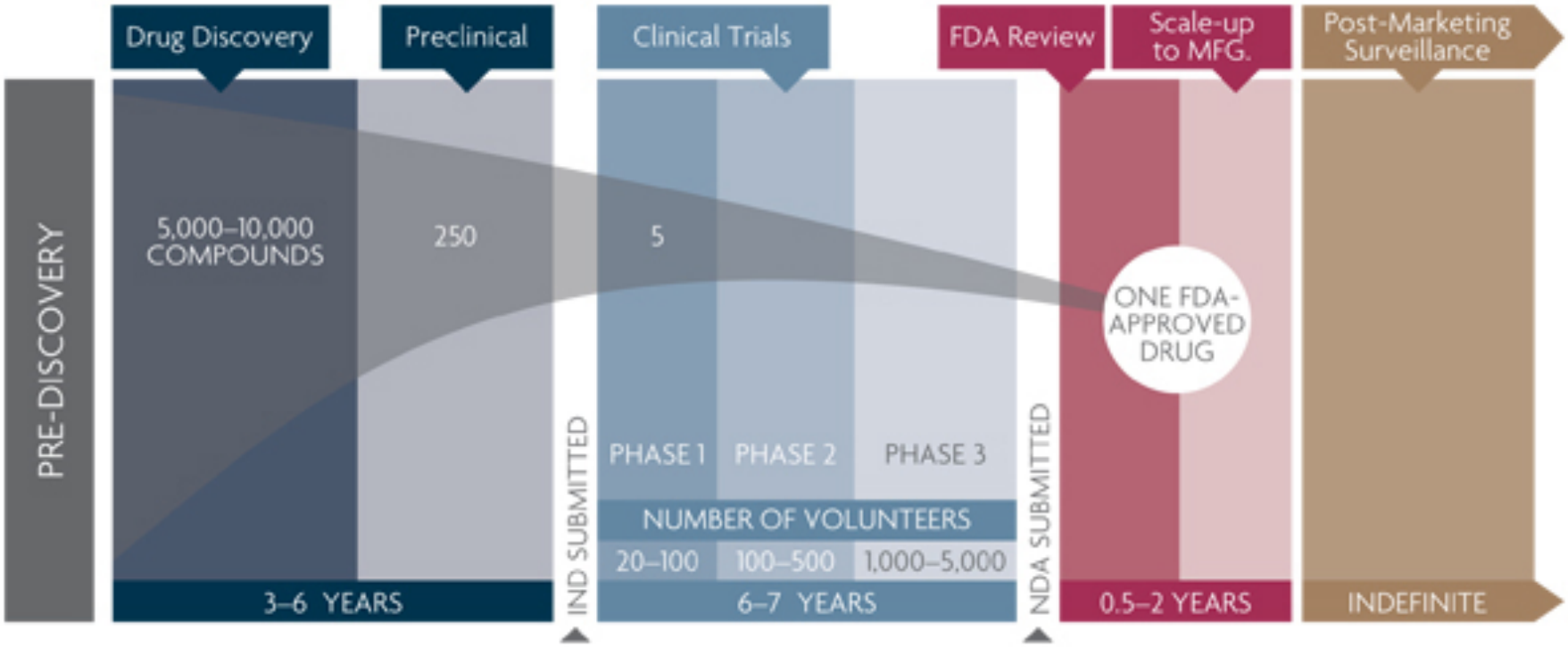
Solution?

