Apprentissage séquentiel de compétences via la motivation intrinsèque et l'apprentissage par renforcement

Hedwin BONNAVAUD
Encadré par Arthur AUBRET et Laëtitia MATIGNON

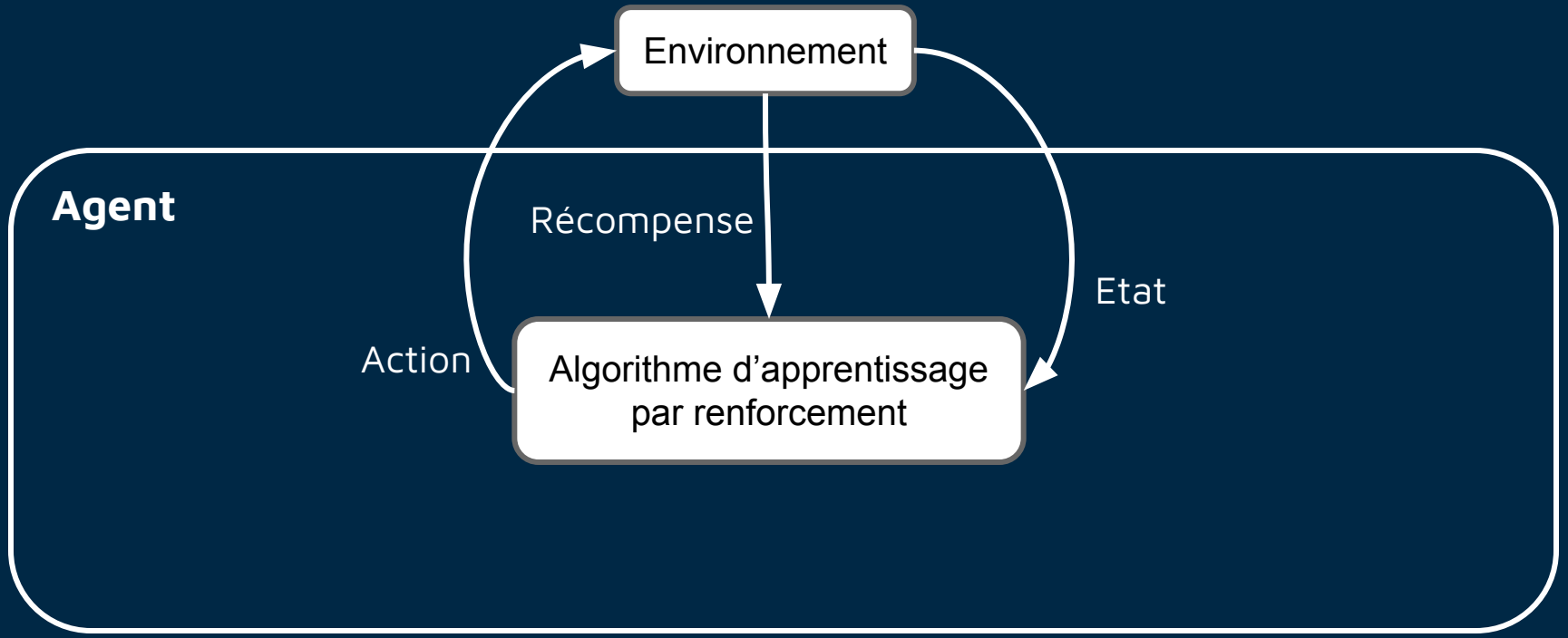
Sommaire

1. **Contexte**
   Apprentissage par renforcement, motivation intrinsèque et apprentissage de compétences
2. **Problématique**
   Problème de l'apprentissage uniforme, problème de l'apprentissage séquentiel, et comment le résoudre.
3. **Orientation de la suite du stage**
   Pistes de recherche

Contexte

Apprentissage par renforcement

Apprentissage de réalisation d'une tâche par essais-erreur.

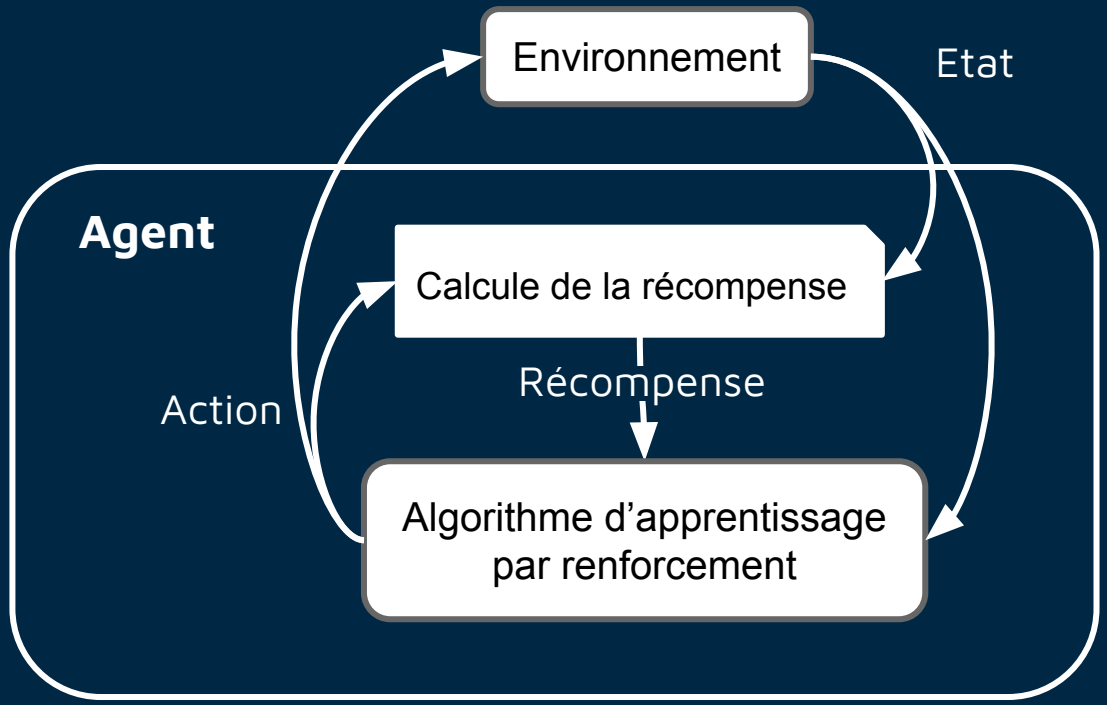


Apprentissage par motivation intrinsèque

Apprentissage de réalisation d'une tâche par essais-erreur de manière non-supervisée.

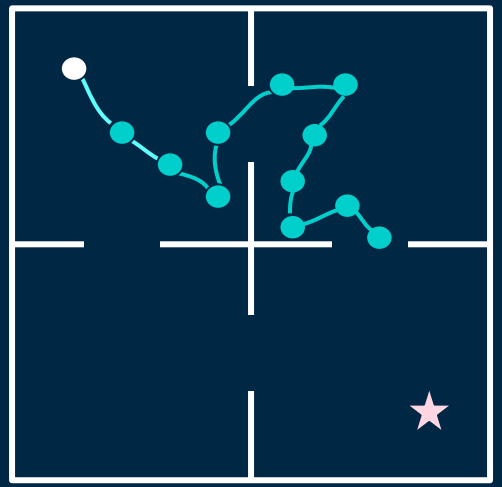
Les motivations intrinsèques s'opposent aux motivations extrinsèques dans la mesure où leur objet n'est pas la satisfaction de besoins liés à des stimuli extérieurs spécifiques, [...] mais l'attrait de certaines activités pour "elles-mêmes".

