Multi-omics data integration by deep learning for phenotype prediction

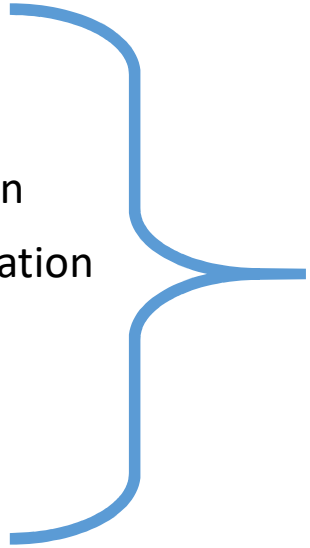
Blaise Hanczar

Laboratoire IBISC
Informatique, Bioinformatique et Systèmes Complexes

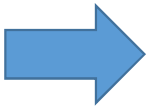
Omics data integration

Supervised learning setting

- Gene expression
- DNA Methylation
- Copy Number Variation
- Single Nucléotide Variation
- Meta-genomics
- Clinical
- Medical images



Predictive Model



- Diagnosis
- Prognostics
- Treatment response

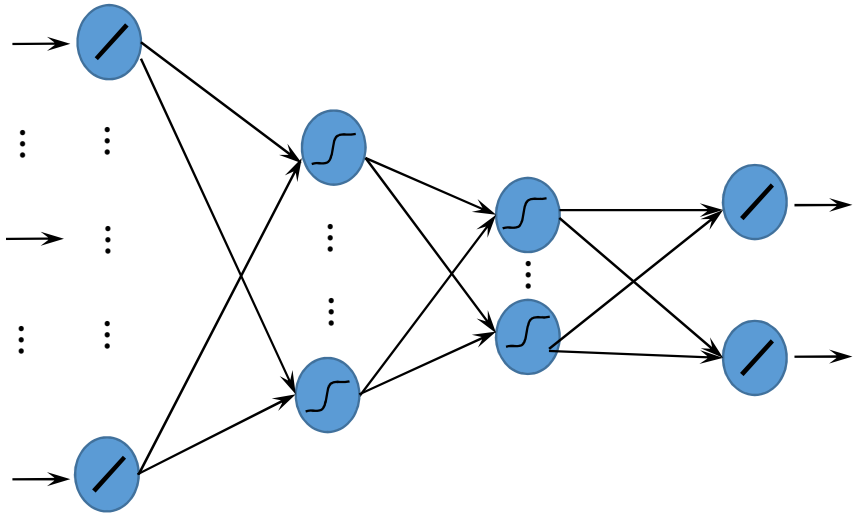
How to integrate different sources of data ?

Numeric, binary, categorial, count, sequences, images

Omics data integration

Supervised learning setting

- Gene expression
- DNA Methylation
- Copy Number Variation
- Single Nucléotide Variation
- Meta-genomics
- Clinical
- Medical images



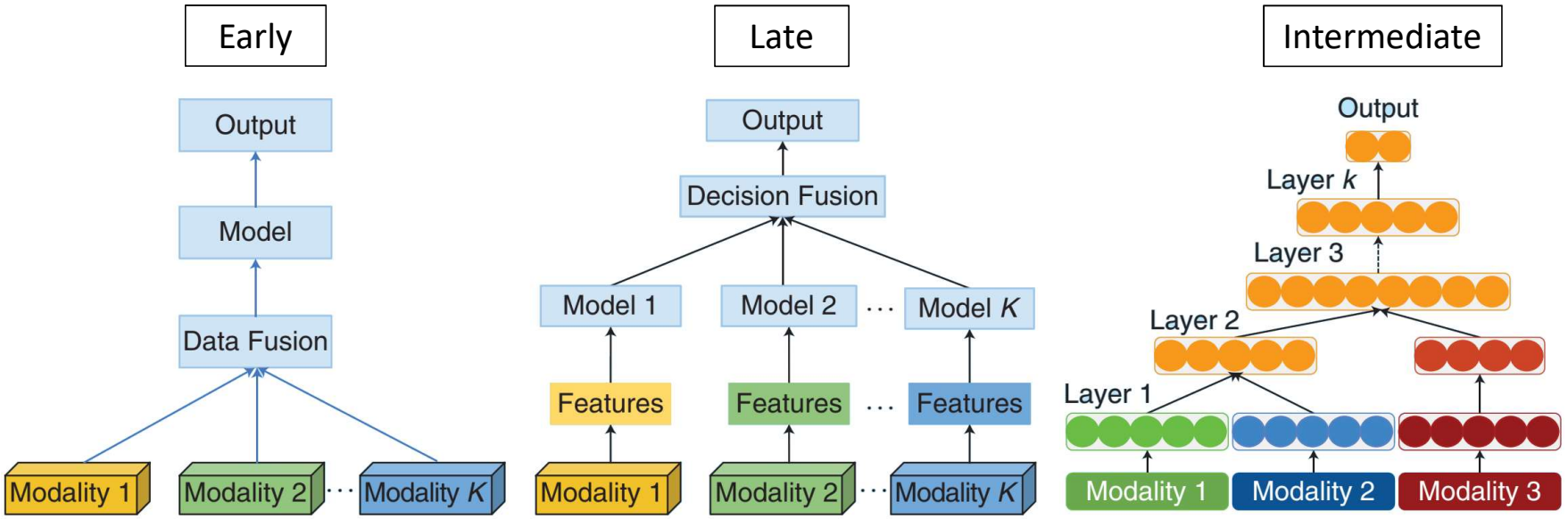
- Diagnosis
- Prognostics
- Treatment response

