Machine Learning for Data Generated from Next-Generation Sequencing

Fangfang Xia

fangfang@anl.gov

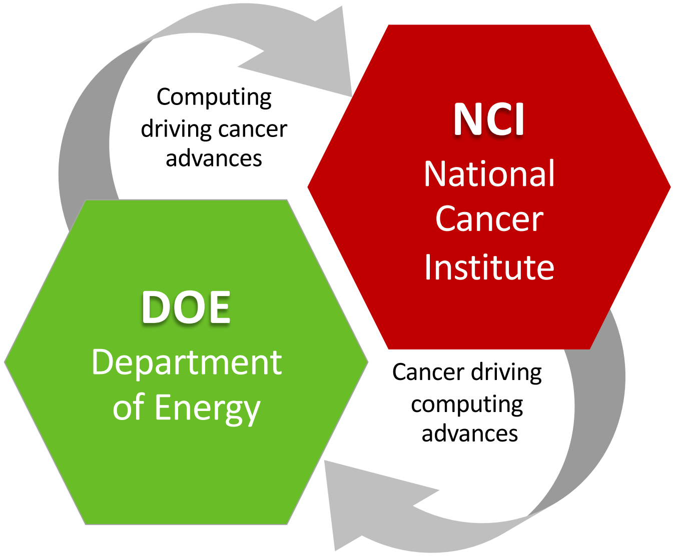
Computer Scientist

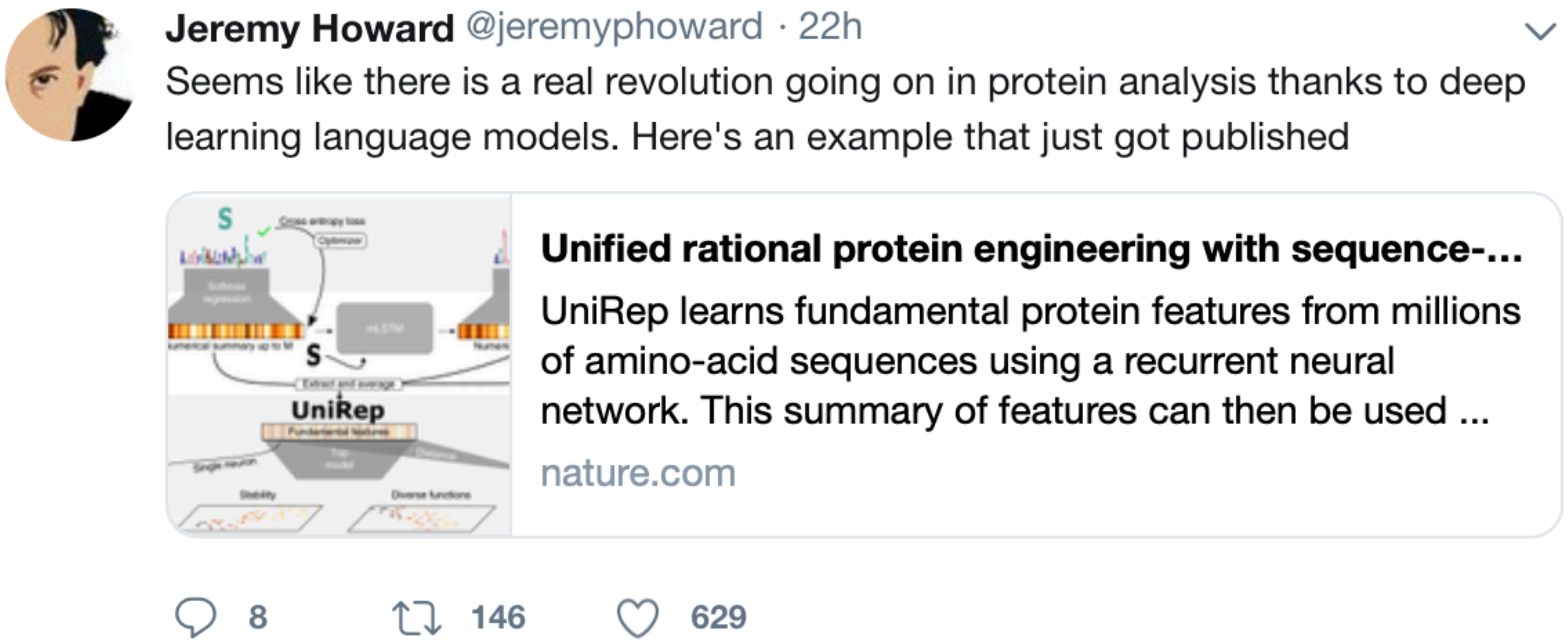
Argonne National Laboratory

2019-10-23

Argonne NATIONAL LABORATORY

DOE-NCI partnership to advance exascale development through cancer research


Computing driving cancer advances
NCI
National Cancer Institute
DOE
Department of Energy
Cancer driving computing advances



Jeremy Howard @jeremyphoward · 22h

Seems like there is a real revolution going on in protein analysis thanks to deep learning language models. Here's an example that just got published

Unified rational protein engineering with sequence-...

UniRep learns fundamental protein features from millions of amino-acid sequences using a recurrent neural network. This summary of features can then be used ...

nature.com

8 146 629

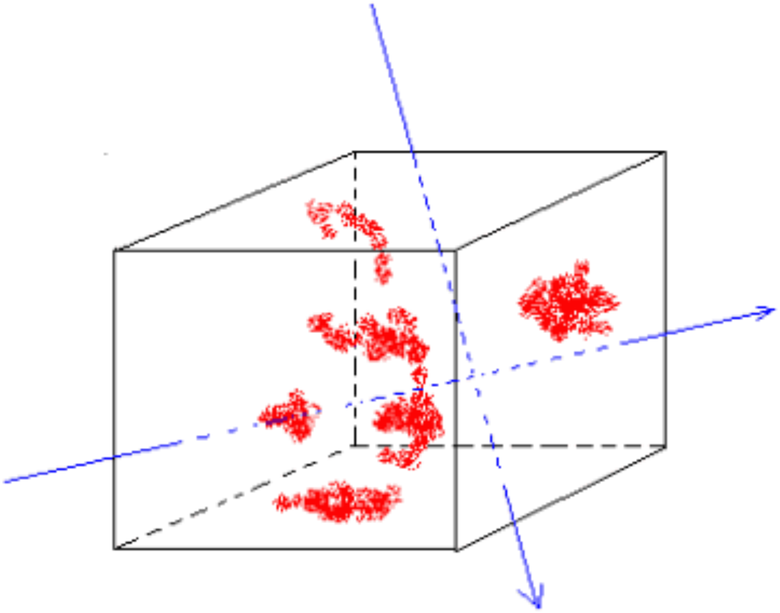
Representation learning

dnaK

- `atgggtaaaataattggtatcgacct`

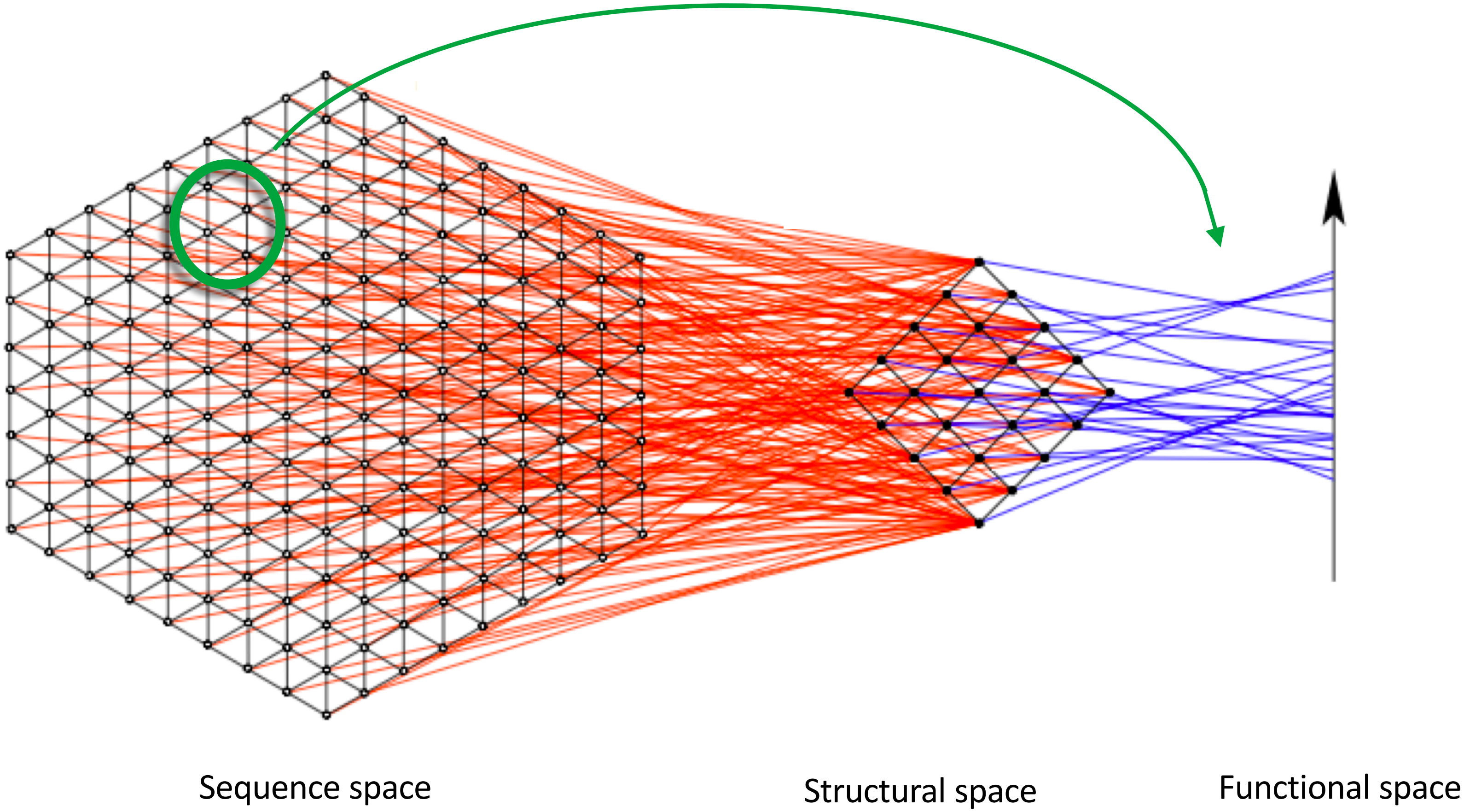
random

- `tgaggtcgtagagtagacca`



- 4^n possible sequences
  - Real sequences occupy a tiny fraction of all possible sequences

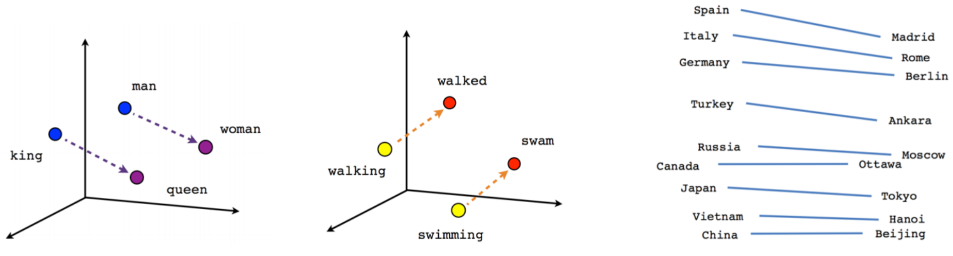
Mapping the spaces: starting from raw seqs



Embedding for categorical variables

puppy	[1, 0, 0, 0]
dog	[0, 1, 0, 0]
kitten	[0, 0, 1, 0]
cat	[0, 0, 0, 1]

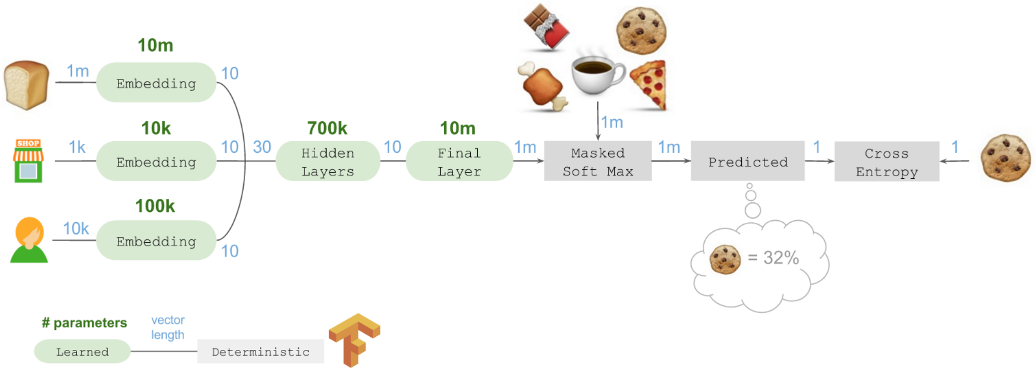
puppy	[0.9, 1.0, 0.0, 0.2]
dog	[1.0, 0.2, 0.0, 0.9]
kitten	[0.0, 1.0, 0.5, 0.1]
cat	[0.0, 0.2, 1.0, 1.0]



Male-Female

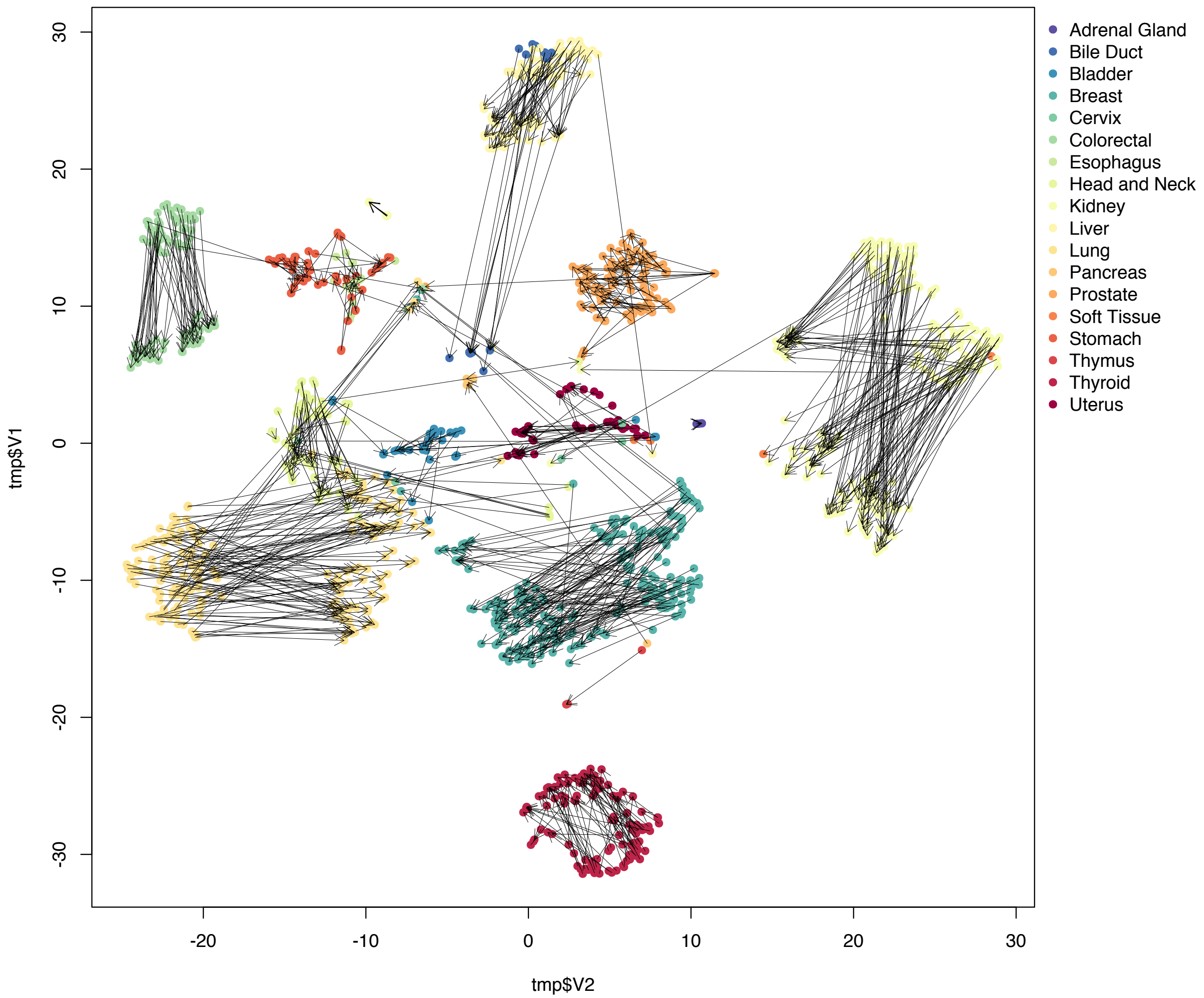
Verb tense

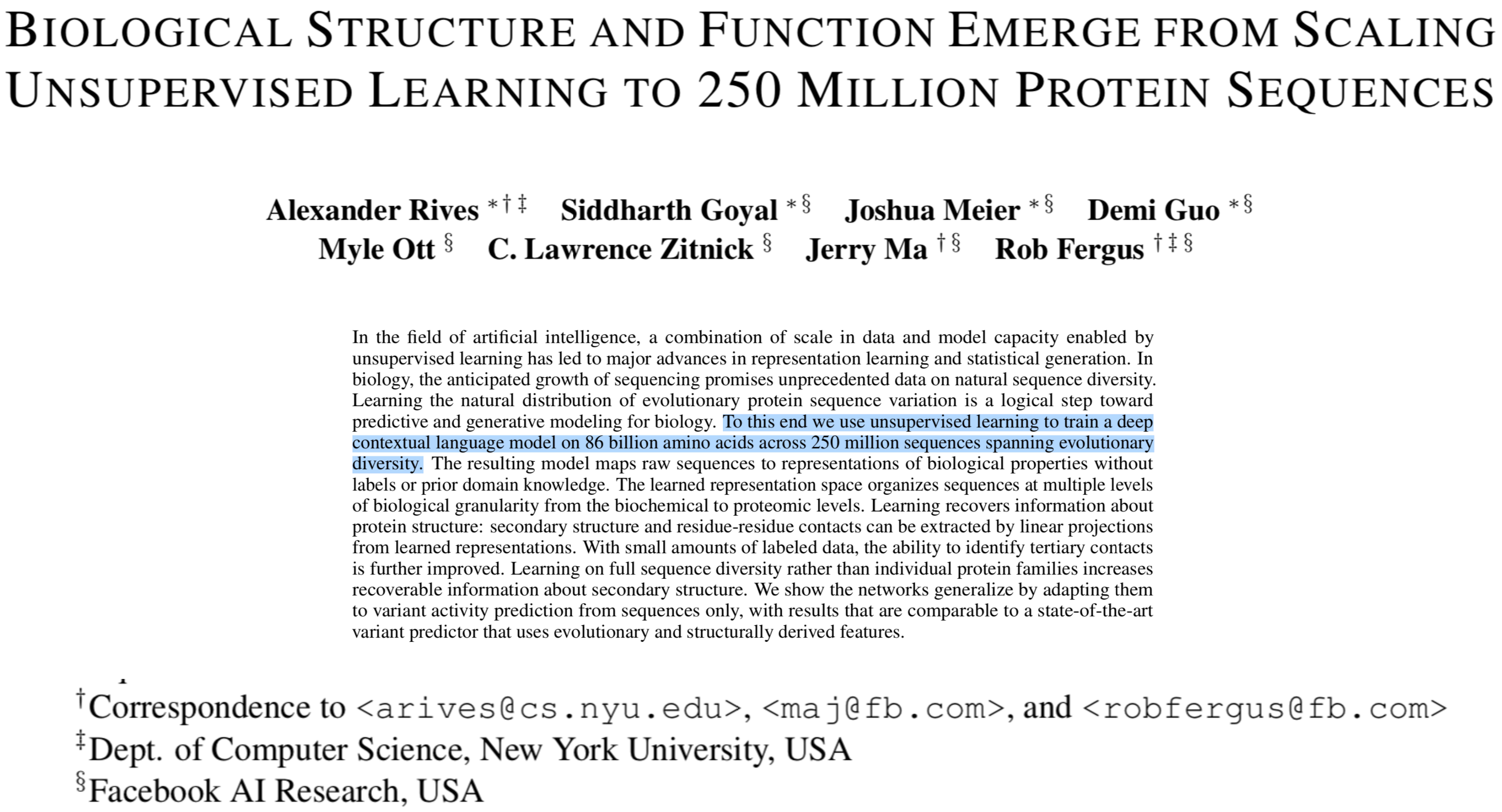
Country-Capital



From the Instacart blog post 'Deep Learning with Emojis (not Math)'

Normal => Tumor



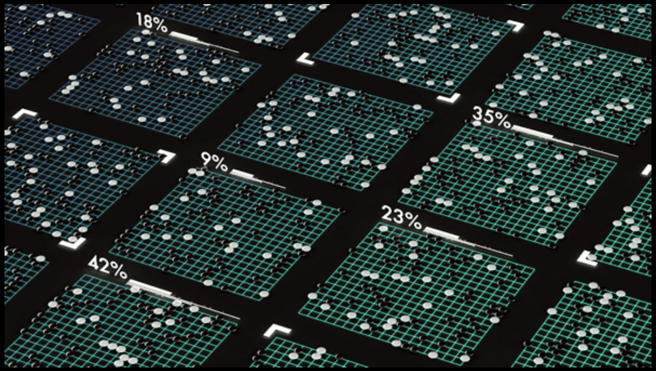


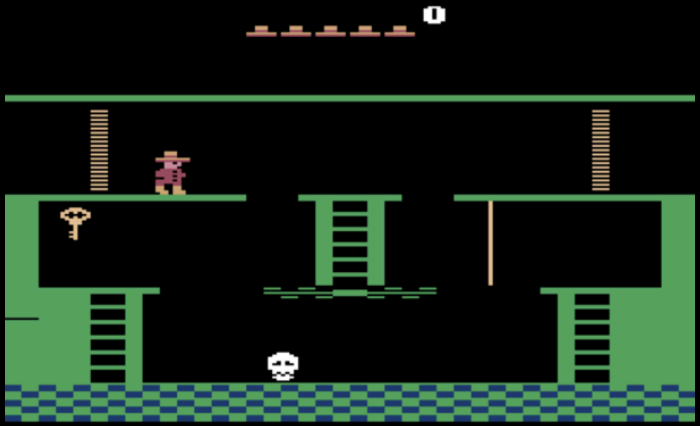
BIOLOGICAL STRUCTURE AND FUNCTION EMERGE FROM SCALING UNSUPERVISED LEARNING TO 250 MILLION PROTEIN SEQUENCES

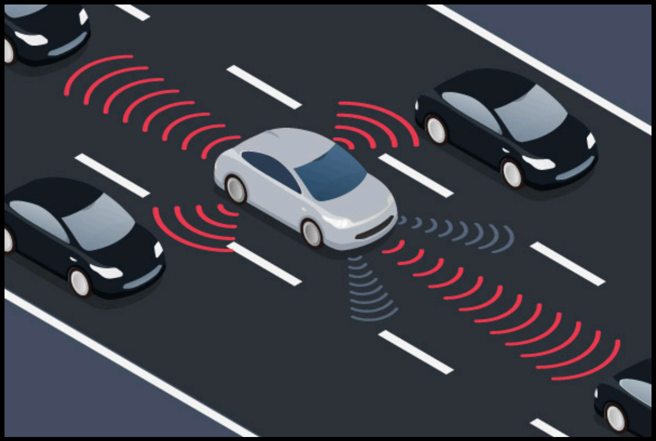
Alexander Rives *†‡ **Siddharth Goyal** *§ **Joshua Meier** *§ **Demi Guo** *§
Myle Ott § **C. Lawrence Zitnick** § **Jerry Ma** †§ **Rob Fergus** †‡§

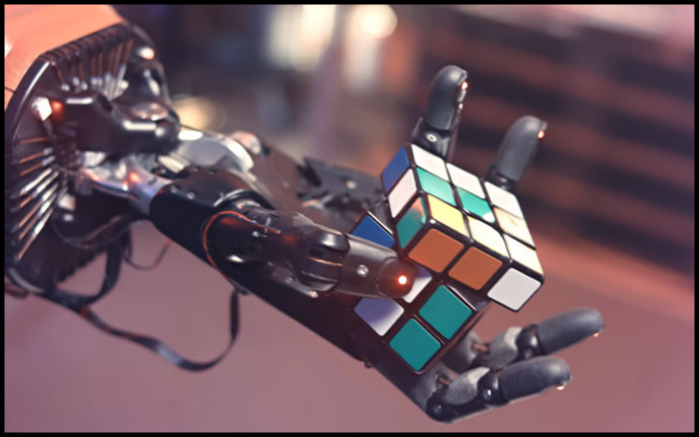
In the field of artificial intelligence, a combination of scale in data and model capacity enabled by unsupervised learning has led to major advances in representation learning and statistical generation. In biology, the anticipated growth of sequencing promises unprecedented data on natural sequence diversity. Learning the natural distribution of evolutionary protein sequence variation is a logical step toward predictive and generative modeling for biology. To this end we use unsupervised learning to train a deep contextual language model on 86 billion amino acids across 250 million sequences spanning evolutionary diversity. The resulting model maps raw sequences to representations of biological properties without labels or prior domain knowledge. The learned representation space organizes sequences at multiple levels of biological granularity from the biochemical to proteomic levels. Learning recovers information about protein structure: secondary structure and residue-residue contacts can be extracted by linear projections from learned representations. With small amounts of labeled data, the ability to identify tertiary contacts is further improved. Learning on full sequence diversity rather than individual protein families increases recoverable information about secondary structure. We show the networks generalize by adapting them to variant activity prediction from sequences only, with results that are comparable to a state-of-the-art variant predictor that uses evolutionary and structurally derived features.

†Correspondence to <arives@cs.nyu.edu>, <maj@fb.com>, and <robfergus@fb.com>
‡Dept. of Computer Science, New York University, USA
§Facebook AI Research, USA


18%
35%
9%
23%
42%







What is Machine Learning?

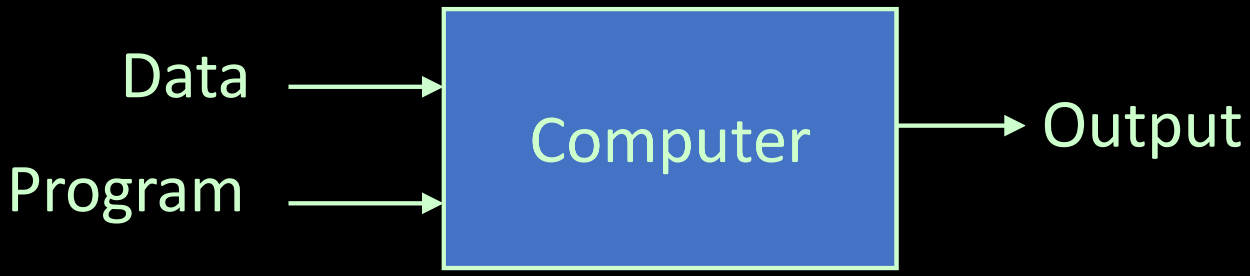
The complexity in traditional computer programming is in the code (programs that people write). In machine learning, algorithms (programs) are in principle simple and the complexity (structure) is in the data. Is there a way that we can automatically learn that structure? That is what is at the heart of machine learning.

-- Andrew Ng

That is, machine learning is the about the construction and study of systems that can learn from data. This is very different than traditional computer programming.

The Cartoon Form

Traditional Programming



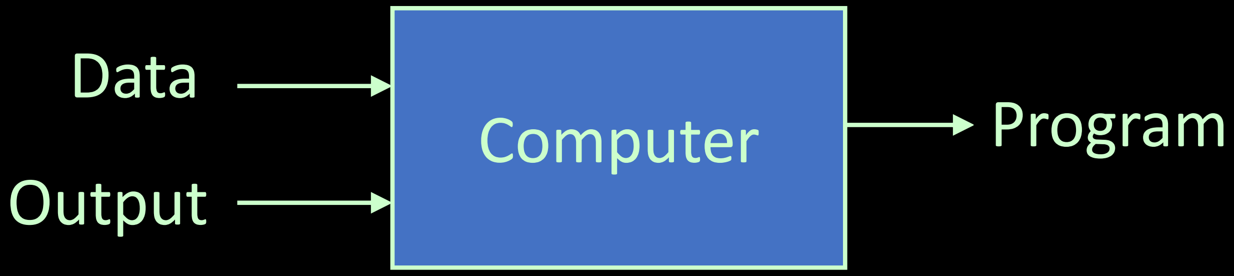
Data

Program

Computer

Output

Machine Learning



Data

Output

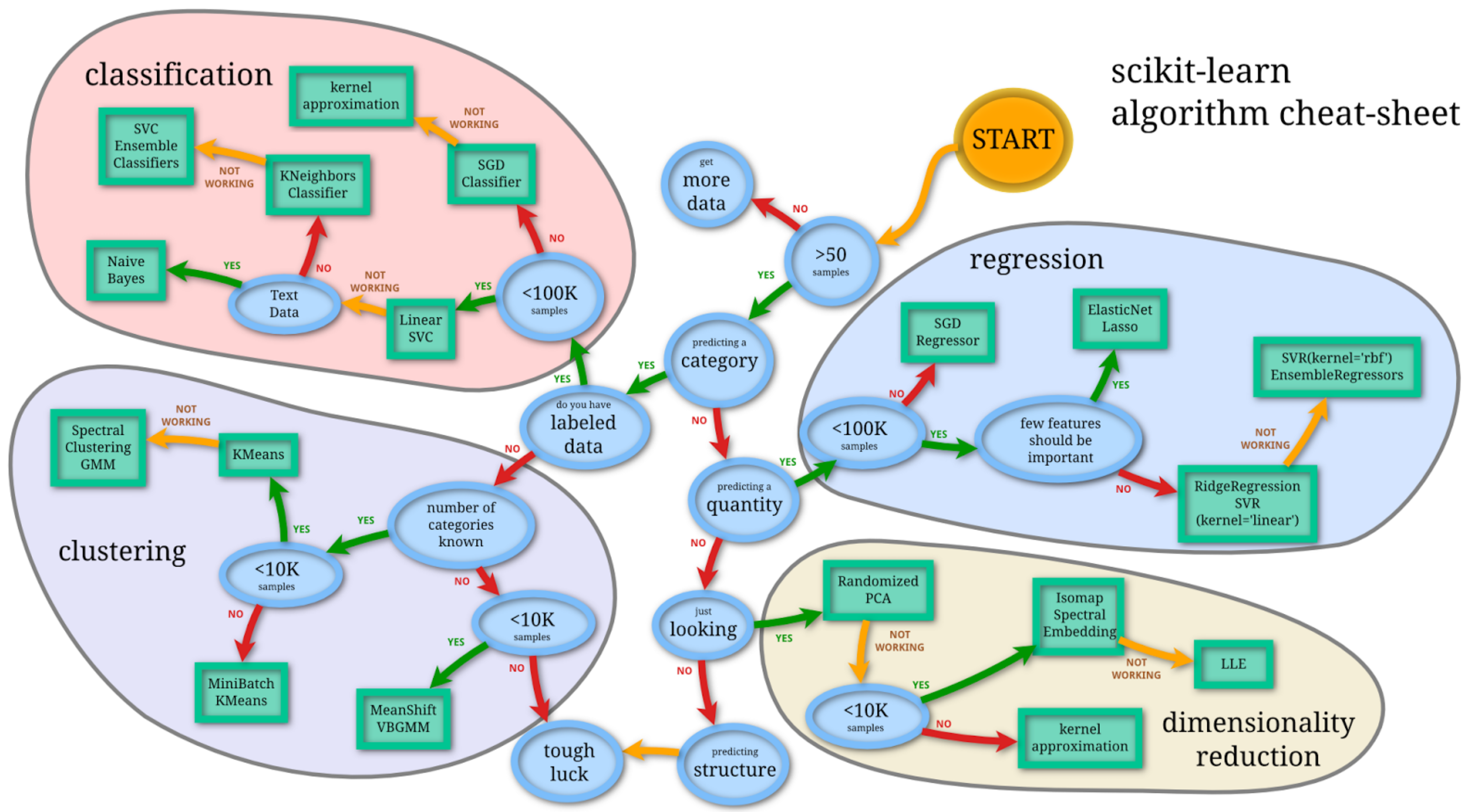
Computer

Program

Examples in applying ML to NGS data

- Classifying cancer type with gene expression profiles
- Removing study bias in tumor gene expression profiles
- Classifying cancer type with SNP data
- Drug response prediction - introduction

Four Typical Problems

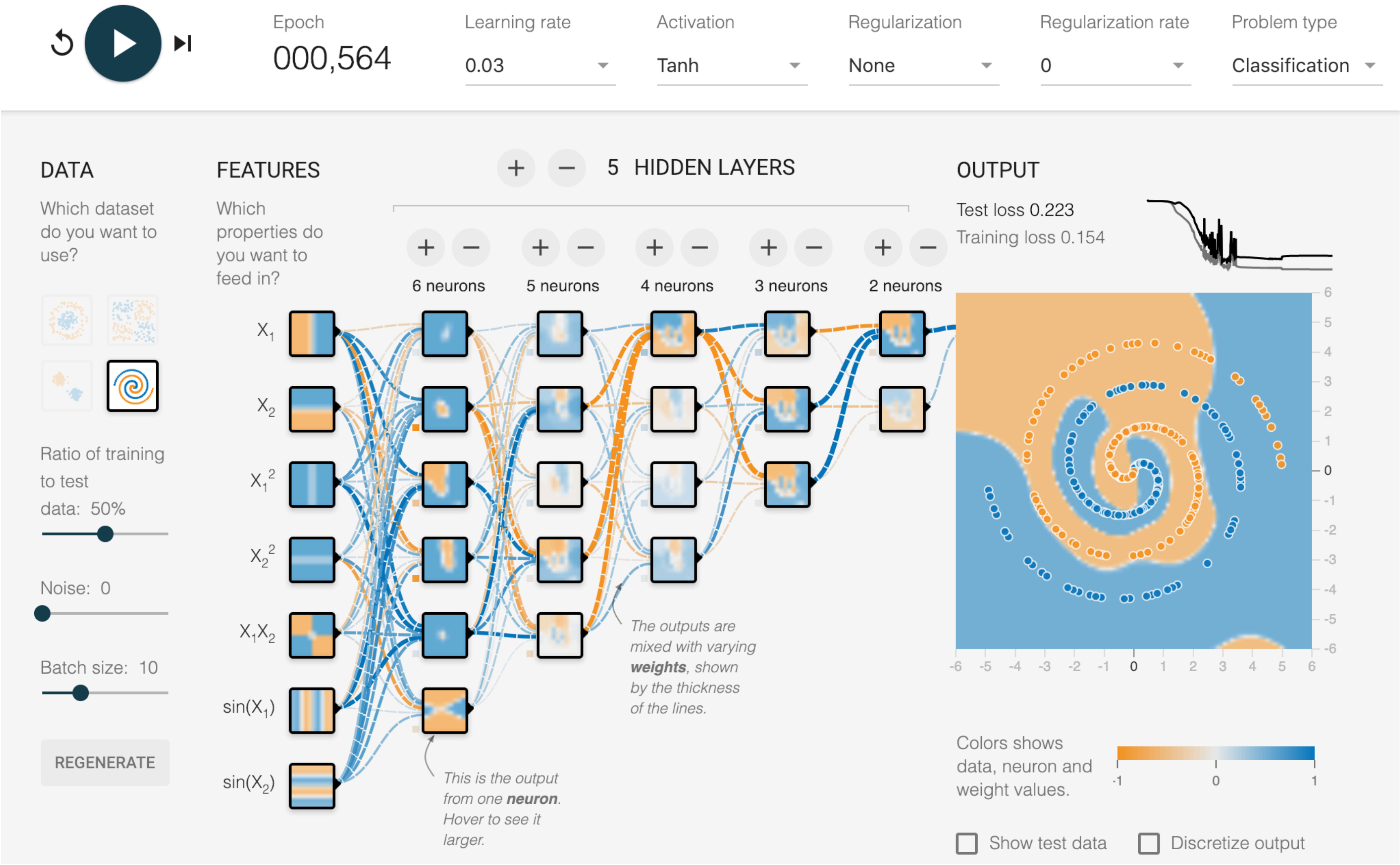


Deep learning in biology and medicine

https://github.com/greenelab/deep-review

Keras.js

Tensorflow playground



Epoch 000,564

Learning rate 0.03

Activation Tanh

Regularization None

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES

Which properties do you want to feed in?

