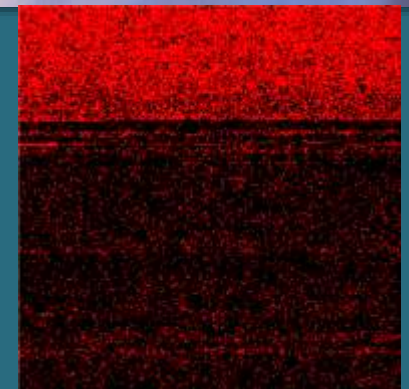


# Cancer Type/Site Classification using Deep-Learning (Preliminary presentation slides)

**S. Ravichandran, Ph.D**  
BIDS, FNLCR



# Acknowledgements

- **NCI-DOE Pilot-1 Team**
- **BIDS**
  - Drs. George Zaki, Andrew Weissman, Mark Jensen and Eric Stahlberg
  - Amar Khalsa, Dr. Deb Hope, Anney Che, Hue Readron, Dr. Yongmei Zhao
  - Colleagues who reviewed the material

# Feel free to follow-along

## Github

- <https://github.com/ravichas/ML-TC1>

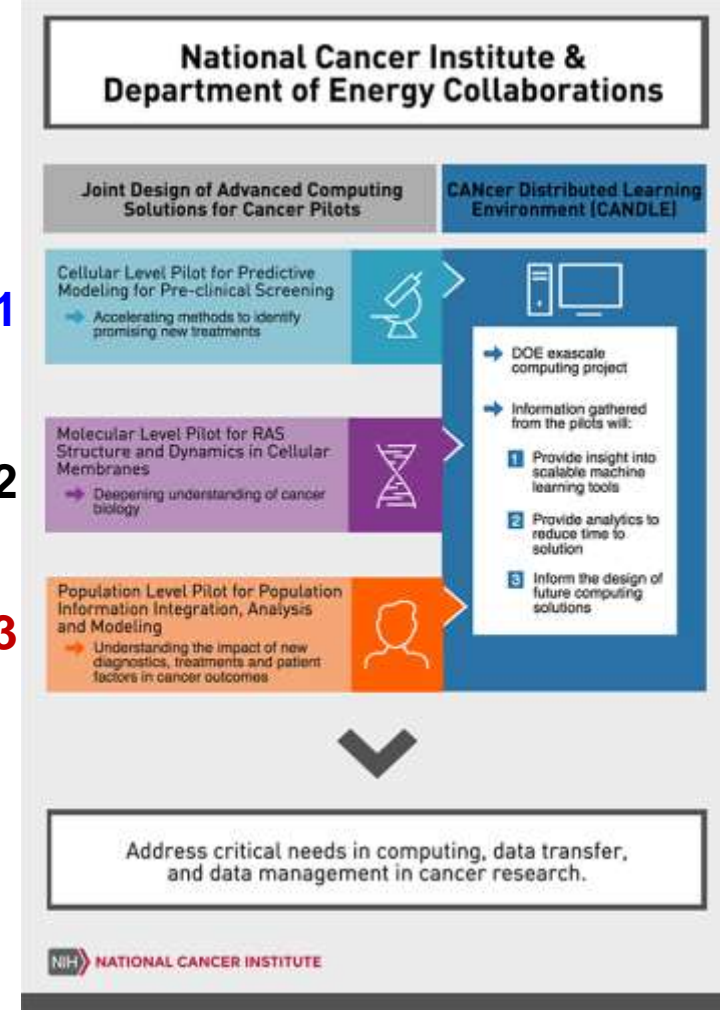
# The Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

- JDACS4C program was created in 2016 to accelerate cancer research using emerging exascale computing capabilities.
- Part of the Cancer Moonshot
- Cross-agency collaboration between NCI and the DOE
- Pilot1:
  - Focuses on developing predictive models, both *computational* and *experimental*, to improve pre-clinical *therapeutic drug screening*.
  - <https://datascience.cancer.gov/collaborations/joint-design-advanced-computing/cellular-pilot>

Pilot1

Pilot2

Pilot3



# Introduction

- **Goal is to share tools/techniques/solutions for cancer related problems. We often take a test-case and show how it works**
- **You would be able to take our test-case (code/scripts) and tune it to your needs**
- **We want to hear from you, please send us your feed-back**

# Motivation: Cancer Prediction vs Cancer Detection

- Cancer Prediction has been the major focus
  - Prognosis, Recurrence, Susceptibility
- Cancer Detection (classification of tumors/cancers) is lagging behind Prediction and we would like to share an application that might be useful
  - Detect/Identify cancer type at an early stage

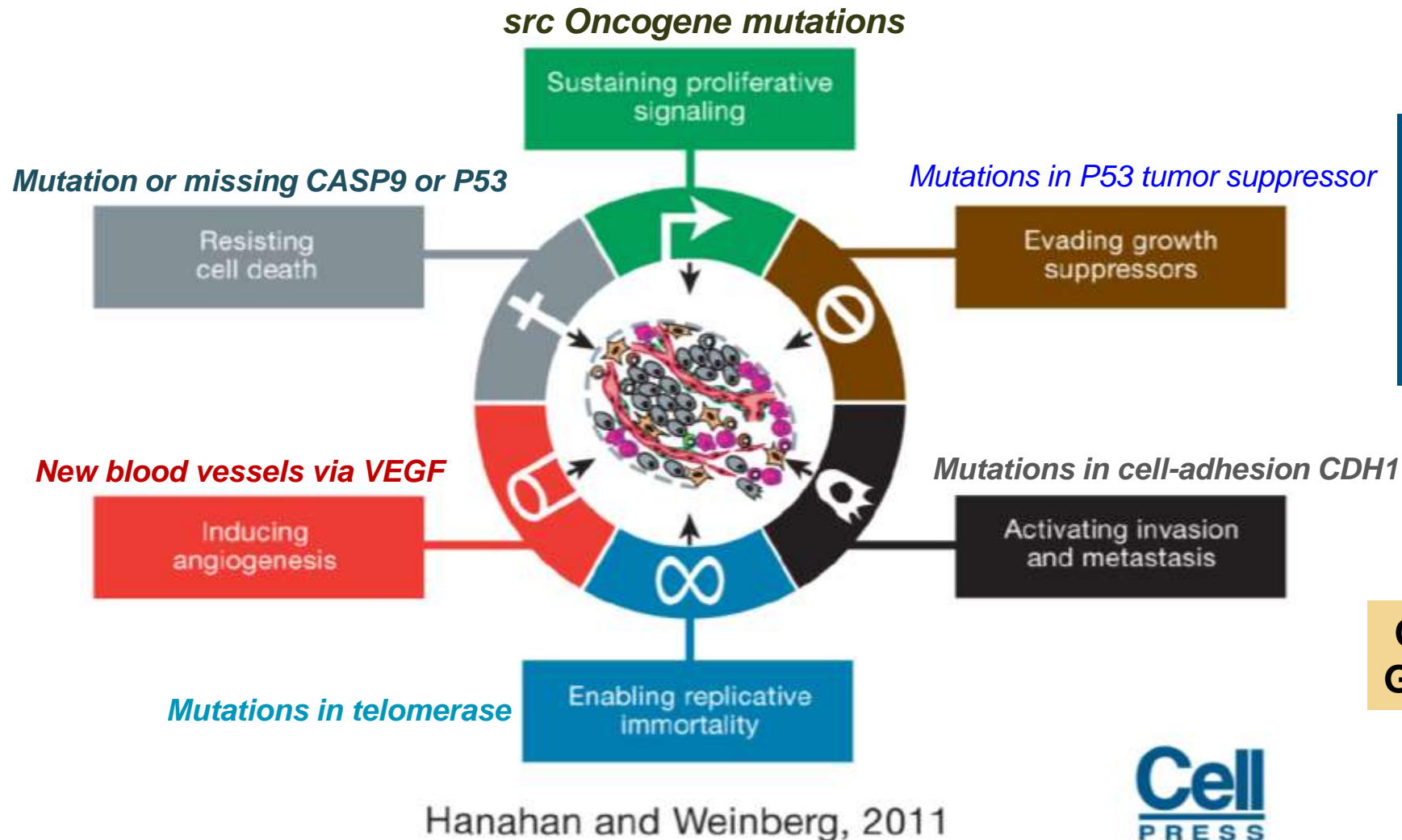
## Goal(s)/Questions

- **Take genomic expression data from tumor/cancer samples and apply Deep-Learning to create cancer types/site(s) classifier models**
- **Are the expression profiles unique to be used for early cancer detection?**
  - Improving chance of early detection cure/survival?



# Hallmarks of cancer: Integral Components of Most Forms of Cancer (Acquired Capabilities)

Hallmarks of Cancer: The Next Generation



REVIEW | VOLUME 100, ISSUE 1, P57-70, JANUARY 07, 2000

## The Hallmarks of Cancer

Douglas Hanahan • Robert A Weinberg

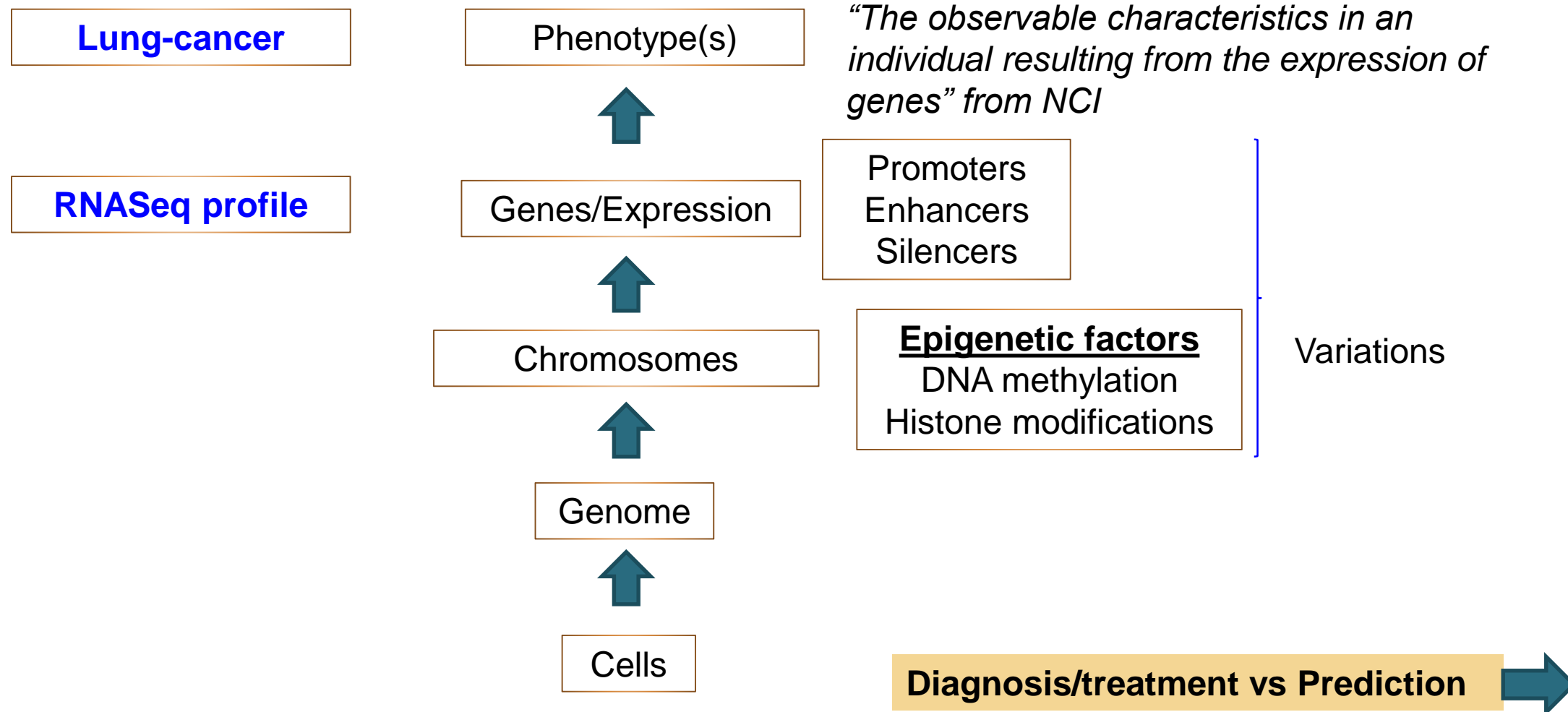
Open Archive • DOI: [https://doi.org/10.1016/S0092-8674\(00\)81683-9](https://doi.org/10.1016/S0092-8674(00)81683-9)

Overview of  
Genotype/phenotypes?





# Influence of genomic features on phenotypes: An overview



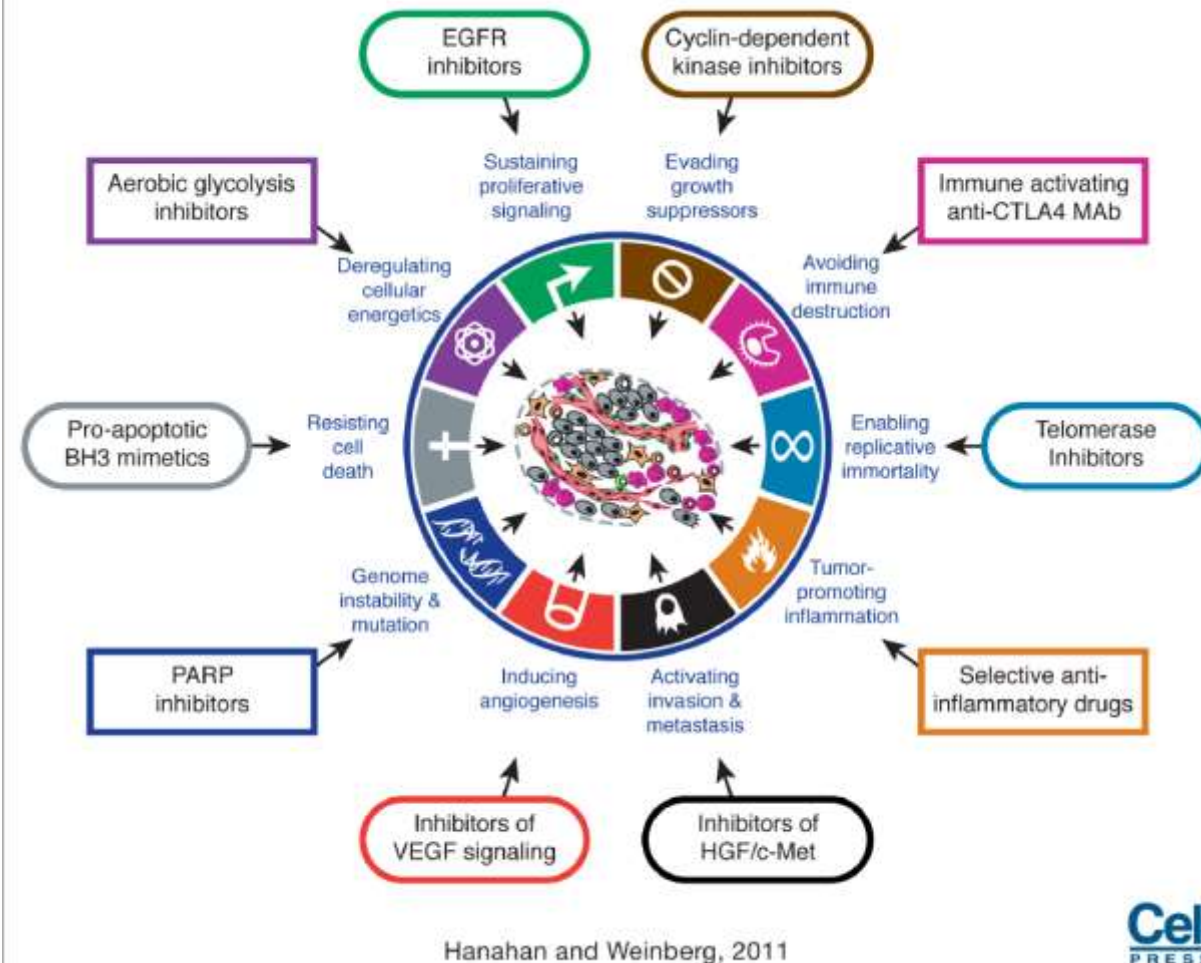
# Treatment vs Type-Prediction

## • Treatment

- Gene-centric (or a slice of pathway)
- Disease:
  - Tumor is called a gastrointestinal stromal tumor, or GIST
  - Medicine/inhibitor: Imatinib targeting BCR/KIT