How to integrate different sources of data ?

Numeric, binary, categorial, count, sequences, images

With a neural network

Integration levels



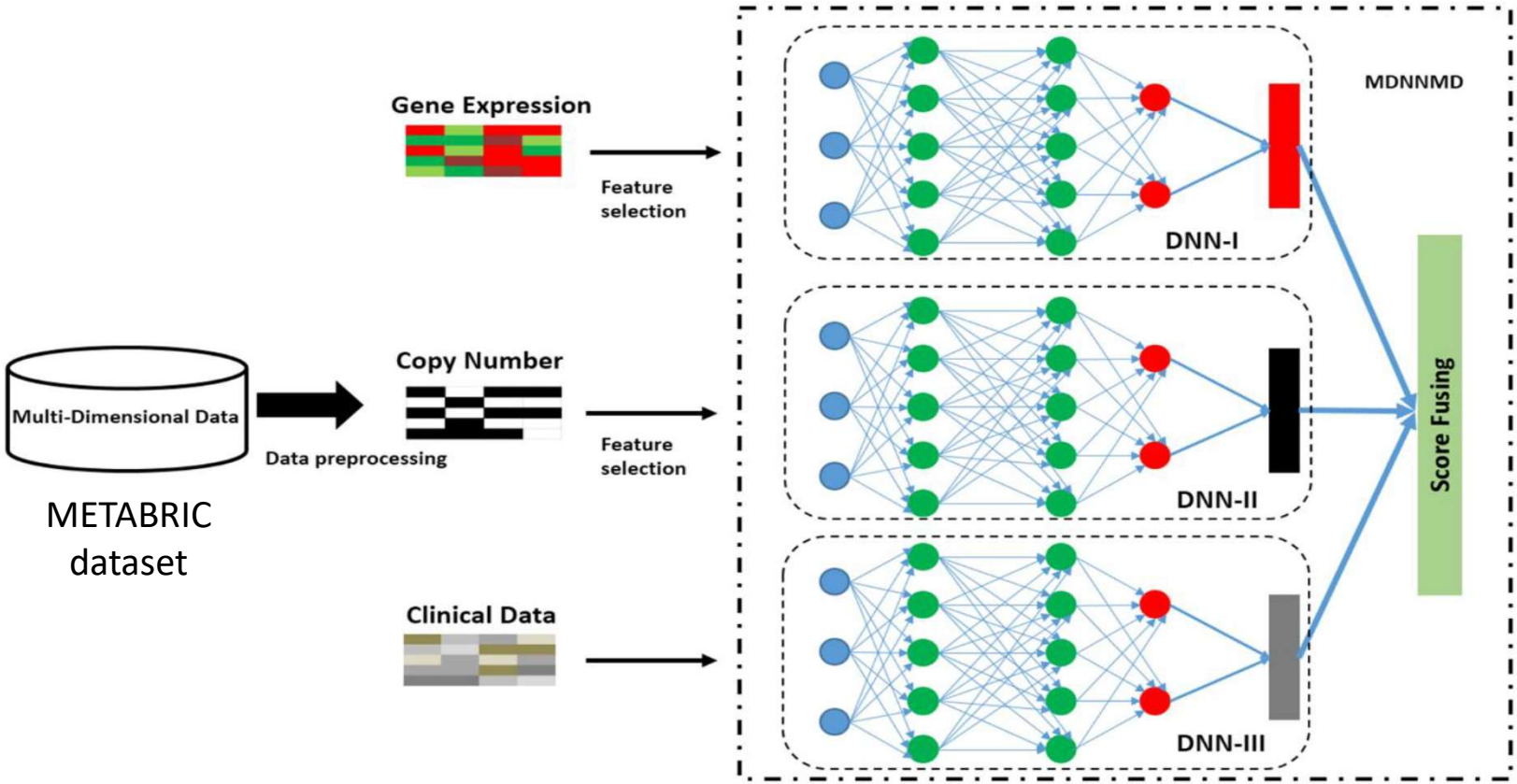
- Preprocessing
- Feature extraction
- Redundancy
- Note fully exploit the complementary nature of the modalities

