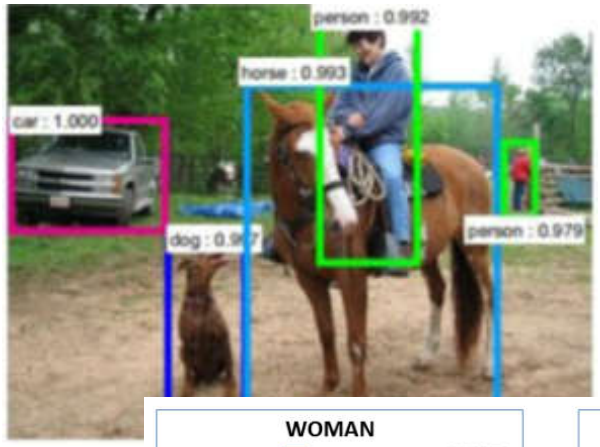
Apprentissage profond pour la prédiction de phénotypes à partir de données d'expression

Blaise Hanczar

Laboratoire Informatique, Bioinformatique
et Systèmes Complexes

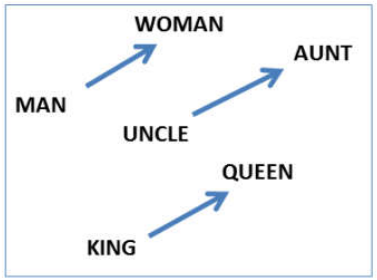
Journées PEPI IBIS 2019 – 6 juin 2019

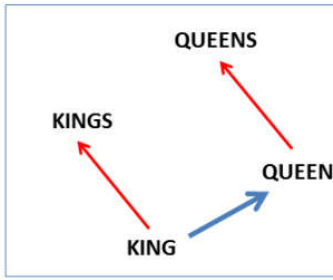
L’apprentissage profond



- Analyse d’images

  Classification, reconnaissance d’objets, segmentation,…

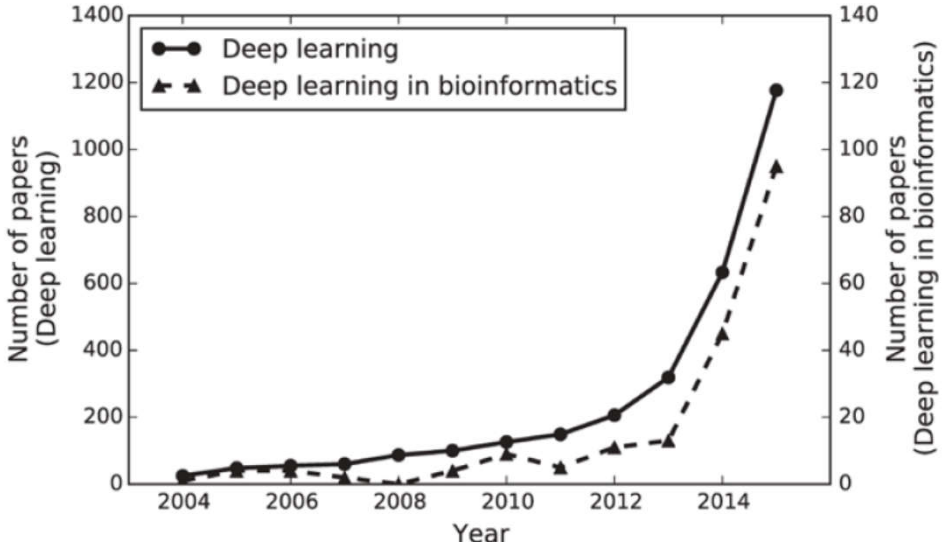


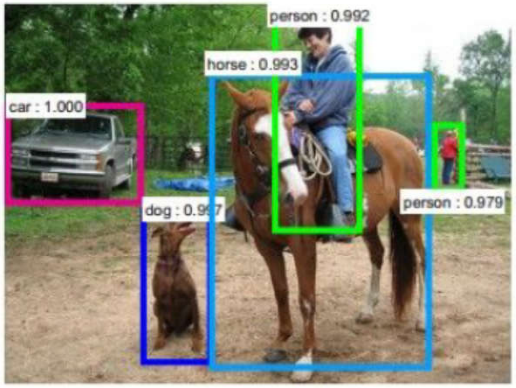


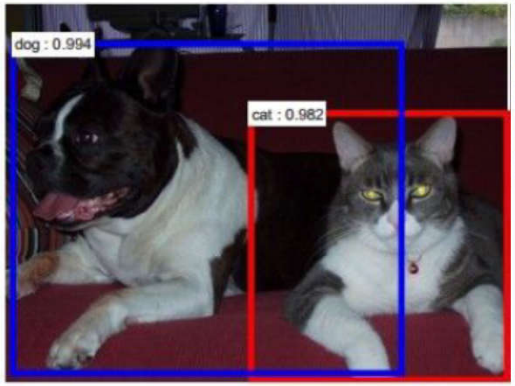
- Traitement automatique de langage naturel

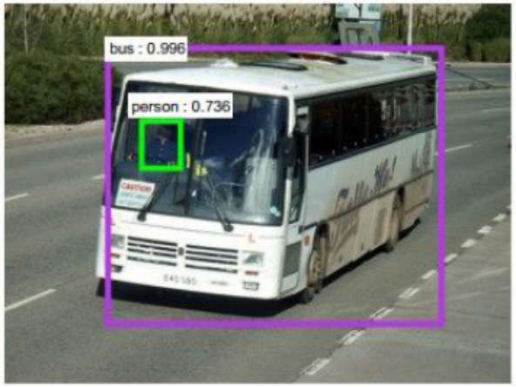
  Traduction, classification, génération automatique de textes

- Nombreux champs d’application futur :
  Sciences, Ingénierie, Marketing, Finance,
  Médecine, Bioinformatique ...



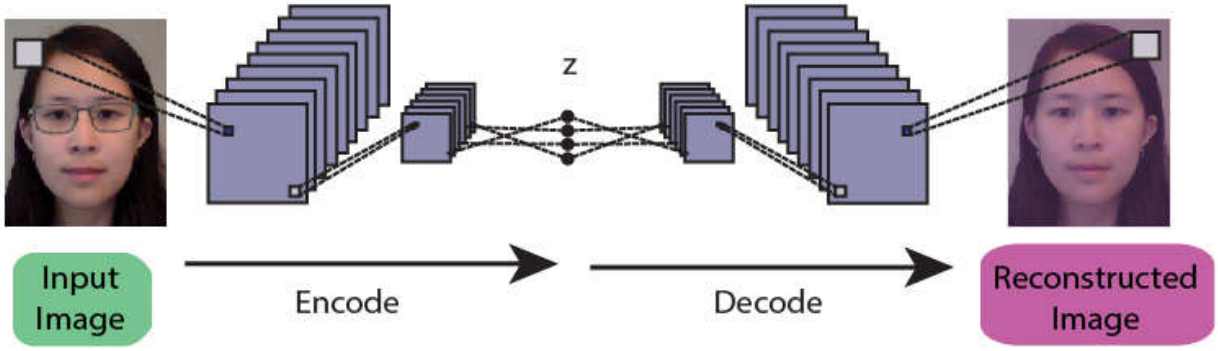






Face **With** Glasses

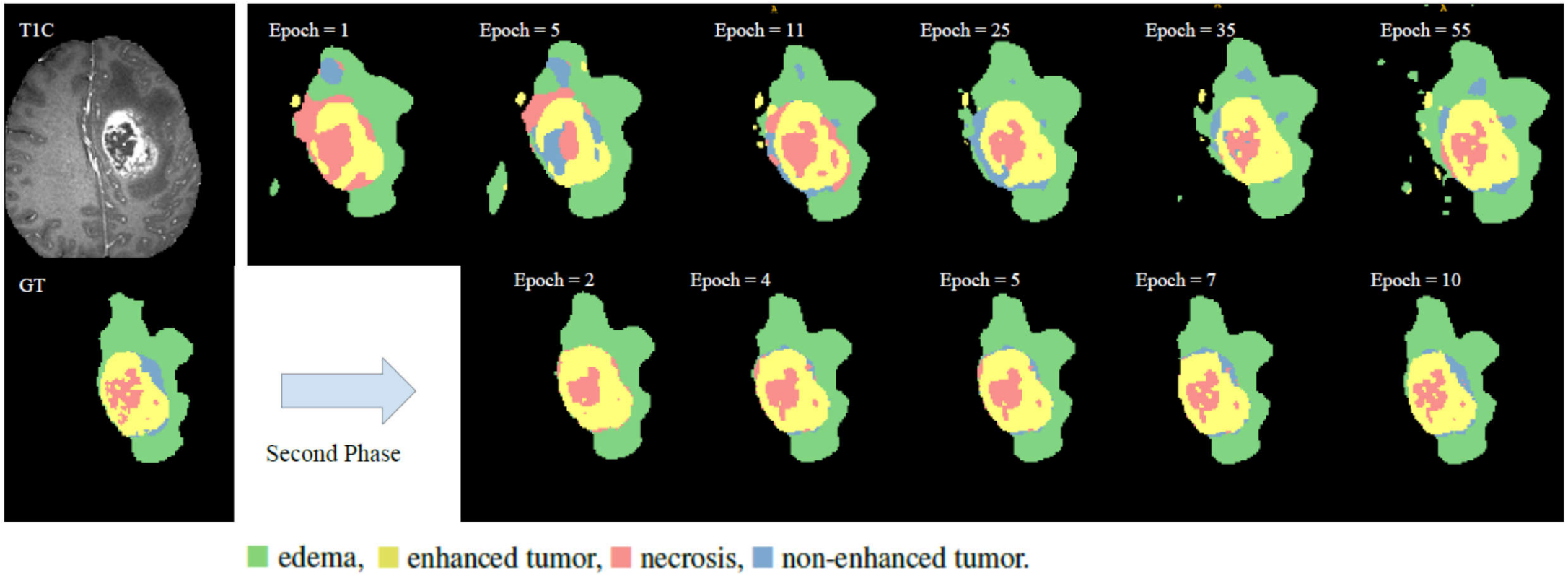
Face, **No** Glasses



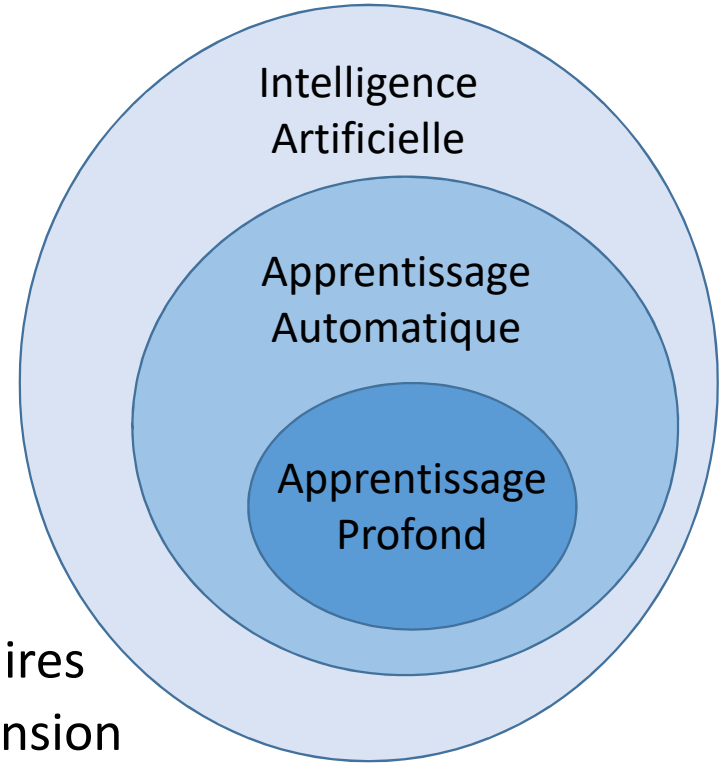
Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

Segmentation d'images médicales

M. Havaei et al. Brain Tumor Segmentation with Deep Neural Networks. *Medical Image Analysis 2016.*



Apprentissage profond

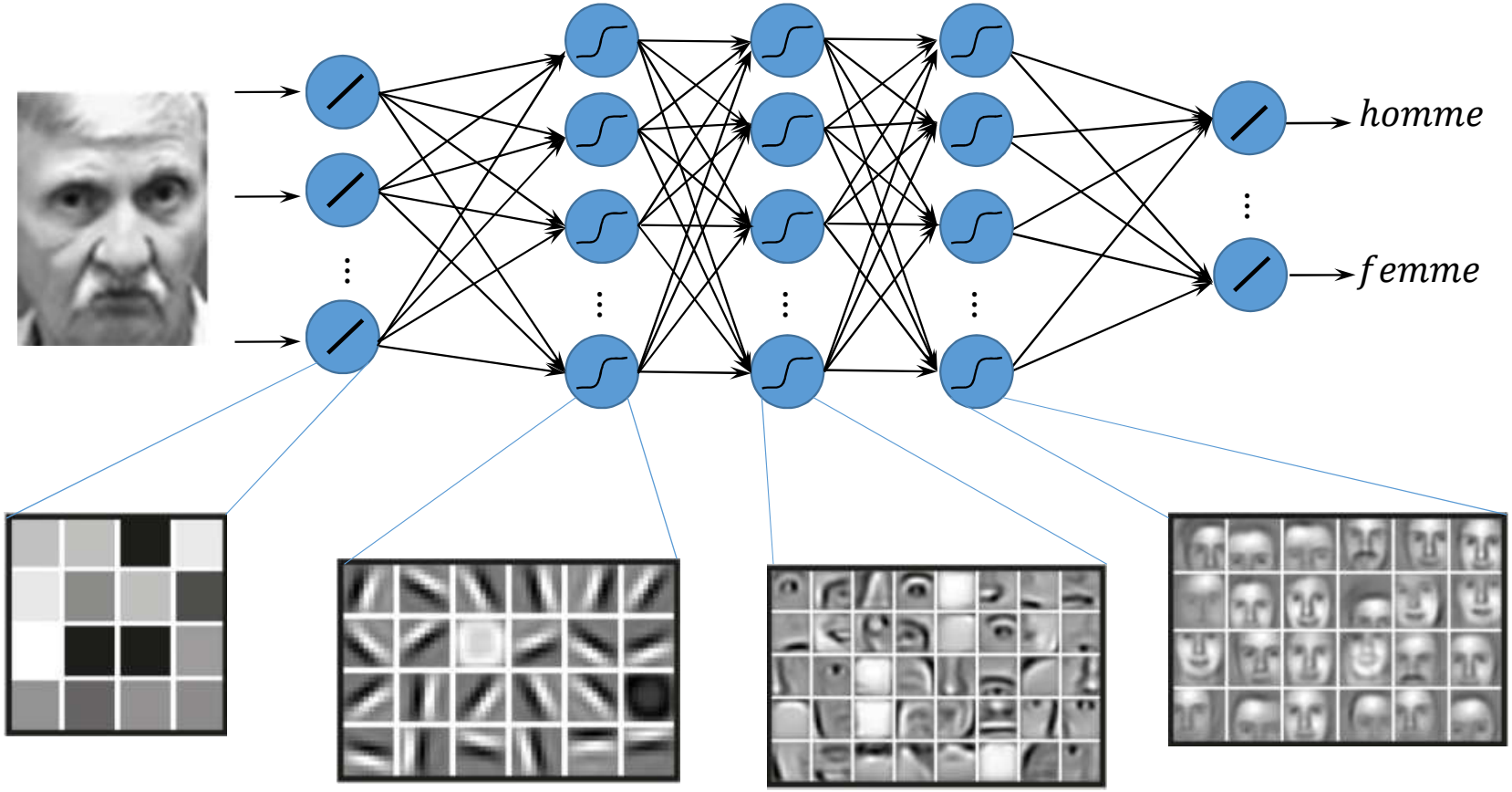


- Sous domaine de l'apprentissage automatique (machine learning)

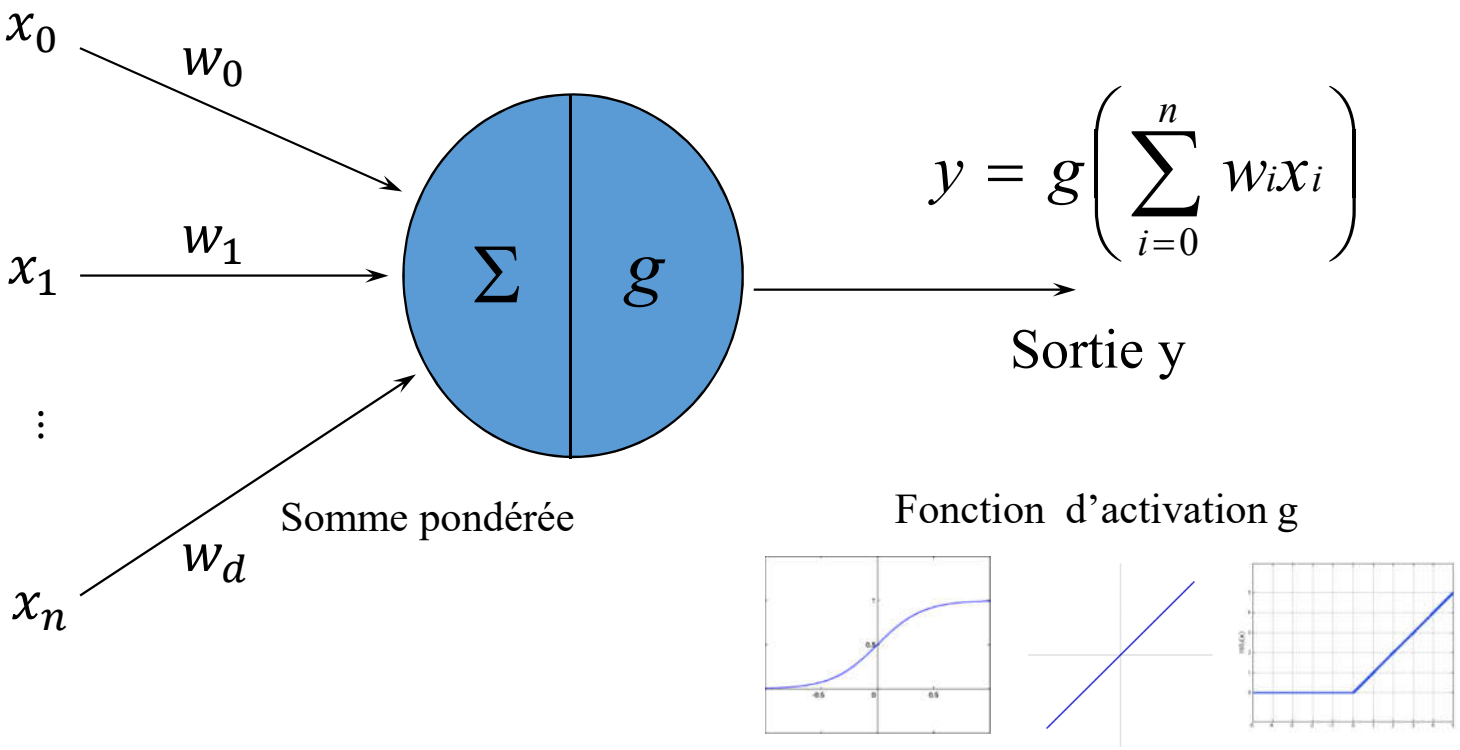
- Réseaux de neurones de grande taille
  - Composition de multiples transformations non linéaires
  - Découvrir des structures complexes en grande dimension

- Construction d'une nouvelle représentation des données
  - Plusieurs niveaux d'abstraction
  - Donner un sens aux données

