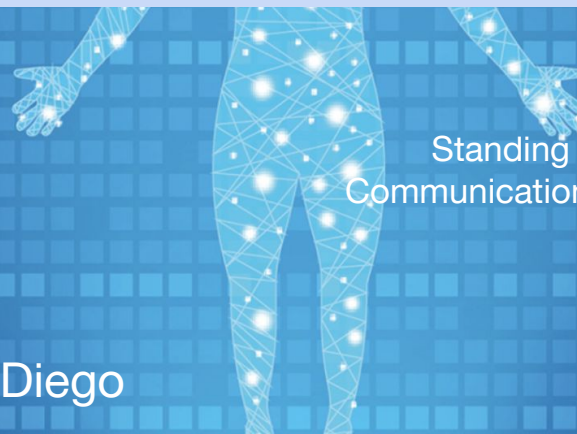




Systems Biology & Visible Machine Learning

Trey Ideker, PhD

Professor of Genetics
Department of Medicine
University of California San Diego



The National Academies
Standing Committee on Evidence Synthesis and
Communications in Diet and Chronic Disease Relationships

Declaration of Interests:
Data4Cure (coFounder, SAB)
Ideaya (SAB)

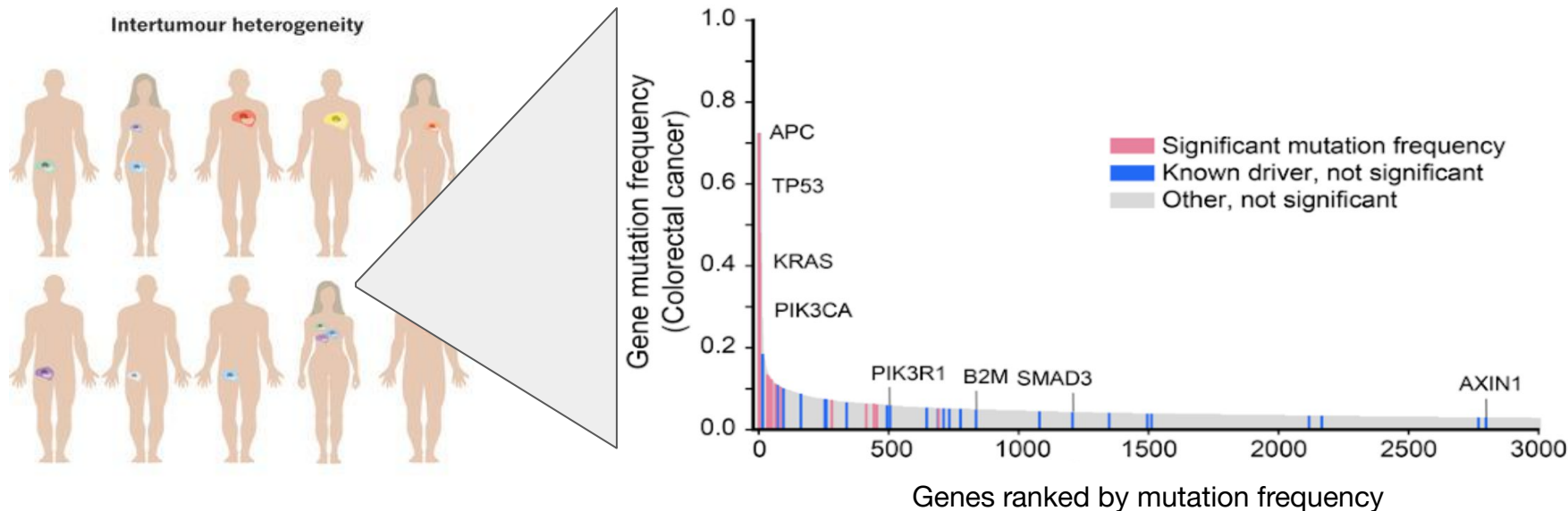




“Siri, my patient’s cells have genetic mutations in PTEN and DAAM1, what phenotypes should I expect?”

So why don't we have this technology?

