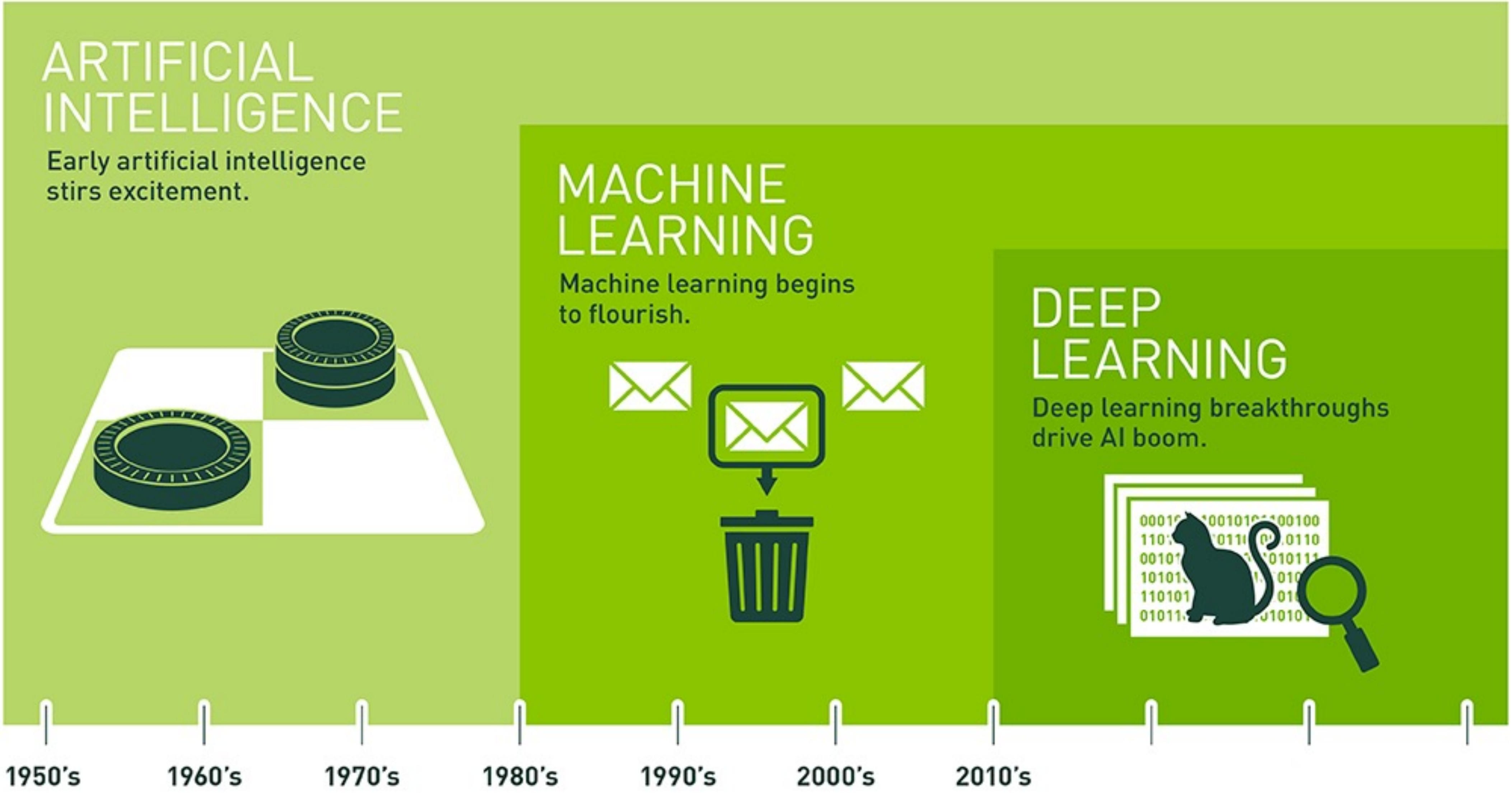
Introduction to AI in life science

Bin Chen
Associate Professor
Dept. of Pediatrics and Human Development
Dept. of Pharmacology and Toxicology
College of Human Medicine
Michigan State University
Bin.Chen@hc.msu.edu @DrBinChen
http://binchenlab.org

July 2024

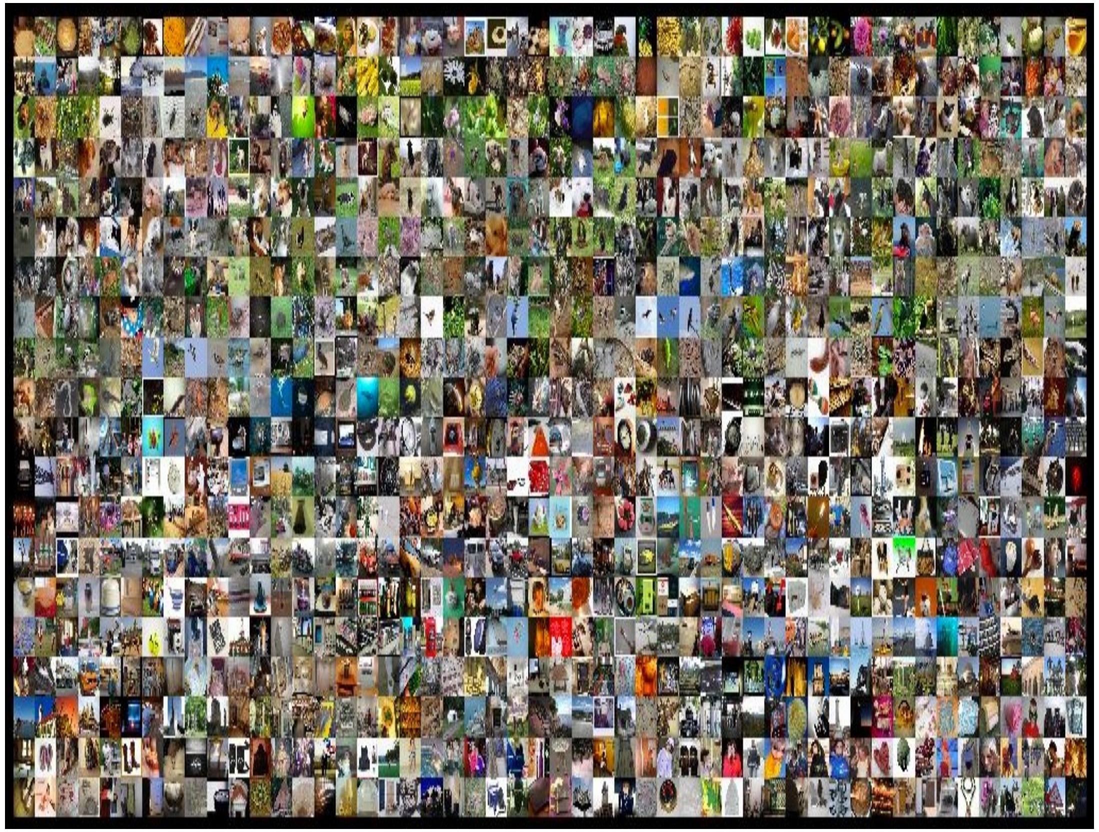

ALPHAGO
LEE SEDOL
01:59:31
AlphaGo
Lee Sedol

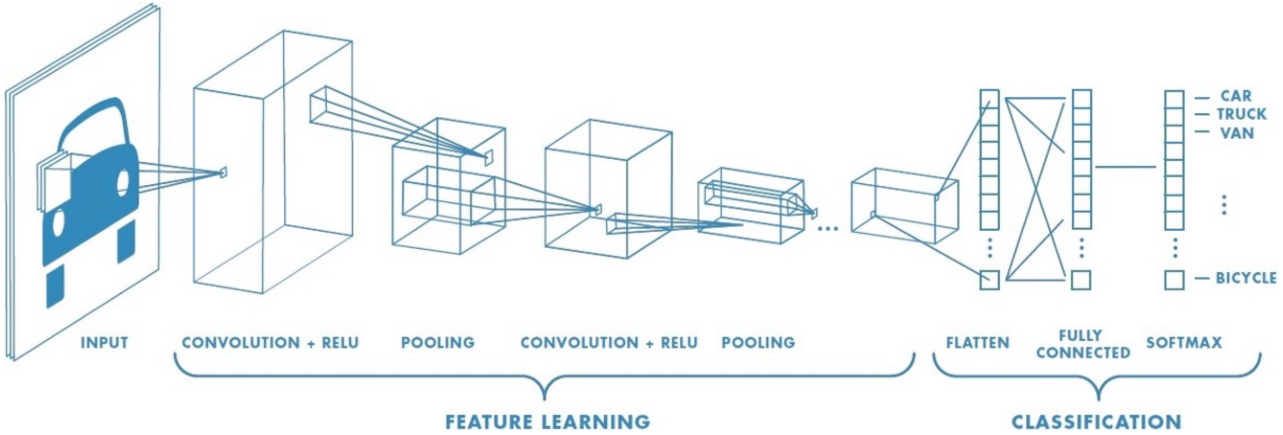


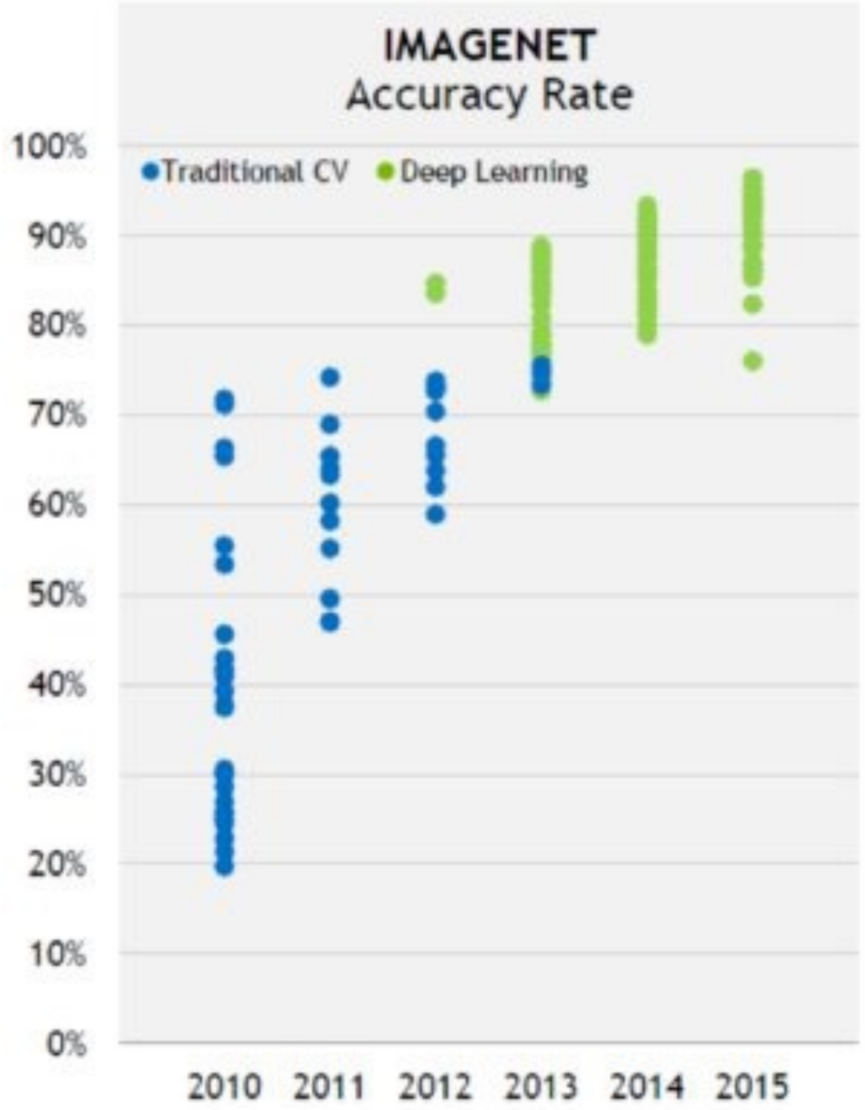
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://goo.gl/qJ7U32