Apprentissage profond

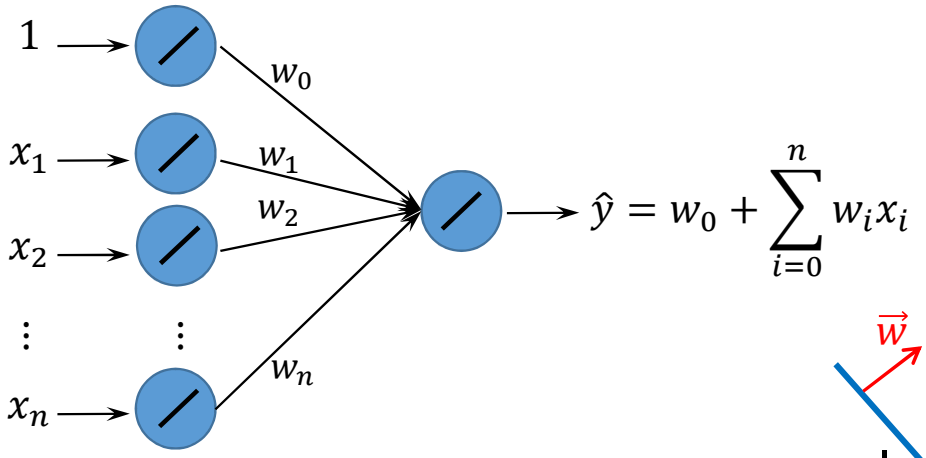


Neurone formel

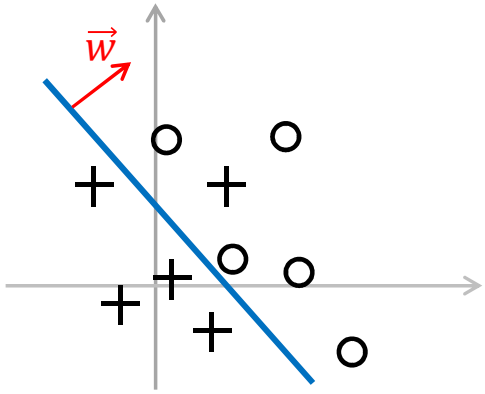


Perceptron

1ere génération de réseaux de neurones (50-60's): perceptron

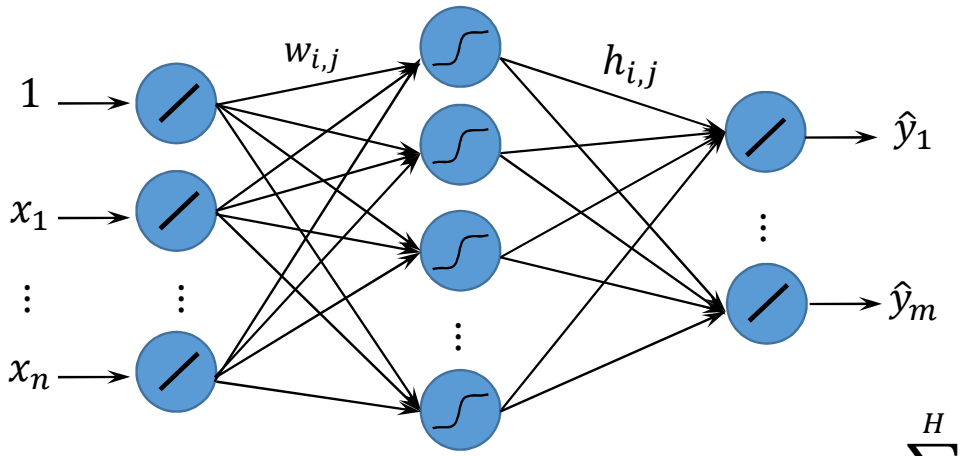


Limité a des modèles linaires [Rosenblatt 62]



Perceptron multicouches

2ème génération de réseaux de neurones (80's-90's) : perceptron multicouches

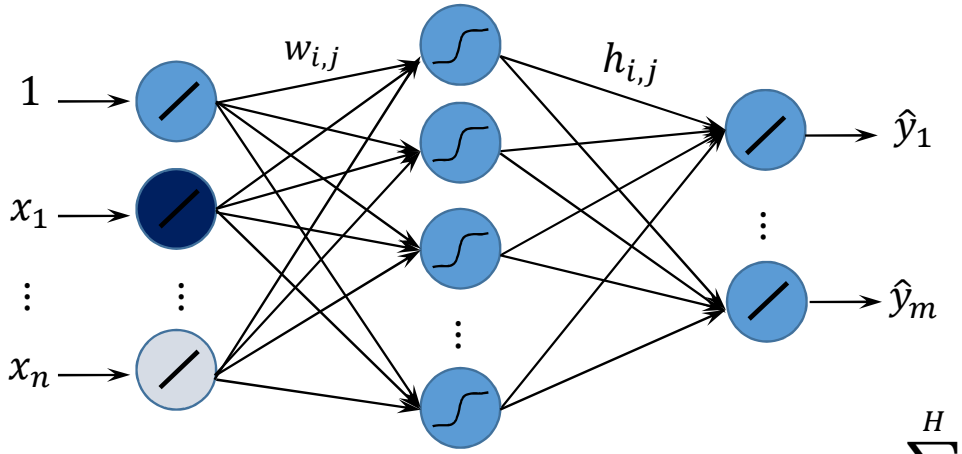


$$\hat{y}_i = h_{i,0} + \sum_{j=1}^{H} h_{i,j} . g\left(w_{j,0} \sum_{k=1}^{n} w_{jk} x_k \right)$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Propagation d'un exemple d'apprentissage $\{(x^{(1)}, \dots, x^{(n)}); (y^{(1)}, \dots, y^{(m)})\}$

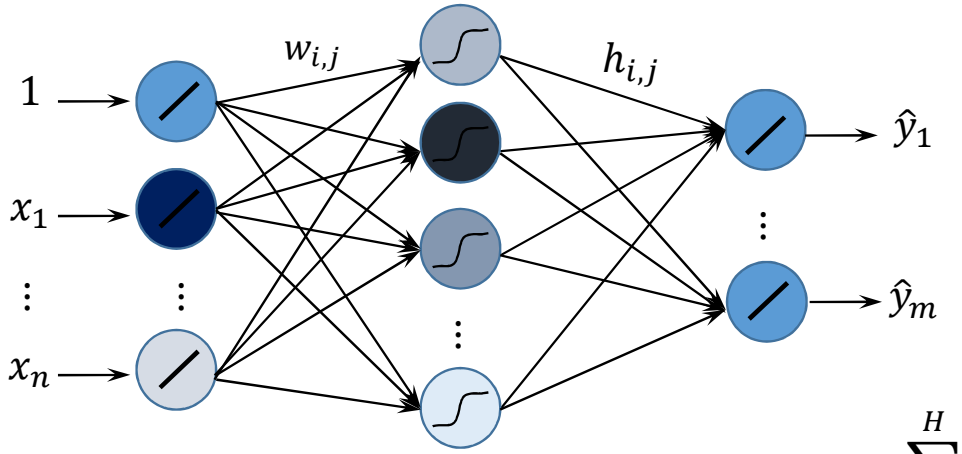


$$\hat{y}_i = h_{i,0} + \sum_{j=1}^{H} h_{i,j} . g\left(w_{j,0} \sum_{k=1}^{n} w_{jk} x_k\right)$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Propagation d'un exemple d'apprentissage $\{(x^{(1)}, \dots, x^{(n)}); (y^{(1)}, \dots, y^{(m)})\}$

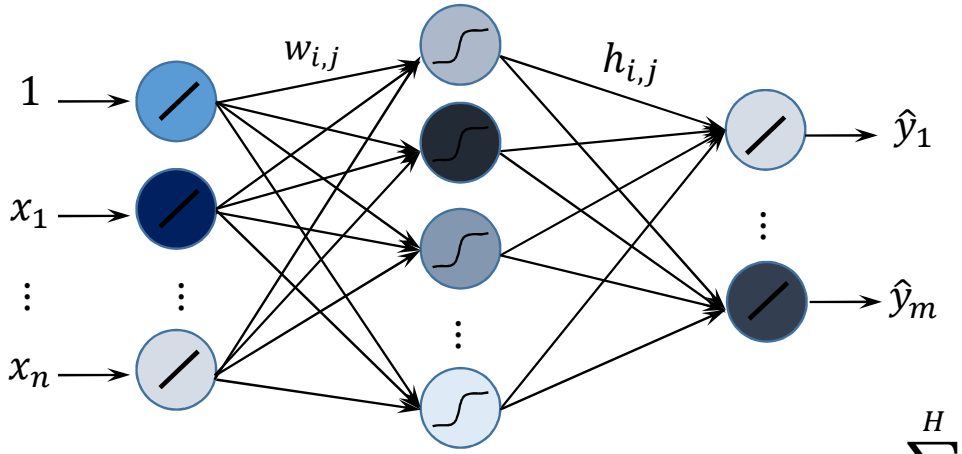


$$\hat{y}_i = h_{i,0} + \sum_{j=1}^{H} h_{i,j} . g\left(w_{j,0} \sum_{k=1}^{n} w_{jk} x_k\right)$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Propagation d'un exemple d'apprentissage $\{(x^{(1)}, \dots, x^{(n)}); (y^{(1)}, \dots, y^{(m)})\}$

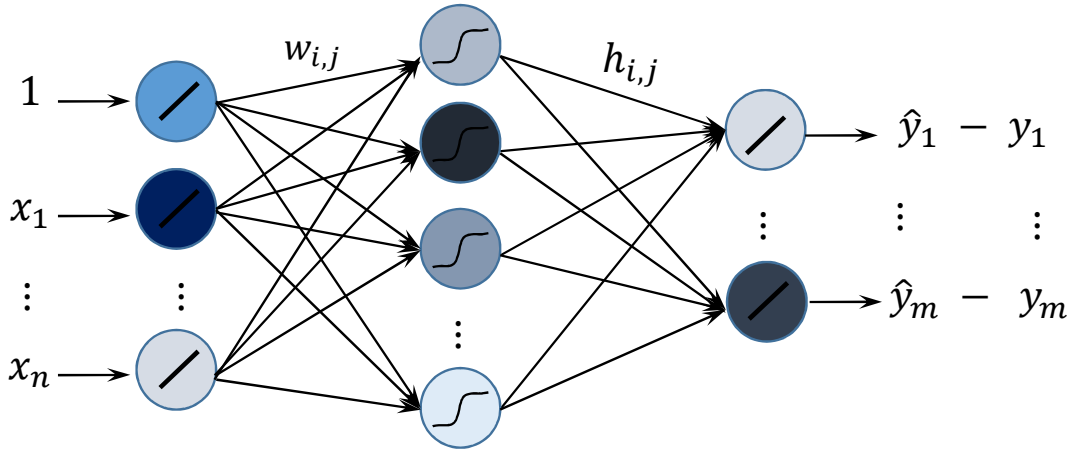


$$\hat{y}_i = h_{i,0} + \sum_{j=1}^{H} h_{i,j} . g\left(w_{j,0} \sum_{k=1}^{n} w_{jk} x_k\right)$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Calcul d' erreur de prédiction sur $\{(x_1, \dots, x_n); (y_1, \dots, y_m)\}$

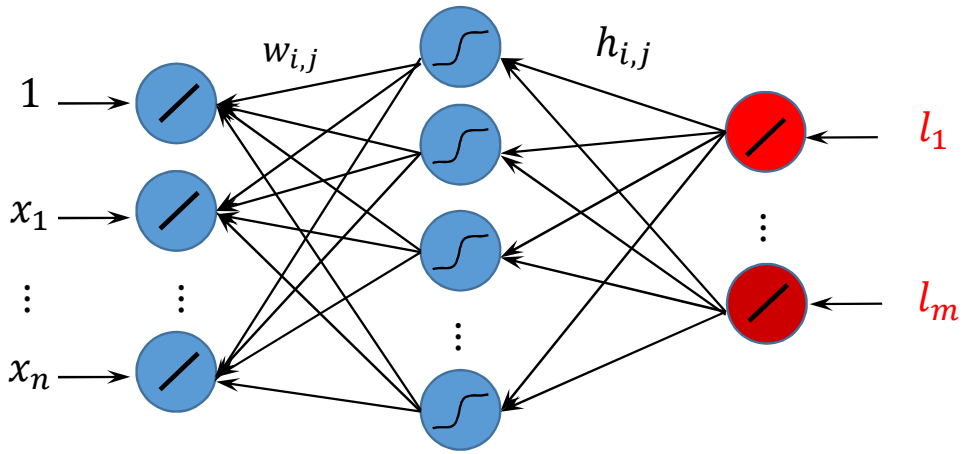


$$L(Y, \hat{Y}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Rétro-propagation de l'erreur et correction des connexions

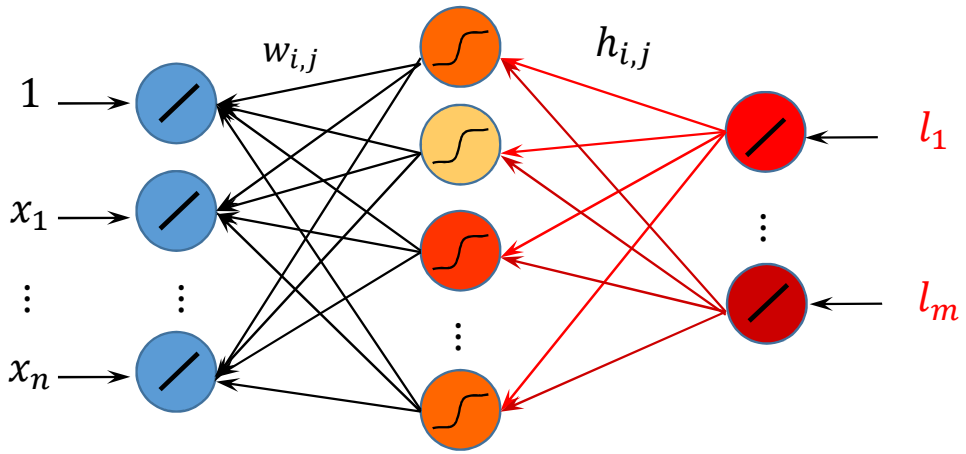


$$L(Y,\hat{Y}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Rétro-propagation de l'erreur et correction des connexions

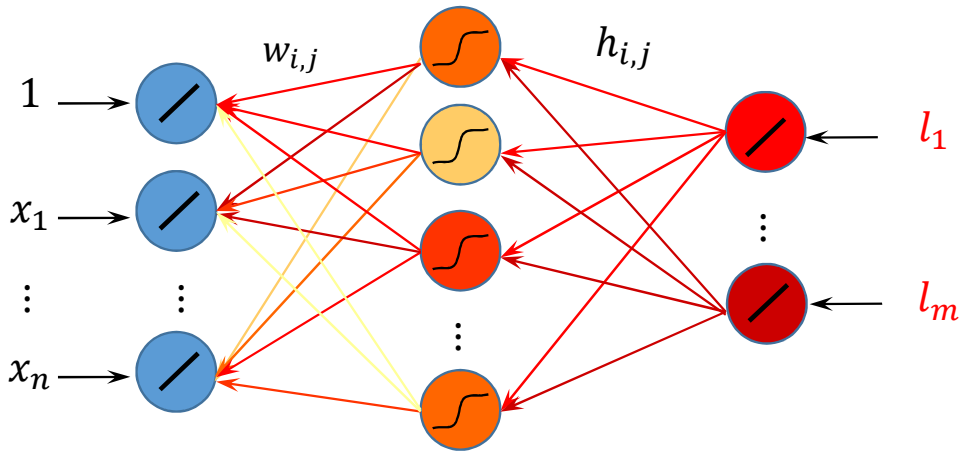


$$L(Y,\hat{Y}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Rétro-propagation de l'erreur et correction des connexions



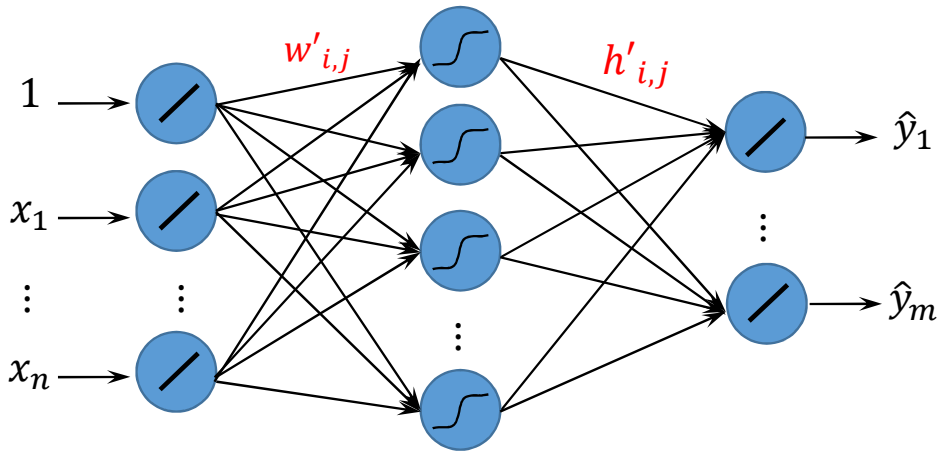
$$w'_{i,j} = w_{i,j} - \eta \frac{\partial L}{\partial w_{i,j}} \qquad h'_{i,j} = h_{i,j} - \eta \frac{\partial L}{\partial h_{i,j}}$$

$$L(Y, \hat{Y}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Perceptron multicouches

Construction à l'aide d'un ensemble d'apprentissage

Rétro-propagation de l'erreur et correction des connexions



$$w'_{i,j} = w_{i,j} - \eta \frac{\partial L}{\partial w_{i,j}} \qquad h'_{i,j} = h_{i,j} - \eta \frac{\partial L}{\partial h_{i,j}}$$

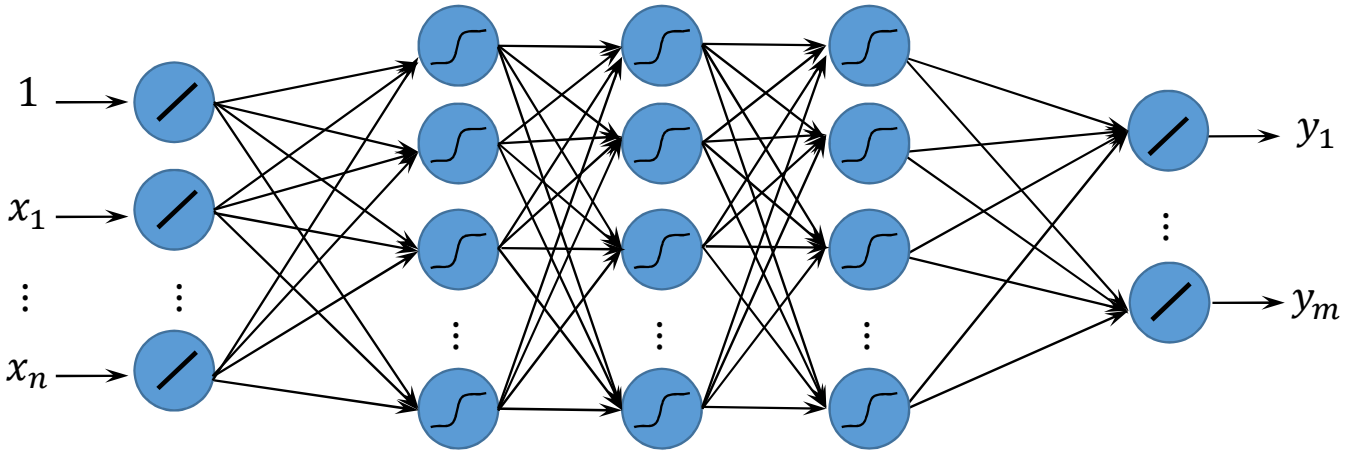
Limitations

- Difficulté à apprendre des réseaux de grande taille
  - Sur-apprentissage
  - Minimum local
  - Temps de calcul

- Performances surpassées par d'autre méthodes
  - SVM, boosting, random forest, …

Apprentissage profond

3ème génération de réseaux de neurones (2010's) : deep learning



Apprentissage profond

Pourquoi l'émergence de l'apprentissage profond maintenant ?