- Use ***active*** machine learning
- Choose experiments not to ***prove*** model but to ***improve*** model

Typical drug development: consider each target separately

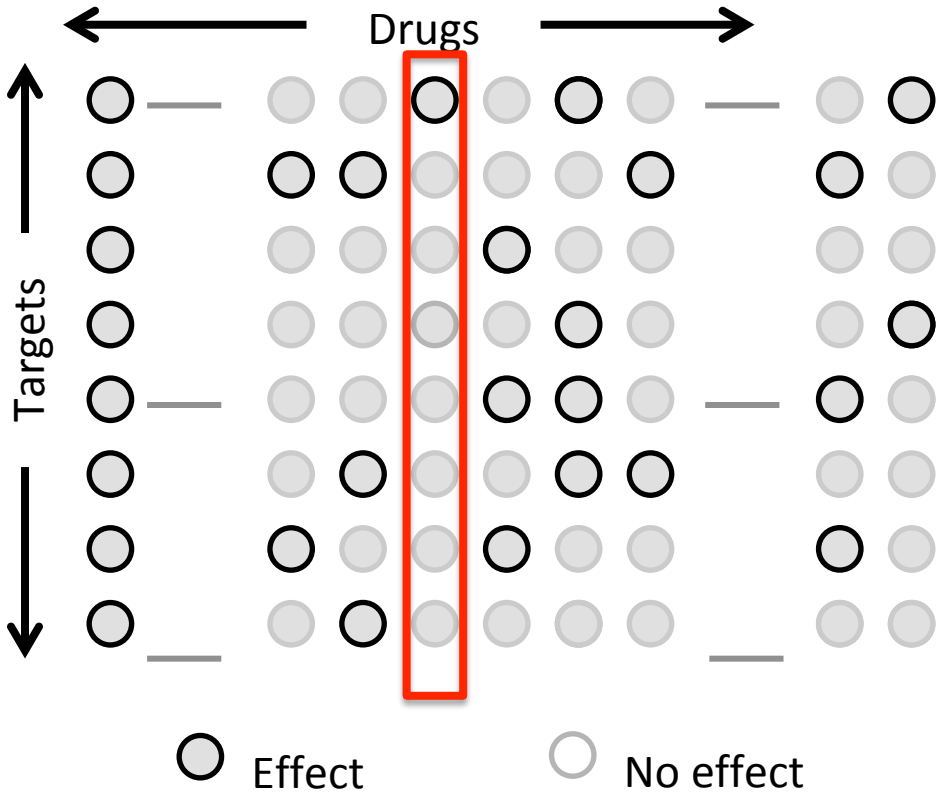


But it is not just about finding hits…



Source: PhRMA[1]

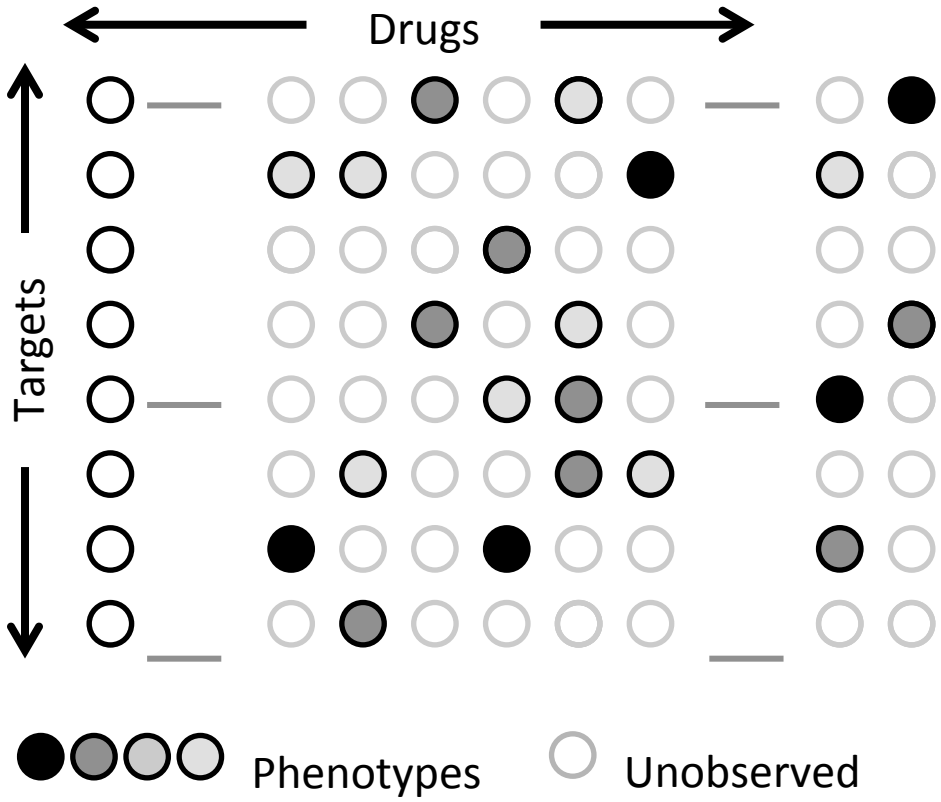
Where we'd like to be: measure all drugs for all targets