Every genome is distinct, with many driving variants & mutations rare

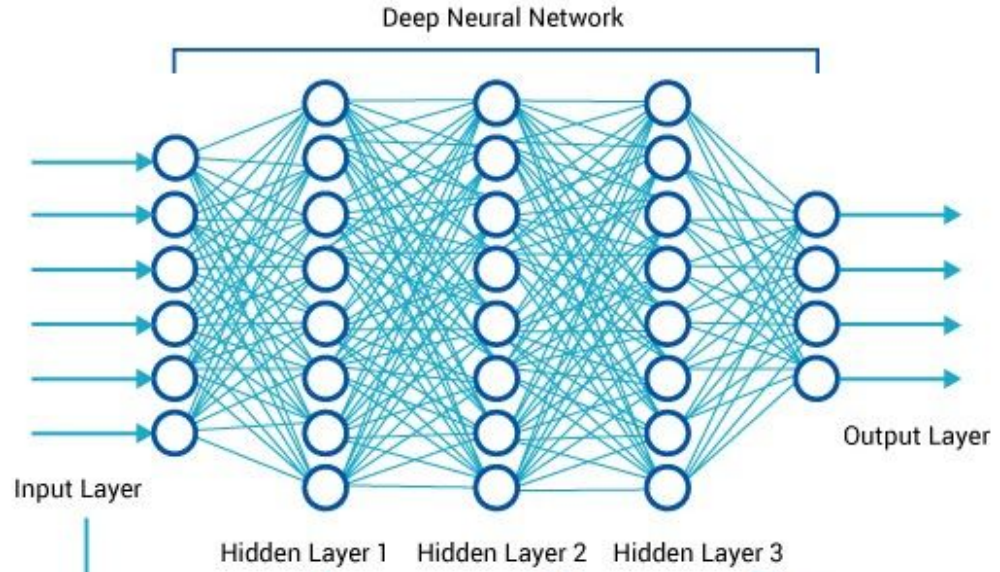


Separately, the fields of genomics & artificial intelligence are rapidly maturing, but they have not yet fully integrated.

Enormous opportunities (and pitfalls) lie in decoding the effects of genes on health & complex genetic diseases like diabetes & cancer

(Deep) Machine Learning is All the Rage for Learning to Translate Inputs to Outputs

- Game playing (Chess, Go, Jeopardy)
- Speech recognition & text processing
- Image recognition & computer vision
- Robotics & self-driving vehicles
- ...
- Biomedical imaging
- Chemoinformatics & drug discovery
- Proteomics and protein folding
- Gene expression & transcriptional binding
- **What about genotype to phenotype?**
(e.g. effect of diet on BMI & diabetes, chronic diseases like cancer)



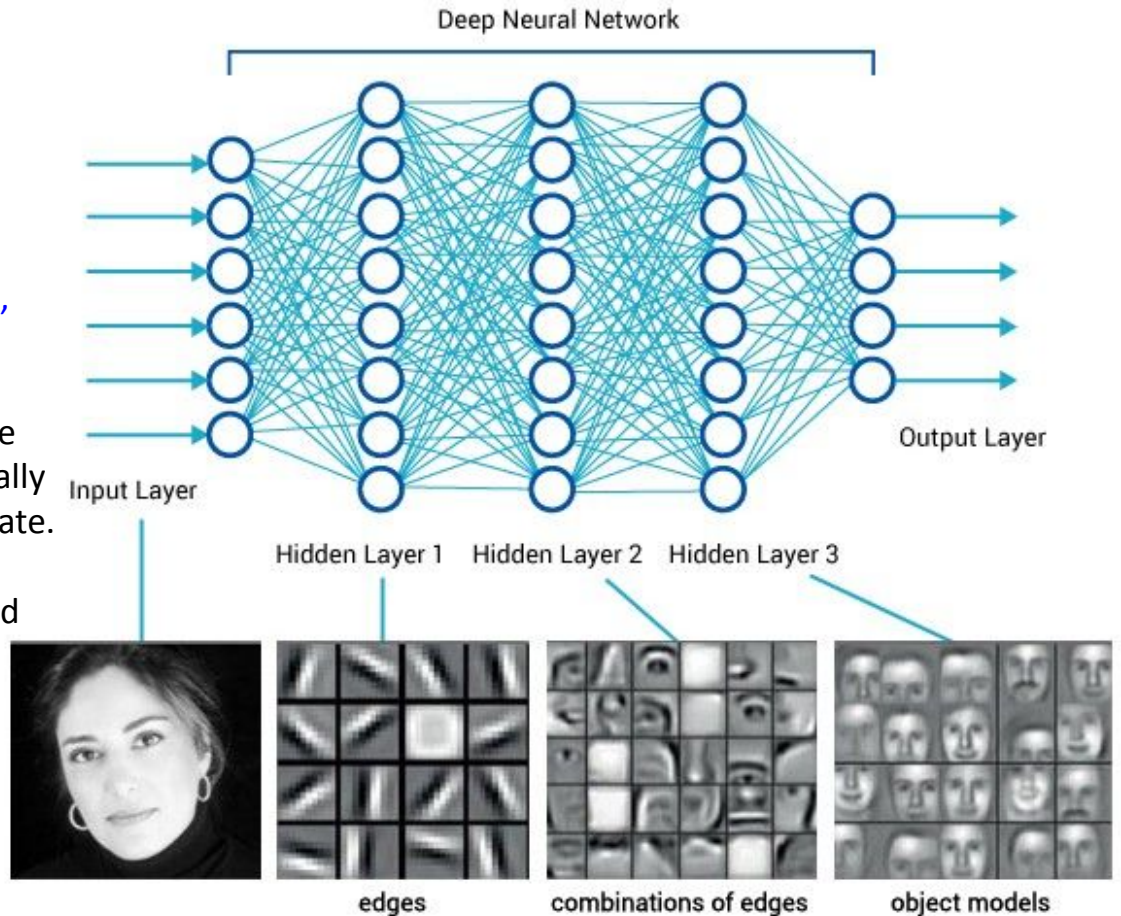
Input: Profiles of individuals
(here faces, but also **molecular profiles of genetic variants and other 'omic changes**)

Output: Type of individual
(here identity, but also **health/disease status, disease subtype, response to therapy**)

Hidden Layers: Between input and output are “hidden” layers of artificial neurons (potentially many), each having a numerical activation state.