Kaplan et Oudeyer, 2007



Apprentissage de compétences

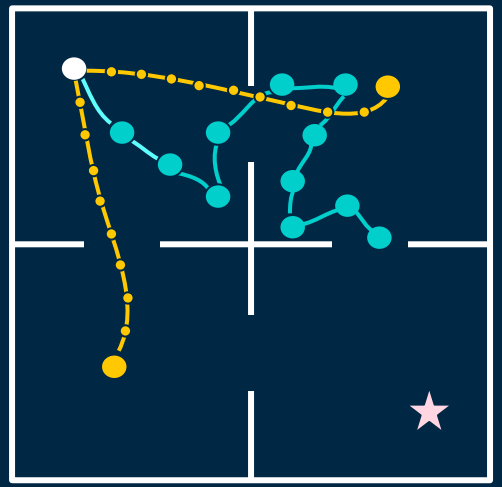
- Il peut être difficile d’atteindre un but dans un environnement complexe.



Apprentissage de compétences

- Apprendre des **compétences** permettrait de résoudre ce problème.

Une compétence est une politique réutilisable et indépendante de la la tâche à réaliser.

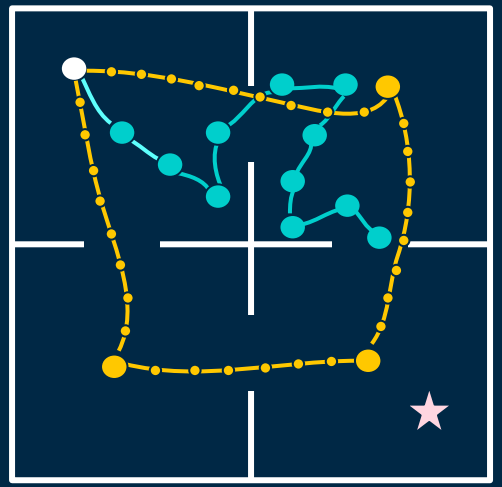


- L'exploration est simplifiée.
  - L'agent vas se rendre dans des états différents pour chaque compétences

Apprentissage de compétences

- Apprendre des **compétences** permettrait de résoudre ce problème.

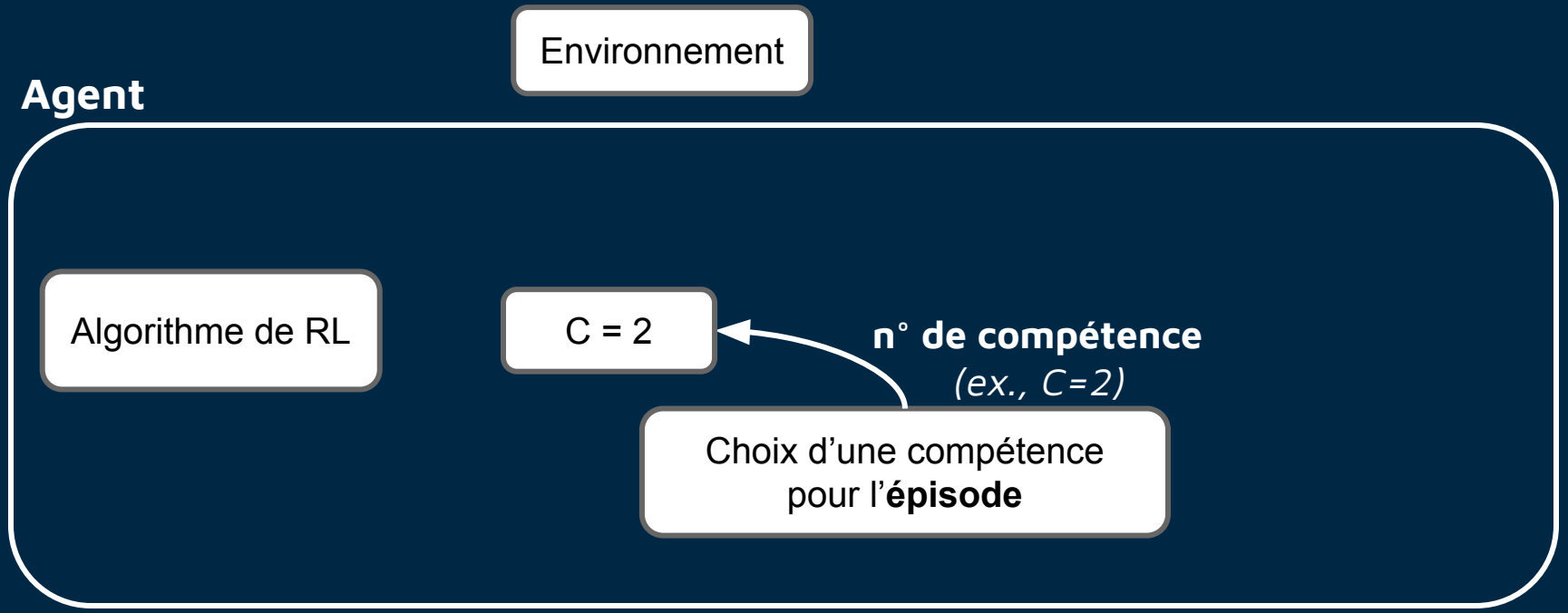
Une compétence est une politique réutilisable et indépendante de la la tâche à réaliser.



- L'exploration est simplifiée.
  - L'agent vas se rendre dans des états différents pour chaque compétences
- Credit assignment plus rapide
  - L'agent vas plus rapidement propager les valeurs associés à chaque actions.

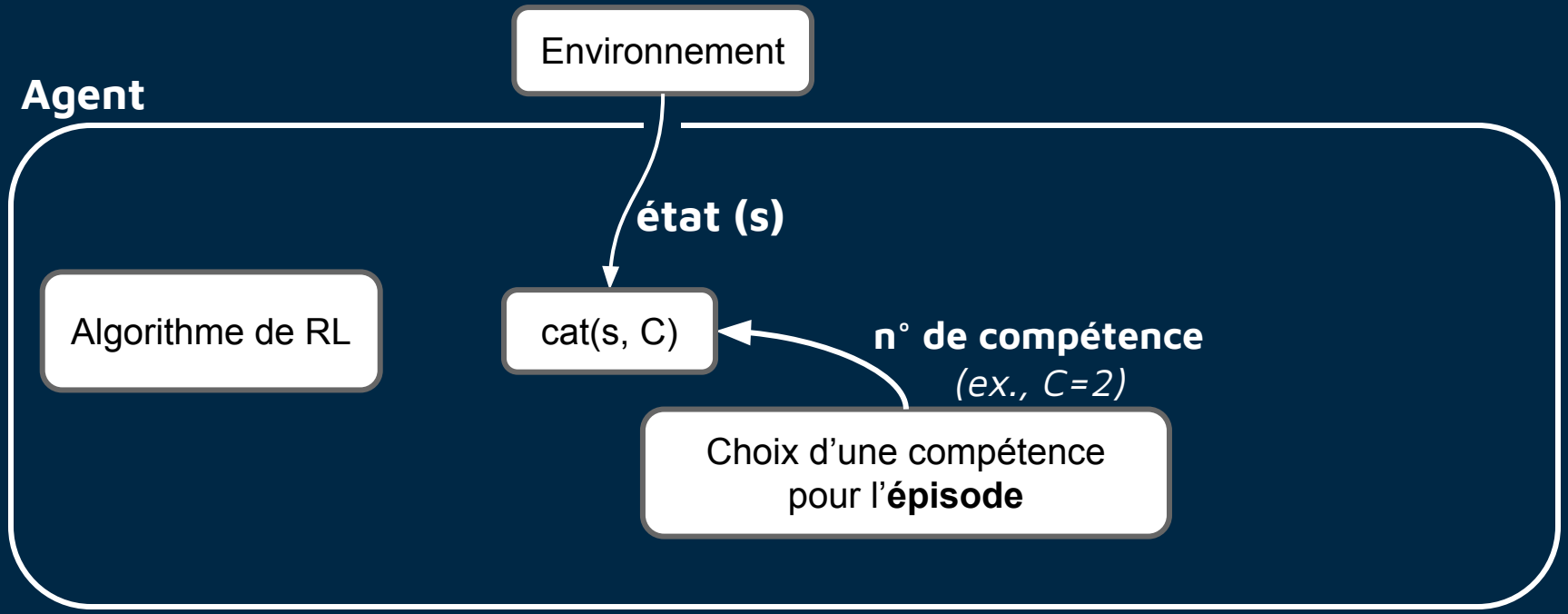
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



