

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
 - Colleagues who reviewed the material

Feel free to follow-along

CBIIT

- <https://cbiit.github.io/sdsi/workshops> (landing site; creation in progress)

Github

- <https://github.com/ravichas/ML-TC1> (in progress)

Introduction

- **This is part of the NCI-DOE knowledge/capability transfer efforts**
- **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 the test-case (code/scripts) and tune it to your needs**
- **We want to hear from you**

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

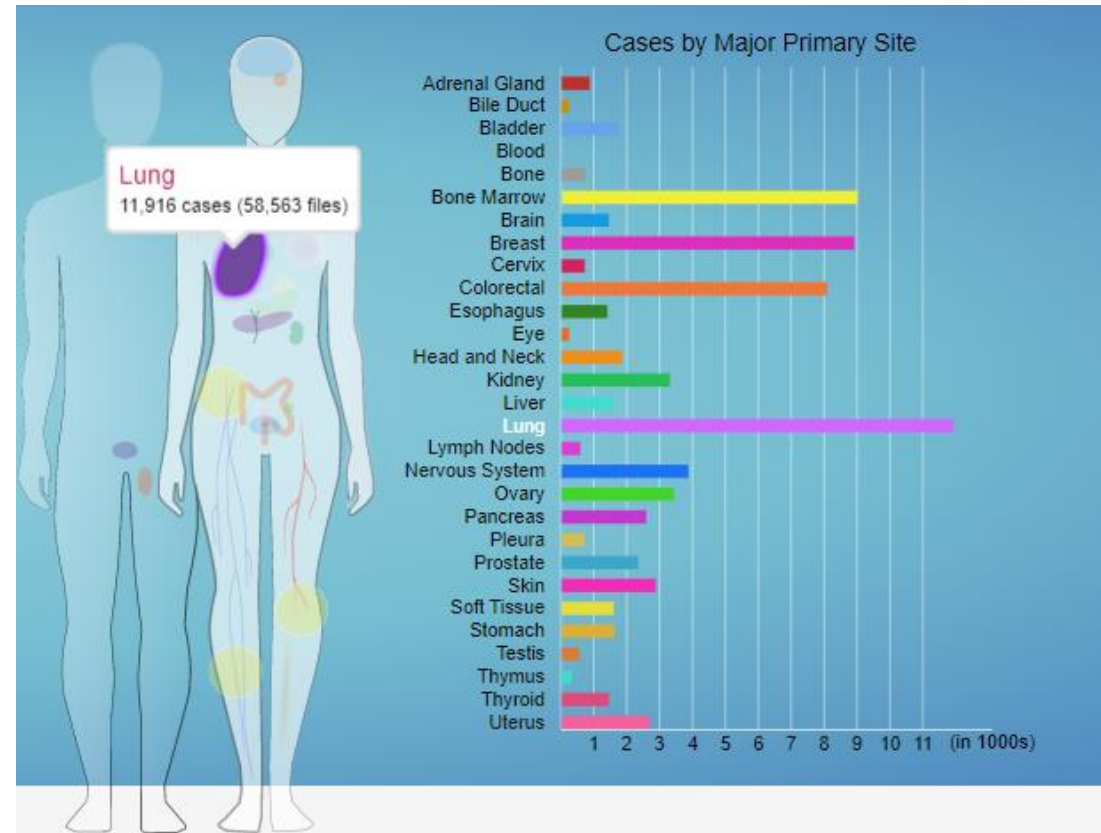
Goal(s)/Questions

- **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**
 - ~1.8M new cancer cases are expected
 - ~600K deaths will occur

Figure from Genomic Data Commons



Expected New Cases/Deaths in 2020

New Cancer Cases

Between 2010 and 2020, we expect the number of new cancer cases in the United States to go up about 24% in men to more than 1 million cases per year, and by about 21% in women to more than 900,000 cases per year.

US population gender	Cancers that are expected to increase
Men	Prostrate, Kidney, Liver and Bladder
Women	Lung, Breast, Uterine and Thyroid

Dynamic genomic changes result in Cancer

Somatic alterations in oncogenes are the source of Transcript alterations

Article

Genomic basis for RNA alterations in cancer

<https://doi.org/10.1038/s41586-020-1970-0>

Received: 29 March 2018

Accepted: 11 December 2019

Published online: 5 February 2020

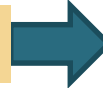
Transcript alterations often result from somatic changes in cancer genomes. Various forms of RNA alterations have been described in cancer, including overexpression, altered splicing and gene fusions; however, it is difficult to attribute these to underlying genomic changes owing to heterogeneity among patients and tumor types, and the relatively small cohorts of patients for whom samples have been analyzed by both transcriptome and whole-genome sequencing.

Somatic alterations in the human cancer genome

Barbara Weir, Xiaojun Zhao, and Matthew Meyerson*

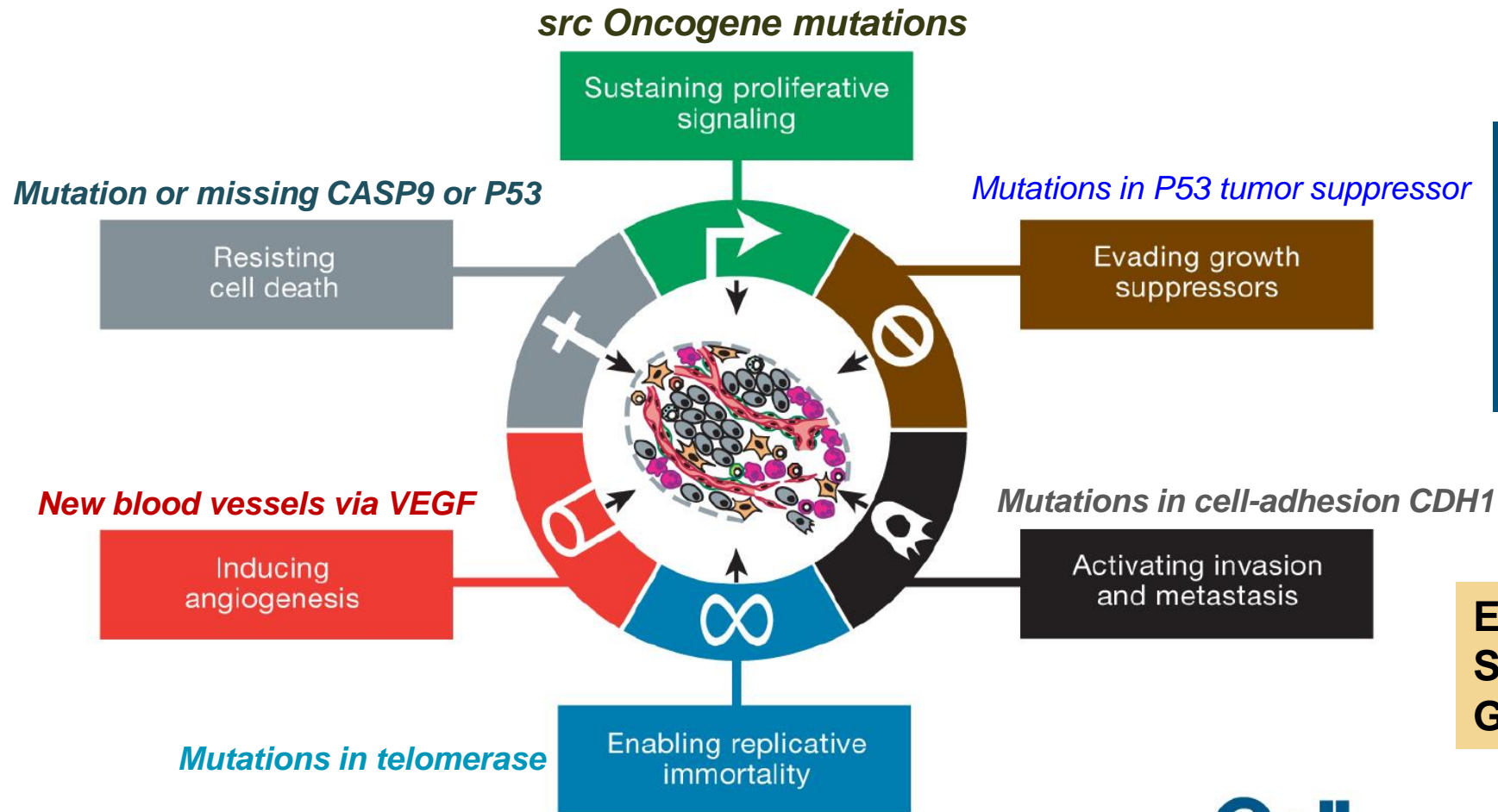
CANCER CELL : NOVEMBER 2004 · VOL. 6 · COPYRIGHT © 2004 CELL PRESS

Expression changes in oncogenes; What type of changes?



