Deep Learning Analytics for Genomic Medicine in AML

Andy Nguyen, M.D., M.S.

Professor of Pathology and Laboratory Medicine,
University of Texas-Houston, Medical School

UTHealth®
The University of Texas
Health Science Center at Houston
McGovern
Medical School

Outline of talk

- Define application of Deep Learning method, a technological breakthrough, to big-data analytics
- Define 2 major areas in pathology where Deep Learning method, a disruptive technology, is predicted to be an integrated part in future practice (molecular diagnosis with next-gen sequencing, and morphological diagnosis with whole slide digital imaging)
- Describe our 2 studies using Deep Learning algorithm to:
  -Find correlation between FLT3-ITD mutation status and proteomics in acute myeloid leukemia
  -Find correlation between prognosis and cytogenetics, age, and 23 most common mutations in acute myeloid leukemia

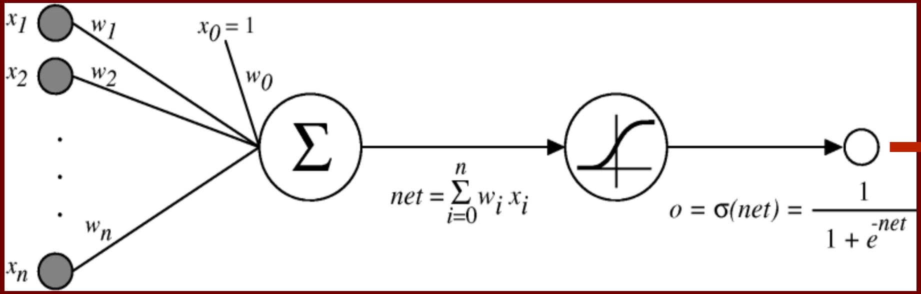
Financial Disclosures:
No relevant financial relationships with commercial interests to disclose

Big-Data Analytics and Deep Learning

- Big companies are analyzing large volumes of data for business analysis and decisions, using Deep Learning technology (Google's search engine, Google Photo, automobile companies: self-driving cars, IBM's Watson)

- Big data analytics in cancer proteomics and genomics can significantly be benefited from Deep Learning ***("We are drowning in information and starving for knowledge", Rutherford D. Roger)***

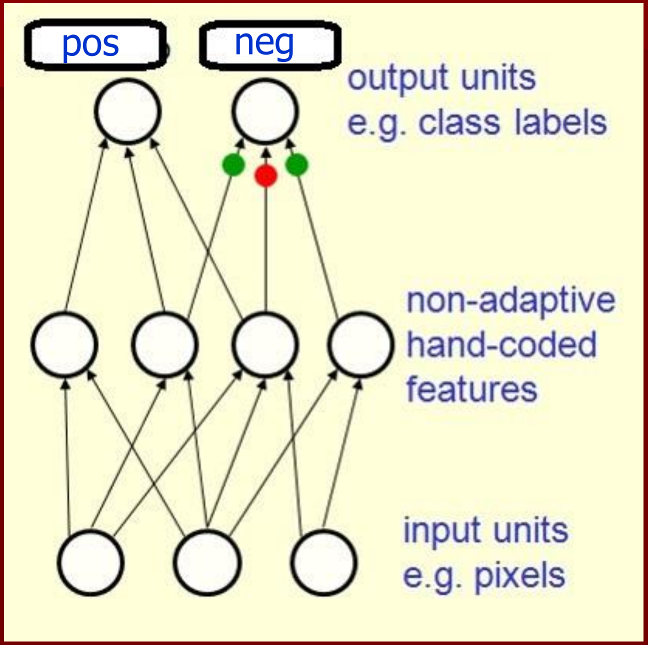
- Deep Learning is based on artificial neural networks (inspired by biological neural networks): artificial nodes ("neurons“) are connected together to form a network for prediction/classification tasks


x_1
w_1
$x_0 = 1$
x_2
w_2
w_0
Σ
w_n
x_n
$net = \sum_{i=0}^{n} w_i x_i$
$o = \sigma(net) = \frac{1}{1 + e^{-net}}$

