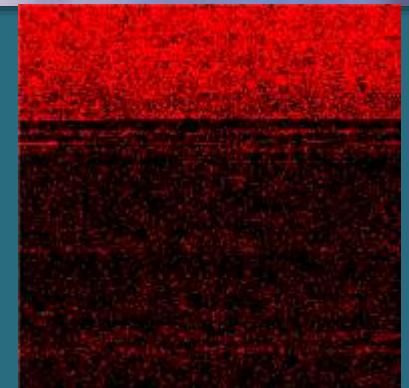


Cancer Type/Site Classification using Deep-Learning (Preliminary presentation slides)

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>

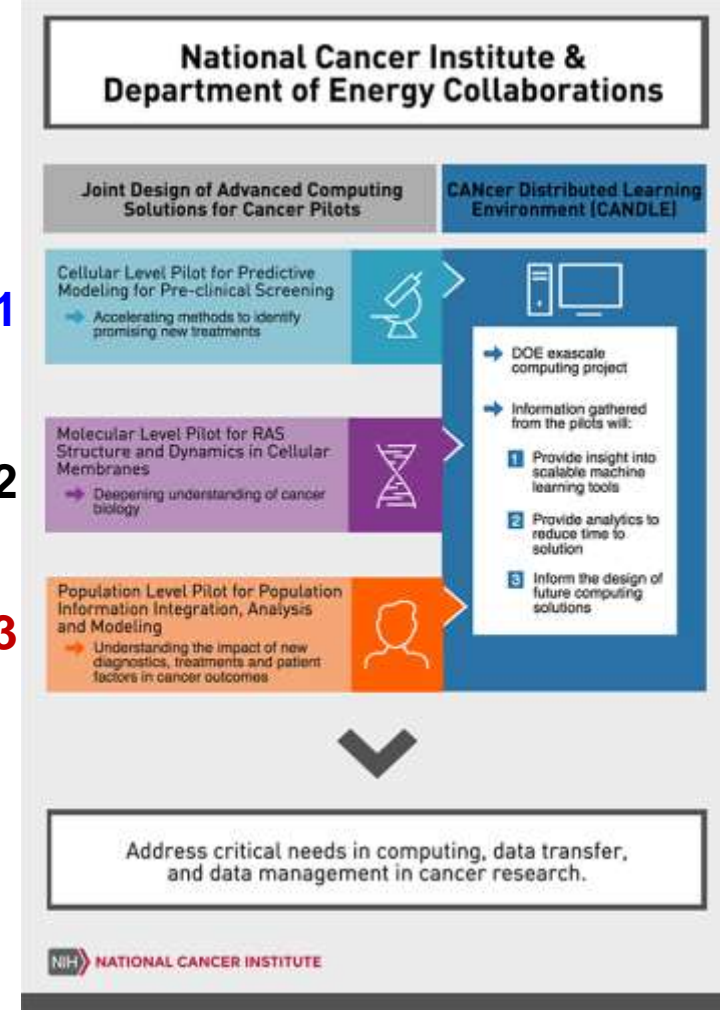
The Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

- 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

- **This workshop is part of the NCI-DOE Pilot project knowledge/capability transfer efforts**
- **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

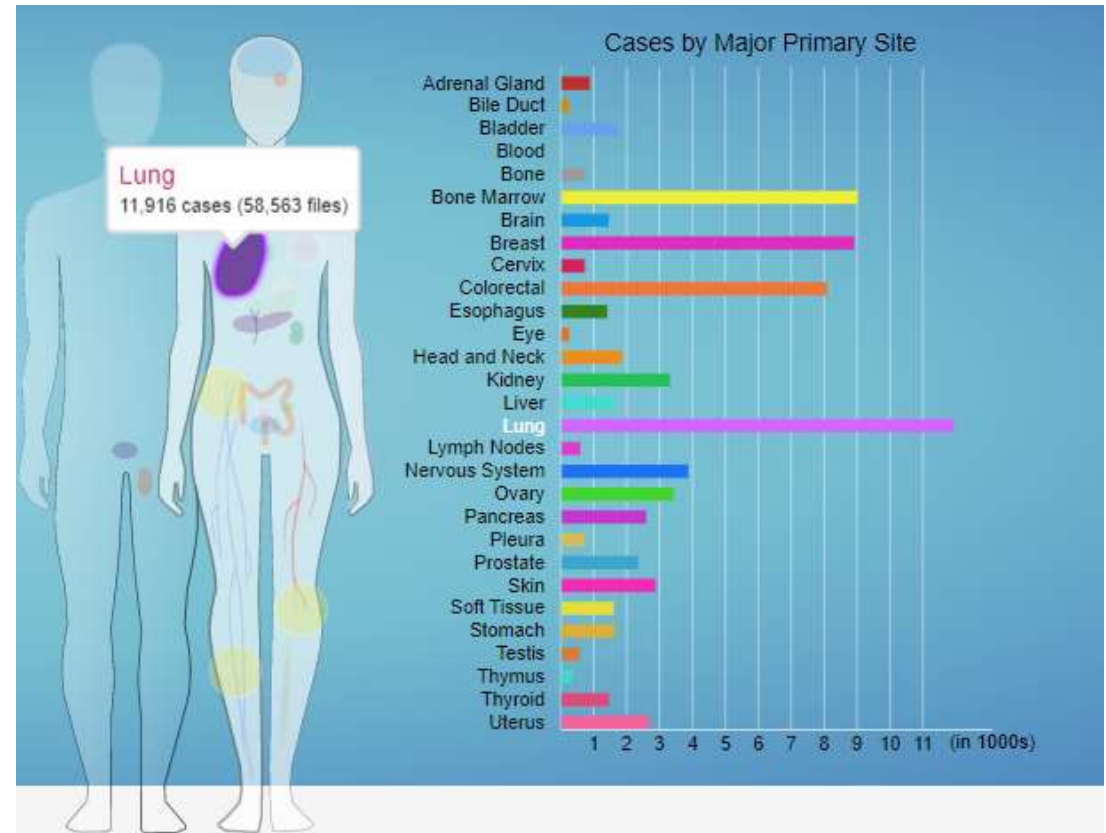
Goal(s)/Questions

- **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?**
- **Can we use the model as early cancer type detection**
 - Improving chance of early detection cure/survival?

Cancer Burden

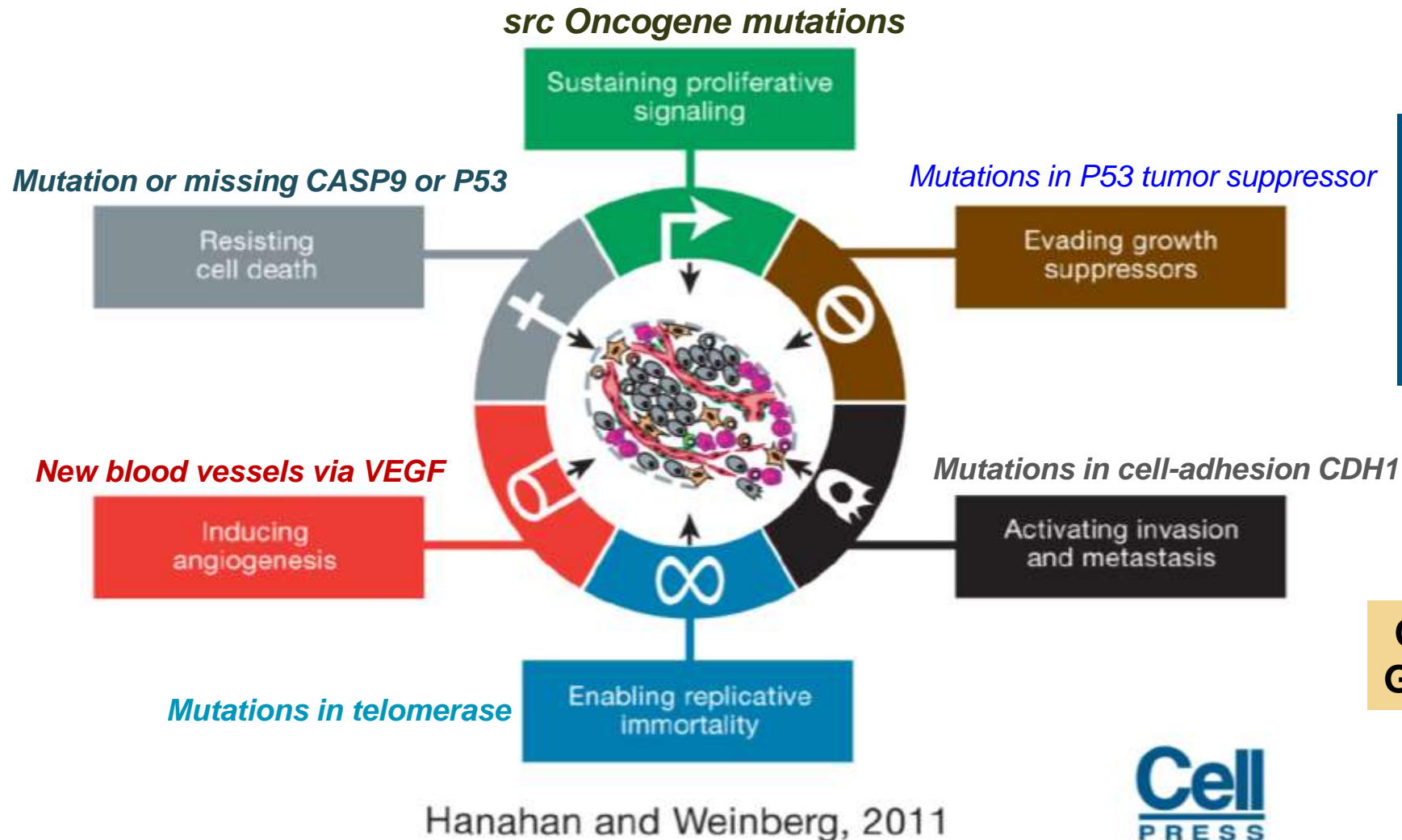
- **Cancer is a group of diseases with world-wide risk**
- **Acquired or somatic changes causes 90-95% of cancer (all types)**
 - *Source TCGA*
- **~ 200 forms of cancer**
 - *DOI: 10.5114/wo.2014.47136*
- **For 2020 in USA**
 - ~1.8M new cancer cases are expected
 - ~600K deaths will occur
 - *Source: American Cancer Society*