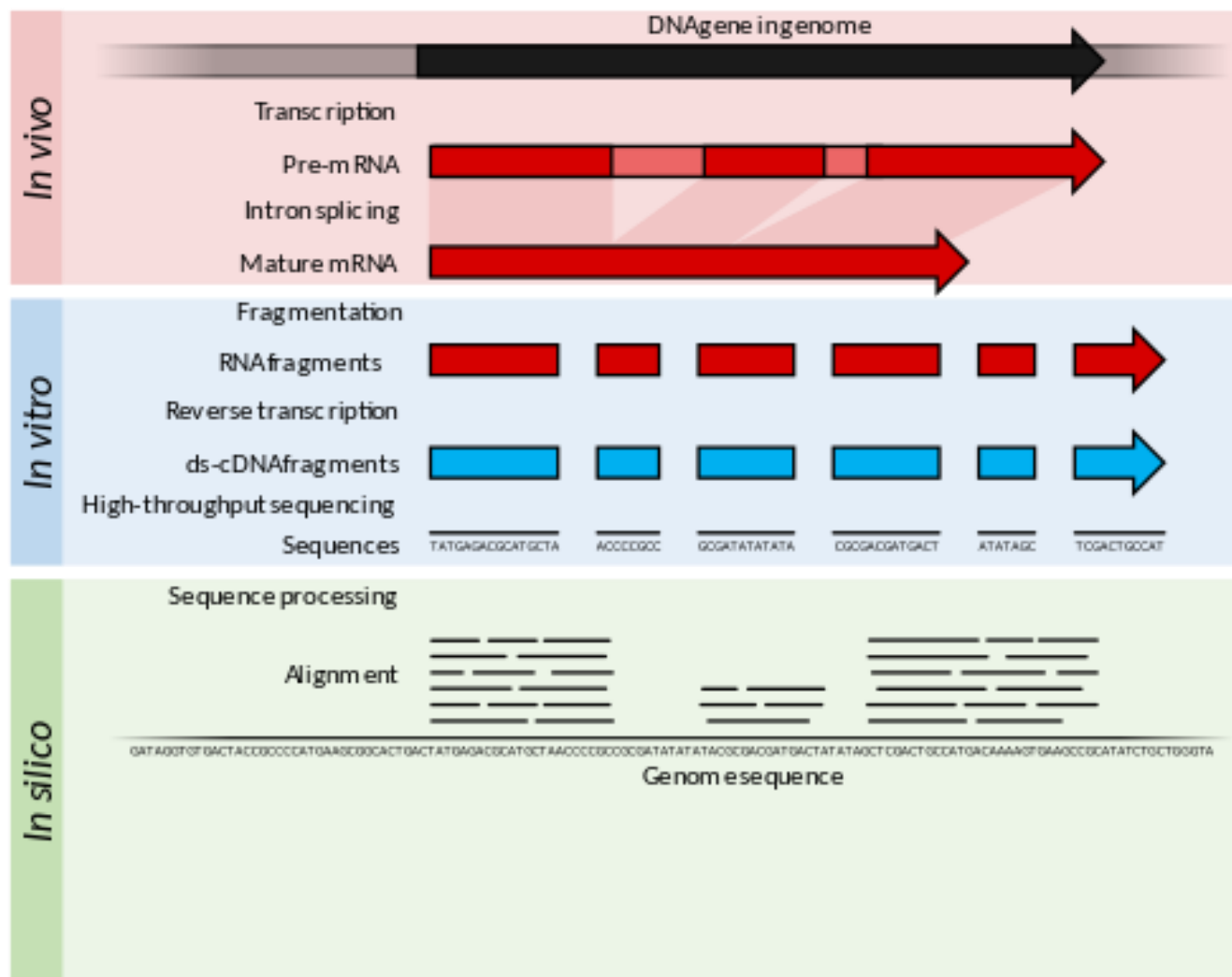
## • Detecting Type

- Genomic instability in Cancer Cells → Random mutations → rare genetic changes that can orchestrate hallmark capabilities. (Hanahan and Weinberg 2011)
- “The architecture of occurring genetic aberrations such as somatic mutations, CNVs, changed gene expression profiles, and different epigenetic alterations, is unique for each type of cancer.”, DOI: 10.5114/wo.2014.47136
- <https://pubmed.ncbi.nlm.nih.gov/26963104/> (PLOS, 2016)



# Expression data

NGS



Spliced to become mature mRNA  
mRNA is extracted

mRNA captured/fragmented/copied  
into stable ds-cDNA  
**Sequenced**

**Reference Genome**

NGS

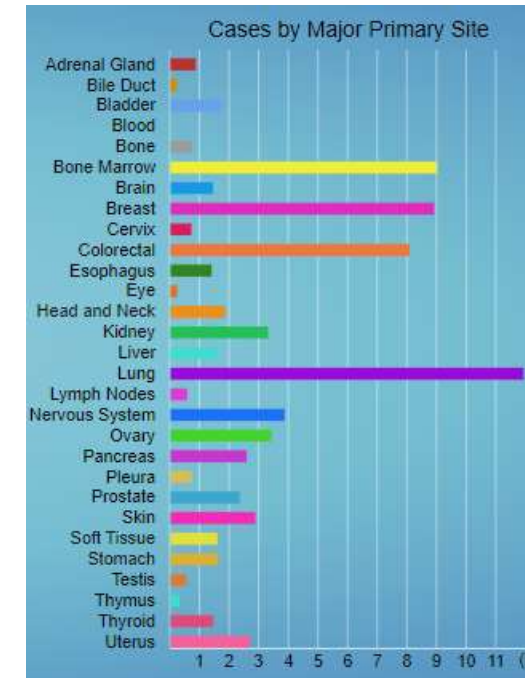
## Data source: The Cancer Genome Atlas (TCGA)

- NIH launched TCGA Pilot Project – a public funded project
- Goal of creating a comprehensive “atlas” of cancer genomic profiles.
- Large cohorts of over 30 human tumors through large-scale genome sequencing and integrated multi-dimensional analyses.
- Contains Microarray and NGS data
  - RNASeq
  - miRNA seq
  - SNP based platforms
  - .....
- TCGA data is available via GDC

<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>

# Data Harmonization: GDC ( <https://gdc.cancer.gov/> )

- Data and metadata is submitted to the GDC in standard data types and file formats. Other data sources (Ex. TCGA) are also included
- Data are harmonized against a common reference genome (GRCh38)
- For this workshop, we will focus on TCGA Genomic expression data from GDC



# Expression Data Quantification

- $RC_g$ : Number of reads mapped to the gene
- $RC_{g75}$ : The 75th percentile read count value for genes in the sample
- $L$ : Length of the gene in base pairs; Calculated as the sum of all exons in a gene

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

FASTQ

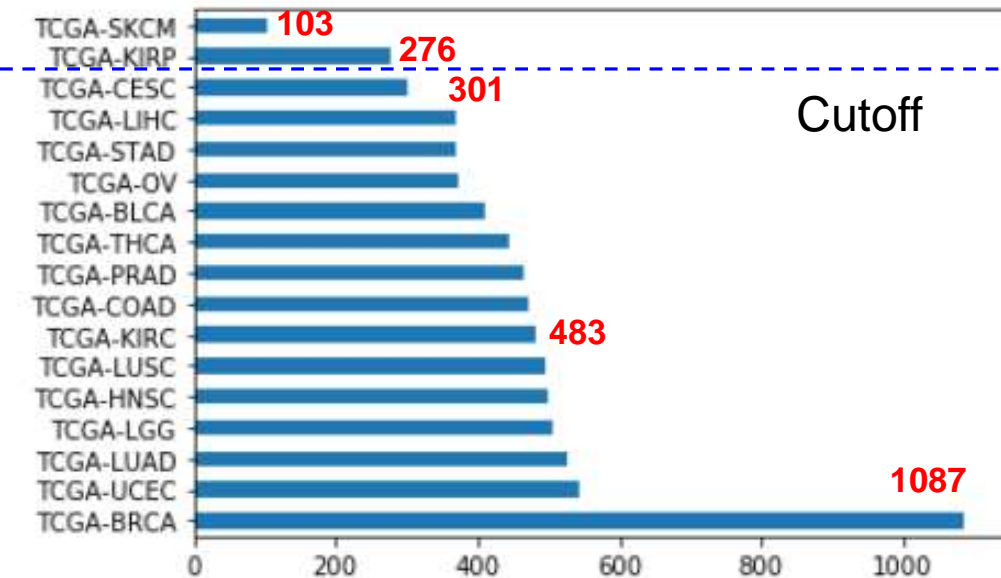
Alignment to Ref  
Genome (SAM/BAM)

Quantification HTSeq

Gene Expression  
(FPKM-UQ)

Fragments **P**er **K**ilobase of transcript per **M**illion mapped reads

# How much data for modeling?



CODE	Cancer Site/Type
BRCA	Breast invasive carcinoma
UCEC	Uterine Corpus Endometrial Carcinoma
LUAD	Lung adenocarcinoma
LGG	Brain Lower Grade Glioma
HNSC	Head and Neck squamous cell carcinoma
LUHSC	Lung squamous cell carcinoma
KIRC	Kidney renal clear cell carcinoma
PRAD	Prostate adenocarcinoma
COAD	Colon adenocarcinoma
THCA	Thyroid carcinoma
BLCA	Bladder Urothelial Carcinoma
OV	Ovarian serous cystadenocarcinoma
STAD	Stomach adenocarcinoma
LIHC	Liver hepatocellular carcinoma
CEC	Cervical squamous cell carcinoma and endocervical adenocarcinoma

**300  
samples  
each**



# Expression data from a sample

## TCGA-BRCA

Genes	Expression
ENSG00000242268.2	1658.464179
ENSG00000270112.3	460.2343433
ENSG00000167578.15	52440.10096
ENSG00000273842.1	0
ENSG00000078237.5	68165.45626
ENSG00000146083.10	255959.2351
ENSG00000225275.4	0
ENSG00000158486.12	104.9473768
ENSG00000198242.12	4968556.658
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000201788.1	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	3484.027373
ENSG00000172137.17	41485.9507
ENSG00000167700.7	226717.4208
ENSG00000234943.2	2082.245035
ENSG00000240423.1	310.5246749
ENSG00000060642.9	155863.9216
ENSG00000271616.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000231105.1	1583.312582
ENSG00000243044.1	0
ENSG00000182141.8	45538.60648
ENSG00000269416.4	119.0847054
ENSG00000264981.1	0

60,483  
transcripts

Gene: AC090241.2 ENSG00000270112

Description

novel transcript, antisense to ST8SIA5

Location

[Chromosome 18: 46,756,487-46,802,449](#) forward strand.  
 GRCh38:CM000680.2

About this gene

This gene has 8 transcripts ([splice variants](#))

Transcripts

Hide transcript table

Gene: DNAH3 ENSG00000158486

Description

dynein axonemal heavy chain 3 [Source:HGNC Symbol;Acc:[HGNC:2949](#)]

Gene Synonyms

DKFZp434N074, DLP3, Dnahc3b, Hsadhc3

Location

[Chromosome 16: 20,933,111-21,159,441](#) reverse strand.  
 GRCh38:CM000678.2

About this gene

This gene has 6 transcripts ([splice variants](#)), [371 orthologues](#), [14 paralogues](#) and is a member of [1 Ensembl protein family](#).

Transcripts

Hide transcript table

# Data Preparation

# Sample1

## Sample2

## Sample3

## Sample4

Sample297

Sample298

Sample299

Sample300

# Breast Cancer

60,484  
transcripts

Genes	Expression
ENSG00000232608.2	3658.464179
ENSG00000270122.3	404.213443
ENSG00000167578.15	52440.130038
ENSG00000148423.1	104.484243
ENSG00000278237.5	88116.426356
ENSG00000140683.10	25569.25131
ENSG00000252275.4	0
ENSG00000146842.12	34773.76758
ENSG00000128242.12	406856.658
ENSG00000125983.1	6108.999052
ENSG00000231981.3	0
ENSG00000257788.2	0
ENSG00000257788.2	0
ENSG00000134108.11	957330.2056
ENSG00000263089.1	4384.07773
ENSG00000172137.17	495.9917
ENSG00000172137.17	226717.4048
ENSG00000234943.2	2082.24055
ENSG00000240423.1	130.5246749
ENSG00000172137.17	555863.9216
ENSG00000275161.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.4755669
ENSG00000140575.1	581.312582
ENSG00000234944.1	1
ENSG00000128124.18	45318.86648
ENSG00000269146.4	118.087054

Genes	Expression
ENSG0000024268.2	1658.464179
ENSG0000027101.3	460.234143
ENSG0000027578.15	5240.130403
ENSG0000027184.1	1264.1
ENSG0000027037.5	6185.464262
ENSG0000027103.10	2559.95125
ENSG0000027527.4	0
ENSG0000027488.12	947.761768
ENSG0000021984.12	4096556.58
ENSG0000025983.1	6108.999052
ENSG0000023198.1	0
ENSG0000027188.1	0
ENSG0000027188.1	0
ENSG0000014108.11	957310.2056
ENSG0000028308.1	3148.627793
ENSG0000027137.17	147.000000
ENSG0000027030.7	22617.4028
ENSG0000024943.2	2082.246505
ENSG0000024042.1	103.52684789
ENSG0000027156.11	155863.9216
ENSG0000027156.11	0
ENSG0000023488.1	0
ENSG0000023488.1	394.4755669
ENSG0000023488.1	1581.312156
ENSG0000023488.1	0
ENSG0000021824.8	45338.06438
ENSG0000026416.4	118.078454