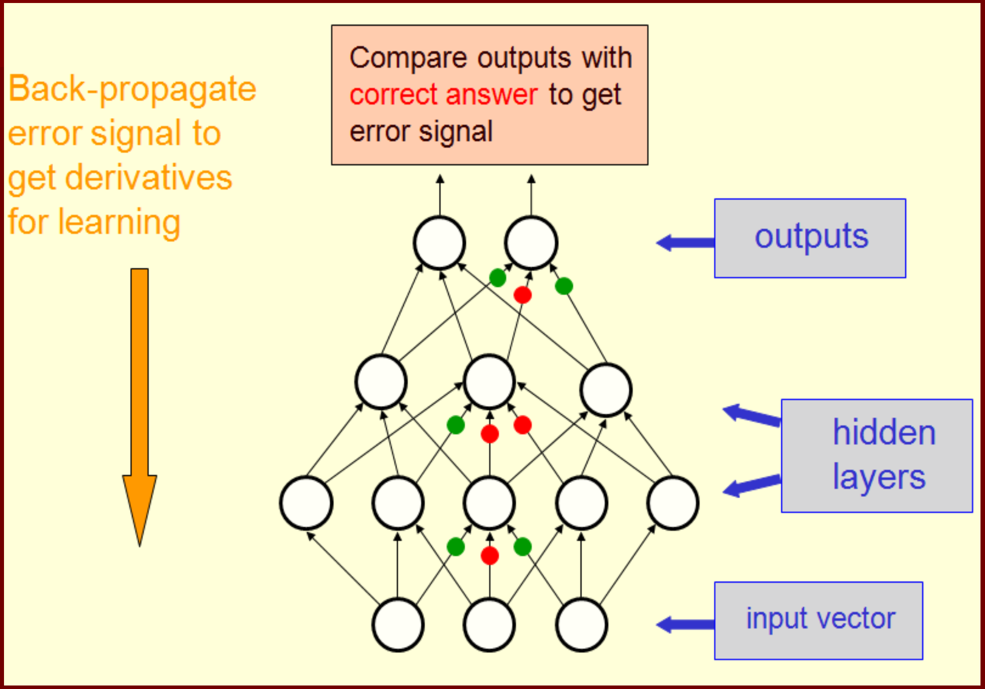
→ Fire/not Fire

- In traditional programming, an engineer writes explicit, step-by-step instructions for computers to follow. In neural network, they do not encode software with instructions; instead they train the software

Early Generations of Neural Networks with Supervised Training (model is trained with known outcomes)



1st gen (1960's)



2nd gen (1980's)

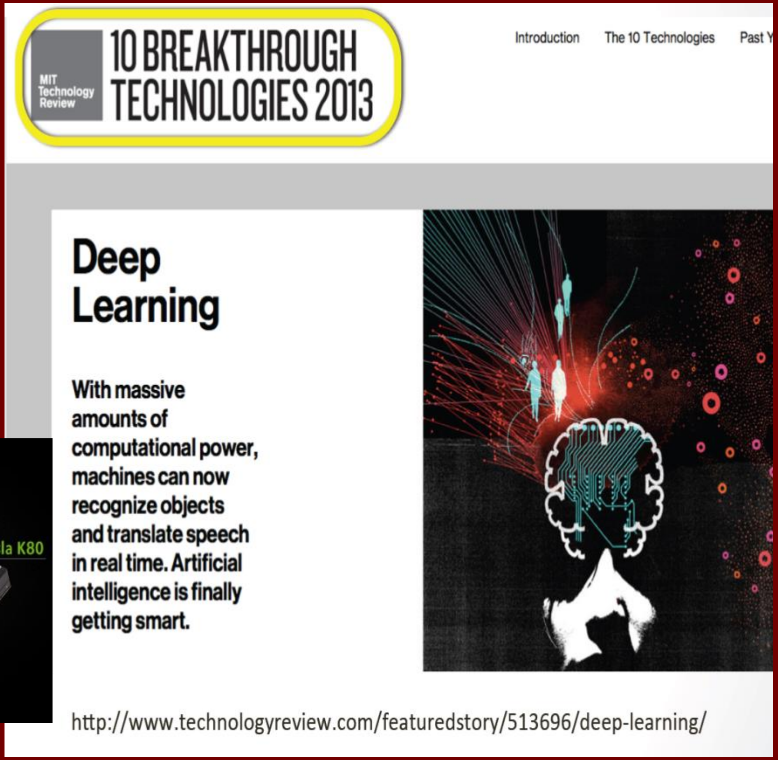
- Early neural networks were based on supervised training often too difficult to train and they were found to be less effective than other methods.

Deep Learning (3rd Gen Neural Network)

- A major breakthrough in 2006: Hinton (U of Toronto) won a contest held by Merck to identify molecules that could lead to new drugs. The group used deep learning to zero in on the molecules most likely to bind to their targets.