IMAGE-NET

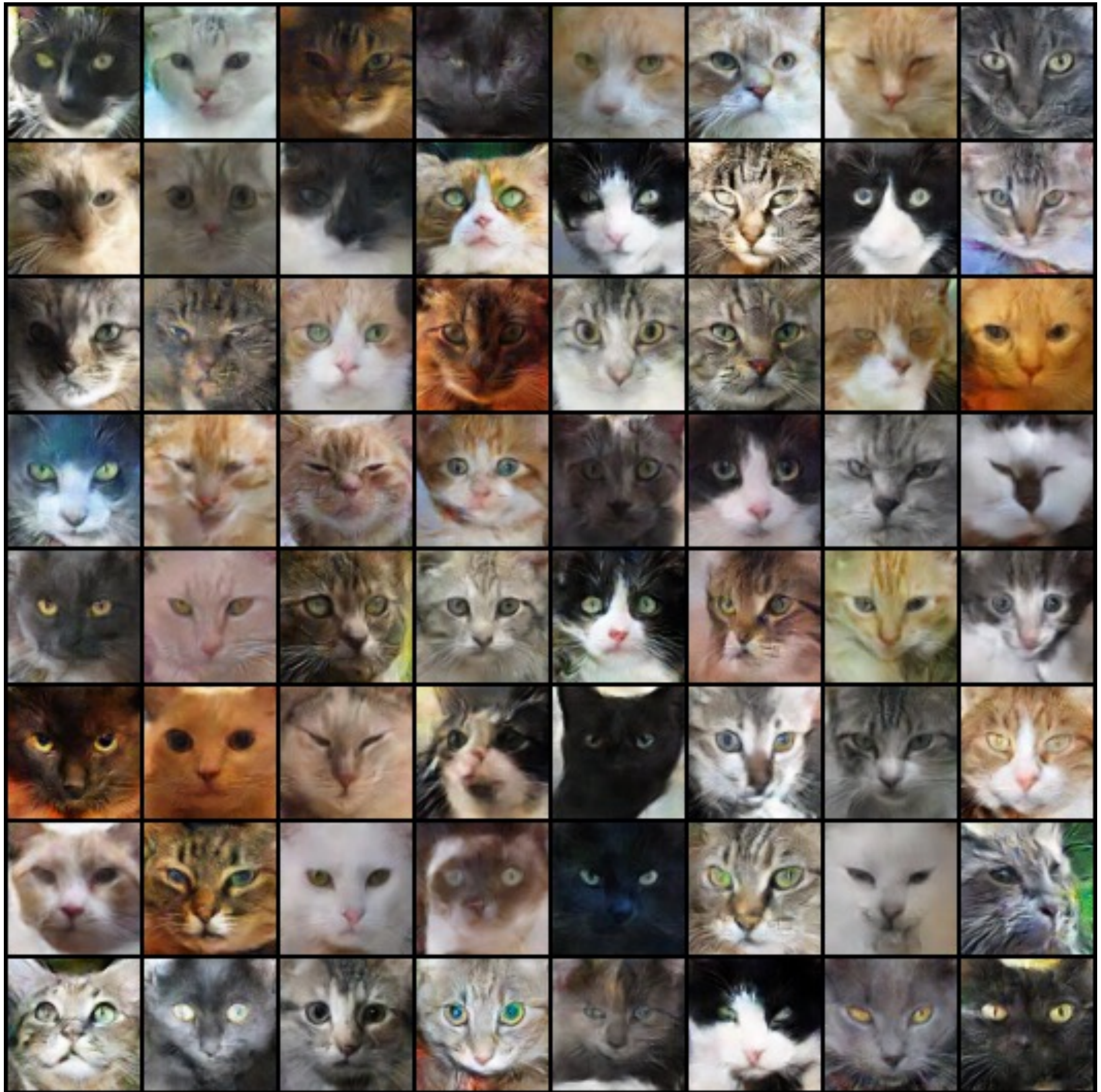




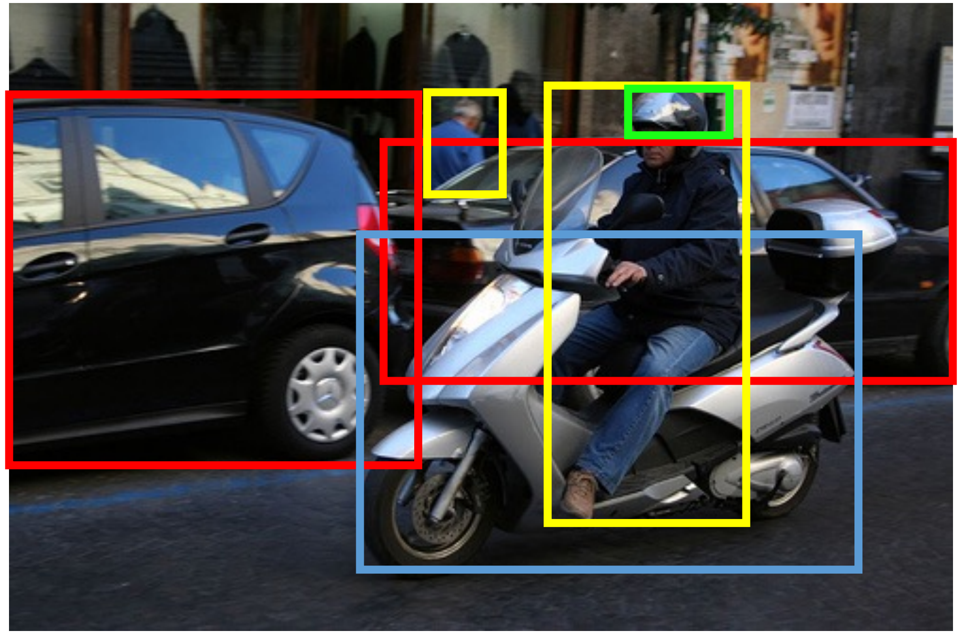


Deep Neural Network

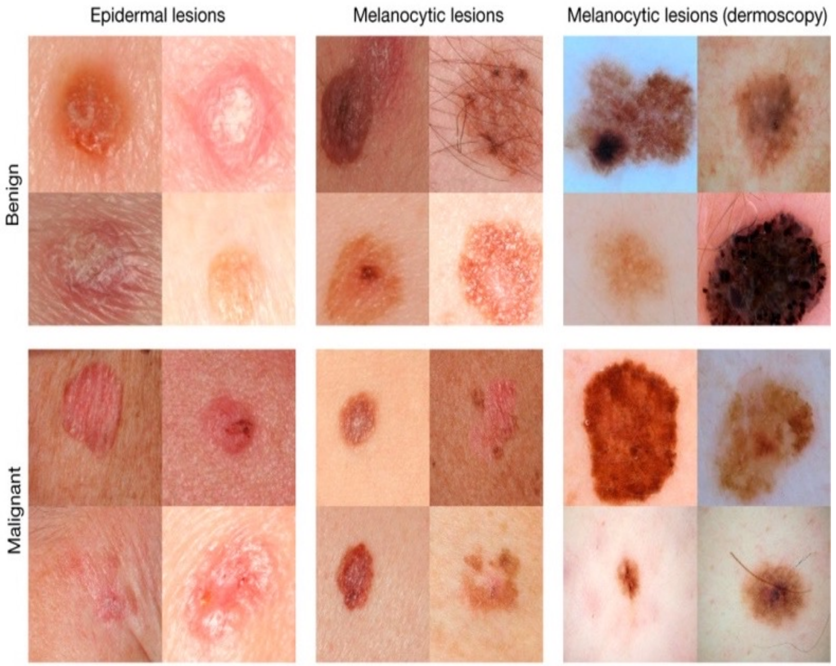
http://www.image-net.org/

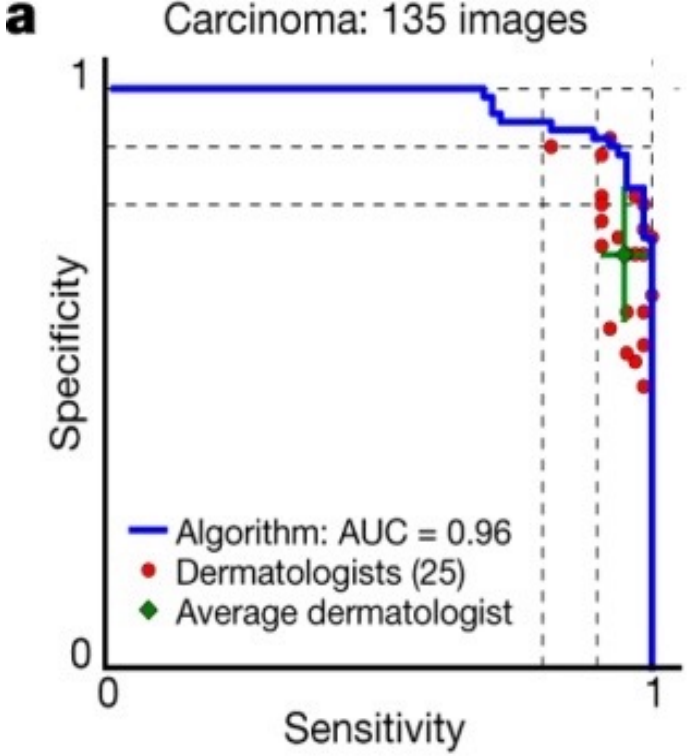


https://junyanz.github.io/CycleGAN/



Person
Car
Motorcycle
Helmet



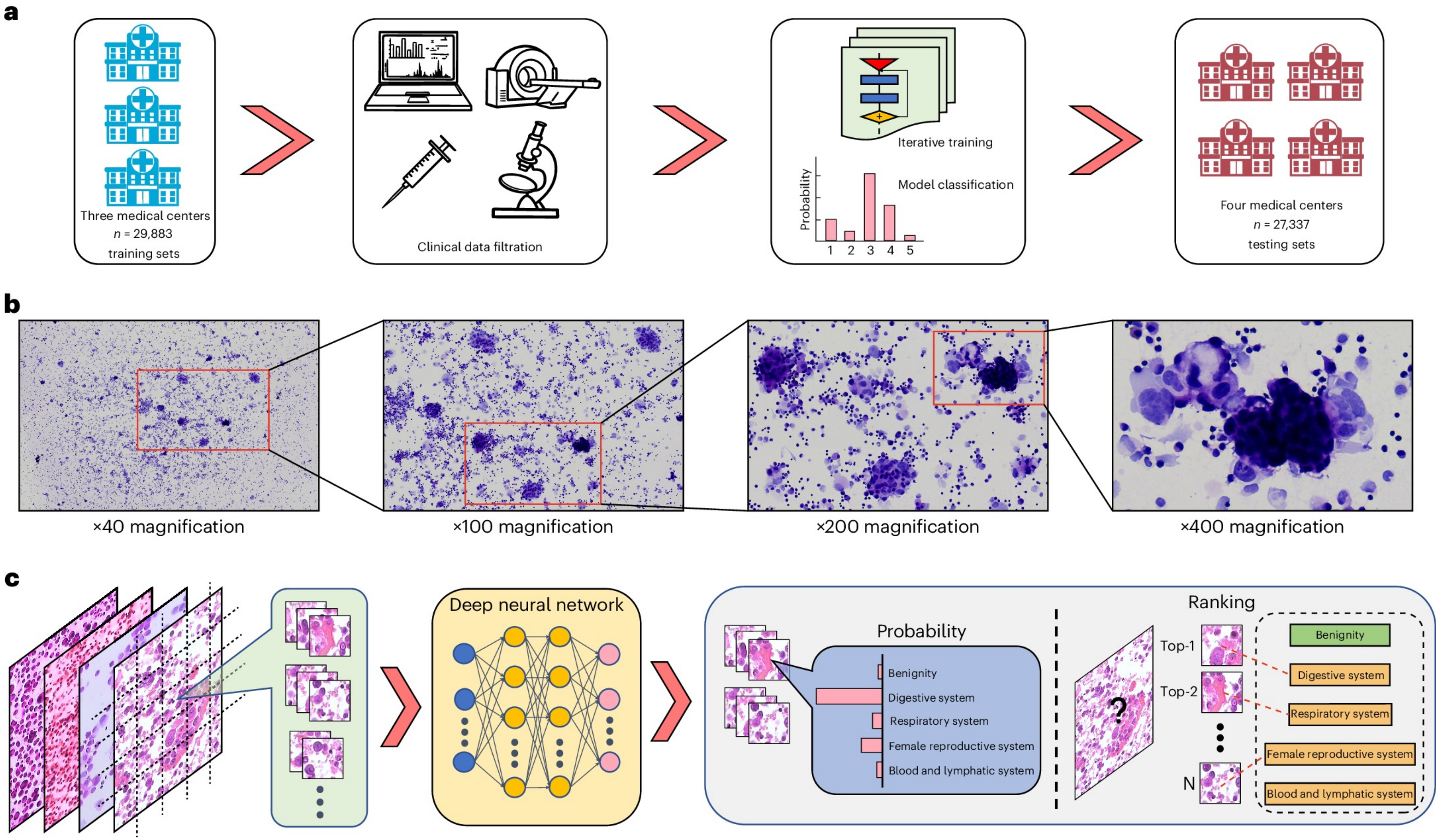


Andre Esteva, Nature, 2017

Specialty	Images	Publication
Radiology	CT head for brain hemorrhage	Arbabshirani, Nature Digital Medicine, 2018
Pathology	Breast cancer	Bejnordi, JAMA,2017
	Brain tumors	Capper, Nature 2018
Dermatology	Skin cancers	Esteva, Nature 2017
	Melanoma	Haenssle, Annals of Oncology, 2018
Ophthalmology	Diabetic retinopathy	Gulshan, JAMA 2016
	Congenital cataracts	Long, Nature Biomedical Engineering 2017
	Macular degeneration	Burlina, JAMA Ophthalmology, 2018
	Retinopathy of Prematurity	Brown, JAMA Ophthalmology, 2018
	AMD and diabetic retinopathy	Kermany, Cell, 2018
Cardiology	Echocardiography	Madani, Nature Digital Medicine, 2018