5 HIDDEN LAYERS

The outputs are mixed with varying ***weights****, shown by the thickness of the lines.*

This is the output from one ***neuron****. Hover to see it larger.*

OUTPUT

Test loss 0.223

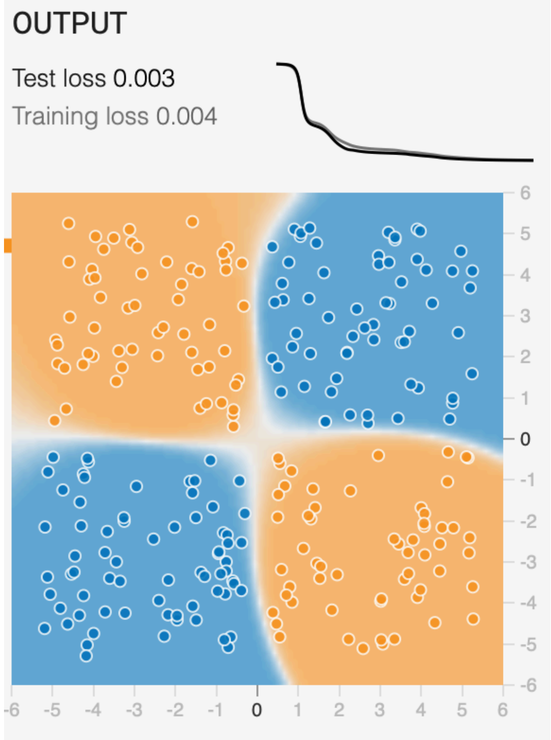
Training loss 0.154

Colors shows data, neuron and weight values.

□ Show test data

□ Discretize output

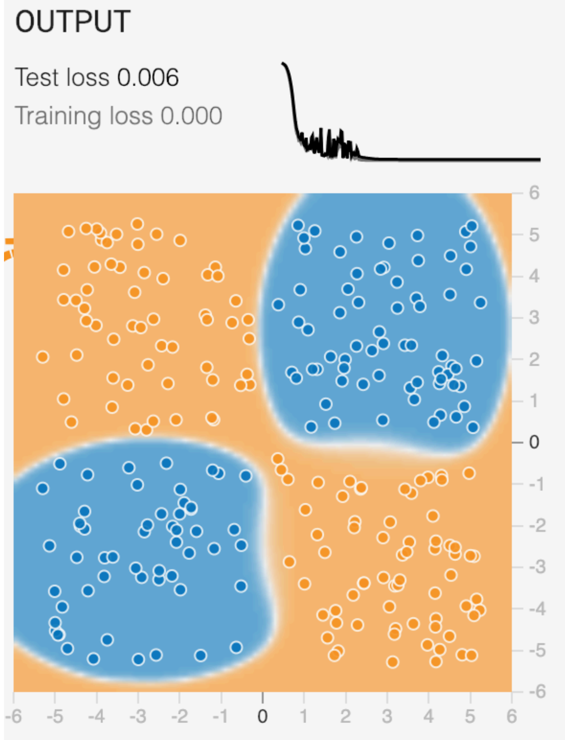
A few “good” runs



OUTPUT

Test loss 0.003

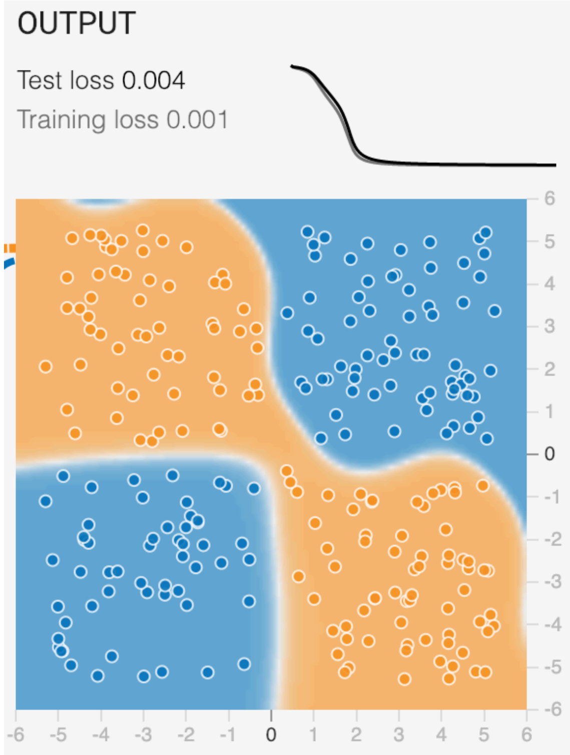
Training loss 0.004



OUTPUT

Test loss 0.006

Training loss 0.000

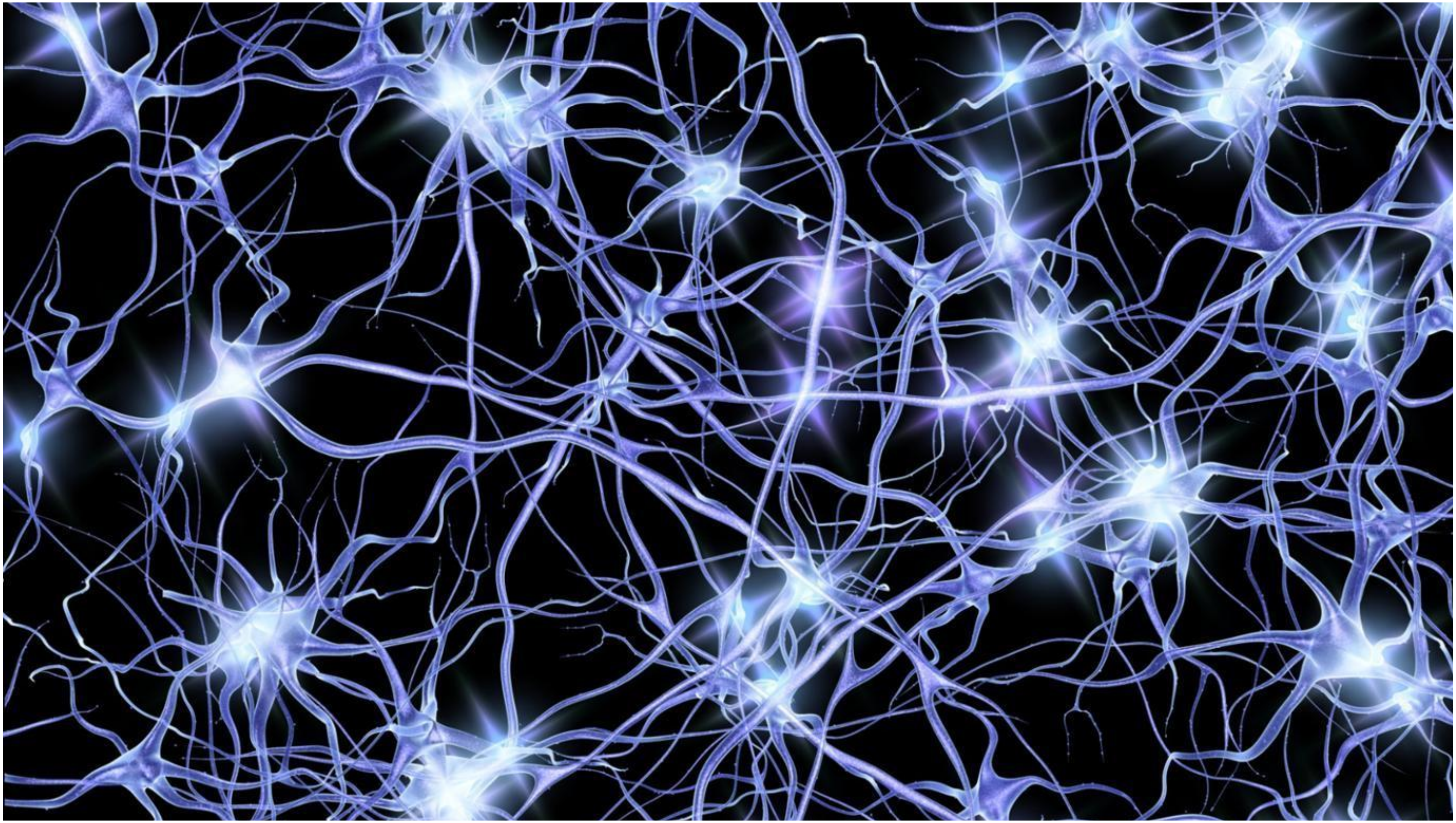


OUTPUT

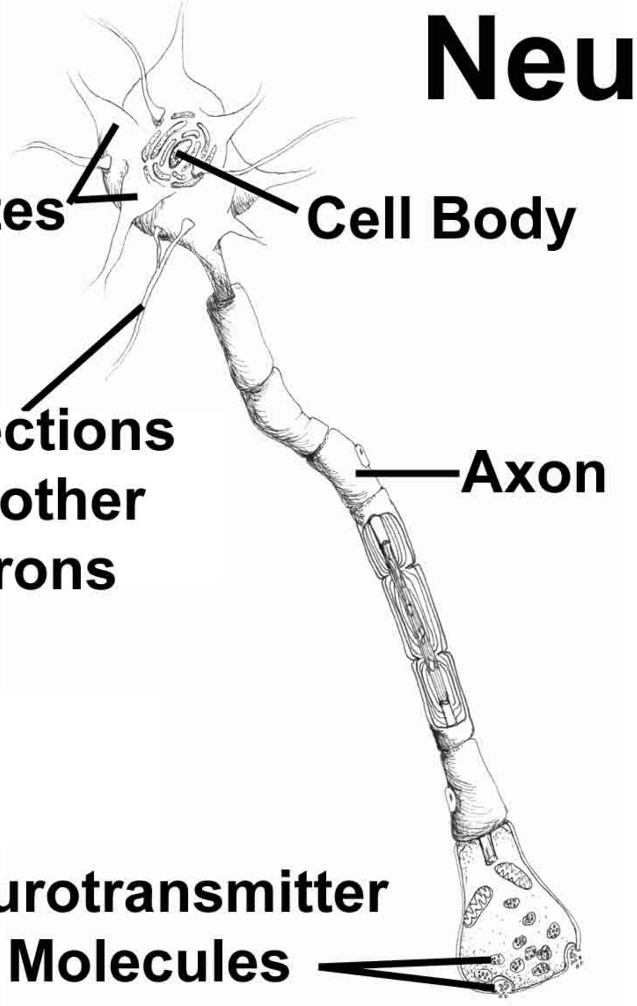
Test loss 0.004

Training loss 0.001

Deep Learning Basics



Neuron



Dendrites

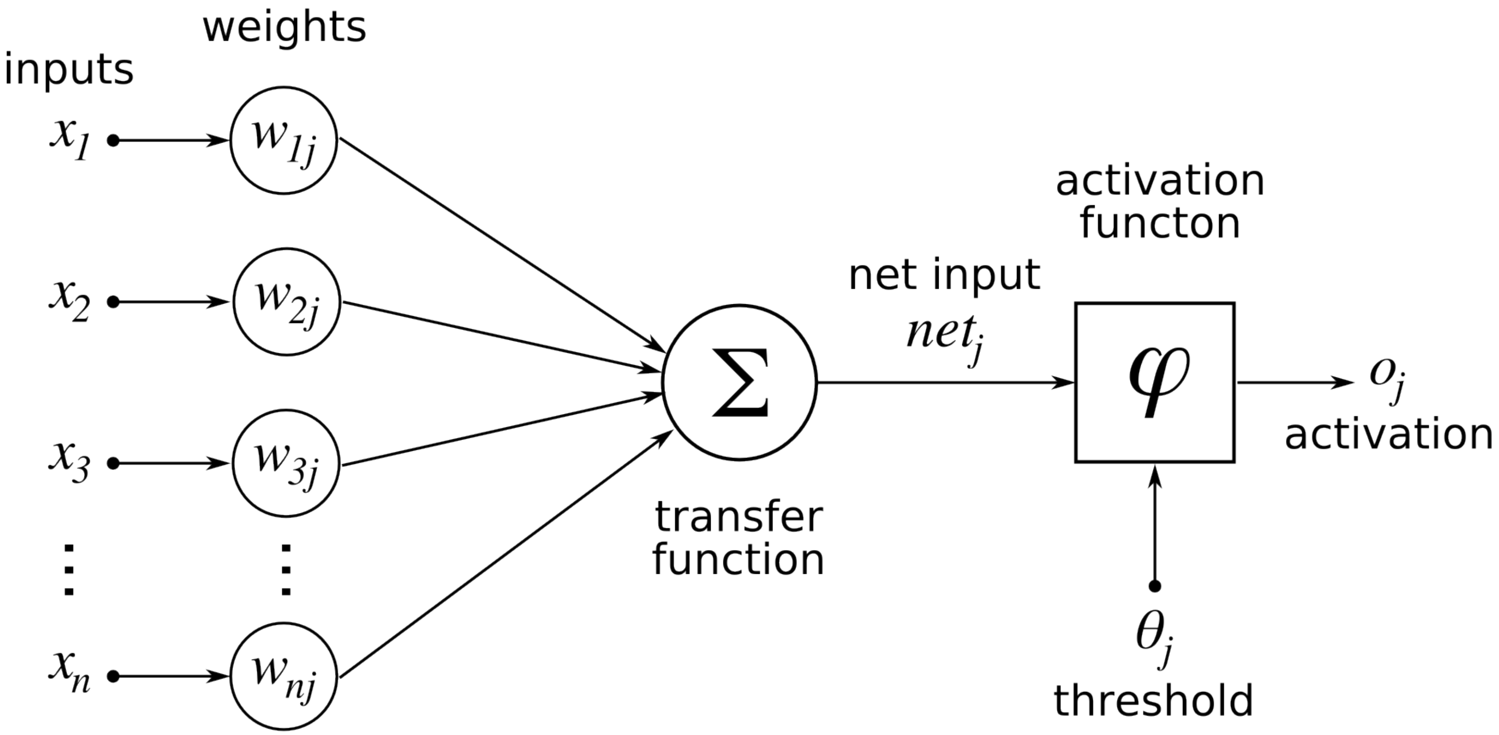
Cell Body

Connections from other Neurons

Axon

Neurotransmitter Molecules

Mathematical Model of a Neuron

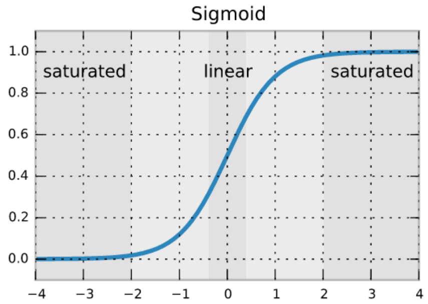


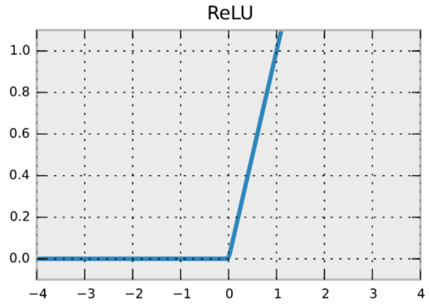
$z = w \cdot x + b$

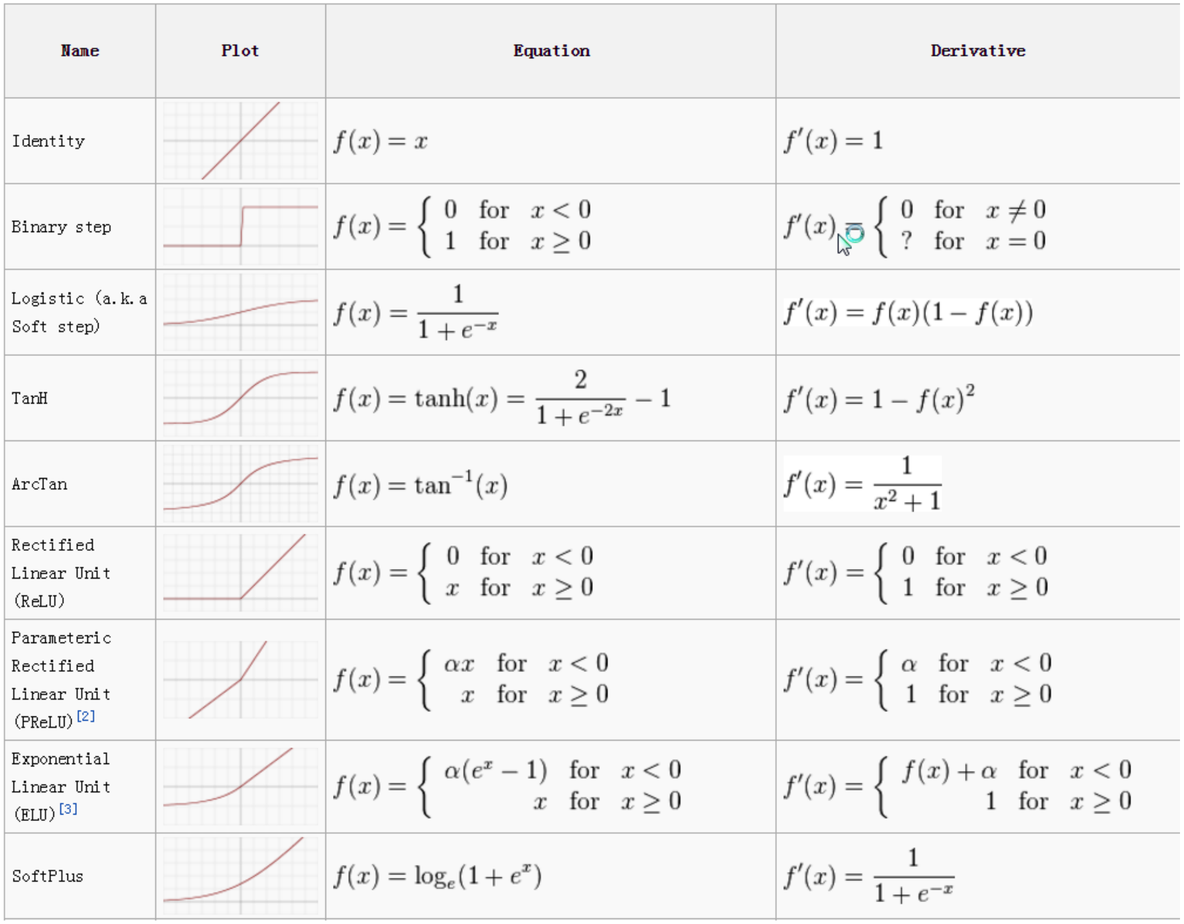
$o = \varphi(z)$

Activation

$$\sigma(z) \equiv \frac{1}{1+e^{-z}}.$$







Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1 - f(x))$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU)[2]		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU)[3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$

Loss and Accuracy

Regression problem

y_true: [-50, -10, +80]
y_pred: [-30, +10, +60]

MSE = 0.04

Classification ***view***

[1, 1, 0]

[1, 0, 1]

ACC = 0.67

y_true: [-50, -10, +80]
y_pred: [-50, -10, +45]

MSE = 0.04

[1, 1, 0]

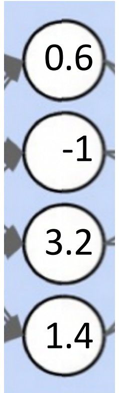
ACC = 1.00

Classification loss

- Softmax
- Cross entropy

How different are the two probability distributions?

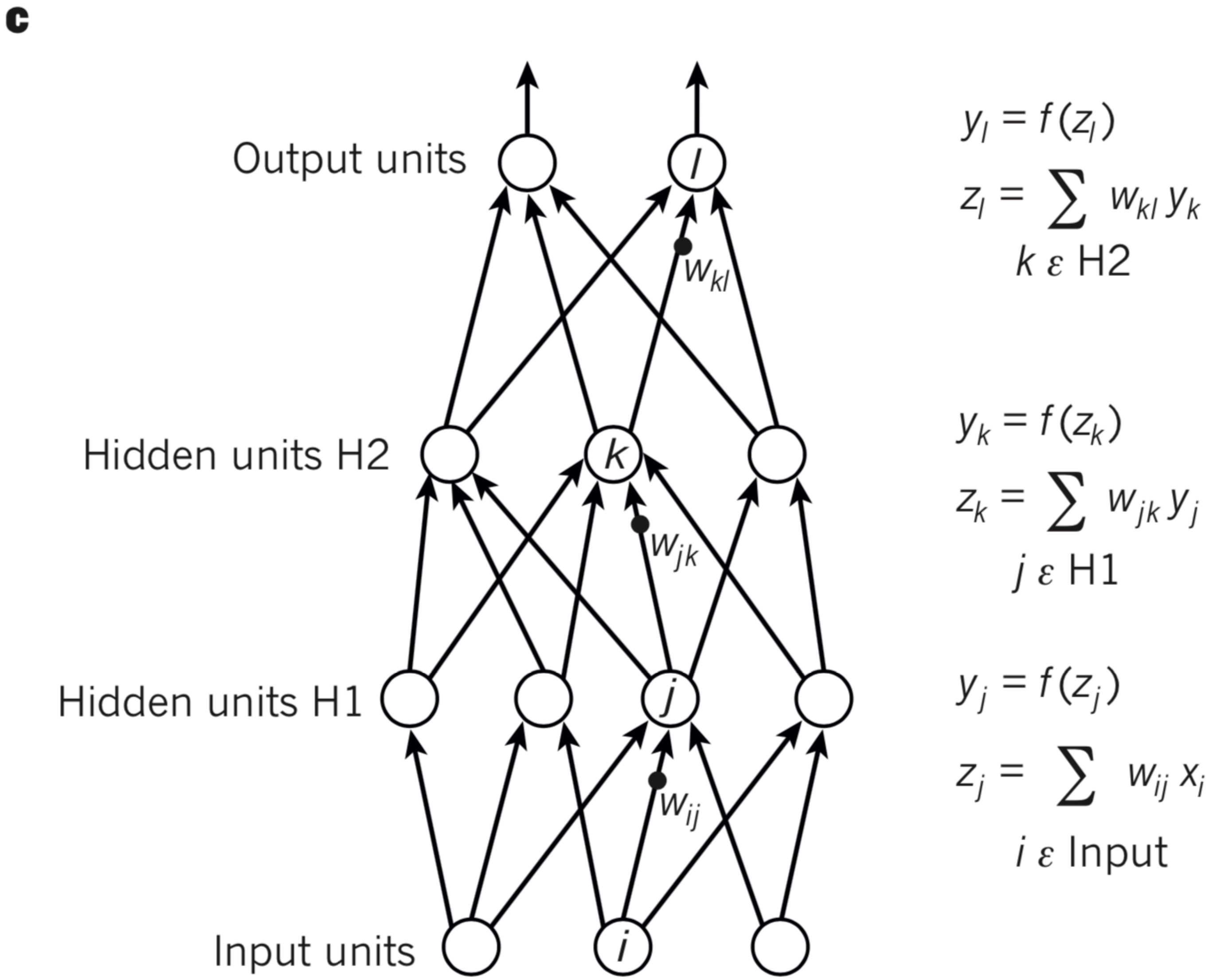
scores $\xrightarrow{\text{softmax}}$

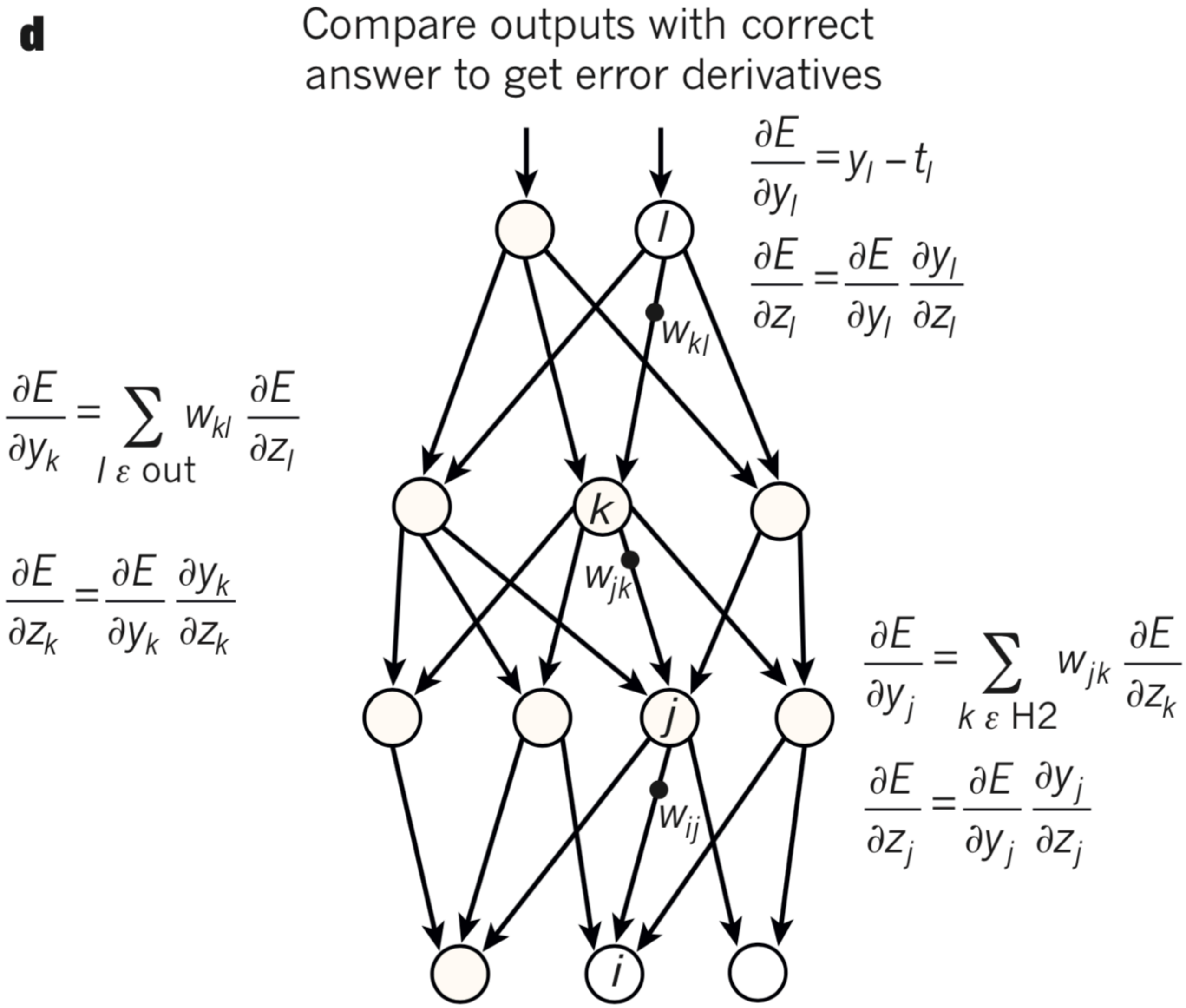


$$\frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

probabilities	true probabilities	cross entropy
0.059	0.1	0.866
0.012	0.2	
0.797	**0.6**	
0.132	0.1	

$$H(p,q) = -\sum_{x} p(x)\, \log q(x).$$


c
Output units
l
w_{kl}
Hidden units H2
k
w_{jk}
Hidden units H1
j
w_{ij}
Input units
i
$y_l = f(z_l)$
$z_l = \sum_{k \, \varepsilon \, \mathrm{H2}} w_{kl} \, y_k$
$y_k = f(z_k)$
$z_k = \sum_{j \, \varepsilon \, \mathrm{H1}} w_{jk} \, y_j$
$y_j = f(z_j)$
$z_j = \sum_{i \, \varepsilon \, \mathrm{Input}} w_{ij} \, x_i$



Compare outputs with correct answer to get error derivatives

$$\frac{\partial E}{\partial y_l} = y_l - t_l$$

$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l} \frac{\partial y_l}{\partial z_l}$$

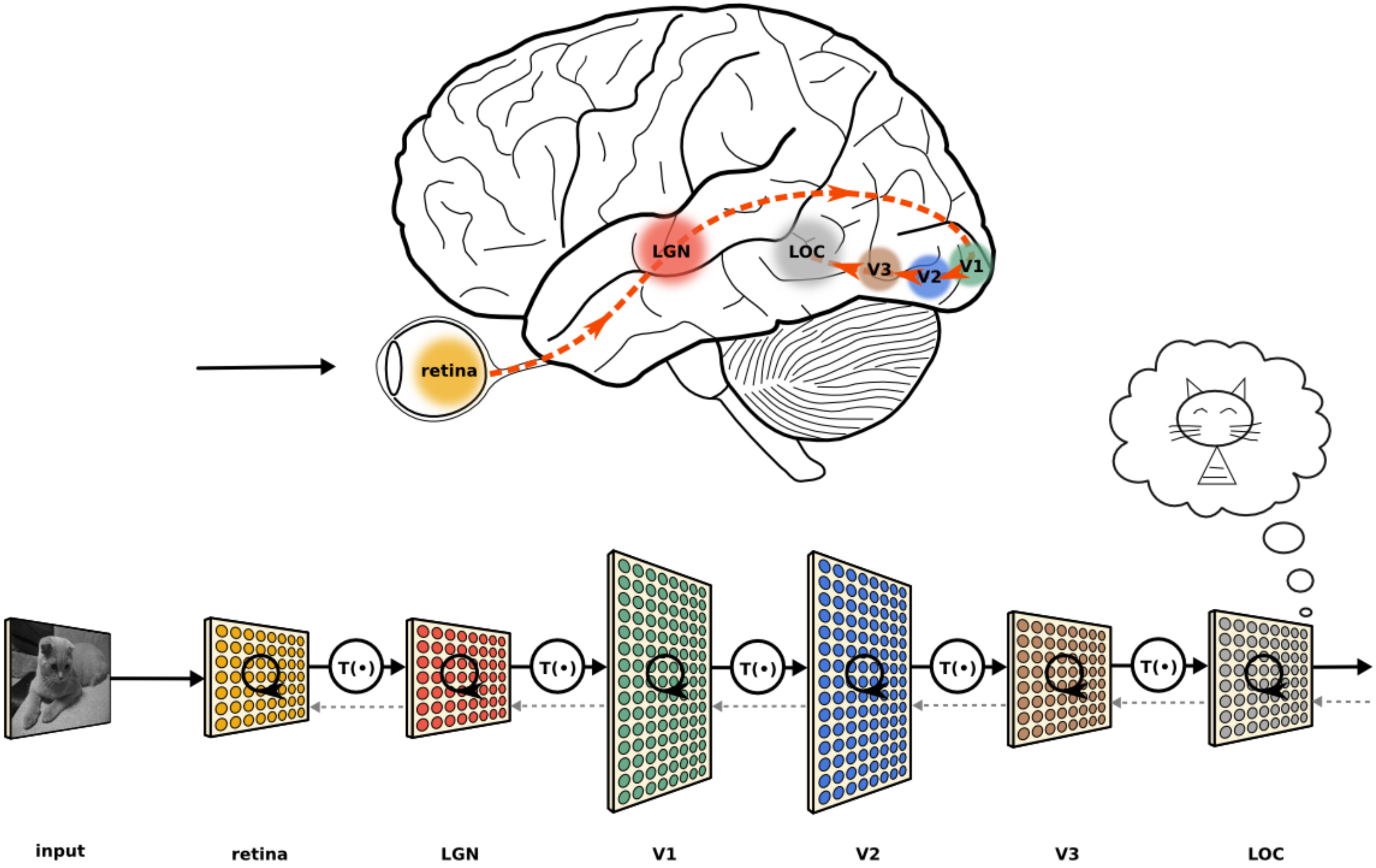
$$\frac{\partial E}{\partial y_k} = \sum_{l \, \varepsilon \, \mathrm{out}} w_{kl} \frac{\partial E}{\partial z_l}$$