But again, too many combinations

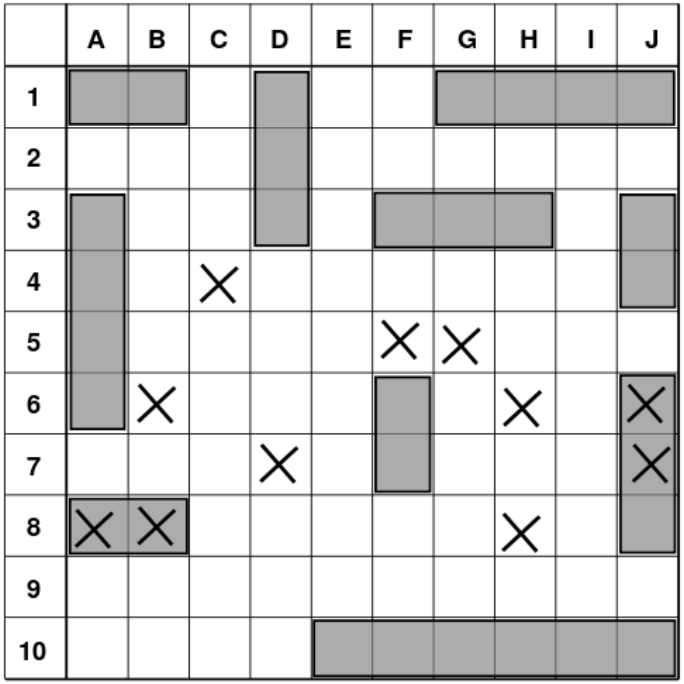
- Approximately 10,000 targets
- Approximately 1,000,000 potential drugs
- How would active learning help?

Goal: build a predictive model for all drugs and targets



Dempster et al (1977)
Hill et al. (1995);
Lee & Seung (1999);
Buchanan & Fitzgibbon (2005);
Salakhutdinov & Mnih (2008);
Mitra (2010);
Gönen (2012); …

Playing Battleship with Drugs and Cells



Source: Wikipedia

Testing retrospectively (with existing data)