Figure from Genomic Data Commons



Hallmarks of cancer: Integral Components of Most Forms of Cancer

Hallmarks of Cancer: The Next Generation



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

The Hallmarks of Cancer

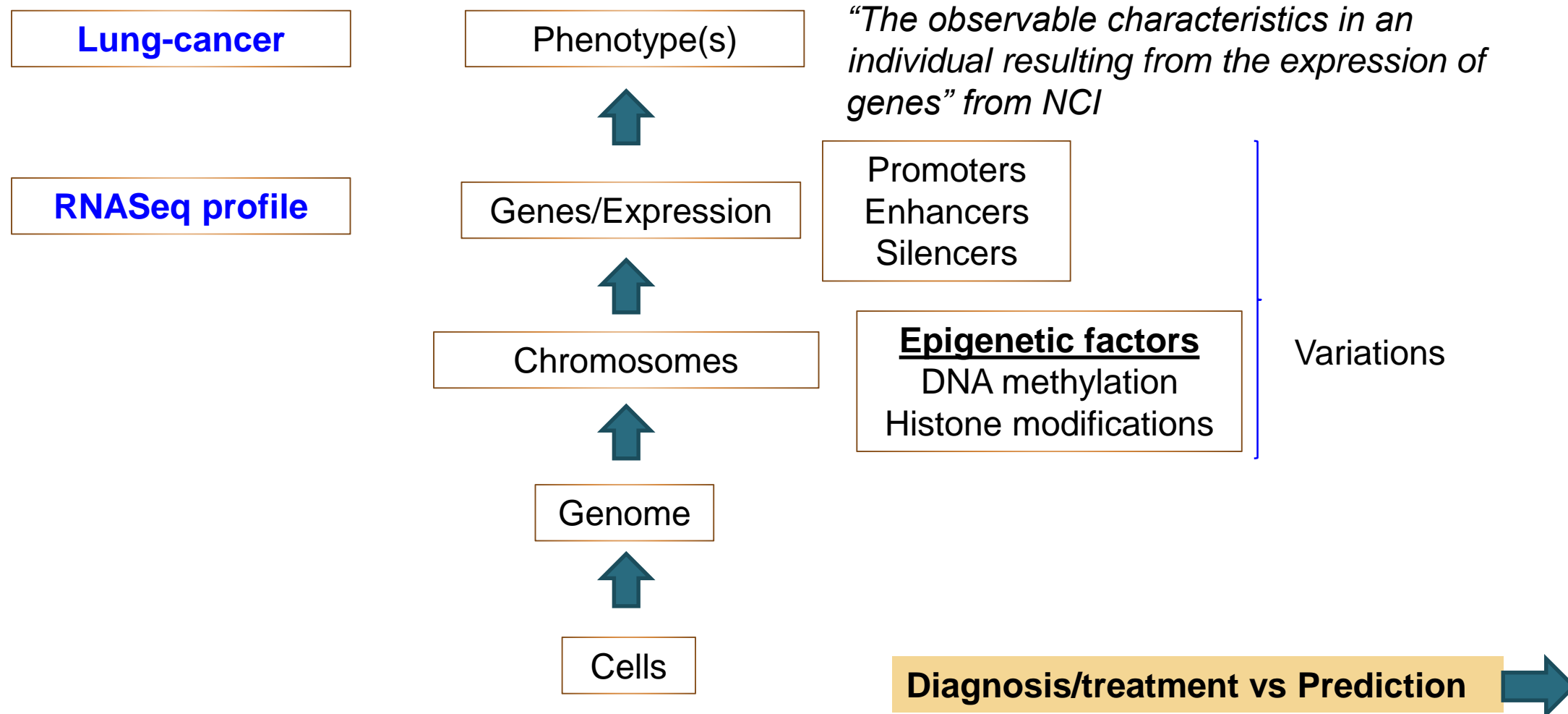
Douglas Hanahan • Robert A Weinberg

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



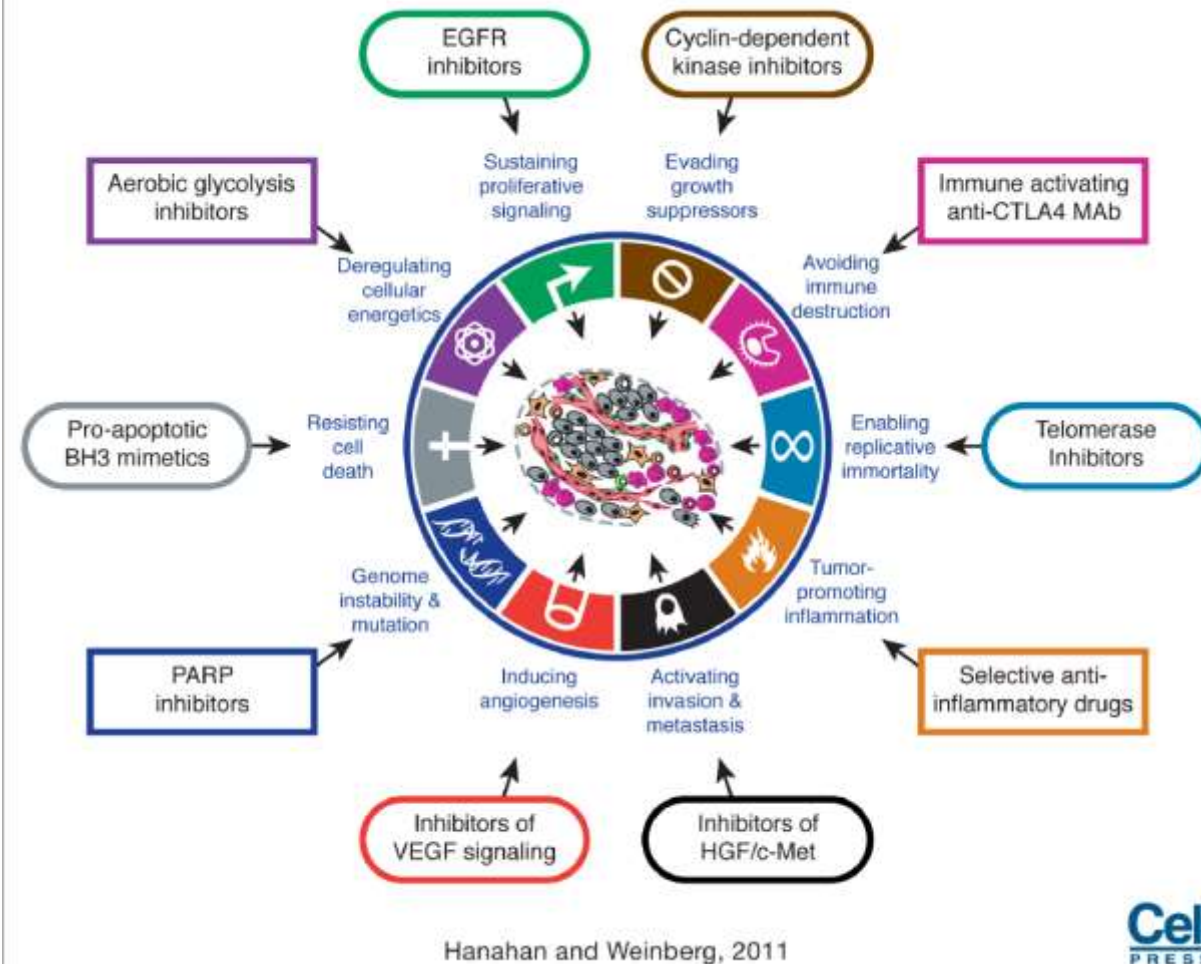
Treatment vs Type-Prediction

- **Treatment**

- Gene-centric (or a slice of pathway)
- Imatinib targeting BCR/KIT

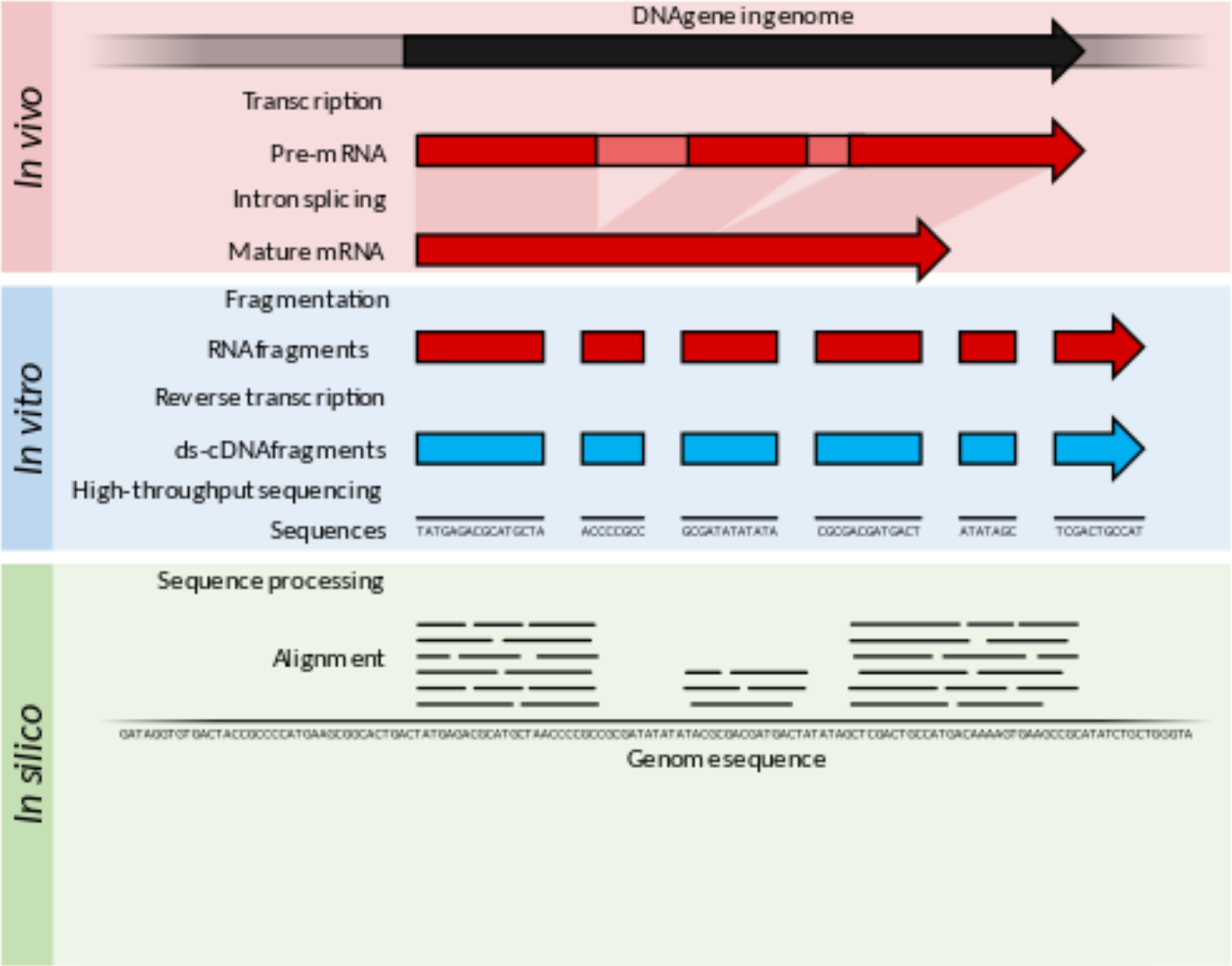
- **Detecting Type**

- 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

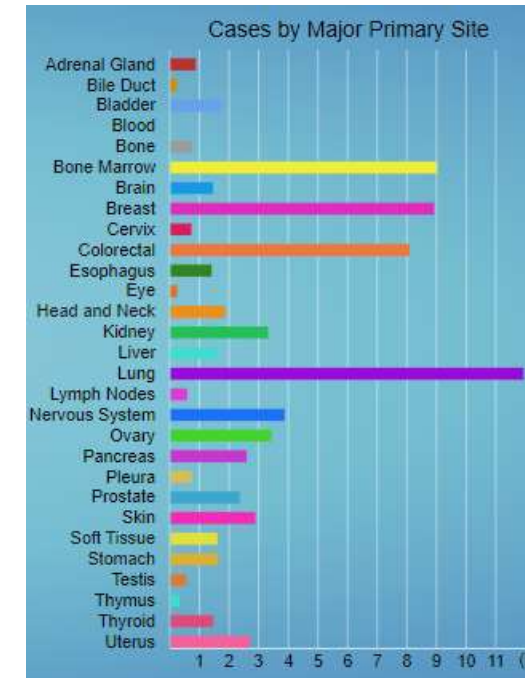
Data source: The Cancer Genome Atlas (TCGA)

- 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

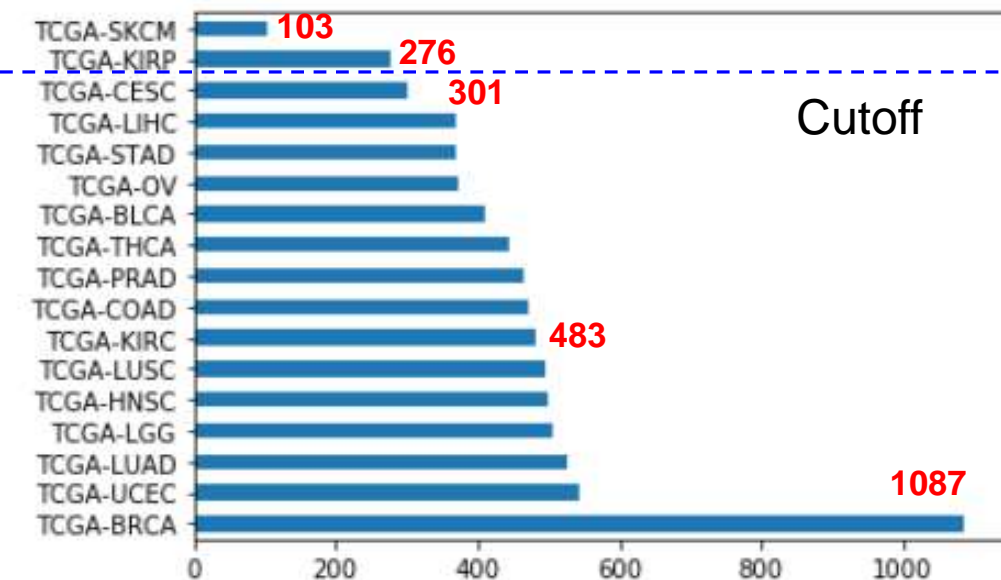
Alignment to Ref
Genome (SAM/BAM)