Hallmarks of cancer: Acquired capabilities (mutations) that drive cancer

Hallmarks of Cancer: The Next Generation



Hanahan and Weinberg, 2011



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

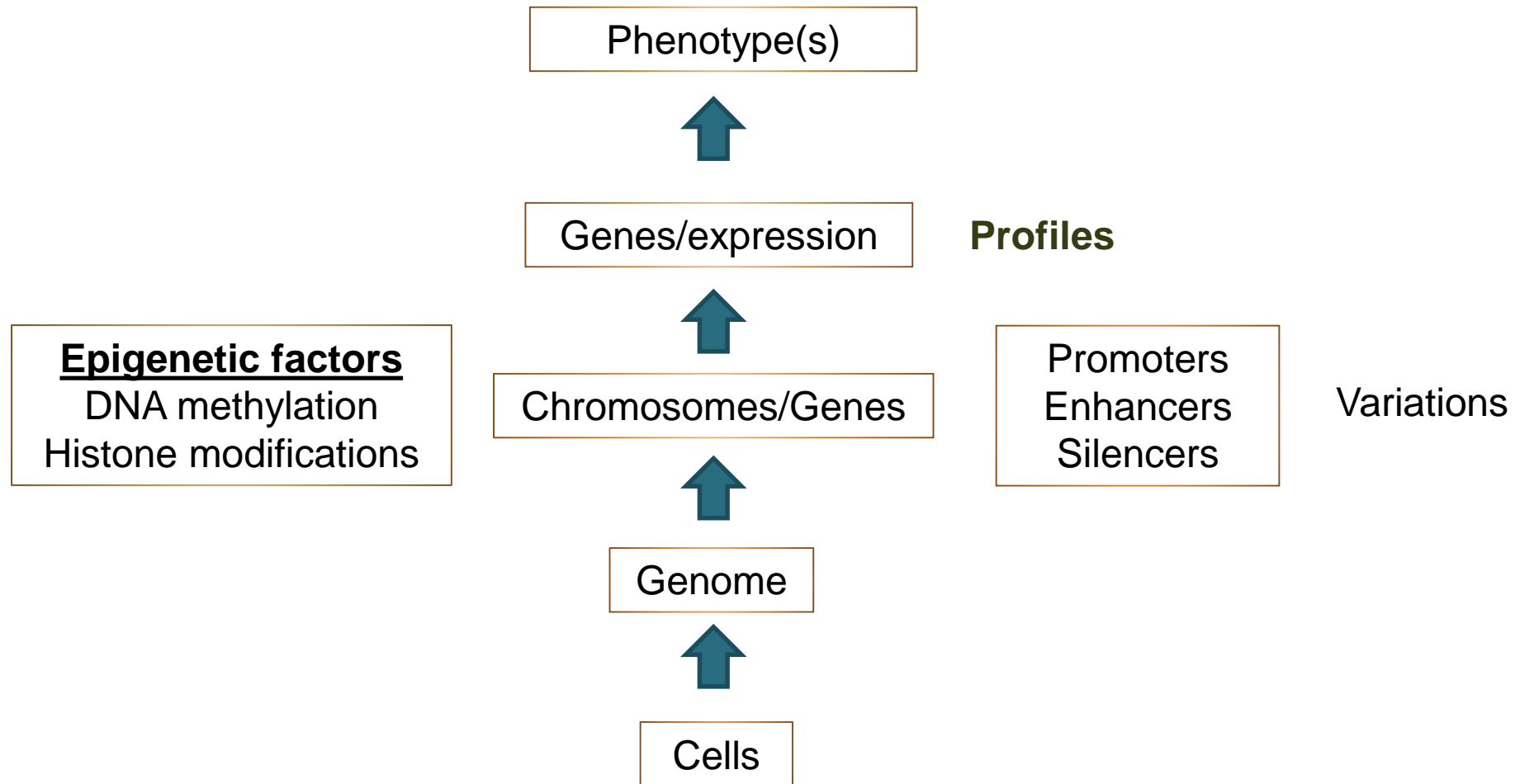
The Hallmarks of Cancer

Douglas Hanahan • Robert A Weinberg

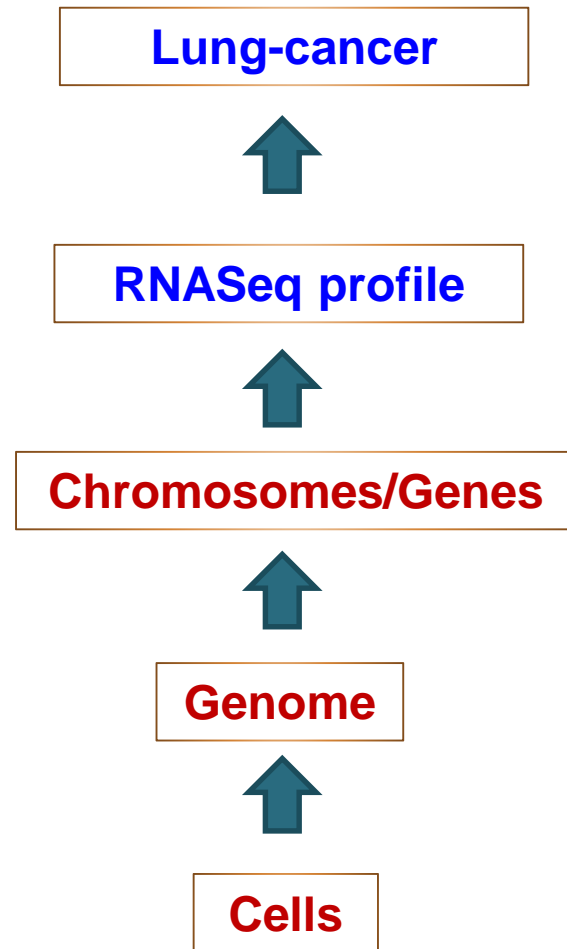
Open Archive • DOI: [https://doi.org/10.1016/S0092-8674\(00\)81683-9](https://doi.org/10.1016/S0092-8674(00)81683-9)

Expression changes in oncogenes;
Six capabilities; Overview of
Genotype/phenotypes?

Influence of genomic features on phenotypes: An overview



Influence of genomic features on phenotypes: An overview



Diagnosis/treatment vs Prediction →

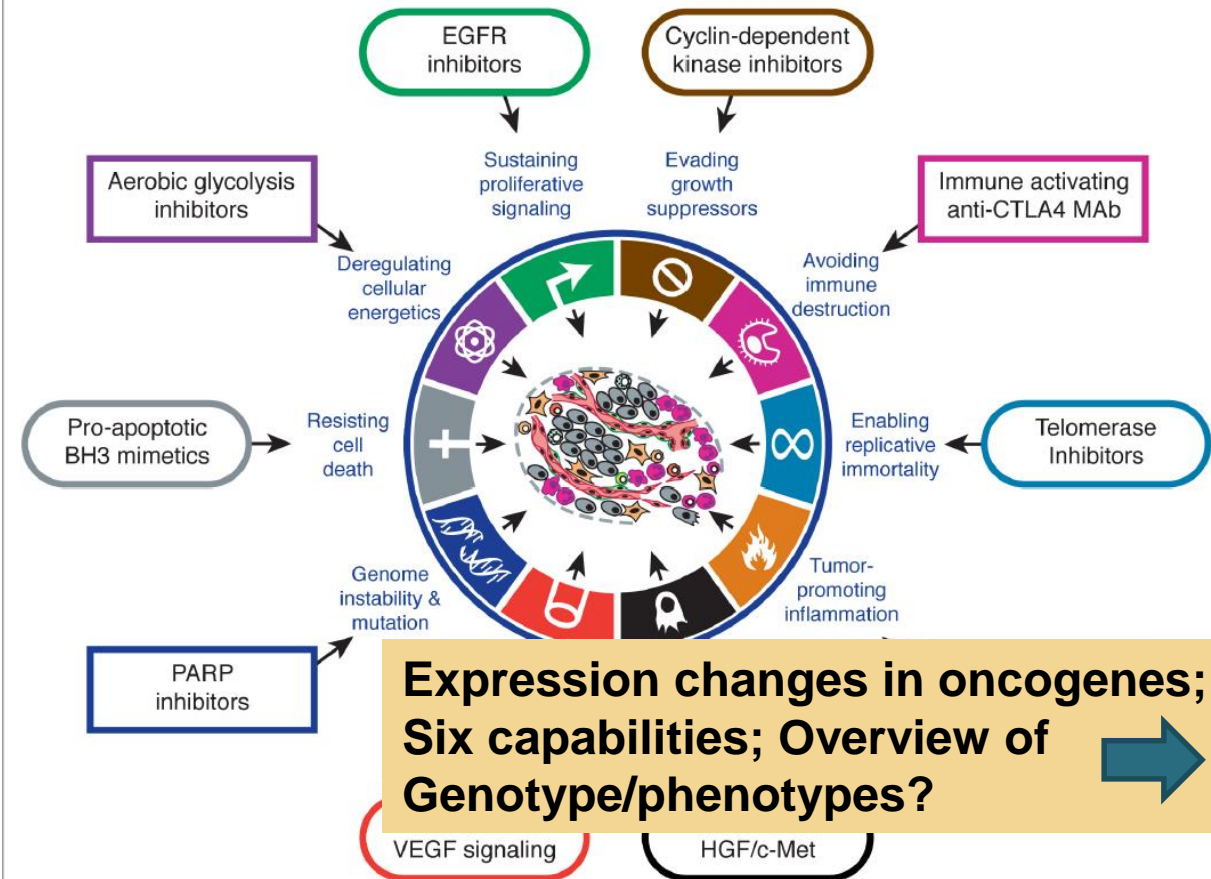
Treatment vs Type-Prediction

- **Treatment**

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

- **Detecting Type**

- “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
- Complex
- Multi-gene centric