Genes	Expression
ENSG0000024268.2	3058.46479
ENSG00000210718.1	404.23443
ENSG00000107576.5	52440.33043
ENSG0000013861.1	10.38611
ENSG0000027827.5	6815.48626
ENSG0000014608.1	25599.2515
ENSG0000025275.4	0
ENSG0000018488.12	947.94768
ENSG0000019842.12	409556.58
ENSG0000025983.1	6108.999052
ENSG0000011983.1	0
ENSG0000020475.2	0
ENSG0000017688.1	0
ENSG0000013410.11	957310.2056
ENSG0000020689.10	3484.02773
ENSG0000017317.17	2467.1047
ENSG0000017903.7	22617.408
ENSG0000023404.2	2082.245439
ENSG0000014042.11	320.534679
ENSG0000017545.1	155863.9212
ENSG0000027165.1	0
ENSG0000023488.1	0
ENSG0000020601.3	394.475569
ENSG0000011105.1	5581.31582
ENSG0000024044.1	0
ENSG0000018214.18	45318.06048
ENSG0000019246.1	118.087454

Genes	Expression
ENS00000242068.2	3658.484179
ENS00000201723.1	460.2341343
ENS00000270316.5	52440.10096
ENS00000218482.1	1748.21
ENS00000278273.5	6835.45626
ENS00000460810.0	25059.2551
ENS00000252574.0	4
ENS00000244882.12	492.187268
ENS00000298422.12	406856.68
ENS00000259833.1	6168.99052
ENS00000219833.0	4
ENS00000296572.0	4
ENS00000219833.0	4
ENS00000114821.1	957330.2056
ENS00000263099.1	3484.027373
ENS00000271017.0	41449.57
ENS00000271017.0	23827.14207
ENS00000424453.2	2002.244503
ENS00000240423.1	130.5248749
ENS00000271017.0	155863.9276
ENS00000274561.0	4
ENS00000234881.1	4
ENS00000236040.1	334.255683
ENS0000021105.1	1581.312582
ENS0000023044.1	304.0
ENS0000023141.8	45338.60548
ENS00000292466.0	1184.087504

Genes	Expression
ENSG0000024268.2	12508.404179
ENSG00000070112.3	460.234413
ENSG00000178812.5	52.1404.30096
ENSG00000178275.5	68.0546.36266
ENSG00000146081.30	20.5559.2351
ENSG00000252574.0	
ENSG00000168682.1	124.9437768
ENSG00000190422.12	49.0856.6858
ENSG00000158831.4	63.9409.552
ENSG00000131981.3	
ENSG00000104973.2	
ENSG00000114081.1	
ENSG00000143811.5	95.7830.2056
ENSG00000130081.1	3484.6277379
ENSG00000172117.1	41.1414.30096
ENSG00000167700.2	22.9717.4208
ENSG00000140432.3	20.82.2405
ENSG00000140343.3	120.5246749
ENSG00000160629.9	15.5683.152
ENSG00000171651.1	
ENSG00000144881.1	
ENSG00000136401.1	334.4755669
ENSG00000135411.5	158.63.31252
ENSG00000130481.1	
ENSG00000182141.6	45.538.40648
ENSG00000196416.4	119.087054

Genes	Expression
ENSG00000242282	1658.464179
ENSG00000270113	460.234343
ENSG00000276728	12404.10096
ENSG00000274412	10.000000
ENSG00000278375	68405.4456
ENSG00000246035	255959.2351
ENSG00000252974	0.000000
ENSG00000256815	12.9472738
ENSG00000259842	490856.5616
ENSG00000259833	6108.59952
ENSG00000231983	0.000000
ENSG00000259832	0.000000
ENSG0000021788	0.000000
ENSG0000034308	957330.266
ENSG0000036089	3484.077373
ENSG00000371217	17.000000
ENSG00000367700	228751.4208
ENSG0000024943	2082.2405
ENSG0000040243	3015.524670
ENSG0000040493	53863.5212
ENSG0000027526	0.000000
ENSG0000023488	1.000000
ENSG0000033604	394.475569
ENSG0000033604	1383.112582
ENSG0000033604	0.000000
ENSG00000321418	45338.60648
ENSG0000044646	119.087034

Gene	Expression
ENSG00000242058.2	1658.464179
ENSG00000270712.1	40.1214343
ENSG00000275125.1	52440.10096
ENSG00000274142.1	1.000000000
ENSG00000276237.5	68165.46626
ENSG00000240380.3	205591.2351
ENSG00000252575.4	0
ENSG00000256816.12	234.947358
ENSG00000238042.2	408956.5856
ENSG00000259883.1	6208.990962
ENSG00000231981.3	0
ENSG00000256945.2	0
ENSG00000217408.1	0
ENSG00000213801.15	97310.26106
ENSG00000256309.1	3484.02773
ENSG00000271217.1	1481.4952
ENSG00000256700.7	22671.42420
ENSG00000234493.2	2082.24055
ENSG00000240423.1	30152.64799
ENSG00000256442.9	55863.91262
ENSG00000275266.1	0
ENSG00000234881.1	0
ENSG00000236401.1	394.475569
ENSG00000236411.1	1583.112582
ENSG00000234004.1	0
ENSG00000282146.8	4533.60648
ENSG00000259444.4	129.087054

Case	Expression
ENGG0000024268.2	1658.464179
ENGG0000012712.3	40.213443
ENGG0000016017.8	53440.1006
ENGG0000014824.1	10.000000
ENGG0000016708.2	6818.45626
ENGG00000146083.0	255939.2351
ENGG00000125275.4	0
ENGG0000014668.12	234.947378
ENGG0000018924.12	4098568.558
ENGG000001259883.1	6108.995902
ENGG000001219881.3	0
ENGG000001368975.2	0
ENGG00000125788.1	0
ENGG00000134108.11	967330.2656
ENGG00000126308.1	3484.02773
ENGG00000126308.17	4418.4697
ENGG00000126700.7	2267.41208
ENGG00000123493.42	2802.45035
ENGG000001240623.1	335.5426740
ENGG000001240623.9	155836.932
ENGG00000125756.1	0
ENGG000001234884.1	0
ENGG000001260401.1	394.455669
ENGG00000125581.1	1583.31282
ENGG00000123404.1	0
ENGG000001282141.8	45338.4608
ENGG000001269416.4	119.087054

# Data Preparation

Sample1	Sample2	Sample3	Sample4		Sample297	Sample298	Sample299	Sample300
---------	---------	---------	---------	--	-----------	-----------	-----------	-----------

# Breast Cancer

Games	Expression
ENSG00000242826.2	3558.464179
ENSG00000270123.1	404.234313
ENSG00000257125.1	52440.1300
ENSG0000021842.1	184.01
ENSG00000273172.3	2815.45626
ENSG00000214087.5	689.999052
ENSG00000252554.0	120.923481
ENSG00000248122.12	94.9473768
ENSG00000238242.2	406.85658
ENSG00000229883.1	6108.999052
ENSG00000213983.1	
ENSG00000239472.2	
ENSG00000217881.1	
ENSG000002114108.11	95.733026
ENSG00000260819.1	4848.027373
ENSG00000271712.17	22.617
ENSG00000217700.7	2267.470
ENSG00000240423.1	2082.140525
ENSG00000240921.3	130.5246740
ENSG00000240922.9	50.852316
ENSG00000274656.1	
ENSG00000244881.1	
ENSG00000240340.1	374.455660
ENSG00000211105.1	158.312582
ENSG00000243044.1	
ENSG00000242141.8	455.3810845
ENSG00000240194.18	119.684704

Games	Expression
ENSG00000242268.2	1658.464179
ENSG00000071718.1	460.234343
ENSG00000150735.1	52460.13005
ENSG0000013842.1	1842.1
ENSG00000178217.5	68155.46526
ENSG00000140681.0	25519.9.235.1
ENSG00000252575.4	0
ENSG00000148448.12	194.947368
ENSG00000192842.12	4068556.55
ENSG00000215881.1	6108.999052
ENSG00000219893.0	0
ENSG00000234852.1	0
ENSG00000117488.1	0
ENSG00000114301.18	973730.2656
ENSG00000263809.1	3468.627373
ENSG00000172312.17	48545.9051
ENSG00000174207.17	22617.4107
ENSG00000214943.2	2082.45025
ENSG00000240243.1	130.5246749
ENSG00000192909.9	550683.3216
ENSG00000274661.1	0
ENSG00000218581.0	0
ENSG00000248811.0	0
ENSG00000216040.1	394.475569
ENSG0000011105.1	1583.312582
ENSG00000213044.1	0
ENSG00000182144.8	45338.06348
ENSG000002189416.1	118.087054

Genes	Expression
ENSG0000024268.2	1608.46749
ENSG0000021711.3	460.234343
ENSG0000027025.1	52440.30096
ENSG0000027841.2	1884.2
ENSG0000027827.5	68305.4656
ENSG0000024081.0	25599.291
ENSG0000025275.4	10.0
ENSG0000025486.1	914.947378
ENSG0000028426.12	408955.616
ENSG0000025983.1	6180.99052
ENSG0000023198.3	0.0
ENSG0000026047.2	0.0
ENSG00000210788.1	0.0
ENSG0000023410.11	957330.256
ENSG0000026308.1	3484.62773
ENSG0000027217.1	431.76717
ENSG0000026790.7	22617.400
ENSG0000023494.3	2062.40525
ENSG0000024042.1	121.5246749
ENSG0000026042.9	25663.522
ENSG0000027656.1	0.0
ENSG0000023481.1	0.0
ENSG0000023040.1	394.4756209
ENSG0000021105.1	493.312582
ENSG0000024040.1	45538.6046
ENSG0000024945.4	119.084704

Genes	Log	Expression
ENG000000242262	1258	4841.79
ENG000000271218	42	294.1343
ENG000000270735	13	52440.1300
ENG000000278842	1	784.21
ENG000000278275	5	68455.4562
ENG000000410810	10	25059.2351
ENG000000252754	0	2385.1
ENG000000284828	14	947.9378
ENG000000284122	12	498565.58
ENG000000259831	6	618.999052
ENG000000319813	0	338.0
ENG000000294572	2	400.000000
ENG000000217881	0	1.0
ENG000001141011	11	95730.2056
ENG000001389001	1	348.67773
ENG000001721717	17	27.17
ENG000001677707	7	22817.4100
ENG000002140432	1	2082.4585
ENG000002404231	13	52464.90
ENG000002745611	5	150683.922
ENG000002745611	0	338.0
ENG00000248811	0	1.0
ENG000002404041	1	394.475569
ENG000002110511	1	158.453282
ENG000002430404	1	1.0
ENG000002914168	18	45338.60648
ENG000002624618	18	108.87054

Genes	Time	Expression
ENSG00000242688.2	12658.2	1568.464179
ENSG00000270123.1	60	204.234433
ENSG00000271719.1	18	52440.10096
ENSG00000271842.1	18	184.234433
ENSG00000278275.1	68	68165.46626
ENSG00000280148.1	20	25599.2351
ENSG00000252574.0	6	0
ENSG00000254575.1	18	194.945738
ENSG00000280426.12	12	4089565.68
ENSG00000259831.1	6	6198.590952
ENSG00000251983.1	0	0
ENSG00000254075.2	2	0
ENSG00000217881.1	0	0
ENSG00000214108.11	15	95730.2056
ENSG00000260898.1	18	3484.657173
ENSG00000271217.1	18	174.234433
ENSG00000257907.1	70	226717.42049
ENSG00000240443.1	20	2082.26525
ENSG00000244942.1	130	5246749
ENSG00000246429.9	15	55683.5216
ENSG00000276561.1	6	0
ENSG00000248811.1	0	0
ENSG00000246041.1	104	475.475475
ENSG0000021155.1	18	158.464179
ENSG00000230404.1	38	33.333333
ENSG00000292148.8	45	45338.40648
ENSG00000269414.8	119	10847054

Genes	Log2 Expression
ENSG0000024268.2	1658.464179
ENSG00000270113	460.2934431
ENSG0000027812.3	52440.10096
ENSG0000027821.5	68265.45646
ENSG0000014800.1	20599.91281
ENSG0000025274.7	0
ENSG0000025866.1	52.9472768
ENSG0000028412.2	498826.5816
ENSG0000025983.1	6308.999052
ENSG0000023198.3	0
ENSG0000028472.9	0
ENSG0000001780.1	0
ENSG0000013408.1	957330.2056
ENSG0000000800.1	3484.027173
ENSG0000017211.7	147.21171
ENSG0000016700.2	22973.47048
ENSG0000024943.2	2082.240525
ENSG0000024351.3	310.5246749
ENSG0000026624.9	55083.92126
ENSG0000027456.1	0
ENSG0000023488.1	0
ENSG0000023040.1	394.4755477
ENSG0000011026.1	833.312582
ENSG0000004304.1	0
ENSG0000018214.8	45358.60648
ENSG0000004945.6	138.087034

Gene	Exp	Expression
ENSG00000242088.2	1658	4641.79
ENSG00000270112.3	460	23443.03
ENSG00000257815.3	2444	10096
ENSG00000246412.3	5	142.41
ENSG00000248235.5	68165	45626
ENSG00000260730.3	205919	2531
ENSG00000252574.5	0	0
ENSG00000254686.12	12	987.58
ENSG00000258042.2	94	49957.68
ENSG00000258083.1	6388	99002.952
ENSG00000251988.1	0	0
ENSG00000250425.2	0	0
ENSG00000201788.1	1400	11
ENSG00000213401.11	95730	2056
ENSG00000203089.1	3468	62773
ENSG00000217117.12	48	952
ENSG00000256700.7	22671	2420
ENSG00000248493.2	2082	2405
ENSG00000240523.1	531	52649.79
ENSG00000240626.2	18	52662
ENSG00000272565.1	0	0
ENSG00000234881.1	0	0
ENSG00000236040.1	39	47556.69
ENSG00000211815.1	8383	31262
ENSG00000240444.1	0	0
ENSG00000282148.6	45338	60648
ENSG00000249446.4	128	48705.4

Gene	Expression
ENSG00000244208.2	1658.464179
ENSG000002701.123	460.234343
ENSG000002785.23	10424.30096
ENSG000002784.2	178.08423
ENSG000002786.15	68165.46526
ENSG000002400.130	25599.2511
ENSG000002252.574	50
ENSG000002248.12	184646.12
ENSG000002382.12	194.947378
ENSG000002382.12	408955.68
ENSG000002598.11	6108.99052
ENSG000002319.13	40
ENSG000002345.2	100
ENSG00000230788.1	100
ENSG00000213408.111	957330.266
ENSG000002008.11	3484.027173
ENSG000002721.17	172.117
ENSG00000235700.7	22817.44078
ENSG000002349.42	2082.14052
ENSG000002403.1	310.5246749
ENSG000002362.42	236863.122
ENSG0000027156.12	100
ENSG000002348.11	100
ENSG000002340.1	394.475627
ENSG00000231125.1	833.112562
ENSG000002400.14	100
ENSG000002321.418	45338.40548
ENSG000002394.16	119.084704