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

Breast Cancer
60,484
transcripts

Data Preparation

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

Breast Cancer

Games	Expression
ENSG00000242826.2	3558.464179
ENSG00000270123.1	404.2343413
ENSG00000257125.1	52440.13006
ENSG0000021842.1	108.4118421
ENSG00000273172.3	2815.456626
ENSG00000214608.1	105.2999341
ENSG00000252554.0	61.9822385
ENSG00000248122.12	94.9473768
ENSG00000238242.2	406.85658
ENSG0000023983.1	6108.990052
ENSG00000215983.1	
ENSG00000239472.2	
ENSG00000217881.1	
ENSG000002114108.11	95.7373026
ENSG00000260819.1	4848.027373
ENSG00000271712.17	22.47148
ENSG00000217700.2	2267.61708
ENSG00000240423.1	2082.140525
ENSG00000240921.1	103.5246740
ENSG00000240922.0	50.8532316
ENSG00000274656.1	
ENSG00000248811.1	
ENSG00000240301.1	374.357560
ENSG00000211105.1	158.312582
ENSG00000243044.1	
ENSG00000242141.8	455.3084058
ENSG00000240194.18	119.6084704

Games	Expression
ENSG00000242268.2	1658.464179
ENSG00000071718.1	460.234343
ENSG00000150735.1	52440.1006
ENSG0000013842.1	1842.1
ENSG00000178217.5	68155.4626
ENSG00000140681.0	25519.9.235.1
ENSG00000252575.4	0
ENSG00000148448.12	194.947368
ENSG00000192842.12	4068556.55
ENSG00000215881.1	6108.999052
ENSG00000219883.0	0
ENSG00000234852.1	0
ENSG00000117488.1	0
ENSG00000141301.18	973730.2656
ENSG00000263809.1	3484.627373
ENSG00000172312.17	48545.9071
ENSG00000174207.17	22617.4017
ENSG00000214943.2	2082.45025
ENSG00000240243.1	130.5246749
ENSG00000192909.9	550683.3216
ENSG00000274661.1	0
ENSG00000218581.0	0
ENSG00000248811.0	0
ENSG00000216040.1	394.475567
ENSG00000111057.1	1583.312582
ENSG00000243041.1	0
ENSG00000182144.8	45338.0638
ENSG00000218416.18	118.087054

Genes	Expression
ENSG0000024268.2	1608.49749
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	914.9473788
ENSG0000028426.12	4089655.68
ENSG0000025983.1	6180.99052
ENSG0000023198.3	0.0
ENSG0000026047.2	0.0
ENSG00000210788.1	0.0
ENSG00000234108.11	957330.256
ENSG0000026308.1	3484.62773
ENSG0000027217.1	413.77777
ENSG0000026790.7	22617.400
ENSG00000234943.2	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 ₂	Expression
ENSG00000242262	1258.46	4841.79
ENSG00000270218	1258.46	234.13
ENSG00000271735	1258.46	1000.00
ENSG00000271842	1258.46	1000.00
ENSG00000278275	1258.46	468.56
ENSG00000410801	1258.46	25059.23
ENSG00000252754	1258.46	1000.00
ENSG00000258428	1258.46	1947.93
ENSG00000218422	1249.85	616.89
ENSG00000259831	6128.99	9990.52
ENSG00000259831	6128.99	1000.00
ENSG00000259872	6128.99	1000.00
ENSG00000217881	6128.99	1000.00
ENSG00000114101	19737.0	2056.0
ENSG00000138901	3484.0	6773.73
ENSG00000271717	2287.1	400.00
ENSG00000167707	2287.1	400.00
ENSG00000214043	2082.4	205.65
ENSG00000240243	1231.5	5246.49
ENSG00000274561	1506.83	32.00
ENSG00000274561	1506.83	1000.00
ENSG00000214881	1506.83	1000.00
ENSG00000240404	1394.47	565.00
ENSG00000211051	1394.47	32.00
ENSG00000240404	1394.47	1000.00
ENSG00000219148	4533.8	606.48
ENSG00000242464	1138.07	8754.0

Genes	Time	Expression
ENSG00000242688.2	12658.2	1568.464179
ENSG00000270123.1	60	204.234433
ENSG00000271719.1	18	52440.10096
ENSG00000271842.1	18	184.23212
ENSG00000278275.1	6	68.16514656
ENSG00000280148.1	20	25.9599.2351
ENSG00000252574.0	6	0
ENSG00000254575.1	18	194.9457788
ENSG00000280426.12	12	40.89556.65
ENSG00000259831.1	6	61.6909552
ENSG00000251983.1	0	0
ENSG00000254075.2	20	0
ENSG00000217881.1	0	0
ENSG00000214108.11	15	95.7330.2505
ENSG00000280898.1	18	348.657173
ENSG00000271217.1	18	174.1717.17
ENSG00000257907.1	7	22.6717.4040
ENSG00000240443.1	20	20.62.2655
ENSG00000240542.1	10	53.6467.90
ENSG00000276561.1	15	55.663.5216
ENSG00000248811.1	0	0
ENSG00000246041.1	18	394.475567
ENSG0000021155.1	18	158.45.45
ENSG00000230404.1	18	33.375849
ENSG00000292148.1	6	45.538.60548
ENSG00000269414.1	18	118.087054

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
ENSG0000025883.1	6308.999052
ENSG0000023198.3	0
ENSG0000028472.9	0
ENSG0000001780.1	0
ENSG0000013408.1	957330.2056
ENSG0000000081.1	3484.027173
ENSG0000017211.7	247.7171
ENSG0000016700.2	2281.737405
ENSG0000024943.2	2082.240525
ENSG0000024353.2	310.5246749
ENSG0000026624.9	550683.9216
ENSG0000027456.1	0
ENSG0000023488.1	0
ENSG0000023040.1	394.4755477
ENSG0000011026.1	833.312582
ENSG0000000404.1	0
ENSG0000018214.8	45358.60648
ENSG0000000494.6	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	2351
ENSG00000252574.5	0	0
ENSG00000254886.12	12	987.58
ENSG00000258042.2	94	49657.68
ENSG00000258083.1	6388	99002.952
ENSG00000251988.1	0	0
ENSG00000250425.2	0	0
ENSG00000201788.1	1400	11
ENSG00000213401.11	95730	2056
ENSG00000203808.11	3468	62773
ENSG00000217117.12	48	952
ENSG00000256700.7	22671	2420
ENSG00000243493.2	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	60548
ENSG00000249446.4	128	48705.4

Gene	Expression
ENSG00000244228.2	1658.464179
ENSG000002701.123	460.234343
ENSG000002785.78.25	2404.10096
ENSG000002784.2	178.08423
ENSG000002786.15.7	68165.46526
ENSG000002400.30	25599.2511
ENSG000002252.574	50
ENSG000002248.126	1846.5812
ENSG000002382.122	194.947378
ENSG000002382.122	408955.68
ENSG000002598.81	6108.99052
ENSG000002319.83	3
ENSG000002345.2	100
ENSG00000230788.1	0
ENSG00000213408.116	957330.266
ENSG000002008.11	3484.02773
ENSG000002721.17	172.117
ENSG00000267700.7	22817.440
ENSG000002349.42	2082.14055
ENSG000002403.23	310.524679
ENSG000002362.42.9	23663.122
ENSG0000027156.261	1
ENSG000002348.81	1
ENSG000002340.40	394.475627
ENSG00000231125.1	83.8312562
ENSG000002400.44	1
ENSG000002321.418	45338.40548
ENSG000002394.16	139.087474

Lung Cancer

Genes	Population
ENSG000002428262	3558 484179
ENSG000002717123	460 2343433
ENSG000002505125	52440 10096
ENSG000002717123	460 2343433
ENSG000002408105	18165 46626
ENSG00000214608105	255959 2351
ENSG00000225274	0
ENSG000002448148	13 1974768
ENSG000002358412	406856 658
ENSG000002598311	6108 99052
ENSG000002198313	0
ENSG000002358412	406856 658
ENSG000002378811	0
ENSG000002378811	0
ENSG00000134108.11	957330 20656
ENSG00000283080.1	4448 50737
ENSG000002733713	377 17317
ENSG000002358412	226717 406856
ENSG000002404212	2081 240545
ENSG000002358412	130 5246740
ENSG000002358412	151863 9216
ENSG00000274561.1	0
ENSG00000234881.1	0
ENSG000002402041	394 3425560
ENSG0000021105.1	583 317562
ENSG00000243044.1	1
ENSG00000231214.8	455 38 60648
ENSG00000249164.6	118 0874054

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
ENSG00000206809.1	3448.02773
ENSG00000217317.17	4454.5667
ENSG00000207902.7	22977.4208
ENSG00000214943.12	2082.43565
ENSG00000214042.13	130.5246749
ENSG00000205842.1	1593.06132
ENSG00000217464.1	0
ENSG00000214881.1	0
ENSG00000206040.1	394.4755669
ENSG0000021105.1	1583.312582
ENSG00000204044.1	0
ENSG00000218244.18	455.38.60648
ENSG00000208916.14	118.087054

Genes	Annotation
ENSEG0000024268.2	3526.464719
ENSEG00000210713.1	460.234343
ENSEG000002675.8	52440.30096
ENSEG000002782.7	6835.46466
ENSEG000002827.5	25599.251
ENSEG0000025274.4	0
ENSEG000002668.1	154.947378
ENSEG000002842.12	490658.658
ENSEG0000025883.1	6148.99052
ENSEG0000023168.1	0
ENSEG000002675.2	0
ENSEG0000021788.1	0
ENSEG0000013408.11	957330.265
ENSEG0000026308.1	3484.02773
ENSEG0000027117.17	225717.4008
ENSEG000002700.7	225717.4008
ENSEG0000023494.3	2082.24051
ENSEG0000024043.1	130.5246749
ENSEG0000023419.9	153681.9216
ENSEG0000027616.1	0
ENSEG0000023488.1	0
ENSEG0000023040.1	394.4755669
ENSEG0000023165.1	538.332582
ENSEG000002841.8	45338.60648
ENSEG0000029246.8	118.087454

Genotype	Frequency
ENSG00000242268	2.358E-464179
ENSG00000270173	4.06E-234343
ENSG00000267125	5.240E-10096
ENSG00000278215	6.883E-4656
ENSG00000282175	2.5599E-2351
ENSG00000252574	0
ENSG00000268485	1.54E-947738
ENSG00000280422	4.0965E-6816
ENSG00000258831	6.108E-99052
ENSG00000231963	0
ENSG00000249752	2
ENSG00000217081	0
ENSG00000234028	9.7533E-2056
ENSG00000230021	3.484E-22713
ENSG00000271217	4.514E-4085
ENSG00000267800	2.2737E-4208
ENSG00000234943	2.082E-2405
ENSG00000240433	3.10E-524670
ENSG00000231919	5.158E-13262
ENSG00000271656	0
ENSG00000234881	0
ENSG00000236040	394.4755669
ENSG00000231535	3.835E-33252
ENSG00000240044	0
ENSG00000282146	4.5358E-4068
ENSG00000259446	1.19E-087454