3 raisons à l'émergence du deep learning :

- Données massives
- Puissance calcul (GPU)
- Avancées algorithmiques

Beaucoup d'architectures :

- Perceptron multi-couches
- Réseaux de convolution (VGG, ResNet, ….)
- Réseaux récurrents (LSTM,GRU,…)
- Réseaux génératifs (VAE, GAN, …)
- …

Kelley *et al.* Basset: Learning the regulatory code of the accessible genome. *Genome research 2015*

Prédiction de 164 facteurs de transcription et « sites de fixation »

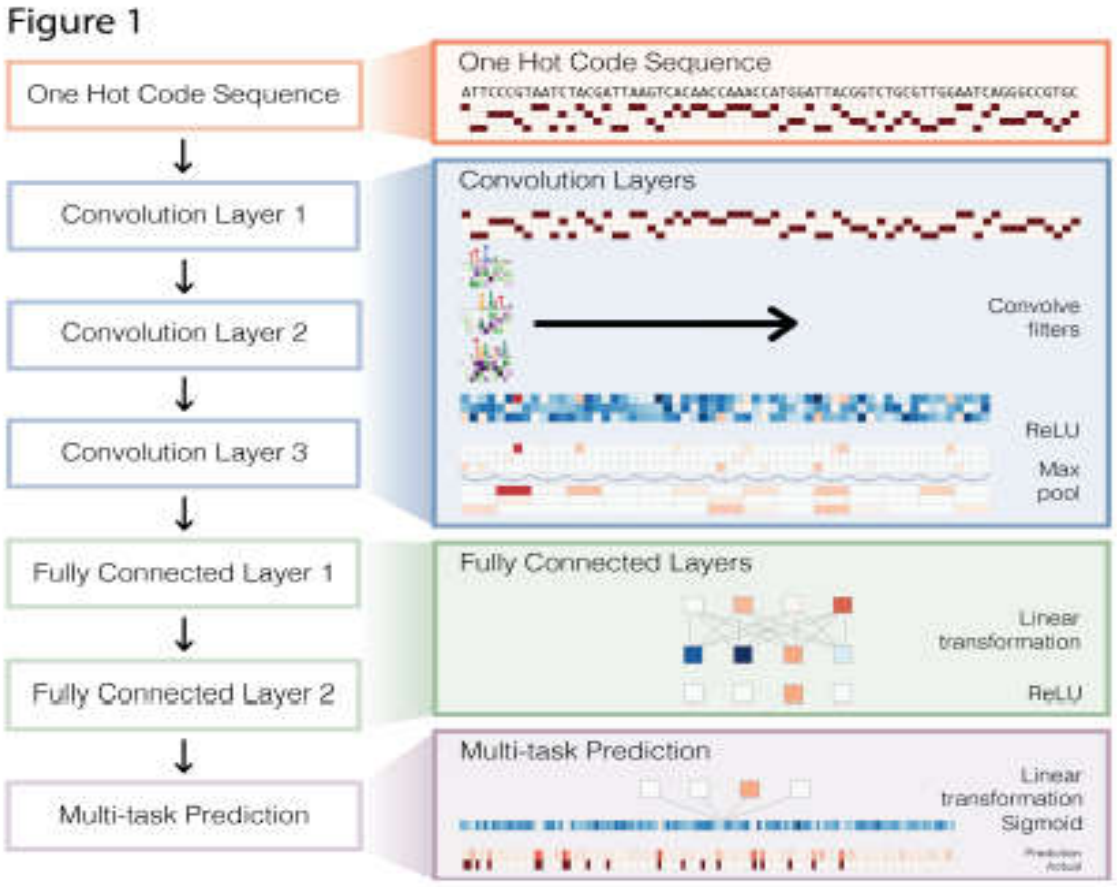
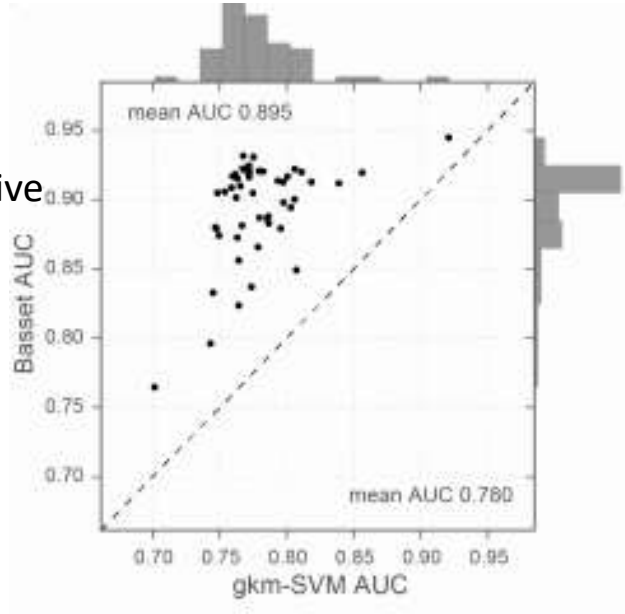
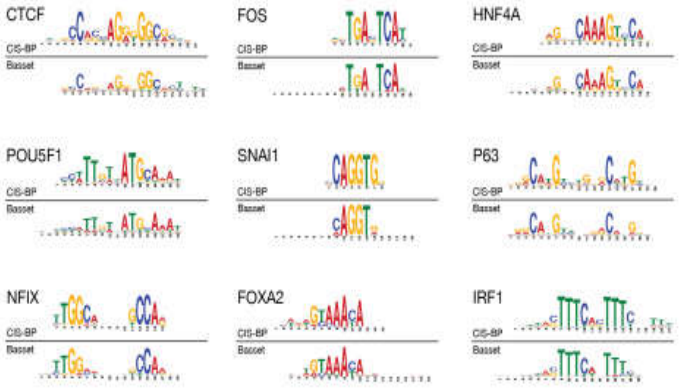


Figure 1 - Deep convolutional neural network for DNA sequence analysis

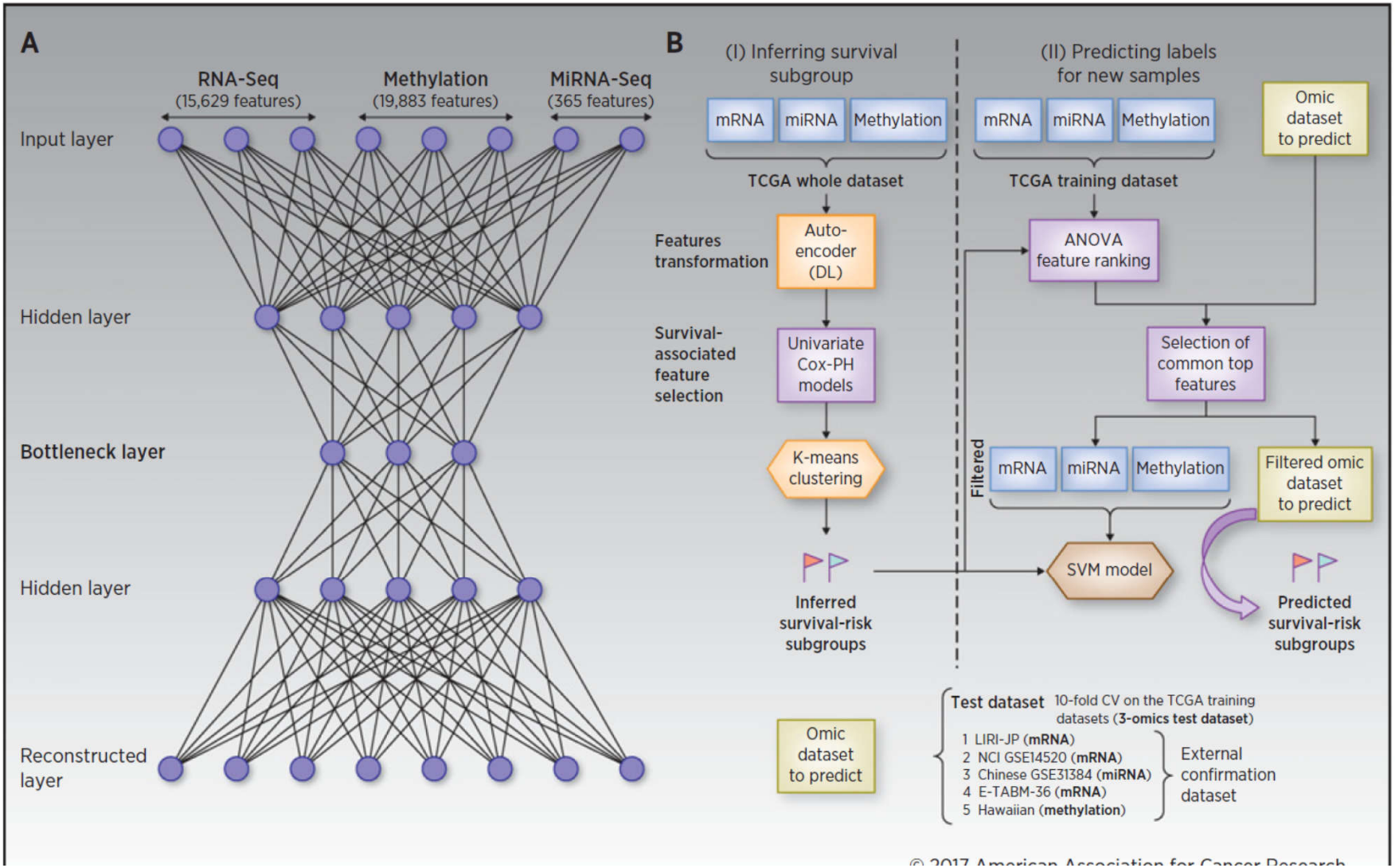
Amélioration significative des performances de prédiction



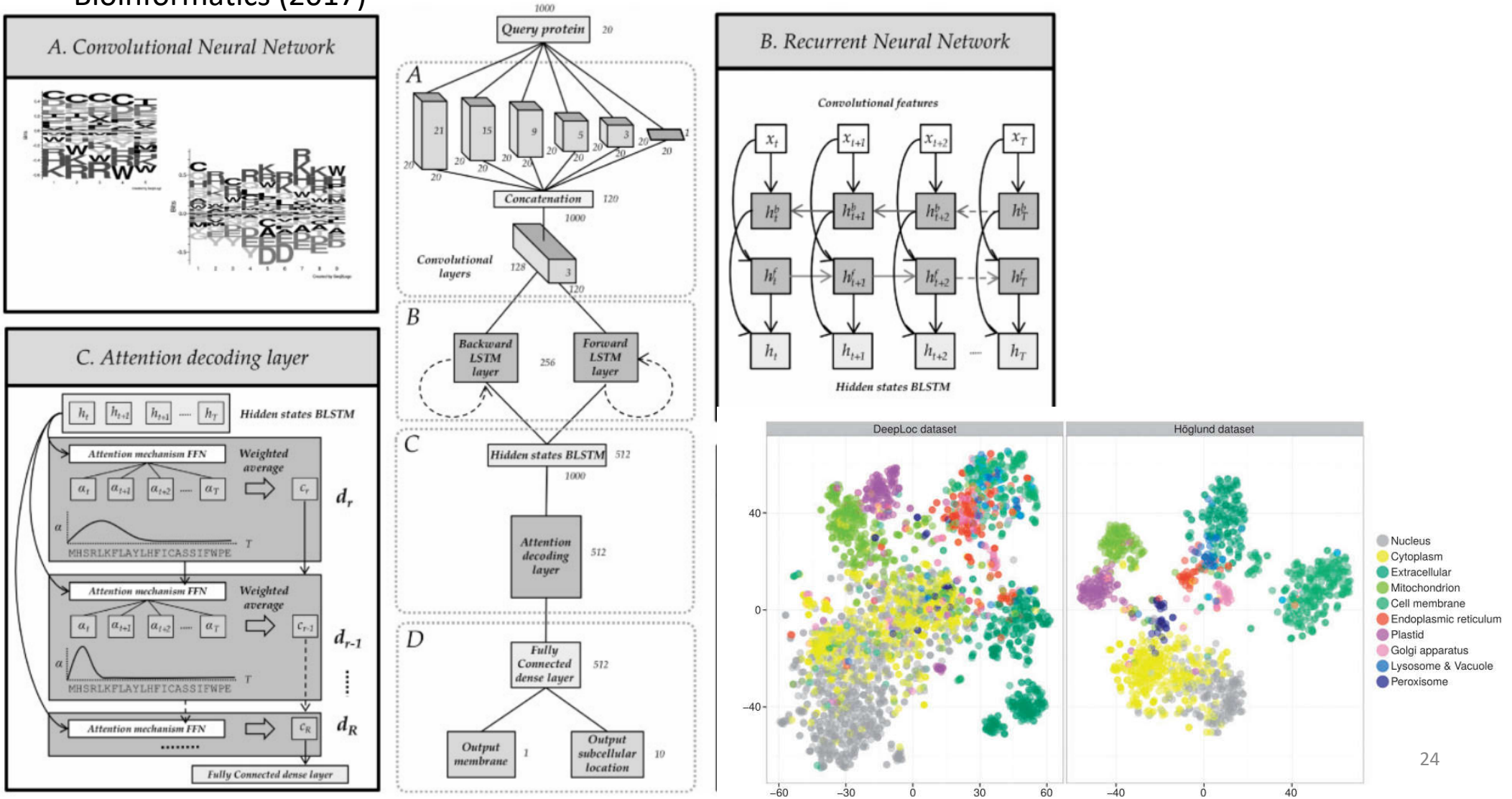
Mise en évidence dans les couches de convolution de séquences connues



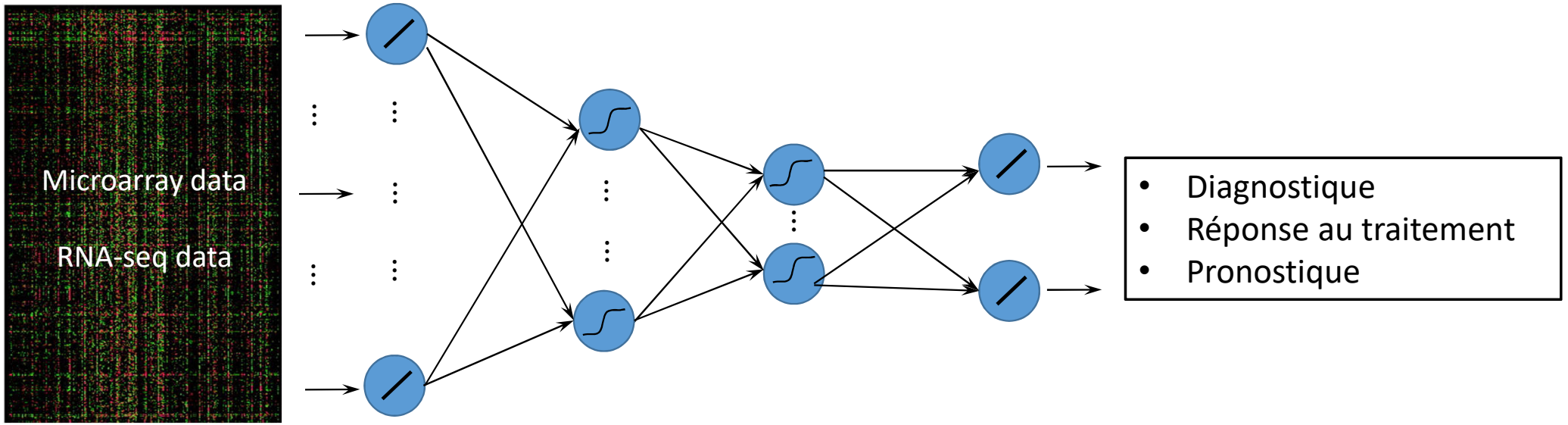
Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, *24*(6), 1248-1259.



Almagro Armenteros DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics (2017)



Prédiction de phénotypes à partir de données d'expression de gènes



Données d'expression de gènes :

- Grand nombres de variables (10,000-50,000)
- Petits nombres de patients (100 – 2,000)
- Grande variabilités des données

Verrous scientifiques

- Apprentissage avec peu de patients
- Intégration des connaissances du domaine
- Robustesse des modèles prédictifs
- Intégration avec d'autres sources de données
- Interprétation biologiques des réseaux de neurones
- Ethique, Equité

Verrous scientifiques

Apprentissage avec peu de patients

- Problème de sur-apprentissage

➢ Comment intégrer des données extérieurs ?

Interprétation biologiques des réseaux de neurones

- Effet boite noire

➢ Comprendre ce que fait un réseau de neurones

➢ Comment expliquer la prédiction d'un modèle ?

Apprentissage avec peu d’exemples

Apprentissage par transfert

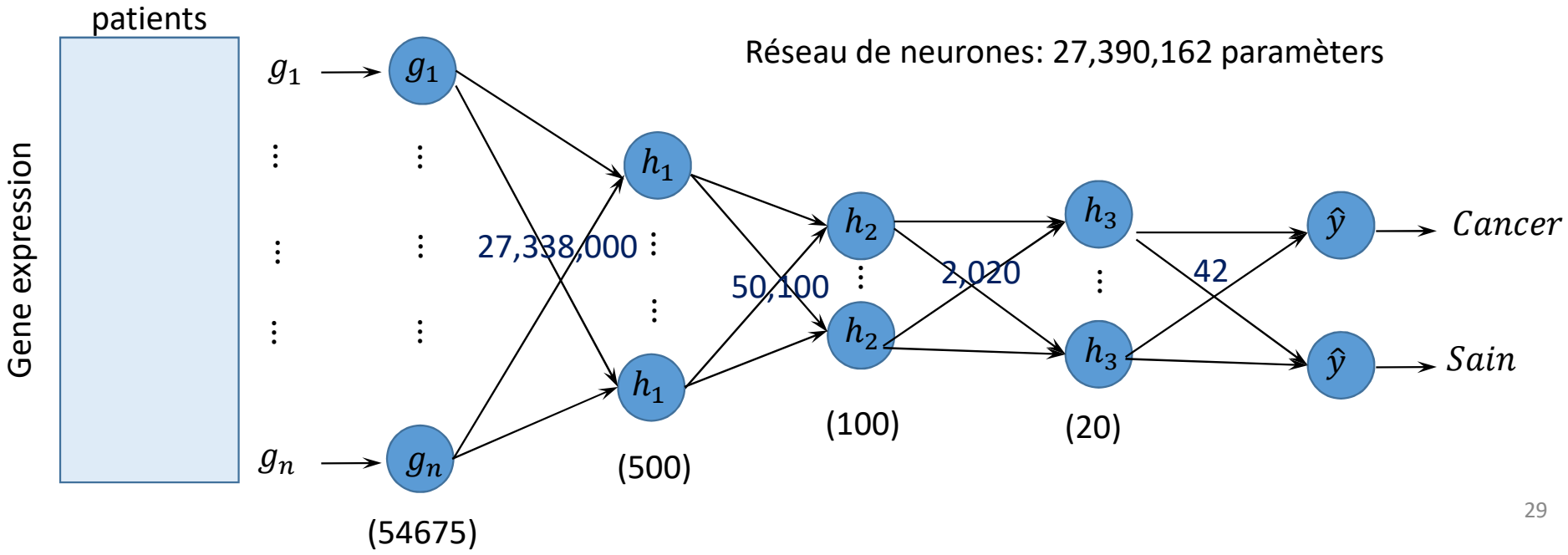
Pré-apprentissage non supervisé

Apprentissage semi-supervisée

Données d'expression

Jeux de données issues de ArrayExpress
Microarray Affymetrix HG-U133A : 54675 probes

Classification à 2 classes (cancer / sain)



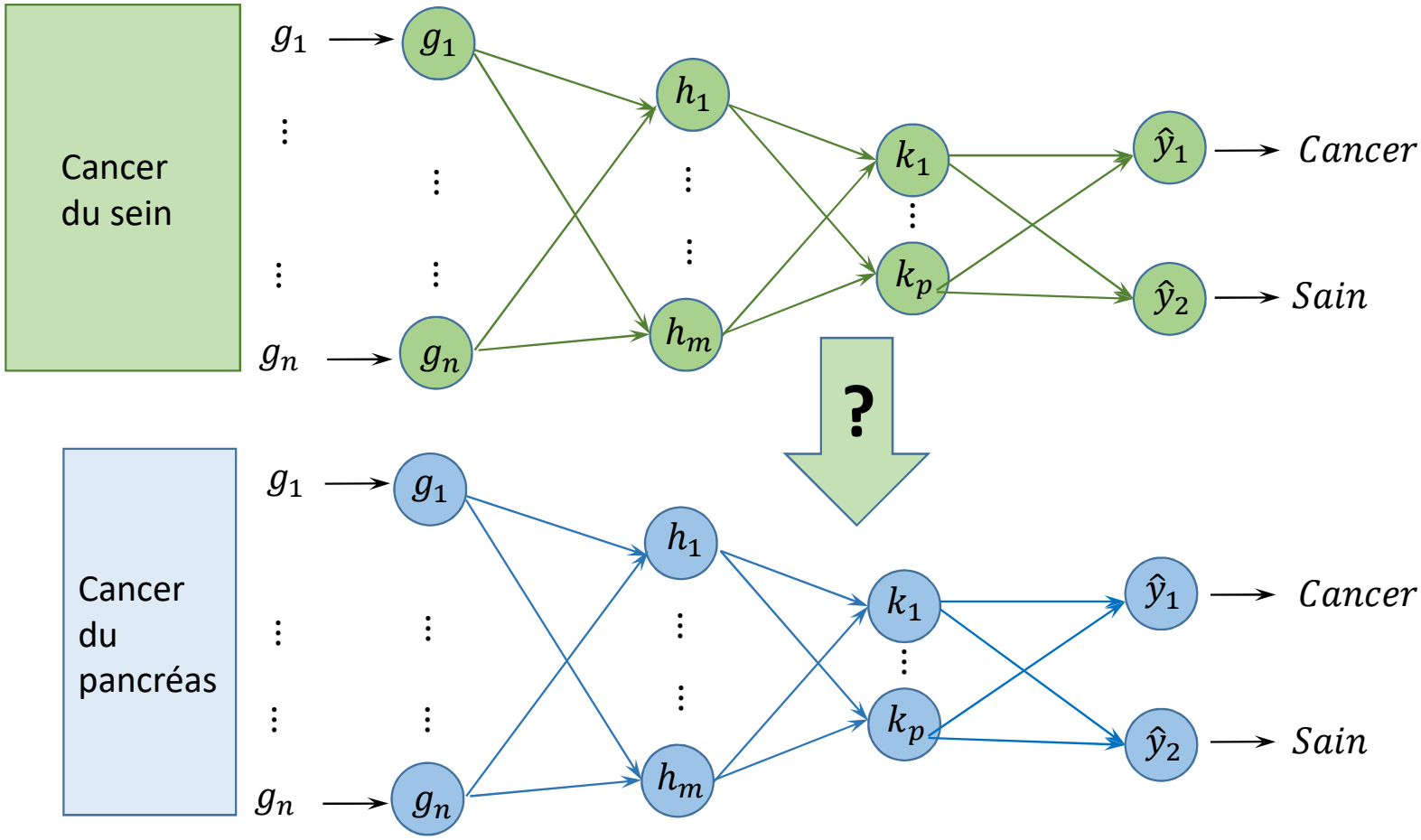
Données d'expression

Jeux de données issues de ArrayExpress
Microarray Affymetrix HG-U133A : 54675 probes

Datasets	#patients	Proportion	Précision
Abdomen	142	0,64	82,4
Sein	2171	0,86	98,0
Rein	657	0,61	93,3
Foie	730	0,82	87,1
Poumon	1415	0,58	85,7
Pancréas	243	0,74	79,6
Peau	835	0,54	75,1
Multi-tissues	27887	0,66	95,2

Torrente, A. et al. (2016). Identification of cancer related genes using a comprehensive map of human gene expression. PLoS One, 11(6).

