



Réseaux de Neurones pour la prédiction de trajectoires de voitures

Jean Mercat

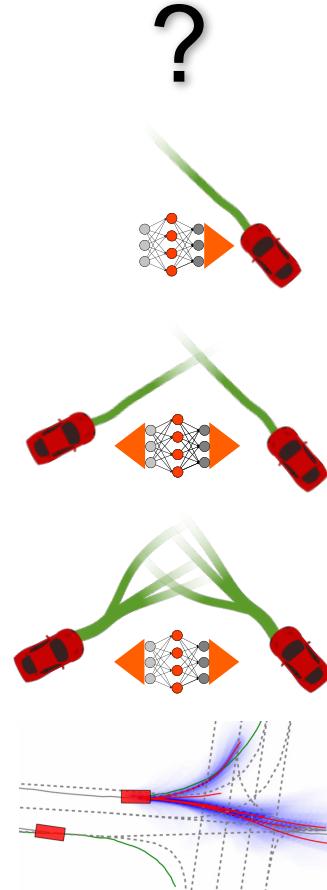
Définitions des objectifs

Prédiction avec des réseaux neuronaux simples

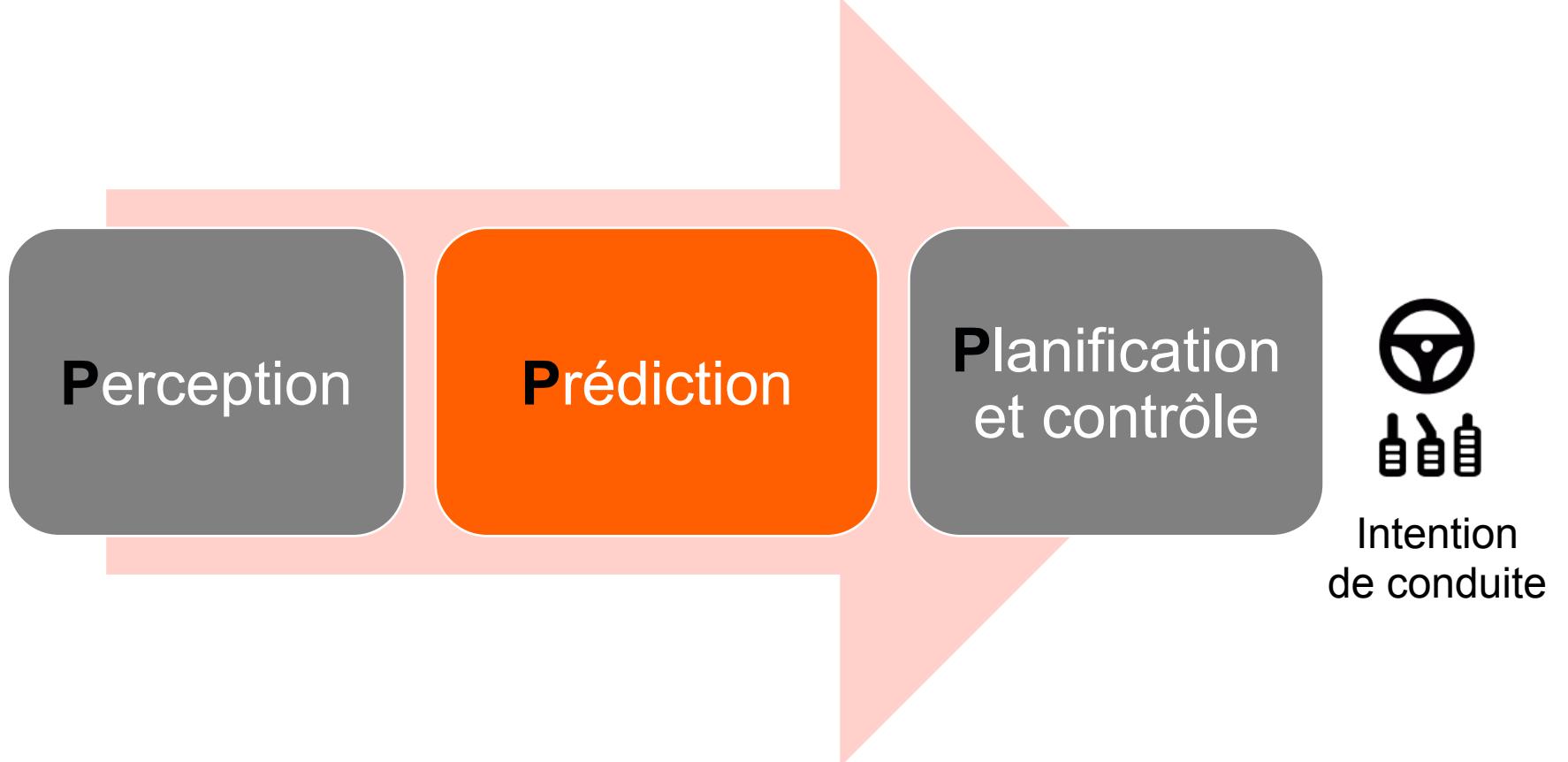
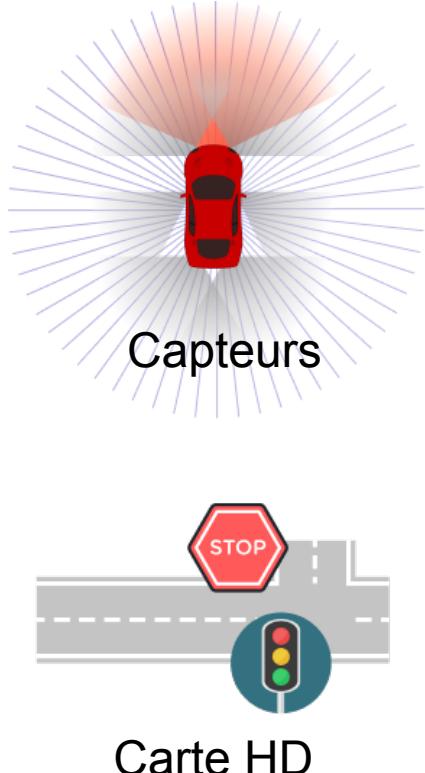
Réseaux neuronaux pour les scènes routières complexes

Prédiction de plusieurs futurs possibles

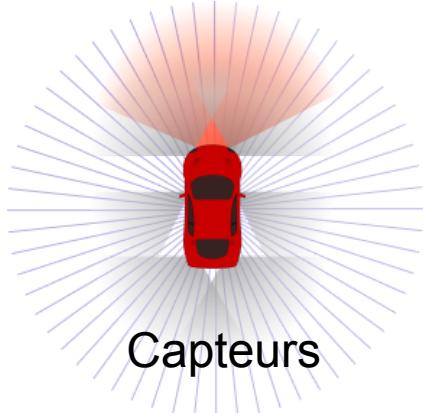
Application



La modularité PPP



Perception de la scène routière



Perception



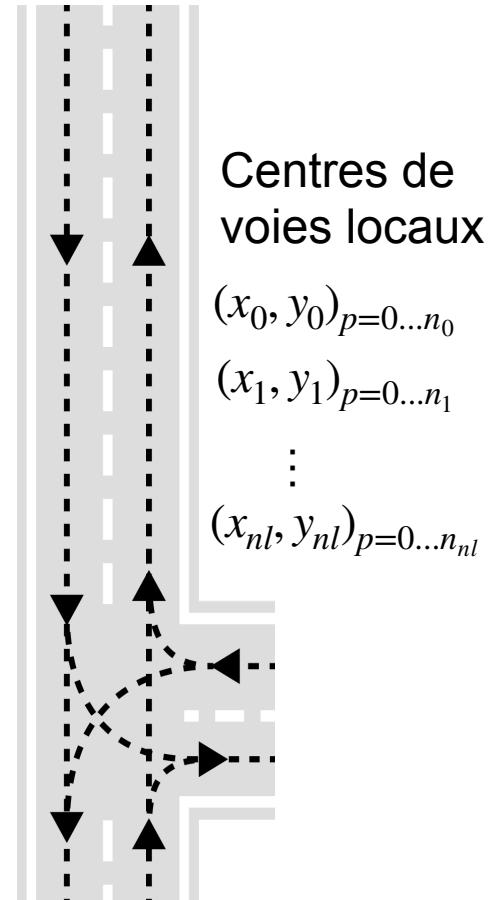
Observations ego
 $(x_0, y_0)_{t=-t_h \dots 0}$



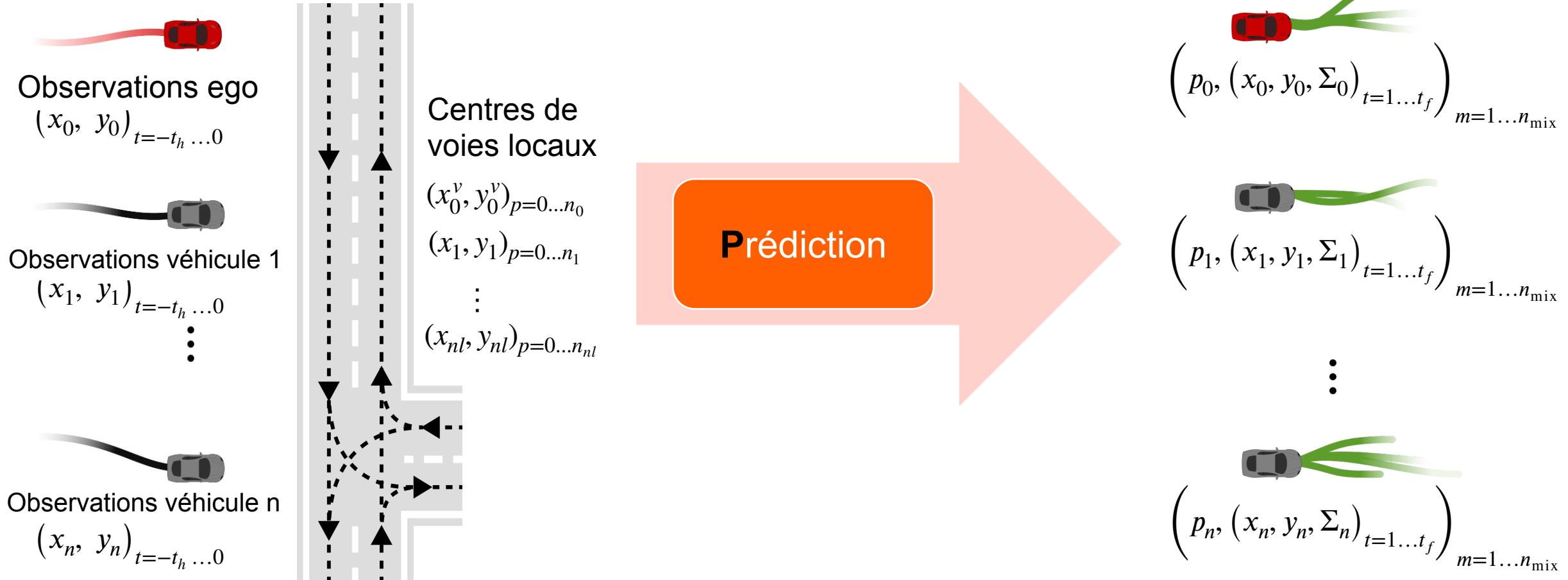
Observations véhicule 1
 $(x_1, y_1)_{t=-t_h \dots 0}$
⋮



Observations véhicule n
 $(x_n, y_n)_{t=-t_h \dots 0}$



Résultats attendus de la prédiction

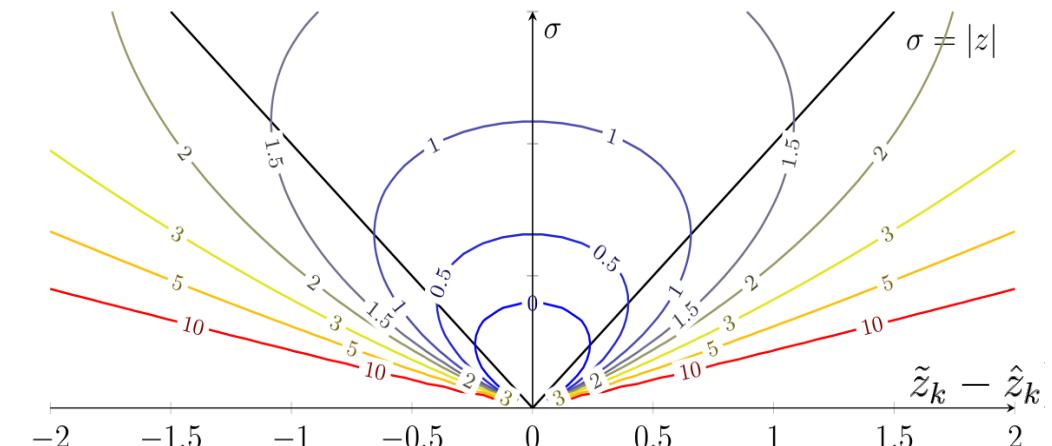


Coût multi-modal pour l'apprentissage

Log-vraisemblance négative :

$$\text{NLL}(k) = -\ln \left(f_{\tilde{\mathbf{Z}}_k | \tilde{\mathbf{Z}}_h}(\tilde{\mathbf{z}}_k) \right) \quad f_{\tilde{\mathbf{Z}}_k | \tilde{\mathbf{Z}}_h}(\tilde{\mathbf{z}}_k) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp \left(-\frac{1}{2} (\tilde{\mathbf{z}}_k - \hat{\mathbf{z}}_k)^T \Sigma_k^{-1} (\tilde{\mathbf{z}}_k - \hat{\mathbf{z}}_k) \right)$$

$$\begin{aligned} \text{NLL}_k^{(i)}(dx, dy, \Sigma) &= \frac{1}{2} \underbrace{\frac{1}{(1-\rho^2)} \left(\frac{d_x^2}{\sigma_x^2} + \frac{d_y^2}{\sigma_y^2} - 2\rho \frac{d_x d_y}{\sigma_x \sigma_y} \right)}_{(\tilde{\mathbf{z}}_k - \hat{\mathbf{z}}_k)^T \Sigma_k^{-1} (\tilde{\mathbf{z}}_k - \hat{\mathbf{z}}_k)} \\ &\quad + \underbrace{\ln \left(\sigma_x \sigma_y \sqrt{1-\rho^2} \right)}_{\ln(\sqrt{|\Sigma_k|})} + \ln(2\pi) \end{aligned}$$



$$\text{NLL}_k^{(i)} \left(\{dx, dy, \Sigma, p\}_{m \in [1, n_{\text{mix}}]} \right) = -\ln \left(\sum_{m=1}^{n_{\text{mix}}} p_m e^{-\text{NLL}_k^{(i)}(dx_m, dy_m, \Sigma_m)} \right)$$

Critères d'évaluation multi-modaux pour N séquences

- K : Nombre de prédictions
- minFDE: Moyenne des distances finales minimales parmi les K prédictions

$$\text{FDE}(t = 3s) = \frac{1}{N} \sum_{i=1}^N \sqrt{(x_t^i - \hat{x}_t^i)^2 + (y_t^i - \hat{y}_t^i)^2}$$

- minADE: Moyenne des distances minimales parmi les K prédictions

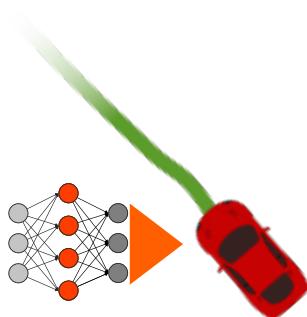
$$\text{ADE} = \frac{1}{N \times T} \sum_{t=0, dt=0.1}^3 \sum_{i=1}^N \sqrt{(x_t^i - \hat{x}_t^i)^2 + (y_t^i - \hat{y}_t^i)^2}$$

- MR: Taux de prédictions dont la distance minimale est supérieures à 2 mètres

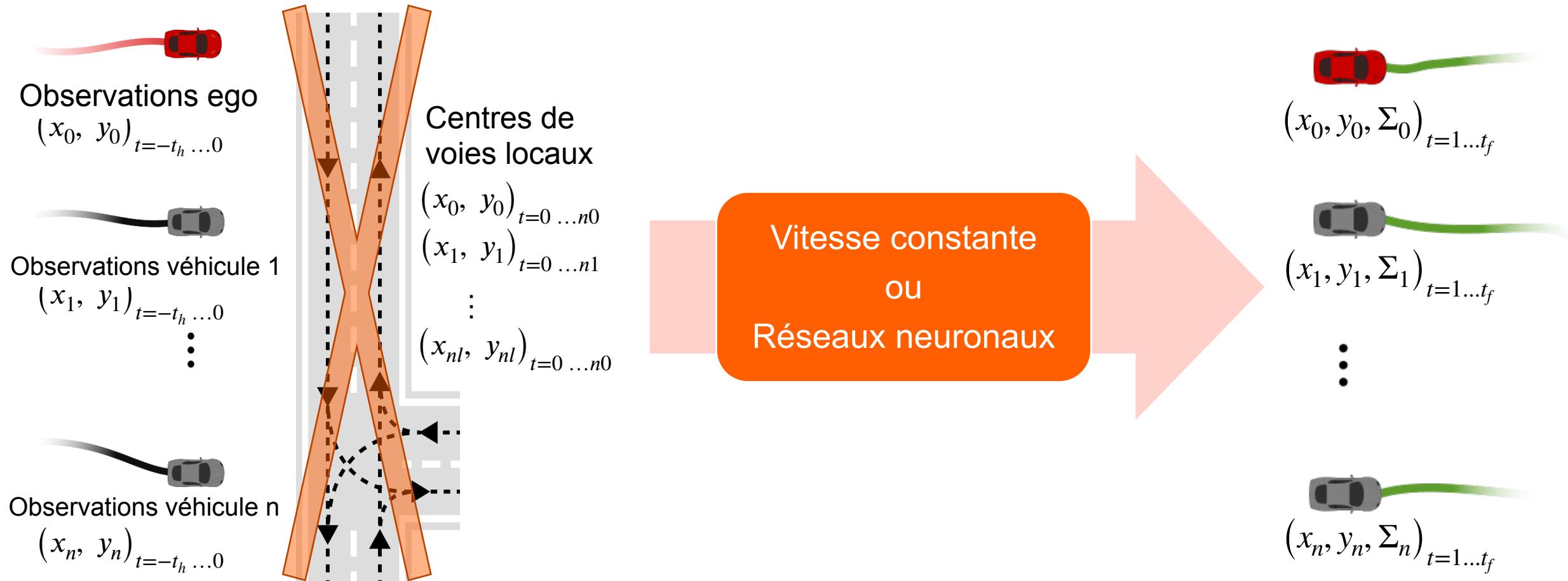
$$\text{MR}(t = 3s) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\sqrt{(x_t^i - \hat{x}_t^{*i})^2 + (y_t^i - \hat{y}_t^{*i})^2} > 2}$$