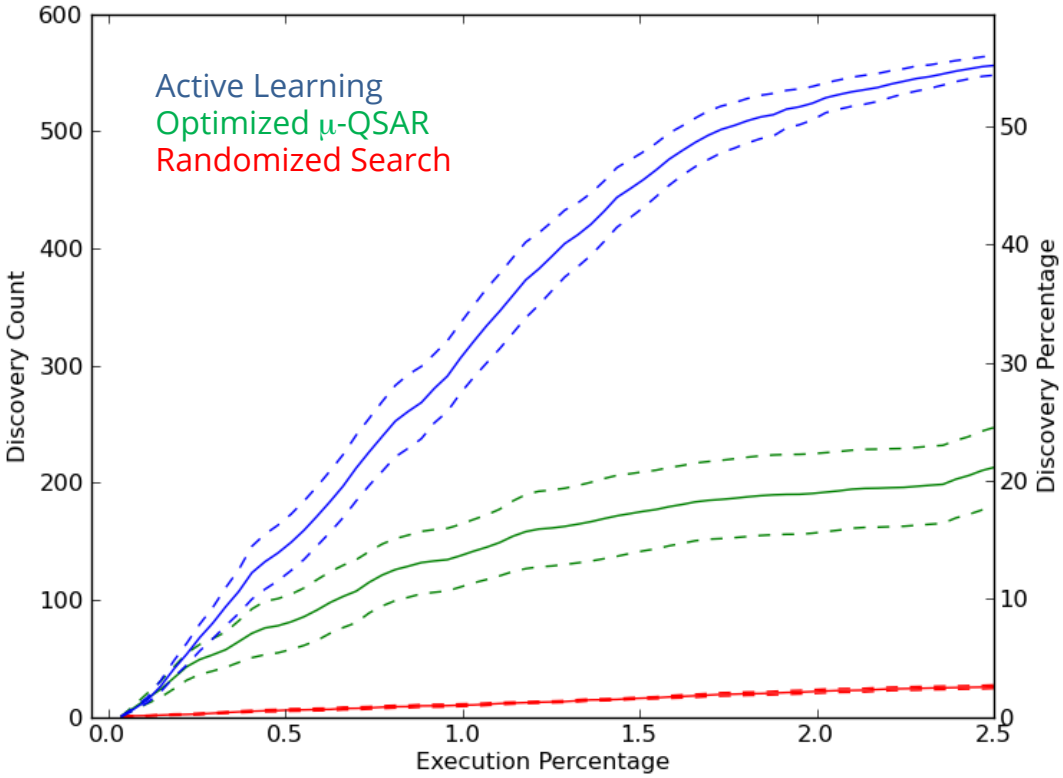
- Large database on effects of drugs on targets
- Very expensive to generate
- Would active learning have been able to save time and money?

Testing retrospectively (with existing data)

- "Hide" the PubChem data (like in Battleship) and only reveal the results when asked
  - as if we were doing that experiment for the first time
- Use different methods to choose what experiments to do

With only **2.5%** of the matrix covered, we can identify **57%** of the active compounds!

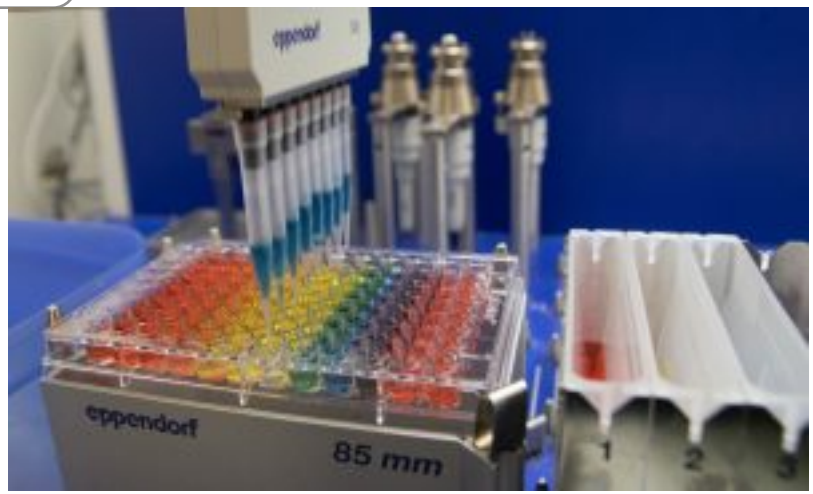
Kangas, Naik, Murphy, BMC Bioinformatics 2014



