Bienvenue!

ÉCOLE D'HIVER FRANCOPHONE EN APPRENTISSAGE PROFOND

5 - 9 mars 2018

IVADO
HEC Montréal
Polytechnique Montréal
Université de Montréal

MILA

Institut
québécois
d’intelligence
artificielle

MILA

Université de Montréal

McGill

Institut
des algorithmes
d'apprentissage
de Montréal
MILA

Introduction
Apprentissage profond

Gaétan Marceau Caron
gaetan.marceau.caron@rd.mila.quebec

Apprentissage automatique

- Tâche

- Base d'apprentissage

- Modèle

- Algorithme d'apprentissage

- Évaluation du modèle

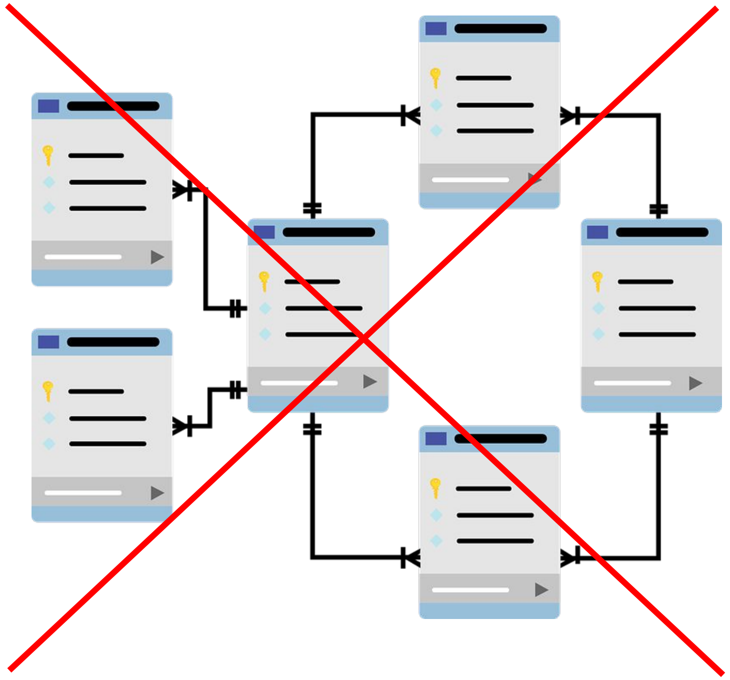
Définition de la tâche

Types d'apprentissage:

- **Supervisé (X, y)**
- **Non-supervisé (X)**
- Par renforcement (Actions, récompenses)

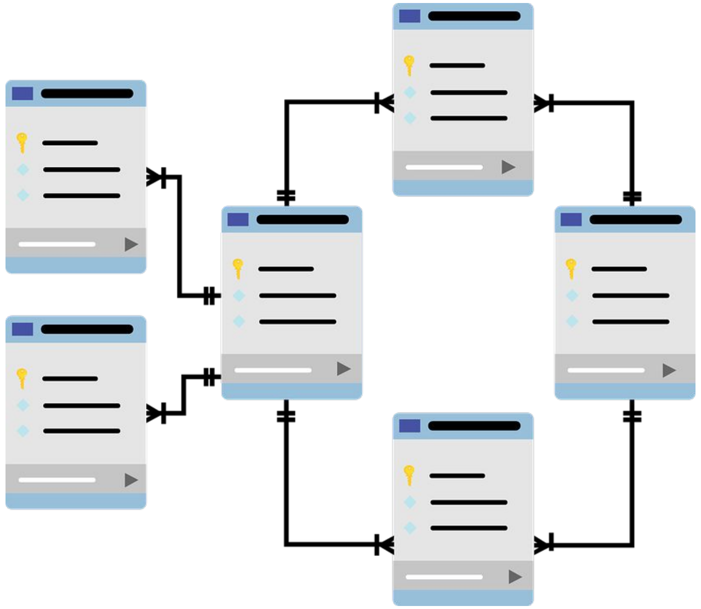
Base d'apprentissage

Une base de données **n'est pas** une base d'apprentissage!

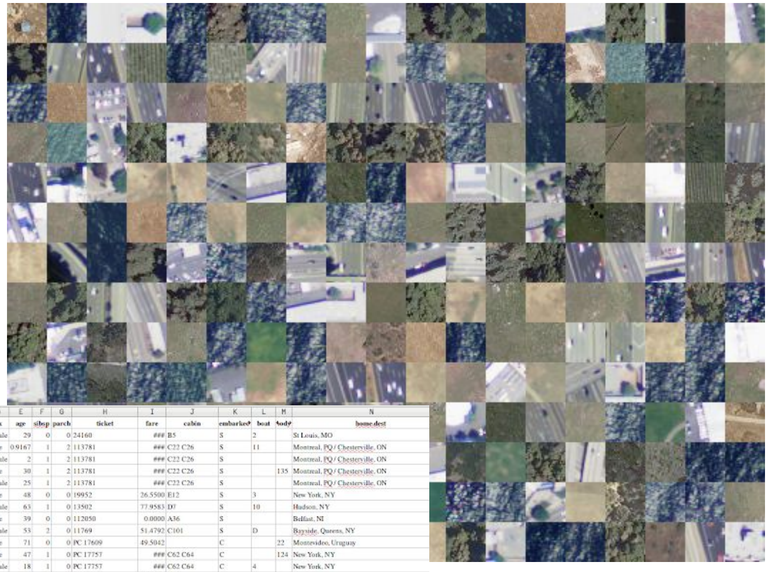


Base d'apprentissage

(X, y)



Extraction de données



Base d’apprentissage

$\mathbf{x} \in \mathbb{R}^4$ **Fisher's Iris Data** $\mathbf{Y}$

$n = 150$

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*

etc...

5.7	2.8	4.5	1.3	*versicolor*
6.3	3.3	4.7	1.6	*versicolor*
4.9	2.4	3.3	1.0	*versicolor*
6.6	2.9	4.6	1.3	*versicolor*
5.2	2.7	3.9	1.4	*versicolor*
5.0	2.0	3.5	1.0	*versicolor*

etc...

7.7	3.0	6.1	2.3	*virginica*
6.3	3.4	5.6	2.4	*virginica*
6.4	3.1	5.5	1.8	*virginica*
6.0	3.0	4.8	1.8	*virginica*
6.9	3.1	5.4	2.1	*virginica*

etc...

Iris setosa

Iris versicolor

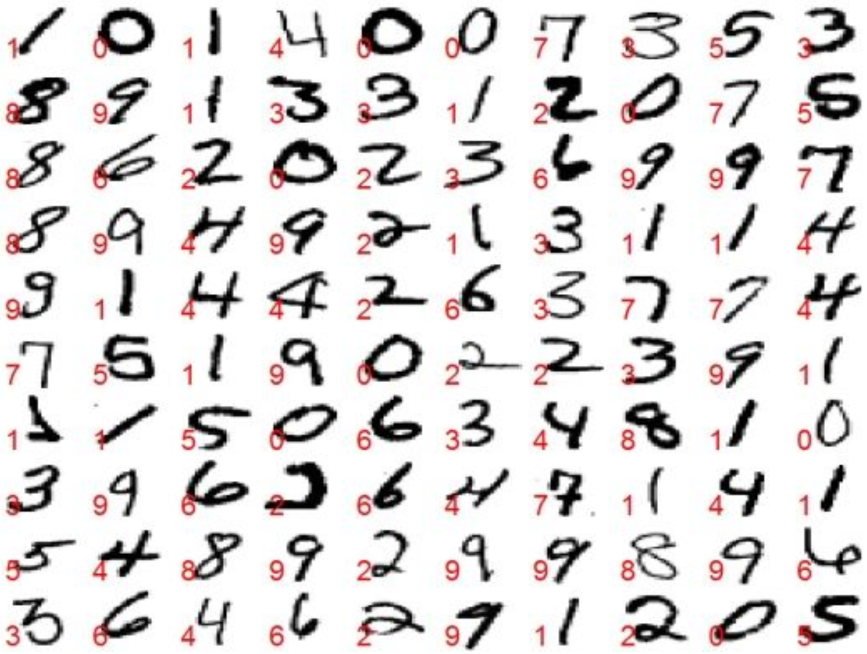
Iris virginica

Ronald Fisher
The use of multiple measurements in taxonomic problems (1936)

MILA

Statistique

Indépendant et identiquement distribué (iid)

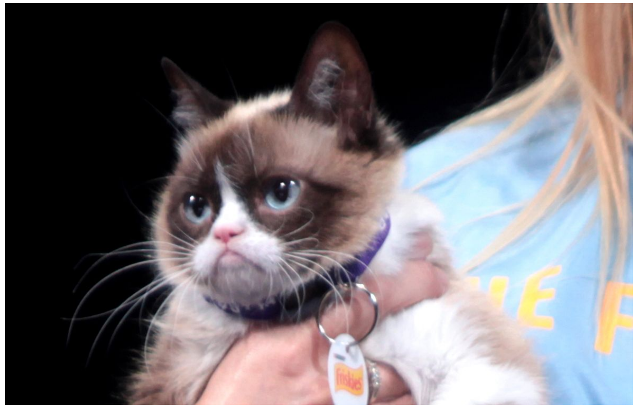


Base d’apprentissage

Ensemble
d’entraînement

Déploiement

Nouvel
exemple



Non-stationnarité

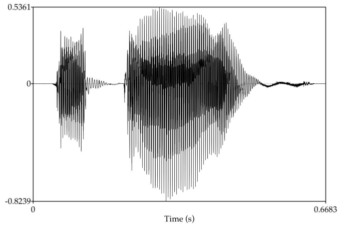
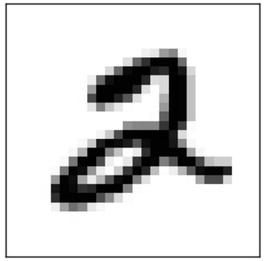
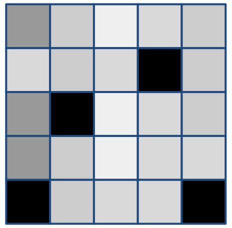
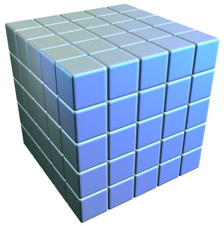
Data augmentation

Symétries

1 exemple

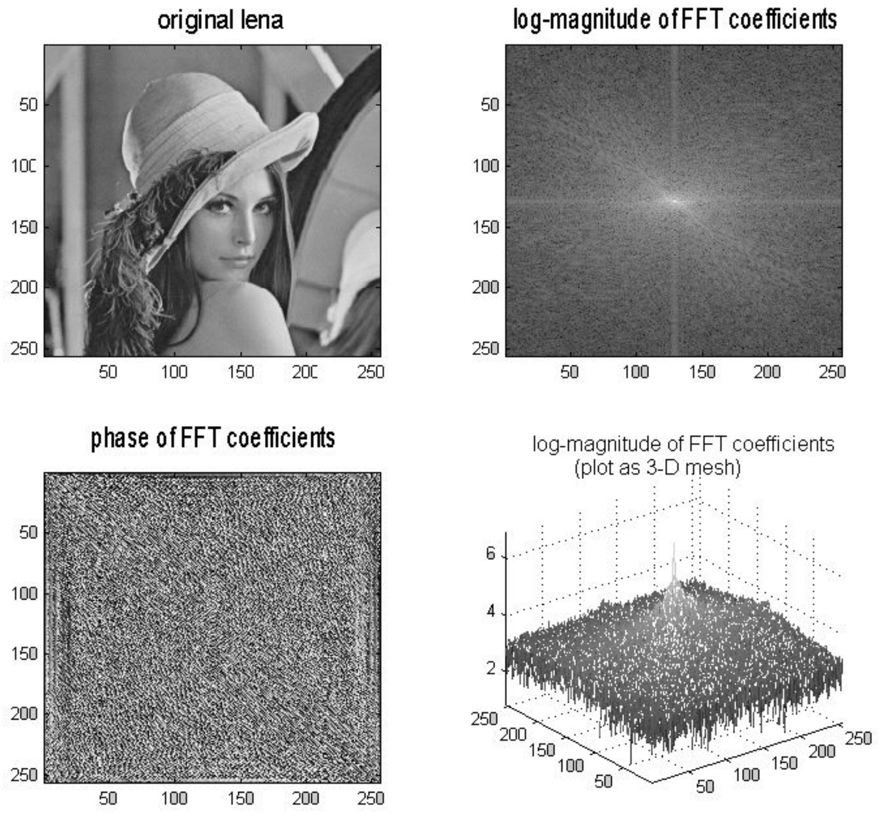
4 exemples

Représentation de l’entrée

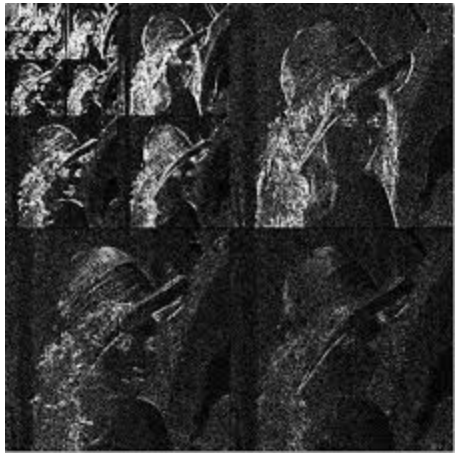
	Représentation naturelle	Représentation à l’entrée du modèle
Son		 Séquence de valeurs réelles
Image		 1D  2D  3D
Catégoriel	\x48\x65\x6c\x6c\x6f (UTF8)	 Vecteur One-hot
Document	\x48\x65\x6c\x6c\x6f… (UTF8)	 Séquence de vecteurs One-hot; 0 0 1 6 0 2 0 2 2 0 0 Bag-of-words Ou tf-idf

MILA

La meilleure représentation?



Wavelet decomposition



...

https://www.ee.columbia.edu/~xlx/courses/dip/imgfun/fft2.html

Preprocessing

- Normaliser les données X et les données y
  - Standardisation $\tilde{x} = \frac{x-\mu}{\sigma}$
  - Normalisation $\tilde{x} = \frac{x-x_{min}}{x_{max}-x_{min}}$

MILA

Base d'apprentissage

Décomposition en trois ensembles:

- ensemble **d'entraînement**
- ensemble de **validation**
- ensemble de **test**

MILA

Apprentissage supervisé

Ingrédients:

- Base d'apprentissage
- **Modèle**
- Algorithme d'apprentissage

Modèle

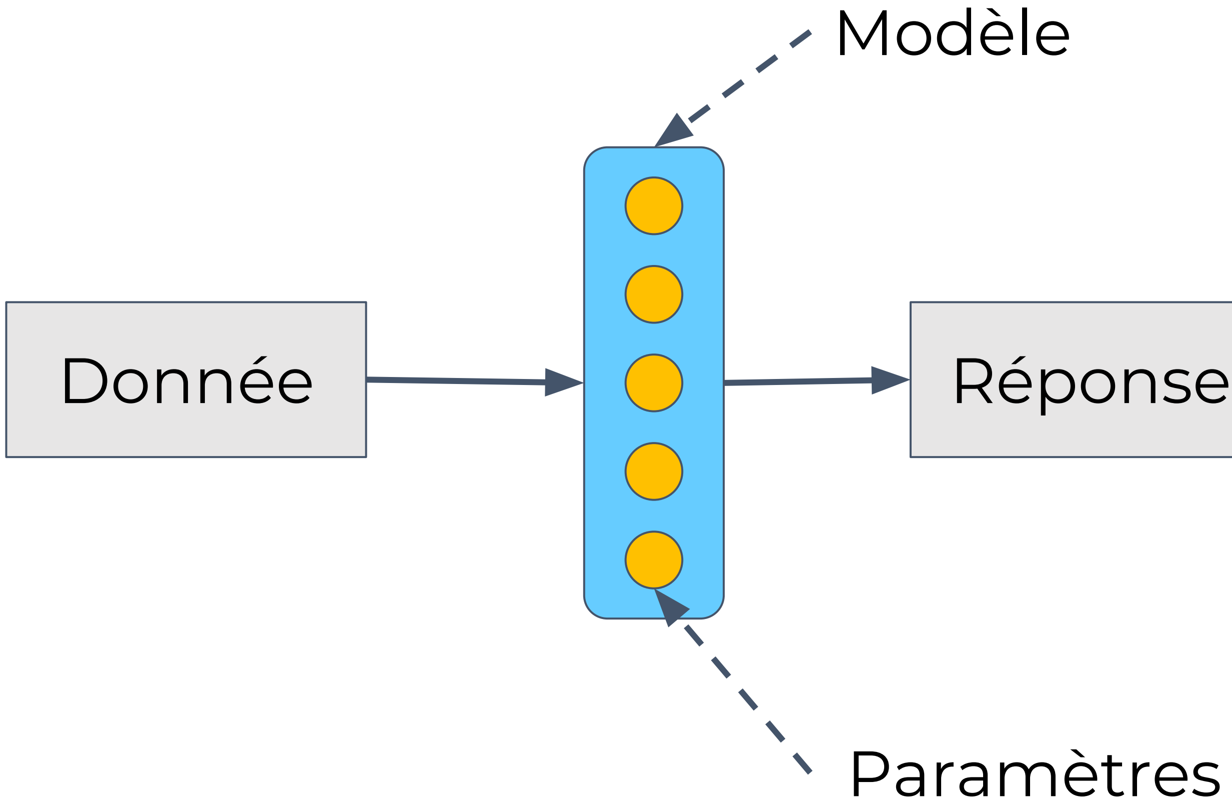
Interprétation **informatique**

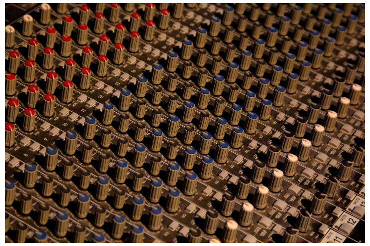
- Programme: f(X) -> y

Interprétation **probabiliste**

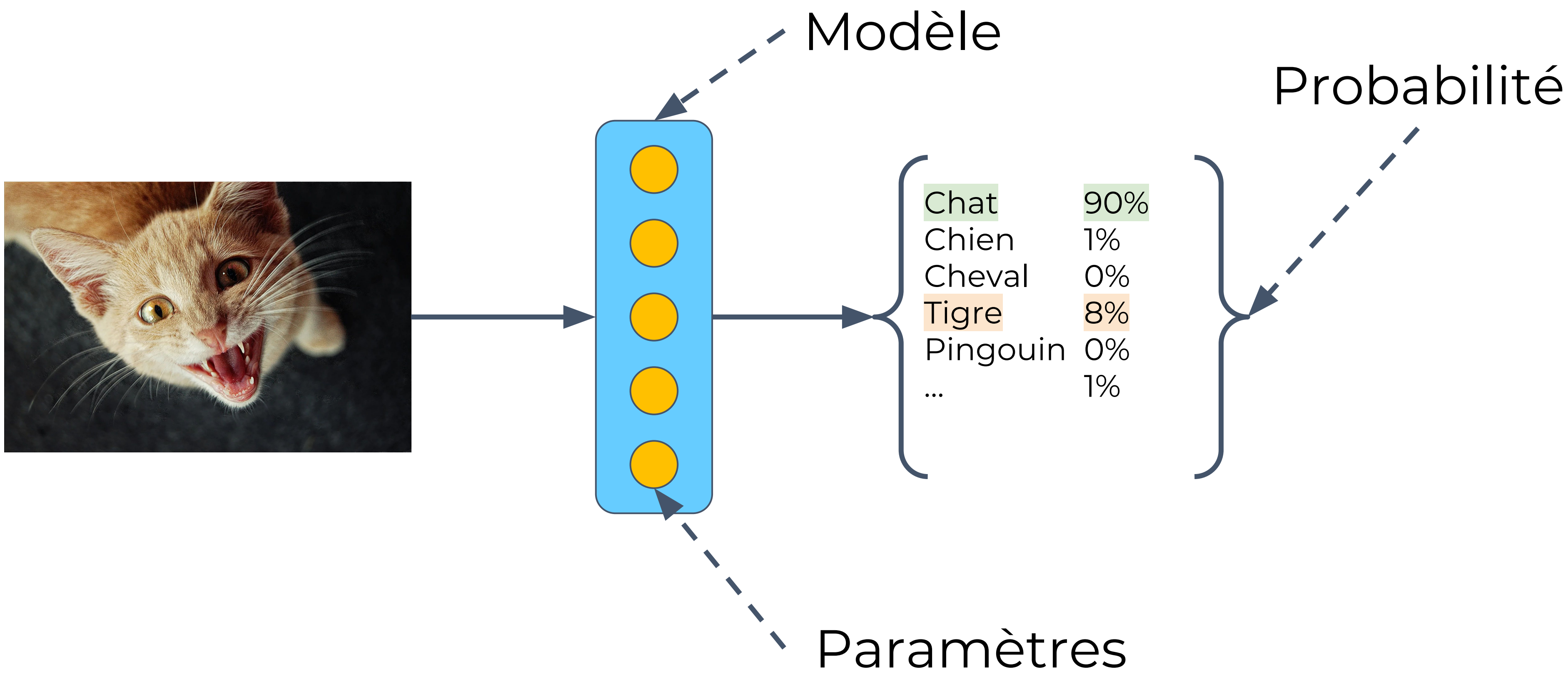
- Distribution de probabilité: p(y|X)

Définition du modèle

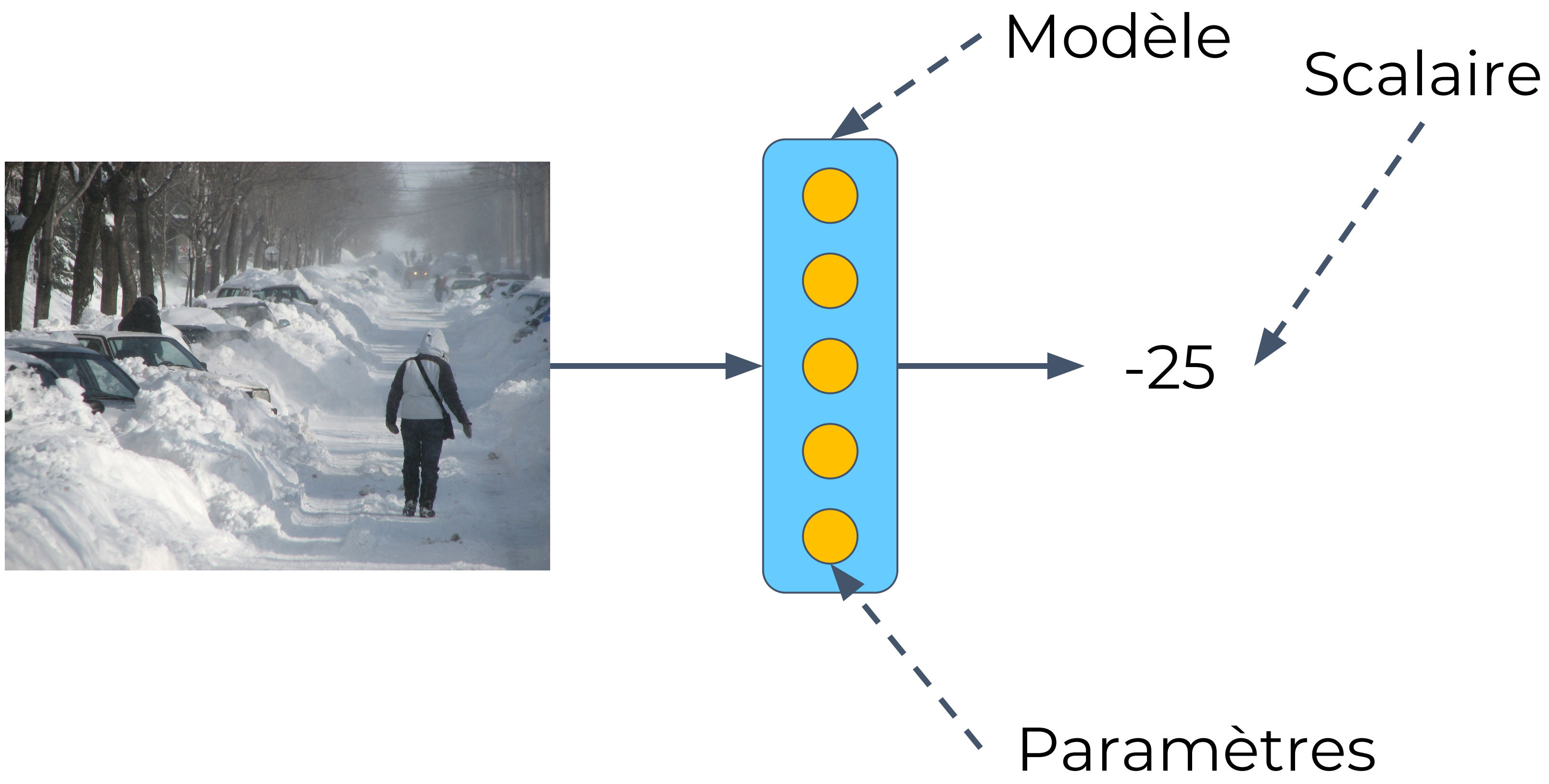




Exemple: déterminer l’animal

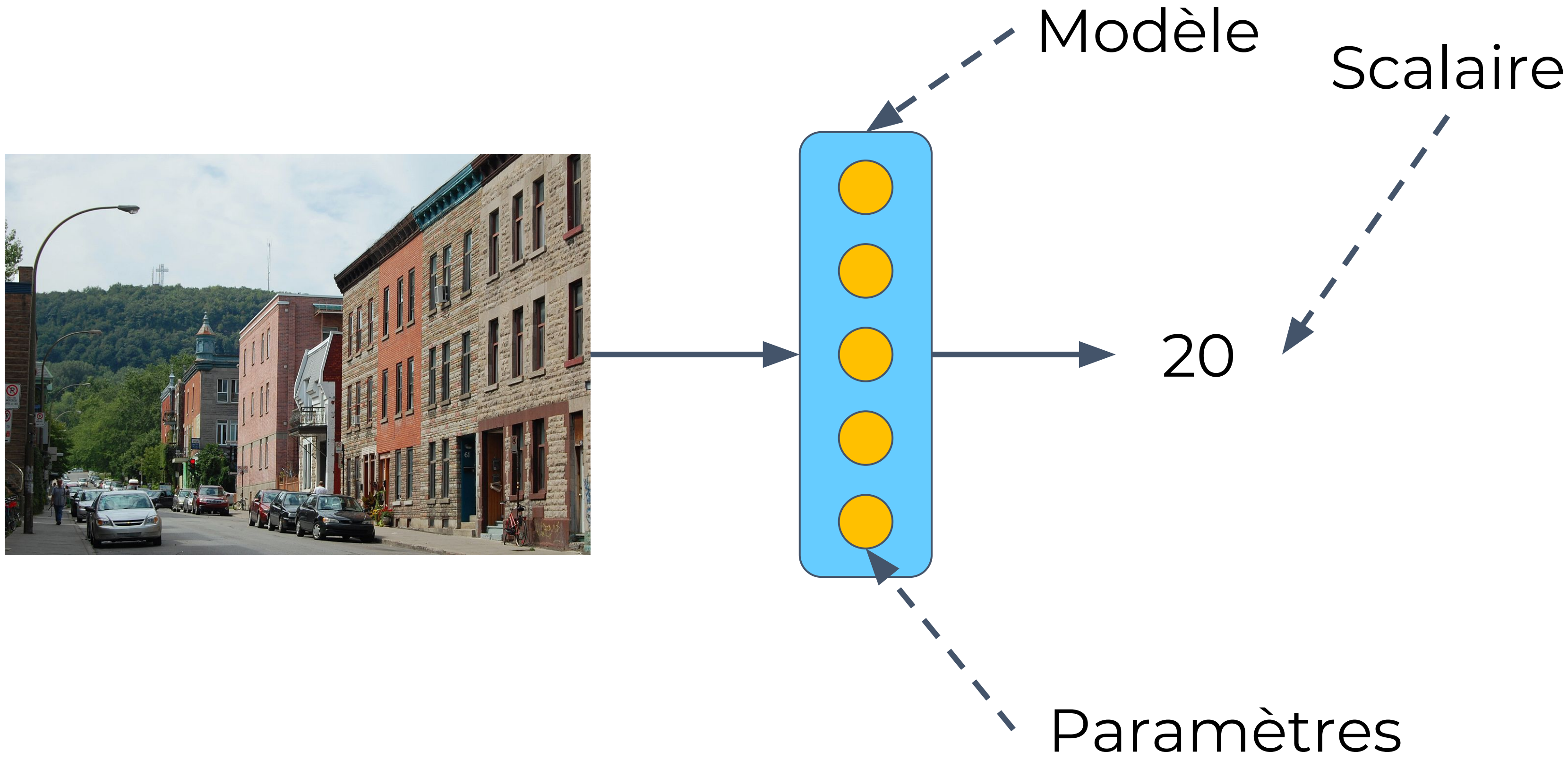


Exemple: déterminer la température



MILA

Exemple: déterminer la température

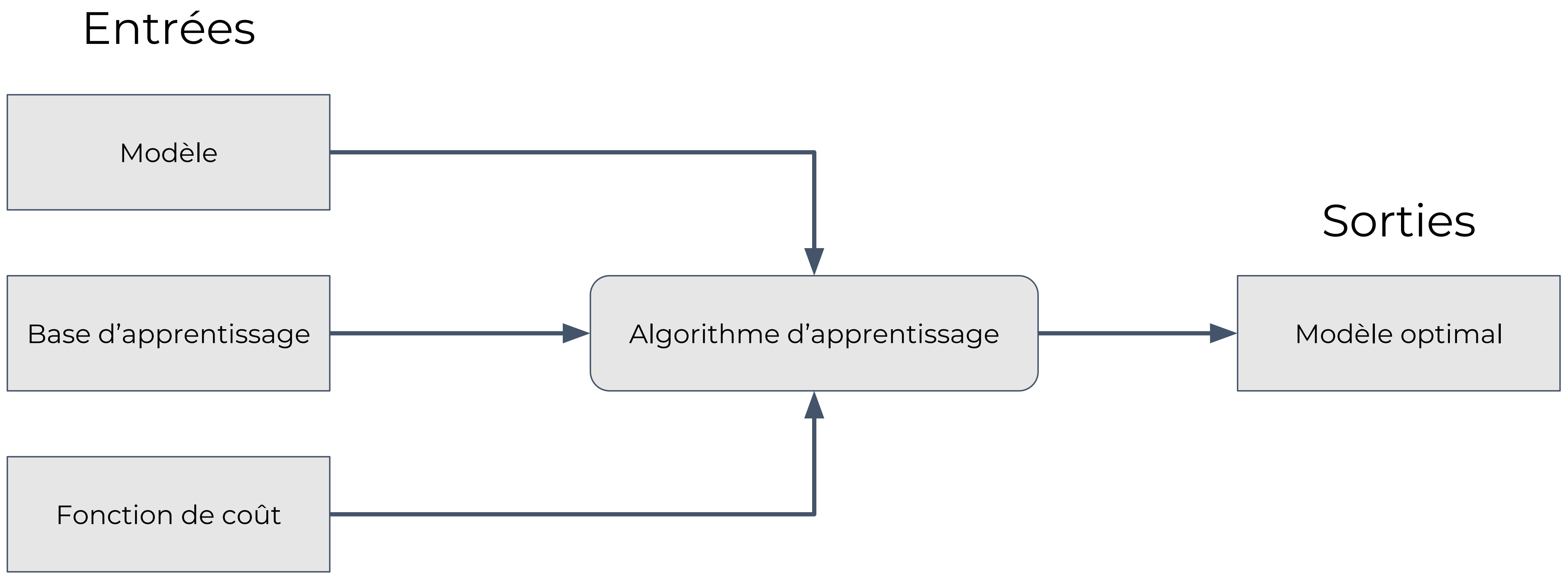


Apprentissage supervisé

Ingrédients:

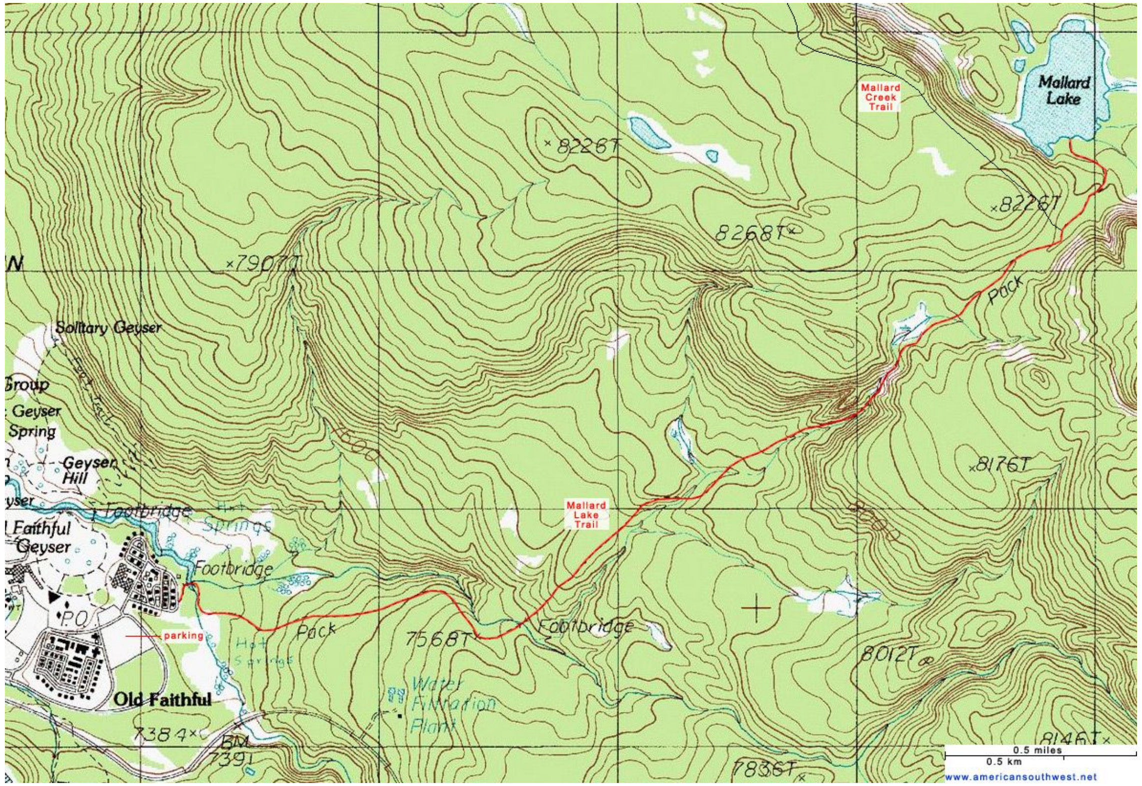
- Base d'apprentissage
- Modèle
- **Algorithme d'apprentissage**