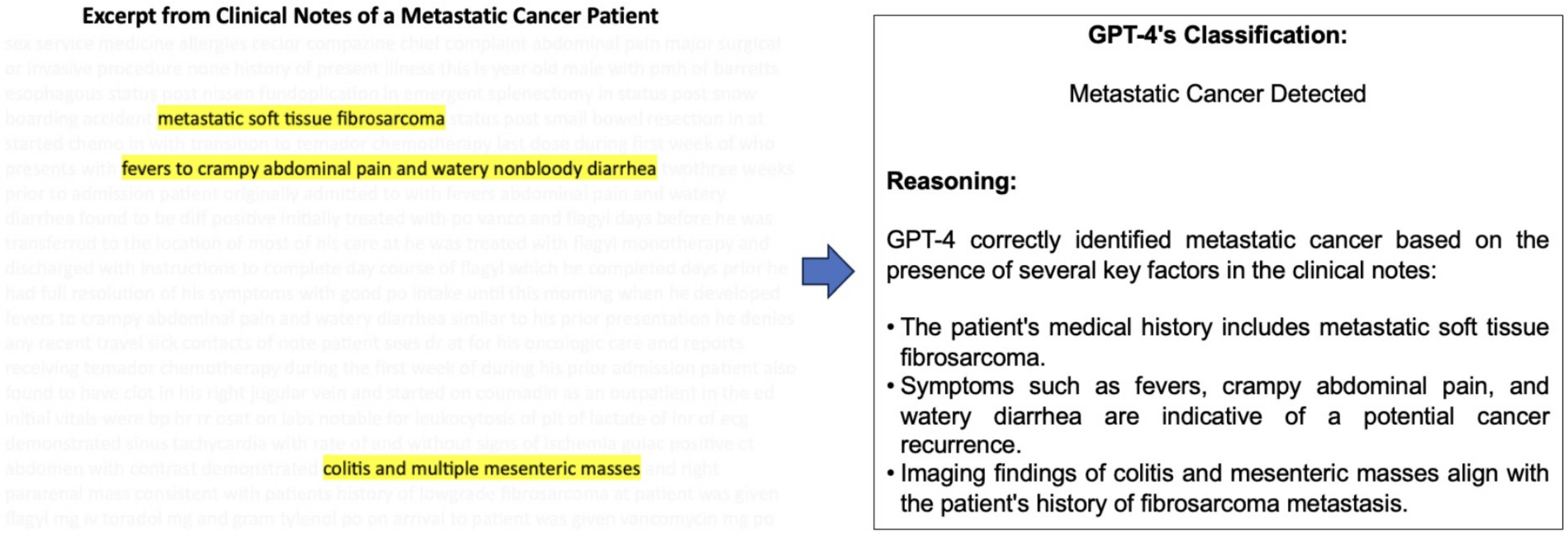
Eric Topol's tweet

Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning

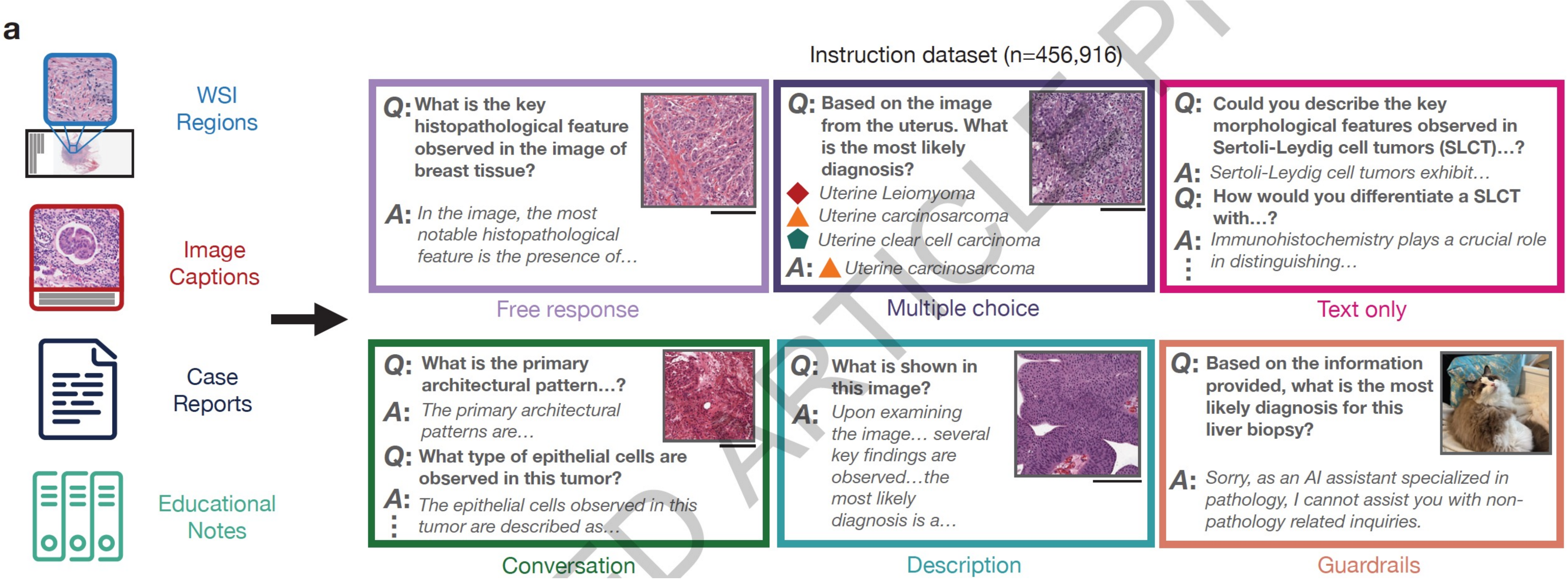


Tian, 2024, Nature Medicine


OpenAI
CHATGPT

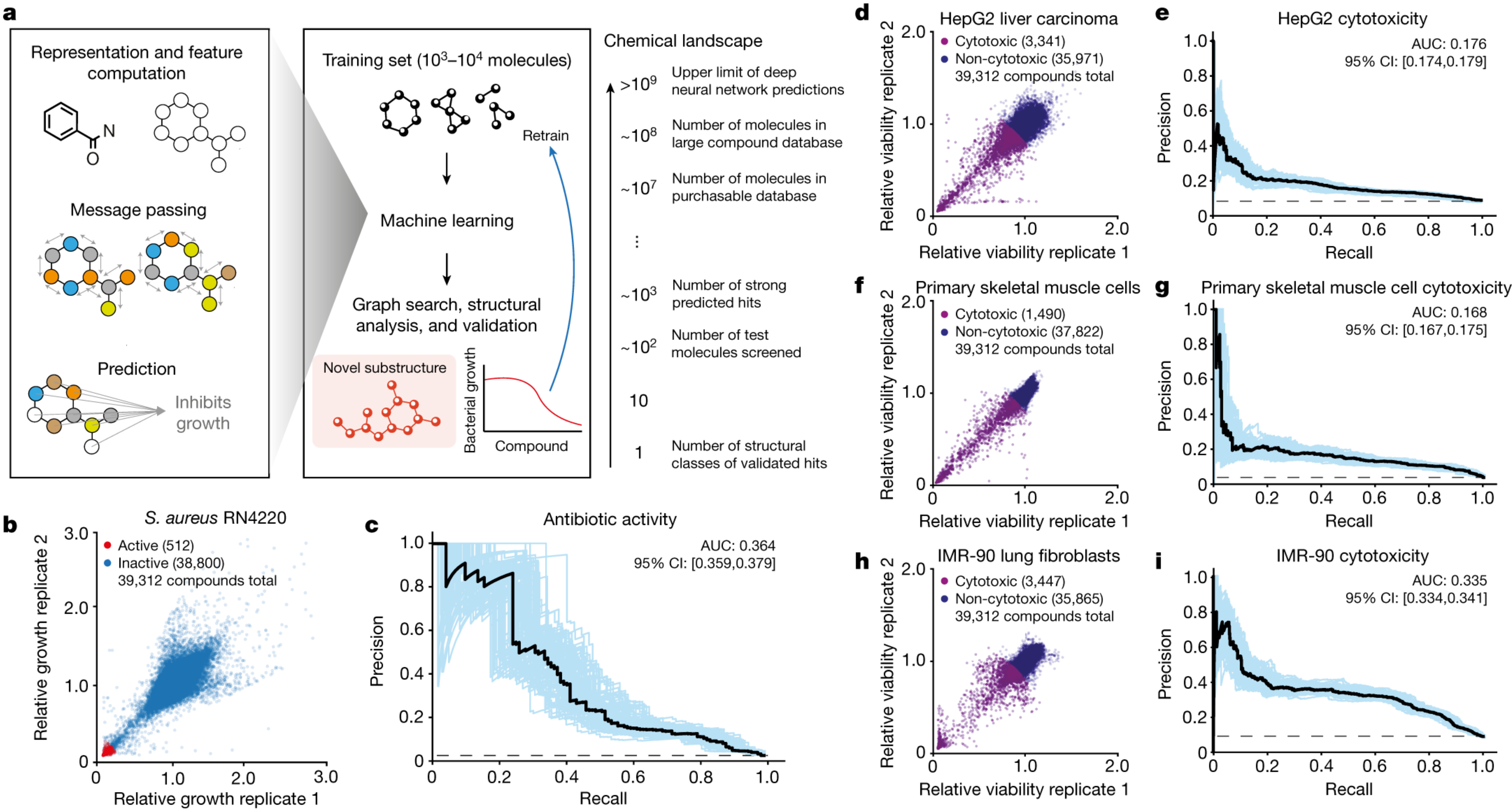

Excerpt from Clinical Notes of a Metastatic Cancer Patient
sex service medicine allergies ceclor compazine chief complaint abdominal pain major surgical or invasive procedure none history of present illness this is year old male with pmh of barretts esophagous status post nissen fundoplication in emergent splenectomy in status post snow boarding accident metastatic soft tissue fibrosarcoma status post small bowel resection in at started chemo in with transition to temador chemotherapy last dose during first week of who presents with fevers to crampy abdominal pain and watery nonbloody diarrhea twothree weeks prior to admission patient originally admitted to with fevers abdominal pain and watery diarrhea found to be diff positive initially treated with po vanco and flagyl days before he was transferred to the location of most of his care at he was treated with flagyl monotherapy and discharged with instructions to complete day course of flagyl which he completed days prior he had full resolution of his symptoms with good po intake until this morning when he developed fevers to crampy abdominal pain and watery diarrhea similar to his prior presentation he denies any recent travel sick contacts of note patient sees dr at for his oncologic care and reports receiving temador chemotherapy during the first week of during his prior admission patient also found to have clot in his right jugular vein and started on coumadin as an outpatient in the ed initial vitals were bp hr rr osat on labs notable for leukocytosis of plt of lactate of inr of ecg demonstrated sinus tachycardia with rate of and without signs of ischemia guiac positive ct abdomen with contrast demonstrated colitis and multiple mesenteric masses and right pararenal mass consistent with patients history of lowgrade fibrosarcoma at patient was given flagyl mg iv toradol mg and gram tylenol po on arrival to patient was given vancomycin mg po
GPT-4's Classification:
Metastatic Cancer Detected
Reasoning:
GPT-4 correctly identified metastatic cancer based on the presence of several key factors in the clinical notes:
• The patient's medical history includes metastatic soft tissue fibrosarcoma.
• Symptoms such as fevers, crampy abdominal pain, and watery diarrhea are indicative of a potential cancer recurrence.
• Imaging findings of colitis and mesenteric masses align with the patient's history of fibrosarcoma metastasis.

