

Exploration interne de la structure conceptuelle des données

Une image contenant texte

Description générée automatiquementMaxime Haurel | L3 MIASHS parcours Sciences Cognitives | Mai - Juillet 2022

Sous la direction de : Mathieu D’Aquin

# Table des matières

[Table des matières 1](#_Toc108540954)

[Introduction 2](#_Toc108540955)

[Présentation de l’entreprise 3](#_Toc108540956)

[Présentation du projet global 5](#_Toc108540957)

[Prérequis 5](#_Toc108540958)

[Enjeux de l’interprétabilité 5](#_Toc108540959)

[Description du projet 5](#_Toc108540960)

[Travaux existants 5](#_Toc108540961)

[Présentation du travail réalisé 6](#_Toc108540962)

[Science des données 6](#_Toc108540963)

[Récolte des données 6](#_Toc108540964)

[Traitement des données 7](#_Toc108540965)

[Exploration des activations du modele 8](#_Toc108540966)

[logiciel de visualisation 8](#_Toc108540967)

[Conclusion 8](#_Toc108540968)

[competences developpees 8](#_Toc108540969)

[Glossaire 9](#_Toc108540970)

[Annexes 10](#_Toc108540971)

[Remerciements 11](#_Toc108540972)

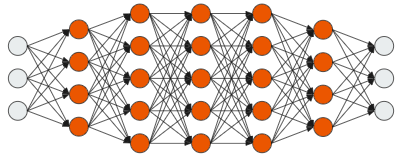
[Quatrième de couverture (changer titre) 12](#_Toc108540973)

# Introduction

Le domaine de l’intelligence artificielle (IA) est en plein essor et n’a presque jamais connu d’hiver. Le nombre de recherches en IA n’a jamais été aussi conséquent et les applications de l’IA touchent tous les domaines, de la santé à l’agriculture en passant par l’éducation.

Alors que les systèmes experts utilisant des règles bien définies étaient les systèmes rois il y a 20 ans, aujourd’hui c’est les modèles statistiques qui sont les plus utilisés dans l’industrie et les plus connus sont les réseaux de neurones artificiels. Ces modèles fonctionnent très bien pour un grand nombre de tâches et c’est la principale tendance en IA connue sous le nom de Deep Learning.

Bien que ces réseaux de neurones soient largement utilisés de partout, ils ont aussi un ensemble de désavantages dont un est **l’inexplicabilité**. En effet, les réseaux de neurones permettent d’abstraire les entrées reçues dans les couches cachées qui composent le réseau.



Ainsi, comme dans la psychologie behavioriste, on ne s’intéresse qu’aux entrées et aux sorties et les couches cachées du réseau (en orange ci-dessus) sont la boîte noire dont, à l’heure d’aujourd’hui, nous ne sommes pas en mesure de comprendre les rouages.

Ces réseaux peuvent être utilisés en production pour prédire le comportement d’un utilisateur. Mais bien que ce cas soit relativement basique, pour une utilisation dans le domaine de la santé, on aimerait savoir en détail pourquoi et comment le réseau a qualifié une radiographie pulmonaire comme laissant présager un cancer proche par exemple.

C’est donc presque dans cet intérêt que le stage s’inscrit. La vision de mon maître de stage est portée sur le comparaison des activations d’une couche cachée d’un réseau de neurones à une connaissance bien définie.  
Dans cette optique, il est possible de penser plus loin en essayant par exemple de corriger le réseau sur des individus qu’il a mal classifié.

# Présentation de l’entreprise

L’entité dans laquelle j’ai fait mon stage est le Laboratoire Lorrain de Recherche en Informatique et ses Applications, abrégé en [LORIA](https://www.loria.fr/fr/) et qui a été créé en 1997.  
Le laboratoire se situe à Villers-lès-Nancy et représente un des plus grands laboratoires en recherche informatique sur le territoire français et un des plus grands laboratoires toutes thématiques confondues dans la région lorraine.

Au sein du laboratoire, on compte 400 personnes dont la plupart sont des ingénieurs et des chercheurs faisant partie de l’Université de Lorraine, de l’INRIA et du CNRS et les recherches sont divisées en 5 départements :

* Département 1 : Algorithmique, calcul, image et géométrie
* Département 2 : Méthodes formelles
* Département 3 : Réseaux, systèmes et services
* Département 4 : Traitement automatique des langues et des connaissances
* Département 5 : Systèmes complexes, intelligence artificielle et robotique

Quant à ma position dans le laboratoire, je me place en tant que stagiaire dans une des équipes les plus récentes, [l’équipe K](https://k.loria.fr/), dont le chef d’équipe est Mathieu d’Aquin.  
L’équipe K est donc composée de :

* Mathieu D’Aquin
* Aurélie Bannay
* Jean Lieber
* Nicolas Jay
* Emmanuel Nauer
* Nicolas Lasolle (doctorant)

Cette équipe s’intéresse à l’intelligence artificielle symbolique et notamment au systèmes basés sur les connaissances dont les principaux problèmes sont de savoir comment :

* Concevoir des moteurs d’inférence
* Construire des bases de connaissances utilisées par les moteurs d’inférence

L’équipe se centre donc sur la connaissance qui donne lieu à 3 domaines de recherche principaux :

* La science des données (ainsi que la découverte de connaissances)
* L’ingénierie des connaissances
* Le raisonnement (hypothétique et déductif)

Dans le cadre de leurs recherches, les membres utilisent majoritairement le langage Python qui est le langage par référence pour traiter des données.