- Deep Learning algorithms:
  (1) Unsupervised learning ->allows a network to be fed with raw data (no known outcomes) and to automatically discover the representations needed for detection or classification

  (2) Extract high-level & complex data representations through multiple layers.
  -> allows for less interferences by background noise



MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

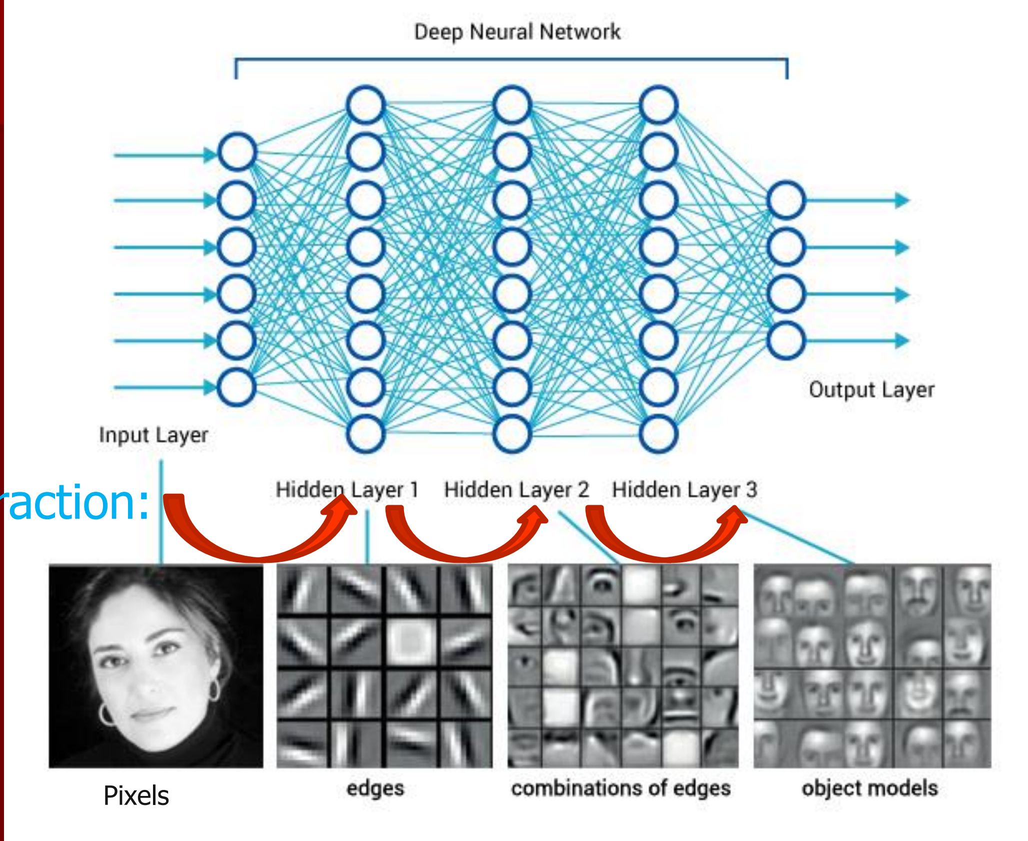
Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

http://www.technologyreview.com/featuredstory/513696/deep-learning/

- Supporting hardware: multiple graphics processing units (GPU)

A Deep Learning Neural Network to Detect Image: Extracting higher-level Features With Unsupervised Learning