Algorithme d’apprentissage



Une balade en million de dimensions

Optimisation	Alpinisme
Paramètre	Coordonnée
Coût	Altitude
Gradient	Pente
Mise à jour	Déplacement



Fonctions de coût

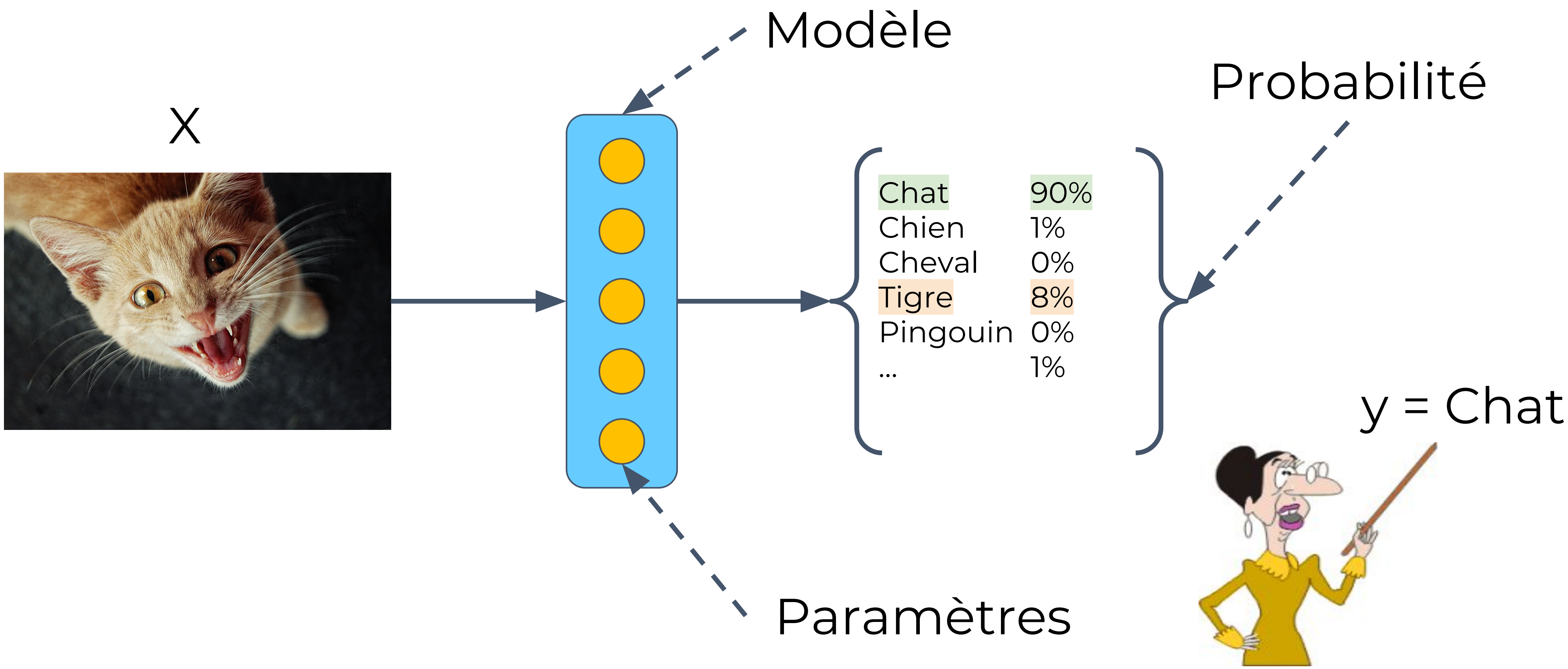
- Maximum de vraisemblance

- Méthode des moindres carrés

- Théorie de la décision

- ...

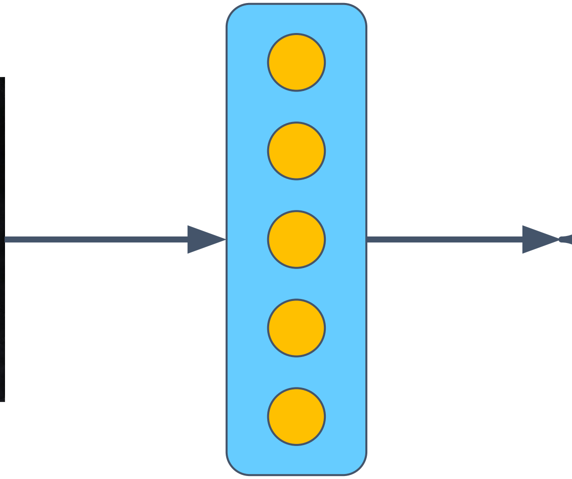
Exemple: maximiser la probabilité



Exemple: maximiser la probabilité

Negative log-likelihood

X



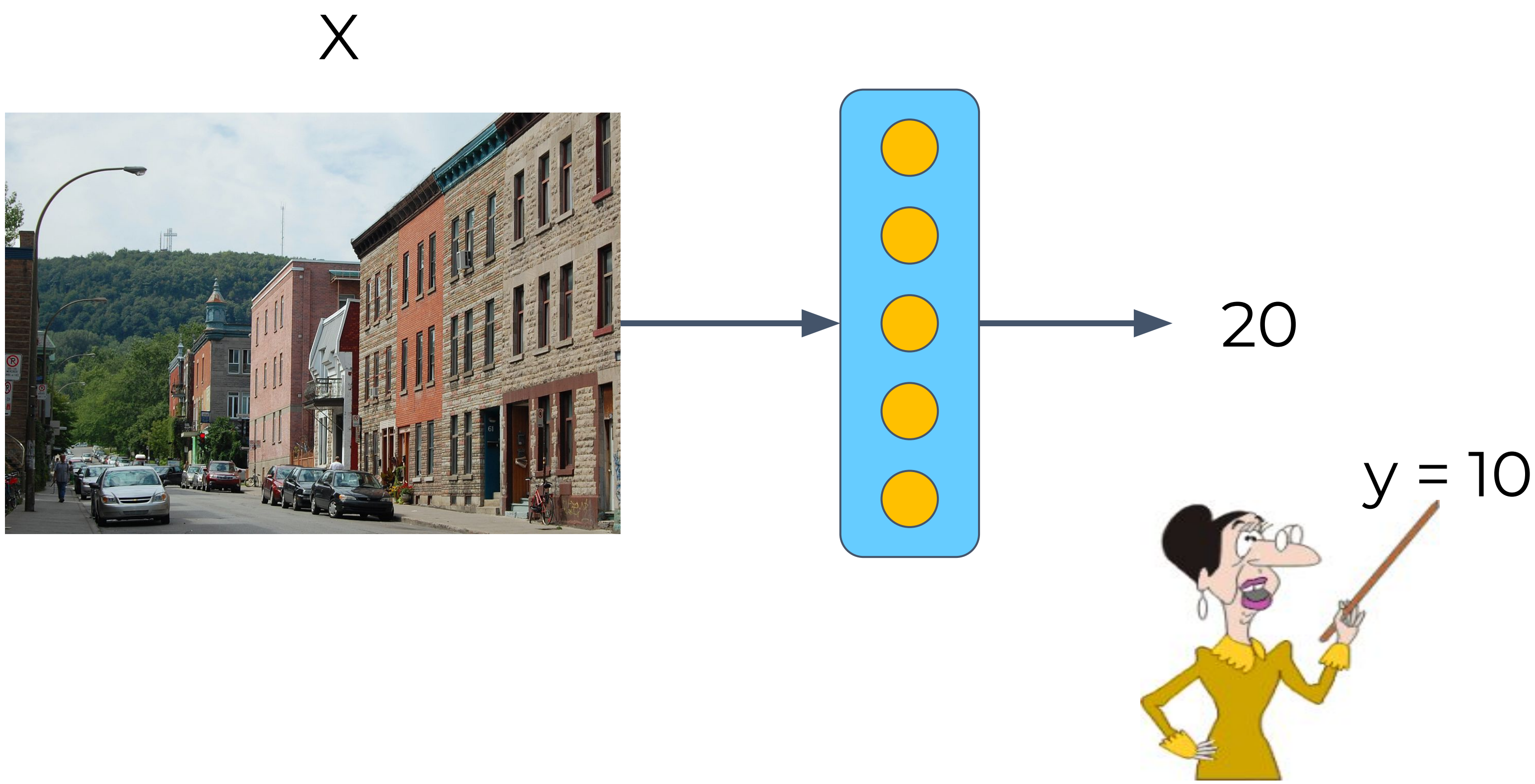
Chat	90%
Chien	1%
Cheval	0%
Tigre	8%
Pinguin	0%
...	1%

Coût: -ln(0.9)

y = Chat

MILA

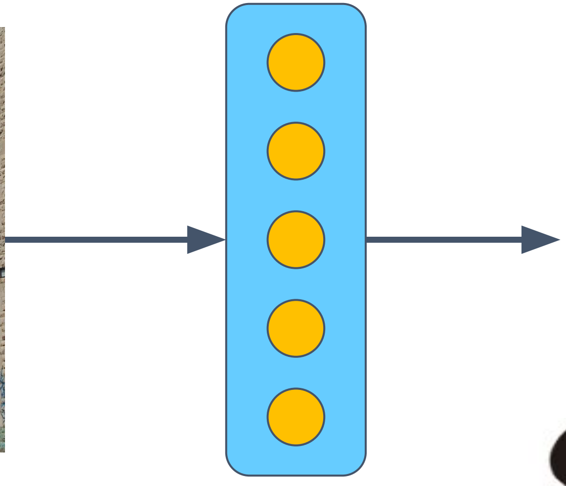
Exemple: réduire l’erreur au carrée



Exemple: réduire l'erreur au carrée

Mean Squared Error

X



20

Coût: $(20-10)^2$

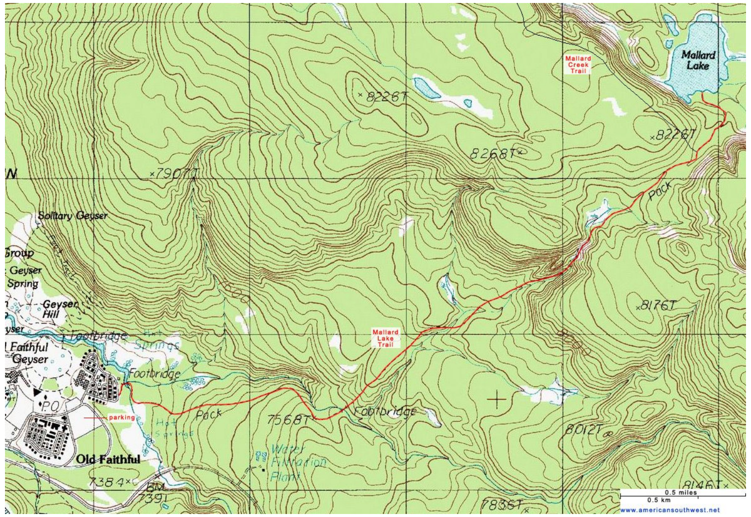
y = 10

Fonction de coût

- Rétroaction sur la réponse du modèle

- Moyenne sur tous les exemples de l’ensemble d’entraînement

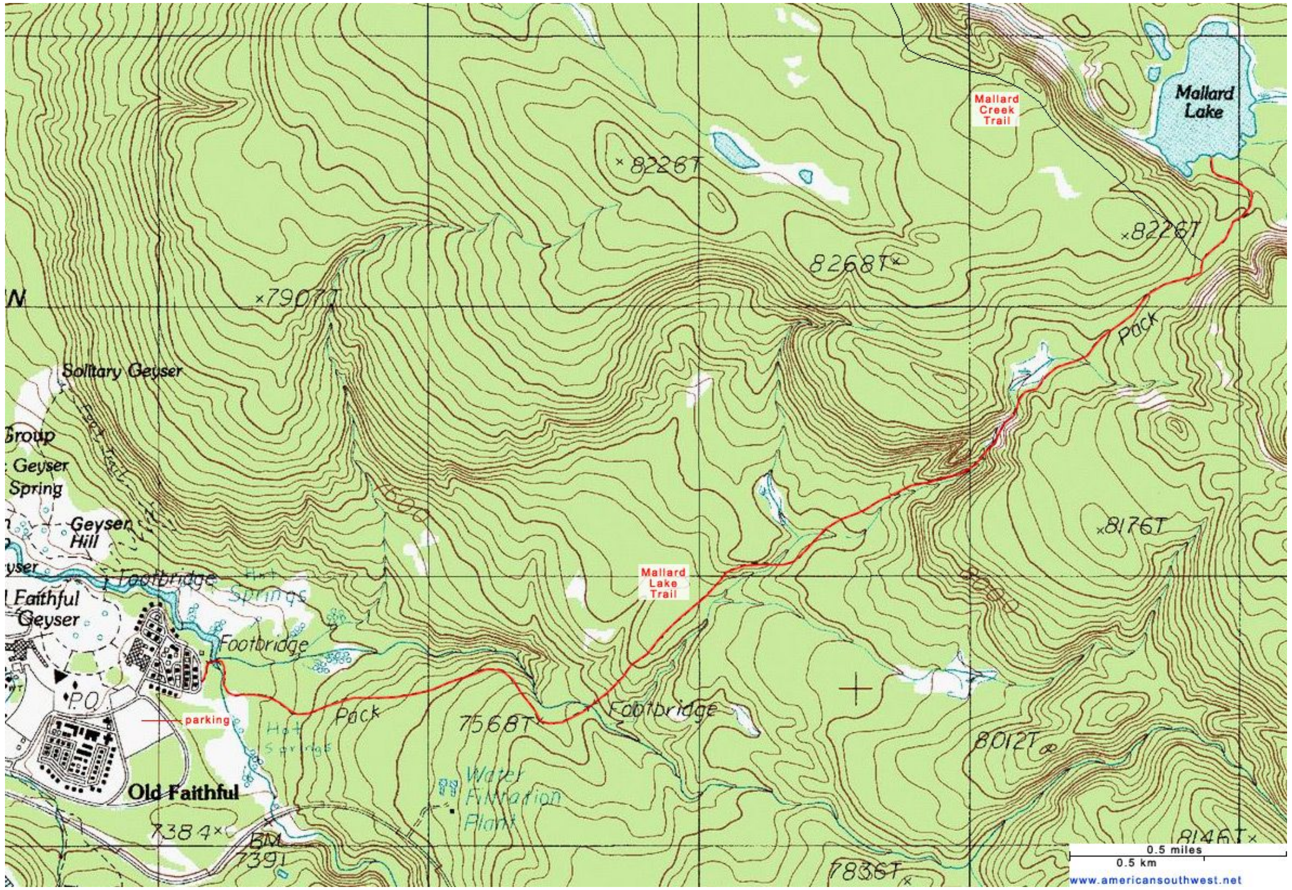
- Définie les courbes de niveaux (altitude)



MILA

Calcul de la pente

- Un petit pas dans chaque direction



MILA

Gradient: calcul de la pente

- Fonction différentiable: la fonction s’approxime bien avec une fonction linéaire (gradient)

- Théorème: le gradient donne la direction de la plus forte pente.