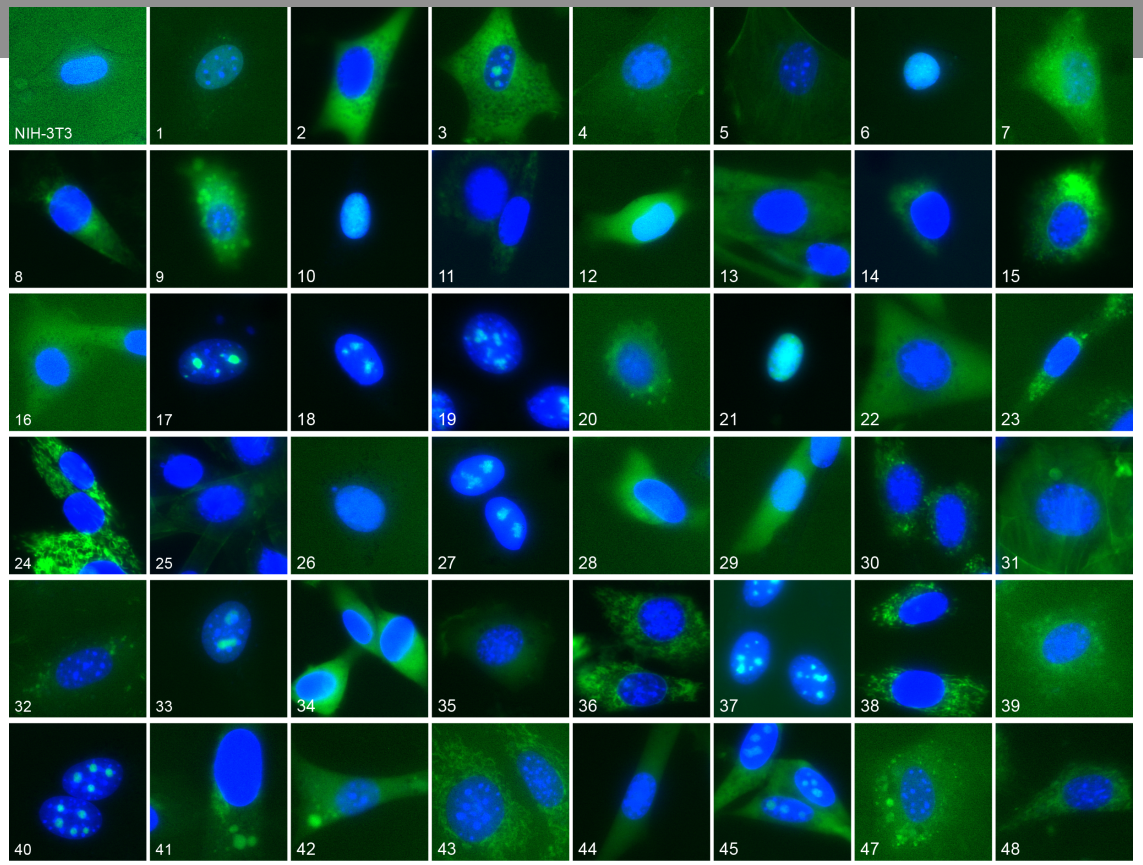
Now try this ***prospectively*** for an even harder problem

Use liquid handling robots and automated microscope to execute experiments chosen by an active learner