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$$

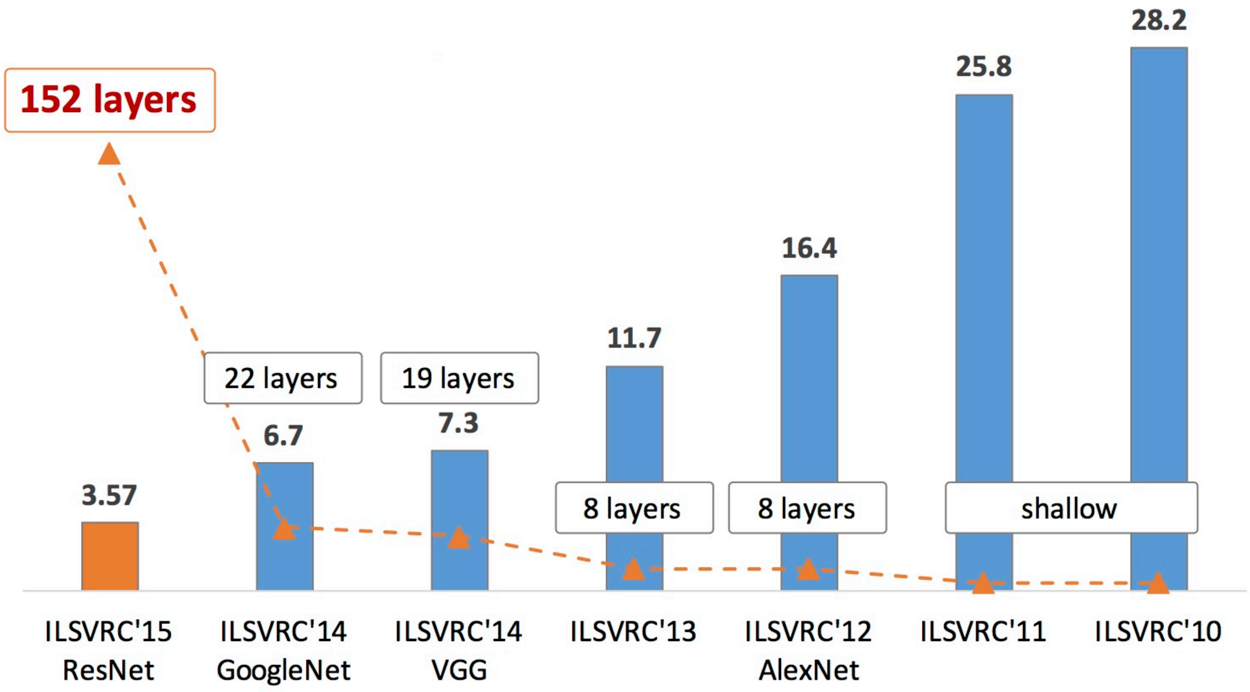
$$\frac{\partial E}{\partial y_j} = \sum_{k \, \varepsilon \, \mathrm{H2}} w_{jk} \frac{\partial E}{\partial z_k}$$

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial z_j}$$

Human Vision System



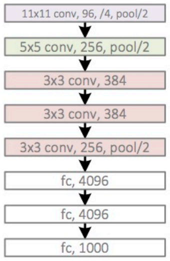
Increasing Depth Works…



ImageNet Classification top-5 error (%)

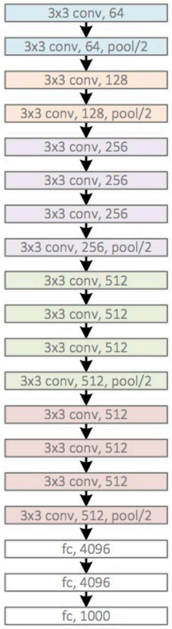
Example CNNs Structures from ILSVRC Winners

AlexNet, 8 layers
(ILSVRC 2012)



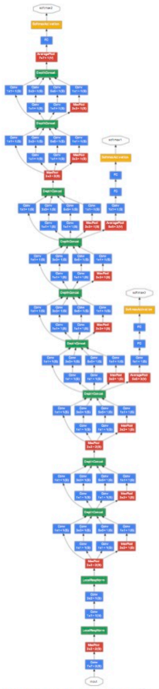
11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

VGG, 19 layers
(ILSVRC 2014)



3x3 conv, 64
3x3 conv, 64, pool/2
3x3 conv, 128
3x3 conv, 128, pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
fc, 4096
fc, 4096
fc, 1000

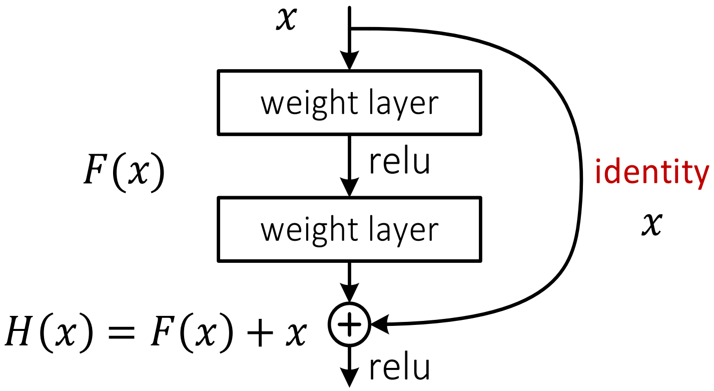
GoogleNet, 22 layers
(ILSVRC 2014)



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016.

Deep Residual Learning

- $F(x)$ is a residual mapping w.r.t. identity



- If identity were optimal, easy to set weights as 0
- If optimal mapping is closer to identity, easier to find small fluctuations

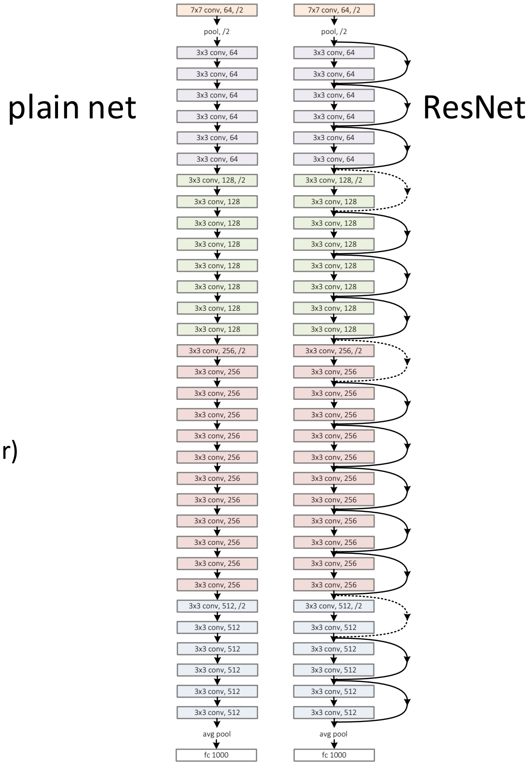
Deep Residual Learning

Network "Design"

- Keep it simple

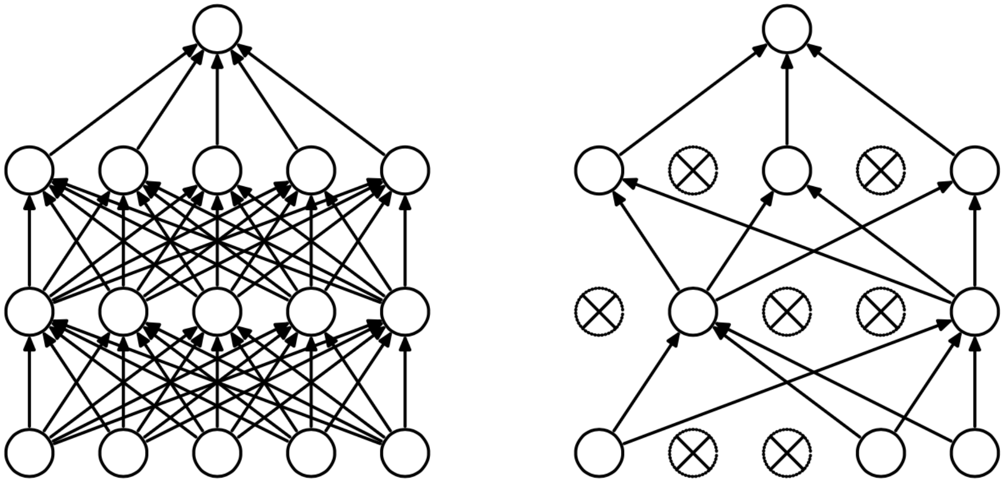
- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2 (~same complexity per layer)
  - Simple design; just deep!

- Other remarks:
  - no hidden fc
  - no dropout



Dropout

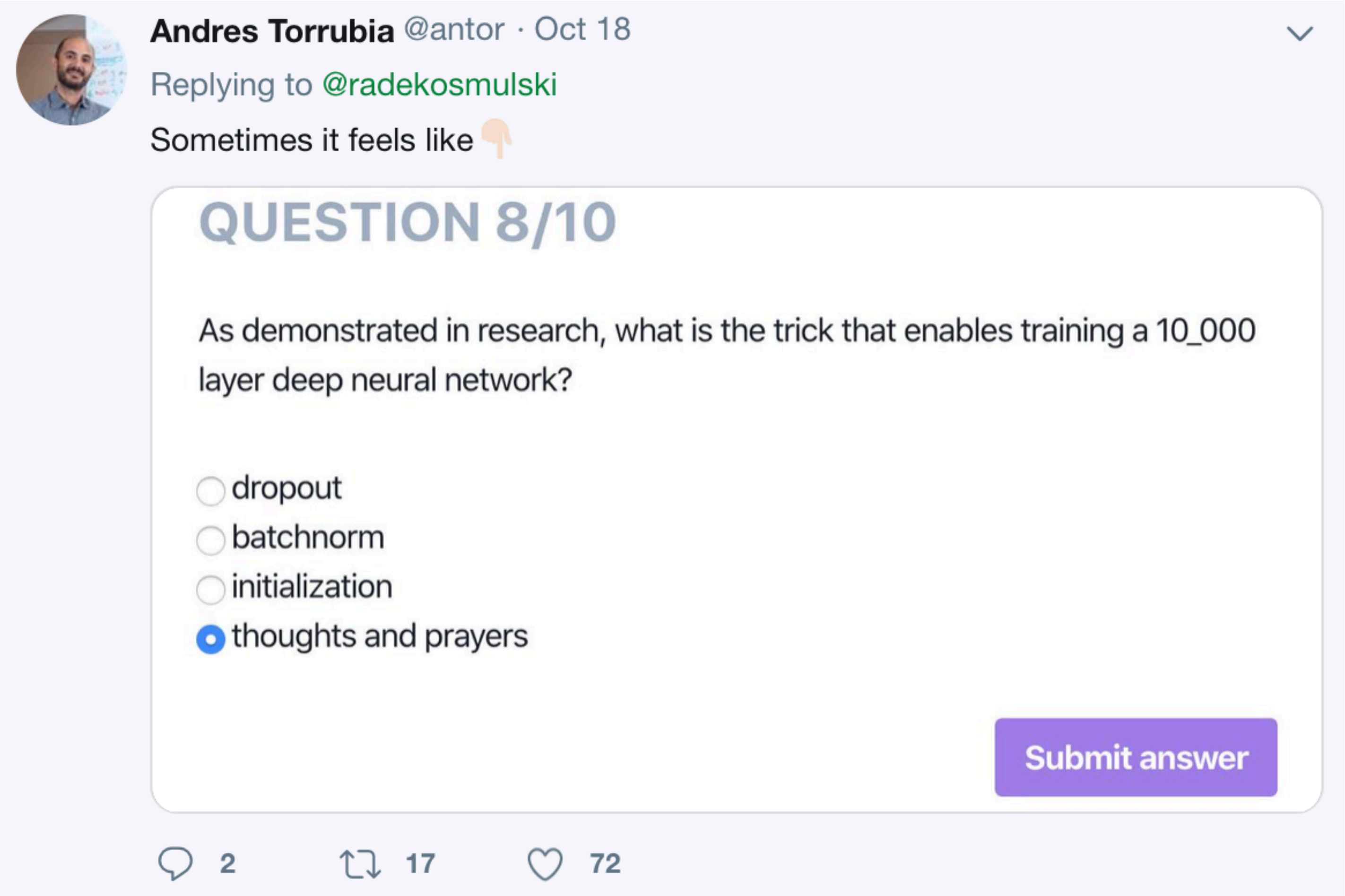
Srivastava, Hinton, Krizhevsky, Sutskever and Salakhutdinov



(a) Standard Neural Net

(b) After applying dropout.

Figure 1: Dropout Neural Net Model. **Left**: A standard neural net with 2 hidden layers. **Right**: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.



Andres Torrubia @antor · Oct 18

Replying to @radekosmulski

Sometimes it feels like 👇🏻

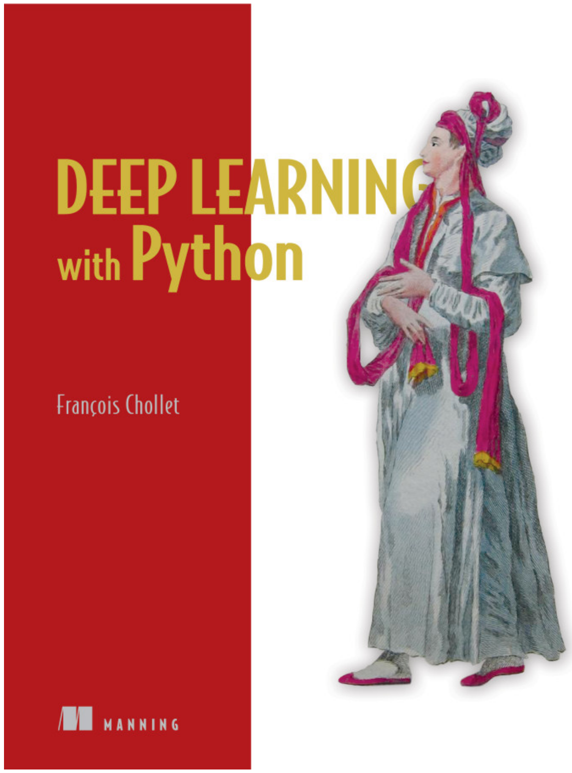
QUESTION 8/10

As demonstrated in research, what is the trick that enables training a 10_000 layer deep neural network?

○ dropout
○ batchnorm
○ initialization
● thoughts and prayers

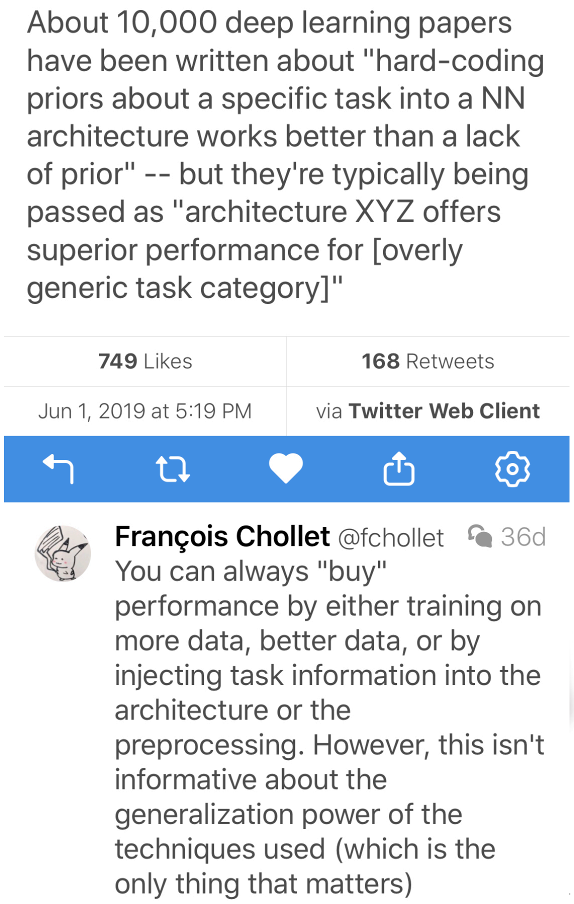
Essential Deep Learning Resources

- Books
  - Michael Nielsen’s free online book on deep learning
  - Deep Learning textbook (Goodfellow et al, 2016)
  - Deep Learning with Python (Keras book)
- Online courses
  - FastAI
  - Stanford CS231n
  - Andrew Ng
- Python
  - Python Cookbook, 3rd Edition
- Other resources
  - PyTorch tutorials
  - Kaggle kernels / discussions; No free hunch interviews
  - Twitter
  - arxiv-sanity



Improving Models with Domain Knowledge

- Books
  - Michael Nielsen’s free online book on deep learning
  - Deep Learning textbook (Goodfellow et al, 2016)
  - Deep Learning with Python (Keras book)
- Online courses
  - FastAI
  - Stanford CS231n
  - Andrew Ng
- Python
  - Python Cookbook, 3rd Edition
- Other resources
  - PyTorch tutorials
  - Kaggle kernels / discussions; No free hunch interviews
  - Twitter
  - arxiv-sanity



About 10,000 deep learning papers have been written about "hard-coding priors about a specific task into a NN architecture works better than a lack of prior" -- but they're typically being passed as "architecture XYZ offers superior performance for [overly generic task category]"

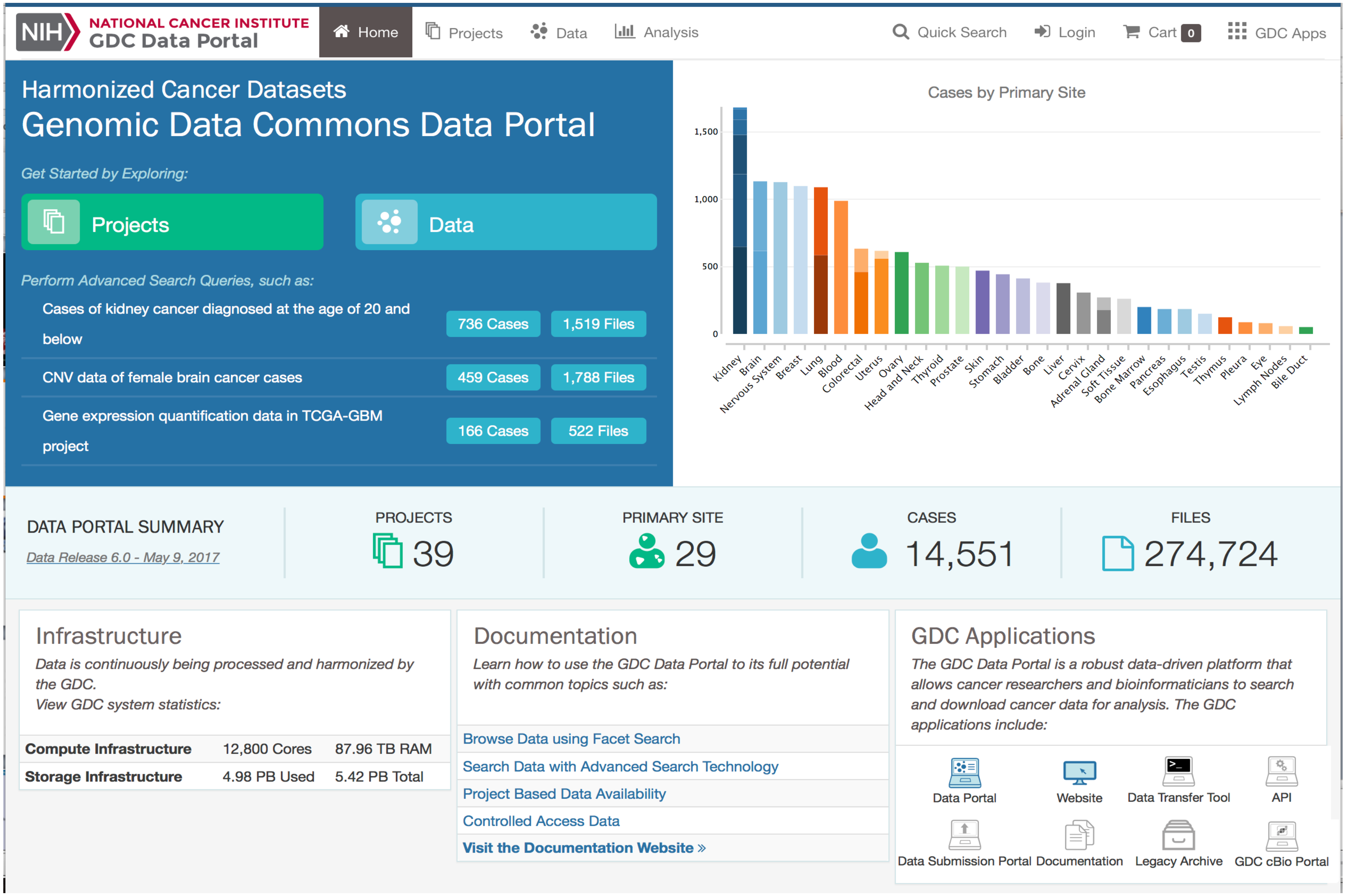
749 Likes | **168** Retweets

Jun 1, 2019 at 5:19 PM | via **Twitter Web Client**

François Chollet @fchollet 36d

You can always "buy" performance by either training on more data, better data, or by injecting task information into the architecture or the preprocessing. However, this isn't informative about the generalization power of the techniques used (which is the only thing that matters)

Cancer Type Classification with RNAseq



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Data | Analysis | Quick Search | Login | Cart 0 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	736 Cases	1,519 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files

Cases by Primary Site

DATA PORTAL SUMMARY	PROJECTS	PRIMARY SITE	CASES	FILES
Data Release 6.0 - May 9, 2017	39	29	14,551	274,724

Infrastructure

Data is continuously being processed and harmonized by the GDC.
View GDC system statistics:

Compute Infrastructure	12,800 Cores	87.96 TB RAM
Storage Infrastructure	4.98 PB Used	5.42 PB Total

Documentation

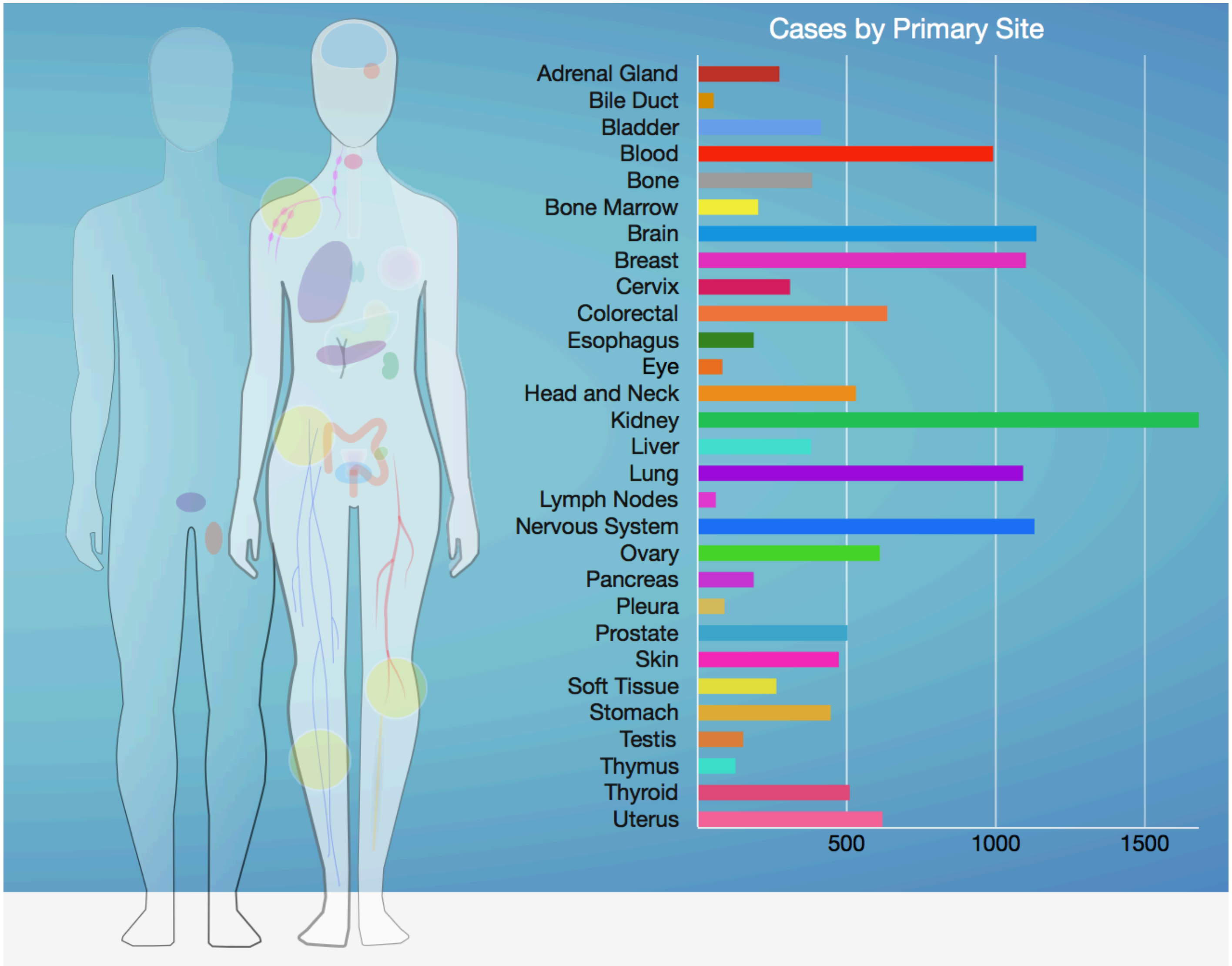
Learn how to use the GDC Data Portal to its full potential with common topics such as:

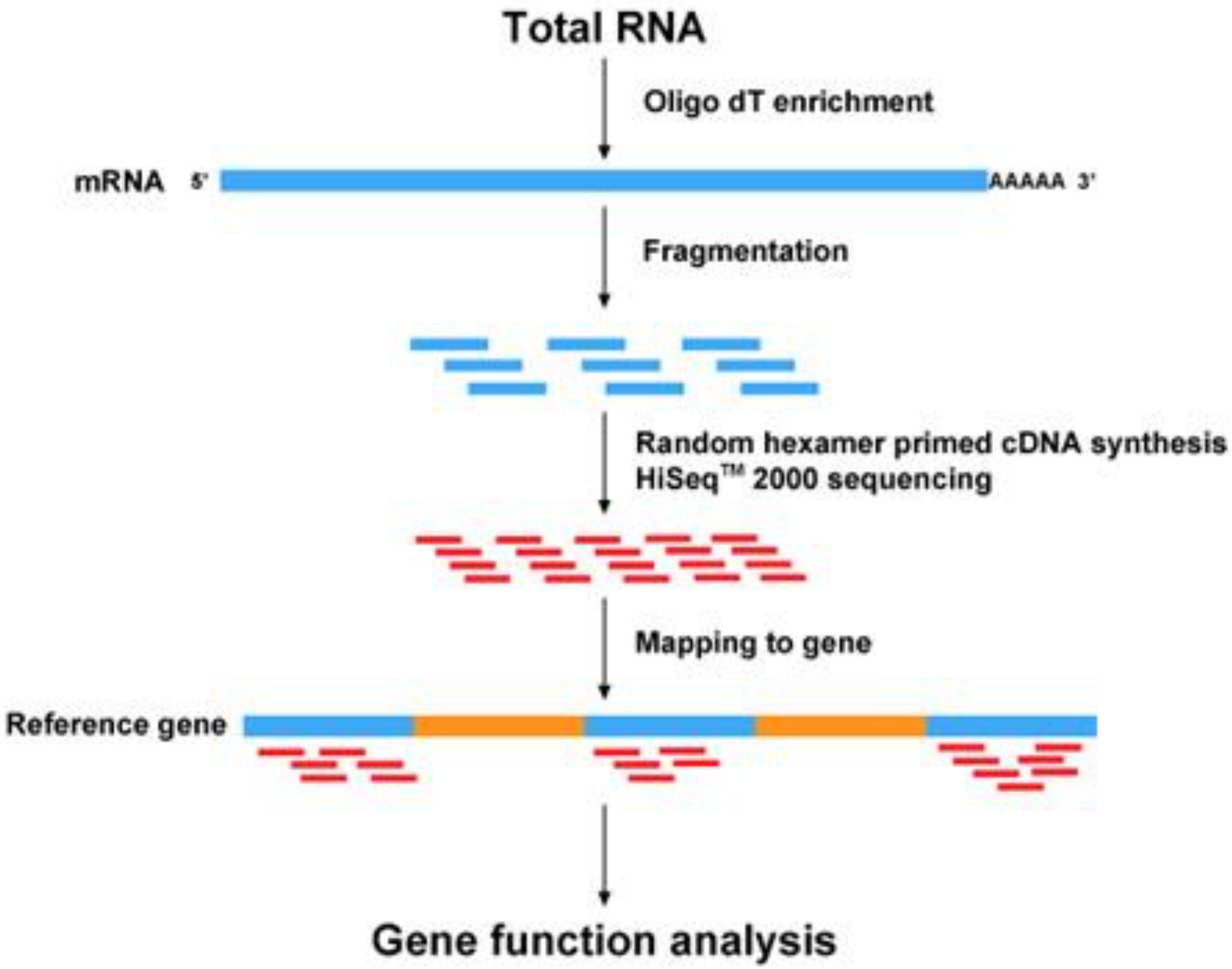
- Browse Data using Facet Search
- Search Data with Advanced Search Technology
- Project Based Data Availability
- Controlled Access Data
- **Visit the Documentation Website »**

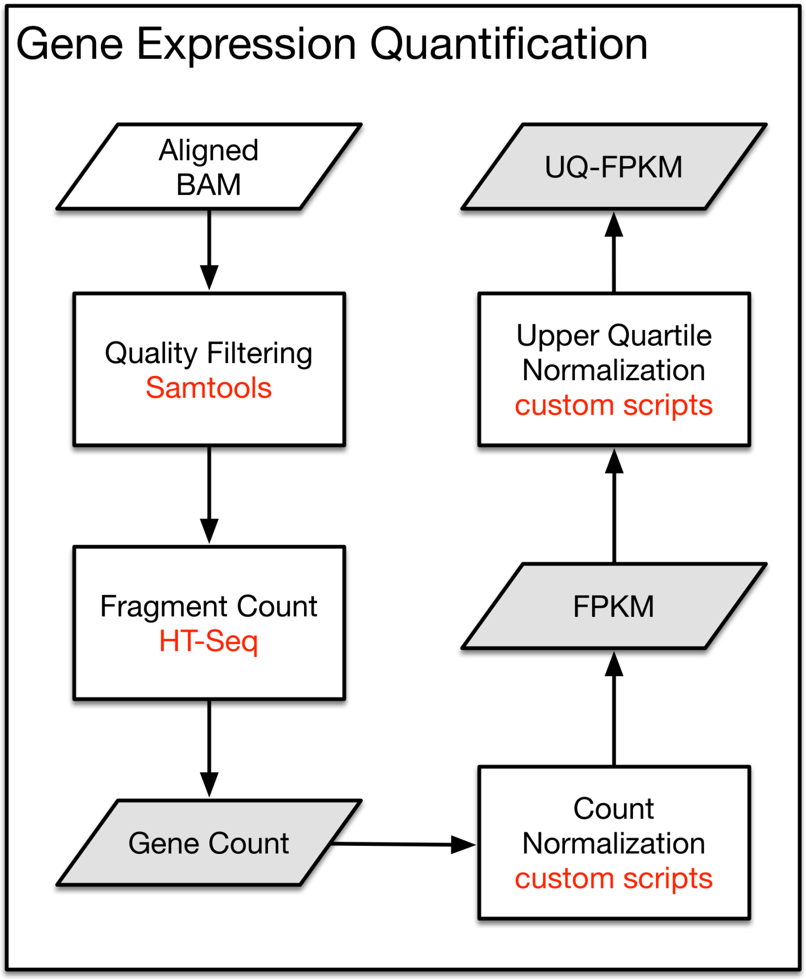
GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

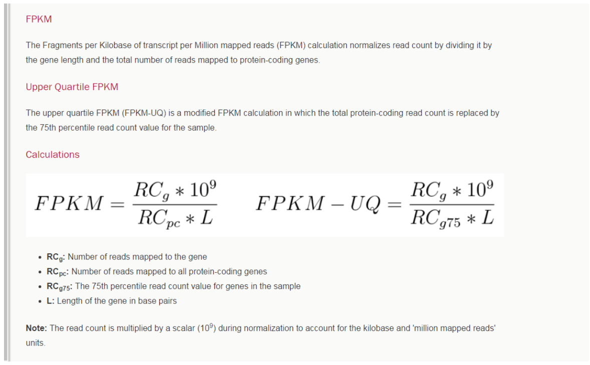
Data Portal | Website | Data Transfer Tool | API
Data Submission Portal | Documentation | Legacy Archive | GDC cBio Portal


Cases by Primary Site
Adrenal Gland
Bile Duct
Bladder
Blood
Bone
Bone Marrow
Brain
Breast
Cervix
Colorectal
Esophagus
Eye
Head and Neck
Kidney
Liver
Lung
Lymph Nodes
Nervous System
Ovary
Pancreas
Pleura
Prostate
Skin
Soft Tissue
Stomach
Testis
Thymus
Thyroid
Uterus
500
1000
1500


Total RNA
Oligo dT enrichment
mRNA
5'
AAAAA
3'
Fragmentation
Random hexamer primed cDNA synthesis
HiSeq™ 2000 sequencing
Mapping to gene
Reference gene
Gene function analysis



RPKM (reads per kilobase per million mapped reads)
Upper Quantile (UQ)



FPKM

The Fragments per Kilobase of transcript per Million mapped reads (FPKM) calculation normalizes read count by dividing it by the gene length and the total number of reads mapped to protein-coding genes.