Weights: Neuron values are sent *via* weighted connections to neurons in the next layer

Progressive layers capture increasingly abstract understanding of the input



Vast diversity in machine learning problems

Learning Problems

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Statistical Inference

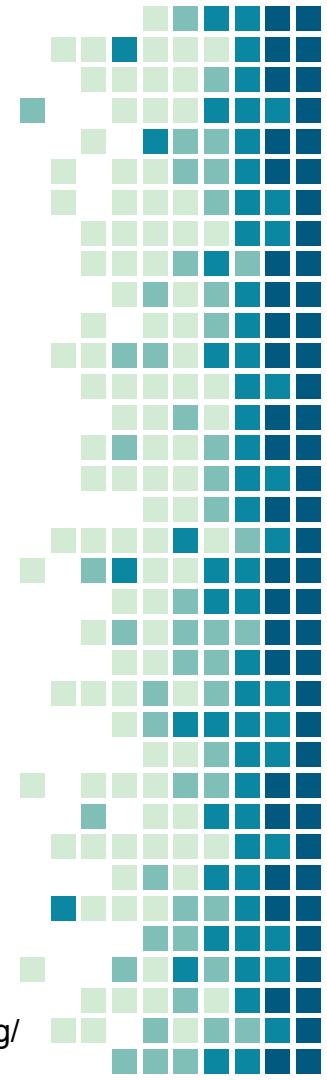
- Inductive Learning
- Deductive Inference
- Transductive Learning

Hybrid Learning Problems

- Semi-Supervised Learning
- Self-Supervised Learning
- Multi-Instance Learning

Learning Techniques

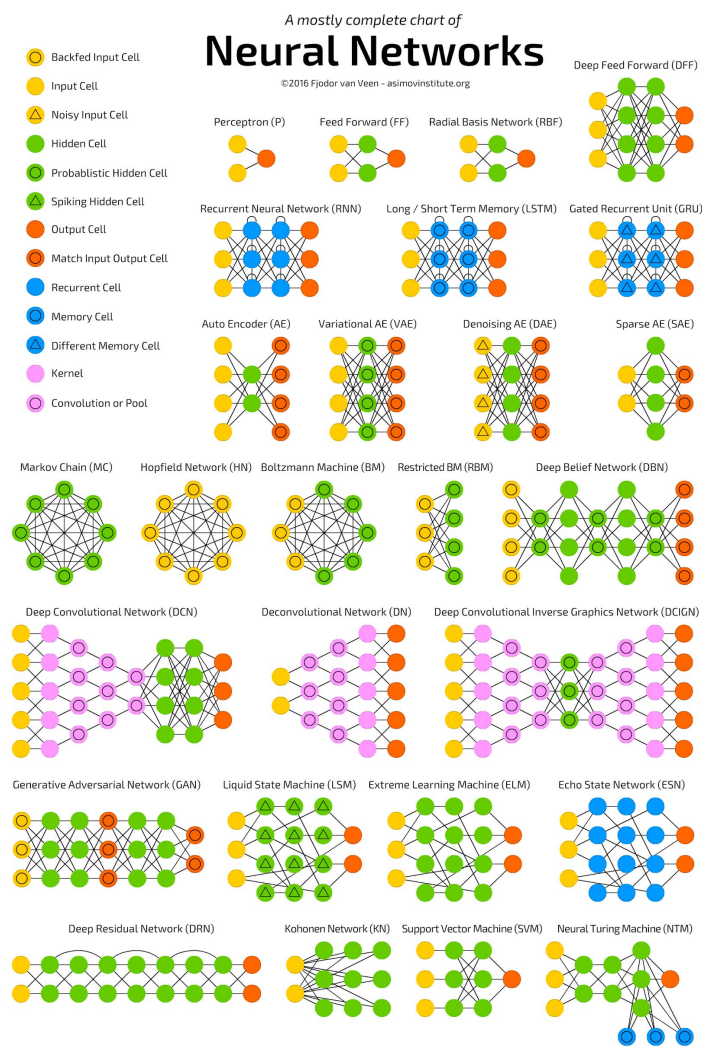
- Multi-Task Learning
- Active Learning
- Online Learning
- Transfer Learning
- Ensemble Learning