Genes	Log2FC	Log10P
ENSG00000242588.2	3258.4	4642.79
ENSG00000210713.1	460	2343.43
ENSG000002752.5	52440	30996
ENSG000002788.1	282	2121
ENSG000002782.5	68365	46626
ENSG000004108.1	25919	2518
ENSG00000275274.4	0	0
ENSG000002784.1	14488	132
ENSG000004107.2	146	94377.98
ENSG0000041026.12	409656	658
ENSG0000025883.1	6386	39905.52
ENSG00000231983.0	0	0
ENSG00000269752.2	0	0
ENSG000002788.1	0	0
ENSG0000041048.11	957330	2656
ENSG0000038109.1	3484	6273.73
ENSG00000272137.17	44435	277.9
ENSG0000027260.7	22677	4208
ENSG0000024043.1	2082	2465
ENSG000004313.3	130	52647.49
ENSG0000027119.9	153683	1262
ENSG0000027656.1	0	0
ENSG00000234881.1	0	0
ENSG0000043040.1	394	47556.69
ENSG0000023155.2	533	3335.81
ENSG0000041814.8	45338	60648
ENSG0000029246.6	119	38470.54

Genotype	Approximation
ENTG0000024268.2	15626.46479
ENTG0000027123.18	5460.234343
ENTG0000026075.75	12420.10096
ENTG0000028182.2	0
ENTG0000028275.75	68355.46456
ENTG0000046018.30	215999.2814
ENTG0000025275.4	0
ENTG0000028488.13	194.9673798
ENTG0000028942.12	49865.5616
ENTG0000025883.1	6108.99252
ENTG0000023198.13	0
ENTG0000025672.2	0
ENTG0000021788.1	0
ENTG0000023408.11	957330.2656
ENTG0000023008.1	3484.02733
ENTG0000027217.47	414.48576
ENTG0000027800.3	22571.47048
ENTG0000024943.2	2082.24051
ENTG0000024042.3	130.5246749
ENTG0000024112.9	15186.13216
ENTG0000027165.1	0
ENTG0000024881.1	0
ENTG0000023040.1	394.4755639
ENTG0000023105.1	583.312582
ENTG0000023004.1	0
ENTG0000028241.8	45738.50648
ENTG0000029546.6	119.087054

ENSG00000242568.2	Exp
ENSG00000242568.2	1658.464179
ENSG00000270123.3	460.234343
ENSG00000275718.1	25404.10096
ENSG00000276235.5	68165.46656
ENSG00000268307.10	205919.2351
ENSG00000252925.4	0
ENSG00000256618.12	94.9473768
ENSG00000258042.12	409856.5568
ENSG00000259831.1	6308.990052
ENSG00000231983.1	0
ENSG00000256525.1	0
ENSG00000251788.1	0
ENSG00000234208.11	95730.2656
ENSG00000236009.1	3844.02773
ENSG00000272145.1	485.9507
ENSG00000270072.7	23571.0
ENSG00000234943.2	2082.2405
ENSG00000240523.1	350.524679
ENSG00000234432.1	53863.192
ENSG00000275561.1	0
ENSG00000234881.1	0
ENSG00000236040.1	394.475567
ENSG00000235565.1	3833.112562
ENSG00000230044.1	0
ENSG00000258241.6	45338.0648
ENSG00000269456.4	129.087054

ENSG00000242568.2	Expression
ENSG00000242568.2	1658.46479
ENSG00000270112.3	462.234343
ENSG00000268218.3	2504.10096
ENSG00000276373.5	68365.46656
ENSG00000268103.2	25959.2151
ENSG00000252575.4	0
ENSG00000256866.13	24.9473718
ENSG00000258042.12	40985.56
ENSG00000259883.1	6198.99052
ENSG00000231983.3	0
ENSG00000259572.2	0
ENSG00000201788.1	0
ENSG00000234028.11	957330.256
ENSG00000230911.1	3484.02793
ENSG00000272700.1	14485.407
ENSG00000249007.1	23971.4708
ENSG00000234493.2	2082.24055
ENSG00000234943.3	130.526479
ENSG00000259443.3	5186.912
ENSG00000275364.1	0
ENSG00000234881.1	0
ENSG00000236940.1	394.45765
ENSG00000243043.1	383.312582
ENSG00000243044.1	0
ENSG00000282148.6	45338.60648
ENSG00000269446.6	129.087054

Kidney Cancer

Genes	Expression
ENSG00000242682	1566.454179
ENSG000002701123	460.2343433
ENSG000001575815	52440.10096
ENSG000001575815	52440.10096
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000001424851	124.9473788
ENSG000001924012	409856.58
ENSG000002798831	6100.99052
ENSG000001198133	0
ENSG00000194752	0
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG000002308914	3484.027373
ENSG000001733123	41495.97057
ENSG000001028681	3000.02861
ENSG000001340018	95730.256
ENSG00000273375	68165.48646
ENSG000001460818	25559.93125
ENSG00000252754	0
ENSG000	

Genes	Expression
ENSG0000024268.2	1658.464179
ENSG00000207012.13	460.234343
ENSG00000269758.7.5	52440.10096
ENSG00000269758.7.5	52440.10096
ENSG00000278237.5	6815.456262
ENSG00000246038.10	25899.65155
ENSG00000252375.4	0
ENSG00000244646.12	1349.747368
ENSG00000201928.12	496856.58
ENSG00000259883.1	6108.999052
ENSG00000231981.3	0
ENSG00000264075.11	0
ENSG0000020788.1	0
ENSG00000214048.11	95730.23056
ENSG00000263089.1	3484.027733
ENSG00000217317.17	41485.907
ENSG00000207476.2	23917.42847
ENSG00000234943.2	2824.56035
ENSG00000204024.11	130.524769
ENSG00000206482.9	155861.9216
ENSG00000271616.11	0
ENSG00000204681.9	0
ENSG00000238040.1	394.475569
ENSG00000231105.11	3513.31282
ENSG00000243044.1	0
ENSG00000208445.1	45338.05648
ENSG00000269164.6	1481870.754

Genes	Expression
ENSEG0000004268.2	1658.464719
ENSEG0000001711.2	46.2043443
ENSEG0000007052.5	5244.10096
ENSEG000000782.3	68.051466
ENSEG0000004837.5	20.5599125
ENSEG00000025274.0	
ENSEG0000004686.1	124.9473768
ENSEG0000009244.2	49.6805168
ENSEG00000059883.1	6.00819952
ENSEG00000031983.0	0.1470752
ENSEG00000020788.1	
ENSEG00000034058.1	95.73362056
ENSEG00000012899.8	3.48420773
ENSEG00000071317.1	14.4851957
ENSEG00000040432.1	2.87617058
ENSEG00000034943.1	2.4035105
ENSEG00000040421.3	120.5246749
ENSEG00000046042.9	15.08611924
ENSEG00000071664.6	
ENSEG00000034881.1	
ENSEG00000036040.1	394.4755669
ENSEG00000031050.1	15.8312582
ENSEG00000049444.4	
ENSEG00000039444.4	15.35810548
ENSEG00000039444.4	119.0470442

Genes	Expression
ENSG00000242368.2	1658.464179
ENSG00000210711.3	406.234343
ENSG00000270825.1	53440.10096
ENSG00000272821.2	178.724121
ENSG00000276273.2	6816.62565
ENSG00000146038.5	25569.25121
ENSG0000025275.4	0
ENSG0000026615.2	943.747768
ENSG00000198242.12	49056.5616
ENSG00000159883.1	6108.999821
ENSG00000231981.3	0
ENSG0000026475.2	0
ENSG0000025768.1	0
ENSG00000134108.11	957330.2056
ENSG00000271397.1	3484.027273
ENSG00000162187.1	41485.907
ENSG00000232878.1	238.029238
ENSG00000234943.2	2802.240535
ENSG00000244042.4	130.1526749
ENSG00000266042.9	15586.19212
ENSG00000274816.1	0
ENSG00000234881.1	0
ENSG00000236100.1	394.47553269
ENSG00000130205.1	1583.312582
ENSG00000249464.1	0
ENSG00000241418.9	6752.65548
ENSG00000259166.1	119.0847054

ENSG0000015768.1	52440	10096
ENSG000001727.2	40	29431.3
ENSG0000017382.1	1	0
ENSG0000017927.5	6885	45626
ENSG000146928.10	255	25599.251
ENSG0000012572.4	0	0
ENSG000145846.12	104	947.148
ENSG000109284.12	496856	658
ENSG000121983.1	618	9990.52
ENSG000121981.3	0	0
ENSG000120947.2	0	0
ENSG000120081.1	0	0
ENSG000121418.11	95730	2056
ENSG000121417.10	348	62773
ENSG000121731.17	41485	597
ENSG000165700.7	7267	40
ENSG000124943.2	208	2450.5
ENSG000124923.1	316	52667.9
ENSG000124924.1	5863	9215
ENSG000121636.1	0	0
ENSG000121483.1	0	0
ENSG000124040.1	394	475669
ENSG000121105.1	183	312582
ENSG000124304.1	0	0
ENSG000121214.8	45338	4066
ENSG0001209426.4	119	1084750.4

[illegible]

ENSG00000242762	181	68419
ENSG00000242763	181	4004343
ENSG00000242765	15	53440
ENSG00000278402	1	
ENSG00000278215	5	68416
ENSG00000246262	181	255959
ENSG00000225275	5	255959
ENSG00000150486	128	104
ENSG00000284122	128	49865
ENSG00000259811	6	1108
ENSG00000211818	1	618
ENSG00000209475	2	
ENSG00000278811	1	
ENSG00000141018	115	95730
ENSG00000242764	181	3684
ENSG00000272137	147	41845
ENSG00000257007	7	22871
ENSG00000234942	2	2082
ENSG00000242513	131	52467
ENSG00000242516	131	52468
ENSG00000271651	1	
ENSG00000234811	1	
ENSG00000240061	394	47556
ENSG00000211525	1	131
ENSG00000340041	1	
ENSG00000212148	453	40843
ENSG00000240426	4	

[illegible]

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	3038.404179
ENSG00000276112.3	400.7345413
ENSG0000026978.15	52440.1006
ENSG0000027840.1	0
ENSG0000029121.1	6085.4526
ENSG0000024293.10	25099.2351
ENSG0000025277.4	0
ENSG0000025486.12	104.9473768
ENSG00000219842.12	406856.658
ENSG0000021085.1	6108.19052
ENSG0000021038.3	0
ENSG0000020471.2	0
ENSG00000201788.1	0
ENSG0000021428.11	90730.2056
ENSG0000020208.1	2484.03713
ENSG00000271217.17	41485.9507
ENSG00000207780.7	22672.4208
ENSG0000020484.2	2002.240505
ENSG00000240423.1	305.5246749
ENSG00000200342.9	121863.1216
ENSG00000271816.1	0
ENSG00000214881.1	0
ENSG00000218046.1	394.475669
ENSG00000211105.1	1583.112582
ENSG00000240464.1	0
ENSG00000215141.8	45338.40648
ENSG00000209416.4	119.0847054
ENSG00000204911.1	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^{1,2}, Brian A Williams^{1,2}, Kenneth McCue¹, Lorian Schaeffer¹ & Barbara Wold¹