# Lung Cancer

Genes	Population
ENSG000002428262	3526 484179
ENSG000002717123	460 2343433
ENSG000002505125	52440 10096
ENSG000002717123	460 2343433
ENSG000002408105	18165 46626
ENSG000002146078	105 25599 2351
ENSG00000225274	0
ENSG000002448148	13 1974738
ENSG000002358422	12 406856 6
ENSG000002578831	6108 99052
ENSG000002198313	0
ENSG000002408105	18165 46626
ENSG000002378811	0
ENSG000001340811	95730 20656
ENSG000002830801	1 4448 5077
ENSG000002733717	348 372173
ENSG000002408105	18165 46626
ENSG000002404212	2081 245405
ENSG000002404212	2081 245405
ENSG000002746511	151863 9216
ENSG000002746511	151863 9216
ENSG000002348811	1
ENSG000002404212	2081 245405
ENSG000002110511	583 3175562
ENSG000002430411	1
ENSG000002321418	455 38 60648
ENSG000002891644	118 087054

Genes	Expression
ENSG00000204268.2	3658.484179
ENSG00000201778.15	460.2343433
ENSG00000205753.25	52440.10096
ENSG00000201842.1	1364.2
ENSG00000201775.7	68165.46626
ENSG00000214082.10	255959.2351
ENSG00000252574.0	0
ENSG00000204848.12	1947.947768
ENSG00000219832.12	406856.658
ENSG00000205885.1	6108.990052
ENSG00000211983.0	0
ENSG00000209821.2	0
ENSG00000205788.1	0
ENSG00000204118.11	957330.256
ENSG00000206089.01	3448.02773
ENSG00000217317.17	4454.5667
ENSG00000207902.7	22977.4208
ENSG00000214943.12	2082.43565
ENSG00000214042.13	510.5246749
ENSG00000205842.1	1593.06132
ENSG00000217165.1	0
ENSG00000214881.1	0
ENSG00000206040.1	394.4755669
ENSG0000021105.1	1583.312582
ENSG00000204044.1	0
ENSG000002182145.8	455.380648
ENSG00000208416.4	1184.087054

Genes	Annotation
ENSEG0000024268.2	2566.464279
ENSEG0000021071.3	460.234343
ENSEG000002675.8	52440.30096
ENSEG000002782.7	6835.46466
ENSEG000002827.5	25599.251
ENSEG000002527.4	0
ENSEG000002668.3	154.947378
ENSEG000002842.12	490658.658
ENSEG0000025883.1	6148.99052
ENSEG0000023168.3	0
ENSEG000002675.2	0
ENSEG0000021788.1	0
ENSEG0000013408.11	957330.265
ENSEG0000026308.1	3484.02773
ENSEG0000027217.17	226717.4008
ENSEG000002700.7	226717.4008
ENSEG0000023494.3	2082.24051
ENSEG000002404.3	130.5246749
ENSEG0000023419.9	153683.9216
ENSEG0000027616.1	0
ENSEG0000023488.1	0
ENSEG0000023040.1	394.4755669
ENSEG0000023165.1	5383.32352
ENSEG000002844.8	45338.60648
ENSEG0000029216.8	118.087454

Genotype	Count	Proportion
ENSG00000242268	2658	4641.79
ENSG00000270123	460	2343.43
ENSG00000267115	5240	30096
ENSG00000278215	68365	46456
ENSG00000282175	255959	2361
ENSG00000252574	0	0
ENSG00000268488	514	9473.78
ENSG00000280422	409656	68
ENSG00000598831	6108	399952
ENSG00000319613	0	0
ENSG00000405752	2	0
ENSG00000217881	0	0
ENSG00000340811	95730	2056
ENSG00000263021	3484	2273.43
ENSG00000271217	424	4085.43
ENSG00000267007	225737	4208
ENSG00000234943	2082	2405
ENSG00000404313	310	5246.79
ENSG00000281119	513683	3252
ENSG00000271666	0	0
ENSG00000244881	0	0
ENSG00000260401	394	47556.69
ENSG00000261316	3983	33252
ENSG00000240044	0	0
ENSG00000282146	45358	4068
ENSG00000259446	119	10874.04

[illegible]

Genotype	Approximation
ENTG0000024268.2	15626.464719
ENTG0000027123.18	5460.234343
ENTG0000026075.15	12420.10096
ENTG0000028182.2	0
ENTG0000028217.5	68355.46456
ENTG0000046018.30	215999.2811
ENTG0000025275.4	0
ENTG0000028488.13	154.9673798
ENTG0000028942.12	49865.56
ENTG0000025883.1	6108.99252
ENTG0000023198.13	0
ENTG0000025672.2	0
ENTG0000021788.1	0
ENTG0000023408.11	957330.2656
ENTG0000030081.1	3484.027313
ENTG0000027217.47	414.485
ENTG0000027800.3	22571.47048
ENTG0000024943.2	2082.24051
ENTG0000024042.3	130.5246749
ENTG0000024112.9	15186.13216
ENTG0000027165.1	0
ENTG0000024881.1	0
ENTG0000030401.4	394.4755639
ENTG0000028105.1	583.312582
ENTG0000030044.1	0
ENTG0000028241.8	45738.50648
ENTG0000029546.6	119.087054

ENSG00000242568.2	Exp	Expression
ENSG00000242568.2	1658	464179
ENSG00000270123.3	460	234343
ENSG00000275718.1	2544	10096
ENSG00000276235.5	68165	46656
ENSG00000268307.0	205919	2351
ENSG00000252925.4	0	0
ENSG00000256688.12	94	9473768
ENSG00000258042.12	4098565	58
ENSG00000259883.1	6308	9900562
ENSG00000231983.0	0	0
ENSG00000256525.1	0	0
ENSG000002501788.1	0	0
ENSG00000234308.11	95730	2656
ENSG00000236009.1	3844	207733
ENSG00000272145.1	485	9507
ENSG00000250770.7	23571	0
ENSG00000234943.2	2082	2405
ENSG00000240423.3	130	524679
ENSG00000234432.9	15863	921
ENSG00000275566.1	0	0
ENSG00000234881.1	0	0
ENSG00000236040.1	394	475638
ENSG00000238550.1	3833	112582
ENSG00000230044.1	0	0
ENSG00000258241.6	45338	60548
ENSG00000269456.4	129	1807504

ENSG00000242568.2	Expression
ENSG00000242568.2	1658.46479
ENSG00000270112.3	462.234343
ENSG00000270112.3	25404.10096
ENSG00000270112.3	0
ENSG00000270112.3	68365.4656
ENSG00000270112.3	259599.2151
ENSG00000252575.4	0
ENSG00000252575.4	124.9473718
ENSG00000252575.4	409865.56
ENSG00000259883.1	6198.99052
ENSG00000259883.1	0
ENSG00000259883.1	0
ENSG00000259883.1	957330.256
ENSG00000260019.1	3484.02793
ENSG00000260019.1	14485.407
ENSG00000267700.7	23971.4708
ENSG00000269423.2	2082.2405
ENSG00000269423.2	130.5246749
ENSG00000269423.2	5186.912
ENSG00000275516.1	0
ENSG00000275516.1	0
ENSG00000275516.1	0
ENSG00000275516.1	394.45765
ENSG00000275516.1	3833.312582
ENSG00000284044.1	0
ENSG00000284148.6	45338.60648
ENSG00000284148.6	1219.087654

# Kidney Cancer

Genes	Expression
ENSG0000024268.2	1566.464179
ENSG00000210112.3	460.234343
ENSG0000014578.15	52440.0000
ENSG0000027384.1	0
ENSG0000014578.15	68165.45426
ENSG00000140831.10	25959.8121
ENSG0000022578.4	0
ENSG0000014548.1	194.937768
ENSG0000021212.12	4948465.618
ENSG0000025988.1	2190.699952
ENSG0000025988.1	0
ENSG0000028475.2	0
ENSG0000027801.1	0
ENSG0000013408.11	957330.2056
ENSG0000013089.1	3484.027373
ENSG0000017137.17	14548.9507
ENSG0000017008.7	28671.42836
ENSG0000014948.1	2922.34015
ENSG0000024042.1	130.5246789
ENSG0000020642.9	155816.9126
ENSG0000027616.1	0
ENSG0000014548.1	0
ENSG0000023040.1	394.4755669
ENSG00000231105.1	183.21852
ENSG0000024042.1	0
ENSG0000014141.8	455.38.405648

Genes	Expression
ENSG00000242068.2	1655.464179
ENSG00000270112.3	460.334343
ENSG00000647678.5	52440.300095
ENSG00000273842.1	0
ENSG00000273842.1	6185.465262
ENSG00000146083.1	25599.251391
ENSG00000252375.4	0
ENSG00000154846.12	104.947378
ENSG00000241432.1	46836.581676
ENSG00000259883.1	6208.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG00000207988.1	0
ENSG00000134408.11	95730.2056
ENSG00000263039.1	3484.027377
ENSG00000171317.1	41485.9507
ENSG00000697003.7	2267.742088
ENSG00000274422.1	286.140005
ENSG00000130422.1	130.524769
ENSG00000266482.9	155861.9216
ENSG00000274636.1	0
ENSG00000260682.1	0
ENSG00000230401.1	394.475569
ENSG00000231105.1	15831.31282
ENSG00000243044.1	0
ENSG00000252141.8	45538.40545
ENSG00000252141.8	110.082054

Genes	Expression
ENSG00000242682	1658.464179
ENSG00000100713	406.234343
ENSG00000167518	5240.100126
ENSG00000173842	1
ENSG00000167518	68816.452626
ENSG000001408310	25599.2181
ENSG00000125724	5
ENSG00001584612	104.947378
ENSG0000012312	276.676.688
ENSG00000198811	6308.998991
ENSG000001319813	1
ENSG00000164975	1
ENSG00000127886	1
ENSG00000131018	97330.2056
ENSG000001630891	3484.027031
ENSG00000172137	14185.9507
ENSG00000176027	22671.42470
ENSG00000148811	2292.24505
ENSG00000140031	130.536749
ENSG000001606429	15086.15216
ENSG000001718616	1
ENSG00000148811	1
ENSG000001360401	394.475669
ENSG000001312051	11058.131282
ENSG00000144444	1
ENSG000001821448	45336.02648
ENSG000001409448	10.882676

Genes	Expression
ENSG0000024268.2	1055.484179
ENSG0000020112.3	406.234343
ENSG00000275718.5	52440.10093
ENSG0000027894.1	0
ENSG0000027893.1	18315.45262
ENSG00000246031.8	25599.2911
ENSG00000252745.2	0
ENSG00000215886.4	104.9471768
ENSG00000214212.2	166.926.458
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000269475.2	0
ENSG0000022705.6	0
ENSG00000234108.11	957330.2056
ENSG00000263891.8	3484.027737
ENSG00000271387.1	141485.907
ENSG00000266703.7	226747.548
ENSG00000234492.2	246035.2
ENSG00000240423.4	310.342639
ENSG00000266042.9	155863.1922
ENSG00000278163.6	0
ENSG00000234481.1	0
ENSG00000231050.1	394.4755669
ENSG00000231001.5	1583.312582
ENSG00000243044.1	0
ENSG00000282144.1	40.925645
ENSG00000249849.1	110.084764

ENS000002042288.2	1698	484.73
ENS000002042289.1	465	284.93
ENS000002042290.1	1535	5244.00
ENS000002042292.1	0	0
ENS000002042307.5	6885	456.26
ENS000002044088.10	25599	2351
ENS000002044090.12	0	0
ENS000002044148.2	104	947.978
ENS000002044182.12	409656	58
ENS000002044883.1	68	999052
ENS000002044900.1	0	0
ENS000002044975.2	0	0
ENS000002045188.1	0	0
ENS000002045188.11	95730	2056
ENS000002045188.12	348	67379
ENS000002047137.17	43485	527
ENS000002047700.7	226717	420
ENS000002048493.2	2802	245035
ENS000002049023.1	310	5246749
ENS00000204942.9	15563	8276
ENS000002049648.1	0	0
ENS000002049801.1	0	0
ENS000002049801.1	394	475669
ENS0000021105.11	158	312582
ENS00000213484.1	0	0
ENS00000212114.8	45338	60648
ENS00000209416.4	0	0

ENS000000242628.2	1638.464618
ENS000000242629.2	406.2345433
ENS000000242630.2	12046.30096
ENS000000273842.1	0
ENS000000782327.5	68185.45626
ENS000000440638.10	25759.2531
ENS000000158848.12	0
ENS000000158486.12	304.9471268
ENS000000198042.12	46986556.12
ENS000000259883.1	0.18.999052
ENS000000231988.3	0
ENS000000269475.2	0
ENS000000201788.1	0
ENS000000134018.11	95730.2036
ENS000000262089.1	9430.427373
ENS000000272717.17	426.85.2857
ENS000000264700.7	22817.24028
ENS000000234943.2	2808.240355
ENS000000240423.1	10.5246748
ENS000000265642.9	155863.1823
ENS000000271616.1	0
ENS000000234881.1	0
ENS000000238604.1	394.475669
ENS000000231105.1	1531.21282
ENS000000243944.1	0
ENS000000182141.8	453338.6048
ENS000000269164.6	119.0847054

ENSG00000242818	1608.4494129
ENSG00000270112	440.7340463
ENSG00000293435	52440.10006
ENSG0000027842.1	0
ENSG00000278275	68165.46265
ENSG00000240108	20599.95123
ENSG00000275164	0
ENSG00000250488.12	104.9473788
ENSG00000280422.12	48965.61658
ENSG00000258881.1	6108.1909052
ENSG00000280422.1	0
ENSG00000280475.2	0
ENSG00000287888.1	0
ENSG00000234108.11	957330.26610
ENSG00000268018	0
ENSG00000250818	41485.46527
ENSG00000257700	22872.42740
ENSG00000234943.2	2082.145035
ENSG00000240435.1	301.5267499
ENSG00000240109.9	150383.92126
ENSG00000271635.1	0
ENSG00000234881.1	0
ENSG00000238040.1	394.4755669
ENSG00000230145	1593.112542
ENSG00000243044	0
ENSG00000232148.8	4535.40836
ENSG00000298436	119.0847054

ENSG00000204262.2	10	6846.44179
ENSG00000204263.2	10	20.245433
ENSG00000204264.2	10	127.1738
ENSG00000204265.2	10	5244.10096
ENSG0000020782.1	0	
ENSG0000020782.5	68	4616.4626
ENSG0000020803.0	10	25599.2351
ENSG0000020803.4	10	25599.2351
ENSG0000020846.1	10	104.947378
ENSG0000020842.12	10	49085.6158
ENSG0000020988.1	6	1018.99592
ENSG0000020988.2	6	1018.99592
ENSG00000209475.2	0	
ENSG00000210788.1	0	
ENSG00000214308.11	10	97530.20616
ENSG0000021569.1	10	348.02773
ENSG0000021569.2	17	1.485.46527
ENSG00000216700.7	7	22671.74408
ENSG00000218483.2	10	2802.45035
ENSG00000240433.1	10	310.544261
ENSG0000024043.2	10	15586.9216
ENSG00000271636.1	0	
ENSG00000274881.1	0	
ENSG00000276901.1	39	494.47659
ENSG0000028115.1	15	1583.11282
ENSG0000028304.1	10	10.000000
ENSG0000028321.8	48	4538.46048
ENSG00000295416.6	19	119.87048