... and evolving approaches

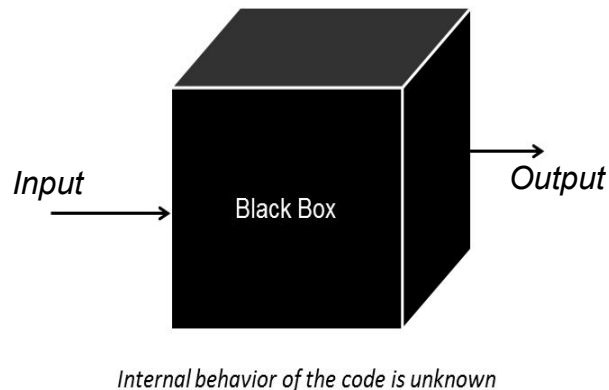
- Classic linear regression
- Ridge regression, LASSO, Elastic net
- Random forest
- Support Vector Machine
- K-nearest-neighbor
- Neural network
 - Conventional
 - Convolutional
 - Generative adversarial network
 - Long short-term memory
 - Recurrent
- Many, many others

<https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464>



Challenge 1:

What's wrong with “black box” learning?



- Modern ‘black box’ machine learning methods model the input/output function of a system
- They do not model its internal structure or function
- This aspect makes models difficult to interpret and also harder to validate and generalize
- **In contrast, high-stakes applications, including those in biomedicine, seek to understand internal mechanisms and reasons for prediction**

Cynthia Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 206–215(2019)

Example: Dogs on the Internet

Google

dog



All Images News Videos Shopping More

Settings Tools

Collections SafeSearch



The Best Dogs of BBC Earth | Top 5 ...
youtube.com



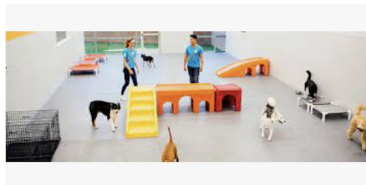
Dog Boarding Facilities in Temecula ...
careah.com



scientists explain puppy dog eyes ...
theguardian.com



Dog - Wikipedia
en.wikipedia.org



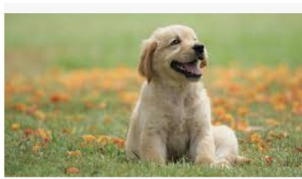
Dog Daycare, Boarding, Spa Temecula ...
dogtopia.com



How to Keep Your Dog Cool in the S...
pets.webmd.com



What animal welfare problem...
kb.rspca.org.au



The 25 Cutest Dog Breeds - Most ...
goodhousekeeping.com



convert dog years to human years ...
sciencemag.org



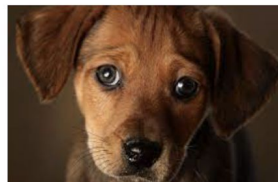
How dogs contribute to your health and ...
medicalnewstoday.com



9 reasons to own a dog
businessinsider.com



Most Popular Breeds - American Ken...
akc.org



But does the machine really understand what is a dog?

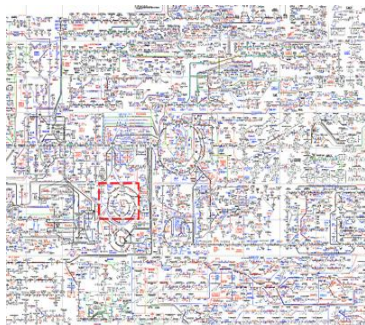


Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier, arXiv:1602.04938v3 9 Aug 2016
<http://becominghuman.ai/its-magic-i-owe-you-no-explanation-explainableai-43e798273a08>