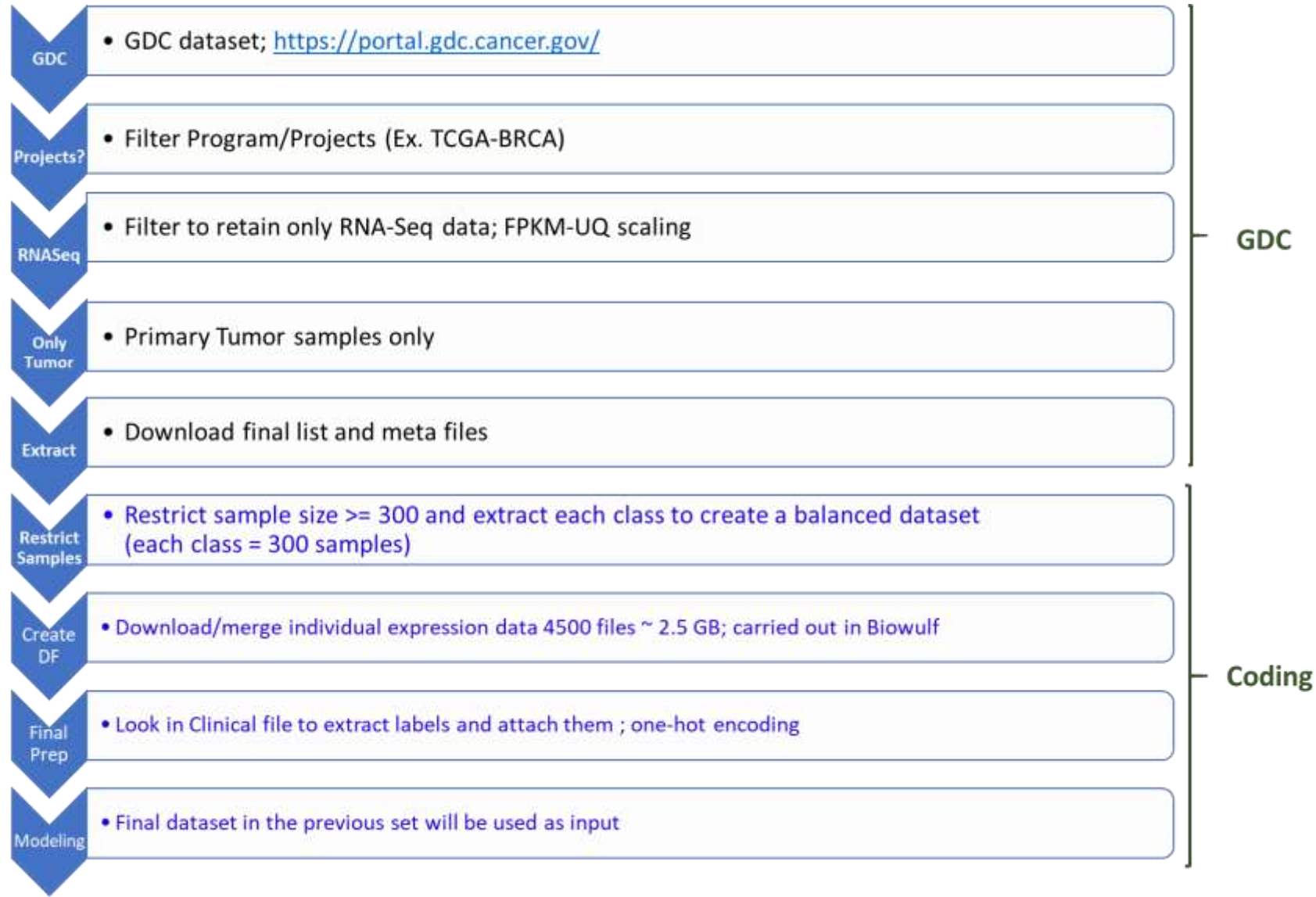
One-hot encoding to convert Cancer types to numbers

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



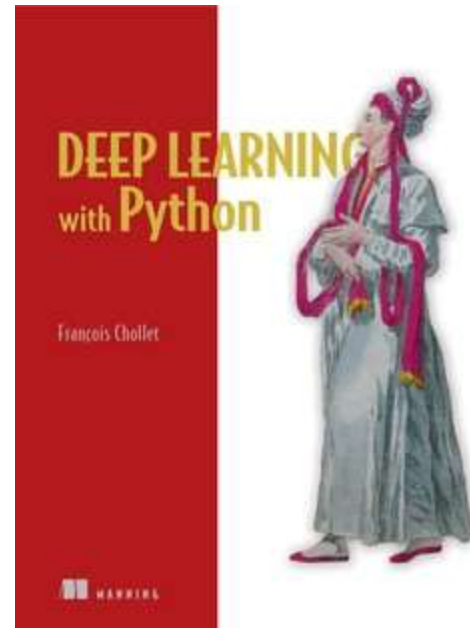
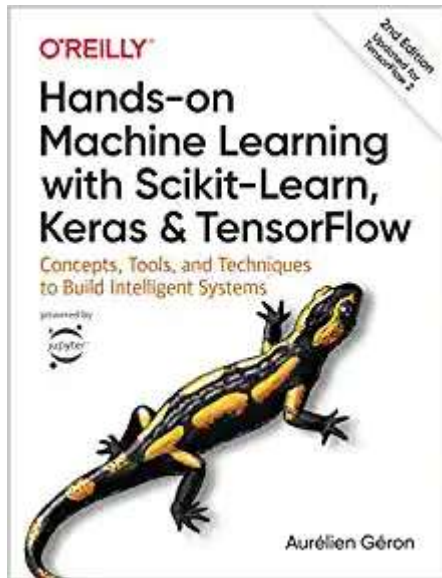
- **Will use Jupyter Notebook for sharing the code**
 - With little effort one can convert the Python code into R and still use Jupyter Notebook

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.

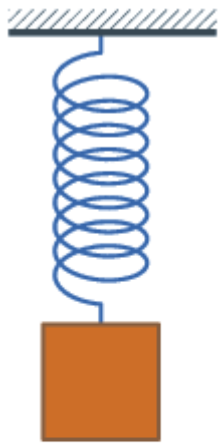
INPUT

ALGORITHM

OUTPUT

A Simple Network

Input: Mass or M (kg)
Output: Length or L (m)

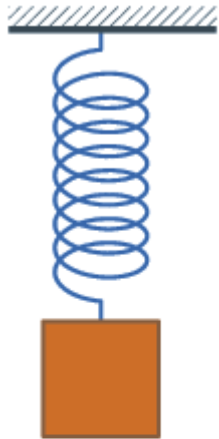


M	L
Input	Output
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	???

[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](#)

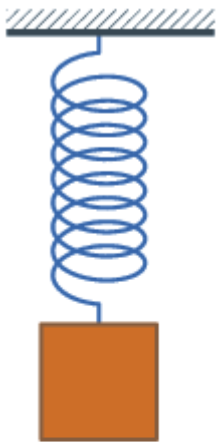
Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003

A Simple Network

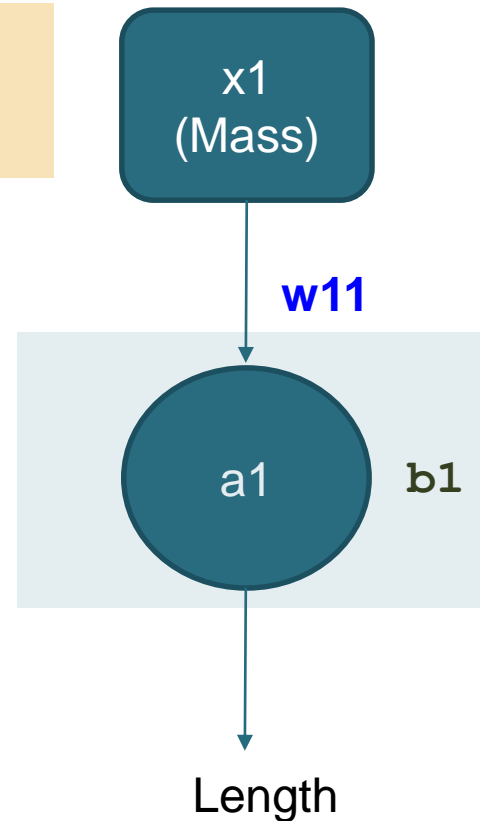
$$a1 = x1 * w11 + b1$$

$$L = M * 0.1 + 0.38$$

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



Hidden Layer



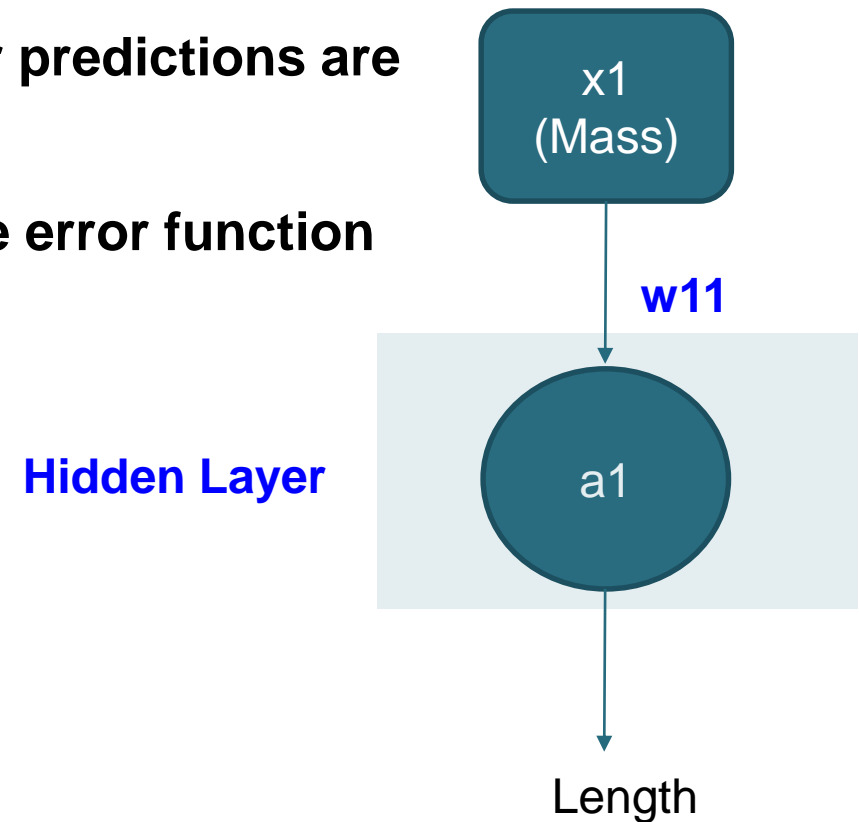
M	L
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	0.68