- Decisions fusion
- Related to ensemble methods
- Need uncorrelated errors

- Shared representation of the data
- Flexible fusion
- Majority of the deep mutli-modal methods

Ramachandram, D., & Taylor, G. W. (2017). *IEEE Signal Processing Magazine*, *34*(6), 96-108.

Late integration

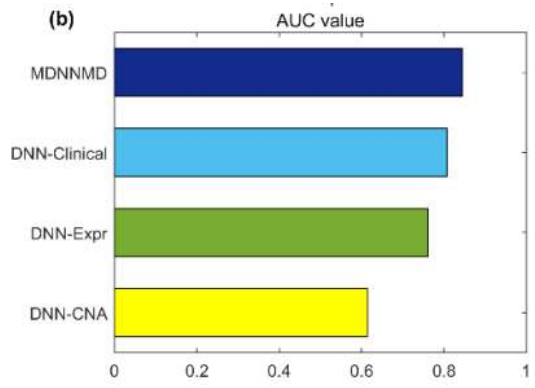


- DNN learnt separately
- Final decision :

$$o_{DNNMD} = \alpha * o_{DNN-Expr} + \beta * o_{DNN-CNA} + \gamma * o_{DNN-Clinical}$$

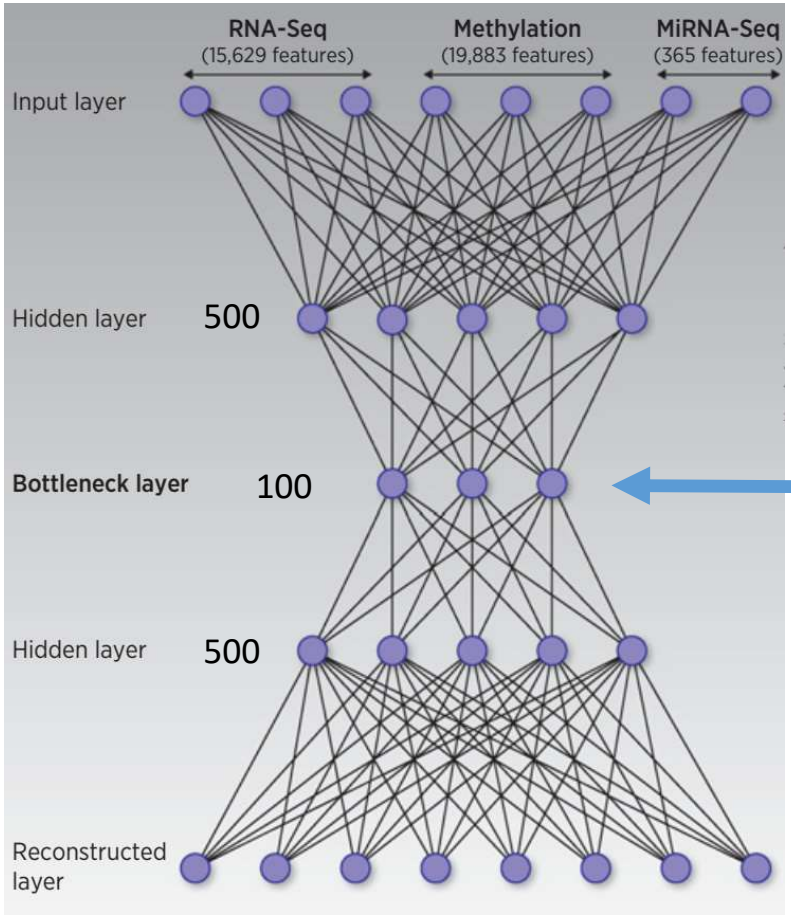
$$s.t.\ \alpha + \beta + \gamma = 1, \alpha \geq 0,\ \beta \geq 0, \gamma \geq 0$$

- Prognosis prediction :



Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, *16*(3), 841-850.