-

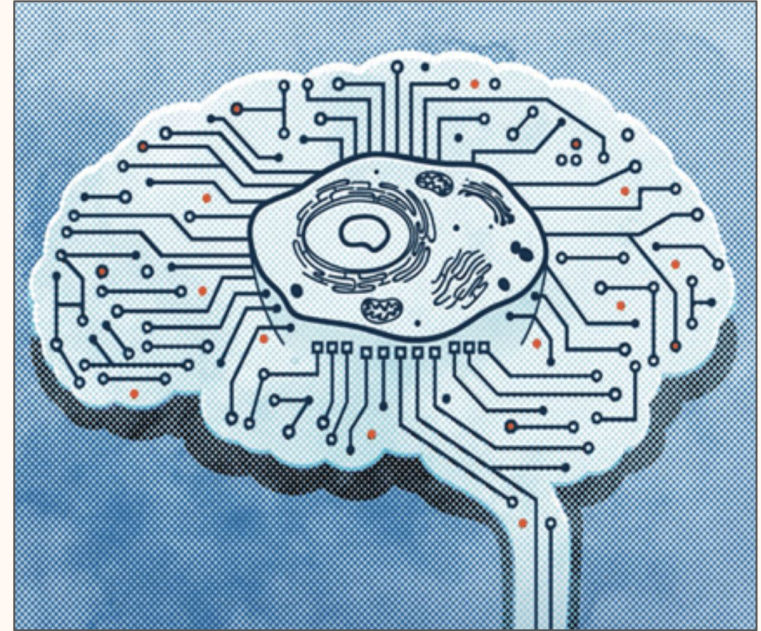
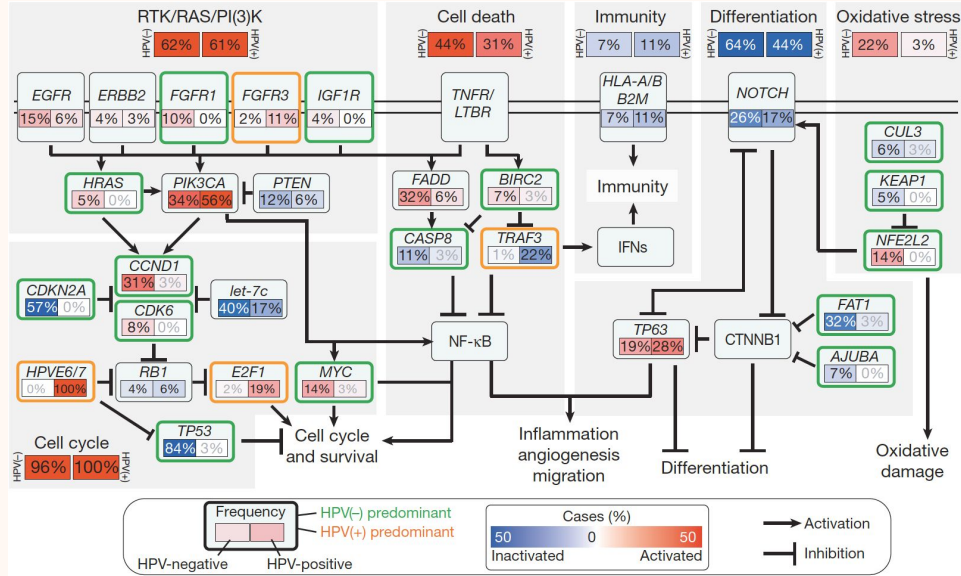


3. Biological complexity exceeds that of playing Go or driving cars



- Machine learning has become very powerful at simulating human tasks in specific domains
- Examples: game playing (Chess, Go), image and speech recognition, driving
- These tasks are governed by a small set of rules
- **Biomedical research is mostly about discovering things humans don't yet know or do**
- **In these cases, we don't know most of the rules**

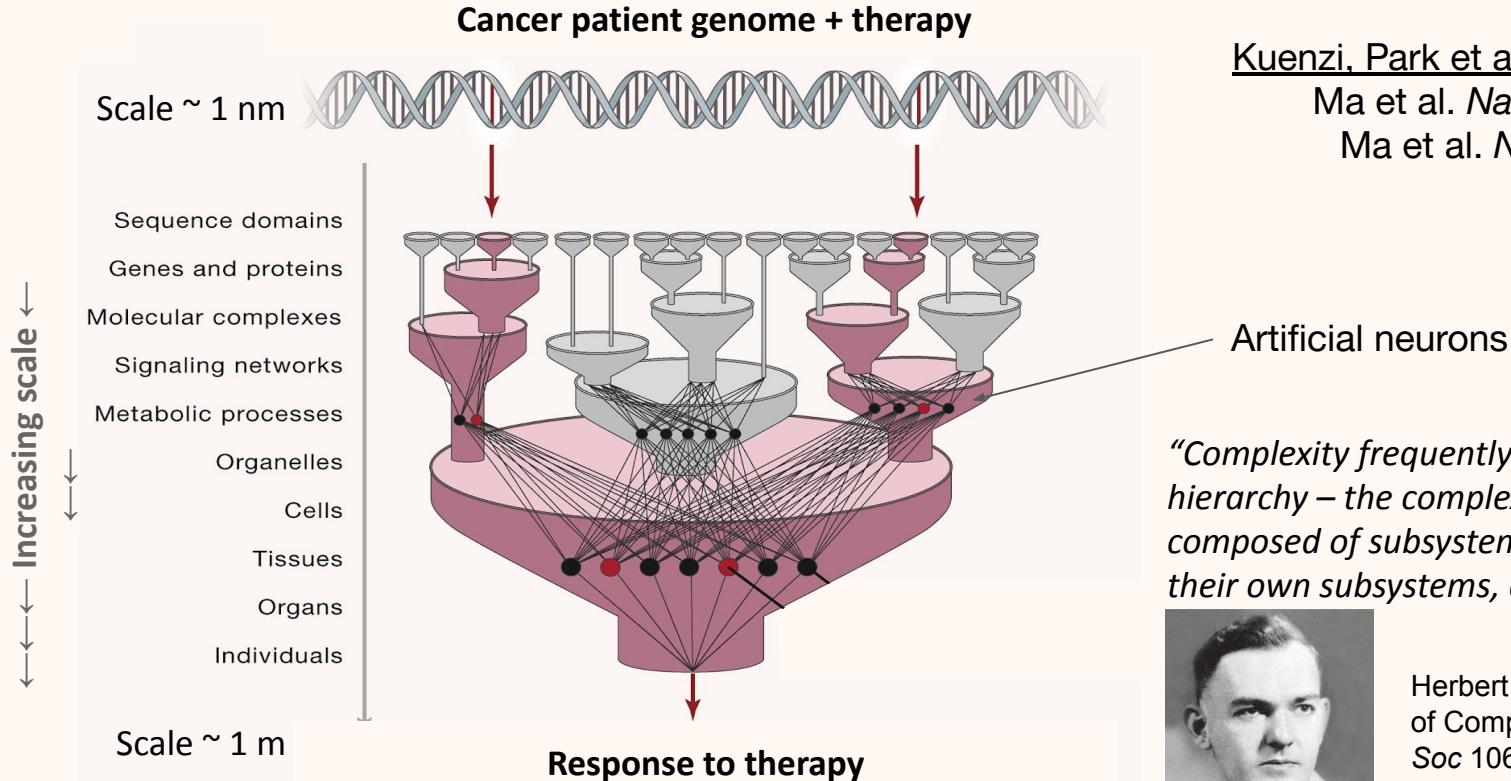
Combining two key approaches for biological machine learning