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

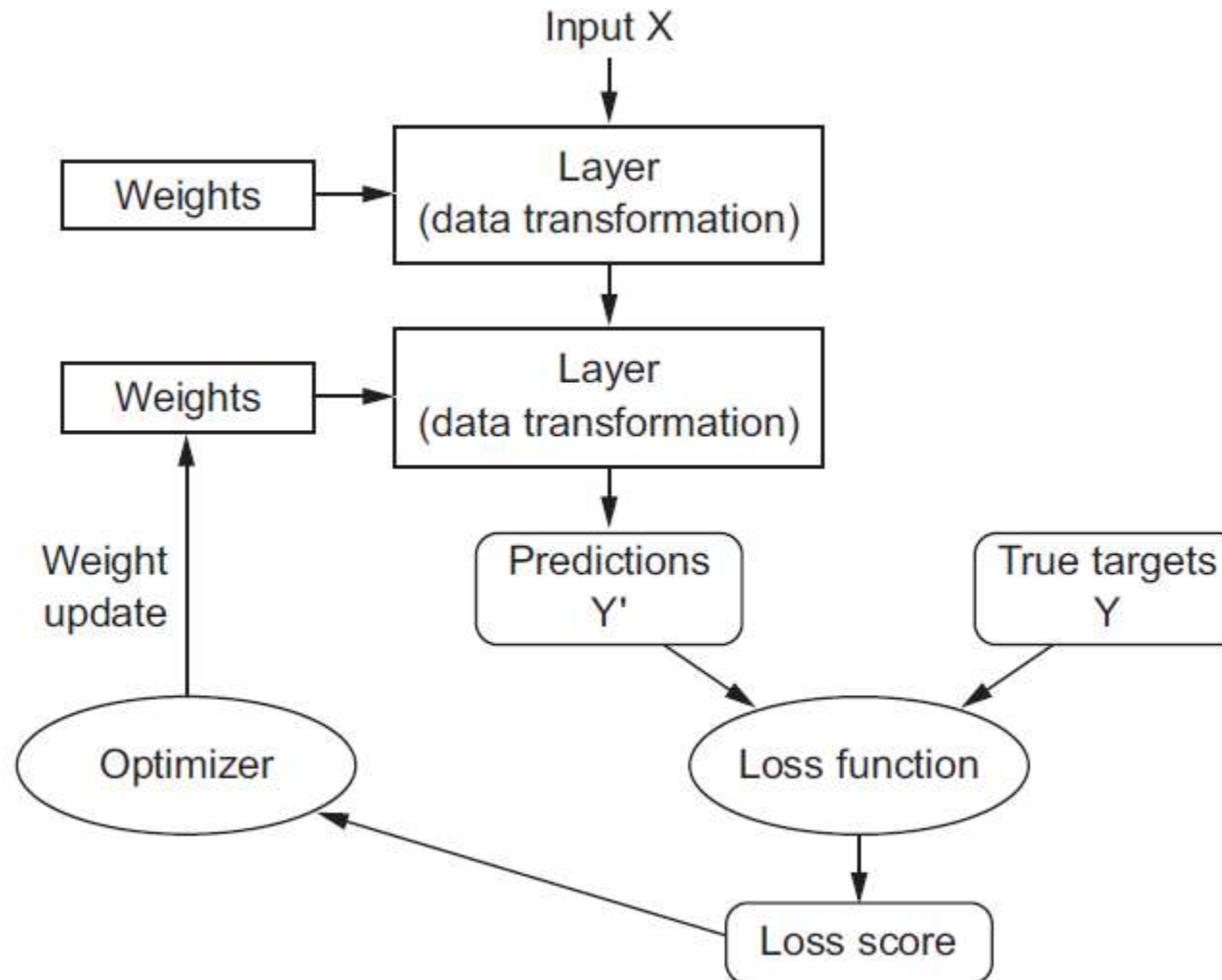
Error minimization

- 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

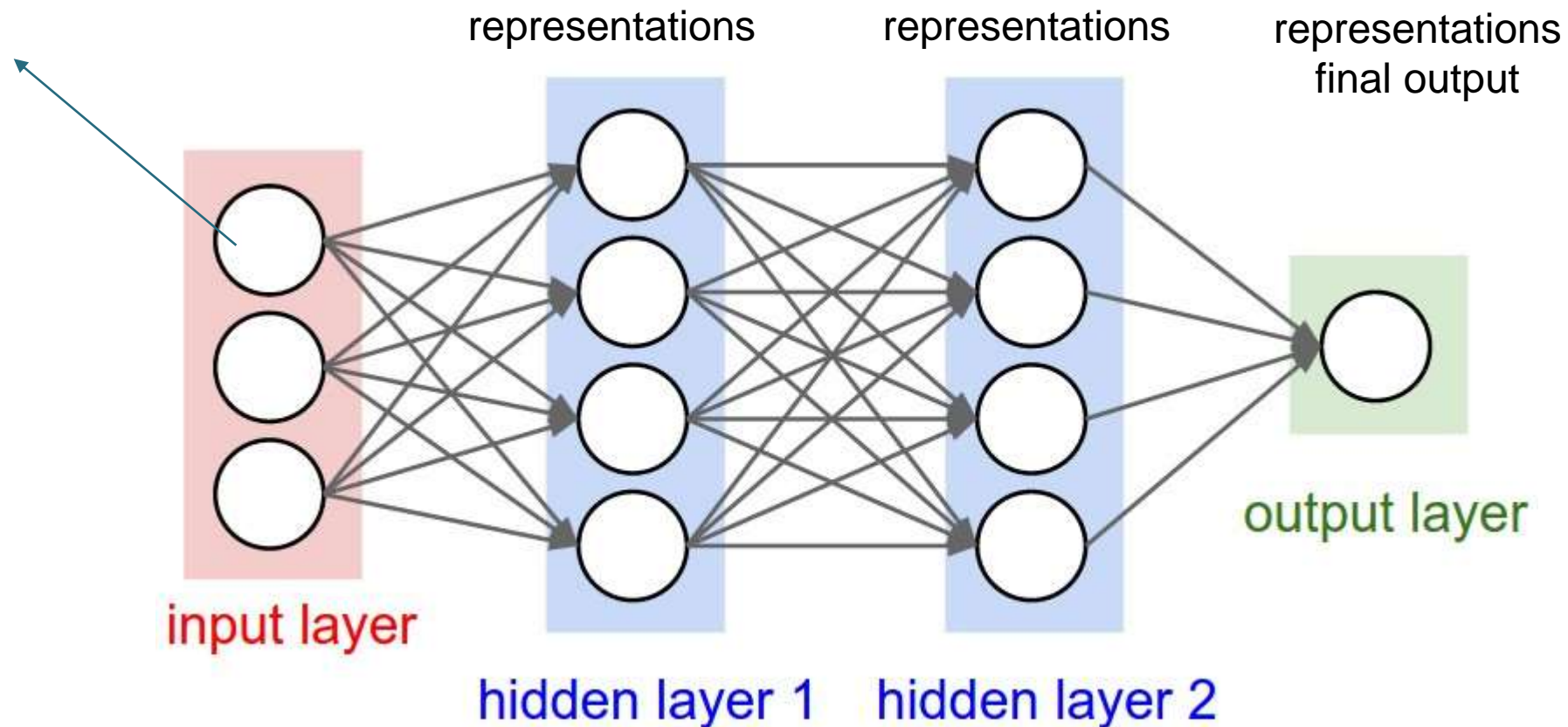
Taken from Deep Learning with Keras book



Vanilla network

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

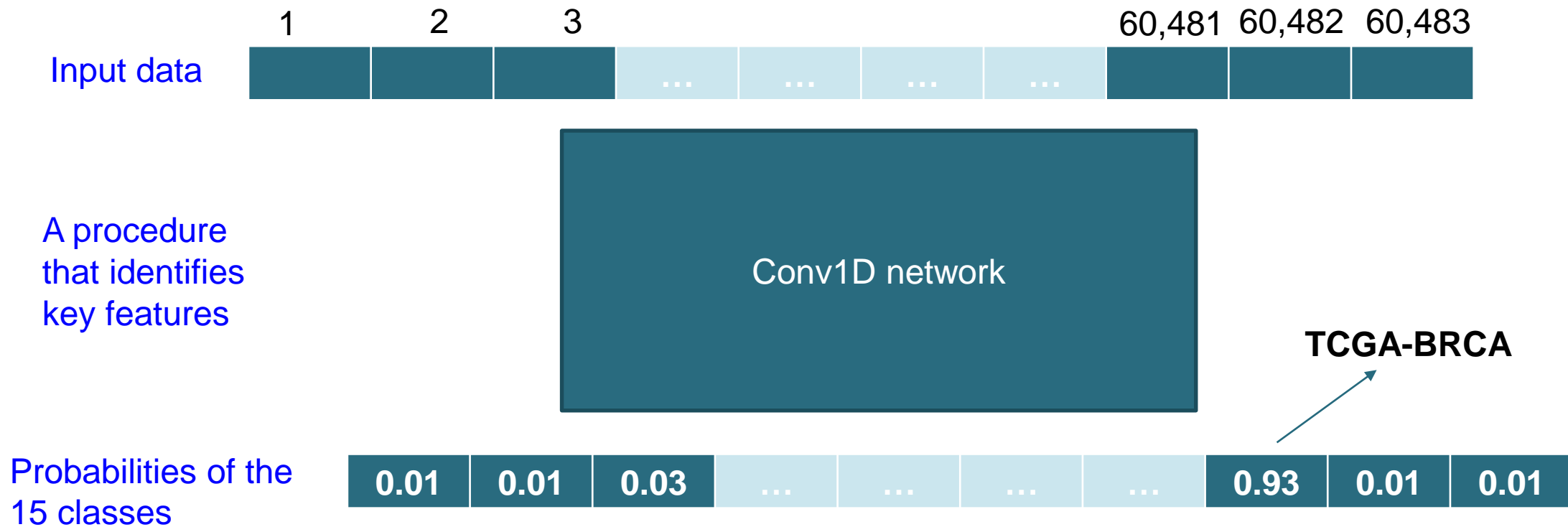
Neuron: a unit that holds a number



[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

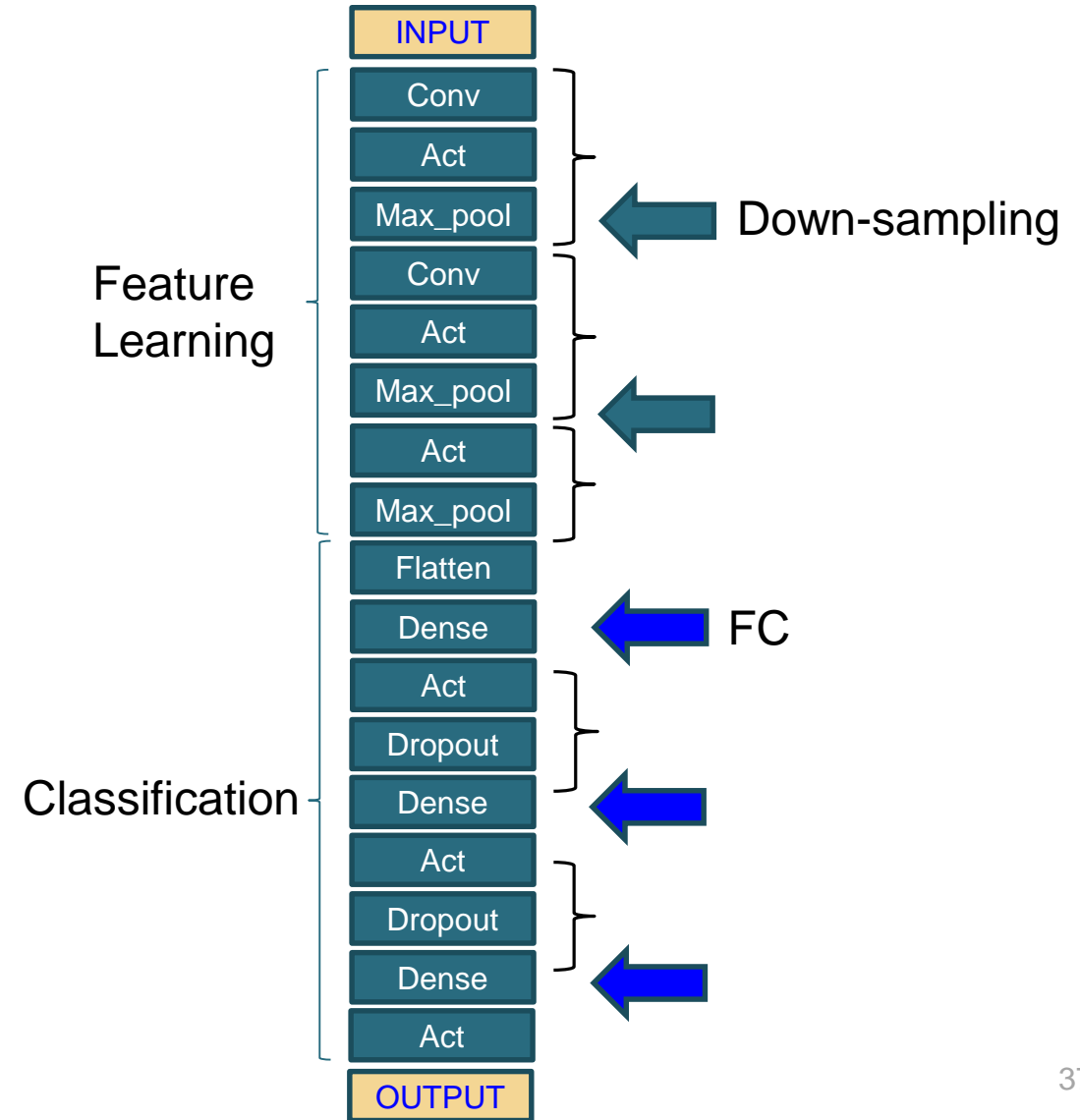


Components of conv1D

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

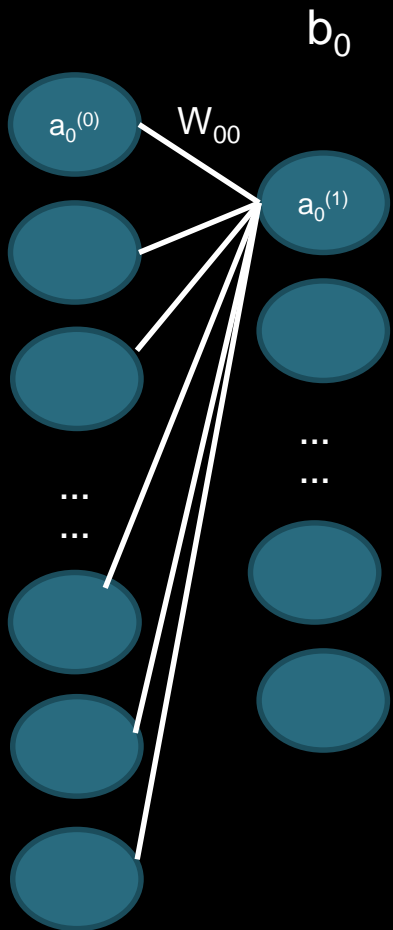
Topology of a network defines a “hypothesis space”