# Merged Sample Expression Data

Genes

SAMPLES

	0	1	2	3	4	5	6	7	8	9	...	60474	60475	60476	60477	60478	60479	60480	60481	60482	submitter_id
0	574548	2263.14	983212	69718	54834.9	19718.1	175853	735123	38662.4	233190	...	0	0	0	0	0	0	0	0	0	TCGA-04-1331-01A-01R-1569-13
1	352295	4592.37	663107	39745.4	36553.5	41147.1	241313	396423	37567	128693	...	0	0	0	0	0	0	0	0	0	TCGA-04-1332-01A-01R-1564-13
2	295162	649.026	1.21115e+06	57385.5	33097.4	58051.8	228615	346066	105567	408267	...	0	0	0	0	0	0	0	0	0	TCGA-04-1338-01A-01R-1564-13
3	329580	1835.59	1.08437e+06	33812.3	24516.1	22330.6	42134.4	895558	56178	83847.3	...	0	0	0	0	0	0	0	0	0	TCGA-04-1341-01A-01R-1564-13
4	289269	40061.7	2.44837e+06	26399.5	18248	49610	74761.1	571992	71951.9	98726.4	...	0	0	0	0	0	0	0	0	0	TCGA-04-1343-01A-01R-1564-13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4495	1.18093e+06	0	1.01139e+06	67877.2	15005.7	50527.3	6.21536e+06	1.47373e+06	459656	167488	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CD-01A-11R-A37K-07
4496	929228	0	869800	95607.5	17188.6	9352.12	7.61121e+06	196838	354465	138074	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CE-01A-11R-A37K-07
4497	469276	476.683	516938	110051	34469.4	37334.7	5.95811e+06	427832	323833	154861	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CF-01A-11R-A38B-07
4498	2.44119e+06	18282.7	853547	79288.7	106926	42593.9	4.80111e+06	955338	331924	177020	...	0	0	0	0	0	0	0	0	0	TCGA-ZS-A9CG-01A-11R-A37K-07
4499	259853	505.488	591328	74253.7	42553.5	118772	148978	508465	153862	170412	...	0	0	0	0	0	0	0	0	0	TCGA-ZX-AA5X-01A-11R-A42T-07

4500 rows × 60484 columns

Transpose and  
add as a row

Genes	Expression
ENSG0000024298.2	3038.484179
ENSG00000276112.3	480.734143
ENSG0000026978.15	52440.1006
ENSG0000027840.1	0
ENSG0000028121.1	6885.4526
ENSG0000024293.10	25099.2351
ENSG0000025271.4	0
ENSG0000025486.12	104.947378
ENSG0000021842.12	484856.458
ENSG0000021881.1	618.19052
ENSG0000021881.3	0
ENSG0000028471.2	0
ENSG0000026788.1	0
ENSG0000023428.11	90730.2056
ENSG0000023428.1	2484.0373
ENSG0000027217.17	41485.9507
ENSG00000257780.7	22672.4208
ENSG0000023484.2	2882.24035
ENSG0000024042.1	305.5246749
ENSG0000028044.9	121863.1216
ENSG00000271816.1	0
ENSG00000214881.1	0
ENSG00000238041.1	384.475669
ENSG00000231101.1	1583.112582
ENSG0000024044.1	0
ENSG0000021411.8	45338.40648
ENSG00000289416.4	119.0847054
ENSG0000025481.1	0

# Quantifying mRNA abundance and Scaling

- GDC harmonization data is provided in FPKM-UQ
- In our code, FPKM-UQ is rescaled to TPM using the following formula.

$$\text{TPM}_i = \left( \frac{\text{FPKM}_i}{\sum_j \text{FPKM}_j} \right) \cdot 10^6$$

- TPM has nice mathematical properties and a stable entity

<https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/>

Mapping and quantifying mammalian transcriptomes  
by RNA-Seq

Ali Mortazavi<sup>1,2</sup>, Brian A Williams<sup>1,2</sup>, Kenneth McCue<sup>1</sup>, Lorian Schaeffer<sup>1</sup> & Barbara Wold<sup>1</sup>

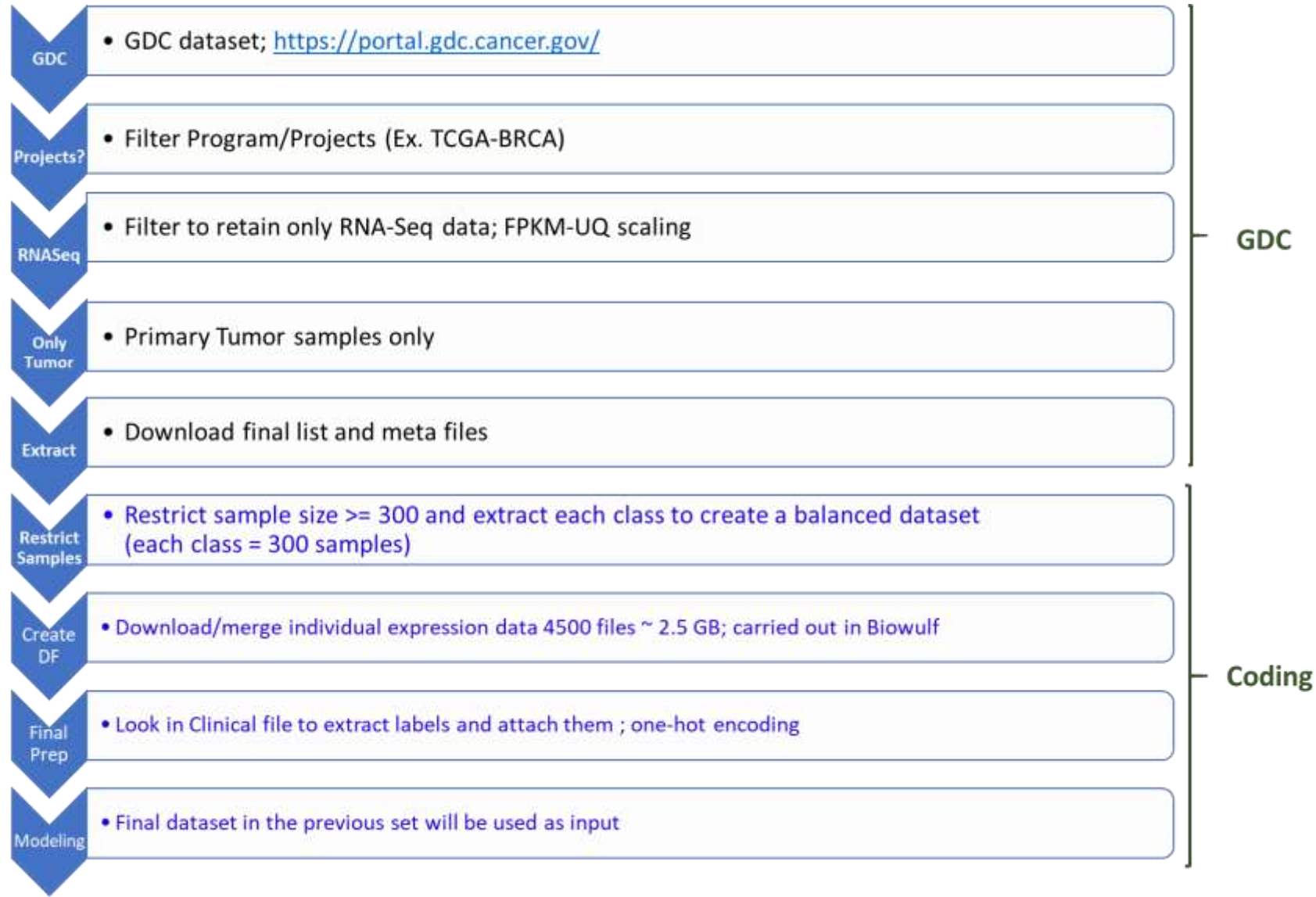
# One-hot encoding to convert Cancer types to numbers

- Convenient to transform categorical variables into a numerical quantity for computations
  - BRCA to 0 ; LUAD to 1 etc.
  - 0, 1, 2, 3, ..., 13, 14

TCGA-CESC  
TCGA-LIHC  
TCGA-STAD  
TCGA-OV  
TCGA-BLCA  
TCGA-THCA  
TCGA-PRAD  
TCGA-COAD  
TCGA-KIRC  
TCGA-LUSC  
TCGA-HNSC  
TCGA-LGG  
TCGA-LUAD  
TCGA-UCEC  
TCGA-BRCA

```
>>> encoded
array([[1., 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., 0., 0., 0., 0., 0., 0.],
       [0., 0., 1., 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., 0., 0., 0., 0.],
       [0., 0., 0., 0., 1., 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., 0., 0.],
       [0., 0., 0., 0., 0., 0., 1., 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.],
       [0., 0., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 0., 0., 0., 0., 0., 1., 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., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.]],
      dtype=float32)
```

# Data preparation steps summary





## Before we break for hands-on

- **Python as the programming language for this workshop, but similar libraries are available in R or other languages**



- **Will use Jupyter Notebook for sharing the code**
  - With little effort one can convert the Python code into R and still use Jupyter Notebook

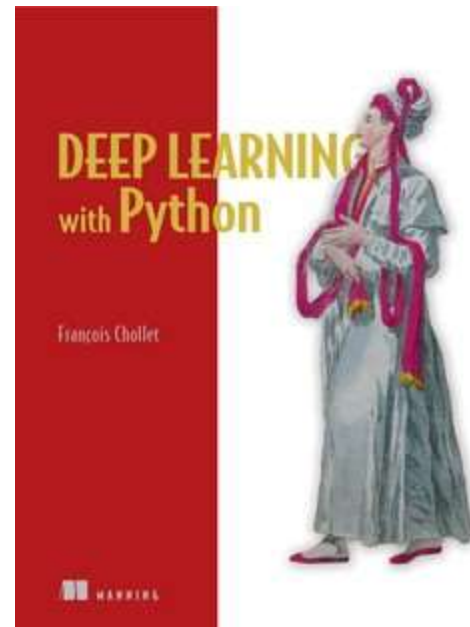
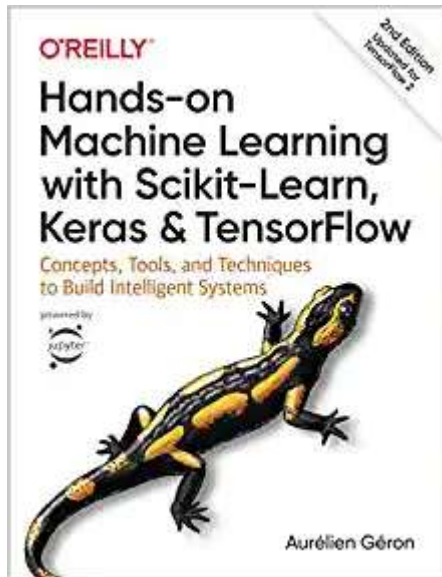
## To be continued after hands-on

---

<https://github.com/ravichas/ML-TC1>

## Before we begin the modeling section ...

- Due to lack of time, I won't be covering the basics of Neural Network



*Keras is a high-level NN package that is built on top of popular high-level libraries (TF, Theano). Works well with CPU/GPU*



These are good books for beginners and up

Figure from Deep Learning with Python

# Supervised Learning

- Goal
  - Construct a model that takes in input features/target pair to return a prediction for target/outcome
- Train a machine learning
  - Model refers to learning its parameters, which typically involves minimizing a loss function on training data with the aim of making accurate predictions on unseen (test) data

## Supervised Learning:

Data:  $(x,y)$  ; where  $x$  is the genomic expression profile ;  $y$  is the cancer classes

Goal? Learn the function that maps  
 $x \rightarrow y$

# Terminology

	0	1	2	3	4	5	6	7	8	9	...	60474	60475	60476	60477	60478	60479	60480	60481	60482	submitter_id
0	574548	2263.14	983212	69718	54834.9	19718.1	175853	735123	38662.4	233190	...	0	0	0	0	0	0	0	0	0	TCGA-04-1331-01A-01R-1569-13
1	352295	4592.37	663107	39745.4	36553.5	41147.1	241313	396423	37567	128693	...	0	0	0	0	0	0	0	0	0	TCGA-04-1332-01A-01R-1564-13
2	295162	649.026	1.21115e+06	57385.5	33097.4	58051.8	228615	346066	105567	408267	...	0	0	0	0	0	0	0	0	0	TCGA-04-1338-01A-01R-1564-13
3	329580	1835.59	1.08437e+06	33812.3	24516.1	22330.6	42134.4	895558	56178	83847.3	...	0	0	0	0	0	0	0	0	0	TCGA-04-1341-01A-01R-1564-13
4	289269	40061.7	2.44837e+06	26399.5	18248	49610	74761.1	571992	71951.9	98726.4	...	0	0	0	0	0	0	0	0	0	TCGA-04-1343-01A-01R-1564-13

- **Columns**
  - input variables or features or attributes
- **Outcome column**
  - Outcome variables or targets
- **Rows**
  - Training example or instance
- **Whole table Training data set**

# What is different about Neural Network?

- If you know the equation (algorithm), then you feed in the **input** and you get the **output**.  
You can code the function yourself

```
def function(x):  
    y = 2.0 + 5.0 * x  
    return(y)
```

- You can choose to use linear modeling and use the data to figure the relationship

```
Model ← lm( y ~ x)
```

- Neural Network using the data learn the algorithm.

**INPUT**

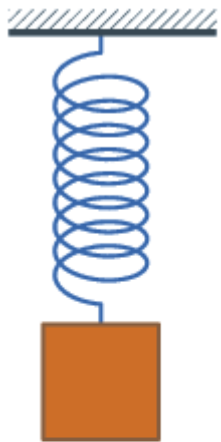
**ALGORITHM**

**OUTPUT**

# A Simple Network

**Input: Mass or  $M$  (kg)**

**Output: Length or  $L$  (m)**



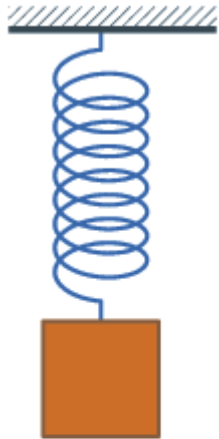
M	L
Input	Output
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	???