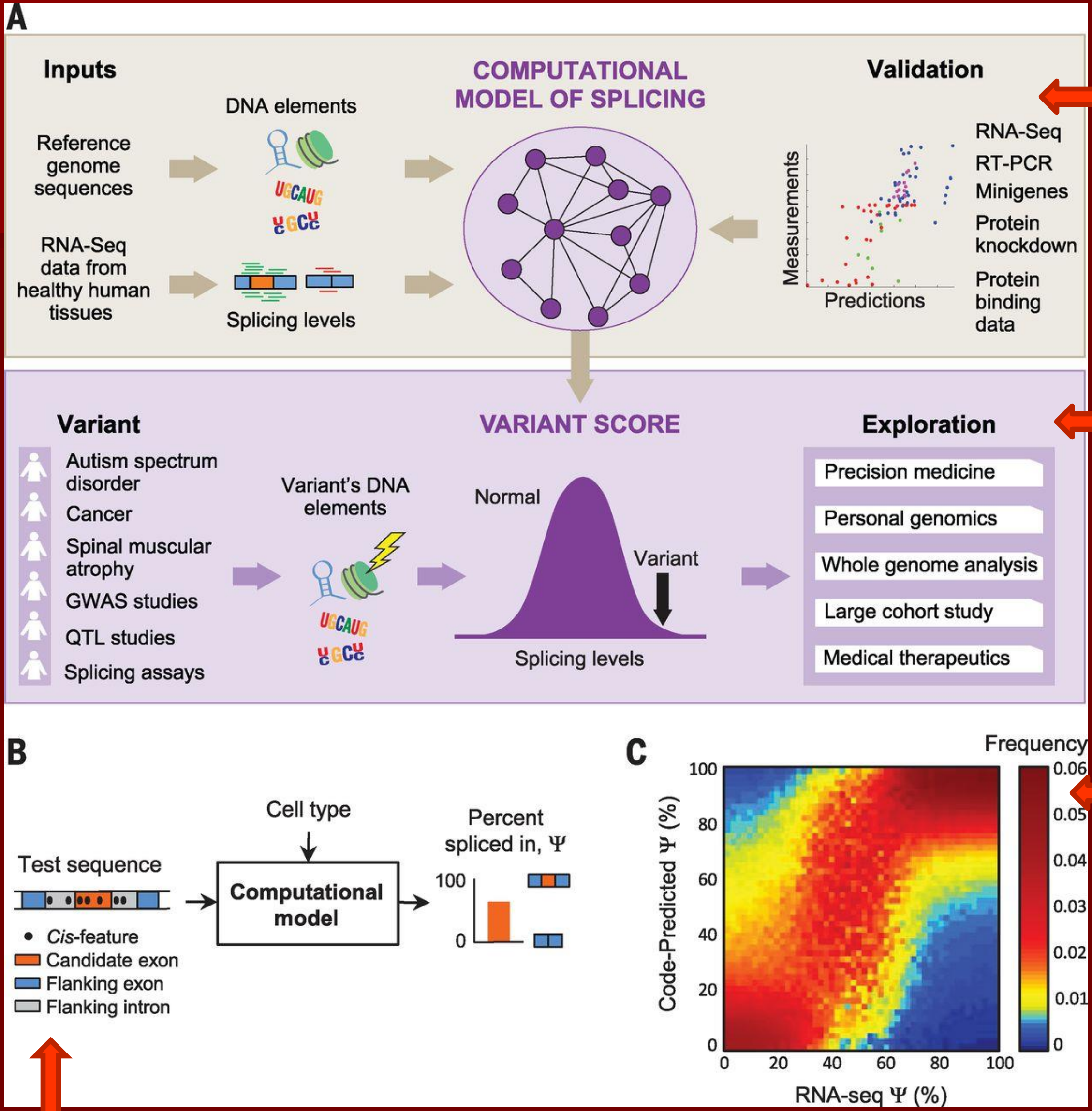


-Each hidden layer applies a nonlinear transformation on its input to transform the input to higher level of representation in its output.

-Multiple levels of abstraction of the image: from pixels to complex shapes and objects defining a human face

Detecting pathologic genetic variants using a deep learning model of splicing:

Mutations in MLH1 and MSH2 arising in patients with colorectal cancer **(U. of Toronto)