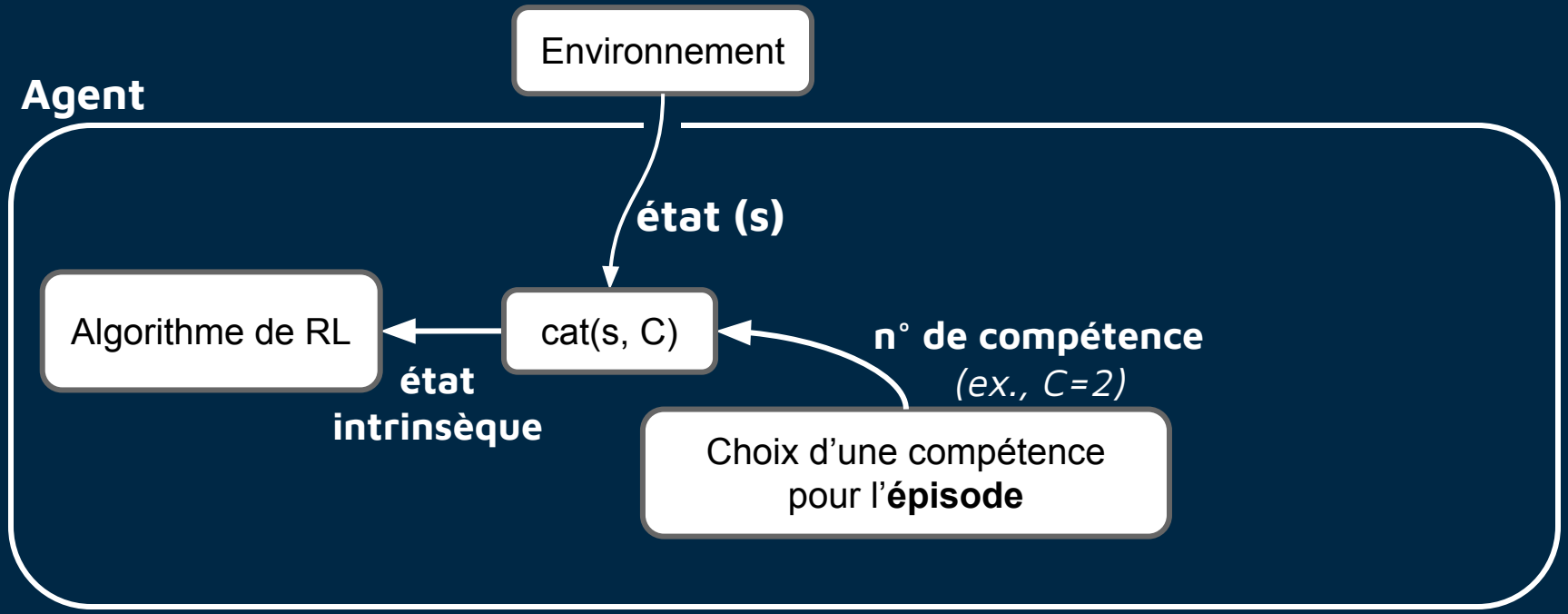
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



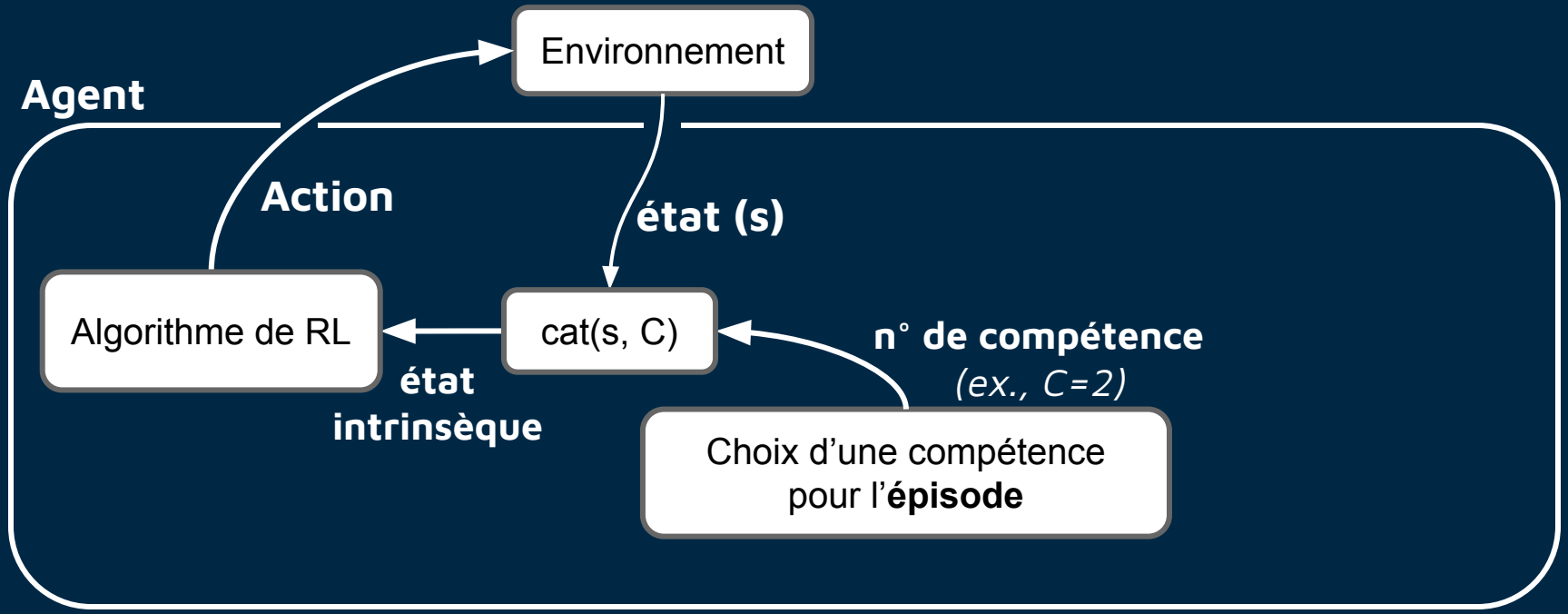
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