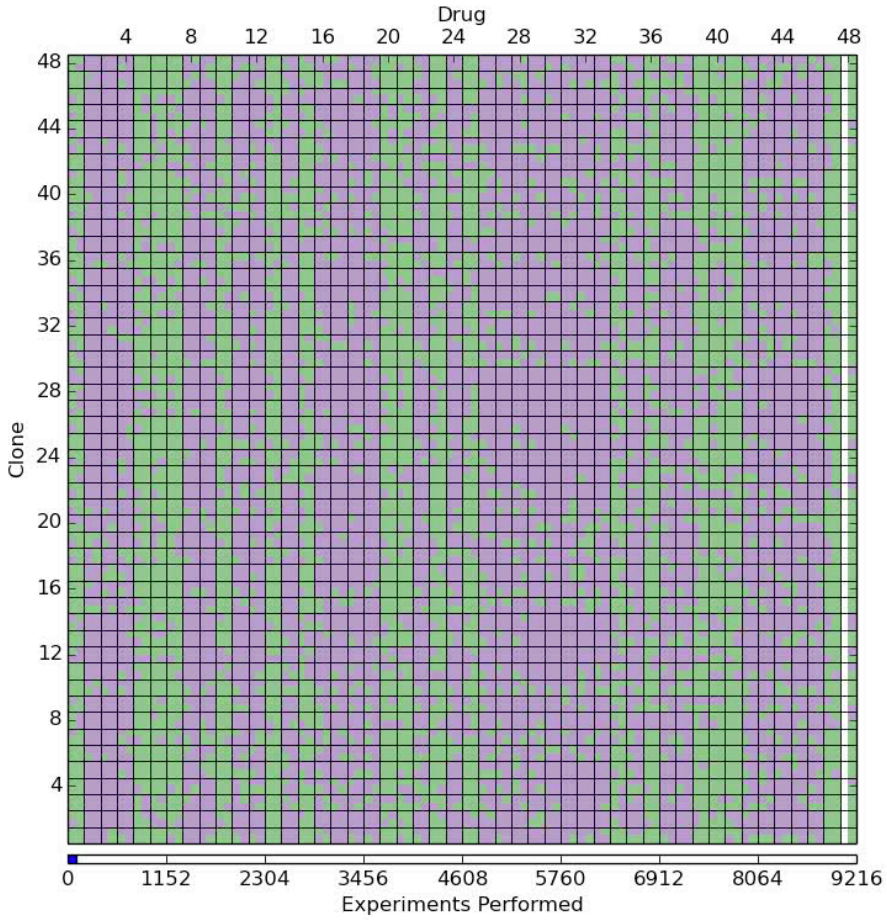
Source: Beckman Coulter



Try to learn the effects of 96 drugs upon 96 GFP-tagged proteins, *without doing experiments for all drugs and proteins,* and where the *kinds* of effects drugs might have are unknown