[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Based on Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003



# A Simple Network



M	L
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	0.68

$$L = 0.1 * Mass + 0.38$$

[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

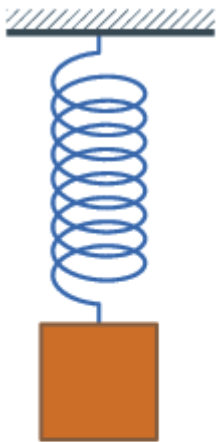
Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003

# A Simple Network

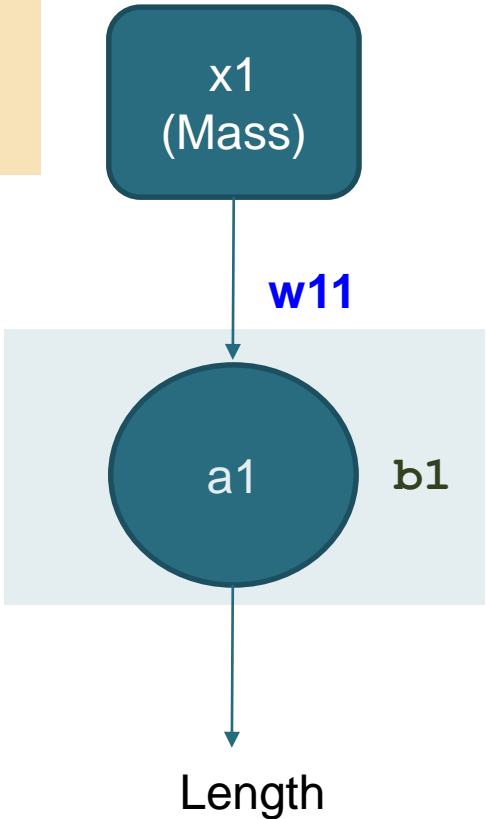
$$a1 = x1 * w11 + b1$$

$$L = M * 0.1 + 0.38$$

[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)



Hidden Layer



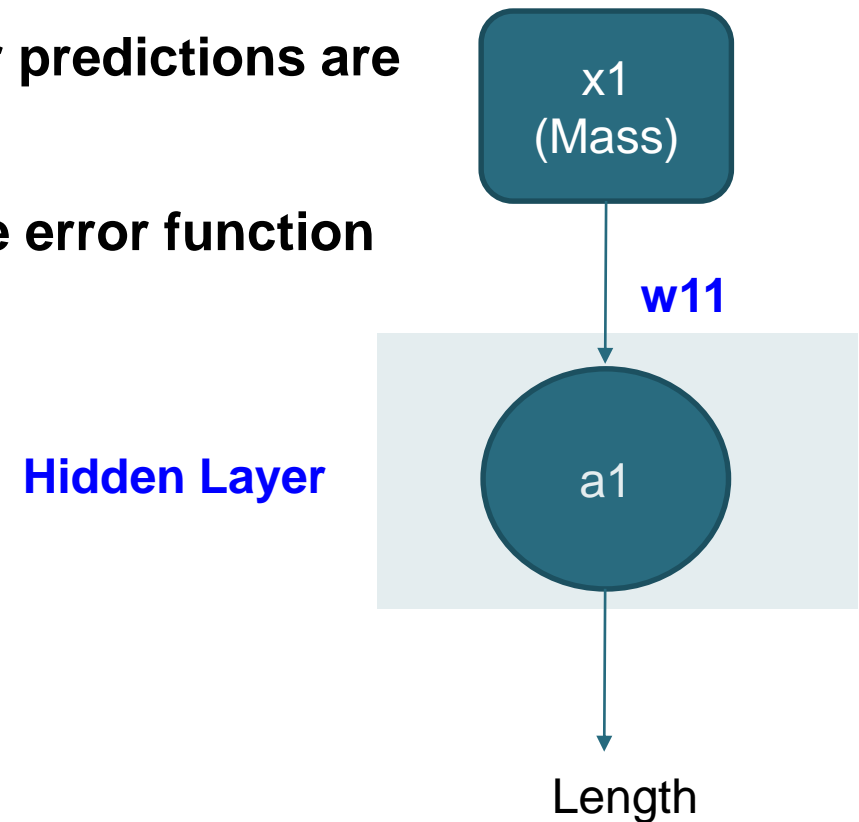
M	L
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	0.68

These are the model variables: `[array([[0.10058284]], dtype=float32), array([0.37793916], dtype=float32)]`

Based on Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003

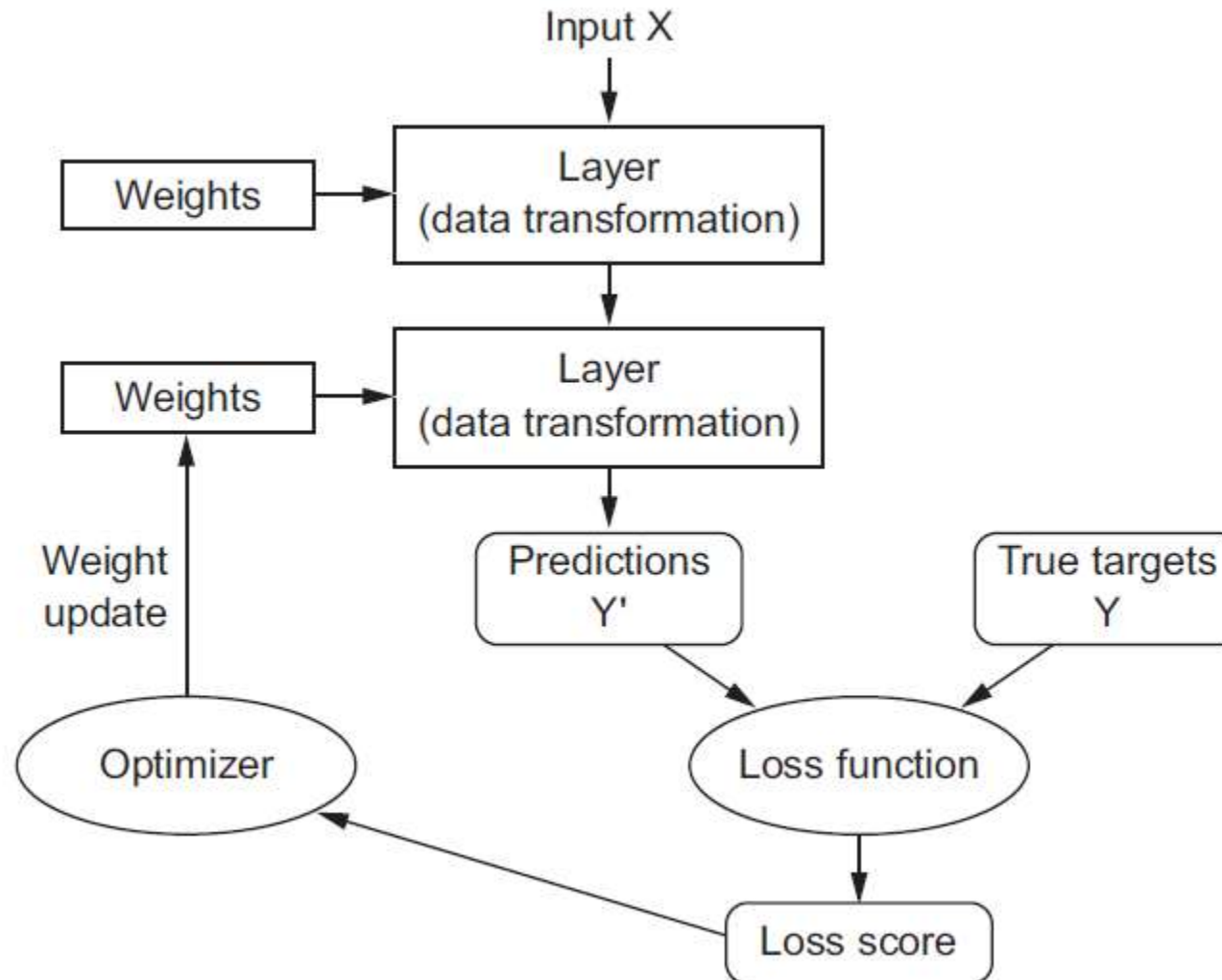
# Error minimization

- Goal is to choose  $W$ s such that predictions of the network should be close to  $y$
- Error function or cost function a measure how good our predictions are
- Eventually, we want to pick a set of  $w$  that minimizes the error function



# Deep Learning Procedure

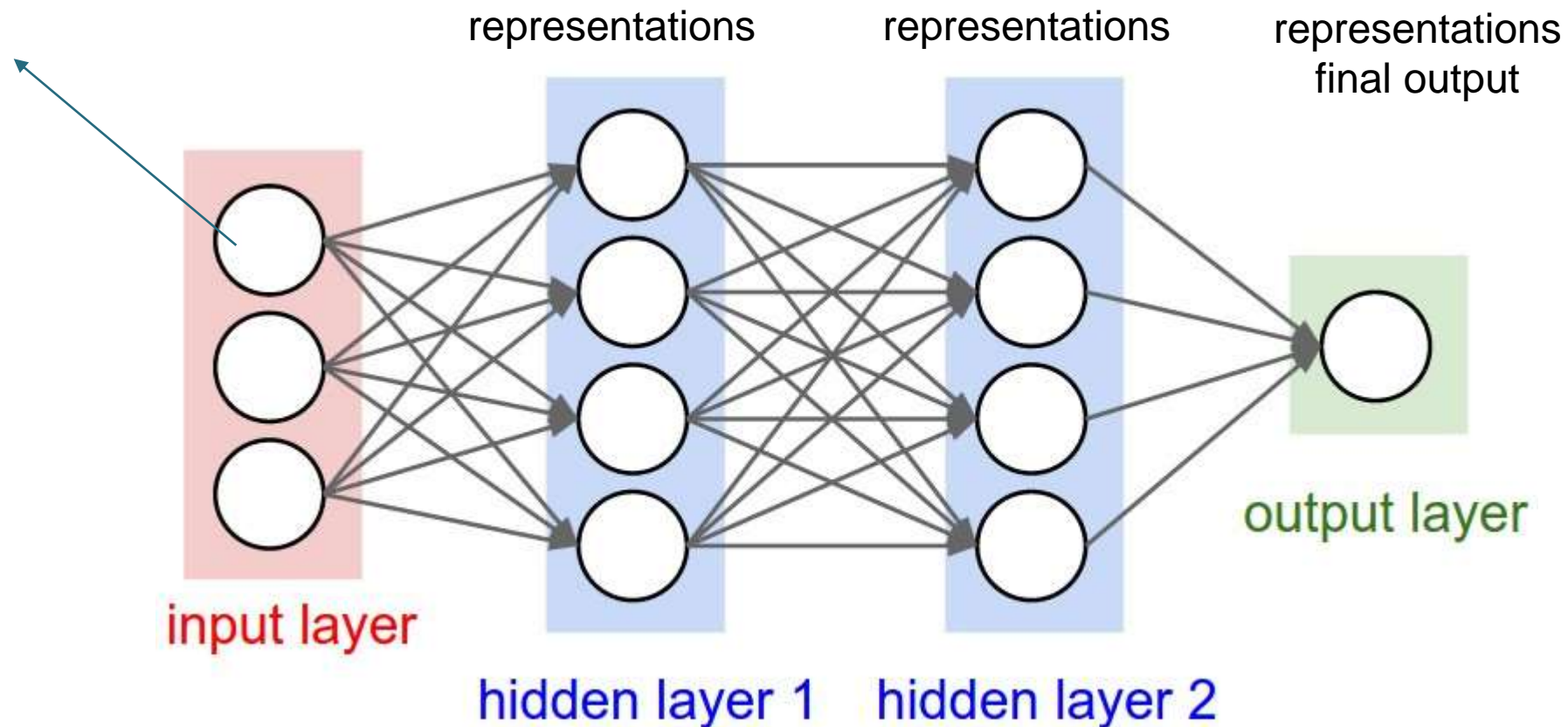
Taken from Deep Learning with Keras book



# Vanilla network

Each neuron receives input from all the neurons in the previous layer (densely connected)

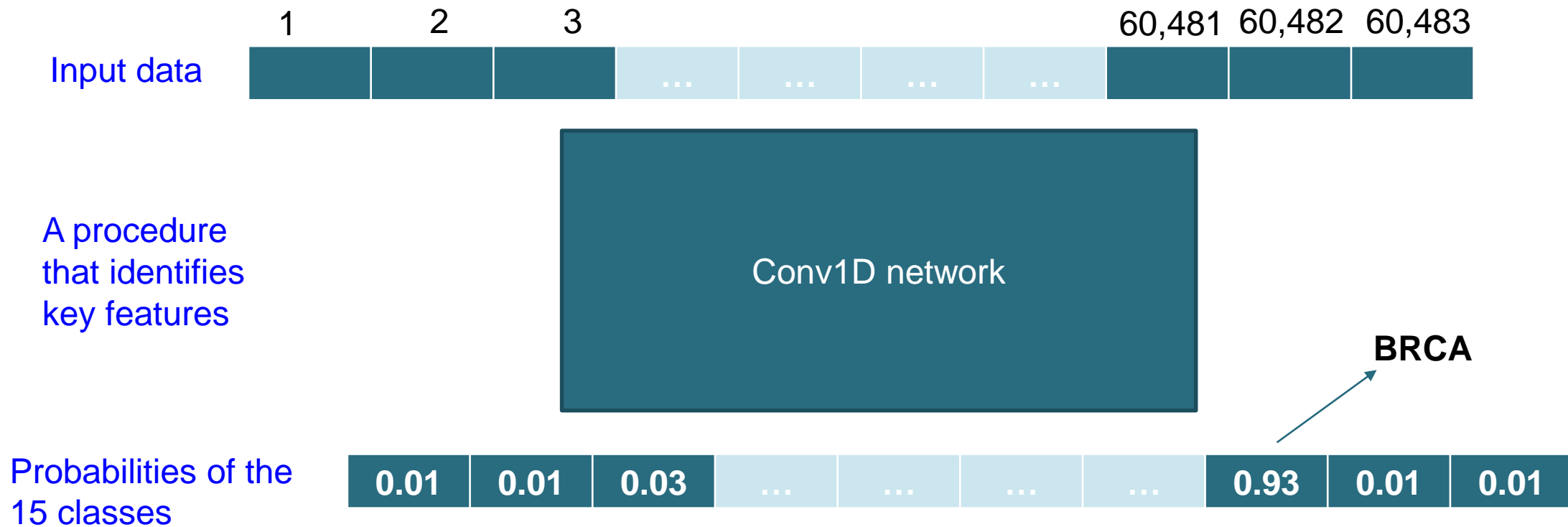
Neuron: a unit that holds a number



[This Photo](#) by Unknown Author is licensed under [CC BY-NC](#)

# Convolutional Neural Network

- We are going to take a vector of genomic expression values and feed them into a network with a series of operations to create a model
- Model is what we call convolutional-1D network