- Attention: information locale

Terminologie

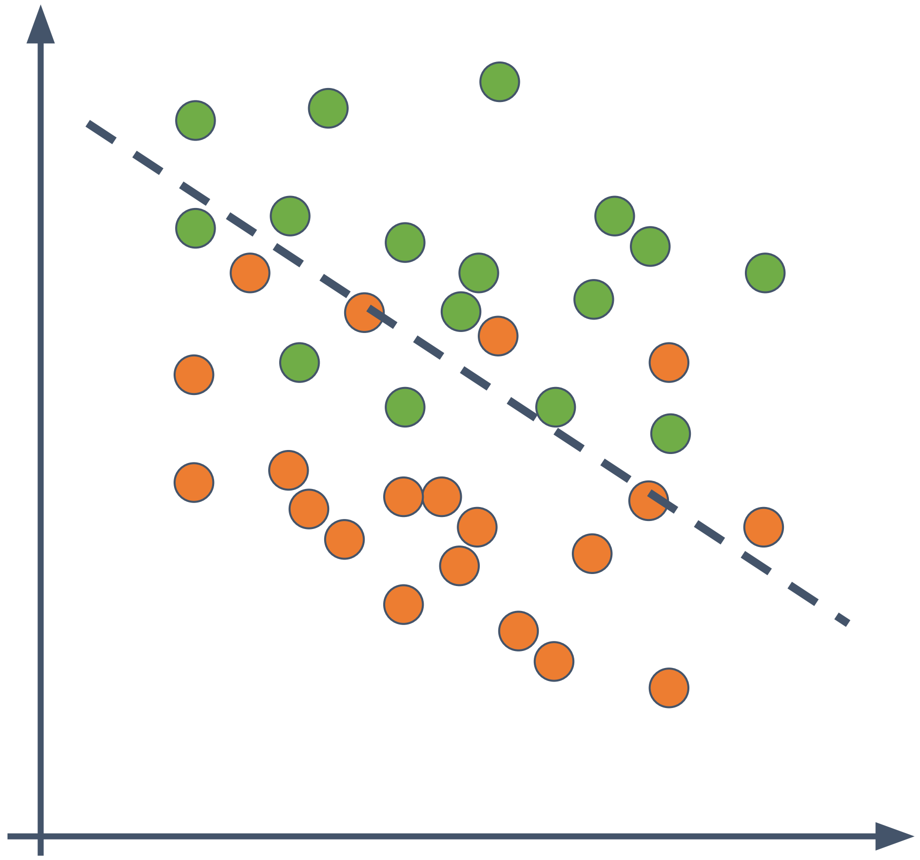
- Calcul du gradient: backpropagation
- Se déplacer avec le gradient: optimiseur

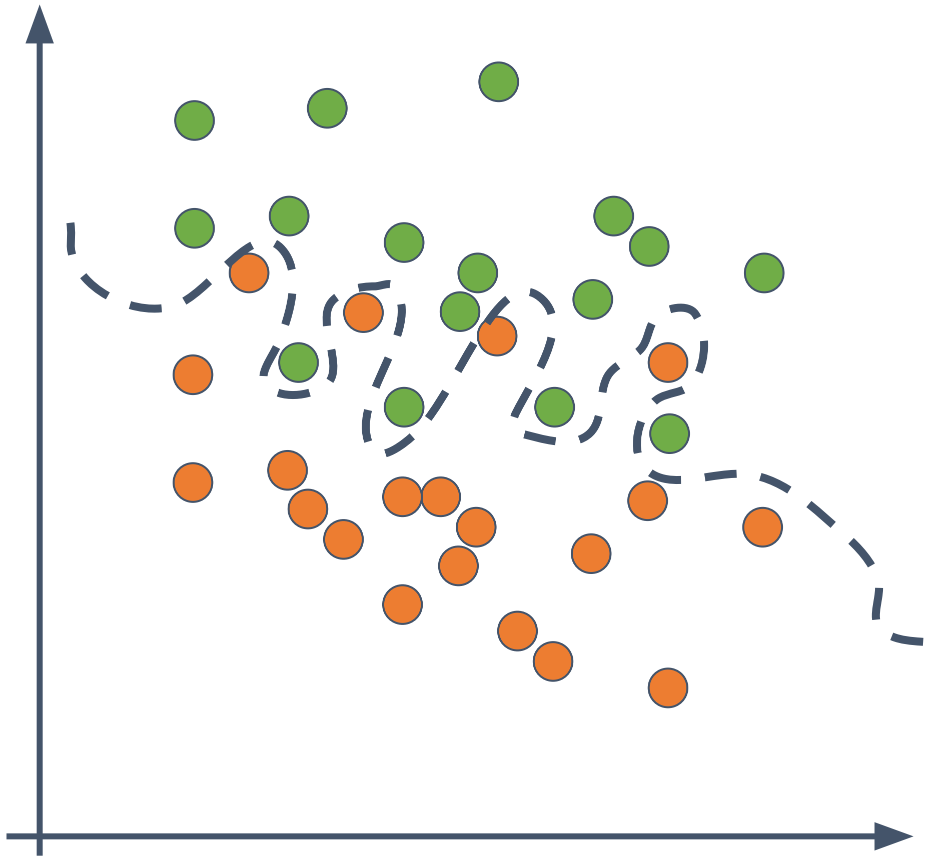
Base d’apprentissage

Décomposition en trois ensembles:

- ensemble **d’entraînement**
- ensemble de **validation**
- ensemble de **test**