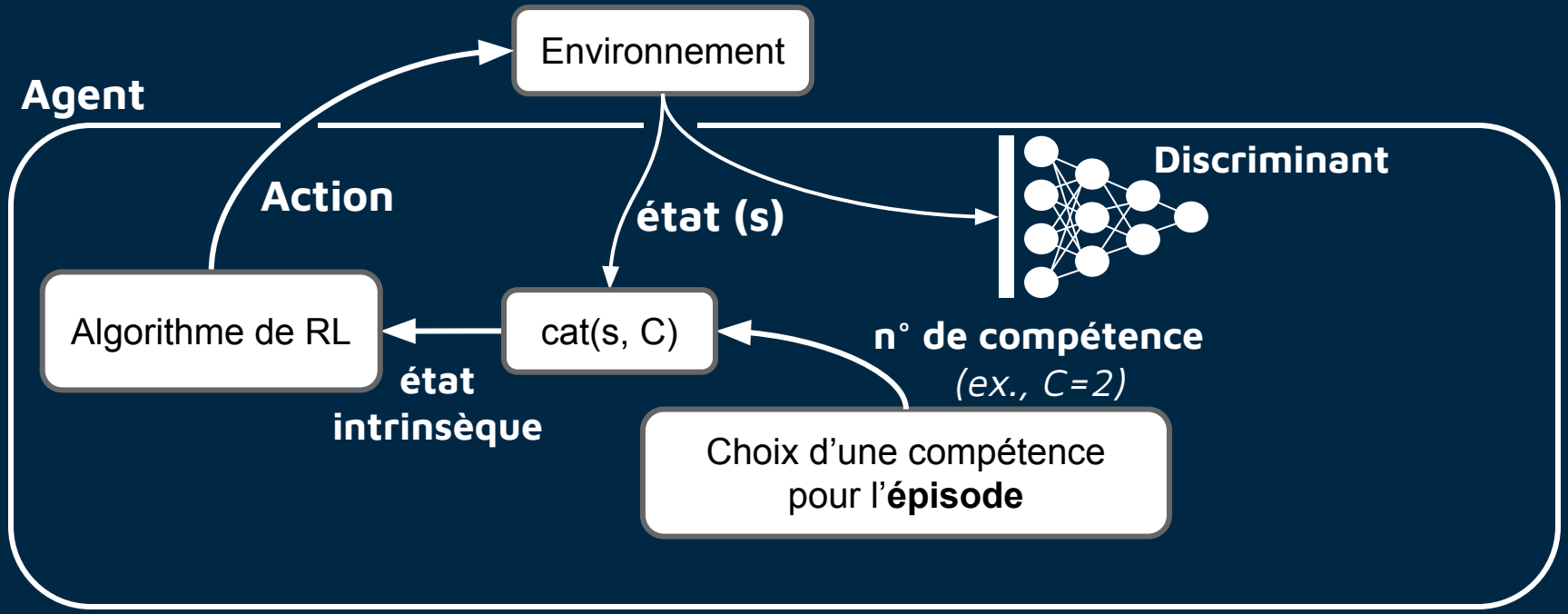
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



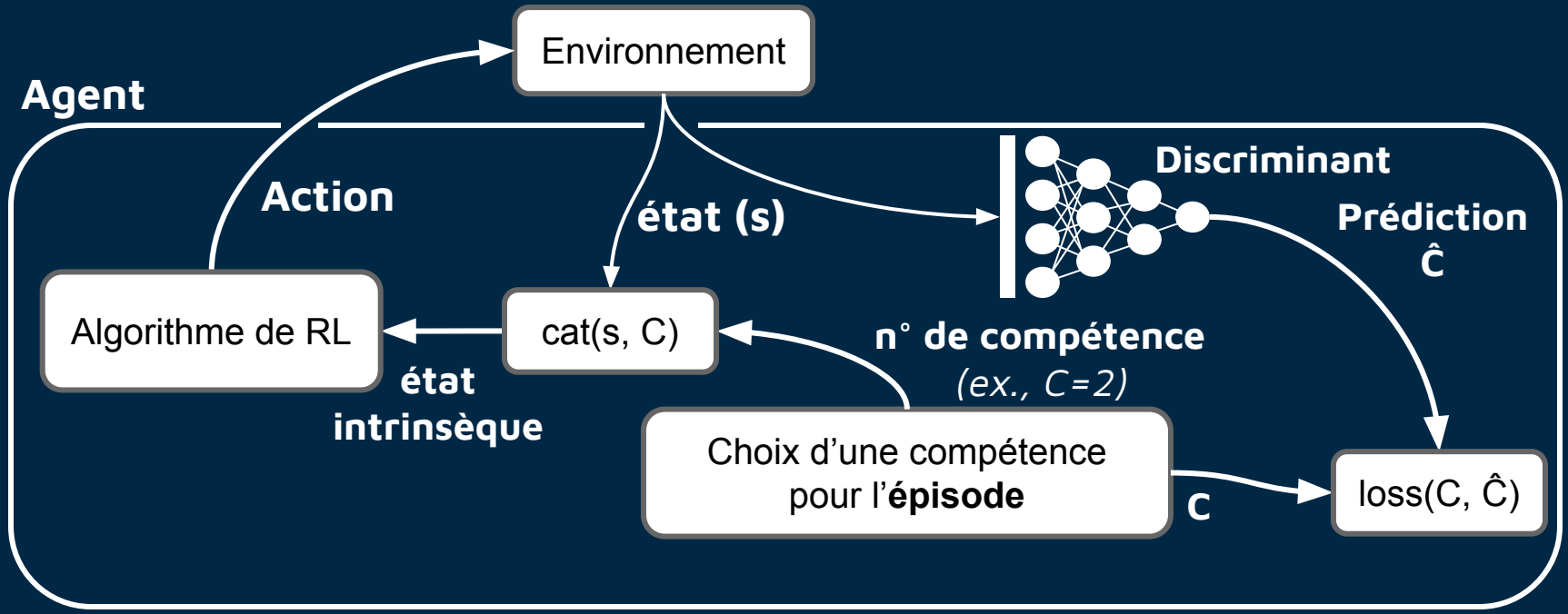
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