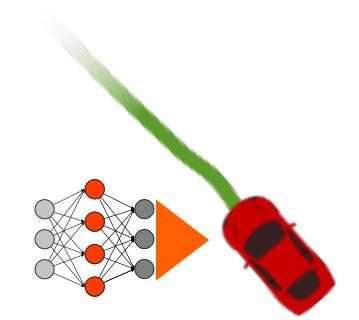
Prédiction

Premiers modèles et base de comparaison



Prédiction de trajectoire avec des réseaux neuronaux

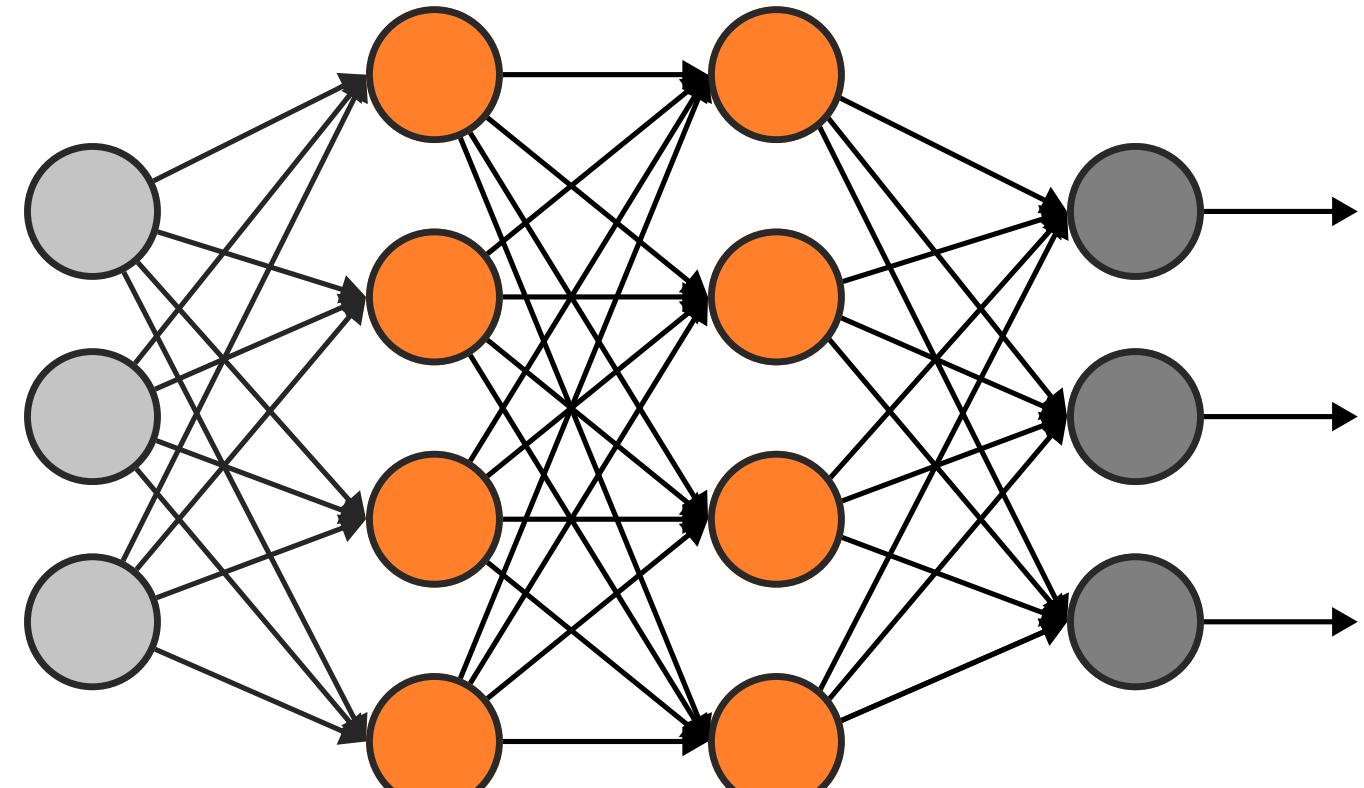
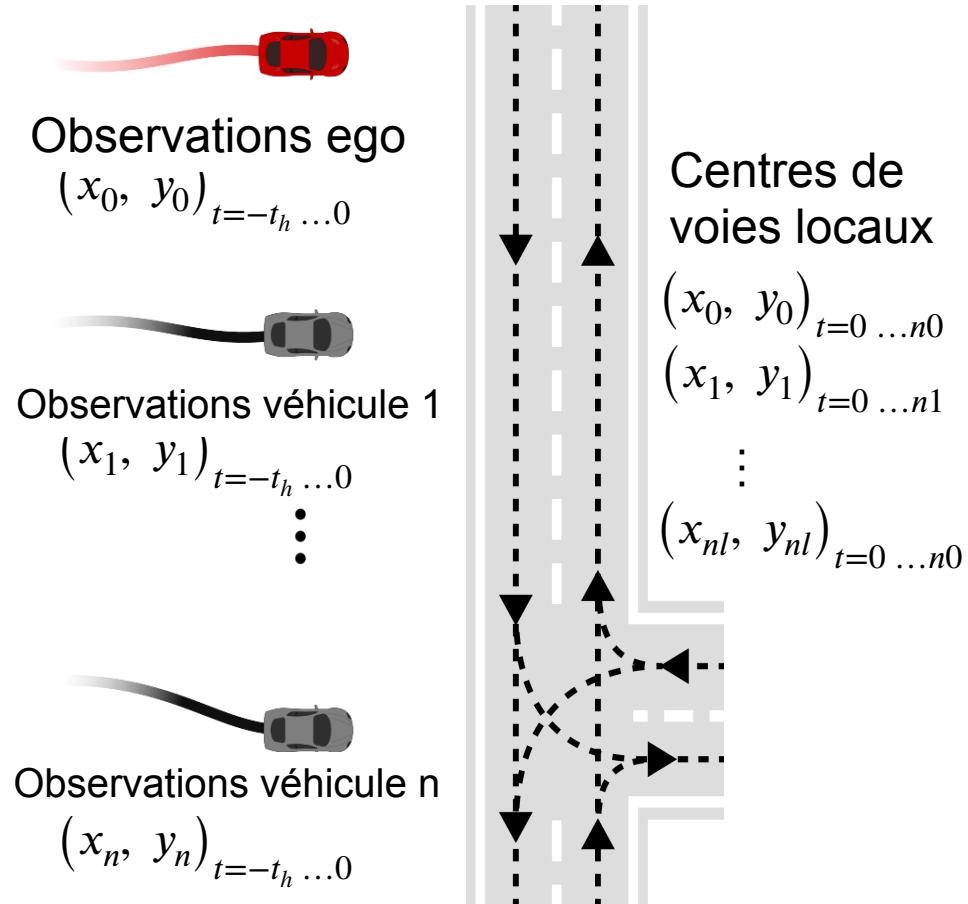




Prédiction de trajectoire avec des réseaux neuronaux

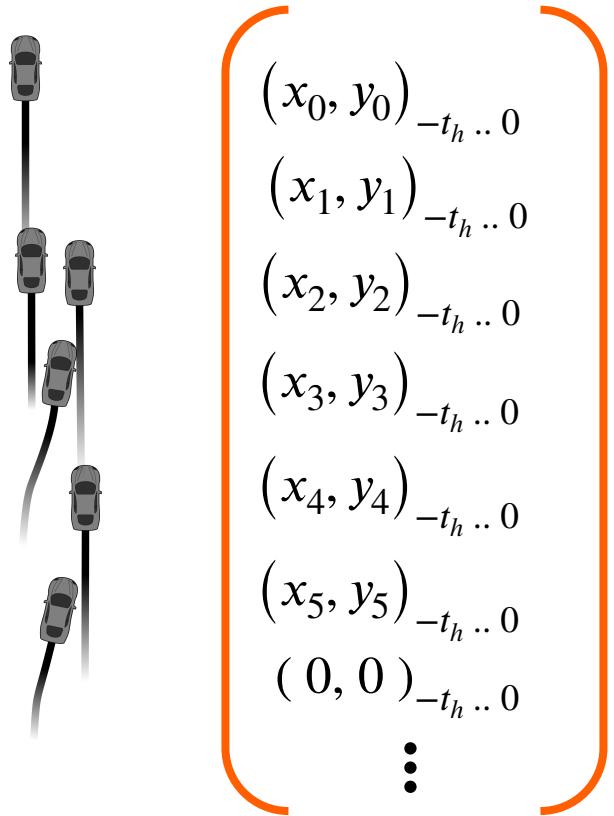
	Time horizon	1s	2s	3s	4s	5s
FDE (m)	Constant velocity	0.46	1.24	2.27	3.53	4.99
	Fully connected [size8]	0.46	1.24	2.27	3.52	4.97
	Fully connected [size16]	0.47	1.24	2.26	3.51	4.95
	Convolutional [size8]	0.47	1.25	2.29	3.54	5.00
	Convolutional [size16]	0.46	1.23	2.25	3.50	4.96
	Recurrent [size8]	0.48	1.26	2.30	3.57	5.07
	Recurrent [size16]	0.47	1.25	2.28	3.55	5.05
NLL	Constant velocity	0.81	2.31	3.22	3.91	4.46
	Fully connected [size8]	0.40	2.04	3.01	3.71	4.27
	Fully connected [size16]	0.37	2.02	2.98	3.68	4.24
	Convolutional [size8]	0.38	2.04	3.01	3.71	4.27
	Convolutional [size16]	0.32	2.00	2.97	3.67	4.22
	Recurrent [size8]	0.48	2.09	3.05	3.74	4.30
	Recurrent [size16]	0.33	2.01	2.99	3.70	4.26
MR	Constant velocity	0.02	0.20	0.44	0.61	0.71
	Fully connected [size8]	0.02	0.19	0.43	0.62	0.75
	Fully connected [size16]	0.02	0.19	0.44	0.62	0.74
	Convolutional [size8]	0.02	0.20	0.44	0.63	0.74
	Convolutional [size16]	0.02	0.19	0.43	0.62	0.74
	Recurrent [size8]	0.02	0.20	0.45	0.63	0.75
	Recurrent [size16]	0.02	0.19	0.44	0.64	0.75

Prédiction de trajectoire avec des réseaux neuronaux

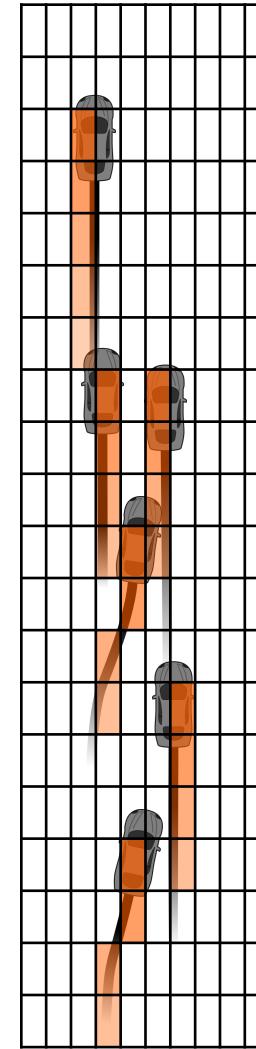


Vecteur d'entrée de taille constante

Liste de features



- Nombre maximum d'entrées fixé complété par des 0
- Comment trier la liste
 - Distance
 - Grille

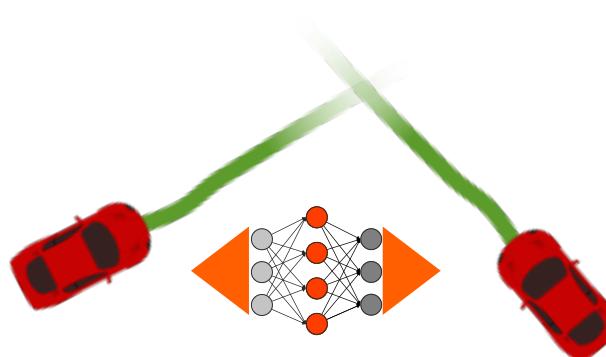


Carte quadrillée

- Grille de pixels de taille fixe
- Pas de tri des entrées
- Perte de précision avec la discréttisation
- Représentation avec une faible densité d'information

Réseaux neuronaux pour les scènes routières complexes

Prendre en compte les interactions entre véhicules sans fixer d'ordre de tri ni de nombre maximum de véhicules

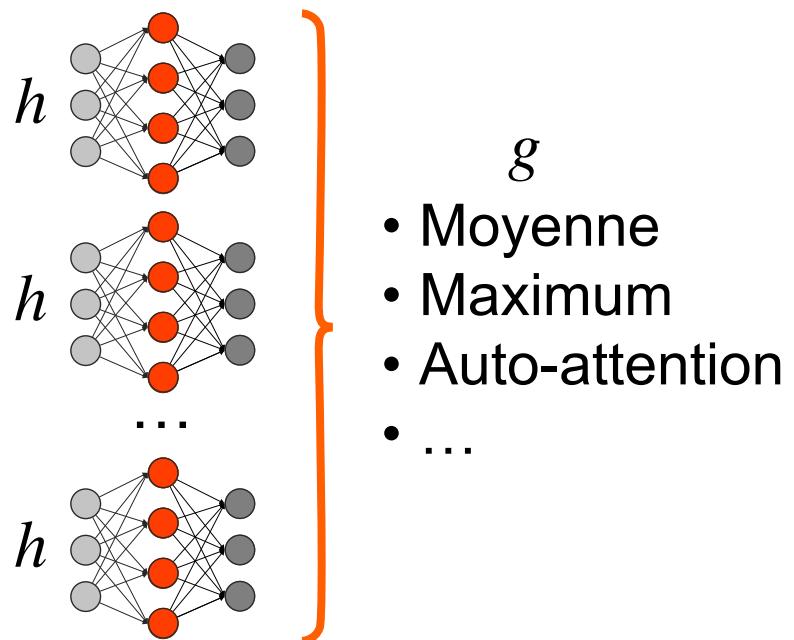


Interactions entre un nombre dynamique de véhicules

$$f\left((x_0, y_0)_{t=-t_h \dots 0}, \dots, (x_n, y_n)_{t=-t_h \dots 0}\right) = g\left(h\left((x_0, y_0)_{t=-t_h \dots 0}\right), \dots, h\left((x_n, y_n)_{t=-t_h \dots 0}\right)\right)$$

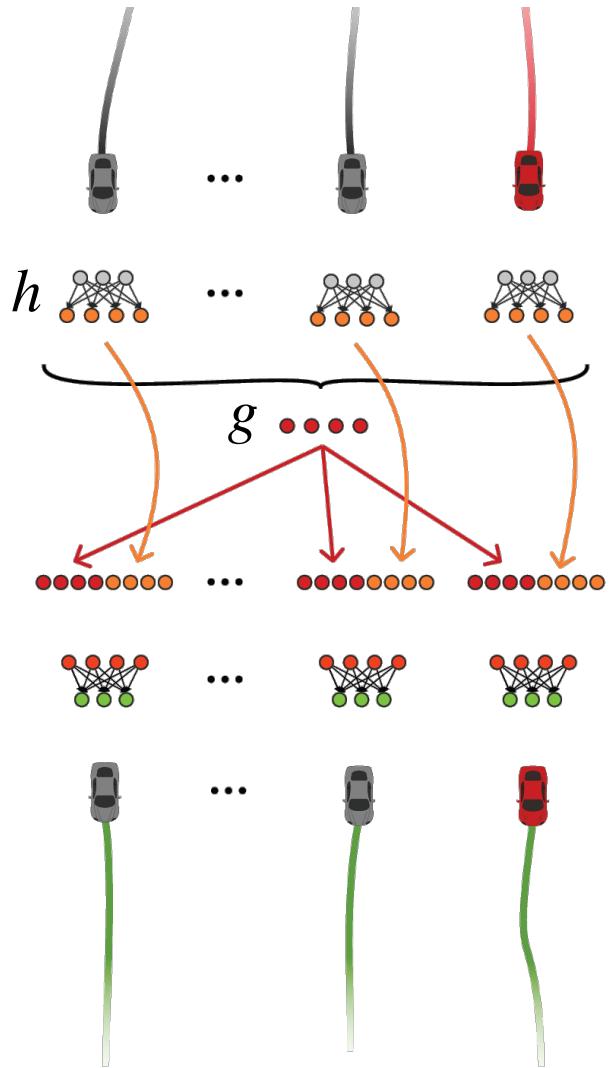
Avec $f : \mathbb{R}^{2 \times t_h \times n} \rightarrow \mathbb{R}^{2 \times t_h \times n}$, $h : \mathbb{R}^{2 \times t_h} \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{2 \times t_f \times n}$
 g est équivariante pour l'ordre et définie pour tout nombre de véhicules n .