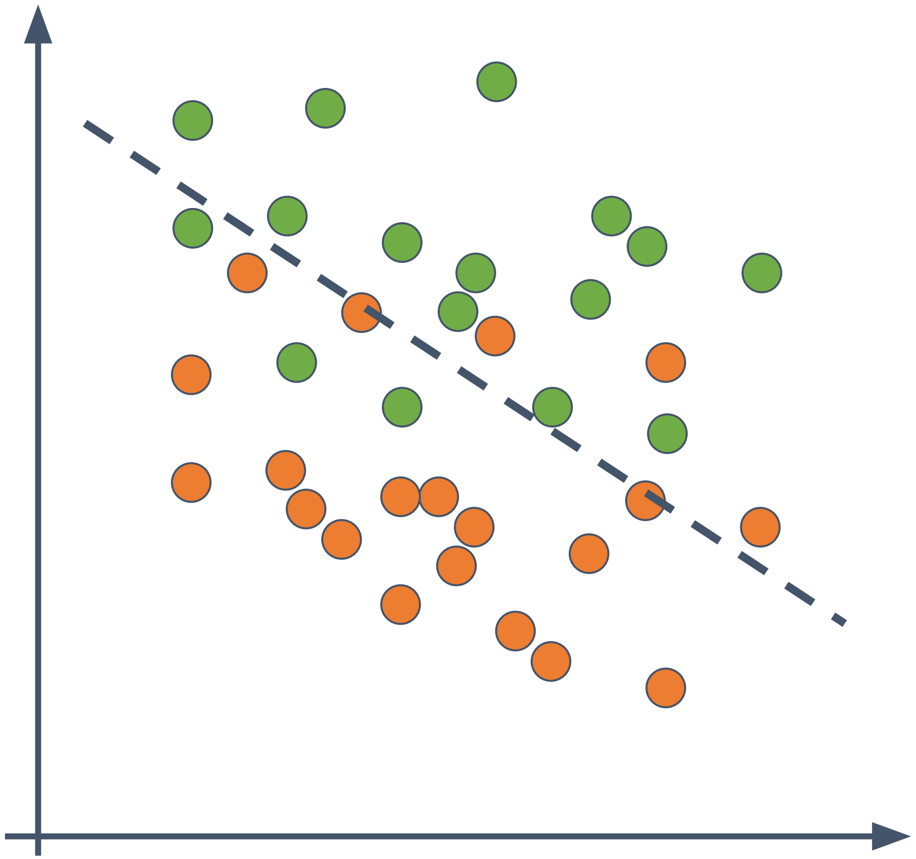
Sur-apprentissage: classification

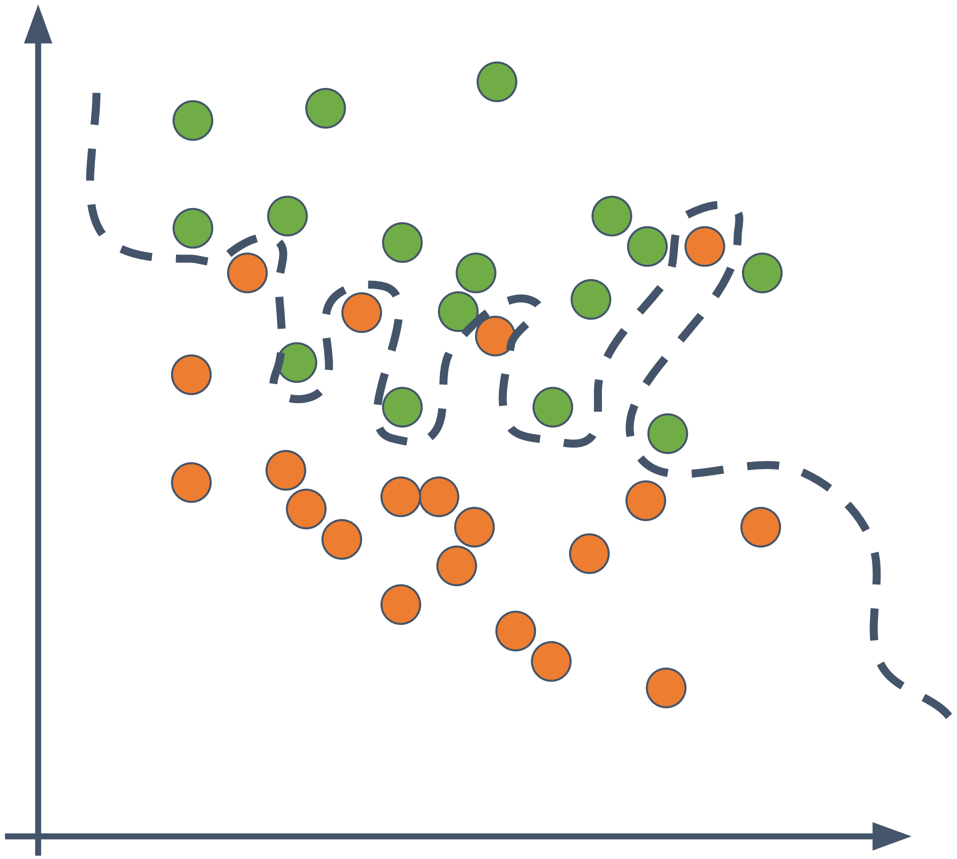




MILA

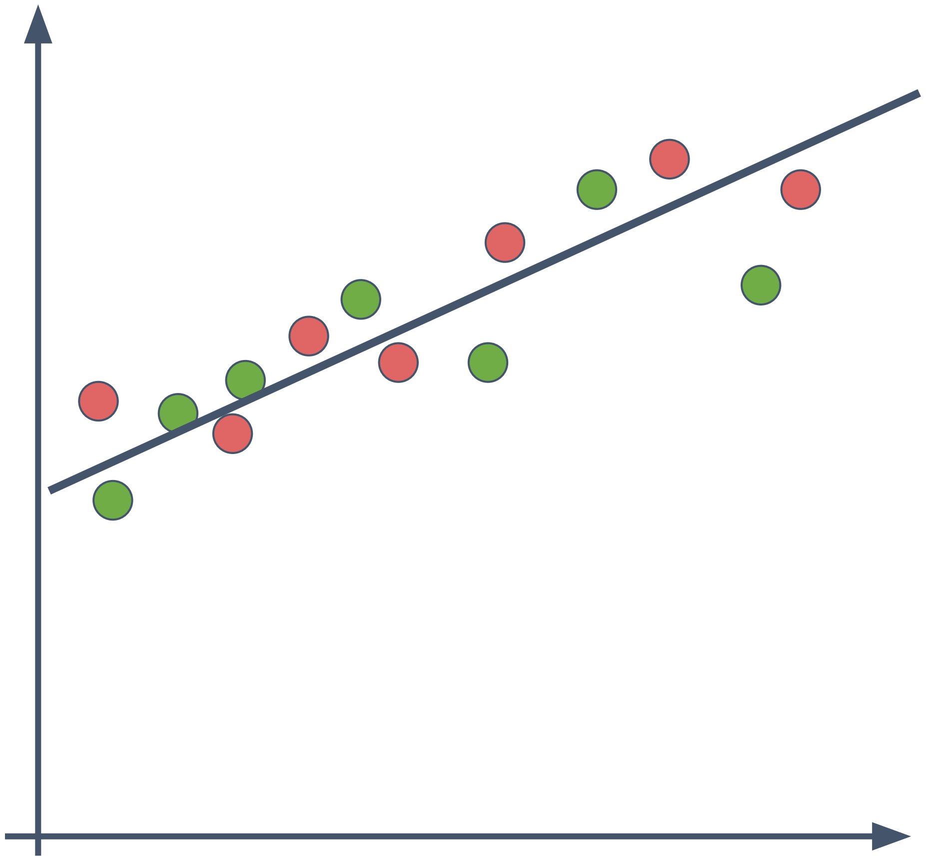
Sur-apprentissage: classification

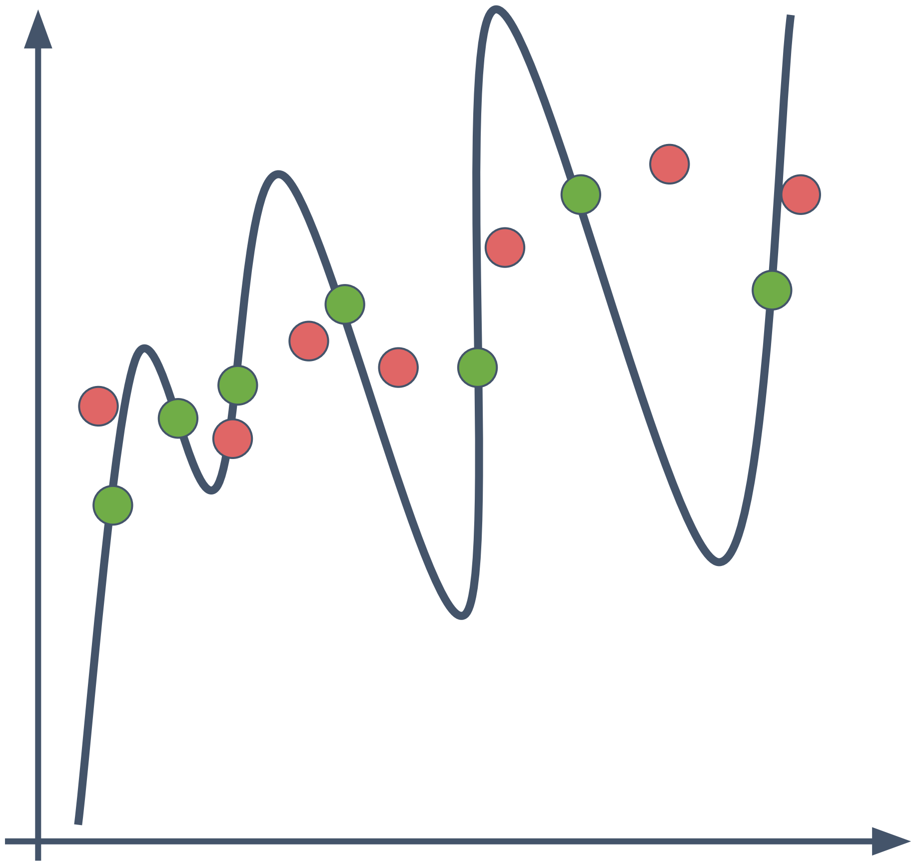




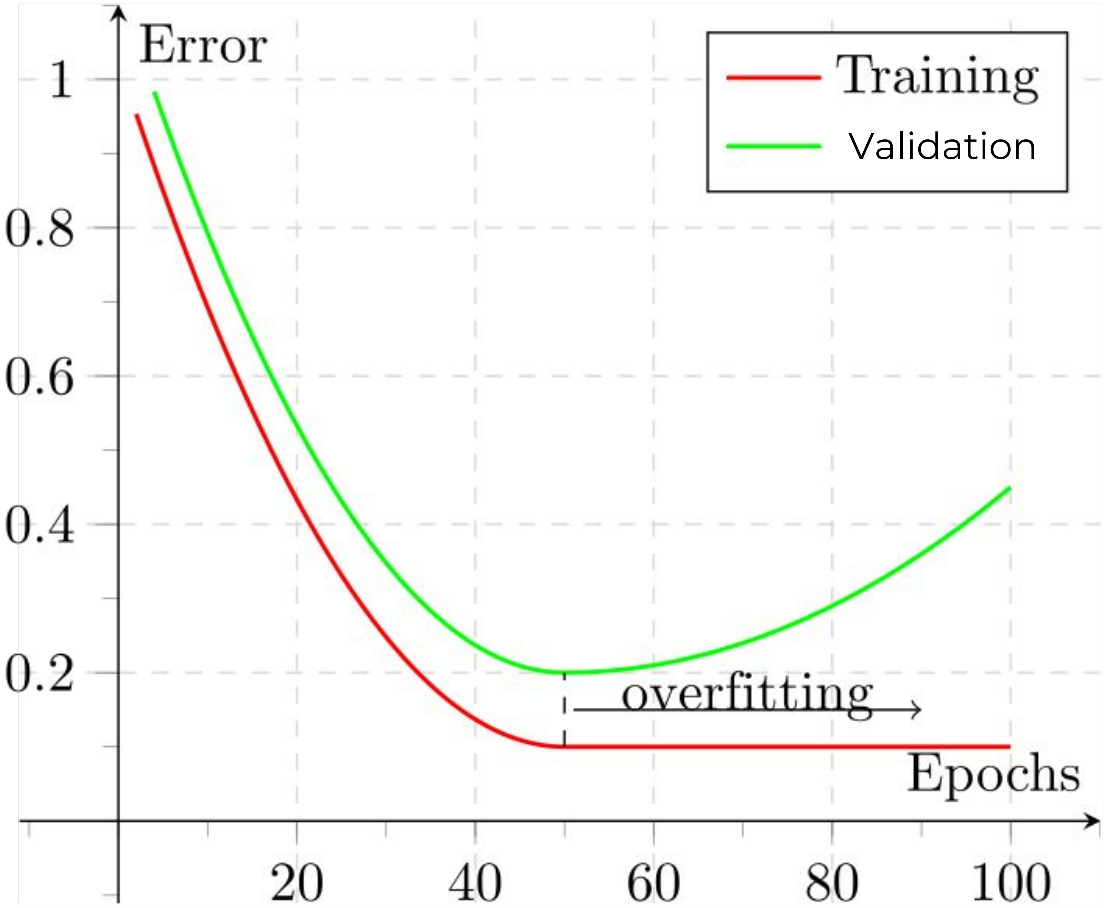
MILA

Sur-apprentissage: régression



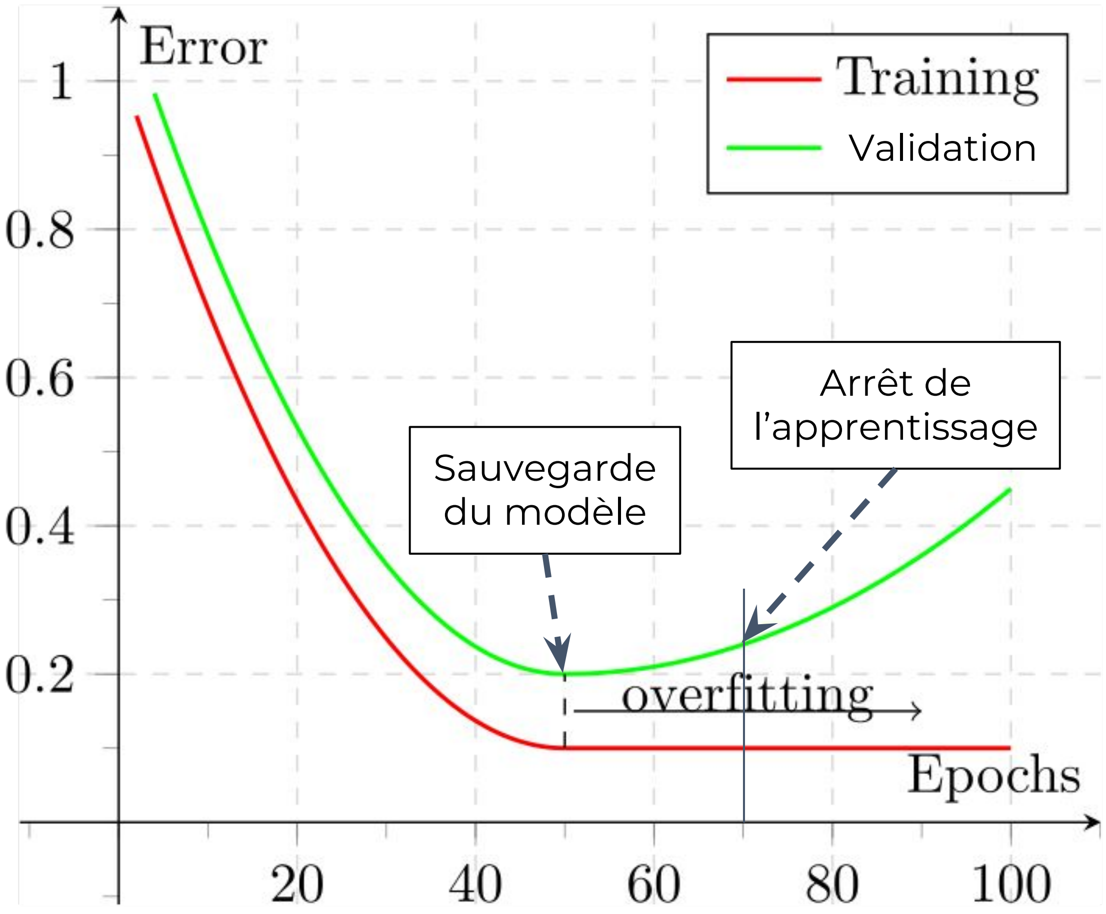


Early Stopping



MILA

Early Stopping



Fonction de coût: régularisation

- Lutter contre le sur-apprentissage

- Pénalité sur les paramètres (coordonnées)

- Norme L1 ou L2

- D'autres types de régularisations plus efficaces

Hyperparamètres

- Choix du modèle

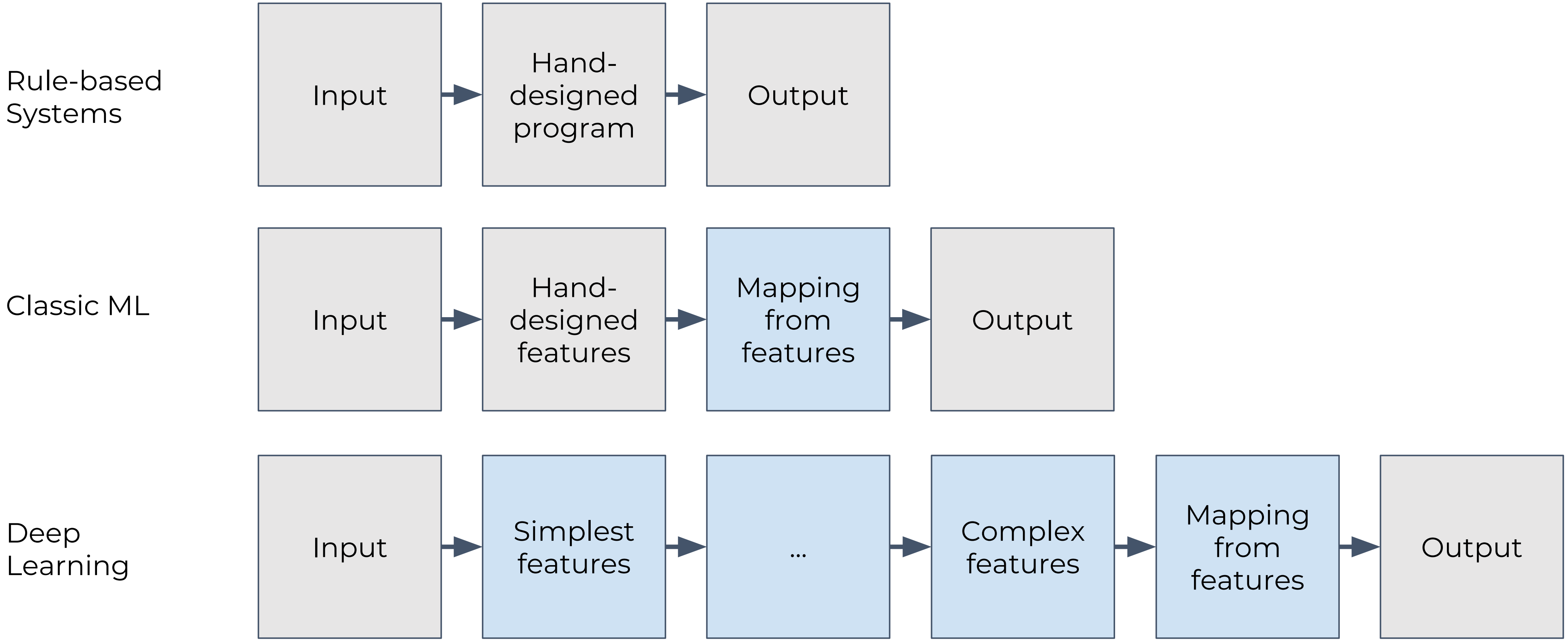
- Structure du modèle

- Paramètres de l'optimiseur

Apprentissage profond

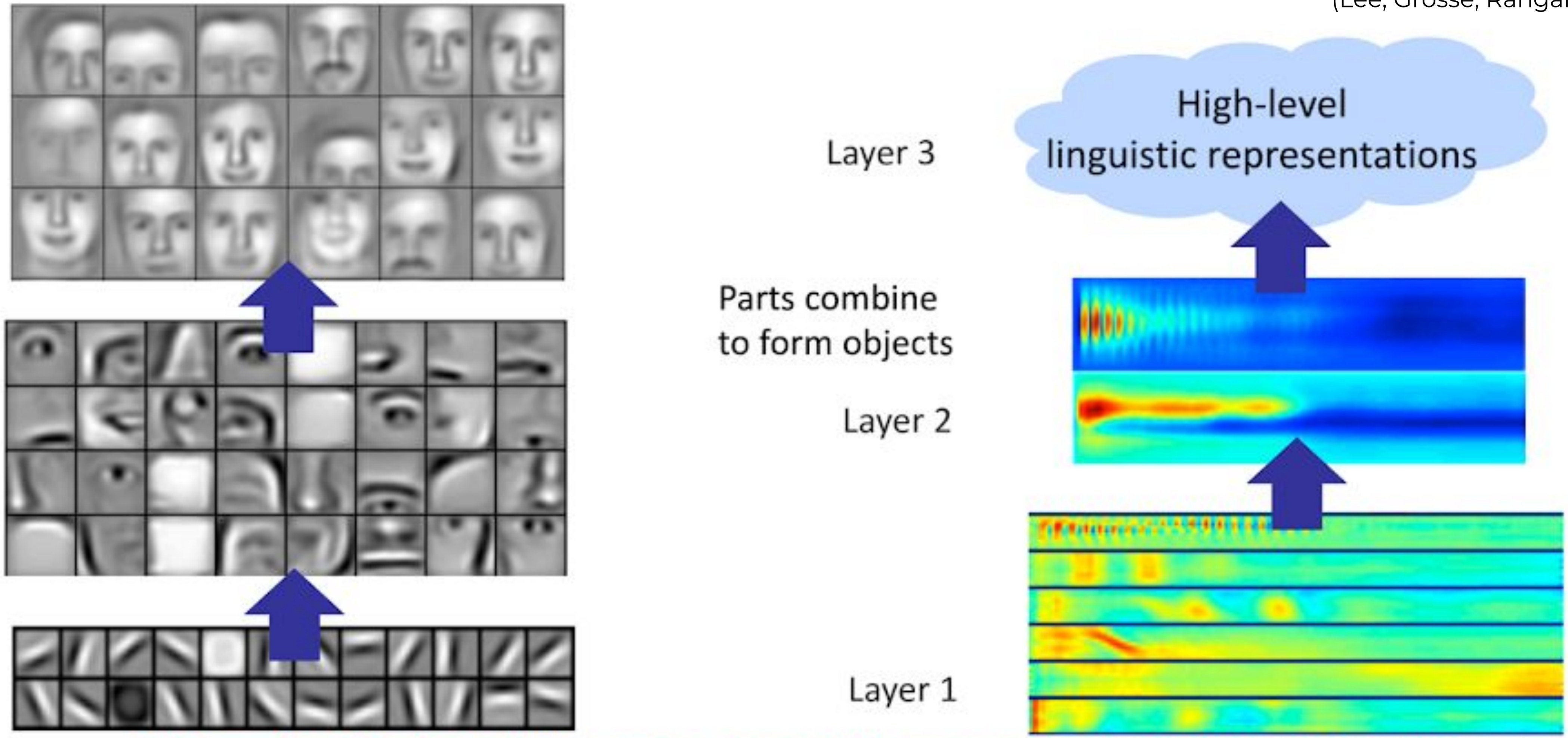
(ou apprentissage de représentations)

Modèle en apprentissage profond



Niveaux de représentations

(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)



MILA

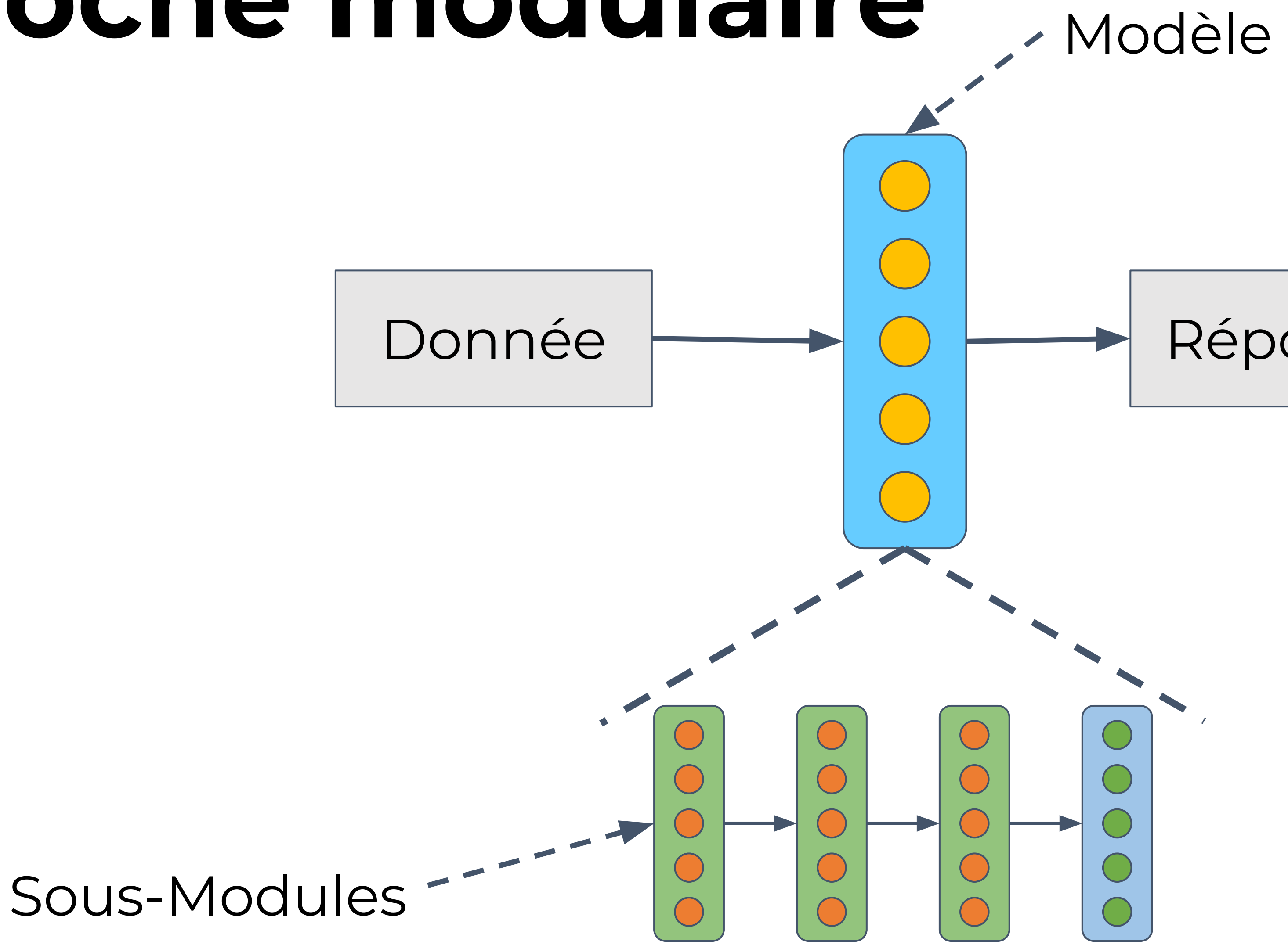
Principe de compositionnalité

MILA

Ingrédients au DL

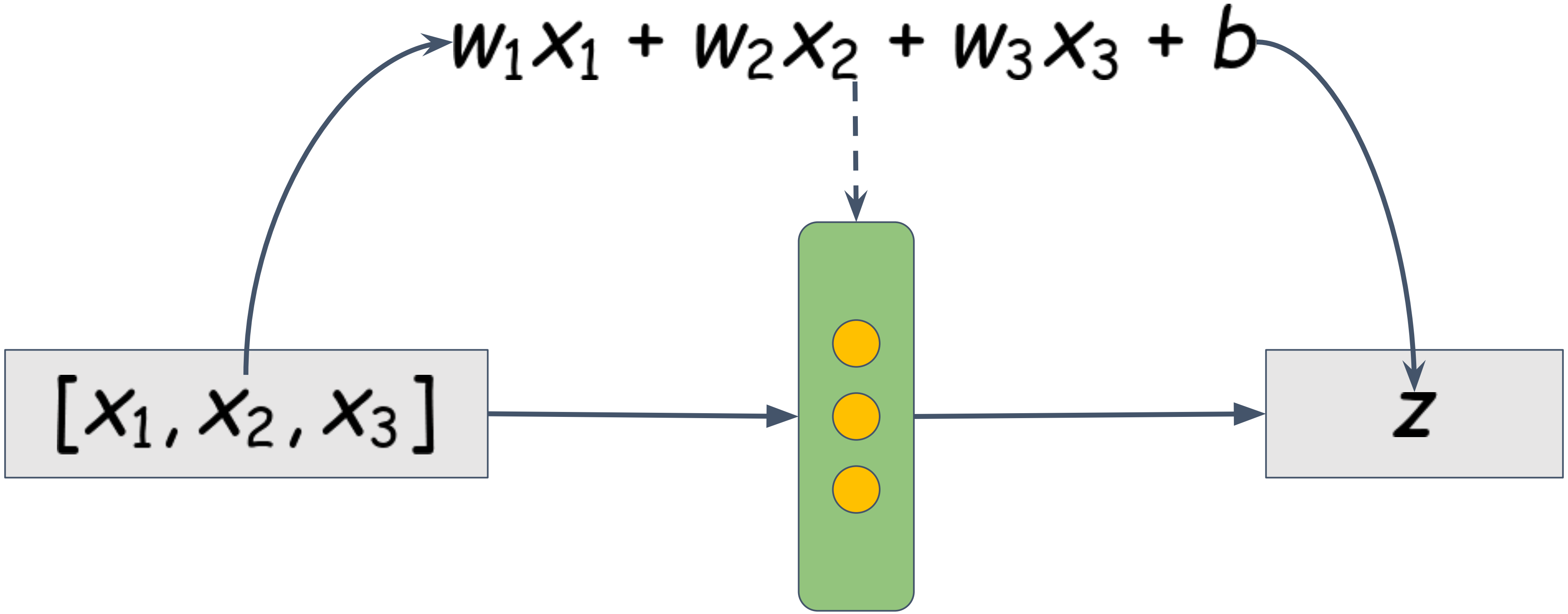
- Une **quantité importante** de données
- Graphe computationnel **différentiable** de bout en bout
- Un algorithme de calcul de gradient: **backpropagation**
- Des fonctions d'activation **non-linéaires**
- Un **optimiseur** itératif de paramètres