Hanahan and Weinberg, 2011

The architecture of occurring genetic aberrations such as somatic mutations, CNV, changed gene expression profiles, and different epigenetic alterations, is unique for each type of cancer

DOI: [10.5114/wo.2014.47136](https://doi.org/10.5114/wo.2014.47136)

PERSPECTIVE

Understanding Genotype-Phenotype Effects in Cancer via Network Approaches

Yoo-Ah Kim, Dong-Yeon Cho, Teresa M. Przytycka*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

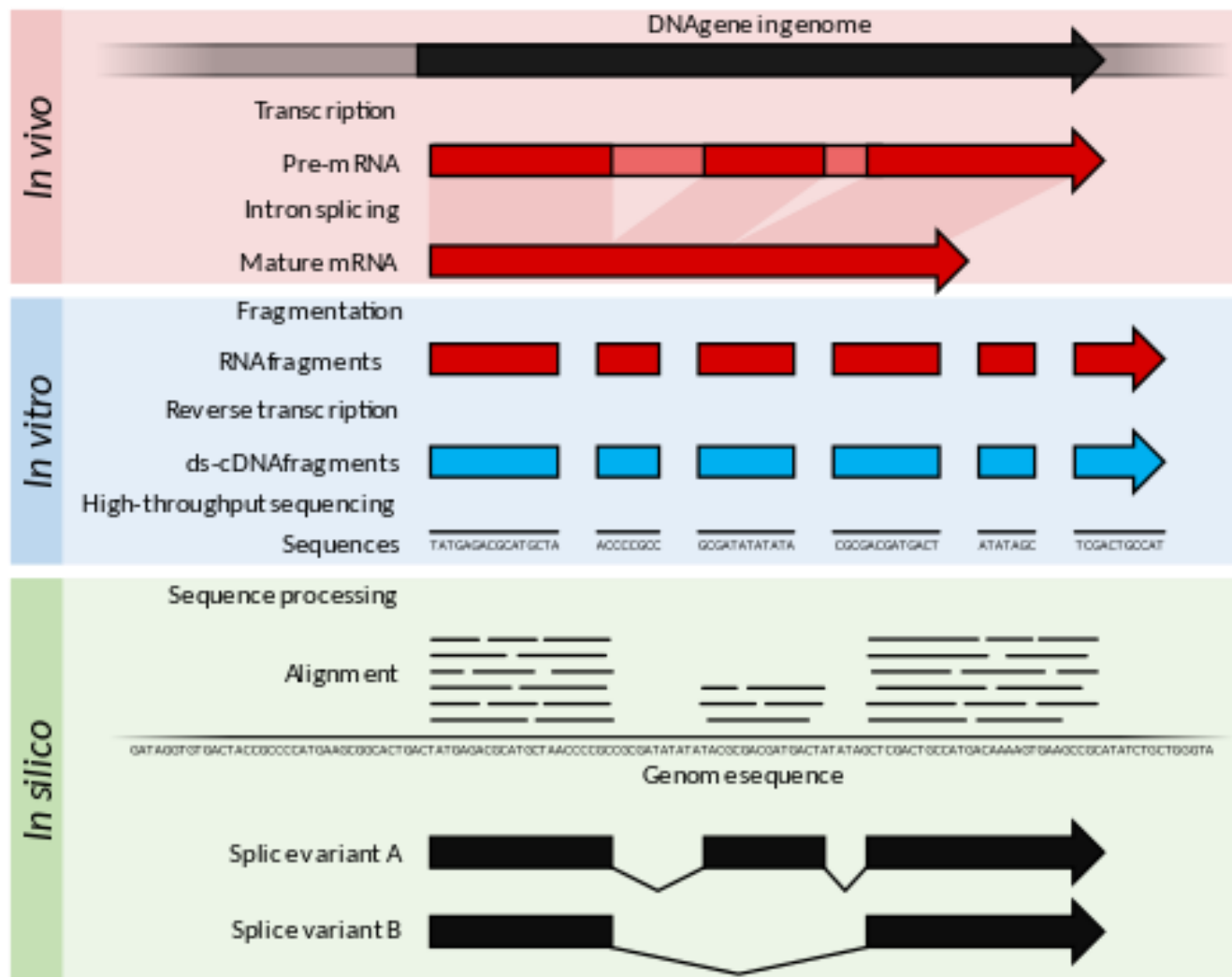
* przytyck@ncbi.nlm.nih.gov

Author Summary

Cancer is now increasingly studied from the perspective of dysregulated pathways, rather than as a disease resulting from mutations of individual genes. A pathway-centric view acknowledges the heterogeneity between genomic profiles from different cancer patients while assuming that the mutated genes are likely to belong to the same pathway and cause similar disease phenotypes. Indeed, network-centric approaches have proven to be helpful for finding genotypic causes of diseases, classifying disease subtypes, and identifying drug targets. In this review, we discuss how networks can be used to help understand patient-to-patient variations and how one can leverage this variability to elucidate interactions between cancer drivers.

What kind of data do we need?

NGS



READS

NGS

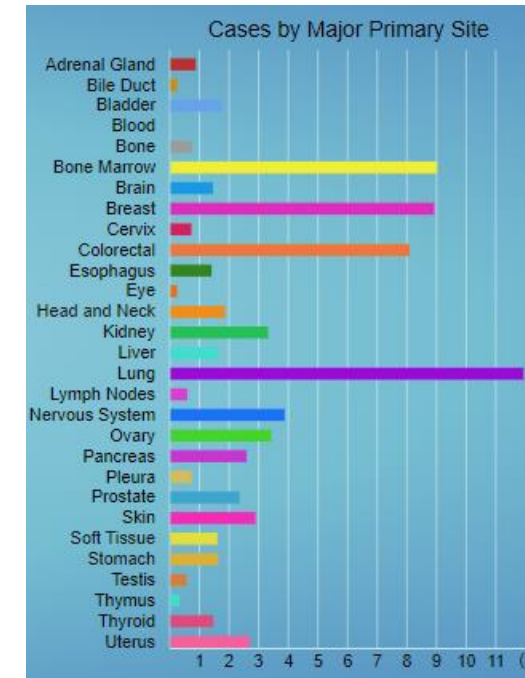
Data source: The Cancer Genome Atlas (TCGA)

- 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



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

 Projects

 Exploration

 Analysis

 Repository

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

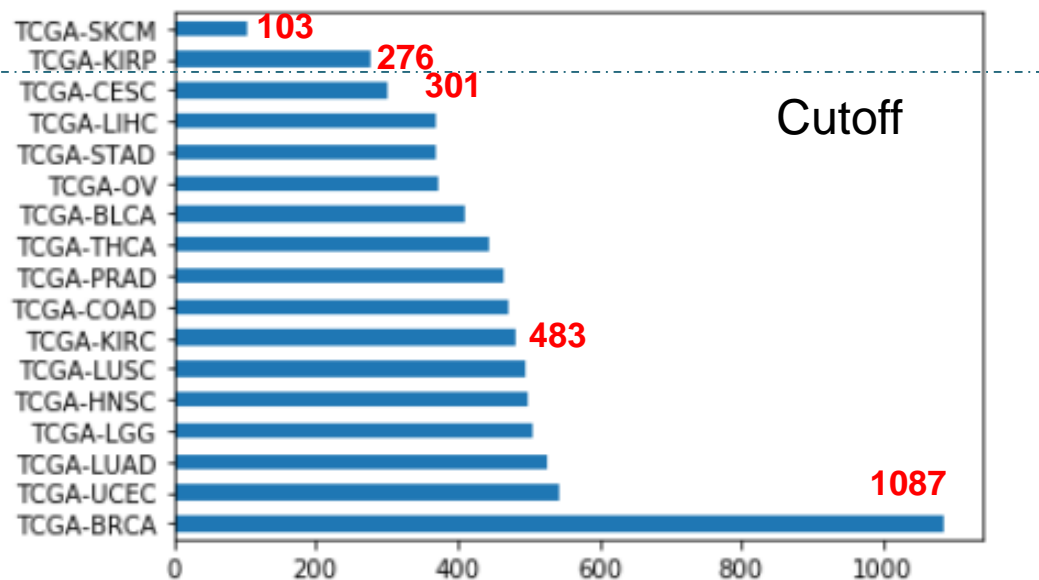
Alignment to Ref
Genome (SAM/BAM)

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](#)

Case	Expression
ENGG0000024268.2	1658.464179
ENGG0000012712.3	40.213443
ENGG0000016017.5	53440.1006
ENGG0000014824.1	10.000000
ENGG0000016708.2	6818.45626
ENGG00000146083.0	255939.2351
ENGG0000025275.4	0
ENGG0000014668.12	254.947378
ENGG0000018922.12	4986858.658
ENGG000001259883.1	6108.995092
ENGG000002119881.3	0
ENGG000001368975.2	0
ENGG00000125788.1	0
ENGG00000134108.11	967330.2656
ENGG00000126308.1	3484.02773
ENGG0000016708.17	1484.4697
ENGG0000016700.7	2267.41208
ENGG00000123493.42	2802.45035
ENGG000001240923.1	330.5246740
ENGG000001240923.1	155836.932
ENGG00000127526.1	0
ENGG000001234884.1	0
ENGG00000126040.1	394.455669
ENGG000001234884.1	1583.31282
ENGG000001234934.1	0
ENGG000001282141.8	45338.4608
ENGG000001269416.4	119.087054