Upper Quartile FPKM

The upper quartile FPKM (FPKM-UQ) is a modified FPKM calculation in which the total protein-coding read count is replaced by the 75th percentile read count value for the sample.

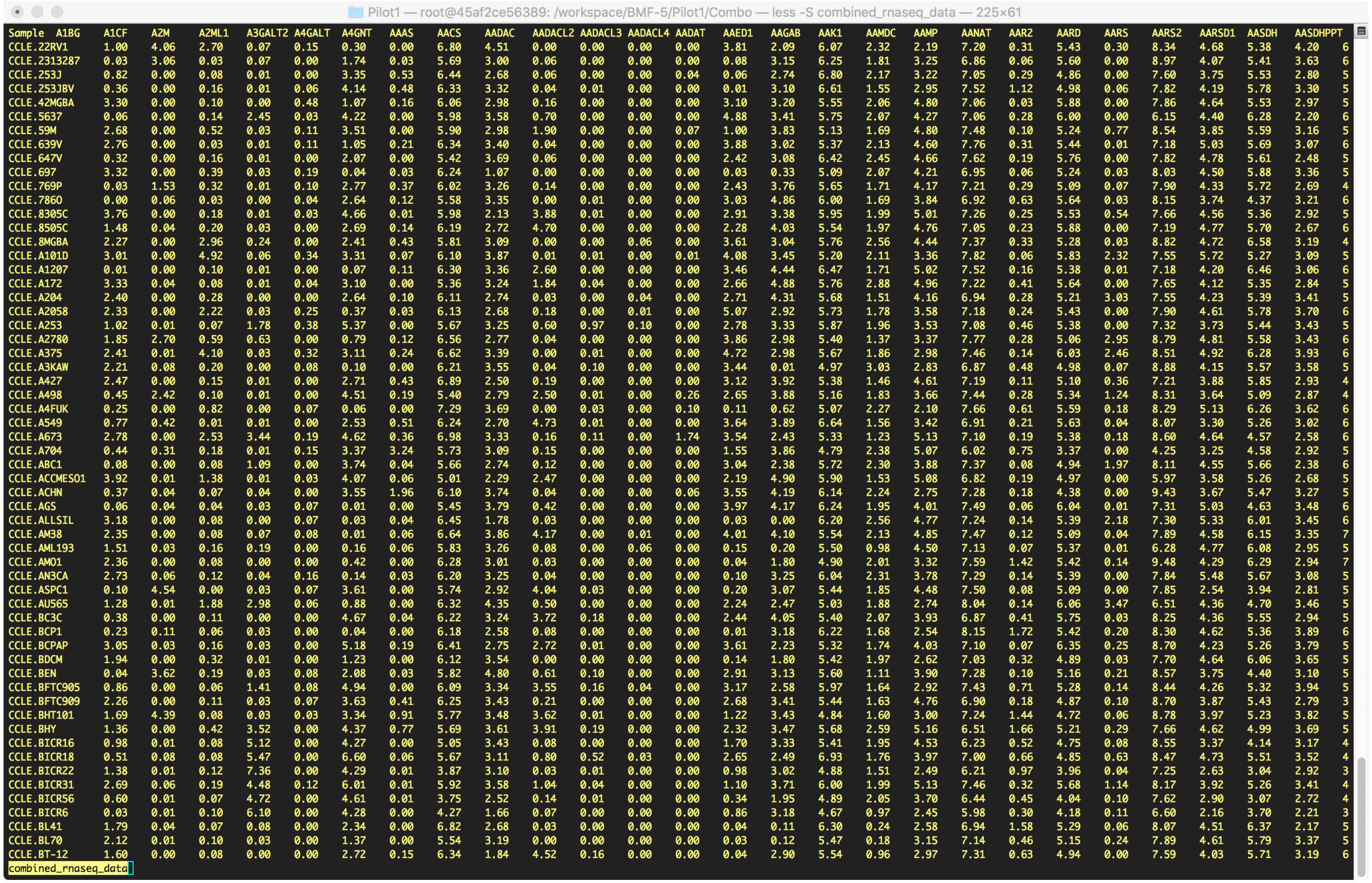
Calculations

$$FPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \qquad FPKM - UQ = \frac{RC_g * 10^9}{RC_{g75} * L}$$

- **RC_g:** Number of reads mapped to the gene
- **RC_{pc}:** Number of reads mapped to all protein-coding genes
- **RC_{g75}:** The 75th percentile read count value for genes in the sample
- **L:** Length of the gene in base pairs

Note: The read count is multiplied by a scalar (10^9) during normalization to account for the kilobase and 'million mapped reads' units.

Each Sample has > 60,000 columns

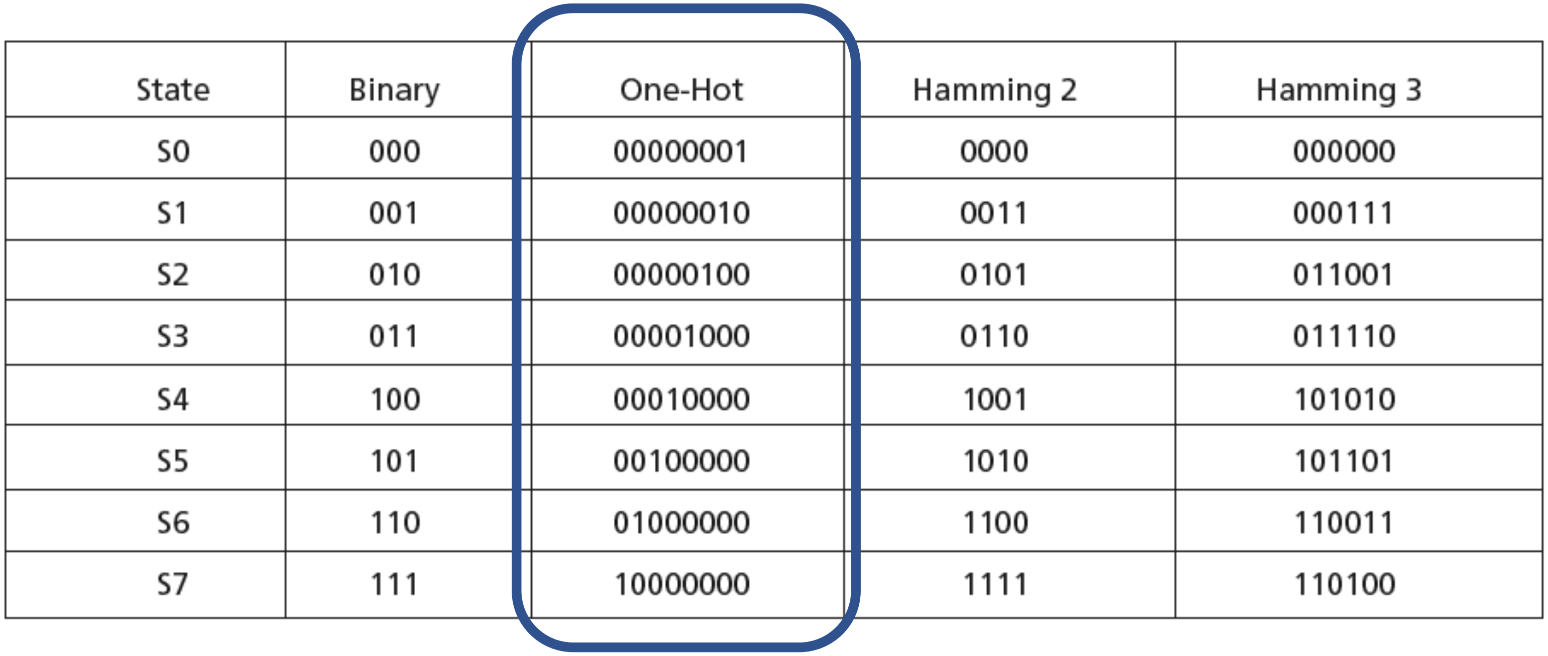

Pilot1 — root@45af2ce56389: /workspace/BMF-5/Pilot1/Combo — less -S combined_rnaseq_data — 225×61

```
Sample  A1BG    A1CF    A2M     A2ML1   A3GALT2 A4GALT  A4GNT   AAAS    AACS    AADAC   AADACL2 AADACL3 AADACL4 AADAT   AAED1   AAGAB   AAK1    AAMDC   AAMP    AANAT   AAR2    AARD    AARS    AARS2   AARSD1  AASDH   AASDHPPT
CCLE.22RV1      1.00    4.06    2.70    0.07    0.15    0.30    0.00    6.80    4.51    0.00    0.00    0.00    0.00    3.81    2.09    6.07    2.32    2.19    7.20    0.31    5.43    0.30    8.34    4.68    5.38    4.20    6
CCLE.2313287    0.03    3.06    0.03    0.07    0.00    1.74    0.03    5.69    3.00    0.06    0.00    0.00    0.00    0.08    3.15    6.25    1.81    3.25    6.86    0.06    5.60    0.00    8.97    4.07    5.41    3.63    6
CCLE.253J       0.82    0.00    0.08    0.01    0.00    3.35    0.53    6.44    2.68    0.06    0.00    0.00    0.04    0.06    2.74    6.80    2.17    3.22    7.05    0.29    4.86    0.00    7.60    3.75    5.53    2.80    5
CCLE.253JBV     0.36    0.00    0.16    0.01    0.06    4.14    0.48    6.33    3.32    0.04    0.01    0.00    0.00    0.01    3.10    6.61    1.55    2.95    7.52    1.12    4.98    0.06    7.82    4.19    5.78    3.30    5
CCLE.42MGBA     3.30    0.00    0.10    0.00    0.48    1.07    0.16    6.06    2.98    0.16    0.00    0.00    0.00    3.10    3.20    5.55    2.06    4.80    7.06    0.03    5.88    0.00    7.86    4.64    5.53    2.97    5
CCLE.5637       0.06    0.00    0.14    2.45    0.03    4.22    0.00    5.98    3.58    0.70    0.00    0.00    0.00    4.88    3.41    5.75    2.07    4.27    7.06    0.28    6.00    0.00    6.15    4.40    6.28    2.20    6
CCLE.59M        2.68    0.00    0.52    0.03    0.11    3.51    0.00    5.90    2.98    1.90    0.00    0.00    0.07    1.00    3.83    5.13    1.69    4.80    7.48    0.10    5.24    0.77    8.54    3.85    5.59    3.16    5
CCLE.639V       2.76    0.00    0.03    0.01    0.11    1.05    0.21    6.34    3.40    0.04    0.00    0.00    0.00    3.88    3.02    5.37    2.13    4.60    7.76    0.31    5.44    0.01    7.18    5.03    5.69    3.07    6
CCLE.647V       0.32    0.00    0.16    0.01    0.00    2.07    0.00    5.42    3.69    0.06    0.00    0.00    0.00    2.42    3.08    6.42    2.45    4.66    7.62    0.19    5.76    0.00    7.82    4.78    5.61    2.48    5
CCLE.697        3.32    0.00    0.39    0.03    0.19    0.04    0.03    6.24    1.07    0.00    0.00    0.00    0.00    0.03    0.33    5.09    2.07    4.21    6.95    0.06    5.24    0.03    8.03    4.50    5.88    3.36    5
CCLE.769P       0.03    1.53    0.32    0.01    0.10    2.77    0.37    6.02    3.26    0.14    0.00    0.00    0.00    2.43    3.76    5.65    1.71    4.17    7.21    0.29    5.09    0.07    7.90    4.33    5.72    2.69    4
CCLE.786O       0.00    0.06    0.03    0.00    0.04    2.64    0.12    5.58    3.35    0.00    0.01    0.00    0.00    3.03    4.86    6.00    1.69    3.84    6.92    0.63    5.64    0.03    8.15    3.74    4.37    3.21    6
CCLE.8305C      3.76    0.00    0.18    0.01    0.03    4.66    0.01    5.98    2.13    3.88    0.01    0.00    0.00    2.91    3.38    5.95    1.99    5.01    7.26    0.25    5.53    0.54    7.66    4.56    5.36    2.92    5
CCLE.8505C      1.48    0.04    0.20    0.03    0.00    2.69    0.14    6.19    2.72    4.70    0.00    0.00    0.00    2.28    4.03    5.54    1.97    4.76    7.05    0.23    5.88    0.00    7.19    4.77    5.70    2.67    6
CCLE.8MGBA      2.27    0.00    2.96    0.24    0.00    2.41    0.43    5.81    3.09    0.00    0.00    0.06    0.00    3.61    3.04    5.76    2.56    4.44    7.37    0.33    5.28    0.03    8.82    4.72    6.58    3.19    4
CCLE.A101D      3.01    0.00    4.92    0.06    0.34    3.31    0.07    6.10    3.87    0.01    0.01    0.00    0.01    4.08    3.45    5.20    2.11    3.36    7.82    0.06    5.83    2.32    7.55    5.72    5.27    3.09    5
CCLE.A1207      0.01    0.00    0.10    0.01    0.00    0.07    0.11    6.30    3.36    2.60    0.00    0.00    0.00    3.46    4.44    6.47    1.71    5.02    7.52    0.16    5.38    0.01    7.18    4.20    6.46    3.06    6
CCLE.A172       3.33    0.04    0.08    0.01    0.04    3.10    0.00    5.36    3.24    1.84    0.04    0.00    0.00    2.66    4.88    5.76    2.88    4.96    7.22    0.41    5.64    0.00    7.65    4.12    5.35    2.84    5
CCLE.A204       2.40    0.00    0.28    0.00    0.00    2.64    0.10    6.11    2.74    0.03    0.00    0.04    0.00    2.71    4.31    5.68    1.51    4.16    6.94    0.28    5.21    3.03    7.55    4.23    5.39    3.41    5
CCLE.A2058      2.33    0.00    2.22    0.03    0.25    0.37    0.03    6.13    2.68    0.18    0.00    0.01    0.00    5.07    2.92    5.73    1.78    3.58    7.18    0.24    5.43    0.00    7.90    4.61    5.78    3.70    6
CCLE.A253       1.02    0.01    0.07    1.78    0.38    5.37    0.00    5.67    3.25    0.60    0.97    0.10    0.00    2.78    3.33    5.87    1.96    3.53    7.08    0.46    5.38    0.00    7.32    3.73    5.44    3.43    5
CCLE.A2780      1.85    2.70    0.59    0.63    0.00    0.79    0.12    6.56    2.77    0.04    0.00    0.00    0.00    3.86    2.98    5.40    1.37    3.37    7.77    0.28    5.06    2.95    8.79    4.81    5.58    3.43    6
CCLE.A375       2.41    0.01    4.10    0.03    0.32    3.11    0.24    6.62    3.39    0.00    0.01    0.00    0.00    4.72    2.98    5.67    1.86    2.98    7.46    0.14    6.03    2.46    8.51    4.92    6.28    3.93    6
CCLE.A3KAW      2.21    0.08    0.20    0.00    0.08    0.10    0.00    6.21    3.55    0.04    0.10    0.00    0.00    3.44    0.01    4.97    3.03    2.83    6.87    0.48    4.98    0.07    8.88    4.15    5.57    3.58    5
CCLE.A427       2.47    0.00    0.15    0.01    0.00    2.71    0.43    6.89    2.50    0.19    0.00    0.00    0.00    3.12    3.92    5.38    1.46    4.61    7.19    0.11    5.10    0.36    7.21    3.88    5.85    2.93    4
CCLE.A498       0.45    2.42    0.10    0.01    0.00    4.51    0.19    5.40    2.79    2.50    0.01    0.00    0.26    2.65    3.88    5.16    1.83    3.66    7.44    0.28    5.34    1.24    8.31    3.64    5.09    2.87    4
CCLE.A4FUK      0.25    0.00    0.82    0.00    0.07    0.06    0.00    7.29    3.69    0.00    0.03    0.00    0.10    0.11    0.62    5.07    2.27    2.10    7.66    0.61    5.59    0.18    8.29    5.13    6.26    3.62    6
CCLE.A549       0.77    0.42    0.01    0.01    0.00    2.53    0.51    6.24    2.70    4.73    0.01    0.00    0.00    3.64    3.89    6.64    1.56    3.42    6.91    0.21    5.63    0.04    8.07    3.30    5.26    3.02    6
CCLE.A673       2.78    0.00    2.53    3.44    0.19    4.62    0.36    6.98    3.33    0.16    0.11    0.00    1.74    3.54    2.43    5.33    1.23    5.13    7.10    0.19    5.38    0.18    8.60    4.64    4.57    2.58    6
CCLE.A704       0.44    0.31    0.18    0.01    0.15    3.37    3.24    5.73    3.09    0.15    0.00    0.00    0.00    1.55    3.86    4.79    2.38    5.07    6.02    0.75    3.37    0.00    4.25    3.25    4.58    2.92    5
CCLE.ABC1       0.08    0.00    0.08    1.09    0.00    3.74    0.04    5.66    2.74    0.12    0.00    0.00    0.00    3.04    2.38    5.72    2.30    3.88    7.37    0.08    4.94    1.97    8.11    4.55    5.66    2.38    6
CCLE.ACCMESO1   3.92    0.01    1.38    0.01    0.03    4.07    0.06    5.01    2.29    2.47    0.00    0.00    0.00    2.19    4.90    5.90    1.53    5.08    6.82    0.19    4.97    0.00    5.97    3.58    5.26    2.68    5
CCLE.ACHN       0.37    0.04    0.07    0.04    0.00    3.55    1.96    6.10    3.74    0.04    0.00    0.00    0.06    3.55    4.19    6.14    2.24    2.75    7.28    0.18    4.38    0.00    9.43    3.67    5.47    3.27    5
CCLE.AGS        0.06    0.04    0.04    0.03    0.07    0.01    0.00    5.45    3.79    0.42    0.00    0.00    0.00    3.97    4.17    6.24    1.95    4.01    7.49    0.06    6.04    0.01    7.31    5.03    4.63    3.48    6
CCLE.ALLSIL     3.18    0.00    0.08    0.00    0.07    0.03    0.04    6.45    1.78    0.03    0.00    0.00    0.00    0.03    0.00    6.20    2.56    4.77    7.24    0.14    5.39    2.18    7.30    5.33    6.01    3.45    6
CCLE.AM38       2.35    0.00    0.08    0.07    0.08    0.01    0.06    6.64    3.86    4.17    0.00    0.01    0.00    4.01    4.10    5.54    2.13    4.85    7.47    0.12    5.09    0.04    7.89    4.58    6.15    3.35    7
CCLE.AML193     1.51    0.03    0.16    0.19    0.00    0.16    0.06    5.83    3.26    0.08    0.00    0.06    0.00    0.15    0.20    5.50    0.98    4.50    7.13    0.07    5.37    0.01    6.28    4.77    6.08    2.95    5
CCLE.AMO1       2.36    0.00    0.08    0.00    0.00    0.42    0.00    6.28    3.01    0.03    0.00    0.00    0.00    0.04    1.80    4.90    2.01    3.32    7.59    1.42    5.42    0.14    9.48    4.29    6.29    2.94    7
CCLE.AN3CA      2.73    0.06    0.12    0.04    0.16    0.14    0.03    6.20    3.25    0.04    0.00    0.00    0.00    0.10    3.25    6.04    2.31    3.78    7.29    0.14    5.39    0.00    7.84    5.48    5.67    3.08    5
CCLE.ASPC1      0.10    4.54    0.00    0.03    0.07    3.61    0.00    5.74    2.92    4.04    0.03    0.00    0.00    0.20    3.07    5.44    1.85    4.48    7.50    0.08    5.09    0.00    7.85    2.54    3.94    2.81    5
CCLE.AU565      1.28    0.01    1.88    2.98    0.06    0.88    0.00    6.32    4.35    0.50    0.00    0.00    0.00    2.24    2.47    5.03    1.88    2.74    8.04    0.14    6.06    3.47    6.51    4.36    4.70    3.46    5
CCLE.BC3C       0.38    0.00    0.11    0.00    0.00    4.67    0.04    6.22    3.24    3.72    0.18    0.00    0.00    2.44    4.05    5.40    2.07    3.93    6.87    0.41    5.75    0.03    8.25    4.36    5.55    2.94    5
CCLE.BCP1       0.23    0.11    0.06    0.03    0.00    0.04    0.00    6.18    2.58    0.08    0.00    0.00    0.00    0.01    3.18    6.22    1.68    2.54    8.15    1.72    5.42    0.20    8.30    4.62    5.36    3.89    6
CCLE.BCPAP      3.05    0.03    0.16    0.03    0.00    5.18    0.19    6.41    2.75    2.72    0.01    0.00    0.00    3.61    2.23    5.32    1.74    4.03    7.10    0.07    6.35    0.25    8.70    4.23    5.26    3.79    5
CCLE.BDCM       1.94    0.00    0.32    0.01    0.00    1.23    0.00    6.12    3.54    0.00    0.00    0.00    0.00    0.14    1.80    5.42    1.97    2.62    7.03    0.32    4.89    0.03    7.70    4.64    6.06    3.65    5
CCLE.BEN        0.04    3.62    0.19    0.03    0.08    2.08    0.03    5.82    4.80    0.61    0.10    0.00    0.00    2.91    3.13    5.60    1.11    3.90    7.28    0.10    5.16    0.21    8.57    3.75    4.40    3.10    5
CCLE.BFTC905    0.86    0.00    0.06    1.41    0.08    4.94    0.00    6.09    3.34    3.55    0.16    0.04    0.00    3.17    2.58    5.97    1.64    2.92    7.43    0.71    5.28    0.14    8.44    4.26    5.32    3.94    5
CCLE.BFTC909    2.26    0.00    0.11    0.03    0.07    3.63    0.41    6.25    3.43    0.21    0.00    0.00    0.00    2.68    3.41    5.44    1.63    4.76    6.90    0.18    4.87    0.10    8.70    3.87    5.43    2.79    3
CCLE.BHT101     1.69    4.39    0.08    0.03    0.03    3.34    0.91    5.77    3.48    3.62    0.01    0.00    0.00    1.22    3.43    4.84    1.60    3.00    7.24    1.44    4.72    0.06    8.78    3.97    5.23    3.82    5
CCLE.BHY        1.36    0.00    0.42    3.52    0.00    4.37    0.77    5.69    3.61    3.91    0.19    0.00    0.00    2.32    3.47    5.68    2.59    5.16    6.51    1.66    5.21    0.29    7.66    4.62    4.99    3.69    5
CCLE.BICR16     0.98    0.01    0.08    5.12    0.00    4.27    0.00    5.05    3.43    0.08    0.00    0.00    0.00    1.70    3.33    5.41    1.95    4.53    6.23    0.52    4.75    0.08    8.55    3.37    4.14    3.17    4
CCLE.BICR18     0.51    0.08    0.08    5.47    0.00    6.60    0.06    5.67    3.11    0.80    0.52    0.03    0.00    2.65    2.49    6.93    1.76    3.97    7.00    0.66    4.85    0.63    8.47    4.73    5.51    3.52    4
CCLE.BICR22     1.38    0.01    0.12    7.36    0.00    4.29    0.01    3.87    3.10    0.03    0.01    0.00    0.00    0.98    3.02    4.88    1.51    2.49    6.21    0.97    3.96    0.04    7.25    2.63    3.04    2.92    3
CCLE.BICR31     2.69    0.06    0.19    4.48    0.12    6.01    0.01    5.92    3.58    1.04    0.04    0.00    0.00    1.10    3.71    6.00    1.99    5.13    7.46    0.32    5.68    1.14    8.17    3.92    5.26    3.41    4
CCLE.BICR56     0.60    0.01    0.07    4.72    0.00    4.61    0.01    3.75    2.52    0.14    0.01    0.00    0.00    0.34    1.95    4.89    2.05    3.70    6.44    0.45    4.04    0.10    7.62    2.90    3.07    2.72    4
CCLE.BICR6      0.03    0.01    0.10    6.10    0.00    4.28    0.00    4.27    1.66    0.07    0.00    0.00    0.00    0.86    3.18    4.67    0.97    2.45    5.98    0.30    4.18    0.11    6.60    2.16    3.70    2.21    3
CCLE.BL41       1.79    0.04    0.07    0.08    0.00    2.34    0.00    6.82    2.68    0.03    0.00    0.00    0.00    0.04    0.11    6.30    0.24    2.58    6.94    1.58    5.29    0.06    8.07    4.51    6.37    2.17    5
CCLE.BL70       2.12    0.01    0.10    0.03    0.00    1.37    0.00    5.54    3.19    0.00    0.00    0.00    0.00    0.03    0.12    5.47    0.18    3.15    7.14    0.46    5.15    0.24    7.89    4.61    5.79    3.37    5
CCLE.BT-12      1.60    0.00    0.08    0.00    0.00    2.72    0.15    6.34    1.84    4.52    0.16    0.00    0.00    0.04    2.90    5.54    0.96    2.97    7.31    0.63    4.94    0.00    7.59    4.03    5.71    3.19    6
combined_rnaseq_data
```

Encode numeric columns

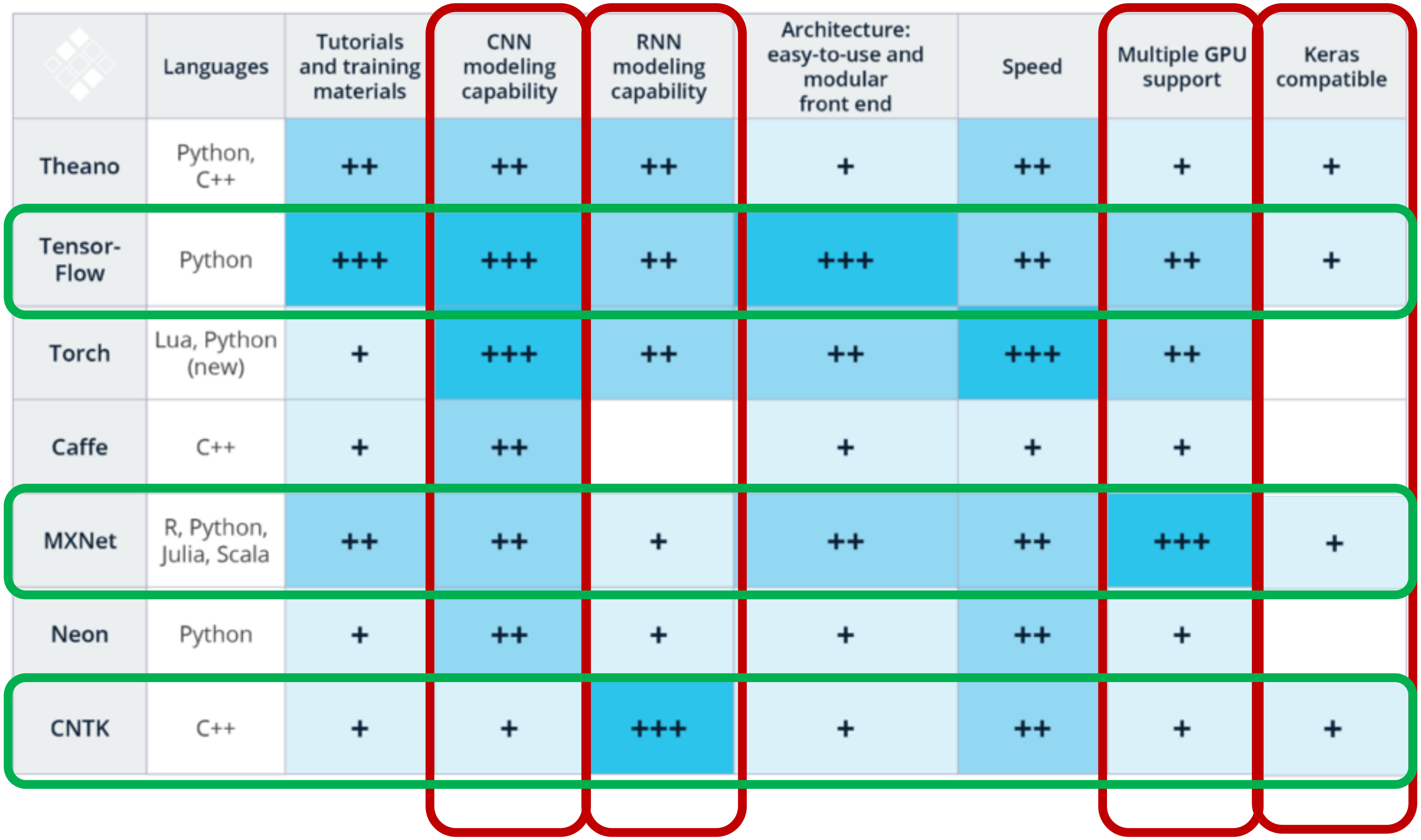
- Log transformation: log (1+x)
- Standard scalar: z score
- Rank => Gaussian
- Discretization
- Mark missing values

One Hot Encoding of Categories



State	Binary	One-Hot	Hamming 2	Hamming 3
S0	000	00000001	0000	000000
S1	001	00000010	0011	000111
S2	010	00000100	0101	011001
S3	011	00001000	0110	011110
S4	100	00010000	1001	101010
S5	101	00100000	1010	101101
S6	110	01000000	1100	110011
S7	111	10000000	1111	110100

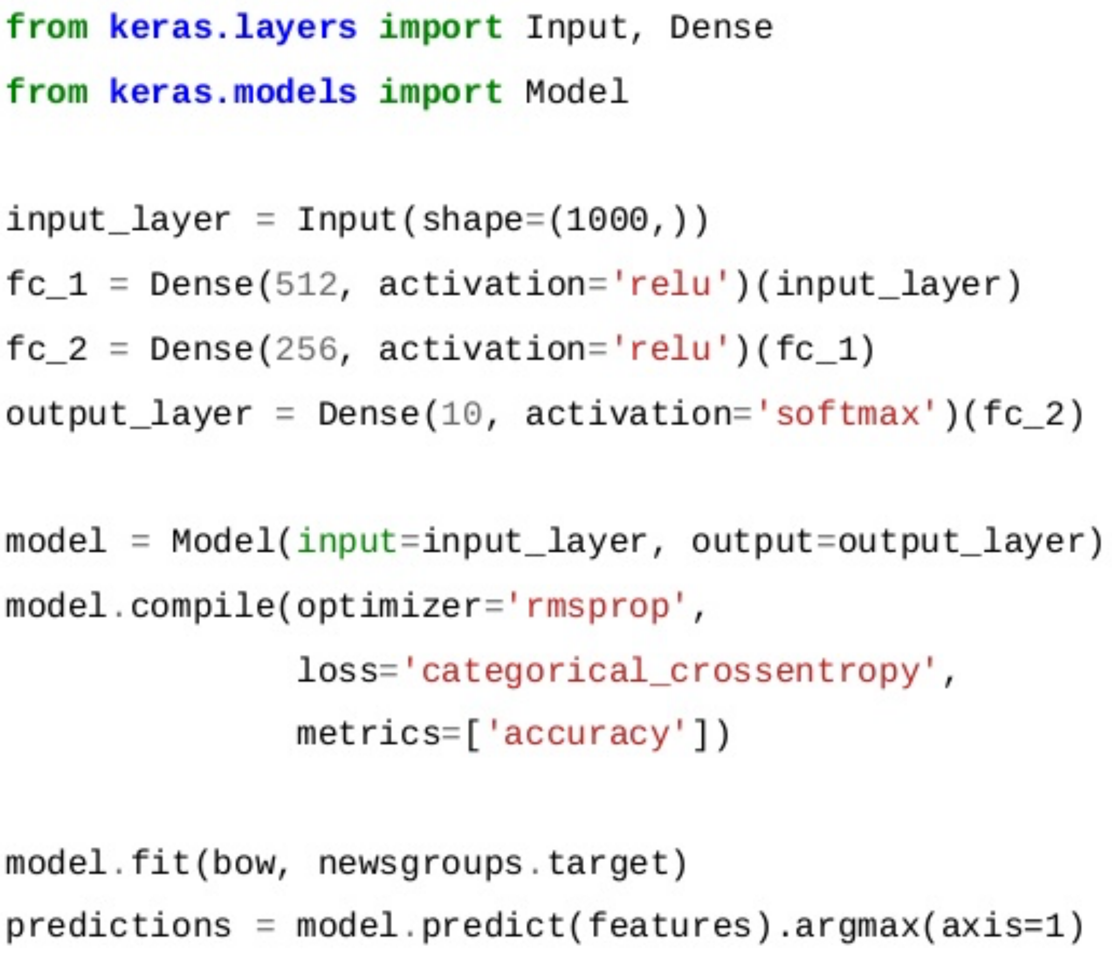
Open Source Framework Comparison



	Languages	Tutorials and training materials	CNN modeling capability	RNN modeling capability	Architecture: easy-to-use and modular front end	Speed	Multiple GPU support	Keras compatible
Theano	Python, C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Lua, Python (new)	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R, Python, Julia, Scala	++	++	+	++	++	+++	+
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	+