A. Machine learning to infer a model of splicing, by correlating DNA elements with splicing levels in healthy tissues.

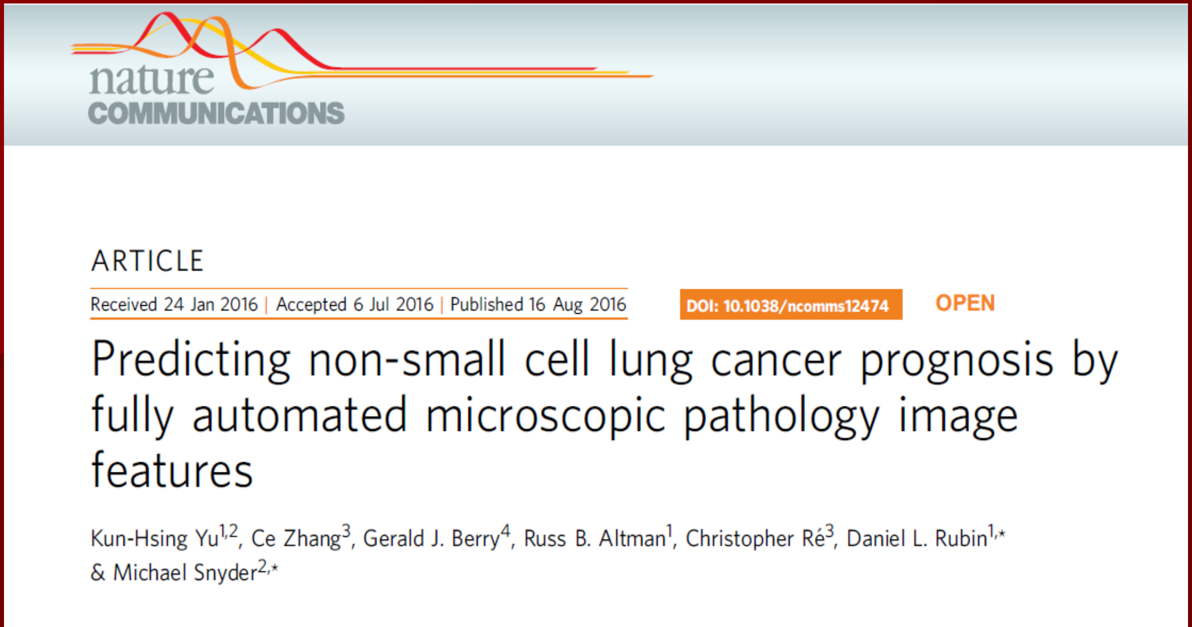
Model to learn variants a/w diseases

C. Predictions are made for 10,689 test exons profiled in 16 tissues -> AUC=94%

**Lynch syndrome, or hereditary nonpolyposis colorectal cancer

[H. Y. Xiong et al. Science 2015;347:1254806]

B. The model extracts the regulatory code from a test DNA sequence and predicts the percentage of transcripts with the central exon spliced in (Ψ)



nature COMMUNICATIONS

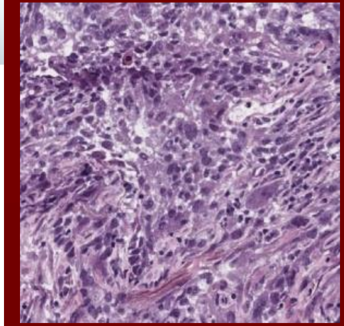
ARTICLE

Received 24 Jan 2016 | Accepted 6 Jul 2016 | Published 16 Aug 2016 DOI: 10.1038/ncomms12474 OPEN

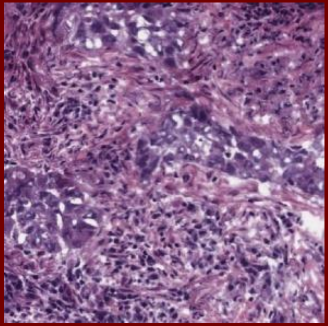
Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features

Kun-Hsing Yu[1,2], Ce Zhang[3], Gerald J. Berry[4], Russ B. Altman[1], Christopher Ré[3], Daniel L. Rubin[1,*] & Michael Snyder[2,*]