Mon intégration à l’équipe s’est faite naturellement puisque les centres d’intérêts de l’équipe et les miens se rejoignent. J’ai pu fréquenter d’autres stagiaires de l’équipe qui travaillaient sur des sujets assez différents du mien.

Par ailleurs, je dirais que le LORIA a été très bénéfique dans ce sens car j’ai aussi pu rencontrer des stagiaires issus d’autres équipes, souvent de thématiques très éloignées, qui m’ont donné une ouverture d’esprit sur leurs domaines de compétences en ayant échangé avec eux.

# Présentation du projet global

L’intuition scientifique de M. Mathieu d’Aquin a conduit à la création de ce stage. Cette intuition de pouvoir révéler des concepts au sein des réseaux de neurones artificiels à partir des activations des couches cachées est très inspirante et pourrait faire avancer de nombreux domaines de recherche.

## Prérequis

Dans le but de pouvoir concrétiser cette intuition, certains prérequis étaient nécessaires, le principal étant de savoir programmer dans le langage Python.  
Heureusement, je connaissais déjà le langage mais les librairies que j’ai du utiliser m’étaient souvent inconnues.  
De plus, malgré mon appétence pour le domaine de la recherche et les sciences dures de manière générale, lire des articles de recherche n’était pas aisé lorsque j’ai commencé ce stage.

## Enjeux de l’interprétabilité

## Description du projet

Il n’est pas encore possible aujourd’hui d’expliquer le pourquoi du comment d’une décision faite par un réseau de neurone, ce qui est frustrant lorsque l’on veut savoir la raison pour laquelle la radiographie A a été classifiée comme montrant une tumeur alors que la radio B ne montre pas de signe de tumeur selon le réseau.

Ce que je viens de décrire par un exemple est un vrai problème. On aimerait savoir ce qu’il se passe à l’intérieur d’un modèle de ce genre, ne serait-ce que pour expliquer au patient que l’IA est légitime de lui donner une tumeur.

Je prends volontairement un cas extrême pour donner une image forte et concrète du problème mais on peut aussi appliquer cela dans le monde du droit.  
Imaginez-vous étant jugé pour un crime que vous n’avez pas commis. Il serait injuste qu’une IA vous punisse pour quelque chose que vous n’avez pas fait n’est-ce pas ?  
Dans ce cas, on aimerait bien inspecter les profondeurs du système intelligent afin de vérifier si l’IA n’a pas fait de jugement hâtif en prenant un raccourci dans son raisonnement. Et bien c’est ce que les chercheurs en explicabilité/interprétabilité cherchent à montrer.

Pour le cas de mon stage, on aimerait savoir comment des concepts existants ou qui regroupent des éléments sont représentés au sein des réseaux de neurones.

Un nouvel exemple serait de voir si un modèle ayant été entraîné à prédire les revenus d’un film en se basant sur sa description textuelle – par exemple « Joker is a 2019 American psychological thriller film directed and produced by Todd Philips, […] winning Best Actor for Phoenix and Best Original Score for Hildur Guðbnadóttir. » - contiendrait, parmi ses activations, une représentation du concept du pays de provenance du film.

|

## Travaux existants

L’interprétabilité en IA est un domaine assez à la mode dans la communauté IA. Beaucoup d’articles sont publiés sur différentes techniques mais les articles qui nous intéressent ici sont ceux traitant de concepts de haut niveau et des activations des couches dans les systèmes d’IA numérique/stochastique.

# Présentation du travail réalisé

## Science des données

### Récolte des données

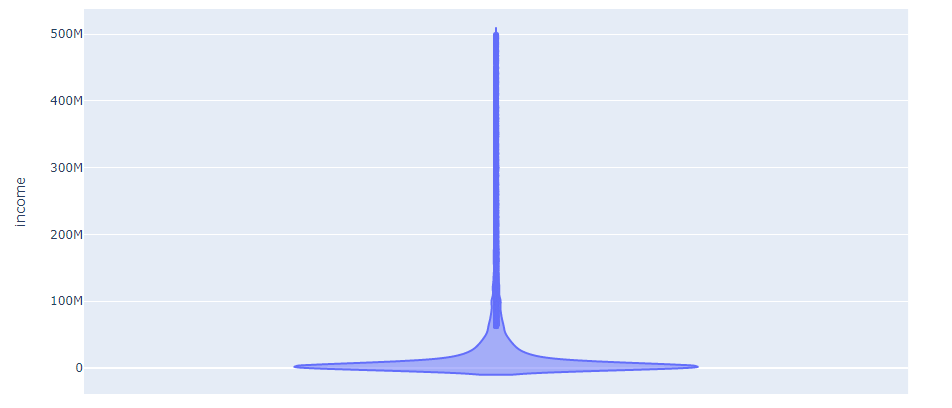
La première partie de mon stage consistait en la récolte de données. L’idée était de pouvoir récupérer des données publiques qui conviendrait au travail futur, c’est-à-dire un jeu de données sur lequel nous pourrions faire des prédictions à partir d’un attribut d’entrée ainsi qu’un autre attribut représentant des connaissances sur chaque individu.