Keras

- https://keras.io/
- Minimalist, highly modular neural networks library
- Written in Python
- Capable of running on top of either TensorFlow/Theano and CNTK
- Developed with a focus on enabling fast experimentation



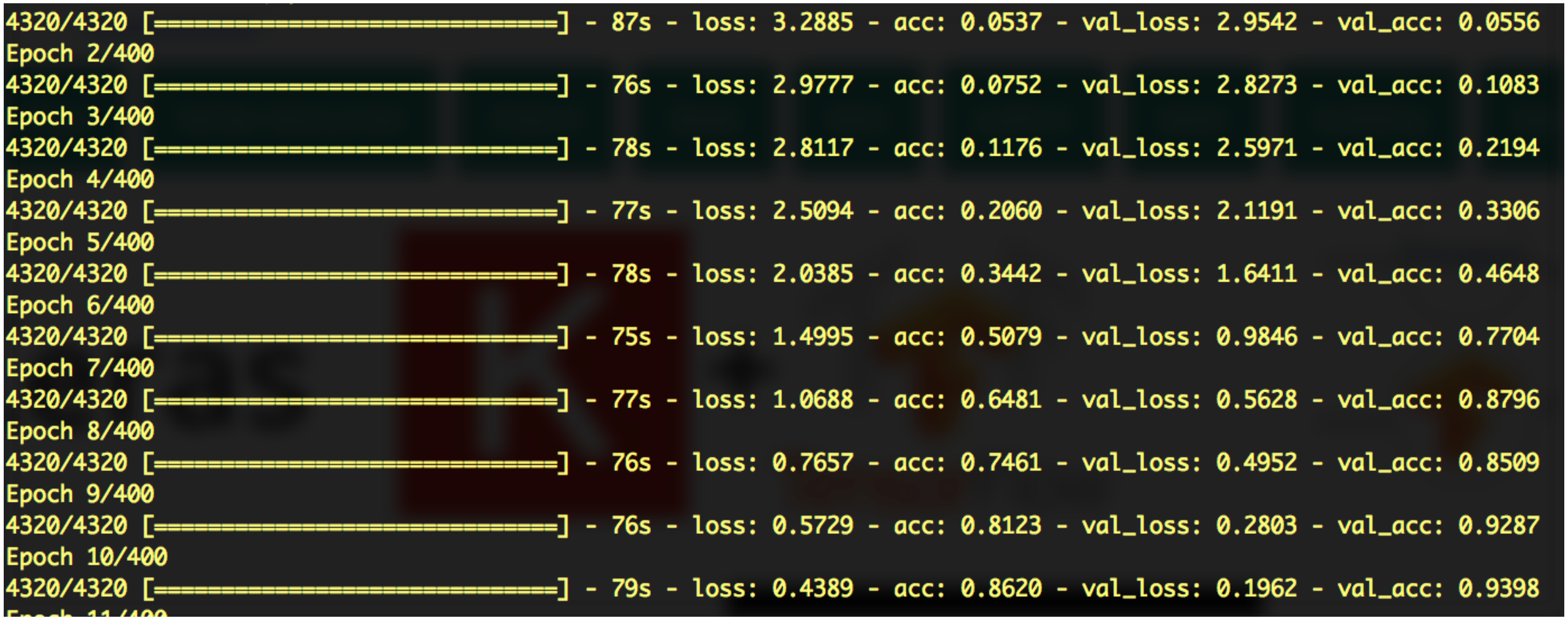
```python
from keras.layers import Input, Dense
from keras.models import Model

input_layer = Input(shape=(1000,))
fc_1 = Dense(512, activation='relu')(input_layer)
fc_2 = Dense(256, activation='relu')(fc_1)
output_layer = Dense(10, activation='softmax')(fc_2)

model = Model(input=input_layer, output=output_layer)
model.compile(optimizer='rmsprop',
              loss='categorical_crossentropy',
              metrics=['accuracy'])

model.fit(bow, newsgroups.target)
predictions = model.predict(features).argmax(axis=1)
```

DNN hyperparameter examples

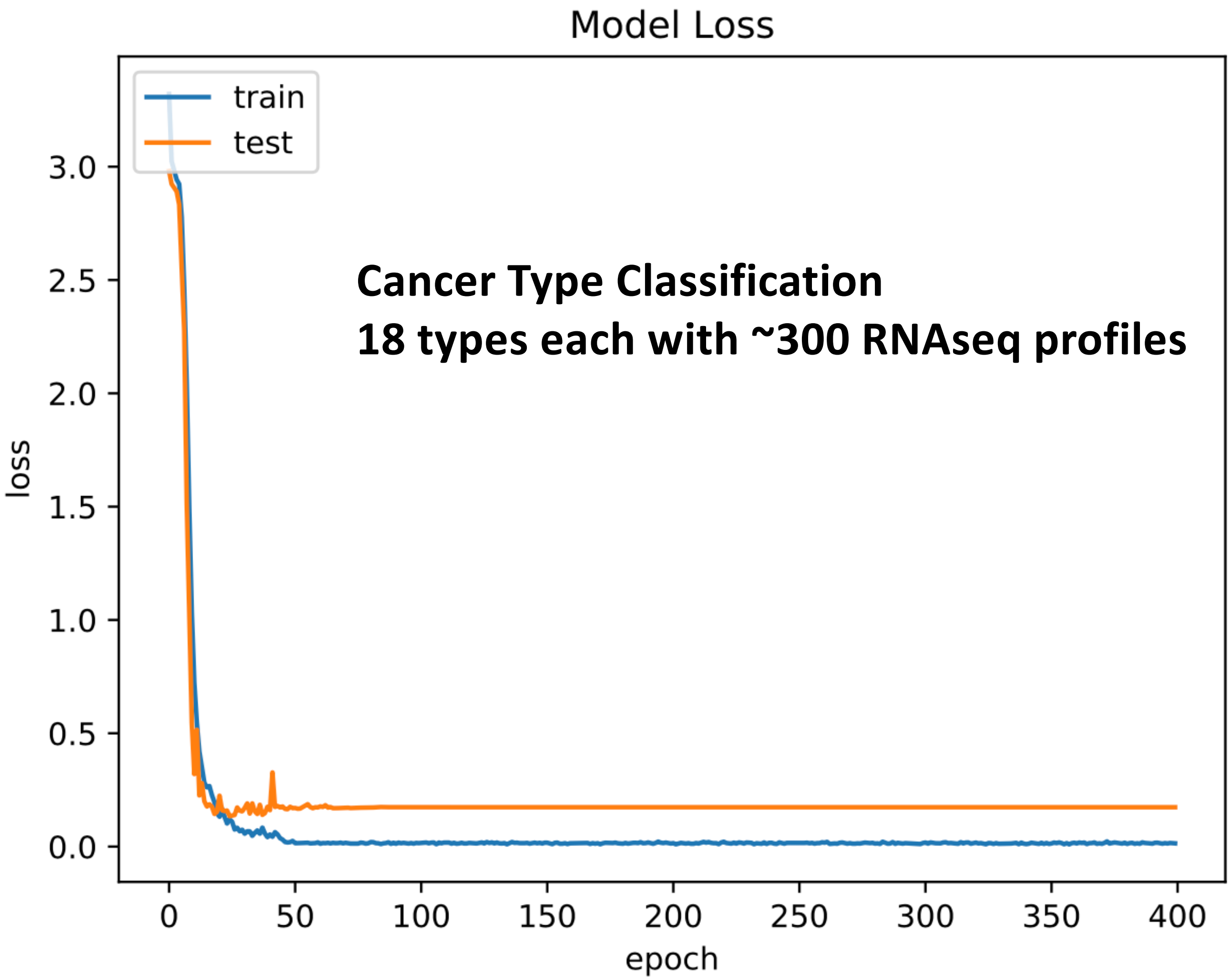
- Data preprocess
  - Positive features: logarithmic transformation y = log(1+x)
  - Mixed features: standard scaler
- Number of hidden layers: 4
- Number of neurons in hidden layers: 4000-2000-1000-1000
- Activation function: ReLU
- Dropouts: Input: 0%; layers 1,2,3: 25%; layer 4: 10%
- Initialization: no unsupervised pretraining
- Optimization: learning rate = 0.05, momentum = 0.9, and weight decay = 0.0001
- Training epochs: as large as possible (dropout can prevent overfitting)

Code Examples

Cancer Type Classification

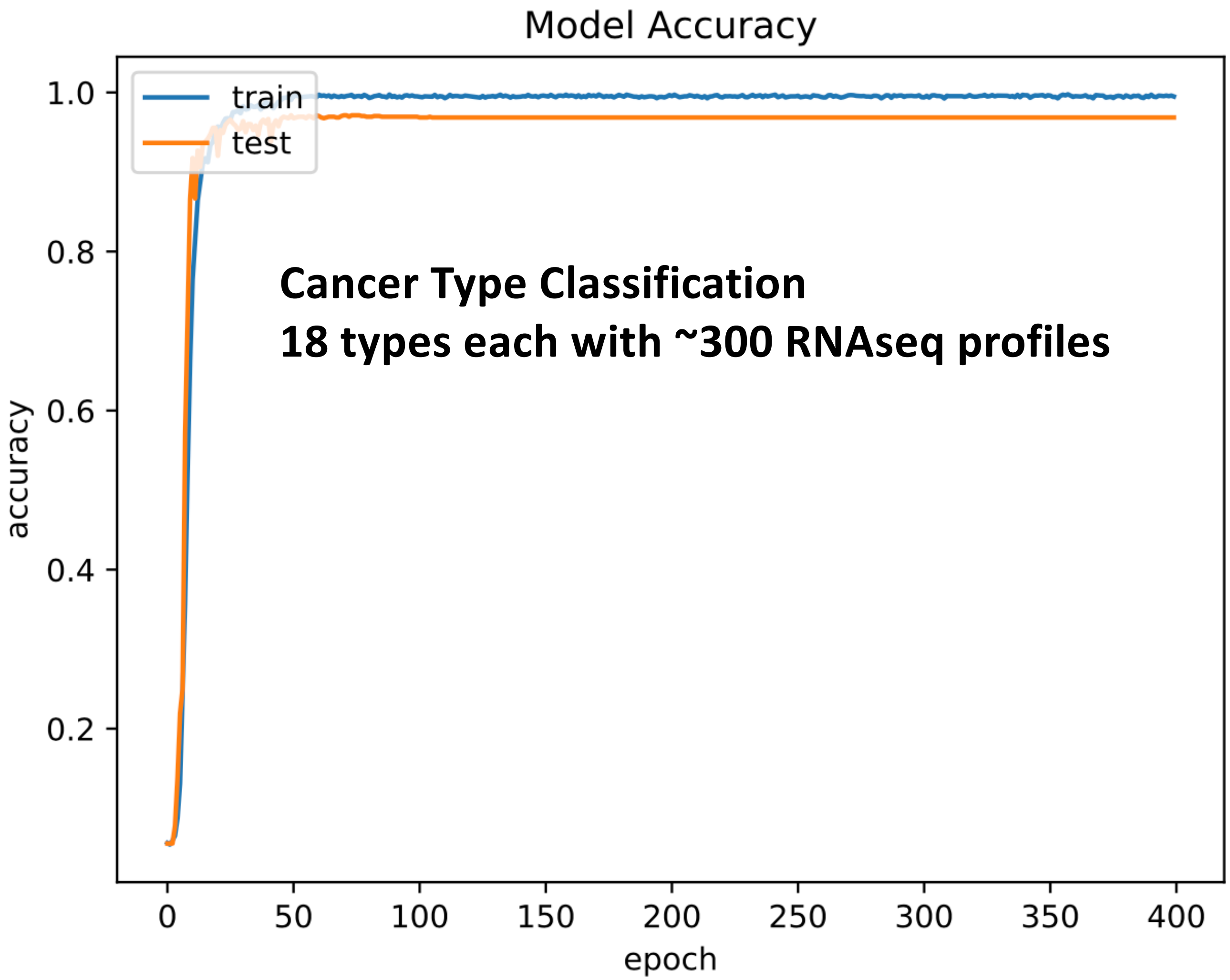


```
4320/4320 [==============================] - 87s - loss: 3.2885 - acc: 0.0537 - val_loss: 2.9542 - val_acc: 0.0556
Epoch 2/400
4320/4320 [==============================] - 76s - loss: 2.9777 - acc: 0.0752 - val_loss: 2.8273 - val_acc: 0.1083
Epoch 3/400
4320/4320 [==============================] - 78s - loss: 2.8117 - acc: 0.1176 - val_loss: 2.5971 - val_acc: 0.2194
Epoch 4/400
4320/4320 [==============================] - 77s - loss: 2.5094 - acc: 0.2060 - val_loss: 2.1191 - val_acc: 0.3306
Epoch 5/400
4320/4320 [==============================] - 78s - loss: 2.0385 - acc: 0.3442 - val_loss: 1.6411 - val_acc: 0.4648
Epoch 6/400
4320/4320 [==============================] - 75s - loss: 1.4995 - acc: 0.5079 - val_loss: 0.9846 - val_acc: 0.7704
Epoch 7/400
4320/4320 [==============================] - 77s - loss: 1.0688 - acc: 0.6481 - val_loss: 0.5628 - val_acc: 0.8796
Epoch 8/400
4320/4320 [==============================] - 76s - loss: 0.7657 - acc: 0.7461 - val_loss: 0.4952 - val_acc: 0.8509
Epoch 9/400
4320/4320 [==============================] - 76s - loss: 0.5729 - acc: 0.8123 - val_loss: 0.2803 - val_acc: 0.9287
Epoch 10/400
4320/4320 [==============================] - 79s - loss: 0.4389 - acc: 0.8620 - val_loss: 0.1962 - val_acc: 0.9398
Epoch 11/400
```



Cancer Type Classification
18 types each with ~300 RNAseq profiles

https://github.com/ECP-CANDLE/Benchmarks/tree/frameworks/Pilot1/TC1

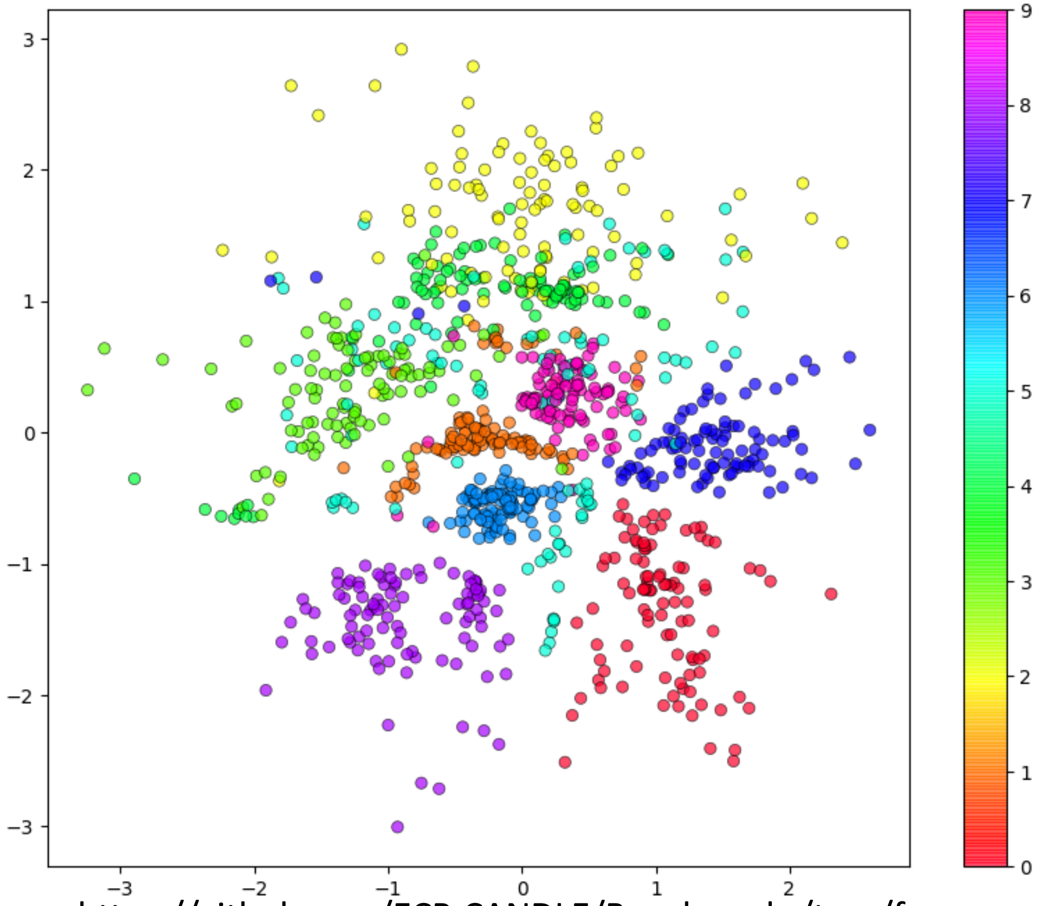


Cancer Type Classification

18 types each with ~300 RNAseq profiles

https://github.com/ECP-CANDLE/Benchmarks/tree/frameworks/Pilot1/TC1

VAE Latent Representation of GDC Expression



Uterine Corpus Endometrial Carcinoma

Thyroid Carcinoma

Skin Cutaneous Melanoma

Prostate Adenocarcinoma

Other Cancer Types

Lung Squamous Cell Carcinoma

Lung Adenocarcinoma

Head and Neck Squamous Cell Carcinoma

Breast Invasive Carcinoma

Brain Lower Grade Glioma

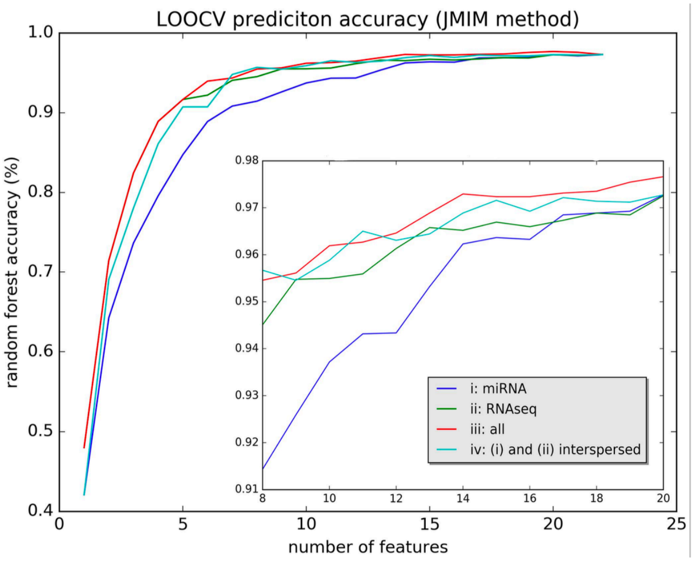
https://github.com/ECP-CANDLE/Benchmarks/tree/frameworks/Pilot1/P1B1

How did we know it might work?

- Build autoencoders first with the features you are going to work with
- If you get reasonable accuracy then the model can learn a representation and that is a good sign
- Class balance seems to matter
- Number of training examples matters > 1000 is good > 10,000 better, > 100,000 much better
- Hyper parameter search is also important once you get something that basically works

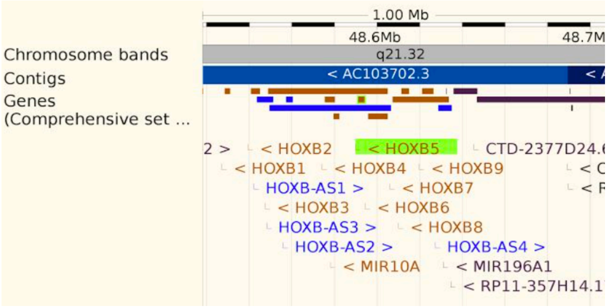
Generate Compact Molecular Signatures

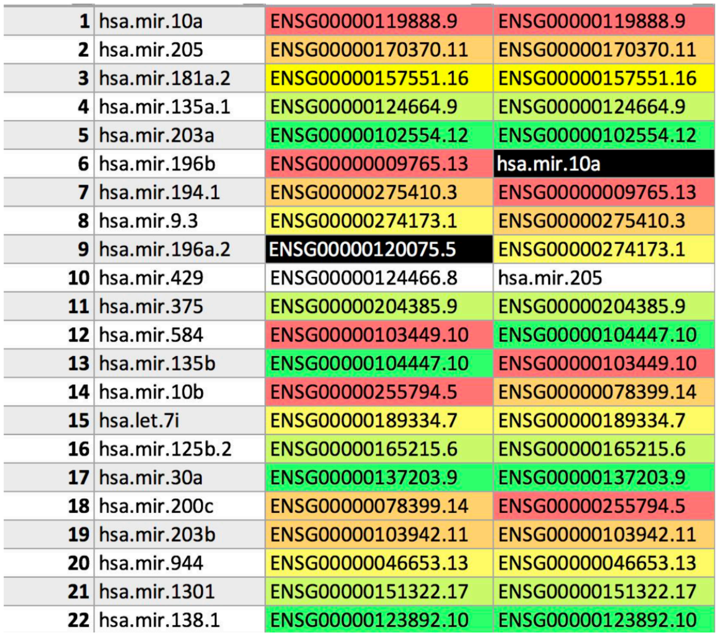
- **For each agent or class of agents we will apply feature selection methods to the models to generate where possible a compact molecular signature that retains prediction performance**
  - Typical reduced signatures include O(10)-O(100) features from >> 50,000 starting features
  - Features may be genes, SNPs, µRNA etc.
- **Developed and applied multiple feature selection methods**
  - Selection criterion: Chi2, Anova, mutual info, ensemble ML, deep neural networks
  - Algorithms: ranking, intersection, iterati maximization
  - Supervised recursive binning
- **Extracted compact features**
  - Features from cancer type prediction
  - 50 features: 0.981 accuracy
  - 20 features: 0.976 accuracy
  - 14 features: 0.973 accuracy
  - RNAseq is more informative than miRN/



Analyze Molecular Signatures to Provide Insight to Potential Mechanisms

- **Started mapping gene features to pathways**
  - Enrichment analysis will be applied to the signatures to identify associated pathways
  - Pathways will be identified that associate with both sensitive and resistant response phenotypes
- **Identified co-located or known interacting pairs of gene and microRNA signatures**
  - Top miRNA feature hsa.mir.10a is co-located with ENSG00000120075.5
  - It has also been experimentally verified that this miRNA downregulates the corresponding HOX genes

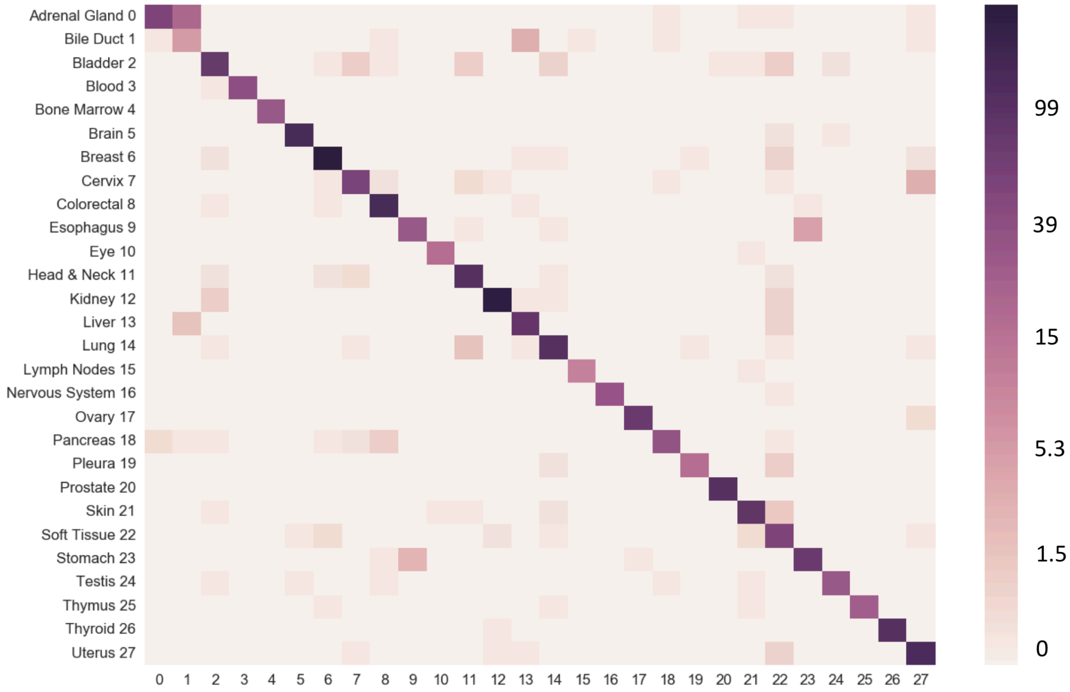




Rank	MIR	RNA	MIR&RNA
1	hsa.mir.10a	ENSG00000119888.9	ENSG00000119888.9
2	hsa.mir.205	ENSG00000170370.11	ENSG00000170370.11
3	hsa.mir.181a.2	ENSG00000157551.16	ENSG00000157551.16
4	hsa.mir.135a.1	ENSG00000124664.9	ENSG00000124664.9
5	hsa.mir.203a	ENSG00000102554.12	ENSG00000102554.12
6	hsa.mir.196b	ENSG00000009765.13	hsa.mir.10a
7	hsa.mir.194.1	ENSG00000275410.3	ENSG00000009765.13
8	hsa.mir.9.3	ENSG00000274173.1	ENSG00000275410.3
9	hsa.mir.196a.2	ENSG00000120075.5	ENSG00000274173.1
10	hsa.mir.429	ENSG00000124466.8	hsa.mir.205
11	hsa.mir.375	ENSG00000204385.9	ENSG00000204385.9
12	hsa.mir.584	ENSG00000103449.10	ENSG00000104447.10
13	hsa.mir.135b	ENSG00000104447.10	ENSG00000103449.10
14	hsa.mir.10b	ENSG00000255794.5	ENSG00000078399.14
15	hsa.let.7i	ENSG00000189334.7	ENSG00000189334.7
16	hsa.mir.125b.2	ENSG00000165215.6	ENSG00000165215.6
17	hsa.mir.30a	ENSG00000137203.9	ENSG00000137203.9
18	hsa.mir.200c	ENSG00000078399.14	ENSG00000255794.5
19	hsa.mir.203b	ENSG00000103942.11	ENSG00000103942.11
20	hsa.mir.944	ENSG00000046653.13	ENSG00000046653.13
21	hsa.mir.1301	ENSG00000151322.17	ENSG00000151322.17
22	hsa.mir.138.1	ENSG00000123892.10	ENSG00000123892.10

Close examination of prediction error

- Confusion matrix
- Local feature importance
- Force plots

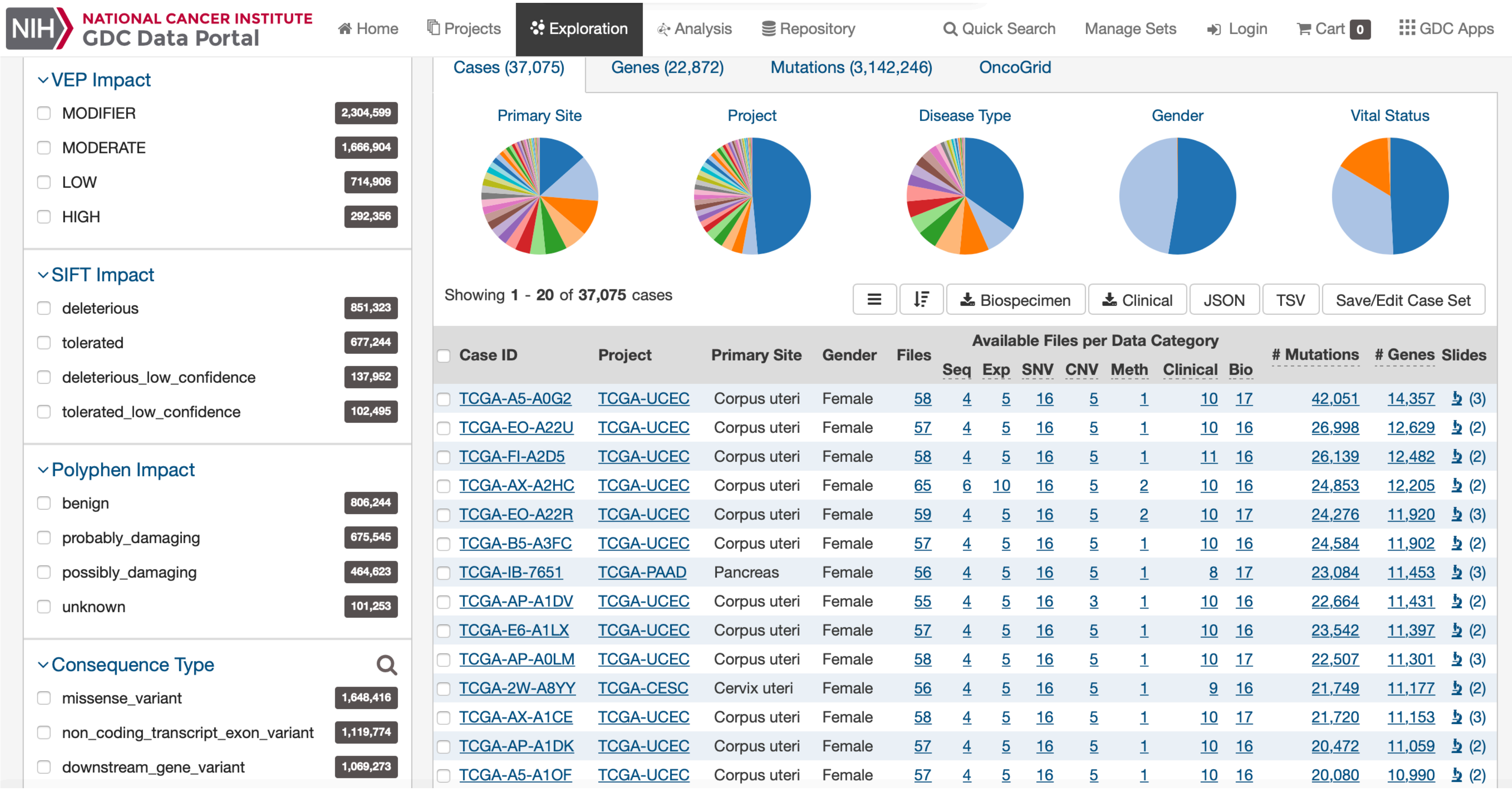


RNAseq Bias Removal

Alex Partin

Cancer Type Classification with SNPs

GDC: 10K samples with 10M mutations



VEP Impact

□ MODIFIER	2,304,599
□ MODERATE	1,666,904
□ LOW	714,906
□ HIGH	292,356

SIFT Impact

□ deleterious	851,323
□ tolerated	677,244
□ deleterious_low_confidence	137,952
□ tolerated_low_confidence	102,495