Early integration



Data integration by multi-modal autoencoder

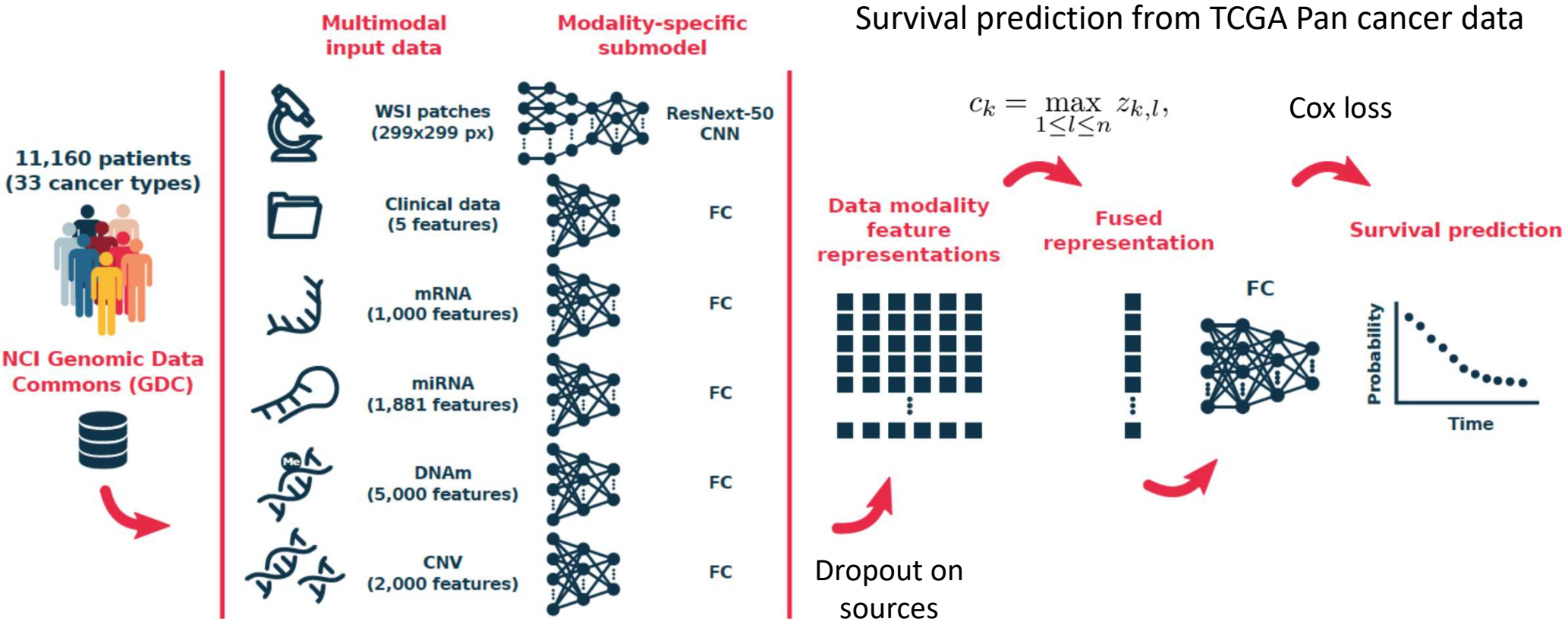
Prediction of liver cancer survival from TCGA dataset