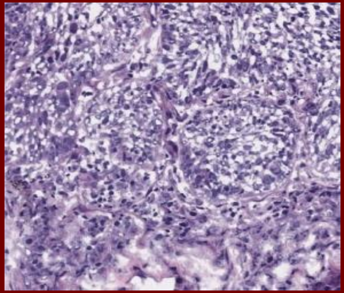
AdenoCA , 1B Gr 3
>99 mo



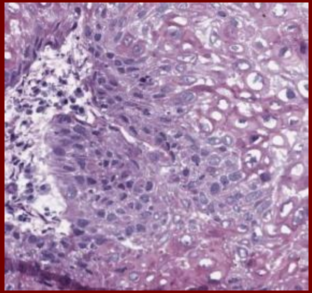
AdenoCA , 1B Gr 3
12 mo



SCC, 1 Gr 1
>70 mo



SCC, 1B Gr 3:
12 mo

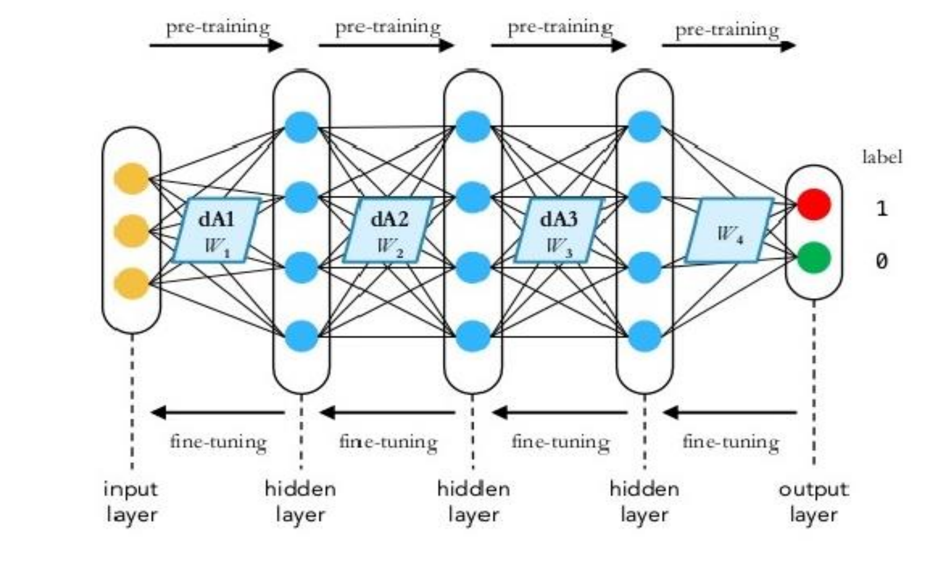


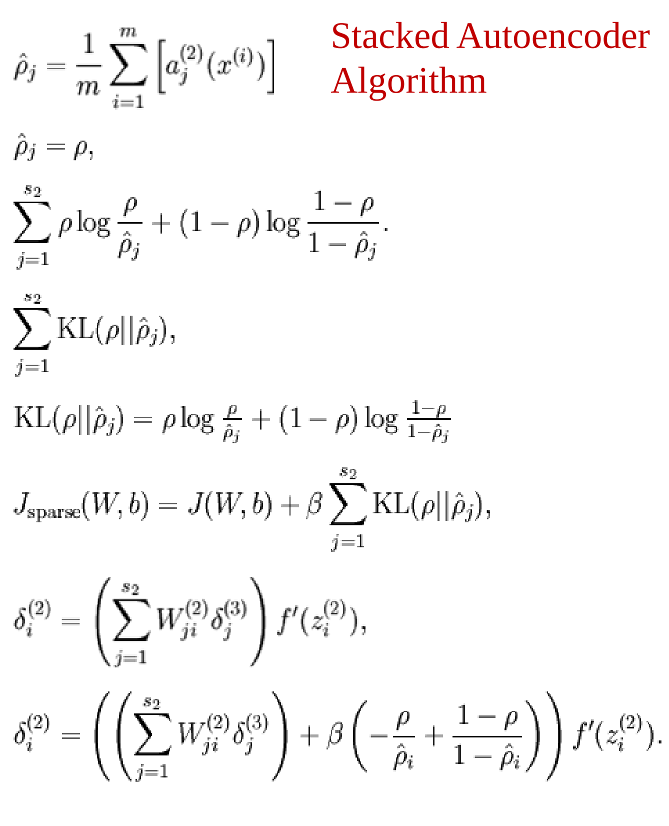
- Tumour stages/grades are insufficient for predicting survival outcomes (diverse)-> room for improvement
- Study conducted by Department of Pathology, Stanford University
- 2,186 whole-slide images (H&E) of lung adenocarcinoma and squamous cell carcinoma patients from The Cancer Genome Atlas (TCGA)
- Extract 9,879 quantitative image features-> scale down to 240 key features -> distinguish shorter-term survivors from longer-term survivors with stage I adenocarcinoma ($p<0.003$) or squamous cell carcinoma ($p=0.023$)
- Methods are extensible to histopathology images of other organs.

Our Deep Learning Projects-Programming Platform

- We design Deep Learning neural networks with stacked (multi-layered) auto-encoder in R language.
- R is a programming language for statistical computing and graphics supported by the R Foundation for Statistical Computing.
- In this study, we use many Deep Learning functions obtained from an R package called "Deepnet" which is available from the Comprehensive R Archive Network, under the GNU General Public License

Stacked Autoencoder Network





Stacked Autoencoder Algorithm

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}(x^{(i)})\right]$$

$$\hat{\rho}_j = \rho,$$

$$\sum_{j=1}^{s_2}\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}.$$

$$\sum_{j=1}^{s_2}\mathrm{KL}(\rho||\hat{\rho}_j),$$

$$\mathrm{KL}(\rho||\hat{\rho}_j) = \rho\log\tfrac{\rho}{\hat{\rho}_j} + (1-\rho)\log\tfrac{1-\rho}{1-\hat{\rho}_j}$$

$$J_{\text{sparse}}(W,b) = J(W,b) + \beta\sum_{j=1}^{s_2}\mathrm{KL}(\rho||\hat{\rho}_j),$$

$$\delta_i^{(2)} = \left(\sum_{j=1}^{s_2}W_{ji}^{(2)}\delta_j^{(3)}\right)f'(z_i^{(2)}),$$

$$\delta_i^{(2)} = \left(\left(\sum_{j=1}^{s_2}W_{ji}^{(2)}\delta_j^{(3)}\right) + \beta\left(-\frac{\rho}{\hat{\rho}_i} + \frac{1-\rho}{1-\hat{\rho}_i}\right)\right)f'(z_i^{(2)}).$$