Polyphen Impact

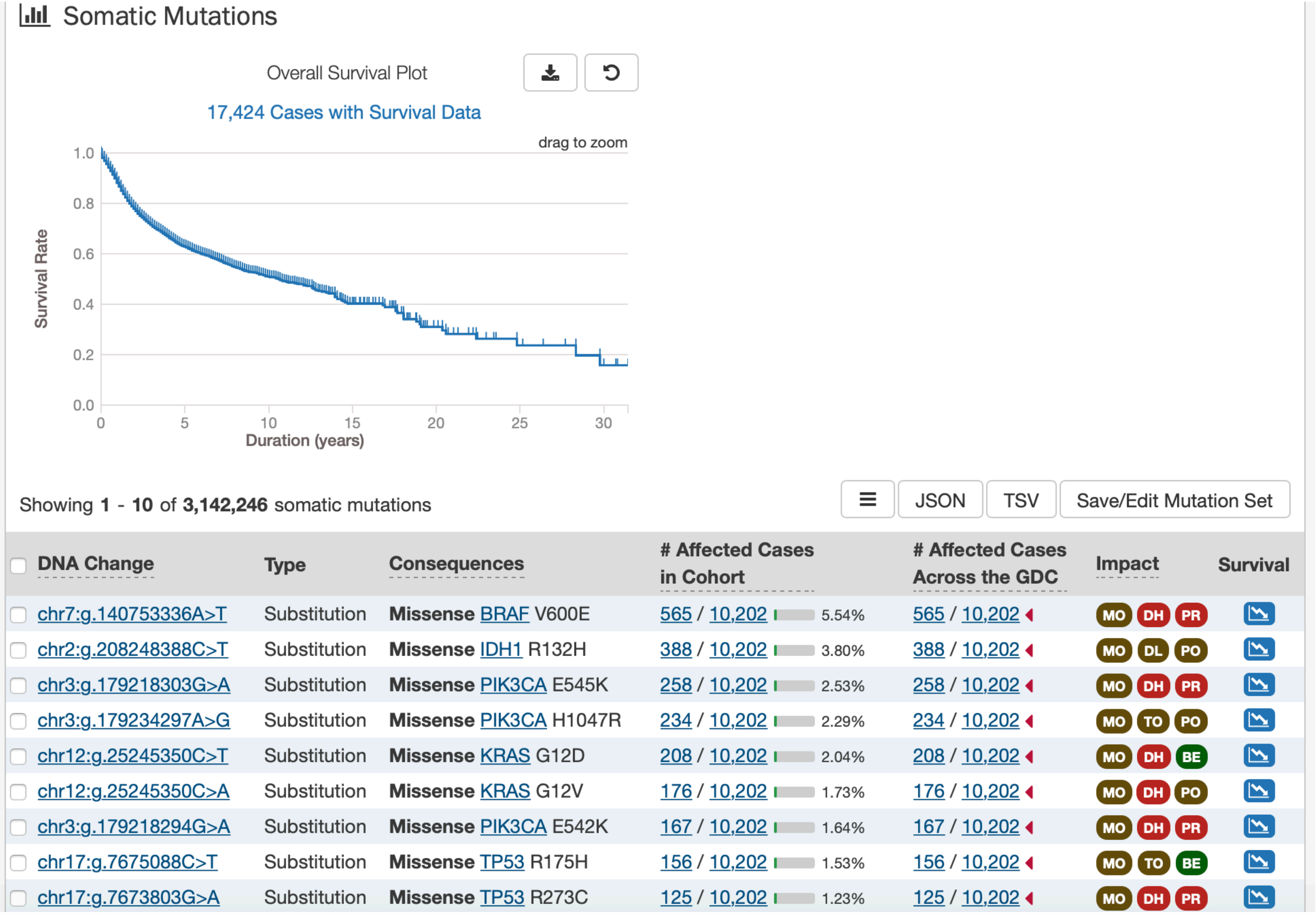
□ benign	806,244
□ probably_damaging	675,545
□ possibly_damaging	464,623
□ unknown	101,253

Consequence Type

□ missense_variant	1,648,416
□ non_coding_transcript_exon_variant	1,119,774
□ downstream_gene_variant	1,069,273

Showing **1** - **20** of **37,075** cases

□	Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mutations	# Genes	Slides
						Seq	Exp	SNV	CNV	Meth	Clinical	Bio			
□	TCGA-A5-A0G2	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	10	17	42,051	14,357	🔬 (3)
□	TCGA-EO-A22U	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	26,998	12,629	🔬 (2)
□	TCGA-FI-A2D5	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	11	16	26,139	12,482	🔬 (2)
□	TCGA-AX-A2HC	TCGA-UCEC	Corpus uteri	Female	65	6	10	16	5	2	10	16	24,853	12,205	🔬 (2)
□	TCGA-EO-A22R	TCGA-UCEC	Corpus uteri	Female	59	4	5	16	5	2	10	17	24,276	11,920	🔬 (3)
□	TCGA-B5-A3FC	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	24,584	11,902	🔬 (2)
□	TCGA-IB-7651	TCGA-PAAD	Pancreas	Female	56	4	5	16	5	1	8	17	23,084	11,453	🔬 (3)
□	TCGA-AP-A1DV	TCGA-UCEC	Corpus uteri	Female	55	4	5	16	3	1	10	16	22,664	11,431	🔬 (2)
□	TCGA-E6-A1LX	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	23,542	11,397	🔬 (2)
□	TCGA-AP-A0LM	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	10	17	22,507	11,301	🔬 (3)
□	TCGA-2W-A8YY	TCGA-CESC	Cervix uteri	Female	56	4	5	16	5	1	9	16	21,749	11,177	🔬 (2)
□	TCGA-AX-A1CE	TCGA-UCEC	Corpus uteri	Female	58	4	5	16	5	1	10	17	21,720	11,153	🔬 (3)
□	TCGA-AP-A1DK	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	20,472	11,059	🔬 (2)
□	TCGA-A5-A1OF	TCGA-UCEC	Corpus uteri	Female	57	4	5	16	5	1	10	16	20,080	10,990	🔬 (2)



Showing **1** - **10** of **3,142,246** somatic mutations

□	DNA Change	Type	Consequences	# Affected Cases in Cohort	# Affected Cases Across the GDC	Impact	Survival
□	chr7:g.140753336A>T	Substitution	**Missense** BRAF V600E	565 / 10,202 5.54%	565 / 10,202	MO DH PR	📉
□	chr2:g.208248388C>T	Substitution	**Missense** IDH1 R132H	388 / 10,202 3.80%	388 / 10,202	MO DL PO	📉
□	chr3:g.179218303G>A	Substitution	**Missense** PIK3CA E545K	258 / 10,202 2.53%	258 / 10,202	MO DH PR	📉
□	chr3:g.179234297A>G	Substitution	**Missense** PIK3CA H1047R	234 / 10,202 2.29%	234 / 10,202	MO TO PO	📉
□	chr12:g.25245350C>T	Substitution	**Missense** KRAS G12D	208 / 10,202 2.04%	208 / 10,202	MO DH BE	📉
□	chr12:g.25245350C>A	Substitution	**Missense** KRAS G12V	176 / 10,202 1.73%	176 / 10,202	MO DH PO	📉
□	chr3:g.179218294G>A	Substitution	**Missense** PIK3CA E542K	167 / 10,202 1.64%	167 / 10,202	MO DH PR	📉
□	chr17:g.7675088C>T	Substitution	**Missense** TP53 R175H	156 / 10,202 1.53%	156 / 10,202	MO TO BE	📉
□	chr17:g.7673803G>A	Substitution	**Missense** TP53 R273C	125 / 10,202 1.23%	125 / 10,202	MO DH PR	📉

How to deal with $n >> p$?

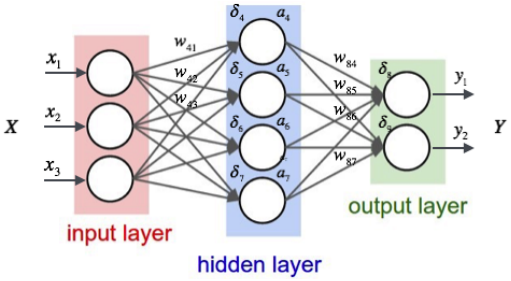
- Dropout
- Regularization
- Locally connected networks

Diet Networks: Thin Parameters for Fat Genomics

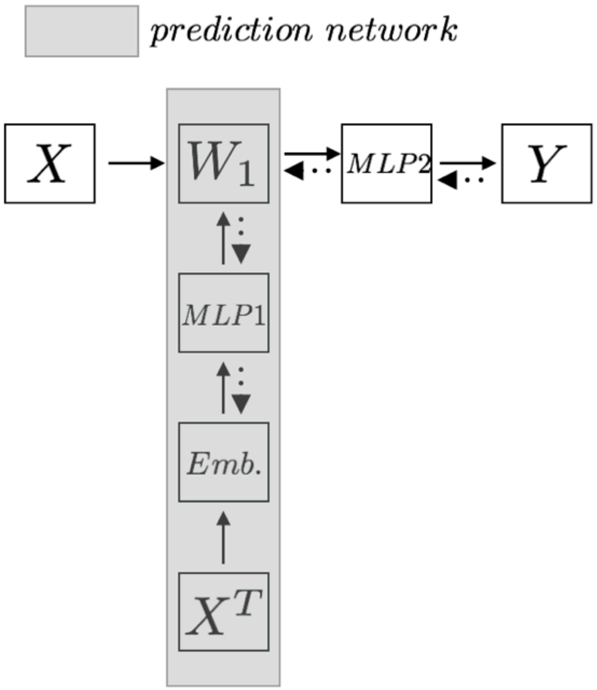
Adriana Romero, Pierre Luc Carrier, Akram Erraqabi, Tristan Sylvain, Alex Auvolat, Etienne Dejoie, Marc-André Legault, Marie-Pierre Dubé, Julie G. Hussin, Yoshua Bengio

Diet Network

- Suppose we have
  - 1000 samples
  - 1,000,000 features
  - 100 neurons in the hidden layer

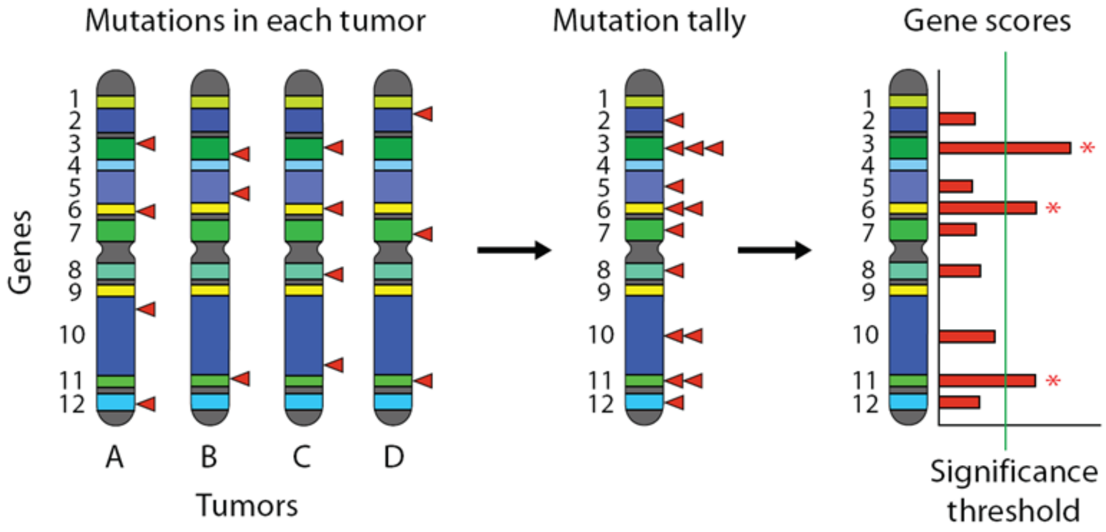


  - Parameters in first layer = 100M



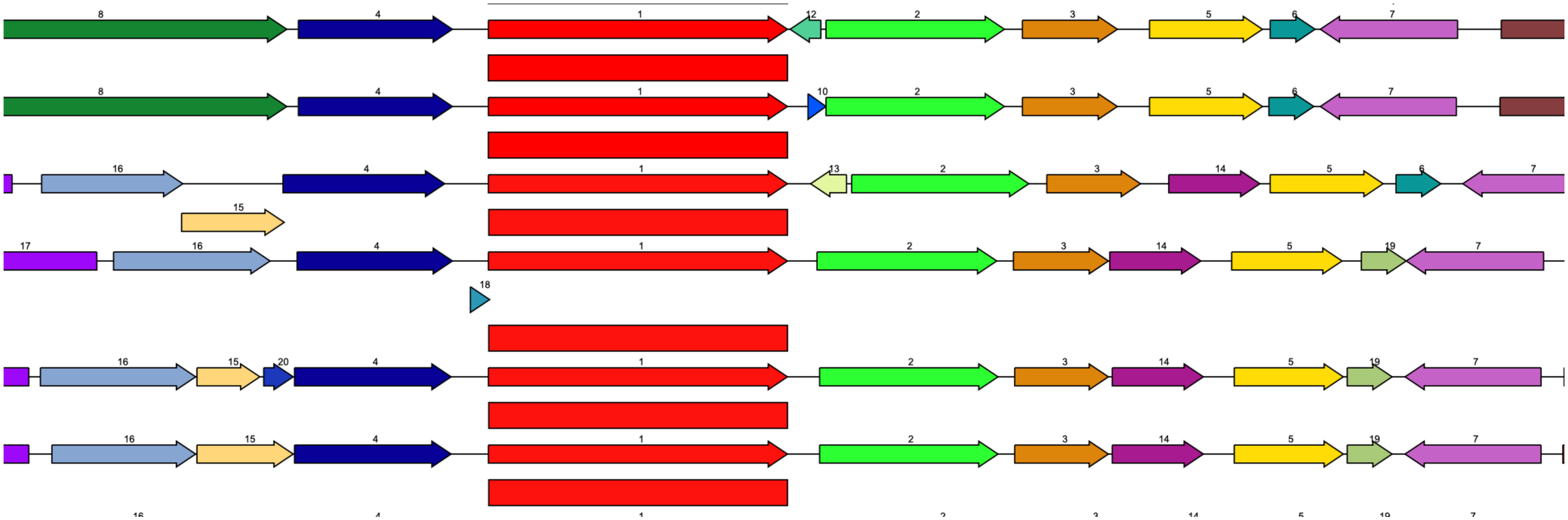
How do we represent the sparse mutations?

- Gene level / pathway level
- Weighting by impact
- Filtering by significance
- Convert to images
  - Variant calling
  - Annotation



MutSigCV

Deep annotation with compare region images



Identification of genomic islands and operons

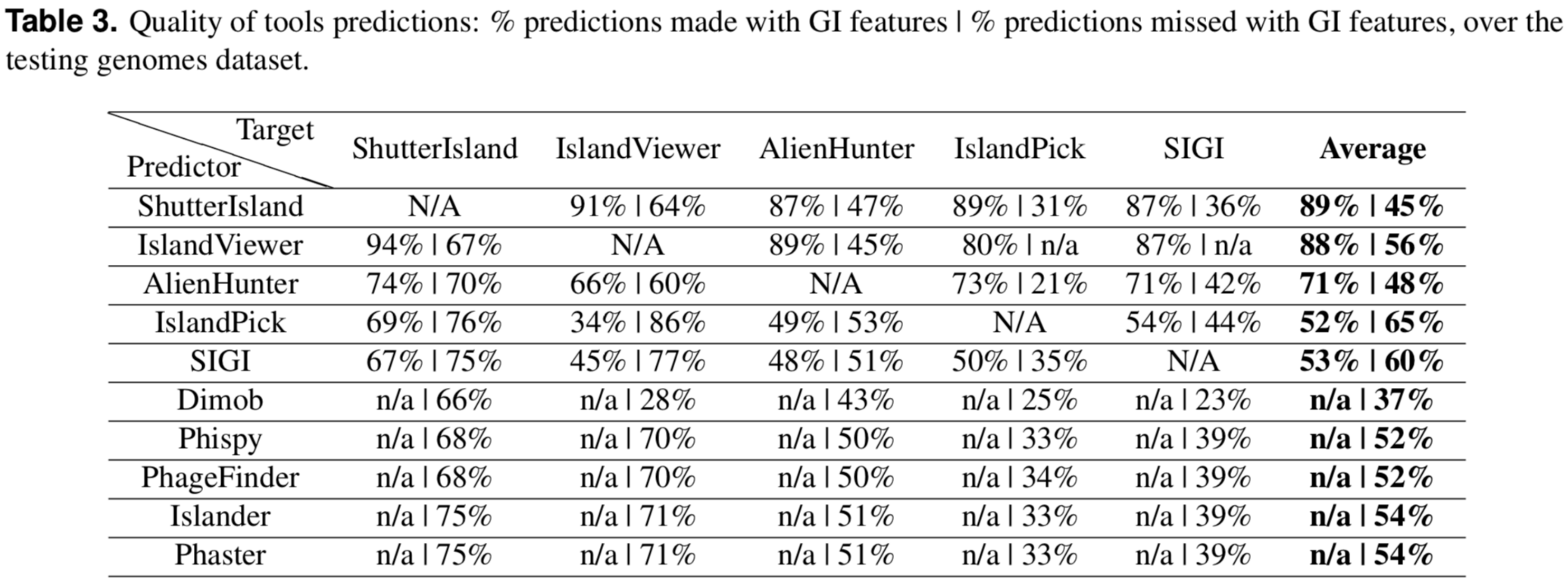
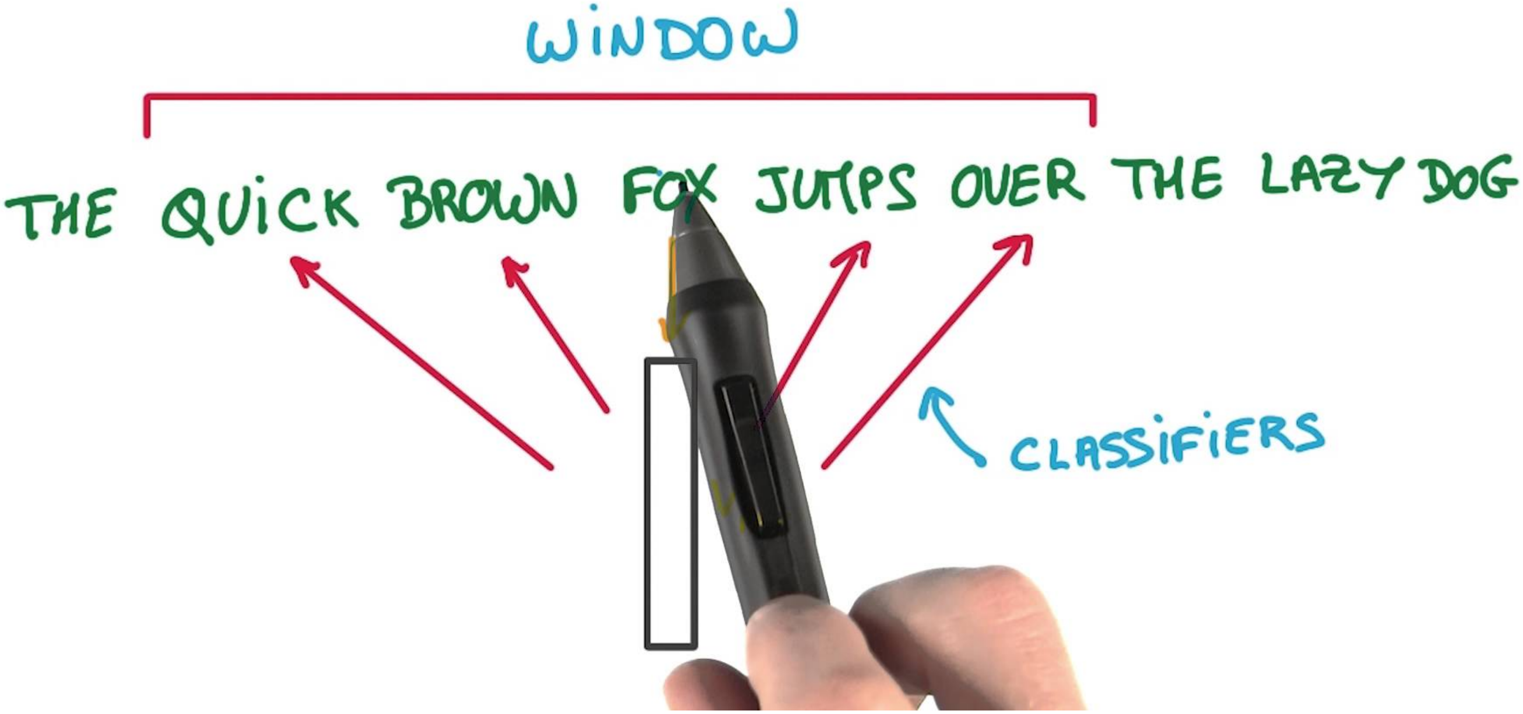


Table 3. Quality of tools predictions: % predictions made with GI features | % predictions missed with GI features, over the testing genomes dataset.

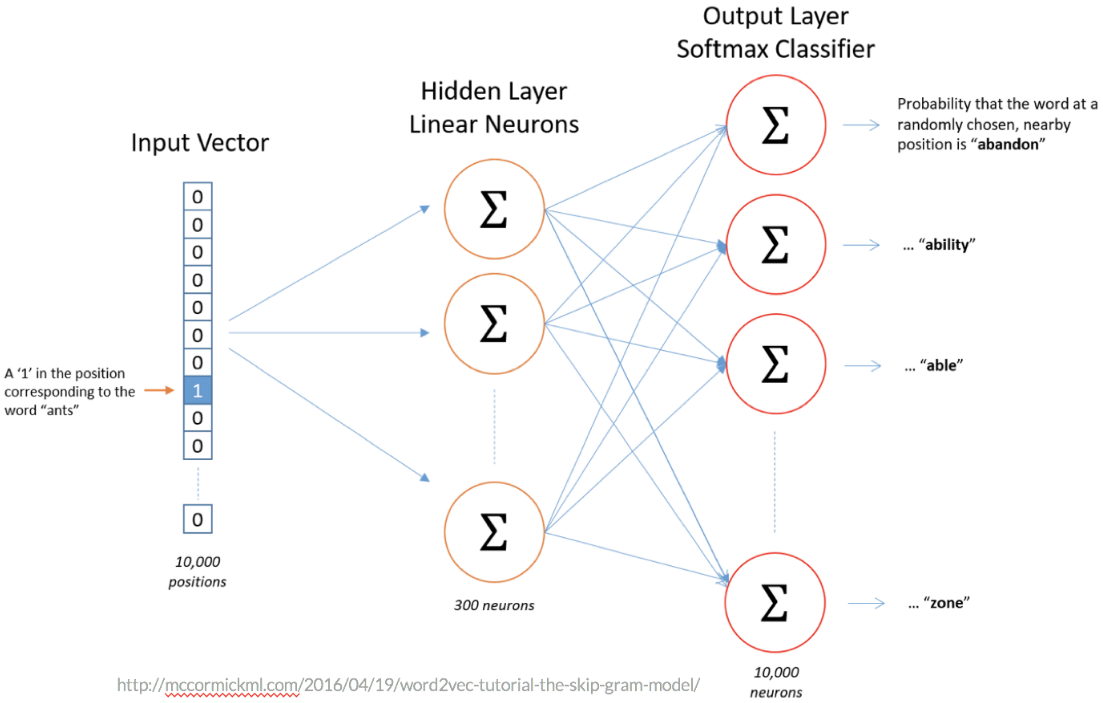
Target Predictor	ShutterIsland	IslandViewer	AlienHunter	IslandPick	SIGI	**Average**
ShutterIsland	N/A	91% | 64%	87% | 47%	89% | 31%	87% | 36%	**89% | 45%**
IslandViewer	94% | 67%	N/A	89% | 45%	80% | n/a	87% | n/a	**88% | 56%**
AlienHunter	74% | 70%	66% | 60%	N/A	73% | 21%	71% | 42%	**71% | 48%**
IslandPick	69% | 76%	34% | 86%	49% | 53%	N/A	54% | 44%	**52% | 65%**
SIGI	67% | 75%	45% | 77%	48% | 51%	50% | 35%	N/A	**53% | 60%**
Dimob	n/a | 66%	n/a | 28%	n/a | 43%	n/a | 25%	n/a | 23%	**n/a | 37%**
Phispy	n/a | 68%	n/a | 70%	n/a | 50%	n/a | 33%	n/a | 39%	**n/a | 52%**
PhageFinder	n/a | 68%	n/a | 70%	n/a | 50%	n/a | 34%	n/a | 39%	**n/a | 52%**
Islander	n/a | 75%	n/a | 71%	n/a | 51%	n/a | 33%	n/a | 39%	**n/a | 54%**
Phaster	n/a | 75%	n/a | 71%	n/a | 51%	n/a | 33%	n/a | 39%	**n/a | 54%**

Assaf et al. 2019

WORD2VEC



WINDOW
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG
CLASSIFIERS



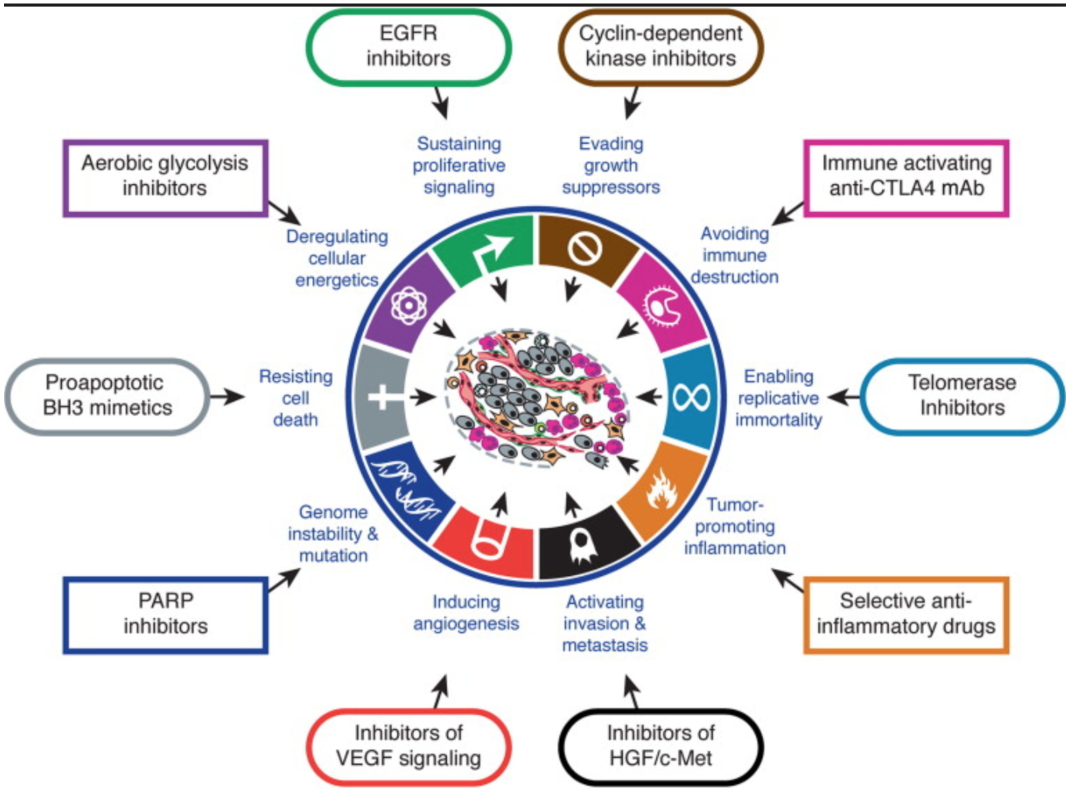
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

"Skip-Gram"

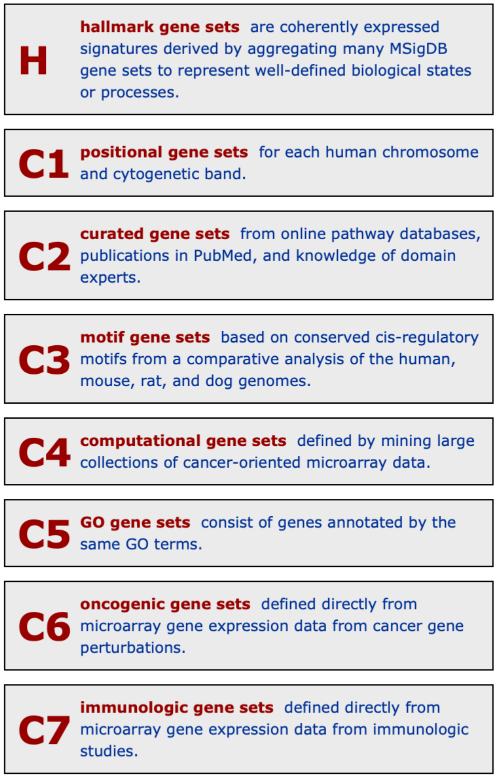
With one-hot encoded centre word, we can predict context words.

Hidden layer creates embeddings

Gene sets from MSigDB


EGFR inhibitors
Cyclin-dependent kinase inhibitors
Aerobic glycolysis inhibitors
Immune activating anti-CTLA4 mAb
Sustaining proliferative signaling
Evading growth suppressors
Deregulating cellular energetics
Avoiding immune destruction
Proapoptotic BH3 mimetics
Resisting cell death
Enabling replicative immortality
Telomerase Inhibitors
Genome instability & mutation
Tumor-promoting inflammation
PARP inhibitors
Inducing angiogenesis
Activating invasion & metastasis
Selective anti-inflammatory drugs
Inhibitors of VEGF signaling
Inhibitors of HGF/c-Met

22,596 gene sets



H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

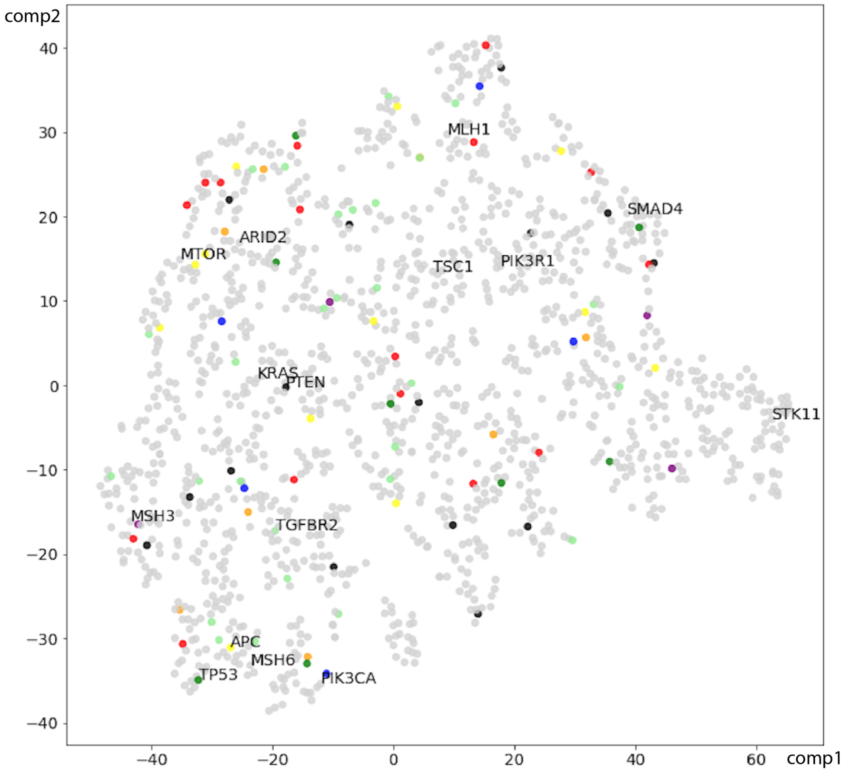
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

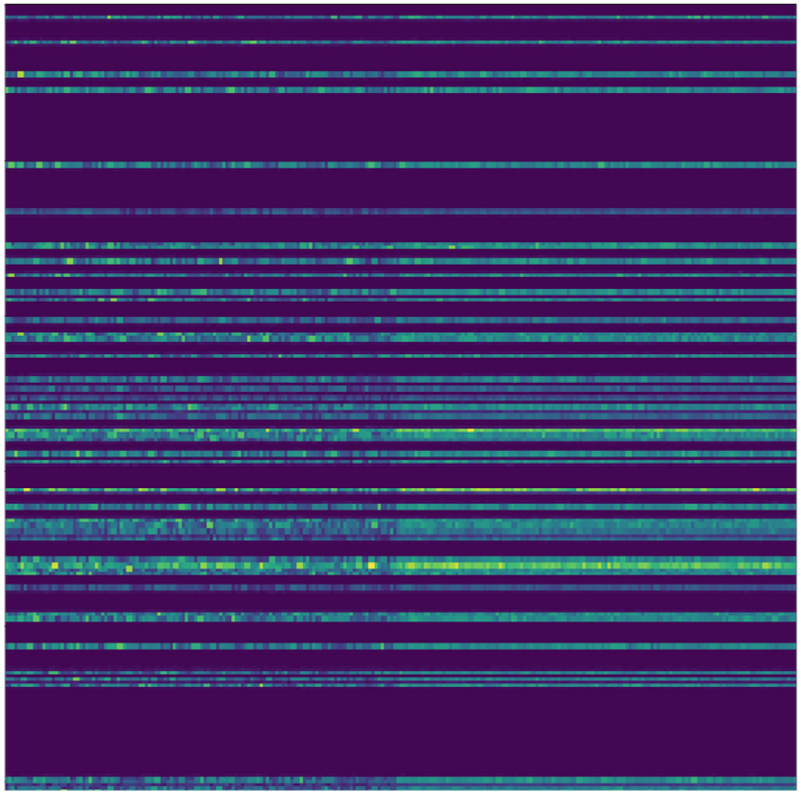
C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