Latent space integrating the information from the three data sources

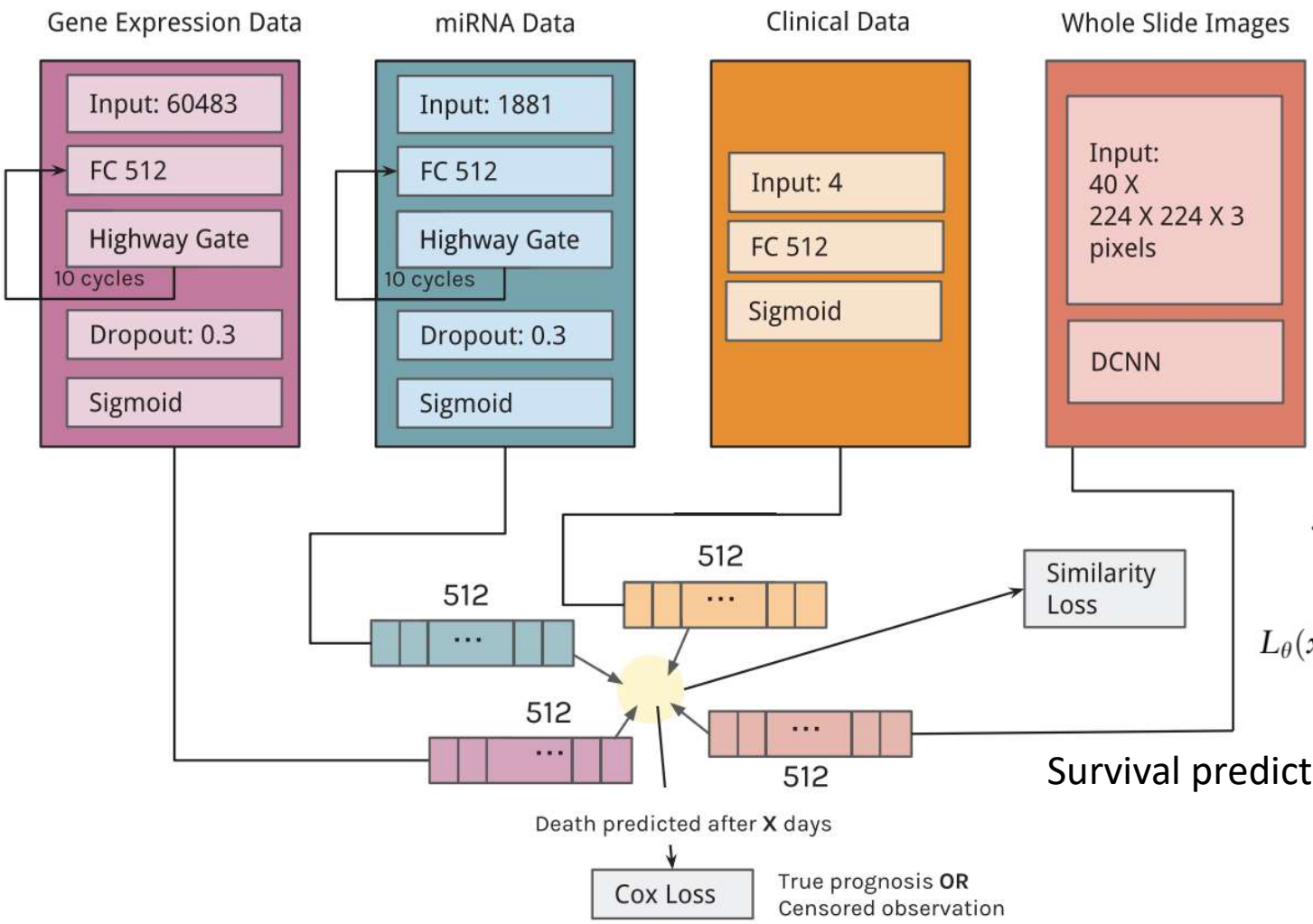
- K-means
- Cox-PH model
- SVM

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, *24*(6), 1248-1259.

Intermediate integration



Vale-Silva, L. A., & Rohr, K. (2020). MultiSurv: Long-term cancer survival prediction using multimodal deep learning. *medRxiv*.



$$sim_\theta(x,y) = \sum_{i,j \in \text{modalities}} \frac{\hat{h}_{\theta,i}(x_i) \cdot \hat{h}_{\theta,j}(y_j)}{|\hat{h}_{\theta,i}(x_i)||\hat{h}_{\theta,j}(y_j)|}$$