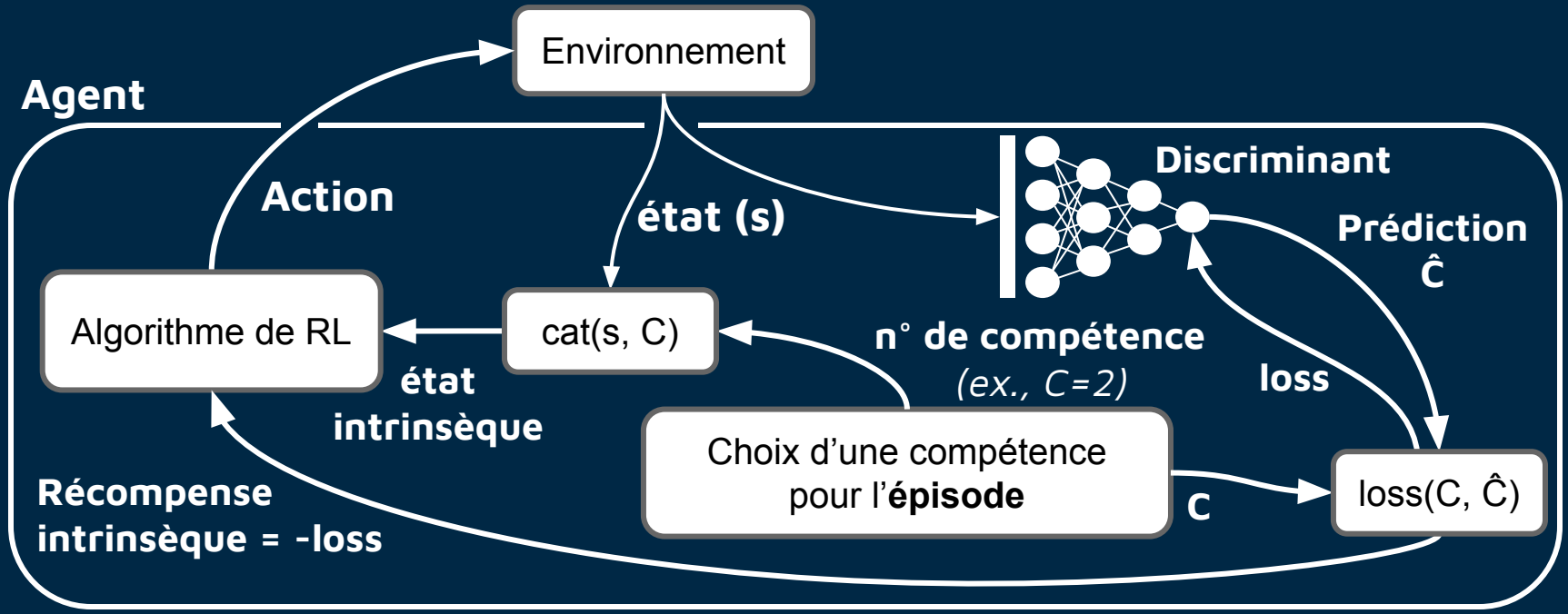
DIAYN *Eysenbach et al., 2018*

Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



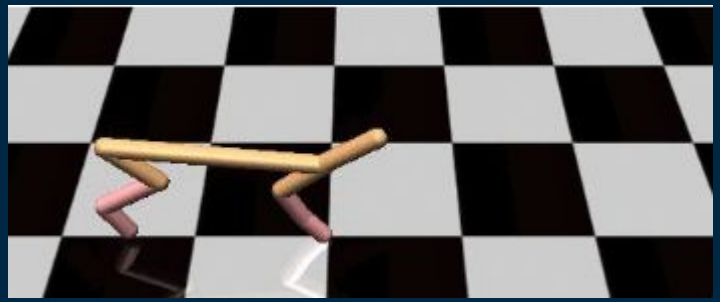
DIAYN *Eysenbach et al., 2018*

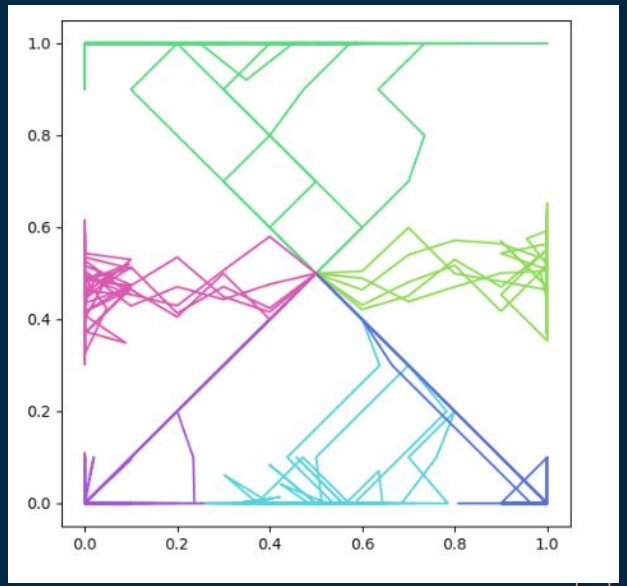
Apprentissage de compétences discrètes de manière uniforme.
L'objectif est ici d'apprendre des compétences menant l'agent dans des états qui leur sont propre.



DIAYN *Eysenbach et al., 2018*

Exemple de 6 compétences apprises par DIAYN.





IM pour l'acquisition de compétences

	Récompense *(simplifiée)*	Utilise un VAE / un discriminant
DIAYN	Max Σ I(S, C)	S > Ĉ
VIC (1)	Max I(Sf, C)	Sf > Ĉ

Légende

- I(X, Y) l'information mutuelle entre X et Y
- S un état
- Sf l'état final d'un épisode
- C une compétence
- Ĉ Version décodée / prédite de Ĉ

1. *Gregor et al., 2016*

IM pour l'acquisition de compétences

	Récompense *(simplifiée)*	Utilise un VAE / un discriminant
DIAYN	Max Σ I(S, C)	S > Ĉ
VIC (1)	Max I(Sf, C)	Sf > Ĉ
VALOR (2)	Max I(τ, C)	C > τ < Ĉ

Légende

- I(X, Y) l'information mutuelle entre X et Y
- S un état
- Sf l'état final d'un épisode
- C une compétence
- Ĉ Version décodée / prédite de Ĉ
- τ la trajectoire de l'agent

1. *Gregor et al., 2016*
2. *Achiam et al., 2018*

IM pour l'acquisition de compétences

	Récompense *(simplifiée)*	Utilise un VAE / un discriminant
DIAYN	Max Σ I(S, C)	S > Ĉ
VIC (1)	Max I(Sf, C)	Sf > Ĉ
VALOR (2)	Max I(τ, C)	C > τ < Ĉ
RIG (3)	Min (C - Cf)	S > C < Ŝ

Légende

- I(X, Y) l'information mutuelle entre X et Y
- S un état
- Sf l'état final d'un épisode
- C une compétence
- Ĉ Version décodée / prédite de Ĉ
- Ŝ Version décodée / prédite de Ŝ
- τ la trajectoire de l'agent
- Cf la compétence encodée à partir de l'état final

1. *Gregor et al., 2016*
2. *Achiam et al., 2018*
3. *Nair et al., 2018*

IM pour l'acquisition de compétences

	Récompense *(simplifiée)*	Utilise un VAE / un discriminant
DIAYN	Max Σ I(S, C)	S > Ĉ
VIC (1)	Max I(Sf, C)	Sf > Ĉ
VALOR (2)	Max I(τ, C)	C > τ < Ĉ
RIG (3)	Min (C - Cf)	S > C < Ŝ
EDL (4)	Max Σ I(S, C)	

Légende

- I(X, Y) l'information mutuelle entre X et Y
- S un état
- Sf l'état final d'un épisode
- C une compétence
- Ĉ Version décodée / prédite de Ĉ
- Ŝ Version décodée / prédite de Ŝ
- τ la trajectoire de l'agent
- Cf la compétence encodée à partir de l'état final

1. *Gregor et al., 2016*
2. *Achiam et al., 2018*
3. *Nair et al., 2018*
4. *V. Campos et al., 2020*

IM pour l'acquisition de compétences

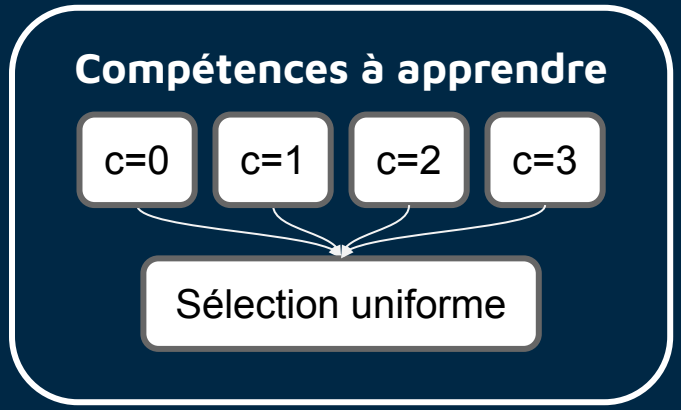
	Récompense *(simplifiée)*	Utilise un VAE / un discriminant
DIAYN	Max Σ I(S, C)	S > Ĉ
VIC (1)	Max I(Sf, C)	Sf > Ĉ
VALOR (2)	Max I(τ, C)	C > τ < Ĉ
RIG (3)	Min (C - Cf)	S > C < Ŝ
EDL (4)	Max Σ I(S, C)	
ELSIM (5)		S > Ĉ

Légende

- I(X, Y) l'information mutuelle entre X et Y
- S un état
- Sf l'état final d'un épisode
- C une compétence
- Ĉ Version décodée / prédite de Ĉ
- Ŝ Version décodée / prédite de Ŝ
- τ la trajectoire de l'agent
- Cf la compétence encodée à partir de l'état final

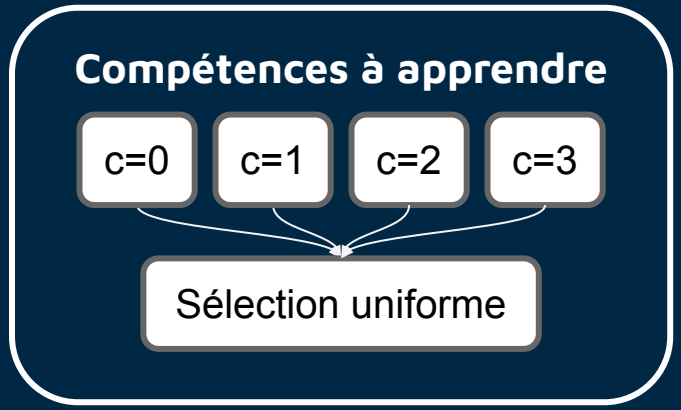
1. *Gregor et al., 2016*
2. *Achiam et al., 2018*
3. *Nair et al., 2018*
4. *V. Campos et al., 2020*
5. *A.Aubret, L.Matignon, S.Hassas, 2020*

Problématique du stage



Les compétences sont apprises de manière uniforme.