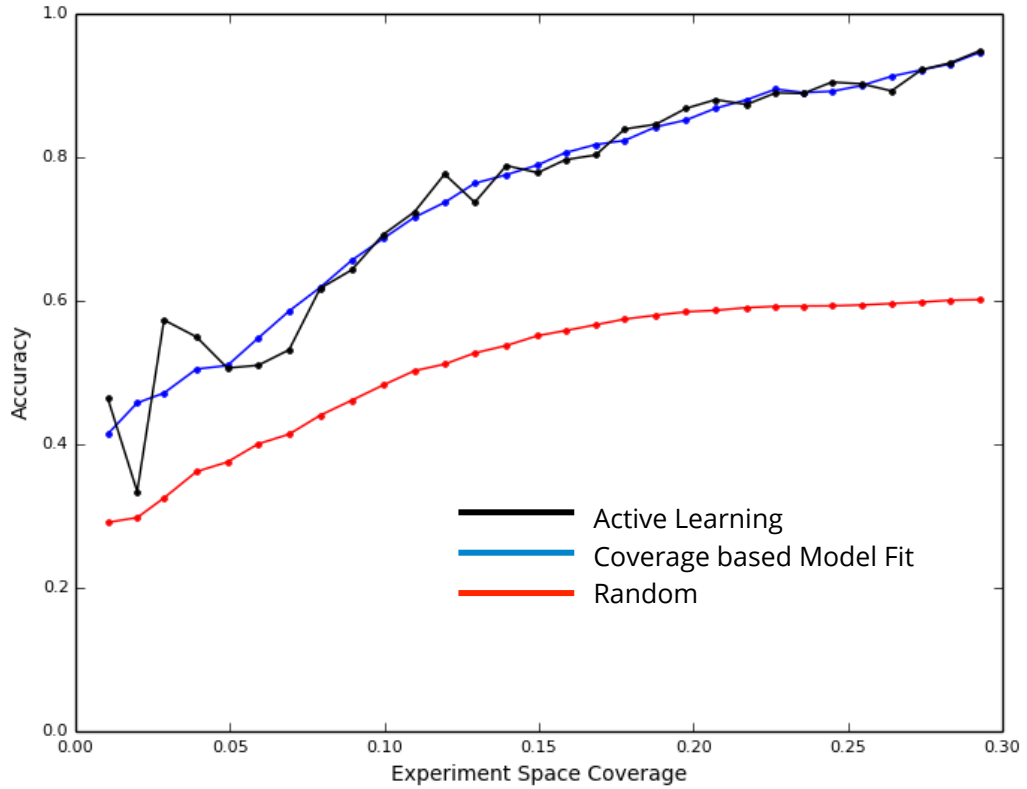


- Each small box is one drug and one target
- Green shows accurate prediction, purple is inaccurate, white shows experiments done



After doing 28% of possible experiments, model is 92% accurate and 40% more accurate than would have been obtained by random choice of experiments

Naik, Kangas, Sullivan, Murphy, eLife 2016



Automated science

- These results provide strong support for the idea of doing "Automated Science" in which not only the ***execution*** of experiments is done robotically but the ***choice*** of experiments is done robotically
- **"Self-driving instruments!"**

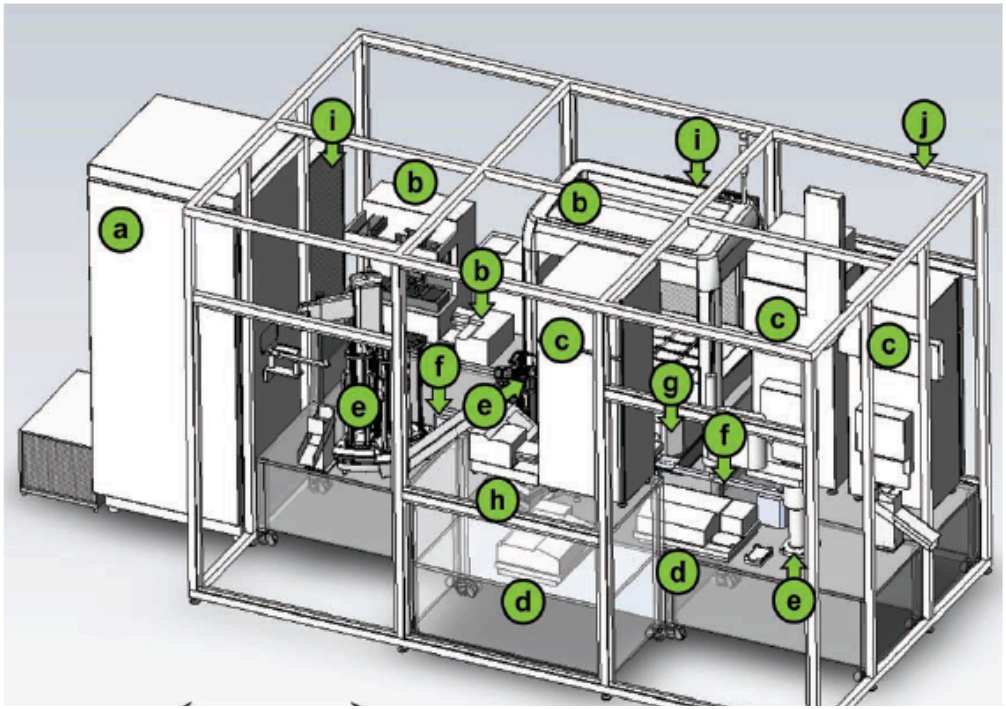
Automated science

- Additional precedent in the work of Ross King and colleagues

Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

[1]*Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK*
[2]*School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK*
[3]*Department of Computing, Imperial College, London SW7 2AZ, UK*
[4]*Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK*
[5]*School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK*



The future

- Embracing complexity in high dimensional models combined with active machine learning to guide experimentation in many areas of biomedical research
- Just like for self-driving cars, human role will be deciding where to go, not how to get there
- Training needed for the Automated Science workforce