1. Convergence of molecular events in hierarchical pathway maps

2. Pattern recognition via deep learning

An AI to translate cancer genomes to therapies, based on the key principle of *genetic convergence*

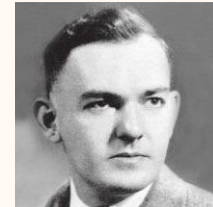


Kuenzi, Park et al. *Cancer Cell* (2020)

Ma et al. *Nature Methods* (2018)

Ma et al. *Nature Cancer* (2021)

“Complexity frequently takes the form of hierarchy – the complex system being composed of subsystems that, in turn, have their own subsystems, and so on.”



Herbert Simon. The Architecture of Complexity *Proc Amer Phil Soc* 106:6 467-482 (1962)

Building deep learning models of cancer cells and their genotype/drug interactions

Key application: matching pts to drugs

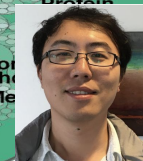
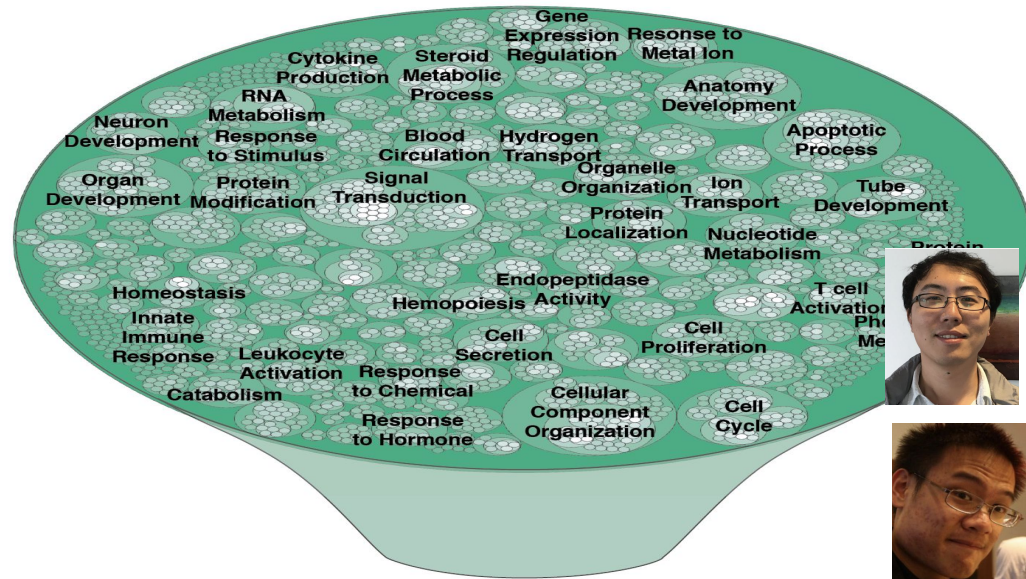
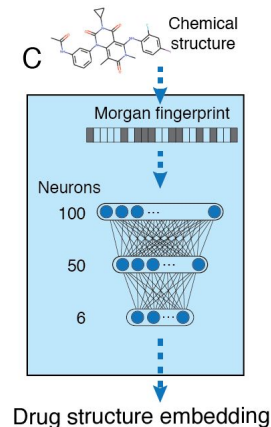
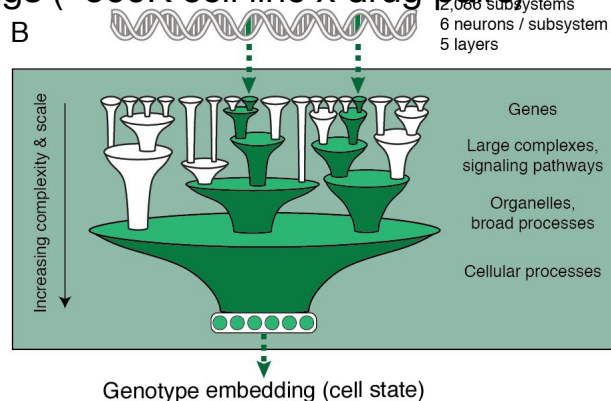
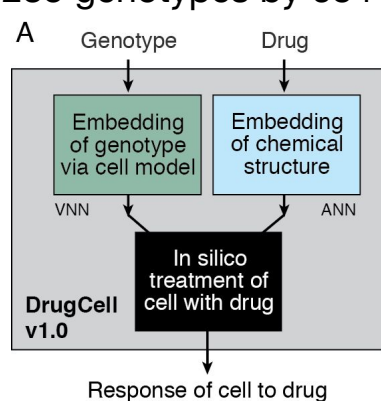
Kuenzi, Park et al. *Cancer Cell* (2020)