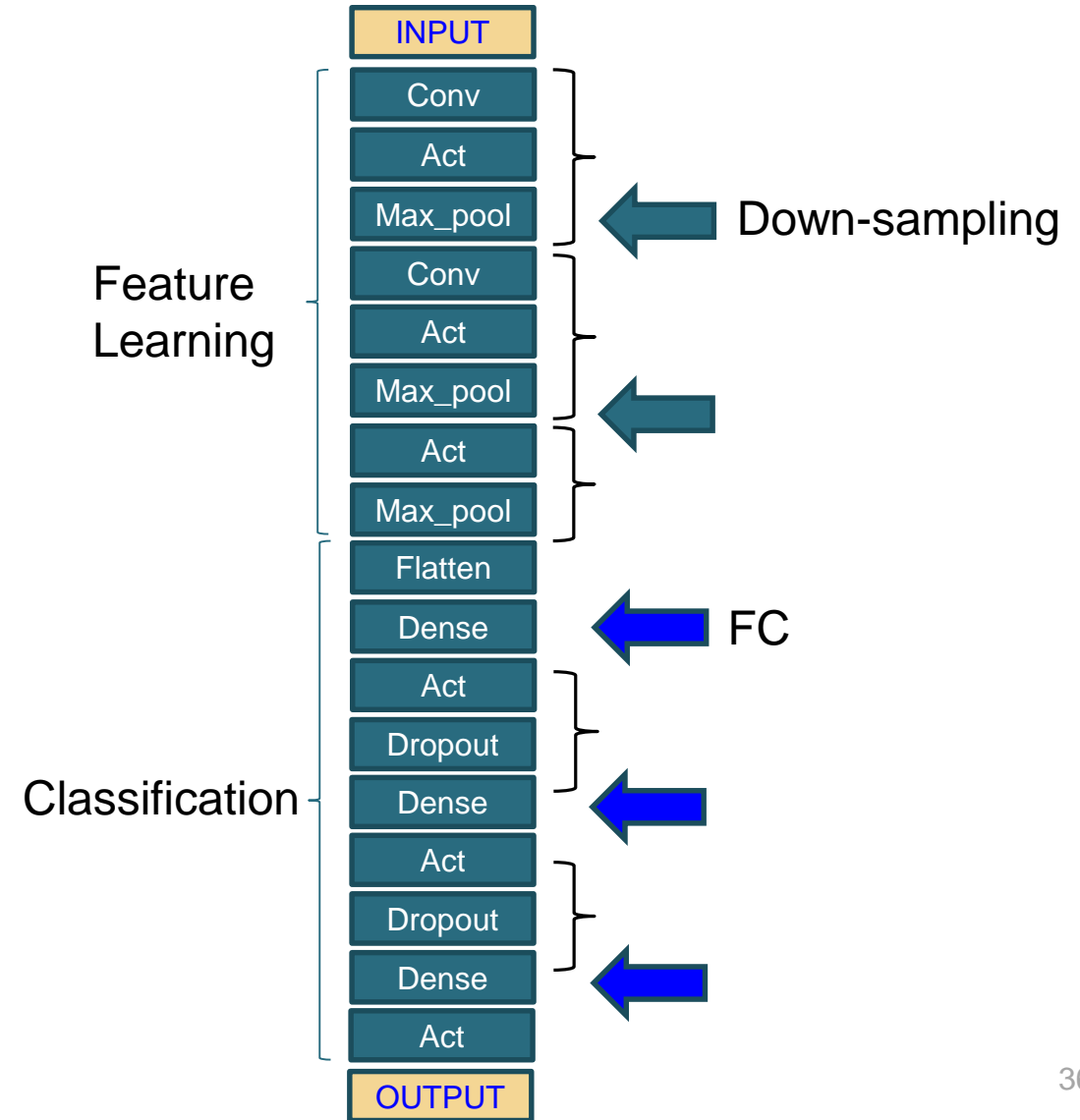


# Components of conv1D

1. **Act: Activation**
2. **Conv: Convolution**
3. **Max\_pool: Maxpooling**
4. **Flatten**
5. **Dense**
6. **Dropout**

Topology of a network defines a “hypothesis space”

Choosing a specific topology is usually not straightforward and comes with practice.





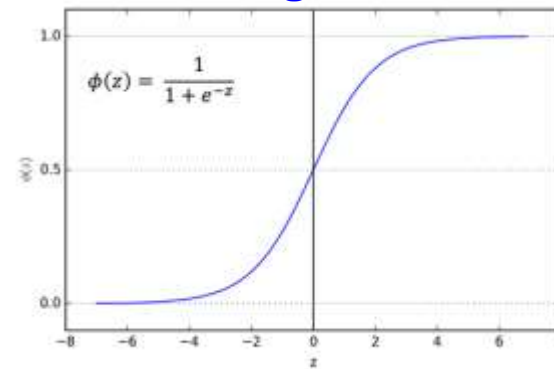
# 1. Activation Function

- Activation functions are included to create non-linearity

[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

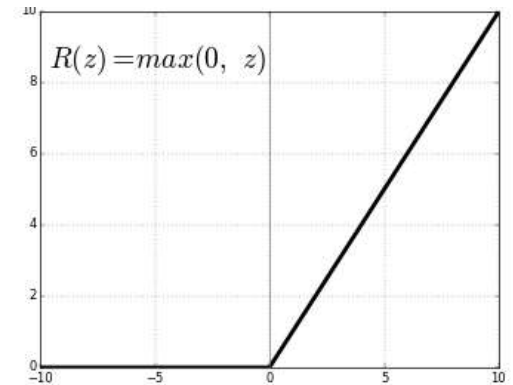
- Sigmoid
- ReLU
- Leaky ReLU
- ELU
- Maxout
- Tanh

Sigmoid

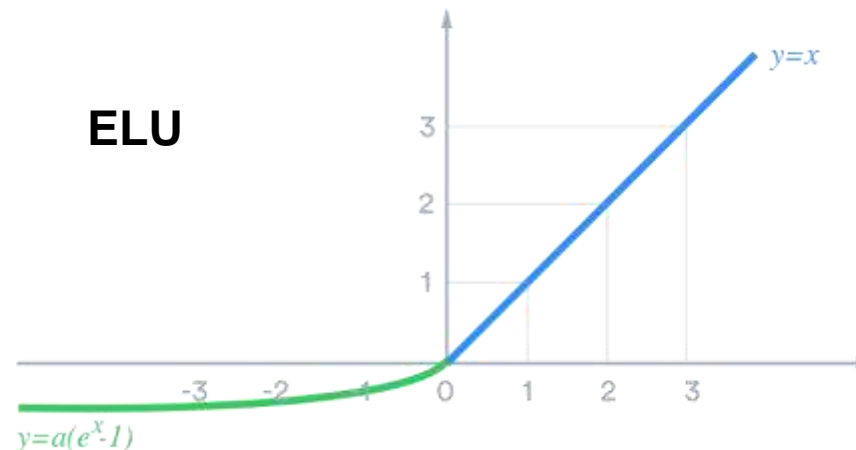


Squashes the #s to [0, 1]

ReLU

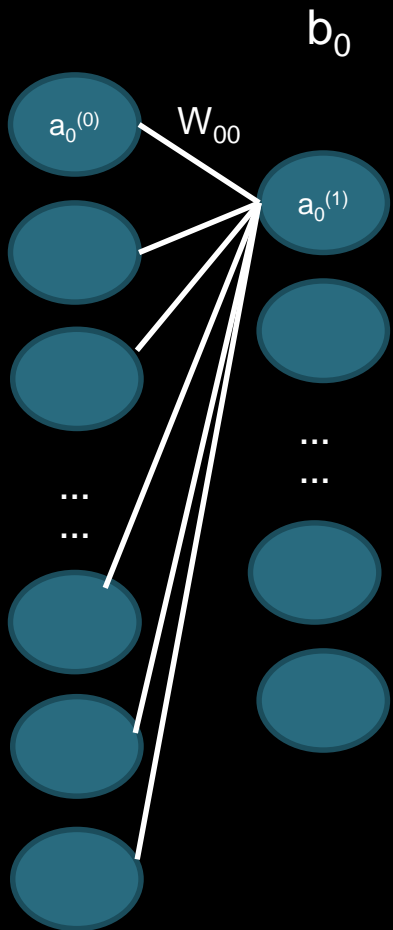


ELU



# 1. Activation function

$$a^{(L)} = \text{ReLU}(w^{(L)}a^{(L-1)} - b^{(L)})$$



ReLU

$$\begin{bmatrix} W_{0,0} & W_{0,1} & \dots & W_{0,n} \\ W_{1,0} & W_{1,1} & \dots & W_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{k,0} & W_{k,1} & \dots & W_{k,n} \end{bmatrix}$$

$$\begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix}$$

+

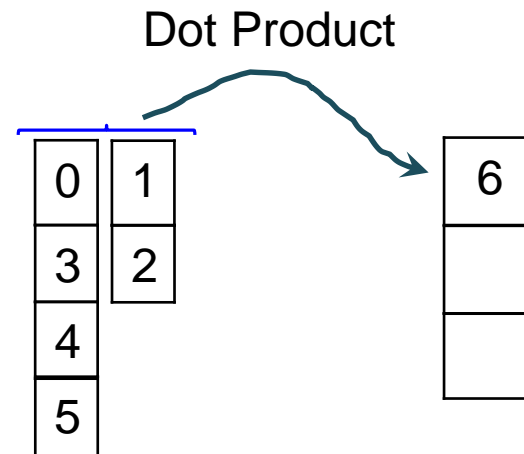
$$\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

$$a_0^{(1)} = \text{ReLU}(W_{00}a_0^{(0)} + W_{0,1}a_1^{(0)} + \dots + W_{0,n}a_n^{(0)} - b_0)$$

## 2. Convolution

Process of applying filter (kernel) to the data for the purpose of subsampling. Kernel is a matrix that has a smaller dimension than the input data creates chunks

Reduces the number of parameters and allow creation of deeper networks

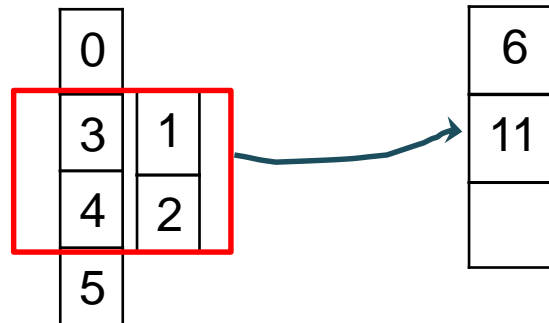


## 2. Convolution

Process of applying filter (kernel) to the data for the purpose of subsampling. Kernel is a matrix that has a smaller dimension than the input data creates chunks

Reduces the number of parameters and allow creation of deeper networks

Dot Product

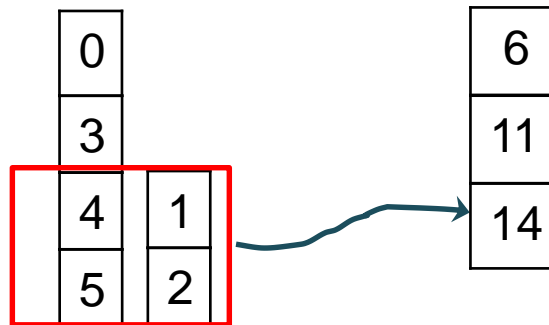


## 2. Convolution

Process of applying filter (kernel) to the data for the purpose of subsampling. Kernel is a matrix that has a smaller dimension than the input data creates chunks

Reduces the number of parameters and allow creation of deeper networks

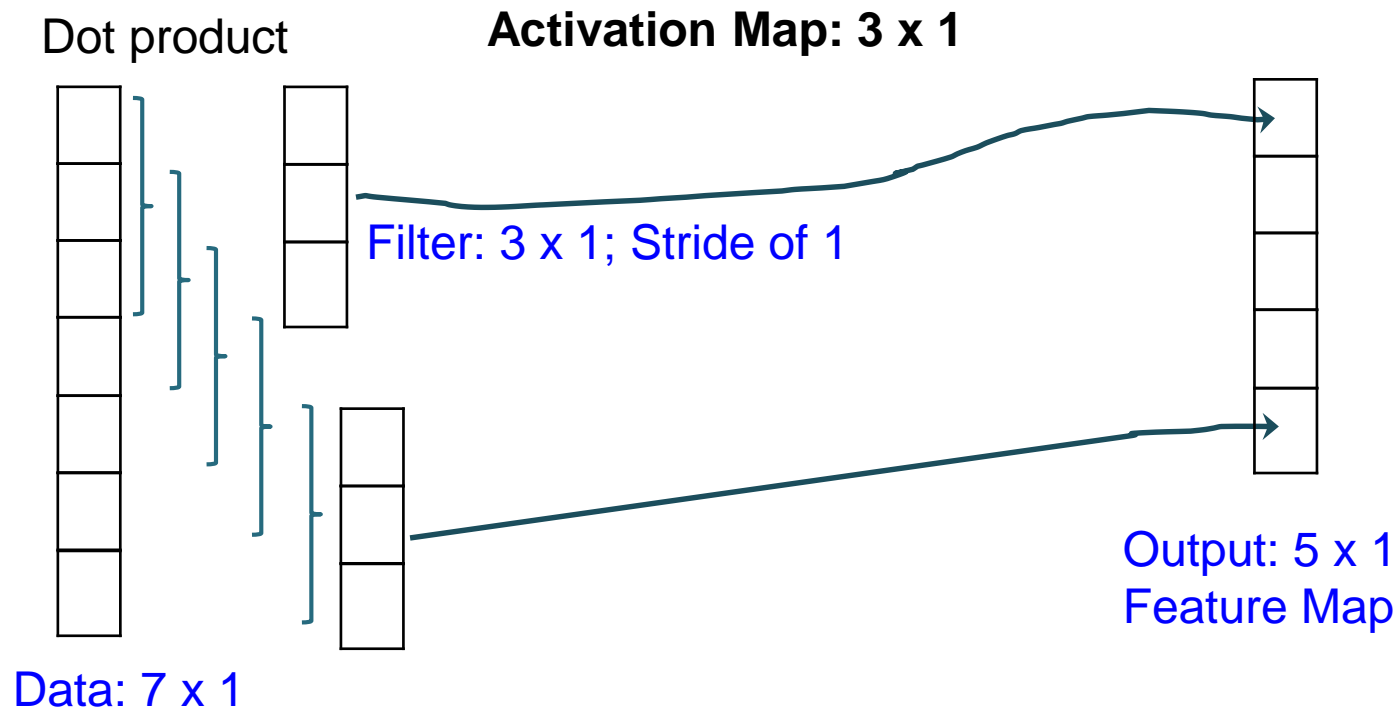
Dot Product



## 2. Convolution

Process of applying filter (kernel) to the data for the purpose of subsampling. Kernel is a matrix that has a smaller dimension than the input data creates chunks

Reduces the number of parameters and allow creation of deeper networks



$(N-F)/\text{stride}+1$  will be the size after filtering

$(7-3)/1+1 = 5$  ;  
zero padding on the border

## 2. Convolution

- **Convolution Layer**
  - Hyperparameters
    - Number of filters
    - Spatial extent
    - Stride
    - Amount of zero padding

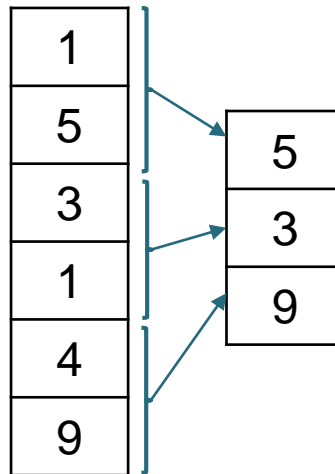
Andrew, one of my colleague, expert in CANDLE, can help you with this.



andrew.weisman@nih.gov

### 3. Pooling

- Pooling makes the representations smaller/manageable (downsampling) by retaining only important features; creates smaller clusters of manageable size
- Each activation map will be pooled separately.
- Common approach is Max Pooling