Approche modulaire



MILA

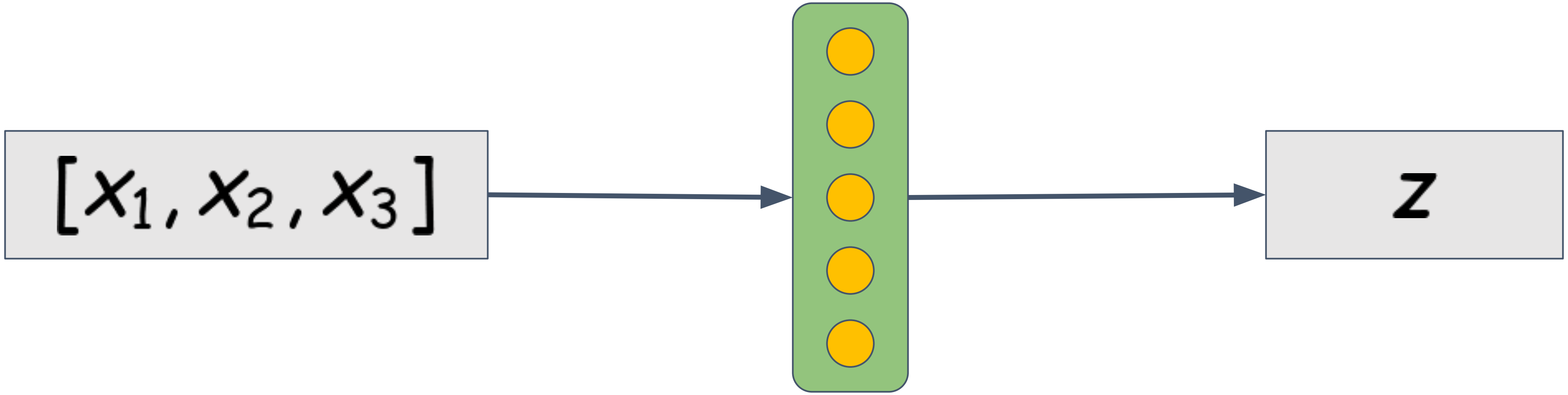
Transformation linéaire



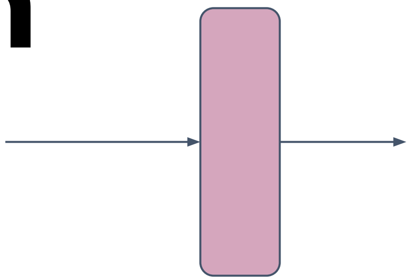
MILA

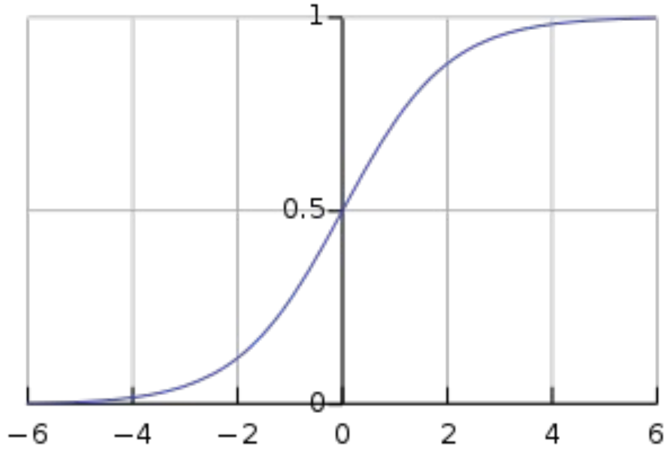
Transformation linéaire

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & b_1 \\ w_{21} & w_{22} & w_{23} & b_2 \\ w_{31} & w_{32} & w_{33} & b_3 \\ w_{41} & w_{42} & w_{43} & b_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix}$$

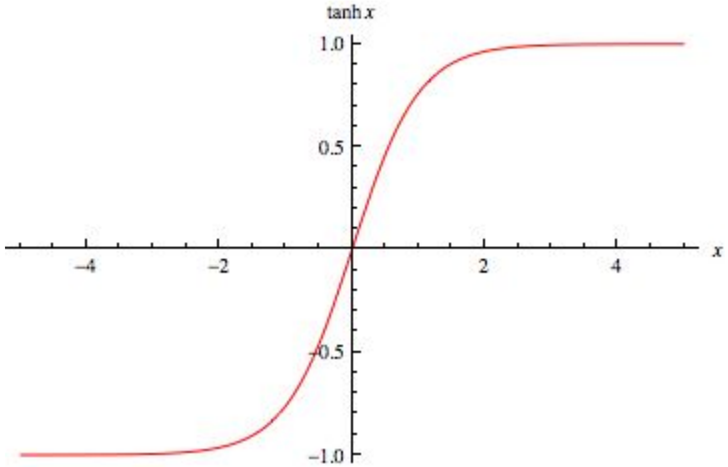


Fonctions d’activation

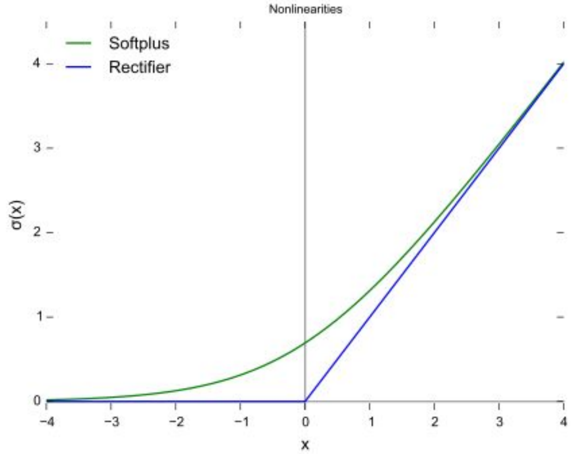




Sigmoid



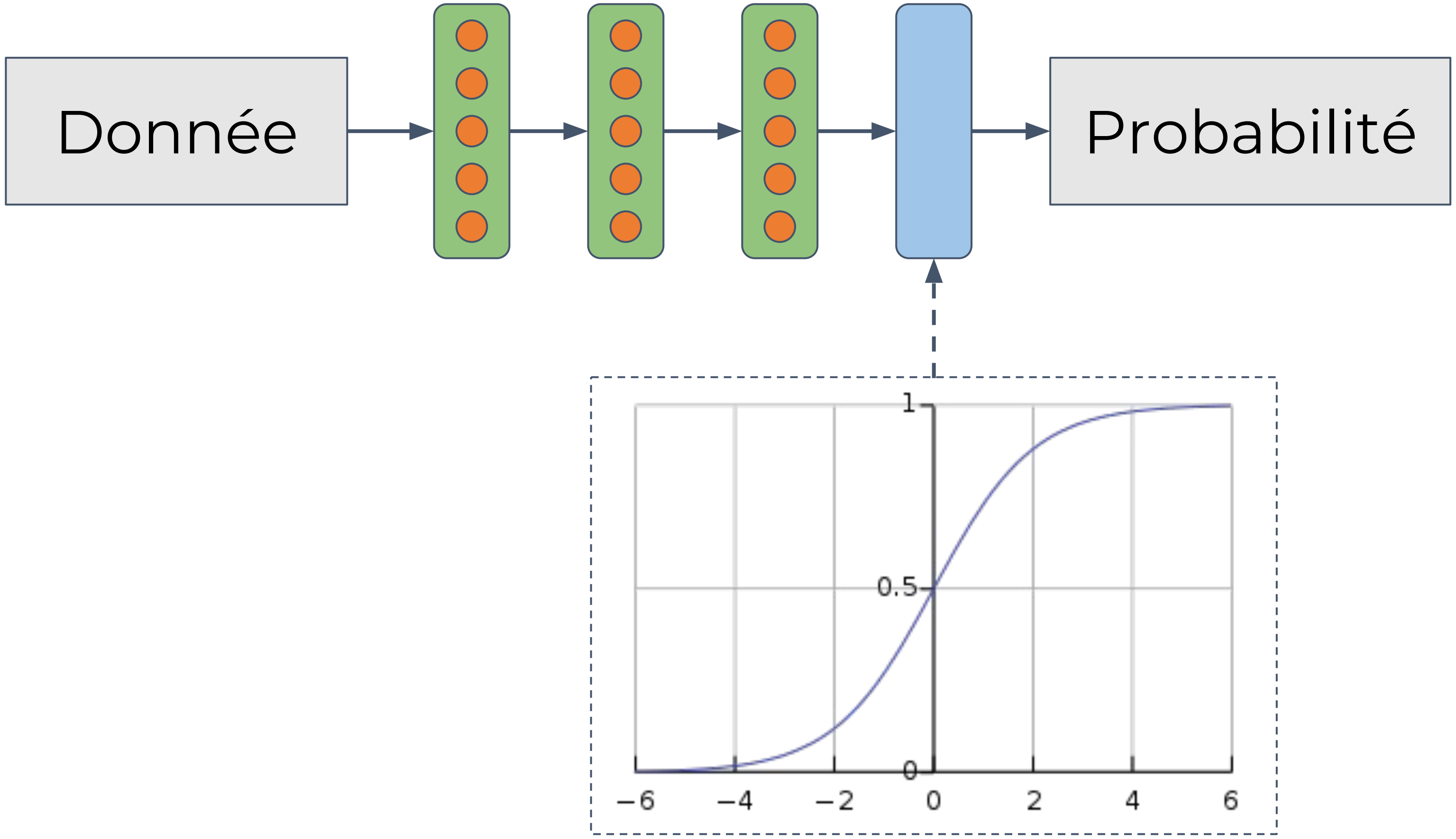
Tanh



ReLU

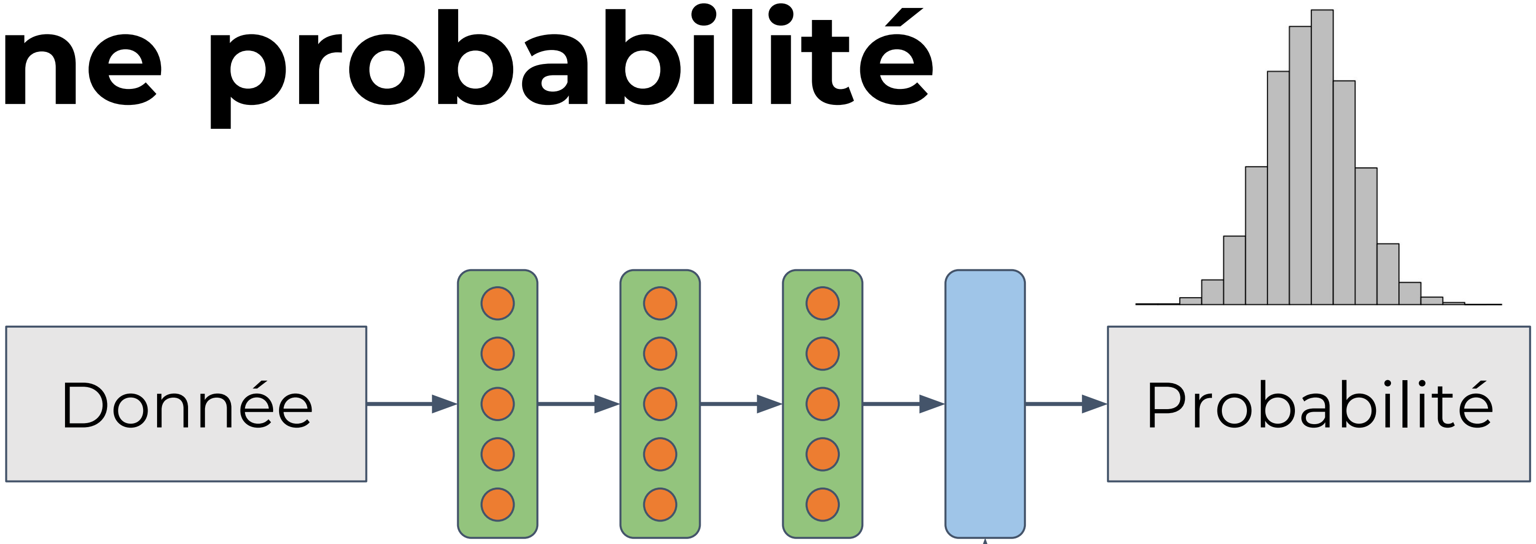
Modéliser une probabilité

- **Sigmoid**
- Softmax
- Modèle gaussien



Modéliser une probabilité

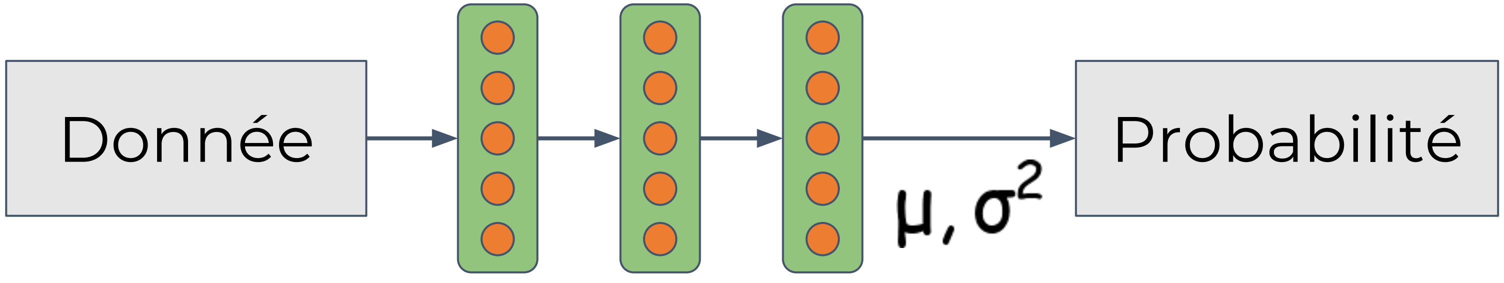
- Sigmoid
- **Softmax**
- Modèle gaussien