Ma, Yu et al. *Nature Methods* (2018)

Ma et al. *Nature Cancer* (2021)

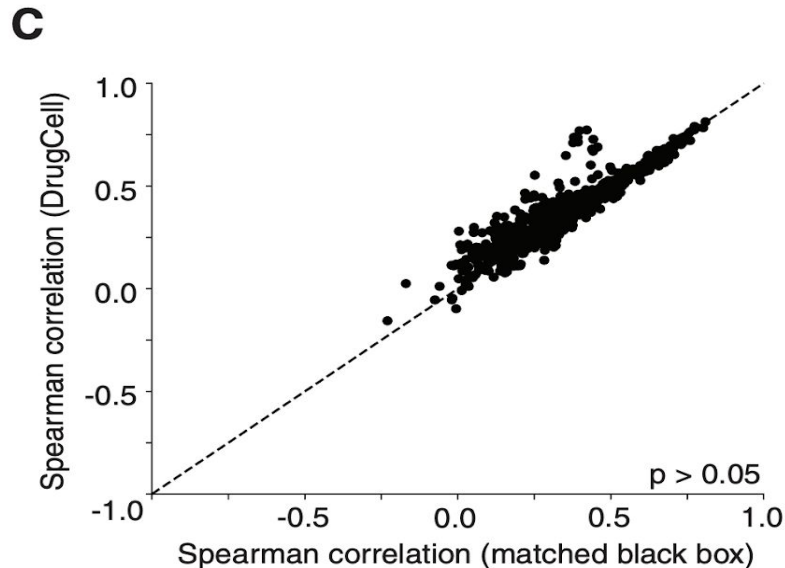
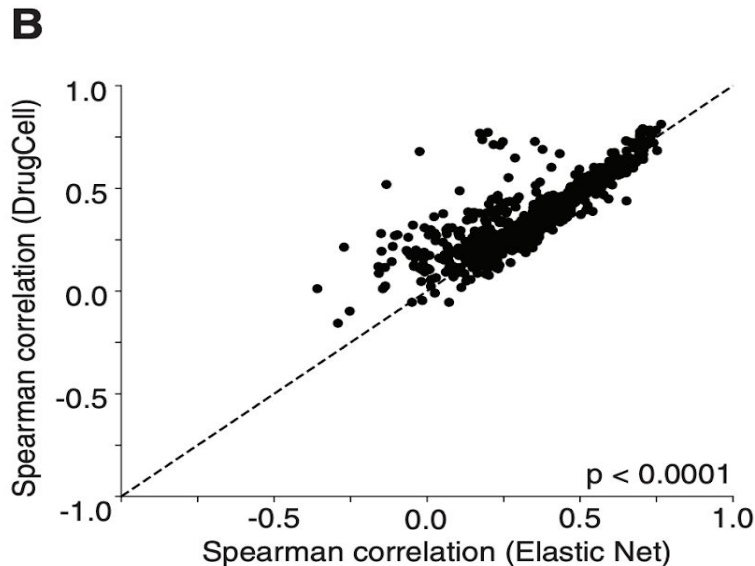
Training data: CTRP & GDSC