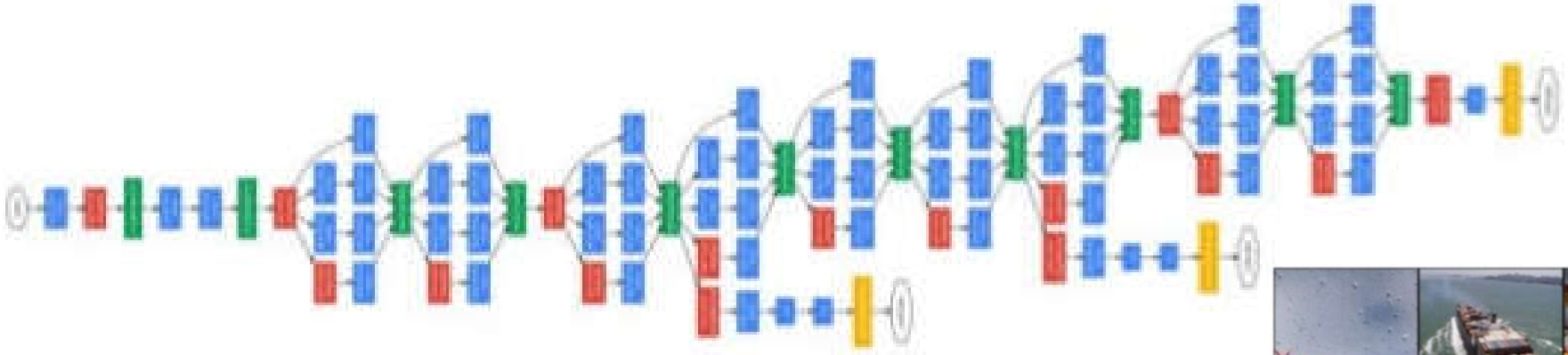
Apprentissage par transfert

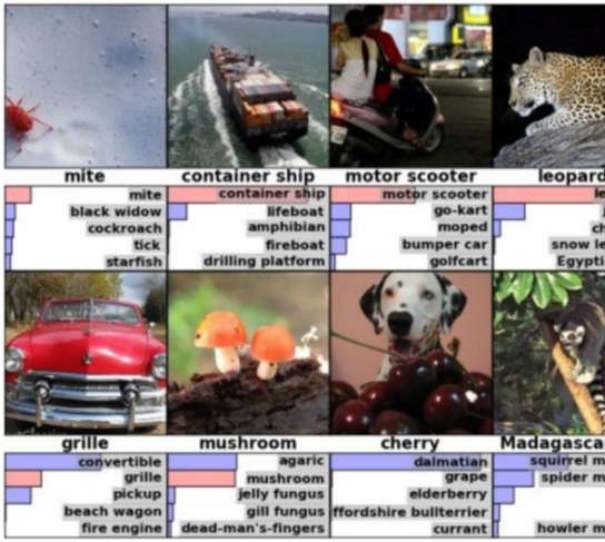


Précision = 98%

Précision = 79,6%

Transfert dans la reconnaissance d'objet

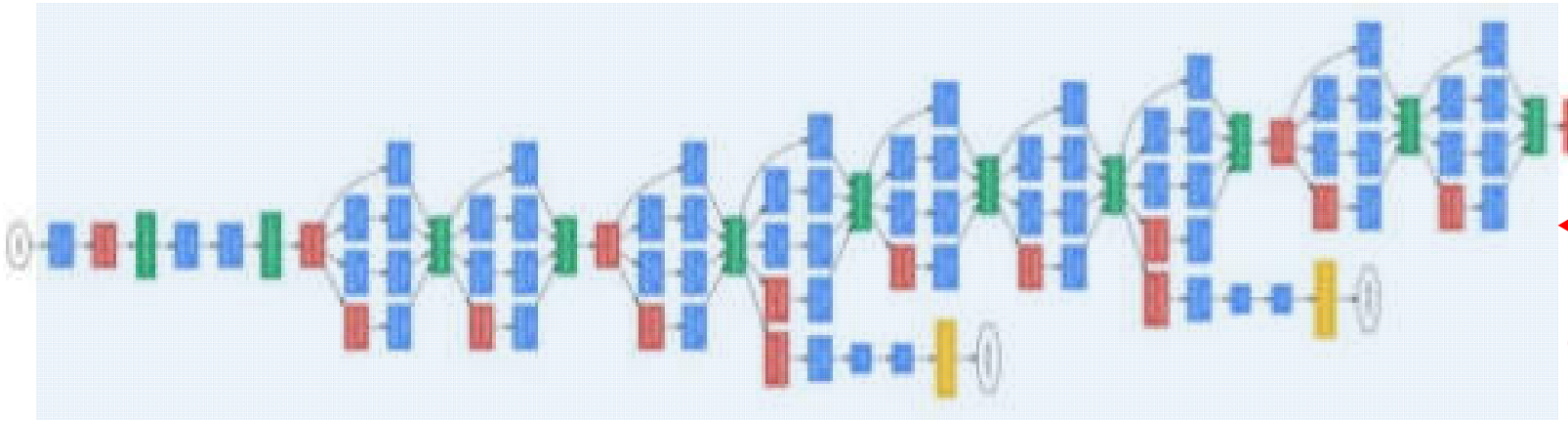




GoogLeNet, ResNet, … :

- Dizaines de couches
- Millions de paramètres
- Millions d'exemples d'apprentissage

Transfert dans la reconnaissance d'objet

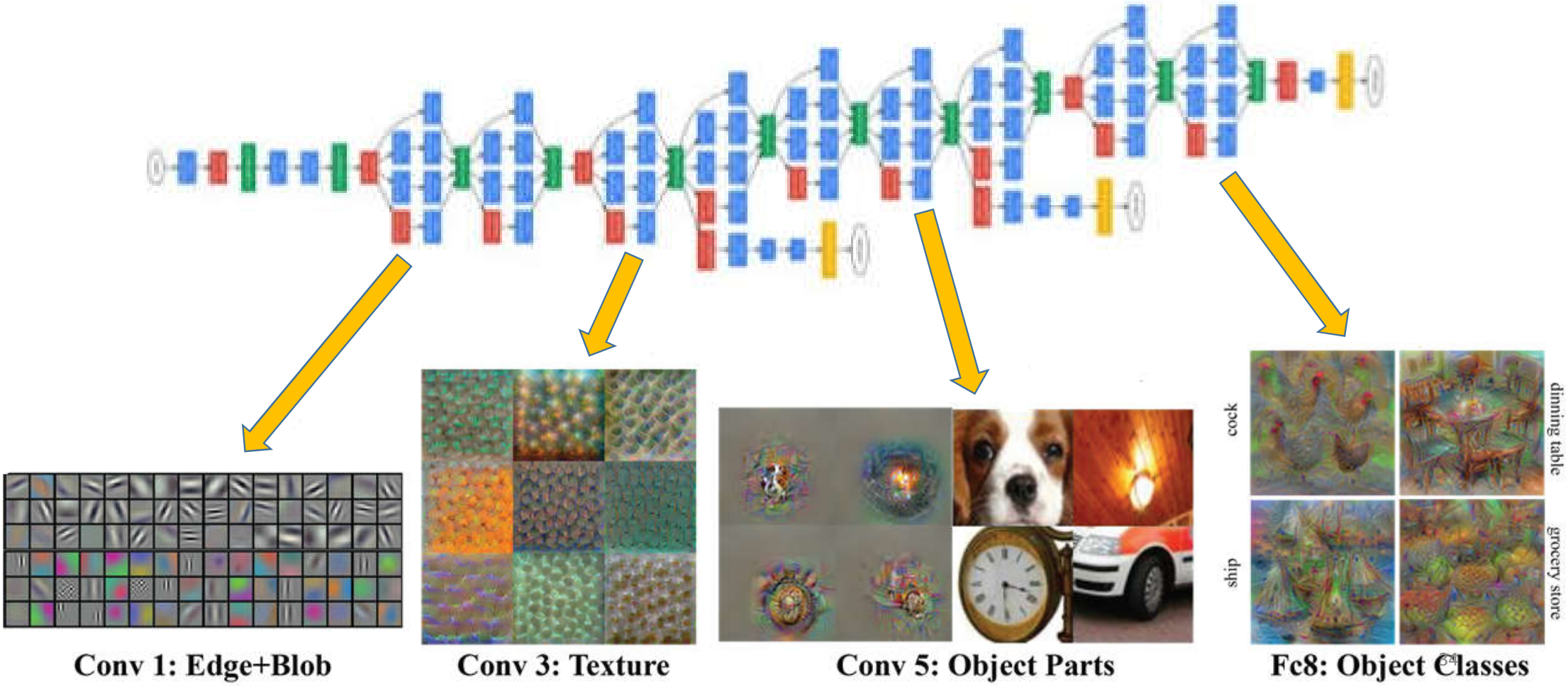


Ré-apprentissage des dernières couches

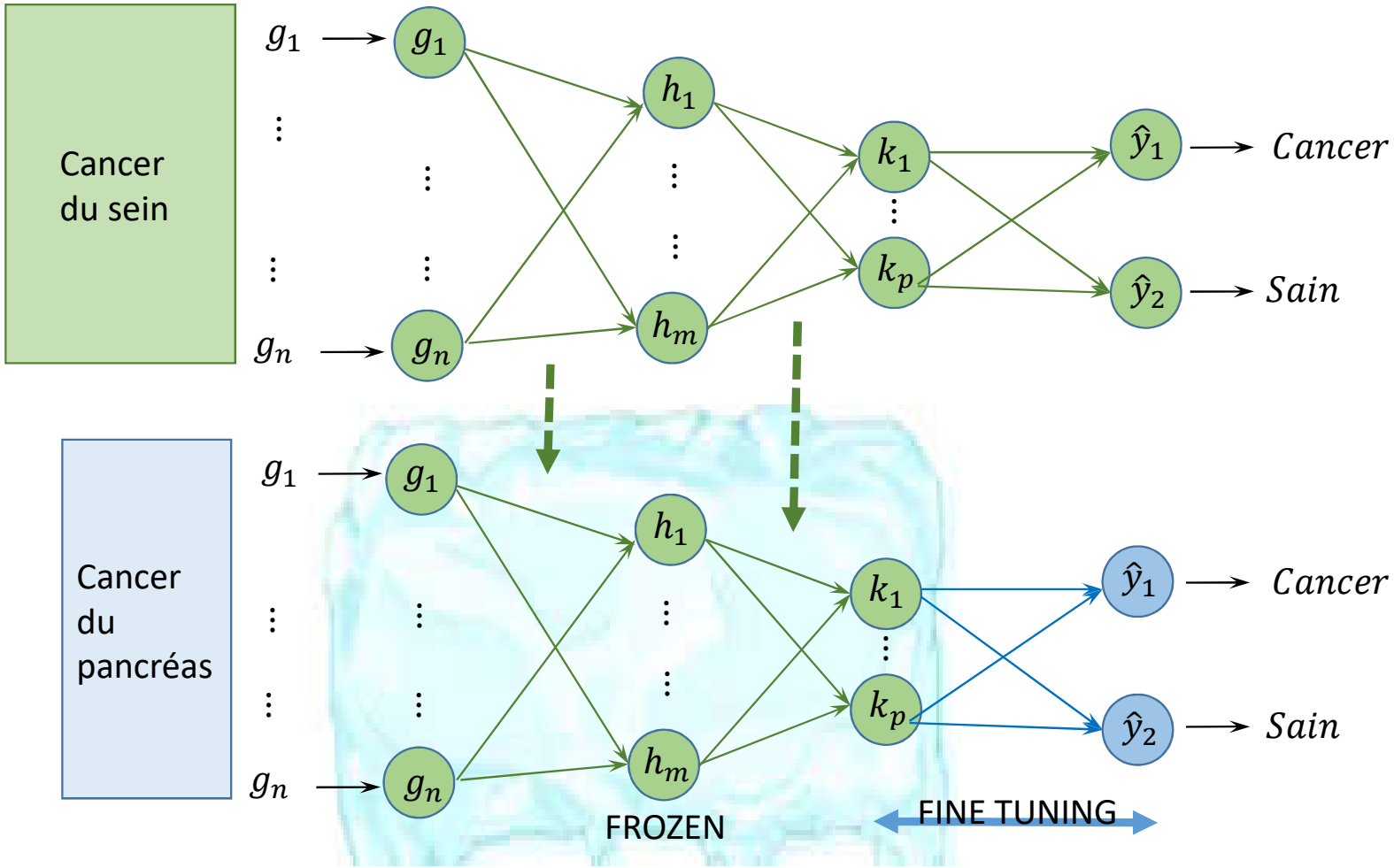
La majorité du réseau est gelée

Réutilisation de la représentation des données apprise dans le réseau d'origine pour une tache spécifique

Transfert dans la reconnaissance d'objet



Apprentissage par transfert

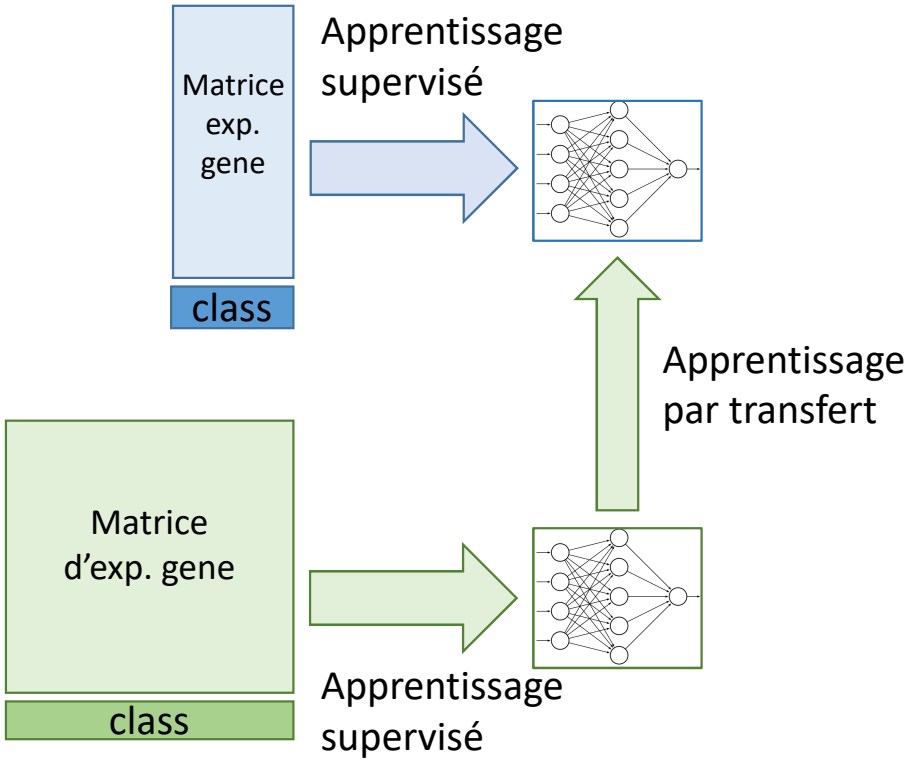


Apprentissage semi-supervisé

Apprentissage par transfert

2 taches de prédiction proches

2 jeux de données étiquetés

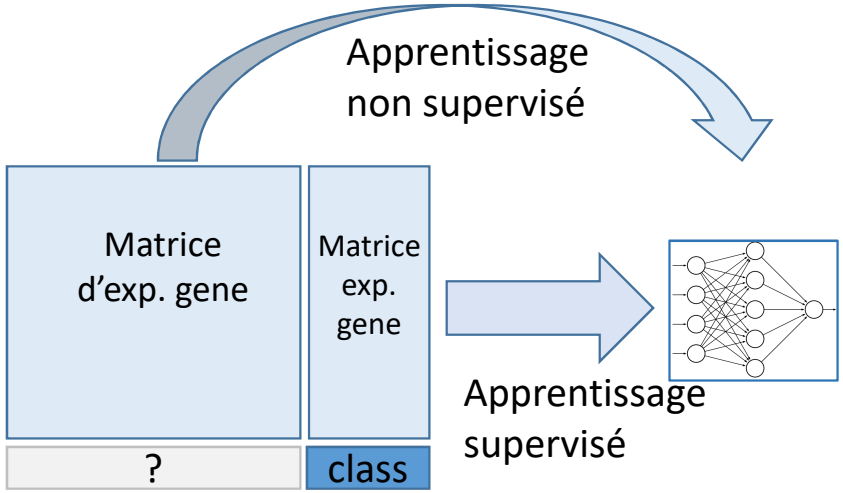


Apprentissage semi-supervisé

1 taches de prédiction

1 jeu de données étiqueté

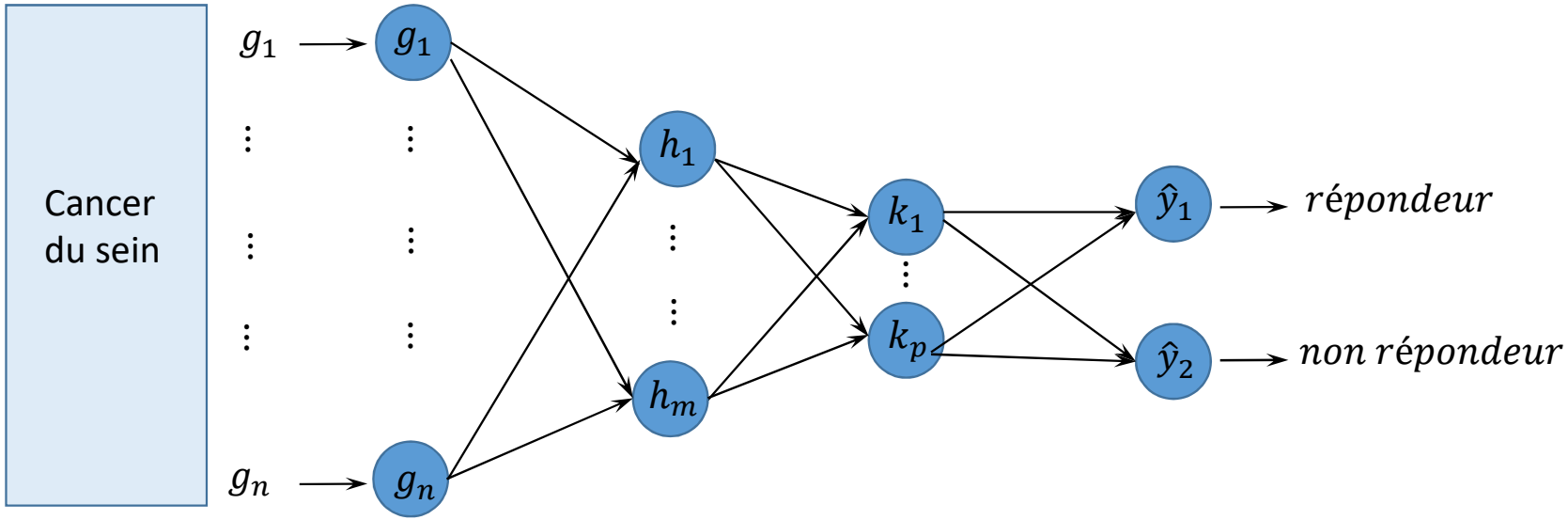
1 jeu de données non étiqueté



Pré-apprentissage non supervisé

Prédiction de la réponse à un traitement au cancer du sein

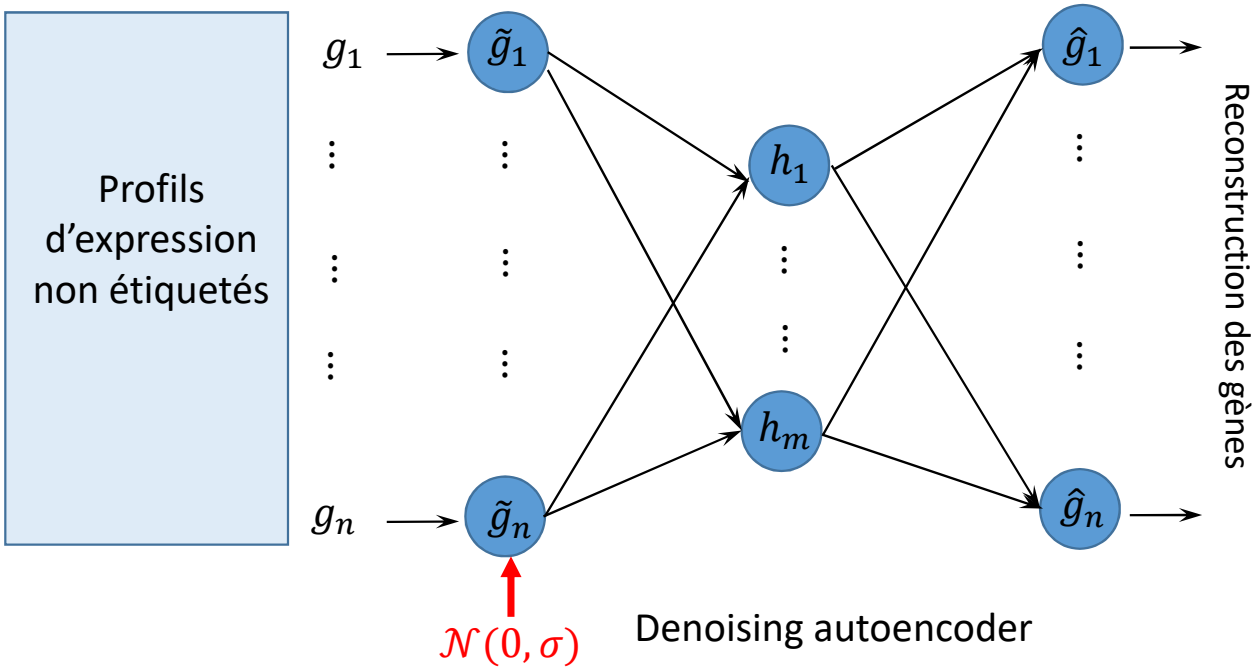
- Jeu de données étiquetés: 388 patients