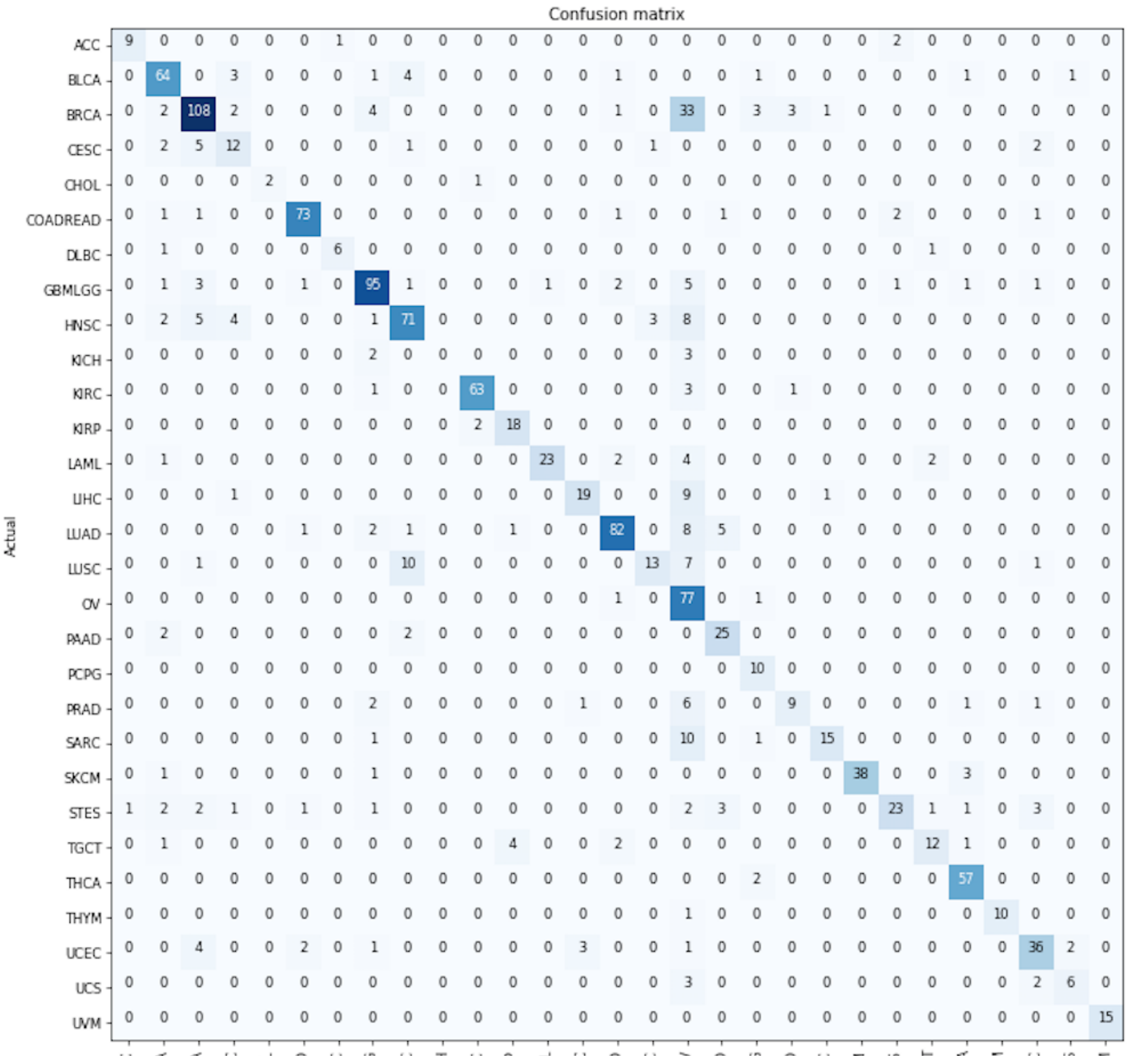
Gene2vec



Sample2image



Alena Harley, The Mystery of the Origin



Confusion matrix

Actual

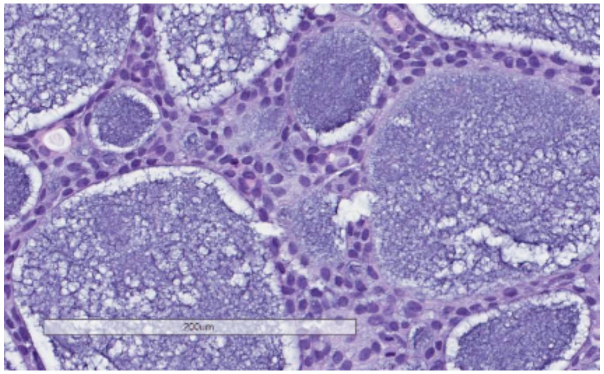
ACC	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
BLCA	0	64	0	3	0	0	0	1	4	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0
BRCA	0	2	108	2	0	0	0	4	0	0	0	0	0	0	1	0	33	0	3	3	1	0	0	0	0	0	0	0	0
CESC	0	2	5	12	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0
CHOL	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
COADREAD	0	1	1	0	0	73	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	2	0	0	0	1	0	0
DLBC	0	1	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
GBMLGG	0	1	3	0	0	1	0	95	1	0	0	0	1	0	2	0	5	0	0	0	0	0	1	0	1	0	1	0	0
HNSC	0	2	5	4	0	0	0	1	71	0	0	0	0	0	0	3	8	0	0	0	0	0	0	0	0	0	0	0	0
KICH	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
KIRC	0	0	0	0	0	0	0	1	0	0	63	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0
KIRP	0	0	0	0	0	0	0	0	0	0	2	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LAML	0	1	0	0	0	0	0	0	0	0	0	0	23	0	2	0	4	0	0	0	0	0	0	2	0	0	0	0	0
LIHC	0	0	0	1	0	0	0	0	0	0	0	0	0	19	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0
LUAD	0	0	0	0	0	1	0	2	1	0	0	1	0	0	82	0	8	5	0	0	0	0	0	0	0	0	0	0	0
LUSC	0	0	1	0	0	0	0	0	10	0	0	0	0	0	0	13	7	0	0	0	0	0	0	0	0	0	1	0	0
OV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	77	0	1	0	0	0	0	0	0	0	0	0	0
PAAD	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0
PCPG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
PRAD	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	6	0	0	9	0	0	0	0	1	0	1	0	0
SARC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	10	0	1	0	15	0	0	0	0	0	0	0	0
SKCM	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	38	0	0	3	0	0	0	0
STES	1	2	2	1	0	1	0	1	0	0	0	0	0	0	0	0	2	3	0	0	0	0	23	1	1	0	3	0	0
TGCT	0	1	0	0	0	0	0	0	0	0	0	4	0	0	2	0	0	0	0	0	0	0	0	12	1	0	0	0	0
THCA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	57	0	0	0	0
THYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	10	0	0	0
UCEC	0	0	4	0	0	2	0	1	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	36	2	0
UCS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	2	6	0
UVM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15

78% accuracy

Alena Harley, The Mystery of the Origin

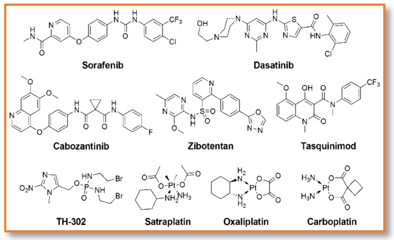
Drug Response Prediction

Modeling Drug Response

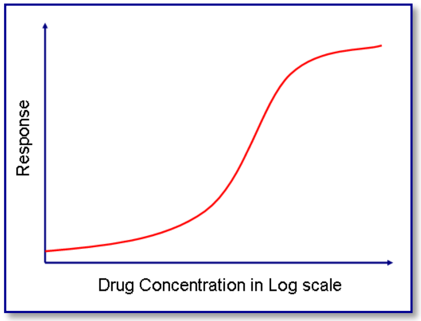


Drug (s)

descriptors
fingerprints
structures
SMILES
dose



$$\mathcal{R} = f(\mathcal{T}, \mathcal{D}_1 , \mathcal{D}_2)$$

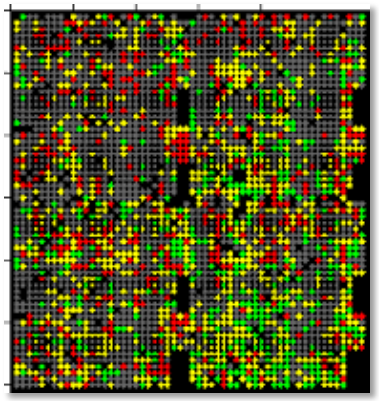


IC50
GI50
% growth
Z-score
AUC

Response

gene expression levels
SNPs
protein abundance
microRNA
methylation

Tumor



Cell Line Features

- NCI-60: 60 cell lines
- Molecular Assays: 20
  - Gene expression array
  - RNA-seq
  - Mutations
  - Protein abundance
  - microRNA

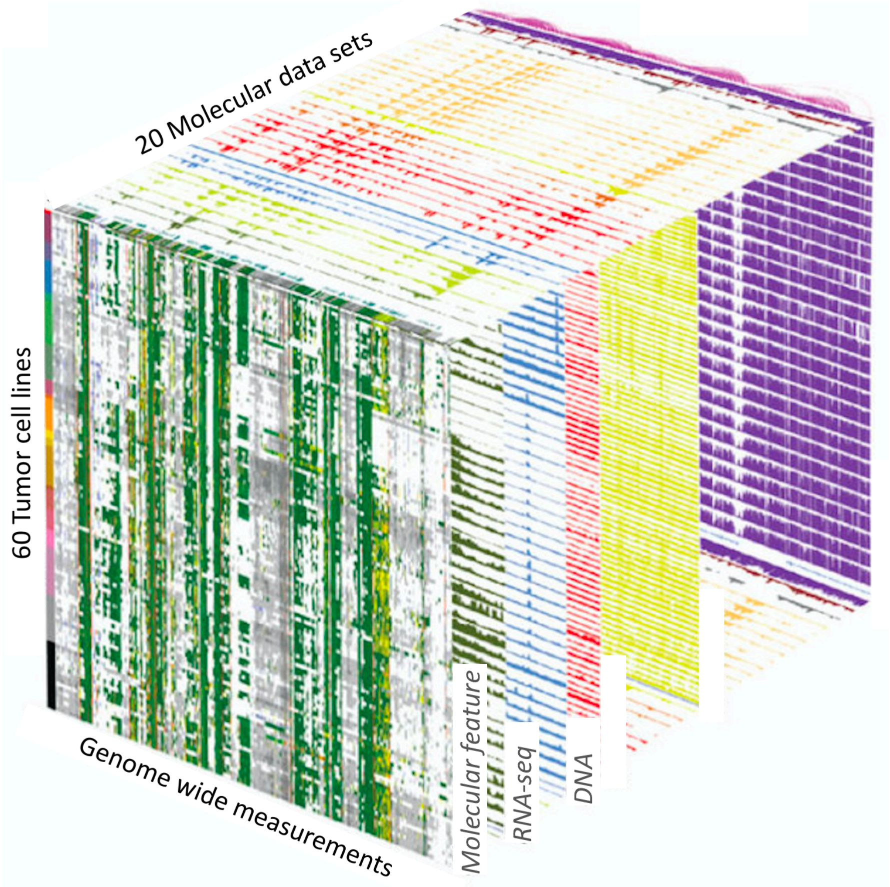
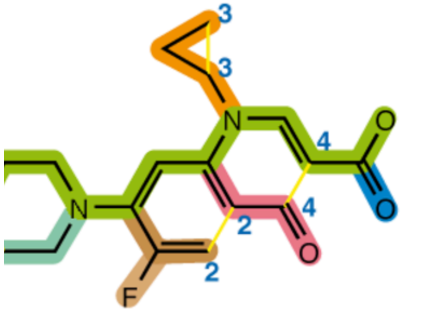


Figure adapted from Kundaje et al. Nature 2015

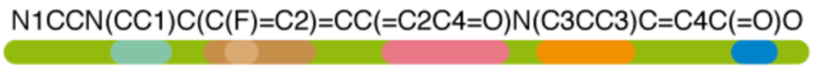
Drug Features

- SMILES strings
- 2D or 3D structures
- Graph convolutions
- Descriptors
- Fingerprints

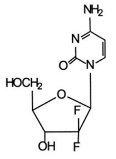
C



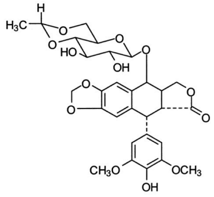
D



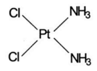
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



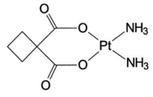
Gemcitabine



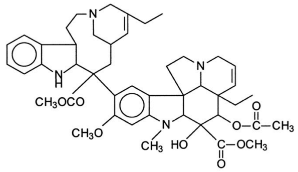
Etoposide (VP-16)



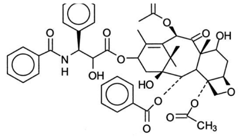
Cisplatin



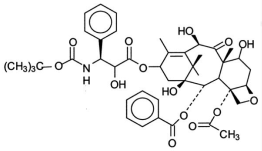
Carboplatin



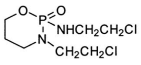
Vinorelbine



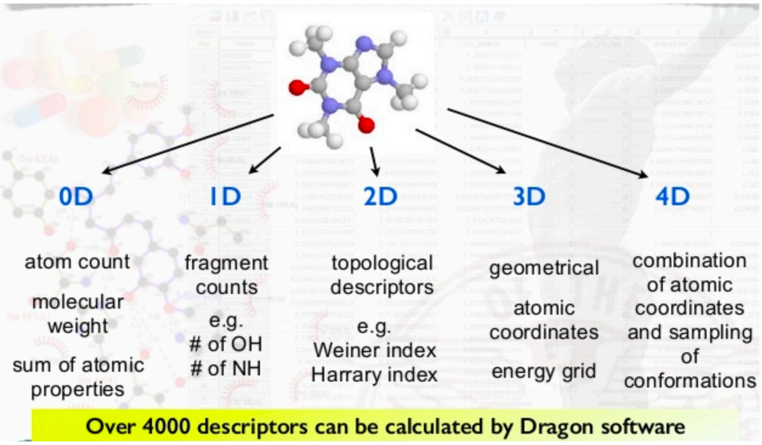
Paclitaxel



Docetaxel



Ifosfamide



0D

atom count

molecular weight

sum of atomic properties

1D

fragment counts

e.g. # of OH # of NH

2D

topological descriptors

e.g. Weiner index Harrary index

3D

geometrical

atomic coordinates

energy grid

4D

combination of atomic coordinates and sampling of conformations

Over 4000 descriptors can be calculated by Dragon software

Deep Learning Model for Drug Pair Response

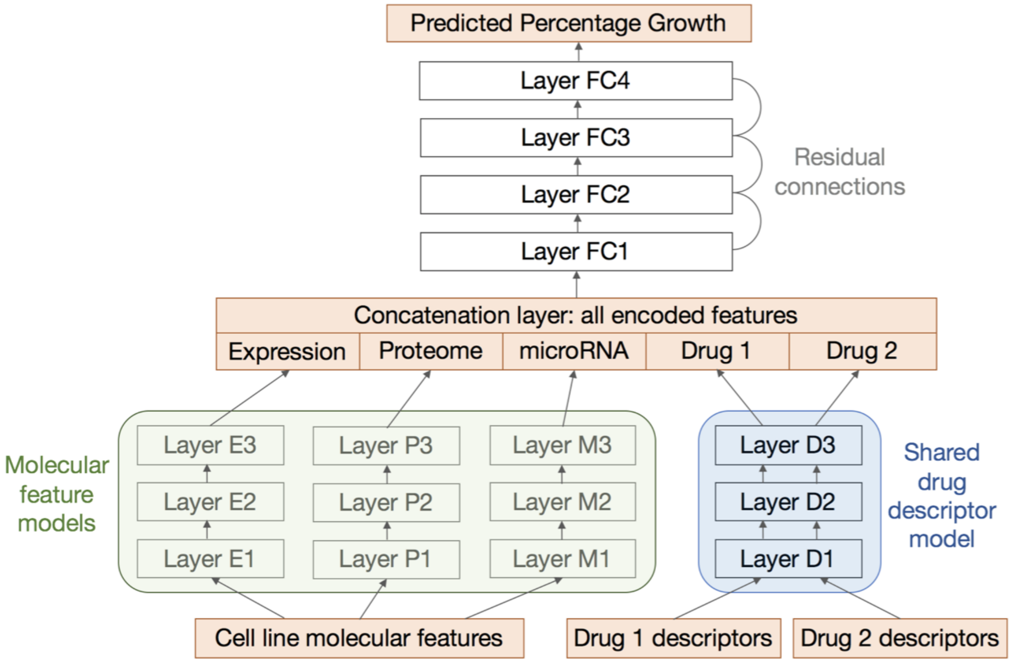
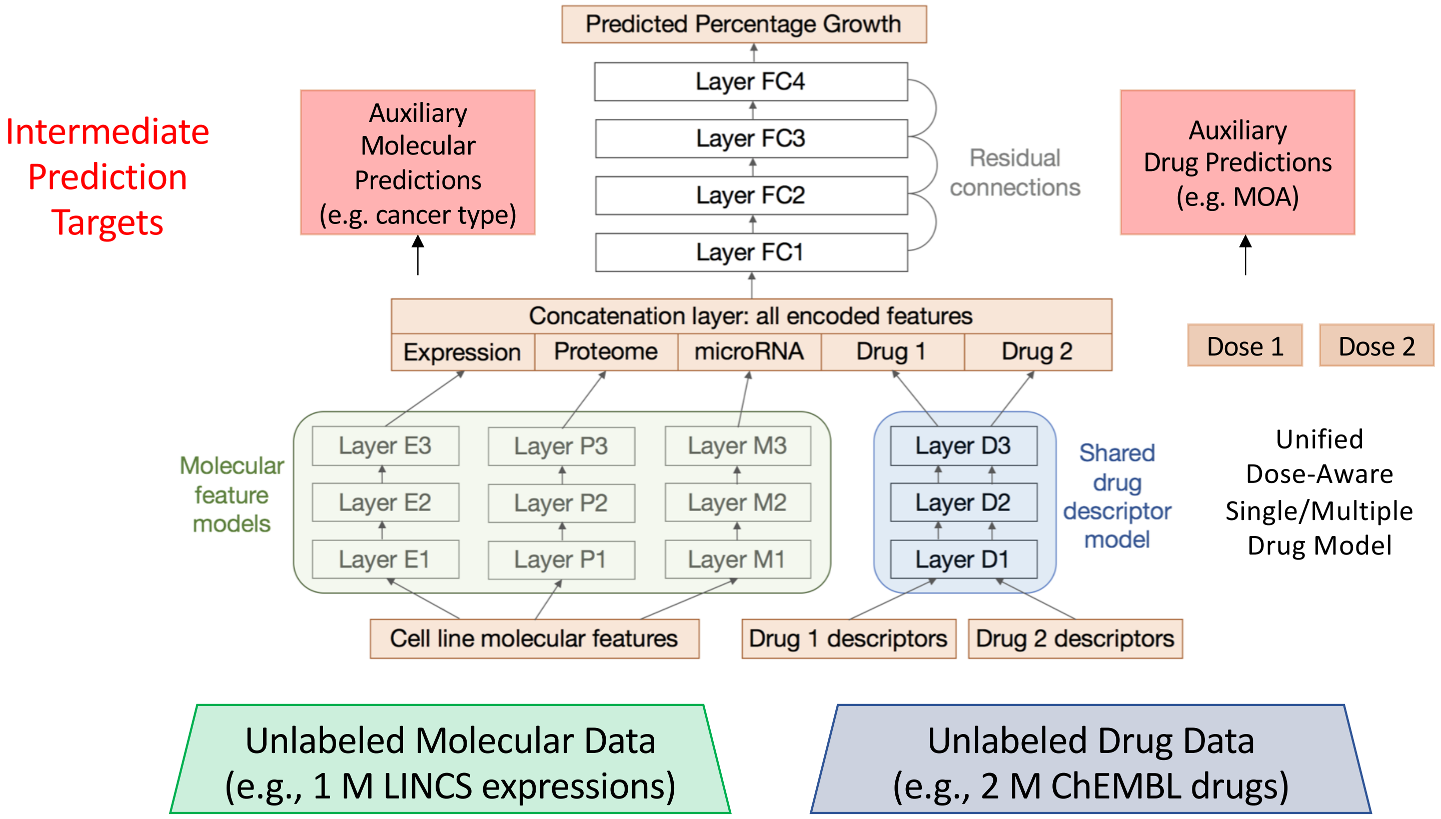


Fig. 2. **Neural network architecture.** The orange square boxes, from bottom to top, represent input features, encoded features, and output growth values. Feature models are denoted by round shaded boxes: green for molecular features and blue for drug features. There are multiple types of molecular features that are fed into submodels for gene expression, proteome, and microRNA. The descriptors for the two drugs share the same descriptor model. All encoded features are then concatenated to form input for the top fully connected layers. Most connecting layers are linked by optional residual skip connections if their dimensions match.

Uno: Predicting Single/Paired Drug Response



Intermediate Prediction Targets

Auxiliary Molecular Predictions (e.g. cancer type)

Predicted Percentage Growth

Layer FC4

Layer FC3

Layer FC2

Layer FC1

Residual connections

Auxiliary Drug Predictions (e.g. MOA)

Concatenation layer: all encoded features

Expression

Proteome

microRNA

Drug 1

Drug 2

Dose 1

Dose 2

Molecular feature models

Layer E3

Layer E2

Layer E1

Layer P3

Layer P2

Layer P1

Layer M3

Layer M2

Layer M1

Layer D3

Layer D2

Layer D1

Shared drug descriptor model

Unified Dose-Aware Single/Multiple Drug Model

Cell line molecular features

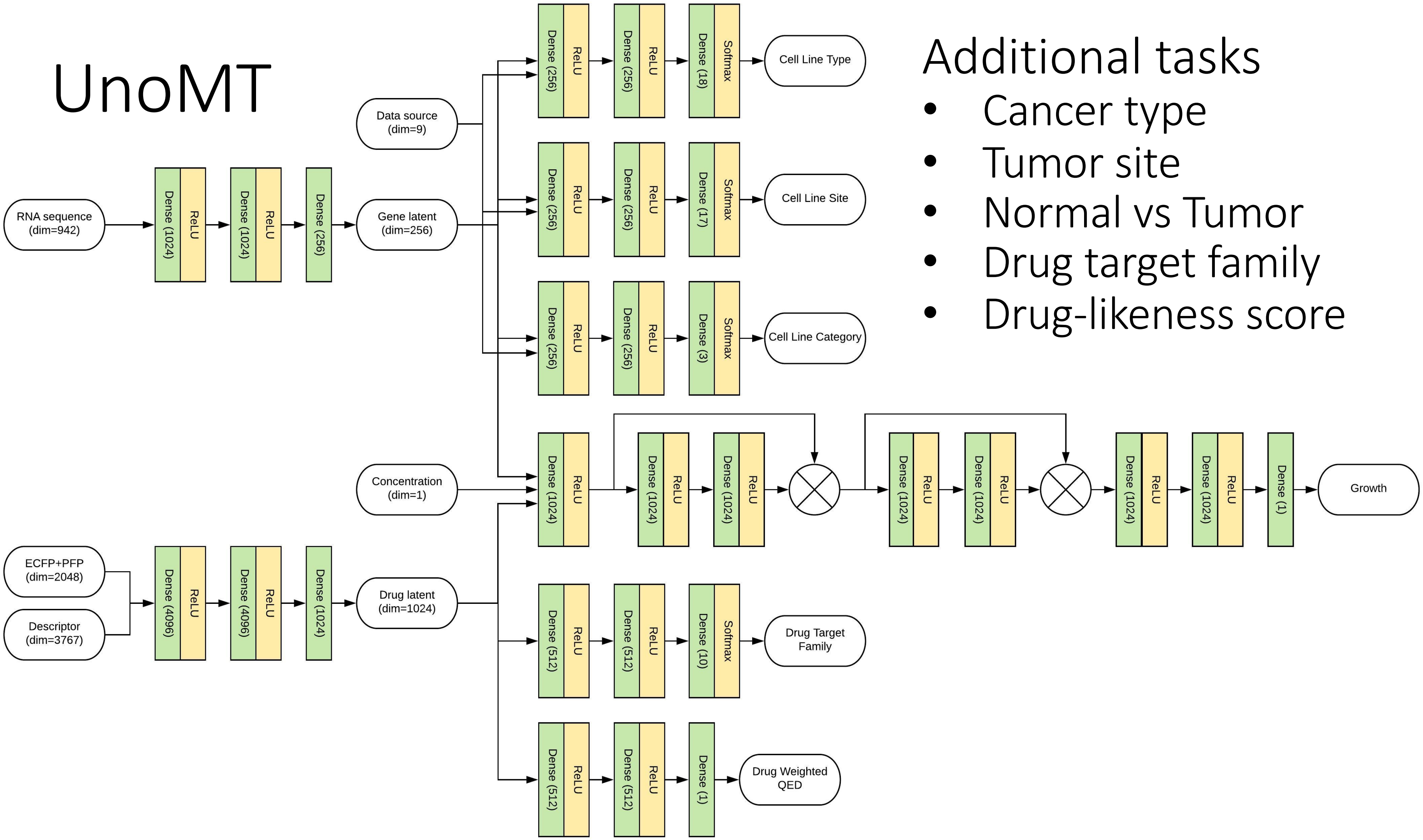
Drug 1 descriptors

Drug 2 descriptors

Unlabeled Molecular Data (e.g., 1 M LINCS expressions)

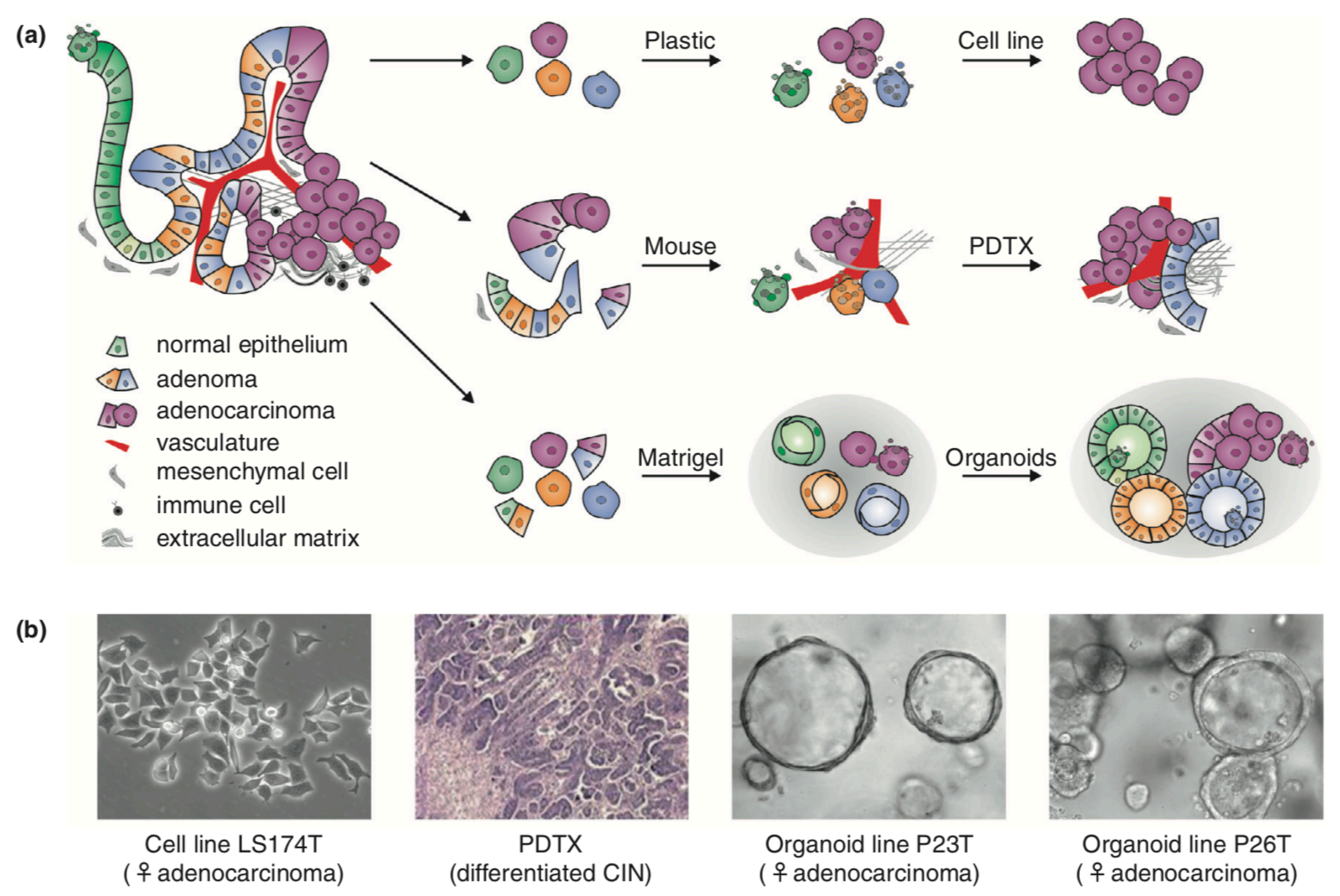
Unlabeled Drug Data (e.g., 2 M ChEMBL drugs)

UnoMT



Additional tasks

- Cancer type
- Tumor site
- Normal vs Tumor
- Drug target family
- Drug-likeness score



Current Opinion in Genetics & Development

Organoid cultures for the analysis of cancer phenotypes, Sachs and Clevers, 2014

Data in place for model training and testing

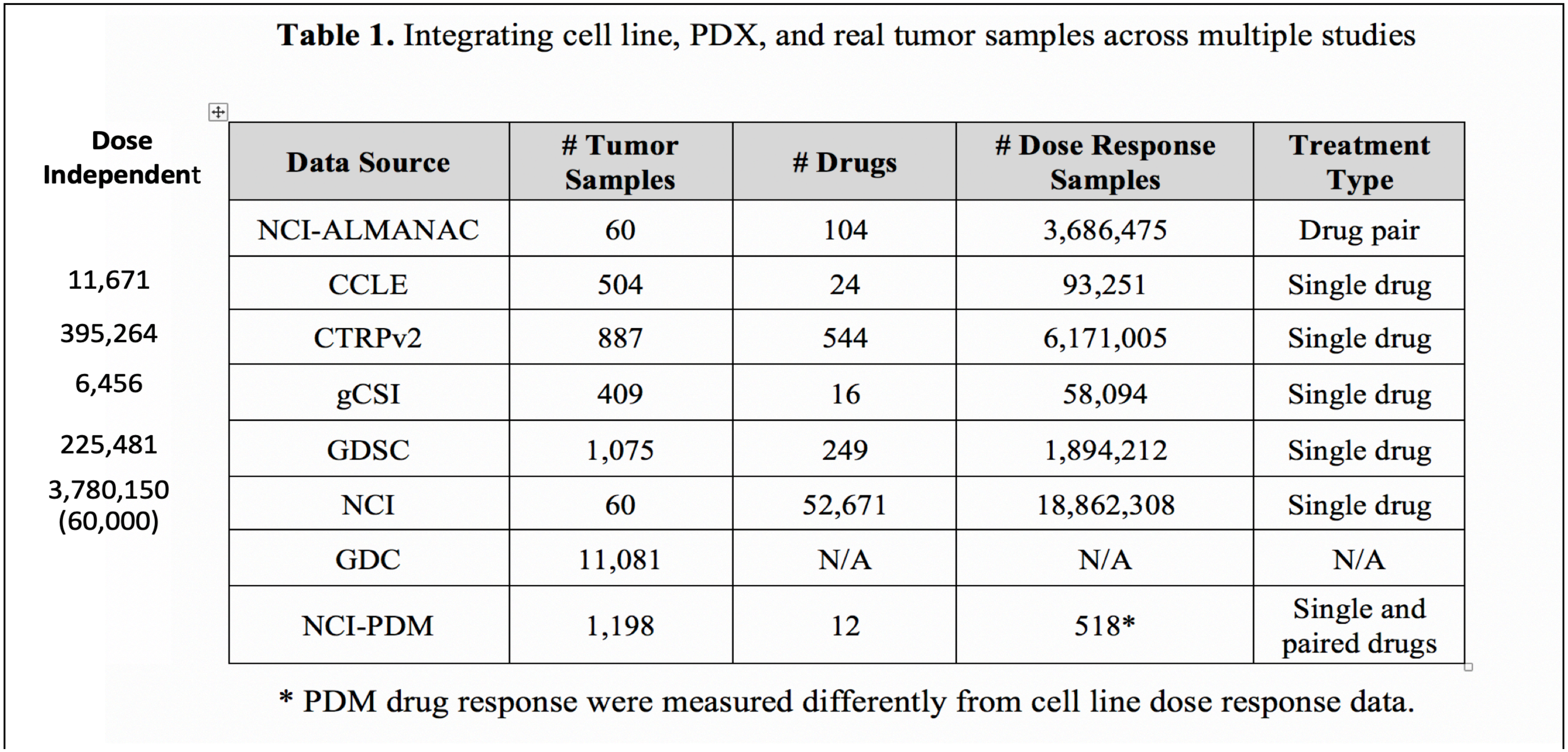


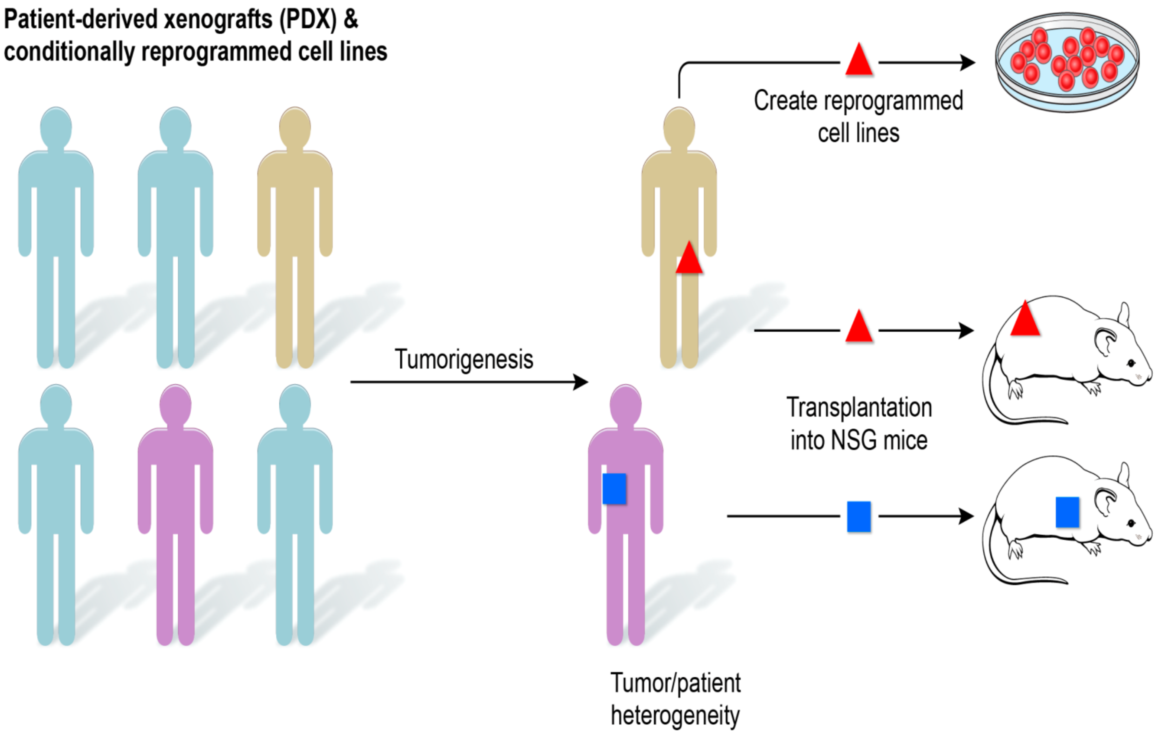
Table 1. Integrating cell line, PDX, and real tumor samples across multiple studies

Dose Independent	Data Source	# Tumor Samples	# Drugs	# Dose Response Samples	Treatment Type
	NCI-ALMANAC	60	104	3,686,475	Drug pair
11,671	CCLE	504	24	93,251	Single drug
395,264	CTRPv2	887	544	6,171,005	Single drug
6,456	gCSI	409	16	58,094	Single drug
225,481	GDSC	1,075	249	1,894,212	Single drug
3,780,150 (60,000)	NCI	60	52,671	18,862,308	Single drug
	GDC	11,081	N/A	N/A	N/A
	NCI-PDM	1,198	12	518*	Single and paired drugs

* PDM drug response were measured differently from cell line dose response data.