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

Genes	Expression
ENSG00000242268.2	1658.464619
ENSG00000270123.1	460.234143
ENSG00000251718.3	52440.13006
ENSG00000271842.1	18.000000
ENSG00000273025.1	68185.46526
ENSG00000246083.10	255939.2515
ENSG00000225274.5	0.000000
ENSG00000254848.12	12.9477378
ENSG00000280426.12	4086556.58
ENSG00000259813.1	6108.899052
ENSG00000231988.3	0.000000
ENSG00000234972.2	0.000000
ENSG00000270881.1	0.000000
ENSG00000214108.11	973730.2056
ENSG00000200898.1	3488.027373
ENSG00000271217.12	227.4017
ENSG00000287702.7	22647.4018
ENSG00000241042.14	2082.245629
ENSG00000240423.1	310.5246740
ENSG00000246503.1	305525.3276
ENSG00000276561.1	0.000000
ENSG00000214851.1	0.000000
ENSG00000204041.1	394.315660
ENSG00000211195.1	158.312583
ENSG00000243040.1	198.472583
ENSG00000282148.8	45338.60548
ENSG00000249416.4	118.087054

Genome	Population
ENSGM00000242682	1658.484179
ENSGM00000247123	460.234343
ENSGM00000250175	52440.10096
ENSGM00000250175	0
ENSGM00000278175	68165.48626
ENSGM00000280230	205969.25181
ENSGM00000292794	0
ENSGM00000300881	147.947768
ENSGM00000302412	490856.658
ENSGM00000308831	6128.99052
ENSGM00000319813	0
ENSGM00000325542	0
ENSGM00000327081	0
ENSGM00000340281	957330.256
ENSGM00000360810	3484.027773
ENSGM00000372137	4545.9267
ENSGM00000373477	27571.4008
ENSGM00000394432	2082.24953
ENSGM00000400343	110.524679
ENSGM00000401919	153681.1261
ENSGM00000716161	0
ENSGM00000748811	0
ENSGM00000804041	394.475569
ENSGM00000810315	553.332582
ENSGM00000830044	0
ENSGM00000824148	45338.60848
ENSGM00000945614	1184.07054

Genes	Expression
ENSG00000242582	1508.404179
ENSG00000270112.3	406.234343
ENSG00000270112.5	52440.10096
ENSG00000270112.6	52440.10096
ENSG00000270125.7	688.65
ENSG0000046083.10	25559.6251
ENSG00000252784.1	105.94
ENSG00000258456.12	104.947378
ENSG00000258456.12	698.656168
ENSG00000258811.1	6108.1995
ENSG00000251981.3	6108.1995
ENSG00000304972.5	0
ENSG00000304972.5	0
ENSG00000314018.11	95730.2056
ENSG00000330891.1	3484.0273
ENSG00000272117.17	41485.9037
ENSG00000280770.2	285.714308
ENSG00000280770.2	285.714308
ENSG00000280770.3	130.524679
ENSG00000260422.9	15581.9216
ENSG00000278561.6	0
ENSG00000248121.1	0
ENSG00000230401.1	394.475569
ENSG00000231105.1	4531.82048
ENSG0000030441.1	4531.82048

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
ENSG0000028121.1	68285.4526
ENSG0000024298.2	25099.2351
ENSG0000025277.4	0
ENSG0000025486.12	104.9473768
ENSG00000219842.12	406856.458
ENSG0000021085.1	6108.19052
ENSG0000021038.3	0
ENSG0000020879.2	0
ENSG00000201788.1	0
ENSG00000214108.11	90730.2056
ENSG0000021028.1	2484.03713
ENSG00000271217.17	41485.9507
ENSG00000207780.7	226712.4208
ENSG0000020484.2	2982.24055
ENSG00000240423.1	305.5246749
ENSG00000200342.9	121863.1216
ENSG00000271816.1	0
ENSG00000214081.1	0
ENSG00000218046.1	394.475669
ENSG00000211105.1	1183.112582
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

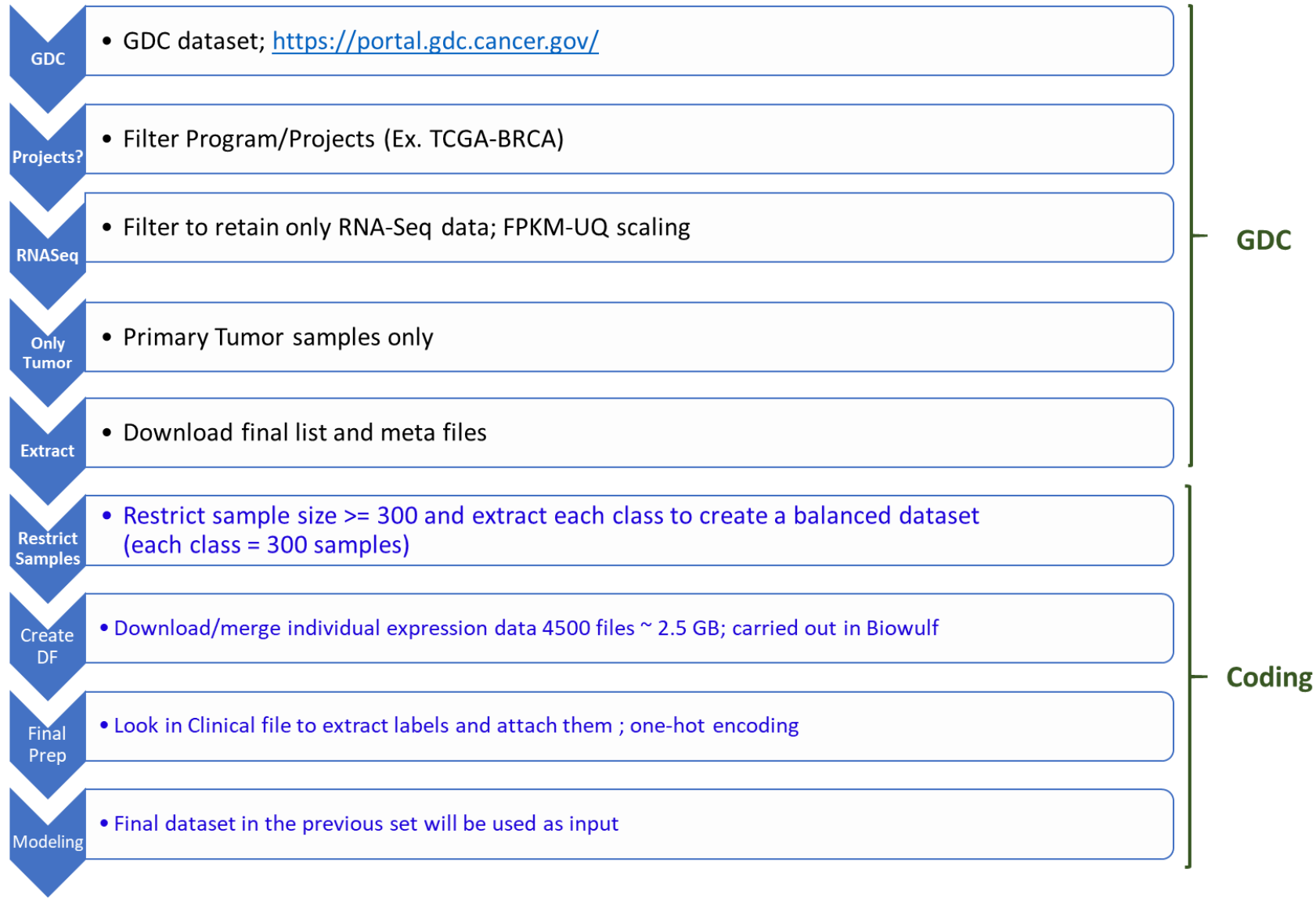
Ali Mortazavi^{1,2}, Brian A Williams^{1,2}, Kenneth McCue¹, Lorian Schaeffer¹ & Barbara Wold¹