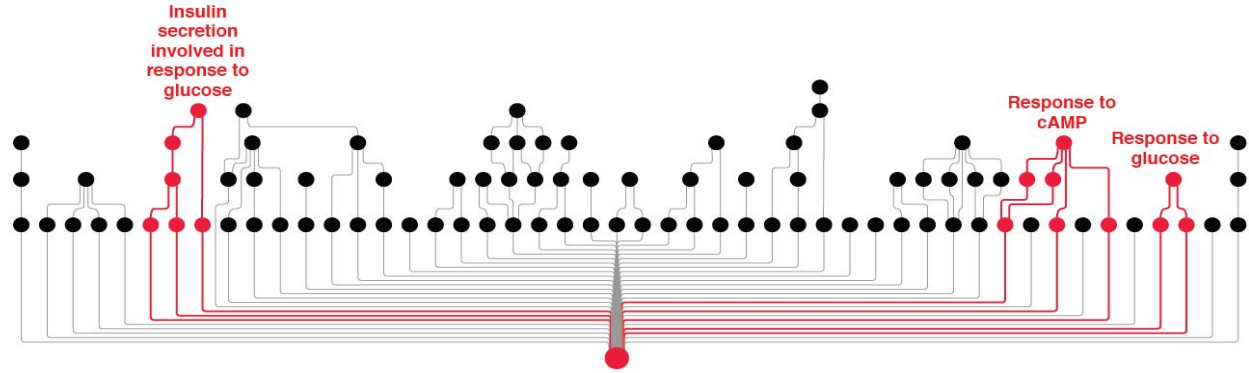
1,235 genotypes by 684 drugs (~500K cell line x drug pairs)



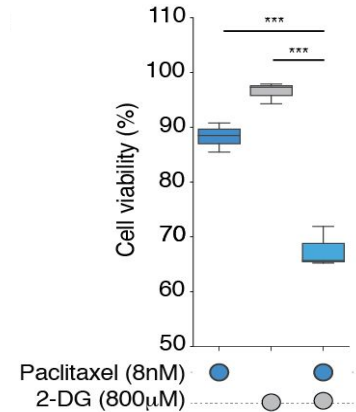
Jianzhu Ma
Mike Yu
Jisoo Park
Brent Kuenzi

Predictions of drug response are comparable or slightly better than linear models or black box neural networks



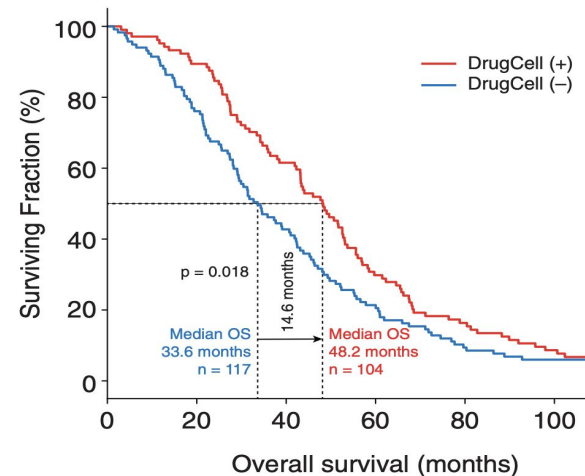
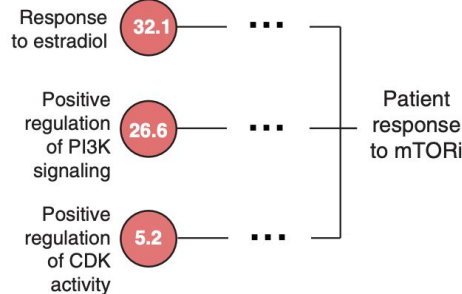
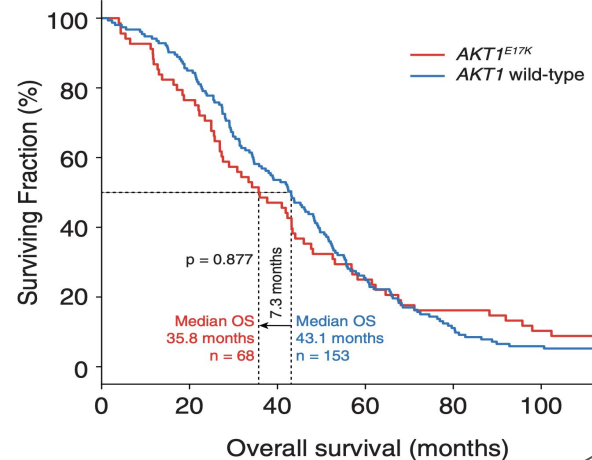


Glycolysis mediates response to paclitaxel, suggesting a combination treatment (glucose mimetics, 2-DG)



Application to clinical trial data for ER+ metastatic breast cancer: Fulvestrant (ER antagonist) +/- Everolimus (mTOR inhibitor)

PI3K → AKT → mTOR → tumor progression



Response mutations
DrugCell(+)

AKT

Non-response mutations
DrugCell(-)

CDH1, TP53, ESR1,
ATM, NTRK1, RUNX1,
CBFB, CREBBP, PIK3CA

AACR GENIE project
Smyth et al. 2020

THE END



Funding

National Institutes of Health - NIDA, NCI, NIMH, NHGRI, NIEHS
Industry - Pfizer, Merck, Ideaya
Private - Luddy Fdn, Roddenberry Fdn

Resources

Networks: NDEx, PCNet
Network visualization: Cytoscape
Hierarchies: DDOT Toolkit