Même calcul pour chaque entrée



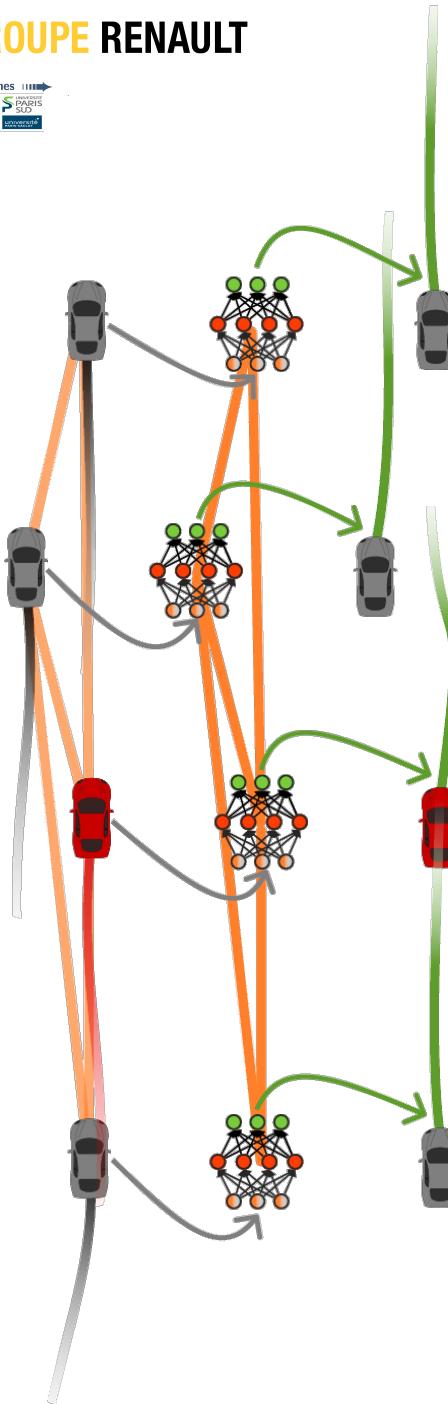
Fonction à nombre d'entrées dynamique et équivariante pour l'ordre

Architectures adaptées aux interactions



PointNet

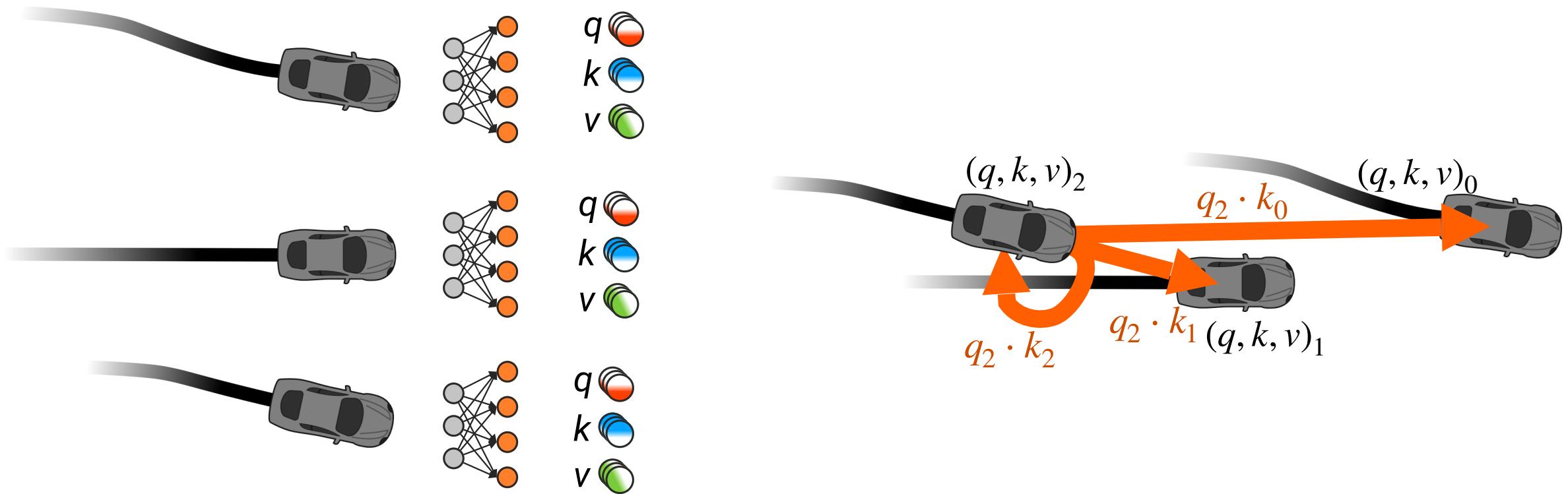
- Mécanisme d'agrégation globale
- Même contexte pour tous:
 - Grande dimension
 - Faible spécificité



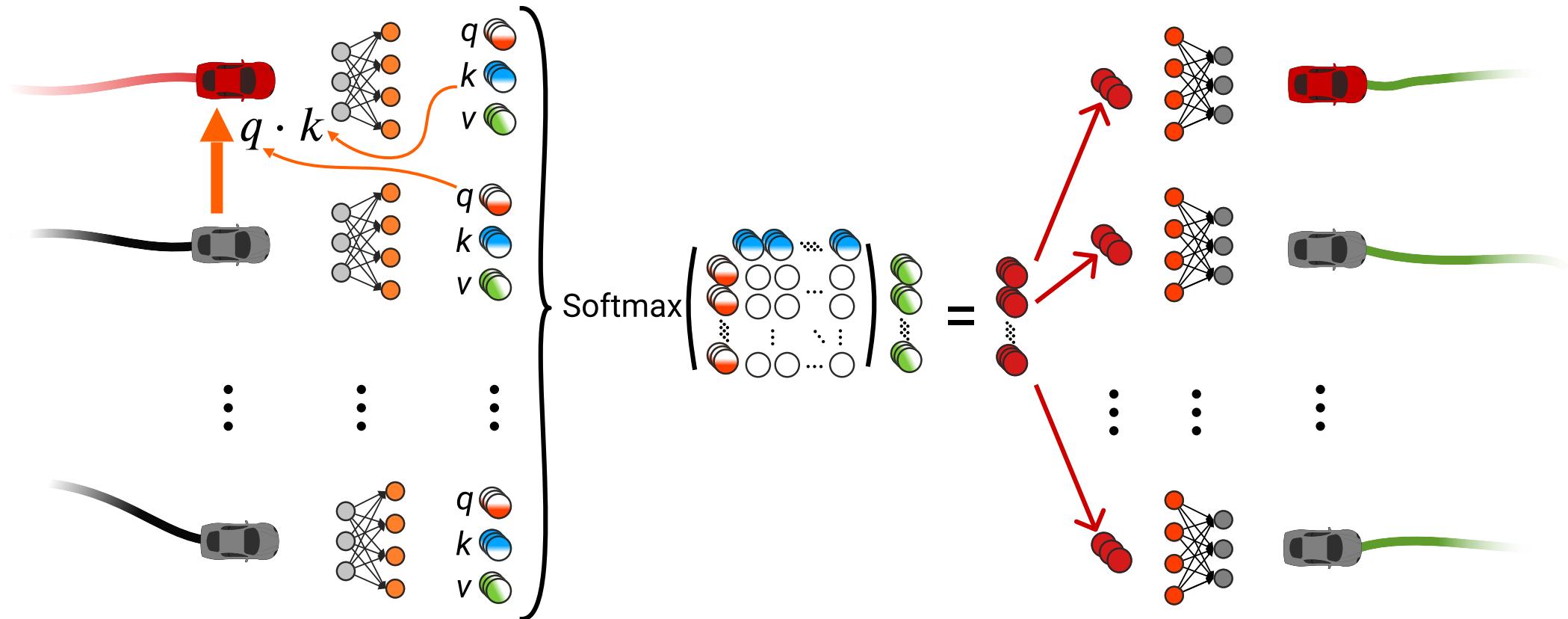
Réseaux neuronaux de graphes

- Définition du graphe à fournir
- Difficile à implémenter efficacement

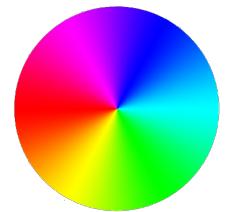
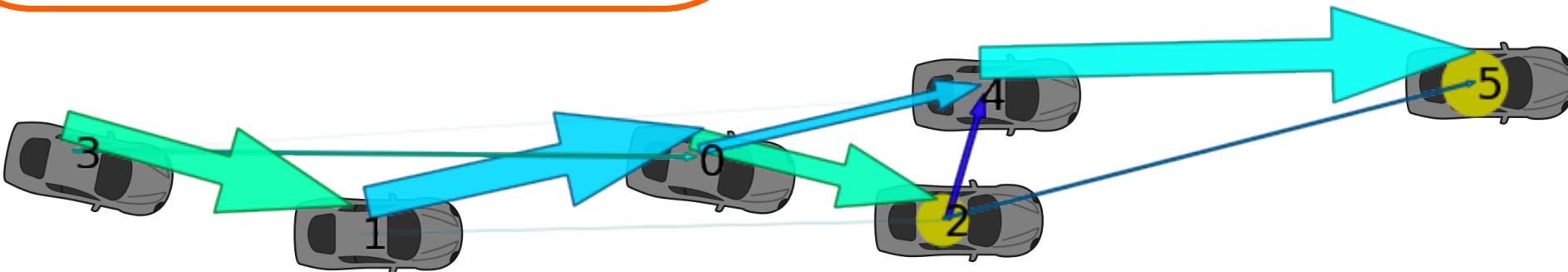
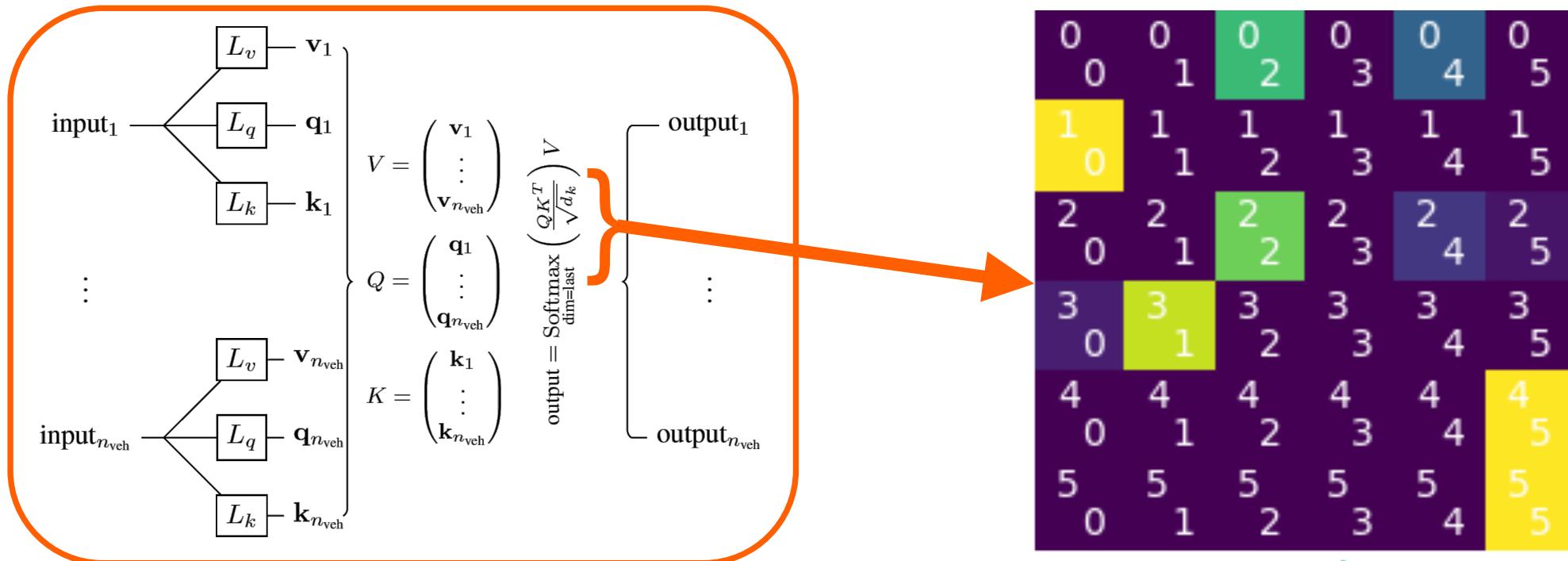
Le mécanisme d'auto-attention définit un graphe



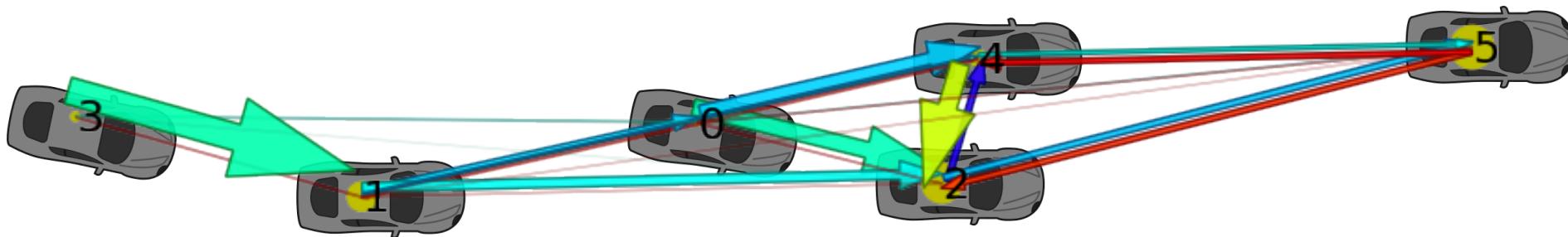
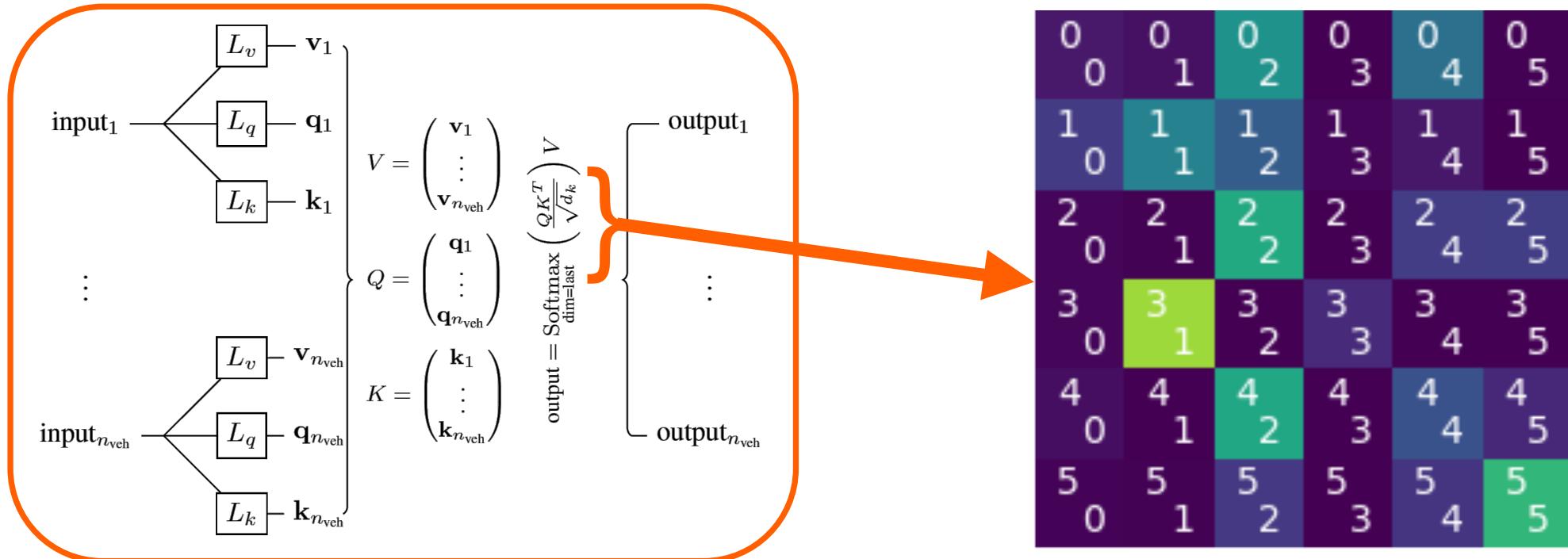
Mécanisme d'auto-attention représentant les interactions