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, 15

```
>>> 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**



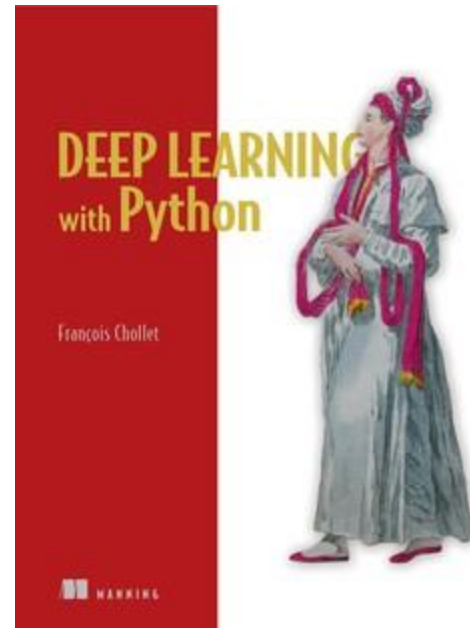
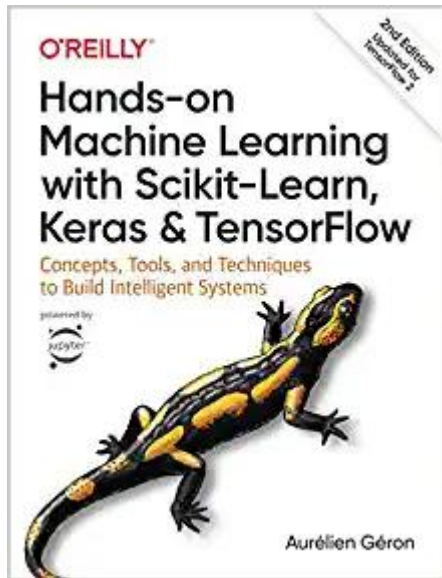
- **Will use Jupyter Notebook for sharing the code**
 - With little effort one can convert the Python code into R and still use Jupyter Notebook

To be continued after hands-on

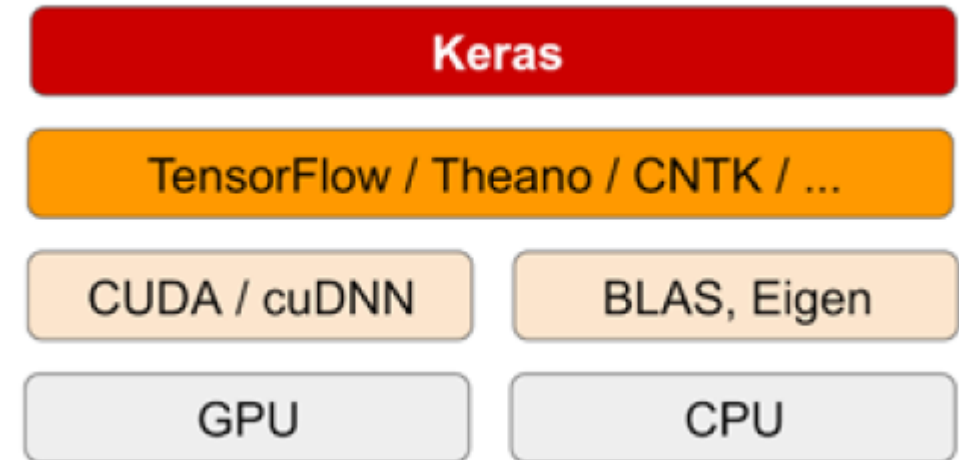
<https://github.com/ravichas/ML-TC1>

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

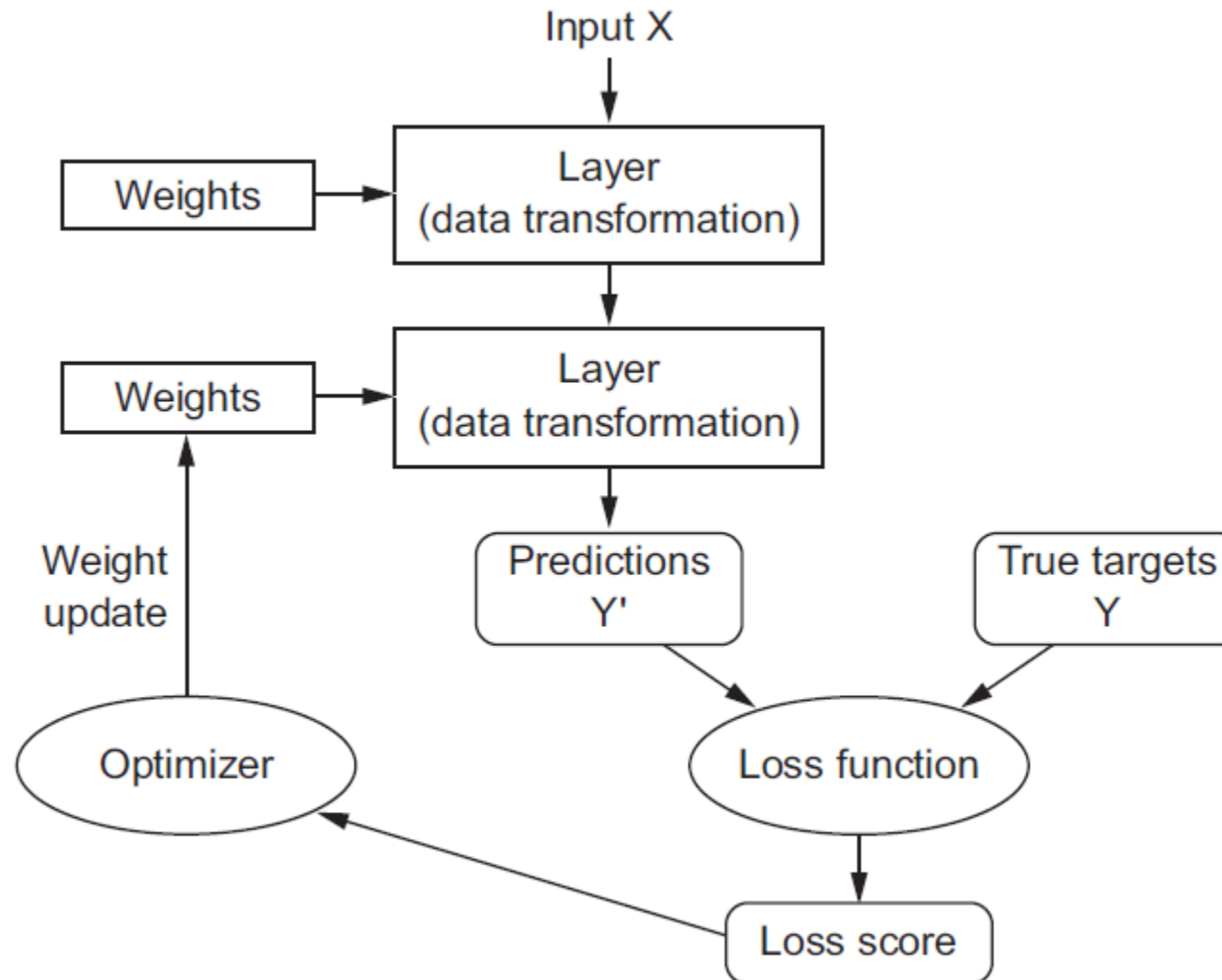
Figure from Deep Learning with Python

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

Deep Learning Procedure

Taken from Deep Learning with Keras book



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**

A Simple Network

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

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

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(m):  
    L = 0.1 * m + 0.38  
    return(L)
```

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

```
Model ← lm( L ~ m)
```

- Neural Network using the data learn the algorithm.

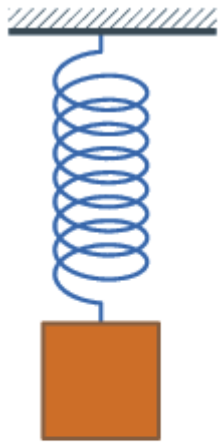
INPUT

ALGORITHM

OUTPUT

A Simple Network

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

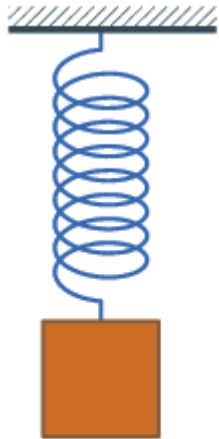


M	L
Input	Output
0.125	0.39
0.25	0.40
0.5	0.43
1	0.48
2	0.58
3	???

[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](#)

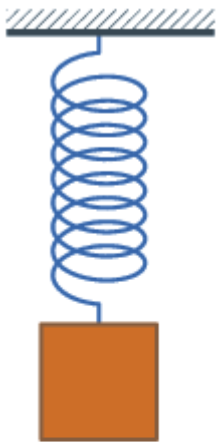
Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003