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



1. Activation function

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



ReLU

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

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

+

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

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

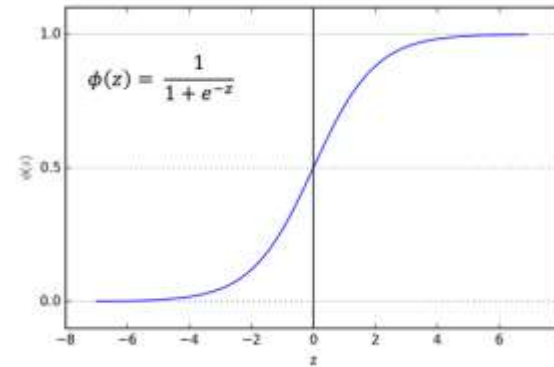
1. Activation Function

- Activation functions are included to create non-linearity

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

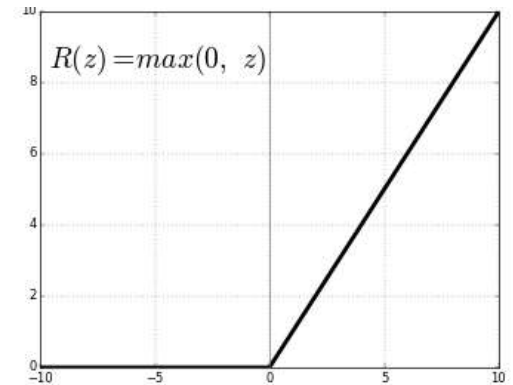
- Sigmoid
- ReLU
- Leaky ReLU
- ELU
- Maxout
- Tanh

Sigmoid

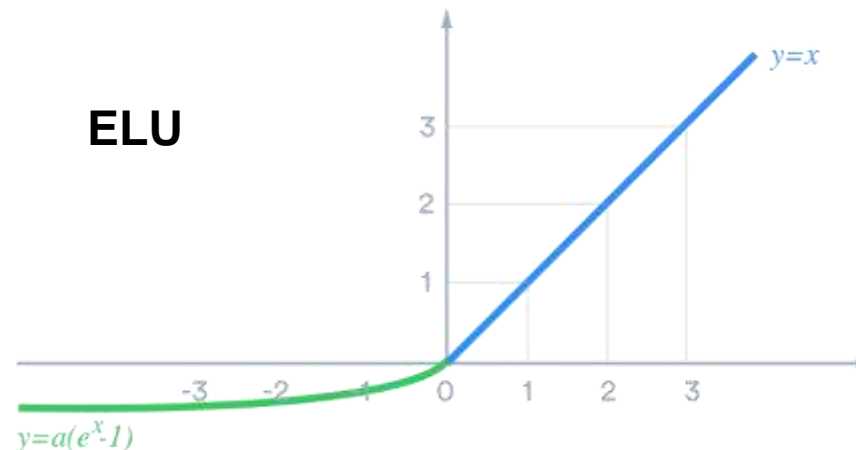


Squashes the #s to [0, 1]

ReLU



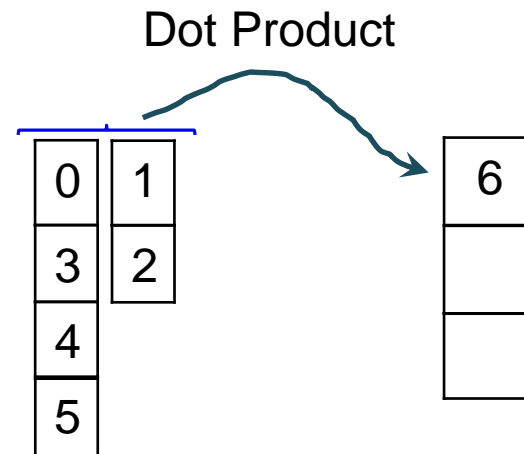
ELU



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

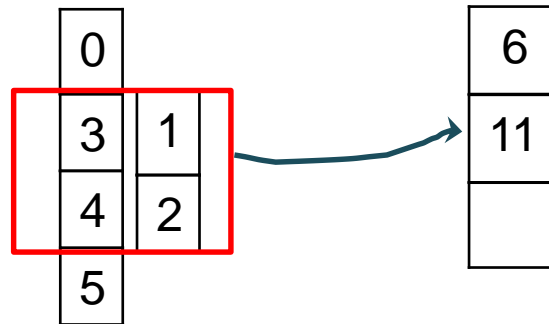


2. Convolution

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

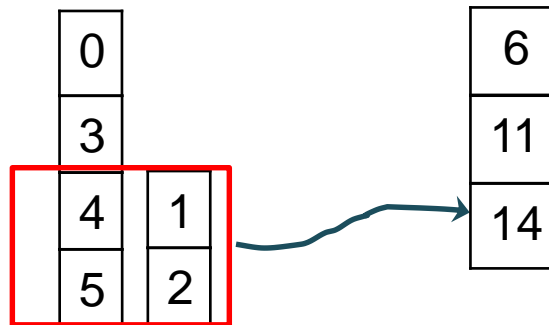


2. Convolution

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

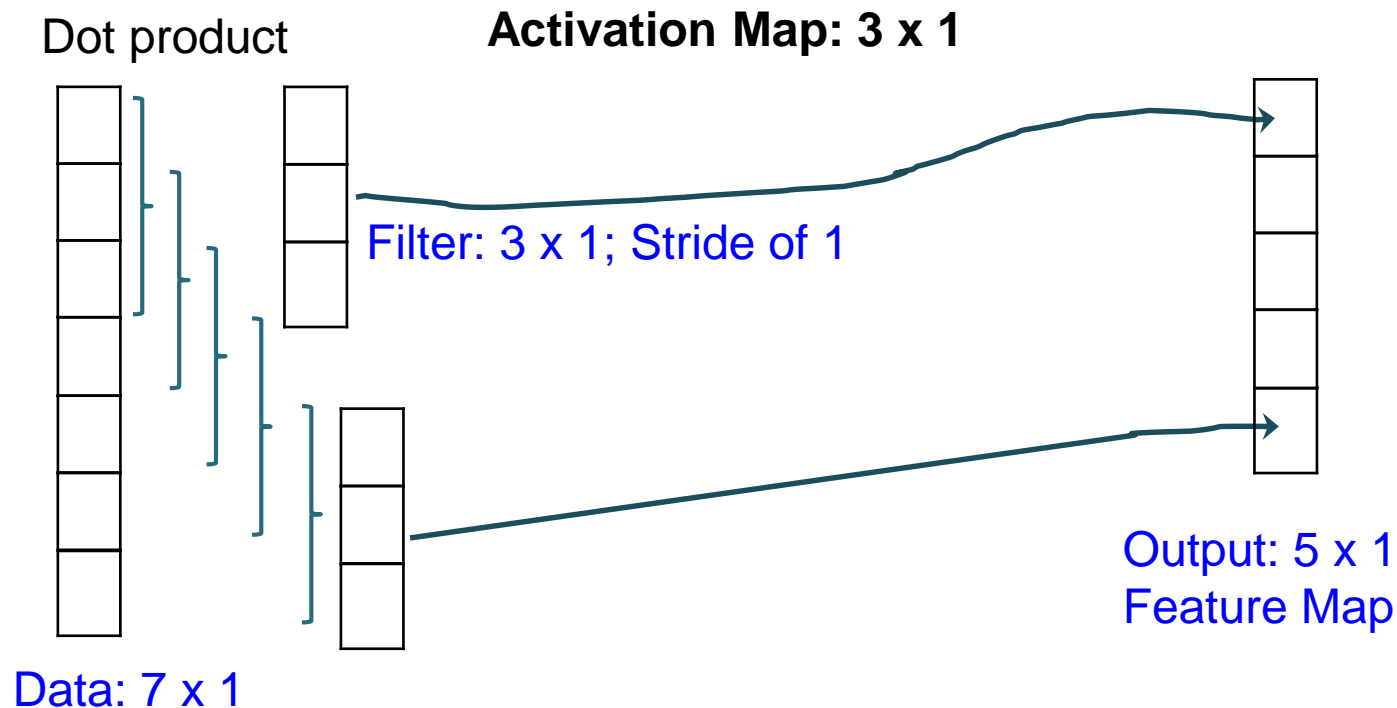
Dot Product



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

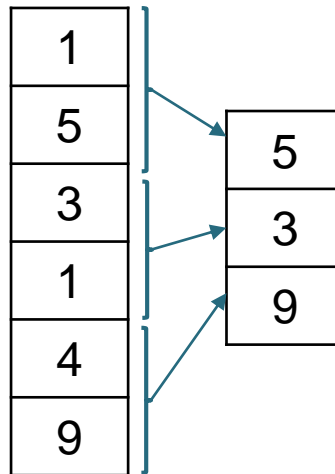
$(7-3)/1+1 = 5$;
zero padding on the border

2. Convolution

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

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

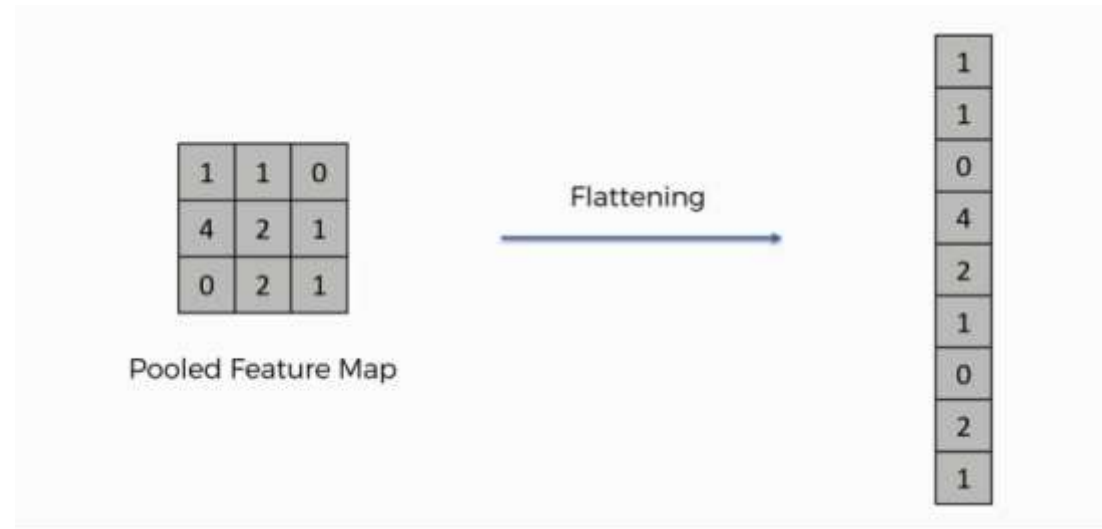
Each of these are activation and we are looking for