Problématique du stage



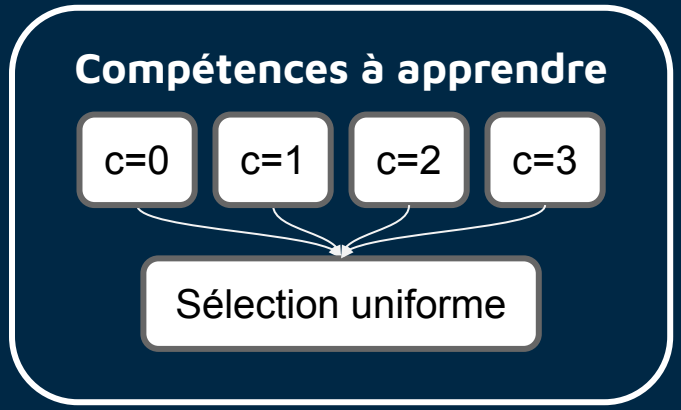
Les compétences sont apprises de manière uniforme.

Ce réapprentissage uniforme pose un problème de complexité.

- Réapprentissage inutile,
- Batch très grand pour un grand nombre de compétences
- Stockage de données

Accuracy à l'épisode t	
Compétence	Accuracy
0	0.6
1	0.99
2	0.4

Problématique du stage



Les compétences sont apprises de manière uniforme.

L'apprentissage séquentiel mène à un problème d'oubli catastrophique.

McCloskey and Cohen, 1989

Accuracy à l'épisode t	
Compétence	Accuracy
0	0.6
1	0.99
2	0.4

Ce réapprentissage uniforme pose un problème de complexité.

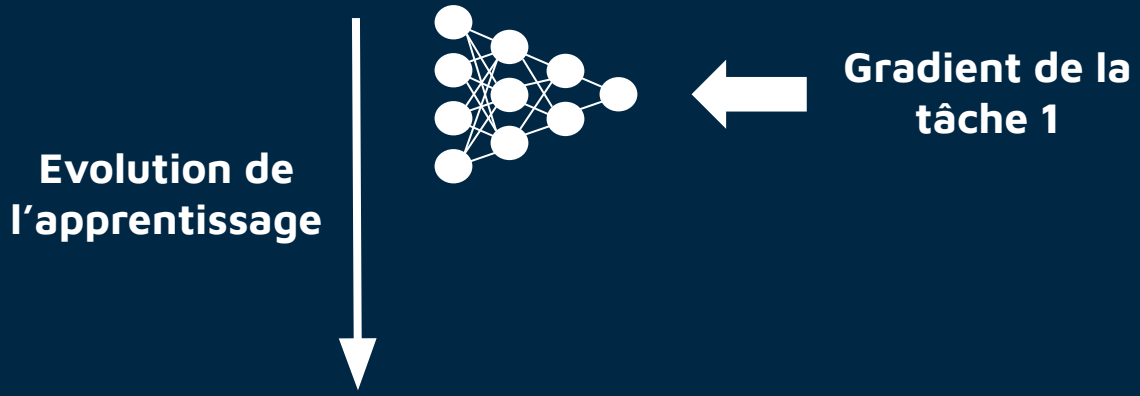
- Réapprentissage inutile,
- Batch très grand pour un grand nombre de compétences
- Stockage de données

