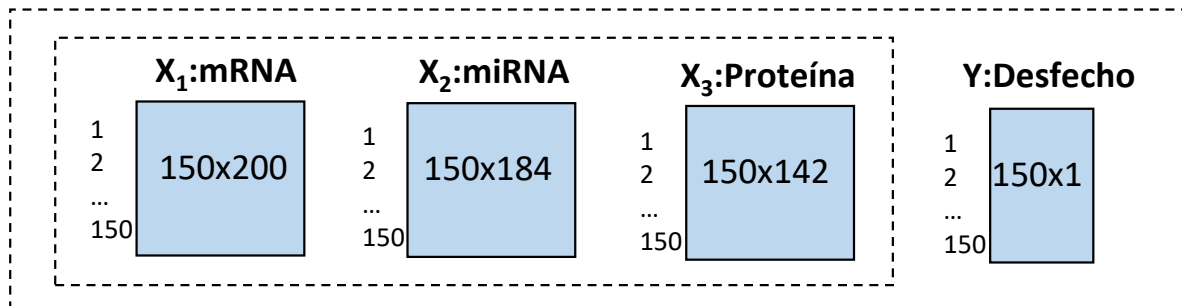


## MAE5776- ANÁLISE MULTIVARIADA DE DADOS 2º Sem/2020

Considere os dados de 3 bancos de dados ( $X_1$ =mRNA,  $X_2$ =miRNA e  $X_3$ =Protein) avaliados em 150 indivíduos. Estes dados estão disponíveis no MixOmics do R e constam de  $p_1=200$  variáveis de mRNA (RNA mensageiro: expressão de 200 fragmentos de DNA),  $p_2=184$  variáveis de miRNA (micro RNA) e de  $p_3=142$  variáveis que correspondem à abundância de proteínas alvos do estudo. Além disso, está disponível informação sobre um desfecho de interesse, variável  $Y$ , que corresponde ao subtipo de câncer de pulmão.

```
> data('breast.TCGA')
> # extrair dados (treinamento) de 3 BD
> data = list(mRNA = breast.TCGA$data.train$mrna,
+            miRNA = breast.TCGA$data.train$mirna,
+            proteomics = breast.TCGA$data.train$protein)
> lapply(data, dim)
$mRNA
[1] 150 200
$miRNA
[1] 150 184
$proteomics
[1] 150 142
> # extrair dados de grupo (análises supervisionadas)
> y = breast.TCGA$data.train$subtype
> summary(y)
Basal Her2 LumA
  45   30   75
```

A figura a seguir ilustra o formato destes dados.



Explore os recursos do R (veja Aula18-ComandosR) na realização das seguintes análises de redução de dimensionalidade e integração destes dados:

1. Análise de Componentes Principais (via Coordenadas Principais) e Componentes Principais Esparsos considerando cada banco de dados ( $X_1$ ,  $X_2$  e  $X_3$ ): calcule as cargas e os escores para os dois primeiros componentes e compare as duas análises (CP e sCP). Construa gráficos para visualização dos resultados de sua análise.
2. Análise Discriminante Esparsa: calcule as cargas e os escores das duas direções discriminantes. Obtenha as matrizes de classificação correta. Qual dos bancos de dados ( $X_1$ ,  $X_2$  ou  $X_3$ ) melhor prediz os subtipos de cânceres?
3. Análise de Correlação Canônica Esparsa (sCCA): Considere os pares de bancos de dados ( $X_1\_X_2$ ,  $X_1\_X_3$  e  $X_2\_X_3$ ) e realize a integração de bancos de dados via sCCA padrão, sCCA por permutação (sCCA.p) e sCCA com predição multiclasse (sCCA.y).
4. Finalmente, comente o racional dessas análises com soluções esparsas para a redução de dimensionalidade em espaços  $n \ll p$ .

Boa Sorte ☺