Study #1: FLT3-ITD and Proteomics in AML

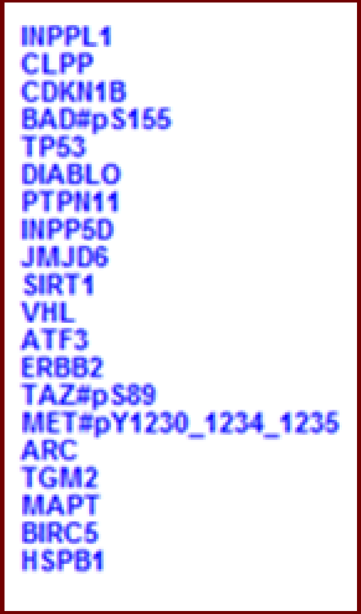
- We explore how Deep Learning can be utilized for proteomics analysis in AML. Specifically we attempt to determine the correlation between FLT3-ITD mutation status with serum level of 231 proteins in newly-diagnosed AML patients
- Dimensional reduction was initially performed to reduce the number of critical proteins from 231 down to 20.

DREAM 9 Challenge Data:

-Hosted by Rice University

-Data were provided by Dr. S. Kornblau from The University of Texas MD Anderson Cancer Center and were obtained through Synapse syn2455683 as part of the AML DREAM Challenge

20 top-ranked proteins



INPPL1
CLPP
CDKN1B
BAD#pS155
TP53
DIABLO
PTPN11
INPP5D
JMJD6
SIRT1
VHL
ATF3
ERBB2
TAZ#pS89
MET#pY1230_1234_1235
ARC
TGM2
MAPT
BIRC5
HSPB1

FLT3-ITD and Proteomics-Results

- We show how Deep Learning which incorporates unsupervised feature training can be used to find excellent correlation between FLT3-ITD mutation with levels of these 20 proteins
  (an accuracy of 97%, sensitivity of 90% and specificity of 100%).

231 Protein Cross Validation Data Sets		20 Protein Cross Validation Data Sets	
1	80%	1	100%
2	90%	2	100%
3	70%	3	80%
4	80%	4	100%
5	80%	5	100%
6	90%	6	100%
7	80%	7	100%
Mean=	81%	Mean=	97%

sensitivity of 90%,
and specificity of 100%

Study #2: AML Prognosis

- We explore how Deep Learning can be utilized for predicting prognosis in AML. Specifically we attempt to determine the correlation between prognosis and cytogenetics, age, and mutations in acute myeloid leukemia
- Materials: 56 AML cases from TCGA database. Data include cytogenetics, age, 23 most common mutations, prognosis (days to death)
- Cytogenetics : t(8;21), inv(16), t(15;17), t(9:11), t(9;22), trisomy 8, del (7), del (5), del (20), complex
- Mutations: FLT3, NPM1, DNMT3A, IDH2, IDH1, TET2,RUNX1, TP53, NRAS, CEBPA, WT1, PTPN11, KIT, U2AF1, KRAS, SMC1A, SMC3, PHF6, STAG2, RAD21, FAM5C, EZH2, HNRNPK

Study #2: AML Prognosis-Results

- Deep Learning which incorporates unsupervised feature training find excellent correlation between prognosis (DTD) with cytogenetics, age, 23 most common mutations
  (**an accuracy of xx%, sensitivity of xx%, and specificity of xx%**).