Ainsi, le jeu de données que j’ai récolté comportait des films et pour chacun des films, sa description textuelle, ses revenus et la/les catégorie(s) à laquelle il appartient.  
Ce jeu de données nous permettait bien de pouvoir prédire à partir de la description d’un film, son revenu séparé en 3 classes :

* medium-low
* medium-high
* exceptional

Ce besoin de séparer les revenus en 3 classes distinctes relève du fait que nous ne pouvons pas prédire une valeur bien définie pour chaque film, c’est-à-dire faire une régression puisque les valeurs sont trop disparates.

[violin plot]



Ce jeu de données a été récolté par le biais de DBpedia qui est une base de données indexée sur Wikipédia donc avec beaucoup de données accessibles et ce, en faisant une requête dans le langage SPARQL.

select distinct ?film ?income ?cat ?desc where {  
?film a <http://dbpedia.org/ontology/Film> .  
?film <http://dbpedia.org/ontology/gross> ?income .  
?film <http://dbpedia.org/ontology/abstract> ?desc .  
 {  
 ?film <http://purl.org/dc/terms/subject> ?cat .   
 } UNION {  
 ?film <http://purl.org/dc/terms/subject> ?scat .   
 ?scat <http://www.w3.org/2004/02/skos/core#broader> ?cat.  
 }   
 filter (lang(?desc) = "en")  
 filter (lang(?film) = "en")  
} group by ?film ?cat ?desc LIMIT 3 OFFSET 100000

### Traitement des données

Les données étant récupérés sur DBpedia, elles ne sont pas prêtes à l’emploi. Elles ont besoin d’être traitées de manière réfléchie pour que le modèle les interprète correctement.

C’est notamment le cas des descriptions qui sont au format texte. On ne peut pas donner de texte à un réseau de neurones ou à tout autre modèle de Machine Learning puisque ces modèles sont avant tout des modèles statistiques donc qui fonctionnent avec des entrées numériques. Alors, afin de nettoyer le texte, j’ai appliqué des méthodes courantes de traitement automatique des langues comme :

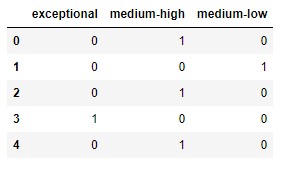
* Transformer chaque lettre en minuscule
* Enlever la ponctuation, les liens, les entités numériques, ainsi que les mots très fréquents (stopwords) du genre *the, a, been, ...*.

Précision : le jeu de données étant en anglais, les mots très fréquents sont évidemment les mots fréquents en langue anglaise.

Ensuite, j’ai dû remanier les revenus pour créer les classes de revenus que j’ai évoqué plus tôt. Puisque le modèle ne fonctionne qu’avec des valeurs numériques, les classes de revenus *(medium-low, medium-high et exceptional)* ont été encodé de la manière suivante :

* On crée une colonne dans le jeu de données pour chaque classe de revenus
* Pour chaque film et chaque colonne des classes de revenus on place un 0 si le film ne fait pas partie de cette classe, ou un 1 si justement le film fait partie de cette classe de revenus

On a alors un tableau présenté tel que :



[...]

## Exploration des activations du modele

La direction de la recherche de la présence de concepts au sein même du réseau de neurones a été orientée par l’intuition que les activations d’une couche cachée d’un réseau pouvaient être porteuses d’un ou de plusieurs concept(s).

Il n’existe pas de fonction préconstruite pour afficher les activations d’une couche d’un modèle au sein des librairies Tensorflow et Keras. Et pour cause, il suffit de reconstruire le réseau jusqu’à la couche que l’on veut explorer et on re-exécute des prédictions avec les entrées originales et on obtient les activations.

De cette manière il semble possible d’automatiser très facilement ce processus d’obtention des activations pour chaque couche et ce peu importe le modèle tant que l’on a accès aux données qui ont servies à son entraînement.

## logiciel de visualisation

# Conclusion

## competences developpees

# Glossaire

**Intelligence artificielle symbolique** : sorte d’IA se basant sur des règles / connaissances en imitant le raisonnement humain dans le but de produire des décisions.

**Système à base de connaissance** : programme utilisant une base de connaissances afin de résoudre des problèmes. Ce genre de système est en général composé d’au moins un moteur d’inférence et au moins une base de connaissances.

# Annexes

# Remerciements

Je tiens à remercier chaleureusement M. Mathieu d’Aquin qui m’a permis de réaliser ma première expérience de recherche dans une entité que j’affectionne tout particulièrement.

# Quatrième de couverture (changer titre)

* Mathieu D’Aquin : [mathieu.daquin@loria.fr](mailto:mathieu.daquin@loria.fr)