Pilot 1

Patient Derived Xenograft Models

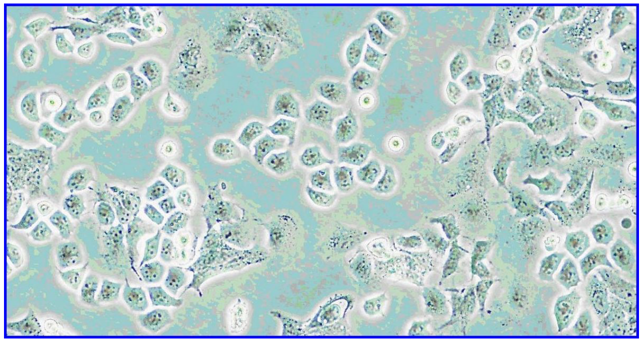


Patient-derived xenografts (PDX) & conditionally reprogrammed cell lines

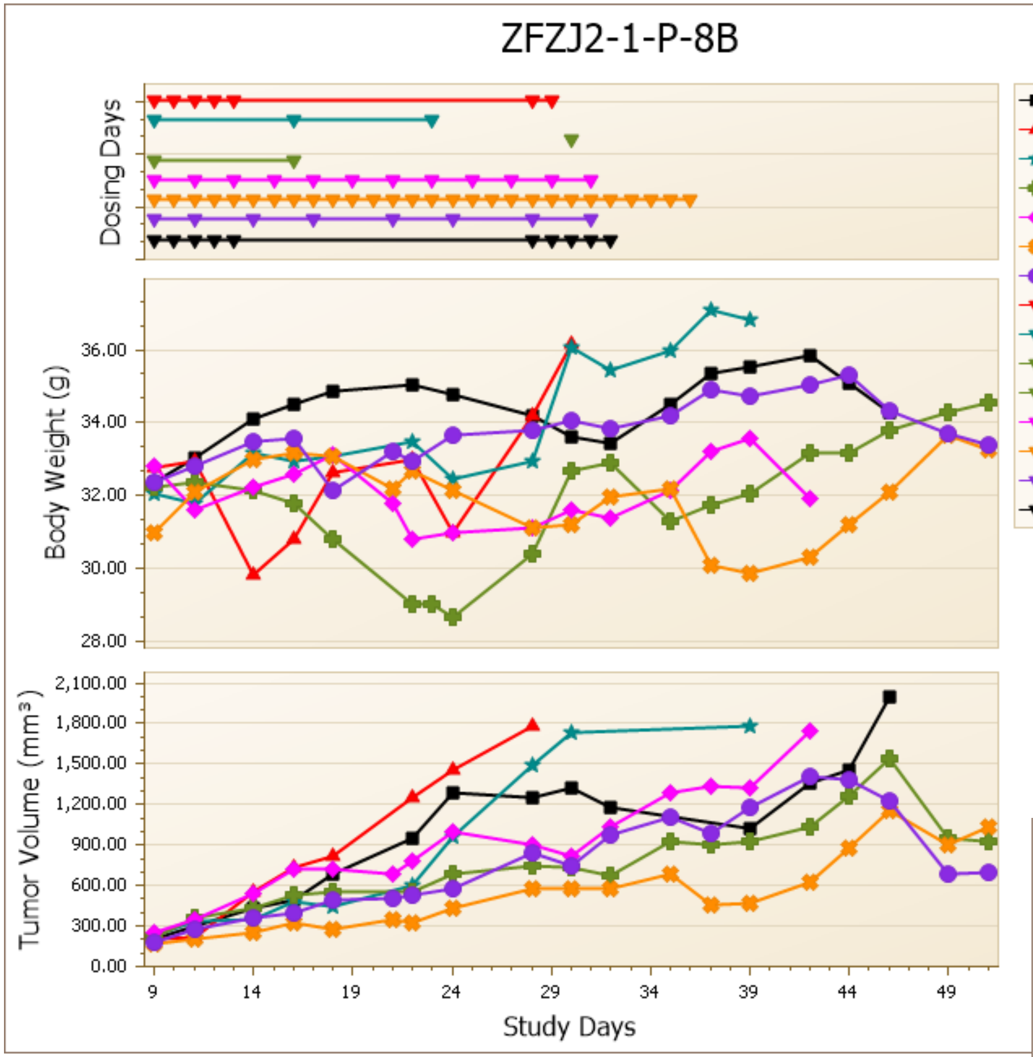
Tumorigenesis
Create reprogrammed cell lines
Transplantation into NSG mice
Tumor/patient heterogeneity

Nature Rev. Clin. Oncol. 11: 649-662, 2014.

Cancer Cell Lines



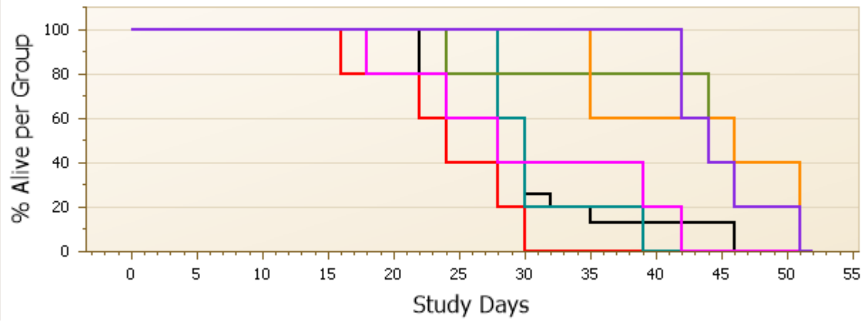
CL and PD Xenografts



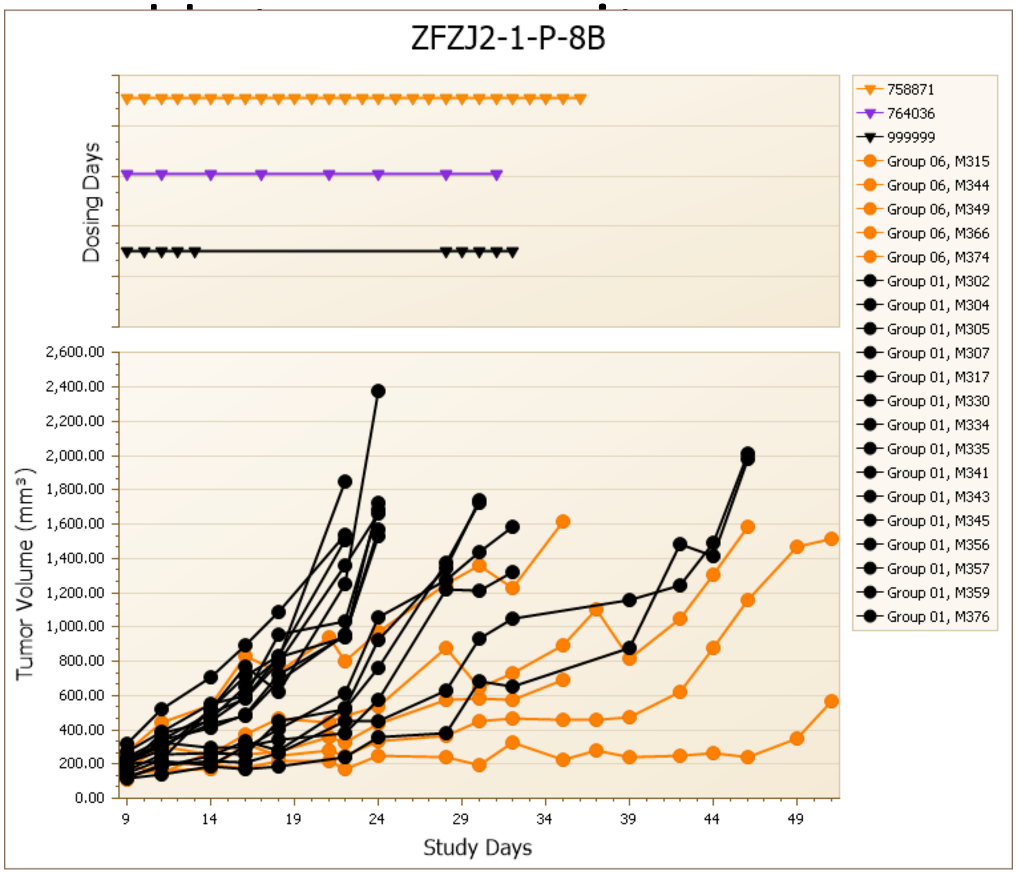
ZFZJ2-1, 146476-266-R, Urothelial/bladder ca

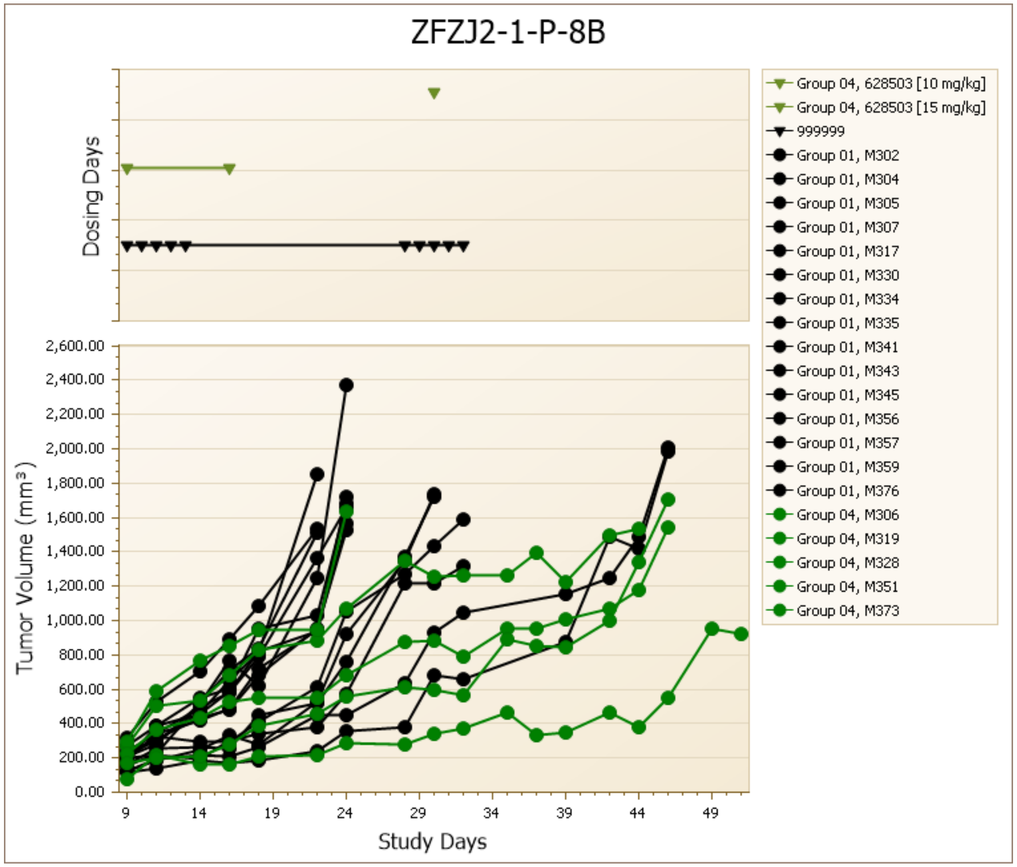
AZD8055 ~2.2x delay. No regression

Paclitaxel	125973
AZD8055	758871
Dinaciclib	764036
Docetaxel	628503
GSK-461364	754354
Methotrexate	740



ZFZJ2-1, 146476-266-R, Urothelial/bladder ca



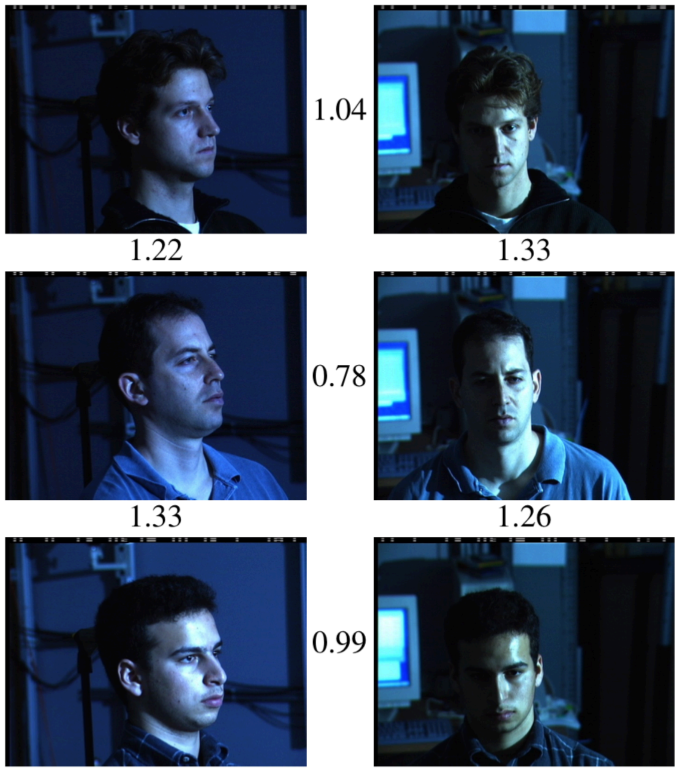


Op Meeting:

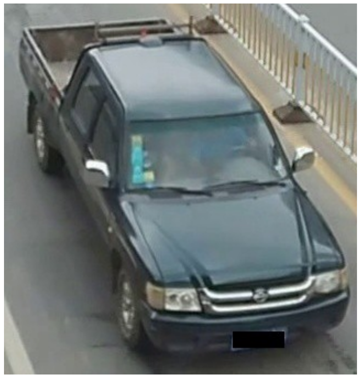
CANDLE prediction analysis notebook

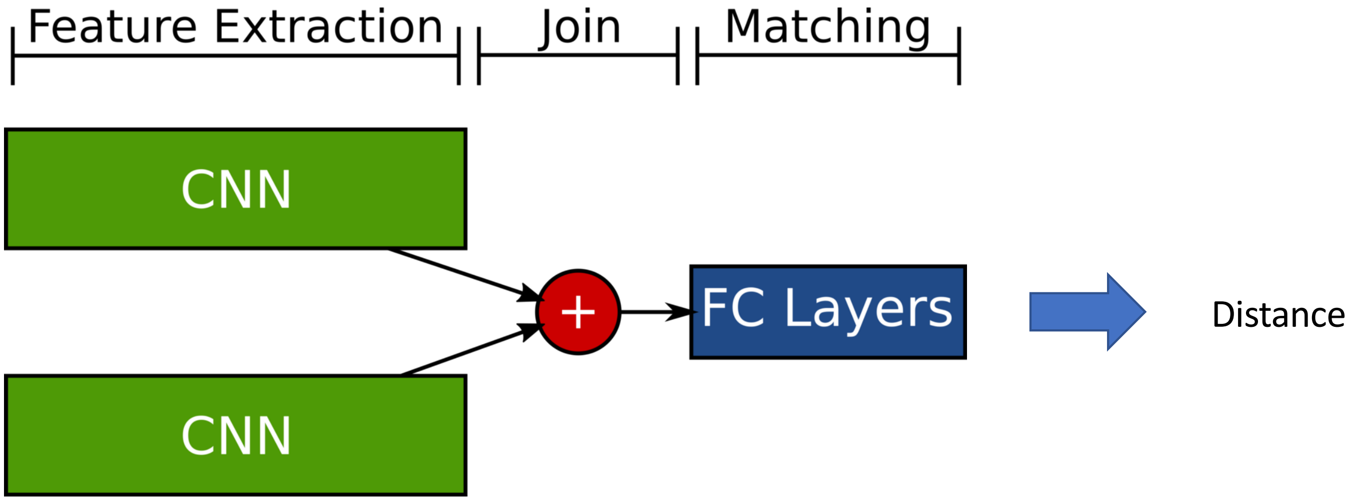
Perceptual distance vs data distance

- Metric learning
- Representation learning
- Feature encoding
- Embedding
- WordVec, ProteinVec



Siamese network





Pairs could be gene expression replicates or samples from the same cancer type

Focusing on the difficult parts

- Mine the difficult samples

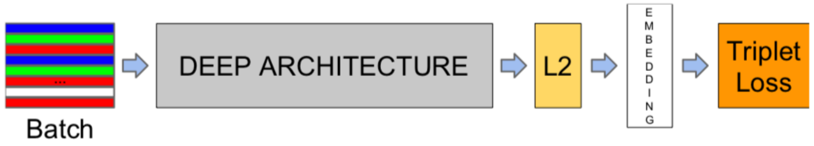


Figure 2. **Model structure.** Our network consists of a batch input layer and a deep CNN followed by L_2 normalization, which results in the face embedding. This is followed by the triplet loss during training.

- Change loss function

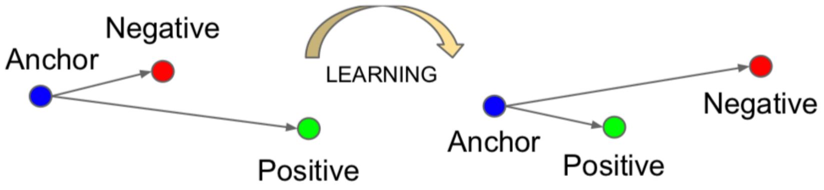
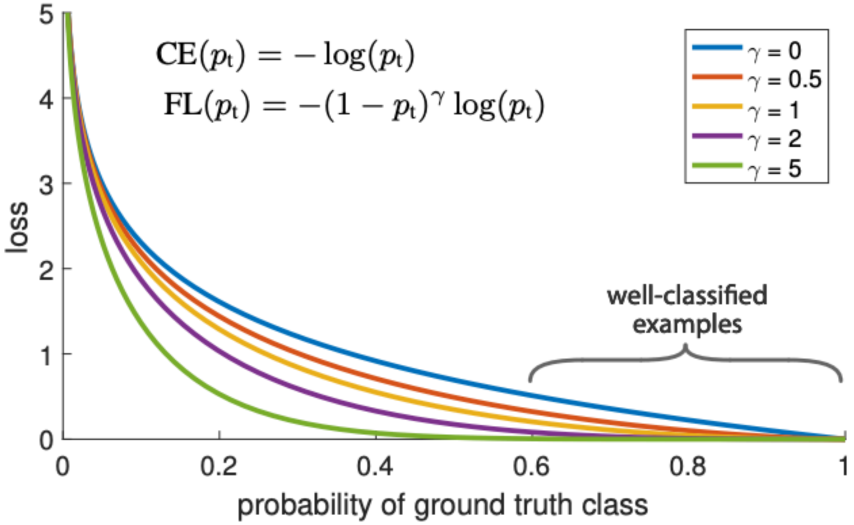
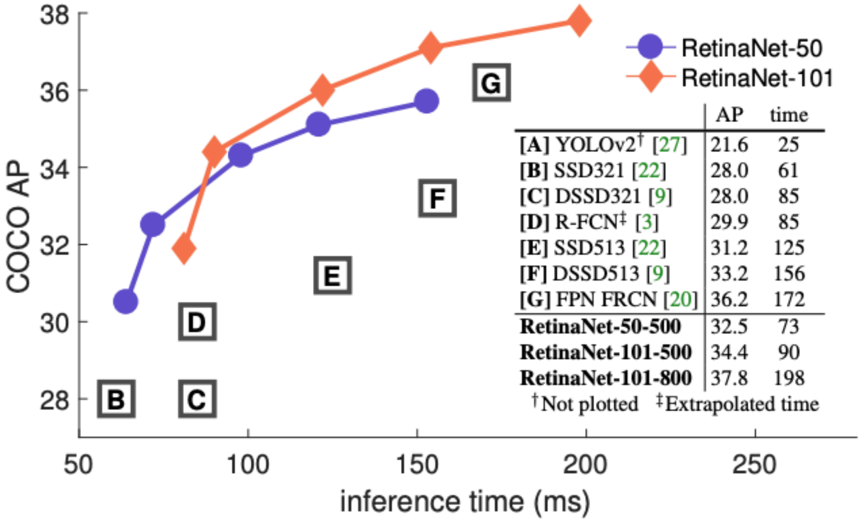


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

Focal Loss for Dense Object Detection

Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He Piotr Dollár
Facebook AI Research (FAIR)





	AP	time
[A] YOLOv2[†] [27]	21.6	25
[B] SSD321 [22]	28.0	61
[C] DSSD321 [9]	28.0	85
[D] R-FCN[‡] [3]	29.9	85
[E] SSD513 [22]	31.2	125
[F] DSSD513 [9]	33.2	156
[G] FPN FRCN [20]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

[†] Not plotted [‡] Extrapolated time

DL Challenge:

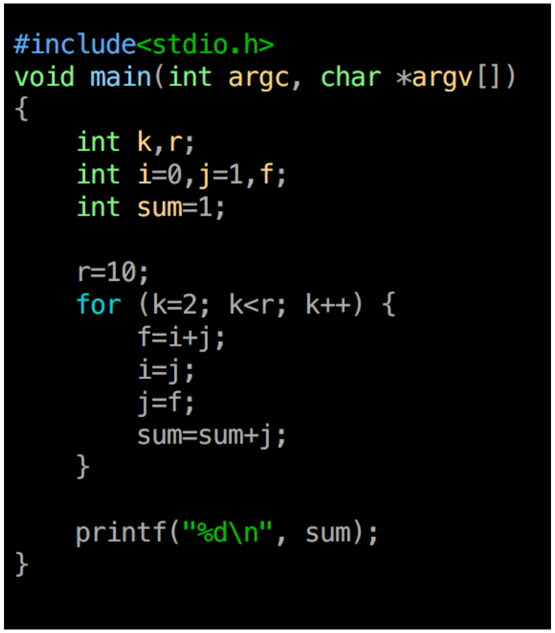
from surface pattern recognition
to deep mechanistic understanding

Just fancy regression?

Sequence to sequence models

Input		Output
1+1	**=**	**2**
19+28	**=**	**47**
577+45	**=**	**622**
👍+👍	=	✌️
1️⃣9️⃣+2️⃣8️⃣	=	4️⃣7️⃣
五七七+四五	=	六二二

Function approximation



```
#include<stdio.h>
void main(int argc, char *argv[])
{
    int k,r;
    int i=0,j=1,f;
    int sum=1;

    r=10;
    for (k=2; k<r; k++) {
        f=i+j;
        i=j;
        j=f;
        sum=sum+j;
    }

    printf("%d\n", sum);
}
```

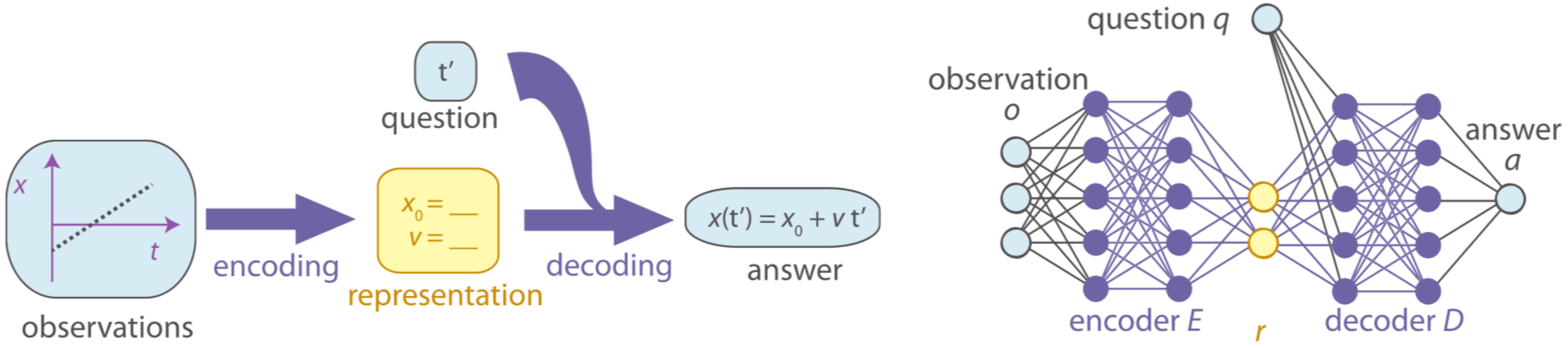
=> 88

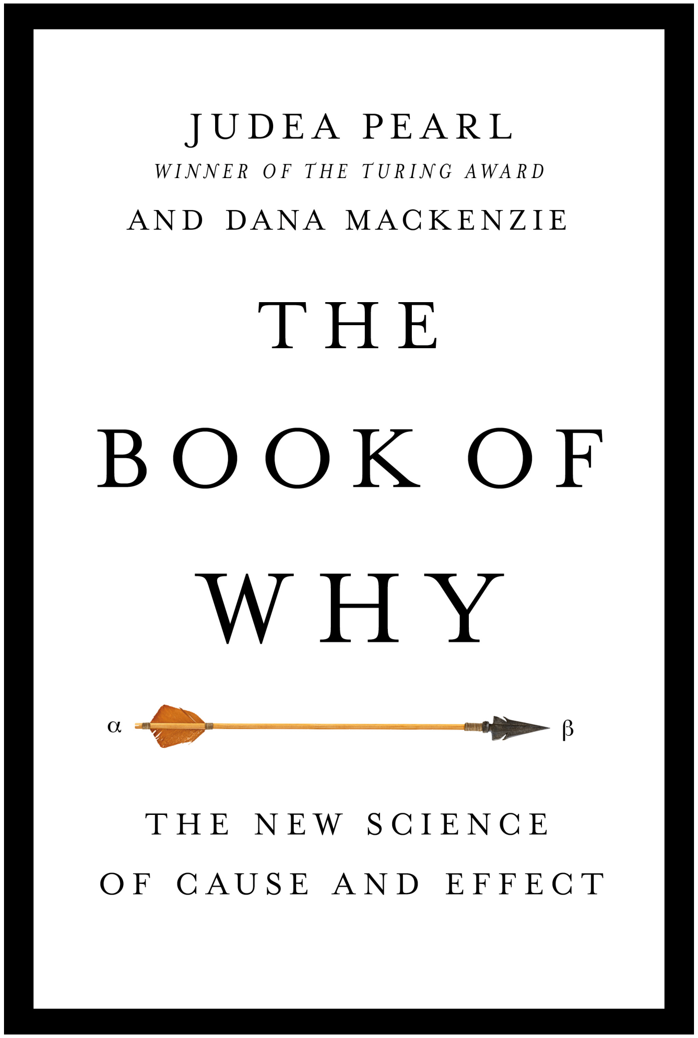
Discovering physical concepts with neural networks

Raban Iten, Tony Metger, Henrik Wilming, Lidia del Rio, Renato Renner

(Submitted on 26 Jul 2018 (v1), last revised 29 Sep 2018 (this version, v2))

We introduce a neural network architecture that models the physical reasoning process and that can be used to extract simple physical concepts from experimental data without being provided with additional prior knowledge. We apply the neural network to a variety of simple physical examples in classical and quantum mechanics, like damped pendulums, two-particle collisions, and qubits. The network finds the physically relevant parameters, exploits conservation laws to make predictions, and can be used to gain conceptual insights. For example, given a time series of the positions of the Sun and Mars as observed from Earth, the network discovers the heliocentric model of the solar system - that is, it encodes the data into the angles of the two planets as seen from the Sun. Our work provides a first step towards answering the question whether the traditional ways by which physicists model nature naturally arise from the experimental data without any mathematical and physical pre-knowledge, or if there are alternative elegant formalisms, which may solve some of the fundamental conceptual problems in modern physics, such as the measurement problem in quantum mechanics.





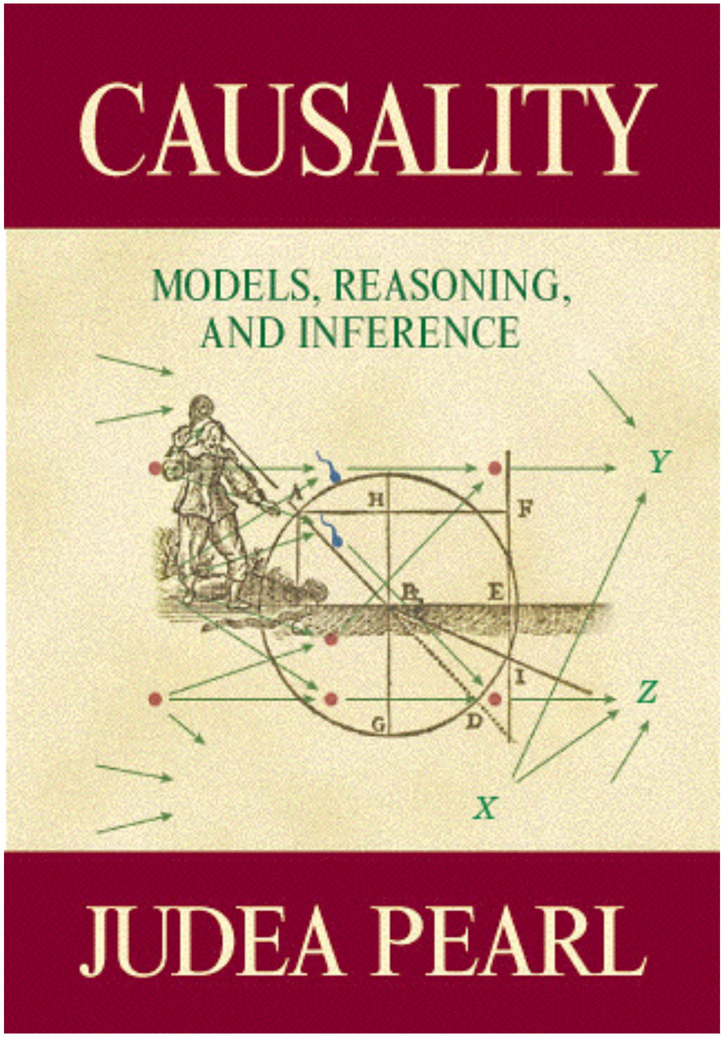
JUDEA PEARL

WINNER OF THE TURING AWARD

AND DANA MACKENZIE

THE BOOK OF WHY

THE NEW SCIENCE OF CAUSE AND EFFECT



CAUSALITY

MODELS, REASONING, AND INFERENCE

JUDEA PEARL

Thank you