Oubli catastrophique

Résoudre le problème de l'apprentissage séquentiel

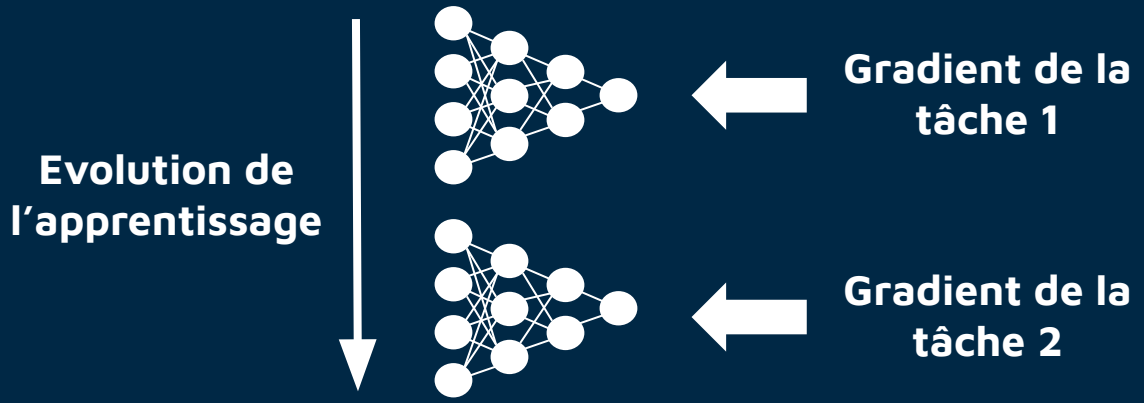
Oubli catastrophique

McCloskey and Cohen, 1989



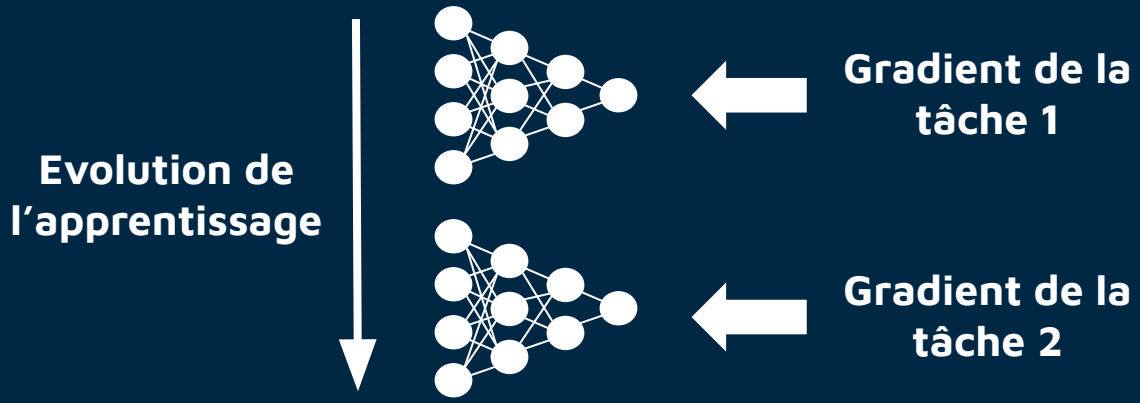
Oubli catastrophique

McCloskey and Cohen, 1989



Oubli catastrophique

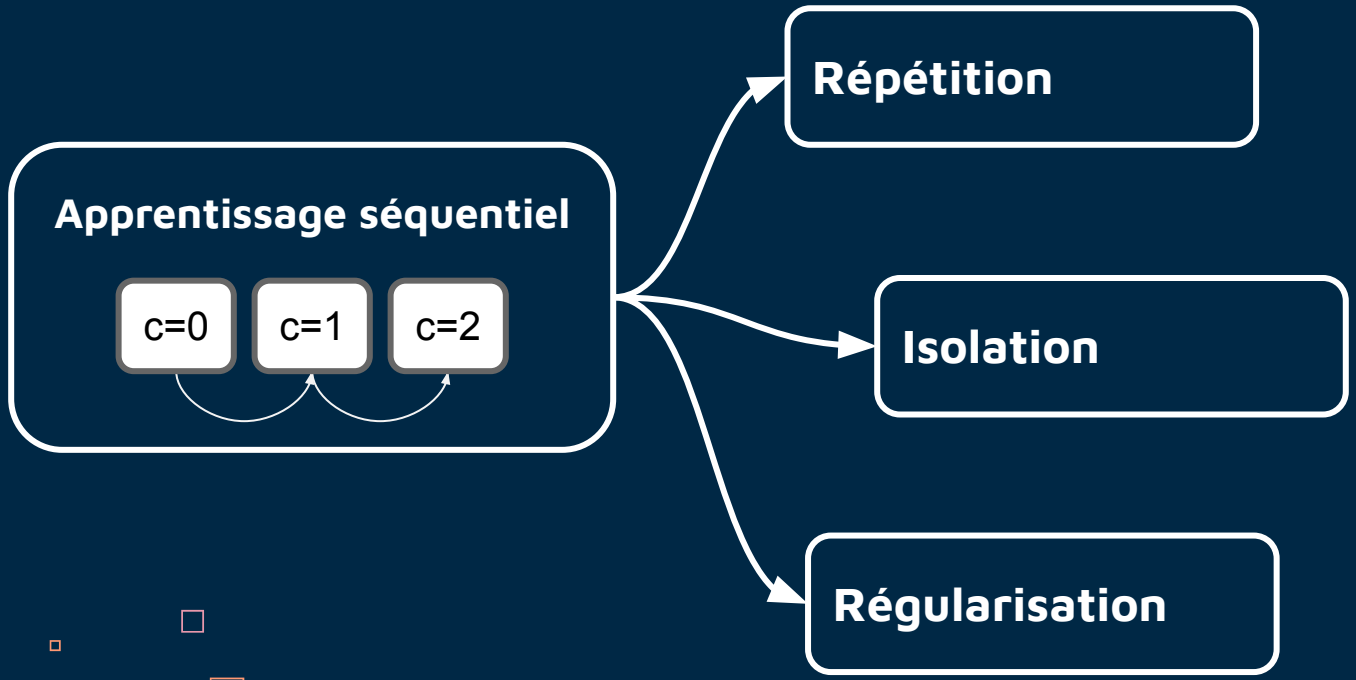
McCloskey and Cohen, 1989



Deux oublis catastrophiques en apprentissage de compétences :

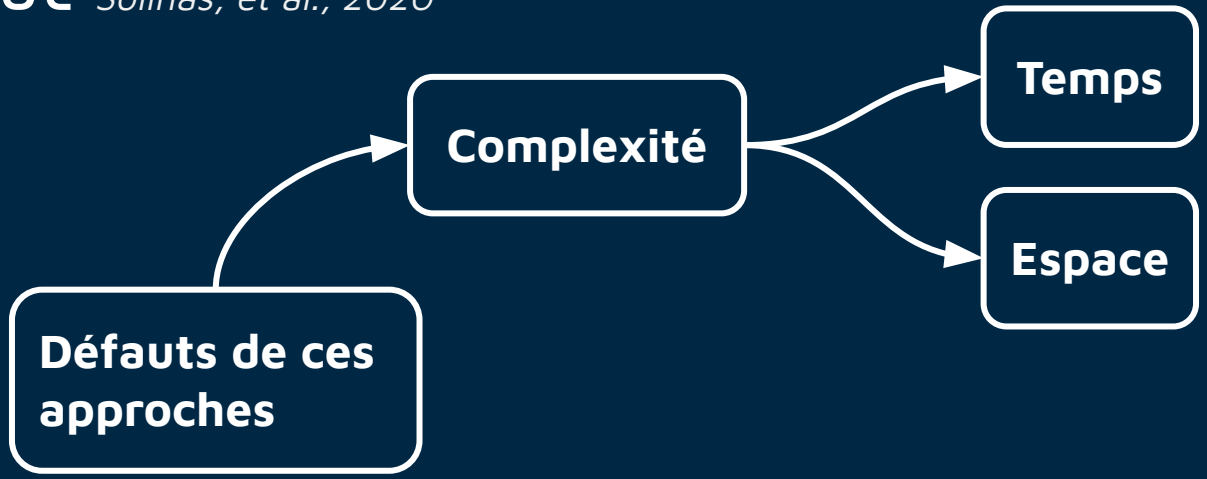
- Au niveau de l'algorithme de RL.
- Au niveau du discriminant ou du VAE.