Find us | msas-help@cbd.cmu.edu

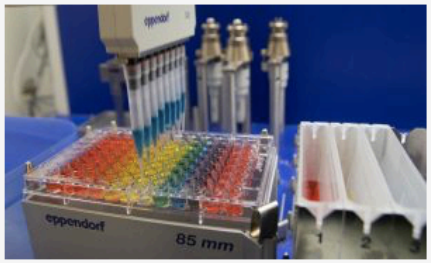
MSAS

Carnegie Mellon University
M.S. in Automated Science

Home | About | People | Curriculum | Admissions

Carnegie Mellon University's New Automated Science Program

Oct 9, 2018 | Recents News



Carnegie Mellon is pleased to announce the launch of a new graduate program: The Masters of Science in Automated Science (MSAS). MSAS graduates will be the leaders in the emerging paradigm of **Automated Science** – the combination of **robotic scientific instruments**, **Machine Learning**, and **Artificial Intelligence** for iteratively interpreting data and selecting experiments.

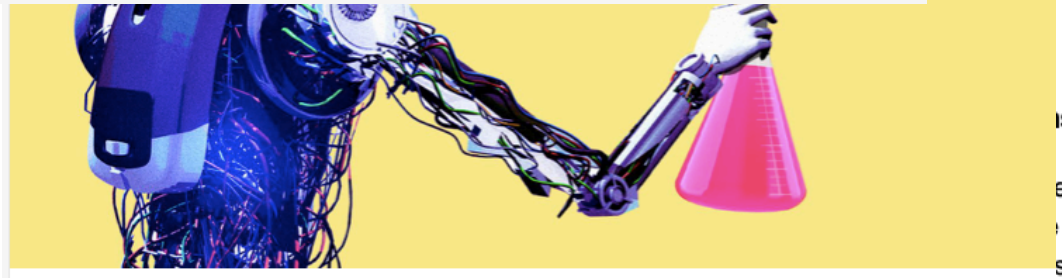


Illustration: Sarah Grillo/Axios

...eing

...s Researchers for AI-

myScience

...s in biological experiments

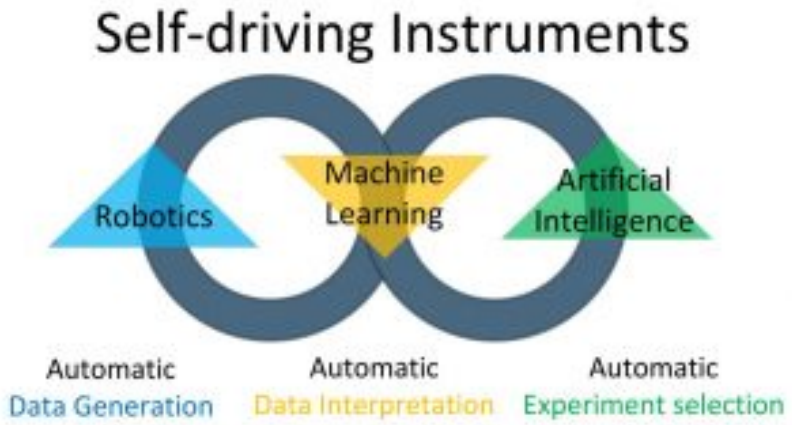
...entify and select experiments

... Mellon University has created a two-

...s needed to design, configure,

M.S. in Automated Science Curricular Goals

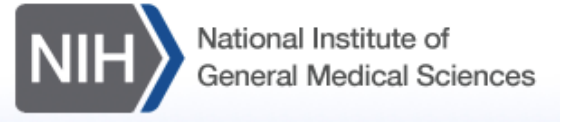
- Hands on training with automated equipment
- Experience creating predictive models from experimental data
- Expertise in active machine learning methods for using predictive models to choose future experiments



Acknowledgments

- Armaghan Naik
- Joshua Kangas
- Christopher Langmead
- Aarti Singh
- Nina Balcan
- Jeff Schneider
- Jaime Carbonell
- Andrew Moore

Supported by:



@murphy2537