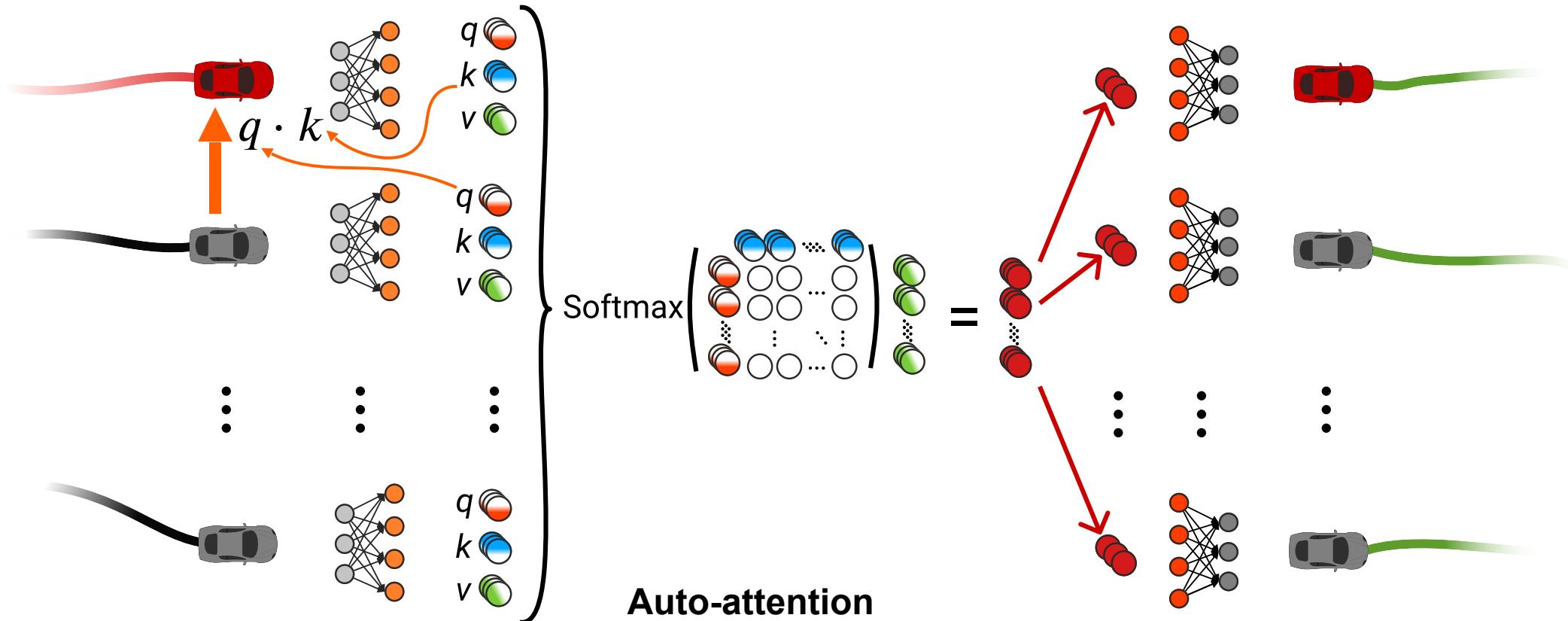
L'auto-attention forme un graphe complet orienté



L'auto-attention forme un graphe complet orienté

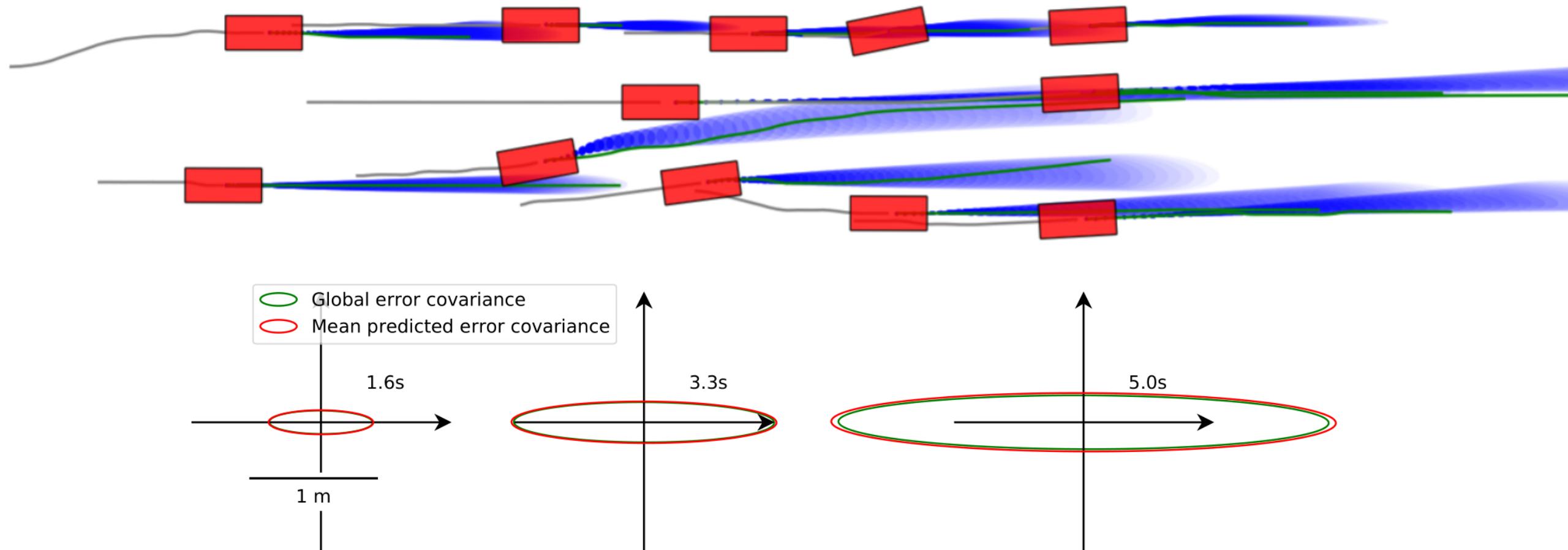


Mécanisme d'auto-attention représentant les interactions

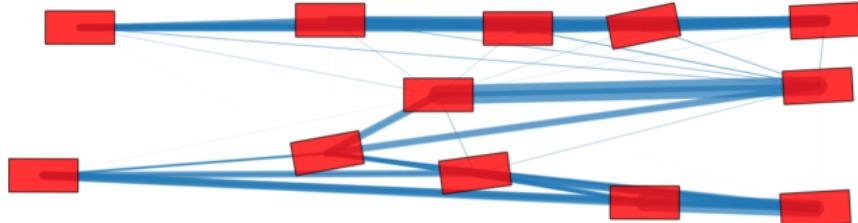


- Complexité de $O(n^2)$ pour n véhicules
- Apprend des corrélations mais pas de «causalité»

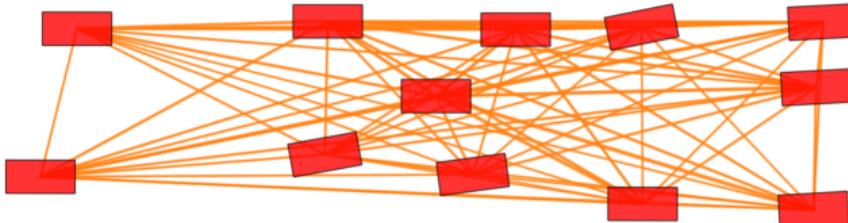
Résultats de validation d'expérience



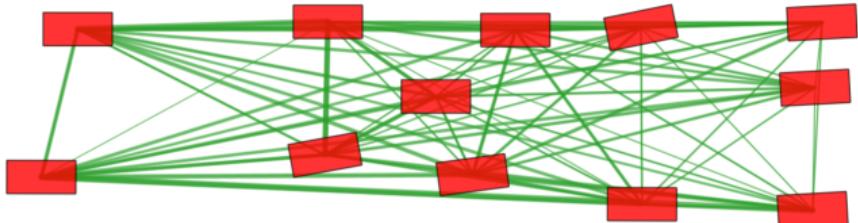
Représentation des graphes d'attention



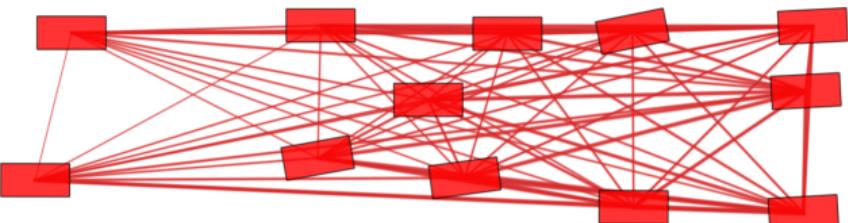
(a) Graphe de la première tête



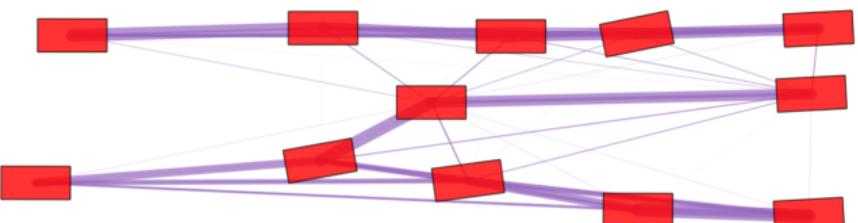
(b) Graphe de la deuxième tête



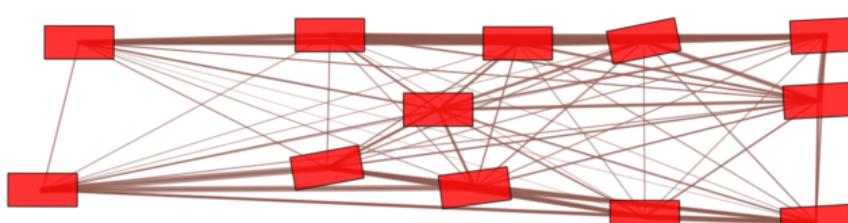
(c) Graphe de la troisième tête



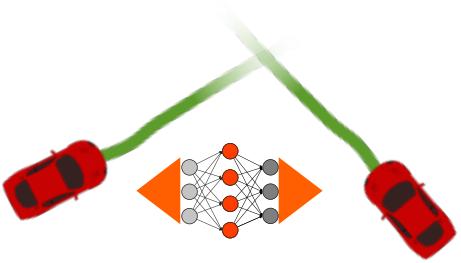
(d) Graphe de la quatrième tête



(e) Graphe de la cinquième tête

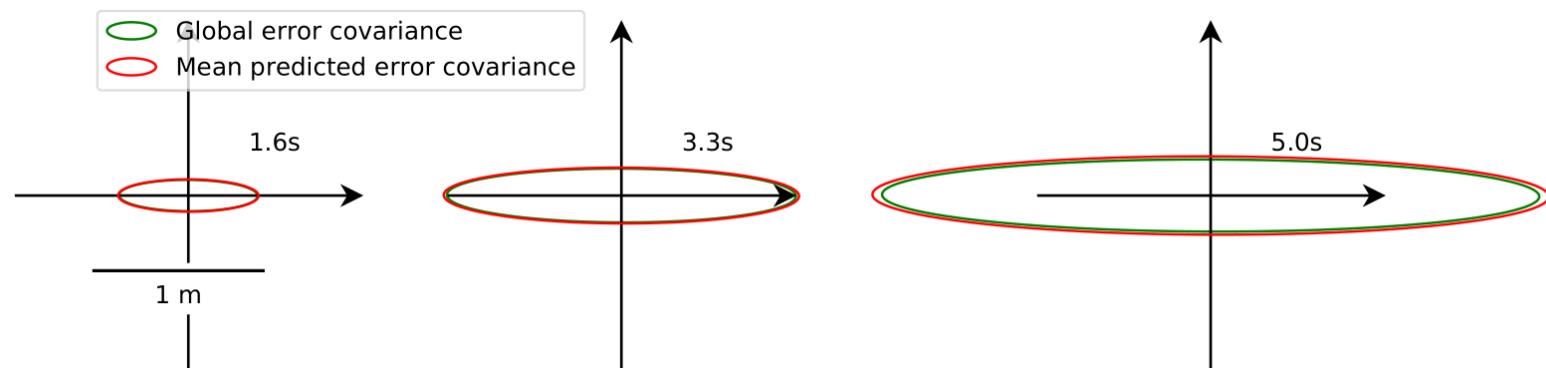


(f) Graphe de la sixième tête

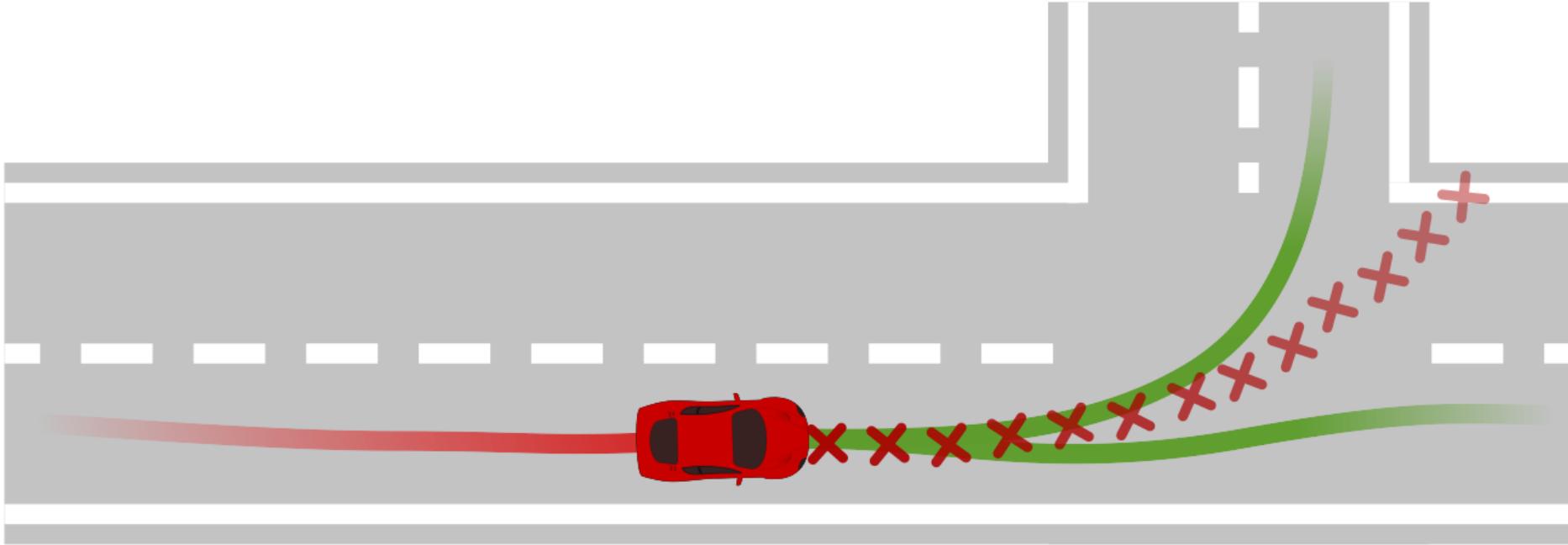


Résultats avec auto-attention

Time horizon		1s	2s	3s	4s	5s
FDE (m)	Constant velocity	0.46	1.24	2.27	3.53	4.99
	List	0.45	1.12	1.97	3.04	4.32
	Self-attention	0.31	0.78	1.36	2.07	2.95
NLL	Constant velocity	0.81	2.31	3.22	3.91	4.46
	List	0.24	1.67	2.55	3.19	3.71
	Self-attention	-0.57	0.99	1.90	2.56	3.09
MR	Constant velocity	0.02	0.20	0.44	0.61	0.71
	List	0.01	0.15	0.37	0.56	0.68
	Self-attention	0.01	0.06	0.22	0.39	0.54