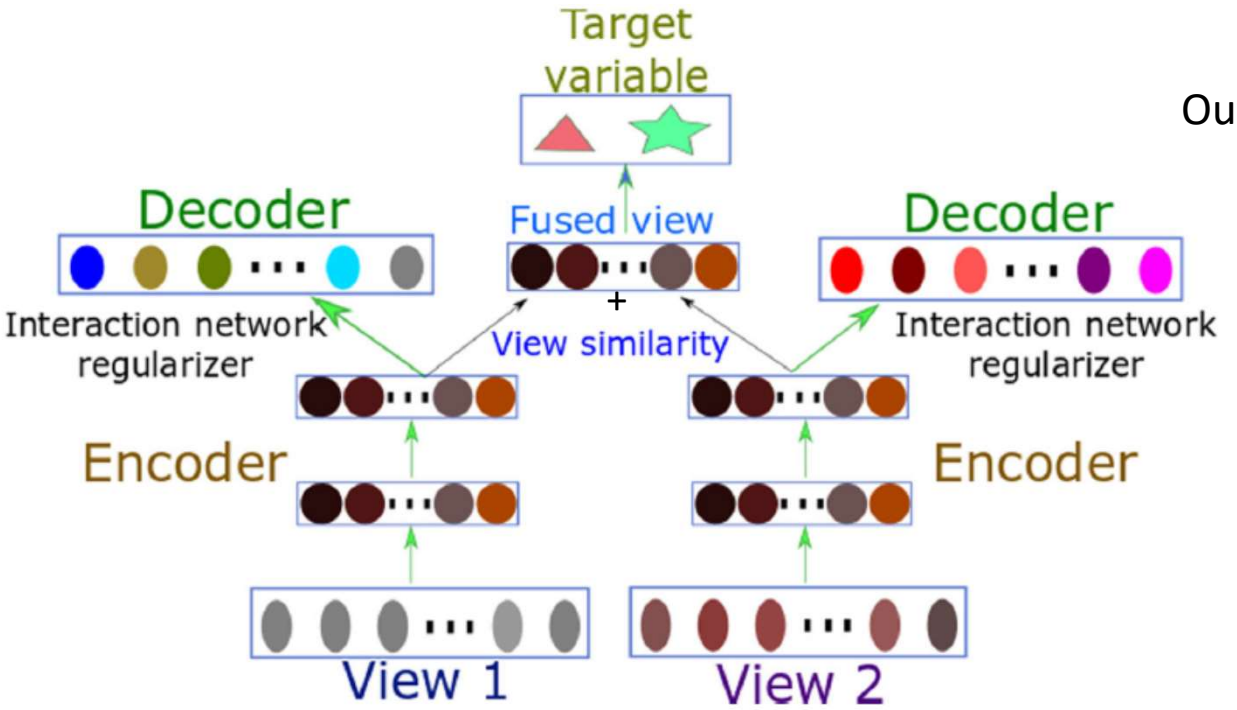
$$L_\theta(x,y) = \max(0, M - sim_\theta(x,y) + sim_\theta(x,x))$$

Survival prediction from TCGA Pan cancer data

$$l(\theta) = l_{sim}(\theta) + l_{cox}(\theta)$$

Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, *35*(14), i446-i454.

Multi omics and knowledge integration



RNA-seq, miRNS-seq, Protein, DNA methylation

Outcome prediction from TCGA Pan cancer data

$$\arg\min_{\{\Theta_{\mathbf{e}}^{(\nu)},\mathbf{Y}^{(\nu)}\}} \mathcal{L}(\mathbf{T}, \left(\sum_{\nu=1}^{V} Encoder\left(\mathbf{M}^{(\nu)}, \Theta_{\mathbf{e}}^{(\nu)}\right)\right) \cdot \mathbf{C})$$
$$+ \eta \sum_{\nu=1}^{V} \left(\left\| \mathbf{M}^{(\nu)} - Encoder\left(\mathbf{M}^{(\nu)}, \Theta_{\mathbf{e}}^{(\nu)}\right) \cdot \mathbf{Y}^{(\nu)} \right\|_F^2 \right)$$
$$+ \alpha \sum_{\nu=1}^{V} Trace\left(\mathbf{Y}^{(\nu)} \cdot \mathbf{L}_{G^{(\nu)}} \cdot \mathbf{Y}^{(\nu)^T}\right)$$
$$+ \beta \sum_{\nu=1}^{V} Trace\left(\mathbf{X}^{(\nu)^T} \cdot \mathbf{L}_S \cdot \mathbf{X}^{(\nu)}\right)$$

Ma, T., & Zhang, A. (2019). Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC genomics*, *20*(11), 1-11.

Some results

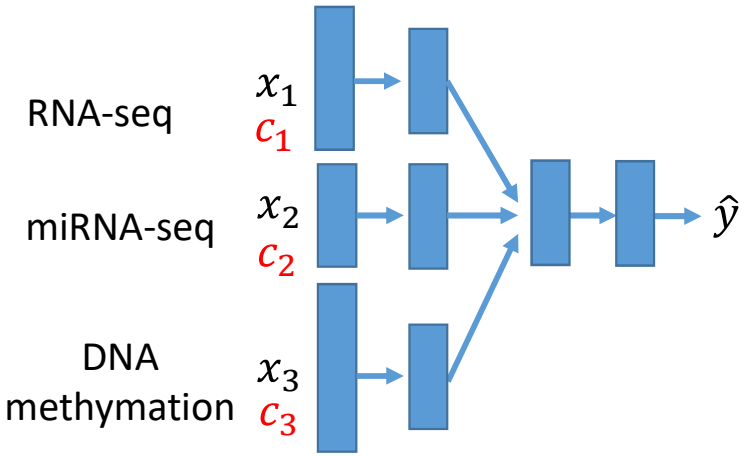
Limitations of multi omics models:

- More data sources does not necessarily mean better accuracy
- More data sources implies more **costly** models

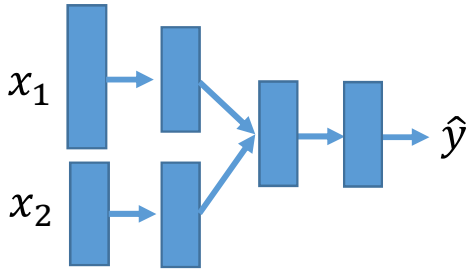
Included data modalities						C^{td} (95% CI)	IBS (95% CI)
Clinical	mRNA	DNAm	miRNA	CNV	WSI		
●	●					**0.822** (0.805-0.837)	0.138 (0.126-0.150)
●		●				0.808 (0.791-0.826)	**0.134** (0.125-0.148)
●			●			0.792 (0.775-0.810)	0.147 (0.136-0.161)
●				●		0.795 (0.778-0.812)	0.140 (0.131-0.152)
●					●	0.801 (0.783-0.817)	0.148 (0.140-0.158)
●	●	●				0.810 (0.793-0.829)	0.146 (0.135-0.158)
●	●	●	●			0.798 (0.781-0.815)	0.153 (0.139-0.168)
●	●	●	●	●		0.802 (0.748-0.820)	0.149 (0.136-0.162)
●	●	●	●	●	●	0.787 (0.769-0.806)	0.152 (0.140-0.166)

Clinical: tabular clinical data; mRNA: gene expression; DNAm: DNA methylation; miRNA: microRNA expression; CNV: gene copy number variation; WSI: whole-slide images.

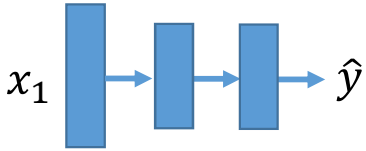
Which data sources to use ?



Cost = $c_1 + c_2 + c_3$



Cost = $c_1 + c_2$



Cost = c_1

- Tradeoff between accuracy and cost
- Some patients are easy to predict :
  - Do no need all data sources
  - Prediction could be less expensive with the same accuracy

Budget learning for multi omics data

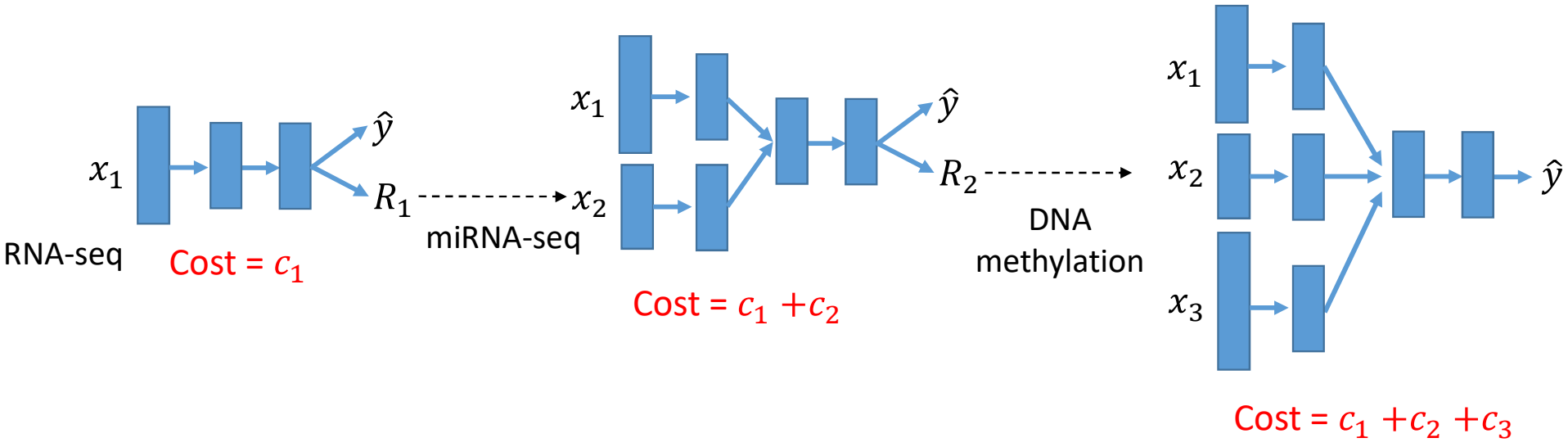
- Objectives in classification:
  - Maximize the accuracy
  - Minimize the prediction cost