A Multimodal Generative AI Copilot for Human Pathology

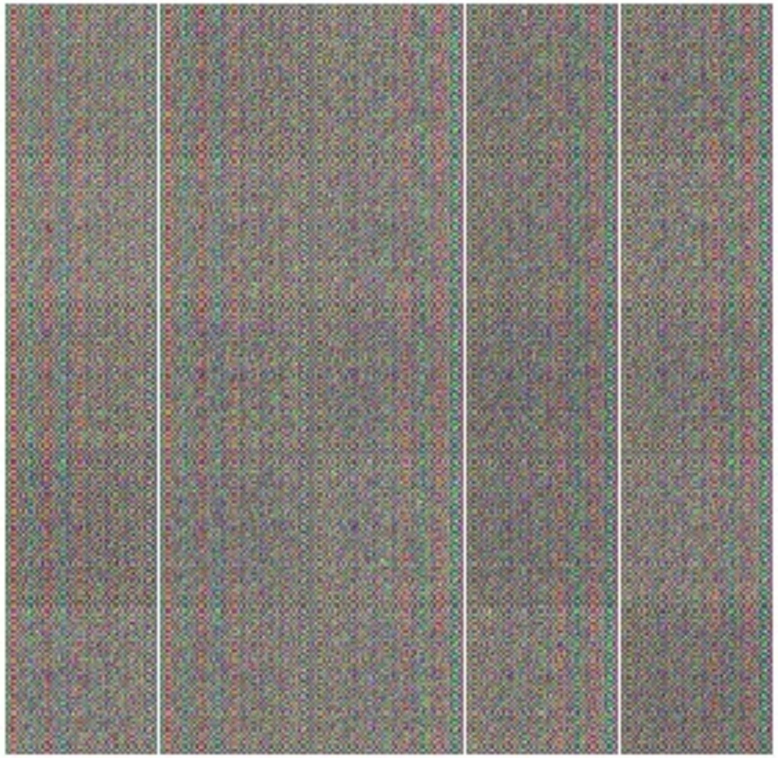


Lu, Nature, 2024

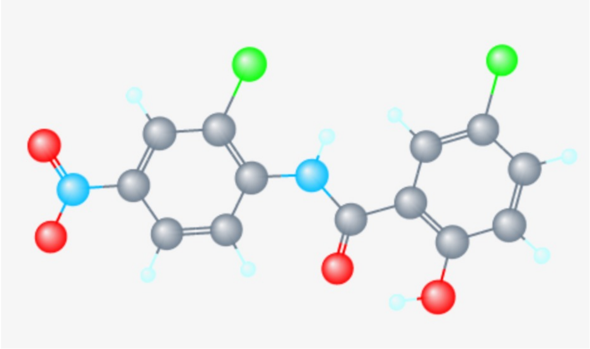
Discovery of a structural class of antibiotics with explainable deep learning (Nature, 2024)



Automated chemical structure generation



Epoch 1



Keith Schmitt

What’s machine learning

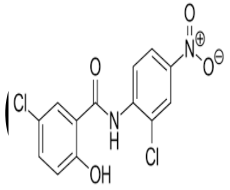
$$Y = f(X)$$

X: input data (e.g., chemical structure)
Y: output labeled data (e.g., solubility, biological activities)
f: parameters learned from the data by machine learning (e.g., beta in linear regression)

The goal of developing a machine learning model is to search the optimal parameters from the training data so that those parameters could be used to predict new data.

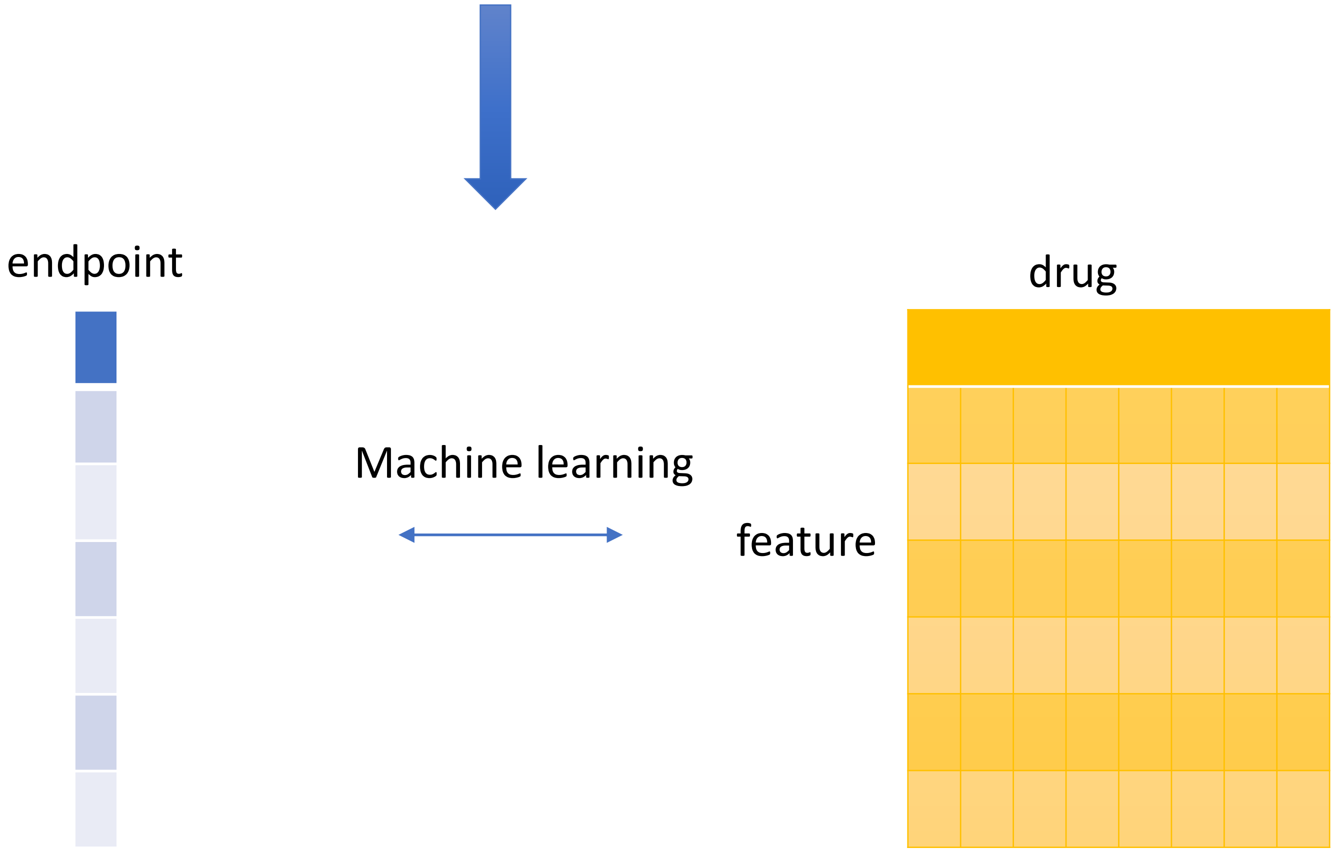
endpoints

Cytotoxicity
Genotoxicity
Mitochondrial toxicity
hERG channel assay
Nuclear receptor assay (PPAR)
Cell Viability
EGFR binding
…

Safety /Efficacy = f()

drug properties

Structural fingerprint
Physical properties
Pharmacophore
Drug-targets
Drug-induced gene expression
…



Model evaluation

- Split the data into training set and test set
- Classification
  - Accuracy, Precision, Recall, Area under the ROC curve
- Regression
  - Mean squared error, correlation

Linear/logistic model

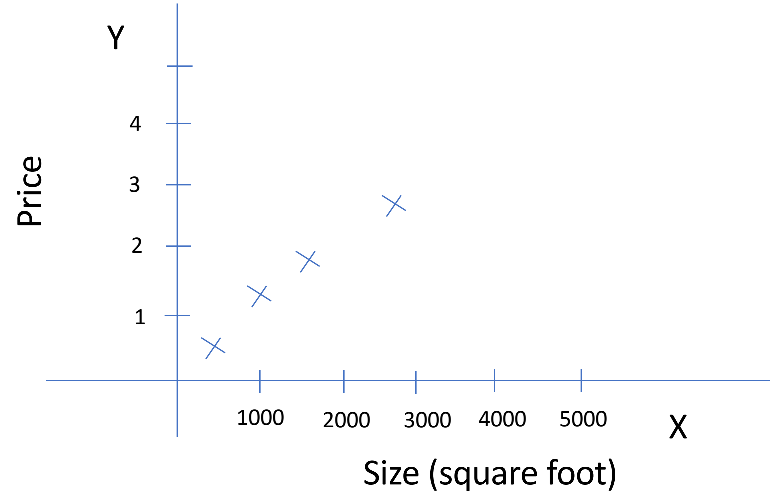
House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

2500?

Linear/logistic model

House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

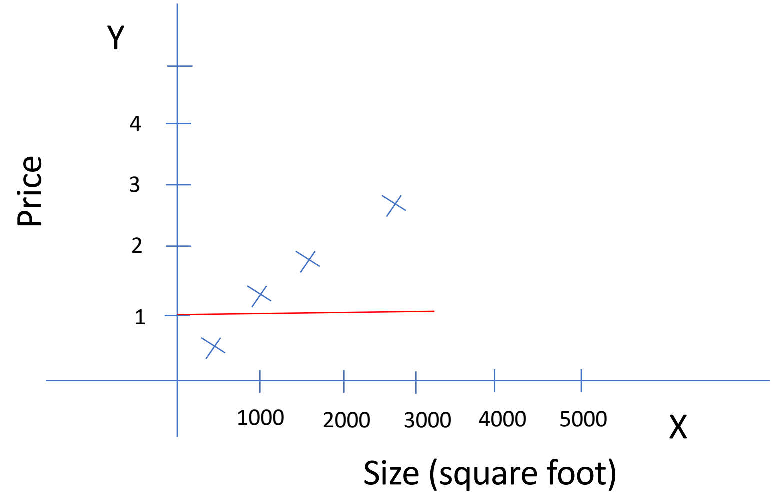
2500?



Linear/logistic model

House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

2500?



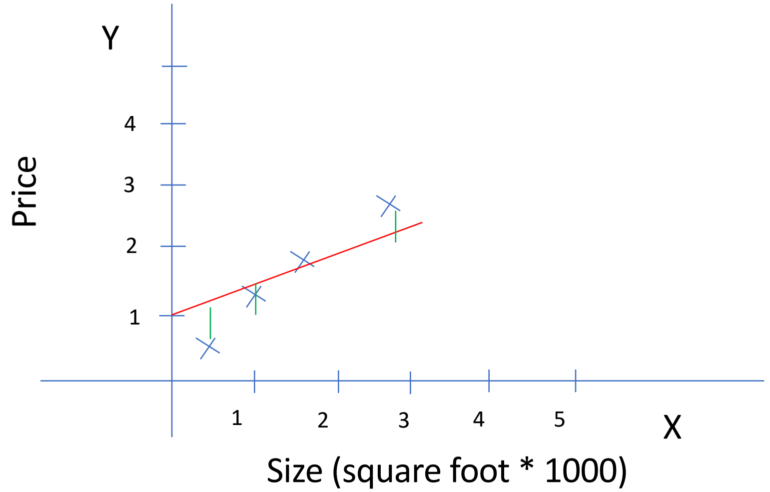
y = 1

Error = sum(y predict – y real)^2 = 2

Linear/logistic model

House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

2500?