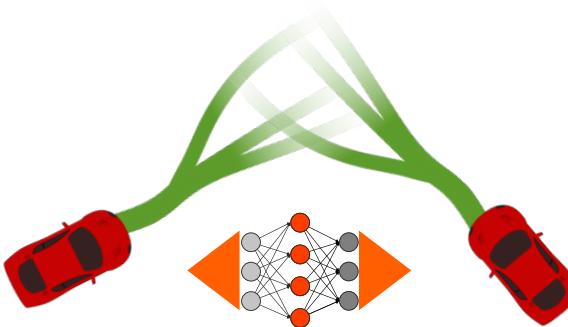


Plusieurs futurs sont possibles

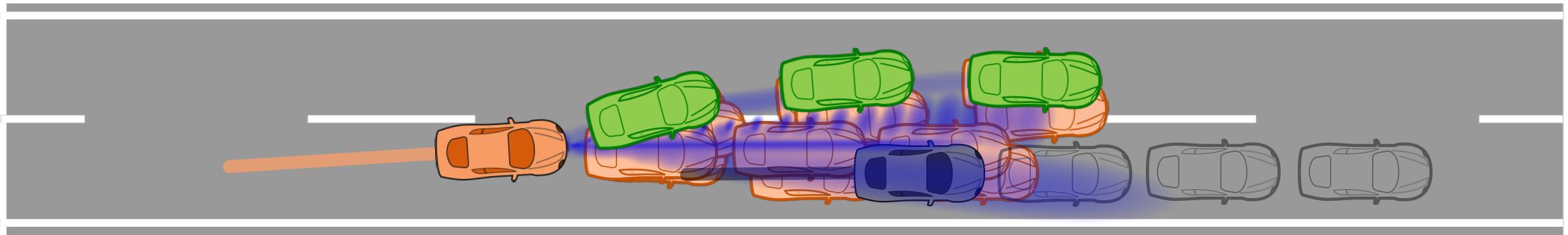


Prédire différents futurs possibles

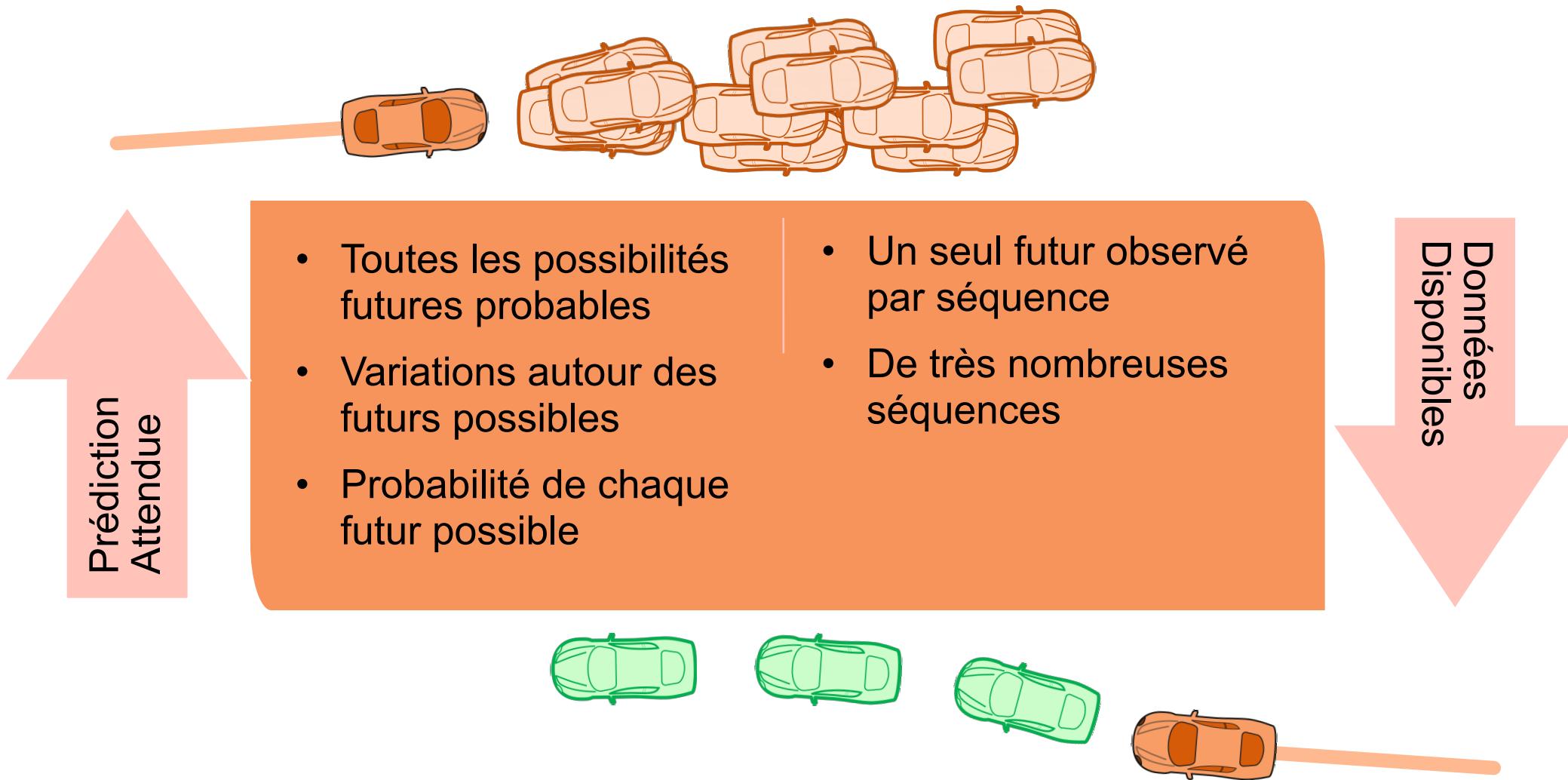
Il y a plusieurs futurs potentiels mais un seul sera observé



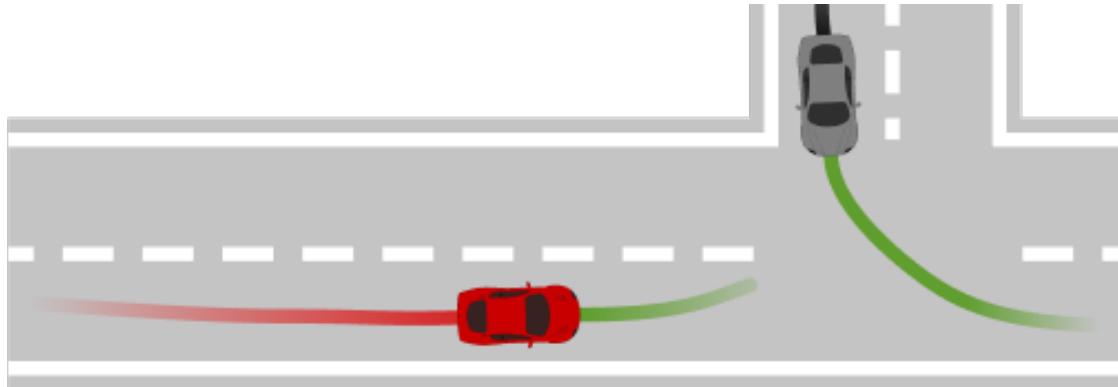
Un seul futur observable



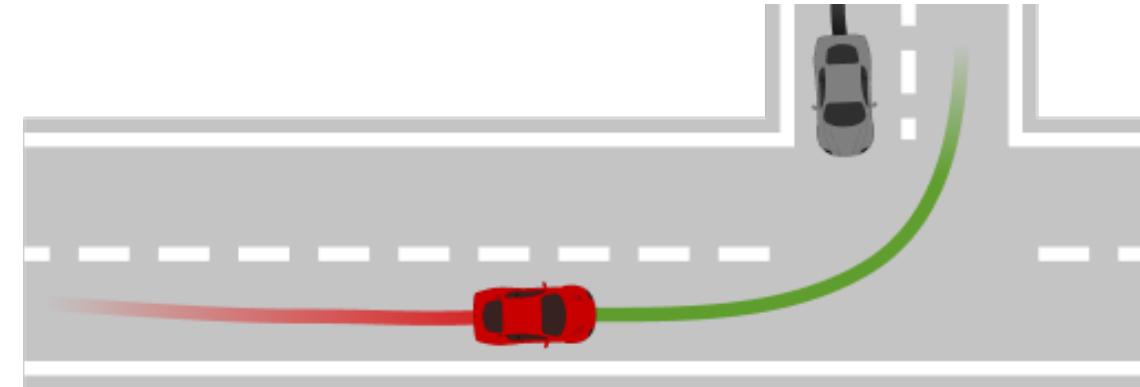
Le résultat multi-modal attendu est différent de l'observation



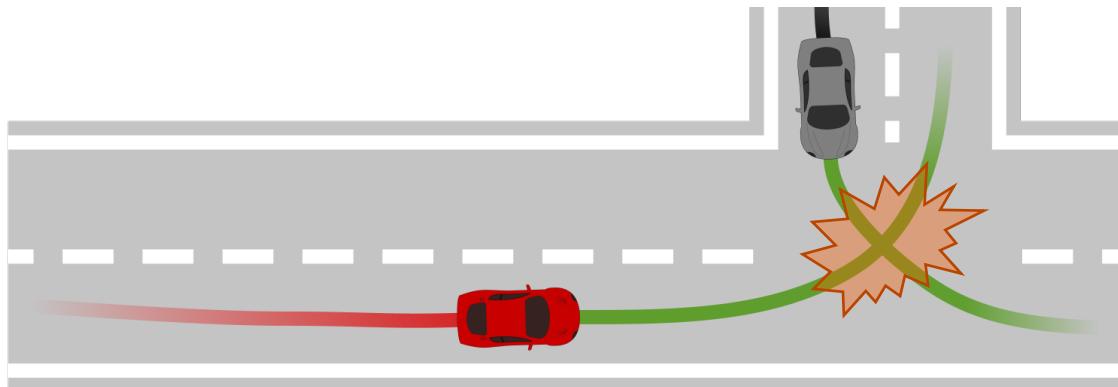
Limites des prédictions multi-modales et multi-agents



Cas 1: Le véhicule gris tourne à gauche et le véhicule rouge lui cède le passage



Cas 2: Le véhicule gris cède le passage et le véhicule rouge tourne à gauche

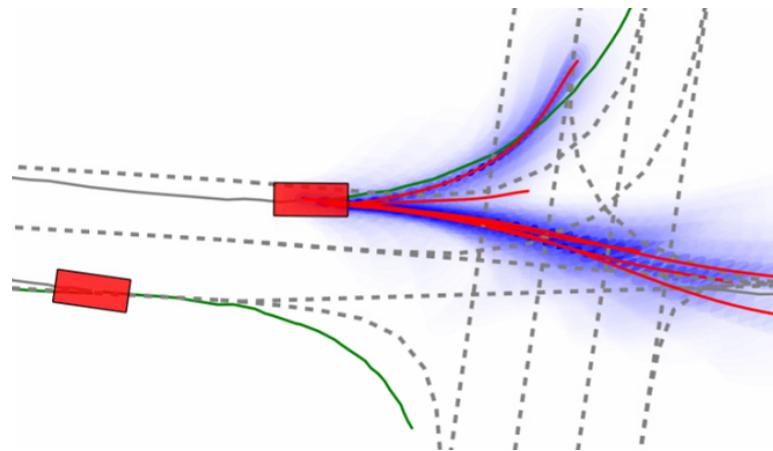


Cas peu probable : Les deux véhicules tournent à gauche en même temps...

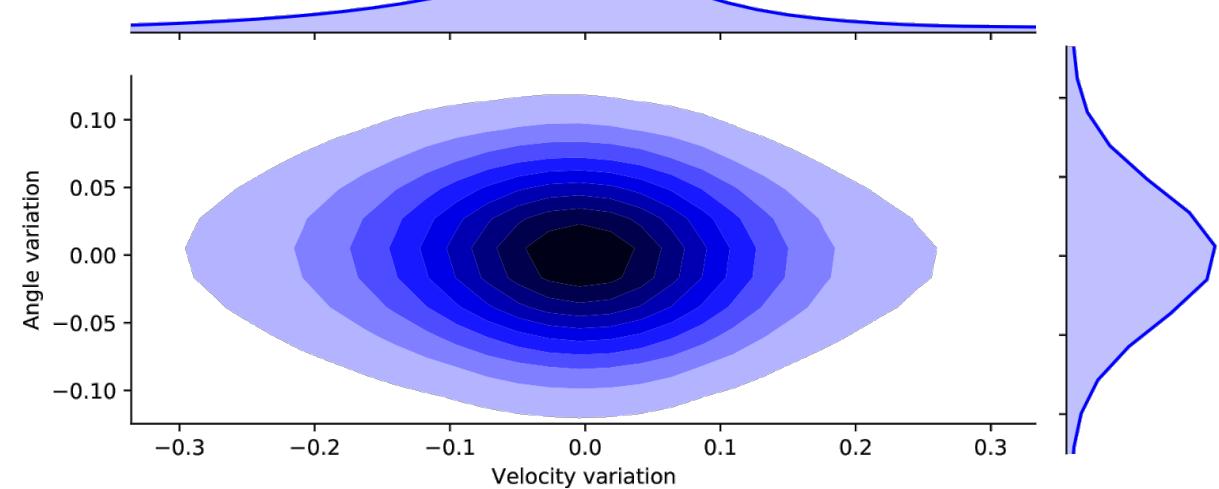
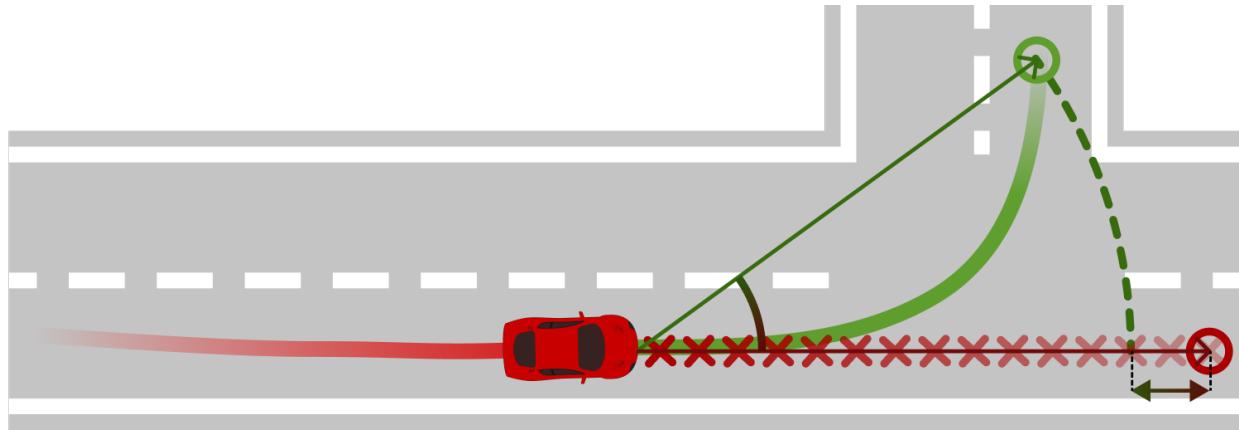
Modes pour chaque objet	Modes pour la scène
<ul style="list-style-type: none"> • Cas croisés peu probables • Pas de négociation 	<ul style="list-style-type: none"> • Seulement des situations probables • Négociations
<ul style="list-style-type: none"> • Peu de modes • Facile à exprimer • Facile à utiliser 	<ul style="list-style-type: none"> • Nombreux modes • Difficile à exprimer • Difficile à utiliser • Difficile à valider

Application de prédiction multi-modale

Définition d'une base de comparaison et d'un modèle complet



Modèle multi-modal à vitesse constante



$$\sin(\alpha) \approx \alpha \Rightarrow dy \approx v_{\text{true}} \times t \times (\alpha - \alpha_{\text{true}})$$
$$dv = \frac{v}{v_{\text{true}}} - 1 \Rightarrow dx = v_{\text{true}} \times t \times dv$$

Résultats comparés du modèle multi-modal à vitesse constante



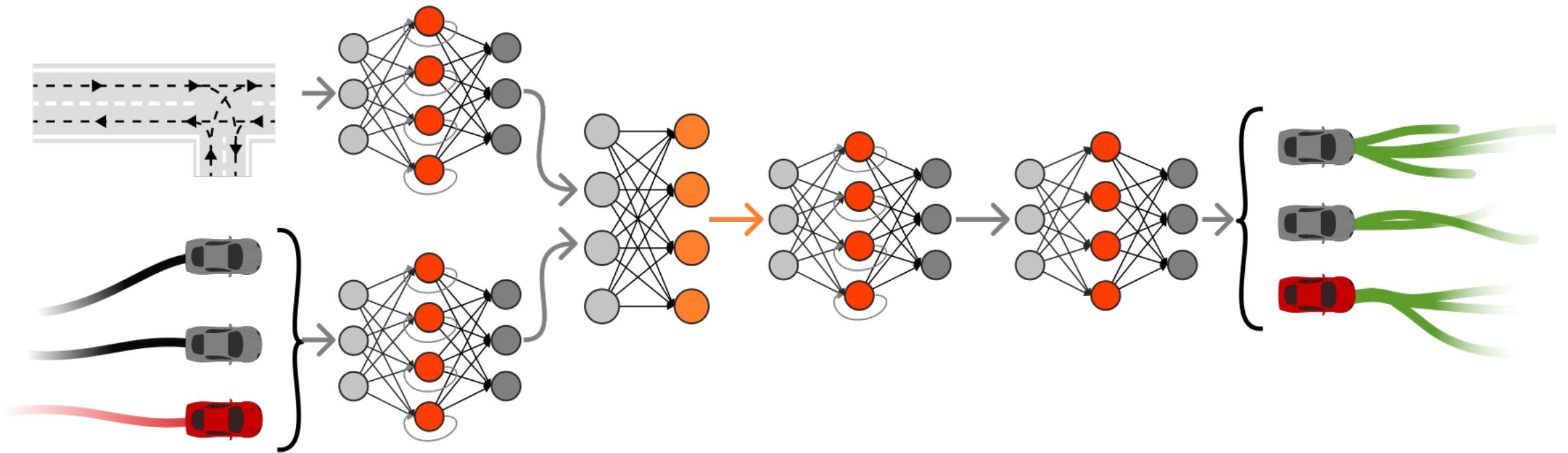
Résultats des modèles **uni-modaux**
avec différentes dynamiques

Time horizon		1s	2s	3s	4s	5s
FDE (m)	Constant velocity	0.46	1.24	2.27	3.53	4.99
	List	0.45	1.12	1.97	3.04	4.32
	Self-attention	0.31	0.78	1.36	2.07	2.95
NLL	Constant velocity	0.81	2.31	3.22	3.91	4.46
	List	0.24	1.67	2.55	3.19	3.71
	Self-attention	-0.57	0.99	1.90	2.56	3.09
MR	Constant velocity	0.02	0.20	0.44	0.61	0.71
	List	0.01	0.15	0.37	0.56	0.68
	Self-attention	0.01	0.06	0.22	0.39	0.54

Résultats du modèle **multi-modal**
à vitesse constante (K=6)