Max-pooling  
with filter size  
of 2x1 and  
stride of 2

#### Max Pooling Intuition:

Enhancing the signals by looking at a region and pick the maximum activation value

Each of these are activation and we are looking for

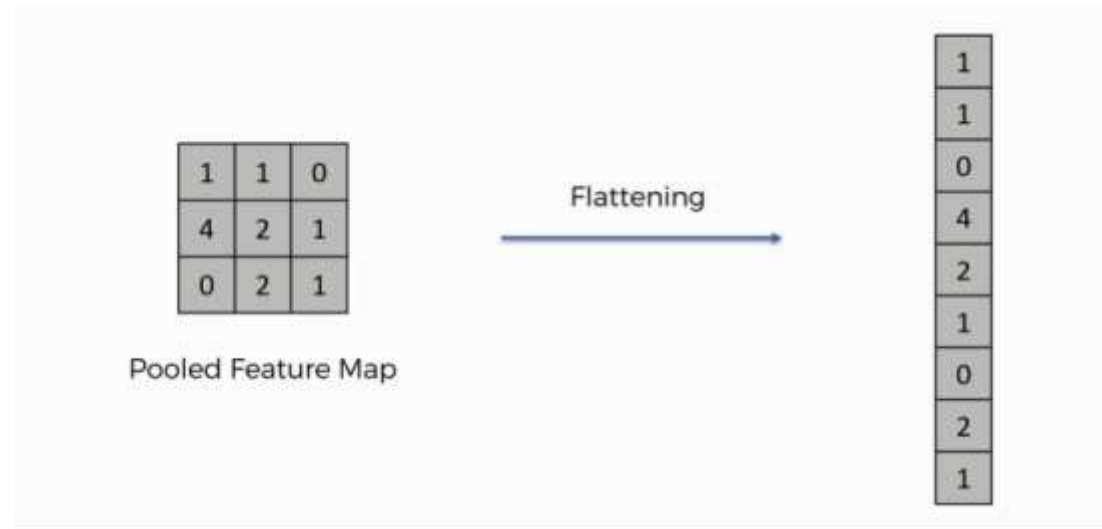
Research shows that zero-padding is not followed.  
Because we are interested in down-sampling

Common setting for filter 2 or 3



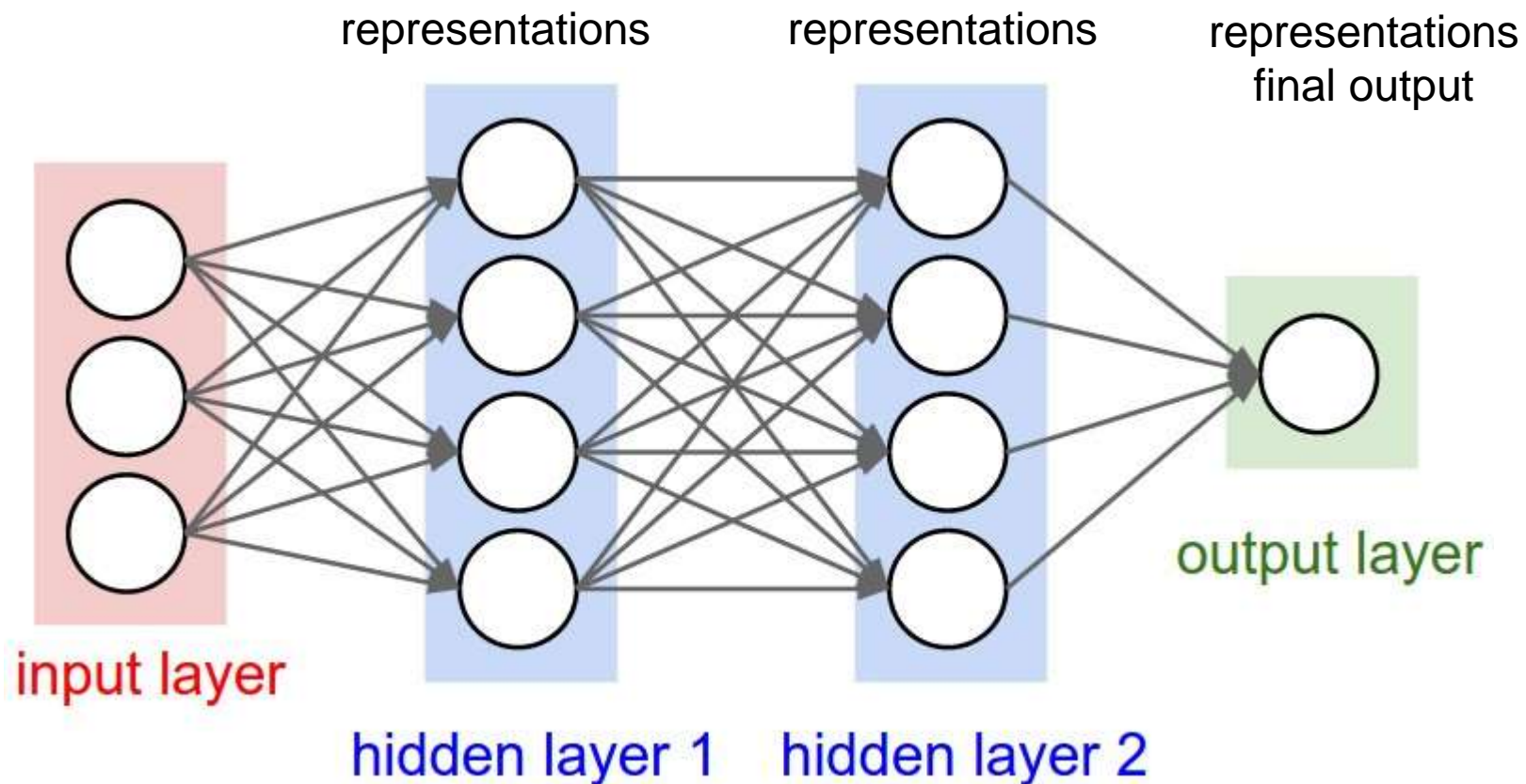
## 4. Flatten

**Procedure to transform a 2D matrix (features) to a 1D vector which in turn can be fed into a fully-connected layer (dense)**



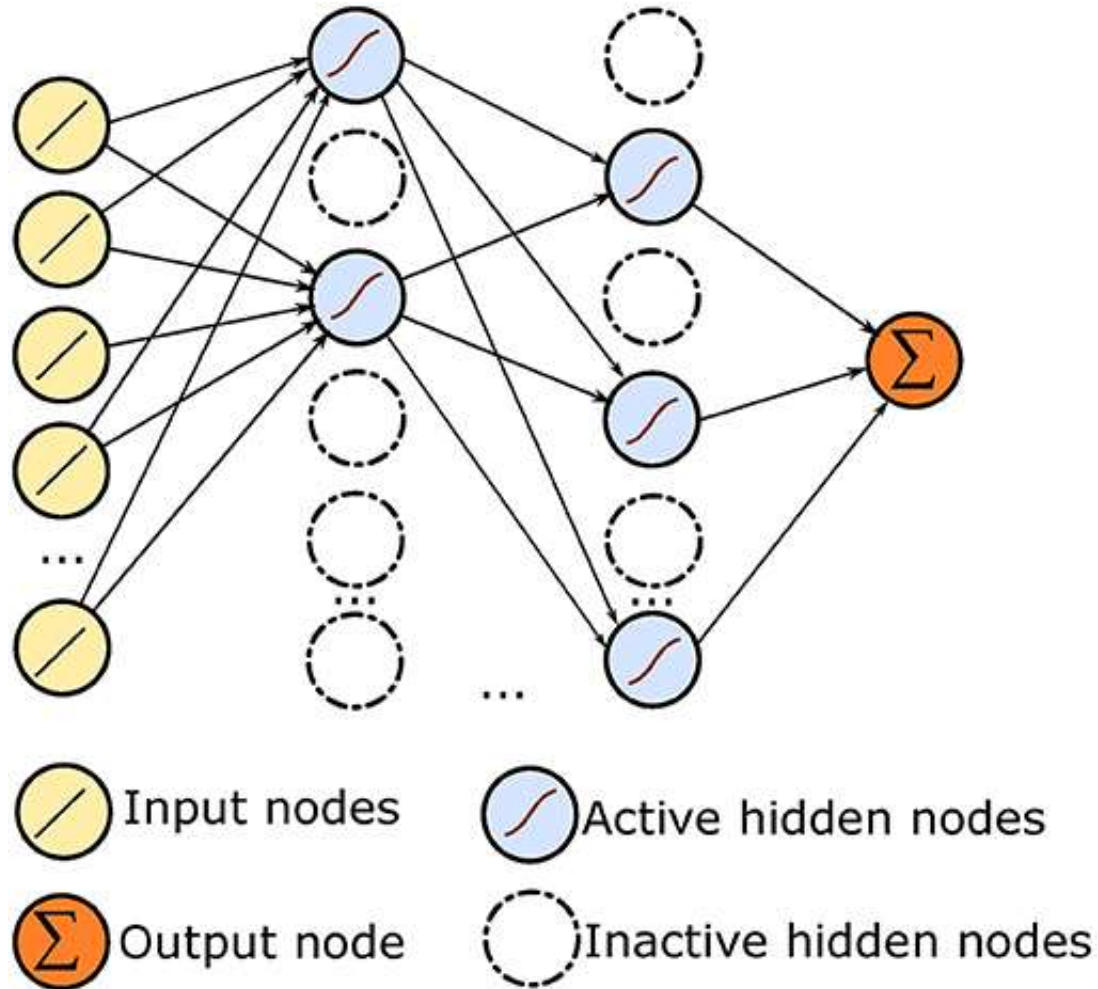
## 5. Dense

Each neuron receives input from all the neurons in the previous layer (densely connected)



[This Photo](#) by Unknown Author is licensed under [CC BY-NC](#)

## 6. Dropout



**Imbalance in the weights among the nodes can lead to some node weights not contributing to the learning**

**One solution:  
Remove a random proportion of selection of neurons in a neural network during training**

**Can help weak learners become strong learners**

## 6. Dropout



[This Photo](#) by Unknown Author is licensed under [CC BY-NC](#)

# Model Summary

~ 154 M parameters

```
1.0 128 10 1
Model: "sequential_1"
```

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 60464, 128)	2688
activation_1 (Activation)	(None, 60464, 128)	0
max_pooling1d_1 (MaxPooling1D)	(None, 60464, 128)	0
conv1d_2 (Conv1D)	(None, 60455, 128)	163968
activation_2 (Activation)	(None, 60455, 128)	0
max_pooling1d_2 (MaxPooling1D)	(None, 6045, 128)	0
flatten_1 (Flatten)	(None, 773760)	0
dense_1 (Dense)	(None, 200)	154752200
activation_3 (Activation)	(None, 200)	0
dropout_1 (Dropout)	(None, 200)	0
dense_2 (Dense)	(None, 20)	4020
activation_4 (Activation)	(None, 20)	0
dropout_2 (Dropout)	(None, 20)	0
dense_3 (Dense)	(None, 15)	315
activation_5 (Activation)	(None, 15)	0

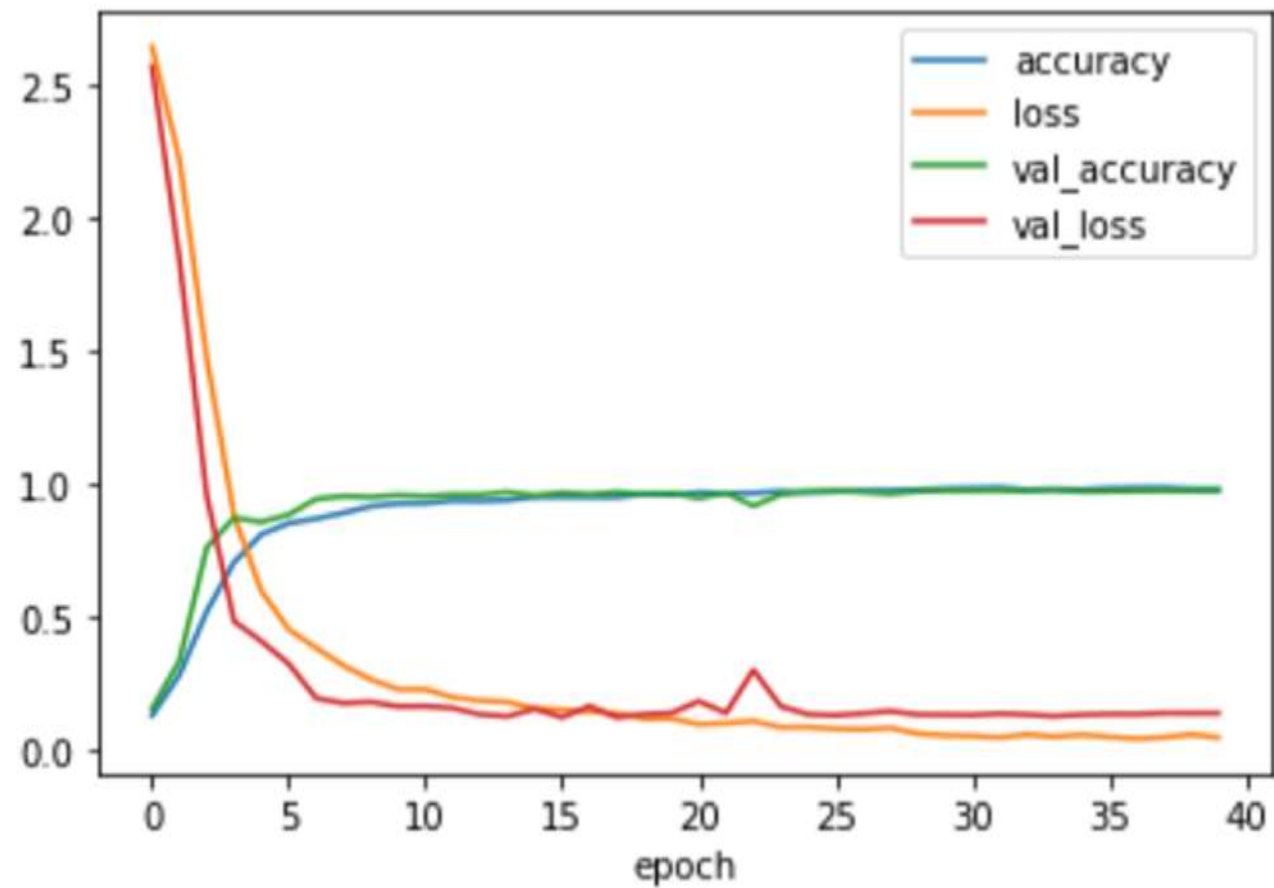
```

Total params: 154,923,191
Trainable params: 154,923,191
Non-trainable params: 0

```



# Model Performance



# Thank you!

[This Photo](#) by Unknown Author is licensed under [CC BY-SA-NC](#)

## Questions/Comments

---

S. Ravichandran  
[ravichandrans@mail.nih.gov](mailto:ravichandrans@mail.nih.gov)