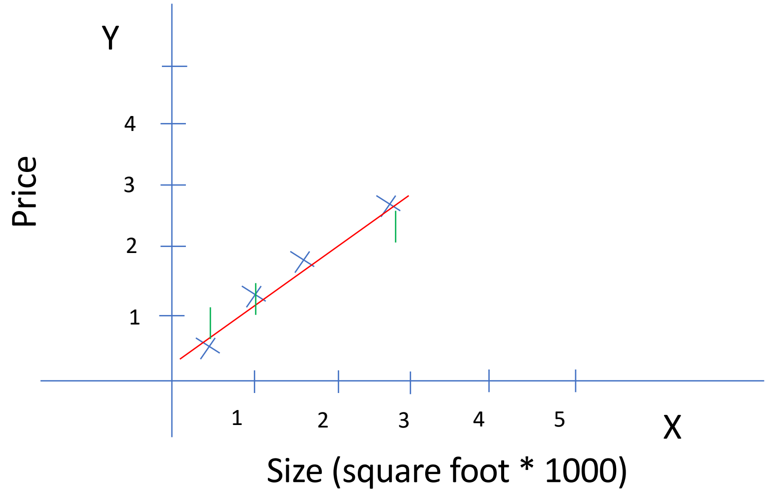
$y = 1 + 0.5*x$

Error = sum(y predict – y real)^2 = 0.8

Linear/logistic model

House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

2500?



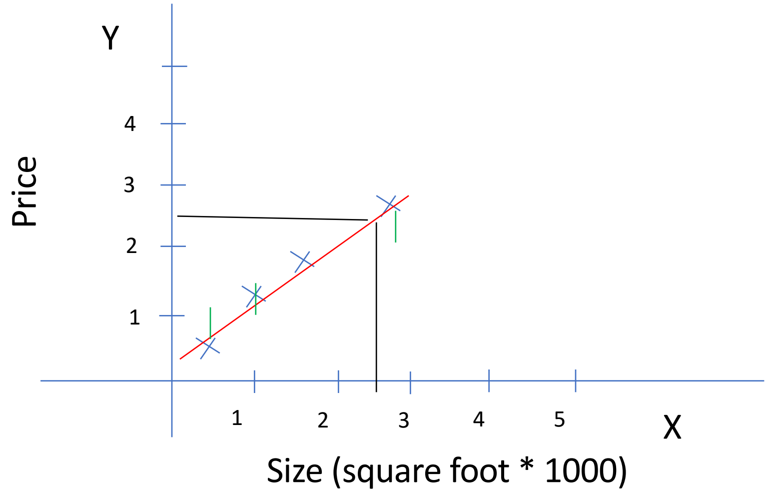
$y = 0.1 + 0.9 * x$

$\text{Error} = \text{sum}(y \text{ predict} - y \text{ real})^2 = 0.2$

Linear/logistic model

House size (sq.)	Price (M)
500	0.7
1000	1.1
2000	1.9
3000	2.8

2500?



$y = 0.1 + 0.9 * x$

Error = sum(y predict – y real)^2 = 0.2

Linear/logistic model

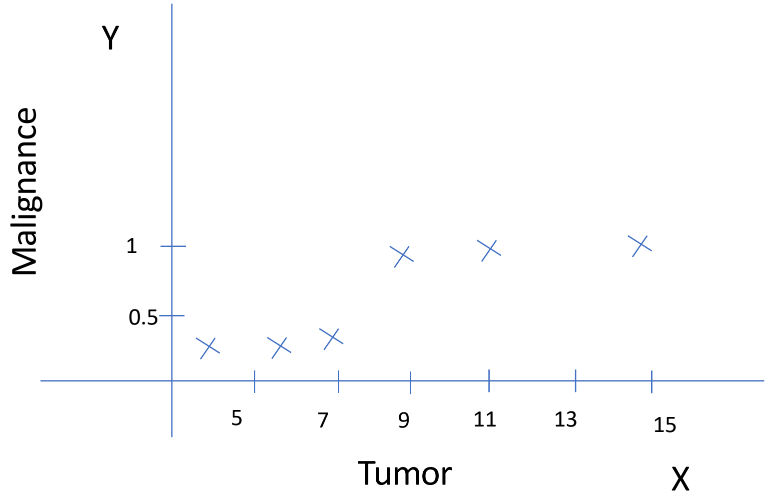
Tumor Size (um)	malignan t
5	No
8	No
10	No
13	Yes
15.1	Yes
16	Yes

12?

Linear/logistic model

Tumor Size (um)	malignant
5	No
8	No
10	No
13	Yes
15.1	Yes
16	Yes

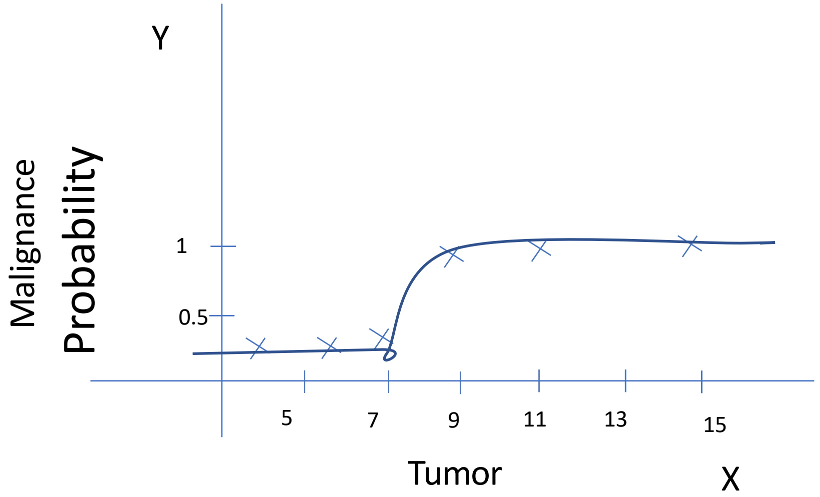
12?



Linear/logistic model

Tumor Size (um)	malignant
5	No
8	No
10	No
13	Yes
15.1	Yes
16	Yes

12?

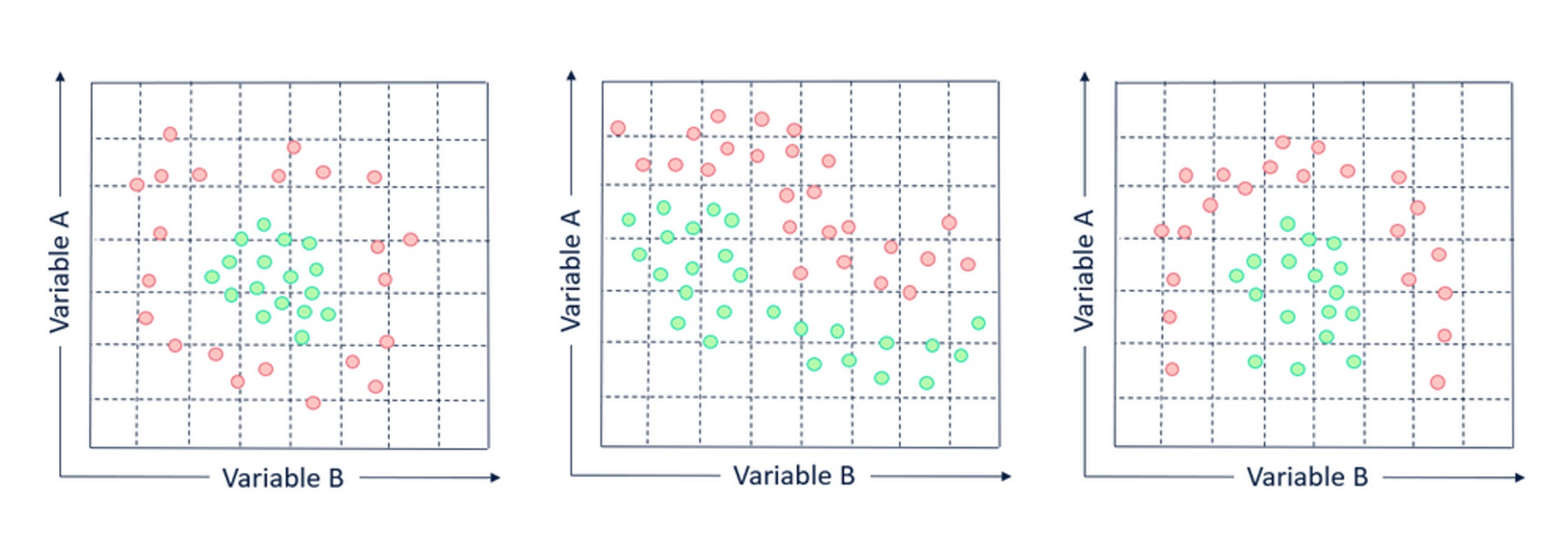


$P = 1/(1 + e^{-(b0 + b1x)})$

$=1/(1 + e^{-(3 + 2x)})$

$\text{Log}\ (P/(1-P)) = 3 + 2x$

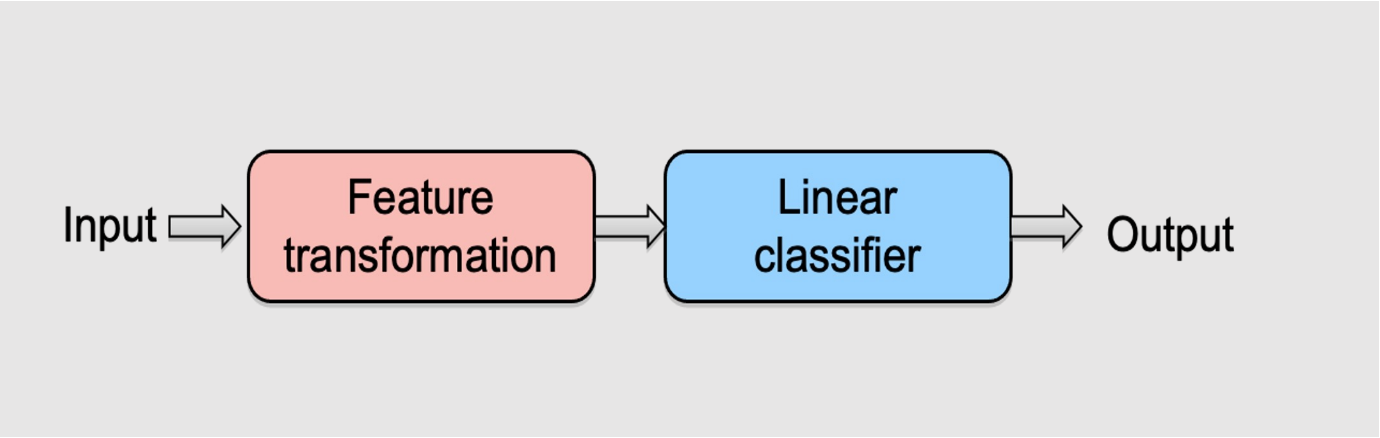
Non Linear Models



Internet Image Credit: https://community.alteryx.com/t5/Data-Science/And-For-My-Next-Trick-An-Introduction-to-Support-Vector-Machines/ba-p/360762

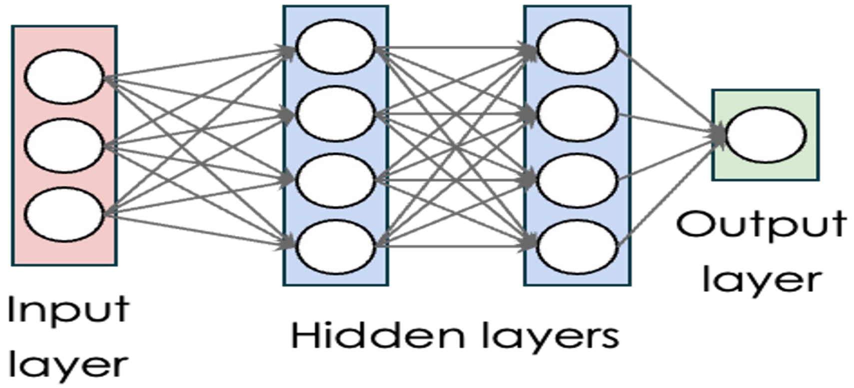
Simple Nonlinear Prediction

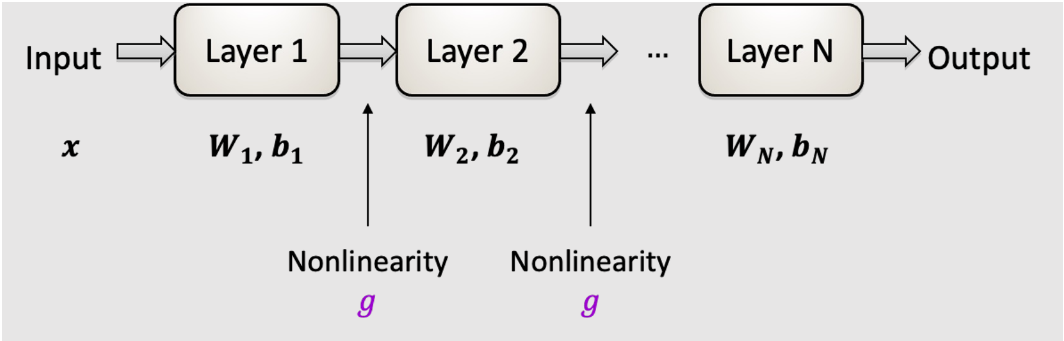
The “shallow” approach: nonlinear feature transformation (often by hand), followed by a linear classifier.