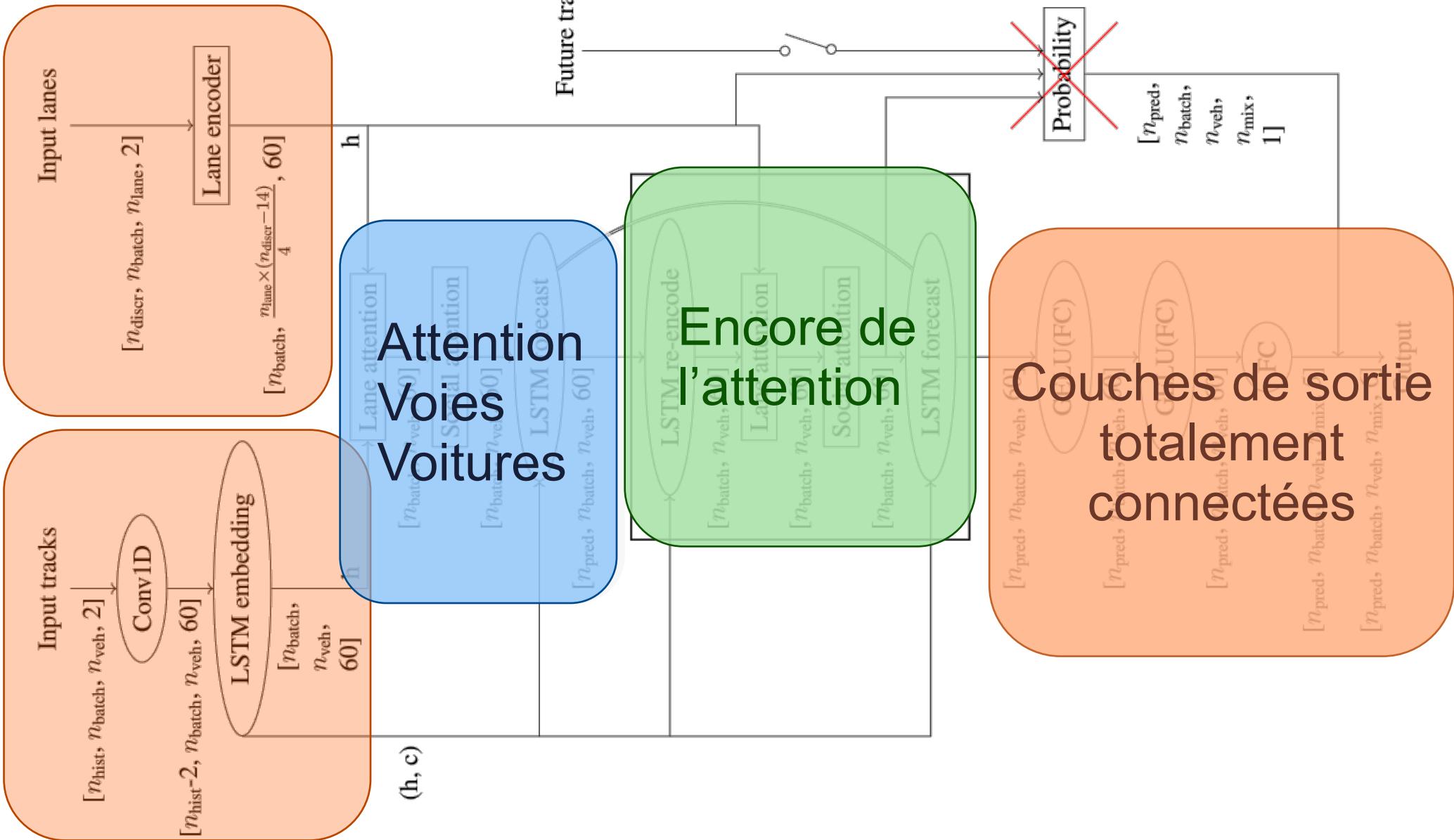
Time horizon	1s	2s	3s	4s	5s
pFDE (m)	0.47	1.24	2.27	3.53	5.00
FDE (m)	1.40	2.87	4.43	6.11	7.89
minFDE (m)	0.64	1.08	1.42	1.74	2.21
NLL	2.21	3.08	3.89	4.59	5.22
MR	0.02	0.13	0.21	0.25	0.29

Modèle complet de prédiction multi-modale

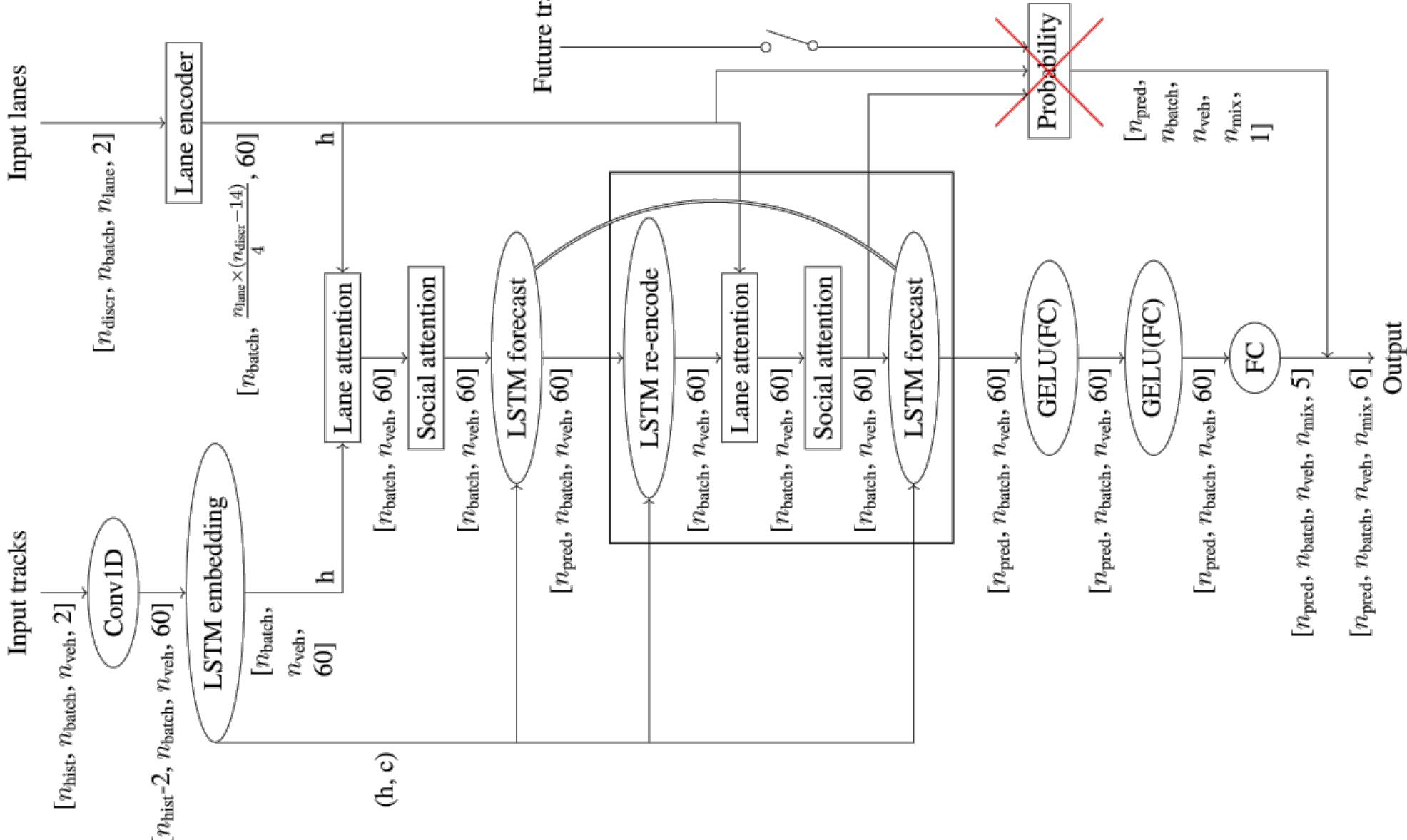


Données d'entrée à encoder

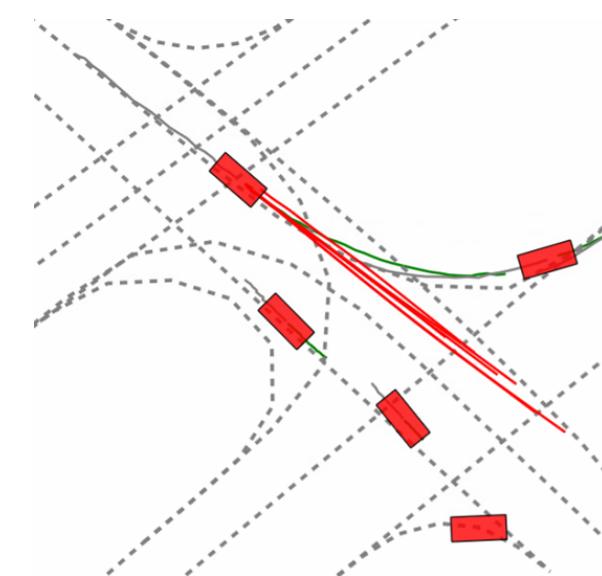
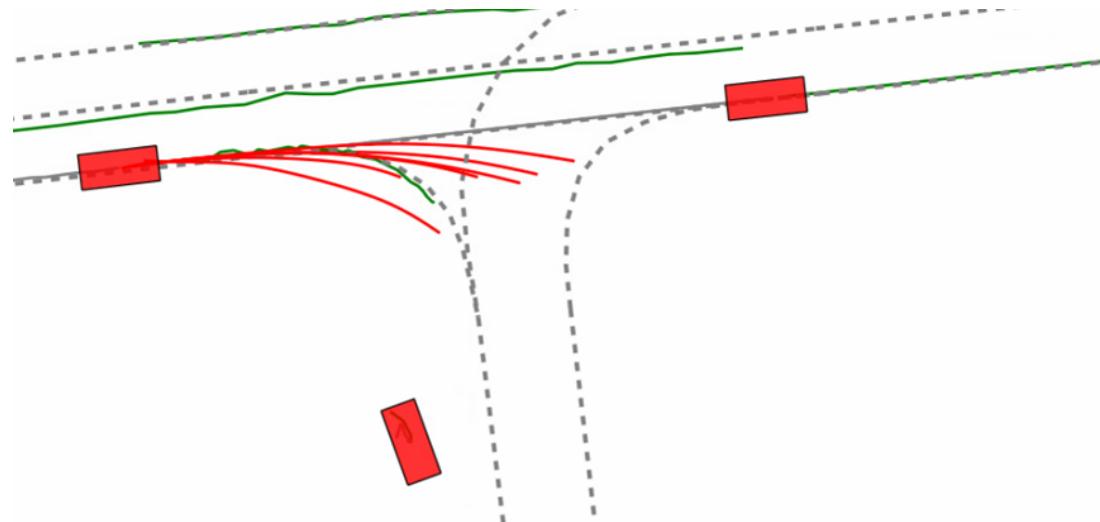
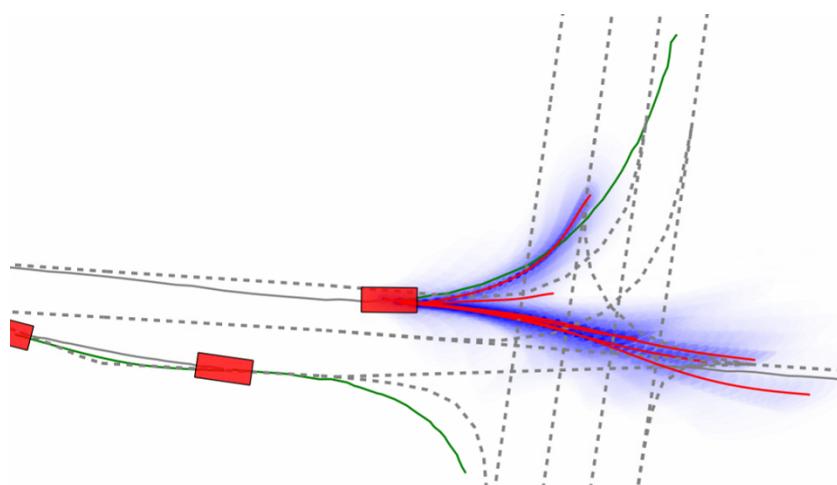
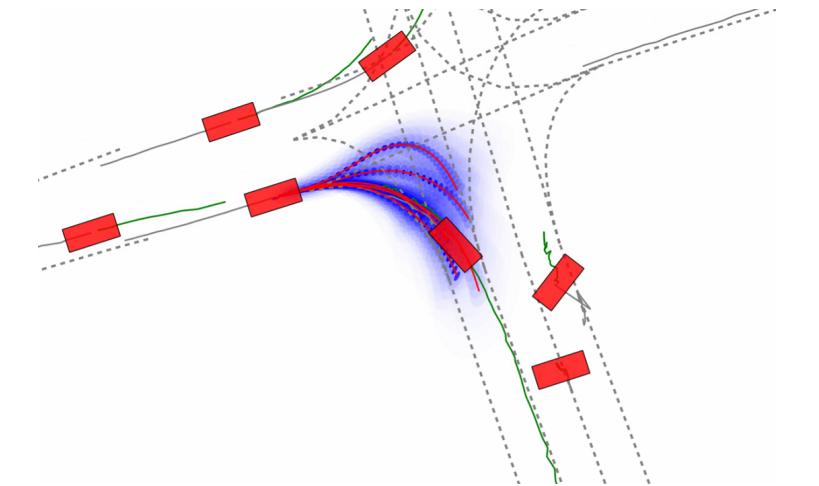
Modèle utilisé



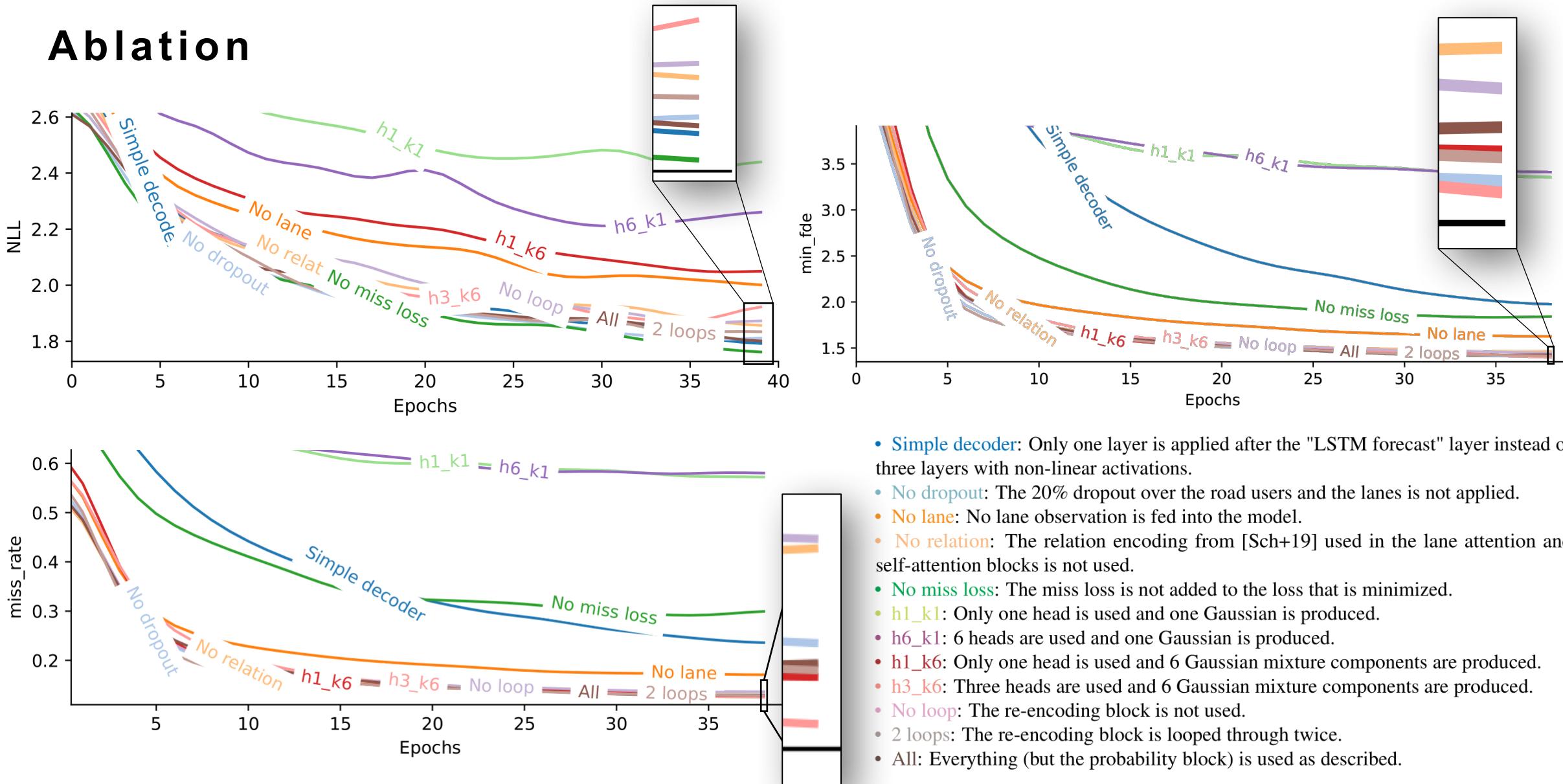
Modèle utilisé



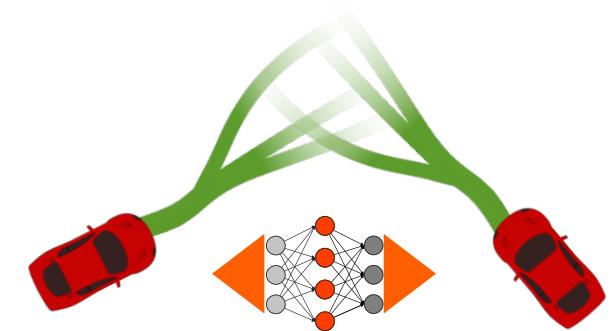
Quelques résultats



Ablation



- **Simple decoder:** Only one layer is applied after the "LSTM forecast" layer instead of three layers with non-linear activations.
- **No dropout:** The 20% dropout over the road users and the lanes is not applied.
- **No lane:** No lane observation is fed into the model.
- **No relation:** The relation encoding from [Sch+19] used in the lane attention and self-attention blocks is not used.
- **No miss loss:** The miss loss is not added to the loss that is minimized.
- **h1_k1:** Only one head is used and one Gaussian is produced.
- **h6_k1:** 6 heads are used and one Gaussian is produced.
- **h1_k6:** Only one head is used and 6 Gaussian mixture components are produced.
- **h3_k6:** Three heads are used and 6 Gaussian mixture components are produced.
- **No loop:** The re-encoding block is not used.
- **2 loops:** The re-encoding block is looped through twice.
- **All:** Everything (but the probability block) is used as described.