Pré-apprentissage non supervisé

Utilisation des données non-étiquetés pour initialiser (pré-apprendre) le réseau

Construction de nouvelles variables représentant les données

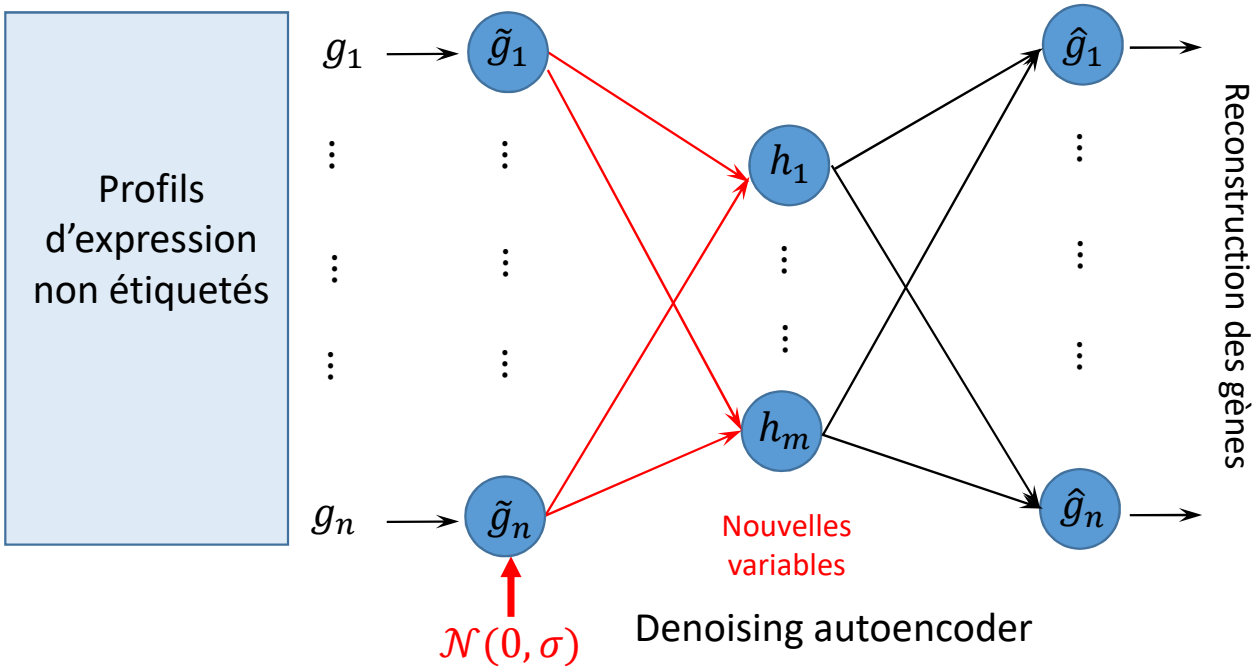


$$L(Y,\hat{Y}) = \frac{1}{2}\sum_{i=1}^{n}(g_i - \hat{g}_i)^2 + \lambda\Omega(W)$$

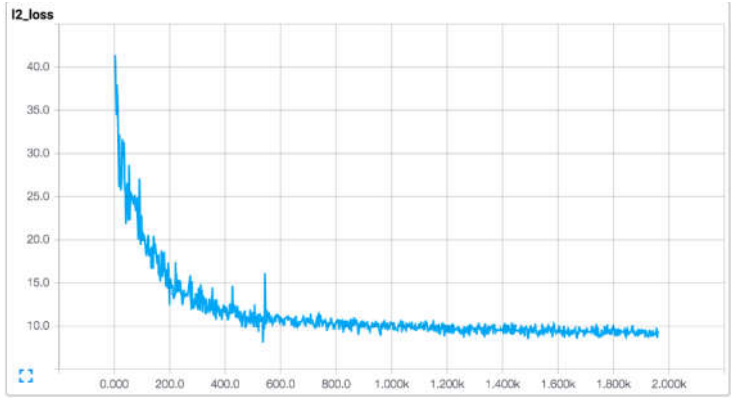
Pré-apprentissage non supervisé

Utilisation des données non-supervisée pour initialiser (pré-apprendre) le réseau

Construction de nouvelles variables représentant les données



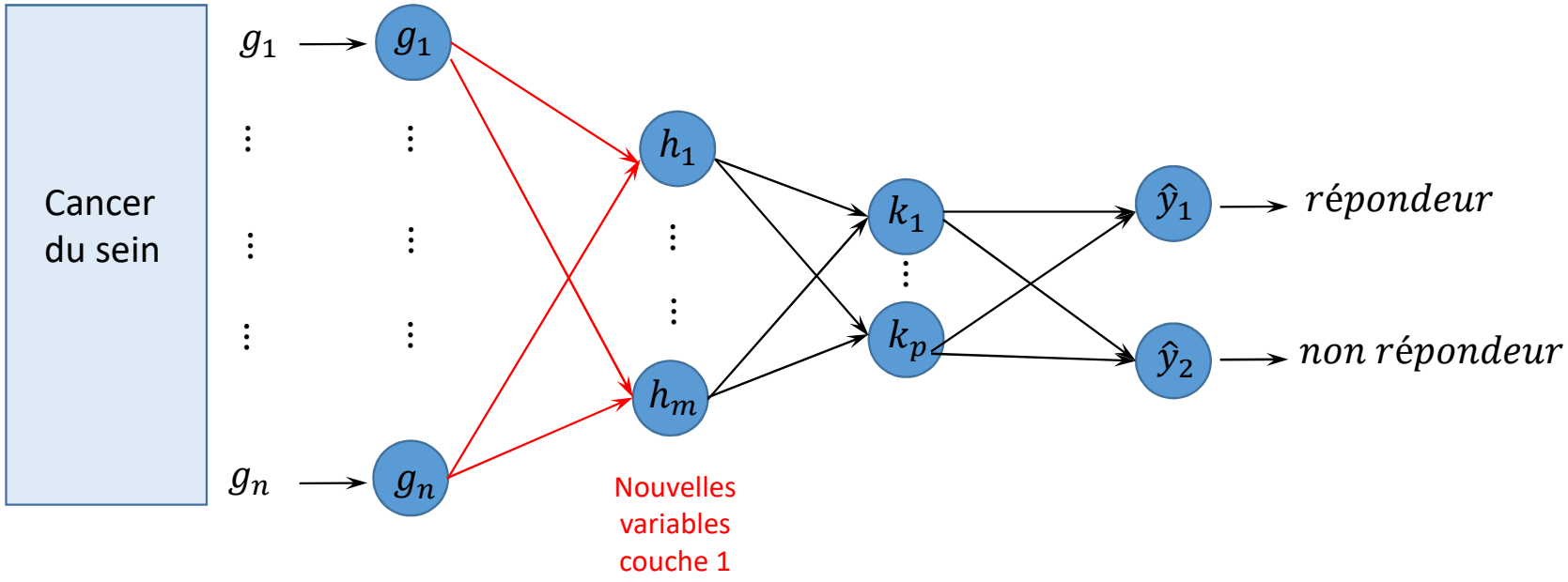
$$L(Y,\hat{Y}) = \frac{1}{2}\sum_{i=1}^{n}(g_i - \hat{g}_i)^2 + \lambda\Omega(W)$$



Pré-apprentissage non supervisé

Utilisation des données non-supervisée pour initialiser (pré-apprendre) le réseau

Construction de nouvelles variables représentant les données

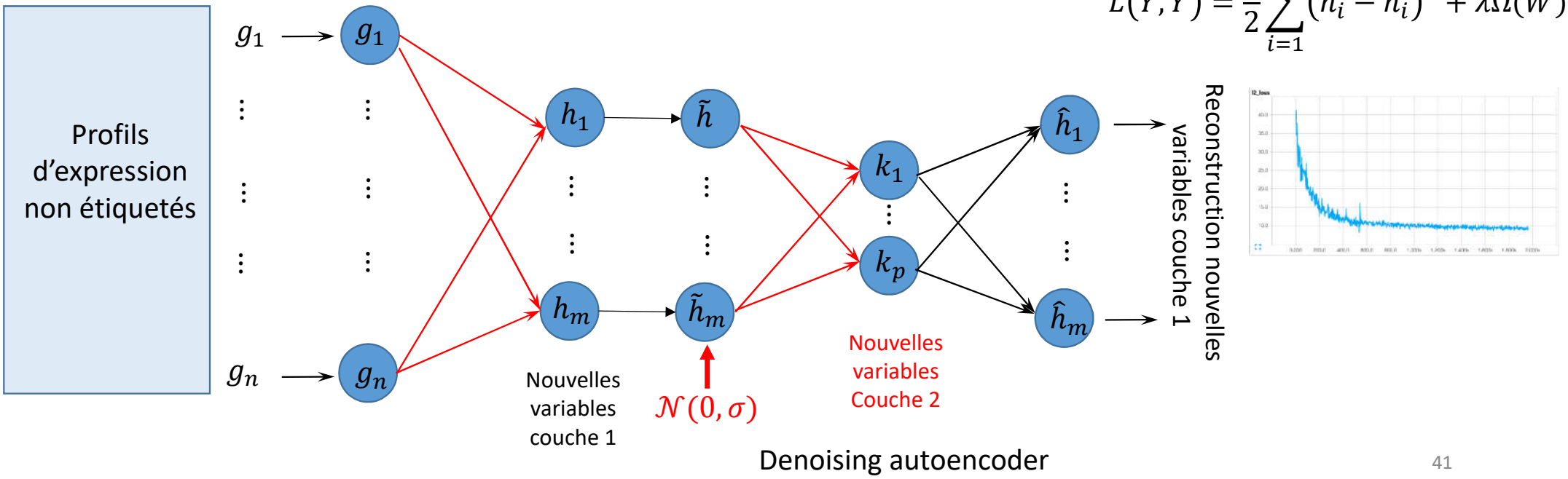


Pré-apprentissage non supervisé

Utilisation des données non-supervisée pour initialiser (pré-apprendre) le réseau

Pré-apprentissage couche par couche

$$L(Y, \hat{Y}) = \frac{1}{2}\sum_{i=1}^{m}(h_i - \hat{h}_i)^2 + \lambda\Omega(W)$$

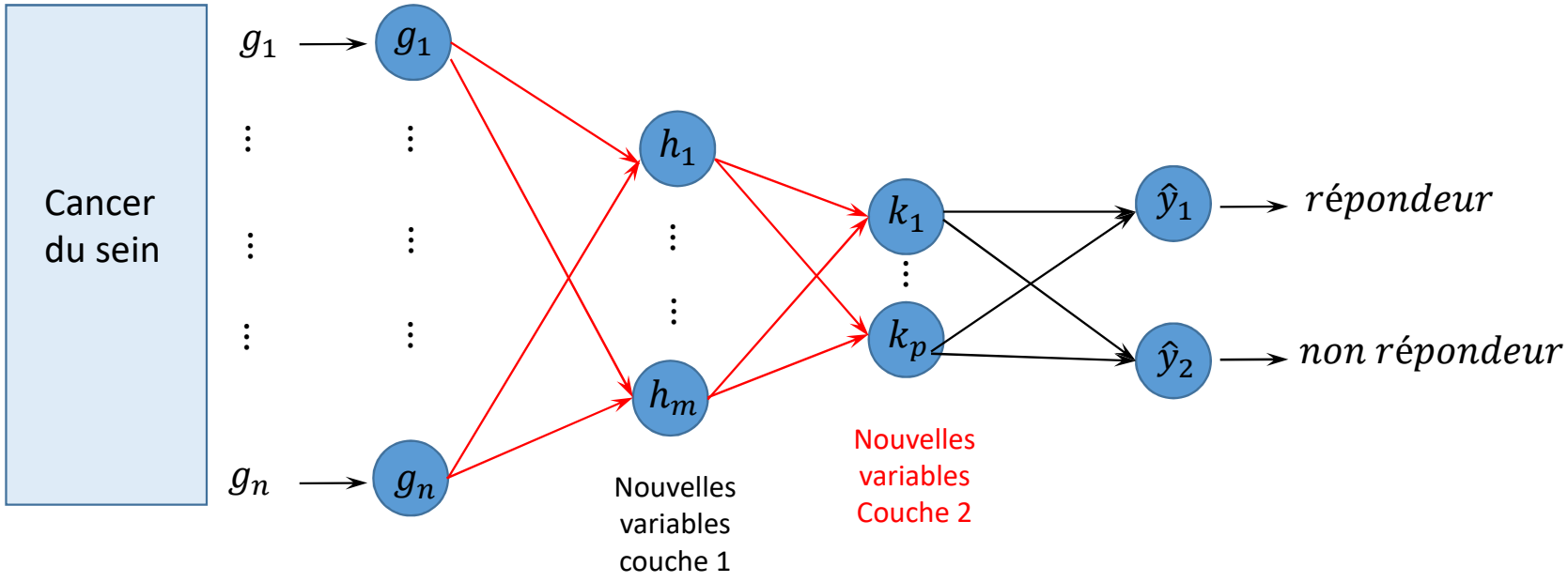


Denoising autoencoder

Pré-apprentissage non supervisé

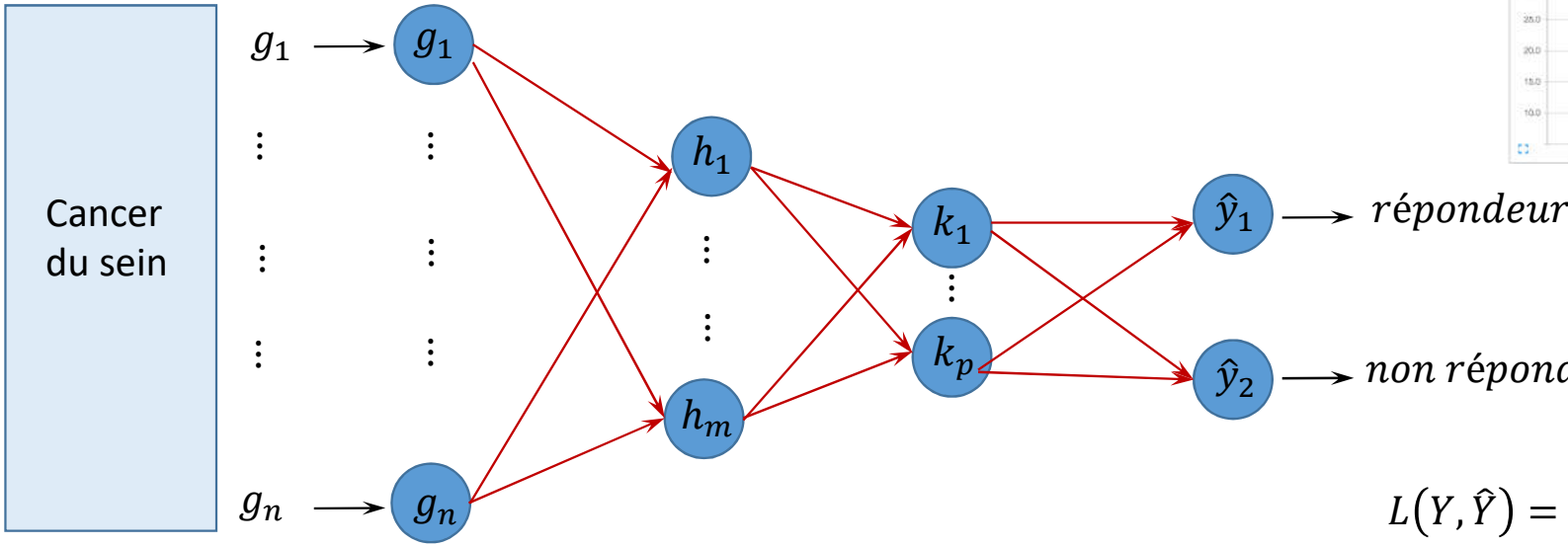
Utilisation des données non-supervisée pour initialiser (pré-apprendre) le réseau

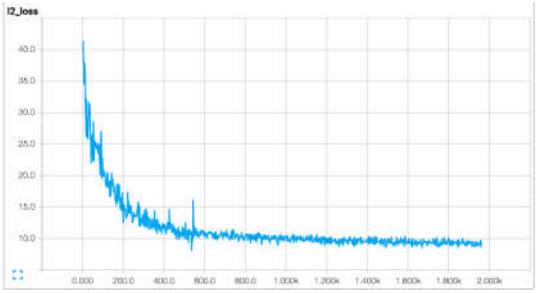
Pré-apprentissage couche par couche



Pré-apprentissage non supervisé

Apprentissage supervisée à partir des nouvelles variables



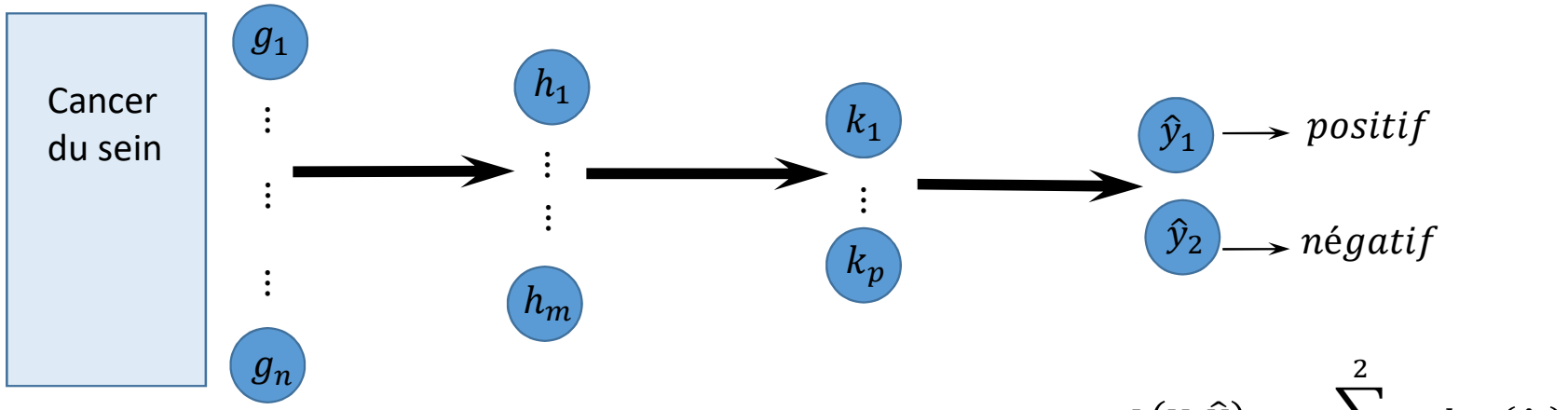


$$L(Y, \hat{Y}) = -\sum_{i=1}^{2} y_i . log(\hat{y}_i) \quad + \lambda\Omega(W)$$