- Cost is mainly the cost of acquisition of the variables
  - Money
  - Time
  - Secondary effect
  - Any non-infinite ressource

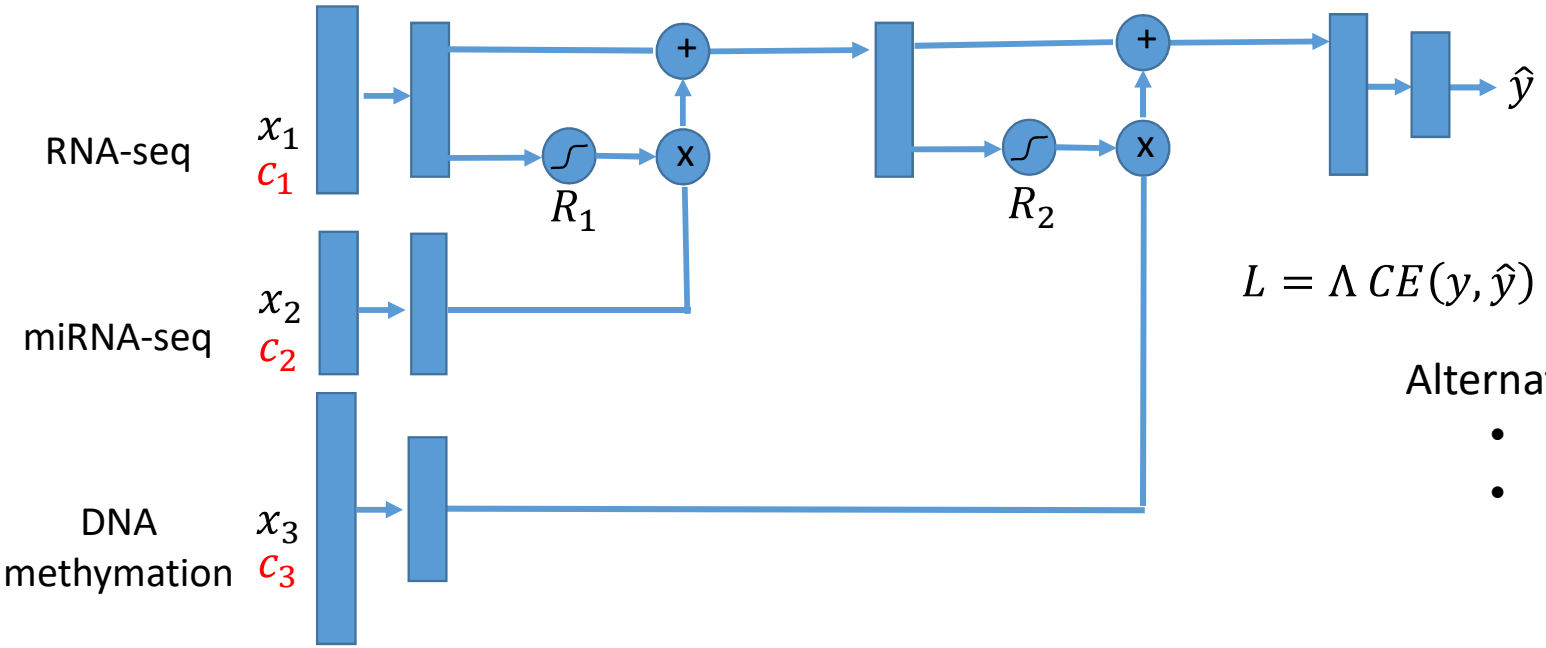
- Adaptative integration of omics data

Adaptative integration of omics data

- Sequence of neural networks with reject option of increasing cost
- At each level, we compute :
  - Prediction from current data sources
  - Decision to acquire the next data source



End-to-End learning



$$L = \Lambda\, CE(y, \hat{y}) + [C_1 + R_1 C_2 + R_2 C_3]$$

Alternate training:

- $min_{w,b}\ L$
- $min_{R_1,R_2}\ L$

- R_1 and R_2 control the integration of x_1 and x_2
- The order of the data sources is important