Résultats globaux sur la base de test du concours Argoverse

Method	ADE	minADE	FDE	minFDE	MR
Our model	1.68	0.97	3.73	1.42	0.13
Waymo Poly [Cha+19a]	1.71	0.89	3.85	1.50	0.13
Waymo TNT [Zha+20]	1.78	0.94	3.91	1.54	0.13
Alibaba	1.97	0.92	4.35	1.48	0.16
Uber ATG-LaneGCN [Lia+20]	1.71	0.87	3.78	1.36	0.16
Argo CMU Wimp [Kha+20]	1.82	0.90	4.03	1.42	0.17



Conclusion

Mes prédictions de trajectoire pour la prédiction de trajectoires

Développements futurs

Futur proche

- Types d'agents (piétons, 2-roues, camions...)
 - Mécanisme d'attentions entre agents hétérogènes
 - Simplement des encodeurs spécifiques ?
- Types d'objets (passages piétons, feux...)
 - Carte avec objets dynamiques
 - Simplement des encodeurs spécifiques ?
- Modes rares
 - Pénalisation ?
 - Trajectoires de référence ?
 - Apprentissage biaisé ?

Futur distant

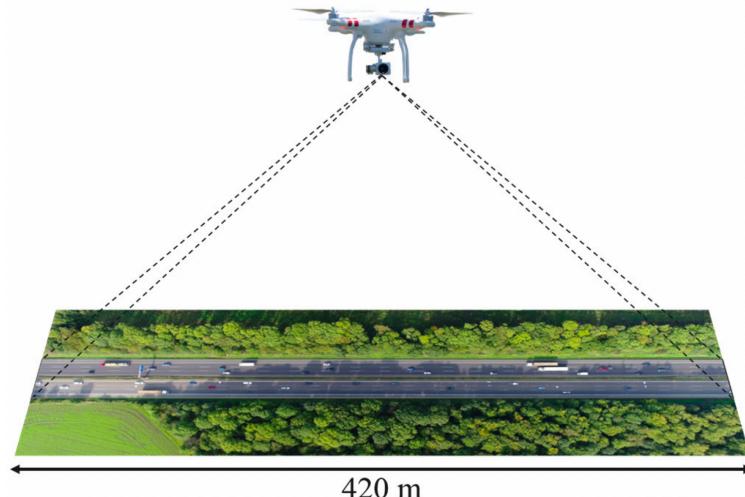
- Métriques satisfaisantes
 - La prédiction pour la prédiction n'améliore pas forcément le système global
 - Une métrique fondée sur l'espace libre pourrait avoir du sens
- Modèles invariants
 - Orientation globale de la scène
 - Localité des modèles
- Causalité et négociations
 - Trouver une interface entre planification et prédiction pour des requêtes spécifiques

Annexes

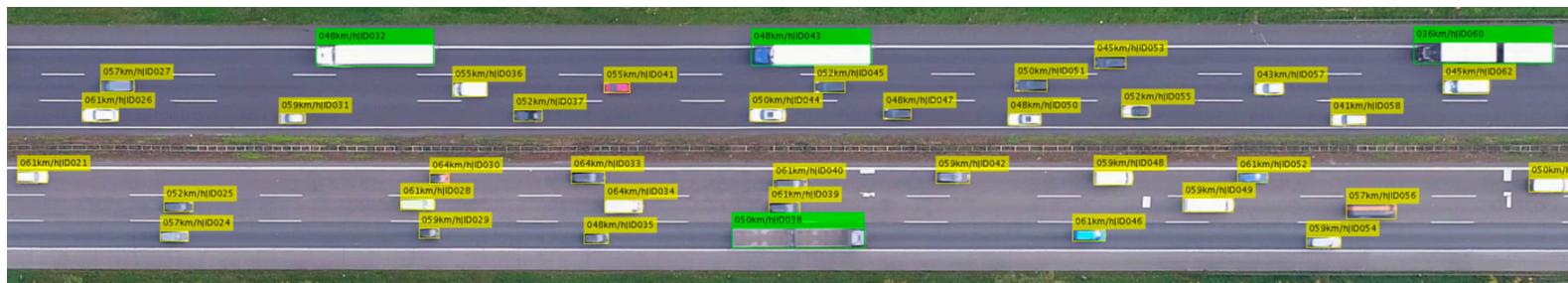


NGSIM et HighD

Base de données NGSIM et HighD



Attribute	Dataset	
	<i>NGSIM</i>	<i>highD</i>
Recording Duration [hours]	1.5	16.5
Lanes (per direction)	5-6	2-3
Recorded Distance [m]	500-640	400-420
Vehicles	9206	110 000
Cars	8860	90 000
Trucks	278	20 000
Driven distance [km]	5071	45 000
Driven time [h]	174	447

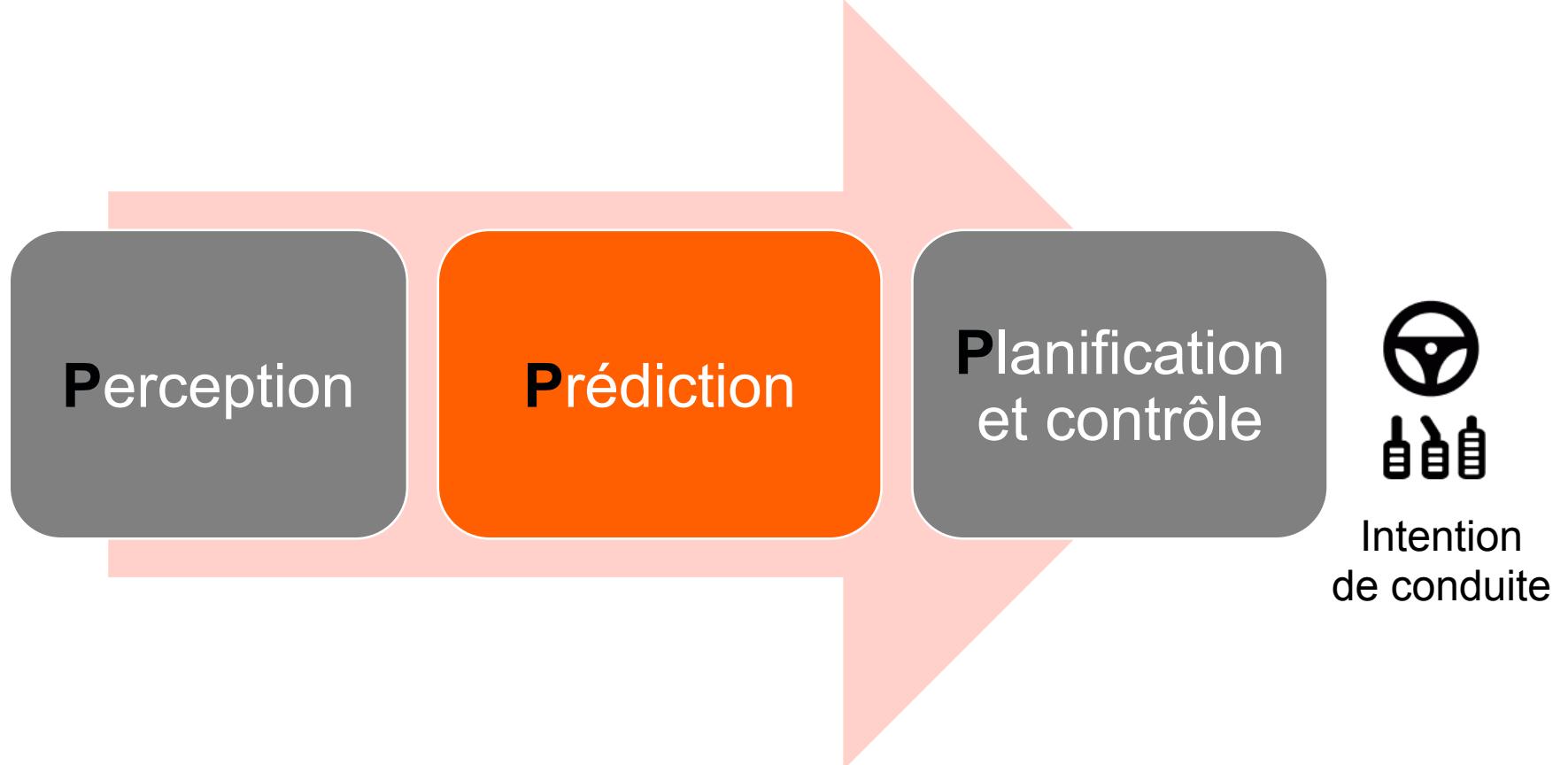
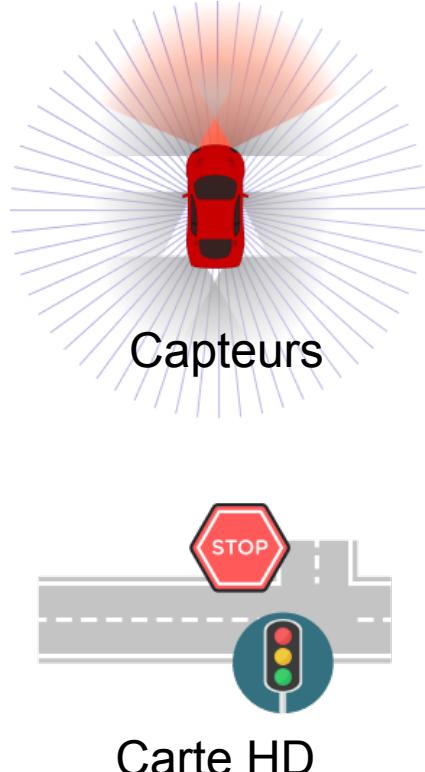




Prédiction et approche modulaire

Pourquoi vouloir prédire les mouvements des scènes routières ?

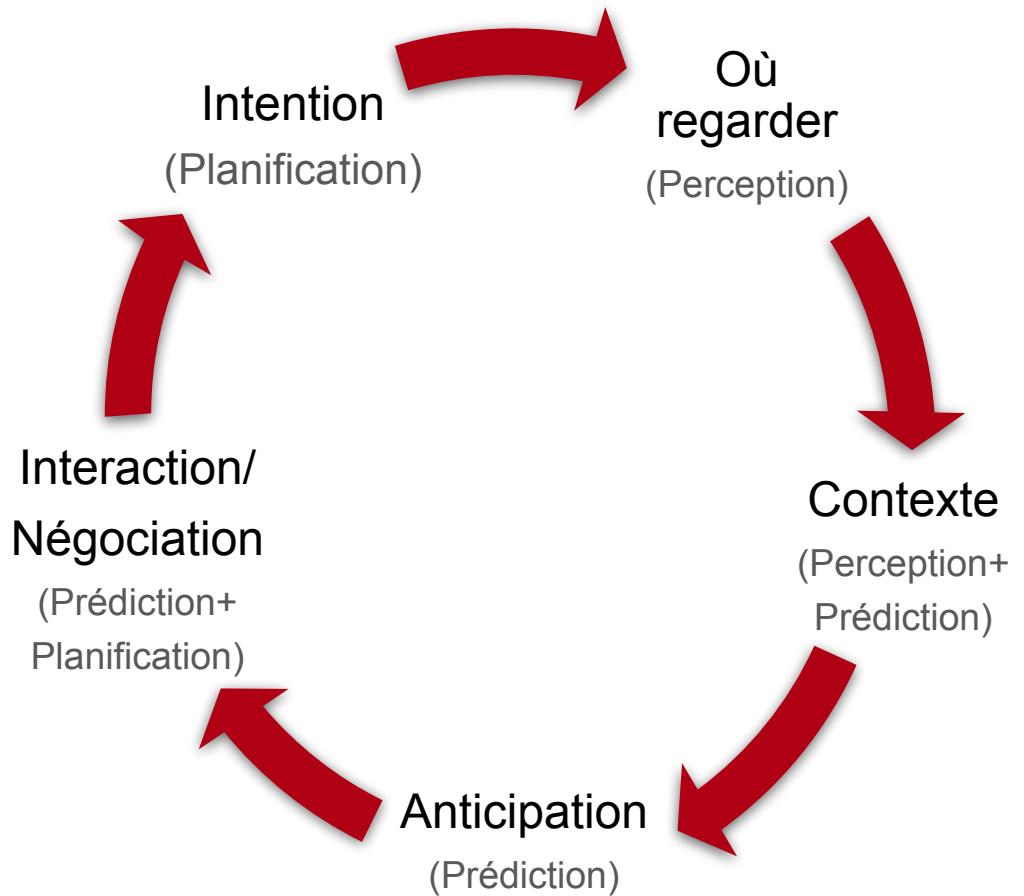
La modularité PPP



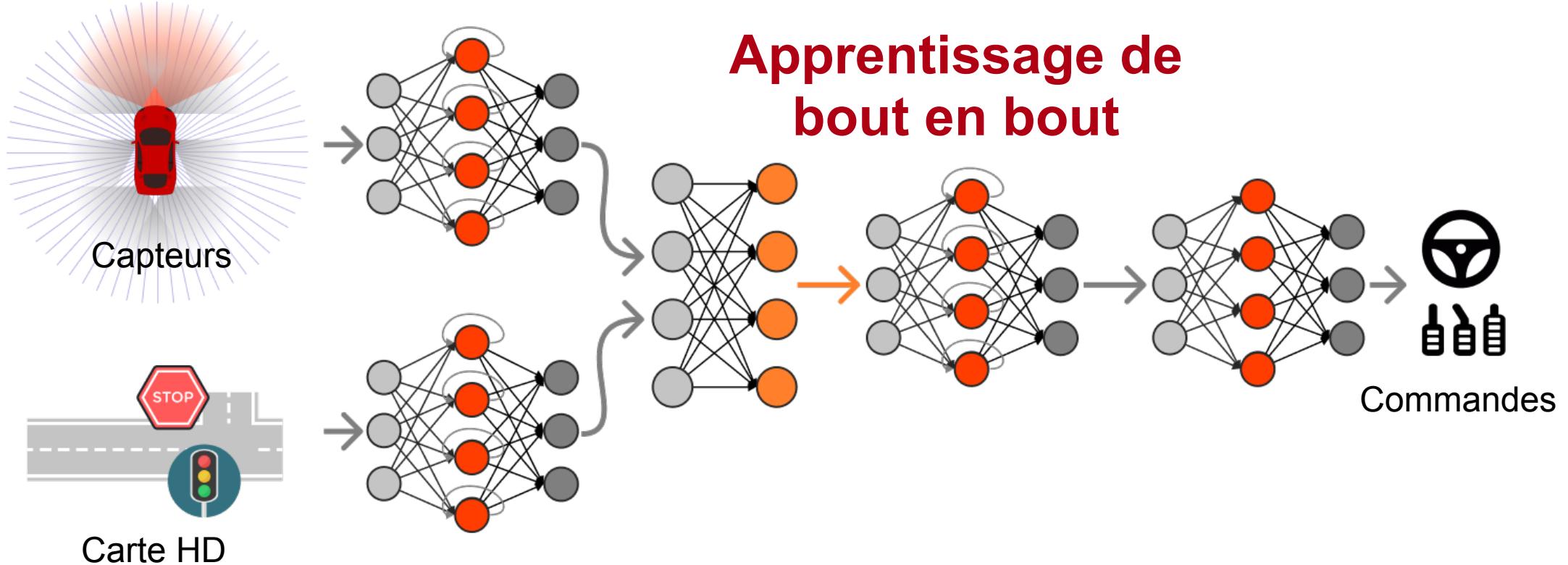
Les modules (prédiction incluse) sont inutiles

Le découpage en module est arbitraire
- Exemple pour la prédiction :

- Prédire ou anticiper n'est pas la même chose
- Tout n'a pas à être prédit et tout ne peut pas être prédict
- Les négociations au cas par cas demandent une interface complexe entre prédiction et planification