A Simple Network

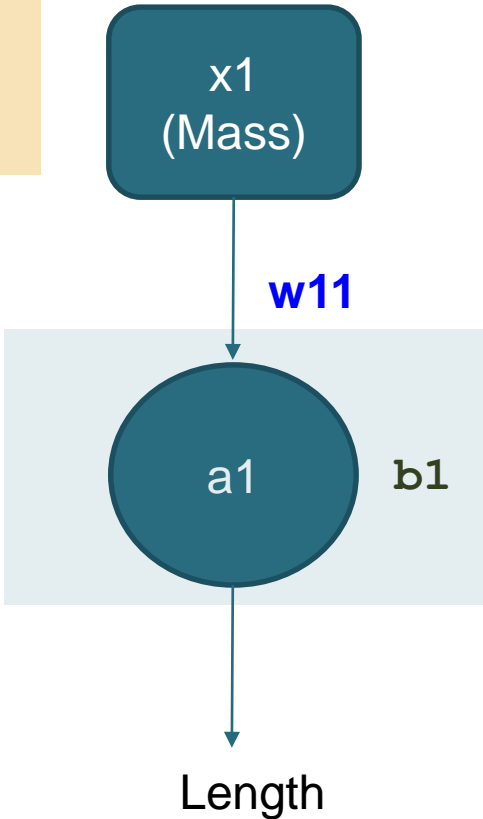
$$a1 = x1 * w11 + b1$$

$$L = M * 0.1 + 0.38$$

[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)]`

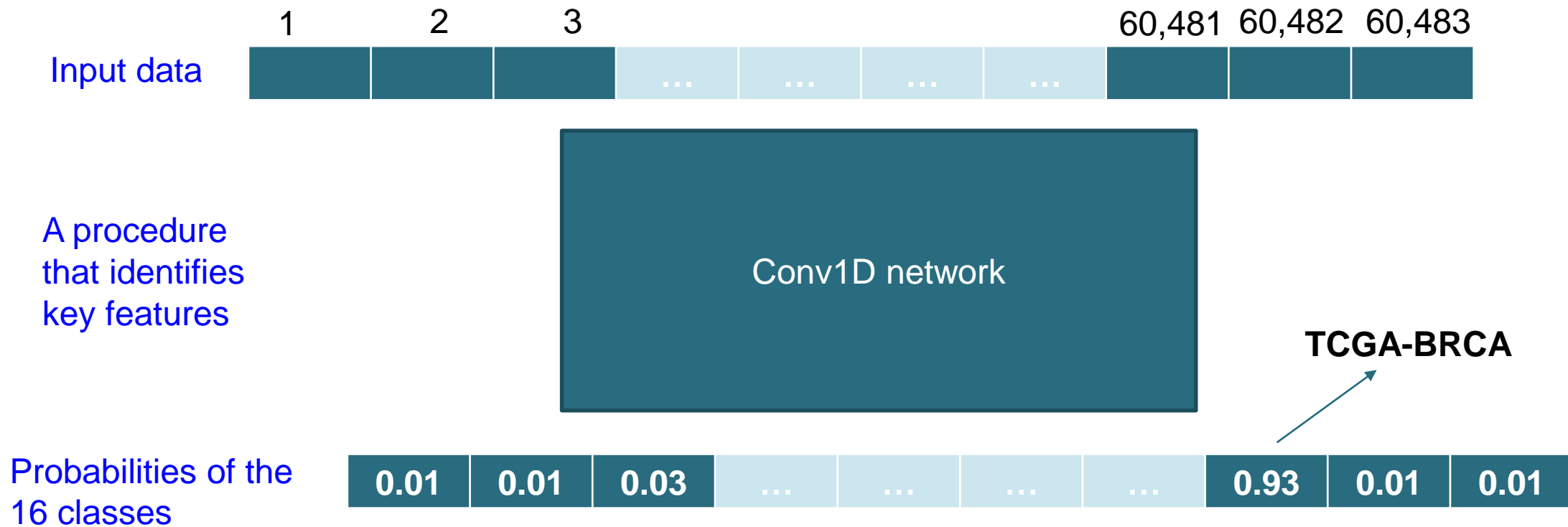
Based on Mary Attenborough, in [Mathematics for Electrical Engineering and Computing](#), 2003

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

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

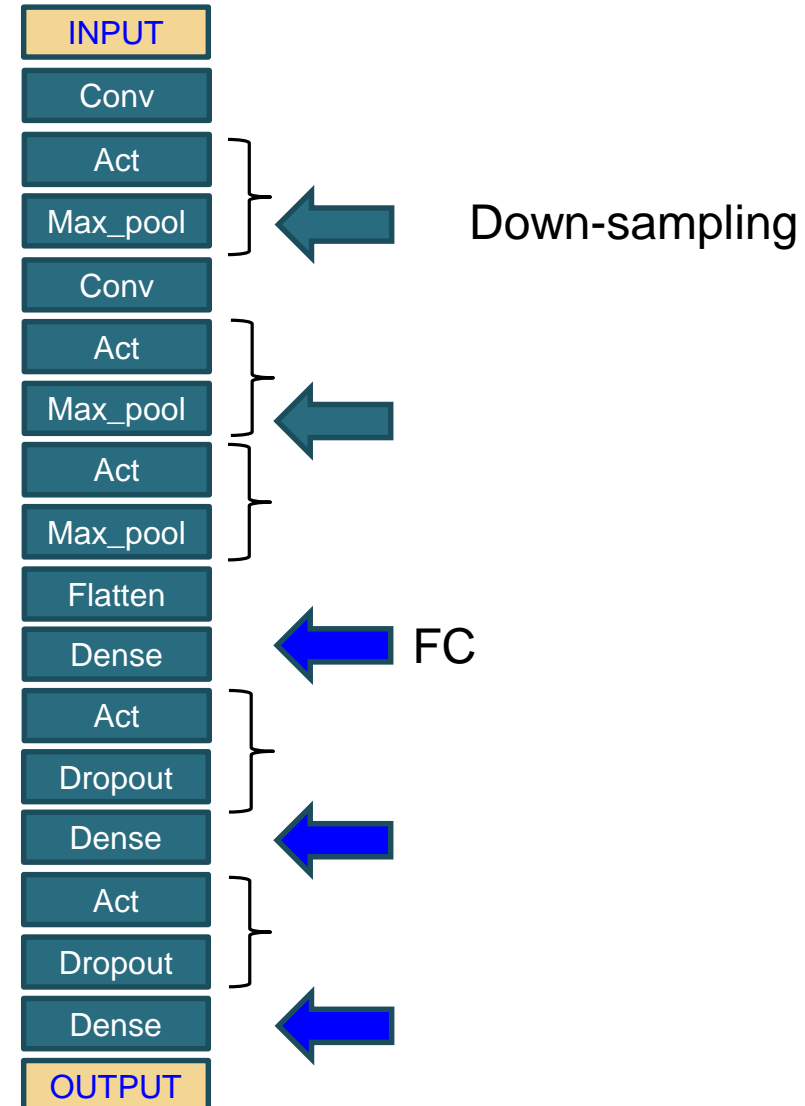


Components of conv1D

1. Activation (Act)
2. Convolution (Conv)
3. Maxpooling (Max_pool)
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.

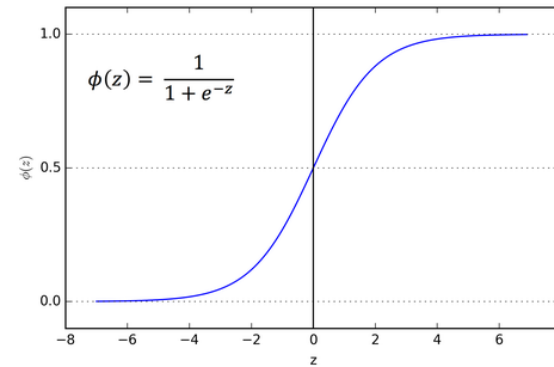


1. Activation Function

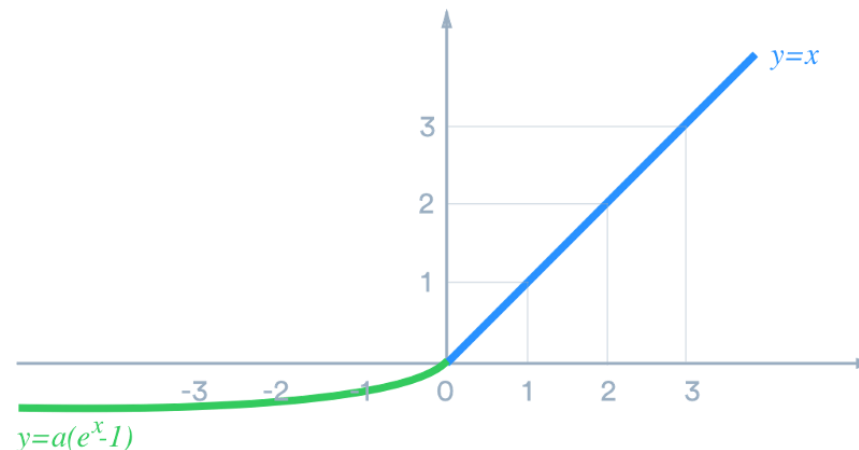
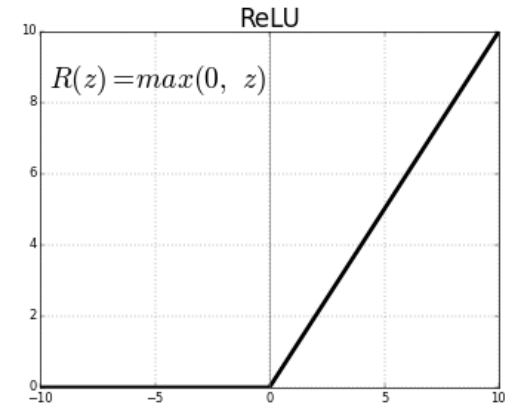
- Activation functions are included to create non-linearity

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

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



Squashes the #s to [0, 1]