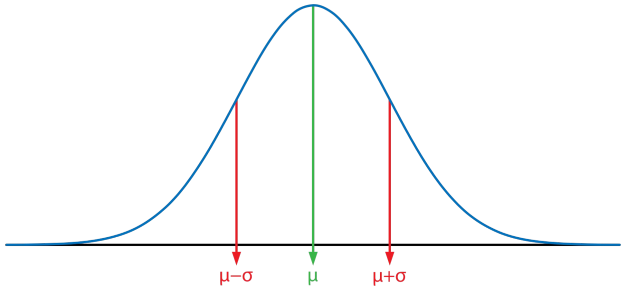


$$p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

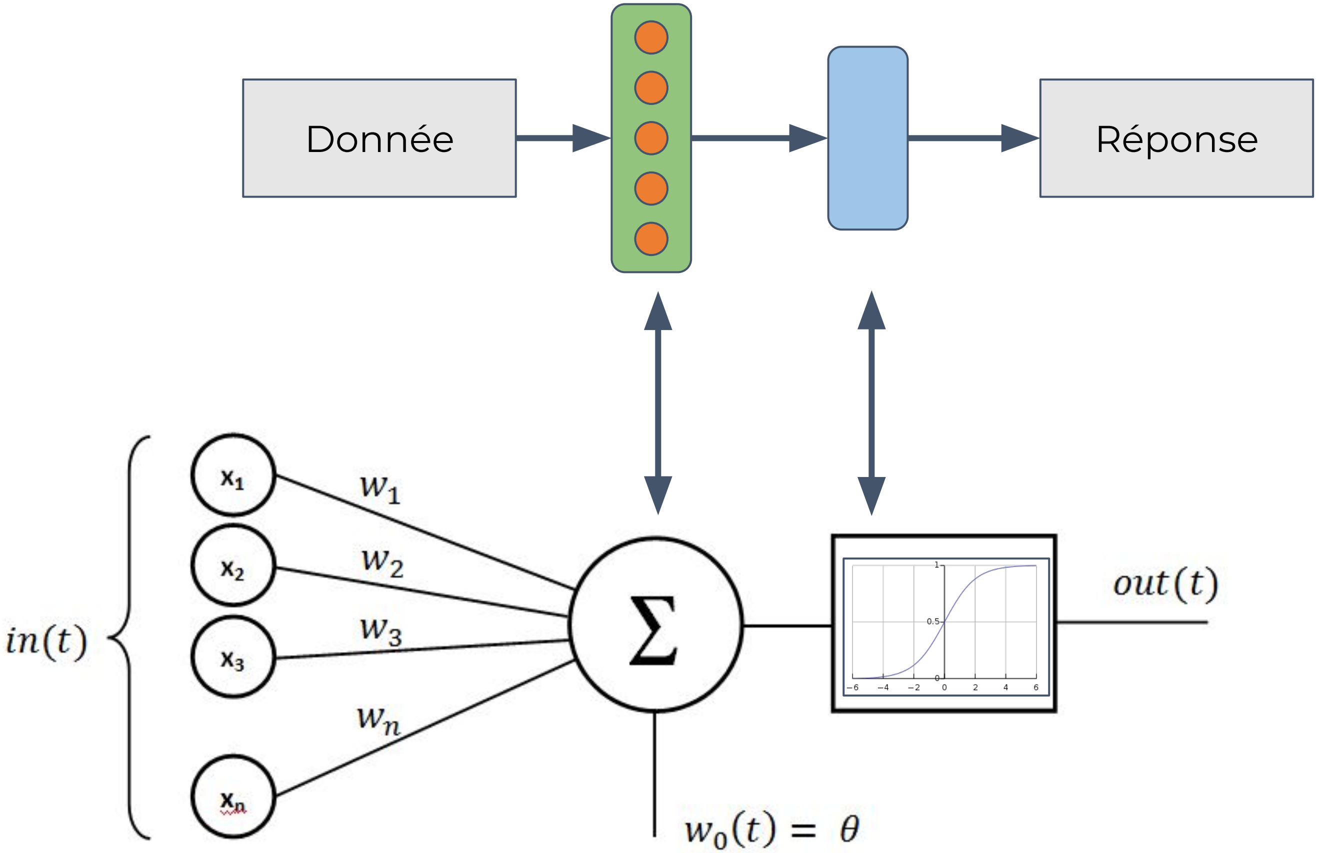
Modéliser une probabilité

- Sigmoid
- Softmax
- **Modèle gaussien**

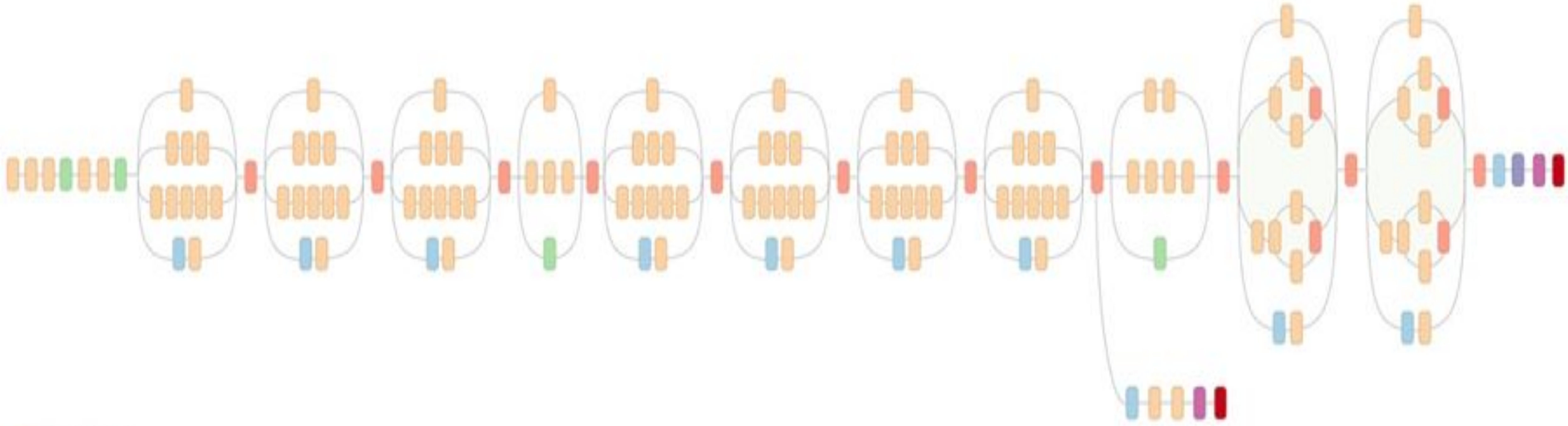




Exemple: régression logistique

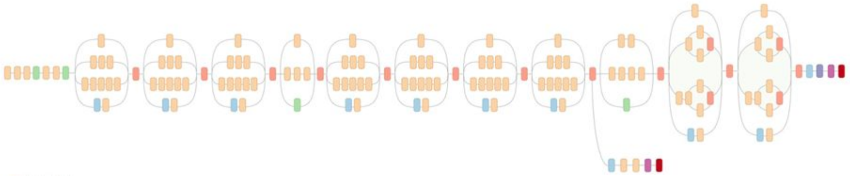


Exemple: Inception V1



MILA

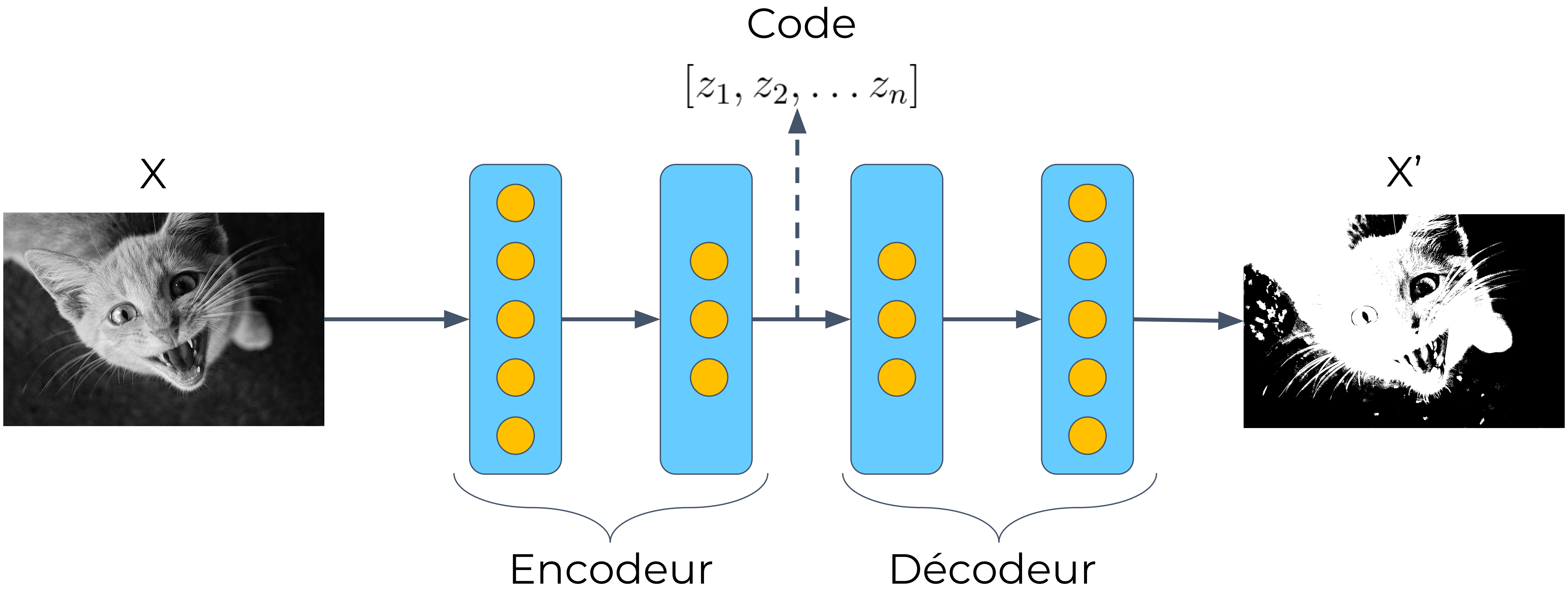
State of the science

	Empirisme	Théorie
DL		?
Aviation		

MILA

Exemples

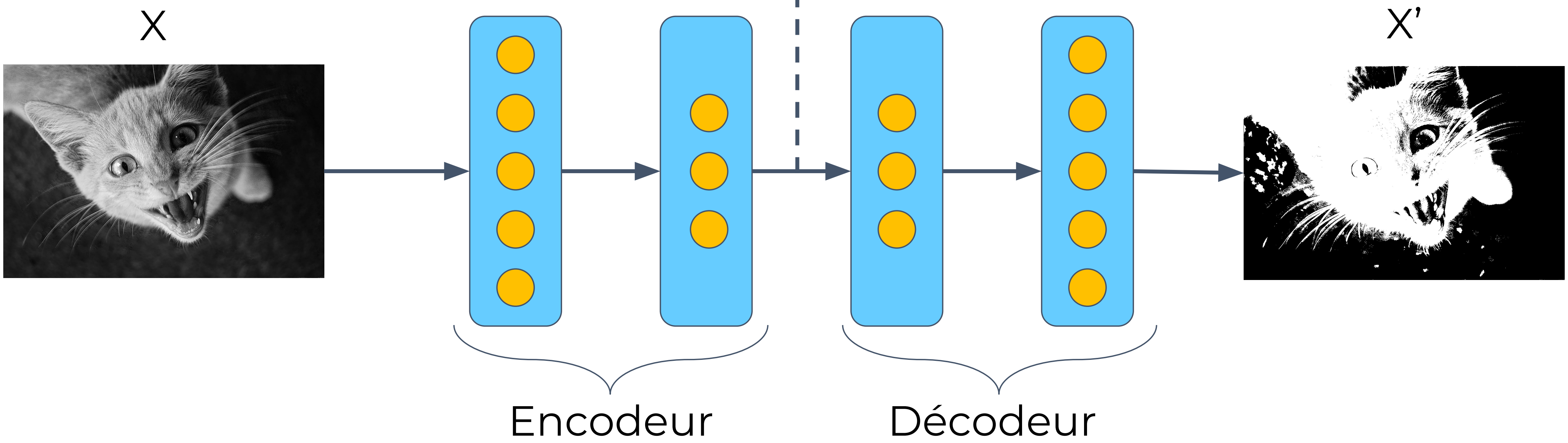
Exemple: auto-encodeur



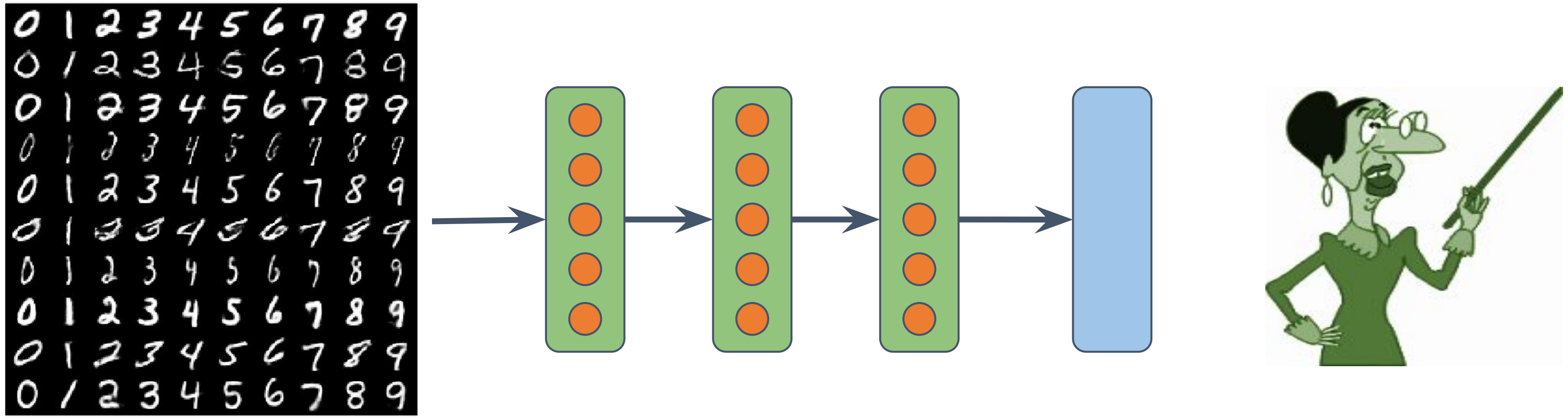
Exemple: auto-encodeur

Code

$[z_1, z_2, \ldots z_n]$



Transfer learning

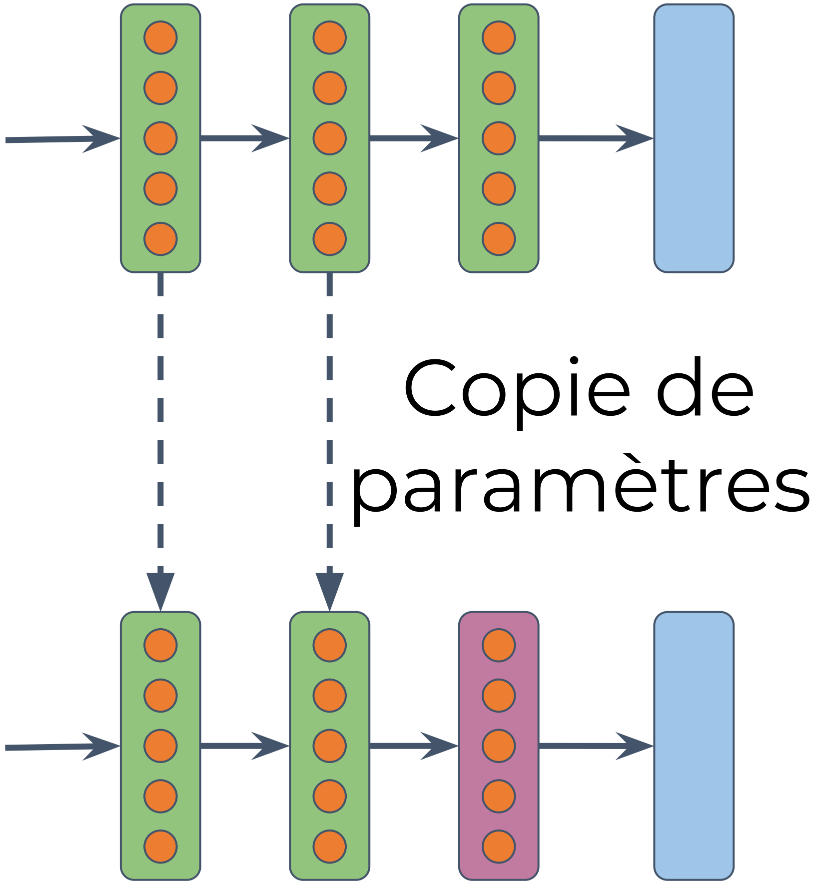


(a) MNIST

Transfer learning

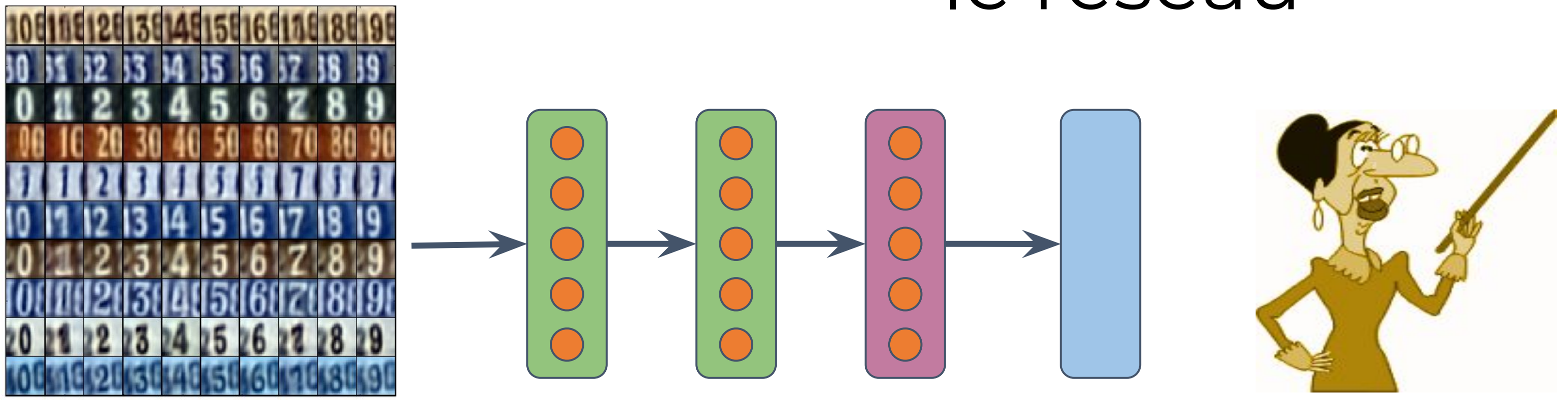
(a) MNIST

(b) SVHN



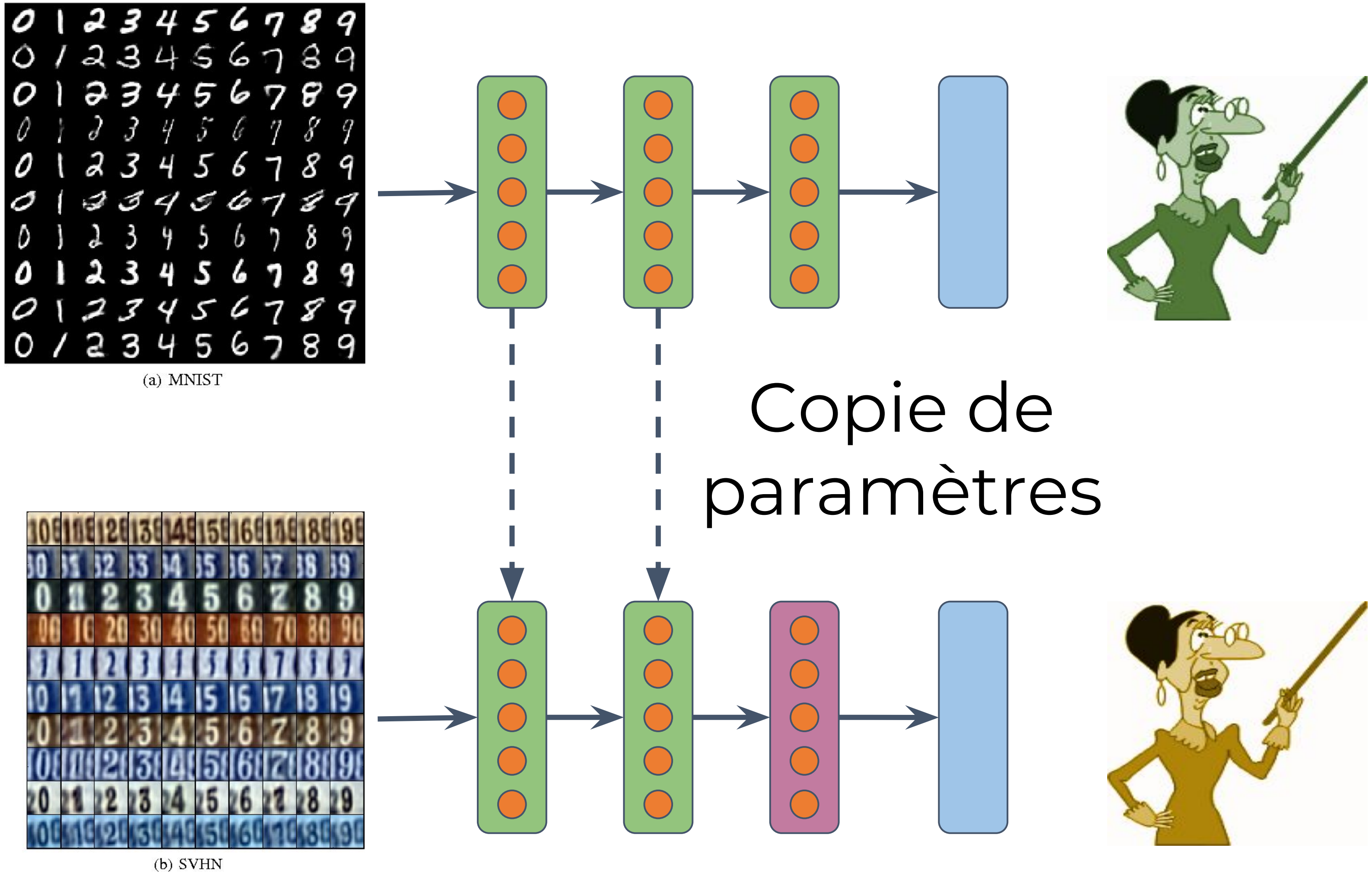