Les modules (prédiction incluse) sont inutiles



Modulaire vs. Bout en bout

Modulaire

Bout en bout

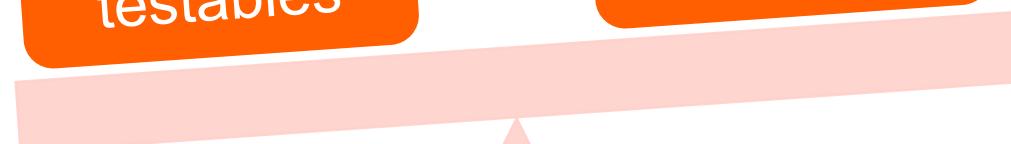
Tâches faisables

Tâches distribuées

Blocs testables

Procédé automatisé

Supérieur en théorie

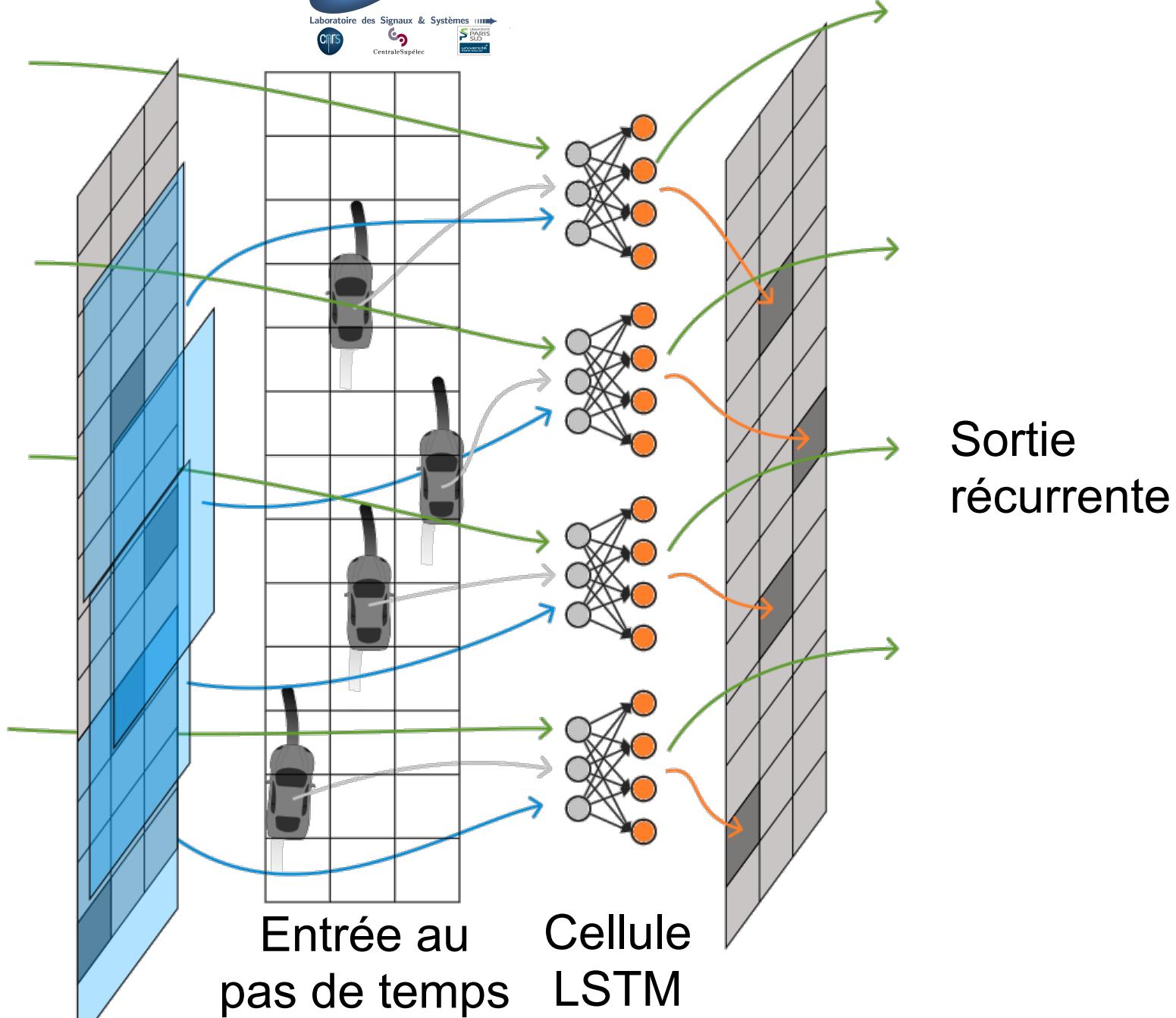




Modèles à grilles larges

Social LSTM

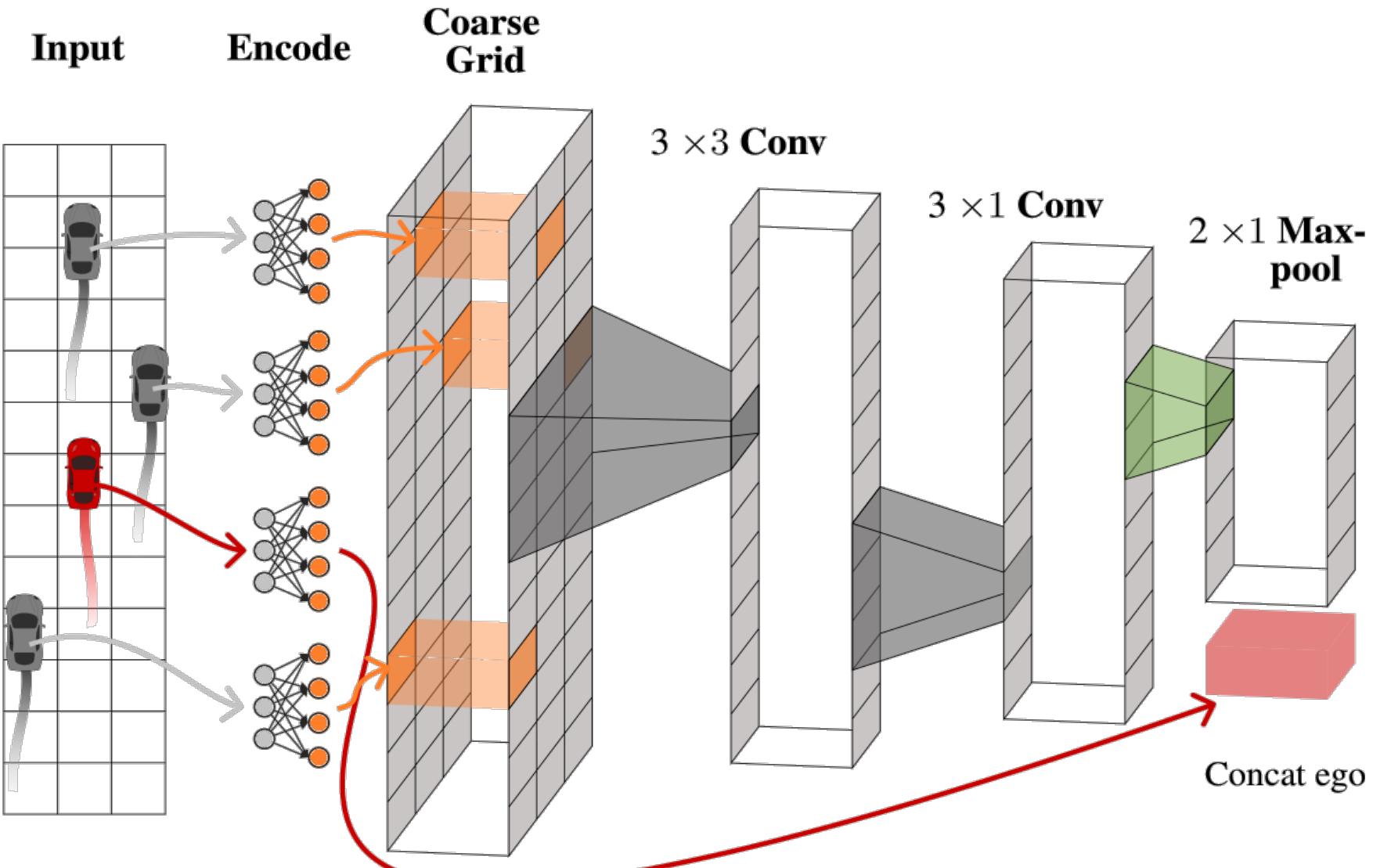
Entrée
récurrente



Entrée au
pas de temps Cellule
LSTM

Sortie
récurrente

Convolutional Social LSTM





Coûts et critères d'évaluation

Les métriques posent problème

“

Loi de Goodhart :
Quand une mesure devient un objectif, elle cesse d'être une bonne mesure.

! Aucune métrique ne mesure réellement ce que l'on veut accomplir



François Chollet 
@fchollet



If your classifier is "99% accurate", either you're using the wrong metric (a metric this high is not informative), or you have an overfitting or leakage problem.

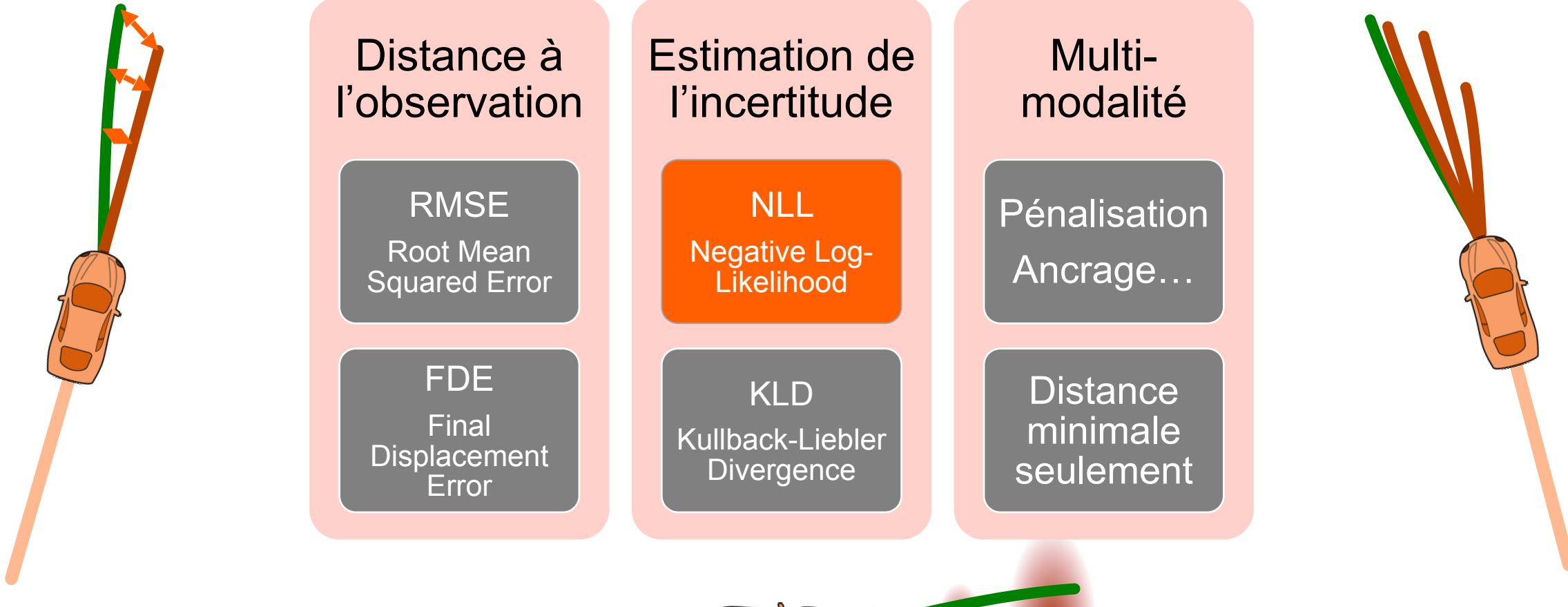
Metrics are feedback points on the way towards better models. Not trophies to show off. They should be actionable.

2 809 00:34 - 25 sept. 2019



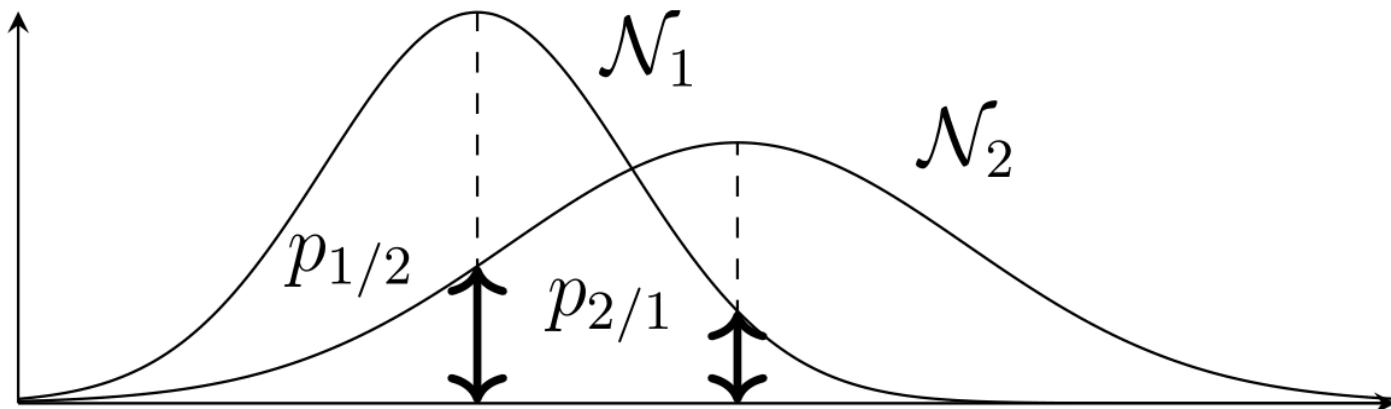
”

Chaque objectif implique sa métrique



Critère de multi-modalité additionnel

Indicateur de similarité :



$$\text{SIM}(k) = \frac{1}{n_{\text{mix}}(n_{\text{mix}} - 1)} \sum_i^{n_{\text{mix}}} \sum_{j \neq i}^{n_{\text{mix}}} p_{i/j}(k) p_{j/i}(k)$$

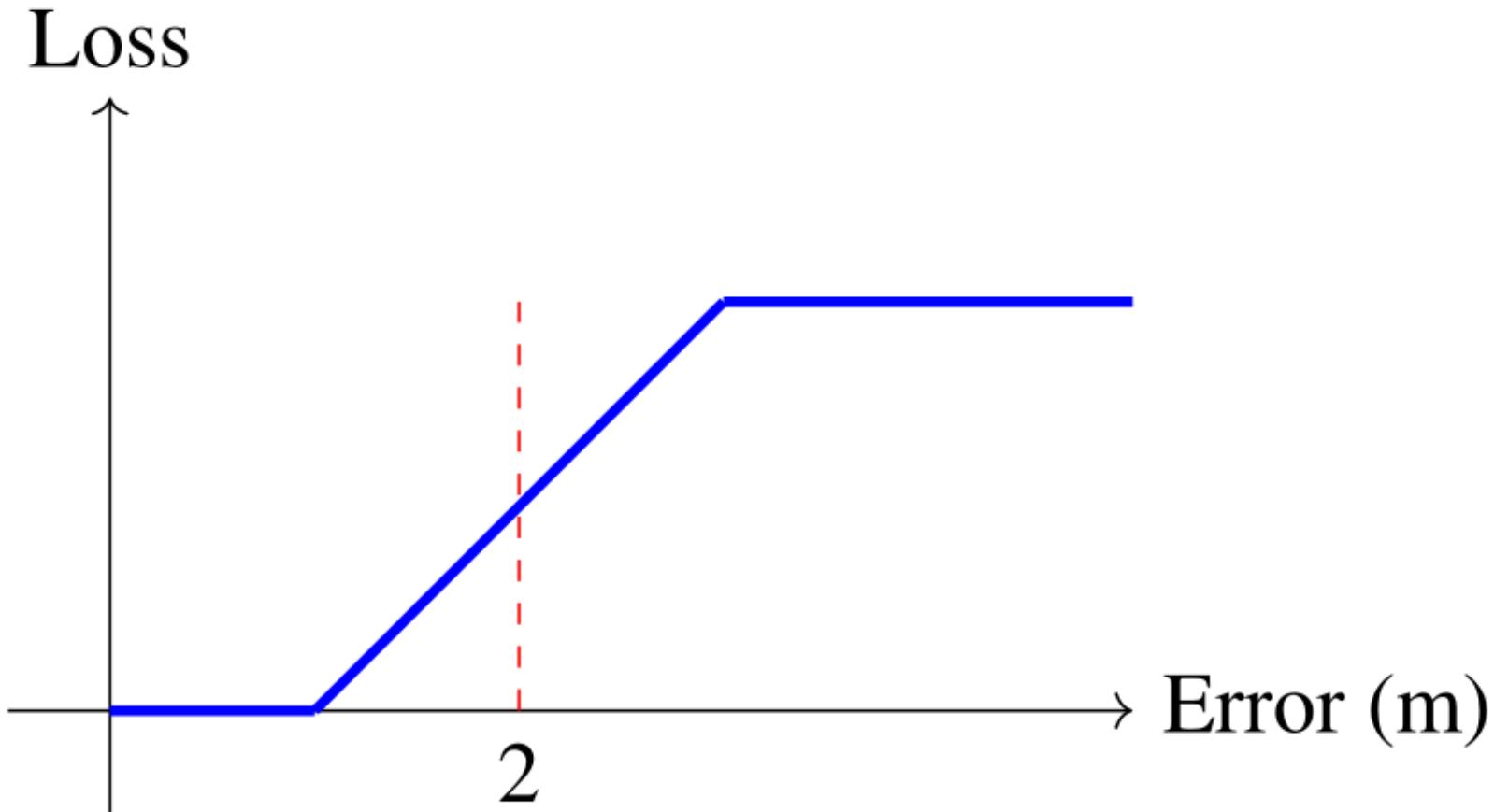


Figure 9.9: Graph of the miss loss.