Multi-layer Neural Networks

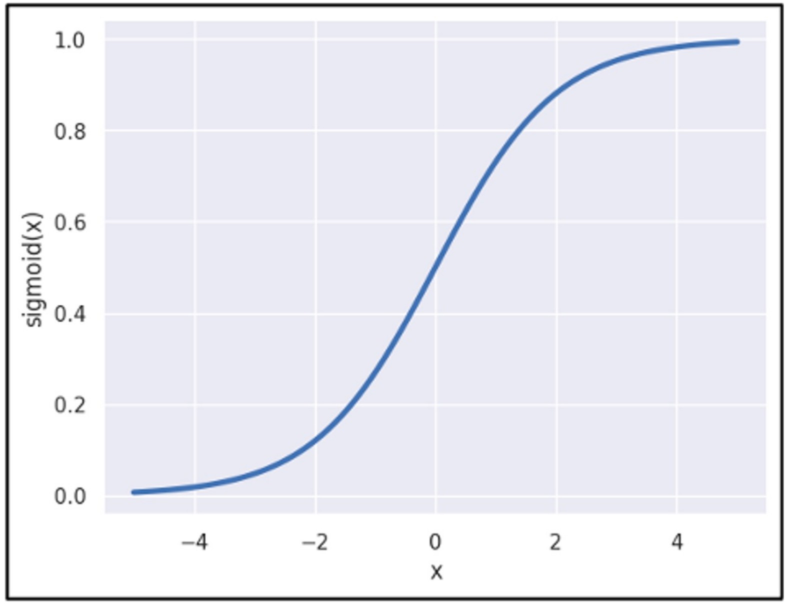
- The “deep” approach: stack multiple layers of linear transformations interspersed with nonlinearities





- (now for a N layers neural network)
  - $f(\boldsymbol{x})=\boldsymbol{W}_N\, g(\,\ldots\, g(\,\boldsymbol{W}_2\, g(\,\boldsymbol{W}_1\,\boldsymbol{x}+\boldsymbol{b_1}\,)+\boldsymbol{b_2}\,)\ldots) + \boldsymbol{b}_N$

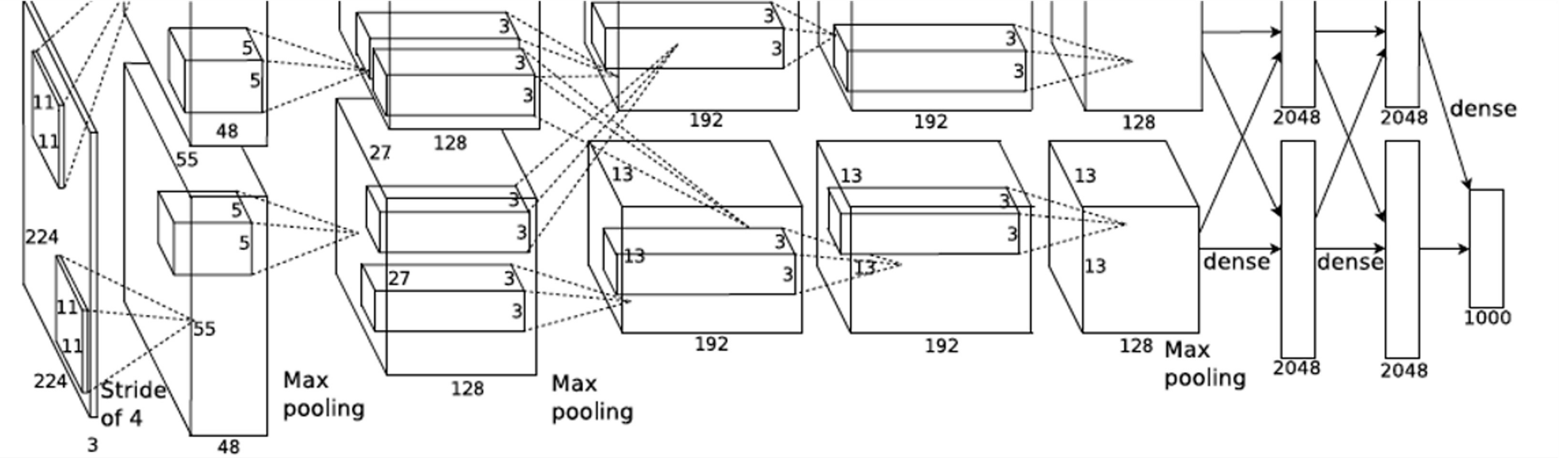
Nonlinearity



Sigmoid: $g(x)$=1/(1+e^($-x$))

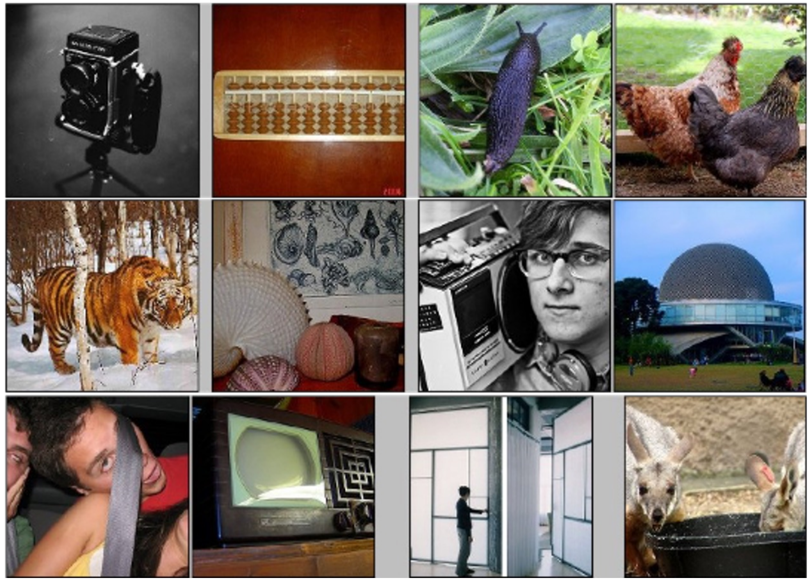
- Historically very popular
- Squashes numbers to range [0,1]
- Saturated neurons "kill" the gradients

AlexNet



- AlexNet is the winner of ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012:
  - Classification task with 1.2 million training images, 1000 classes
  - Images gathered from Internet
  - Human labels via Amazon MTurk