Apprentissage semi-supervisé : Ladder Network

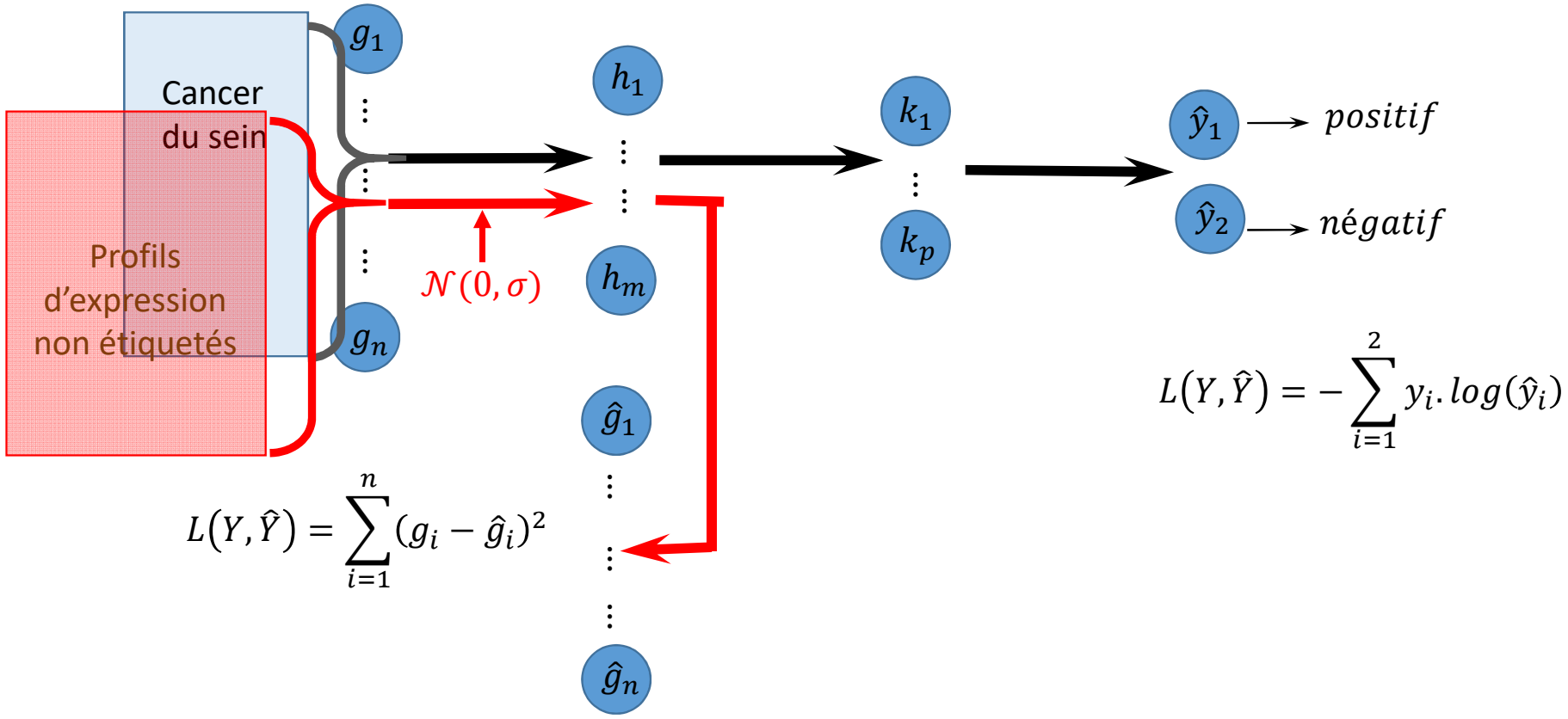
- Utiliser les données étiquetées et non étiquetées simultanément
- Données non étiquetées : apprendre la représentation des connaissances
- Données étiquetées : apprendre à discriminer les classes

Ladder network

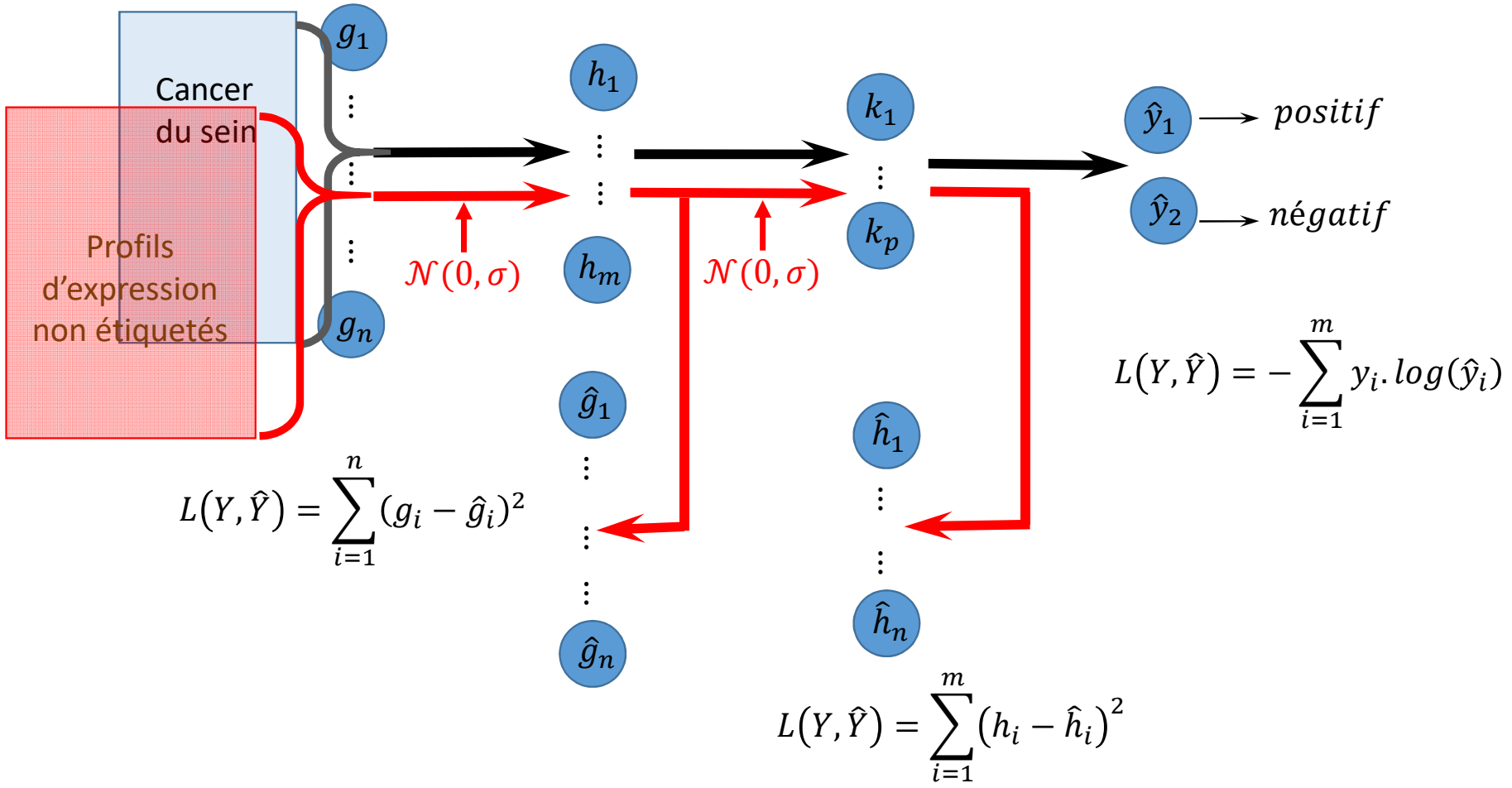


$$L(Y, \hat{Y}) = -\sum_{i=1}^{2} y_i . log(\hat{y}_i)$$

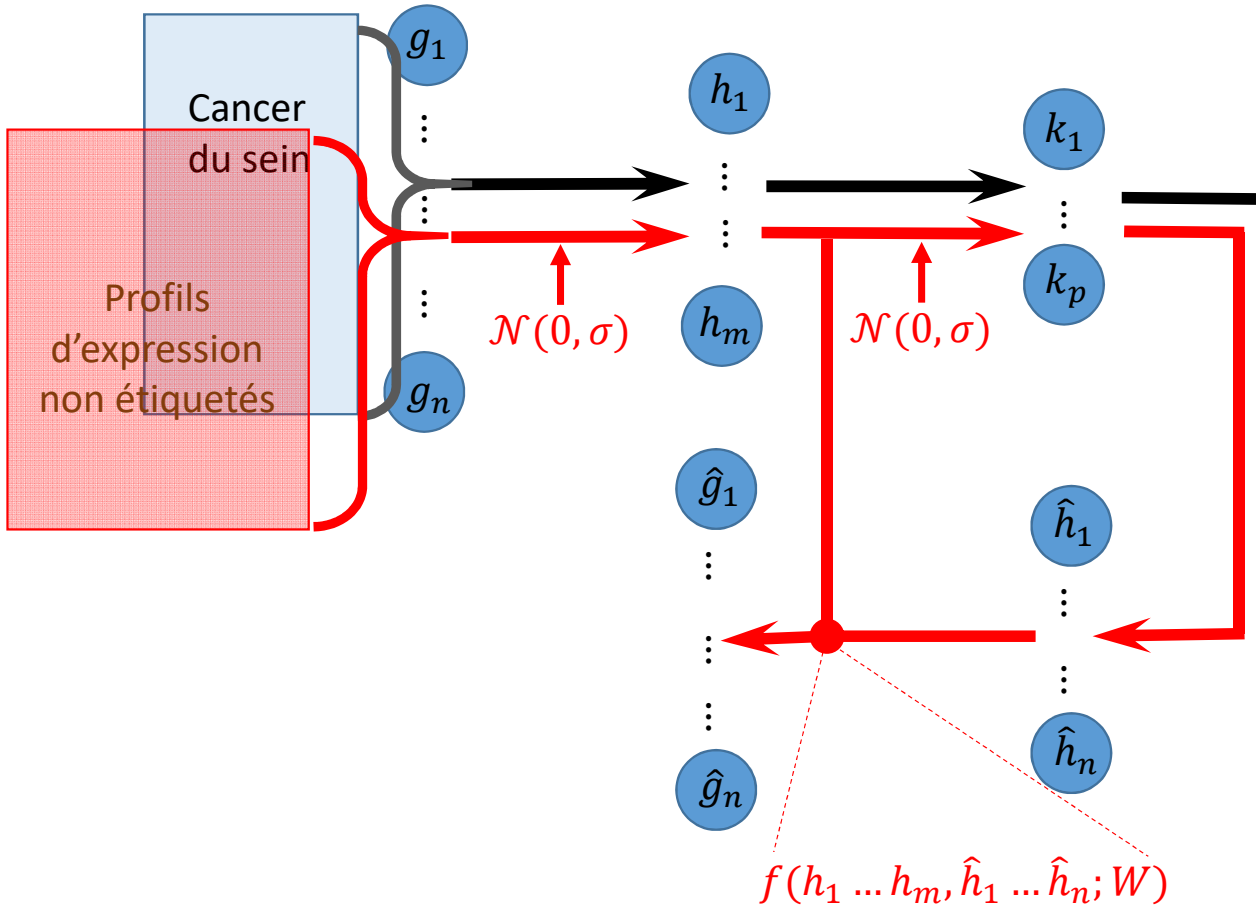
Ladder network



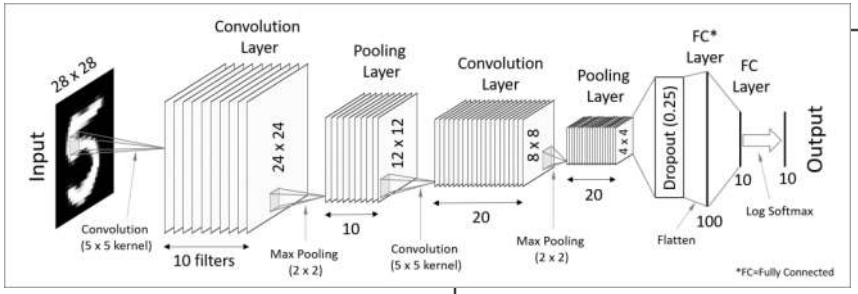
Ladder network



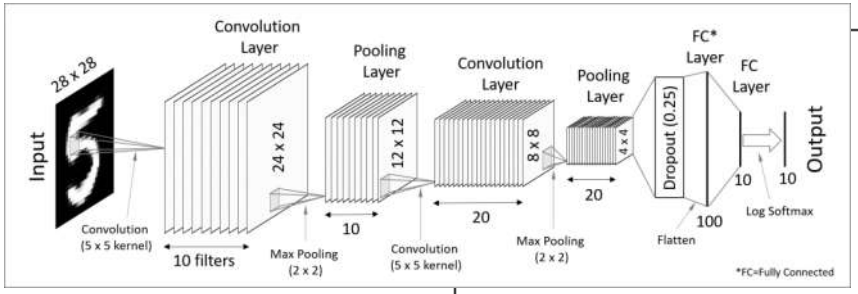
Ladder network



$$L(Y,\hat{Y}) = -\sum_{i=1}^{2} y_i . log(\hat{y}_i)$$
$$+\alpha_1 \sum_{i=1}^{m} (g_i - \hat{g}_i)^2$$
$$+\alpha_2 \sum_{i=1}^{p} (h_i - \hat{h}_i)^2$$


Input
28 x 28
Convolution Layer
24 x 24
10 filters
Convolution (5 x 5 kernel)
Pooling Layer
12 x 12
10
Max Pooling (2 x 2)
Convolution Layer
8 x 8
20
Convolution (5 x 5 kernel)
Pooling Layer
4 x 4
20
Max Pooling (2 x 2)
Dropout (0.25)
FC* Layer
100
Flatten
FC Layer
10
Log Softmax
10
Output
*FC=Fully Connected

MNIST t-SNE visualization
component 1
component 0
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

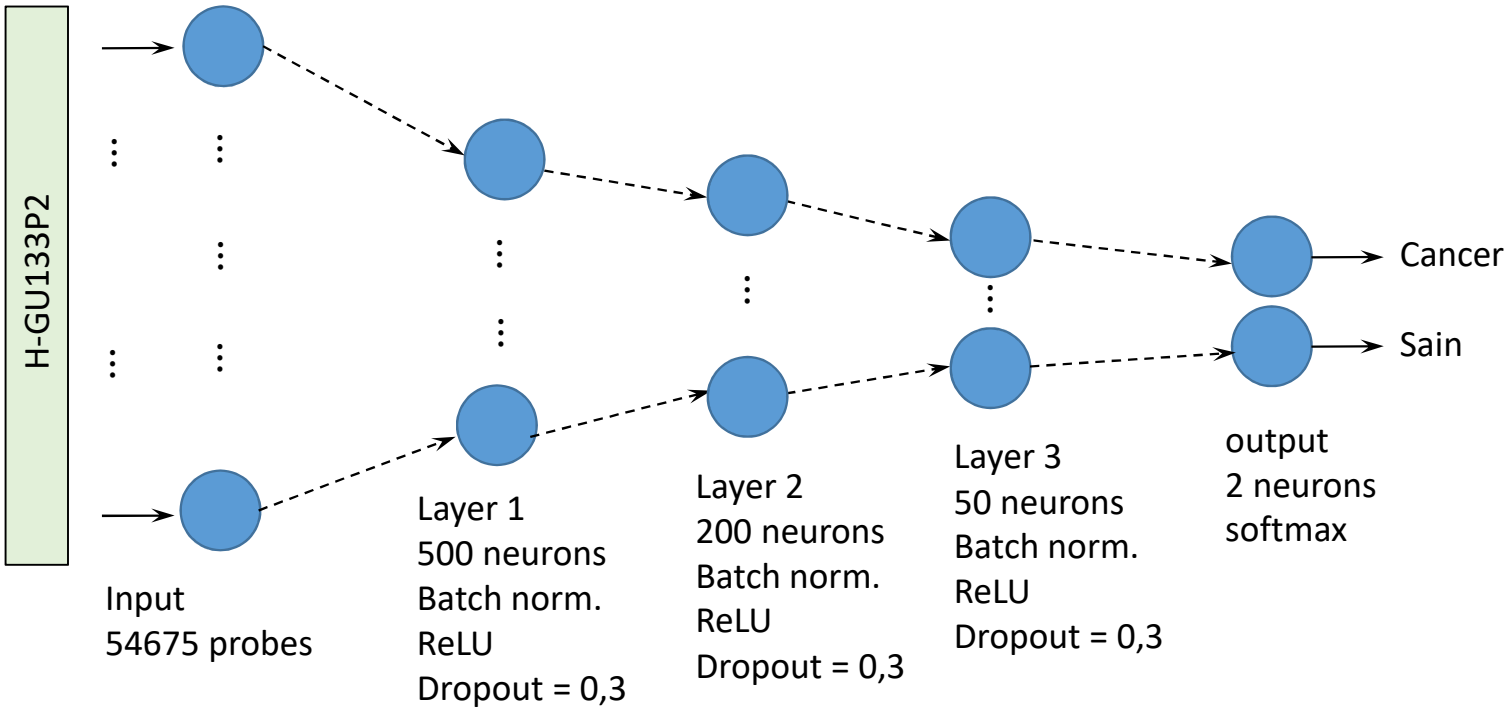

Input
28 x 28
Convolution Layer
24 x 24
Convolution (5 x 5 kernel)
10 filters
Pooling Layer
12 x 12
Max Pooling (2 x 2)
10
Convolution Layer
8 x 8
Convolution (5 x 5 kernel)
20
Pooling Layer
4 x 4
Max Pooling (2 x 2)
20
Dropout (0.25)
FC* Layer
Flatten
100
FC Layer
10
Log Softmax
10
Output
*FC=Fully Connected

MNIST t-SNE visualization
0
6
4
5
3
7
2
1
component 1
component 0
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

Interprétation biologique des réseaux de neurones

Analyse du réseau de neurones

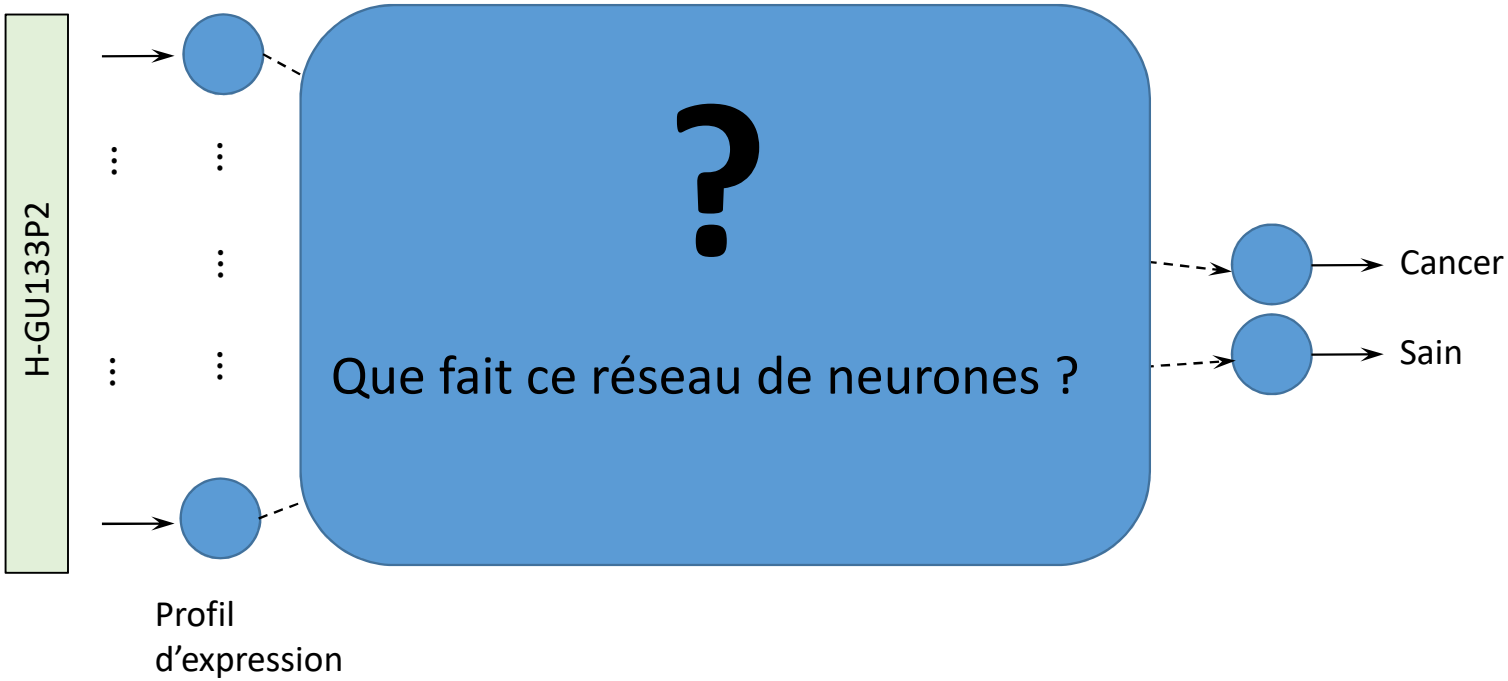
Prediction de la présence d'un cancer à 95%



Hyper-parameters :
Initialization Glorot
Adam (10^{-4})
Batch size = 16
Early Stopping
Régularisation L1 (10^{-4})

Analyse du réseau de neurones

Prediction de la présence d'un cancer à 95%



Enjeux de l'interprétation biologique

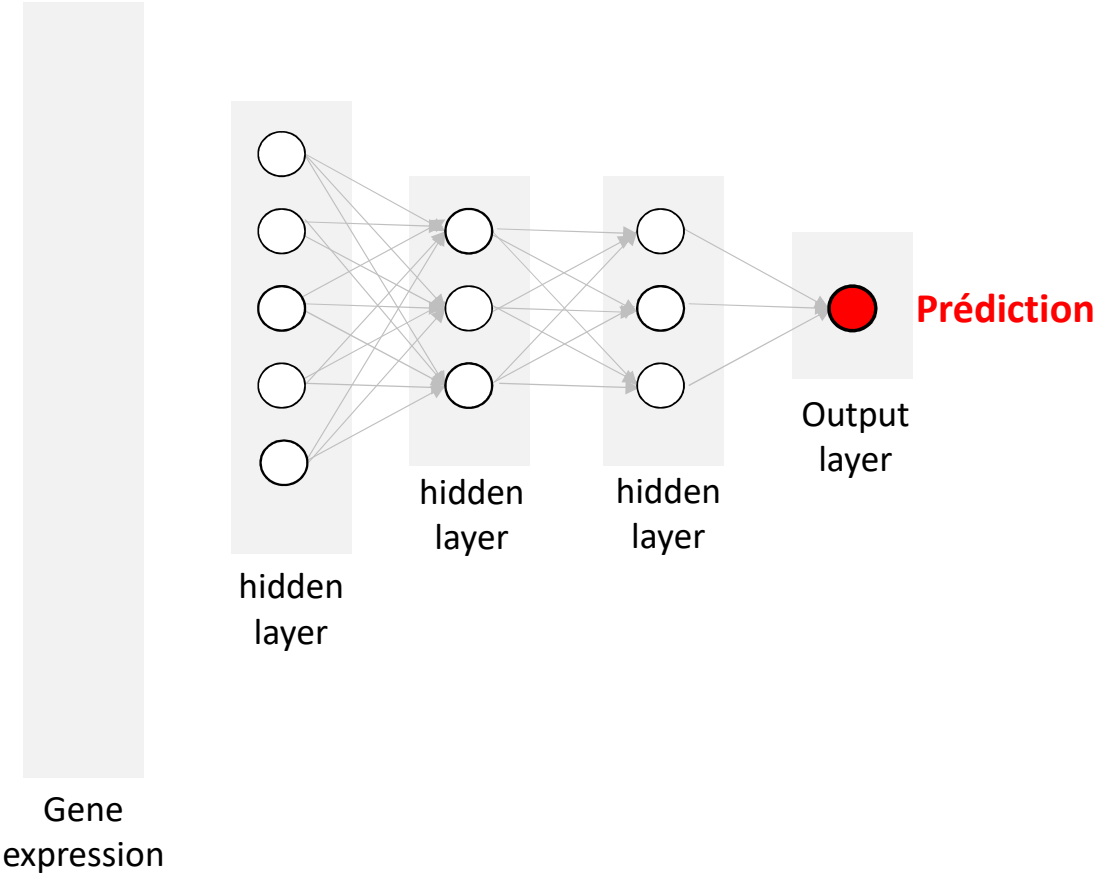
- Explication des prédictions
  - Pourquoi le réseau prédit un cancer à Mme X ?
  - Quels sont les gènes et fonctions impliqués dans cette prédiction ?