1. Activation function

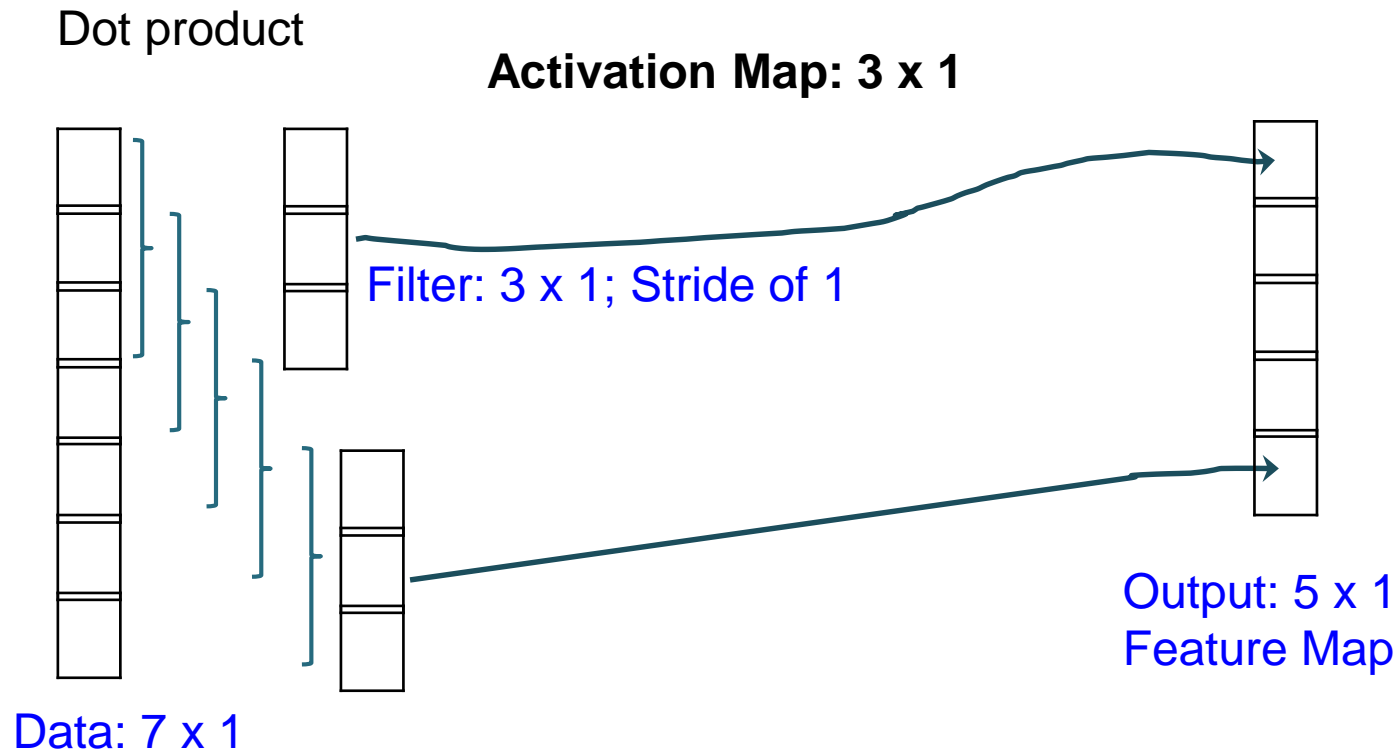
$$a^{(L)} = \sigma(w^{(L)}a^{(L-1)} + b^{(L)})$$

$$\sigma \left(\begin{bmatrix} W_{0,0} & W_{0,1} & \dots & W_{0,n} \\ W_{1,0} & W_{1,1} & \dots & W_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{k,0} & W_{k,1} & \dots & W_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \right)$$

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



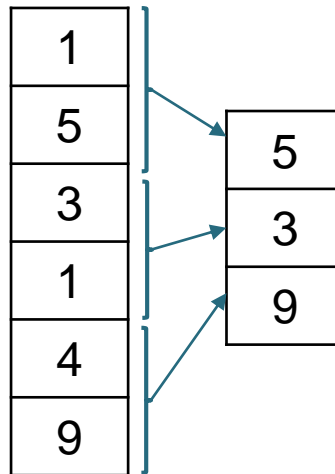
Convolutional Layer

of HP

- # 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:

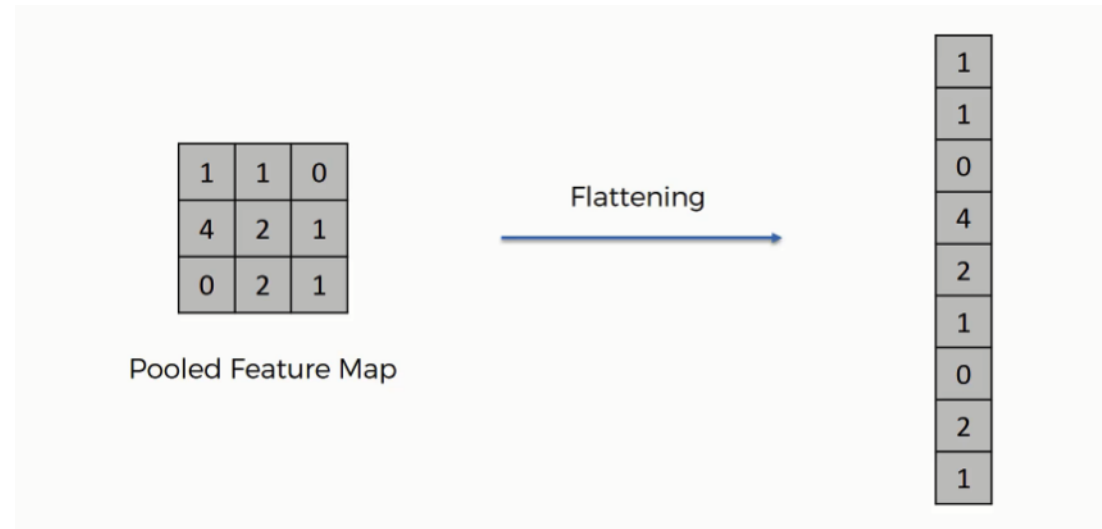
Enhancing the signals by looking at a region and pick the maximum activation value