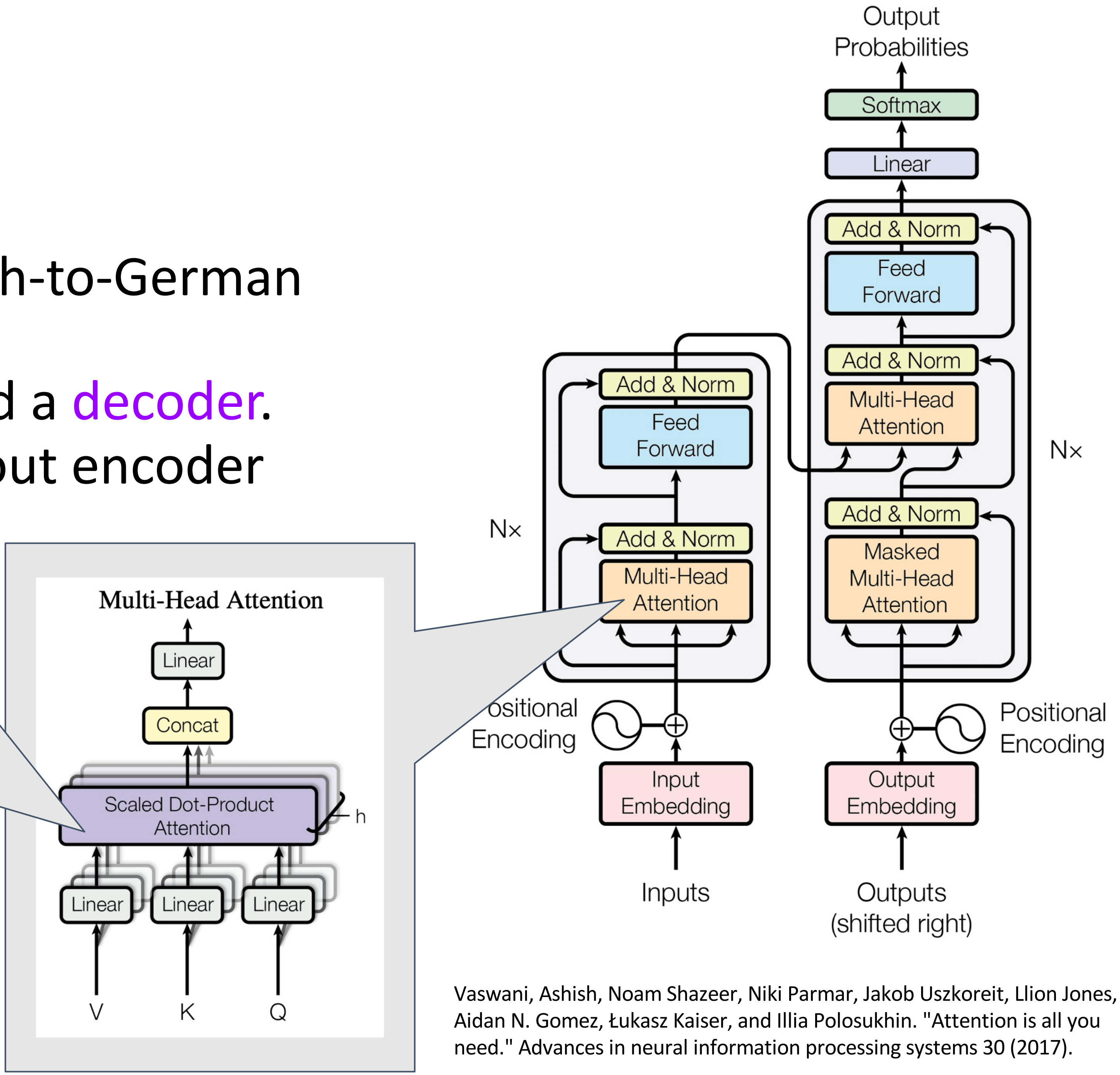
IMAGENET

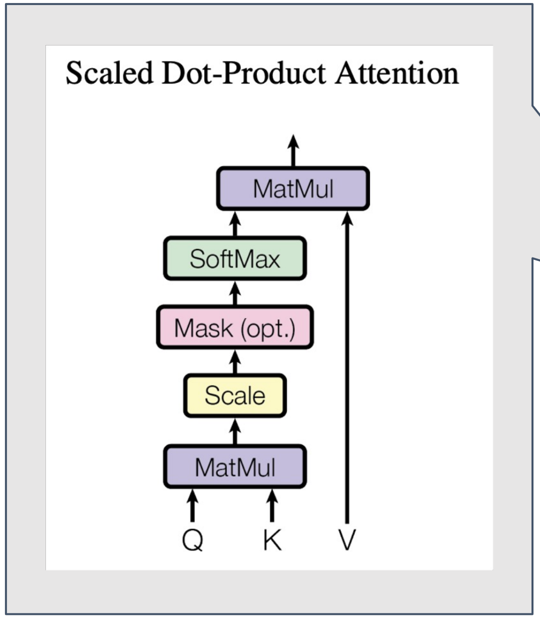


www.image-net.org/challenges/LSVRC/

Transformers

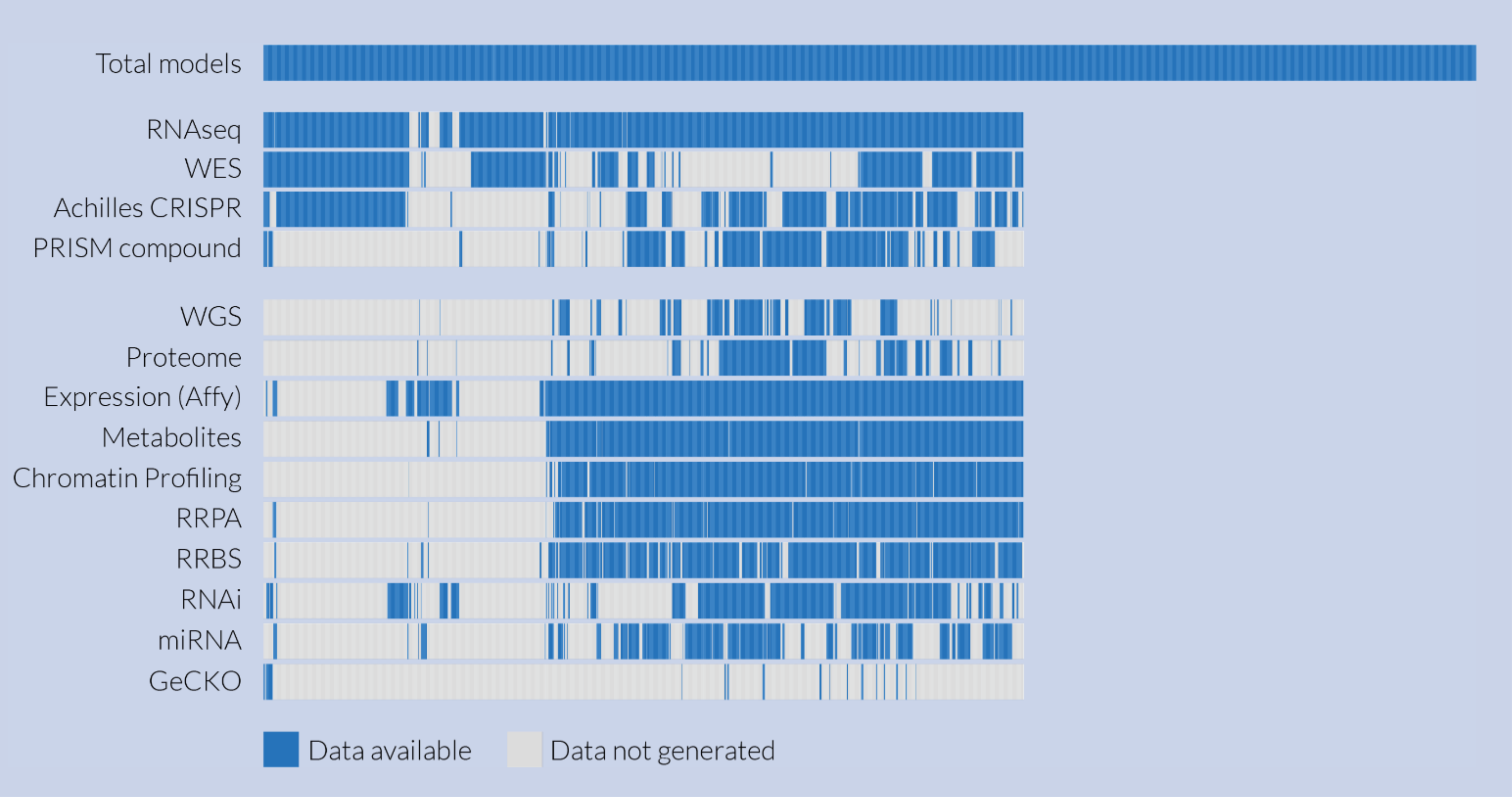
- English-to-French and English-to-German language translation
- Utilized both an encoder and a decoder.
- Where did we also learn about encoder and decoder?





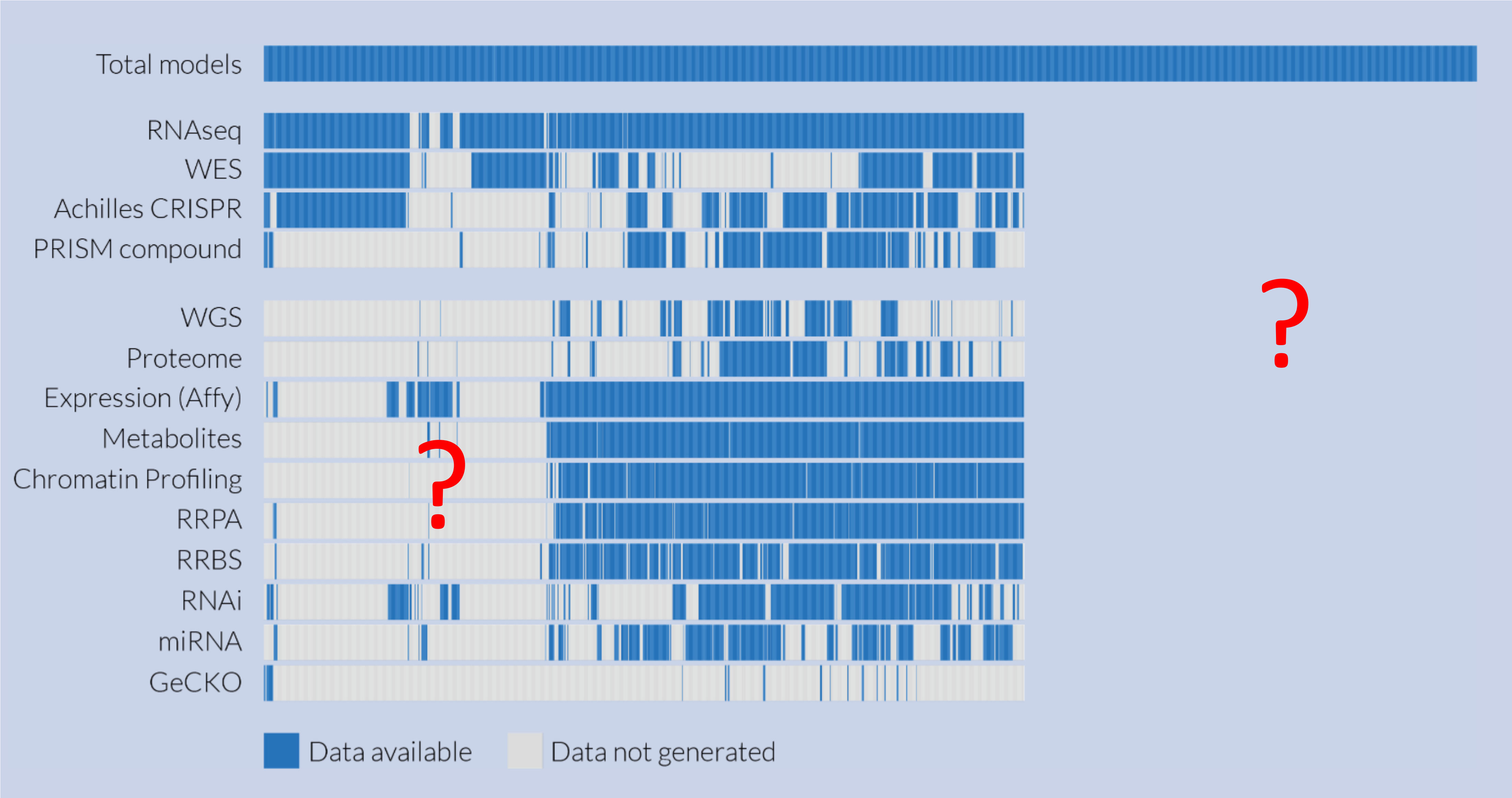
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