Cross Validation Data Sets	
1	xx%
2	xx%
3	xx%
4	xx%
5	xx%
6	xx%
7	xx%
Mean=	85%

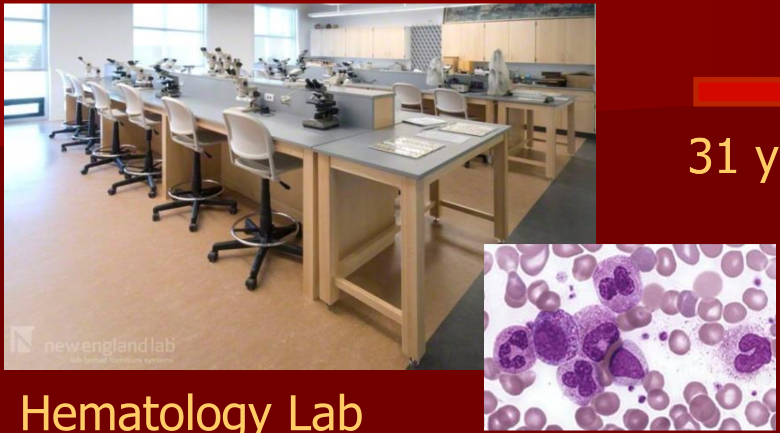
sensitivity of xx%,
and specificity of xx%

SUMMARY

- Deep Learning method, a disruptive technology, is predicted to be an integrated part in future practice in these areas:

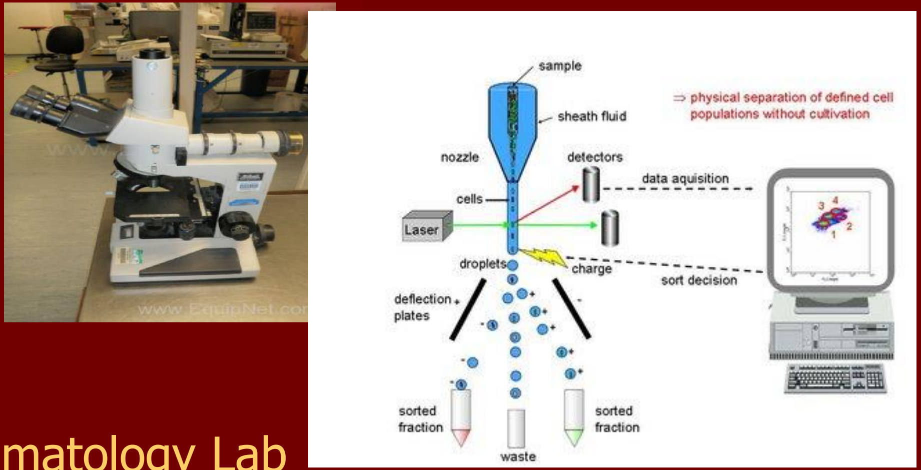
  - Molecular diagnosis & prognosis prediction using next-gen sequencing data (our 2 studies belong to this area)
  - Morphological diagnosis with whole slide digital imaging

Looking Forward: Deep Learning as a Disruptive Technology



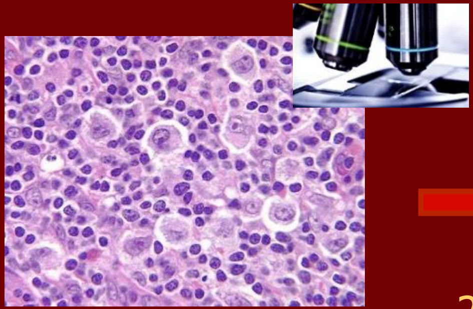
Hematology Lab
(600 bed hospital, circa 1985)
-Rudimentary CBC instruments
-10 microscope stations
for WBC differential counts

31 years

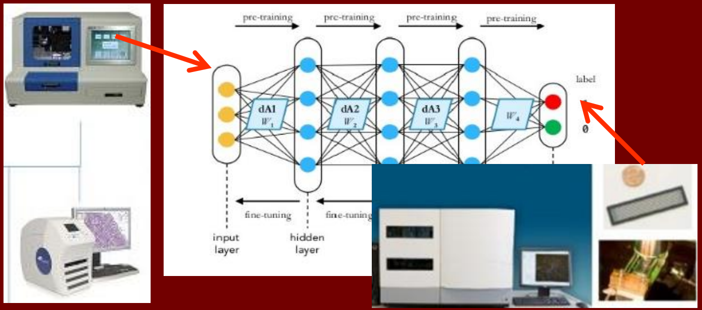


Hematology Lab
(900 bed hospital, circa 2016)
-Sophisticated CBC instruments
that release most WBC diff counts
-1 microscope station to check on WBC flags

What will pathology examination be like in another 31 years?



2016: microscope
->H&E, IHCs



2047: Deep Learning ?
-Digital whole slide imaging -> histol DX
-NGS-> genomic analysis -> mol DX

Ha Long Bay

Ninh Binh Province

Vietnam, March 2016