- Améliorer la confiance envers le modèle
  - Est-ce que les décisions se basent sur des éléments cohérents avec les connaissances des médecins ?

- Découverte de nouveaux biomarqueurs
  - Pourquoi tels gènes où tels fonctions sont utilisés par le réseau ?

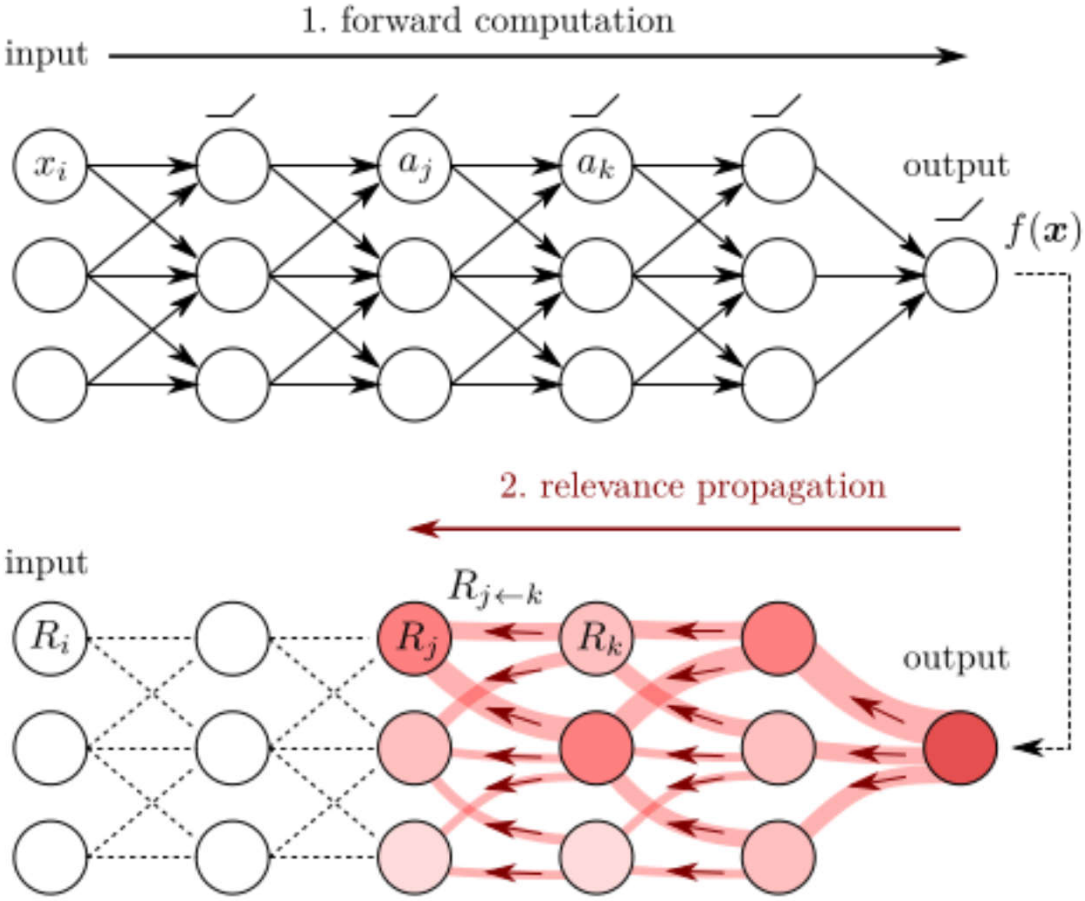
Méthode d’interprétation du réseau

Interprétation pour une prédiction donnée



LRP : Layer Relevance Propagation

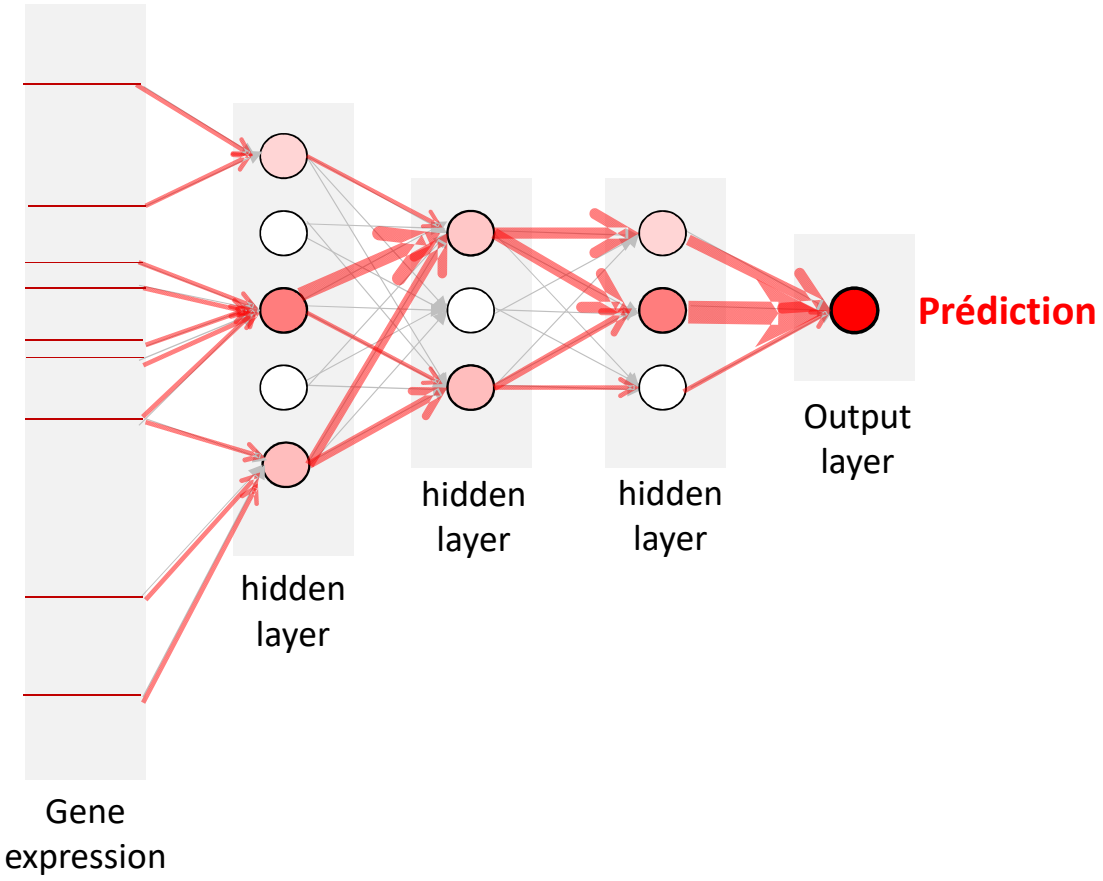
- Rétro-propagation de la prédiction dans le réseau
- Calcul d'un score de pertinence pour chaque neurone et connexion du réseau



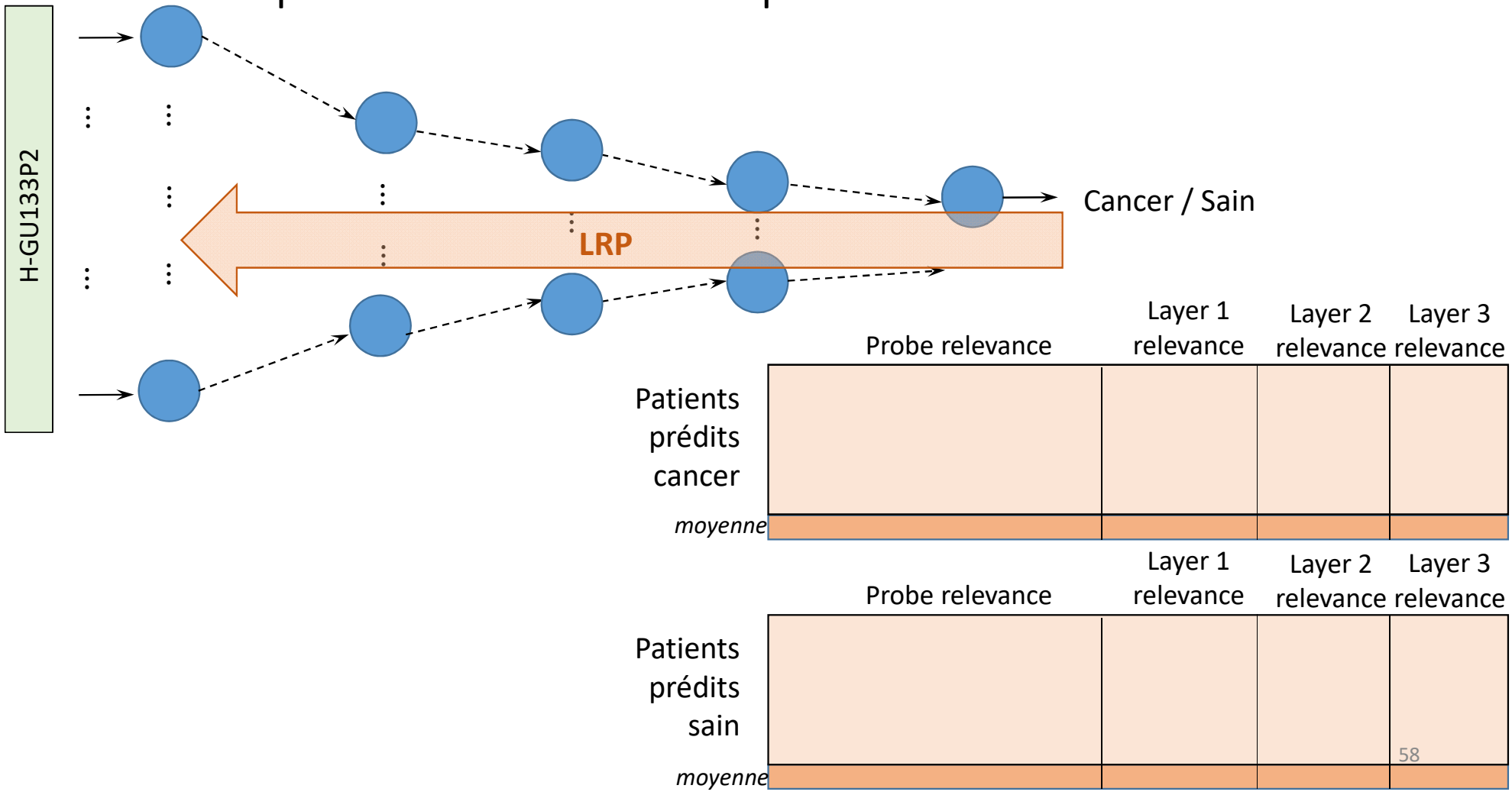
Méthode d'interprétation du réseau

Interprétation pour une prédiction donnée

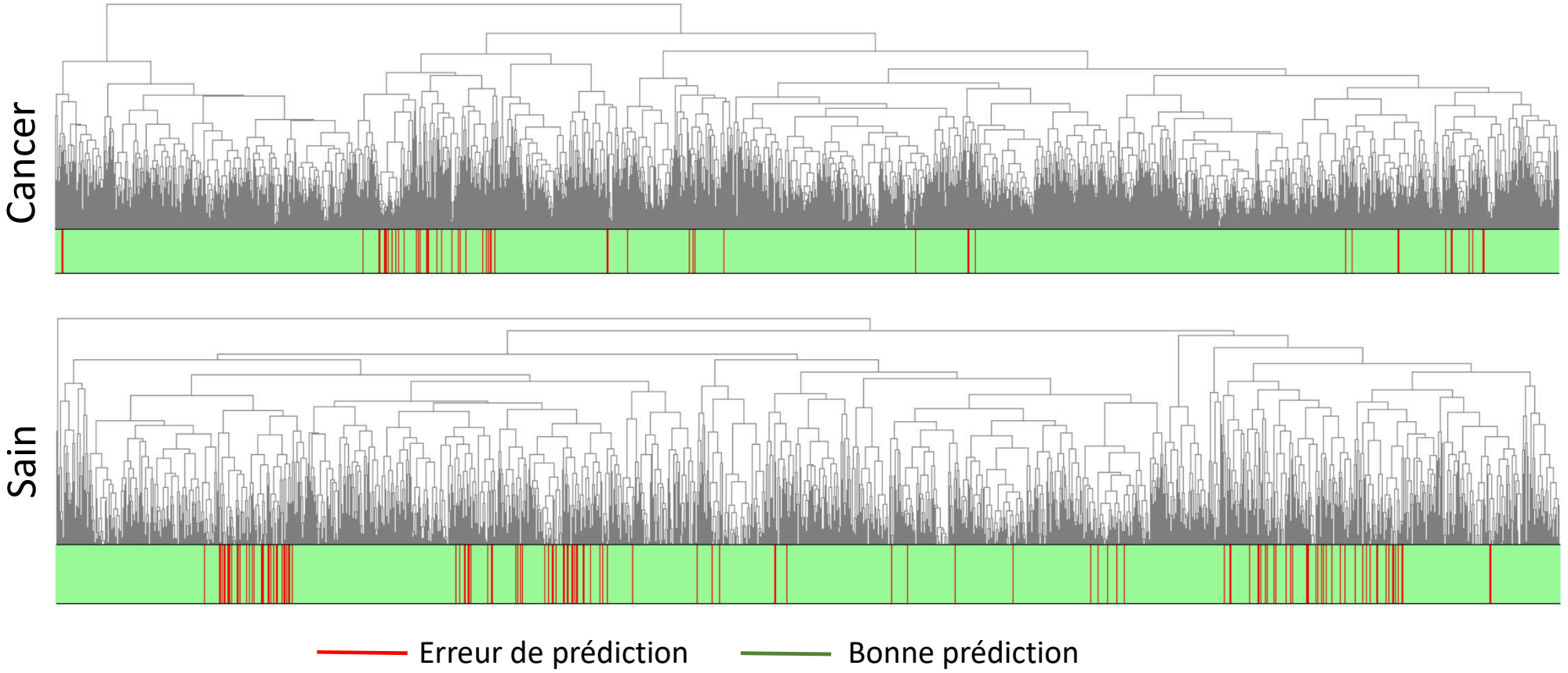
1. Evaluation de l'impact de chaque neurone dans la prédiction



Calcul de la pertinence de chaque neurone



Classification hiérarchique des patients basée sur le profil de pertinence



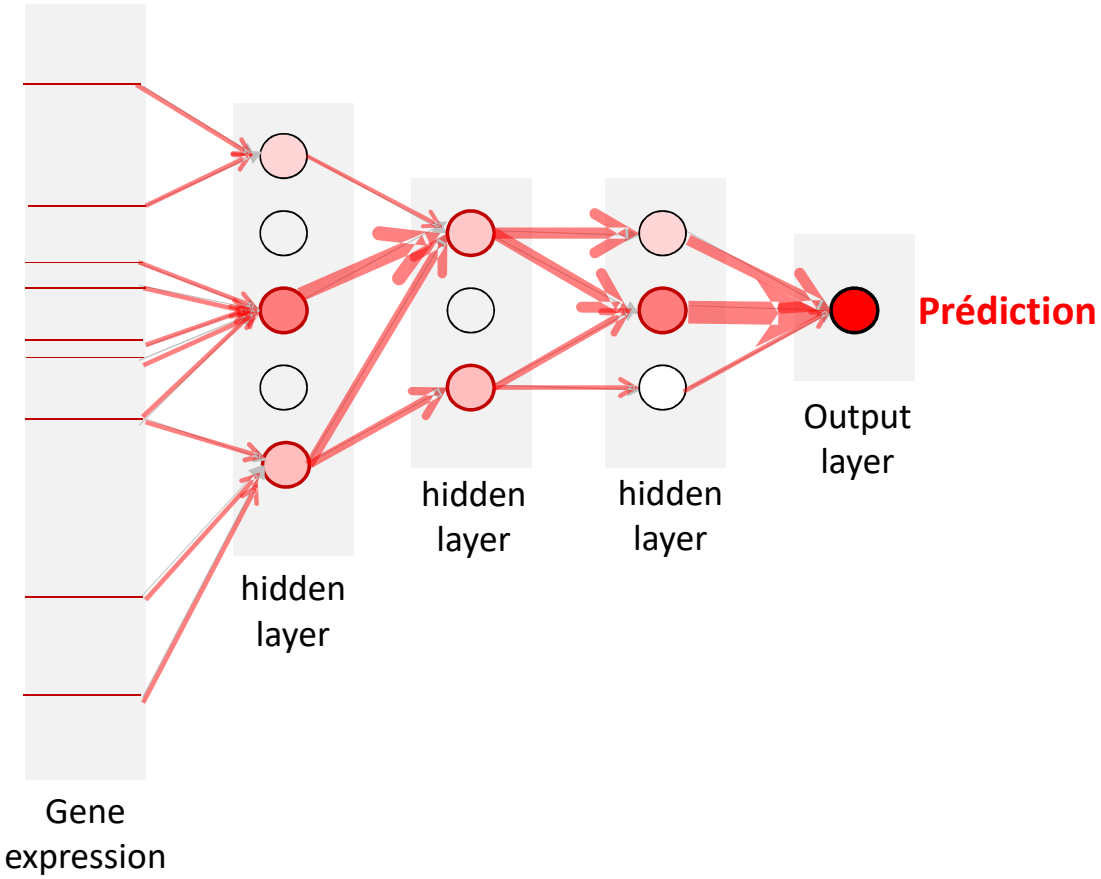
Les erreurs de prédiction sont regroupées dans des clusters

Méthode d'interprétation du réseau

Interprétation pour une prédiction donnée

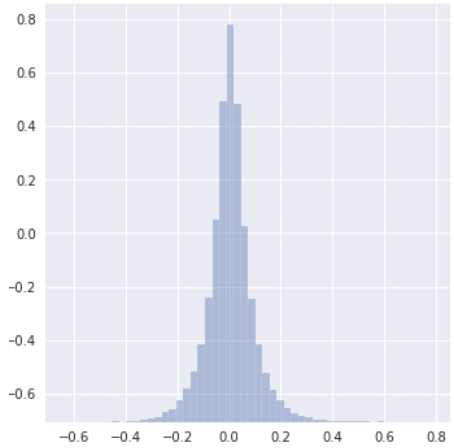
1. Evaluation de l'impact de chaque neurone dans la prédiction