DepMap: large-scale molecular characterization of cancer cell lines



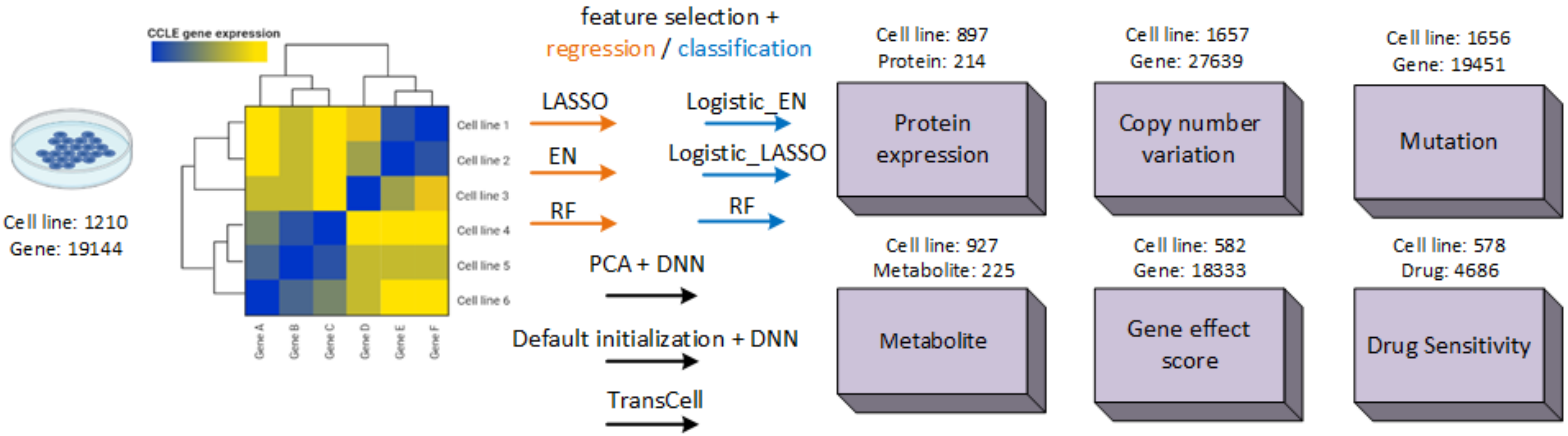
https://depmap.org/portal/home/#/

but there are too many cell lines with incomplete data

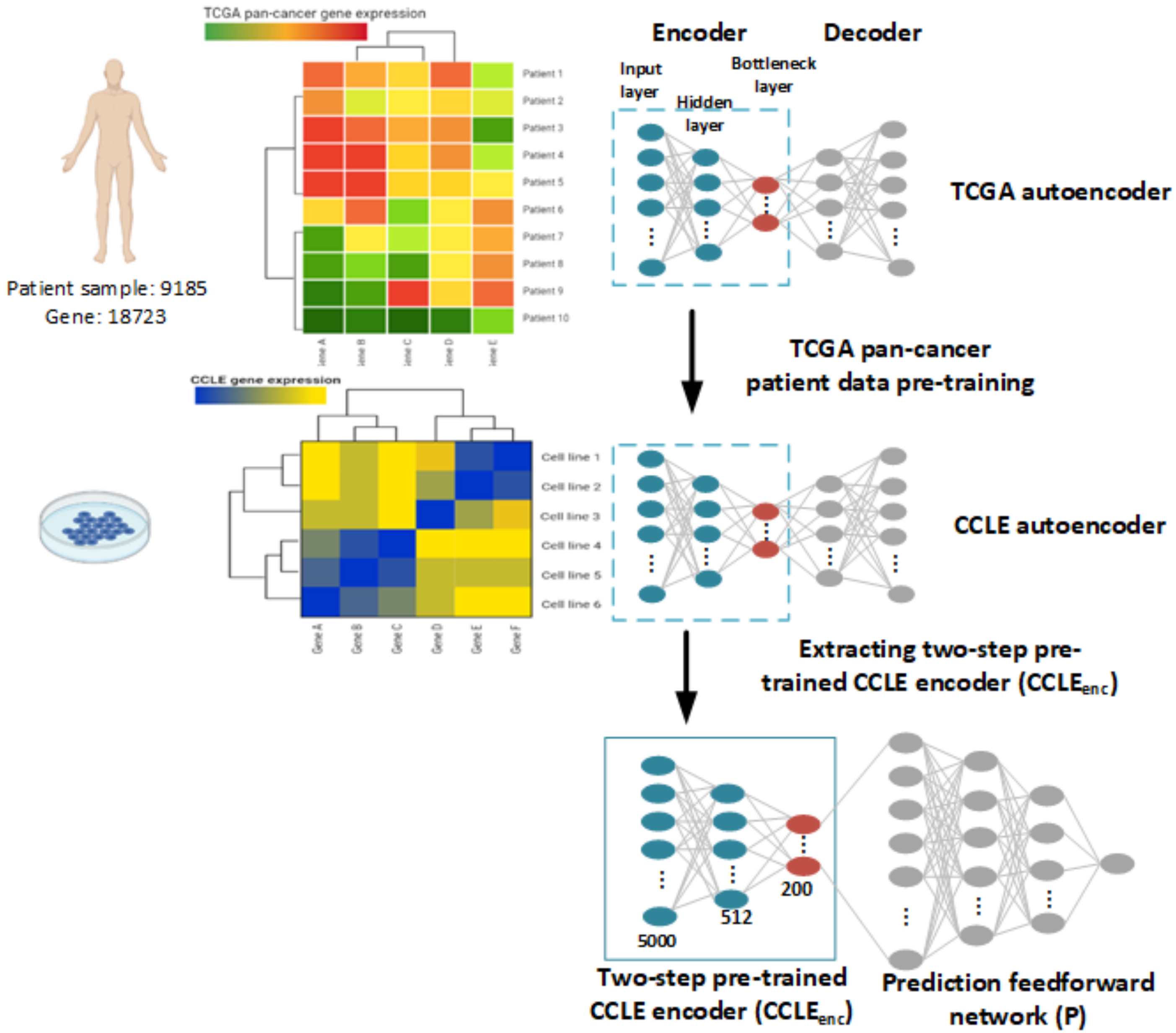


https://depmap.org/portal/home/#/

TransCell: Predicting genomic features based on gene expression

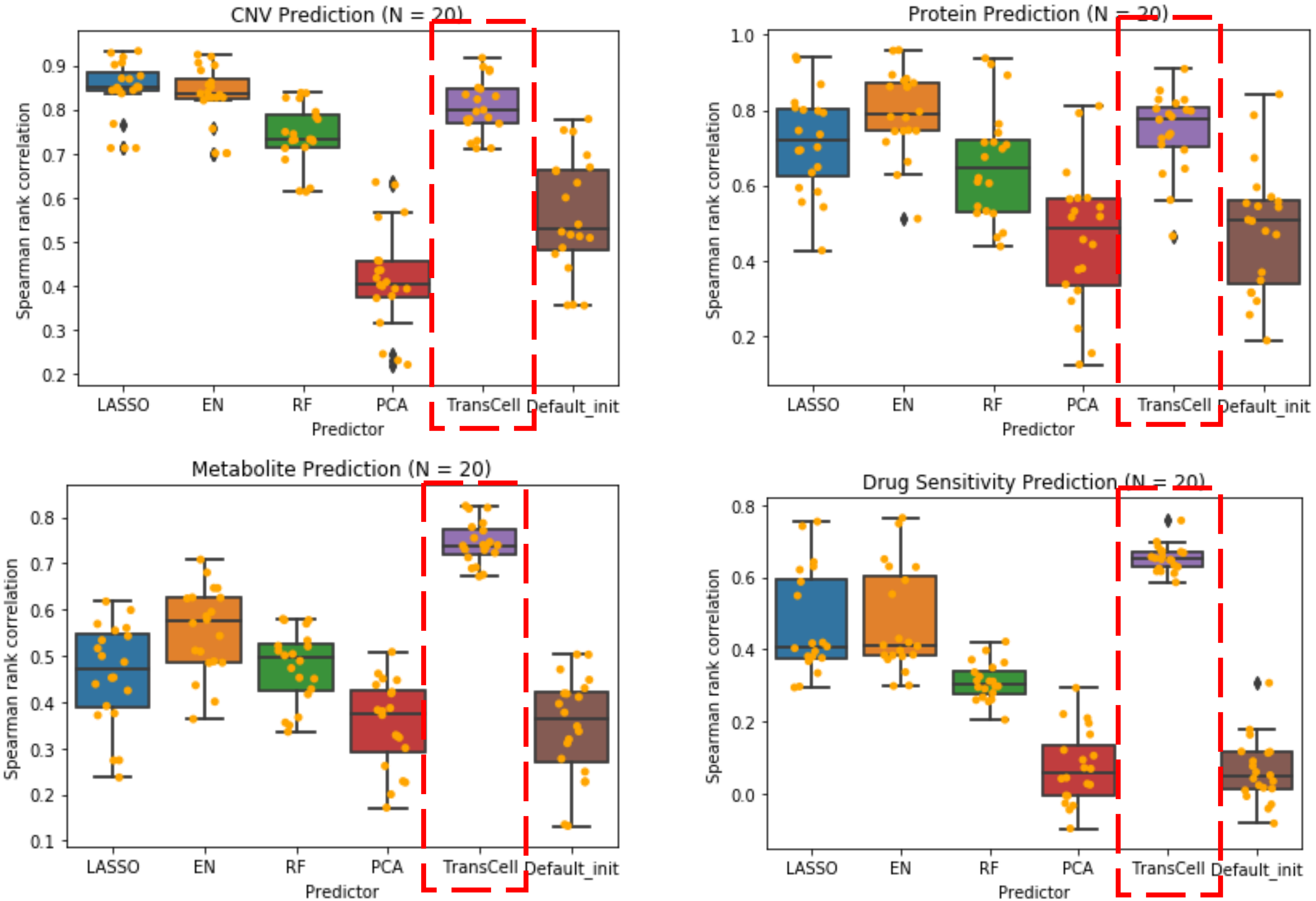


TransCell architecture

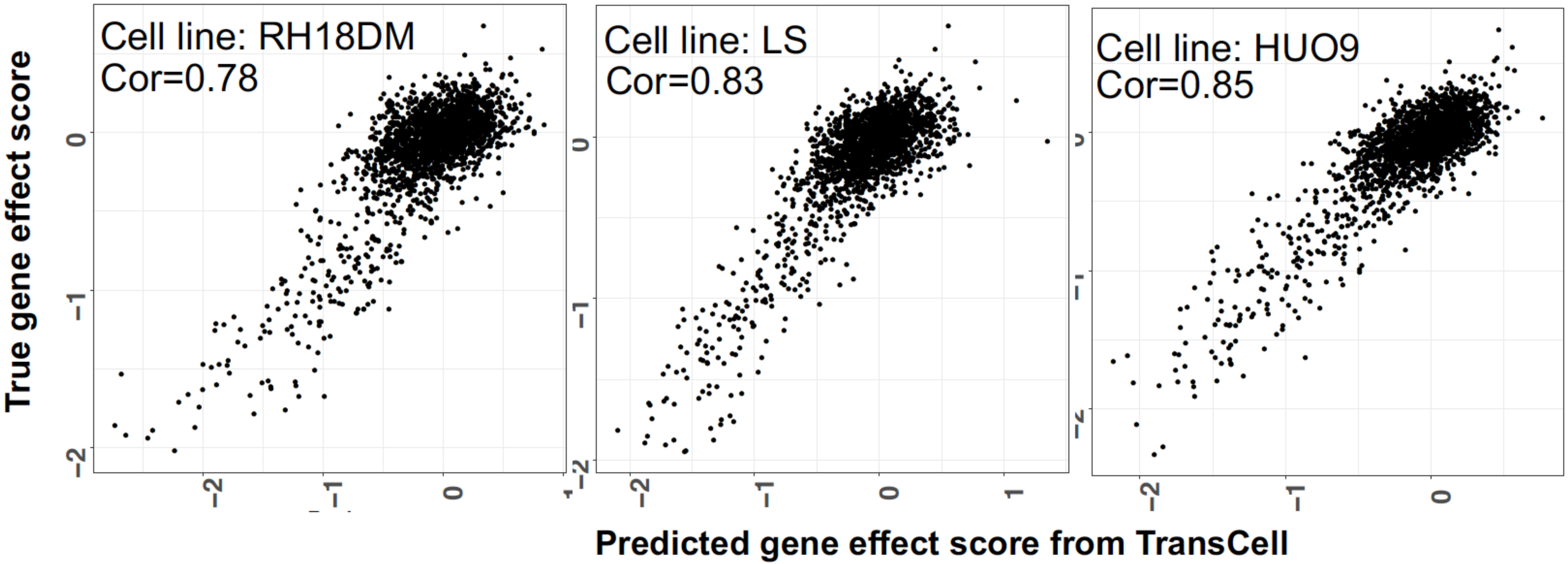


Two-step transfer learning

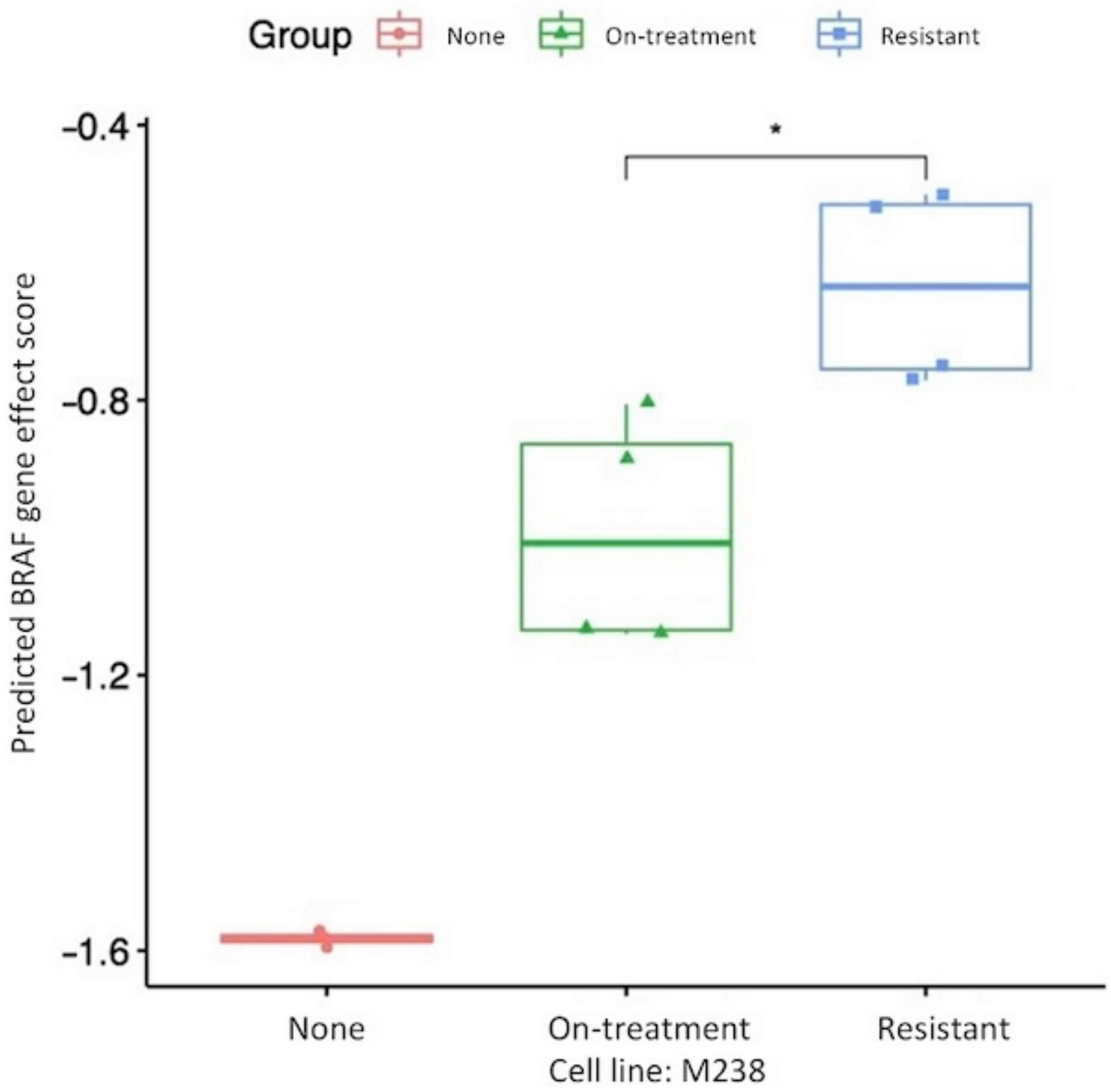
Gene expression can be used to predict genomic features and cellular responses



TransCell predicted gene effect scores (genetic dependency) for new pediatric cancer cell lines



In silico CRISPR knockout captured drug resistances



TransCell Demo

http://apps.octad.org/transcell/