Transfer learning

On entraîne tout
le réseau



(b) SVHN

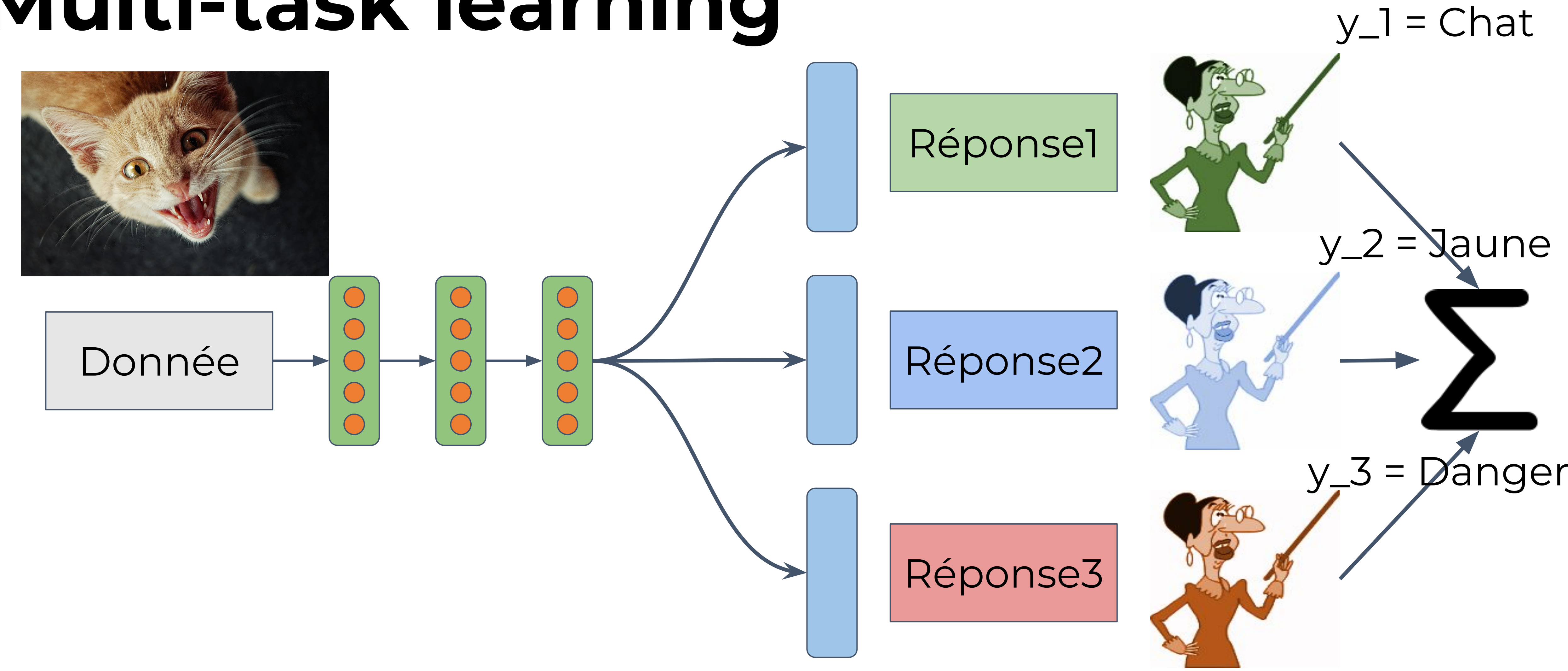
Transfer learning



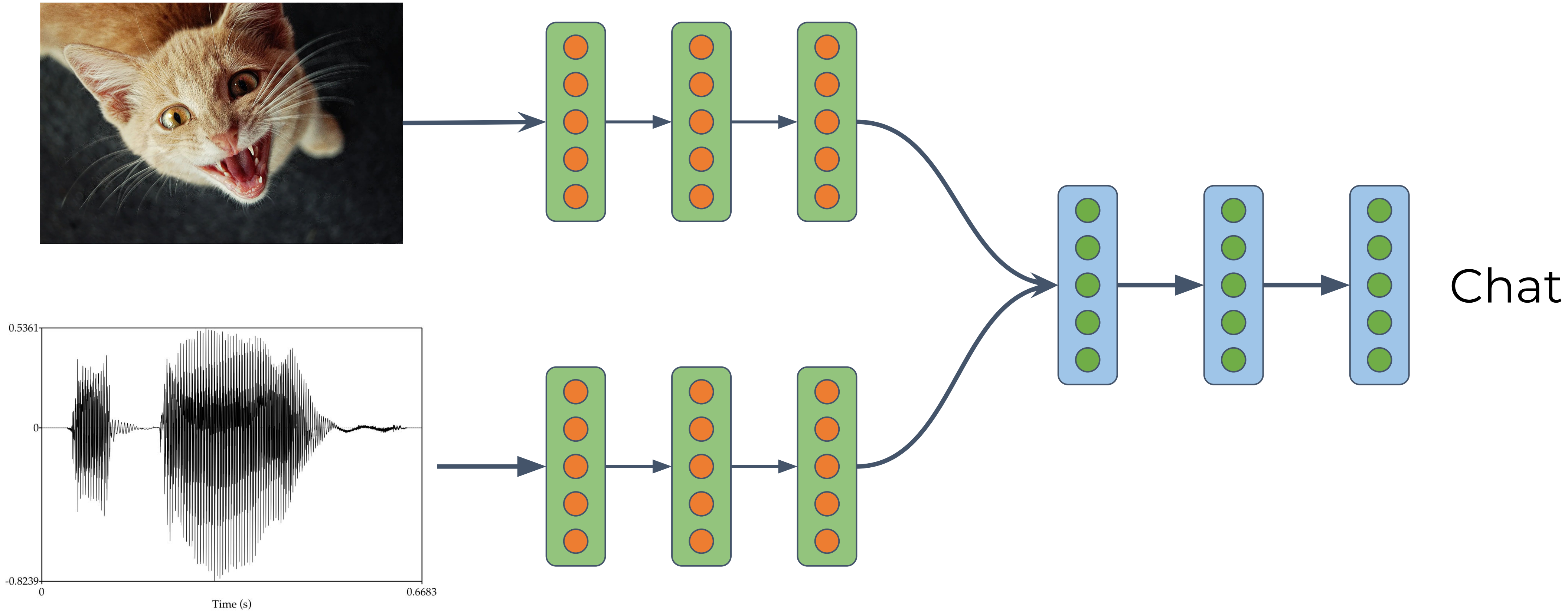
(a) MNIST

(b) SVHN

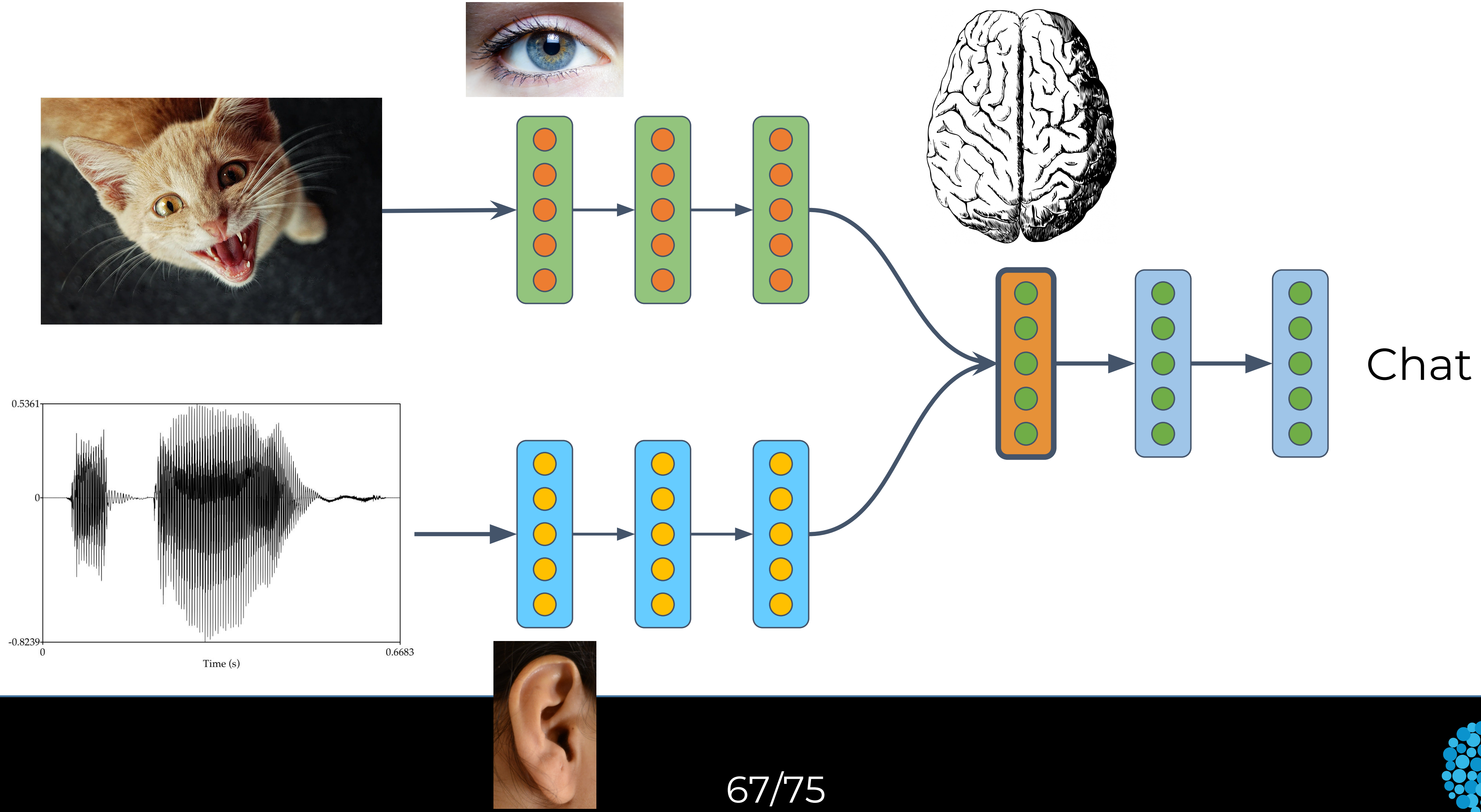
Multi-task learning



Représentation latente commune

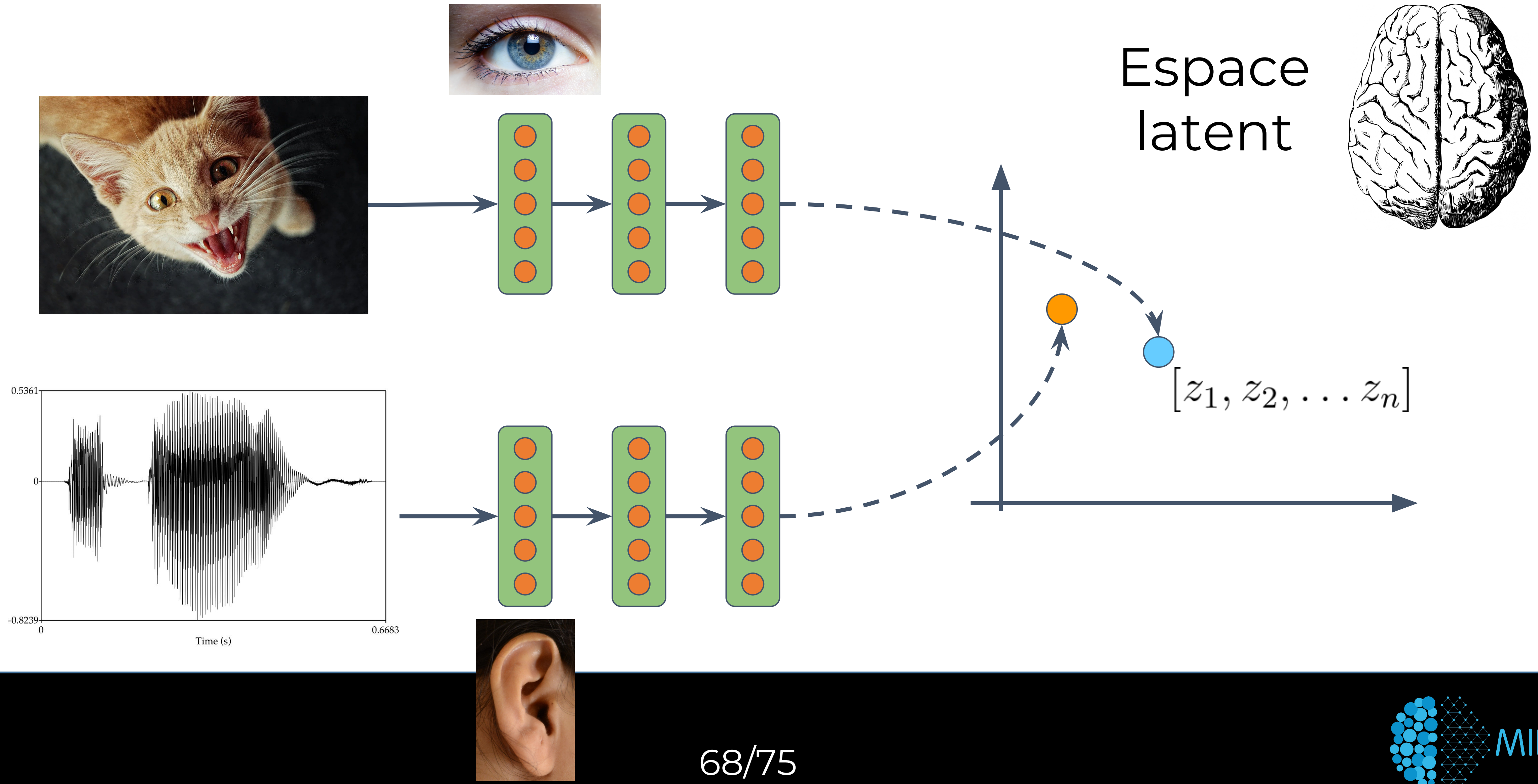


Représentation latente commune



MILA

Représentation latente commune



Références

- Leçon inaugurale de Stéphane Mallat au Collège de France: https://goo.gl/7Lg3g6

- 3Blue1Brown: But what *is* a Neural Network? | Chapter 1, deep learning: https://goo.gl/BMRoPo

Merci pour votre attention!

Contacts

https://mila.quebec/

Gaétan Marceau Caron, Membre MILA R&D,
gaetan.marceau.caron@rd.mila.quebec

Myriam Côté, Directrice équipe MILA R&D,

myriam.cote@rd.mila.quebec

MILA

Merci pour votre attention!