Résoudre l'oubli catastrophique



Classification des approches inspirée de celle de Solinas, et al., 2020

Evaluation des approches résolvant l'OC *Solinas, et al., 2020*

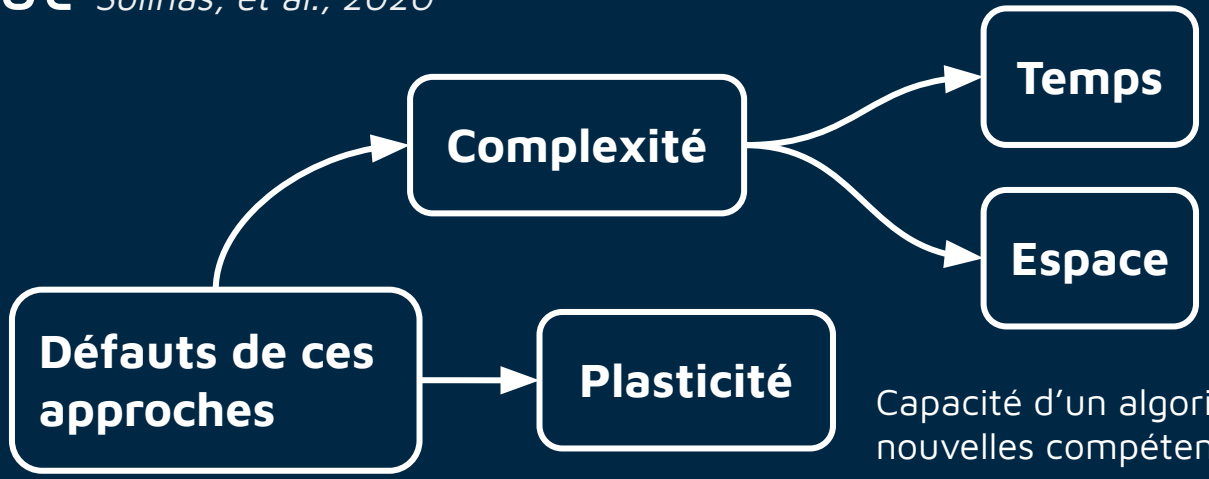


Nombre d'opérations nécessité par l'algorithme

Quantité d'informations stockées par l'algorithme

Liste des défauts inspiré de celle de Solinas, et al., 2020

Evaluation des approches résolvant l'OC *Solinas, et al., 2020*



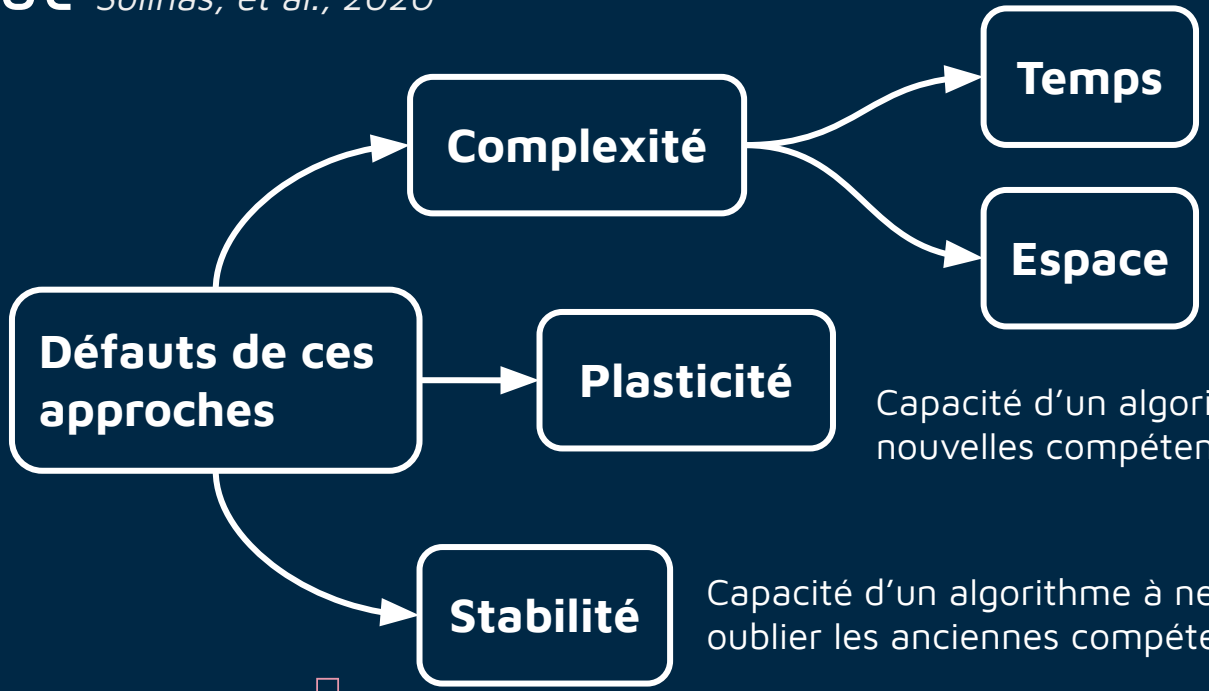
Nombre d'opérations nécessité par l'algorithme

Quantité d'informations stockées par l'algorithme

Capacité d'un algorithme à apprendre de nouvelles compétences

Liste des défauts inspiré de celle de Solinas, et al., 2020

Evaluation des approches résolvant l'OC *Solinas, et al., 2020*



Nombre d'opérations nécessité par l'algorithme

Quantité d'informations stockées par l'algorithme

Capacité d'un algorithme à apprendre de nouvelles compétences

Capacité d'un algorithme à ne pas oublier les anciennes compétences

Liste des défauts inspiré de celle de Solinas, et al., 2020

Approches par répétition

- **Curious** *Colas et al., 2019*
  - utilise le progrès d'apprentissage pour ré-apprendre de façon pertinente, en ne ré-apprenant pas les compétences maîtrisées.

- **Pseudo-répétition** *Bernard ANS, Stéphane ROUSSET - 1997*
  - Utilise des réseaux de neurones pour classifier une donnée et générer une nouvelle donnée visant à imiter la donnée en entrée.

- **Combined Replay** *M. Solinas, et al., - 2020*
  - Utilise le même principe que la pseudo-répétition mais génère de nouvelles données à partir de données réelles conservées dans un batch, et non à partir d'un bruit.

Approches de répétition

Principaux défauts des méthodes de répétition **classique**

- Complexité en temps
- Complexité en espace

	Complexité		Plasticité	Stabilité
	Temps	Espace		
Répétition classique				
CURIOUS				
Pseudo-répétition				
Combined Replay				

Approches d'isolation

- **PNN** (Progressive Neural Networks)
  - Utilise un réseaux de neurone par tâches et une connexion entre eux pour le transfert d'information inter-tâches. *Rusu et al., 2016*
- **PathNet**
  - utilisent un algo. gen. pour trouver des chemins associés à chaque tâche dans le réseau de neurones. *Fernando et al., 2017*
- **PackNet**
  - Gèle une partie des poids et les réserve pour une tâche donnée. *Mallya et al., 2018*

Approches d'isolation

Geler des poids crée un problème de **plasticité** ou de **complexité en espace** selon le caractère dynamique ou statique de la méthode

	Complexité		Plasticité	Stabilité
	Temps	Espace		
PNN				
PathNet				
PackNet				

Approches de régularisation

- **EWC** (Elastic Weight Consolidation)
  - Régule l'erreur en fonction de l'importance moyenne des poids pour les anciennes tâches. *Kirkpatrick et al., 2017*
  - **WVA** (weight velocity attenuation) Réduit le gradient au niveau d'un poid en fonction de l'activation qui le franchit. *Kutalev, 2020*

- **Weight Friction** *Gabrielle K. Liu - 2019*
  - freine l'apprentissage des poids en fonction de leur importance.

- **A-GEM** *Chaudhry et al., 2018*
  - Fait tendre le gradient de la nouvelle tâche vers le gradient correspondant à l'entraînement sur des mémoires épisodiques correspondant aux anciennes tâches.

Approches de régularisation

Caractéristiques des méthodes de régularisation

- Bonne complexité en temps
- Freine l'apprentissage

	Complexité		Plasticité	Stabilité
	Temps	Espace		
EWC				
WVA				
Weight Friction				
A-GEM				

Orientation de la suite du stage

Application à l'apprentissage de compétences

Certaines de ces méthodes peuvent être adaptés pour être appliqués à l'apprentissage de compétences

- Réduire le gradient en fonction de la complexité d'une compétence
  - Nombre d'actions qui la compose par exemple.
  - Application de EWC à l'apprentissage de compétences (AC)
- Utiliser le discriminant comme générateur de données
  - pour réapprendre les compétences passées, et sa classification pour apprendre plusieur compétences en même temps.
  - Application de Combined-Replay à l'AC
- Conserver les états dans des mémoires épisodiques selon leur compétence
  - Application de A-Gem à l'AC
- ...

Planning de fin de stage

Du 13/04 au 23/04	Etude des pistes en détail et **choix de l'une d'entre elles**.
Du 23/04 au 13/06	Développement et étude de la solution choisie.
Du 13/06 au 30/06	Selon les résultats, rédaction d'un article.

Merci de votre attention !

Avez-vous des questions ?

Références

Achiam, J., Edwards, H., Amodei, D., & Abbeel, P. (2018). Variational option discovery algorithms. arXiv preprint arXiv:1807.10299.

Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie, 320(12), 989-997.

Aubret, A., Matignon, L., & Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. arXiv preprint arXiv:1908.06976.

Aubret, A., Matignon, L., & Hassas, S. (2020). ELSIM: End-to-end learning of reusable skills through intrinsic motivation. arXiv preprint arXiv:2006.12903.

Chaudhry, A., Ranzato, M. A., Rohrbach, M., & Elhoseiny, M. (2018). Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420.

Références

Colas, C., Fournier, P., Chetouani, M., Sigaud, O., & Oudeyer, P. Y. (2019, May). CURIOUS: intrinsically motivated modular multi-goal reinforcement learning. In International conference on machine learning (pp. 1331-1340). PMLR.

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. arXiv preprint arXiv:1802.06070.

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., ... & Wierstra, D. (2017). Pathnet: Evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734.

Gregor, K., Rezende, D. J., & Wierstra, D. (2016). Variational intrinsic control. arXiv preprint arXiv:1611.07507.

Kaplan, F., & Oudeyer, P. Y. (2007). Un robot motivé pour apprendre: le role des motivations intrinseques dans le developpement sensorimoteur. Enfance, 59(1), 46-58.

Références

Liu, G. K. (2019). Weight Friction: A Simple Method to Overcome Catastrophic Forgetting and Enable Continual Learning. arXiv preprint arXiv:1908.01052.

Mallya, A., & Lazebnik, S. (2018). Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7765-7773).

Nair, A., Pong, V., Dalal, M., Bahl, S., Lin, S., & Levine, S. (2018). Visual reinforcement learning with imagined goals. arXiv preprint arXiv:1807.04742.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.

Kutalev, A. (2020). Natural Way to Overcome the Catastrophic Forgetting in Neural Networks. arXiv preprint arXiv:2005.07107.

Références

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation (Vol. 24, pp. 109-165). Academic Press.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... & Hadsell, R. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.

Solinas, M., Rousset, S., Cohendet, R., Bourrier, Y., Mainsant, M., Molnos, A., ... & Mermillod, M. (2021). Beneficial Effect of Combined Replay for Continual Learning.