2. Identification des neurones importants



Sélection of the important neurons

La distribution des pertinences moyennes des neurones semble Gausienne



Sélection basée sur un test-T (p-value<0,05)
avec correction de Bonferoni

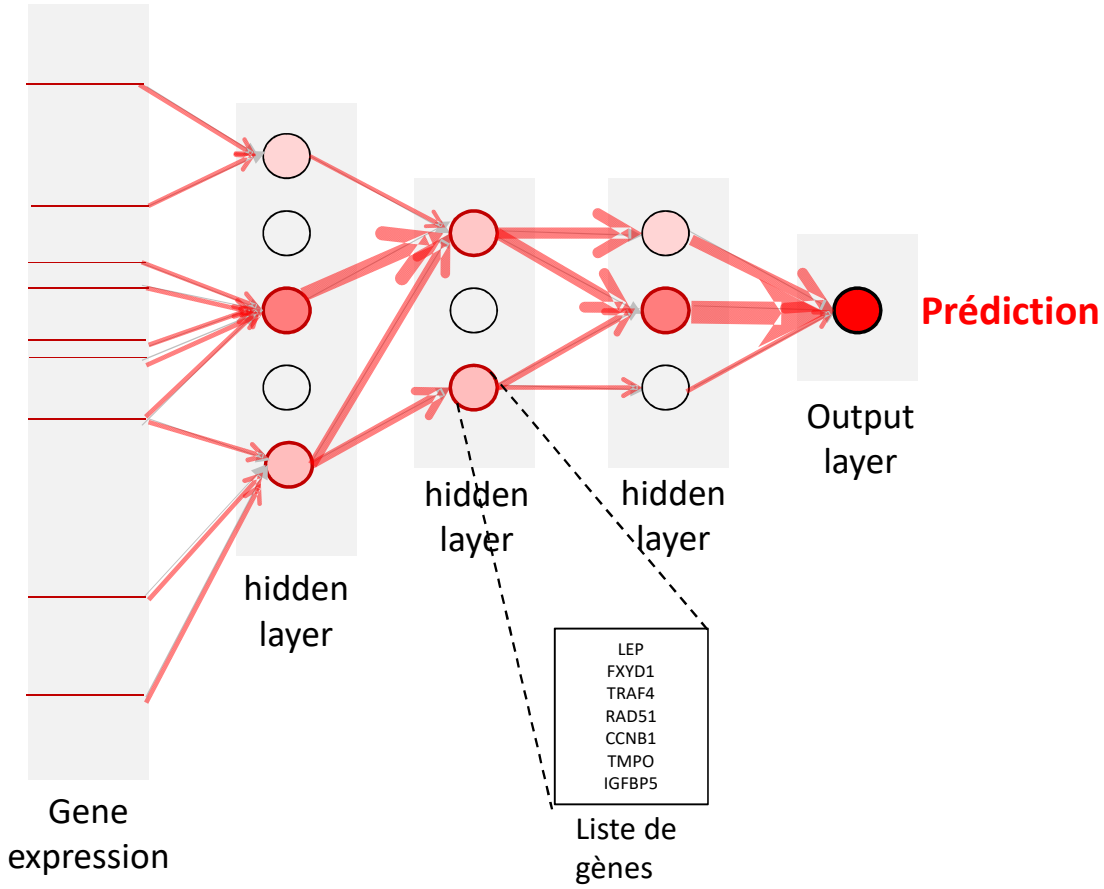
Gènes et neurones importants pour les prédictions:

	Probes (54675)	Layer 1 (500)	Layer 2 (200)	Layer 3 (50)
Cancer	3752	20	7	3
Sain	2988	26	14	4

Méthode d'interprétation du réseau

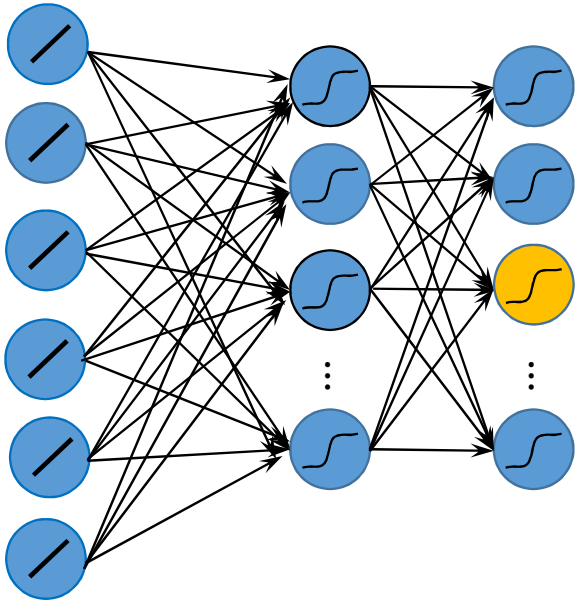
Interprétation pour une prédiction donnée

1. Evaluation de l'impact de chaque neurone dans la prédiction
2. Identification des neurones importants
3. Association neurone – liste de gènes



Analyse des neurones importants

Interprétation des neurones pertinents

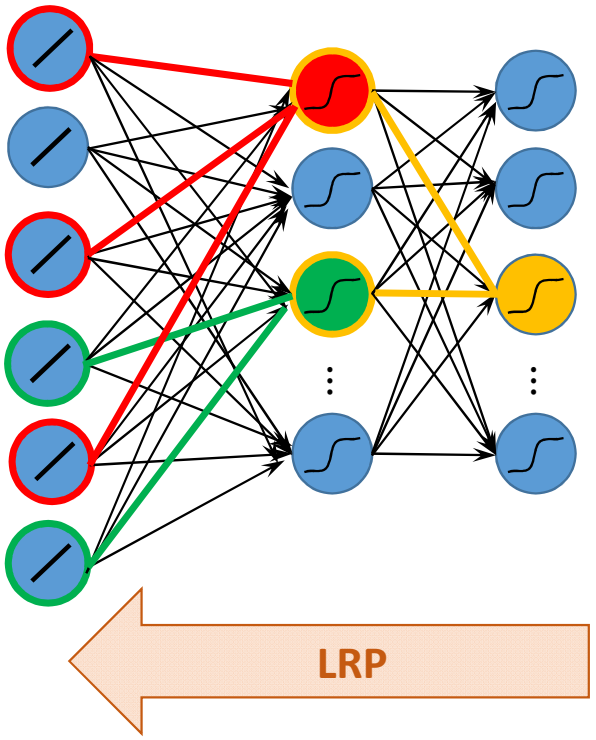


Analyse des neurones importants

Interprétation des neurones pertinents

- LRP sur la sortie du neurone
  Pertinences des gènes par rapport au neurone

- Sélection des gènes les plus pertinents



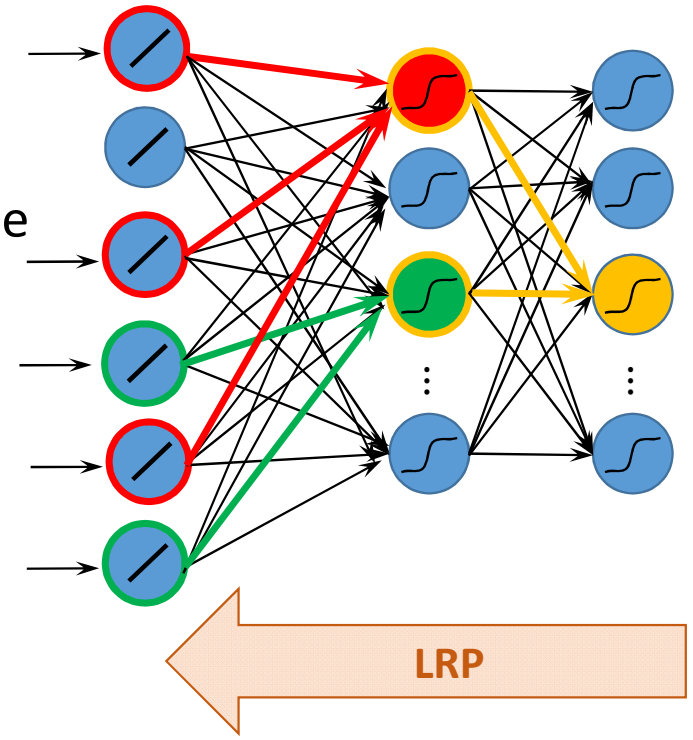
Analyse des neurones importants

Interprétation des neurones pertinents

- LRP sur la sortie du neurone
  Pertinences des gènes par rapport au neurone

- Sélection des gènes les plus pertinents

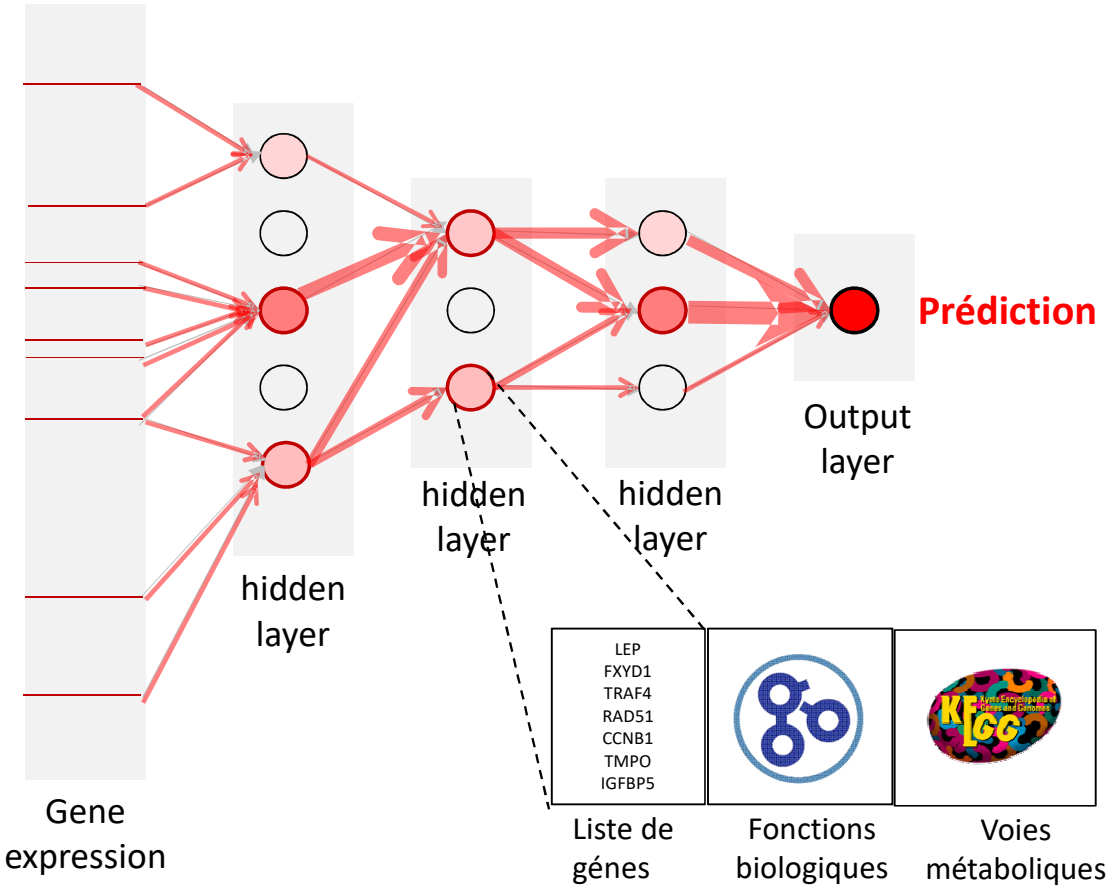
➢ Association : neurone – liste de gènes



Méthode d'interprétation du réseau

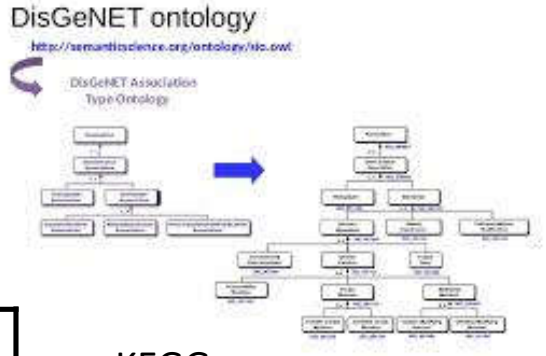
Interprétation pour une prédiction donnée

1. Evaluation de l'impact de chaque neurone dans la prédiction
2. Identification des neurones importants
3. Association neurone – liste de gènes
4. Association avec des connaissances biologiques (GO, KEGG)

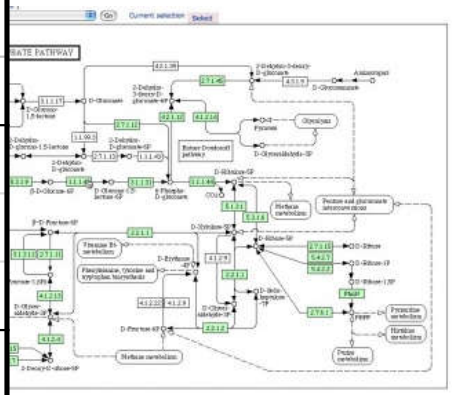


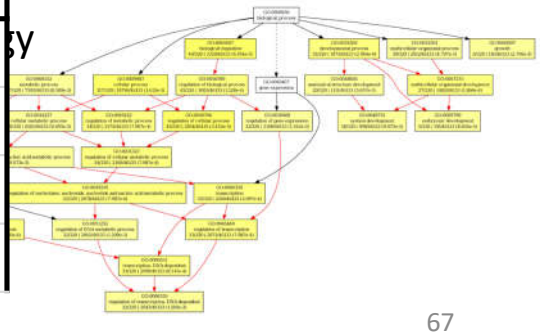
Analyse du réseau de neurones

Cancer	couche3	neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
	couche2	neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
	couche1	neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
Sain		neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>
		neurone	<liste de gènes>	<GO>	<KEGG>

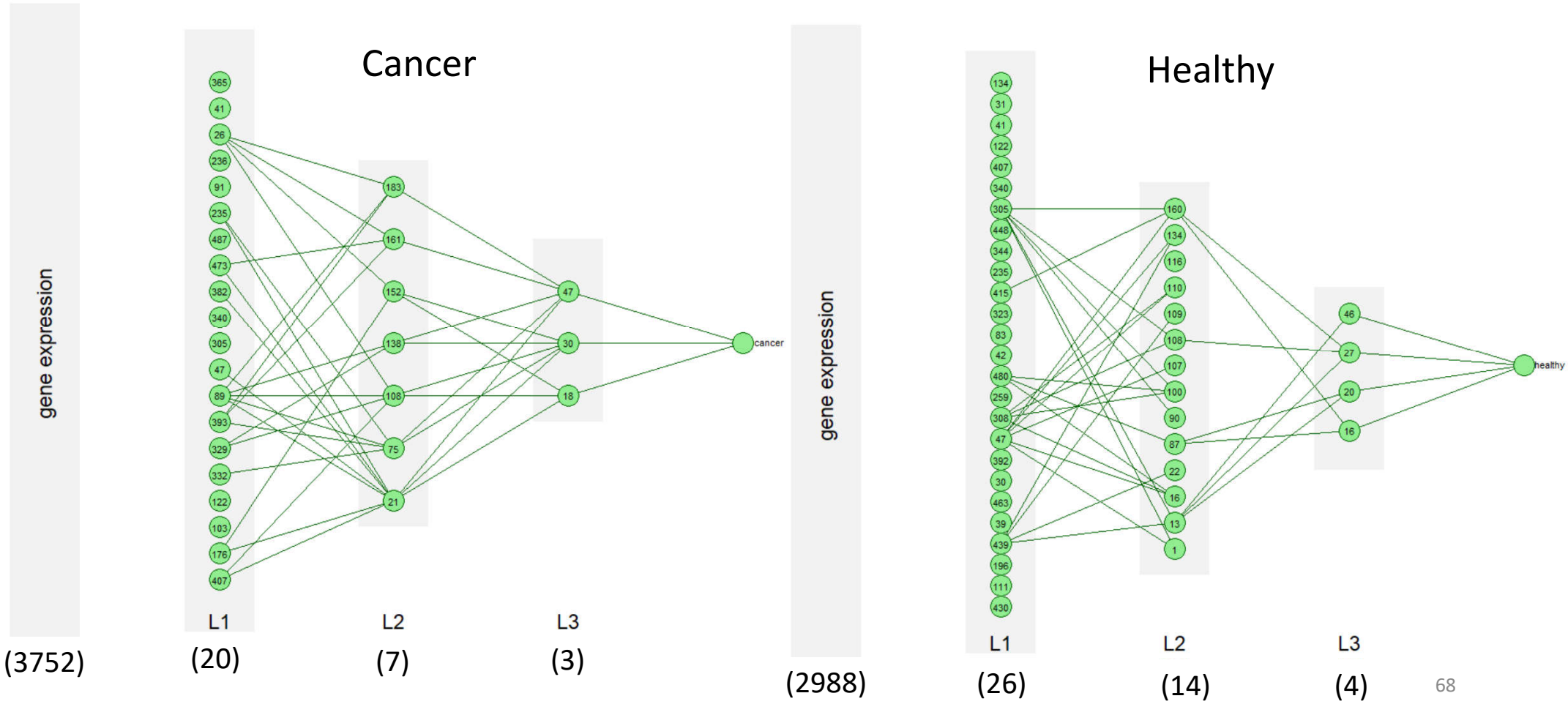


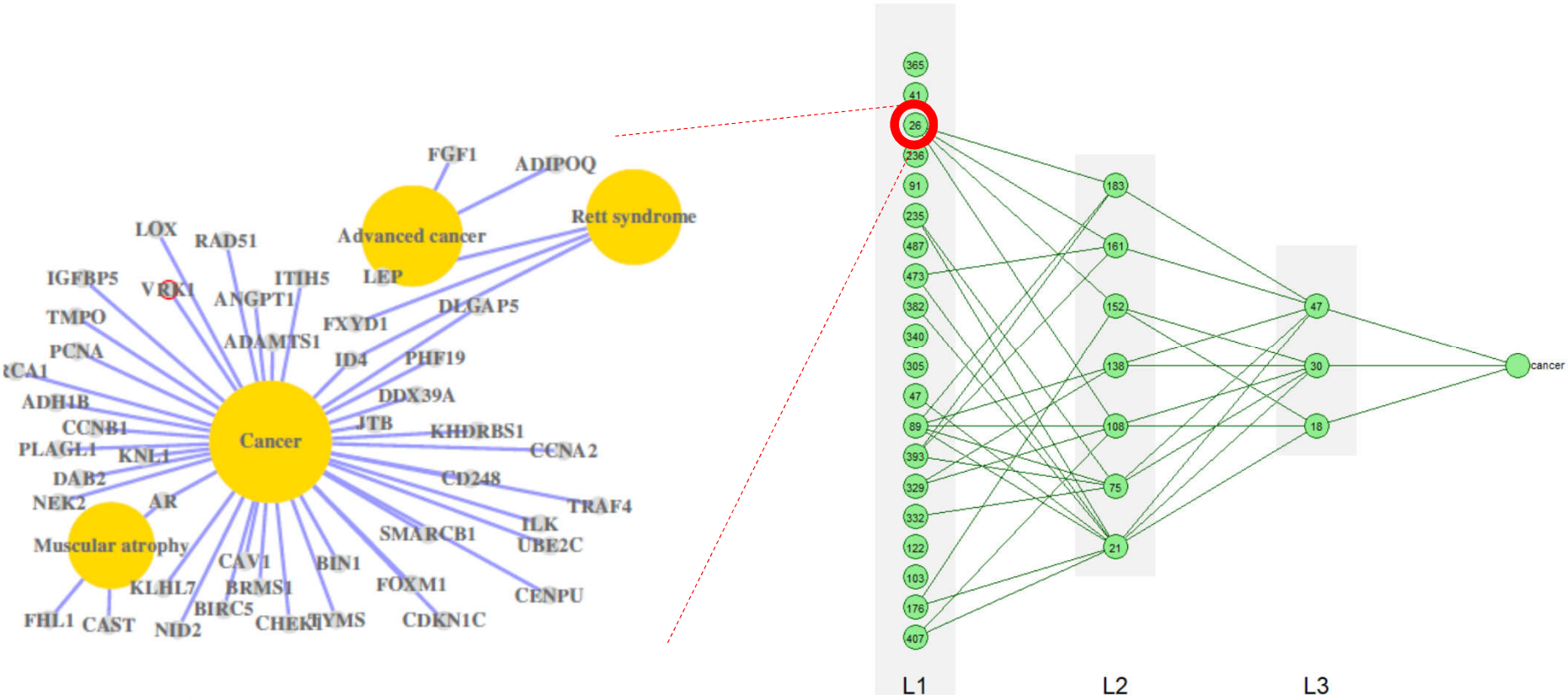
KEGG





Neurones et connexions importants





	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0031648	0.0000379419	15.602205	0.38857143	5	32	protein destabilization
2	GO:0036120	0.0002690861	31.172932	0.13357143	3	11	cellular response to platelet-derived growth factor stimulus
3	GO:0030097	0.0003149311	2.762361	7.24928571	18	597	hemopoiesis
4	GO:0036119	0.0003555988	27.706767	0.14571429	3	12	response to platelet-derived growth factor
5	GO:0002520	0.0004328961	2.610320	8.09928571	19	667	immune system development
6	GO:0060648	0.0004356304	165.119403	0.03642857	2	3	mammary gland bud morphogenesis
7	GO:0034103	0.0004408562	8.759542	0.64357143	5	53	regulation of tissue remodeling
8	GO:0048534	0.0006209209	2.596202	7.67428571	18	632	hematopoietic or lymphoid organ development
9	GO:0046850	0.0007271245	11.145455	0.41285714	4	34	regulation of bone remodeling
10	GO:0060603	0.0007271245	11.145455	0.41285714	4	34	mammary gland duct morphogenesis
11	GO:0061180	0.0009801412	7.242695	0.76500000	5	63	mammary gland epithelium development

Thank you for your attention

People involved in this project :

- Farida Zehraoui
- Mathieu Arles
- Tina issa
- Mohamed Ben Hamdoune
- Victoria Bourgeais