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

Common setting for filter 2 or 3

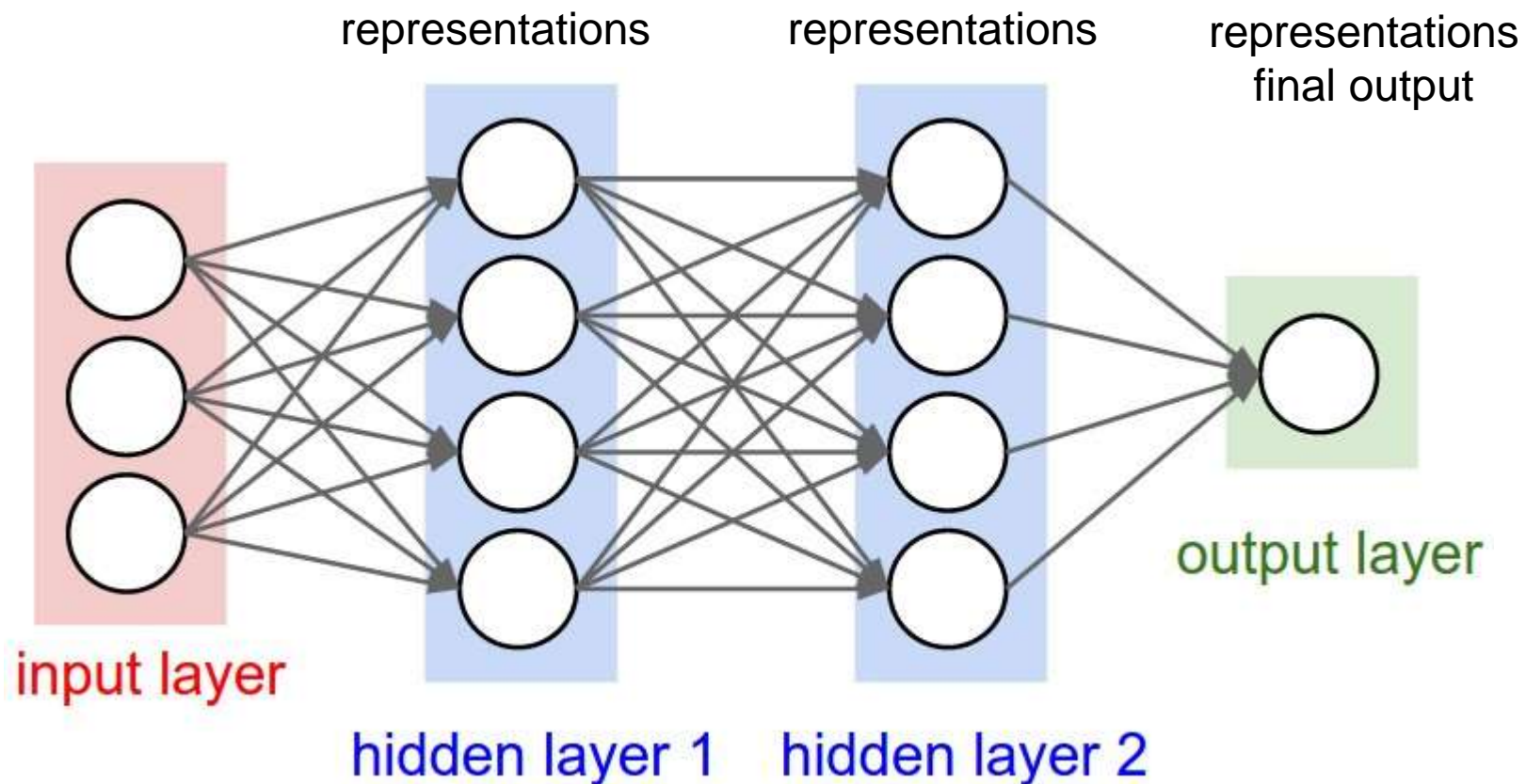
4. Flatten

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



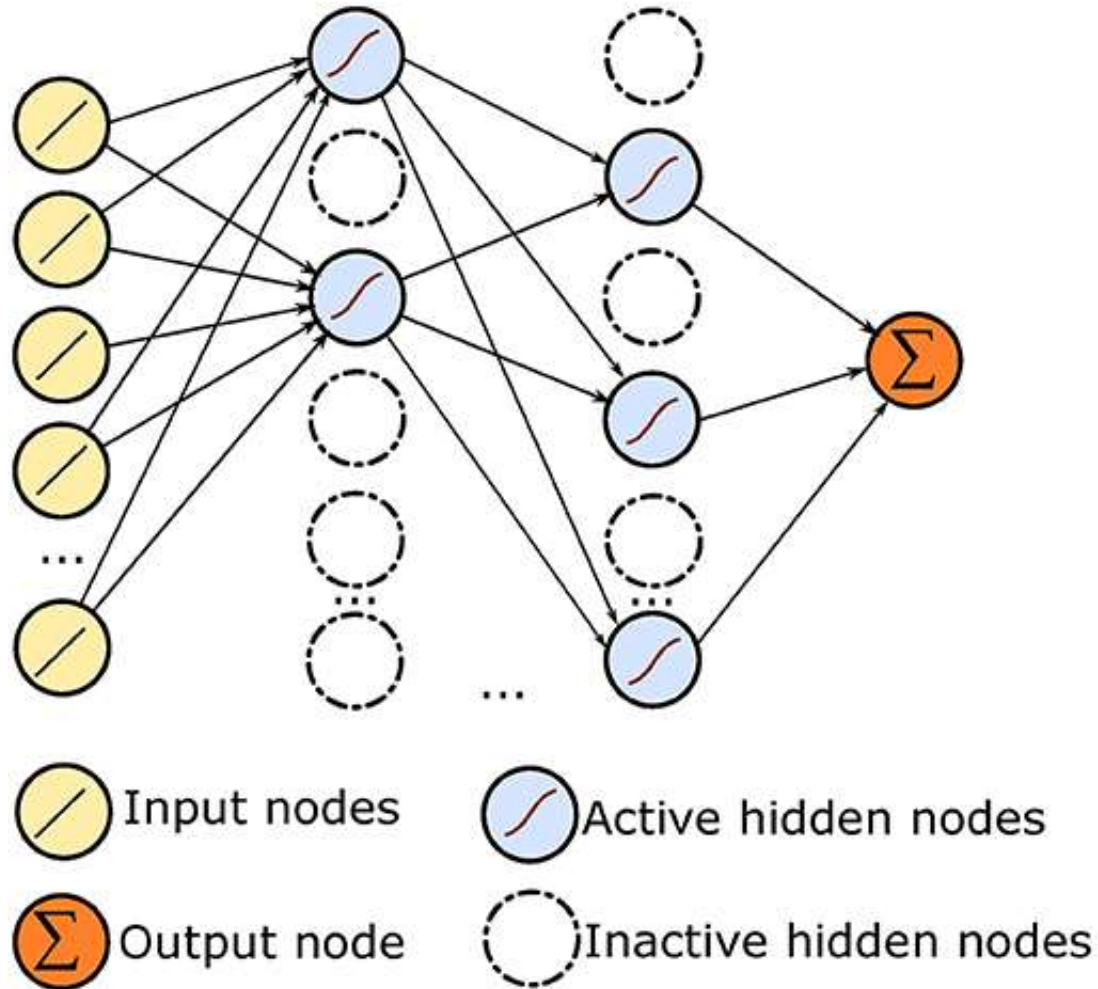
5. Dense

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

```



Code execution and progress

```
Epoch 00001: val_loss improved from inf to 2.56791, saving model to Pilot1.h5

Epoch 2/400
3375/3375 [=====] - 228s 68ms/step - loss: 2.2202 - acc: 0.2821 - val_loss: 1.8444 - val_acc:
Epoch 00002: val_loss improved from 2.56791 to 1.84441, saving model to Pmodel.h5

Epoch 3/400
3375/3375 [=====] - 228s 68ms/step - loss: 1.4736 - accuracy: 0.5206 - val_loss: 0.9554 - val_acc:
Epoch 00003: val_loss improved from 1.84441 to 0.95540, saving model to Pmodel.h5

Epoch 4/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.8795 - accuracy: 0.7058 - val_loss: 0.4835 - val_acc:
Epoch 00004: val_loss improved from 0.95540 to 0.48347, saving model to Pmodel.h5

Epoch 5/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.5968 - accuracy: 0.8107 - val_loss: 0.4083 - val_acc:
Epoch 00005: val_loss improved from 0.48347 to 0.40829, saving model to Pmodel.h5

Epoch 6/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.4529 - accuracy: 0.8519 - val_loss: 0.3236 - val_acc:
Epoch 00006: val_loss improved from 0.40829 to 0.32363, saving model to Pmodel.h5

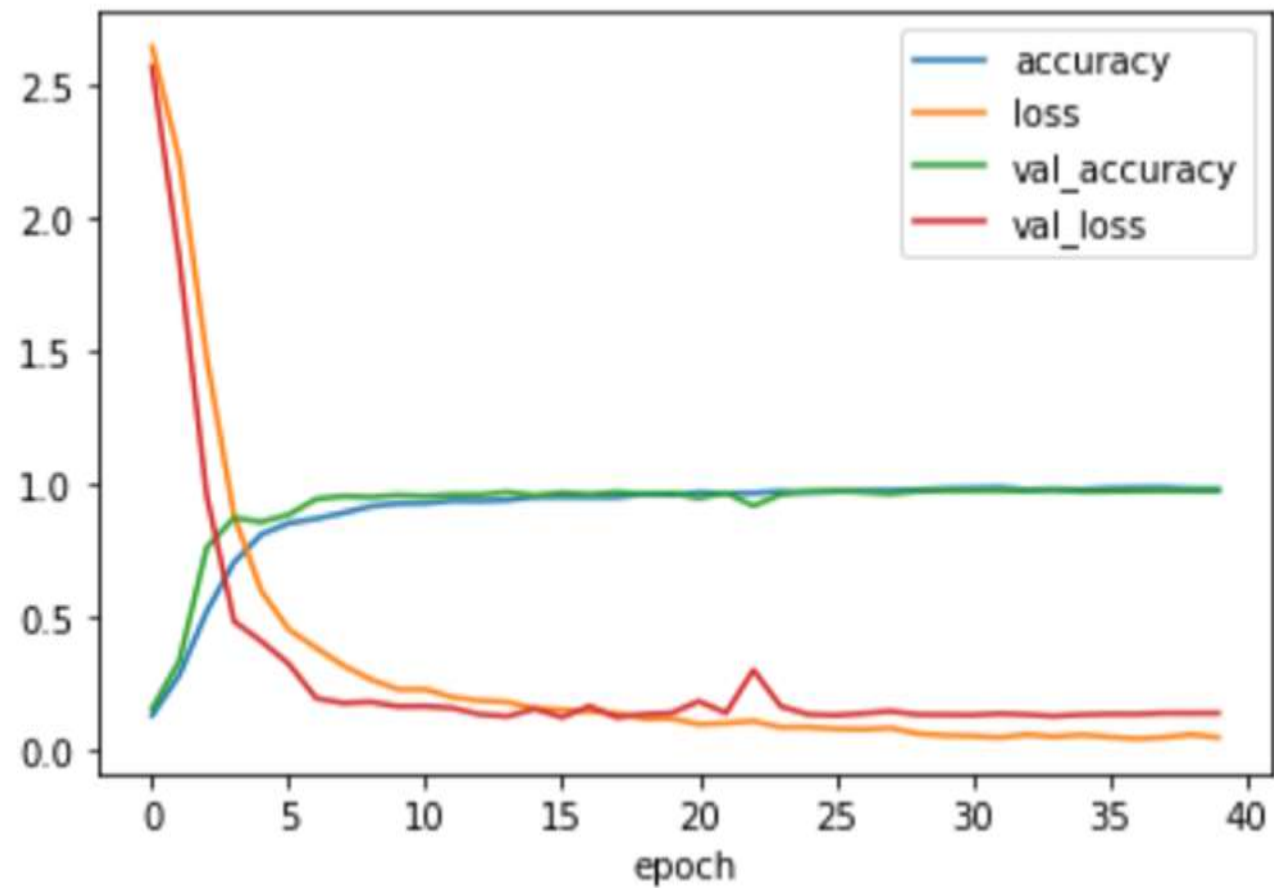
Epoch 7/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.3835 - accuracy: 0.8690 - val_loss: 0.1944 - val_acc:
Epoch 00007: val_loss improved from 0.32363 to 0.19439, saving model to Pmodel.h5

Epoch 8/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.3170 - accuracy: 0.8910 - val_loss: 0.1754 - val_acc:
Epoch 00008: val_loss improved from 0.19439 to 0.17536, saving model to Pmodel.h5

Epoch 9/400
3375/3375 [=====] - 228s 67ms/step - loss: 0.2647 - accuracy: 0.9156 - val_loss: 0.1800 - val_acc:
Epoch 00009: val_loss did not improve from 0.17536

Epoch 10/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.2276 - accuracy: 0.9265 - val_loss: 0.1632 - val_acc:
Epoch 00010: val_loss improved from 0.17536 to 0.16323, saving model to Pmodel.h5
```

Model Performance



Thank you!

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

Questions/Comments

S. Ravichandran
ravichandrans@mail.nih.gov