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

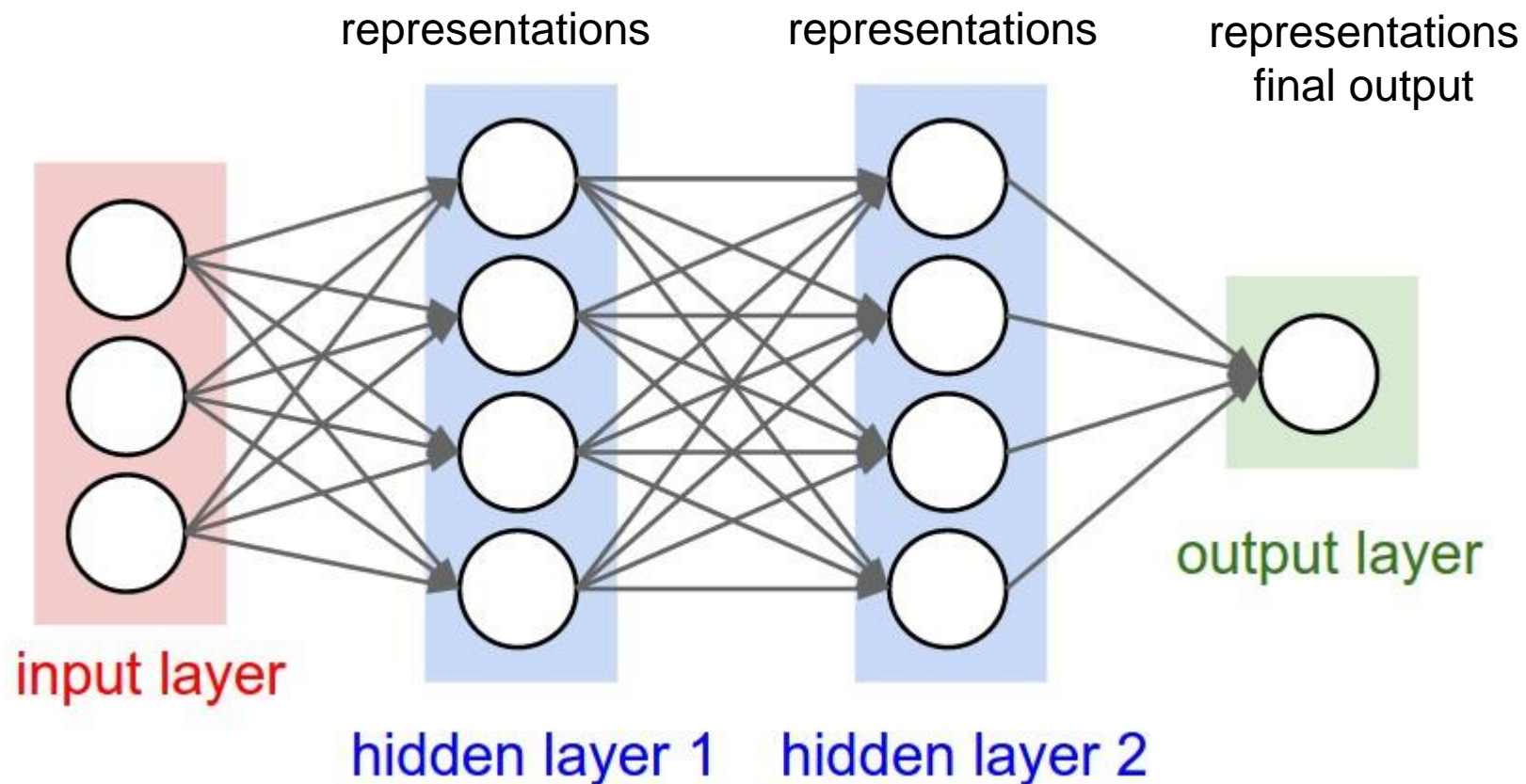
4. Flatten

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



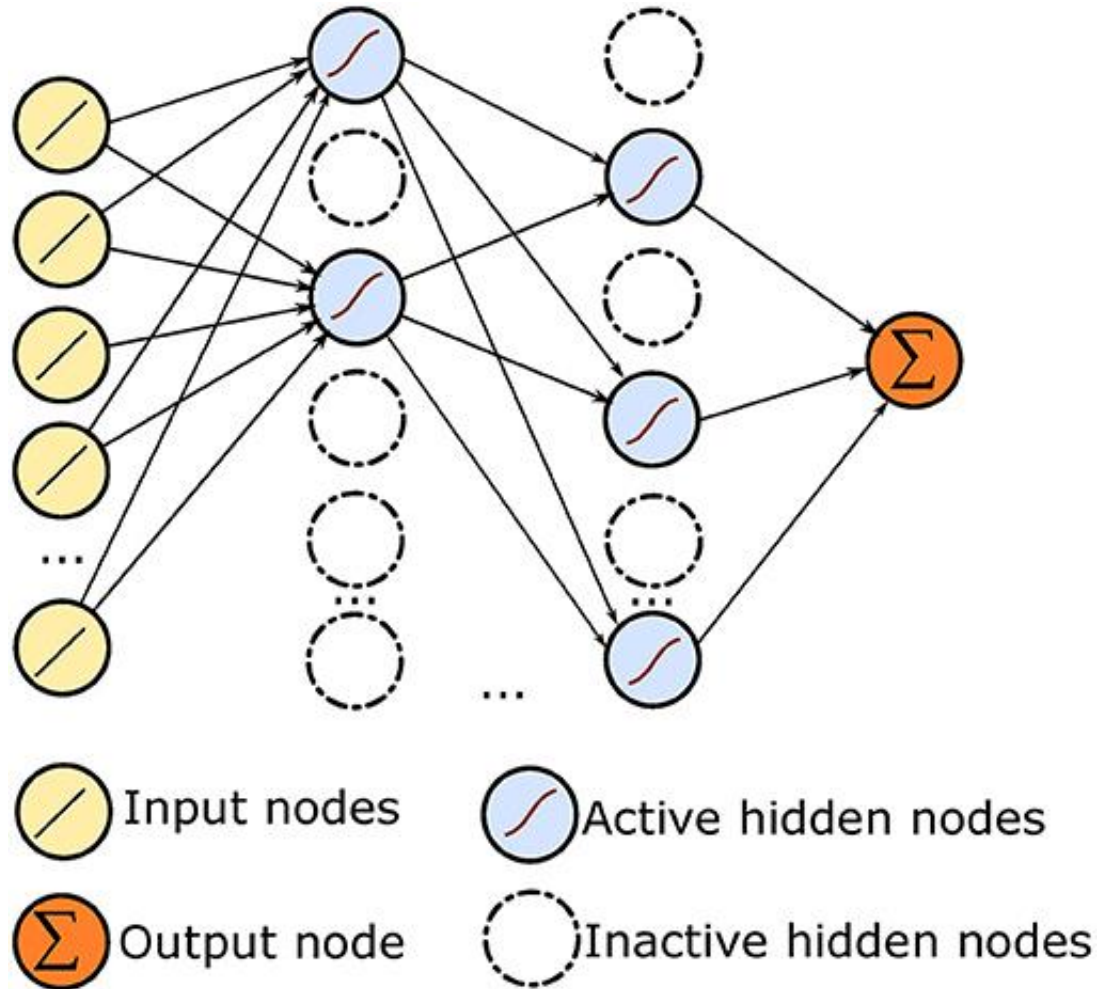
5. Dense

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

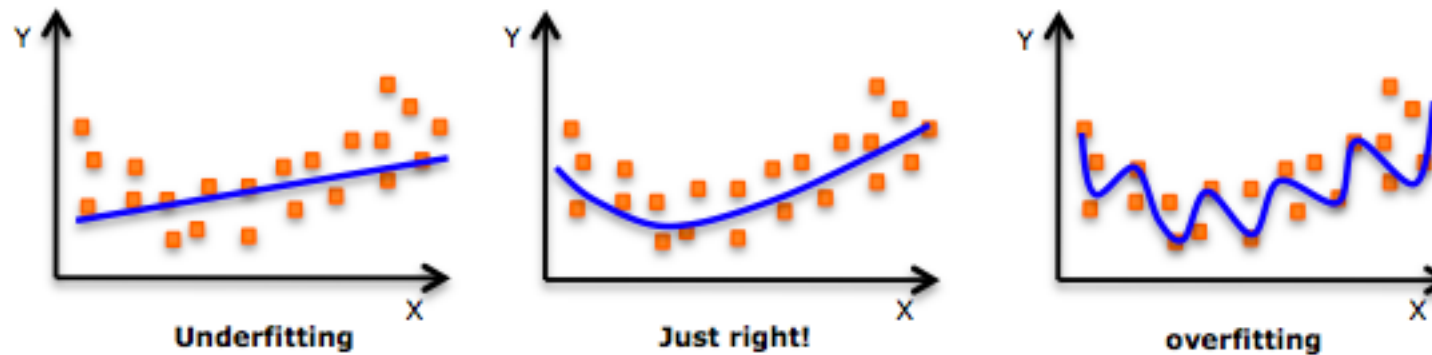


Imbalance in the weights among the nodes can lead to some node weights not contributing to the learning

**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

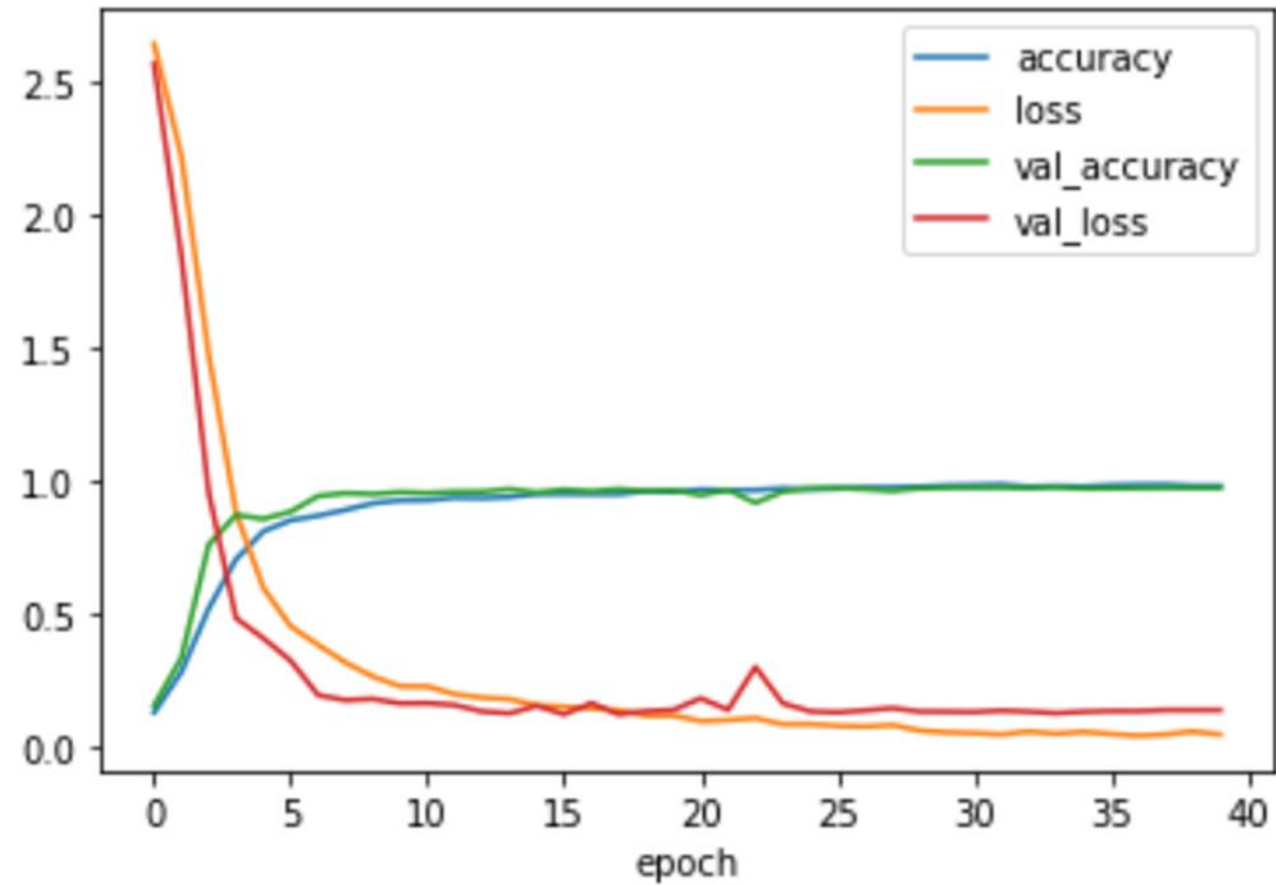
Epoch 7/400
3375/3375 [=====] - 228s 68ms/step - loss: 0.3835 - accuracy: 0.8690 - val_loss: 0.1944 - val_acc:
Epoch 00007: val_loss improved from 0.32363 to 0.19439, saving model to Pmodel.h5

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